Data Protection and Privacy

University of Genoa

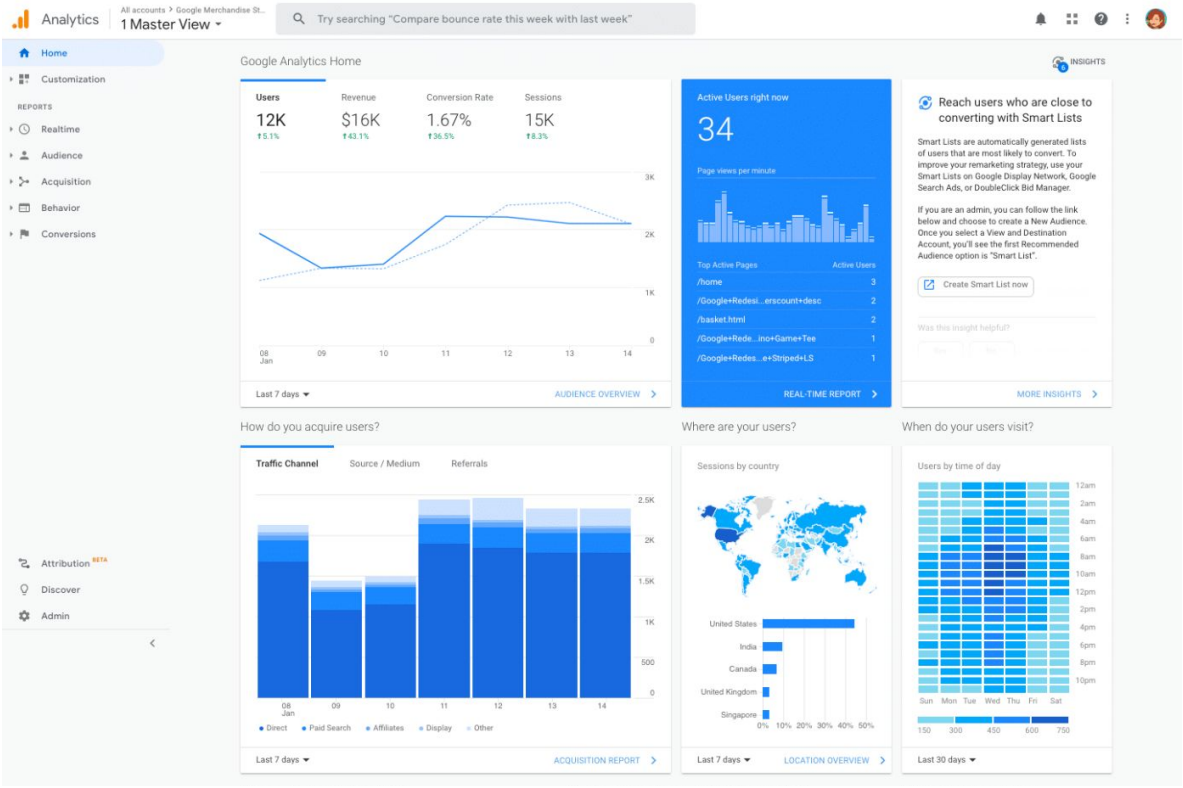
Lesson 5: Privacy Preserving Data Mining

Gaspare Ferraro <ferraro@gaspa.re>

Data Mining

- **Massive amounts of data** are being collected by companies in different ways: online trackers, smart devices, ...
- These data are an **asset** to the companies, and they are mined to **extract knowledge**
- Data mining is a process where critical business data are analyzed to gain new insights about customers, businesses, and markets.
- This new knowledge gained can be used to:
  - improve customer relationships
  - improve website navigation
  - define advertising plans
  - produce better-quality products and services

Data Mining - Website analytics



Data Mining - Data structures

- Mined data are *generally* stored in a relational or multidimensional format and stored in companies' central data **warehouses**.
- But with the evolution of the enterprise, a diverse set of **data structures** have come to be used:
  - graph data, which could feed from social network sites
  - time series data
  - longitudinal data
  - semistructured data, such as XML
  - unstructured data
  - big data

