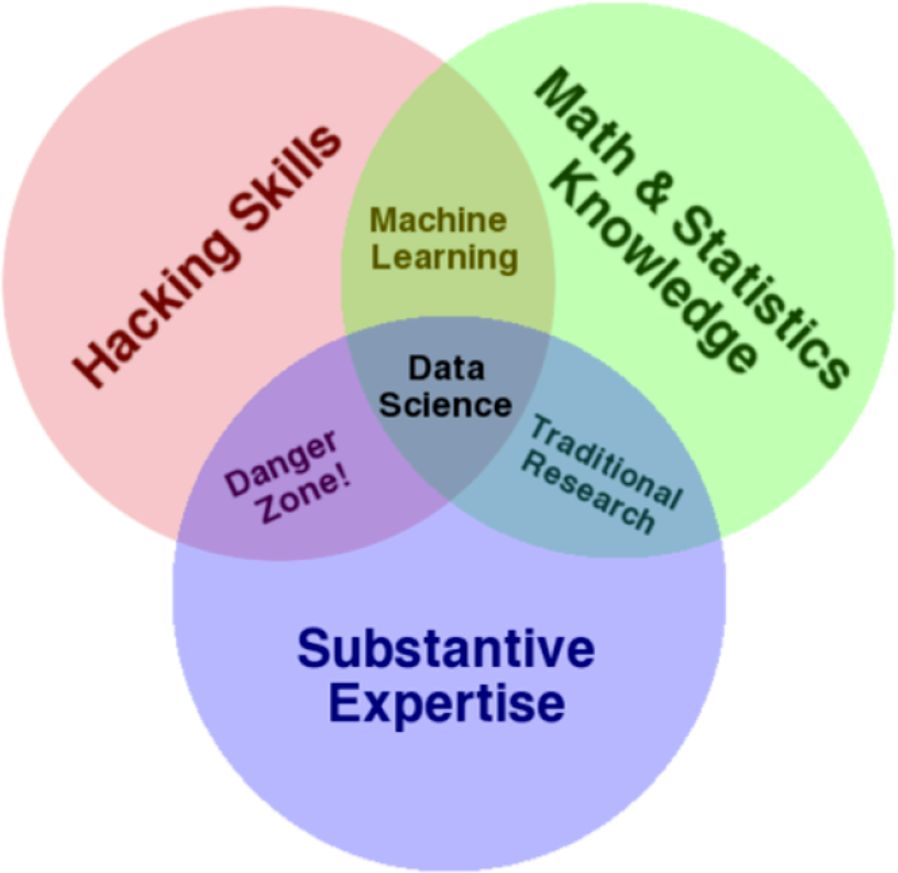
Introduction to Statistical Learning

ACTL3142 & ACTL5110 Statistical Machine Learning for Risk Applications

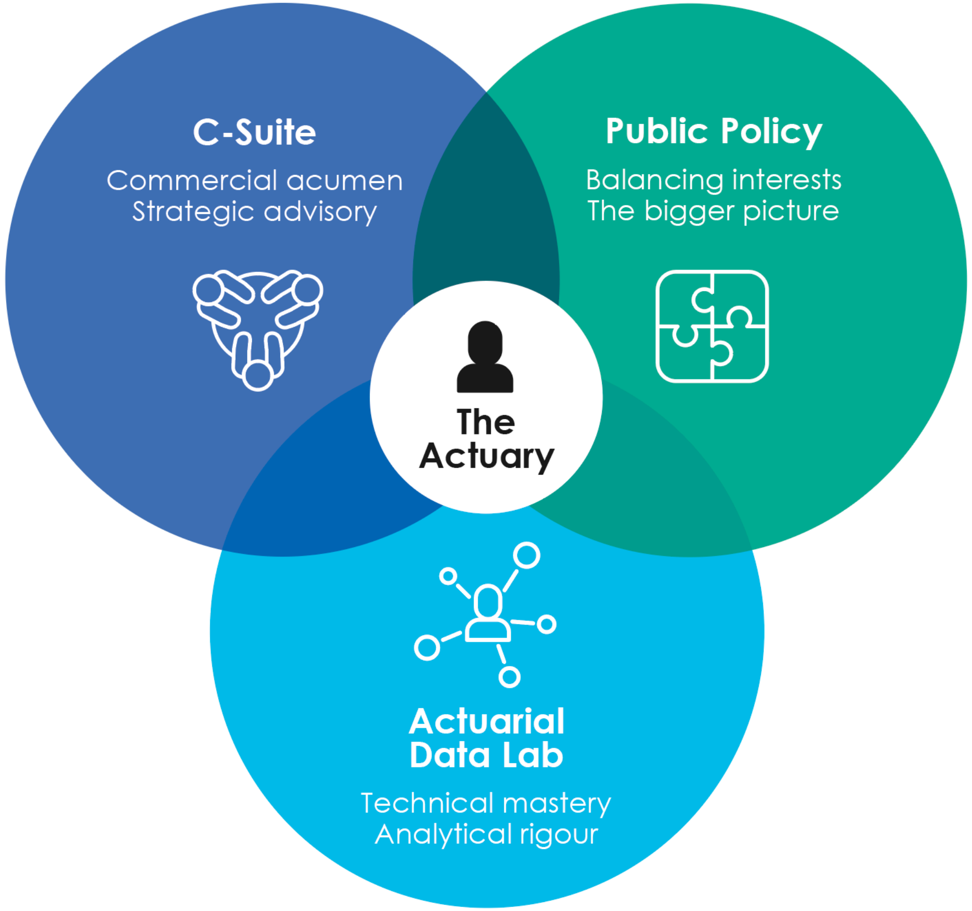
Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Data Science Skills



