Machine Learning and Data Analytics

Eric Medvet Matilde Trevisani

A.A. 2018/2019

Section 1

General information

Lecturers

- Matilde Trevisani
  - $\approx$ 24 h (12 CFU only) + $\approx$ 24 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 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`
- 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 $\rightarrow$ lab activities

- $\approx$ 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 analyzie in depth at least one solution
- form small (2–4) groups
  - possibly with different background
  - peer-tutoring

Exam

Either:

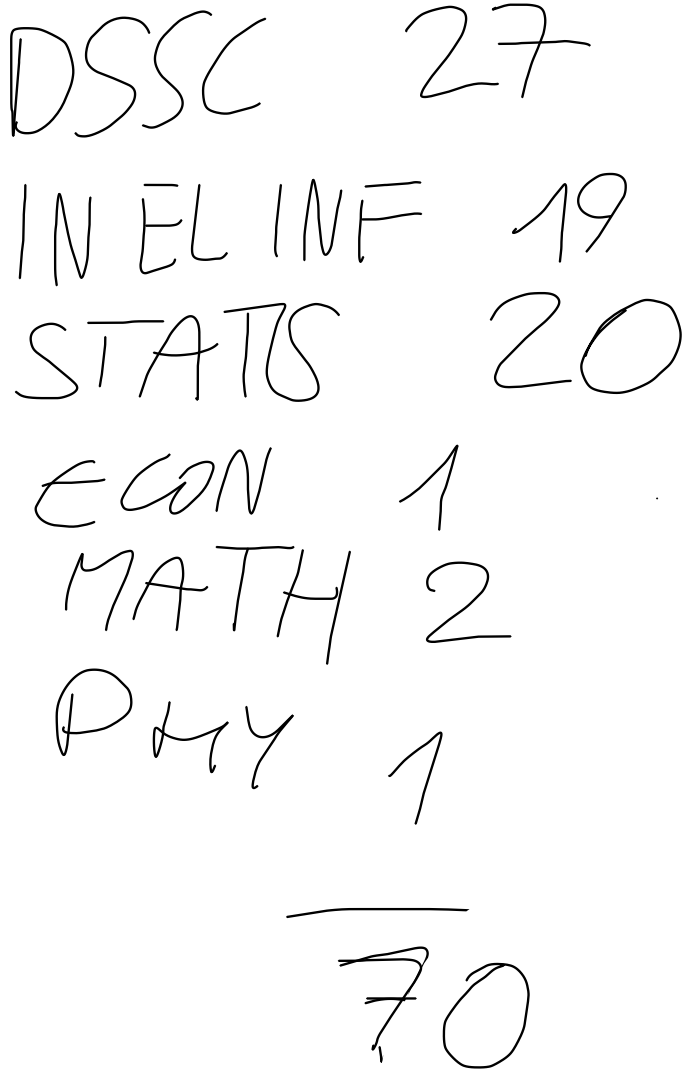
- a student project **and** a written test
- a larger written test

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`)

You?

Who are you?



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 $\langle$input, output$\rangle$ pairs are available
- learn relationships and structures in data (*unsupervised learning*)

Machine Learning everyday

Example problem: spam

Discriminate between spam and non-spam emails.

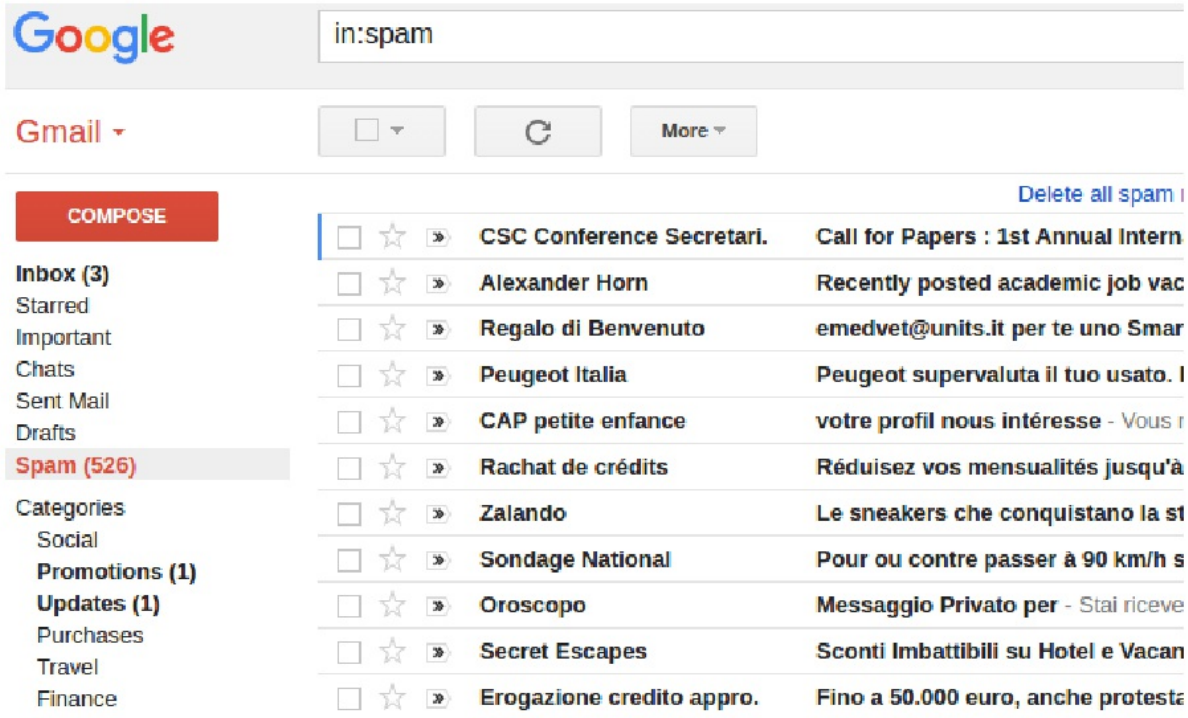


Figure: Spam filtering in Gmail.

Machine Learning everyday

Example problem: flight trajectories

Do flights over the same pair $\langle$origin, destination$\rangle$ follow the "same" trajectory? Why?

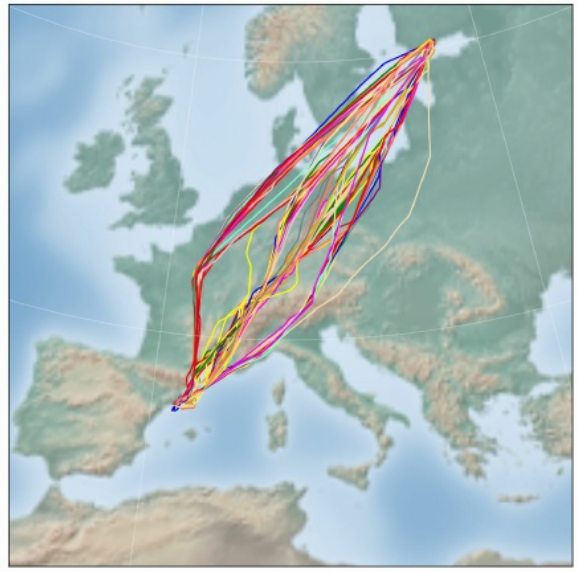


Figure: Clustering of flight trajectories.

