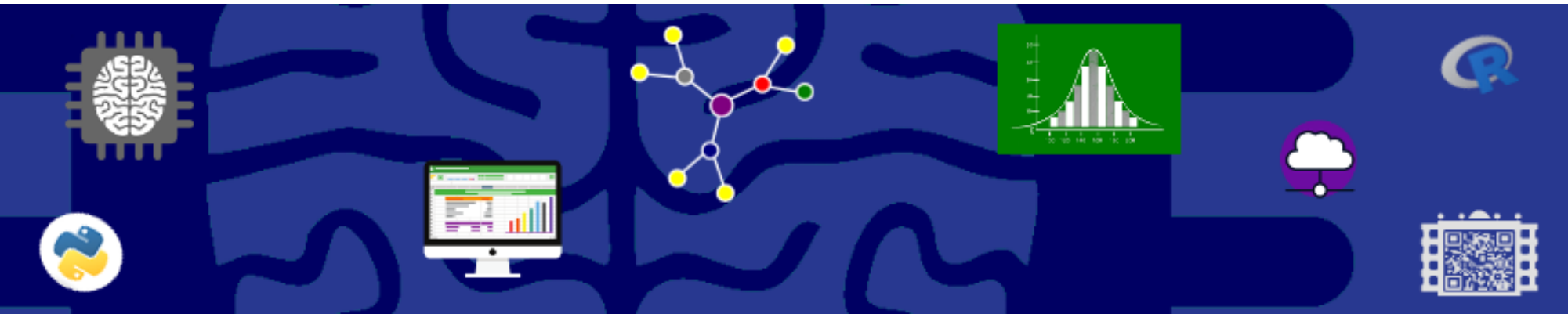
Introduction to Data Science

Winter Semester 2019/20

Oliver Ernst

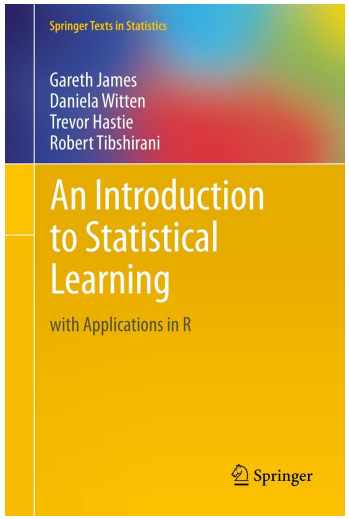
TU Chemnitz, Fakultät für Mathematik, Professur Numerische Mathematik

Lecture Slides



Organizational Issues

- Module: Introduction to Data Science (M24, Einführung in Data Science)
- Class web page: www.tu-chemnitz.de/mathematik/numa/lehre/ds-2018/
- Lecturer: Oliver Ernst (oernst@math.tu-chemnitz.de)
- Class meets Mon 9:15 and Thu 13:45 in Rh70 Room B202.
- Course Assistant: Jan Blechschmidt (jan.blechschmidt@math.tu-chemnitz.de)
- Lab exercises Mon 13:45 Rh39/41 Room 738.
- 50 % of coursework must be handed in correctly to register for exam.
- Oral exam, 30 minutes, at end of teaching term (February)
- **Textbook:** James, Witten, Hastie &Tibshirani. Introduction to Statistical Learning. Springer, 2013.



Note: the majority of images appearing in these lecture slides are taken from this textbook; permission for their use for teaching is gratefully acknowledged.

Contents I

Contents II

Contents III

Contents

What is Data Science?

Point of departure

- **Explosion of data:** sensor technology (e.g. weather); purchase histories (customer loyalty programs, fraud detection); soon every person on the planet ($\approx$ 8 billion) will have a smartphone and generate GPS traces, trillions of photos each year; DNA sequencing ...

  Nagel (2018):

  - Gigabyte (10^9); your hard disk
  - Terabyte (10^{12}); Facebook - 500 TB/day[1]
  - Petabyte (10^{15}); CERN - Large Hadron Collider: 15 PB/year
  - Zettabyte (10^{21}); mobile network traffic 2016
  - Yottabyte (10^{24}); event analysis
  - Brontobyte (10^{27}); sensor data from the IoT (Internet of Things)

[1] Facebook currently has 2.2 billion users worldwide,
Can Mark Zuckerberg Fix Facebook Before It Breaks Democracy? New Yorker, 09/2018

What is Data Science?

Point of departure

- **Explosion of data:** sensor technology (e.g. weather); purchase histories (customer loyalty programs, fraud detection); soon every person on the planet ($\approx$ 8 billion) will have a smartphone and generate GPS traces, trillions of photos each year; DNA sequencing ...

  Nagel (2018):

  - Gigabyte (10^9); your hard disk
  - Terabyte (10^{12}); Facebook - 500 TB/day[1]
  - Petabyte (10^{15}); CERN - Large Hadron Collider: 15 PB/year
  - Zettabyte (10^{21}); mobile network traffic 2016
  - Yottabyte (10^{24}); event analysis
  - Brontobyte (10^{27}); sensor data from the IoT (Internet of Things)
- **Heterogeneity:** besides big, data unstructured, noisy, heterogeneous, not collected systematically.
- **Enabling technologies:** storage capacity; computing hardware; algorithms; 350 years of statistics; 100 years of numerical analysis.

