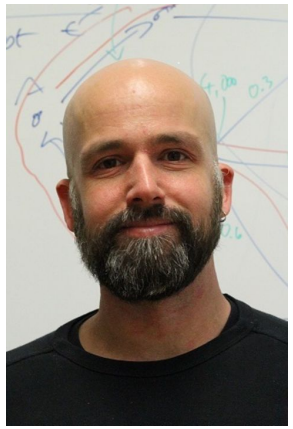
UNIVERSITÄT MANNHEIM

Data Mining I
Introduction and Course Outline

Heiko Paulheim

Hello



- Prof. Dr. Heiko Paulheim
  - Chair for Data Science
- Research Interests:
  - Knowledge Graphs on the Web and their Applications
  - Data Quality and Data Cleaning on Knowledge Graphs
  - Using Knowledge Graphs in Data Mining
  - Societal Impact of Artificial Intelligence
- Room: B6 26, B0.22
- Consultation: Tuesdays 9-10
  - Please make an appointment with Bianca Lermer upfront
- Heiko will teach the lectures

Hello

- M.Sc. Nicolas Heist
- Graduate Research Associate
- Research Interests:
  - Semantic Web Technologies
  - Knowledge Graphs and Linked Data
- eMail: nico@informatik.uni-mannheim.de
- Nico will teach the noon exercise
  and co-supervise the team projects.

Hello

- M.Sc. Ralph Peeters
- Graduate Research Associate
- Research Interests:
  - Entity Matching using Deep Learning
  - Product Data Integration
- eMail: ralph.peeters@uni-mannheim.de
- Ralph will teach the 1:45 pm exercise and co-supervise the team projects.

Hello

- M.Sc. Andreea Iana
- Graduate Research Associate
- Research Interests:
  - Recommender systems
  - Natural language processing
  - Information Retrieval
- eMail: andreea.iana@uni-mannheim.de
- Andreea will teach the 3.30 pm exercise and co-supervise the team projects.

Introduction and Course Outline

- Course Outline and Organization
- What is Data Mining?
- Methods and Applications
- The Data Mining Process

Course Organization

- Lecture
  - introduces the principle methods of data mining
  - discusses how to evaluate generated models
  - presents practical examples of data mining applications from the corporate and Web context
- Exercise
  - students experiment with data sets using Python
- Project Work
  - teams of five students realize a data mining project
  - teams may choose their own data sets and tasks (in addition, we will propose some suitable data sets and tasks)
  - write summary about project, present project results
- Final grade
  - 75 % written exam
  - 25 % project work (20% report, 5% presentation)

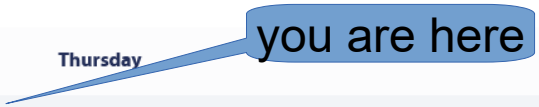
Warning: handing an empty/crossed out exam might still make you pass (with a bad grade)

Important Note on the Exam

- There is only **one** exam per semester
  - i.e., no retake date!
  - the next exam date is at the end of the upcoming FSS
- Please consider this if...
  - you are only here for one semester
  - you are planning to be abroad
  - ...
- Upon failure, you will have to redo **both** the project **and** the exam in another semester
  - unfortunately, we cannot carry over your project mark

Course Outline

Week	Wednesday	Thursday
04.09.2023	Lecture: Introduction	Exercise: Introduction to Python / Preprocessing & Visualization
11.09.2023	Lecture: Clustering	Exercise: Clustering
18.09.2023	Lecture: Classification I	Exercise: Classification I
25.09.2023	*Kick Off Team Project*	--
02.10.2023	Lecture: Classification II	Exercise: Classification II
09.10.2023	Lecture: Regression	Exercise: Regression
16.10.2023	Lecture: Text Mining	Exercise: Text Mining
23.10.2023	Lecture: Association Analysis	Exercise: Association Analysis
30.10.2023	*Team Project Feedback*	--
06.11.2023	*Team Project Feedback*	--
13.11.2023	*Team Project Feedback*	--
20.11.2023	*Team Project Feedback*	--
27.11.2023	*Results Presentation*	--
04.12.2023	*Results Presentation*	--



Deadlines

- Submission of project work proposal
  - Sunday, Oct 1st, 23:59
- Submission of final project work report
  - Friday, Dec 8th, 23:59
- Project presentations
  - schedule to be announced
  - everyone has to attend

Course Organization

- Lecture Webpage: Slides, Announcements
  - https://www.uni-mannheim.de/dws/teaching/course-details/courses-for-master-candidates/ie-500-data-mining
  - hint: look at version tags!
- Additional Material
  - ILIAS eLearning System, https://ilias.uni-mannheim.de/
- Time and Location
  - Lecture: Wednesdays, 10.15 – 11.45, SN 169
  - Exercises: Thursdays, ...
    12.00 – 13.30 (Nico), B6 26, D0.07
    13.45 – 15.15 (Ralph), B6 26, D0.07
    15.30 – 17.00 (Andreea), B6 26, D0.07
    - these are three parallel groups, you only have to attend one

Course Organization

- Registration
  - you have registered via Portal2
  - and been added to ILIAS
- There is a waiting list
  - if you decide not to continue, please resign from the course in Portal2
  - to waive your spot for people on the waiting list

Usage of ChatGPT

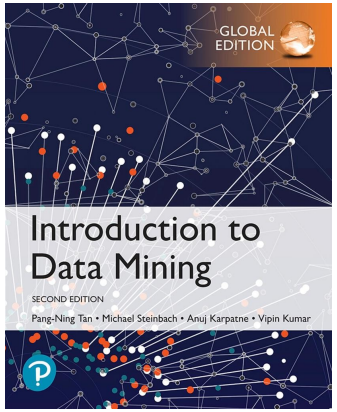
- We will experiment with using ChatGPT in the exercise to
  - discuss suitable methods and parameter settings for different use cases
  - generate and debug Python code for experimenting with the methods
  - generate multiple-choice and open questions for self-assessment

Source: New York Times

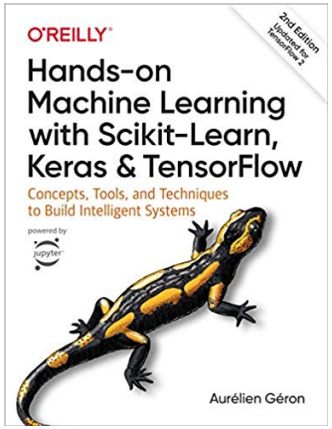
OpenAI

Literature & Slide Sources

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar:
  - Introduction to Data Mining.
    2nd Edition, Pearson / Addison Wesley.

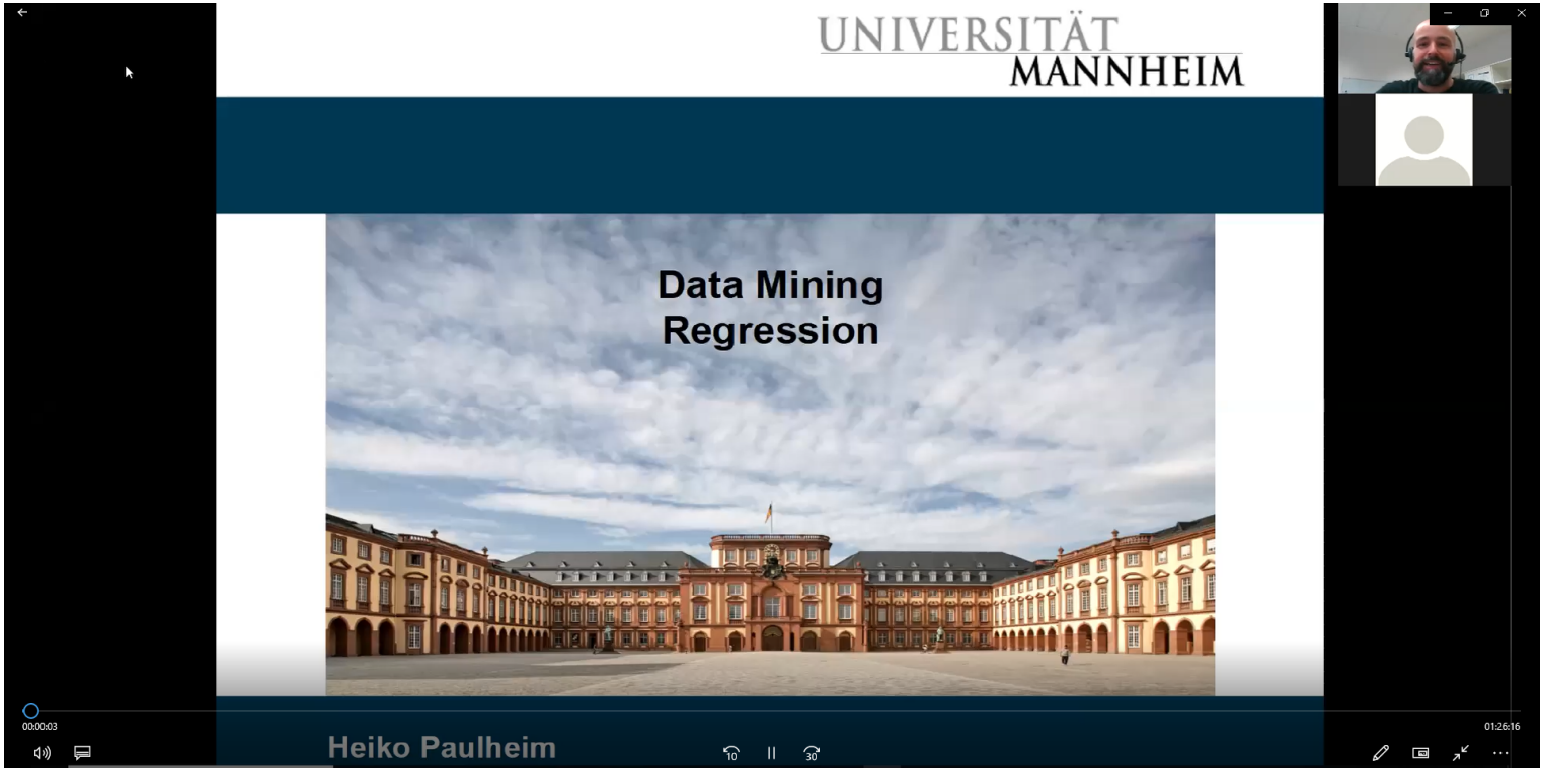


- Aurélien Géron:
  - Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow.
  - 2nd or 3rd Edition, O'Reilly



Additional Material

- Video recordings from HWS 2020
  - https://www.uni-mannheim.de/dws/teaching/lecture-videos/ (VPN!)

