Data Mining

S2

NOVA-IMS 2025/2026

Fernando Lucas Bação

bacao@isegi.unl.pt

http://www.isegi.unl.pt/fbacao

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

1

Agenda

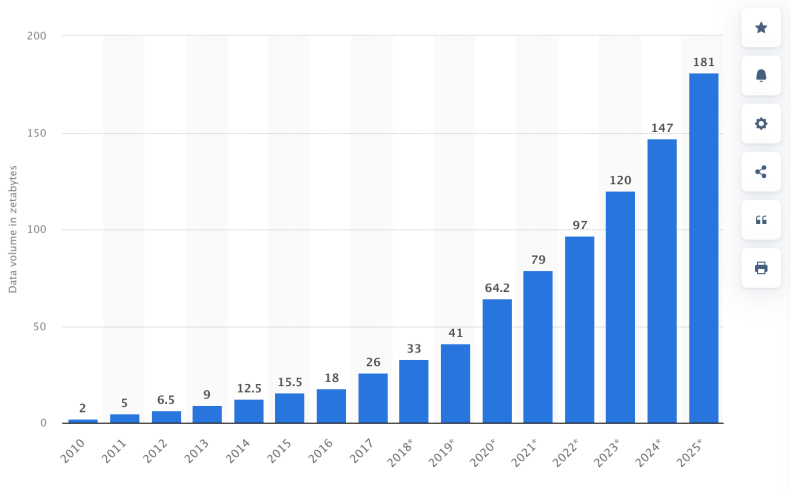
- Growth of the digital universe
  - Big Data
  - Artificial Intelligence
  - Machine Learning
- Data Science
  - Different roles
- How to build models
- Relevance of data
  - Building features
  - Statistics and data science
- The canonical tasks in data mining
  - Supervised learning
  - Unsupervised learning

2

The Growth of the Digital Universe

3

The Digital Universe



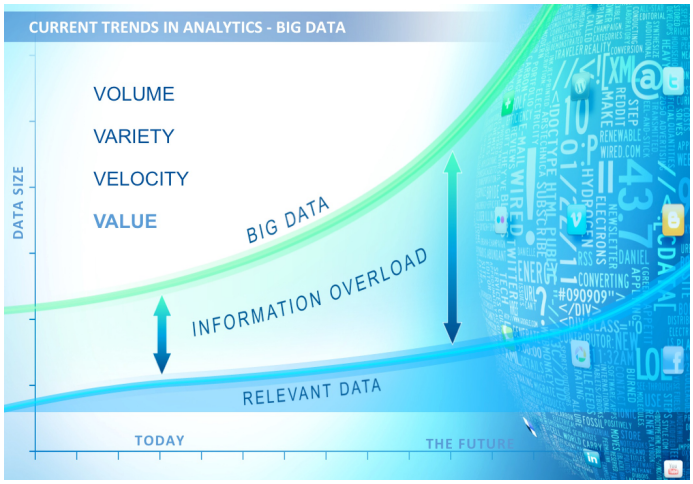
4

The Digital Universe

Source: https://www.sap.com/uk/insights/what-is-big-data.html

5

The Data Gap



6

The Economist: The data deluge

http://www.economist.com/node/15579717

7

Big Data

8

Big Data

'we will simply take big data to mean datasets that are too large for traditional data-processing systems and that therefore require new technologies."

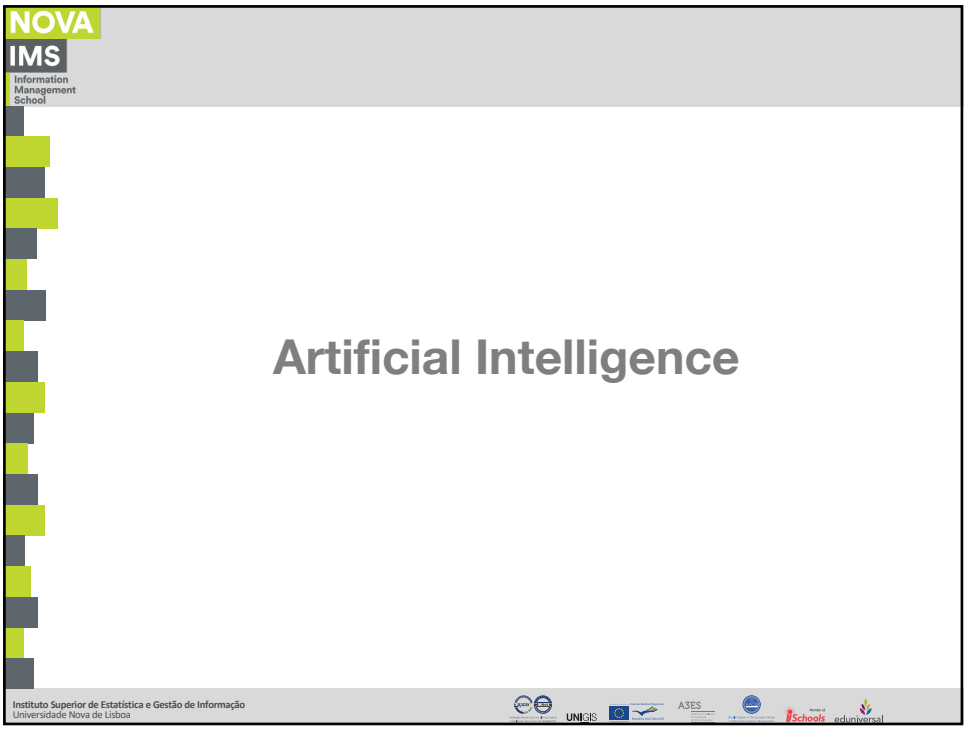
9

Big Data

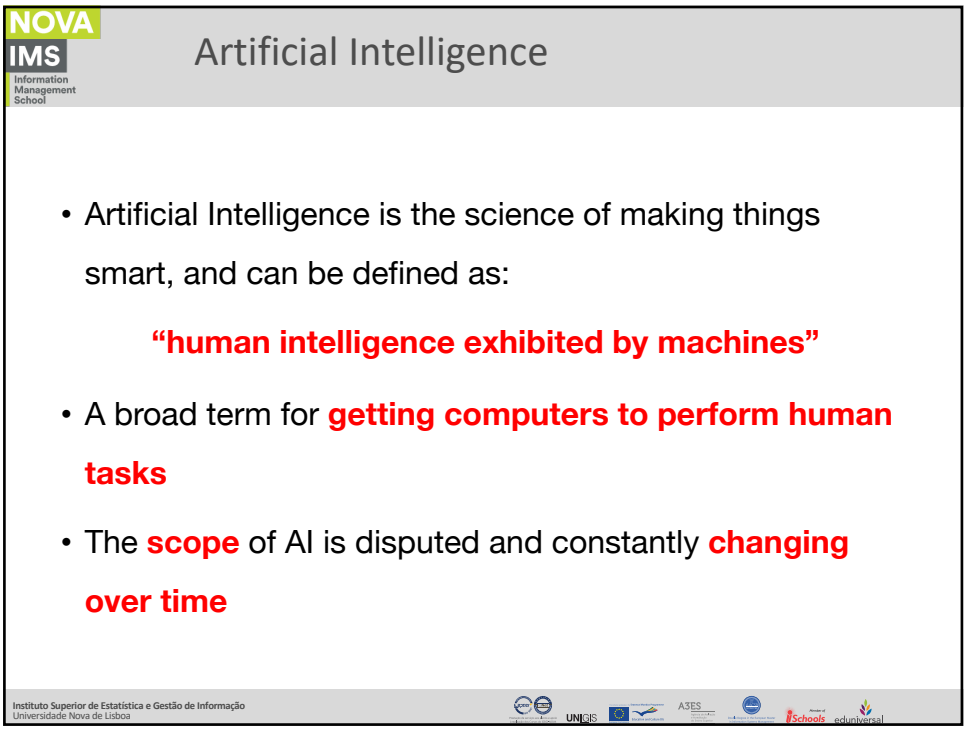
- **Volume**: While volume is by no means the only component that makes Big Data "big," it is certainly a **primary feature**. To fully manage and utilise Big Data, advanced algorithms and AI-driven analytics are required.
- **Velocity**: In the past, any data that was generated **had to later be entered into a traditional database system – often manually** – before it could be analysed or retrieved. Today, Big Data technology allows databases to process, analyse, and configure data while it is being generated – sometimes within milliseconds.
- **Variety**: Data sets that are comprised solely of structured data are not necessarily Big Data, regardless of how voluminous they are. Big Data is typically comprised of combinations of structured, unstructured, and semi-structured data. Traditional databases and data management solutions lack the flexibility and scope to manage the complex, disparate data sets that make up Big Data.

