Data & Web Mining

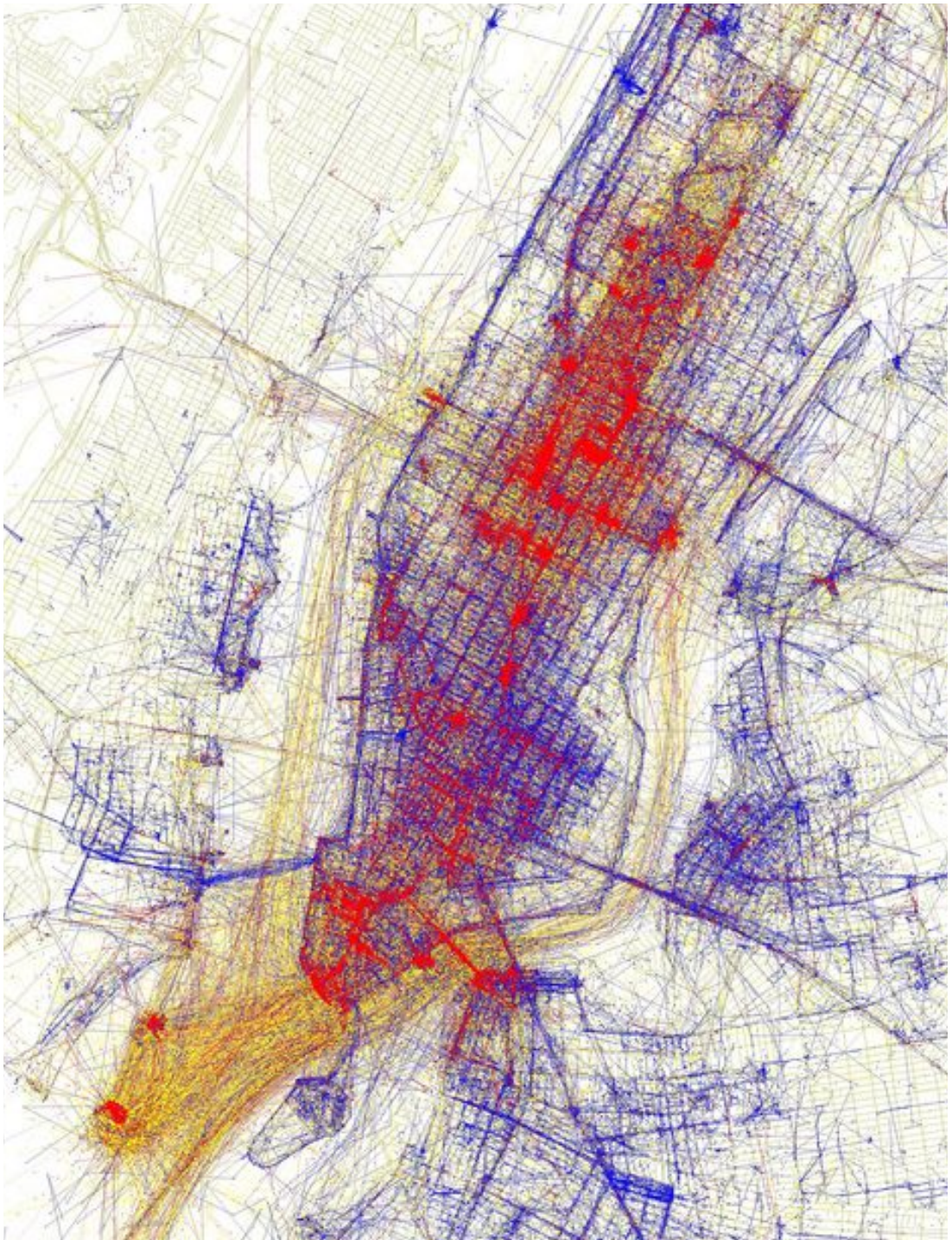
Claudio Lucchese claudio.lucchese@unive.it

General information

DATA AND WEB MINING

Anno accademico	2021/2022
Titolo corso in inglese	**DATA AND WEB MINING**
Codice insegnamento	CT0509 *(AF:337527 AR:178736)*
Modalità	In presenza
Crediti formativi universitari	6
Livello laurea	Laurea
Settore scientifico disciplinare	INF/01
Periodo	I Semestre
Anno corso	3
Sede	VENEZIA
Spazio Moodle	**Link allo spazio del corso**

- Moodle
- Software:
  - Jupyter Notebooks: Python-based environment
  - You may bring your laptop with
  - Google Colab https://colab.research.google.com/
  - or Anaconda https://www.anaconda.com/products/individual
  - or DataSpell https://www.jetbrains.com/dataspell
  - ot VSCode https://code.visualstudio.com/
- Exam: please wait for the last slide
- Teaching Material
  - Introduction to Data Mining, Global (or Second) Edition, Kumar et al
  - Book excerpts. Check moodle and references at the end of each set of slides.
- Contact
  - claudio.lucchese@unive.it
  - *Always check the moodle !!!!!!!*



The Data Deluge

- "When the **Sloan Digital Sky Survey** started work in **2000**, its telescope in New Mexico collected ***more data in its first few weeks than had been amassed in the entire history of astronomy***. Now, a decade later, its archive contains a whopping 140 terabytes of information. ***A successor***, the Large Synoptic Survey Telescope, due to come on stream in Chile in **2016**, ***will acquire that quantity of data every five days***."
- "[..] **Wal-Mart**, a retail giant, handles more than 1m customer transactions every hour, feeding databases estimated at ***more than 2.5 petabytes, the equivalent of 167 times the books in America's Library of Congress*** […]"
- "**Facebook**, a social-networking website, is home to ***40 billion photos.***"

