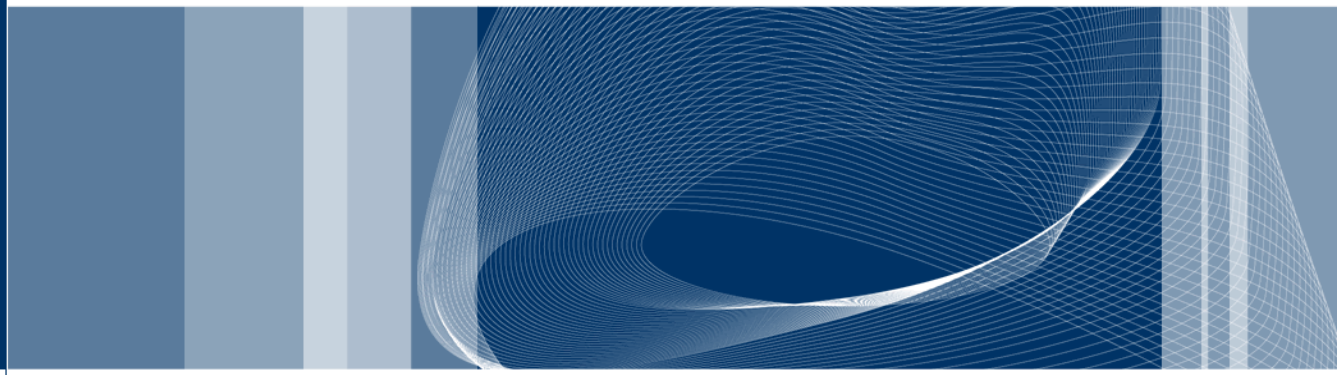
POLITECNICO DI MILANO



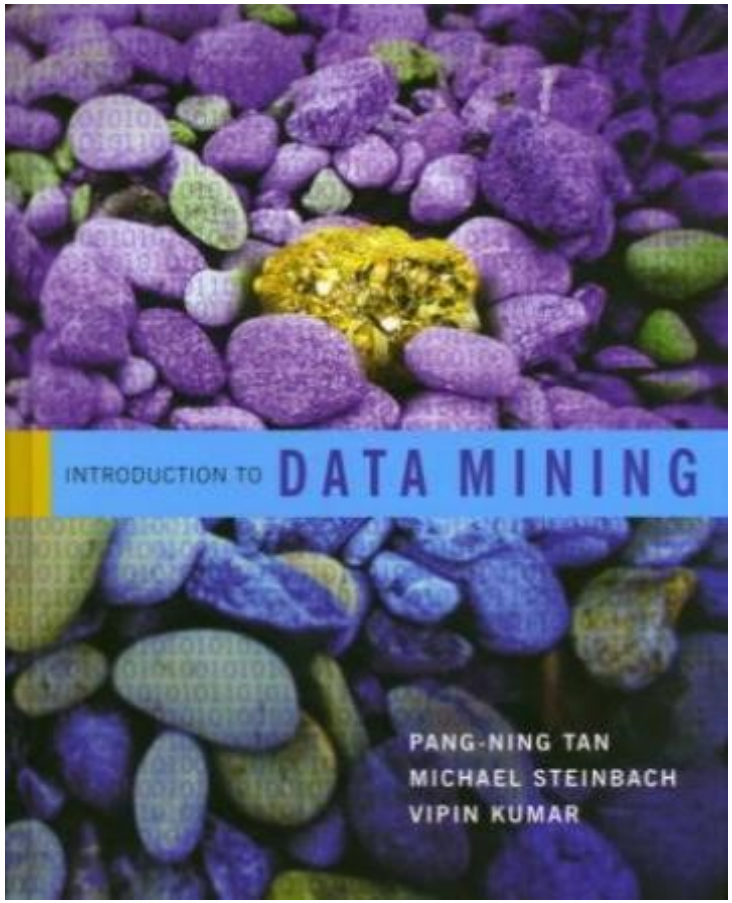
Data Mining

Information Retrieval and Data Mining

Prof. Matteo Matteucci

Slide Credits and References

- These slides have been heavily taken from:
  - Resources for Instructors and Students for "Introduction to Data Mining" by Pang-Ning Tan, Michael Steinbach, Vipin Kumar (2004)
  - Data Mining and Text Mining (UIC 583 @ Politecnico di Milano) course by Pier Luca Lanzi (2013)
  - Slides from "CS345A: Data Mining on the Web" Stanford course by Jeffrey D. Ullman



Do not blame just me if you will not like the content ☺

Evolution of Technology

- 1960s: data collection, database creation, & network DBMS
- 1970s: relational data model, relational DBMS implementation
- 1980s: RDBMS, advanced data models (extended-relational, OO, deductive, etc.); application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s: data mining, data warehousing, multimedia databases, and Web databases
- 2000s: stream data management and mining, web technology (XML, data integration), global information systems
- 2010s: social networks and linked data
- 2020s: "Personalized Genomics!" (… *according to some renowned colleagues*)

Why Data Mining? Pattern Analysis?

"Necessity is the mother of invention"

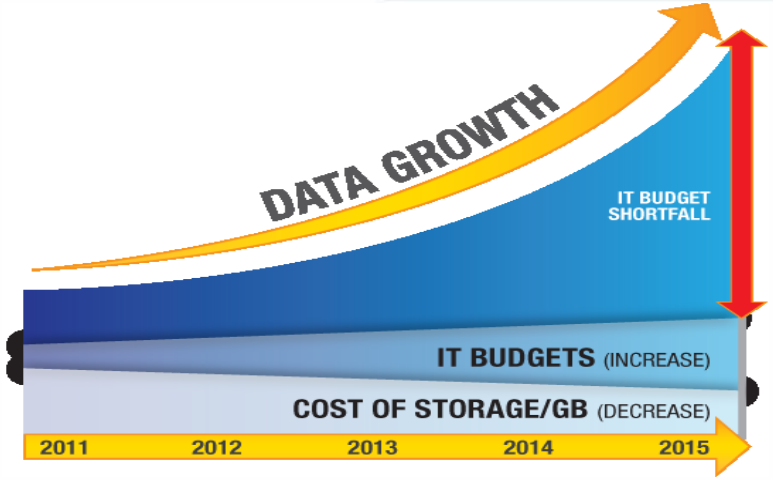
Emerged in the late 1980s

Major developments in the mid 1990s

Explosive Growth of Data

Pressing need for the automated analysis of massive data

Mining Large Datasets



- Much of the data is never analyzed at all!