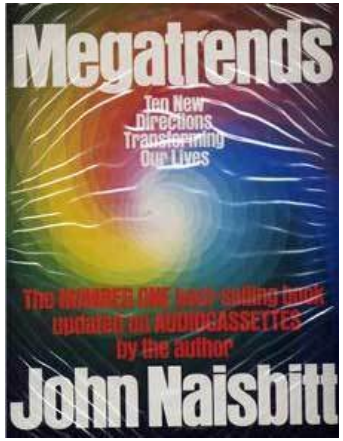


Questions?

A Bit of History

- *We are drowning in data, but starving for knowledge.*

(John Naisbitt, 1982)



- *Computers have promised us a fountain of wisdom but delivered a flood of data.*
- *It has been estimated that the amount of information in the world doubles every 20 months.*

(Frawley, Piatetsky-Shapiro, Matheus, 1992)

“We are Drowning in Data...”

More and more data is generated:

- Transaction data from banking, telecommunication, e-commerce
- Scientific data from astronomy, physics, biology
- All interactions with the Web
- Social network sites
- Application logs
- GPS tracking logs
- ...

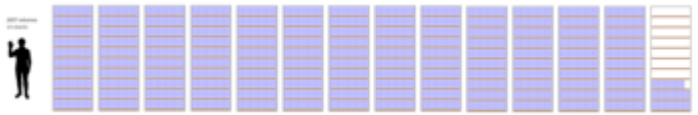
“We are Drowning in Data...”

The following slides are taken from Aidan Hogan's course on “Massive Data Processing”

WIKIPEDIA
The Free Encyclopedia

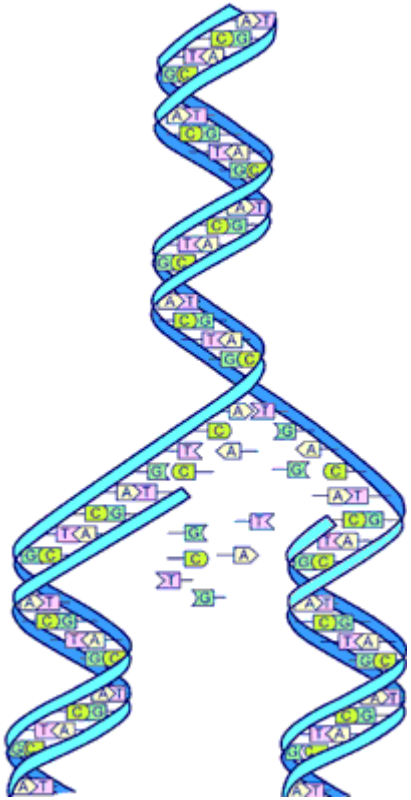
Wikipedia (en, text only)
≈ 22 GB of data

1 Wiki = 1 Wikipedia



https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

“We are Drowning in Data...”



Human Genome

≈ 4 GB/person

≈ 0.18 Wiki/person

≈ 1.5M Wiki/humankind

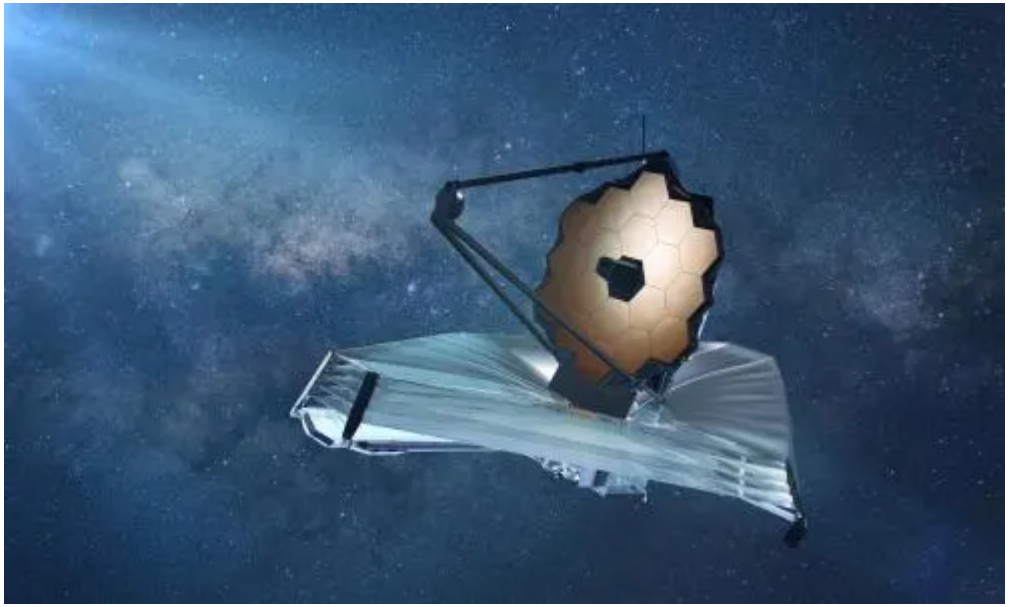
“We are Drowning in Data...”

US Library of Congress

≈235 TB archived

≈11 thousand Wiki

“We are Drowning in Data...”



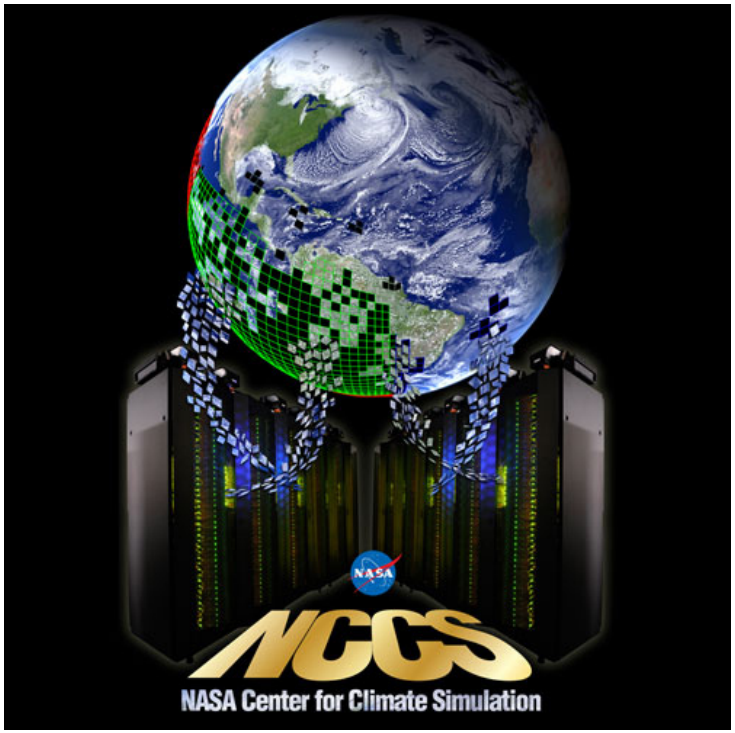
James Webb Telescope

≈57 GB/day

≈21 TB/year

≈1 thousand Wiki/year

“We are Drowning in Data...”



NASA Center for Climate Simulation

≈ 32 PB archived

≈ 1.5M Wiki

“We are Drowning in Data...”

Facebook

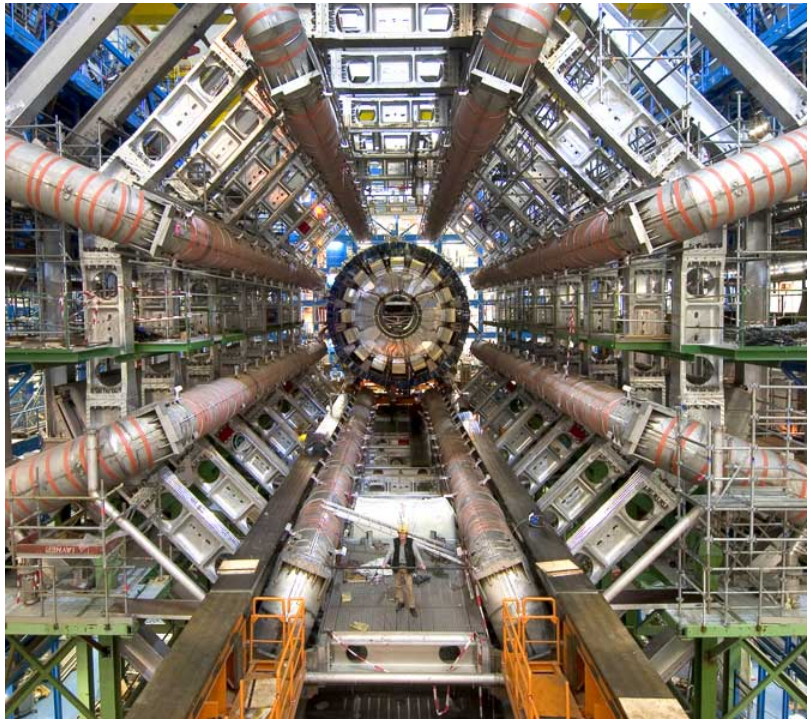
≈12 TB/day added

≈550 Wiki/day

≈200k Wiki/year

(*as of Mar. 2010*)

“We are Drowning in Data...”



Large Hadron Collider
≈15 PB/year
≈680k Wiki/year

“We are Drowning in Data...”

Google

≈20 PB/day processed

≈900k Wiki/day

≈332M Wiki/year

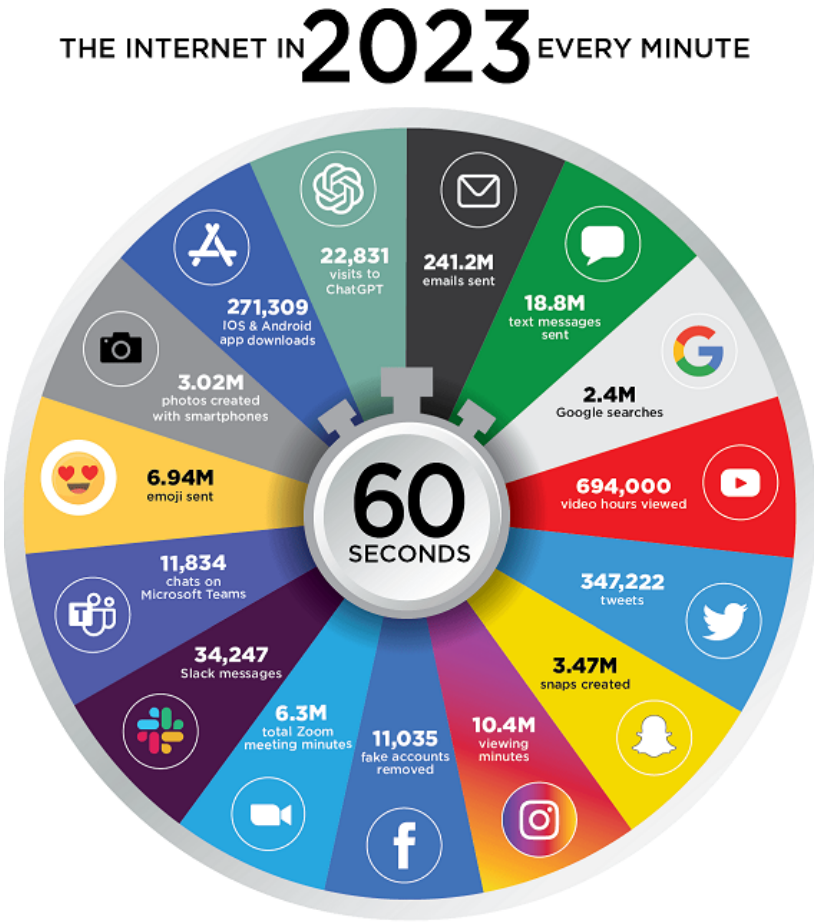
(Jan. 2010)

“We are Drowning in Data...”