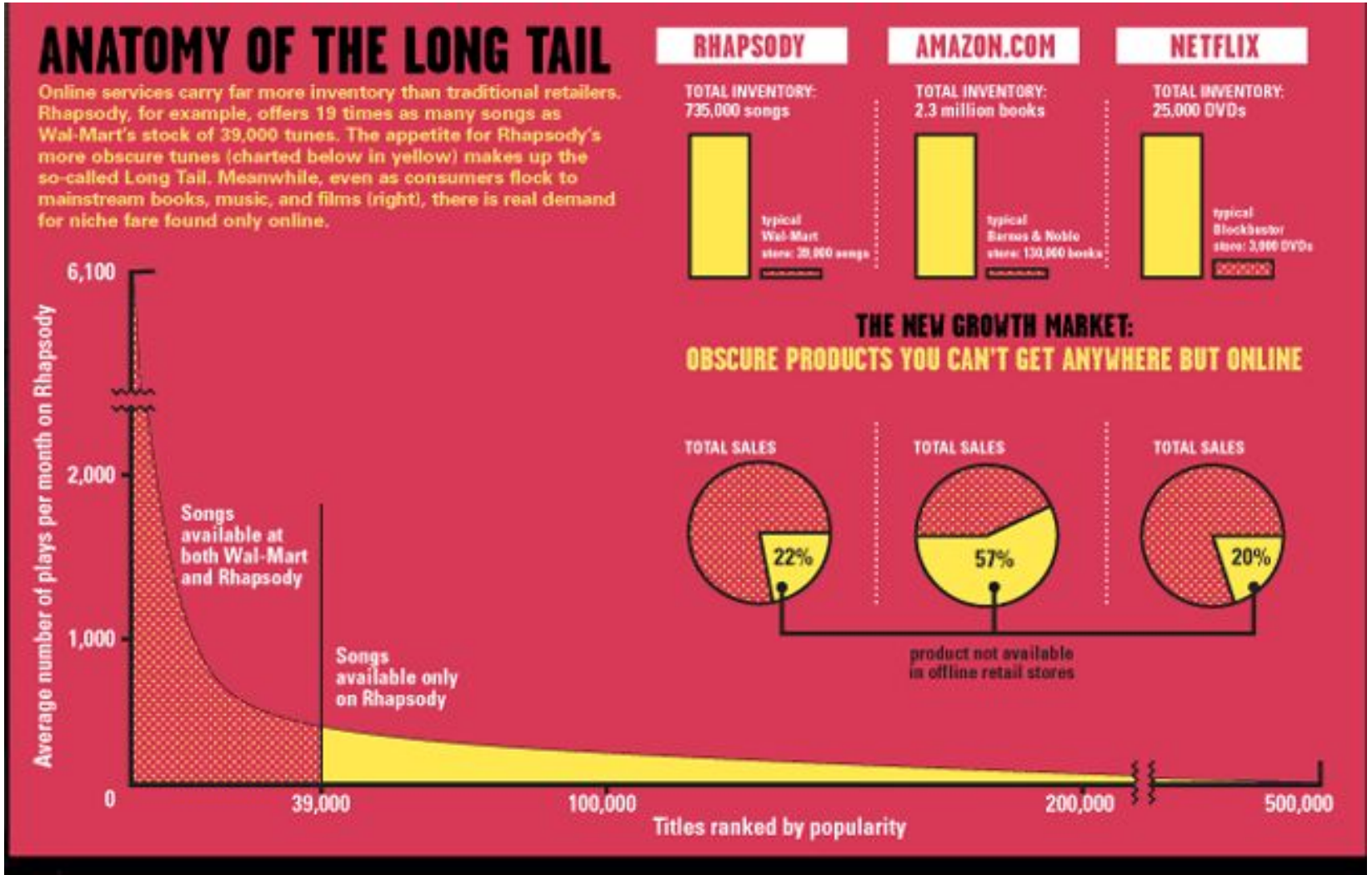

EVERY
1 MINUTE
OF THE
DAY
PRESENTED BY DOMO
TWITTER
USERS POSTS
575k
TWEETS
TIKTOK
USERS WATCH
167M
VIDEOS
GOOGLE
CONDUCTS
5.7M
SEARCHES
DISCORD
USERS SEND
668k
MESSAGES
12M
PEOPLE SEND AN
IMESSAGE
INSTAGRAM
USERS SHARE
65k
PHOTOS
CLUBHOUSE
CREATES 208 ROOMS
FACEBOOK
USERS SHARE 240k PHOTOS
SNAPCHAT
USERS SEND 2M SNAPCHATS
FACEBOOK
LIVE RECEIVES 44M VIEWS
AMAZON
CUSTOMERS SPEND $283k
YOUTUBE
USERS STREAM
694k
HOURS
6M PEOPLE SHOP
ONLINE
NETFLIX
USERS STREAM
452k
HOURS
STRAVA
ATHLETES SHARE
1.5k
ACTIVITIES
TEAMS
CONNECTS
100k
USERS
ZOOM
HOSTS
856
MINUTES OF WEBINARS
SLACK
USERS SEND
148k
MESSAGES
VENMO
USERS SEND
$304k
INSTACART
USERS SPEND
$67k

Plucking the diamond from the waste

- "**Credit-card companies** monitor every purchase and can identify fraudulent ones with a high degree of accuracy, using rules derived by crunching through billions of transactions."
  - Also check https://www.bloomberg.com/news/articles/2018-08-30/google-and-mastercard-cut-a-secret-ad-deal-to-track-retail-sales
- "**Mobile-phone operators**, meanwhile, analyse subscribers' calling patterns to determine, for example, whether most of their frequent contacts are on a rival network."
- "[…] **Cablecom**, a Swiss telecoms operator. ***It has reduced customer defections from one-fifth of subscribers a year to under 5% by crunching its numbers.***"

- "***Retailers, offline as well as online, are masters of data mining.***"

