CCMI Data-driven Modelling week

Introduction to data science and statistical inference

Vahid Shahrezaei
Department of Mathematics

IMPERIAL

"Intelligence is not just about pattern recognition and function approximation. It's about modeling the world". — Josh Tenenbaum, NeurIPS 2021.

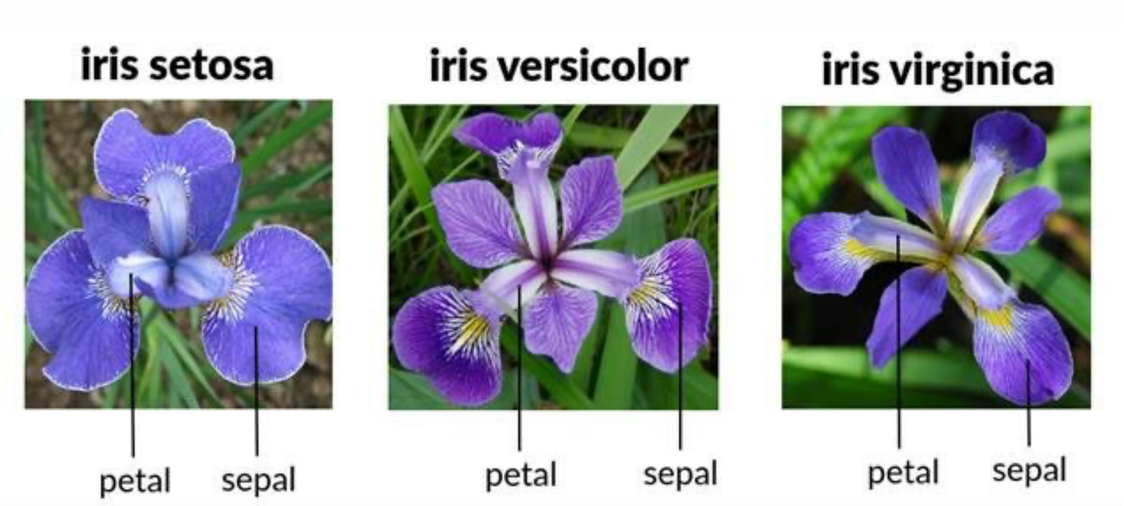
We live in the age of large and complex data that is ever more abundant for example in biomedicine such as single cell RNA-sequencing data that we work with later in this week

We also have large and complex mathematical models to help us make sense of the data and make predictions

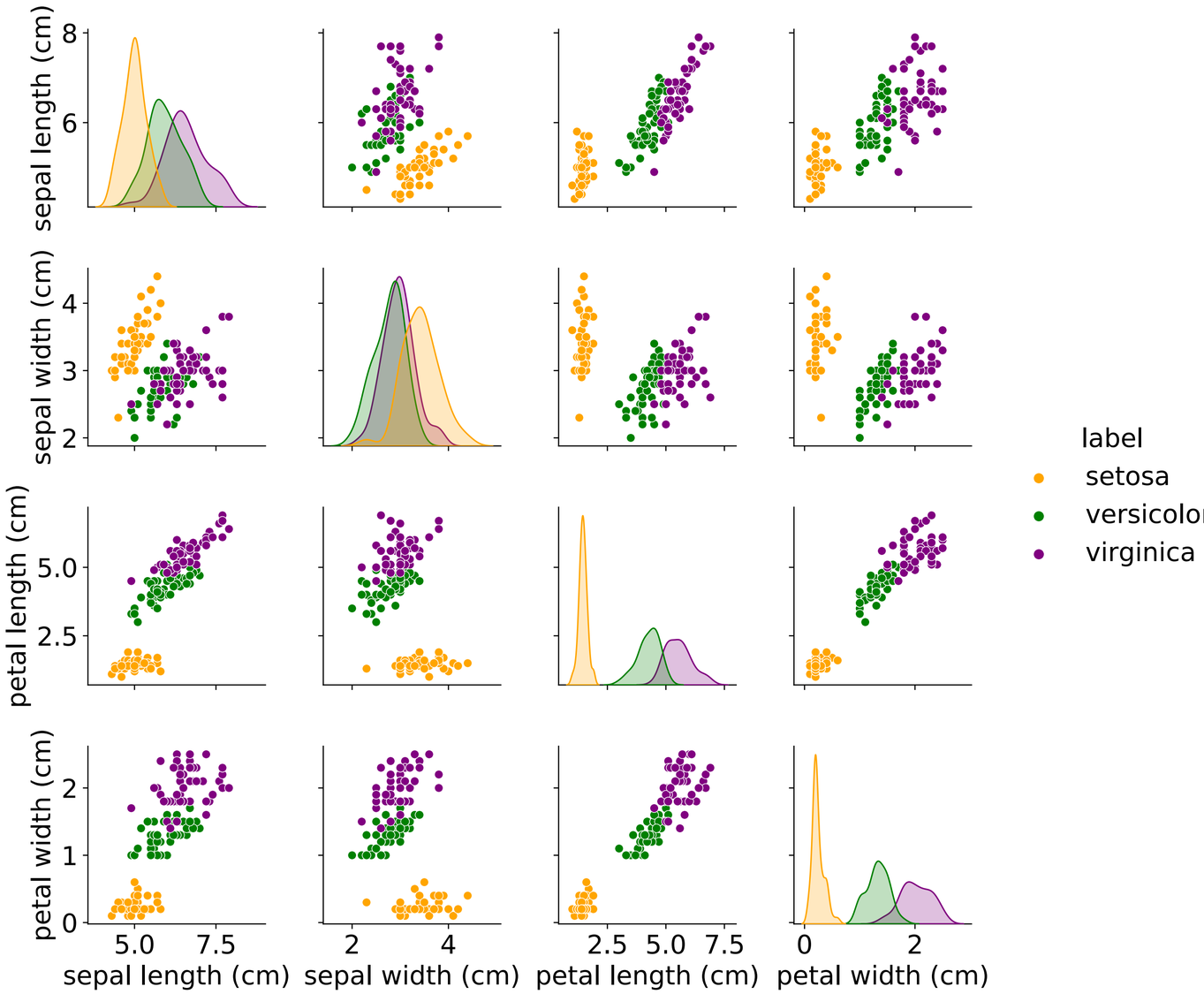
Blackbox vs Mechanistic models, prediction vs understanding!

Exploratory data analysis: visualisation

e.g. pair-plots for low dimensional tabular data

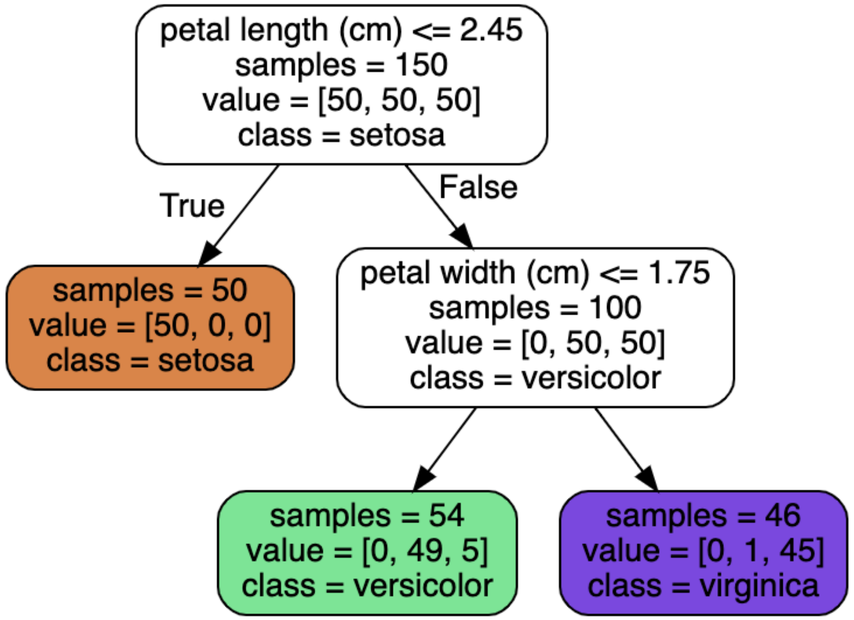


https://doi.org/10.26438/ijsrmss/v9i6.110

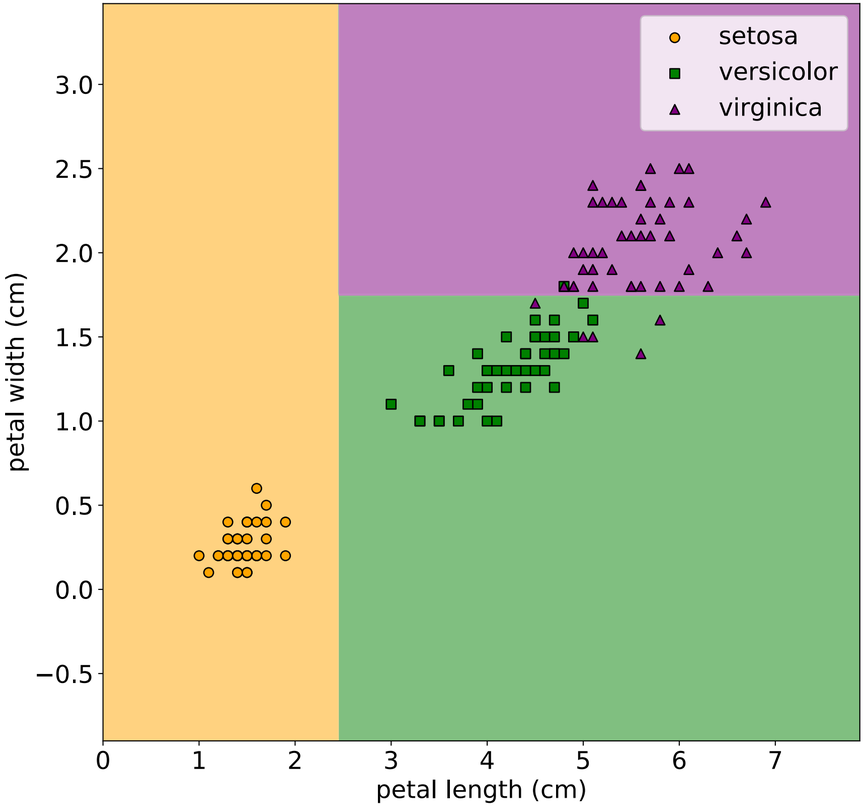


Supervised learning: Classification

Can we learn/predict the labels from the features?

