Statistical Machine Learning

Pier Francesco Palamara
Department of Statistics
University of Oxford

Slide credits and other course material can be found at:
http://www.stats.ox.ac.uk/~palamara/SML20_BDI.html

Course Info

Course material:
`http://www.stats.ox.ac.uk/~palamara/SML20_BDI.html`

About me:
`https://palamaralab.github.io`

Course Aims

1. Understand statistical fundamentals of machine learning
   - Overview of unsupervised learning.
   - Supervised learning.
2. Understand difference between generative and discriminative learning frameworks.
3. Learn to identify and use appropriate methods and models for given data and task.
4. Learn to use the relevant R or python packages to analyse data, interpret results, and evaluate methods.

What is Machine Learning?

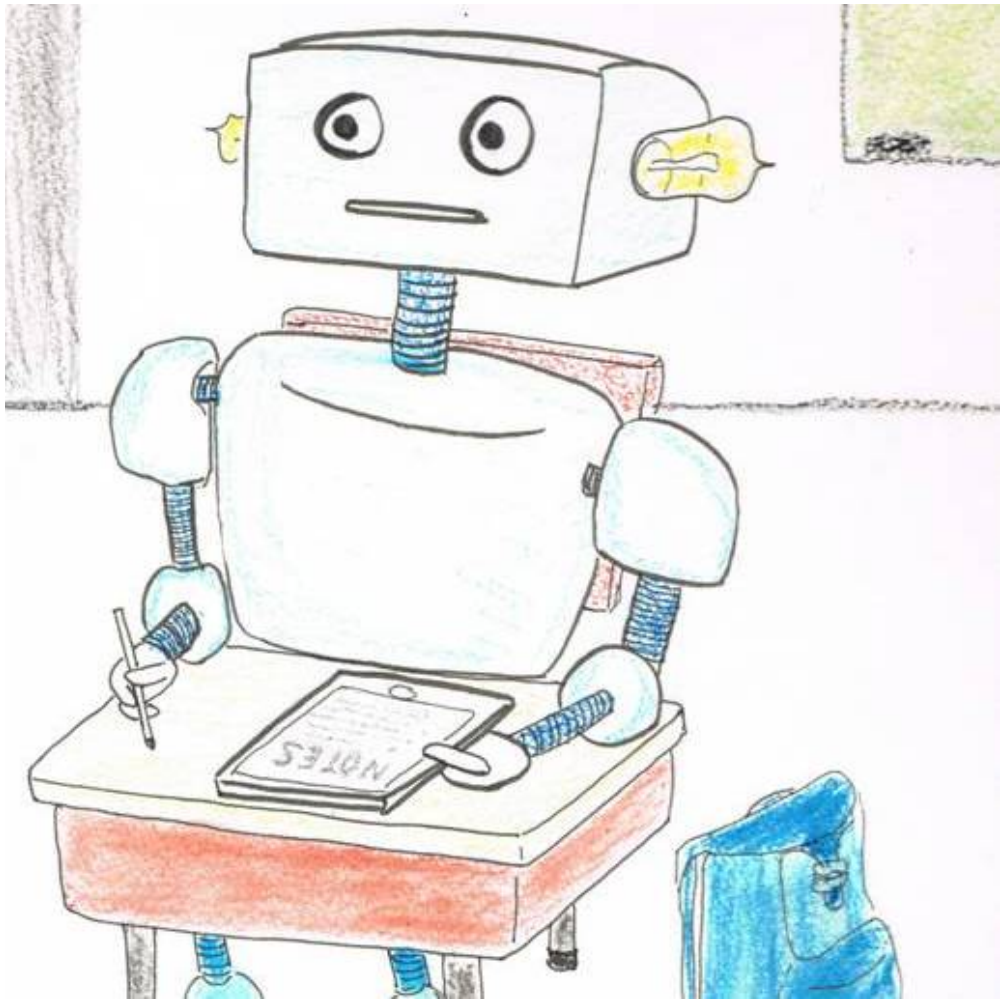


Image: www.gureckislab.org

Bottom-up, data-driven approach. Incoming data “improves” a model.

What is Machine Learning?

Arthur Samuel, 1959

Field of study that gives computers the ability to **learn** without being explicitly programmed.

What is Machine Learning?

Arthur Samuel, 1959

Field of study that gives computers the ability to **learn** without being explicitly programmed.

Tom Mitchell, 1997

Any computer program that **improves its performance** at some task **through experience**.

What is Machine Learning?

Arthur Samuel, 1959

Field of study that gives computers the ability to **learn** without being explicitly programmed.

