NOVA IMS Information Management School

Data Mining

S3

NOVA-IMS 2021/2022

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

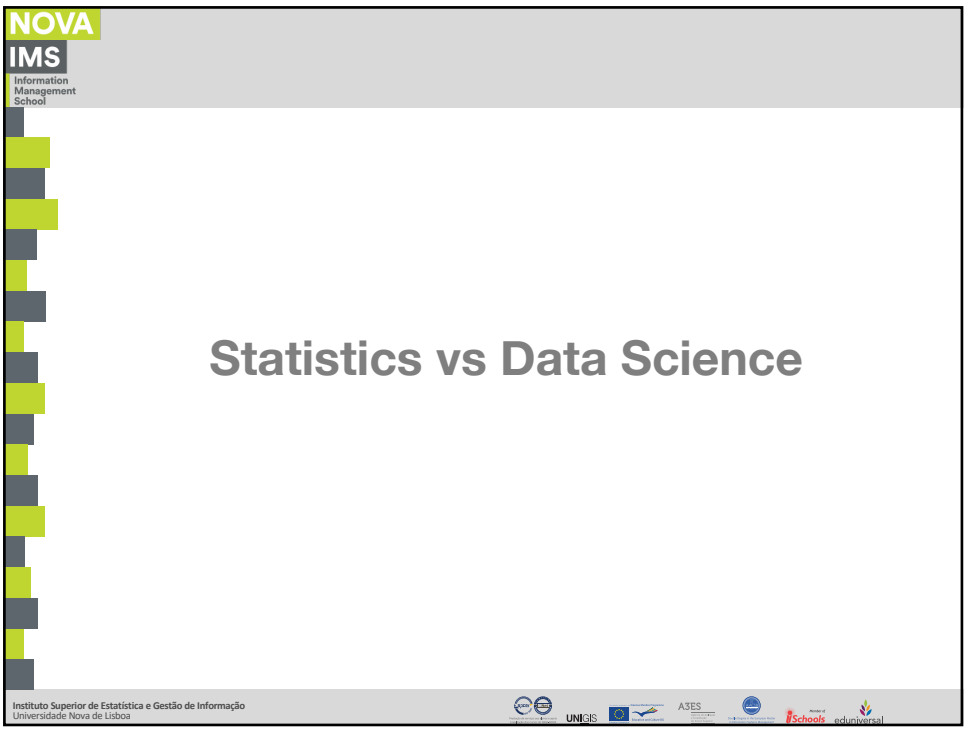
1

NOVA IMS Information Management School

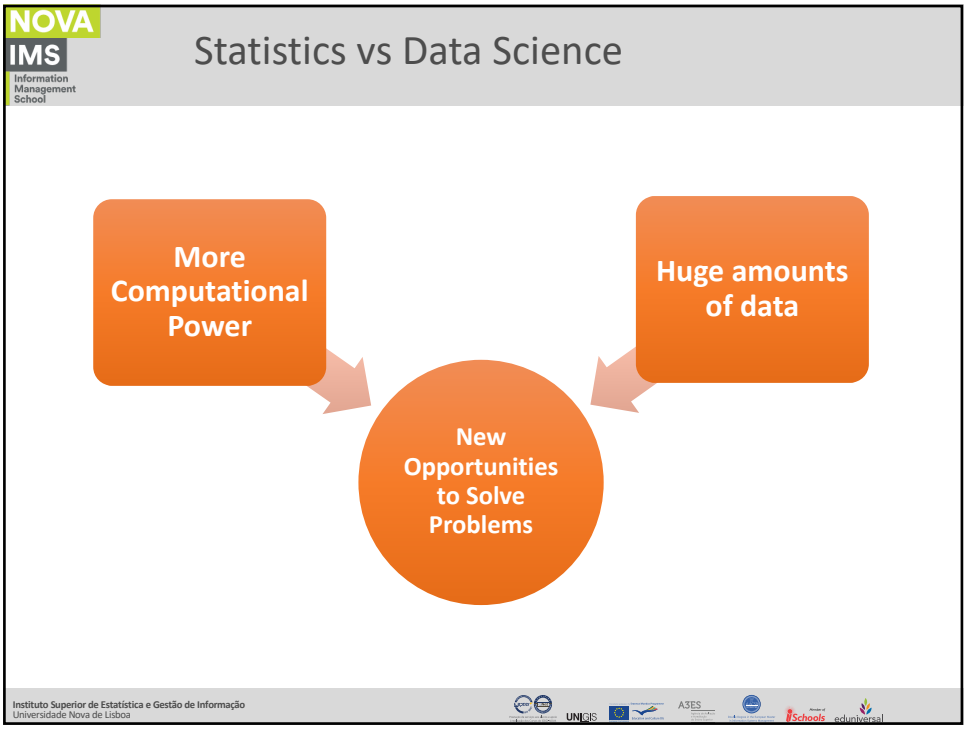
Agenda

- Data Mining
  - Statistics vs data science
  - The canonical tasks in data mining
  - Exercise
  - The data mining process
  - General aspects of problem definition
  - General aspects of data collection

2


NOVA
IMS
Information Management School
Statistics vs Data Science
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

3


NOVA
IMS
Information Management School
Statistics vs Data Science
More Computational Power
Huge amounts of data
New Opportunities to Solve Problems
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

4

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

- Statistics might be described as being characterized by **data sets which are small and clean**, which are **static**, which were **sampled in an iid manner**, which were often **collected to answer the particular problem** being addressed, and which are **solely numeric**.
- None of these apply in the data mining context…..

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

- Size of the data sets
  - Data will **not all fit into the main memory** of the computer this means that, if all of the data is to be processed during an analysis, **adaptive or sequential techniques** have to be developed (nonstatistical communities especially to those working in pattern recognition and machine learning).
  - Data sets may be large because the number of records is large or because the number of variables is large (**deep and large**).
  - Data may not be **stored as the single flat** file so but as multiple interrelated flat files.

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

Incremental (online)

- Examples are presented one at a time and the structure of representation changes.
- In the online learning, the system will handle each instance incrementally, the algorithm itself is updatable, and the knowledge will be updated by every instance in time.