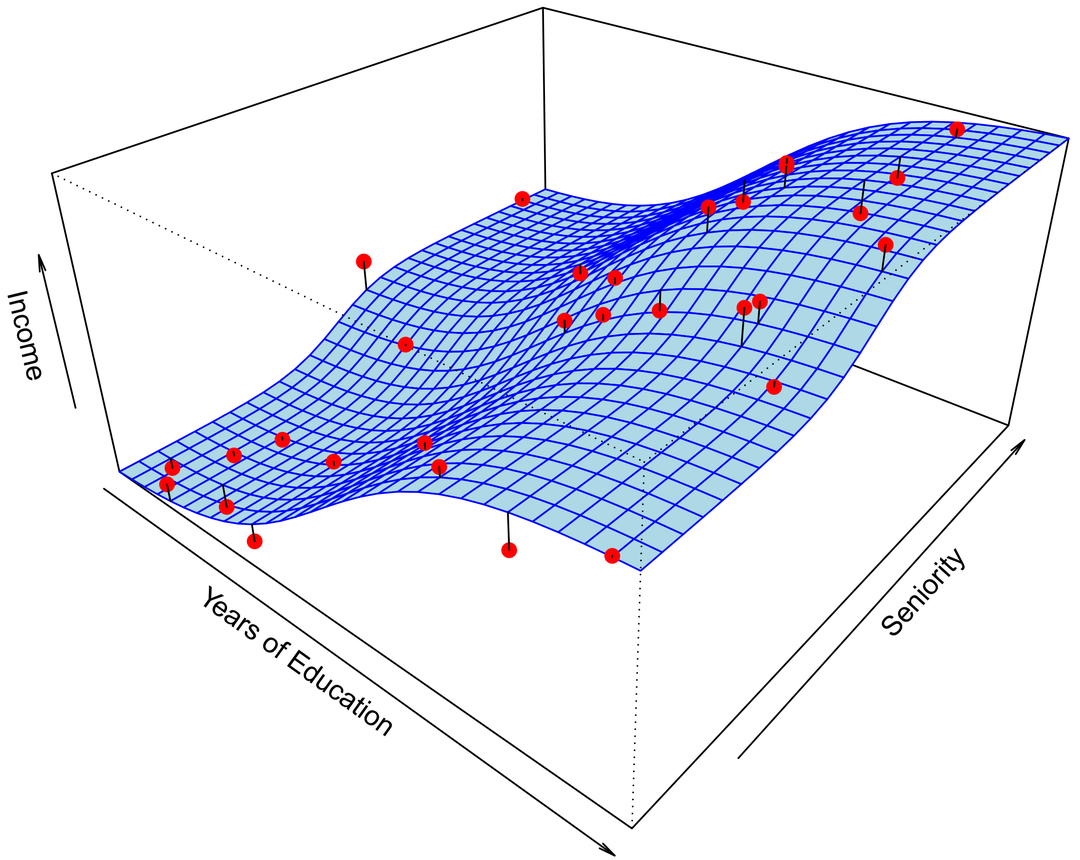


(a)



(b)

Supervised learning: Regression



A set of predictors: $X = (X_1, X_2, \ldots, X_p)$

A quantitative response: $Y = f(X) + \epsilon$

If we have some observations (training data), we can estimate function *f(X)* using a statistical learning algorithm.

Why do we want to estimate function *f(X)*?

1. To make predictions

$$\hat{Y} = \hat{f}(X)$$

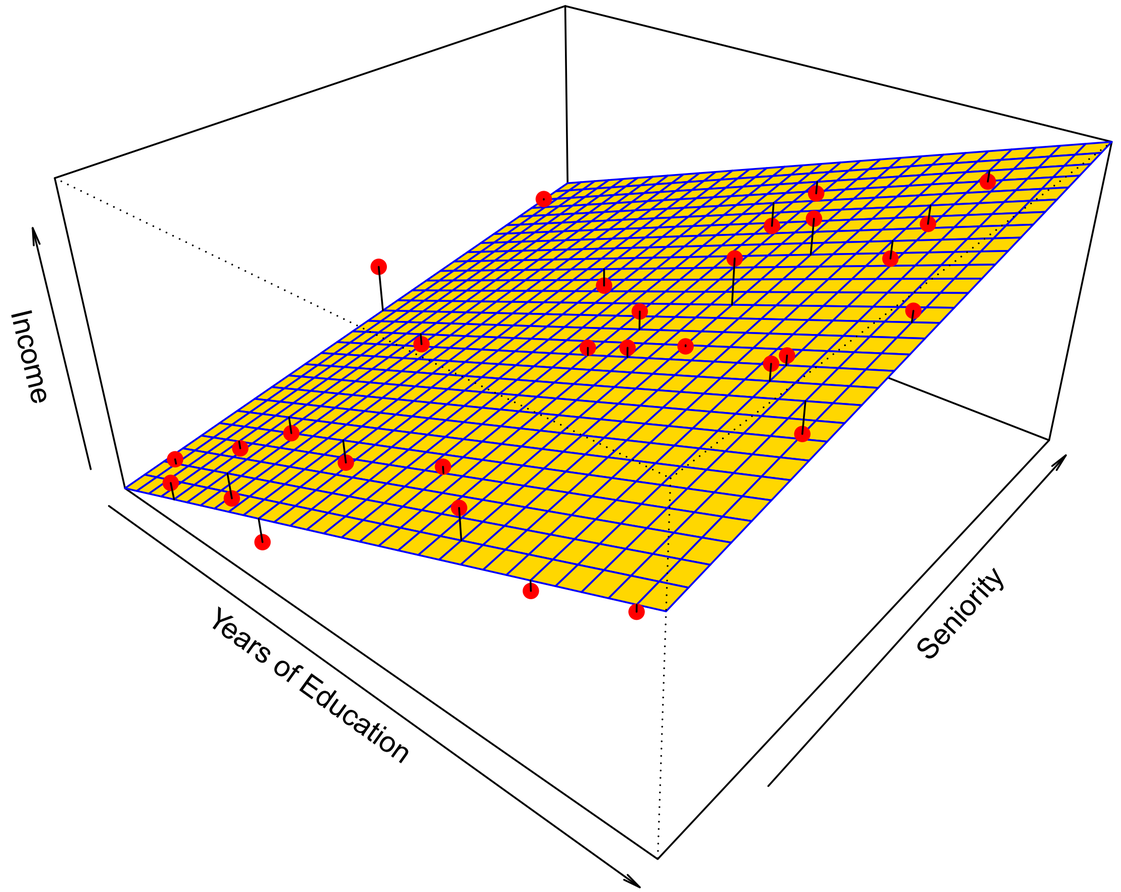
$$\begin{aligned} \mathrm{E}(Y - \hat{Y})^2 &= \mathrm{E}[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

2. To make inference

- How strongly a particular feature *X* controls the response *Y* (Parameter inference?)
- What is the overall mechanism that shapes the response (Model selection)?
- How accurate are the predictions (confidence intervals)?
- How much do we learn from the data about *f(X)*? (non-identifiability)?

How to obtain estimates of *f(X)*?

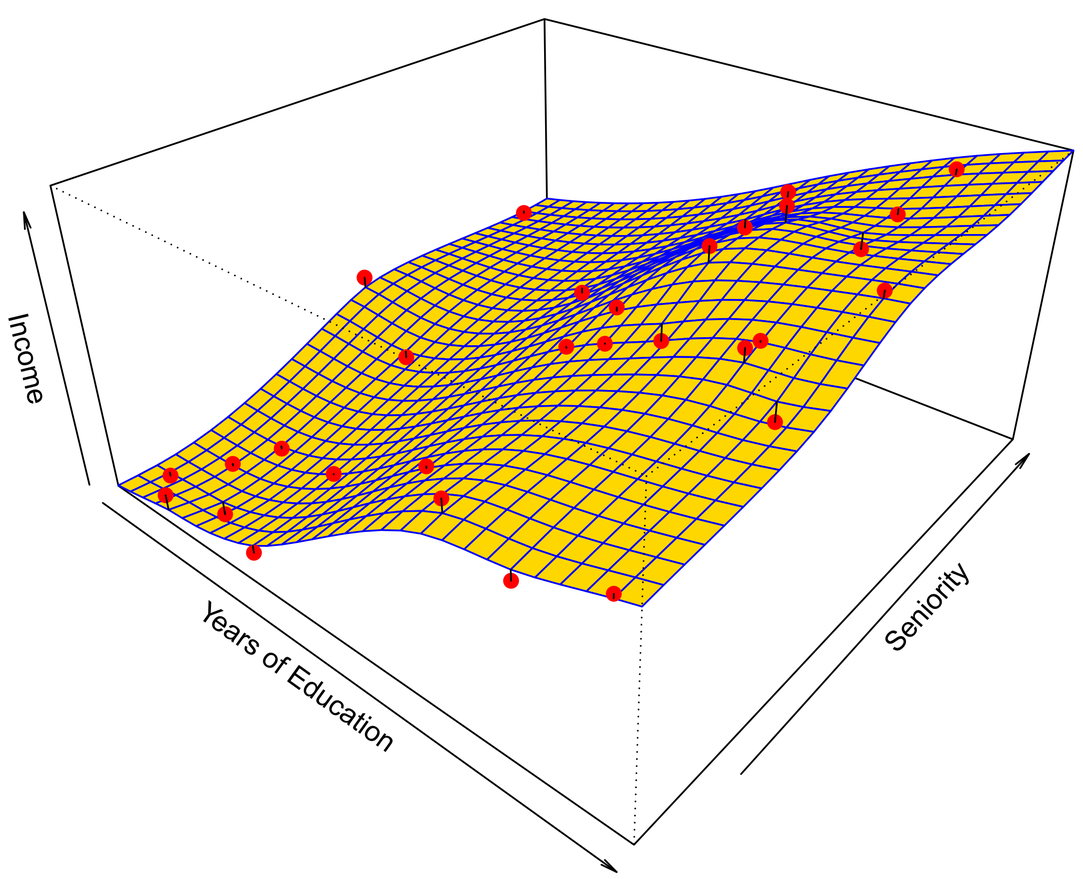
Using a parametric model, e.g. a linear model