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

Actuaries use data for good



C-Suite

Commercial acumen
Strategic advisory

Public Policy

Balancing interests
The bigger picture

The Actuary

Actuarial Data Lab

Technical mastery
Analytical rigour

Source: Actuaries Institute

Do data better with an Actuary

Source: Actuaries Institute

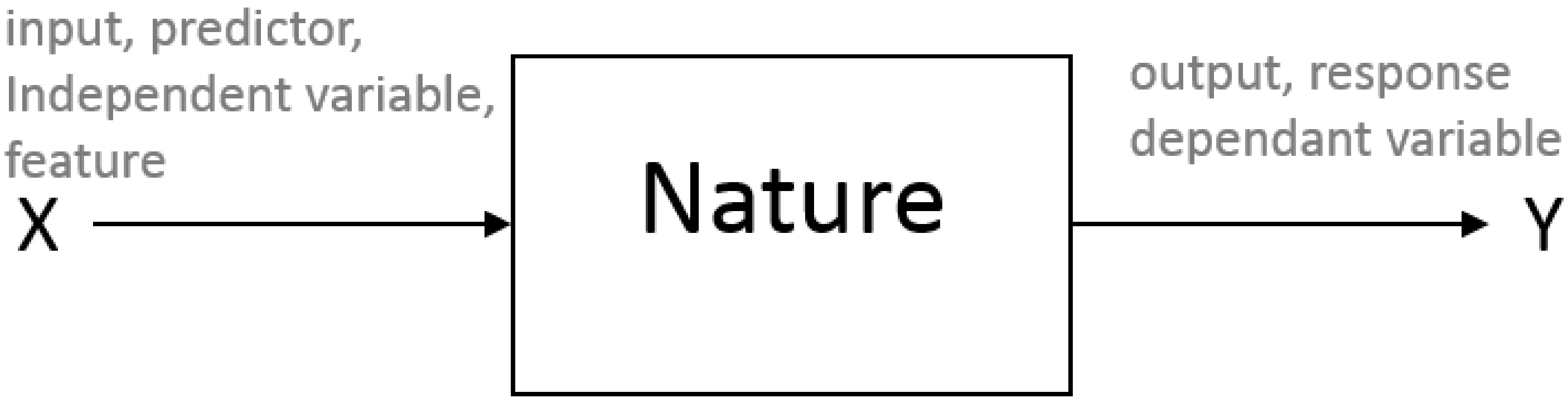
Lecture Outline

- **Statistical learning**
- Assessing model accuracy

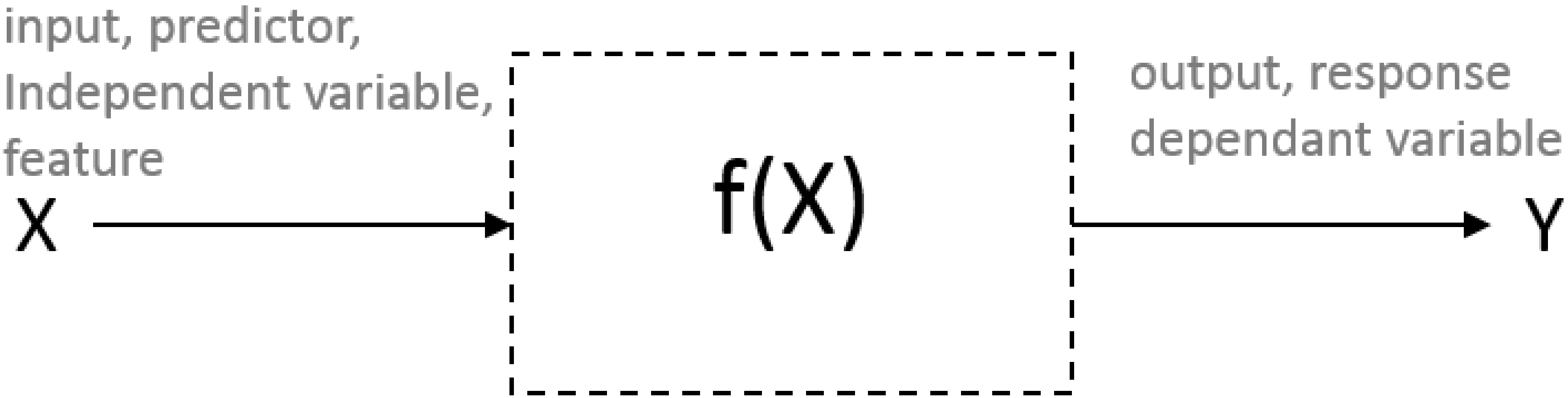
Statistical Learning / Predictive Analytics

- 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

What is statistical (machine) learning?



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

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

- 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

⇒ 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?

How to estimate f?

Parametric

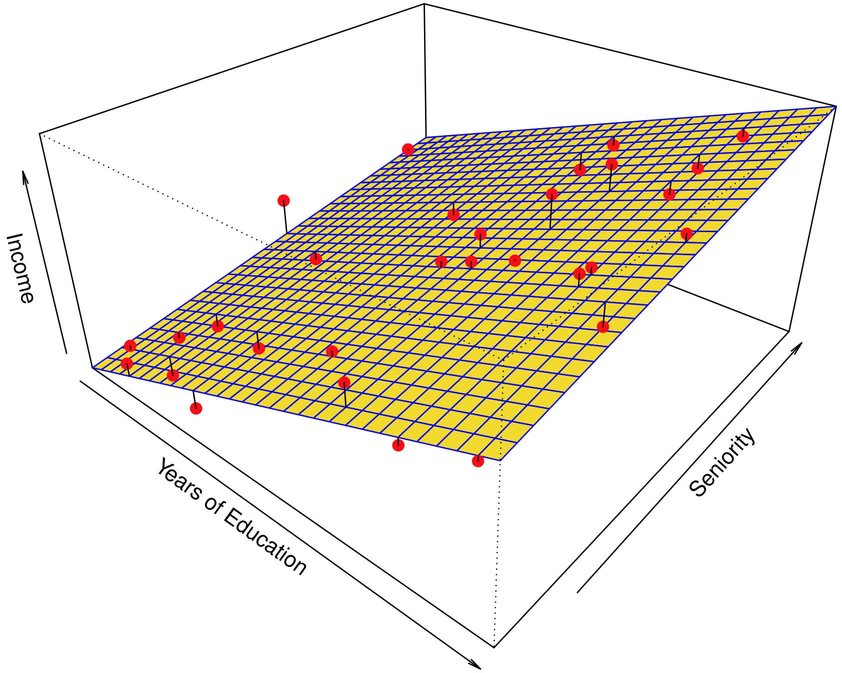
- Make an assumption about the shape of f
- Problem reduced down to estimating a few parameters
  - Works fine with limited data, provided assumption is reasonable
- Assumption strong: tends to miss some signal

Non-parametric

- Make no assumption about f's shape
- 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