Source: https://www.sap.com/uk/insights/what-is-big-data.html

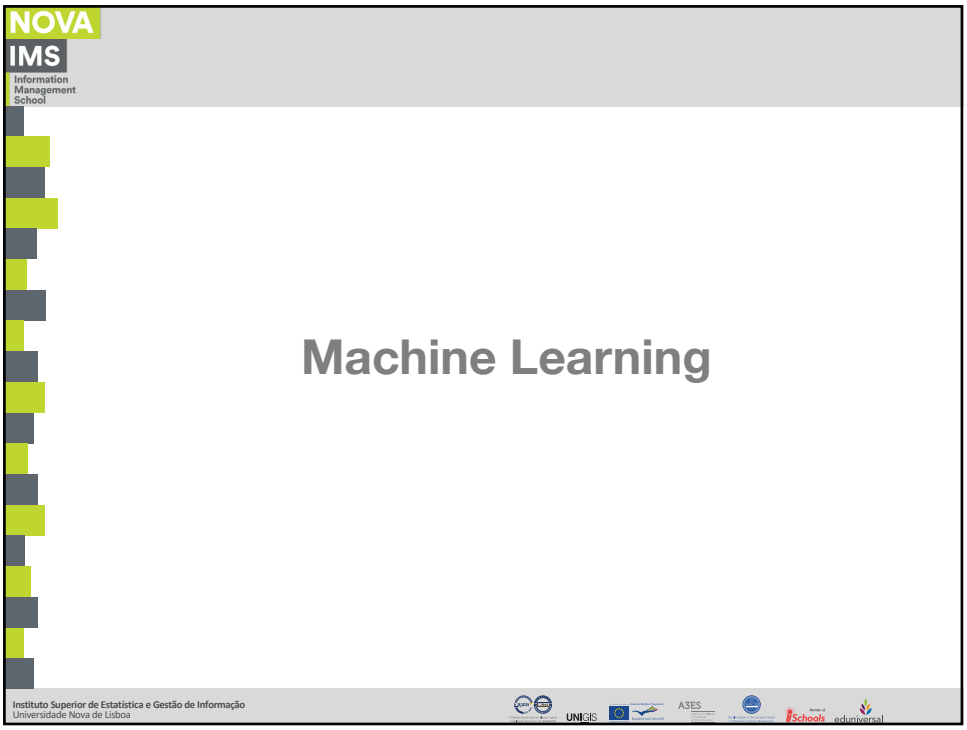
10


NOVA
IMS
Information Management School
Artificial Intelligence
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

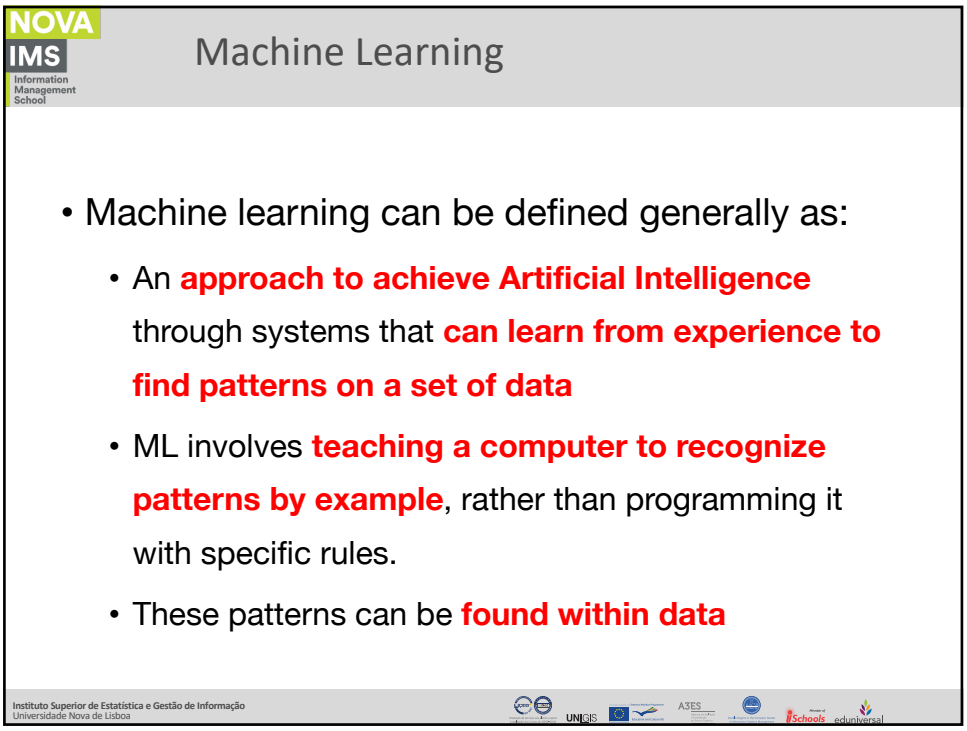
11


NOVA
IMS
Information Management School
Artificial Intelligence
• Artificial Intelligence is the science of making things smart, and can be defined as:
"human intelligence exhibited by machines"
• A broad term for getting computers to perform human tasks
• The scope of AI is disputed and constantly changing over time
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

12


NOVA
IMS
Information
Management
School
Machine Learning
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

13


NOVA
IMS
Information
Management
School
Machine Learning
• Machine learning can be defined generally as:
• An approach to achieve Artificial Intelligence through systems that can learn from experience to find patterns on a set of data
• ML involves teaching a computer to recognize patterns by example, rather than programming it with specific rules.
• These patterns can be found within data
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

14

Machine Learning

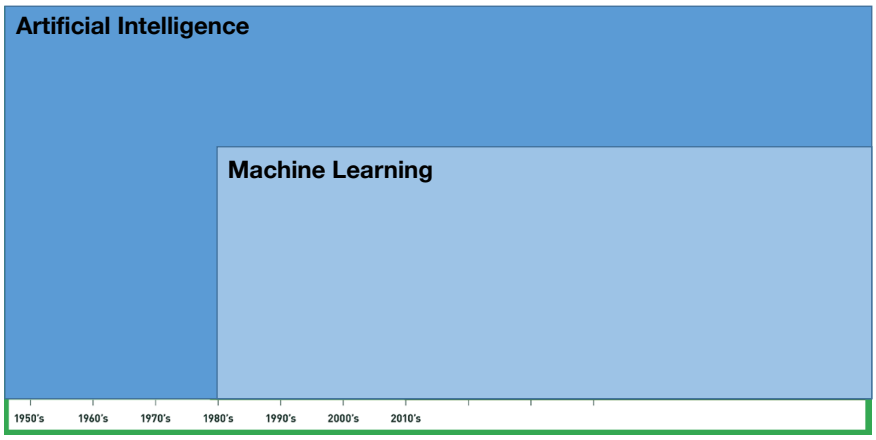
- What is extremelly cool about ML is that it learns by itself from the data

Traditional Programming	Machine Learning Programs

Write a computer program with **explicit rules** to follow
`if email contains V!agrå`
`then mark is-spam;`
`if email contains …`
`if email contains …`
	Write a computer program to **learn from examples**
	`try to classify some emails;`
	`change self to reduce errors;`
	`repeat;`

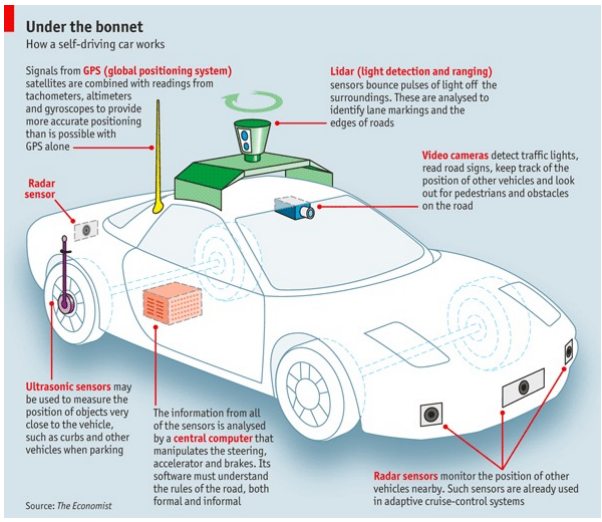
15

Machine Learning



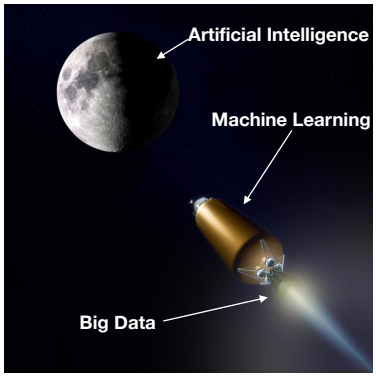
16

Artificial Intelligence vs Machine learning



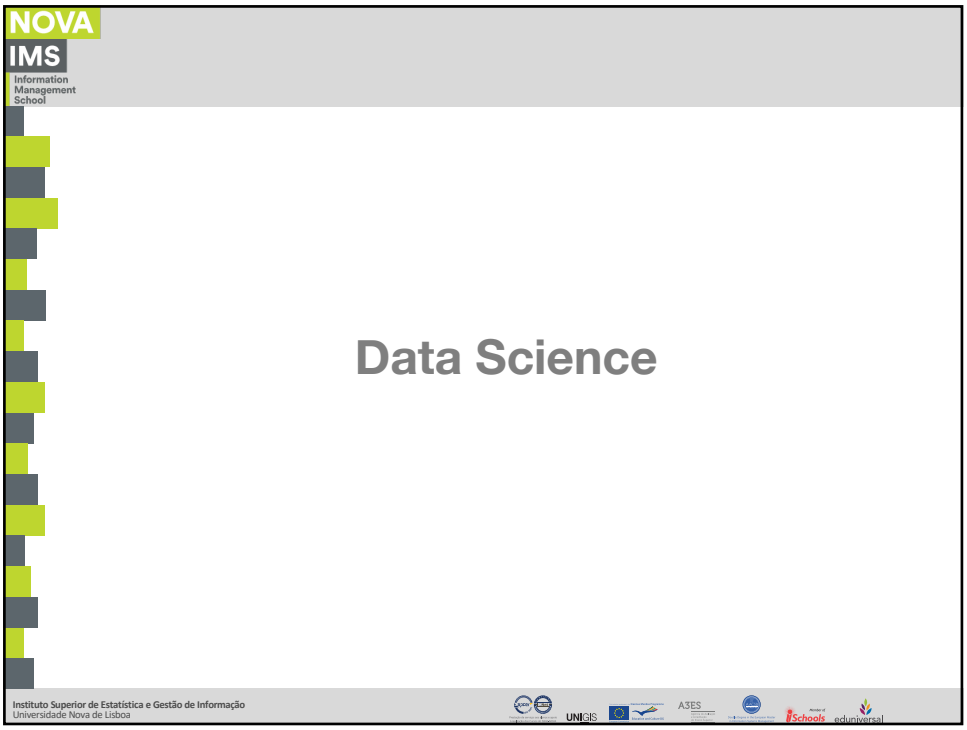
17

Artificial Intelligence and Machine Learning



Source: AI Luminary Series: Pedro Domingos Explains Machine Learning https://www.intel.com/content/www/us/en/analytics/ai-luminary-pedro-domingos-video.html

