Challenges for Transparent and Trustworthy Machine Learning

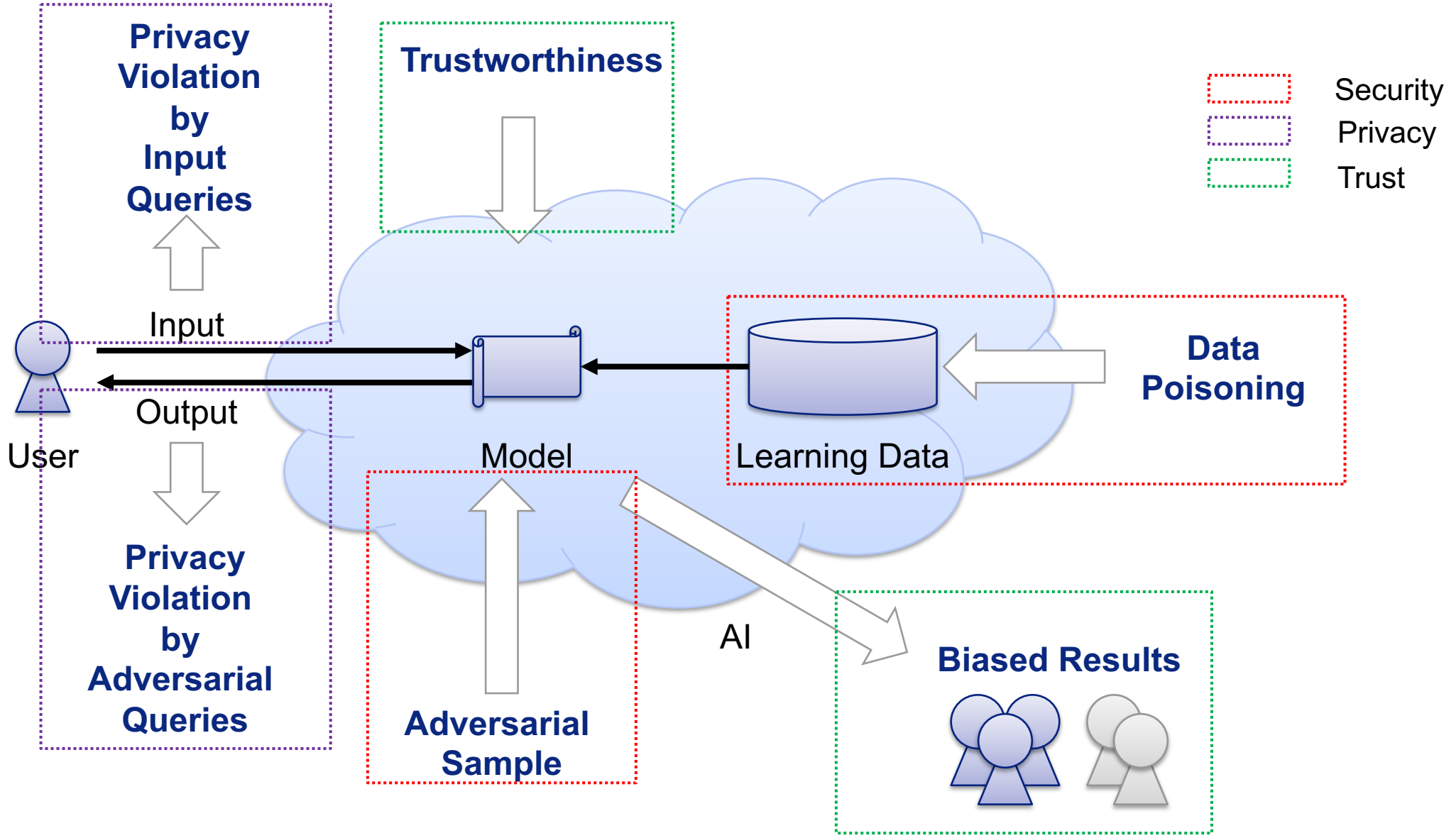
Vanessa Bracamonte

2019/01/21

KDDI Research, Inc.

