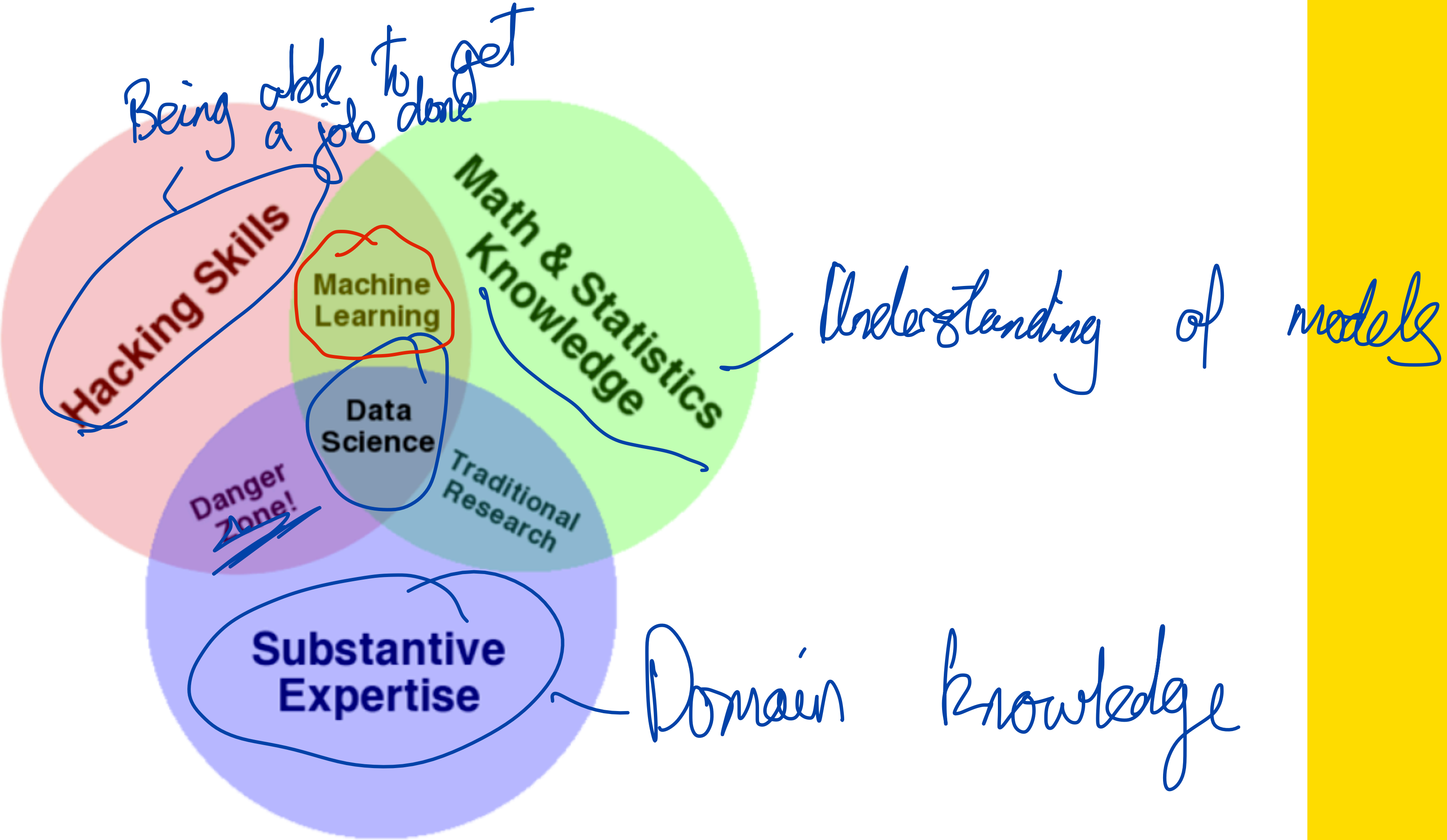
Introduction to Statistical Learning

ACTL3142 & ACTL5110 Statistical Machine Learning for Risk Applications

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

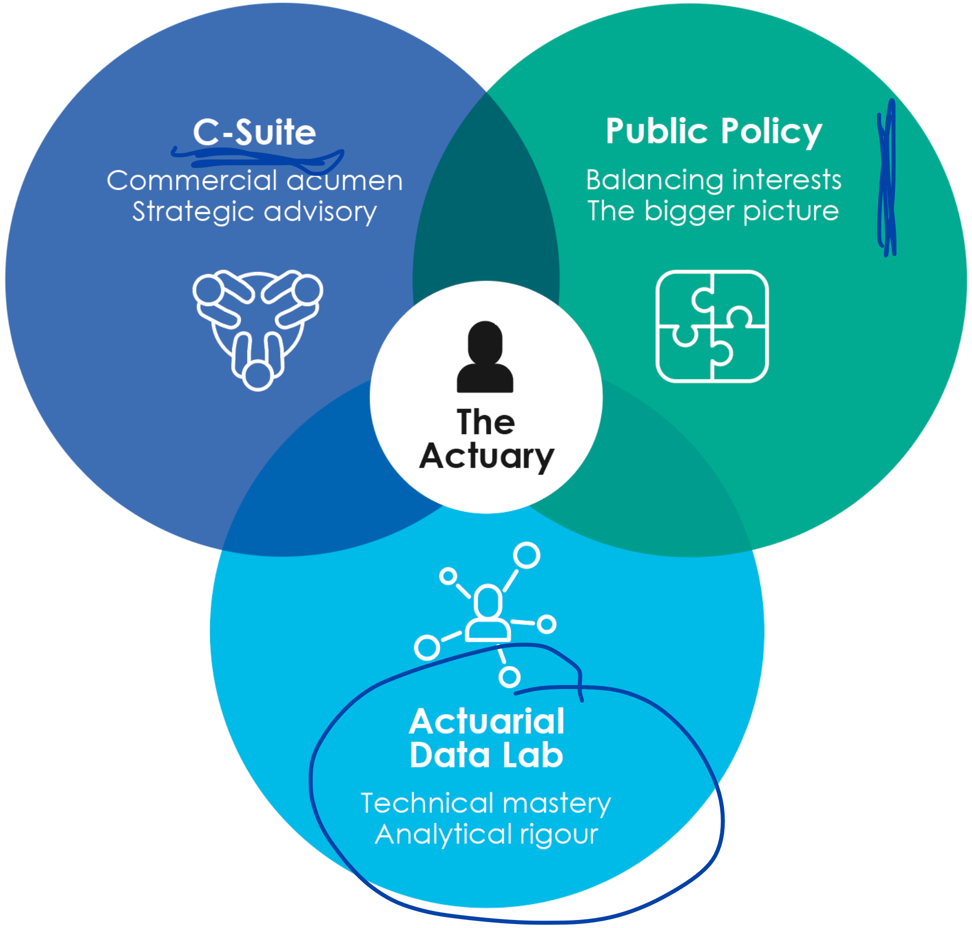
UNSW SYDNEY

Data Science Skills



Source: THE DATA SCIENCE VENN DIAGRAM, Drew Conway Data Consulting

Actuaries use data for good



Source: Actuaries Institute

Do data better with an Actuary

Source: Actuaries Institute

Lecture Outline

- **Statistical learning**
- Assessing model accuracy

Statistical Learning / Predictive Analytics

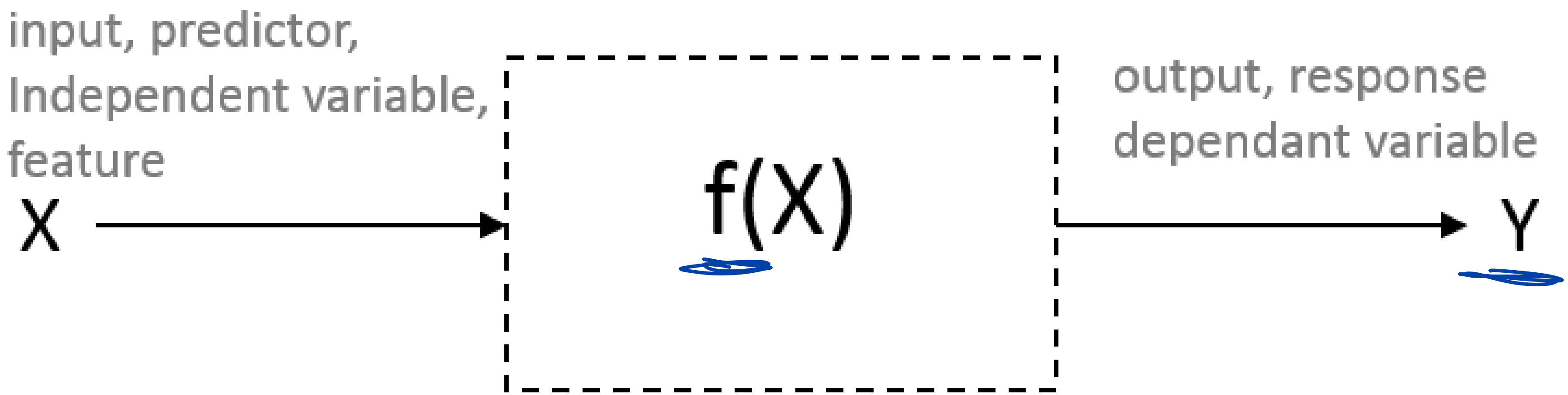
- A vast set of tools for understanding data.
- Other names used to refer to similar tools (sometimes with a slightly different viewpoint) - machine learning, predictive analytics
- Techniques making significant impact to actuarial work especially in the insurance industry
- Historically - started with classical linear regression techniques
- Contemporary extensions included
  - better methods to apply regression ideas
  - non-linear models
  - unsupervised problems
- Facilitated by powerful computation techniques and also accessible software such as R

generative AI

What is statistical (machine) learning?

- Historical stock price, macro data, industry specific → Stock price in 3-months time
- Location, temperature, clouds, atmospheric pressure → Weather

What is statistical (machine) learning?



Prediction

- Predict outcomes of Y given X
- What it means isn't as important, it just needs accurate predictions
- Models tend to be more complex

Inference