Machine Learning everyday

Example problem: image understanding

Recognize objects in images.

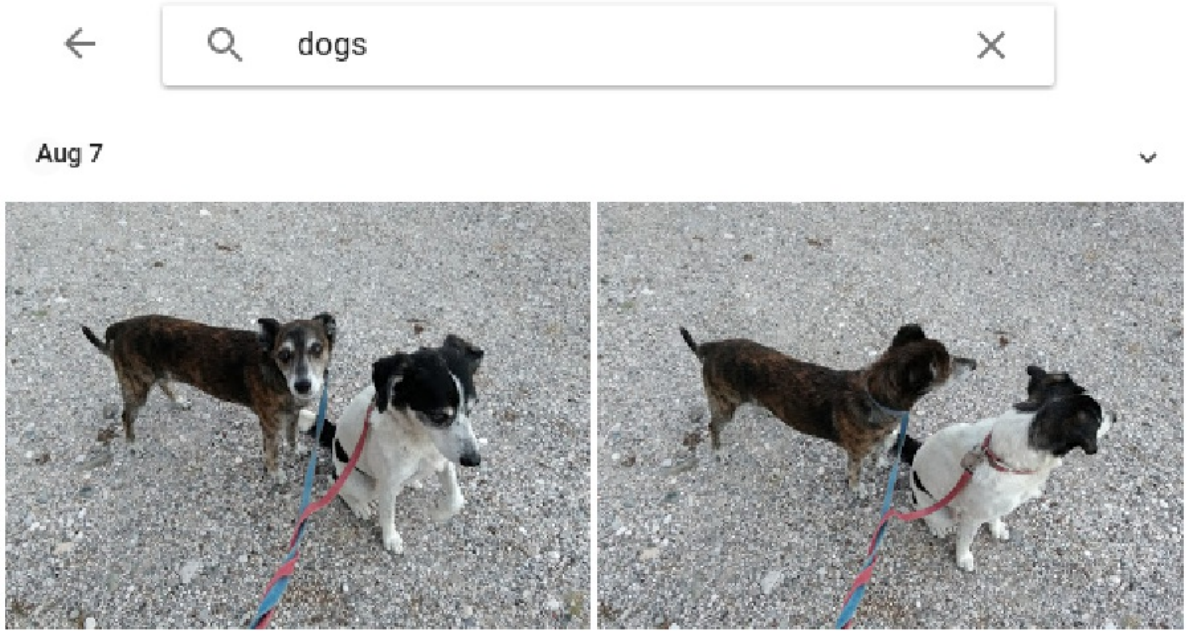


Figure: Object recognition in Google Photos.

Machine Learning everyday

Example problem: authoring regular expressions

Write a regular expression from matching examples.

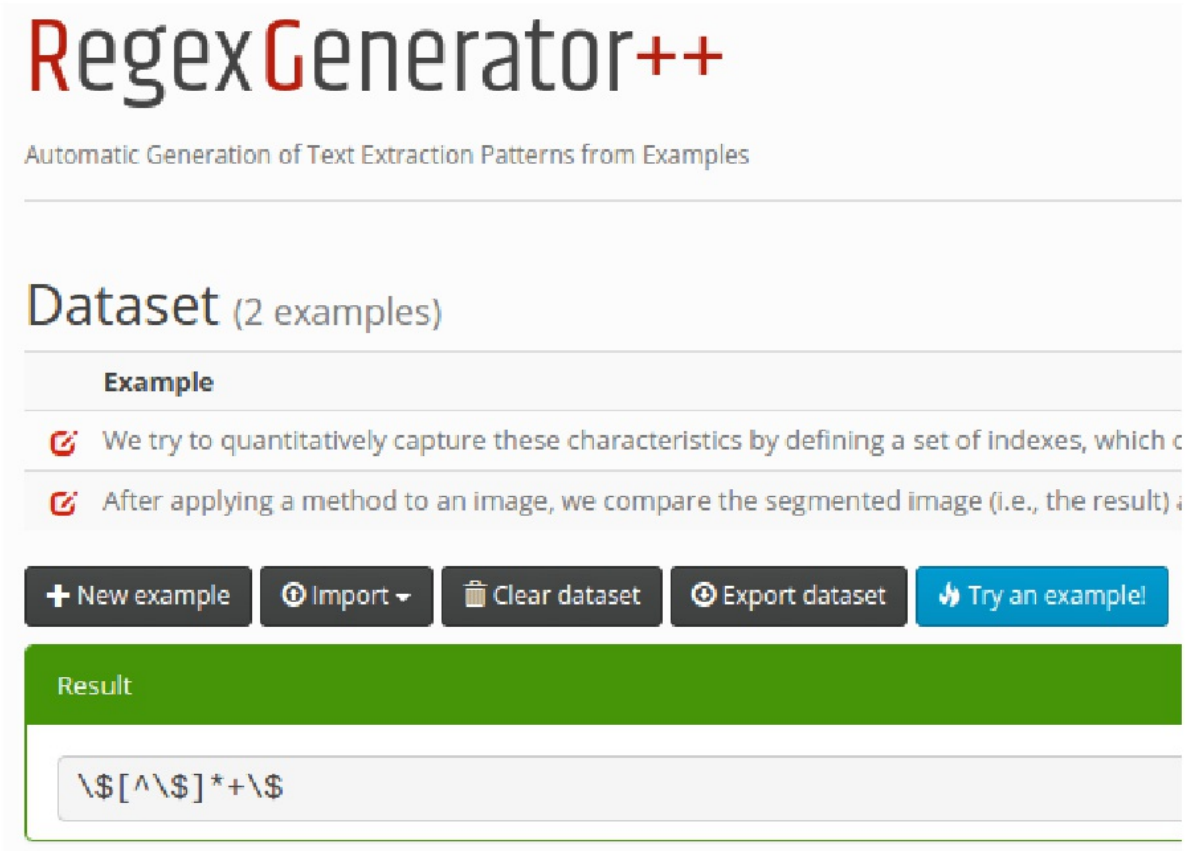


Figure: Regex generation with `http://regex.inginf.units.it/`.