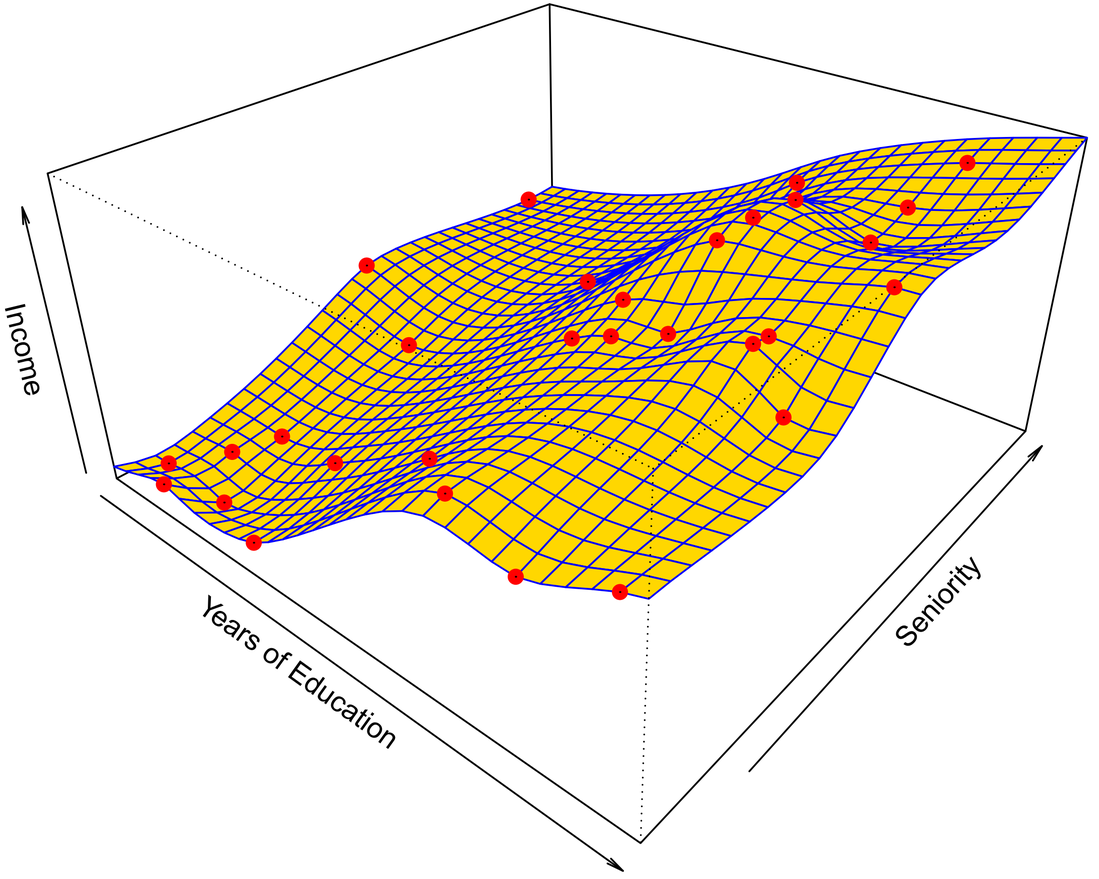


$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}.$$

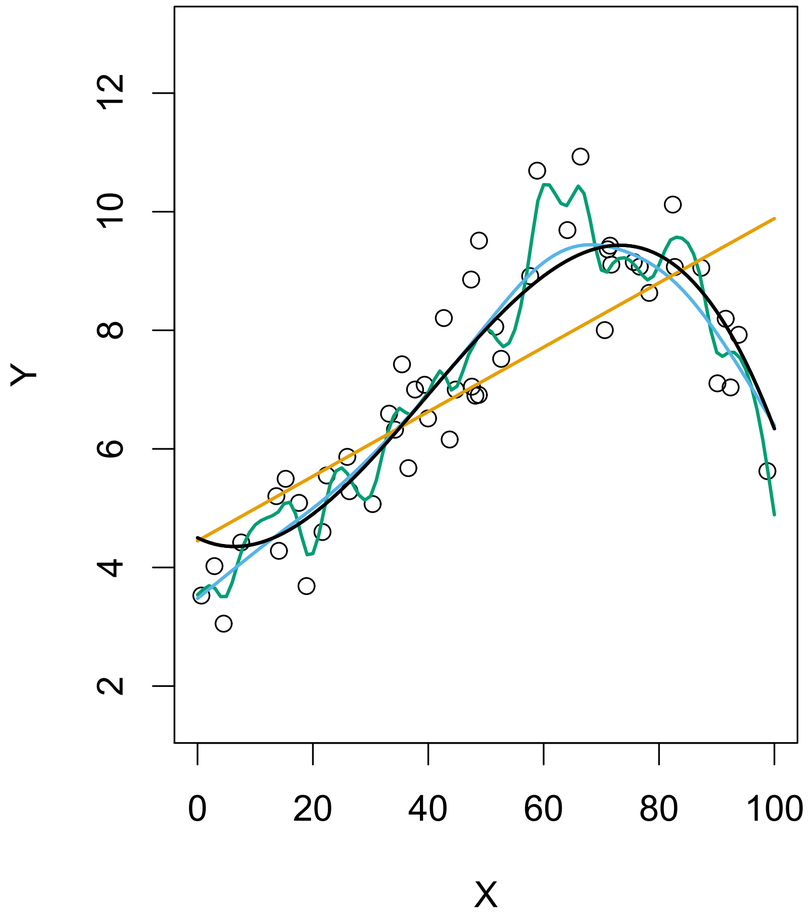
How to obtain even better estimates of *f(X)*?

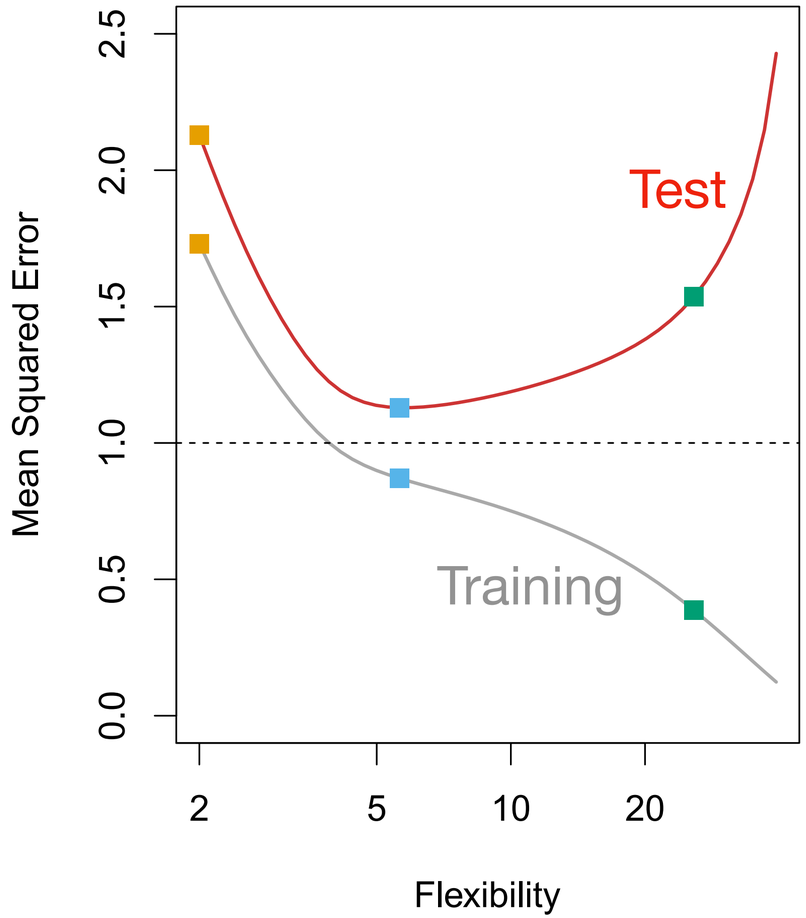
Using a non-parametric model, e.g. a thin-plate spline fit





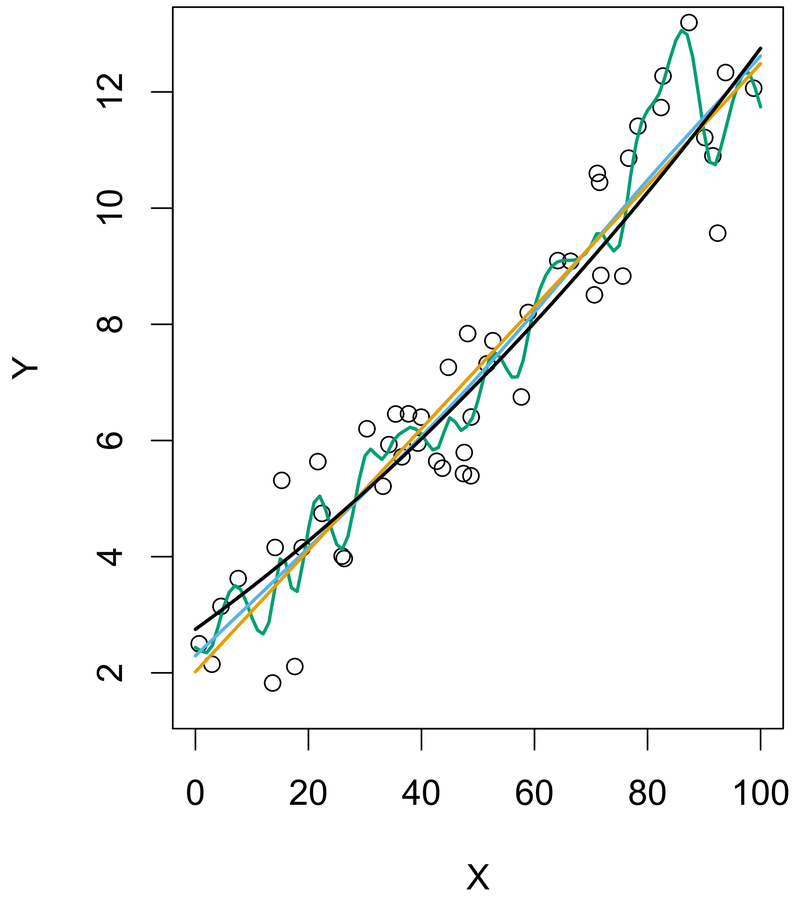
Danger of over-fitting, using training and test data

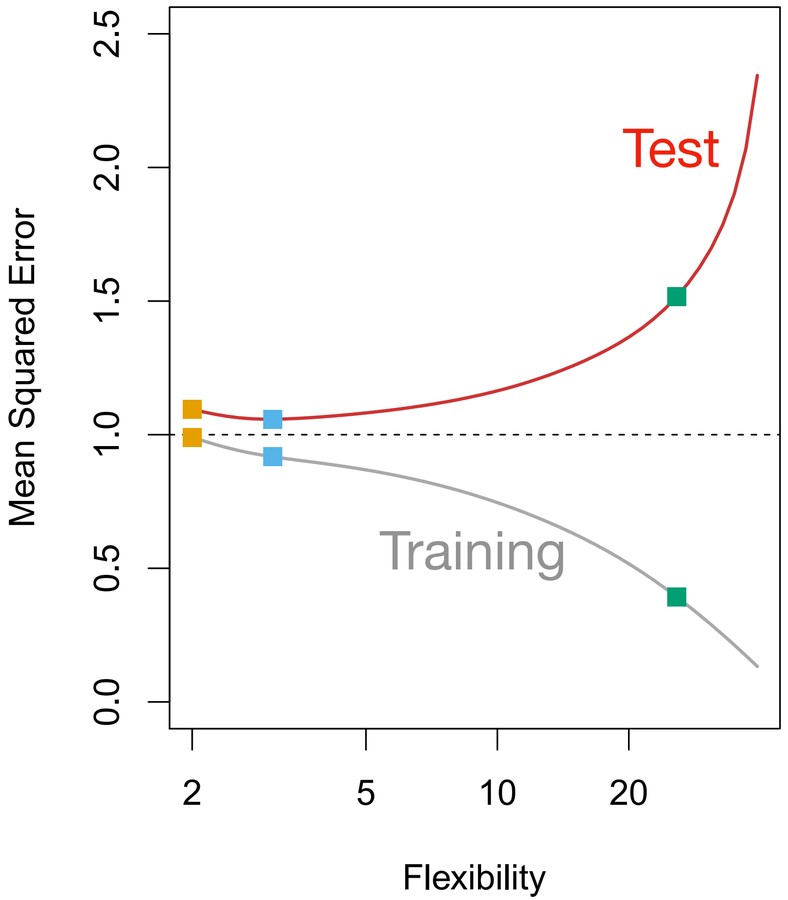




$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

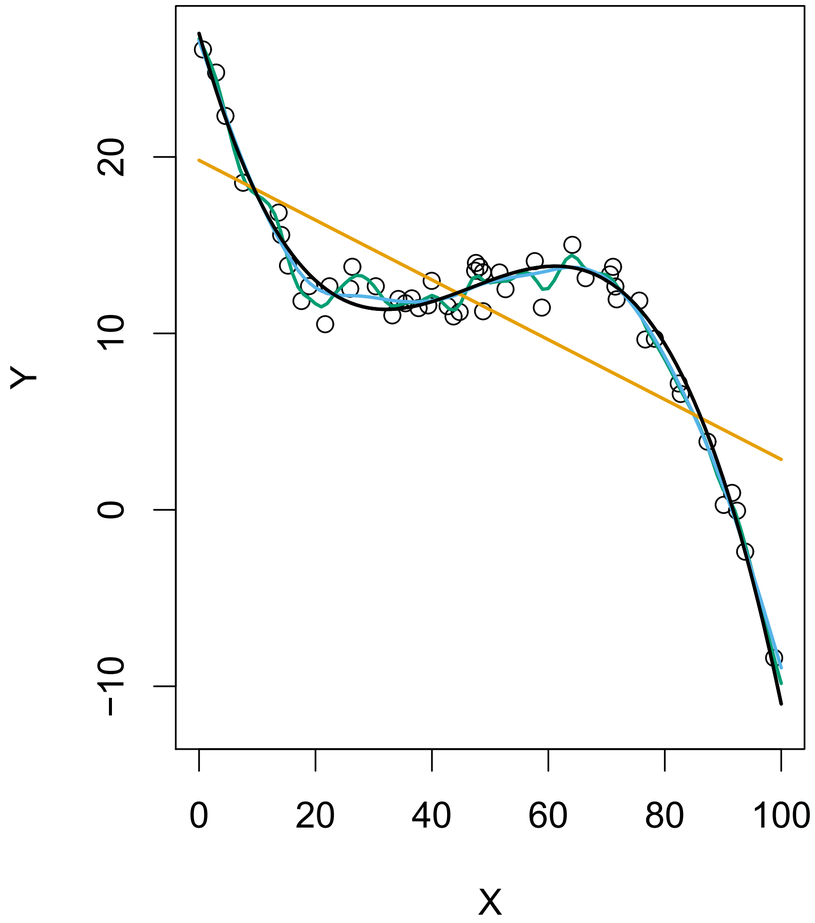
Danger of over-fitting, using training and test data