- Understand how Y is affected by X
- Which predictors do we add? How are they related?
- Models tend to be simpler

The Two Cultures

	Statistical Learning	Machine Learning
Origin	Statistics	Computer Science
f(X)	Model	Algorithm
Emphasis	Interpretability, precision and uncertainty	Large scale application and prediction accuracy
Jargon	Parameters, estimation	Weights, learning
Confidence interval	Uncertainty of parameters	No notion of uncertainty
Assumptions	Explicit a priori assumption	No prior assumption, we learn from the data

See Breiman (2001) and Why a Mathematician, Statistician, & Machine Learner Solve the Same Problem Differently

What is statistical (machine) learning?

Recall that in regression, we model an outcome against the factors which might affect it

error

$$Y = f(X) + \epsilon$$

Proxy for nature

- Y is the outcomes, response, target variable
- $X := (X_1, X_2, \ldots, X_p)$ are the features, inputs, predictors
- ϵ captures measurement error and other discrepancies

$\Rightarrow$ Our objective is to **find** an **appropriate** f for the problem at hand. Harder than it sounds

- What Xs should we choose?
- Do we want to predict reality (prediction) or explain reality (inference)?
- What's signal and what's noise?

X

error you observe

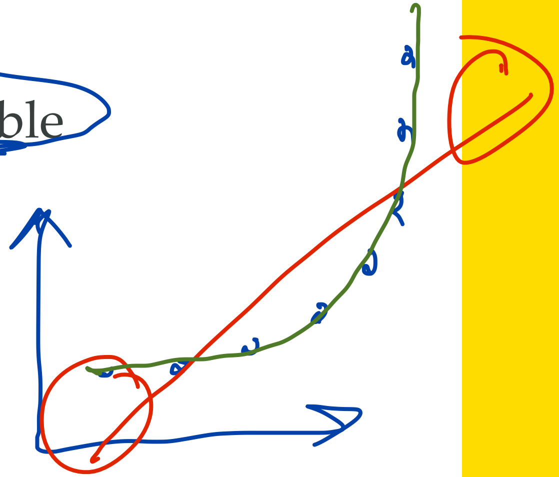
How to estimate f?

f is a line

$f(x) = mx + b$

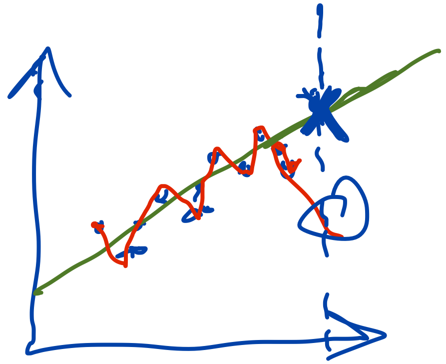
Parametric

- Make an assumption about the shape of f
- Problem reduced down to estimating a few parameters
  - Works fine with limited data, provided assumption is reasonable
- Assumption strong: tends to miss some signal