Machine Learning everyday

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

- spam PROBABLY 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

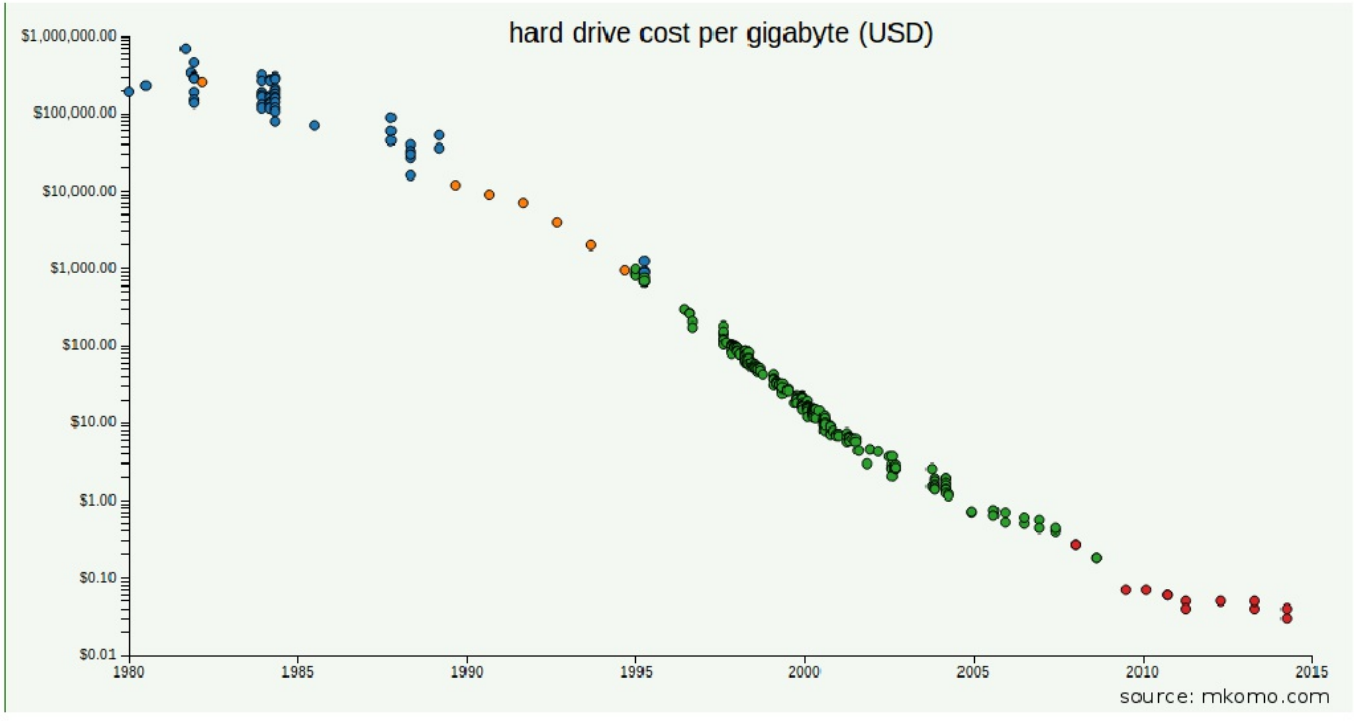


Figure: From `http://www.mkomo.com/cost-per-gigabyte-update`.

ML/DM is popular!

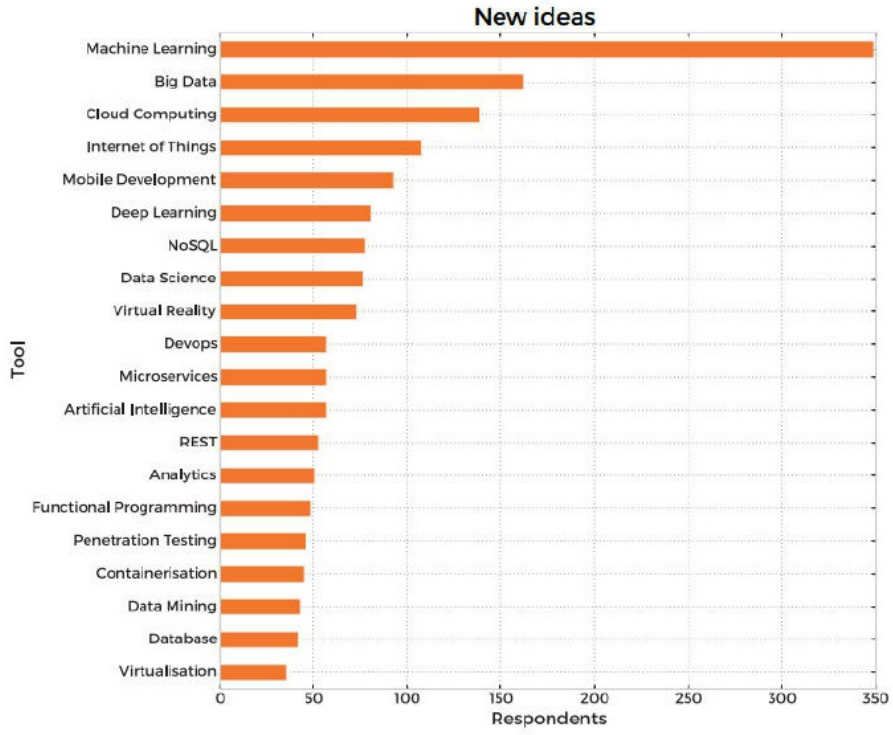


Figure: Popular areas of interest, from the Skill Up 2016: Developer Skills Report[2]

[1] https://techcus.com/p/r1zSmbXut/top-5-highest-paying-programming-languages-of-2016/.

[2] 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 specie 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 specie 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 specie 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 specie 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 specie 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 specie 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

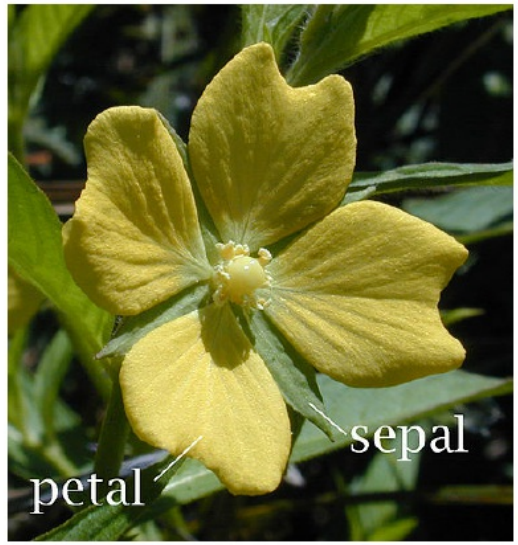


Figure: Sepal and petal.