18



19

Data Science

- Data science is **the study** of where **information** comes from, what it represents and **how it can be turned into a valuable resource** in the creation of business and IT strategies.
- "Data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others," — Mike Loukides

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

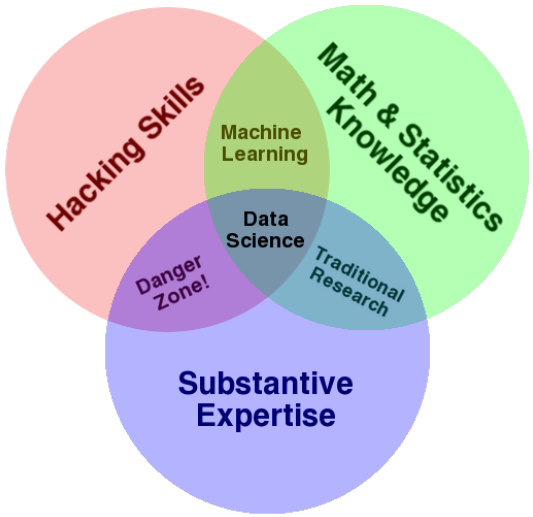
20

Data Science vs Data Mining

- **Data Science**
  - Data science is a **set of fundamental principles** that support and guide the principled extraction of information and knowledge from data.
  - Data science **involves more than just data-mining algorithms**. Successful data scientists must be able to view business problems from a data perspective.
- **Data mining**
  - The **actual extraction of knowledge** from data via technologies that incorporate the data science principles.
  - There **are hundreds of different data-mining algorithms**, and a great deal of detail to the methods of the field.

21

Data Science



22

Data Science

- **iPhone Operations Data Scientist, Apple - Santa Clara Valley - California - US**
- **Key Qualifications**
  - The position requires a software programming skill set, utilization of statistical techniques, experience managing data integrity, and implementing automated solutions. The Data Scientist will need to have an understanding of relational database management systems, design, and structured query language.
  - Excellent analytical skills, high level of statistics with the ability to identify and predict trends and anomalies.
  - Experience in data mining extremely large data sets, high proficiency in SQL (Teradata, Oracle, or MySQL), relational database management systems and design.
  - Strong programming, with excellent object-oriented and dynamic scripting language skills. Python, Java/C#/Objective-C, HTML5, CSS3, JavaScript, and Unix shell scripting are strongly desired.
  - Experience with statistical tools like JMP, Minitab, and Stata
  - Strong ability to manage multiple tasks concurrently and in a timely manner, including large, complex projects.
  - Effective presentation skills and be able to explain complex data and charts in a clear manner to large audiences. Outstanding communication skills, both verbal and written.

23

Data Science

- **Data & Applied Scientist, Microsoft - Redmond, WA, US**
- **Job requirements:**
  - Programming skills related to data using technologies like Python, PERL, C#, etc.
  - Stats and data analysis experience working with advanced tools like R, SAS and **advanced Excel**
  - Experience using data to build ML and statistical models that impact a product or business
  - Ability to access, reduce, and join large data sets programmatically
  - Natural curiosity that extends beyond your daily activities to help stitch together different areas of the company, market and technology in new ways
  - Ability to think algorithmically about product and business issues
  - Willingness to work in a start-up organization with evolving responsibilities and a wide variety of work
  - Understanding that getting value from imperfect data and systems is a core virtue for a data scientist
  - Creativity to find pragmatic paths through ambiguity
  - Attention to detail and accuracy
  - Ability to collaborate with people representing diverse points of view
  - Solid writing, presentation and data visualization skills
  - Other desirable skills and experiences:
  - Engineering experience building large data systems based on SQL, Hadoop, etc.
  - Experience with product and service telemetry systems
  - Experience with Feedback/Text analysis
  - Experience conducting multivariate experiments such as A/B testing
  - Ability to bring broad product, customer, and market context into data analyses
  - Ability to interact with senior leaders to drive product and business impact
  - 2 years' experience in data exploration, analysis, programming, and modeling
  - Degree in computer science, machine learning, statistics, math, economics, business or other scientific or quant-focused field, MS or PHD preferred.

24

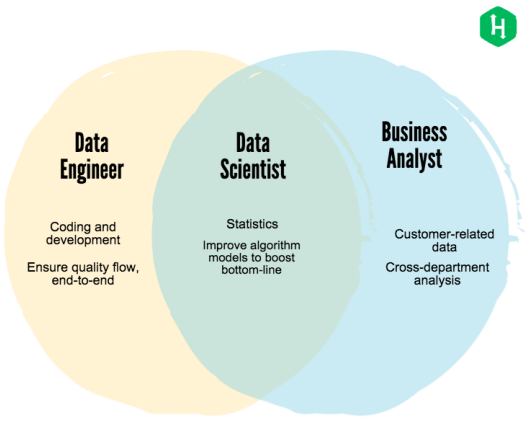
NOVA IMS Information Management School

Different Roles

25

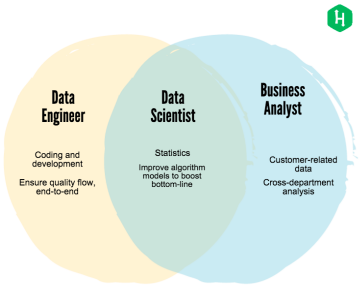
NOVA IMS Information Management School

Data Science – Different Roles



26

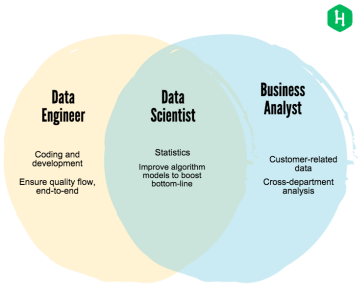
Data Science – Different Roles



- Business Analyst:
  - Business analysts' strengths lie in their business acumen.
  - They can communicate well with both the data scientist and C-suite to help drive data-driven decisions faster.
  - The best business analysts also have skills in statistics to be able to glean interesting insights from past behavior.

27

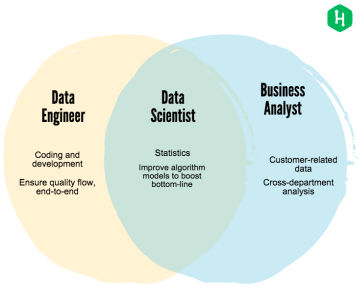
Data Science – Different Roles



- Data Scientist:
  - Data science is largely rooted in statistics, data modeling, analytics and algorithms.
  - They focus on conducting research, optimizing data to help companies get better at what they do.
  - The minds behind recommended products on Amazon.

28

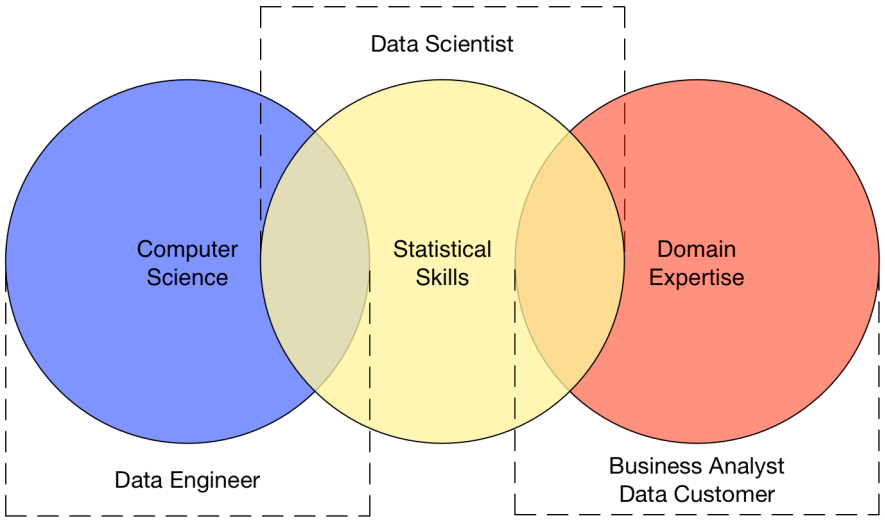
Data Science – Different Roles



- Data Engineer:
  - While data scientists dig into the research and visualization of data, data engineers ensure the data is powered and flows correctly through the pipeline.
  - They're typically software engineers who can engineer a strong foundation for data scientists or analysts to think critically about the data.

