Addressing the environmental challenges created by use of the latest AI models

Malcolm Mason

Sure Business

Where is machine learning used?

Automotive – assisted and fully automated driving, driver behaviour analytics

Automotive Insurance – Black box, Mobile and ODB2 insurance data evaluates drive behaviour. Patterns in data drives lower premiums

Life insurance – Wearable sensor data feeds back body vital signs, enabling predictive risks

Banking – The banking sector increasingly relies on machine learning for decision making

Healthcare – Genome decoding, cancer treatments, predictive sports injury, actual life expectancy

Financial Services – Converting unstructured data e.g. legal agreements and financial reports to structured data for analytics

Media and Entertainment – All user inputs are recorded and analysed across the internet, Sky, AmazonTV etc

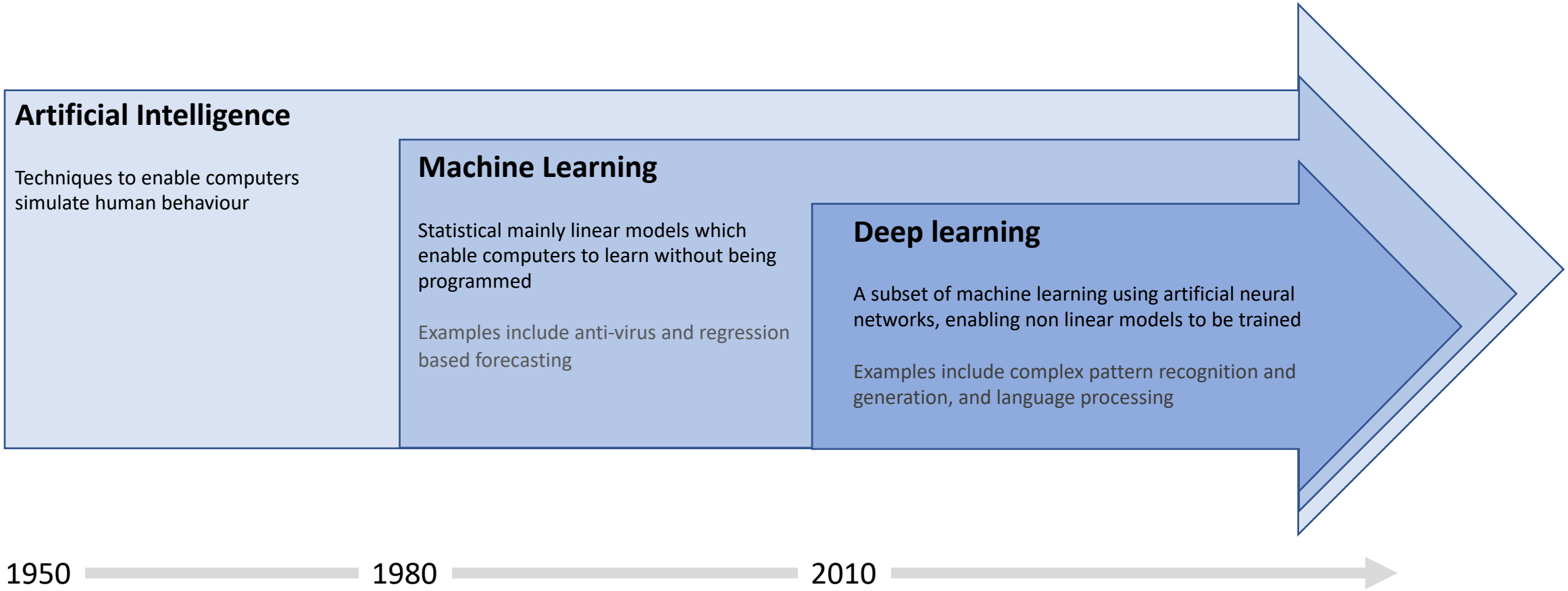
Weather – Prediction uses some of the largest computers in the world right now

