

# Machine Learning and Data Analytics

Eric Medvet    Matilde Trevisani

A.A. 2018/2019

# Section 1

## General information

# Lecturers

- ▶ Matilde Trevisani
  - ▶  $\approx 24\text{ h}$  (12 CFU only) +  $\approx 24\text{ h}$
  - ▶ Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche (DEAMS)
  - ▶ <http://www.units.it/persone/index.php/from/abook/persona/8754>
- ▶ Eric Medvet
  - ▶  $\approx 48\text{ h}$
  - ▶ Dipartimento di Ingegneria e Architettura (DIA)
  - ▶ <http://medvet.inginf.units.it/>

# Course materials

- ▶ Lecturer's slides (my part)
  - ▶ [http://medvet.inginf.units.it/teaching/  
machine-learning-and-data-analytics-2018-2019](http://medvet.inginf.units.it/teaching/machine-learning-and-data-analytics-2018-2019)
- ▶ Suggested textbooks (for further reading)
  - ▶ Gareth James et al. *An introduction to statistical learning*. Vol. 6. Springer, 2013
  - ▶ Kenneth A De Jong. *Evolutionary computation: a unified approach*. MIT press, 2006
- ▶ Other material:
  - ▶ I'll point you to some scientific papers for discussing examples of application or specific details—just a “chat”

Everything you are required to know is in the lecturer's slides

# How to attend lectures

Everything you are required to know is in the lecturer's slides

**But:** slides are designed assuming that **you** are attending the lecture and **taking notes**

However:

- ▶ lectures will be recorded: <https://videocenter.units.it>
- ▶ during the lectures, I'll (hopely) use the interactive whiteboard for writing annotations
- ▶ including answers to questions posed in the slides, e.g.,
  - ▶ **Q:** is this working? *YES*
- ▶ the annotated slides will be available on my website

# How to attend lectures: lab activities

Focus is on methodology, rather than on theory behind techniques:  
how to tackle a problem with ML?

Practicing (in tackling problems) is crucial → lab activities

- ▶ ≈ 13h
- ▶ mainly **design**, then implementation
  - ▶ **you** practice
  - ▶ I am available any time during/before/after for advising
    - ▶ there's a tutor: Marco Zullich
  - ▶ in general, there is no **one** solution; you make the solution better or worse while (virtually) presenting it
  - ▶ we'll analyze in depth at least one solution
- ▶ form small (2–4) groups
  - ▶ possibly with different background
  - ▶ peer-tutoring

# Exam

Either:

- ▶ a student project **and** a written test
- ▶ a larger written test

Written test: questions on theory and application with medium- and short-length open answers

Project: design, develop, and assess an ML system, choosing among a few options (see

[http://medvet.inginf.units.it/teaching/  
machine-learning-and-data-analytics-2018-2019/  
student-project](http://medvet.inginf.units.it/teaching/machine-learning-and-data-analytics-2018-2019/student-project))

You?

DSSC 27

IN EL INF 19

STATS 20

Who are you?

ECON 1

MATH 2

PHY 1

---

70

## Section 2

### Introduction

# What is Machine Learning?

## Definition

**Machine Learning** is the science of getting computer to learn without being explicitly programmed.

## Definition

**Data Mining/Analytics** is the science of discovering patterns in data.

## In practice

A set of mathematical and statistical tools for:

- ▶ building a model which allows to predict an output, given an input (*supervised learning*)
  - ▶ example ⟨input, output⟩ pairs are available
- ▶ learn relationships and structures in data (*unsupervised learning*)

# Machine Learning everyday

Example problem: spam

Discriminate between spam and non-spam emails.

The screenshot shows a Gmail inbox search results page for the query "in:spam". The search bar at the top contains "in:spam". Below it, the inbox header includes "Gmail" with a dropdown arrow, a compose button, and a "More" button. The left sidebar has a red "COMPOSE" button and lists categories: Inbox (3), Starred, Important, Chats, Sent Mail, Drafts, Spam (526) (which is highlighted in red), Categories, Social, Promotions (1), Updates (1), Purchases, Travel, and Finance. The main area displays 10 search results for spam emails, each with a checkbox, a star icon, and a subject line followed by a snippet of text. A "Delete all spam" link is visible at the top right of the list.

Subject Line	Snippet
CSC Conference Secretari.	Call for Papers : 1st Annual Intern
Alexander Horn	Recently posted academic job vac
Regalo di Benvenuto	emedvet@units.it per te uno Smar
Peugeot Italia	Peugeot supervaluta il tuo usato. I
CAP petite enfance	votre profil nous intéresse - Vous r
Rachat de crédits	Réduisez vos mensualités jusqu'à
Zalando	Le sneakers che conquistano la st
Sondage National	Pour ou contre passer à 90 km/h s
Oroscopo	Messaggio Privato per - Stai riceve
Secret Escapes	Sconti Imbattibili su Hotel e Vacan
Erogazione credito appro.	Fino a 50.000 euro, anche protesta

Figure: Spam filtering in Gmail.