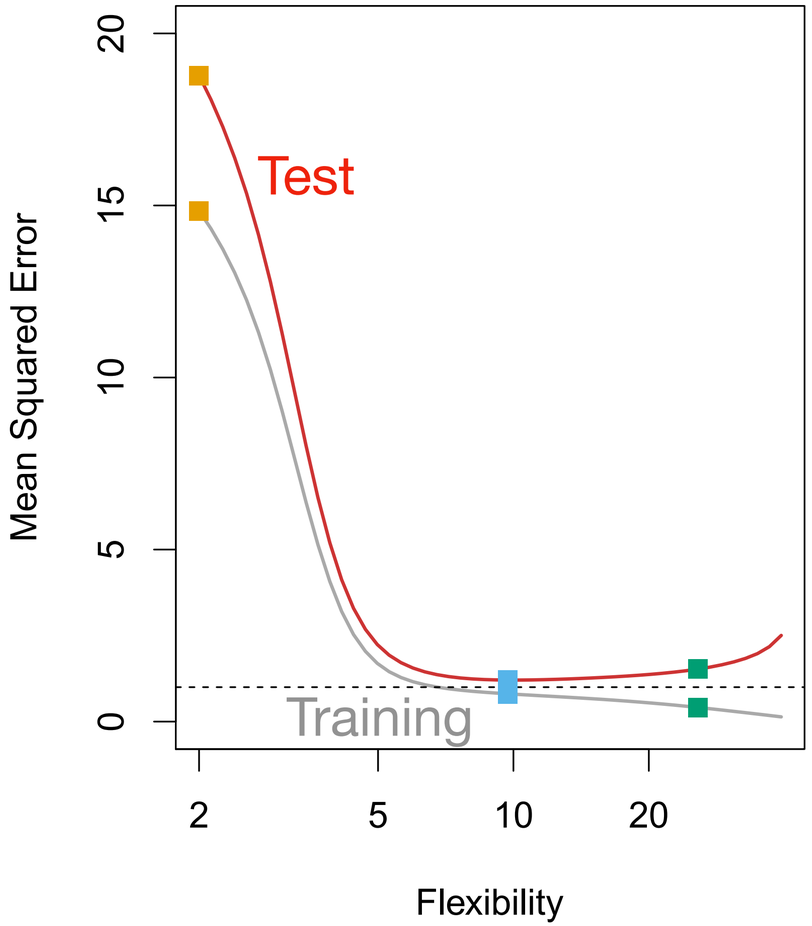




$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Danger of over-fitting, using training and test data

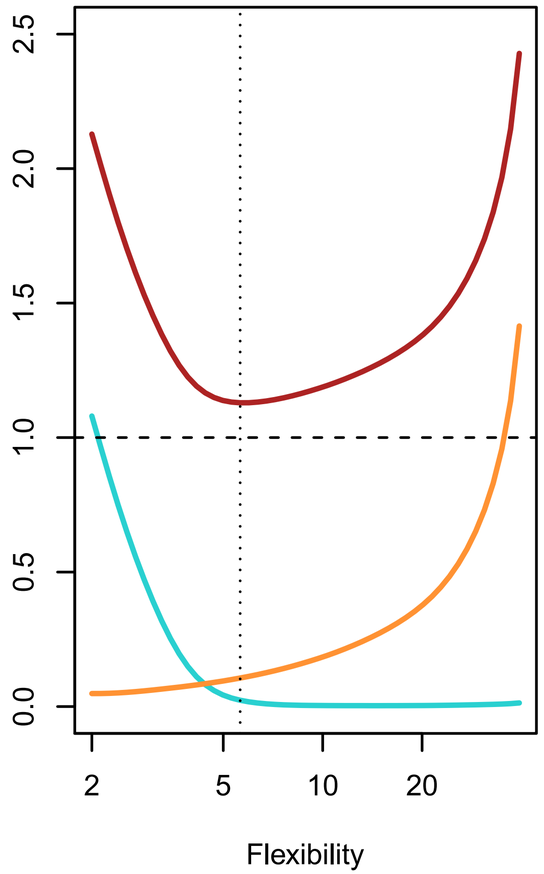


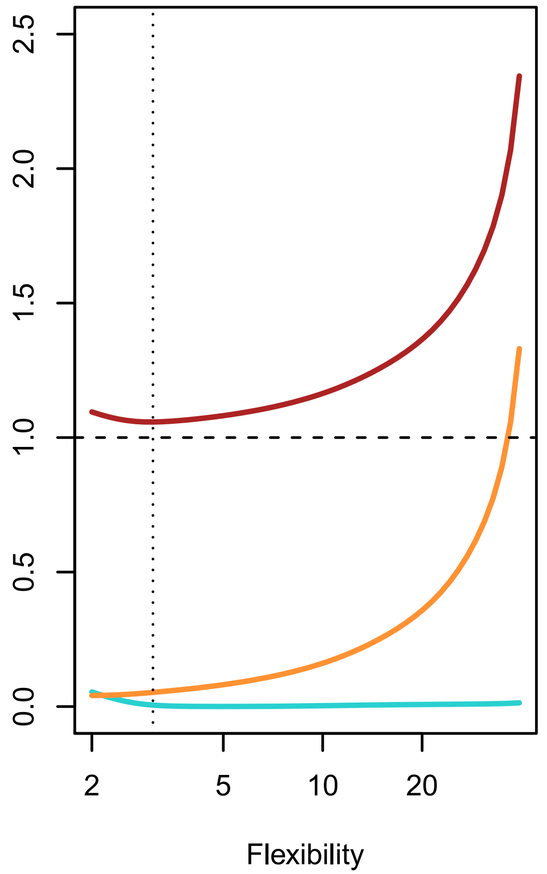


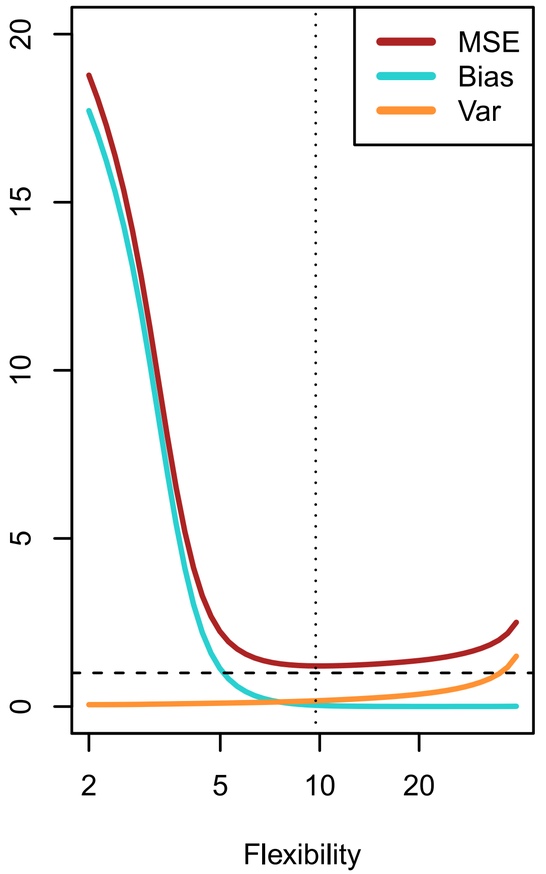
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

