

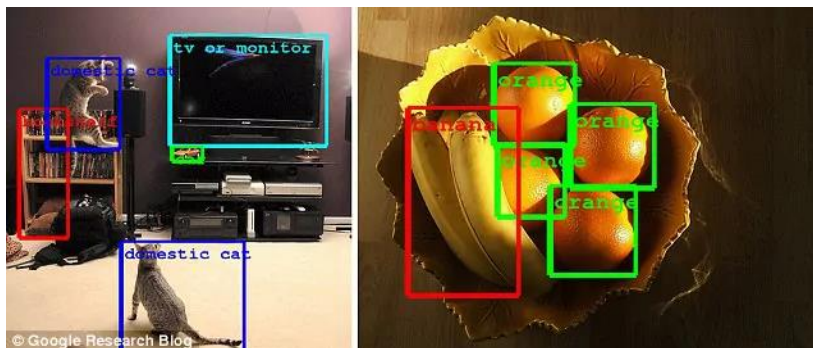


An introduction to Machine Learning

Dr Andrew Johnson
gyamj@leeds.ac.uk



UNIVERSITY OF LEEDS

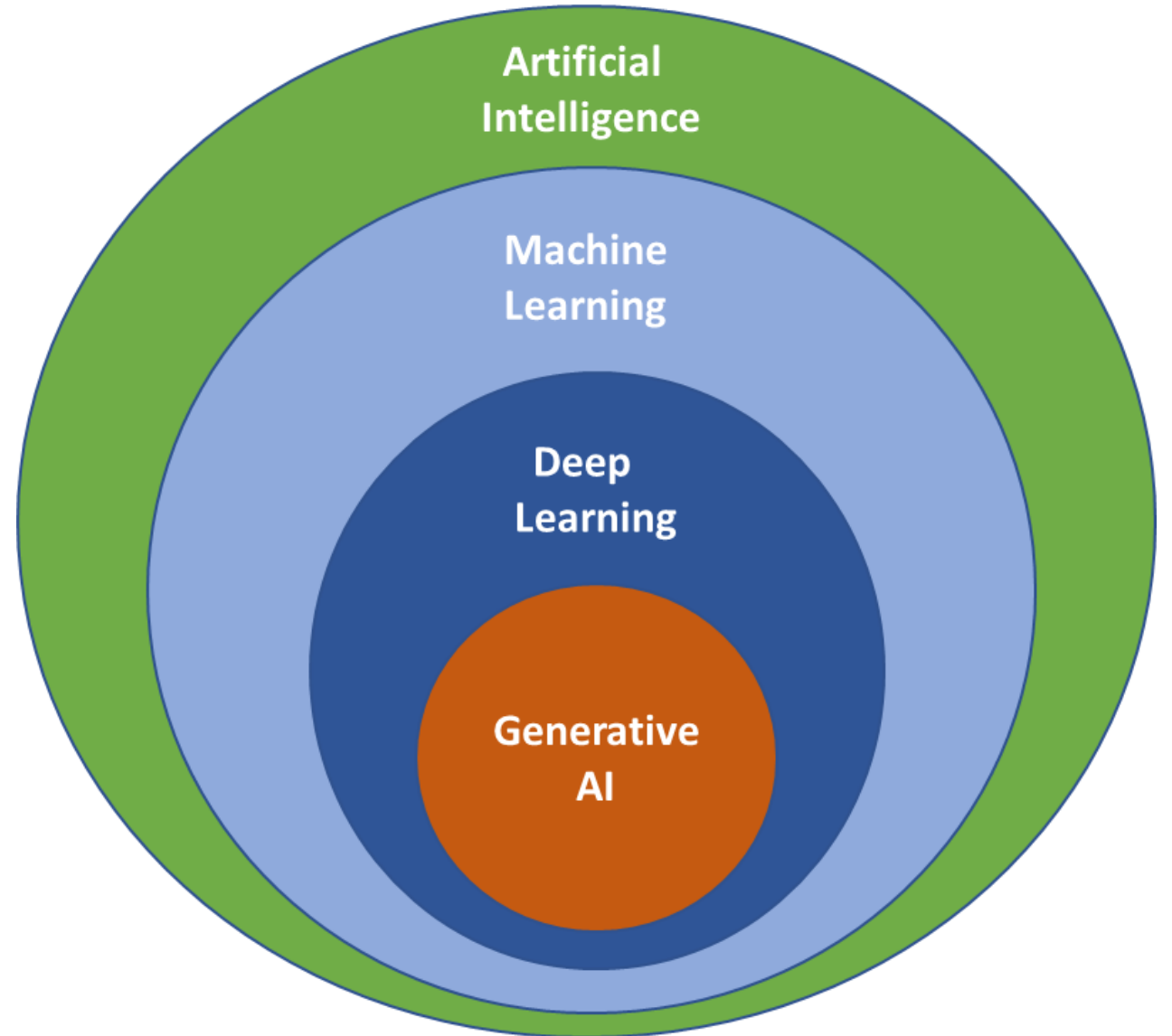


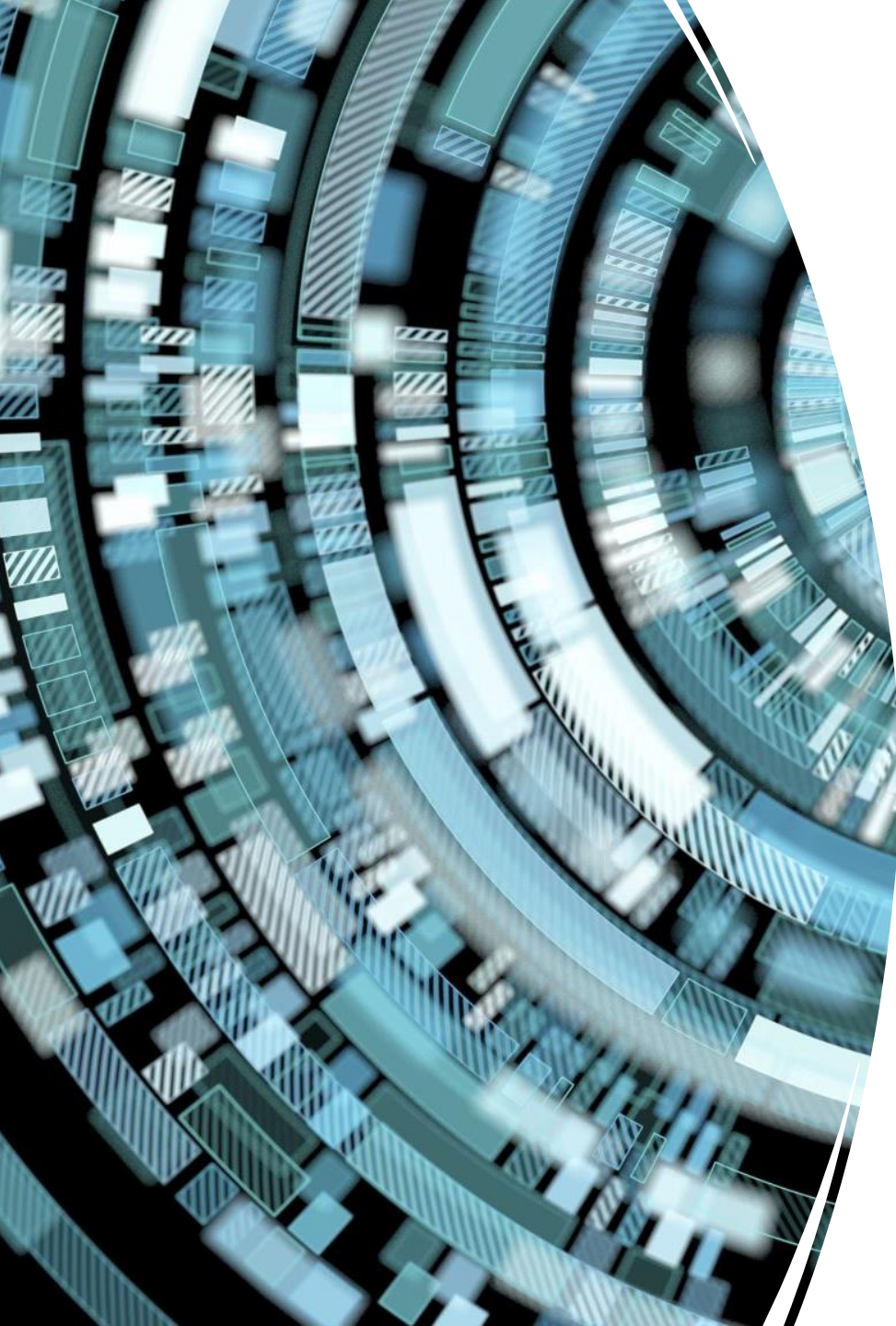
AlphaGo

ChatGPT

Definitions

- Artificial intelligence: making computers do tasks that traditionally requires human intelligence
- Machine learning: giving computers the ability to learn based on pre-specified rules
- Deep learning: a subset of machine learning using advanced algorithms that replicate how humans learn
- Generative AI: deep learning models that are powered by large foundation models





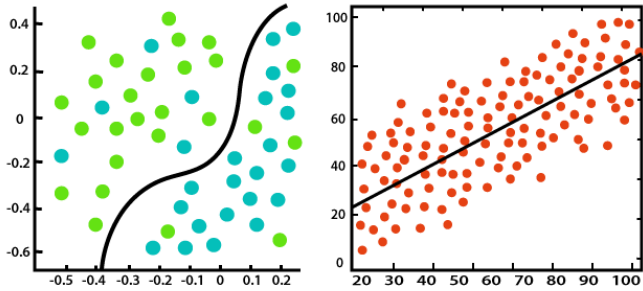
What is machine learning?

- Machine learning is a type of artificial intelligence where computers learn from experience and data to make predictions or take actions
