Technische Universität Berlin

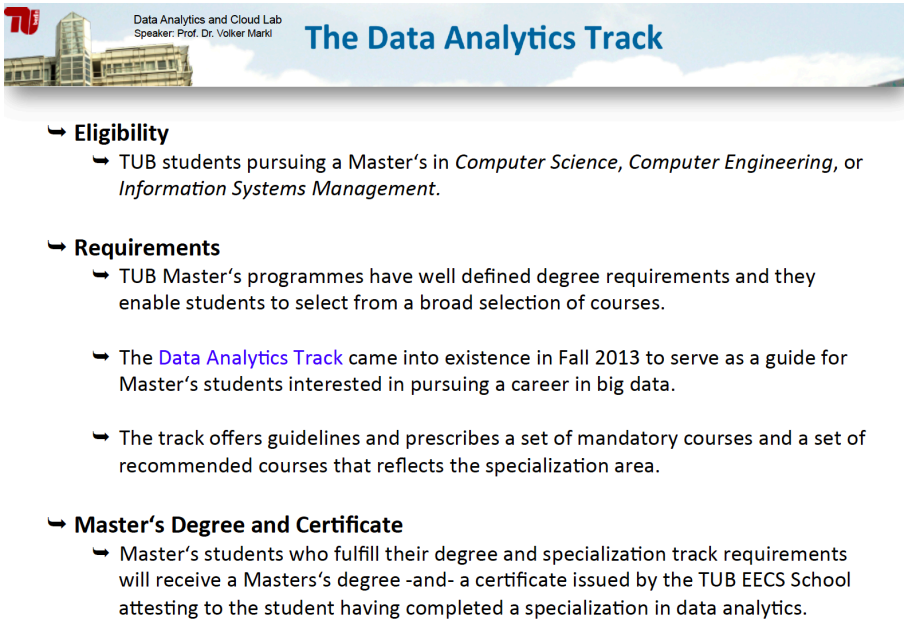
Introduction to Data Science

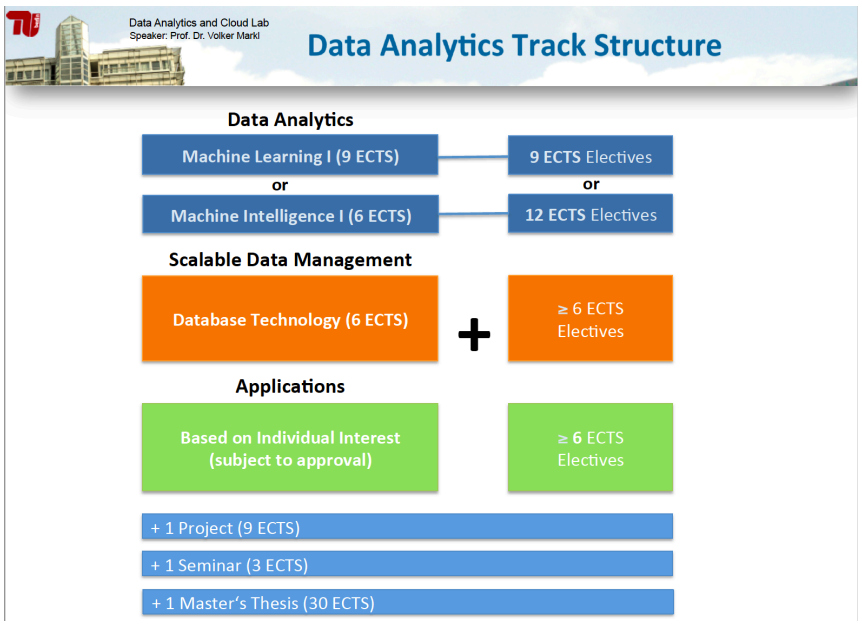
Juan Soto, juan.soto@tu-berlin.de, 28th April 2017

Technische Universität Berlin

Comment About the TU Berlin Data Analytics Track

For more details click here.





Technische Universität Berlin

Pioneer Perspectives

Technische Universität Berlin

John Hopcroft, Turing Award Winner (1986)

- **Future Directions in Computer Science Research**

  John Hopcroft, Cornell University

  20th September 2013, Big Data Workshop,

  Data Analytics Laboratory, TU Berlin

Computer Science is changing

Early years

- Programming languages
- Compilers
- Operating systems
- Algorithms
- Data bases

Emphasis on making computers useful

The future years

- Tracking the flow of ideas in scientific literature
- Tracking evolution of communities in social networks
- Extracting information from unstructured data sources
- Processing massive data sets and streams
- Extracting signals from noise
- Dealing with high dimensional data and dimension reduction
- The field will become much more application oriented

Technische Universität Berlin

Foundations of Data Science Book

- Foundations of Data Science, A. Blum, **J. Hopcroft**, and R. Kannan, 9th June 2016
- Fundamental change is taking place in CS and the focus is more on applications
- Increasingly researchers of the future will be involved with using computers to understand and extract usable information from massive data arising in applications
- Major change, switch from discrete math to probability, statistics, numerical methods
- Emphasis of the book is the mathematical foundations rather than applications
- Key areas include: **mathematics** (matrix algebra, SVD), **random graphs** to analyze large structures (e.g., web, social networks), **machine learning**, **streaming** and **sampling** (drawing good samples efficiently, estimating statistical and linear algebra quantities with such samples), and **clustering**.

Technische Universität Berlin

A Statistician's Perspective: Tukey Centennial Workshop

- 50 Years of Data Science, Princeton, Sept. 2015
- Greater Data Science

1. Data Exploration/Preparation
2. Data Representation/Transformation
3. Computing with Data
4. Data Modeling
5. Data Visualization and Presentation
6. **Science about Data Science**

Meta-analysts learned that > 5% of the conclusions in the scientific literature are incorrect, effects overstated, results irreproducible

50 years of Data Science

David Donoho

Sept. 18, 2015
Version 1.00

Abstract

More than 50 years ago, John Tukey called for a reformation of academic statistics. In 'The Future of Data Analysis', he pointed to the existence of an as-yet unrecognized *science*, whose subject of interest was learning from data, or 'data analysis'. Ten to twenty years ago, John Chambers, Bill Cleveland and Leo Breiman independently once again urged academic statistics to expand its boundaries beyond the classical domain of theoretical statistics; Chambers called for more emphasis on data preparation and presentation rather than statistical modeling; and Breiman called for emphasis on prediction rather than inference. Cleveland even suggested the catchy name "Data Science" for his envisioned field.

A recent and growing phenomenon is the emergence of "Data Science" programs at major universities, including UC Berkeley, NYU, MIT, and most recently the Univ. of Michigan, which on September 8, 2015 announced a $100M "Data Science Initiative" that will hire 35 new faculty. Teaching in these new programs has significant overlap in curricular subject matter with traditional statistics courses; in general, though, the new initiatives steer away from close involvement with academic statistics departments.

This paper reviews some ingredients of the current "Data Science moment", including recent commentary about data science in the popular media, and about how/whether Data Science is really different from Statistics.

The now-contemplated field of Data Science amounts to a superset of the fields of statistics and machine learning which adds some technology for 'scaling up' to 'big data'. This chosen superset is motivated by commercial rather than intellectual developments. Choosing in this way is likely to miss out on the really important intellectual event of the next fifty years.

Because all of science itself will soon become data that can be mined, the imminent revolution in Data Science is not about mere 'scaling up', but instead the emergence of scientific studies of data analysis science-wide. In the future, we will be able to predict how a proposal to change data analysis workflows would impact the validity of data analysis across all of science, even predicting the impacts field-by-field.

Drawing on work by Tukey, Cleveland, Chambers and Breiman, I present a vision of data science based on the activities of people who are 'learning from data', and I describe an academic field dedicated to improving that activity in an evidence-based manner. This new field is a better academic enlargement of statistics and machine learning than today's Data Science Initiatives, while being able to accommodate the same short-term goals.

Technische Universität Berlin

Guide

1. Ubiquity of Data
2. Big Data
3. Data Mining (DM) Basics
4. Real World Challenges
5. Ten DM Challenges
6. Analytics Libraries / Systems
7. Conclusion

Technische Universität Berlin

1. Ubiquity of Data

Data is the New Oil, Enabling a Data Driven Economy

Technische Universität Berlin

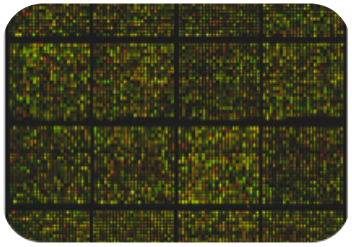
Data Arises from Diverse Sources

What are some examples?