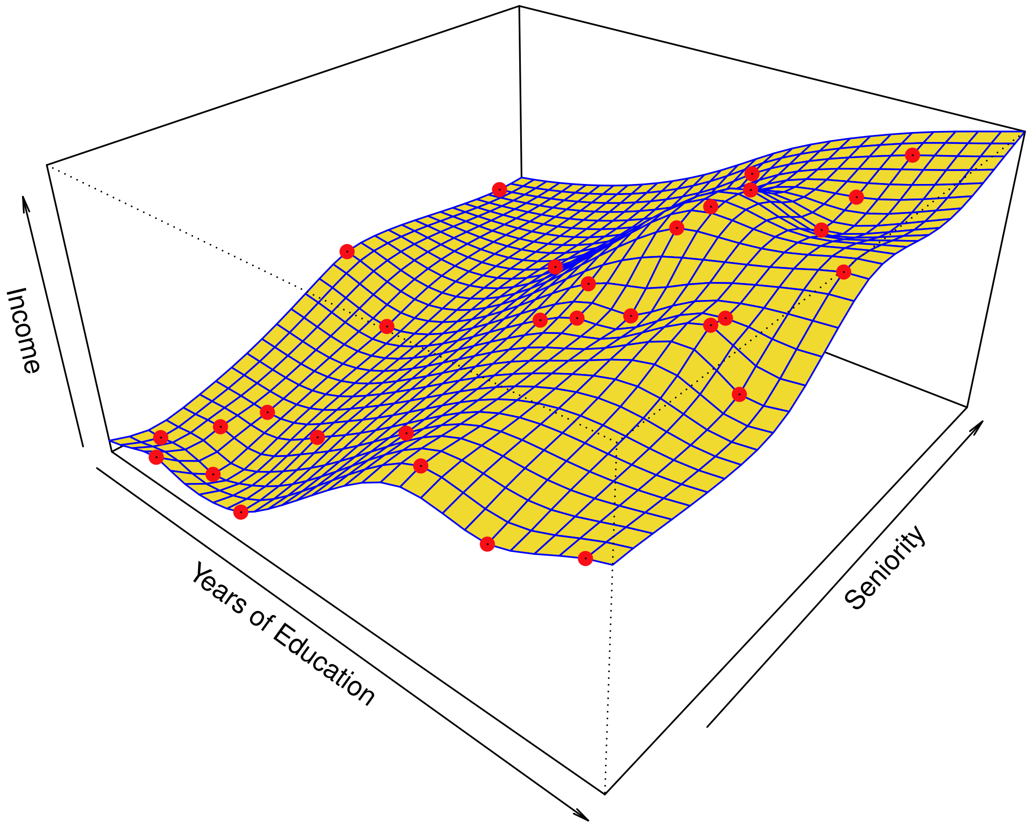
Example: Linear model fit on `income` data

