



Introduction to Data Science

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



Comment About the TU Berlin Data Analytics Track

For more details click [here](#).

Data Analytics and Cloud Lab
Speaker: Prof. Dr. Volker Markl

The Data Analytics Track

Eligibility

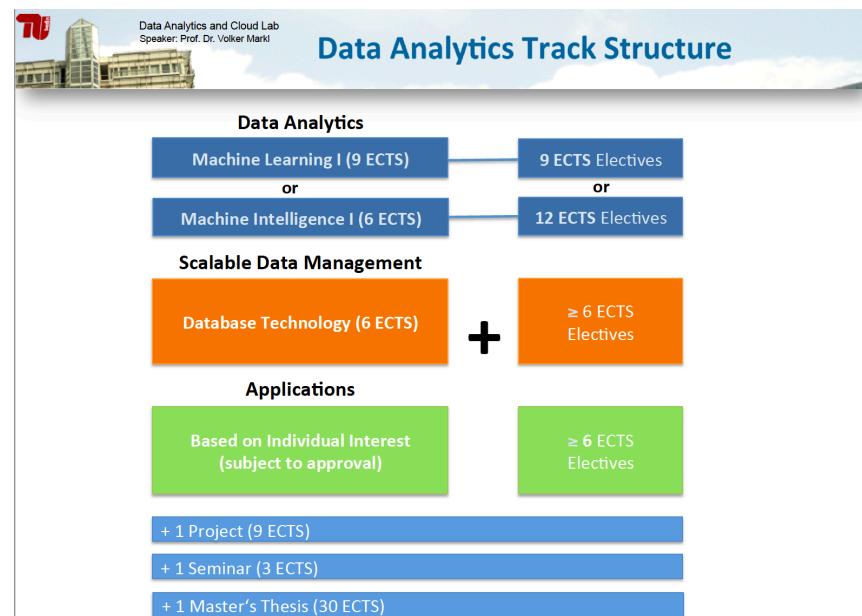
- TUB students pursuing a Master's in *Computer Science*, *Computer Engineering*, or *Information Systems Management*.

Requirements

- TUB Master's programmes have well defined degree requirements and they enable students to select from a broad selection of courses.
- The [Data Analytics Track](#) came into existence in Fall 2013 to serve as a guide for Master's students interested in pursuing a career in big data.
- The track offers guidelines and prescribes a set of mandatory courses and a set of recommended courses that reflects the specialization area.

Master's Degree and Certificate

- Master's students who fulfill their degree and specialization track requirements will receive a Masters's degree -and- a certificate issued by the TUB EECS School attesting to the student having completed a specialization in data analytics.





Pioneer Perspectives



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

The future years

Early years

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

Emphasis on making computers useful

- 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



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**.



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.



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



1. Ubiquity of Data

Data is the New Oil, Enabling a Data Driven Economy



Data Arises from Diverse Sources

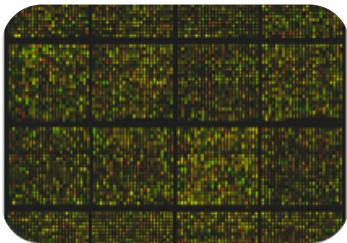
What are some examples?



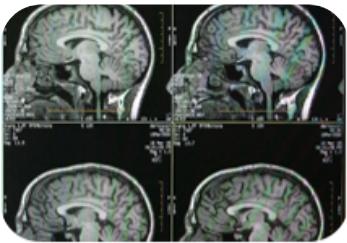
Examples include ...



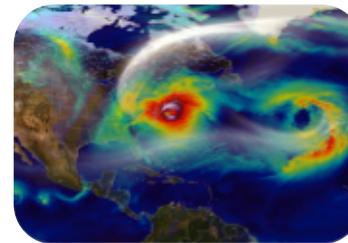
Web



Genomics



Neuroscience



Simulations



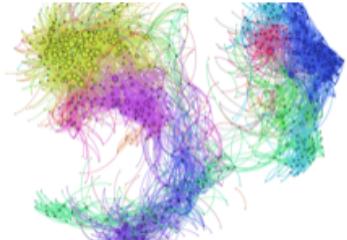
Sensors



Medical
Devices



Mobile
Devices



Social
Networks

UR	Foreign	Rating
K	Independent	5 stars
K	Comedy	5 stars
K	Comedy	5 stars
K	Action & Adventure	5 stars
K	Drama	5 stars
K	Comedy	5 stars
K	PG-13	5 stars
K	Thriller	5 stars
K	Classics	5 stars