The Long Tail



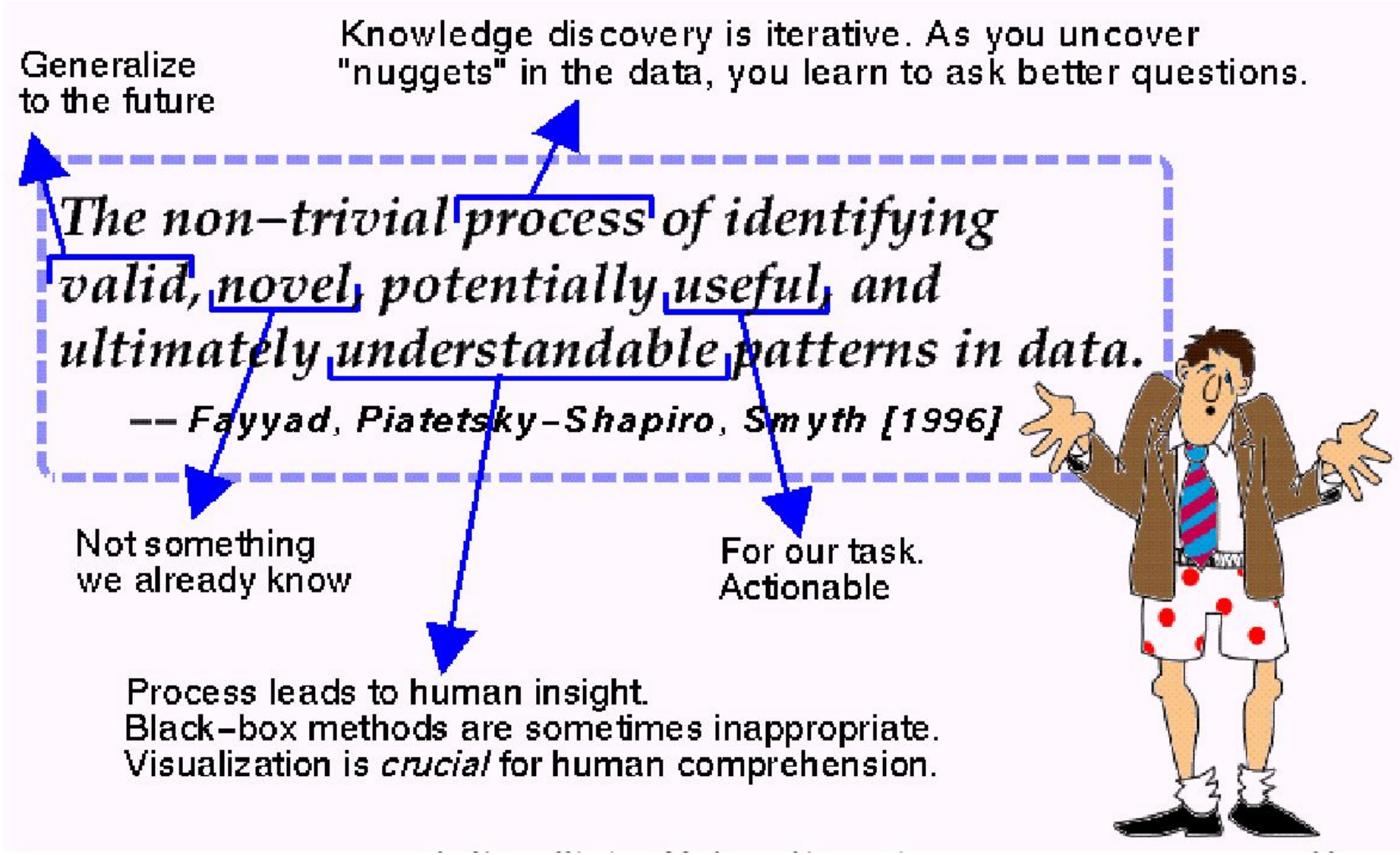
The Data Deluge

- "[…] mankind created 150 exabytes (billion gigabytes) of data in 2005. This year, it will create 1,200 exabytes.

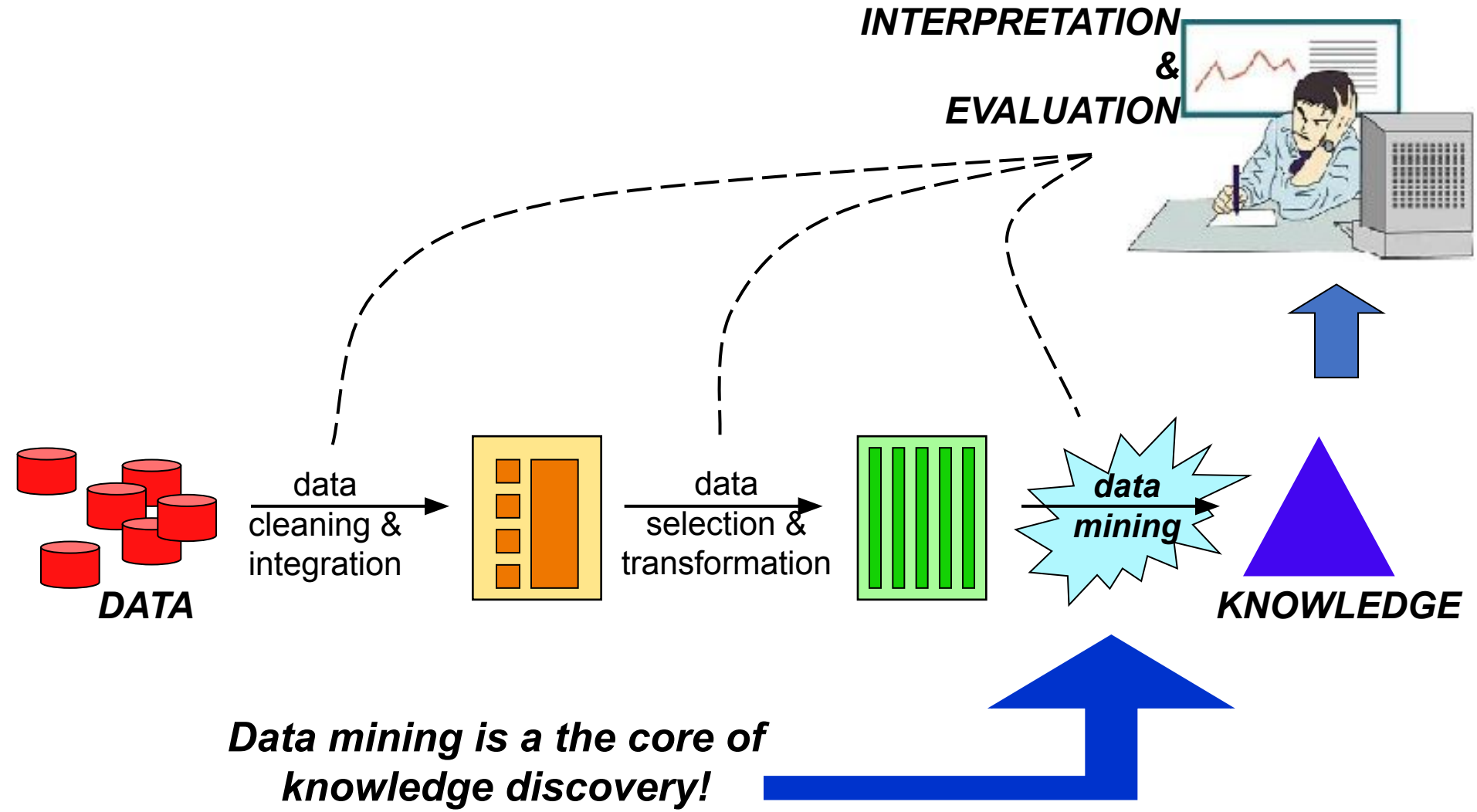
  Merely keeping up with this flood, and storing the bits that might be useful, is difficult enough.
  Analysing it, to spot patterns and extract useful information, is harder still."

