Biological Sciences faculty
Biophysics Department

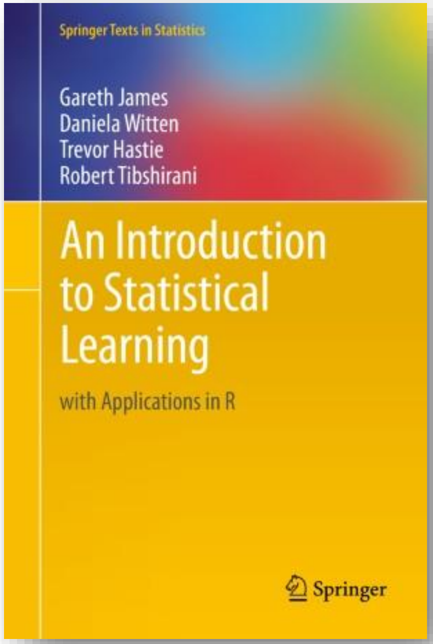
Introduction to Applied Machine Learning

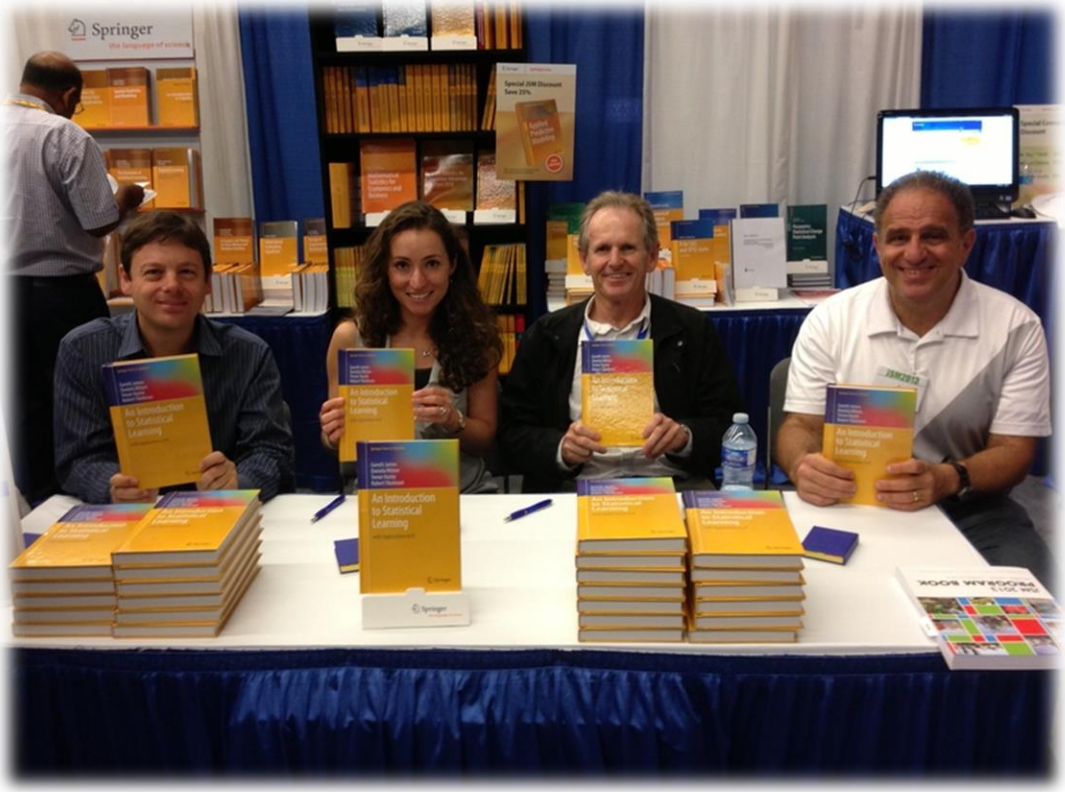
Presented By
Alireza Doustmohammadi
Graduate Student in Bioinformatics

