

Introduction to Data Science

Experiment Design for Data Science: Block 1, Lecture 1

Allan Hanbury

Institute for Information Systems Engineering,
TU Wien

Data Science:

To gain insights into
data through
computation, statistics,
and visualization

“A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.”



Joshua Blumenstock
(University of California, Berkeley)

- Identifying, evaluating and implementing innovative business models

Big Data Business Developer

- Infrastructure development
- Infrastructure provision
- Scalable memory
- Massive infrastructure

Big Data Technologist

- Innovative fusion of data
- Machine Learning
- Statistics and Mathematics

Big Data Analyst

Data Scientist

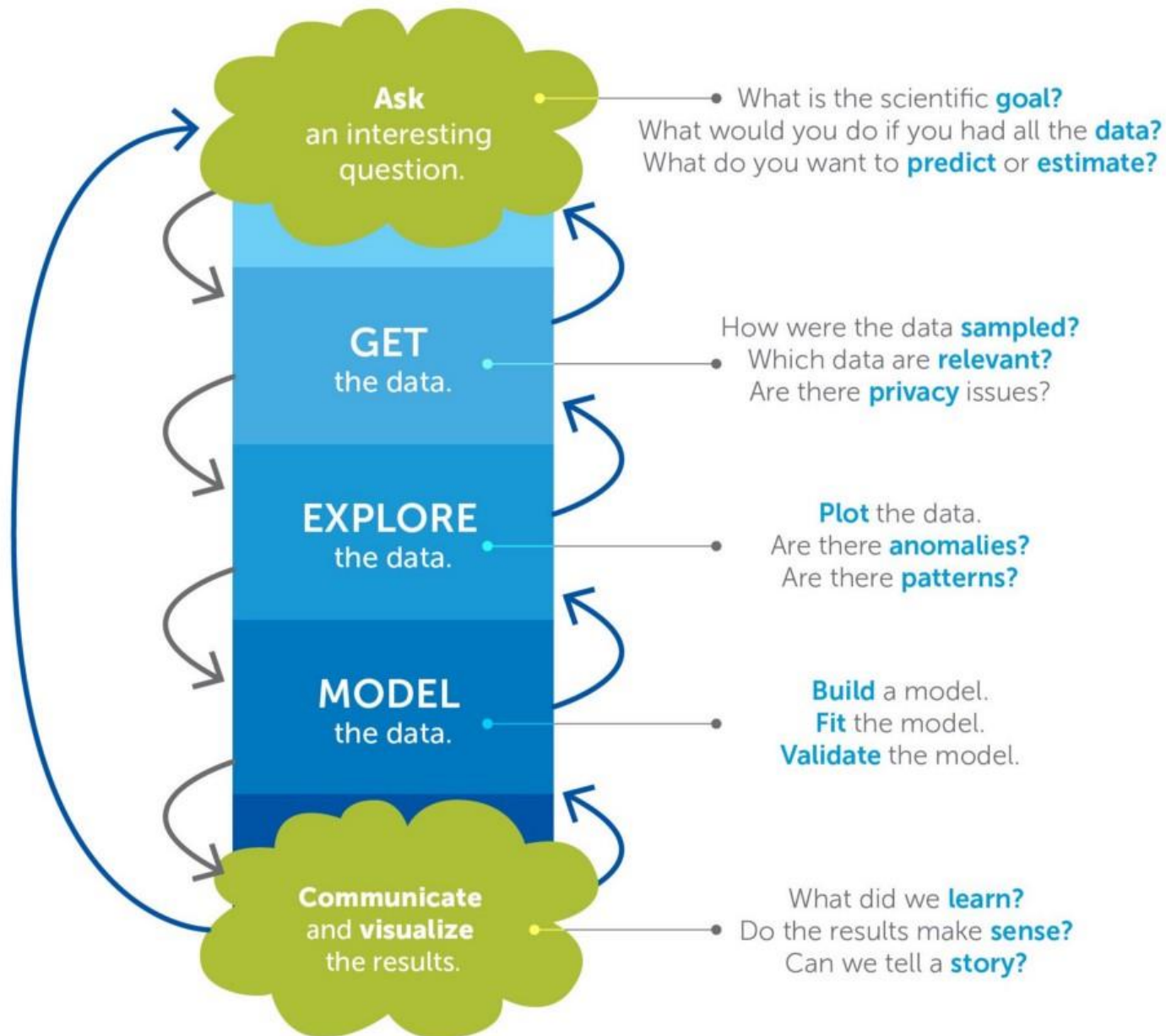
Big Data Developer

- Scalable programming
- Machine Learning
- Data management

Big Data Artist

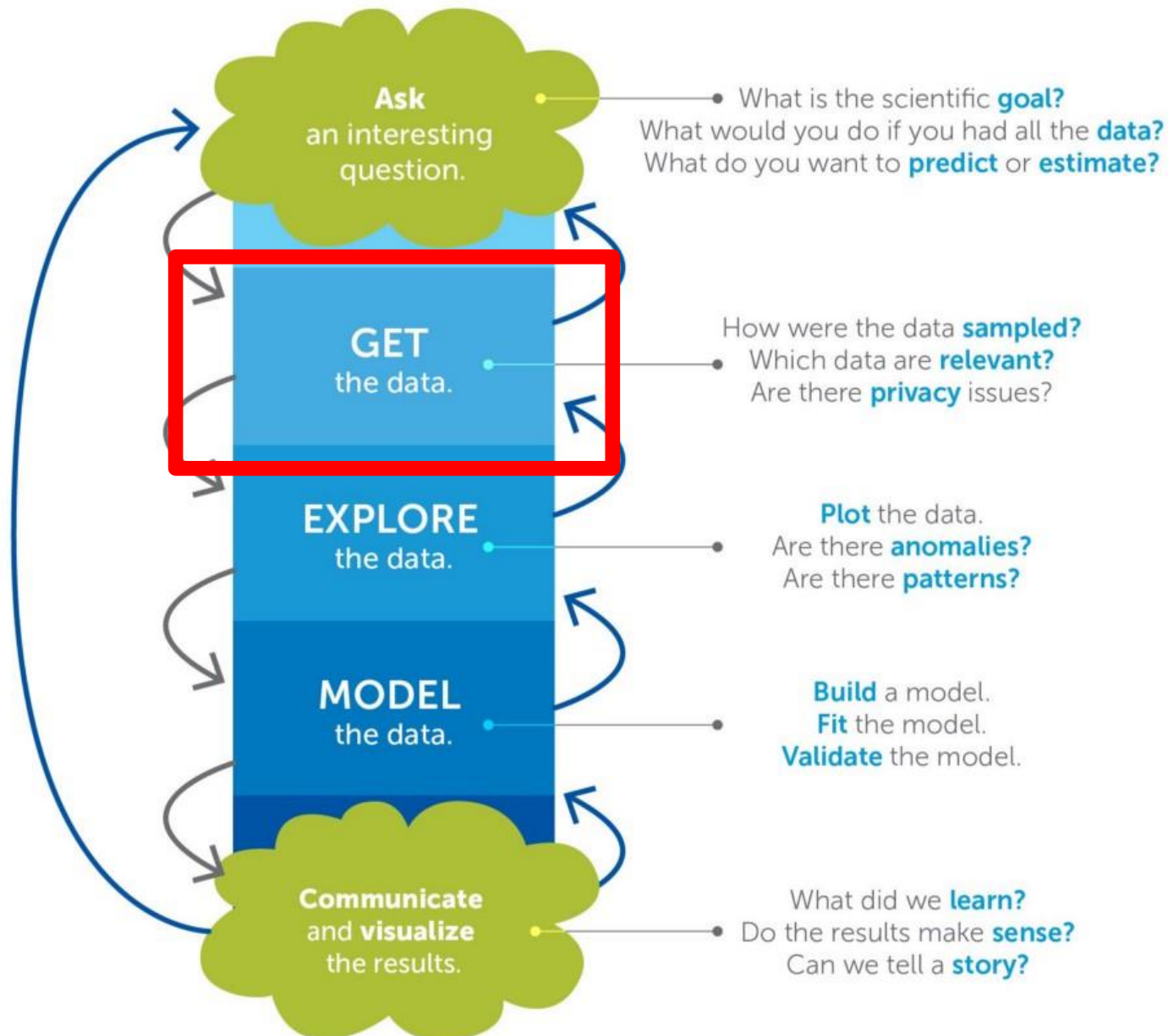
- Data visualisation
- Graphic design
- Communication
- Psychology

The Data Science Process



Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course <http://cs109.org/>.

The Data Science Process



Derived from the work of Joe Blitzstein and Hanspeter Pfister, originally created for the Harvard data science course <http://cs109.org/>.

- Structured
- Unstructured
- Natural language
- Machine-generated
- Graph-based
- Audio, video, and images
- Streaming

- Depends on a data model
- Resides in a fixed field within a record
- Often stored as tables within databases or Excel files
- SQL (Structured Query Language) usually used to manage and query structured data
- Hierarchical data (e.g. family tree) is structured, but not easy to put into a database or Excel

1	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Inte
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 years	2008	88.6%		87.7%
7	214390834	Aged 45-54 years	2008	86.3%		85.1%
8	214390835	Aged 55-64 years	2008	91.5%		90.4%
9	214390840	Aged 65 years and over	2008	94.6%		93.8%
10	214390837	Aged 65-74 years	2008	93.6%		92.4%
11	214390838	Aged 75-84 years	2008	95.6%		94.4%
12	214390839	Aged 85 years and over	2008	96.0%		94.0%
13	214390841	Male (Age-adjusted)	2008	72.2%		71.1%
14	214390842	Female (Age-adjusted)	2008	76.8%		75.9%

- Unstructured data is data that isn't easy to fit into a data model because the content is context-specific or varying
- Could contain structured elements:
 - sender
 - title
 - body text
 - ...



- Example: text in e-mails, letters, Tweets, blogs, scientific papers
- Beware of “unnatural language” – language that is very specific to a domain, e.g. legal texts



20256736-092602

Attorney Docket No. APL1P223/P2727

PATENT APPLICATION

FOR

METHOD AND APPARATUS FOR ACCELERATED SCROLLING

Inventors: 1. Robert Tsuk
2. Jeffrey L. Robbin

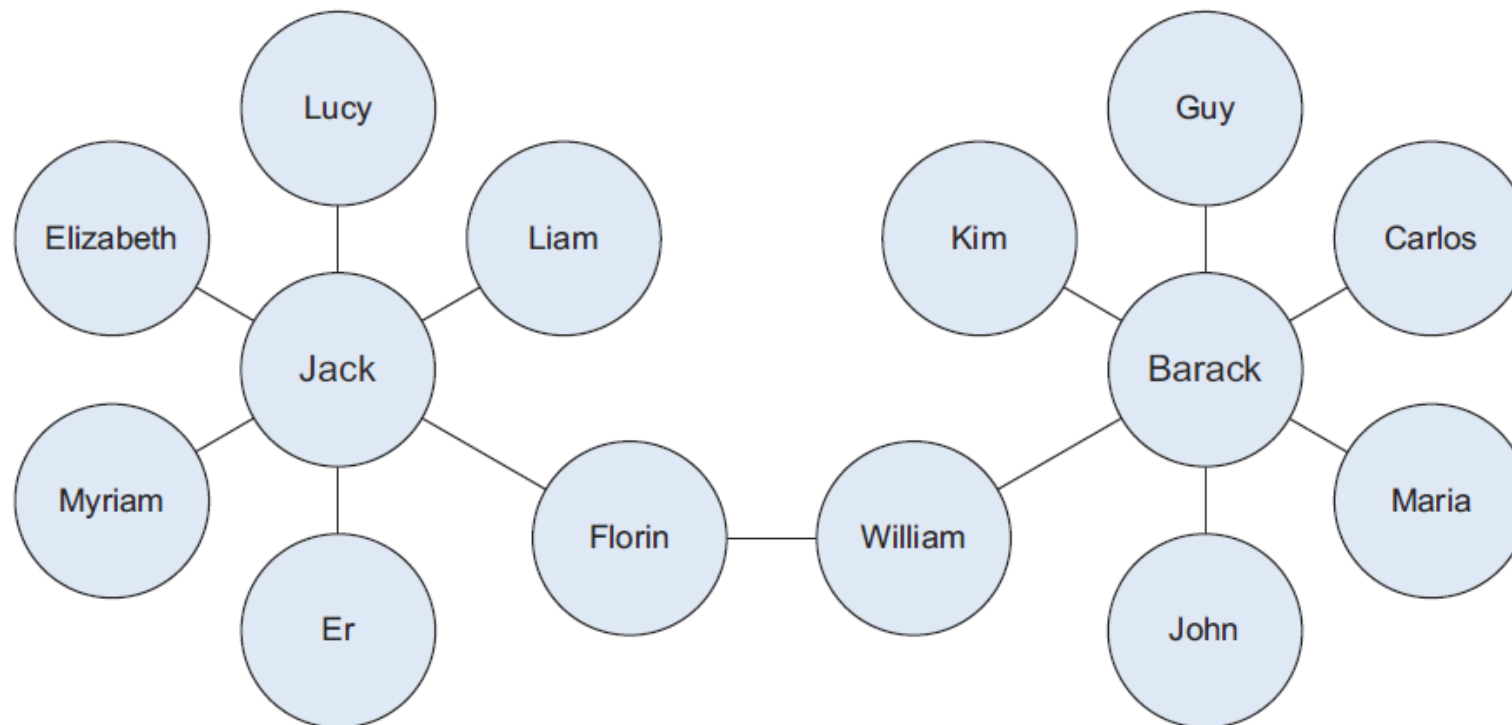
Assignee: Apple Computer, Inc.

BEYER WEAVER & THOMAS, LLP
(650) 961-8300

- Information automatically created by a computer, process, application, or other machine without human intervention
- Industrial Internet, Internet of Things, ...
- Usually high volume and speed
- Examples: web server logs, call detail records, network event logs, and telemetry

CSIPERF:TXCOMMIT;313236	
2014-11-28 11:36:13, Info	CSI 00000153 Creating NT transaction (seq
69), objectname [6]"(null)"	
2014-11-28 11:36:13, Info	CSI 00000154 Created NT transaction (seq 69)
result 0x00000000, handle @0x4e54	
2014-11-28 11:36:13, Info	CSI 00000155@2014/11/28:10:36:13.471
Beginning NT transaction commit...	
2014-11-28 11:36:13, Info	CSI 00000156@2014/11/28:10:36:13.705 CSI perf
trace:	
CSIPERF:TXCOMMIT;273983	
2014-11-28 11:36:13, Info	CSI 00000157 Creating NT transaction (seq
70), objectname [6]"(null)"	
2014-11-28 11:36:13, Info	CSI 00000158 Created NT transaction (seq 70)
result 0x00000000, handle @0x4e5c	
2014-11-28 11:36:13, Info	CSI 00000159@2014/11/28:10:36:13.764

- Graph-based data is a natural way to represent e.g. social networks: nodes, edges, weights
- Its structure allows one to calculate specific metrics such as the influence of a person and the shortest path between two people
- Overlapping graphs with the same nodes but different information in connections is powerful



- More complex to process than text data
- Huge steps made recently in this area, in particular with deep learning

Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification

Kaiming He

Xiangyu Zhang

Shaoqing Ren

Jian Sun

Microsoft Research

{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

*Based on our PReLU networks (PReLU-nets), we achieve **4.94%** top-5 test error on the ImageNet 2012 classification dataset. This is a 26% relative improvement over the ILSVRC 2014 winner (GoogLeNet, 6.66% [29]). To our knowledge, **our result is the first to surpass human-level performance** (5.1%, [22]) on this visual recognition challenge.*

February 2015,
<http://arxiv.org/abs/1502.01852>

- Streaming data can take almost any of the previous forms
- Streaming data flows into the system when an event happens instead of being loaded into a data store in a batch
- Examples: “What’s trending” on Twitter, live sporting or music events, and the stock market

Can you get the data?

Legal issues

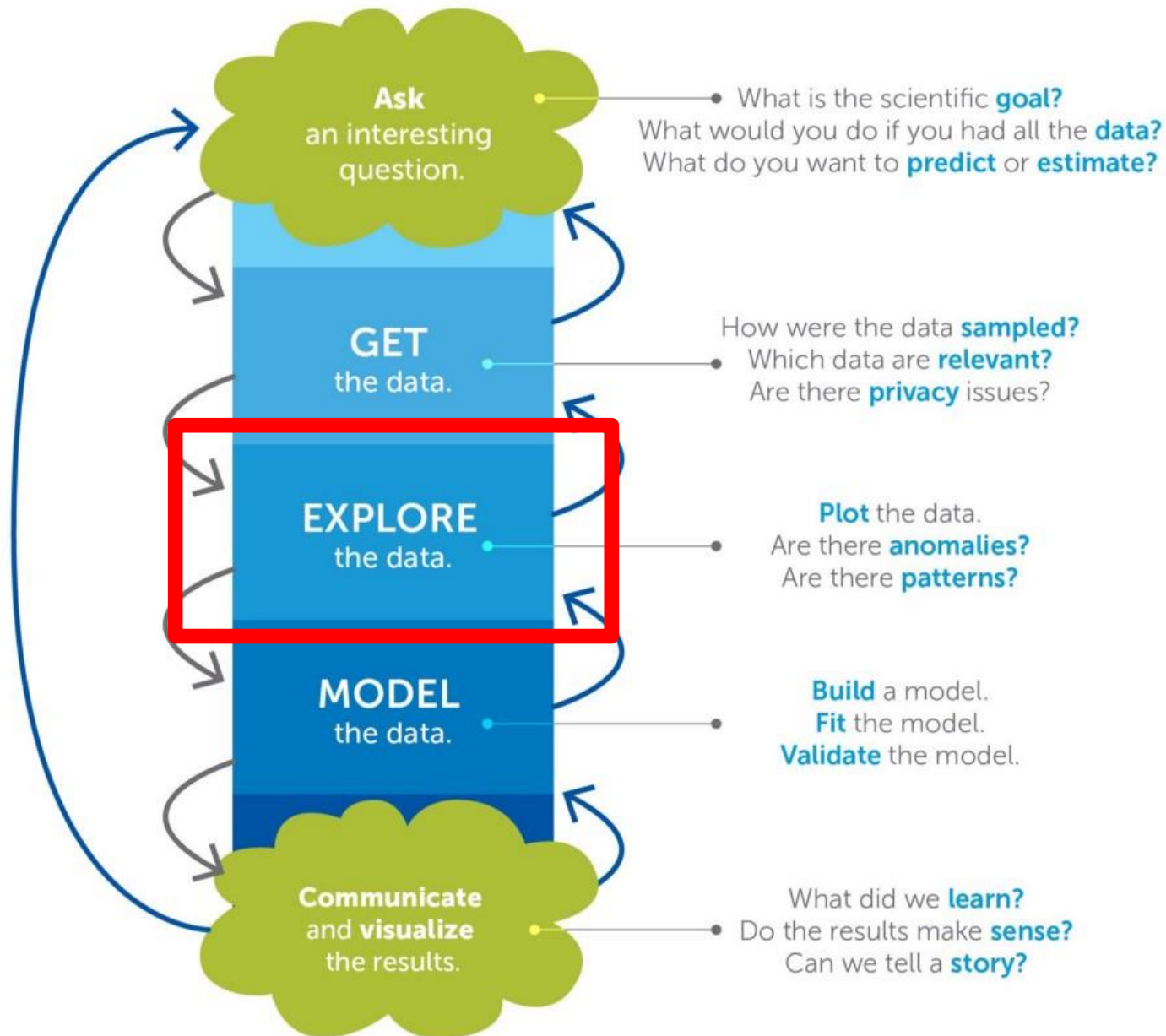
Privacy issues



Is the data suitable for answering the question and of sufficient quality?

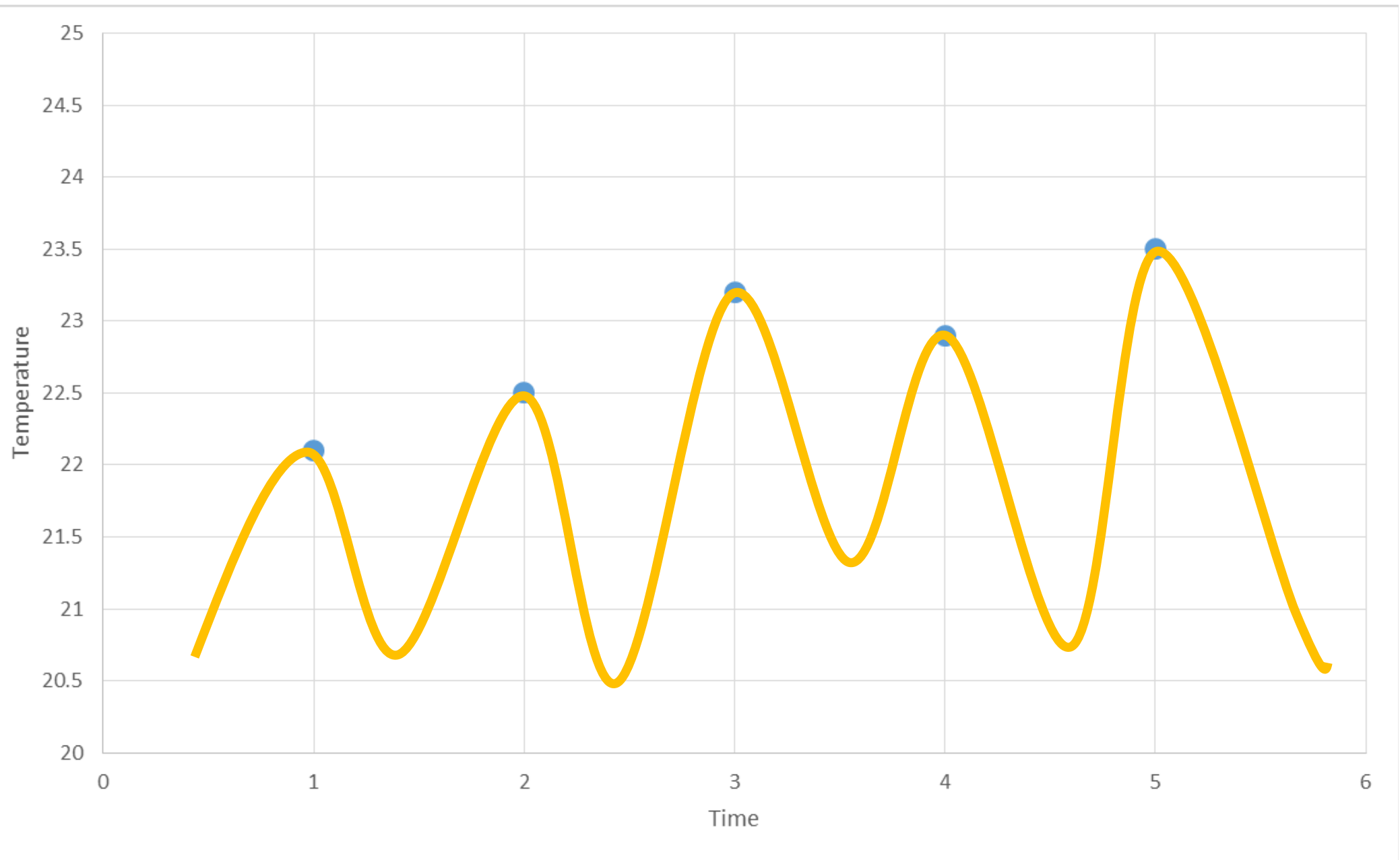
- Missing column headings
- Data is not what is stated in the column headings
- Missing values
- Data from different sources with different time ranges
- Anonymisation has removed necessary information
- Data is scanned PDFs of printed out Word documents
- ...

The Data Science Process

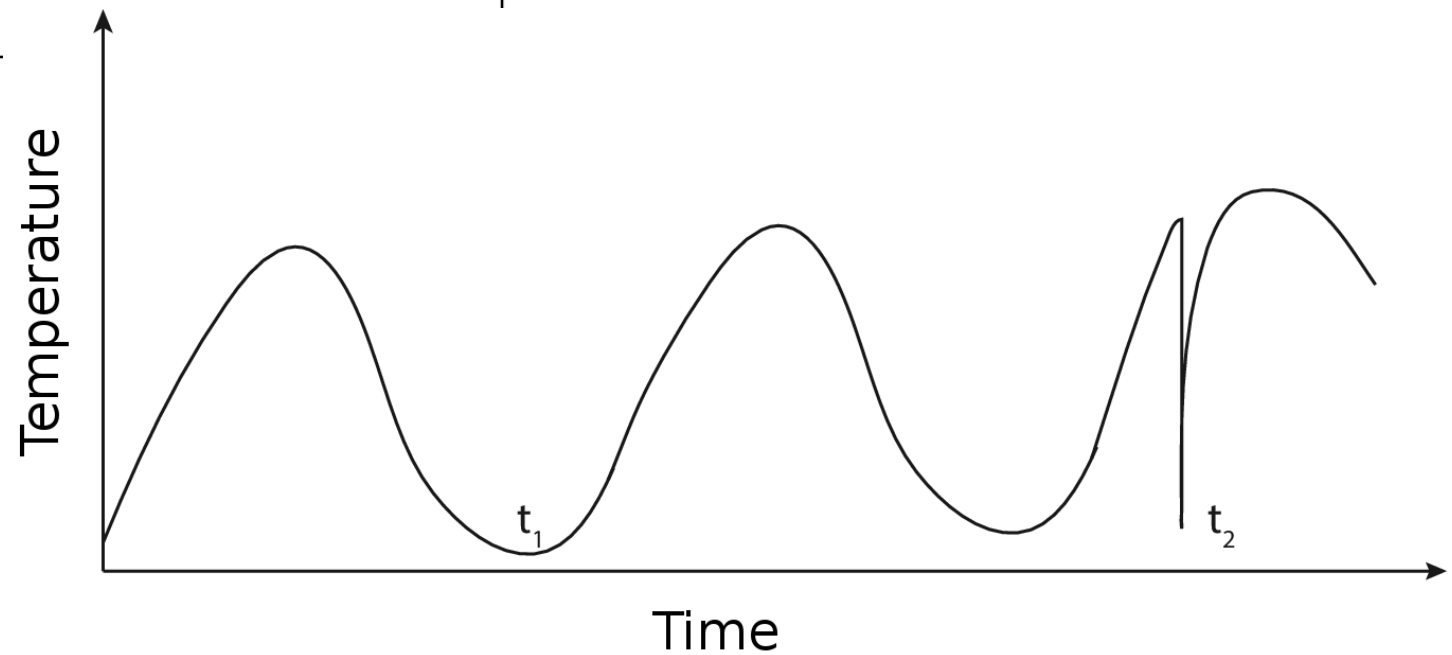
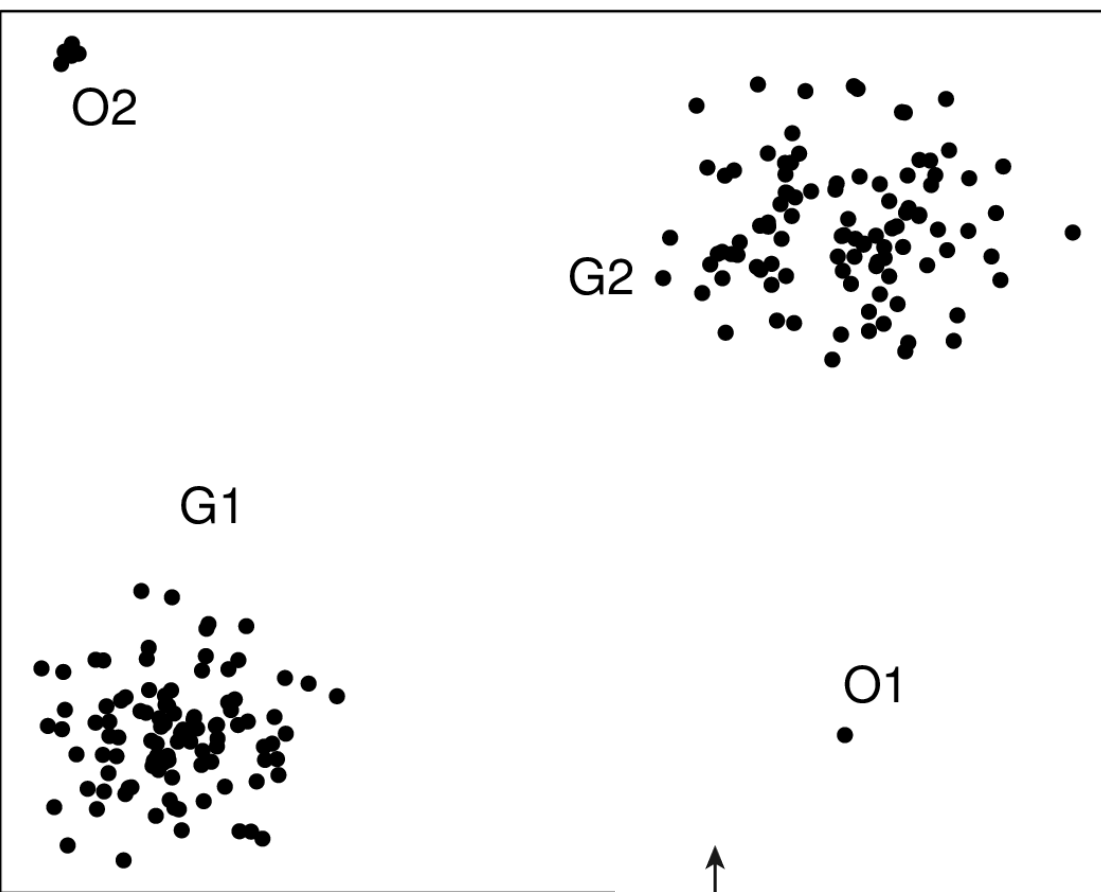


Derived from the work of Joe Blitzstein and Hanspeter Pfister, originally created for the Harvard data science course <http://cs109.org/>.

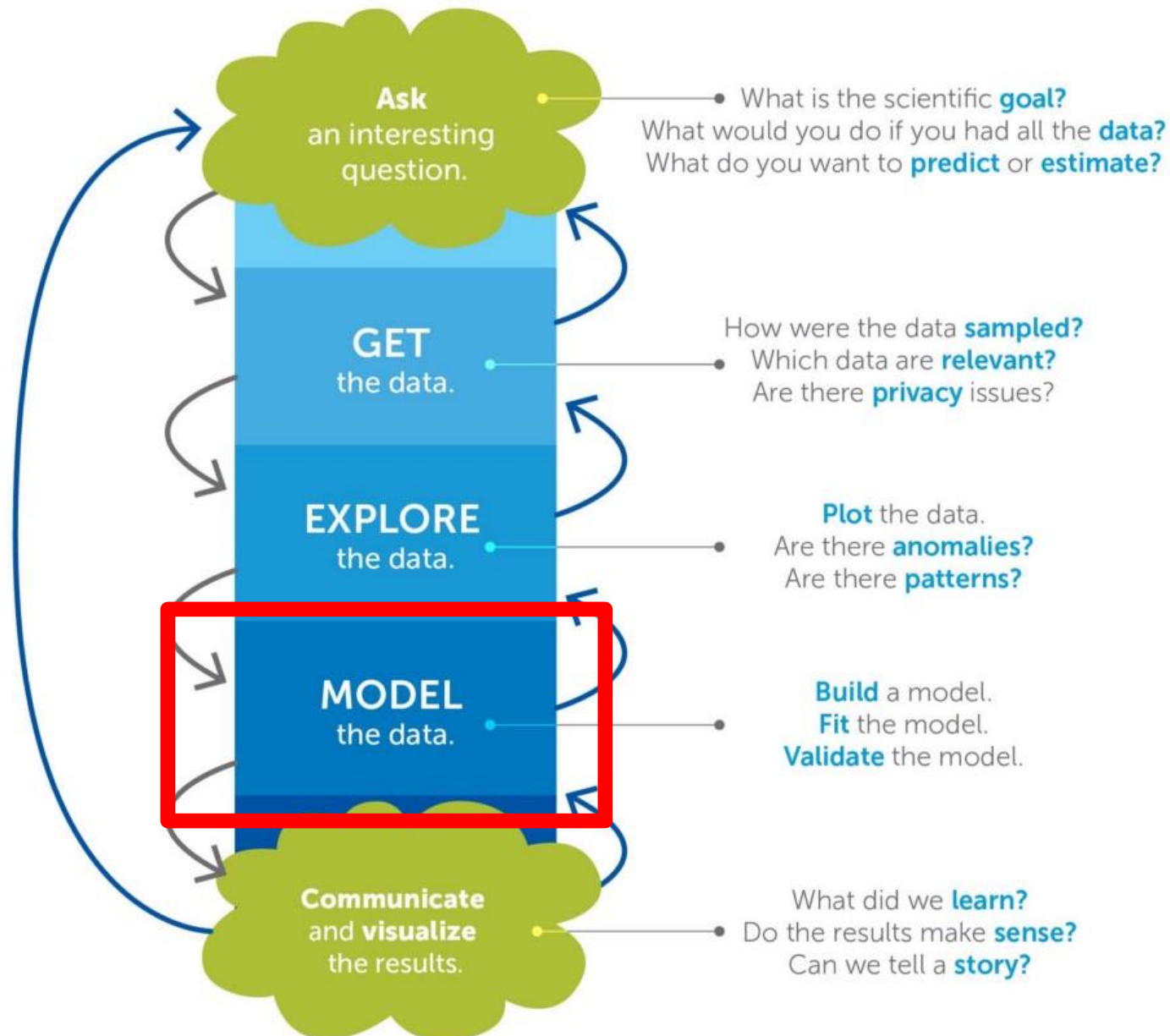
Do you understand the data fully?



Outliers / Anomalies



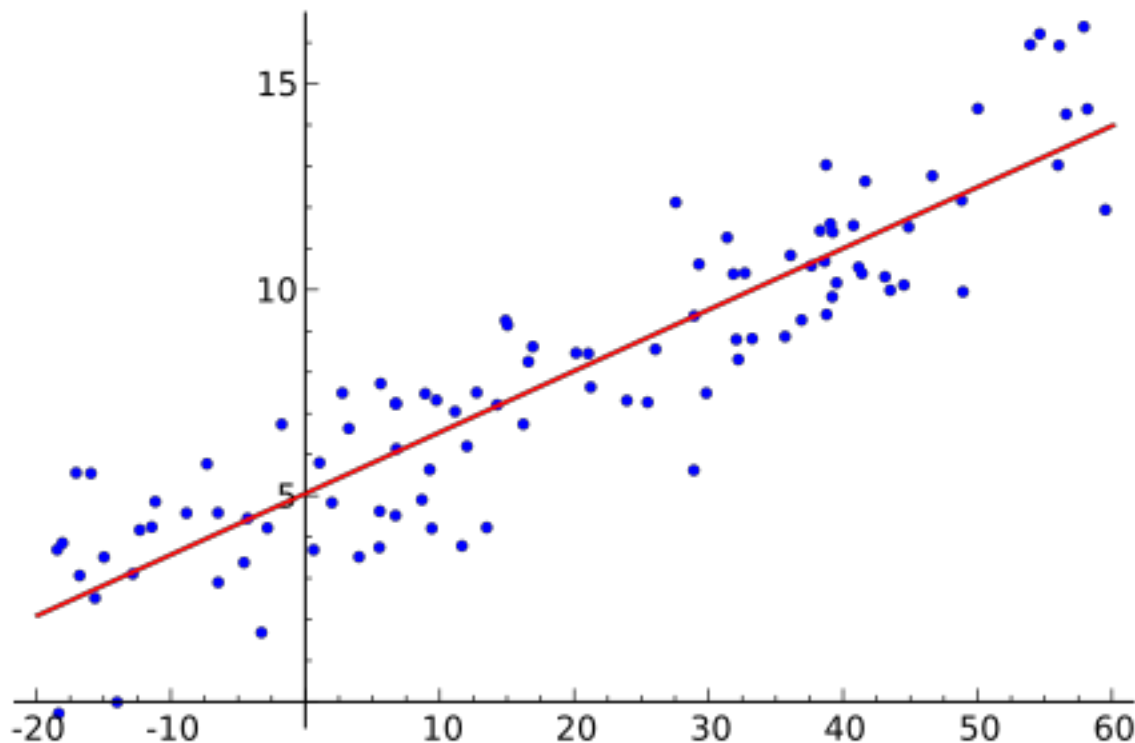
The Data Science Process



Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course <http://cs109.org/>.