29

Data Science – Different Roles



Fonte: https://towardsdatascience.com/data-engineer-vs-data-scientist-vs-business-analyst-b68d201364bc

30

NOVA IMS Information Management School

How to Build Models

31

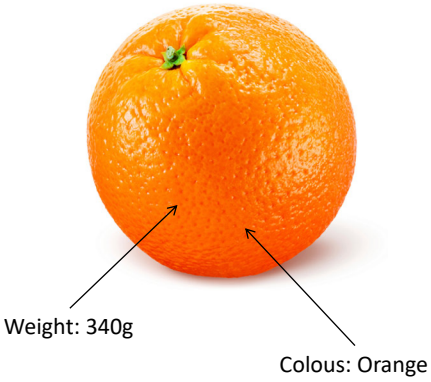
NOVA IMS Information Management School

Find/Build Attributes

32

Find/Build Attributes

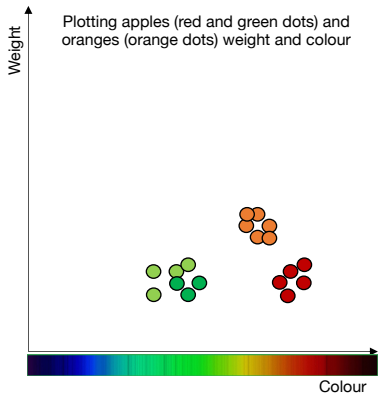
- **Features/Attributes**
  - **Features** are fundamental to train an ML system
  - They are the **properties of the things you're trying to learn about**



33

Find/Build Attributes

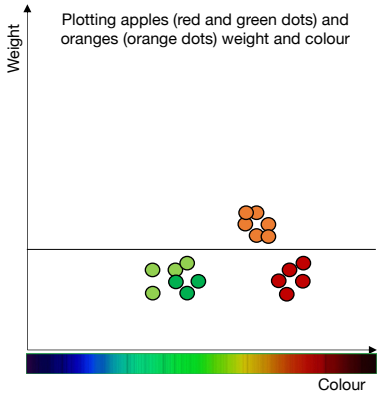
- **Features/Attributes**
  - Features of a fruit might be **weight and colour**. 2 features would mean 2 dimensions
  - 2 dimensions can be plotted in a graphic provided they are expressed numerically



34

Find/Build Attributes

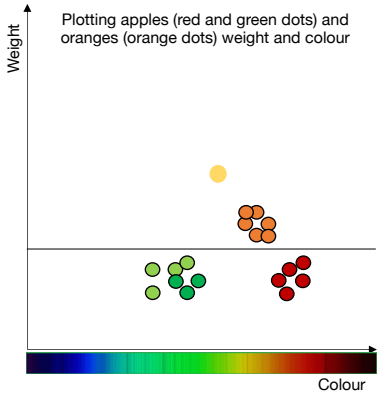
- **Features/Attributes**
  - With these features the **ML system can learn** to split data up with a **line to separate oranges from apples**
  - This can now be **used to make future classifications** when we plot new points the system has not seen



35

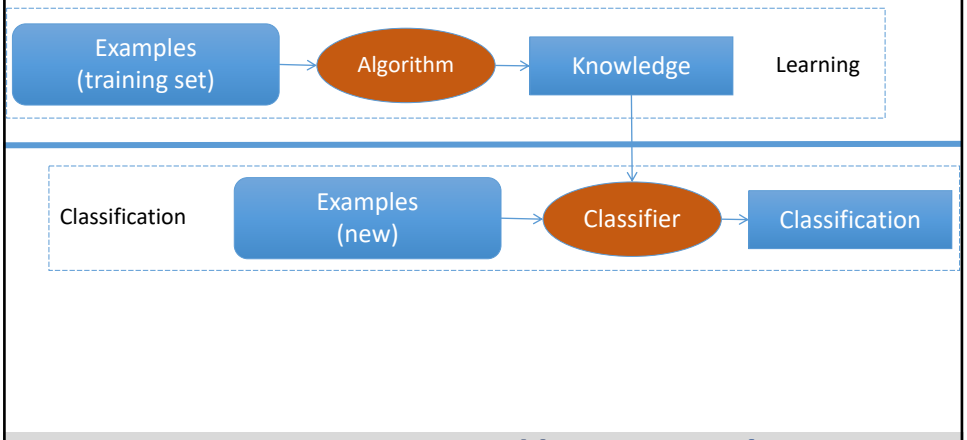
Find/Build Attributes

- **Features/Attributes**
  - An ML system cannot predict about stuff it does not know about
  - Classify a papaya
  - This is because it only knows about apples and oranges and this was the closest match



36

Find/Build Attributes



37

Let's build a model!!!

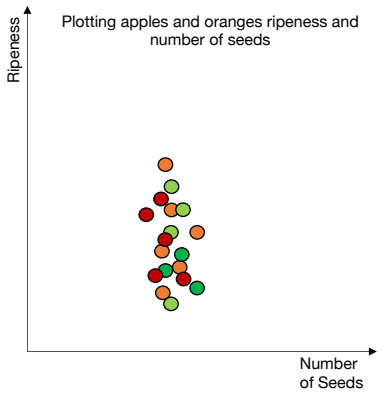
38

The Relevance of Data

39

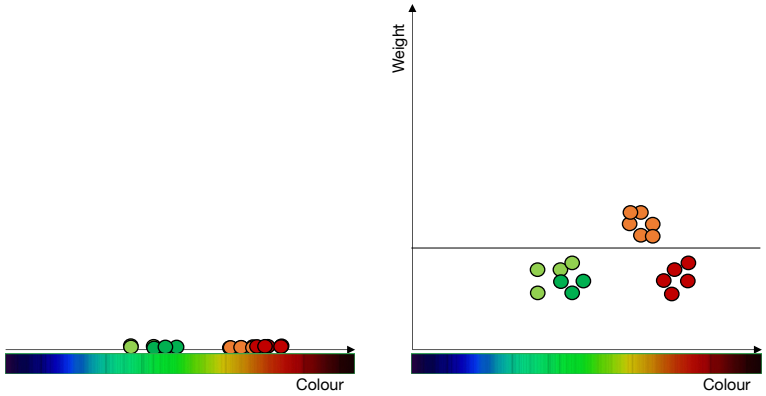
Find/Build Attributes

- **Features/Attributes**
  - Choosing the **appropriate features** has a major impact on the performance of any ML system
  - **Some features will never allow** the system to produce **good results**
  - How to choose? Practice and **knowledge about the problem**



40

Find/Build Attributes



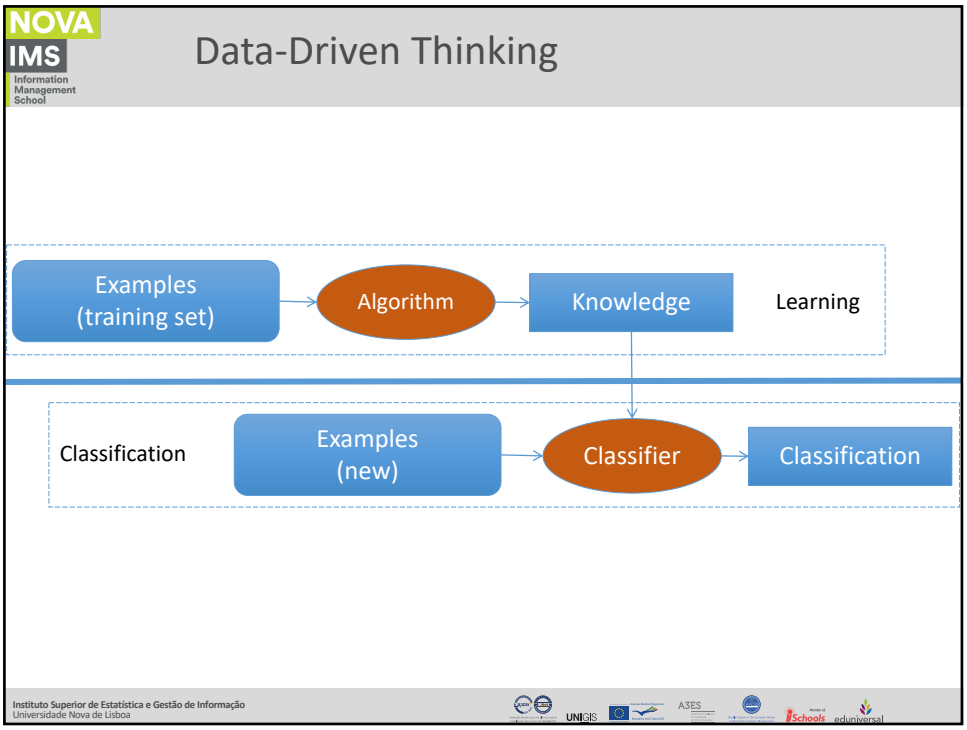
More features = higher probability of discriminating (up to a point)