A bias-variance trade-off

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

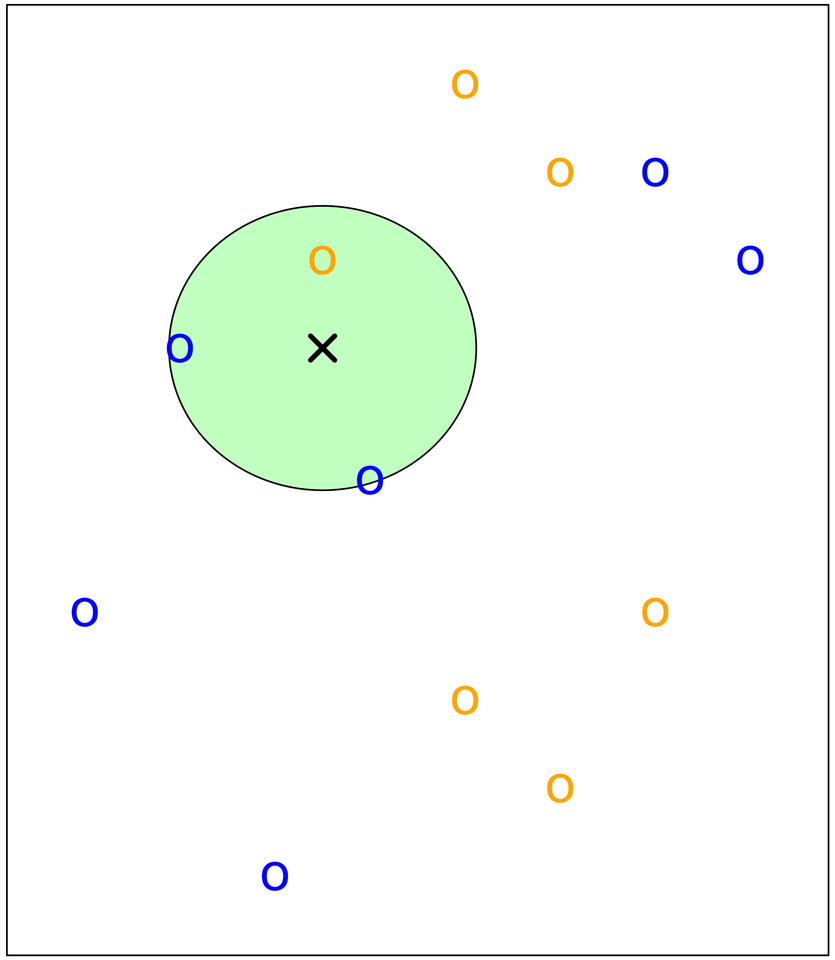


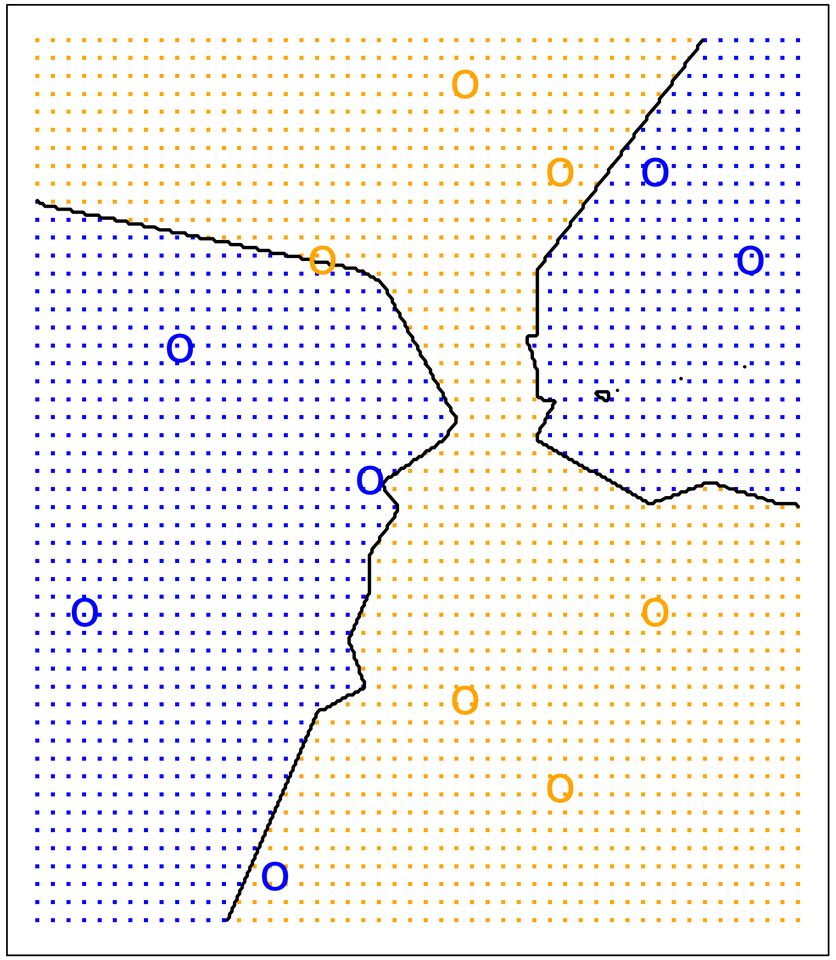




Supervised learning: Classification

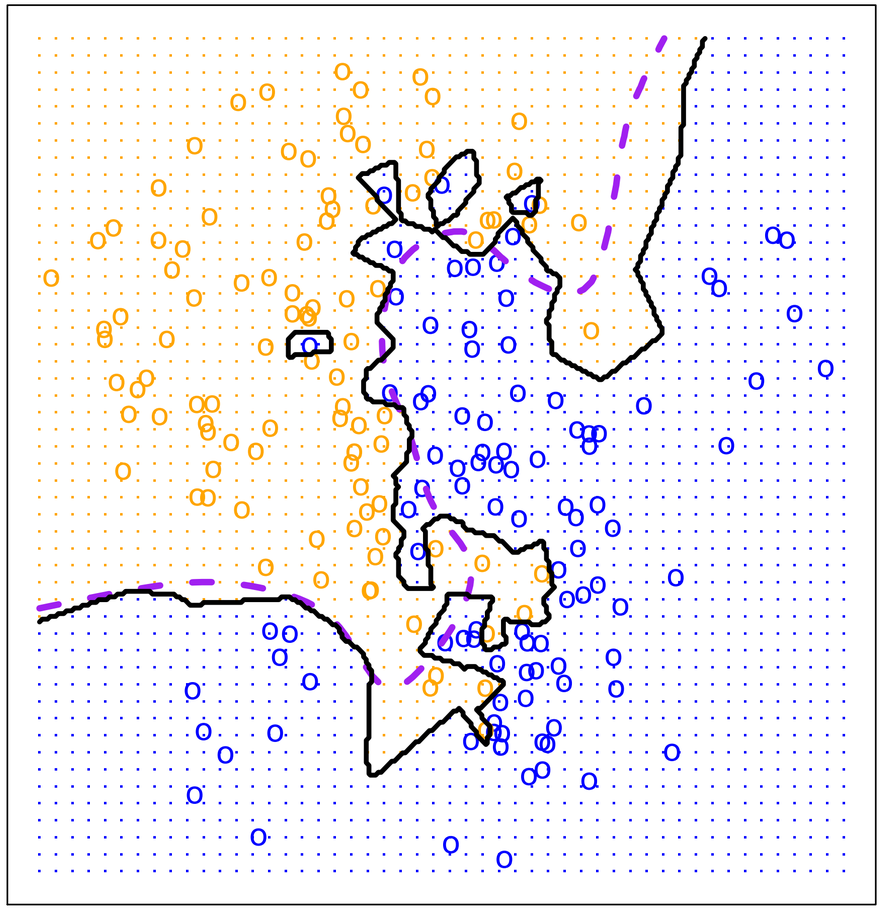
K-nearest neighbours (Knn) classifier



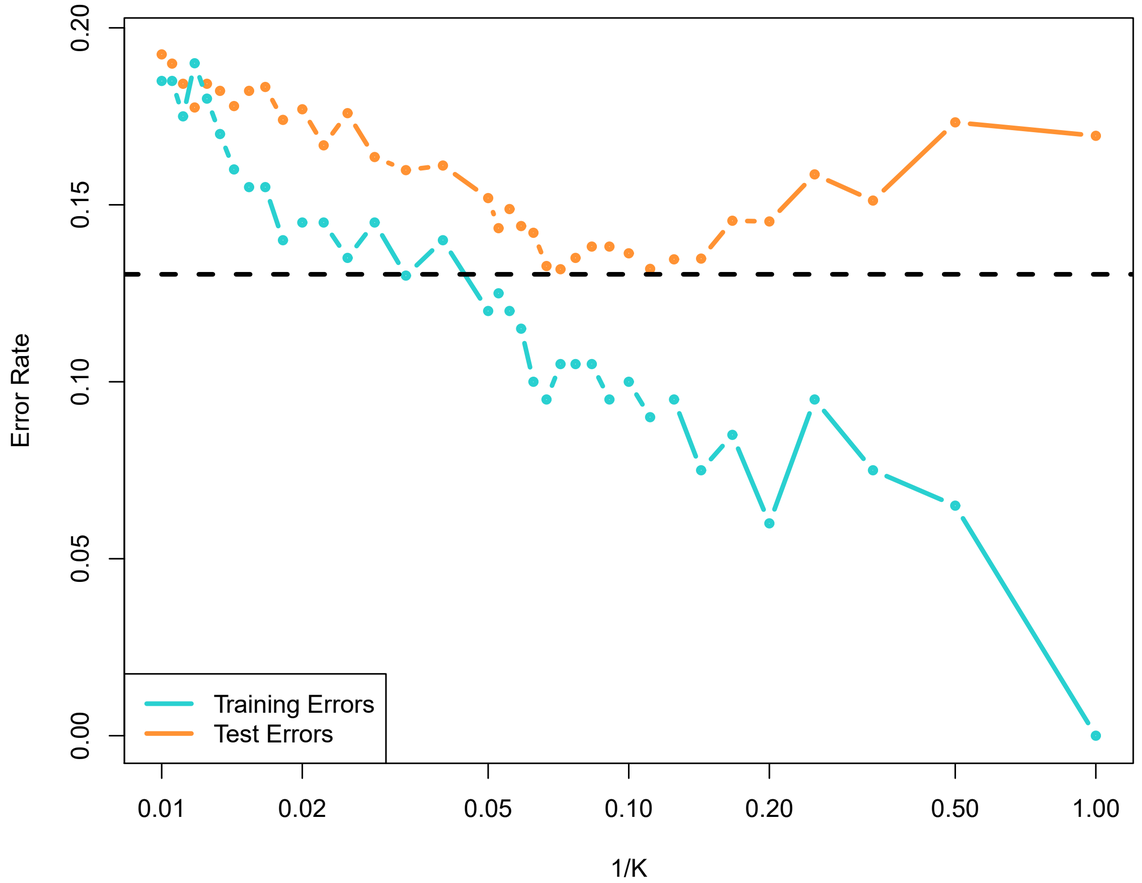


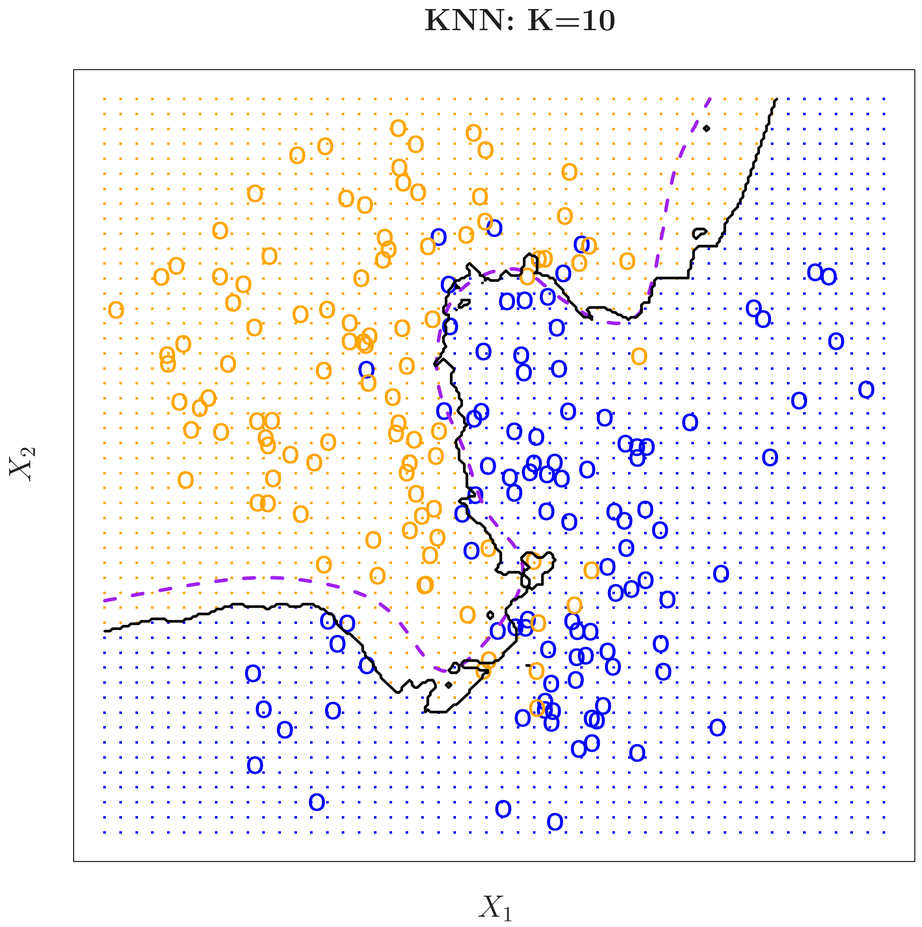
K = 3

KNN: K=1



The bias-variance trade-off



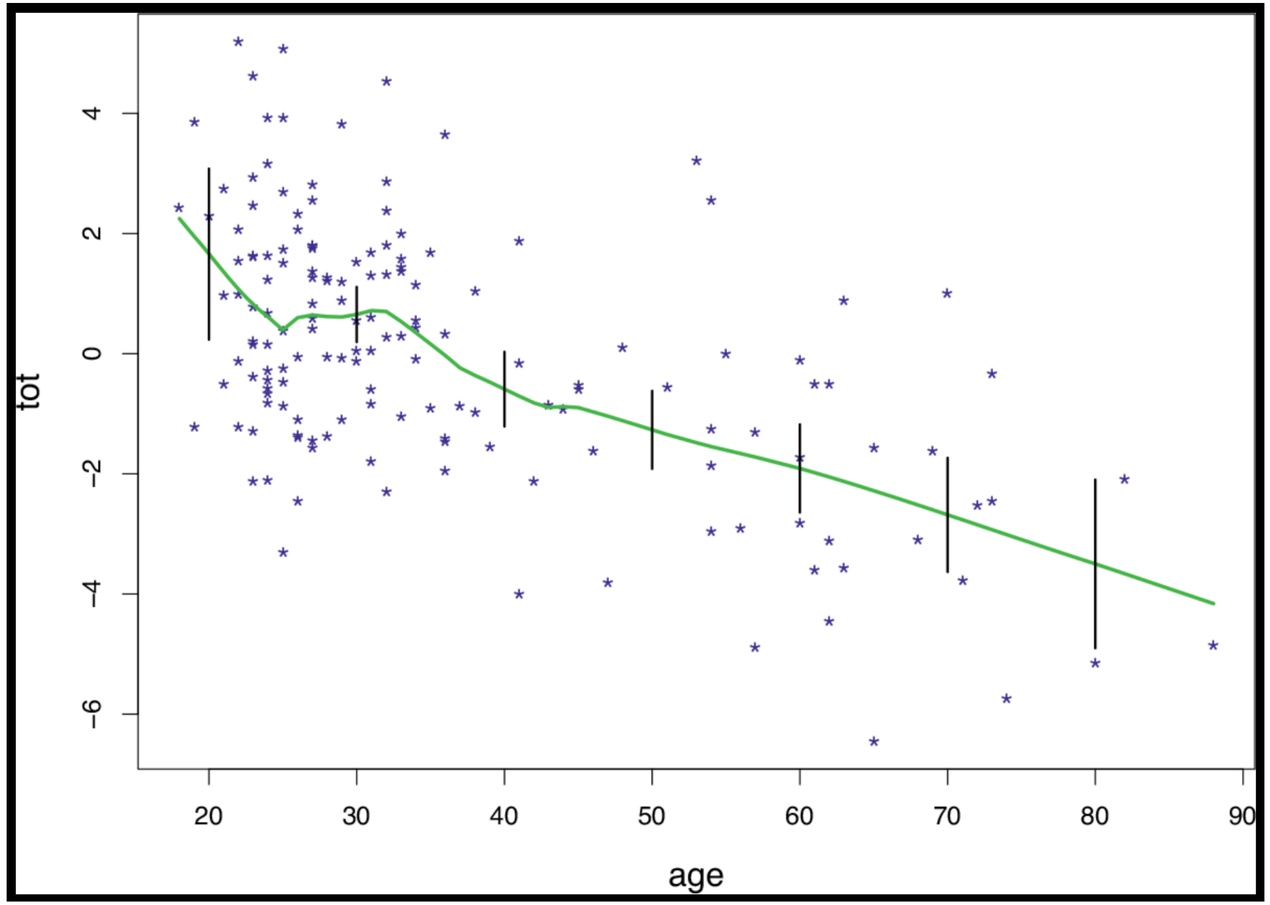


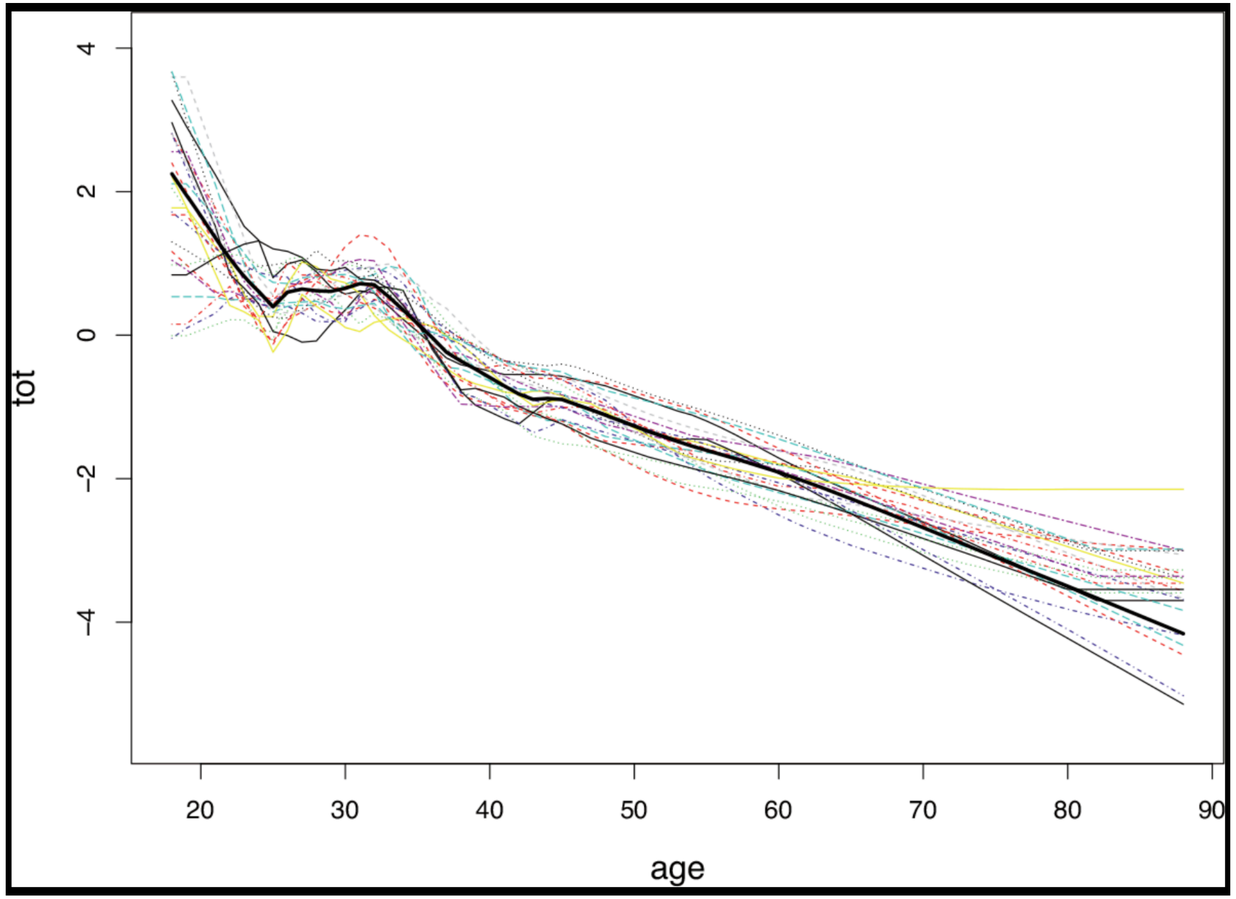
Parametric and non-parametric models/methods

Parametric models: Have fixed assumptions and a (small) fixed number of parameters. They are less flexible, so could produce larger bias, but easier to fit and require less data. There are better in providing inference and understanding. This is specifically the case for a particular type of parametric models we name **mechanistic models.**