Non incremental (batch)

- Examples are presented all at the same time and are considered together.

Incremental learning algorithms are usually faster than non-incremental algorithms, and for extremely large data sets, non-incremental algorithms may not be applicable at all.

7

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

- Size of the data sets
  - In the past, in many situations where statisticians have classically worked, the problem has been one of **lack of data** rather than abundance.
  - However, when data exists in the **superabundance** the results of tests (significance tests) will lead to **very strong evidence that even tiny effects exist**, effects which are so minute as to be of doubtful practical value.
  - In place of statistical significance, we need to consider more carefully substantive significance: **is the effect important or valuable or not?**

8

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

- Contaminated data
  - In the data mining context, when the analysis is necessarily **secondary**, data is always “dirty”.
  - When the data sets are large, it is practically certain that some of the data will be **invalid in some way**.
  - This is especially true when the **data describe human interactions** of some kind, such as marketing data, financial transaction data, or human resource data.

9

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

- Nonstationarity, Selection Bias, and Dependent Observations
  - Very large data sets are **unlikely to arise in an iid** manner;
  - **Population drift**, can arise because the underlying population is changing (for example, the population of applicants for bank loans may evolve as the economy heats and cools). Supermarket transactions or Telco phone calls occur every day, not just one day, so that the database is a constantly evolving entity
  - **Selection bias**, it arises when developing scoring rules, typically in this situation comprehensive data is available only for those previous applicants who were graded good risk by some previous rule. Those graded bad would have been rejected and hence their true status never discovered.

