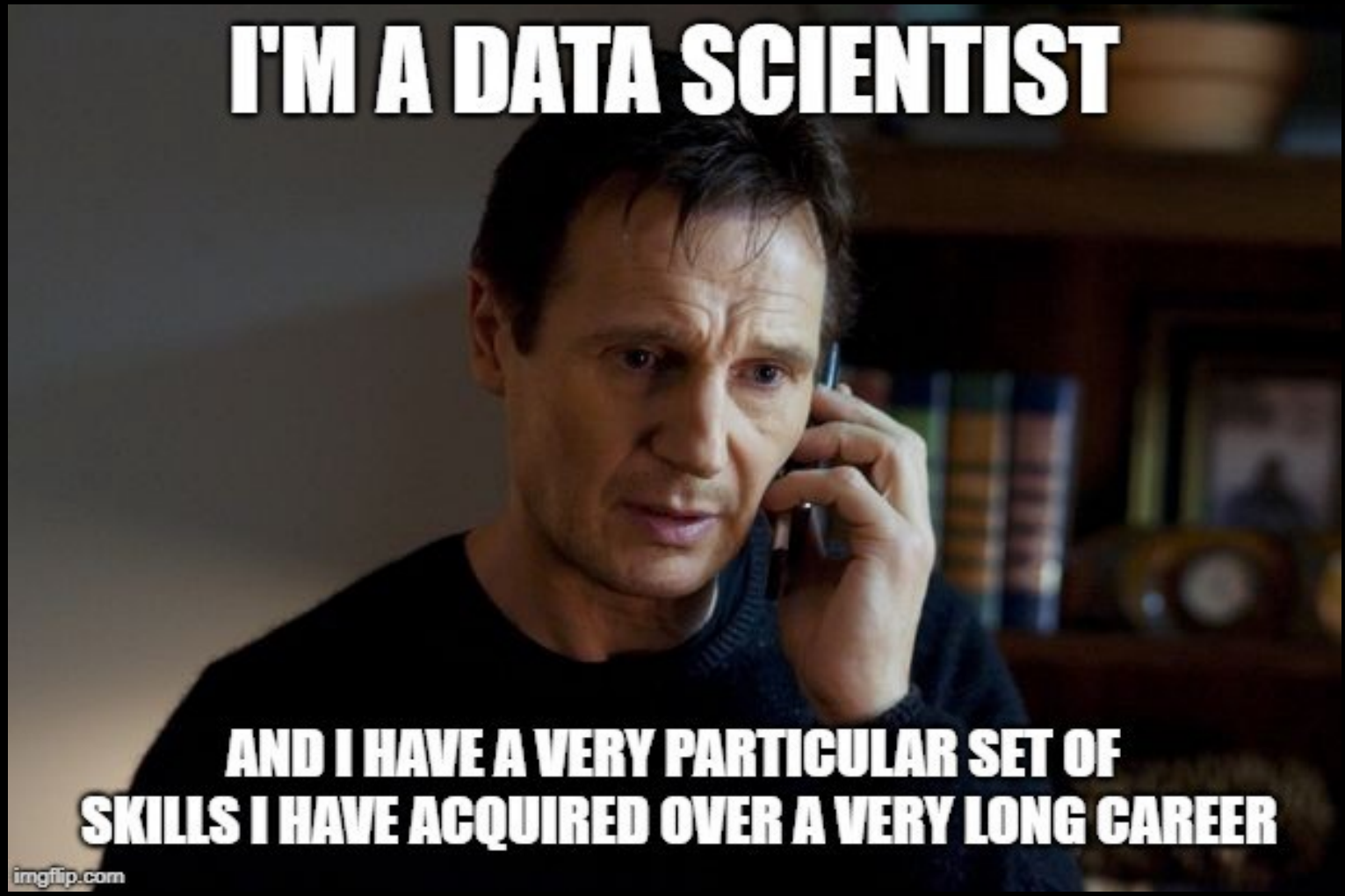
The Data Mining Process

Mining Massive Datasets

Prof. Carlos Castillo — https://chato.cl/teach

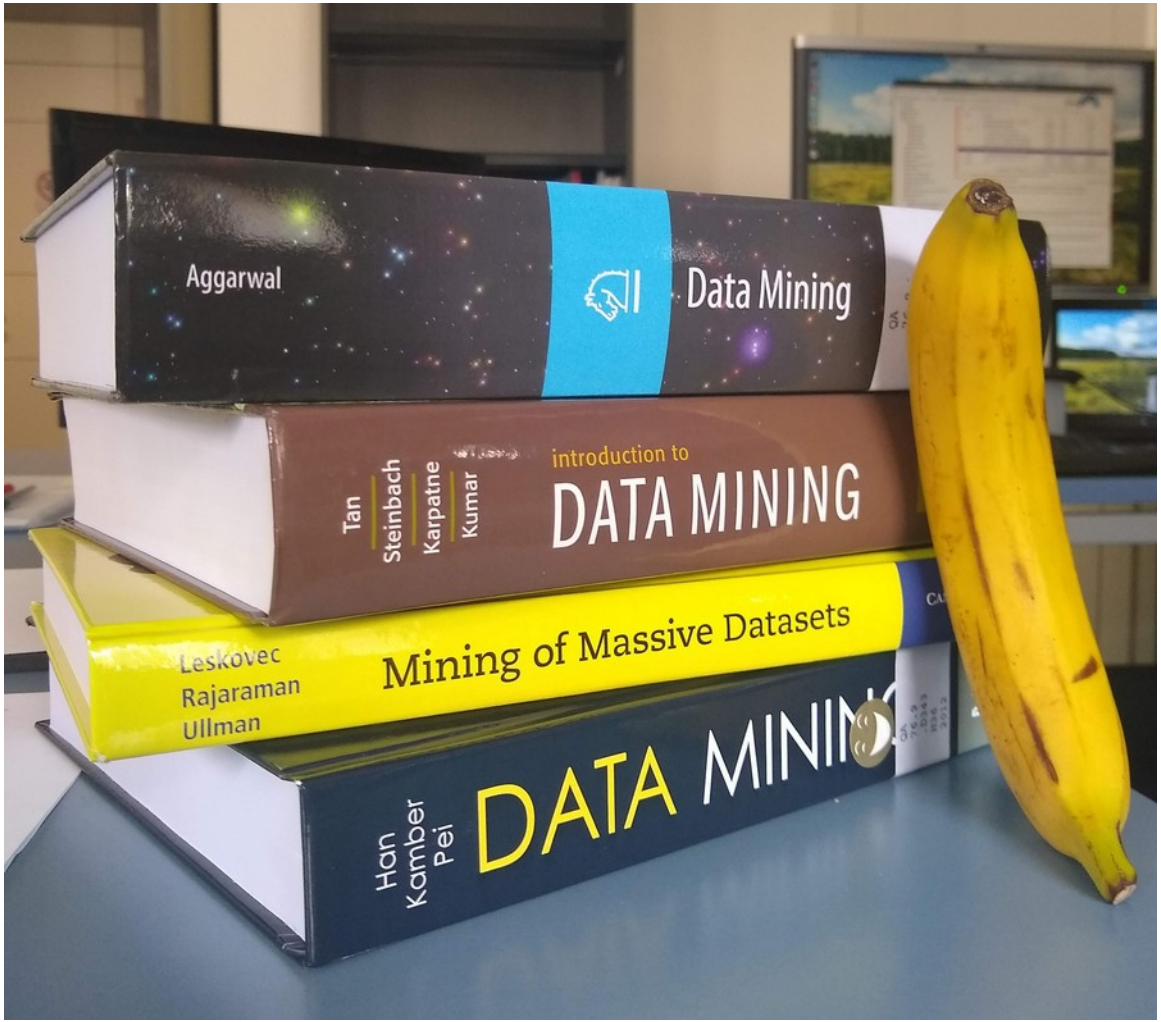
upf. Universitat Pompeu Fabra Barcelona


I'M A DATA SCIENTIST
AND I HAVE A VERY PARTICULAR SET OF
SKILLS I HAVE ACQUIRED OVER A VERY LONG CAREER
imgflip.com

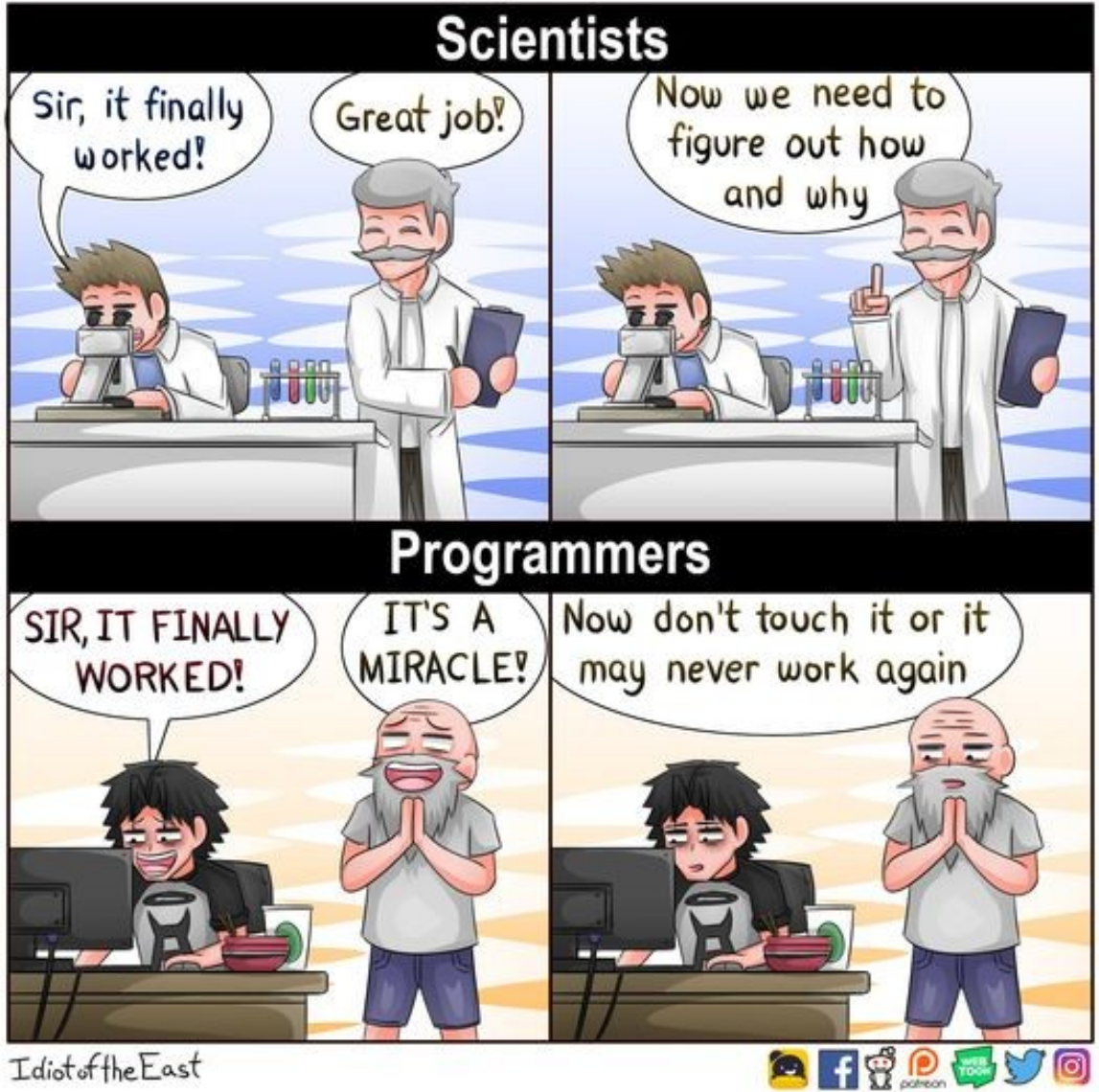
Taken (2008)

Main Sources

- Data Mining, The Textbook (2015) by Charu Aggarwal (Chapter 1) + slides by Lijun Zhang
- Mining of Massive Datasets, 2nd edition (2014) by Leskovec et al. (Chapter 1)
- Data Mining Concepts and Techniques, 3rd edition (2011) by Han et al. (Chapters 1-2)



(Banana for scale)


Scientists
Sir, it finally worked!
Great job!
Now we need to figure out how and why
Programmers
SIR, IT FINALLY WORKED!
IT'S A MIRACLE!
Now don't touch it or it may never work again
Idiot of the East

Data Mining

What do these have in common?

