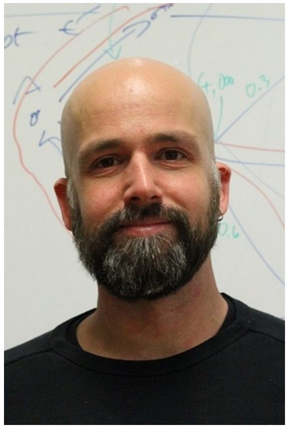
UNIVERSITÄT
MANNHEIM

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 *RapidMiner* exercises and co-supervise the team projects.

Hello

- M.Sc. Sven Hertling
- Graduate Research Associate
- Research Interests:
  - Semantic Technologies / Semantic Web
  - Linked Data
  - Knowledge Graphs
- eMail: sven@informatik.uni-mannheim.de
- Sven will teach the *Python* exercises 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@uni-mannheim.de
- Ralph will teach the *Python* exercises 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 RapidMiner *or* 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)

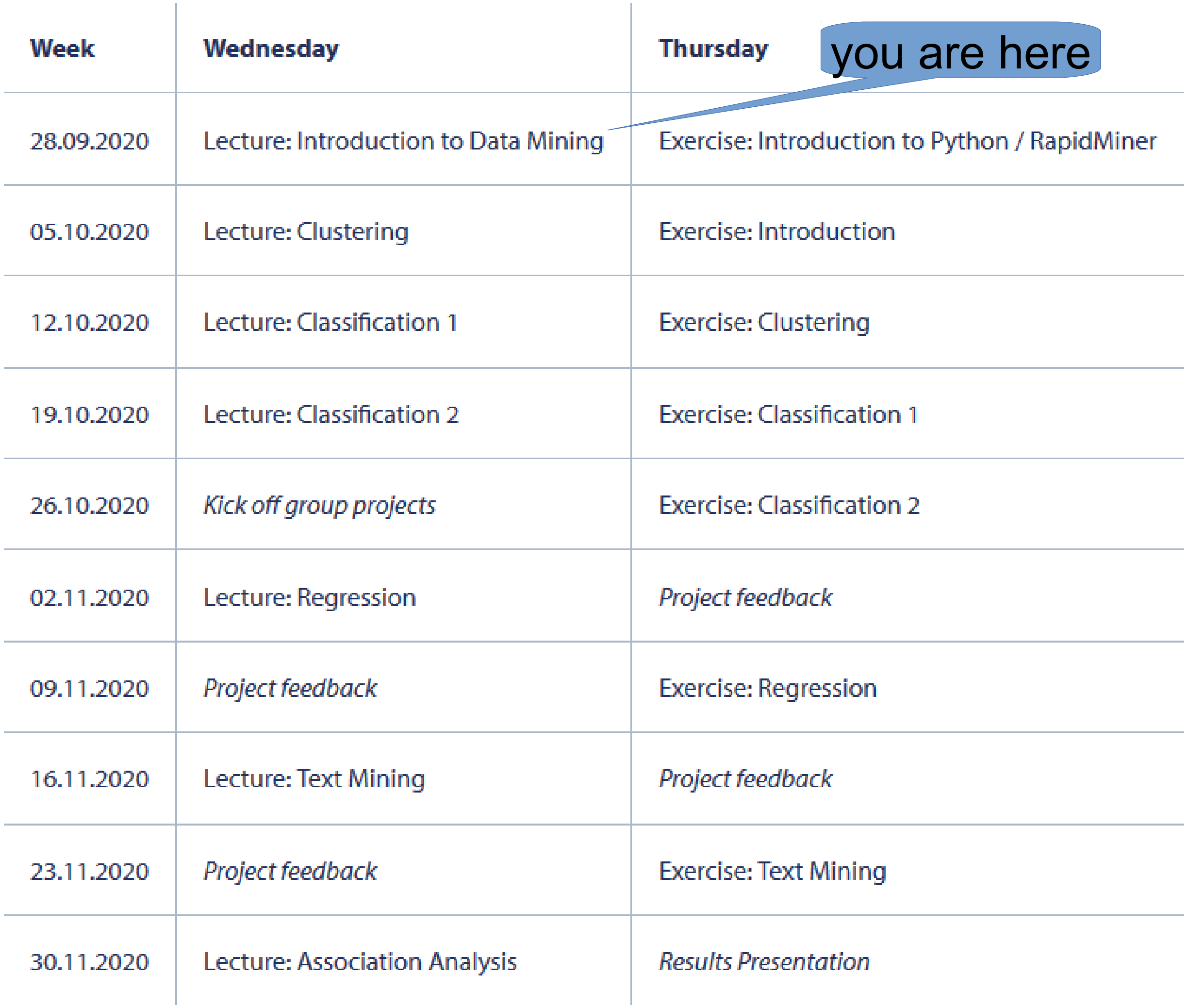
If you fail the exam, but do a good project, you may still pass.

Exercises of Your Choice

- Exercises in RapidMiner
  - Thursday, 12 – 13.30
  - Requires no programming knowledge
- Exercise in Python
  - Thursday, 13.45 – 15.15 and 15.30 – 17.00
  - Requires programming knowledge
- Exercises start tomorrow!

Introduction to Python and Jupyter Notebooks today, 15.30, in this room!

Course Outline



Week	**Wednesday**	**Thursday**
28.09.2020	Lecture: Introduction to Data Mining	Exercise: Introduction to Python / RapidMiner
05.10.2020	Lecture: Clustering	Exercise: Introduction
12.10.2020	Lecture: Classification 1	Exercise: Clustering
19.10.2020	Lecture: Classification 2	Exercise: Classification 1
26.10.2020	*Kick off group projects*	Exercise: Classification 2
02.11.2020	Lecture: Regression	*Project feedback*
09.11.2020	*Project feedback*	Exercise: Regression
16.11.2020	Lecture: Text Mining	*Project feedback*
23.11.2020	*Project feedback*	Exercise: Text Mining
30.11.2020	Lecture: Association Analysis	*Results Presentation*

you are here

Deadlines

- Submission of project work proposal
  - Monday, Nov 2nd, 23:59
- Submission of final project work report
  - Firday, Dec 23rd, 23:59
- Project presentations
  - schedule to be announced
  - everyone has to attend

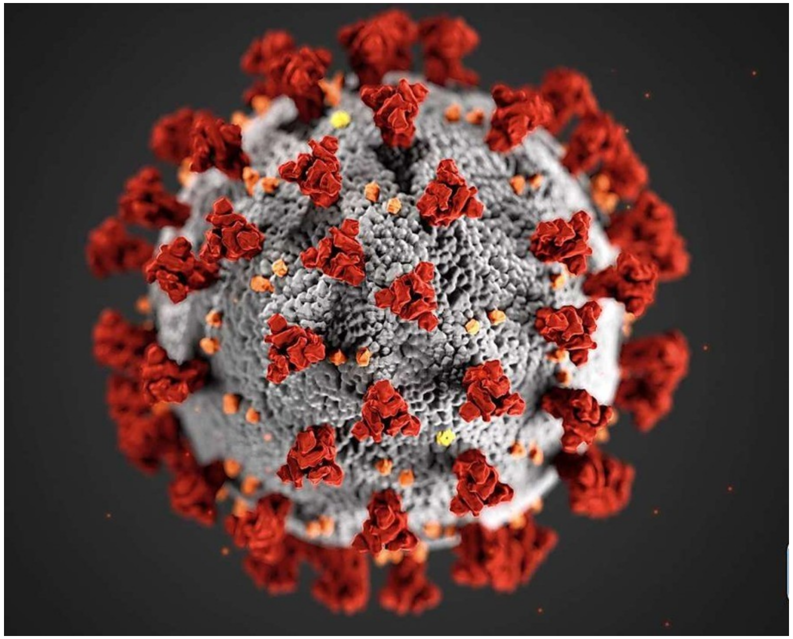
Course Organization

- 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: Wednesday, 10.15 – 11.45, WIM-ZOOM-02
  - Exercises: Thursdays:
    12.00 – 13.30 (RapidMiner w/ Nicolas), WIM-ZOOM-02
    13.45 – 15.15 (Python w/ Sven), WIM-ZOOM-02
    15.30 – 17.00 (Python w/ Ralph), WIM-ZOOM-02
    - 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 email Ms. Czanderle
  - we will reassign your place

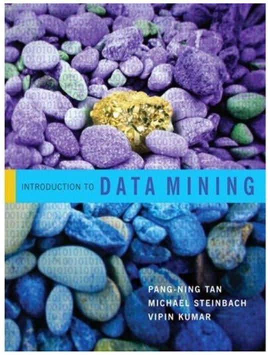
Course Organization – Corona Specials



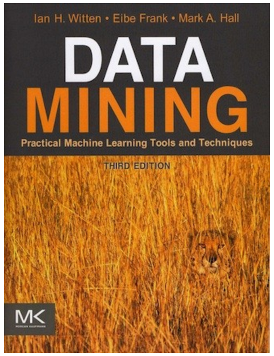
- Lectures and Exercises
  - take place via ZOOM
- Lectures and Exercises are streamed live
  - We will **try to** record lectures and provide the recordings
  - We will **not** record exercises for legal reasons
- Project coaching and presentations
  - will take part via ZOOM
- The written exam will taken place on campus
  - At least as of today...

