



Tópicos Especiais SI

Fundamentos de Ciência de Dados

PROF SERGIO SERRA E JORGE ZAVALETA

{SERRA, ZAVALETA} @PET-SAI.UFRRJ.BR

2020.3

Datas Importantes

Atos acadêmicos no SIGA - Calendário Trimestral	1º Trimestre	2º Trimestre	3º Trimestre	4º Trimestre
Início de atividades	06/07/2020	13/10/2020	01/02/2021	----- X -----
Rematricula de matrícula trancada (destrancamento de matrícula)	Até 27/06/2020	Até 05/10/2020	Até 23/01/2021	----- X -----
Previsão de turmas	Até 19/06/2020	Até 26/09/2020	Até 15/01/2021	----- X -----
Trancamento de matrícula	Até 10/08/2020	Até 17/11/2020	Até 08/03/2021	----- X -----
Pedido de inscrição em disciplinas	De 06/07/2020 a 24/07/2020	De 11/10/2020 a 17/10/2020	De 30/01/2021 a 05/02/2021	----- X -----
Concordância do pedido de inscrição em disciplina	De 27/07/2020 a 30/07/2020	De 18/10/2020 a 24/10/2020	De 06/02/2021 a 12/02/2021	----- X -----
Efetivação do Pedido de Inscrição (Divisão de Ensino – PR2)	31/07/2020	27/10/2020	15/02/2021	----- X -----
Pedido de alteração de inscrição em disciplina – AID	De 02/08/2020 a 08/08/2020	De 28/10/2020 a 31/10/2020	De 16/02/2021 a 19/02/2021	----- X -----
Concordância do pedido de alteração de inscrição em disciplina - AID	De 09/08/2020 a 12/08/2020	De 01/11/2020 a 07/11/2020	De 20/02/2021 a 26/02/2021	----- X -----
Efetivação De Alteração do Pedido de Inscrição (Divisão de Ensino – PR2)	13/08/2020	10/11/2020	01/03/2021	----- X -----
Pedido de trancamento de inscrição em disciplina (desistência de inscrição)	De 14/08/2020 a 19/08/2020	De 11/11/2020 a 14/11/2020	De 02/03/2021 a 05/03/2021	----- X -----
Concordância do pedido de trancamento de inscrição em disciplina	De 20/08/2020 a 31/08/2020	De 15/11/2020 a 28/11/2020	De 06/03/2021 a 19/03/2021	----- X -----
Efetivação do Trancamento do Pedido de Inscrição (Divisão de Ensino – PR2)	24/08/2020	01/12/2020	22/03/2021	----- X -----
Término de atividades	03/10/2020	16/01/2021	24/04/2021	----- X -----
Notas – Pautas de graus e frequência	De 04/10/2020 a 17/10/2020	De 17/01/2021 a 30/01/2021	De 25/04/2021 a 08/05/2021	----- X -----

Programa

Quintas das ~ 13:30 até ~17:00
Teórico–práticas
Google Meet

Módulo 1:

1. Reprodutibilidade em Pesquisa Computacional
2. Introdução a Proveniência de Dados
3. Gestão de Grandes Volumes de Dados de Pesquisa
4. Ambiente de Programação: python 3, jupyter notebook, JupyterLab, Google Colab, DeepNot pacotes e github
5. Python I: tipos de dados, sequências e operações, estruturas de controle e repetição
6. Prática dos conteúdos estudados: construindo e operando listas e strings

Módulo 2:

1. Técnicas de coleta e preparação de dados
2. Numpy I: array, slicing, fancy index, copy and view
3. Pandas I: dataframes, series, index, Pandas I/O (csv, json, excel)
4. Prática dos conteúdos estudados: Processando e extraindo informações de arquivos csv, Jason, rdf

Módulo 3:

1. Técnicas de análise de dados
2. Numpy II e Matplotlib: operações com array, broadcasting, construção de gráficos usuais
3. Pandas II: estatísticas básicas
4. Prática dos conteúdos estudados: manipulando dados de saúde, ambiente, agricultura, cidades inteligentes

Módulo 4:

1. Introdução a técnicas de modelagem de fluxo de dados
2. Algoritmos e técnicas de extração inteligente de conhecimento
3. Scikit learn: introdução a mecanismos de regressão, classificação, clustering e PCA
4. Prática dos conteúdos estudados: clusterização e predição

Módulo 5:

1. *Seminários sobre Ciência de Dados aplicados domínio específicos (e.g. Saúde, Educação, Sustentabilidade, Agricultura, Cidades Inteligentes, covid-19, entre outros)*
2. *Apresentação de trabalhos – proposta de artigos*

Avaliação e Atendimento

Critérios de aprovação são os do PPGI/UFRJ.

A avaliação da disciplina consiste em participação em sala de aula (P); exercícios e/ou protótipos desenvolvidos (E); apresentações/ /escritas de artigos (A).

$$MF = 0.2 * P + 0.2 * E + 0.6 * A$$

O aluno que desejar atendimento deverá requisitar o mesmo por e-mail e um horário será agendado pelos responsáveis para o atendimento.



serra@pet-si.ufrj.br



zavaleta@pet-si.ufrj.br

Bibliografia

Materiais apresentados em sala de aula (português + inglês conforme o caso)

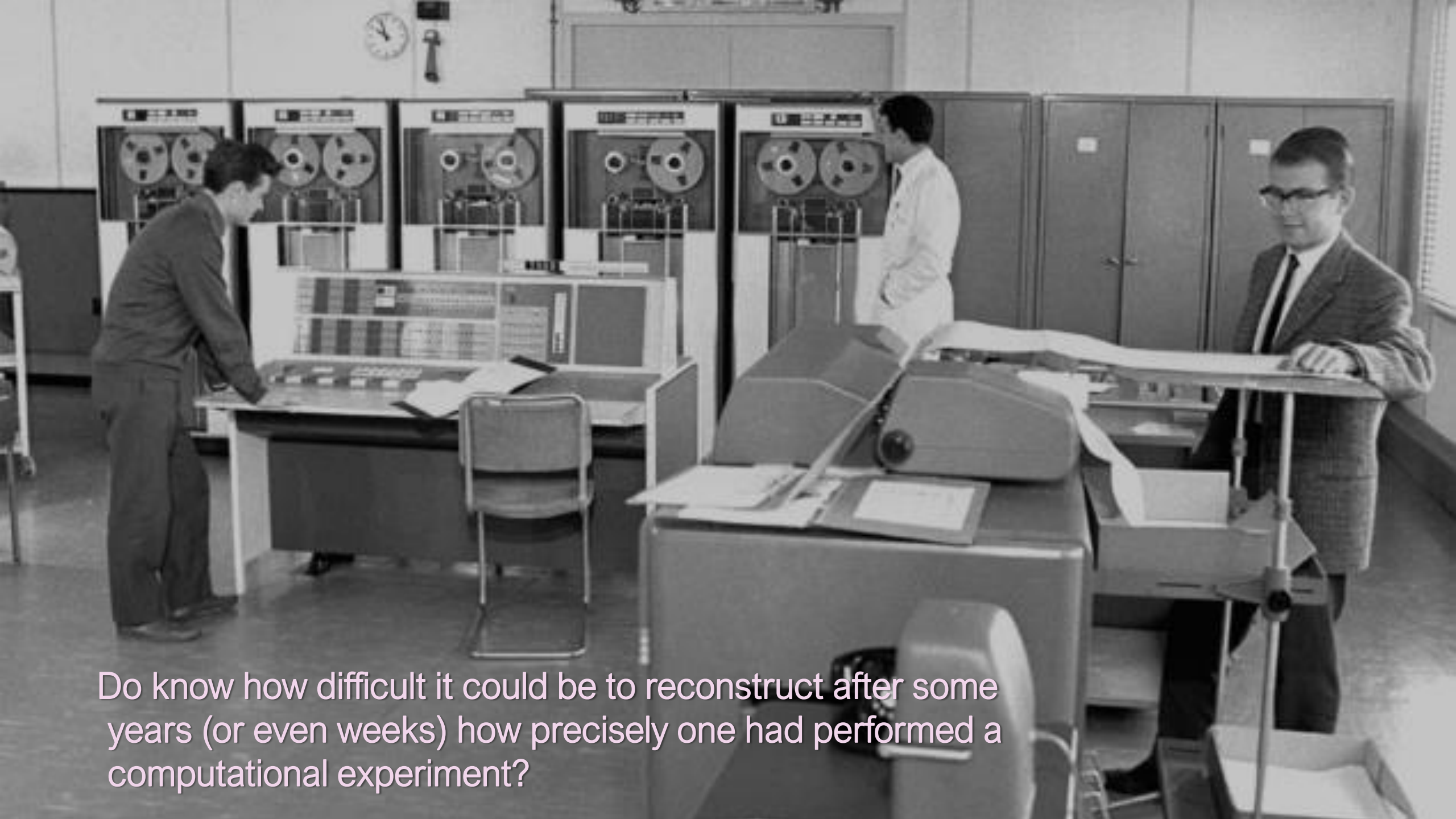
Básica

- 1- National Academies of Sciences, Engineering, and Medicine. Reproducibility and Replicability in Science. Washington, DC: The National Academies Press, 1st Edition, 2019.
- 2- Victoria Stodden, Friedrich Leisch, Roger D. Peng, Implementing Reproducible Research, CRC Press, 1st Edition, 2014.
- 3- Kleppmann, M., Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems, O'Reilly, 2017.
- 4- Taylor, E. Deelman, D.B. Gannon, M. Shields (Eds.), Workflows for e-Science: Scientific Workflows for Grids, Springer, 2006.
- 5- Wes McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd edition O'Reilly Media, 2017
- 6- Mark Lutz, Learning Python, 5th Edition, O'Reilly Media, 2013
- 7- Jonh Hearty, Advanced Machine Learning with Python. Packt Publishing, 2016.
- 8- Andreas C. Mueller and Sarah Guido, Machine Learning with Python. O'Reilly Media, 2016.
- 9- John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT, 2015.
- 10- Artigos ou apresentações selecionados

Introduction to Data Science

MODULE I

REPRODUCIBILITY X REPLICABILITY



Do know how difficult it could be to reconstruct after some years (or even weeks) how precisely one had performed a computational experiment?

