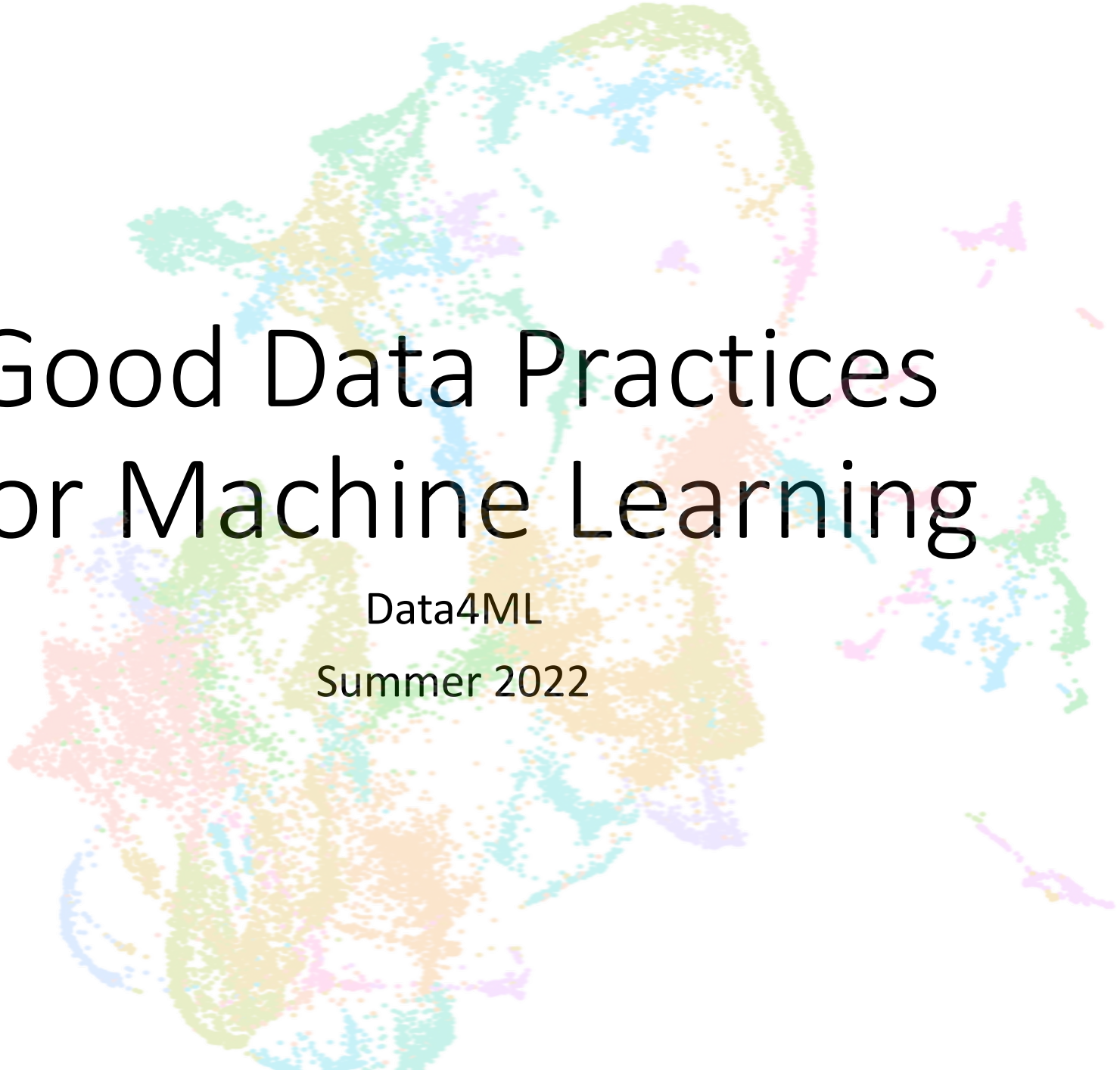


Good Data Practices For Machine Learning

Data4ML

Summer 2022

Ask an AI

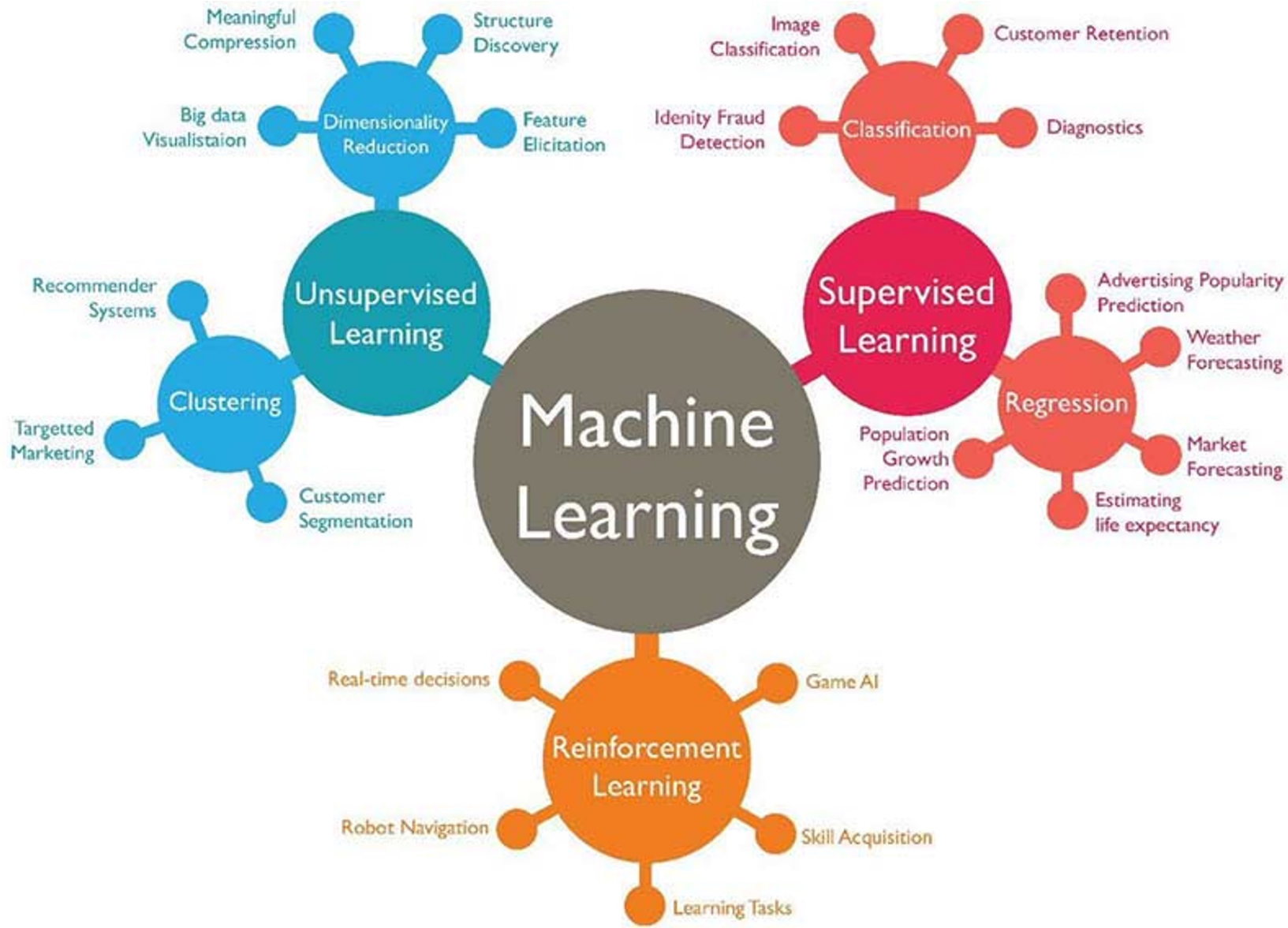
The most important thing to review at the beginning of a machine learning workshop is the data that you will be working with. This is because the data will determine what kind of machine learning algorithms can be used, and how the data should be prepared for use with those algorithms.

The most important thing to consider when using sequencing data for machine learning projects is the quality of the data. Poor quality data can lead to inaccurate results and poor performance from machine learning models.

When working with sequencing data, it is important to check the quality of the data before using it for machine learning. There are a number of ways to assess data quality, including looking at the overall sequence quality, assessing the quality of the reads, and checking for any biases in the data.

Once the quality of the data has been assessed, it is important to pre-process the data to ensure that it is ready for use in machine learning. This may involve quality control steps, such as removing low-quality reads, trimming adapters, and normalizing the data.

After the data has been pre-processed, it is ready to be used in machine learning. There are a variety of machine learning algorithms that can be used with sequencing data, including classification, regression, and clustering algorithms.


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

ML Experiment

1) Define your problem
2) Gather preliminary or existing data
3) Fit a test model
4) Use a learning curve fit to estimate how much more data you might need
5) Use importance or attribution to look for confounders that might affect your data
6) Gather data and experiment with new models
7) Deploy/Publish – Return to 6

Data Concepts

- Good Machine Learning starts with good datasets
- What makes a dataset ‘good’
  - Reproducible
  - ‘FAIR’ – Finable Accessible Interoperable and Reusable
  - AI-Ready
- How to plan out your data taking
- What makes a data set ‘AI-Ready’

AI-Ready Data

- Clear Data Splits
  - Designate some examples as training data
  - Designate some examples as testing data
- Structed Data is ‘Tidy’ - https://vita.had.co.nz/papers/tidy-data.pdf
  - Every column is a variable.
  - Every row is an observation.
  - Every cell is a single value
- The Dataset is Clean
  - No missing values or ‘Nans’
  - No ‘bad’ data
  - Consistent as Possible

ML Data Content

- ML models are programed with data
- Beyond technical requirements the content of the data is important
- What is in your data determines how your model works
  - Will it generalize
  - Will it be able to train
  - How accurate will it be