  The Economist, Feb 2010

Knowledge Discovery in Database

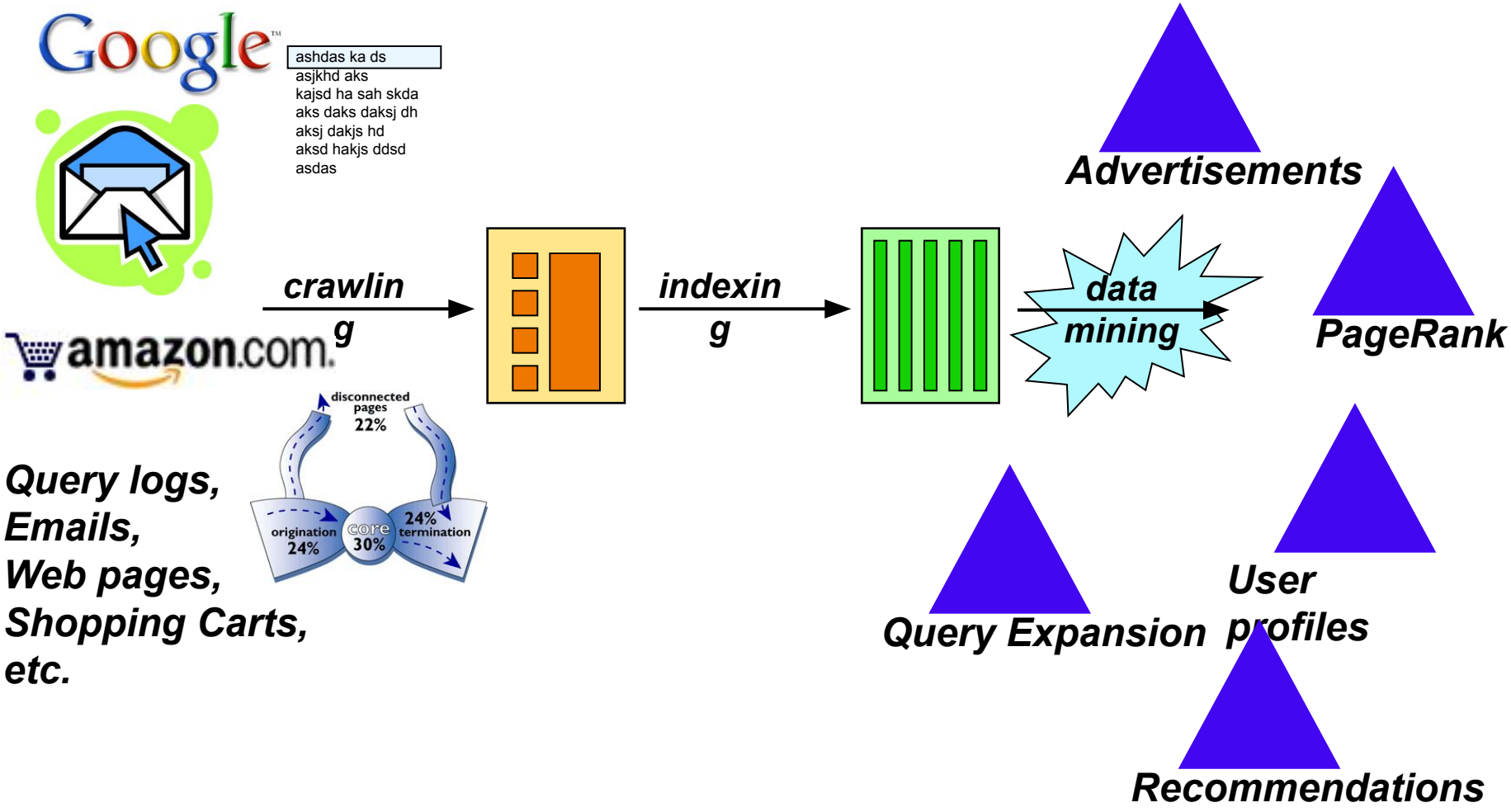


The KDD process



Data mining is a the core of knowledge discovery!

In Web search there are plenty of KDD processes



The Netflix Contest

- Netflix Company
  - DVD rental and video streaming
- Cinematch:
  - "Its job is to predict whether someone will enjoy a movie based on how much they liked or disliked other movies"
  - "To qualify for the **$1,000,000** Grand Prize, the accuracy of your submitted predictions on the qualifying set must be at least 10% better than the accuracy Cinematch"
  - "Contest begins October 2, 2006"
  - Winners awarded on September 21, 2009.

Yahoo! Learning to Rank Challenge

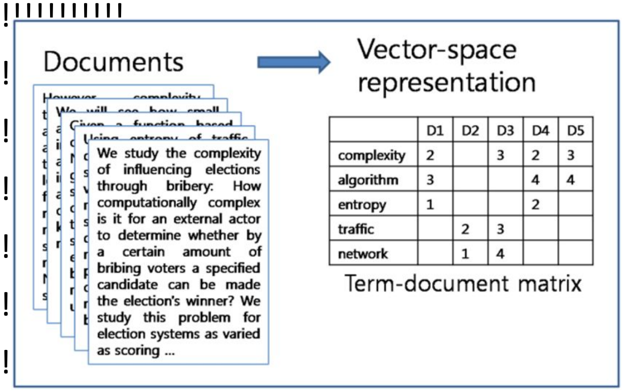
- 2010 scientific contest (no money)
- One of the datasets included:
  - 20K queries, 470K document, 519 features
  - This means ~24 candidate results per query
  - Each document has a label in [0,4]
- The goal is to predict the correct ranking on an unseen test set

- In this course, we will study the winning algorithm of this contest.

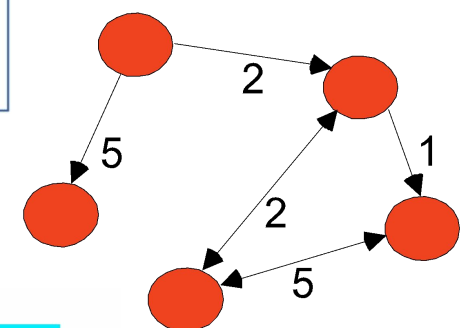
Which kind of data can we mine ?

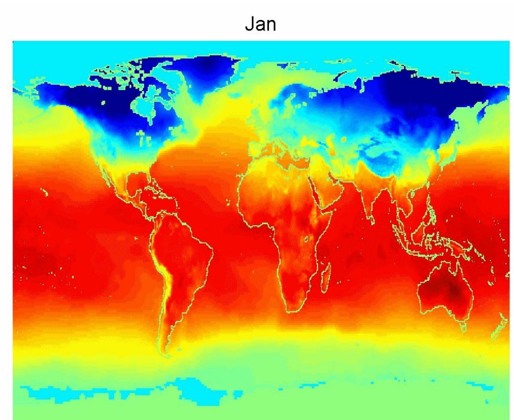
Tid	Home Owner	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- ALL KINDS !!!
- ALL KINDS !!
- ALL KINDS !!
- ALL KINDS !!
- ALL KINDS !!
- ALL KINDS !!
- ALL KINDS !!
- ALL KINDS !!!
- ALL KINDS !!!
- ALL KINDS !!!
- ALL KINDS !!!
- ALL KINDS !!!
- ALL KINDS !!!
- ALL KINDS !!!

