ITUPublications

International Telecommunication Union
Standardization Sector

ITU-T Focus Group Technical Report

(11/2023)

Focus Group on Artificial Intelligence for Natural Disaster Management

Innovative approaches to natural disaster management: Leveraging AI for data-related processes

ITU-T FG-AI4NDM-Data Technical Report

Innovative approaches to natural disaster management: Leveraging AI for data-related processes

Summary

This Technical Report focuses on the standardization of data-related processes, including but not limited to vocabulary, data custodianship, acquisition, and management; data supply chains; data curation and delivery; and data processing for AI/ML applications within the domain of AI for natural disaster management.

Acknowledgement

This Technical Report was prepared under the leadership of Ms Monique Kuglitsch, Chair of FG-AI4NDM (Fraunhofer HHI, Germany) and Mr Rustem Arif Albayrak WG-Data Co-Chair (NASA Goddard Space Flight Center, United States) and Ms Allison Craddock WG-Data Co-Chair (NASA Jet Propulsion Laboratory, United States).

It is based on the contributions of various authors who participated in the Focus Group activities.

Mr Rustem Arif Albayrak WG-Data Co-Chair (NASA Goddard Space Flight Center, United States) and Ms Allison Craddock WG-Data Co-Chair (NASA Jet Propulsion Laboratory, United States) served as the main Editors of this Technical Report. Ms Mythili Menon (FG-AI4NDM Advisor) and Ms Hiba Tahawi (FG-AI4NDM Assistant) served as the FG-AI4NDM Secretariat.

Keywords

Artificial Intelligence, data management, data processing, disaster management cycle, disaster recover, disaster response, Internet of Things, natural disasters, natural hazards, standards.

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Editors:	Rustem Arif Albayrak NASA GSFC United States	Email: rustem.a.albayrak@nasa.gov
	Allison Craddock NASA JPL United States	Email: allison.b.craddock@jpl.nasa.gov

Contributors:
(in alphabetic order)

A. Johnson
Los Alamos National Laboratory, U.S.A.

Ahmad Wani
One Concern, U.S.A

Alec van Herwijnen
SLF, Switzerland

Alejandro Marti
Mitiga Solutions & National Supercomputing Center, Spain

Alexandra Moutinho
Universidade de Lisboa, Portugal

Allison Craddock
NASA Jet Propulsion Laboratory/California Institute of Technology, U.S.A.

Anais Couasnon
Vrije Universiteit Amsterdam, the Netherlands

Anugandula Naveen Kumar
Centre for Development of Telematics, India

Attila Komjathy
NASA Jet Propulsion Laboratory/California Institute of Technology, U.S.A.

Chet Karwatowski
IBM, U.S.A.

Christopher W. Johnson
Los Alamos National Laboratory, U.S.A.

Constantinos Heracleous
KIOS CoE, Cyprus

Corentin Caudron
ISTerre, France

David Grzan
University of California, Davis, U.S.A.

Edier Aristizábal
Universidad Nacional de Colombia, Colombia

Elfatih Mohamed Abdel-Rahman
International Centre of Insect Physiology and Ecology, Kenya

Emily Kimathi
International Centre of Insect Physiology and Ecology, Kenya

Eren Erman Ozguven
Florida State University, U.S.A.

Fernando Pech-May
Instituto Tecnológico Superior de los Ríos,
Mexico

Gonrou Dobou Orou Berme Herve Yamagata
University, Japan

Guillermo Cortés
University of Granada, Spain

Guy Schumann
RSS-Hydro, Luxembourg

Ha Trang Nguyen
Yamagata University, Japan

Helen Li
CAICT, MITT, China

Henri E. Z. Tonnang
International Centre of Insect Physiology and
Ecology, Kenya

Hideo Imanaka
NICT, Japan

Jil Christensen
Day One Relief, U.S.A.

Joe Paluska
One Concern, U.S.A

Joger Magaña Govea
Instituto Tecnológico Superior de los Ríos,
Mexico

John Rundle
University of California, Davis, U.S.A

Juan Pablo Ospina
Universidad Nacional de Colombia,
Colombia

Kiyonori Ohtake
NICT, Japan

Larry Lopez
Yamagata University, Japan

Maria João Sousa
IDMEC, Instituto Superior Técnico,
Universidade de Lisboa, Portugal

Maria Michaelopoulou
KIOS CoE , Cyprus

Marius Kriegerowski
QuakeSaver GmbH, Germany

Masugi Inoue
NICT, Japan

Michele Gazzea
Western Norway University, Norway

Miguel Almeida
ADAI, University of Coimbra, Portugal

Noel Garcia Diaz
Instituto Tecnológico de Colima, Mexico

Om Prakash Kumar
Manipal Institute of Technology, Manipal
Academy of Higher Education, India

Panayiotis Kolios
KIOS CoE, Cyprus

Pankaj Kumar Dalela
Centre for Development of Telematics, India

Pantelis Georgiadis
The Cyprus Institute, Cyprus

Rajkumar Upadhyay
Centre for Development of Telematics, India

Raul Aquino
Universidad de Colima, Mexico

Reza Arghandeh
Western Norway University & StormGeo AS,
Norway

Rinku Kanwar
IBM, U.S.A.

Sameena Pathan
Manipal Institute of Technology, Manipal
Academy of Higher Education, India

Sandeep Sharma
Centre for Development of Telematics, India

Saurabh Basu
Centre for Development of Telematics, India

Shweta Vincent
Manipal Institute of Technology, Manipal
Academy of Higher Education, India

Silvia García
Universidad Nacional Autónoma de México,
Mexico

Simon Horton
Avalanche Canada, Canada

Thomas Ward
IBM, United States

Tobias Leidemer
Leibniz Universität Hannover, Germany

Toshiaki Kuri
NICT, Japan

Tsutomu Nagatsuma
NICT, Japan

Valentino Constantinou
NASA Jet Propulsion Laboratory/California
Institute of Technology, U.S.A.

Yago Diez
Yamagata University, Japan

Zack Spica
University of Michigan, U.S.A.

© ITU 2024

Some rights reserved. This publication is available under the Creative Commons Attribution-Non Commercial-Share Alike 3.0 IGO licence (CC BY-NC-SA 3.0 IGO; https://creativecommons.org/licenses/by-nc-sa/3.0/igo).

For any uses of this publication that are not included in this licence, please seek permission from ITU by contacting TSBmail@itu.int.

Table of Contents

Page

ITU-T FG-AI4NDM Technical Report

Innovative approaches to natural disaster management: Leveraging AI for data-related processes

Summary

This Technical Report provides a comprehensive overview of the contributions made under the ITU/WMO/UNEP Focus Group on Artificial Intelligence for Natural Disaster Management (FG-AI4NDM) and proposes a structure for standardizing artificial intelligence (AI)/ machine learning (ML) data-related processes. The report focuses on best practices for collecting, monitoring, and managing data for AI/ML applications in the domain of natural disaster management. It explores the data requirements, potential issues during data collection, and how AI algorithms can be used to enhance data quantity and quality. Additionally, the report covers data archiving, operational workflows, AI/ML and data standards, policy, ethics, legal issues and open data. Its objective is to provide preliminary information that can be used as guidance for standards or "good practices."

The report discusses the use of AI/ML applications in natural disaster management and the data requirements for their implementation. Section 5 highlights current and future technologies to capture the future data formats that will affect the data processing and handling. Some of the future technologies considered are segmentation algorithms, intelligent satellites and drone mapping, which can be used for disaster response. The impact of Internet of things (IoT), blockchain, and digital twin technologies on disaster management is also discussed. This section also brings up the question of what role data plays in combining IoT with blockchain technology for AI/ML in the domain of disaster management.

This report emphasizes two phases of data handling in AI/ML related applications: static and dynamic. Section 6 focuses on the management of static data related to AI/ML datasets and the use of data themes, drawing on the United Nations Global Fundamental Geospatial Data Themes as the foundation to support global geospatial information management. Data themes are a minimum set of concepts used to label datasets, which enables interoperability between different groups, companies, organizations and nations. Using an ML-based information management system, data themes describe the approach in each use case or study. In addition to the data themes concept, section 6 emphasizes the data management throughout the product life cycle, not just at the beginning or during product development. After the data management practices, the report further highlights the importance of data custodianship and management, with a focus on data accessibility through visibility, interoperability, and usability, and it provides best practices and relevant use cases for each of these topics. The data management section emphasizes the significance of metadata and documentation for effective data management and dissemination. Furthermore, this section covers two main topics related to data management: data supply chains (DSC) and data curation and delivery. DSC refers to the procedures and activities that enable the flow of data through an organization. In contrast, data curation and delivery involve maintaining the value of data and delivering it to end-users. This section provides best practices for both topics and refers to two relevant use cases: an intelligent big data analysis system for wildfire management and the utilization of AI and probabilistic modeling for strategic resilience. This section, furthermore, stresses the need to formalize and streamline data supply chains while maintaining data security and privacy in data curation and delivery.

In section 7 of the report, the importance of machine learning operations (MLOps) is discussed. MLOps is crucial in capturing the dynamic flow of data and the life cycle of topic groups. The report explains the different components of an MLOps software stack, which include data sources, AI models, pipelines and software containers. It also touches upon the 5Vs of big data for AI/ML, which are volume, velocity, variety, veracity and value. These characteristics are essential for researchers to manage large or complicated data volumes and extract information systematically. It also contains

information about data preparation and integration, data pre-processing and exploration, and temporal data processing requirements.

Moreover, section 7.2 highlights the importance of big data indexing, which is used to fragment datasets according to frequently used query criteria. Indexes are persistent data structures that store data in a compressed form for rapid processing by machine learning systems. Additionally, the subsections emphasize the importance of looking into datasets before drawing statistical conclusions and stress the need to perform data transforms like normalization and standardization to aid machine learning models.

Following normalization and standardization, section 7.4 covers the pre-processing steps required for handling temporal data, including handling missing values, detecting and handling outliers, and transforming data. It also points to two use cases to support the ideas: a flash flooding monitoring system in Mexico and artificial intelligence modeling tools for monitoring desert locusts.

Section 7.6 goes on to discuss the importance of data validation in ensuring data quality and avoiding errors and vulnerabilities. It also outlines best practices for data validation in various applications and highlights the challenges involved in validating AI systems, including deep learning and traditional machine learning models. This section suggests best practices for validating correlated time series and geospatial data, online AI systems, and explains the importance of explainable and interpretable AI systems.

One of the major usage concerns for AI and ML techniques is bias. As a result, this technical report analyses the concept of bias in AI tools and the importance of understanding datasets well to avoid harmful outcomes with ethical implications throughout to section 7.8. It also covers different types of data bias and how to identify and mitigate them.

Sections 9, 10 and 11 focus on a) data standards, b) ethics and c) open source/open data concepts, respectively. Section 9.1 refers to the Open Geospatial Consortium (OGC), which is a non-profit standards organization dedicated to the industry standardization of geographic information, and its efforts to develop new data standards for sharing Earth Observation (EO) machine learning datasets. Section 10 concludes by highlighting the regulatory, ethical, and legal issues that arise in the creation and use of responsible AI tools and products, emphasizing the need for diverse and cross-cultural considerations in policy making.

Finally, section 12 discusses the structure and acquisition process of the use cases that were used to develop best practices in the field of AI for natural disasters. The use cases were obtained through an open call for proposals and were presented at focus group meetings. The proponents were asked to provide a project summary, project plan, outline of milestones and description of impacts. In total, 31 use cases were adopted and the proponents of these use cases were asked to complete a detailed questionnaire related to the topics covered by the three working group technical reports. Of these 31 use cases, 27 completed questionnaires were received. The final section provides an excerpt from the original responses to these questionnaires.

In conclusion, the rapid advancement of digital technology has significantly impacted how we collect and interpret data. Since the data can be intricate, it is crucial to ensure that the collection process is accurate to avoid biases. Integrating such datasets with advanced AI/ML techniques might result in further biases, rendering the outcomes unreliable for decision-making (garbage in, garbage out).

In short – there is no room for error in disaster management – as the stakes are incredibly high. Therefore, data selection and processing must be carefully executed, and disaster response frameworks must be meticulously tailored to ensure reliability and accuracy.

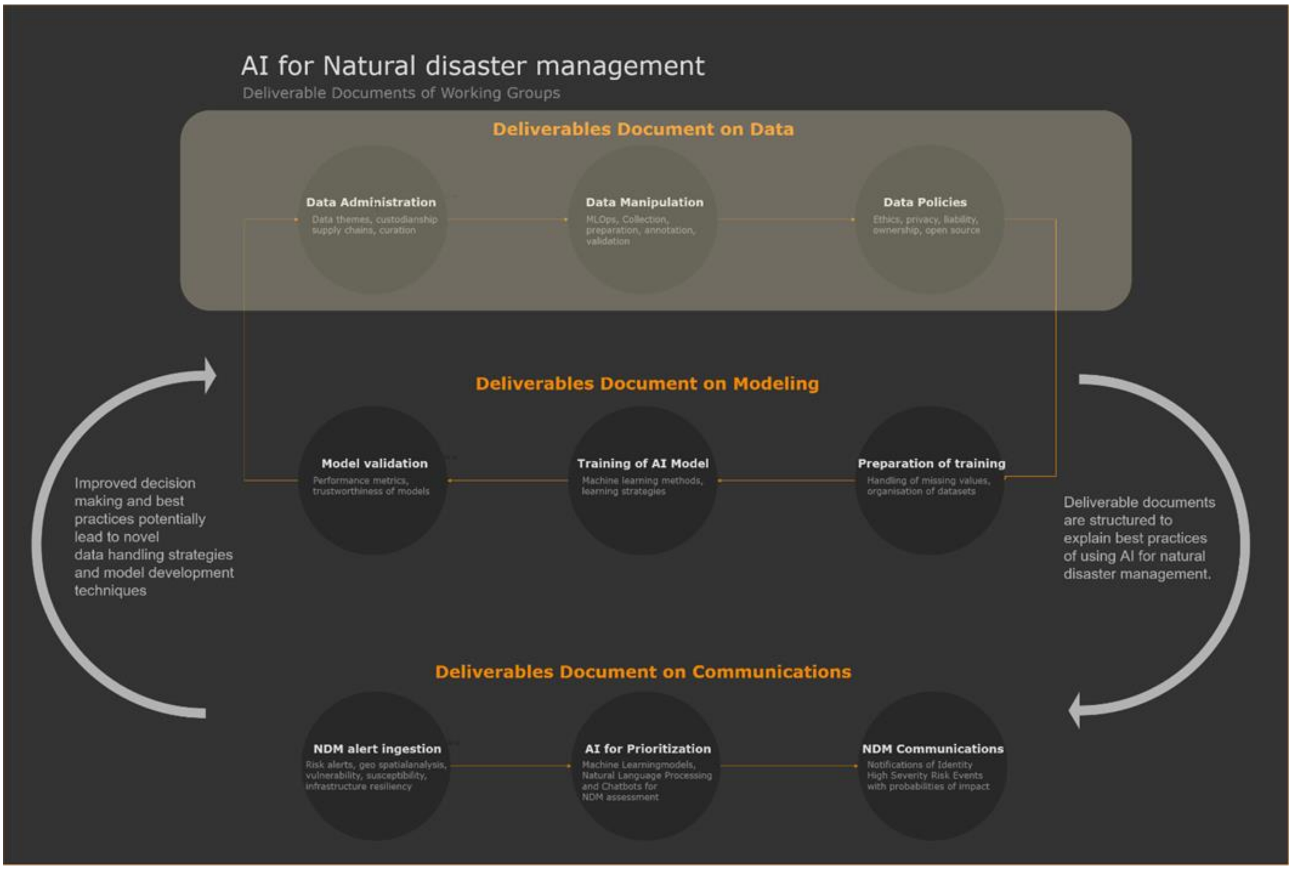
This report takes a practical approach by examining numerous disaster-related AI/ML use cases and drawing from the rich experiences of experts in the field. It provides invaluable insights that summarize the key findings and strategies for the disaster response community, aiding in formulating more effective and informed responses to catastrophic events using ML/AI techniques.

Furthermore, there might be an interest in delving into complementary activities; for instance, curating datasets that can be used to benchmark AI-based tools, contributing towards tools to support reliable annotations, etc.[1]

1 Scope

The Focus Group on AI for Natural Disaster Management (FG-AI4NDM) was established by ITU-T Study Group 2 on "Operational aspects of service provision and telecommunication management" in December 2020. This Focus Group is coordinated in partnership with the World Meteorological Organization (WMO) and UN Environment Programme (UNEP).

The Working Group on Data for AI (WG-Data) is one of three main sub-groups established under FG-AI4NDM.



AI for Natural disaster management
Deliverable Documents of Working Groups
Deliverables Document on Data
Data Administration
Data themes, custodianship supply chains, curation
Data Manipulation
MLOps, Collection, preparation, annotation, validation
Data Policies
Ethics, privacy, liability, ownership, open source
Deliverables Document on Modeling
Model validation
Performance metrics, trustworthiness of models
Training of AI Model
Machine learning methods, learning strategies
Preparation of training
Handling of missing values, organisation of datasets
Improved decision making and best practices potentially lead to novel data handling strategies and model development techniques
Deliverable documents are structured to explain best practices of using AI for natural disaster management.
Deliverables Document on Communications
NDM alert ingestion
Risk alerts, geo spatialanalysis, vulnerability, susceptibility, infrastructure resiliency
AI for Prioritization
Machine Learningmodels, Natural Language Processing and Chatbots for NDM assessment
NDM Communications
Notifications of Identity High Severity Risk Events with probabilities of impact

Figure 1 – Three main deliverables on AI for Natural Disaster Management

This Technical Report explores the best practices in collecting, monitoring and handling data (Figure 1 and Figure 2). Furthermore, the report endeavours to answer questions such as:

- What requirements should data meet when being used to train, test and validate an AI-based algorithm?
- What issues can arise during data collection (bias, insufficient coverage, etc.)?

[1] The scope of this document is limited to best practices in collecting, monitoring, and handling data for AI/ML applications in natural disaster management, and does not fully address AI/ML models, another essential component in the disaster management AI lifecycle. The FG-AI4NDM report on AI for modeling is intended as a companion to this report, and includes higher level details on model development. Users of this report are recommended to read both reports together to obtain a fuller picture of best practices for integrating data into AI/ML-based tools.

- How can AI-based algorithms be used to augment/enhance data quantity and quality?

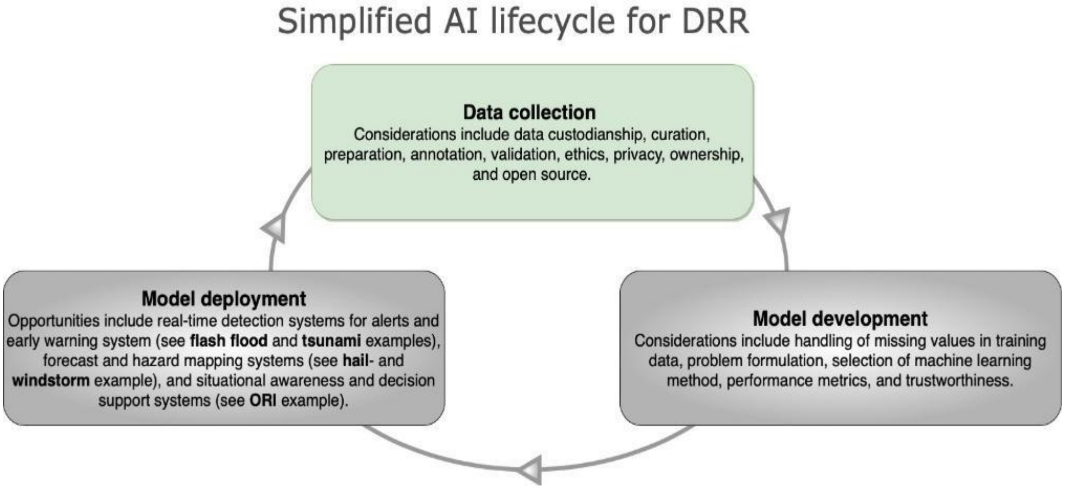
This report delivers sections on two main components of AI/ML and data:

- Data archiving
- Operational workflow: MLOps[2] and data in MLOps stages (see Figure 10).

In addition to the main topics, the report includes, but is not limited to, supporting sections addressing:

- AI/ML and data standards.
- Policy, ethics and legal issues.
- Open-data, open-source guidelines and policies.

Furthermore, the main components of the report are established from three points of view: data management (section 6), operational – ML environment (section 7) and use cases from the topic groups (section 12). The overall intention of this report is to provide preliminary information that may be included in a standard or other "good practices" guidance. ML modeling and communications are addressed at a high level, with greater detail available in dedicated companion reports from other reports prepared under WG-Modeling and WG-Communications of FG-AI4NDM.



Simplified AI lifecycle for DRR

Data collection
Considerations include data custodianship, curation, preparation, annotation, validation, ethics, privacy, ownership, and open source.

Model development
Considerations include handling of missing values in training data, problem formulation, selection of machine learning method, performance metrics, and trustworthiness.

Model deployment
Opportunities include real-time detection systems for alerts and early warning system (see **flash flood** and **tsunami** examples), forecast and hazard mapping systems (see **hail**- and **windstorm** example), and situational awareness and decision support systems (see **ORI** example).

Figure 2 – Simplified AI lifecycle for DRR

2 References

[ITU-T E.102]	Recommendation ITU-T E.102 (2019), *Terms and definitions for disaster relief systems, network resilience and recovery.*
[ITU-T M.3080]	Recommendation ITU-T M.3080 (2021), *Framework of artificial intelligence enhanced telecom operation and management (AITOM).*
[TU-T M.3363]	Recommendation ITU-T M.3363 (2020), *Requirements for data management in the telecommunication management network.*
[ITU-T X.1303]	Recommendation ITU-T X.1303 (2007), *Common alerting protocol (CAP 1.1)*
[ITU-T Y.2113]	Recommendation ITU-T Y.2113 (2009), *Ethernet QoS control for next generation networks.*
[ITU-T Y.3172]	Recommendation ITU-T Y.3172 (2019), *Architectural framework for machine learning in future networks including IMT-2020.*

[2] Combination of machine learning, DevOps and data engineering.

[ITU-T Y.4472]	Recommendation ITU-T Y.4472 (2020), *Open data application programming interfaces (APIs) for IoT data in smart cities and communities.*

3 Abbreviations and acronyms

This Technical Report uses the following abbreviations and acronyms:

5V	Volume, Velocity, Variety, Veracity, Value
AI	Artificial Intelligence
AI4NDM	AI for Natural Disaster Management
API	Application Programming Interface
CNN	Convolutional Neural Network
DBN	Deep Belief Network
DL	Deep Learning
DOI	Digital Object Identifier
DSC	Data Supply Chain
DT	Data Themes
EDA	Exploratory Data Analysis
EO	Earth Observation
FAIR	Findability, Accessibility, Interoperability and Reusability
GDPR	General Data Protection Regulation
GML	Geography Markup Language
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
i.i.d.	independent identically distributed
IoT	Internet of Things
ISO	International Organization for Standardization
ITU	International Telecommunication Union
KML	Keyhole Markup Language
KS	Kolmogorov-Smirnov
ML	Machine Learning
MLOps	Machine Learning Operations
NLP	Natural Language Processing
OGC	Open Geospatial Consortium
OSS	Open-Source Software
PCA	Principal Component Analysis
PII	Personally Identifiable Information
PTHA	Probabilistic Tsunami Hazard Assessment
PyPI	Python Package Index

RMSE	Root Mean Square Error
RS	Remote Sensing
SAR	Synthetic-Aperture Radar
SDO	Standards Development Organization
SHAP	Shapley Additive Explanations
SVM	Support Vector Machine
TEC	Total Electron Content
TG	Topic Group
UN-GGIM	United Nations – Committee of Experts on Global Geospatial Information Management
USGS	United States Geological Survey
WG	Working Group
XML	extensible Markup Language

4 Introduction

Models are tools that enable us to simulate complex processes. They can provide insight into underlying mechanisms and allow for the prediction of outcomes under different circumstances. In the field of natural disaster management, traditional approaches can be categorized as physical-, physics-, or mathematical-based models (see, for example, Figure 3). The former includes graphical (e.g., on maps) and physical representations (e.g., in laboratory experiments). For example, the reader is directed to [b-Mignot], who review laboratory experiments gathering data on urban flooding and [b-Rossetto], who describe a wave generator that is used to reproduce tsunamis. Physics-based models, on the other hand, leverage the laws of nature to capture the behaviour and evolution of natural hazards. For instance, [b-Shaw] applies assumptions on dynamic stress, frictional behaviour and slip rate to the RSQSim platform for simulating earthquakes. In [b-Looper], the physics-based distributed Vflo hydrological model is combined with radar-derived rainfall to forecast flash floods. Finally, mathematical-based models pair equations (e.g., Newton's laws, conservation of energy) with parameterizations (e.g., initial and boundary conditions) and input data (e.g., meteorological data). In [b-Bartelt], for example, a mathematical-based model integrates heat transfer, water transport, vapor diffusion and mechanical deformation equations; new snow, snow melt and wind drift as boundary conditions; and meteorological data to predict avalanche risk. Mathematical-based models are also a popular tool for predicting hazards in the meteorological sciences, with numerical weather predictions often used to forecast storms [b-NWS].

Figure 3 – Mathematical-based models have been used for predicting the risk of avalanche (avalanche scar in Wallis, Switzerland, March 2022)

This report presents the results of these methods, a database consisting of 79 contributions (45 standards and 34 activities), followed by a discussion of the main trends and gaps. An analysis of the results shows that existing standardization activities are mostly oriented towards "reactive" phases of the disaster management cycle and natural (as opposed to man-made) disasters, with geophysical and hydrological hazards as front runners. The most prominent technology mentioned was not AI/ML, but the Internet of things (IoT). Regarding data, topics such as storage and privacy were common; and in terms of modeling, scalability and validation were prevalent. In cases where AI-based methods were applied, supervised techniques seemed most common. Finally, it was found that operational implementation leaned toward forecasting rather than tools such as dashboards, decision support systems, or hazard/susceptibility maps.

In comparison with ongoing focus group activities, it was found that many of the gaps in existing standards are being addressed through the use cases and technical reports being developed within the remit of FG-AI4NDM. However, it is also apparent that – as technology and standardization progresses – this database of standards and related activities will need to be continuously updated through leveraging the newly founded partnerships within the United Nations (UN) and standards development organization (SDO) communities.

In the traditional approaches, earth observations (EOs) serve a supporting role – at times contributing to parameterizations, input data, or model validation. However, as the volume and the dimension of digital EO data has grown, so has the interest to learn from them and to give them a more prominent role in the models. Consequently, data-driven approaches – such as those based on artificial intelligence (AI) – have moved to the forefront of natural hazard and disaster research [b-Sun]. Through the focus group's workshops, analysis of use cases, reviews of literature and internal expertise, it has been observed that many AI-based models in natural disaster management seek to detect events in real time [b-Thüring], forecast events (e.g., the compound and scalable flood prediction system[3] presented by One Concern at the 23 June 2021 focus group workshop) and/or assist in the communication of risks, for instance, via early warning systems, decision support systems and hazard maps [b-Lin]. Although falling outside of the scope of the focus group, many studies have been found applying AI ex post to assess impacts and support recovery [b-Cha].

[3] https://www.itu.int/en/ITU-T/Workshops-and-Seminars/2021/0623/Documents/Feyera%20Hirpa.pdf?csf=1&e=9RIPvx

The performance of these data-driven models depends (among other factors) on the quality, quantity, compatibility and appropriateness of the underlying EO data. Furthermore, certain types of models require additional processing such as annotation. Therefore, a clear set of guidelines on how to acquire, manage and prepare EO data can be an invaluable tool for AI developers (including geoscientists and computer scientists in academia, the private sector, or government agencies such as national meteorological offices) and other stakeholders. From the demographics of the focus group meetings, which are open to the public, these other stakeholders include those regulating (e.g., policymakers at the national and regional government level) and implementing the final tool (e.g., emergency agencies at the local, national and regional level; first responders; and humanitarian organizations).

Therefore, this technical report aims to provide guidelines in a way that is accessible to a diverse readership. First, each section introduces key topics with a general overview. Then, as the section progresses, more detail and complexity are provided along with concrete proposals that align with the approaches shown to be effective in the use cases (see section 12). If the reader seeks additional information on a given topic, they are encouraged to consult the complementary deliverables (e.g., educational materials, glossary, roadmap and online search tool) including other focus group publications.

Some of the key topics to be discussed include: given the volume of EO data, how can they best be administered and managed (see section 6)? How should one clean, analyse and curate a dataset (see section 7)? How can non-graphical and graphical approaches be used to acquire insight into patterns, anomalies and other characteristics of EO data (see section 8)? How do the proposals align with other efforts to standardize data (see section 9)? What policy, ethical and legal issues should be considered when using EO data for AI (see section 10)? What are the benefits of open data and software (see section 11)? How can these approaches be seen in real use cases (see section 12)?

With the ML-based data-related information provided in these documents, the reader will be prepared to continue reading the second technical report in the series ("AI for Modeling"), which explores how ML-ready (co-located, indexed, linked, annotated etc.) EO data can be integrated into AI-based models (Figure 4).

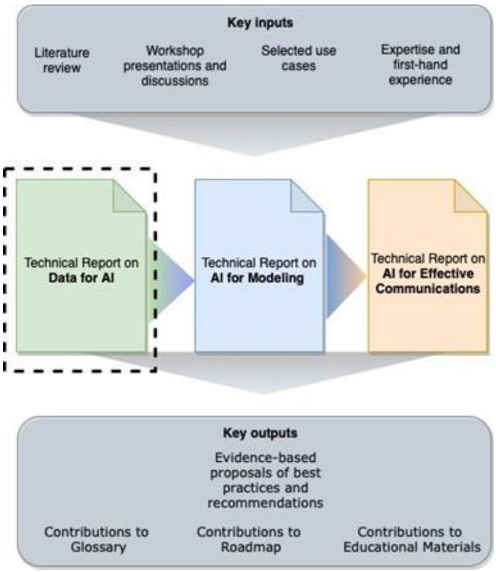

Key inputs
Literature review
Workshop presentations and discussions
Selected use cases
Expertise and first-hand experience
Technical Report on Data for AI
Technical Report on AI for Modeling
Technical Report on AI for Effective Communications
Key outputs
Evidence-based proposals of best practices and recommendations
Contributions to Glossary
Contributions to Roadmap
Contributions to Educational Materials

Figure 4 – The relationship between the three technical reports

5 Understanding technologies that create data for AI-enabled disaster response: Present and future

Currently, AI/ML applications are used in many different areas of natural disaster management (see section 12 for use cases). A roadmap to understanding the data requirements for AI/ML implementations for disaster response can be possible by understanding present and future technologies. [b-Lambert at al.] summarizes some of the technologies:

Present and near future technologies	Mid-term technologies	Distant future technologies
• Segmentation algorithms • Intelligent satellites to detect image changes • Drones mapping and sensing • Vulnerable area detection • Multi-lingual natural language processing	• Drone-enabled communication • Drones moving with visual clues • Intelligent satellites with synthetic-aperture radar (SAR) capability • Networks can combine different modalities • Trustworthy AI and crowdsourcing • Automated image analysis	• Fully automated drone swarms with connectivity • Drones-controlled brain-computer interface • Digital twin technology • Automated intelligent report generation

For example, in the present, it is common to see most drones transmit live video to be analysed by human operators for image classification. In the near future – mid-term case, if image classification is moved to the onboard process without human intervention, data flows might need to be updated because of communication bottlenecks and disc space limitations. For example, drones will be required to load not only a pre-trained AI/ML system but also data will need to be transmitted to the closer local servers (fog servers) instead of the cloud. In the case of distant future technologies, such as fully automated drone swarms with connectivity, each drone with AI capabilities will have to communicate with each other to obtain the complete picture. Noting that today's ML systems are central, this will require the implementation of distributed AI/ML systems where data flow may be restructured.

Some other important technologies that probably affect the data formats are (a) Internet of things (IoT) networks of physical objects embedded with sensors, (b) blockchain technology and (c) digital twins.

The Internet of things refers to a network of heterogeneous physical objects, such as sensors and smartphones, which gather and transmit information in real-time [b-Vermiglio]. Computational intelligence and network connection are incorporated into these devices to enable unified global sensing and communication in real-time [b-Kumar-2]. Social media data can also be combined to provide complementary information about areas affected by a disaster [b-Vermiglio]. Automated and remote IoT networks are especially useful for disaster management when predicting disasters and transmitting early warnings. They can reduce the need to send human beings into dangerous zones for disaster relief operations [b-Davis].

For example, IoT sensors can be crucial in fighting forest fires. Physical sensors measure carbon dioxide and humidity in the atmosphere to trigger alerts when there is a high risk of forest fires and firefighters can use the information gathered from those early detection sensors to have a clearer view of the magnitude and location of the fire [b-Davis].

IoT sensors also play a role in earthquake and tsunami early warning systems through SMART cables, an initiative of the ITU/WMO/UNESCO IOC Joint Task Force, Science Monitoring And Reliable Telecommunications (JTF SMART) Subsea Cables[4], in which environmental sensors (accelerometers and temperature/pressure sensors) are placed on existing telecommunications cables in the ocean [b-

[4] https://www.itu.int/en/ITU-T/climatechange/task-force-sc/Pages/default.aspx

Howe]. This serves to densify the network of existing observation techniques (in-situ sensors on buoys and ships and remote sensing via satellites) by several orders of magnitude and provide real-time, global measurements of various oceanic parameters, such as temperature, mass distribution, sea level rise, circulation, tides and seismic activity. Such systems provide access to the deep ocean and other problematic areas, enabling more timely and reliable forecasting of earthquakes and tsunamis.

SMART cables have already been implemented in a number of projects; one example is the CAM-2 system connecting the Portuguese mainland with the Azores and Madeira, which was shown through modeling to improve tsunami and earthquake early warning capabilities significantly.

While IoT shows many promises, some shortcomings must be considered before adopting it for disaster management. The IoT system architecture should enable the secure integration of heterogeneous devices with various communication protocols while filtering false alarms [b-Kumar-2]. Continuous data collection and transmission raise data privacy concerns, mainly as centralized data collection is vulnerable to security breaches [b-Al-Rayani]. IoT networks require internet connection and power, which can present a challenge in remote or undeveloped areas [b-Davis]. Submarine sensor networks such as those implemented by SMART cables also require consideration of translational relations and it is important that data collected through these networks adhere to FAIR data principles of Findability, Accessibility, Interoperability and Reusability, particularly when used in early warning applications [b-Howe].

[b-Yadavalli] proposes an autonomous modular framework for a disaster management system consisting of low-cost sensor nodes that observe physical conditions such as temperature or soil moisture, an observer node that tracks the status of the sensor nodes, hybrid connectivity, a message forwarding mechanism between the nodes that is scalable for systems of various sizes and any HTTP browser as the client. The simple message-based scheme for communication between the nodes enables scalability and integration with other systems. The low-cost sensors can also be deployed at remote or difficult to access places to be used for applications such as early warning systems for flooding and smart city disaster management.

[b-Kaur] discusses how IoT can be further enhanced in combination with blockchain technology (BIoT). Blockchain is a decentralized system that manages information by recording peer-to-peer transactions in a ledger. BIoT can improve performance and address some of the current shortcomings in IoT for disaster management, such as data integrity, security, communication and context awareness (Figure 5: Comparison of data flow in IoT without and with Blockchain [b-Al-Rayani]).