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*

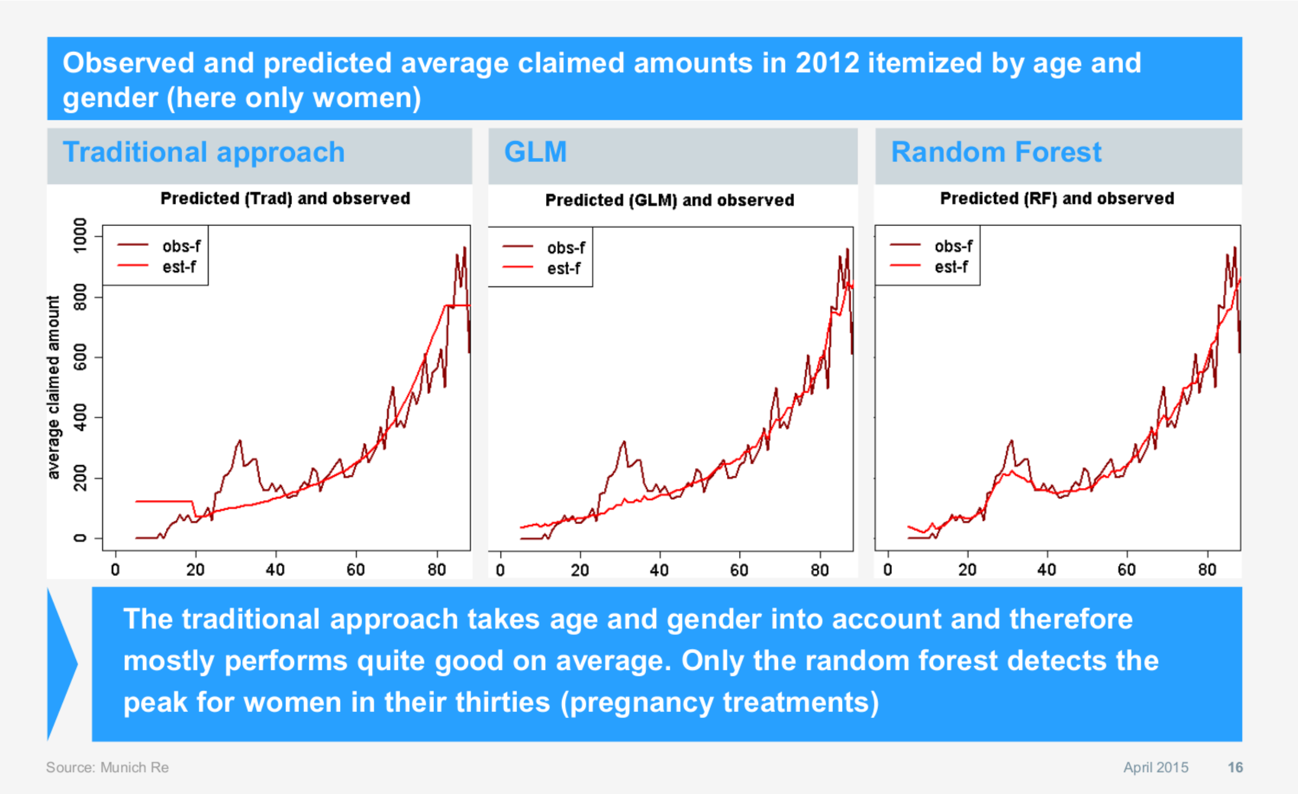
Example: “Perfect” fit on `income` data



- 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



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

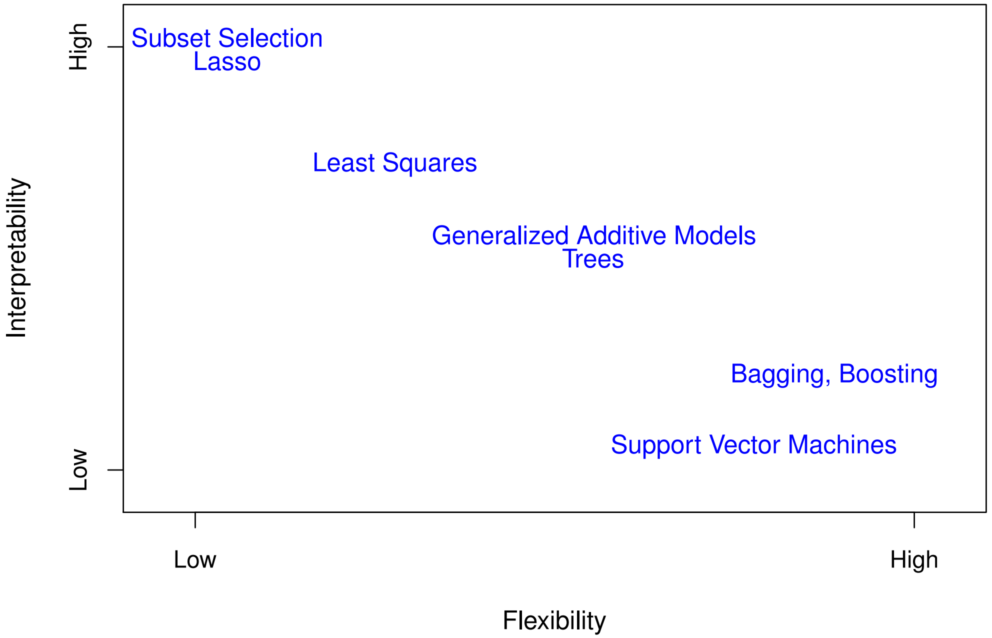
The traditional approach takes age and gender into account and therefore mostly performs quite good on average. Only the random forest detects the peak for women in their thirties (pregnancy treatments)

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:



Discussion Question

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

Supervised vs unsupervised learning

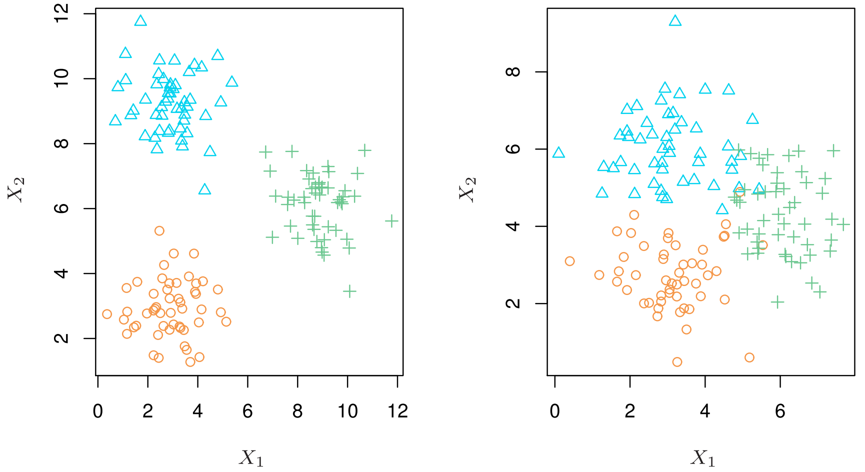
Supervised

- 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

Unsupervised

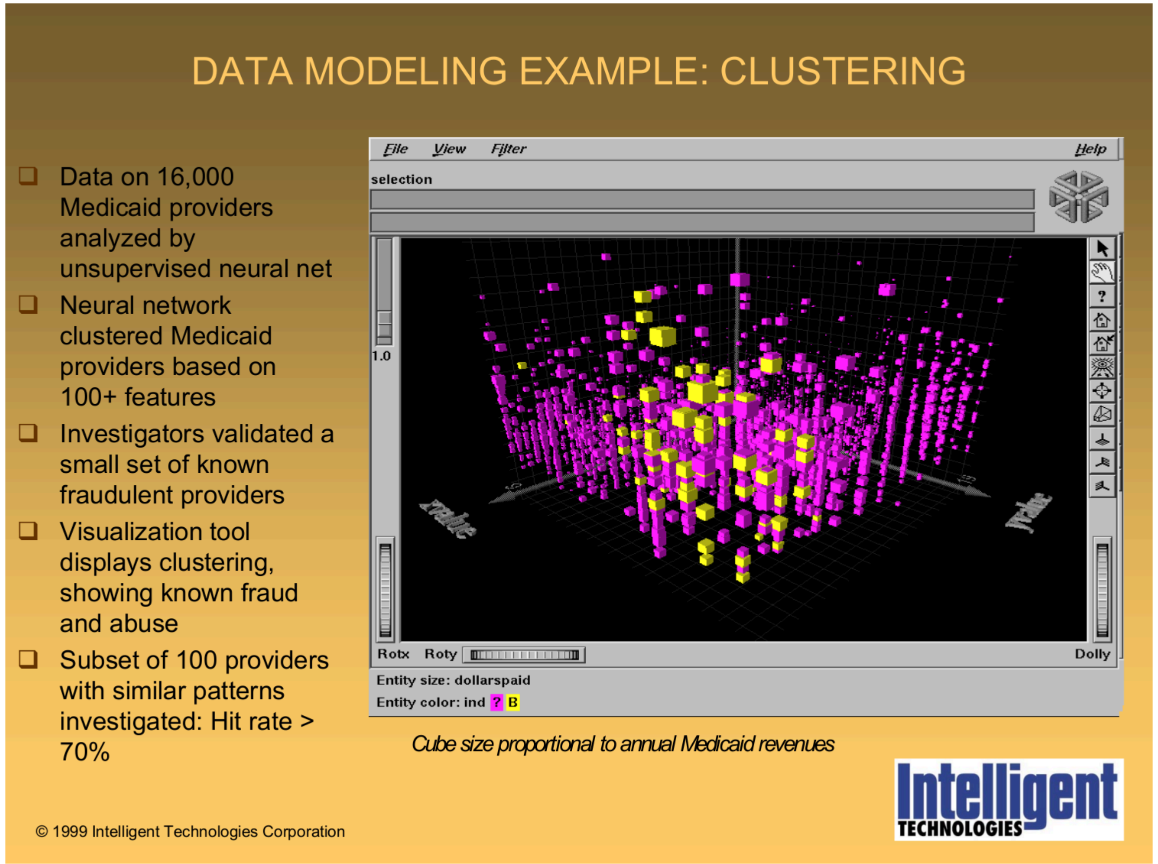
- No y_i, just sets of x_{ji}
- e.g. Cluster analysis
- Can only find associations between predictors

