DAUPHINE UNIVERSITÉ PARIS | PSL

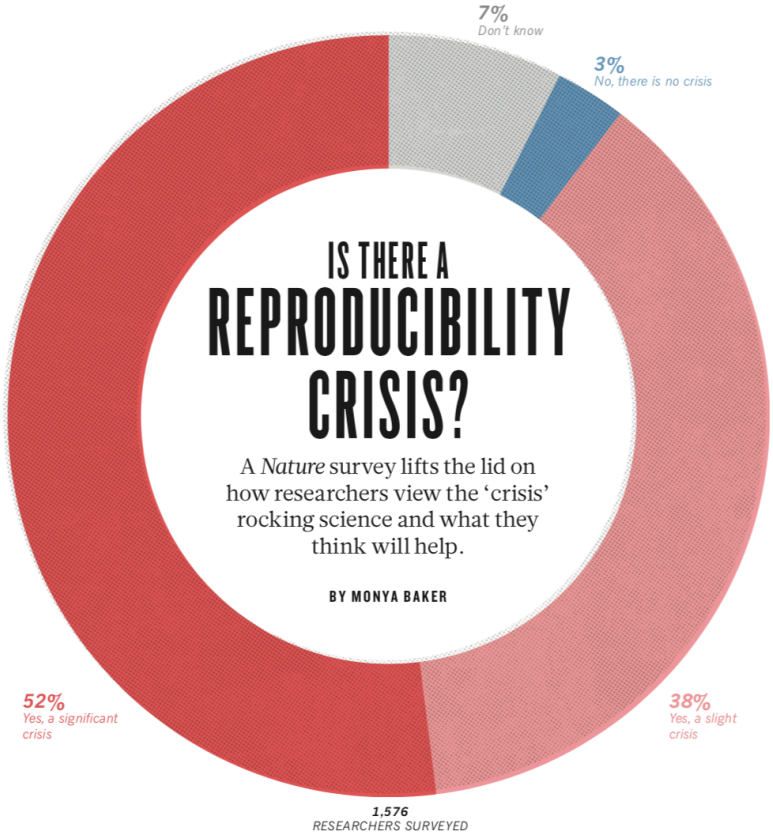
LAMSADE
UMR CNRS 7243
laboratoire d'analyse et modélisation de systèmes pour l'aide à la décision

The State of Reproducibility in Computational and Data-Driven Research?

Khalid Belhajjame
kbelhajj@gmail.com

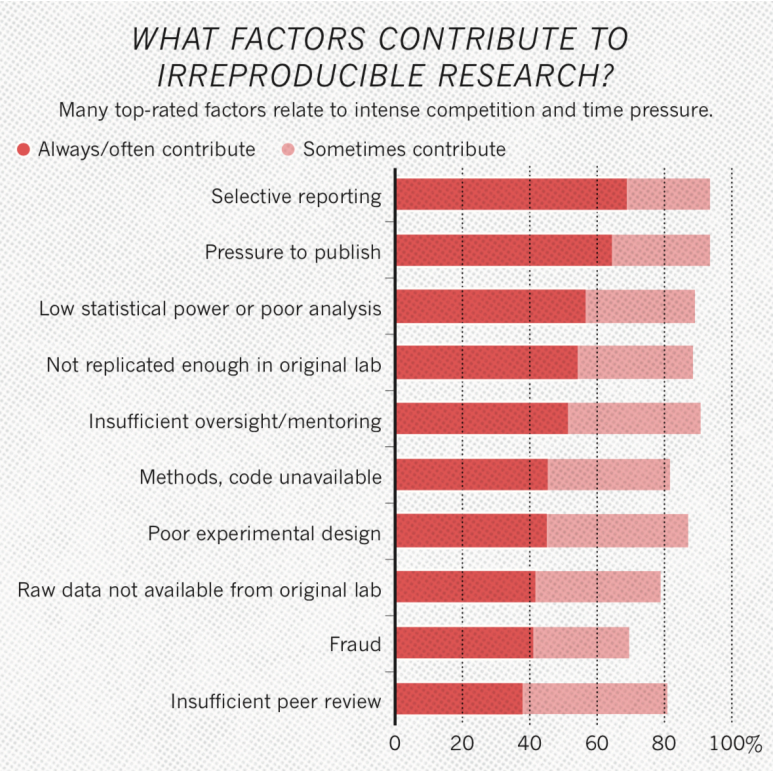
We Know That There is A Crisis in Experimental Sciences

More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments.



Researchers attitude towards reproducibility

73% said that they think that at least half of the papers in their field can be trusted, with physicists and chemists generally showing the most confidence.



Computational and Data-Driven Research

COMPUTERS IN SCIENCE

Publisher
George Laughead

Managing Editor
Michael J. Comendul

Associate Editor
Bud Sadler

Art Director
Christine Destrempes

Editorial Board
Consulting Editors
David Pope
Jeffrey N. Bairstow

Contributing Editors
John W. Root, Ph.D.
Diana J. Gabaldon, Ph.D.
Eugene F. Mallove, Sc.D.
Mickey Williamson