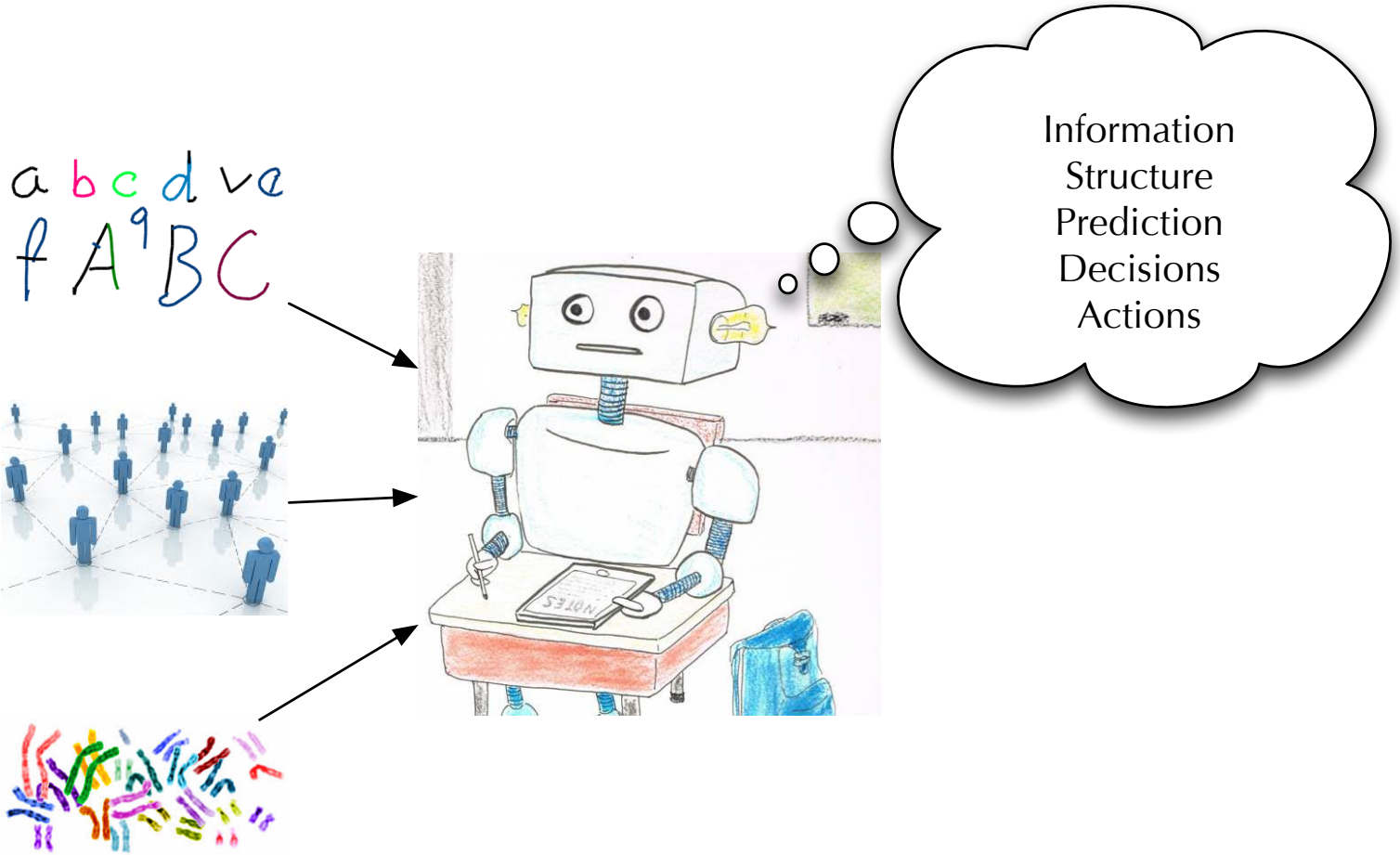
Tom Mitchell, 1997

Any computer program that **improves its performance** at some task **through experience**.

Kevin Murphy, 2012

To develop methods that can **automatically** detect **patterns in data**, and then to use the uncovered patterns to **predict** future data or other outcomes of interest.

What is Machine Learning?



Larry Page about DeepMind's ML systems that can learn to play video games like humans

Machine Learning: Historical Perspective (Stats)

- 1763: Bayes (Prior, Likelihood, Posterior)
- 1920's: Fisher (Maximum Likelihood)
- 1937: Pitman (Exponential Family)
- 1969: Jaynes (Maximum Entropy)
- 1970: Baum (Hidden Markov Models)
- 1978: Dempster (Expectation Maximization)
- 1980's: Vapnik (VC-Dimension)
- 1990: Lauritzen, Pearl (Graphical Models)
- 1990's: ... Kalman Filtering, Hidden Markov Models, Belief Nets, Markov Random Fields, SVMs, Learning Theory, Boosting, Kernels

Machine Learning: Historical Perspective (Bio/AI)