Non-parametric

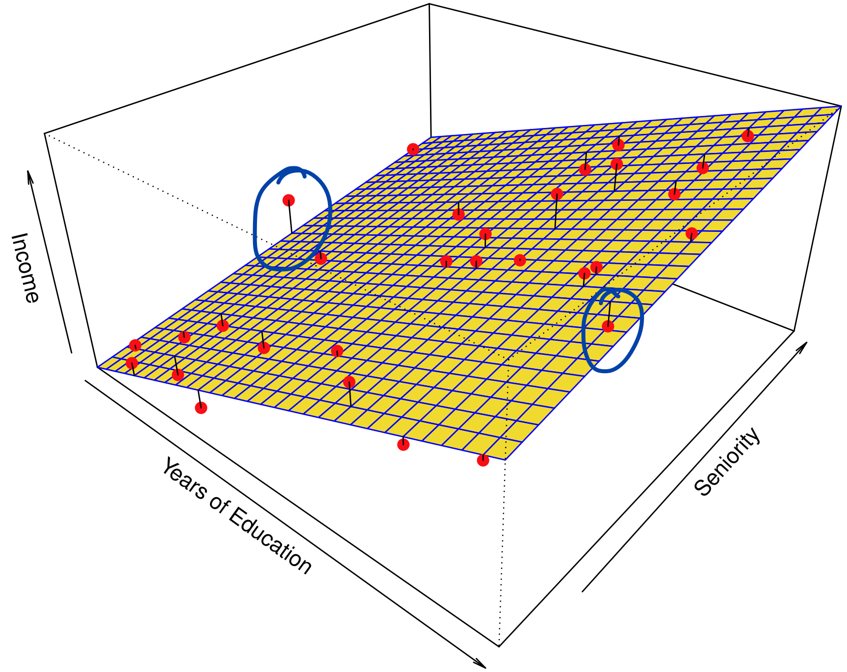
- Make no assumption about f's shape (specifying f from data
- Involves estimating a lot of "parameters"
  - Need lots of data
- Assumption weak: tends to incorporate some noise
- Be particularly careful re the risk of overfitting



f is too close to the data it learned, and not the true data

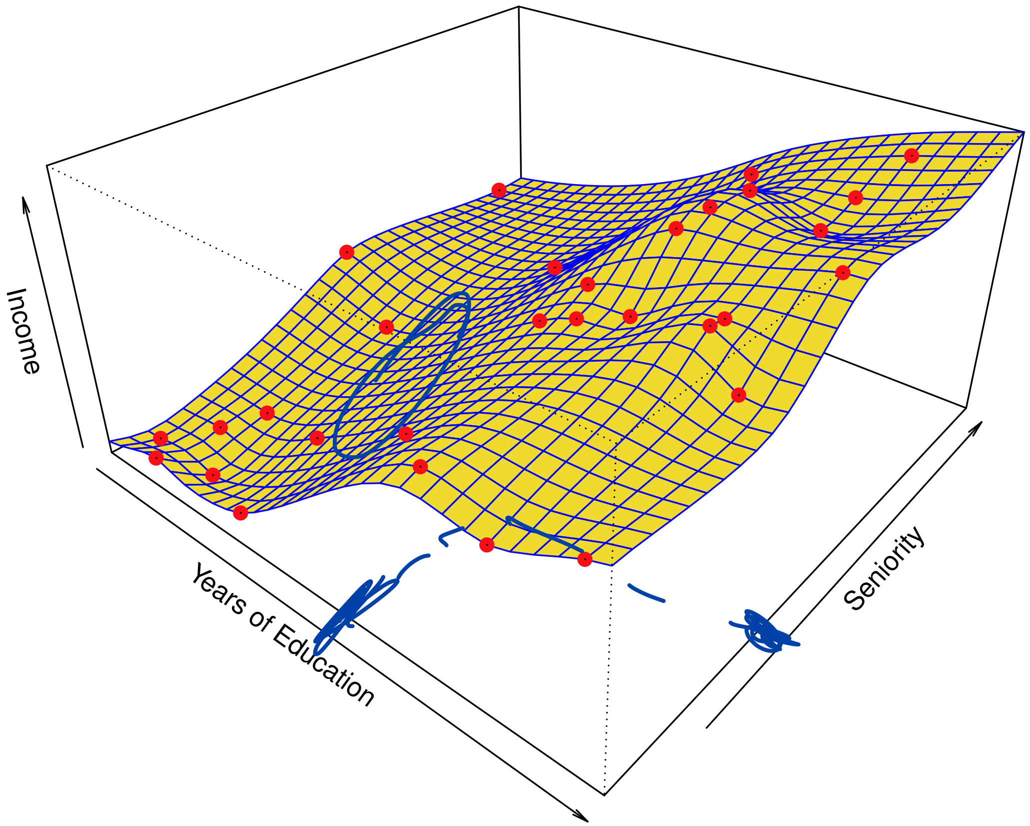
Example: Linear model fit on `income` data

Using Education and Seniority to explain Income:



- Linear model fitted
- Does a pretty decent job of fitting the data, by the looks of it, but doesn't capture *everything*

Example: "Perfect" fit on `income` data



- Non-parametric spline fit
- Fits the data perfectly. This is indicative of overfitting

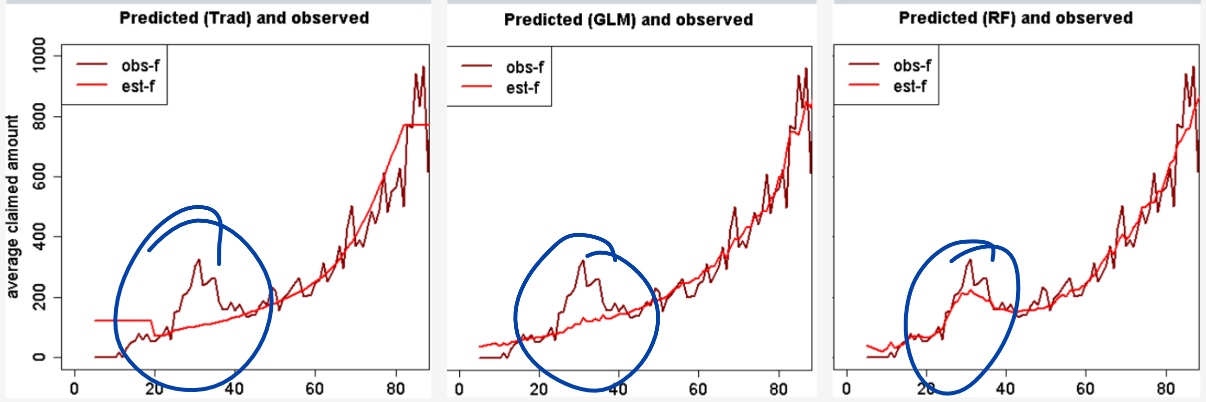
Actuarial Application: Health Insurance model choice

Predicted vs. observed claimed amounts for particular subgroups allows optimal model choice

MUNICH HEALTH   Munich RE

Observed and predicted average claimed amounts in 2012 itemized by age and gender (here only women)

Simpler model — **Traditional approach** | **GLM** | **Random Forest** — Complex



The traditional approach takes age and gender into account and therefore mostly performs quite good on average. Only the random forest detects the peak for women in their thirties (pregnancy treatments)