Stone

Clay

Papyrus

Paper

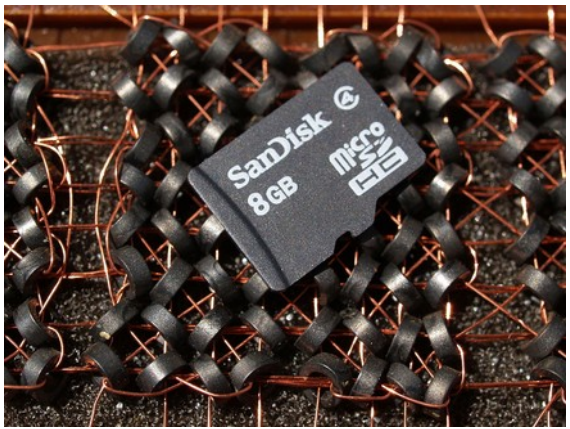
Wax cylinder

Tape

Vinyl

https://en.wikipedia.org/wiki/Writing

What do these have in common?



8GB (front) vs 8B (back)

Floppy disks (8”, 5 1/4”, 3 1/2”)

Compact disk

[Wikipedia]

The age of "Big Data"

The **co-evolution** of **storage** capacity, **transmission** capacity, and **processing** capacity

Visualcapitalist.com (2021)

Wikipedia definition

- **Data mining** is the process of
  - discovering patterns in
  - large data sets
  - involving methods at the intersection of
    - machine learning,
    - statistics, and
    - database systems.

Informal definition

Given **lots of data,** discover **patterns** and **models** that are:

- **Valid** hold on new data with some certainty
- **Useful** should be possible to act on them
- **Unexpected** or **novel** non-obvious
- **Understandable** interpretable
- **Complete** contain most of the interesting information

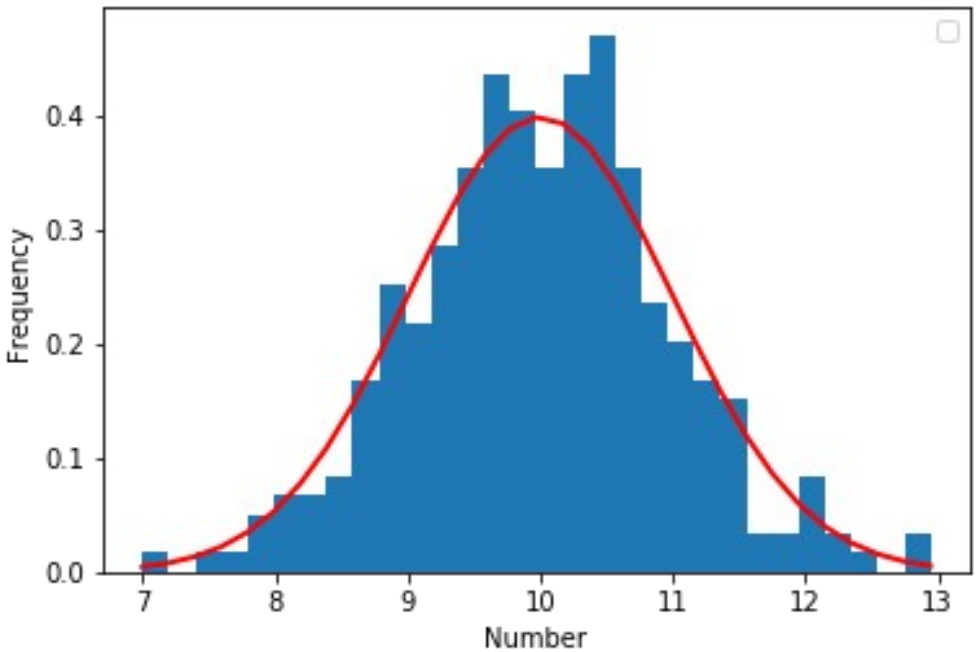
Example : 300 numbers

```
8.5998019   10.82452538 10.25496714  9.9264092  10.26304865  8.80526888  8.96569273  9.00883512  9.82813977 10.19311326
9.6545295   10.83958189 12.20970744 10.41521275 10.15902266  9.86904675 10.17021837 10.58768438 12.07341981  8.45713965
9.62152893  11.2494364   9.30073426 10.12753479 11.06429886  9.80406205  9.74418407 11.15815923 10.87659275 10.39190038
10.52911904 10.84125322 11.98925384 10.63545001  9.07420116 10.48011257 11.32273164  9.4831463  10.67973822 10.87064128
9.35940084   9.51149749 11.13211644  9.23292561  8.4767592   9.64339604  9.91374069  9.84184184  9.85576594  9.18523161
10.27107348  8.7511958   8.70297841 10.50609814 11.1908866  10.59484161 10.60027882  9.06375121 10.48534475  9.34253203
10.37303225  9.27441407 11.27229628 12.88441445  9.80825939  9.09844847 10.82873991  8.89169535 10.43092526  7.43215579
10.29787802  9.87946998  8.3799398  10.21263966  9.93826568  9.17325487 10.22256677 10.04892038 11.01233696  9.6145273
9.9495437   10.51474851  9.19288505  7.87728009  9.987364   10.94639021 10.01814962  9.40505023  8.87242546 10.23686131
8.90710325  10.31678617 10.4571519   9.04315227  9.85321707 11.89885306  6.99926999 10.71534924 10.29215034 10.59516732
9.8807174    9.01321711  8.45289144  9.1739316   7.90909364  9.42165081 10.37087284  9.57754821  9.60350044 10.75691005
8.24594836  10.33419146  9.7779209   9.51609087 10.25712725 12.1256587   9.53397549  9.44765209  9.53901558  9.8006768
9.633075    11.17692346 11.00022919  8.38767624  8.63908897  8.10049333 10.66422258 10.70986552 10.82945121 10.45206684
9.21578565  10.21230495 10.28984339  9.4130091  10.54597988 10.8042254  10.52795479 10.76288124 11.3554357  11.484667
10.36068758 8.18239896  11.20998409  9.88574571  9.8811874  10.64332788  8.67828643  9.23619936 10.71263899  9.36036772
8.80204902  8.84117879   9.60177677  8.82383074  9.85787872 10.30883419 10.09771435 10.33417508  8.94003225  9.63795622
8.88926589  8.51484154  10.61543214 10.10520145 10.23046826 11.22923654 10.25575855 10.4210496   9.79970778  7.70796076
9.56309629 10.82893108  10.4055698  10.12121772  9.38935918  9.48947921  9.53357322  9.87589518 10.5455508   9.98665703
9.440398    9.67368819  12.94191966 10.01303924 12.14295086  9.58399348 10.92799244 10.4654533  10.14613624  9.29818262
9.25613292 11.59370587   8.62517536 10.29703335  9.11065832 10.68766309  9.86507094 10.58314944 10.65232968  8.13400366
11.0414868 10.16883849  10.23649503 11.51859843  9.4754405  10.88103754  8.6249062   9.64581983  8.80660132 10.3794072
11.7687303  9.6768357   10.83753706 12.39138541  9.45756373 10.4746549  11.44321655 10.70109831  8.36186335  8.99123853
10.7221973  9.25735885  10.11287178  9.77908247 10.05372548 12.32358117  9.09128196 10.27487412  8.31704578  9.67337192
11.1712355 11.33146049  10.44967579  9.58649468  9.5908432  10.53829167 10.16738708 10.45433891 10.79223358 11.3936216
9.27709756  8.91159056   8.67186161  7.83968452 11.00207472 10.61085929 11.15868605 10.13873855  9.29370024 10.49794191
10.49884897 9.77150045   8.80503866 10.08775177 11.38167004 10.42724794 11.11626475 10.68890453 10.49280739  9.53675721
9.74560138 10.34343033  10.19711682  9.20212506  9.06407316 10.07228419 11.06791431 12.10523742  8.72119193 10.04645774
11.47090441 8.92472486  10.04585273 10.41149437  9.90118185  9.02229964  8.66708035 11.53976046 11.40609367  9.73014878
8.94607876 11.562354     9.58552216  9.74172847  9.64220948  9.69459042  9.58460199 11.14917832  9.49543794  9.46369271
10.16544667 9.92277128   9.61975057 11.11679747  9.42894032  9.25751891 11.44948256  8.16601628 10.11500258  9.42431821
```

