Tópicos Especiais SI

Fundamentos de Ciência de Dados

PROF SERGIO SERRA E JORGE ZAVALETA

{SERRA, ZAVALETA} @PPGI.UFRJ.BR

2024.2

Datas Importantes

CALENDÁRIO DE ATIVIDADES ACADÊMICAS PARA 2024

Aprovado na Sessão do CEPG de **10/11/2023**
Aprovado na Sessão do CONSUNI de **14/12/2023 (Resolução CONSUNI/CET Nº 126/2023)**

Atos Acadêmicos no SIGA - Calendário Semestral	**1º período**	**2º período**
Início de atividades	11/03/2024	12/08/2024
Rematrícula de matrícula trancada (destrancamento de matrícula)	Até 08/03/2024	Até 09/08/2024
Previsão de turma	Até 23/02/2024	Até 26/07/2024
Trancamento de matrícula	Até 29/03/2024	Até 30/08/2024
Pedido de inscrição em disciplinas	**De 24/02/2024 a 05/03/2024**	**De 27/07/2024 a 06/08/2024**
Concordância do pedido de inscrição em disciplina	De 06/03/2024 a 07/03/2024	De 07/08/2024 a 08/08/2024
Efetivação do Pedido de Inscrição (Divisão de Ensino – PR2)	08/03/2024	09/08/2024
Pedido de alteração de inscrição em disciplina	De 09/03/2024 a 12/03/2024	De 10/08/2024 a 13/08/2024
Concordância do pedido de alteração de inscrição em disciplina	De 13/03/2024 a 14/03/2024	De 14/08/2024 a 15/08/2024
Efetivação de Alteração do Pedido de Inscrição (Divisão de Ensino – PR2)	15/03/2024	16/08/2024
Trancamento do pedido de inscrição (desistência de inscrição)	De 16/03/2024 a 19/03/2024	De 17/08/2024 a 20/08/2024
Concordância do pedido de trancamento de inscrição	De 20/03/2024 a 21/03/2024	De 21/08/2024 a 22/08/2024
Efetivação do Trancamento do Pedido de Inscrição (Divisão de Ensino – PR2)	22/03/2024	23/08/2024
Término de atividades	20/07/2024	14/12/2024
Notas – Pautas de graus e frequência	De 21/07/2024 a 20/08/2024	De 15/12/2024 a 14/01/2025

Programa

Terças das ~ 13:30 até ~17:00 Teórico–práticas
Lab NCE e Google Meet (excepcionalmente)

Módulo 1:

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

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 + artigos*

Avaliação e Atendimento

Critérios de aprovação são os do PPGI/UFRJ.

A avaliação da disciplina consiste em participação em sala de aula (P); protótipos de DS desenvolvidos com boas práticas (E); apresentações e elaboração de Dataset/Executable Paper (A).

$$MF = 0.1 * P + 0.4 * E + 0.5 * A$$

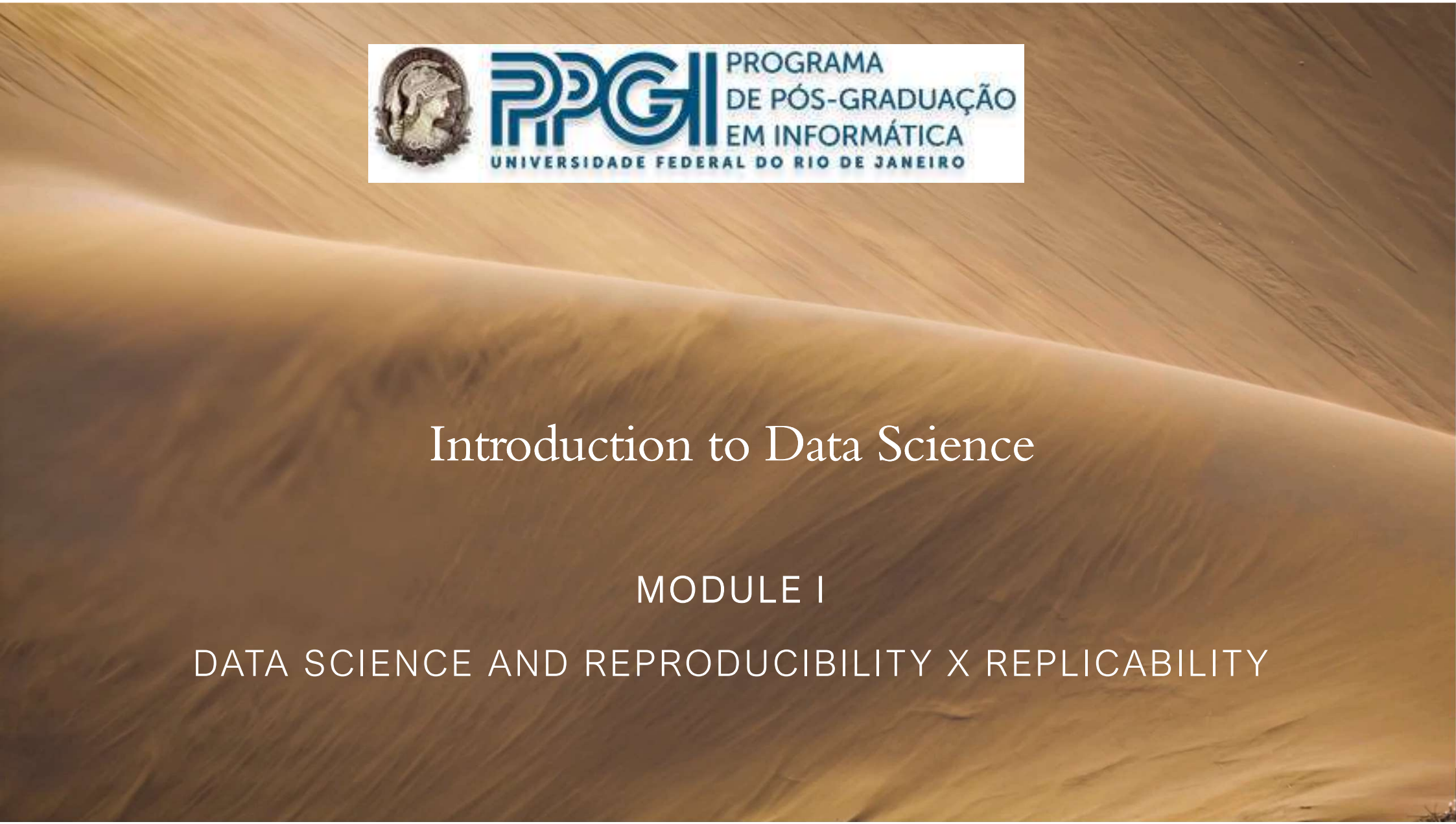
O aluno que desejar atendimento deverá requisitar o mesmo por e-mail e um horário será agendado pelos responsáveis para o atendimento.

serra@ppgi.ufrj.br zavaleta@ppgi.ufrj.br

Bibliografia

Materiais apresentados em sala de aula + extas

1- National Academies of Sciences, Engineering, andMedicine. 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 McKinny, 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 DArcy. 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

DATA SCIENCE AND REPRODUCIBILITY X REPLICABILITY

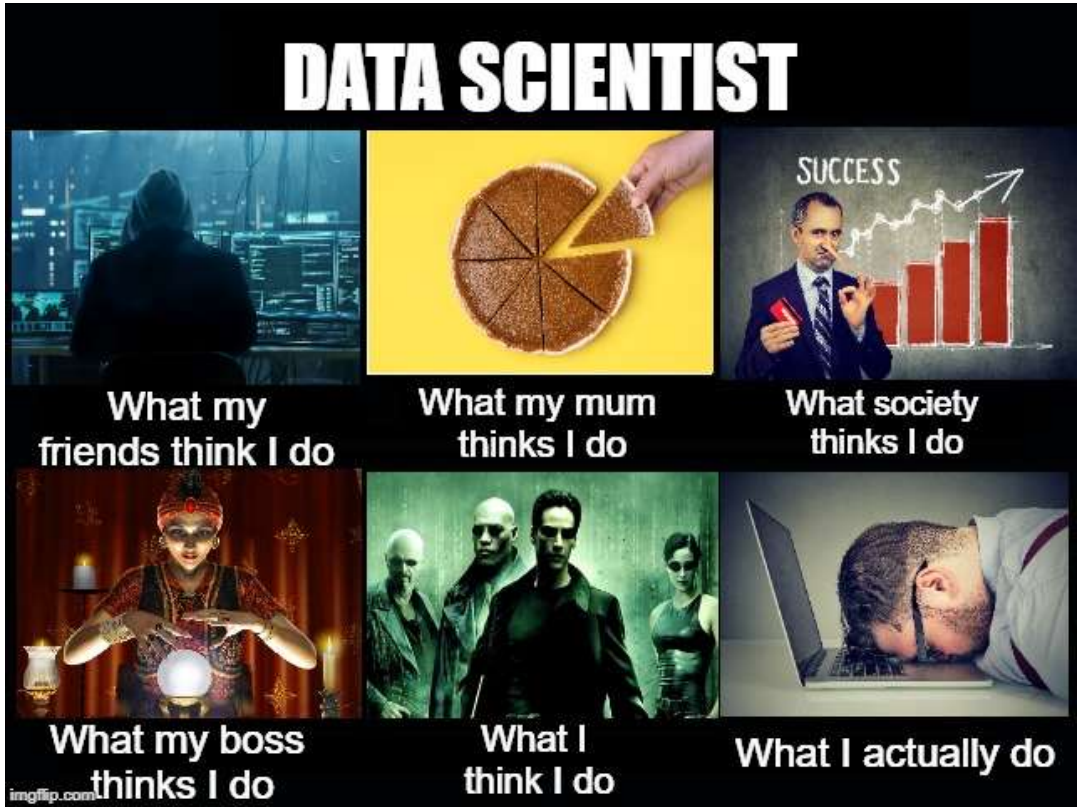
Who are you?

Name?

Class?

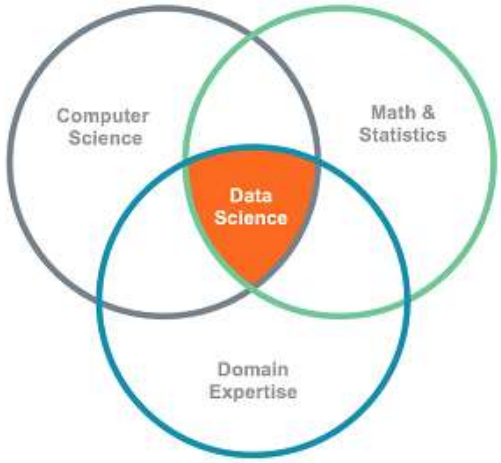
Experience in Data Science?

Expectations?



What Data Science is?

Data science is a powerful combination of various disciplines.



Computer Science Skills

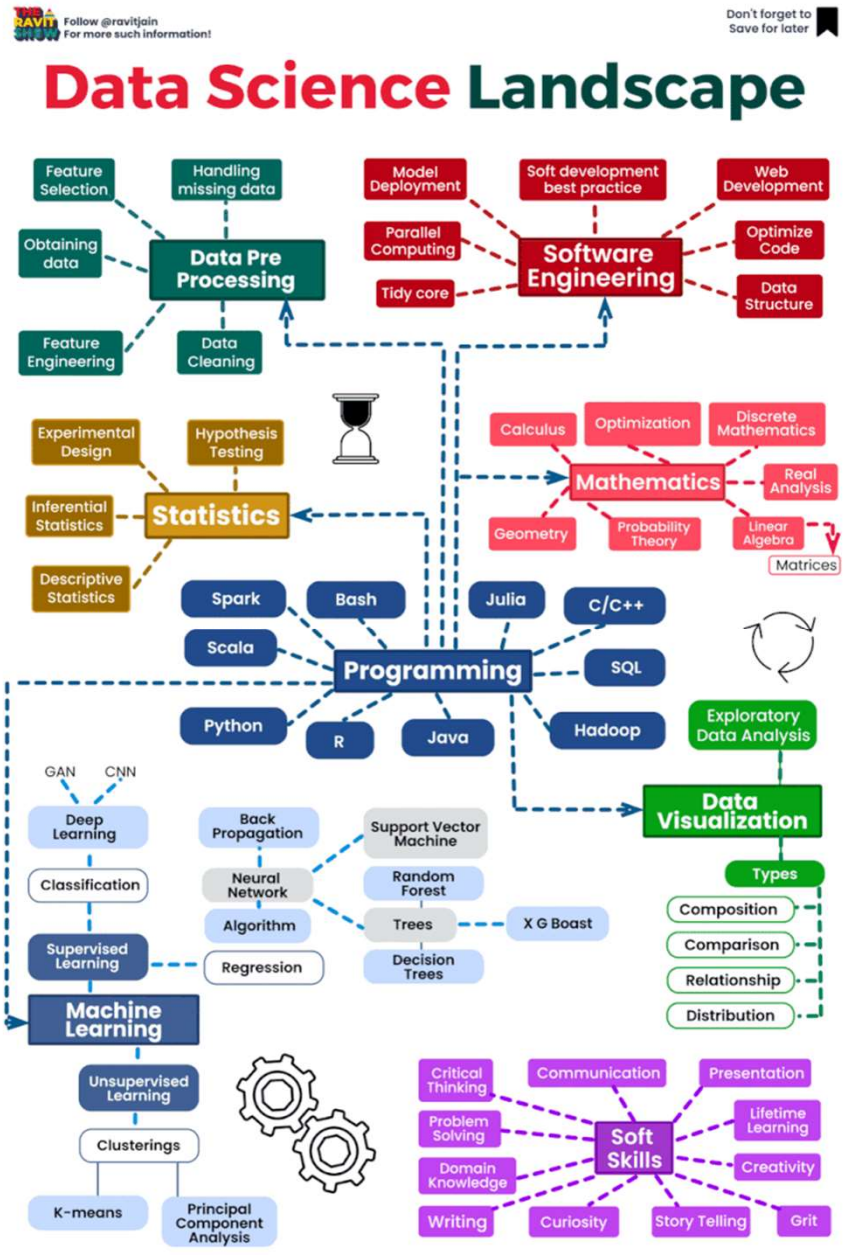
- Programming
- Big data technologies

Math and Statistics Knowledge

- Machine learning
- Ensemble models
- Anomaly detection

Domain Expertise

- Business knowledge
- Expert systems
- User testing



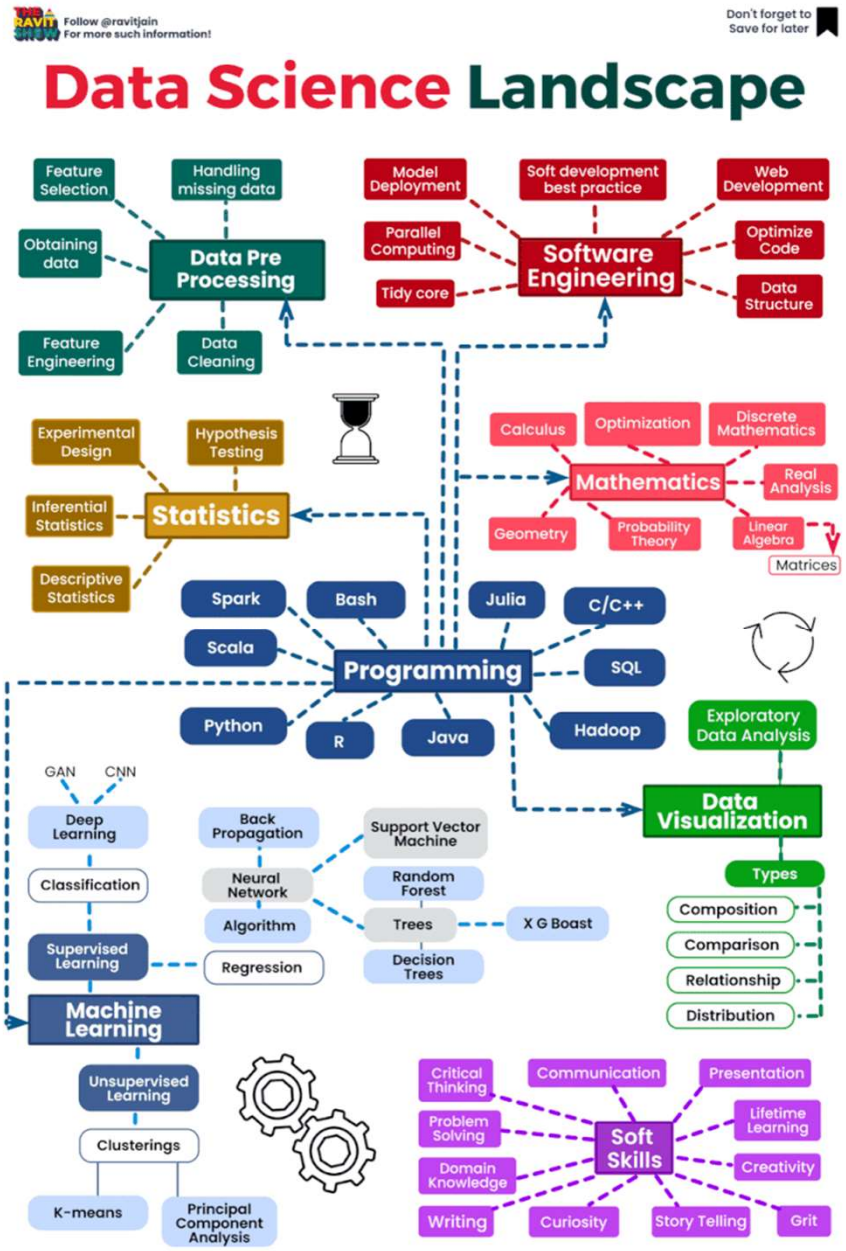
What Data Science is?