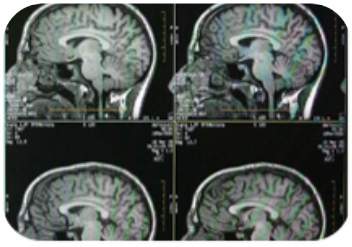
Technische Universität Berlin

Examples include …

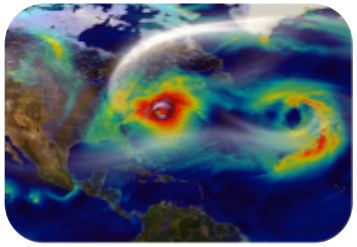
Web



Genomics



Neuroscience

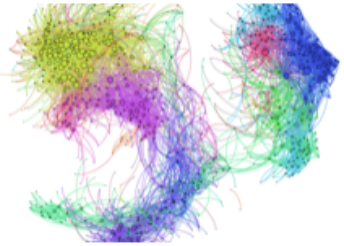


Simulations

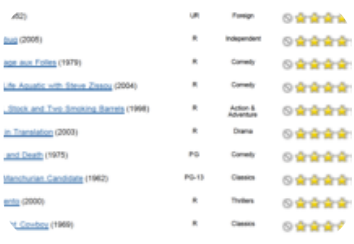
Sensors

Medical Devices

Mobile Devices



Social Networks



User Ratings

Purchase Histories

Technische Universität Berlin

Additional Sources of Data

Healthcare
- home-based monitoring
- integration across providers

Urban Planning
- fusion of high-fidelity geographical data

Intelligent Transportation
- analysis and visualization of live and detailed road network data

Environmental Modeling
- sensor networks ubiquitously collect data

Energy Saving
- unveiling patterns of use

Machine Translation Between Natural Languages
- analysis of large corpora

Education
- online courses

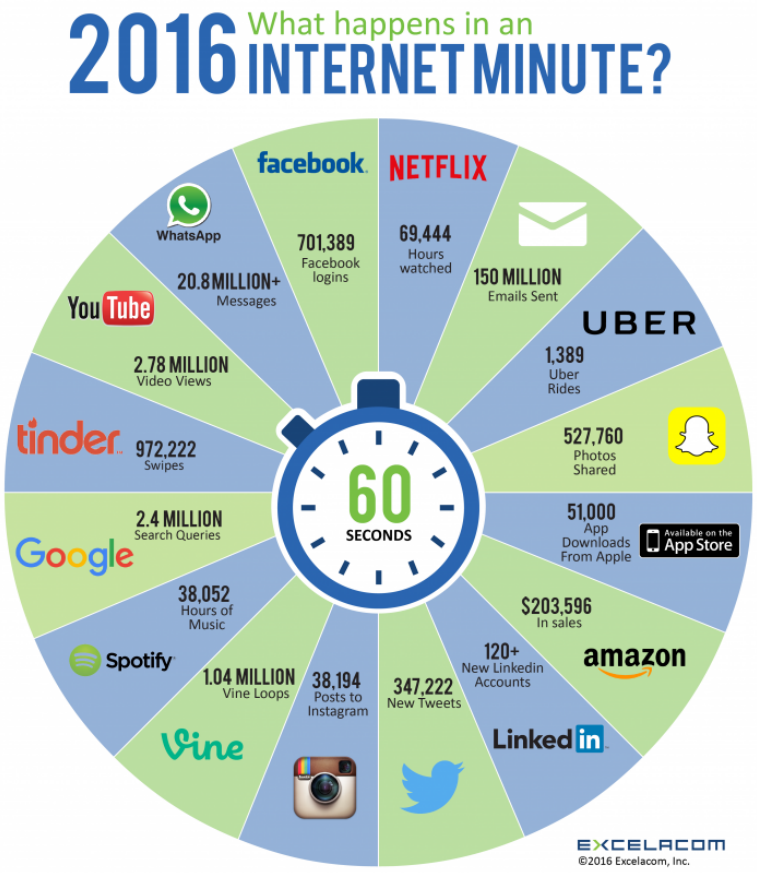
Risk Analysis in Finance
- analysis of a web of contracts to find dependencies between financial entities

Homeland Security
- analysis of social networks
- analysis of financial transactions

Computer Security
- analysis of logged events

Technische Universität Berlin

Data Amounts are Growing at a Rapid Pace



150M emails sent

20.8M+ **WhatsApp** messages

2.4M **Google** search queries

347K new tweets on **Twitter**

51K **Apple** app downloads

701K **Facebook** logins

excelacom.com/resources/blog/2016-update-what-happens-in-one-internet-minute

Technische Universität Berlin

Explosion in the Volume of Digital Data Created

- social networks, mobile applications, cloud computing, sensor networks
- RFID, IoT, imaging technologies, gene sequencing, remote sensing, LHC
- **Google** answers ca. 300K searches, 126 hrs. uploaded to YouTube / min.
- **YouTube** will serve 140K+ video views / min.
- **Twitter** will create ca. 700 new user accounts, 350K+ generate tweets / min.
- **LinkedIn** 11K searches performed / min.
- **Facebook** records 3.2M likes, stores 3.4M+ posts, generates 4GB of data / min.
- **Walmart** handles 1M+ customer transactions/hour, produces 2.5 PB data daily
- **Alibaba** stored 100PB+ processed data (2014)
- **Survey telescope** generates 30 TB+ data daily
- ca. 30M RFID tags are created daily
- **IDC predicts the worldwide volume of data will reach 40 ZB in 2020**

Decimal	
Value	SI
1000	k kilo
1000^2	M mega
1000^3	G giga
1000^4	T tera
1000^5	P peta
1000^6	E exa
1000^7	Z zetta
1000^8	Y yotta

Source: **Big Data 2.0 Processing Systems**, S. Sakr (p.1), Springer, 2016, springer.com/gp/book/9783319387758

Technische Universität Berlin

Data is the New Oil: Enabling a Data Driven Economy

Smart Web (Every Minute)

- > 400 Hours of New Videos on YouTube

Exponential growth of digital data.

In the midst of a data explosion.

Smart Phone

- 60GB generated each year per phone

Smart Factory

- 23000GB/year in smart factories

Smart Grid

- 2016: 343M smart meters, A network w/ 1M smart meters generates 1TB of data/yr.

Smart Mobility

- BMW receives 30GB/day from their cars

Technische Universität Berlin

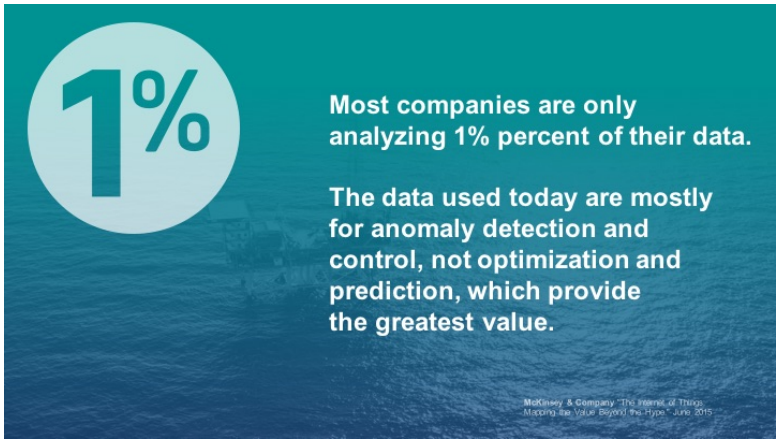
What is dark data?

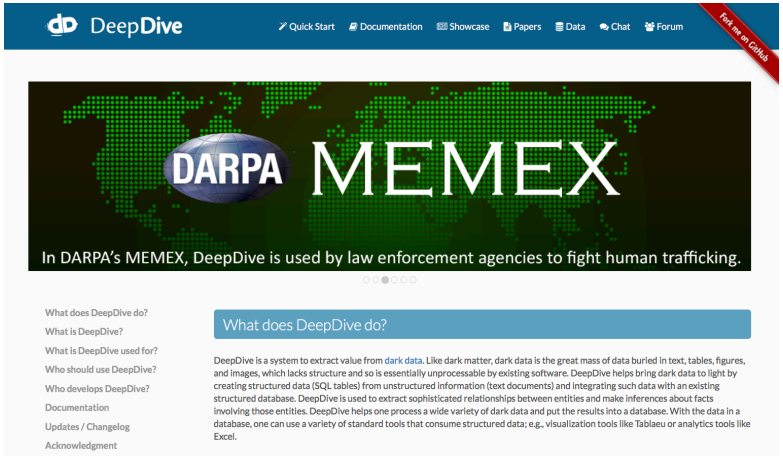
- Guesses?

Technische Universität Berlin

What is dark data?