Literature & Slide Sources

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining, Pearson / Addison Wesley.
  - 10 copies in university library.
  - we provide scans of important chapters via ILIAS

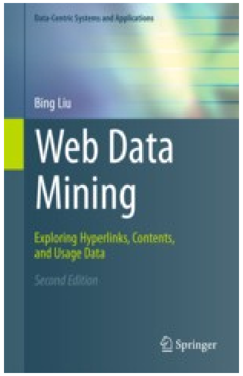


- Ian H. Witten, Eibe Frank, Mark A. Hall: Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann.
  - several copies in university library
  - we provide scans of important chapters via ILIAS


Ian H. Witten • Eibe Frank • Mark A. Hall
DATA MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

Literature & Slide Sources

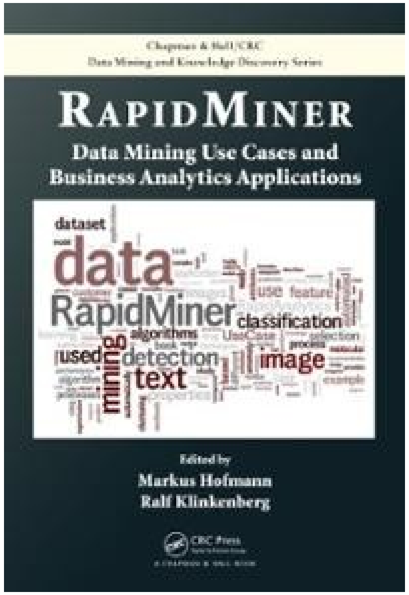
- Bing Liu: Web Data Mining, 2nd Edition, Springer.
  - several copies in university library
  - electronic edition available via the library


Data-Centric Systems and Applications
Bing Liu
Web Data Mining
Exploring Hyperlinks, Contents, and Usage Data
Second Edition
Springer

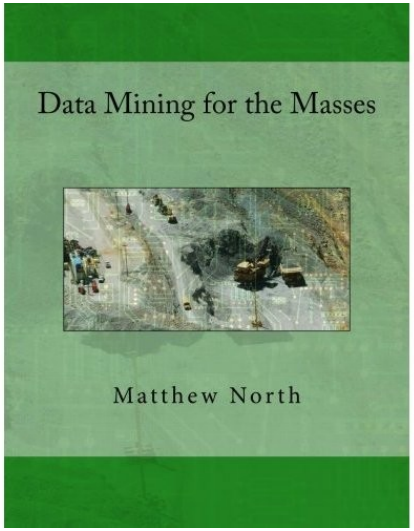
- Gregory Piatetsky-Shapiro, Gary Parker: KDNuggets Data Mining course: http://www.kdnuggets.com/data_mining_course/

Literature – Rapidminer

1. Markus Hofmann, Ralf Klinkenberg:
   RapidMiner: Data Mining Use Cases and Business Analytics Applications.
   Chapman & Hall, 2013.
   - Explains along case studies how to use simple and advanced Rapidminer features.
   - Website with data and processes: http://rapidminerbook.com


Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series
RapidMiner
Data Mining Use Cases and Business Analytics Applications
Edited by
Markus Hofmann
Ralf Klinkenberg
CRC Press

2. Matthew North: **Data Mining for the Masses**.
   Global Text Project, 2012.
   - Free PDF version available online.

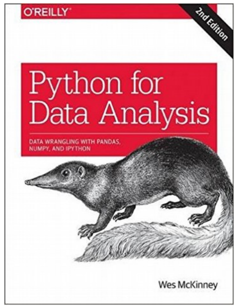


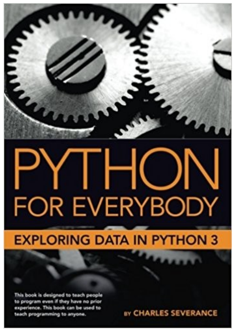
3. **Rapidminer – User Manual**
   - introduction to user interface and basic features
   - http://rapidminer.com/learning/getting-started/

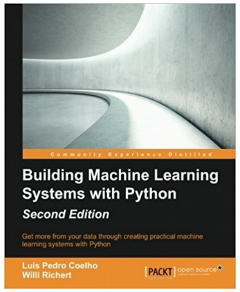
Literature – Python

- McKinney: Python for Data Analysis
- Severance: Python for Everybody:
  Exploring Data in Python 3
- Coelho and Richert: Building Machine Learning Systems with Python – *Free Online Access via university library*

- Online Sources:
  - https://www.learnpython.org/
  - https://docs.python.org/3/tutorial/
  - http://scikit-learn.org/stable/tutorial/index.html







Additional Material

- Video recordings from FSS 2015
  - http://dws.informatik.uni-mannheim.de/en/teaching/lecture-videos/

Data Mining I

Organization and Course Outline

Introduction to Data Mining

Prof. Dr. Christian Bizer

Data and Web Science Group

UNIVERSITY OF MANNHEIM

UNIVERSITÄT MANNHEIM

Data Mining I

Organization and Course Outline

Universität Mannheim – Bizer/Bryl/Lehmberg: Data Mining I

Slide 1

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

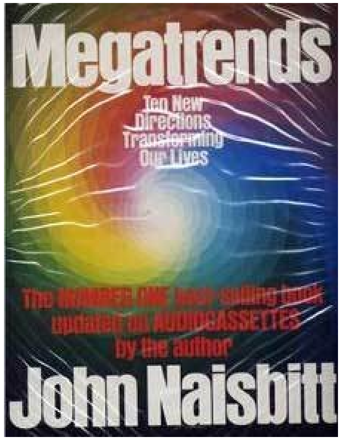
DATA-MINING-CUP
prudsys Anwendertage und Data Mining Wettbewerb

Questions?

A Bit of History

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

(John Naisbitt, 1982)


Megatrends
Ten New Directions Transforming Our Lives
The NUMBER ONE best-selling book updated on AUDIOCASSETTES by the author
John Naisbitt

- *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.*

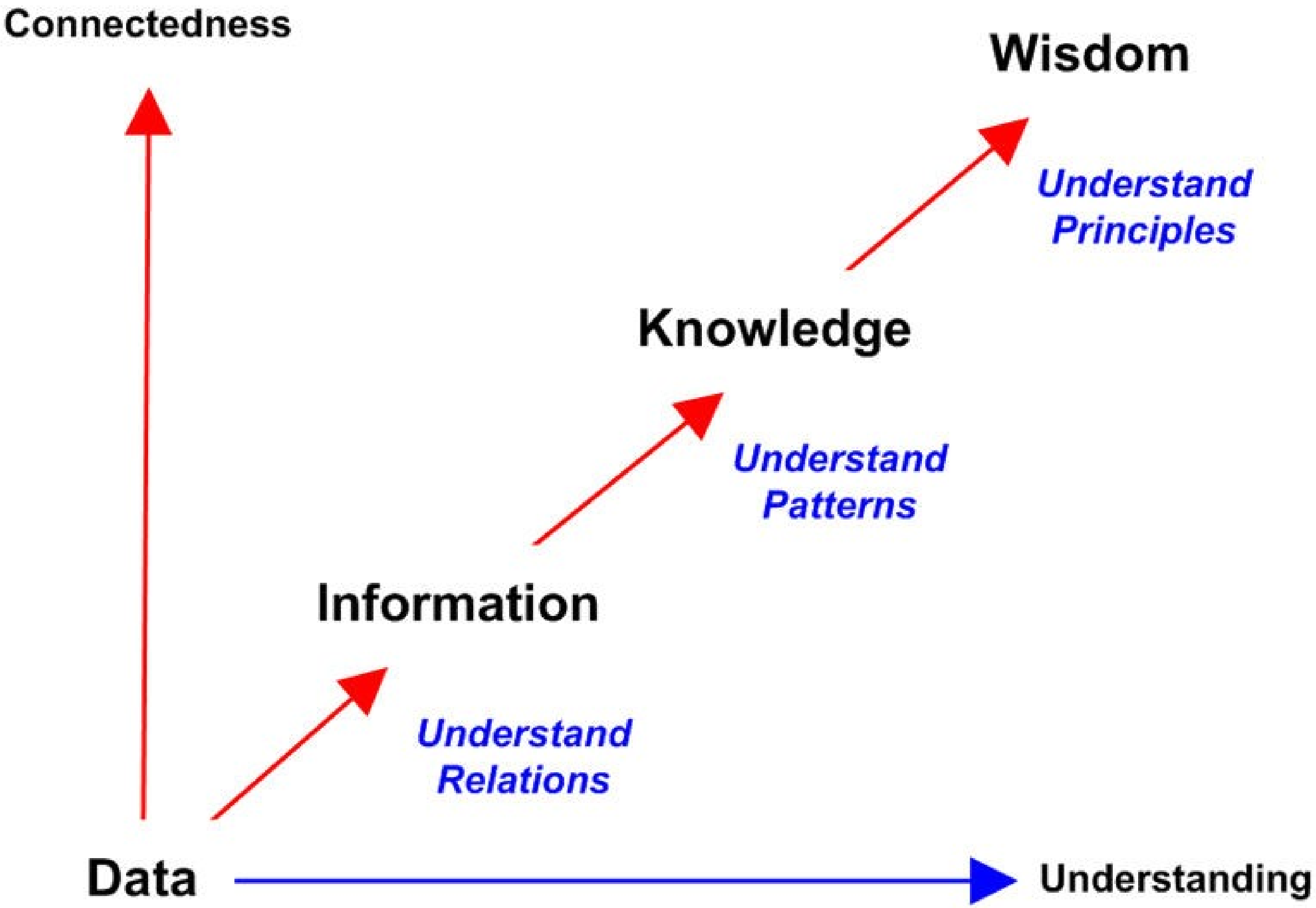
(Frawley, Piatetsky-Shapiro, Matheus, 1992)

“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
- ...

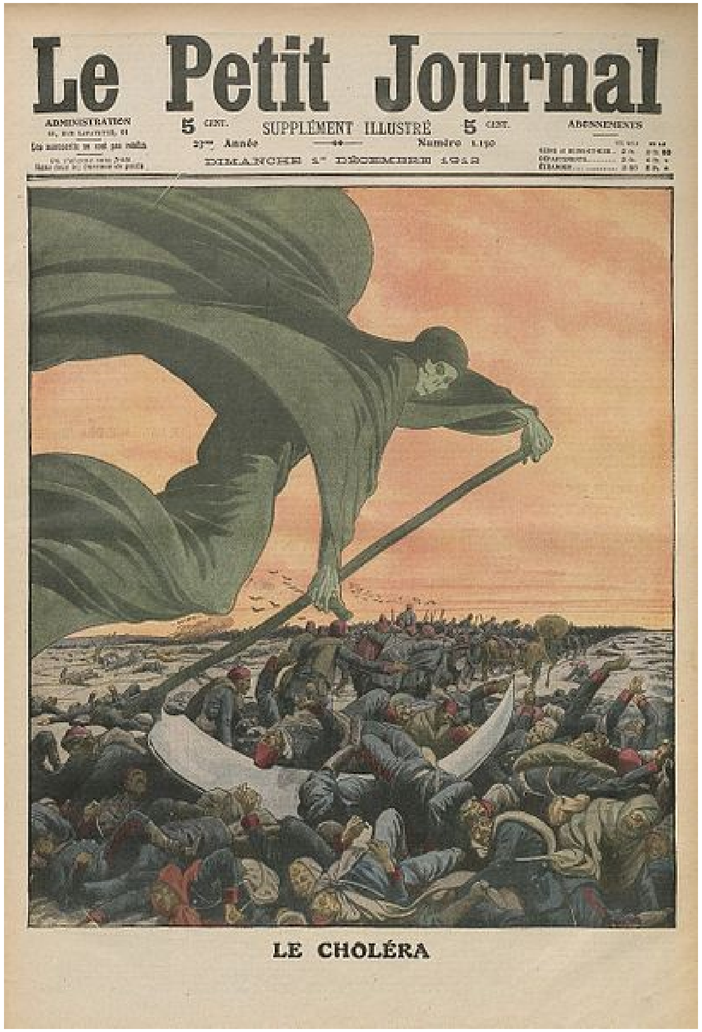
Data, Information, Knowledge, and Wisdom