Data Mining - Threats & Security requirements

- There is a need for **different data repositories** to store all these diverse data
- Analytics is carried out on the data in the repositories
- Access to these data repositories is strictly controlled by access control rights
- **Strict security measures** are employed to secure the data as they are very sensitive and contain customer-identifying information
- Companies need to ensure that the data are anonymized **before** being used for analytics or mining
- Companies share also their data with specialized analytics firms, and the data need to be protected before sharing

Example of Open Data - Almalaurea

Publication year ---

25th Survey (2023) Graduates' Profile 2022	77 universities	281.000 graduates	summary report	query the databank	Conference -Palermo
24th Survey (2022) Graduates' Profile 2021	77 universities	300.000 graduates	summary report	query the databank	Conference -Bologna
23rd Survey (2021) Graduates' Profile 2020	76 universities	291.000 graduates	summary report	query the databank	Conference -Bergamo
22nd Survey (2020) Graduates' Profile 2019	75 universities	290.224 graduates	summary report	query the databank	Conference -Roma
21st Survey (2019) Graduates' Profile 2018	75 universities	280.230 graduates	summary report	query the databank	Conference -Roma
20th Survey (2018) Graduates' Profile 2017	74 universities	276.195 graduates	summary report	query the databank	Conference -Torino
19th Survey (2017) Graduates' Profile 2016	71 universities	272.225 graduates	summary report	query the databank	Conference -Parma

Example of Open Data - Istat

Istat | Istituto Nazionale di Statistica

POPULATION & HOUSEHOLDS | INSTITUTIONS & SOCIETY | EDUCATION & LABOUR | ECONOMY | ENVIRONMENT & TERRITORY | SEARCH IN THIS WEBSITE | A-Z Statistics | Glossary

HOME > ANALYSIS AND PRODUCTS > DATASETS

[ITALIANO]

DATASETS

ANALYSIS AND PRODUCTS
- DATABASES
  - StatBase
- DATASETS
- MICRODATA FILES
  - Recognition
- PRESS RELEASES
- PUBLICATIONS
  - Review of official statistics
- DATA VISUALIZATIONS
  - Dashboard
- INTERACTIVE CONTENTS
  - Baby names
- OPEN DATA IN ISTAT
- A-Z STATISTICS
- SMART STATISTICS FROM BIG DATA