- A computer uses predefined 'algorithms' to learn
- Tasks are often learning relationships from data, making predictions, or finding structure in data
- Common themes:
 - What the model is learning 'from'
 - How the model is learning
 - The structure of the task

Types of machine learning

Supervised learning

Task driven models
Regression or classification

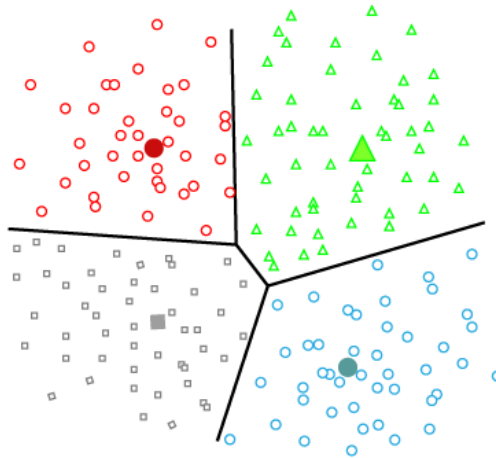


Classification

Regression

Unsupervised learning

Data driven models
Clustering or ordination



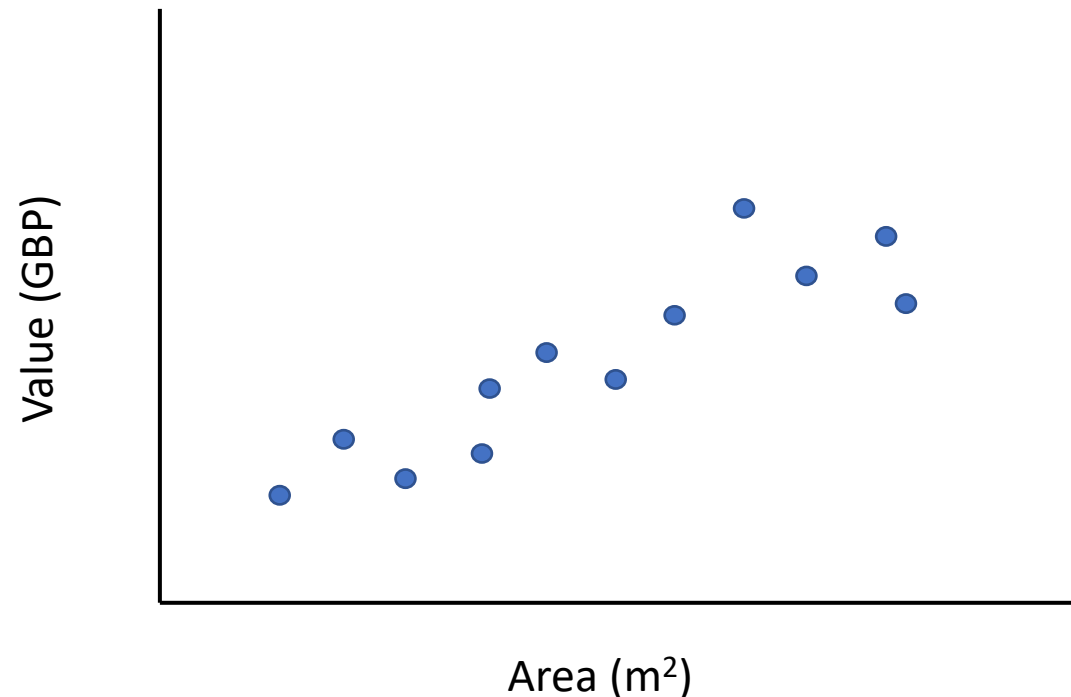
Reinforcement learning

Learning from mistakes
e.g. Playing games



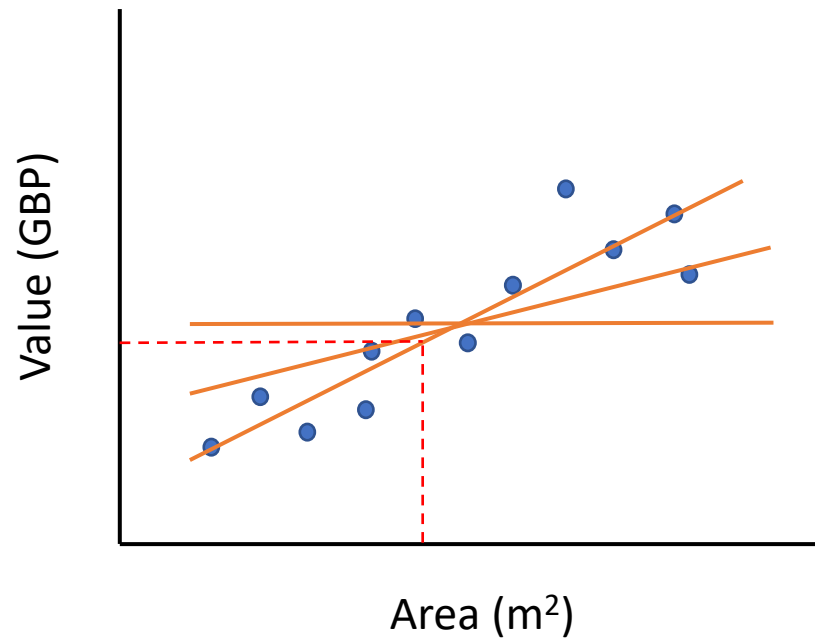
How is the value of a property influenced by area?