Data science is the study of data to gain knowledge and insights that can help inform decisions and predictions for businesses and industries.

It's a multidisciplinary field that combines principles and practices from many scientific disciplines, including mathematics, statistics, business, artificial intelligence, and computer engineering.

Data scientists use tools, methods, and technologies like data analysis, modeling, human-machine interaction, algorithms, and machine learning to extract relevant insights from large amounts of data.

They ask questions like: what happened? why did it happen? what will happen? how can the results be used for planning and decision-making?, etc…



What Data Scientist is?

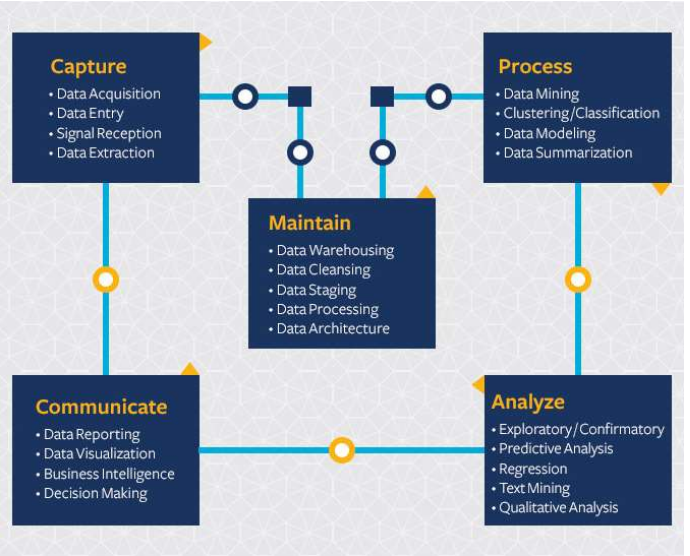
Professional who uses data science principles to solve problems by analyzing and interpreting data to find insights and patterns.

Data scientists often work with analysts and businesses to convert data insights into action.

Create and adjust models to predict future trends, make diagrams, graphs, and charts to represent trends and predictions, and summarize data to help stakeholders understand and implement results.

Data scientists can work in a variety of areas, including: *finance, academia, scientific research, health, retail, information technology, government, and ecommerce.*

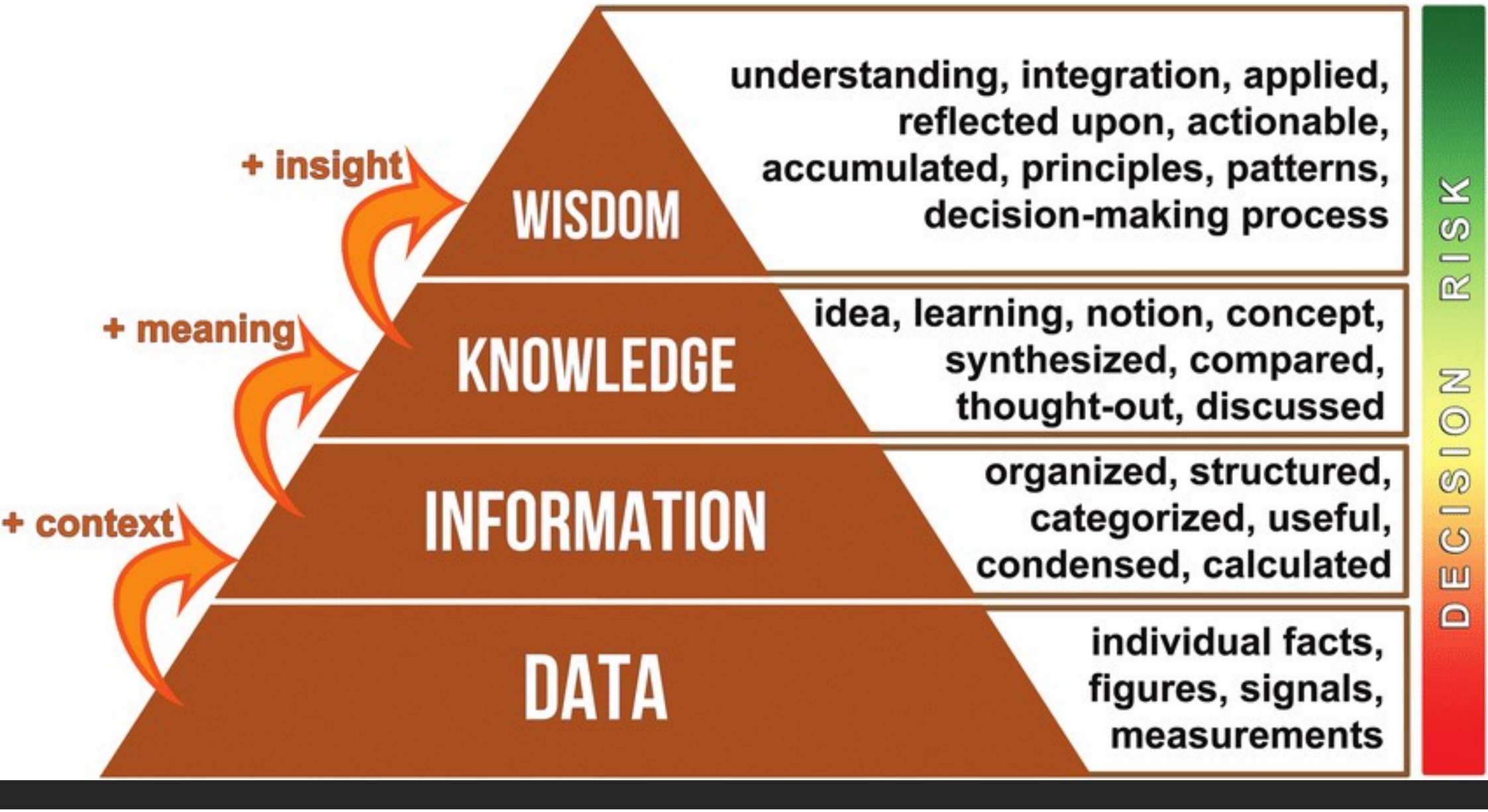
Data scientists need to be effective communicators, leaders, team members, and high-level analytical thinkers. Often need skills in programming languages like Python, R and Machine Learning Techniques.



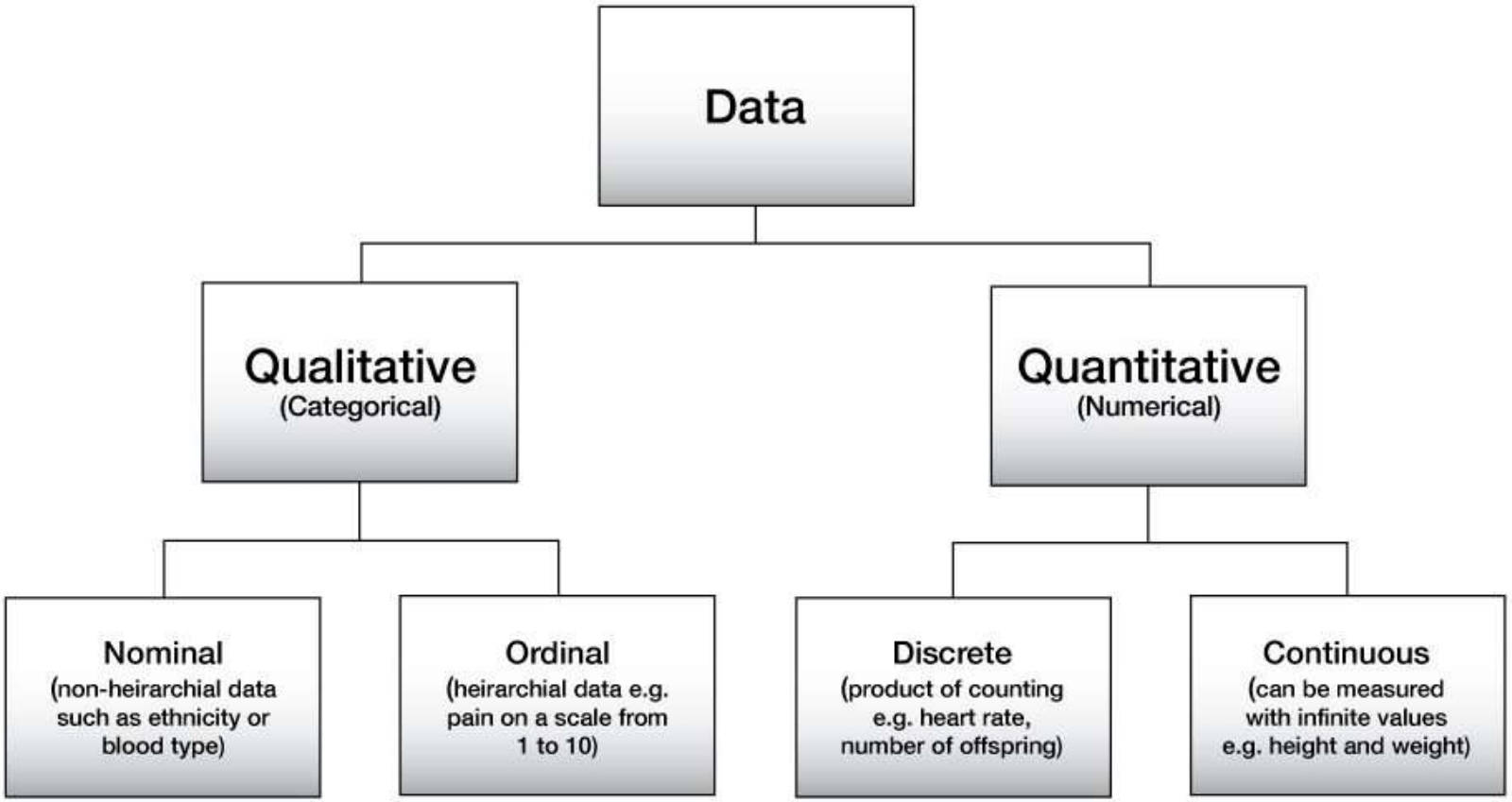
Motivation

1. Data science is a vital part of every business
2. Growing Demand for Data Scientists
3. Data Science Base Salaries are Hard to Beat
4. Data is an Enchanting and an Ever-Evolving Field
5. Diverse Career Growth Options for Data Scientists
6. Contribute to Society

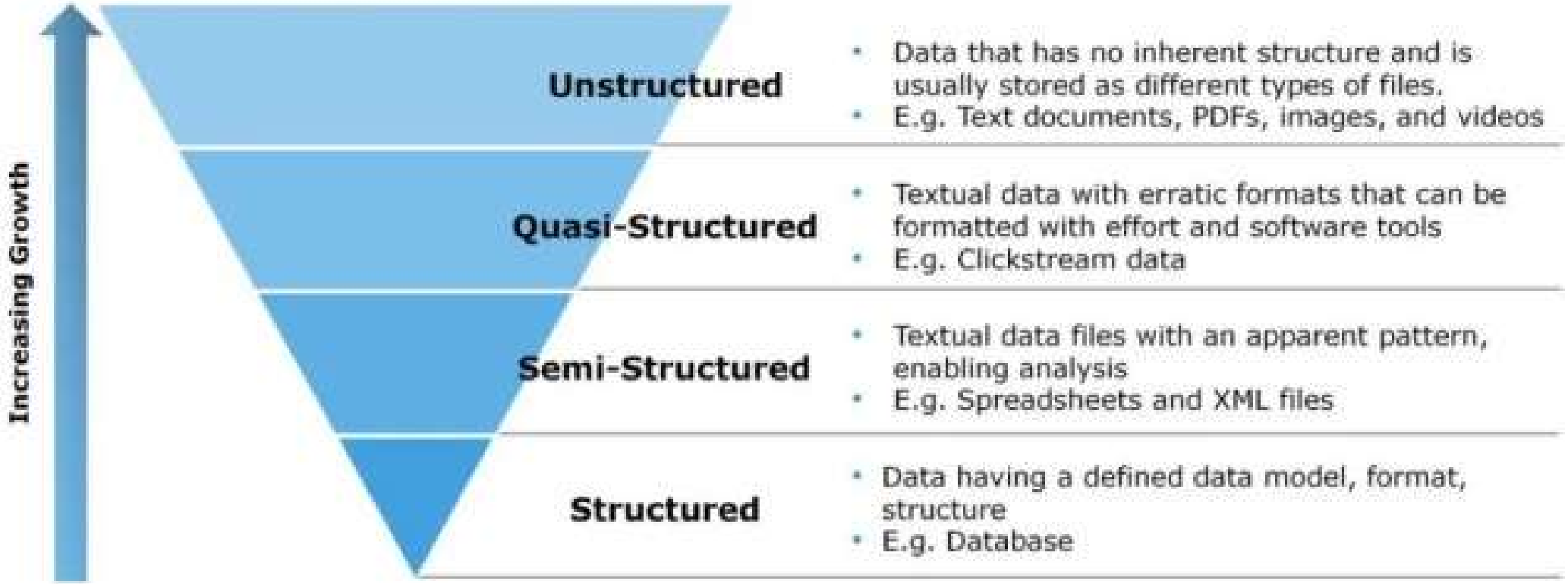
YOU MUST UNDERSTAND THE DIFFERENCE BETWEEN DATA SCIENCE VS. COMPUTER SCIENCE


+ insight
+ meaning
+ context
WISDOM
KNOWLEDGE
INFORMATION
DATA
understanding, integration, applied, reflected upon, actionable, accumulated, principles, patterns, decision-making process
idea, learning, notion, concept, synthesized, compared, thought-out, discussed
organized, structured, categorized, useful, condensed, calculated
individual facts, figures, signals, measurements
DECISION RISK