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

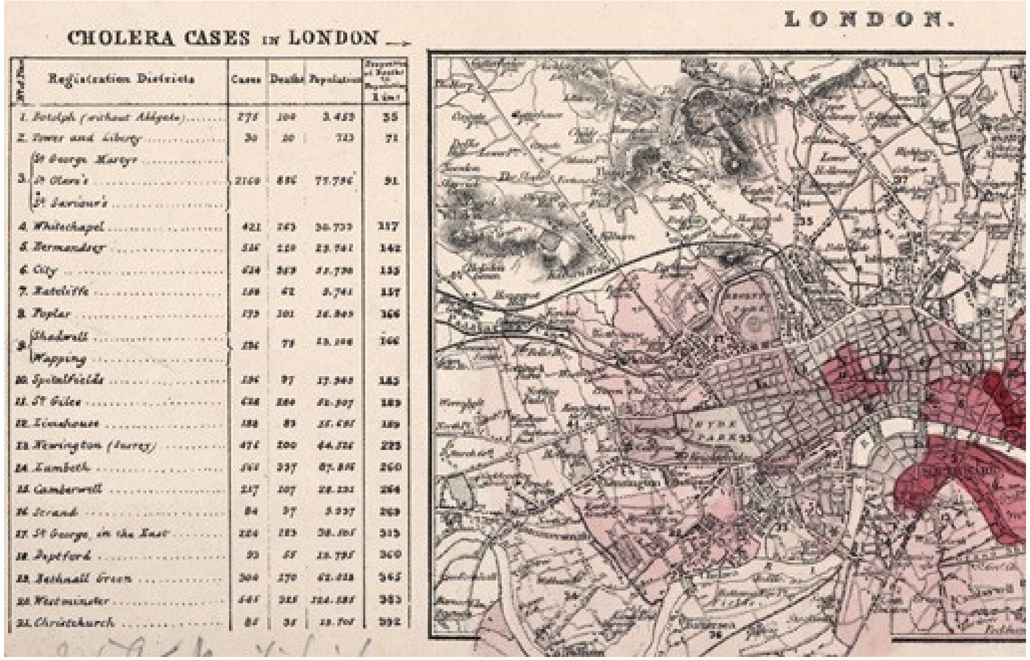
A Historical Example

- 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



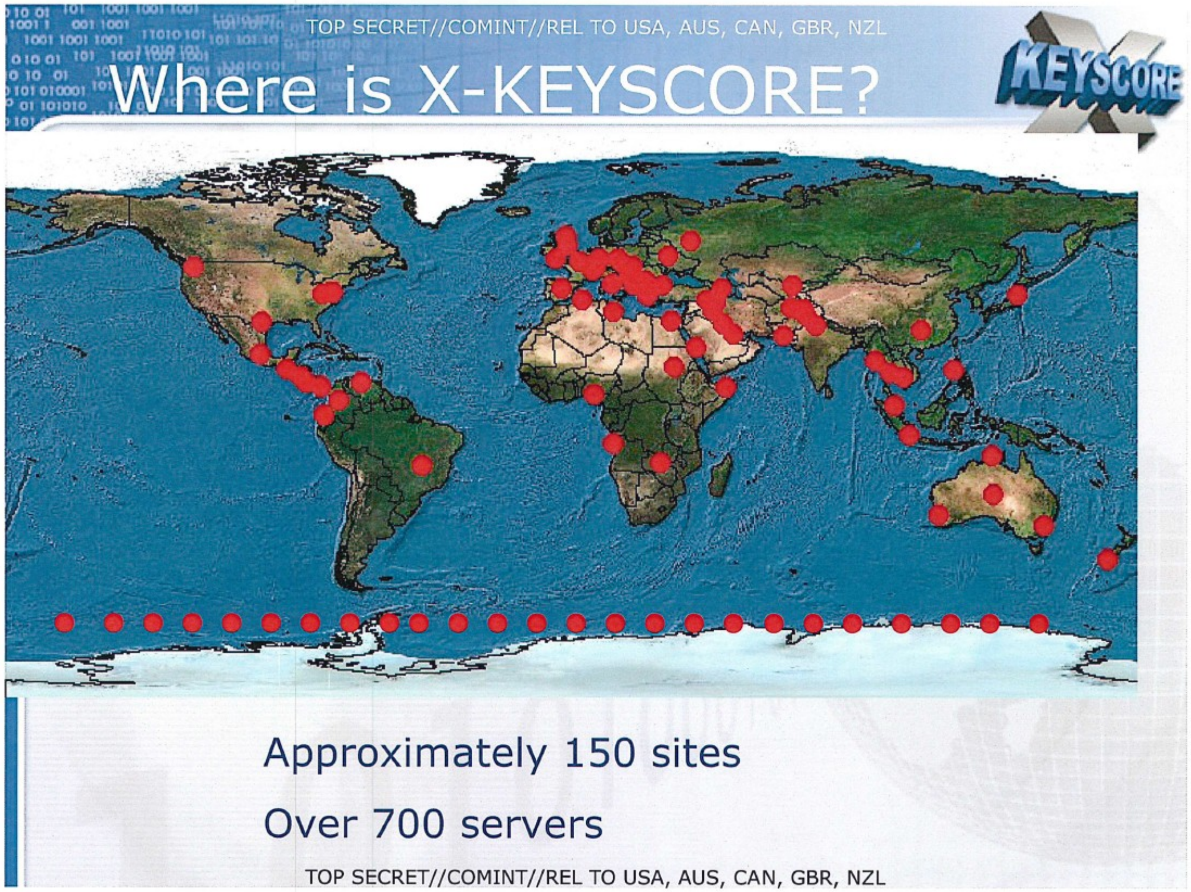
CHOLERA CASES IN LONDON

Registration Districts	Cases	Deaths	Population	Proportion of Deaths to Population 1 in
1. Botolph (without Aldgate)	275	100	3.453	35
2. Tower and Liberty	30	20	723	71
3. St George Martyr St Olave's St Saviour's	2160	886	77.796	91
4. Whitechapel	421	263	30.733	117
5. Bermondsey	526	220	23.761	142
6. City	624	369	51.730	133
7. Ratcliffe	168	62	8.761	157
8. Poplar	575	301	36.843	166
9. Shadwell Wapping	356	73	13.104	166
10. Spitalfields	236	97	17.340	183
11. St Giles	636	280	52.907	183
12. Limehouse	188	83	25.691	189
13. Newington (Surrey)	474	200	44.526	223
14. Lambeth	565	237	87.846	260
15. Camberwell	217	107	28.231	264
16. Strand	84	37	9.937	269
17. St George, in the East	226	183	38.505	315
18. Deptford	93	58	13.795	360
19. Bethnall Green	300	270	62.018	345
20. Westminster	545	318	124.585	383
21. Christchurch	85	38	13.708	392

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



TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

Where is X-KEYSCORE?

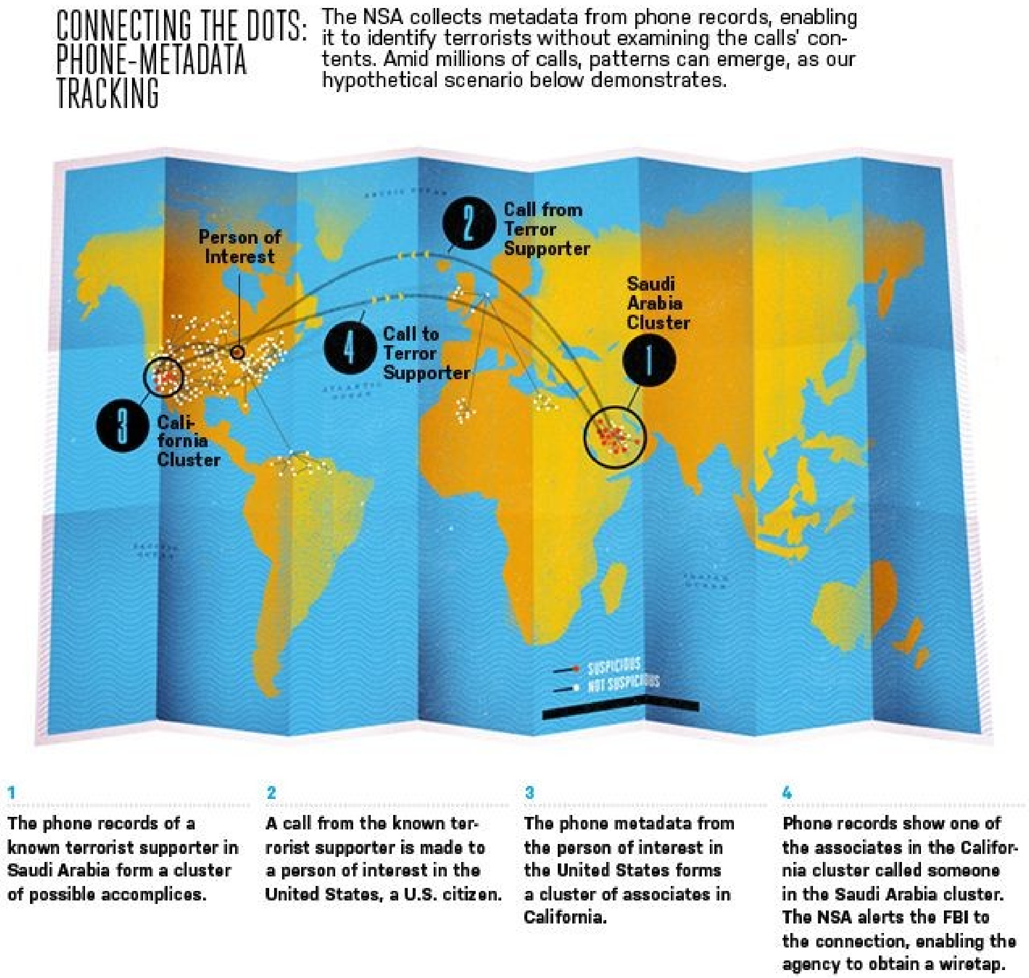
Approximately 150 sites

Over 700 servers

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

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

A Recent Example: the NSA



CONNECTING THE DOTS: PHONE-METADATA TRACKING

The NSA collects metadata from phone records, enabling it to identify terrorists without examining the calls' contents. Amid millions of calls, patterns can emerge, as our hypothetical scenario below demonstrates.

1

The phone records of a known terrorist supporter in Saudi Arabia form a cluster of possible accomplices.

2

A call from the known terrorist supporter is made to a person of interest in the United States, a U.S. citizen.