- General concerns on AI
- Why Transparency
- Transparency as Interpretability in ML Research
- Challenges

General concerns on AI



Use cases

KDDI Research

Real-world examples of the scenarios in this survey

All four of the concepts discussed in the survey are based on real-life applications of algorithmic decision-making and artificial intelligence (AI):

Numerous firms now offer **nontraditional credit scores** that build their ratings using thousands of data points about customers' activities and behaviors, under the premise that "all data is credit data."

States across the country use **criminal risk assessments** to estimate the likelihood that someone convicted of a crime will reoffend in the future.

Several multinational companies are currently using AI-based systems **during job interviews** to evaluate the honesty, emotional state and overall personality of applicants.

http://www.pewinternet.org/2018/11/16/public-attitudes-toward-computer-algorithms/

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

BBC Sign in News Sport Weather Shop Reel Travel Mo

NEWS

Home Video World Asia UK Business Tech Science Stories Entertainment &

Technology

Amazon scrapped 'sexist AI' tool

10 October 2018 Share

https://www.bbc.com/news/technology-45809919

Real-world examples of the scenarios in this survey

All four of the concepts discussed in the survey are based on real-life applications of algorithmic decision-making and artificial intelligence (AI):

Numerous firms now offer **nontraditional credit scores** that build their ratings using thousands of data points about customers' activities and behaviors, under the premise that "all data is credit data."

States across the country use **criminal risk assessments** to estimate the likelihood that someone convicted of a crime will reoffend in the future.

Several multinational companies are currently using AI-based systems **during job interviews** to evaluate the honesty, emotional state and overall personality of applicants.

KDDI Research

Why Transparency

- “**Requirements for Trustworthy AI** from the earliest design phase: Accountability, Data Governance, Design for all, Governance of AI Autonomy (Human oversight), Non- Discrimination, Respect for Human Autonomy, Respect for Privacy, Robustness, Safety, **Transparency**. “

European Commission’s High-Level Expert Group on Artificial Intelligence (AI HLEG) “Ethics guidelines for trustworthy AI (Draft)”

Transparency

■ Specialized Conferences and Workshops

- Fairness, Accountability, and Transparency in Machine Learning Workshop (since 2014)
- Workshop on Human Interpretability in Machine Learning (since 2016)
- Workshop on Explainable Artificial Intelligence (XAI) (since 2017)
- ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*) (since 2018)

■ Documents

- The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems., “Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems. Version 2.” (2017)
- European Commission’s High-Level Expert Group on Artificial Intelligence (AI HLEG) “Ethics guidelines for trustworthy AI (Draft)” (2018)

Transparency in ML technical literature

KDDI Research

■ Black box model

- Inner workings of a model and the reason for its outcomes are not understood

■ **Examples** of models described as ”black boxes”

- neural networks (Ribeiro, Singh, & Guestrin, 2016; Chu, Hu, Hu, Wang, & Pei, 2018)
- decision trees and random forests (Ribeiro, Singh, & Guestrin, 2016; Krause, Perer, & Ng, 2016)
- matrix factorization (Abdollahi & Nasraoui, 2017), latent factor models (Peake & Wang, 2018).

Ribeiro, M. T., Singh, S., & 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* (pp. 1135–1144).

Chu, L., Hu, X., Hu, J., Wang, L., & Pei, J. (2018). Exact and Consistent Interpretation for Piecewise Linear Neural Networks: A Closed Form Solution. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1244–1253).

Krause, J., Perer, A., & Ng, K. (2016). Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5686–5697).

Abdollahi, B., & Nasraoui, O. (2017). Using Explainability for Constrained Matrix Factorization. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (pp. 79–83).

Peake, G., & Wang, J. (2018). Explanation Mining: Post Hoc Interpretability of Latent Factor Models for Recommendation Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2060–2069).