Internet (2016)
≈1.3 ZB/year
≈72M Wiki/year
(2016 IP traffic; Cisco est.)

≈ 2 Wiki/second

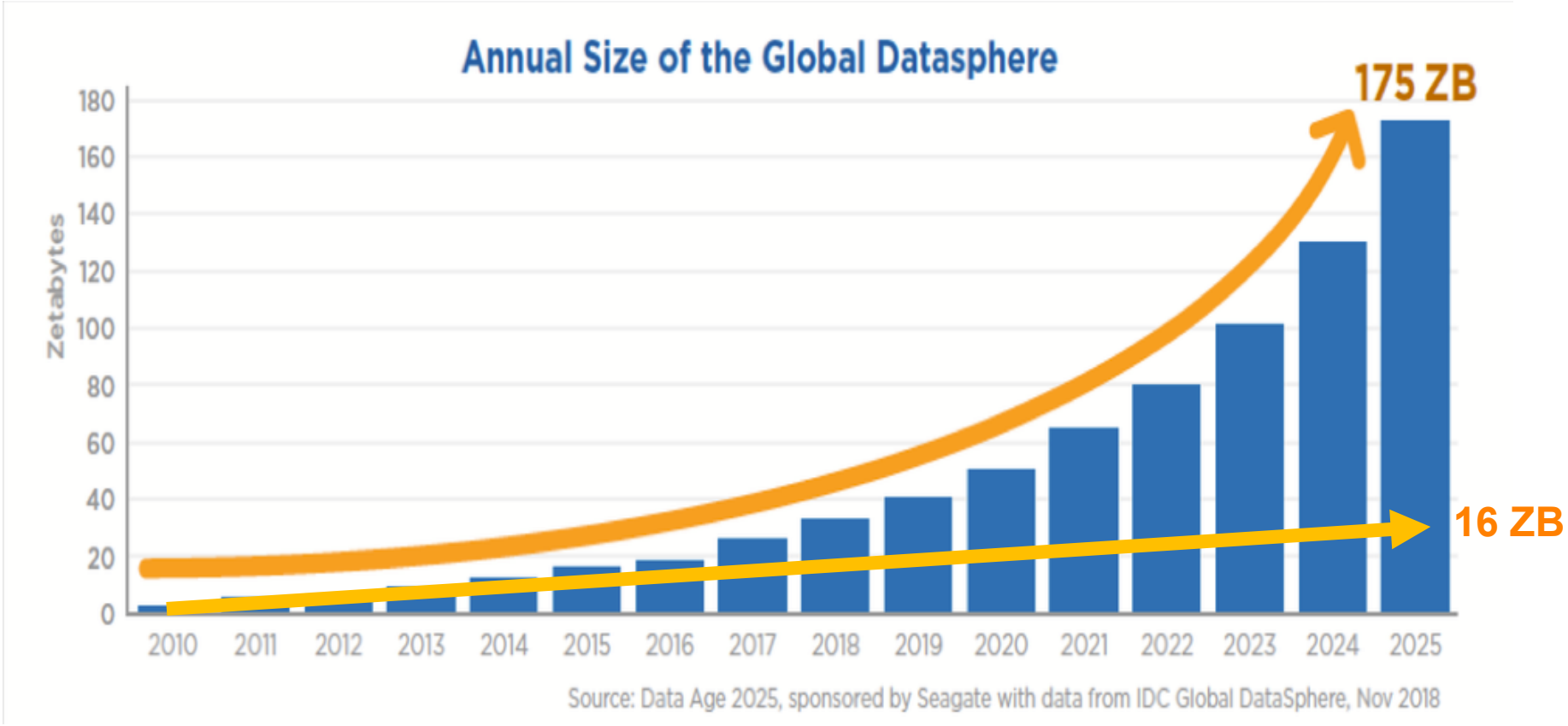
“We are Drowning in Data...”



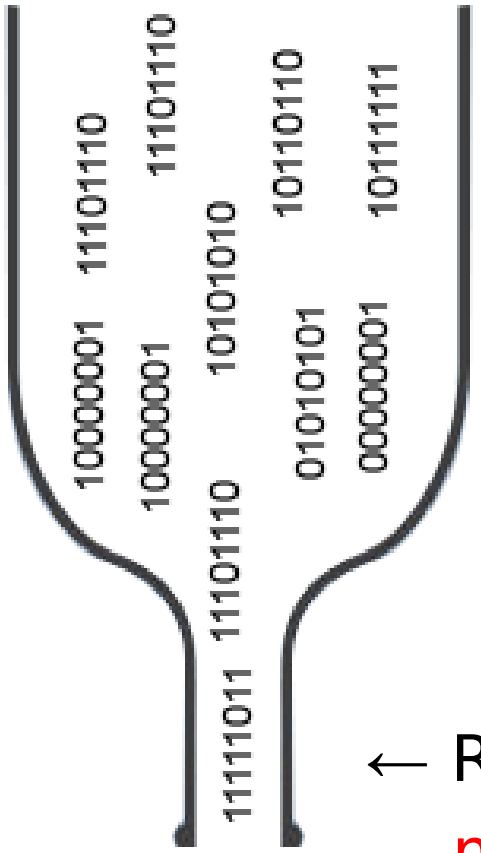
https://ediscoverytoday.com/2023/04/20/2023-internet-minute-infographic-by-ediscovery-today-and-ltmg-ediscovery-trends/

“We are Drowning in Data...”

- Slide based on Keynote on “Data Disposal by Design” given by Tova Milo at ESWC 2022:



...but starving for knowledge!

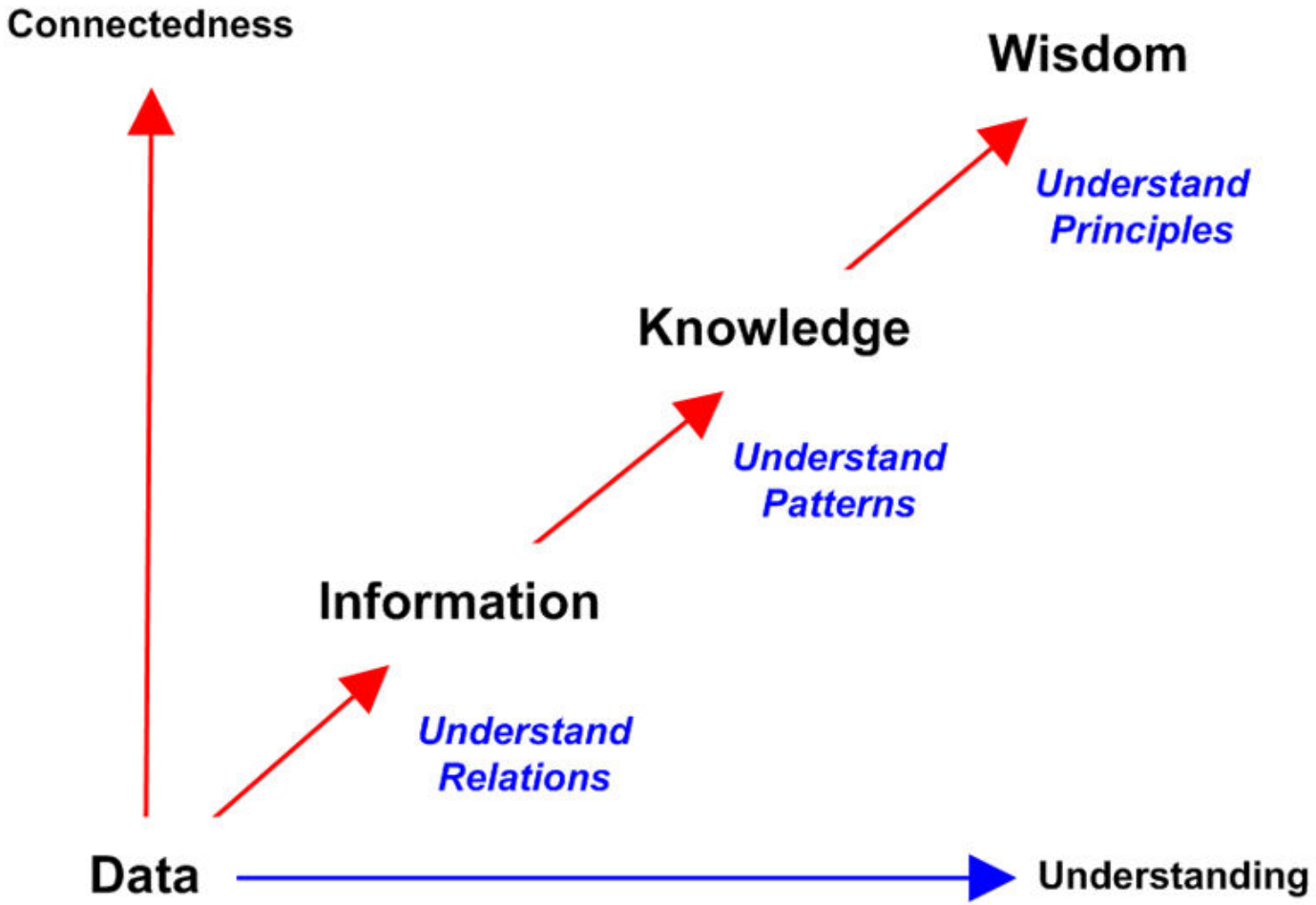


← Rate at which data are produced

← Rate at which data can be understood

manual interpretation is hardly feasible!