In the beginning...

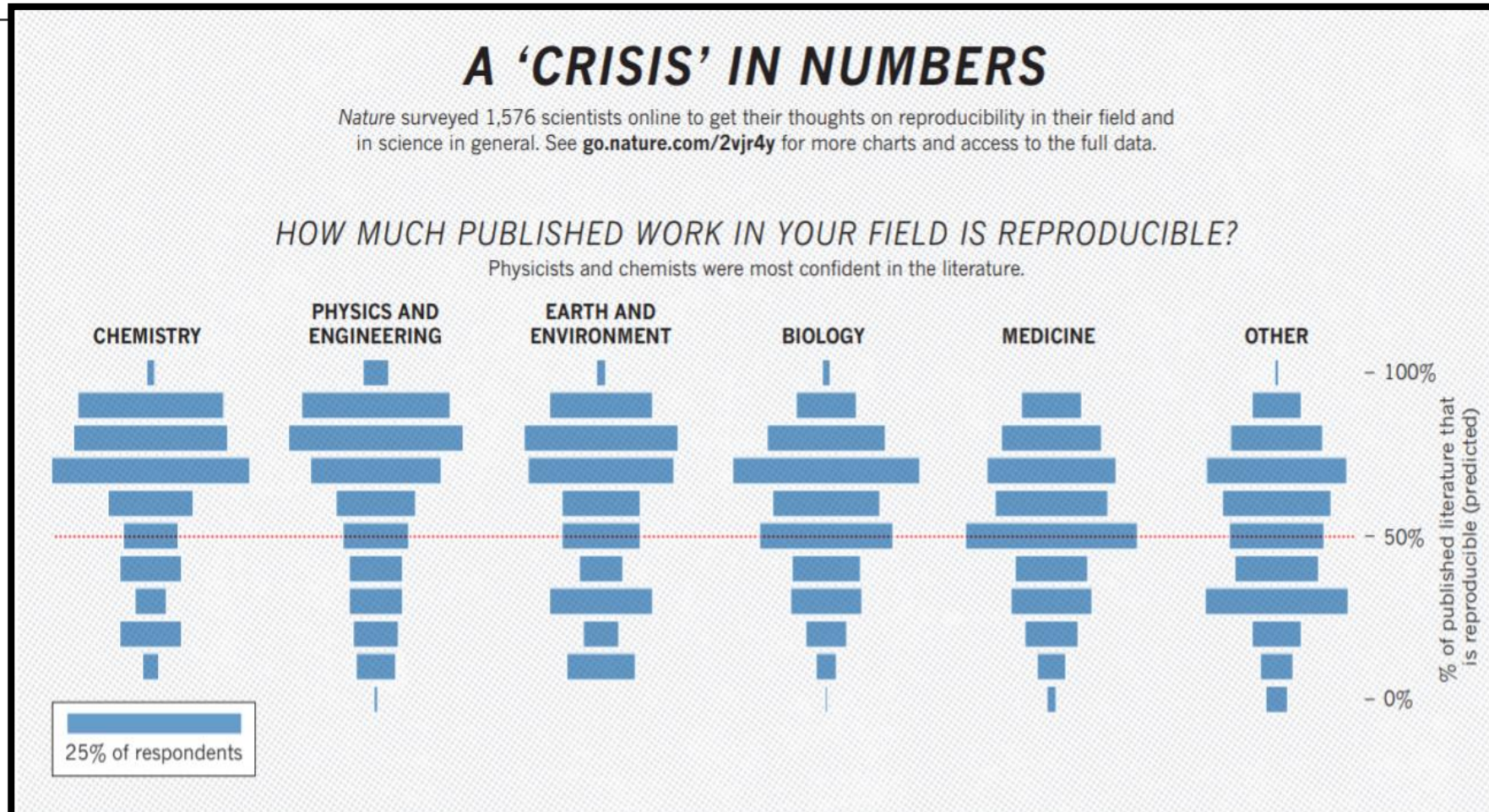
When scientists began to use computers to perform **simulation experiments and data analysis**, attention to **experimental error** took backstage.

- Computers are exact machines, **practitioners apparently assumed that results obtained by computer could be trusted**, provided that the principal algorithms and methods employed were suitable to the problem at hand.

Little attention was paid :

- 1) correctness of implementation,
- 2) potential for error, or
- 3) variation introduced by system soft and hardware.

Have you failed?

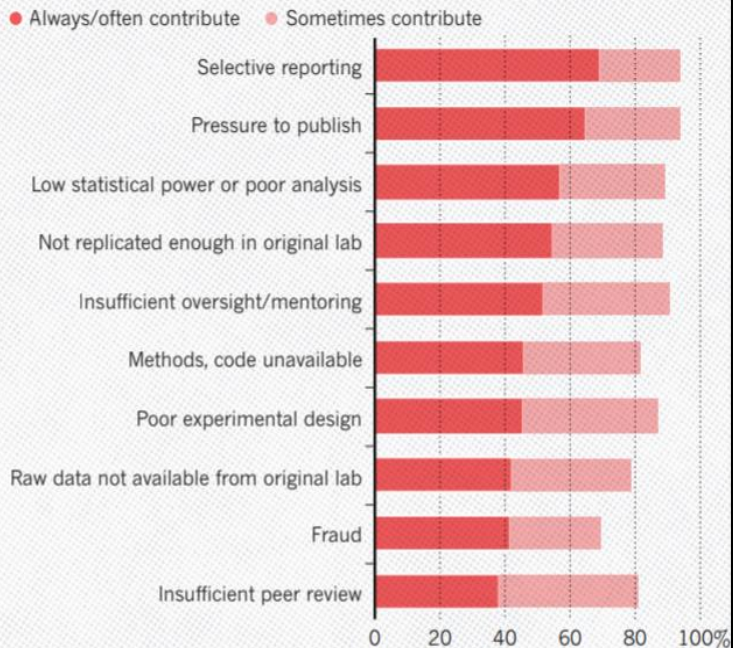


* Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 452-454.

Why you failed?

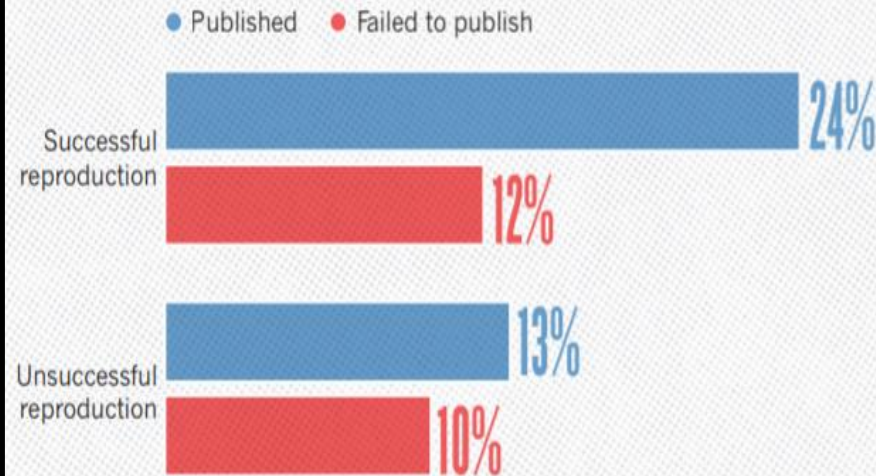
WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.



HAVE YOU EVER TRIED TO PUBLISH A REPRODUCTION ATTEMPT?

Although only a small proportion of respondents tried to publish replication attempts, many had their papers accepted.

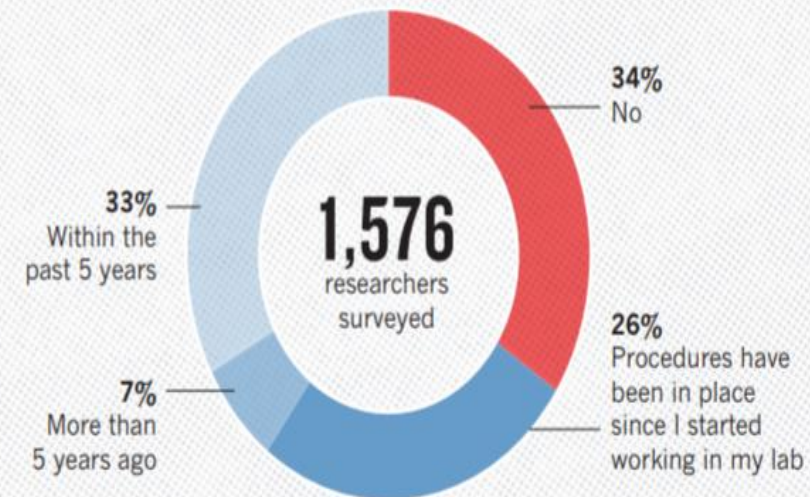


Number of respondents from each discipline:

Biology **703**, Chemistry **106**, Earth and environmental **95**, Medicine **203**, Physics and engineering **236**, Other **233**

HAVE YOU ESTABLISHED PROCEDURES FOR REPRODUCIBILITY?