41

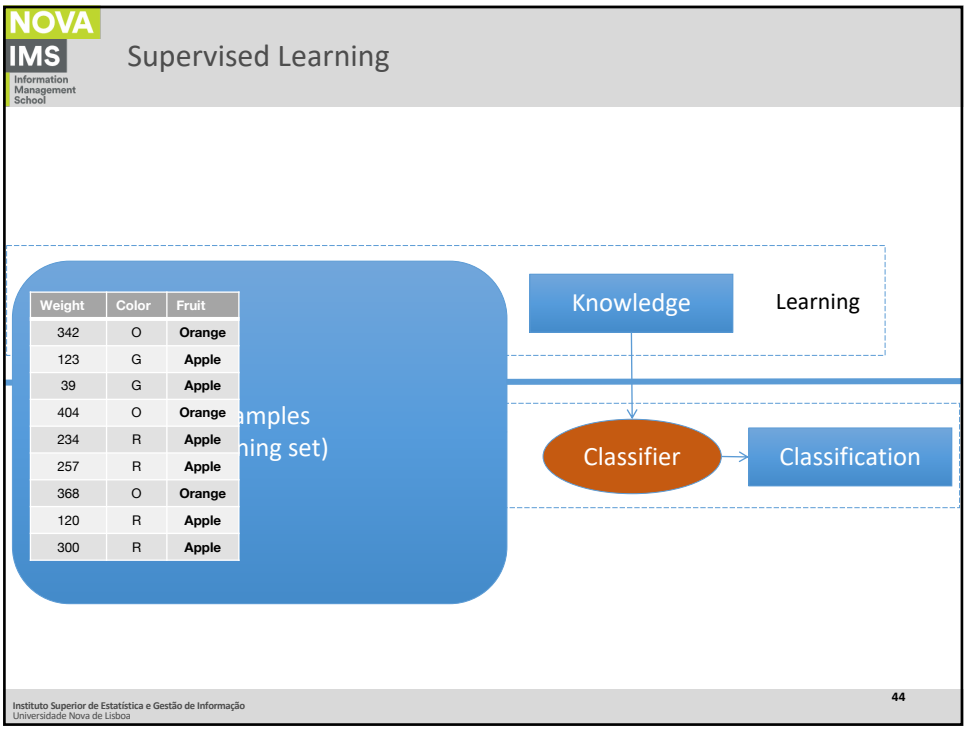
Find/Build Attributes

- Do you know **which examples correspond to apples** and which correspond to oranges?
  - We need the labels (fraud)
- Do you have **enough labeled examples**?
  - We need experience (scarce)
- Do you know **what an orange is**?
  - We need clear cut definitions (churn)

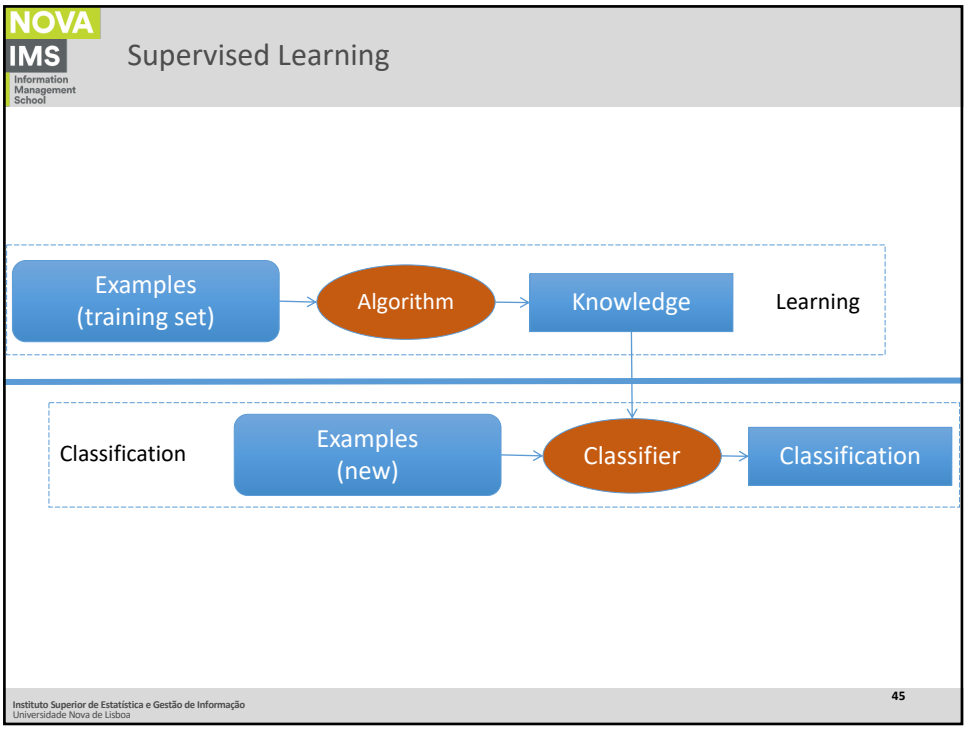
42


NOVA IMS Information Management School
Data-Driven Thinking
Examples (training set)
Algorithm
Knowledge
Learning
Classification
Examples (new)
Classifier
Classification
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

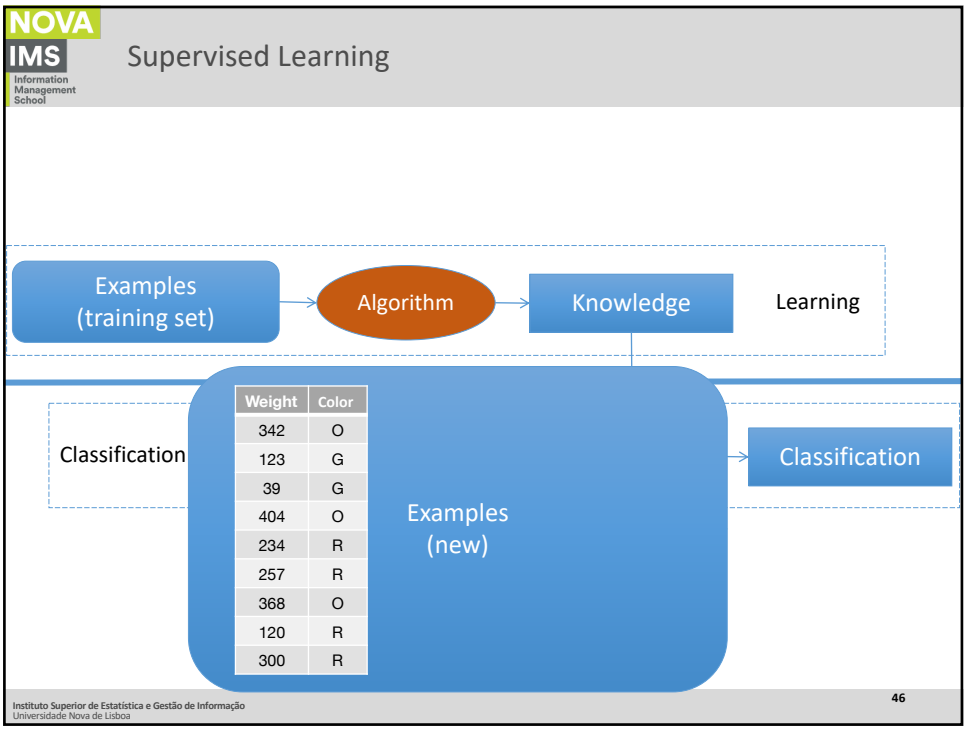
43


NOVA IMS Information Management School
Supervised Learning
Weight Color Fruit
342 O Orange
123 G Apple
39 G Apple
404 O Orange
234 R Apple
257 R Apple
368 O Orange
120 R Apple
300 R Apple
Knowledge
Learning
Classifier
Classification
44
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

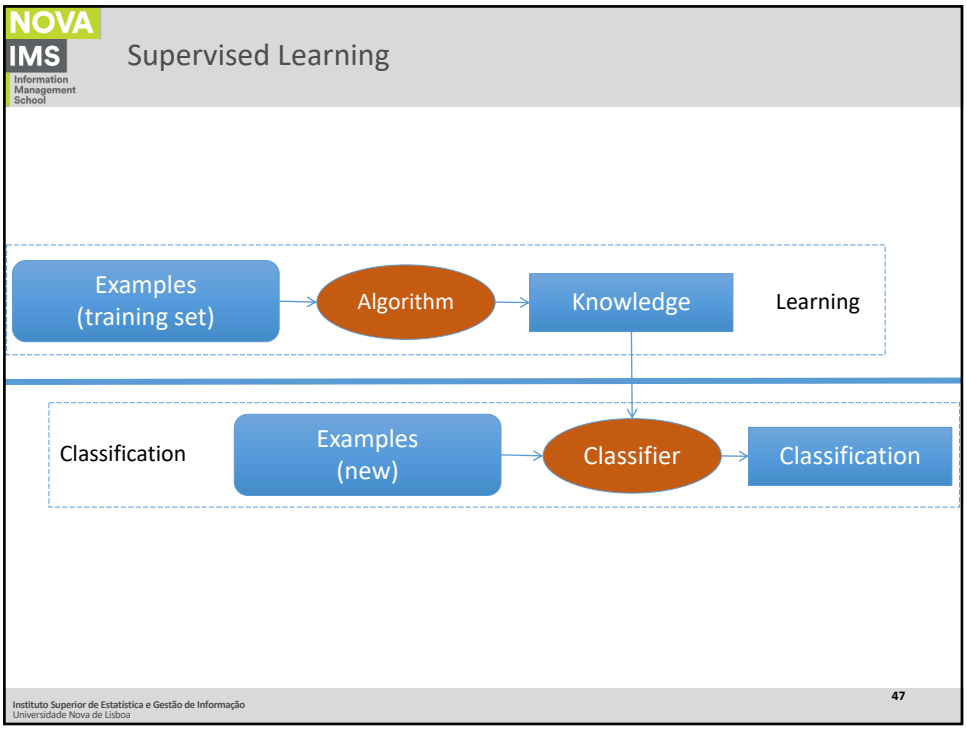
44


NOVA IMS Information Management School
Supervised Learning
Examples (training set)
Algorithm
Knowledge
Learning
Classification
Examples (new)
Classifier
Classification
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa
45