- Task based model
 - Describe the value of a house based on it's area
 - Supervised machine learning
 - Regression model
- As area increases the value also increases
- Different types of regression models could be used
 - The shape of the relationship
 - The structure of the data

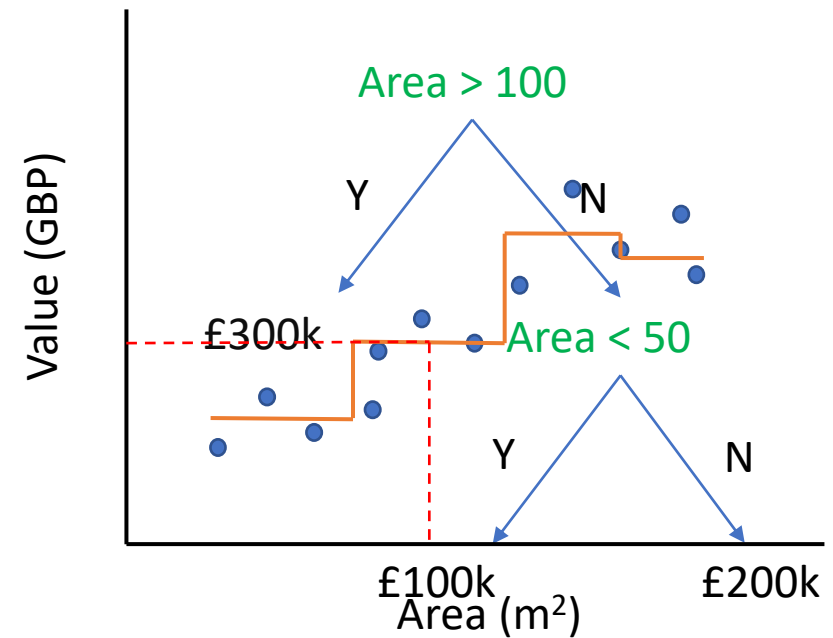


Different models can do the same task

Linear model



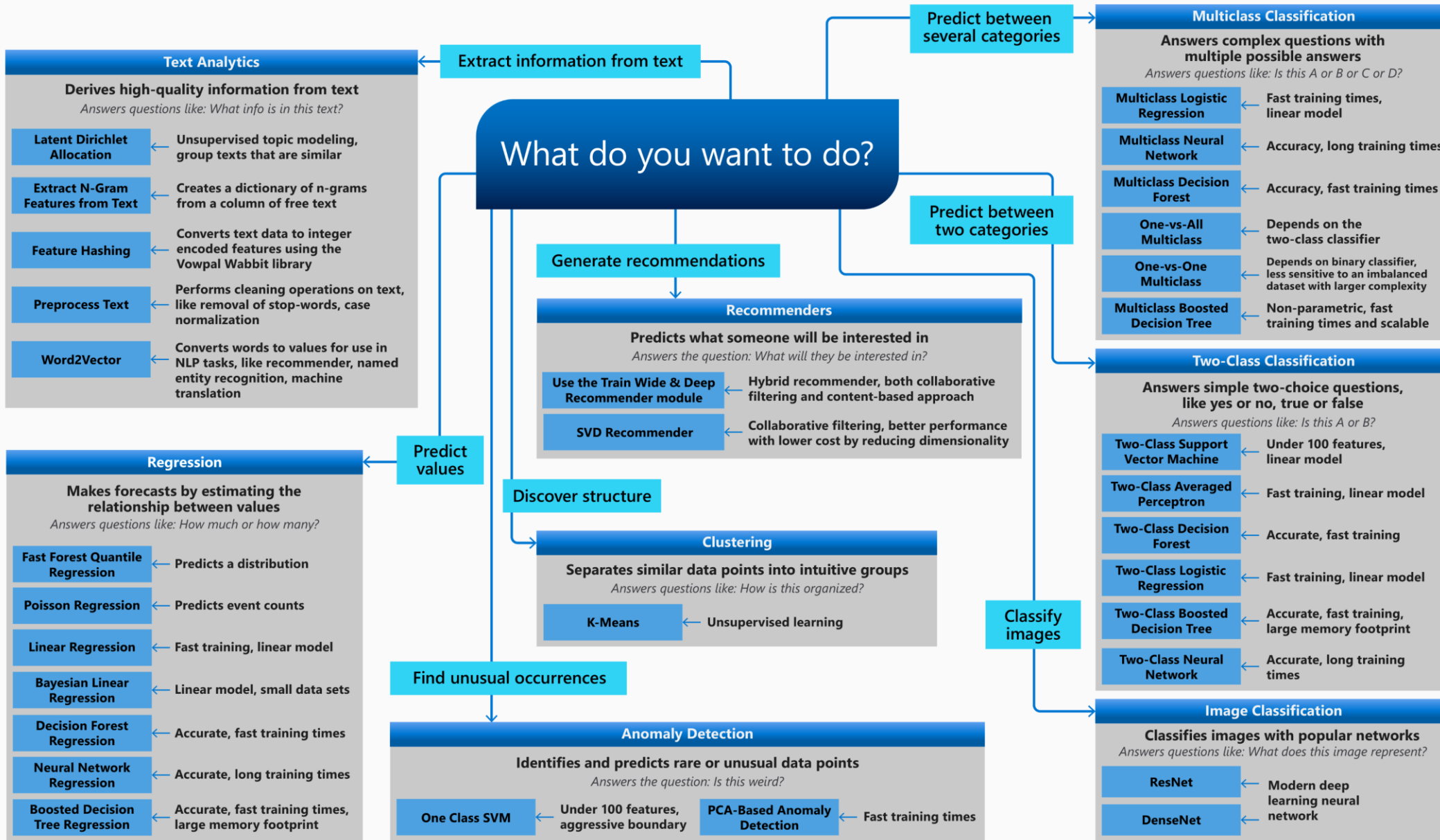
Decision tree





Machine Learning Algorithm Cheat Sheet

This cheat sheet helps you choose the best machine learning algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the goal you want to achieve with your data.



Machine learning and freshwater

Data collection and accessibility

- Automatic data tidying
- Data reconstruction

Empower local communities and citizen science

- Automatic detection and tracing of pollution events
- Event forecasting

Ecological modelling

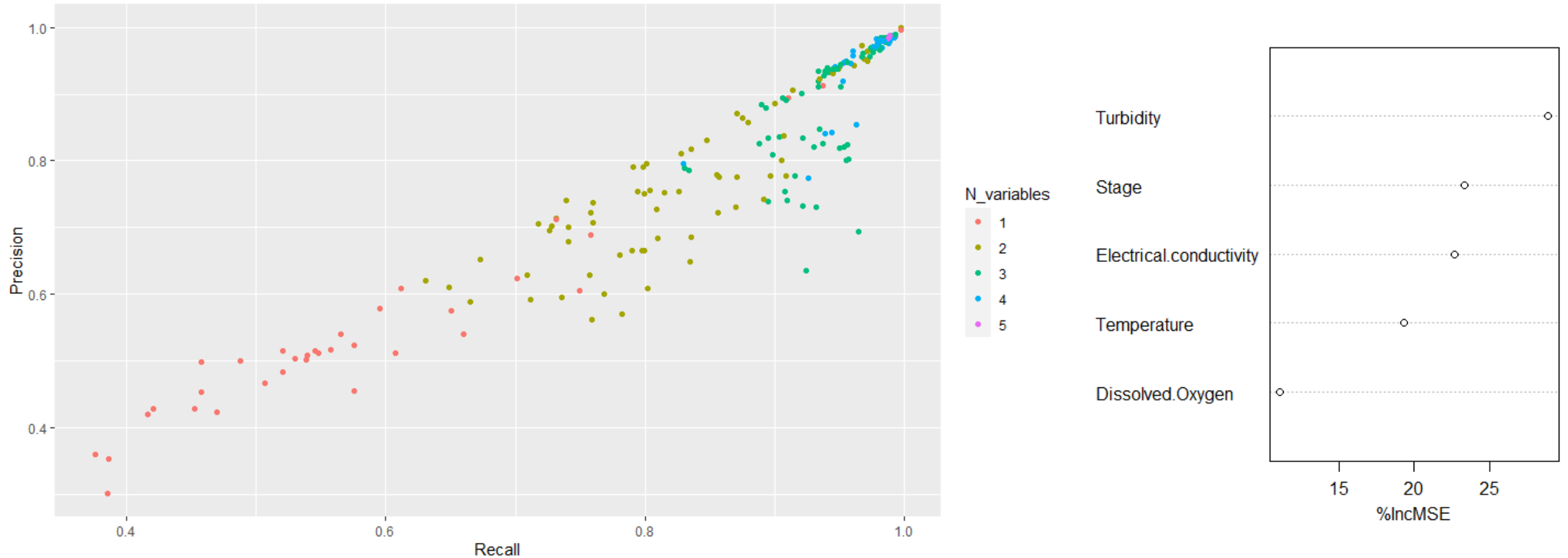
- Predicting the response of communities to disturbances
- Modelling responses to climate change

