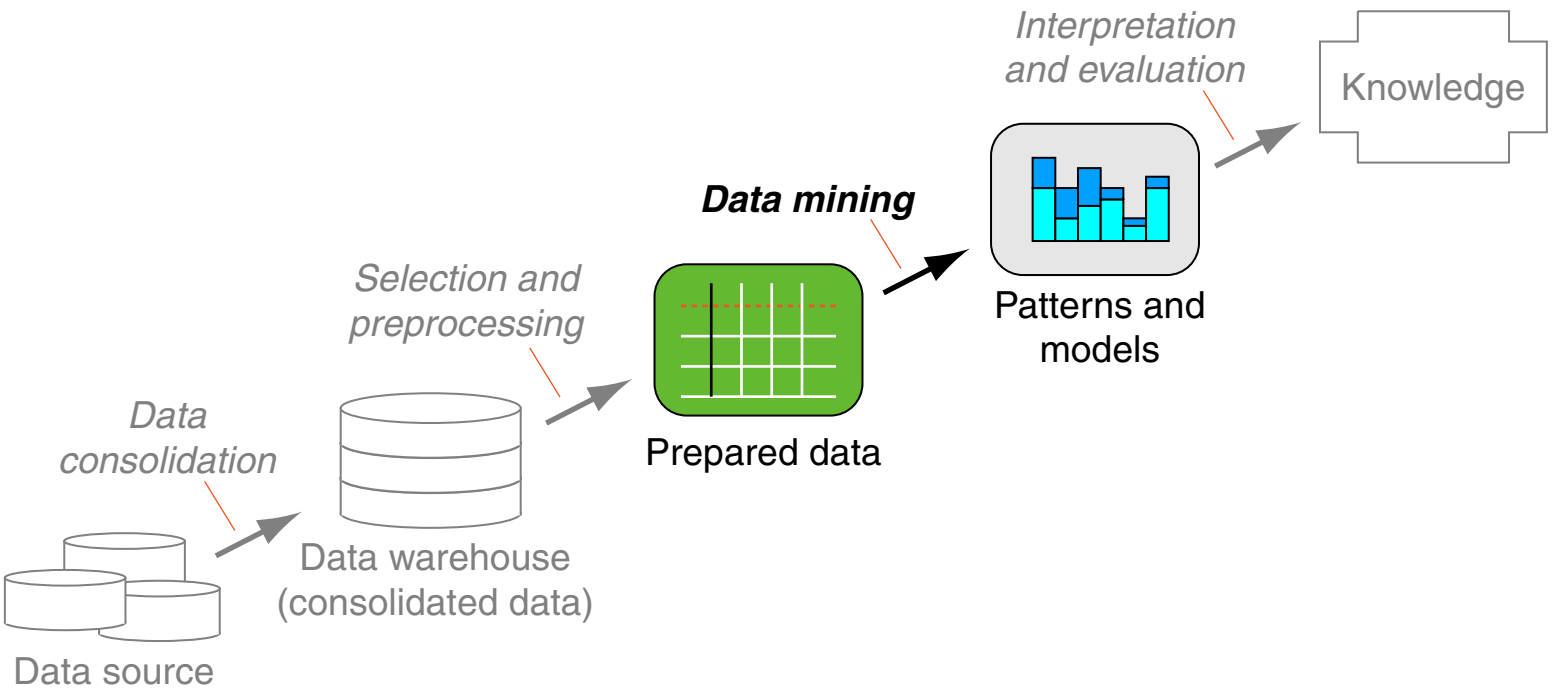
Chapter DM:I

Data Mining Overview

Definition 1 (Knowledge Discovery in Databases, KDD [Fayyad 1996, Wrobel 1998]**)**