Source: Munich Re

April 2015   16

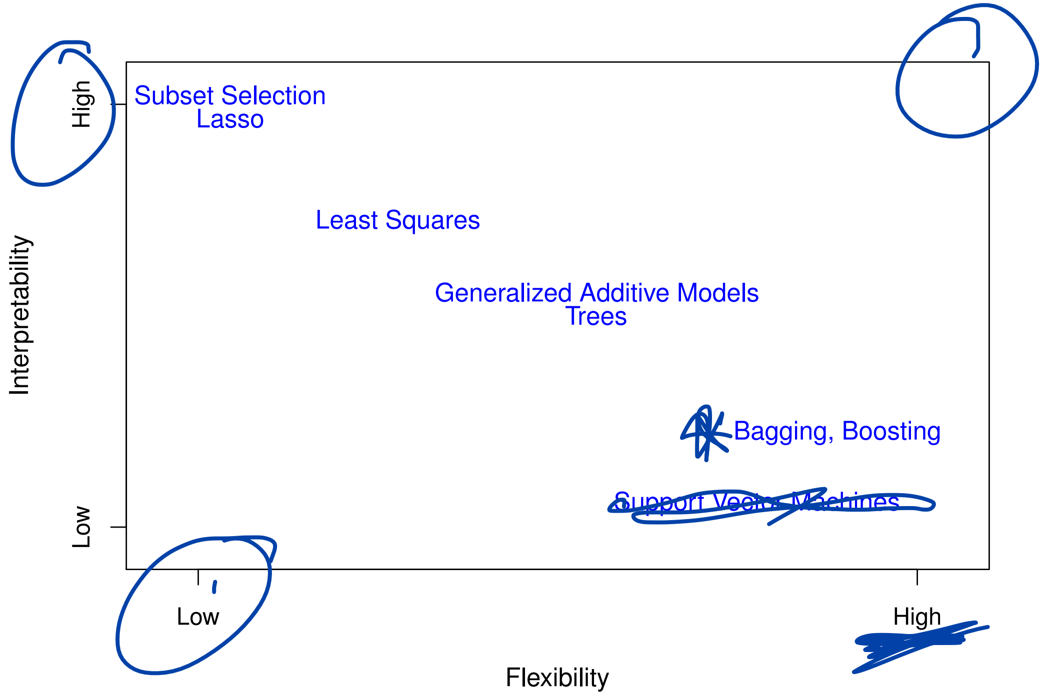
Tradeoff between interpretability and flexibility

- We will cover a number of different methods in this course
- They each have their own (relative) combinations of interpretability and flexibility:

Inference: High interpretability

Prediction: High flexibility

Dream: High flex + interpret



Deep learning

Discussion Question

Suppose you are interested in prediction. Everything else being equal, which types of methods would you prefer?

Supervised vs unsupervised learning

Supervised

- There is a response (y_i) for each set of predictors (x_{ji})
- e.g. Linear regression, logistic regression
- Can find f to boil predictors down into a response

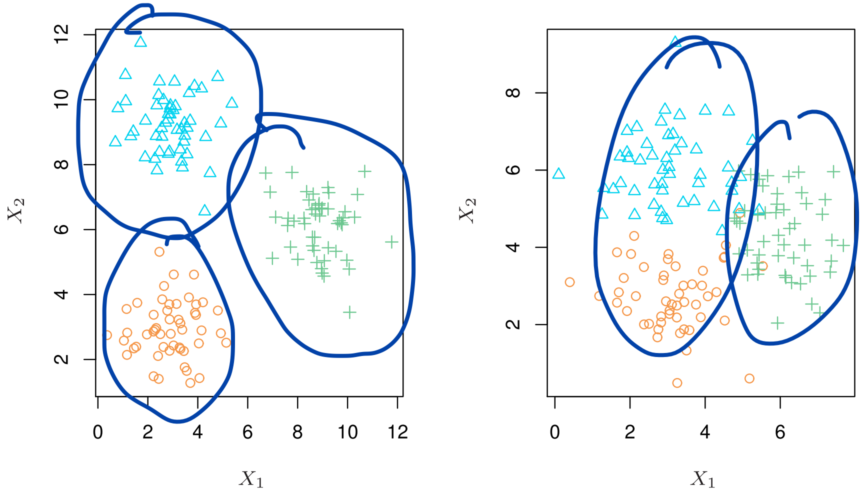
Weeks 1-9

Unsupervised

- No y_i, just sets of x_{ji}
- e.g. Cluster analysis
- Can only find associations between predictors

UNSW SYDNEY

Cluster analysis is a form of unsupervised learning



Week 10

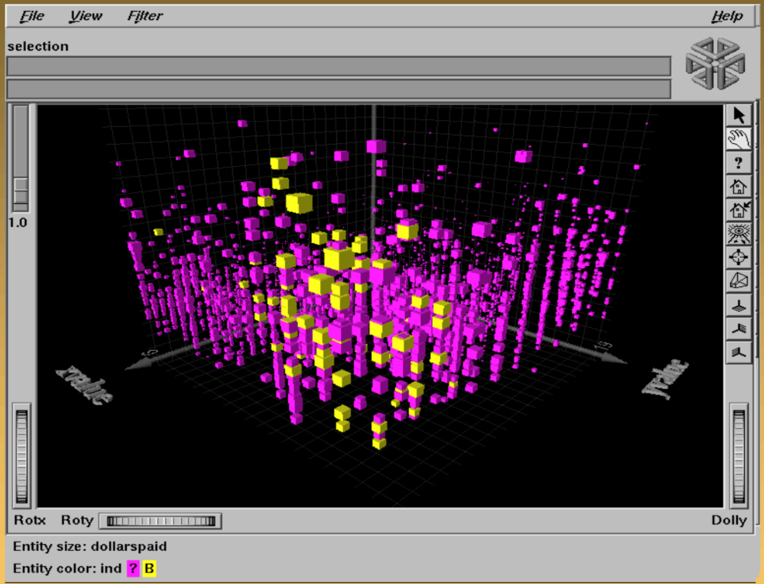
- For illustration we have provided the real groups (in different colours)
- In reality the actual grouping is not known in an unsupervised problem
- Hence idea is to identify the clusters.
- The example of the right will be more difficult to cluster properly

UNSW SYDNEY

Actuarial Application: predict claim fraud and abuse

DATA MODELING EXAMPLE: CLUSTERING

- Data on 16,000 Medicaid providers analyzed by unsupervised neural net
- Neural network clustered Medicaid providers based on 100+ features
- Investigators validated a small set of known fraudulent providers
- Visualization tool displays clustering, showing known fraud and abuse
- Subset of 100 providers with similar patterns investigated: Hit rate > 70%