# Machine Learning everyday

Example problem: flight trajectories

Do flights over the same pair  $\langle \text{origin}, \text{destination} \rangle$  follow the “same” trajectory? Why?

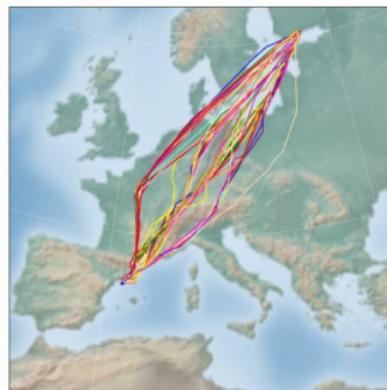


Figure: Clustering of flight trajectories.

# Machine Learning everyday

Example problem: image understanding

Recognize objects in images.

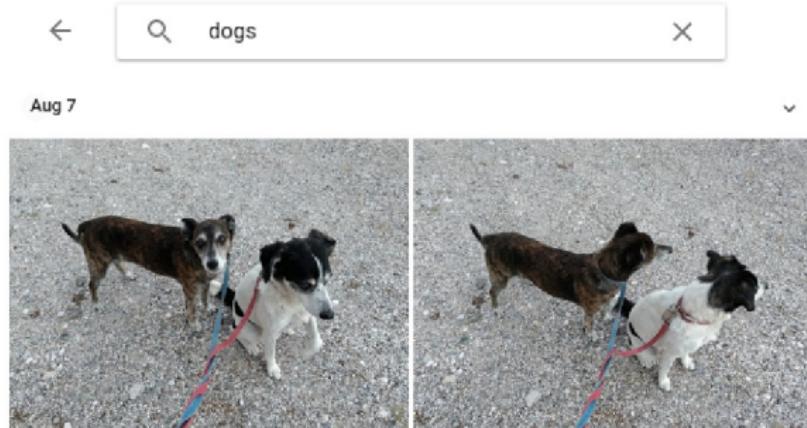


Figure: Object recognition in Google Photos.

# Machine Learning everyday

Example problem: authoring regular expressions

Write a regular expression from matching examples.

## RegexGenerator++

Automatic Generation of Text Extraction Patterns from Examples

### Dataset (2 examples)

#### Example

- We try to quantitatively capture these characteristics by defining a set of indexes, which can be used to quickly compare different images.
- After applying a method to an image, we compare the segmented image (i.e., the result) i

+ New example

Import

Clear dataset

Export dataset

Try an example!

#### Result

```
\$[^\$]*+\$
```

Figure: Regex generation with <http://regex.inginf.units.it/>.

## Machine Learning everyday

**Q:** what type of learning (supervised/unsupervised) is in the examples? ANSWERED

- ▶ examples?  
prob ably supervised
  - ▶ spam supervised
  - ▶ image understanding supervised
  - ▶ flight trajectories unsupervised
  - ▶ authoring regular expressions none of the 2

# Why ML/DM “today”?

- ▶ we collect more and more data (*big data*)
- ▶ we have more and more computational power

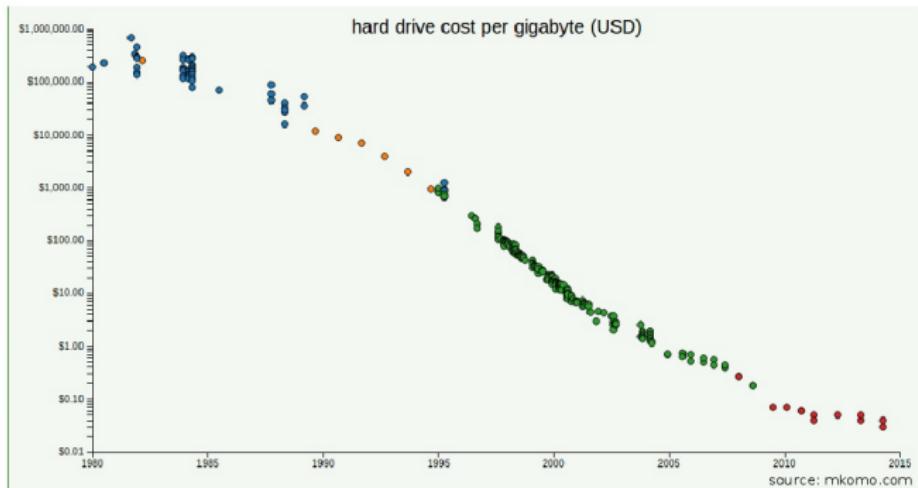


Figure: From <http://www.mkomo.com/cost-per-gigabyte-update>.