Among the most popular strategies was having different lab members redo experiments.



The value of Reproducibility...

Illuminating the black box

Note to biologists: submissions to *Nature* should contain complete descriptions of materials and reagents used.

nature

Reporting Checklist For Life Sciences Articles

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. For more information, please read [Reporting Life Sciences Research](#).

A Biostatistic Paper Alleges Potential Harm To Patients In Two Duke Clinical Studies

By Paul Goldberg

Biostatistics journals aren't usually the place to go for sensational allegations. The most recent issue of the *Annals of Applied Statistics* is an

Human lives

published on this journal's website alleges that cancer patients died by being placed on two Duke University clinical trials that rely on biomarkers to select therapies.

The paper is the culmination of efforts by Keith Baggerly and Kevin Coombes, biostatisticians at M.D. Anderson Cancer Center, to verify the work by a group of Duke researchers led by Anil Potti and Joseph Nevins.

In multiple publications, the Duke team has claimed that microarray (Continued to page 2)

COMPUTER SCIENCE

Accessible Reproducible Research

Jill P. Mesirov

Scientific publications have at least two goals: (i) to announce a result and (ii) to convince readers that the result is correct. Mathematics papers are expected to contain a proof complete enough to allow edgeable readers to fill in any details. For experimental sciences, should...

As use of computation in research grows, new tools are needed to expand record reporting, and reproduction of methods and data.

Science POLICYFORUM

The New York Times

NYTimes: [Home](#) - [Site Index](#) - [Archive](#) - [Help](#)

Scientific integrity

Friday, December 2, 2011 As of 12:00 AM New York 43°/34°

THE WALL STREET JOURNAL. HEALTH

HEALTH INDUSTRY | DECEMBER 2, 2011

Scientists' Elusive Goal: Reproducing Study Results

In September, Bayer published a study describing how it had halted nearly two-thirds of its early drug target projects because in-house experiments failed to match claims made in the literature.

Trust

Financial

Reliability

Nobel Laureate Retracts Two Papers Unrelated to Her Prize

by KENNETH CHANG
Published: September 23, 2010

Linda B. Buck, who shared a 2004 Nobel Prize in Physiology or Medicine, apologized for

Reproducibility x replicability



[Front Neuroinform.](#) 2017; 11: 76.

PMCID: PMC5778115

Published online 2018 Jan 18. doi: [10.3389/fninf.2017.00076](https://doi.org/10.3389/fninf.2017.00076)

PMID: [29403370](https://pubmed.ncbi.nlm.nih.gov/29403370/)

Reproducibility vs. Replicability: A Brief History of a Confused Terminology

[Hans E. Plesser](#)^{1,2,*}

► [Author information](#) ► [Article notes](#) ► [Copyright and License information](#) [Disclaimer](#)

- **Reproducibility** – can you recreate the same result using original data and code?
- **Replicability** – can you recreate the same result using new data but same experimental design?

Why reproducibility? Is it useful?

1. You can come back to your own analysis after a break (think peer review) or on a new machine
2. You can verify and extend other people's analyses

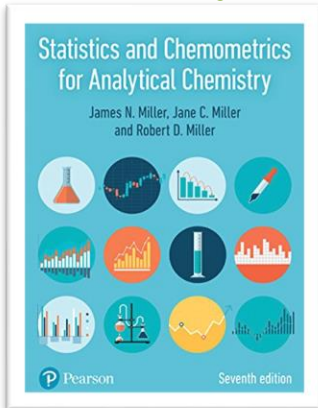
The Pioneers...

- ### Claerbout and Karrenbach, (1992)

- “Reproducing” means “running the same software on the same input data and obtaining the **same results**”
- “Replicating” means “writing and then running new software based on the description of a computational model or method provided in the original publication, and obtaining **results that are similar enough**”

- ### Donoho et al., (2009)

- ### Peng (2011)



Back to the Future (1985)

ACM definitions...

Repeatability (Same team, same experimental setup):

- The measurement can be obtained with stated precision by the **same team using the same measurement** procedure, the same measuring system, under the same operating conditions, in the same location on **multiple trials**.
- For computational experiments, this means that a researcher can reliably repeat her own computation.

Problems associated with repeatability

- Unaffordable cost of experimental procedure
- Genetic differences in experimental model
- Variation in experimental condition
- More variable involves, more error
- Restriction of using same instrument
- Biological process provides additional source of variability

ACM definitions...

Replicability (Different team, same experimental setup):

- The measurement can be obtained with stated precision by **a different team using the same measurement** procedure, the same measuring system, under the same operating conditions, in the same or a different location on **multiple trials**.
- For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.

ACM definitions...

Reproducibility (Different team, different experimental setup):

The measurement can be obtained with stated precision by a **different team**, a **different measuring system**, in a **different location** on **multiple trials**.

- For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

Criteria for Reproducibility

- Method of measurement
- Principle of measurement
- Observer
- Measuring instrument
- Reference standard
- Location
- Conditions of use
- Time

ACM definitions...

Reproducibility (Different team, different experimental setup):

The measurement can be obtained with stated precision by a **different team**, a **different measuring system**, in a **different location** on **multiple trials**.

- For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

Goodman	Claerbout	ACM
		Repeatability
Methods reproducibility	Reproducibility	Replicability
Results reproducibility	Replicability	Reproducibility
Inferential reproducibility		

Depending on who you ask, these definitions are reversed!

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778115/>

Goodman saves!

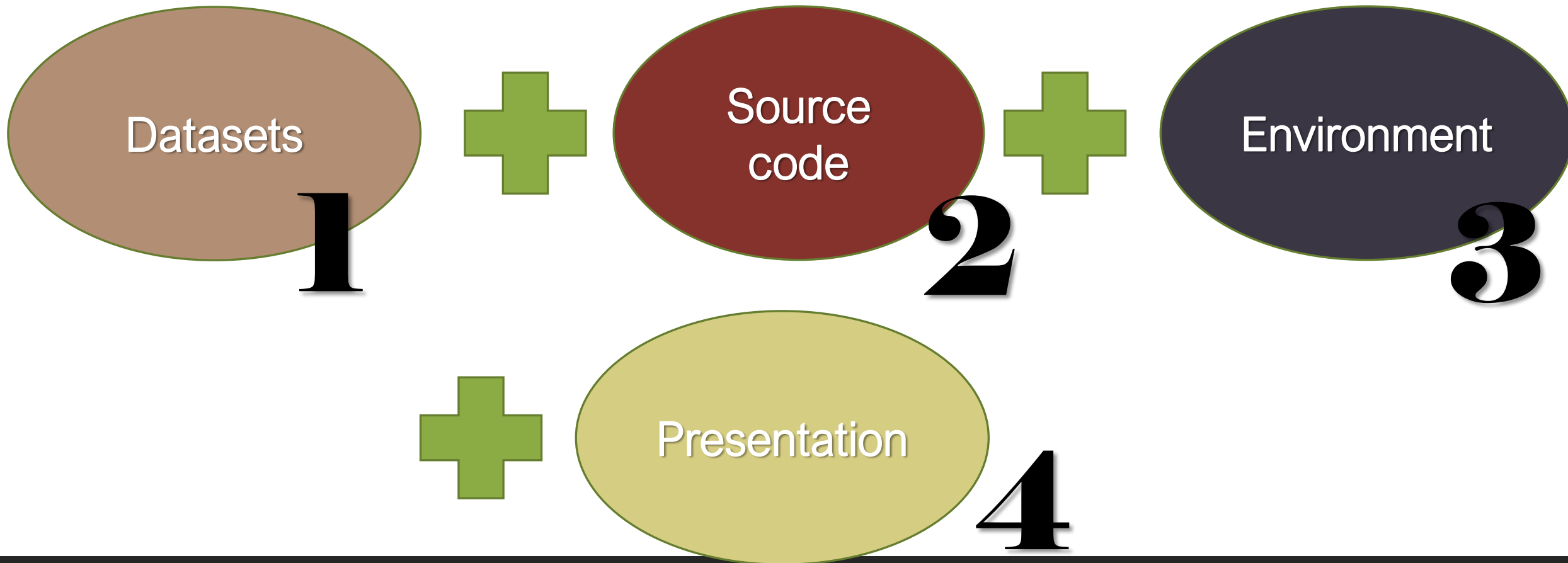
Terminology confusion! Goodman et al. (2016) propose a new lexicon for **research reproducibility**.

Definitions:

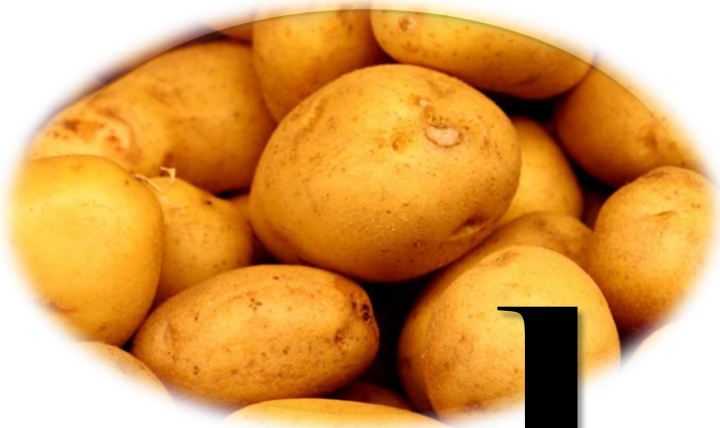
- **Methods reproducibility:** provide sufficient detail about procedures and data so that the same procedures could be exactly repeated.
- **Results reproducibility:** obtain the same results from an independent study with procedures as closely matched to the original study as possible.
- **Inferential reproducibility:** draw the same conclusions from either an independent replication of a study or a reanalysis of the original study.