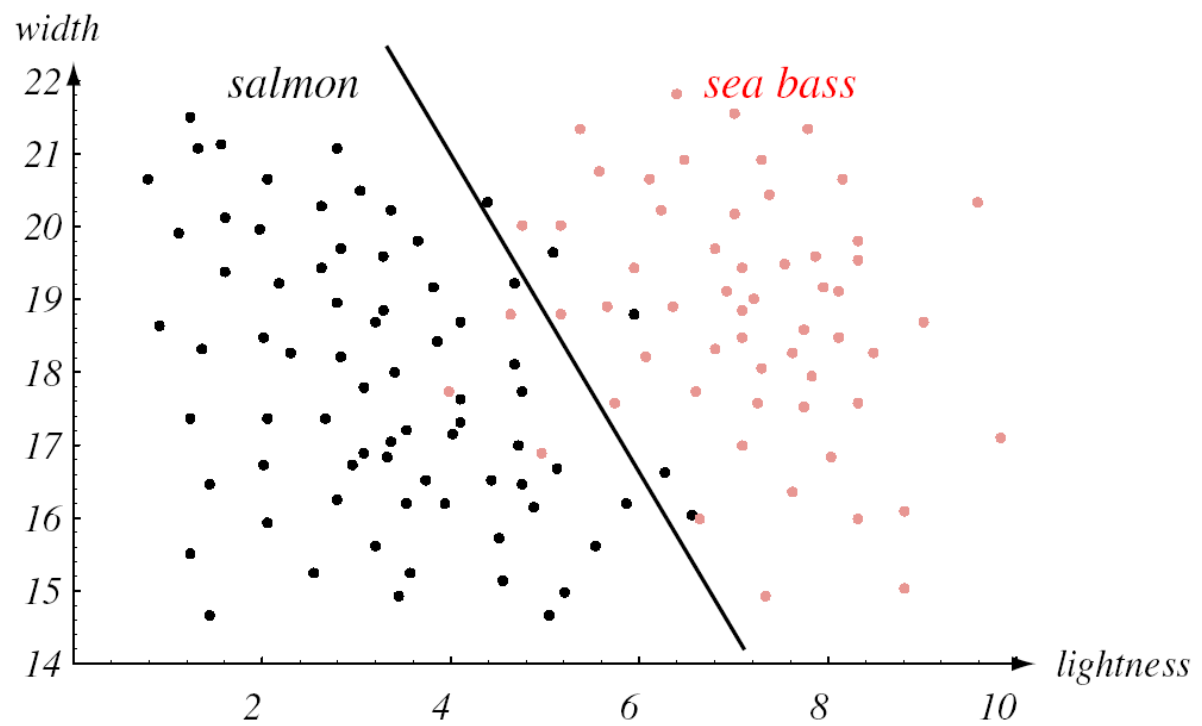
Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known

Regression is the processes of estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors')

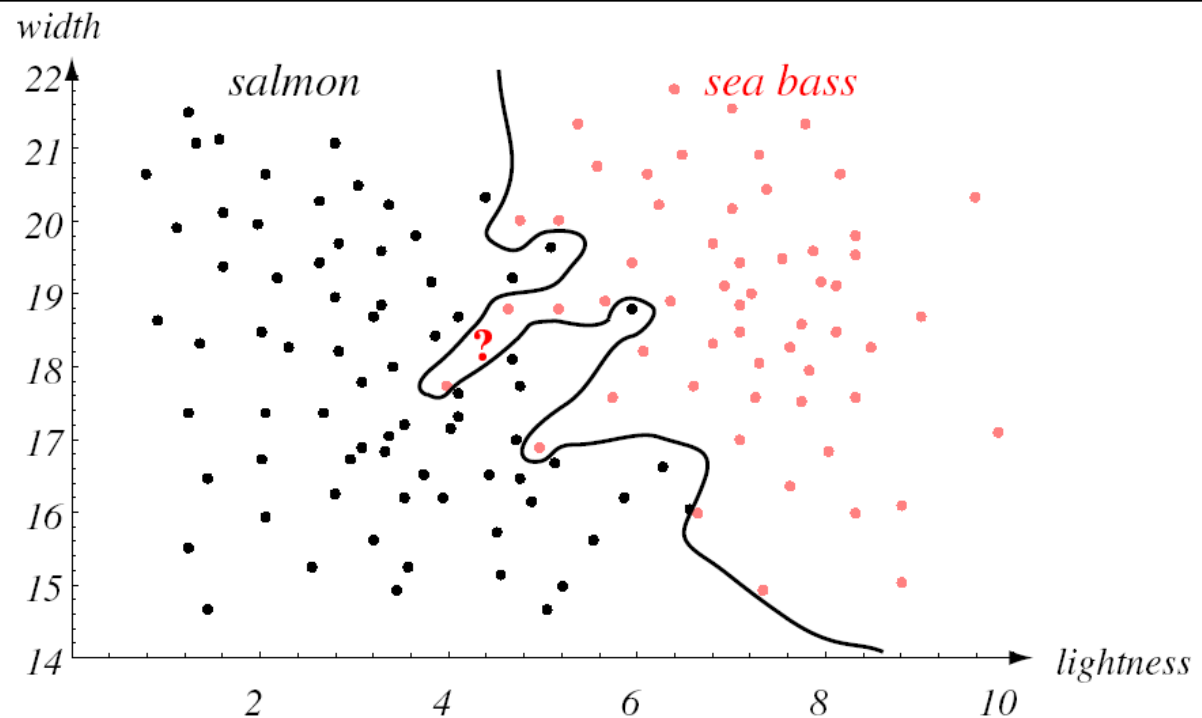


- How well the trained model fits the real world depends on:
 - The model selected
 - The data used for training
- These are interdependent
- A way of thinking about this is with **Bias** and **Variance** of the model

- Bias high
(*underfitting*)



- Variance high
(*overfitting*)

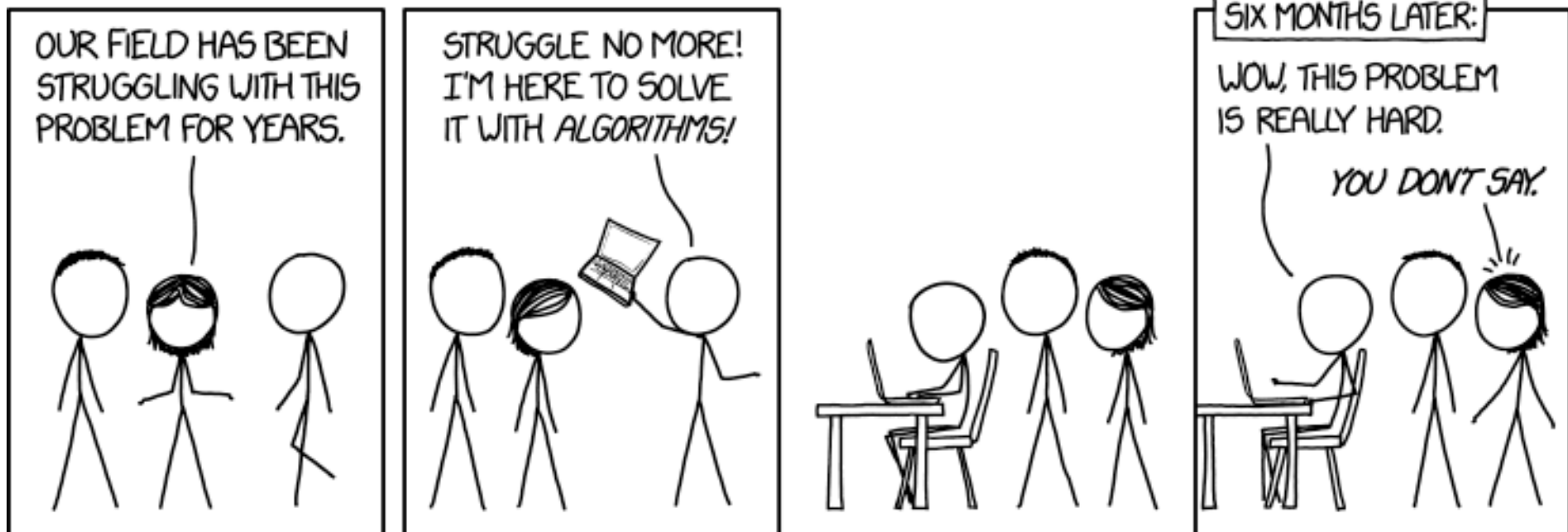


Summary – Bias and Variance in the Model

- Model Bias
 - High for a model that is too simple or inflexible
- Model Variance
 - High for a model that is too flexible with respect to the training data
- Bias and Variance are not independent of each other

What's Hard about Data Science? (1)

- Getting the data (usually)
- Overcoming assumptions
- Communication:
 - with domain experts
 - expectation management for client



What's Hard about Data Science? (2)

- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity



Introduction to Ethical and Legal Aspects

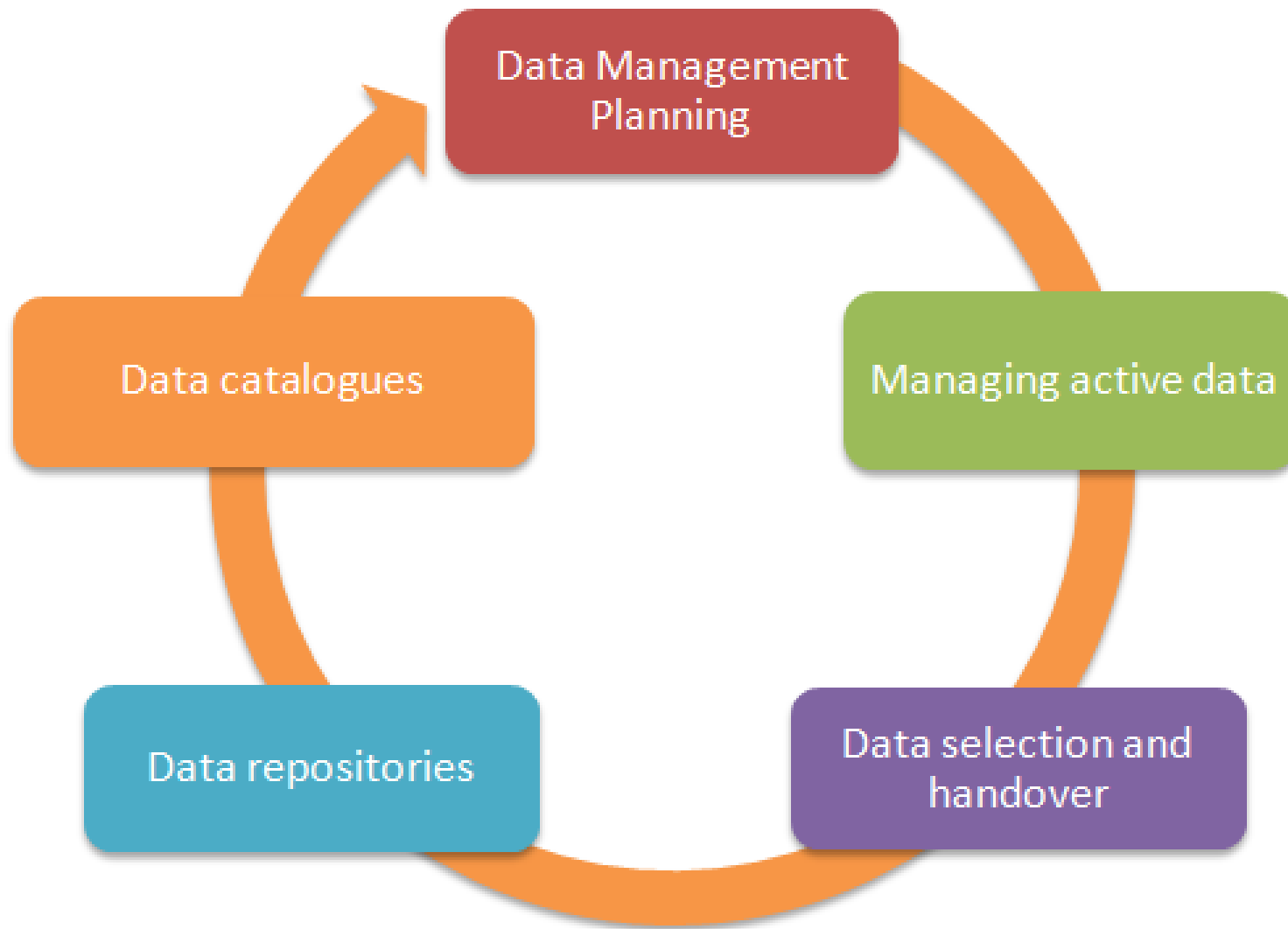
Experiment Design for Data Science: Block 1, Lecture 2

Allan Hanbury

Institute for Information Systems Engineering,
TU Wien

- Data and the Data Lifecycle
- Data Ethics and Legal Aspects
- Algorithm Ethics

DATA AND THE DATA LIFECYCLE





- Data Management Plan (DMP) documents:
 - how the data will be created
 - how it will be documented
 - who will be able to access it
 - where it will be stored
 - who will back it up
 - whether (and how) it will be shared & preserved



- Consider the cases for own hosting or outsourcing



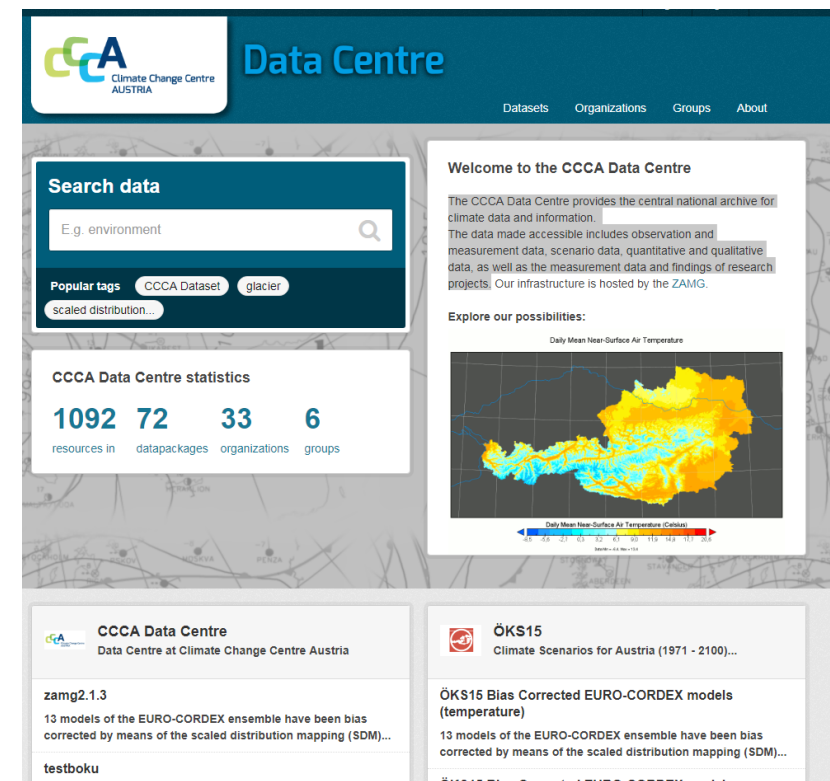
- Where appropriate, make a case for investment to provide additional data storage
- Develop procedures for the allocation and management of data storage
- Provide flexible systems to support the creation, management and sharing of data that meet a diverse range of contexts and needs



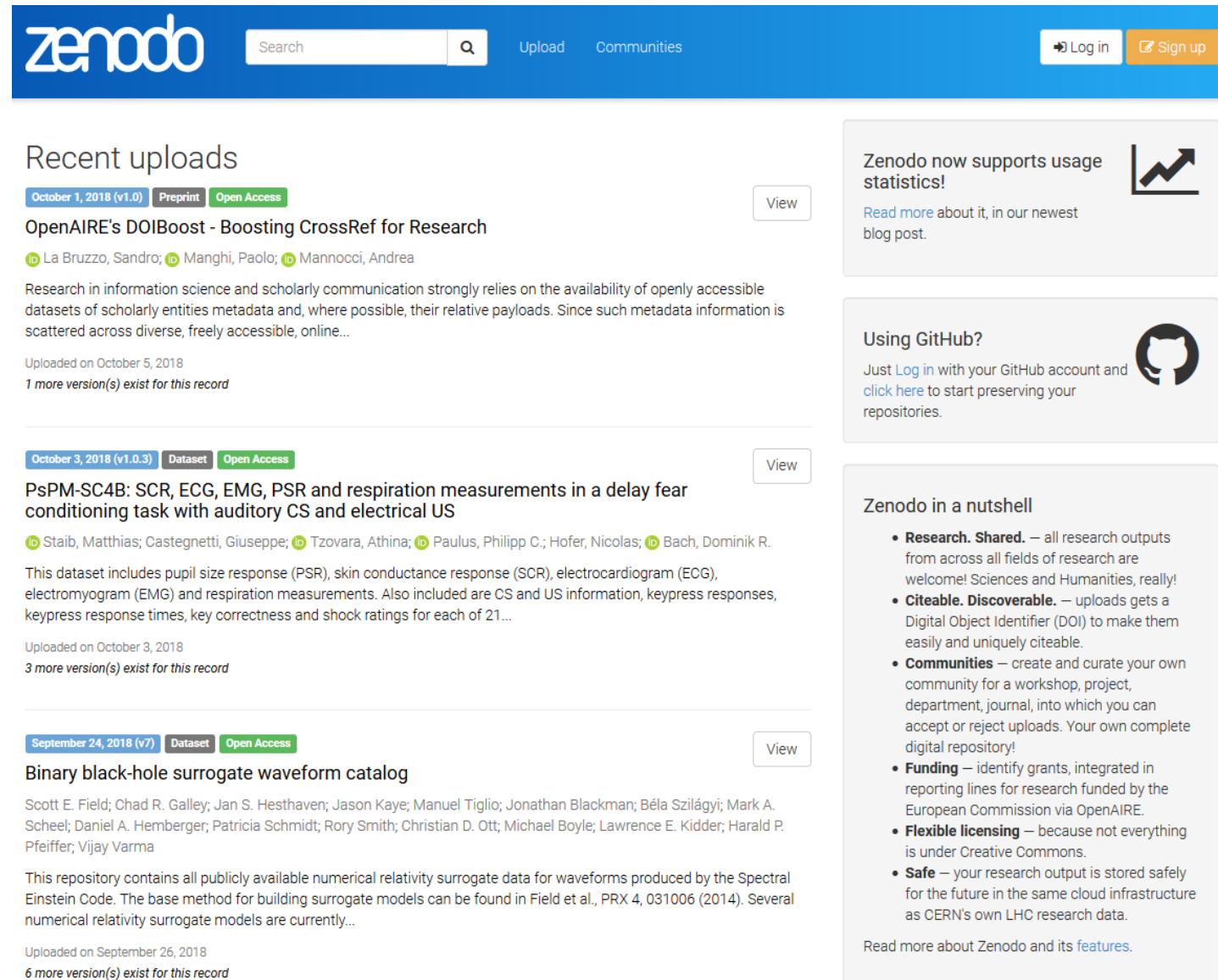
- Once data is no longer actively used, it can be placed in a repository
- Establish criteria to guide decisions on
 - What to keep
 - What to delete
- Ensure that all information/specifications for data reuse also go into the repository

- Consider internal and external repositories
- Consider measures against data loss
- Consider measures for keeping data formats up-to-date

- For scientific data, external repositories exist, e.g. Climate Change Centre Austria
 - observation and measurement data
 - scenario data
 - quantitative and qualitative data
 - measurement data
 - findings of research projects



- Make sure that your data remains findable
- Define metadata to ensure this
- Index metadata in a search engine, or, for open data, expose it for inclusion in national catalogues



The screenshot shows the Zenodo website interface. At the top is a blue header with the Zenodo logo, a search bar, and links for 'Upload' and 'Communities'. On the right side of the header are buttons for 'Log in' and 'Sign up'.

The main content area is divided into two columns. The left column, titled 'Recent uploads', lists three datasets:

- OpenAIRE's DOIBoost - Boosting CrossRef for Research**: Uploaded on October 1, 2018 (v1.0). It is a Preprint and Open Access. The description states: 'Research in information science and scholarly communication strongly relies on the availability of openly accessible datasets of scholarly entities metadata and, where possible, their relative payloads. Since such metadata information is scattered across diverse, freely accessible, online...'. It was uploaded on October 5, 2018, and 1 more version exists for this record.
- PsPM-SC4B: SCR, ECG, EMG, PSR and respiration measurements in a delay fear conditioning task with auditory CS and electrical US**: Uploaded on October 3, 2018 (v1.0.3). It is a Dataset and Open Access. The description states: 'This dataset includes pupil size response (PSR), skin conductance response (SCR), electrocardiogram (ECG), electromyogram (EMG) and respiration measurements. Also included are CS and US information, keypress responses, keypress response times, key correctness and shock ratings for each of 21...'. It was uploaded on October 3, 2018, and 3 more versions exist for this record.
- Binary black-hole surrogate waveform catalog**: Uploaded on September 24, 2018 (v7). It is a Dataset and Open Access. The description states: 'This repository contains all publicly available numerical relativity surrogate data for waveforms produced by the Spectral Einstein Code. The base method for building surrogate models can be found in Field et al., PRX 4, 031006 (2014). Several numerical relativity surrogate models are currently...'. It was uploaded on September 26, 2018, and 6 more versions exist for this record.

The right column contains three informational boxes:

- Zenodo now supports usage statistics!**: Includes a line graph icon and a link to 'Read more about it, in our newest blog post.'
- Using GitHub?**: Includes the GitHub logo and text: 'Just Log in with your GitHub account and click here to start preserving your repositories.'
- Zenodo in a nutshell**: A list of features:
 - Research. Shared.** — all research outputs from across all fields of research are welcome! Sciences and Humanities, really!
 - Citeable. Discoverable.** — uploads gets a Digital Object Identifier (DOI) to make them easily and uniquely citeable.
 - Communities** — create and curate your own community for a workshop, project, department, journal, into which you can accept or reject uploads. Your own complete digital repository!
 - Funding** — identify grants, integrated in reporting lines for research funded by the European Commission via OpenAIRE.
 - Flexible licensing** — because not everything is under Creative Commons.
 - Safe** — your research output is stored safely for the future in the same cloud infrastructure as CERN's own LHC research data.

<https://zenodo.org/>

DATA ETHICS AND LEGAL ASPECTS



(Legislative acts)

REGULATIONS

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

Article 99

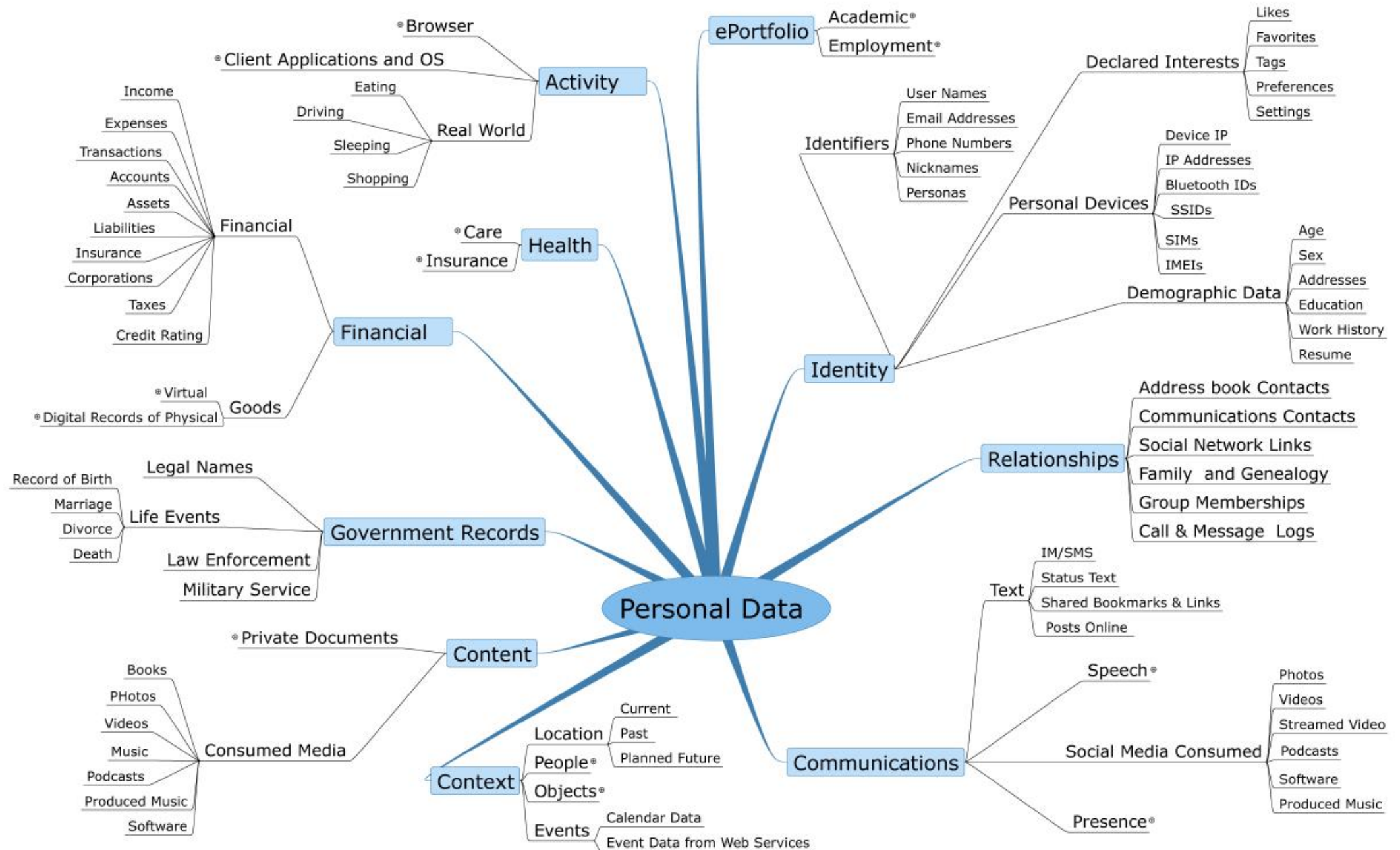
Entry into force and application

1. This Regulation shall enter into force on the twentieth day following that of its publication in the *Official Journal of the European Union*.
2. It shall apply from 25 May 2018.

5. Infringements of the following provisions shall, in accordance with paragraph 2, be subject to administrative fines up to 20 000 000 EUR, or in the case of an undertaking, up to 4 % of the total worldwide annual turnover of the preceding financial year, whichever is higher:

- (a) the basic principles for processing, including conditions for consent, pursuant to Articles 5, 6, 7 and 9;
- (b) the data subjects' rights pursuant to Articles 12 to 22;
- (c) the transfers of personal data to a recipient in a third country or an international organisation pursuant to Articles 44 to 49;
- (d) any obligations pursuant to Member State law adopted under Chapter IX;
- (e) non-compliance with an order or a temporary or definitive limitation on processing or the suspension of data flows by the supervisory authority pursuant to Article 58(2) or failure to provide access in violation of Article 58(1).

Types of personal data



How careful are you with your personal data?

How much is your personal data worth?

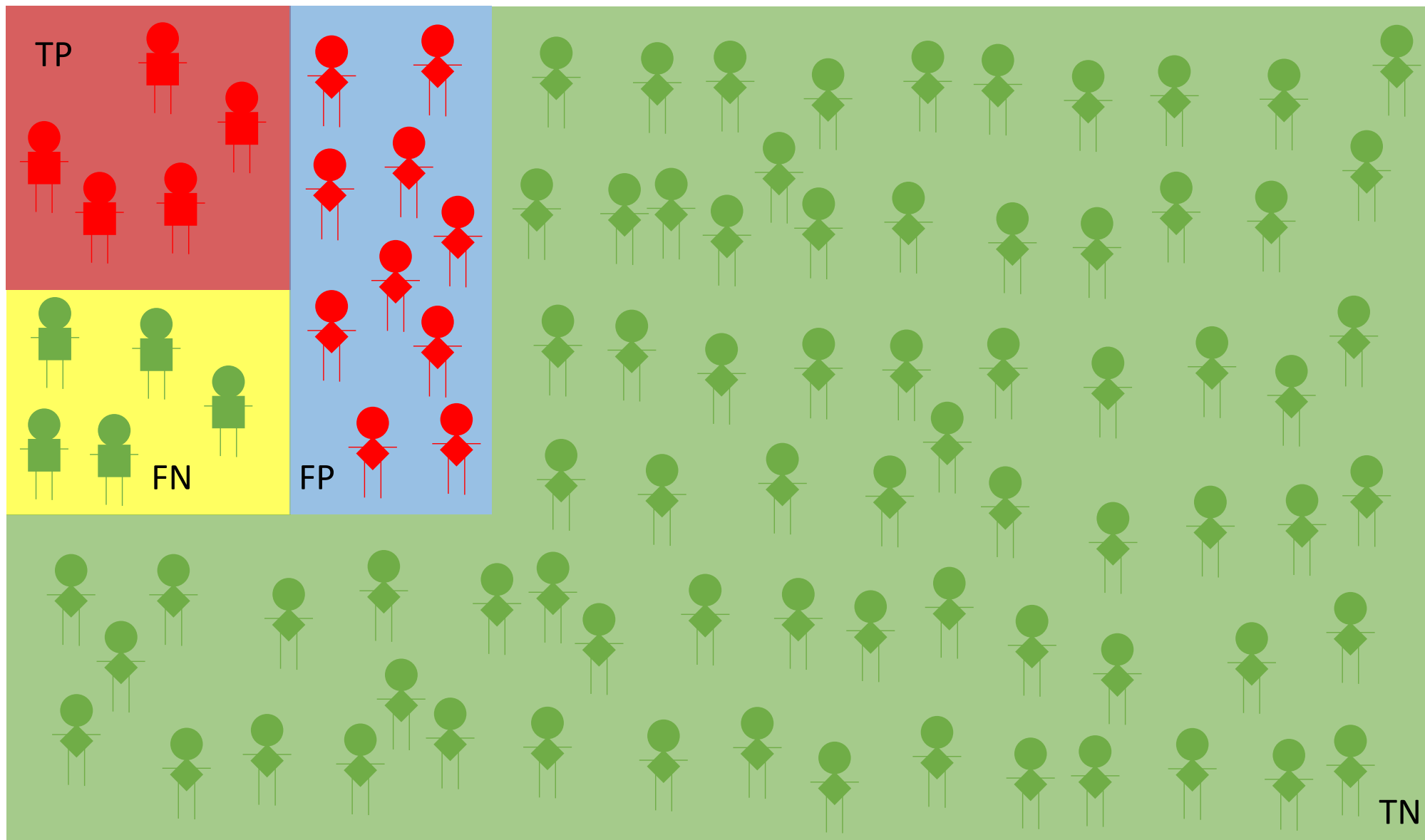
- (26) The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.

- Should medical records be used for medical research?
- Should dead people's records automatically become public?
- Should medical records be given to companies?
- Should medical records be given to Google/Apple?
- What should patients get in exchange?
- ...

ALGORITHM ETHICS

Only considering algorithms that

- Turn data into evidence for a given outcome
- Where this outcome is used to trigger and motivate an action that may not be ethically neutral
- Perform this process in a (semi-)autonomous way



$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Discriminatory Algorithms: GDPR yet again

In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect. Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions.

- List of 350 occupation titles
 - 16 are female-specific (e.g. congresswoman)
 - 20 male-specific (e.g. congressman)
 - rest are gender neutral (e.g. nurse, dancer, bookkeeper)
- Define vectors for female and male, zeros everywhere except for ones at
 - 32 female-specific words (e.g. she, her, woman) in V_f
 - 32 equivalent male-specific words (e.g. he, his, man) in V_m
- Female factor: $\lambda_f(w) = \text{cosine}(V_w; V_f)$
Male factor: $\lambda_m(w) = \text{cosine}(V_w; V_m)$

Where do the biases in algorithms arise?