Advertising Sales
Raymond Low
603/924-9471

James M. Burns
415/328-3470

Advertising Coordinator
Cornelia Taylor
603/924-9471

CW Communications/Peterborough
President
James S. Povec

Vice President/Finance
Roger Murphy

Director of Operations
Matt Smith

Corporate Creative Director
Christine Destrempes

Guest Editorial

The Future of Scientific Computing

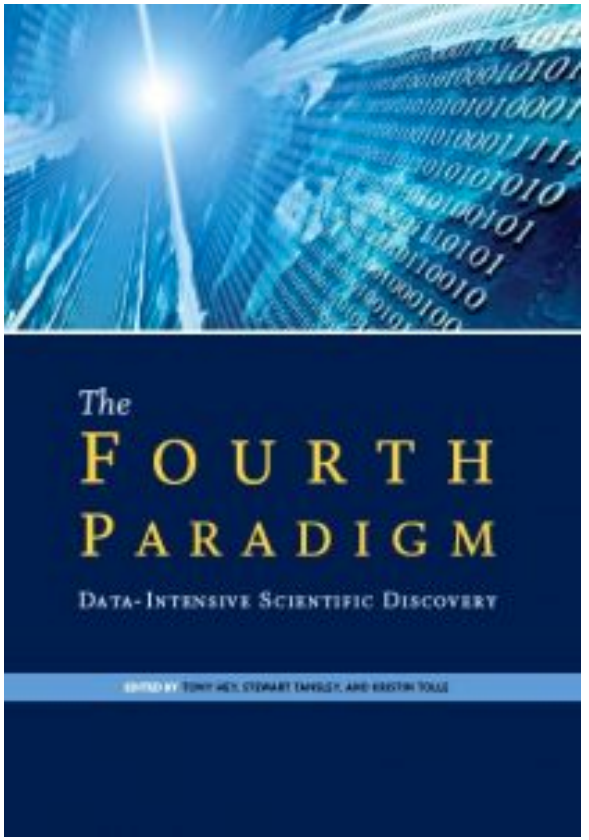
by C. Gordon Bell

Forty-one years after the birth of ENIAC— the first electronic computer—computers are still in their infancy. We are on the verge of a true revolution, when we will see the computer itself "doing science." In the next decade advances in computer-assisted science should dwarf the past historical accomplishments of scientific computing. Ken Wilson, Cornell University's Nobel laureate, points out that computational science is now the third paradigm of science, supplementing theory and experimentation.

This powerful computational science has only recently emerged with the development of the large-scale supercomputer able to carry out over 1 billion floating-point operations per computer. With over 200 times the power of the VAX and 60,000 times the power of a personal computer, the emergence of a supercomputer offers a significant qualitative and structural change in the way science is carried out.

Computers In Science is choosing a propitious moment to begin its chronicle of computer-assisted science. Every field of science is changing—molecular chemistry, biology (computational molecular biology), materials structures, astrophysics (in effect a computational observatory), and every facet of large-scale engineering—all because of the enhanced capabilities of computing.

In the future the scientific computer will simulate new classes of phenomena such as the interaction of mole-



Computational is now considered as the third paradigm of science

Data-Driven Research is the fourth paradigm of science

Computational and Data-Driven Research

- Most of published discoveries today have a computational component.
- Hypothesis-driven research gave way to data-driven research:
- Data are used in the early stages of the research to:
  - Data is not used to simply test the validity or verify a hypothesis at the later stages of a research, but is used in early stages to:
    - Learn insights
    - Detect Correlation
    - Learn Models
    - Check feasibility

Why Care About reproducibility For Computational and Data-Driven Research?

- Verification (repeatability) to increase Trust
- This is a good reason, but is somewhat pointless from the scientific discovery point of view in the sense that we are not reaching new insights

Why Care About reproducibility For Computational and Data-Driven Research?

- Verification (repeatability) to increase Trust
- This is a good reason, but is somewhat pointless from the scientific discovery point of view in the sense that we are not reaching new insights
- Well that is not completely true …
- By making computational research reproducible we have some concrete benefits, by facilitating:
  - Reuse
  - Comparison
  - Debug Errors
  - Allows for constructive and guided scientific discussions

How Can a Computational And Data-Driven Research Made Reproducible?

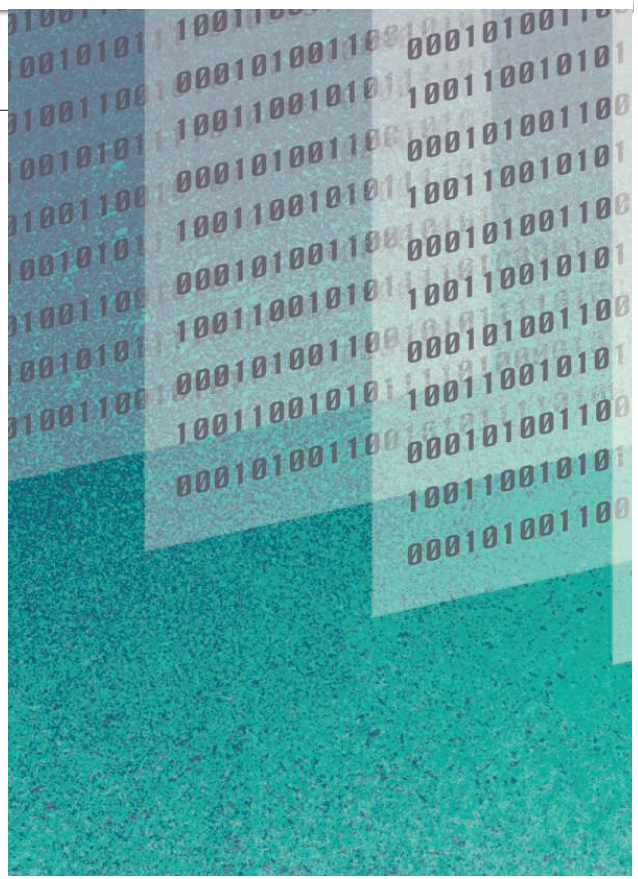
REPRODUCIBILITY

Enhancing reproducibility for computational methods

Data, code, and workflows should be available and cited

By **Victoria Stodden,[1] Marcia McNutt,[2] David H. Bailey,[3] Ewa Deelman,[4] Yolanda Gil,[4] Brooks Hanson,[5] Michael A. Heroux,[6] John P.A. Ioannidis,[7] Michela Taufer[8]**

Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transparency in disclosure of computational methods. Current reporting methods are often uneven, to understanding how computational results were derived and to reconciling any differences that might arise between independent replications (*4*). We thus focus on the ability to rerun the same computational steps on the same data the original authors used as a minimum dissemination standard (*5*, *6*), which includes workflow information that explains what raw data and intermediate results are input to which computations (*7*). Access to the data and code that underlie discoveries can also enable downstream scientific contributions, such as meta-analyses, reuse, and other efforts that include results from multiple studies.



