Technische Universität Braunschweig

BRICS Braunschweig Integrated Centre of Systems Biology

PLRI PETER L. REICHERTZ INSTITUTE FOR MEDICAL INFORMATICS

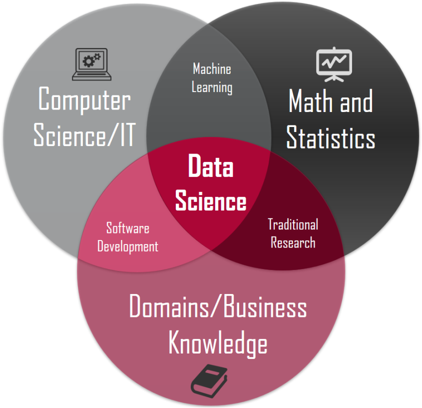
ifis Institut für Informationssysteme Technische Universität Braunschweig

An Introduction to Data Science Projects

Prof. Dr. Tim Kacprowski, Prof. Dr. Wolf-Tilo Balke, April 17, 2023

What is Data Science?

- Data science is the use of statistical and mathematical methods to identify a problem, then find, organize and analyze data to look for patterns that can lead to solutions
- Data science integrates scientific methods from statistics, computer science, and the respective application domain, depending on the problem that is being solved



https://statweb.stanford.edu/~{}tibs/stat101.html

What do data scientists do?

> The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that's going to be a hugely important skill in the next decades.
>
> Hal Varian,
> *chief economist at Google and UC Berkeley professor of information sciences, business, and economics*

What are the main skills required?

- Mathematics and statistics (understanding what types of analysis are possible and the computational techniques involved)
- Data wrangling (cleaning and organizing the data)
- Programming skills (writing code that implements the algorithms that do the work)
- Domain knowledge (knowing the relevant underlying concepts to optimize and adapt the technical work to the unique requirements of the project)
- Communication skills (interact with experts and colleagues)

How to be an effective data scientist?

- identify relevant questions
- collect and organize information from various (appropriate) sources
- translate results into actionable findings
- communicate (not only) findings

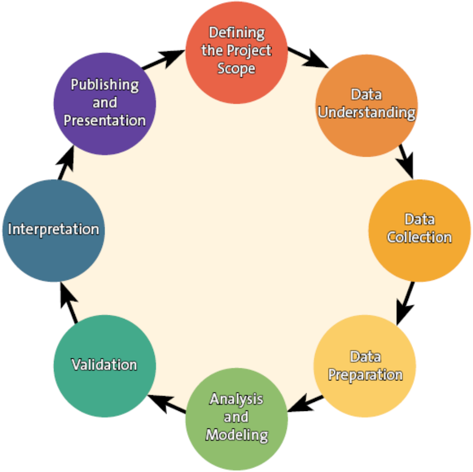
These skills are required in almost all industries, and skilled data scientists are becoming increasingly valuable to companies.

Areas of application

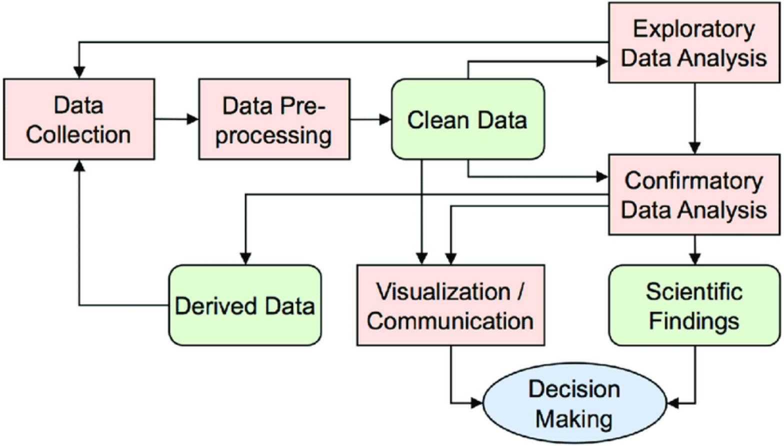
- Engineering
- Medicine
- Biology / Chemistry / Pharmacy
- Image and signal processing
- Finance
- Cybersecurity
- Marketing
- Manufacturing
- Aerospace
- Media
- Government
- Etc.