January 2021

Contents

Reference

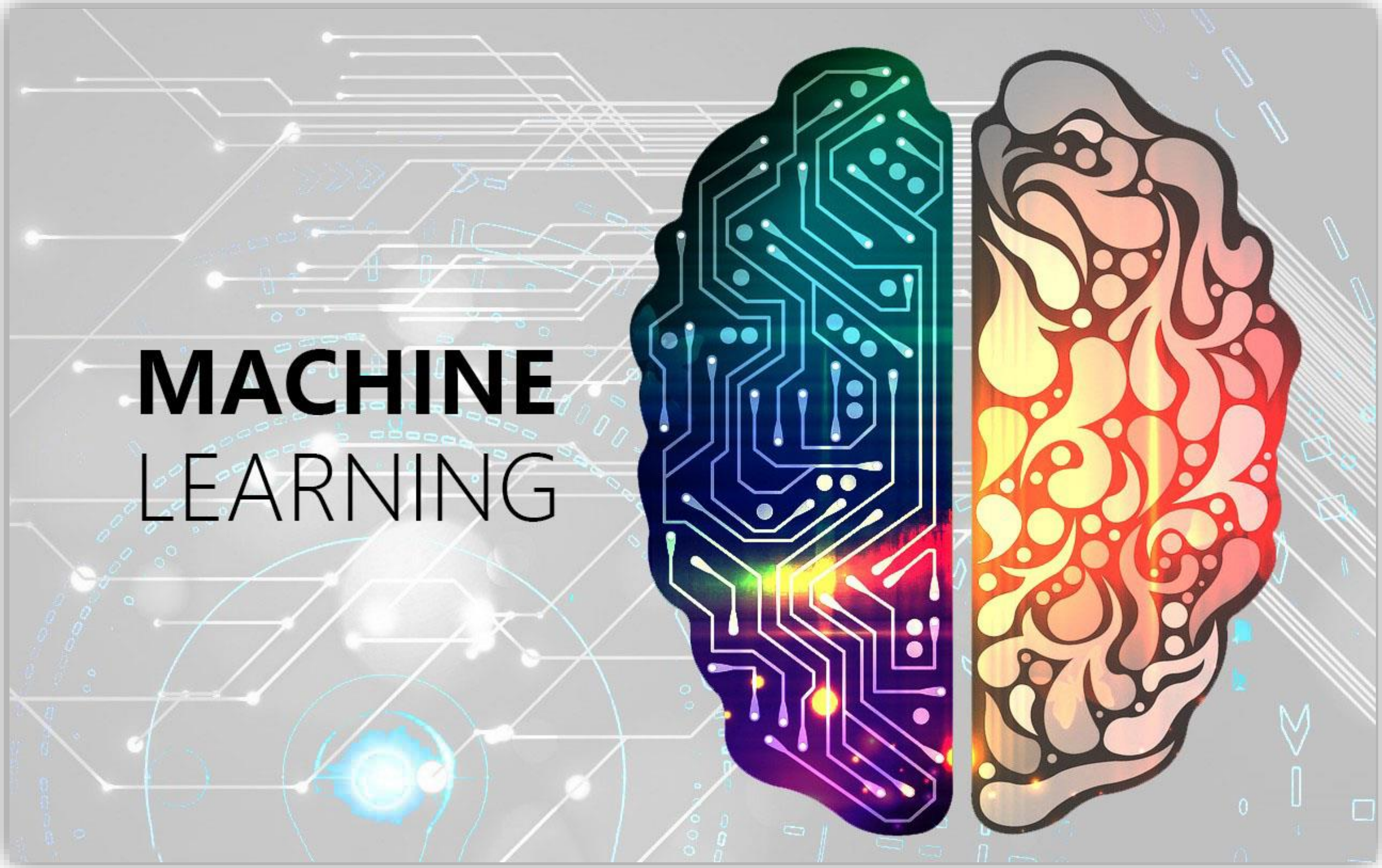

Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer


Springer

Outline

- Supervised Learning
- Unsupervised Learning
- Ensemble Learning

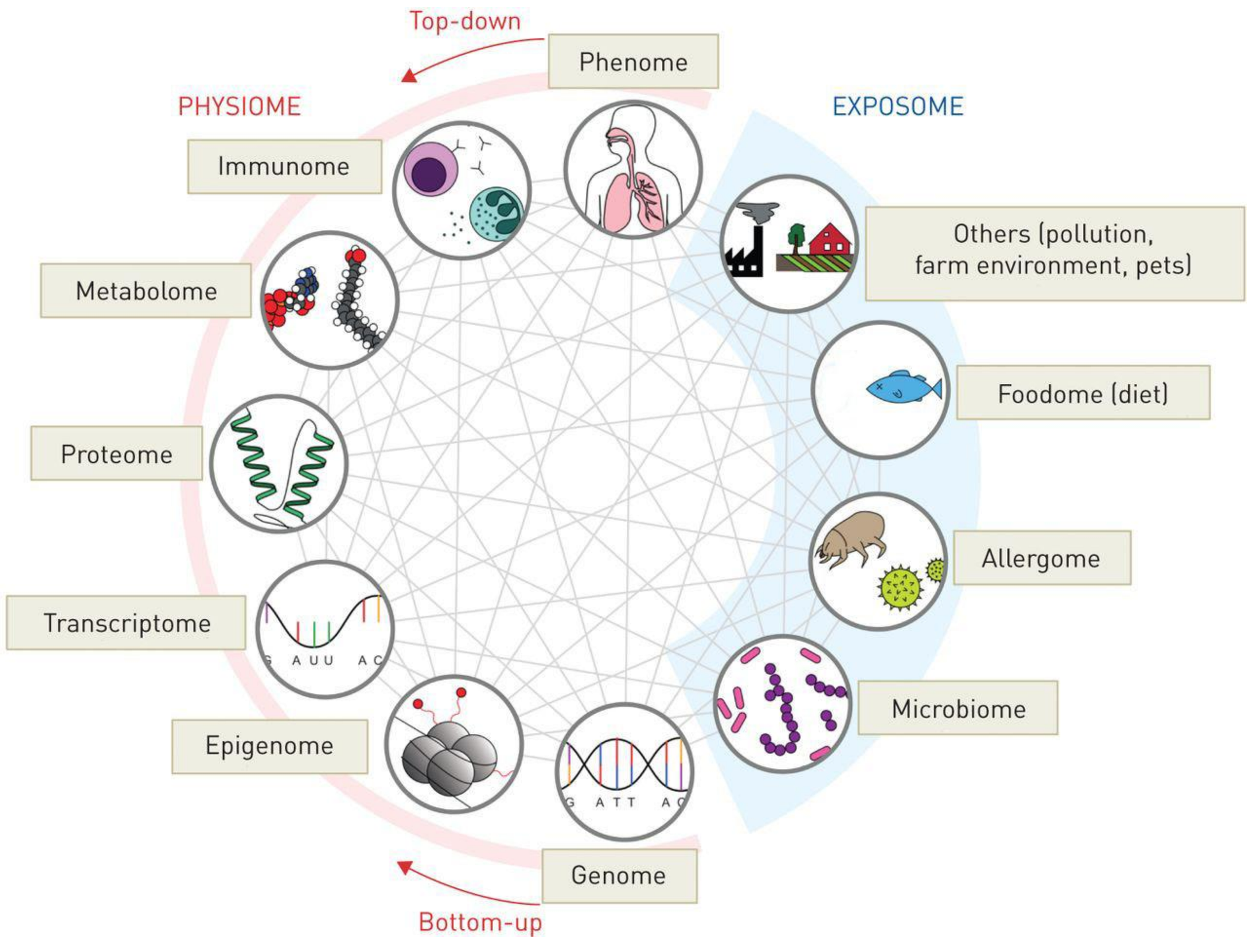
Why do we need to prediction?



Why do we need to prediction?

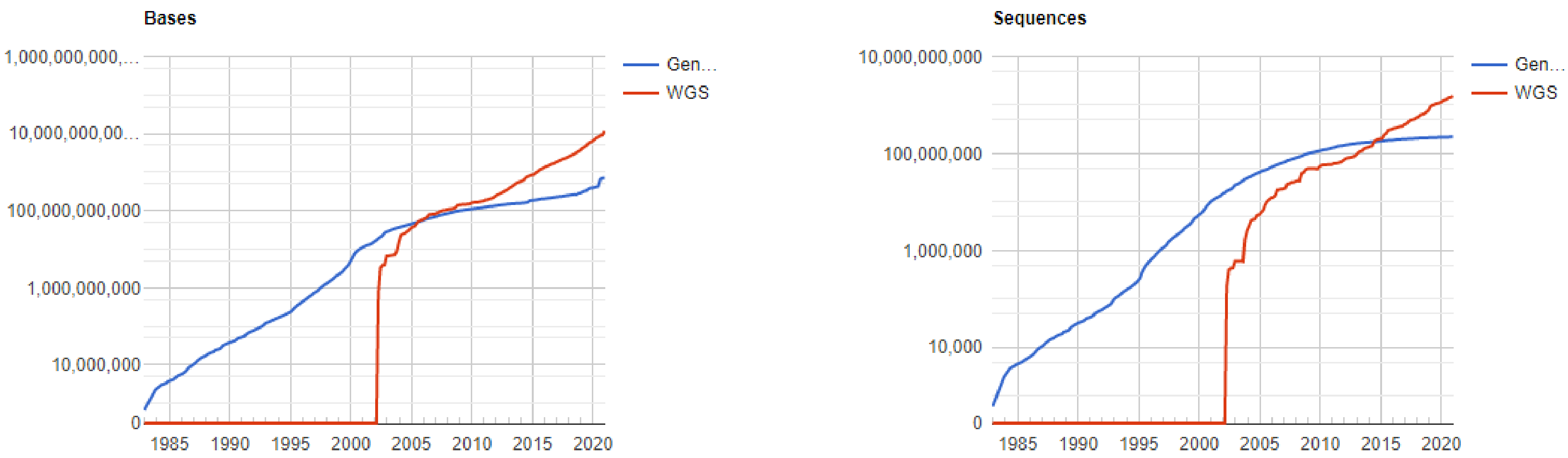
Welcome To de Era of Big Data .....