# ML/DM is popular!

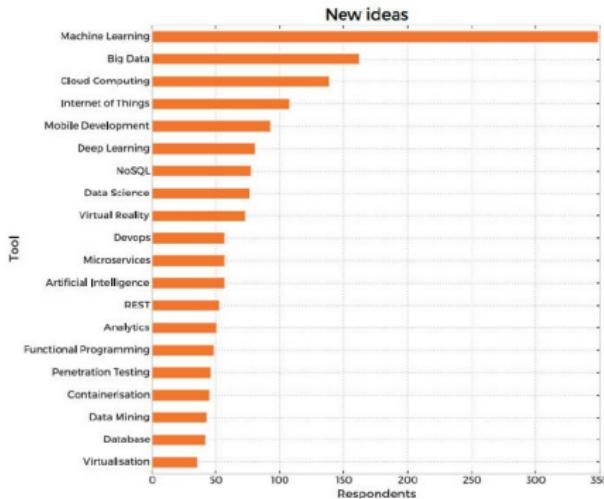


Figure: Popular areas of interest, from the Skill Up 2016: Developer Skills Report<sup>2</sup>

<sup>1</sup><https://techcus.com/p/r1zSmbXut/>

top-5-highest-paying-programming-languages-of-2016/.

<sup>2</sup><https://techcus.com/p/r1zSmbXut/>

top-5-highest-paying-programming-languages-of-2016/.

# Aims of the course

Be able to:

1. design
2. implement
3. assess experimentally

an end-to-end Machine Learning or Data Mining system.

# Aims of the course

Be able to:

1. design
2. implement
3. assess experimentally

an end-to-end Machine Learning or Data Mining system.

- ▶ Which is the problem to be solved? Which are the input and output? Which are the most suitable techniques? How should data be prepared? Does computation time matter?

# Aims of the course

Be able to:

1. design
2. **implement**
3. assess experimentally

an end-to-end Machine Learning or Data Mining system.

- ▶ Which is the problem to be solved? Which are the input and output? Which are the most suitable techniques? How should data be prepared? Does computation time matter?
- ▶ Write some code!

# Aims of the course

Be able to:

1. design
2. implement
3. **assess experimentally**

an end-to-end Machine Learning or Data Mining system.

- ▶ Which is the problem to be solved? Which are the input and output? Which are the most suitable techniques? How should data be prepared? Does computation time matter?
- ▶ Write some code!
- ▶ How to measure solution quality? How to compare solutions? Is my solution general?

# Aims of the course

Be able to:

1. design
2. implement
3. assess experimentally

an end-to-end Machine Learning or Data Mining system.

- ▶ Which is the problem to be solved? Which are the input and output? Which are the most suitable techniques? How should data be prepared? Does computation time matter?
- ▶ Write some code!
- ▶ How to measure solution quality? How to compare solutions? Is my solution general?
  - ▶ Itself: design and implementation

# Aims of the course: communication

Be able to:

1. design
2. implement
3. assess experimentally

an end-to-end Machine Learning or Data Mining system.

And be able to convince the “client” that it is:

- ▶ technically sound
- ▶ economically viable
- ▶ in its larger context

## Subsection 1

Motivating example

## The amateur botanist friend

He likes to collect Iris plants. He “realized” that there are 3 species, in particular, that he likes: *Iris setosa*, *Iris virginica*, and *Iris versicolor*. He’d like to have a tool to automatically *classify* collected samples in one of the 3 species.



Figure: Iris versicolor.

How to help him?

## Let's help him

- ▶ Which is the problem to be solved?

## Let's help him

- ▶ Which is the problem to be solved?
  - ▶ Assign exactly one specie to a sample.

## Let's help him

- ▶ Which is the problem to be solved?
  - ▶ Assign exactly one specie to a sample.
- ▶ Which are the input and output?

## Let's help him

- ▶ Which is the problem to be solved?
  - ▶ Assign exactly one species to a sample.
- ▶ Which are the input and output?
  - ▶ Output: one species among I. setosa, I. virginica, I. versicolor.

## Let's help him

- ▶ Which is the problem to be solved?
  - ▶ Assign exactly one species to a sample.
- ▶ Which are the input and output?
  - ▶ Output: one species among I. setosa, I. virginica, I. versicolor.
  - ▶ Input: the plant sample...

## Let's help him

- ▶ Which is the problem to be solved?
  - ▶ Assign exactly one species to a sample.
- ▶ Which are the input and output?
  - ▶ Output: one species among I. setosa, I. virginica, I. versicolor.
  - ▶ Input: the plant sample...
    - ▶ a description in natural language?

## Let's help him

- ▶ Which is the problem to be solved?
  - ▶ Assign exactly one species to a sample.
- ▶ Which are the input and output?
  - ▶ Output: one species among I. setosa, I. virginica, I. versicolor.
  - ▶ Input: the plant sample...
    - ▶ a description in natural language?
    - ▶ a digital photo?

## Let's help him

- ▶ Which is the problem to be solved?
  - ▶ Assign exactly one species to a sample.
- ▶ Which are the input and output?
  - ▶ Output: one species among I. setosa, I. virginica, I. versicolor.
  - ▶ Input: the plant sample...
    - ▶ a description in natural language?
    - ▶ a digital photo?
    - ▶ DNA sequences?