Retail – Amazon claim to know what we want and deliver it before we realise we want it

Gaming – User interactivity analysed and available for sale

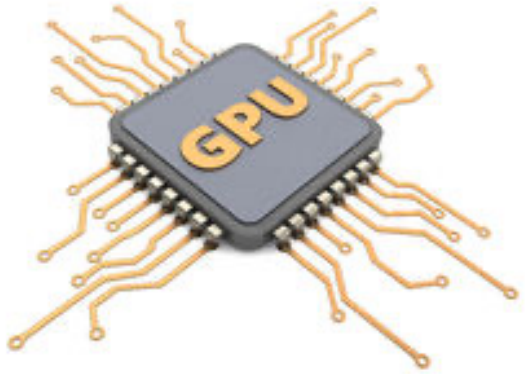
AI is now mainstream, and users rely on trained models daily without realising

Artificial Intelligence timeline



If you look beyond the marketing hype, this is now a very mature space

Why now?



Big Data (and lots of it)

Approx 64 zettabytes of data in 2020

(64,000,000,000,000,000,000,000 bytes)

Compute power

GPU performance doubling every year, enabling deep learning previously not possible

5 Petaflops in a single 3U off the shelf server now possible

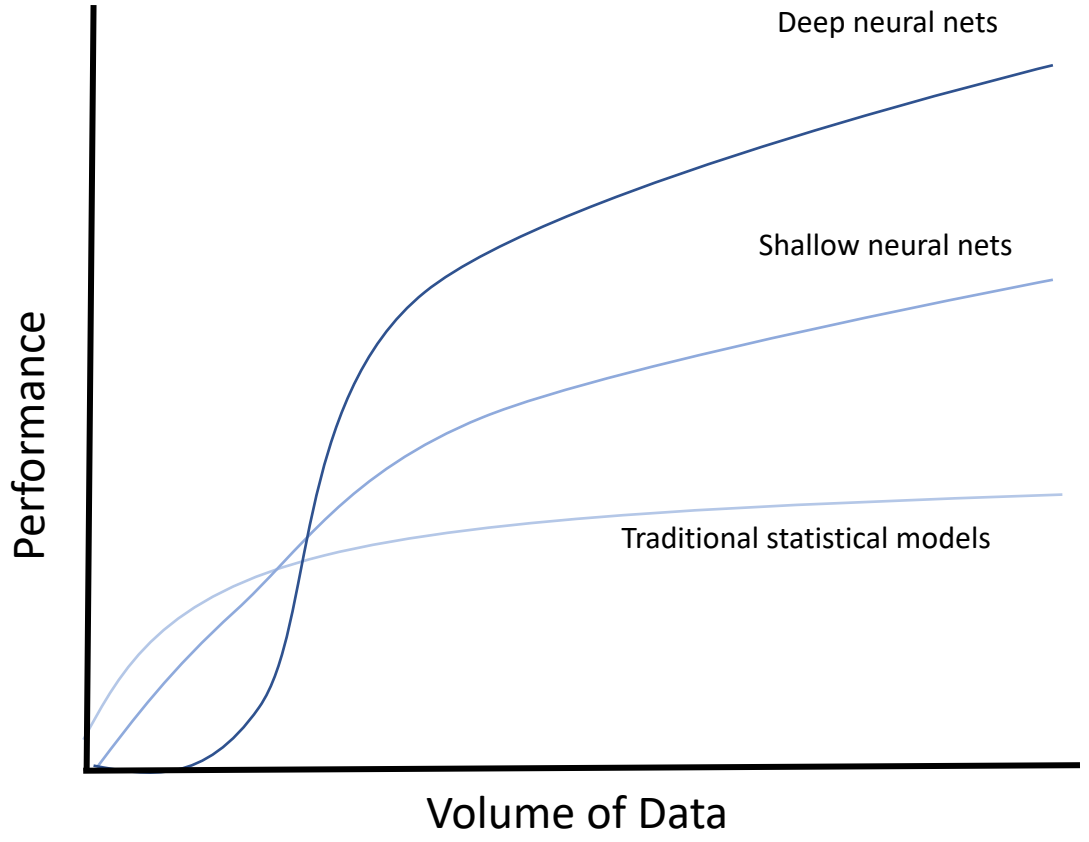
IoT (Connected devices)

Connected devices, from mobile devices, wearables, machines, automotive, entertainment and so on

Estimate 9.1bn by end of 2020

From a human element, what is driving the evolution?