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. setosa

Other information

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

A FLOWER →
B FLOWER →

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	I. setosa
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

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

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

Notation and terminology

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

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

Notation and terminology

$n \ll p$

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

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

Notation and terminology

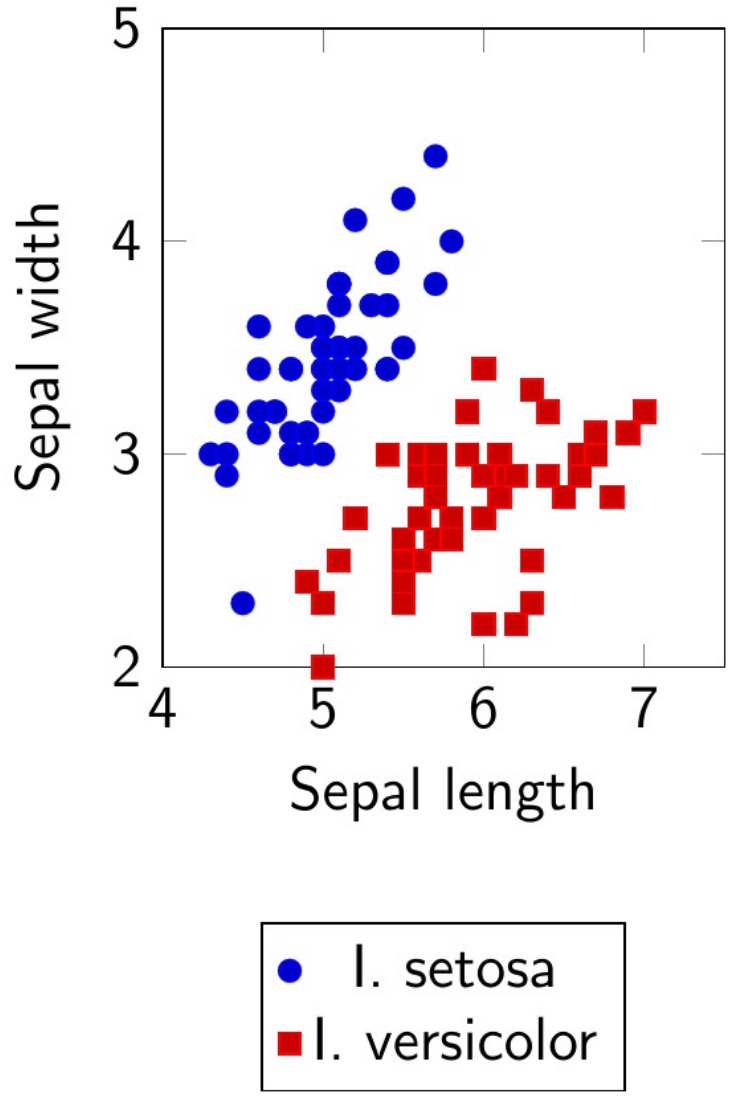
Different communities (e.g., statistical learning vs. machine learning vs. artificial intelligence) use different terms and notation:

- $x_j^{(i)}$ 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

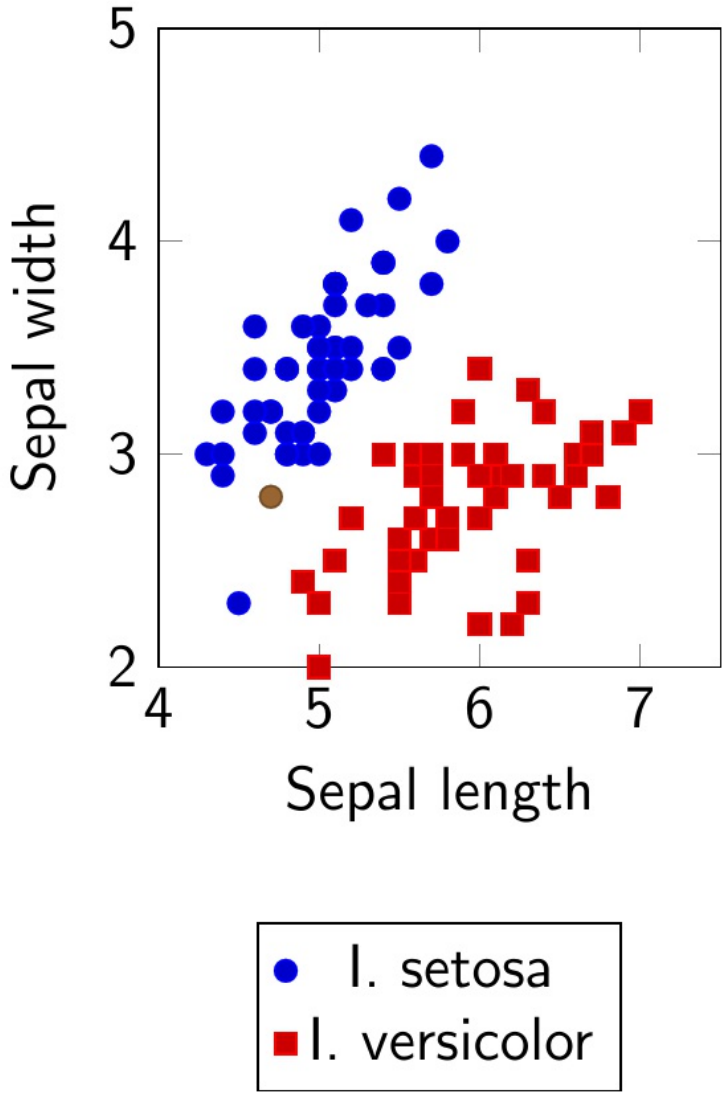
Simplification: forget petal and I. virginica $\rightarrow$ 2 variables, 2 species (binary classification problem).



Iris: visual interpretation

Simplification: forget petal and I. virginica $\rightarrow$ 2 variables, 2 species (binary classification problem).