Non-parametric models/methods: Do not make specific assumptions such as distributional ones, they are more flexible but can overfit. They do have parameters but the number of parameters are usually not fixed and increase with the complexity of data. They need more data to fit typically and are less explainable but could be good at predictions. Some non-parametric methods such as resampling methods could be applied to almost any problem, so one does not need to know a lot of classical statistics!

Bootstrap can produce confidence intervals for any (non-parametric) method





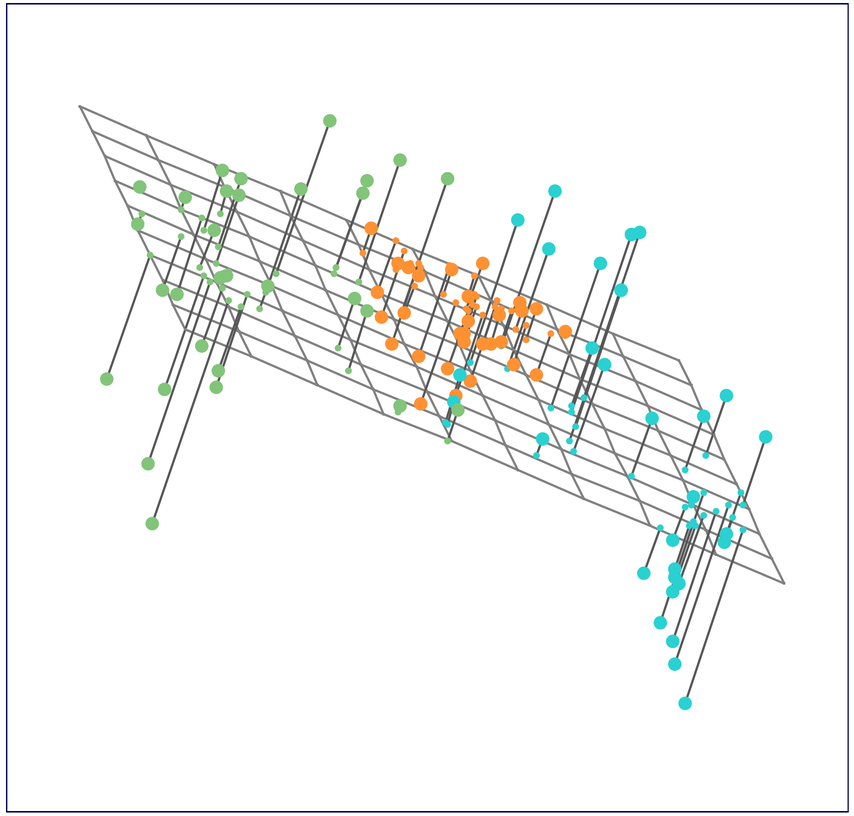
Unsupervised learning

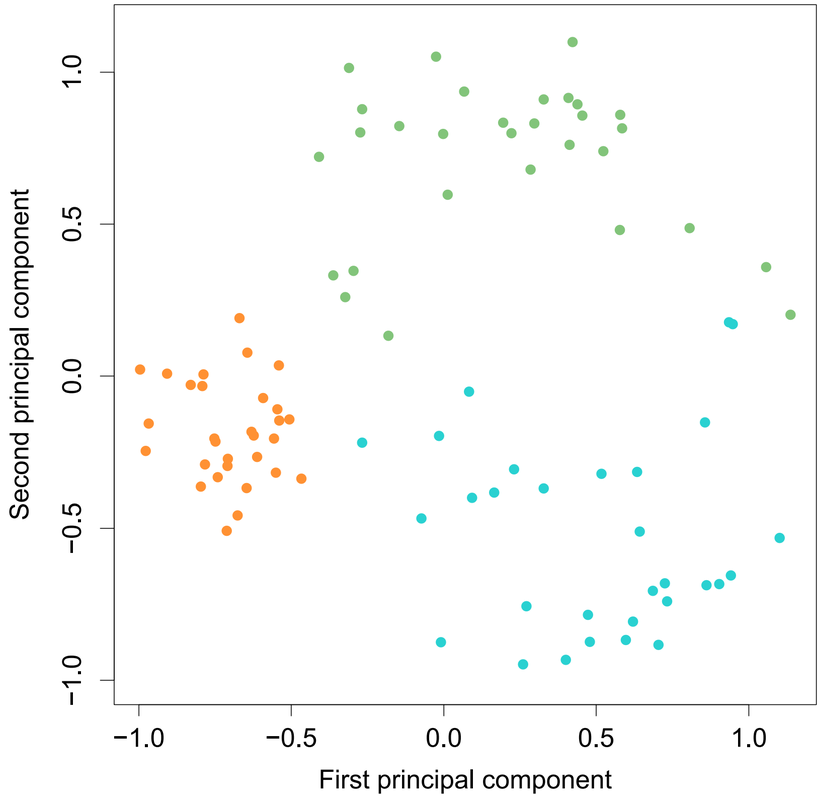
It is harder to assess the accuracy as there are no training/test data

Can be used for exploratory data analysis and/or as preprocessing for supervised learning methods

- Density estimation
  (non-parametric methods e.g. KDE, parametric methods e.g. GMM)
- Clustering
- Dimensional reduction
- Outlier (anomaly) detection

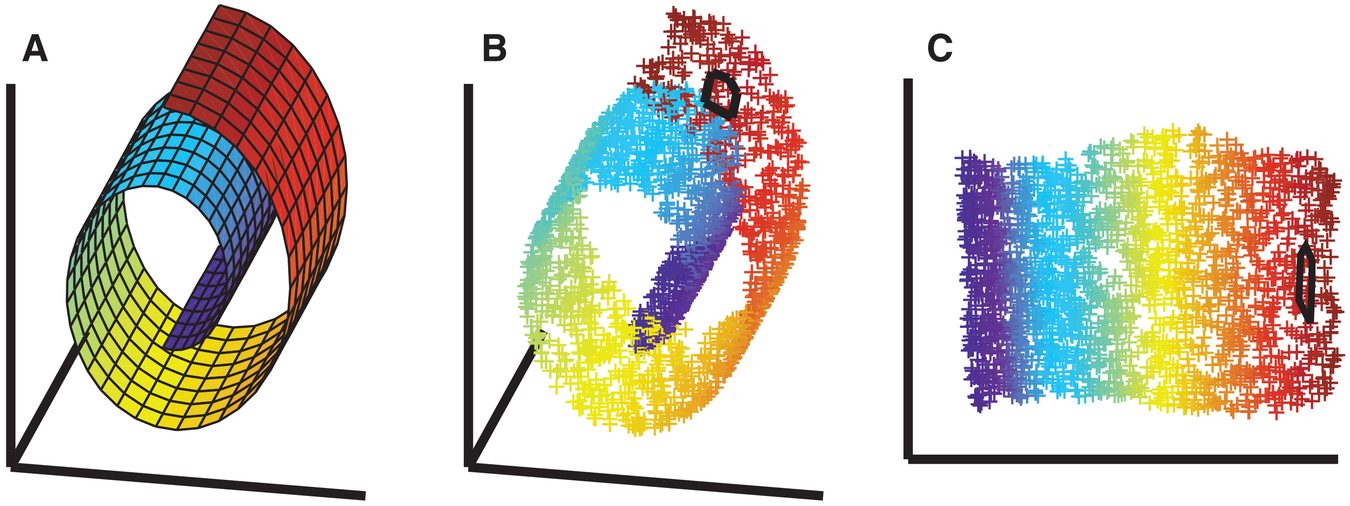
Dimensionality reduction: Principal components analysis (PCA)