Confounders

- Beyond the format of the data good datasets have limited irrelevant confounders
- ML picks up on patterns, so you don’t want to add your own that might not be relevant to the question you’re asking

Columbia River

Riverbend

Example Differences between Reproducible and AI-Ready

Blurry Image results from Bad Calibration

- Useful for your future-self to debug
- Not useful for training an ML model

An instrument is broken, so you're missing a variable

- Useful if you don't need that variable
- Not useful for training an ML model that needs that variable

Contaminated Sequences

- Useful for your future-self to debug
- Not useful for training an ML model

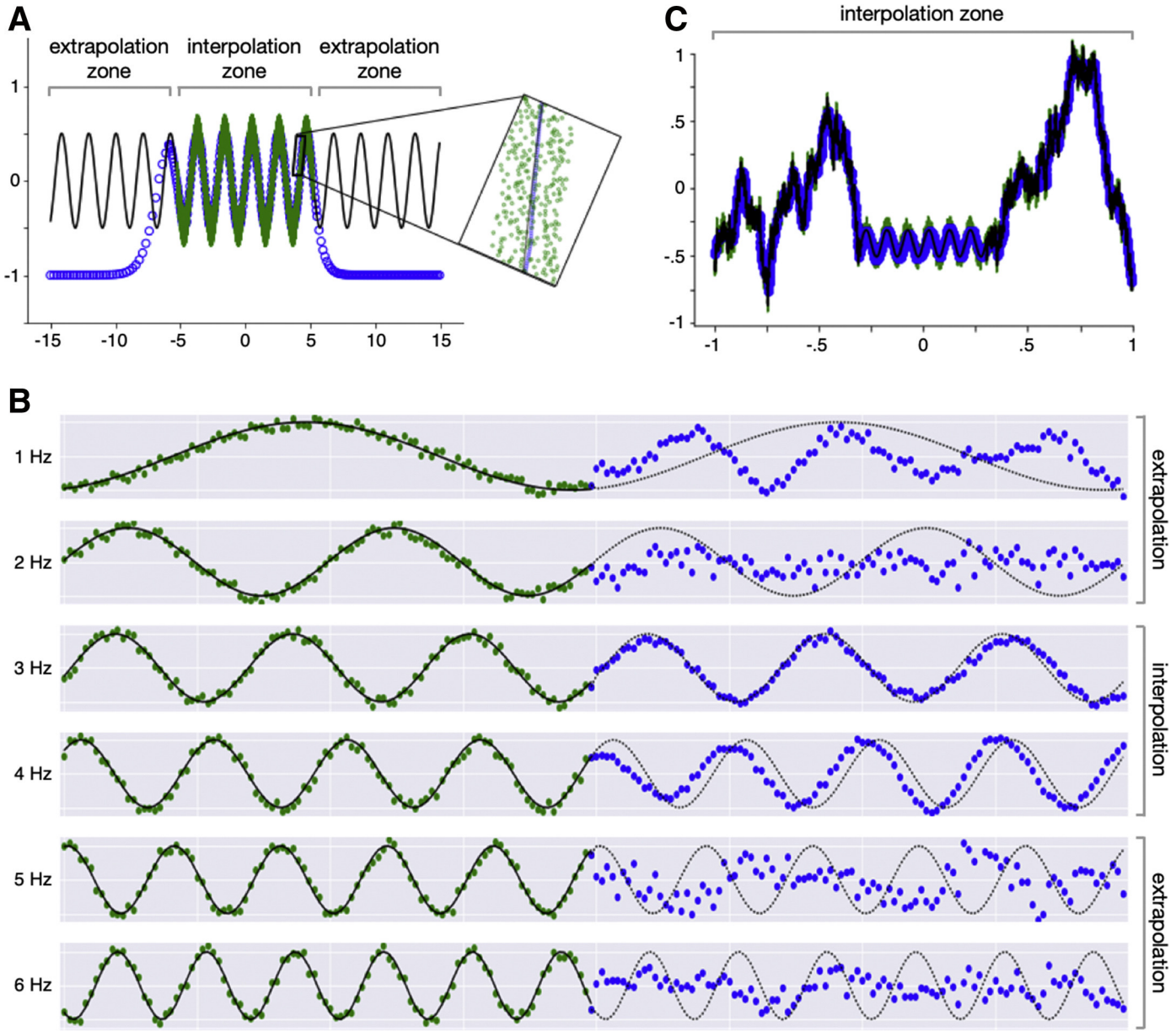
You switch units on an instrument

- Useful it's still the same data
- Not useful for training an ML model you must convert to consistent units in advance

When taking data consider recording whether your confident if that data should be used for science or if there is a possibility something went wrong. This can help you clean your data for ML projects, can True/False, or more fine grained red/yellow/green.

