Data Mining

Università di Pisa

309AA

Shifting focus

- Optimization methods: minimize a function
- Machine learning: learn and assess models
- Algorithms: solve a well-defined problem

In data mining the focus is **the data itself**! We wish to analyze it and understand it.

Task VS Data focus

Università di Pisa

It's January 2020, and you are analyzing health records, e.g., hospital reports, data from Pisa, Frankfurt, and Wuhan.

Task-focus

- Predict discharge date
- Predict mortality
- ...

vs

Data-focus

- Find patients with atypical symptoms
- Find patterns in delayed care
- Find typical patient profiles
- ...

Task VS Data focus

UNIVERSITÀ DI PISA

It's January 2020, and you are analyzing health records, e.g., hospital reports, data from Pisa, Frankfurt, and Wuhan.

Task-focus

- Problem-centric
- Artifact-centric

vs

Data-focus

- Information-centric
- Human-centric

Data and information focus

Università di Pisa

In data mining, the goal is to extract (**human-readable**) **knowledge and insight** from **raw data**.

- Knowledge implies we are often not *just* trying to solve a task
- Insight implies that we should infer *non-obvious* knowledge
- Human-readable implies that knowledge should be (when possible) understood by humans: focus on *interpretability*!
- Raw data implies we'll need to clean it

Data Mining

Università di Pisa

Data Mining
Discipline that studies the efficient extraction and analysis of information and patterns in large data collections, finally inducing information from data.

Large data collections

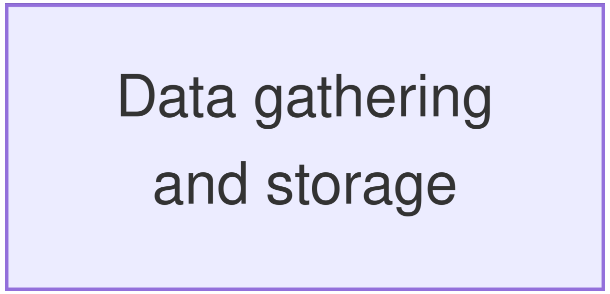
Large collections tend to be heterogeneous in...

- Source, i.e., they often gather different data sources, e.g., data from different labs, e-commerce websites, different cities/states, etc.
- Domain: scientific data, transactional data (e-commerce), traffic data, social networks, sensor data, etc.
- Language: different conventions, scales, encodings, etc.
- Refinement: often data is raw, unprocessed, or noisy

These separate data collections from datasets!

Large data collections

Before we even think of analyzing and extract patterns from such data, we must store it. The first step of Data Mining is data gathering, storage and warehousing.

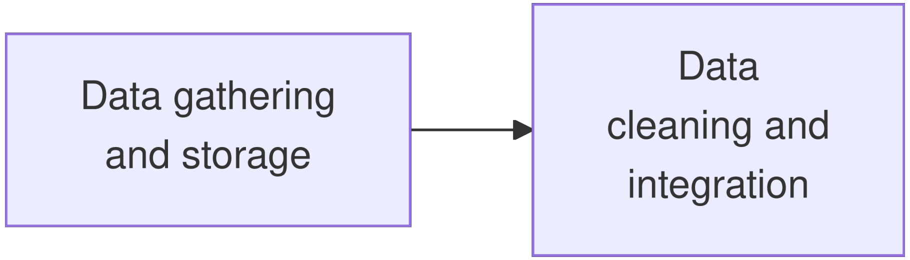


The knowledge discovery pipeline, step 1.

Large data collections

Università di Pisa

Simple storage does not tackle the source heterogeneity, thus we need to properly clean and integrate data. This tackles heterogeneity in language and refinement.



The knowledge discovery pipeline, step 1 and 2.

Large data collections... after data gathering, cleaning and integration

- Sources are integrated
- Language is homogeneous: same conventions, scales, and encodings
- Refinement: data is cleared of noise and outliers, and can be analyzed

Information and patterns... for what?

Data Mining
Discipline that studies the efficient extraction and analysis of information and patterns in large data collections, finally inducing information from data.

Information... for what? For whom?

Information as insight

Look to answer **questions on the data as a stakeholder**. Say it's January 2020, and you are analyzing health records, e.g., hospital reports, data from Pisa, Frankfurt, and Wuhan. What might you ask of the data?

Information as insight

Università di Pisa

- Are there some common patterns in the data? You find a shared influx of new patient with respiratory diseases
- Are there some anomalies? A small set of such patients does not exhibit any common predisposing conditions
- Are there data groups with different behaviors? A group responds well to known treatment, another does not, another worsens

From insight to action

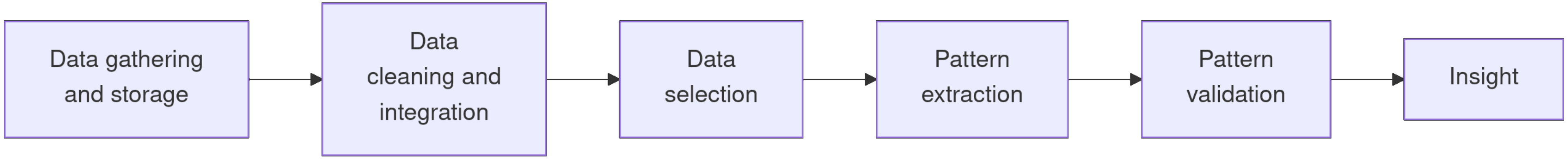
Insight allows a human to **make decisions**, e.g.,