[1] Facebook currently has 2.2 billion users worldwide,
Can Mark Zuckerberg Fix Facebook Before It Breaks Democracy? New Yorker, 09/2018

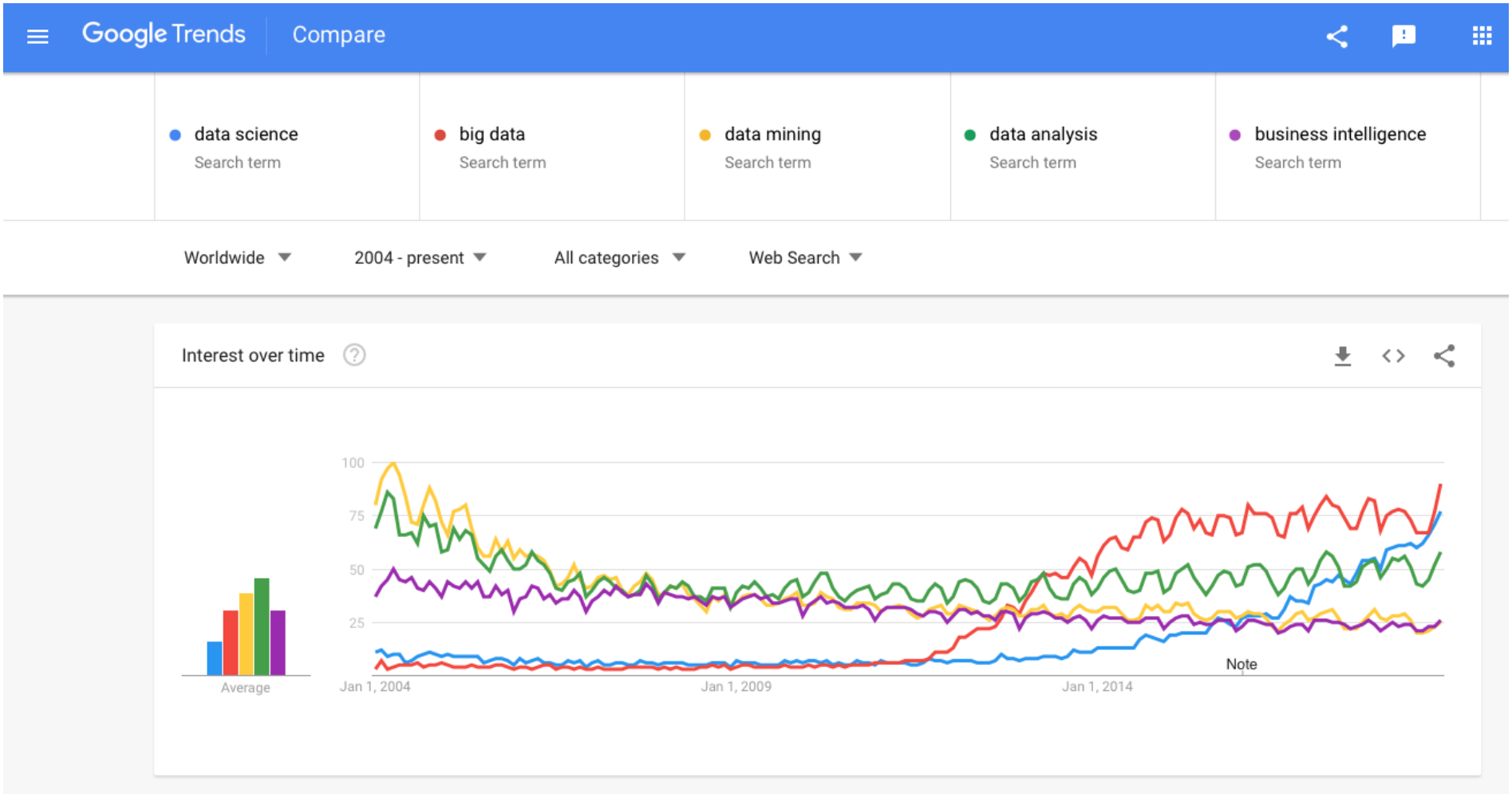
What is Data Science

Buzz words

- Data Science
- Big Data
- Data Mining
- Data Analysis
- Business Intelligence
- Analytics