These definitions **make explicit which aspects of trustworthiness** of a study, **avoid the ambiguity** caused by the fact that “reproducible”, “replicable,” and “repeatable” have very **similar meaning** in everyday language.

Reproducibility



Reproducibility (kitchen analogy!)



1



2

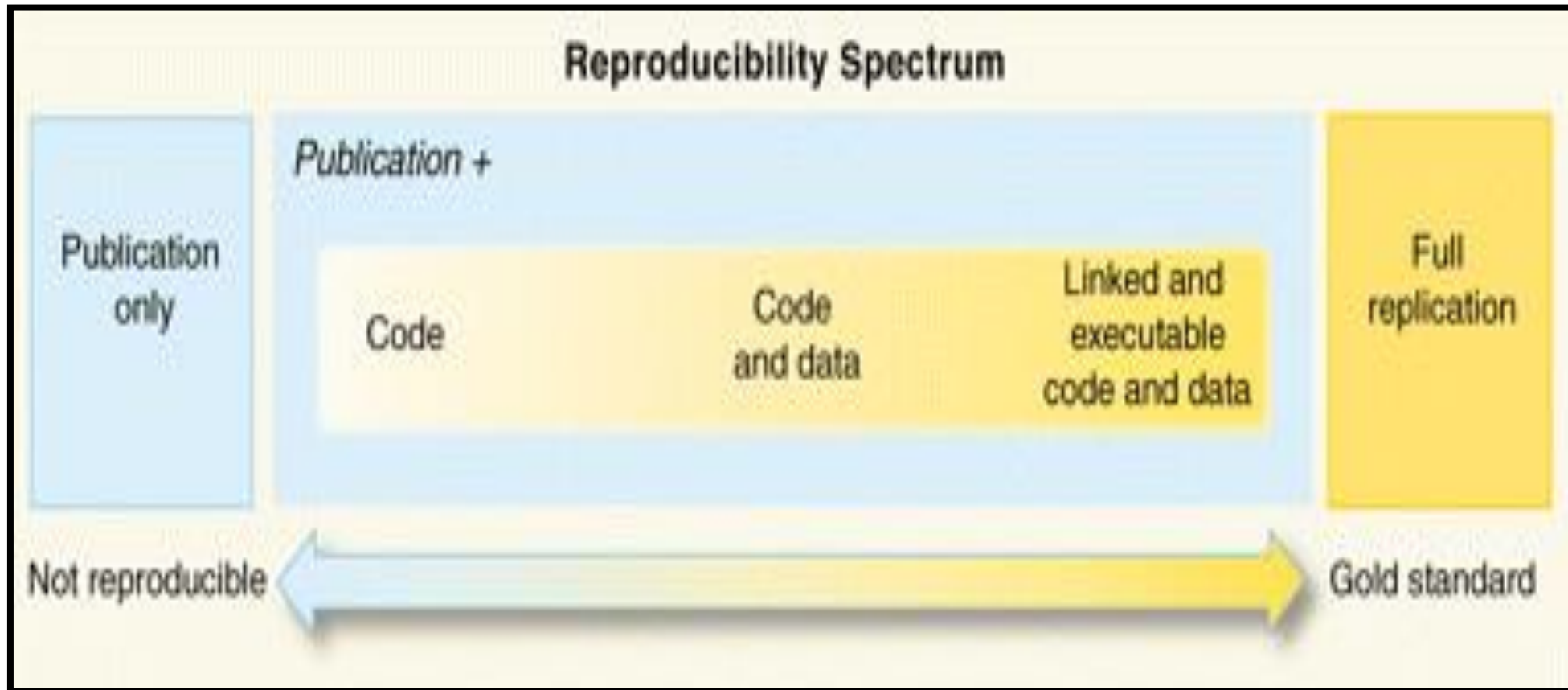


3



4

Reproducibility Spectrum



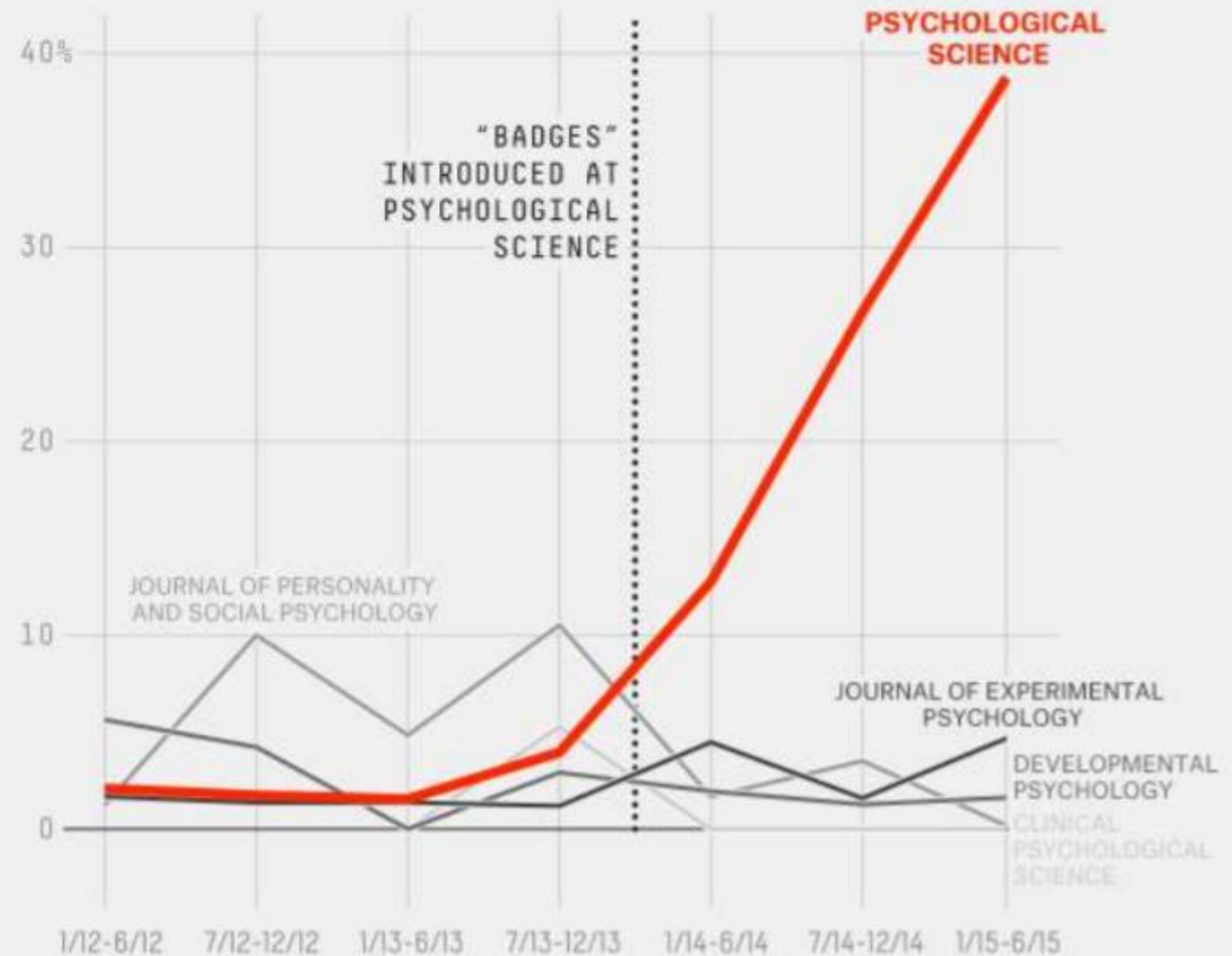
1- Dataset

Access to data is necessary,
but not sufficient for reproducibility!

As discussed in
Scientific Data
Management
course (2020.2)