Interpretability

- For works that describe machine learning models as black boxes, **transparency and interpretability** are closely related, if not the same concept.

- Common approach proposed to address the opacity of models is through improving that interpretability
- **Post hoc interpretability**
  - Aims to explain the resulting prediction of black box models
- **Interpretable models**
  - introduce interpretability in the model itself

Post hoc interpretability

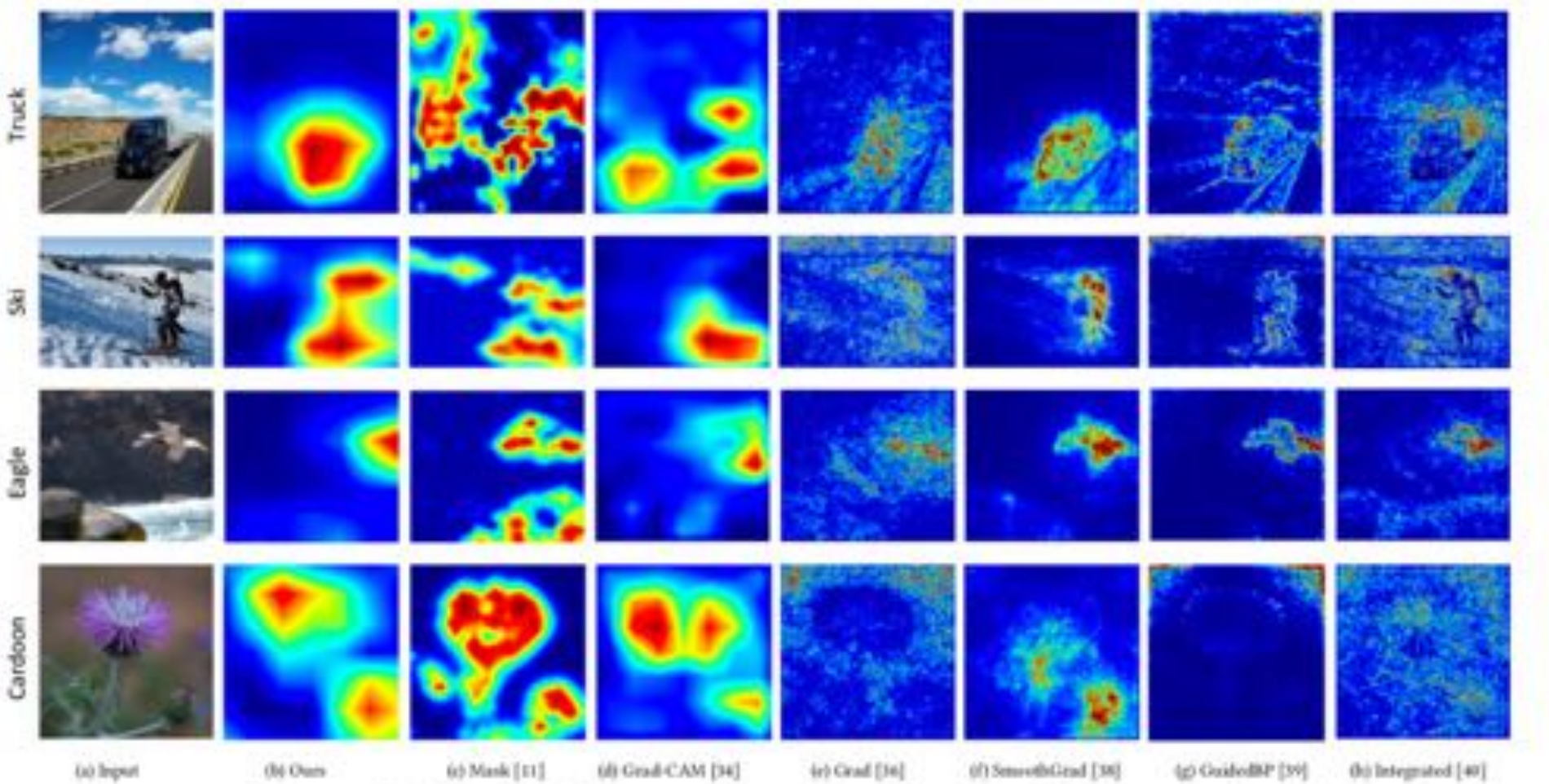


Figure 2: Visualization saliency maps comparing with 6 state-of-the-art methods.