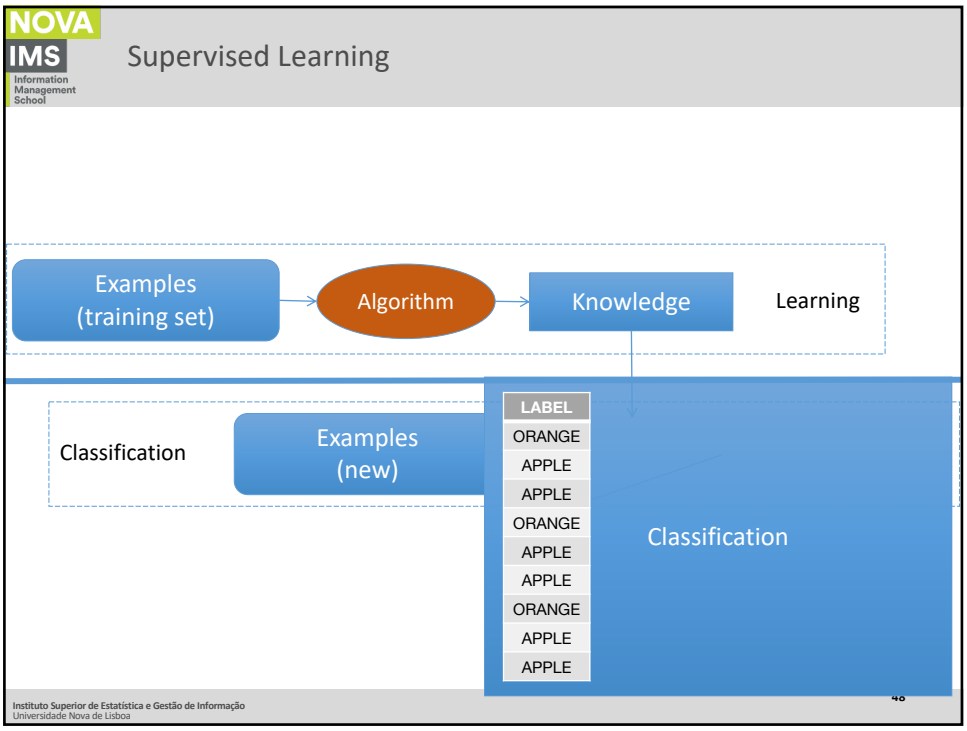
45


NOVA IMS Information Management School
Supervised Learning
Examples (training set)
Algorithm
Knowledge
Learning
Classification
Weight Color
342 O
123 G
39 G
404 O
234 R
257 R
368 O
120 R
300 R
Examples (new)
Classification
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa
46

46


NOVA IMS Information Management School
Supervised Learning
Examples (training set)
Algorithm
Knowledge
Learning
Classification
Examples (new)
Classifier
Classification
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa
47

47

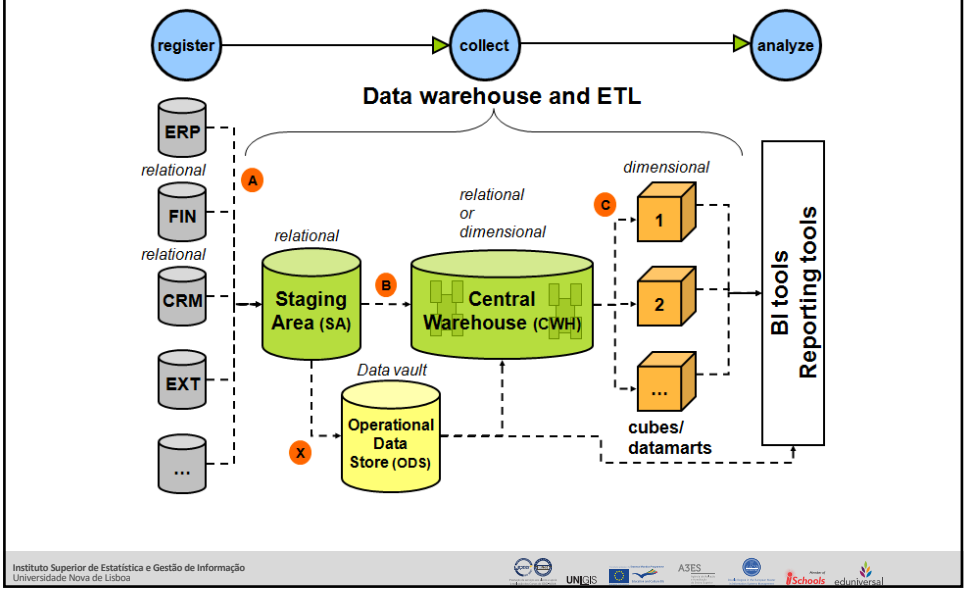

NOVA IMS Information Management School
Supervised Learning
Examples (training set)
Algorithm
Knowledge
Learning
Classification
Examples (new)
LABEL
ORANGE
APPLE
APPLE
ORANGE
APPLE
APPLE
ORANGE
APPLE
APPLE
Classification
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

48

Building Features

49

Building Features



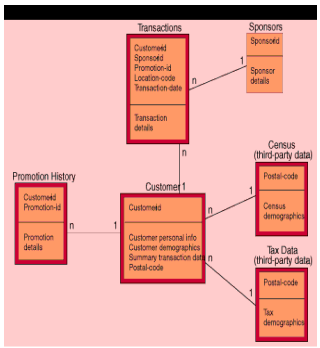
50

Building Features

- ETL (extract, transform and load)
  - Extract
    - To extract and to consolidate data from different sources.
  - Transform
    - Select variables, create new variables, merge, etc.
  - Load
    - Load data, periodicity, replacement, historical.

51

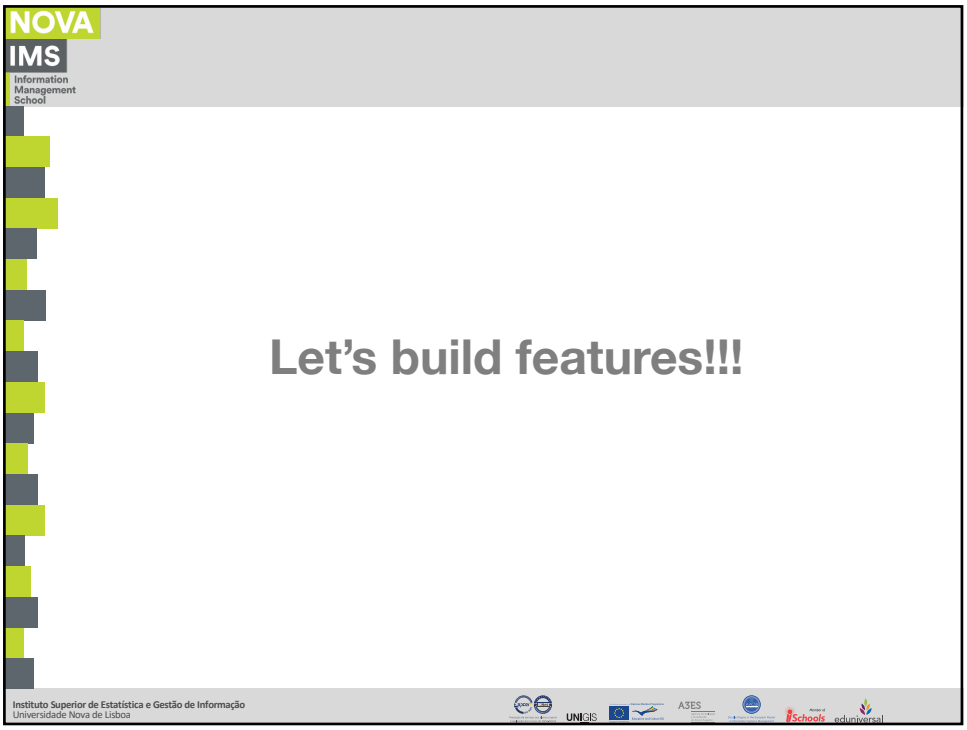
Building Features



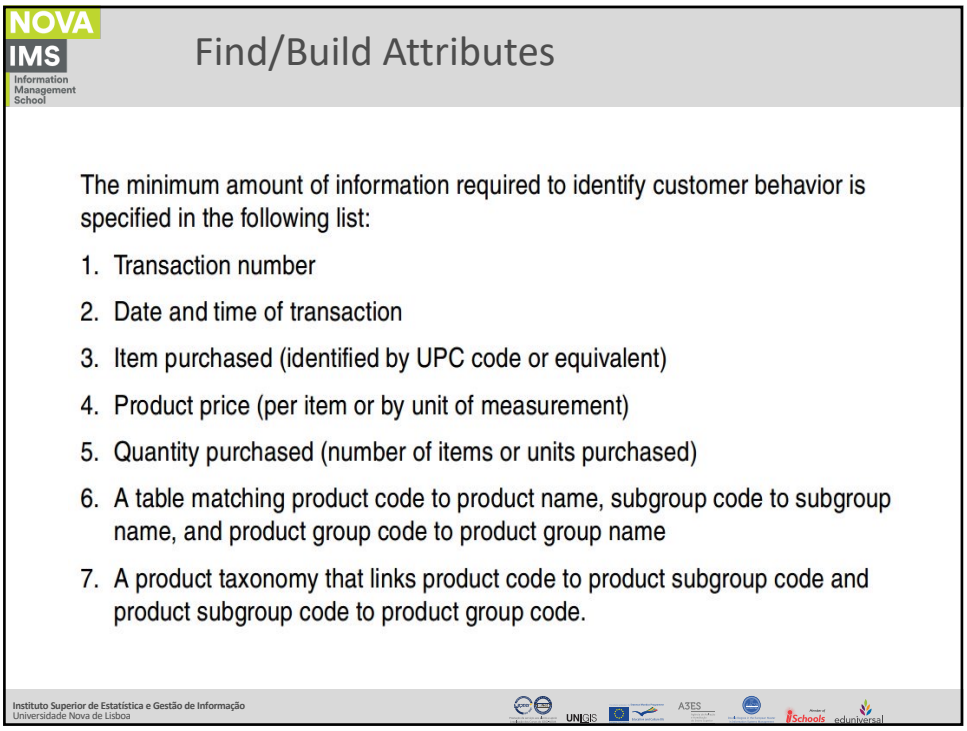
ABT (Analytic Base Table)