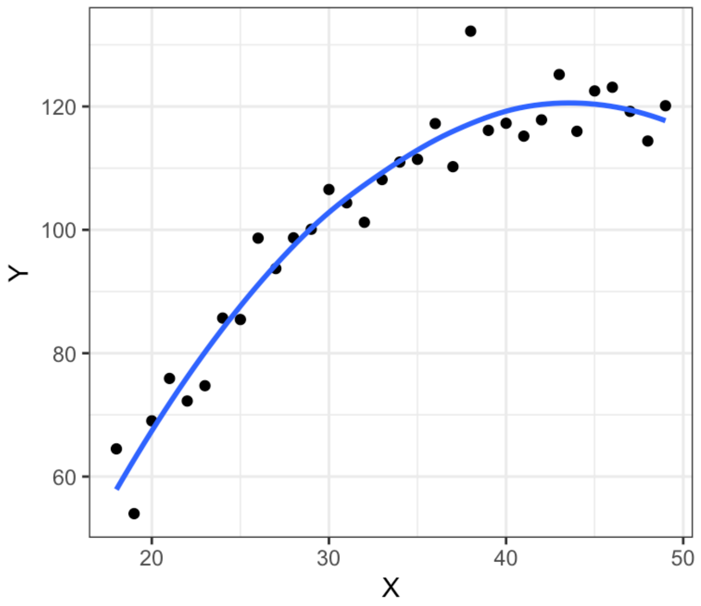


Cube size proportional to annual Medicaid revenues

Intelligent TECHNOLOGIES

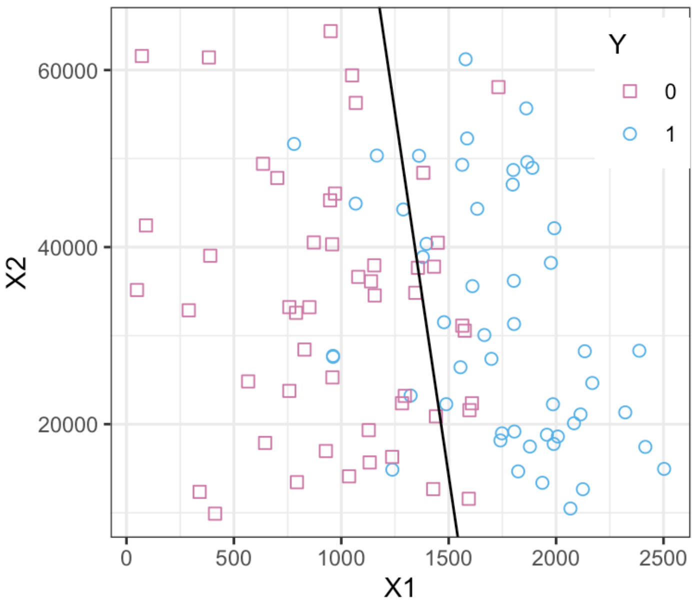
A note re Regression vs Classification problems – Supervised

Regression



- Y is quantitative, continuous
- Examples: Sales prediction, claim size prediction, stock price modelling

Classification



- Y is qualitative, discrete
- Examples: Fraud detection, face recognition, accident occurrence, death

$$Y = \underline{f}(X) + \varepsilon$$

How do we know if f is good?

Lecture Outline

- Statistical learning
- **Assessing model accuracy**

Assessing model accuracy

– Regression problems.

- Measuring the quality of fit and examples
  - Training MSE
  - Test MSE
- Bias-variance trade-off and examples
- Classification setting and example: K-Nearest Neighbors

Assessing Model Accuracy

$\hat{f} = \min_{f} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$

• Subject to restrictions on f.

• e.g. f is linear

- There are often a wide range of possible statistical learning methods that can be applied to a problem
- There is no single method that dominates over all others is all data sets
- How do we assess the accuracy?
  - Quality of Fit: Mean Squared Error

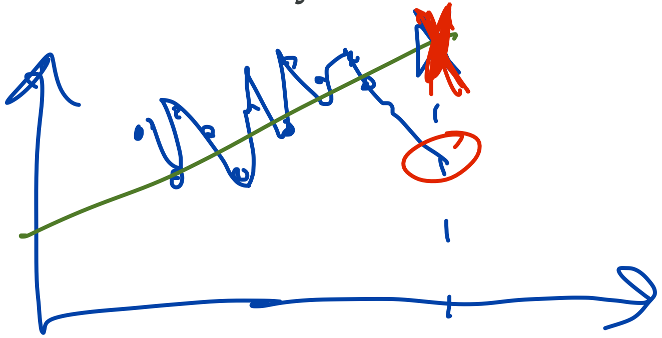
$\hat{f}$ is model estimated

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$

$f(x_i) = y_i$

  - This should be small if predicted responses are close to true responses.
  - Note that if the MSE is computed using the data used to fit the model (which is called the training data) then this is more accurately referred to as the **training MSE**.

$\hat{f}$ is based on this data.



training = data used to fit model.

Discussion question

What are some potential problems with using the training MSE to evaluate a model?

$\hat{f}(x_i) = m x_i + b$

$\hat{g}(x_i) = y_i$

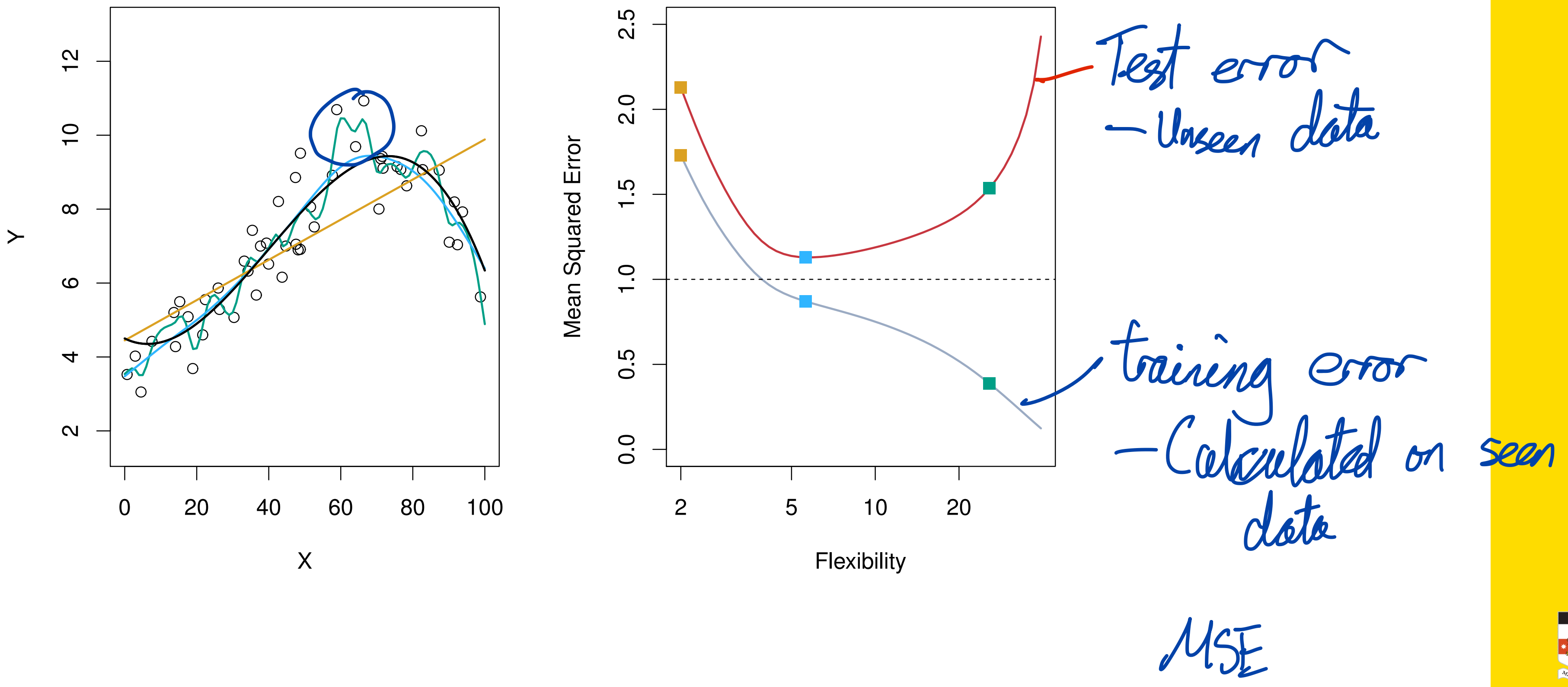
Using training MSE is no guarantee it will perform well on unseen test data

- Relying on training MSE will lead to overfitting as $f(x_i) = y_i$ minimises training MSE.