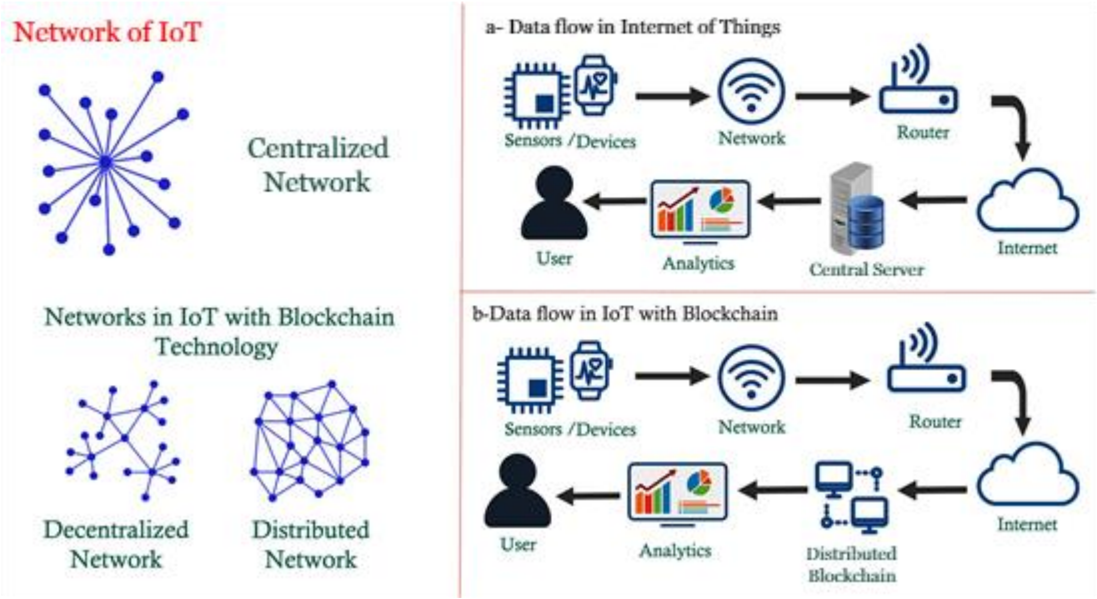


Figure 5 – Comparison of data flow in IoT without and with Blockchain **(adapted from [b-Al-Rayani])**

For example, BIoT can verify that the information comes from reliable sources in the pre-disaster phase. Once the disaster strikes, communication networks and platforms can be created on BIoT to track missing persons, locate and direct aid and assess damages for immediate response. Finally, in the post-disaster phase, BIoT can protect sensitive information such as identities and medical data while also being able to verify claims.

Topic group use case:

***Flash flooding monitoring system in Mexico**. IoT instruments provide water level, weather station and soil moisture data, which are synthesized and processed with other data such as satellite and drone imagery to detect flooding and improve early warning systems. Such IoT networks require secure and reliable infrastructure, such as wireless communication networks to transmit data for real-time applications and protection against equipment theft or damage.*

A *digital twin* is another emerging technology with promising implications for disaster management. A real city can be represented as a digital twin replica, created with sensor data from IoT devices, geospatial information provided by governments and participatory sensing in the form of crowdsourced data [b-Ham]. As shown in the workflow in Figure 6, scenarios are simulated and analysed in the digital twin city, and then used to inform real world decisions, damage estimations and ongoing monitoring [b-Yu], [b-Ford].

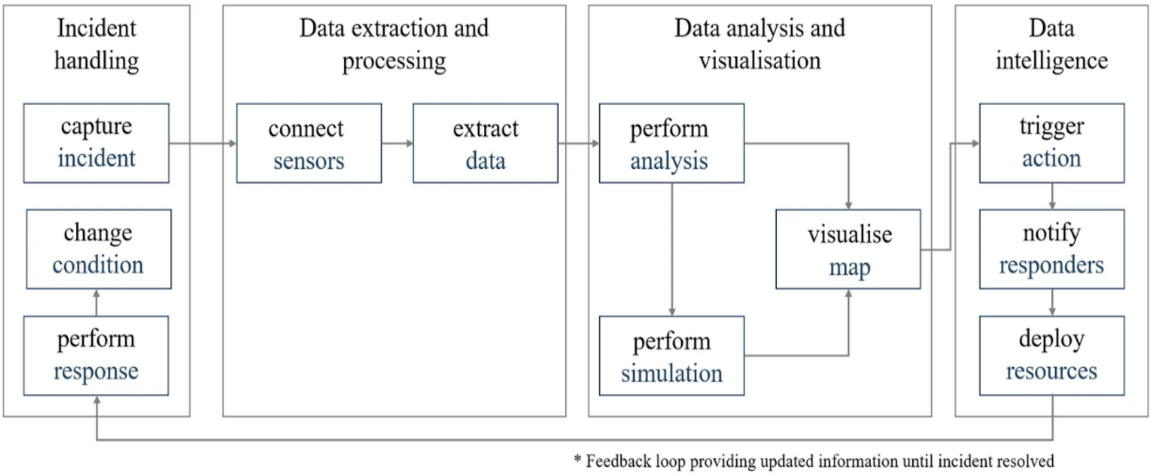


Figure 6 – Digital twin incident response [b-Wolf]

The digital twin city uses AI/ML to bring together a wide range of data sources with near real-time capabilities. For example, images and point clouds obtained from satellite and UAV remote sensing can be analysed with computer vision [b-Lary]; natural language processing and data mining can be used to access data from social media in order to detect, forecast and respond to disasters [b-Arthur]; and crowdsourced visual data can be used to update the digital city model and bring it closer to reality [b-Ham].

These data will be stored in different data structures, so while there will be access to an unprecedented and ever-increasing amount of data, it will be difficult to integrate and analyse them, especially for real-time applications [b-Hashem]. Knowledge graphs and heterogeneous information networks (HIN) can provide a way to integrate and analyse heterogeneous data, ranging from social media posts, aerial imagery, 3D laser scanners, news reports, weather stations, traffic sensors, etc. [b-Fan], [b-Yu]. Once processed, the data can be displayed as a map or used as insights on which to inform responses and resource allocation [b-Wolf].

The processed data informs the decision-making process, which involves multiple actors and stakeholders. Gamified learning environments can draw on realistic information flows and interactions to provide a dynamic simulation space in which actors work together in the digital twin city and train for natural disaster responses [b-Fan].

The digital twin city is a dynamic system that requires constant updates and analysis to improve efficiency and outcomes. AI/ML enables the examination of the missing links or inefficiencies in a disaster response system and can point to solutions specific to the network. Meta-network analysis can be used to investigate the interactions of various actors and entities and eventually find ways to ensure a smoother cooperation [b-Fan].

Topic group use case:

Utilizing AI & probabilistic modeling for strategic resilience. *Data from the natural and built environment are incorporated into a digital twin model, which is then used to predict the extent of damage in the built environment from natural disasters, including for telecommunications infrastructure.*

6 Elements of AI/ML: Data administration point of view

Through the survey, liaison statements and online form, 45 standards could be identified (Table 3). These came from SDOs including ISO, IEC, IEEE, ETSI and ASTAP, with two-thirds of the standards (30) coming from the ITU [primarily, ITU-T (17) and ITU-R (13)]. Across the United Nations system, 34 activities in this domain were also identified. Several sister UN agencies such as the WFP, UNDP, WMO and UNU provided information on projects in the context of disaster reconnaissance (after a disaster strikes). However, many of these activities are not technology-centric; focusing more on the relationship between emergency response and humanitarian aid, with human personnel responsible for interventions.

Liaison statements sent proved to be an effective channel for inviting inputs to the roadmap. Additionally, individual research carried out within WG-Roadmap also yielded positive results in terms of gathering inputs from online databases for different SDOs given in Tables 3 and 4.

Figure 7 describes the structure of section 6.

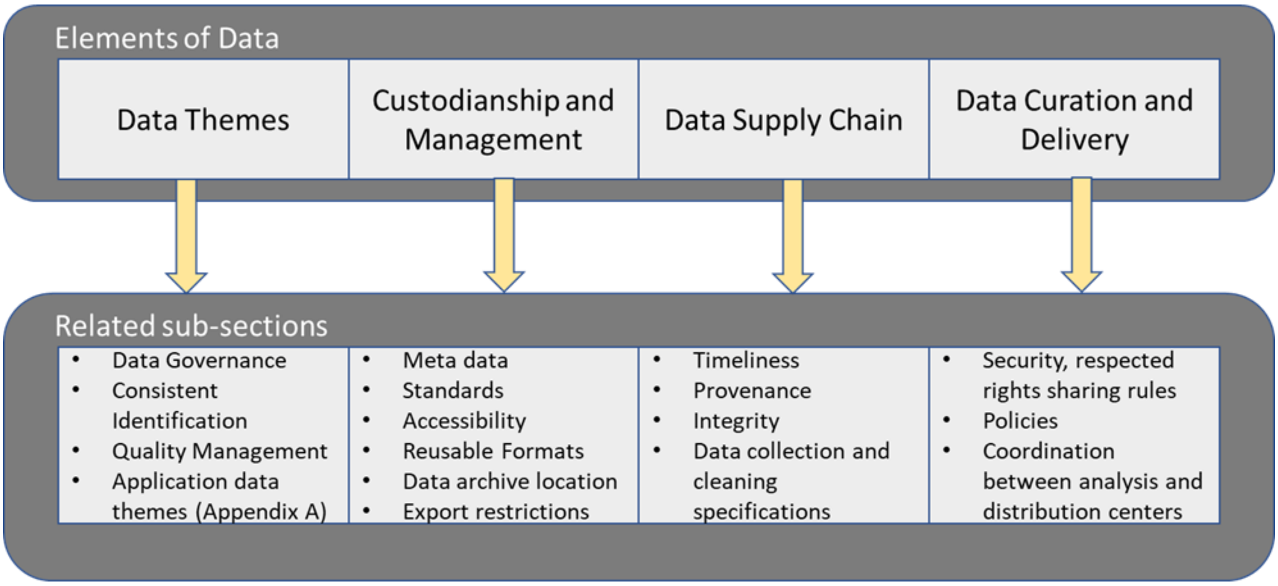

Elements of Data
Data Themes
Custodianship and Management
Data Supply Chain
Data Curation and Delivery
Related sub-sections
• Data Governance
• Consistent Identification
• Quality Management
• Application data themes (Appendix A)
• Meta data
• Standards
• Accessibility
• Reusable Formats
• Data archive location
• Export restrictions
• Timeliness
• Provenance
• Integrity
• Data collection and cleaning specifications
• Security, respected rights sharing rules
• Policies
• Coordination between analysis and distribution centers

Figure 7 – Pathway to section 6 [b-ACL-IJCNLP]

6.1 Data themes

This report utilizes the United Nations Global Fundamental Geospatial Data Themes described by the United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM). It builds its AI/ML-related data concepts around them. These fourteen data themes (Figure 8) are considered fundamental to strengthening geospatial information infrastructure and serve as a foundation to support global geospatial information management, as well as a companion resource to the UN-GGIM World Bank's Integrated Geospatial Information Framework (IGIF).

Each data theme is explained in detail in a publication of the United Nations Department of Economic and Social Affairs, Statistics Division, Global Geospatial Information Management Secretariat: "The Global Fundamental Geospatial Data Themes" [b-UN-GGIM], [b-UN-GGIM-2]. This document provides the argument for each theme's fundamental nature, as well as noting a theme's contributions to the UN Sustainable Development Goals (SDGs). Sources of each theme's data are noted, as are existing relevant data standards.

Definition: Data themes are a minimum set of concepts used to label datasets to enable interoperability between different groups, companies, organizations and nations.

Example of data themes:

Theme	Sub-theme	Dataset
Water	River	Water quality
	Sea	Wave height (sensor)
	Groundwater	Sea ice (imagery)

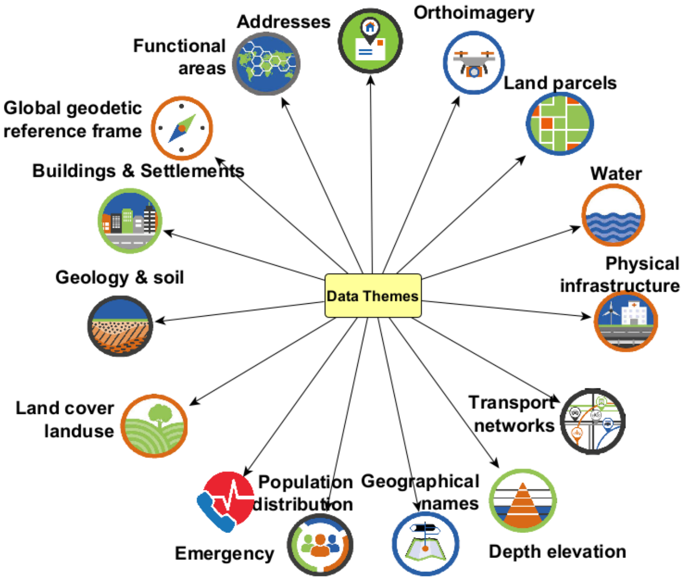


Figure 8 – Data themes

6.1.1 Data themes for AI/ML (use case)

Data themes are an ML-based information management system. A combination of data themes (DT) is used to describe the approach in each use case or study. For example, application DT can describe data that are captured for a specific purpose, such as health and utilities; and socio-economic DT can describe data that are used for demographic studies. Additional information and details about DT can be found in Figure 9.

Best practices:

Researchers are advised to use data themes to interpret their dataset. It is important that the data are managed throughout the product life cycle, not just at the beginning or in product development.

Topic group use case:

Situational awareness system for disaster response using space-based AI (SARA). *Satellite imagery and meteorological data are used to predict the areas in a region that are most vulnerable to natural disasters. The data themes included are described in Table 1.*

Table 1 – Data themes of use case situational awareness system for disaster response using space-based AI

Subject: Situational Awareness System for Disaster Response using Space-Based AI	
Data Theme inputs to MLOps (Feature Vectors)	Data Themes describing the outputs (Labels)
Global Geodetic Reference Frame	Buildings and settlements
Orthoimagery	Physical infrastructure
Land cover, land use	Transportation networks
Population distribution	

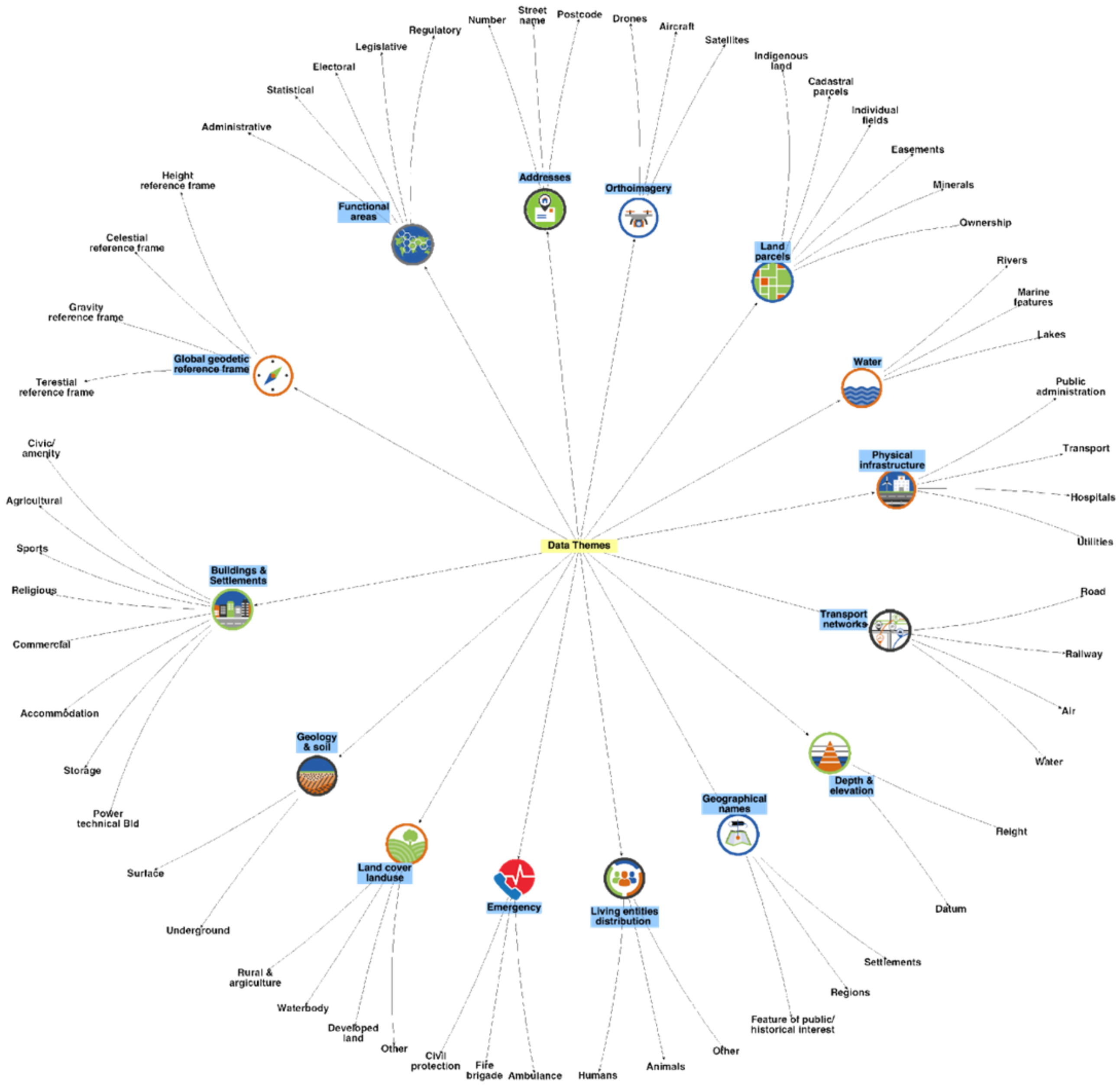


Figure 9 – Data theme examples

6.2 Custodianship acquisition and management

Data custodianship mandates responsibilities for the acquisition, management, maintenance and quality of information [b-UN-GGIM-4].

Data custodianship refers broadly to the cleaning, versioning, access control, transformation, automation, integrity and security of data to improve the speed and reliability of ML/DL applications. Model training relies on large amounts of high-quality data and ML workflows involve a great deal of time and effort on the part of model engineers and operators to manage versions and prepare datasets prior to feeding them into models. Data are transformed and modified according to the application, and it is an enormous but essential undertaking to ensure the quality of evolving data and their dependencies throughout their life cycle, from collection to deployment.

Some challenges in data management identified by [b-Munappy] include:

- Lack of metadata, leading to confusion and poor understanding of datasets.
- Data granularity, in which aggregation of data can result in information loss, making it difficult to combine with other datasets.

- Shortage of diverse samples, in which DL models lack counterexamples on which to be trained.
- Heterogeneous sharing techniques that make it difficult to track data.
- Ethical and legal issues inherent in data storage and the memorization of data by DL algorithms.
- Heterogeneous data formats that require time and effort to transform to a uniform format.
- Data with missing values that lack context from those who collected the data about how they should be handled.
- Categorical data with no predefined sets or standards that similarly lack context on how they should be handled.
- Chronological or other sequences that need to be property aligned with metadata.
- Data drifts and changes in data distribution that cause errors and trigger cascading effects throughout the data pipeline.

A number of data management systems have been created to address these challenges. MLdp from Apple is one such example, focusing on data lineage and provenance, data semantics and formats, integration with diverse frameworks, data exploration, experimentation, reproducibility of model training and compliance with privacy and security regulations [b-Agrawal]. The data is managed through versioning, access control, ensuring compliance and the lineage and dependencies can be tracked among different versions to maintain data provenance.

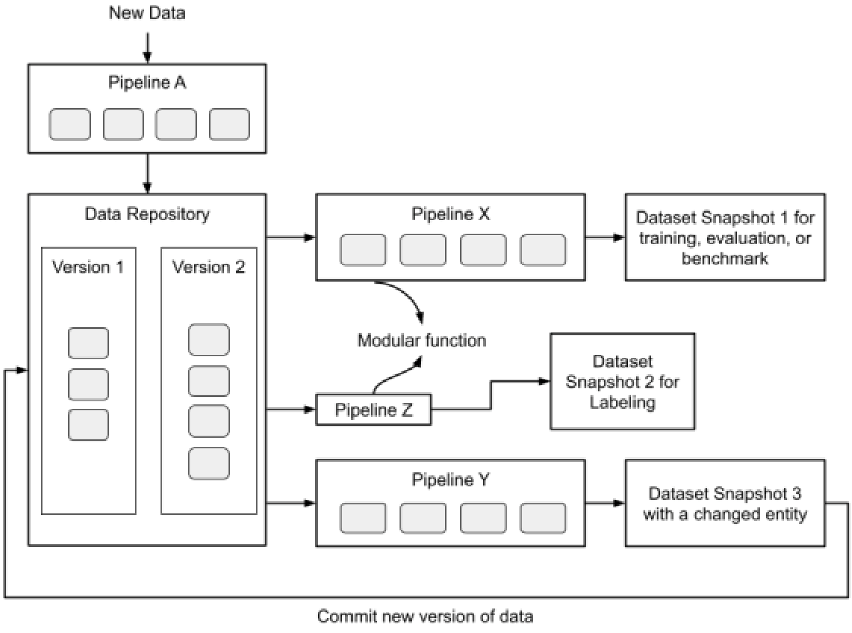

New Data
Pipeline A
Data Repository
Version 1
Version 2
Pipeline X
Dataset Snapshot 1 for training, evaluation, or benchmark
Modular function
Pipeline Z
Dataset Snapshot 2 for Labeling
Pipeline Y
Dataset Snapshot 3 with a changed entity
Commit new version of data

Figure 10 – Dataset manager and workflow manager architecture [b-Mao]

[b-Mao] presents another example architecture consisting of a dataset manager and workflow manager. The dataset manager stores the datasets, manages versions and provides access control that enforces permissions and enables querying by various users. The workflow manager is a user-defined pipeline for transforming data, the output of which can be used for model training and evaluation, or be fed back into the data repository.

The following chapters address in further detail the challenges and potential solutions for data custodianship in the context of AI/ML applications.

6.2.1 Data accessibility

The Data Accessibility Formula is as follows:

Data Accessibility = Visibility + Interoperability + Usability

Data visibility

Data visibility is the degree of ease through which an enterprise can monitor, search, display and analyse data from disparate sources. Data visibility and accessibility are often tied to one another. However, for data to be visible they need to be easily searchable by users. While in academia in the past, researchers would "request access" to large data from a research team (e.g., Linguistic Data Consortium) and wait weeks for approval and data access, many more datasets have become publicly downloadable. In situations where it is not possible to release the entire dataset, researchers can instead make a smaller version of the dataset available for use, and researchers who need the full version can request access.

Best practices:

- Publish the dataset details on platforms; examples include PapersWithCode ("Papers with Code"), Kaggle ("Kaggle") and Huggingface ("Huggingface").
- Describe how data were generated.
- Publish who is maintaining the data.
- Create relevant tags alongside the code on the publishing medium to make it easy for users to find, for example, by creating relevant tags on version-controlled platforms (e.g., GitHub).
- Provide an example notebook or query to illustrate using the dataset.
- Provide metadata information alongside the datasets regarding usage, topic, update frequency and data gaps on multiple platforms outside of the particular one used to host one's data.
- For dynamic datasets available through an application programming interface (API), have a pre-packaged static version available for reference computations and benchmarks.
- Include a specific standard license for a permitted use, such as the MIT license.

Topic group use case:

Artificial Intelligence Modeling Tools for Monitoring Desert Locust (AI-Locust): Breeding Grounds, Hatching Time, Population and Spatio-temporal Distribution. *One goal of this use case is to create an integrated desert local database under the Creative Commons copyright licence that is available for the general public to verify, replicate and reuse. Data is managed through icipe's Research Data Management and Archiving (RDMA) policy to ensure that the data remain Findable, Accessible, Interoperable and Reusable (FAIR). To this end, the data will be accompanied by proper metadata and documentation. Effective data management is employed to disseminate the use case outputs and make them accessible to the public under an open-source by-attribution licence (CC-BY 4.0).*

Data interoperability

Data interoperability is a principle affecting the storage, indexing and defining of data that enables organizations to reduce the barriers to data movement between different analysis environments and user contexts. Data interoperability is crucial to achieving integrated data supply chains. It is brought about by systems and services that create, exchange and consume data with clear, shared expectations for the contents, contexts and meaning of the data. In addition to promoting standardization for data sharing and reuse, interoperable data support multidisciplinary knowledge integration, discovery, innovation and productivity improvements. Interoperability requires commonly agreed formats, language and vocabularies; metadata will also need to use agreed standards and vocabularies and contain links to related information using identifiers.

Best practices:

- Standardized data formats, such as CSV, JSON and NetCDF, should be used to facilitate data exchange across different systems.
- Common data standards, such as the Common Alerting Protocol (CAP) for emergency alerts, should be used to ensure that data remain consistent across different datasets, systems and platforms.
- Metadata should be used to provide context and meaning for a dataset so that it can be searched, shared and reused appropriately.

Topic group use case:

Limitations of predicting snow avalanche hazards in large data sparse regions. *All data are stored in recognized data formats that support interoperability, depending on the data type and application. For example, avalanche observation and hazard assessment data are stored in standardized formats defined by the Canadian Avalanche Association, weather model data are stored in NetCDF files and snowpack model data are stored in a custom developed data object for R software. The group publishes their R software packages on The Comprehensive R Archive Network, with up-to-date codes and documentation.*

Usability

Data usability addresses the ability of an organization to easily access and use the data without much prior domain knowledge or the requirement of uncommon or inaccessible hardware or software.

Data in their raw or loosely processed nature can be arduous and time consuming to pre-process for any machine learning pipeline. Converting them to more easily accessible and usable formats (e.g., from simple comma separated values to formats providing hierarchically nested structures such as TFRecords, HDF5 files and JavaScript Object Notations) can accelerate the usage and minimize the time to get started. These data can be characterized as "machine readable," with high-performance data loaders available in machine learning packages including PyTorch and TensorFlow. In cases where data must be stored in other formats, the webpage / GitHub with instructions on downloading data should contain the command that someone will need to download the data (i.e., client URL commands or download hyperlinks) as well as the Python command to read the data in [(e.g., astropy.io.fits.open ('filename. Fits')] or a Jupyter notebook demonstrating the use of data. The data should provide sufficient context for a user to not only download the data and access it via their ML programming language of choice (typically Python), but also transform the data as needed. This context includes information on recency, metadata and variable descriptions.

Topic group use case:

Satellite images and machine learning for mapping flood. *Images are pre-processed and make use of a convolutional neural network, with implementation in Python and Tensorflow. The languages, tools, applications, libraries, operating system and other technologies are open source, making it easy to access and reference the data for future work and research related to flooding.*

6.2.2 Metadata: Data sheets or better meta sheets for enhanced interoperability and security

During the implementation of ML algorithms, data go through a series of processing levels to reach the ML-ready stage. Despite the importance of data to machine learning, there is currently no standardized process for documenting such a staged process [b-Gebru]. Considering AI/ML's increasing usage in the domains such as natural disasters, critical infrastructure and finance, the lack of information on each stage of the data processing can lead to severe consequences from the loss of human life to economic losses [b-Gebru].

To resolve this issue [b-Gebru] suggested a data sheets concept to facilitate better communication between dataset creators and dataset consumers and encourage the machine learning community to prioritize transparency and accountability. However, it is worth noting that metadata files do not have the full capacity to hold the ML-related information and data sheets are just a collection of questions.

Best practices:

To capture the actions taken on the data during the ML implementation, it is proposed to use a combination of metadata and datasheets; that is meta sheets [b-Microsoft] together.

Metadata + Data Sheets = Meta Sheets

Some of the possible information to include:

- Provenance
- Lineage
- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
- Is there a label or target associated with each instance?
- Is any information missing from individual instances?
- Collection process
- Pre-processing/cleaning/labelling
- Uses
- Distribution

In practice: IBM "FactSheets"

For example, IBM advocates the use of "FactSheets" to communicate the key facts of a model or service for its entire AI lifecycle [b-IBM-2]. This ensures the most effective and appropriate use of the model or service and contributes to the overall goal of improving AI governance and transparency. In a study by [b-Piorkowski], the FactSheets methodology was used to create the documentation for an AI model in the healthcare domain, resulting in overall improvements in AI transparency and a better understanding of the model by its users.

6.3 Data supply chains (DSCs)

The data supply chains refer to the technological procedures and human-involved activities that enable the flow of data through an organization, from its origin to the point of consumption or analysis [b-Bartley]. As noted in the UN-GGIM-World Bank Integrated Geospatial Information Framework (IGIF) – Strategic Pathway on Data [b-UN-GGIM-4], data supply chains and their interlinkages need to be formalized and streamlined to improve the quality of information for end-users.

- Provenance (the origin and quality of data); integrity.

- Data centres: infrastructure, where the data will be analysed, etc.
- Discuss where data originated from, in commercial and/or open datasets, whatever applicable to the group.
- Consider supplying the digital object identifier (DOI), which is a persistent link, and specifically, an identifier allowing data to be traced from production to publication.
- Data collection and cleaning specification [b-Pineau]; reproducibility checklist [b-ACL-IJCNLP].
- Data processing and analysis, including what and how much will be done in (non-AI-based) automated processes or algorithms, what software package(s) are used, which organizations serve as "nodes" in the flow of data, how they add value (or degrade) data before transferring the data/information on to the next node/organization.

The IGIF Strategic Pathway on Data further notes that data supply chains – especially long or multifaceted ones, are often not well integrated and that this poses a risk for data-derived information to become inconsistent, out-of-date, and unusable due to lack of reliability. Resynchronization of datasets is currently a highly manual task and would benefit from formalized and streamlined processes throughout the data supply chain, as data from an identified and authoritative source could save considerable time and effort, minimize total supply chain length and improve the overall quality and timeliness of information for users in disaster situations.

Topic group use case:

An intelligent big data analysis system for wildfire management. *This use case aims to build an intelligent big data analysis system for fire management. Their data are centralized in an independent LAN/San environment. The data centres realize the data curation, including processing, storage, transmission, exchange and management of both metadata and master data through advanced analysis technology and open-source frameworks (such as R, Apache spark, knime, rapidminer).*

6.4 Data curation and delivery

Data curation and delivery refers to the art of maintaining the value of data and delivering it to end users in a way it can be visualized and used. The main purpose of data curation is to ensure that data are retrievable for future purposes or reuse. Important considerations include:

- Security, respected rights, data sharing rules (confidentiality, privacy, intellectual property rights and the protection of sensitive information are preserved) and the sharing by applying these guiding principles; through this, data custodians will be able to manage and share reusable geospatial data, and in doing so, meet their obligations to government and the user community [b-Bartley].
- Policies for "data life" scenarios for deleted or changed data from the original sources – for example, crowdsourced data.
- Analysis centre coordination, data centre coordination and secure file transfer.

Best practices:

- A data security plan should be developed to outline procedures to data access, storage, transfer and processing. This plan should be regularly reviewed and updated to ensure that it remains effective.
- When transferring data, it is recommended to encrypt the data and use secure file transfer protocols such as SFTP or HTTPS. Sharing protocols should be established between data centres.

- Version control should be implemented to track changes to data over time, enabling data users to understand and reproduce results.

Topic group use case:

Utilizing AI & probabilistic modeling for strategic resilience. *The data workflow of this use case, which draws on data that are subject to proprietary or privacy restrictions, incorporates curated quality assurance (QA) at various steps, such as preparation, cleaning, training and calculation. The data are centrally managed by a data operations team, with a data Wiki available for further details. Data privacy is maintained by storing sensitive data on protected cloud services and obfuscating private information such as building locations.*

7 AI/ML data manipulation: Actions taken on data

To capture the dynamic flow of the data and the life cycle of the topic groups, we follow simple steps of machine learning operations (MLOps [b-Merritt]), which is a set of best practices to run AI successfully (Figure 10).

Elements of an MLOps software stack include but are not limited to:

- Data sources and the datasets created from them.
- A repository of AI models tagged with their histories and attributes.
- An automated ML pipeline that manages datasets, models and experiments through their lifecycles.
- Software containers, typically based on Kubernetes, to simplify running these jobs.

In the context of MLOps, topic groups might need to provide the following information about their research/products to their users:

- Are datasets carefully labelled and tracked to enable data scientists to cut and paste as needed?
- Did the topic groups establish flexible sandboxes and solid repositories for data as well as their models (see "AI for Modeling" companion report)?
- Are there ways to work with the ML engineers who run the datasets and models through prototypes, testing and production? This process requires automation and attention to detail so that models can be easily interpreted and reproduced.

Section 7 summarizes the main components of MLOps from data collection to data preparation (Figure 11).

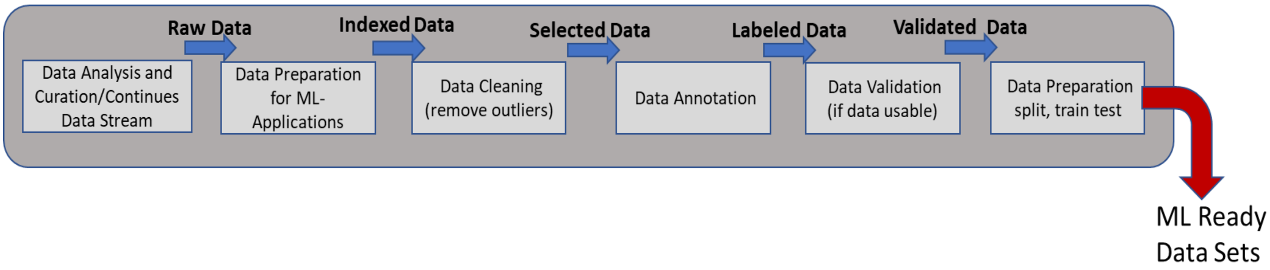

Raw Data
Indexed Data
Selected Data
Labeled Data
Validated Data
Data Analysis and Curation/Continues Data Stream
Data Preparation for ML-Applications
Data Cleaning (remove outliers)
Data Annotation
Data Validation (if data usable)
Data Preparation split, train test
ML Ready Data Sets

Figure 11 – Pathway to section 7

Topic group use case:

***AI and vector-borne diseases**. A Github repository for disease vector data will include scripts for users to download and work with the input data and the trained model in Python. Along with adherence to common protocols for data surveillance and reporting, centralized data repositories such as this one will enable researchers to tailor the AI models to their area of interest.*

7.1 Data collection (raw data stream)

Essentially, collecting data means putting your design for collecting information into operation. It includes declaration of data sources and the datasets created from them.

Data are collected in various ways according to the desired application. They can be collected from multiple sources, automatically transformed and integrated into a unified repository, or ingested without any transformation [b-altexsoft]. They can also be acquired at a single epoch, at regular intervals, or continuously as a data stream.

On the other hand, data discovery methods such as collaborative analysis, web-based systems, data lakes make it easier to search and explore existing datasets to determine if they are suitable for the application at hand [b-Roh]. Such methods should consider scalability, homogeneous data formats and data provenance/metadata.

Following acquisition, a data storage system should be established to maintain security, integrity and traceability [b-altexsoft].

Topic group use case:

***AIDERS: Real-time Artificial Intelligence for DEcision Support via RPAS Data AnalyticS.** AIDERS is a project that exploits the large volume of heterogeneous data available through cameras, LIDAR and sensors aboard remotely piloted aircraft systems to develop algorithms and a mapping platform to improve real-time emergency response systems. Raw data for training and testing are transferred directly to control stations, and then to servers accessible to researchers.*

7.1.1 5Vs of Big Data for AI/ML

Big data is a term that refers to datasets that have been created on a scale that is difficult to obtain, manage and manipulate in a timely manner using traditional computing tools [b-Patel], but are instead handled with a "scalable architecture for efficient storage, manipulation and analysis" [b-Chang], such as that provided by ML and DL. Advances in ML have given rise to the computational processing capacity needed to work with real Big Data [b-Zhou], while its development to DL models has increased its complexity and convolution of the network's dimensions [b-Teri].

The datasets of big data are defined as having at least five[5] [b-Khan] common features (5Vs) [b-Hadi]:

- Volume: Extremely large amounts of data. How much data?
- Velocity: Extremely high data rate. How frequent or close to real time are the data?
- Variety: Extremely wide variety of data. How many different kinds and what kinds of data?
- Veracity: The quality/integrity/credibility/accuracy of the data source, type and processing. How accurate and applicable are they?