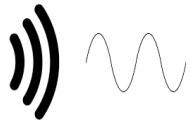
- Business has no choice if they want to compete and grow
- Multi-national companies such as Google and Facebook have open sourced their deep learning frameworks
- As soon as a revolutionary advancement is made, it is simplified so practically anybody can use it
- There are sample data sets for just about everything
- The capability to build, train and inference models is accessible everywhere from the home laptop to the hyperscale cloud providers
- There are education mechanisms for everything over the Internet, including Medium, YouTube, GitHub, Kahn Academy and many more



The ability to train non-linear deep learning models means any digitized data type can be used to drive business value

Data

3.01e-08
2.40e-08
4.44e-08
3.33e-08



Numeric | **Text** | **Images** | **Video** | **Audio** | **……** | **Any sensor data**

Numeric data can be in many forms including integer, floating point etc

Words are converted to numeric tensors, enabling each word in the vocabulary to maintain a relationship with all other words

Images are converted to any understandable format such as RGB, YCrCb, and then to floating point tensors

Video is the same as images, but time based

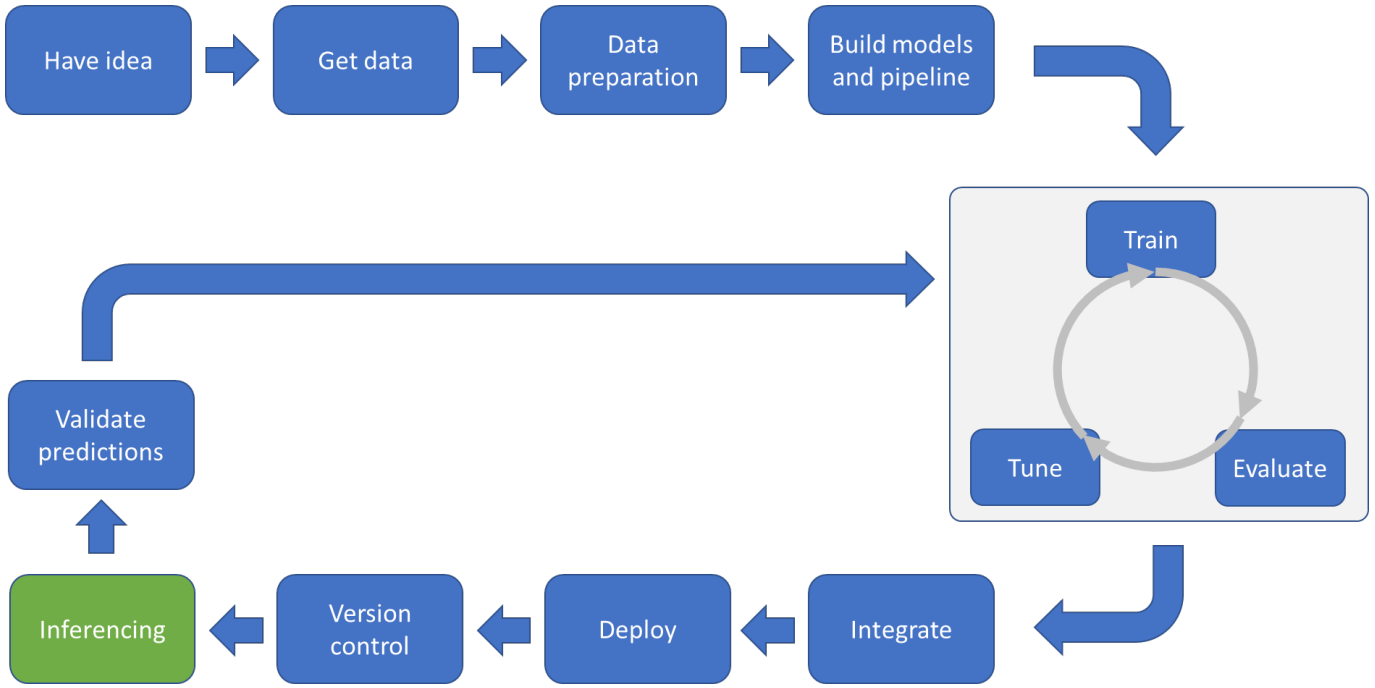
Audio is bit rate and sampling over time

