What can we do with data?

João Mendes Moreira

LIAAD INESC TEC, University of Porto, Portugal

Definition of Analytics

The science that analyze crude data to extract useful knowledge (patterns) from them.

A bit of history

Statistics -> Inductive learning

Reproducing the human behaviour: artificial intelligence

Learning from databases

Machine Learning

Data Mining

Big Data and Data Science

Big Data:

- Volume: How to store large amounts of data whose structure is not known in advance?
- Velocity: How to guarantee that the extracted information uses the data as long the data arrives?
- Variety: how to use together information arriving in different moments, different granularities, different sources?

Big Data and Data Science

Data Science:

- Data science is concerned with the creation of models able to extract patterns from complex data and the use of this models in real life problems.
- Data science extracts meaningful and useful knowledge from data, with the support of suitable technologies. It has a strong relation to analytics and data mining. Data science goes beyond data mining by providing a knowledge extraction framework that also includes statistics and visualization.

Big Data and Data Science

Big Data: gives support to data collection and management;

Data science: applies techniques to these data to discover new and useful knowledge.

Other terms such as **Knowledge discovery** or **knowledge extraction**, **pattern recognition, data analysis, data engineering**, and several others are also used.

The definition we use of data analytics crosses all these areas that are able to extract knowledge from data.

Big Data and architectures

- Distributed systems
- One of the first developed techniques for big data processing using clusters of computers was **MapReduce**.
- MapReduce is a programming model that has two steps: map and reduce.
- MapReduce deals with the previous requirements by dividing the data set into parts, named chunks, and store in each computer in the cluster the chunk of the data set needed by this computer to accomplish its task.
- The most famous implementation of MapReduce is called Haddop.

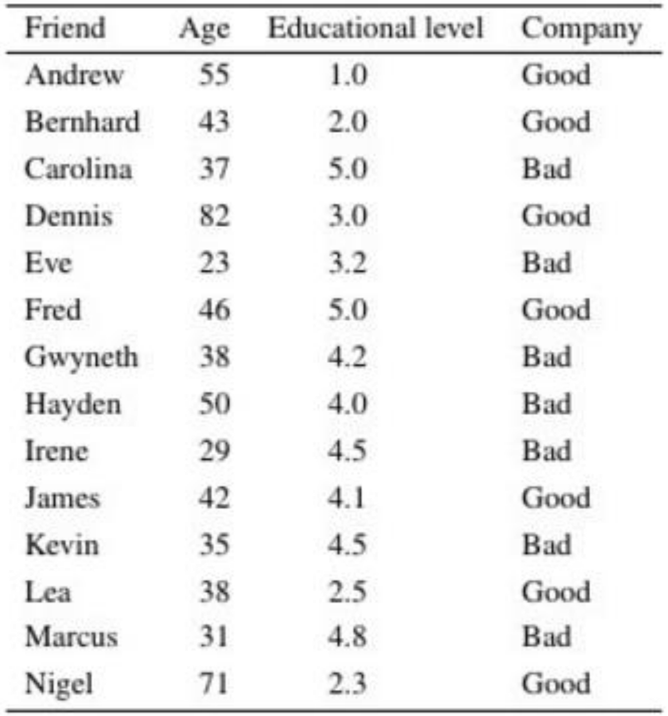
Big Data and architectures

- The distributed system must also:
  - Make sure that no chunk of data is lost and the whole task is concluded. If one or more nodes have a failure, its task, and the corresponding data chunk, is assumed by other computer in the cluster.
  - Repeat the same task, and corresponding data chunk, in more than one cluster computers, which is called redundancy. Thus, if one or more computer fails, the redundant computer carry on with the task and the whole computer system
  - Computers with problems can return to the cluster again when they are fixed.
  - New computers can be easily be relieved from the cluster or be included in the cluster as the processing demand changes.

What is data?

- But what is **Data** about? Data per se, in the Information Technology age, are a large set of bits encoding numbers, texts, images, sounds, videos, etc..
- Unless we add information to data, they are meaningless. When we add information, giving a meaning to them, these data becomes knowledge. But before data become knowledge, typically, they pass through several steps where data are still named data, despite being a bit more organized, i.e., with some information associated.

What is data?



Friend	Age	Educational level	Company
Andrew	55	1.0	Good
Bernhard	43	2.0	Good
Carolina	37	5.0	Bad
Dennis	82	3.0	Good
Eve	23	3.2	Bad
Fred	46	5.0	Good
Gwyneth	38	4.2	Bad
Hayden	50	4.0	Bad
Irene	29	4.5	Bad
James	42	4.1	Good
Kevin	35	4.5	Bad
Lea	38	2.5	Good
Marcus	31	4.8	Bad
Nigel	71	2.3	Good

What is data?