- Gartner defines **dark data** as the **information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes** (e.g., business relationships and direct monetizing).
- Organizations often retain dark data for compliance purposes only.
- Storing and securing data typically incurs more expense (and sometimes greater risk) than value.





http://deepdive.stanford.edu

Technische Universität Berlin

THE VERGE

TRENDING NOW
Here's the tracklist for Drake's new album, which is just called VIEWS now

18 NEW ARTICLES

LOG IN | SIGN UP LONGFORM REVIEWS VIDEO TECH SCIENCE ENTERTAINMENT CARS DESIGN US & WORLD FORUMS

PREVIOUS STORY
The US is dropping 'cyberbombs' on ISIS

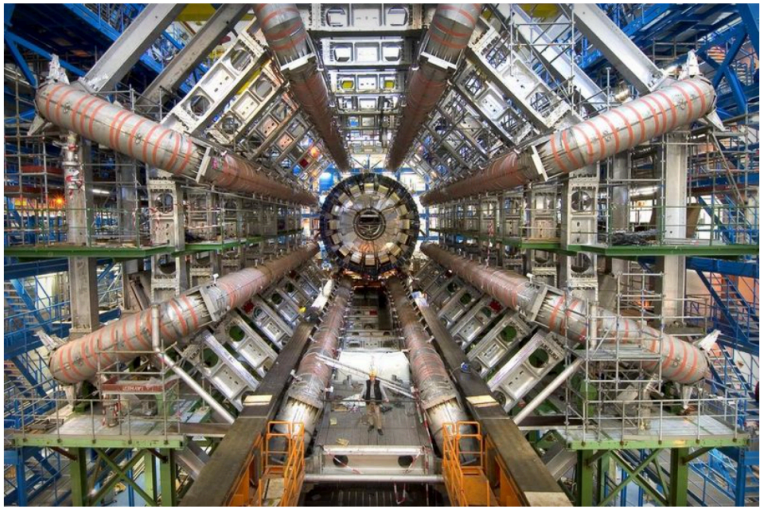
NEXT STORY
Logitech turned 160 gaming keyboards into a massive pixelated display

SCIENCE

11 COMMENTS

You can now download 300TB of data from the Large Hadron Collider

By James Vincent on April 25, 2016 05:24 am @jjvincent



THE LATEST HEADLINES

FCC approves real-time text proposal to aid people with disabilities

The FTC is investigating Venmo over 'deceptive or

theverge.com/2016/4/25/11501078/cern-300-tb-lhc-data-open-access

Name	Symbol	Binary
Kilobyte	KB	2^{10}
Megabyte	MB	2^{20}
Gigabyte	GB	2^{30}
Terabyte	TB	2^{40}
Petabyte[citation needed]	PB	2^{50}
Exabyte[citation needed]	EB	2^{60}
Zettabyte[citation needed]	ZB	2^{70}
Yottabyte[citation needed]	YB	2^{80}

Technische Universität Berlin

European Molecular Biology Laboratory (EMBL)

- Biology has become a data rich science. The amount and complexity of data produced in biology now exceeds that of any other scientific field.
- Individualized medicine based on patient genomes will have an enormous impact on healthcare. With breakthroughs in DNA sequencing technology, the number of sequenced genomes could reach > 1 Million within 5–10 years.
- The simultaneous generation and integration of this associated molecular and clinical data will provide an unprecedentedly rich set of “big data” for basic research and translation. Integration of these data will provide new research opportunities (e.g., identification of novel biomarkers or enabling the identification of causal relationships in molecular biology through analysing complex datasets, but will also come with significant technical and bioethical challenges.)

http://www.embl.de/training/events/2016/BIG16-01/index.html

Technische Universität Berlin

Deutsche Welle (DW) Documentary

http://www.dw.com/en/the-end-of-memory/av-36157732

DOCUMENTARIES

The end of memory?

The human race is producing digital data at an unprecedented rate. How can we store so much data? Storage media such as DVDs, CDs and hard drives have a lifespan of about ten years. Will our data still be preserved in 20 or 30 years' time?

Date 26.10.2016

Duration 42:30 mins.

Homepage Documentaries -

All videos Documentaries

Related Subjects Sci-Tech

Keywords Big Data, Documentation, Science, Data Storage, IT

Share Facebook Twitter Google+ More

Print Print this page

Permalink http://dw.com/p/2RiHE

Technische Universität Berlin

2. Big Data

Definition, Characteristics, Five Dimensions of Big Data,

Difference Between Statistics and Data Science

Technische Universität Berlin

Big Data: A Statistician's Viewpoint

- **Big data** is an *accumulation of data* that cannot be processed / handled using traditional data management processes / tools.

- A *big data management infrastructure* should ensure that the underlying hardware, software, and architecture have *the ability to enable learning (from data) using analytics*.

Technische Universität Berlin

Big Data Characteristics

Typically, how are big data characterized?

What are your thoughts?

Technische Universität Berlin

Big Data Characteristics

volume

- "data at rest"
- the amount of data with respect to the number of observations (size of the data) and the number of variables (dimensionality of the data)

variety

- "data in many forms"
- heterogeneity of data types (i.e., structured, semi-structured, unstructured)
- data sources that are either private or public
- examples include log files, text, web, images, video, audio

Technische Universität Berlin

Big Data Characteristics

velocity

- "data in motion" or "data in transit" (aka *streaming data*)
- … is concerned with the **data generation rate**
- … is concerned with the **rate which data arrives**
- … is concerned with the **timeframe in which they must be acted upon**
- requires appropriate data handling mechanisms

veracity

- "data in doubt"
- … is concerned with noise and processing errors, including the reliability (i.e., quality over time) and validity of the data

Technische Universität Berlin

The Five Dimensions of Big Data

In terms of the impact of big data on our lives …

How can we succinctly capture the impact?

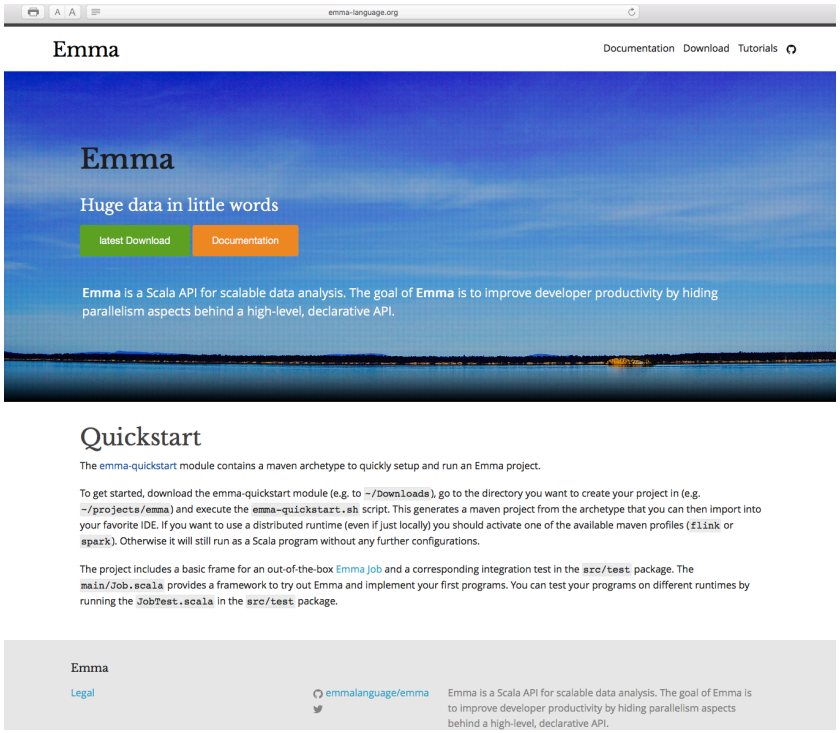
What are five dimensions?

Technische Universität Berlin

Dimension 1: Technology

There is a need for scalable systems and platforms for data analysis, novel data analysis methods, and **technologies to help overcome the skills gap**.

That is, enable data analysis to be accessible to a wider audience.



Technische Universität Berlin

Dimension 2: Applications

Many novel applications are emerging in the information economy, such as information marketplaces, which refine and sell enriched data.

Information marketplaces effectively bootstrap the information economy.

Other examples, include **personalized medicine**, **Industry 4.0**, and **digital humanities.**

Technische Universität Berlin

Dimension 3: Economic

The challenges and opportunities in the economic dimension lie in new business models and content delivery paradigm shifts.

For example, **information pricing** and the **role of open-source software**.

Technische Universität Berlin

Dimension 4: Legal

From a legal perspective, big data will present many challenges with respect to ownership, liability, and insolvency.

In addition to prevalent issues, such as **privacy** and **security**.