## Let's help him

- ▶ Which is the problem to be solved?
  - ▶ Assign exactly one species to a sample.
- ▶ Which are the input and output?
  - ▶ Output: one species among *I. setosa*, *I. virginica*, *I. versicolor*.
  - ▶ Input: the plant sample...
    - ▶ a description in natural language?
    - ▶ a digital photo?
    - ▶ DNA sequences?
    - ▶ some measurements of the sample!

## Iris: input and output



Figure: Sepal and petal.

Input: sepal length and width, petal length and width (in cm)

Output: the class

Example:  $(5.1, 3.5, 1.4, 0.2) \rightarrow I. \text{ setosa}$

## Other information

The botanist friend asked a senior botanist to inspect several samples and **label** them with the corresponding species.

	Sepal length	Sepal width	Petal length	Petal width	Species
A FLOWER →	5.1	3.5	1.4	0.2	I. setosa
B FLOWER →	4.9	3.0	1.4	0.2	I. setosa
	7.0	3.2	4.7	1.4	I. versicolor
	6.0	2.2	5.0	1.5	I. virginica

## Notation and terminology

- ▶ Sepal length, sepal width, petal length, and petal width are **input variables** (or independent variables, or features, or attributes).
- ▶ Species is the **output variable** (or dependent variable, or response).

## Notation and terminology

$$\mathbf{X} = \begin{pmatrix} \textcolor{red}{x_{1,1}} & \textcolor{red}{x_{1,2}} & \cdots & \textcolor{red}{x_{1,p}} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- ▶  $x_1^T = (x_{1,1}, x_{1,2}, \dots, x_{1,p})$  is an **observation** (or instance, or data point), composed of  $p$  variable values;

## Notation and terminology

$$\mathbf{X} = \begin{pmatrix} \textcolor{red}{x_{1,1}} & \textcolor{red}{x_{1,2}} & \cdots & \textcolor{red}{x_{1,p}} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} \textcolor{red}{y_1} \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- ▶  $x_1^T = (x_{1,1}, x_{1,2}, \dots, x_{1,p})$  is an **observation** (or instance, or data point), composed of  $p$  variable values;  $y_1$  is the corresponding output variable value

## Notation and terminology

$m \ll p$

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \textcolor{red}{x_{1,2}} & \cdots & x_{1,p} \\ x_{2,1} & \textcolor{red}{x_{2,2}} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & \textcolor{red}{x_{n,2}} & \cdots & x_{n,p} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- ▶  $x_1^T = (x_{1,1}, x_{1,2}, \dots, x_{1,p})$  is an **observation** (or instance, or data point), composed of  $p$  variable values;  $y_1$  is the corresponding output variable value
- ▶  $\mathbf{x}_2^T = (x_{1,2}, x_{2,2}, \dots, x_{n,2})$  is the vector of all the  $n$  values for the 2nd variable ( $X_2$ ).

# Notation and terminology

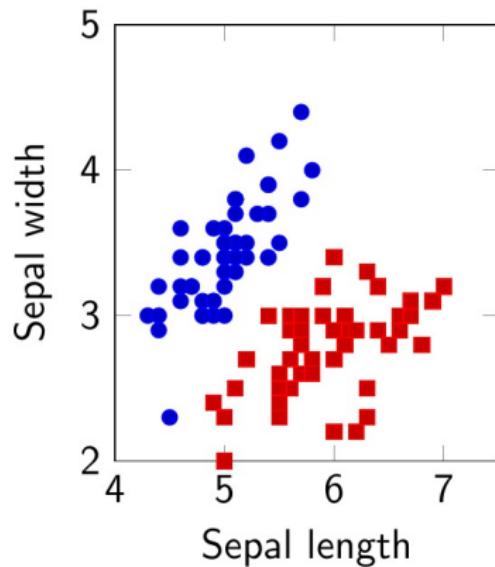
Different communities (e.g., statistical learning vs. machine learning vs. artificial intelligence) use different terms and notation:

- ▶  $x_j^{(i)}$  instead of  $x_{i,j}$  (hence  $x^{(i)}$  instead of  $x_i$ )
- ▶  $m$  instead of  $n$  and  $n$  instead of  $p$
- ▶ ...

Focus on the meaning!

# Iris: visual interpretation

Simplification: forget petal and  
*I. virginica* → 2 variables, 2  
species (**binary classification**  
problem).

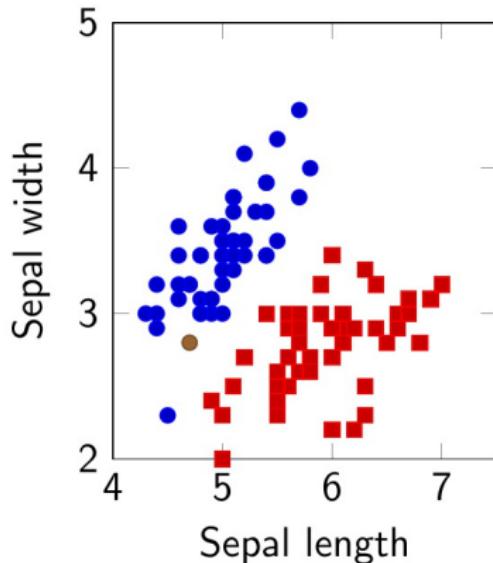


- *I. setosa*
- *I. versicolor*

# Iris: visual interpretation

Simplification: forget petal and I. virginica → 2 variables, 2 species (**binary classification** problem).