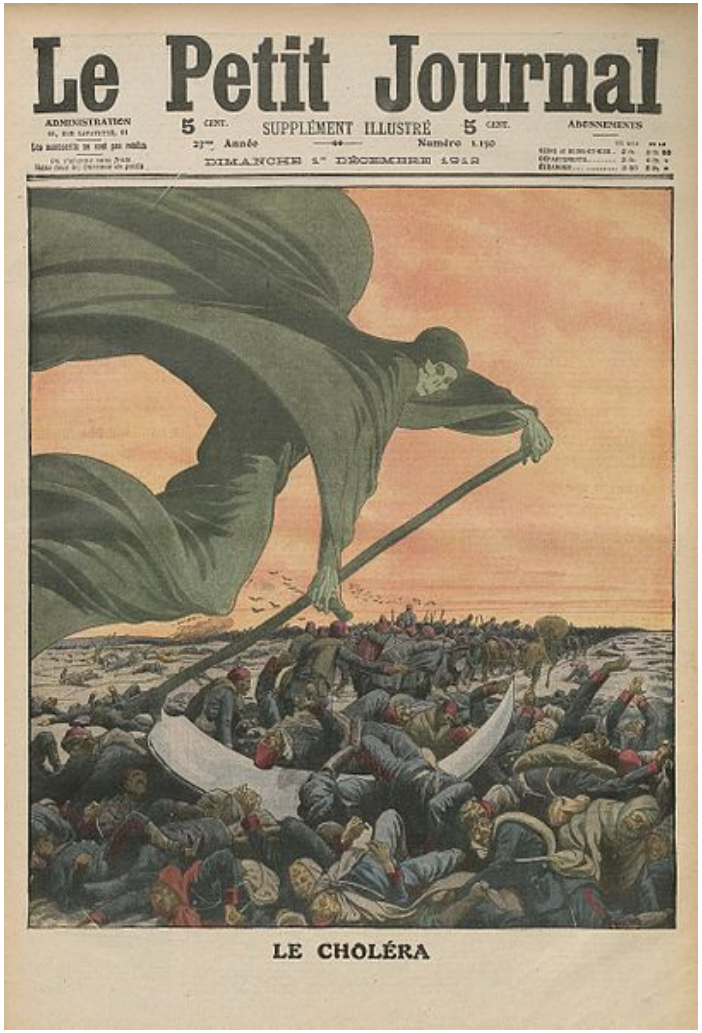
Data, Information, Knowledge, and Wisdom



Gene Bellinger, Durval Castro and Anthony Mills. "Transforming Data to Wisdom."

A Historical Example

- Cholera disease
- From beginning of 19th century
- ~100,000 deaths per year
  - until today!
- For a long time,
  there was little knowledge
  - on ways of infection
  - on causes of the disease



http://fieldnotes.unicefusa.org/2008/09/newsnet_combating_cholera_1.html

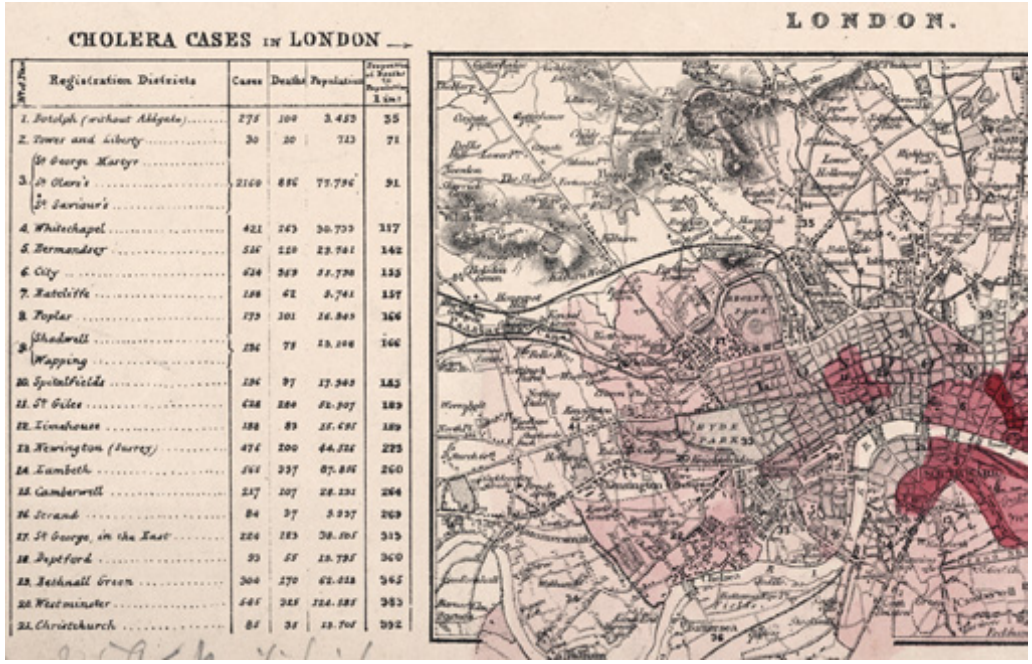
A Historical Example

- August Heinrich Petermann
- 1822-1878
- Geographer and Cartographer
- Geographic maps as a means
  - to understand data
  - to gather knowledge

http://commons.wikimedia.org/wiki/File:August_Heinrich_Petermann.jpg

A Historical Example

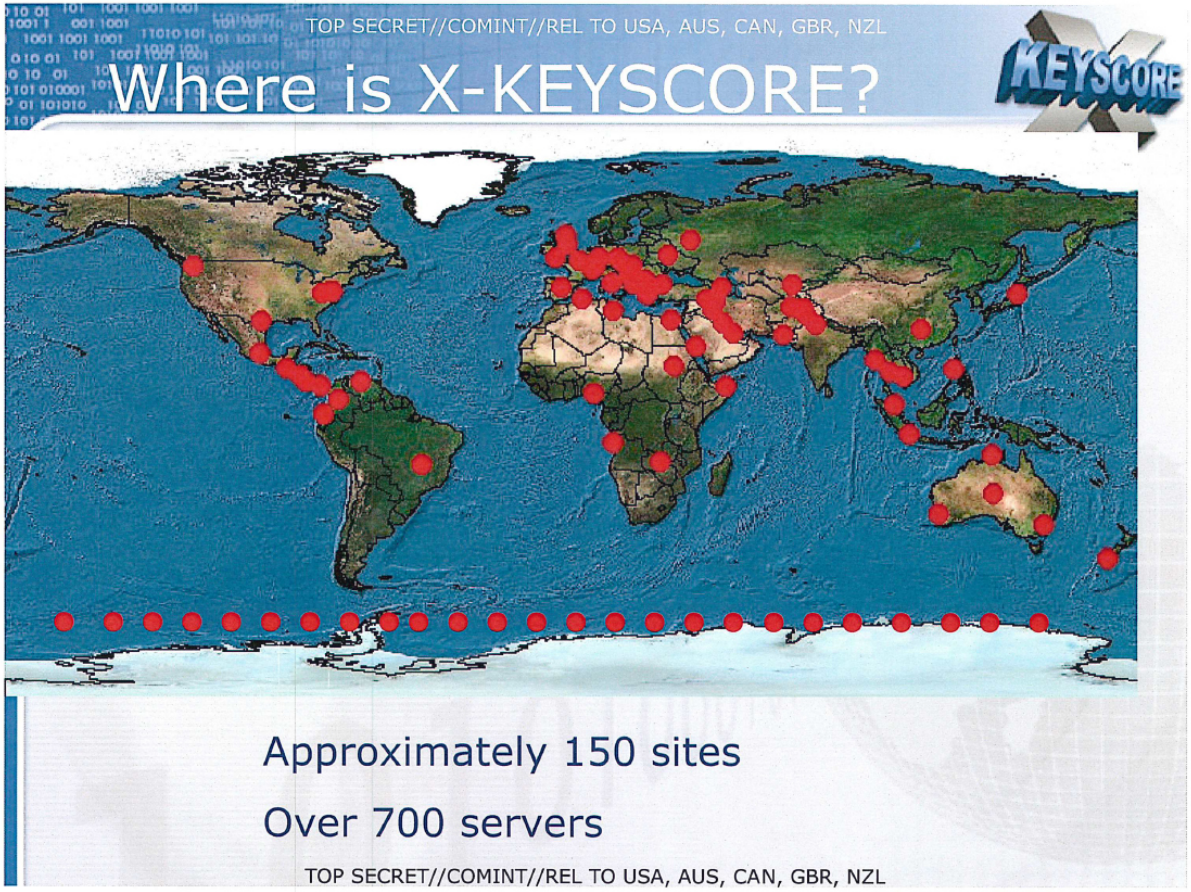
- 1848 map of Cholera deaths in London
  - finding: Cholera is more likely in densely populated areas
  - where there is no functioning sewage system
  - conclusion: Cholera is transmitted through contaminated water



http://www.dgfk.net/index.php?do=dbk&do2=1209

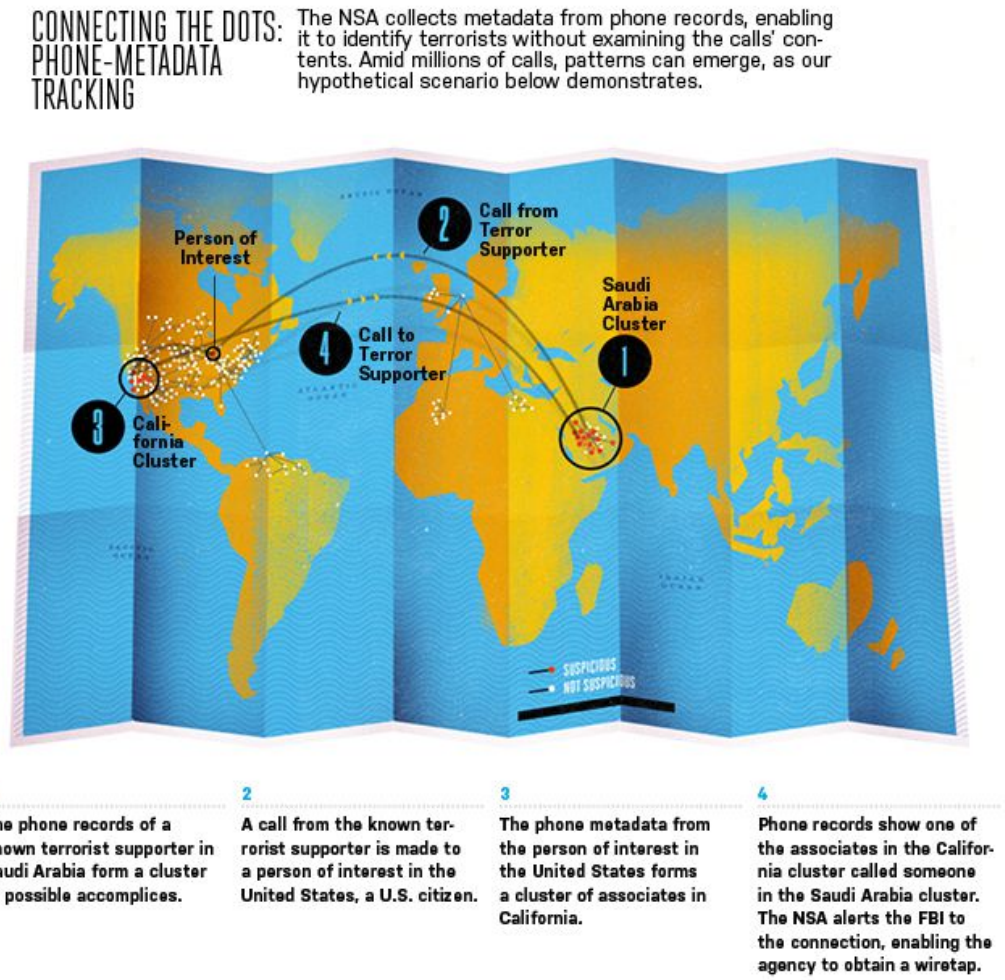
A Recent Example: the NSA

- Communication data from all over the world
- Searching for suspects and terrorists