Types of Data

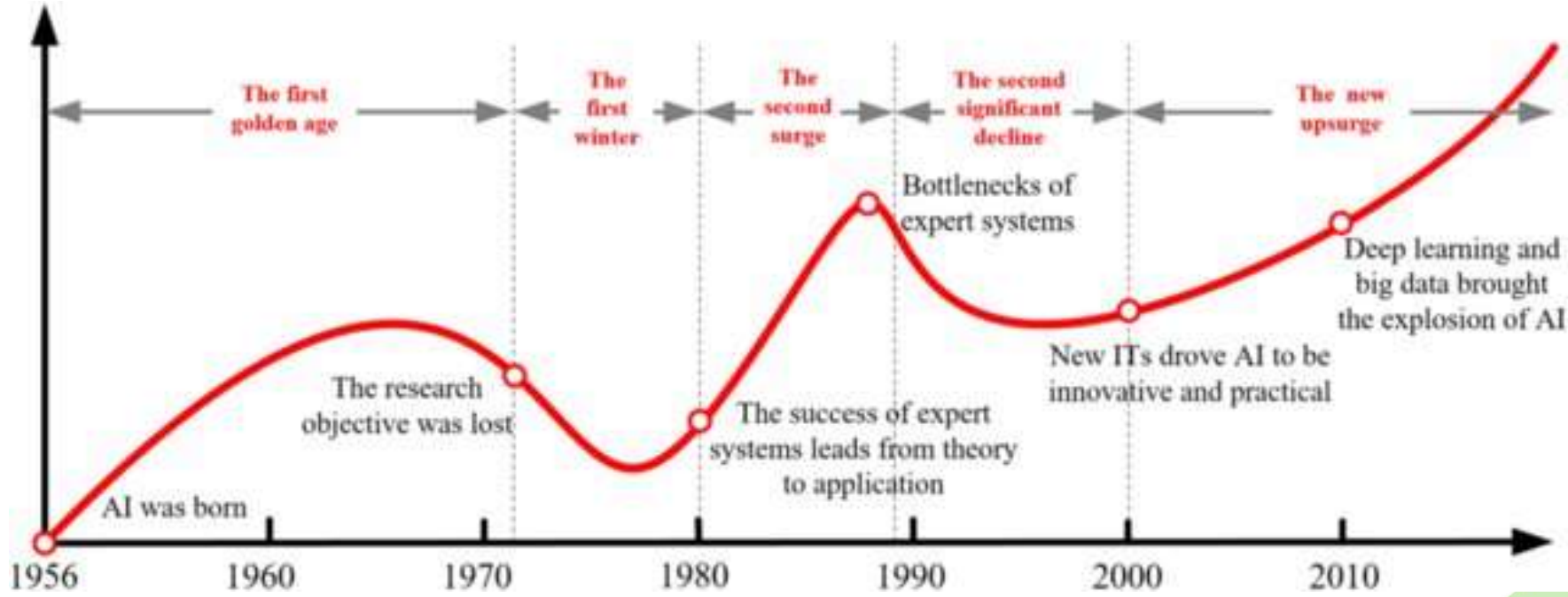


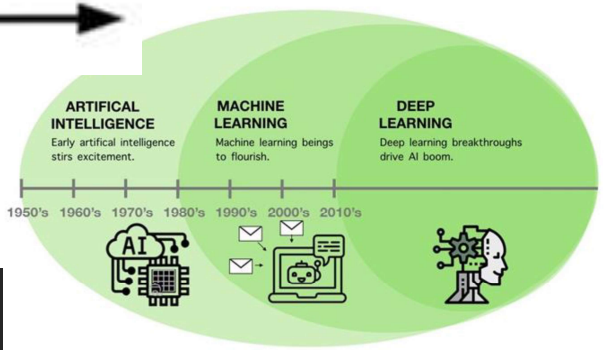
Types of Digital Data



Fonte: https://mycloudwiki.com/san/data-and-information-basics/

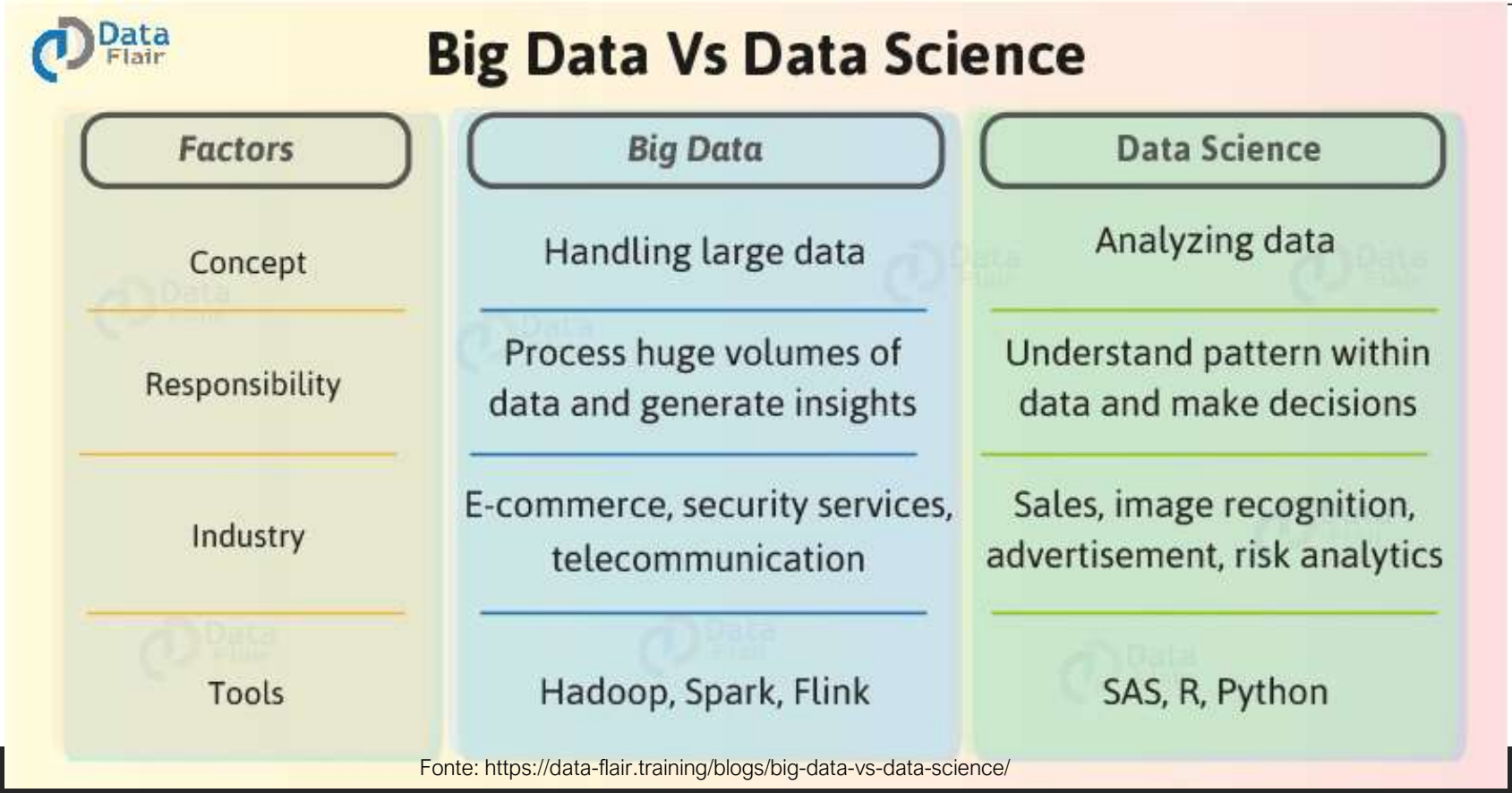
Data Science X AI





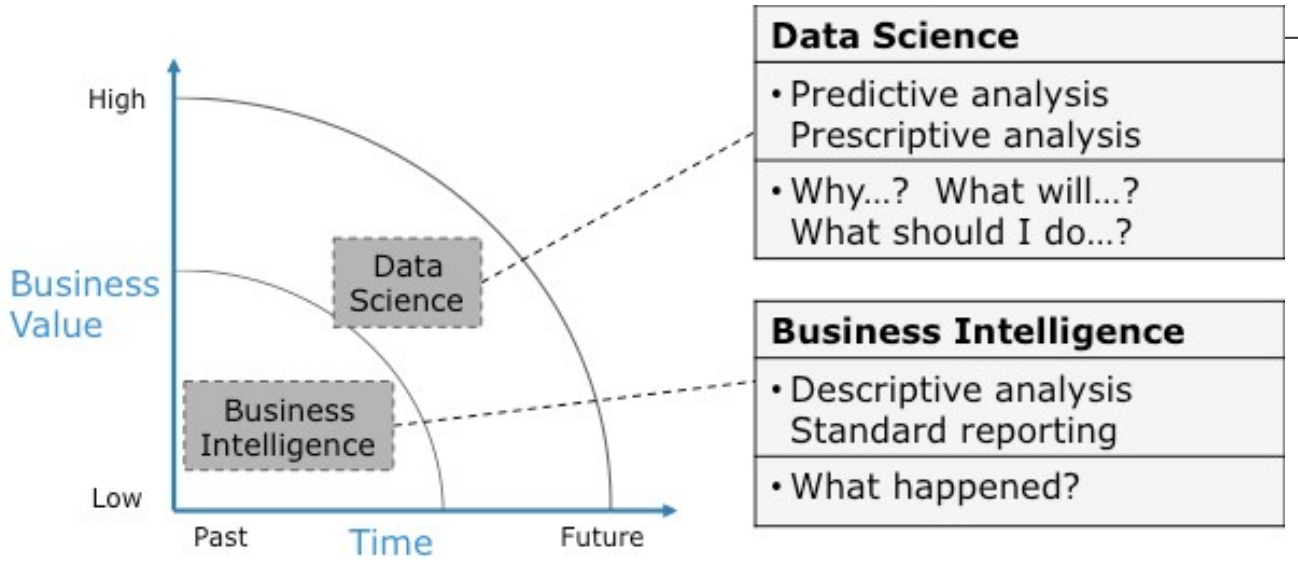
•DOI: 10.1007/s00170-021-06882-1

Data Science x Big Data



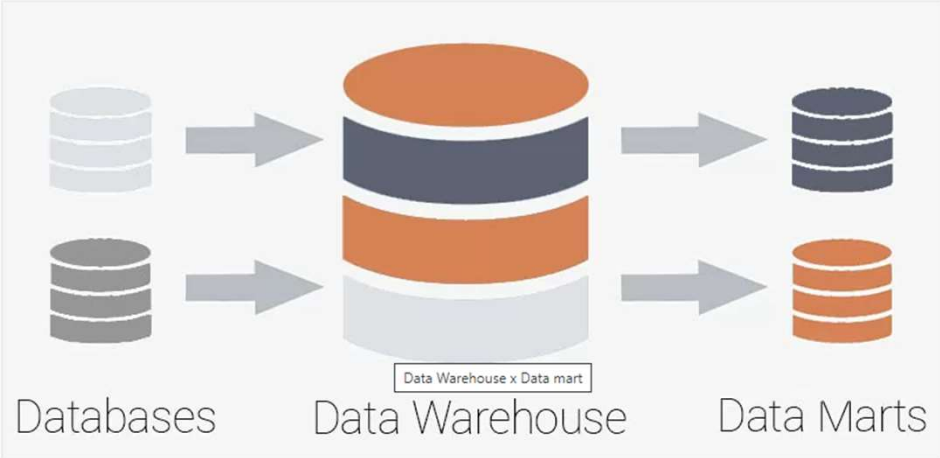
Fonte: https://data-flair.training/blogs/big-data-vs-data-science/

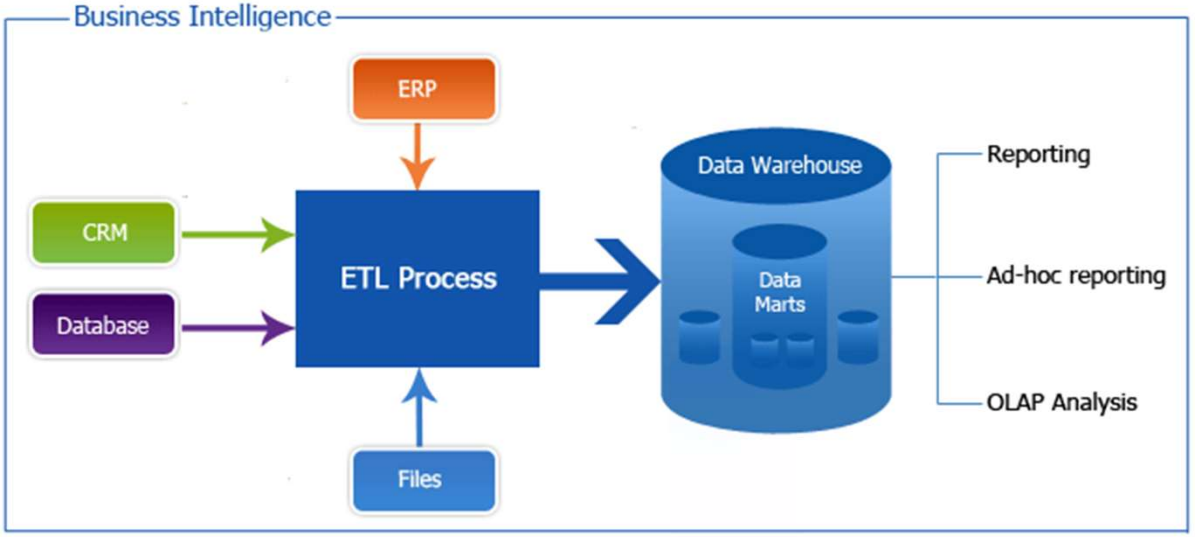
Data Science x BI



https://blog.dsacademy.com.br/qual-a-diferenca-entre-o-analista-de-bi-e-o-cientista-de-dados/

Data Science x DW



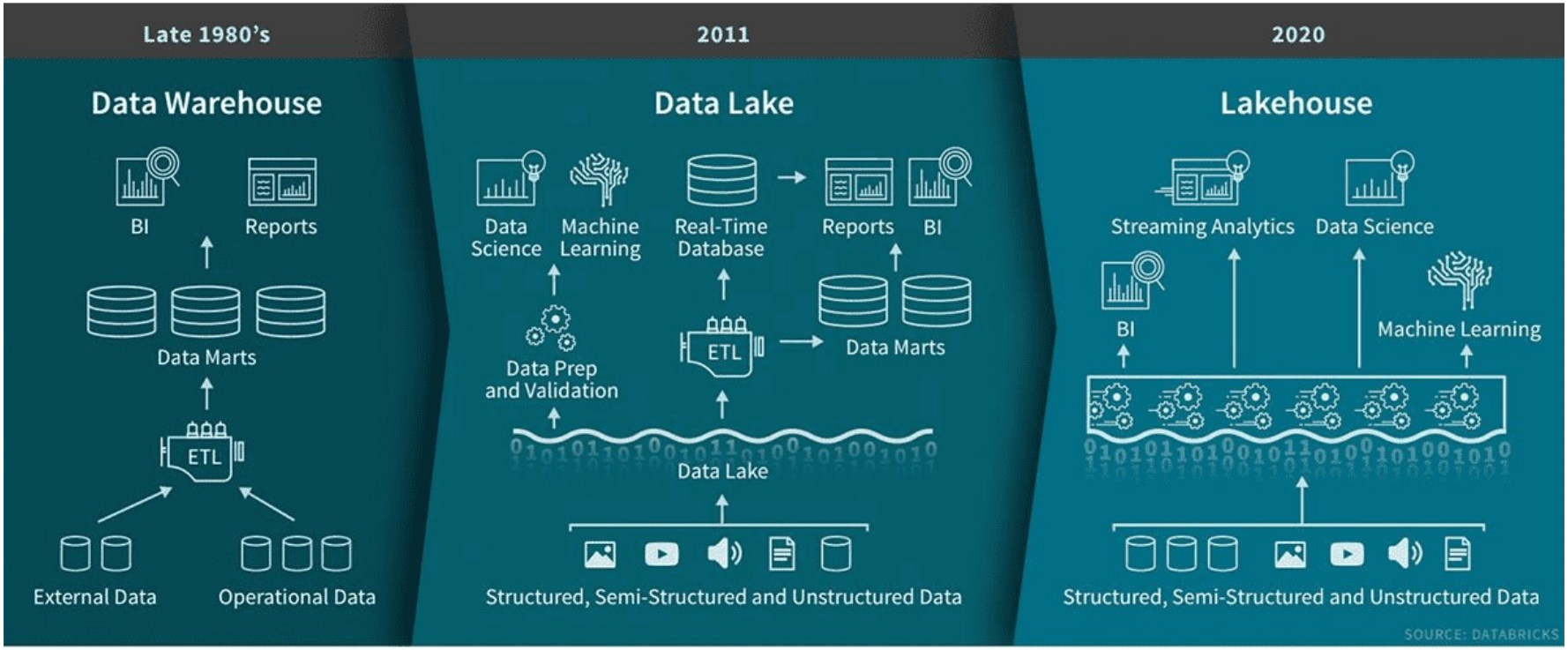


DW x Data Lake

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.