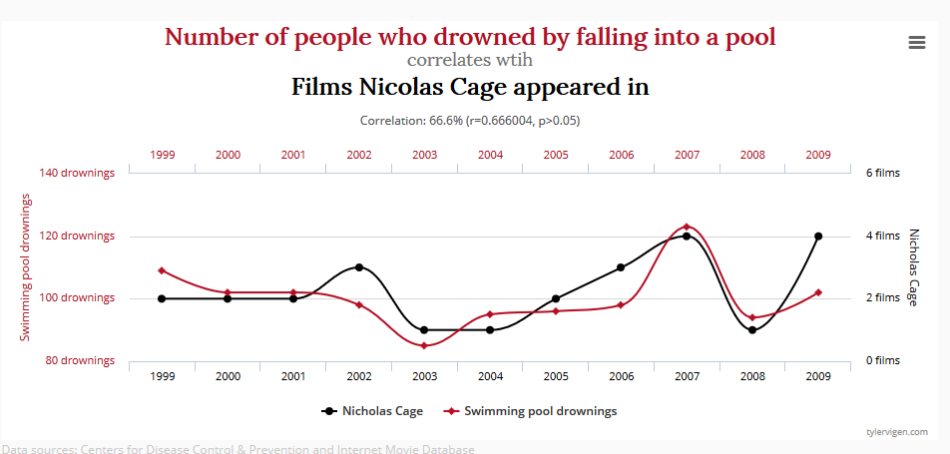
10

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

- Spurious Relationships and Automated Data Analysis
  - Because the pattern searches will throw up a **large numbers of candidate patterns**, there will be a **high probability that spurious** (chance) data configurations will be identified as patterns.
  - The bottom line is that those patterns and structures identified as potentially interesting will be presented to a **domain expert for consideration to be accepted or rejected** in the context of the substantive domain and objectives, and not merely on the basis of internal statistical structure.

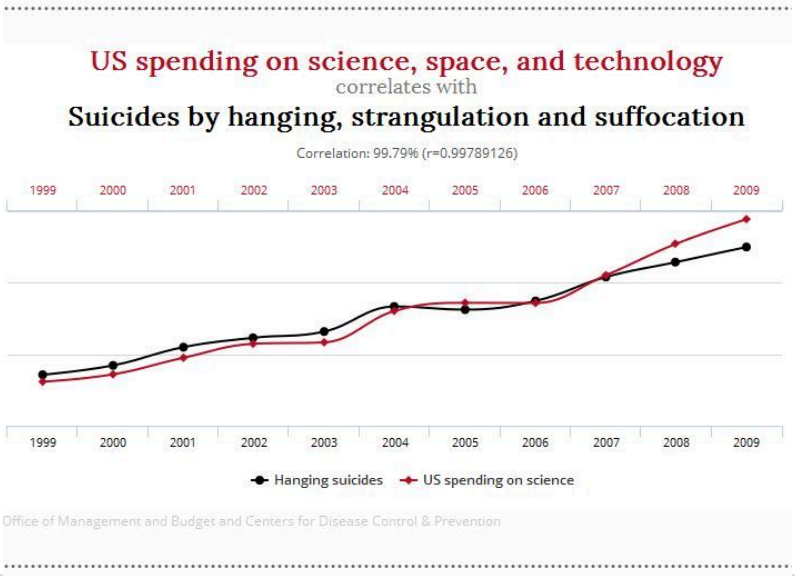
11

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)



12

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)



13

Statistics vs Data Science

	Experimental Primary	Opportunistic Secundary
Purpose	Research	Operational
Value	Scientific	Commercial
Origin	Controlled	Passively observed
Size	Small	Massive
Hygiene	Clean	Dirty
Status	Static	Dynamic

14

What is Data Mining?

*Data mining is a new **discipline lying at the interface** of statistics, database technology, pattern recognition, machine learning, and other areas.*

*It is concerned with **the secondary analysis of large databases** in order to find previously unsuspected relationships which are of interest or value to the database owners.*

*Data analysis is as much an **art as a science**.*

David J. Hand

15

The Canonical Tasks in Data Mining

16

Different Tasks

Data Mining

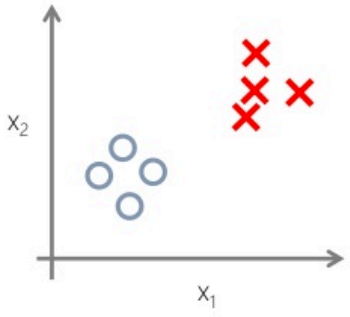
- Predictive modelling* — To learn a decision criterion that allows us to classify new and unknown examples;
- Descriptive modeling** — To describe, to provide a summary of a data set;