[5] Based on the literature there are five more 'V': Viscosity, Variability, Volatility, Viability, Validity.

- Value: The worth of the data being collected for decision-making or improving operations processes. How useful are the data in decision-making? What data do you need to create, and what are their benefits to the deployment[6]?

The characteristics of 5Vs are depicted in Figure 12.

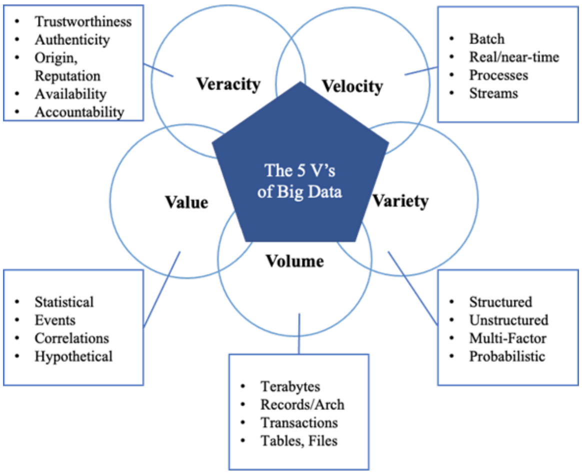


Figure 12 – Characteristics of 5Vs [b-Hadi]

It is crucial for a researcher to know which of the aforementioned 'Vs' are included in their datasets. This will facilitate their work to extract information systematically; deal with cleaning and noise issues; choose advanced, scalable and interoperable tools; follow the appropriate methods for the analysis, curation and annotation of the data; and, finally, give fast responses. The main goal is for each researcher to manage in the most efficient way data volumes that are too large or complicated for standard data processing tools to process.

A case in point of Big Data is the high dimensionality of geospatial data and the complex spatiotemporal relationships involved in a disaster. The constant increase of the spatial (e.g., sub-meter), temporal (hourly) and spectral (hundreds of bands) resolutions creates Big Data as well. Big spatiotemporal data are also generated by geospatial models through numerical simulations of complicated earth systems. IoT devices produce Big Data through the data streams in real time from networked mobile devices, personal computers, sensors, radio-frequency identification markings and cameras throughout the globe [b-Michael].

Best practices:

- Choose the most appropriate format (e.g., the structured data format).
- Leverage the power of distributed computing through, for instance, parallel processing platforms (e.g., Hadoop).
- Recall that Big Data is not synonymous with data quality (accuracy, completeness, redundancy and consistency).
- Pay attention to the storage systems (e.g., need for storage systems that can scale up fast).
- Consider the difficulties in managing the heterogeneity of Big Data.
- Consider new security challenges (e.g., smart devices and privacy concerns).
- Select the data that will be utilized and create the simplest, most uncomplicated database possible.

[6] The 'deployment' corresponds to section 7.5 of the companion report from WG-Modeling.

- Before conducting data analysis and mining (e.g., cluster analysis, classification, machine learning), recall that Big Data require pre-processing to improve quality [b-LaValle], [b-Mayer-Schönberger].
- Deep learning [deep belief network (DBN), convolutional neural network (CNN)] can be used to address Big Data problems such as data tagging and indexing, information retrieval, etc. [b-Jan].
- Construct metadata to describe Big Data and associated procedures [b-Gantz].
- Cloud computing provides virtually unlimited and on-demand processing capacity (e.g., Earth System Grid, Pangeo, Google Earth Engine, CloudGIS, etc.).

Topic group use cases:

- ***Flash flooding monitoring system in Mexico.*** *Data are sent in JSON format to ensure scalability and interoperability.*
- ***Building a coupled earthquake-tsunami*** (Leverage the power of distributed computing)**.** *TEC Simulator in a Parallel HPC Environment. Tsunamis are simulated in a parallel High-Performance-Computing environment using the Tsunami Squares method to create faster forecasts with higher accuracy.*
- ***Geographical data science applied to landslide and debris flow hazard in the Colombian Andes*** (pre-processing to improve quality)**.** *Raster images containing Digital Elevation Models and images of landslide-prone areas are oversampled or undersampled to ensure that the dataset is balanced before training the model, and to reduce the computational cost of the algorithms.*
- ***Satellite images and machine learning for mapping flood.*** *High-resolution Sentinel-1 and -2 satellite images produce a large amount of data, which are pre-processed (noise removal, cleaning), analysed (computation of spectral indices) and interpreted (classifying images to map flood areas) by a Convolutional Neural Network (CNN) architecture.*
- ***Multi-hazard use case for operations risk insights and day one relief for natural disaster response.*** *Operations Risk Insights (ORI) collects and analyses data from WMO-based national alert service feeds and thousands of trusted news source feeds to identify and assess natural disaster risks. The data are ingested, analysed and retained in central hybrid cloud-based IBM databases (IBM's Cognitive Enterprise Data Platform and other cloud centers).*

7.2 Data preparation and integration (indexed data)

The idea of big data indexing is to fragment the datasets according to criteria that will be used frequently in query (e.g., date and time). The fragments are indexed with each containing value satisfying some query predicates [b-Fasolin]. An index is a persistent data structure, which stores data in a suitably abstracted and compressed form to facilitate rapid processing by an application such as an ML system [b-Gani]. In general, indexes are a list of tags, names, subjects, etc., of a group of items that reference where the items occur. An indexing strategy is the design of an access method to a searched item. It also describes how data are organized in a storage system [B-Manning]. Some examples for types of indexing can be summarized as [b-Daniel]: (1) AI-based approaches (e.g., latent semantic indexing, hidden Markov model) and (2) non-AI approaches (e.g., tree-based indexing strategies, custom indexing strategies, inverted indexing strategies).

Topic group use case:

Flash flooding monitoring system in Mexico. *This use case draws on various IoT sensor data streams that measure water level, soil moisture, standard depth, perimeter, hydraulic radius, area, velocity and flow. These data are initially sent to a database server in JSON format, encrypted in MQTT, then indexed by sensor name and date in Excel, CSV and PDF formats and made available to all users as open source and open data on a web application.*

7.3 Data pre-processing and exploration (data cleaning, analysis and curation)

For many purposes, especially in multi-sensor intercomparison and data validation, it is essential to look into the datasets before drawing statistical conclusions. Some of the example cases are summarized below.

- The choice of cost function for a neural network (NN) may require normally distributed residuals in the data domain.
- To handle growing variances in the data, sometimes log transformation may be required.
- The behaviour of outliers might require special attention including discarding of the outlier data.

Topic group use case:

Artificial intelligence modeling tools for monitoring desert locust (AI-Locust): Breeding grounds, hatching time, population and spatio-temporal distribution. *Datasets consisting of long-term satellite-based monthly rainfall, temperature, wind speed, vegetation variables and edaphic factors are pre-processed and harmonized to the same resolution to prepare for further data analysis.*

7.4 Temporal data processing requirements

The use of temporal data for machine learning applications in environmental science typically requires one or more pre-processing steps before ingesting it into the machine learning-based model. Typical requirements include operations related to missing values manipulation, outlier detection and handling, data transformation steps like normalization and standardization, and data type modification. Specifically, when dealing with missing values in a dataset, a practitioner has two alternatives: a) data reduction - by removing the missing data, or b) data augmentation - by creating data points to replace the missing values. The latter often involves extrapolation or another data completion operation. The selection of one alternative over the other depends on several factors, like the number of data points available (i.e., record length) and the feasibility of using an interpolation technique to fill in the missing values (e.g., a missing noon surface temperature should not be interpolated with records from 6:00 and 18:00, as it is likely higher than both). Regarding the outlier detection, like the missing values problem, practitioners have several alternatives, but in this case the users can also decide to keep them unchanged. For instance, after detecting an extreme gust event, the event might be flagged as an outlier by a probabilistic algorithm or if it crosses a user-defined threshold (like the cut-off speed for wind turbine design), here the practitioner can a) remove the outlier, b) change the outlier using interpolation techniques, or c) leave it unchanged. As always, the choice of alternative is application dependent as machine learning models could be used to forecast average conditions or extreme conditions, thus needing different flagging characteristics. Data transformation steps like normalization, on the other hand, relate to mathematical transformations that change the timeseries characteristics, so it looks more Gaussian, often transforming the original timeseries into a new timeseries with zero mean and one standard deviation. These operations are often performed to aid the machine learning models using sigmoidal or tangential activation functions during training, and then follow a back-transformation step when the trained model produces its forecasts, so the forecasts have similar mean and standard deviation to the original data. Other temporal data processing requirements include transformation between numerical types. For example, some machine learning models might expect to use integer data as inputs, other models might require float values. Other common transformations of temporal data include date and time formatting (e.g., Unix time, 24 hr day format), calendar formatting (e.g., Gregorian, Julian, etc.) and scientific programming for specific package transformations, like converting the data to lists, pandas, or arrays in Python, for example.

Best practices:

- The decision to augment missing values in a timeseries or to remove outliers should follow careful consideration of the nature of the dataset or model being used.
- Timeseries should be normalized for training of certain machine learning models, then backtransformed for forecasting to resemble the original data.
- Date, time and calendar formats should be transformed to the most appropriate numerical type or format for the machine learning model.

Topic group use case:

Probing seismogenesis for fault slip and earthquake hazards. *It is generally not possible to record multiple seismic events at a single location, so deep learning models are trained to learn seismic cycles and applied to regional network seismic data. The pre-processing of the timeseries data of these models includes normalizing for unit variance, formatting to equal time resolution, and ignoring the gaps and missing data from the recorded data sets.*

7.5 Data annotation

Supervised learning is one of the most popular paradigms for using machine learning in different applications of AI, ranging from natural language processing to computer vision applications. This type of machine learning model relies upon high-quality and primarily manual data annotation efforts by human users. The annotations or labels with the data objects are used by a learning algorithm to train a computational model, which can infer or predict an annotation class for a given data object automatically [b-Tseng]. Data annotation is also referred to as data labelling and it is a process to assign one or more class labels from a predefined set of labels to a given data object. For instance, assigning a label to an image based on the contained object, or assigning a label to a textual news document for the topic of news such as technology and sports. The annotation task is often conducted with the help of domain experts or crowd-workers on crowdsourcing platforms such as Amazon Mechanical Turk [b-Gancheva]. It is a costly process given the manual effort involved and very important given a high-quality annotated dataset is required to develop supervised learning applications with good performance [b-Lapuschkin].

The data annotation process includes the following steps and decision areas:

1. Definition of the annotation task: identify the type of data object, class labels, required amount of labelled data, etc.
2. Decision on the annotators based on the complexity of task: determine the level of expertise required, that is, domain experts versus crowdsourcing workers.
3. Preparation of annotation guideline: define classes with diverse examples to avoid biases in interpretations.
4. Decision on annotation exercises and platform: consider at least three or more annotators per data object, run the exercise on the unpaid or paid platform such as Amazon Mechanical Turk.

5. Finalizing annotations or labels for the data objects: identify mechanisms to aggregate labels by multiple annotators for a data object, such as through majority voting and inter-rater agreement (many annotation tools and platforms for crowdsourcing provide similar measures).

Best practices:

- Attempt to minimize the subjectivity in the class label definitions, by including a diverse group of experts that help reduce cultural biases.
- Provide different examples for the annotation task, given the annotators might have a different background of culture, social context, etc.
- Critically analyse assumptions on annotations. For instance, labelled data might not always be applicable as ground-truth as the meaning of class labels might change in the future.
- Decide the level of expertise required and accordingly, consider the background of annotators suitable for the task; consider test annotation exercises.
- Create test examples and provide reasoning with the answers before launching an annotation exercise at scale.
- Document the changes in the annotation guideline with time, if any, and include them in the metadata of the resulting annotated dataset.
- One can also refer to this practical guide [b-Hellström] to conduct data annotation projects for AI-infused applications.

Topic group use cases:

- ***Proposal of an AI chatbot use case as a multihazard communication technologies.*** *SNS messages are annotated in a deep learning model, which uses natural language processing to extract and analyse disaster information. First, annotation takes place in a training dataset using simulated SNS messages, and the annotations are evaluated and modified until the desired result is achieved.*
- ***Real-time volcano-independent seismic recognition as volcano monitoring tool.*** *Volcano-seismic (VS) events are manually detected and classified according to their source, then made available as open-access waveform databases from universities, Volcano Monitoring Observatories and the International Federation of Digital Seismograph Networks. These labelled datasets vary in reliability and availability, thus underscoring the need for a standardized quality assurance process.*

7.6 Data validation

Data validation (Figure 13) is a system that assists data providers to check and validate their data against a set of validation rules to ensure data are of the highest quality possible [b-UN-GGIM-5] and is usually carried out prior to processing and importing. The most common sources of significant errors and vulnerabilities are data that have not been validated or have been insufficiently validated [b-Wang]. Therefore, validation has an extremely high impact on data quality and, moreover, its absence will incur expenses for cleaning, converting and storing.

Depending on the application, consideration must be given to whether data were recorded correctly and reflect realistic values, how much error can be tolerated, whether all relevant data are recorded with no missing entries or missing values (completeness checking), deduplication, whether data agree with its format and structure, whether data are unreasonable, and whether data comply with standards.

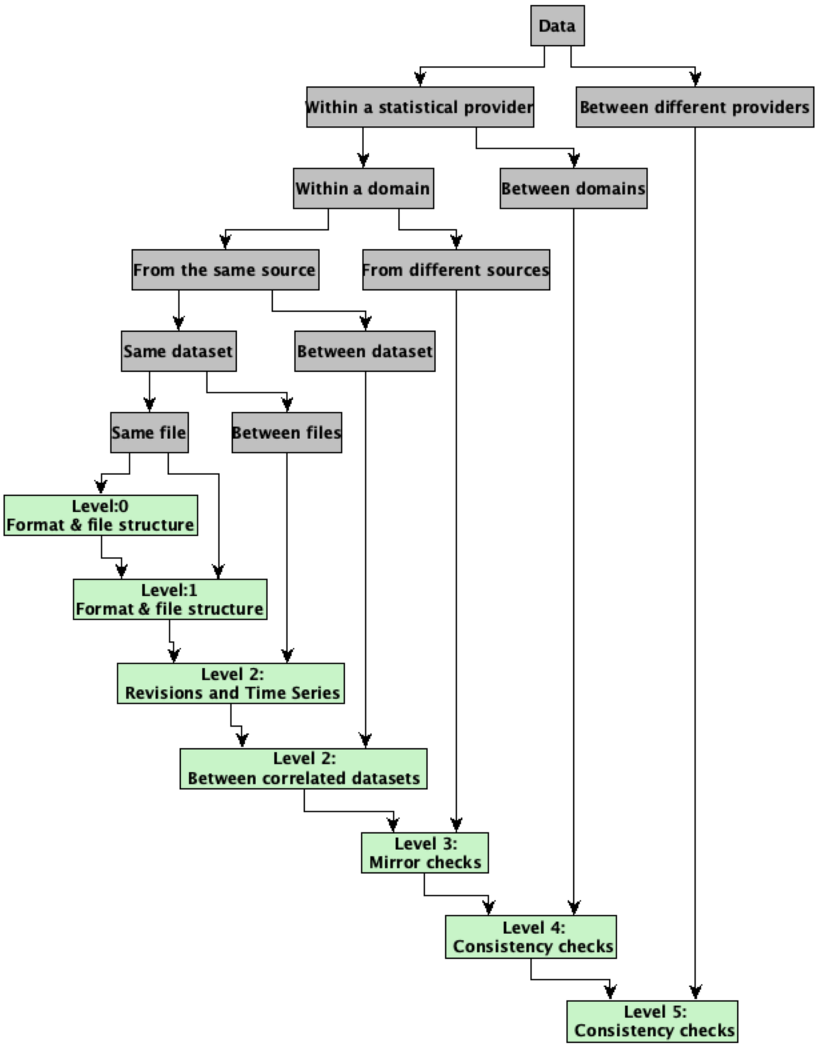


Figure 13 – Visual representation of validation [b-EC-1]

Best practices:

- Check the accuracy of the instruments you use in the field.
- Database cleaning – Perform exploratory data analysis (EDA) (summary statistics and graphical representations) – Make sure that no additional duplicated records are produced.
- Perform a range check to ensure that the input data are inside a predetermined range (e.g., latitude: −90° to 90° and longitude: −180° to 180°).
- Use trusted data sources that follow well-defined data quality assurance standards and procedures.
- Compare your data (information) with similar trusted databases and check the corresponding literature.
- Check to ensure that the original data types are compatible for the processing to follow. Various collecting methods create several types of data.
- Make the appropriate atmospheric corrections for satellite data.
- Use cloud/shadow masks for satellite data.
- Keep one or a low range of data structures and multimedia formats.
- Data validation is not the same as data verification.

Topic group use cases:

- ***Flash flooding monitoring system in Mexico*** *(instrument accuracy)****.*** *The sensors were calibrated before sending information to the system to ensure data validation.*
- ***AI for multi-hazard communications technologies*** (data cleaning)***.*** *Removal of null or missing data and duplications.*
- ***Landslides of masses of soil and rock: Intelligent risk management in areas highly threatened by climate change*** (data cleaning)***.*** *The validation is done through simple rules of continuity in space and response to the factorials (of the EDA).*
- ***Landslides of masses of soil and rock: intelligent risk management in areas highly threatened by climate change*** (data range check)***.*** *The information that comes from analysing field and laboratory tests as well as information from satellite images is reviewed on the grid when it is installed over the area that is being studied. In this stage, the rows, or columns (in the worst case), that do not correspond to natural outliers are discarded.*
- ***Utilizing AI & probabilistic modeling for strategic resilience*** (standards)***.*** *Data came from trusted sources and were reviewed by subject matter experts.*
- ***AI for landslide monitoring and detection*** (standards)***.*** *Map data are validated by the institutions that create them and those that publish them.*
- ***AI and vector-borne diseases.*** *The data were compared to the known presence of the vector as reported in the literature and the ECDC VectorNet database.*
- ***Using ML to reconstruct flooded area under clouds in optical satellite images: The Mozambique use case*** (atmospheric corrections)***.*** *Auxiliary data such as digital elevation models are used to validate cloud-covered low-resolution optical satellite images during the training and inference of the ML algorithm.*

7.7 Data constraints on model/data validation

Data scientists and ML researchers face a constant challenge: validating their data, models and algorithms on real-world data and using these data to optimize and improve their solutions. An appropriate dataset is needed to validate the model to have a successful artificial intelligence/machine learning application. This section will discuss how to validate models for deep learning and traditional machine learning models. The primary consideration will be given to categories of data and the challenges that inherently come with this type of data.

AI systems and, more generally, statistics rely on two assumptions. The first assumption is that the data a system experiences during inference displays the same distribution as the data that were used to train the AI. This concept is named *identical distribution*. The second assumption is that each sample is independent of other samples. This concept is called *independence*. Together independence and identical distribution (i.i.d.) lay the foundation for many basic model validation techniques. Unfortunately, in real-world data, these assumptions are rarely given.

The danger of improper validation of AI systems leads to poor performance in production. Specifically, data that are not independent will overperform during training. This is due to the fact that the correlation of data points can leak information the AI system would not have during inference in production. An example here is any timeseries where an intermediate timestep can be predicted using simple interpolation of surrounding timesteps. In a forecasting scenario, these surrounding data would not be available.

The subsections 7.7.1 to 7.7.4 discuss these different data types and how to properly address the validation.

7.7.1 Validation of independent data types

Every type of AI/ML validation relies on a split of the available training data into at least two subsets. The number of splits is mainly decided by virtue of the modeling approach. The training data split are the only data a model is provided during the optimization of the model. AI systems are particularly good at remembering the training data; hence an independent set of data samples is needed to assess if a model learned generalizable truths about the data and is applicable to new unseen data. This concept is called overfitting, where a model learns to conform to the random fluctuations and noise of individual data samples rather than general relationships between the input and output data.

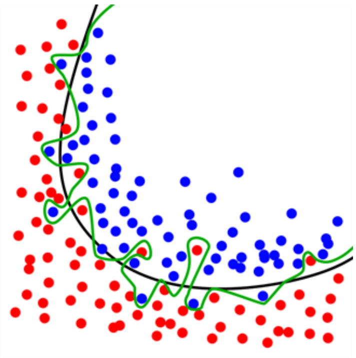


Figure 14 – Overfitted versus regularized model[7] [b-Chabacano]

However, many models and modeling approaches need to optimize parameters and hyperparameters of the AI system. This optimization loop requires its own validation data. Technically, the AI will optimize for the validation data, and it is possible for models to implicitly overfit to that dataset. This is where AI practitioners require a third split: a final test subset, where the optimized model can be tested for generalization. In fact, it is possible to overfit a model manually by tinkering with the parameters of the model. For this reason, the final test set is only used once for the final model validation. In the simplest case where the data are i.i.d., one can randomly sample for the dataset to split out subsets of data. Depending on the amount of data available, it is common to split the data 50%-25%-25% to obtain a sufficiently large test set that covers the distribution of data samples. Section 6.6 of the "AI for Modeling" companion report addresses data validation in further detail.

Best practices:

- The training dataset should be large and diverse enough to capture the underlying relationships and patterns without memorizing the noise.
- It is recommended to split the data into at least three sets: a training set to train the model, a validation set to evaluate the performance of the model during training, and a test set to evaluate the final performance of the model.
- Regularization techniques, such as L1, L2, and dropout regularization, can be used in ML models to prevent overfitting and improve generalization to unseen data. The model architecture design should be as simple as possible to avoid unnecessary complexity and manipulation of parameters that can lead to overfitting.

[7] The green line represents an overfitted model and the black line represents a regularized model. While the green line best follows the training data, it is too dependent on those data and it is likely to have a higher error rate on new unseen data, compared to the black line.

Topic group use cases:

- ***Enabling natural hazards risk information sharing using derived products of export-restricted real-time GNSS data for detection of ionospheric total electron disturbances.*** *This use case takes a supervised learning approach that incorporates SME-labeled training data. To reduce overfitting, the dataset is split into 4 events: 3 for training and testing, and 1 for sample validation.*
- ***Exploring deep learning capabilities for surge predictions in coastal areas.*** *The last layer of the deep learning model employed in this use case includes an L2 regularizer, which prevents overfitting by keeping the model weights small, and dropout, which randomly drops out some of the neurons during training to prevent the neurons from memorizing the training data.*

Cross-validation for data sparse environments

In some cases, the dataset is too small for this type of split. Especially for classic machine learning models like linear and logistic regressions, random forests and support vector machines (SVMs), it is feasible to apply a cross-validation (CV) scheme. This means separating the data into equal parts, so-called folds, and training a model on all folds except for one. The *k*-fold CV approach trains *k* individual models to estimate the best parameters, where this validation fold is rotated in a round robin approach (Figure 15). This validation scheme is computationally more expensive as the models are trained but use the entire training dataset for training. Particularly in hyperparameter optimization, it is common to train a k + 1st model on the best parameters with the entire training dataset to finalize the model.

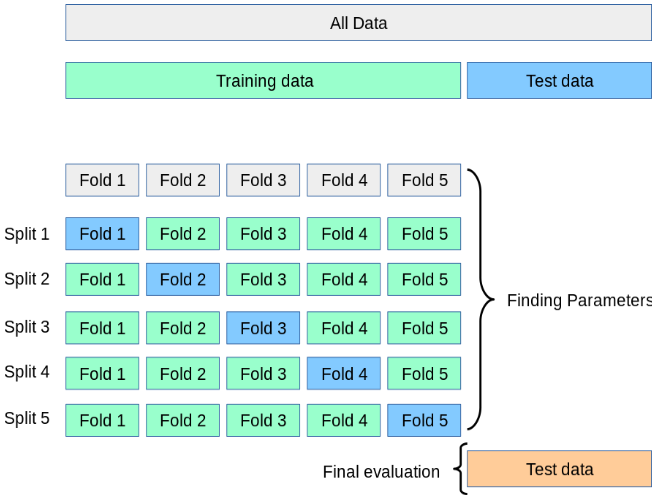


Figure 15 – ross-validation in a 5-fold split with separate test data [b-Scikit]

In practice:

Assessment of pluvial flooding in Venice. *For example, cross-validation reduces overfitting in machine learning models designed to assess pluvial floods in Venice [b-Zanetti]. The spatio-temporal correlation of flood event observations can undermine the assumption of independence between testing and training data. To counter this effect, random cross-validation and leave location and time out cross-validation are used to split data into a training set and a validation set that consider both time and space.*

Stratification for imbalanced data

Random sampling does not only assume that the data are i.i.d.; it usually implies that the target variable is distributed approximately equally over the entire dataset. This is rarely true in the real world and requires special handling of the train-test split. Usually, real-world data have some type of majority class and one or several minority classes. A simple example would be the Earth's surface, where water is the majority class with roughly 70% and forests would be a minority class among other classes. Randomly sampling from these data could mean that certain classes might be missed in the test data, or at least overrepresented. This problem can be addressed by assuming that a dataset is i.i.d. within each individual target class. This is called stratification and performs the random split, for instance, 50%-25%-25%, within each class on which an AI practitioner stratifies. Furthermore, this stratification can be applied to cross-validation as well (Figure 16).

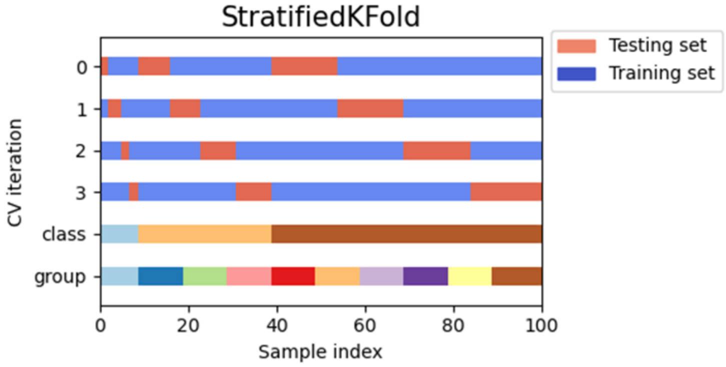


Figure 16 – Stratification of a 4-fold cross-validation scheme [b-Scikit]

In practice:

Modeling Onshore Tsunami Hazards. *For example, [b-Davies] explores various stratification methods to facilitate the use of offshore probabilistic tsunami hazard assessments (PTHAs) as an input to model onshore tsunami hazards at areas of interest. Simulating onshore tsunami hazards from offshore data is a very computationally expensive process that draws upon imbalanced and synthetic datasets. To address these challenges, a stratified sampling scheme is introduced in which tsunami scenarios are binned according to magnitude, with a certain number of random samples being used from each bin. The approach is further improved by sampling a user-defined subset of more important scenarios for the given location and using a non-uniform number of samples from each bin.*

7.7.2 Validation of correlated data samples

Validation of timeseries data

In temporal data, each sample is correlated in one dimension to the time steps before and after each individual sample. This causes problems with random sampling of validation datasets. The one-dimensional correlation of data results in a machine learning model learning from data in a specific time period, albeit subsampled due to the random split. The prediction on the "unseen" data in that same time period is then a mere interpolation of the training data rather than a proper validation on unseen time periods. Naturally, this suggests that contiguous time periods should be used for the train-test split. Timeseries data in AI/ML systems are commonly used for forecasting at inference time. This leads to the special circumstance that a model should be validated on data in the future relative to the training data.

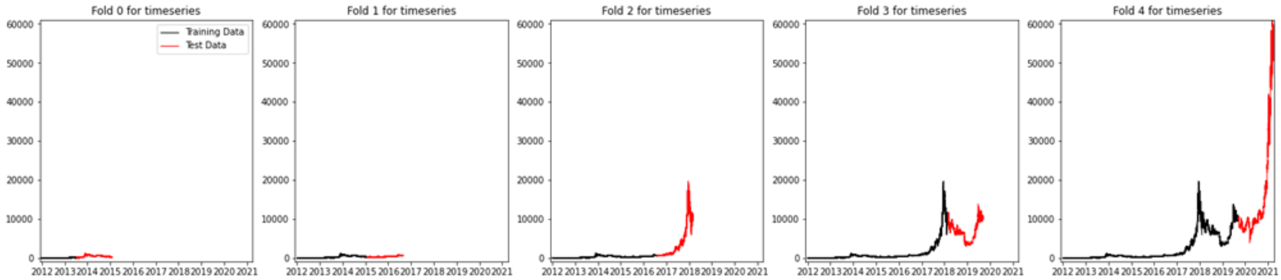


Figure 17 – Forecasting timeseries data with 5-fold cross-validation

Validation of geospatial data

Spatial and especially geospatial data violate the i.i.d. assumption since they are spatially correlated. Tobler's 1st law states that "everything is related to everything else, but near things are more related than distant things" [b-Tobler]. If the validation and test set were to be randomly sampled, this results in the test set containing data samples that are very close to data samples in the training set. Due to the correlation with proximal points in spatial data, this test set is now a poor measure for generalization as the AI system has seen a correlated sample during training time.

The blocked approach to a data split (Figure 18) ensures that the violation of the independence assumption can be treated more as a relaxation of the independence assumption.

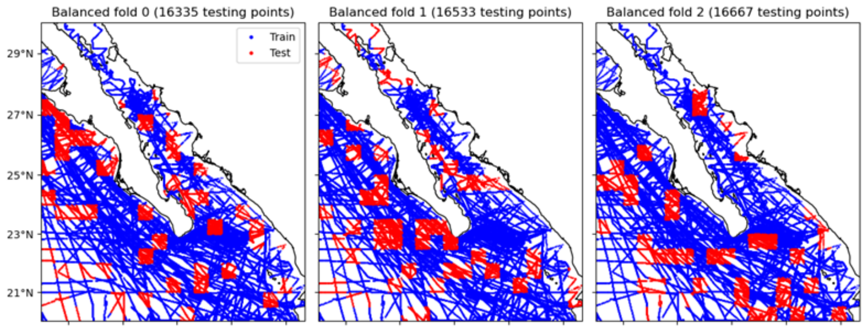

Balanced fold 0 (16335 testing points)
Balanced fold 1 (16533 testing points)
Balanced fold 2 (16667 testing points)
Train
Test
29°N
27°N
25°N
23°N
21°N

Figure 18 – Blocked cross-validation of geospatial data [b-Verde]

7.7.3 Validation of online AI systems

In some cases, real-time or near-real-time processing of data requires the training of machine learning systems within a larger dynamic system. This process of continuous training and training integrated in a dynamic system that generates data is called online training. The simplest way of training and validating these models is to use the online system to generate and save a static dataset. This dataset can be validated in the ways described beforehand. However, this often does not capture the full distribution of the data that will be encountered in the online setting and often underperforms. This subject is covered in further detail in section 6.5 of the "AI for Modeling" companion report.

7.7.4 Additional data consideration for validation

In addition to considerations that need to be made based on the type of data, there are general considerations that can impact the validity of a model. These are often based on how the data were acquired. Moreover, especially systems in production can experience drift, which will deteriorate the model performance over time, due to the ground truth shifting from the training dataset. In the case of systematic data drift, training data and, therefore, the train-validation-test split of the data used for model development are static. The data were acquired, pre-processed, and are not changing anymore. Unfortunately, many natural systems are non-stationary. This means that the distribution of the data fluctuates (see section 7.3 for statistical methods for preliminary analysis of the data). This can be periodic, but there can also be a trend in the data. An example of this is the global mean temperature that is currently experiencing an increase as time progresses due to climate change. This means that future data that will be collected will be different, and the model performance. The validation of data is a key consideration for building reliable AI/ML models. Thus, topics such as data drift, leakage and snooping are covered in greater detail in section 6.6 of the "AI for Modeling" companion report.

Best practices:

- Real-world data are rarely independent and identically distributed (i.i.d.), therefore it is often advised to generate validation groups over the dependent variable, e.g., time, space.
- Forecasting scenarios require special treatment to avoid artificially elevated results due to snooping from future samples.
- Cross-validation is useful to train and test models on the entire dataset to avoid blind spots in the evaluation of the model.
- Avoid data snooping/leakage by gaining a deep understanding of the dataset and its acquisition possibly through communication with domain scientists.
- Imbalanced data can be resampled for balancing during training and should be stratified for validation to ensure an identical distribution in the training and test sets.
- Explainable and interpretable AI systems can explain how machine learning and AI systems come to their decision. These can be used to inspect the model, perform ablation studies, and possibly implement causality. Moreover, it is important to understand how the data were acquired and which assumptions and decisions were made during acquisition of the training dataset.
- Implementing MLOps systems to automatically measure raw deterioration of model predictions on production systems is an important step. For data drift, testing covariate shift to online production data can indicate data shift. The Kolmogorov–Smirnov (KS) test for continuous data and Chi-squared test for categorical data can indicate data drift and be implemented automatically [b-Mohandas].

Topic group use cases:

- ***Utilizing AI & probabilistic modeling for strategic resilience.*** *This use case involves the creation of a digital twin to predict natural damage disasters. The fragmented and incomplete nature of the data makes it difficult to validate the model results directly. Thus, a hybrid physics-based/ML approach that incorporates subject-matter expertise is used to achieve a better estimation of resilience.*
- ***Multimodal databases and artificial intelligence for airborne wildfire detection and monitoring.*** *This use case employs deep neural networks with transfer learning and*

interpretable fuzzy modeling to overcome data limitations and enable assessment of model performance for wildfire detection.

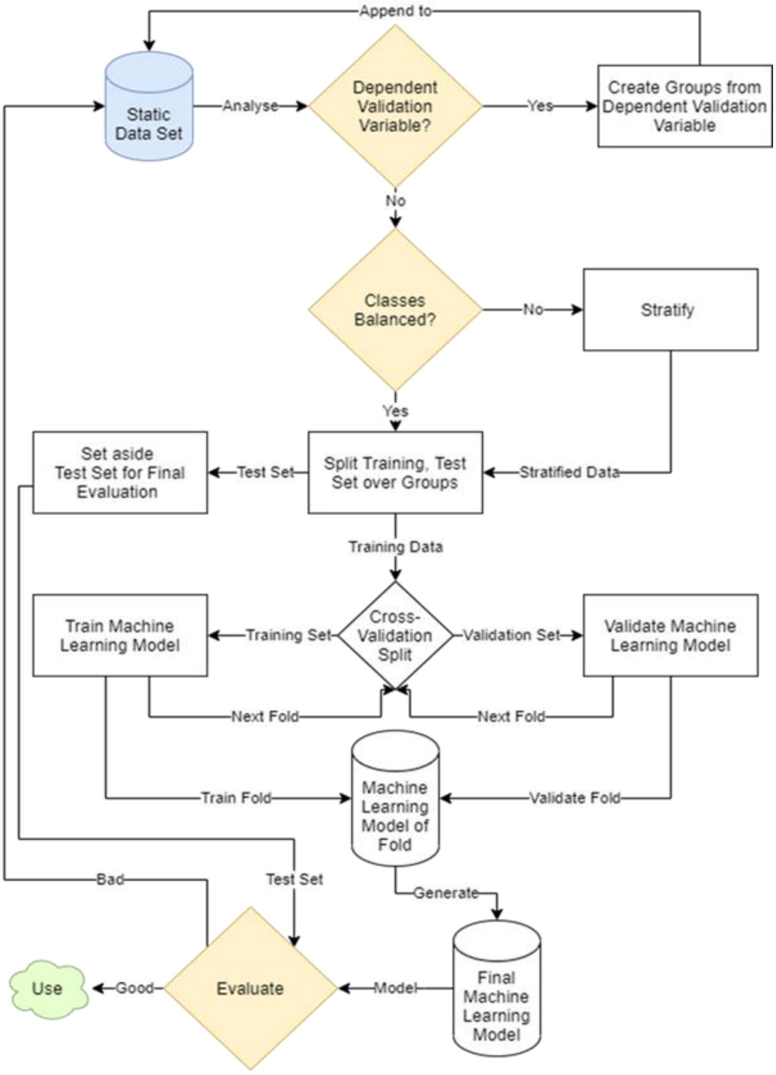

Append to
Static Data Set
Analyse
Dependent Validation Variable?
Yes
Create Groups from Dependent Validation Variable
No
Classes Balanced?
No
Stratify
Yes
Set aside Test Set for Final Evaluation
Test Set
Split Training, Test Set over Groups
Stratified Data
Training Data
Train Machine Learning Model
Training Set
Cross-Validation Split
Validation Set
Validate Machine Learning Model
Next Fold
Next Fold
Train Fold
Machine Learning Model of Fold
Validate Fold
Bad
Test Set
Generate
Use
Good
Evaluate
Model
Final Machine Learning Model

Figure 19 – Flow chart to treat data for machine learning and ensure valid results from real-world data

7.8 Understanding the relation between target, study area, model and sampling strategy

The use of the word bias when referring to AI tools can differ depending on the context in that it is being used, which can include statistical or sociological terminologies [b-Hellström]. Whilst terminology may differ, the essence of its use aims to denote a difference in the way that reality is represented. This can take place as an exclusion, preference, distortion, or manipulation of data (e.g., populations, object types and frequencies). An increasing number of reports have highlighted biased outcomes from AI-based products that have resulted from biased datasets [b-EC-2]. Without correction, biased datasets can result in biased algorithms and products, leading to potentially harmful outcomes with ethical implications for users [b-Goodrum], [b-Wang]. It is important to understand the datasets well before feeding them into any machine learning pipeline. Tools such as "What-If" or "Know Your Data" (from Google PAIR teams) are efforts to minimize such pitfalls and lead researchers towards better-informed machine learning research. It is important that the distribution of bias in data is communicated when a user downloads the data. For data to be usable, users must know their limits and for what they cannot be used.