Extrapolation

- ML does a lot of neat things, but it isn't magic
- Unless specified treat ML algorithms as universal function approximators
  - i.e. 'Linear' regression is not a universal function approximator
- If you're collecting data make sure you collect 'representative' samples

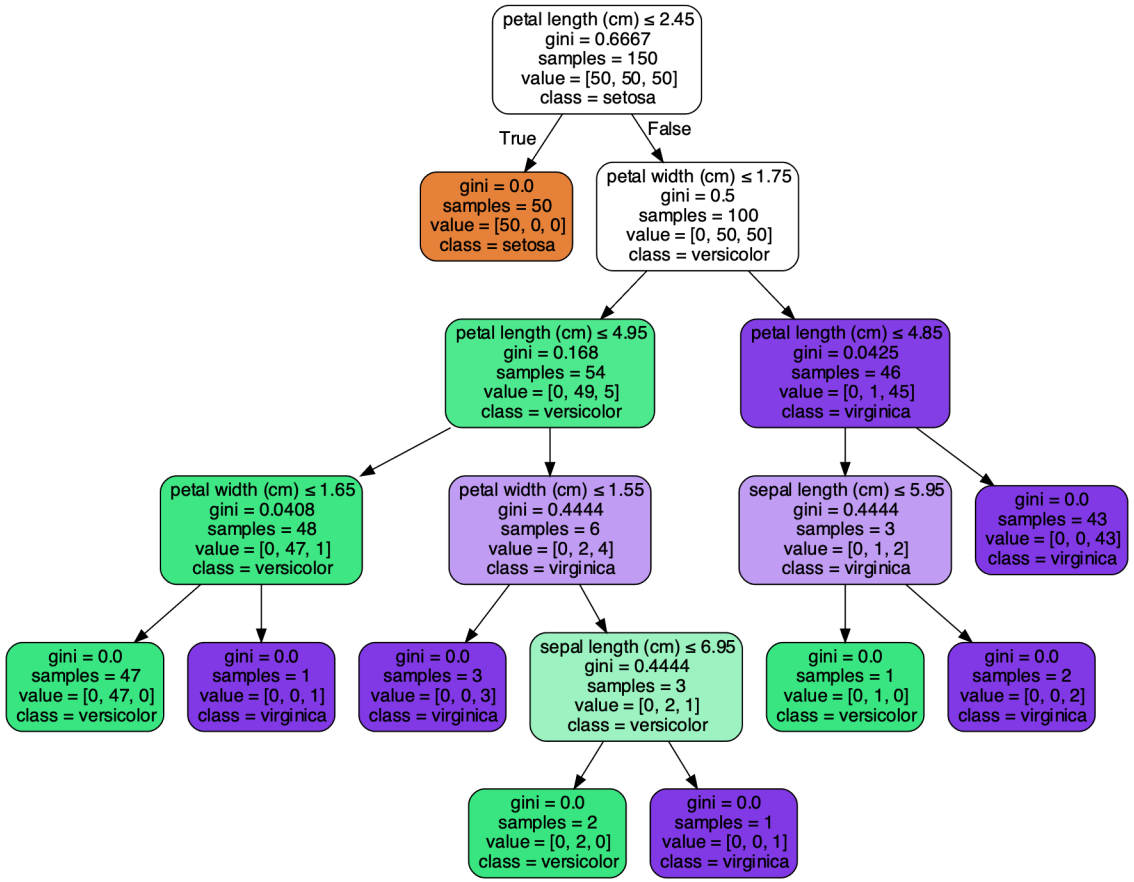


Algorithms Concepts

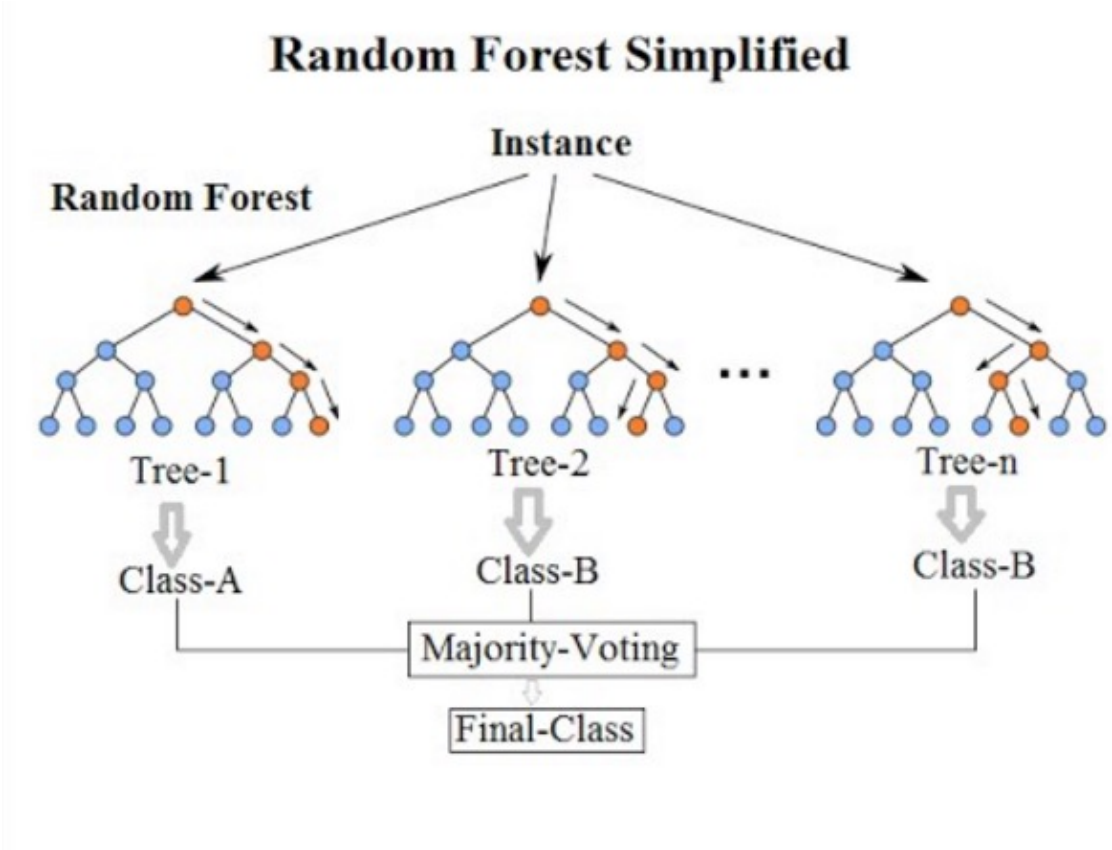
- Algorithms are selected to meet the data
- Structured
  - Anything that would fit in a spread sheet
  - Each row is an example each column is a variable
- Unstructured
  - Text, images, sequences, etc.
  - Variable length or sizes
  - Datasets are a consistent set of examples with optional labels