* aka supervised learning
** aka unsupervised learning

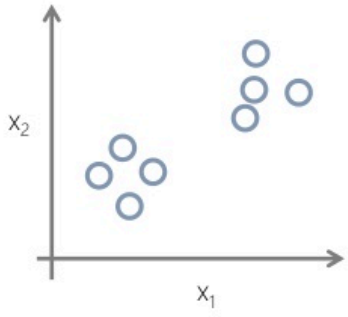
17

Different Tasks

Predictive modelling
Supervised Learning

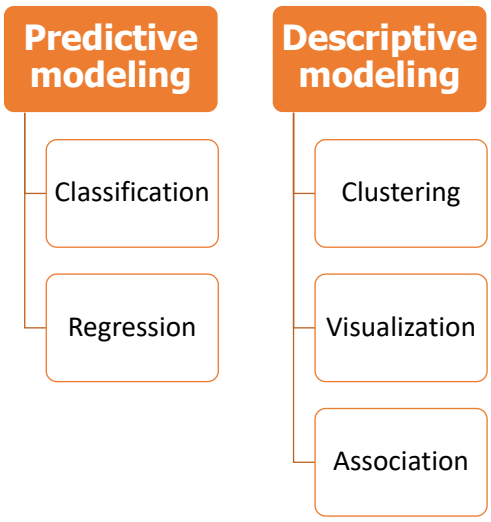


Descriptive modeling
Unsupervised Learning



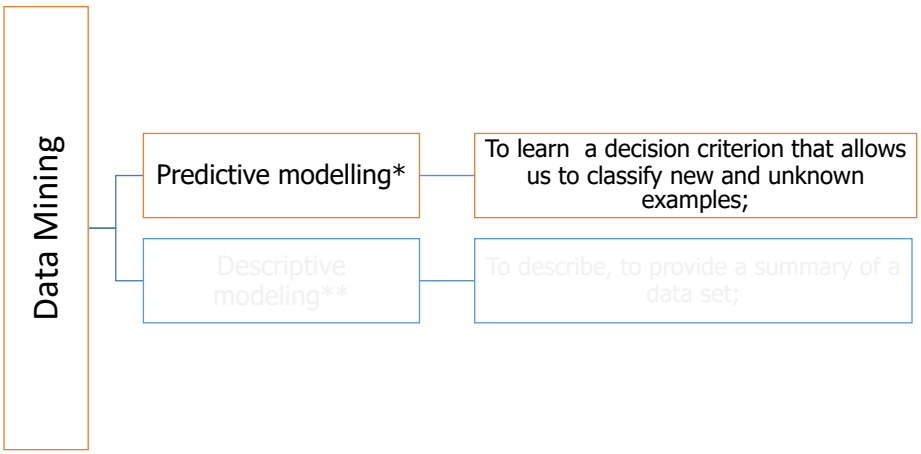
18

Different Tasks



19

Different Tasks



20

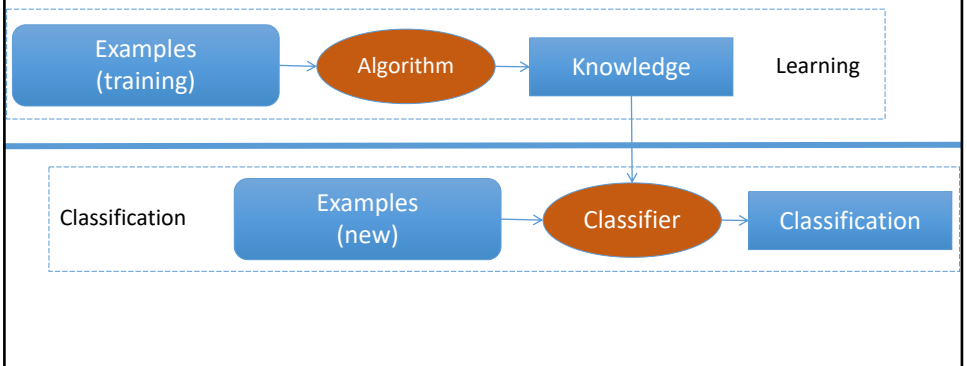
Predictive Modeling

Feature → **Label** →

Height	Weight	Sex	Age	Income	Physical Activity	Insurance Costs
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

21

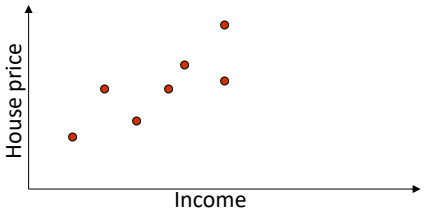
Predictive Modeling

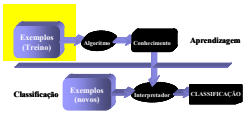


22

Predictive Modeling

- A real estate agency wants to estimate the price range for each customer based on their income;
- Training examples:
  - Historical data;
  - Income vs sold house prices.