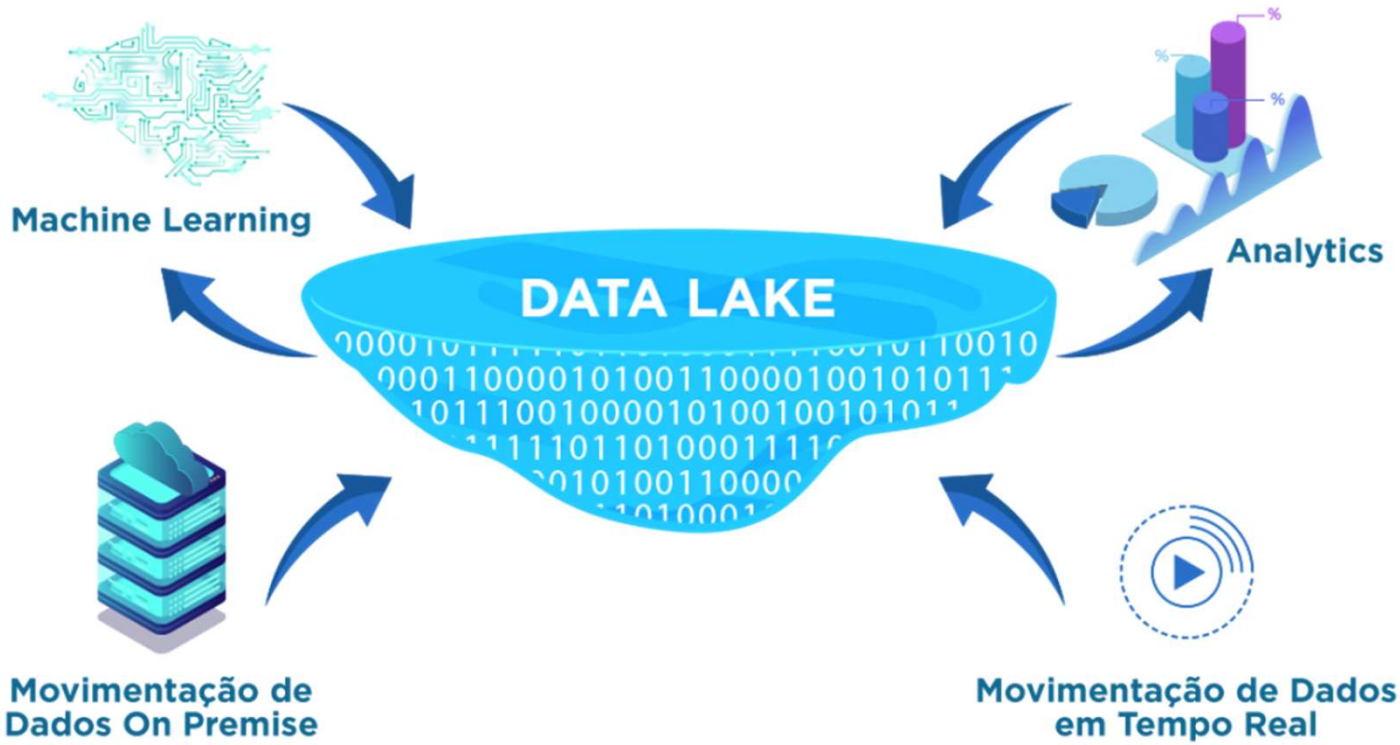
Fonte: https://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html

DW X Dala LakeHouse



Fonte: https://brains.dev/2023/data-warehouse-x-data-lake-x-data-lakehouse/

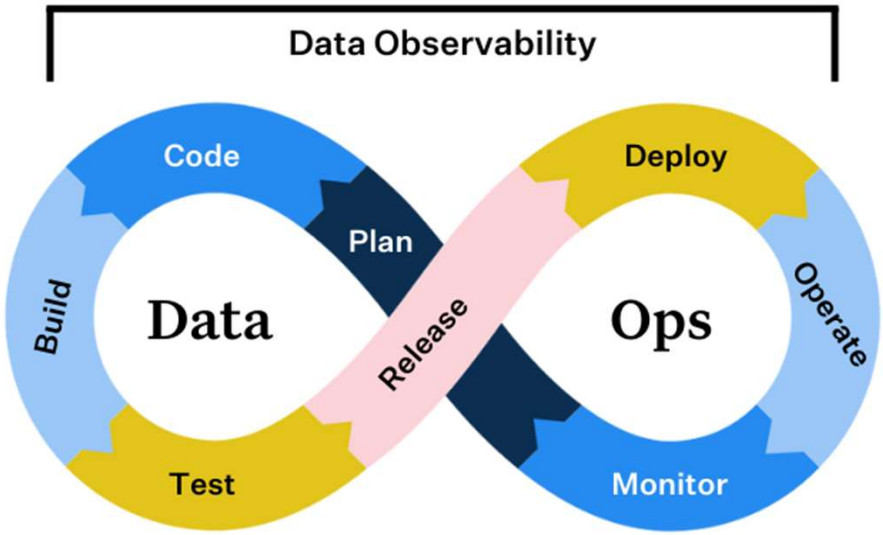
Dara Science x Data Lake



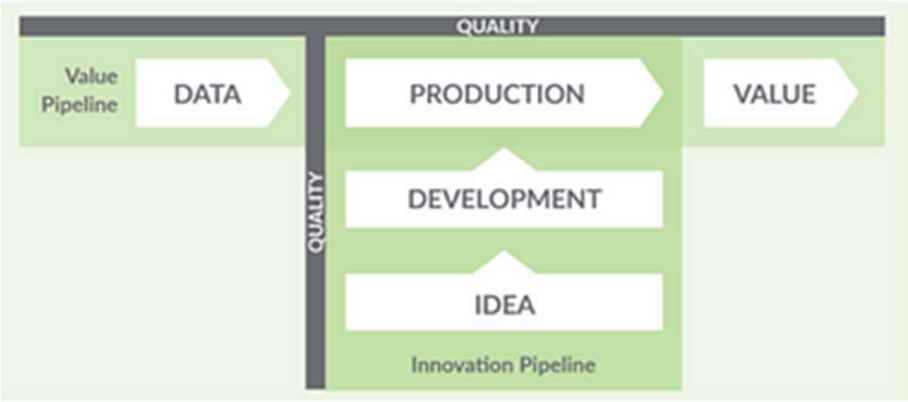
DataOps

DataOps is a discipline that merges **data engineering and data science** teams to support an organization's data needs, similar to how DevOps helps organizations scale software engineering

DataOps is a **set of practices, processes and technologies** that combines an **integrated and process-oriented perspective on data with automation and methods from agile software engineering** to improve quality, speed, and collaboration and promote a culture of continuous improvement in the area of data analytics

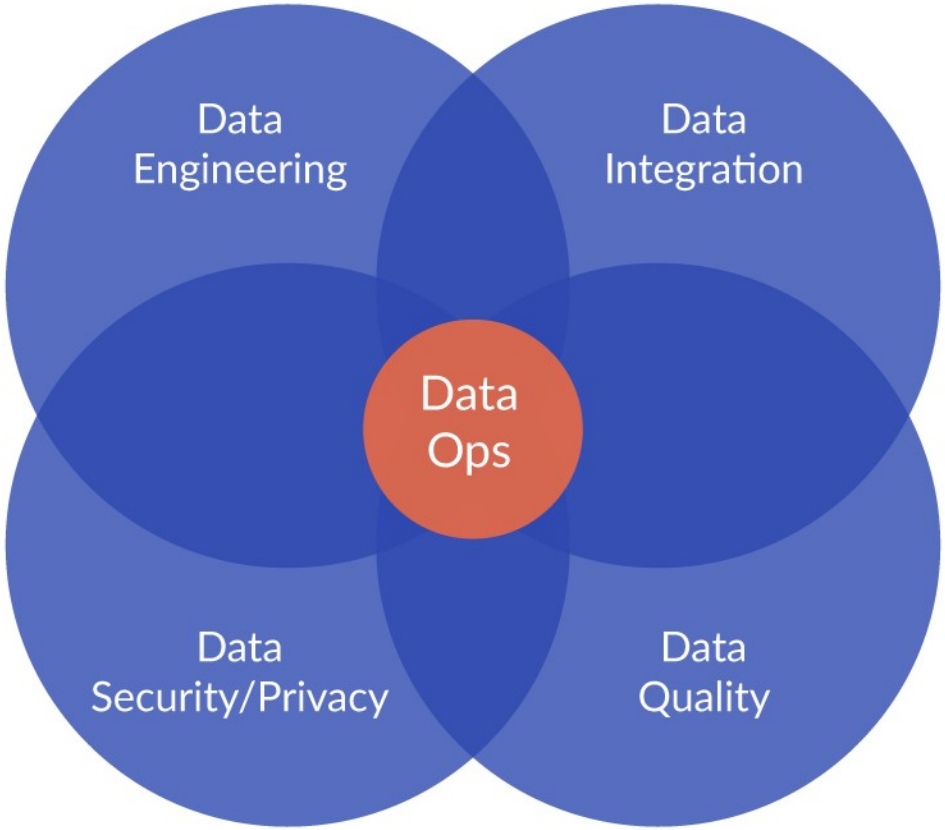


DataOps



DataOps

https://medium.com/data-ops/dataops-is-not-just-devops-for-data-6e03083157b7



Fonte: https://esimplicity.com/works/data-ops-for-field-agents/

Data Engineering

Data Engineers are the link between the management's big data strategy and the data scientists that need to work with data.

What they do is building the platforms that enable data scientists to do their magic.

These platforms are usually used in five different ways:

- Data ingestion and storage of large amounts of data
- Algorithm creation by data scientists
- Automation of the data scientist's machine learning models and algorithms for production use
- Data visualisation for employees and customers
- Most of the time these guys start as traditional solution architects for systems that involve SQL databases, web servers, SAP installations and other "standard" systems

Data Engineering

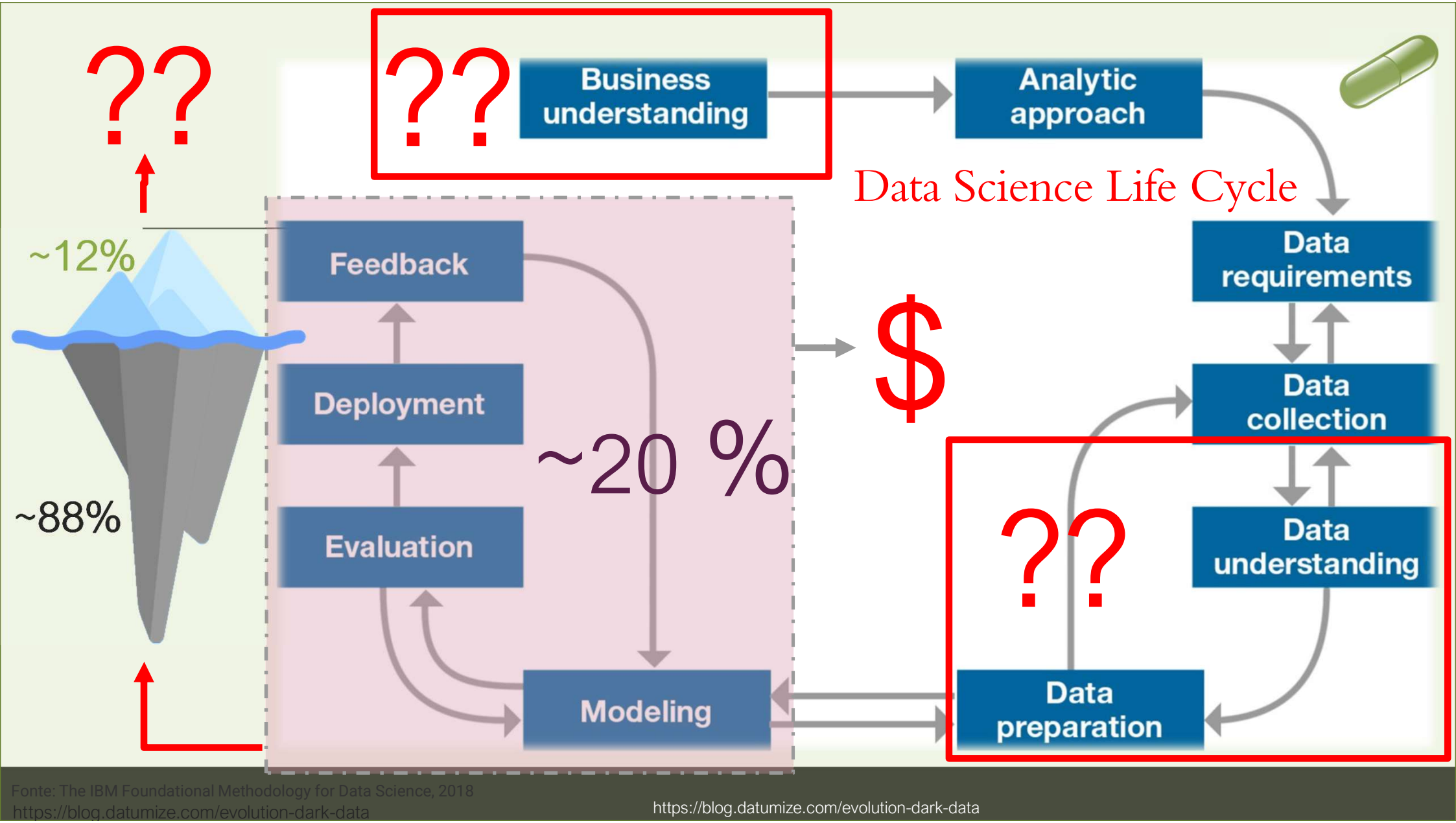
To create big data platforms the engineer needs to be an expert in specifying, setting up and maintaining big data technologies like:

Hadoop, Spark, HBase, Cassandra, MongoDB, Kafka, Redis and more.

What they also need is experience on how to deploy systems on cloud infrastructure like at Amazon or Google or on premise hardware.

Data Science Life Cycle

1. Understand the problem and set goals - What problem am I solving?
2. Collect and analyze the data - What information do I need?
3. Prepare the data - How do I need to process the data?
4. Build the model - What are the patterns in the data that lead to solutions?
5. Evaluate and critique the model - Does the model solve my problem?
6. Present results - How can I solve the problem?
7. Deploy the model - How do I solve the problem in the real world?