METHODS AND TOOLS

INFORMATION AND SERVICES

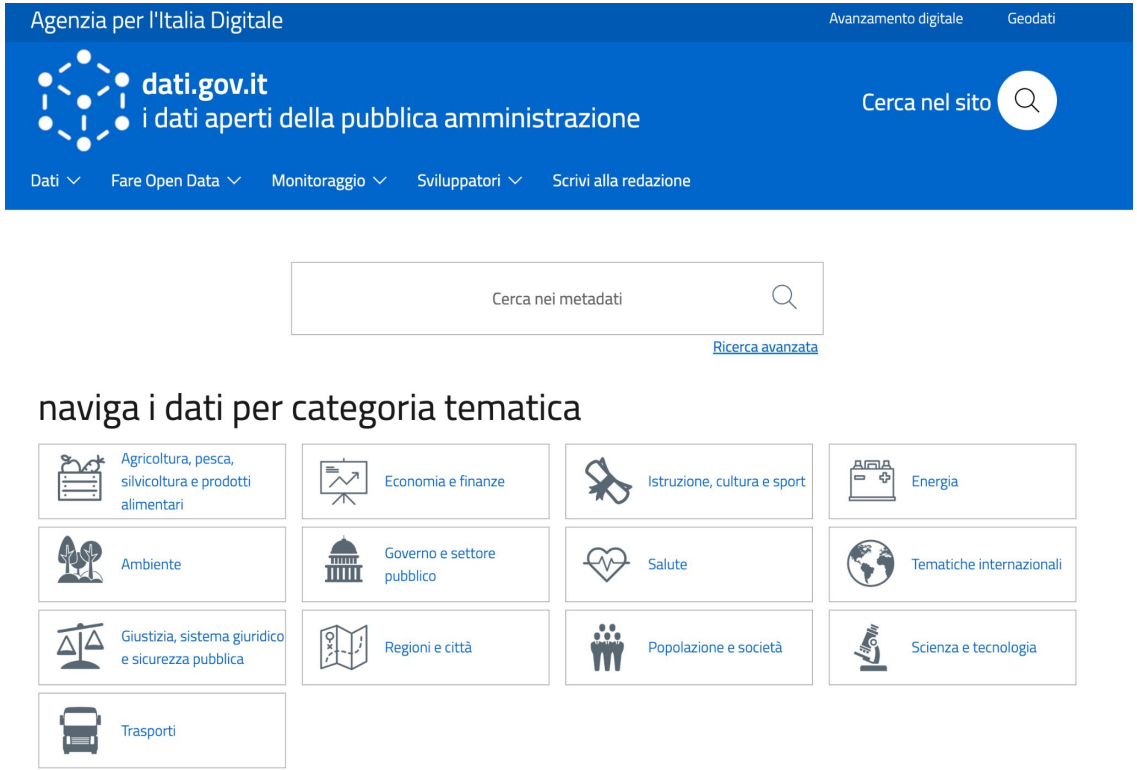
The datasets produced by Istat are collections of data disseminated without a regular frequency. They are generally produced when surveys are concluded, as a preliminary form of publication of the data produced.

Datasets are available on spreadsheet and have introductory and methodological note. They can be downloaded for free.

Mostly, dataset are in italian language.

Last edit: **22 April 2018**

Example of Open Data - Dati Gov



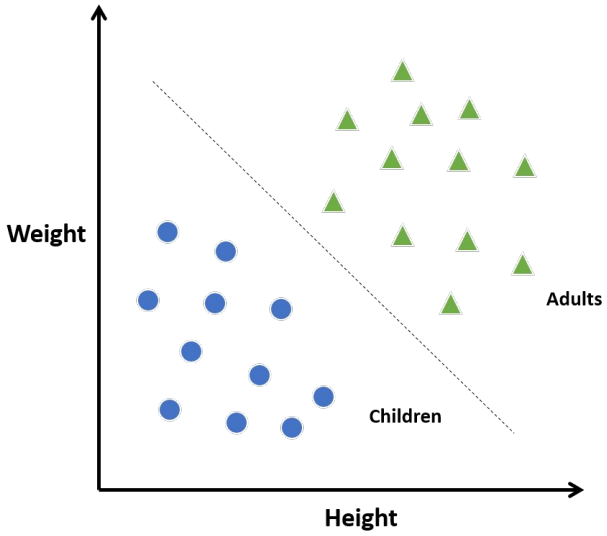
Key feature of Data Mining

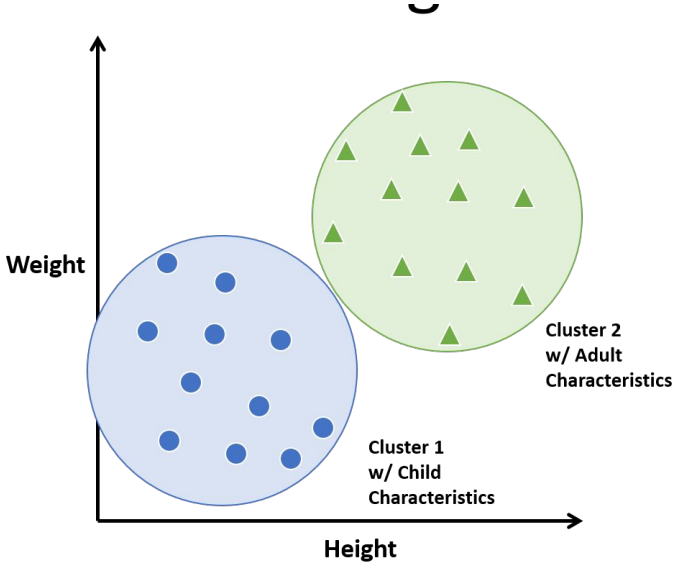
The goal of data mining is to **extract knowledge** from the data

