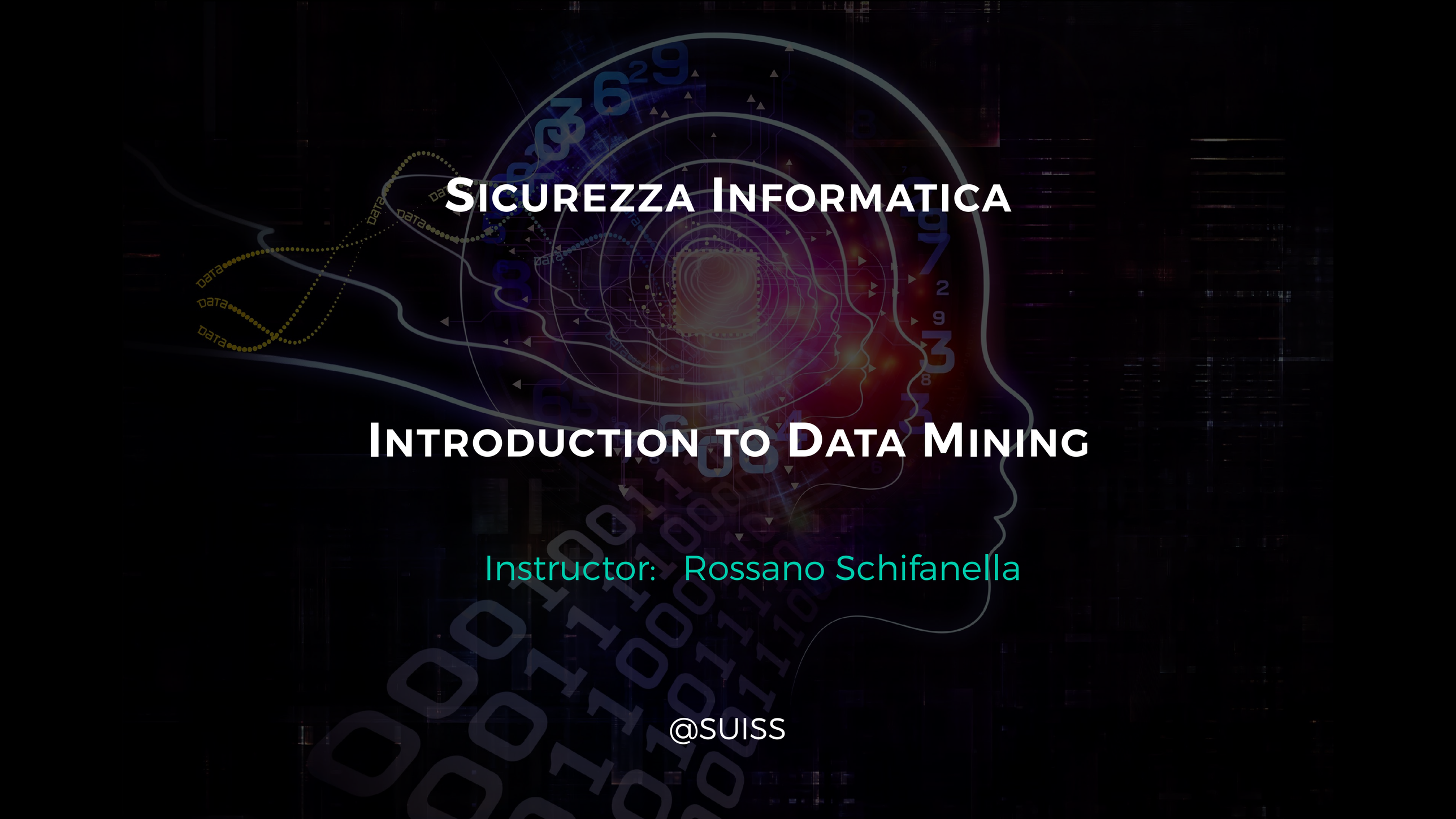


Sicurezza Informatica

Introduction to Data Mining

Instructor: Rossano Schifanella

@SUISS

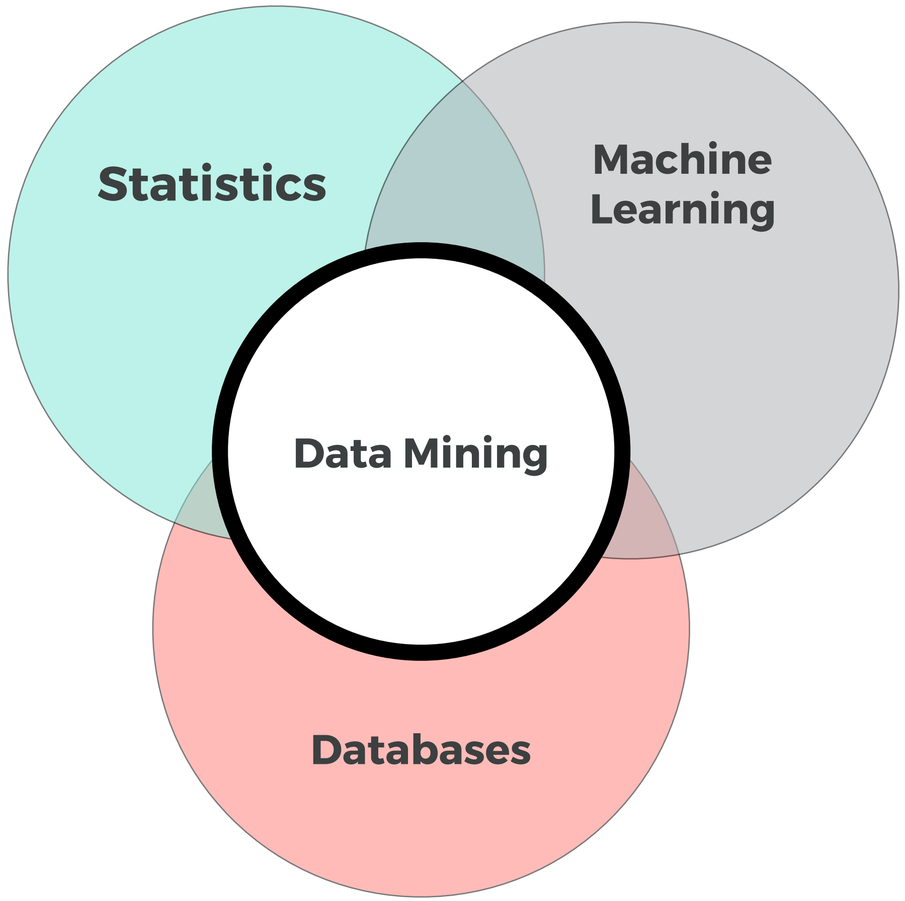
What is Data Mining?

Several Definitions

- **Non-trivial** extraction of **implicit**, **previously unknown** and potentially **useful** information from data.