Synthetic datasets

Sometimes it is valuable to augment a real dataset with synthetic data. When doing so, careful analysis needs to be done to ensure that the distribution of the synthetic data allows for the model to transfer learning to real-world data. Developing synthetic data can come from artificially generated examples created by generative/adversarial models or mathematical models describing what phenomena should look like along with added noise.

Topic group use case:

Probing seismogenesis for fault slip and earthquake hazards. *Most earthquakes occur on decadal or centennial cycles, making it difficult to gather enough in-situ geophysical measurements as input for data-driven ML models to monitor and detect seismic events. Instead, simulated laboratory data and numerical models are applied through transfer learning so that AI models can fill in the necessary data to learn seismic cycles. The simulated data has standardized formatting and includes metadata to facilitate integration with real-world data.*

Definitions of bias in different contexts can be:

- Deviation of the expected value of a statistical estimate from the quantity it estimates.
- Systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others [b-Merriam-Webster].
- The action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence personal judgment [b-Cambridge].
- Statistical bias is the difference between the expected value of an estimator and its estimand [b-Kozyrkov].
- Biased data occur when the dataset is incomplete, too heavily weighted toward specific attributes and/or unrepresentative of all use cases.
- Biased data can lead to biased models, as a model is only as good as the data used in training [b-Capilnean].
- The inclination or prejudice of a decision made by an AI system which is for or against one person or group, especially in a way considered to be unfair [b-Ntoutsi].
- Stereotyping, prejudice, or favouritism towards some things, people, or groups over others. These biases can affect collection and interpretation of data, the design of a system and how users interact with a system.
- Systematic error introduced by a sampling or reporting procedure [b-Google].

Data bias in AI: Data bias in AI systems is an error that occurs when some elements/factors are more represented and/or considered than others [b-Hellström]. The employment of biased data in AI systems results in faulty AI systems. The issues are mainly related to generalization problems with the consequence of creating unfair predictions and promoting some factors over others [b-Smuha]. Data can be considered as a building foundation and the AI system as a building. If the foundation is faulty, the construction is compromised as well. However, in case of data scarcity, it could be inevitable to exploit biased data. Therefore, in this scenario, they must be identified, and bias mitigation strategies should be applied.

How to identify bias on data: The importance of identifying data bias arises from the awareness that it is possible to mitigate depending on the type of bias being dealt with. Thus, pre-processing, in-processing and post-processing approaches may mitigate bias successfully. However, having knowledge of the data generation process is the first key aspect that can help in the identification of a bias, as well as performing exploratory data analysis (EDA) [b-Tukey].

Type of data bias

- Sample bias (selection bias [b-Campolo]; population bias [b-Olteanu]): This happens in the case of a non-random sample of a population. It causes some members of the population to be less likely to be represented than others [b-Heckman].
- Response bias: This happens when most data come from few sources [b-Baeza-Yates], [b-Paulhus].
- Exclusion bias: This occurs when a variable is excluded because it is considered to be irrelevant. In any case, data are omitted.
- Measurement bias: This happens when data are collected that differ from the real-world case with which they are associated, or in case of faulty measurements [b-Rothman]. Here, investigating natural hazards, there are various problems related to sensors. Some examples are resolution on the imagery, image distortions and different sensors (RGB, infrareds, SAR); furthermore, it can occur when validating the ground truth on low-resolution or faulty data.
- Recall bias: This occurs when similar types of data are labelled inconsistently. The classes are too similar and cannot be differentiated by the model.
- Observer bias: This occurs when labellers let their subjective thoughts control their labelling goal, resulting in inaccurate data.

7.8.1 Limitation of data bias mitigation

The ability to remove all bias from datasets may result in outcomes undesirable for the development of an effective tool. Data bias mitigation aims to reduce the real-world implications that arise from bias found in datasets [b-EU-FRA-1]. Therefore, the ability to mitigate or reduce such harms to users is essential for creating more equitable tools whose benefits outweigh harms. However, undertaking the "correcting" or "fixing" of bias found in datasets (such as resampling and homogenizing) requires a substantial number of resources. In depth knowledge of the origins of the dataset, including its features and proxy variables, as well as an understanding and accurate labelling of protected characteristics would also be required [b-EU-FRA-2], [b-GOV.UK-1], [b-Hooker], [b-Gebru] and satisfactory attempts may only be achieved with datasets produced "in-house," rather than readily available open-source datasets [b-EU-FRA-1].

Technical limitations (stability)

- Usability (undesirable products)
- Knowledge of data quality
- Societal and historical presence of bias [b-Barocas]

Best practices:

- Understanding the origins and intentions of the dataset (data disaggregation) [b-UN-OHCHR]
- Establishing the stage of the data pipeline and identifying the biases that are likely to occur
- Acknowledging the existence of bias in society, organisations, etc. [b-Fitter]
- Considering feasible countermeasures to adjust for bias [b-Berthold]

Common issues

- Data diversity
- Data volume versus time for judging its asset
- Diversity of data sources [b-Cai]

Topic group use case:

Landslides of Masses of Soil and Rock: Intelligent Risk Management in Areas Highly Threatened by Climate Change. *The neural networks used in this case study to study complex landslide events must be trained with sufficient "real" cases, but many sites cannot be reached or observed, leading to gaps and biases in the data that must be addressed by the modeler.*

7.8.2 Bias pipeline in data

Each dataset encodes human biases [b-Kumar]. The most common issues that can be faced when dealing with data are:

- Poor organization: Going through data is fundamental for optimizing data.
- Having a huge amount of data: This is usually seen as a good starting point. However, a portion of it could be unusable and lead to a bad generalization.
- Inconsistent data.
- Poorly defined data.
- Incorrect data

7.8.3 Bias in popular datasets

In general, the complexity of the real world is too high to be described by any finite set of samples. Moreover, in most cases the data are collected for a specific task. Therefore, data, in this case, will cover just the specific environment/visual region and will not be sufficient for an exhaustive generalization of the real world [b-Tommasi]. Although popular datasets could be considered as solid and reliable, unfortunately in most cases this is not true. An interesting study carried out by [b-Conner-Simons] at MIT analysed some of the most exploited shared datasets, showing various labelling errors, such as mislabelled images, text sentiment and audio. Here are some examples:

- ImageNet: 20% of images were found to contain multiple objects, 5.83% incorrect labels,
- CIFAR: 5.85% had incorrect labels,
- QuickDraw: 10.12% had incorrect labels, and
- Amazon Reviews: had about 390,000 label errors (4%).

7.8.4 Bias in acquisition

In the context of data acquisition, bias refers to a systematic error or distortion in the data collection process that leads to inaccuracies or skewness in the resulting dataset. Bias can occur due to various factors, including the methodology used, the selection of data sources, or the inherent limitations and assumptions of the data collection process. Following are some of the possible causes for bias in acquisition:

- Non-random sampling errors,
- Landscape biases in United States geological survey (USGS) gage locations [b-Deweber],
- Signal distortions (e.g., satellite imagery acquired in high slope areas, mainly an issue for synthetic aperture radar imagery, but also present in optical imagery), and
- Impossibility for an image to represent all real-world properties.

Best practices:

- The Data Ethics Commission suggests "promoting technologies to enforce the law and uphold ethical principles in the world," stressing the point that harmful outcomes arising from AI tools ought to be avoided [b-DEC]. While the legislation does not yet enforce auditable practices in the use of data in AI, a key step to mitigating data bias is to understand the data that are used and to develop internal practices that document data quality [b-Ada].

- Understand the data used and the influence it has on the product outcome (EDA) [b-AI HLEG-1].
  - ○ Review the quality and limitations of the data.
- What processes are in place to ensure and maintain data integrity? Are all metadata and field names clearly understood? Why are these data needed; can this be explained to members of the public? [b-GOV.UK-2]
- Take measures to include auditable and accountable steps [b-AI HLEG-2].
  - ○ Technical audit: A narrowly targeted test of a particular hypothesis about a system by looking at its inputs and outputs; for instance, seeing if it exhibits racial bias in the outcomes of a decision.
- Accountability and transparency in algorithms relate to how designers of algorithms are held responsible for outcomes of their products, throughout the developmental pipeline [b-Leslie].
  - ○ General Data Protection Regulation (GDPR) states that the data subject has the right to obtain "(in) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved" [b-EU-GDPR-1].

Topic Group use case:

Limitations of predicting snow avalanche hazards in large data sparse regions. *This case study examined the challenges in obtaining ground truth data that could describe the true likelihood of avalanches across space and time. They also stress the importance of communicating complex data and uncertainties to avalanche forecasters, particularly when it comes to incomplete or inconsistent data.*

Iteration and Revision

- Those who collect and use data ought to also maintain and update systems that accommodate data [b-UN-OHCHR].

Resources:

1) Auditability
   - AI Fairness 360 – IBM
   - Fairlearn – GitHub
   - Know Your Data – Google
   - TensorFlow Data Validation
2) Know Your Data
   - Datasheets for Datasets – Microsoft Research
   - AI Blindspot cards – MIT Media Lab

8 Data visualization for AI/ML

Data visualization is the application of algorithms to generate images, maps, animations and videos (graphical and pictorial) from data to facilitate recognition and reaction to that information. However, as with any visual approximation, to be productive, it requires exploring the best option among many, which in several cases could result in the application of a multifunctional instrument with capacities for both visualization and analysis [b-Andrienko]. For the AI, whose objective is for algorithms that "understand" and "answer" to data, mimicking human behavior or doing better, it could be mistakenly understood that as AI matures, the necessity to understand data becomes irrelevant. AI innovation is still a human effort, and the developers must thoroughly exploit data visualization in each of its stages:

- Information pre-processing (interpret databases),
- Model training (clarify learning process), and
- Presentation of results (display the efficiency and the adjustment to a certain task).

Pre-processing data

The development of intelligent applications begins with data exploration or Exploratory Data Analysis (EDA). Exploration and analysis of data refers to the critical process of performing preliminary investigations on data to uncover patterns, to notice anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations [b-Jambu]. EDA is about making sense of data before applying modeling techniques to them. The reasoning from EDA suggests the next logical operations, questions, or areas of research for an AI project. The most accepted process is to investigate a dataset using multiple exploratory techniques for comparing findings and confronting conclusions. The general steps in EDA [b-Mukhiya], [b-Baillie], [b-Heckert] are:

- Identification of variables and data types,
- Analysis of basic metrics,
- Non-graphical univariate analysis,
- Graphical univariate analysis,
- Bivariate (or multi) analysis,
- Variable transformations,
- Missing value treatment,
- Outlier treatment,
- Correlation analysis, and
- Dimensionality reduction.

A deep understanding of the different data types is a crucial prerequisite for using AI models. Almost any aspect in life (in nature and in anthropic environments) can be turned into data and the way they are expressed and how they are introduced as variables will affect the final model. Most data can be categorized into four basic types from an AI perspective (Figure 20): numerical data (continuous or discrete), categorical data (numerical values, class labels), timeseries data (sequence of numbers collected at regular intervals over some period) and text (words). The basic metrics are dependent on the data type and the task (target) of the intelligent model. A simple and easy way to start could be using descriptive statistics, as they are brief descriptive coefficients that summarize a given dataset. Measures of central tendency (mean, median and mode) and measures of variability (standard deviation, variance, minimum and maximum variables, kurtosis and skewness) enable the modeler to start to understand the data [b-Bradley], [b-Downey]. Through the interpretation of the key statistical concepts (tendency, dispersion, symmetry and correlation), EDA is developed (Figure 21).

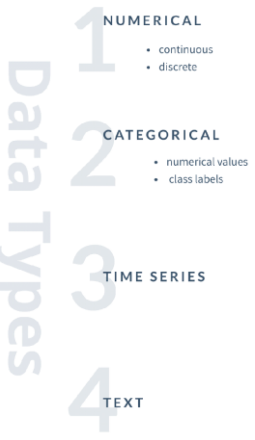


Figure 20 – Four basic data types from an AI/ML perspective[8]

The univariate non-graphical EDA is used for knowing the range of values and frequency of each value. For quantitative variables, for example, the ML expert looks at the location (mean, median), spread (interquartile range, standard deviation, range), modality (mode), shape (skewness, kurtosis) and outliers. The graphical univariate analysis, on the other hand, is applied to define the distribution of data. On that note, an approximation of the data distribution requires that it be of quantitative nature. The most used methods are histograms, density plots (a smoother version of a histogram), bar charts (useful for categorical variables) and boxplots (for numerical data). Boxplots (i.e., box-and-whiskers plots) are excellent in representing information regarding central tendency, symmetry, skew and outliers. When two categorical variables are studied, the most used EDA plot is the grouped bar plot (side-by-side boxplot), but if they are quantitative variables, a scatterplot would be ideal. In some cases, it may easily reveal some pattern like whether a linear, polynomial (even exponential) or other relationship exists between the two quantitative variables in the plot. The pairplot provides the possible combination of scatterplots between permutations of any two quantitative variables. As a bonus, it adds histograms on the main diagonal of individual quantitative variables.

[8] The application of certain techniques to understand variables and intelligent models is dependent on the data type and the task (target) of the intelligent model.

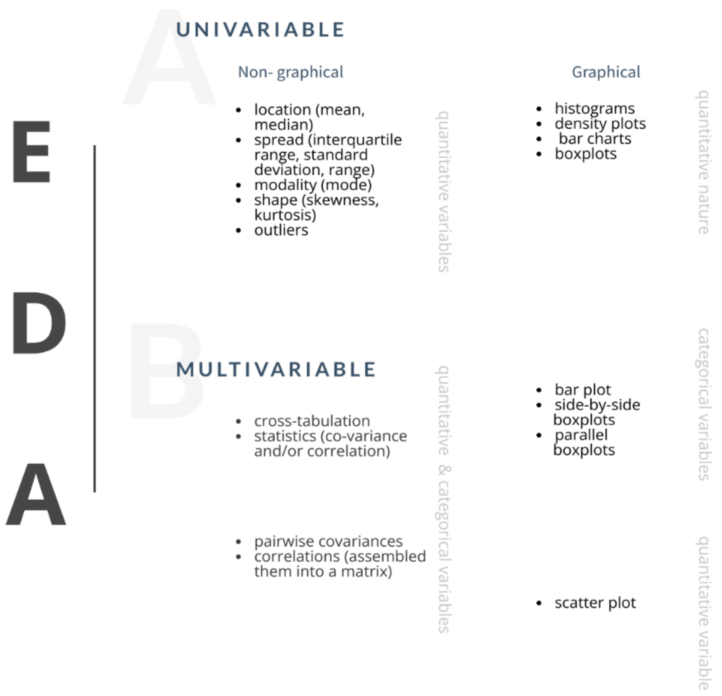


Figure 21 – Graphical and non-graphical EDA techniques[9]

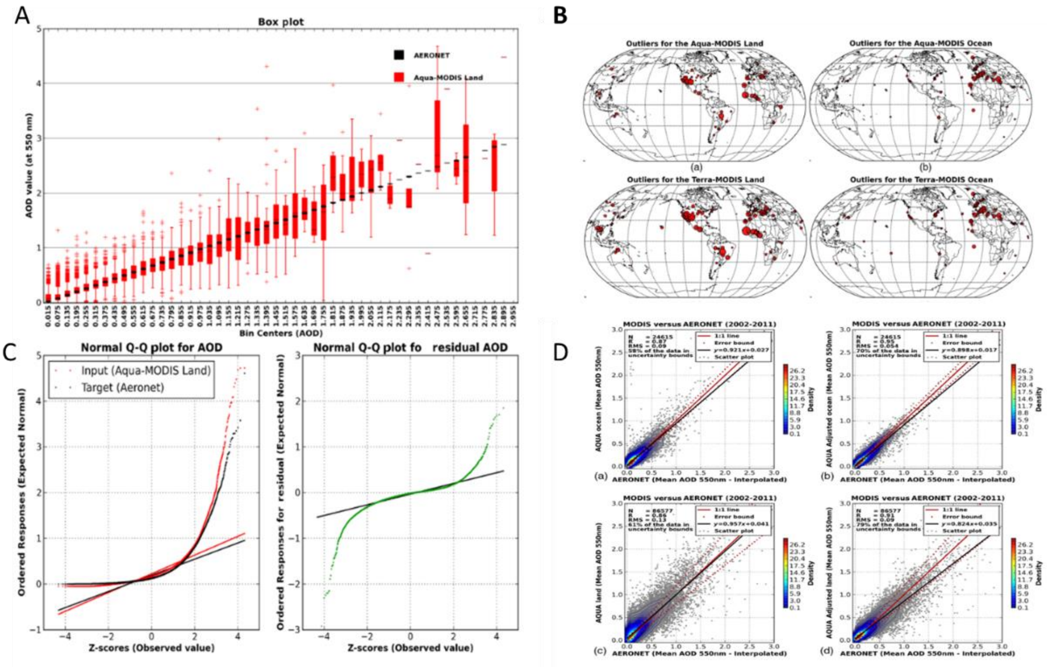


Figure 22 – Example of preliminary analysis of data [b-Albayrak][10]

[9] Their selection and application are dependent on the variable type.

[10] This diagram indicates: (a) box plots for binned AERONET AOD data; (b) global locations of the outliers for Aqua-MODIS Land, Aqua-MODIS Ocean, Terra-MODIS Land, and Terra-MODIS Ocean; (c) quantile–quantile (Q-Q) plots for (left) Aqua-MODIS Land versus AERONET and (right) Aqua-MODIS Land residuals; and (d) Scatter plots for MODIS-Aqua Aerosol Optical Depth at 550 nm where the first row is the ocean, and the second is the land. The x-axis of each plot is the AERONET AOD and y-axis is the MODIS data. (a and c) before NN adjustment; (b and d) after NN adjustment.

Topic group use case:

Landslides of masses of soil and rock: Intelligent risk management in areas highly threatened by climate change. *The integrity and quality of the data are double-checked using EDA to find patterns, notice anomalies, test hypotheses and check assumptions. This enables the modeler to have a better understanding of the data and thus determine the most suitable methods for dealing with data bias through transformations, augmentations and/or reductions.*

On many occasions, it is necessary to convert raw data into a format or structure that would be more suitable for model building (facilitating discovering). For example, when the algorithm is likely to be biased because a data distribution is skewed, scale allows the algorithm to better compare the relative relationship between data points. When implementing supervised algorithms, training and testing data need to be transformed; this is usually done using a transformation algorithm on the training and testing data. The most applied converting techniques include log transformation (for transforming the skewed distribution into normal distribution), clipping methods (for setting up the upper and lower bounds) and data scaling (for converting the data into the same scale). The concept of missing data is implied in the name: data that were not captured reducing the statistical power of the analysis, which can distort the validity of the results. When dealing with missing data, two methods can be used: imputation (developing reasonable guesses for inferring the value; for doing this the percentage of missing data must be low) or elimination of whole observations (entire files or examples). For working with outliers (data points that differ significantly from other observations in given dataset), it is important to define if they are artificial or natural. Entry errors (measurement, experimental, intentional, processing, or sampling) are common causes of artificial ones, while the natural (most real-world data belong to this category) are not considered errors but novelties[11] in data. Some techniques used to deal with outliers are transforming values, imputation, separately treating, or deleting observations [b-Aguinis].

The tendency of simultaneous variation between two variables, called correlation or covariation, must be detected before training any intelligent model. The degree of association or relationship between two variables is measured with an index termed as correlation. Some of the methods to compute this value are the scatter diagram method, Pearson's product moment and Spearman's rank order [b-Hansford]. Dimensionality reduction refers to techniques that shrink the number of input variables in a dataset. More input features often make a predictive modeling task more challenging, generally referred to as the curse of dimensionality. There are mainly two types of dimensionality reduction methods, the one that keeps the most important features in the dataset and removes the redundant features (e.g., no-transformation, backward elimination, forward selection and random forests) and the method that finds a combination of new features. This last method can be further divided into linear [e.g., principal component analysis (PCA), factor analysis, linear discriminant analysis and truncated singular value decomposition] and non-linear methods [kernel PCA, *t*-distributed stochastic neighbour embedding (SNE), multidimensional scaling and isometric mapping, among others] [b-Jiang].

Presentation of results

To ensure the safe and effective functioning of AI models, developers need to be able to evaluate how well the model performs at its designated task. Data visualization is especially helpful in this evaluation because models often work on a range of behaviours whose outcome cannot be evaluated based at a single point, but rather as a trade-off curve or surface or hyper-surface, normally understandable only qualitatively (via visualization) rather than numerically as a score. As examples,

[11] Novelties are also **anomalies in data, but they only exist in new instances**. They do not reside in the original dataset. The presence of outliers, anomalies, or novelties does not imply a change in the underlying data distribution or regime.

receiver operating characteristic (ROC) curves are used to evaluate the results of classification algorithms (Figure 23) and silhouette plots (Figure 24) are used for clustering.

Once an AI system has been developed and performs to the satisfaction of its creators, a final critical hurdle needs to be cleared up before it can be used to automate any real-world tasks: humans must be convinced that this is safe and profitable. People who are involved in approving AI systems are often those who currently perform similar tasks and want to know why an AI system responds to data the way it does, couched in terms they rationalize. This "interpretability" requirement has generated alternatives as Rivelo [b-Tamagnini] or local interpretable model-agnostic explanations (LIME) [b-Ribeiro] that visually explain individual predictions of very complex models (regardless of the model structure). The value of providing explainable AI (XAI) tools and services is multi-fold, and especially relevant for responding to current and future world-wide regulations [b-EU-GDPR-2] that mandate the provision of explanations for individuals who may be affected by automated decisions. Every day new XAI tools appear. Most of them try to clarify the results, particularly those that have to do with behaviour prediction, in terms of the weight of the inputs over the outputs in multidimensional environments that are easy (as much as possible) to understand. Shapley additive explanations (SHAP) [b-Lundberg], for example, explain individual predictions based on the game theoretically optimal Shapley values, providing global and local interpretations. For more information about XAI, please see the "AI for Modeling" companion report.

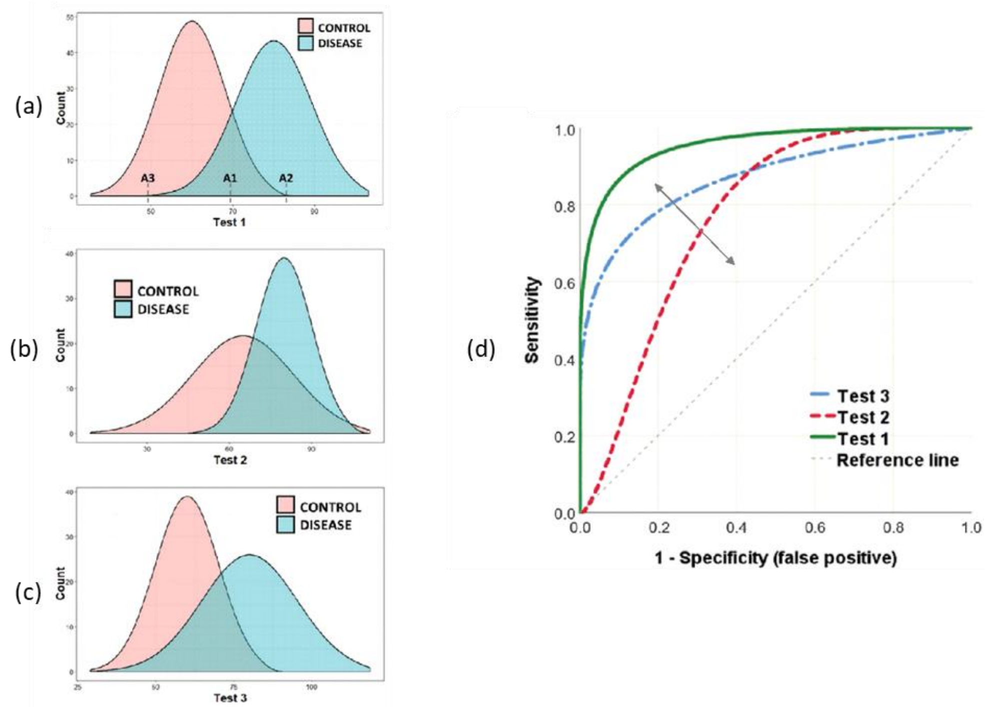
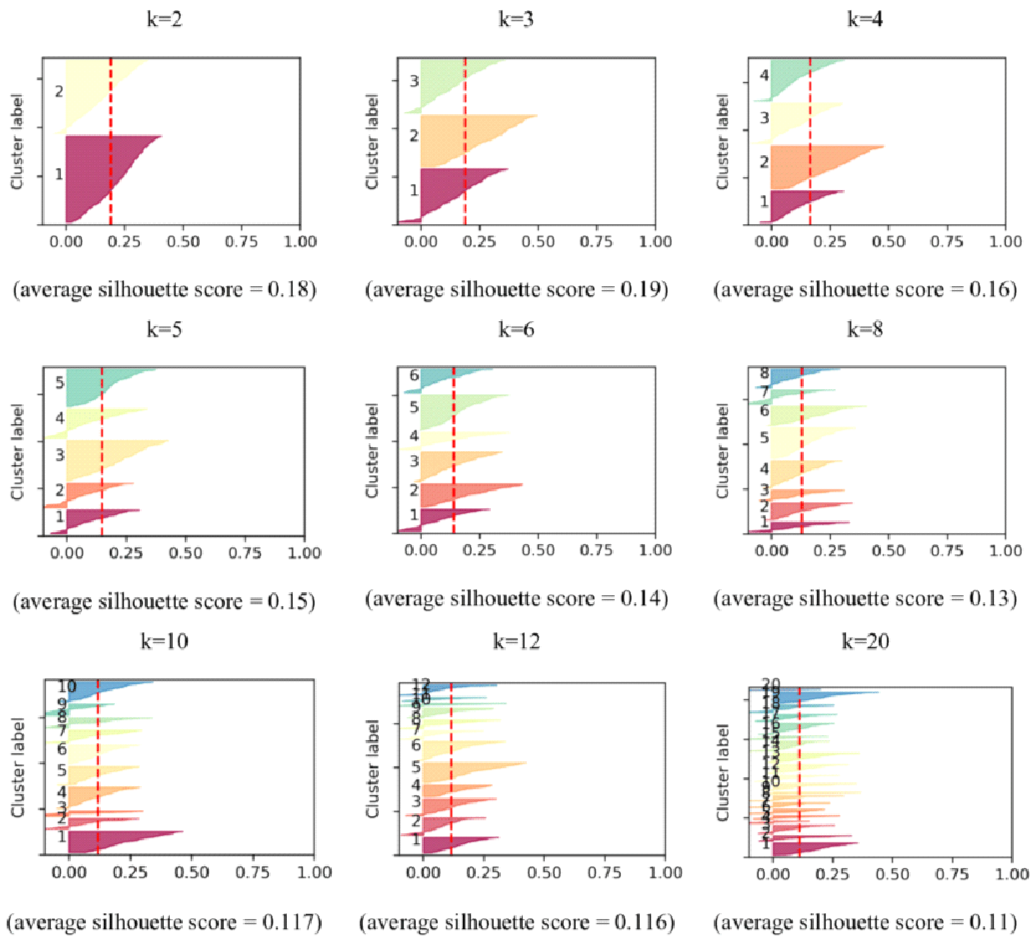


Figure 23 – Example of an ROC[12]

[12] (a)-(c) Illustrations of the distribution curves for the results of tests (tests 2 and 3) to classify patients as sick or healthy. (d) The results of tests 1, 2, and 3, plotting their sensitivities against their false positive rates.



(average silhouette score = 0.18) (average silhouette score = 0.19) (average silhouette score = 0.16)

(average silhouette score = 0.15) (average silhouette score = 0.14) (average silhouette score = 0.13)

(average silhouette score = 0.117) (average silhouette score = 0.116) (average silhouette score = 0.11)

Figure 24 – Example of electricity readings (in Watts from apartments located in western Switzerland)[13]

Best practices:

Visually depicting data often makes it easier to understand and draw insights. In AI/ML, data visualization is an effective means of making data more accessible but also of making the performance of algorithms transparent for the purpose of identifying anomalies and generating trust. The first stage where data visualization algorithms create images from raw data and display hidden correlations must be completed with additional steps of evaluation or metrics for learning and model functioning. Normally, an NDM intelligent model involves decisions, and a deep understanding of reasons behind them requires absolutely connecting data to conclusions. This includes how they were included as inputs and why (what variations were tried and their effects on the outputs), why an AI platform delivers a certain answer, and what the dependencies of confidence are on the results.

Concerns about potential misuse or unintended consequences of AI have prompted efforts to examine and develop more algorithms, tools and standards to create building blocks for reliable, robust and trustworthy AI systems. In NDM, AI/ML modelers must exploit the visualization tools in every stage of the development of their systems not only to generate better models and confidence in the outputs, but also to recognize the risks specifically created by the AI applications.

[13] Their silhouette analysis for k-means clustering using the daily load shape. Varying the number of the cluster (from k = 2 to k = 20), these plots show the silhouette score for every load profile grouped and coloured by the cluster label (vertical line shows the average silhouette score for all the observations); silhouette score is highest when "k" is equal to 3.

Topic group use case:

Landslides of masses of soil and rock: intelligent risk management in areas highly threatened by climate change. *Data visualization is utilized at each stage of the process: first to interpret databases in the pre-processing stage, to clarify the learning process in the model training stage, and finally to present the results. Instructions and documentation are also included to ensure that other users are able to evaluate the outputs and make use of it for their specific purposes.*

9 AI/ML and data standards

9.1 OGC

The benefits of data standardization are consensus on data format, semantics and interpretation, providing interoperable tools for organizations, businesses and markets. Data standards can help organizations of all sizes quickly adapt data to AI systems with ML components and reduce risk through unambiguous testing. Transparency in the data lifecycle based on data standards can also promote innovation.

The Open Geospatial Consortium (OGC) is an international, non-governmental, non-profit standards organization dedicated to the industry standardization of geographic information, founded in 1994. The OGC provides a global hub for geospatial communities to develop open and market-driven standards for geospatial content and services, data processing and data sharing. The OGC standards cover not only fundamental concepts and architectures as OGC Abstract Specifications but also technical structures and interfaces as OGC Implementation Standards. A typical OGC architecture is the OGC Web Services (OWS) architecture, which includes data models, encoding formats, access services, processing services, discovery services, visualization and sensor web enablement. As data format standards, for example, geography markup language (GML) [b-OGC-3] and keyhole markup language (KML) [b-OGC-6] specify extensible markup language (XML) grammars for encoding geographic vector and raster data and transmitting it via the Web protocol. GeoPackage [b-OGC-1] describes a compact storage format for platform-independent data transfer by leveraging an SQLite database. GeoTIFF [b-OGC-2] is a data format for sharing georeferenced or geocoded raster imagery based on Tagged Image File Format (TIFF), a platform-independent file format for image data. The OGC GeoTIFF is commonly used to distribute satellite and aerial imagery within the geospatial science community, as well as NetCDF [b-OGC-7] and HDF5 [b-OGC-4]. NetCDF provides a data model and a machine-independent format for array-oriented scientific data, especially representing spatially and temporally varying geographic phenomena. HDF5 also defines a data model and a data storage format for spatial- and time-varying phenomena by employing multidimensional numeric arrays. An OGC discussion paper [b-OGC-5] addresses the problem of data formats to share point cloud data with point-level semantic label information annotated to each point, defined as labelled point clouds and proposes the HDF5 profile as an open standard data format. The HDF5 is one of the most used formats for non-tabular or numerical data in AI systems, not only in geospatial domains, because of its flexibility in representing structured and unstructured data with complex relationships. While many data formats have been developed and standardized by the OGC, they are not designed or optimized for training, validating, or testing machine learning data.

In 2018, OGC established a new Domain Working Group (DWG), Artificial Intelligence in Geoinformatics DWG (GeoAI DWG), to identify and deliver new requirements of data standards with various use cases and applications related to AI in geospatial domains such as healthcare, smart home and autonomous cars. The GeoAI DWG found three main challenges and issues as follows:

- Data-driven learning models: Modern deep learning (DL) relies more on the data than program logic. Data standards that support interoperability, reusability, discoverability and quality from heterogeneous geospatial data are increasingly essential to efficiently adapting geospatial data to DL algorithms.

- Compositionality of constituents to assemble or transfer deep stacks of learning: The main difference between traditional machine learning and DL approaches is the composition of models assembled in different applications. Thus, standardization of model data is also an important factor in ensuring interoperability, adaptability, discoverability and quality of geospatially oriented DL models.
- Not enough benchmark datasets and business use cases that can offer considerable benefits: Although many geospatial data are published and shared based on OGC data standards, there are not enough benchmark training datasets like ImageNet. Many stakeholders are still struggling to identify business use cases to promote considerable benefits. Best practices and guidelines with common AI toolchains and geospatial data can help developers to build reliable and robust AI systems in geospatial domains.

With new requirements for training data in geospatial analysis based on machine learning, OGC has initiated a new Standard Working Group (SWG), Training Data Markup Language for AI SWG (aka TrainingDML SWG), to develop new OGC data standards for sharing Earth Observation (EO) machine learning datasets. The new data standards will formalize and document ML training data, considering content, metadata, data quality and provenance, etc. OGC is actively working to build a strong foundation of open standards for ML data sharing and encourages the development of open tools, workflows and best practices to ensure AI technology's safe, responsible and ethical use in geospatial applications with OGC standards.

Topic group use case:

Exploring deep learning capabilities for surge predictions in coastal areas. *NetCDF files containing atmospheric variables are downloaded from ERA-517, a high-resolution climate reanalysis dataset from the European Centre for Medium-Range Weather Forecasts. The datasets and outputs from this study are available as open data, including the predictor and predicted variables which are stored as NetCDF files to facilitate interoperability.*

9.2 ISO/TC 211

The International Organization for Standardization's (ISO) Technical Committee on Geographic Information (TC211) aims to develop international standards for geospatial information. The concept of standardization in ISO/TC211 is not to unify spatial data formats, but to standardize the use of diverse types of spatial data in different GIS applications. ISO/TC211 published core concepts that can be applied to other international standards and technical specifications in the field of geographic information. For example:

- ISO 19101-1 defines the ISO reference model, which guides structuring geographic information standards to enable the universal usage of digital geographic information.
- ISO 19103 provides rules and guidelines for using a conceptual schema language (e.g., the Unified Modeling Language) within the context of geographic information.
- ISO 19104 specifies requirements for developing terminological entries, such as the collection, management and publication of terminology, in the field of geographic information.
- ISO 19105 specifies the framework, concepts and methodology for conformance testing and criteria, based on concepts defined in ISO standards regarding geographic information.
- ISO 19106 defines the concept of a profile of the ISO geographic information standards developed by ISO/TC 211 and guides creating such profiles.