The (Ideal) Life Cycle of a Data Science Project

- Defining the project scope
- Data understanding
- Data collection
- Data preparation
- Analysis and modeling
- Validation
- Interpretation
- Publishing and presentation

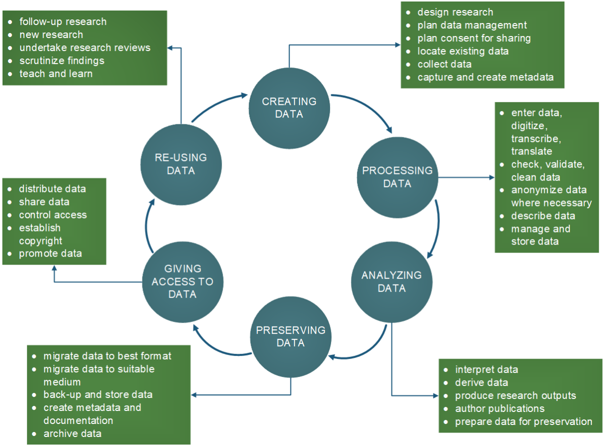


The (More Realistic) Life Cycle of a Data Science Project



Ma, X., et al. (2017). ISPRS International Journal of Geo-Information.

The Life Cycle of Data



https://libguides.mines.edu/c.php?g=949816&p=6850016

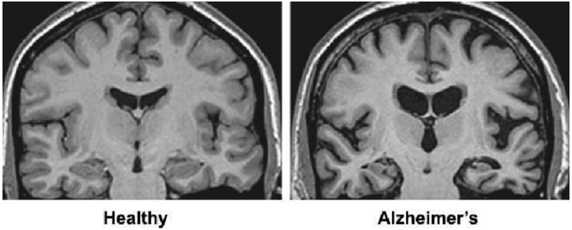

Defining the Project Scope
Data Understanding
Data Collection
Data Preparation
Analysis and Modeling
Validation
Interpretation
Publishing and Presentation
Defining the Project Scope

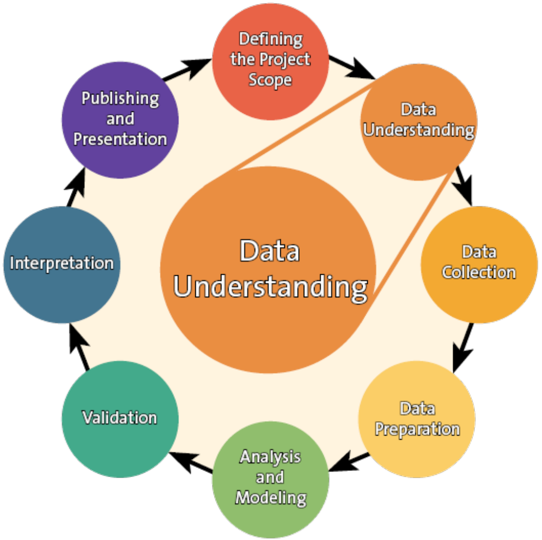
Defining the Project Scope

- What are the main problems that you aim to resolve?
- What is the output or target variable that should be predicted?
- How do you plan to achieve them?

Defining the Project Scope: Example

- Project on Alzheimer's disease (AD)
- Aim: early detection of the disease
- Methods: machine learning algorithm
- Output: probability of development of AD later in life



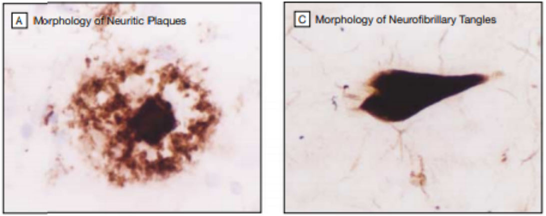

Defining the Project Scope
Data Understanding
Data Collection
Data Preparation
Analysis and Modeling
Validation
Interpretation
Publishing and Presentation
Data Understanding

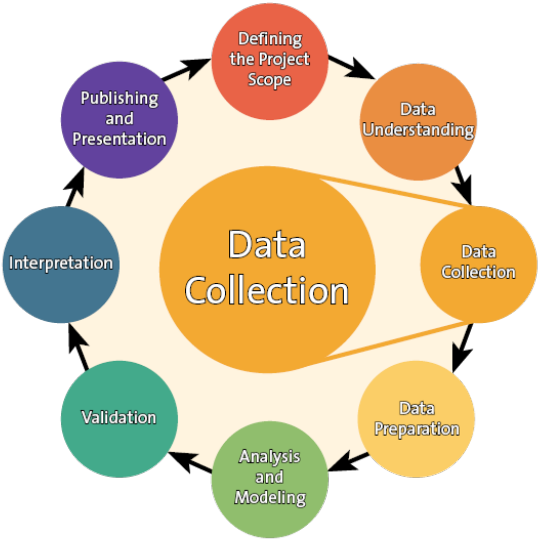
Data Understanding

- Data Science is **multi-disciplinary**
- Get familiar with relevant domain knowledge
- Understand basic concepts
- Communicate with collaborators
  → Where does the data come from? What does it mean? What are possible "problems" with the data?

Data Understanding: Example

- AD is a neurodegenerative disorder
- Plaques (β-amyloid) and tangles (tau) form in the brain
- Data: blood samples from healthy elderly people and elderly people with AD
- Regular exchange with neurologists
- Can the future progression to AD be determined from blood samples?