- **Exploration** & **analysis**, by **automatic** or semi-automatic means, of **large quantities** of data in order to discover **meaningful patterns**.

Origins

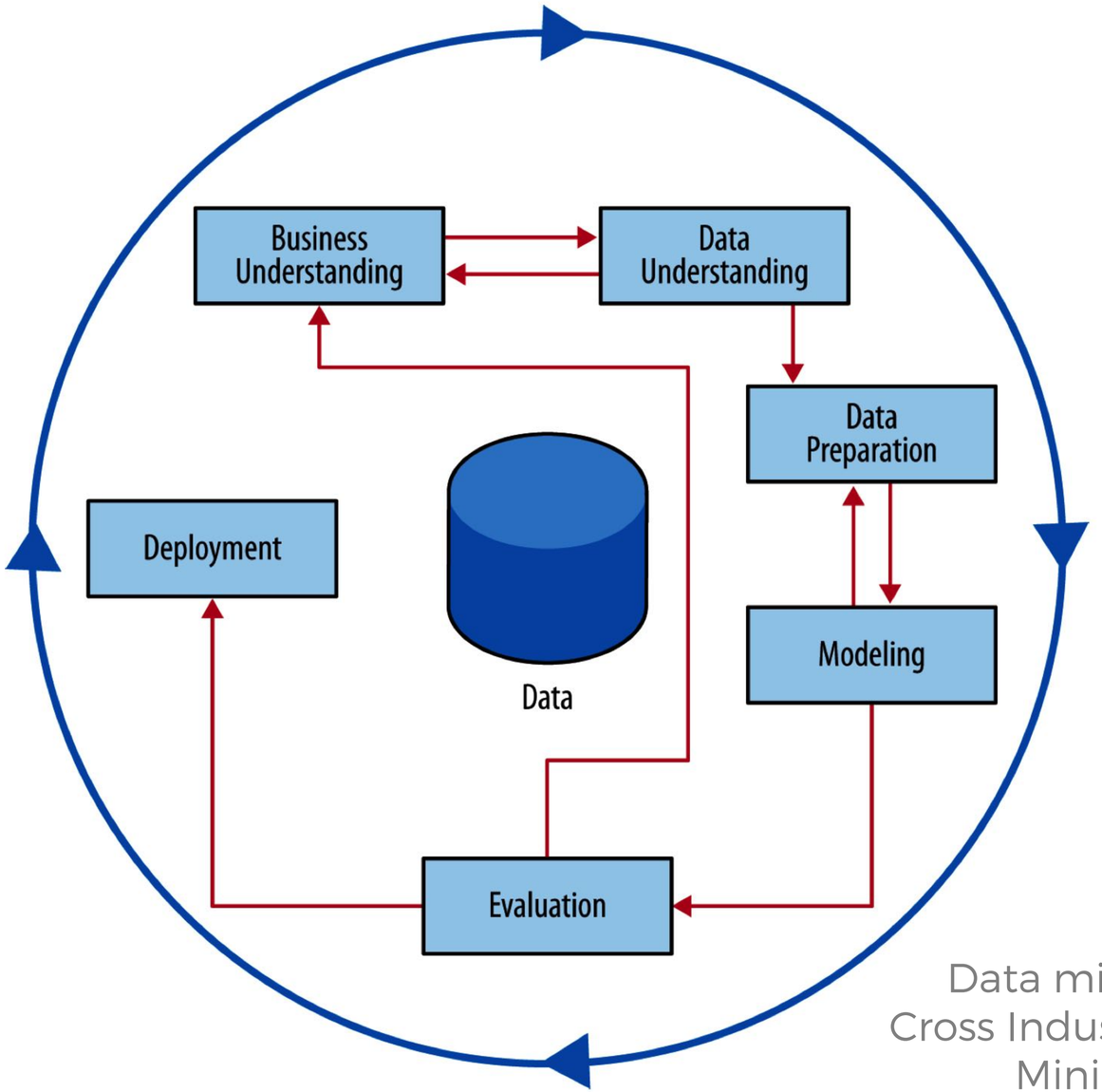
- Draws ideas from **machine learning**, **pattern recognition**, **statistics**, and **database systems**
- Traditional techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



Recap: What is Data Mining?