- 1917: Karel Capek (Robot)
- 1943: McCullogh & Pitts (Bio, Neuron)
- 1947: Norbert Weiner (Cybernetics, Multi-Disciplinary)
- 1949: Claude Shannon (Information Theory)
- 1950: Minsky, Newell, Simon, McCarthy (Symbolic AI, Logic)
- 1957: Rosenblatt (Perceptron)
- 1959: Arthur Samuel Coined “Machine Learning” (Learning Checkers)
- 1969: Minsky & Papert (Perceptron Linearity, no XOR)
- 1974: Werbos (BackProp, Nonlinearity)
- 1986: Rumelhart & McLelland (Connectionism, MLP, Verb-Conjugation)
- 1980’s: NeuralNets, Genetic Algos, Fuzzy Logic, Black Boxes
- Recent years: “Big Data”, Deep Learning, Machine Learning ubiquitous

Statistics vs Machine Learning

Traditional Problems in Applied Statistics

- Well formulated question that we would like to answer.
- Expensive data gathering and/or expensive computation.
- Create specially designed experiments to collect high quality data.

Information Revolution

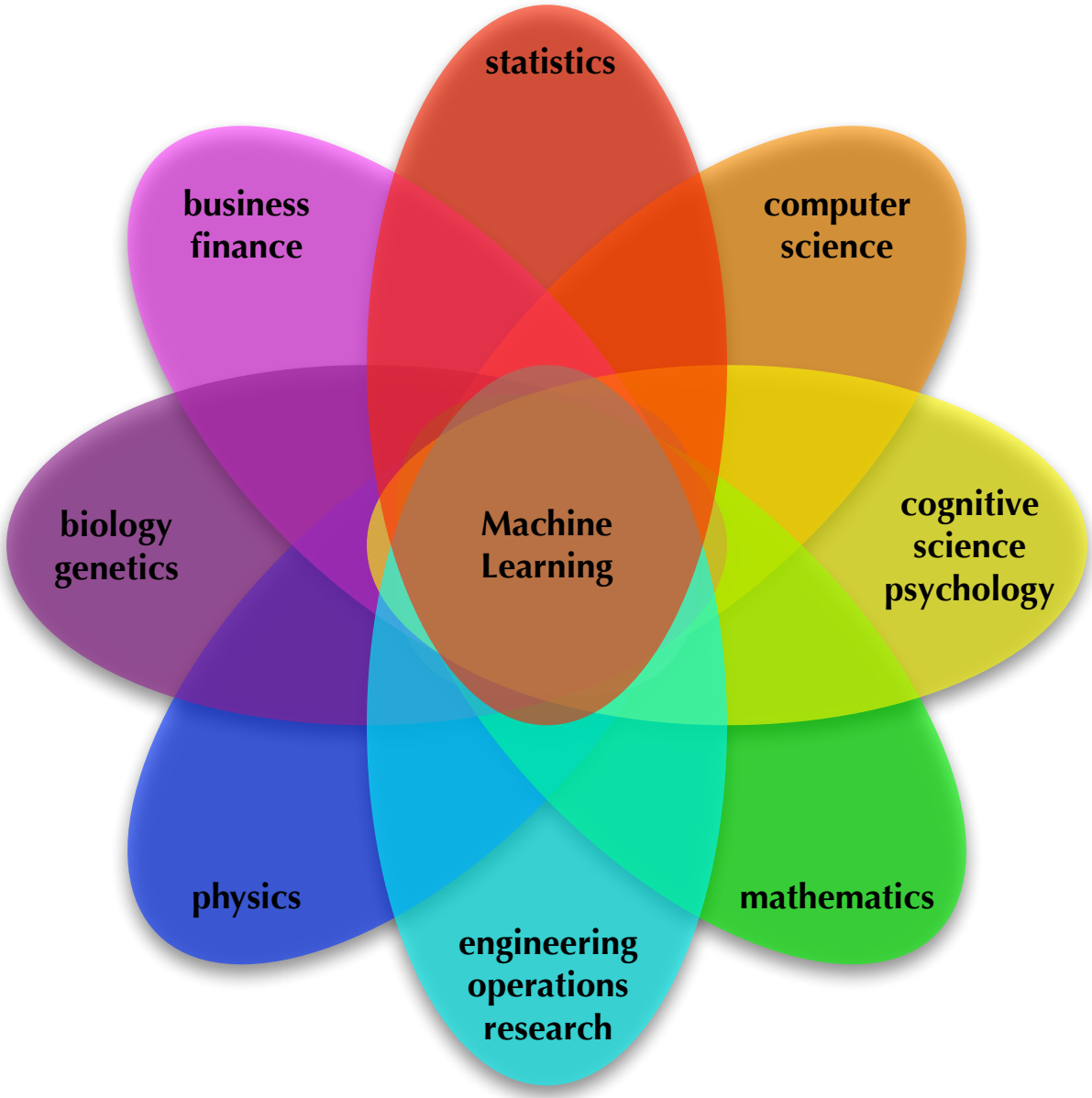
- Improvements in data processing and data storage.
- Powerful, cheap, easy data capturing.
- Lots of (low quality) data with **potentially valuable** information inside.