Detection of pollution events

- Can we distinguish between different types of pollution events?
 - Task driven model
 - Supervised classification model
- How do we build the model:
 - Historical labelled data of pollution events
 - Use water quality variables from sensors
 - Random forest: supervised classification and regression technique
- Real-time implementation:
 - Anomalous reading detected
 - Trained model identifies the event
 - Send out alert to citizen scientists



Distinguishing between anomalies



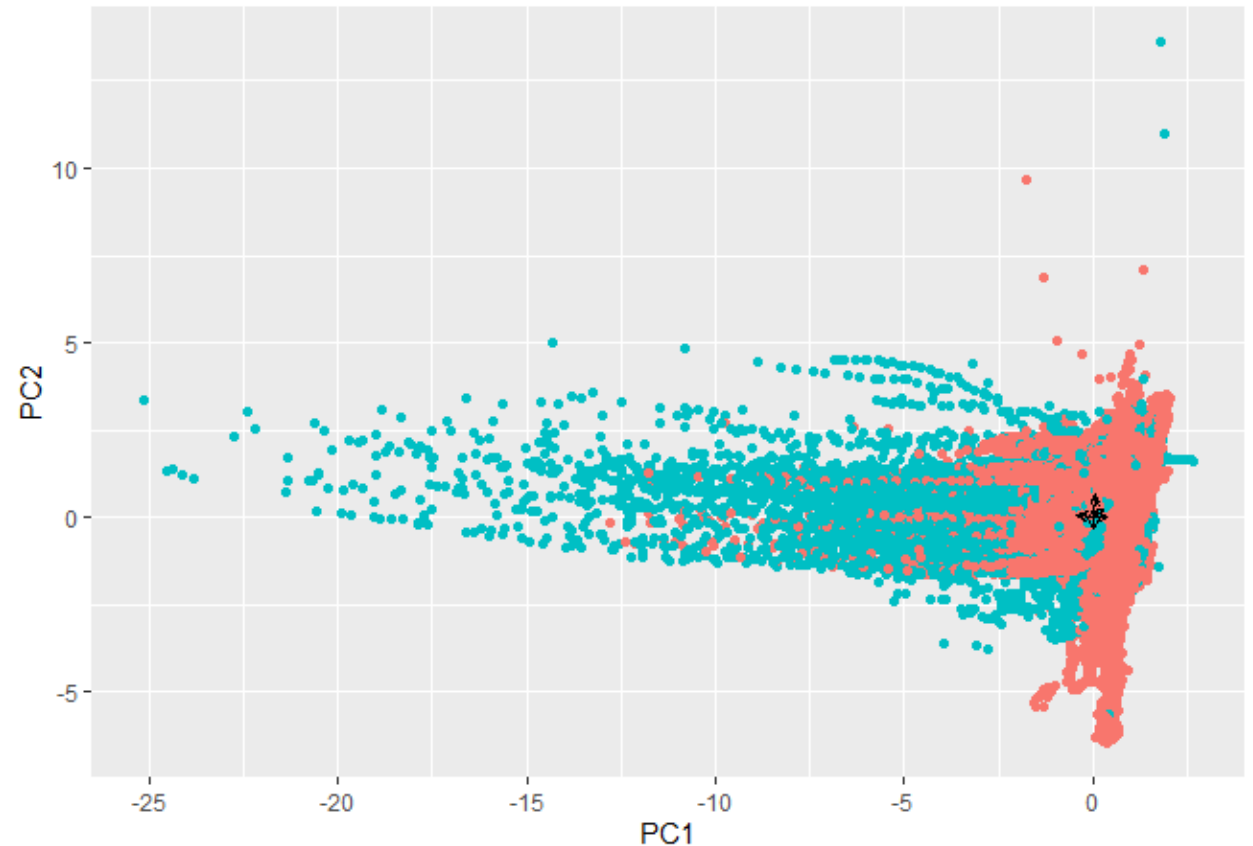


Data reconstruction

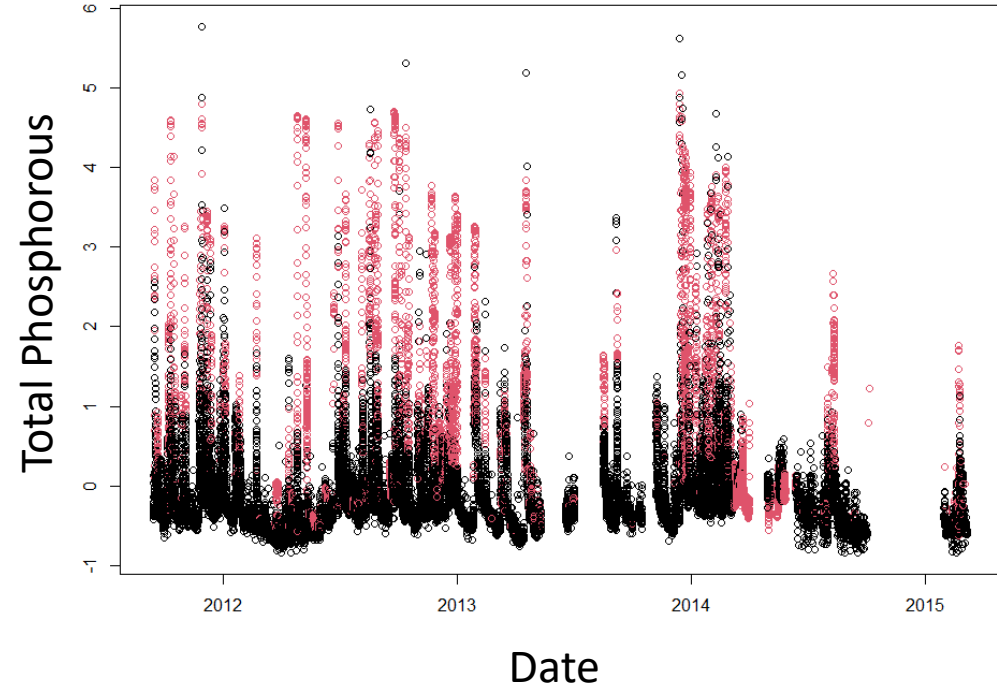
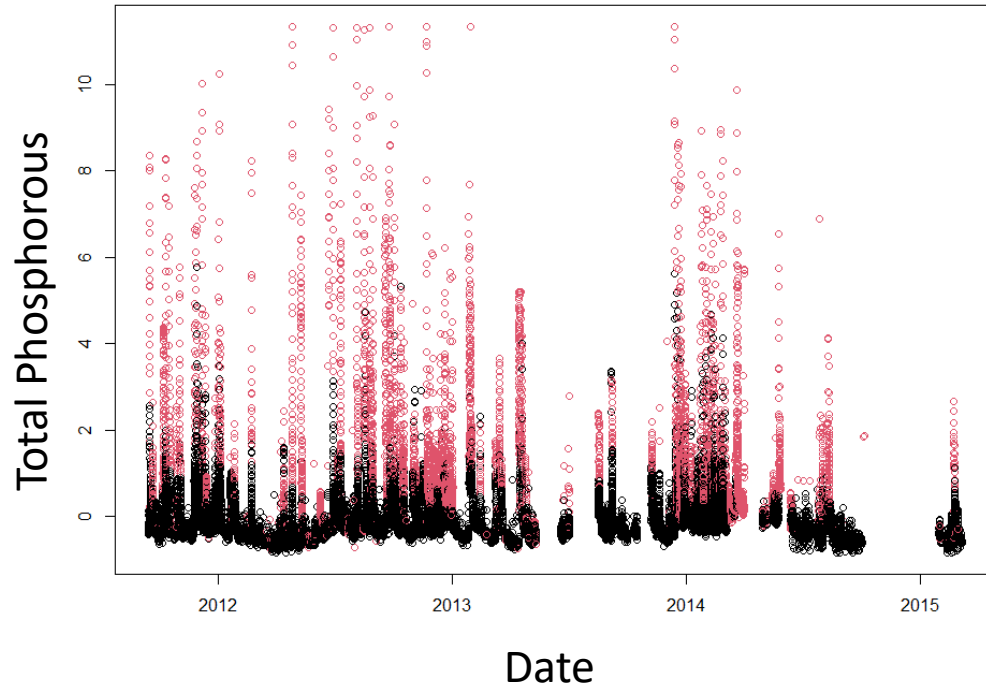
- Can we reconstruct data from malfunctioning sensors?
 - Task driven model
 - Supervised regression model
- How do we build the model?
 - Historical water quality data
 - Train the model on labelled normal data
 - Predict the malfunctioning variable
- Real-time implementation
 - The previous model detects a broken sensor
 - Predict what the broken sensor values should be
 - Send an alert to get the sensor repaired