??
??
Business understanding
Analytic approach
Data Science Life Cycle
~12%
~88%
Feedback
Deployment
Evaluation
~20 %
Modeling
$
Data requirements
Data collection
Data understanding
??
Data preparation
Fonte: The IBM Foundational Methodology for Data Science, 2018
https://blog.datumize.com/evolution-dark-data
https://blog.datumize.com/evolution-dark-data

The roles of human in the cycle

1 - Precise formulation of objectives

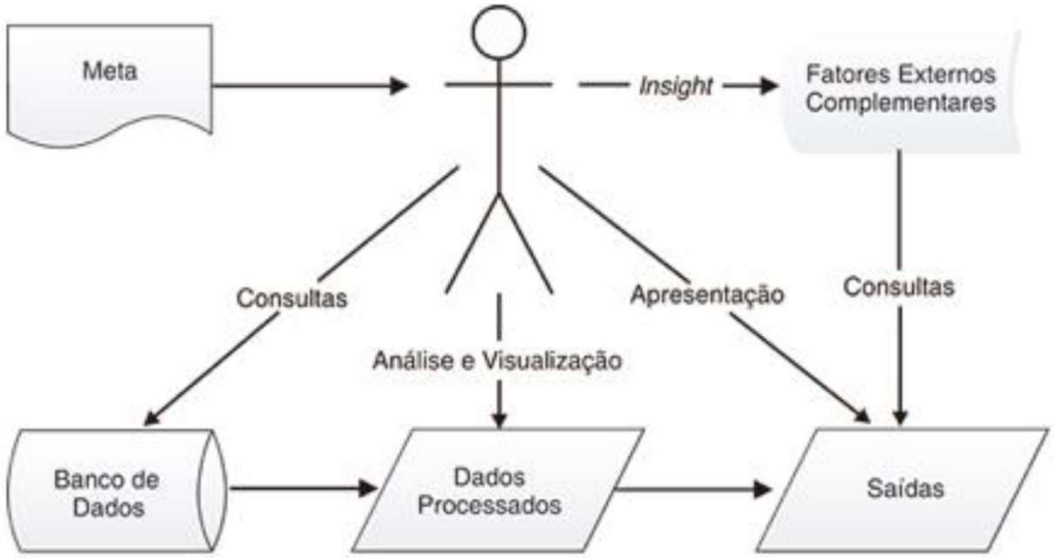
2 - Choice of algorithms

3 - Choice of data preprocessing techniques

4 - Parameterization of algorithms

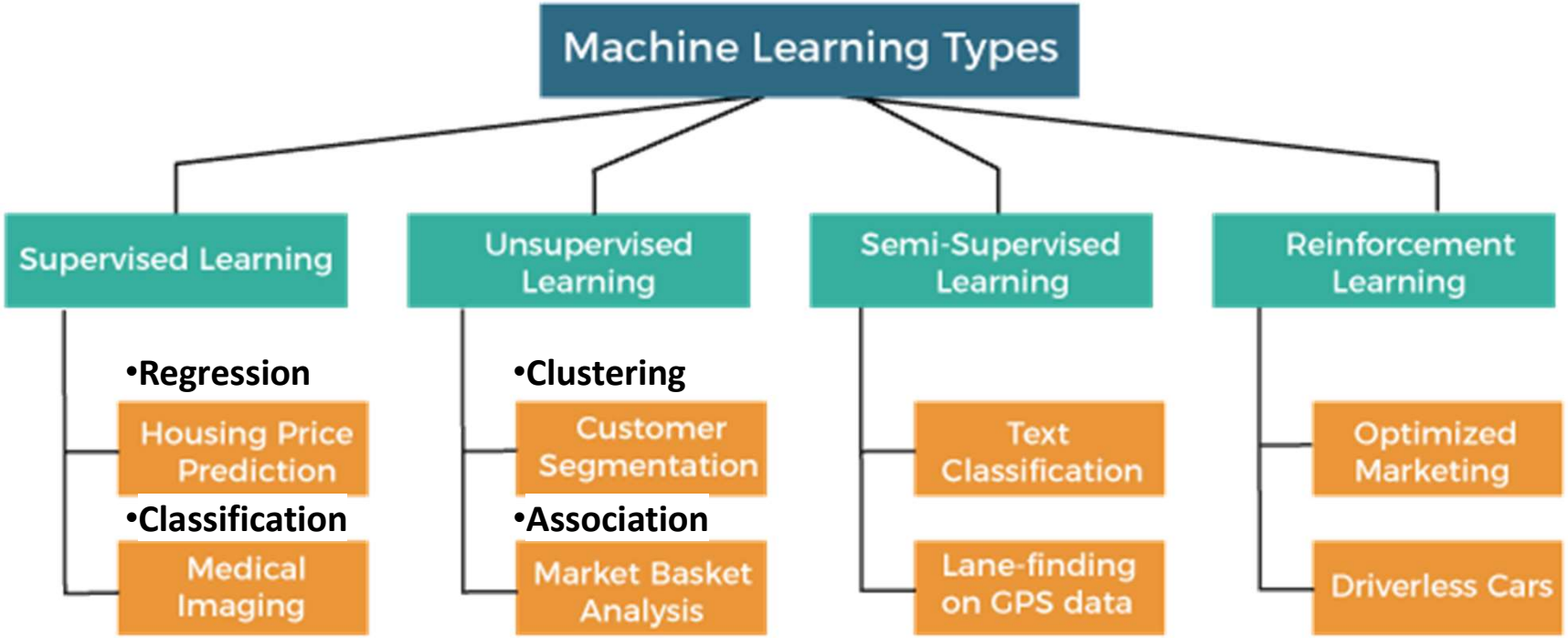
5 - Experience and Knowledge

6 - Intuition



The expert's sense cannot be ignored,even if the most sophisticated techniques are used.

Type of Data Science Problems (basic)



https://www.javatpoint.com/types-of-machine-learning

Supervised Machine Learning

The supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output.

First, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

a) Classification

- Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "Yes" or No, Male or Female, Red or Blue, etc. The classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms are Spam Detection, Email filtering, etc.

b) Regression

- Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Supervised Machine Learning

Some popular classification algorithms:

1. Random Forest Algorithm
2. Decision Tree Algorithm
3. Logistic Regression Algorithm
4. Support Vector Machine Algorithm

Some popular regression algorithms :

1. Simple Linear Regression Algorithm
2. Multivariate Regression Algorithm
3. Decision Tree Algorithm
4. Lasso Regression

Advantages and Disadvantages of Supervised Learning

Advantages:

Since supervised learning work with the labelled dataset so we can have an exact idea about the classes of objects.
These algorithms are helpful in predicting the output on the basis of prior experience.

Disadvantages:

These algorithms are not able to solve complex tasks.
It may predict the wrong output if the test data is different from the training data.
It requires lots of computational time to train the algorithm.

Unsupervised Machine Learning

Unsupervised learning is different from the Supervised learning; as its name suggests, there is no need for supervision. It means, the machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision.

In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision.

The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences. Machines are instructed to find the hidden patterns from the input dataset.

- **Clustering**
- **Association**

Unsupervised Machine Learning

1) Clustering

- The clustering technique is used when we want to find the inherent groups from the data. It is a way to group the objects into a cluster such that the objects with the most similarities remain in one group and have fewer or no similarities with the objects of other groups. An example of the clustering algorithm is grouping the customers by their purchasing behaviour.

Popular clustering algorithms :

- **K-Means Clustering algorithm**
- **Mean-shift algorithm**
- **DBSCAN Algorithm**
- **Principal Component Analysis**
- **Independent Component Analysis**

2) Association

Association rule learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset. The main aim of this learning algorithm is to find the dependency of one data item on another data item and map those variables accordingly so that it can generate maximum profit. This algorithm is mainly applied in Market Basket analysis, Web usage mining, continuous production, etc.

Popular clustering algorithms :

- **Apriori Algorithm**
- **Eclat**
- **FP-growth algorithm**

Unsupervised Machine Learning

1) Clustering

- The clustering technique is used when we want to find the inherent groups from the data. It is a way to group the objects into a cluster such that the objects with the most similarities

2) Association

Association rule learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset. The main aim of this learning algorithm is to find the

Advantages and Disadvantages of Unsupervised Learning

Advantages:

Algorithms can be used for complicated tasks compared to the supervised ones because these algorithms work on the unlabeled dataset.

Unsupervised algorithms are preferable for various tasks as getting the unlabeled dataset is easier as compared to the labelled dataset.

Disadvantages:

The output of an unsupervised algorithm can be less accurate as the dataset is not labelled, and algorithms are not trained with the exact output in prior.

Working with Unsupervised learning is more difficult as it works with the unlabelled dataset that does not map with the output.

Semi-Supervised Learning

Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning.

It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabeled datasets during the training period.

The main is to effectively use all the available data, rather than only labelled data like in supervised learning. Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labelled data. It is because labelled data is a comparatively more expensive acquisition than unlabeled data.

Semi-Supervised Learning

Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning.

It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabeled datasets during the training period.

Advantages and Disadvantages of Unsupervised Learning

Advantages:
It is simple and easy to understand the algorithm.
It is highly efficient.
It is used to solve drawbacks of Supervised and Unsupervised Learning algorithms.

Disadvantages:
Iterations results may not be stable.
We cannot apply these algorithms to network-level data.
Accuracy is low.

Reinforcement Learning

Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance.

Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards.

In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only.

The reinforcement learning process is similar to a human being; for example, a child learns various things by experiences in his day-to-day life. An example of reinforcement learning is to play a game, where the Game is the environment, moves of an agent at each step define states, and the goal of the agent is to get a high score. Agent receives feedback in terms of punishment and rewards.

Reinforcement Learning

Reinforcement learning is categorized mainly into two types of methods/algorithms:

- Positive Reinforcement Learning: Positive reinforcement learning specifies increasing the tendency that the required behaviour would occur again by adding something. It enhances the strength of the behaviour of the agent and positively impacts it.
- Negative Reinforcement Learning: Negative reinforcement learning works exactly opposite to the positive RL. It increases the tendency that the specific behaviour would occur again by avoiding the negative condition.

Advantages and Disadvantages of Unsupervised Learning

Advantages:

It helps in solving complex real-world problems which are difficult to be solved by general techniques.

The learning model of RL is similar to the learning of human beings; hence most accurate results can be found.

Helps in achieving long term results.

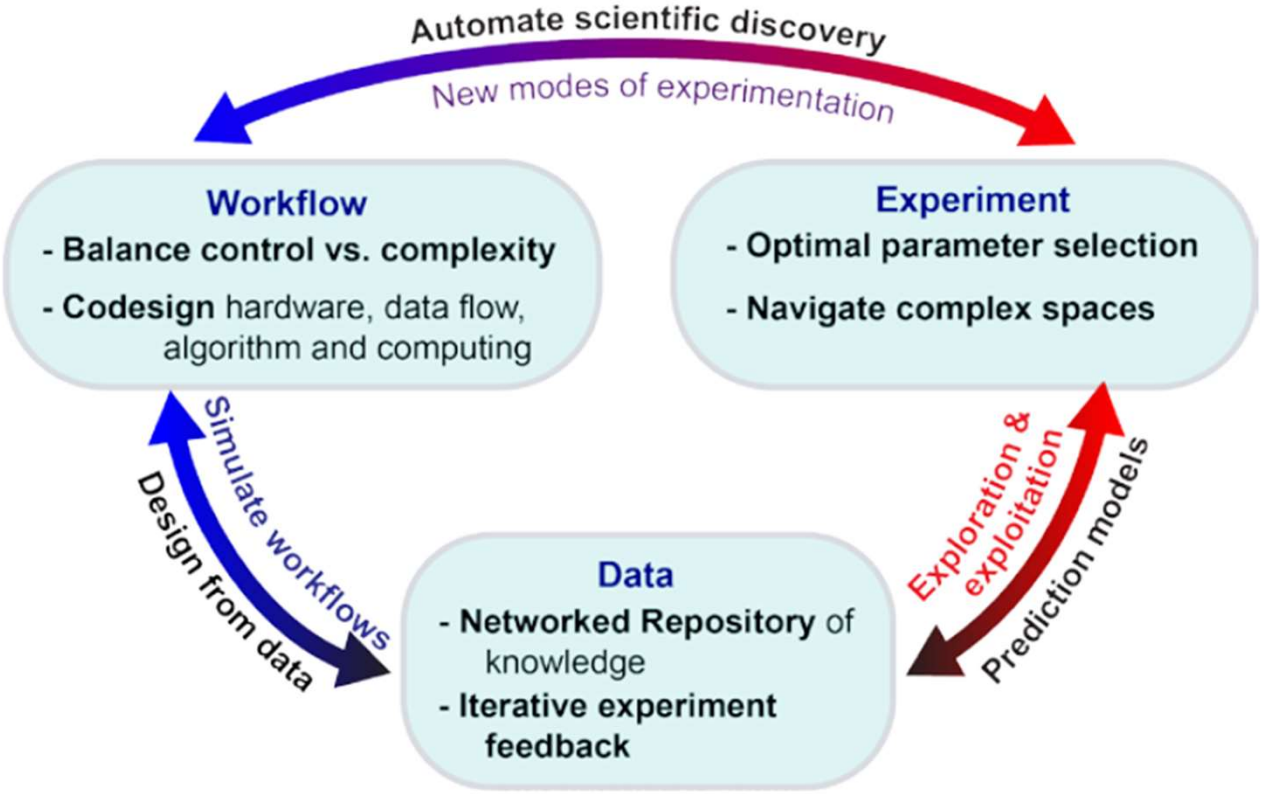
Disadvantages:

RL algorithms are not preferred for simple problems.

RL algorithms require huge data and computations.

Too much reinforcement learning can lead to an overload of states which can weaken the results.

Data Science in 21st Century…



nature reviews chemistry

Explore content | About the journal | Publish with us | Subscribe

nature > nature reviews chemistry > expert recommendation > article

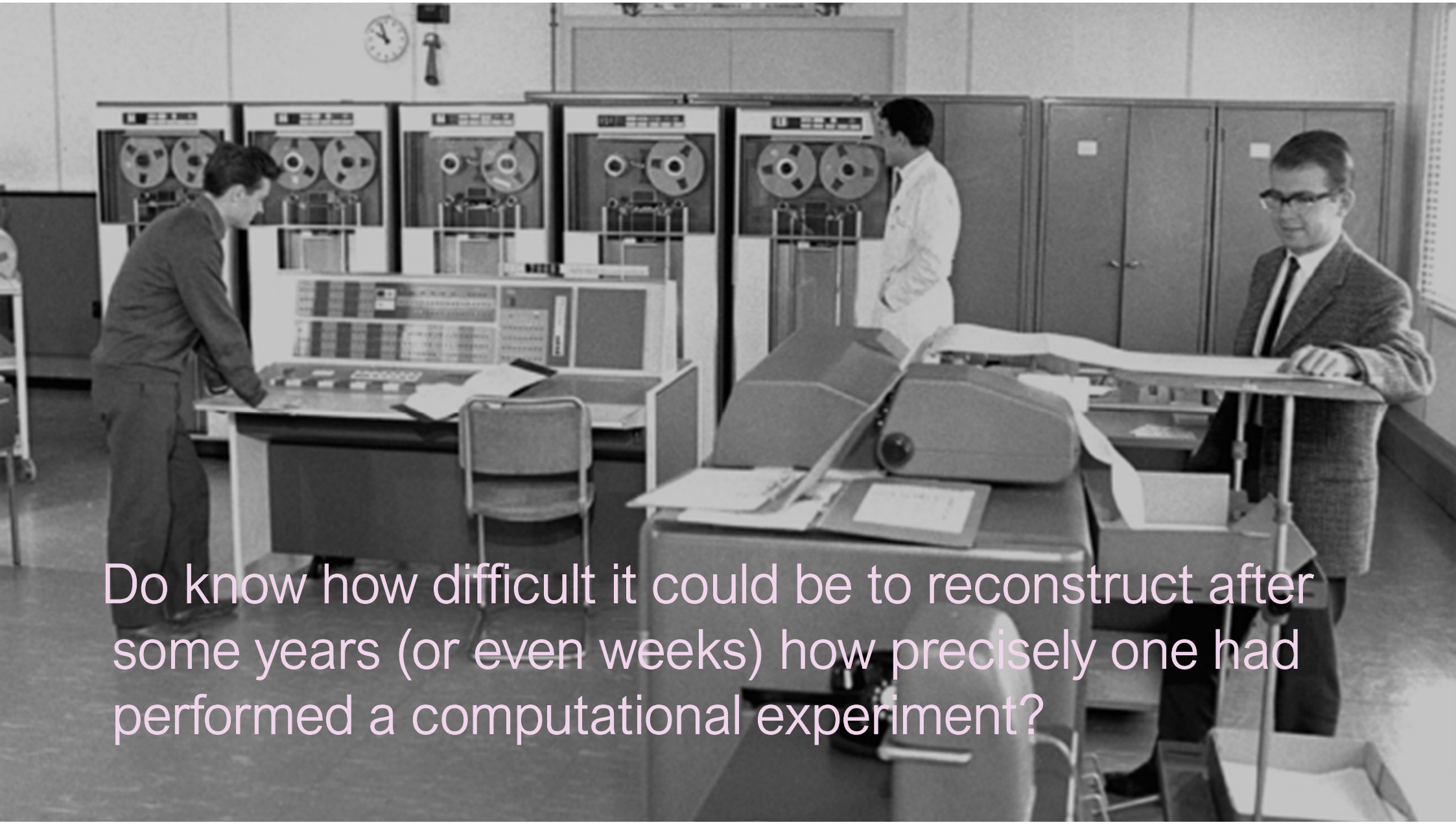
Expert Recommendation | Published: 21 April 2022

The case for data science in experimental chemistry: examples and recommendations

Junko Yano, Kelly J. Gaffney, John Gregoire, Linda Hung, Abbas Ourmazd, Joshua Schrier, James A. Sethian & Francesca M. Toma

Nature Reviews Chemistry **6**, 357–370 (2022) | Cite this article

4642 Accesses | **39** Citations | **31** Altmetric | Metrics


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...