- Algorithm
- Data

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

68

by James Vincent | @jjvincent | Mar 24, 2016, 6:43am EDT



SHARE



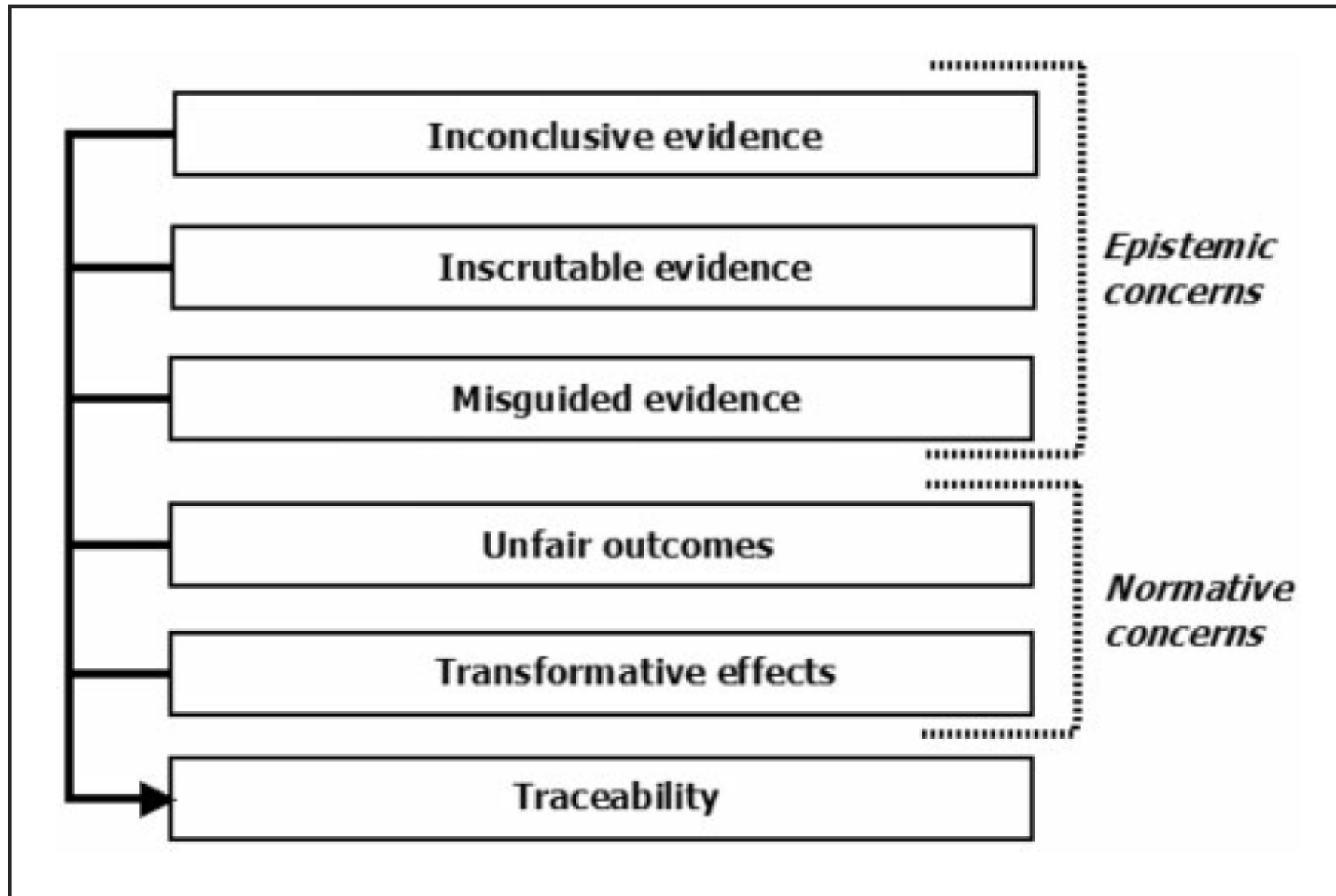
TWEET



LINKEDIN



Six types of ethical concerns raised by algorithms

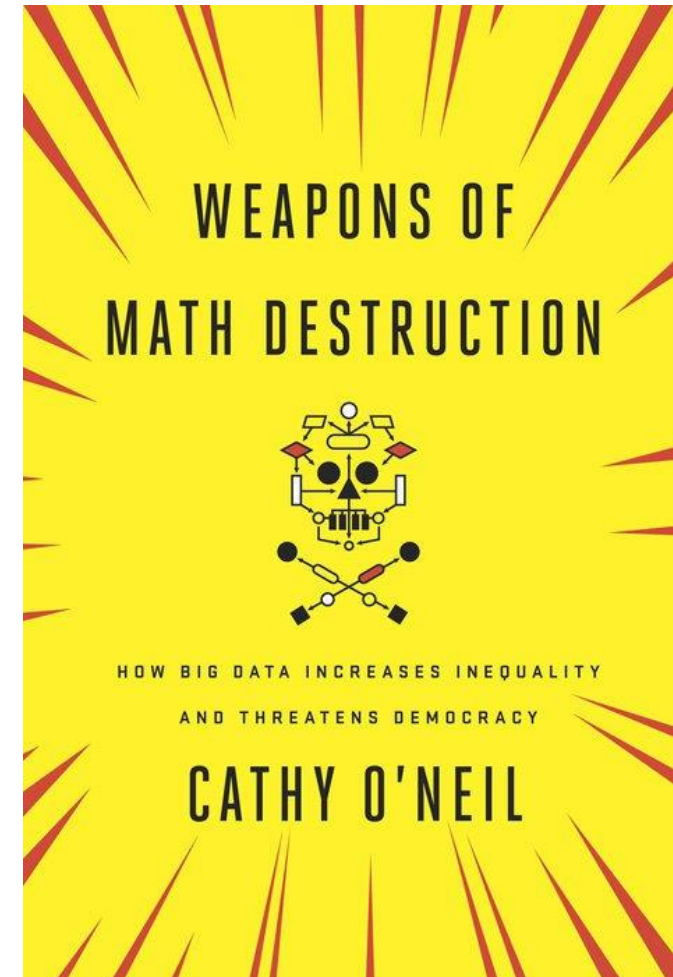


Six types of ethical concerns raised by algorithms

- **Inconclusive evidence**
 - Inferential statistics and/or machine learning leads to uncertain knowledge
- **Inscrutable evidence**
 - Connection between data and conclusion is not obvious/accessible
- **Misguided evidence**
 - Conclusions can only be as reliable (and neutral) as the data they are based on (GIGO)
- **Unfair outcomes**
 - Actions and their effects driven by algorithms are judged to be “unfair” (observer-dependent)
- **Transformative effects**
 - Algorithms can affect how we conceptualise the world, and modify its social and political organisation (e.g. profiling)
- **Traceability**
 - For an identified problem, ethical assessment requires both the cause and responsibility for the harm to be traced

- Justice
- Selection of applicants for a job
- Getting credit
- Getting insurance
- Policing
- Dispute resolution
- Clinical decision support
- ...

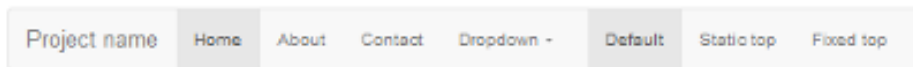
Algorithms in Use



- What are the advantages in algorithms making decisions?
- What are the disadvantages in algorithms making decisions?
- What are the advantages of making algorithms transparent?
- What are the disadvantages of making algorithms transparent?

Who is responsible
when an algorithm
makes an error?





Welcome to our website

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

[Learn more](#)

Click rate: **52 %**



Welcome to our website

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

[Learn more](#)

72 %

"Every data subject should [...] have the right to know and obtain communication [...] (of) the logic involved in any automatic personal data processing"

Article 63

- Data Processing and Profiling
 - Organizations may not use personal data for a purpose other than the original intent without securing additional permission from the consumer
 - Robust anonymization processes must be used where possible
- Right to an Explanation
 - Not yet clear which decisions are subject to this right
 - There are good reasons for data scientists to use interpretable techniques, in particular to avoid bias
 - GDPR should not limit the techniques used to train predictive models
- Bias and Discrimination
 - Ensure fair and transparent processing
 - Use appropriate mathematical and statistical procedures
 - Establish measures to ensure the accuracy of subject data employed in decisions
 - Take into account data with potentially implicit bias, e.g. residential area

7 reasons why Data Science lacks ethics

1. Users, and some operators, give data and analysis an inflated level of objectivity
2. Models hide the truth (SVM, Neural Network)
3. Data hides the truth
4. Data Scientists hide the truth
5. Users hide the truth
6. Models explain how to maintain the *status quo* but don't address the question of whether it should be maintained.
7. Science is the first casualty when running short of time (shortly followed by documentation and testing)

Ten simple rules for responsible big data research

1. Acknowledge that data are people and can do harm
2. Recognize that privacy is more than a binary value
3. Guard against the re-identification of your data
4. Practice ethical data sharing
5. Consider the strengths and limitations of your data; big does not automatically mean better
6. Debate the tough, ethical choices
7. Develop a code of conduct for your organization, research community, or industry
8. Design your data and systems for auditability
9. Engage with the broader consequences of data and analysis practices
10. Know when to break these rules

Conceptual Experiment Design

Experiment Design for Data Science - Block 2
Lecture 3

Peter Knees

peter.knees@tuwien.ac.at

Institut für Information Systems Engineering, TU Wien



- A definition of **science**:
*“knowledge or a system of knowledge covering general truths or the operation of general laws especially as obtained and tested through **scientific method**”*

- Definition of **scientific method**:
“principles and procedures for the systematic pursuit of knowledge involving the recognition and formulation of a problem, the collection of data through observation and experiment, and the formulation and testing of hypotheses”

- Definitions of **experiment**:

*“an operation or procedure carried out under **controlled conditions** in order to discover an unknown effect or law, to **test or establish a hypothesis**, or to illustrate a known law”*



*“a procedure carried out to support, refute, or validate a hypothesis. Experiments provide insight into cause-and-effect by demonstrating **what outcome occurs when a particular factor is manipulated**. Experiments vary greatly in goal and scale, but always rely on **repeatable procedure and logical analysis of the results**.”*



- **Field experiments**

observations in natural settings → possibly more validity;
experimental conditions difficult to control; cf. social sciences

- **Natural experiments** (“quasi experiments”)

mere observation of variables, no controlled manipulation;
collection of evidence to test hypotheses; cf. economics, meteorology

- **Controlled experiments** (“lab conditions”)

Based on manipulation of experimental (independent) variables and control (or measurement) of other factors of experiment;
outcome: dependent variable

→ hypothesis: prediction of effect of independent variable on a dependent variable

Which Experiments in Data Science?

- Q: Which of these experiments are done in Data Science?
- A: **In the end controlled experiments**, but a bit of all...

Situation in practice:

- We often do not collect specific data in order to test a hypothesis, but have to deal with data that happens to be available/automatically generated
- But also partly just controlled experiments on collected, observational data; collecting data to support a hypothesis
- Machine learning experiments → *controlled, repeatable*

- E.g, Image classification
- Classic machine learning setup
 - Data points described by features (e.g. after feature extraction)
 - Target class (or value) for each data point
 - Machine learning algorithm to build a model that can predict target class or value from features (classification, regression, resp.)
- Possible hypotheses
 - Feature set X predicts targets better than feature set Y
 - Algorithm A predicts targets better than algorithm B



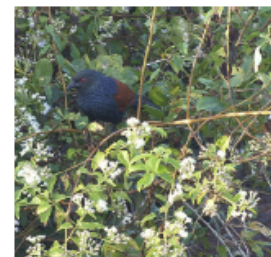
GT: horse cart
 1: horse cart
 2: minibus
 3: oxcart
 4: stretcher
 5: half track



GT: birdhouse
 1: birdhouse
 2: sliding door
 3: window screen
 4: mailbox
 5: pot



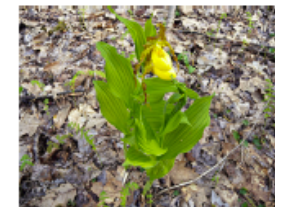
GT: forklift
 1: forklift
 2: garbage truck
 3: tow truck
 4: trailer truck
 5: go-kart



GT: coucal
 1: coucal
 2: indigo bunting
 3: lorikeet
 4: walking stick
 5: custard apple



GT: komondor
 1: komondor
 2: patio
 3: llama
 4: mobile home
 5: Old English sheepdog



GT: yellow lady's slipper
 1: yellow lady's slipper
 2: slug
 3: hen-of-the-woods
 4: stinkhorn
 5: coral fungus



GT: torch
 1: stage
 2: spotlight
 3: torch
 4: microphone
 5: feather boa



GT: banjo
 1: acoustic guitar
 2: shoji
 3: bow tie
 4: cowboy hat
 5: banjo



GT: go-kart
 1: go-kart
 2: crash helmet
 3: racer
 4: sports car
 5: motor scooter

- Hypothesis: **testable (!)** proposed explanation of a phenomenon not yet scientifically satisfactorily explained
 - *“how independent variable(s) affect dependent variable”*
- Needs to meet conditions of cause and effect:
 - presumed cause and presumed effect
 - cause must take place before the effect
 - rule out or take into account other (extraneous) variables
- Control: at least 2 different settings of independent variable to compare

- E.g., testing the effects of a drug
- Hypothesis: **Treatment with drug** **alleviates symptoms**
(=independent var) (=dependent var)
- Control (= different settings of independent var)
 - Group A (treatment group): patients receive drug
 - Group B (control group): patients receive placebo (no drug)

Control group

- accounts for extraneous variables:
effects of procedure, suggestion, expectation, etc.
- allows to calculate effects of the extraneous variables
- allows to remove these effects from the treatment effect

- Testing: measuring (positive) effect on patients' symptoms
...comparing outcomes: $\text{health}(A) > \text{health}(B)$?

- E.g., comparing two retrieval systems (search engines)
- Hypothesis: **system X outperforms system Y**
(e.g. Google) (e.g. Bing)
- Independent var: system
- Dependent var: performance indicator
- Control: system X vs. system Y
- Testing:
 - System X retrieves more relevant documents than system Y for the same query (or set of queries)
 - $\text{performance indicator}(X) > \text{performance indicator}(Y)$

And another, concrete example...

A set of movie ratings given as $\langle \text{user_id}, \text{item_id}, \text{rating} \rangle$ is sampled from MovieLens (movie recommender website)

- Hypothesis:
SVD algorithm better to predict ratings than User KNN algorithm
- Independent var: prediction algorithm
- Dependent var: error of prediction
- Control:
 - Setting A: SVD
 - Setting B: User KNN
 - every other aspect must be identical when running the experiment!
- Performance criteria: RMSE on test data
- Testing: under the assumption that the set is representative for the task, train an SVD model using data, train a User KNN model using the exact same data and measure prediction error of each on the same testing data. If the error of SVD is smaller than that of User KNN, the hypothesis is confirmed.

- Controlled variables in ML:
 - Model: k-NN, decision tree, SVM, neural network, ...
 - Algorithm: optimization criteria, implementation, parallelization, ...
 - Parameters: model parameters, learning rate, initialization, ...
 - Selected features
 - Training data
 - Runtime environment: architecture, OS, number format, ...
- Dependent variable(s):
 - System performance
 - Expressed as evaluation criteria: accuracy, precision, recall, F1, AUC, error, RMSE, etc.
 - Which one to choose needs to be justified by data scientist (does the number really measure what we want to test?)

BASICS IN ML (WITH OLD-SCHOOL EXAMPLES)

cf. Tom M. Mitchell, Machine Learning, McGraw Hill, 1997.

- **Nominal**

Labeling variables, belonging to different classes, categorical, no hierarchy; examples: gender (M/F/X), country

- **Ordinal**

Order of values is known, but no numerical difference; example: very bad < bad < neutral < good < very good

- **Interval**

Numeric scale, difference of values known, but no reference point! (zero point); example: time (consistent, measurable increments)

- **Ratio**

True zero reference → meaningful multiplication and division, statistical moments are measurable; examples: height, weight

- Different scales require different strategies for comparison!

- D : sample drawn from data

target value

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0	1
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0	0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0	0
40,M,205,0,115,90,37,18,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1	1

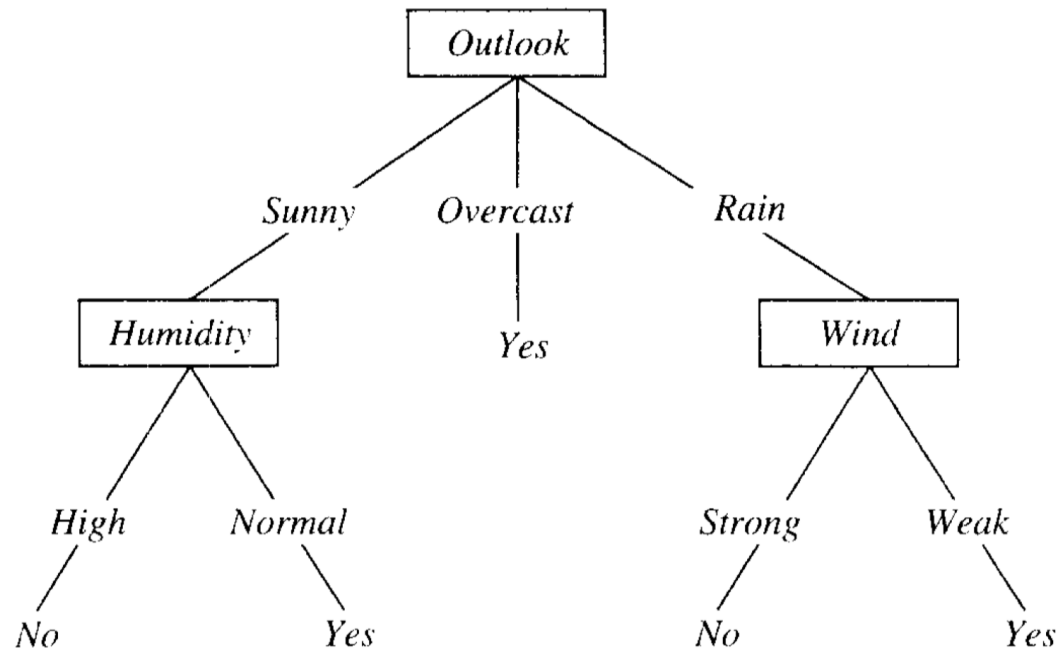
- Supervised Learning ... learn a model that predicts the correct target value/attribute given other attributes
- $y=F(x)$: true function (usually not known)
- $G(x)$: model learned from sample $D \sim F(x)$
- Goal: $E<(F(x)-G(x))^2>$ is small (near zero) for future samples drawn from $F(x)$

Supervised Learning Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

How to predict whether Tennis can be played based on other attributes?
 → Learning a classifier: train model from given data

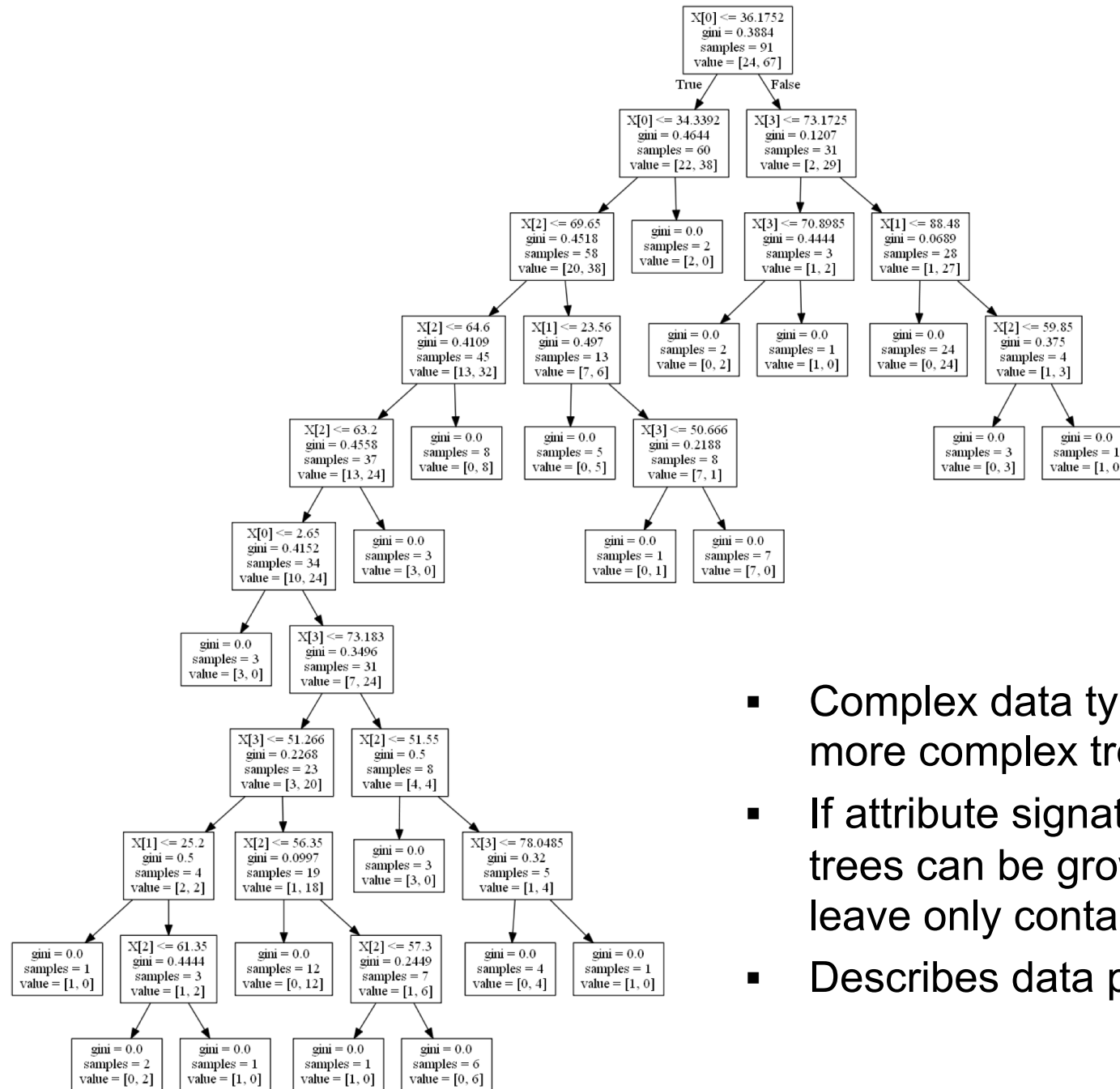
A Simple Decision Tree



Day	Outlook	Temp	Humid	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

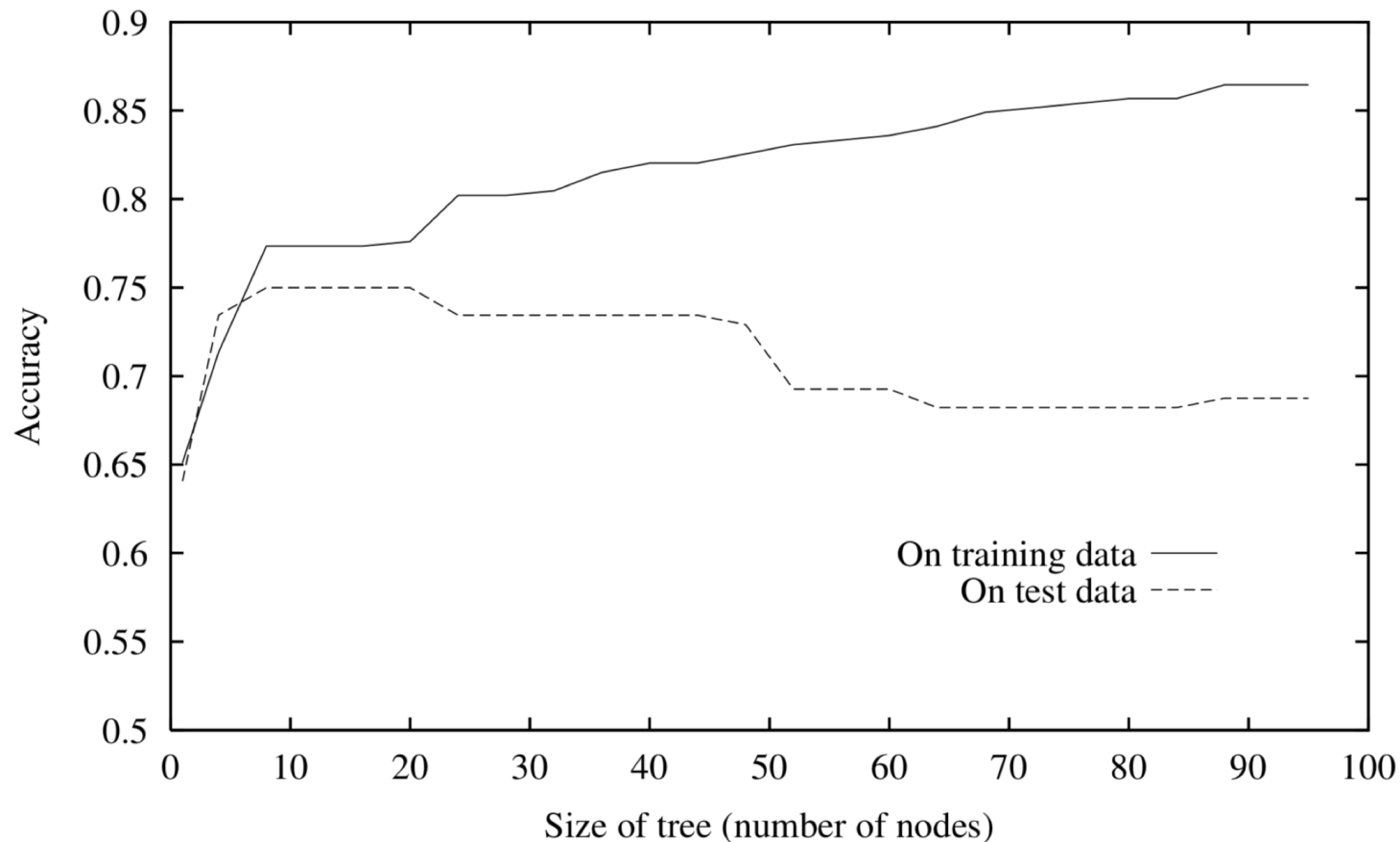
- Choose splitting attributes to maximize Information Gain wrt. target class (= reducing entropy)
- In this example, all data points are perfectly classifiable and classified

Slightly more realistic example of Decision Tree



- Complex data typically results in more complex trees
- If attribute signatures are unique, trees can be grown such that every leaf only contains one class only
- Describes data perfectly

Overfitting in Decision Tree Learning



- For complex data, growing the tree describes the data better
- However, on **new, unseen data**, performance goes down
- **Overfitting**: model “corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably” *[Wikipedia]*

- Goal: learned model must **generalize** to also fit previously unseen data
- Performance on training data gives no information on this
- Need to simulate a more realistic scenario and find model that performs well on unseen data
- Idea: hold back some of the training data from training and use to test performance
- Data used for training and for testing should resemble each other (similar properties, same distribution)
- No data used for testing (or information extracted from it) must ever be used in the training of a model!

EXPERIMENTAL DESIGNS

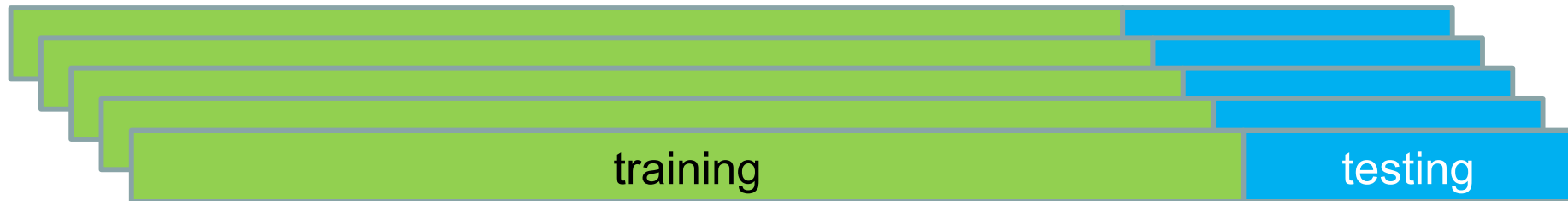
- To test model on new data, draw a **random sample**

1. Random shuffle of the data!
2. Partition data into part for training and part for estimating performance (testing)



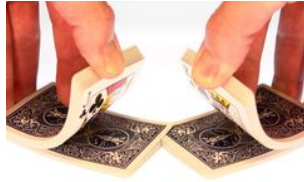
3. Calculate success criteria (performance metric) on testing set
- Issues: small number of testing instances and by chance, we might not evaluate important instances (=bias)
 - (we need as much data as possible for training...)

- Repeat that process n times (always shuffle anew!)



- Results in n models and n performance scores
- Aggregate scores (e.g. mean)
- Actually sample of scores from underlying distribution
- Compare models (parameters) via aggregated scores and chose the best one (a final model can be trained with the best settings using all the data)
- Issue: we might happen to favor some instances for testing performance (might appear in test sets more often)

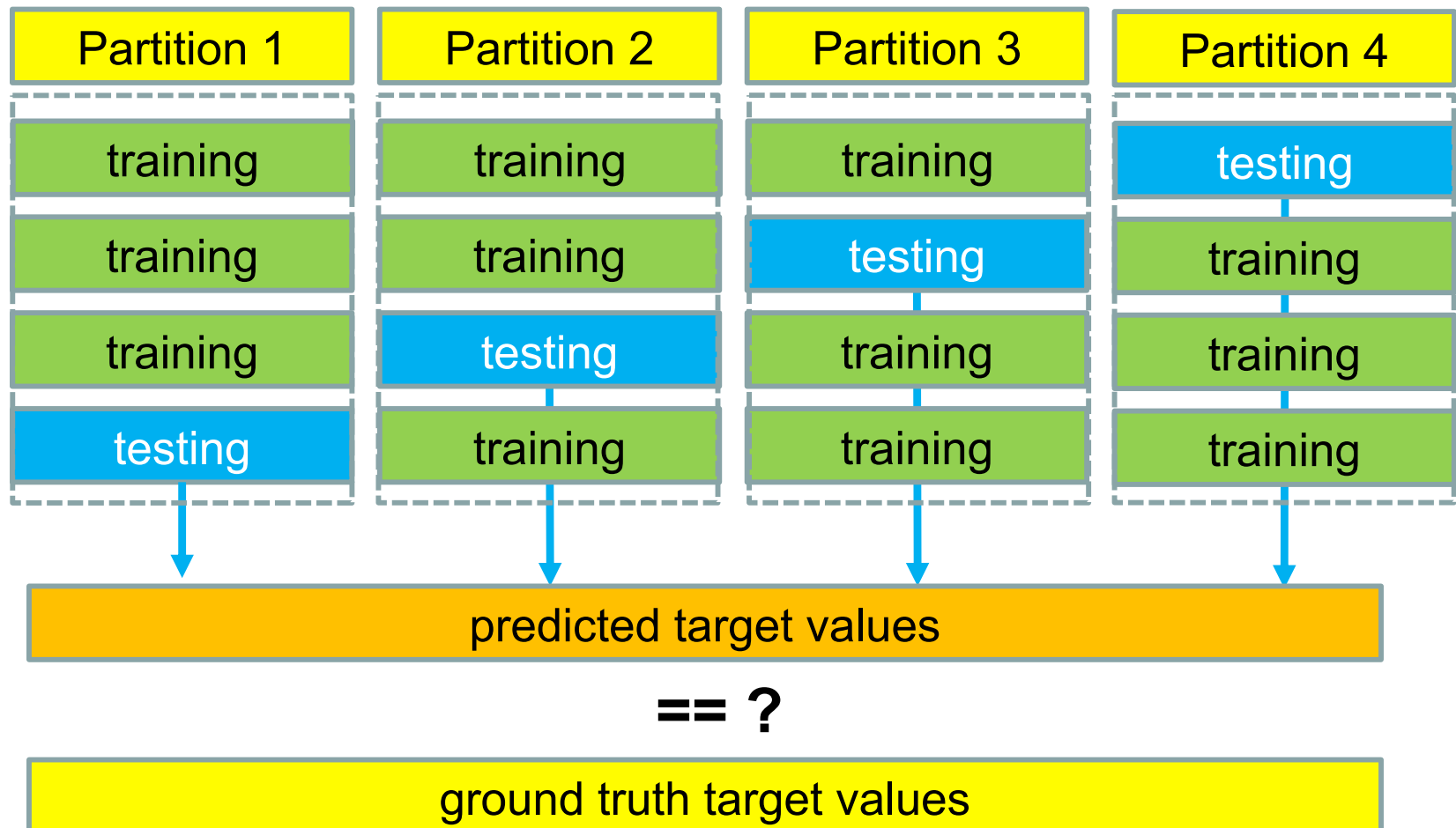
- To make use of all instances for testing and give equal impact of instances on performance measure
- **k -fold cross validation**
 1. Shuffle data (!)
 2. Partition data into k (near-)equally sized subsets
 3. Train k models such that
 - $k-1$ subsets are used for training
 - The remaining subset is used for testing
 - No two models are tested on the same subset
 4. Prediction available on all subsets (thus each instance used exactly once for testing)
 5. Calculate performance over full predicted set
 6. Optionally: repeat process multiple times



shuffled instances

split into e.g. $k=4$ (almost) equally sized subsets/partitions

- fold 1
- fold 2
- fold 3
- fold 4

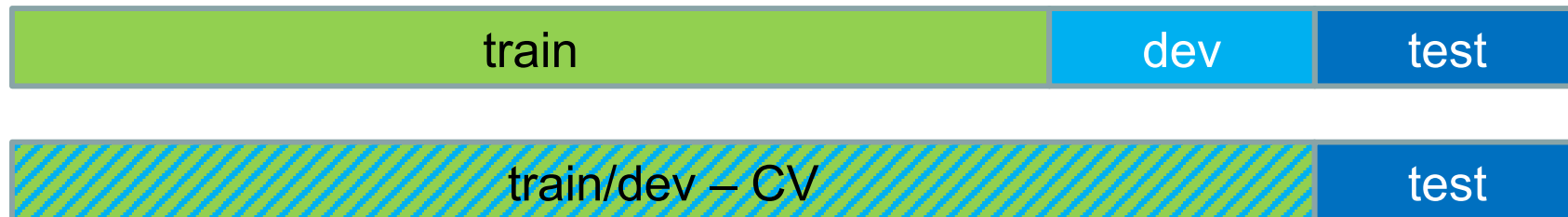


- Typical range for k : $[2, 10]$
- Special variant: leave-one-out CV
 - $k = \#instances$
 - Closest simulation (models very similar to final model)
 - Typically too expensive, not seen very often anymore
- Variant: Stratified k -fold cross validation
 - each partition should resemble same distribution of target classes than overall class distribution
 - Only possible if $k \leq \#instances \text{ of least frequent class}$

Common mistake made with CV:

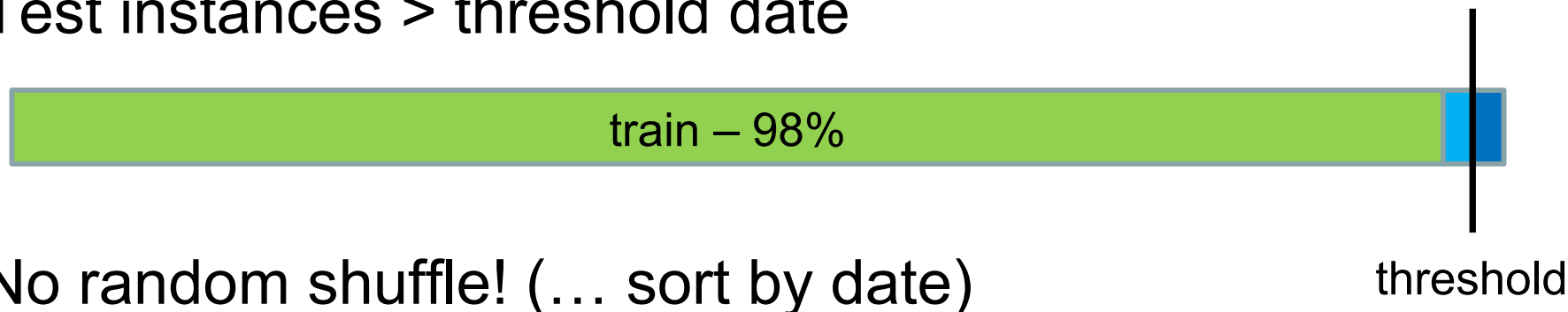
- Estimation of parameters, feature selection, dimensionality reduction, normalization etc. performed prior to CV
- → information of data used for testing leaks into training!
- **All these steps need to be carried out for each fold individually!**
- Might make CV very expensive to calculate
- Also: same splits should be used when comparing models/settings!

