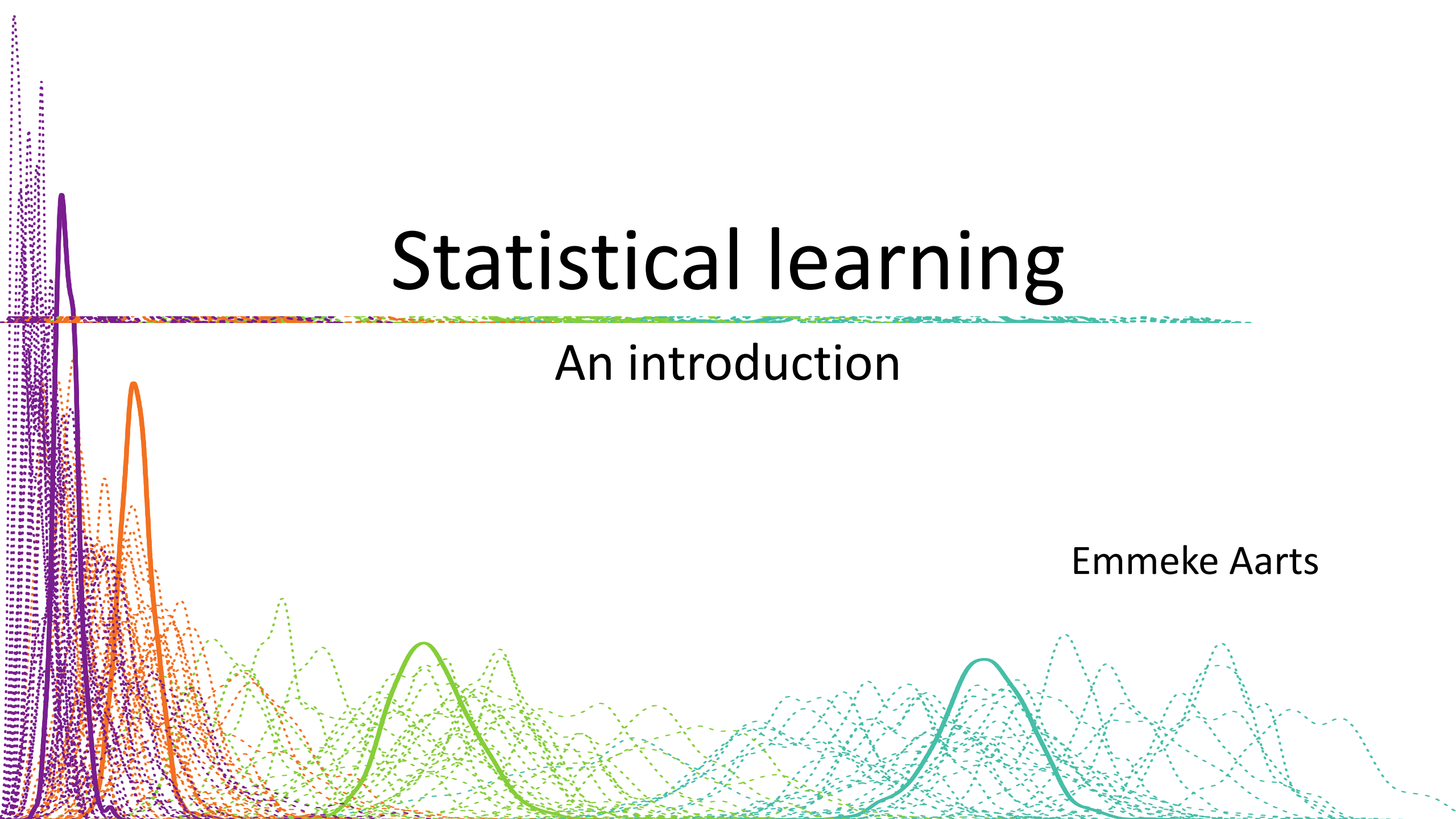
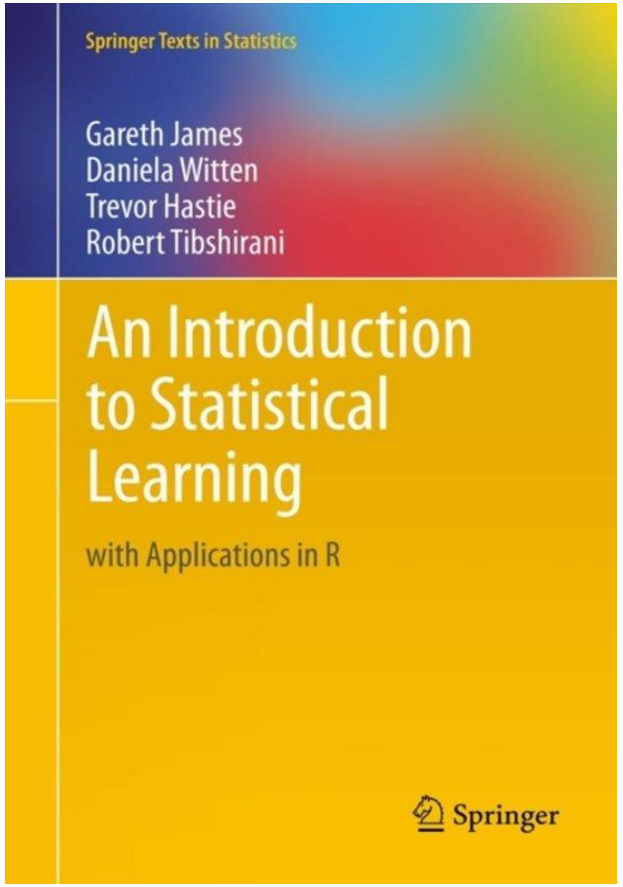