Sufficient metadata should be provided for someone in the field to use the shared digital scholarly objects without resorting to contacting the original authors (i.e., http://bit.ly/2fVwjPH). Software metadata should include, at a minimum, the title, authors,

How Can a Computational And Data-Driven Research Made Reproducible?

- To answer this question in a systematic manner considering the different fields of computational sciences, I decided to perform an umbrella review.

- Umbrella review refers to review compiling evidence from multiple reviews into one accessible and usable review. Focuses on broad condition or problem for which there are competing interventions and highlights reviews that address these interventions and their results [Grant and Booth, 2009].

Types of Papers Included in the Study

- Systematic Reviews with a focus on computational reproducibility
- The reviews included usually cover a specific scientific module (e.g. Computational simulation, biomechanics, etc.)
- We also considered papers that attempts to reproduce/repeat existing solutions.

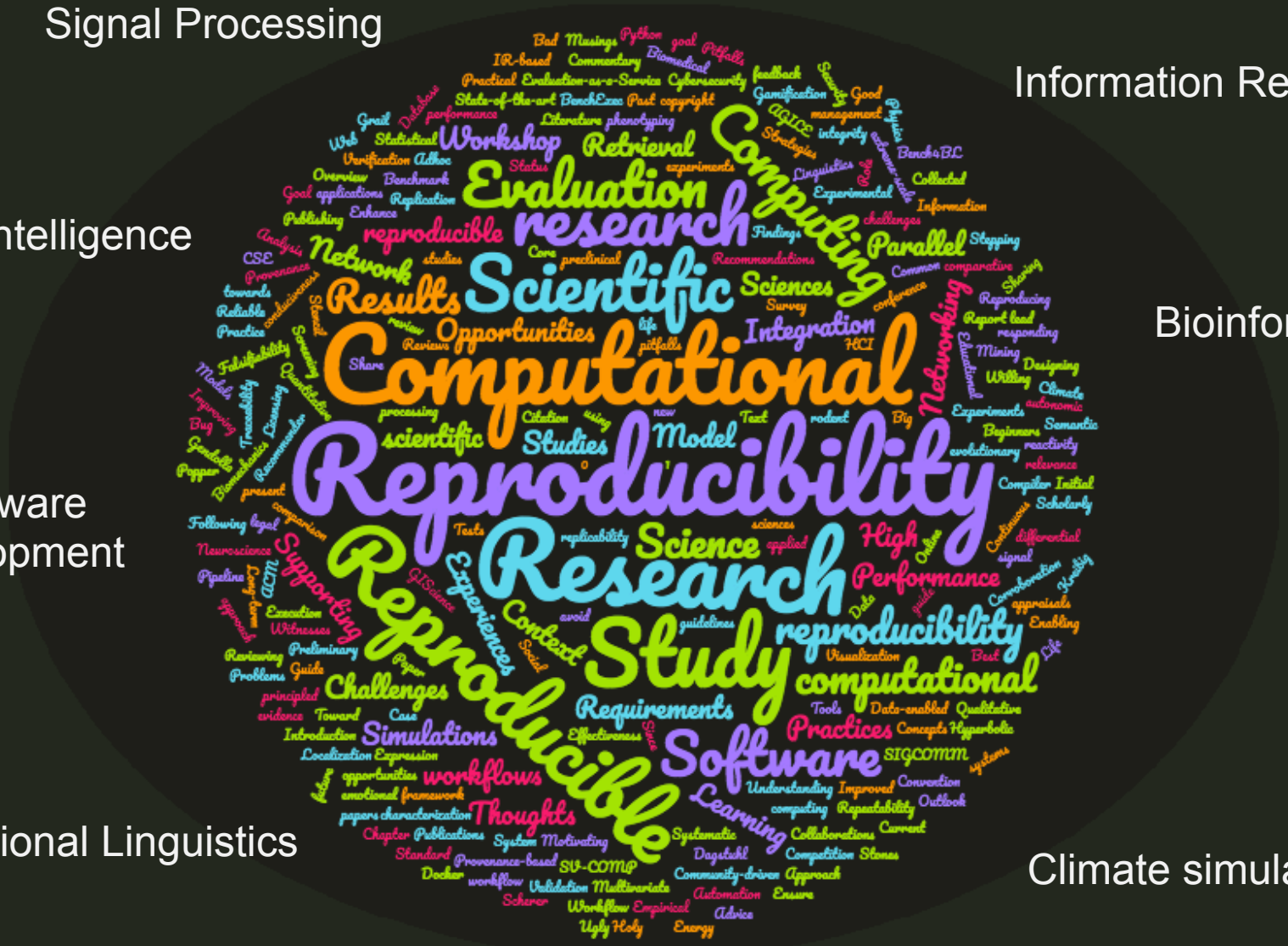
Papers Selected

- We used three digital libraries
  - ACM DL,
  - IEEE Xplore DL, and
  - ScienceDirect
- We confined our search to papers published in the last ten years: 2009-2019
- We manually filtered the papers and selected 51 ones to examine