UNSW SYDNEY

Discussion question

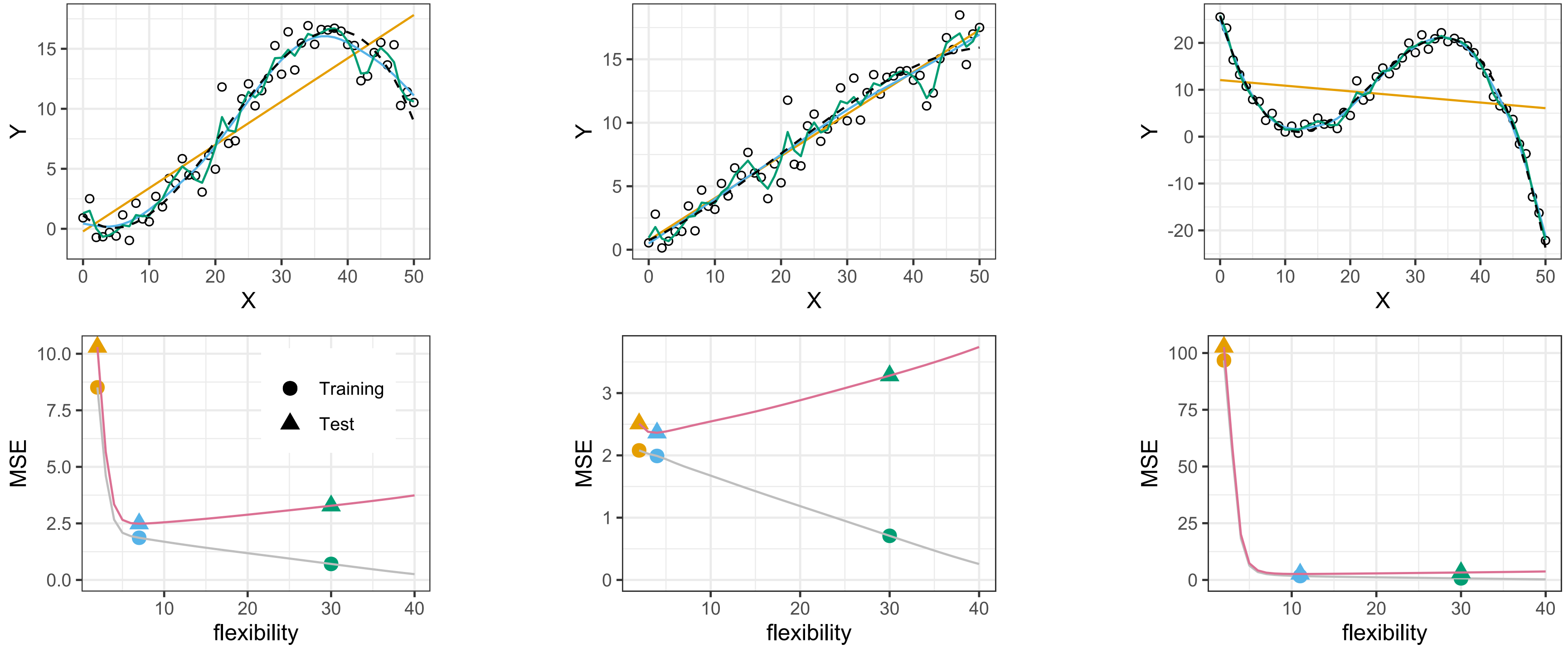
Consider the example below. The true model is black, and associated 'test' data are identified by circles. Three different fitted models are illustrated in blue, green, and orange. Which would you prefer?



e

Examples: Assessing model accuracy

The following are the training and test errors for three different problems:



train MSE min. at $\hat{f}(x_i) = y_i$

Bias-Variance Tradeoff

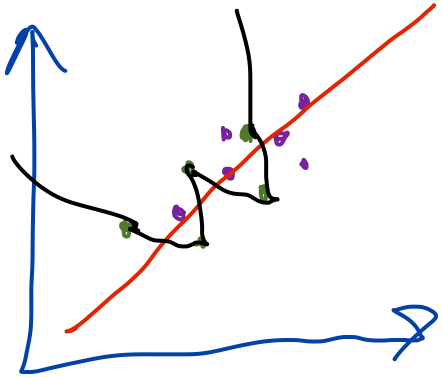
x_0 is new data not seen before.

The expected test MSE can be written as:

$$\mathbb{E}\big(y_0 - \hat{f}(x_0)\big)^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon)$$

- $\text{Var}(\hat{f}(x_0))$: how much $\hat{f}$ would change if a different training set is used } control this
- $[\text{Bias}(\hat{f}(x_0))]^2$: how much the model is off by
- $\text{Var}(\epsilon)$: irreducible error ✗ cannot control