ISO/TC211 has also conducted common interpretations of geospatial phenomena and guides the conceptual data models and encoding schemas for digital geographic information. For example:

- ISO 19107 specifies conceptual schemas for describing the spatial characteristics of geographic entities (especially "vector" geometry and topology) and a set of spatial operations consistent with these schemas.
- ISO 19108 defines concepts for describing temporal characteristics of geographic information, such as temporal feature attributes (and operations, associations) and temporal aspects of metadata.
- ISO 19109 defines rules for creating and documenting application schemas, including principles for determining features, which is a fundamental concept of geographic data.
- ISO 19115 describes the metadata schema for cataloguing geospatial datasets and services to support data discovery.
- ISO 19123 defines a conceptual schema for the spatial characteristics of coverages, including the relationship between the domain of a coverage and an associated attribute range.
- ISO 19137 defines a core (which is intentionally small and limited to be easy to understand) profile of the geometry part of the spatial schema specified in ISO 19107, in accordance with ISO 19106.

Input data quality directly affects the quality assurance of AI systems by incorporating ML components. ISO 19157 describes a conceptual model of geographic data quality and data quality measures but does not consider the use cases of AI systems.

9.3 ISO/IEC JTC 1/SC 42

With new requirements for promoting highly reliable AI, in October 2017, ISO and IEC Joint Technical Committee (ISO/IEC JTC1), which is responsible for international standardization of information technology (IT), decided to establish the Subcommittee on Artificial Intelligence (SC 42). The scope of SC 42 is to "Serve as the focus and proponent for JTC 1's standardization program on Artificial Intelligence" and "Provide guidance to JTC 1, IEC and ISO committees developing Artificial Intelligence applications" [b-ISO-1]. ISO/IEC JTC 1/SC 42 has published three international standards and six technical reports related to AI. Also, two international standards and three technical reports related to Big Data were published by SC 42 after disbanding WG 9 Big Data. A number of standards are in development by SC 42 working groups [b-ISO-1]. In particular, WG 2 Data is responsible for standardization related to data in the context of artificial intelligence, Big Data and data analytics.

Figure 25 shows the past and current standardization projects by SC 42/WG 2. The ISO/IEC 5259 series will provide a holistic approach to handling data quality issues in the context of data analysis and machine learning, comprising foundation concepts, data quality measures, data quality management requirements and guidelines, data quality process frameworks for various types of analytic and machine learnings, and data quality governance frameworks. ISO/IEC 8183 will provide "an overarching data life cycle framework that is instantiable for any AI system from data ideation to decommission" [b-ISO-2]. It is applicable for the use of data across different levels of an organization with common terminology and processes.

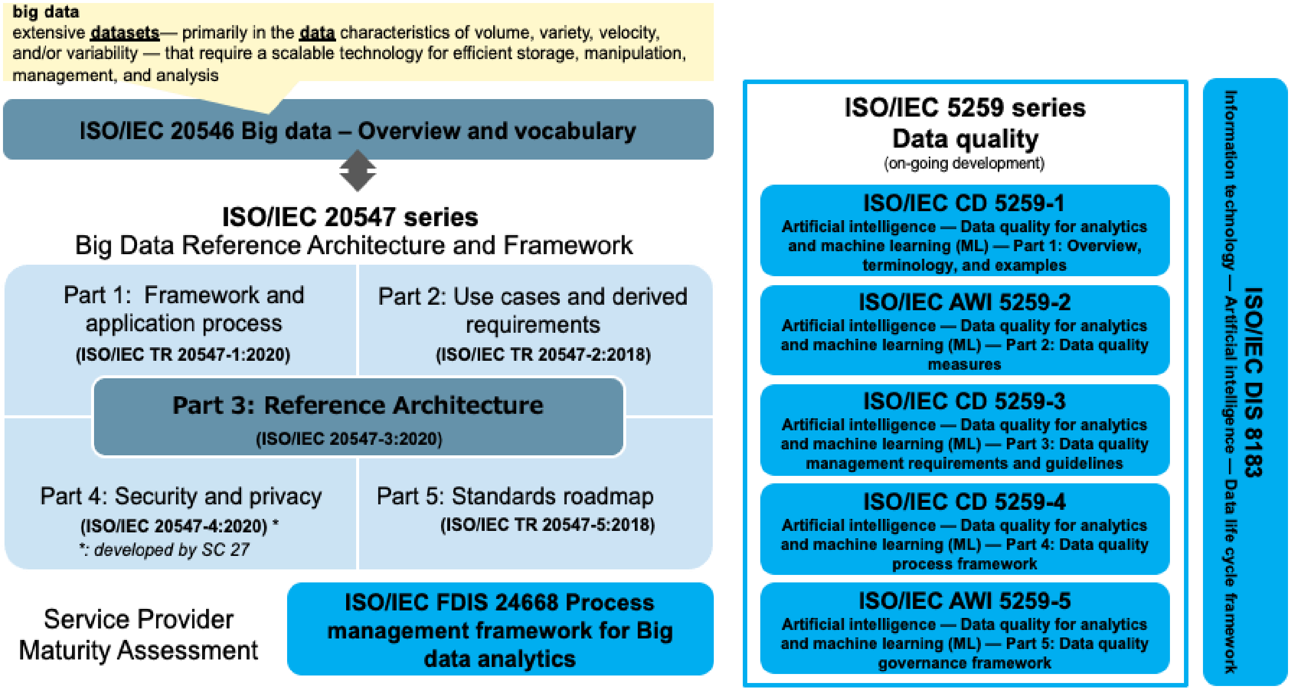

big data
extensive datasets— primarily in the data characteristics of volume, variety, velocity, and/or variability — that require a scalable technology for efficient storage, manipulation, management, and analysis
ISO/IEC 20546 Big data – Overview and vocabulary
ISO/IEC 20547 series
Big Data Reference Architecture and Framework
Part 1: Framework and application process
(ISO/IEC TR 20547-1:2020)
Part 2: Use cases and derived requirements
(ISO/IEC TR 20547-2:2018)
Part 3: Reference Architecture
(ISO/IEC 20547-3:2020)
Part 4: Security and privacy
(ISO/IEC 20547-4:2020) *
*: developed by SC 27
Part 5: Standards roadmap
(ISO/IEC TR 20547-5:2018)
Service Provider Maturity Assessment
ISO/IEC FDIS 24668 Process management framework for Big data analytics
ISO/IEC 5259 series
Data quality
(on-going development)
ISO/IEC CD 5259-1
Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples
ISO/IEC AWI 5259-2
Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures
ISO/IEC CD 5259-3
Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines
ISO/IEC CD 5259-4
Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework
ISO/IEC AWI 5259-5
Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 5: Data quality governance framework
ISO/IEC DIS 8183
Information technology — Artificial intelligence — Data life cycle framework

Figure 25 – Past and current standardization projects by SC 42/WG 2

10 Policy, ethics and legal issues

As policies, recommendations and guidelines relating to the creation and use of responsible AI tools and products continue to be published, data play a fundamental role in achieving that aim. Policy, ethics and legal concerns that centre around data governance include considerations of how data are collected, stored and used, whilst adhering to legal regulations such as General Data Protection Regulation (GDPR) and applying trustworthy or fair approaches (Figure 26). These concerns are multi-faceted, with each presenting a specific interpretation owing to geopolitical, cultural, or regulatory sensitivities. This section will provide a descriptive overview of the subject matter as it pertains to data.

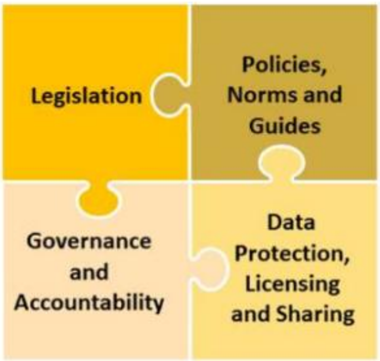

Legislation
Policies, Norms and Guides
Governance and Accountability
Data Protection, Licensing and Sharing

Figure 26 – Overview of key elements for the responsible and effective use of geospatial information [b-UN-GGIM-3]

The resulting impacts of AI are now global; therefore, considerations to policy, ethics and legal approaches ought to include diverse and cross-cultural considerations. Legal aspects, such as adopting human rights law, such as the Universal Declaration on Human Rights (UDHR), provide a baseline for other national legal frameworks to be placed upon [b-Latonero]. Enactment of policies cross laterally fosters collaboration between nations and government bodies, increasing the diversity of stakeholders and their contributions [b-Galindo]. AI affects human interaction and, in turn, shapes the way that cultures and values are expressed [b-IEEE].

Geopolitics will shape how data are used for AI, with resulting tools and products influencing the course of geopolitics. For example, facial recognition products that are driven by personal biometric data are facing increased regulation in Europe under GDPR law due to instances of the private collection of biometric data being acquired for commercial purposes without a lawful basis, breaching fundamental rights [b-Kroet]. However, owing to their extensive use in law enforcement the international exchange of biometric data and facial recognition products is a multi-billion-dollar industry [b-Newcombe]. This global exchange can cause oversight or lapse in the enactment of home or host nations' legal or policy-related frameworks [b-ANE].

Frequently cited considerations:

- **Data governance:** This includes data collection, data ownership, data sharing and data access. Effective policies and regulations help to address issues that arise from the transference and transformation of data and their consequential impact across societal, political and economic factors. Most notably, the geospatial implications of the usage of data in AI require robust governance to mitigate national exposure to risk [b-UN-GGIM-3].
- **Trustworthiness:** Throughout the life cycle of data used in AI, legal, ethical and robust measures should be met to increase levels of trustworthiness and mitigate intentional or unintentional outcomes. This can be achieved through auditability practices, technical robustness and data governance (Figure 27) [b-AI HLEG-2].
- **Accountability:** Accountability seeks to ensure that responsible and auditable uses of data are maintained both during and after the development of the tool or product. This can be achieved through standards, frameworks, or certification [b-AI HLEG-2].
- **Explainability:** Differing degrees of explainability are required to provide insight into how data are used and their functionality in the AI tool or product. This level of context allows stakeholders to better understand the outcomes of finalized products and to challenge any possible negative factors [b-UN].
- **Human rights and democratic values:** Human rights and democratic values cover aspects relating to human dignity, autonomy, data protection and privacy to ensure the adherence to fundamental principles. For data, this may include best practices regarding data acquisition, upholding granted permissions and the repurposing of data for other means [b-UN].
- **Safe/secure AI:** Safety seeks to mitigate harm and risk from malicious use such as data breaches or cyberwarfare. These harms can be the result of intentional or unintentional misuse and, therefore, require cyclical assessment [b-UN].
- **Data protection:** AI is relevant to the data protection principles embodied in the General Data Protection Regulation (GDPR). Since transparency is required by data protection (law), ethical data collection practices are needed. Additionally, GDPR places a restriction on the unwarranted mass collection of data beyond reasonable use [b-UN].
- **User data rights:** The protection of user power over access and uses of data ensures that digital autonomy and control remains within the reach of users. This can be achieved through permissions being explicitly granted or restricted if data are to be used for purposes outside of the agreed initial intention of the user [b-IBM].
- **Harm avoidance:** It is important to ensure that the data-based product or service, while creating the intended values for clients and provider, does not harm individuals. This can be achieved by training and testing datasets that are representative and relevant, as well as ensuring that impacts and outcomes of developed products are measurable [b-Leslie].

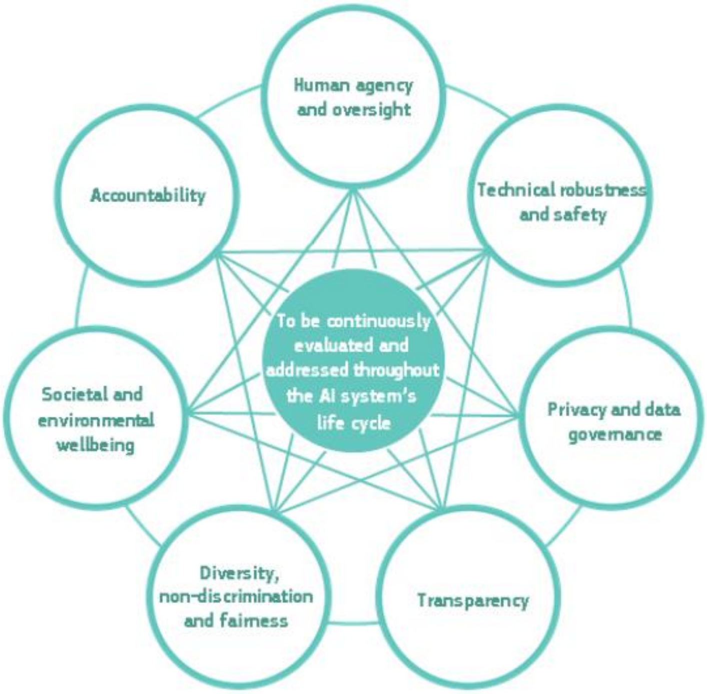

Human agency and oversight
Technical robustness and safety
Privacy and data governance
Transparency
Diversity, non-discrimination and fairness
Societal and environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle

Figure 27 – Requirements for trustworthy AI [b-AI HLEG-2]

Topic group use case:

Enabling natural hazards risk information sharing using derived products of export-restricted real-time GNSS data for detection of ionospheric total electron disturbances. *A data governance system is adopted to verify the integrity of training data from global navigation satellite system (GNSS) satellites and labels assigned by experts. Ultimately, the model will be regularly re-trained via a human-in-the-loop process in which experts contribute feedback to improve the model and keep it up to date.*

11 Open-data, open-source guidelines and policies

The scientific community and the general public thrive on the foundation of open-source software (OSS). OSS has contributed to major scientific findings and collaborations from around the world and has enabled many modern conveniences such as mobile phones. OSS is utilized heavily because their communities maintain and extend functionality in transparent, well-documented ways. Research and development of AI-based tools and software for natural hazards detection will – by employing readily available and documented approaches with newly developed approaches – significantly leverage existing OSS available on platforms such as (but not limited to) GitHub, GitLab, R-CRAN, the Python Package Index (PyPI), Anaconda and others.

AI-based tools and software developed for natural hazards detection, forecasting and communication that contribute to and publish OSS maintain high standards of documentation, transparency and reproducibility of work demanded by OSS communities and users and defined by many existing software development standards. Additionally, tools and software development under OSS principles and practices often develop into higher levels of capability when compared to tools developed exclusively within organizations. Applications of high-stakes AI [b-Sambasivan] like those used in natural hazards detection benefit from employing OSS principles by helping to reduce the likelihood of data cascades. At the very least, research and development of AI-based natural hazards systems should be well documented with results reproducible by outside parties, with any data and related code shared publicly on a readily accessible third-party distribution platform. Ideally, the platforms leveraged for OSS conceptualization, research, development and support provide a) code with an appropriate license and pointers (links) to data and documentation in a publicly accessible way, b) tools to report issues (e.g., bugs and todos), c) high standards of documentation (e.g., transparency, reproducibility, usability and reliability), which is required by a variety of stakeholders (users, maintainers, developers and acquirers) and d) encouragement of community participation and engagement between users and developers of OSS. While many foundations such as the Apache Foundation self-host code and communities for their software development, utilizing a third-party, well-known distribution platform will encourage greater community participation and engagement. Examples of existing distribution platforms include (but are not limited to) GitHub, GitLab, Anaconda, the Python Package Index (PyPI), or R-CRAN. While some distribution platforms are agnostic to language or computing platforms (e.g., GitHub, GitLab), others such as R-CRAN or PyPI are specific to the R and Python programming languages, respectively. With respect to AI specifically, platforms such as Kaggle could be leveraged to encourage the community to develop new and improved AI models through competition. Kaggle competitions leverage available data, thorough documentation and a description of the end goal (usually with prizes) to encourage community-based AI development and participation. Open-source licenses allow software to be freely used, modified, further developed and shared with developers and users. Several open-source licenses have strong communities, with well-understood definitions of use: the Apache 2.0 license, the MIT license and the BSD 3-Clause license to name a few. Open-source licenses provide those developing and sponsoring OSS with the means to not only freely share work, but also to share work in a way the "rules of engagement" are understood by everyone. An in-depth resource of available OSS licenses can be found at Opensource.org.

The versioning scheme utilized in OSS is software specific and chosen based on each software's unique considerations. One common versioning scheme is semantic versioning [b-Preston-Werner]. In semantic versioning, major versions are first incremented (e.g., 1.x.x), followed by minor revisions (e.g., 1.2.x) and patches (e.g., 1.2.3). The vast majority of OSS leverages semantic versioning.

Topic group use case:

Artificial intelligence modeling tools for monitoring desert locust (AI-Locust): Breeding grounds, hatching time, population and spatio-temporal distribution. *This use case emphasizes the importance of proper data management when adhering to open science and open access principles. Their project results and products will be available under an open-source by-attribution license (CC-BY 4.0) with proper metadata and documentation so that users can verify, replicate and reuse the data.*

12 Conclusion

This Technical Report aims to introduce the current methods and technologies used for data collection and provide an overview of how AI-based methods can be utilized for processing and analysing data. It also proposes best practices to improve data management, curation, bias, security, ethics and other important considerations. Furthermore, the report gives a glimpse of the near future and future technologies that might impact the 5V's of data.

This topic is particularly timely. In the last few years, both the volume of data and the complexity of algorithms used to process and analyse data have grown exponentially. In parallel, the increased frequency and magnitude of extreme weather events due to climate change and ambitious goals of the United Nations Early Warnings for All Initiative have brought great demand for AI-based tools. However, in the high-stakes realm of natural disaster management, such tools must meet certain standards to ensure their safety. For this reason, efforts to standardize and provide best practices guidance on how to best utilize large quantities of data and complex AI methods are long overdue for this domain. This report seeks to provide a foundational first step in this direction, and is intended to serve as a reference for those working with AI-based methods for data collection, management and analysis; those evaluating such systems (e.g., in policy and regulation); and those deploying such tools.

As with the complementary technical reports on "AI for Modeling" and "AI for Communications," this report was bound by the broader limitations faced by the focus group in terms of time, resources and scope, and by no means represents an all-encompassing, final word on this topic. Thus, this technical report will continue to evolve to include future developments and implementations in the field. Future research areas of interest include topics such as Explainable AI (XAI), federated learning techniques, and advances in Earth sensing technologies (intelligent satellites, smartphone accelerometer data, etc.). Projects such as MedEWSa, an EU Horizon project tasked with developing an interoperable Multi-hazard impact-based Early Warning Systems covering the European-Mediterranean-African region, provide opportunities to test and update these best practices. The outcomes of such implementations, as well as developments in the topic group use cases, are expected to provide further guidance for future iterations of this report and enable the creation of more comprehensive standards.

13 Topic group (TG) use cases

WG-Data suggests the following structure (see Figure 28) for TG to consider under their corresponding sections. WG-Data is aware that requirements might change depending on the problem in hand. For further definitions and explanations, please go to the corresponding section.

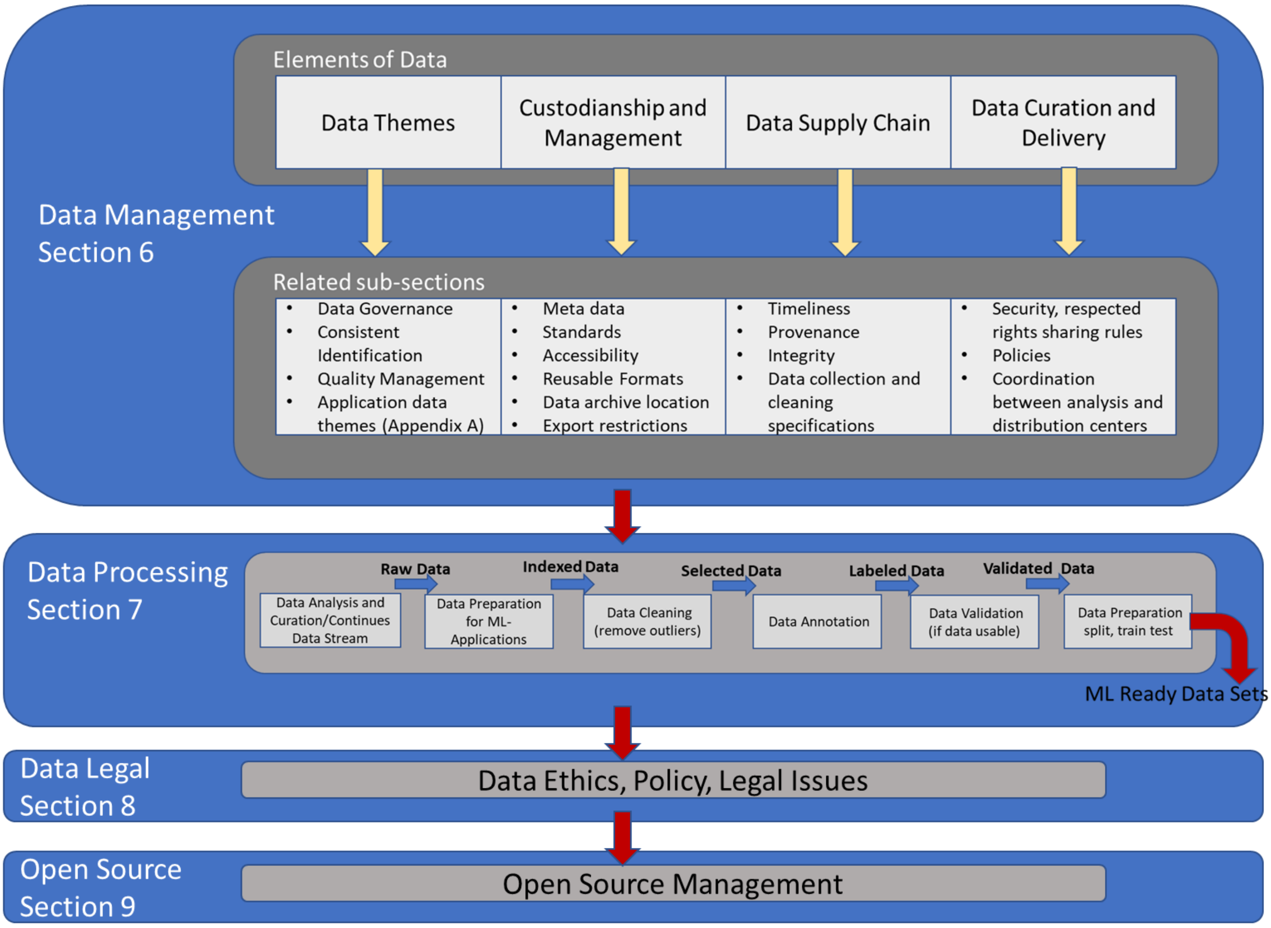

Elements of Data
Data Themes
Custodianship and Management
Data Supply Chain
Data Curation and Delivery
Data Management Section 6
Related sub-sections
• Data Governance
• Consistent Identification
• Quality Management
• Application data themes (Appendix A)
• Meta data
• Standards
• Accessibility
• Reusable Formats
• Data archive location
• Export restrictions
• Timeliness
• Provenance
• Integrity
• Data collection and cleaning specifications
• Security, respected rights sharing rules
• Policies
• Coordination between analysis and distribution centers
Data Processing Section 7
Raw Data
Indexed Data
Selected Data
Labeled Data
Validated Data
Data Analysis and Curation/Continues Data Stream
Data Preparation for ML-Applications
Data Cleaning (remove outliers)
Data Annotation
Data Validation (if data usable)
Data Preparation split, train test
ML Ready Data Sets
Data Legal Section 8
Data Ethics, Policy, Legal Issues
Open Source Section 9
Open Source Management

Figure 28 – Path for TGs

As shown in Figure 4, a key source of information used to derive best practices is the topic group use cases. These use cases were acquired through an open call for proposals that was put on the focus group website in advance of meetings A (16-17 March 2021), B (24-25 June 2021), C (31 August - 2 September 2021), D (20 October 2021), E (26-28 January 2022) and F (7-9 June 2022). To facilitate the systematic analysis of the use case proposals (for relevance, maturity, etc.), proponents were provided a template. Specifically, the proponents were requested to provide a project summary (a half page that describes the project and aspect being considered – data, modelling, or communications – for a given natural disaster type), a two-page project plan, a one-page outline of milestones and a one-page description of impacts. For the project summary, information about the research question and context, the method, the data and the evaluation were requested. These use case proposals were presented by the proponents at the respective focus group meeting. Following a discussion, the focus group decided whether to adopt the use case for inclusion in its activities. In total, 31 use cases were adopted. In a next step, the proponents of these 31 use cases were requested to complete a detailed questionnaire containing questions pertinent to the three working group technical reports (on "Data for AI," on "AI for Modeling," and on "AI for Effective Communications"). Out of the 31 use cases, 27 provided responses to these detailed questionnaires. An excerpt of these original responses can be found in the Annex.

Annex A

"Earthquake Disaster Mitigation through AI on Smart Seismic Networks"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Earthquake Monitoring, Detection, and Forecasting
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Earthquake Disaster Mitigation Through AI on Smart Seismic Networks
c. Please provide a short description of the use case.	A lot of details can be retrieved from the MSc. Thesis "A hybrid deep-learning approach for reliable real-time assessment of high magnitude earthquakes" by Viola Hauffe (University Magdeburg, Germany). This project tackles earthquake preparedness by developing artificial neural networks to be deployed on affordable smart seismic household sensors. The purpose of these is to (1) quickly identify if a signal is a seismic event or a different source of noise (2) analyse the vulnerability of a building within which the sensor was installed and (3) analyse a potential structural damage while and after a significant earthquake occurred.
d. Please provide a short description of the datasets.	Continuous time series recorded by publicly available seismic stations (hosted at https://geofon.gfz-potsdam.de) and seismic data acquired by QuakeSaver GmbH. The data sets are continuously recording 100 samples per second accelerometer data.
e. Please provide a short description of the model/method.	Deep convolutional neural networks trained on aforementioned continuous data to detect events, locate clustered events and pick first onsets of events recorded by the stations.
f. Please provide a short description of communications technologies that benefit or result from this use case.	The described technology allows to improve earthquake early warning in terms of speed and robustness against network failure due to the distributed computation (no single point of failure). Also in case of an event only relevant information from a large number of stations can be transmitted (time of first onset, maximum shaking intensity, damage reports) in a highly compressed data format.
2. Data-related questions	
a. Please provide information about data themes.	The training data carried information of seismic events recorded in Chile, Japan and Germany.

High-Level Questions	Responses
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	The majority of data is publicly available and open for further analysis via https://geofon.gfz-potsdam.de . A minority of data samples require clearance by QuakeSaver GmbH.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	Data can be retrieved using the FDSN standard from https://geofon.gfz-potsdam.de
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Waveform data and meta data were downloaded via FDSN from https://geofon.gfz-potsdam.de. The data center is coordinated by the German Research Centre for Geosciences GFZ in Potsdam. Data curation has not been carried out. Data was used in miniSEED format, the standard file format used in seismology. Meta data are provided in seed format. Event and seismic onset catalogs of this specific research have been provided after personal communication with Dr. Christian Sippl (sippl@ig.cas.cz).
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	Waveform, event and onset data were read using ObsPy. Waveforms were then restituted from their inherent measurement system (counts/s) to acceleration using the meta data mentioned above. The waveforms were high and lowpass filtered. Event catalogs were loaded together with their associated onset picks. Based on this information the waveforms were windowed and prepared to be fed into a deep neural network.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	Data is owned by GFZ Potsdam. Most of the waveforms are open access.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	Most data is open access at https://geofon.gfz-potsdam.de
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	The main problem in seismological applications of AI with respect to early warning is the bias of magnitude and frequency. Large mega thrust events are very rare as data but are the most interesting aspect.

"Probing Seismogenesis for Fault Slip and Earthquake Hazards"

High-Level Questions	Responses
1. General information about the use case	

High-Level Questions	Responses
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Earthquake Monitoring, Detection, and Forecasting
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Probing Seismogenesis for Fault Slip and Earthquake Hazards
c. Please provide a short description of the use case.	For active seismic fault systems, particularly when located near dense urban environments, predicting instantaneous and future characteristics of fault slip has long been a fundamental goal of geoscientists from an earthquake hazards perspective, but also to improve the basic understanding of fault mechanics. However, on natural faults the repeat cycles for all but the smallest earthquakes can span timescales on the order of decades to hundreds of years. Thus, in-situ geophysical measurements as input for data-driven ML models are generally not available or sufficiently complete for more than a portion of a single earthquake cycle. Transfer learning for AI models is the focus of this case use and may provide a tractable means of bringing the success of data-driven machine-learning approaches for predicting fault-slip characteristics in the laboratory to natural fault systems in the Earth.
d. Please provide a short description of the datasets.	Laboratory experiment data is routinely collected and a viable source of information to train models for application to nature fault systems. Numerical simulation data is available that matches the laboratory results and more simulations are needed to broaden the variance in the numerical results. With future application to faults in seismically active regions, obtaining sufficient training data is a challenge. In Earth systems data generally only exists for a portion of an interseismic slip cycle on a fault. Many data exist for continuous recording, but repeating seismic cycles at a single location, i.e., multiple large magnitude events within a decade, is not generally available. Transfer learning applications and cross-training techniques with the laboratory and numerical data are the solution to produce deep learning models of the necessary data to learn the seismic cycle. The trained model is applied to regional network seismic data.
e. Please provide a short description of the model/method.	The model combines data recorded in a laboratory setting to simulate earthquake rupture and numerical models to describe earthquake rupture. These data are combined in a convolutional encoder-decoder modeling framework to train the deep learning model with the numerical simulation data and then apply transfer learning with the laboratory data to fine tune the model. The final model is applied to new laboratory data to test if the evolving material properties are described directly from the input waveforms.

High-Level Questions	Responses
f. Please provide a short description of communications technologies that benefit or result from this use case.	NA
2. Data-related questions	
a. Please provide information about data themes.	Continuous seismic waveform data, laboratory acoustic emissions, and numerical simulations.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centres).	Laboratory and simulation data is published in refereed journals and available. Seismic waveform data utilized is freely available from regional networks operating in seismically active areas and collected as a community data product for earthquake hazard monitoring.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	Utilize open URLs
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Laboratory and simulation data is formatted at equal time resolution for application to the models. Seismic waveform data is formatted to standards described by the Incorporated Research Institutions for Seismology for standardized timing and quality with meta data included.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	Time series data is used as a continuous series into the models. Preprocessing typically includes normalizing for unit variance. Models are designed to ignore gaps and missing data that occur in continuous recorded data collected in the Earth.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	Data is openly available.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	Data is openly available.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	The main challenge is designing a data set and model that generalizes to all applications, which is not necessarily the primary goal if a location specific model is applicable.

A.1 TG-AI for flood monitoring and detection

"Flash Flooding Monitoring System in Mexico"

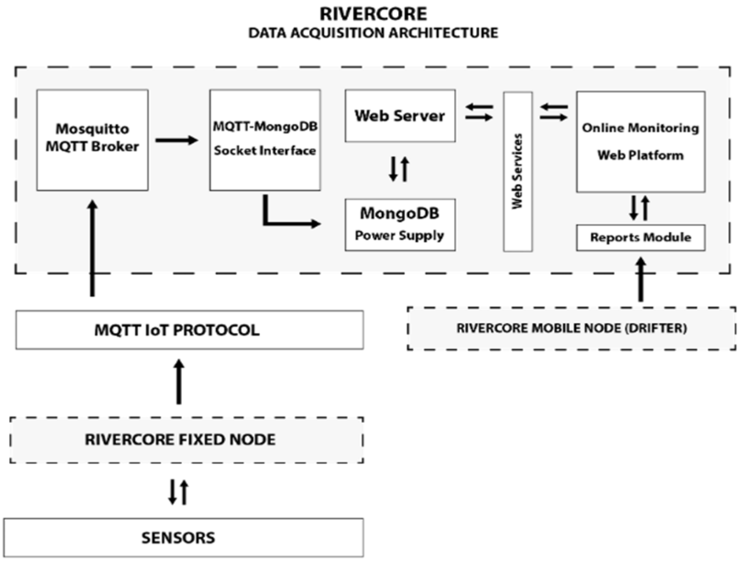


Figure A.1 – Data acquisition architecture

"Flash Flooding Monitoring System in Mexico"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Flood Monitoring and Detection
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Flash Flooding Monitoring System in Mexico
c. Please provide a short description of the use case.	The use case explores artificial intelligence to synthesize streams of instrumental (including sensor) data in real-time and detect features indicative of floods in Mexico. Using the EWIN - IoT network in Colima, we have three types of data (water level, weather station data, and soil moisture) to train machine learning models. The results of these machine learning models are compared with those of hydrological/hydraulic models, and performance metrics include root mean square error (RMSE). Such a system can be used to improve early warning systems. The study area is in Colima, Mexico, from 2018 to the present.
d. Please provide a short description of the datasets.	Our dataset includes the following information: device name, date, water level, soil moisture, standard depth, perimeter, hydraulic radius, area, velocity, and flow for 2019, 2020, and 2021. We have 3,286,062 records and the

High-Level Questions	Responses
	collection period starts on the 13 of June 2019 to the 26 of September 2021 with approximately 656 days.
e. Please provide a short description of the model/method.	In process.
f. Please provide a short description of communications technologies that benefit or result from this use case.	Technologies 3G or 4G and Lora are used for this use case. However, 5G and other wireless technologies such as Sigfox, Wi-Fi, or even Zigbee can be employed.
2. Data-related questions	
a. Please provide information about data themes.	Data themes are geographical position, water level, soil moisture, standard depth, perimeter, hydraulic radius, area, velocity, and flow.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	Water level, weather stations, and soil moisture represent data input acquisition, and their outputs are exported in Excel, SCV, and PDF formats.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	Sensors send data through 3G or 4G to Mosquito, Debian, JavaScript, and Database local servers. Data is stored in a database for future use, and real-time data is visualized in a dashboard.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Data is sent in JSON format and encrypted in MQTT; information such as Geographical position, water level, soil moisture, standard depth, perimeter, hydraulic radius, area, velocity, and Flow are stored in a database server. Different formats are allowed: Excel, SCV, and PDF.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	Data is collected by water level, weather stations, and soil moisture sensors and sent in JSON format, encrypted in MQTT until the mosquito MQTT broker server. Data is stored in a database server and indexed in Excel, SCV, and PDF formats.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	Data is open to everybody; we do not have personal information or data ownership issues.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	The information is open to everybody; this is open source and open data.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what	Our topic group has several flash flood monitoring and detection alternatives, such as satellites, drones, and IoT technology. One challenge is the theft of infrastructure in developing countries like Mexico. In our case, it is

High-Level Questions	Responses
recommendations can you offer to someone who intends to apply AI?	necessary to deploy a vast number of sensors, often located in places of difficult access or danger.