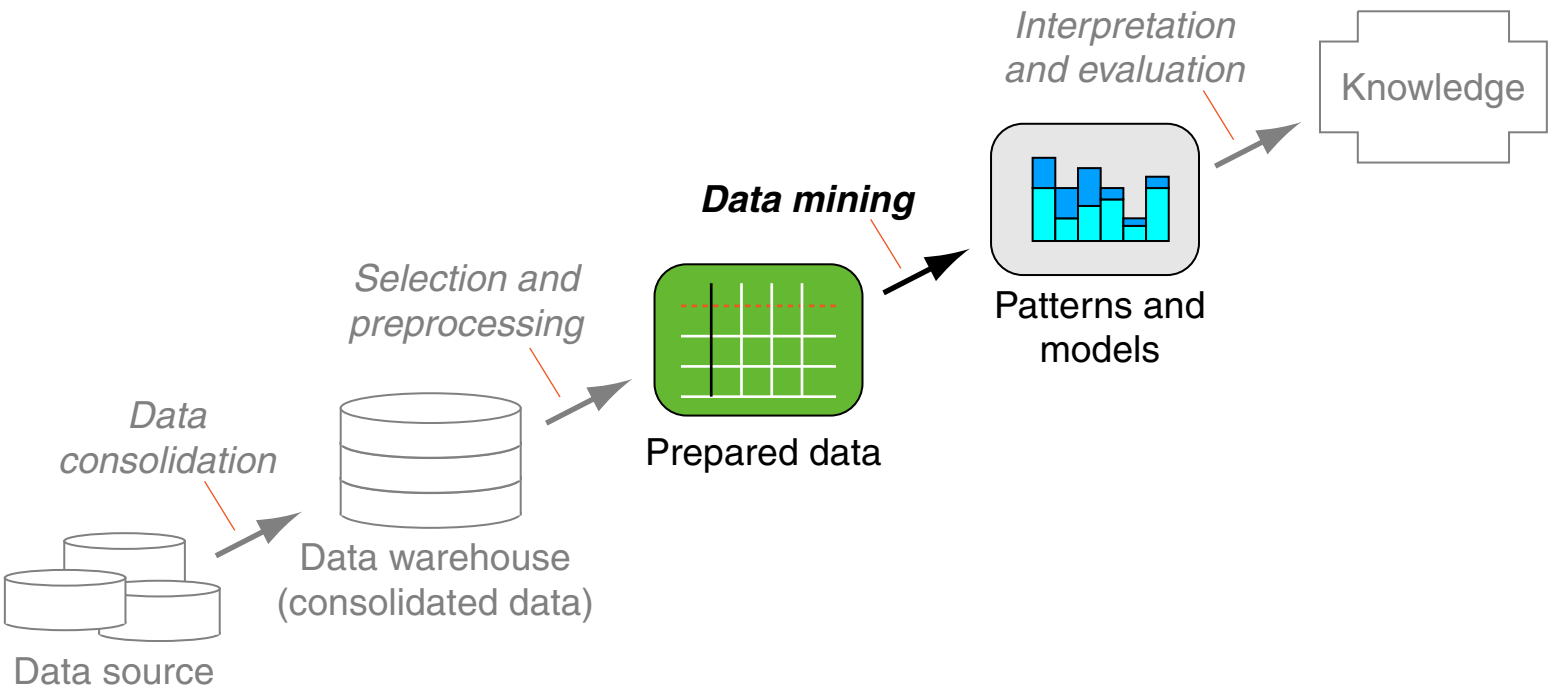
Knowledge Discovery in Databases is the process of identifying valid, new, relevant, and interpretable patterns in huge data sets.



Data Mining Overview

Definition 1 (Knowledge Discovery in Databases, KDD [Fayyad 1996, Wrobel 1998])

Knowledge Discovery in Databases is the process of identifying valid, new, relevant, and interpretable patterns in huge data sets.