- Human analysts may take weeks to discover useful information

- There is often information "hidden" in the data that is not readily evident

What is the Commercial Viewpoint?

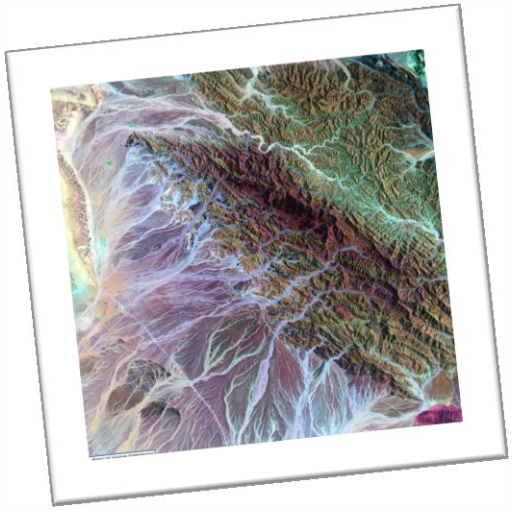
- Huge amounts of data is being collected and warehoused everyday
  - Web data, e-commerce
  - Purchases at department stores
  - Bank/Credit Card transactions

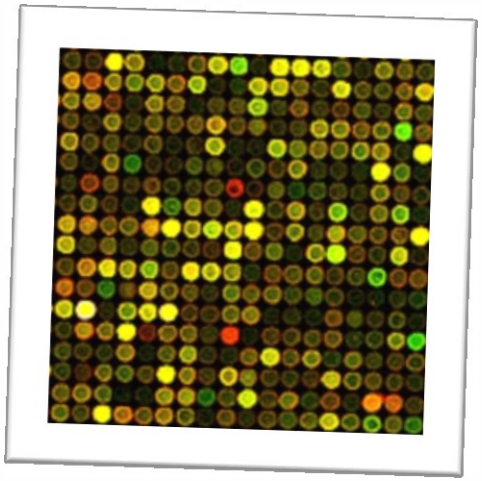
- Computers have become cheaper and more powerful
  (not to talk about storage)

- Competitive pressure is strong
  to provide better, customized services
  (e.g., CRM or Customer Relationship Management)

What is the Scientific Viewpoint?

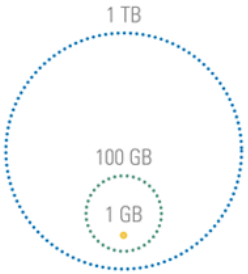
- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation

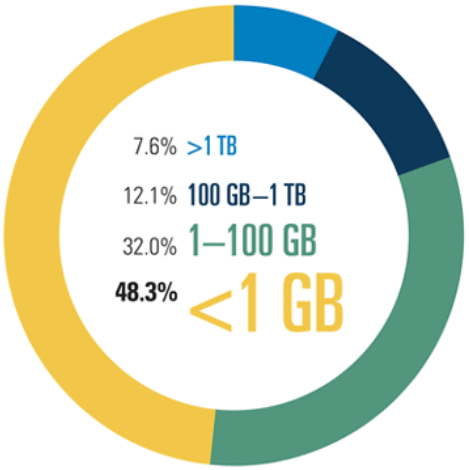




Science 11 February 2011 "Dealing with Data"

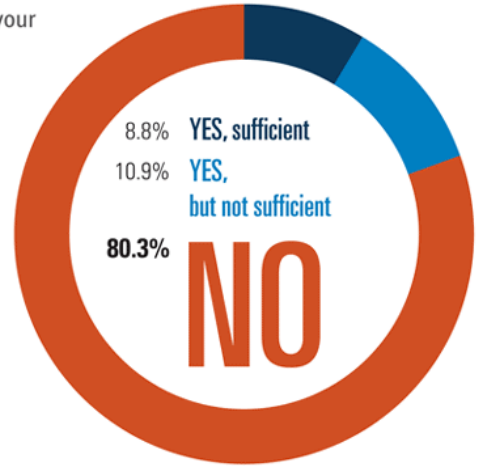
What is the size of the largest data set that you have used or generated in your research?





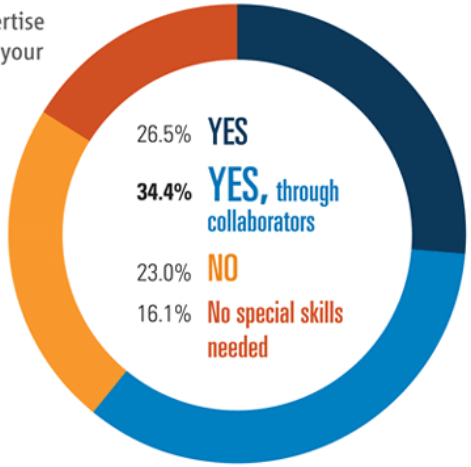
Is there sufficient funding for your lab or research group for data curation?

"There are many tales of early archaeologists burning wood from the ruins to make coffee. If we fail to curate the environmental archives **we collect from nature at public expense**, we essentially repeat those mistakes."

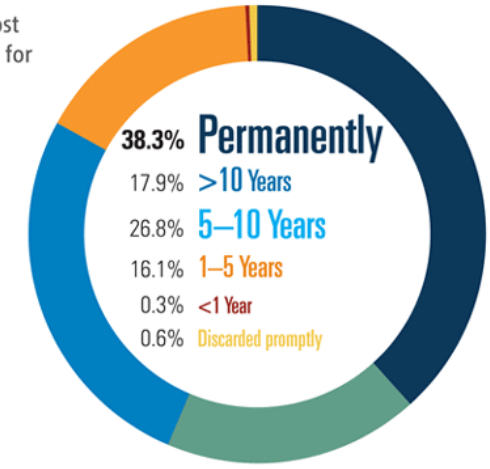


Do you have the necessary expertise in your lab or group to analyze your data in the way you want?