Du, M., Liu, N., Song, Q., & Hu, X. (2018). Towards Explanation of DNN-Based Prediction with Guided Feature Inversion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1358–1367

Interpretable Models

- Interpretability included or as a characteristic of the model

Figure 5: Decoded prototypes when we include R_1 and R_2.

- In this example, the model includes two interpretability regularization terms (“R1 helps make the prototypes meaningful, and R2 keeps the explanations faithful”)

Figure 6: Decoded prototypes when we remove R_1 and R_2.

Figure 7: Decoded prototypes when we remove R_1.

Figure 8: Decoded prototypes when we remove R_2.

Li, O., Liu, H., Chen, C., & Rudin, C. (2018). Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*

Interpretability and Trustworthiness

- If the results of a model can be interpreted, that provides information that can help decide on the trustworthiness of the model

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

Ribeiro, M. T., Singh, S., & 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* (pp. 1135–1144).

Challenges

- Clarification of
  - Scope and definitions
  - Stakeholders
  - What information to show and how to show it?

- Transparency Risks

Scope and Definitions

- Focus seems to be on transparency as model interpretability, but **transparency is a wider** concept.

- Other aspects of transparency include **data transparency**, for example.
  - Machine learning models are highly dependent on the data that they were trained on, and therefore understanding the characteristics of that data is important to understand the outcomes as well.

- **Social, legal and political** considerations play a important role.
  - In particular, **transparency** for the purposes of obtaining information that will help stakeholders such as government or regulators decide on the application of a machine learning model for people is a complex issue beyond the difficulty of understanding the inner workings or the outcomes of a model.

Scope and Definitions

■ Principle of Explicability: “Operate transparently”

- Explicable: **Intelligible** and **explainable**
- “Technological transparency implies that AI systems be **auditable, comprehensible** and **intelligible by human beings** at varying levels of comprehension and expertise.” (AI HLEG, 2018)

European Commission’s High-Level Expert Group on Artificial Intelligence (AI HLEG) “Ethics guidelines for trustworthy AI (Draft)”

Scope and Definitions

- Interpretability techniques are important for transparency
- Interpretability is considered in relation to people —**human interpretability**, but it it's **not clearly defined** (or not defined at all) in many cases
- Without clear definitions, it is **difficult to evaluate** if interpretability has indeed been achieved

Stakeholders

- If the goal is human/user interpretability, who is the user?
- Works on interpretability focus too narrowly on the developer as the stakeholder, with rare exceptions.
- Other stakeholders should be considered
  - Role-based stakeholders (Tomsett et al., 2018): creators, operators, executors, decision-subjects, data-subjects and examiners.
  - Each of these have their own perspective and needs, and correspond to developers, regulators, or the general public, who **may require different types of information**.

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. Presented at the ICML Workshop on Human Interpretability in Machine Learning (WHI 2018).

Stakeholders

- Once the identified, the question becomes whether these users can interpret or understand the meaning of the machine learning model outcomes.
- However, often **interpretability is not validated** by the stakeholders
  - Ties to the challenge of definition
- Claims that techniques improve interpretability are often founded on **assumptions of how obvious the information** appears to be.
- While this approach may be serviceable for use cases such as image recognition, it may not be enough for more **complex cases**.

Desired level of information

- **What information to show?**
- **How to show information?**

- Not enough research on usable and practical explanations (Abdul, Vermeulen, Wang, Lim, & Kankanhalli, 2018)

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 582:1–582:18).