Why do we need to prediction?



Why do we need to prediction?

GenBank and WGS Statistics



[https://www.ncbi.nlm.nih.gov/genbank/statistics/]

Why do we need to prediction?

PDB Data Distribution by Experimental Method and Molecular Type:

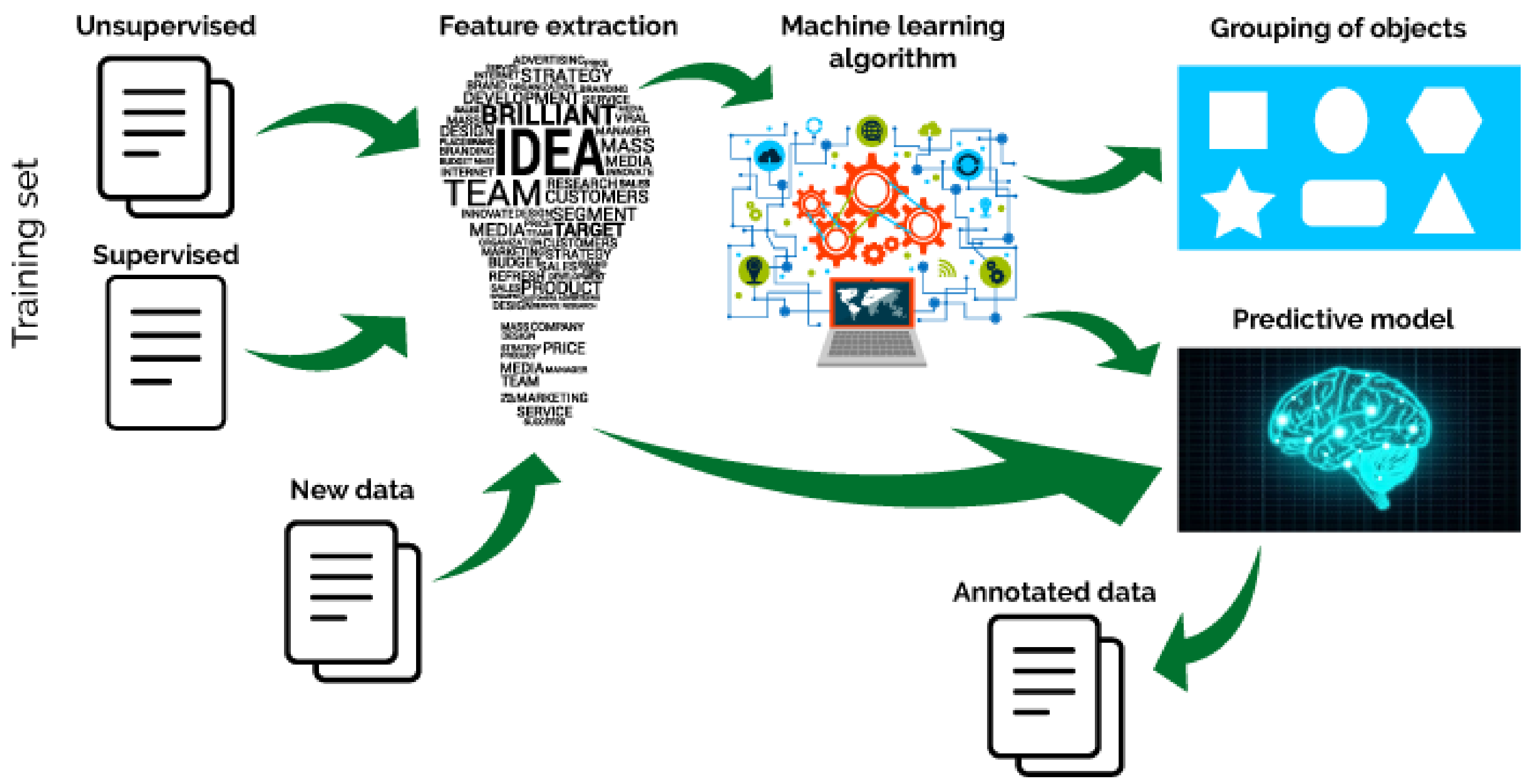
Molecular Type	X-ray	NMR	EM	Multiple methods	Neutron	Other	Total
Protein (only)	135896	36576	4544	165	67	36	152280
Protein/NA	7177	269	1603	3	0	0	9052
Nucleic acid (only)	2158	1360	53	7	2	1	3561
Other	149	31	3	0	0	0	183
Total	153600	13653	6814	181	69	37	173754

[https://www.rcsb.org/stats/summary]

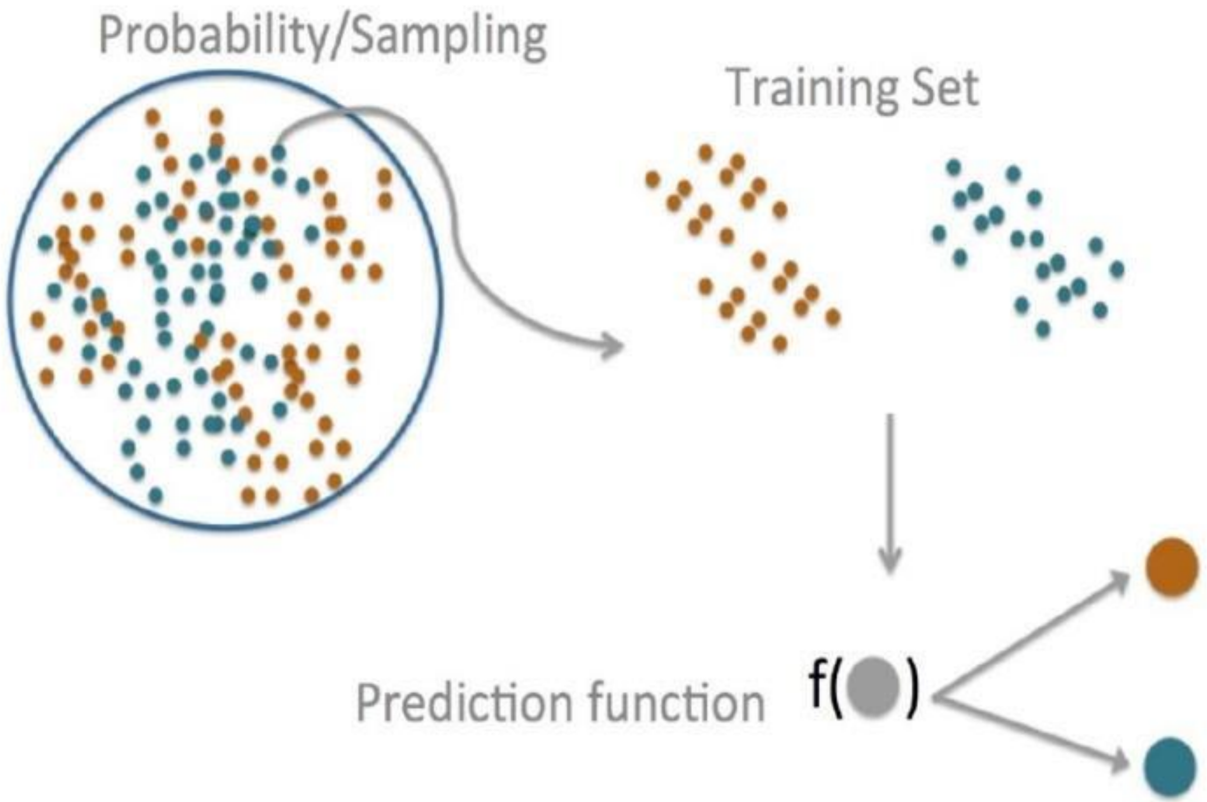
Why do we need to prediction?