3

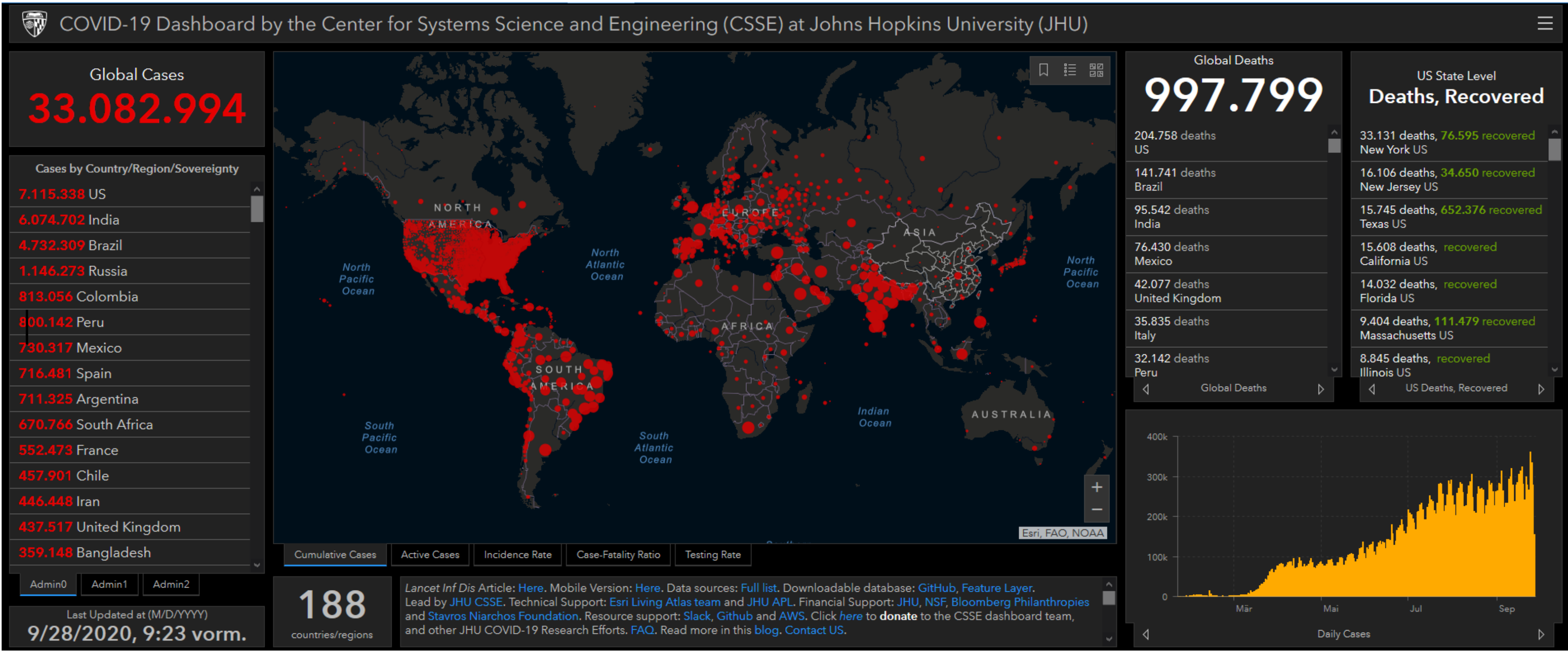
The phone metadata from the person of interest in the United States forms a cluster of associates in California.

4

Phone records show one of the associates in the California cluster called someone in the Saudi Arabia cluster. The NSA alerts the FBI to the connection, enabling the agency to obtain a wiretap.

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

A Very Recent Example: CoViD-19



COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)

Global Cases

33.082.994

Cases by Country/Region/Sovereignty

7.115.338	US
6.074.702	India
4.732.309	Brazil
1.146.273	Russia
813.056	Colombia
800.142	Peru
730.317	Mexico
716.481	Spain
711.325	Argentina
670.766	South Africa
552.473	France
457.901	Chile
446.448	Iran
437.517	United Kingdom
359.148	Bangladesh

Last Updated at (M/D/YYYY)

9/28/2020, 9:23 vorm.

188

countries/regions

Lancet Inf Dis Article: Here. Mobile Version: Here. Data sources: Full list. Downloadable database: GitHub, Feature Layer. Lead by JHU CSSE. Technical Support: Esri Living Atlas team and JHU APL. Financial Support: JHU, NSF, Bloomberg Philanthropies and Stavros Niarchos Foundation. Resource support: Slack, Github and AWS. Click *here* to **donate** to the CSSE dashboard team, and other JHU COVID-19 Research Efforts. FAQ. Read more in this blog. Contact US.

Global Deaths

997.799

204.758 deaths	US
141.741 deaths	Brazil
95.542 deaths	India
76.430 deaths	Mexico
42.077 deaths	United Kingdom
35.835 deaths	Italy
32.142 deaths	Peru

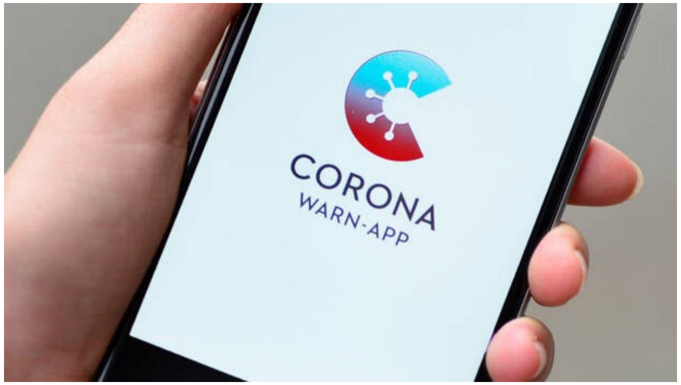
US State Level

Deaths, Recovered

33.131 deaths, 76.595 recovered	New York US
16.106 deaths, 34.650 recovered	New Jersey US
15.745 deaths, 652.376 recovered	Texas US
15.608 deaths, recovered	California US
14.032 deaths, recovered	Florida US
9.404 deaths, 111.479 recovered	Massachusetts US
8.845 deaths, recovered	Illinois US

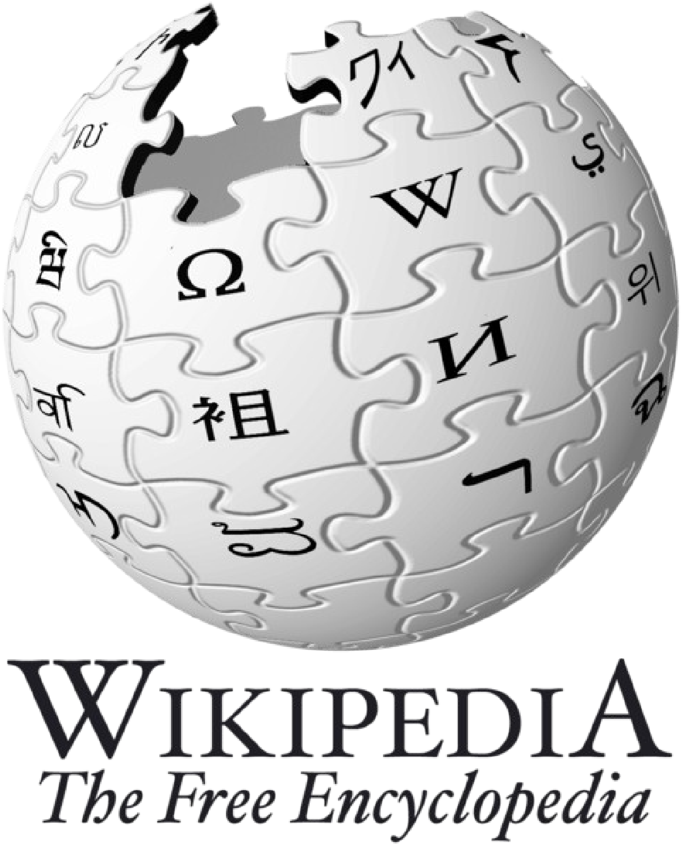
A Very Recent Example: CoViD-19

- 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



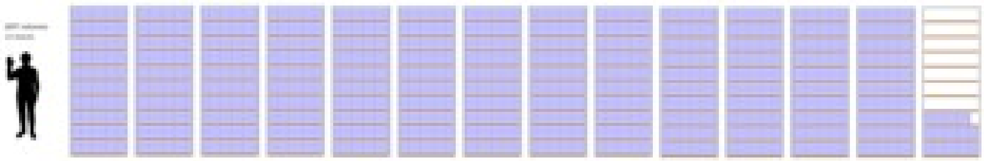
“We are Drowning in Data...”

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

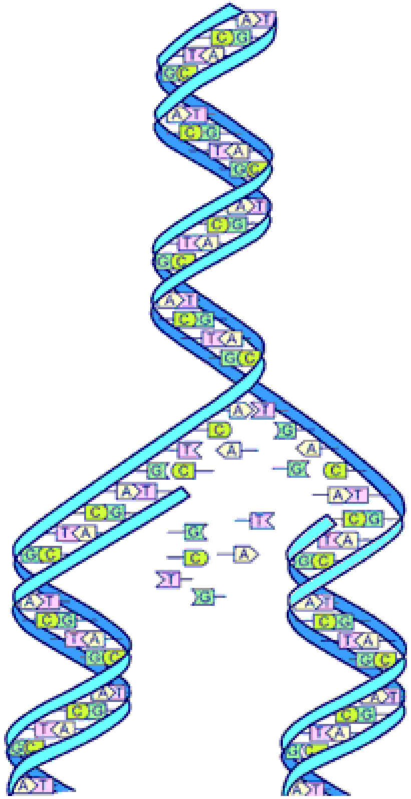


Wikipedia (en, text only)
≈ 20 GB of data

1 Wiki = 1 Wikipedia



“We are Drowning in Data...”



Human Genome

≈ 4 GB/person

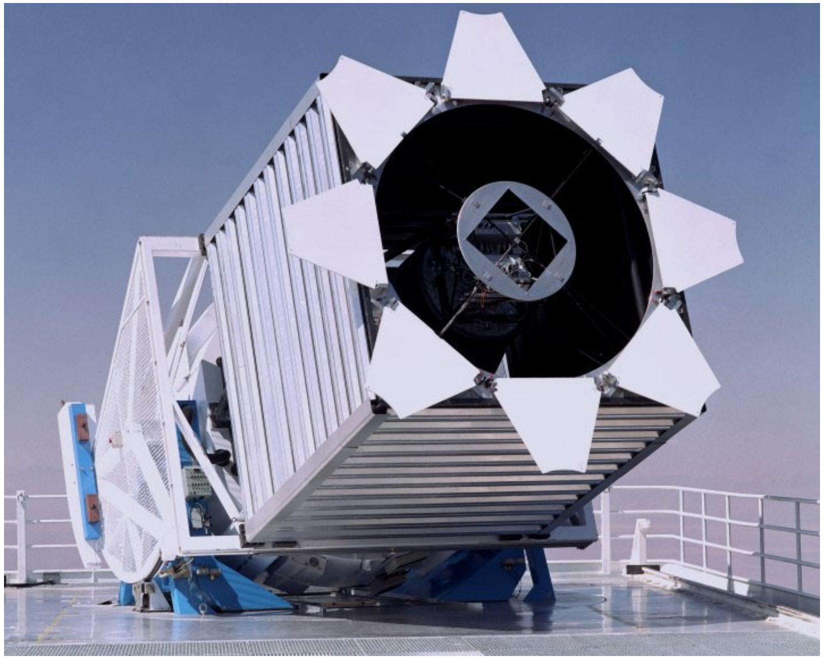
≈ 0.2 Wiki/person

≈ 1.6M Wiki/humankind

“We are Drowning in Data...”