Statistical learning

An introduction

Emmeke Aarts



Date	Topic	Presented by
8 november	Statistical learning: an introduction	Dr. Emmeke Aarts
29 november	Regression from the data science perspective	Dr. Dave Hessen
6 december	Classification	Dr. Gerko Vink
10 januari	Resampling Methods	Dr. Gerko Vink
7 februari	Regularization	Dr. Maarten Cruijf
7 maart	Moving beyond linearity	Dr. Maarten Cruijf
4 april	Tree-Based models	Dr. Emmeke Aarts
9 mei	Support vector Machines	Dr. Daniel Oberski
6 juni	Unsupervised learning	Prof. Dr. Peter van der Heijden

Outline

- What is statistical learning
- Accuracy versus interpretability
- Supervised versus unsupervised learning
- Regression versus classification
- Model accuracy & bias-variance trade off
- Potential benefits for social scientist
- Software

What is statistical learning – Big data

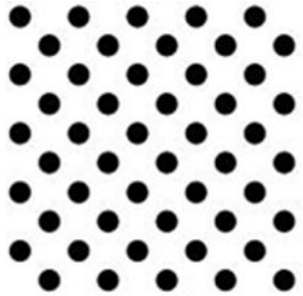
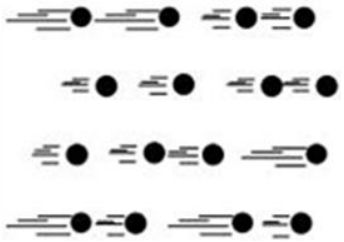
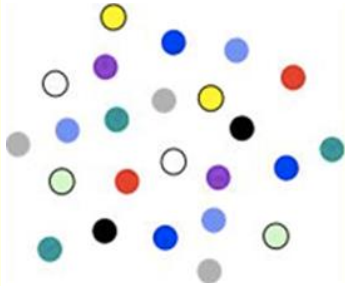
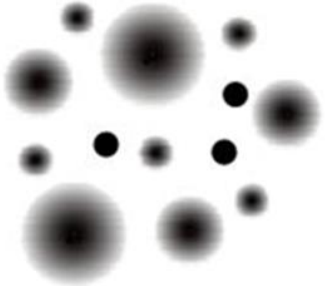
- *‘[B]ig data’ […] refers to*
  - *large,*
  - *diverse,*
  - *complex,*
  - *longitudinal,*
  - *and/or distributed data sets*

generated from

- *instruments,*
- *sensors,*
- *Internet transactions,*
- *email,*
- *video,*
- *click streams,*
- *and/or all other digital sources available[.]*

(NSF, NIH 2012)

What is statistical learning – Big data

Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data into Money**
Terabytes to Exabytes of existing data to process	**Streaming data, requiring milliseconds to seconds to respond**	**Structured, unstructured, text, multimedia,…**	**Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations**	**Business models can be associated to the data**

Source: http://informationcatalyst.com

Adapted by a post of Michael Walker on 28 November 2012

What is statistical learning – Big data

So, how different from e.g., the massive data sets arising in physics?

1. *'Big data' [is] the amassing of huge amounts of statistical information on **social** and **economic** trends and **human behavior**. (M. Chen)*

   data on people

2. *Granularity:* documents of social phenomena at the granularity of *individual* people and their activities. (M.I. Jordan)

Issues regarding ethics, privacy, bias, fairness, and inclusion.

For a nice overview on this, see Hanna Wallach on *Medium*: Big data, machine learning and the social sciences: Fairness, accountability, and Transparancy

Why should social scientist bother

- Science: “minimal evidence of emerging computational social science engaged in quantitative modeling of these new kinds of digital traces.” (Lazer, Science)

- Industry & government: computational social science is occurring on **a large scale**, in places like
  - Google
  - Yahoo
  - the National Security Agency

See e.g.: D. Lazer et el. (2009). Life in the network: the coming age of computational social science. *Science*

What is statistical learning

Machine learning (ML): Allowing computers to learn for themselves without explicitly being programmed

- Google: AlphaGo, computer that defeated world champion Go player
- Apple & android: Siri voice assistant

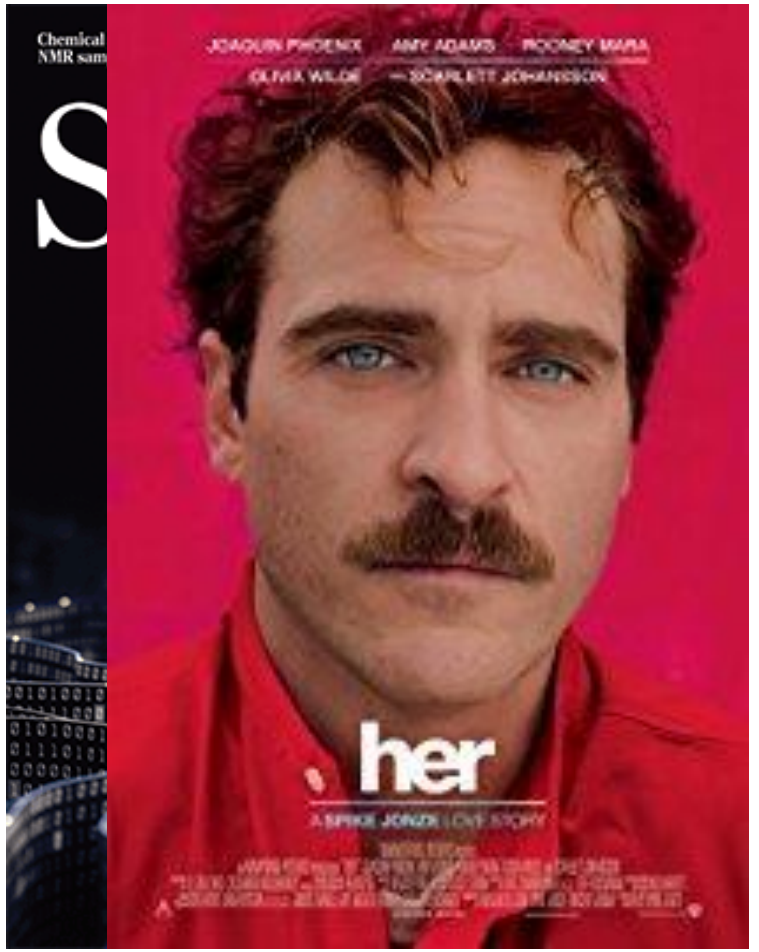
Train a system by showing examples of input-output behavior, instead of