Basic Concepts & Nomenclatures

Central Dogma of Prediction



Central Dogma of Prediction



Central Dogma of Prediction

- Defining the Questions
- Data Collection
- Feature Extraction
- Preprocessing & Feature Selection
- Algorithm (Classifier)
- Evaluation
- Redesign the Algorithm (Parameter Tuning)

Features

- ✓ Good representation of data
- ✓ Data Compression
- ✓ Need to expert’s knowledge

Data Matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \qquad \mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \end{pmatrix}$$

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FAL	
...	rainy	75.0	80.0	FAL	
...	sunny	75.0	70.0	TRU	
...	overcast	72.0	90.0	TRU	
...	overcast	81.0	75.0	FAL	
...	rainy	71.0	91.0	TRU	

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

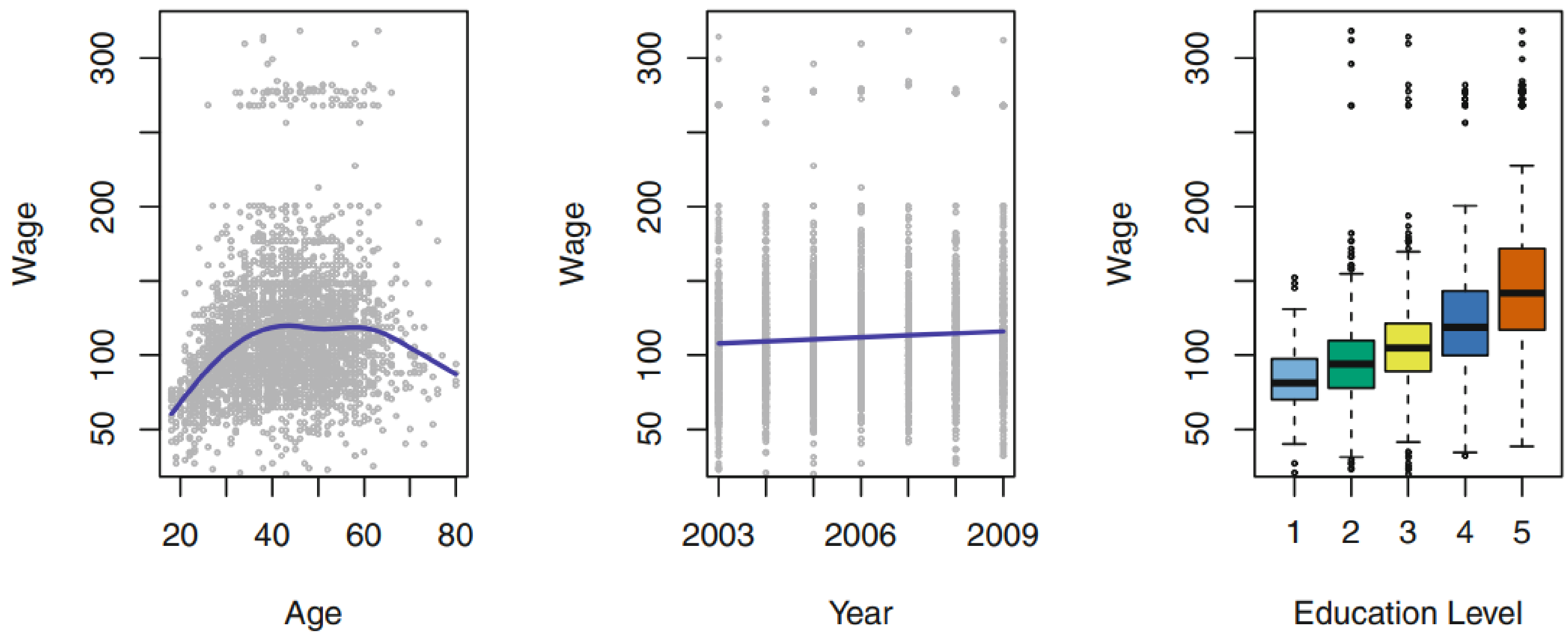
X , Y Relation

- We assume there is a **relationship between Y and X=** $\left(\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \dots, \boldsymbol{X_p}\right)$, witch can be written in the very general form:

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

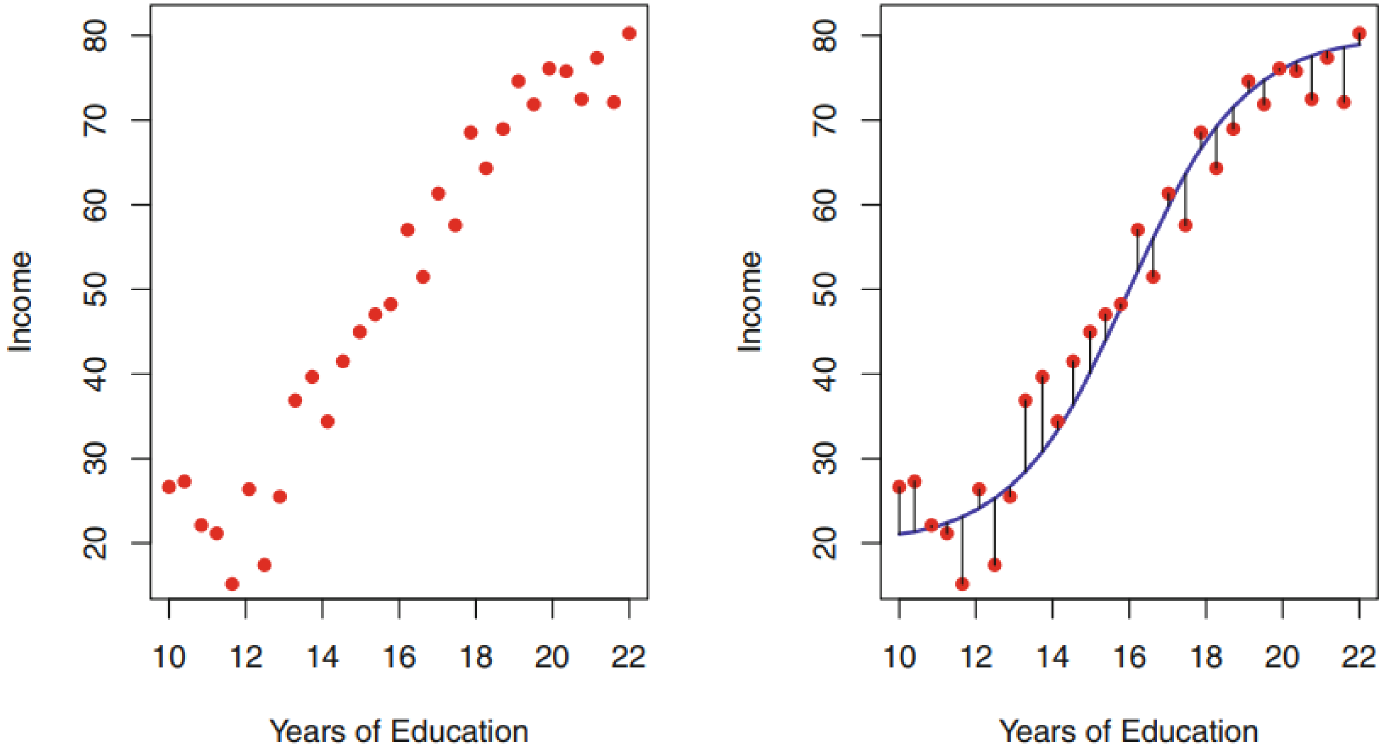
- Here $\boldsymbol{f}$ is some **fixed but unknown** function of $X_1, X_2, \dots, X_p$, and $\boldsymbol{\varepsilon}$ **is a random error term**, witch is **independent of** $\boldsymbol{X}$ and has mean zero.