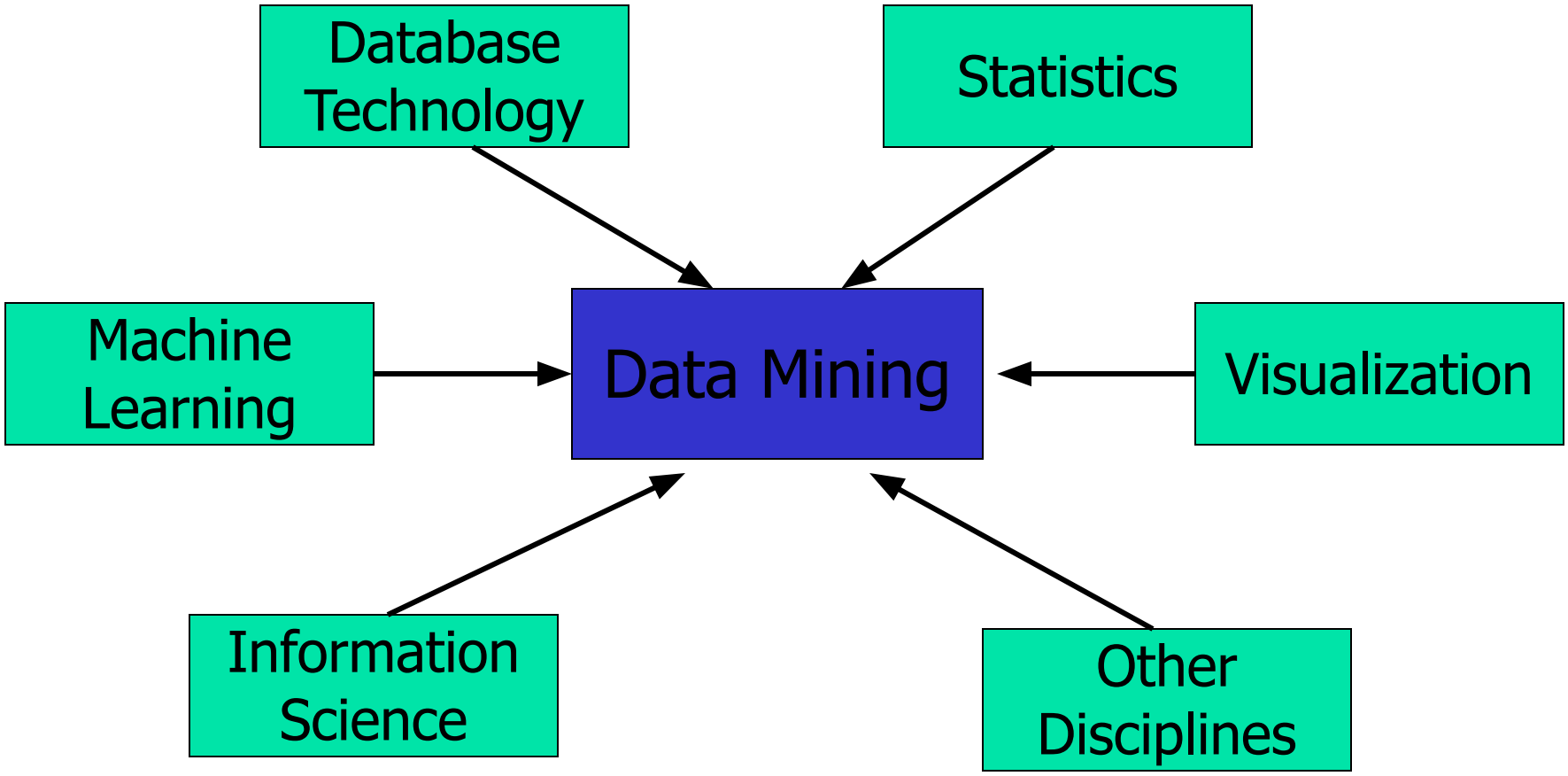


	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		





KDD is the meeting point of several disciplines



Data Mining allows for a new kind of data analysis

- Data-driven
  - The computer, i.e., the DM algorithms, generates and tests million of hypothesis and presents the bests ones
  - New knowledge is extracted automatically with a smaller contribution from the analyst, and may generate novel and unexpected knowledge

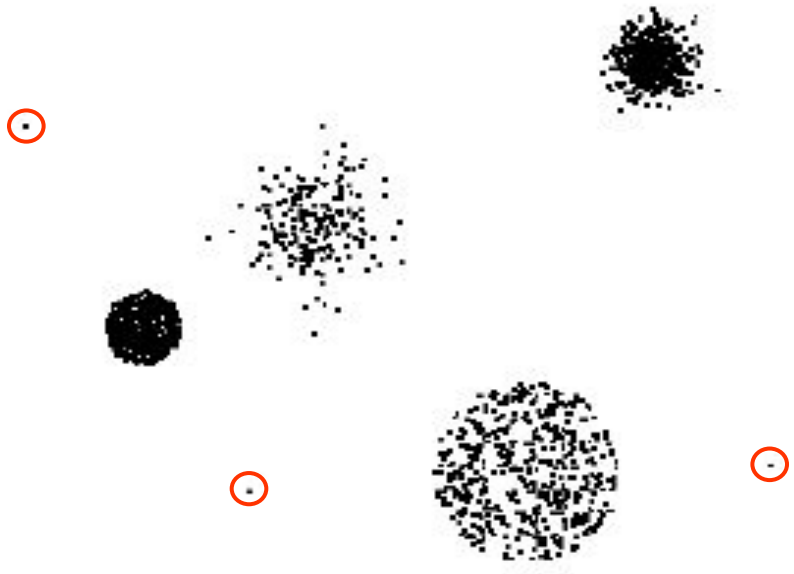
- Example:
  - A big company (e.g., Amazon, Spotify), exploits a data mining algorithm to find groups of users with similar interests. Those users are likely to purchase the same items. An item purchased by a user can be recommended to the other users in the same group. (profiling for recommentation)

Data Mining versus Statistics

- Statistics:
  - Primary analysis: data is collected for a specific analysis, the design of the data collection is part of the process
  - Random samples, statistically significant samples
  - Usually small amounts of data
  - Statistical significance
- Data mining:
  - Secondary analysis: analysis is run on data usually collected for another purpose
  - Convenience sample
  - Large amounts of data
  - Other measures of interest (including human understandability)

Data Quality Issues

- Examples:
  - Noise: original values are altered
  - Missing:
  - Duplicate data
    - Solutions:
      - Removal of noisy data,
      - Estimate missing values,
      - Discard duplicates.
  - Outliers
    - Objects considerably different from the others