programming it manually by anticipating the desired response for all possible inputs

What is statistical learning

- Artificial intelligence (AI): Constructing machines (robots, computers) to think and act like human beings
- ML is a subset of AI

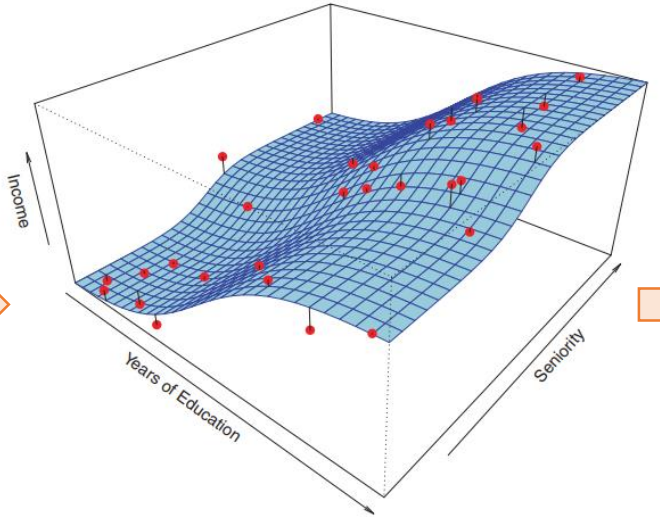
- Statistical learning (SL): a set of approaches for estimating *f*; a function that represents our data that can be used for e.g., prediction and/or inference
- SL is a subset of ML



Supervised versus unsupervised learning

Supervised learning

Input **Statistical model** **Output**



Building a statistical model for **predicting** / estimating an **output** based on one or more **inputs**

Graph from: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning

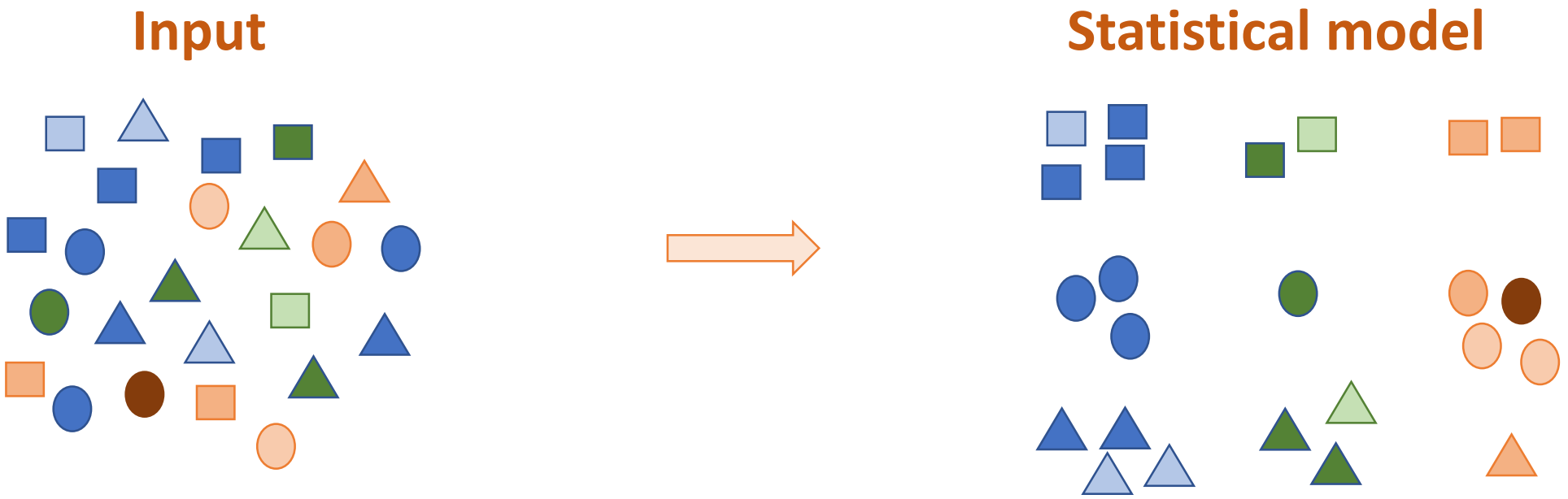
Supervised versus unsupervised learning

Supervised learning

- most widely used machine-learning methods are supervised
  - spam classifiers of e-mail
  - face recognizers over images
  - medical diagnosis systems for patients
- Methods include
  - decision trees
  - (logistic) regression
  - support vector machines
  - neural networks
  - Bayesian classifiers