X , Y Relation



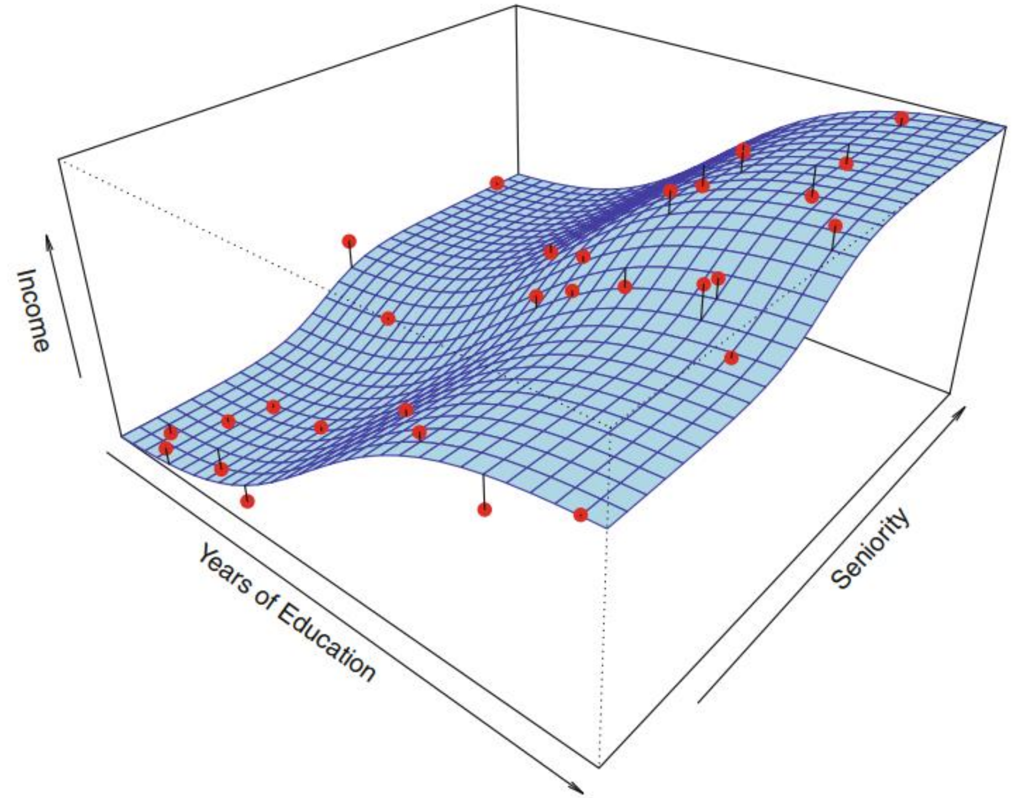
Wage
300
200
100
50
20
40
60
80
Age
Wage
300
200
100
50
2003
2006
2009
Year
Wage
300
200
100
50
1
2
3
4
5
Education Level

X , Y Relation



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

X , Y Relation



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

Reducible and irreducible error

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

$$\boldsymbol{E\left(Y - \hat{Y}\right)^2 = E\left[f(X) + \varepsilon - \hat{f}(X)\right]^2}$$

$$\boldsymbol{= \underbrace{\boxed{\left[f(X) - \hat{f}(X)\right]^2}}_{\text{Reducible}} + \underbrace{\boxed{Var(\varepsilon)}}_{\text{Irreducible}}}$$