"Satellite Images and Machine Learning for Mapping Flood"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Flood Monitoring and Detection
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Satellite Images and Machine Learning for Mapping Flood
c. Please provide a short description of the use case.	In Mexico, different regions suffer from floods every year, affecting economic activities, human health, agriculture, livestock, among others. This makes it important to monitor water bodies and areas affected by floods to help reduce risks and make decisions in response to these disasters. Consequently, obtaining data that is very useful for mapping risk areas is very useful for agriculture, fishing, population settlement and different human activities. There are satellites that generate large amounts of data on the Earth and tools for processing large volumes of images and that are very useful for monitoring floods, detecting forest areas, crop areas and bodies of water, classification of land use, among others. Machine learning, particularly deep learning, has been used for the analysis of satellite images with satisfactory results, which has allowed the development of methods for land cover classification, flood detection, etc. In this research proposal, the mapping of the flooded areas and bodies of water is proposed, in the Los Ríos region of the state of Tabasco, made up of the municipalities of Balancán, Emiliano Zapata and Tenosique, in the period 2018-2022, through images. Sentinel-1 and Sentinel-2 satellites and deep learning algorithms. This, in order to collaborate in reducing the damage caused by floods and considerably reduce direct and indirect economic losses in municipalities vulnerable to this phenomenon.
d. Please provide a short description of the datasets.	SAR Sentinel-1 and Multispectral Sentinel-2 images will be used in this study. Images will be collected from the study area, from the municipalities of Balancán,

High-Level Questions	Responses
	Tenosique and Emiliano Zapata for the years 2018, 2019, 2020, 2021 and part of 202q. The Google Earth Engine platform will be used for this purpose.
e. Please provide a short description of the model/method.	The methodology proposed mapping flood using SAR and multispectral satellite images and deep learning consists of 5 stages: 1) input data, obtain datasets of images from the sentinel satellite, 2) Sentinel images selection: It is proposed to combine Sentinel-1 and Sentinel-2 images, 3) Images preprocessing: in order to obtain a collection of cleaner and sharper images, 4) Deep learning model, use convolutional neural networks (CNN) to analyse images, 5) Evaluate interpretability, interpret the data obtained with CNN and 6) finally classify the images to mapping flood areas
f. Please provide a short description of communications technologies that benefit or result from this use case.	Satellite technologies: Sentinel-1 and Sentinel 2 images. Machine learning algorithms (deep learning). Hardware for data processing and algorithm training, graphics processing unit, GPU, and TensorFlow
2. Data-related questions	
a. Please provide information about data themes.	They are the Sentinel-1 and Sentinel-2 images of the study area that will be obtained. Time series for the years 2018-2022 will be established. The images will be processed to remove noise and obtain cleaner and sharper data sets. Likewise, computation of spectral indices will be made between the digital levels stored in two or more spectral bands of the same satellite image.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	The satellite data (images) will be acquired on the Copernicus platform (Copernicus Open Access Hub) or from the Google Earth Engine platform. Images are raster data. Rasters are made up of an array of pixels, each with a value that represents the conditions of the area covered by that cell.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	There are different Sentinel image repositories. However, they will be obtained mainly from the Copernicus Open Access Hub repository and through Google Earth Engine. High-performance computing resources are required for image processing and training of the Deep Learning algorithm. To do this, the use of resources from a research center or an organization such as Huawei will be managed.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Annual time series are established considering the data from the National Water Commission (CONAGUA, a Mexican organization that provides data on rainfall) and the rainy seasons of the study area. Images will be processed with machine learning and results will be

High-Level Questions	Responses
	compared with GeoPDF files (maps stored in PDF files that have georeferenced properties)
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	All images will be pre-processed to obtain a clean and sharp collection of images and will use a Convolutional Neural Network (CNN, commonly consisting of input, convolutional, pooling, connection and output layers) architecture with Sentinel-1 images and Sentinel-2. All networks will be trained and tested with TensorFlow and GPU. The proposed algorithm will be implemented in Python and the use of Tensorflow software.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	The data sets to be created will be open data so that they can be used for subsequent research and serve as a reference for future work related to floods.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	The languages, tools, applications, libraries, operating system and other technologies are open source, so the products and datasets to be created will be totally free of copyright and license.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	Pending

"Using ML to Reconstruct Flooded Area under Clouds in Optical Satellite Images: The Mozambique Use Case"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Flood Monitoring and Detection
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Using ML to Reconstruct Flooded Area under Clouds in Optical Satellite Images: the Mozambique Use Case
c. Please provide a short description of the use case.	The Machine Learning algorithm developed by the RSS-Hydro team requires as inputs a cloud-covered low-resolution optical (e.g. Sentinel-2) satellite flood image and auxiliary data, both during the training and the inference phase. During training, the model additionally requires a ground-truth flood map. Auxiliary data like, for

High-Level Questions	Responses
	example, digital elevation model and derived datasets such as slope and topographic wetness, help the FloodSENS algorithm learn the correlation between flooded areas and their surrounding topography.
d. Please provide a short description of the datasets.	Within this study we categorize feature data into two different types: • Static data, such as the Copernicus DEM, has been acquired or generated for a particular point in time, generally before a given flood event. • Continuous data, such as Sentinel-2 images, generally exist in the form of time series, and have a cycle that covers pre- and post-event dynamics. Technically these data sources come with specific properties concerning the flood mapping. Considering as an example the properties of a static DEM for the mapping of a dynamic event, which are not reflected in the dataset, this DEM still offers indirectly fluvial forms that can serve as proxy, even if acquired totally independently of such event. On the other hand, a Sentinel-2 time series might suffer from impenetrable cloud cover after flood events, rendering the data obsolete even if available. Two types of input data are required for training and deployment; optical data and static auxiliary data.
e. Please provide a short description of the model/method.	It is important to note that at this stage in the project, all the pre-processing part as well as the data for the training and references is completed. We are now at the stage where we train the model architecture on different use cases and test it for generalization. The ML algorithm will go through two separate phases namely training and inference. Training an effective algorithm is the main challenge and the next three sub chapters are focusing on training related aspects of the project. In a first instance a static trained algorithm will be deployed on WASDI, meaning once deployed the weights are frozen and will not be changing. A major source of information lies in the propagation auxiliary data. Tiling them could be enough for good results since the auxiliary dataset of the flow accumulation numbers is in itself a propagation of information from other tiles (in the same hydrological basin). Our goal is to grow our model. This means, we have a live model, that is continuously fine-tuned on a growing database of cases and study sites, and which will improve iteratively its transferability,
f. Please provide a short description of communications technologies that benefit or result from this use case.	NA

High-Level Questions	Responses
2. Data-related questions	
a. Please provide information about data themes.	Any data chosen for the development of FloodSENS needs to fulfill a few criteria to successfully train and deploy an algorithm that responds to customer needs.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	Input data needs to be available globally, for free and shortly after flood events. Reference data needs to be of the highest possible quality to avoid introducing errors to the ML algorithm.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	Freely available data from ESA and NASA
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Static data, such as the Copernicus DEM, has been acquired or generated for a particular point in time, generally before a given flood event. Continuous data, such as Sentinel-2 images, generally exist in the form of time series, and have a cycle that covers pre- and post-event dynamics.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	All data are pre-processed to the same specifications, including spatial resolution and tile size
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	The following third-party products/rights are planned to be used in this product development: Python, Sentinel-2 or Landsat optical imagery and Sentinel-1 SAR; new Copernicus DEM; Microsoft Azure OR/AND WASDI Development Platform; high-resolution flood imagery.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	In particular for the Python programming software: all Open-Source components/libraries and distribution of created products is free of copyright and licensing. Satellite imagery & DEM: all free and open, including commercial use.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	Pending

"Exploring Deep Learning Capabilities for Surge Predictions in Coastal Areas"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Flood Monitoring and Detection
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Exploring Deep Learning Capabilities for Surge Predictions in Coastal Areas
c. Please provide a short description of the use case.	This use case applies tide station data (from GESLA-2) and atmospheric conditions (from ERA-5) to train four types of deep learning models (artificial neural networks, convolutional neural networks, long short-term memory layer, and a combination of the latter two) to predict hourly storm surge ensembles at a global scale. The models are assessed using minimum absolute error as the selected loss function as well as Continuous Ranked Probability Score for the ensemble of models.
d. Please provide a short description of the datasets.	For the predictand variable, we used observed sea levels from the Global Extreme Sea-Level Analysis Version 2 database (GESLA-2). We selected stations with a high temporal frequency (15 min to one hour) which resulted in 736 stations spread globally. This dataset is already controlled for potential errors and has been used in many coastal studies. We extracted the storm surge from the total sea levels by detrending sea levels and subsequently applying a harmonic analysis. For the predictor variables, we extracted the selected atmospheric variables (mean sea level pressure, meridional, zonal wind at 10 m) from the most recent ECMWF high resolution climate reanalysis dataset, ERA-517. This global dataset has a spatial resolution of 0.25° and an hourly temporal resolution. While it is documented to have some biases, its increased temporal and spatial resolution resulted in considerable improvements in performance over its predecessor ERA-Interim.
e. Please provide a short description of the model/method.	In our study, we compared four neural network (NN) models. The input layer is connected to the following hidden layer: • ANN A fully connected layer with an l2 kernel regularizer. • LSTM a stateless LSTM layer with a hard sigmoid recurrent activation function. • CNN a 2D convolution layer. Each filter has a kernel size of 3 × 3 with the same padding and the convolution step is followed by a max-pooling layer with a kernel size of 2 × 2.

High-Level Questions	Responses
	- ConvLSTM a 2D convolution layer following a stateless LSTM layer with a hard sigmoid recurrent activation function. Each filter has a kernel size of 3 × 3 with the same padding and the convolution step is followed by a max-pooling layer with a kernel size of 2 × 2. All of the NN models are activated using the ReLu activation function as is common in NNs. In the cases of the LSTM and ConvLSTM, a hard sigmoid function is used for the recurrent activation. The last hidden layer is a fully connected layer with an l2 weight regularizer and a dropout is added. We select the Adam optimizer algorithm for the learning rate optimization algorithm and train the NN model to minimize the mean absolute error, the selected loss function, between observed and predicted surge. The output layer, with one node only, represents the predicted surge levels.
f. Please provide a short description of communications technologies that benefit or result from this use case.	Forecasting systems, critical infrastructure
2. Data-related questions	
a. Please provide information about data themes.	"Elevation and Depth": storm surge "Water": predicting sea storm surge "Addresses": location of the tide gauge stations Not added in the minimum list if global fundamental geospatial data themes but this case study used climate data as input variables
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	This study uses openly available climate data from: ECMWF (European Centre for Medium-Range Weather Forecasts) for the input variables with their own data handling system (LINK). GESLA-2 dataset: not clearly stated from their dataset The produced storm surge time series from the models and the models are stored on Zenodo, an open repository maintained by CERN. Zenodo is compliant with the data management requirements of Horizon Europe, the ERC and other EU research and innovation funding programmes
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	More information can be found on the respective websites: For Zenodo: https://about.zenodo.org/infrastructure/ For ECMWF: https://www.ecmwf.int/en/computing/our-facilities/supercomputer-facility For the GESLA-2 dataset – this information is not available

High-Level Questions	Responses
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Specific data curation has already been performed by the ECMWF, and the GESLA-2 dataset as described on their respective platforms: ECMWF: Netcdf or grib files to be downloaded from the CDS data store. GESLA-2 dataset: text files with similar header information (see link) Produced time series and models from the case studies can be downloaded from Zenodo, where data is stored using FAIR principles (Findable – Accessible – Interoperable – Reusable) (see link)
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	Data collection of climate variables (predictor variables) and total sea levels (predictand variables). Both datasets have already been cleaned. Detrending climate data. Extraction of the storm surge by removing the tide and detrending. Storing of predictor and predictand variables as netcdf files.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	No ethical considerations needed for this study. Legal considerations refer to the license of the data used: ECMWF data (ERA5 dataset): Licence to use Copernicus Products GESLA-2 data: free for research (license) Time series and model produced in this case study: Creative Commons Attribution 4.0 International
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	All datasets used in this case study are open source and open data for research purposes, including all the outputs from this case study. See previous point for specific license
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	The main challenge is not to "misuse" the model and apply it for purposes outside of its original design/application. While some models can be modified for other applications, it is often difficult to do so from unforeseen logistical applications (for example, data is not updated frequently enough, etc).

A.2 TG-AI for geodetic enhancements to tsunami monitoring and detection

"Deep Learning Detection of Elasto-Gravity Signals for Earthquake and Tsunami Early Warning"

For this use case, no completed questionnaire was received by the submission deadline. Therefore, the details of this use case have been omitted during the derivation of best practices in this technical report.

"Enabling Natural Hazards Risk Information Sharing Using Derived Products of Export-Restricted Real-Time GNSS Data for Detection of Ionospheric Total Electron Disturbances"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Geodetic Enhancements to Tsunami Monitoring and Detection
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Enabling Natural Hazards Risk Information Sharing Using Derived Products of Export-Restricted Real-Time GNSS Data for Detection of Ionospheric Total Electron Disturbances
c. Please provide a short description of the use case.	Tsunamis can trigger internal gravity waves (IGWs) that propagate to the ionosphere, causing a perturbation in the natural Total Electron Content (TEC). These perturbations are often referred to as Traveling Ionospheric Disturbances (TIDs) and are detectable through the Global Navigation Satellite System (GNSS) signals. In this interdisciplinary work, we describe a framework for leveraging slant total electron content (sTEC) produced by the VARION (Variometric Approach for Real-Time Ionosphere Observation) algorithm and Convolutional Neural Networks (CNNs) in a process which trains a generalized model for TID detection, applicable across various atmospheric conditions and geographic areas.
d. Please provide a short description of the datasets.	Slant total electron content (sTEC) time-series data produced by the VARION (Variometric Approach for Real-Time Ionosphere Observation) algorithm was used for initial trials. Future versions of this work will leverage data from the GUARDIAN system.
e. Please provide a short description of the model/method.	Time-series sTEC data is transformed into images using an approach called Gramian Angular Difference Fields (GADFs). These images are subsequently used to train a Convolutional Neural Network (CNN), a type of deep learning network that leverages computer vision techniques. This combined methodology of using GADFs together with a CNN results in an approach that's robust to missing data.
f. Please provide a short description of communications technologies that benefit or result from this use case.	N/A
2. Data-related questions	

High-Level Questions	Responses
a. Please provide information about data themes.	The sTEC data is produced by VARION from raw TEC data from pairs of satellites and ground stations, and is itself a multivariate time series.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	With respect to the deep-learning based component of the overall system, a data governance system will be adopted in a production setting that verifies the integrity of any new model training data produced by GNSS satellites and labels provided by subject matter experts (SMEs). In a production setting (with further development of the system), TID detections made by the system will be labeled as true, false, or be adjusted by SMEs for subsequent continued retraining of the deep learning model. This ensures that the model remains up to date over time and improves over time with additional information provided by SMEs. Detections made by the system will be accessible by a REST API in the future, providing a means to access the information programmatically.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	The data originates from JPL's GDGPS (GUARDIAN) system and is high-quality, regular and reliable. The system provides near real-time TEC data for various GNSS systems.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Training data for machine and deep learning (Artificial Intelligence, AI) models require labeled training data in the case of supervised learning. Our deep learning based approach is a supervised learning approach, leveraging SME-labeled sTEC data indicating which perturbations and time periods correspond to tsunami-generated TIDs. Our existing methodology has been prototyped on a dataset with 4 events, 3 used for training and testing and one for out of sample validation. Future versions of these systems should consider Human-in-the-Loop processes whereby SMEs regularly contribute feedback to the system's classifications, resulting in an improved system with regular re-training.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	Model Training: Data collection by GUARDIAN, GUARDIAN produced TECs, TEC data produced, historical data labeled, data cleaned and transformed into images for model training, model trained, model validated in simulation of real-time scenario metrics produced. In Production: Data collection by GUARDIAN, GUARDIAN produced TECs, TEC data produced, data converted to image, image fed into trained model, predictions produced positive classifications, result in alert.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	No concerns about personally identifiable information (PII). TEC and sTEC produced and owned by the Jet Propulsion Laboratory (JPL). Trained models do not

High-Level Questions	Responses
	contain data from GUARDIAN, and possibly (with proper approvals) could be shared internationally.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	Authors and developers have the desire to open source 1) data used in model training 2) models trained and 3) any code used to produce those trained models, in the spirit of transparency and reproducibility. Production-level, in-use trained systems using data at scale will not be shared (only experimental validations, models).
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	Continued data curation and educating scientists on the importance of providing funding, support and continued labeling of data to ensure the effective use of AI systems.

"Building a Coupled Earthquake-Tsunami-TEC Simulator in a Parallel HPC Environment"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Geodetic Enhancements to Tsunami Monitoring and Detection
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Building a Coupled Earthquake-Tsunami-TEC Simulator in a Parallel HPC Environment
c. Please provide a short description of the use case.	The project here represents advancements made towards the creation of a neural network-based tsunami warning system which can produce fast inundation forecasts with high accuracy. This was done by first improving the waveform resolution and accuracy of Tsunami Squares, an efficient cellular automata approach to wave simulation. It was then used to create a database of precomputed tsunamis in the event of a magnitude 9+ rupture of the Cascadia Subduction Zone. Our approach utilized a convolutional neural network which took wave height data from buoys as input and proved successful as maps of maximum inundation could be predicted for the town of Seaside, OR with a median error of ~0.5 m. Other hypothetical configurations of buoys were tested and compared to determine the lowest number of buoys necessary in order to make such a prediction.

High-Level Questions	Responses
d. Please provide a short description of the datasets.	For this project, three datasets were created via simulation. These include a dataset of 3000 earthquakes, 3000 tsunamis, and 3000 inundation maps. The earthquakes range in magnitude from 8.9 to 9.4. The tsunami simulations were used to generate time series wave height data from buoys and acted as the input for the neural network. The inundation maps acted as the output for the neural network.
e. Please provide a short description of the model/method.	A convolutional neural network was utilized to predict inundation maps by analyzing off-shore wave height data collected by buoys. Datasets for training and testing data were simulated. In addition to existing buoys, various hypothetical configurations of buoys were tested to determine the most optimal amount and placement of said buoys. This was done using a sensitivity test to determine which buoys were prioritized more by the neural network.
f. Please provide a short description of communications technologies that benefit or result from this use case.	N/A
2. Data-related questions	
a. Please provide information about data themes.	Data was simulated using Tsunami Squares.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	For any real-time scenario, access to off-shore buoys for the US West Coast would have to be obtained and fed into the network in order to make an inundation prediction.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	Real-time data would originate from the National Oceanic and Atmospheric Administration's (NOAA's) National Data Buoy Center.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Data for the training and test set would be simulated by Tsunami Squares since it is fast compared to typical finite difference methods. Simulations are stored in NetCDF4 format.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	First, two simulation regions are selected: a large area encompassing most of the coast, and a smaller one for the specific city used to simulate inundation. Stochastically generated earthquakes are then simulated off the coast in order to build a database of seafloor uplift values corresponding to each scenario. Then, using the seafloor uplift as an initial condition, the resulting tsunami is simulated over the large area defined previously. The wave is simulated towards the coast where the smaller area wave simulation takes over and propagates the wave the remaining distance to the coast, producing an

High-Level Questions	Responses
	inundation map. The inundation map is converted to CSV format and cut down to reduce the total number of variables. For instance, cutting out areas where the elevation is particularly high, where no water will reach. Then, the large simulation is used to produce the simulated buoy data. This is also stored in CSV format.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	N/A
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	We have the desire to be open source.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	The most important part of any project involving AI is having a sound database to train your model with.

A.3 TG-AI for insect plague monitoring and detection

"Identification and Classification of Pest Infested Coniferous Forest Using AI"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Insect Plague Monitoring and Detection
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Identification and Classification of Pest Infested Coniferous Forest Using AI
c. Please provide a short description of the use case.	In this use case, we aim at developing a system that uses Deep Learning and UAV-acquired forest images that can identify individual tree health conditions (defoliation rate) in areas of hundreds or thousands of hectares to comprehensively evaluate the health of diverse forest ecosystems.
d. Please provide a short description of the datasets.	The data of the tree health were divided into training and testing datasets for DL classification

High-Level Questions	Responses
e. Please provide a short description of the model/method.	The use case uses Deep Neural Network to automatically identify different categories of tree healths including 1. healthy, no defoliation; 2. Very low. < 10 % defoliation; 3. Low, 10–25 % defoliation; 4. Medium, 26–50 % defoliation 5. High, 51–75 % defoliation and 6. Very High (Dead), > 75 % defoliation.
f. Please provide a short description of communications technologies that benefit or result from this use case.	This project presents the development of an automatic tree health classification method based on UAV-acquired very high-resolution images for training of a deep learning model that is unprecedented in terms of practical application and generalization potential.
2. Data-related questions	
a. Please provide information about data themes.	The data theme was in mountainous forest where the coniferous trees were infested by bark beetles
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	The data were collected by UAV, 70m height to the take off point, 90% side and front overlaps. The orthomosaics, DSM, CHM exported from tiff to jpg to annotate. The annotated data were stored according to sites.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	Not applicable
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Data was transferred locally by hard drives and clouds. Coordination was done verbally or by email.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	Data collection was done by drone. The data were standardized in GIS software. Subsequently, parts of the data were extracted.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	The data were owned by PI Larry Lopez and the team
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	Open source software: QGIS, GIMP, Fusion LDV Licensed software: ESRI ArcGIS, Agisoft Metashape and Global Mapper
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when	Require numerous data for each degree of infestation which is not easy to get in the case of natural hazard.

High-Level Questions	Responses
using AI for this application and what recommendations can you offer to someone who intends to apply AI?	

"Artificial Intelligence Modeling Tools for Monitoring Desert Locust (AI-Locust): Breeding Grounds, Hatching Time, Population and Spatio-temporal Distribution"

proposed by Elfatih Mohamed Abdel-Rahman (1), Emily Kimathi (1), and Henri E. Z. Tonnang (1)

(1) International Centre of Insect Physiology and Ecology, Kenya

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Insect Plague Monitoring and Detection
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Artificial Intelligence Modeling tools for Monitoring Desert Locust (AI-Locust): Breeding Grounds, Hatching Time, Population Dynamics and Spatio-temporal Distribution
c. Please provide a short description of the use case.	The use case aims to develop an early warning and decision support system for monitoring desert locusts for sustainably managing its impact in Eastern Africa and Sahel-Maghreb regions. The use case will build an innovative platform essentially based on the use of cross-cutting artificial intelligence (AI) tools and algorithms (e.g., Artificial neuro Fuzzy) and means of near-real-time and long-term (> 30 years) Earth observation tools viz., satellite-based systems. We will use readily available climate, soil, and vegetation datasets, and AI-analytics to forecast desert locust outbreaks. The use case will utilize long-term desert locust observations that are readily available from the DLIS-FAO hub and other sources. Specifically, the use case will predict locust breeding grounds, hatching time, spatial distribution, and forecast its outbreaks. We will roll out the AI-model outputs to assess the site-specific risk of locust breeding and predict future migratory patterns and intensity of desert locusts; improve the locust monitoring system; determine the economic, food security, health, and environmental burden of the locust invasion. We will also study the impact of climate change on locust resurgence.
d. Please provide a short description of the datasets.	The use case will combine datasets from various sources for AI-analytics. Specifically, we will use long-term (> 30 years) satellite-based monthly rainfall, temperature, wind speed, and vegetation variables; and edaphic factors to predict and forecast desert locust breeding sites and outbreaks. The rainfall and temperature datasets are freely

High-Level Questions	Responses
	available from Envidat (https://www.envidat.ch/#/metadata/chelsa_cmip5_ts). The Envidat provides mean monthly maximum and minimum temperatures, as well as monthly precipitation at ~5 km spatial resolution globally for the years 1850-2100. While the wind speed will be obtained from the worldclim database (https://www.worldclim.org/data/worldclim21.html) and the edaphic factors include soil moisture (1985 – 2021) and sand content at 0 – 20 cm depth at 4 km spatial resolution from Terraclimate (https://climate.northwestknowledge.net/TERRACLIMATE/index_directDownloads.php). All these variables will be pre-processed and harmonized at 5 x 5 km resolution. The desert locust observations (adult and nymph occurrence data) are available from the DLIS-FAO data hub (https://locust-hub-hqfao.hub.arcgis.com/). This dataset compiles ground survey observations spanning 36 years, from 1985 to 2021, covering ~ 29 million km^2. We will use records for both desert locust nymphs and adult occurrence for 36 years (1985 and 2021). The desert locust data will be explored using open data science approaches and procedures. A grid of different sizes (5 x 5, 10 x 10, …, 50 x 50 km) will be applied to the entire study area which covers the desert locust occurrence observation points. Data sets within the grid that provide the most spatio-temporal desert locust observations over the 36 years will be used for calibrating the AI-modeling experiment. Socio-economic and other variables will be sourced from individual countries' databases.
e. Please provide a short description of the model/method.	The proposed use case will employ different machine learning (ML) and AI analytics to predict desert locust breeding grounds and forecast its outbreak. Specifically, we will use the maximum entropy (MaxEnt) approach to assess the suitable habitats for desert locust breeding grounds. The MaxEnt model is a machine learning model that uses the entropy approach to predict species distribution. The MaxEnt model outputs (desert locust suitability maps) together with the climate, soil, and vegetation variables to be utilized to develop the AI-based model (AI-Locust). We will use the artificial neuro-fuzzy algorithm for developing the AI-model. Among multiple hybrid modeling approach, the evolutionary adaptive-Network-based Fuzzy Inference System (GA-ANFIS) that integrates the benefit of Fuzzy logic, Neural network (NN) and Genetic algorithm (GA) appears to be the most promising due to its high degree of diagnostic accuracy, which is justified by its application in various fields. This technique will be widely used in our use case.

High-Level Questions	Responses
f. Please provide a short description of communications technologies that benefit or result from this use case.	Our main communication tools will be scientific, publications, policy briefs, reports, interviews etc. We further plan to use mobile and digital technology to disseminate our findings.
2. Data-related questions	
a. Please provide information about data themes.	This use case will build tools that enable the integration of historical and ongoing collections of desert locust. These data will include documents and scripts (plain text (.txt), MS word, MS Excel, Open documents, Rich Text format, HTML, PDF etc.); geospatial data (vector and raster) ESRI Shapefile (.shp, .shx, .dbf;), ESRI Geodatabase format (.mdb), MapInfo Interchange Format (.mif) for vector data, Geo-referenced TIFF (.tif, .tfw), CAD data (.dwg), global positioning system (GPS) location files (.gpx), Keyhole Mark-up Language (KML) (.kml). We will also add digital image and video data in TIFF, JPEG and Adobe Portable document format, MPEG-4 High Profile and JEPEG 2000. Quantitative tabular data with extensive metadata (dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data) and quantitative tabular data with minimal metadata (matrix of data with or without column headings or variable names or labeling) will also be used.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	The use case data are based on open data and citizen science principles. All data will be acquired and downscaled from freely available sources. Also, the data will be processed and analysed using open data processing tools such as Google Earth Engine and python language. The data will be managed through the icipe's Research Data Management and Archiving (RDMA) policy, which aims to make the data Findable, Accessible, Interoperable and Reusable (FAIR). Moreover, the use case data will be managed using the following data engineering and pipelines: • Common Ontology - http://dmmg-co.icipe.org • Data Management Plan - http://dmmg-dmp.icipe.org • Data Warehouse (CKAN) - http://dmmg.icipe.org/dataportal • Version Control (GitHub) - https://github.com/icipe-official
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	We will ensure that the data and related materials, both digital and non-digital, must be accompanied by proper metadata and documentation in a way that facilitates the verification, replication and, if possible, reuse of the data. We will generate an integrated resourceful desert locust database for wider public use under the Creative Commons copyright licenses. The data will offer the

High-Level Questions	Responses
	public a pool of data that can be copied, distributed, edited, remixed and built upon.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	This use case will provide the foundation for the team to strengthen data governance, data warehousing, data architecture and repository, data integration, data classification, data quality management, data security, data standards and corporate and regulatory compliance. This will further guide effective decision making and taking actions that will maximize benefit to the proposed use case especially in the delivery of the outputs. Improving its capacity in data management should position the team to use evidence-based and data-driven insights that will inform local, national, regional, and global policies in desert locust management and advisory services to improve food security in affected regions. By strengthening these areas, the team involved in the use case will improve their skills and knowledge in data management to effectively deliver the use case outputs.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatio temporal requirements; and structure of feature vectors).	N/A
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	The use case does not involve collation and analysis and publishing of human data and in all cases; we will follow the international standards and will abide by icipe data management and sharing policy which uses FAIR and open data principles. Likewise, the future desert locust predictions and forecasts using AI and ML algorithms do not involve personal information and data privacy. Moreover, the project shall strive to be gender-responsive during implementation by equally involving men, women, and youth in the implementation.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	This use case is in line with icipe's research data management and archiving policies, which is requested by researchers, donors, partners, and publishers to provide the appropriate information infrastructure (policies, guidelines, ICT connectivity) to derive the most value from the research activities. Therefore, we will implement this use case through a strategy that is designed around the needs, resources, and structures of using 'FAIR' guiding principles for scientific data management and stewardship. This use case will completely adhere to open science and open access requirements. Proper data management workflow is the primary procedure that will enable us to ensure the accuracy, completeness, and integrity of all information that will be collated, analysed, and published.

High-Level Questions	Responses
	The results and products of this project, namely the geo-referenced desert locust outbreaks and AI-analytics will be available under an open source by-attribution license (CC-BY 4.0). All data will be deposited in previously mentioned icipe's software platforms. The dissemination of the use case outputs, and knowledge will be conducted, through publications, reports, emails, social media, and internet links. Overall, the AI-Locust platform itself will serve as means of communication and networking. We will ensure that the data and outputs will be accompanied by proper metadata and documentation in a way that facilitates the verification, replication and, if possible, reuse of the data.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	The main challenge is having access to good quality data. Today, many AI tools and algorithms exist and are very powerful to extract knowledge from data and produce raisable outputs which can help in decision making and then transform the society. We should establish and nurture solid teamwork, disseminate best practices in data management which are aligned with the FAIR and open data principles and promote policies that are favorable to the use and application of advanced analytics for knowledge discovery.

A.4 TG-AI for landslide monitoring and detection

"Landslides of Masses of Soil and Rock: Intelligent Risk Management in Areas Highly Threatened by Climate Change"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Landslide Monitoring and Detection
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Landslides of Masses of Soil and Rock: Intelligent Risk Management in Areas Highly Threatened by Climate Change
c. Please provide a short description of the use case.	To handle the complex dynamics of the factors involved – with temporal and spatial dependence – Data Science (factorial analysis, fuzzy clustering, and CART) and Artificial Intelligence (Neural Networks) are used to study landslides events (as cause-effects) from geology, geomorphology, geotechnics, and climate data (the threat is rainfall -extreme-). The neural model shows remarkable capacities to spatially quantify the impact of geomorphological, anthropic, and hydric variables on

High-Level Questions	Responses
	mass removal processes. Mud and debris flows, as well as other destructive processes in mountainous areas are associated with the existence of rural developments and civil infrastructure to define integral risk scenarios and to measure the impact of deforestation (and other harmful human activities) on natural environment stability. Based on the results, vulnerability and exposure maps are constructed (at useful scales) for the poorest southern states of Mexico, but the methodology is general and can be extrapolated to other world regions.
d. Please provide a short description of the datasets.	According to the universe of descriptors, this research is based on information from government offices, academic/research institutions and civil organizations linked to the NDM. The main source of data is the CENAPRED (National Center for Disaster Prevention), an institution that compiles information from the army, navy, civil protection offices and the national university (UNAM) in questionnaires that describe the process in an organized way (footprint, approximate volume slide, materials on the foot, date, etc.). The CENAPRED is also in charge of reviewing and publishing geological, geotechnical and relief maps, among others. To categorize the threat (rain) we have agreements with the Mexican government to open the information from the hydrometeorological stations in the studied areas.
e. Please provide a short description of the model/method.	Once the area (poorest southern regions in Mexico) and the events (hurricane and cyclone season in the Mexican Pacific) are descriptive, the information is analysed with Data Science to define the best representation of the variables. At this stage a CART for getting the most efficient training set for the intelligent models, for example, examples of geographical situations that slide compared with the number of situations that do not, force the modeler to define the best proportion of the YES/NO occurrence (slide) for the NN. Also, the CART is used for integrating boundaries or limits of application. Then a neural network (multilayer feed-forward, quick propagation, supervised learning) is trained to predict i) if a "patch", or a group of them, slides (a patch is the best spatial unit to characterize the environment and to measure the effects of the hydrometeorological phenomenon), ii) to characterize the inputs effects and iii) to define the dependence between the rainfall and the event. These patches are conceptualized as 3D (voxels) and are communicated in 2D (pixels, maps) where each unit is filled with information of the exposition and susceptibility.
f. Please provide a short description of communications technologies that benefit or result from this use case.	The first communication is through high-resolution static hazard maps (that could be migrated to dynamic ones). The alert system of the Mexican government is benefited

High-Level Questions	Responses
	with the model outputs because it informs when the rainfall is approaching high levels, so the risk of sliding in susceptible areas will also be high and the specialized team must be mobilized. The disaster manager receives alarm messages to different recipients, and it should use different communication mediums. The model gradually qualifies the warning messages, being the most important ones sent directly to the targeted populations. Because the studied areas are poor regions, the communication follows the restrictions of infrastructure and security.
2. Data-related questions	
a. Please provide information about data themes.	Geo-data: geology (materials, homogeneity, geometry), geomorphology (shape, topography, relief), geotechnics (material, stratigraphy -general-, angle of repose -equilibrium-), and climate data (rainfall, temperature, humidity). Slide-data: footprint, volume -approximate-, step (slope-high ratio), materials on the foot, object description (shoulder, foot, material on the foot, material on the shoulder, dynamics -speed-, signals -for anticipation-) and date. Anthropic data: deforestation, roads construction, alterations to water currents, chaotic development of communities (existence of non-functional drainage networks).
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	The CENAPRED information questionnaires contain a large percentage of the information on Geo-data and Landslide-data. In addition, government sources are consulted on parameters, at the regional level, which are labeled on the topographic maps of the INEGI (Statistics and Geography Institute of Mexico), as some of the Anthropic-data. The data is fully accessible, however most of the map parameters must be manipulated to transform them from bits (in a colorimetric scale) to numbers. The concentration of data in matrices that can be studied with AI is designed in such a way that it can be exported to GIM and BIM (geo information and building information modeling), that, at a certain moment, could serve as the basis for the development of digital twins of the area. The data and results concentrated as maps are exportable in all formats used in geo-frameworks. The instructions that are generated from CART and NN must be translated for exploitation with and for sources other than the modeler. The models, datasets and results in general are disponible through petition to the UNAM, who is responsible for evaluating the posterior use.