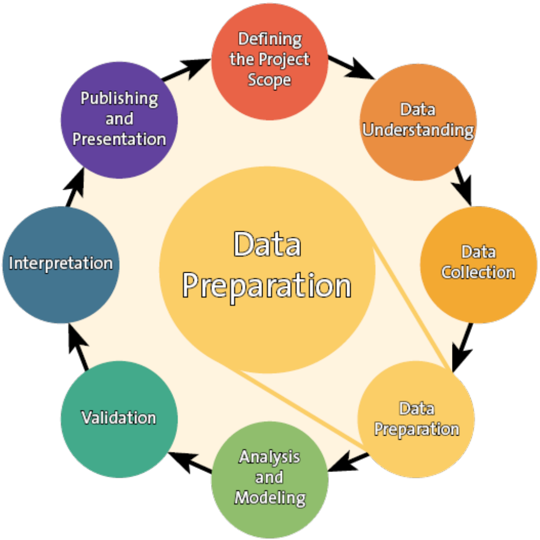



Defining the Project Scope
Data Understanding
Data Collection
Data Preparation
Analysis and Modeling
Validation
Interpretation
Publishing and Presentation
Data Collection

Data Collection

- Gather the data needed for the project
- Examples:
  - Electronic health records (EHRs)
  - Gene expression data
  - SNP information
  - DNA Methylation records
  - **Blood composition**
  - Imaging data (MRI, PET, ...)
  - Questionnaires
  - ...

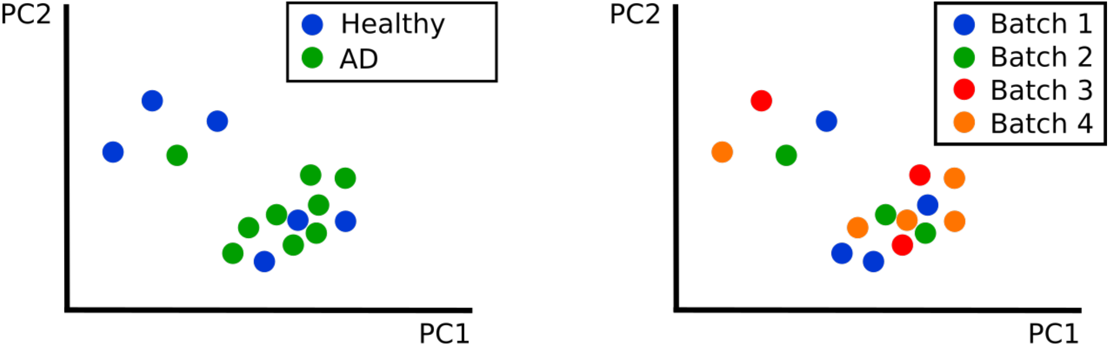
Test Name	Result	Flag	Reference Range	Lab
HEPATIC FUNCTION PANEL				
PROTEIN, TOTAL	6.1		6.1-8.1 g/dL	EN
ALBUMIN	4.3		3.6-5.1 g/dL	EN
GLOBULIN	1.8	LOW	1.9-3.7 g/dL (calc)	EN
ALBUMIN/GLOBULIN RATIO	2.4		1.0-2.5 (calc)	EN
BILIRUBIN, TOTAL	0.6		0.2-1.2 mg/dL	EN
BILIRUBIN, DIRECT	0.2		< OR = 0.2 mg/dL	EN
BILIRUBIN, INDIRECT	0.4		0.2-1.2 mg/dL (calc)	EN
ALKALINE PHOSPHATASE	61		40-115 U/L	EN
AST	27		10-35 U/L	EN
ALT	19		9-46 U/L	EN


Defining the Project Scope
Data Understanding
Data Collection
Data Preparation
Analysis and Modeling
Validation
Interpretation
Publishing and Presentation
Data Preparation

Data Preparation

- Raw data has to be cleaned and preprocessed before being used:
  - Handling missing values
  - Duplicates
  - Imputation
  - Remove low-quality data
  - Remove outliers
  - Formatting
  - Normalization of inconsistent data
  - (Feature engineering)
- Exploratory Data Analysis can help in identifying the structure, outliers, anomalies and patterns present in the data.

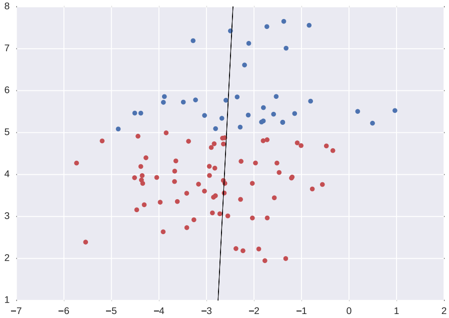
Exploratory Data Analysis: Example



PCA *vs.* PLS

PCA maximizes variance **within the data** (irrespective of classes)!