Supervised versus unsupervised learning

Unsupervised learning



Inputs, but **no outputs**. Try to learn structure and relationships from these data

Supervised versus unsupervised learning

Unsupervised learning

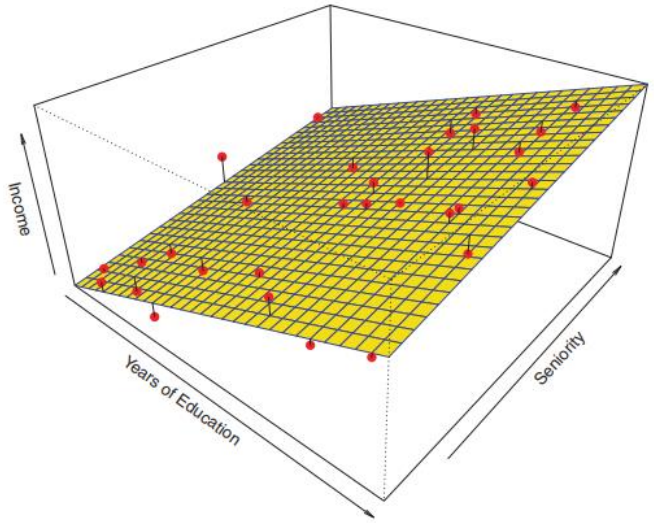
- assumptions about structural properties of the data

- Dimension reduction methods
  - principal components analysis
  - factor analysis
  - random projections