- Discover from data patterns and models that are:
  - **Valid**
    - hold on new data with some certainty
  - **Useful**
    - should be possible to act on the item and solve a problem
  - **Unexpected**
    - non-obvious to the system
  - **Understandable**
    - humans should be able to interpret the pattern

An alternative view



Data mining process according to the Cross Industry Standard Process for Data Mining (CRISP-DM; Shearer, 2000)

Business understanding and problem definition

The process starts always with the **definition of the problem** to solve

e.g., predicting customer churn: a telco company you work for has a major problem with customer retention in their wireless business.

Why Mine Data?

Scientific view point

- Data:
  - remote sensors on a satellite
  - telescopes scanning the skies
  - gene expression
  - scientific simulations generating terabytes of data
  - health and medical data
- Data mining may help scientists
  - in classifying and segmenting data
  - in hypothesis formation

Industrial view point

- Data:
  - web data, e-commerce
  - purchases at department/ grocery stores
  - bank/credit card transactions
- Data mining may help in
  - providing better, and customized services for competitive advantage

Example of applications

- **Marketing**
  - targeted marketing, online advertising, recommendations for cross-selling
- **Finance**
  - credit scoring and trading, fraud detection, workforce management
- **Retailers**
  - supply-chain management
- **Social media**
  - personalization, ranking

Example: Hurricane Frances

- **Why** data-driven prediction might be useful in this scenario?
- Amount of increase in **past similar events**
  - obvious (e.g., water)
  - not localized (e.g., a particular DVD)
- **Unusual** local demand
  - strawberry PopTarts, beer

Data Mining Tasks

- **Predictive**
  - Use some variables to predict unknown or future values of other variables.
- **Descriptive**
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

What Is Data?

48 k
apps are downloaded by Apple users

3.8 mln
Google searches are performed

149.513
emails are sent

400 hours
of new videos are uploaded to YouTube

3.3 mln
pieces of content are posted on Facebook

Report from IBM states that **90%** of the data in the world today has been created in the last two years alone

IDC says that worldwide revenues for big data and business analytics will grow **from $130.1 billion in 2016** to more than $203 billion in 2020

In 2000, only 738 million people used the internet, but by 2017, this number grew to **3,6 billion**

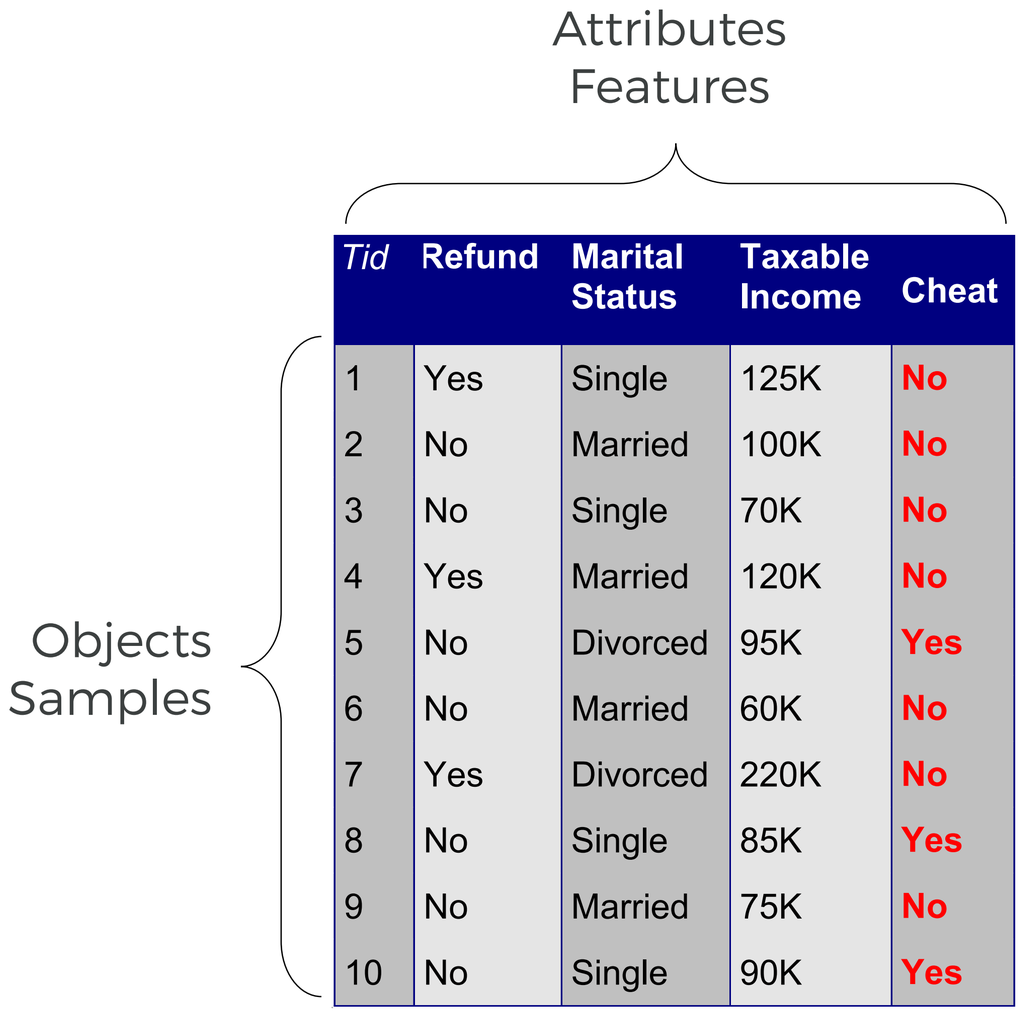
What is Data?