What is Data Science

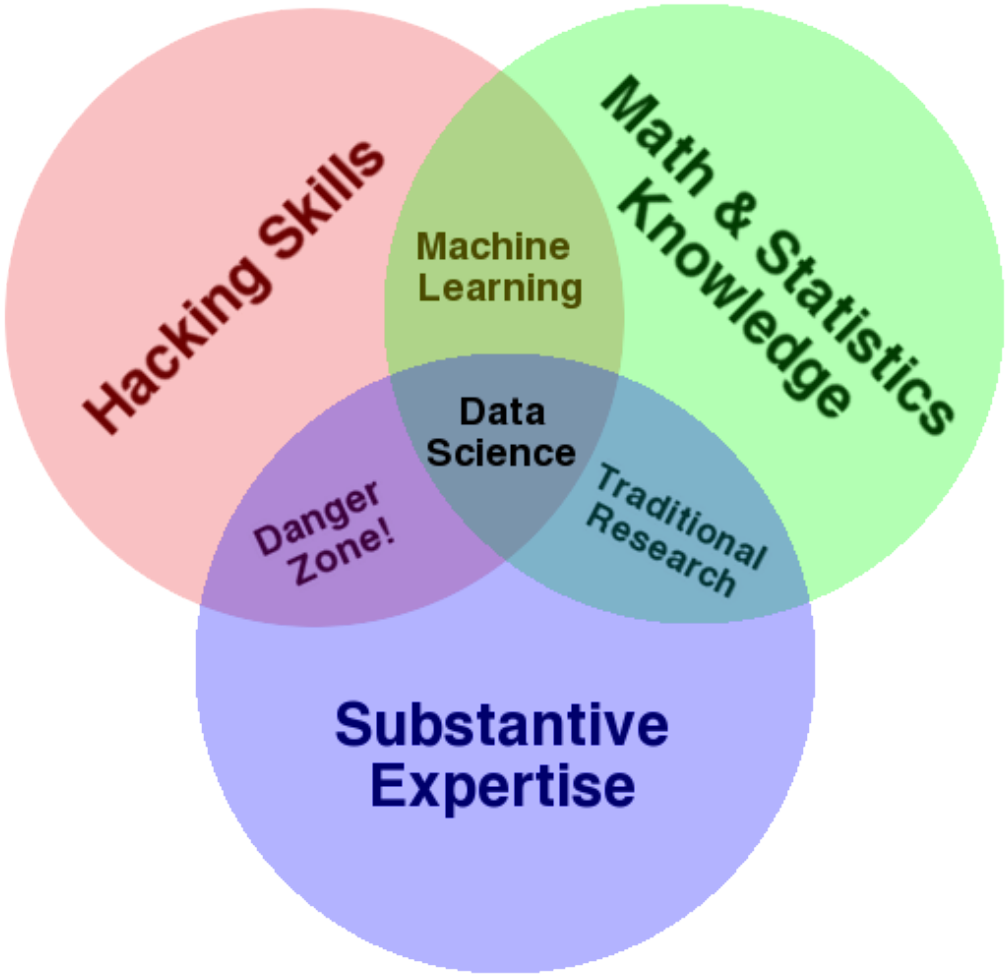
Buzz words



trends.google.com

What is Data Science

Conway's data science Venn diagram (2010) ...

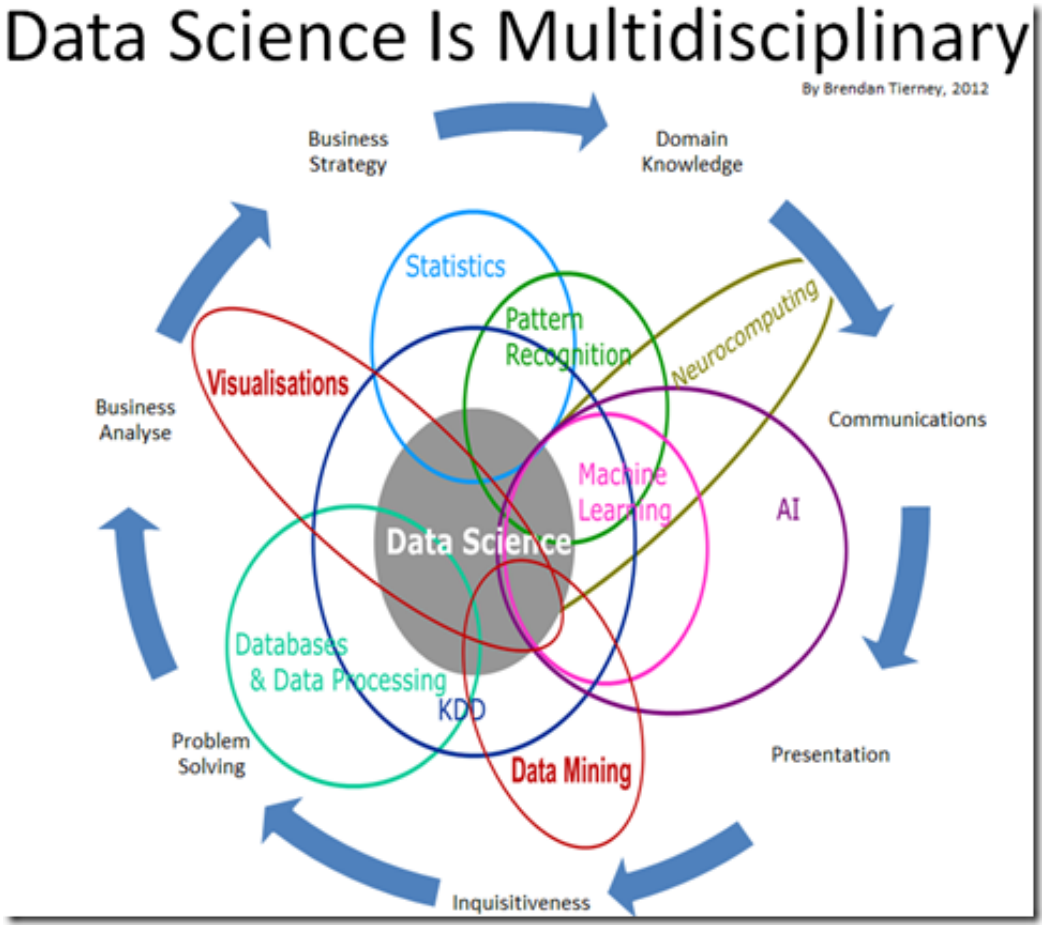


http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

What is Data Science

Conway's data science Venn diagram (2010) . . . and variations

Brendan Tierney (2012)