Item	Description	Weight	Price
Cabot Vermont Cheddar	16 oz	0.50 lb	\$7.99/lb
Diary	Farmhouse Lactose Free Cheddar (16oz)		\$2.99/lb
	Nature's Pride Grade A Double Broken Eggs (2 dozen)		\$2.49/doz
	Kettle Foods Whole Milk Yogurt (32oz)		\$2.49/doz
	Stonyfield Farm Organic Plain Yogurt (32oz)		\$2.79/doz
Fruit	Ajape Pears (Farm Fresh, Med)	1.76 lb	\$2.49/lb
	Cartelapple (Farm Fresh, Med)		\$2.99/lb
Grocery	Fantastic World Foods Organic Whole Wheat Crackers (12oz)		\$1.99/box
	Great Value! Blue Earth Chips (16oz)		\$2.49/box
	Great Value! Sodium Chloride (15.75oz)		\$0.99/box
	Marsell 2-Ply Paper Towels, White (20ct)		\$1.99/box
	Muir Glen Organic Tomato Paste (16oz)		\$0.99/box
	Starkist Solid White Albacore Tuna in Spring Water (4oz)		\$1.89/box

User
Ratings

Purchase
Histories



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

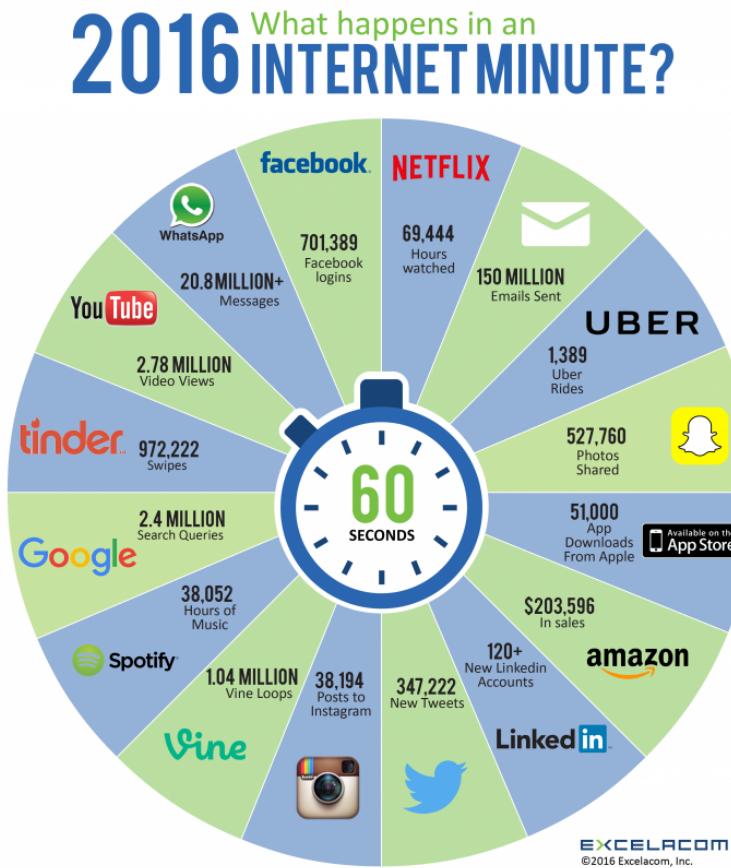
- analysis of social networks
- analysis of financial transactions

Computer Security

- analysis of logged events



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

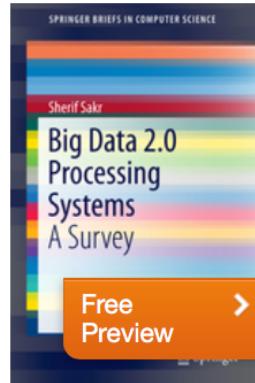
excelacom.com/resources/blog/2016-update-what-happens-in-one-internet-minute



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





Data is the New Oil: Enabling a Data Driven Economy

Smart Web (Every Minute)

- > 400 Hours of [New Videos](#) on YouTube

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

Exponential growth of digital data.

In the midst of a data explosion.



What is dark data?

- Guesses?



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.



The infographic features a large teal circle containing the number "1%". To its right, a statement reads: "Most companies are only analyzing 1% percent of their data." Below this, another statement says: "The data used today are mostly for anomaly detection and control, not optimization and prediction, which provide the greatest value." At the bottom left, there's a screenshot of the DeepDive website with a globe icon and the text "DeepDive". On the right, there's a screenshot of the DARPA MEMEX project, featuring a world map and the text "In DARPA's MEMEX, DeepDive is used by law enforcement agencies to fight human trafficking." A sidebar on the DeepDive site lists questions like "What does DeepDive do?", "What Is DeepDive?", etc. A footer at the bottom right of the slide includes the URL <http://deepdive.stanford.edu>.