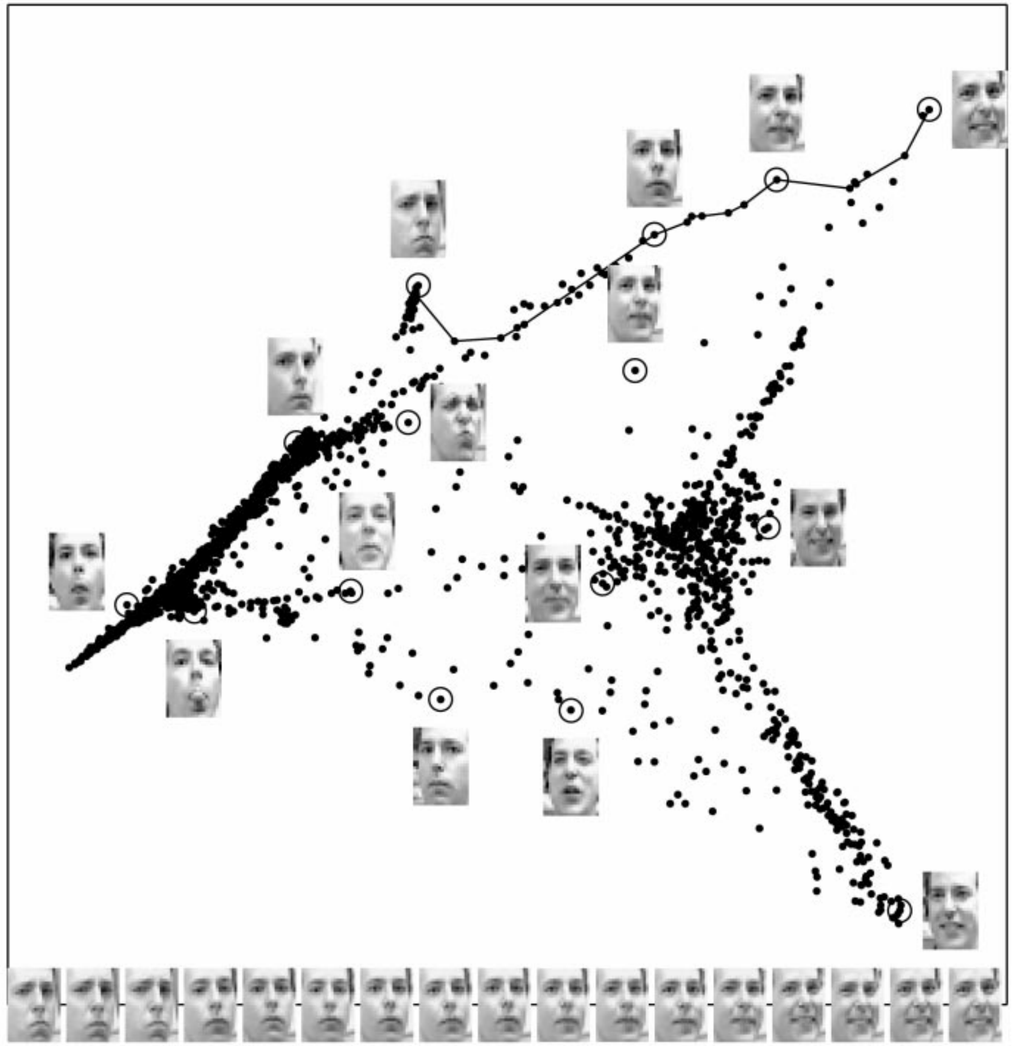




Nonlinear Dimensionality reduction: Locally linear embedding (LLE)

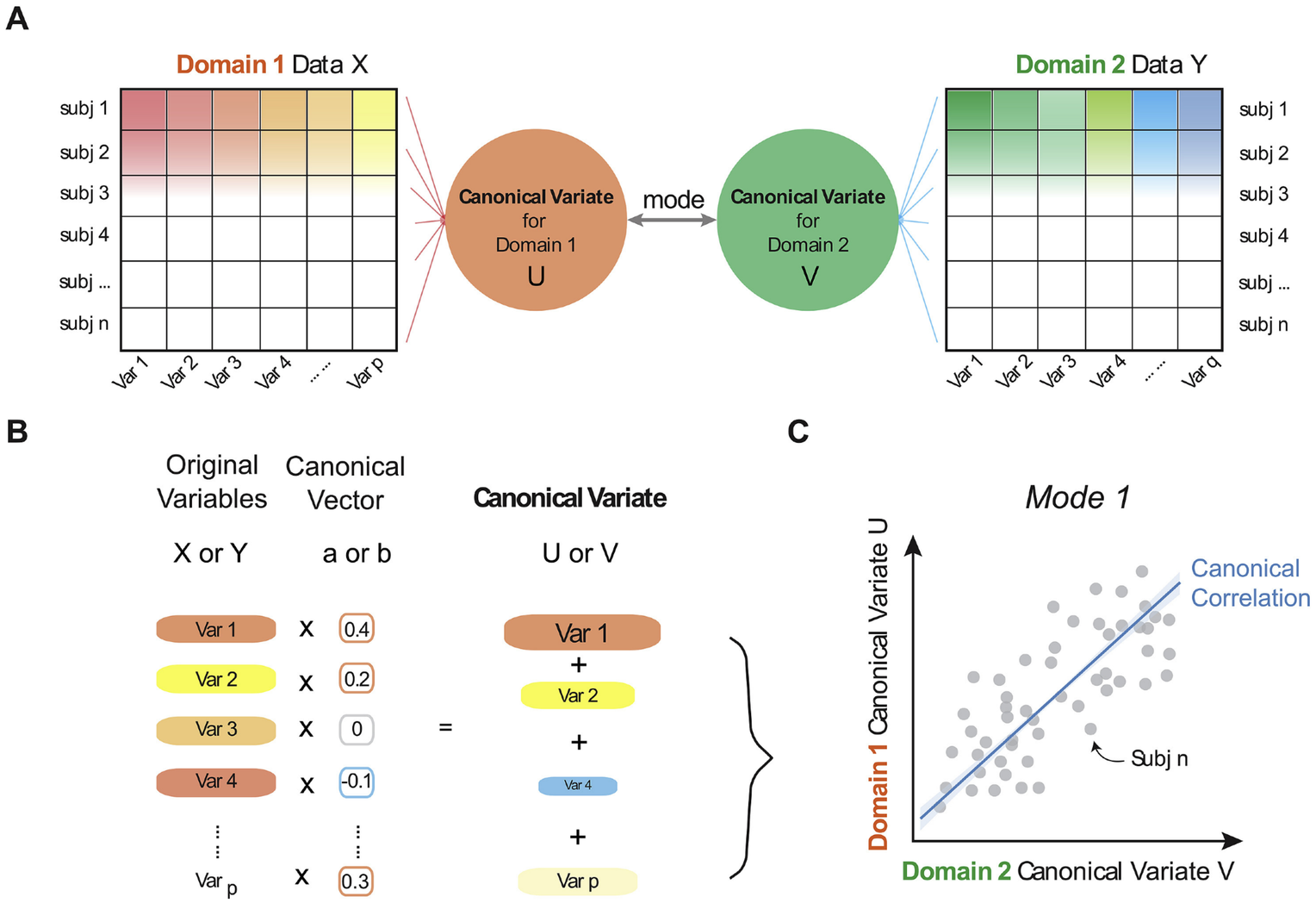
Isomap, tSNE, Diffusion map, UMAP, etc.





Dimensionality reduction (variable selection) for Multi-view datasets

Canonical Correlation Analysis



https://doi.org/10.1016/j.neuroimage.2020.116745

How to “fit” models? Frequentist or Bayesian

Unsupervised learning: $\mathcal{D} = \{(\boldsymbol{y}_n) : n = 1 : N\}$

Supervised learning: $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n) : n = 1 : N\}$

Maximum Likelihood Estimation (MLE): (e.g. minimising MSE under normal noise assumption)

$$\hat{\boldsymbol{\theta}}_{\text{mle}} \triangleq \underset{\boldsymbol{\theta}}{\text{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})$$

Bayes theorem:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})}{p(\mathcal{D})} = \frac{p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}')p(\mathcal{D}|\boldsymbol{\theta}')d\boldsymbol{\theta}'}$$

Posterior predictive distribution:

$$p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D}) = \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$$

MODIFIED BAYES' THEOREM:

$$P(H|X) = P(H) \times \left(1 + P(C) \times \left(\frac{P(X|H)}{P(X)} - 1\right)\right)$$

H: HYPOTHESIS

X: OBSERVATION

P(H): PRIOR PROBABILITY THAT H IS TRUE

P(X): PRIOR PROBABILITY OF OBSERVING X

P(C): PROBABILITY THAT YOU'RE USING BAYESIAN STATISTICS CORRECTLY

Choosing prior distribution

Informative, uninformative (minimally informative) and conjugate priors

Hierarchical priors (multi-level model): $\boldsymbol{\phi} \rightarrow \boldsymbol{\theta} \rightarrow \mathcal{D}$

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathcal{D}) = p(\boldsymbol{\phi}) p(\boldsymbol{\theta}|\boldsymbol{\phi}) p(\mathcal{D}|\boldsymbol{\theta})$$

Empirical Bayes, a frequentist-Bayesian hybrid approach based on the idea of learning prior from the data itself (maximising marginal likelihood):

$$\hat{\boldsymbol{\phi}}_{\mathrm{mml}}(\mathcal{D}) = \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\, p(\mathcal{D}|\boldsymbol{\phi}) = \underset{\boldsymbol{\phi}}{\operatorname{argmax}} \int p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\phi}) d\boldsymbol{\theta}$$

Sparsity-inducing prior can play the role of shrinkage in MLE: e.g. in Bayesian Lasso we use a Laplace prior.

False discovery rate (FDR): large-scale multiple hypothesis test in cDNA microarray data

	BRCA1				BRCA2			
	1	2	—	7	1	2	—	8
Gene1	−1.29	−1.41	—	−0.55	−0.70	1.33	—	1.14
Gene2	2.03	0.58	—	−0.12	0.23	−0.91	—	−0.39
Gene3	0.32	−0.44	—	1.25	0.53	−0.96	—	−0.51
Gene4	−1.31	−0.98	—	0.24	−0.24	0.28	—	2.13
Gene5	−0.66	−0.07	—	1.22	−0.41	−0.88	—	−0.83

$p_1 = \text{Prob}\{\text{Different}\}$ $f_1(y)$ density of Y_i if gene_i "Different"

$p_0 = \text{Prob}\{\text{Not Different}\}$ $f_0(y)$ density of Y_i if gene_i "Not Different".