Cluster analysis is a form of unsupervised learning



- 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

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%

Entity size: dollarspaid

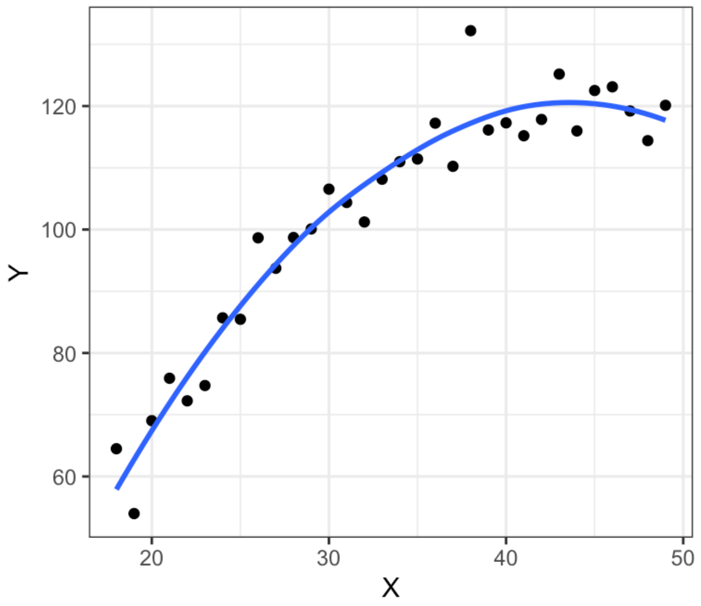
Entity color: ind ? B

Cube size proportional to annual Medicaid revenues

Intelligent TECHNOLOGIES

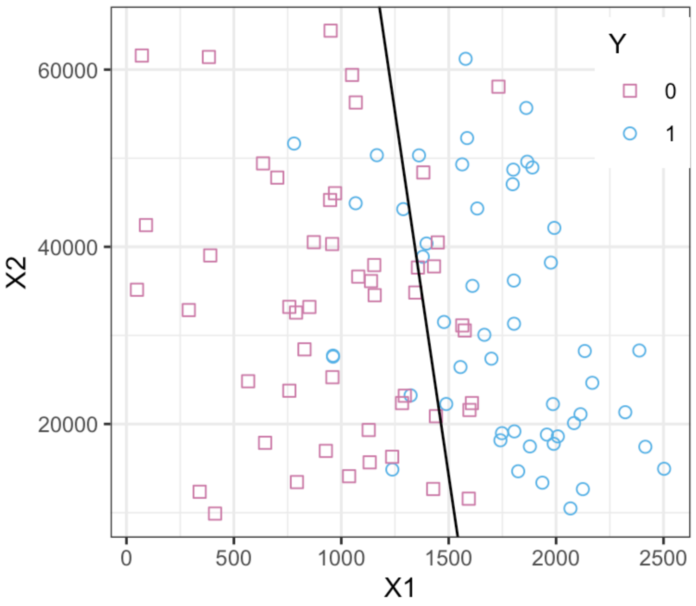
A note re Regression vs Classification problems

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

Lecture Outline

- Statistical learning
- **Assessing model accuracy**

Assessing model accuracy

- 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

- 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

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

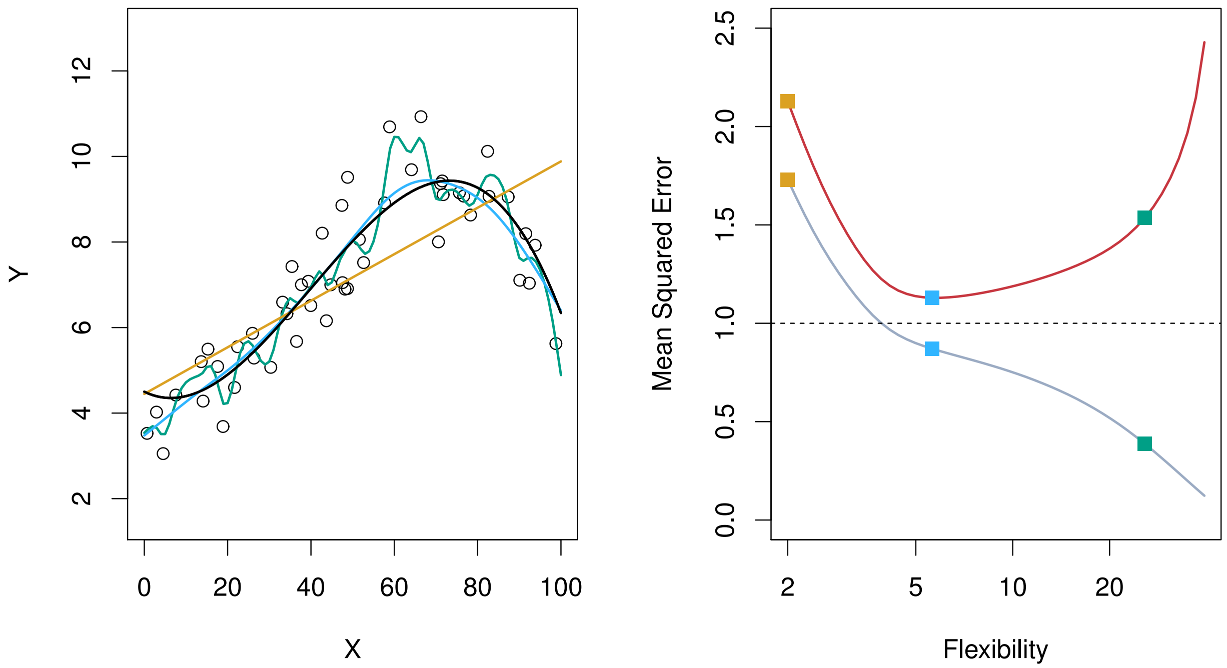
  - This should be small if predicted responses are close to true responses.
  - Note that if the MSE is computed using the data used to fit the model (which is called the training data) then this is more accurately referred to as the **training MSE**.