Signal Processing

Information Retrieval

Artificial Intelligence



Software development

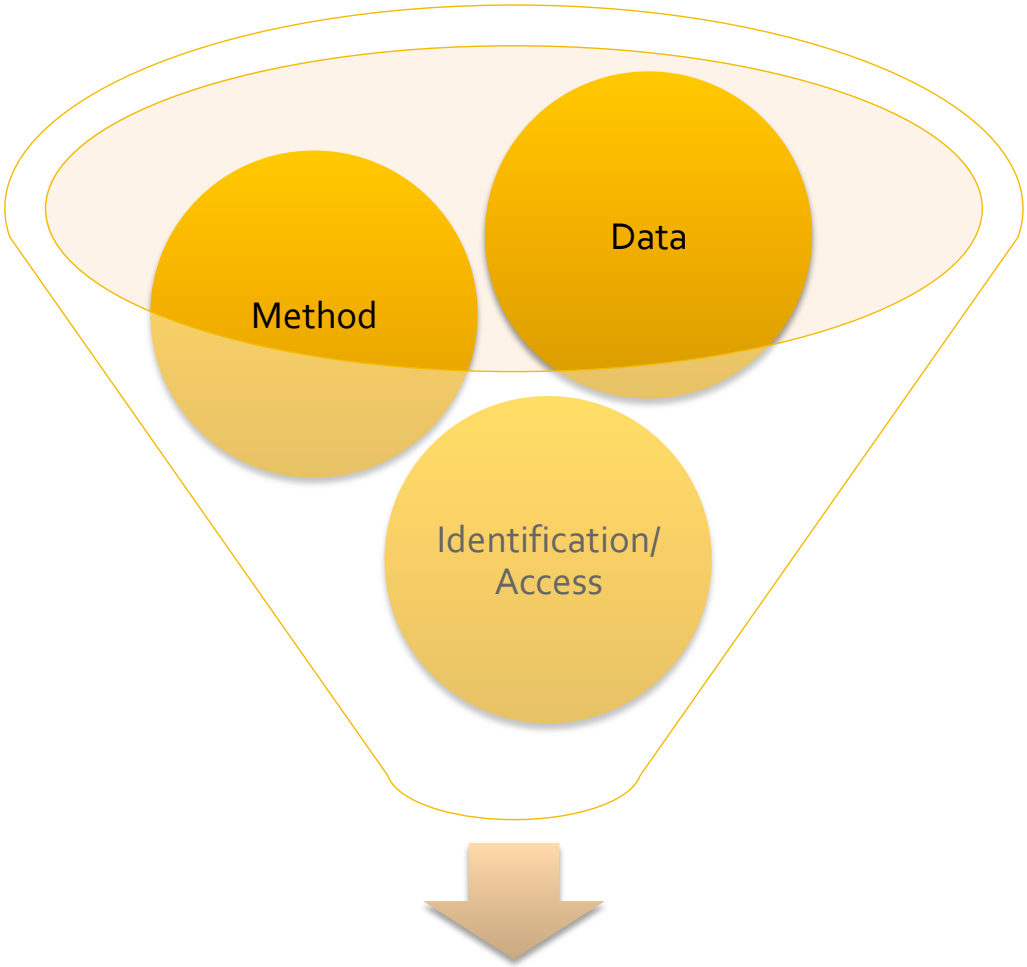
Bioinformatics

Computational Linguistics

Climate simulations

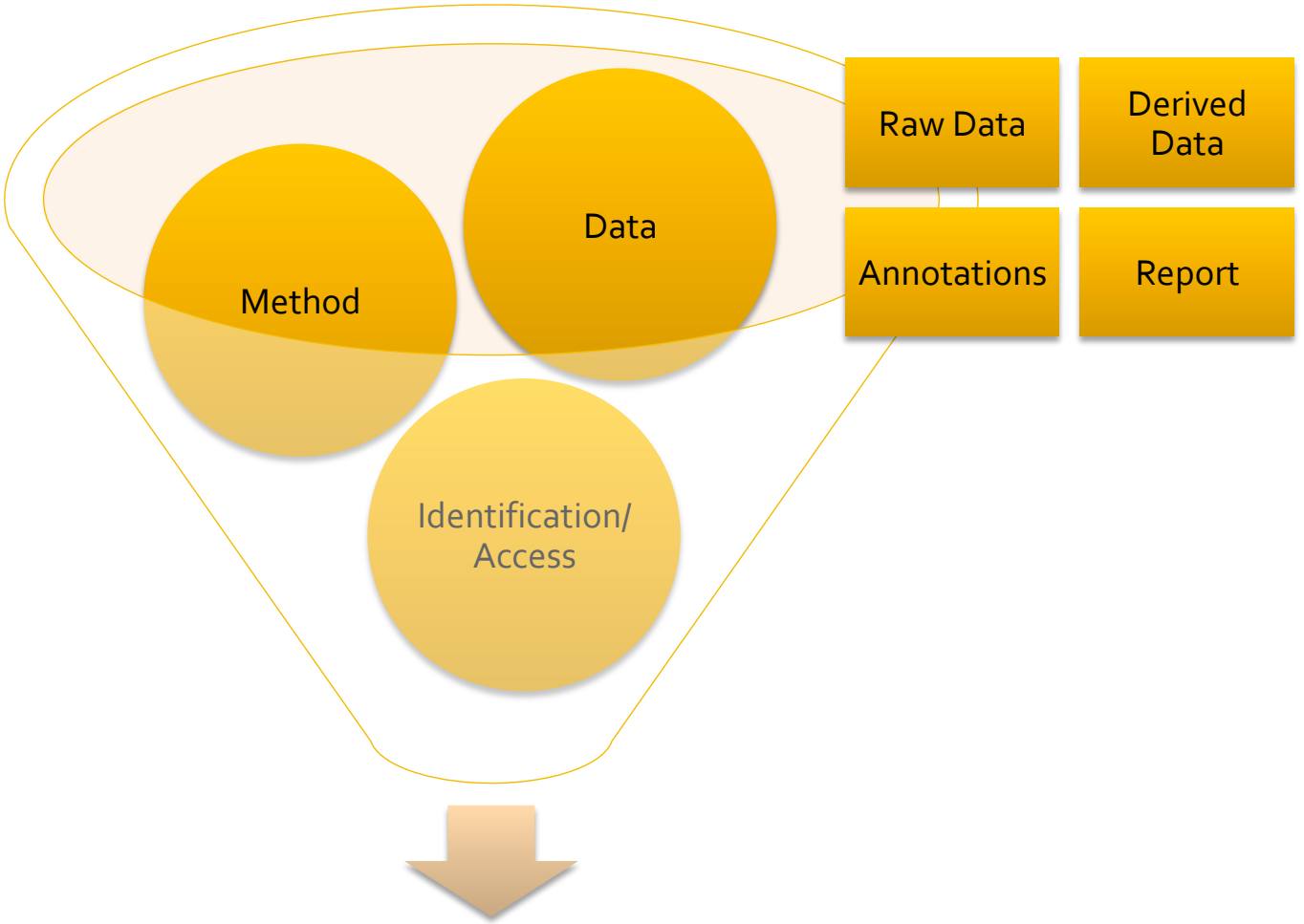
Wordcloud obtained using the titles of the selected articles