Collection of data objects and their attributes

An **attribute or feature or variable** is a property or characteristic of an **object or sample**

e.g., eye color of a person, temperature, etc.

A **collection of attributes** describe an object



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- **Attribute values are numbers or symbols assigned to an attribute**
- Distinction between **attributes** and **attribute values**
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Types of Attributes

Quantitative

Temperature

Qualitative

Taste

- **Discrete Attribute**
  - Finite or countably infinite values
  - zip codes, counts, or set of words in a collection of documents
- **Continuous Attribute**
  - Real numbers as values
  - height, weight

Types of Attributes

Categorical	mutual exclusive, not ordered, categories e.g., five different genotypes
Ordinal	order matters but not the difference e.g., movie ratings
Interval	difference between two values is meaningful e.g., temperatures in Celsius, a temperature of 100 degrees C is not twice as hot as 50 degrees C
Ratio	as interval but has a clear definition of 0.0 e.g., temperature in Kelvin,

Types of Attributes

	Nominal	Ordinal	Interval	Ratio
frequency distribution.	Yes	Yes	Yes	Yes
median and percentiles.	No	Yes	Yes	Yes
add or subtract.	No	No	Yes	Yes
mean, standard deviation, standard error of the mean.	No	No	Yes	Yes
ratio, or coefficient of variation.	No	No	No	Yes

Examples

Age in years [discrete, quantitative, ratio]

Number of patients in a hospital [discrete, quantitative, ratio]

Brightness measured by a light meter [continuous, quantitative, ratio]

Brightness measured by people's judgments [discrete, qualitative, ordinal]

ISBN numbers for books [discrete, qualitative, nominal]

Types of data sets

Record

- Data Matrix
- Document Data
- Transaction Data

Graph

- World Wide Web
- Molecular Structures

Ordered

- Spatial Data
- Temporal Data
- Sequential Data
- Genetic Sequence Data

Characteristics of Structured Data

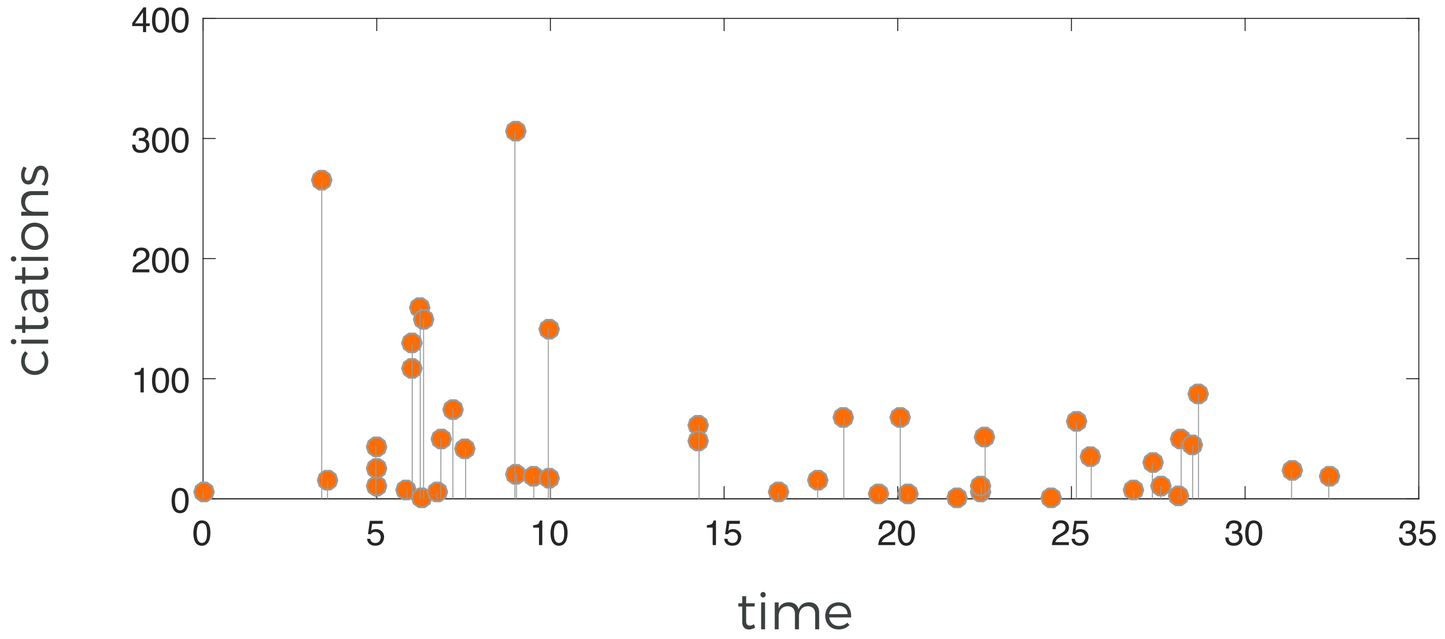
- **Dimensionality**
  - Curse of dimensionality
- **Sparsity**
  - Only presence counts
- **Resolution**
  - Patterns depend on the scale