- Information as presented in the table, usually named **tabular data**, is characterized by the way data are organized. In tabular data, data are organized in rows and columns where each column represents a characteristic of the data and each row represents an occurrence of the data. According to this organization, a column is named an **attribute** or, with the same meaning, a **feature**, while a row is named an **instance**, or with the same meaning, an **object**.

What is data?

Definition of **Instance** or **Object:**

- Instances, also named objects, are examples of the concept we want to characterize.

Definition of **Attribute** or **Feature:**

- Attributes, also named features, are characteristics present in the instances.

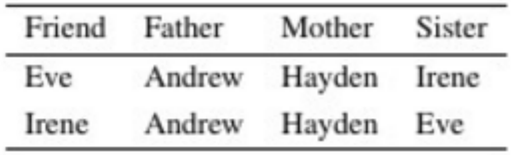
What is data?

The majority of the contents of this course expect the data to be in tabular format, i.e., already organized by rows and columns where each row represents an instance and each column represents an attribute. However, a table can be organized differently, having the instances per column and the attributes per row.

There are, however, data that are not possible to represent in a single table.

What is data?

If some of the **friends are relatives of** other friends, a second table representing the familiar relations between the friends would be necessary.



Friend	Father	Mother	Sister
Eve	Andrew	Hayden	Irene
Irene	Andrew	Hayden	Eve

You should note that each person referred in this table also exists in the previous one, i.e., there are relations between attributes of different tables.

Data sets represented by several tables, making clear the relations between these tables, are named **relational data sets**. This is done easily using relational databases. In this course only simple forms of relational data will be used. This is discussed in each chapter whenever necessary. The majority of the methods imply the use of **tabular data**.

A short taxonomy on data analytics

A natural taxonomy that exists in data analytics is:

Descriptive Analytics: summarize or condensate data to extract patterns;

Predictive Analytics: extract models from data to be used for future predictions.

A short taxonomy on data analytics

In descriptive tasks, the result of a given method or technique is obtained directly by applying an algorithm to the data. The result can be a statistic, such as an average, a plot, a set of groups with similar instances, among other things that we will see along this course.

Definition of **Method** or **Technique**

A method or technique is a systematic procedure that allows to achieve an intended goal.

A short taxonomy on data analytics

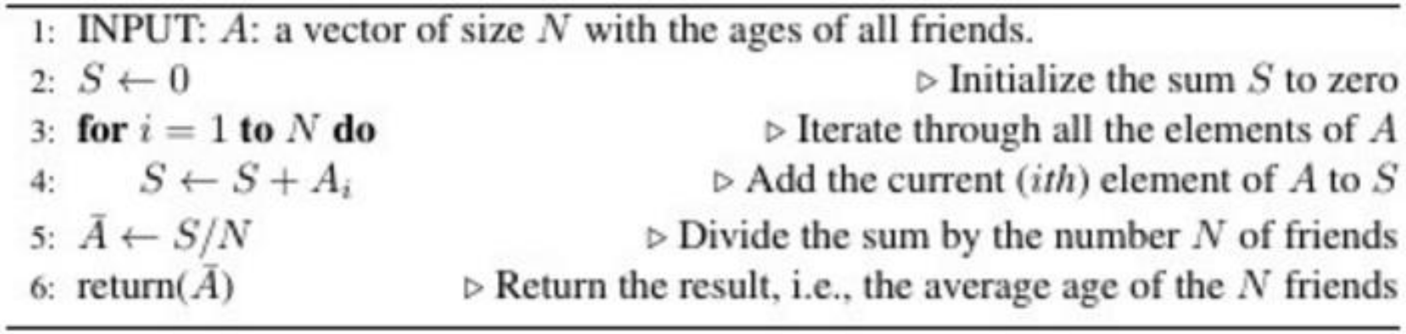
A method shows how to perform a given task. But in order to use a language closer to the language computers can understand, it is necessary to describe the method/technique through an algorithm.

Definition of **Algorithm**

Algorithm is a self-contained step-by-step set of instructions easily understandable by humans, allowing the implementation of a given method. They are self-contained in order to be easily translated to an arbitrary programming language.