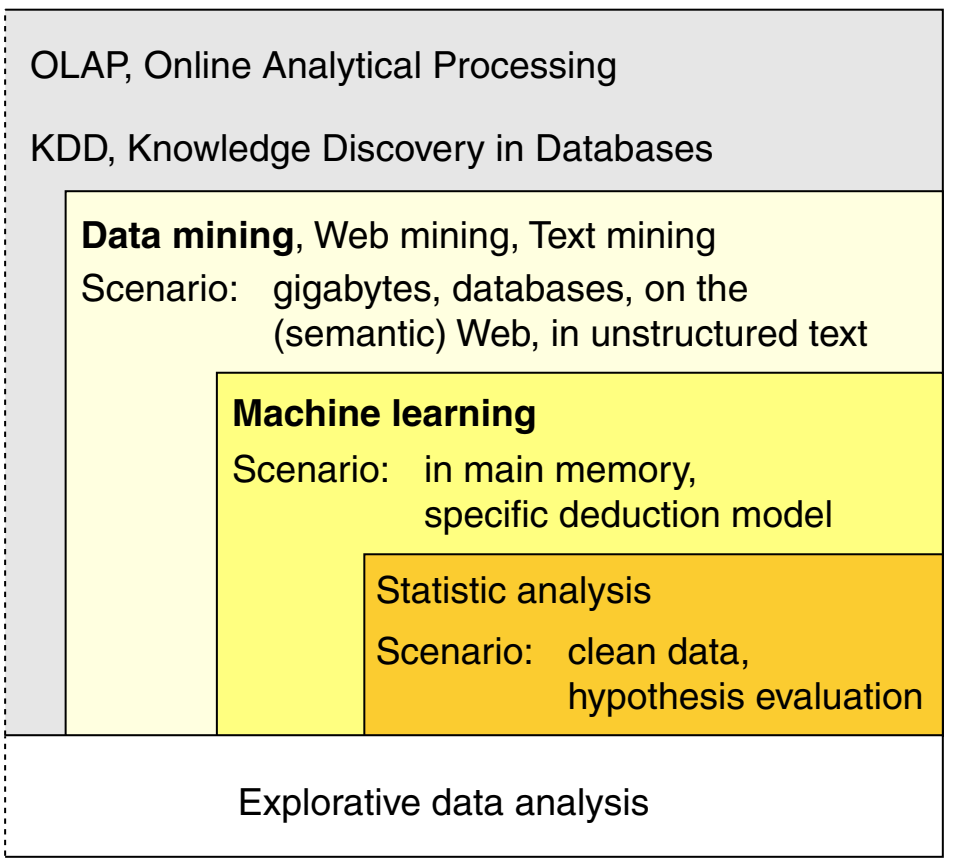
Definition 2 (Data Mining)

Data mining is the systematic, usually automated or semi-automated discovery and extraction of so far unknown relations from huge data sets.