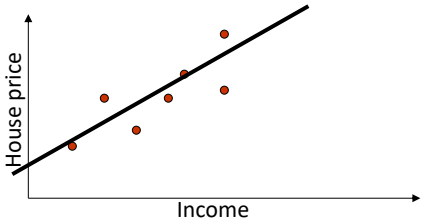


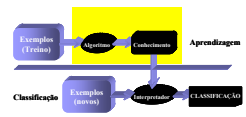


23

Predictive Modeling

- Algorithm
  - Linear regression
- Knowledge representation
  - Regression line (slope and origin)

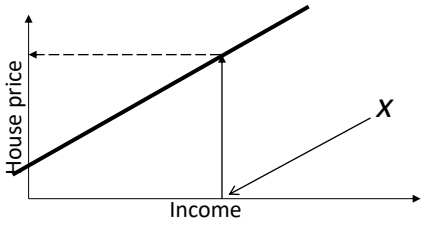


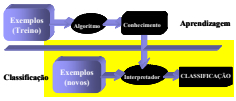


24

Predictive Modeling

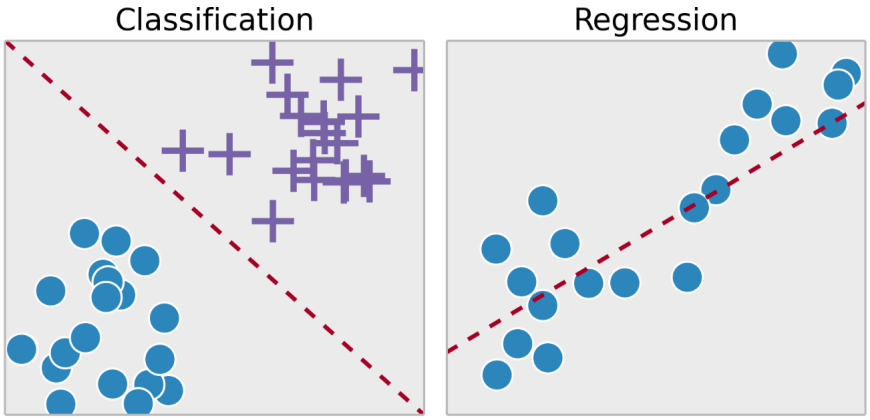
- New examples
  - A customer with an income of *x*
- Interpretation
  - Use the line (prediction method) to obtain an estimate





25

Predictive Modeling

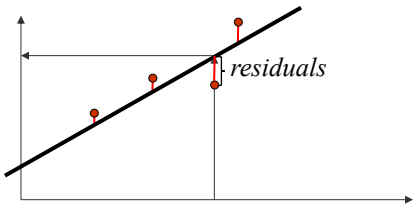


Source: http://ipython-books.github.io/featured-04/

26

Predictive Modeling

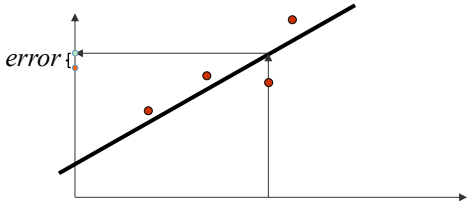
- If we are facing a regression problem, then we have to consider the average deviations produced by the model.



27

Predictive Modeling

- If we are facing a regression problem, then we have to consider the average deviations produced by the model.



28

Predictive Modeling

- If we are facing a classification problem, then we only need to count the number of times that the model was wrong;

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

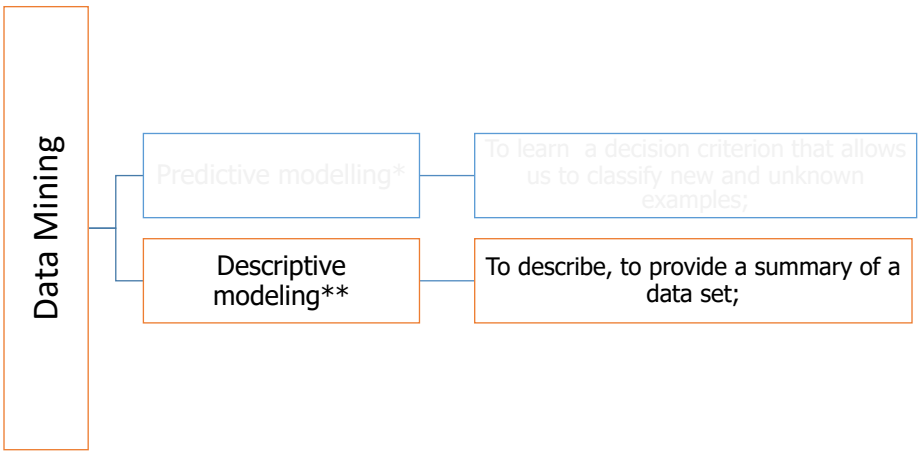
$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