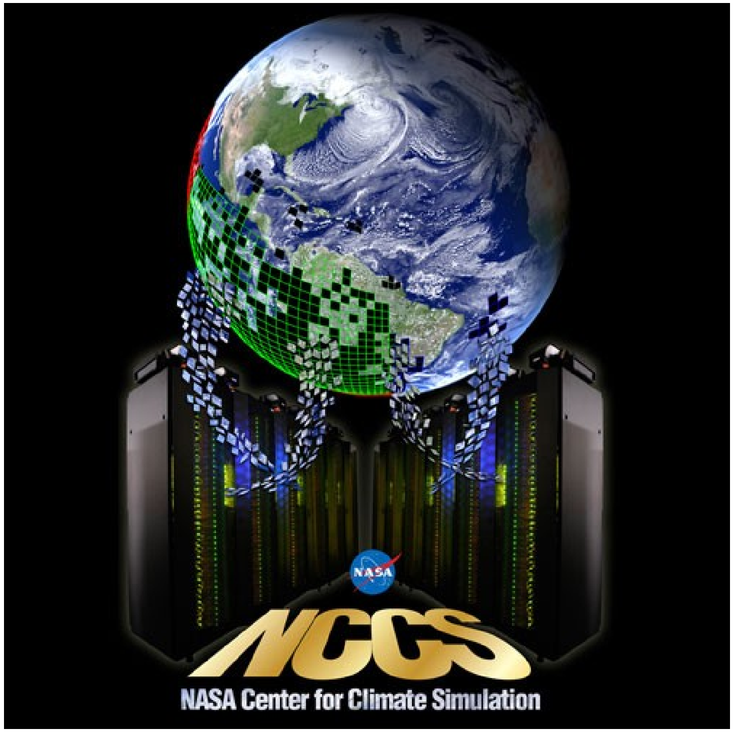
US Library of Congress
≈ 235 TB archived
≈ 11.7M Wiki

“We are Drowning in Data...”



Sloan Digital Sky Survey
≈ 200 GB/day
≈ 73 TB/year
≈ 3.7k Wiki/year

“We are Drowning in Data...”



NASA Center for Climate Simulation

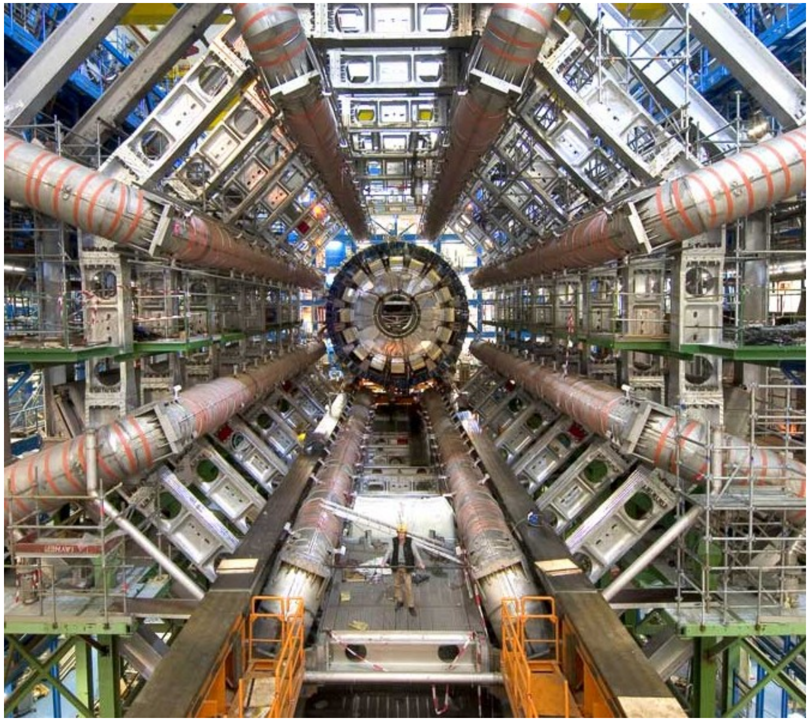
≈ 32 PB archived

≈ 1.6M Wiki

“We are Drowning in Data...”

Facebook
≈ 12 TB/day added
≈ 600 Wiki/day
≈ 219k Wiki/year
(*as of Mar. 2010*)

“We are Drowning in Data...”



Large Hadron Collider
≈ 15 PB/year
≈ 750k Wiki/year

“We are Drowning in Data...”

Google
≈ 20 PB/day processed
≈ 1M Wiki/day
≈ 365M Wiki/year
(Jan. 2010)

“We are Drowning in Data...”

Internet (2016)

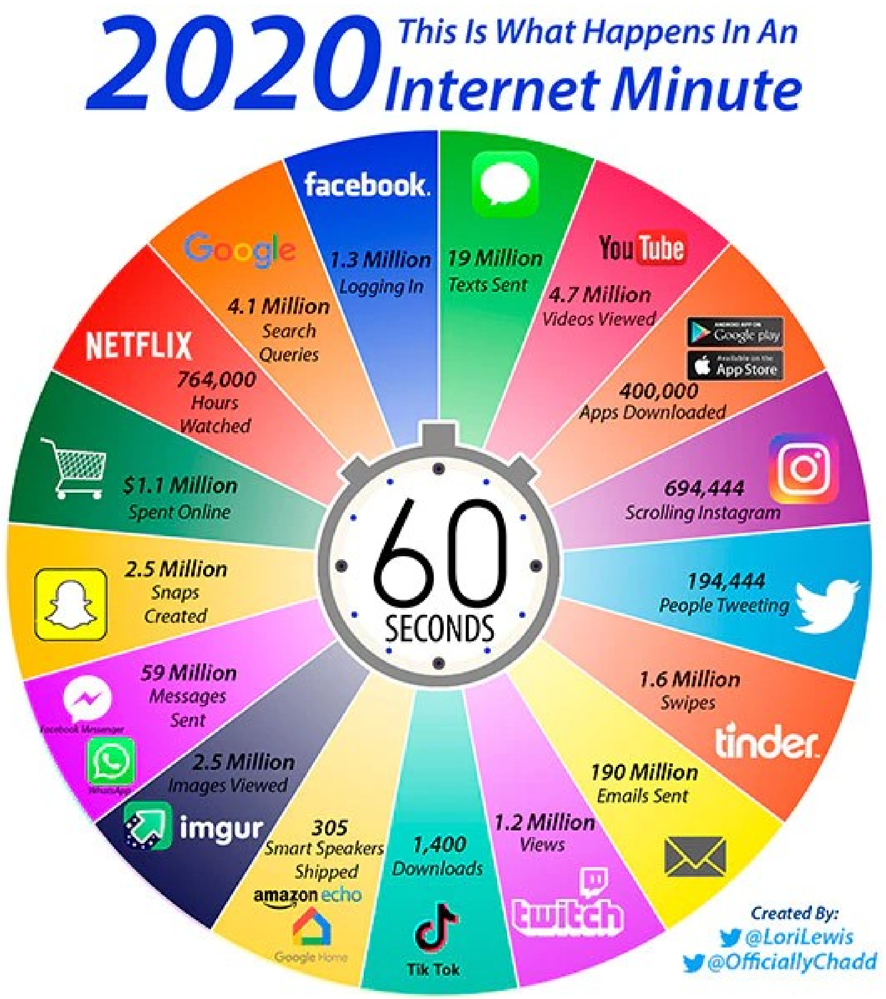
≈ 1.3 ZB/year

≈ 65M Wiki/year

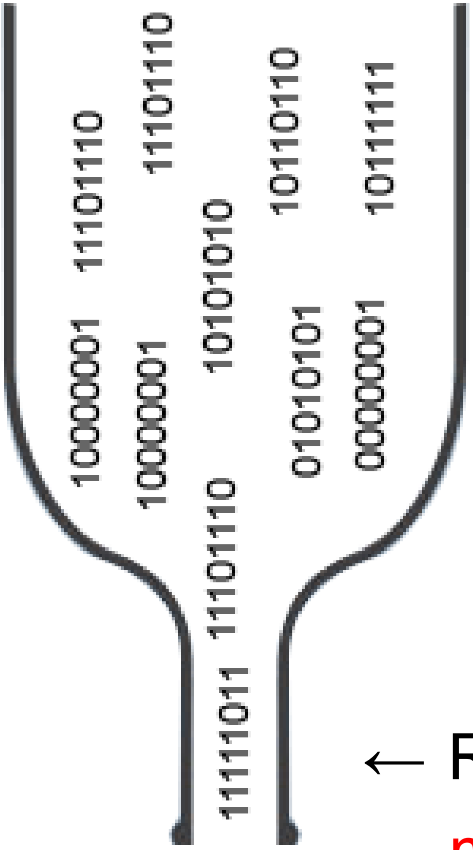
(2016 IP traffic; Cisco est.)

≈ 2 Wiki/second

“We are Drowning in Data...”



...but starving for knowledge!