"The next few years [particularly in medicine] **the volume of data we need to analyze will expand exponentially.**"



For how long do you store most data generated in your lab or for your research associated with your publications?



What is Data Mining?

What is Data Mining?

- The non-trivial process of identifying (1) valid, (2) novel, (3) potentially useful, and (4) understandable patterns in data.

- Alternative names,
  - Data Fishing, Data Dredging (1960-)
  - Data Mining (1990-), used by DB and business people
  - Knowledge Discovery in Databases (1989-), used by AI
  - Business Intelligence, Information Harvesting, Information Discovery, Knowledge Extraction, ...
  - Currently, ***Data Mining*** and ***Knowledge Discovery*** are used interchangeably

- Data Mining ***is not*** looking up in the phone directory, it ***is not*** querying a Web search engine for information about "Amazon"

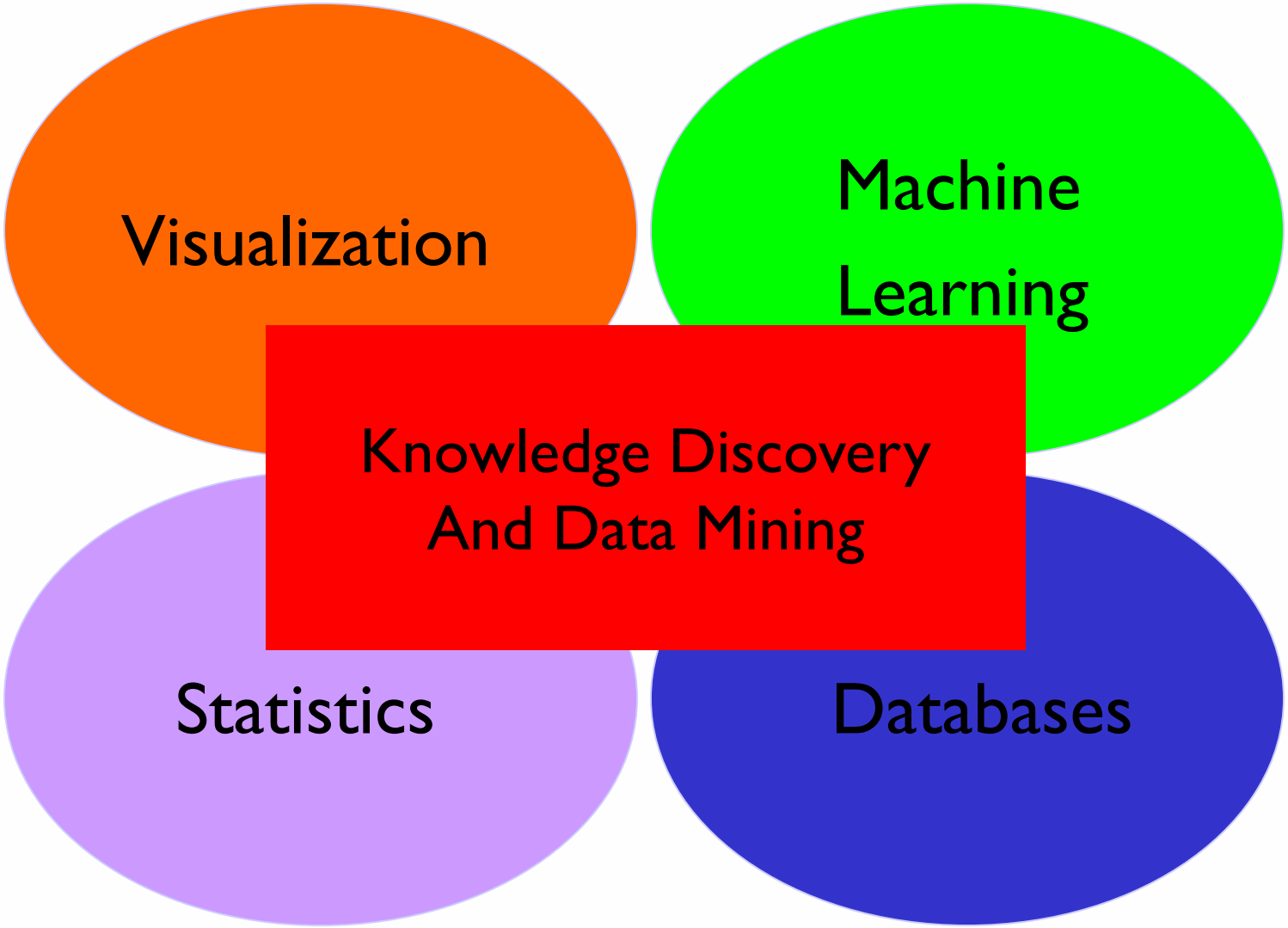
What is the General Idea?

- Build computer programs that navigate through databases automatically, seeking regularities or patterns

- There will be problems
  - Most patterns are trivial and uninteresting
  - Most patterns are spurious, inexact, or contingent on accidental coincidences in the particular dataset used
  - Real data is imperfect: some parts will be garbled, and some will be missing

- Algorithms need to be robust enough to cope with imperfect data and to extract regularities that are inexact but useful

What are the Related Research Fields? (Who does Data Mining?)



Statistics, Machine Learning, or Data Mining

- *Statistics* is more theory-based, focuses on hypotheses testing,

  "Strong results come with strong assumptions!"