29

Different Tasks



30

Different Tasks

Feature

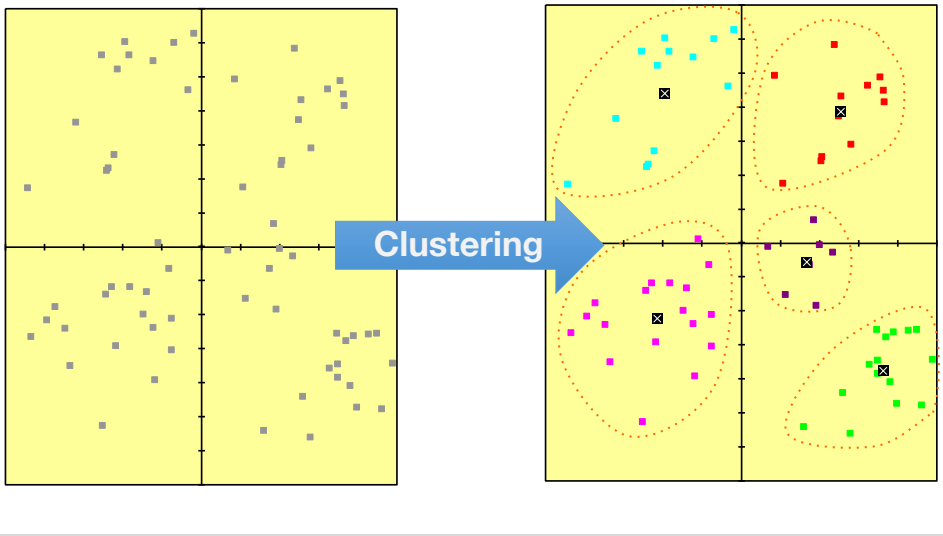
Height	Weight	Sex	Age	Income	Physical Activity
1.60	79	M	41	3000	S
1.72	82	M	32	4000	S
1.66	65	F	28	2500	N
1.82	87	M	35	2000	N
1.71	66	F	42	3500	N