Technische Universität Berlin

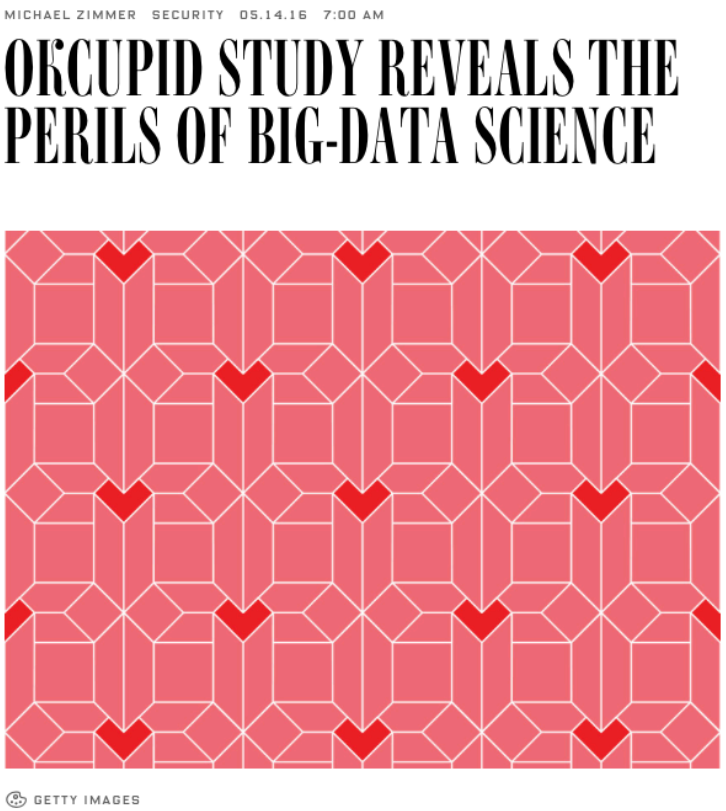
Dimension 5: Social

Data driven innovation will have a profound impact on society as a whole with respect to **social interaction**, **news**, and **democratic processes**, among others.

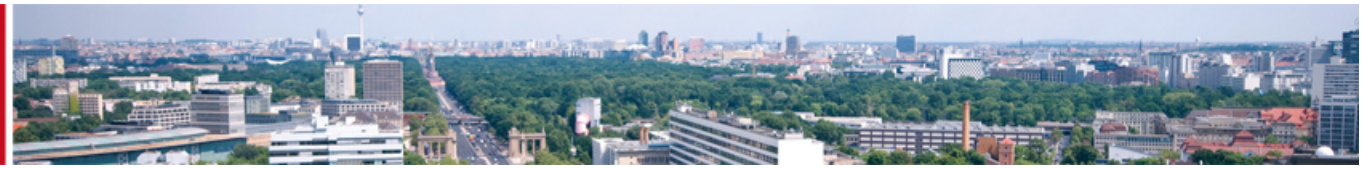
Technische Universität Berlin

Study Reveals the Perils of Big-Data Science (WiReD)

- **8 May 2016**: Researchers publicly released a dataset of nearly 70,000 users of the online dating site **OkCupid**, including usernames, age, and gender. Asked whether the researchers attempted to anonymize the dataset, the Aarhus U. graduate student replied: "No. Data is already public."
- "… this data scraping study - lacks any sense of ethics, ... it gives technologists a bad name and sets back the important work of genuine data science. Small wonder that politicians hate us and ordinary people don't trust us."
  – L. Weinstein, 14 May 2016



wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/



The Five Dimensions of Big Data

Read the article located here:

bbdc.berlin/media-press/blog-articles/#c1896

Technische Universität Berlin

Statistics v. Data Science

What are the key differences?

Technische Universität Berlin

Statistics v. Data Science

- **Statistics** is concerned with analyzing primary (e.g., experimental) data that have been collected to **explain** and **check** the **validity** of **specific ideas** (hypotheses).
  - primary data analysis or top-down (explanatory and confirmatory) analysis
  - i.e., **idea (hypothesis) evaluation or testing**

http://statoo.com/BigDataDataScience/#oesg
(Presentation, 21.10.2015, Slide 36)

Technische Universität Berlin

Statistics v. Data Science

- **Data science** (or data mining) in contrast is concerned with analyzing secondary (e.g., observational or 'found') data that have been collected for other reasons (and not 'under the control' of the investigator) to create new ideas (hypotheses).
  - secondary data analysis or bottom-up (exploratory and predictive) analysis
  - i.e., **idea (hypothesis) generation**

http://statoo.com/BigDataDataScience/#oesg
(Presentation, 21.10.2015, Slide 36)

Technische Universität Berlin

Caveat

The roots of data science can be traced back over several decades.

In this lecture, we will use the terms *data mining methods (process)* and *data science analytics (process)* interchangeably for simplicity.

For the time being, let us put aside the stark differences (e.g., principally, the tools used for data storage, processing, and analysis), among other things.

Technische Universität Berlin

3. Data Mining

Viewpoints, Algorithms, Characteristics

Technische Universität Berlin

What is Data Mining?

Technische Universität Berlin

Data Mining Encompasses …

- finding hidden information in a database
- conducting data driven discovery
- performing exploratory data analysis (EDA)
- What is EDA?

Technische Universität Berlin

Data Mining Encompasses …

- finding hidden information in a database
- conducting data driven discovery
- performing exploratory data analysis (EDA)

EDA is an approach (philosophy) for data analysis, which employs a variety of primarily graphical techniques to:

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important factors;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop parsimonious(?) models;

Technische Universität Berlin

Data Mining Encompasses …

- finding hidden information in a database
- conducting data driven discovery
- performing exploratory data analysis (EDA)

EDA is an approach/philosophy for data analysis, which employs a variety of primarily graphical techniques to:

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important factors;
4. detect outliers and anomalies;
5. test underlying assumptions:
6. develop parsimonious models;

A **parsimonious model** is a model that accomplishes a desired level of explanation or prediction with as few predictor variables as possible.

Technische Universität Berlin

Data Mining Encompasses …

- finding hidden information in a database
- conducting data driven discovery
- performing exploratory data analysis (EDA)

EDA is an approach/philosophy for data analysis, which employs a variety of primarily graphical techniques to:

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important factors;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop parsimonious models; and
7. determine optimal factor settings.

Technische Universität Berlin

EDA Principles (1 of 2)

- Insight is more important than summary statistics.
- "You can see a lot just by looking." Always use graphics.
- Never be content with just a quantitative analysis.
- Know the data (vs. trust the statistic/statistician).
- Always plot the raw data.
- Keep the statistics simple.
- Plots of simple summary statistics are a powerful combo.
- Parsimony: Do not overcomplicate (overfit) the model.
- Always validate a fitted model.
- General conclusions must be true in general.
- Make sure conclusions are not approach-dependent.
- Always plot subsets.

Drawn from a workshop held at NIST and offered by Dr. James Filliben, a descendant of John Tukey, Princeton Univ.

Technische Universität Berlin

EDA Principles (2 of 2)

- Validity of conclusions depends on validity of assumptions.
- Use techniques which have fewer assumptions.
- Every plot should have a lead-in question.
- Every plot should have a conclusion.
- For larger data sets, graphics is even more important.
- For every true conclusion, there exists a raw data plot that will convincingly show such
- To focus on the importance of a single factor, neutralize all other factors via 1) residuals or 2) subsetting
- Subsetting is the #1 tool for determining if a conclusion is robustly true.
- “The essence of data analysis is comparison, and the essense of comparison is juxtaposition (side-by-side).”