- *Machine Learning* is more based on heuristic, focuses on building program that learns, more general than Data Mining

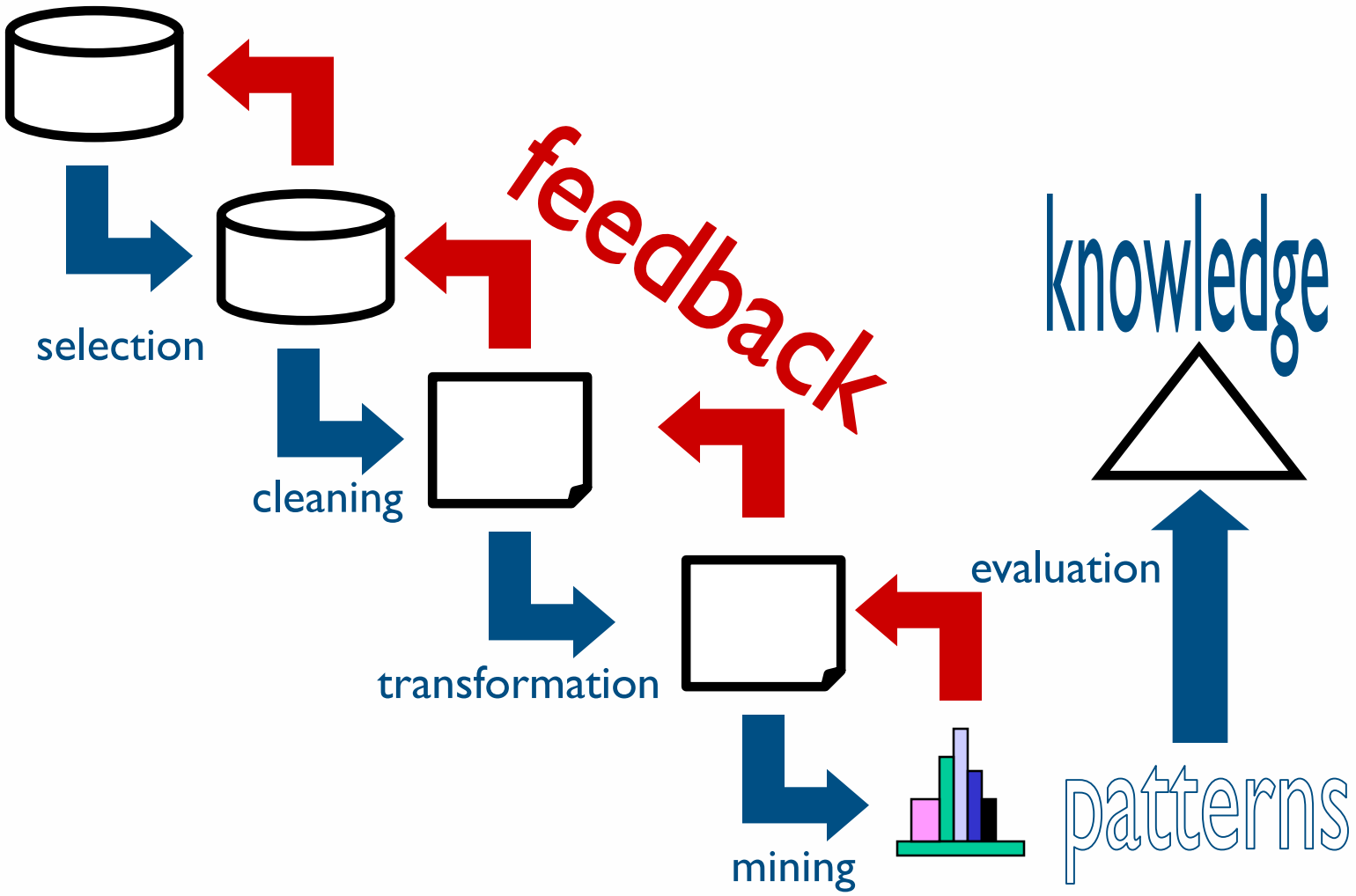
- *Knowledge Discovery*
  - integrates theory and heuristics
  - focus on the entire process of discovery, including data cleaning, learning, integration and visualization

Distinctions are blurred!

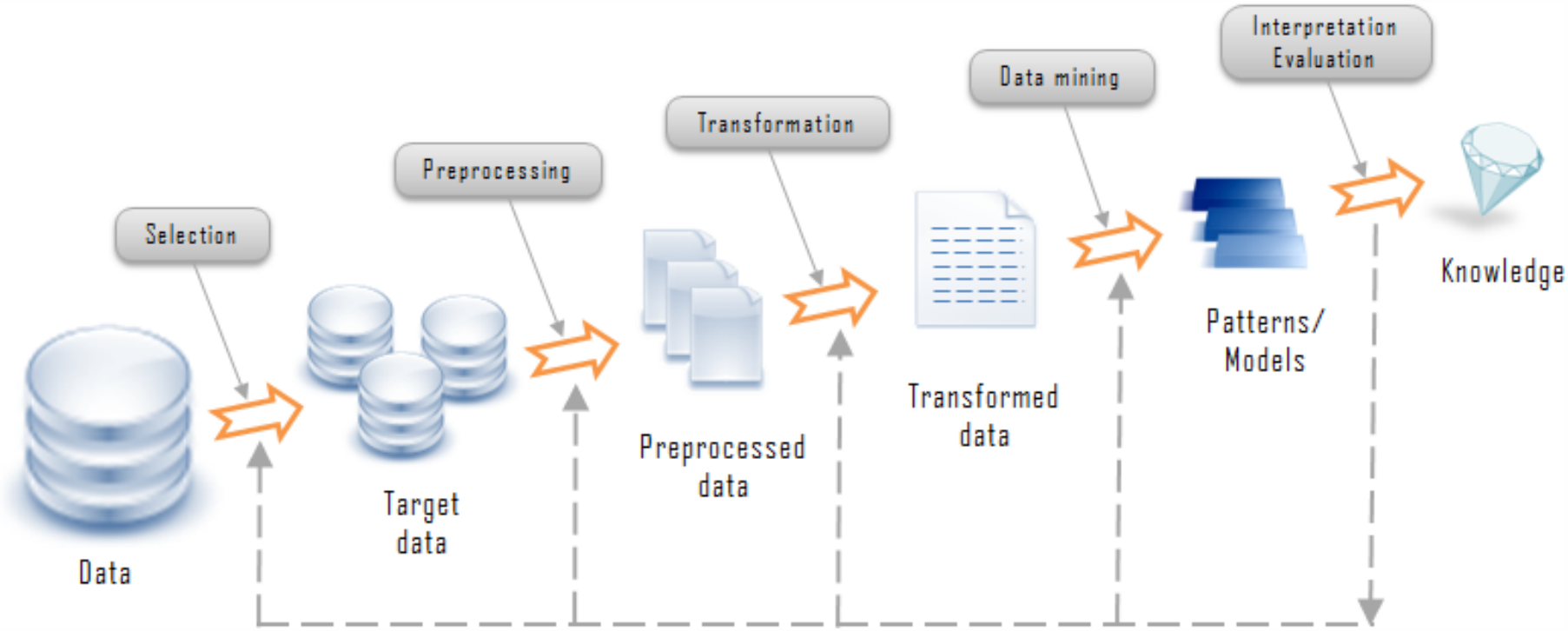
Why Not Traditional Data Analysis?