- ▶ *Problem:* given any new observation, we want to automatically assign the species.

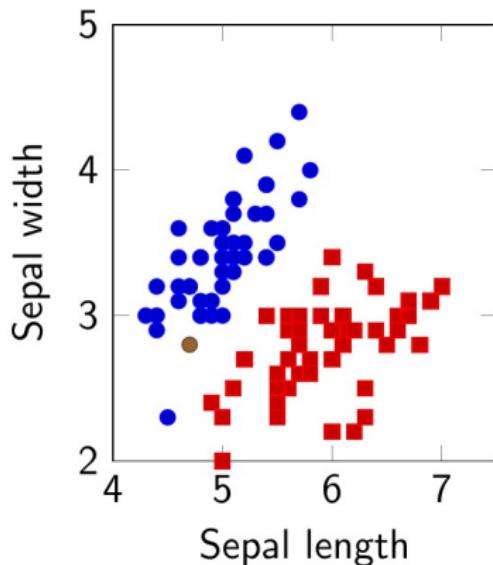


- I. setosa
- I. versicolor

# Iris: visual interpretation

Simplification: forget petal and I. virginica → 2 variables, 2 species (**binary classification** problem).

- ▶ *Problem:* given any new observation, we want to automatically assign the species.
- ▶ *Sketch of a possible solution:*

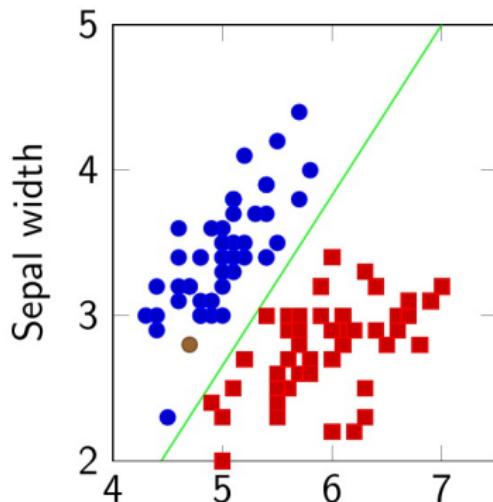


- I. setosa
- I. versicolor

# Iris: visual interpretation

Simplification: forget petal and I. virginica → 2 variables, 2 species (**binary classification** problem).

- ▶ *Problem:* given any new observation, we want to automatically assign the species.
- ▶ *Sketch of a possible solution:*
  1. learn a model (**classifier**)



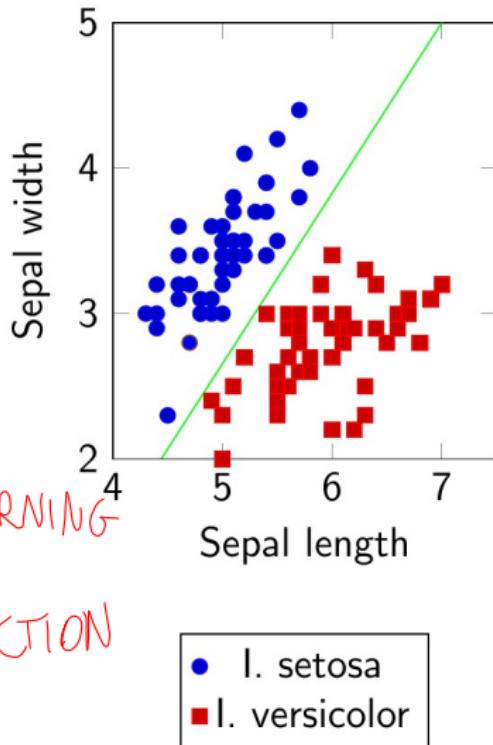
- I. setosa
- I. versicolor

## Iris: visual interpretation

Simplification: forget petal and  
*I. virginica* → 2 variables, 2  
species (**binary classification**  
problem).

- ▶ *Problem:* given any new observation, we want to automatically assign the species.
  - ▶ *Sketch of a possible solution:*

1. learn a model (**classifier**)
  2. “use” model on new observations



# “A” model?

There could be many possible models:

- ▶ how to choose?
- ▶ how to compare?

**Q:** a model of what?

A MODEL OF IRIS FLOWERS

## Choosing the model

The choice of the model/tool/technique to be used is determined by many factors:

- ▶ Problem size ( $n$  and  $p$ )
- ▶ Availability of an output variable ( $y$ )
- ▶ Computational effort (when learning or “using”)
- ▶ Explicability of the model
- ▶ ...

We will see some options.

## Comparing many models

WILL IT WORK IN THE FUTURE?

Experimentally: does the model work well on (new) data?

# Comparing many models

Experimentally: does the model work well on (new) data?

Define “works well”:

- ▶ a single performance index?
- ▶ how to measure?
- ▶ repeatability/reproducibility...