- Repeated splitting of same data for optimization might lead again to overfitting
- Final estimate of real-world performance should be made on another, independent test set
- Again, reserve part of the data with relevant properties for later testing, e.g.,



- NB: name of set for param. optimization (development set) not consistent in literature; e.g., often called **validation set**
- Moreover, inconsistent naming of what is validation and what is test set..
→ Make sure to always describe what it is used for to avoid confusion!

- Depending on the scenario to evaluate, other experiment setups might be more relevant (bias desired)
- Consider a recommender system:
 - Goal: predict future behavior from collected data
 - Simulation: predict interactions after a chosen point in time by learning from interactions before
 - Potentially for each user individually
- **Time-based split**
- Training instances (+dev?) \leq threshold date
Test instances $>$ threshold date



- No random shuffle! (... sort by date)

MEASURING PERFORMANCE

- Goal of experiments/success **needs to be quantifiable**
- **Operationalization**: the process of strictly defining variables into measurable factors
- The better the data is understood and the clearer the goal can be phrased (operationalized), the more effective optimization will be
- **Goal and relevant performance measure need to be defined clearly before starting experimentation**
(and defining the experimental setup...)

- We assess performance by **comparing predictions made by a model with the actual ground truth**
- Depending on task and goal (hypothesis), different performance criteria are relevant
 - Spam filtering: does not delete non-spam messages
 - Patent search: finds all relevant patents
- Not all criteria are directly expressible in a performance measure
 - Product recommendation: users are satisfied (?)
- Find a suitable approximation, ideally ruling out all other influences/variables (such as, e.g., user context)

- For regression tasks (prediction of numerical value), the residual between true value y_i and prediction \hat{y}_i is a typical performance measure, e.g.,

- Mean absolute error:
$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$

- Root mean squared error:
$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

→ large errors are disproportionally penalized by squaring the difference

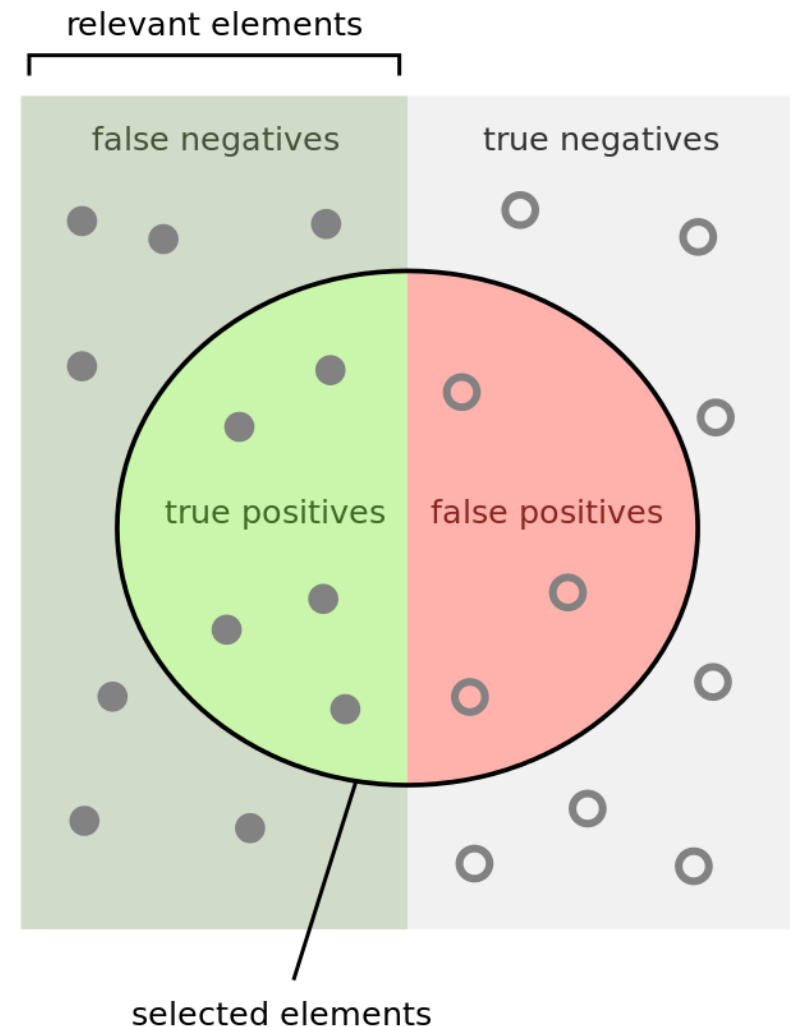
(n ... number of test instances)

- Quality of classifier: how well can the true labels be predicted?

- Accuracy:** percentage of correctly predicted instances

- $$Acc = \frac{TP+TN}{TP+FP+FN+TN}$$

	Classified positive	Classified negative
Actual positive	TP	FN
Actual negative	FP	TN



Source: [Wikipedia]

Classification Performance

- Example:

<i>Classifier 1</i>	Classified positive	Classified negative
Actual positive	10	15
Actual negative	25	100

$$Acc = \frac{10+100}{10+25+15+100} = 73.3\%$$

→ looks pretty ok!

- Now: a really stupid classifier (“computer says no”)

<i>Naysayer</i>	Classified positive	Classified negative
Actual positive	0	25
Actual negative	0	125

$$Acc = \frac{0+125}{10+25+15+100} = 83.3\%$$

→ looks even better!

- Class distribution can not be ignored
- Performance measures should always be reported together with a **baseline** reference score
 - e.g., “intelligent guessing” (predict always the most frequent class)
 - e.g., same algorithm without special optimizations (control!)
 - e.g., the currently best performing algorithm (state of the art)
- Accuracy alone is not often a good performance indicator

Gain more understanding of what classifier does

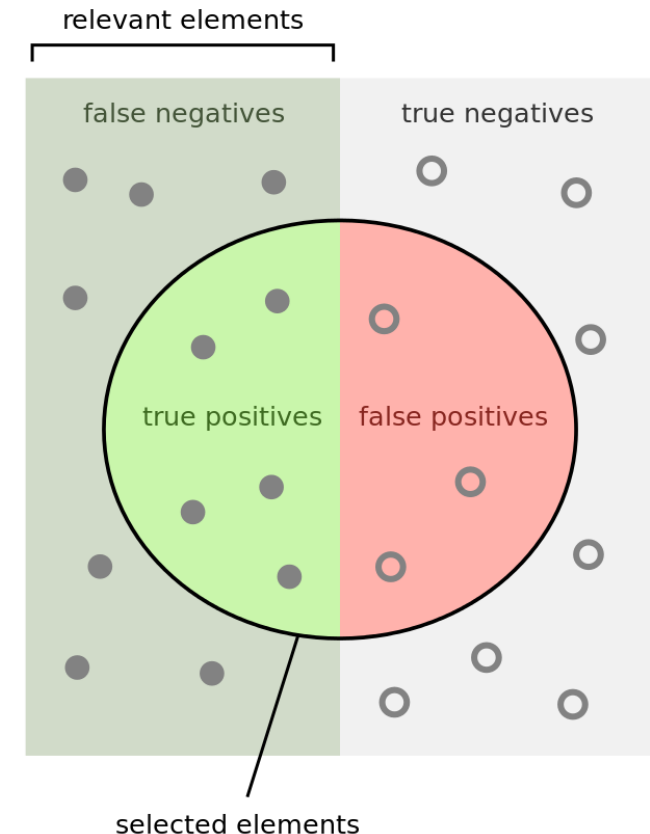
- **Precision:** how many of those predicted as class x are actually correct?

$$Prec = \frac{TP}{TP + FP}$$

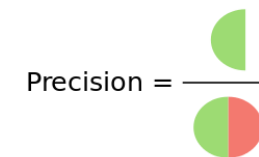
- **Recall:** how many of the instances of class x were actually predicted as such?

$$Rec = \frac{TP}{TP + FN}$$

- Precision and recall can be calculated for each class
- Based on parameter tuning, precision can be sacrificed in favor of recall and vice versa (cf. “always-no” example)

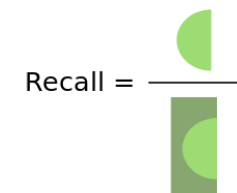


How many selected items are relevant?



Precision =

How many relevant items are selected?



Recall =

Source: [Wikipedia]

→ Which performance metrics would we look to optimize for
a) a spam detector and b) a patent search system?

- If both precision and recall are important, what is the optimization objective?
- Combined measure to balance precision and recall (and punish low values of either)
- **F-measure**/F1-score: harmonic mean of precision and recall

$$F = 2 \cdot \frac{Prec \cdot Rec}{Prec + Rec}$$

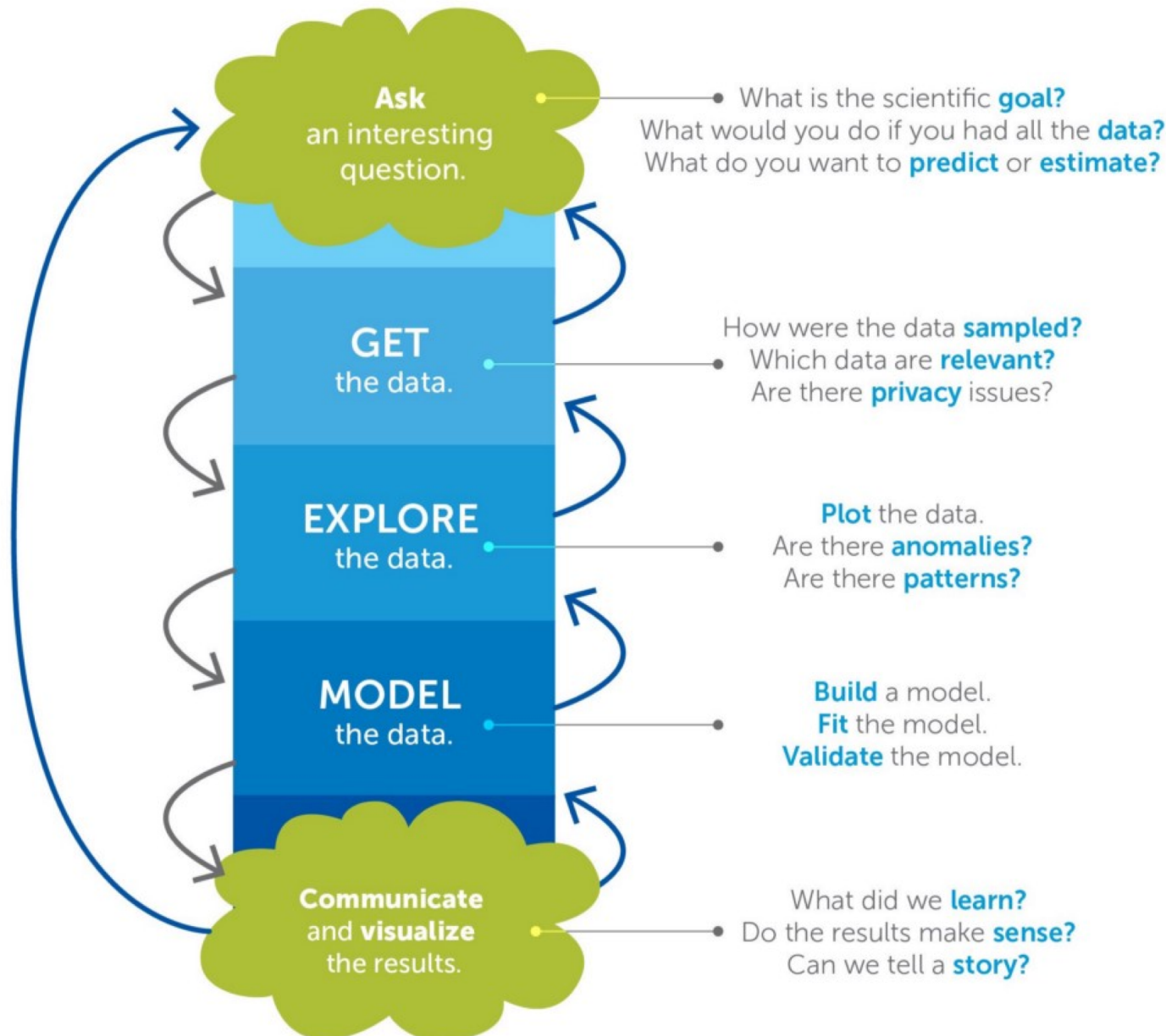
- General F_β -measure can be tuned to favor Prec or Rec
- Precision, recall, F-measure and variants thereof often used as criteria in information retrieval

- The model (parameter combination) yielding the best performance according to the chosen criteria is used
- In practice, combinations of variations over several parameter ranges are explored automatically (**grid search**)
 - Brute force approach
 - More efficient heuristics exist, trying to find minima in the parameter space
- In order to compare two models, we should not look at just one number (e.g. a mean value) but compare range of outputs (e.g., variance over several runs, cf. later)
- Calls for **statistical significance testing** (t-test, U test, etc.)

Classifier	Accuracy	Runtime
A	90%	80ms
B	92%	95ms
C	95%	1,500ms

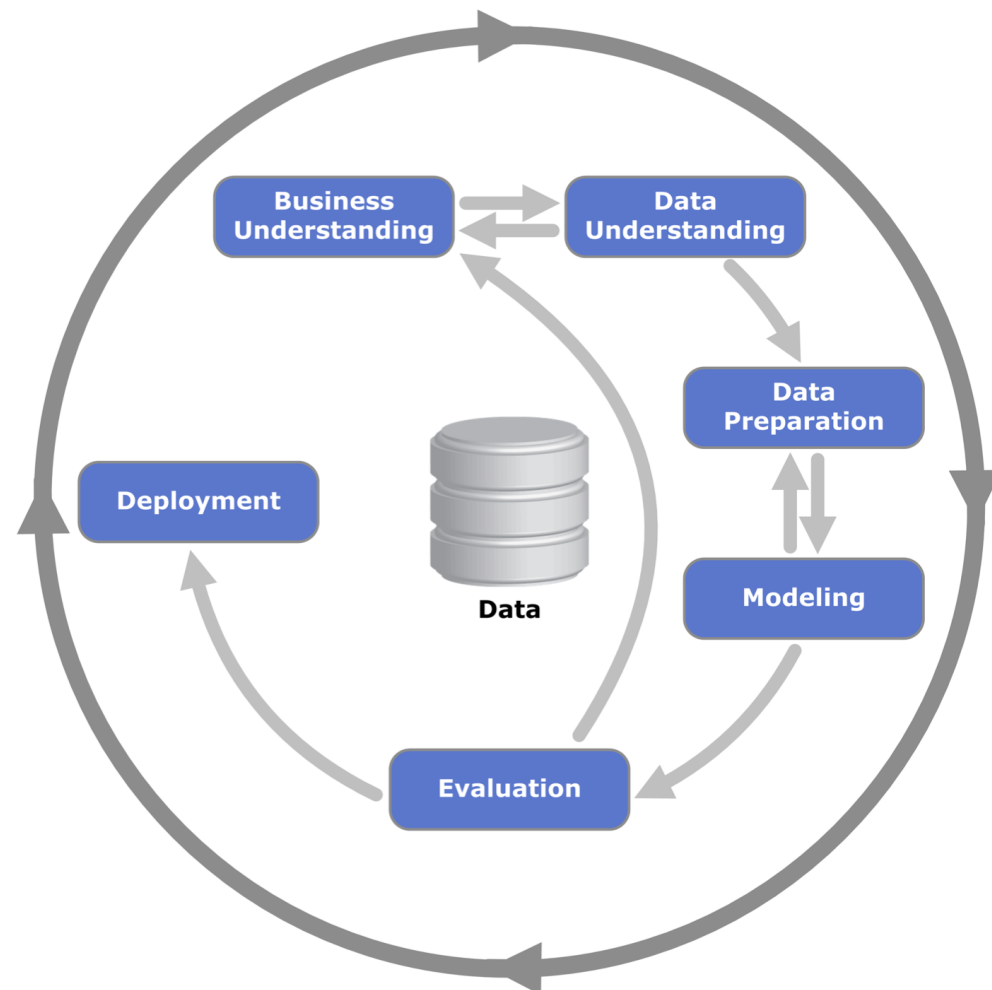
- Optimize for accuracy as primary metric in this example
- However, other factors might be relevant, e.g. runtime (classifier C performs best but takes long)
- One strategy: $\text{value} = \text{accuracy} - 0.5 \times \text{runtime}$
- Alternative: maximize accuracy subject to $\text{runtime} \leq 100\text{ms}$
 - Accuracy: **optimizing metric**
 - Runtime: **satisficing metric**

- Data Science is an **empirical science**
 - Investigate data transformation processes using scientific methods
 - Methods originally applied to natural objects (fundamental particles, chemicals, living organisms) or individuals and social groups
- Strategies from Data Mining and Machine Learning experimentation
 - Definition of target criteria measuring success
 - Preparation/selection of training, development, and test sets
- Iterative process
 1. Construct hypotheses/build (approximate) theories
 2. Test with empirical experiments
 3. Refine hypotheses and modelling assumptions



- Build hypothesis, define target metric
- Check data, understand origin and distribution, prepare for experiments
- Learn model, optimize model
- Interpret results
- Try again...

- Cross-industry standard process for data mining (CRISP-DM) from 1996
- Business-oriented, iterative process developed to organize data mining
- 6 phases:
 1. Business understanding
 2. Data understanding
 3. Data preparation
 4. Modeling
 5. Evaluation
 6. Deployment



1. Business understanding

assessing the situation (business requirements, risks, cost, etc.), determining data mining goals, producing project plan (cf. “hypothesis building”)

2. Data understanding

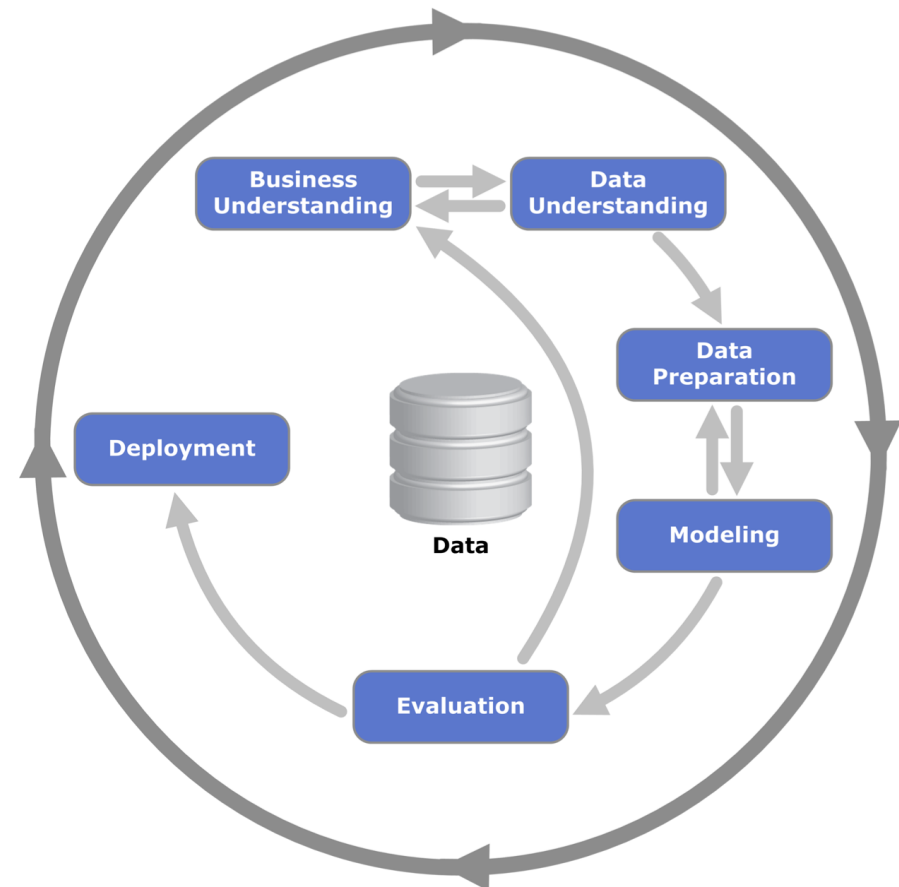
collecting, describing, exploring, verifying data

3. Data preparation

selecting, cleaning, constructing data

4. Modeling

selecting model, generating test design, building + assessing model

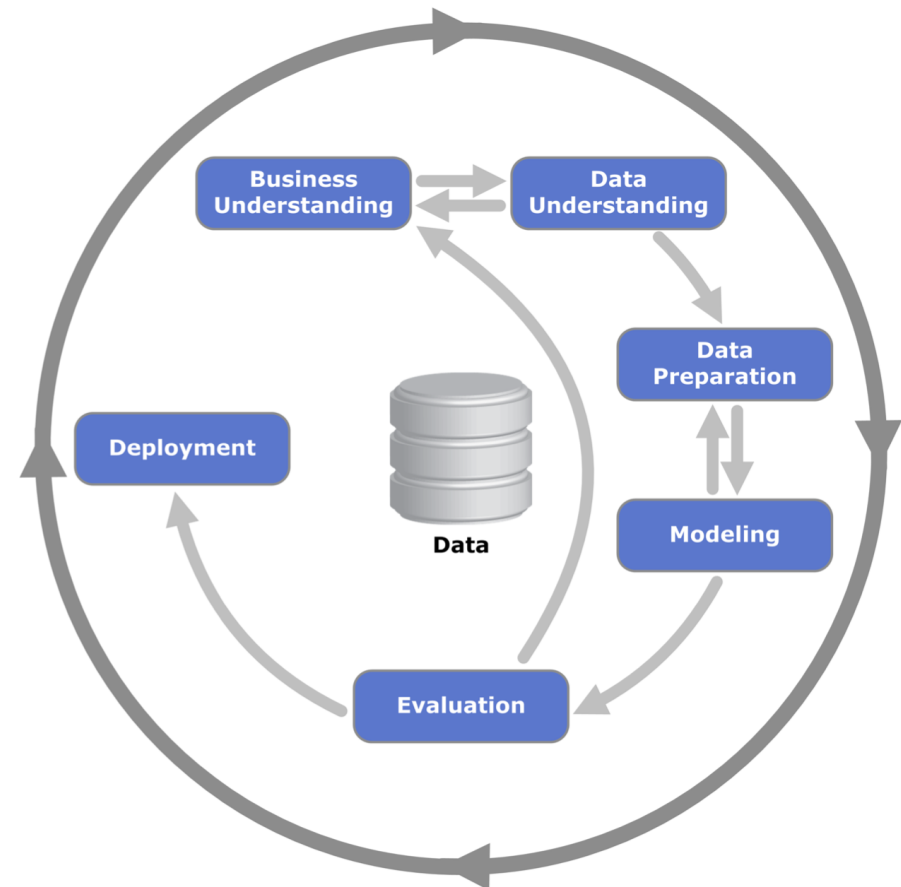


5. Evaluation

evaluating results, reviewing process, determining next steps

6. Deployment

planning deployment, monitoring, maintenance, reporting



■ References:

- Colin Shearer, *The CRISP-DM model: the new blueprint for data mining*, Journal of Data Warehousing, 5(4), pp.13–22, 2000
- IBM SPSS Modeler CRISP-DM Guide:
<ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf>

- **Pilot experiments** are important!
- Not necessarily designed to test the hypothesis but to test the experimental apparatus
- i.e., check whether pipeline actually can test the hypothesis
- In pilot experiments:
 - Provide preliminary data
 - Check if protocol works
 - Check for plausible results and outcomes
 - Try out statistical analysis
 - Find bugs!

Scientific Workflow Environments

Experiment Design for Data Science - Block 2
Lecture 4

Peter Knees

peter.knees@tuwien.ac.at

Alexander Schindler

schindler@ifs.tuwien.ac.at

Institut für Information Systems Engineering, TU Wien

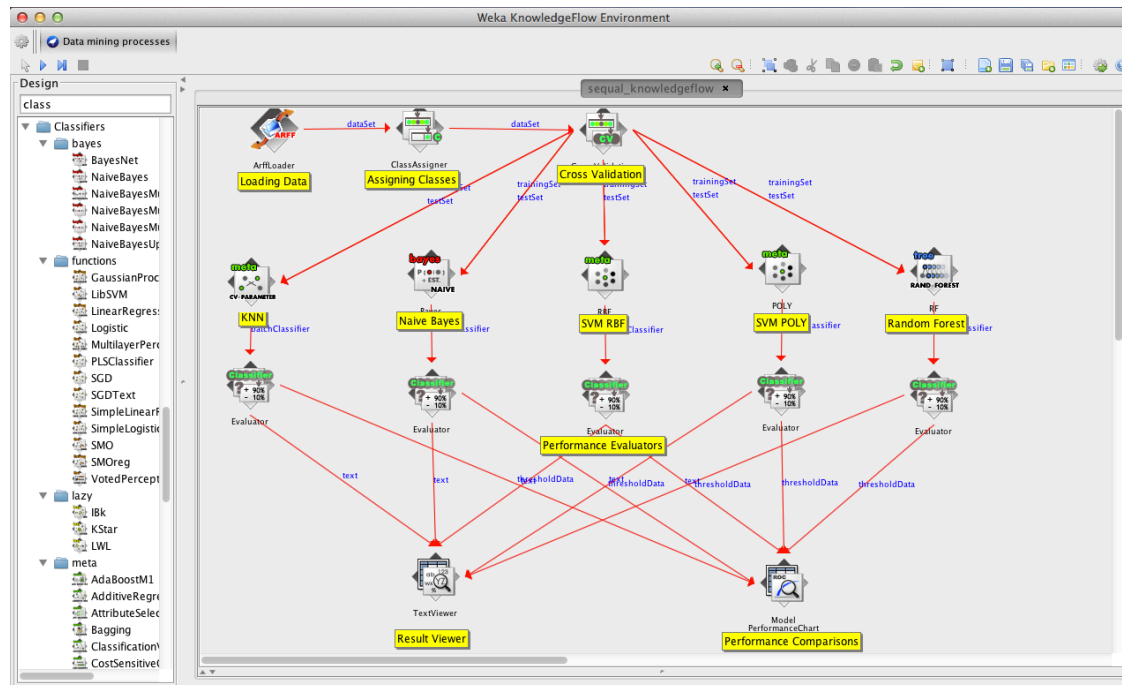
Tools for Scientific Experimentation

- Machine Learning and Workflow Environments:
WEKA + WEKA KnowledgeFlow, RapidMiner, Orange, Taverna, Kepler + Outlook: provenance
- Code/Script Environments:
IPython/Jupyter/Colaboratory, R, MATLAB, Julia

- Experimentation with a focus on different machine learning algorithms and parameters
- **WEKA**: Waikato Environment for Knowledge Analysis
- Reference implementations of a variety of algorithms for preprocessing, clustering, classification, regression, feature selection, visualization etc.
- Extensions through package manager
- Interfaces for systematic comparison, parameter space search
- Java-based, GNU licensed
- <https://www.cs.waikato.ac.nz/ml/weka/index.html>

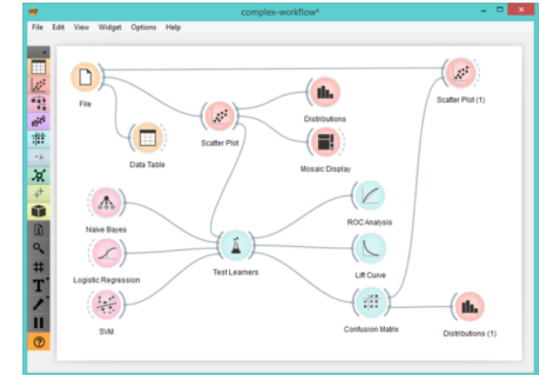


- Tools with graphical representation of workflow
- Individual steps encapsulated and structured
- WEKA KnowledgeFlow:
graphical representation of WEKA workflows



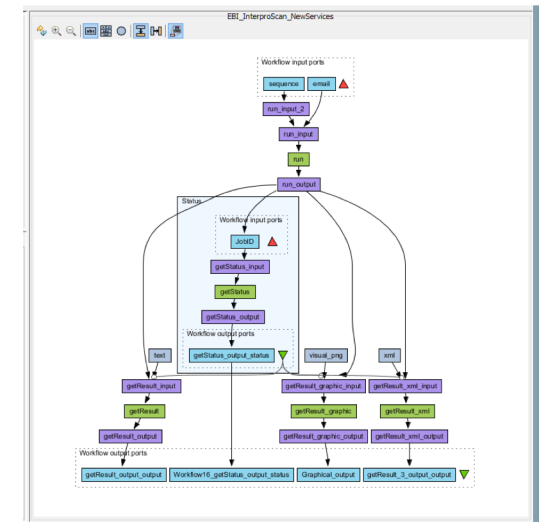
Machine Learning focus

- RapidMiner: commercial, integrates WEKA
- Orange: Python, C++ based
- Microsoft Azure, AWS: commercial, cloud-based



Scientific Data Workflow Management

- Kepler: <https://kepler-project.org>
- Taverna: <http://www.taverna.org.uk/> (now Apache)
- designing, executing, reusing, evolving, archiving, and sharing open source workflows
- focus on integration of data repositories, often from life sciences, medicine, astronomy
- often include capabilities to track data as it is being processed (provenance)



- Historical record of data and its origins
- Inputs + influencing entities, systems, and processes
- Benefits of capturing and sharing provenance information:
 - Provision of detailed account of how results were derived given inputs → intermediate results, workflow steps, parameter settings
 - Facilitates transparency and reproducibility of workflows
 - Internal use for scientists
e.g., to trace sources of errors and debug workflows
- Feature of scientific workflow managements systems
- For scripting languages:
 - noWorkflow (not only workflow): Python runtime profiling, generate provenance traces of script processing history
 - YesWorkflow: using annotations to make data flow dependencies explicit and visualize in graph form
 - RDataTracker: provenance library for R

- Exploratory and fast prototyping
- Interpreter languages and data-oriented programming, e.g., Python, R, MATLAB
- Often interactive, web-based environments, e.g., Jupyter Notebooks, Colaboratory, R Markdown Notebooks
- Combines code and output

IP[y]: IPython
Interactive Computing

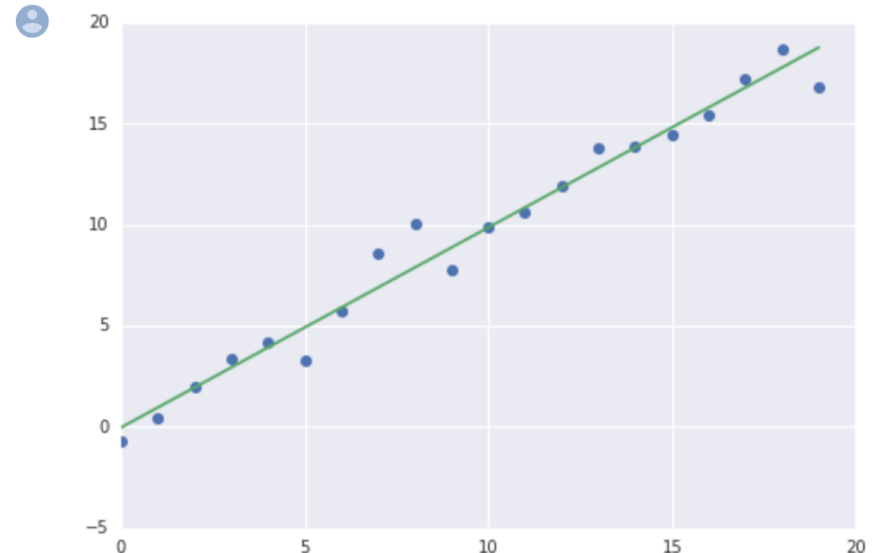


jupyter



```
[ ] import matplotlib.pyplot as plt
import numpy as np

x = np.arange(20)
y = map(lambda x: x + np.random.randn(1), x)
a, b = np.polyfit(x, y, 1)
plt.plot(x, y, 'o', np.arange(20), a*np.arange(20)+b, '-');
```



IPython Demo

- Kaggle [House Prices](#) challenge
- Data exploration part from akquinet blog
 - <https://blog.akquinet.de/2017/09/19/predicting-house-prices-on-kaggle-part-i/>
 - <https://blog.akquinet.de/2017/10/25/predicting-house-prices-on-kaggle-a-gentle-introduction-to-data-science-part-ii/>
- Experimental design for prediction

Further Resources

- Installing [Jupyter Notebook](#) (for convenience: [Anaconda](#))
- [Python and Numpy Basic Introduction](#)
- [Running the Jupyter Notebook](#)
- [Notebook Basics](#)

Experiment Error Analysis and Statistical Testing

Experiment Design for Data Science - Block 2
Lecture 5

Peter Knees

peter.knees@tuwien.ac.at

Institut für Information Systems Engineering, TU Wien

- Hypothesis: prediction of effect of independent variable on a dependent variable
- e.g., machine learning algorithm X yields better results in terms of F-measure than machine learning algorithm Y for classifying images
- Independent var: machine learning algorithm
- Dependent var: performance indicator F-measure
- Control: varying independent var (X vs. Y)
- Testing: $F(X) > F(Y)$?

- In factorial experiments, examine every combination of factors (independent variables, typically 2-3)
- Vary and sample factor x at all levels:
 - x ... categorical variable: categories or combined/higher categories
→ sampled levels of x
 - x ... interval/ratio variable: > 2 levels of x
 - x ... ordinal variable: > 2 levels of x
- Also, when already running factorial experiments (typically expensive), measure effect on more than one dependent var. (e.g., prec., rec., F)

- Testing hypotheses by translating them into statistics about data
- Statistics: observations of random variables from known distributions
- Statistical inference: process of drawing a conclusion about unseen *population* from a *sample* (relatively small)
- Classic setup:
 - H_0 ... **null hypothesis**: default position
 - H_1 ... **alternative hypothesis**: differs from default
- Using statistics to decide whether we can reject H_0 or not

- System to extract facts from news stories for summarization

Table 4.1 Recall scores in five trials of between eight and ten news stories.