Discussion question

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

Discussion question

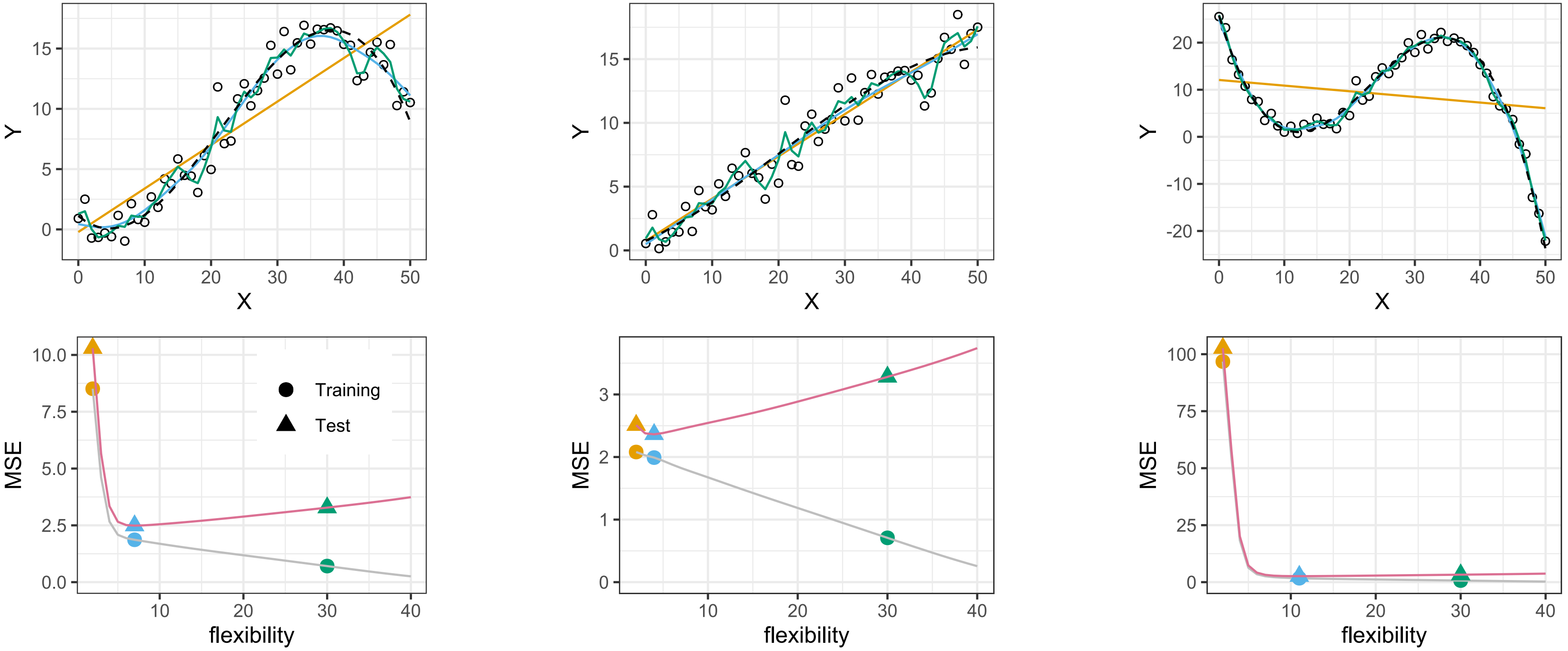
Consider the example below. The true model is black, and associated ‘test’ data are identified by circles. Three different fitted models are illustrated in blue, green, and orange. Which would you prefer?



e

Examples: Assessing model accuracy

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



Bias-Variance Tradeoff

The expected test MSE can be written as:

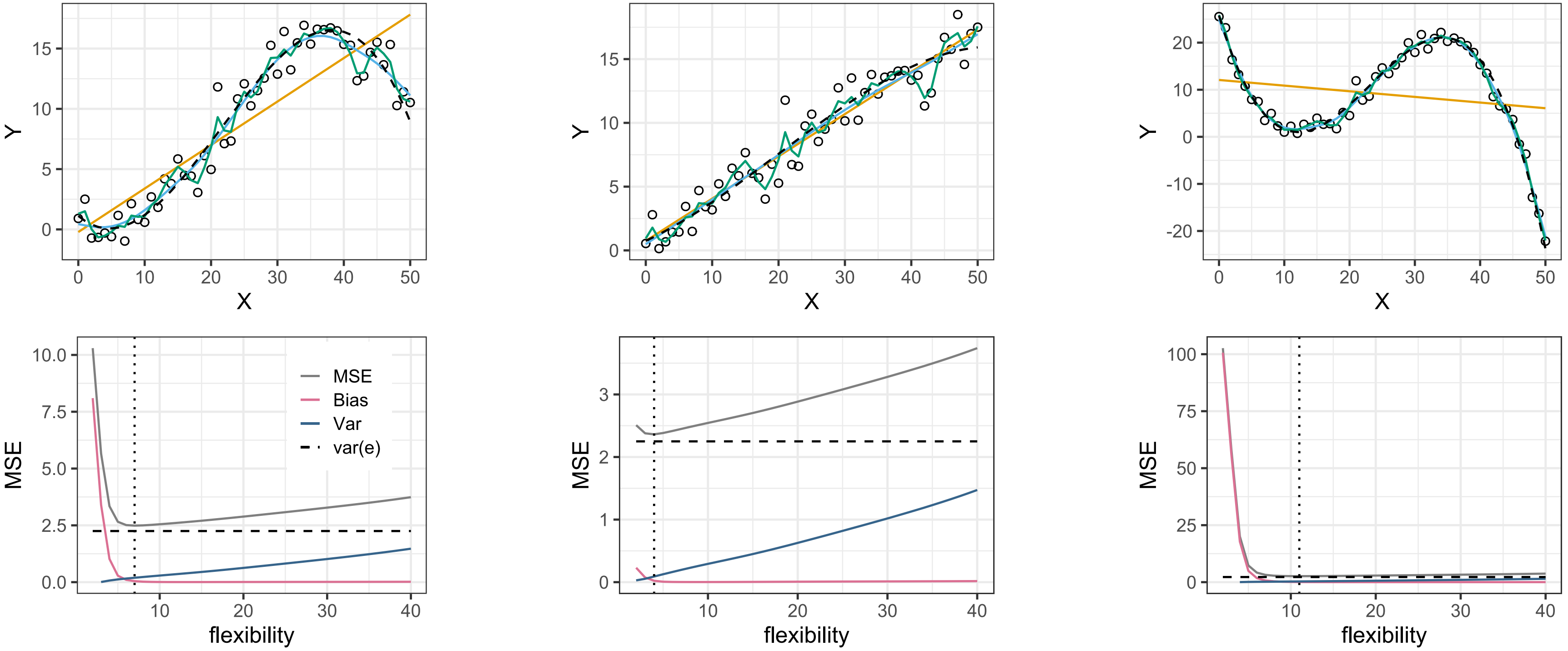
$$\mathbb{E}\big(y_0 - \hat{f}(x_0)\big)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