- Clustering
  - K-means clustering

How do we learn

Parametric models

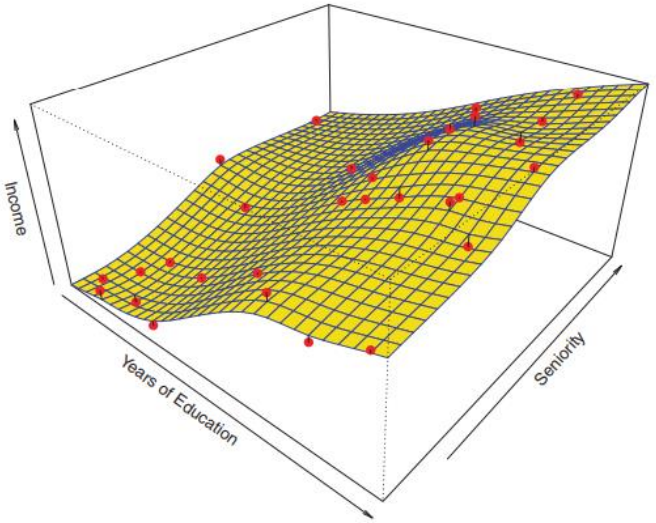


Linear model

Restrictive

Inference: interpretable

Non-parametric models



Smooth thin-plate spline model

Flexible

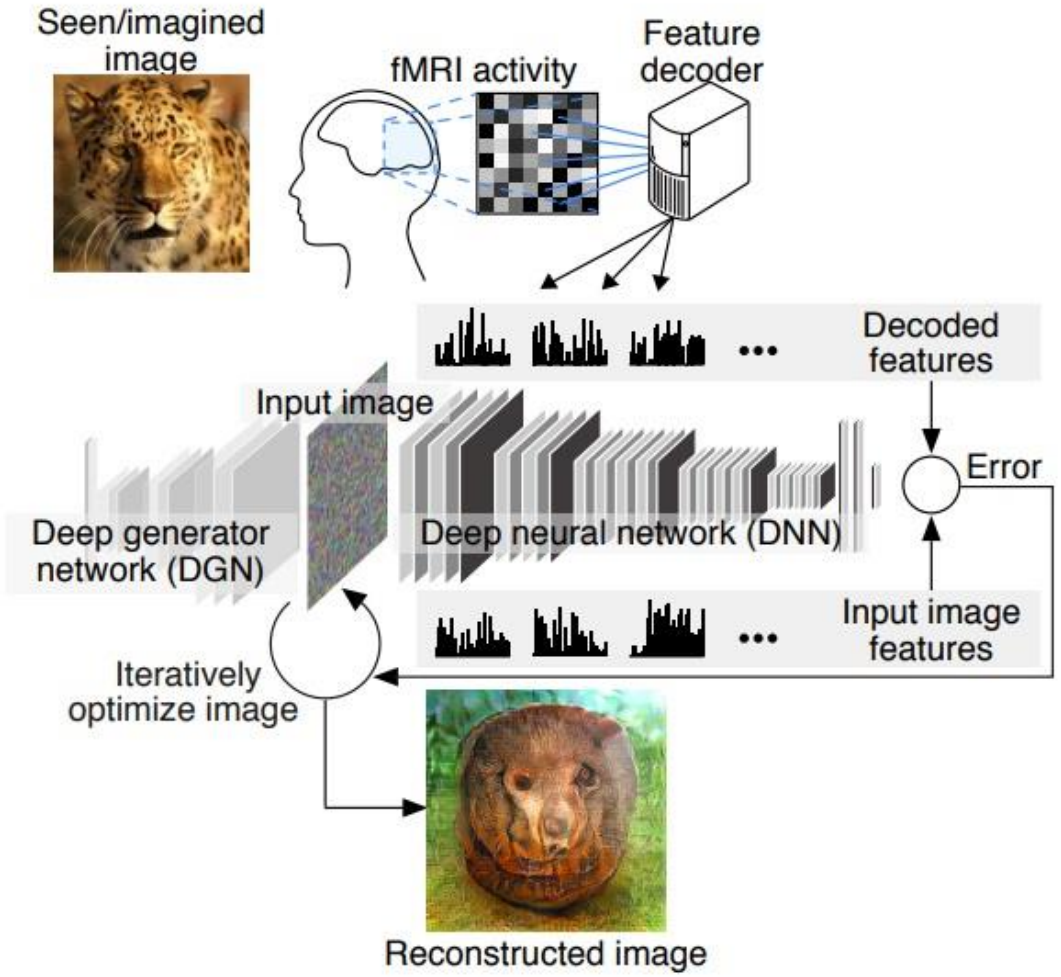
Not so interpretable

Graphs from: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning

Accuracy versus interpretability

Image from: https://blog.siftscience.com/2017/machine-learning-black-box/

Black box example from neuroscience

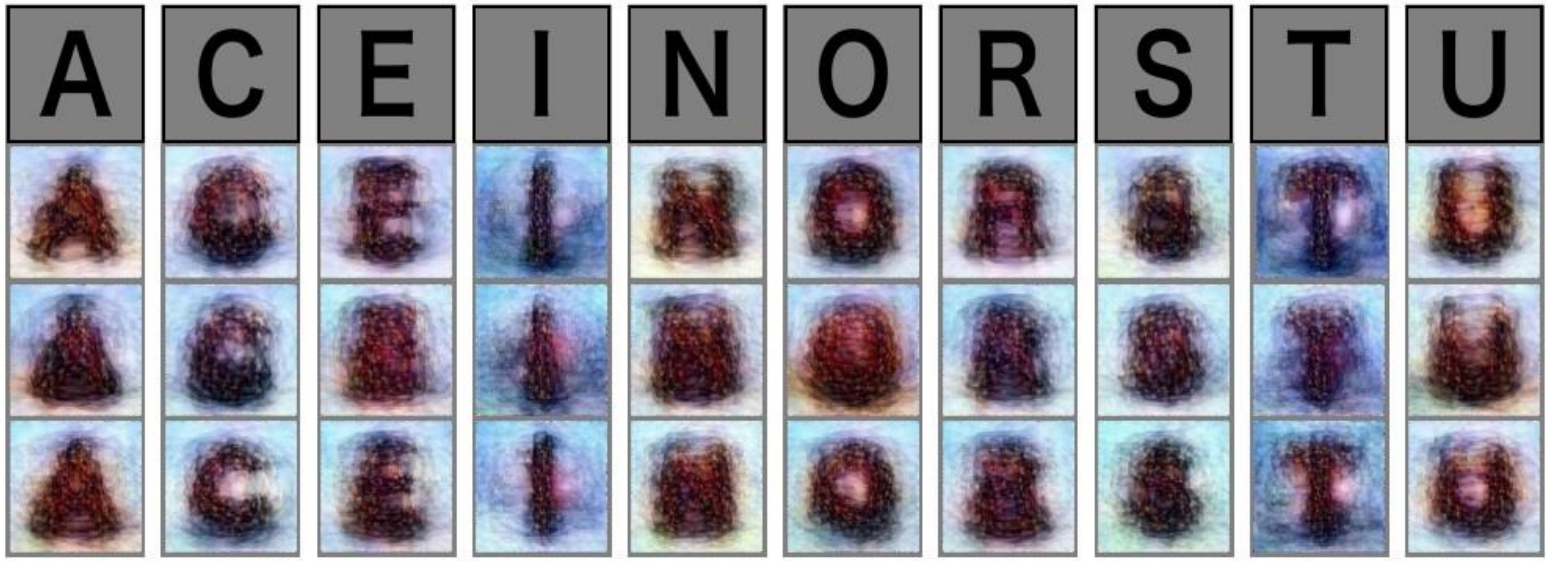


Deep image reconstruction from human brain activity

G. Shen*, T. Horikawa*, K. Majima*, and Y. Kamitani

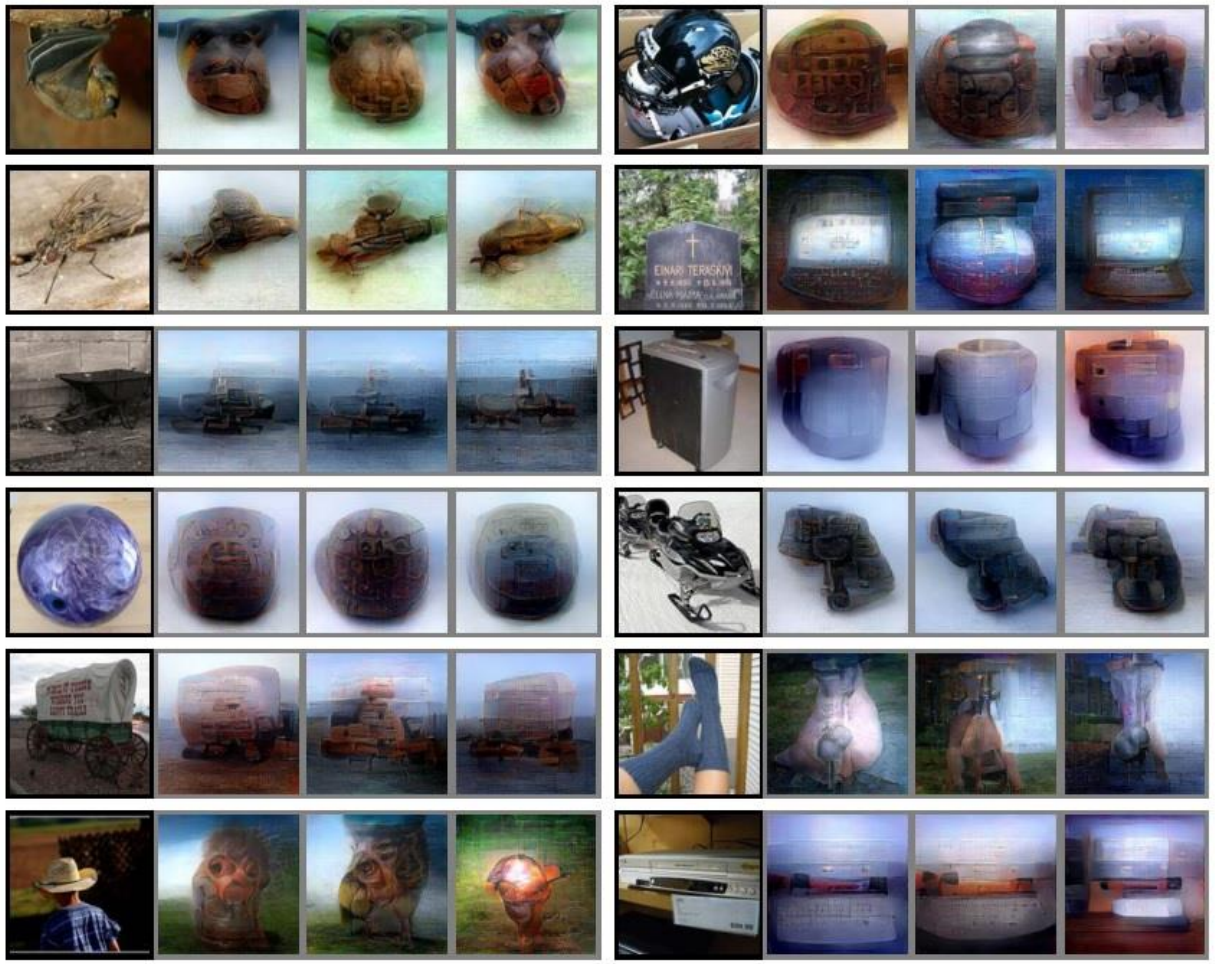
2017

Black box example from neuroscience



Supplementary Figure 9 | All examples of alphabetical letter reconstructions. Images with black and gray frames show presented and reconstructed images, respectively (reconstructed from VC activity without the DGN). Three reconstructed images correspond to reconstructions from three subjects.

Black box example from neuroscience



Supplementary Figure 2 | Other examples of natural image reconstructions obtained with the DGN. Images with black and gray frames show presented and reconstructed images, respectively (reconstructed from VC activity using all DNN layers). Three reconstructed images correspond to reconstructions from three subjects.