- Tremendous amount of data
  - High scalability to handle terabytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations

Knowledge Discovery Process

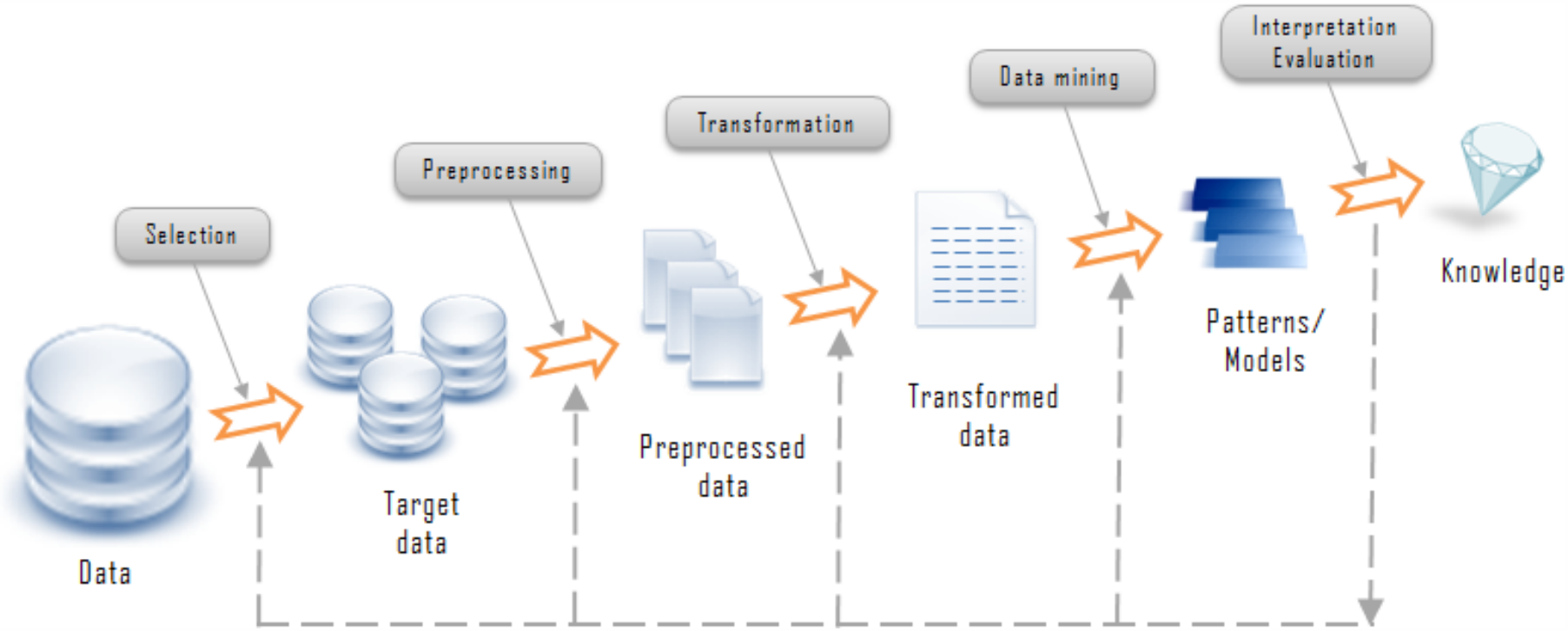


What are the (human) main steps?