Data Quality

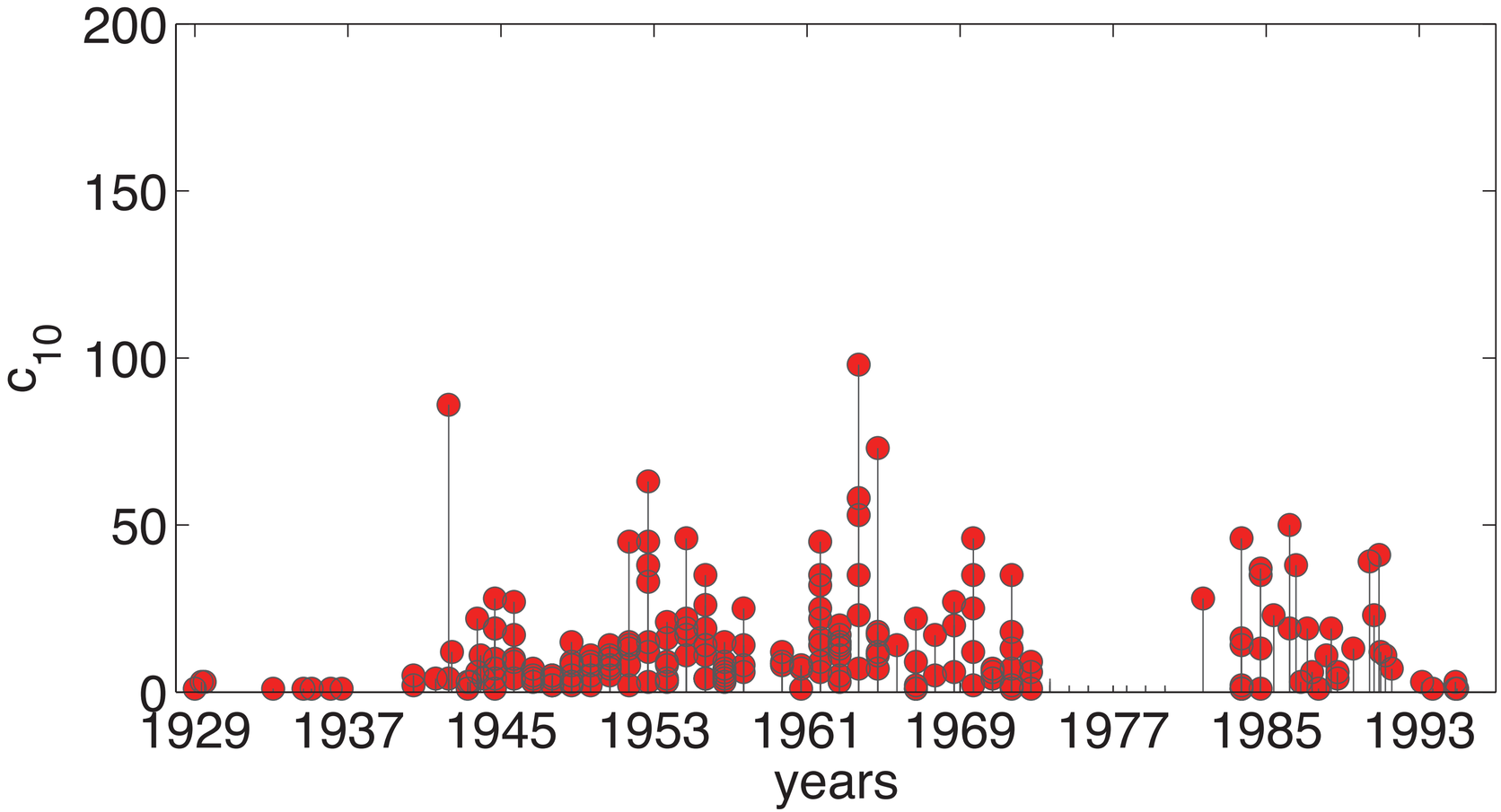
- Big data not always means high quality data
  - presence of inconsistencies and errors