THE VERGE

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

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

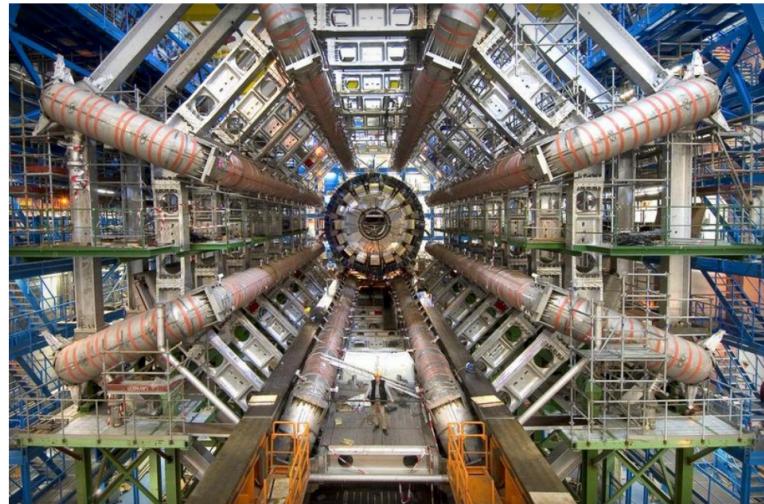
18 NEW ARTICLES

PREVIOUS STORY NEXT STORY
[The US is dropping 'cyberbombs' on ISIS](#) [Logitech turned 160 gaming keyboards into a massive pixelated display](#)

SCIENCE

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

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



11 COMMENTS

Erkennen statt nur Betrachten - darin liegt die Ausdrucks Kraft des Bildes.
JETZT ENTDECKEN
gettyimages®

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	PB	2^{50}
Exabyte	EB	2^{60}
Zettabyte	ZB	2^{70}
Yottabyte	YB	2^{80}



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>



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](#)
[g+](#) [Google+](#) [More](#)

Print [Print this page](#)

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



2. Big Data

Definition, Characteristics, Five Dimensions of Big Data,
Difference Between Statistics and Data Science



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*.



Big Data Characteristics

Typically, how are big data characterized?

What are your thoughts?



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



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



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?



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.

The screenshot shows the homepage of the Emma project at emma-language.org. The page has a blue header with the word 'Emma'. Below the header is a large blue banner with the text 'Emma' and 'Huge data in little words'. There are two buttons: 'latest Download' (green) and 'Documentation' (orange). A descriptive paragraph below the banner states: 'Emma is a Scala API for scalable data analysis. The goal of Emma is to improve developer productivity by hiding parallelism aspects behind a high-level, declarative API.' The background of the page features a photograph of a body of water under a blue sky.

Quickstart

The `emma-quickstart` module contains a maven archetype to quickly setup and run an Emma project.

To get started, download the `emma-quickstart` module (e.g. to `~/downloads`), go to the directory you want to create your project in (e.g. `~/projects/emma`) and execute the `emma-quickstart.sh` script. This generates a maven project from the archetype that you can then import into your favorite IDE. If you want to use a distributed runtime (even if just locally) you should activate one of the available maven profiles (`elink` or `spark`). Otherwise it will still run as a Scala program without any further configurations.

The project includes a basic frame for an out-of-the-box `Emma Job` and a corresponding integration test in the `src/test` package. The `main/Job.scala` provides a framework to try out Emma and implement your first programs. You can test your programs on different runtimes by running the `JobTest.scala` in the `src/test` package.

The screenshot shows the footer of the Emma project website. It includes links for 'Legal', 'emmalanguage/emma' (with icons for GitHub and Twitter), and a brief description: 'Emma is a Scala API for scalable data analysis. The goal of Emma is to improve developer productivity by hiding parallelism aspects behind a high-level, declarative API.'



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**.



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**.



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**.



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.

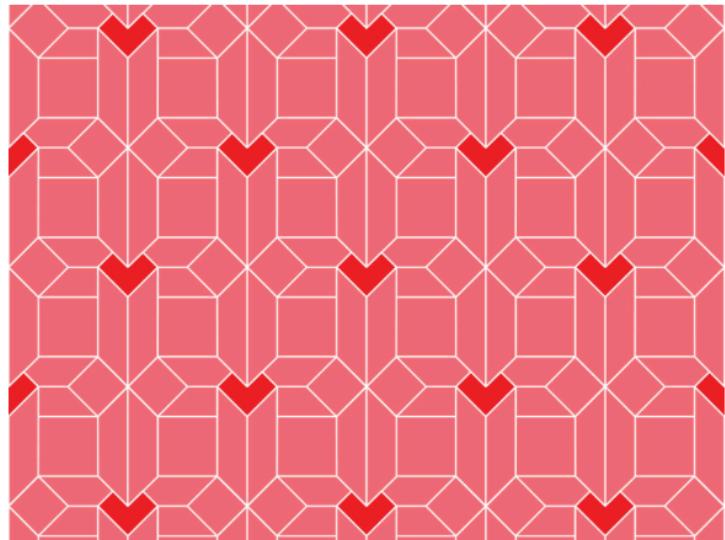


Study Reveals the Perils of Big-Data Science (WiReD)

- **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

MICHAEL ZIMMER SECURITY 05.14.16 7:00 AM

OKCUPID STUDY REVEALS THE PERILS OF BIG-DATA SCIENCE



© GETTY IMAGES

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



Statistics v. Data Science

What are the key differences?



