Introduction to Data Analytics
School on Scientific Data Analytics and Visualization

Roberta Turra, Cineca

8 June 2015
Data analytics

the process of extracting useful insights from raw data

Same as … **Data Mining** (also known as Knowledge Discovery in Databases - KDD): the process of discovering valuable information from very large databases using algorithms that discover hidden patterns in data (1995)
Why is it challenging

**Volume**
- Data at Rest
  - Terabytes to exabytes of existing data to process

**Velocity**
- Data in Motion
  - Streaming data, requiring milliseconds to respond

**Variety**
- Data in Many Forms
  - Structured, unstructured, text, multimedia,…
Going back to the definition ...

the **process** of extracting valuable information from raw **data** using **algorithms** that discover hidden patterns

It’s an **explorative approach** or **data driven approach** in contrast with “traditional” data analysis (statistics) that could also be hypothesis driven.
Topics

Data analytics

- data
- process
  - pre-processing
- algorithms / techniques

- applications
Data

The number and rate of data produced in any particular discipline now exceed our ability to effectively treat and analyse them.

Sources:
- digital instruments
- high resolution cameras
- medical scanners
- simulations
- transactional data
- social media
- …
Data typologies

- **structured data**
  - data matrix
  - transactional data

- **graph**
  - web and social networks
  - molecular structures

- **ordinal data**
- **spatial data**
- **time series**
- **sequences**
  - genetic sequences
- **unstructured data**
  - textual documents
  - images
  - audio and videos (multimodal)
The Data Mining Process

- **Selection**
  - Target data
  - Cleaned data
  - Transformed data
  - Patterns/models
  - Knowledge
- **Preprocessing and cleaning**
- **Transformation and reduction**
- **Data Mining**
- **Visualization/Evaluation**

**Database/Data Warehouse**
CRISP-DM reference model
Cross Industry Standard Process for Data Mining
Is it still the reference model? (1)

New challenges

 далее. Новое. Модель CRISP отражает перспективу управления данными, где все важные данные можно хранить и очищать до того, как начать дальнейшую обработку. В этом предположении легко нарушить в тех случаях, когда поток данных слишком обширен для полного хранения (фильтрации/сжатия данных перед тем, как это можно будет сделать, потребуется некоторое понимание ожидаемого анализа) или когда имеются ограничения по времени.

 далее. Модель CRISP предлагает плоский подход. Мастерство данных разнообразия и сложности требует нескольких уровней анализа, объединяя результаты различных инструментов обработки, чтобы получить сложные модели или структуры, которые могут быть использованы для создания иерархической зависимости между шагами проведенной работы.
Is it still the reference model? (2)

New challenges

- In complex applications, the design of an analytical process is actually a multi-disciplinary effort that involves actors with different backgrounds.

- The computational complexity requires new scalable algorithms and the distribution of workloads on clusters (e.g., MapReduce) or on cloud.

- Big Data Analytics often involve the use of personal data, ranging from medical records to location information, activity records on social networks, web navigation and searching history, etc. All this calls for mechanism that ensure that the information flow employed in the analyses does not harm the privacy of individuals.
Is it still the reference model? (3)

New challenges

- **Data integration** from multiple and heterogeneous sources.
- **Data quality**.
- Models **fast adapting** to temporal changes.

New emphasis on

- **Re-purposing data** that was collected for a different purpose.
- **Re-purposing algorithms** (e.g. page rank on graphs).
- **Data products**: data driven applications (e.g. spell checkers, machine translation, recommendation systems, …) interactive visualizations, online databases.

Not just answering the question once, empower others to use data in new ways
Is it still the reference model? (4)

e-Science

4th paradigm of scientific inquiry:
to acquire massive data sets from instruments or from simulations

- e-Science is **driven by data** more than by the computation
- **Data analysis** has replaced data acquisition as the new bottleneck to discovery
Another way of describing the process (BDVA) data analysis output can be input for other higher level analysis.
Pre-processing

- data understanding and data quality assessment
  (evaluation of data accuracy and reliability, completeness, consistence, … correlation)
  - Presence of missing values, outliers, inconsistencies
  - Level of noise
  - Redundance

- data preparation
  - Cleaning
  - Transformation (normalization, discretization, aggregation, new variables computation…)
  - Feature extraction
  - Selection / filtering
Pre-processing

Why is it useful - a few examples

- L’Equité: high peak of 96 years old insured
  - missing birth dates had been codified 1/1/1900
- Trento University: a high number of students with very low grades in the high school diplomas
  - grades in the high school diplomas have undergone a scale change (from 60 as a maximum to 100)
- Local Health Service: high consumption of cardiovascular drugs in diabetics
  - the quantity of active ingredient for cardiovascular drugs was in milligrams (instead of grams)
- Eurostat: visual patterns of outliers
  - the Country was a key variable in international trade outliers identification
Pre-processing
Ask the right question
Data representation
Analysis matrix

<table>
<thead>
<tr>
<th></th>
<th>$X_{11}$</th>
<th>$X_{12}$</th>
<th>$X_{13}$</th>
<th>...</th>
<th>$X_{1d}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{21}$</td>
<td>$X_{22}$</td>
<td>$X_{23}$</td>
<td>...</td>
<td>$X_{2d}$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$X_{n1}$</td>
<td>$X_{n2}$</td>
<td>$X_{n3}$</td>
<td>...</td>
<td>$X_{nd}$</td>
<td></td>
</tr>
</tbody>
</table>

variable

observation
Coal: data structure

Association analysis

Customer segmentation
Coal: customer segmentation matrix

- variables describing the buyer behavior:
  - items list (only the characterizing, distinguishing items)
  - number of receipts
  - average number of items per receipt
  - average expense
  - percentage of items having a promotion

- socio-demographic variables:
  - genre
  - age
  - job
  - marital status
  - number of sons
  - number of children
  - cats
  - dogs

“active” variables

“descriptive” variables
Tasks and techniques

descriptive

- clustering
  - k-means
  - relational analysis
  - Self Organizing Maps
  - ...
- association rules
- sequential patterns
- graph and network analysis
  - ...

predictive

- classification (machine learning)
  - Naive Bayes
  - Decision Trees
  - Neural Networks
  - KNN
  - Rocchio
  - Support Vectors Machine
  - ...
- regression

Unsupervised learning
training samples have no class information
guess classes or clusters in the data

Supervised learning
use training samples with known classes
to classify new data
Terminology

- **Supervised learning** ("Training")
  - we are given examples of inputs and associated outputs
  - we learn the relationship between them

- **Unsupervised learning** (sometimes "Mining")
  - we are given inputs but no outputs
    - unlabeled data
  - we learn the "latent" labels
    (e.g. clustering, dimensionality reduction)
Tasks, techniques and applications

**Descriptive**
- clustering
  - k-means
  - relational analysis
  - Self Organizing Maps
  -...
- association rules
- sequential patterns
- graph and network analysis

**Predictive**
- classification (machine learning)
  - Naive Bayes
  - Decision Trees
  - Neural Networks
  - KNN
  - Rocchio
  - Support Vectors Machine
  -...
- regression

- Churn analysis
- Fraud detection
- Prospect identification
- Recommendation systems
- Document classification
  -...