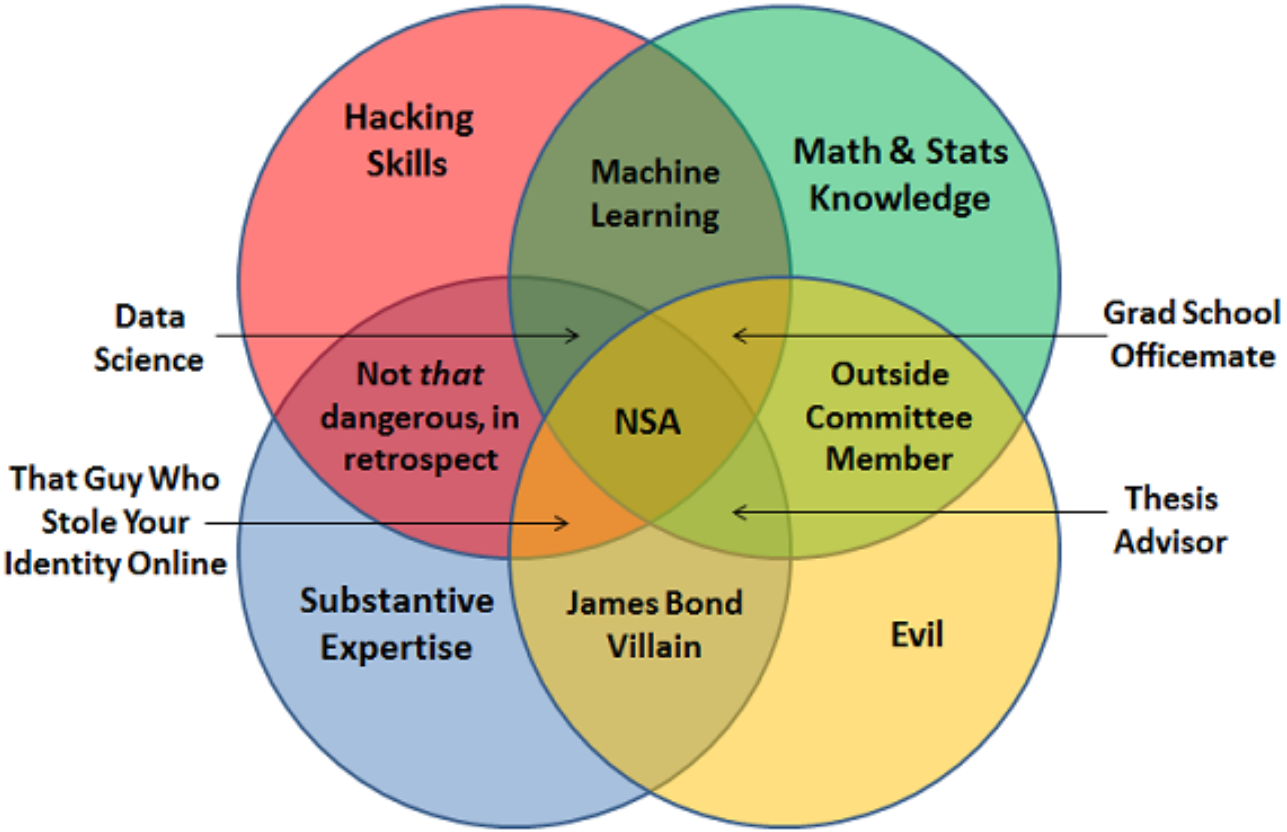


http://www.prooffreader.com/2016/09/battle-of-data-science-venn-diagrams.html

What is Data Science

Conway's data science Venn diagram (2010) . . . and variations

Joel Grus (2013), at the time of the Snowden revelations

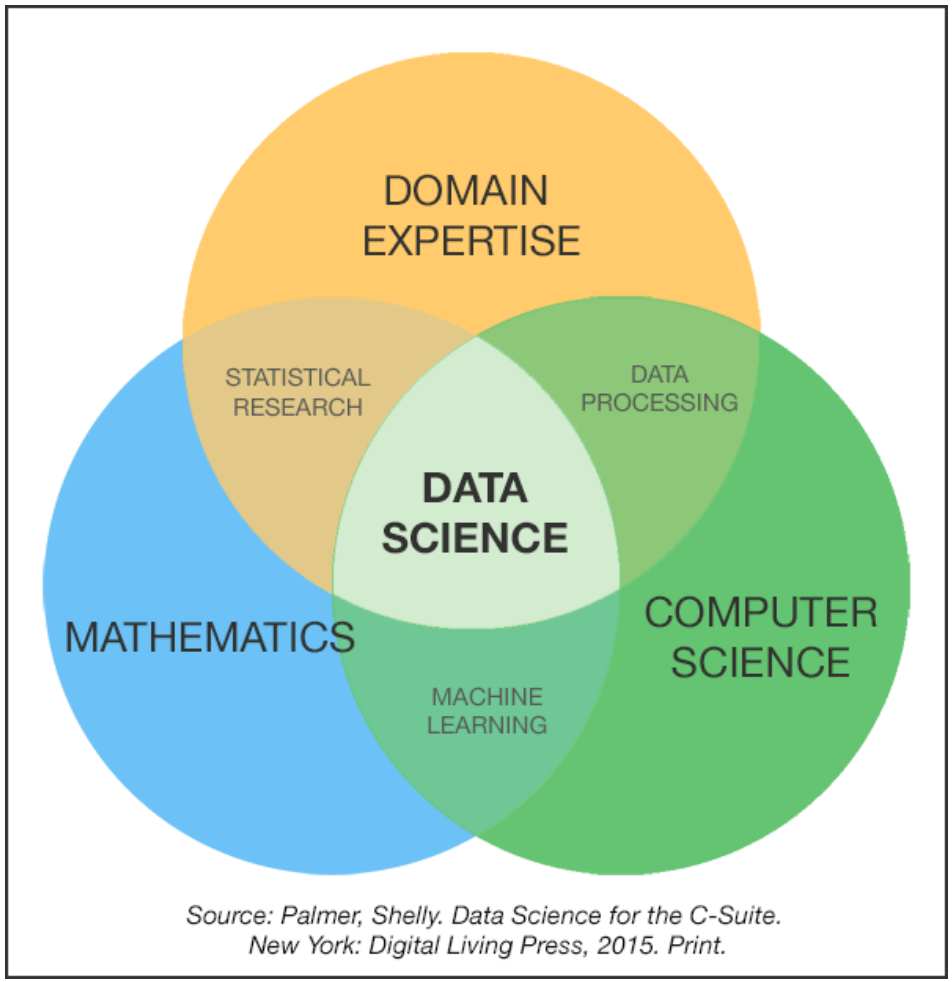


http://www.prooffreader.com/2016/09/battle-of-data-science-venn-diagrams.html

What is Data Science

Conway's data science Venn diagram (2010) . . . and variations

Shelly Palmer (2015)



http://www.prooffreader.com/2016/09/battle-of-data-science-venn-diagrams.html

What is Data Science

. . . and from the twitterverse:

"A data scientist is a statistician who lives in San Francisco."

What is Data Science

. . . and from the twitterverse:

"A data scientist is a statistician who lives in San Francisco."

"Data Science is statistics on a Mac"

What is Data Science

. . . and from the twitterverse:

"A data scientist is a statistician who lives in San Francisco."

"Data Science is statistics on a Mac"

"A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician." *Josh Wills, Director of Data Science at Cloudera*

What is Data Science

. . . and from the twitterverse:

"A data scientist is a statistician who lives in San Francisco."

"Data Science is statistics on a Mac"

"A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician."

Josh Wills, Director of Data Science at Cloudera

"A data scientist is someone who is worse at statistics than any statistician and worse at software engineering than any software engineer."

Will Cukierski, Data Scientist at Kaggle

https://twitter.com/cdixon/status/428914681911070720

What is Data Science

Donoho's first definition

David Donoho (2015)[2]

Classifies the activities of *Greater Data Science* (GDS) into six divisions.

(GDS1) Data Gathering, Preparation and Exploration
experimental design; collect; reformat, treat anomalies, expose unexpected features.

(GDS2) Data Representation and Transformation
databases; represent e.g. acoustic, image, sensor, network data.

(GDS3) Computing with Data
SW engineering; cluster computing; developing workflows.

(GDS4) Data Visualization and Presentation
classical EDA plots; high-dimensional or streaming data

(GDS5) Data Modeling
generative vs. predictive modeling

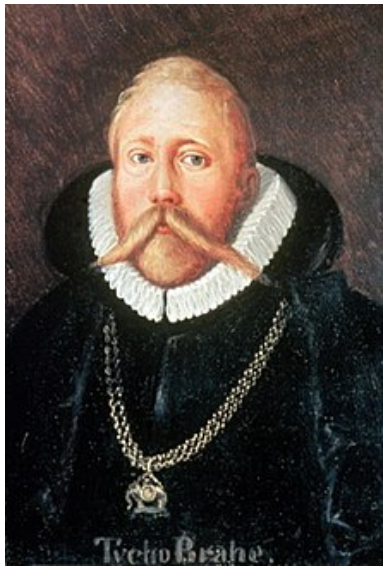
(GDS6) Science about Data Science
evaluate results, processes, workflows; analysis of methods.