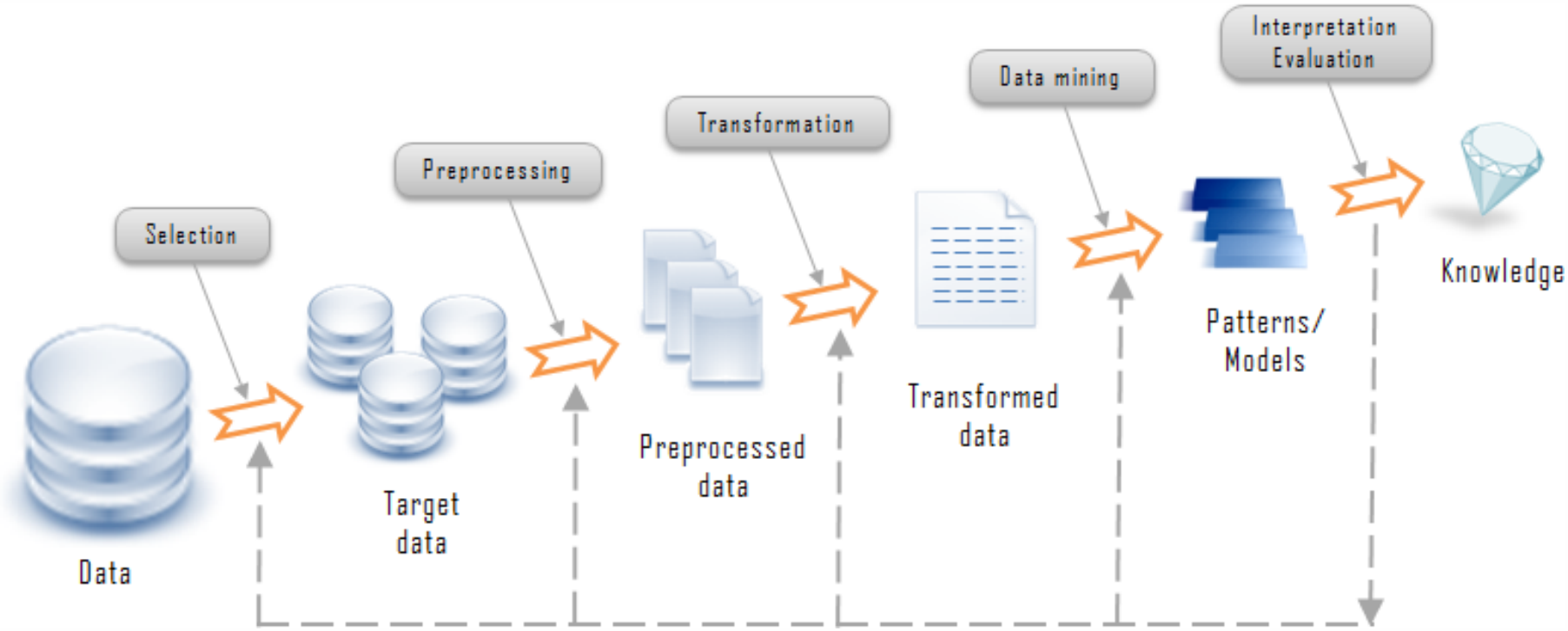
- Learn the application domain to extract relevant prior knowledge and define the goals for the mining
- Prepare the data for the mining
  - data selection
  - data cleaning
  - data reduction and transformation

What are the (human) main steps?



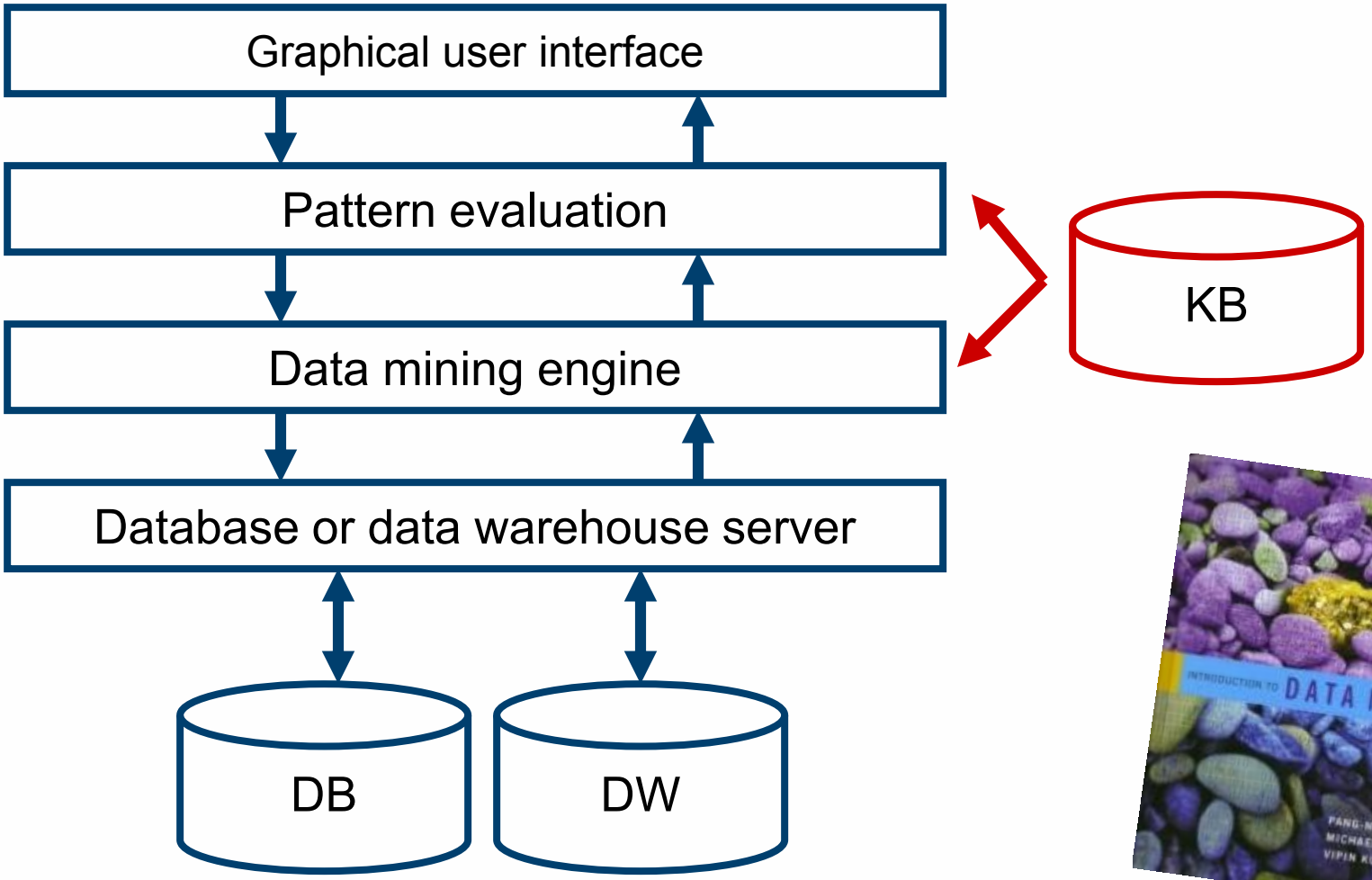
- Mining
  - Select the mining approach: classification, regression, association, clustering, etc. (this is related to the *potential use* of the result)
  - Choose the mining algorithm(s)
  - Perform mining: search for patterns of interest

What are the (human) main steps?