Information: what and how

■ Visualization of results/explanation

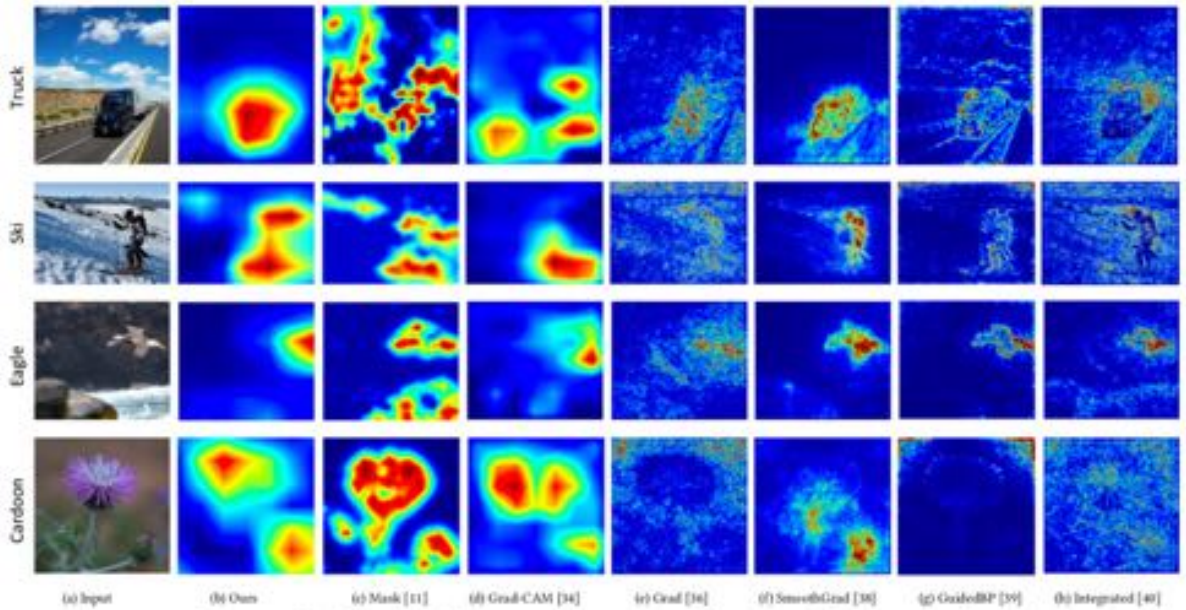
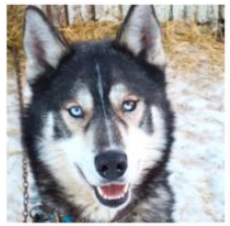


Figure 2: Visualization saliency maps comparing with 6 state-of-the-art methods.

Du, M., Liu, N., Song, Q., & Hu, X. (2018).



(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016).