▶ **Q:** what's the difference?

We will see/discuss some options.

→ ONE AND REPEAT  
THE EXP PROCEDURE

→ AND OBTAINS THE  
SAME RESULTS

# It does not work well...

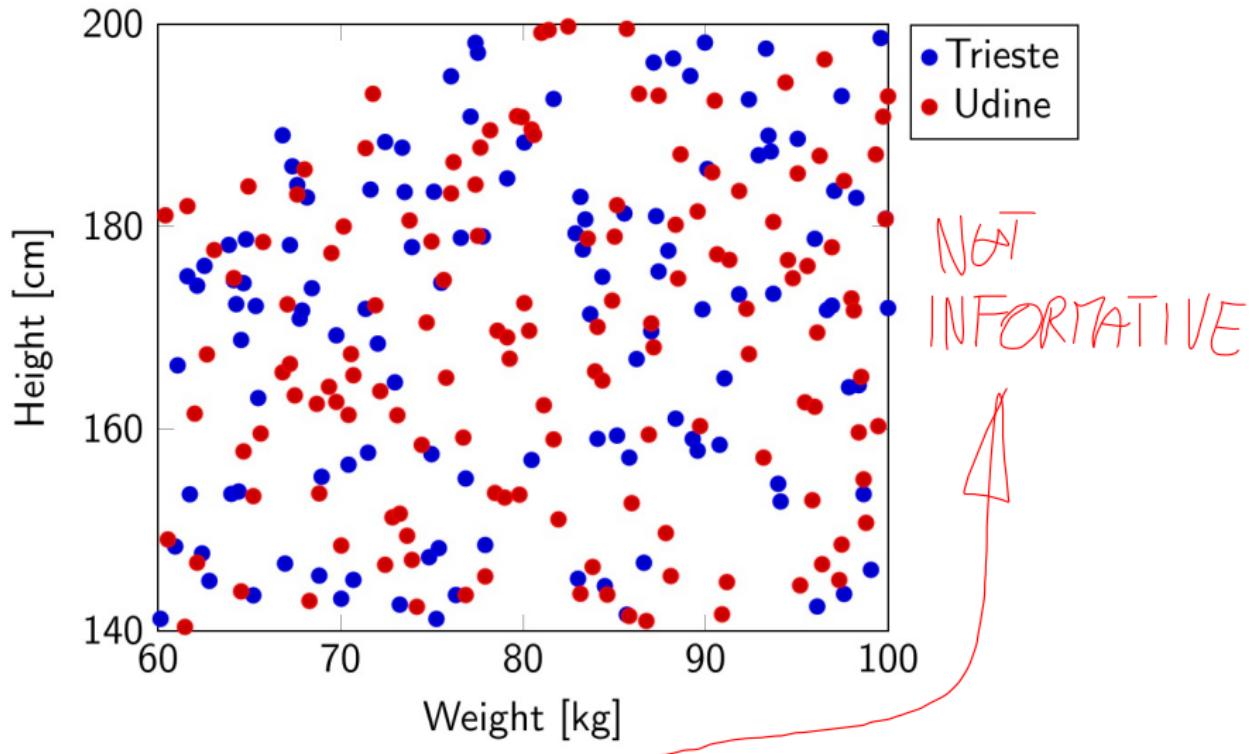
Why?

- ▶ the data is not informative
- ▶ the data is not representative
- ▶ the data has changed
- ▶ the data is too noisy

We will see/discuss these issues.

# ML is not magic

*Problem:* find birth town from height/weight.



**Q:** which is the data issue here?

# Implementation

When “solving” a problem, we usually need:

- ▶ explore/visualize data
- ▶ apply one or more ML technique
- ▶ assess learned models

“By hands?” No, with software!

# ML/DM software

Many options:

- ▶ libraries for general purpose languages:
  - ▶ Java: e.g., <http://haifengl.github.io/smile/>
  - ▶ Python: e.g., <http://scikit-learn.org/stable/>
  - ▶ ...
- ▶ specialized sw environments:
  - ▶ Octave: [https://en.wikipedia.org/wiki/GNU\\_Octave](https://en.wikipedia.org/wiki/GNU_Octave)
  - ▶ R: [https://en.wikipedia.org/wiki/R\\_\(programming\\_language\)](https://en.wikipedia.org/wiki/R_(programming_language))
- ▶ from scratch

# ML/DM software: which one?

- ▶ production/prototype
- ▶ platform constraints
- ▶ degree of (data) customization
- ▶ documentation availability/community size
- ▶ ...
- ▶ previous knowledge/skills

# ML/DM software: why?

In all cases, sw allows to be more productive and concise.  
E.g., learn and use a model for classification, in Java+Smile:

```
1 double[][] instances = ...;
2 int[] labels = ...;
3 RandomForest classifier = (new RandomForest.Trainer()).train(
    instances, labels);
4 double[] newInstance = ...;
5 int newLabel = classifier.predict(newInstance);
```

In R:

```
1 d = ...
2 classifier = randomForest(label~, d)
3 newD = ...
4 newLabels = predict(classifier, newD)
```

We will work with R.