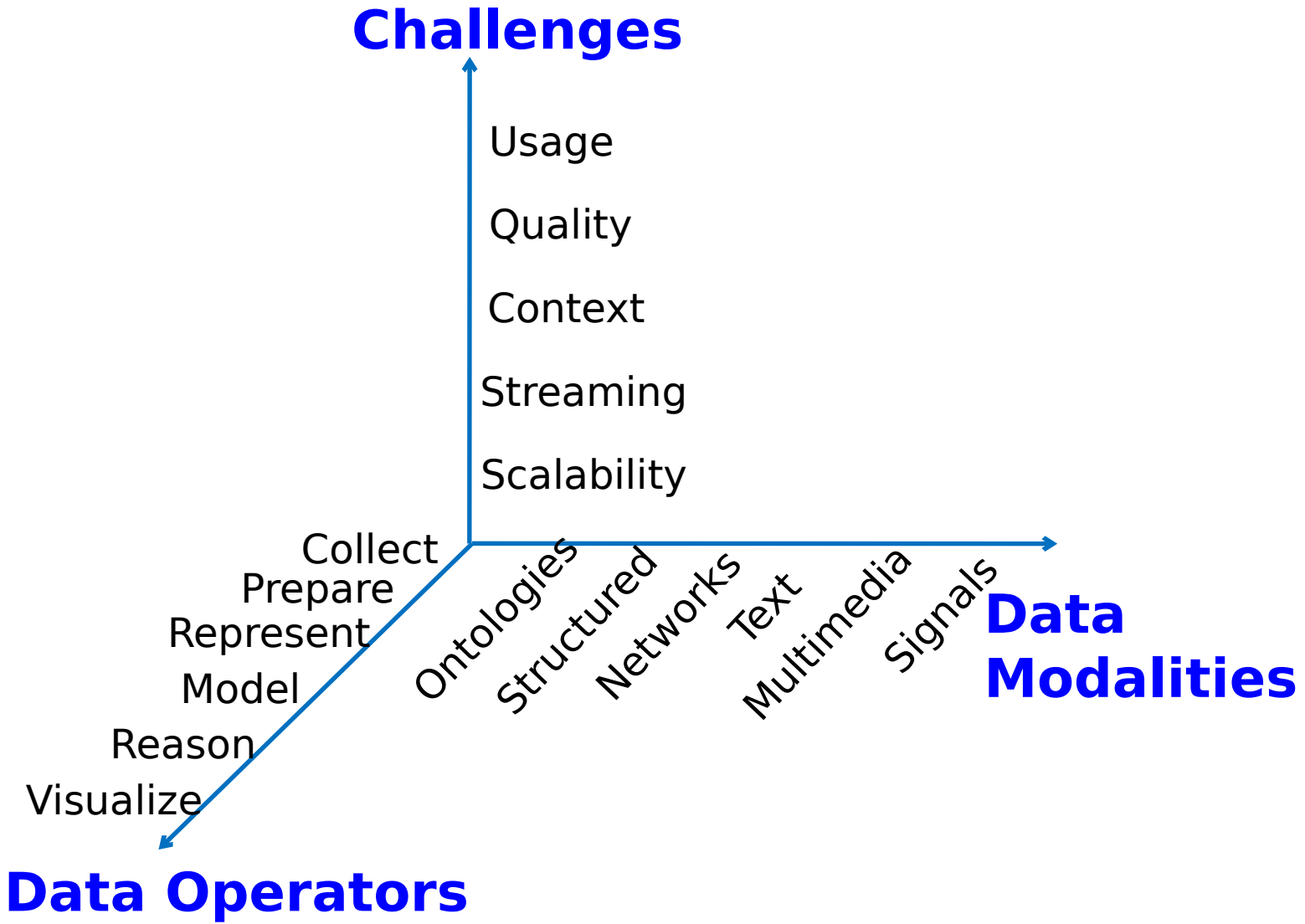
What are these numbers?