Checking the Data

WEB OF SCIENCE™

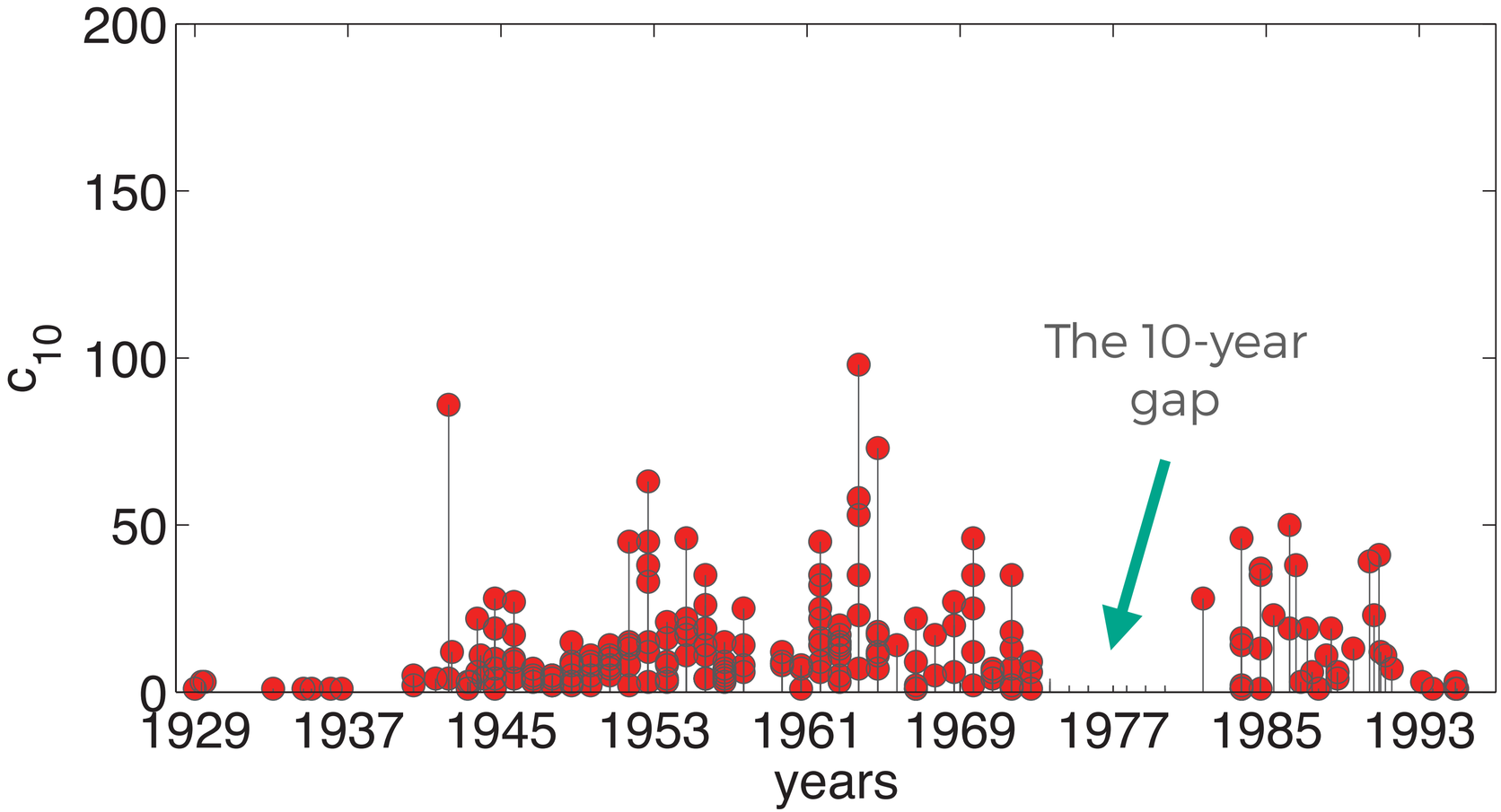


Checking the Data



Subrahmanyan Chandrasekhar

Checking the Data



Subrahmanyan Chandrasekhar

Checking the Data

True?