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/#oesq>
(Presentation, 21.10.2015, Slide 36)



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/#oesq>
(Presentation, 21.10.2015, Slide 36)



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.





3. Data Mining

Viewpoints, Algorithms, Characteristics



What is Data Mining?





Data Mining Encompasses ...

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



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;



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.



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.



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.



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 essence of comparison is juxtaposition (side-by-side).”



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](#)

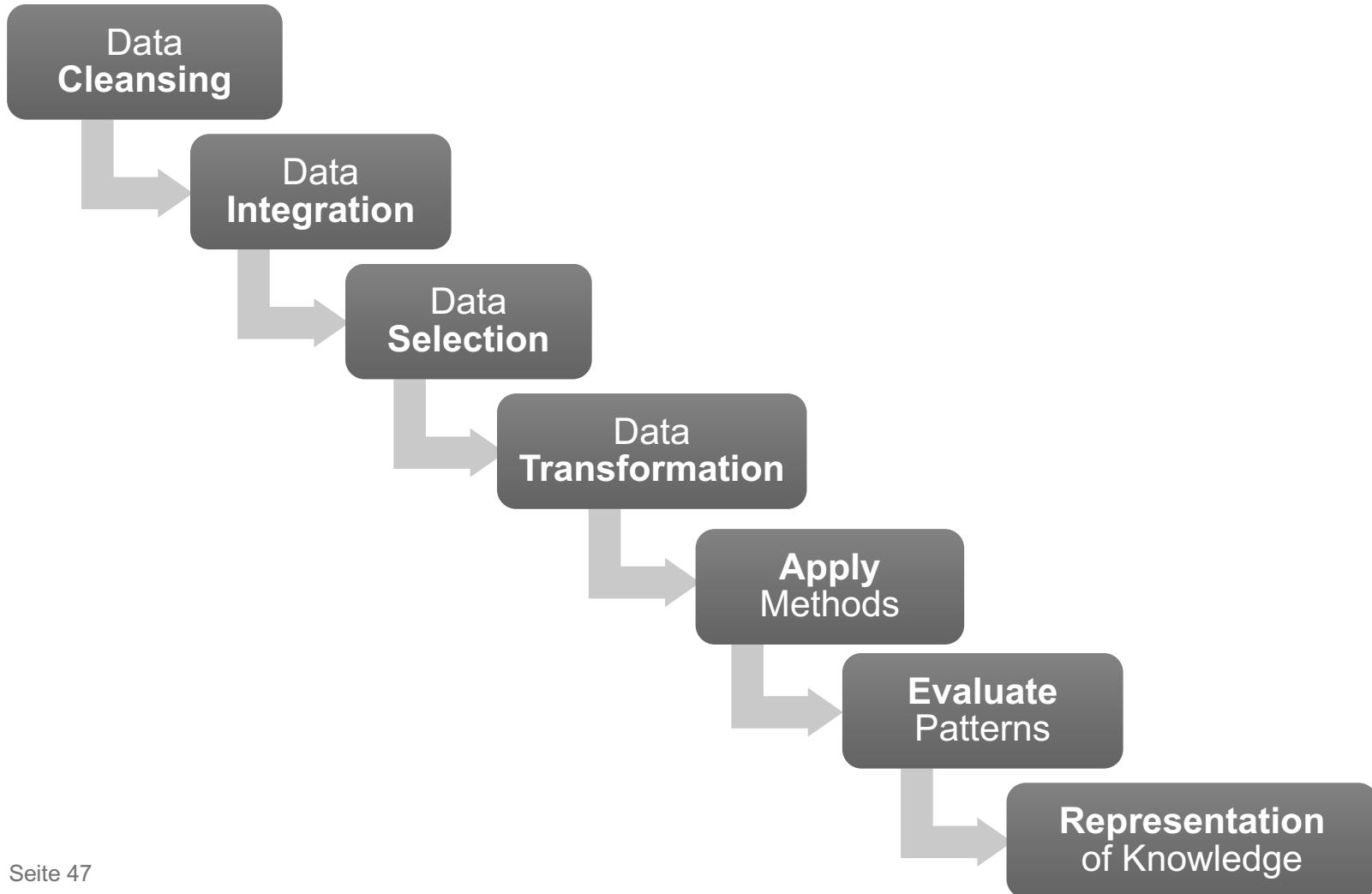


The Number of Data Mining Steps

- How many are there?



The Seven Data Mining Steps





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
- ...



Step 2 Data Integration

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



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



Step 4 Data Transformation

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



Step 4 Data Transformation

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



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



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



What is a data analytic?