Accuracy versus interpretability

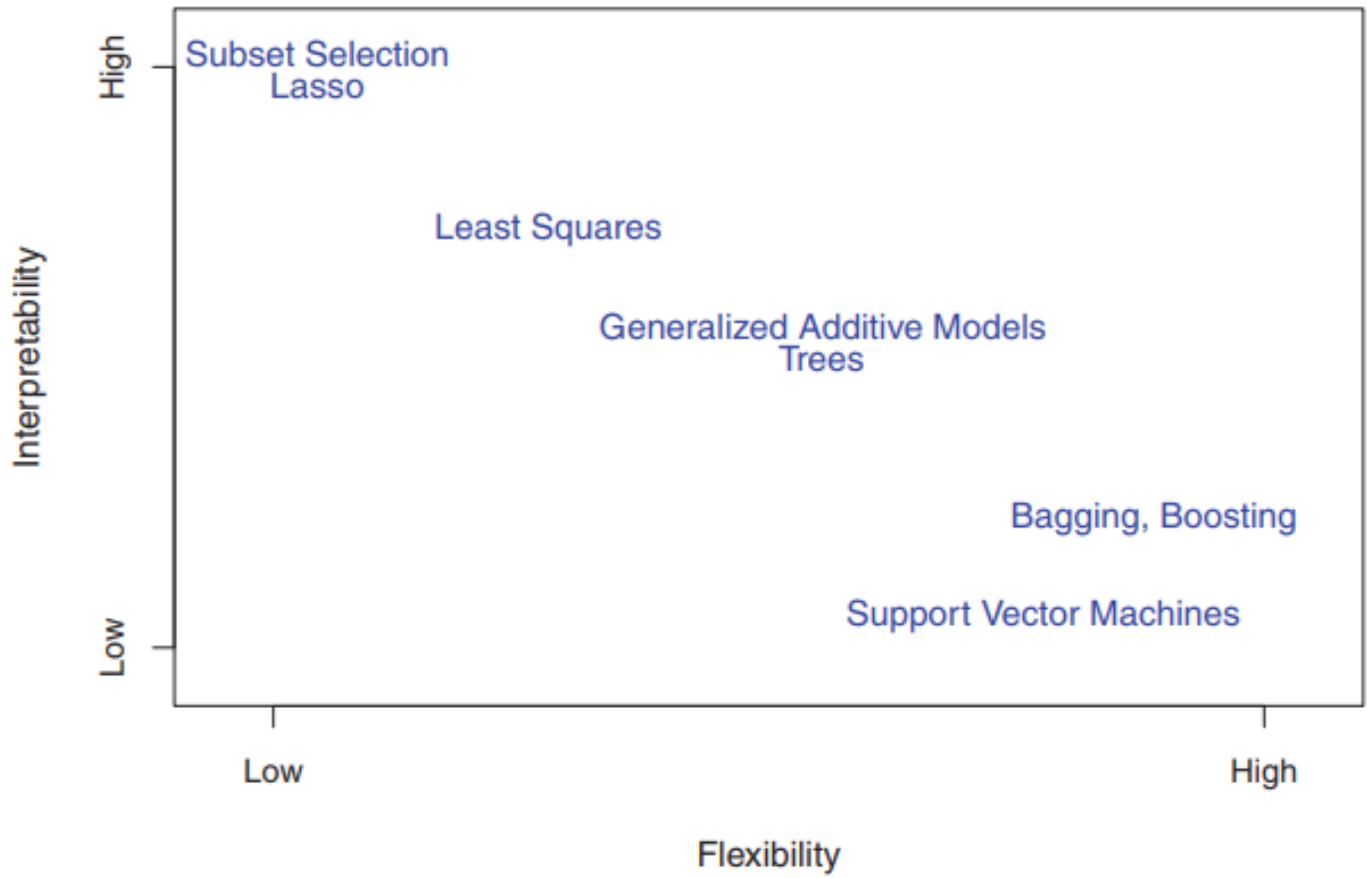


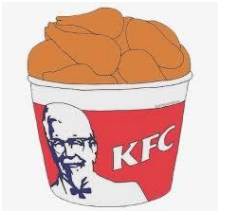
Illustration from: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning

Regression versus classification

Quantitative outcomes

Predict an quantitative outcome -> regression

Qualitative outcomes



Predict to which category an observation belongs -> classification

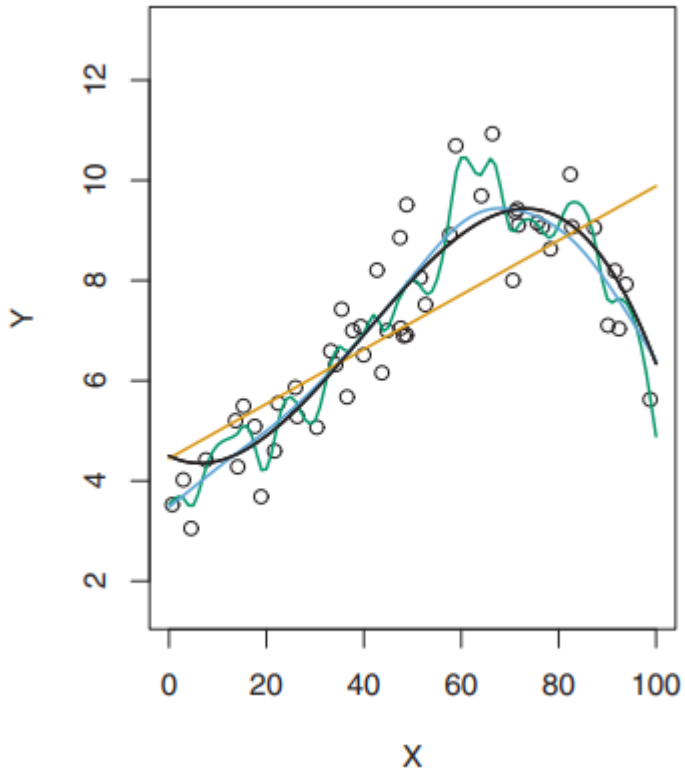
Model accuracy

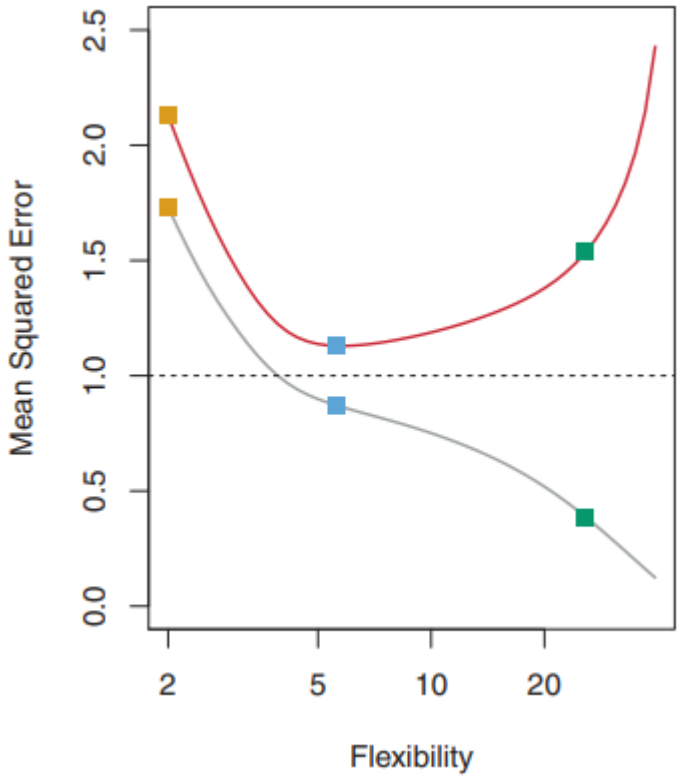
Model accuracy: Mean Squared Error

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

To obtain the MSE, we use **training set** and a **test set**

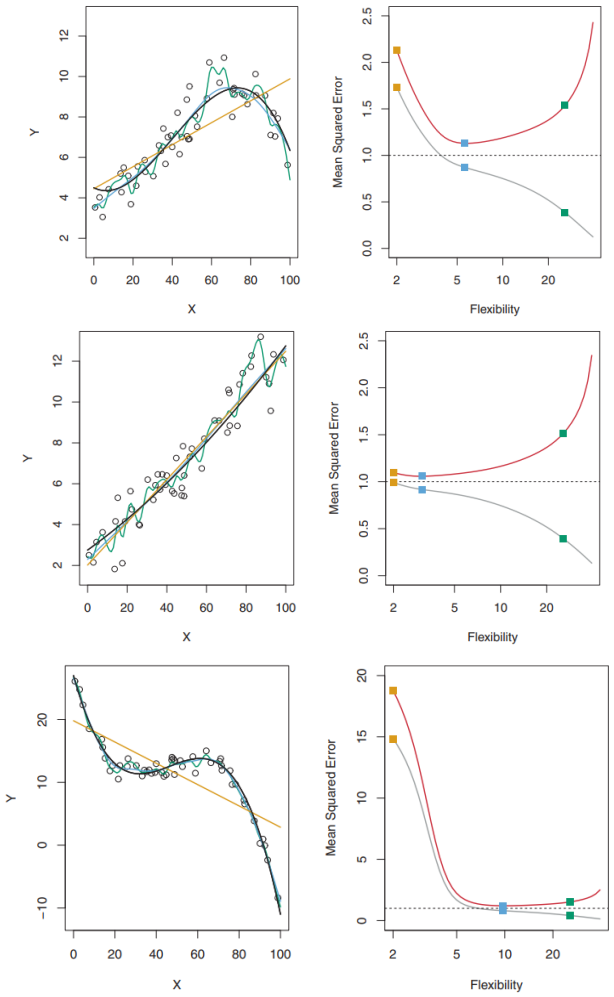
Model accuracy

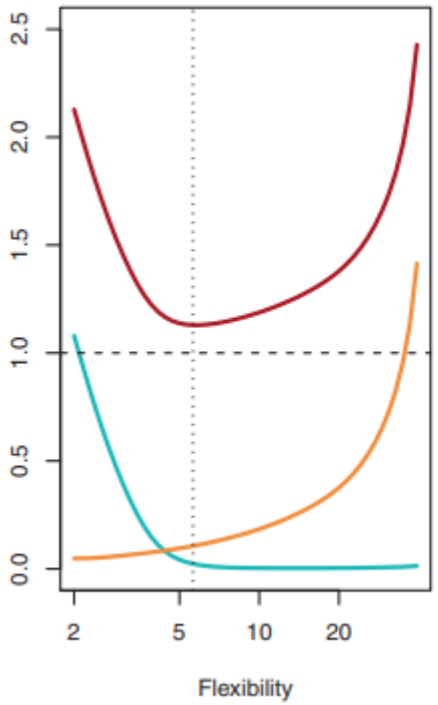


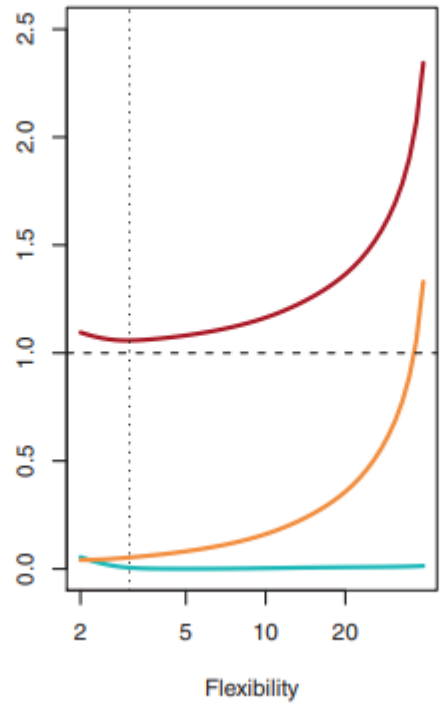


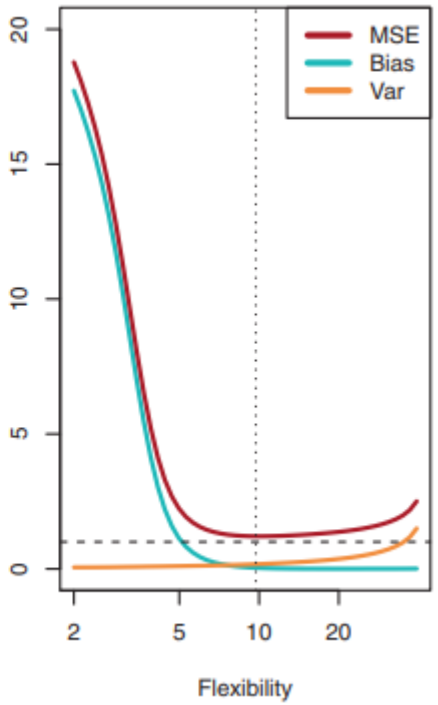
Graphs from: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning

Bias-variance trade off









Graphs from: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning

Potential benefits for social scientist

- Solutions to overfitting (what we call the replication bias)

1. uncover patterns and structure embedded in data
2. test and improve model specification and predictions
3. perform data reduction

Software

- R and Python: core of machine learning development
- Matlab has a ML toolbox, but lacks customizability
- Some techniques available in SPSS / Stata
- Specific programs for specific techniques, e.g., Tensorflow

References

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: springer
- Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Life in the network: the coming age of computational social science. *Science*, *323*(5915), 721.
- Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2017). Deep image reconstruction from human brain activity. *bioRxiv*, 240317.
- https://medium.com/@hannawallach/big-data-machine-learning-and-the-social-sciences-927a8e20460d