Biographical information

Looking at raw data

Manual inspection

Cross checking

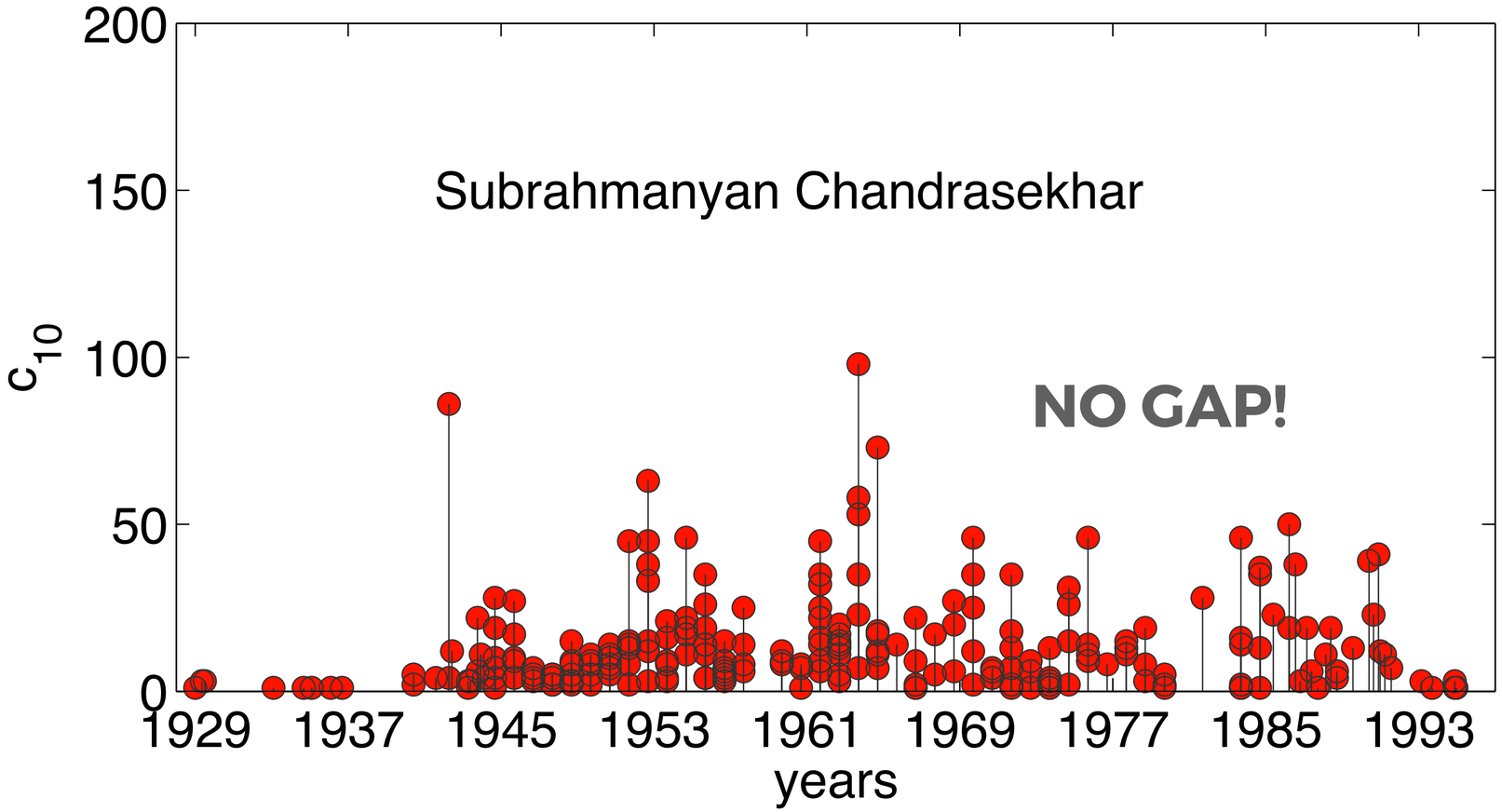
Checking the Data

True?

Solution:

for 10 years the name Chandrasekhar was abbreviated in Chandras.

Checking the Data



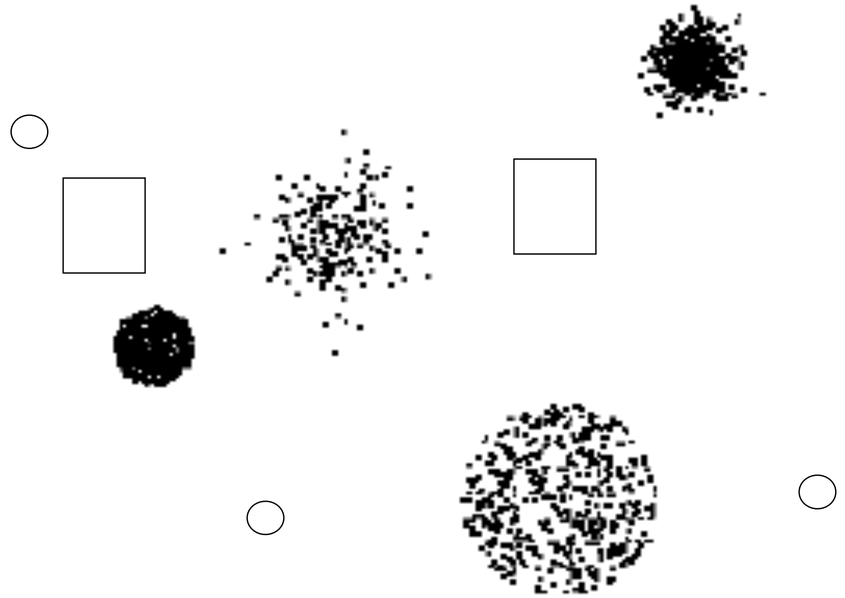
Subrahmanyan Chandrasekhar

Data Quality

- Big data not always means high quality data
  - presence of inconsistencies and errors
- Examples of data quality problems:
  - **Inconsistent labelling**
  - **Noise and outliers**
  - **Missing values**
  - **Duplicate data**

Outliers

Data objects with characteristics that are **considerably different** than most of the other data objects in the data set