A short taxonomy on data analytics

The method to obtain the average age of my friends uses the ages of each friend (we could use other methods, such as using the number of friends for each different age). A possible algorithm for this very simple example is shown next.



```
INPUT: $A$: a vector of size $N$ with the ages of all friends.
$S \leftarrow 0$                          ▷ Initialize the sum $S$ to zero
for $i = 1$ to $N$ do                     ▷ Iterate through all the elements of $A$
    $S \leftarrow S + A_i$                ▷ Add the current ($ith$) element of $A$ to $S$
$\bar{A} \leftarrow S/N$                  ▷ Divide the sum by the number $N$ of friends
return($\bar{A}$)                         ▷ Return the result, i.e., the average age of the $N$ friends
```

In the limit, a method can be straightforward being possible, in many cases, to express it as a formula instead of an algorithm.

$$\bar{A}= \frac{\sum_{i=1}^{N} A_i}{N}$$

A short taxonomy on data analytics

Differently from descriptive methods, the result of applying an algorithm on a predictive method to given data, is typically a model. Let us see what a model is.

Definition of Model

A **model** in data analytics is a generalization obtained from data that can be used afterwards to generate predictions for new given instances. It can be seen as a prototype that can be used to make predictions. Thus, model induction is a predictive task.

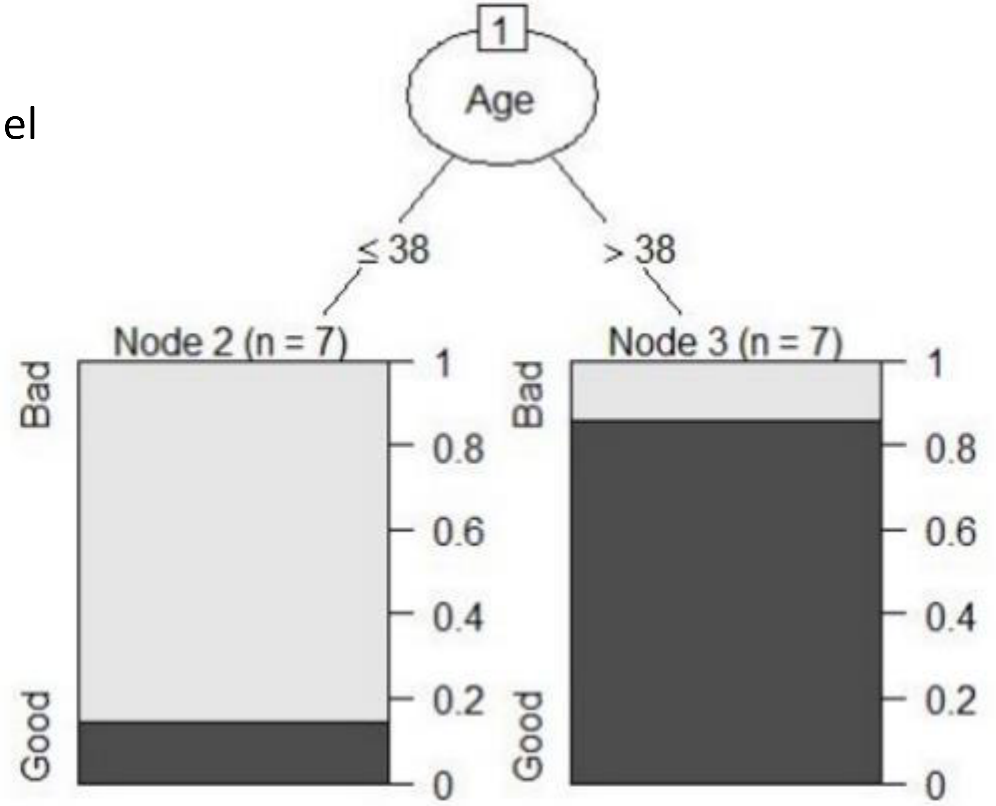
A short taxonomy on data analytics

If we apply an algorithm for induction of decision trees to provide an explanation of who, among our friends, is a good company, we obtain a model, named decision tree, like the one presented in the next figure, where it can be seen that people with more than 38 years are typically better company than those whose age is smaller than or equal to 37.

Indeed, the authors strongly agree that more than 80% of the people with an age smaller than or equal to 38 is bad company, while more than 80% of people older than 38 years is good company. This model could be used to predict whether a new friend is or not a good company. It would be enough to know the age of that new friend.

A short taxonomy on data analytics

A model


1
Age
≤ 38
> 38
Node 2 (n = 7)
Node 3 (n = 7)
Bad
Good
Bad
Good
1
0.8
0.6
0.4
0.2
0
1
0.8
0.6
0.4
0.2
0