Goal: Minimizing the reducible error

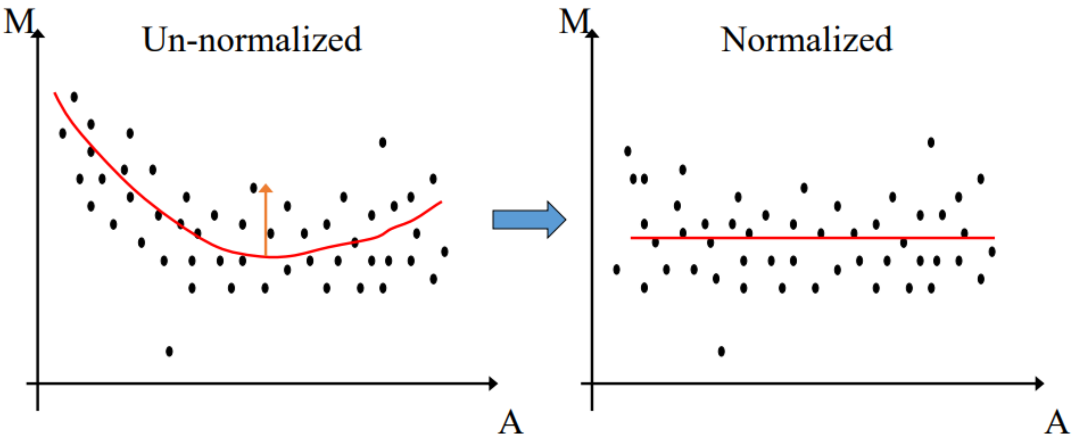
Preprocessing & Feature Selection

Preprocessing & Feature Selection

	MMT00000044	MMT00000046	MMT00000051	MMT00000076	MMT00000080	MMT00000102	MMT00000149
F2_2	-0.01810000	-0.077300000	-0.02260000	-0.00924000	-0.04870000	0.17600000	0.07680000
F2_3	0.06420000	-0.029700000	0.06170000	-0.14500000	0.05820000	-0.18900000	0.18600000
F2_14	0.00006440	0.112000000	-0.12900000	0.02870000	-0.04830000	-0.06500000	0.21400000
F2_15	-0.05800000	-0.058900000	0.08710000	-0.04390000	-0.03710000	-0.00846000	0.12000000
F2_19	0.04830000	0.044300000	-0.11500000	0.00425000	0.02510000	-0.00574000	0.02100000
F2_20	-0.15197410	-0.093800000	-0.06502607	-0.23610000	0.08504274	-0.01807182	0.06222751
F2_23	-0.00129000	0.093400000	0.00249000	-0.06900000	0.04450000	-0.12500000	0.22600000
F2_24	-0.23600000	0.026900000	-0.10200000	0.01440000	0.00167000	-0.06820000	0.31100000
F2_26	-0.03070000	-0.133000000	0.14200000	0.03630000	-0.06800000	0.12500000	-0.20700000
F2_37	-0.02610000	0.075700000	-0.10200000	-0.01820000	0.00567000	0.00998000	0.12100000
F2_42	0.07370589	-0.009193803	0.06428929	0.47787460	-0.07534868	-0.03736660	0.18534580
F2_43	-0.04660000	-0.007500000	0.01690000	0.14400000	-0.06730000	-0.04020000	-0.13800000

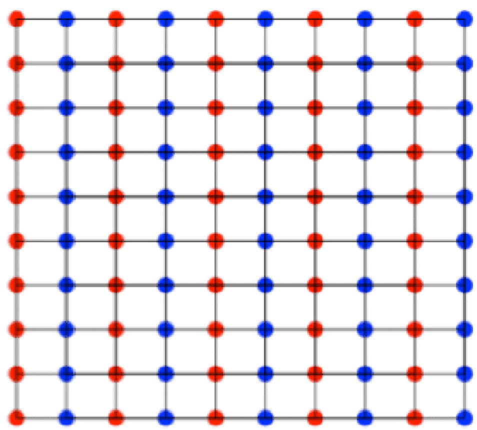
Data Challenges:

- Miss Value
- Low-frequency variant Features
- Outliers

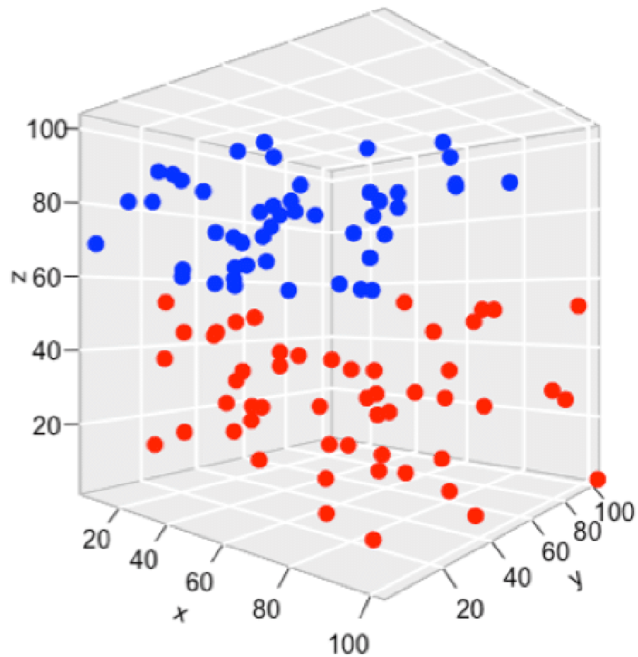


Blessing and Curse of Dimensionality

(A) 1-D



(B) 2-D



(C) 3-D

Algorithms

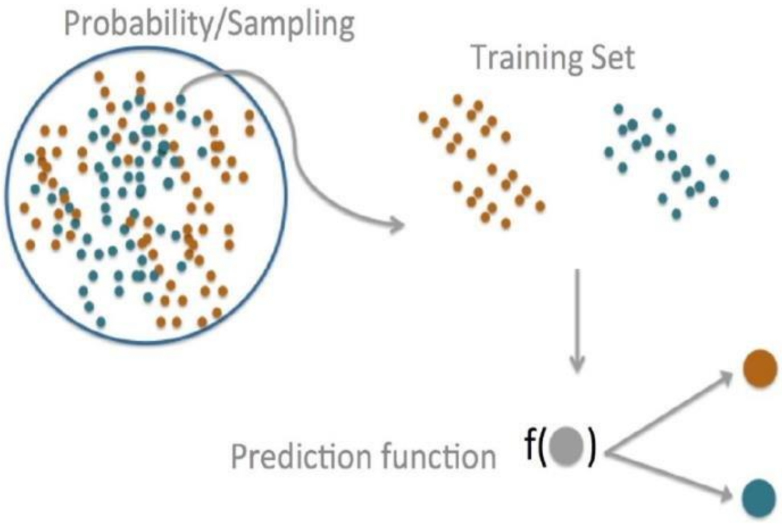
Algorithms

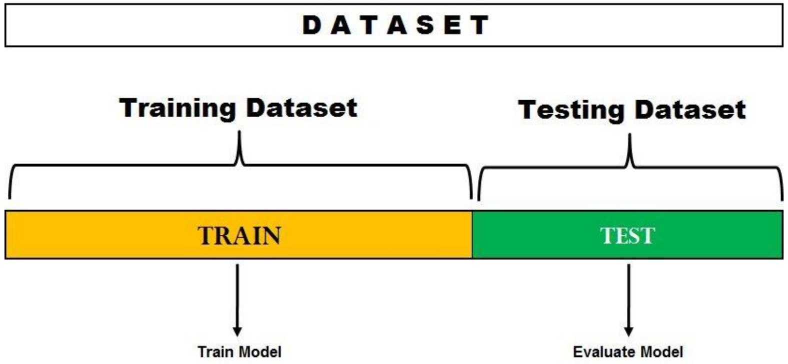
- Interpretable
- Simple
- Fast
- Accurate
- Scalable