Dimensions



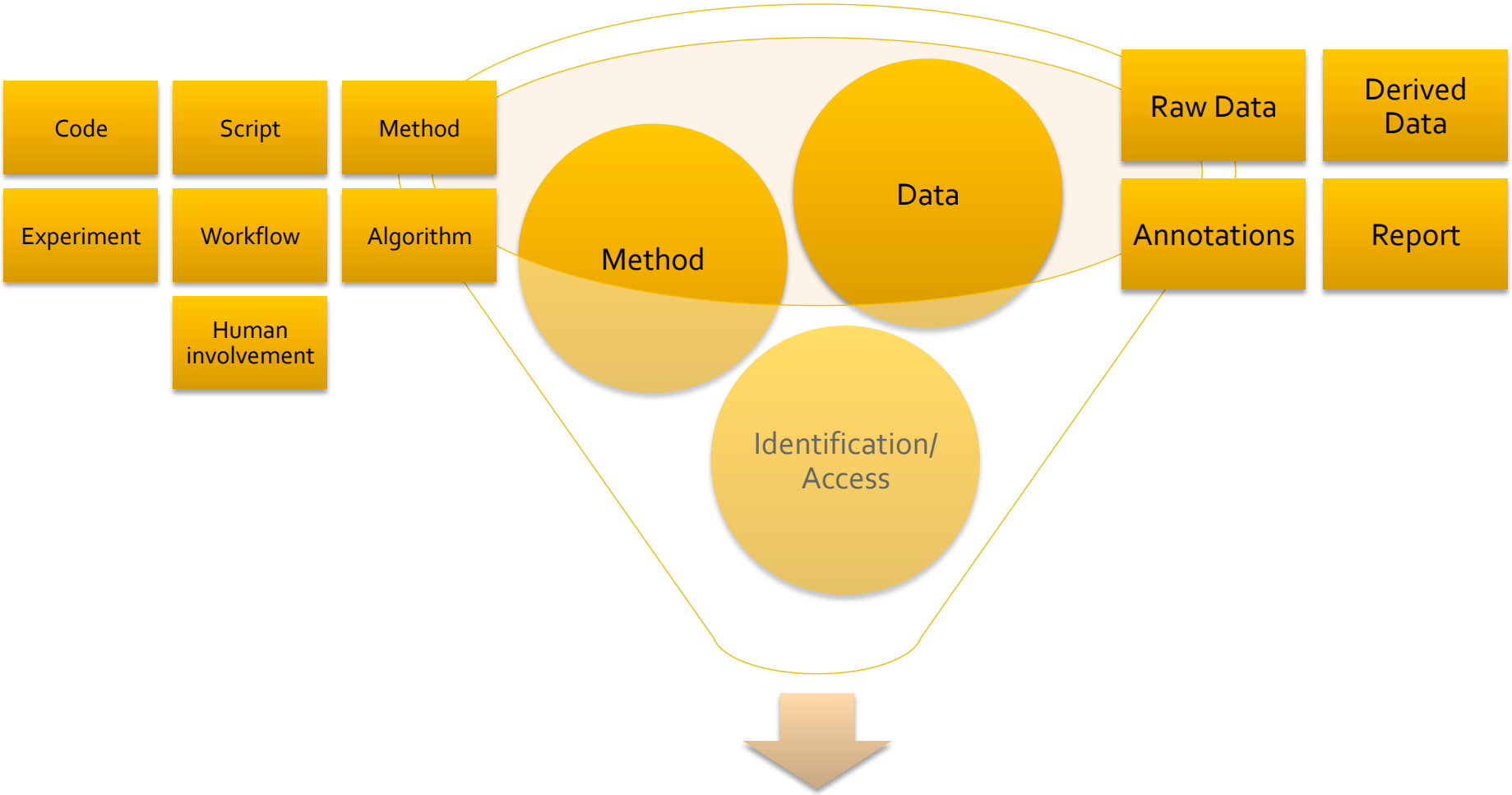
Reproducibility

Dimensions



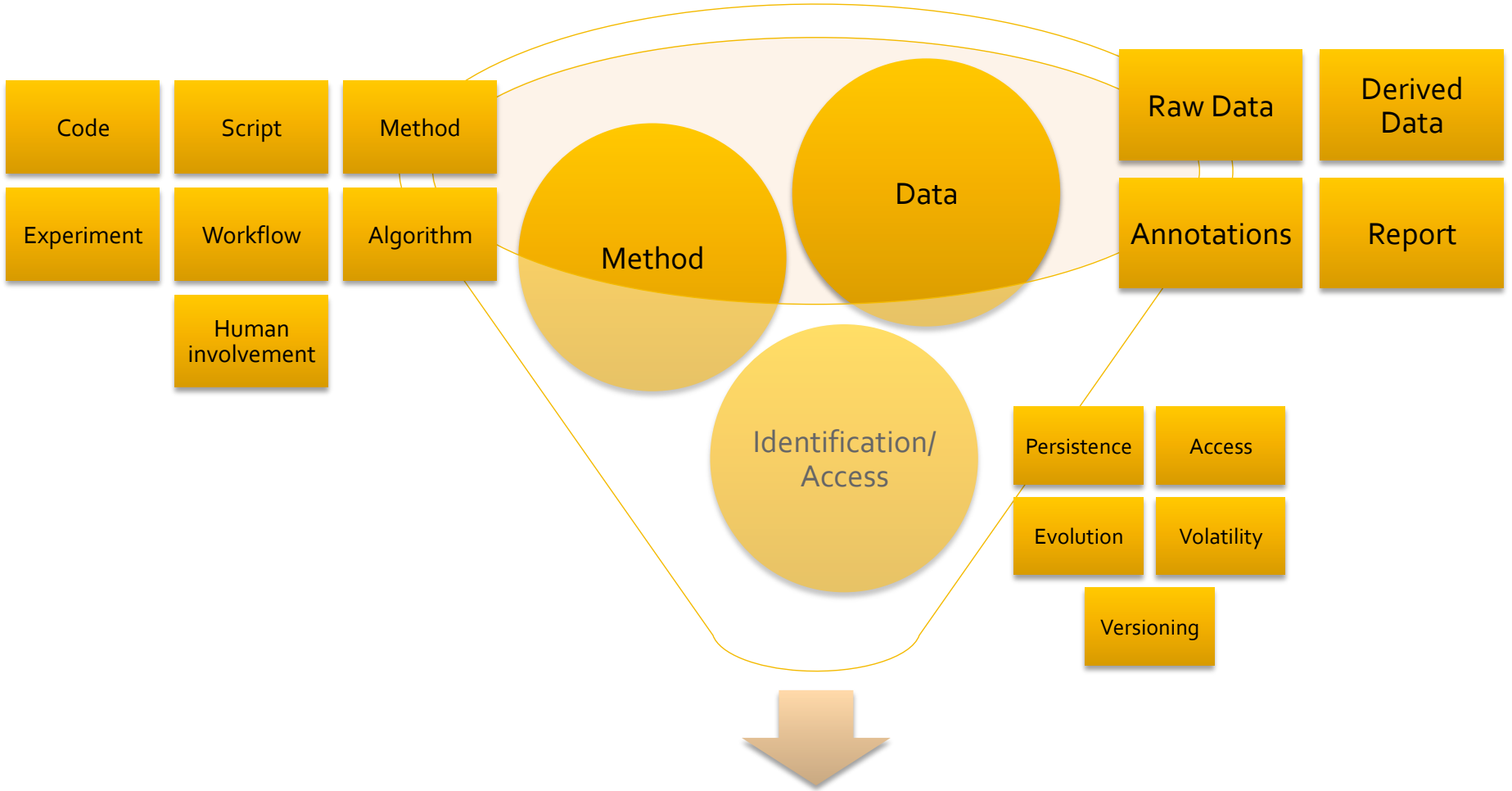
Reproducibility

Dimensions



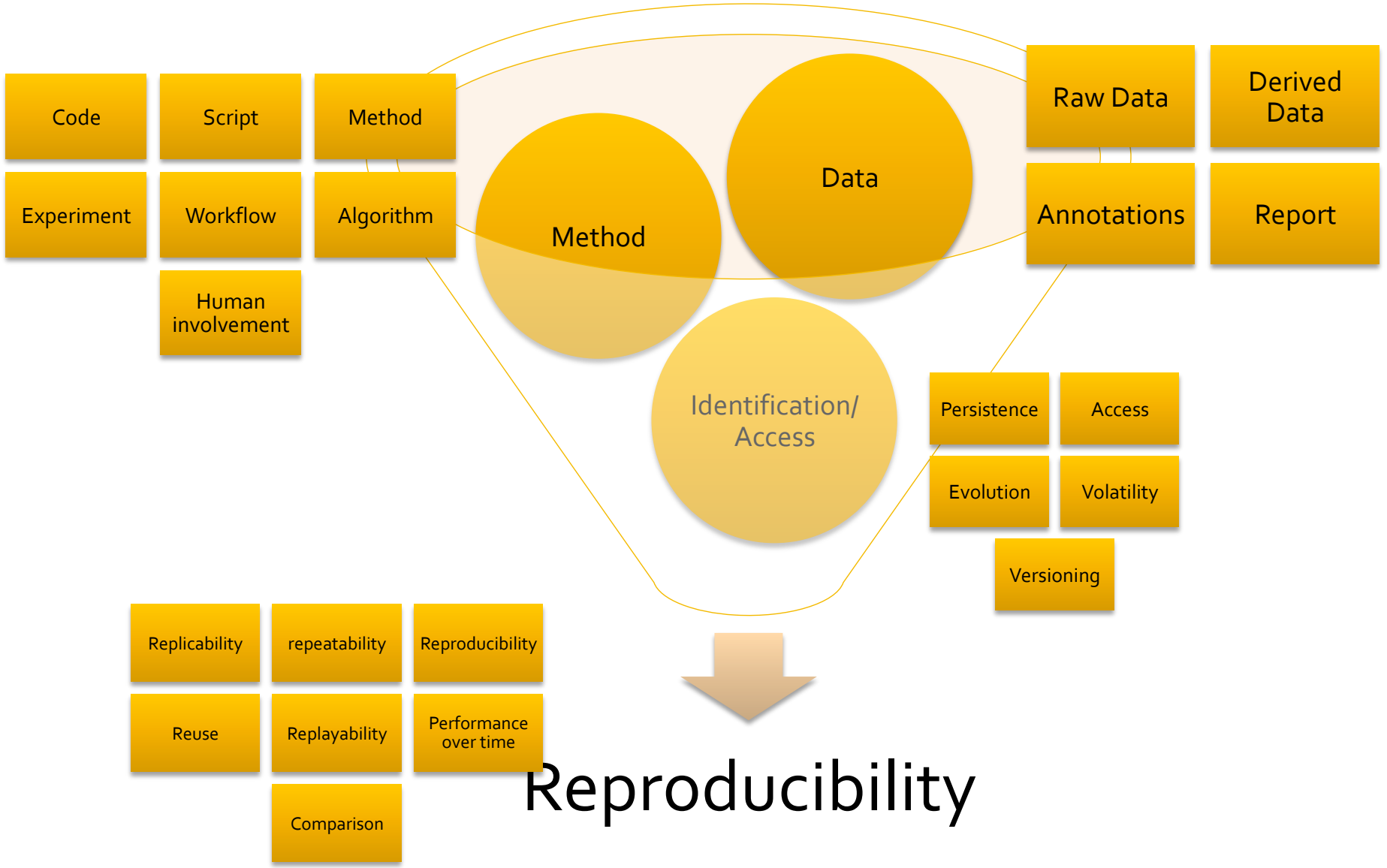
Reproducibility

Dimensions



Reproducibility

Dimensions


Code
Script
Method
Experiment
Workflow
Algorithm
Human involvement
Method
Data
Identification/ Access
Raw Data
Derived Data
Annotations
Report
Persistence
Access
Evolution
Volatility
Versioning
Replicability
repeatability
Reproducibility
Reuse
Replayability
Performance over time
Comparison
Reproducibility

Data and Code

- It is more likely for a scientific paper to contains pointers to the data than a code.