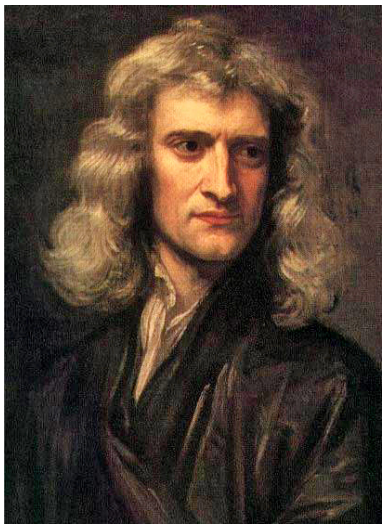
[2] 50 Years of Data Science. Journal of Computational and Graphical Statistics. 26 (2017)

Some Examples

Celestial motion

Brahe, Kepler & Newton:





- Tycho Brahe (1546–1501): Accurate and detailed observations of positions of celestial bodies over many years. (data collection)
- Johannes Kepler (1571–1630): Based on Brahe's observations, derives laws of orbital motion. (model fitting)
- Isaac Newton (1643–1727): Derives Newtonian mechanics, from which Kepler's laws follow. (deeper underlying truths)

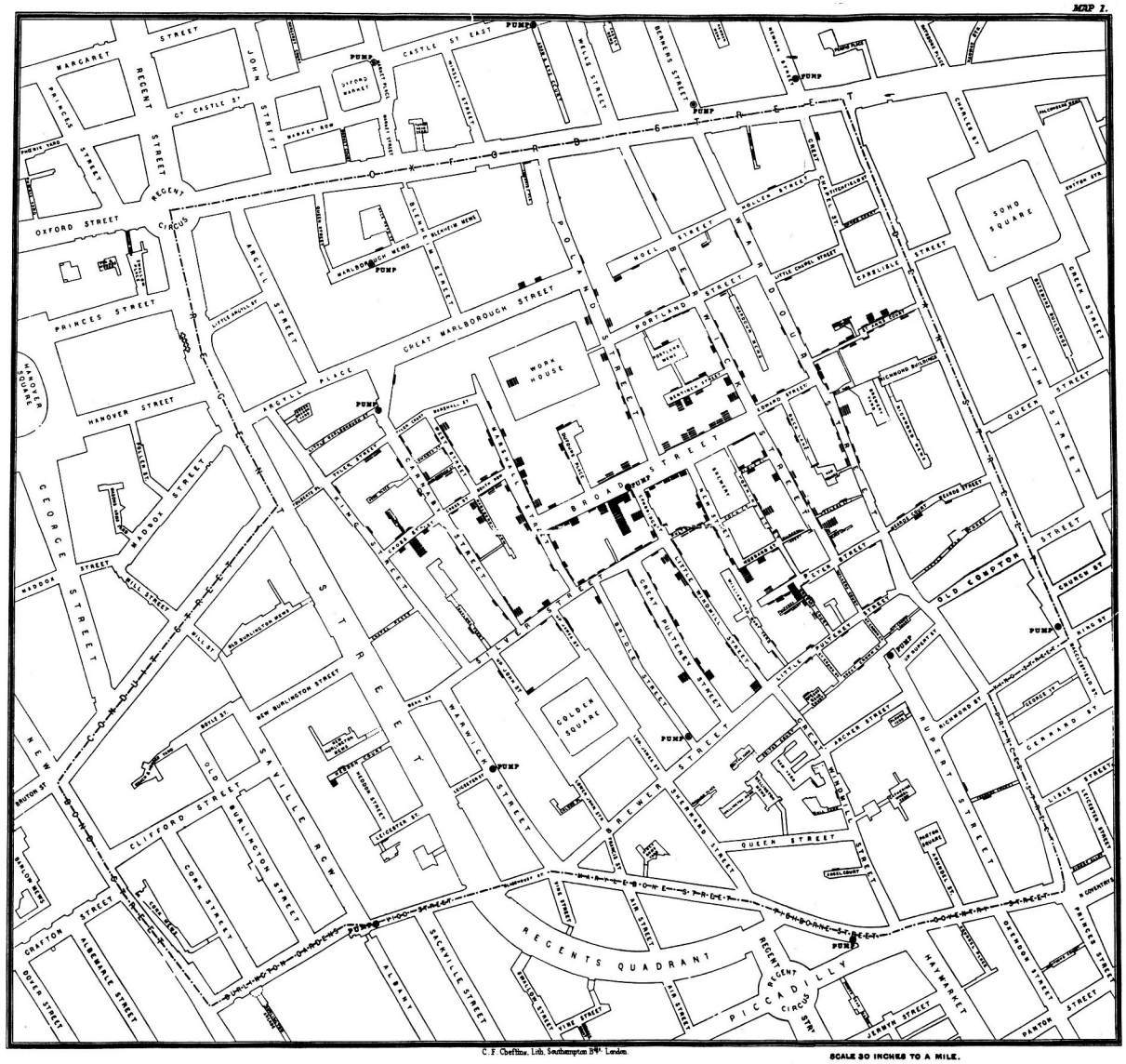
Some Examples

1854 cholera outbeak, Broad Street, London

- Mid 19-th century London was plagued by devastating outbreaks of Cholera.
- Competing medical theories of the time: miasma ("bad air") vs. as yet unidentified microbes.
- London doctor John Snow, microbe theory proponent, conjectured that Cholera-inducing microbes spread via drinking water.
- During a particularly severe outbreak in Soho district in 1854, Snow recorded the deaths on a *dot map*.
- The clustering pattern evident from the map pointed to the *Broad Street pump* as the likely source of contamination.
- Based on this evidence, he was able to convince the authorities to disable the pump, and Cholera incidence in the neighborhood declined.