Most Common Algorithms

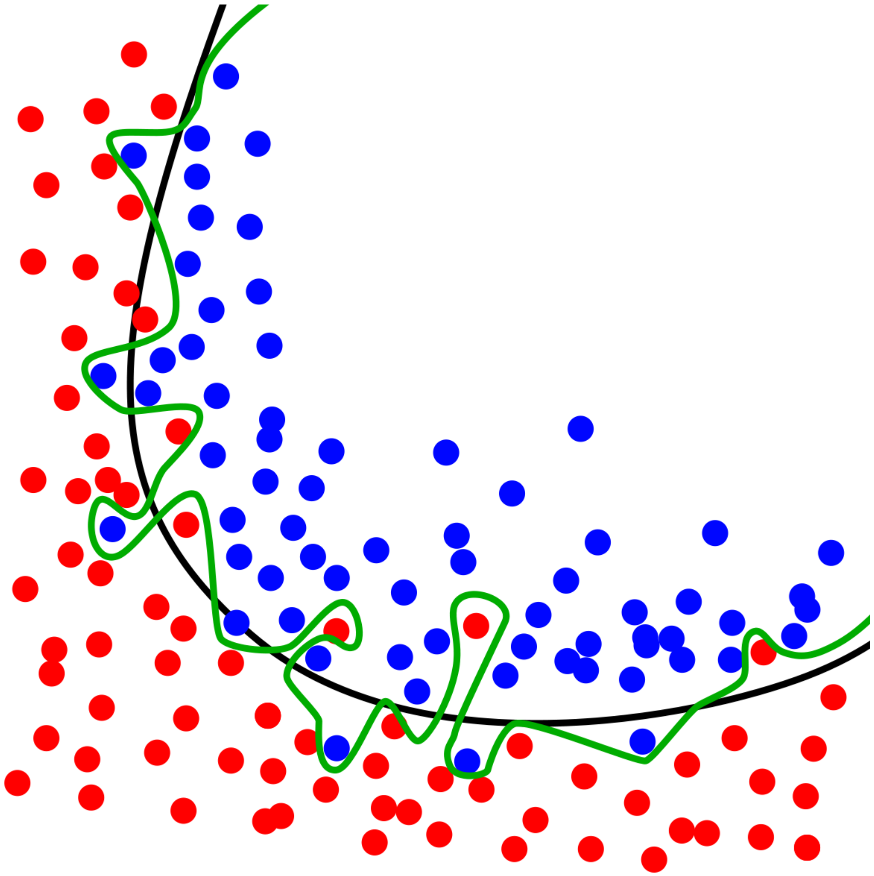
- KNN
- Regression
- MLP
- ANN
- SVM
- Decision Tree
- Random Forest
- Bayesian Net