Case study for data reconstruction

- Anomalous readings from a phosphate probe in the Eden DTC
- Models were trained on normal data
- Predict phosphate from other known variables at the same sensor



Time-series reconstruction



A satellite-style image showing a complex network of rivers and streams flowing through a lush green landscape. The rivers are dark blue/black, contrasting with the vibrant green of the surrounding vegetation. The image is partially cut off on the right side.

Remote sensing of river habitat quality

- Can we determine water quality from remote sensing data?
 - Task driven model
 - Supervised regression or classification model
- Remote sensing often uses image classification
 - Historical satellite images
 - Label with water quality data
 - Convolutional Neural Network
- Merging and combining data allows us to answer more complex questions in novel manners

Links

Training courses

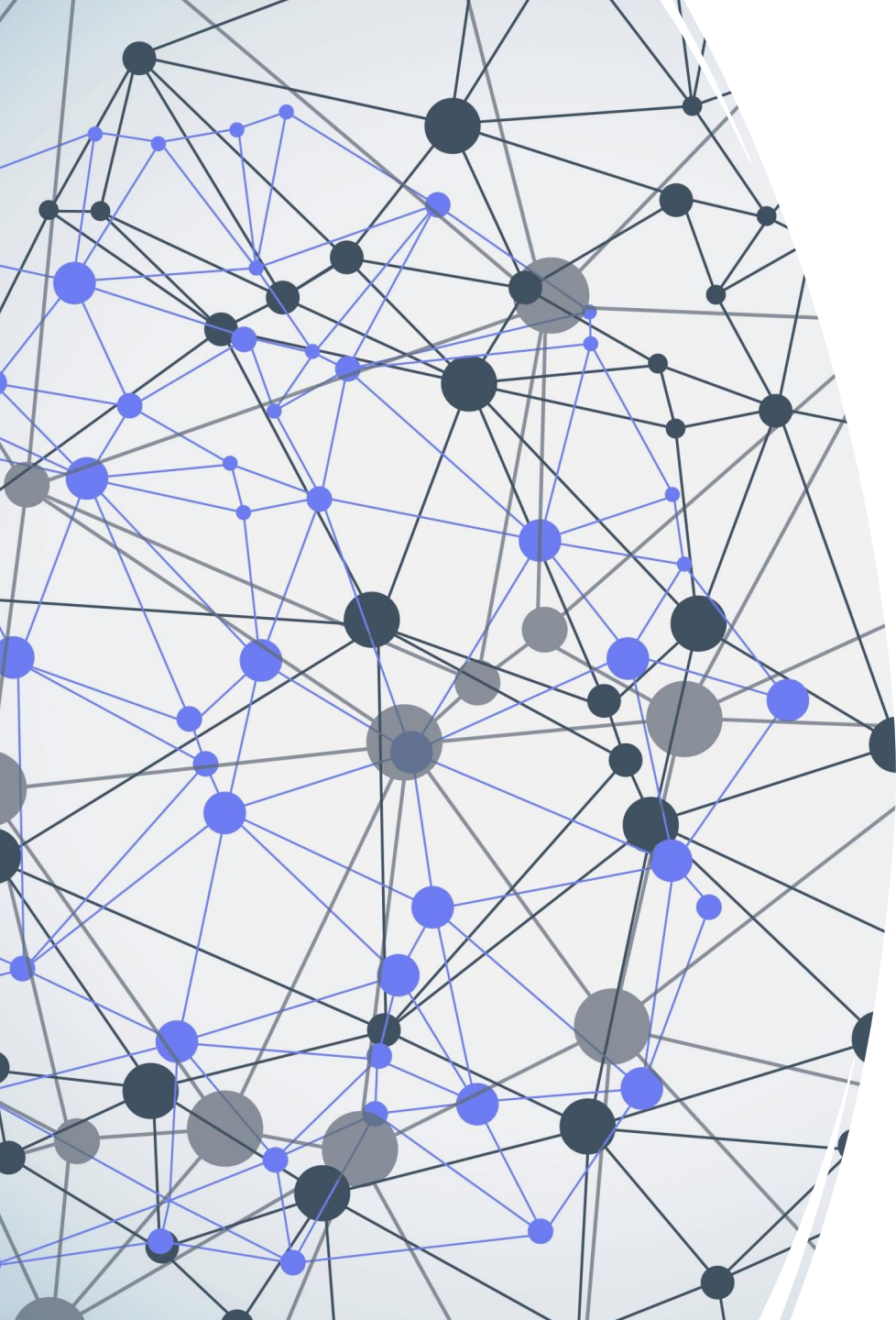
- <https://www.datacamp.com/tracks/understanding-data-topics>
- <https://www.coursera.org/specializations/machine-learning-introduction>
- <https://pll.harvard.edu/course/cs50-introduction-computer-science?delta=0>

Cheat sheets

- <https://learn.microsoft.com/en-us/azure/machine-learning/algorithm-cheat-sheet>
- <https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-supervised-learning>
- <https://medium.com/machine-learning-in-practice/cheat-sheet-of-machine-learning-and-python-and-math-cheat-sheets-a4afe4e791b6>