## Section 3

### Fundamentals of R

SKIP

# R software

- ▶ R
  - ▶ a programming language
  - ▶ a software environment
- ▶ RStudio
  - ▶ an IDE built on R

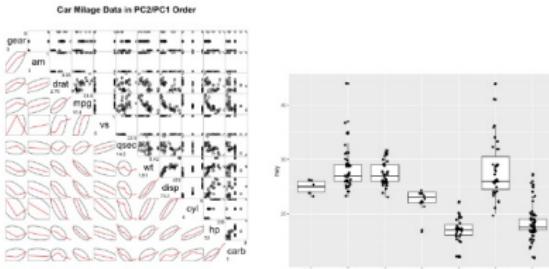
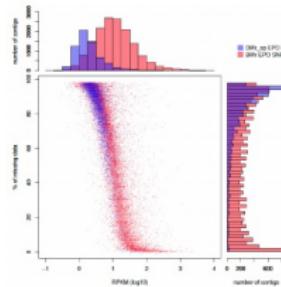
## Section 4

### Plotting data: an overview

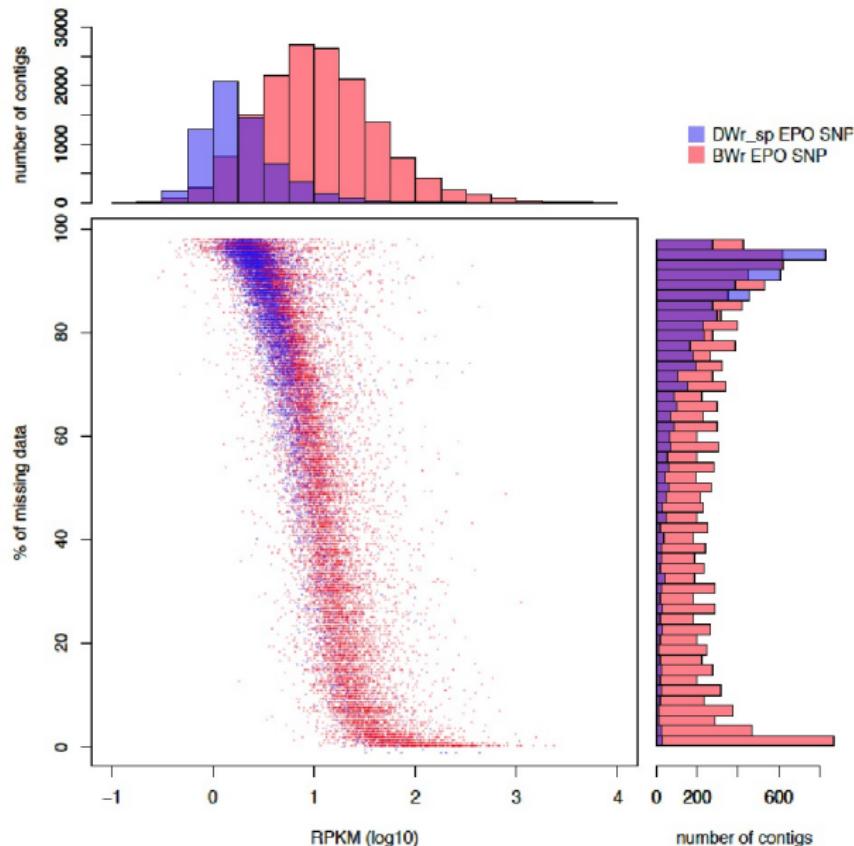
# Advanced plotting

- ▶ many packages (e.g., ggplot2)
- ▶ many options

Which is the most proper chart to support a thesis?