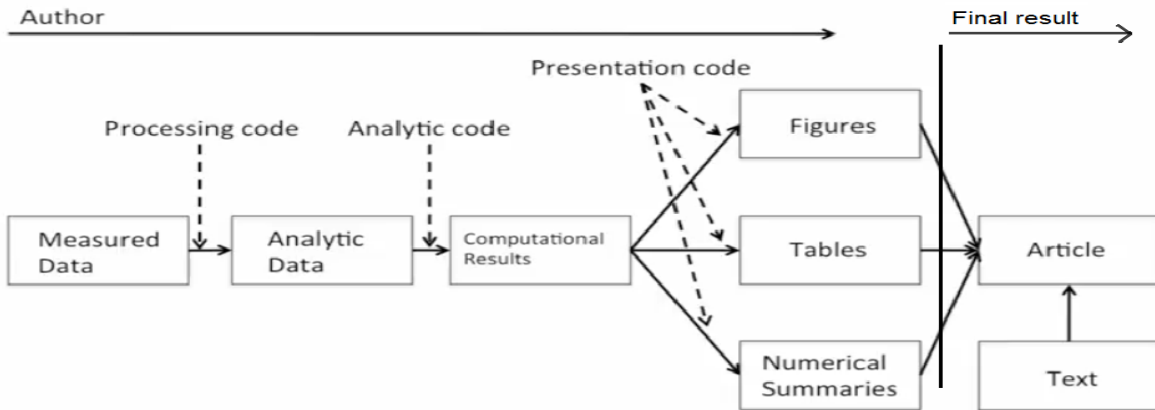
Data sharing rose when it was rewarded

Share of papers in four psych journals reporting open data

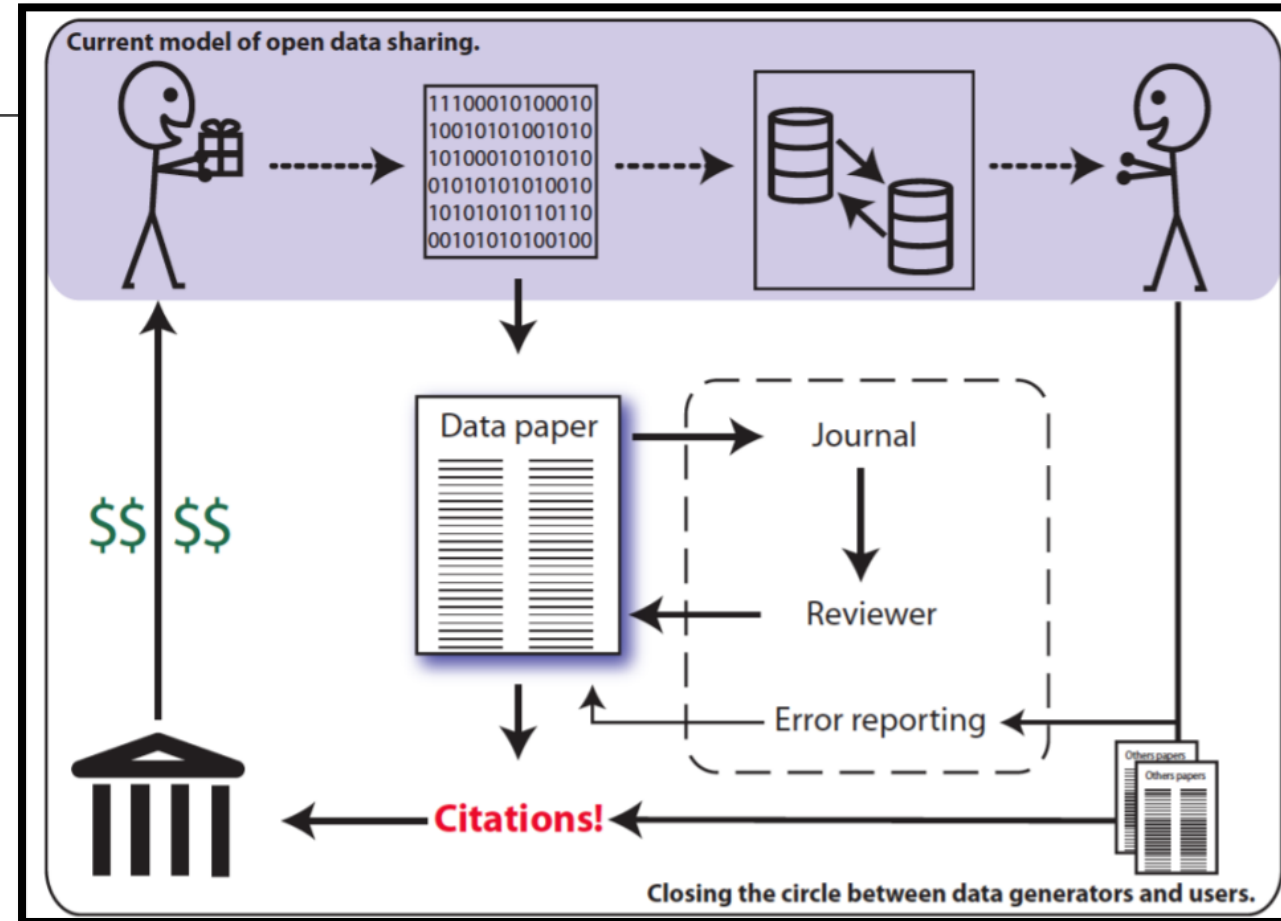


1- Dataset

Research Pipeline



The Wolf of Wall Street (2013)



Heidorn, P. B. (2008) Shedding Light on the Dark Data in the Long Tail of Science. Trends 57(2):280-299 DOI: 10.1353/lib.0.0036

1- Data

“Three types” of data (e.g. BioInfo)

1- Source data

short read datasets, microarrays, etc

2- Support data

Reference genomes, gene annotations, etc

3- Transformed data

Alignments, gene counts, etc

1- Data : Source Data

- Raw, unprocessed data is always preferred
 - E.g. untrimmed reads directly from sequencer
- Data should be deposited in a publicly available repository
 - E.g. GEO, dbGaP, figshare, zenodo
- Repository should have a plan for longevity → FAIR Data Principles
 - No personal servers!

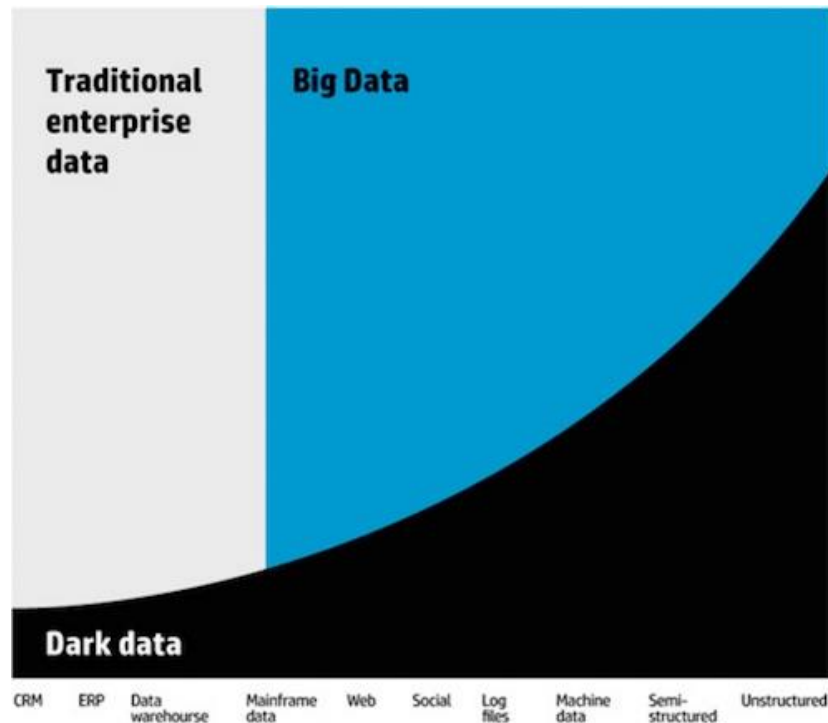
1– Data : Support Data (a.k.a Metadata)

- Data used to condense, process, annotate, and interpret source data
- Most support data are maintained in persistent repositories with consistent formats
- If there is a persistent link, specify it!
- If not, download and store data yourself

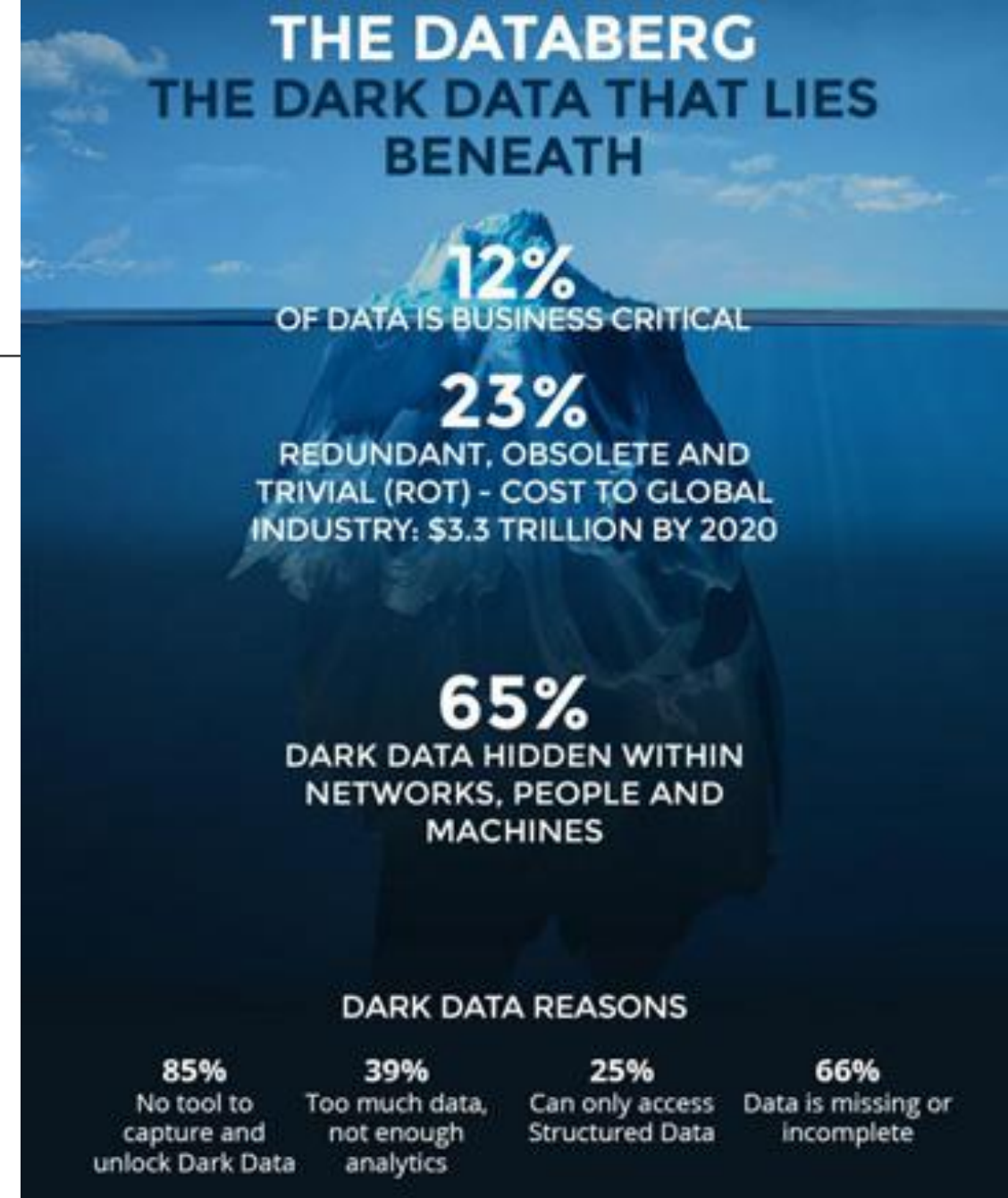
1- Data : Transformed Data

- Derived from source(+support) data
- Usually what we use to interpret our results
- Code is a recipe for creating transformed data from source data
 - Should not be maintained as part of your **workflow**
- EXCEPT the final transformed form of the data used to make interpretations