Some of the key functions of data mining are as follows:

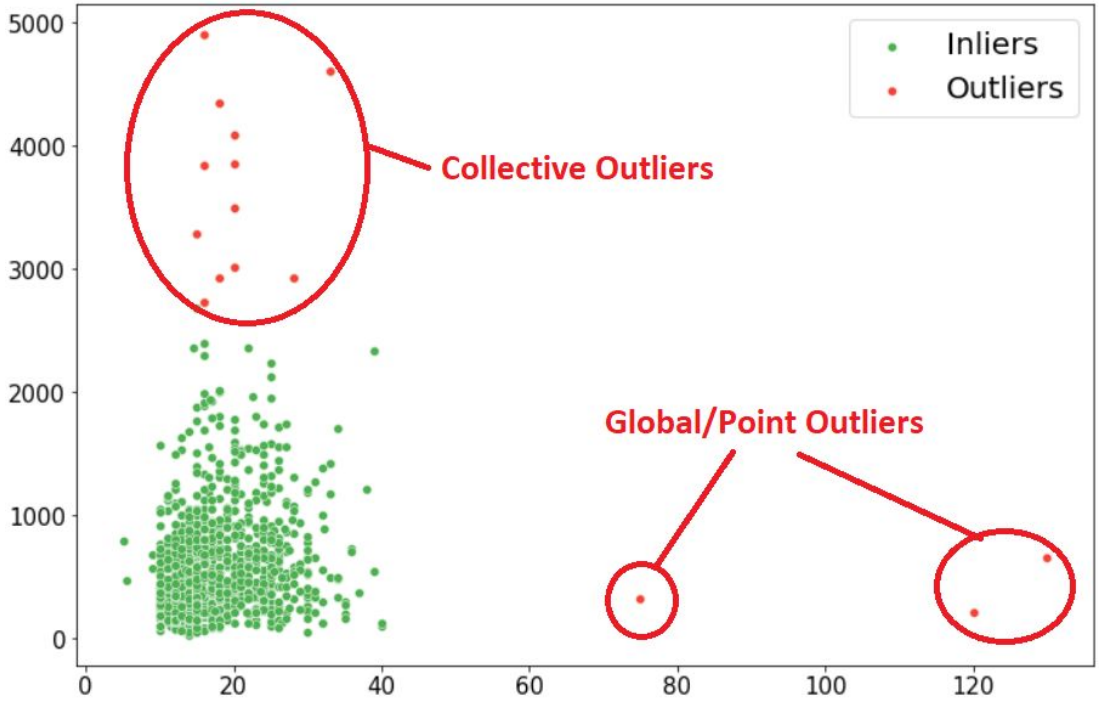
- **Clustering** → partitioning a data set into clusters of similar data
- **Classification** → Classification is used for prediction. In predictive modeling, a model is built to predict a value of a single variable based on the values of the other variables
- **Association rule** → find associations between the transactions of a customer
- **Outliers** → Identifying outlying data, that is, the data whose value is way outside or away from other data values

Classification vs Clustering





Outliers


Inliers
Outliers
Collective Outliers
Global/Point Outliers
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
120

Threats

- Clustering, classification, and association rule mining, generate an output that does not contain any customer data but generalized models
  - No threats to de-identification
- However, they should be protected in any case as:
  - They can be provided to third parties.
  - It is impossible to make assumptions on the background knowledge of an attacker
  - There are regulatory compliance needs