Videos and reading

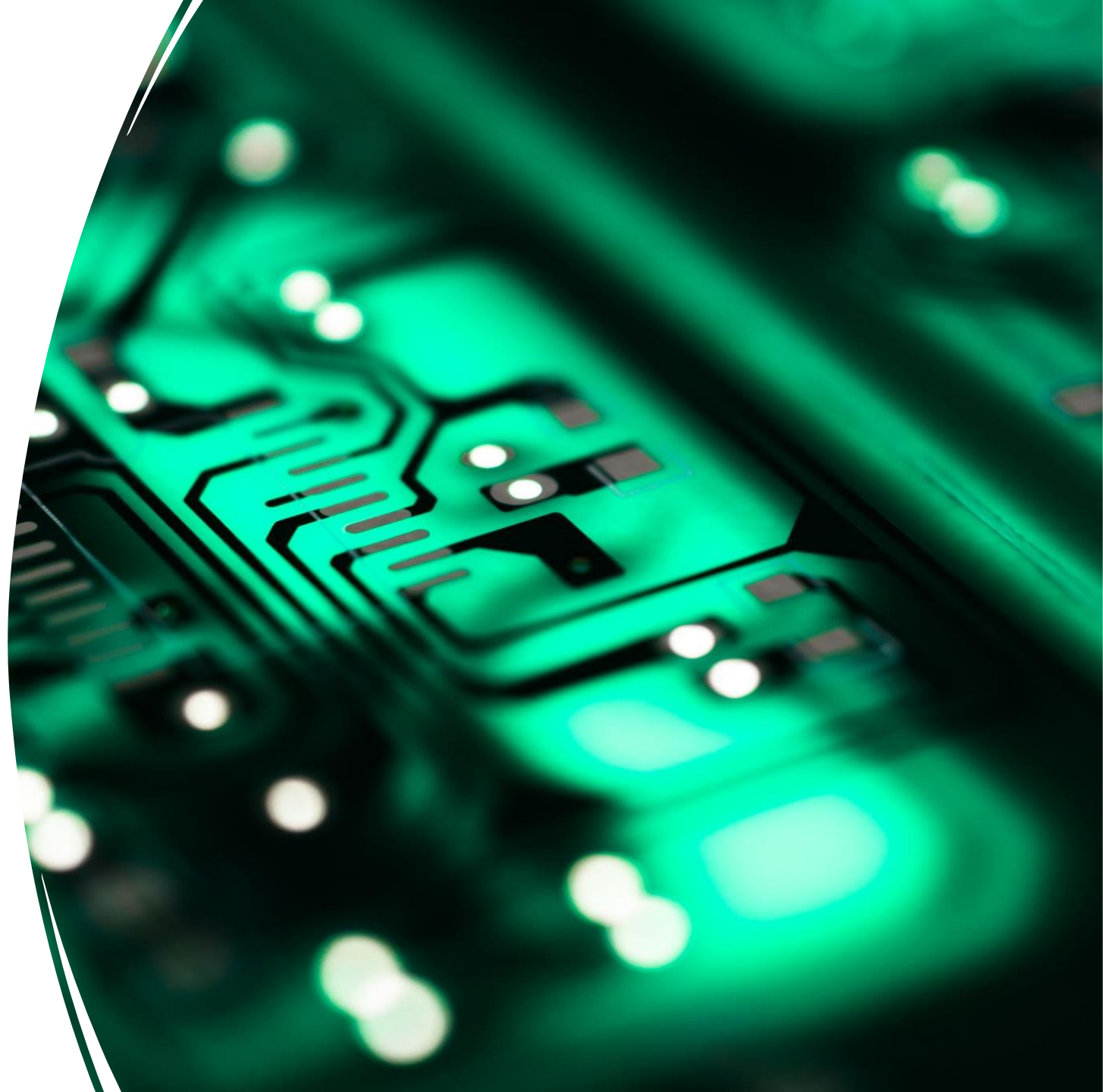
- <https://www.youtube.com/@statquest>
- Fast.ai
- <https://medium.com/tag/machine-learning>



An overview of Generative AI

What is Generative AI?

- Sophisticated algorithms that are capable of generating new content based on prompts or images
- Largely built on supervised machine learning algorithms
- New methodologies and techniques have resulted in a rapid improvement in capabilities



Background knowledge

What are neural networks
and how are they used?