Random Forest

Decision Tree

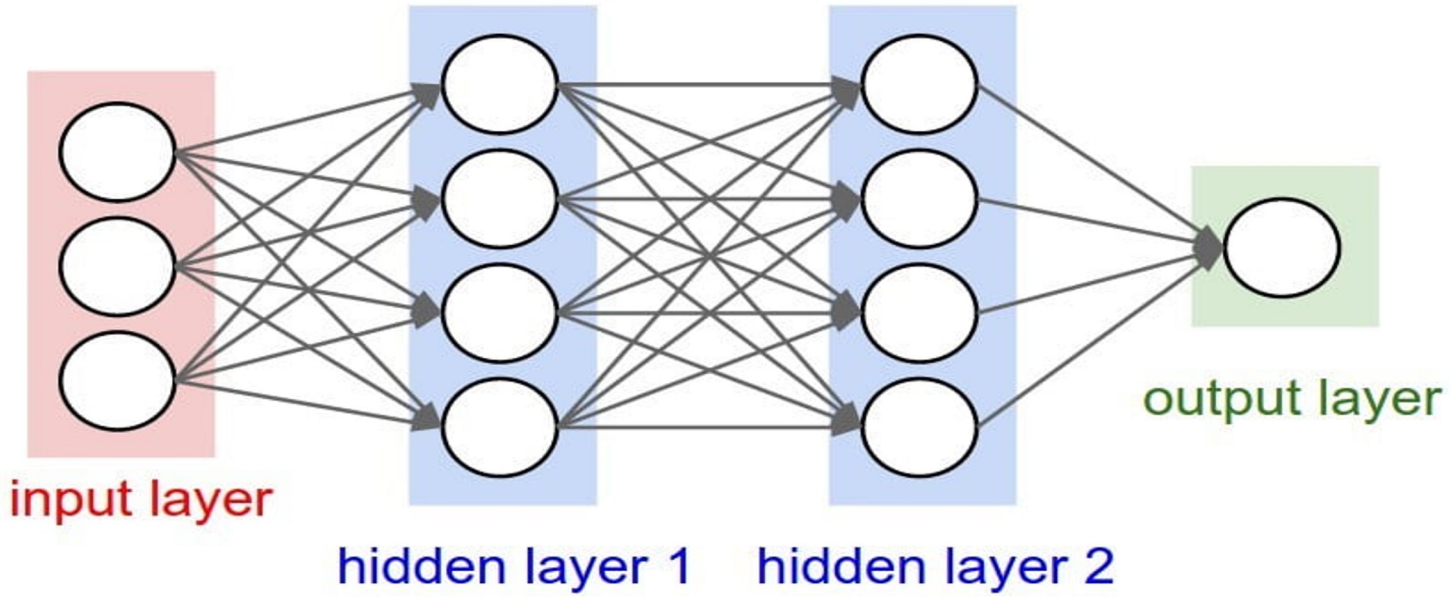


Decision Forest

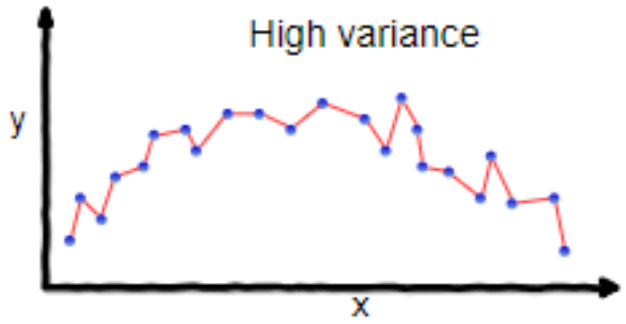


Neural Networks

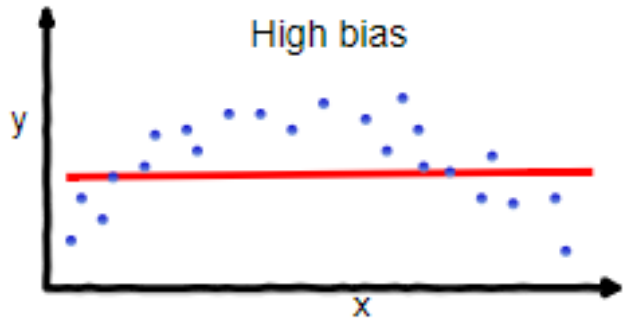
- A bit tricker to work with than decision tree
  - Need to be careful about data normalization
- Training may not actually be better than a boosted decision tree
- Why use them?
  - Offers tremendous flexibility
  - Can use any differential loss function
  - Has excellent performance if tuned