Technische Universität Berlin

Data Mining is a Process

1. "DM is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." – U. Fayyad
2. "[The process of] finding interesting structure (patterns, statistical models, relationships) in databases." – U. Fayyad, S, Chaduri, and P. Bradley
3. "… a knowledge discovery process of extracting previously unknown, actionable information from very large databases." – A. Zornes
4. "… a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions." – H. Edelstein

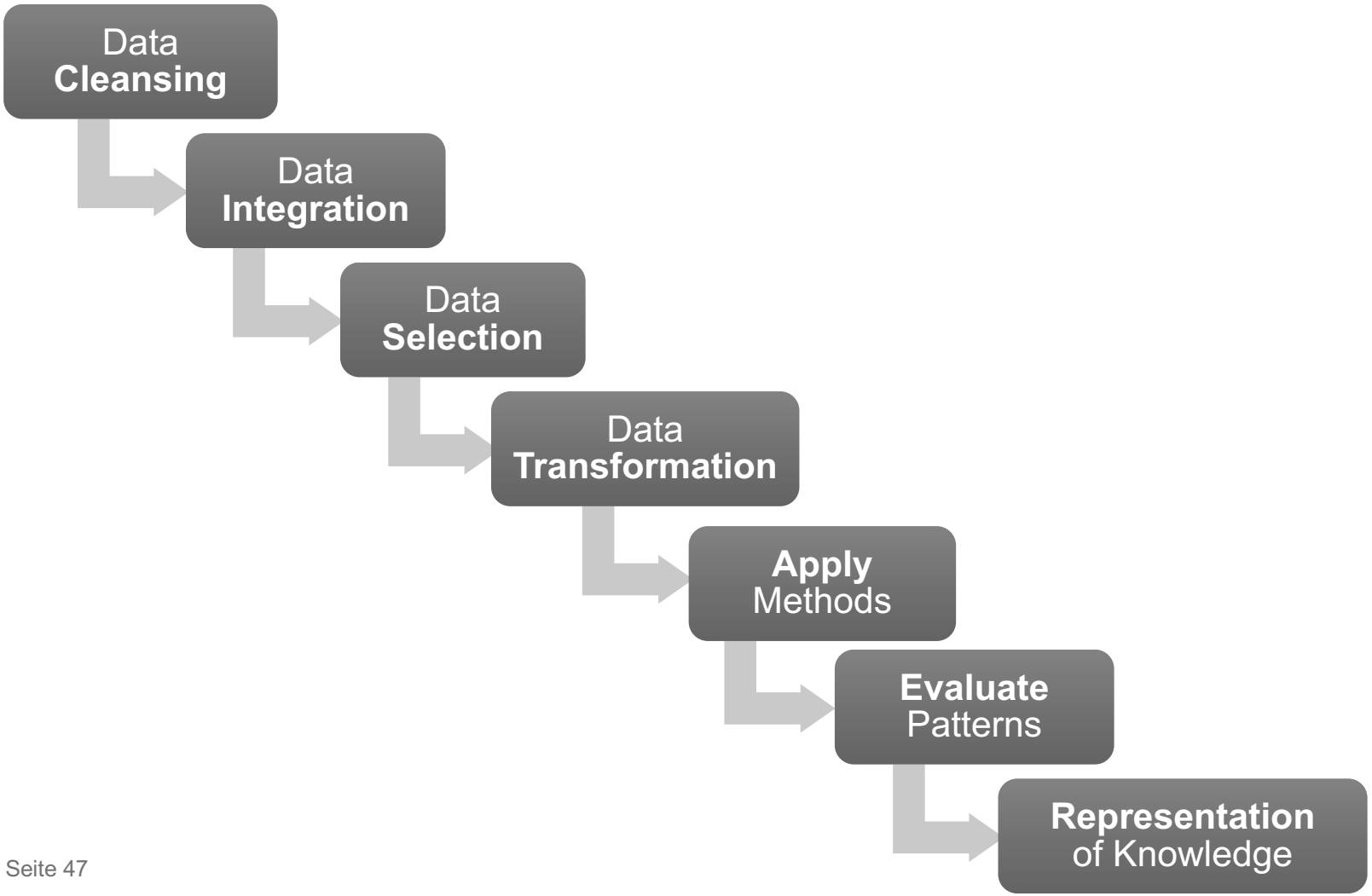
Technische Universität Berlin

The Number of Data Mining Steps

- How many are there?

Technische Universität Berlin

The **Seven** Data Mining Steps



Technische Universität Berlin

Step 1 Data Cleansing/Cleaning/Munging/Scrubbing/Wrangling

- concerned with data quality
- lack of integrity constraints, poor schema design, uniqueness
- data entry errors, misspellings, redundancy, duplicates
- contradictory / inconsistent values
- aims to remove noisy and irrelevant data
- …

Technische Universität Berlin

Step 2 Data Integration

- seeks to combine multiple data sources
- provide users with a unified view
- quite laborious and time intensive task

Technische Universität Berlin

Step 3 Data Selection

- the focus is on appropriately selecting data for an analysis
- selecting data subsets is trivial
- however, selecting the relevant data is non-trivial in many cases

Technische Universität Berlin

Step 4 Data Transformation

- the focus in on the transformation of the data selected in the previous step into a forms that are appropriate for the method employed
- examples of transformations include
  - ???

Technische Universität Berlin

Step 4 Data Transformation

- the focus in on the transformation of the data selected in the previous step into a forms that are appropriate for the method employed
- examples of transformations include
  - linear, logarithmic, square root mappings
  - performing summary or aggregation operations

Technische Universität Berlin

Steps 5-7 Methods, Evaluation, and Representation

Step 5 Applying the Analytics/Methods

- employing the algorithms on data to extract potentially useful patterns

Step 6 Pattern Evaluation

- identifying interesting patterns that represent knowledge

Step 7 Knowledge Representation

- discovered knowledge is visually represented to the user
- uses visualization techniques to help users understand and interpret the data mining results

Technische Universität Berlin

Data Mining: Algorithmic Matters

scalability

- algorithms that don't scale are of limited use

real-world data is …

- noisy, have missing attributes
- algorithms must be able to cope

update

- many data mining algorithms work with static data
- this is not always realistic

ease of use

- algorithms must be accessible

Technische Universität Berlin

What is a data analytic?

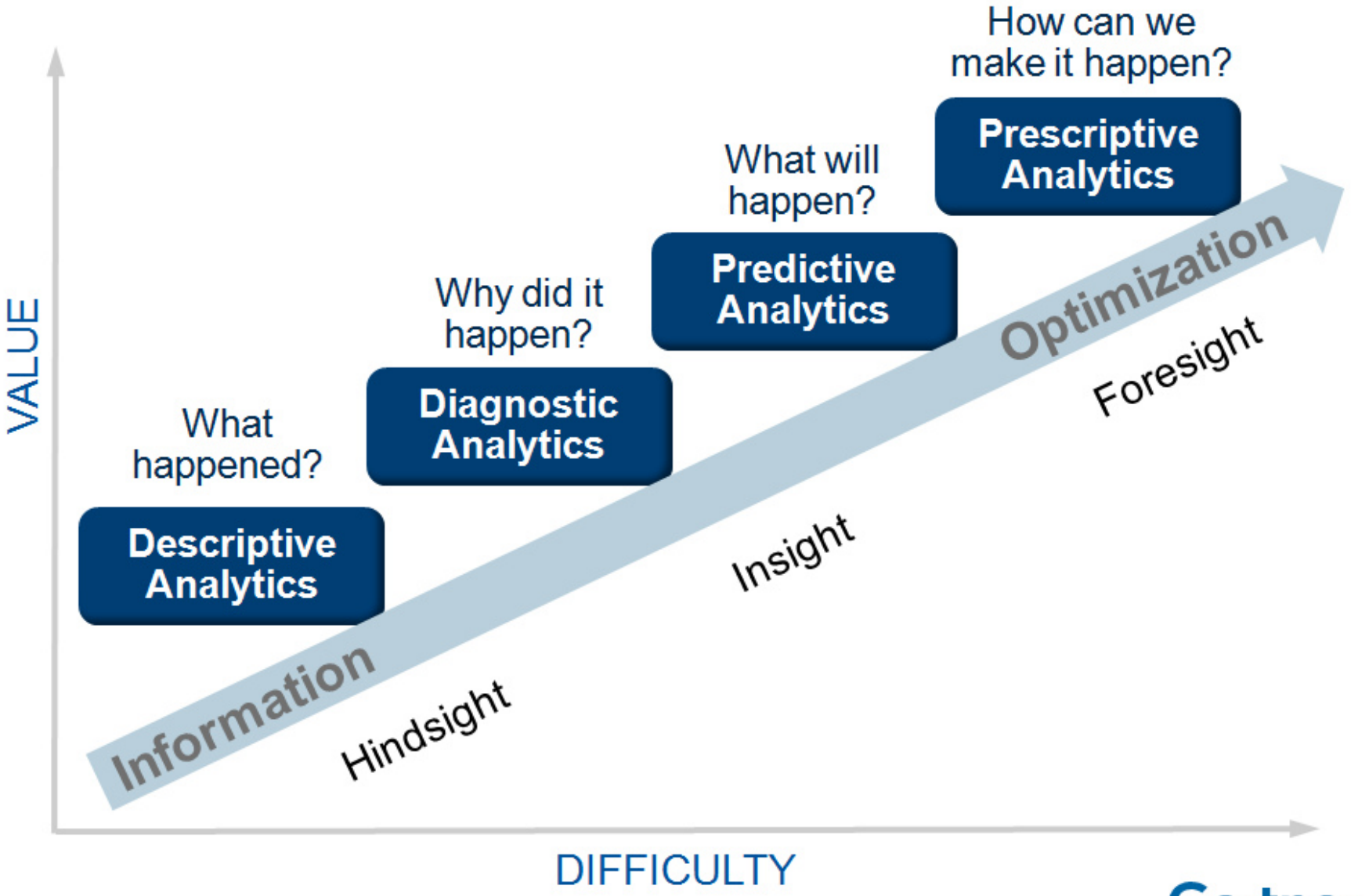
- Comments, thoughts, perspectives?