http://www.theguardian.com/world/2013/jul/31/nsa-top-secret-program-online-data

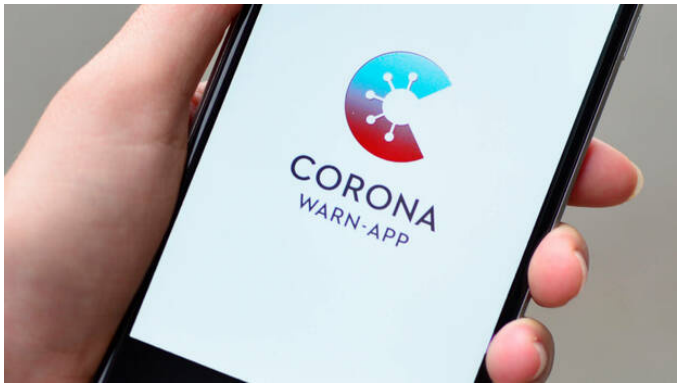
A Recent Example: the NSA



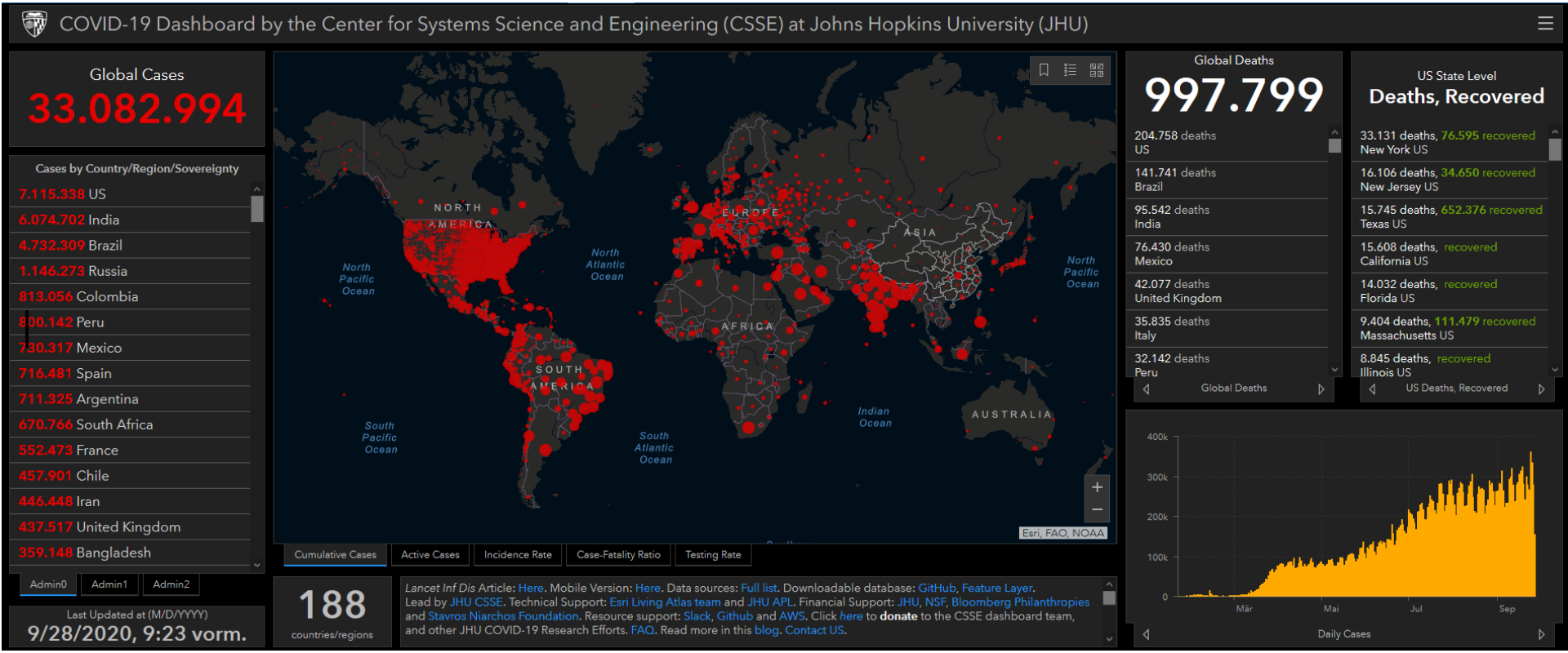
https://www.popularmechanics.com/military/a9465/nsa-data-mining-how-it-works-15910146/

A Very Recent Example: CoViD-19

- Data Mining can help understanding
  - pathways and chains of infection
  - critical preconditions of patients
    - previous diseases
    - medications
    - genetic preconditions
  - effectiveness of prevention strategies
    - e.g., famous hammer & dance paper
  - vulnerable factors in health infrastructures

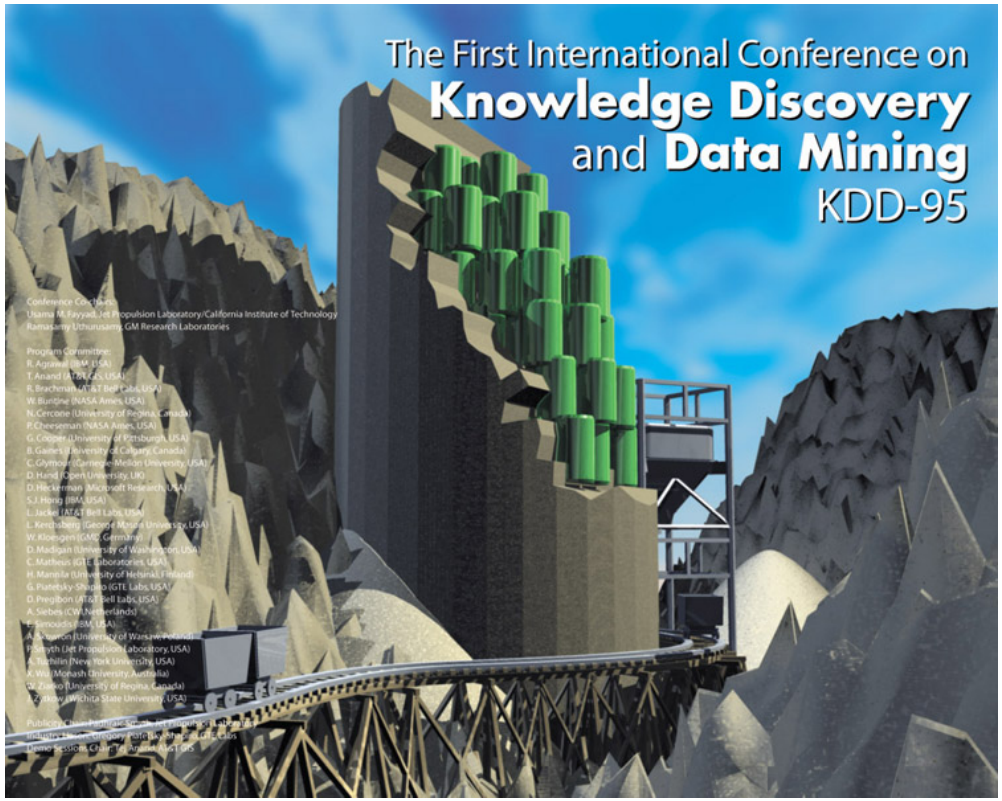


A Very Recent Example: CoViD-19



Data Mining: Definitions

- Idea: mountains of data
  - where knowledge is mined



Data Mining: Definitions

- Data Mining is a non-trivial process of identifying
  - valid
  - novel
  - potentially useful
  - ultimately understandable

  patterns in data.

(Fayyad et al. 1996)

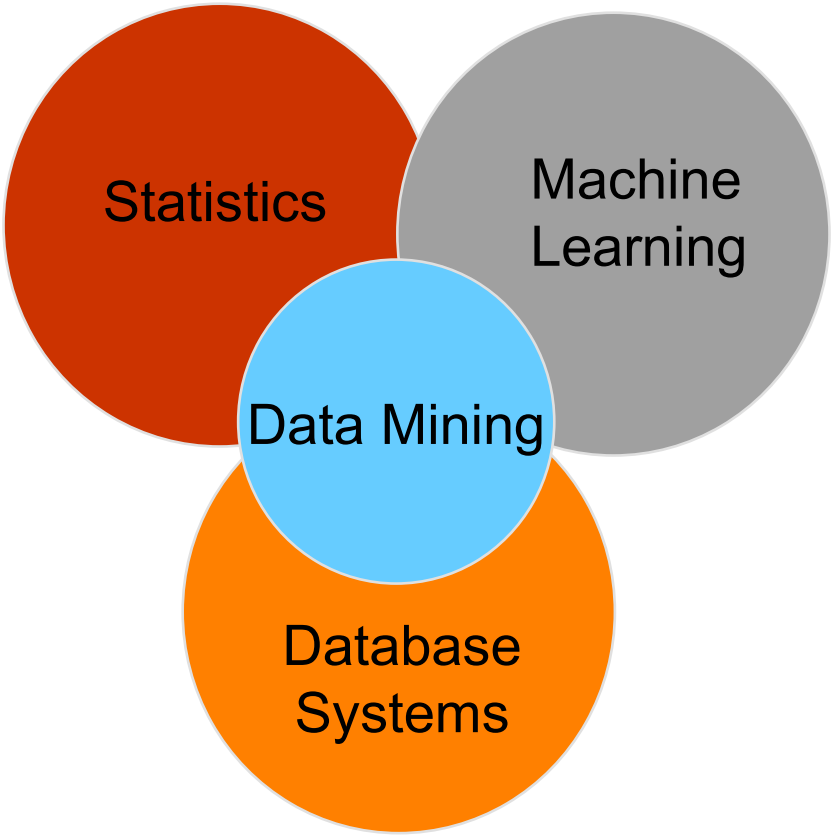
- Data mining is nothing else than torturing the data until it confesses

(Fred Menger, year unknown)

- ...and if you torture it enough, you can get it to confess to anything.