- Example of computational linguistics

Distribution of data and code availability in both 2011 and 2016.

	2011: data		2016: data		2011: code		2016: code	
Data / code available	**116**	**75.8%**	**196**	**86.3%**	**48**	**33.1%**	**131**	**59.3%**
- working link in paper	98	64.1%	179	78.9%	27	18.6%	80	36.2%
- link sent	11	7.2%	15	6.6%	17	11.7%	50	22.6%
- repaired link sent	7	4.6%	2	0.9%	4	2.8%	1	0.5%
Data / code unavailable	**37**	**24.2%**	**31**	**13.7%**	**97**	**66.9%**	**90**	**40.7%**
- sharing impossible	19	12.4%	14	6.2%	46	31.7%	42	19.0%
- no reply	17	11.1%	12	5.3%	43	29.7%	32	14.5%
- good intentions	0	0.0%	2	0.9%	5	3.4%	12	5.4%
- link down	1	0.7%	3	1.3%	3	2.0%	4	1.8%
Total	***153***	***100%***	***227***	***100%***	***145***	***100%***	***221***	***100%***
No data/code used	11		4		19		10	
Total nr. of papers	164		231		164		231	

Raw vs. Derived Data

- The conclusions reported on in a scholarly papers are made based on interpretation of the derived data.
- Often, it is the derived data (that is data used in the charts shown in the paper), that is made available.
- The raw data, and the processing performed in order to get rid of the outliers is not reported on.
  - This can be essential for debugging or discussing the results.

Method, Algorithm, Workflow, Script, Code

- They are used for different purposes
- They have different levels of abstractions
- In some scientific fields we need all of them, e.g., signal processing, AI applications
- In scientific papers, we often describe the method, and sketch the algorithm (for space sake ☺), the code is often overlooked …

Hi! I am also working on a project related to X. I have implemented your algorithm but unable to get the same results as described in your paper. Which values should I use for parameters Y and Z?"

Method, Algorithm, Workflow, Script, Code

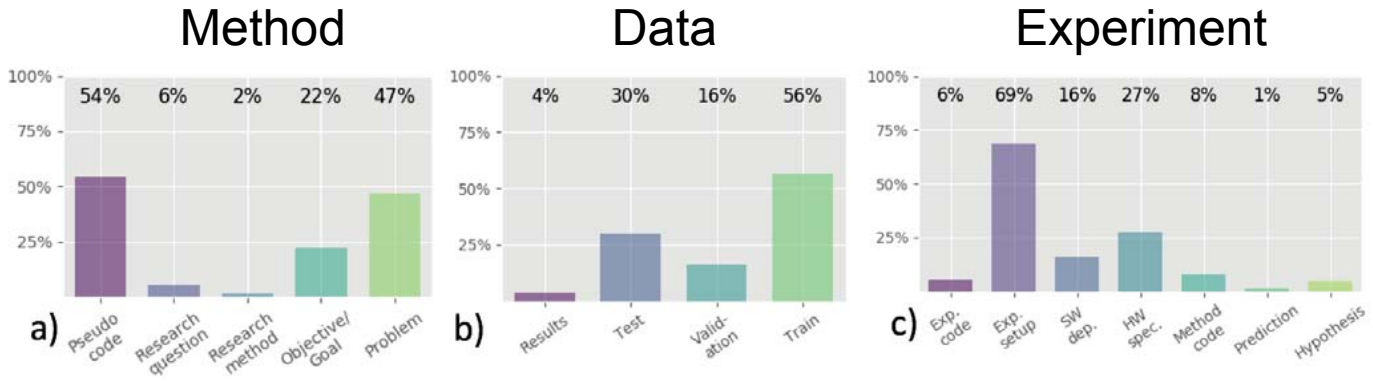
- They are used for different purposes
- They have different levels of abstractions
- In some scientific fields we need all of them, e.g., signal processing
- In scientific papers, we often describe the method, and sketch the algorithm (for space sake ☺), the code is often overlooked …

[TABLE 1] RESULTS OF REPRODUCIBILITY STUDY ON *IEEE TRANSACTIONS ON IMAGE PROCESSING* PAPERS PUBLISHED IN 2004. AVERAGE SCORES OVER THE 134 PAPERS ARE PRESENTED.

ALGORITHM						CODE		DATA		
DETAILS	PARAMETER VALUES	BLOCK DIAGRAM	PSEUDO-CODE	PROOFS	COMPARISON	IMPLEM. DETAILS	CODE AVAIL.	EXPLANATION OF DATA	SIZE DATA SET	DATA AVAIL.
0.84	0.71	0.37	0.33	0.27	0.64	0.12	0.09	0.83	0.47	0.33

Method, Algorithm, Workflow, Script, Code

- They are used for different purposes
- They have different levels of abstractions
- In some scientific fields we need all of them, e.g., signal processing