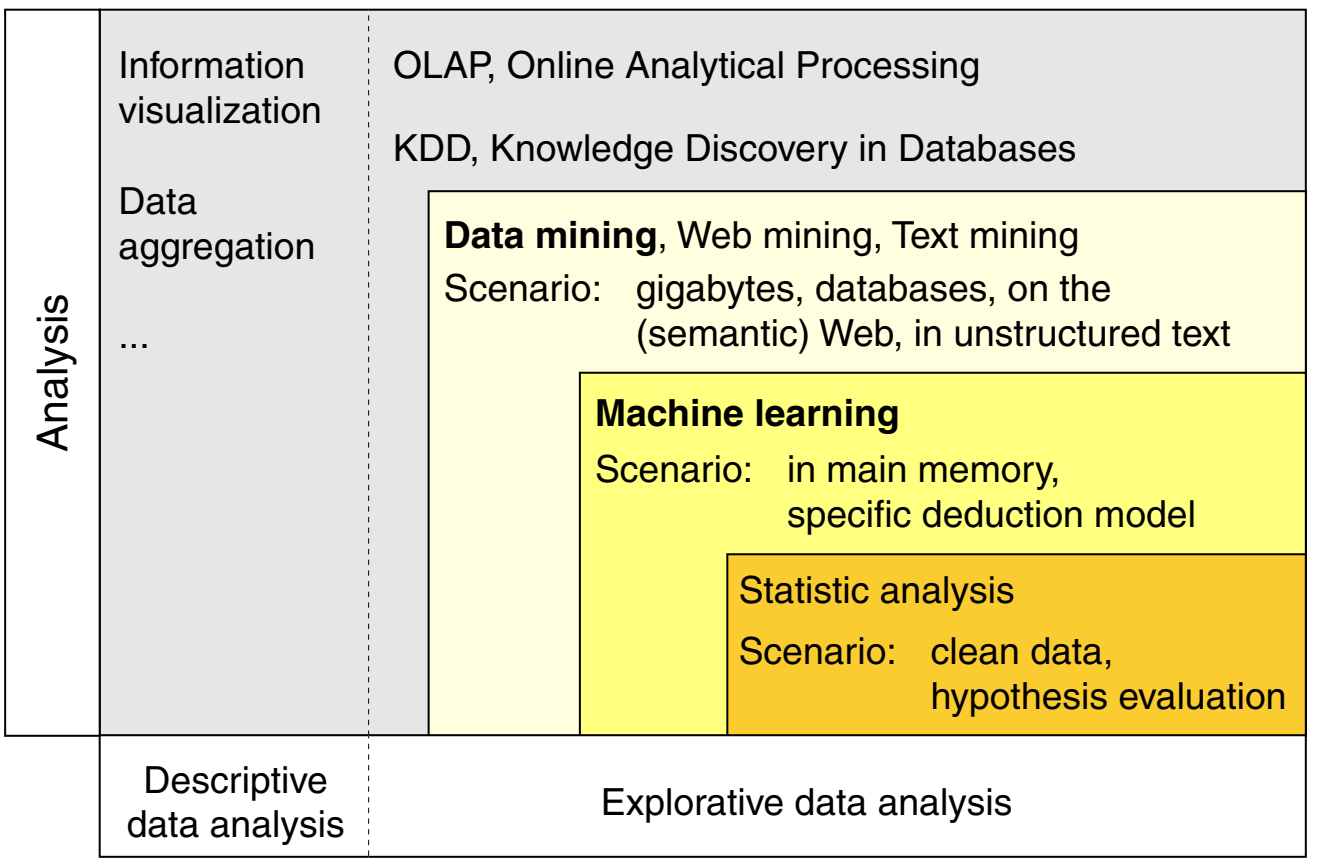
Remarks:

- Data mining technology belongs to the field of *explorative data analysis*. Explorative data analysis deals with both data presentation and search for structures, peculiarities, and anomalies. It is applied if the research question is fuzzy or if the choice of the statistical model is unclear.
- The data mining definition does not use the notion of “information”: under the viewpoint of semiotics, data mining operates on the sigmatic layer only.
  The *interpretation* of discovered patterns, i.e., the examination of information with regard to new findings and a subjective knowledge gain, which happens on the pragmatic layer, belongs to the field of KDD.
- In the business world, the terms data mining and knowledge discovery in databases, KDD, are used synonymously.
  Note however, that data mining is only a *single step* within a KDD process, namely the analysis step for pattern recognition.
- Web data mining is the transfer and usage of data mining technology for information extraction on the Internet and especially the World Wide Web. Text mining is the identification of relevant information in text.