Central Dogma of Prediction



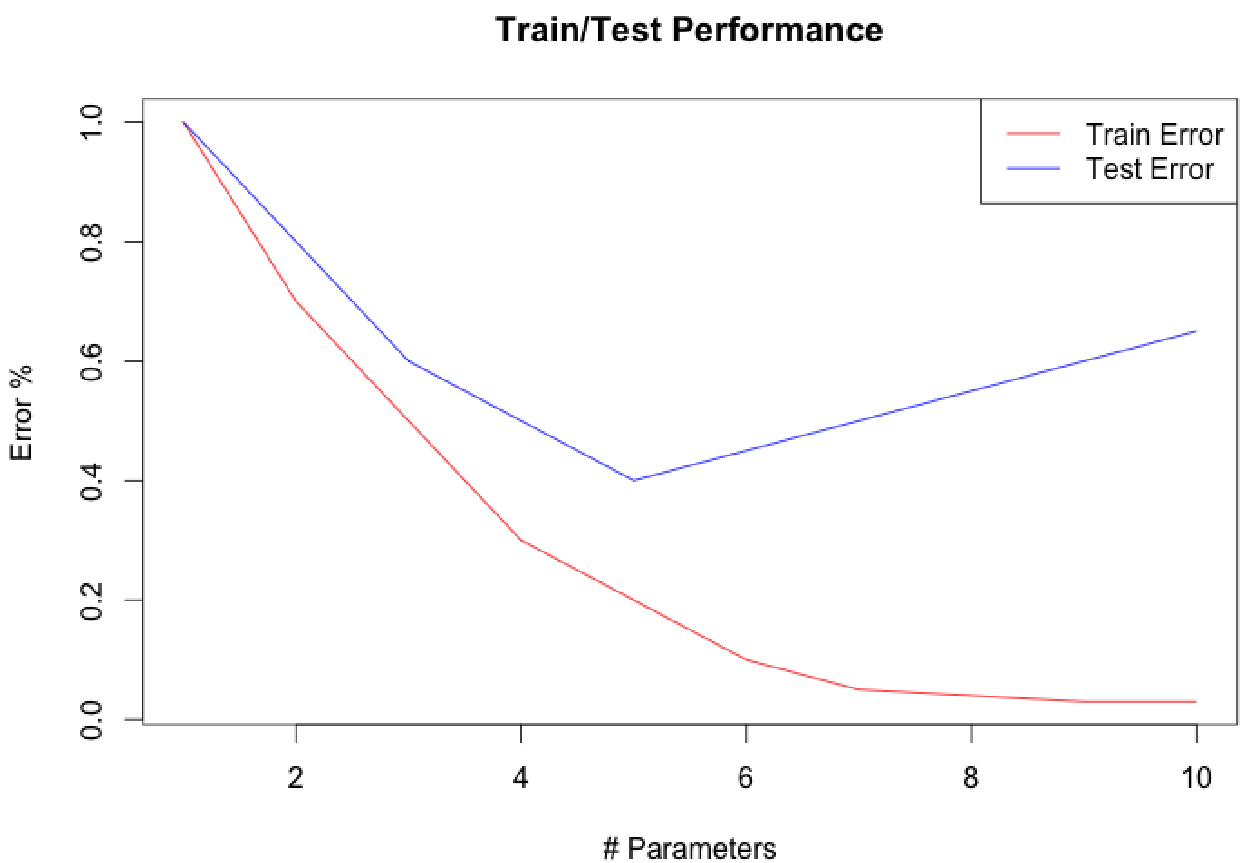


Algorithm (Model Selection): over fitting

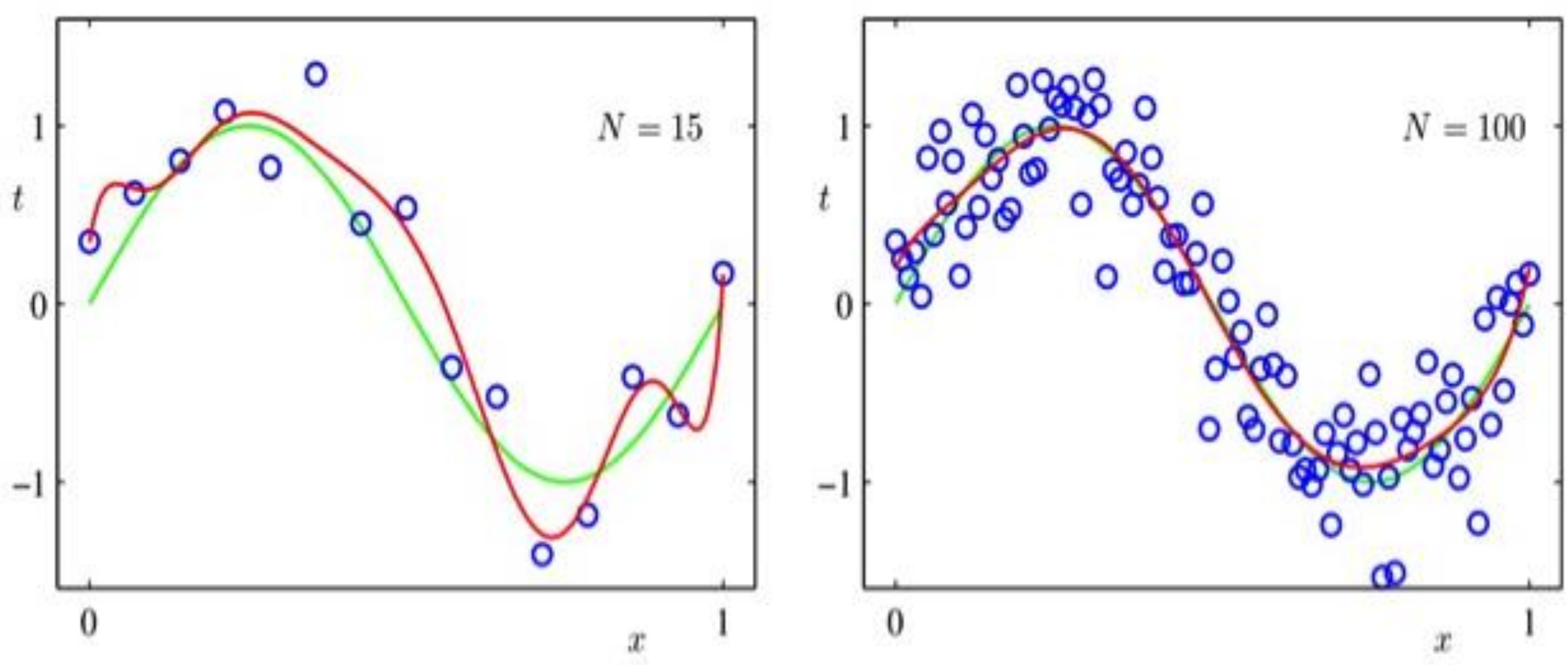


“All models are wrong,
but some are useful.”
George Box, British Statistician
1919-213

Algorithm (Model Selection): over fitting

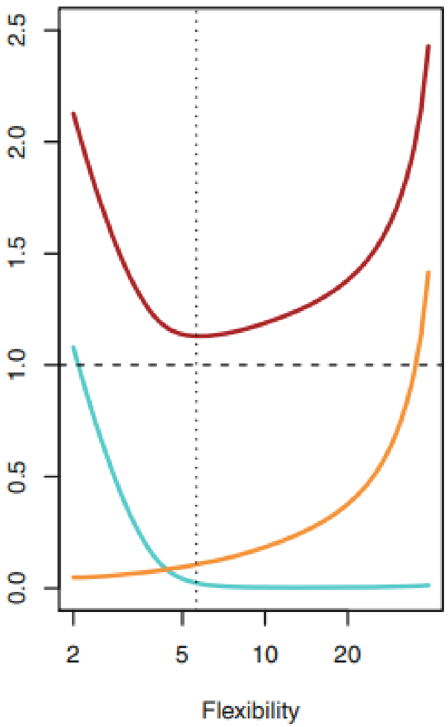


Algorithm (Model Selection): over fitting



Increasing the size of the data set may reduce the over-fitting

Algorithm (Model Selection): over fitting



Increase Flexibility:

- **Bias** tends to **initially decrease** faster than **variance increases**
- **At some point** has **little impact on the bias** but **starts to significantly increase the variance**.

Bias – variance Trade off

The Bias-Variance Trade-off

Expected error can always be **decomposed** into the sum of three fundamental quantities:

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

$$\boldsymbol{Use\ more\ flexible\ mothods\ \rightarrow \uparrow variance\ , bias \downarrow}$$

In Sample Error Vs on Sample error

In sample error

- Train data
- **Bias**

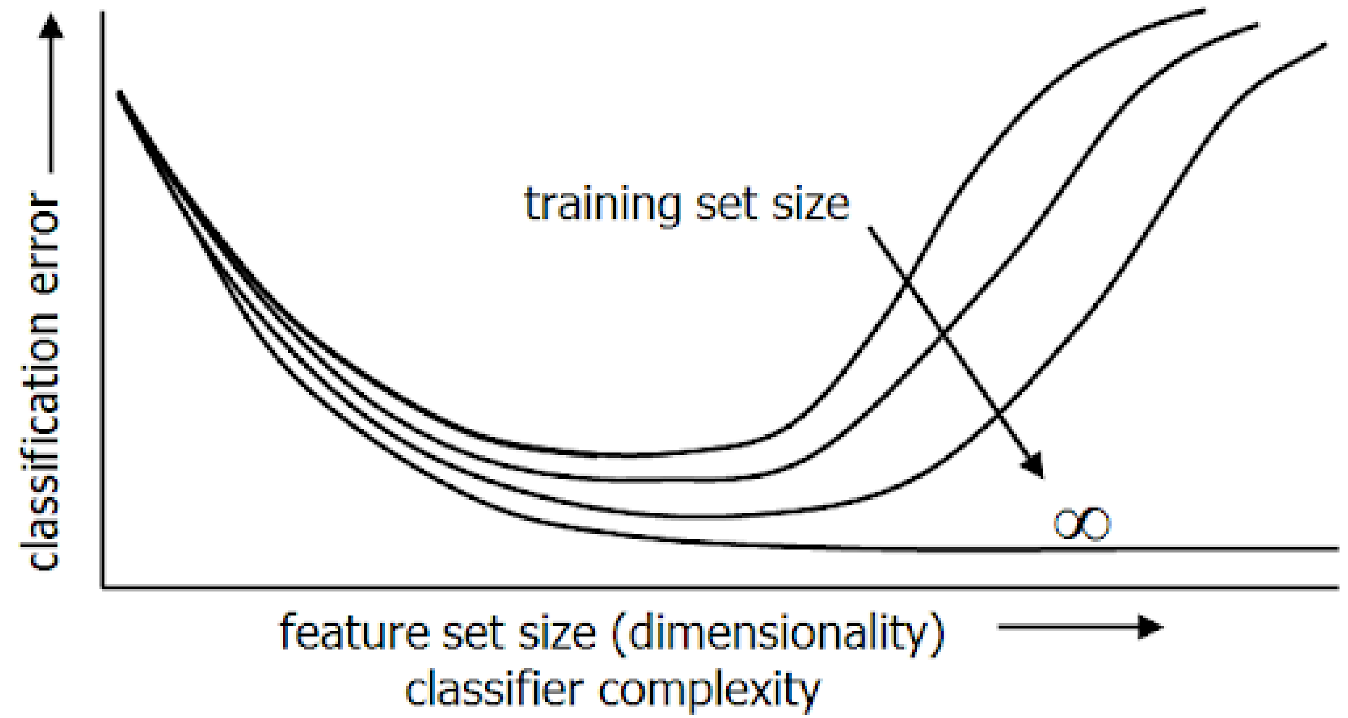
Out sample error

- Test data
- **Variance**

$$Usually\ Out\ sample\ error > In\ sample\ error$$

$$care\ about\ out\ sample\ error$$

The peaking paradox



“Machine learning is the next internet”

-Anthony Tether

Director, DARPA (Defense Advanced Research Projects Agency, USA).