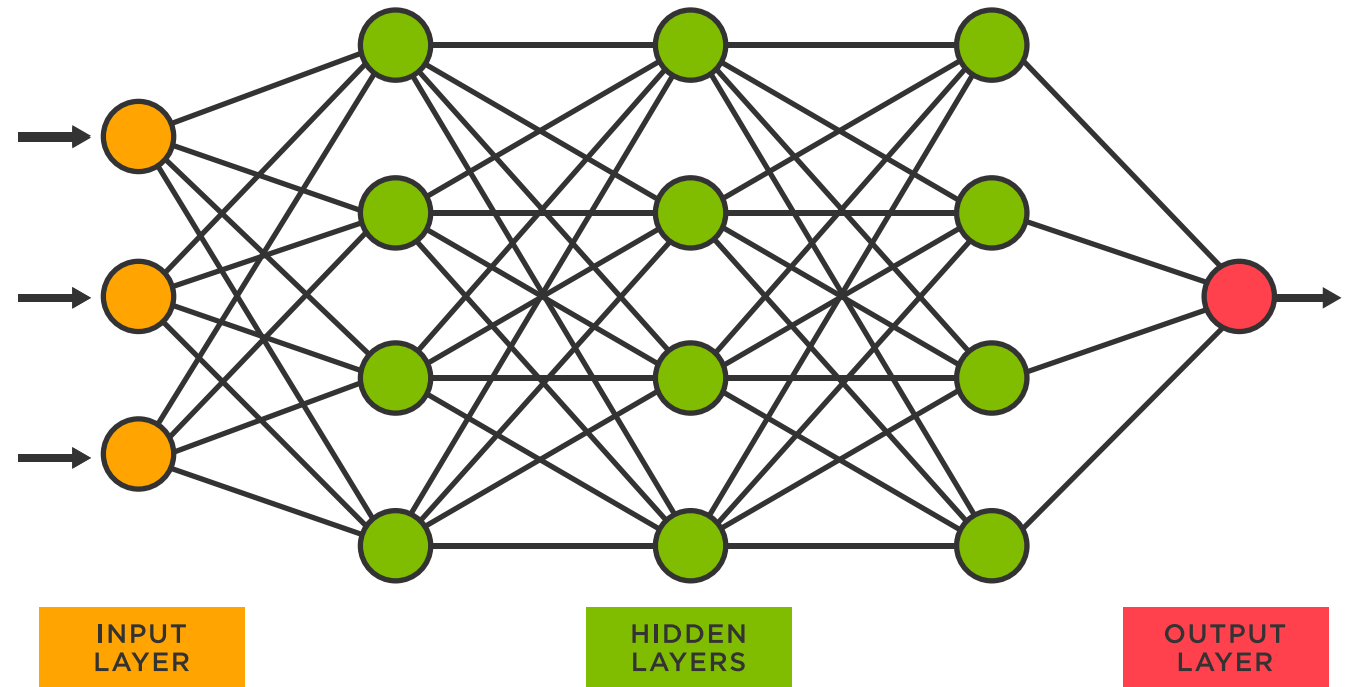
The use of foundation
models in AI

Advantages and Risks of
Generative AI

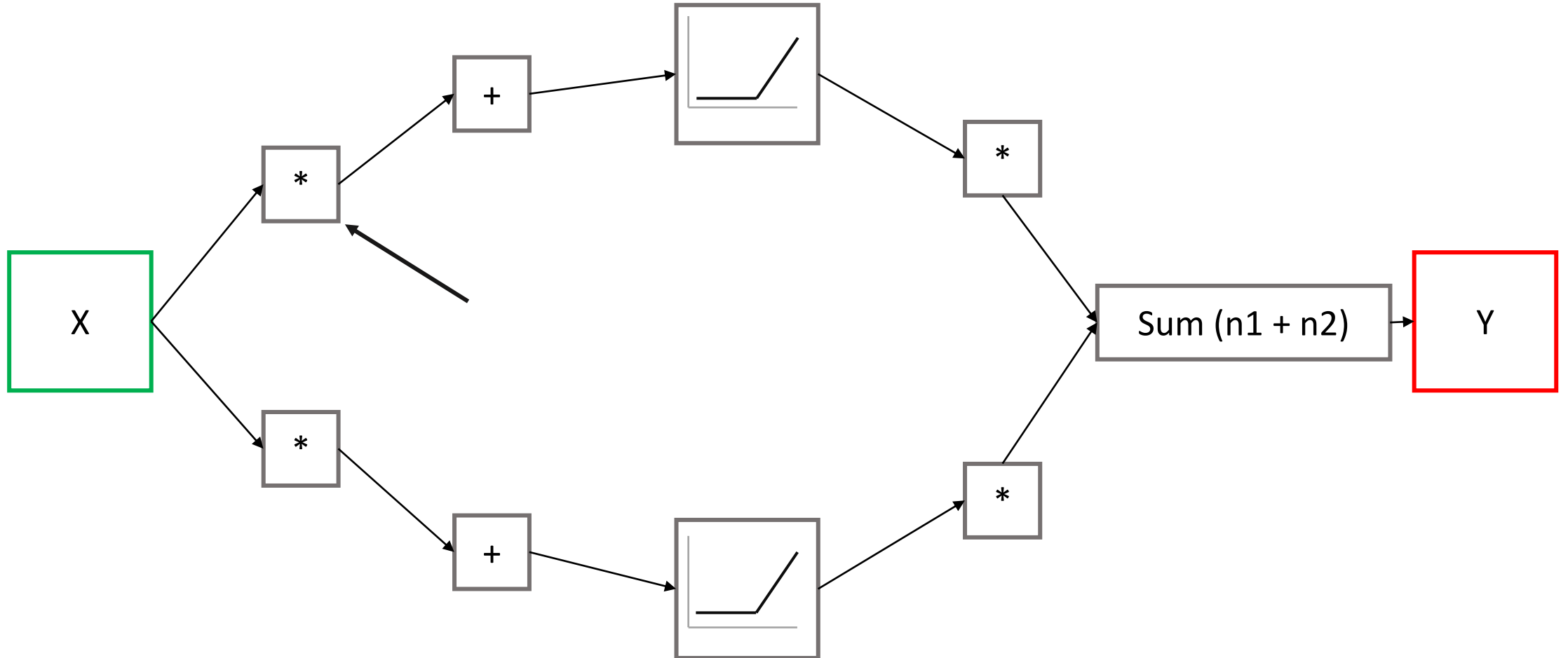
Generative AI in the
freshwater environment

Neural Networks

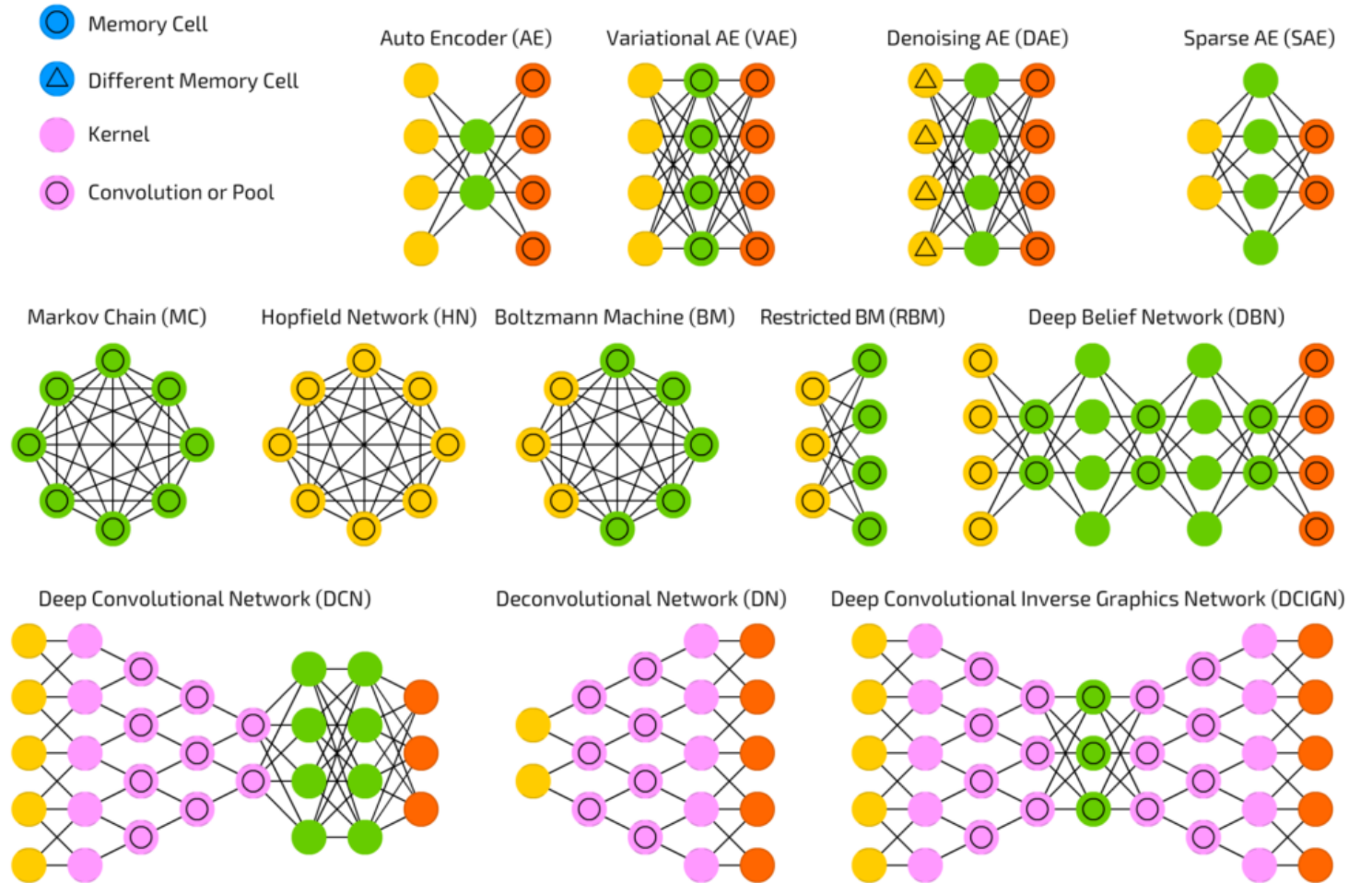
- Neural networks are a form of supervised machine learning
- Independent variables are fed into a series of hidden layers
- Calculations occur in the hidden layers to relate the independent variables to the dependent
- There are different structures of neural networks for different types of data