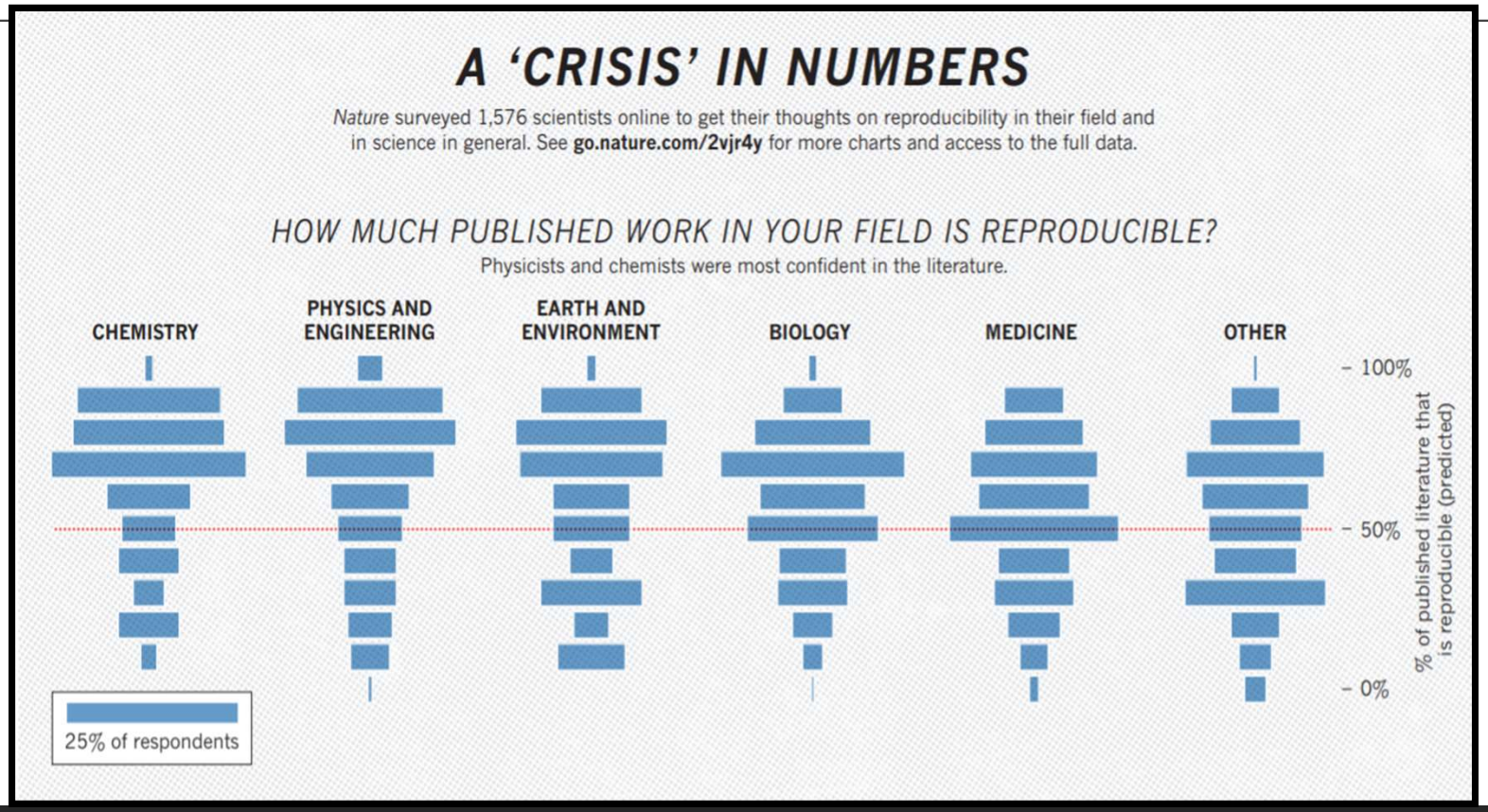
When scientists began to use computers to perform simulation experiments and data analysis, attention to experimental error took backstage.

- 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 :

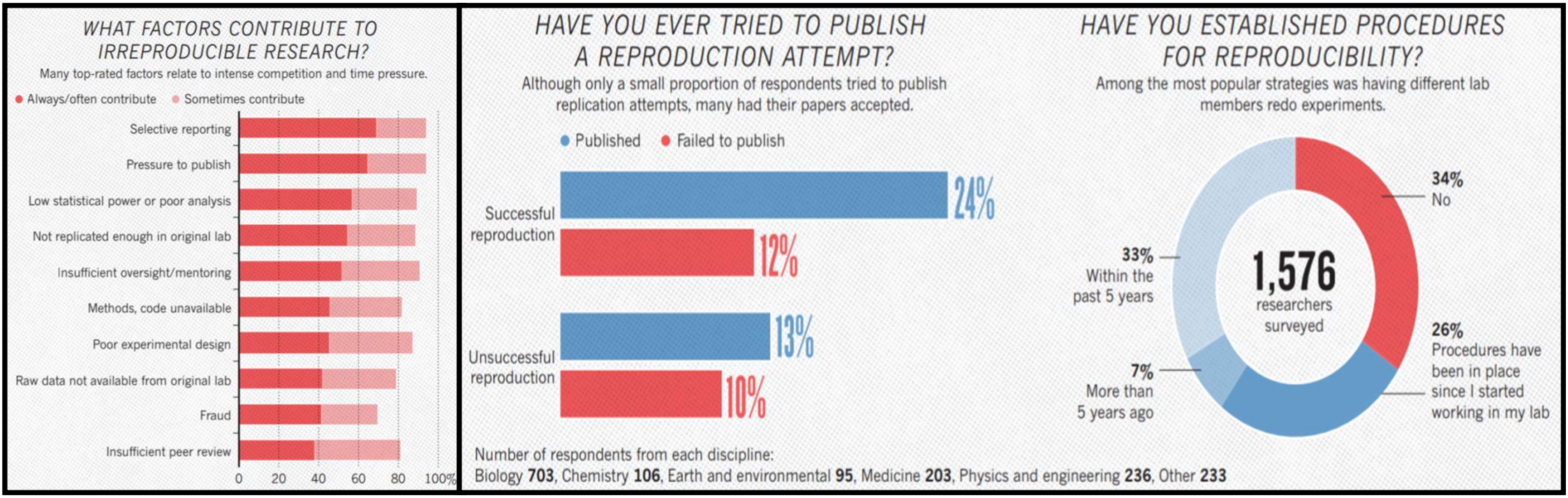
- 1) correctness of implementation,
- 2) potential for error, or
- 3) variation introduced by system soft and hardware.

Have you failed?



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

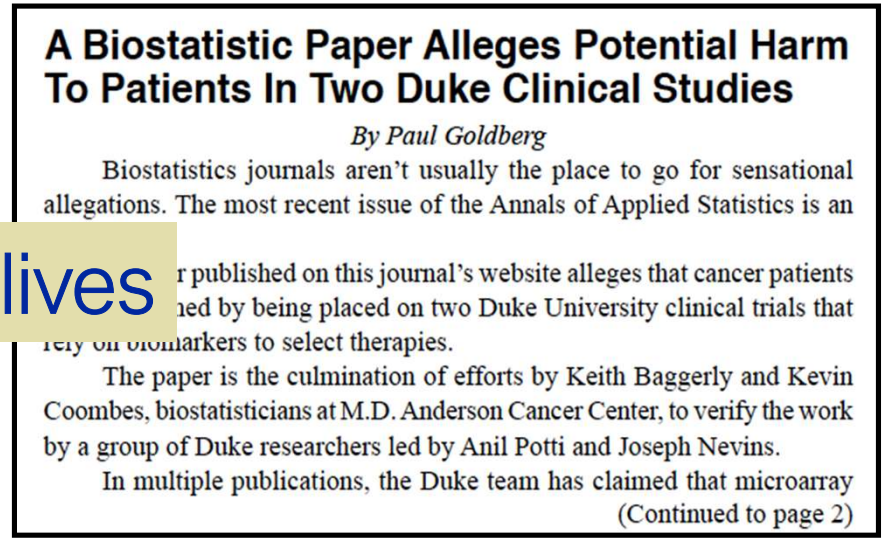
Why you failed?



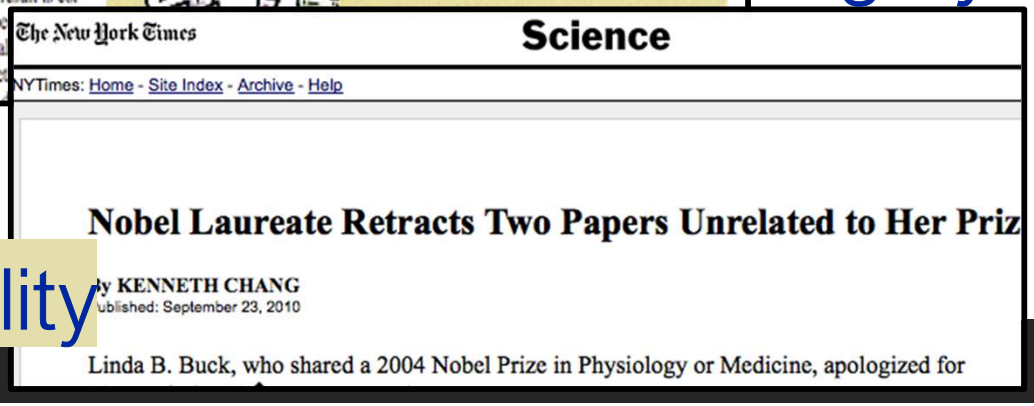
* Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*,*533*(7604), 452-454.

The value of Reproducibility…

Human lives



Scientific integrity

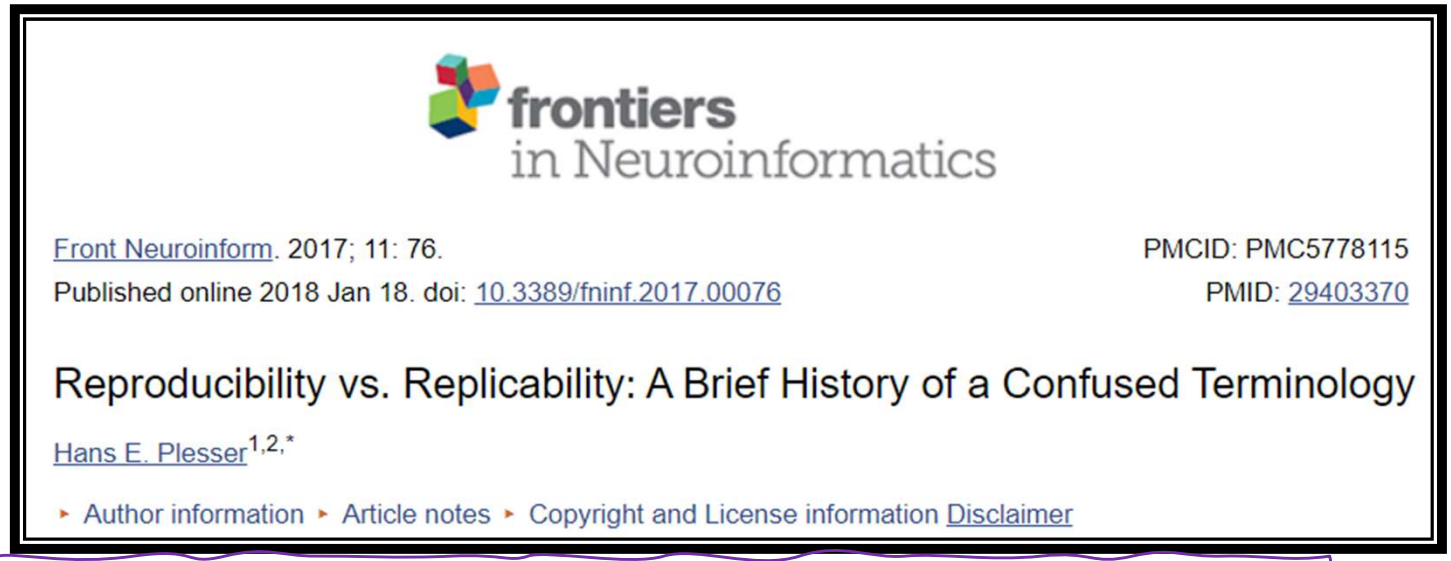


Reliability

Trust

Financial

Reproducibility x replicability



frontiers in Neuroinformatics

Front Neuroinform. 2017; 11: 76.
Published online 2018 Jan 18. doi: 10.3389/fninf.2017.00076
PMCID: PMC5778115
PMID: 29403370

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

Hans E. Plesser[1,2,*]

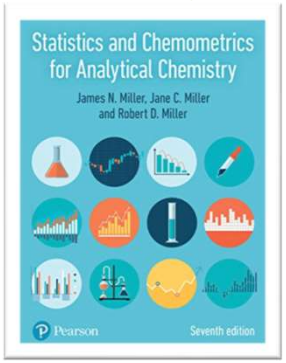
- **Reproducibility** – can you recreate the same result using original data and code?
- **Replicability** – can you recreate the same result using new data but same experimental design?

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

Association for Computing Machinery (2016). Artifact Review and Badging. Available online at: https://www.acm.org/publications/policies/artifact-review-badging

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.

Association for Computing Machinery (2016). Artifact Review and Badging. Available online at: https://www.acm.org/publications/policies/artifact-review-badging

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

Association for Computing Machinery (2016). Artifact Review and Badging. Available online at: https://www.acm.org/publications/policies/artifact-review-badging

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		

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

Depending on who you ask, these definitions are reversed!

Association for Computing Machinery (2016). Artifact Review and Badging. Available online at: https://www.acm.org/publications/policies/artifact-review-badging

Goodman saves!

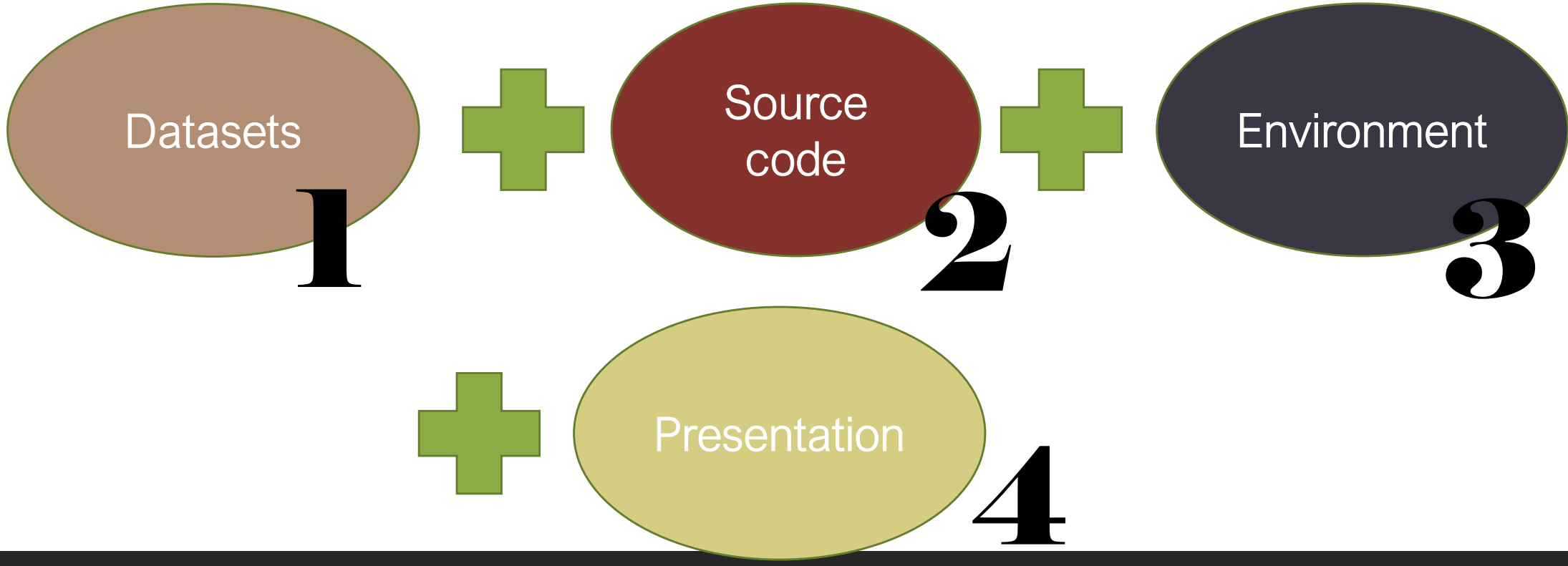
Terminology confusion! Goodman et al. (2016) propose a new lexicon for research reproducibility. Definitions:

- 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.