- Are there some common patterns in the data? Then maybe my heterogeneous sources are observing a common phenomenon: study said phenomenon
- Are there some anomalies? Then maybe there is a problem with my data, or I've found something new: check my data sources
- Are there data groups with different behaviors? Then I may want to study them separately

Information as insight

UNIVERSITÀ DI PISA

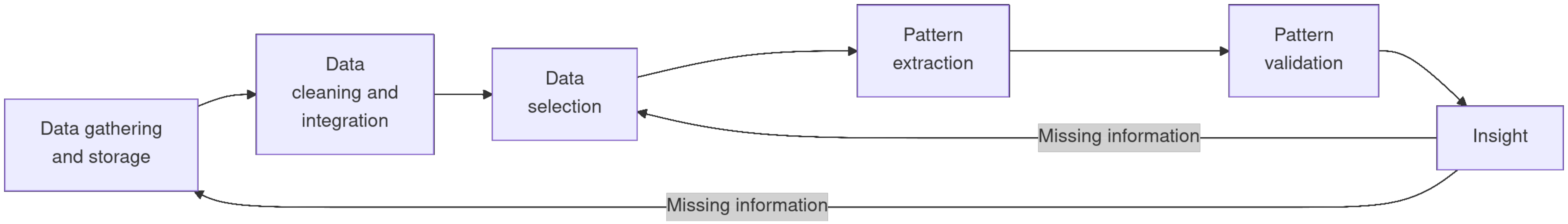
Information is extracted from filtered data from which patterns are extracted. Not all patterns are equally useful, thus a pattern evaluation step is required.



The knowledge discovery pipeline.

Information as insight

Missing data may lead us to go back to gathering.



The knowledge discovery loop.

Thought exercise

Università di Pisa

You are given a cycling data collection, with data gathered from different sources, covering all tours of thousands of cyclists from 2018 to 2024.

Sources

- A social network for competitive non-professional cyclists
- A training platform for professional cyclists
- A social network for non-competitive, non-professional cyclist that cycle to explore nature

Features

- Speed
- Cadence
- Bike used
- Track, e.g., length, elevation, climbs
- Info on the cyclist, e.g., age

Steps 1 and 2: data cleaning and integration

In data cleaning and integration, we look to find...

Looking for...	Action
Missing values	Impute them, or drop the feature
Out of range values	Standardize the sources, or drop them
Different data scales	Standardize them
Non-informative or redundant features	Drop them
Data semantics	Understand distributions and general patterns

Steps 1 and 2: data cleaning and integration

Università di Pisa

Looking for...	Action	Insight?
Missing values	Analyze missing values	Malfunctioning sensors
Out of range values	Standardize the sources	Professionists are much faster, but e-bikes exists
Different data scales	Miles and kilometers conversion, and adjustment for different bikes	Bikes do not really impact performance as much
Non-informative or redundant features	Drop them	
Data semantics	Understand distributions and general patterns	Little improvement over time for amateurs

Steps 3: data selection and transformation

Università di Pisa

Not all data is useful to extract any patterns, and must be processed accordingly.

Looking to...	Action
Find anomalous data	Remove from the analysis, or analyze separately
Aggregate data	Extract higher-level patterns
Generate novel features	Study specific phenomena

Steps 3: data selection and transformation

Not all data is useful to extract any patterns, and must be processed accordingly.

Looking to...	Action	Insight
Find anomalous data	Run an anomaly detection algorithm	Exceptional cyclists follow a steeper improvement curve
Aggregate data	Group by occupation and source	Non-professional cyclists reach a performance plateau much later
Generate novel features	Create a climb difficulty index	Tracks with lots of continuous climbing reduce performance between male and female cyclists

Steps 4 and 5: pattern extraction and evaluation

Università di Pisa

What patterns are we trying to extract?

Looking to extract...	Patterns
Profiles of cyclists	Clusters
Descriptive rules	Rules

Steps 4 and 5: pattern extraction and evaluation

Clusters

- Profile 1: Cyclists very good in flat terrain
- Profile 2: Cyclists very good in mountainous terrain
- Profile 3: Cyclists jack of all trades, not excelling in anything in particular

Steps 4 and 5: pattern extraction and evaluation

Università di Pisa

Rules

```
Slim and short cyclists -> Good on mountains
Heavy, burly cyclists -> Good on flat terrains
```

Summing up: data mining tasks

Task	Goal
Data understanding	Understand data behavior
Data transformation	Cleaning and enriching data
Outlier detection	Find anomalous data
Rule mining	Finding rule-like patterns
Clustering	Find profiles and groups within data
Modeling	Predict on future data