- CS and Stats forced **back together**: unified framework of data, inferences, procedures, algorithms
  - statistics taking computation seriously
  - computing taking statistical risk seriously

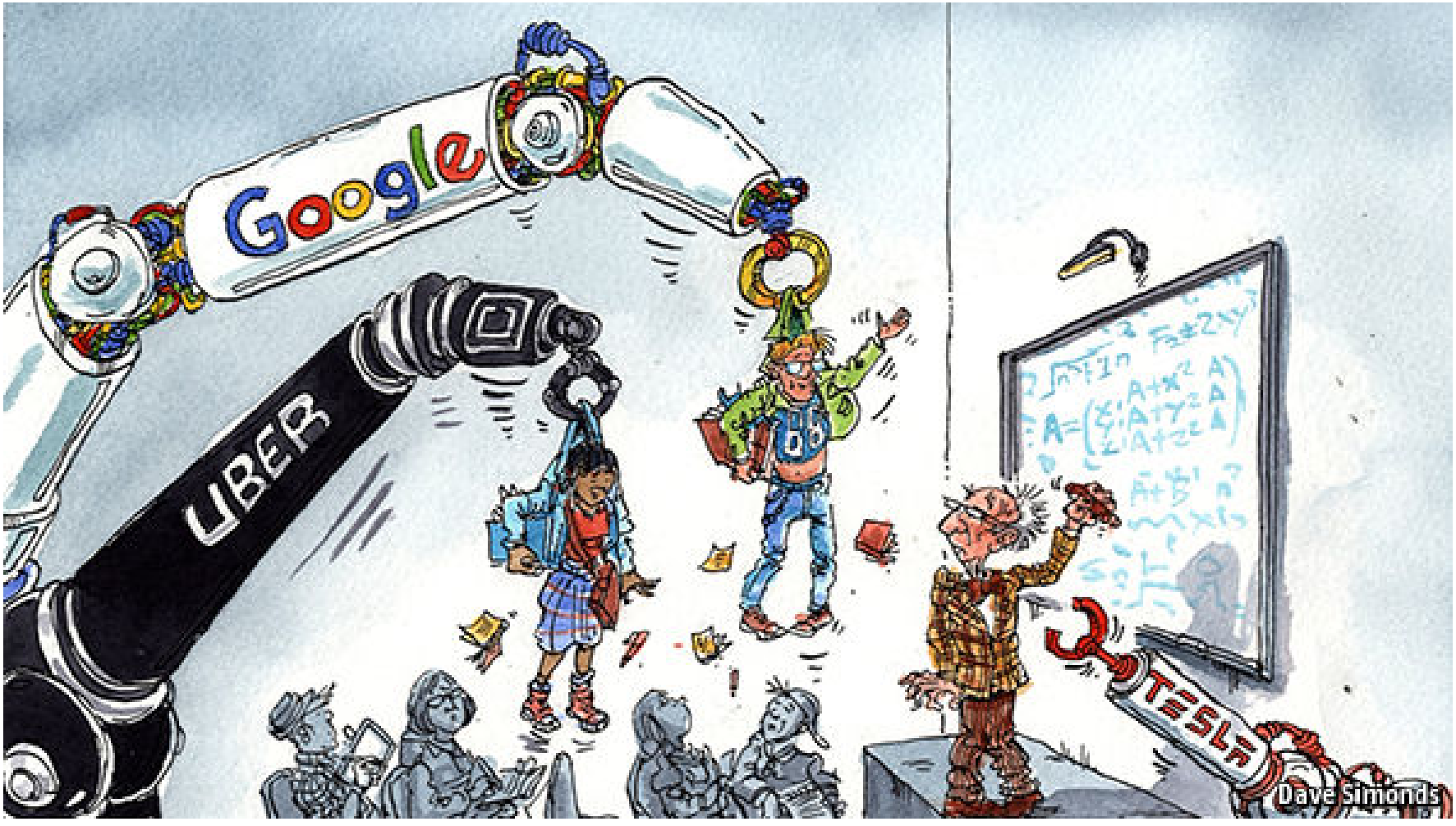
Michael I. Jordan: On the Computational and Statistical Interface and "Big Data"

Max Welling: Are Machine Learning and Statistics Complementary?