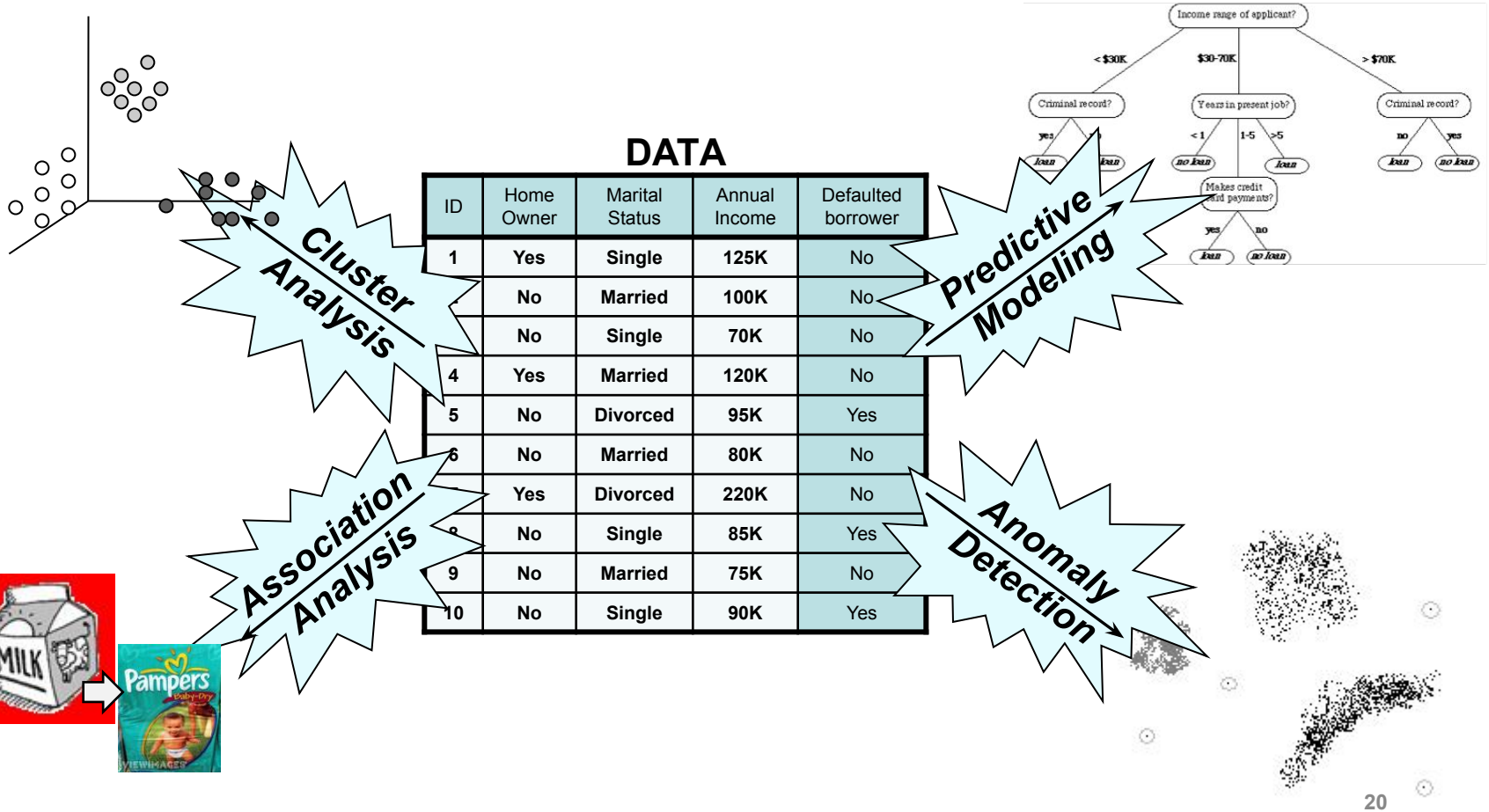


Typologies of DM tasks

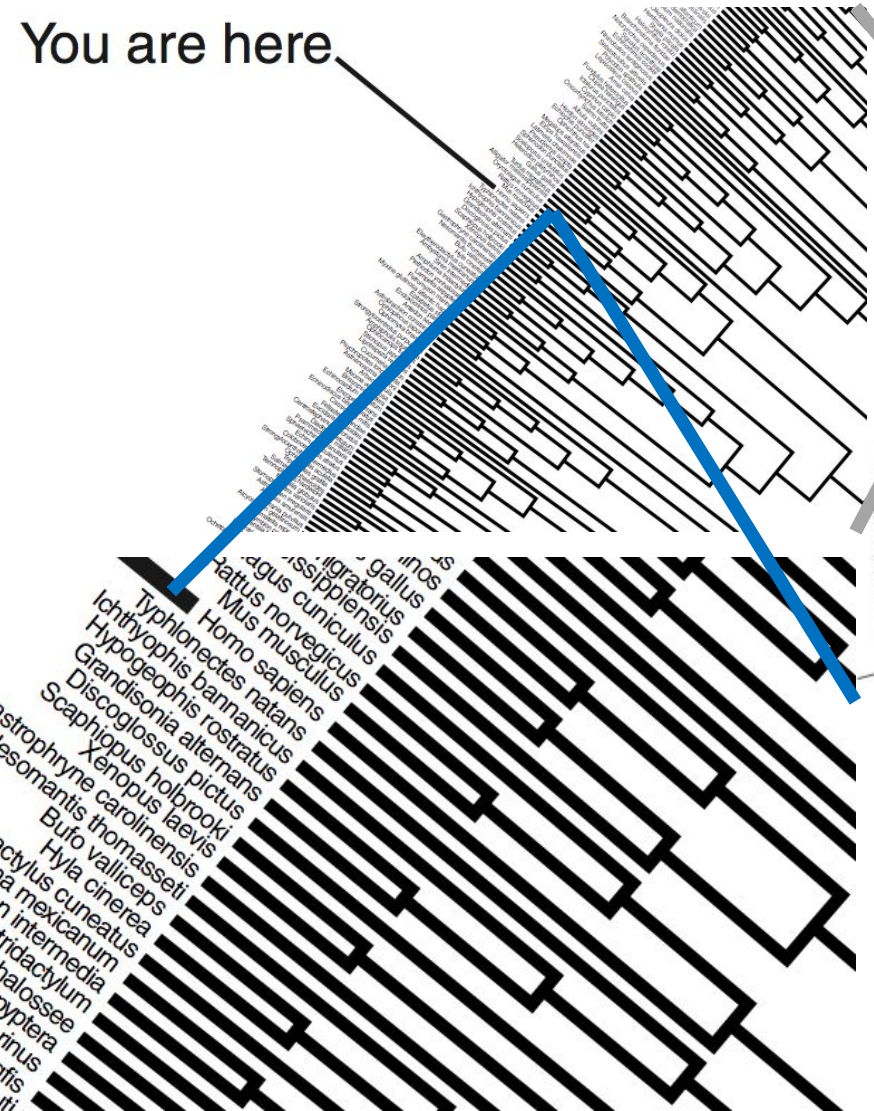
- **Prediction Methods**
  - Use some variables to predict unknown or future values of other variables.