Height	Weight	Sex	Age	Inc.	PA	Insurance Cost
1.60	79	M	41	3000	S	N
1.72	82	M	32	4000	S	N
1.66	65	F	28	2500	N	N
1.82	87	M	35	2000	N	S
1.71	66	F	42	3500	N	S

52


NOVA
IMS
Information Management School
Let's build features!!!
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

53


NOVA
IMS
Information Management School
Find/Build Attributes
The minimum amount of information required to identify customer behavior is specified in the following list:
1. Transaction number
2. Date and time of transaction
3. Item purchased (identified by UPC code or equivalent)
4. Product price (per item or by unit of measurement)
5. Quantity purchased (number of items or units purchased)
6. A table matching product code to product name, subgroup code to subgroup name, and product group code to product group name
7. A product taxonomy that links product code to product subgroup code and product subgroup code to product group code.
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

54

Find/Build Attributes

- **Some relevant variable**
  - **Recency** – day since last visit/purchase
  - **Frequency** – number of transactions per customer
  - **Monetary Value** – total value of sales (different from profit)
  - **Average Purchase** – average of the purchase per visit
  - **Most Frequent Store**
  - **Average Time Between Transactions** – Transaction Interval
  - **Standard Deviation of Transactional Interval**
  - **Customer Stability Index** - Standard Deviation of Transactional Interval/Average Time Between Transactions
  - **Relative Spend on Each Product**

55

Agenda

- The canonical tasks in data mining
  - Supervised learning
  - Unsupervised learning
- Statistics vs data Science
- The data mining process
- General aspects of problem definition
  - Input space
  - Curse of dimensionality
  - Input space coverage
  - Separability and Bayes error
  - Different types of data
  - Spurious correlations and confounding variable

56

The Canonical Tasks in Data Mining

57

Different Tasks

Data Mining

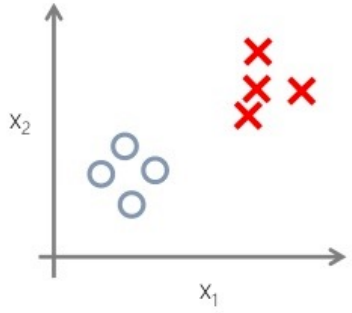
- Predictive modelling* — To learn a decision criterion that allows us to classify new and unknown examples;
- Descriptive modeling** — To describe, to provide a summary of a data set;

* aka supervised learning
** aka unsupervised learning

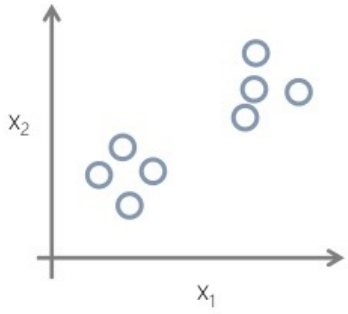
58

Different Tasks

Predictive modelling
Supervised Learning



Descriptive modeling
Unsupervised Learning

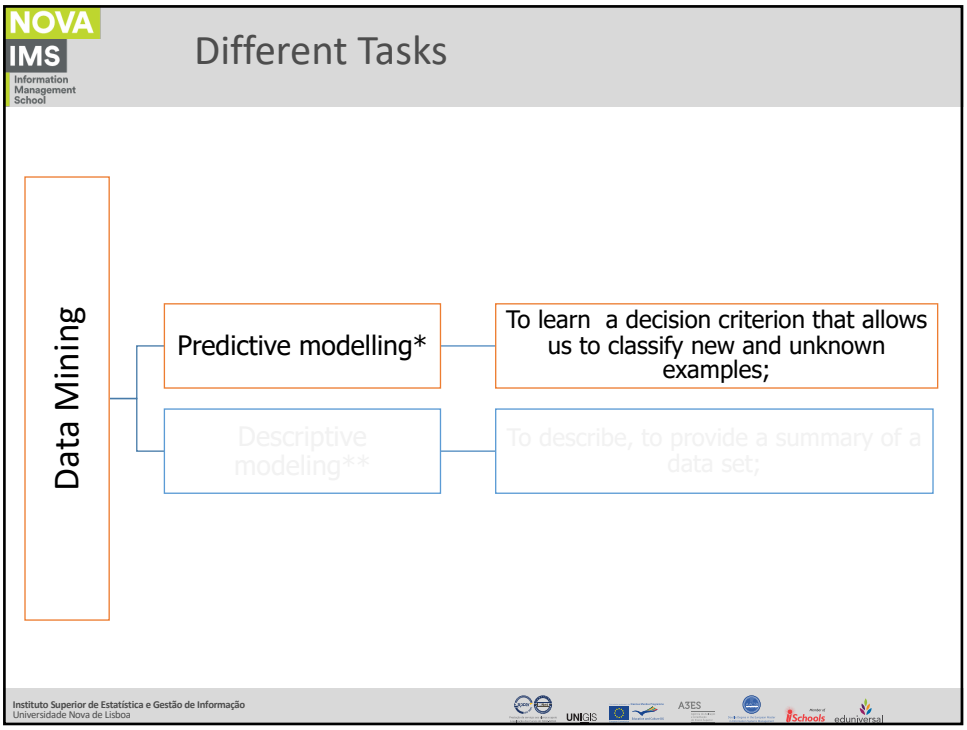


59

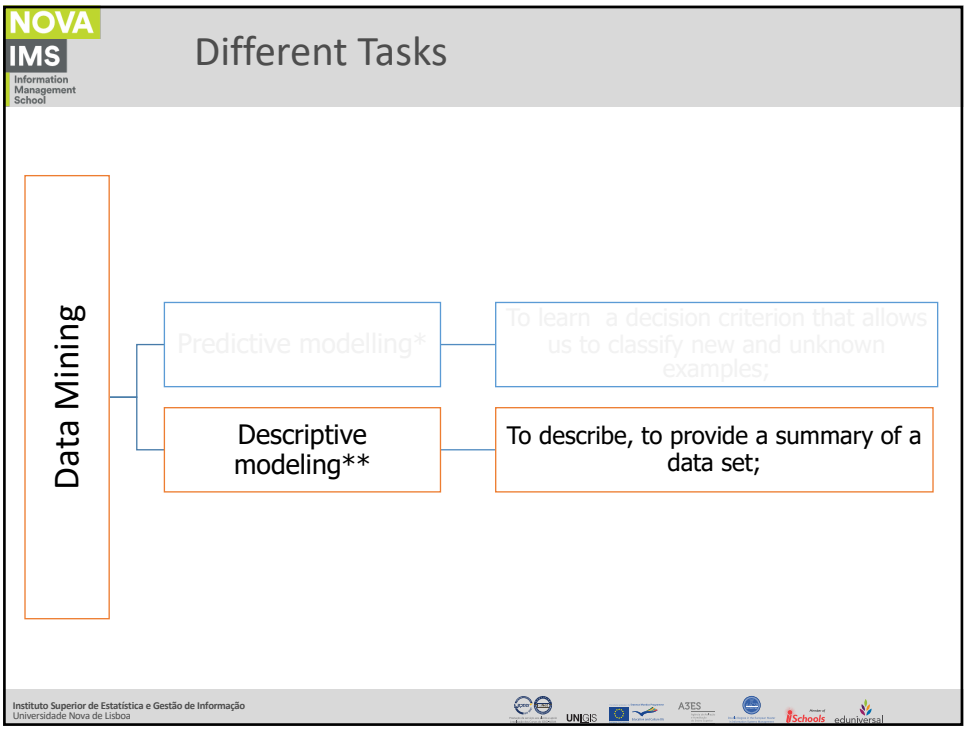
Different Tasks

- **Predictive modeling**
  - Classification
  - Regression
- **Descriptive modeling**
  - Clustering
  - Visualization
  - Association

60


NOVA IMS Information Management School
Different Tasks
Data Mining
Predictive modelling*
To learn a decision criterion that allows us to classify new and unknown examples;
Descriptive modeling**
To describe, to provide a summary of a data set;
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

61


NOVA IMS Information Management School
Different Tasks
Data Mining
Predictive modelling*
To learn a decision criterion that allows us to classify new and unknown examples;
Descriptive modeling**
To describe, to provide a summary of a data set;
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