PLS maximizes varaince **between classes**!



Explorative anlaysis always with PCA!

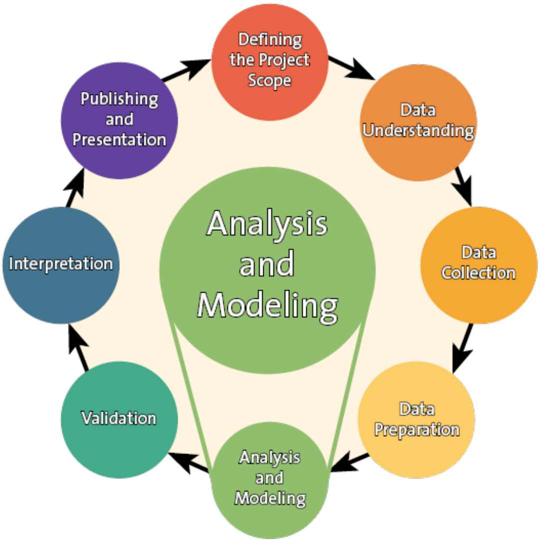
Data Preparation: Example

Data has missing values → (mean) imputation

ID	Molecule A	Molecule B
0	0.05	2.0
1	NaN	3.3
2	0.72	NaN
3	0.25	1.7
4	NaN	4.0
5	0.42	NaN
6	0.39	2.7
7	0.59	2.0
8	0.10	1.3

→

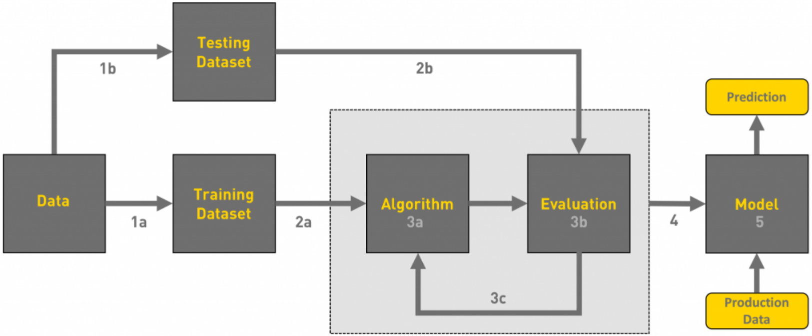
ID	Molecule A	Molecule B
0	0.05	2.0
1	**0.36**	3.3
2	0.72	**2.4**
3	0.25	1.7
4	**0.36**	4.0
5	0.42	**2.4**
6	0.39	2.7
7	0.59	2.0
8	0.10	1.3


Defining the Project Scope
Data Understanding
Data Collection
Data Preparation
Analysis and Modeling
Validation
Interpretation
Publishing and Presentation
Analysis and Modeling

- **Core of a data science project**
- Find patterns in data
- Which machine learning algorithms are suitable for the project?
  - Linear regression
  - Random forests
  - Support vector machines
  - Deep learning
  - Clustering methods
  - Other emerging computational methods

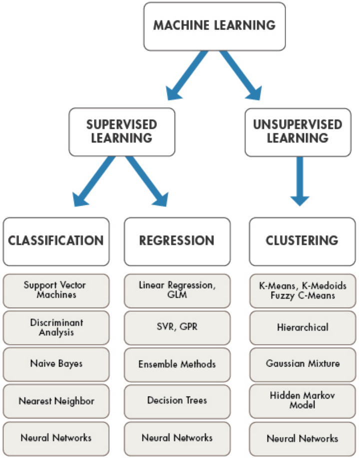
Machine Learning: Example

Build a model capable of predicting the probability of AD developing
Train it with preprocessed data