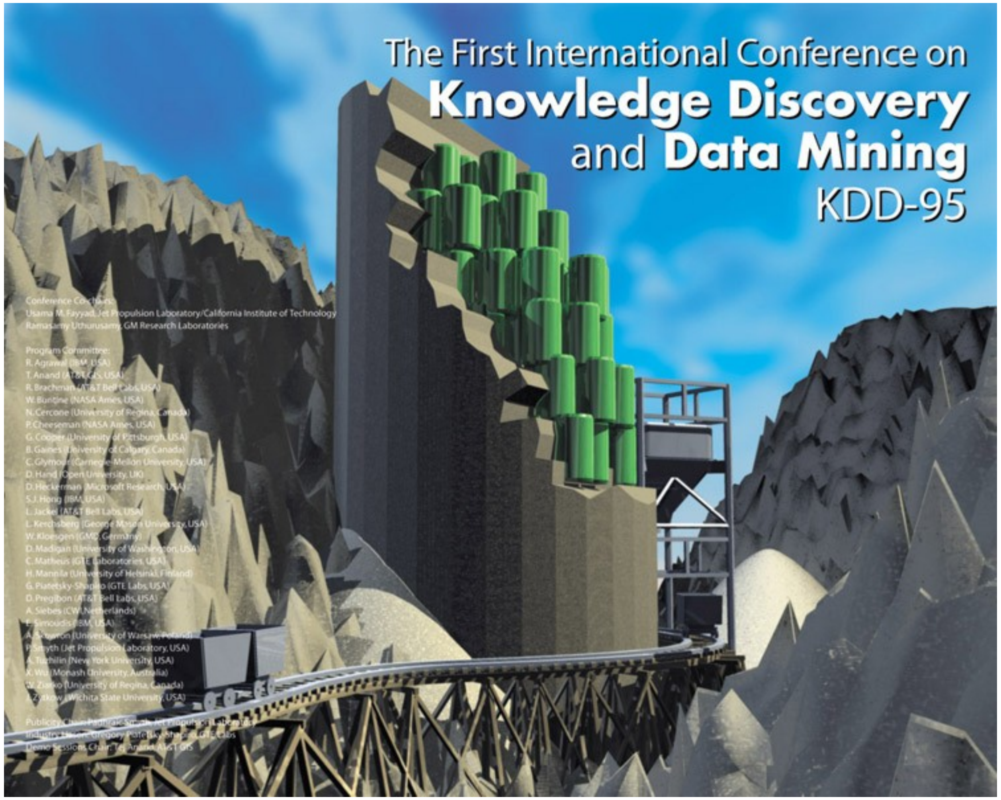
← Rate at which data are produced

← Rate at which data can be understood

manual interpretation is hardly feasible!

Data Mining: Definitions

- Idea: mountains of data
  - where knowledge is mined


The First International Conference on
Knowledge Discovery
and Data Mining
KDD-95
Conference Co-chairs:
Usama M. Fayyad, Jet Propulsion Laboratory/California Institute of Technology
Ramasamy Uthurusamy, GM Research Laboratories
Program Committee:
R. Agrawal (IBM, USA)
T. Anand (AT&T GIS, USA)
R. Brachman (AT&T Bell Labs, USA)
W. Buntine (NASA Ames, USA)
N. Cercone (University of Regina, Canada)
P. Cheeseman (NASA Ames, USA)
G. Cooper (University of Pittsburgh, USA)
B. Gaines (University of Calgary, Canada)
C. Glymour (Carnegie-Mellon University, USA)
D. Hand (Open University, UK)
D. Heckerman (Microsoft Research, USA)
S.J. Hong (IBM, USA)
L. Jackel (AT&T Bell Labs, USA)
L. Kerchsberg (George Mason University, USA)
W. Kloesgen (GMD, Germany)
D. Madigan (University of Washington, USA)
C. Matheus (GTE Laboratories, USA)
H. Mannila (University of Helsinki, Finland)
G. Piatetsky-Shapiro (GTE Labs, USA)
D. Pregibon (AT&T Bell labs, USA)
A. Siebes (CWI, Netherlands)
E. Simoudis (IBM, USA)
A. Skowron (University of Warsaw, Poland)
P. Smyth (Jet Propulsion Laboratory, USA)
A. Tuzhilin (New York University, USA)
X. Wu (Monash University, Australia)
W. Ziarko (University of Regina, Canada)
J. Zytkow (Wichita State University, USA)
Publicity Chair: Padhraic Smyth, Jet Propulsion Laboratory
Industry Liaison: Gregory Piatetsky-Shapiro, GTE Labs
Demo Sessions Chair: Tej Anand, AT&T GIS

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)

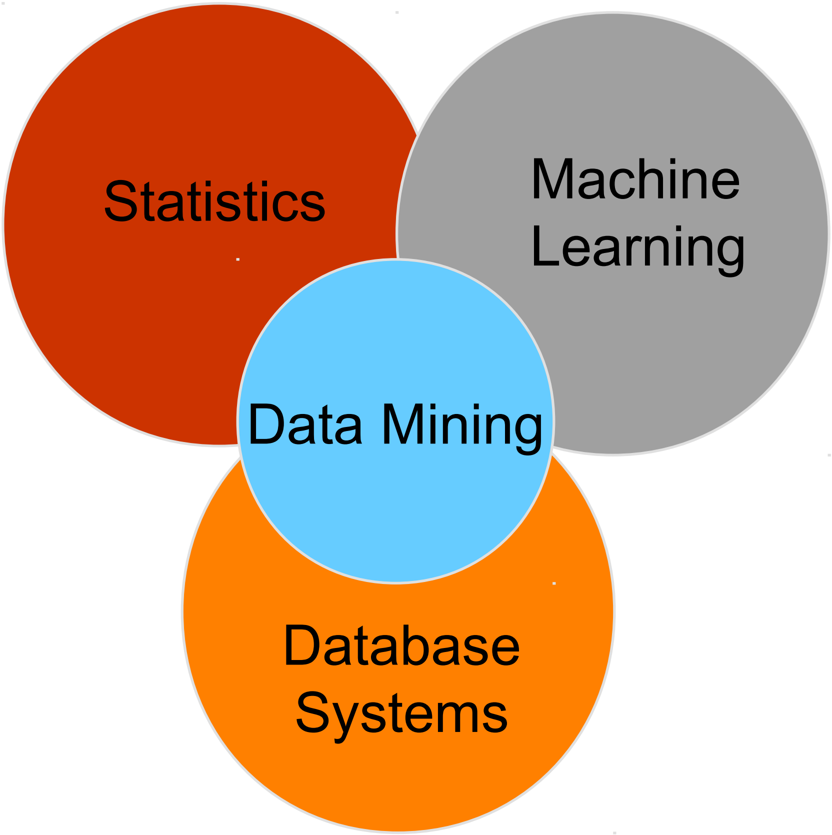
- Data mining is nothing else than torturing the data until it confesses

  (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

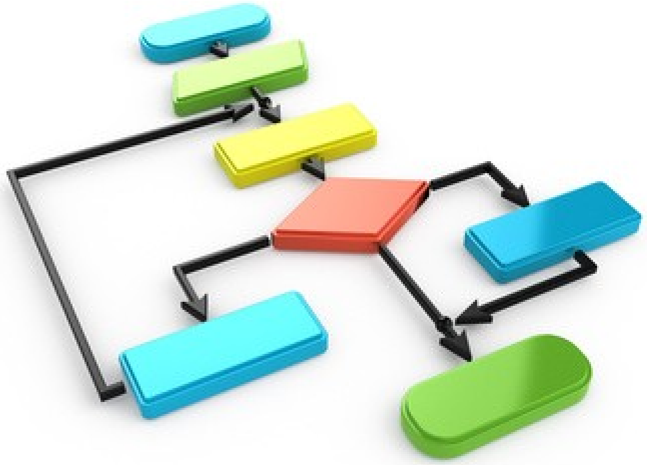


Data Mining Application Fields

- 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

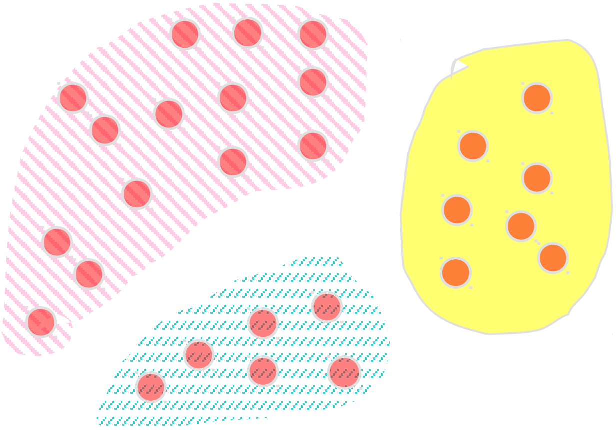
Data Mining Tasks

- 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)

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

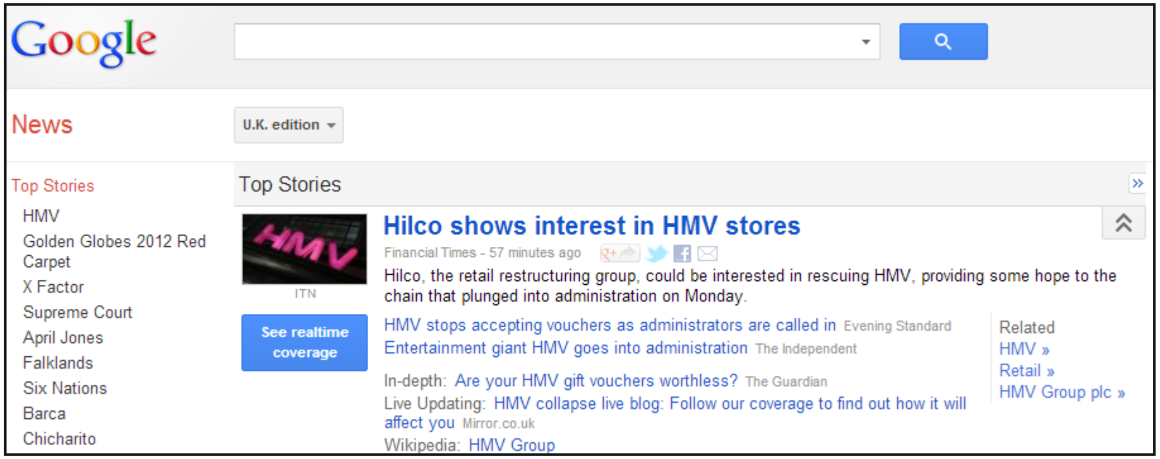


Clustering: Applications

- 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



Hilco shows interest in HMV stores

Financial Times - 57 minutes ago

Hilco, the retail restructuring group, could be interested in rescuing HMV, providing some hope to the chain that plunged into administration on Monday.