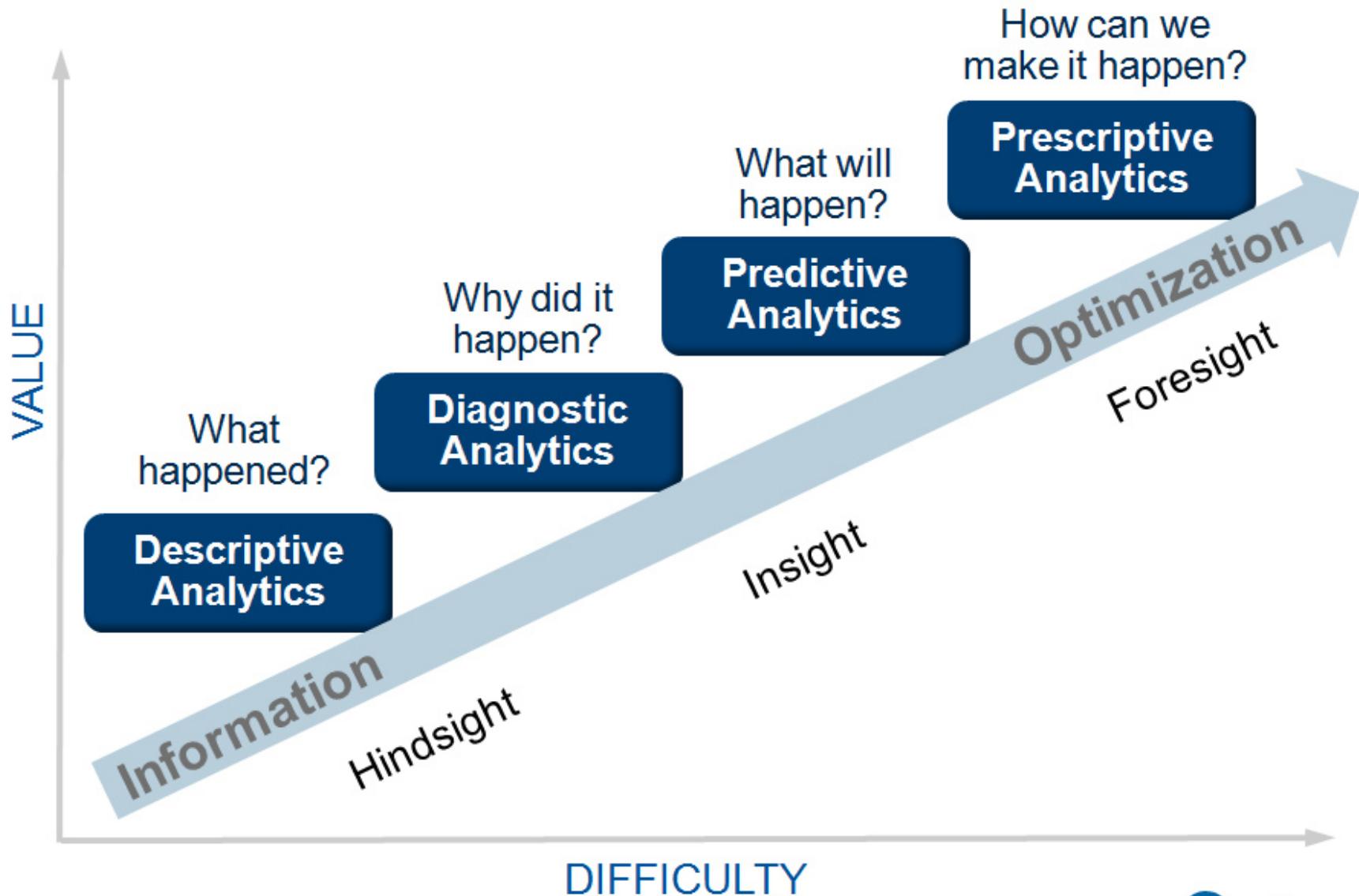
- Comments, thoughts, perspectives?



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



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.



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



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



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



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

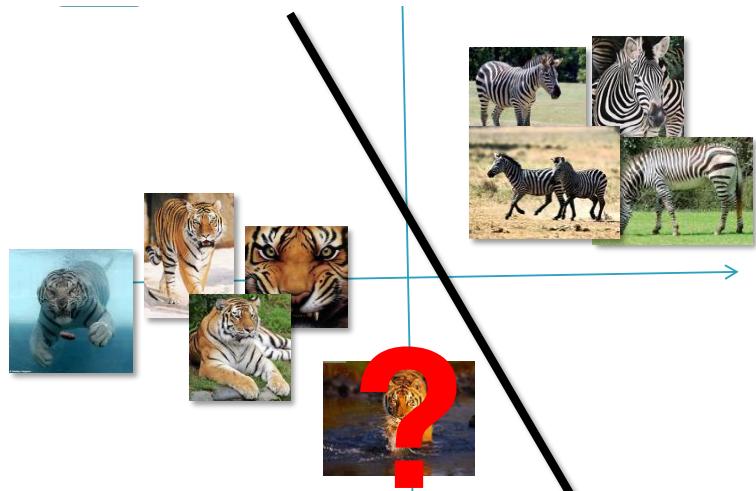
‘tiger’