There is often a tradeoff between Bias and Variance

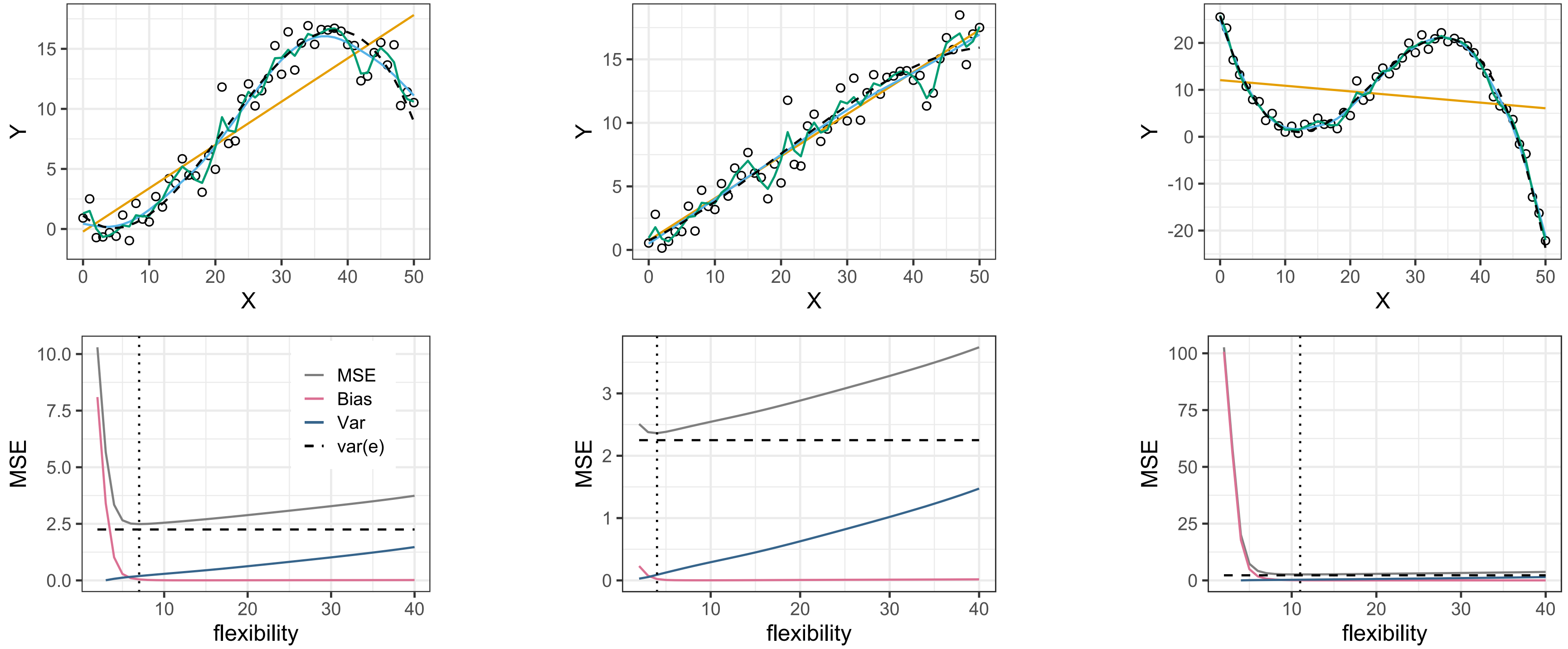


draw black for purple

– Did red line change much whether we use green or purple? black?

Examples: Bias-variance tradeoff

The following are the Bias-Variance tradeoff for three different problems:



Classification

$\frac{1}{n}\sum_{i=1}^{n}(\text{'Red'} - f(x_i))^2$

Objective

- Place data point into a category (Y) based on its predictors (X_i)
- Test Error is the proportion of times the estimate is wrong