Machine Learning is a highly interdisciplinary field

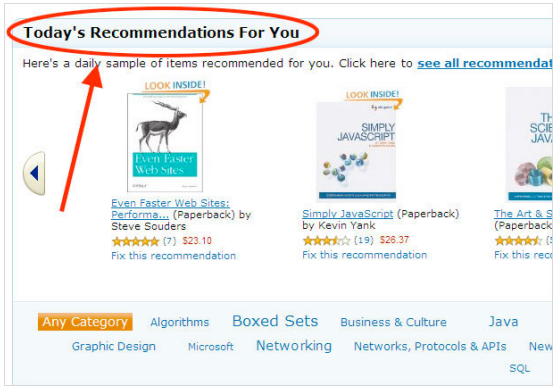


Machine Learning + industry



Applications of Machine Learning

spam filtering



recommendation systems

fraud detection

self-driving cars

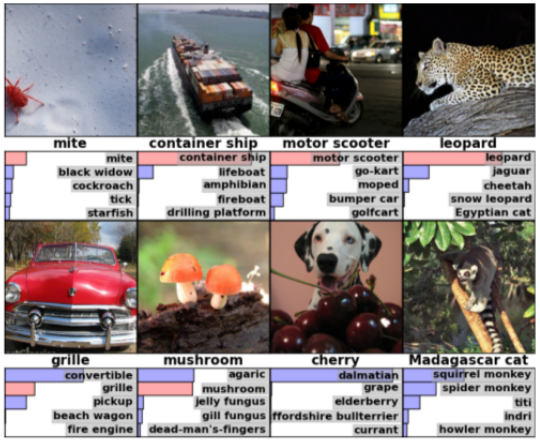


image recognition

stock market analysis

ImageNet: Computer Eyesight Gets a Lot More Accurate, Krizhevsky et al, 2012

New applications of ML: Machine Learning is Eating the World

Types of Machine Learning

Unsupervised learning

- Extract key features of the “unlabelled” data
- clustering, signal separation, density estimation
- Goal: **representation, hypothesis generation, visualization**

Supervised learning

- Data contains “labels”: every example is an input-output pair
- classification, regression
- Goal: **prediction on new examples**

Types of Machine Learning

Semi-supervised Learning

A database of examples, only a small subset of which are labelled.

Multi-task Learning

A database of examples, each of which has multiple labels corresponding to different prediction tasks.

Reinforcement Learning

An agent acting in an environment, given rewards for performing appropriate actions, learns to maximize their reward.

Software

- Python: scikit-learn, mlpy, Theano
- R
- Matlab/Octave
- Weka, mlpack, Torch, Shogun, TensorFlow...

Syllabus I

Part I: Introduction to unsupervised learning (~ 3 lectures)

- Dimensionality reduction
  - Principal component analysis, SVD, Biplots, Multidimensional scaling, Isomap
- Clustering
  - K-means
  - Hierarchical clustering

Syllabus II

Part II: Supervised learning (~ 7 lectures)

- Empirical risk minimization
- Regression
  - Linear
  - Non-linear basis functions
- Overfitting, cross-validation
- Regularization
- Bias/variance tradeoff
- Classification
  - Linear discriminant analysis
  - Logistic regression
  - Naïve Bayes
  - Perceptron
  - K-nearest neighbors
- Generative vs discriminative methods
- Performance evaluation

Syllabus III

Part III: Useful algorithms for supervised learning (~ 5 lectures)

- Decision trees
- Ensamle methods
  - Bagging
  - Random forests
  - Boosting
- Problems / other topics TBD if time allows

Unsupervised Learning: Visualisation and Dimensionality Reduction

Unsupervised Learning

Goals:

- Find the variables that summarise the data / capture relevant information.
- Discover informative ways to visualise the data.
- Discover the subgroups among the observations.

It is often much easier to obtain unlabeled data than labeled data!

Exploratory Data Analysis

Notation

- Data consists of p variables (features/attributes/dimensions) on n examples (items/observations).
- $\mathbf{X} = (x_{ij})$ is a $n \times p$-matrix with $x_{ij} :=$ the j-th variable for the i-th example

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix}.$$

- Denote the i-th data item by $x_i \in \mathbb{R}^p$ (we will treat it as a column vector: it is the transpose of the i-th row of $\mathbf{X}$).
- Assume $x_1, \dots, x_n$ are **independently and identically distributed** samples of a **random vector** X over $\mathbb{R}^p$. The j-th dimension of X will be denoted $X^{(j)}$.

Crabs Data ($n = 200$, $p = 5$)

Campbell (1974) studied rock crabs of the genus **leptograpsus**. One species, **L. variegatus**, had been split into two new species according to their colour: orange and blue. Preserved specimens lose their colour, so it was hoped that morphological differences would enable museum material to be classified.