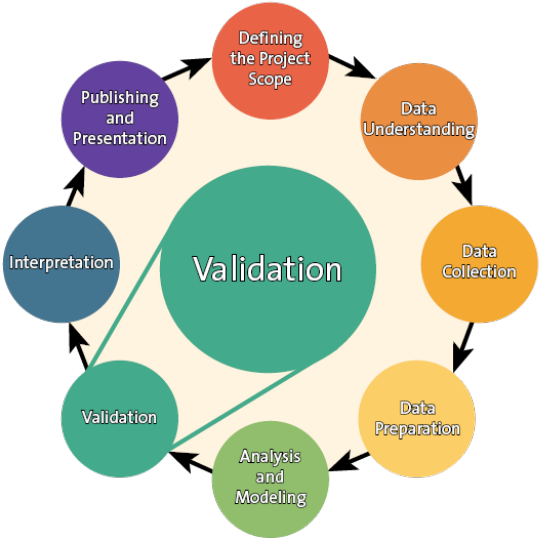


https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94

Types of Machine Learning Approaches



https://de.mathworks.com/help/stats/machine-learning-in-matlab.html

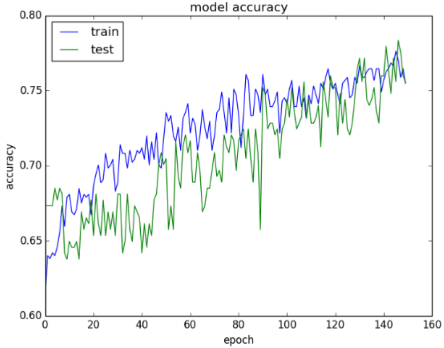

Defining the Project Scope
Data Understanding
Data Collection
Data Preparation
Analysis and Modeling
Validation
Interpretation
Publishing and Presentation
Validation

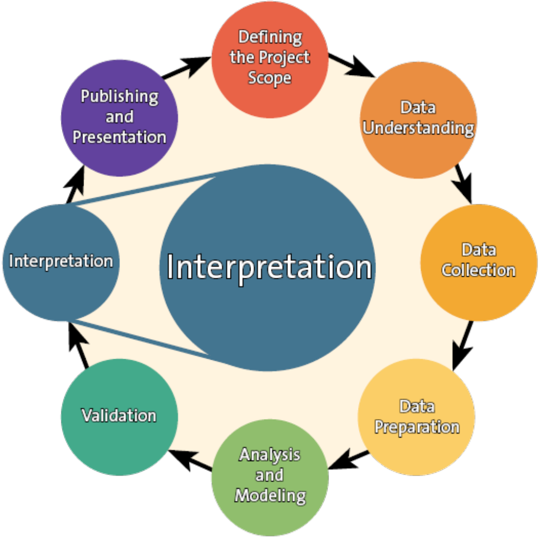
Data Modeling - Validation

- Are the results actually valid?
- Are the findings robust and reproducible?
- For machine learning models: does the model generalize well?

Validation: Example

- The model has an accuracy of up to 0.78
- It can predict the probability of the development of AD in other populations → robust
- Model generalizes well and is not overfitted




Defining the Project Scope
Data Understanding
Data Collection
Data Preparation
Analysis and Modeling
Validation
Interpretation
Publishing and Presentation
Interpretation

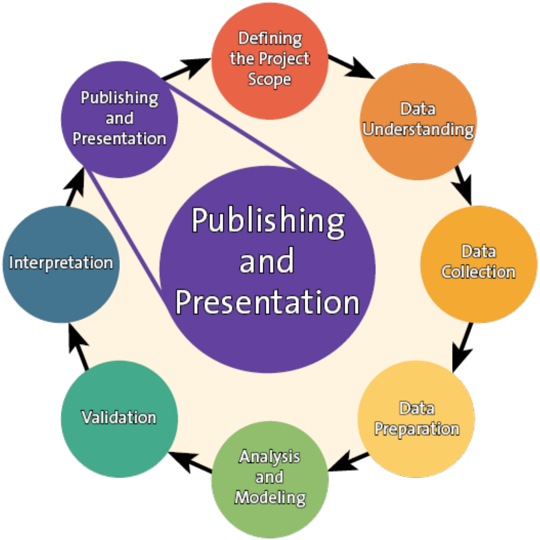
Interpretation

- How do we make sense of the results?
- What are the insights that can be derived from the results?

Interpretation: Example

- Blood samples were examined
- A model was trained with levels of many molecules
- The model was able to predict the probability of developing AD based on levels of molecule A and B

$\rightarrow$ **Molecule A and B may be involved in the development of AD**


Defining the Project Scope
Data Understanding
Data Collection
Data Preparation
Analysis and Modeling
Validation
Interpretation
Publishing and Presentation
Publishing and Presentation

Publishing and Presentation

- **"Data storytelling"**
- Summarize the results to make it easier to understand
- Visualize your results
- Communicate your findings to your peers, the scientific community, and to a wider audience