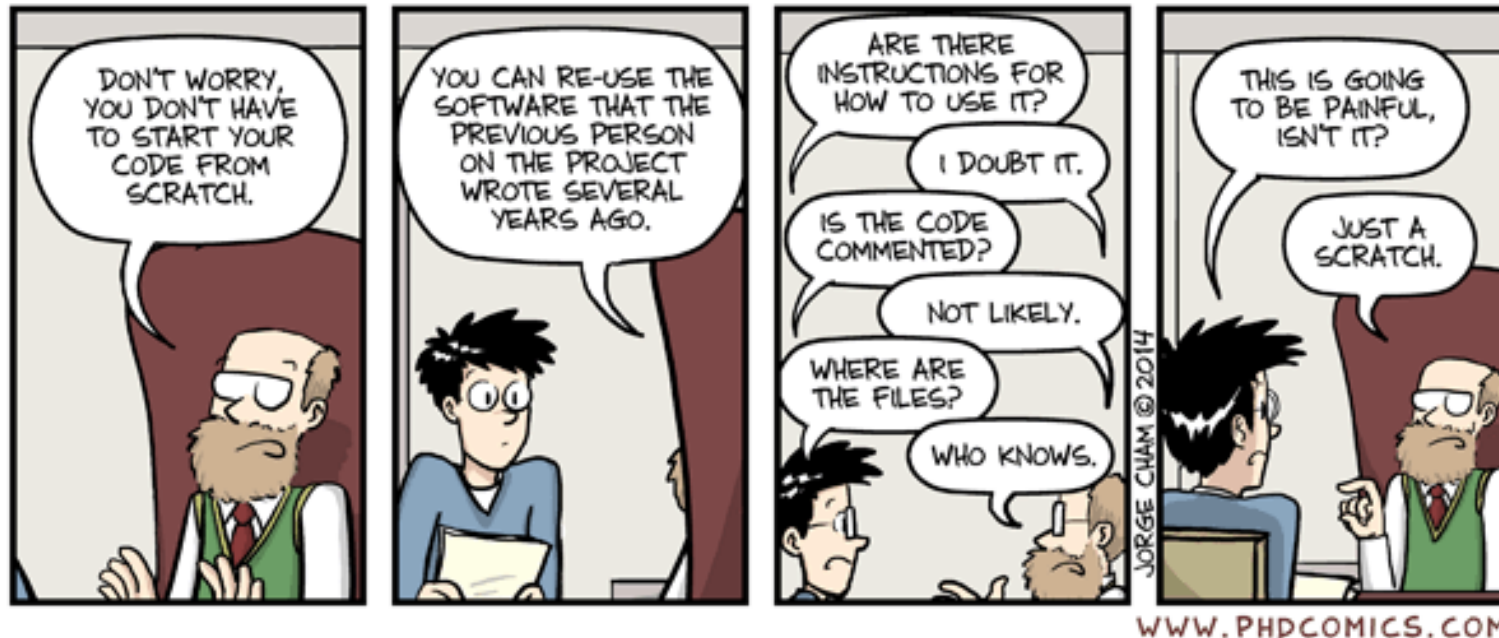
1- Dataset: Darkness



<https://medium.com/untrite/what-is-the-dark-data-and-why-should-organisations-start-looking-into-it-61cdba7aab8f>



2 – Code: Avoid darkness...



Must describe software and versions and their dependencies and their versions to fully recreate an environment!

“Dark Software” is the counterpart of “Dark Data” (Heidorn 2008)

2 – Code: Embrace automation

Automate your data analysis!

- If you do something **twice** write code for it. If you need to run, share it many times **create a workflow**
- Use **containerized and versioned preprocessing tools**:

- **Generic Tools**

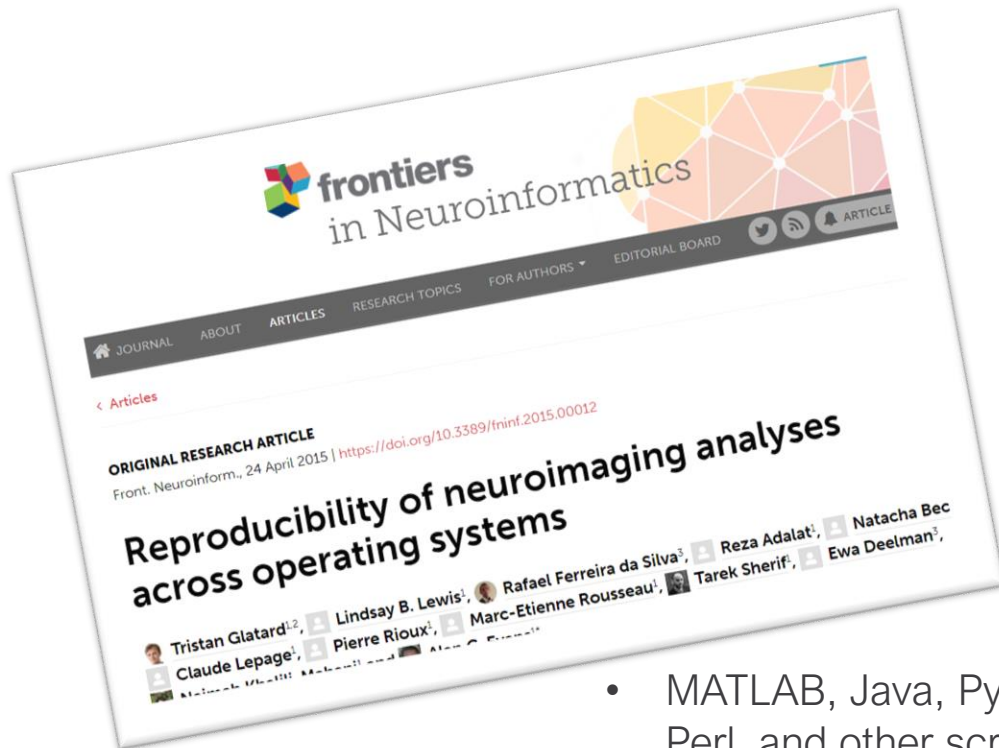
- Jupyter notebook,
- Jupyter Lab,
- Google Colab,
- DeepNote

- **Domains Tools**

- Examples...aa - <http://automaticanalysis.org>, C-PAC - <https://fcp-indi.github.io>, FMRIPREP – <http://fmriprep.or>



2- Code: Embrace automation



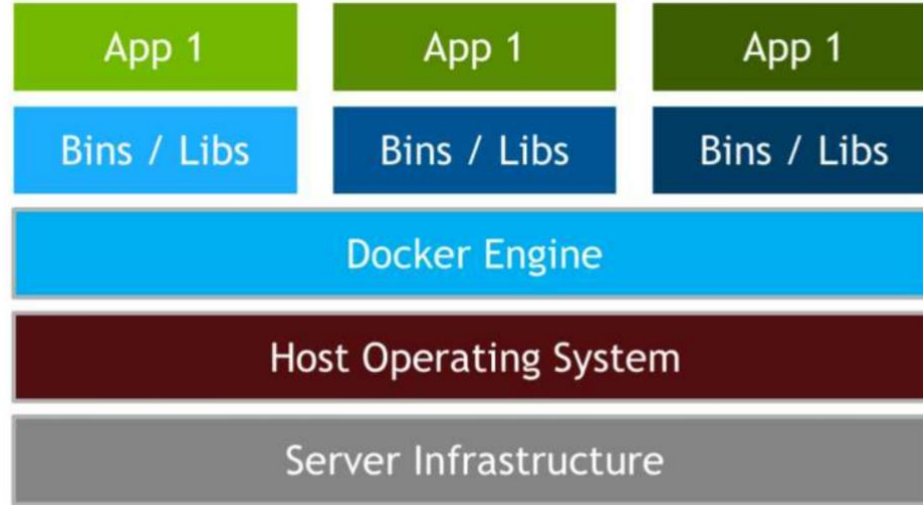
- MATLAB, Java, Python, Perl, and other scripting languages,
- CentOS 4 vs CentOS 6

This paper reports experiments with three of the neuroimaging tools (FMRIB Software Library Freesurfer and CIVET).

- Quantify the **reproducibility of tissue classification** (cortical and subcortical), resting-state fMRI analysis, and cortical thickness extraction, using different builds of the tools, **deployed on different versions of GNU/Linux**. We also identify some causes of these differences, using **library-call and system-call interception**.
- The paper closes with a discussion suggesting directions to address the **identified reproducibility issues**.