Story number:	1	2	3	4	5	6	7	8	9	10	\bar{x}
Day 1	51	63	59	60	62	63	60	62	60	54	59.4
Day 2	49	53	54	64	66	42	45	69	61	50	55.3
Day 3	55	57	54	65	68	51	49	61			57.5
Day 4	52	61	63	49	44	56	65	63	42		55.0
Day 5	66	61	58	51	46	61	42	55	57		55.2

- Recall sample mean for day 1

$$\bar{x}_1 = \frac{51 + 63 + 59 + 60 + 62 + 63 + 60 + 62 + 60 + 54}{10} = 59.4.$$

- Updated system (“better recognition”?) applied on day 121
- Recall sample mean for day 121

$$\bar{x}_{new} = \frac{63 + 44 + 61 + 72 + 74 + 47 + 72 + 56 + 68 + 71}{10} = 62.8.$$

- Higher score than old system, but is it really an improved system? Or did it just happen to operate on simpler stories?
- → statistical hypothesis testing

1. Hypotheses:

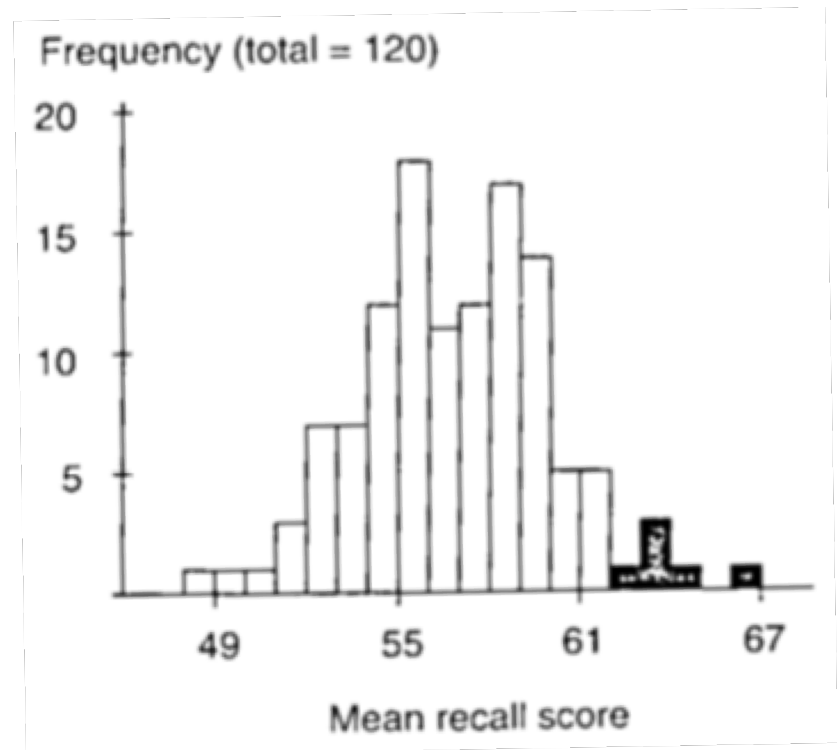
- H_0 ... systems are equal: no difference in mean recall performance
- H_1 ... updated system is more effective: difference exists

1. Hypotheses:
 - H_0 ... systems are equal: no difference in mean recall performance
 - H_1 ... updated system is more effective: difference exists
2. Determine probability of obtaining a sample mean of 62.8 given H_0
3. \rightarrow if very unlikely, H_0 probably wrong
4. We can
 - a. reject H_0 with some confidence (H_1 still might be true) or
 - b. maintain belief in H_0 (new sample mean is just very improbable)

NB: strategy similar to proof by contradiction:

1. negate proposition ($\cong H_0$) — 2. show contradiction (\cong low probability) \rightarrow prove proposition (\cong bound probability you are wrong)

- What's the probability of a sample mean of 62.8 given H_0 ?
- Let's look at the sample means of the first 120 days (=old system)
- On 6 out of 120 days, mean recall > 62
(\rightarrow *empirical sampling distribution*)
- Under H_0 , probability of achieving score > 62 is $6/120 = 0.05$
- Hence, when observing a score of 62.8 and therefore rejecting H_0 , there is a 5% chance this is wrong ("rejecting the null at the 0.05 level")

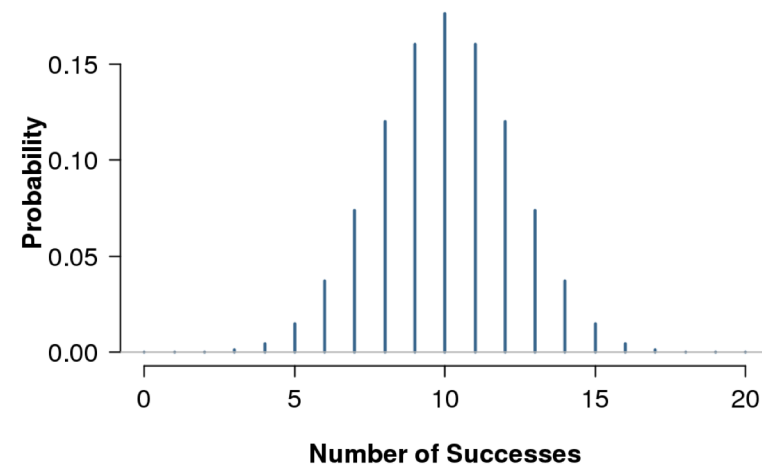


- ! Statistical hypothesis testing **does not prove the *null* false**; it bounds the probability of incorrectly asserting (based on some observation) that the null is false
- If H_0 is rejected, H_1 may be accepted
- **Type I Error**: falsely reject H_0 ; conclude that the observed differences are significant although they are not (“false positive”)
- **Type II Error**: accept H_0 when in fact it is false, i.e. not detecting significant performance differences (“false negative”)
- Before testing, define at which level (=for which extreme outliers) we consider an observation to be very improbable:
Level of significance α (common values: $\alpha = .05$ or $\alpha = .01$)
- → Probability of Type I Error: level of significance α
- → Probability of Type II Error: ? (...depends on power)

- In the previous example, we used the empirical sampling distribution
- Typically, distributions are *calculated exactly* or *estimated analytically*
- → **Parametric statistics**
based on assumptions about the probability distributions of the variables being assessed; parameters of model either known or estimated
- Alternatively: **Nonparametric statistics**
not based on parameterized families of probability distributions; parameters determined by data
examples: order statistics (rank-based)

Example with Exact Sampling Distribution

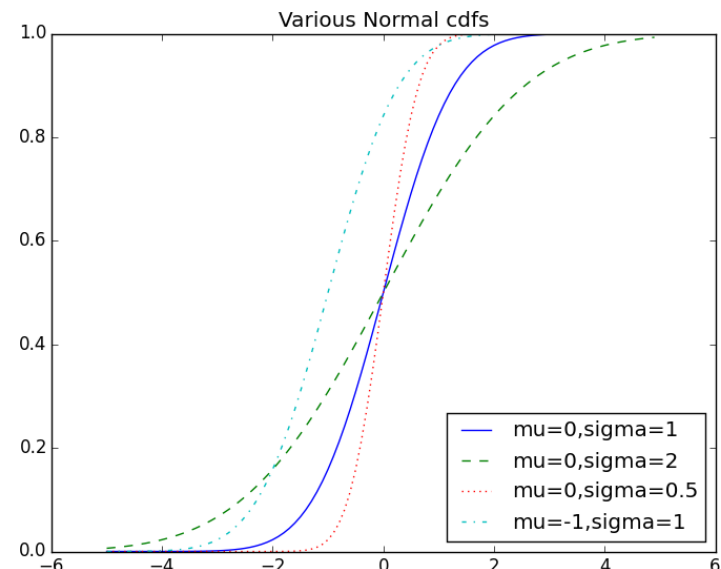
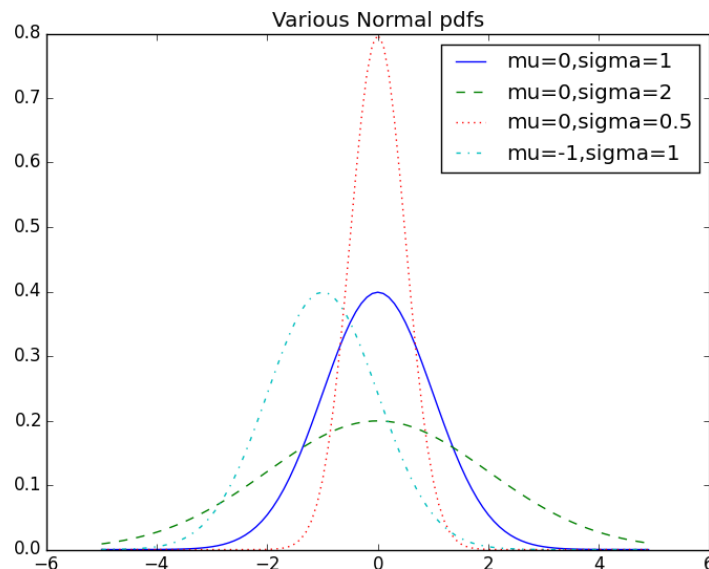
- Observation: 20 coin tosses, 16 times heads → is coin fair?
- H_0 : coin is fair, probability of heads $\rho=0.5$; H_1 : coin not fair
- Sampling distribution of proportion p of heads from N coin tosses under H_0
- Exact probabilities for p : $\frac{0}{N}, \frac{1}{N}, \dots, \frac{N}{N}$
- (Discrete) probability distribution over p :
Binomial distribution (two parameters: N, r)
- $$P\left(p = \frac{i}{N}\right) = \frac{N!}{i!(N-i)!} r^i (1-r)^{N-i}$$
- $$P\left(p = \frac{16}{20} = 0.8\right) = \frac{20!}{16!(20-16)!} 0.5^{20} = 0.0046$$
- $< \alpha = .05$ and $< \alpha = .01 \rightarrow$ reject H_0



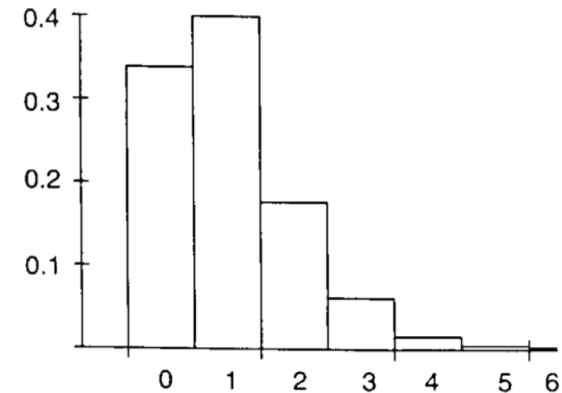
- For continuous distributions, no probability can be assigned to individual points as in discrete distributions
- E.g., *sampling distribution of mean* (draw all possible samples of size N from a given population and calc means) vs. sampling distribution of proportion
- Continuous distributions give probabilities for **ranges of outcomes**
- Represented with a *probability density function* (pdf); the *cumulative distribution function* (cdf) gives the probability that a random variable is less than or equal to a certain value
- Probability of observing value in interval = integral of pdf over the interval (or delta of cdf)
- Hypotheses are tested by asking about probability of an observation at least as extreme as a particular observation

- The *sampling distribution of the mean* of samples of size N **approaches a normal distribution** as N increases
- If the samples are drawn from a population with mean μ and standard deviation σ , then the mean of the sampling distribution is μ and its standard deviation is σ/\sqrt{N}
- True, irrespective of shape of population distribution from which samples are drawn
- i.e., provided the samples are large ($N \geq 30$), drawing from any population (independent, random variable), the sampling distribution of the sample mean \bar{x} is normal
- mean of sampling distribution approaches population mean μ as N increases ($\rightarrow \infty$)

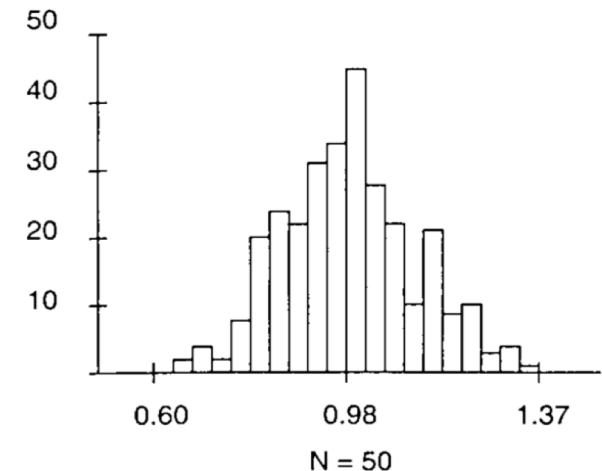
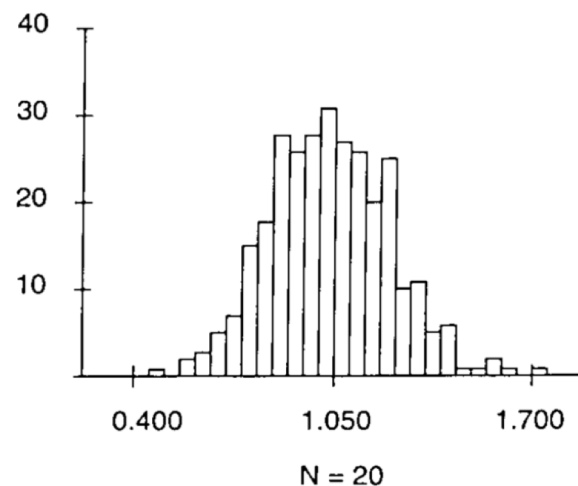
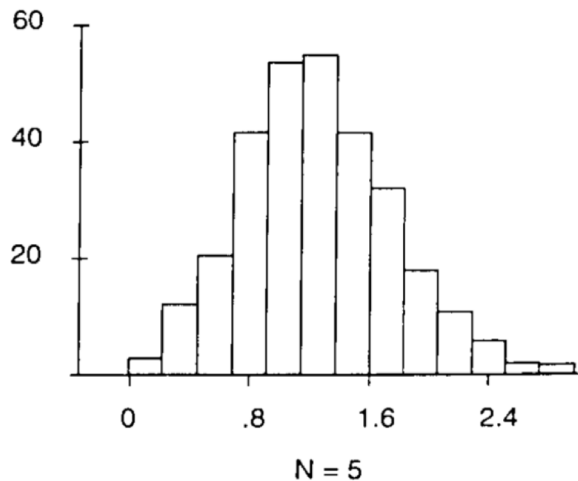
- Bell curve-shaped distribution; determined by two parameters: mean μ (center) and standard deviation σ (“wideness”)
- pdf:
$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
- Standard normal distribution: $\mu = 0, \sigma = 1$
- If X is a normal random variable with mean μ and standard deviation σ , $Z = \frac{(X-\mu)}{\sigma}$ is a standard normal variable



- Example:
skewed population distribution
($\mu_p = 1.0$, $\sigma_p = 0.948$),
discrete distribution, integer values [0, 6]

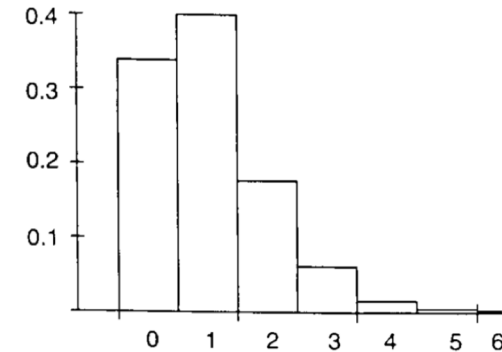


- Randomly draw 300 samples of size N from this distribution

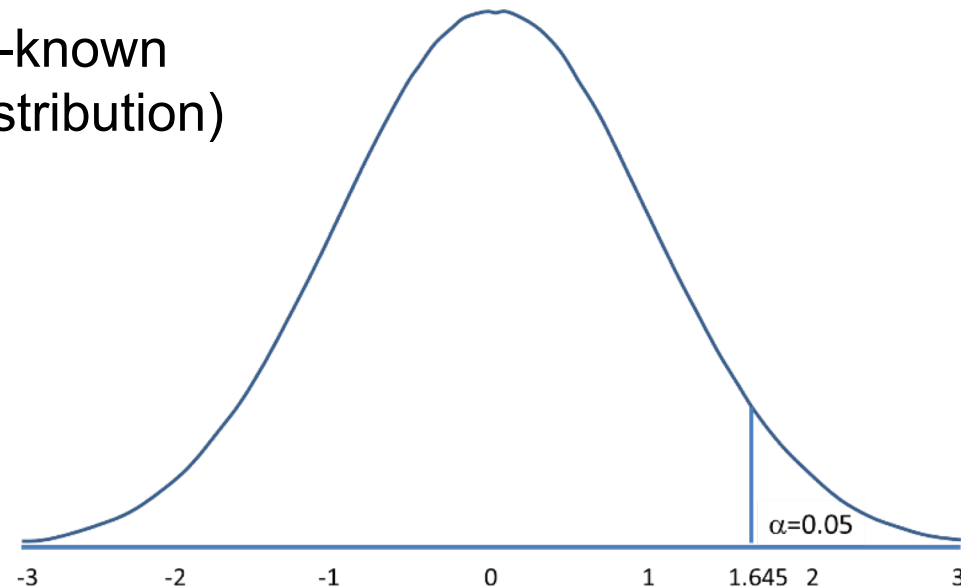


- The larger N , the lower the uncertainty of μ (*standard error* = std. of sampling distribution)

- Using the **Z-test** to test whether there are *significant differences* in means
- Same example from previous slide, representing rate of task failures of a system, e.g. no failure ($x=0$) in 34%
- $N=25$ new tasks (hard) $\rightarrow \bar{x}_{new} = 2.8$
- Hard tasks cause more system failures than ordinary tasks?
 $\rightarrow \mu_{new}$ significantly higher than population mean $\mu_p = 1.0$?
 $\rightarrow H_0: \mu_p = \mu_{new} = 1.0; \quad H_1: \mu_p < \mu_{new}; \quad \alpha = .05$
- CLT: sampling distribution of \bar{x}_p , thus \bar{x}_{new} (stemming from same population under H_0), is normal with mean μ_p and std $\sigma_{\bar{x}} = \sigma_p / \sqrt{N} = \frac{0.948}{\sqrt{25}} = 0.19$
- Next: “standardize” using Z score



- Observed mean of failures in hard tasks $\bar{x}_{new} = 2.8$, expected $\mu_p = 1.0$
- Difference $\bar{x}_{new} - \mu_p = 1.8$ units above sampling distribution mean
- In relation to $\sigma_{\bar{x}}$:
- Standard score / Z score:** $Z = \frac{(\bar{x}_{new} - \mu_p)}{\sigma_{\bar{x}}}$
- $Z = \frac{1.8}{0.19} = 9.47 \rightarrow$ sample result 9.47 std units above expected value
 \rightarrow reject H_0
- Z score normalizes values to well-known standard normal distribution (Z distribution)
- One-tailed test:**
 “rejection region” upper 5%:
 reject H_0 if $Z > 1.645$
- For lower 5% ($H_1: \mu_p > \mu_{new}$):
 reject H_0 if $Z \leq -1.645$

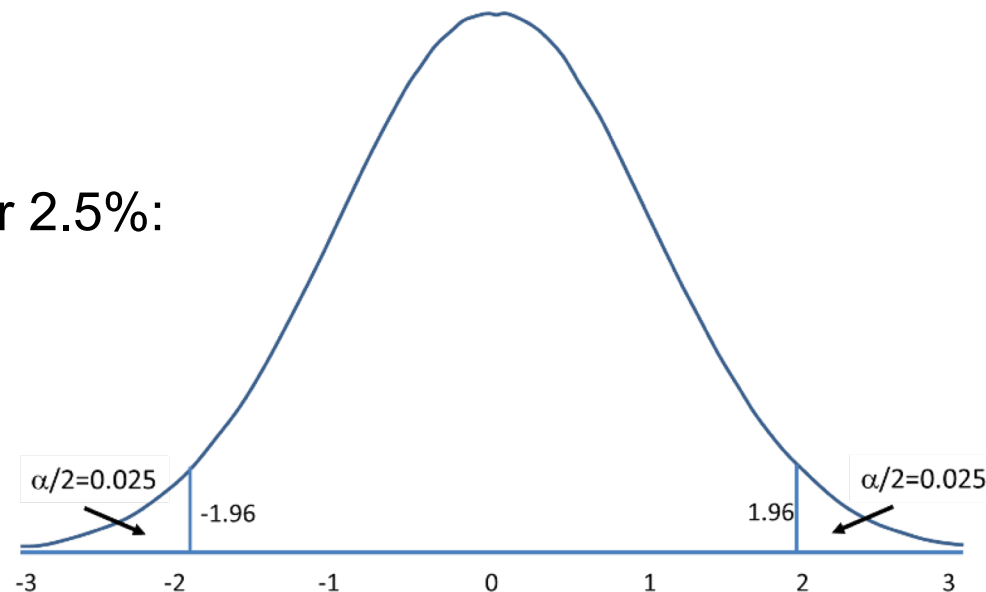


- Same example, different scenario: system applied to new environment
- Sample of $N=25$ different tasks $\rightarrow \bar{x}_{dif} = 1.35$
- Are environments significantly different in terms of failure rate?
 $\rightarrow \mu_{dif}$ significantly different than population mean $\mu_p = 1.0$?
 $\rightarrow H_0: \mu_p = \mu_{dif} = 1.0; \quad H_1: \mu_p \neq \mu_{dif}; \quad \alpha = .05$

- $$Z = \frac{1.35 - 1.0}{0.19} = 1.842$$

- H_1 undirected: **Two-tailed test**
“rejection region” upper and lower 2.5%:
reject H_0 if $Z > 1.96$ or $Z \leq -1.96$

- In example: H_0 not rejected,
no significant difference in
environments



Critical values

- Values of sample mean sufficient to reject H_0 at particular confidence level α (\rightarrow go back from Z score to value range)
e.g., for $H_1: \mu_p > \mu_{new}$, $\alpha = .05$: $\bar{x}_{crit} = \mu_p - 1.645\sigma_{\bar{x}}$
- in example above $\bar{x}_{crit} = 16.75$, i.e., if $\bar{x} > 16.75$, reject H_0
- = comparison to percentiles

p Values

- Probability of sample result given H_0
- Report significant result with actual p value rather than α
- ! value refers to area of sampling distribution bounded by sample result, not area defined by, e.g. $\alpha < .05$

Scenario: **population standard deviation is unknown**

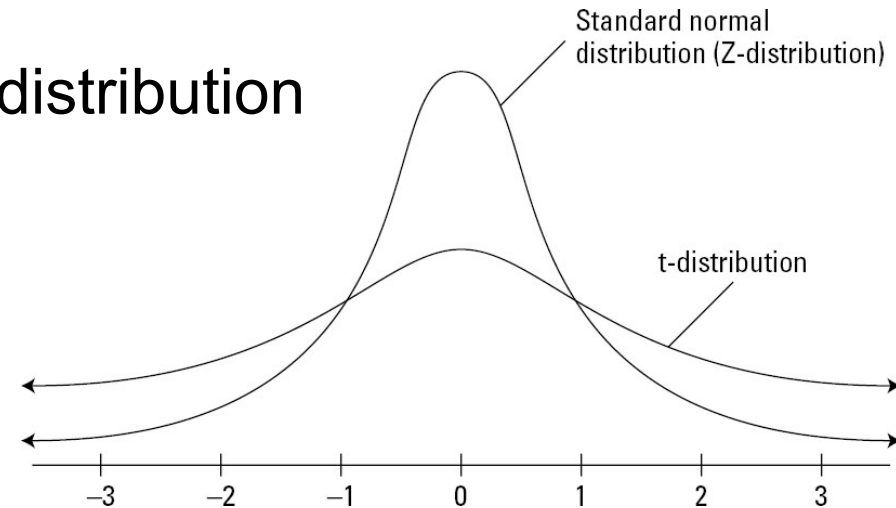
- typically, this will be the case
- Estimate σ from sample standard deviation s : $\hat{\sigma} = s$
- For standard error and Z score substitute $\hat{\sigma}$ for σ

Scenario: **all population parameters are unknown**

- Compare a sample against a chosen threshold/arbitrary value, e.g. a targeted performance measure
- Threshold represents the mean of an imagined null hypothesis distribution
- Z test can be performed using sample estimates (see above)
- This is not ideal and should not be done

Scenario: **N is small** ($N < 30$)

- Use *t distribution* as sampling distribution
- Similar to normal distribution, “heavier tails”, extreme values more likely
- Accounts for less confidence due to smaller N
- (Almost) same procedure as with Z score
- $$t = \frac{(\bar{x} - \mu)}{\hat{\sigma}_{\bar{x}}} = \frac{(\bar{x} - \mu)}{s/\sqrt{N}}$$
- However, t is a family of distributions; one distribution for each value of N ($N > 1$)
- Refer to t distribution with $N-1$ degrees of freedom



- Example: burglar looking for places to rob...
- Finds area with $N=5$ expensive cars:
mean price $\bar{x} = 20,270$, sample standard dev $s = 5,811$
- Mean cost of cars in town $\mu = 12,000$
- Worth robbing? (=significantly richer area?)

- Test statistic $t = \frac{20270 - 12000}{5811/\sqrt{5}} = 3.18$
- Look up in table for $df = N-1=4$
- Test statistic t between values
for $t_{.025}$ and $t_{.01}$
→ has p value < 0.025 and > 0.01
- Area significantly more expensive
($p < 0.025$)

Table 4.3 A fragment of a table of t distributions.

Degrees of freedom	$t_{.05}$	$t_{.025}$	$t_{.01}$	$t_{.005}$
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.955	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.709
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
⋮	⋮	⋮	⋮	⋮
∞	1.645	1.965	2.330	2.570

- Comparing the means of two samples
- e.g., performance indicators from two algorithms
- Identical test logic, slightly different t statistic
- Sampling distribution of difference of means**

$$H_0 : \mu_1 = \mu_2;$$

$$H_1 : \mu_1 \neq \mu_2 \text{ (two-tailed test)}$$

$$H_1 : \mu_1 > \mu_2 \text{ (one-tailed test)}$$

$$H_1 : \mu_1 < \mu_2 \text{ (one-tailed test)}$$

- Starting from one sample t-Test:
→ SS ... sum of squares

$$\hat{\sigma}_{\bar{x}} = \sqrt{s^2/N} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1} \cdot \frac{1}{N}} = \sqrt{\frac{SS}{df \cdot N}}$$

- In two sample case: two sample stds to estimate standard error of sampling distribution of difference of means

$$\sigma_{\bar{x}_1 - \bar{x}_2}^2 = \sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2$$

$$\hat{\sigma}_{pooled}^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}, \quad \hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\hat{\sigma}_{pooled}^2 \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}$$

(same form but distribution of differences and two sample sizes)

- Test statistic: $t_{\bar{x}_1 - \bar{x}_2} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$

Two Sample t-Test Example (Cohen, 1995)

- Comparison of two algorithms A and B, each running on half of 50 tasks

- Formulate a null hypothesis and an alternative hypothesis:

$$\mu_A = \mu_B; \text{ equivalently } \mu_{A-B} = 0;$$

$$\mu_A \neq \mu_B; \text{ equivalently } \mu_{A-B} \neq 0.$$

- In $N_A = 25$ and $N_B = 25$ trials, the sample means for A and B are:

$$\bar{x}_A = 127, s_A = 33;$$

$$\bar{x}_B = 131, s_B = 28.$$

- Determine the sampling distribution of the difference of the sample means given the null hypothesis. The mean of the sampling distribution is $\mu_{A-B} = 0$. The estimated standard error is derived from the pooled variance:

$$\hat{\sigma}_{pooled}^2 = \frac{(N_A - 1)s_A^2 + (N_B - 1)s_B^2}{N_A + N_B - 2} = \frac{(24)1089 + (24)784}{48} = 936.5.$$

$$\hat{\sigma}_{\bar{x}_A - \bar{x}_B} = \sqrt{\hat{\sigma}_{pooled}^2 \left(\frac{1}{N_A} + \frac{1}{N_B} \right)} = \sqrt{936.5 \left(\frac{1}{25} + \frac{1}{25} \right)} = 8.65.$$

The test statistic is

→ look up t in table to obtain p value → H_0 not rejected

$$t_{\bar{x}_A - \bar{x}_B} = \frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}_{\bar{x}_A - \bar{x}_B}} = \frac{127 - 131}{8.65} = -0.46.$$

expected as for $df=48$ (>30) t similar to Z
and value in bounds of Z

Paired sample t-test

- Similar to two sample t-test, but paired samples
- **Sampling distribution of mean of differences** of paired values
- Advantages:

- samples are correlated; performance measures are estimated from same samples e.g., by using identical cross-validation for all methods
- Minimizes variance (half the test problems), increases confidence!

- Record difference and calculate mean and std of differences δ

- If no difference, mean expected to be zero: $H_0 : \mu_\delta = 0$

- Test statistic against one sample t-test

- $$t_\delta = \frac{\bar{x}_\delta - \mu_\delta}{\hat{\sigma}_\delta}, \quad \hat{\sigma}_\delta = \frac{s_\delta}{\sqrt{N_\delta}}$$

Table 4.4 The design of t tests for two samples.

Trial	A	B
1	p_1	p_6
2	p_2	p_7
3	p_3	p_8
4	p_4	p_9
5	p_5	p_{10}
	\bar{x}_A	\bar{x}_B

(4.4a) Two-sample t test.

Trial	A	B	$\delta = A - B$
1	p_1	p_1	
2	p_2	p_2	
3	p_3	p_3	
4	p_4	p_4	
5	p_5	p_5	
			\bar{x}_δ

(4.4b) Paired sample t test.

- Task: music genre classification
- Approach:
 - Content features
 - Gaussian Mixture Models (GMMs)
 - Nearest neighbor classification based on GMM similarity function
- Question: *“Do GMMs with mixtures of 30 Gaussians (GMM30) achieve better genre classification accuracy results than GMMs with mixtures of 10 Gaussians (GMM10)?”*
- Test setup:
 - 10-fold cross validation
 - Comparison of different settings (GMM10 vs GMM30) on same subsets/folds

Case Study by Flexer (2006)

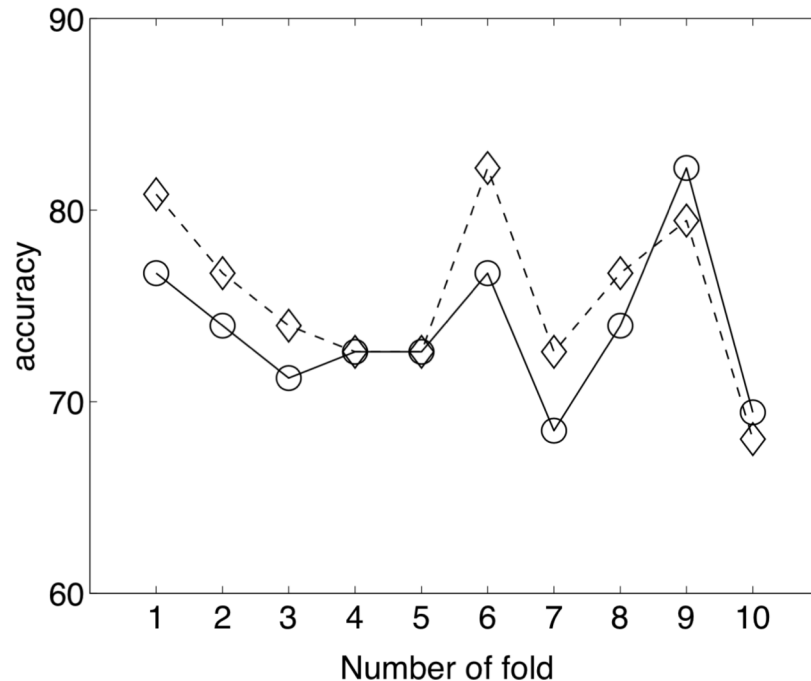


Figure 1: Results of the 10-fold cross-validation. Number of fold on x-axis, accuracy in percent on y-axis, solid line with circles for GMM10, broken line with diamonds for GMM30.

- High correlation (0.8) of performance across the 10 folds
- Parallelization of samples important!

mean accuracy \pm var:

- GMM10: 73.79% \pm 16.00
- GMM30: 75.57% \pm 19.39

Judging only from mean accuracy, GMM30 is the better system

Table 2: Summary of different accuracy results depending on evaluation method used. Results are given in percent correct classification.

Evaluation method	GMM10	GMM30
resubstitution	99.86	100.00
cross-validation range	68.49 - 82.19	68.06 - 82.19
cross-validation best	82.19	82.19
cross-validation mean	73.79	75.57
cross-validation confidence interval	70.93 - 76.65	72.42 - 78.72

- $N < 30$ (...10-folds \rightarrow 10 sampled values)
- Use paired sample t-test

we have two classifiers A and B and we perform a cross validation with N folds. $x_{A,i}$ and $x_{B,i}$ are the achieved accuracies on the i th test fold.

$$t = \frac{\bar{d}}{s_{\bar{d}}}$$

$$\bar{d} = \frac{1}{N} \sum_{i=1}^N d_i$$

$$s_{\bar{d}} = \sqrt{\frac{\sum_{i=1}^N d_i^2 - (\sum_{i=1}^N d_i)^2 / N}{N(N-1)}}$$

$$d_i = x_{A,i} - x_{B,i}$$

We compute the t -value and examine the observed performance difference at an appropriate level of significance $\alpha = 1$ or 5% (i.e. a probability of 95 or 99%) and with degrees of freedom $df = N - 1$ for significance with the help of a t -table (for the two-tailed test, i.e. H_0 will be rejected when the t -value is either sufficiently small or sufficiently large.).

The difference in genre classification accuracy between GMM10 and GMM30 is not significant: $|t| = |-2.1077| < t_{(95, df=9)} = 2.26$ (the same of course holds true for the stricter 99% error level).

- Paul R. Cohen, Empirical Methods for Artificial Intelligence, MIT Press, 1995.
- Arthur Flexer, Statistical Evaluation of Music Information Retrieval Experiments, Journal of New Music Research 35 (2), 113-120, 2006.
- Joel Grus, Data Science from Scratch: First Principles with Python, O'Reilly, 2015.