Example: 300 numbers (cont.)

Through *statistical modeling* we can find the data comes from a Normal distribution with mean 10 and standard deviation 1



- **Normal(μ=10,σ=1)** is a *model* for the data


Challenges
Usage
Quality
Context
Streaming
Scalability
Collect
Prepare
Represent
Model
Reason
Visualize
Ontologies
Structured
Networks
Text
Multimedia
Signals
Data Modalities
Data Operators

Describing vs Predicting

Descriptive methods

- Find human-interpretable patterns that describe the data
- Example: Clustering

Predictive methods

- Use some variables to predict unknown or future values of other variables
- Example: recommender systems

Characterizing vs Distinguishing

Data characterization methods

- A summary of the general characteristics or features of a target class of data

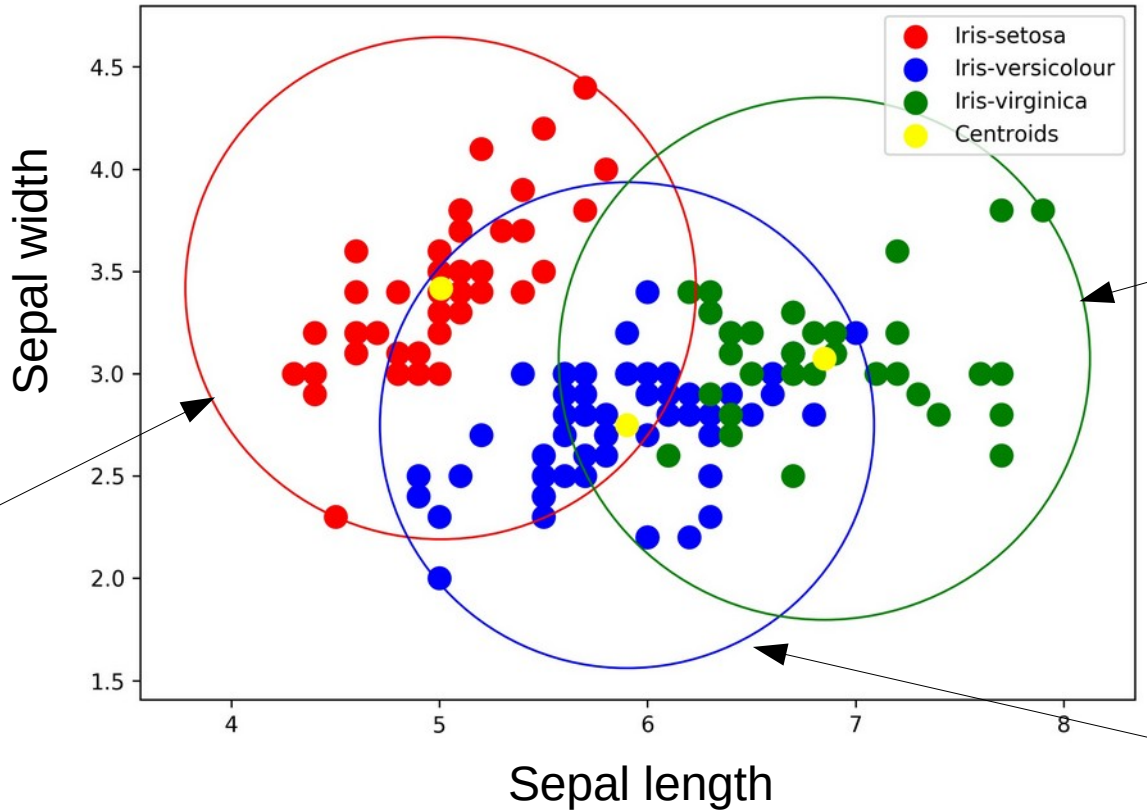
Data discrimination methods

- A comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes

Data mining has several goals

- To produce a **model**
  - E.g., a regression model for a numerical variable, or a classification model for a categorical variable
- To create a **summary**
- To extract **prominent features**