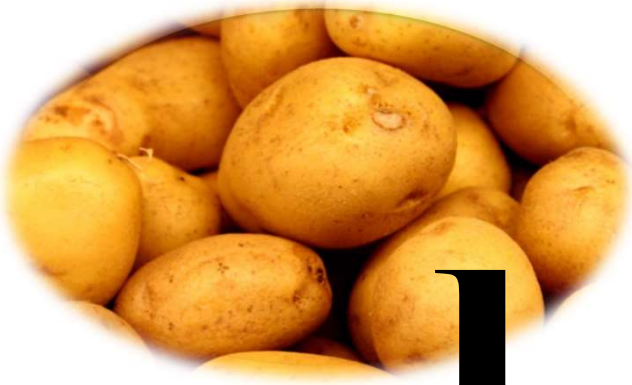
These definitions make explicit which aspects of trustworthiness of a study, avoid the ambiguity caused by the fact that “reproducible”, “replicable,” and “repeatable” have very similar meaning in everyday language.

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

Reproducibility



Reproducibility (kitchen analogy!)



1

+

2

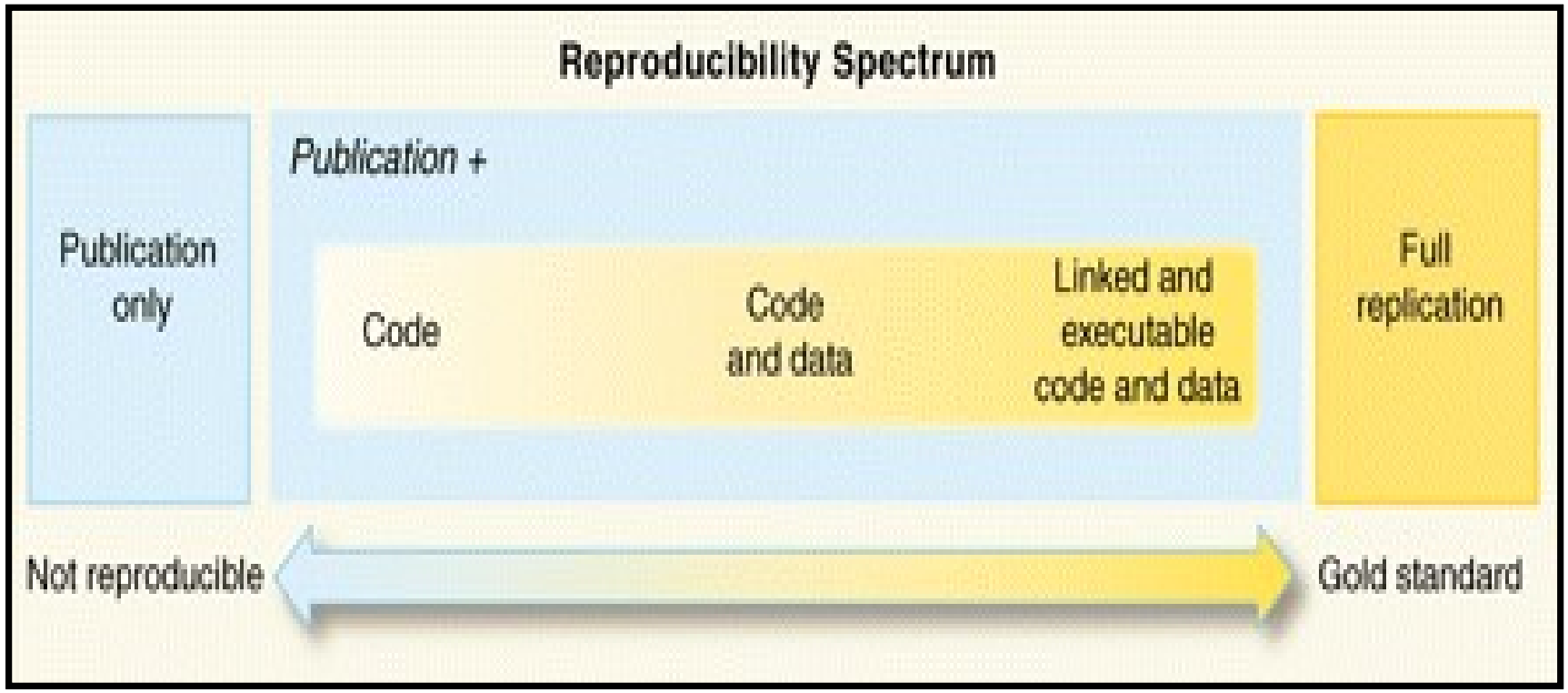
+

3

=

4

Reproducibility Spectrum

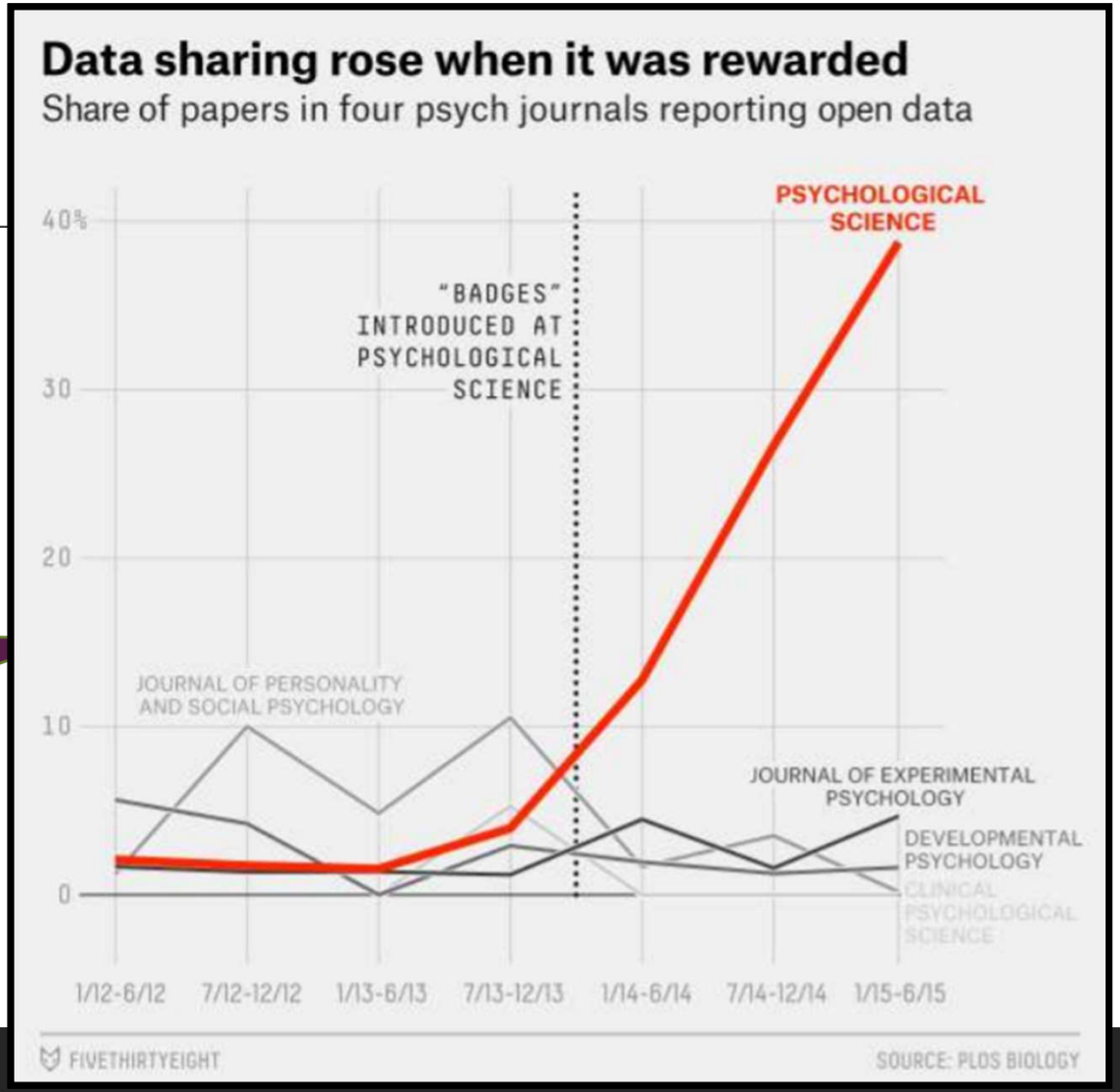


Peng, R. D. (2011). Reproducible research in computational science. *Science (New York, Ny)*, *334*(6060), 1226.

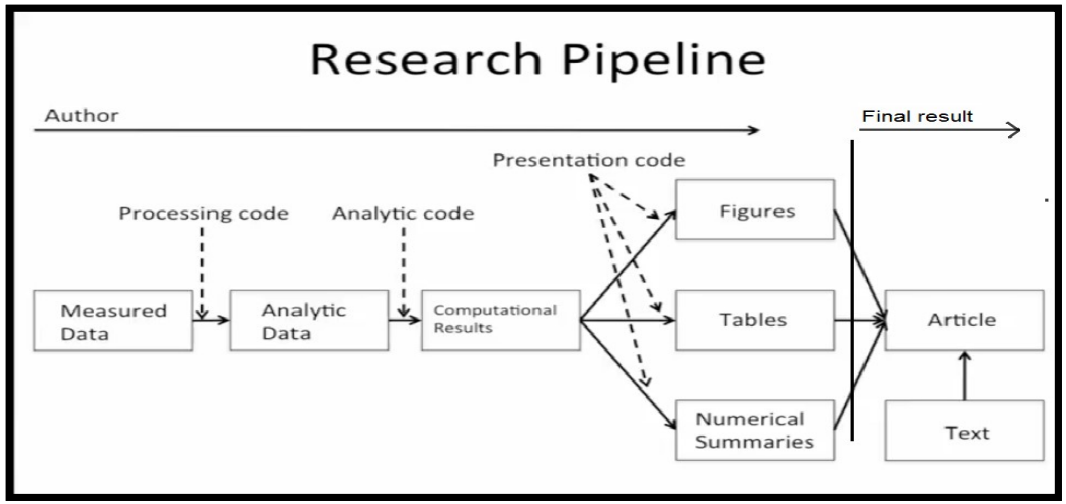
1- Dataset

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

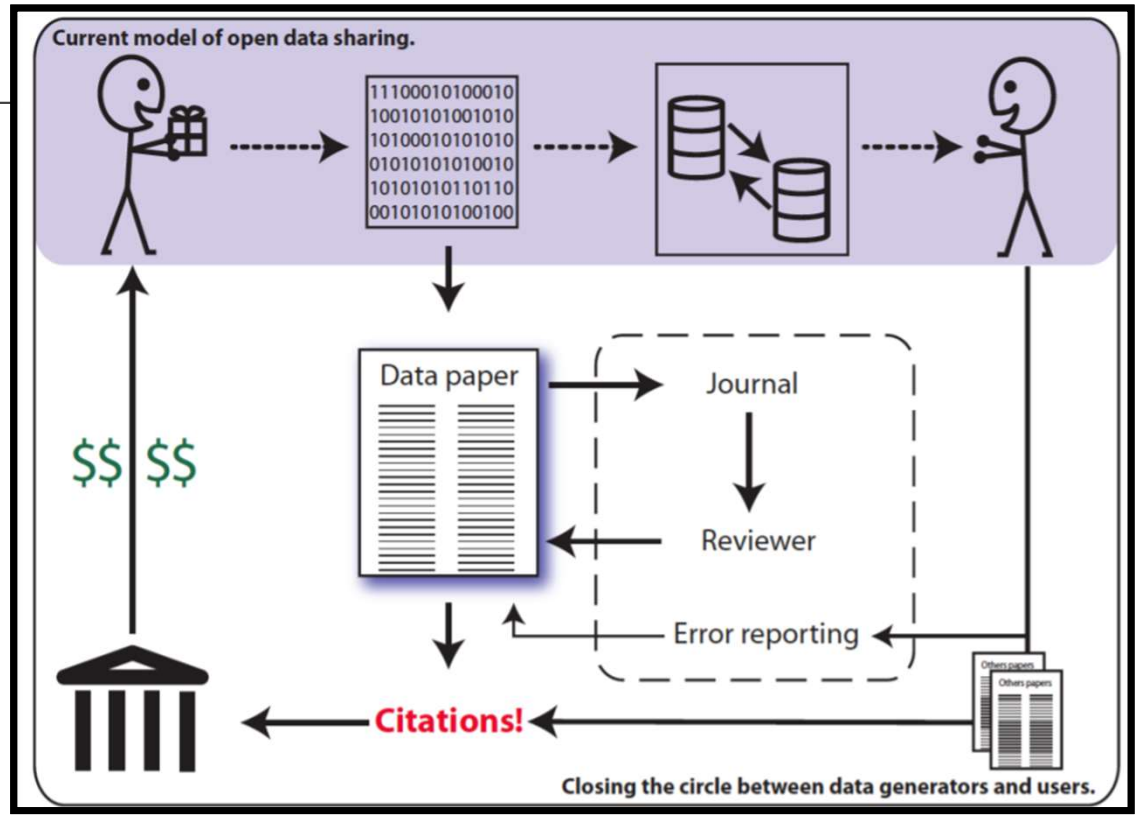
As discussed in Scienctific Data Managenet course (2020.2)



1- Dataset



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!

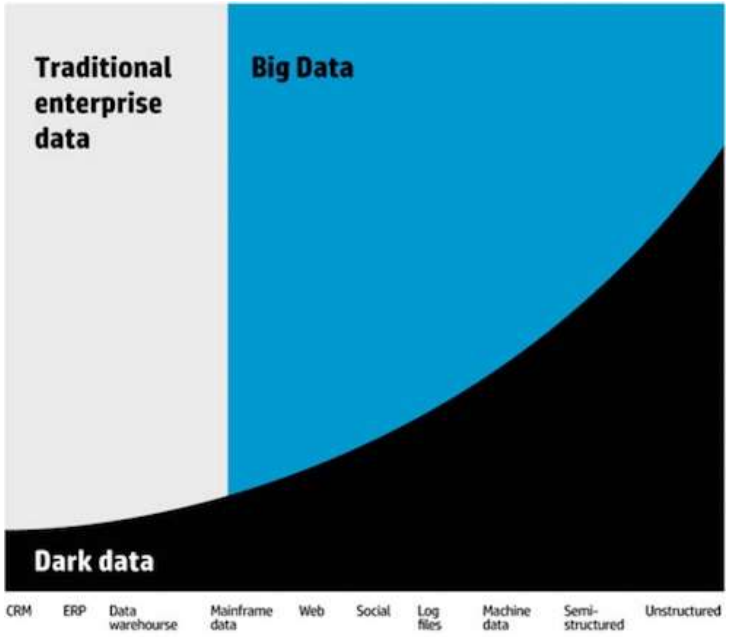
1- Data : Support Data (a.k.a Metadata)