Any sensor data to tensors

Features | **Features** | **Features** | **Features** | **Features** | **Features**

Model

Why are the latest strategies so inefficient?



- Models key to driving predictions once trained are built using experimentation
- State of the art non linear models take a huge amount computing power when training, evaluating, fine tuning regardless of where trained
- Driving optimum accuracy often means bigger is better, and state of the art means we are not only going deeper, but now layering models on models
- The current AI strategy failure is between 60% and 80%

For businesses, an AI strategy in present day is as important as the personal computer from 1980's onwards

The environmental impact of state of the art models

Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-GEN (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	–
2021	Switch-C [43]	1.57E+12	745GB

- AI deep learning models are doubling in size every 3 – 4 months
- Guesses estimate GPT-3 generated over 283 tonnes of CO2 during training – (approx. 5 family car lifetimes)
- Non-linear models demand state of the art deep learning to achieve useful results, meaning millions of tailored company AI strategies
- If 0.1% of all business trained a model the size of BERT, estimating doubling model size every 3 months, CO2 output from training would exceed current global CO2 output in next 5 years

How can the impact on the environment be offset - utilisation

Some examples where ensuring the carbon benefits when deploying AI solutions outweigh their environmental impact

Smart cities	• Training real time sensor data and modelling in a digital twin has the ability to enable massive operational efficiency in the real world
Financial Services	• Using natural language processing with contractual data enabling immediate answers to queries, document summarisation, translation, markup sentiment analysis and more, enables users and IT to work more efficiently
Healthcare	• The use of AI in healthcare enables efficiency in countless areas through automating tasks and huge data set analysis, driving significant efficiencies in care services. • For the vulnerable, combined with additional real time sensor data, it can enable care to be delivered faster and more accurately
Technology	• Predictions in IT behaviour ensures IT energy consumption can be scaled to meet business and user requirements

Consider any area using technology where people can use IT smarter, faster and greener to achieve their aims

How can the impact on the environment be offset - implementation?

Efficiencies when implementing AI solutions

Starting with clear realistic goals

- Having a well researched strategy with clear outcomes to the business enables focus to complete the task in hand and prevents waste through failure. Scope creep and lack of business understanding is a major contributor to AI strategy failure

Pre-trained model re-use

- Hyperscale IT service providers deliver hugely accurate models to meet a wealth of tasks, which are often open source. This enables businesses to create state of the art solutions through re-using models and fine tuning for their own needs

Measure and track everything

- The latest state of the art deep learning models require experimentation. Measuring outcomes when hyperparameter and model tuning drives faster results than flying blind

Don’t reinvent the wheel

- Millions of data engineers and data scientists are pushing the boundaries constantly to push the art of the possible. Learn from their success.

Conclusion

AI is now mainstream, and is a necessity to enable businesses to compete

More data, more compute and more IoT devices demands more compute power and larger models to drive performance

Businesses should be encouraged to consider the overall carbon footprint pre and post AI deployment, both through implementation and their impact on the task they are aiming to solve

When used effectively, the use of AI has the capability to significantly reduce carbon emissions for our planet