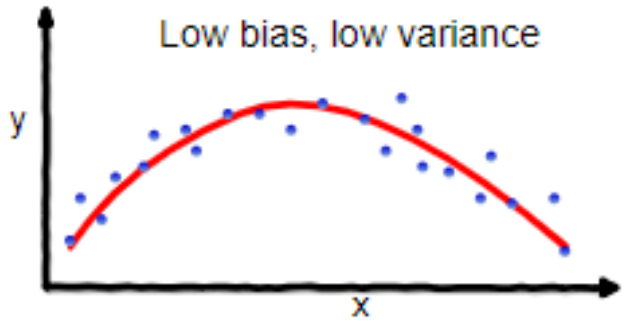
Bias Variance Tradeoff



overfitting



underfitting



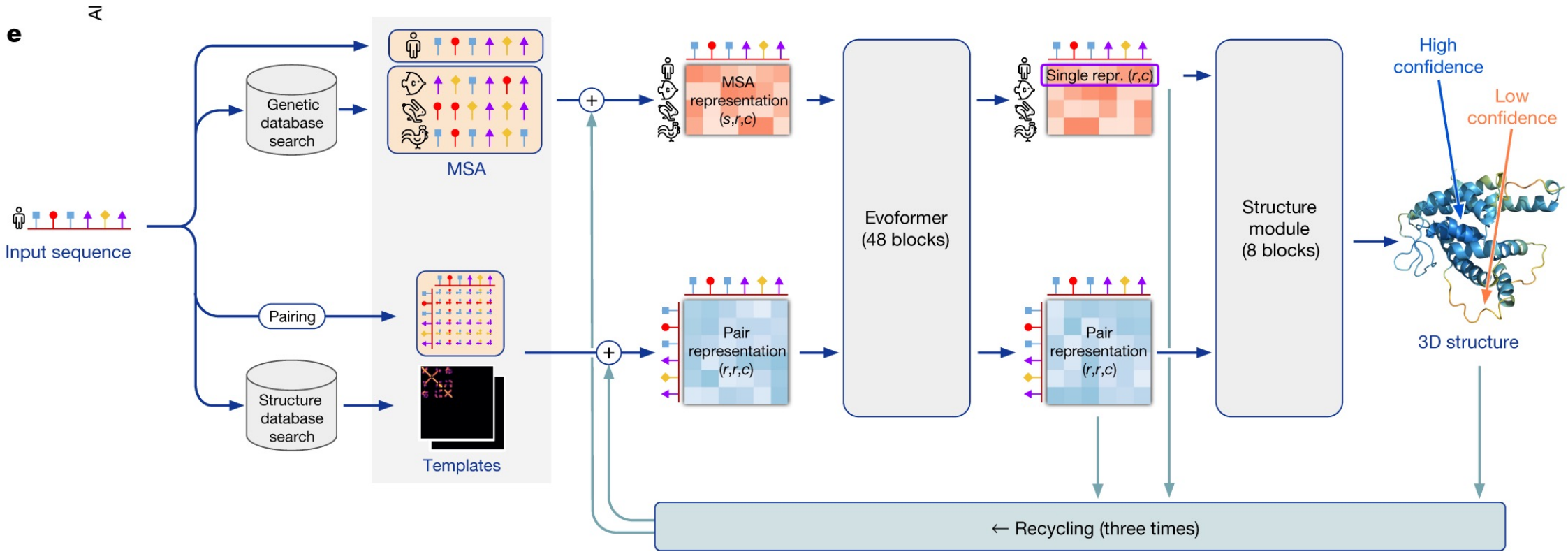
Good balance

https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

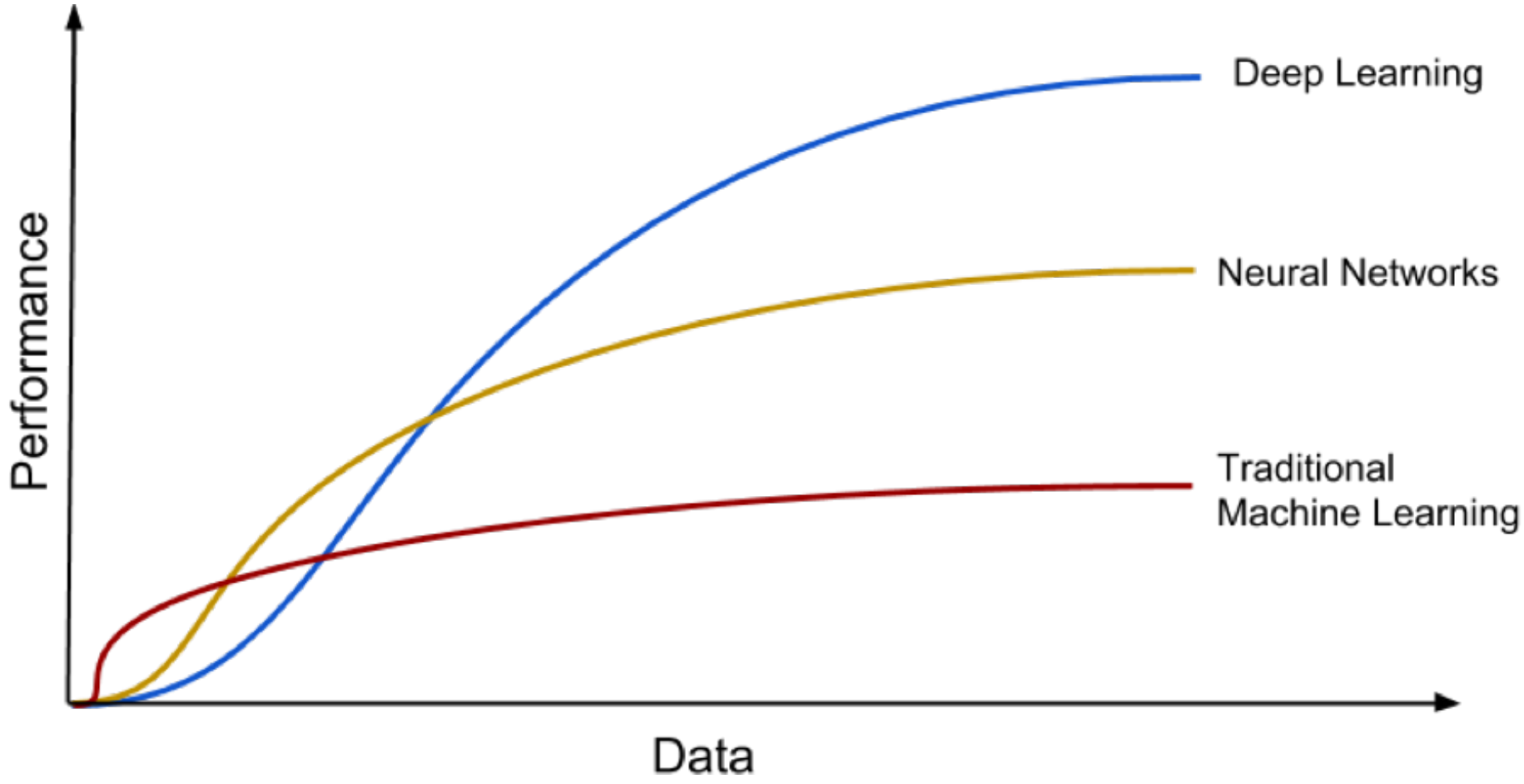
Deep Learning

- Unlike the above algorithms Deep Learning doesn't neatly fit into an R packages
- It’s better to think of it as a box of Lego blocks
- Deep Neural networks are made up of layers
- Layers exist for text, images, structured data, etc.
- Allows for building and training custom models
- This models are often made available

Example Alpha Fold (Protein Structure Prediction)