Experiment Error Analysis and Statistical Testing

Experiment Design for Data Science - Block 2
Lecture 6

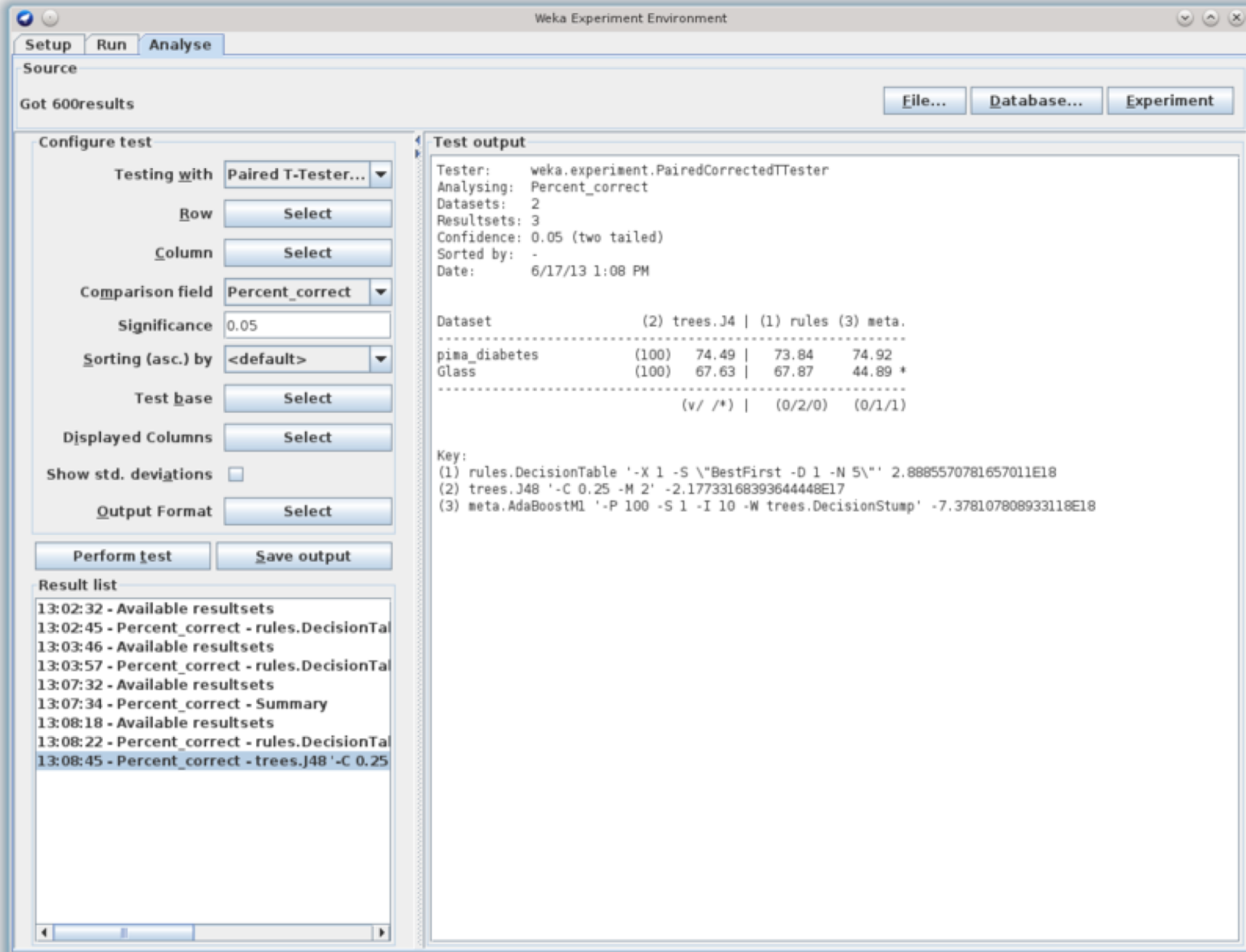
Peter Knees

peter.knees@tuwien.ac.at

Institut für Information Systems Engineering, TU Wien

- Two sample t-test for paired samples
- **Sampling distribution of mean of differences** of paired values
- Advantages:
 - Samples are correlated; performance measures are estimated from same samples
 - Minimizes variance (half the test problems), increases confidence!
- Test statistic against one sample t-test
- Preferred scenario
 - compare performance measures on same splits to perform paired sample t-tests

- **WEKA Experimenter**
- Automated comparison of different classifiers on different datasets
- Generate sample values by repeatedly running experiments (train/test split, n-fold cross validation) → *Iteration Control*
- Selection of performance measure → *Comparison Field*
- Test for significant difference between algorithms via paired t-test
- → better: corrected paired t-test [Nadeau & Bengio, 2003]: takes into account variability of training set (not only test set) due to sampling, e.g. in cross-validation



Weka Experiment Environment

Setup Run **Analyse**

Source
Got 600 results

File... Database... Experiment

Configure test

Testing with: Paired T-Tester...

Row: Select

Column: Select

Comparison field: Percent_correct

Significance: 0.05

Sorting (asc.) by: <default>

Test base: Select

Displayed Columns: Select

Show std. deviations: ☐

Output Format: Select

Perform test Save output

Test output

Tester: weka.experiment.PairedCorrectedTTester
Analysing: Percent_correct
Datasets: 2
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by: -
Date: 6/17/13 1:08 PM

Dataset	(2) trees.J4	(1) rules	(3) meta.
pima_diabetes	(100) 74.49	73.84	74.92
Glass	(100) 67.63	67.87	44.89 *

(v/ /*) | (0/2/0) (0/1/1)

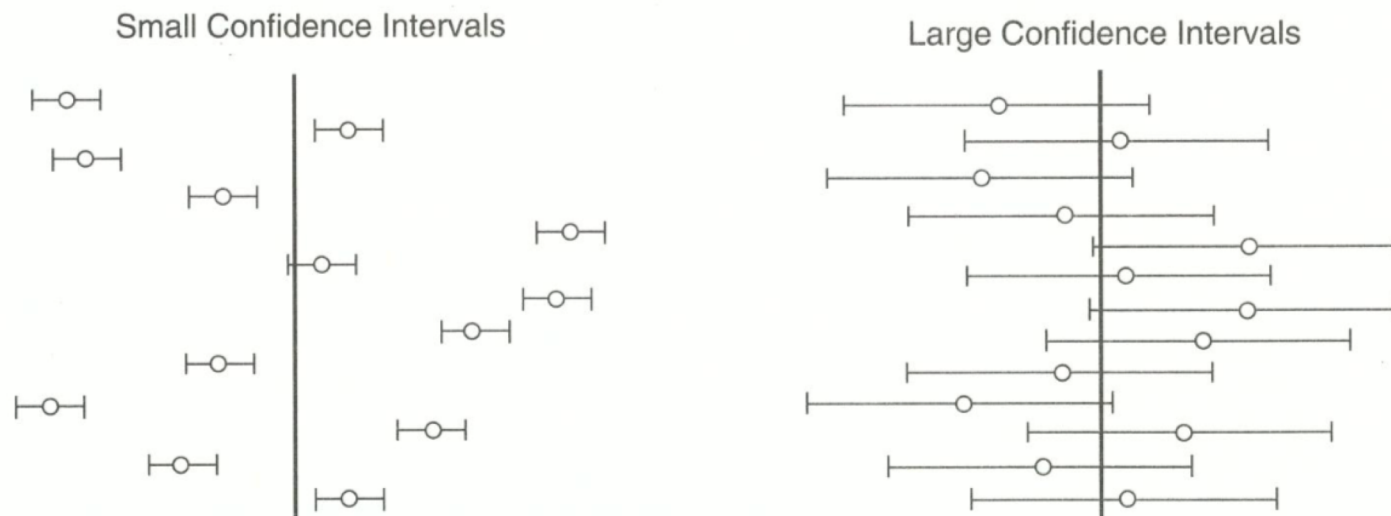
Key:
(1) rules.DecisionTable '-X 1 -S \"BestFirst -D 1 -N 5\"' 2.8885570781657011E18
(2) trees.J48 '-C 0.25 -M 2' -2.17733168393644448E17
(3) meta.AdaBoostM1 '-P 100 -S 1 -I 10 -W trees.DecisionStump' -7.378107808933118E18

Result list

- 13:02:32 - Available resultsets
- 13:02:45 - Percent_correct - rules.DecisionTa
- 13:03:46 - Available resultsets
- 13:03:57 - Percent_correct - rules.DecisionTa
- 13:07:32 - Available resultsets
- 13:07:34 - Percent_correct - Summary
- 13:08:18 - Available resultsets
- 13:08:22 - Percent_correct - rules.DecisionTa
- 13:08:45 - Percent_correct - trees.J48 '-C 0.25

- Characterize the accuracy of parameter estimate
- E.g., estimate population mean μ via sample mean \bar{x}

$$\mu = \bar{x} \pm \varepsilon$$
- If ε small, \bar{x} is good estimate for μ
- Wider confidence interval around \bar{x} , more likely to contain μ , but also low precision to estimate parameter



Black line: true population mean μ

- CLT: draw large number of samples of size $N \rightarrow \sim 95\%$ of sample means will fall within interval of 2×1.96 standard deviations around population mean μ
 - $\rightarrow \bar{x} = \mu \pm 1.96\sigma_{\bar{x}}$ for 95% of the means \bar{x}
- as well as
- for $\varepsilon = 1.96\sigma_{\bar{x}}$, confidence interval $\bar{x} \pm \varepsilon$ contains μ for 95% of samples

! Confidence interval $\bar{x} \pm 1.96\sigma_{\bar{x}}$

- **does not mean:** $\Pr(\mu = \bar{x} \pm 1.96\sigma_{\bar{x}}) = 0.95$ (=“I am 95% sure about the true value of μ .”)
- But “**I am 95% sure that this interval around \bar{x} contains μ .**”
- (no probability of value of μ , population mean μ is a constant)

- The bigger the better ...right?
- For **parameter estimation**, larger samples narrow confidence intervals ✓
- For **hypothesis testing**, if we have enough confidence already, nothing is gained by increasing sample sizes ✗
- Increasing sample size N can boost any effect to significance (reduces standard error)
- Hence, the quality seal “statistically significant difference” often required in machine learning research can be engineered by drawing larger samples
- → can show non-existing effects of independent variable
- Chose reasonable value for sample size N !

Errors associated to rejecting H_0 :

- $\alpha = \Pr(\text{type I error}) = \Pr(\text{reject } H_0 \mid H_0 \text{ is true})$
- $\beta = \Pr(\text{type II error}) = \Pr(\text{fail to reject } H_0 \mid H_0 \text{ is false})$
- α easy to estimate based on decision rule when to reject H_0
- For β , less obvious... refers to infinite number of alternative hypothesis subsumed under “ H_0 is false”

- Repeated testing “inflates” probability of making a type I error (*multiplicity effect*)
- If H_0 is not rejected and new data is obtained/sampled, overall error increases as samples are not independent
- $Pr(\text{overall type I}) = Pr(\text{type I in test 1}) + Pr(\text{type I in test 2} \mid \text{no type I in test 1}) + \dots$
- e.g., for repeated testing with $\alpha = 0.05$

# interim analyses	1	2	5	10	∞
Probability of a type I error	0.05	0.08	0.14	0.19	1.00

- “If we keep testing, we will find significant results”
- One strategy: divide α by number of tests performed
→ makes it, however, difficult to obtain significant results

Increased probability of type I

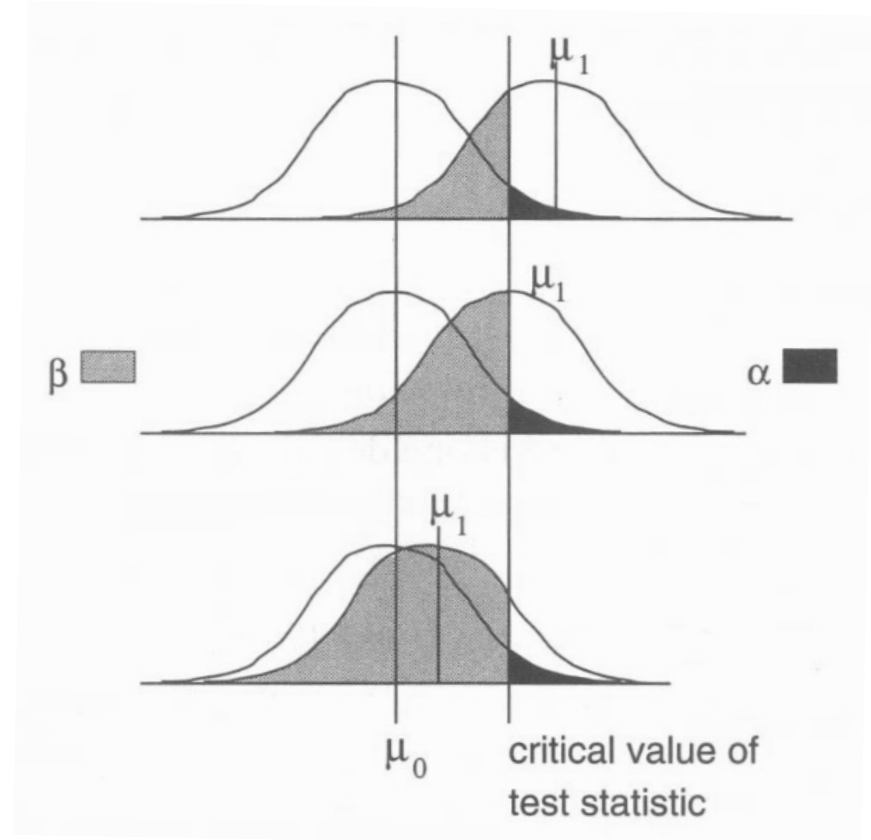
- In cross-validation
 - no independence of samples (e.g., in 5-fold CV, each pair of training set shares 80% of data)
 - Alternative: non-parametric test (e.g., McNemar)
- When comparing multiple sample means

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

Alternative: Analysis of variances (ANOVA)

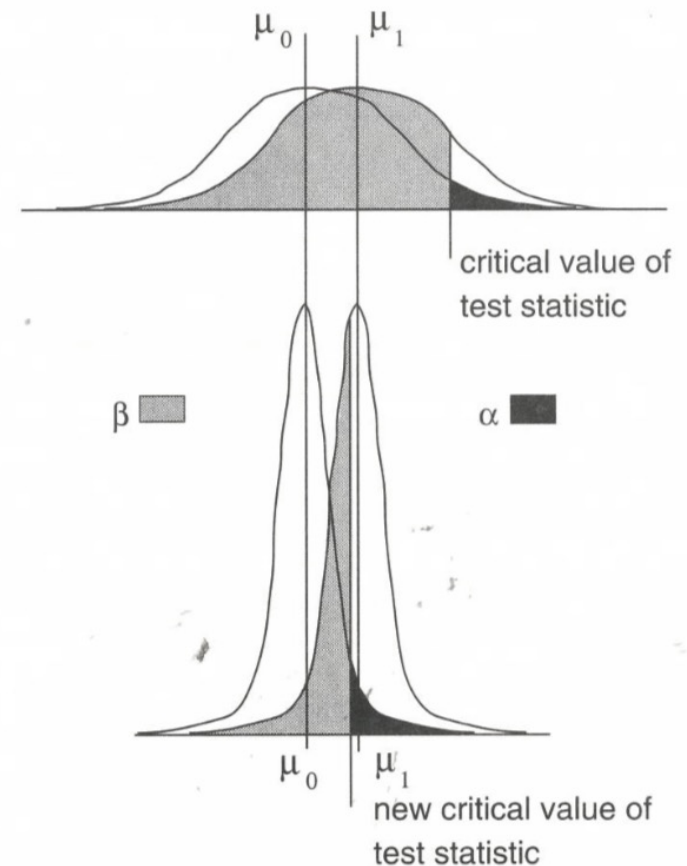
- Tests for equality; H_1 : at least one μ is different
- If H_0 rejected: which to be determined in pairwise post-hoc tests

- $\beta = \Pr(\text{type II error}) = \Pr(\text{fail to reject } H_0 \mid H_0 \text{ is false})$
- Suppose H_0 is false and true sample mean is μ_1 rather than μ_0
- Under H_1 , sampling distribution will be centered about μ_1 (otherwise identical)
- α (type I) depends on μ_0
- β (type II) depends on μ_1 and critical value of test statistic under H_0
- $\Pr(\text{reject } H_0 \mid H_0 \text{ is false}) = 1 - \beta$
- $1 - \beta \dots \text{power of test}$



Relationship between α , β , and $\delta = \mu_0 - \mu_1$

- For fixed α , when δ decreases, β increases and the power decreases:
→ more difficult to discriminate H_0 and H_1 as δ gets smaller
- For fixed δ , only way to:
 - increase power of test is to increase α
 - decrease α is to decrease power of test
- Hence, again, for small δ , power is small
- Remedy: smaller variances
→ increase sample sizes
- → increasing N decreases the probabilities of type I and type II errors



Power of test depends on

- The α level of the test
- δ ... the degree of separation between the H_0 and H_1 distributions
- The variance(s) of the population(s) the samples are drawn from
- The sample size N , affecting the standard error of the sampling distributions under H_0 and H_1

- Different tests have different power; tests with higher power are to be preferred
- Typically, parametric tests have increased power over non-parametric tests (due to knowledge of distribution parameters)

- In order to apply parametric tests, certain conditions need to be fulfilled
- e.g., ANOVA (F-test) assumes
 - Normality: population sample distribution must be normal
 - Independence: sampled observations must be independently/randomly selected from each population
 - Homogeneity (of variance): populations the samples are selected from have equal variance
- e.g., t-test assumes normality, independence, interval/ratio scale level, no outliers
- Statistical tests to test for these conditions to be fulfilled
- **In practice, conditions are often ignored and/or violated**

(tests are quite robust; improper methodology and error is accepted in exchange for automatable decision making)

- Non-parametric statistics make **no assumptions about the probability distributions** of the investigated variables
- E.g, “*two unspecified continuous distributions are identical*”
→ hypothesis makes no assumption on underlying form of distribution, nor on any parameters (e.g., mean, variance)
- For non-interval/ratio data, ordinal scales; data with outliers
- Typically **rank-based methods** used for statistical testing
- Note: *non-parametric models* typically refers to models defined by data, including, e.g. histograms (cf. distribution), k-NN classifier (lazy learner), or support vector machine (large-margin classifier)

- Tests whether matched pair samples are drawn from distributions with **equal medians**
- → test the hypothesis that the difference between matched pair samples X and Y has zero median
- Assumptions: $X, Y \dots$ continuous distributions, at least ordinal scale, paired samples
- Let $p = \Pr(X > Y)$, test $H_0: p = 0.50$
- Sample pairs (x_i, y_i) , ignore pairs with $x_i = y_i \rightarrow m$ pairs
- $W \dots$ # pairs with $y_i - x_i > 0$
→ under H_0 , $W \sim$ binomial dist $b(m, 0.5)$.
- Binomial test (or normal approximation for $m > 25$)
- Critical values: $\Pr(W \leq w)$ and/or $\Pr(W \geq w)$

▪ Mann-Whitney U test

- determine whether two *independent* samples X , Y were selected from populations having the same distribution
- test $H_0: Pr(X > Y) = Pr(X < Y)$
- Alternative to *t-test*, e.g., when data not normal

▪ Wilcoxon signed-rank test

- whether paired samples stem from populations with same distributions
- test whether population mean ranks of paired samples differ
- Alternative to *paired t-test*, e.g., when data not normal
- Requires interval scale (ordinal possible); higher power than sign test

▪ **Kruskal-Wallis test**

- Comparing two or more independent samples with equal or non-equal sample sizes (extension of U test for >2 groups)
- Alternative to *one-way ANOVA*
- testing whether at least one median is different from the others

▪ **Friedman Test**

- Comparing two or more dependent samples
- detect differences in results across multiple test attempts
- Alternative to *two-way ANOVA*

- Paul R. Cohen, Empirical Methods for Artificial Intelligence, MIT Press, 1995.
- Janez Demsär, Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research, 7 pp. 1–30, 2006.
- Chris Drummond, Machine Learning as an Experimental Science, AAAI workshop on evaluation methods for machine learning, pp. 1-5, 2006.
- Claude Nadeau and Yoshua Bengio, Inference for the Generalization Error, Machine Learning, 52, pp. 239–281, 2003.

Reproducibility

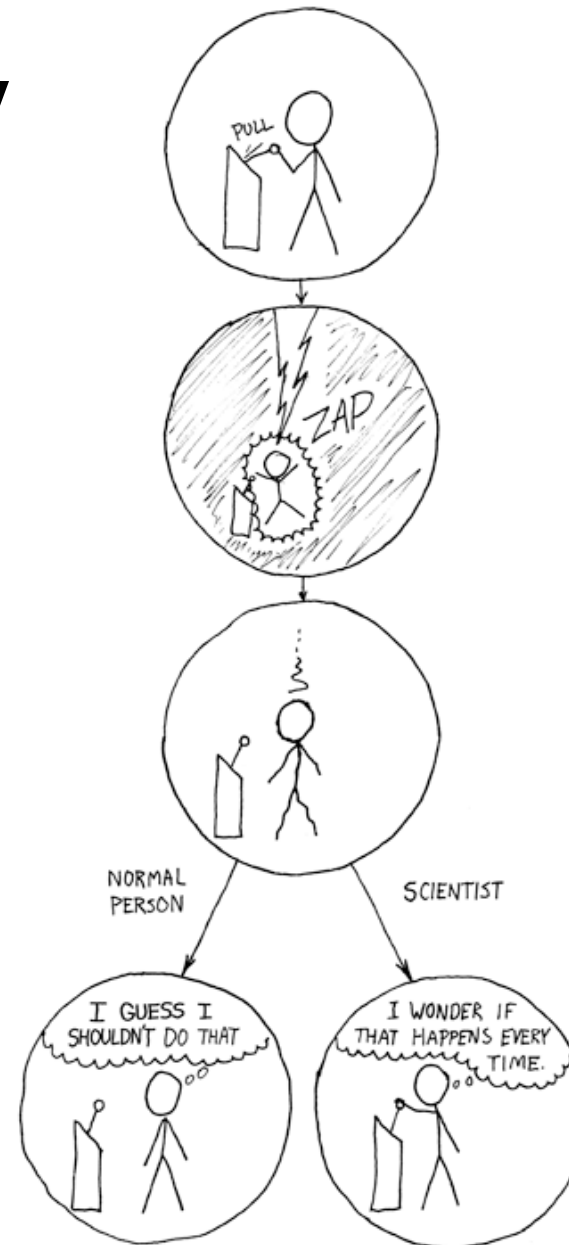
Andreas Rauber

Vienna University of Technology
Favoritenstr. 9-11/188
1040 Vienna, Austria
rauber@ifs.tuwien.ac.at
<http://www.ifs.tuwien.ac.at/~andi>

-
- Reproducibility
 - What are the challenges in reproducibility?
Why do we need it? Why is it difficult?
 - How to address the challenges of complex processes?
 - (How to ensure reproducibility with dynamic data?)
 - Summary
-

Reproducibility

- Reproducibility is core to the scientific method
- Focus not on misconduct – but on complexity and the will to produce good work
- Should be easy
 - Get the code, compile, run, ...
 - Why is it difficult?



Reproducibility in “Small Data”

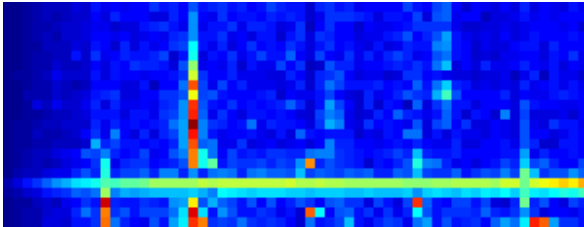
- Carmen M. Reinhart and Kenneth S. Rogoff (2010) vs. Thomas Herndon, Michael Ash, Robert Pollin (2013)
- **Original spreadsheet provided**
 - Some data excluded on purpose
 - Questionable statistical procedures
 - **Excel error**
 - Accidentally missed 5 rows of data!
 - Average Annual Growth changed from -0.1 to 2.2 after correction
- Lead to prominent coverage on importance of transparency, reproducibility



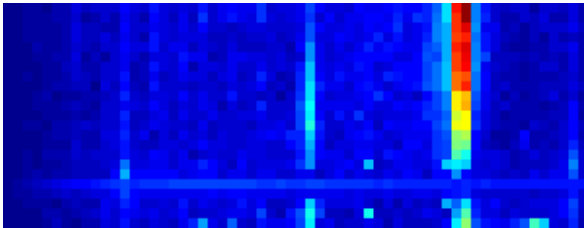
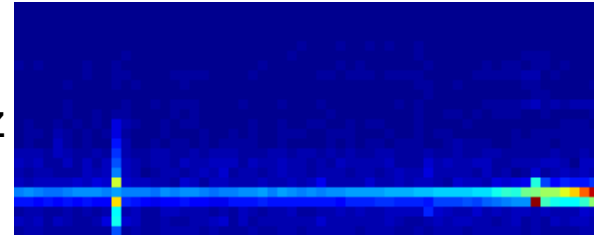
<https://www.newyorker.com/news/john-cassidy/the-reinhart-and-rogooff-controversy-a-summing-up>
<https://www.nytimes.com/2013/04/19/opinion/krugman-the-excel-depression.html>

Challenges in Reproducibility

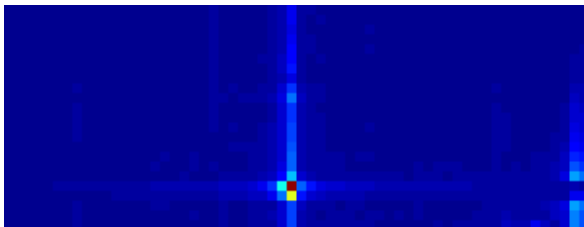
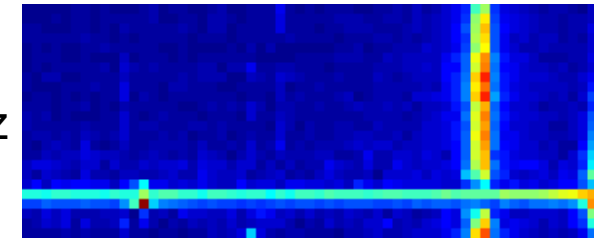
- Excursion: scientific processes



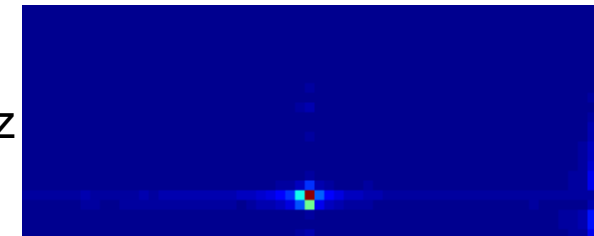
set1_freq440Hz_Am11.0Hz



set1_freq440Hz_Am12.0Hz



set1_freq440Hz_Am05.5Hz



Java

Matlab

A simpler example

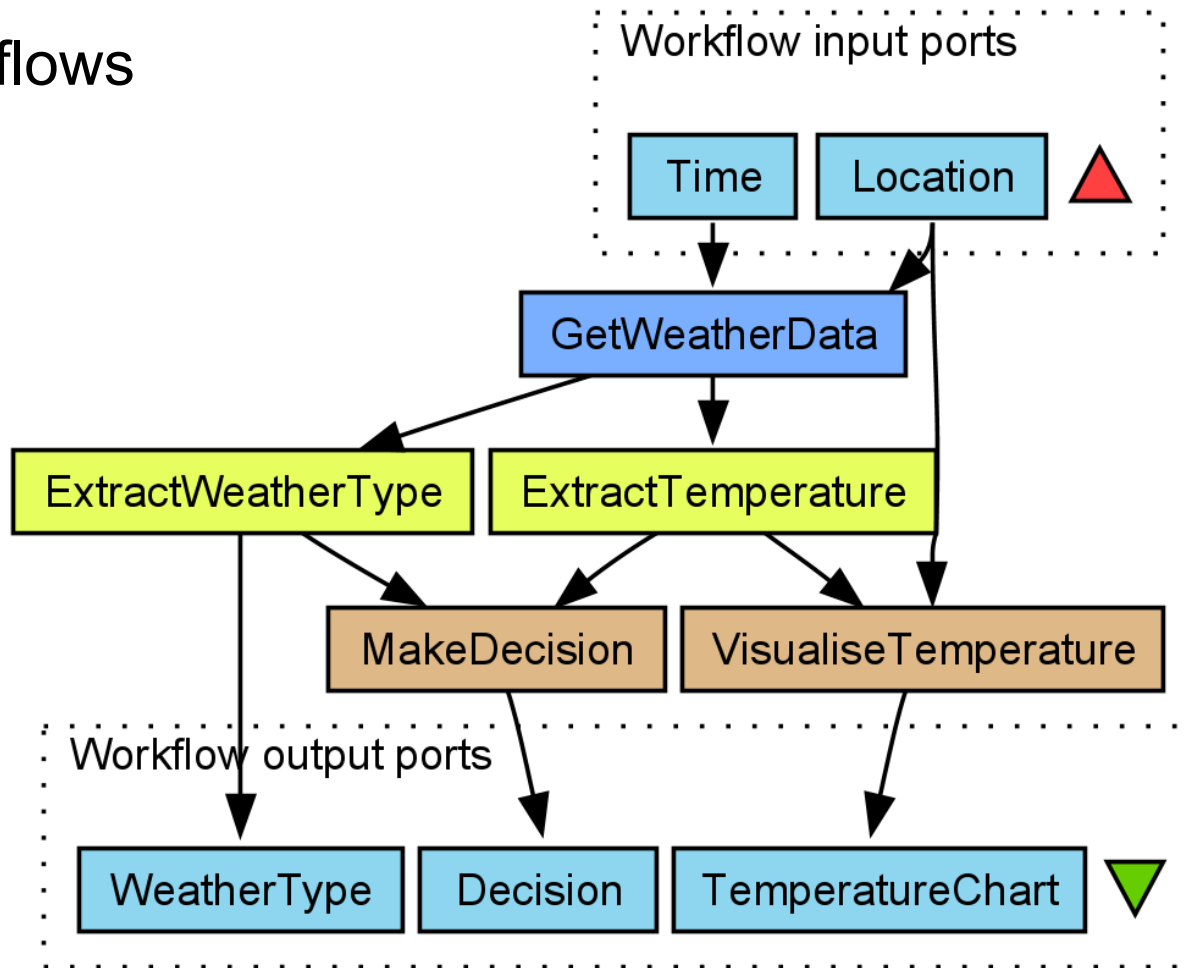
- Image conversion from jpg to tiff using *ImageMagick*

	<i>View Path #1</i>	<i>View Path #2</i>
Data formats	Raw JPEG Stream (fmt/41);Portable Network Graphics (fmt/13)	Raw JPEG Stream (fmt/41);Portable Network Graphics (fmt/13)
Application	ImageMagick 6.8.9-7 Q16 Microsoft Visual C++ 2010	ImageMagick 6.8.9-7
JVM	Java SE 6 Update 45	Java SE 7 Update 10
Operating System	Windows 7 Enterprise SP1	OS X 10.9.4
Hardware	3,3GHz Intel Core i3 8GB 1600MHz DDR3 NVIDIA GT630 2GB	2,3GHz Intel Core i5 4GB 1333MHz DDR3 Intel HD Graphics 3000 384MB

Challenges in Reproducibility

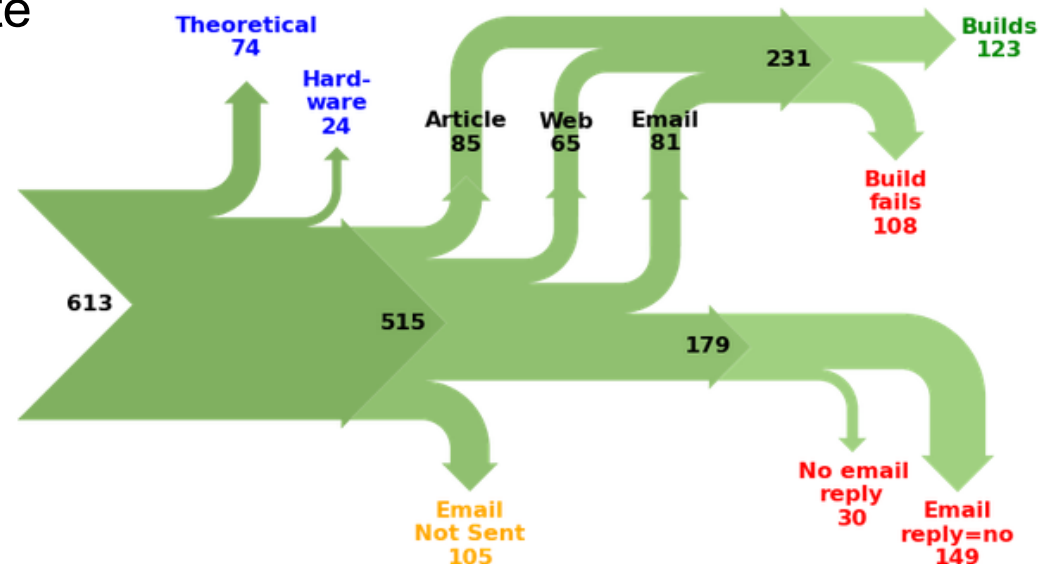
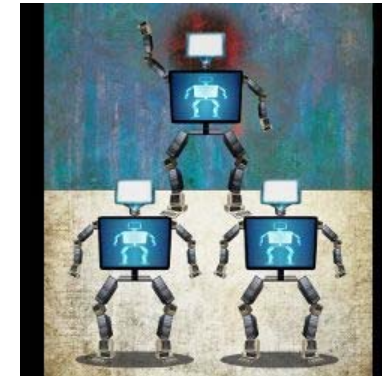
- Workflows

Taverna



Challenges in Reproducibility

- 613 papers in 8 ACM conferences
- Process
 - download paper and classify
 - search for a link to code (paper, web, email twice)
 - download code
 - build and execute



Christian Collberg and Todd Proebsting. "Repeatability in Computer Systems Research," CACM 59(3):62-69.2016

■ ACM Statement on Algorithmic Transparency and Accountability, May 25 2017

http://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf

1. **Awareness:** potential bias
2. **Access and redress:** for individuals and groups
3. **Accountability:** responsible for decisions made by algorithms
4. **Explanation:** encouraged to explain procedures, decisions
5. **Data Provenance:** data collection, bias analysis, ...
6. **Auditability:** models, data, algorithms recorded
7. **Validation and Testing:** rigorous, routinely, public



Excursion: Ethics & Privacy

How can we address this, support us in proper behavior?

■ Steps towards solutions:

- Automated documentation, provenance
- Data versioning, reproducibility
- Monitoring data quality, data drift,
- Defining triggers, roles and responsibilities

■ Open questions

- “Ethical algorithms by design” ?
- Run-time monitoring for ethical behavior of algorithms?
- Automated bias-testing for sensitive attributes?
- Ontology of likely correlated attributes?
- Can we encode ethical rules/behavior?
- Role of randomness in human decision making?

Excursion: Ethics & Privacy

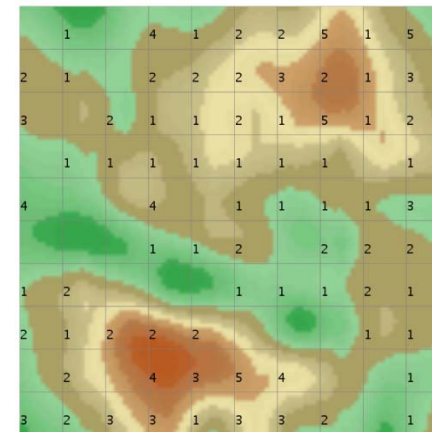
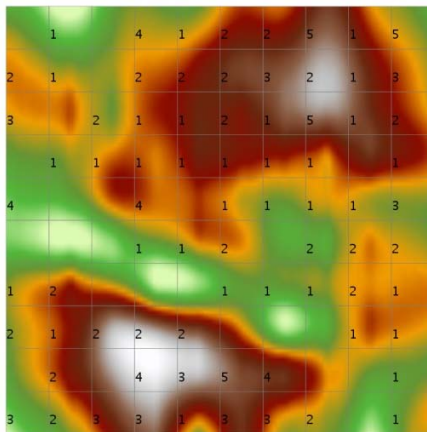
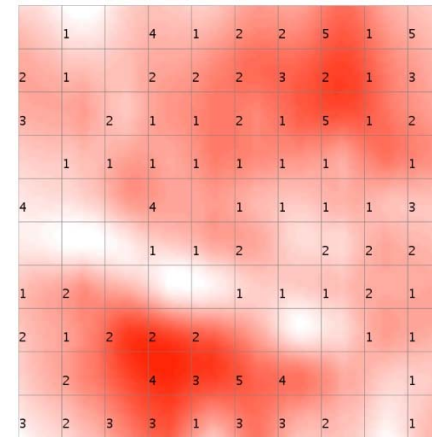
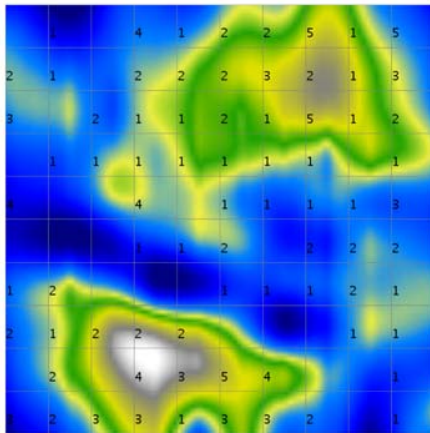
- Issues in Data Mining
 - Used for decision making
 - Use of some attributes not acceptable for some applications
 - religion; sex; race
 - sometimes, these attributes are “hidden” (e.g. race in address)
 - Same attributes ok in other applications (e.g. medical)
 - Document type of decisions permissible based on attribute analysis
 - Precision, over-fitting, generalizing from too few samples (support)
 - Partially, ethical issues arise because of simplicity with which results can be obtained (e.g. information retrieval)

Examples

- Self-driving / connected cars
 - Minimizing the impact of accidents
 - Optimizing routing / driving behavior: global / local optimization
- Service provision
 - From elevators to self-driving cars
 - Infrastructure planning
 - Credit scoring
- Social media-based / crowd decision support (Manipulation and social dynamics)
 - Chatbots
 - Recommender Systems, Information retrieval / filters (hate speech)
 - Wikipedia (edit wars) -> input to algorithms -> ...