Technische Universität Berlin

What is a data analytic?

- Comments, thoughts, perspectives?
- Do you see a difference among the following terms or are they analogous?
  - data analysis technique
  - data mining method
  - machine learning algorithm
  - data analytic

Analytic Value Escalator



Gartner

Technische Universität Berlin

Data Analytics Definitions

- **Descriptive Analytics** – identify patterns in the data
  - examples include *frequencies*, *distributions*, *tabulations*, and *visualizations*
- **Predictive Analytics –** make predictions using historical data
  - examples include *classification*, *(non)linear regression*, *support vector machines*
- **Prescriptive Analytics** – recommend actions so as to maximize the expected value (e.g., utility) associated with an outcome
  - examples include stochastic models of uncertainty and optimal solutions
  - suitable for optimizing *production*, *scheduling*, and *inventory* in supply chains to ensure the right products are delivered at the right time, and optimize experience

Ref.: "Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions," National Academy of Sciences, 2016.

Technische Universität Berlin

Data Mining Algorithms: Broad Perspective

- fit models to designated data
- employ some criteria is used to select the best model
- fall under varying labels
  - descriptive v. predictive, supervised v. unsupervised
- particular classes of algorithms, include
  - classification, regression, time series, clustering
  - summarization, association rules, sequence discovery
  - anomaly/outlier detection, ranking, collaborative filtering

Technische Universität Berlin

Supervised v. Unsupervised Methods

Supervised Methods

- find patterns in observed data, then predict something from newly observed data
  - observe a collection of email messages that are categorized into spam and not spam
  - after learning something about them, we want to automatically categorize newly incoming messages

Technische Universität Berlin

Supervised v. Unsupervised Methods

Supervised Methods

- find patterns in observed data, then predict something from newly observed data
  - observe a collection of email messages that are categorized into spam and not spam
  - after learning something about them, we want to automatically categorize newly incoming messages

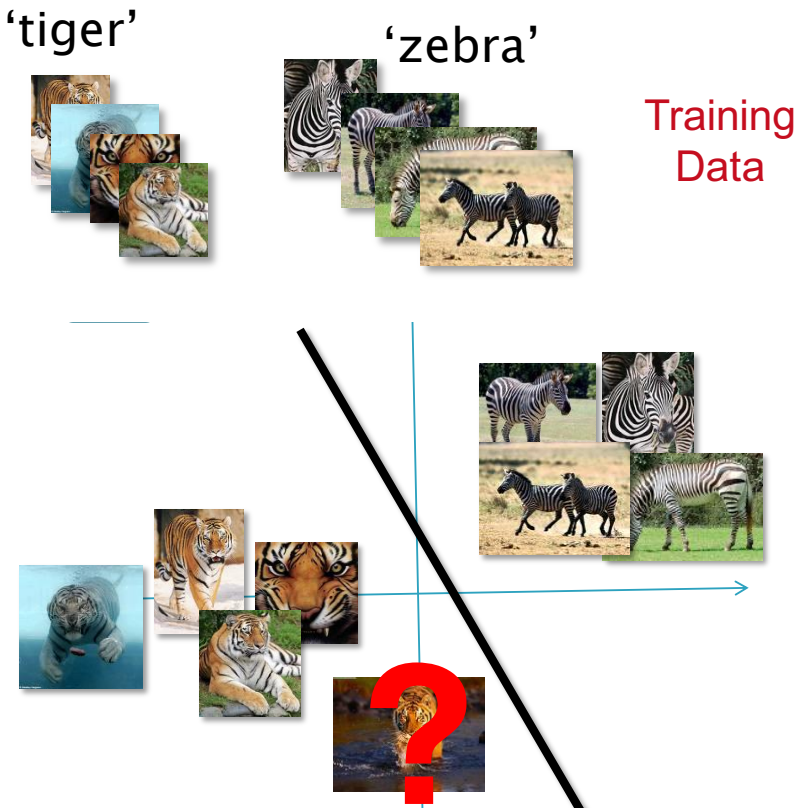
Unsupervised Methods

- find hidden structure in data
  - museum has images in their collection that they want to group by similarity into clusters
  - more difficult to evaluate than supervised learning

Technische Universität Berlin

Classification

- maps data into predefined *classes*
- i.e., *supervised learning,* since classes are determined before examining the data
- classes are defined based on data attributes / characteristics



Technische Universität Berlin

Types of Classification Algorithms

1. Decision Trees
2. K-Nearest Neighbors
3. Multinomial Logistic Regression
4. Random Forest
5. Artificial Neural Networks
6. Support Vector Machines

Classification

from data to discrete classes

Due to Prof. David Sontag, NYU

Spam filtering

data

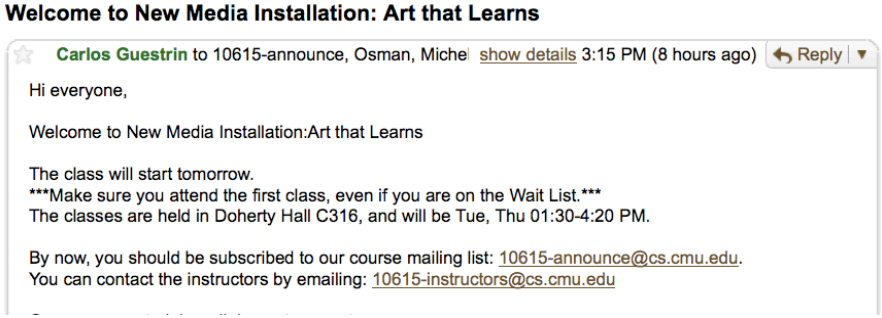
prediction

Osman Khan to Carlos show details Jan 7 (6 days ago) Reply

sounds good
+ok

Carlos Guestrin wrote:
Let's try to chat on Friday a little to coordinate and more on Sunday in person?

Carlos



Welcome to New Media Installation: Art that Learns

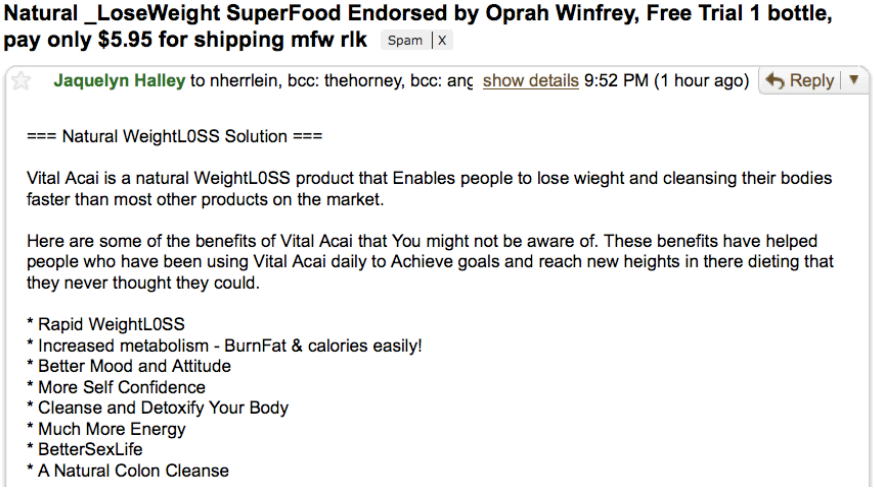
Carlos Guestrin to 10615-announce, Osman, Miche show details 3:15 PM (8 hours ago) Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu



Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | X

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: anç show details 9:52 PM (1 hour ago) Reply