- Pattern evaluation and knowledge presentation
  - Visualization, transformation, removing redundant patterns, etc.
- Depending on the outcome
  - Use of discovered knowledge
  - Repeat the process from any of the previous step

Architecture of a Typical Knowledge Discovery System

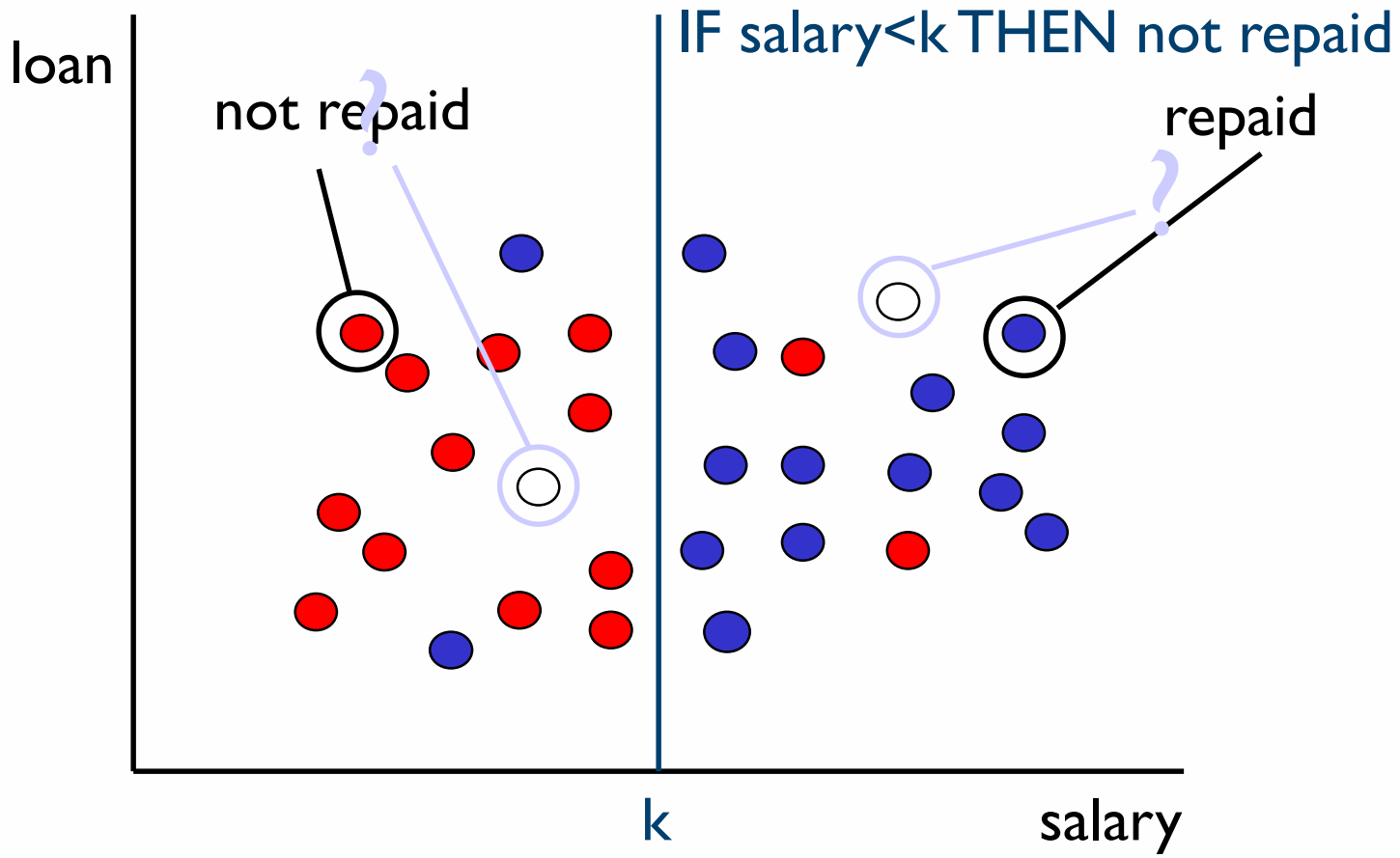


What are the typical Data Mining tasks?

What are the Major Data Mining Tasks?

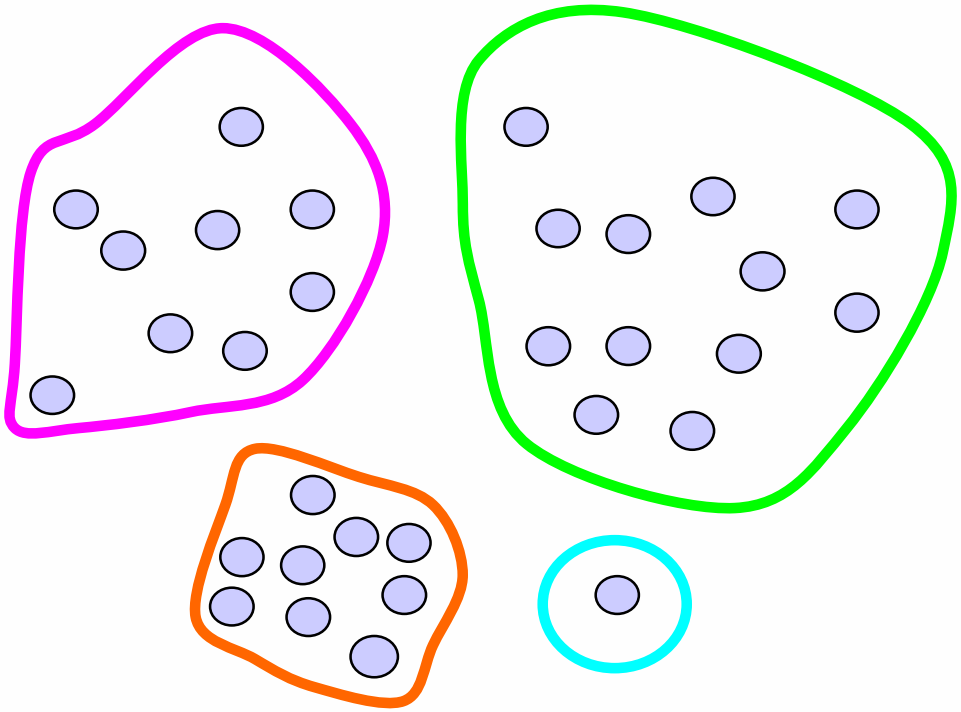
- Classification: predicting an item class / category / outcome
- Clustering: finding clusters / groups in data
- Associations: detect frequent occurring events…
- Visualization: to facilitate human discovery
- Summarization: describing a group of data in a meaningful way
- Deviation Detection: finding changes in normal data patterns
- Estimation/Regression: predicting a continuous value
- Link Analysis: finding relationship (e.g., social media, page-rank)
- But many appears as time goes by …
  - Outlier analysis, rare event analysis
  - Trend and evolution analysis, sequential pattern mining
  - Text Mining, Graph Mining, Data Streams
  - Sentiment analysis, Reputation analysis, Opinion mining
  - …