Publishing and Presentation

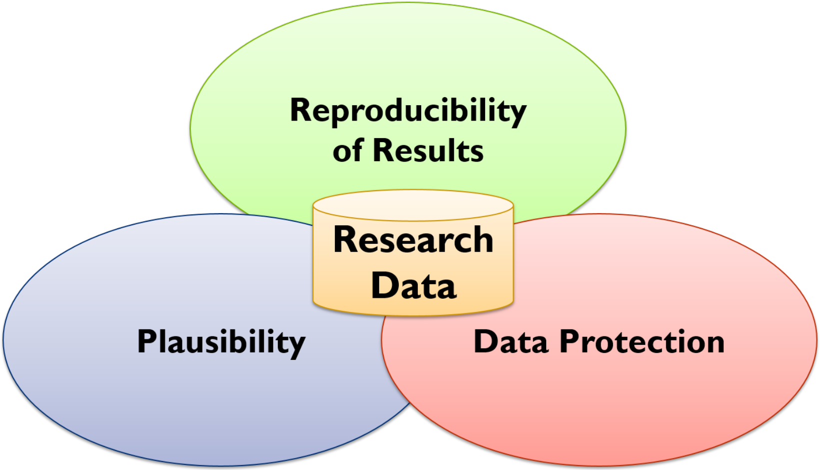
I don't strive for scientific perfection; that would be impossible. I focus on the big story. You can't change people's views with facts alone. That's a fallacy that many scientists fall into. People don't think in facts, they think in stories. To change human thinking, you have to be able to construct an alternative narrative.

Yuval Noah Harari
Historian, Philosopher, Best-Seller Author, Hebrew University of Jerusalem Professor

Publishing and Presentation

- Design meaningful figures and posters
- Publish a paper about your findings in a research journal
- Attend a conference and present your work

So, what could possibly go wrong..?!



Reproducibility of Results: Let's look at Research Data

Research data is essential for **reproducibility**

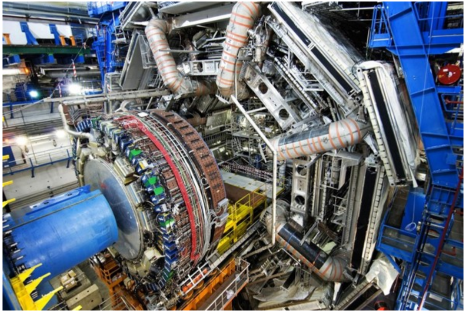
- Producing research data is often very costly
- Cross checks are needed to exclude methodological errors or bias in data
- Can I trust research results?

Best practices in some limited areas

- CERN data centers
- PANGAEA
- Re3data.org

NFDI Initiative

- NFDI4Life, NFDI4Ing,. . .



Example: Geo-Referenced Disease Control

Introducing **Google Flu Trends**

- **Basic idea:** when influenza hits an area, search queries using flu-related terms will be posed more often

- "Harnessing the collective intelligence of millions of users, Google Web Search logs can provide one of the most timely, broad-reaching influenza monitoring systems available today."

Example: Geo-Referenced Disease Control

A typical **Big Data** problem

- **Volume & Velocity:** Google processes about 50 million different search terms in about 3 billion queries per day
- **Variety:** search queries are textual information (i.e. unstructured), but the service aims at predicting the number of actual influenza cases
- **Veracity:** data is not cleaned, especially spelling mistakes ("pnumonia") and changes in user behavior may happen

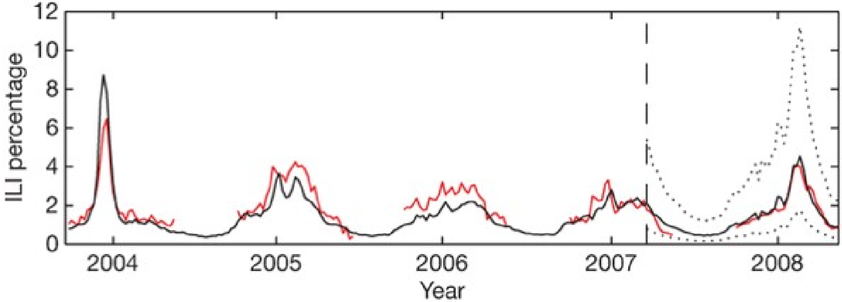
Example: Geo-Referenced Disease Control

Let's take a closer look...

- Google decided for an **automated method** of selecting flu-related search queries
  - Establish historical ground truth: yearly data e.g., from US Centers for Disease Control and Prevention (CDC)
  - For the time span of ground truth take each query term from user logs individually and test predictive value
  - Take the first few highest-scoring terms (about 45) and build a more complex model
  - Retrain once a year. . .

Example: Geo-Referenced Disease Control

Results: Google model (black) vs. ground truth (red)



- *J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, L. Brilliant: Detecting influenza epidemics using search engine query data. Nature, vol. 457(7232), 2009.*

Big Data rules!!!

Well,... Yes and No...

- Google Flu Trends is a good example of a **failed data science project**, because after the initial success predictions were far off
  - It estimated more than double the number of real cases
  - Among the top 100 predictive terms were terms such as *"OSCAR nominations"* and *"high school basketball"*..?!
- Later research found **severe problems** in measurements, reliability, stability, construct validity, and data dependencies
  - *D. Lazer, R. Kennedy, G. King, A. Vespignani: The Parable of Google Flu: Traps in Big Data Analysis. Science, vol. 343 (6176), 2014.*

What happened?