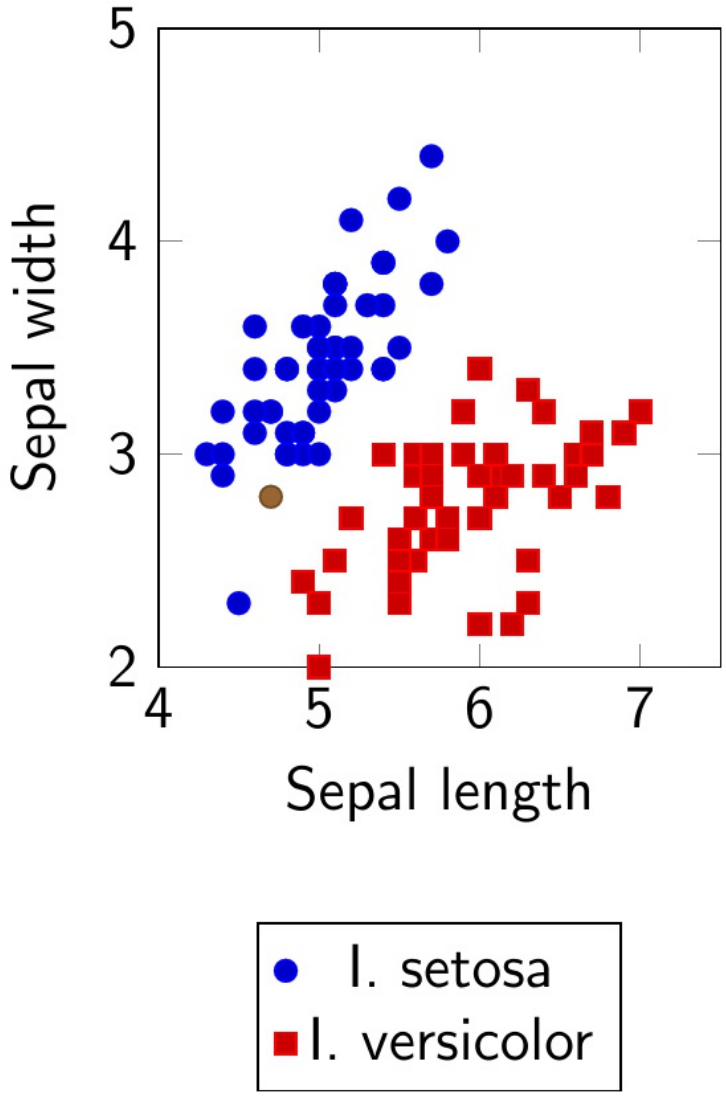
- *Problem*: given any new observation, we want to automatically assign the species.



Iris: visual interpretation

Simplification: forget petal and I. virginica $\rightarrow$ 2 variables, 2 species (binary classification problem).

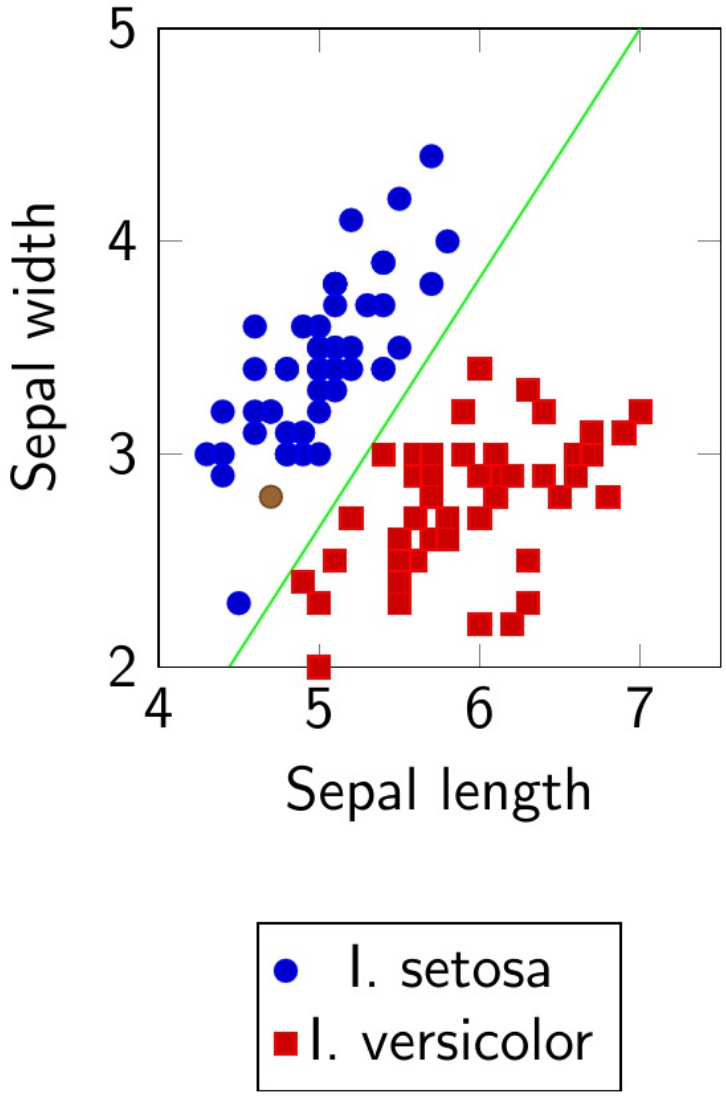
- *Problem*: given any new observation, we want to automatically assign the species.
- *Sketch of a possible solution*:



Iris: visual interpretation

Simplification: forget petal and I. virginica $\rightarrow$ 2 variables, 2 species (binary classification problem).

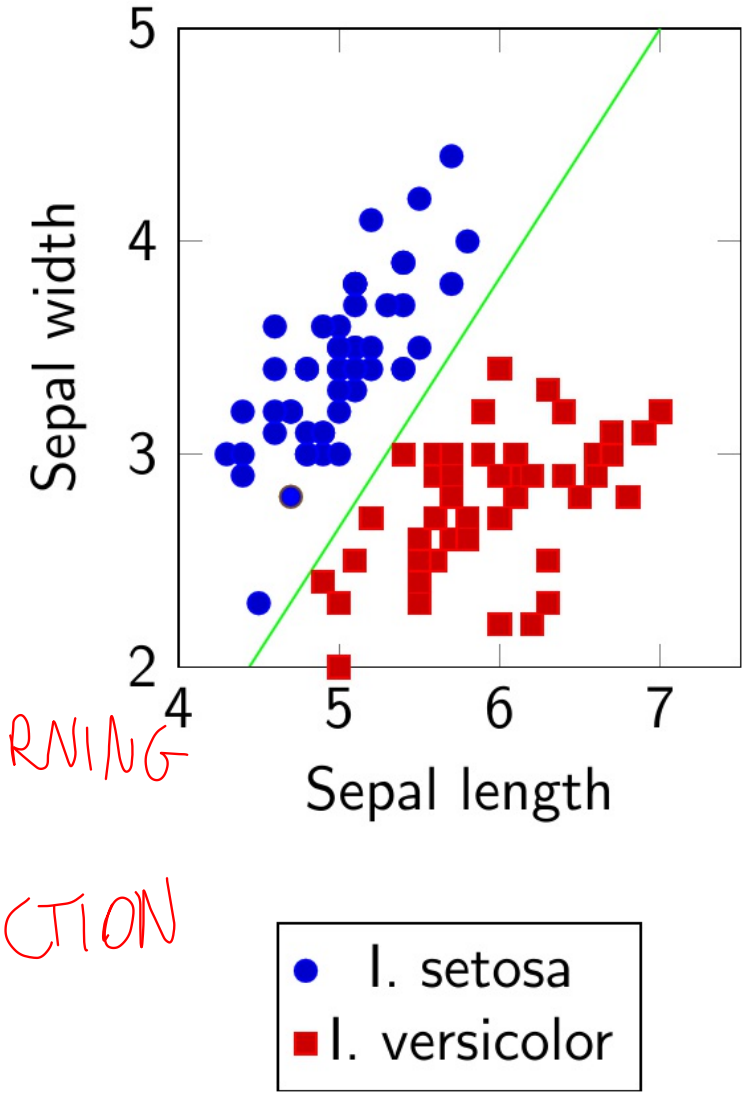
- *Problem*: given any new observation, we want to automatically assign the species.
- *Sketch of a possible solution*:
  1. learn a model (classifier)



Iris: visual interpretation

Simplification: forget petal and I. virginica $\rightarrow$ 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?