‘zebra’



Training Data





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

★ 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, Michel 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: ang show details 9:52 PM (1 hour ago) [Reply](#) | ▾

==== Natural WeightLOSS Solution ===

Vital Acai is a natural WeightLOSS 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 WeightLOSS
- * 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

prediction



Spam
vs.
Not Spam

Due to Prof. David Sontag, NYU

Face recognition

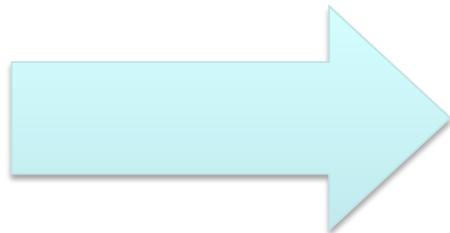


Example training images
for each orientation



Due to Prof. David Sontag, NYU

Weather prediction

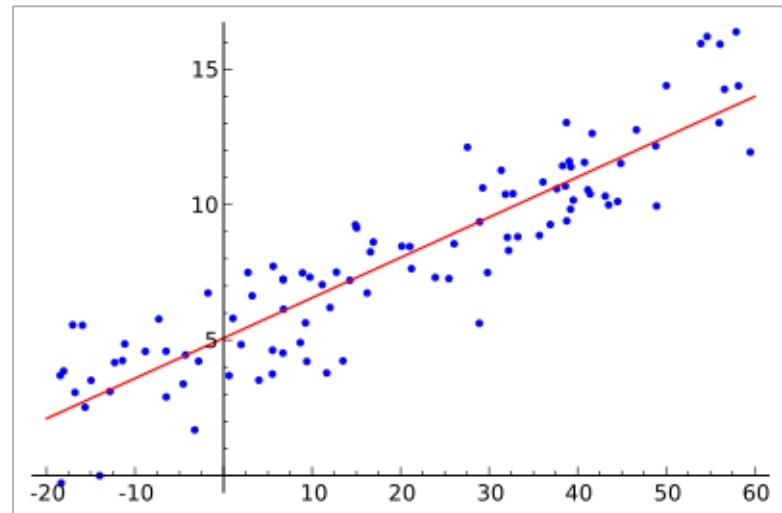


Due to Prof. David Sontag, NYU



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:
weight_i = c₀ + c₁ height_i



Regression Based Algorithms

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

Regression

predicting a numeric value

Stock market



Due to Prof. David Sontag, NYU

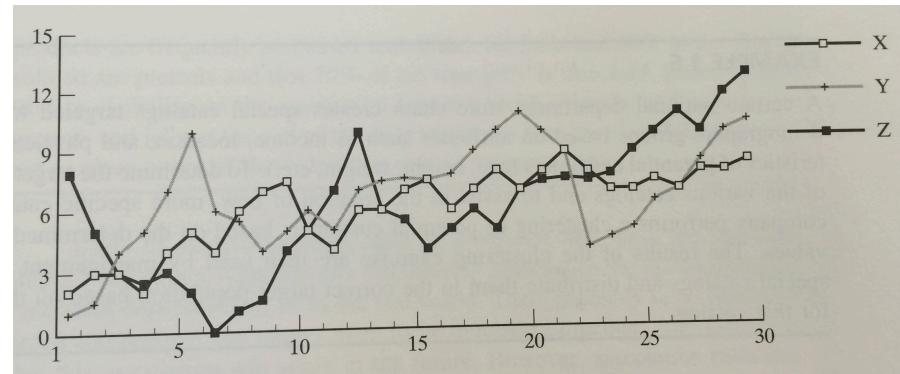
Weather prediction revisited





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



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



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)