Example: Credit Risk



Data Mining Tasks: Clustering

- The class label is unknown
- Group data to form new classes, e.g., cluster houses to find distribution patterns
- Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity



Data Mining Tasks: Associations

Bread
Peanuts
Milk
Fruit
Jam

Bread
Jam
Soda
Chips
Milk
Fruit

Steak
Jam
Soda
Chips
Bread

Is there something interesting to be noted?

Jam
Soda
Chips
Milk
Bread

Fruit
Soda
Chips
Milk

Fruit
Soda
Peanuts
Milk

Fruit
Peanuts
Cheese
Yogurt

Beer and Diapers

- A
- T
- I
- A

ould
well.
to
ales

Surprise examination !!!!

- The teacher needs some rest (already!) so please:
  - Find a data analysis which is a classification task
  - Find a data analysis which is a clustering task
  - Find a data analysis which is a association mining task
- What the most important step in the knowledge discovery process?

Is the result of Data Mining Meaningful?

Meaningfulness of Answers

- A big data mining risk is that you will "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.
  - The Rhine Paradox: a great example of how not to conduct scientific research.
  - A big objection to TIA was that it was looking for so many vague connections that it was sure to find things that were bogus and thus violate innocents' privacy.

Credits for the following slides should go to Jeffrey D. Ullman.

TIA stands for DARPA Total Information Awareness (now Terrorism Information Awareness)

Rhine Paradox

- Joseph Rhine was a parapsychologist in the 1950's who hypothesized that some people had Extra-Sensory Perception.
- He devi… …ts were a…
- He disc… …e able to get …
- He told these people they had ESP and called them in for another test of the same type.
- Alas, he discovered that almost all of them had lost their ESP.
  - What did he conclude?
  - What did you conclude?

Understanding Statistics and Bonferroni's principle will help you look a little less stupid than a parapsychologist!

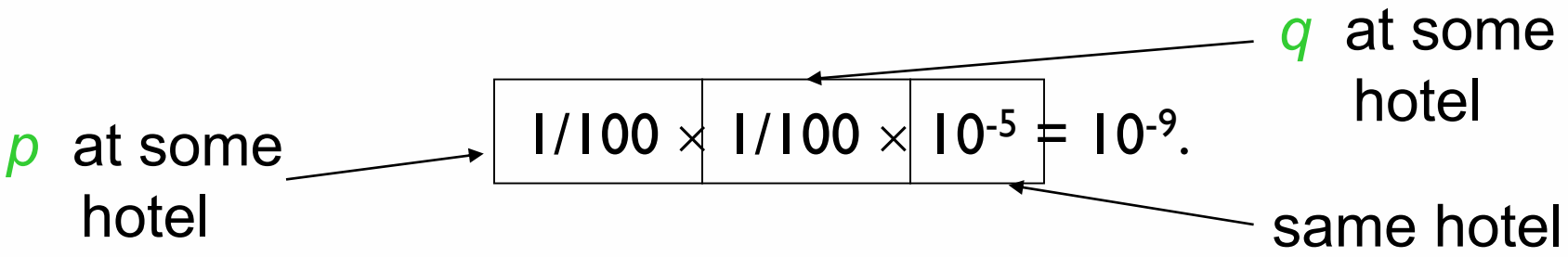
(Let's finish with bla bla bla! Press here to skip the TIA example)

The TIA Story

- Suppose we believe that certain groups of evil-doers are meeting occasionally in hotels to plot doing evil.
- We want to find (unrelated) people ***who at least twice have stayed at the same hotel on the same day***.
  - 10^9 people being tracked.
  - 1000 days.
  - Each person stays in a hotel 1% of the time (10 days out of 1000).
  - Hotels hold 100 people on average (so 10^5 hotels).
- If everyone behaves randomly (i.e., no evil-doers) will the data mining detect anything suspicious?

The TIA Story: Let's do the math

- Probability that given persons p and q will be at the same hotel on given day d :



- Probability that p and q will be at the same hotel on given days d_1 and d_2:

$$10^{-9} \times 10^{-9} = 10^{-18}.$$

- Pairs of days (in the 1000 observed):

$$5 \times 10^{5}.$$

- Probability p and q will be at the same hotel on some 2 days:

$$5 \times 10^{5} \times 10^{-18} = 5 \times 10^{-13}.$$

The TIA Story: Conclusion

- Pairs of people (out of the 10^9 tracked):

  5×10^{17}

- Expec

Make sure not to allow so many possibilities in your query that enough random data will surely produce facts "of interest."

- Suppose now there are (say) 10 pairs of evil-doers who definitely stayed at the same hotel twice.

- Analysts have to sift through 250,010 candidates to find the 10 real cases … it is not gonna happen!

(So, what is this Bonferroni principle about?)