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 CAN REPEAT THE EXP PROCEDURE

AND ORTAINS THE SAME RESULTS

It does not work well. . .

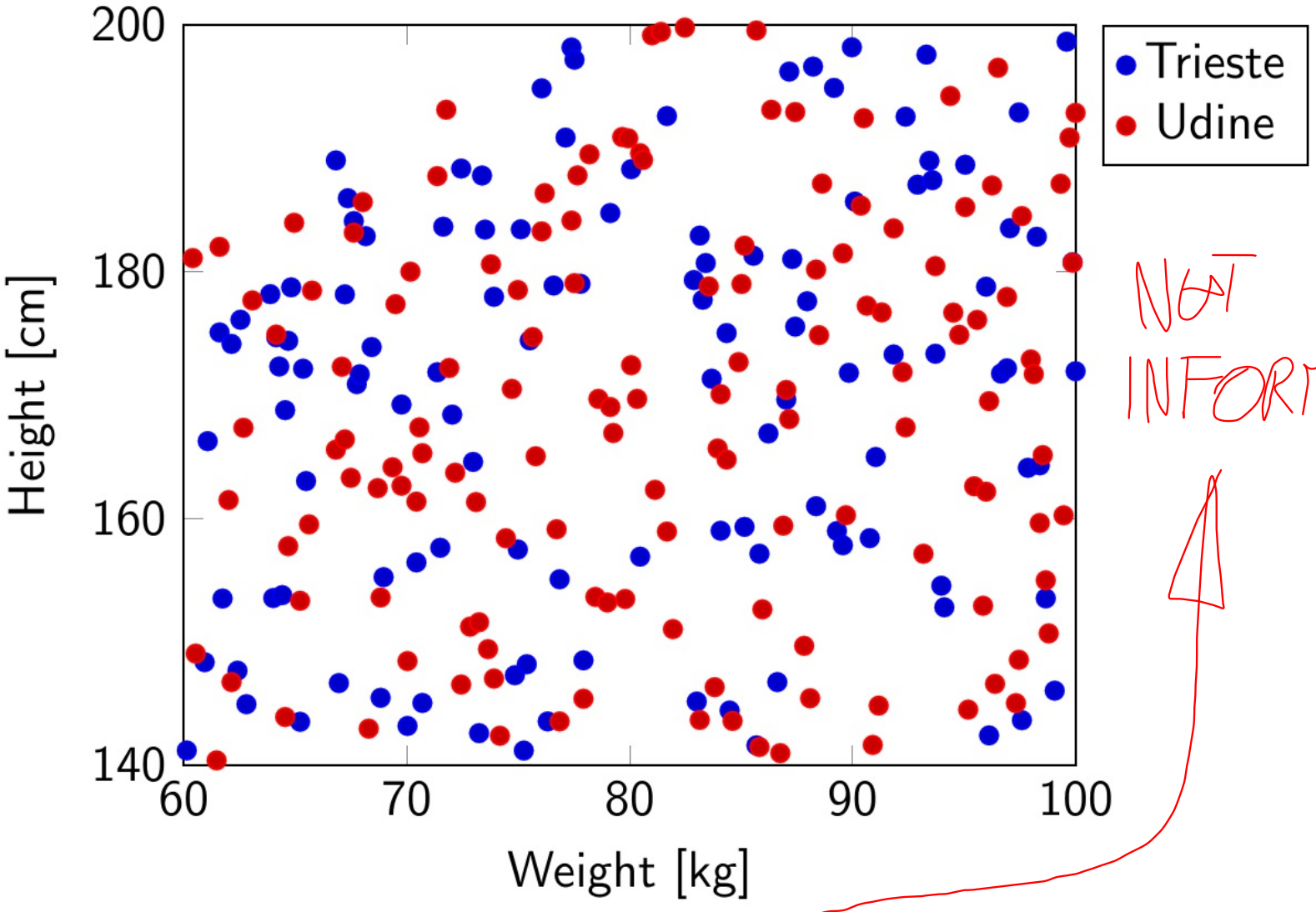
Why?

- 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`
  - R: `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