Data are available on 50 specimens of each sex of each species. Each specimen has measurements on:

- the width of the frontal lobe `FL`,
- the rear width `RW`,
- the length along the carapace midline `CL`,
- the maximum width `CW` of the carapace, and
- the body depth `BD` in mm.

photo from: inaturalist.org

in addition to colour/species and sex (we will later view these as labels, but will ignore for now).

Crabs Data

```
## load package MASS containing the data
library(MASS)

## extract variables we will look at
varnames<-c("FL","RW","CL","CW","BD")
Crabs <- crabs[,varnames]

## look at raw data
Crabs
```

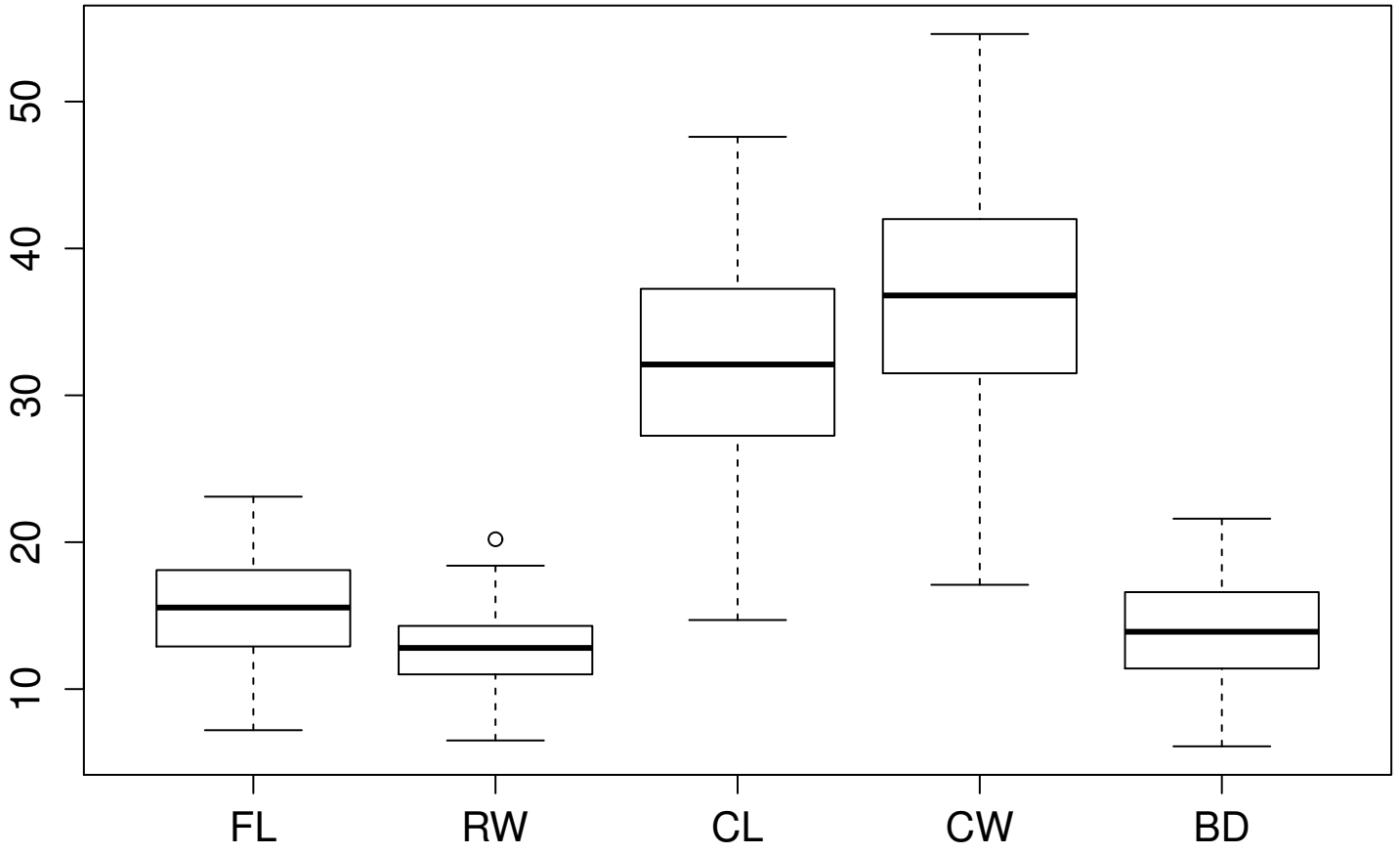
Crabs Data

```
## look at raw data
Crabs
```