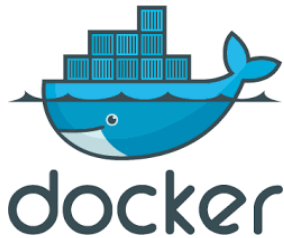
Excursion: Issues with Coloring

- Example: Interpolating over units: color palettes



Reproducibility – solved! (?)

- Provide source code, parameters, data, ...
- Wrap it up in a container/virtual machine, ...



...

- Why do we want reproducibility?
- Which levels of reproducibility are there?
- What do we gain by different levels of reproducibility?
- A simple “re-run” is usually not enough
– otherwise, video would be sufficient....

Types of Reproducibility

- The **PRIMAD Model**¹: which attributes can we “prime”?
 - Data
 - Parameters
 - Input data
 - Platform
 - Implementation
 - Method
 - Research Objective
 - Actors
- What do we gain by “priming” one or the other?

[1] Juliana Freire, Norbert Fuhr, and Andreas Rauber. Reproducibility of Data-Oriented Experiments in eScience. Dagstuhl Reports, 6(1):108-159, 2016.

http://drops.dagstuhl.de/opus/volltexte/2016/5817/pdf/dagrep_v006_i001_p108_s16041.pdf

Types of Reproducibility and Gains

Label	Data		Platform / Stack	Implementation	Method	Research Objective	Actor	Gain
	Parameters	Raw Data						
Repeat	-	-	-	-	-	-		Determinism
Param. Sweep	x	-	-	-	-	-		Robustness / Sensitivity
Generalize	(x)	x	-	-	-	-		Applicability across different settings
Port	-	-	x	-	-	-		Portability across platforms, flexibility
Re-code	-	-	(x)	x	-	-		Correctness of implementation, flexibility, adoption, efficiency
Validate	(x)	(x)	(x)	(x)	x	-		Correctness of hypothesis, validation via different approach
Re-use	-	-	-	-	-	x		Apply code in different settings, Re-purpose
Independent x (orthogonal)							x	Sufficiency of information, independent verification

Reproducibility Papers

- Aim for reproducibility: for one's own sake – and as Chairs of conference tracks, editor, reviewer, supervisor, ...
 - Review of reproducibility of submitted work (material provided)
 - Encouraging reproducibility studies
 - (Messages to stakeholders in Dagstuhl Report)
- Consistency of results, not identity!
- Reproducibility studies and papers
 - Not just re-running code / a virtual machine
 - When is a reproducibility paper worth the effort / worth being published?
 - Issues with peer review and verification...

Peer Review and Verification

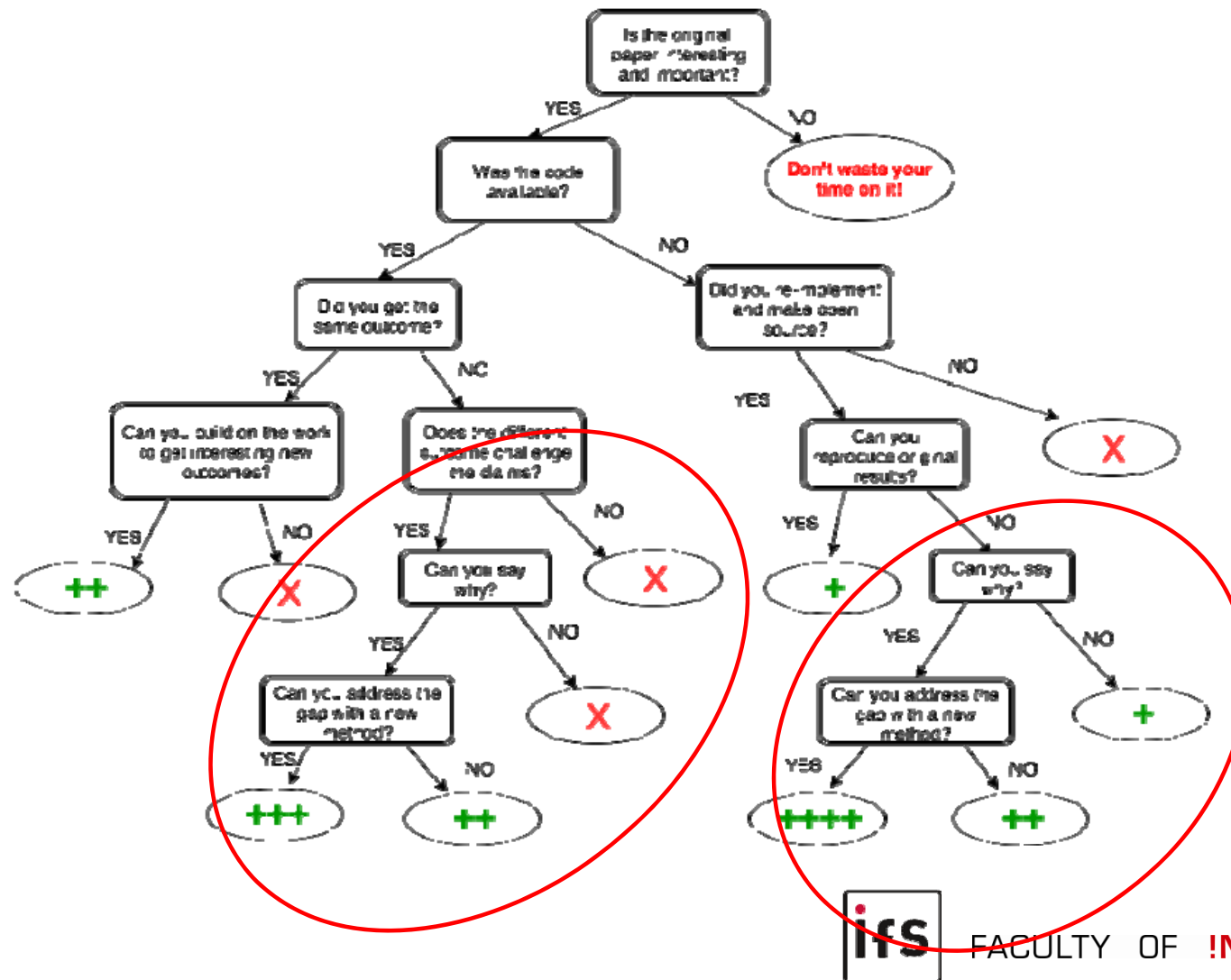
- Peer review is an established process
 - Focused on publications mainly
 - Hardly any data quality reviews
 - Even less reproducibility studies
- Reproducing or replicating experiments is not considered original research
 - No recognition
 - No money
 - A lot of work
- Encourage reproducibility studies
- **Needed beyond science!**

Peer Review and Verification

- Encourage reproducibility studies -> **How?**
- Dagstuhl Seminar:
Reproducibility of Data-Oriented Experiments in e-Science, January 2016, Dagstuhl, Germany
http://drops.dagstuhl.de/opus/volltexte/2016/5817/pdf/dagrep_v006_i001_p108_s16041.pdf
- Call for action to conference Organizers, Editors, ...
- Several conferences include reproducibility tracks

Reproducibility Papers



- When is a Reproducibility paper worth being published?



Learning from Non-Reproducibility

- Do we always want reproducibility?
 - Scientifically speaking: yes!
- Research is addressing challenges:
 - Looking for and learning from non-reproducibility!
- Non-reproducibility if
 - Some (un-known) aspect of a study influences results
 - Technical: parameter sweep, bug in code, OS, ... -> fix it!
 - Non-technical: input data! (specifically: “the user”)

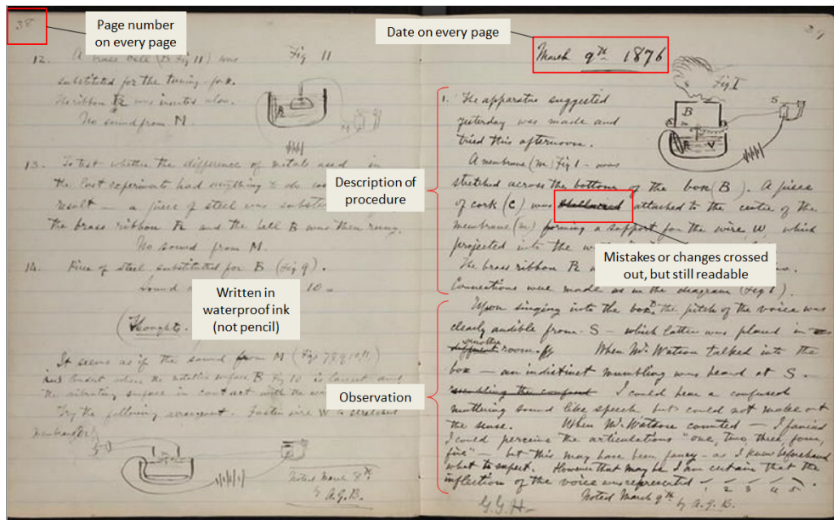
Challenges in MIR – “things don’t seem to work”

- Virtual Box, Github, *<your favourite tool>* are starting points
- Same features, same algorithm, different data -> 
- Same data, different listeners -> 
- Understanding “the rest”:
 - Isolating unknown influence factors
 - Generating hypotheses
 - Verifying these to understand the “entire system”, cultural and other biases, ...
- Benchmarks and Meta-Studies

-
- Reproducibility
 - What are the challenges in reproducibility?
 - How to address the challenges of complex processes?
 - Data Management & Citation
 - Digital Preservation
 - Summary
-

And the solution is...

- Standardization and Documentation
 - Standardized components, procedures, workflows
 - Documenting complete system set-up across entire provenance chain
- How to do this – efficiently?



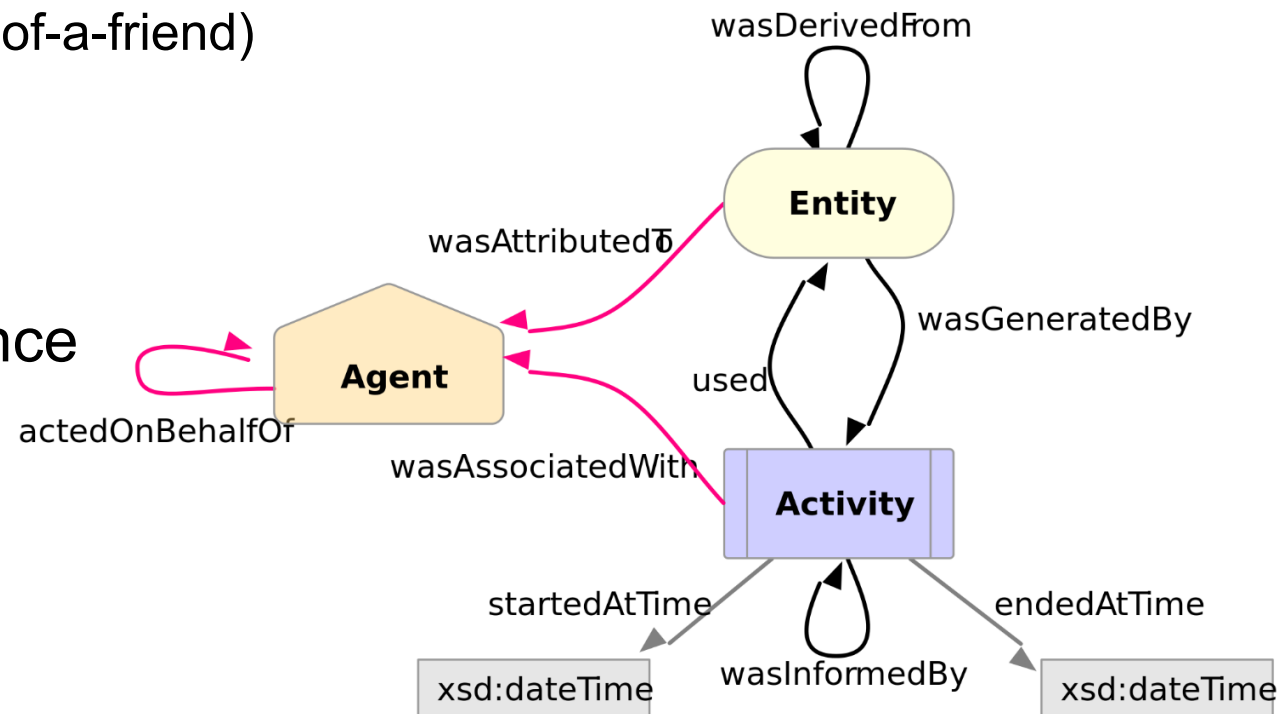
Alexander Graham Bell's Notebook, March 9 1876

Pieter Bruegel the Elder: De Alchemist (British Museum, London)

https://commons.wikimedia.org/wiki/File:Alexander_Graham_Bell's_notebook,_March_9,_1876.PNG

PROV-O

- W3C Recommendation
<https://www.w3.org/TR/prov-o/>
- Ontology to represent provenance information
- May use other languages
 - FOAF (friends-of-a-friend)
 - Dublin Core
 - PREMIS
- (Alternative:
Open Provenance
Model)

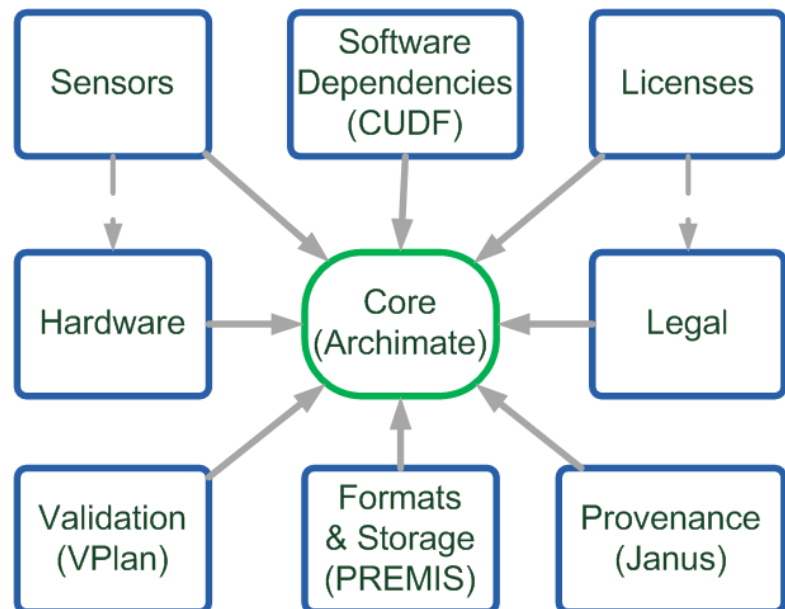


And the solution is...

- Standardization and Documentation
 - Standardized components, procedures, workflows
 - Documenting complete system set-up across entire provenance chain
- **How to do this – efficiently!?**
- **Ideally:**
 - **Processing pipeline documents provenance automatically**
- **Reality:**
 - **Combination**
 - **automatic documentation / logging**
 - **monitoring behaviour of the system**

Documenting a Process

- Context Model: establish what to document and how
- Meta-model for describing process & context
 - Extensible architecture integrated by core model
 - Reusing existing models as much as possible
 - Based on ArchiMate, implemented using OWL
- Extracted by static and dynamic analysis



Context Model – Static Analysis

- Analyses steps, platforms, services, tools called
- Dependencies (packages, libraries)
- HW, SW Licenses, ...

```
#!/bin/bash

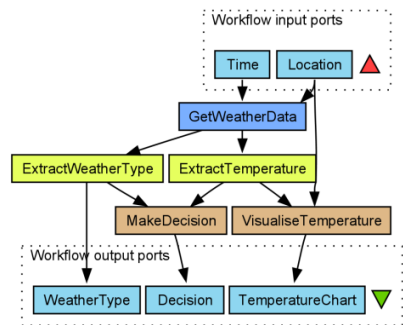
# fetch data
java -jar GestBarragensWSClientIQData.jar
unzip -o IQData.zip

# fix encoding
#iconv -f LATIN1 -t UTF-8 iq.r > iq_utf8.r

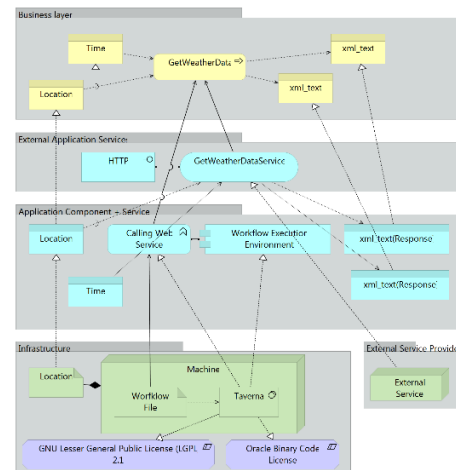
# generate references
R --vanilla < iq_utf8.r > IQout.txt

# create pdf
pdflatex iq.tex
pdflatex iq.tex
```

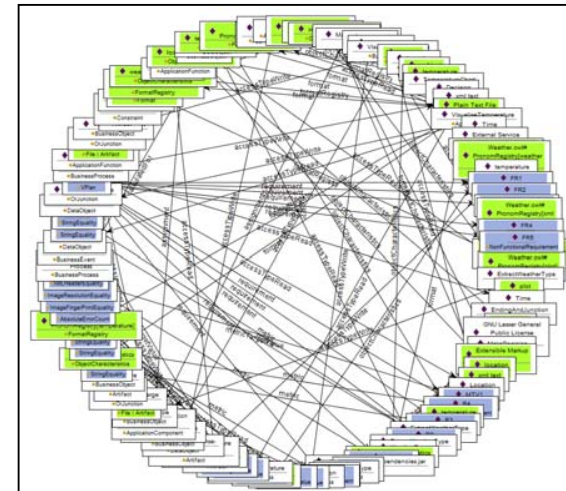
Script



Taverna Workflow



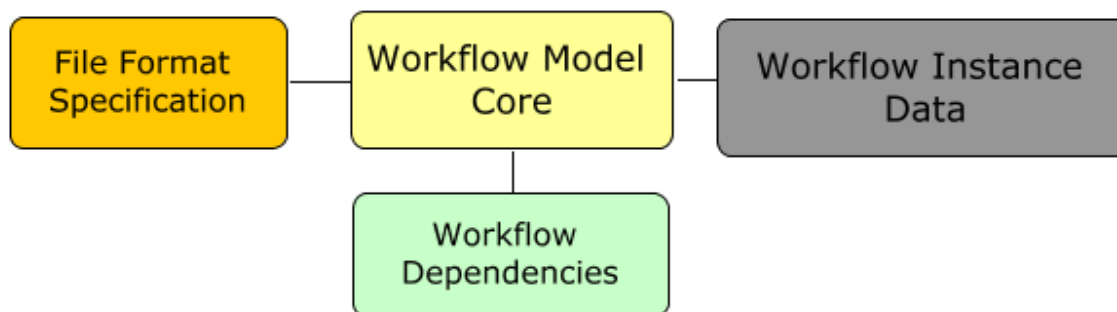
ArchiMate model



Context Model
(OWL ontology)

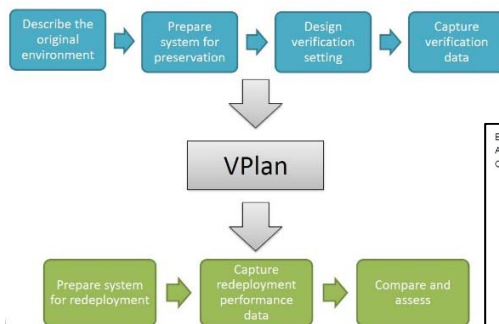
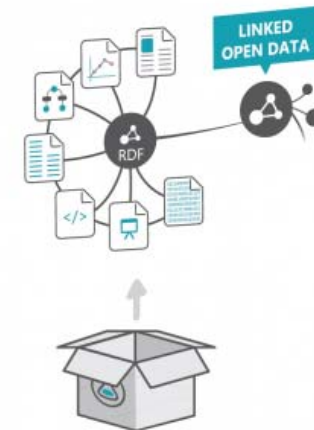
Context Model – Dynamic Analysis

- Process Migration Framework (PMF)
 - designed for automatic redeployments into virtual machines
 - uses *strace* to monitor system calls
 - complete log of all accessed resources (files, ports)
 - captures and stores process instance data
 - analyse resources (file formats via PRONOM, PREMIS)



Preservation and Re-deployment

- „Encapsulate“ as complex Research Object (RO)
- DP: Re-Deployment beyond original environment
 - Format migration of elements of ROs
 - Cross-compilation of code
 - Emulation-as-a-Service
- Verification upon re-deployment



Evaluation result: PASS
All Significant Properties are OK. All metrics were fulfilled.
Comparison performed using following workflow execution traces

Original Workflow
ID: 70264734-cdda-4630-8ecd-27ba30f11d8f
Timestamp: 2015-04-21 13:39:03.499

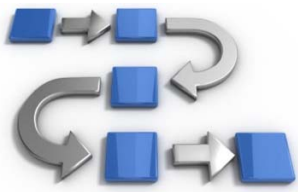
Compared Workflow
ID: 70264734-cdda-4630-8ecd-27ba30f11d8f
Timestamp: 2015-04-21 13:39:03.499

Table 1: Overview of significant properties

Significant Property	Description	Is Fulfilled
SP1_hove2_input	The workflow step hove2_input has identical outputs	True
SP2_WorkflowCorrectInputs	The inputs to the workflow are the same	True
SP3_BeanshellCopy	The workflow step BeanshellCopy has identical outputs	True
SP4_WorkflowCorrectOutputs	The outputs of the workflow are the same	True
SP5_hove2_output	The workflow step hove2_output provides the	True

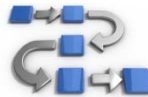
VFramework

Original environment



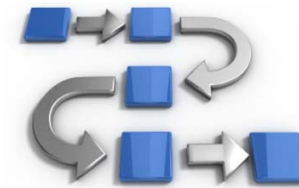
Preserve

Repository



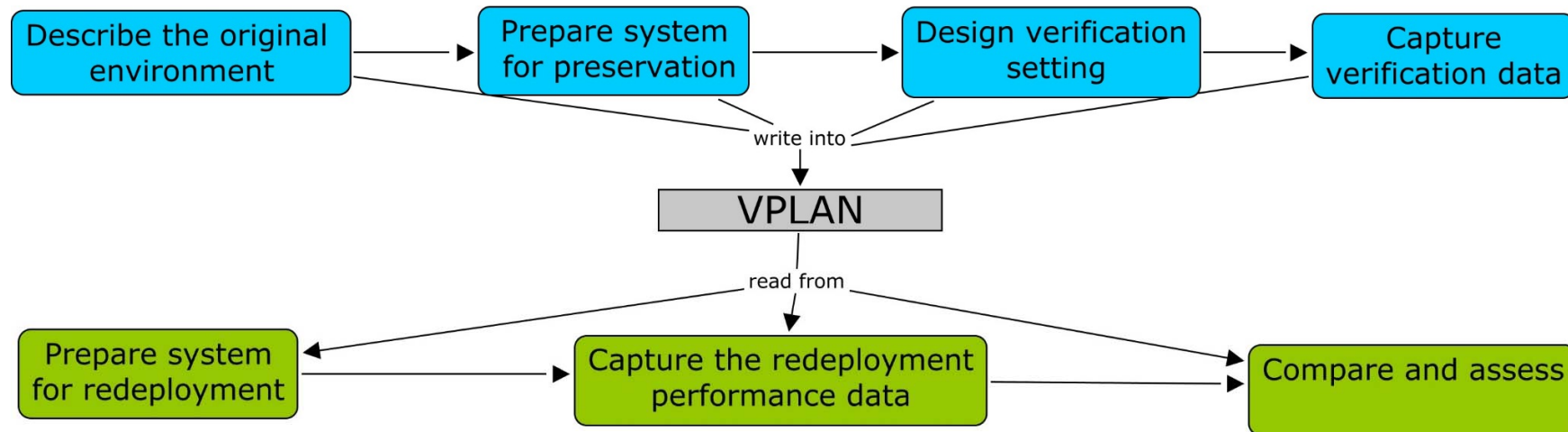
Redeploy

Redeployment environment



Are these processes the same?

VFramework



VFramework

- Documents system set-up and process execution
- Represents data in ontology
- Can be used as provenance documentation
- Can be used to verify re-execution
- Can be used to trace causes for differing behaviour
- Tomasz Miksa, Andreas Rauber. Using ontologies for verification and validation of workflow-based experiments, Web Semantics: Science, Services and Agents on the World Wide Web, 43:25-45, March 2017. <https://doi.org/10.1016/j.websem.2017.01.002>
- Tomasz Miksa, Andreas Rauber, Eleni Mina. Identifying Impact of Software Dependencies on Replicability of Biomedical Workflows. Journal of Biomedical Informatics 64:232-254, 2016. <https://doi.org/10.1016/j.jbi.2016.10.011>

Data Management & Citation, and Digital Preservation

Andreas Rauber

Vienna University of Technology
Favoritenstr. 9-11/188
1040 Vienna, Austria
rauber@ifs.tuwien.ac.at
<http://www.ifs.tuwien.ac.at/~andi>

■ Reproducibility

- What are the challenges in reproducibility?
Why do we need it? Why is it difficult?
- How to address the challenges of complex processes?
- How to ensure reproducibility with dynamic data?

■ Summary

-
- Reproducibility
 - Data Management & Citation
 - Why should we cite data?
 - How can Data Management Plans help?
 - What is so difficult about citing data?
 - How should we do it?
 - Digital Preservation
 - Summary
-

Why to cite data?

- Reproducibility on the process level
 - But: processes driven by data
- Data is the basis for almost everything
 - eScience, digital humanities,
 - Industry 4.0
 - Driving policies, society, ...
- Why should we manage & cite data?
 - Prevent scientific misconduct (“extrinsic”) ?

Why to cite data?

- Data is the basis for almost everything
 - eScience, digital humanities,
 - Industry 4.0
 - Driving policies, society, ...

- Why should we cite data?
 - Prevent Scientific misconduct (“extrinsic”) ?
 - Give credit (“altruistic”) ?
 - Show solid basis (“egoistic”) ?
 - Enable reproducibility, re-use (extrinsic + altruistic + egoistic) ?
 - Because it’s what you do if you do good work, speeding up the process of scientific discovery, efficiency! (“intrinsic”)



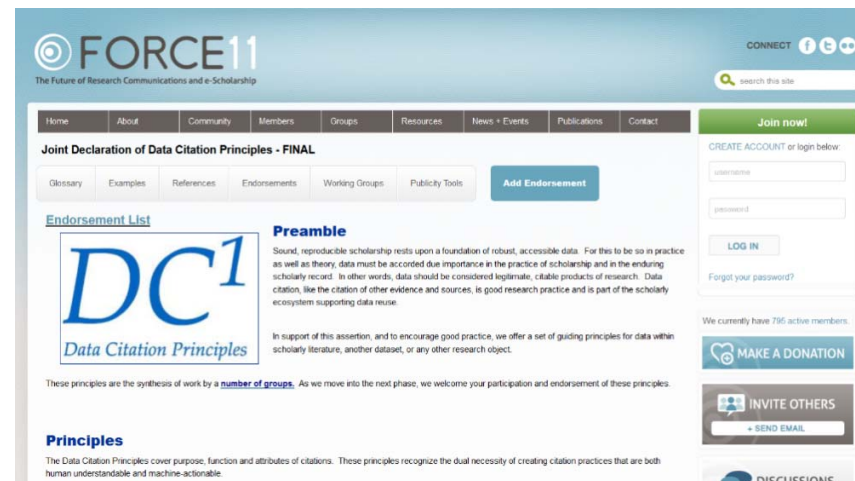
Why to cite data?

- It's what you do! – Lots of benefits
 - Makes live easier because you can build on a solid foundation
 - Speeds up the process because you can re-use existing stuff
 - Helps avoiding / detecting mistakes, improves quality, comparability
 - Reuse increases citations, visibility (“currency”)

- But:
 - To achieve this it must be easy, straightforward, “automatic”
 - Citing Papers is easy...
 - ...what about data?
(more about this later... first: “we should just do it”)

Joint Declaration of Data Citation Principles

- 8 Principles created by the Data Citation Synthesis Group
- <https://www.force11.org/datacitation>
- The Data Citation Principles cover purpose, function and attributes of citations
- Goal: Encourage communities to develop practices and tools that embody uniform data citation principles



1) Importance

Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance as publications.

2) Credit and Attribution

Data citations should facilitate giving credit and normative and legal attribution to all contributors to the data.

3) Evidence

Whenever and wherever a claim relies upon data, the corresponding data should be cited.

4) Unique Identification

A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

5) Access

Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

6) Persistence

Unique identifiers, and metadata describing the data, and its disposition, should persist - even beyond the lifespan of the data they describe.

7) Specificity and Verifiability

Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific time slice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

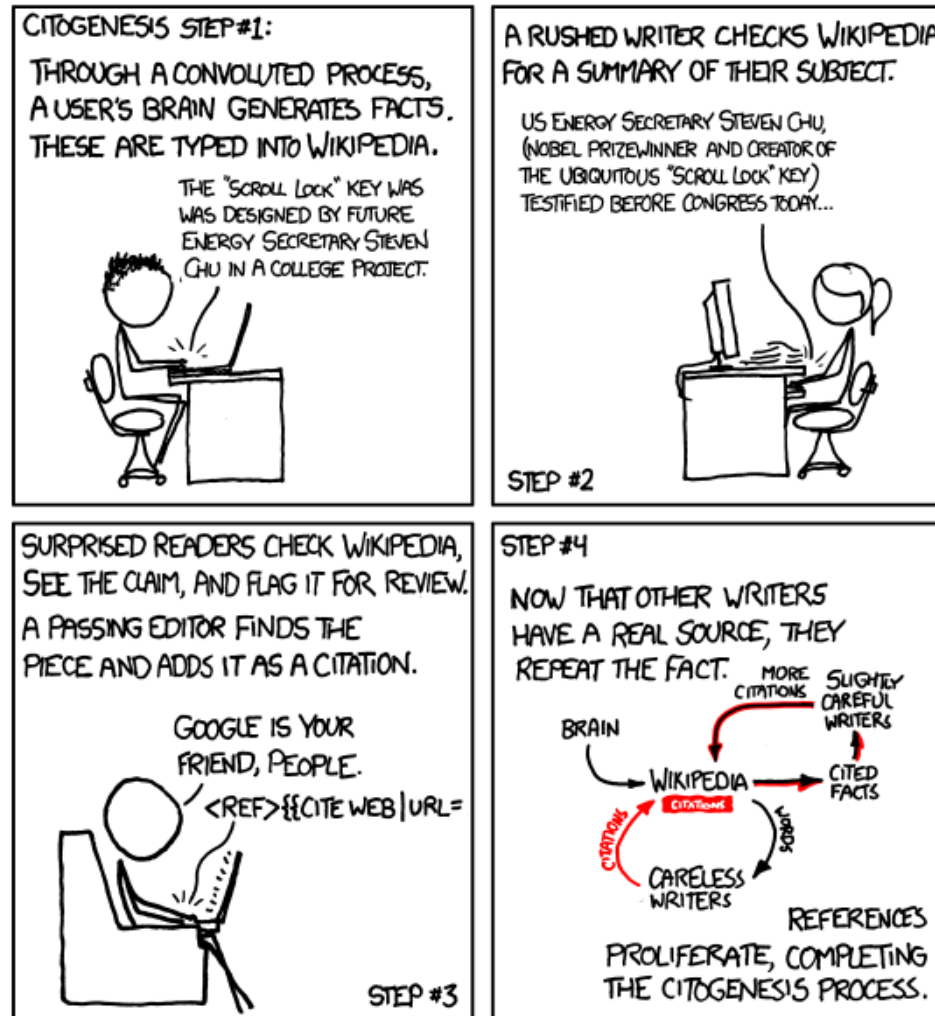
8) Interoperability and flexibility

Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

Benefits of Citation

- Identification
- Documentation
- Context
- Impact
- Transparency
- Reproducibility
- Reuse

WHERE CITATIONS COME FROM:



<https://xkcd.com/978/>

-
- Reproducibility
 - Data Management & Citation
 - Why should we cite data?
 - How can Data Management Plans help?
 - What is so difficult about citing data?
 - How should we do it?
 - Digital Preservation
 - Summary
-

Data Management

- For data citation to work we need the data
 - -> Data Management
- This should be planned early-on
 - -> Data Management Plans
- Required in most research settings
- Relevant in most industry settings as well

<https://xkcd.com/978/>

Data Management Plans

- A DMP is a brief plan to define:
 - how the data will be created
 - how it will be documented
 - who will be able to access it
 - where it will be stored
 - who will back it up
 - whether (and how) it will be shared & preserved

What is in a DMP?

- It depends...
 - on the institution requiring a DMP
 - field of research
- DMP is
 - usually a written document
 - usually has an enforced structure
- Most templates overlap
- Level of details varies
- DMP creation facilitated by
 - questionnaires, guidance documents, checklists, etc.