2- Code: Capturing dependencies

VM



CONTAINERS



2- Code:Version control

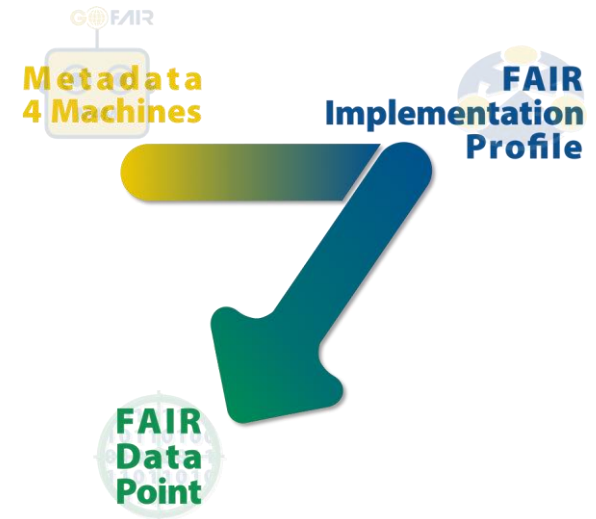
Git and GitHub and Bitbucket are useful for everyday work

- They also provide a way for you to share the code
- Use tags/releases = Metadata Standards
- Public dissemination of analysis code upon publication

Zenodo.org and FAIR Data Points for archival

LaTEX et al. for editing

Slack et al. as a workplace communication tool



GitHub

zenodo

2- Code: Documentation

1. Code should be well documented in comments and README files, also:
2. Code should be well documented in comments and README files, then:
3. Code should be well documented in comments and README files, also then:

Future you will thank current you for it

2- Code: Version control (e.g. Machine Learning/Data Science)

Improving and automate work and optimize processes on workflow with ML/DS Projects

- **Neptune** is a lightweight experiment management and collaboration tool. It is flexible, works with many other frameworks, and has a stable user interface you can effectively systematize your ML experiments and improve management.
- **Pachyderm** is a complete version-controlled data science platform that helps to control an end-to-end machine learning life cycle.
- **Delta Lake** is an open-source storage layer that brings reliability to data lakes. Delta Lake provides ACID transactions, scalable metadata handling, and unifies streaming and batch data processing.
- **R studio** development environment for R programming language.
- **Python is life!**

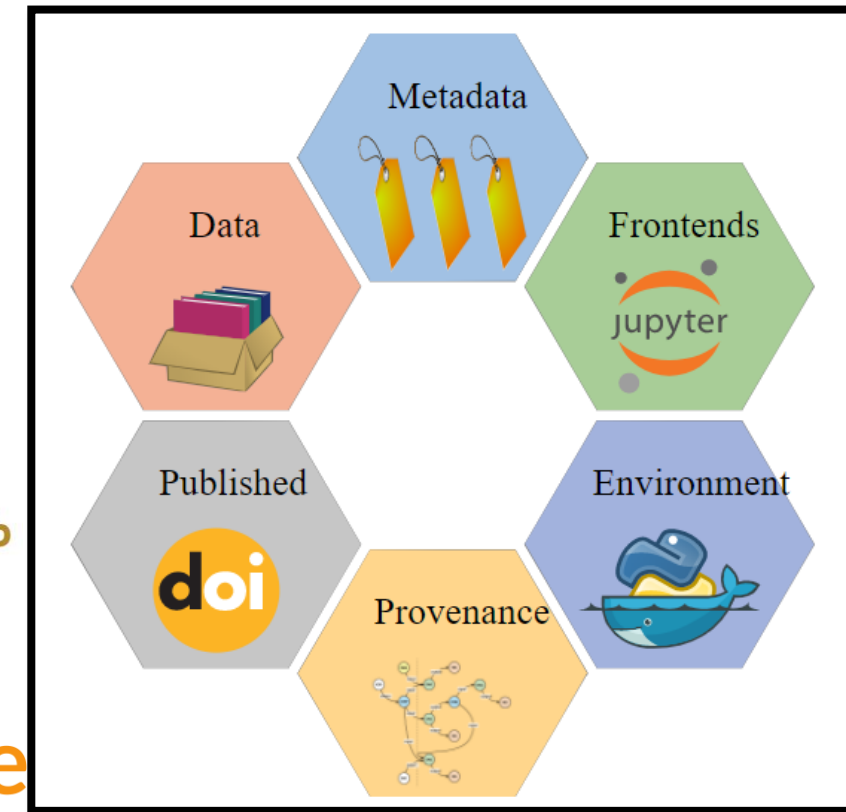
2- Code: Package control (e.g. Machine Learning/Data Science)

- anaconda: python distribution and environment management suite
- miniconda: just environment management suite
 - create **environment** that **installs and describes** a set of packages with versions
 - software organized into channels, e.g. bioconda
- **Strengths:** easy to use, good environment description
- **Disadvantages:** packages must already be in channels, when it fails it fails hard...



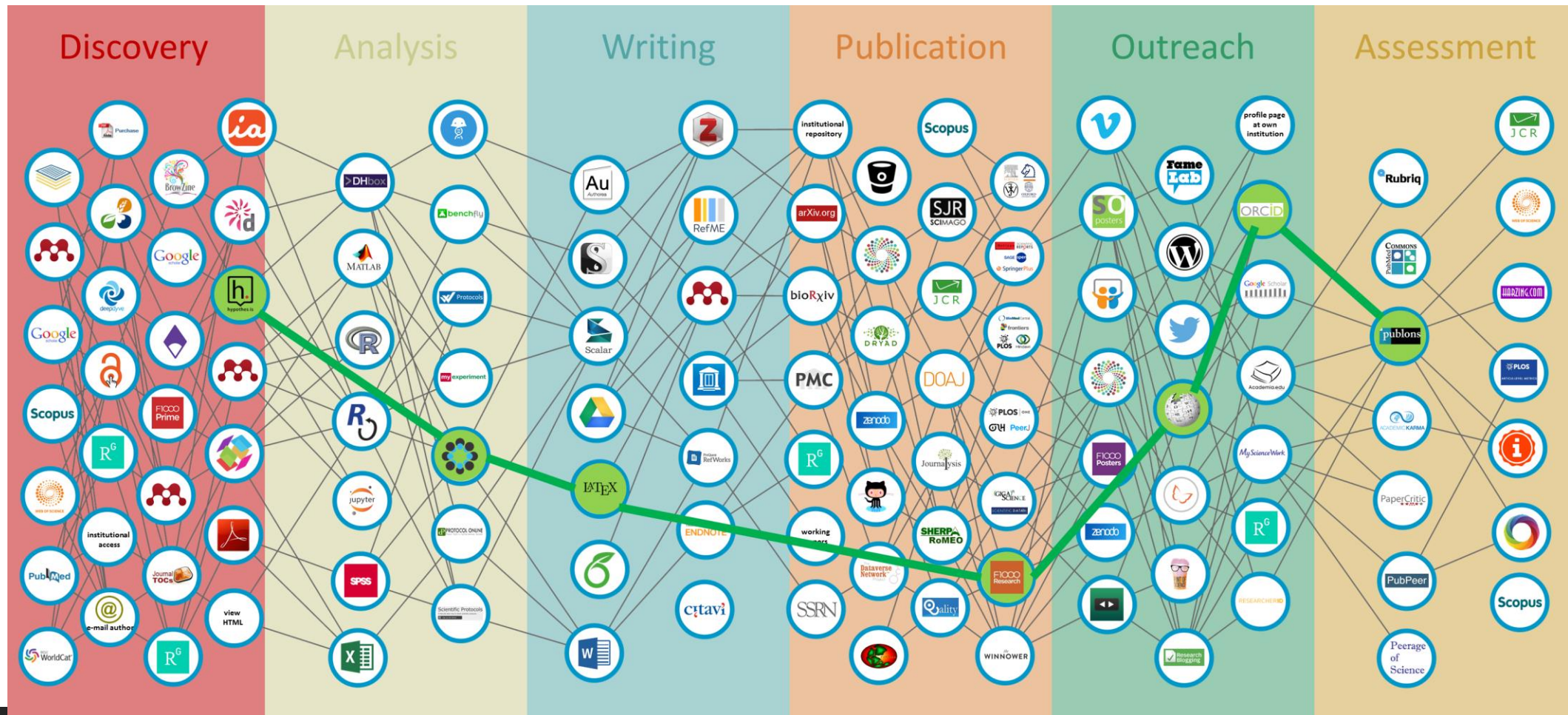
3- Environments

instrumented desktop tools
hosted services
packaging and archiving
repositories, catalogues
online sharing platforms
integrated authoring
integrative frameworks



Environment management = organization and configuration of a set of software

3- Environments, is there a formula?



4 – Presentation

Tables and Figures

- Main vehicles of scientific communication - guide readers through manuscripts
- Visualizing data is very powerful, important, ...and very challenging
- ALL the data underlying a figure must be available in textual form as well

Tables and Tabular Data

- Avoid copy and paste whenever possible!
- Tabular data should (almost) always be included as supplementary materials
- Standardize formats (CSV, not excel!)
- Human- and machine-readable:
 - consistent , controlled textual values, column headers, no comment rows, no irregular formatting, etc

4 – Presentation

- Create figures programmatically from transformed data whenever possible
- Invest in learning plotting libraries
 - matplotlib,
 - seaborn,
 - ggplot,
 - ploty
 - etc
- Output to Scalable Vector Format (SVG) rather than bitmap formats

Replicability

Replication is the best way for the community to verify credibility of a finding

Replication Awards <> www.humanbrainmapping.org/
The purpose:

- Promote replications, by highlighting the best replications studies and their authors
- Cash award of \$2,500 USD and an engraved plaque.