Finally let $f(y)$ be the mixture density

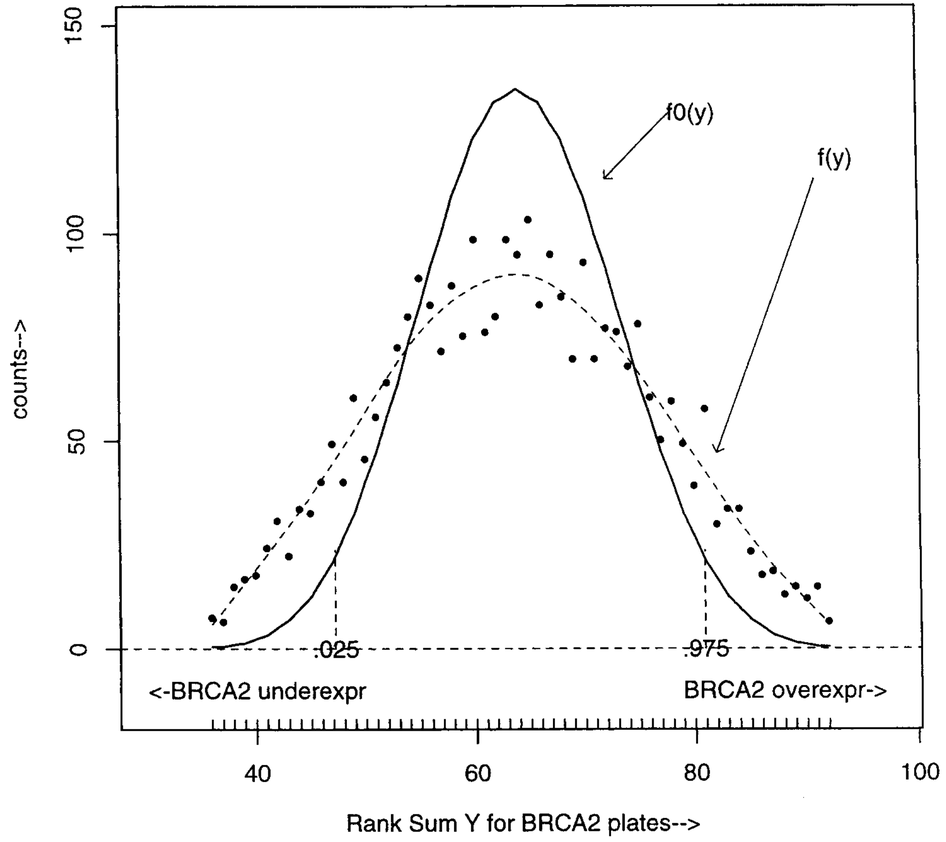
$$f(y) = p_0 f_0(y) + p_1 f_1(y).$$

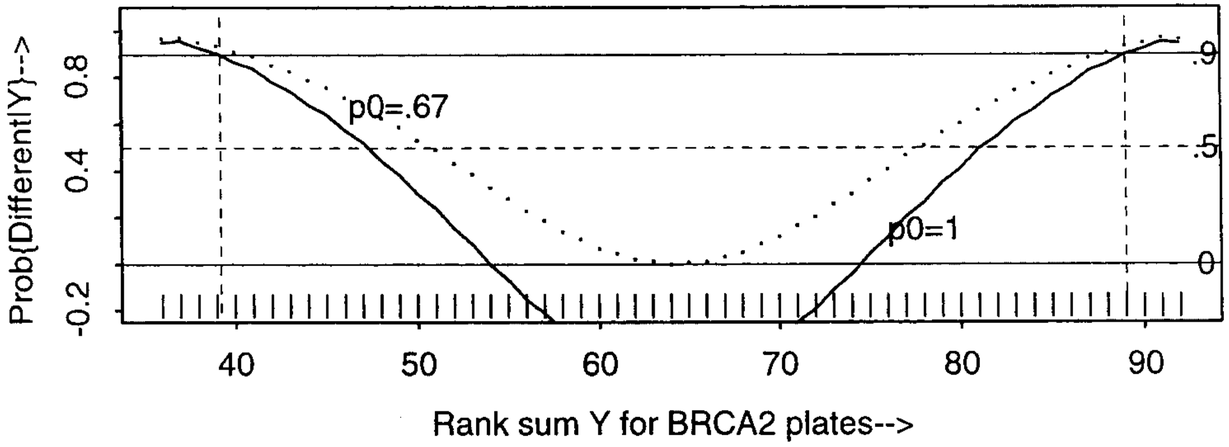
A direct application of Bayes' theorem gives *a posteriori* probabilities

$$p_1(y) \equiv \text{Prob}\{\text{Different} \mid Y_i = y\} = 1 - p_0 f_0(y)/f(y)$$

and

$$p_0(y) \equiv \text{Prob}\{\text{Not Different} \mid Y_i = y\} = p_0 f_0(y)/f(y).$$





Efron & Tibshirani, Genetic epidemiology 2002

Challenges and fruits of “fitting” mechanistic models

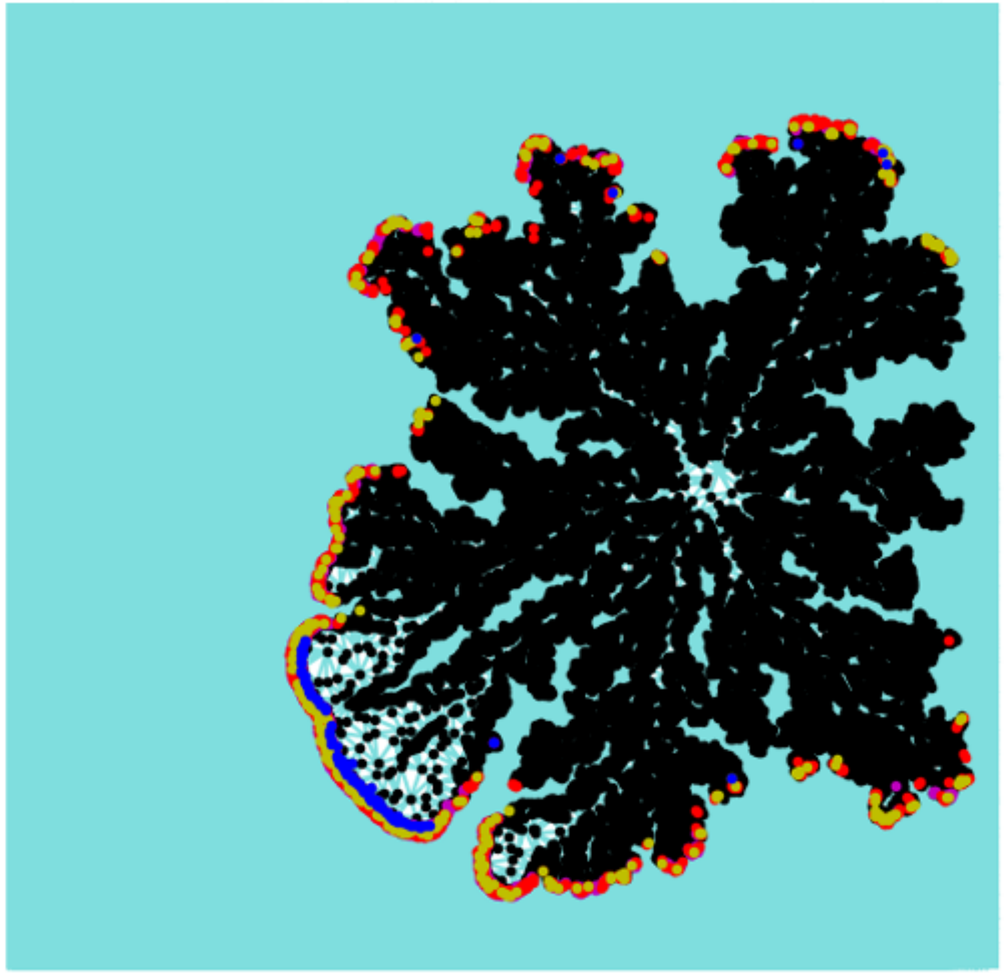
Learn about underlying physics of the world and can start to understand (and control) natural phenomenon.

However, these models are typically complex, can only be (inefficiently) simulated and are not analytically solvable. Fitting requires many evaluations of the model for a range of parameter values.

The first solution is to use mathematical approximations or better numerical algorithms to “efficiently” solve these models (CCMI).

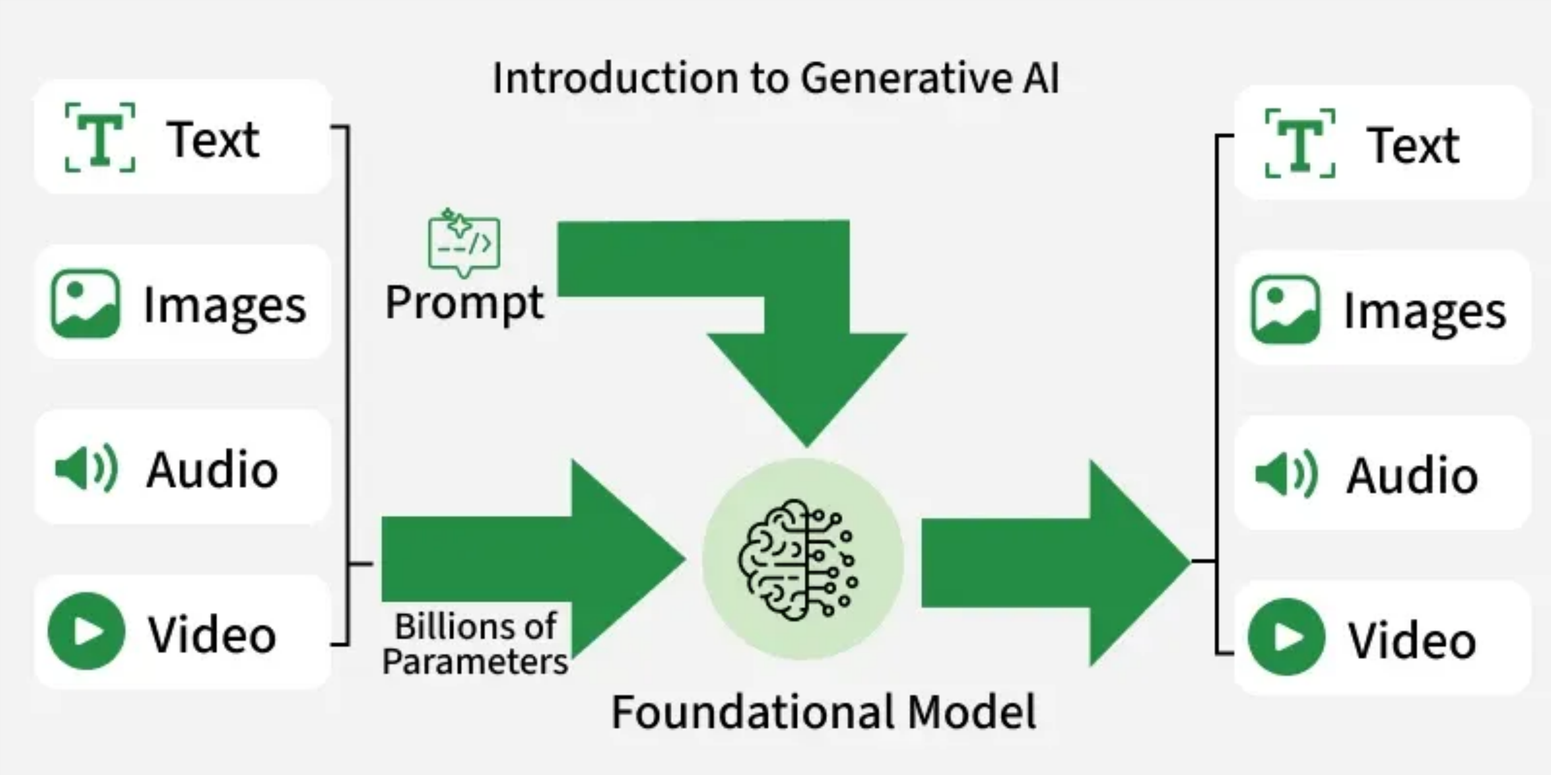
The second solution is to use “efficient” simulation-based inference methods, such as Approximate Bayesian Computation or machine-learning enhanced posterior density estimation.

Agent-based model of tumour growth



Joergensen et al, PLOS Computational Biology 2022

Generative AI, semi-supervised and self-supervised learning



https://www.geeksforgeeks.org/artificial-intelligence/what-is-generative-ai/

References

1. James, Witten, Hastie, Tibshirani and Taylor, An introduction to Statistical Learning with Applications in R or Python, Springer Texts in Statistics (free download from https://www.statlearning.com/).

2. Eforn and Hastie, Computer Age Statistical Inference, Cambridge University Press.

3. Murphy, Probabilistic Machine Learning, An Introduction, MIT Press.

4. Murphy, Probabilistic Machine Learning, Advanced topics, MIT Press.

5. Handbook of Approximate Bayesian Computation, Edited by Sisson, Fan and Beaumont, Chapman & Hall/CRC Press.

6. Cranmer et. al. The frontier of simulation-based inference. PNAS 2020 117, 30055 (https://doi.org/10.1073/pnas.1912789117).