Common themes in DMPs

1. Description of data to be collected / created
2. Methodologies for data collection & management
3. Ethics and Intellectual Property
4. Plans for data sharing and access
5. Strategy for long-term preservation

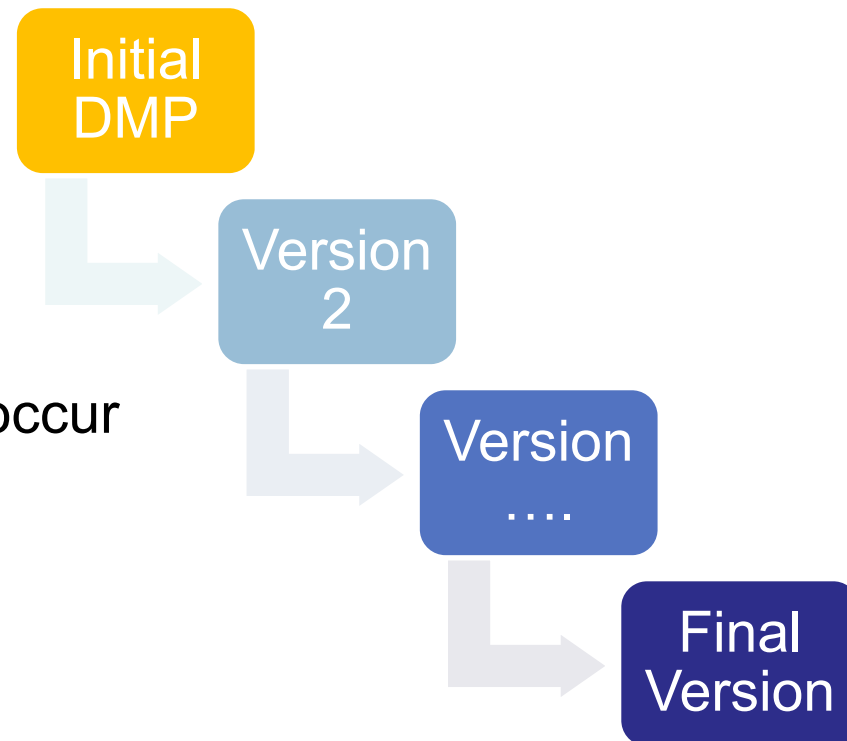
How to create a DMP?

- Most cases by
 - filling out a template
 - answering questions from a checklist
 - in most cases of limited use beyond awareness raising...
- Using software tools
 - users choose appropriate funder's template
 - only relevant questions and guidance is presented
 - results can be exported or directly submitted

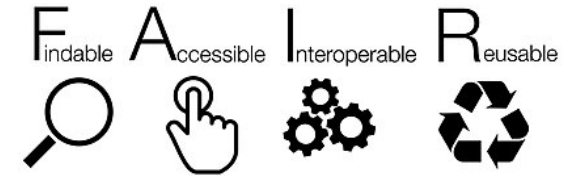


EC Horizon 2020 DMP versions

- DMP is a living document
- First version
 - within the first 6 months
- Updated versions
 - when significant changes occur
 - new datasets
 - changes in policies
 - periodic reporting
 - project reviews
 - end of project



FAIR Principles



- **Findable**
 - contains metadata that facilitates search
- **Accessible**
 - access conditions are specified
 - software needed to interpret data is known
- **Interoperable**
 - Follow standards and domain specific conventions
- **Reusable**
 - clear license and documentation
 - 'sum of the three other rules'
- **FAIR Metrics:** <http://fairmetrics.org>

Resources:

- <https://www.nature.com/articles/sdata201618>
- <https://www.go-fair.org/fair-principles/>



What should I write in fact?

- Most templates require
 - Data set description
 - Standards and metadata
 - Data sharing
 - Archiving and preservation



Data set description

- Type
 - text, spreadsheets, software, models, images, movies, audio, patient records, etc.
- Source
 - human observation, laboratory, field instruments, experiments, simulations, compilations, etc.
- Volume
 - total volume of data, number of files, etc.
- Data and file formats
 - non-proprietary formats
 - used within community

Standards and metadata

- Metadata
 - helps to understand and interpret data
 - provides details about experiment setup
 - who, when, in which conditions, tools, versions, etc.
 - helps identify and discover new data
- Use community standards to enable interoperability
 - Dublin Core
 - PREMIS
 - ...

<http://www.dcc.ac.uk/resources/metadata-standards>

Data sharing

- Which data will be shared?
 - final result?
 - intermediate data?
- Where will the data be deposited?
 - not all of the data must be shared in the same way
- Are there any embargo periods?
- Who will have access?
- Note: increasingly a trend to “data visiting”!

Archiving and preservation

- Which data needs to be preserved?
 - What has to be kept?
 - e.g. data underlying publications
 - What cannot be recreated?
 - e.g. environmental recordings
 - What is potentially useful to others?
 - What has scientific, cultural or historical value?
 - What legally must be destroyed?
- For how long?
- What is the cost and who will pay for it?

Persistent Identifiers

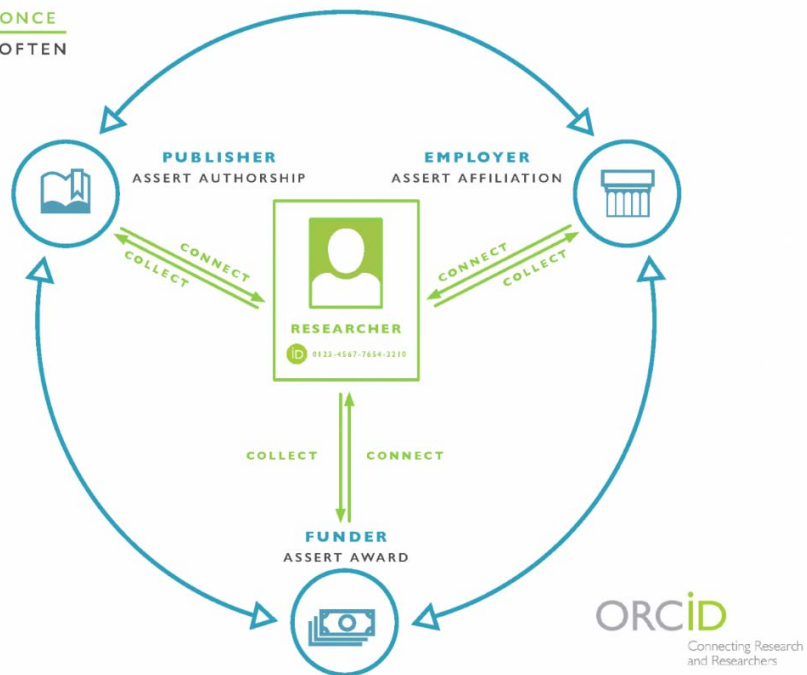
- Digital Object Identifier (DOI)
 - Uniquely identify objects
 - handle system
 - DOI assigned once
 - Physical location of data can change



<http://dx.doi.org/10.1016/j.websem.2017.01.002>

- ORCID
 - Unique user ID

ENTER ONCE
REUSE OFTEN



Tips for writing DMPs

- DMP can reveal how solid your research (proposal) is
- Seek advice - consult and collaborate
- When answering questions from checklists write coherent text
- Be specific when referring to tools and standards
- Assign responsibilities and name responsible personnel

Machine-actionable DMPs

- Writing these DMPs is tedious work
 - Researchers do not like this – neither do analysts in industry
 - It's cumbersome
 - It's error-prone
 - Institutions / repositories don't like it
 - It's error-prone
 - It's natural language text -> no automation of processes
- Need automation of DMP creation and processing
- Machine-actionable DMPs (**maDMPs**)

<https://xkcd.com/978/>

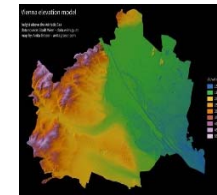
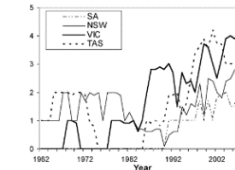
-
- Reproducibility
 - Data Management & Citation
 - Why should we cite data?
 - How can Data Management Plans help?
 - What is so difficult about citing data?
 - How should we do it?
 - Digital Preservation
 - Summary
-

Motivation

- Research data is fundamental for science/industry/...
 - Data serves as input for workflows and experiments
 - Data is the source for graphs and visualisations in publications
 - Decisions are based on data
- Data is needed for Reproducibility
 - Repeat experiments
 - Verify / compare results
- Need to provide specific data set
 - Service for data repositories

1. Put data in data repository,
2. Assign PID (DOI, Ark, URI, ...)
3. Make is accessible
→ done!?

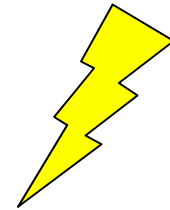
Fig. 4 The average number of high-elevation stations operating in January of the listed year. High-elevation stations are defined as those above 1500 metres in NSW and Victoria, above 1000 metres in Tasmania and above 700 metres in South Australia.



<https://commons.wikimedia.org/w/index.php?curid=30978545>

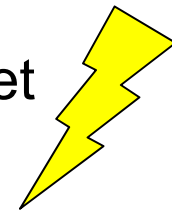
Identification of Dynamic Data

- Usually, datasets have to be static
 - Fixed set of data, no changes:
no corrections to errors, no new data being added
 - But: (research) data is **dynamic**
 - Adding new data, correcting errors, enhancing data quality, ...
 - Changes sometimes highly dynamic, at irregular intervals
 - Current approaches
 - Identifying entire data stream, without any versioning
 - Using “accessed at” date
 - “Artificial” versioning by identifying batches of data (e.g. annual), aggregating changes into releases (time-delayed!)
- Would like to identify precisely the **data as it existed at a specific point in time**



Granularity of Subsets

- What about the **granularity** of data to be identified?
 - Enormous amounts of data
 - Researchers use specific subsets of data
 - Need to identify precisely the subset used
- Current approaches
 - Storing a copy of subset as used in study -> scalability
 - Citing entire dataset, providing textual description of subset
-> imprecise (ambiguity)
 - Storing list of record identifiers in subset -> scalability,
not for arbitrary subsets (e.g. when not entire record selected)



- Would like to be able to identify precisely the **subset of (dynamic) data used** in a process

Data Citation – Requirements

- Dynamic data
 - corrections, additions, ...
- Arbitrary subsets of data (granularity)
 - rows/columns, time sequences, ...
 - from single number to the entire set
- Stable across technology changes
 - e.g. migration to new database
- Machine-actionable
 - not just machine-readable,
definitely not just human-readable and interpretable
- Scalable to very large / highly dynamic datasets
 - But: should also work for small and/or static datasets!

-
- Reproducibility
 - Data Management & Citation
 - Why should we cite data?
 - How can Data Management Plans help?
 - What is so difficult about citing data?
 - How should we do it?
 - Digital Preservation
 - Summary
-

We have: Data + Means-of-access

**Dynamic Data Citation:
Cite (dynamic) data dynamically via query!**

Steps:

1. Data → versioned (history, with time-stamps)

Researcher creates working-set via some interface:

2. Access → **store & assign PID to “QUERY”**, enhanced with

- **Time-stamping** for re-execution against versioned DB
- **Re-writing** for normalization, unique-sort, mapping to history
- **Hashing** result-set: verifying identity/correctness

leading to landing page

S. Pröll, A. Rauber. **Scalable Data Citation in Dynamic Large Databases: Model and Reference Implementation**. In IEEE Intl. Conf. on Big Data 2013 (IEEE BigData2013), 2013

http://www.ifs.tuwien.ac.at/~andi/publications/pdf/pro_ieeebigdata13.pdf

Data Citation – Deployment

- Researcher uses workbench to identify subset of data
- Upon executing selection („download“) user gets
 - Data (package, access API, ...)
 - PID (e.g. DOI) (Query is time-stamped and stored)
 - Hash value computed over the data for local storage
 - Recommended citation text (e.g. BibTeX)
- PID resolves to landing page
 - Provides detailed metadata, link to parent data set, subset,...
 - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
 - Query is re-executed against time-stamped and versioned DB
 - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

- Note: query string provides excellent subset of data
- provenance information on the data set! user gets

- Data (package)
- PID (e.g. DOI)
- Hash value
- Recommended citation text (e.g. PID-TEXT)

This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!

- PID resolves
 - Provides details
 - Option to retrieve
- Upon activating PID associated with a data citation
 - Query is re-executed against time-stamped and versioned DB
 - Results as above are returned
- Query store aggregates data usage

Identify which parts of the data are used.
If data changes, identify which queries
(studies) are affected

Data Citation – Recommendations

Preparing Data & Query Store

- R1 – Data Versioning
- R2 – Timestamping
- R3 – Query Store

When Resolving a PID

- R11 – Landing Page
- R12 – Machine Actionability

When Data should be persisted

- R4 – Query Uniqueness
- R5 – Stable Sorting
- R6 – Result Set Verification
- R7 – Query Timestamping
- R8 – Query PID
- R9 – Store Query
- R10 – Citation Text

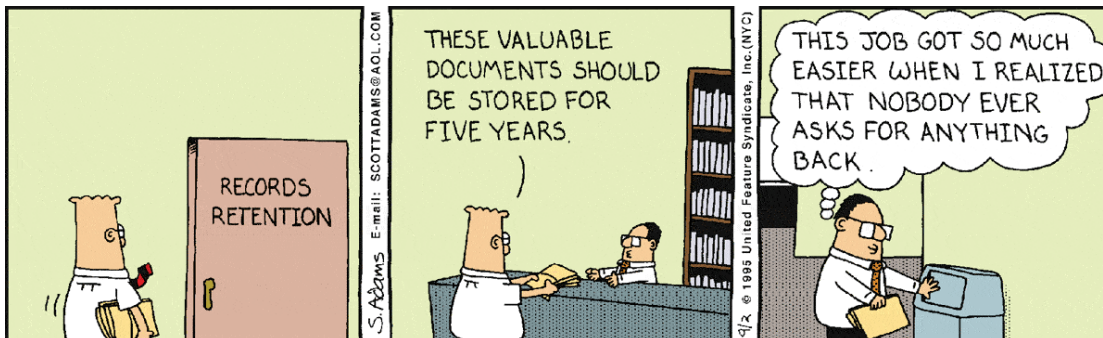
Upon Modifications to the Data Infrastructure

- R13 – Technology Migration
- R14 – Migration Verification



R1: Data Versioning

- **Apply versioning to ensure earlier states of the data can be retrieved**
- Versioning allows tracing the changes (static data: no changes – principle still applies)
- No in-place updates or deletes
 - Mark record as deleted, re-insert new record instead of update
 - Keep old versions – only way to be able to “go back”
- Do we really need to keep everything?
 - (*“changes that were never read never existed”*)



Src: <http://dilbert.com/strip/1995-09-02>

R2: Data Timestamping

- **Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp**
- Timestamping is closely related to versioning
- Granularity depends on
 - Change frequency / tracking requirements
 - Per individual operation
 - Batch-operations
 - Grouped in-between read accesses
(*“changes that were never read do not matter”*)
 - System (data storage, databases)
 - e.g. FAT 2 seconds, NTFS 100 ns, EXT4 1 ns



https://www-03.ibm.com/ibm/history/exhibits/cc/cc_T30.html

R1 & R2: Versioning / Timestamping

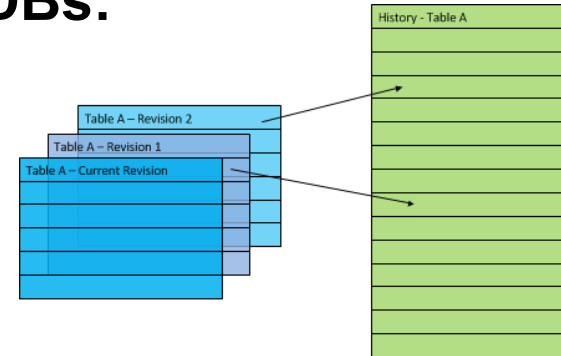
Note:

- R1 & R2 are already pretty much standard in many (RDBMS-) research databases
- Different ways to implement, depending on
 - data type / data structure: RDBMS, CSV, XML, LOD, ...
 - data volume
 - amount and type of changes
 - number of APIs, flexibility to change them
- Distributed settings:
 - synchronized clocks, or:
 - each node keeps individual, local timetime-stamps for distributed queries based on local times

R1 & R2: Versioning / Timestamping

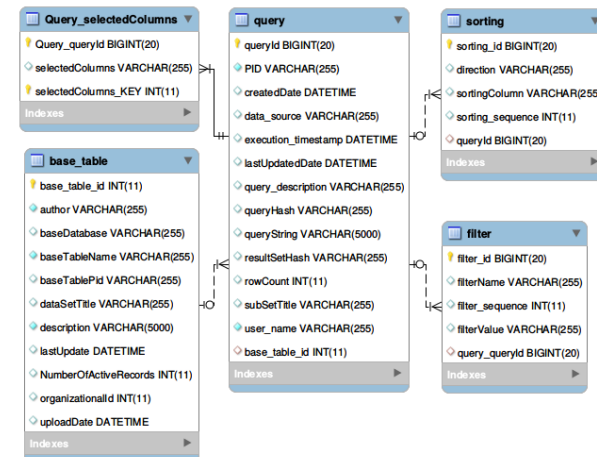
Implementation options for e.g. relational DBs:

- History Table
 - Utilizes full history table
 - Also inserts reflected in history table
 - Doubles storage space, no API adaptations
- Integrated
 - Extend original tables by temporal metadata
 - Expand primary key by timestamp/version column
 - Minimal storage footprint, changes to all APIs
- Hybrid
 - Utilize history table for deleted record versions with metadata
 - Original table reflects latest version only
 - Minimal storage footprint, some API change, expensive query re-writes
- Solution to be adopted depends on trade-off
 - Storage Demand
 - Query Complexity
 - Software/API adaption



R3: Query Store

- Provide means for storing queries and the associated metadata in order to re-execute them.
- Approach is based upon queries.
 - Therefore we need to preserve the queries
 - Original and re-written (**R4**, **R5**), potentially migrated (**R13**)
 - Query parameters and system settings
 - Execution metadata
 - Hash keys (multiple, if re-written) (**R4**, **R6**)
 - **Persistent identifier(s)** (**R8**)
 - Citation text (**R10**) ...
- Comparatively small, even for high query volumes



R4: Query Uniqueness

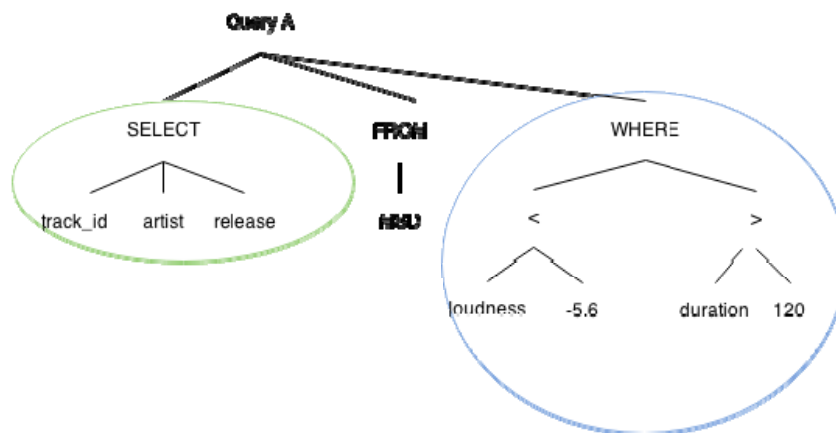
- **Re-write the query to a normalized form so that identical queries can be detected.**
Compute checksum of the normalized query to efficiently detect identical queries
- Detecting identical queries can be challenging
 - Query semantics can be expressed in different ways
 - Different queries can deliver identical results
 - Interfaces can be used for maintaining a stable query structure
- Best effort, no perfect solution
- Usually not a problem if queries generated via standardized interfaces, e.g. workbench – optional!
- Worst case: two PIDs for semantically equivalent queries

R4: Query Uniqueness

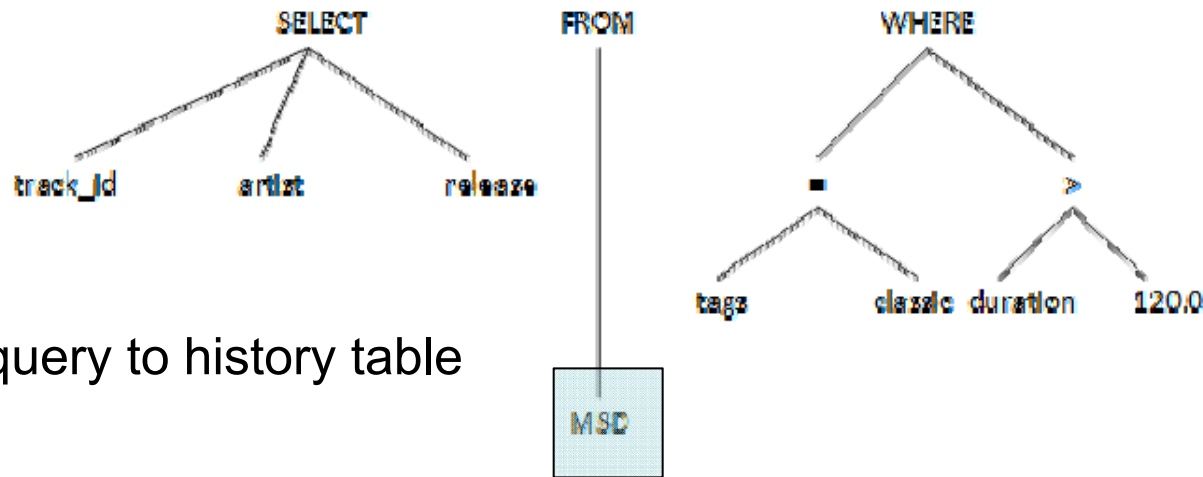
- Query re-writing needed to
 - **Standardization/Normalization** of query to help with identifying semantically identical queries
 - upper/lower case spelling, sorting of filter parameters, ...
 - Re-write to **adapt to versioning approach** chosen (versioning in operational tables, separate history table, ...), e.g. **identify last change to result set touched** upon (i.e. select including elements marked deleted, check most recent timestamp, to determine correct PID assignment)
 - **Add timestamp** to any select statement in query
 - **Apply unique sort** to any table touched upon in query prior to query to ensure unique sort (see **R5**)

R4: Query Uniqueness

- Normalizing queries to detect identical queries
 - WHERE clause sorted
 - Calculate query string hash
 - Identify semantically identical queries
 - → non-identical queries: columns in different order



R4: Query Uniqueness



- Adapt query to history table

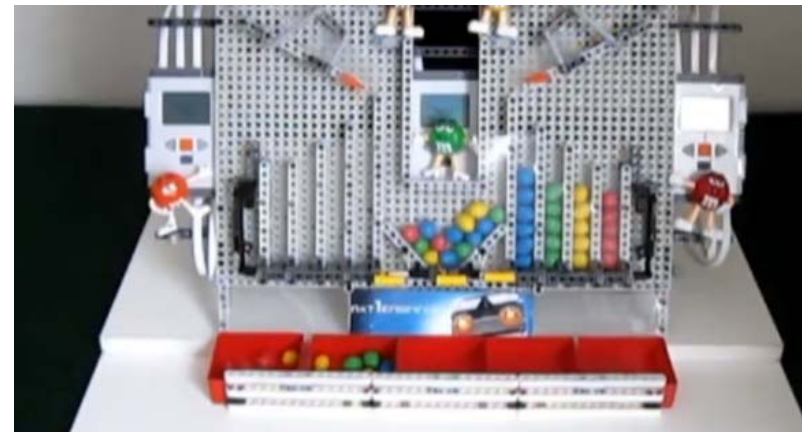
```

SELECT results.track_id, results.artist, results.release
FROM MSD AS results JOIN (
  SELECT track_id, max(timestamp) AS latestTimestamp
  FROM MSD
  WHERE timestamp <= (SELECT @queryExecutionTimestamp)
  AND (track_id NOT IN
    (SELECT track_id FROM MSD AS deletedRecords
     WHERE deletedRecords.status_mark = 'deleted'
     AND (deletedRecords.timestamp < @queryExecutionTimestamp))
  )
  GROUP BY track_id
) AS version ON results.track_id = version.track_id AND results.timestamp = version.latestTimestamp

WHERE
  results.tags = 'classic' AND results.duration > 120
ORDER BY results.track_id;
  
```

R5: Stable Sorting

- **Ensure that the sorting of the records in the data set is unambiguous and reproducible**
- The sequence of the results in the result set may not be fixed, but data processing results may depend on sequence
 - Many databases are set based
 - The storage system may use non-deterministic features
- If this needs to be addressed, apply default sort (on id) prior to any user-defined sort
- Optional!



<http://www.geek.com/>

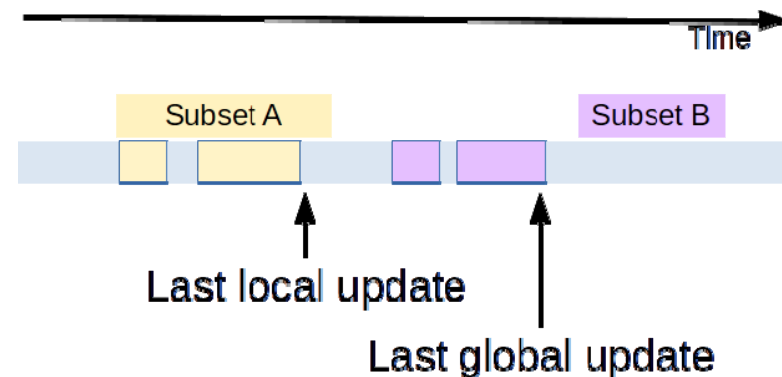
R6: Result Set Verification

- **Compute fixity information (also referred to as checksum or hash key) of the query result set to enable verification of the correctness of a result upon re-execution.**
- **Correctness:**
 - No record has changed within a data subset
 - All records which have been in the original data set are also in the re-generated data set
- **Compute a hash key**
 - Allows to compare the completeness of results
 - For extremely large result sets: potentially limit hash input data, e.g. only row headers + record id's



R7: Query Timestamping

- Assign a timestamp to the query based on the last update to the entire database (or the last update to the selection of data affected by the query or the query execution time).
- Allows to map the execution of a query to a state of the database
 - Execution time: default solution, simple, potentially privacy concerns?
 - Last global update: simple, **recommended**
 - Last update to affected subset: complex to implement
- All equivalent in functionality! (transparent to user)



R8: Query PID

- **Assign a new PID to the query if either the query is new or if the result set returned from an earlier identical query is different due to changes in the data. Otherwise, return the existing PID of the earlier query to the user.**
- **Existing PID:** Identical query (semantics) with identical result set, i.e. no change to any element touched upon by query since first processing of the query
- **New PID:** whenever query semantics is not absolutely identical
(irrespective of result set being potentially identical!)

R8: Query PID

- Note:
 - Identical result set alone does not mean that the query semantics is identical
 - Will assign different PIDs to capture query semantics
 - Need to normalize query to allow comparison
- Process:
 - Re-write query to adapt to versioning system, stable sorting, ...
 - Determine query hash
 - Execute user query and determine result set hash
 - Check query store for queries with identical query hash
 - If found, check for identical result set hash

R9: Store the Query

- **Store query and metadata (e.g. PID, original and normalised query, query and result set checksum, timestamp, superset PID, data set description, and other) in the query store.**
 - Query store is central infrastructure
 - Stores query details for long term
 - Provides information even when the data should be gone
 - Responsible for re-execution
 - Holds data for landing pages
 - Stores sensitive information

R10: Create Citation Texts

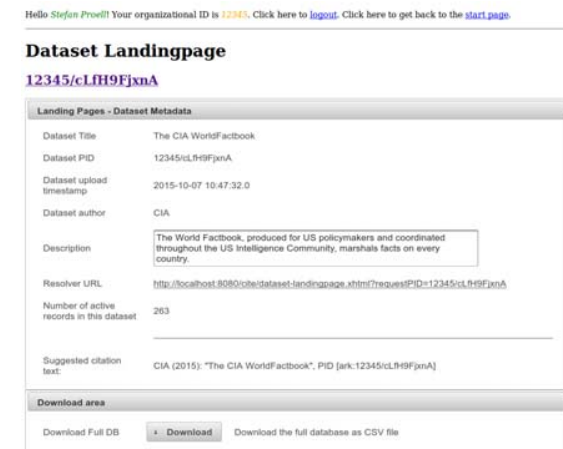
- **Generate citation texts in the format prevalent in the designated community for lowering the barrier for citing and sharing the data.**
Include the PID in the citation text snippet.
- Researchers are “lazy”/efficient
 - Support citing by allow them to copy and paste citations for data
 - Citations contain text including PIDs and timestamps
 - Adapted for each community
- **2 PIDs!**
 - Superset: the “database” and it’s holder (repository, data center)
 - Subset: based on the query
 - Accumulate credits for subset and (dynamic) data collection/holder

Suggested citation
text:

Stefan Proell (2015) "Austria Facts" created at 2015-10-07 10:51:55.0, PID
[ark:12345/qmZi2wO2vv]. Subset of CIA: "The CIA WorldFactbook", PID
[ark:12345/cLfH9FjxnA]

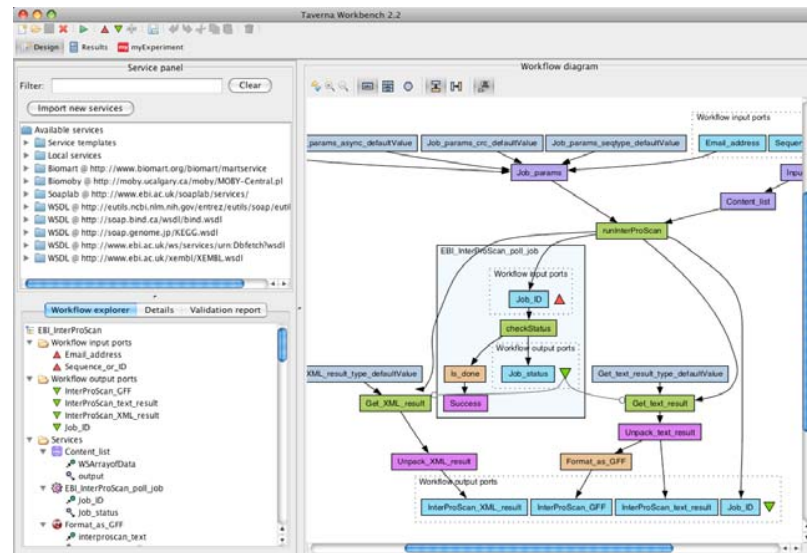
R11: Landing Page

- **Make the PIDs resolve to a human readable landing page that provides the data (via query re-execution) and metadata, including a link to the superset (PID of the data source) and citation text snippet.**
 - Data sets and subsets uniquely identifiable by their PID, which resolves to a human readable landing page.
 - Landing page reachable by a unique URL, presented in a Web browser
 - Not all information needs to be provided on landing page (e.g. query strings frequently not relevant / potential security threat)



R12: Machine Actionability

- **Provide an API / machine actionable landing page to access metadata and data via query re-execution.**
 - Experiments are increasingly automated
 - Machines most likely to consume data citations
 - Allows machines to resolve PIDs, access metadata and data
 - Note: does NOT imply full / automatic access to data!
 - Authentication
 - Load analysis
 - Handshake, content negotiation, ...
 - Allows automatic meta-studies, monitoring, ...



R13: Technology Migration

- **When data is migrated to a new representation (e.g. new database system, a new schema or a completely different technology), migrate also the queries and associated fixity information.**
 - Technology evolves and data may be moved to a new technology stack
 - Query languages change
- **Migration required**
 - Migrate data and the queries (both are with the data center!)
 - Adapt versioning, re-compute query hash-keys
 - Maybe decide to keep “original” queries in the provenance trace
- **Note: such data migrations constitute major projects, usually happen rarely – require all APIs to be adapted, ...**

R14: Migration Verification

- **Verify successful data and query migration, ensuring that queries can be re-executed correctly.**
- Sanity check: After migration is done, verify that the data can still be retrieved correctly
- Use query and result set hashes in the query store to verify results
- If hash function is incompatible/cannot be computed on new system as hash input data sequence cannot be obtained, pairwise comparison of subset elements
 - May constitute new PID / data subset in this case, as subsequent processes will not be able to use it as input if result set presentation has changed, breaks processes

RDA Recommendations - Summary

- Building blocks of supporting dynamic data citation:
 - Uniquely identifiable data records
 - Versioned data, marking changes as insertion/deletion
 - Time stamps of data insertion / deletions
 - “Query language” for constructing subsets
- Add modules:
 - Persistent query store: queries and the timestamp (either: <when issued> or <of last change to data>)
 - Query rewriting module
 - PID assignment for queries that enables access
- Stable across data source migrations (e.g. diff. DBMS), scalable, machine-actionable

RDA Recommendations - Summary

■ *Benefits*

- Allows **identifying, retrieving and citing the precise data subset** with minimal storage overhead by only storing the versioned data and the queries used for extracting it
- Allows retrieving the data both **as it existed** at a given point in time as well as the **current view** on it, by re-executing the same query with the stored or current timestamp
- It allows to cite even an **empty set**!
- The query stored for identifying data subsets provides valuable **provenance data**
- Query store collects **information on data usage**, offering a basis for data management decisions
- **Metadata** such as checksums support the verification of the correctness and **authenticity** of data sets retrieved
- The same principles work for **all types of data**

■ *Some considerations and questions*

- May data be deleted?
Yes, of course, given appropriate policies. Queries may then not be re-executable against the original timestamp anymore
- Does the system need to store every query?
No, only data sets that should be persisted for citation and later re-use need to be stored.
- Can I obtain only the most recent data set?
Queries can be re-executed with the original timestamp or with the current timestamp or any other timestamp desired.
- Which PID system should be used?
Any PID system can, in principle, be applied according to the institutional policy.

-
- Reproducibility
 - Data Management & Citation
 - Digital Preservation
 - What are the Challenges in Digital Preservation?
 - How can we address them?
 - Summary
-

Why do we need Digital Preservation?

Questions / discussion:

- What is *Digital Preservation*?

Why do we need Digital Preservation?

1. Physical Preservation (Bit-stream preservation)

- Transferring to current storage systems
 - note: transfer may not be trivial
(file systems, encodings, relative references, copy protection,...)
- Ensure redundancy
 - technologically
 - geographic spread
- Access, security
- Error detection, recovery, disaster planning

Why do we need Digital Preservation?

2. Logical Preservation

- Digital Objects require specific environment to be accessible :
 - Files need specific programs
 - Programs need specific operating systems (-versions)
 - Operating systems need specific hardware components
- SW/HW environment is not stable:
 - Files cannot be opened anymore
 - Embedded objects are no longer accessible/linked
 - Programs won't run
 - Information in digital form is lost
(usually total loss, no degradation)
- Digital Preservation aims at maintaining digital objects authentically usable and accessible for long time periods.

Why do we need Digital Preservation?

3. Semantic Layer: information object

- How to interpret the data (information?) in the objects?
 - terminology changes:
changes in country names, borders, connotation of words,...
 - concept changes:
drunk driving: before 1998: 0.8‰ , afterwards 0.5‰
 - transformations: currencies/exchange rates, sensor resolutions,
 - provenance: actions applied to objects
sources: who? / which sensor?, transformations, post-processing
 - context of objects:
understanding the context of decisions, side-effects, quotations,
calibration timestamps
- For preserving digital information, all 3 layers
need to be addressed

Why do we need Digital Preservation

- The goal of Digital Preservation is to **maintain digital objects accessible and usable in an authentic manner for a long term** into the future.

Digital Preservation - Summary

- Is a complex task
- Requires a concise understanding of the objects, their intellectual characteristics, the way they were created and used and how they will most likely be used in the future
- Requires a continuous commitment to preserve objects to avoid the „digital dark hole“
- Requires a solid, trusted infrastructure and workflows to ensure digital objects are not lost
- Is essential to maintain electronic publications & data accessible
- Will become more complex as digital objects become more complex
- Needs to be defined in a preservation plan