High-Level Questions	Responses
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	The origin and quality of data is sustained by the CENAPRED, UNAM and INEGI institutions, all of them recognized as prestigious. The integrity of databases is double-checked by the engineers' team through the application of some Exploratory Data Analysis (EDA) tools. At UNAM, particularly at its Instituto de Ingeniería, the datasets and the analyses. The university has enough infrastructure to analyse the data and to develop the analysis and modeling. The developers of this project exploit data visualization in each of its stages: information pre-processing (interpret databases), model training (clarify learning process) and presentation of results (display the efficiency and the adjustment to the task).
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	For input-data sharing there are no restrictions about confidentiality, privacy, or intellectual property rights, but for results-data the custodians of the maps and GIM sets are the university authorities and the researchers involved in the investigation (particularly the Technical Responsible). Also, the expressed recommendations to the government authorities are not available to share without special permissions. On the other hand, two methods/systems are used in this research that are patented by the Mexican government in favor of the UNAM (algorithms to massively parameterize scarcely instrumented natural masses), so the data that comes from their application may not be shareable.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	The first step, the data collection, is an effort between the CENAPRED and UNAM in which the questionnaires are being migrated to a data mart created not only for this research but to become a versatile and exploitable repository by other engineering professionals. This part of the compilation is essential for the prediction since it represents the presentation of the SI examples, that is, of the cases or examples that have slipped (under what conditions and how the mass-removal dynamics developed). Its identification is necessary as a first step since the analysis grid size (or patch grid) is proposed from the footprint and slip volume values. Subsequently, the map information on the sizes of the patches is analysed to "fill" them with the relevant information. This is conceived as the matrix of cases where each row means a georeferenced patch (the geo-coordinates is, as a couple, the ID for each patch, a group of patches are identified as subzones and a group of zones as areas), and the columns are the descriptor parameters of that geo-space. To achieve this, affordable routines have been created to transform, according to solid criteria (the result of applying CART and

High-Level Questions	Responses
	clustering), the colored pixels to suitable numerical values. Then for preparing data some tools from Exploratory Data Analysis EDA are applied to uncover patterns, to notice anomalies, to test hypotheses and to check assumptions. For the spatial-temporal nature of this phenomenon, it is particularly important to detect the more suitable variable transformations, the missing value treatment, and the dimensionality reduction.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	The input data is open and there are no conflicts of confidentiality. However, it is important to highlight that in some cases, when the Mexican government sponsors research to generate parameter maps, the publication or sharing of information is subject to legal requirements regarding waiting times, discretion in interpretation and communication or, in the most complex of cases, if the value of the land suffers some decrease due to the publication of risk maps, the government has the power to prevent them from being opened to the public and for the Secretariat of Civil Protection to verify situations in response to demand expression of the citizens.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	It must be declared that some of the research items (methods that solve problems in input-output systems) are susceptible to patenting and are closed until the legal process is completed, so the data that resulted from them are not open. However, the general methodology, the basic data, and open maps can be shared and analysed with other researchers and organizations that require them, upon an official request to UNAM.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	One of the greatest challenges to study landslides is the survey of the events. The supervised training of neural networks with "real" cases allows us to discover relationships between very important parameters that conventional models (those that calculate the susceptibility to landslide based on topography, relief, and surface geo-materials) cannot handle. Constructing risk maps based on few or wide-ranging parameters means that the information is not useful for micro-regions or small communities that are strongly threatened. For this, it is necessary to summon sufficient and competent authorities to go to the field and fill out the questionnaires that the disaster prevention centers have built to study this phenomenon. Unfortunately, this is not always possible, either because of the difficulty in reaching the affected sites or because of the economic limitations to bring observation crews. Then the data has biases from various sources that the modeler must understand and address with appropriate tools. Another important consideration is that the meteorological stations selected to manage the diffuse system (alerts)

High-Level Questions	Responses
	must be maintained and operated in optimal conditions, so it must be protected from vandalism, supplied with energy, and financed so that it works and communicates without loss of information. On the modeling side, the inputs and outputs constitute a challenge by themselves since they have different natures. Some are vectors relative to depth, for some their meaning is in the plane, in others the categorization is too general (regional maps) and when it is lowered to small areas it loses resolution or relevance and must be discarded. Some of the parameters change on time and this must be introduced in the model. On the other hand, when the displaced volume is measured, sufficiently precise tools are not always available, and the data may cause inconsistencies in the model. Also, and very important is that this project is based on information cause-effect from an historical perspective, i.e., using simple and easy to get information from past events. In order to increase the predictive capabilities of the model, it is necessary to instrument specific geo-situations where movements are expected and from which closer symptoms could be obtained (displacement monitors, humidity, for example), behaviors more related to the dynamics could be observed and survey of more comprehensive scenarios could be developed. Through the histories thus recorded, the understanding of the susceptibility and the beginning of the movement because of intense rains could be improved.

"Geographical Data Science Applied to Landslide and Debris Flow Hazard in the Colombian Andes"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Landslide Monitoring and Detection
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Geographical Data Science Applied to Landslide and Debris Flow Hazard in the Colombian Andes
c. Please provide a short description of the use case.	Landslides are one of the most naturally occurring phenomena with the highest human and economic losses around the world, making susceptibility and hazard assessment a fundamental tool for land use planning. There is a wide range of Artificial Intelligence algorithms in the recent literature with completely different approaches to establish the

High-Level Questions	Responses
	relationship between the independent variable (predictors) and the dependent variable (landslide inventory). In the present study, a wide range of algorithms were used for the La Miel creek basin, in the Colombian Andes, and the methodology implemented for this type of data-based modeling is presented in detail and step by step. The results obtained show that the assembled boosting models present the best values in terms of performance and predictability. Contrasting with the linear parametric models, pointing dataset was derived from two sources: (see below)
d. Please provide a short description of the datasets.	1. 5 m x 5 m digital elevation model from which from ArcGis the variables of slope, aspect, roughness, profile curvature, plane curvature, standard curvature, elevation, Stream Power Index (SPI), Topographic Wetness Index (TWI) and flow accumulation were obtained. 2. The landslide inventory was obtained from the photo-interpretation of aerial images of the area at a scale of 1:10000 and the historical events reported by the Colombian Geological Service through SIMMA (Information System of Mass Movements) in the basin area.
e. Please provide a short description of the model/method.	The models used to predict the susceptibility maps were: stochastic gradient boosting, random forest, support vector machines, xgboost, decision tree, adaboosting, linear discriminant analysis, artificial neural network, logistic regression, K nearest neighbors.
f. Please provide a short description of communications technologies that benefit or result from this use case.	The main result of the project is the mass movement susceptibility map, with the best model built with the available data. This map can be used by decision makers as an input for a more complete risk analysis involving temporal and economic factors, and eventually in land-use planning.
2. Data-related questions	
a. Please provide information about data themes.	The predictor variables were initially selected according to Reichenbach et al., (2018). A total of 14 predictor variables of continuous and categorical type were constructed using Arcgis 10.5 software. The continuous variables were: slope, aspect, roughness, profile curvature, planar curvature, standard curvature, elevation curvature, standard curvature, elevation, Stream Power Index (SPI), Topographic Wetness Index (TWI) and flow accumulation; these variables were obtained from the Digital Elevation Model (DEM) with a spatial

High-Level Questions	Responses
	resolution of 5 m x 5 m developed by the Instituto Geográfico Agustín Codazzi (IGAC) in the CartoAntioquia project. The categorical variables used were: geology, elaborated at a scale of 1:10,000 by the Aburrá Valley Seismic Microzonification Study (AMVA), distance to faults, distance to drainage, and finally, the land cover map, elaborated with the map of land cover elaborated with aerial photographs of the years 2010-2011, which were georeferenced and digitized.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	The data was obtained from the Digital Elevation Model with a spatial resolution of 5 m x 5 m developed by the Agustín Codazzi Geographic Institute (IGAC) in the CartoAntioquia project. The categorical variables used were: geology, elaborated at a scale of 1:10000 by the Aburrá Valley Seismic Microzonification Study (AMVA, 2006), which is publicly available.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	The data was analysed and processed on the authors' personal computers.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	For input-data sharing there are no restrictions about confidentiality, privacy, or intellectual property rights.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	The methodological sequence implemented in the present study is summarized in the following five steps: (i) Preparation of the mass movement inventory, (ii) Exploratory data analysis and variable selection, (iii) Application of ML algorithms, (iv) Model validation and hyperparameter optimization, and (v) Generation of susceptibility maps.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	The input data is open and there are no conflicts of confidentiality.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	The data was provided by IGAC in the CartoAntioquia project. As for software licenses, ArcGIS and Python (which is open and free) were used.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	The most important thing to keep in mind when applying AI to this type of problem is that the database must be robust and effectively represent the reality of the target variable. In our case, we consider it fundamental to perform a proper photo interpretation and to be sure of the incorporation of the historical databases that have events. If there is not a good landslide inventory, there simply will not be good results, since in machine learning it is well known that

High-Level Questions	Responses
	"trash in, trash out", so if we do not have a solid base we will only receive bad results. Regarding the implementation of the algorithms it is important to keep in mind when using geospatial data that we are working with big data. Due to the high amount of pixels that raster images have, it is vital to oversample or undersample, since it is an unbalanced problem (the pixels of mass movements are much smaller than those that are not). This affects the learning of the model because if the dataset is not balanced with some technique it will predict only the cells that are not landslides, which would not have any relevance. On the other hand, performing these subsampling techniques has a huge impact on the computational cost of the algorithms, so they are highly recommended to be performed.

"Improving Landslide Prediction by Machine Learning and Deep Learning"

For this use case, no completed questionnaire was received by the submission deadline. Therefore, the details of this use case have been omitted during the derivation of best practices in this technical report.

"Soft Computing Paradigm for Landslide Monitoring and Disaster Management"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Landslide Monitoring and Detection
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Soft Computing Paradigm for Landslide Monitoring and Disaster Management
c. Please provide a short description of the use case.	The Remote Sensing of disasters such as landslides is one of the most important forms of gathering information prior to the occurrence of a catastrophe. The use case is the usage of the space-borne technique for creation of landslide susceptibility maps (LSM) for the region of Nainital, India using machine learning algorithms.
d. Please provide a short description of the datasets.	In our study of the region of Nainital, the Landslide Inventory Map (LIM) has been downloaded for the region from the Bhukosh portal provided by the Geological Survey of India at a scale of 1:100000.

High-Level Questions	Responses
	Geological data of various regions in India can be downloaded from this portal.
e. Please provide a short description of the model/method.	The machine learning algorithms of Maximum Likelihood, ISO and Random Forest are used for creation of the landslide susceptibility map of Nainital.
f. Please provide a short description of communications technologies that benefit or result from this use case.	Not Applicable
2. Data-related questions	
a. Please provide information about data themes.	Not Applicable
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	Data for LCF creation has been downloaded from the Bhukosh portal provided by the Geological Survey of India at a scale of 1:100000. The DEM model has been downloaded obtained from the LANDSAT 7 dataset.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	Bhukosh portal of Geological Survey of India and LANDSAT 7 dataset.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Bhukosh portal of Geological Survey of India provides landslide inventory in shapefile format with vector points. DEM data is downloaded in the .tif format.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	The landslide inventory data downloaded from the Bhukosh portal is projected in ArcMap. It is a shape file with landslide points. The DEM data is also overlaid on the shape file of the study area. DEM is in the raster format with the resolution of 30 m. The different landslide conditioning elements such as slope, aspect, altitude, curvature, Sediment Transport Index (STI), Wetness Index (WI) have been derived from DEM using different ArcMap functions. Land use land-cover map of the study area is prepared using the unsupervised ISO clustering algorithm.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	Humans or animals are not included in the experimentation.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	Both the datasets used in this use case, landslide inventory points and digital elevation data are available as open data in Bhukosh and Bhuvan portal respectively.

High-Level Questions	Responses
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	Using the AI based machine learning models to build LSM requires accurate and sufficient data for training. The amount of data used in training also matters while assessing the efficiency of the model. Too many or too few landslide points might lead to overfitting and underfitting problems respectively. There are different sources from which the dataset of landslide inventory and DEM can be downloaded. Verification of these data sources for correctness is important before using it in our implementation.

A.5 TG-AI for snow avalanche monitoring and detection

"AI for Snow Avalanche Monitoring and Detection"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Snow Avalanche Monitoring, Detection, and Forecasting
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	AI for Snow Avalanche Monitoring and Detection
c. Please provide a short description of the use case.	In this use case, we focus on the use of AI to improve avalanche detection methods to obtain more accurate and reliable avalanche data. Such AI methods are poised to drastically change operational avalanche forecasting.
d. Please provide a short description of the datasets.	We use data from ground-based detections systems (radar, infrasound and seismic) and avalanche observations from automatic camera systems and field surveys.
e. Please provide a short description of the model/method.	We intend to use machine learning models (e.g. random forest) to automatically detect avalanche signals.
f. Please provide a short description of communications technologies that benefit or result from this use case.	Results from our work will be used in operational avalanche forecasting, will be published in open access papers, and will be disseminated to avalanche professionals in courses.
2. Data-related questions	

High-Level Questions	Responses
a. Please provide information about data themes.	Global Geodetic Reference Frame, Elevation and depth, Geology and Soils, Orthoimagery
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	The data we use are collected by systems deployed by the SLF.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	The data from our detection systems is streamed in real-time to the Swiss Seismological service, where we can download it for further analysis.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Data is stored in miniseed format, a standard format for seismological data.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	The raw data are first filtered with a frequency bandpass filter, then features are extracted (amplitude, frequency content, duration, etc..)
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	Not applicable at this stage
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	Once we publish a paper, the data are published as well.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	The main challenge in our field is obtaining reliable ground truth data to train our models. Avalanches are relatively rare events, and mostly occur during periods of bad visibility.

"Limitations of Predicting Snow Avalanche Hazards in Large Data Sparse Regions"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Snow Avalanche Monitoring, Detection, and Forecasting

High-Level Questions	Responses
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Limitations of Predicting Snow Avalanche Hazards in Large Data Sparse Regions
c. Please provide a short description of the use case.	Our use case explores relationships in snow avalanche datasets including observation, model, and expert assessment data, with findings that highlight limitations of using AI methods to predict avalanches in large data-sparse regions.
d. Please provide a short description of the datasets.	Our data includes expert assessments of avalanche danger and character from western Canada as well as relevant snowpack and weather datasets (both from field observations and model generated datasets).
e. Please provide a short description of the model/method.	We explore relationships with classification trees (e.g., conditional inference trees).
f. Please provide a short description of communications technologies that benefit or result from this use case.	Our work has informed operational avalanche forecasters about inconsistencies in their assessments and supported the development of dashboards that illustrate uncertainties in their datasets.
2. Data-related questions	
a. Please provide information about data themes.	Global geodetic reference frame, elevation and depth, geology and soils
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	Data is managed by Avalanche Canada, Parks Canada, the Canadian Avalanche Association, and Simon Fraser University
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	Observation and hazard assessment data is continually submitted to online databases daily each winter. Model data is generated on AWS and Compute Canada servers.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Observation and hazard assessment data are stored in standardized formats defined by the Canadian Avalanche Association, weather model data in netcdf files, and snowpack model data in a custom developed data object for R software.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	We run scripts to compile our datasets on a daily basis during the winter season to support operational forecasting and also in annual batches during the summer for research projects and testing model upgrades.

High-Level Questions	Responses
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	Some avalanche occurrence data is proprietary, and we are only allowed to share it in aggregate form for research purposes.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	Data supporting our research papers is published on the Open Science Framework and we have published several R software packages on CRAN.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	Our challenge is obtaining ground truth data that describes the true likelihood of avalanches across space and time. We also deal with challenges of communicating the complex data and uncertainties to avalanche forecasters.

A.6 TG-AI for wildfire monitoring and detection

"An Intelligent Big Data Analysis System for Wildfire Management"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Wildfire Monitoring and Detection
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	An Intelligent Big Data Analysis System for Wildfire Management
c. Please provide a short description of the use case.	Our existing work is to build an intelligent big data analysis system for fire management, which uses IoT equipment and AI technology to monitor potential fire risks in real time and assess the risks in key areas. This system has been applied in China's provincial regions and is extending to forest fire management.
d. Please provide a short description of the datasets.	Training and testing data mainly come from public and private datasets, which include popular image datasets like ImageNet, COCO and data collected from remote sensing satellites, monitoring devices and social media. AI models pre-trained on top datasets like ImageNet, COCO and DOTA display high accuracy in wildfire detecting. Datasets of remote sensing forest images and monitoring pictures are important in risk assessment, which contain forest terrain, plant species, dryness, tree density and distribution, as well as plant growth and leaf oil composition. Now tremendous existing

High-Level Questions	Responses
	data sources like DOTA, RSSCN7, which include remote sensing data for forest and trees guarantee the accuracy of wildfire predicting models.
e. Please provide a short description of the model/method.	By applying computer vision (CV) and natural language processing (NLP) techniques, AI systems can help to reduce wildfire loss significantly. In detail, the wildfire AI system includes an object detection model, image classification model, image segmentation model etc. Additionally, AI systems can assess wildfire damage and generate restoration plans precisely after a disaster. For the purposes of this proposal, however, we are focusing on wildfire detection and risk mapping.
f. Please provide a short description of communications technologies that benefit or result from this use case.	There are several IoT equipment (remote sensing satellites, monitors, social media apps, etc.) for supporting the wildfire detection and risk assessment system. These communications technologies reduce labor and business costs by predicting wildfire risk and marking high risk areas in advance, reporting wildfire immediately, predicting wildfire spread and guiding wildfire fighting accurately, rescuing trapped people quickly and safely.
2. Data-related questions	
a. Please provide information about data themes.	1. The forest dataset includes remote sensing forest images and monitoring pictures, which contain forest terrain, plant species, dryness, tree density and distribution, as well as plant growth and leaf oil composition. 2. The wildfire dataset contains wildfire and smoke pictures and remote sensing data includes fire reflectance, radiance and emissivity.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	The data is centralized in an independent LAN / San environment, and each server cabinet and area has direct cable connection. Moreover, data governance includes management and master data management (mcm) enables production monitoring, controlling and data collection.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	The data centers realize the centralized processing, storage, transmission, exchange and management for data.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer,	The data curation and delivery include extract-transform-load, streaming ingestion, and data wrangling, by the help of advanced analysis

High-Level Questions	Responses
creation of metafiles for data search, and data formats).	technology and open source frameworks (such as R, Apache spark, knime, rapidminer) at the same time.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	There are several approaches for data processing, including data collection, data cleaning, feature engineering, data classification and data visualization. In order to improve the accuracy of data processing, the project uses data algorithms for instance data augmentation algorithms, data labeling algorithms, data cleaning algorithms, etc.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	Data property rights should emphasize the priority property rights of individuals to data, so as to restrict the data utilization and transaction behavior of enterprises. The commercialization of data generated by data transactions will bring great harm to personal privacy and unpredictable information security problems. A wide range of uncontrolled data transactions will also provide a hotbed for illegal activities.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	The application of artificial intelligence for natural disasters, especially for fire management, is still in the exploratory stage, the application is relatively scattered, the available data and standard AI model is lacking. Therefore, there are many challenges that we have to face. Based on the experience of AI systems for wildfire management, we hope to summarize a system architecture to provide reference for AI application and research in natural disaster in the future, including innovative core applications, data requirements, and standard AI methods.

"Wildland Fire Detection and Strategic Intelligence from Camera and Satellite Data Analyzed Using AI"

For this use case, no completed questionnaire was received by the submission deadline. Therefore, the details of this use case have been omitted during the derivation of best practices in this technical report.

"Multimodal Databases and Artificial Intelligence for Airborne Wildfire Detection and Monitoring"

High-Level Questions	Responses
1. General information about the use case	

High-Level Questions	Responses
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Wildfire Monitoring and Detection
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Multimodal Databases and Artificial Intelligence for Airborne Wildfire Detection and Monitoring
c. Please provide a short description of the use case.	AI methods for wildfire detection and monitoring and data annotation pipelines
d. Please provide a short description of the datasets.	Multimodal datasets comprising thermal and visible range data for airborne wildfire detection and monitoring
e. Please provide a short description of the model/method.	Deep neural networks using transfer learning and interpretable fuzzy modeling approaches
f. Please provide a short description of communications technologies that benefit or result from this use case.	N/A
2. Data-related questions	
a. Please provide information about data themes.	N/A
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	N/A
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	N/A
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	N/A
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	N/A
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	N/A

High-Level Questions	Responses
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	N/A
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	N/A

A.7 TG-AI for vector borne disease forecasting

"AI and Vector-Borne Diseases"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Vector borne Disease Forecasting
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	AI and Vector-Borne Diseases
c. Please provide a short description of the use case.	In this use-case, field data from surveillance efforts for mosquitoes which are able to transmit diseases to humans (i.e., act as vectors of disease) are used to train machine learning models. The models are able to predict the spatio-temporal distribution and seasonality of certain mosquito species, which in turn can aid in vector control strategies. The ultimate aim is to mitigate the risk of vector-borne disease outbreaks.
d. Please provide a short description of the datasets.	Climate data (e.g. CMIP6 or ERA5), land-use (LUH2) and population density data can be used to spatio-temporally characterize a grid for which field surveillance data are available, in order to train the models and perform predictions.
e. Please provide a short description of the model/method.	The field surveillance data are summarized into monthly presence/absence form for each grid cell/month, which are characterized by the climate, land-use and population density data. A binary classification machine learning model is then trained on this data, to predict whether a grid cell in a specific point in time has the appropriate conditions for the vector to survive (i.e., predict habitat suitability).

High-Level Questions	Responses
f. Please provide a short description of communications technologies that benefit or result from this use case.	The models create forecasts of vector habitat suitability with a monthly temporal resolution, which stakeholders can use for policy decision support.
2. Data-related questions	
a. Please provide information about data themes.	Feature Vectors: - Global Geodetic Reference Frame - Land Use - Population Density Labels: - Boolean indicator of habitat suitability
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	The data used for the feature vectors are available through open-access databases. The field surveillance data are a mixture of open-access data and restricted data acquired from individuals performing field-surveillance schemes in the AIM-COST initiative.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	The input feature vectors were obtained through the Copernicus DataStore and NASA-NEX GDP database to the Cyprus Institute High Performance Computing (HPC) facility, where they were treated accordingly to match the field surveillance spatial and temporal resolution.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	The data used in the study are saved in the CyI's HPCF, which adheres to industry standard security protocols. The HPC facility is open to external applications for acquiring access to it and the end user can access the input vector data without restrictions. Some restrictions apply on labeled output classified data.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	Data from the NASA-NEX-GDP database (CMIP6 climate data) and Land Use Harmonization 2 project (LUH2) are transferred to the CyI's HPC facility, where they are treated accordingly to match a 720*1440 (lat-lon) regular grid with a monthly temporal resolution. The processing of the netcdf files was performed using the Cyclone system of HPCF.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	Some restrictions apply on field-surveillance data, as the ownership lies with the respective group which performed the field surveillance effort. The datasets contain no personal information and are, thus, free of any legal or ethical considerations.

High-Level Questions	Responses
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	We plan to share a github repository which contains the scripts to download and build the input vector data, as well as the trained model in Python, for users to perform predictions.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	Since vectors of diseases have a diverse biology and, especially the tiger mosquito, have been demonstrated to adapt to their local environment extremely effectively, the biological variability has to be taken into account and it's extremely difficult to obtain reliable data from several regions plagued by such problems, such as Africa, Asia and Latin America. A centralized repository for data gathering and management and established common protocols for surveillance and data reporting are crucial for researchers to be able to formulate effective AI models, which are tailor-made for specific regions and vector species.

A.8 TG-AI for volcanic eruption forecasting

"Towards Forecasting Eruptions Using Machine Learning of Volcano Seismic Data"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Volcanic Eruption Forecasting
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Towards Forecasting Eruptions Using Machine Learning of Volcano Seismic Data
c. Please provide a short description of the use case.	We try to locate volcanic tremor associated with the 2018 Lower East Rift Zone Eruption in Hawai'i
d. Please provide a short description of the datasets.	We use earthquake catalogs and provide volcanic tremor locations
e. Please provide a short description of the model/method.	We train a regression model based on seismic amplitudes (features) and earthquake locations (target). We then locate the tremor associated with the 2018 Lower East Rift zone eruption using this model.
f. Please provide a short description of communications technologies that benefit or result from this use case.	N/A

High-Level Questions	Responses
2. Data-related questions	
a. Please provide information about data themes.	All the data is freely available through the IRIS platform. https://www.iris.edu/hq/
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	No restriction
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	Data centers providing catalogs freely. We only use open source software
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	N/A
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	Data collection, preparation, processing until tremor location is reached
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	N/A
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	IRIS network for data and USGS for catalogs
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	The results are promising but we would like to test how they would apply in areas monitored with less sensors.

"Real-time Volcano-Independent Seismic Recognition as Volcano Monitoring Tool"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Volcanic Eruption Forecasting

High-Level Questions	Responses
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Real-time Volcano-Independent Seismic Recognition as Volcano Monitoring tool
c. Please provide a short description of the use case.	Proposal of a real-time seismic-based monitoring system for any volcano using statistical models built by other volcanoes with the ultimate aim of forecasting eruptions and detecting dangerous volcano-seismic (VS) events (such as collapses, floods, explosions…) for people living nearby.
d. Please provide a short description of the datasets.	Waveform data bases (DBs) labeled (a.k.a: manually classified in VS types) of ~ 10 volcanoes and open-access data from internet servers of seismic networks.
e. Please provide a short description of the model/method.	Statistical classification models, built by the labeled DBs, are used to classify continuous VS data remotely retrieved from a monitoring network of one given volcano. Automatic VS-catalogs are built by the classification output and are analysed to detect eruption precursors patterns and VS events involving population safety.
f. Please provide a short description of communications	Even if they're not scheduled in the project, a subsystem of SMS-cell phone warnings (in case of dangerous VS event detection) could be designed - as already exists in other monitoring systems.
2. Data-related questions	
a. Please provide information about data themes.	Not (directly) applicable… I guess that the more appropriate might be Geographical Names referring to the Volcano name to be monitored BUT the real used data are seismic waveforms recorded by a seismometer network deployed around a volcano area.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	Data custodianship directly depends on the data supplier policy. Data used to build our models often are privately shared by our partners and, also, open-access databases. The data used to monitor a given / selected volcano are delivered in 'nearly' real-time (or with a delay of 5-60 minutes) from public servers with no custodianship.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	Data required for model building (system design) are supplied by research centers (universities) and/or Volcano monitoring Observatories (VOs). Data for online (real-time) monitoring (once our system is designed and operative) are freely obtained by public VS FDSN servers. The system can also work to monitor data from non-public VS servers via internet

High-Level Questions	Responses
	private connection. Data for offline monitoring can be supplied by both VOs and research institutions.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Curation and delivery are taken into account by the supplied centers, often involving hard work in designing, deployment and maintenance of the seismic network. Data is delivered via internet protocols, such as stfp, ftp or www – queries and digitally stored.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	1. Data filtering: often in the band of [1-50] Hz 2. Data denoising (standardization) 3. Data (manual) labeling (i.e., VS-cataloging): VS events are manually detected and classified according to their physical source (lahars, floods, tectonic earthquakes, ashfall, etc.) 4. Feature Extraction (data dimensionality reduction): the 1D-continuous data waveform stream is converted into a sequence of feature vectors. Often each feature has about 5-40 components (values) describing a waveform frame lasting 1-20 secs. Stage 2 is not always performed, and stage 3 achieves automatic labeling once the system is running.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	Data ownership is not always clear. It depends on the data supplying center.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	Input data (a.k.a. training DB or labeling data) are freely available only when data suppliers are FDSN servers, and, occasionally, research centers. Output data: (a.k.a VS catalogs automatically built by the VI.VSR proposed system) are planned to be open-access (but in case of monitoring a volcano with non-public data or restricted under the VO policy conditions).
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	Main challenge (in our proposal, and I assume in others similar) is the availability of reliable, open access labeled data to be used to design the AI system. A QA on these DBs is crucial. This QA process may be controlled or taken into account under a standardized protocol. In spite of that, recommendations are clear: – Open data, open-access and open software. – Standardization of evaluation indexes for AI, recognition-based systems (as F1-score, accuracy or similar metrics). – Open-access resources and corpus to compare and evaluate diverse technologies performing the same tasks.

A.9 TG-AI for hail and windstorm hazard mapping

"Unified Methodology for Windstorm and Hailstorm Hazard Modeling and Mapping"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Hail and Windstorm Hazard Mapping
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Unified Methodology for Windstorm and Hailstorm Hazard Modeling and Mapping
c. Please provide a short description of the use case.	AI-based software tool that predicts the probability of observing a convective event for a specific day at a given location under certain atmospheric conditions.
d. Please provide a short description of the datasets.	Tabular dataset of more than 50 years of reported events in the US including location, time, intensity, etc. Reanalysis data providing historical hourly estimates of a large number of atmospheric, land and oceanic climate variables.
e. Please provide a short description of the model/method.	The models used are binary classifiers (yes or no) of different types. Each classifier is given a score depending on its performance, and an ensemble classifier is created using the outputs of the original ones weighted by their scores.
f. Please provide a short description of communications technologies that benefit or result from this use case.	Effective communication of the risks derived from severe convective events to society and stakeholders in the shape of maps of probability of occurrence and return periods.
2. Data-related questions	
a. Please provide information about data themes.	Tabular dataset of more than 50 years of reported events in the US including location, time, intensity, etc. Reanalysis data providing historical hourly estimates of a large number of atmospheric, land and oceanic climate variables.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	Reported events dataset downloaded directly from the Storms Events Database web portal. Reanalysis datasets acquired via the Copernicus Climate Data Store api.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure	N/A

High-Level Questions	Responses
file transfer, creation of metafiles for data search, and data formats).	
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	1. Data gathering 2. Data conversion and cleaning 3. Data enrichment (feature engineering) 4. Data labeling 5. Data fusion
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	N/A
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	N/A
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	The main challenges are: first, to build the labeled data set for model training using sparse and sketchy observational datasets; second, to overcome the extreme data imbalance using resampling techniques.

"Predicting Hail with XBoost in Switzerland"

For this use case, the proponent withdrew the use case. Therefore, the details of this use case have been omitted during the derivation of best practices in this technical report.

A.10 TG-AI for multi-hazard communications technology

"Utilizing AI & Probabilistic Modeling for Strategic Resilience"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Multi-hazard Communications Technologies
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Utilizing AI & Probabilistic Modeling for Strategic Resilience
c. Please provide a short description of the use case.	One Concern combines AI/machine learning and probabilistic modeling with data from the natural and manmade environment to create a digital twin of target regions. The digital twin is used to predict damage to the built environment from natural disasters.

High-Level Questions	Responses
d. Please provide a short description of the datasets.	Data comes from 4 sources: Data vendors (e.g., Corelogic, Estated); Open source directly related (e.g., available from municipalities); Open source indirectly related (e.g., satellite images); Direct collection.
e. Please provide a short description of the model/method.	The model uses k-nearest neighbor and statistical imputation to fill in missing building features. It uses ML techniques (e.g., PCA and logistic regression) to predict damage probabilities given building features and hazard intensities. Random Forest is used to detect potential flood levee locations to construct synthetic levee data for locations with missed ground truth data.
f. Please provide a short description of communications technologies that benefit or result from this use case.	One Concern uses automated emails to communicate about predicted damage during and following a disaster. More broadly, telecommunications infrastructure could be included in the modeling of the digital twin, enabling the estimation of technological resilience.
2. Data-related questions	
a. Please provide information about data themes.	Building characteristics, including addresses and land parcels; lifeline network characteristics, including nodes and edges for graph representation; hazard and peril intensities; climate-change related scenarios; elevation and depth
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	Current data from vendors are subject to contractual restrictions; some municipal data are subject to basic data custodianship; in the future, differential-privacy-preserving data collection platforms are envisioned to collect proprietary data that can improve all aspects of the calculation process; all sensitive Japanese data is stored on cloud services physically located in Japan.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	Current data supply chain makes use of protected cloud services that are then managed within the calculation process subject to relevant data privacy restrictions.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Data curation starts in the data science team supported by the research team. All data is managed centrally. For finalized calculation processes, the data supply chain is developed and maintained by the engineering team. A data Wiki is available for details and a data operations team manages data and ensures data protection requirements are enforced.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data	Ingestion & collection; Preparation & evaluation as to data control requirements; Cleaning & curation including some iteration with data science & research

High-Level Questions	Responses
cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	teams; Curated quality assurance (QA); Format into relevant data stores; Training & calibration; Final calculation; Final QA
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	Primary issue relates to contractual restrictions related to data vendor contracts and some restrictions related to use of open source data. In some jurisdictions (e.g., Europe), other considerations such as GDPR become a constraint on data use. Building locations are obfuscated to protect privacy.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	Follow generally accepted approaches to licensing requirements for using open source code & data
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	The primary challenge for our use cases arises from the fact we cannot directly validate all resilience analytics. ML can be used to synthesize data and generate simulations based on the hybrid physics-based/ML approach. Unsupervised AI/ML and even non-hybrid, supervised AI/ML do not work in this space given the fragmented and incomplete nature of the data. A mix of hybridized modeling and subject-matter expertise is essential to iterate (in a Bayesian manner) useful models to quantify resilience in a consistent, comparable, and benchmarkable manner.

"AI Enabled Citizen-centric Decision Support System for Disaster Managers"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Multi-hazard Communications Technologies
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	AI Enabled Citizen-centric Decision Support System for Disaster Managers
c. Please provide a short description of the use case.	The use case explores how AI can assist disaster managers to use communication tools in an effective way. Using data from the C-DOT developed Integrated Alert System and other media types, the decision support system provides text classification, prediction, and transfer learning through neural network and supervised learning approaches to: (a) Filter information: the model categorizes received information into actionable classes for disaster managers from social networks and other agencies.

High-Level Questions	Responses
	(b) Predict alert scope: the model informs the disaster manager of the best way to target a message to different recipients and with different communication mediums (c) Message content analyser: the model determines the effectiveness of warning messages to be sent to the targeted populations.
d. Please provide a short description of the datasets.	The system uses alert feeds from PAN India Integrated Alert System developed by C-DOT. The data is also prepared from social networking feeds for filtering information. For predicting alert scope, tele-density and other infrastructure data is taken from respective concerned organization's sources in India.
e. Please provide a short description of the model/method.	The system uses various supervised learning algorithms as well as Natural language processing based pre-trained models like BERT.
f. Please provide a short description of communications technologies that benefit or result from this use case.	The decision support system will benefit the disaster managers in effective utilization of communication media like SMS, Internet based notifications, Radio, TV, social media, etc. for alerting vulnerable populations.
2. Data-related questions	
a. Please provide information about data themes.	The use case uses latitude and longitude values, population distribution, and functional areas related to communication services.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	The collected data is stored, analysed, and retained in a centralized database server for modeling and analysis purposes.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	The system is hosted on dedicated servers.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	The data consumers are within India and access to data is available only to authorized users based on their roles.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatio temporal requirements; and structure of feature vectors).	Data collection, Preparation, Pre-processing, and Feature extraction

High-Level Questions	Responses
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	No personally identifiable information is stored. The data is owned by C-DOT, India
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	The data related to labeled disaster related tweets can be found at https://crisisnlp.qcri.org/ The terms of use of data have been stated at https://crisisnlp.qcri.org/terms-of-use.html
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	The main challenge is the data availability and correctness of available data with respect to the ground situation.

"Proposal of an AI Chatbot Use Case as a Multihazard Communication Technologies"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Multi-hazard Communications Technologies
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Proposal of an AI Chatbot Use Case as a Multihazard Communication Technologies
c. Please provide a short description of the use case.	The NICT solution contains: 1. DISAANA: a disaster information analyser, which uses natural language processing (Question & Answering) to discover relevant information from Japanese SNS data (Twitter). 2. D-SUMM: an information summarizer, which uses the "BERT" natural language processing model to derive situational awareness for a specified area. 3. SOCDA: a chatbot system, which uses a rule-based method to distribute and collect disaster information about victims, damage areas, and evacuation places and communicates with first responders. Collected texts are analysed by both DISAANA and D-SUMM, and a big-picture of a damaged area can be drawn with collected disaster-related information. DISAANA and D-SUMM are freely available at https://disaana.jp/ and they have proven to be useful for disaster response of local governments in actual disasters. SOCDA is also freely available at LINE ID:@socda and it is in the process of conducting a demonstration

High-Level Questions	Responses
	test. In addition, some local governments in Japan started to use commercial versions of SOCDA that are customized to each local government.
d. Please provide a short description of the datasets.	Japanese SNS (Twitter and LINE) messages and manually created texts that simulate SNS messages in disaster situations. We prepared a training dataset by annotating these messages to build a ML model.
e. Please provide a short description of the model/method.	Supervised machine learning methods, especially SVMs and BERT, are used. We have been using SVMs until now, but we are now developing it into deep learning such as BERT.
f. Please provide a short description of communications technologies that benefit or result from this use case.	By appealing as a fast-paced medium, SNS can benefit from this use case. Chatbot technology also results from this use case.
2. Data-related questions	
a. Please provide information about data themes.	Text messages on SNSs (Twitter and LINE), which include "Addresses", "Buildings and Settlements", "Geographical Names", "Transport Networks", and "Water". LINE is an SNS similar to WhatsApp, and is very popular in Japan.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	SNS providers, Twitter and LINE. NICT follows each data license agreement.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	(as for real-time data) SNS messages are delivered to NICT via the Internet.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	Both Twitter and LINE provide SNS messages in JSON format. NICT receives SNS messages at an NICT building in Japan, and automatically analyses them and the analysed results are opened for public via Web applications such as https://disaana.jp/
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatio temporal requirements; and structure of feature vectors).	Basic workflows are the same for Twitter and LINE. (1) set the conditions of the data collection, such as the size (the number of SNS messages), time range, sampling conditions, and so on. (2) extract plain text for annotation. (3) carry out annotation. (4) train an AI model and evaluate. (5) observe the evaluated results. If the annotation has a problem, modify the annotation (data cleaning). Repeat (4) and (5). Now we are developing a deep learning model and data set for the model.