=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

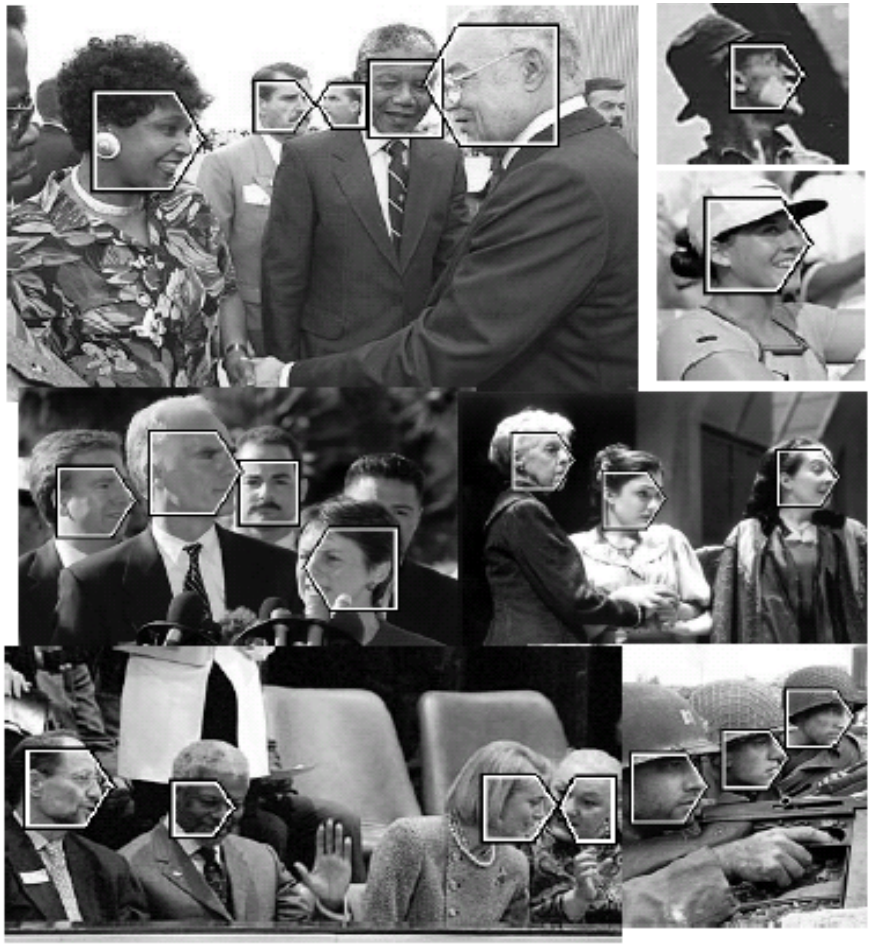
* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy
* BetterSexLife
* A Natural Colon Cleanse

Spam
vs.
Not Spam

Due to Prof. David Sontag, NYU

Face recognition

Example training images
for each orientation



Due to Prof. David Sontag, NYU

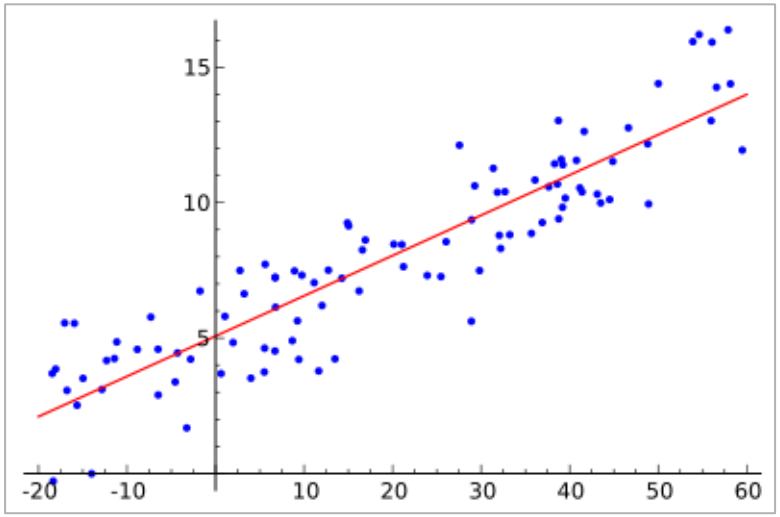
Weather prediction

Due to Prof. David Sontag, NYU

Technische Universität Berlin

Regression

- establish relationship between a dependent variable (response) and independent variables (factors)
  - useful in predicting a child's weight (response) based on height (factor)
- goal is to predict a response based on factors approaches include
  - linear regression (numerical data)
  - logistic regression (categorical data)
- identifies best fit that minimizes the error



linear regression assumes linear relationship exists:

$\mathbf{weight}_i = c_0 + c_1\ \mathbf{height}_i$

Technische Universität Berlin

Regression Based Algorithms

1. Linear Regression
2. Logistic Regression
3. Ridge Regression

Regression

predicting a numeric value

Due to Prof. David Sontag, NYU

Stock market

Due to Prof. David Sontag, NYU

Weather prediction revisited

Temperature

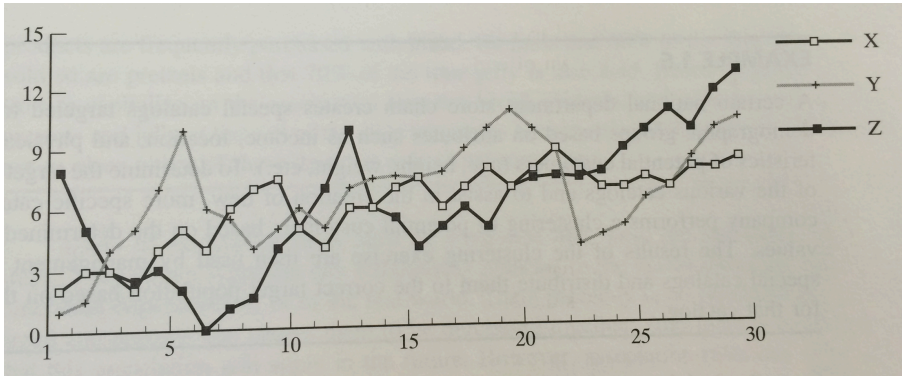
72° F

Due to Prof. David Sontag, NYU

Technische Universität Berlin

Time Series Analysis

- attribute examined over time (hourly, daily)
- plot used to visualize time series
- basic functions performed
  - distance measures used to determine similarity between time series (e.g., Euclidean distance)
  - structure of the line is examined to determine behavior (e.g., autocorrelation)
  - time series plot based on historical data used to predict future values



Time series plot: *day of the month* vs. *stock price* ($)

Observations

1. co. X is less volatile than co. Y and co. Z
2. stocks for Y and Z have similar behavior
3. Y behavior between days 6-20
4. Z behavior between days 13-27

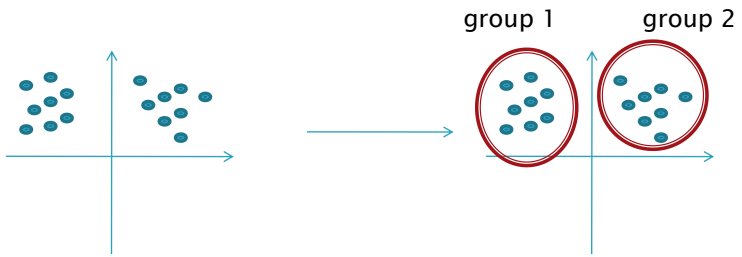
Technische Universität Berlin

Time Series Based Algorithms

1. Autoregressive Models
2. Integrated Models
3. Moving Average Models
4. Autoregressive Integrated Moving Average (ARIMA) Models
5. Autoregressive Moving Average (ARMA) Models

Technische Universität Berlin

Clustering



- i.e., unsupervised learning
- similar to classification, except the groups are not predefined
- clusters determined by the data
- partitions the data into groups that may or may not be disjoint
- clustering accomplished by determining similarity among the data on predefined attributes

Example

- store **creates catalogs** targeting demographic groups based on *location*, *customer characteristics*
- company **clusters** potential **customers** based on **attributes**
- results are used to **create custom catalogs** and distributed to the cluster (target population)