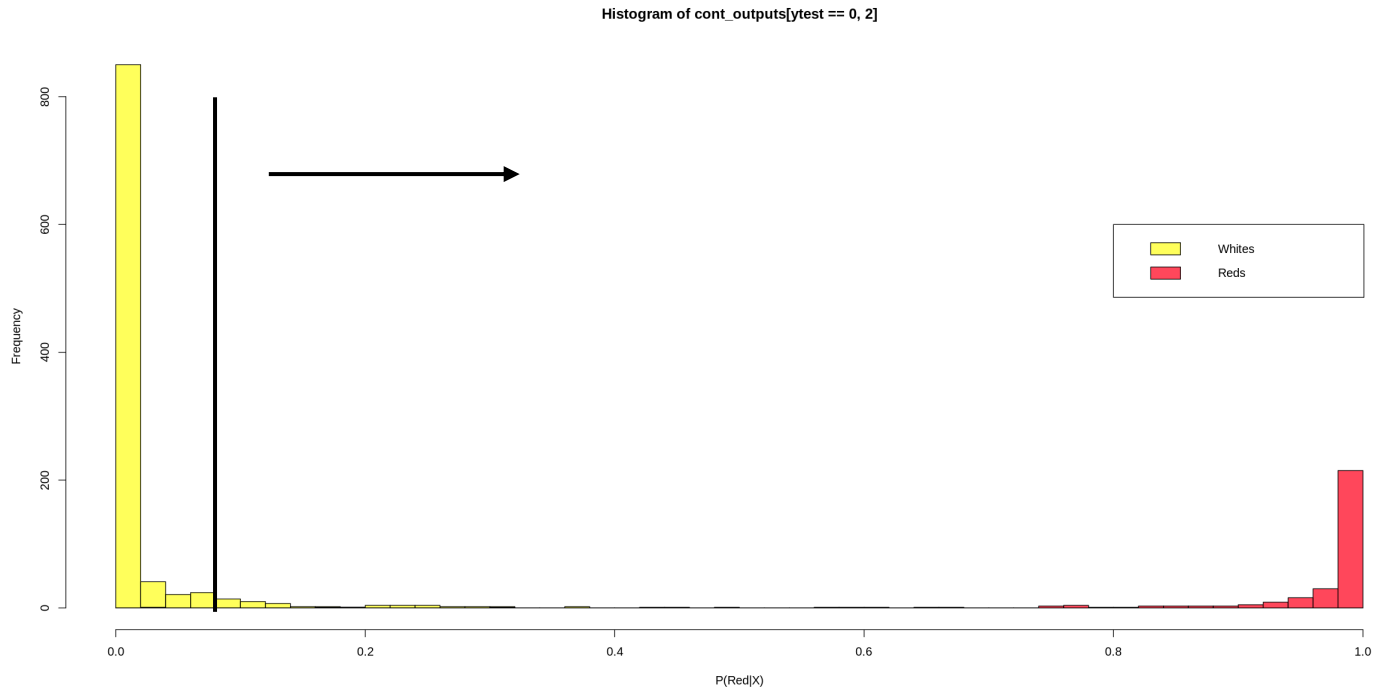


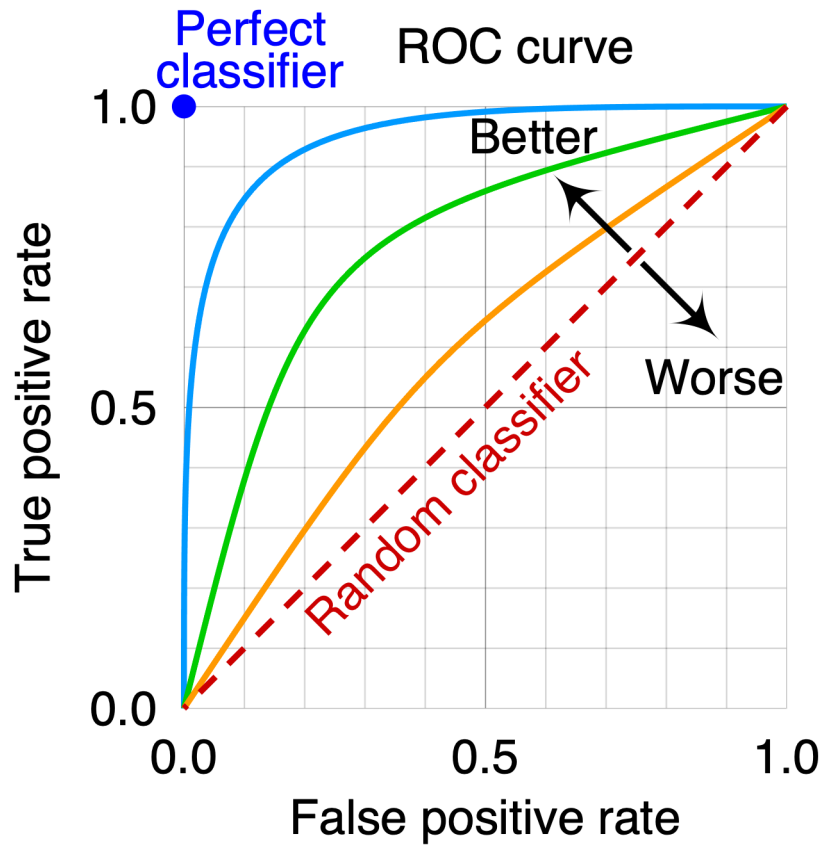
Models have different learning curves



Model Evaluation ROC Curves

- It can be a bit more complicated as we often don't just know the decision we can get a probability
- By default most tools assume >0.5 is True, but it isn't required
- You can give this Receiver Operator Characteristic
- **AUC = Area under Curve with 1 being the best**





This week

- We'll working on preparing sequencing data
  - The kinds of data and how much you need
- We'll apply some of the algorithms we use before on sequencing data
- We'll write a Deep Learning Algorithm
- We'll explore common tasks in gene expression data and unsupervised techniques to make them work.