Some Examples

1854 cholera outbeak, Broad Street, London



Original Snow map of 1854. Stacked rectangles indicate Cholera cases. (Wikipedia)

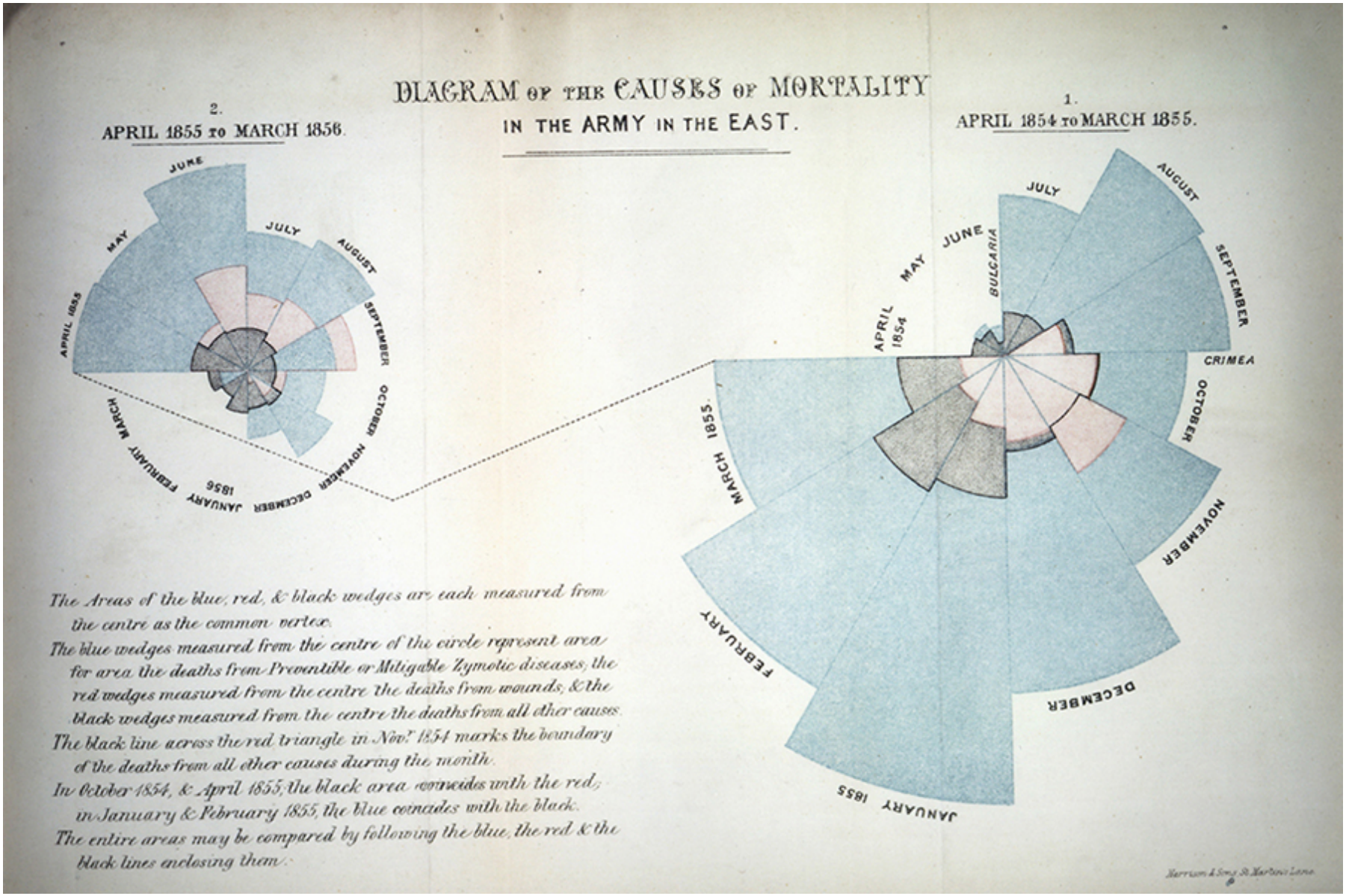
Some Examples

Florence Nightingale's data visualizations

- During the *Crimean War* (1853–1856), Florence Nightingale, a well-educated British social reformer and statistician, was in charge of a group of nurses tending the wounded British soldiers at Selimiye Barracks near Istanbul.
- She systematically documented the horrific medical conditions she encountered on arrival and reported her findings to the authorities, effecting dramatic hygenic improvements.
- Nightingale pioneered the graphical display of statistical data in pie-chart-like diagrams known then as *"coxcombs"*.
- In 1859, Nightingale was elected the first female member of the Royal Statistical Society. An honorary membership in the American Statistical Association followed in 1874.

Some Examples

Florence Nightingale's data visualizations



Some Examples

Data science today

- **Precision medicine:** the cost of sequencing the human genome has dropped by a factor of one million (3.5 billion to 1000 USD) in the last decade. New algorithms achieve a full genome analysis in 24 hours, looking at 6 billion genes, vs. standard technology taking 3 months, accounting for 5 genes only.