Technische Universität Berlin

Clustering Algorithms

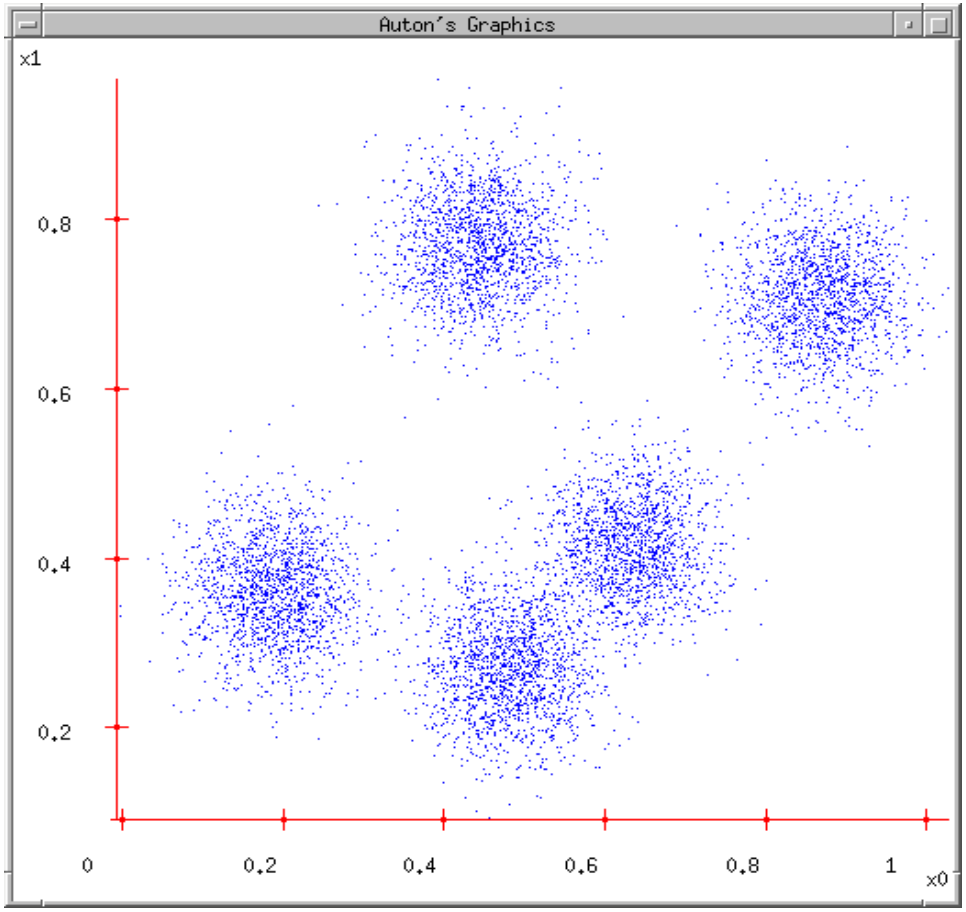
1. K-means
2. BFR (Bradley-Fayyad-Reina, K-means Variant)
3. CURE (Clustering Using Representatives)
4. Gaussian Mixture Model (GMM)
5. Hierarchical Clustering
6. Spectral Clustering

Clustering

discovering structure in data

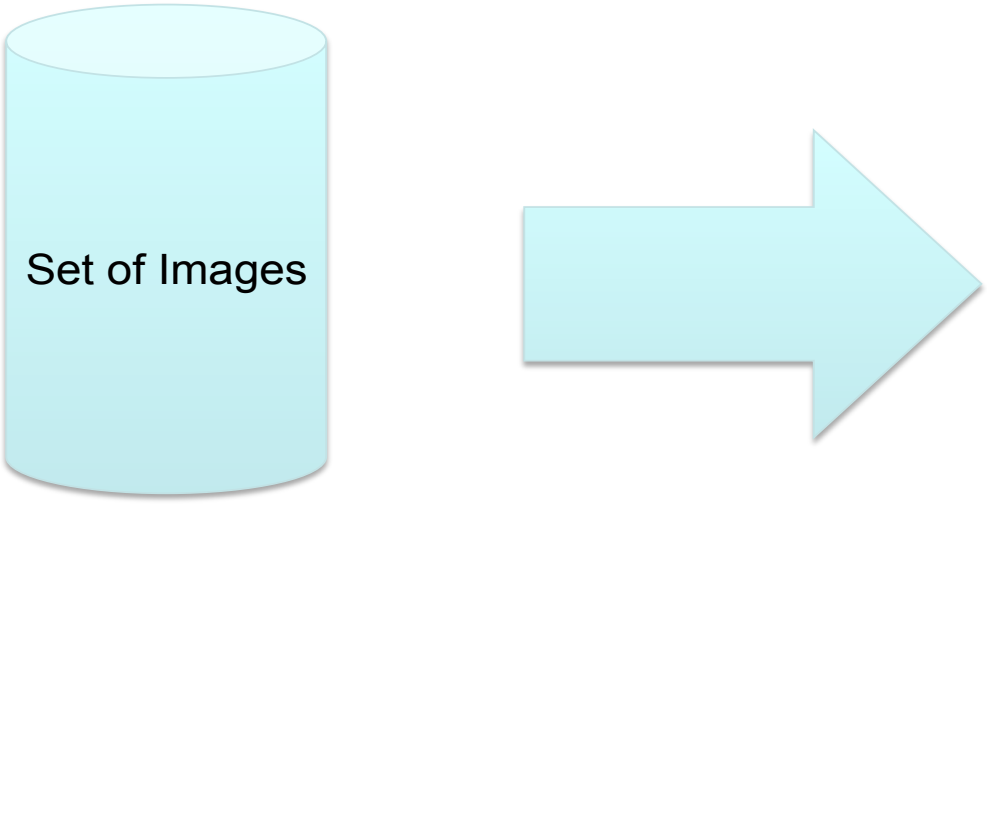
Due to Prof. David Sontag, NYU

Clustering Data: Group similar things



Due to Prof. David Sontag, NYU

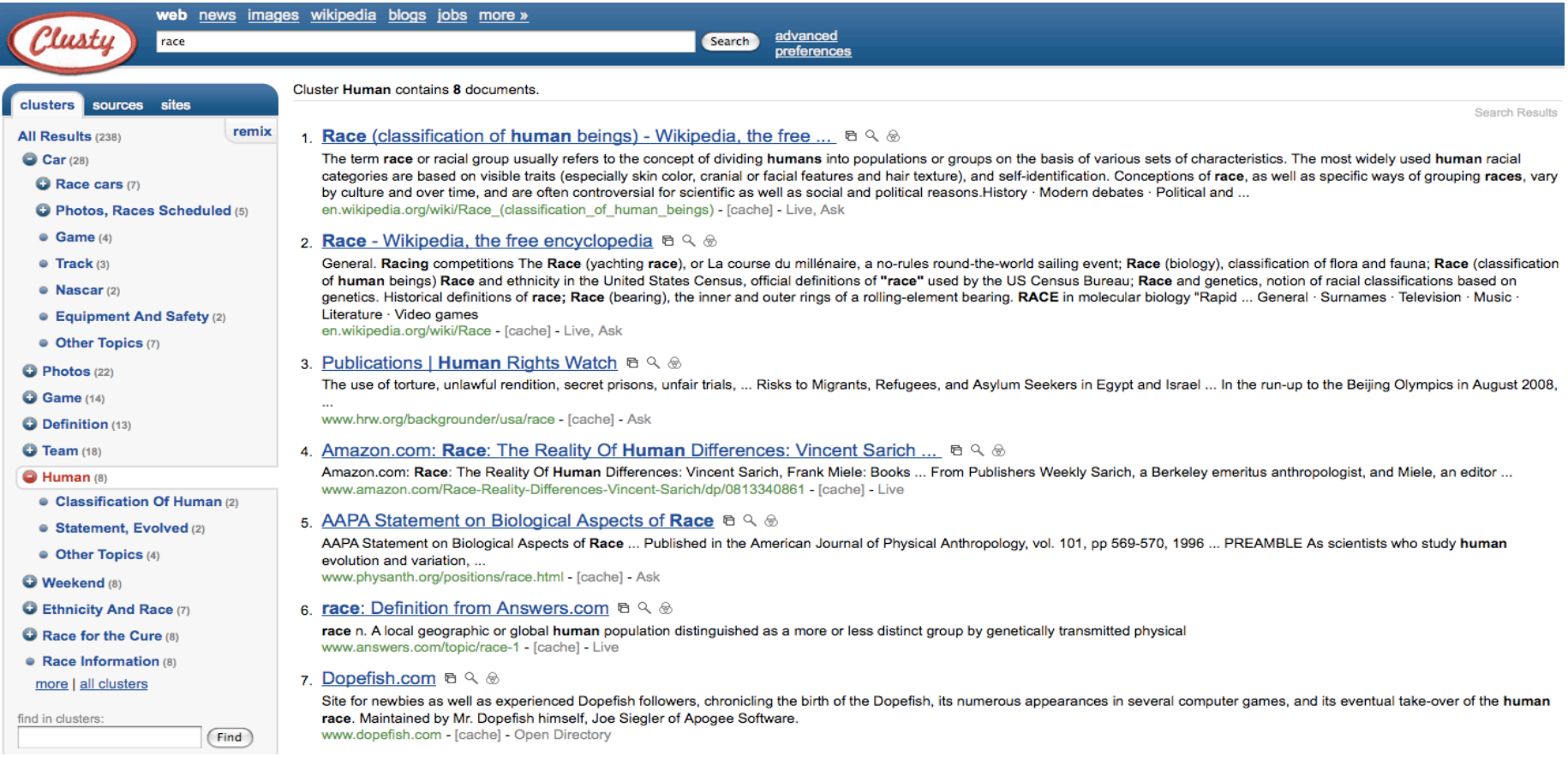
Clustering images



[Goldberger et al.]

Due to Prof. David Sontag, NYU

Clustering web search results



Due to Prof. David Sontag, NYU