Missing Values

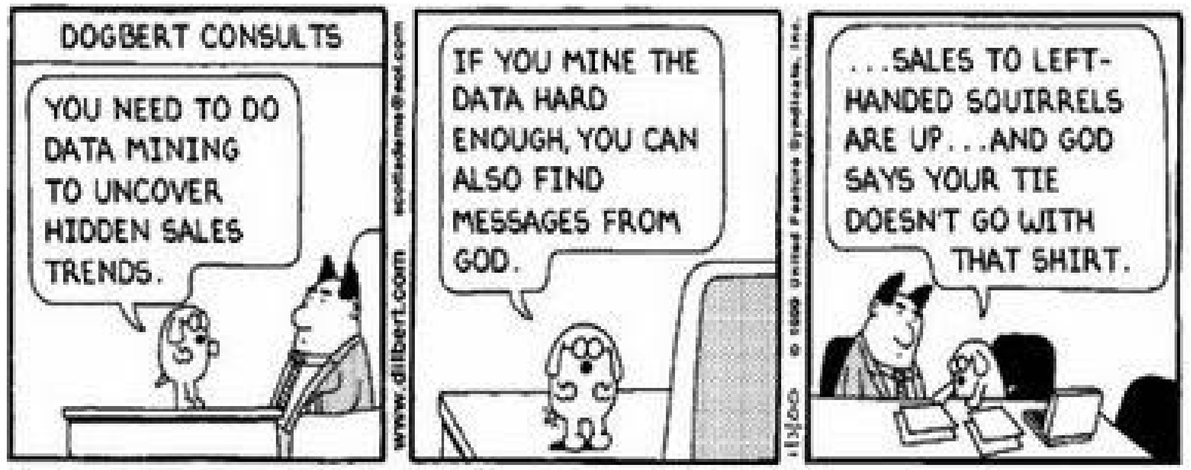
- **Reasons for missing values**
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- **Handling missing values**
  - Eliminate Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogenous sources
- Examples:
  - Same person with multiple email addresses

Meaningfulness of Analytic Answers

- A risk with Data mining is that an analyst can discover patterns that are meaningless
- Statisticians call it **Bonferroni's principle**:
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap



Whenever you work with data ask...

could this occur “randomly”?

are there confounding factors?

Approaches to learning

- **supervised**
- **unsupervised**
- **semi-supervised**
- **reinforcement learning**

Tasks

- **Predictive**
  - Use some variables to predict unknown or future values of other variables.
- **Descriptive**
  - Find human-interpretable patterns that describe the data.

Tasks

- **Classification** [Predictive]
- **Regression** [Predictive]
- **Anomaly Detection** [Predictive]
- **Clustering** [Descriptive]
- **Association Rule** [Descriptive]
- **Sequential Pattern Mining** [Descriptive]

Classification

A classifier is a mapping:

$$\hat{c} : \mathscr{X} \rightarrow \mathscr{C}$$

where

$$\mathscr{C} = \{C_1, C_2, \ldots, C_k\}$$

the “hat” over the name of the classifier denotes that the classifier is an approximation of the true but unknown function **c**.

An example is a pair:

$$(x, c(x)) \in \mathscr{X} \times \mathscr{C}$$

where **x** is an “instance” and **c(x)** is the true class of the instance (possibly contaminated by noise).

Classification

Learning a classifier involves constructing the function such that it matches **c** as closely as possible.

Example: Email Spam Detection

Scoring classifier

A scoring classifier is a mapping:

$$\hat{\mathbf{s}} : \mathscr{X} \to \mathbb{R}^k$$

the boldface notation denotes that the output is a vector, i.e.:

$$\hat{\mathbf{s}}(x) = (\hat{s}_1(x), \ldots, \hat{s}_k(x))$$

where the i-th component is the score assigned to class $\mathbf{C_i}$ for instance **x**. If we only have two classes, it usually suffices to consider the score for only one of the classes.

Here the scores are a measure of the **confidence** the classifier has in its prediction.

Example: Advertisement

Class probability estimation

A class probability estimator is a scoring classifier that outputs probability vectors over classes, i.e., a mapping:

$$\hat{\mathbf{p}} : \mathscr{X} \to [0,1]^k$$

We write:

$$\hat{\mathbf{p}}(x) = \left(\hat{p}_1(x), \ldots, \hat{p}_k(x)\right)$$

where the i-th component is the probability assigned to class $\mathbf{c_i}$ and $\sum_{i=1}^{k} \hat{p}_i(x) = 1$

If we have only two classes, then $\hat{p}(x)$ denotes the estimated probability for the positive class.

Example: Fraud Detection

Regression

A regressor is a mapping

$$\hat{f} : \mathscr{X} \rightarrow \mathbb{R}$$

The regression learning problem is to learn a function estimator from examples $\mathbf{(x_i, f(x_i))}$.

Note that we switched from a low-resolution target variable to one with infinite resolution. It's highly likely that some part of the target values in the examples is due to fluctuations that the model is not able to capture.

Since often the examples are noisy, it is reasonable to assume that the estimator is only intended to capture the general trend of the function.

Example: Stock Price

Association Rules

Let $\mathbf{I=\{i_1, \ldots, i_n\}}$ be a set of binary attributes called **items**.

Let $\mathbf{D=\{t_1, \ldots, t_m\}}$ be a set of **transactions** called the database.

Each transaction in D contains a subset of the items in I.

A rule is defined as an implication of the form:

$$\mathbf{X \Rightarrow Y}$$

where $\mathbf{X,Y \subseteq I}$.

Example: Supermarket

Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

{butter, break} ⇒ {milk}

Sequential Pattern Mining

Find **statistically relevant patterns** between data examples where the values are delivered in a **sequence**.

Usually with values that are discrete

e.g., strings

Example: DNA Sequences

Models

- Models can be distinguished according to their approach:
- **Geometric**
  - models use intuitions from geometry such as separating (hyper-)planes, linear transformations and distance metrics.
- **Probabilistic**
  - models view learning as a process of reducing uncertainty, modelled by means of probability distributions.
- **Logical**
  - models are defined in terms of easily interpretable logical expressions.

Questions?

@rschifan

schifane@di.unito.it

http://www.di.unito.it/~schifane