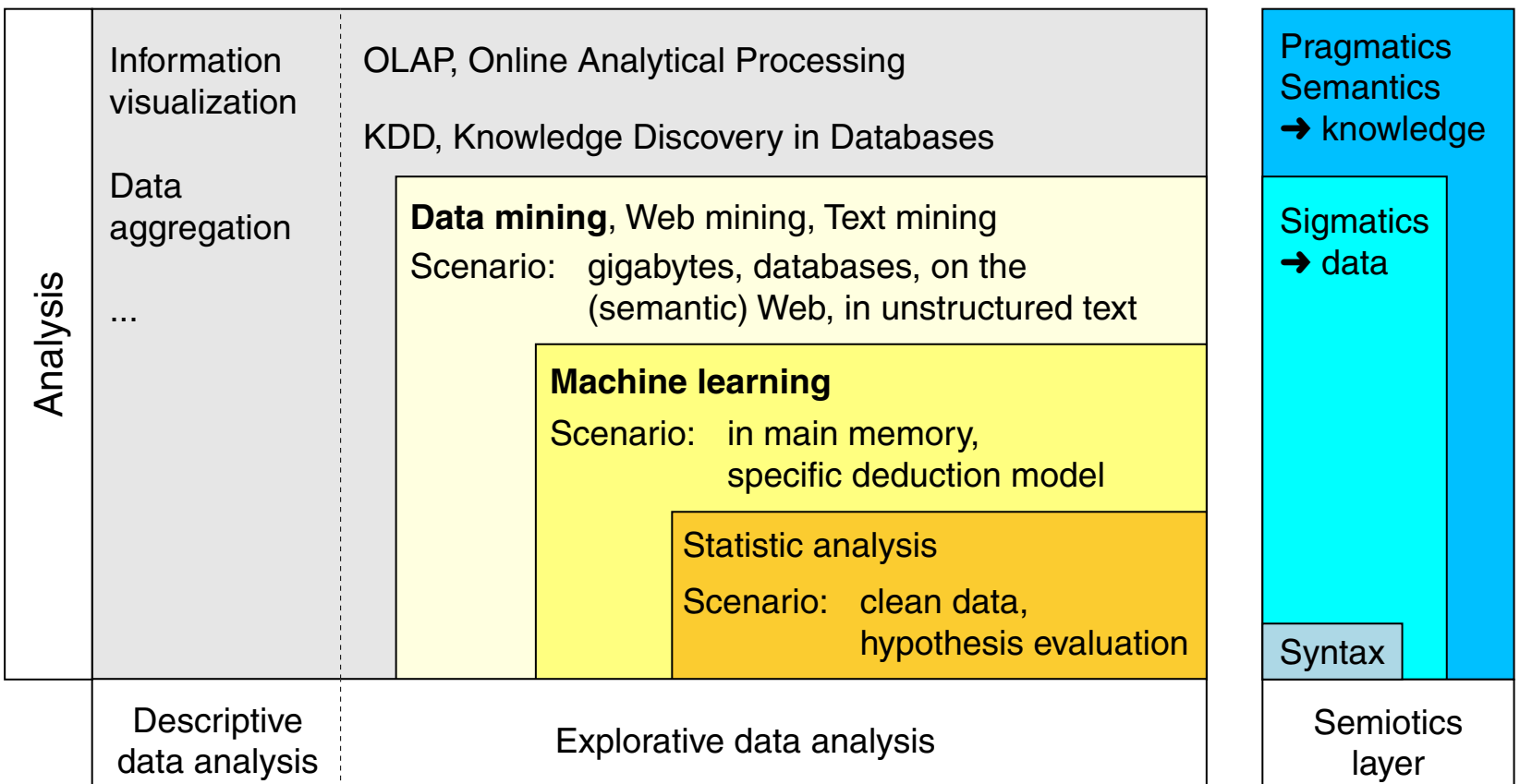
Data Mining Overview



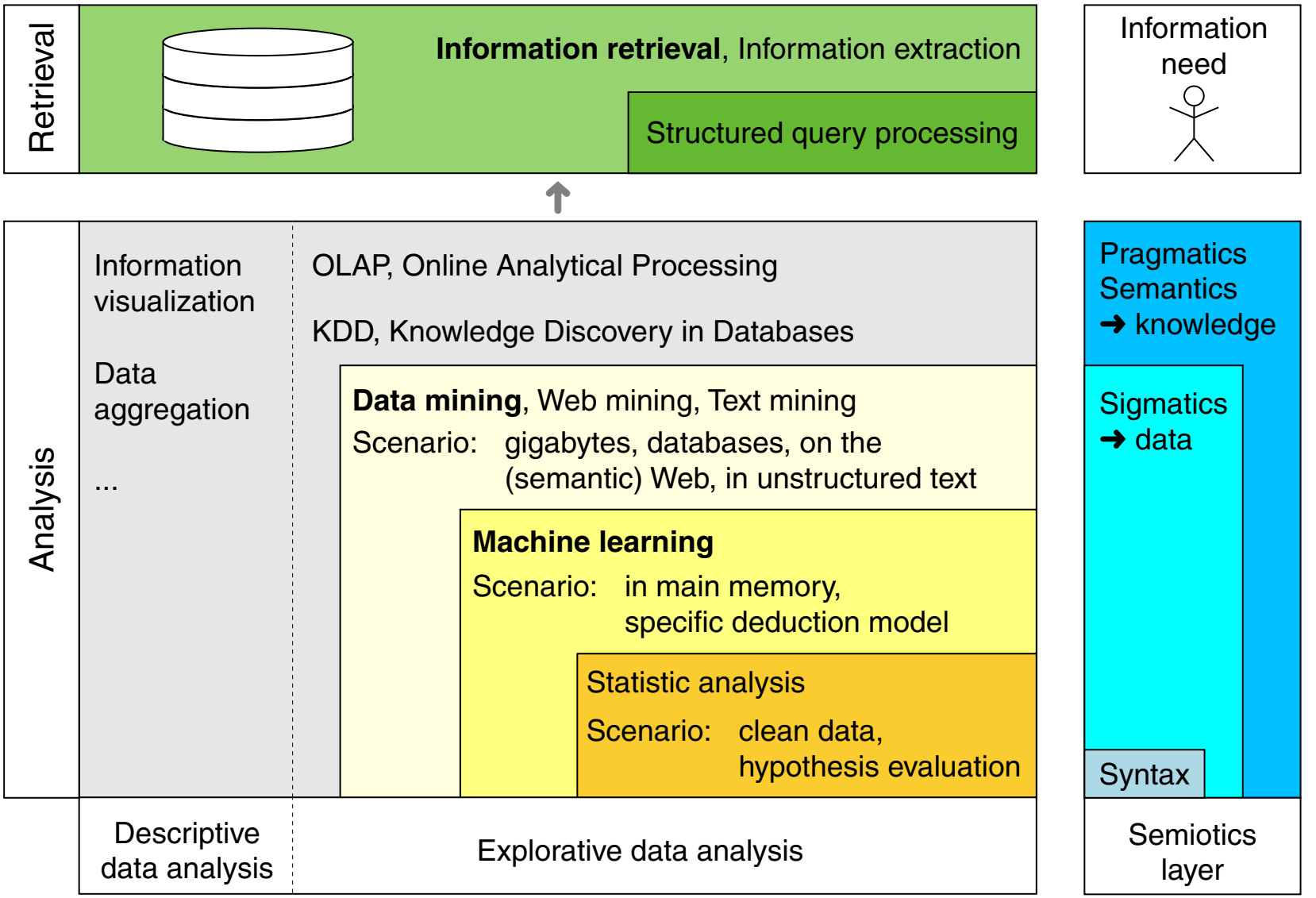
Data Mining Overview



Data Mining Overview



Data Mining Overview



Remarks:

- A clear separation between machine learning and data mining is not always possible. A key difference, however, results from the sizes of the analyzed data sets: machine learning applications are usually executed in main memory. The field of data mining arose from the necessity to apply analysis methods to large data bases.
- The foci of machine learning are the processes and theories of learning and deduction, such as analogical reasoning, learning from examples, or reinforcement learning. The major driving force behind data mining is the business world with their large data bases.
- The following count to relevant data mining problems:
  - undirected association analysis to identify dependencies between consumer products (market basket analysis)
  - cluster analysis and categorization
  - filtering of process data
  - forecasting and prediction

Data Mining Overview

Relevant Data Mining Methods

- cluster analysis
- learning of propositional or description-logical rules – example:

  ```
  IF status=married AND house_owner=true THEN creditor=good
  ```

- learning of association rules – example:

  “75% of the buyers of product A will buy the products B, C, and D as well.”

- principal component analysis (PCA), factor analysis
- multi-dimensional scaling (MDS)