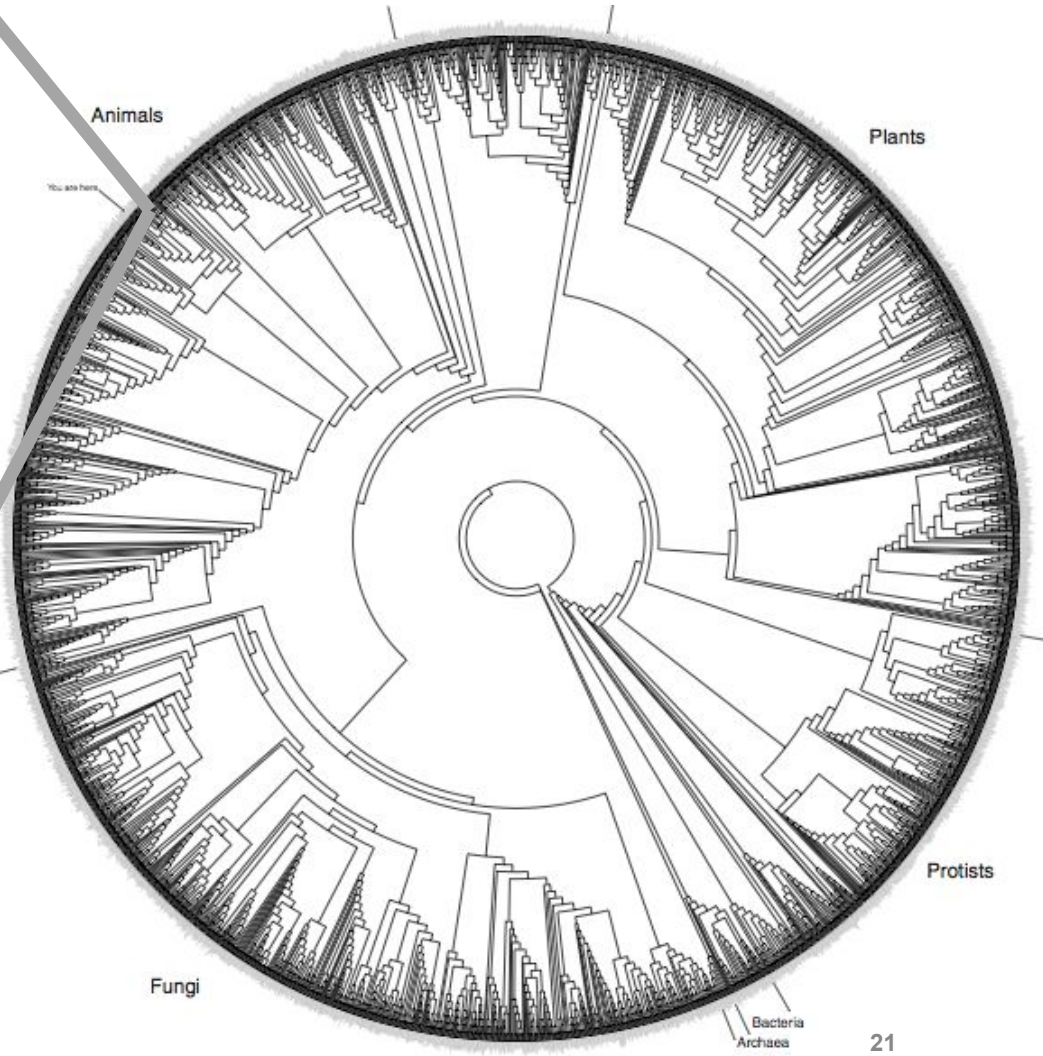
- **Description Methods**
  - Find human-interpretable patterns that describe the data.

Data Mining tasks



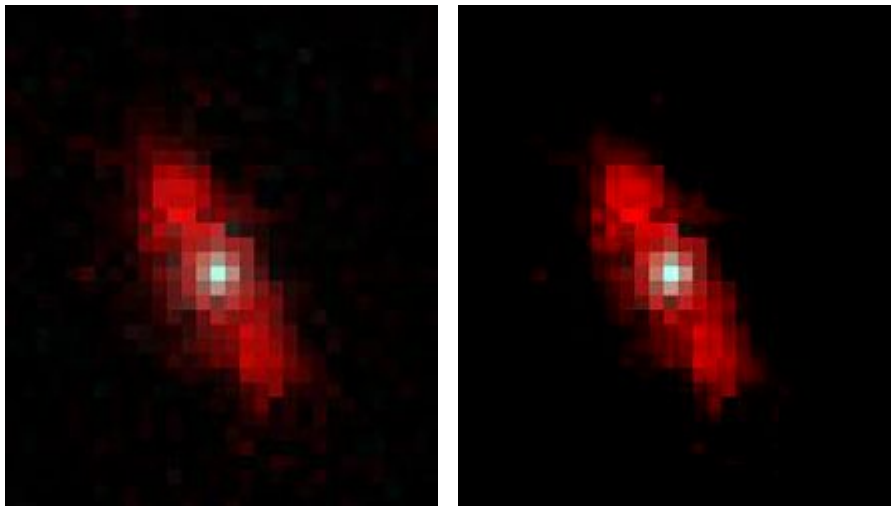
Phylogenetic trees



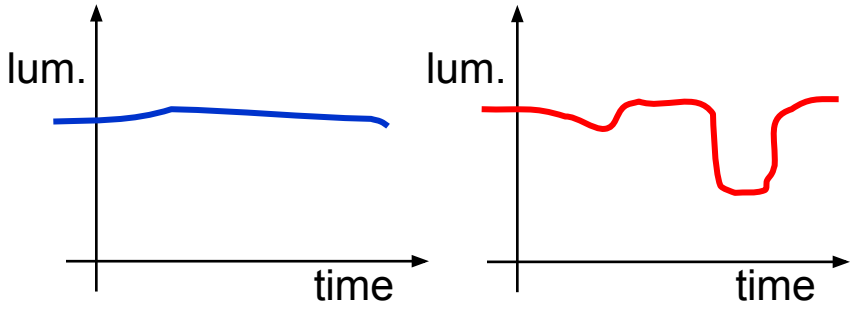


An example from astronomy

- Input:
  - A large image database coming from radar measurements of stars luminosity over time
- Data Mining:
  - clustering: segmentation in groups of similar elements
- Output:
  - A group of images unexpectedly different from the others…