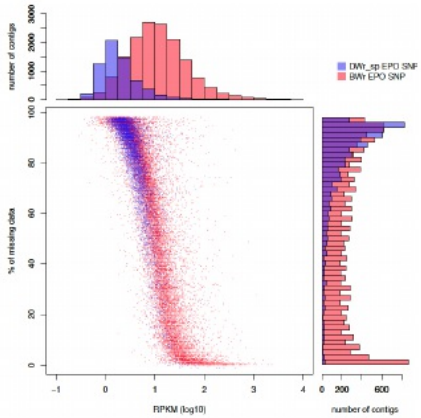
Section 4

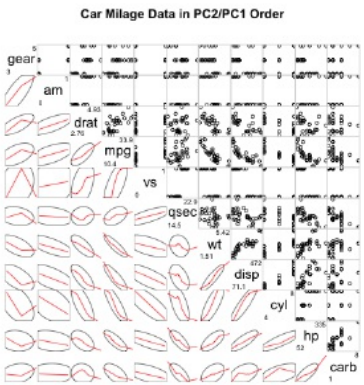
Plotting data: an overview

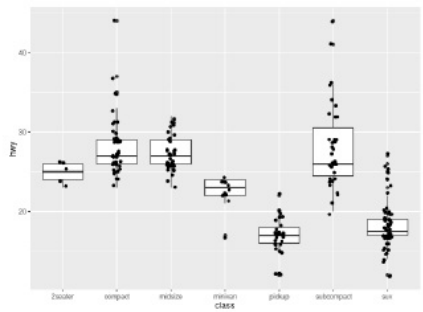
Advanced plotting

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

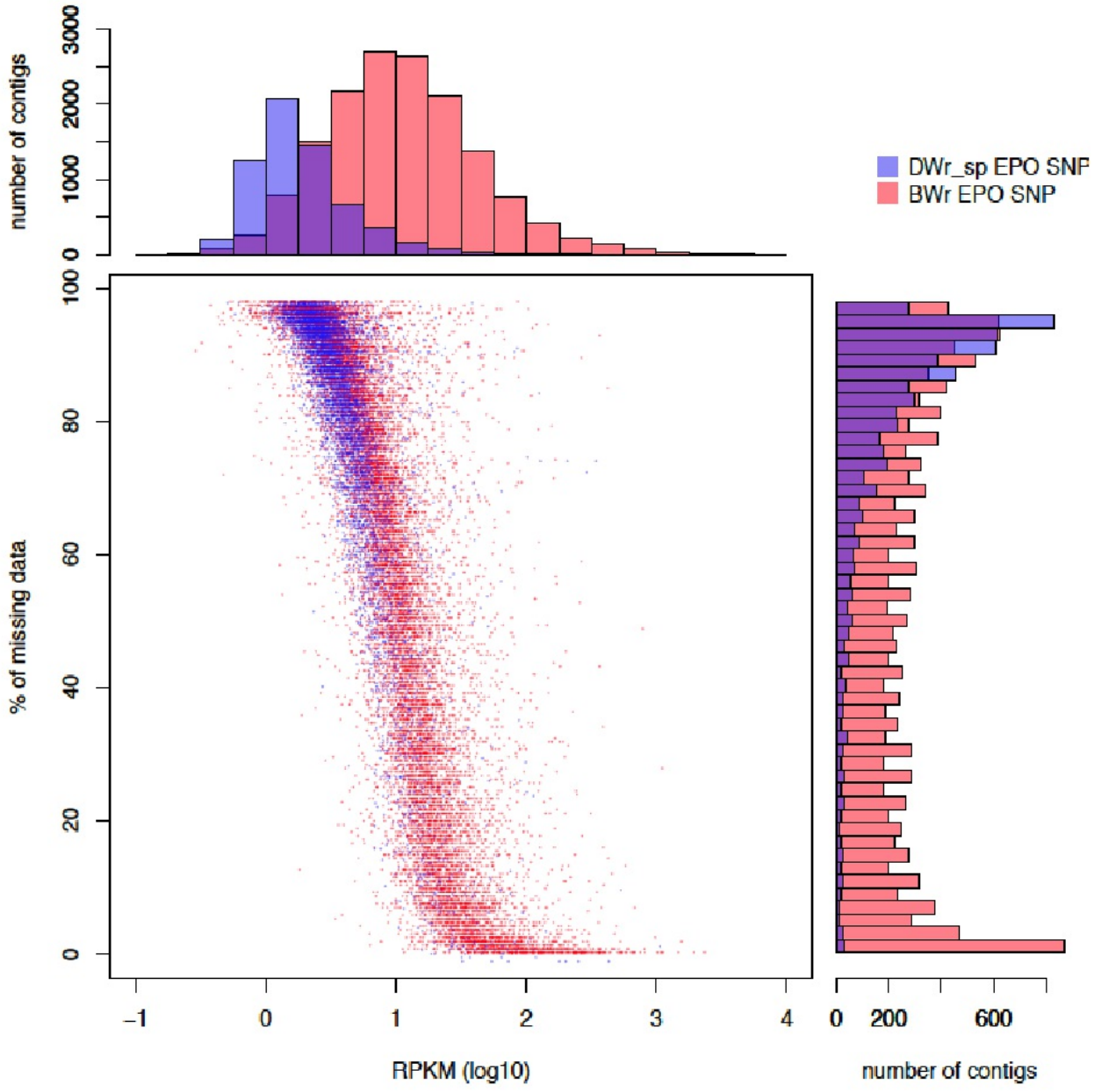
Which is the most proper chart to support a thesis?