Clustering Algorithms

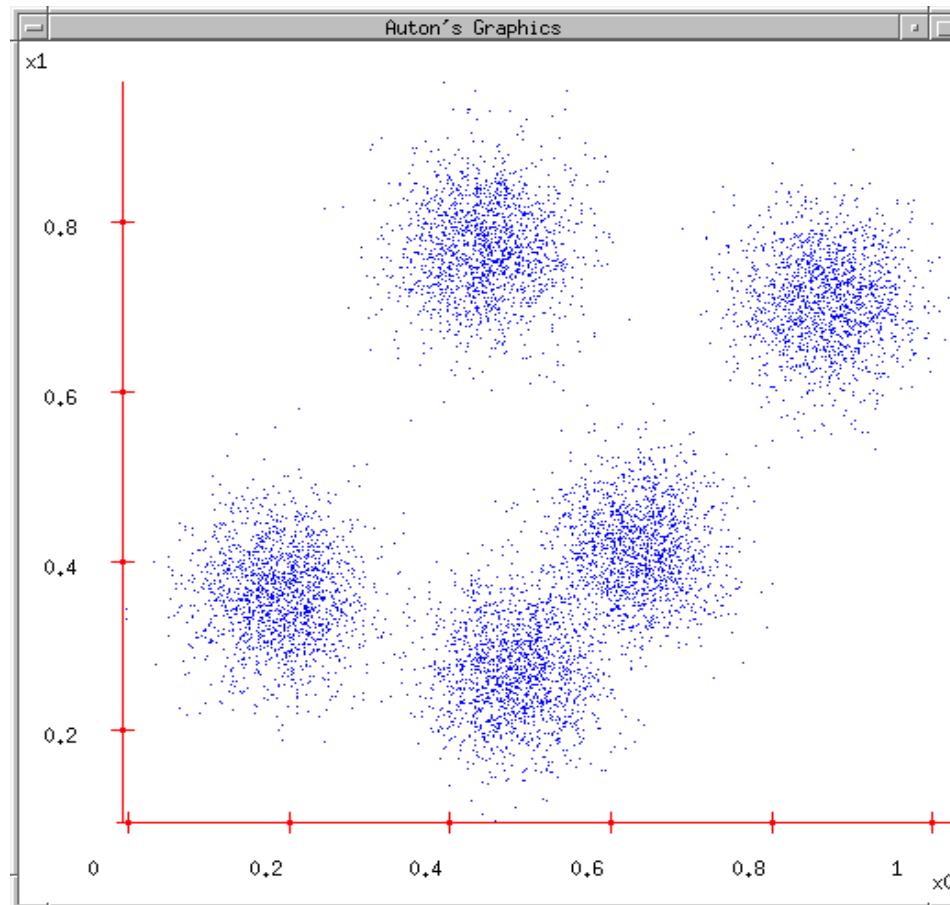
1. K-means
2. BFR (Bradley-Fayyad-Reina, K-means Variant)
3. CURE (Clustering Using Representatives)
4. Gaussian Mixture Model (GMM)
5. Hierarchical Clustering
6. Spectral Clustering

Clustering

discovering structure in data

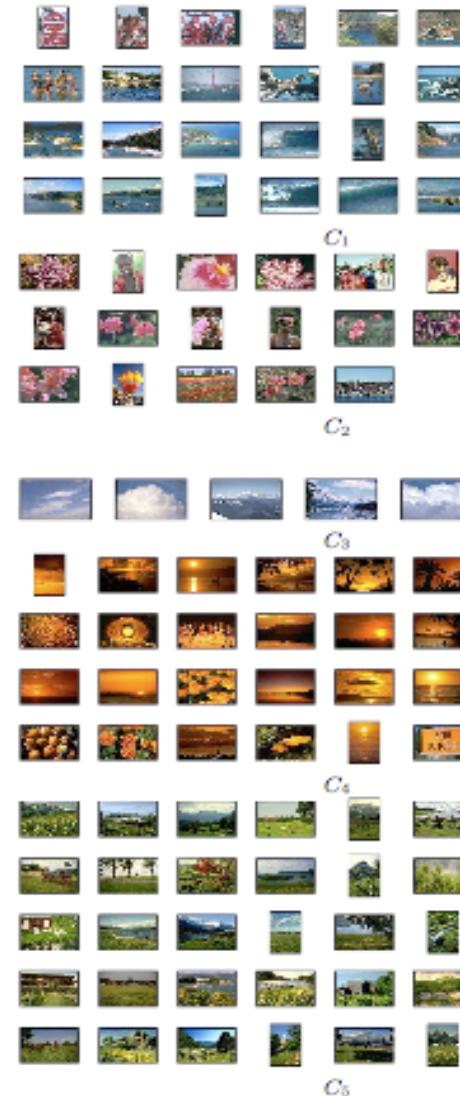
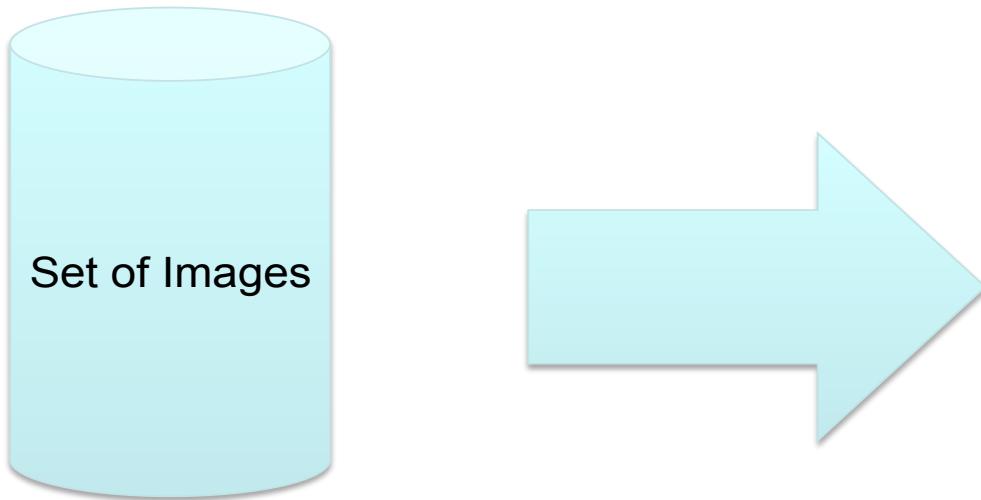
Due to Prof. David Sontag, NYU