$$\text{Ave}\left(I(y_0 \neq \hat{y}_0)\right)$$

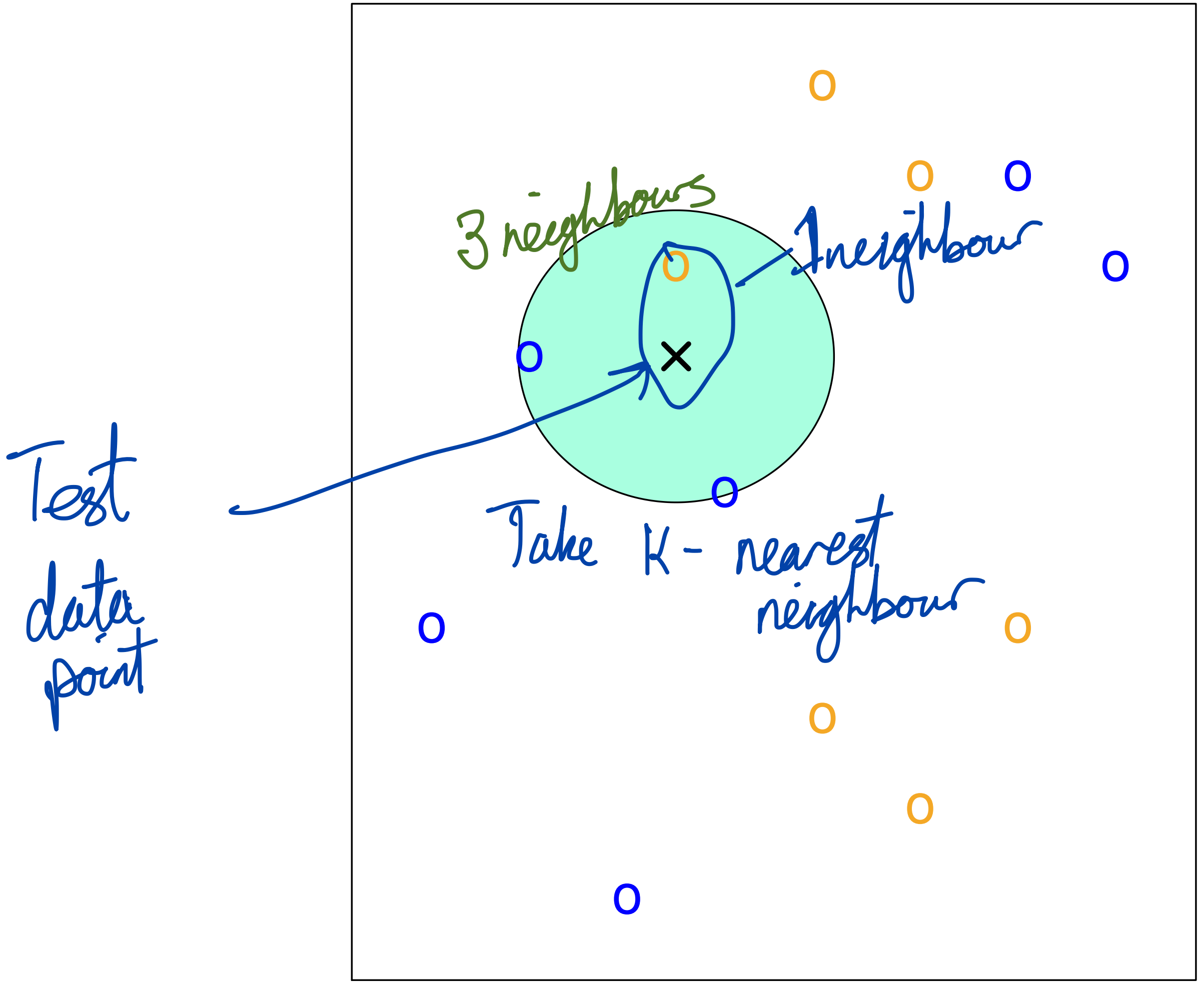
"1" if condition is satisfied

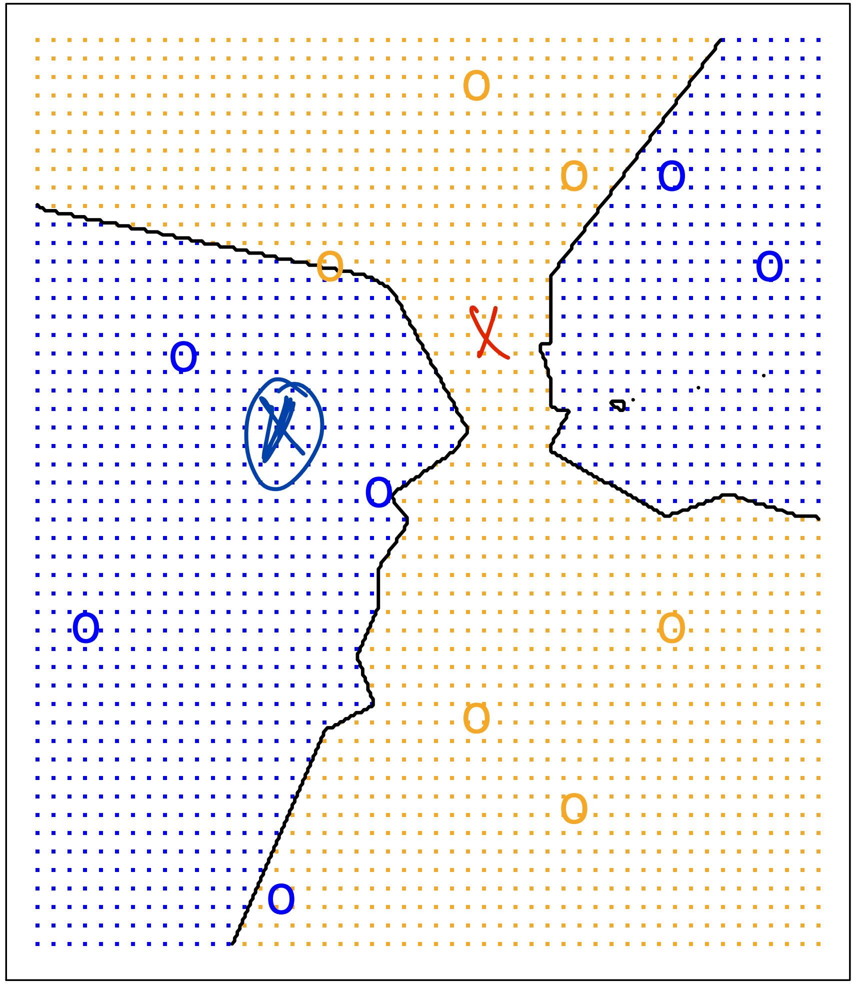
Bayes' Classifier

- Assigns a prediction x_0 to the class j which maximises $\mathbb{P}(Y = j|X = x_0)$ — Cannot calculate this
- In the case of two classes, this would be the one where $\mathbb{P}(Y = j|X = x_0) > 0.5$
- Theorectically the optimum, but in reality do not know the conditional probabilities.
- A simple alternative is the K nearest neighbors (KNN) classifier

UNSW SYDNEY

K-nearest neighbours - illustration





e

K-nearest neighbours

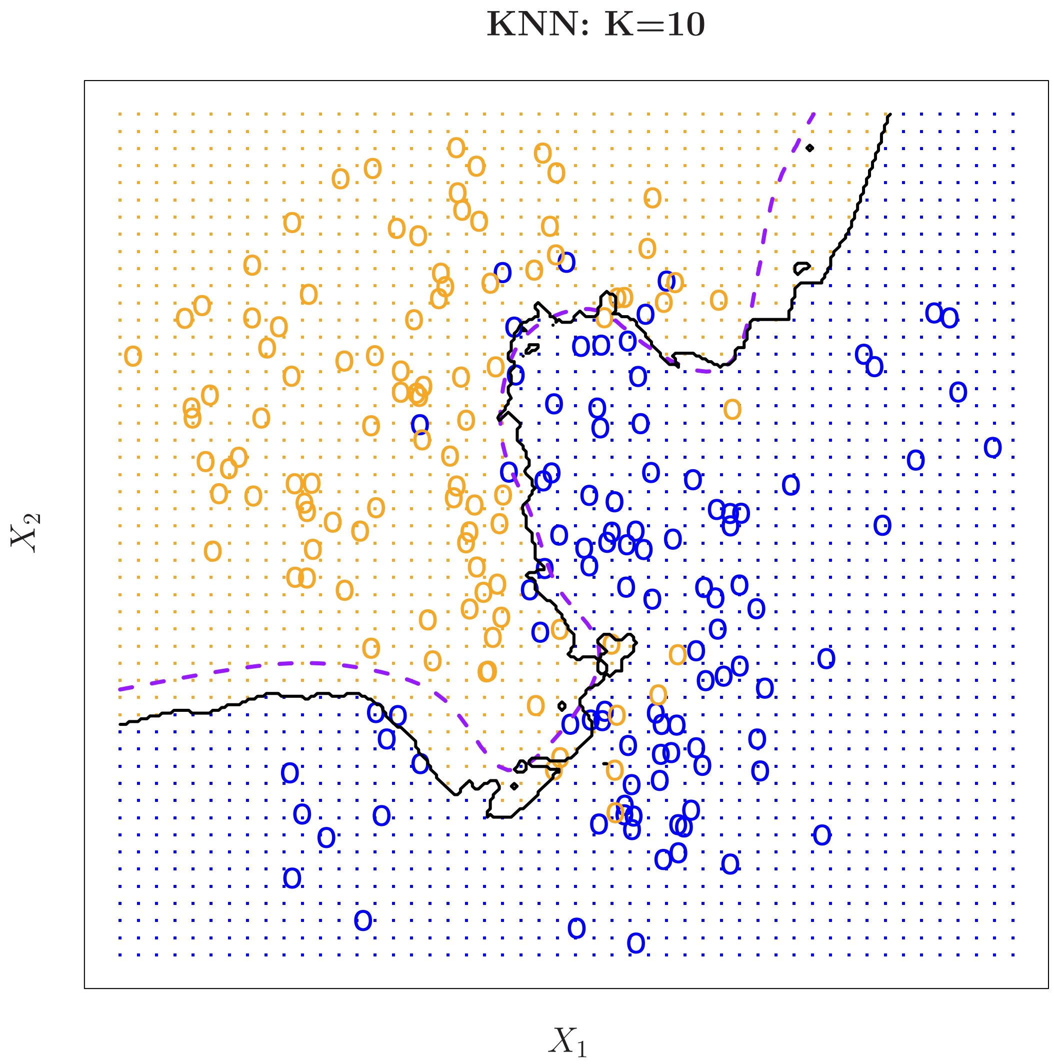
K-nearest neighbours

- Looks at a new observation's K-nearest (training) observations
  - In other words, it maximises

$$\mathbb{P}(Y = j | X = x_0) = \frac{1}{K} \sum_{i=1} \mathbb{I}(y_i = j)$$

- New observation's category is where the majority of its neighbours lie
- High K: less variance but more bias, fit missing signal - too close to a global average
- Low K: less bias but more variance, fit too noisy - assuming less relationship between close-by data points than there is
- Intelligent choice of K is key: too low and you overfit, too high and you miss important information

K-nearest neighbours example, K=10



(purple is the Bayes boundary, black is the KNN boundary with K=10)

K-nearest neighbours example, K=1, K=100