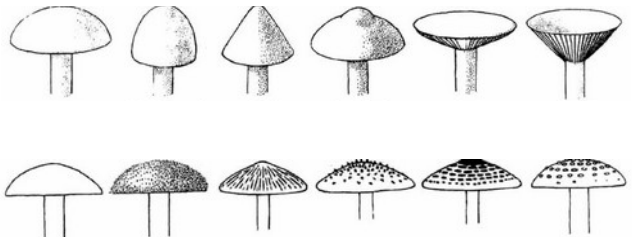
Example summary: clustering

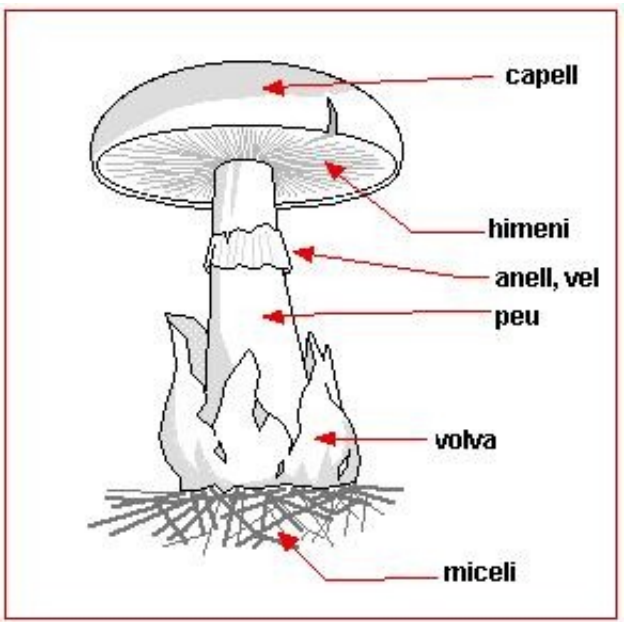


Picking the right features

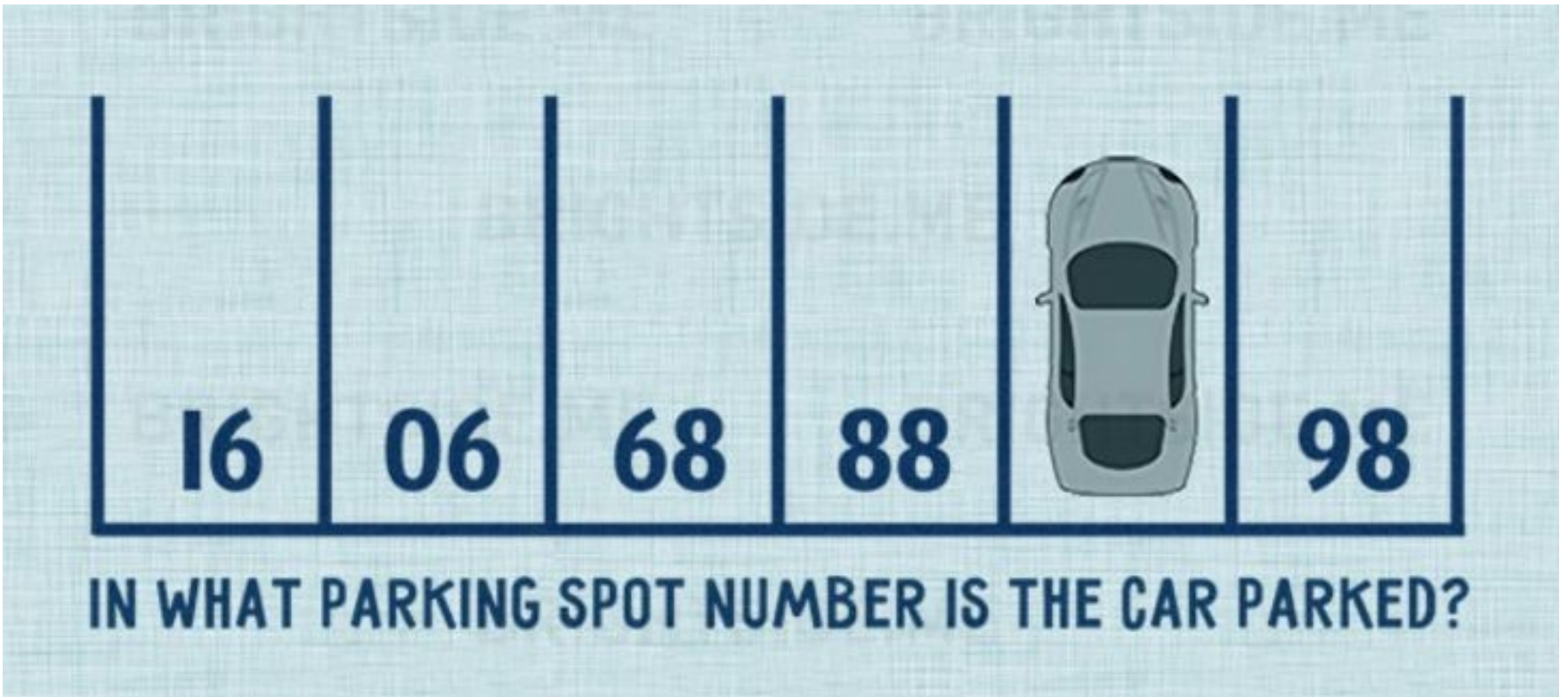
- Representing these flowers by their *petal length* and *sepal length* was key
  - These are good features for this task
- Other features such as color or number of leaves may not be so good
- Feature selection is key!

Features: a matter of life or death





Another pattern-finding example



Source: Centauro Blog (2017)

Example: complex features

- Given shopping baskets of previous customers, determine:
  - **Frequent itemsets** (bought together)
  - **Similar items** (e.g., for recommendations)

Risk #1: Spurious patterns

- A risk with "Data mining" is that an analyst can "discover" patterns that are **meaningless**
- *If you look in more places for interesting patterns than your amount of data will support, you are bound to find something (~Bonferroni principle)*

If you interrogate data
hard enough it will tell you
what you want to hear

Risk #2: Surveillance state

- Attention-grabbing evil actions are also very rare, with consequences:
  - Suppose 1 in a million in a suicide bomber
  - Catching one suicide bomber a year on average means examining 999.999 innocent people
- A system with 1% false positive rate will flag ~10K people as potential suicide bombers