	FL	RW	CL	CW	BD
1	8.1	6.7	16.1	19.0	7.0
2	8.8	7.7	18.1	20.8	7.4
3	9.2	7.8	19.0	22.4	7.7
4	9.6	7.9	20.1	23.1	8.2
5	9.8	8.0	20.3	23.0	8.2
6	10.8	9.0	23.0	26.5	9.8
7	11.1	9.9	23.8	27.1	9.8
8	11.6	9.1	24.5	28.4	10.4
9	11.8	9.6	24.2	27.8	9.7
10	11.8	10.5	25.2	29.3	10.3
11	12.2	10.8	27.3	31.6	10.9
12	12.3	11.0	26.8	31.5	11.4
13	12.6	10.0	27.7	31.7	11.4
14	12.8	10.2	27.2	31.8	10.9
15	12.8	10.9	27.4	31.5	11.0
16	12.9	11.0	26.8	30.9	11.4
17	13.1	10.6	28.2	32.3	11.0
18	13.1	10.9	28.3	32.4	11.2
19	13.3	11.1	27.8	32.3	11.3
20	13.9	11.1	29.2	33.3	12.1

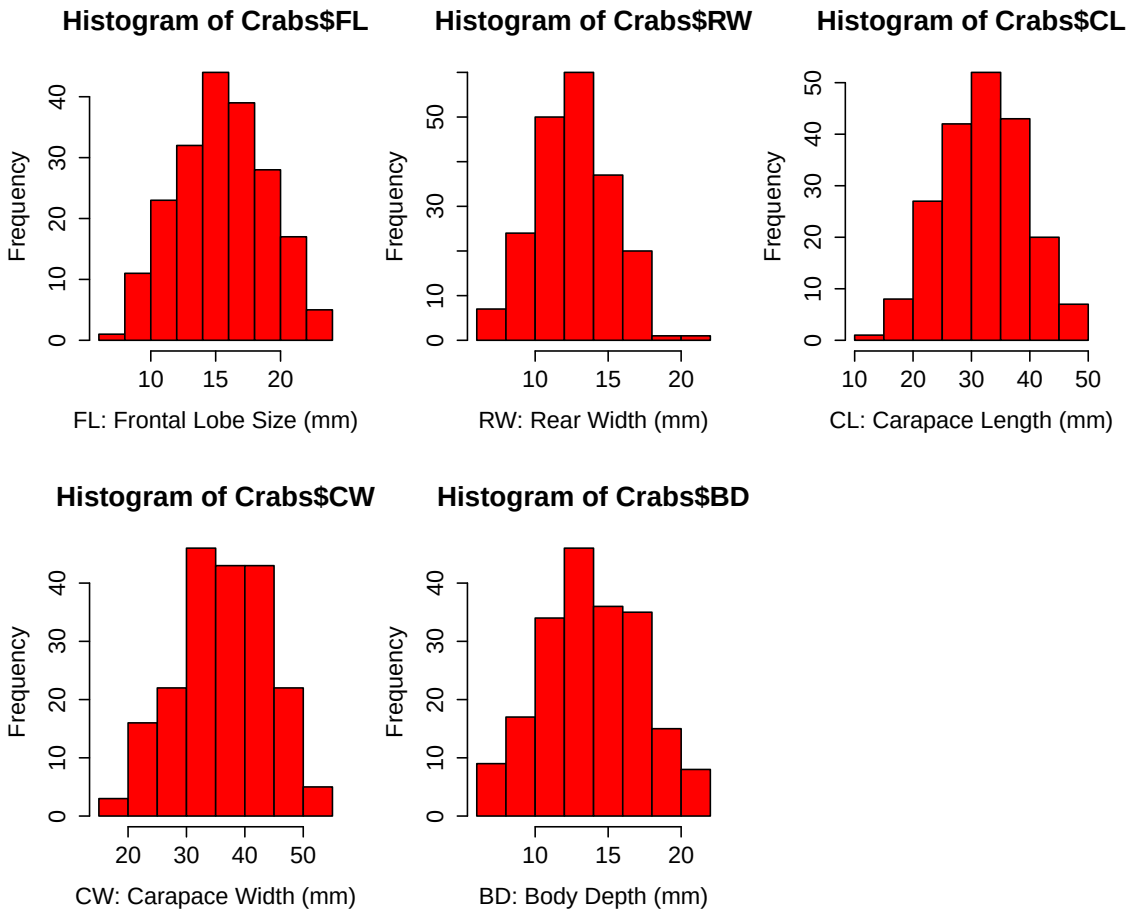
Univariate Boxplots

```
boxplot(Crabs)
```