If you **throw Big Data at small problems** you may get any correlation that you might want to find...

- Google Flu Trends threw 50 million search terms against only **1152 data points** measured by the CDCs
  - Thus, overfitting a relatively small number of cases
  - Actually, simple statistical interpolations of the data points result in better and more stable estimates...
- Changing **environments** may influence results
  - For instance, media hypes will spark lots of queries without actual cases

Conclusions — Reproducibility of Results

Data sharing is crucial for finding out whether a result really is meaningful

- Scientists need to check results and compare with other (similar) data sets

The actual process of sharing **needs to be supported**

- Infrastructure is needed, like support by libraries or centralized repositories for communities
- Sharing and data generation cost time and are of scientific value
- Commercial use of data calls for clear data exploitation policies

Plausibility: Towards Building new Theories

Pass the Easter Egg! New study reveals that eating chocolate doesn't affect your Body Mass Index ... and can even help you LOSE weight!

- New research from Roy Morgan reveals there's no proof that chocolate consumption affects BMI
- Currently two thirds of Australians eat chocolate at least once a month
- A study from German researchers has also found there's a connection between cocoa diets and increased weight loss
- Chocolate also found to benefit brain, heart and stress levels

By SAM BAILEY FOR DAILY MAIL AUSTRALIA

PUBLISHED: 01:22 EST, 31 March 2015 | **UPDATED:** 16:14 EST, 31 March 2015

Share 1k shares 55 View comments

From the endless chocolate blocks passed around the office, to the glaring supermarket aisles and the family relatives who miraculously appear with baskets of eggs, Easter can be a minefield to navigate if you're trying to watch your waistline.

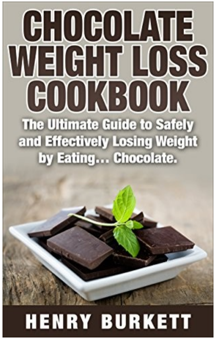
But according to new research, there's no need to go easy on the eggs this week, with a Roy Morgan study revealing there is no direct connection between chocolate consumption and an increasing Body Mass Index (BMI).

This should come as sweet relief for chocoholics when according to Roy Morgan, two thirds of Australians admit to munching on chocolate at least once a month.

- Article in Daily Mail Australia, March 31st, 2015

The Damage Done...

"Bitter chocolate tastes bad, therefore it must be good for you!"



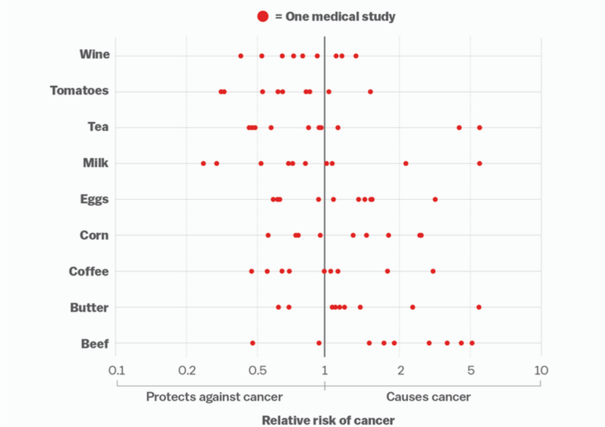
And there is more...

Jonathan Schoenfeld and John Ioannidis performed an **interesting experiment**

- Select 50 common ingredients from random recipes in a cookbook
- Query PubMed to identify recent studies that evaluate the relation of each ingredient to cancer risk
- Extract information regarding author conclusions and relevant effect estimates
  - Increased/decreased cancer risk
  - Use the ten most prominent articles for each ingredient
- *J. D. Schoenfeld, J. P. A. Ioannidis: Is everything we eat associated with cancer? A systematic cookbook review. Am. Journal Clinical Nutrition, vol. 97(1), 2013.*

Let's look at what the Clinical Trials say

Aggregated results



What happened?

All the clinical trials considered had **loads of variables** that could not be fully controlled

- Statistical significance? Missing or weak in 75% of studies...
- Incomplete descriptions of the test subjects vs. control groups
- Many studies highlight implausibly large effects, even though evidence is weak
- See for instance *http://retractionwatch.com/*

Conclusions — Plausibility of Results

Fitting new results into the **current world view** is crucial for building valid theories

- Methodologies need to be checked for soundness
- Research data management may also need to archive all evaluation tools

Methodological competences need to be taught stricter within scientific education

- Strong interdisciplinarity is needed as a service between communities

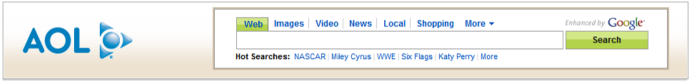
Data Protection: Your Data is Safe with us!