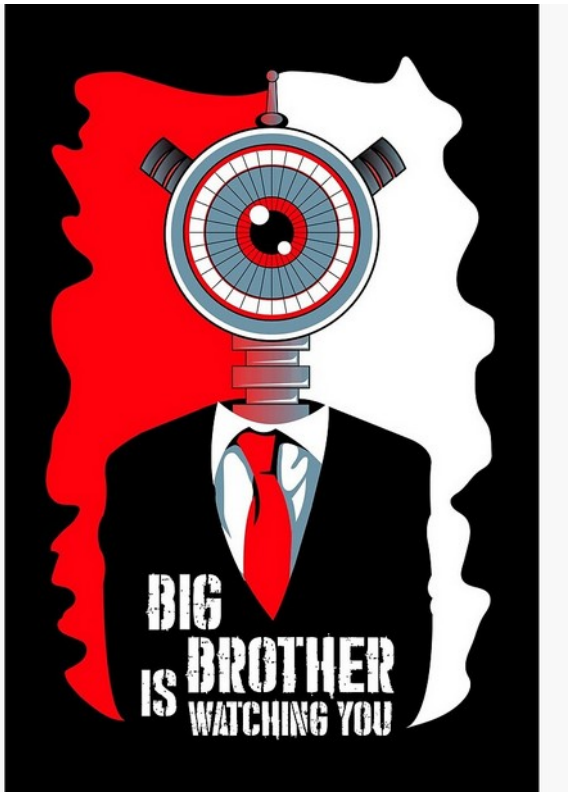


Image: Red Bubble

Data mining (DM) vs other disciplines

- For a database person, DM means analytic processing
- For a machine learning person, DM means modeling
- For an algorithms person, DM means ensuring scalability

Our focus will be on **scalable algorithms**

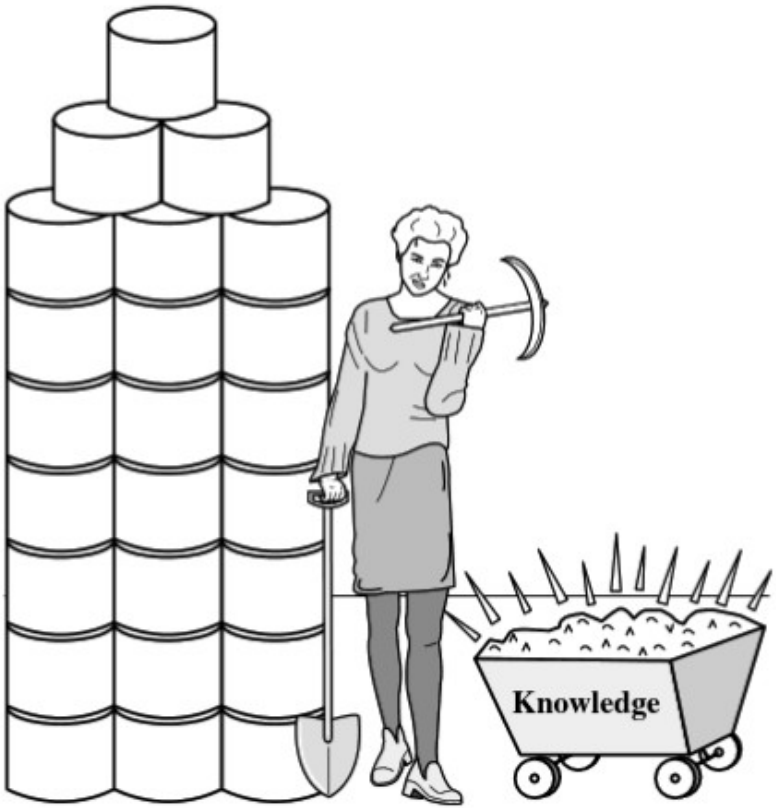
Data rich but information poor

- Fast-paced data streams become data archives that become data tombs
- Decisions could be better made by using data that already exists but is hard to "mine"

Data Mining Concepts and Techniques, 3rd edition (2011) by Han et al.

Knowledge Discovery from Data

- KDD, a popular acronym
  - "Discovery" is Data Mining
- Other names: knowledge mining from data, knowledge extraction, data/pattern analysis



Data Mining Concepts and Techniques, 3rd edition (2011) by Han et al.