Origins of Data Mining

- Draws ideas from machine learning, statistics, and database systems.
- Traditional techniques may be unsuitable due to
  - large amount of data
  - high dimensionality of data
  - heterogeneous, distributed nature of data

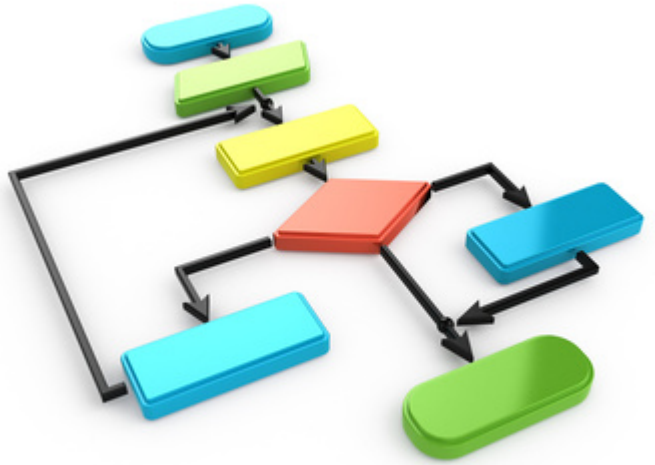


Data Mining Application Fields

- Business
  - Customer relationship management, e-commerce, fraud detection, manufacturing, telecom, targeted marketing, health care, …
- Science
  - Data mining helps scientists to analyze data and to formulate hypotheses.
  - Astronomy, physics, bioinformatics, drug discovery, …
- Web and Social Media
  - advertising, search engine optimization, spam detection, web site optimization, personalization, sentiment analysis, …
- Government
  - surveillance, crime detection, profiling tax cheaters, …

Data Mining Methods

- Descriptive methods
  - find patterns in data
  - e.g., *which products are often bought together?*
- Predictive methods
  - predict unknown or future values of a variable
    - given observations (e.g., from the past)
  - e.g., *will a person click an ad?*
    - given his/her browsing history
- Machine learning terminology:
  - descriptive = unsupervised
  - predictive = supervised



Data Mining Tasks

- Clustering (descriptive)
- Classification (predictive)
- Regression (predictive)
- Association Rule Mining (descriptive)
- Text Mining (both descriptive and predictive)
- Covered in Data Mining 2
  - Anomaly Detection (descriptive)
  - Sequential Pattern Mining (descriptive)
  - Time Series Prediction (predictive)
- Covered in other lectures
  - Textual data (in depth)
  - Image data
  - Process data

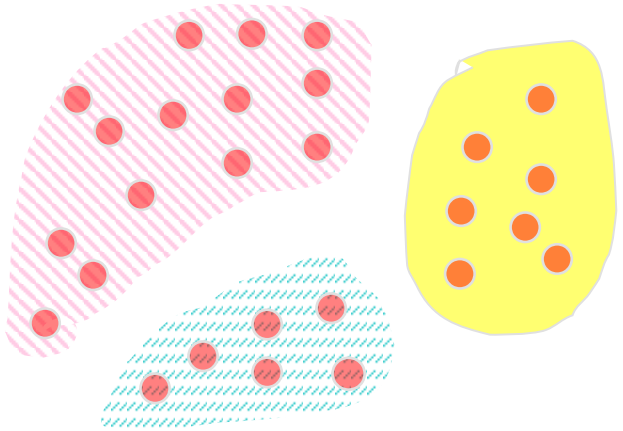
Outlook: Data Mining II

- Taught every FSS
- Topics
  - Sequential Pattern Mining, Time Series Prediction
  - Neural Networks and Deep Learning
  - Anomaly Detection
  - Online Data Analysis
  - Advanced Data Preprocessing
- Practical project
  - The annual Data Mining Cup
  - Worldwide competition of student teams
  - Real-world data mining tasks

DATA-MINING-CUP
prudsys Anwendertage und Data Mining Wettbewerb

Clustering

- Given a set of data points, and a similarity measure among them, find clusters such that
  - Data points in one cluster are similar to one another
  - Data points in separate clusters are different from each other
- Result
  - a descriptive grouping of data points

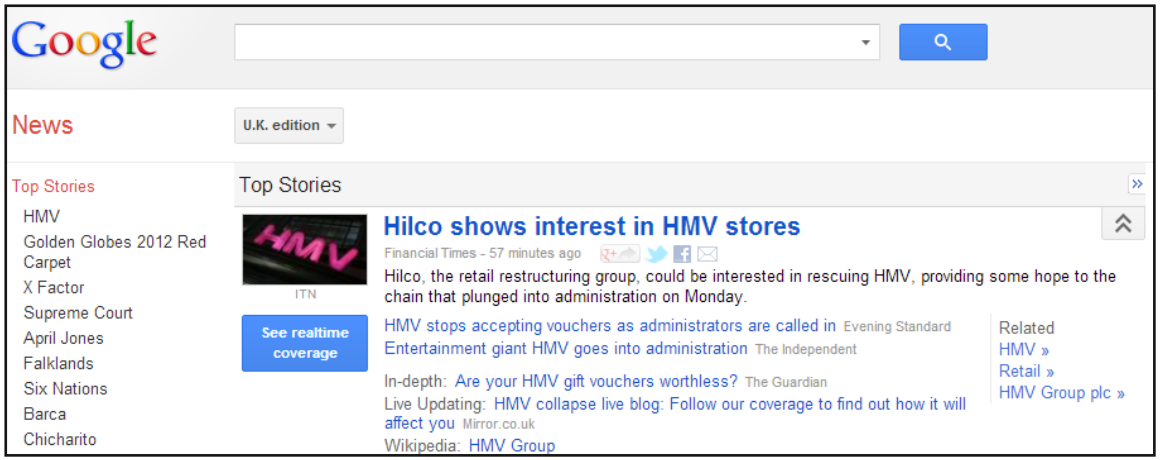


Clustering: Applications

- Application area: Market segmentation
- Goal: Subdivide a market into distinct subsets of customers
  - where any subset may be conceived as a marketing target to be reached with a distinct marketing mix
- Approach:
  - Collect information about customers
  - Find clusters of similar customers
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters

Clustering: Applications

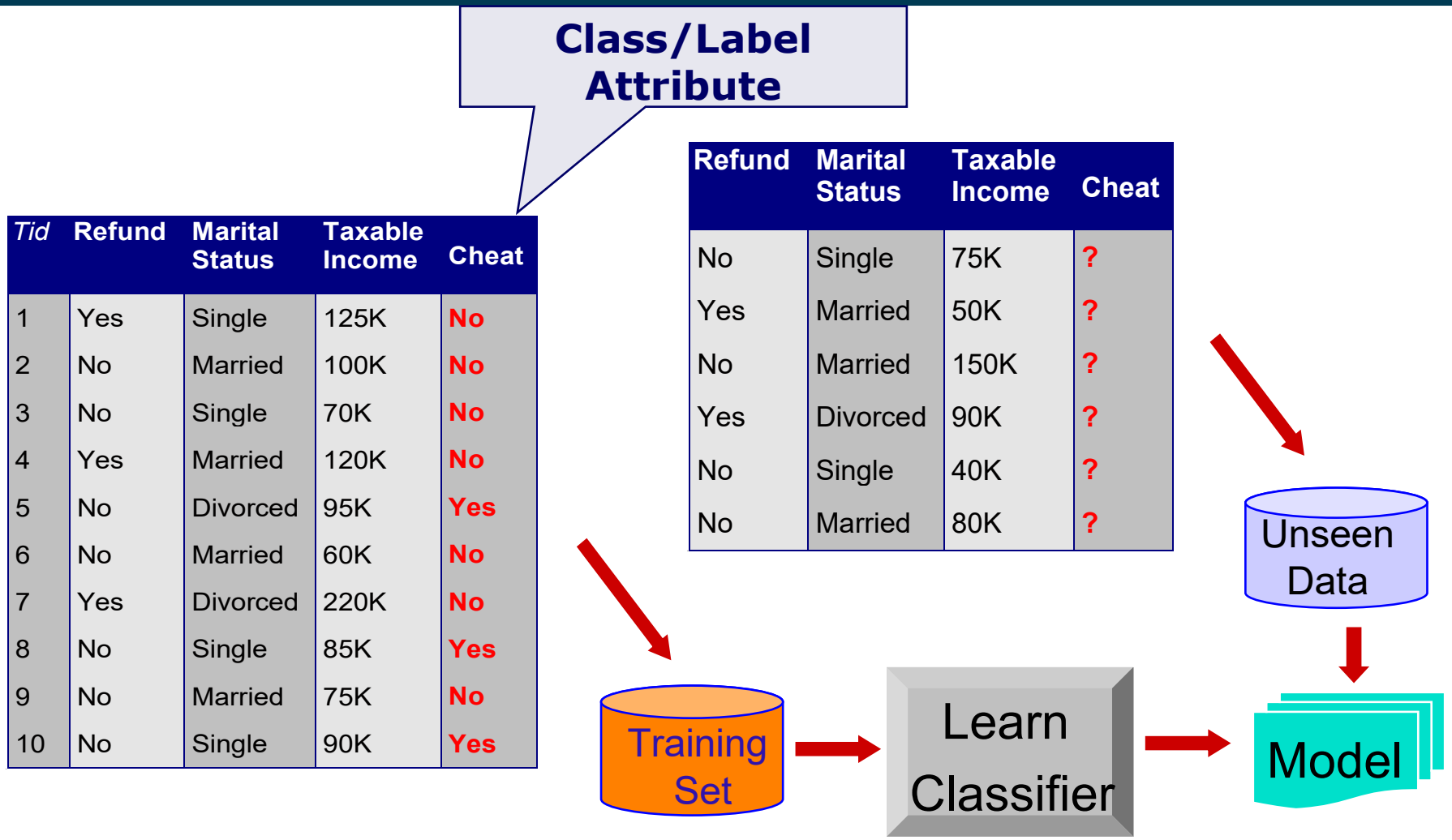
- Application area: Document Clustering
- Goal: Find groups of documents that are similar to each other based on the important terms appearing in them
- Approach
  - Identify frequently occurring terms in each document
  - Define a similarity measure based on the frequencies of different terms
- Application Example: Grouping of stories in Google News



Classification

- Given a collection of records (training set)
  - each record contains a set of attributes
  - one of the attributes is the class (label) that should be predicted
- Find a *model* for class attribute as a function of the values of other attributes
- Goal: previously unseen records should be assigned a class as accurately as possible
  - A test set is used to validate the accuracy of the model
  - Training set may be split into training and validation data