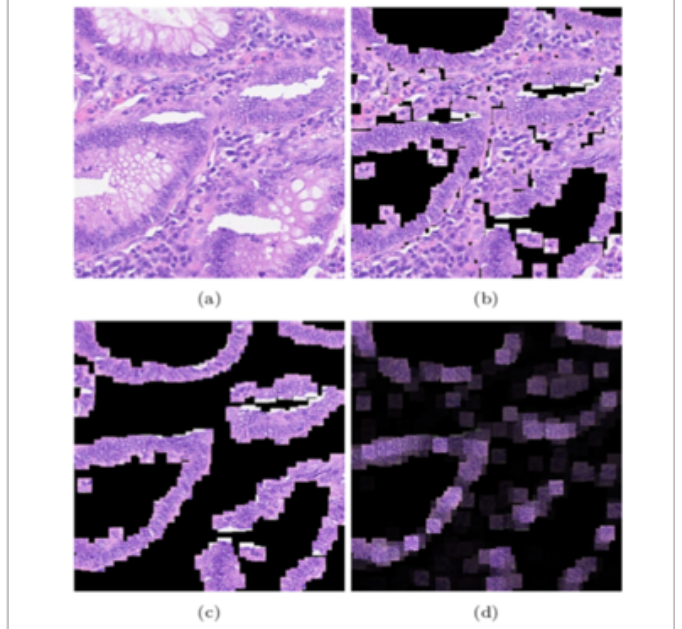


Figure 5. (a) H&E stained histology image. (b) 27×27 patches centered around all marked nuclei. (c) Ground truth: Patches that belong to the class epithelial. (d) Heatmap: Every patch from (b) multiplied by its corresponding attention weight, we rescaled the attention weights using $a'_k = (a_k - \min(\mathbf{a}))/(\max(\mathbf{a}) - \min(\mathbf{a}))$.

Ilse, M., Tomczak, J., & Welling, M. (2018). Attention-Based Deep Multiple Instance Learning. In *International Conference on Machine Learning* (pp. 2127–2136).

