Porto Business School
/ University of Porto

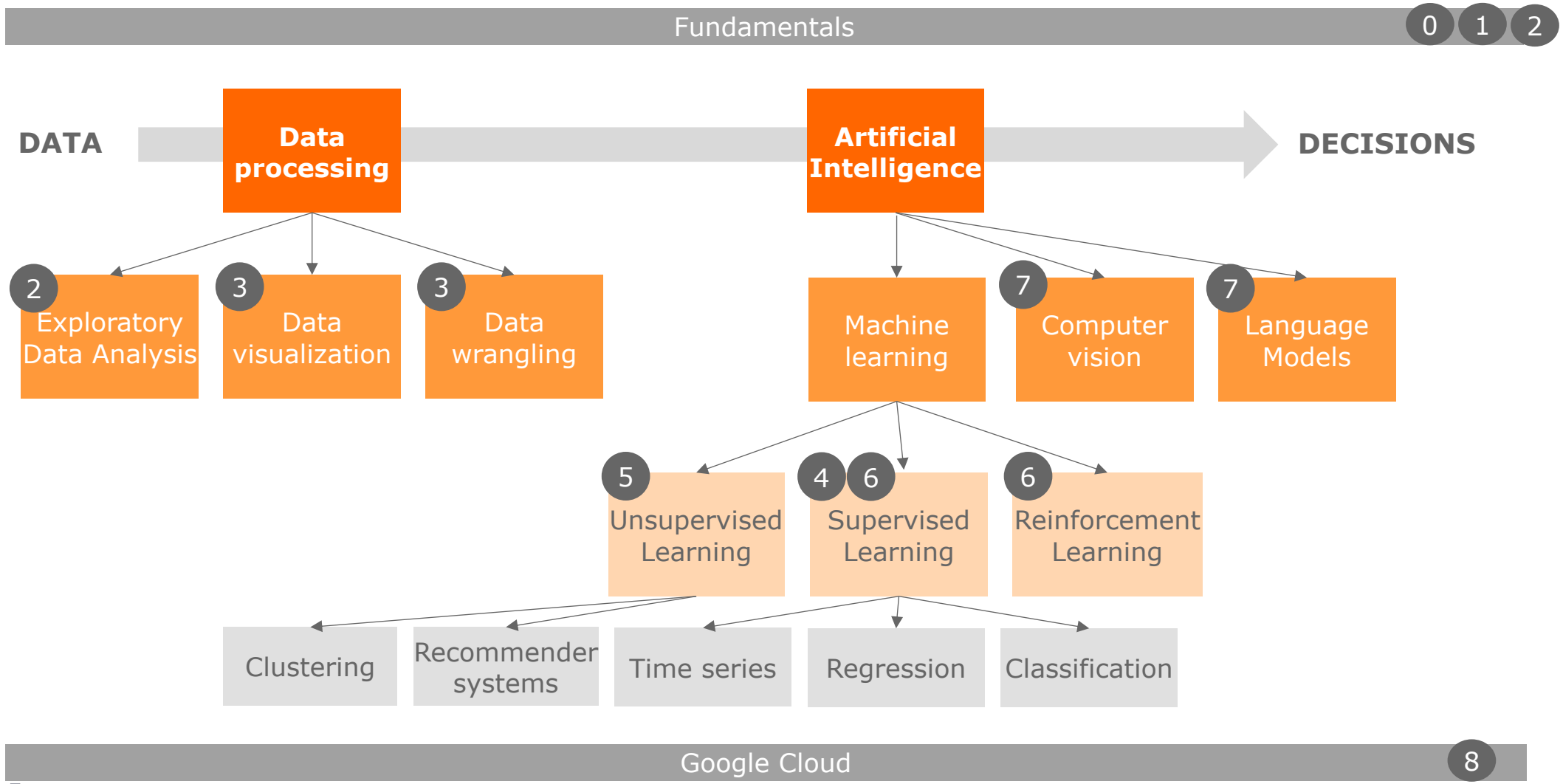
BPI Grupo CaixaBank

LTP

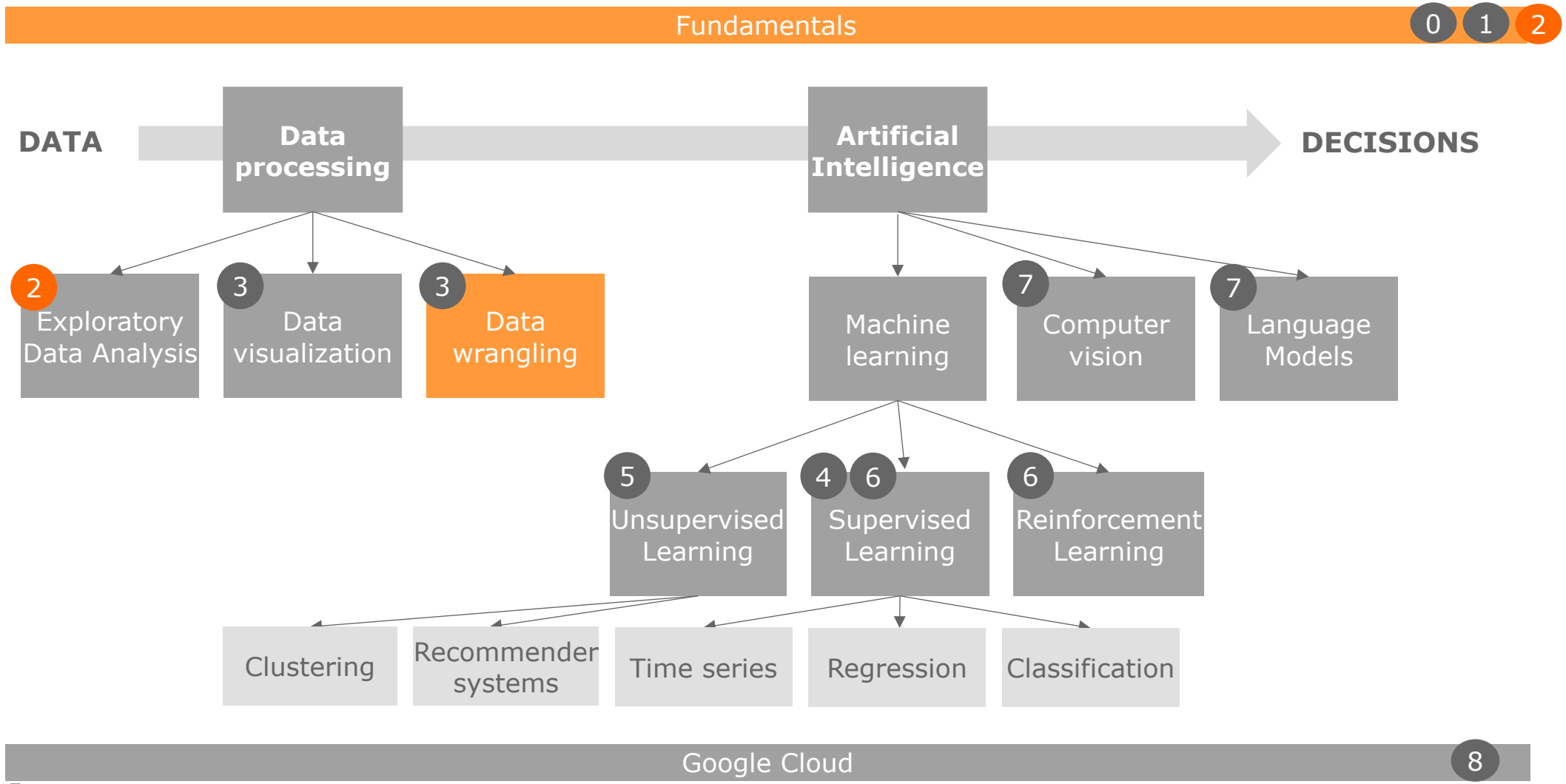
Data Engineering

Mário Amorim Lopes/Fábio Neves Moreira

Academia DTI



Academia DTI



Data Wrangling and Data Visualization

Data Engineering	Data Visualization	EDA in Python	Adv. Topics in Dataviz
- Different types of data (unstructured, semi-structured and structured data) - Data importing - Data cleaning - Data transformation (date parsing, character encodings, etc.)	- Visualizing data in Python using Seaborn - Line charts - Bar charts and heat maps - Scatter plots - Histograms and density plots	- Building a data processing pipeline (import, transform, visualize) - Descriptive statistics of a dataset - Visual outlier detection and correction	- Examples of advanced data visualization - Univariate visualization - Multivariate visualization - Whole dataset visualisations

Data Wrangling and Data Visualization

Session 1

09h30	Introduction and setup
10h00	Data cleaning (tutorial and exercise)
10h30	Parsing dates (tutorial and exercise)
11:00	**Break**
11h15	Character encodings (tutorial and exercise)
12h00	Data importing (tutorial and exercise)
12h45	**Wrap-up**

Introduction - Types of data

Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
 </Student>
....
</University>
```

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Introduction - Typical activities in data cleaning

Wrap-up

- ✓ **Data rarely comes clean and ready-to-use** — that means we need to prepare (clean, transform) it
- ✓ **There is no single way to do it** — that means you should practice a lot and build up your own style
- ✓ **Experience will save you time** — that means that in the future you will be able to anticipate the most common data issues early in the project

Make change happen