Inside the black box



Neural networks for different data



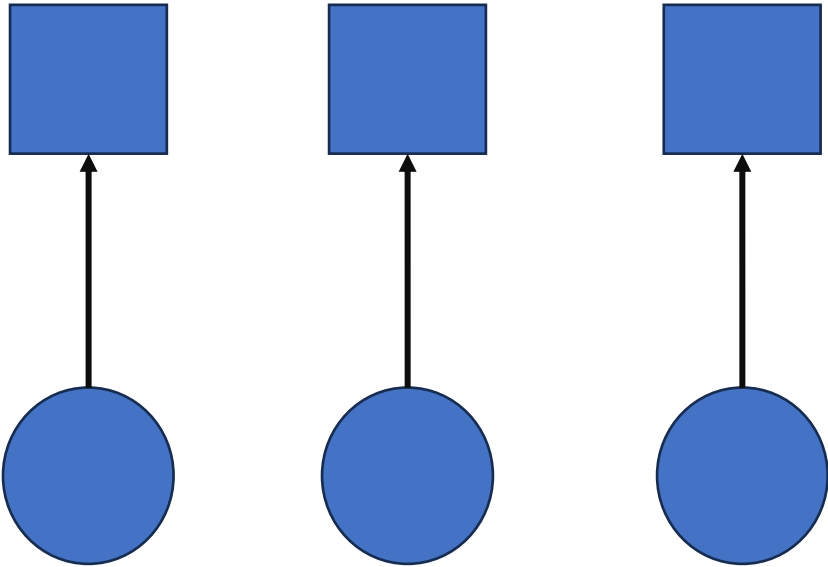


Key features

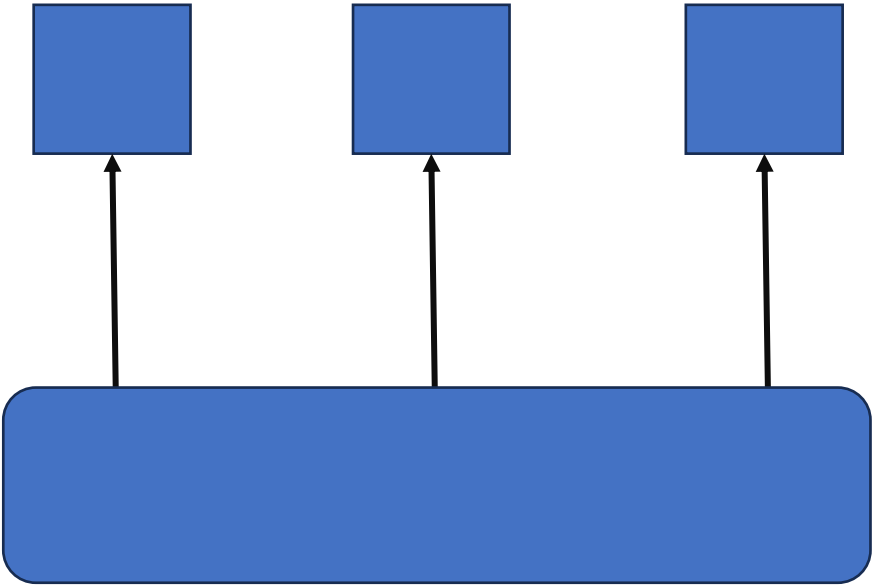
- Flexibility and adaptability
- Large neural networks are more powerful than smaller networks
 - Better control of overfitting
 - Emergent properties
- Supervised machine learning
 - Regression or classification
 - Gradient descent algorithm

The development of foundation models

Traditional data science



Foundation models



The development of foundation models



Train a very large neural network on massive datasets to perform non-specific tasks

Predict the next word of a sentence based on previous words

Predict the caption of an image based on the content



Tune the foundation model to a new task

Large Language Models

Identify pollution in photos of rivers

The quick brown fox jumps over the lazy **dog**

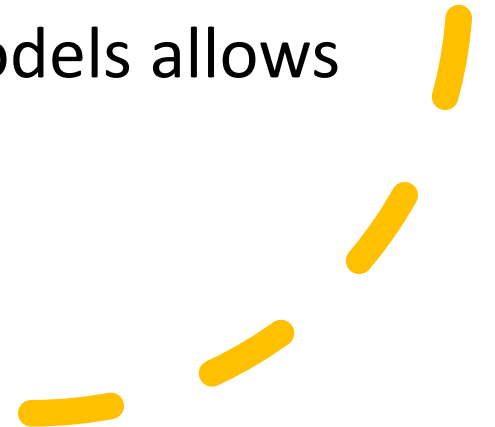
Let the cat out of the **bag**

Which way to the **library?**

What actually is machine **learning?**

Foundation models and generative AI


- Generative AI is powered by foundation models
 - GPT-4 is used to create ChatGPT
 - Microsoft Azure AI vision suite uses the Florence foundation model
 - DALL.E 2 can generate novel art
- Generative AI is a form of supervised machine learning
- The large scale of foundation models allows emergent properties to develop





Advantages Generative AI

- Drawing on large datasets allows more accuracy and emergent capabilities
- Foundation models are relatively fast to build
- Models can be fine tuned for specific problems
- Novel architectures are continually being developed and datasets always increasing



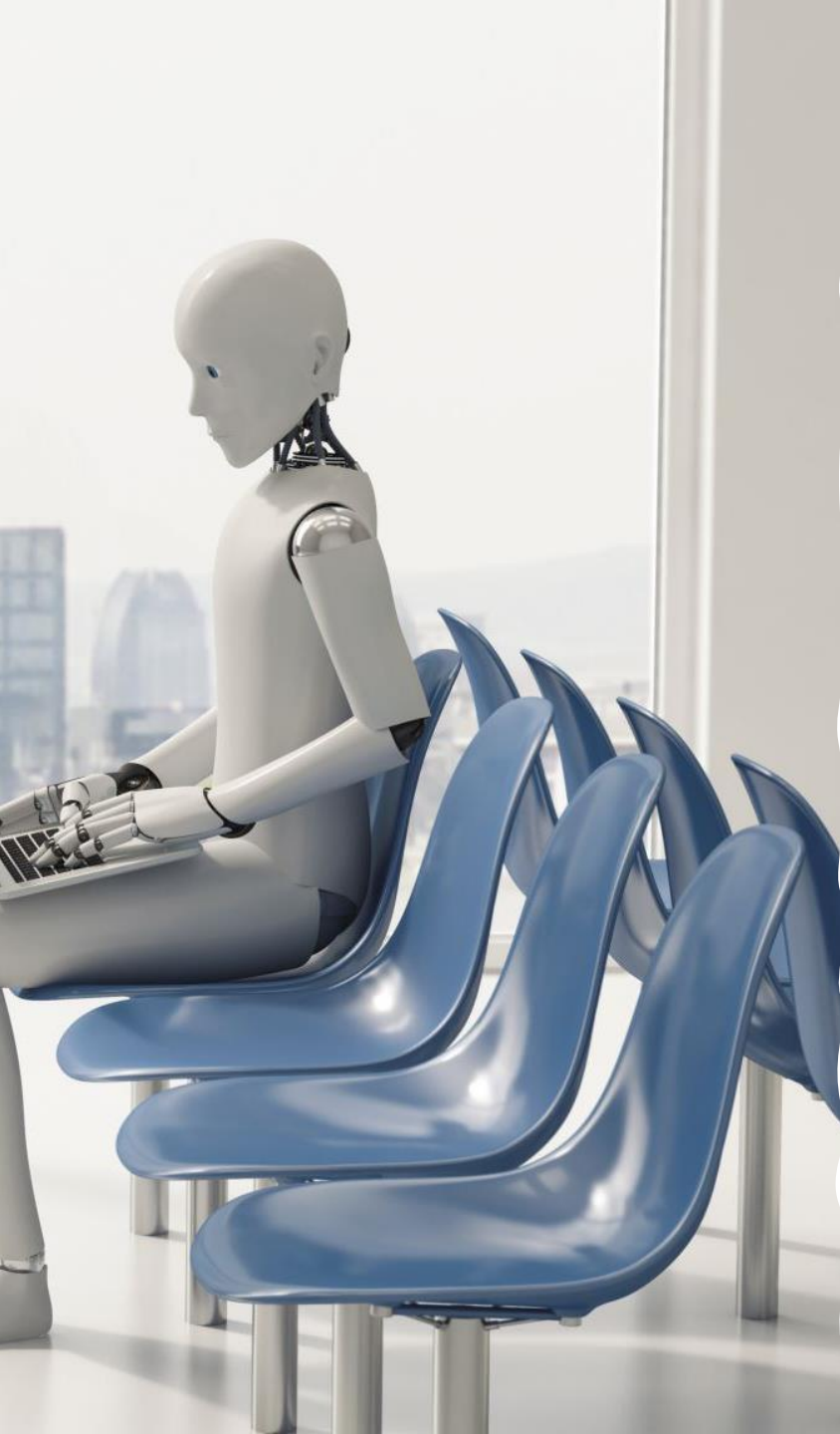
Limitations of Generative AI

- Bias can be introduced based on the quality and quantity of the training data
- Models lack true understanding and simply predict the most likely option
- The size of foundation models results in heavy energy and financial costs
- Ethical and social concerns regarding job security and fake data

Generative AI and foundation models in the freshwater environment

- Computer vision-based tasks:
 - Identifying pollution from citizen science-based images
 - Monitoring river habitat quality
- Semantic analysis of surveys
 - Measure the tone and meaning of written survey answers
- Modelling ecological trends over time
- A large challenge in freshwater is the amount and distribution of data





Closing remarks

- Generative AI is a powerful subset of machine learning
- Models can be used to speed up work and improve our analysis of small datasets
- Computer vision-based problems are a prime candidate in freshwater
- Generative AI and foundation models are expensive, use the smallest model you can



Any questions?