Information: what and how

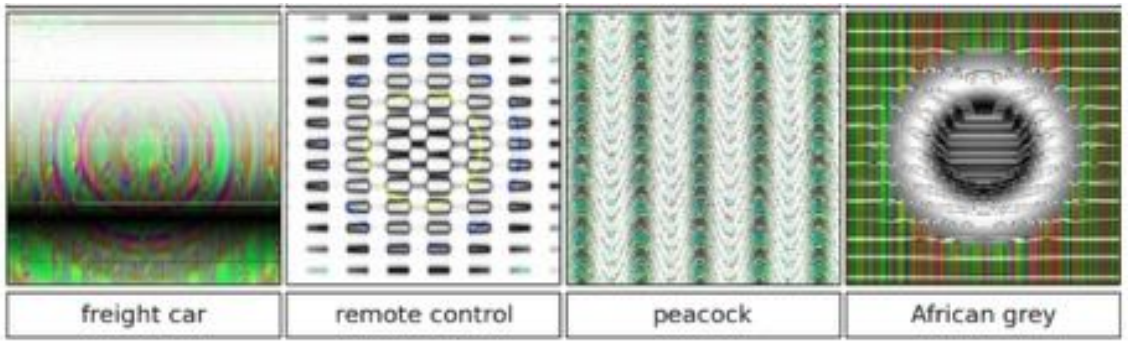


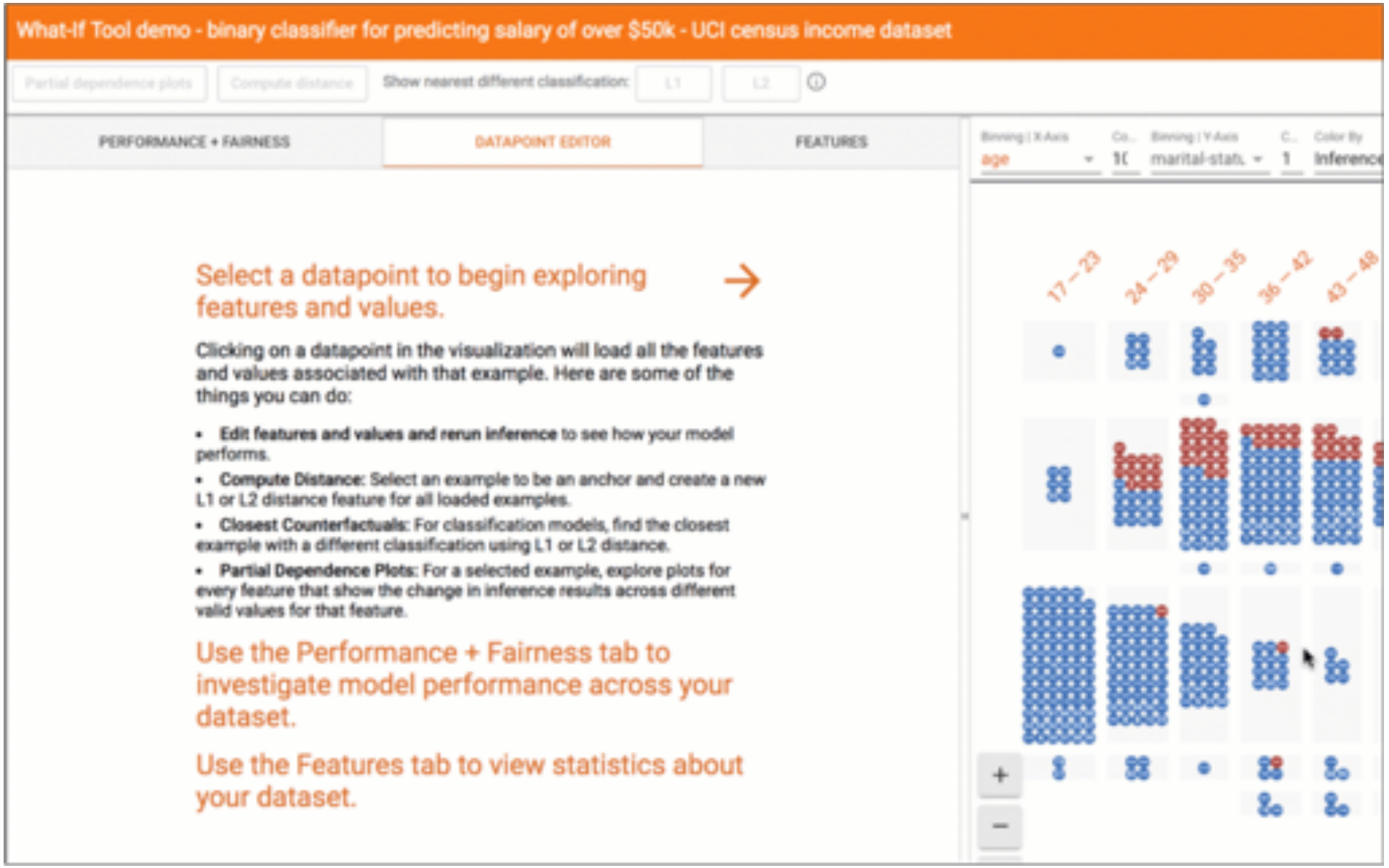
Figure 1. Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with $\geq$ 99.6% certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Images are either directly (*top*) or indirectly (*bottom*) encoded.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 427–436).

- High confidence of an obviously erroneous result
- Influence on people's perception

- Most works based on images

■ Google What-If Tool



Transparency and Standardization

Potential Items for International Standards

■ Transparency evaluation guidelines

- Data
- Business/Management model
- Technical framework

Transparency Risks

Security

- Transparency may make **models vulnerable to attack** by increasing the understanding regarding how the results are obtained (Papernot, McDaniel, Sinha, & Wellman, 2018).

Privacy

- Transparency could reveal private information. Rule disclosure may be prohibited by law if it involves private information. (Kroll et al. 2016; Ananny & Crawford 2018)
- Should consider what, how and to whom the information is revealed.

Intellectual property (Papernot, McDaniel, Sinha, & Wellman, 2018).

Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and Privacy in Machine Learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS P)* (pp. 399–414)

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989.

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. U. Pa. L. Rev., 165, 633.

Summary

KDDI Research

- Transparency is important for trustworthy ML/AI

- Interpretability is important for improving transparency in machine learning.
  - However, it does not equal transparency.

- For interpretability, future challenges are to
  - **Clarify scope and definitions**, which will help set measurable objectives.
  - **Consider stakeholders** beyond developer and conduct user evaluation. Without validation, it is difficult to say whether interpretability has been achieved.
  - What **information** to provide and how to provide it.

- Challenges for transparency in general
  - **Evaluate risks**, vulnerabilities
  - Standardization