31

Clustering

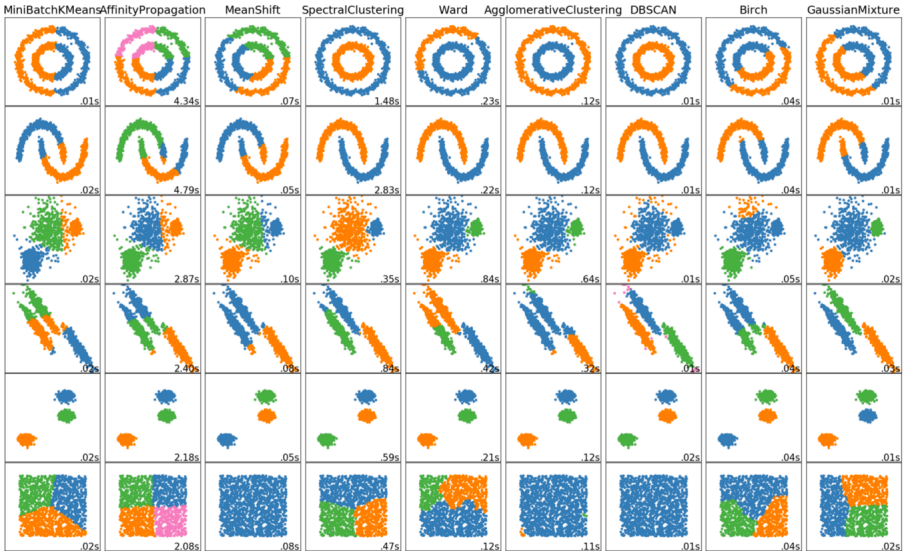
32

Different Tasks



33

Different Tasks



34

NOVA IMS Information Management School

Association Rules

35

NOVA IMS Information Management School

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

36

NOVA IMS Information Management School

Visualization

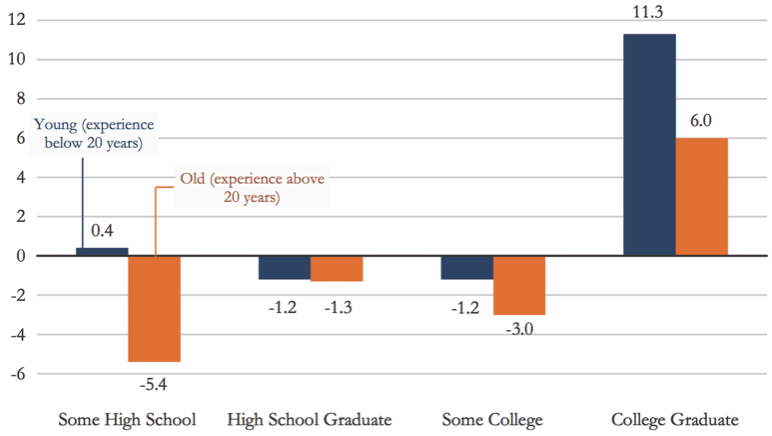
37

NOVA IMS Information Management School

Different Tasks

Flattening a 3D Chart

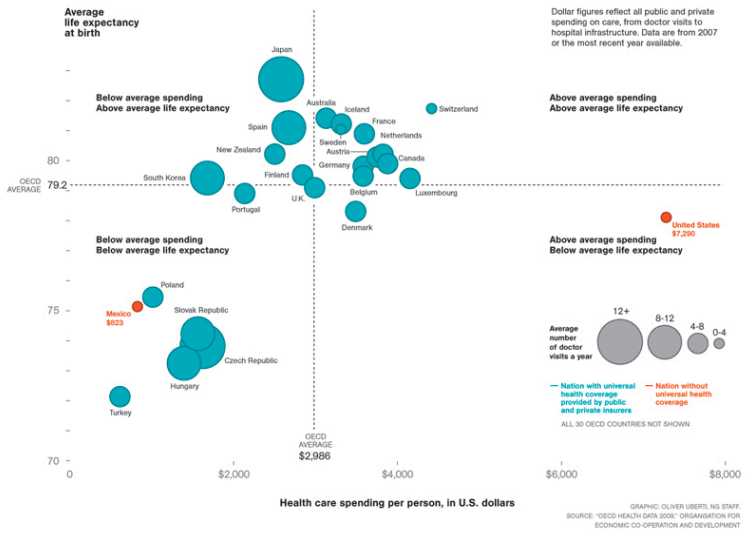
Change in real weekly wages of US-born workers by group, 1990–2006
(Percent)



38

NOVA IMS Information Management School

Different Tasks



39

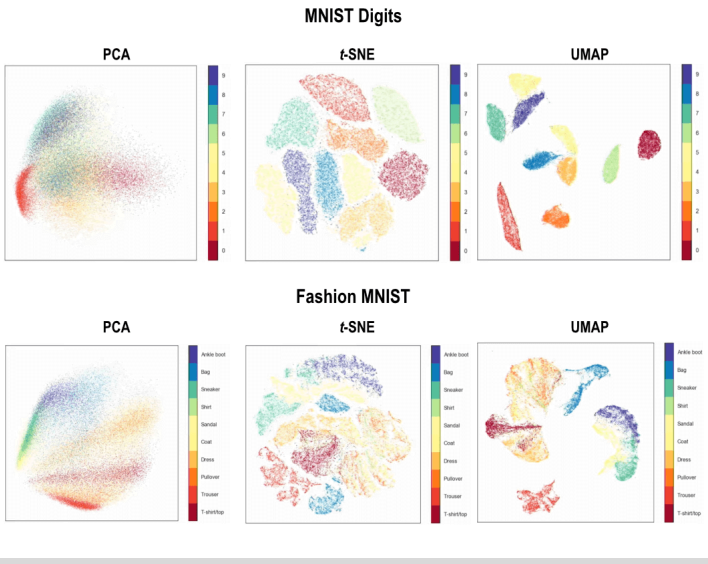
NOVA IMS Information Management School

Different Tasks

40

NOVA IMS Information Management School

Different Tasks



41

NOVA IMS Information Management School

Let's do a predictive model!!!

42

43