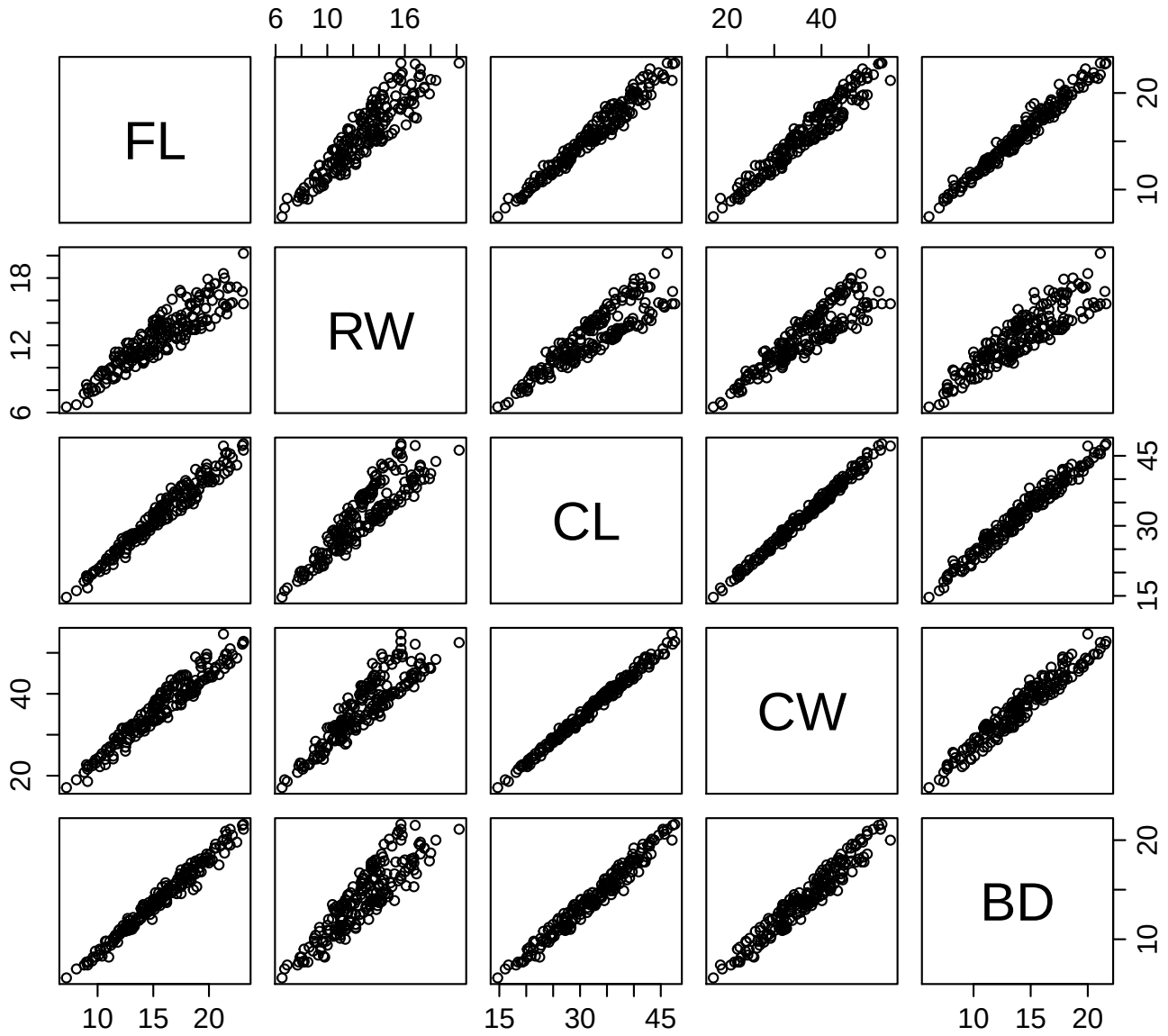
Univariate Histograms

```
par(mfrow=c(2,3))
hist(Crabs$FL,col="red",xlab="FL: Frontal Lobe Size (mm)")
hist(Crabs$RW,col="red",xlab="RW: Rear Width (mm)")
hist(Crabs$CL,col="red",xlab="CL: Carapace Length (mm)")
hist(Crabs$CW,col="red",xlab="CW: Carapace Width (mm)")
hist(Crabs$BD,col="red",xlab="BD: Body Depth (mm)")
```



Simple Pairwise Scatterplots

```
pairs(Crabs)
```



Visualisation and Dimensionality Reduction

The summary plots are useful, but limited use if the dimensionality p is high (a few dozens or even thousands).

- Constrained to view data in 2 or 3 dimensions
- Approach: look for 'interesting' projections of $\mathbf{X}$ into lower dimensions
- Hope that even though p is large, considering only carefully selected $k \ll p$ dimensions is just as informative.

Dimensionality reduction

- For each data item $x_i \in \mathbb{R}^p$, find its lower dimensional representation $z_i \in \mathbb{R}^k$ with $k \ll p$.
- Map $x \mapsto z$ should preserve the **interesting statistical properties** in data.