High-Level Questions	Responses
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	SNS messages may include personal information.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	NICT will not open the dataset for AI due to the original SNS messages license agreement. NICT can license the software for both business and R&D.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	

"AIDERS: Real-time Artificial Intelligence for DEcision Support via RPAS Data AnalyticS"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Multi-hazard Communications Technologies
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	AIDERS: Real-time Artificial Intelligence for DEcision support via RPAS data analyticS
c. Please provide a short description of the use case.	The AIDERS project aims at developing application-specific algorithms and a novel mapping platform that will harness the large volume of data that first responders are now able to collect through heterogeneous sensors (including visual, thermal, and multispectral cameras, LIDAR, CBRN sensors, etc.) on-board RPAS units, and converting that data into actionable decisions for improved emergency response.
d. Please provide a short description of the datasets.	The AIDERS project uses datasets for training and testing its AI solution acquired from multiple sensors attached as payloads to RPAS units. The datasets include RGB images, thermal images, multispectral images, elevation, structural data from lidar sensors, and multi-gas detection data from CBRNE sensors.
e. Please provide a short description of the model/method.	The AIDERS project utilizes machine learning models such as the Darknet Framework for training, and the tiny version of the YoloV4 Neural Network model for real-time object detection.

High-Level Questions	Responses
f. Please provide a short description of communications technologies that benefit or result from this use case.	Dashboards and Emergency Services Network (ESN) are benefited from the AIDERS use case.
2. Data-related questions	
a. Please provide information about data themes.	The AIDERS AI toolkit uses various data themes for providing rapid situational awareness The data themes include Buildings and Settlements, Elevation and Depth, Geographical Names, Orthoimagery, Physical Infrastructure, Population Distribution, Transport Networks, and Water.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	All the data used in the AIDERS project are either open source or available to first responders' organizations by their respective national data management agencies and can be loaded into the AIDERS AI toolkit.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	Not applicable
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	The data used for training and testing are acquired by sensors onboard RPAS. The raw data are transferred directly from RPAS units to the control station (pc running the AIDERS AI toolkit) and subsequently to KIOS CoE dedicated servers. The raw data from the server can be accessed by the researchers that are working on AIDERS project.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	The image data is collected in real time while the UAVs fly. Then, the images are resized from 4K resolution to FHD and are then fed into the Convolutional Neural Network, where the detection of the objects of interest happens.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	For data acquisition, the privacy and data protection legal framework that is active in each country must be followed accordingly. For the AIDERS project the European Union standardized drone regulations across the continent (Regulation (EU) 2019/947 and Regulation (EU) 2019/945) are followed.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	The AI toolkit is available as an open source software through the project website
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this	The main challenge is the appropriate data acquisition by the RPAS units during emergencies that are then used by the AI solution to provide the necessary output.

High-Level Questions	Responses
application and what recommendations can you offer to someone who intends to apply AI?	

"Situational Awareness System for Disaster Response Using Space-based AI (SARA)"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Multi-hazard Communications Technologies
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Situational Awareness System for Disaster Response Using Space-based AI (SARA)
c. Please provide a short description of the use case.	The use case explores the potential of satellite images, meteorological data and AI to increase the situational awareness against natural disasters. The output is a Geographical Information System (GIS) map showing the most vulnerable areas in a region (e.g., a city) before the event, which can be conveyed into dashboards for early warning and immediate response.
d. Please provide a short description of the datasets.	High-resolution satellite images are acquired for the study area. Typical images have from 4 to 8 spectral bands (ranging from blue to infrared) and a resolution between 0.5 and 3 meters/pixel. Meteorological dataset consists of hourly weather data (wind and precipitations) will be investigated in future. Infrastructure datasets (building locations, roadways, emergency stations, etc.) are shapefiles with geographical coordinates and attributes.
e. Please provide a short description of the model/method.	The main model for satellite image analysis is a Unet-based deep learning model. The model characterizes tree structure and land use properties. In future, we will design a new strategy to make the training procedure less dependent on data via self-learning and consistent learning with unlabeled data, for example using cross-pseudo regression technique (CPR).
f. Please provide a short description of communications technologies that benefit or result from this use case.	Highly-vulnerable geographical locations are conveyed into a GIS dashboard for early warning and immediate response by emergency responders and municipality operators.
2. Data-related questions	

High-Level Questions	Responses
a. Please provide information about data themes.	Physical Infrastructure (e.g., Transportation Networks, Power Lines), Orthoimagery, Land Cover, Demographics and Socioeconomics, and Land Use, Buildings and Settlements
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	Infrastructure and Census population data are open access and are acquired directly from local or state online platforms. Low-resolution satellite images are open access. High-resolution satellite images are purchased from commercial providers (e.g. Maxar, Airbus, Planet). Each provider has different license terms.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	See previous answer
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure file transfer, creation of metafiles for data search, and data formats).	High-resolution satellite images are provided by operators through a secure SSH protocol. Along with images, operators provide already .xml metafiles. The data format is GeoTIFF, which consists of an array of numbers (the actual image) and a header where metadata like geographical reference system projection, no-data values, affine between pixels and geographical units are stored. Infrastructure and Census data are curated and maintained by relevant government agencies on open-access platforms.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	Data is acquired from different sources (satellite providers, state offices, etc.). Satellite images are pre-processed. All data is integrated into a GIS platform for sanity check.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	There is no personal data in this use case. Most of the data is open access. High-resolution satellite images acquired by commercial operators should be generally kept private.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	See answer to b.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	One of the challenges is the transferability: make sure that an approach built for a use case can still work (in case with little modifications) with another use case. Shortage of data to train models is often also a limitation.

"Multi-hazard Use Case for Operations Risk Insights and Day One Relief for Natural Disaster Response"

High-Level Questions	Responses
1. General information about the use case	
a. Please provide the name of the associated topic group (e.g., TG-AI for Flood Monitoring and Detection).	TG-AI for Multi-hazard Communications Technologies
b. Please provide the name of the use case from the proposal (e.g., Flash Flooding Monitoring System in Mexico).	Multi-hazard Use Case for Operations Risk Insights and Day One Relief for Natural Disaster Response
c. Please provide a short description of the use case.	ORI aggregates global, country, regional, or local risk alert data from many trusted sources. ORI applies natural language processing and machine learning to identify higher impact risks.
d. Please provide a short description of the datasets.	ORI uses alert feeds from GDACS, USGS, TWC, NWS, Meteo and many other WMO based national alert services. Plus, ORI ingests and analyses news feeds from 1000's of trusted news sources.
e. Please provide a short description of the model/method.	ORI uses Natural Language Processing for finding and aligning new data to high and medium severity risk events. ORI uses a Support Vector Machine (SVM) – linear programming-based machine learning model for high model result transparency.
f. Please provide a short description of communications technologies that benefit or result from this use case.	ORI uses automated email and Slack based user notifications. Aggregated alert, severity and geospatial location details can also be obtained via API.
2. Data-related questions	
a. Please provide information about data themes.	The data theme for ORI is to collect, aggregate and analyse data from trusted global sources of natural disaster alerts and the news or social media which provide better context for these alerts.
b. Please provide information about data custodianship, acquisition, and management (e.g., export restrictions and analysis centers).	ORI data is ingested, analysed and retained in a central hybrid cloud-based database for the purpose of creating and improving a machine learning model to identify and assess the forecasted severity of these NDM alerts.
c. Please provide information about data supply chains (e.g., data centers and infrastructure).	ORI is hosted in multiple IBM cloud data centers. Data is stored in IBM's Cognitive Enterprise Data Platform (CEDP) in Zurich and the cloud centers.
d. Please provide information about data curation and delivery (e.g., analysis center coordination, data center coordination, secure	ORI data ingestion and primary data consumers are located in the US and India. Data is encrypted upon ingestion and transfer. Metadata for data search is

High-Level Questions	Responses
file transfer, creation of metafiles for data search, and data formats).	available to application and ML model developers in CEDP.
e. Please describe your data workflow process until the AI-ready dataset step (e.g., data collection; data preparation and indexing; data cleaning, analysis, curation, and exploration; data spatiotemporal requirements; and structure of feature vectors).	ORI ingests public and commercial data sources into the CEDP drop zone. Data is accessed for authorized users through the CEDP landing zone. Historical records of impacting risk alerts and the points of interest, plus geospatial details are retained in CEDP.
f. Please comment on legal and ethical considerations (e.g., personal information, data ownership).	ORI is GDPR compliant with an IBM Global Privacy Assessment refreshed annually.
g. Please provide information about open source and open data (e.g., sharing plans, software licenses).	ORI uses a substantial amount of open-source code and publicly available data sources. See the data section for more details.
Based on your use case and your understanding of the other use cases in your topic group, what are the challenges when using AI for this application and what recommendations can you offer to someone who intends to apply AI?	The main challenges to be overcome are the inconsistent reporting and granularity of data globally to develop and maintain an application like ORI. Specifically, a good county or district level of data granularity is available for the US, much of Europe and other developed countries. But, less developed countries typically only have details at a country level. So, deep insights and forecasts are much more challenging for parts of Africa, SE Asia, South America and other regions.

Bibliography

[b-ACL-IJCNLP] The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (2021). Reproducibility Checklist. https://2021.aclweb.org/calls/reproducibility-checklist/ .

[b-Ada] Ada Lovelace Institute, AI Now Institute and Open Government Partnership. (2021). Algorithmic Accountability for the Public Sector. https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/ .

[b-Andrienko] Andrienko, N. and Andrienko, G. (2005). Exploratory Analysis of Spatial and Temporal Data. A Systematic Approach. Springer.

[b-Agrawal] Agrawal, P., et al., (2019). Data platform for machine learning. In Proceedings of the 2019 International Conference on Management of Data (SIGMOD '19). pp. 1803-1816. https://doi.org/10.1145/3299869.3314050 .

[b-Aguinis] Aguinis, H., Gottfredson, R. K., and Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. Organizational Research Methods. 16(2), pp. 270-301. https://doi.org/10.1177/1094428112470848 .

[b-AI HLEG-1] High-Level Expert Group on Artificial Intelligence. (2019). A Definition of AI: Main Capabilities and Disciplines. European Commission.

[b-AI HLEG-2] High-Level Expert Group on AI. (2019). Ethics guidelines for trustworthy AI. European Commission.

[b-Al-Rayani] Al-Rayani, B., Al-Harbi, J., and Al-Ghamdi, M. (2022) Enhancing Security of IoT by Using Blockchain. Open Access Library Journal. 9, 1-14. doi: 10.4236/oalib.1109148.

[b-Albayrak] Albayrak, A., Wei, J., Petrenko, M., Lynnes, C. S., and Levy, R. C. (2013). Global bias adjustment for MODIS aerosol optical thickness using neural network. Journal of Applied Remote Sensing. 7(1), 073514. https://doi.org/10.1117/1.JRS.7.073514 .

[b-altexsoft] altexsoft. (2023, June 26). Data Collection for Machine Learning: Steps, Methods, and Best Practices. https://www.altexsoft.com/blog/data-collection-machine-learning/

[b-ANE] Association of Nordic Engineers (2021). Addressing ethical dilemmas in ai: Listening to engineers. https://nordicengineers.org/wp-content/uploads/2021/01/addressing-ethical-dilemmas-in-ai-listening-to-the-engineers.pdf .

[b-Arthur] Arthur, R., Boulton, C. A., Shotton, H., and Williams, H. T. P. (2018). Social sensing of floods in the UK. PLoS ONE. 13(1), e0189327. https://doi.org/10.1371/journal.pone.0189327 .

[b-Baeza-Yates] Baeza-Yates, R. (2018). Bias on the Web. Communications of the ACM. 61(6), pp. 54-61. https://doi.org/10.1145/3209581 .

[b-Baillie] Baillie, M., le Cessie, S., Schmidt, C. O., Lusa, L., and Huebner, M., for the Topic Group "Initial Data Analysis" of the STRATOS Initiative. (2022) Ten simple rules for initial data analysis. PLoS Computational Biology. 18(2), e1009819. https://doi.org/10.1371/journal.pcbi.1009819 .

[b-Barocas] Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org.

[b-Bartelt] Bartelt, P. and Lehning, M. (2002). A physical SNOWPACK model for the Swiss avalanche warning: Part I: numerical model. Cold Regions Science and Technology. 35(3), pp. 123-145. https://doi.org/10.1016/S0165-232X(02)00074-5 .

[b-Bartley] Bartley, P. (2021). DataOps Dilemma: Survey Reveals Gap in the Data Supply Chain Demand for Data Is Growing, But So Are DataOps Challenges. S&P Global Market Intelligence.

[b-Berthold] Berthold, M. (2020, November 14). You can't eliminate bias from machine learning, but you can pick your bias. Venture Beat. https://venturebeat.com/ai/you-cant-eliminate-bias-from-machine-learning-but-you-can-pick-your-bias/ .

[b-Bradley] Bradley, E. and Trevor, H. (2016). Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Institute of Mathematical Statistics Monographs, Band 5. Cambridge University Press.

[b-Cai] Cai, L. and Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal. 14(2). http://doi.org/10.5334/dsj-2015-002 .

[b-Cambridge] Cambridge. (n.d.). Bias. Cambridge.org dictionary. https://dictionary.cambridge.org/dictionary/english/bias .

[b-Campolo] Campolo, A., Sanfilippo, M., Whittaker, M., and Crawford, K. (2017). AI Now 2017 Report. The AI Now Institute.

[b-Capilnean] Capilnean, T. (2021, June 21). On The Semantics Of Data Bias: Reducing Bias Versus Creating Inclusive AI. Forbes. https://www.forbes.com/sites/forbescommunicationscouncil/2021/06/21/on-the-semantics-of-data-bias-reducing-bias-versus-creating-inclusive-ai/?sh=45e408bff42d .

[b-Cha] Cha, Y.-J., Choi, W., and Büyüköztürk, O. (2017). Deep learning-based crack damage detection using convolutional neural networks. Computer-Aided Civil Infrastructure and Engineering. 32(5), pp. 361-378. https://doi.org/10.1111/mice.12263 .

[b-Chabacano] Chabacano. (2008). Overfitting.svg. Wikipedia. https://commons.wikimedia.org/wiki/File:Overfitting.svg .

[b-Chang] Chang, W. L. and Grady, N. (2019). NIST Big Data Interoperability Framework: Volume 1, Definitions. Special Publication (NIST SP), National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1500-1r2 .

[b-Conner-Simons] Conner-Simons, A. (2021). Major ML datasets have tens of thousands of errors. MIT CSail. https://www.csail.mit.edu/news/major-ml-datasets-have-tens-thousands-errors .

[b-Daniel] Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., and Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. ACM Computing Surveys. 51(1), 7. https://doi.org/10.1145/3148148 .

[b-Davies] Davies, G., Weber, R., Wilson, K., and Cummins, P. (2022). From offshore to onshore probabilistic tsunami hazard assessment via efficient Monte Carlo sampling. Geophysical Journal International. 230(3), pp. 1630–165. https://doi.org/10.1093/gji/ggac140 .

[b-Davis] Davis, B. (2022, July 27). Role of IoT in Disaster Management and Emergency Planning. MYTECHMAG. https://www.mytechmag.com/iot-in-disaster-management/ .

[b-DEC] Data Ethics Commission. (2018). Recommendations of the Data Ethics Commission for the Federal Government's Strategy on Artificial Intelligence. Federal Ministry of the Interior and Community (BMI).

[b-Deweber] Deweber, J. T. et al. (2014). Importance of Understanding Landscape Biases in USGS Gage Locations: Implications and Solutions for Managers. Fisheries. 39(4), pp. 155-163. https://doi.org/10.1080/03632415.2014.891503

[b-Downey] Downey, A. B. (2014). Think Stats, Exploratory Data Analysis in Python Version 2.2. Needham, MA: Green Tea Press.

[b-EC-1] European Commission. (2018). ESS Handbook Methodology for data validation 1.1 Revised edition 2018. Collaboration in Research and Methodology for Official Statistics (CROS).

[b-EC-2] European Commission. (2020). Open Data and Data Bias. https://data.europa.eu/de/news/open-data-and-data-bias

[b-EU-FRA-1] European Union Agency for Fundamental Rights. (2019). Data quality and artificial intelligence: mitigating bias and error to protect fundamental rights. Publications Office. doi: 10.2811/546219

[b-EU-FRA-2] European Union Agency for Fundamental Rights. (2010). EU Charter of Fundamental Rights, Title 3, Article 21: Non-discrimination. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:12010P021 .

[b-EU-GDPR-1] EU General Data Privacy Regulation. (n.d.) Article 15 - Right of access by the data subject.

[b-EU-GDPR-2] EU General Data Privacy Regulation. (n.d.) Article 22 – Automated individual decision-making, including profiling.

[b-Fan] Fan, C., Zhang, C., Yahja, A., and Mostafavi, A. (2021). Disaster City Digital Twin: A vision for integrating artificial and human intelligence for disaster management. International Journal of Information Management. 56, 102049. https://doi.org/10.1016/j.ijinfomgt.2019.102049 .

[b-Fasolin] Fasolin, K. et al. (2013). Efficient Execution of Conjunctive Complex Queries on Big Multimedia Databases. 2013 IEEE International Symposium on Multimedia. pp. 536-543. doi: 10.1109/ISM.2013.112.

[b-Fitter] Fitter, F., Hunt, S. T. (n.d.) How AI can end bias. SAP. https://www.sap.com/insights/viewpoints/how-ai-can-end-bias.html

[b-Ford] Ford, D. N. and Wolf, C. M. (2020). Smart Cities with Digital Twin Systems for Disaster Management. Journal of Management in Engineering. 36(4). https://doi.org/10.1061/(ASCE)ME.1943-5479.0000779 .

[b-Galindo] Galindo, L., Perset, K. and Sheeka, F. (2021), An overview of national AI strategies and policies. OECD Going Digital Toolkit Notes, No. 14. https://doi.org/10.1787/c05140d9-en .

[b-Gancheva] Gancheva, V. (2011). Data Security and Validation Framework for a Scientific Data Processing SOA Based System. 2011 Developments in E-systems Engineering. pp. 576-580, doi: 10.1109/DeSE.2011.75.

[b-Gani] Gani, A., Siddiqa, A., Shamshirband, S., and Hanum, F. (2016). A survey on indexing techniques for big data: taxonomy and performance evaluation. Knowledge and Information Systems. 46, pp. 241-284. https://doi.org/10.1007/s10115-015-0830-y .

[b-Gantz] Gantz, J. and D. Reinsel. (2012). The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. IDC iView: IDC Analyze the Future 2007. pp. 1-16.

[b-Gebru] Gebru, T., et al. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.

[b-Goodrum] Goodrum, W. (2017). Statistical & Cognitive Biases in Data Science: What is Bias?. Elder Research. https://www.elderresearch.com/blog/statistical-cognitive-biases-in-data-science-what-is-bias/

[b-Google] Google Developers. (n.d.) Machine Learning Glossary: Fairness. https://developers.google.com/machine-learning/glossary/fairness

[b-GOV.UK-1] GOV.UK. (n.d.) Discrimination: your rights. https://www.gov.uk/discrimination-your-rights

[b-GOV.UK-2] GOV.UK. (2018). Data Ethics Framework. https://www.gov.uk/government/publications/data-ethics-framework

[b-Hadi] Hadi, H. J., Shnain, A. H., Hadishaheed, S., and Ahmad, A. H. (2015). Big Data and Five V's Characteristics. International Journal of Advances in Electronics and Computer Science. 2(1), pp. 16-23.

[b-Ham] Participatory Sensing and Digital Twin City: Updating Virtual City Models for Enhanced Risk-Informed Decision-Making. May 2020. https://www.researchgate.net/publication/341081867_Participatory_Sensing_and_Digital_Twin_City_Updating_Virtual_City_Models_for_Enhanced_Risk-Informed_Decision-Making

[b-Hansford] Hansford, B. C. and Hattie, J. A. (1982). The Relationship Between Self and Achievement/Performance Measures. Review of Educational Research. 52(1), pp. 123-142. https://doi.org/10.3102/00346543052001123 .

[b-Hashem] Hashem, I. A. et al. (2016). The role of big data in smart city. International Journal of Information Management. 36(5), pp. 748–758. https://doi.org/10.1016/J.IJINFOMGT.2016.05.002 .

[b-Heckert] Heckert, N. and Filliben, J. (2003). NIST/SEMATECH e-Handbook of Statistical Methods; Chapter 1: Exploratory Data Analysis.

[b-Heckman] Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. Econometrica. 47(1), pp. 153-161. https://doi.org/10.2307/1912352 .

[b-Hellström] Hellström, T., Dignum, V., and Bensch, S. (2020). Bias in Machine Learning – What is it Good for? arXiv preprint arXiv:2004.00686.

[b-Hooker] Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". Patterns. 2(4), 100241. https://doi.org/10.1016/j.patter.2021.100241 .

[b-Howe] Howe, B. M. e al. (2022). SMART subsea cables for observing the earth and ocean, mitigating environmental hazards, and supporting the blue economy. Frontiers in Earth Science. 9, 775544. https://doi.org/10.3389/feart.2021.775544

[b-IBM] IBM. (2022). Everyday Ethics for Artificial Intelligence. https://www.ibm.com/design/ai/ethics/everyday-ethics/ .

[b-IBM-2] IBM. (n.d.) Introduction to AI FactSheets. https://aifs360.mybluemix.net/introduction .

[b-IEEE] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html .

[b-ISO-1] International Standard for Organization. (2017). ISO/IEC JTC 1/SC 42 Artificial intelligence. https://www.iso.org/committee/6794475.html .

[b-ISO-2] International Standard for Organization. (Under Development). ISO/IEC FDIS 8183 Information technology – Artificial intelligence – Data life cycle framework. https://www.iso.org/standard/83002.html .

[b-Jambu] Jambu, M. (1991). Exploratory and Multivariate Data Analysis. Academic Press.

[b-Jan] Jan, B. et al. (2017). Deep learning in big data Analytics: A comparative study. Computers and Electrical Engineering. 75, pp. 275-287. https://doi.org/10.1016/j.compeleceng.2017.12.009 .

[b-Jiang] Jiang, X. (2009). Asymmetric Principal Component and Discriminant Analyses for Pattern Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence. 31(5), pp. 931-937. doi: 10.1109/TPAMI.2008.258.

[b-Kaur] Kaur, M., Kaur P. D., and Sood, S. K. (2022). BIoT (Blockchain-based IoT) Framework for Disaster Management. 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India. pp. 318-323. doi: 10.1109/Confluence52989.2022.9734193.

[b-Khan] Khan, N. et al. (2018). The 10 Vs, Issues and Challenges of Big Data. In Proceedings of the 2018 International Conference on Big Data and Education (ICBDE '18). pp. 52-56. https://doi.org/10.1145/3206157.3206166 .

[b-Kozyrkov] Kozyrkov, C. (2019, January 24). What is bias?. Towards Data Science. https://towardsdatascience.com/what-is-ai-bias-6606a3bcb814 .

[b-Kroet] Kroet, C. (2021, October 20). Facial recognition probed across Europe under GDPR ahead of new AI rules. MLex Market Insight. https://mlexmarketinsight.com/news-hub/editors-picks/area-of-expertise/data-privacy-and-security/facial-recognition-probed-across-europe-under-gdpr-ahead-of-new-ai-rules .

[b-Kuglitsch] Kuglitsch, M. et al. (2022). Artificial intelligence for disaster risk reduction: Opportunities, challenges, and prospects. WMO Bulletin. 71(1).

[b-Kumar] Kumar, B. S., Chandrabose, A., and Chakravarthi, B. R. (2021). An Overview of Fairness in Data – Illuminating the Bias in Data Pipeline. In Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion. pp. 34-45.

[b-Kumar-2] Kumar, S. (2022). Collaborative Processing Using the Internet of Things for Post-Disaster Management. Internet of Things and Its Applications. Springer. https://doi.org/10.1007/978-3-030-77528-5_20 .

[b-Lambert] Lambert, J. D. et al. (2021). Ready for the Next Storm; AI-Enabled Situational Awareness in Disaster Response. Johns Hopkins University Applied Physics Laboratory.

[b-Lapuschkin] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. and Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. Nature Communications 10(1096). https://doi.org/10.1038/s41467-019-08987-4 .

[b-Lary] Lary, D. J., Alavi, A. H., Gandomi, A. H., and Walker, A. L. (2016). Machine learning in geosciences and remote sensing. Geoscience Frontiers. 7(1), pp. 3-10. https://doi.org/10.1016/j.gsf.2015.07.003 .

[b-Latonero] Latonero, M. (2018). Governing artificial intelligence: Upholding human rights & dignity. Data & Society.

[b-LaValle] LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., and Kruschwitz, N. (2013). Big data, Analytics and the Path from Insights to Value. MIT Sloan Management Review. 52(2). http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/ .

[b-Leslie] Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

[b-Lin] Lin, J.-T., Melgar, D., Thomas, A. M., and Searcy, T. J. (2021). Early warning for great earthquakes from characterization of crustal deformation patterns with deep learning. Journal of Geophysical Research: Solid Earth. 126, e2021JB022703. https://doi.org/10.1029/2021JB022703 .

[b-Looper] Looper, J. P. and Vieux, B .E. (2012). An assessment of distributed flash flood forecasting accuracy using radar and rain gauge input for a physics-based distributed hydrologic model. Journal of Hydrology. (412-413), pp. 114-132. https://doi.org/10.1016/j.jhydrol.2011.05.046 .

[b-Lundberg] Lundberg, S., Lee S.-I. (2017). A Unified Approach to Interpreting Model Predictions. https://arxiv.org/abs/1705.07874 .

[b-Manning] Manning, C. and Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

[b-Mao] Mao, Z., Xu, Y., and Suarez, E. (2023). Dataset Management Platform for Machine Learning. arXiv preprint arXiv:2303.08301.

[b-Mayer-Schönberger] Mayer-Schönberger, V. and Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt.

[b-Merriam-Webster] Merriam-Webster. (n.d.). Bias. Merriam-Webster.com dictionary. https://www.merriam-webster.com/dictionary/bias

[b-Merritt] Merritt, R. (2020). What is MLOps?. NVIDIA. https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/

[b-Michael] Michael, K. and Miller, K. W. (2013). Big Data: New Opportunities and New Challenges [Guest Editors" Introduction]. Computer 46 (6), pp. 22-24. doi: 10.1109/MC.2013.196.

[b-Microsoft] Microsoft Research. (n.d.) Data Documentation. https://www.microsoft.com/en-us/research/project/datasheets-for-datasets/ .

[b-Mignot] Mignot, E., Li, X., and Dewals, B. (2019). Experimental modelling of urban flooding: A review. Journal of Hydrology. 568, pp. 334-342. https://doi.org/10.1016/j.jhydrol.2018.11.001 .

[b-Mohandas] Mohandas, G. (2022). Monitoring – Made With ML. https://madewithml.com

[b-Mukhiya] Mukhiya, S. K. and Ahmed, U. (2020). Hands-On Exploratory Data Analysis with Python. Packt Publishing.

[b-Munappy] Munappy, A., Bosch, J., Olsson, H. H., Arpteg, A., and Brinne, B. (2019). Data management challenges for deep learning. In 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). IEEE. pp. 140-147. doi: 10.1109/SEAA.2019.00030.

[b-Newcombe] Newcombe, T. (2019, July/August). Facing the Privacy Costs of Biometric Identification. Government Technology. https://www.govtech.com/products/facing-the-privacy-costs-of-biometric-identification.html .

[b-Ntoutsi] Ntoutsi, E. et al. (2020). Bias in data-driven artificial intelligence systems – An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 10(3), e1356. https://doi.org/10.1002/widm.1356

[b-NWS] National Weather Service Juneau. (n.d.) Numerical Weather Prediction (Weather Models). National Oceanic and Atmospheric Administration NOAA.

[b-OGC-1] Open Geospatial Consortium. (2014). GeoPackage Encoding Standard. https://www.geopackage.org/ .

[b-OGC-2] Open Geospatial Consortium. (2023). GeoTIFF Standard. https://www.ogc.org/standards/geotiff .

[b-OGC-3] Open Geospatial Consortium. (2023). Geography Markup Language. https://www.ogc.org/standards/gml .

[b-OGC-4] Open Geospatial Consortium. (2023). Hierarchical Data Format Version 5 (HDF5®) Standard. https://www.ogc.org/standards/HDF5 .

[b-OGC-5] Open Geospatial Consortium. (2022). The HDF5 profile for labeled point cloud data. https://docs.ogc.org/dp/21-077.html .

[b-OGC-6] Open Geospatial Consortium. (2023). Keyhole Markup Language. https://www.ogc.org/standards/kml .

[b-OGC-7] Open Geospatial Consortium. (2023). network Common Data Form (netCDF) standards suite. https://www.ogc.org/standards/netcdf .

[b-Olteanu] Olteanu, A., Castillo, C., Diaz, F., and Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data. 2. https://doi.org/10.3389/fdata.2019.00013 .

[b-Patel] Patel, A. B., Birla, M., and Nair, U. (2012). Addressing big data problem using Hadoop and Map Reduce. 2012 Nirma University International Conference on Engineering, NUiCONE. pp. 1–5. https://doi.org/10.1109/NUICONE.2012.6493198 .

[b-Paulhus] Paulhus, D. L. (1991). Measurement and Control of Response Bias. Measures of Personality and Social Psychological Attitudes, pp. 17–59. Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X .

[b-Pineau] Pineau, J. et al. (2020). Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). https://arxiv.org/abs/2003.12206 .

[b-Piorkowski] Piorkowski, D., Richards, J., Hind, M. (2022). Evaluating a Methodology for Increasing AI Transparency: A Case Study. https://arxiv.org/abs/2201.13224 .

[b-Preston-Werner] Preston-Werner, T. (n.d.) Semantic Versioning. https://semver.org/ .

[b-Ribeiro] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). pp. 1135–1144. https://doi.org/10.1145/2939672.2939778 .

[b-Roh] Roh, Y., Heo, G., and Whang, S. E. (2019). A survey on data collection for machine learning: a big data-ai integration perspective. https://doi.org/10.48550/arXiv.1811.03402 .

[b-Rossetto] Rossetto, T., Allsop, W., Charvet, I., and Robinson, D. I. (2011). Physical modelling of tsunami using a new pneumatic wave generator. Coastal Engineering. 58(6), pp. 517-527. https://doi.org/10.1016/j.coastaleng.2011.01.012 .

[b-Rothman] Rothman, K. J., Greenland, S., and Lash, T. L. (2015). Modern Epidemiology. Wolters Kluwer Health/Lippincott Williams & Wilkins.

[b-Sambasivan] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 39, pp. 1-15. https://doi.org/10.1145/3411764.3445518 .

[b-Scikit] Scikit-learn. (n.d.) 3.1. Cross-validation: evaluating estimator performance. https://scikit-learn.org/stable/modules/cross_validation.html .

[b-Shaw] Shaw, B. E., Milner, K. R., Field, E. H., Richards-Dinger, K., Gilchrist, J. J., and Dieterich, J. H. (2018). A physics-based earthquake simulator replicates seismic hazard statistics across California. Science Advances. 4(8). doi: 10.1126/sciadv.aau0688.

[b-Smuha] Smuha, N. A. (2019). The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence. Computer Law Review International. 20(4), pp. 97-106. https://doi.org/10.9785/cri-2019-200402 .

[b-Sun] Sun, W., Bocchini, P., and Davison, B. D. (2020). Applications of artificial intelligence for disaster management. Nat Hazards. 103, 2631–2689. https://doi.org/10.1007/s11069-020-04124-3 .

[b-Tamagnini] Tamagnini, P., Krause, J., Dasgupta, A., and Bertini, E. (2017). Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations. In Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics (HILDA'17). 6, pp. 1–6. https://doi.org/10.1145/3077257.3077260 .

[b-Teri] Teri, S.S. and Musliman, I.A. (2019). MACHINE LEARNING IN BIG LIDAR DATA: A REVIEW. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XLII-4/W16, pp. 641-644. https://doi.org/10.5194/isprs-archives-XLII-4-W16-641-2019 .

[b-Thüring] Thüring, T., Schoch, M., van Herwijnen, A., and Schweizer, J. (2015). Robust snow avalanche detection using supervised machine learning with infrasonic sensor arrays. Cold Regions Science and Technology. 111, pp. 60-66. https://doi.org/10.1016/j.coldregions.2014.12.014 .

[b-Tobler] Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography. 46, pp. 234–240. https://doi.org/10.2307/143141 .

[b-Tommasi] Tommasi, T., Patricia, N., Caputo, B., and Tuytelaars, T. (2017) A Deeper Look at Dataset Bias. Domain Adaptation in Computer Vision Applications. Advances in Computer Vision and Pattern Recognition. pp. 37-55. https://doi.org/10.1007/978-3-319-58347-1_2 .

[b-Tseng] Tseng, T., Stent, A., and Maida, D. (2020). Best Practices for Managing Data Annotation Projects. arXiv preprint arXiv:2009.11654.

[b-Tukey] Tukey, J. W. (2020). Exploratory data analysis. Hoboken, Nj: Pearson.

[b-UN] United Nations. (2021). Resource Guide on Artificial Intelligence (AI) Strategies.

[b-UN-GGIM] United Nations Committee of Experts on Global Geospatial Information Management. (2019). The Global Fundamental Geospatial Data Themes.

[b-UN-GGIM-2] United Nations Committee of Experts on Global Geospatial Information Management. (2018). E/C.20/2018/7/Add.1 ANNEX 1 – Minimum List of Global Fundamental Geospatial Data Themes.

[b-UN-GGIM-3] United Nations Committee of Experts on Global Geospatial Information Management. (2020). Integrated Geospatial Information Framework (IGIF), Part 2 Global Consultation Draft, Strategic Pathway 2, Policy and Legal.

[b-UN-GGIM-4] United Nations Committee of Experts on Global Geospatial Information Management. (2020). Integrated Geospatial Information Framework (IGIF), Part 2 Global Consultation Draft, Strategic Pathway 4, Data.

[b-UN-GGIM-5] United Nations Committee of Experts on Global Geospatial Information Management. (2020). Integrated Geospatial Information Framework (IGIF), Part 2 Global Consultation Draft, Strategic Pathway 4, Data, Appendices.

[b-UN-OHCHR] United Nations Human Rights Office of the High Commissioner. (2018) A Human Rights Based Approach to Data - Leaving No One Behind in the 2030 Agenda for Sustainable Development.

[b-Verde] The Verde Developers. (2022). K-Fold cross-validation with blocks. https://www.fatiando.org/verde/dev/gallery/blockkfold.html .

[b-Vermiglio] Vermiglio, C., Noto, G., Rodríguez Bolívar, M.P. and Zarone, V. (2022). Disaster management and emerging technologies: a performance-based perspective. Meditari Accountancy Research. 30(4), pp. 1093-1117. https://doi.org/10.1108/MEDAR-02-2021-1206 .

[b-Wang] Wang, Z. et al. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. arXiv:1911.11834.

[b-Wolf] Wolf, K., Dawson, R., Mills, J., Blythe, P., and Morley, J. (2022). Towards a digital twin for supporting multi-agency incident management in a smart city. Scientific Reports. 12, 16221. https://doi.org/10.1038/s41598-022-20178-8 .

[b-Yadavalli] Yadavalli, K. K. and Gudino, L. J. (2022). An Autonomous, Scalable and Low-Cost IoT Based Framework for Disaster Management System. 2022 13th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP), Porto, Portugal. pp. 619-624. doi: 10.1109/CSNDSP54353.2022.9907972.

[b-Yu] Yu, D. and He, Z. (2022). Digital twin-driven intelligence disaster prevention and mitigation for infrastructure: advances, challenges, and opportunities. Natural Hazards. 112, pp. 1–36. https://doi.org/10.1007/s11069-021-05190-x .

[b-Zanetti] Zanetti, M., Allegri, E., Sperotto, A., Torresan, S., and Critto, A. (2022). Spatio-temporal cross-validation to predict pluvial flood events in the Metropolitan City of Venice. Journal of Hydrology. 612, Part B. 128150. https://doi.org/10.1016/j.jhydrol.2022.128150 .

[b-Zhou] Zhou, L., Pan, S., Wang, J. and Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. Neurocomputing. 237, pp. 350-361. https://doi.org/10.1016/j.neucom.2017.01.026 .