Classification Example



Classification: Applications

- Application area: Direct Marketing
- Goal: Reduce cost of mailing by targeting a set of consumers which are likely to buy a new cell phone
- Approach:
  - Use the data for a similar product introduced before
  - We know which customers decided to buy and which did not
  - Collect various demographic, lifestyle, and company-interaction related information about all such customers
    - Type of business, where they stay, how much they earn, etc.
  - Use this information as input attributes to learn a classifier model

Classification: Applications

- Application area: Fraud Detection
- Goal: Recognize fraudulent cases in credit card transactions
- Approach:
  - Use credit card transactions and the information on its account-holder as attributes
    - When and where does a customer buy? What does s/he buy?
    - How often s/he pays on time? etc.
  - Label past transactions as *fraud* or *fair* transactions
    This forms the *class attribute*
  - Learn a model for the class of the transaction
  - Use this model to detect fraud by observing credit card transactions on an account

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
- produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered
{Diaper, Milk} → {Beer}
{Milk} → {Coke}

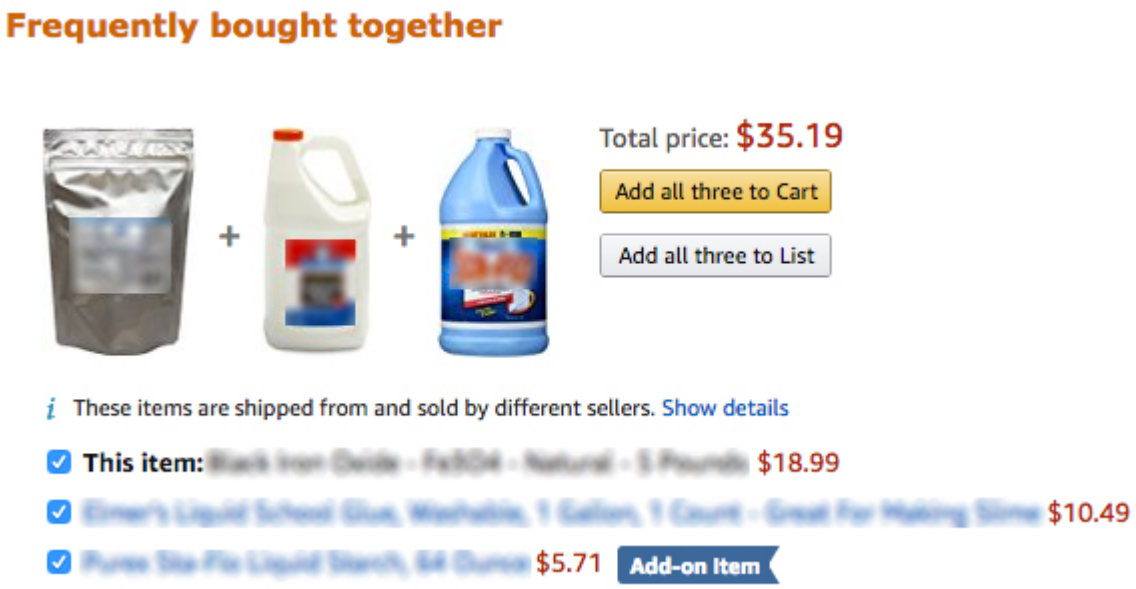
Association Rule Discovery: Applications

- Application area: Marketing and Sales Promotion
- Example rule discovered:

  {Bagels, Coke} --> {Potato Chips}
- Insights:
  - promote bagels to boost potato chips sales
  - if selling bagels is discontinued, this will affect potato chips sales
  - coke should be sold together with bagels to boost potato chips sales

Association Rule Discovery: Applications

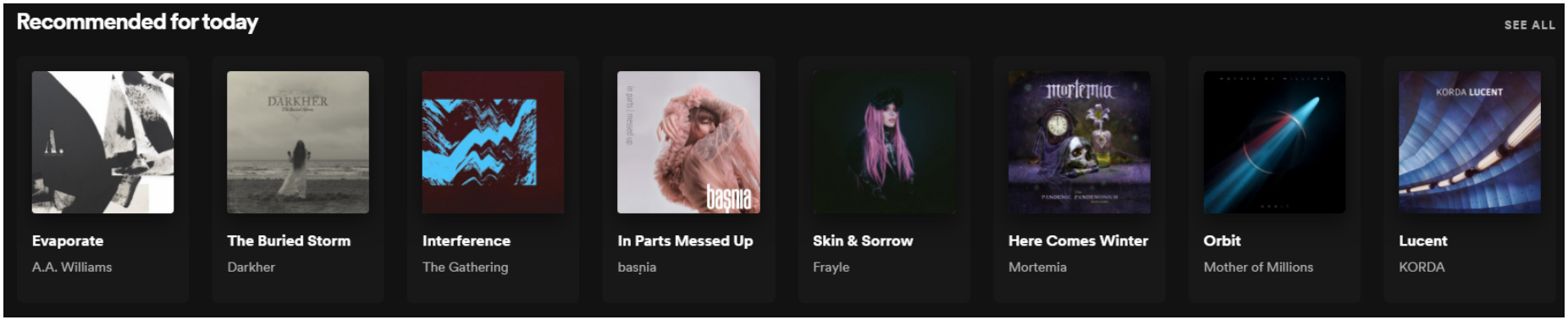
- Customers who bought this product also bought…
  - ...do terrorists order bomb building parts on Amazon?



http://thenewdaily.com.au/news/world/2017/09/21/amazon-bomb-explosives-ingredients-algorithm-frequently-bought-together/

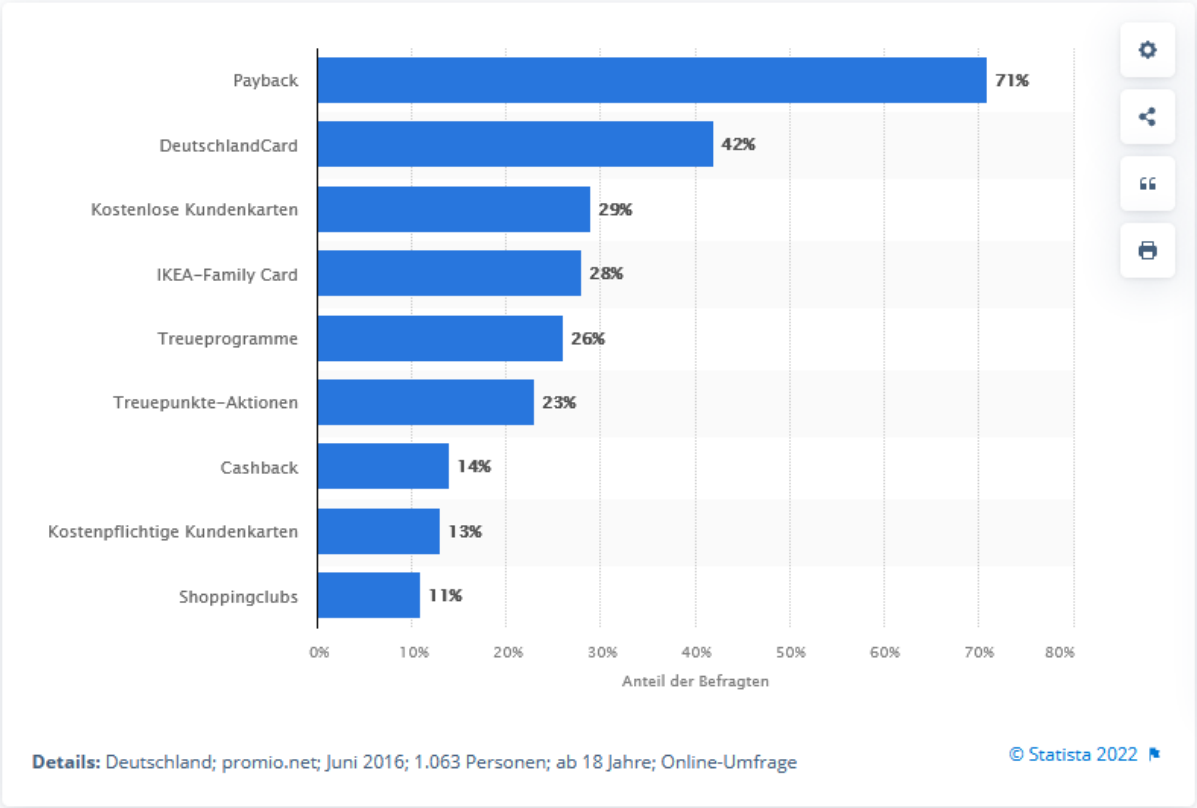
Association Rule Discovery: Applications

- Content-based recommendation
  - requirement: much data
  - e.g., Amazon transactions, Spotify logfiles



Association Rule Discovery: Applications

- Real world example:
  - Customer loyalty programs



http://de.statista.com/statistik/daten/studie/36618/umfrage/anzahl-herausgegebener-bonuskarten-mehrere-partnerunternehmen/

Association Rule Discovery: Applications

- Real example:
  - Target (American grocery store)
  - Analyzes customer buying behavior
  - Sends personalized advertisement

- Famous case in the USA:
  - Teenage girl gets advertisement for baby products
  - ...and her father is mad

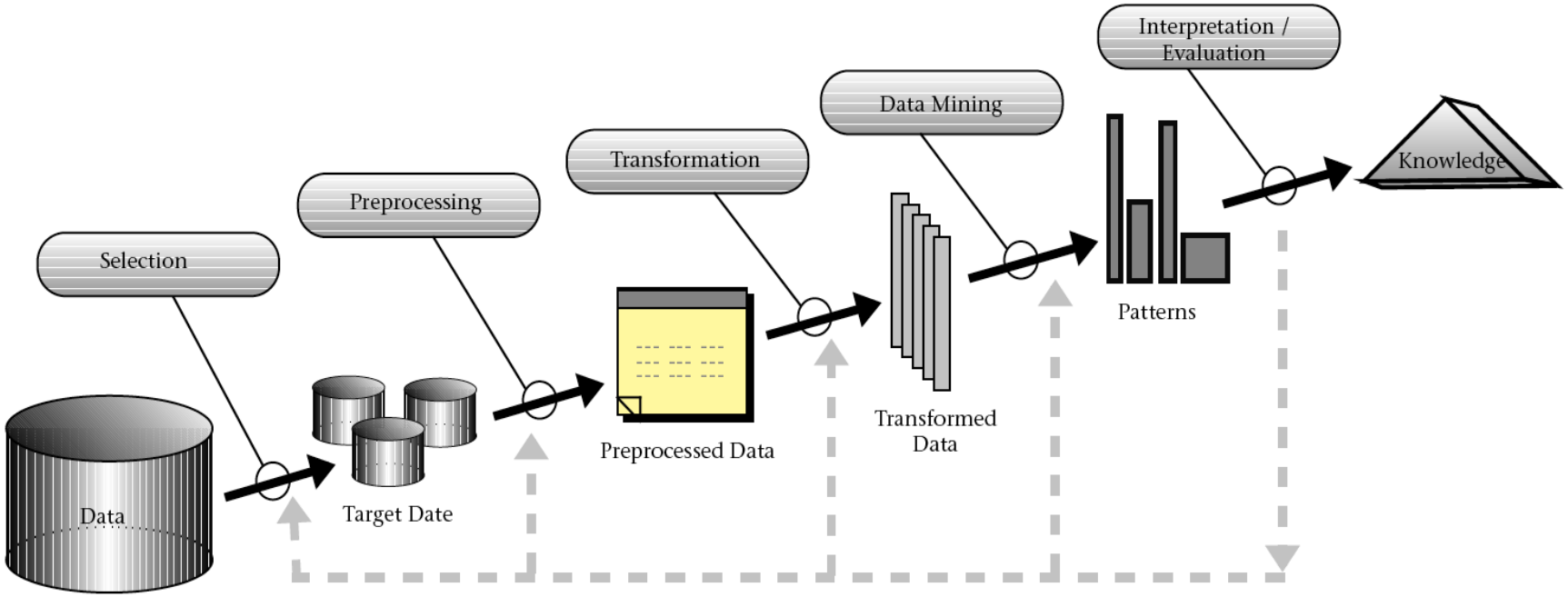
http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