400 research papers from the conference series IJCAI and AAAI have been surveyed

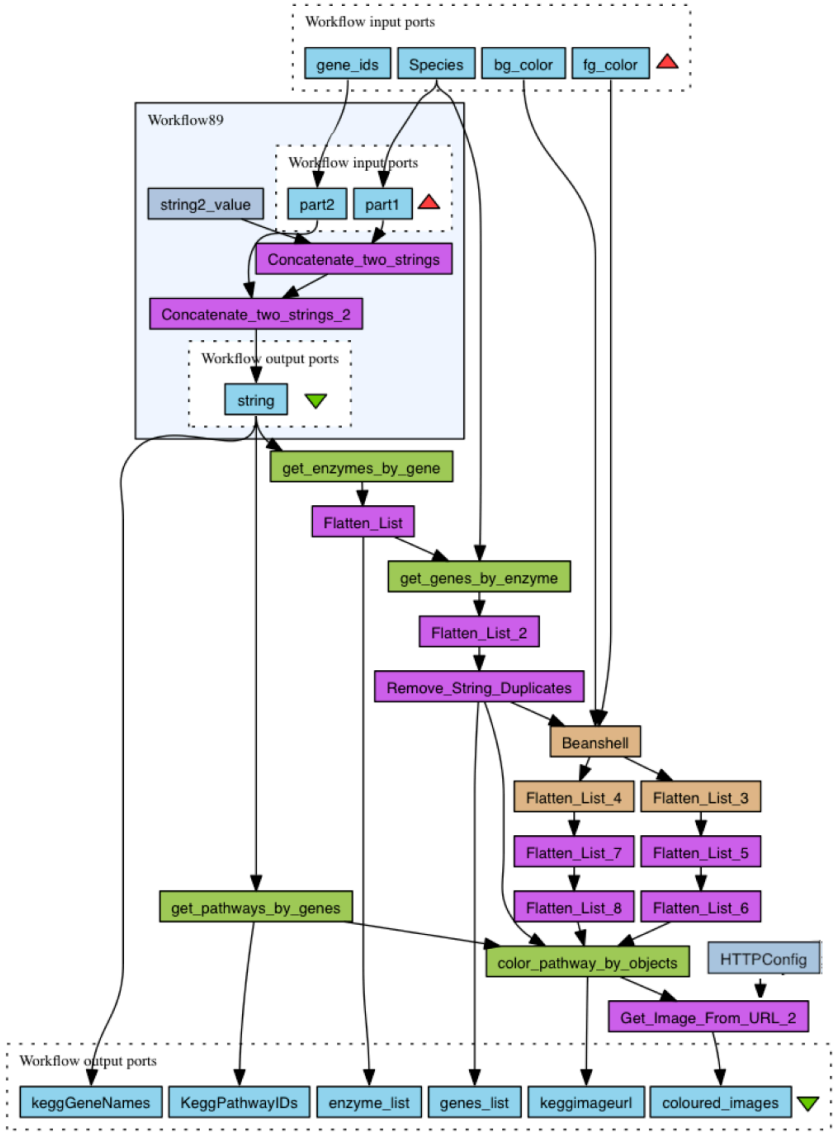
Odd Erik Gundersen, Sigbjørn Kjensmo: State of the Art: Reproducibility in Artificial Intelligence. AAAI 2018: 1644-1651

Identification, Persistence

- The URLs provided within papers works for few months
- The software too
  - Can anyone guarantee that github or bitbucket will exist 10 years from now?
- The API
  - For example, Facebook and Twitter provides fettered access to their content using API, with consequences on online social network studies. In addition a license agreement needs to be honored.
- Services too
  - Impact on Workflows.

The Proportion of Decay in Workflows

- 75% of the 92 tested workflows failed to be either executed or produce the same result (if testable)
- Those from earlier years (2007-2009) had 91% failure rate



Human Involved Computation

- **Cost**: Repeating an experiment that involve humans can be costly.
- **Sampling strategy**: When conducting user studies, it is important to know whether the authors were investigating a certain population, or whether they intend their findings to be generally applicable to a wider population, as this has implications for how participants are recruited for replications.
- **Consent**: The issue of obtaining informed consent when conducting online research is contentious
- **Participant briefing**: As with the acquisition of con- sent, the briefing and debriefing experience is an important ethical consideration when conducting human subjects research.

Privacy and Reproducibility

- Different techniques for ensuring data privacy with different protection levels
  - Pseudo-anonymization
  - Generalization/k-anonymity
  - Differential privacy

https://gdpr.eu/data-privacy/

- From reproducibility point of view, it is certainly better to have the data in its pure form without it being anonymized at all.
- That said, a certain of reproducibility is possible even with anonymized data, viz. inferential reproducibility
- Inferential reproducibility through replayability. : The drawing of qualitatively similar conclusions by replayability, which allows the investigator to “go back and see what happened”. It does not necessarily involve execution or enactment of processes and services. It places a requirement on provenance of data.

Need to strike right the balance between reproducibility and privacy

Evaluation of Performance Over Time

- Some of the reproducibility test papers that we reviewed, went beyond the definition of repeatability or replicability to assess the performance of systems overs time.
- For example, in IR, Armonstrong et al., 2009, performed experiments on 5 search engines to assess their effectiveness regarding the processing of Ad-Hoc queries between 1994 and 2005.
- Their starting hypothesis was that they would observe an upward trend in effectiveness.
- They found no evidence that the retrieval models were improved from 1994 to 2005.
- Their follow-up study further analyzed the retrieval results published at SIGIR and CIKM from 1998-2008, pointed out the baselines used in these publications were generally weak, and concluded that the ad hoc retrieval is not measurably improving.

Comparison of Performance

- Another application of reproducibility, that was investigated in IR is the comparison of performance of IR functions using benchmark datasets, as opposed to those used by the authors in the original paper (see Yang and Fang, 2016).

- This is an interesting case **for automatically evaluating** the performance of new solutions given the state of the art.

Assessing the Impact of New Hardware/ Software on Reproducibility

- Impact of new versions of the software on the reproducibility of the results of a method.
- In climate simulation, for example, the nature of computer architecture layouts result in solutions with round-off differences.
- Round-off differences are generally caused by the order of a sequence of computations, which may depend on the order of messages arriving from different parallel processes.
- To assess the impact of round-off differences, the authors investigated if changes in the hardware or software (versioning) result in tolerant round-offs in the expected results.

Conclusions Reached This Far

- Computational reproducibility has different requirements depending on the application domain

- Beyond establishing trust, reproducibility have the potential of facilitating advances in the state of the art through increased reuse, comparison, et re-evaluation of performances over time.

DAUPHINE UNIVERSITÉ PARIS | PSL

LAMSADE
UMR CNRS 7243
laboratoire d'analyse et modélisation de systèmes pour l'aide à la décision

The State of Reproducibility in Computational and Data-Driven Research?

Khalid Belhajjame
kbelhajj@gmail.com