HMV stops accepting vouchers as administrators are called in Evening Standard

Entertainment giant HMV goes into administration The Independent

In-depth: Are your HMV gift vouchers worthless? The Guardian

Live Updating: HMV collapse live blog: Follow our coverage to find out how it will affect you Mirror.co.uk

Wikipedia: HMV Group

Classification

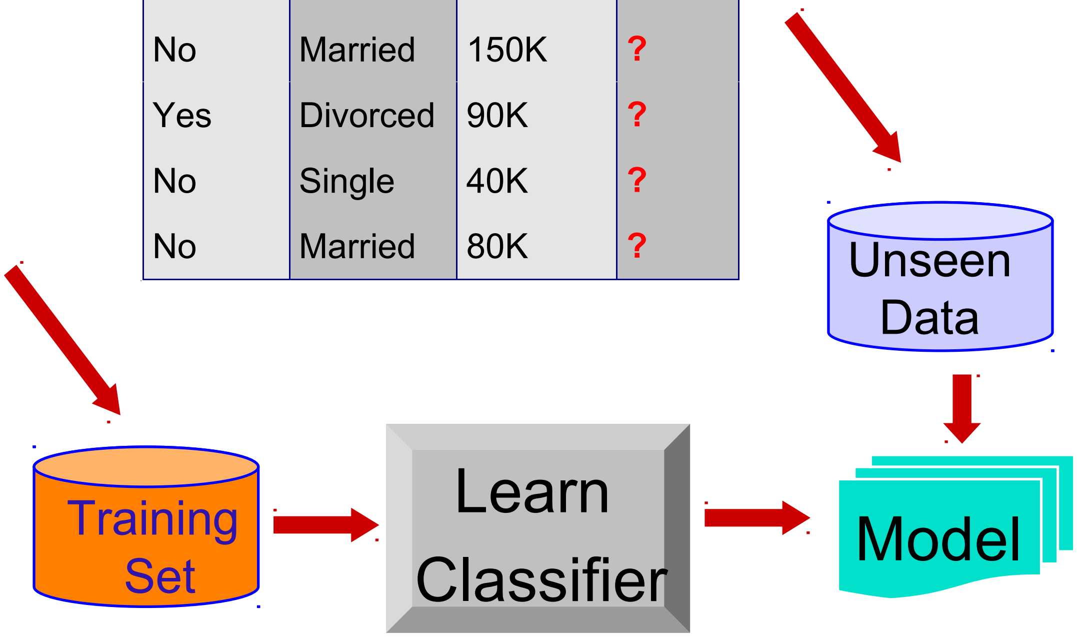
- Given a collection of records (training set)
  - each record contains a set of attributes
  - one of the attributes is the class (label) that should be predicted
- Find a *model* for class attribute as a function of the values of other attributes
- Goal: previously unseen records should be assigned 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

Class/Label Attribute

Tid	Refund	Marital Status	Taxable Income	Cheat
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

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



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

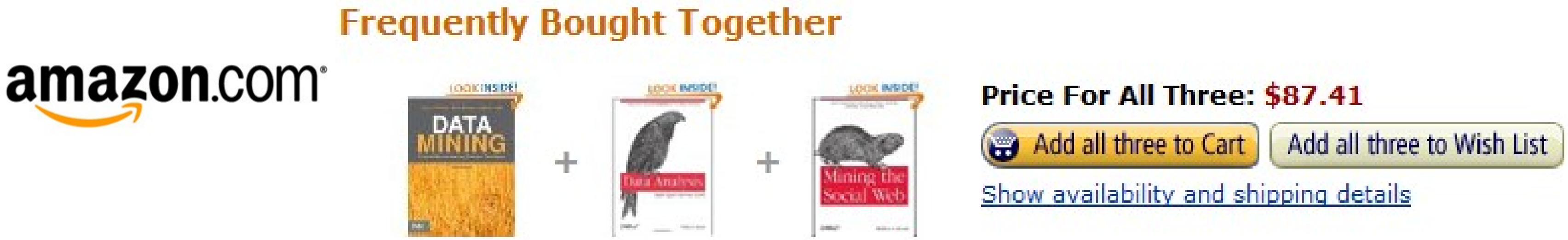
{Diaper, Milk} → {Beer}

{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

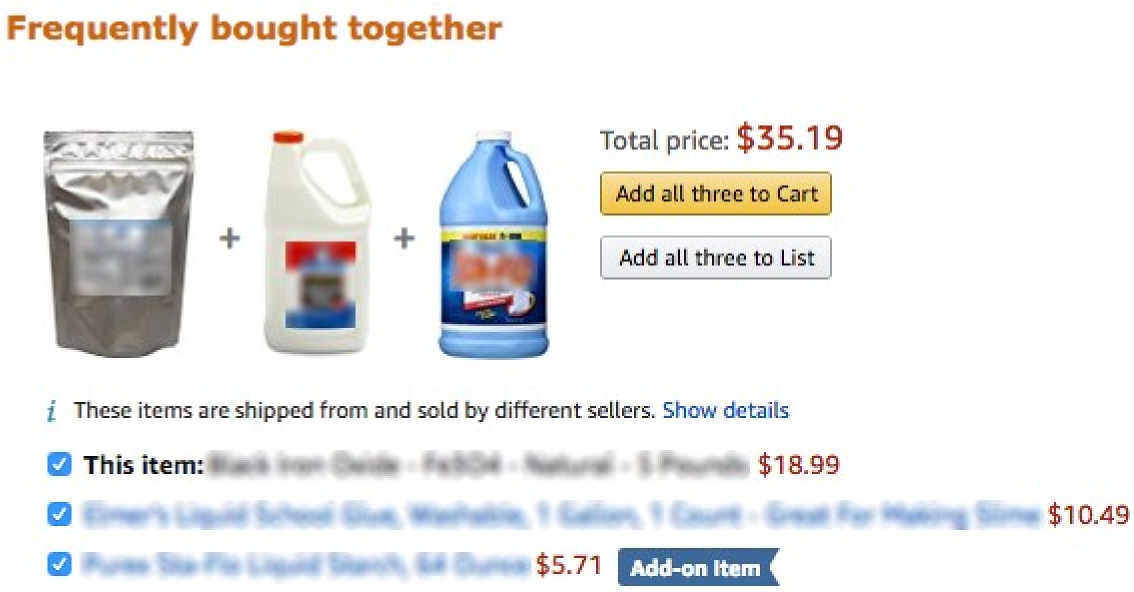


Frequently Bought Together

Price For All Three: $87.41

Association Rule Discovery: Applications

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



Frequently bought together

Total price: $35.19

i These items are shipped from and sold by different sellers. Show details

☑ **This item:** [redacted] $18.99

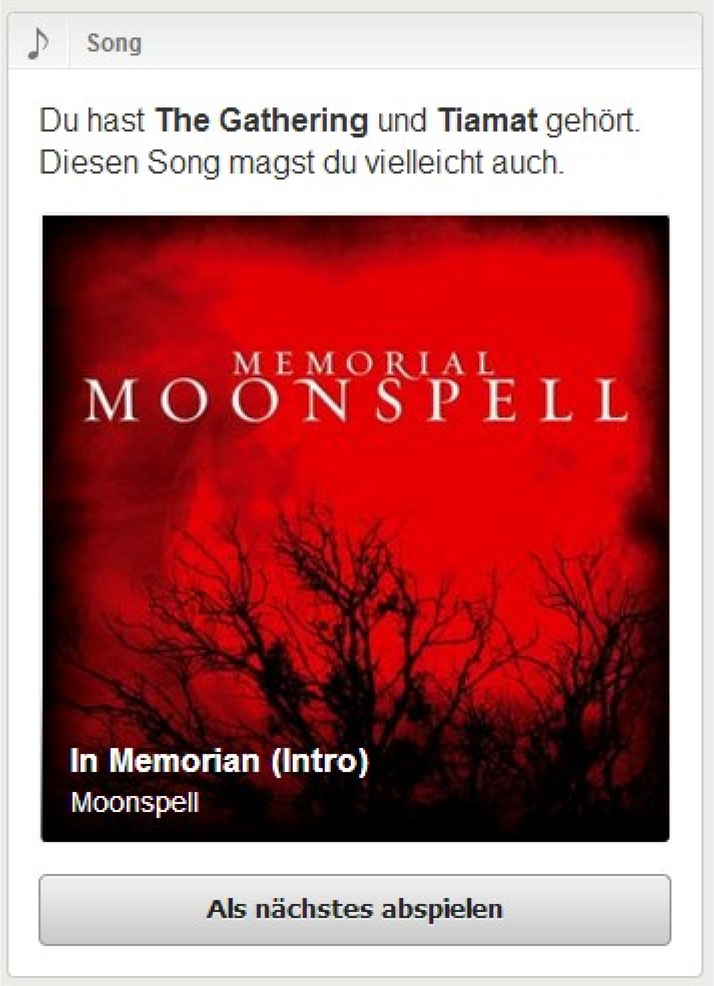
☑ [redacted] $10.49

☑ [redacted] $5.71 Add-on Item

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

Association Rule Discovery: Applications

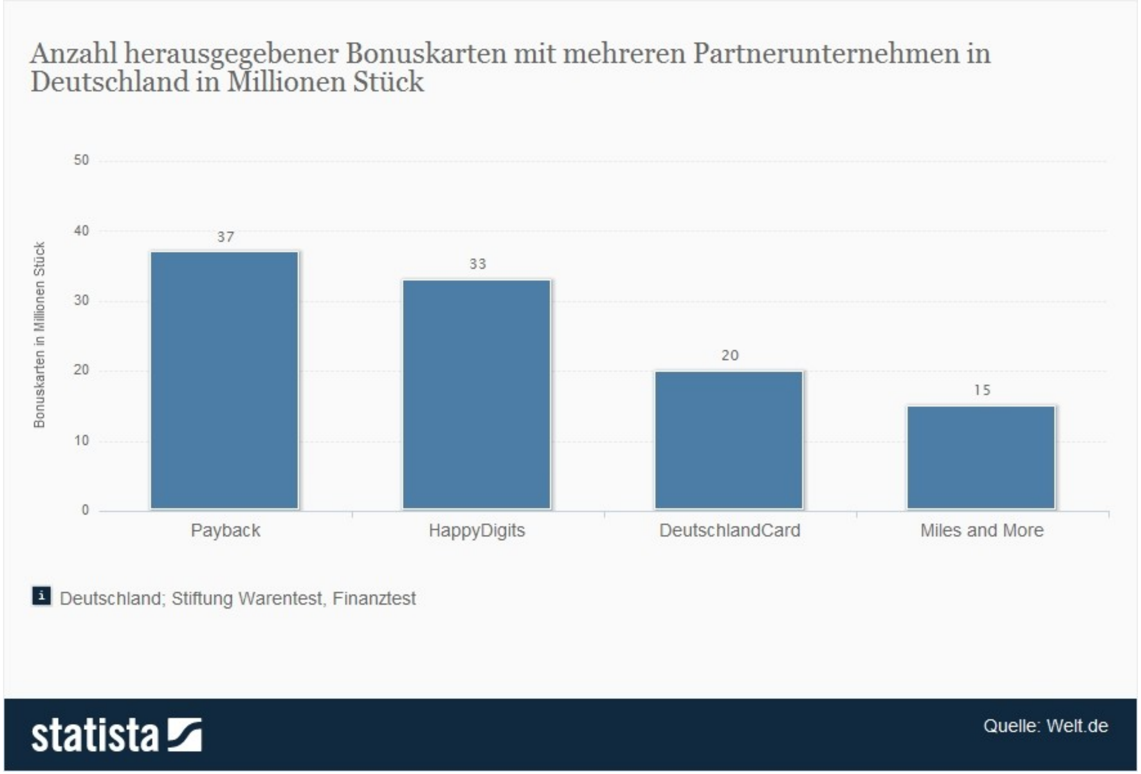
- Content-based recommendation
  - requirement: much data
  - e.g., Amazon transactions, Spotify logfiles



Du hast **The Gathering** und **Tiamat** gehört. Diesen Song magst du vielleicht auch.