Association Rule Discovery: Applications

- Bottom line of the Target teenage girl story:
  - Janet Vertesi, Princeton university
  - Tried to hide her pregnancy from computers
- Measures taken:
  - using Tor for online surfing
  - no social media posts about her pregnancy
  - paying all pregnancy/baby related products in cash
  - a fresh Amazon account delivering to a local locker
    - paying with cash-payed gift cards
- Outcome:
  - massive buying of gift cards in a convenience store was reported to tax authorities

read the full story at
http://mashable.com/2014/04/26/big-data-pregnancy/

The Data Mining Process



Source: Fayyad et al. (1996)

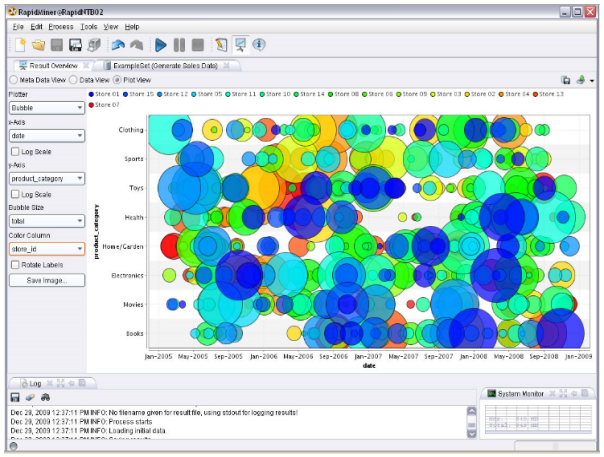
The Data Mining Process

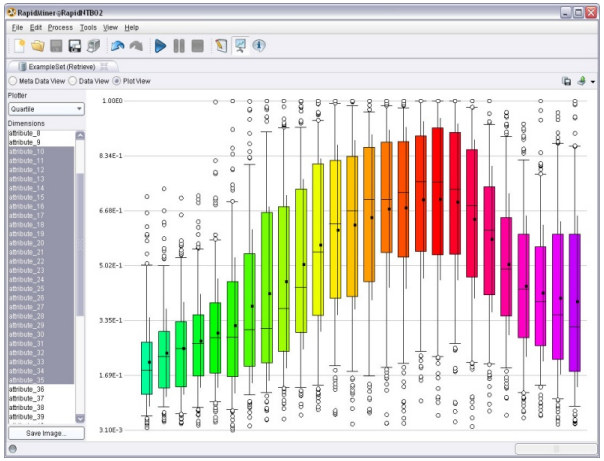
- Note that none of those steps actually requires a computer
- Recall Petermann's Cholera maps
  - Data Selection: find data on cholera deaths
  - Data Preprocessing: organize data by geographic area
  - Transformation: draw data on a map
  - Data Mining: look at the map and find patterns
    - possibly step back: add more data (population, water system, ...)
  - Interpretation: Cholera is transmitted via contaminated water

- However, computers make things easier
  - mainly: scalability (size of datasets, number of patterns)
  - avoiding human bias

Selection and Exploration

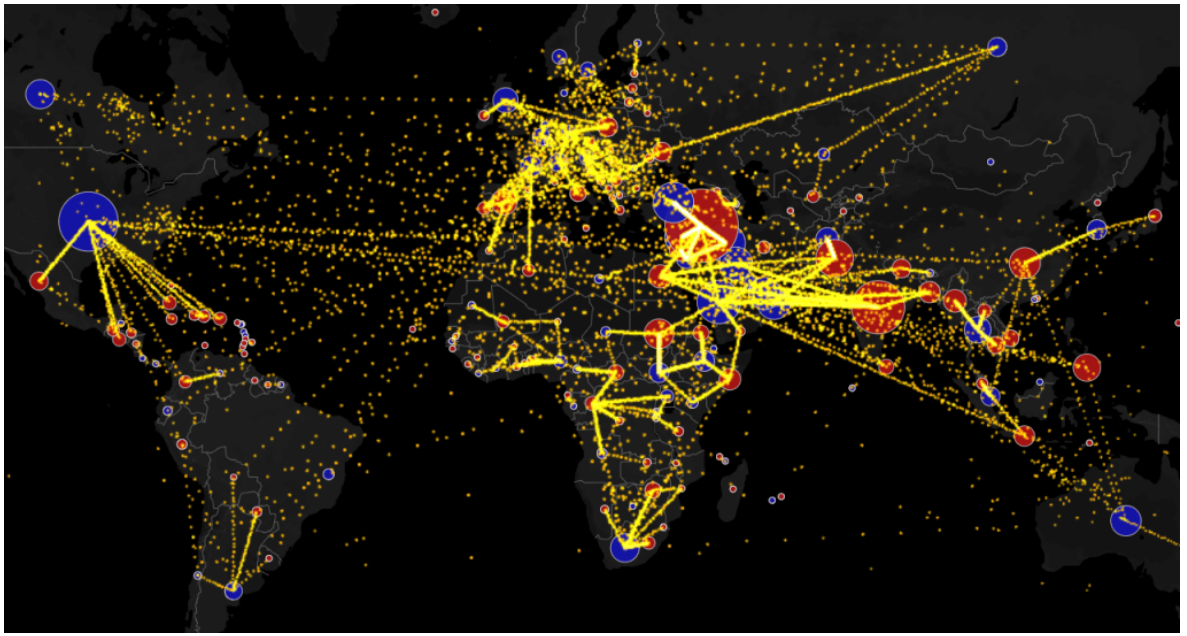
- Selection
  - What data is available?
  - What do I know about the provenance of this data?
  - What do I know about the quality of the data?
- Exploration
  - Get an intitial understanding of the data
  - Calculate basic summarization statistics
  - Visualize the data
  - Identify data problems such as outliers, missing values, duplicate records





Selection and Exploration

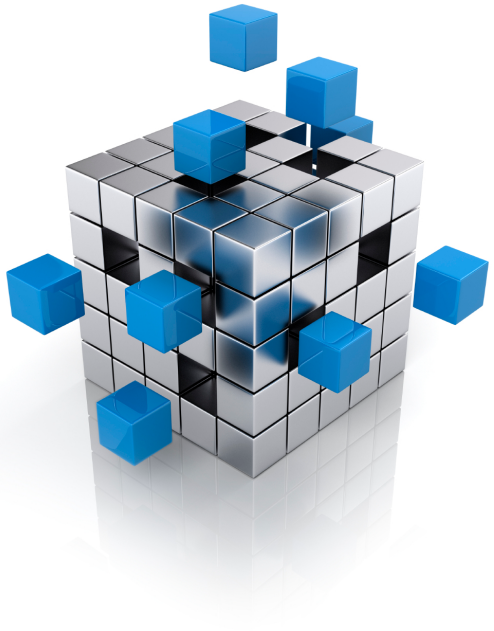
- Visual Data Mining
  - For example as maps
  - Example: Map showing migration streams and net migration of different countries



http://metrocosm.com/global-migration-map.html

Preprocessing and Transformation

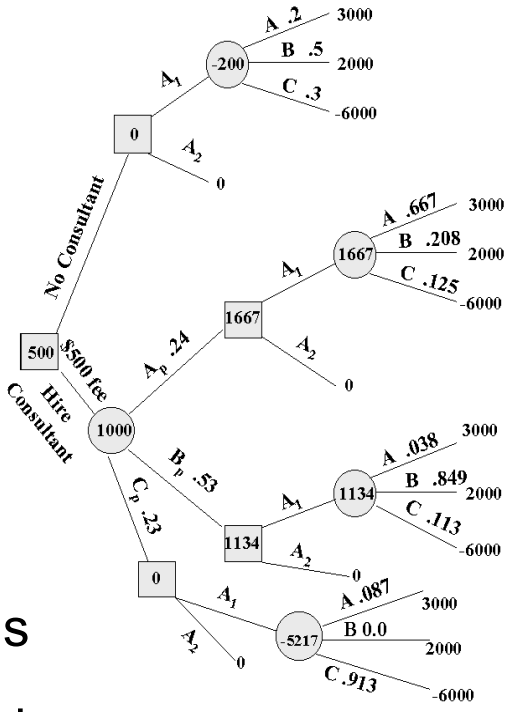
- Transform data into a representation that is suitable for the chosen data mining methods
  - number of dimensions
  - scales of attributes (nominal, ordinal, numeric)
  - amount of data (determines hardware requirements)
- Methods
  - Aggregation, sampling
  - Dimensionality reduction / feature subset selection
  - Attribute transformation / text to term vector
  - Discretization and binarization



- Good data preparation is key to producing valid and reliable models
- Data preparation estimated to take 70-80% of the time and effort of a data mining project!

Data Mining

- Input: Preprocessed Data
- Output: Model / Patterns



1. Apply data mining method
2. Evaluate resulting model / patterns
3. Iterate:
   - Experiment with different parameter settings
   - Experiment with different alternative methods
   - Improve preprocessing and feature generation
   - Combine different methods

Interpretation / Evaluation

- Output of Data Mining
  - Patterns
  - Models

- In the end, we want to derive value from that, e.g.,
  - gain knowledge
  - make better decisions
  - increase revenue

What you will learn in this lecture

- Common data mining tasks
  - How they work
  - When and how to apply them
  - How to interpret their output

Data is the New...

- Oil (2006)

- CO_2 (2019)

Questions?