Typical stages of KDD

1)Data Cleaning

2)Data Integration

3)Data Selection

4)Data Transformation

5)Data Mining $\leftarrow$ application of a DM algorithm

6)Pattern Evaluation

7)Knowledge Presentation

Typical stages of KDD

1)Data Cleaning

2)Data Integration

3)Data Selection

4)Data Transformation

Pre-processing phase

5)Data Mining

6)Pattern Evaluation

7)Knowledge Presentation

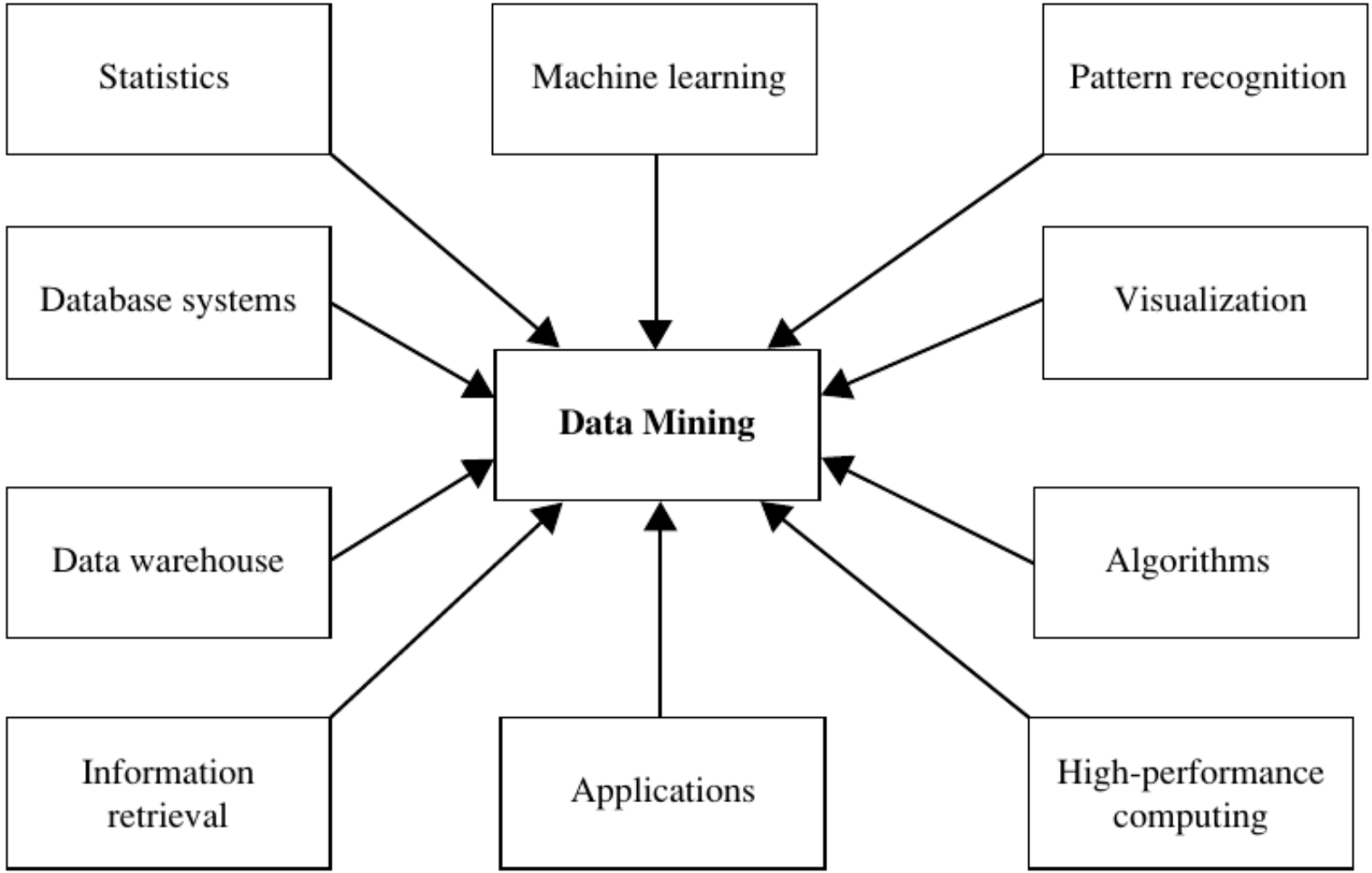
Analytical phase

Summary

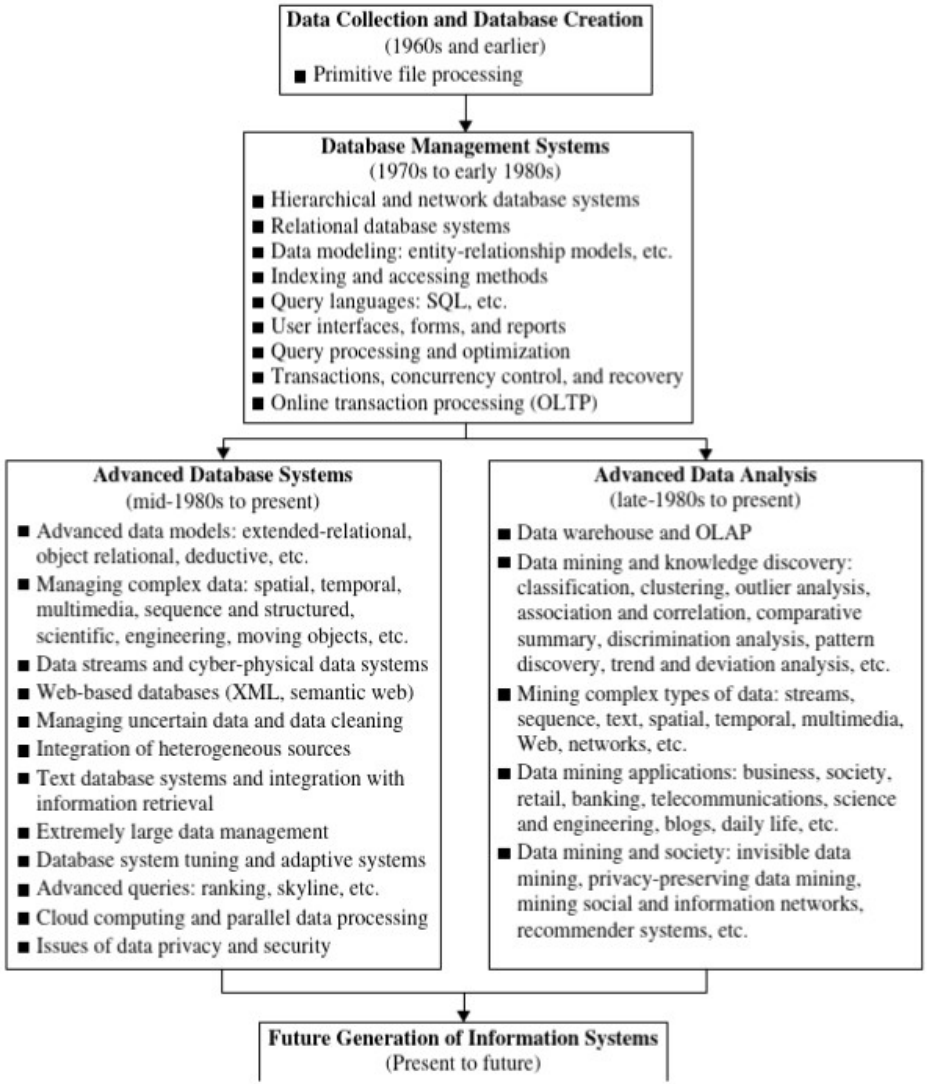
Things to remember

- Define and contrast:
  - Describing vs Predicting
  - Characterizing vs Discriminating
- Describe the stages of the KDD process

Additional contents
(not included in exams)



Data Mining Concepts and Techniques, 3rd edition (2011) by Han et al.



Data mining is a descendant of methods for Online Analytical Processing (OLAP) done over Data Warehouses

Data Mining Concepts and Techniques, 3rd edition (2011) by Han et al.