Clustering Data: Group similar things



Due to Prof. David Sontag, NYU

Clustering images



[Goldberger et al.]

Clustering web search results

The screenshot shows the Clusty search interface. The top navigation bar includes links for web, news, images, wikipedia, blogs, jobs, and more. A search bar contains the query 'race'. Below the search bar are links for advanced preferences and search results. The main content area displays a cluster titled 'Cluster Human' containing 8 documents. The documents listed are:

- Race (classification of human beings) - Wikipedia, the free ...**
The term **race** or racial group usually refers to the concept of dividing **humans** into populations or groups on the basis of various sets of characteristics. The most widely used **human racial** categories are based on visible traits (especially skin color, cranial or facial features and hair texture), and self-identification. Conceptions of **race**, as well as specific ways of grouping **races**, vary by culture and over time, and are often controversial for scientific as well as social and political reasons. History · Modern debates · Political and ...
[en.wikipedia.org/wiki/Race_\(classification_of_human_beings\)](http://en.wikipedia.org/wiki/Race_(classification_of_human_beings)) - [cache] - Live, Ask
- Race - Wikipedia, the free encyclopedia**
General. **Racing** competitions The **Race** (yachting **race**), or La course du millénaire, a no-rules round-the-world sailing event; **Race** (biology), classification of flora and fauna; **Race** (classification of **human beings**) **Race** and ethnicity in the United States Census, official definitions of "race" used by the US Census Bureau; **Race** and genetics, notion of racial classifications based on genetics. Historical definitions of **race**; **Race** (bearing), the inner and outer rings of a rolling-element bearing. **RACE** in molecular biology "Rapid ... General · Surnames · Television · Music · Literature · Video games
en.wikipedia.org/wiki/Race - [cache] - Live, Ask
- Publications | Human Rights Watch**
The use of torture, unlawful rendition, secret prisons, unfair trials, ... Risks to Migrants, Refugees, and Asylum Seekers in Egypt and Israel ... In the run-up to the Beijing Olympics in August 2008, ...
www.hrw.org/backgrounder/usa/race - [cache] - Ask
- Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich ...**
Amazon.com: **Race**: The Reality Of Human Differences: Vincent Sarich, Frank Miele: Books ... From Publishers Weekly Sarich, a Berkeley emeritus anthropologist, and Miele, an editor ...
www.amazon.com/Race-Reality-Differences-Vincent-Sarich/dp/0813340861 - [cache] - Live
- AAPA Statement on Biological Aspects of Race**
AAPA Statement on Biological Aspects of **Race** ... Published in the American Journal of Physical Anthropology, vol. 101, pp 569-570, 1996 ... PREAMBLE As scientists who study **human** evolution and variation, ...
www.physanth.org/positions/race.html - [cache] - Ask
- race: Definition from Answers.com**
race n. A local geographic or global **human** population distinguished as a more or less distinct group by genetically transmitted physical
www.answers.com/topic/race-1 - [cache] - Live
- Dopefish.com**
Site for newbies as well as experienced Dopefish followers, chronicling the birth of the Dopefish, its numerous appearances in several computer games, and its eventual take-over of the **human race**. Maintained by Mr. Dopefish himself, Joe Siegler of Apogee Software.
www.dopefish.com - [cache] - Open Directory

On the left sidebar, there is a tree view of clusters under 'All Results (238)'. Clusters include: Car (28), Race cars (7), Photos, Races Scheduled (5), Game (4), Track (3), Nascar (2), Equipment And Safety (2), Other Topics (7), Photos (22), Game (14), Definition (13), Team (18), Human (8), Classification Of Human (2), Statement, Evolved (2), Other Topics (4), Weekend (8), Ethnicity And Race (7), Race for the Cure (8), Race Information (8), and more | all clusters. At the bottom of the sidebar is a 'find in clusters:' input field with a 'Find' button.

Due to Prof. David Sontag, NYU