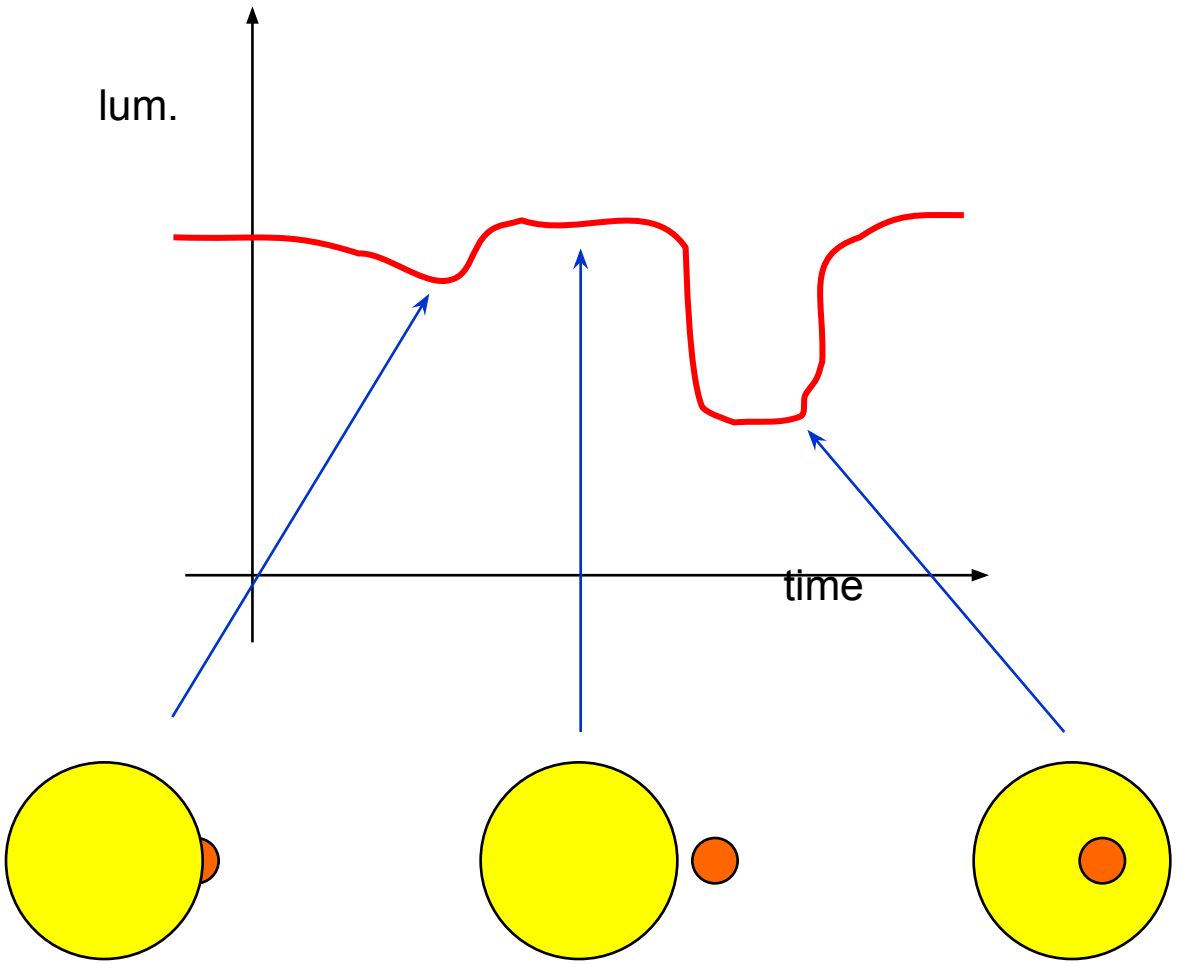


A B



A B

An example from astronomy

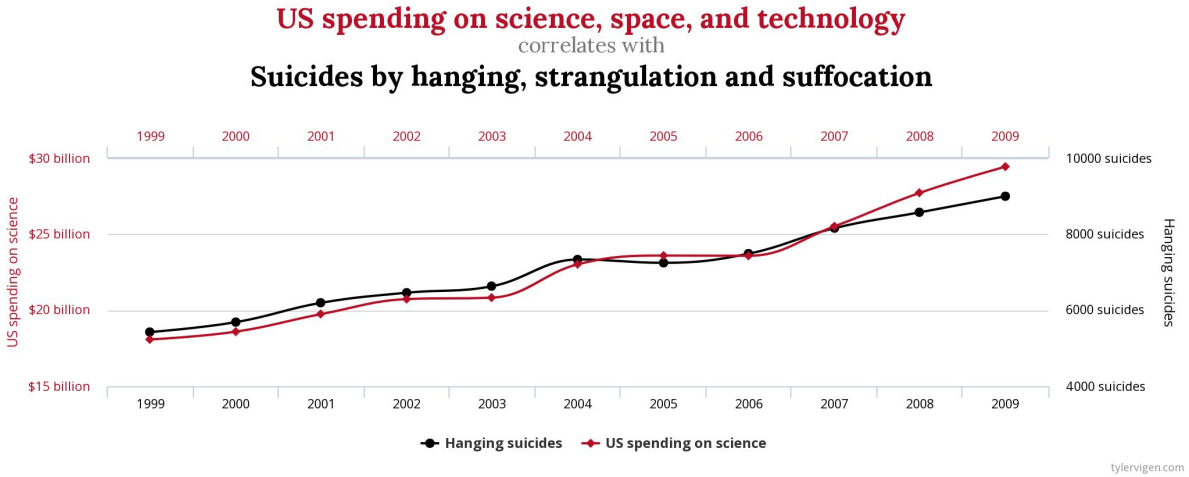


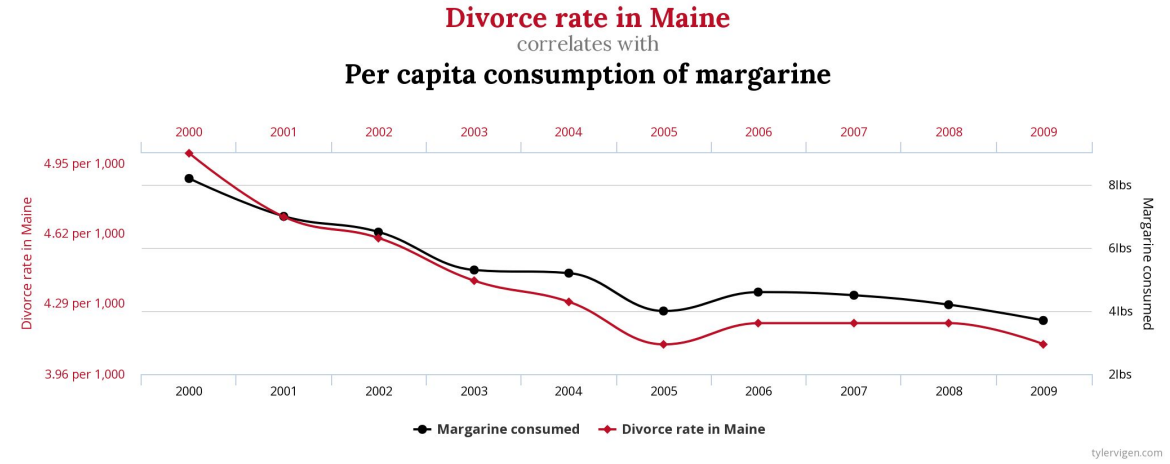
Clustering found binary stars

The motto

- "Data! Data! Data!" he cried impatiently.
  "I can't make bricks without clay."

Correlation vs. Causation





- Both are 99% correlated!
  - Check http://www.tylervigen.com/spurious-correlations for more

Written + Oral Exam

- Written Exam
  - ~ 6 questions/exercises about the notions, methods and concepts discussed
- Oral Exam: Lab project discussion
  - Only if written exam is sufficient
  - *Next Week!*
  - Groups are allowed
    - larger groups = more work
  - To be delivered
    - Jupyter/Colab Notebooks and 5-page report via Moodle
  - Evaluation
    - Quality of the report, quality of the code, number of methods experimented, depth of the analysis
  - Deadline
    - Same day of the written exam

Want to start?

try