  E. A. Ashley. Towards Precision Medicine. Nature Reviews Generics (2017).
- **Predicting elections:** in 2008, the American statistician Nate Silver correctly predicted the outcome in 49 of 50 states in the US presidential election. In 2012 he correctly predicted in all 50 states.

  N. Silver. The Signal and the Noise. Penguin (2012).
- **Influencing elections:** A classical model for personality traits was combined with data scraped from the accounts of a large number of Facebook users to produce microtargeted ads in recent political campains, including the 2016 US presidential election and the UK referendum on leaving the EU. Besides litigation, this has led to concerns over privacy issues and the possibilities for systematic disinformation based on data science techniques.

  The Data that Turned the World Upside Down. Motherboard Vice (2017).

Some Examples

Data science today

- **Beating humans at games**. 1997 IBM's Deep Blue beat world chess grandmaster Garry Kasparov. 2011 IBM's Deep Blue won at the TV game show 'Jeopardy'. In 2016 Google's Deep Mind beat world champion Lee Sedol at the ancient chinese game 'Go'.

  B. Dickson.. All the important games artificial intelligence has conquered. (2018).
- **Autonomous vehicles:** Automobiles already contain multiple software systems for control, navigation, communication and entertainment. By combining multiple sensor input (cameras, LIDAR, RADAR), artificial intelligence systems are moving ever closer to cars which can drive autonomously. Daimler has promised driverless cars by 2020, Ford by 2021.
- **Medical diagnostics:** In 2017, the Stanford Machine Learning Group introduced the algorithm CheXNet, a neural network which outperformed practicing radiologists in detecting pneumonia from chest X-ray images. In 2018, the same group presented a machine learning system which can summarize the key findings of radiology reports for x-ray images nearly as well as radiologists can.
- **Natural language processing:** Google, Google translate, Google ngrams,

What is Data Science?

Donoho's second definition/paradigm

David Donoho. Data Science: The End of Theory?
Lecture at Data Science Kickoff event, U Vienna, 22.06.2018

> **Def.** *Data Science concerns the recognition, formalization and exploitation of data phenomenology emerging from digital transformation of business, society and science itself.*

Hallmarks of DS activities:

Developments	Examples
Accumulating digital assets Emergent data/methodology universe Emergent technique of exploitation	

https://www.youtube.com/watch?v=PekBM76z2qE

What is Data Science?

Donoho's second definition/paradigm

In his Vienna lecture Donoho provides a series of examples for activities which qualify as DS under this paradigm, among these

- a study of false discovery rates in medical literature [Jager & Leek, 2014],
- a cross-model study on detection methods for ovarian cancer,
- the enterprise of machine learning, as well as
- the use of deep learning to accelerate existing algorithms.

What is Data Science?

Donoho's second definition/paradigm

Biostatistics (2014), **15**, 1, *pp.* 1–12
doi:10.1093/biostatistics/kxt007
Advance Access publication on September 25, 2013

R

An estimate of the science-wise false discovery rate and application to the top medical literature

LEAH R. JAGER

Department of Mathematics, United States Naval Academy, Annapolis, MD 21402, USA

JEFFREY T. LEEK*

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

jleek@jhsph.edu

SUMMARY

The accuracy of published medical research is critical for scientists, physicians and patients who rely on these results. However, the fundamental belief in the medical literature was called into serious question by a paper suggesting that most published medical research is false. Here we adapt estimation methods

What is Data Science?

Donoho's second definition/paradigm

Biostatistics (2014), **15**, 1, *pp.* 1–12
doi:10.1093/biostatistics/kxt007
Advance Access publication on September 25, 2013

R

An estimate of the science-wise false discovery rate and application to the top medical literature

LEAH R. JAGER
Department of Mathematics, United States Naval Academy, Annapolis, MD 21402, USA
JEFFREY T. LEEK*
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA
jleek@jhsph.edu

SUMMARY

The accuracy of published medical research is critical for scientists, physicians and patients who rely on these results. However, the fundamental belief in the medical literature was called into serious question by a paper suggesting that most published medical research is false. Here we adapt estimation methods

Developments	Examples
Accumulating digital assets	database, SW for text scraping
Emergent data/methodology universe	all Pub-Med abstracts
Emergent technique of exploitation	stats about scientific activity

What is Data Science?

The end of theory?

- In 2008 an article in Wired Magazine[3] suggested that emerging data science techniques would render the traditional scientific method based on hypotheses obsolete, implying that automated data mining will lead directly to new discoveries.
- This is particularly relevant for areas of research such as bioinformatics, systems biology, epidemiology and ecology.
- This prompted an intense discussion on data-driven vs. hypothesis-driven research.
- The "No-Free-Lunch theorem" states that there can be no learning without knowledge.
- Currently, a "mathematization" of data science is setting in, leaving ample room for theoretical analysis, likely leading to improved methods. (Compare with the development of other computational disciplines such as the finite element method or uncertainty quantification)

[3] The end of theory:the data deluge makes the scientific method obsolete. Wired 6/2008.