- $\mathrm{Var}(\hat{f}(x_0))$: how much $\hat{f}$ would change if a different training set is used
- $[\mathrm{Bias}(\hat{f}(x_0))]^2$: how much the model is off by
- $\mathrm{Var}(\epsilon)$: irreducible error

There is often a tradeoff between Bias and Variance

Examples: Bias-variance tradeoff

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



Classification

Objective

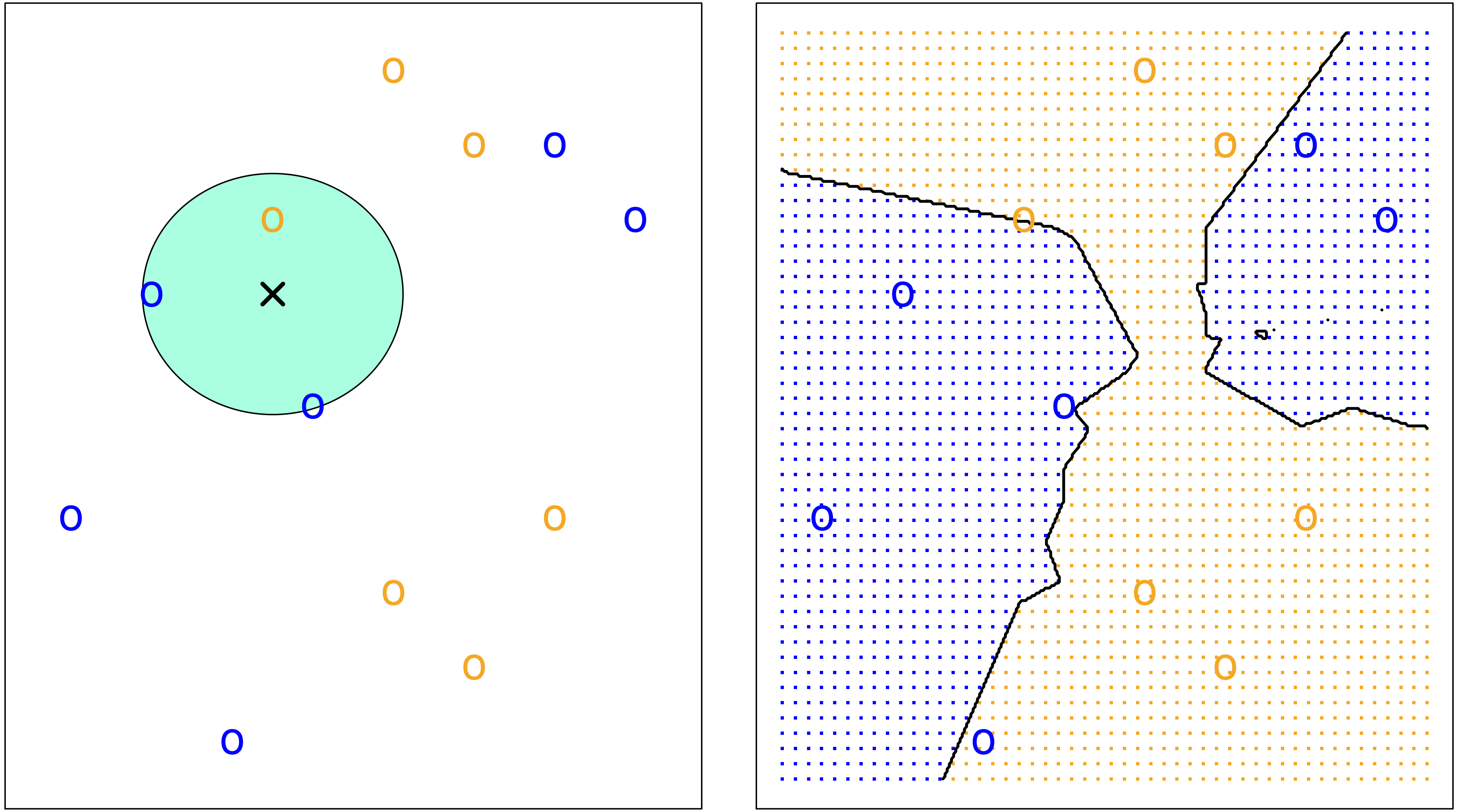
- Place data point into a category (Y) based on its predictors (X_i)
- Test Error is the proportion of times the estimate is wrong

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

Bayes' Classifier

- Assigns a prediction x_0 to the class j which maximises $\mathbb{P}(Y = j | X = x_0)$
- 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