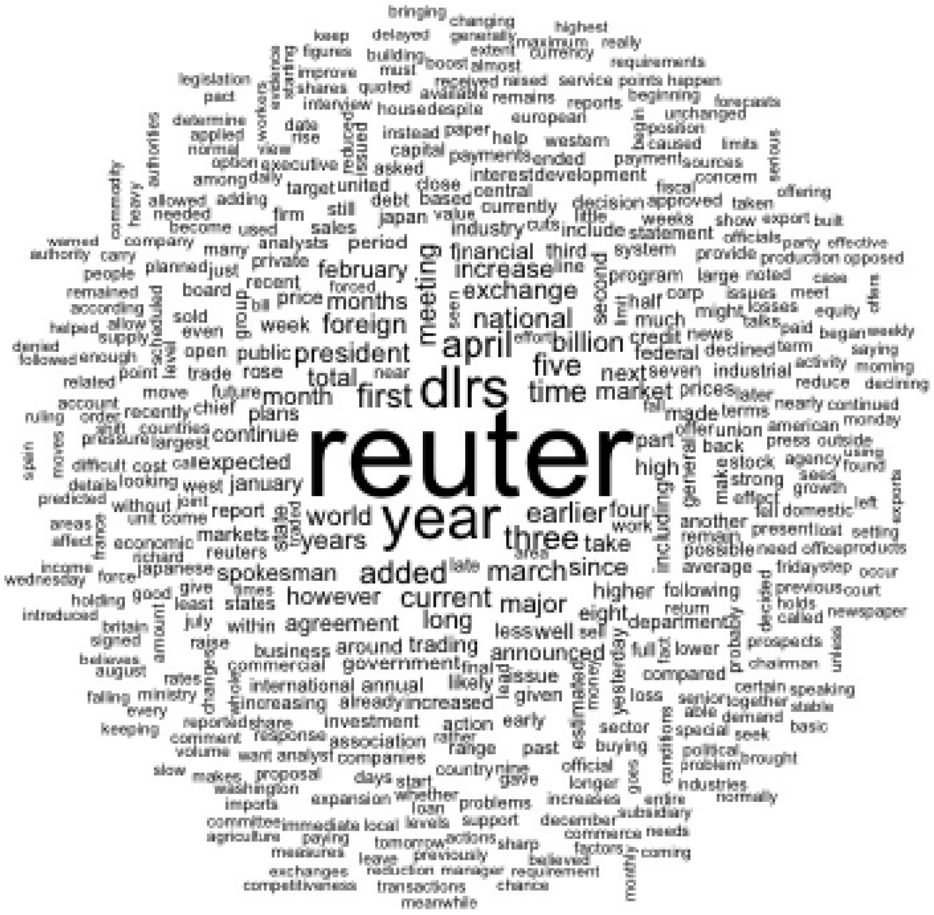




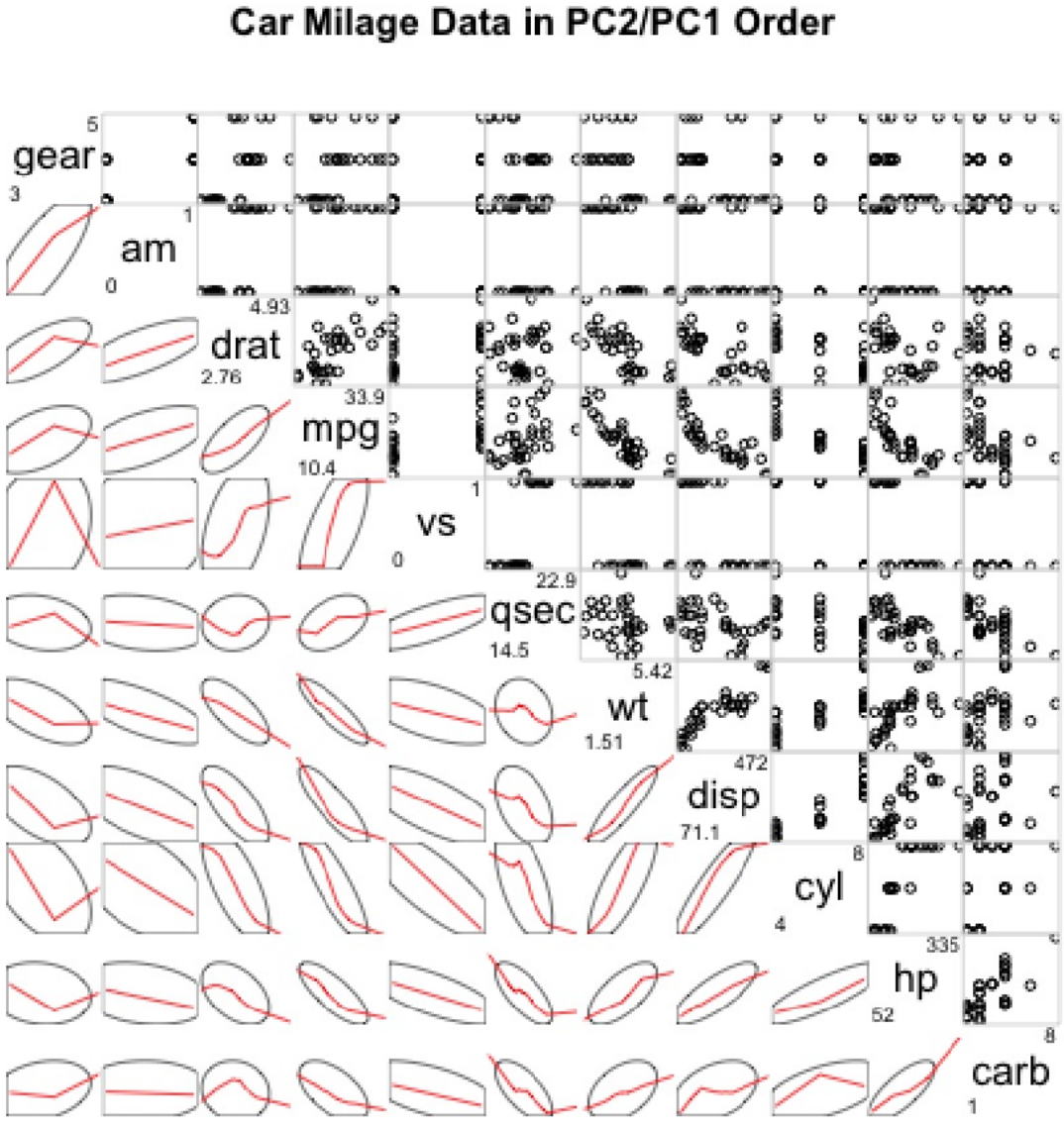
Aim of a plot: examples



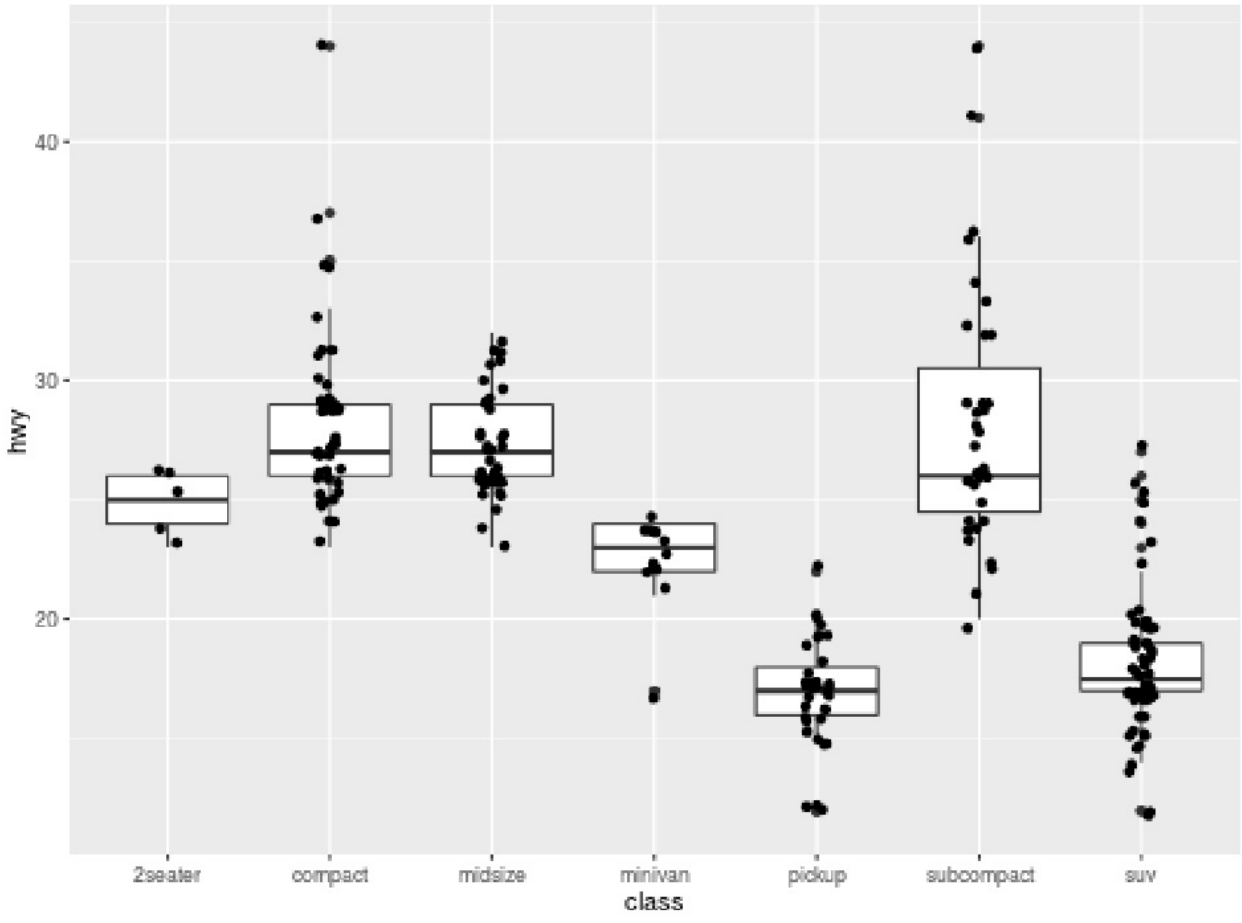
Aim of a plot: examples



Aim of a plot: examples



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:
- n? 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`, `tidyr`

CONSIDER CLASS. OF SPECIES

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