62

Different Tasks

Feature →

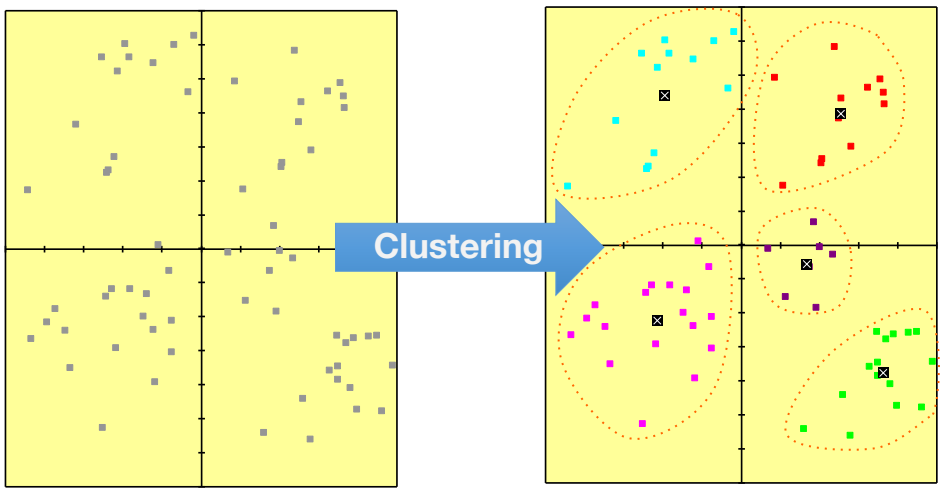
Height	Weight	Sex	Age	Income	Physical Activity
1.60	79	M	41	3000	S
1.72	82	M	32	4000	S
1.66	65	F	28	2500	N
1.82	87	M	35	2000	N
1.71	66	F	42	3500	N

63

Clustering

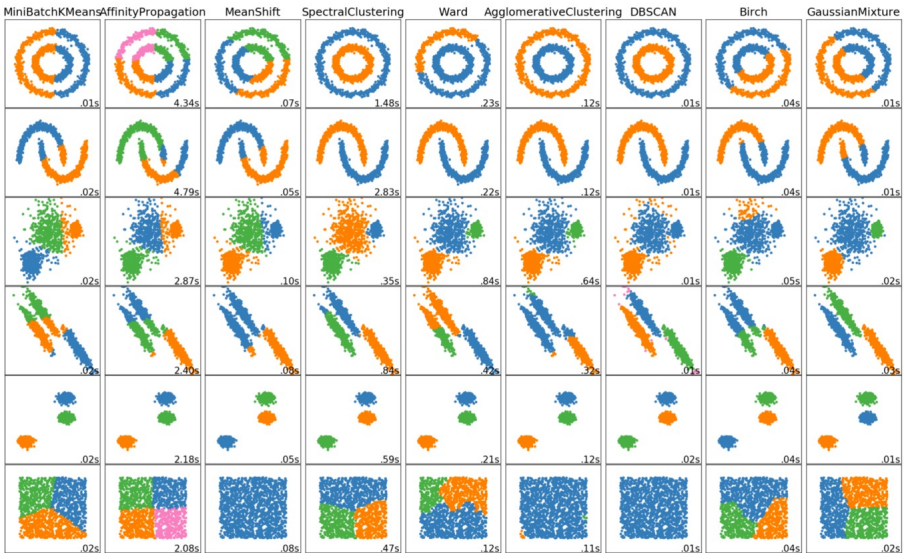
64

Different Tasks



65

Different Tasks



66

Association Rules

67

Different Tasks

Transaction Table

1,000,000 Total Transactions

200,000 Shoes

50,000 Socks

20,000 Shoes and Socks

Rule

If a customer purchases shoes, then 10% of the time he or she will purchase socks.

Evaluation Criteria:

Confidence: 20,000/200,000 = 10%

Support 20,000/1,000,000 =2%

Expected Confidence= 50,000/1,000,000 = 5%

Lift= Confidence/Expected Confidence =2

Note: The confidence factor with socks on the left-hand side and shoes on the right-hand side is 40% (20,000/50,000).

The lift value of two implies that you are twice as likely to buy socks if you bought shoes than if you did not buy shoes.

Figure 1. Association Discovery Statistics Example

68

NOVA IMS Information Management School

Visualization

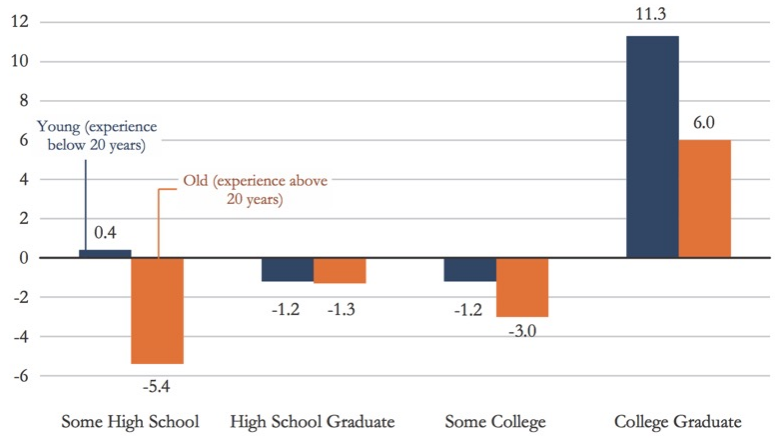
69

NOVA IMS Information Management School

Different Tasks

Flattening a 3D Chart

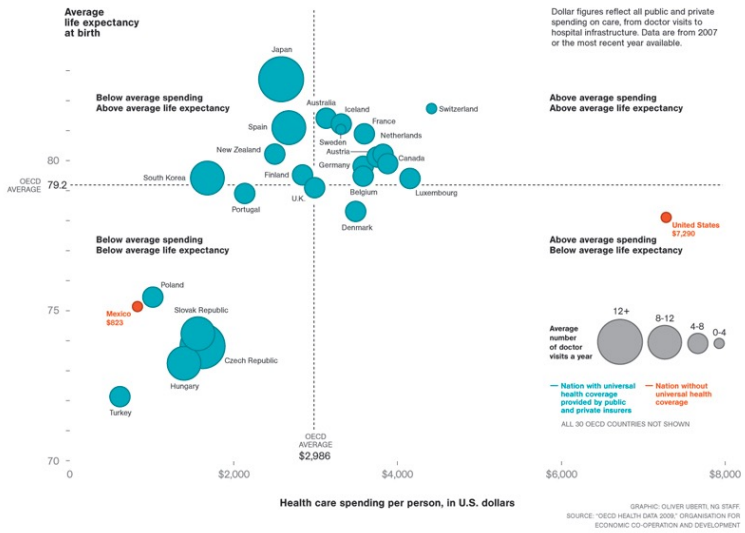
Change in real weekly wages of US-born workers by group, 1990–2006
(Percent)



70

NOVA IMS Information Management School

Different Tasks



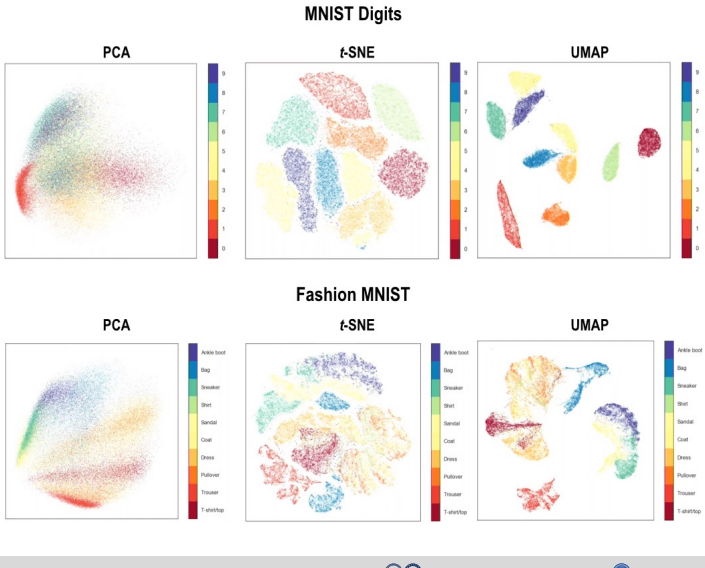
71

NOVA IMS Information Management School

Different Tasks

72

Different Tasks

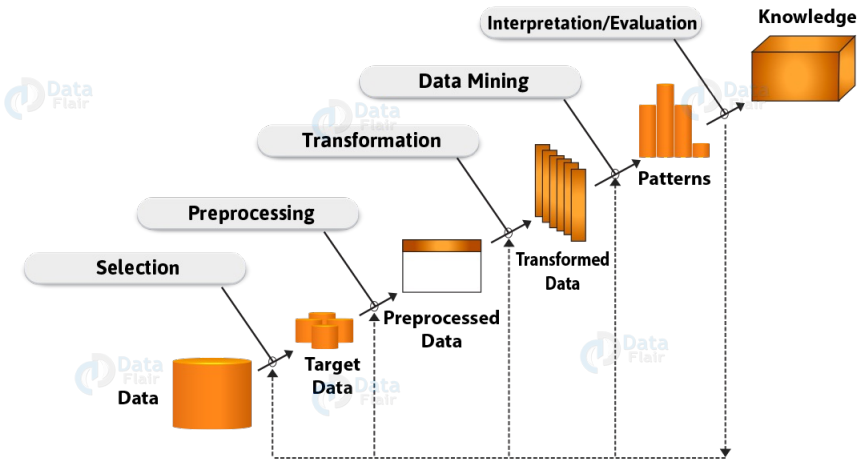


73

The Data Mining Process

74

Data Mining and Knowledge Discovery Database (KDD) Process



75

CRISP-DM, Cross Industry Standard Process for Data Mining

76

77