BEHAVIORAL AND BRAIN SCIENCES (2018), Page 1 of 61
doi:10.1017/S0140525X17001972, e120

Making replication mainstream

Rolf A. Zwaan

Department of Psychology, Education, and Child Sciences, Erasmus University Rotterdam, 3000 DR Rotterdam, The Netherlands
zwaan@essb.eur.nl
<https://www.eur.nl/essb/people/rolf-zwaan>

Alexander Etz

Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100.
etz.alexander@gmail.com
<https://alexanderetz.com/>

Richard E. Lucas

Department of Psychology, Michigan State University, East Lansing, MI 48824
lucasri@msu.edu
<https://www.msu.edu/user/lucasri/>

M. Brent Donnellan¹

Department of Psychology, Texas A&M University, College Station, TX 77843
donnel59@msu.edu
<https://psychology.msu.edu/people/faculty/donnel59>

Abstract: Many philosophers of science and methodologists have argued that the ability to repeat studies and obtain similar results is an essential component of science. A finding is elevated from single observation to scientific evidence when the procedures that were used to obtain it can be reproduced and the finding itself can be replicated. Recent replication attempts show that some high profile results – most notably in psychology, but in many other disciplines as well – cannot be replicated consistently. These replication attempts have generated a considerable amount of controversy, and the issue of whether direct replications have value has, in particular, proven to be contentious. However, much of this discussion has occurred in published commentaries and social media outlets, resulting in a fragmented discourse. To address the need for an integrative summary, we review various types of replication studies and then discuss the most commonly voiced concerns about direct replication. We provide detailed responses to these concerns and consider different statistical ways to evaluate replications. We conclude there are no theoretical or statistical obstacles to making direct replication a routine aspect of psychological science.

What makes a good replication?

Dimension 1 (**Importance**): The need for replicating the original finding (1-5).

- Is the original finding used in policy making?
- Did the original finding open a new subfield of research?
- Is there a debate about the original finding?
 - Are there studies undermining the original finding?
 - Are there studies confirming the original finding?

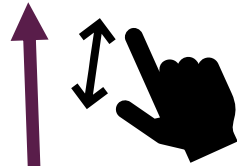
What makes a good replication?

Dimension 2 (**Quality**): Quality of the replication attempt (1-5).

- Was the replication study **pre-registered**?
- Was the study protocol discussed with the original researchers prior to acquiring data and/or performing analysis?
- Was the replication performed by **an independent team of researchers** or was it done by the same people?
- Was the **sample size sufficient** considering the originally reported effect size?
- Were the methods used in the replication attempt in accordance with **current academic standards**?
- Would the **departures from the original protocol** in the replication attempt change the conclusion of the original study if they were applied originally?

What makes a good replication?

IMPORTANCE



Important Topic
POOR QUALITY



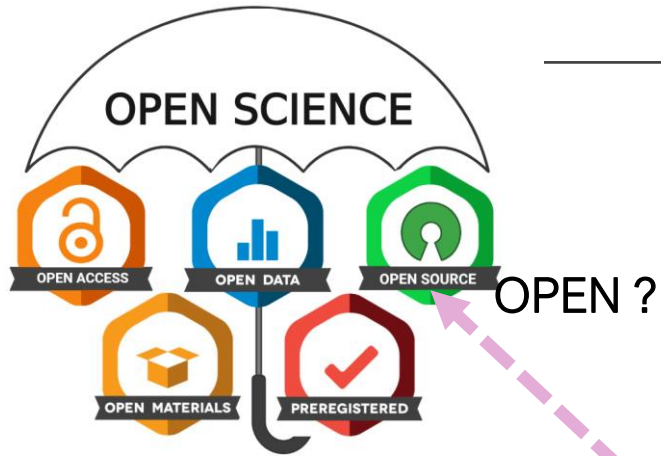
Important Topic
Great Execution



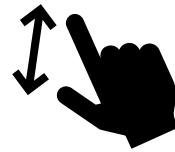
Great Execution
NICHE TOPIC

QUALITY

What makes a good replication?



IMPORTANCE



Important Topic
POOR QUALITY



Important Topic
Great Execution



Great Execution
NICHE TOPIC

QUALITY

FAIR ?

???



Fostering Fair Data Practices in Europe

FAIRsharing.org
standards, databases, policies

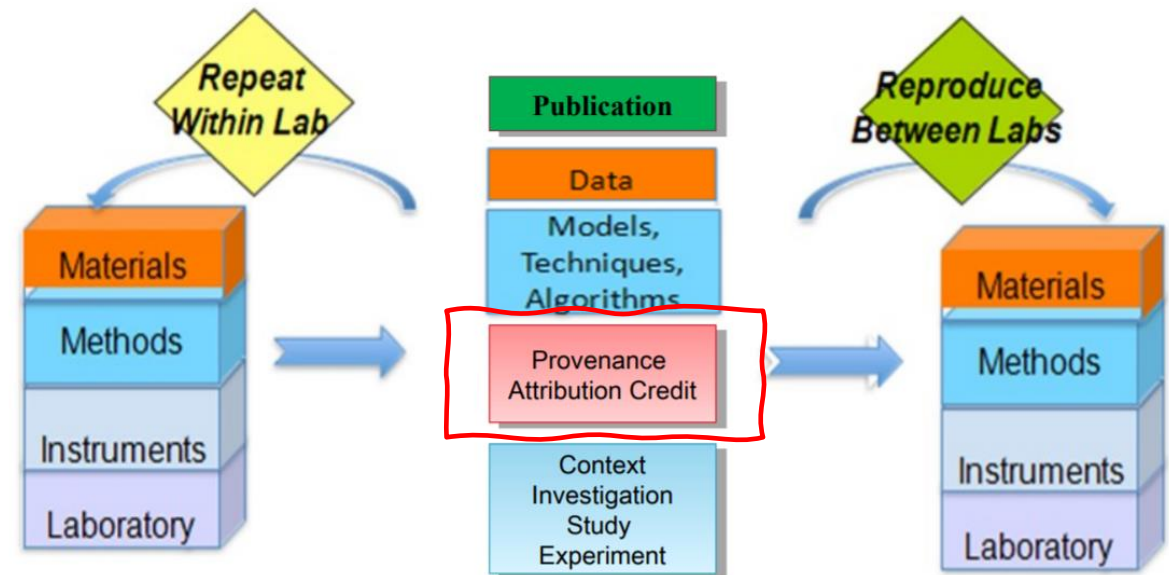
Summary

- High level of reproducibility can be achieved with


- Data sharing
- Code version control
- Software containers


Replication studies

- Require careful planning
- Are a great fit for Registered Reports





 Why GitHub? ▾ Team Enterprise Explore ▾ Marketplace Pricing ▾ Search


 [zavaleta](#) / [Fundamentos_DS](#)

<> Code ⓘ Issues 🔗 Pull requests ⏮ Actions 📁 Projects ⓘ Security 📈 Insights


🔑 main ▾ 🔗 1 branch 🏷 0 tags

Go to file

📄 Code ▾


 **zavaleta V2.1.1** ...

27cf380 18 minutes ago ⌚ 9 commits

 .ipynb_checkpoints


V2.0

1 hour ago

 imagens

V2.0

1 hour ago

 pdf

V2.1

20 minutes ago

Fundamentos de Ciência de Dados

Professores:

Sergio Serra	Jorge Zavaleta
	
serra@pet-si.ufrj.br	zavaleta@pet-si.ufrj.br

Ementa:

Introdução a reprodutibilidade em pesquisa, proveniência de dados e gestão de grandes volumes de dados científicos. Coleta e preparação de dados. Algoritmos de exploração e análise de dados. Métodos de modelagem fluxo de dados. Elaboração de relatórios de resultados através de documentos com código Python incluindo gráficos e tabelas.

Módulo 1:

- Reprodutibilidade em Pesquisa Computacional
- Introdução a Proveniência de Dados
- Gestão de Grandes Volumes de Dados de Pesquisa
- Ambiente de Programação: python 3, jupyter notebook, JupyterLab, Google Colab, DeepNote, pacotes e github. PDF: [Teoria](#)
- Python I: tipos de dados, sequências e operações, estruturas de controle e repetição. Tipos de Dados em Python: [Tipos](#)
- Prática dos conteúdos estudados: construindo e operando listas e strings.
- Aulas: [PDF]

Módulo 2:

- Técnicas de coleta e preparação de dados
- Numpy I: array, slicing, fancy index, copy and view
- Pandas I: dataframes, series, index, Pandas I/O (csv, json, excel)
- Prática dos conteúdos estudados: Processando e extraindo informações de arquivos csv, Jason, rdf
- Aulas: [PDF]

https://github.com/zavaleta/Fundamentos_DS

References

Claerbout J. F., Karrenbach M. (1992). Electronic documents give reproducible research a new meaning. SEG Expanded Abstracts 11, 601–604. 10.1190/1.1822162

Delescluse, Matthieu, et al. (2012). Making neurophysiological data analysis reproducible: Why and how?. Journal of Physiology-Paris 106.3 159-170.

Donoho D. L., Maleki A., Rahman I. U., Shahram M., Stodden V. (2009). 15 Years of reproducible research in computational harmonic analysis. Comput. Sci. Eng. 11, 8–18. 10.1109/MCSE.2009.15

Goodman S. N., Fanelli D., Ioannidis J. P. A. (2016). What does research reproducibility mean? Sci. Transl. Med. 8:341ps12. 10.1126/scitranslmed.aaf5027

Peng R. D. (2011). Reproducible research in computational science. Science 334, 1226–1227. 10.1126/science.1213847