K-nearest neighbours - illustration



e

K-nearest neighbours

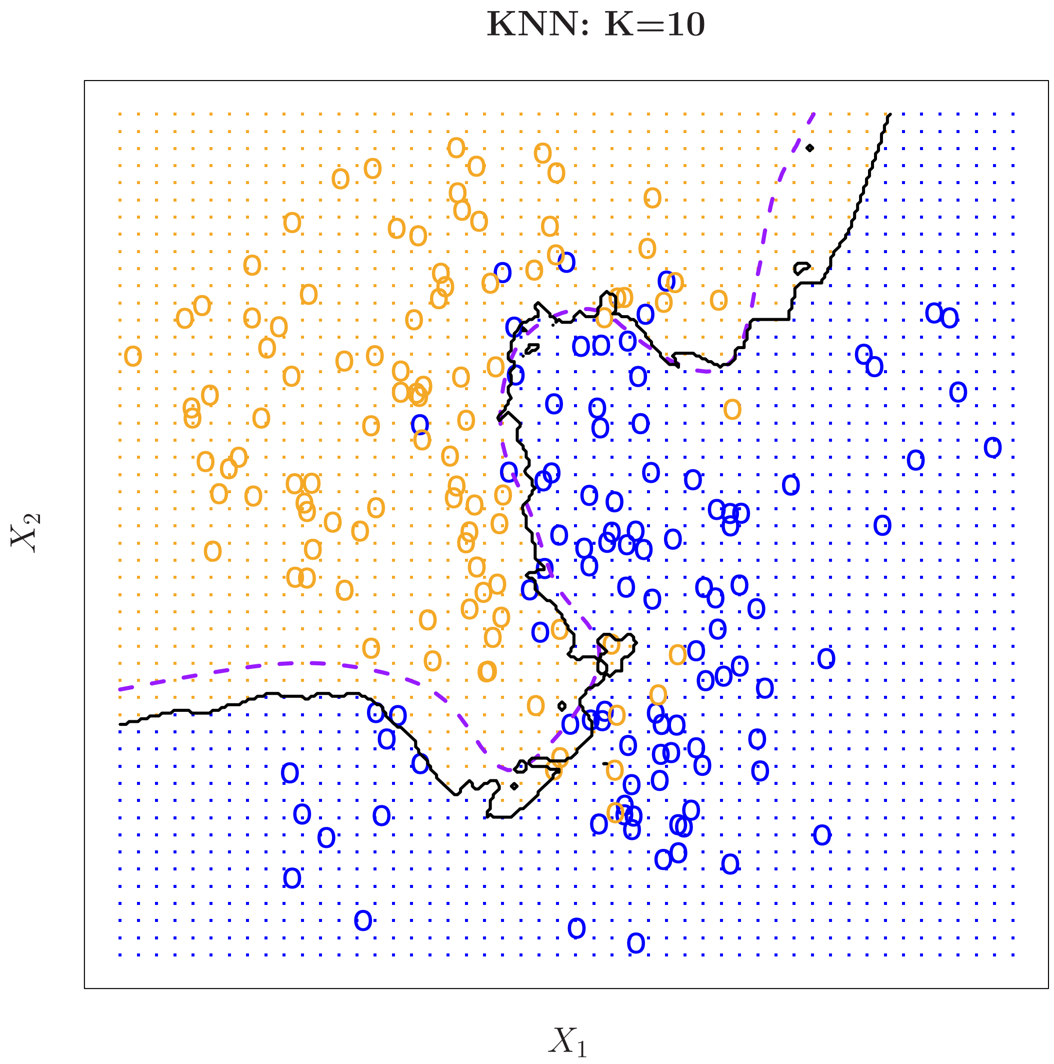
K-nearest neighbours

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

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

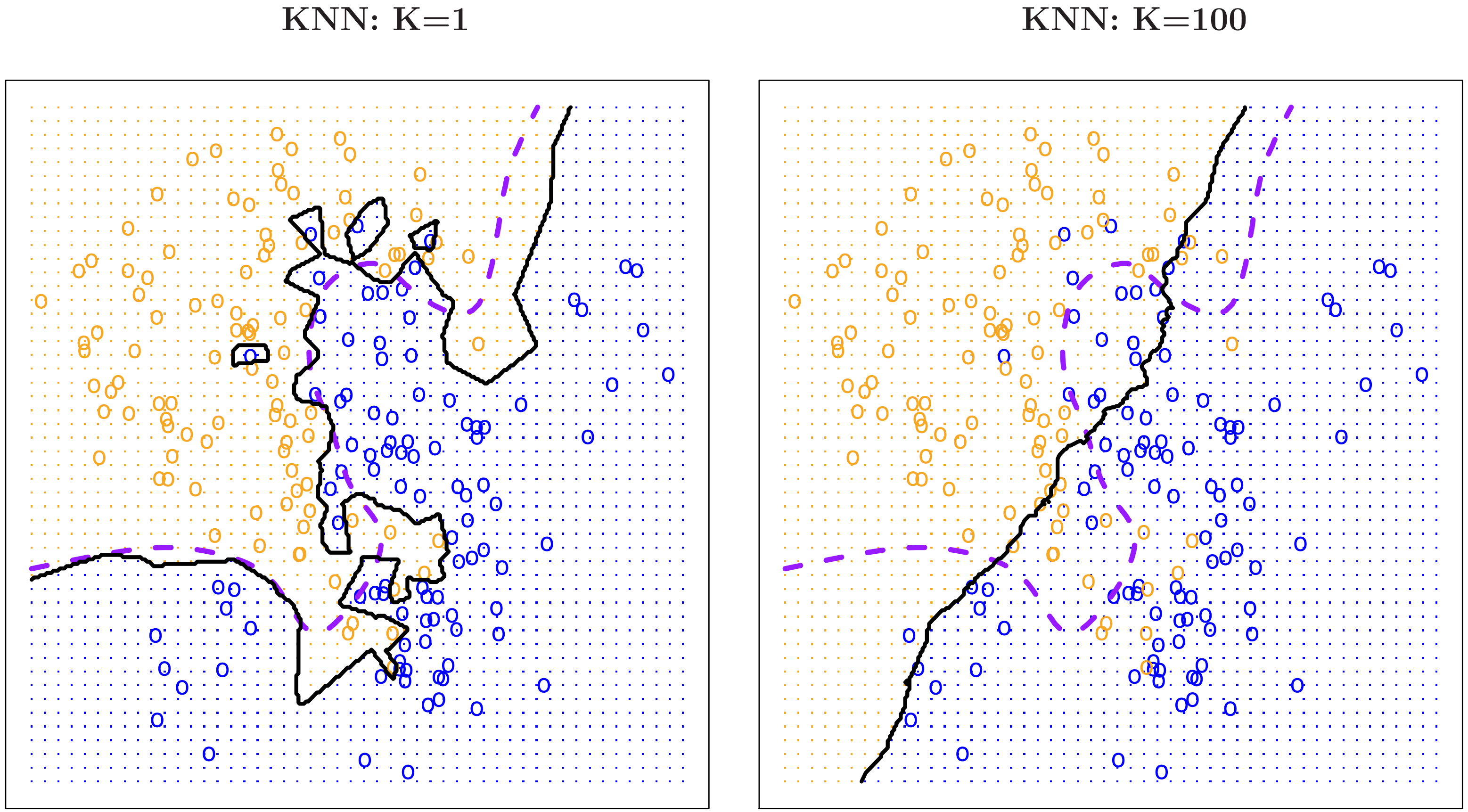
- New observation's category is where the majority of its neighbours lie
- High K: less variance but more bias, fit missing signal - too close to a global average
- Low K: less bias but more variance, fit too noisy - assuming less relationship between close-by data points than there is
- Intelligent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


KNN: K=1
KNN: K=100