## Aim of a plot: examples



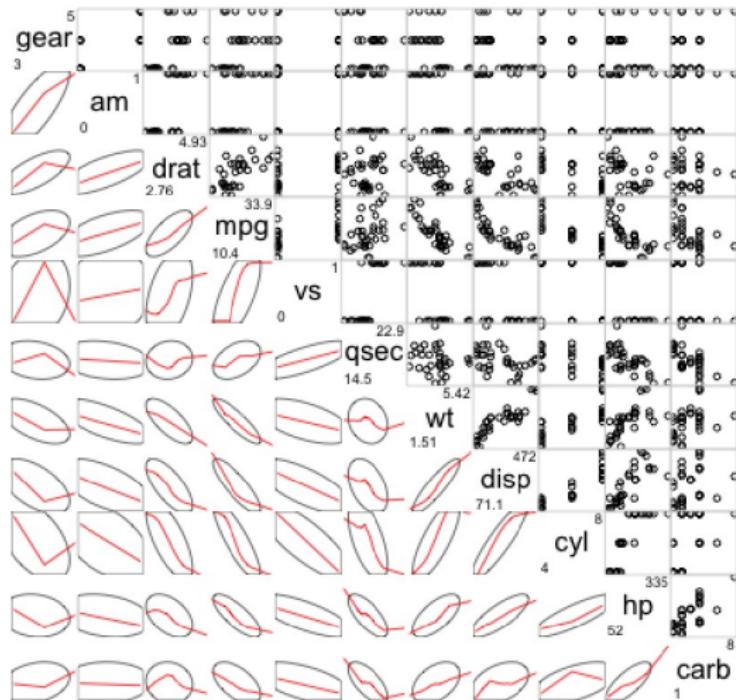
# Aim of a plot: examples

bringing changing highest  
delayed general maximum really  
keep figures building extent currency requirements  
improve must boost almost  
legislation paid  
workers entrance  
determine interview  
shares quoted market remains reports long-term  
normal view instead paper western position  
option executive asked  
among daily target united close interest development fiscal concern  
allow adding needed debt basis central  
become used firm still japan value industry cuts include statement official party effective  
commodity company sales export built  
warned many analysts period financial third system provide production opposed  
authorities planned adjust private february increasing program large noted  
people remained board price months exchange corp issues meet  
remained forced according sold even group week foreign national set through might losses equity claim  
helped allow even group week foreign national set through might losses equity claim  
denied followed enough open public president april billion credit news paid began weekly  
related point local trade rose total near five next seven industrial activity morning  
move difficult move future month first dtrs time market prices value reduce declining  
ruling country recently chief plans part back price outside  
skin pressure largest continue made terms offer union american money  
moves difficult cost expected data looking west january part back price outside  
prediction will come report world world years  
area unit come report world years  
affect economic markets world years  
in richard reuters spokesman added late march since  
wednesday force japanese spokesman added late march since  
holding good give times higher following previous court  
introduced july within agreement long major eight departments  
introduced july within agreement long major eight departments  
britain within less well sell department  
size amount raise business around trading announced full lower prospects  
belgium commercial government final issue compared to chairman  
august rates international annual likely issue  
falling ministry gunnars already increased given  
every repudiation investment action early  
keeping comment response association range past buying political  
volume want analyst companies country give longer opening  
slow makes propose days start conditions normally  
whether expansion whether problems increases entire industry  
import committee immediate local needs support december  
agriculture saying tomorrow actions sharp commerce needs  
measures leave previously believe factors coming  
exchanges reduction manager requirement  
competitiveness transactions chance  
meanwhile

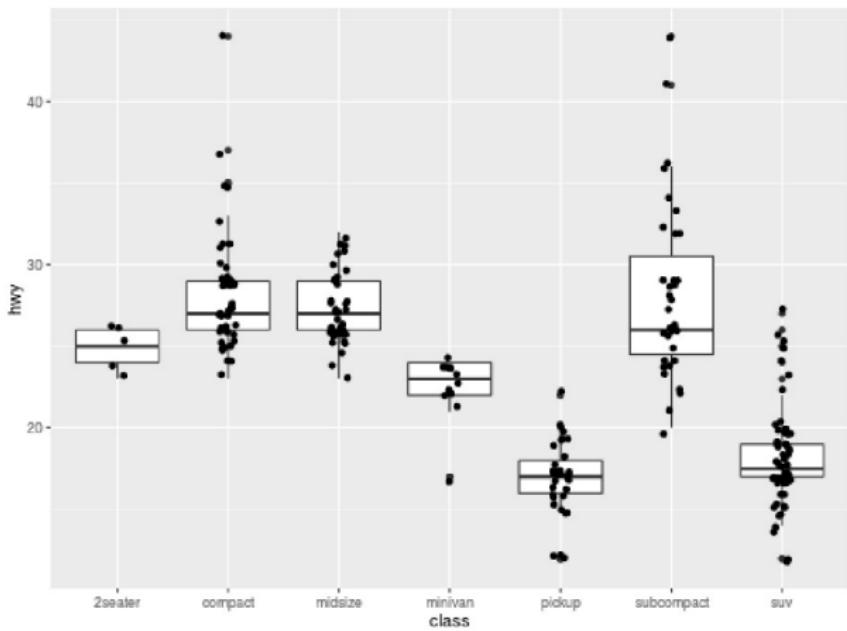
**reuter year**

## Aim of a plot: examples

## Car Milage Data in PC2/PC1 Order



## Aim of a plot: examples



# Lab: let's know iris (1 h) HERE SPECIES IS THE CLASS

JUST WRITE IRIS IN R  
(OR d=iris)

1. get iris data
2. know basic info about it ([summary](#))
3. plot iris and play with it

QUESTIONS:

- m? p?
- VALUES OF ATTRIBUTES?
- APPARENT RELATIONS?
- RISK OF NON INFORM. DATA?

Hints:

```
dr = iris %>% group_by(Species) %>% summarise(Avg.Sepal.Ratio=mean(Sepal.Length/Sepal.Width), Avg.Petal.Ratio=mean(Petal.Length/Petal.Width))
```

```
dr %>% gather(ratio, value, -Species)
```

Packages: ggplot2, dplyr, tidyverse

CONSIDER CLASS. OF SPECIES

IS VAR A "SIGNIFIC." DIFFERENT IN SETOSA  
WRT VERSICOLOR?