Association Rule Mining I

- Goal: find associations between the transactions of a customer
- Problem: find relationships among items in a database D
- let $I = \{i_1, i_2, ..., i_m\}$ be a set of items
- let $T = \{t_1, t_2, ..., t_n\}$ be a set of transactions on the database D where
  - $t_i \subseteq I$ i.e. t_i is a subsets of the available items
- A relationship is defined as:

$$X \rightarrow Y \text{ where } X \subseteq I,\ Y \subseteq I \text{ and } X \cap Y = \varnothing$$

- **Support**: the number of transactions containing X. Low support implies that the transaction randomly occurs → a minimum support (minSup) should be defined to prune rare transactions.
- **Confidence**: the percentage of transactions in T that contain X and that also contain Y . Low confidence implies that it is impossible to predict Y from X → minConf should be defined to remove weak associations.

Association Rule Mining II

Consider:

- I = {Bread, butter, eggs, cheese, ...}
- T = {t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8}
- X = {Bread, butter}
- Y = {Eggs}

We have that:

- Support = 6/8 (t_1, t_2, t_4, t_5, t_6, t_7)
- Confidence = 3/6 (t_1, t_4, t_7)

t_1	Bread, butter, eggs, cheese, chocolates
t_2	Chocolates, bread, butter, cheese
t_3	Eggs, flour, butter
t_4	Bread, butter, eggs
t_5	Bread, butter, cheese
t_6	Bread, butter, meat, beer
t_7	Bread, butter, eggs, milk
t_8	Eggs, flour, chocolates

Association Rule Mining III

Sample Transaction Database

Transaction ID	Bread		Butter	Eggs	Milk	Chocolate	Cheese	Flour	Beer	Meat
	i_1	i_2								i_m
t_1	1		1	1		1	1			
t_2	1		1	1		1		1		
t_3			1	1						
t_4	1		1	1						
t_5	1		1				1			
t_6	1		1						1	1
t_7	1		1	1	1			1		
t_8				1		1				

Association Rule Mining IV

- Transactions are recorded in a transaction database
- Companies have hundreds of products or items that determine the dimensions of the database
- A transaction database is therefore of high dimension and is sparsely filled with binary data
- Is there a privacy risk in sharing a transaction database? Definitely not.
- A privacy risk comes only when the transaction database is tagged or joined with customer identity data
- When a transaction table is associated with customer data, then the table becomes sensitive
- The challenges to privacy preservation are high dimensionality, no fixed schema, and Boolean data

Clustering I

- Data clustering is a method of creating groups of objects in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct
- Data clustering is also referred to as unsupervised learning in ML
- Clustering is exploratory in nature → there is no right or wrong approach
- Each cluster has a center point
- The goal of clustering is to find the intrinsic grouping of data for which a distance function is used

Clustering II

- Given m_i the mean of a group, the cluster is made by all the data that has an Euclidean distance less than a given threshold
- Consider that the mean of a group is denoted by m_i and the data in the group is denoted by x_i
- The distance between x_i and m_i in the Euclidean distance is:

$$\text{dist}(x_i, m_i) = ||x_i - m_i||$$

$$= \left(\sum (x_i - m_i)^2 \right)^{1/2}$$

Clustering III

- Data points similar to one another and also close to the mean come together to form a cluster
- This brings up some important aspects of a cluster such as cluster quality, which indicates that similar data points form a cluster and dissimilar points are in different groups of clusters
- One of the features of cluster quality is similarity (similarity function)
- Similar data points constitute a cluster, and dissimilar points are not in the same cluster
- A cluster has a center point, and other points in the cluster are close to it (distance measure) and the structure of the cluster

Clustering IV

- A cluster quality is controlled by the:
  - Similarity measure
  - Center
  - Distance measure
  - Structure
- These aspects of cluster quality are important when privacy preservation techniques are applied before clustering
- When an organization wants to carry out data mining activities such as clustering, they generally outsource the task to specialized analytics firms
- Outsourcing data has a major issue: data need to be protected before outsourcing