The tale of an anonymized dataset: **AOL Search**

- One of the major web search and content portals
- AOL serves millions of searches per day

Of course AOL has a **privacy policy**…

- However, internally records of all user searches are kept
- Search records are very valuable for improving algorithms
- In 2006, search data of 650,000 users over a 3-month period was published for free use by the IR research community



Your Data is Safe with us!

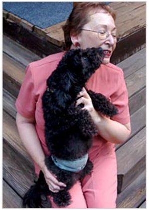
The New York Times, August 9th, 2006

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-MAIL THIS
PRINT
SINGLE PAGE
REPRINTS

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.



Erik S. Lesser for The New York Times
Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

What does Search Data Tell?

Most prominent example: **User #4417749**

- Thelma Arnold, 62-year-old, widowed, lives in Lilburn, Georgia
  - Is looking for a new partner in his 60s
  - Has at least one dog randomly pissing on furniture
  - Has problem with trembling fingers and aches in her back
  - Is worried about the safety of her neighborhood
  - Wonders about problems of the world, like hunger in Africa or children in war-torn Iraq

What does Search Data Tell?

But there are others: **User #311045**

- how to change brake pads on scion xb
- 2005 us open cup florida state champions
- how to get revenge on a ex
- how to get revenge on a ex girlfriend
- how to get revenge on a friend who f***ed you over
- replacement bumper for scion xb
- florida department of law enforcement
- crime stoppers florida

What does Search Data Tell?

Or how about: **User #11574916**

- cocaine in urine
- asian mail order brides
- states reciprocity with florida
- florida dui laws
- extradition from new york to florida
- mail order brides from lagos
- will one be extradited for a dui
- cooking jobs in french quarter new orleans
- will i be extradited from ny to fl on a dui charge

The Legal Aftermath

AOL **immediately** removed the dataset

- But it is still around on various mirrors and databases
- "Browse others' AOL data – hours of fun guaranteed"

In September 2006, a **class action lawsuit** was filed

- Seeks at least $5,000 for each person involved – i.e., 3.250 Billion Dollars!

The lawsuit was finally **settled in 2013**

- AOL pays up to $5 Million (U.S. citizens only)
- Tier 1 – $50 if you belief in good faith that your data was included
- Tier 2 – $100 if you were identified
- Plus individual settlements for tier 2 persons suffering harm

The Limits of Anonymization

The **Massachusetts Group Insurance Commission** had a bright idea back in the mid-1990s – it decided to release "anonymized" data on state employees

- The data showed all hospital visits after **removing all obvious identifiers** such as name, address, and social security number
- At the time of release, **William Weld**, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting all identifiers

Commonwealth of Massachusetts Group Insurance Commission

The Limits of Anonymization

Latanya Sweeney (Harvard U) started hunting for the Governor's hospital records

- Governor Weld resided in Cambridge, MA, a city of 54,000 residents and seven ZIP codes
  - For twenty dollars, she purchased the complete voter rolls from the city, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter
- By combining this data with the GIC records, Sweeney found Governor Weld quite easily
  - Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code
- Finally she sent the Governor's health records (which included diagnoses and prescriptions) to his office...

The Bottom Line

De-Anonymization by cross-referencing usually works amazingly well, not only in medical contexts...

- 87% of all Americans can be uniquely identified using only three pieces of information: ZIP code, birthdate, and sex
  - *L. A. Sweeney: Simple Demographics Often Identify People Uniquely. Carnegie Mellon Univ. Data Privacy Working Paper 3, 2000.*
- Failures of anonymization in system interactions are not really understood by users
  - *P. Ohm: Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Rev. 57(1701), 2010.*

The Bottom Line

Even worse... Information fusion in **Social Networks**

- Every interaction with social networks, every piece of information, every decision for accepting/rejecting a friendship request tells something about the user and can (and will) be used as an attribute for later classification
  - What Facebook „likes" tell about its members
    M. Kosinski, D. Stillwell, T. Graepel: Private traits and attributes are predictable from digital records of human behavior. PNAS 110(15), 2013.
  - What Facebook friendships tell about your sexual orientation
    C. Jernigan, B. F. T. Mistree: Gaydar: Facebook Friendships Expose Sexual Orientation. First Monday Vol. 14(10), 2009.

Conclusions — Data Protection

Data protection within data science projects is obviously a **necessary task**, but. . .

- Does anonymization really work?
- Are scientists also data protection experts?
- Who is liable in case something actually happens?

Data protection laws may be a great nuisance when working on personal data, but they have to be respected!

Thanks for your Attention!

Acknowledgements: all materials prepared by

- Lisa-Marie Bente
- Gihanna Galindez
- Wolf-Tilo Balke
- Tim Kacprowski

If there are any questions, don't hesitate to contact us at t.kacprowski@tu-braunschweig.de or balke@ifis.cs.tu-bs.de