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/

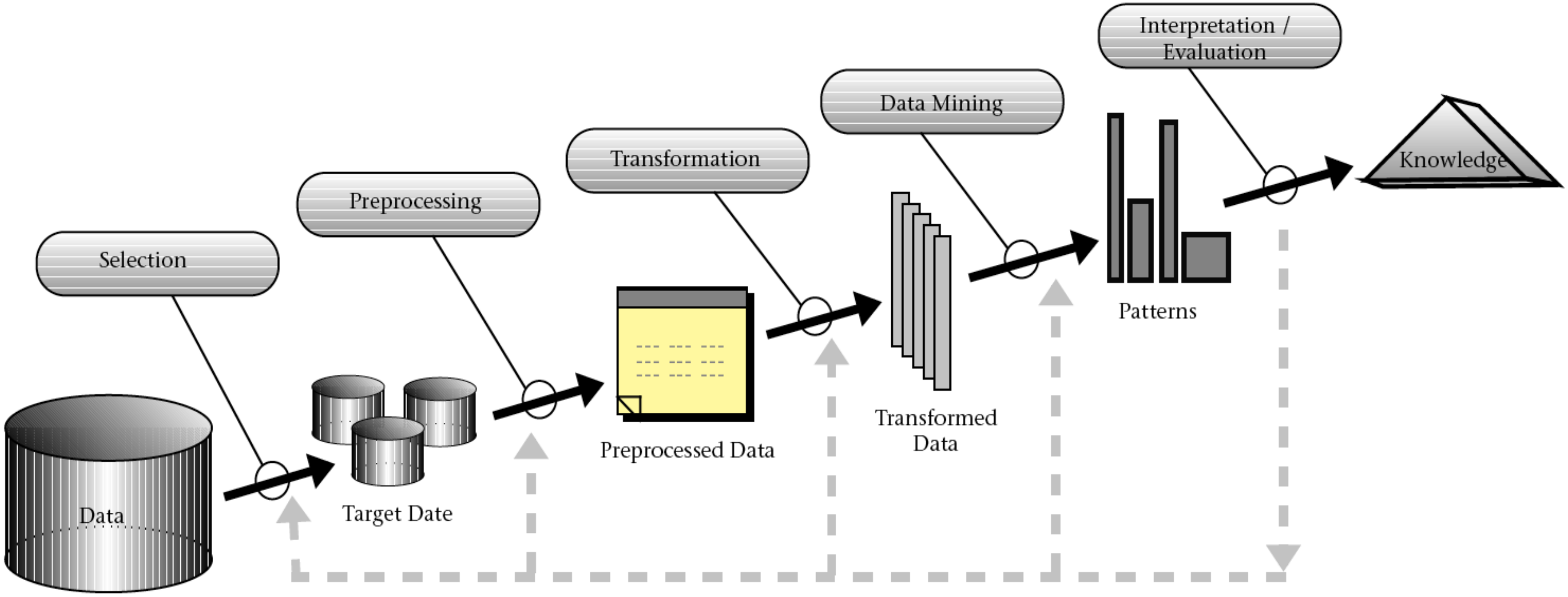
Association Rule Discovery: Applications

- 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

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

- Outcome:
  - massive buying of gift cards in a convenience store was reported to tax authorities

The Data Mining Process



Source: Fayyad et al. (1996)

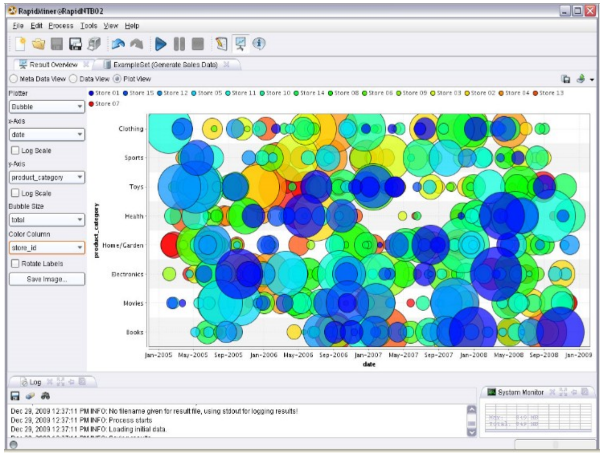
The Data Mining Process

- 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

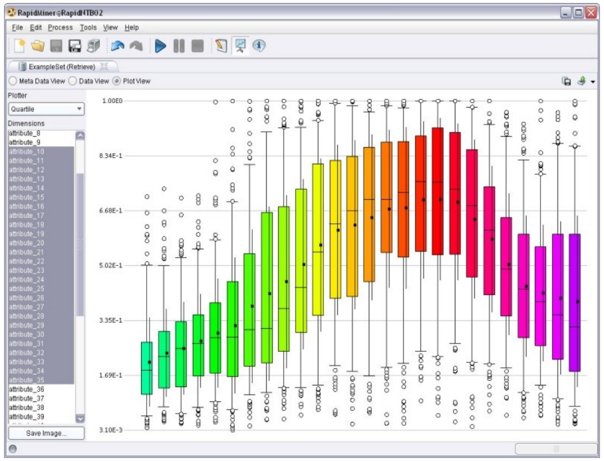
Selection and Exploration

- 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


RapidMiner@RapidNTB02
File Edit Process Tools View Help
Result Overview
ExampleSet (Generate Sales Data)
○ Meta Data View ○ Data View ● Plot View
● Store 01 ● Store 15 ● Store 12 ● Store 05 ● Store 11 ● Store 10 ● Store 14 ● Store 08 ● Store 06 ● Store 09 ● Store 03 ● Store 02 ● Store 04 ● Store 13 ● Store 07
Clothing
Sports
Toys
Health
Home/Garden
Electronics
Movies
Books
product_category
Jan-2005 May-2005 Sep-2005 Jan-2006 May-2006 Sep-2006 Jan-2007 May-2007 Sep-2007 Jan-2008 May-2008 Sep-2008 Jan-2009
date
Save Image...
Log
System Monitor

Plotter: Bubble
x-Axis: date
□ Log Scale
y-Axis: product_category
□ Log Scale
Bubble Size: total
Color Column: store_id
□ Rotate Labels

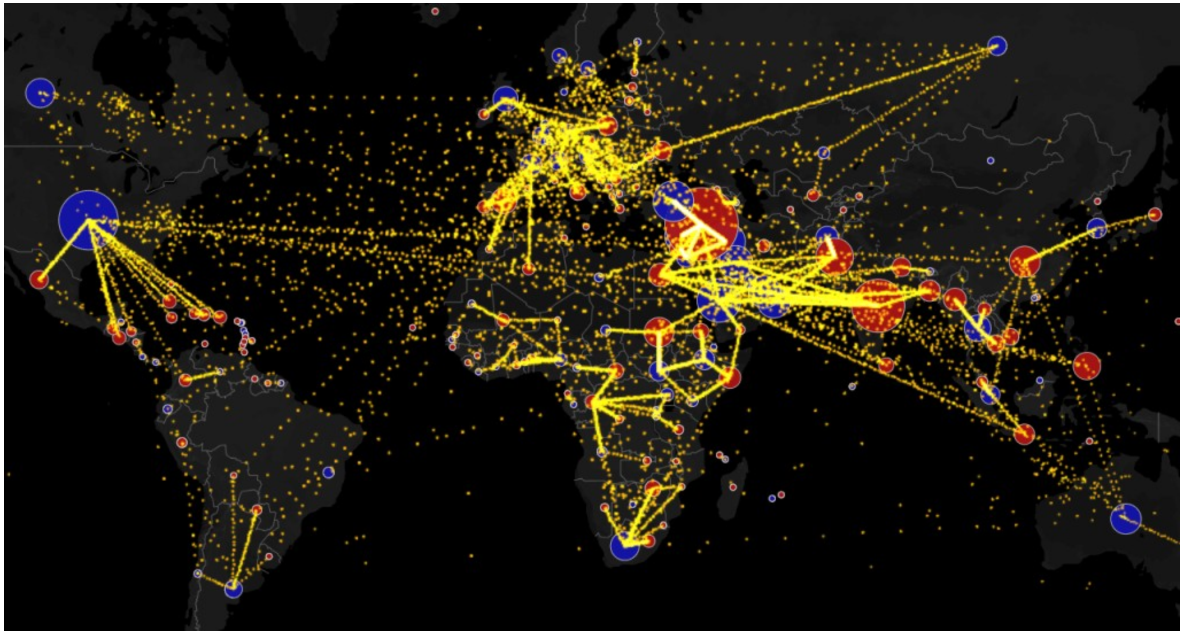
```
Dec 29, 2009 12:37:11 PM INFO: No filename given for result file, using stdout for logging results!
Dec 29, 2009 12:37:11 PM INFO: Process starts
Dec 29, 2009 12:37:11 PM INFO: Loading initial data.
```



Plotter: Quartile
Dimensions: attribute_8, attribute_9, attribute_10, attribute_11, attribute_12, attribute_13, attribute_14, attribute_15, attribute_16, attribute_17, attribute_18, attribute_19, attribute_20, attribute_21, attribute_22, attribute_23, attribute_24, attribute_25, attribute_26, attribute_27, attribute_28, attribute_29, attribute_30, attribute_31, attribute_32, attribute_33, attribute_34, attribute_35, attribute_36, attribute_37, attribute_38, attribute_39

Selection and Exploration

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



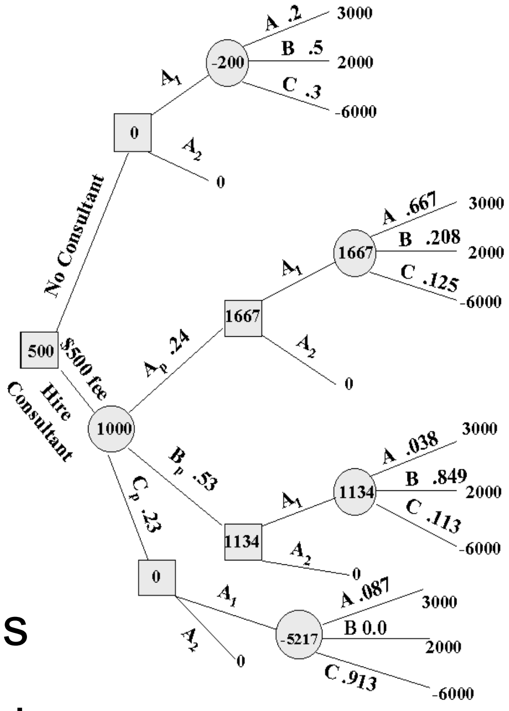
http://metrocosm.com/global-migration-map.html

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?