- 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

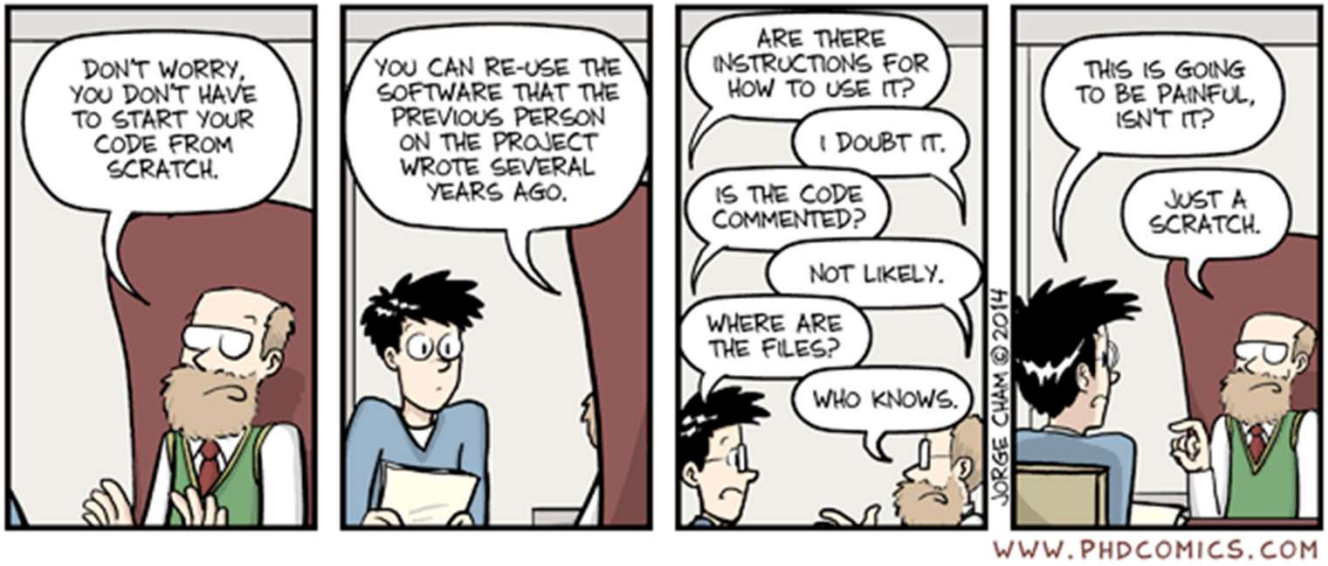
1- Dataset: Darkness



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

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

2 – Code: Avoid darkness…



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

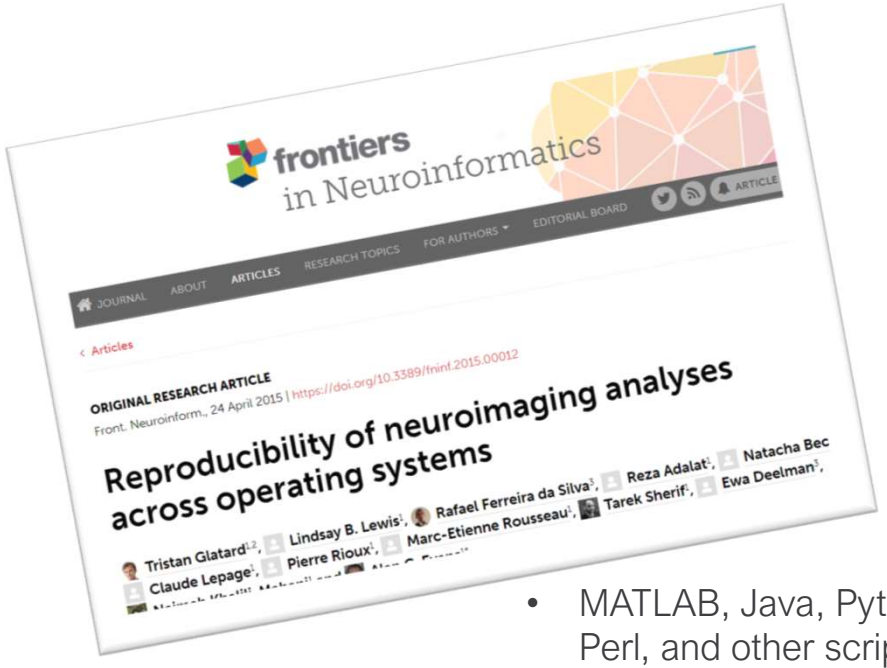
"Dark Software" is the counterpart of "Dark Data" (Heidorn 2008)

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

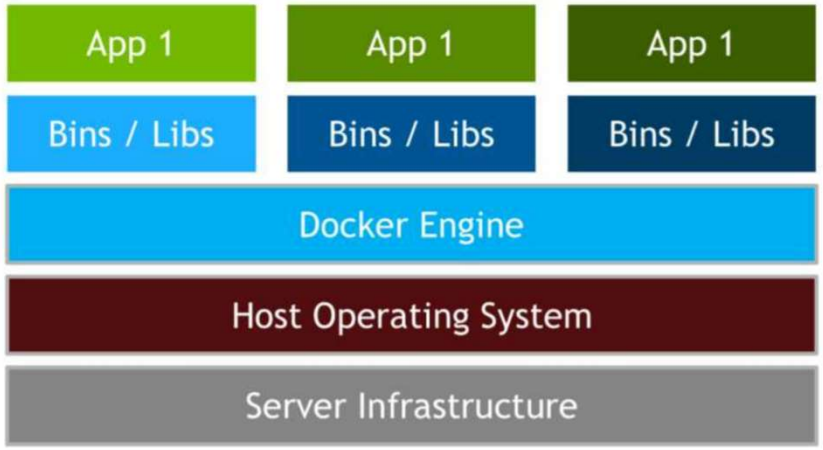
This paper reports experiments with three of the neuroimaging tools (FMRIB Software Library Freesurfer and CIVET).

- 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**.

https://www.frontiersin.org/articles/10.3389/fninf.2015.00012/full

2- Code: Capturing dependencies

VM



CONTAINERS

ReproZip

2- Code:Version control

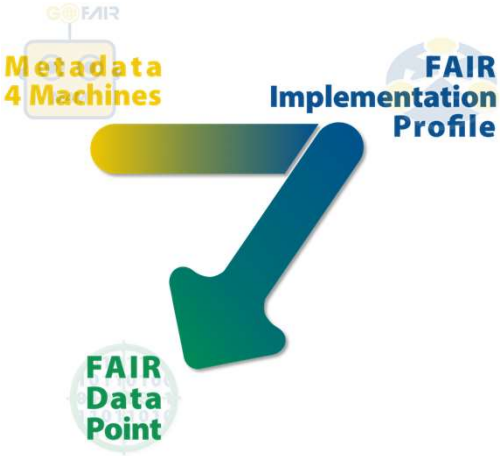
Git and GitHub ann Bitbucket are useful for everyday work

- 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



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!

https://neptune.ai/blog/best-7-data-version-control-tools-that-improve-your-workflow-with-machine-learning-projects

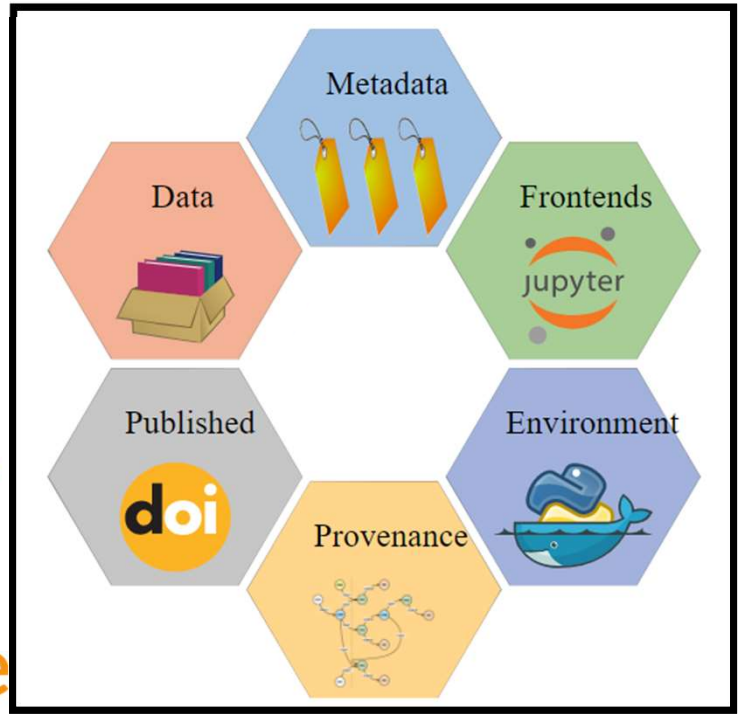
2- Code: Package control (e.g. Machine Learning/Data Science)

- anaconda: python distribution and environment management suite
- miniconda: just environment management suíte

  - 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...

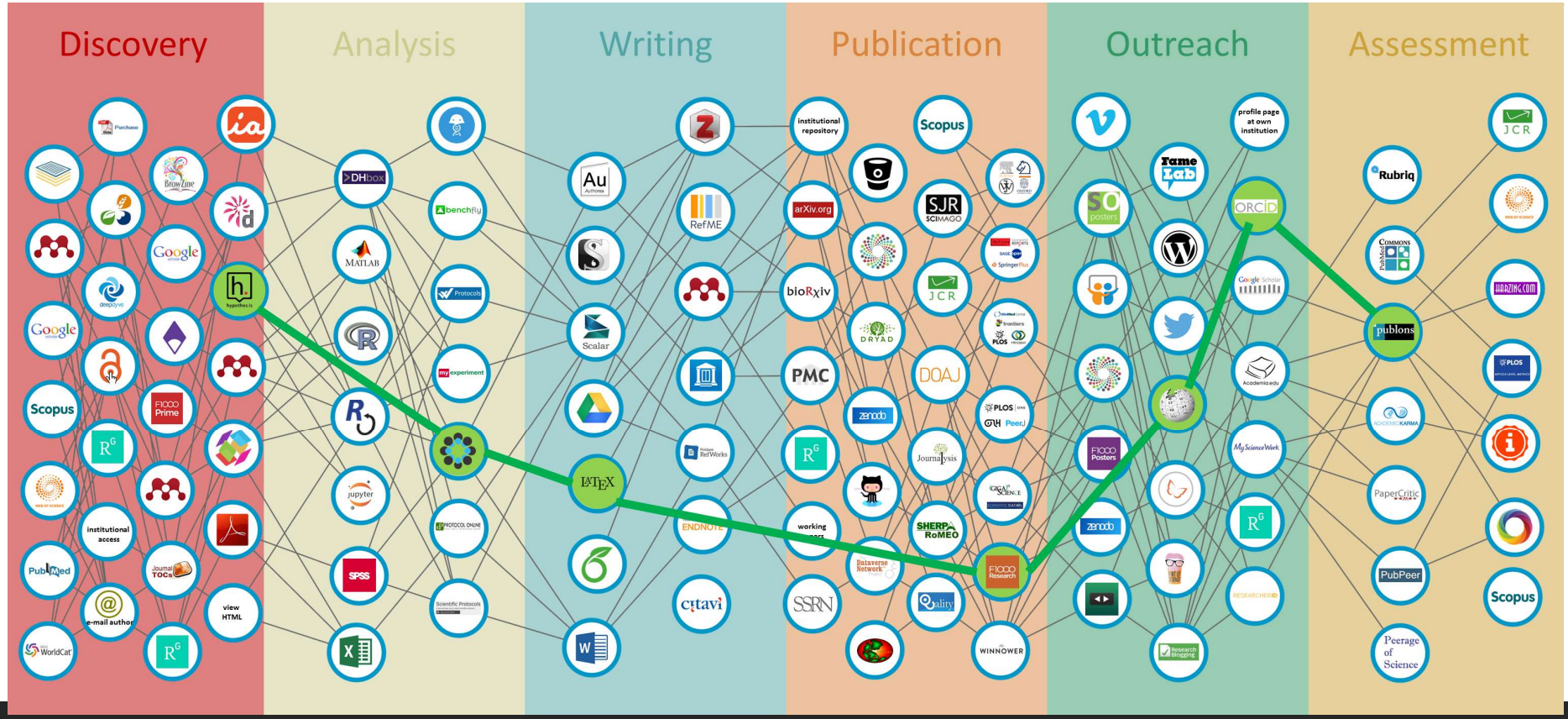
3- Environments

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?



https://101innovations.wordpress.com/workflows/

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 Donnellan[1]
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).

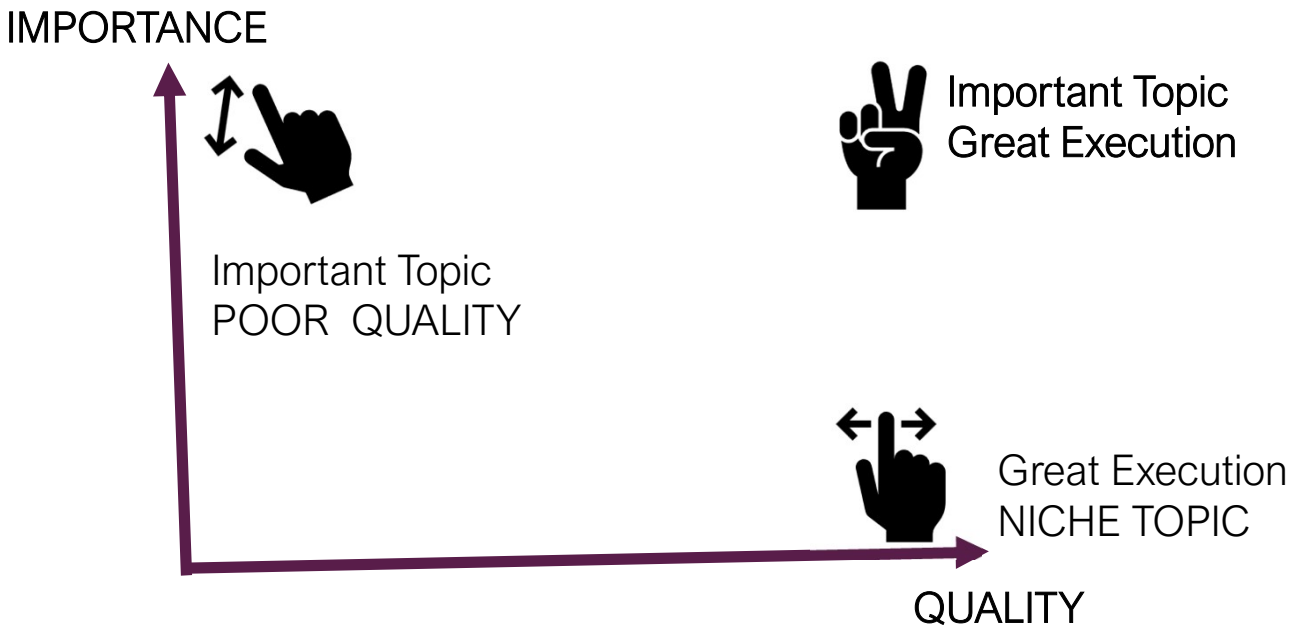
- Is the original finding used in **policy making**?
- Did the original finding **open a new subfield of research**?
- Is there a **debate about the original finding**?
  - Are there studies undermining the original finding?
  - Are there studies confirming the original finding?

What makes a good replication?

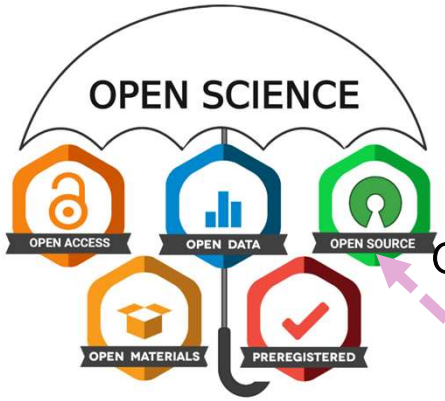
Dimension 2 (Quality): Quality of the replication attempt (1-5).

- 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?



What makes a good replication?



OPEN ?

IMPORTANCE

Important Topic
POOR QUALITY

Important Topic
Great Execution

Great Execution
NICHE TOPIC

QUALITY

???

FAIR ?

FAIRSFAIR
Fostering Fair Data Practices in Europe

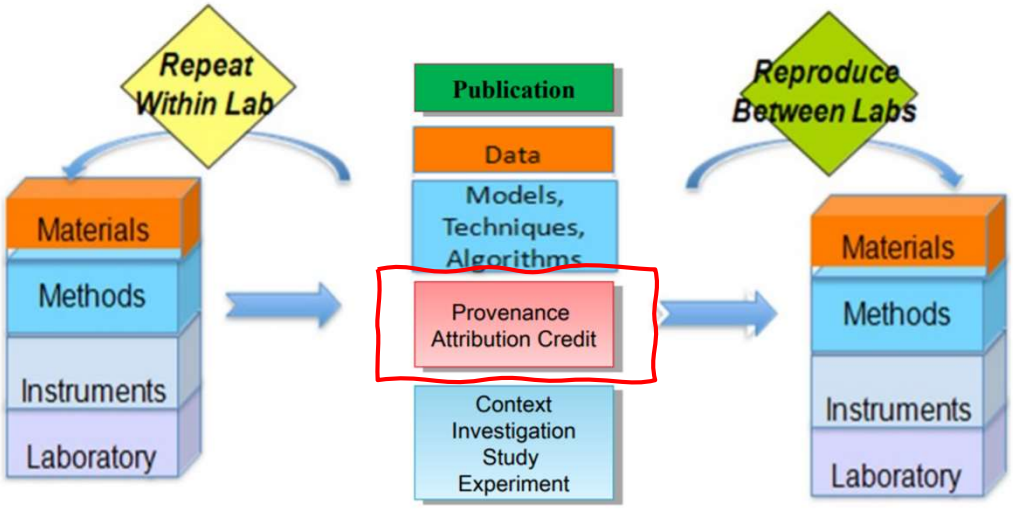
FAIRsharing.org
standards, databases, policies

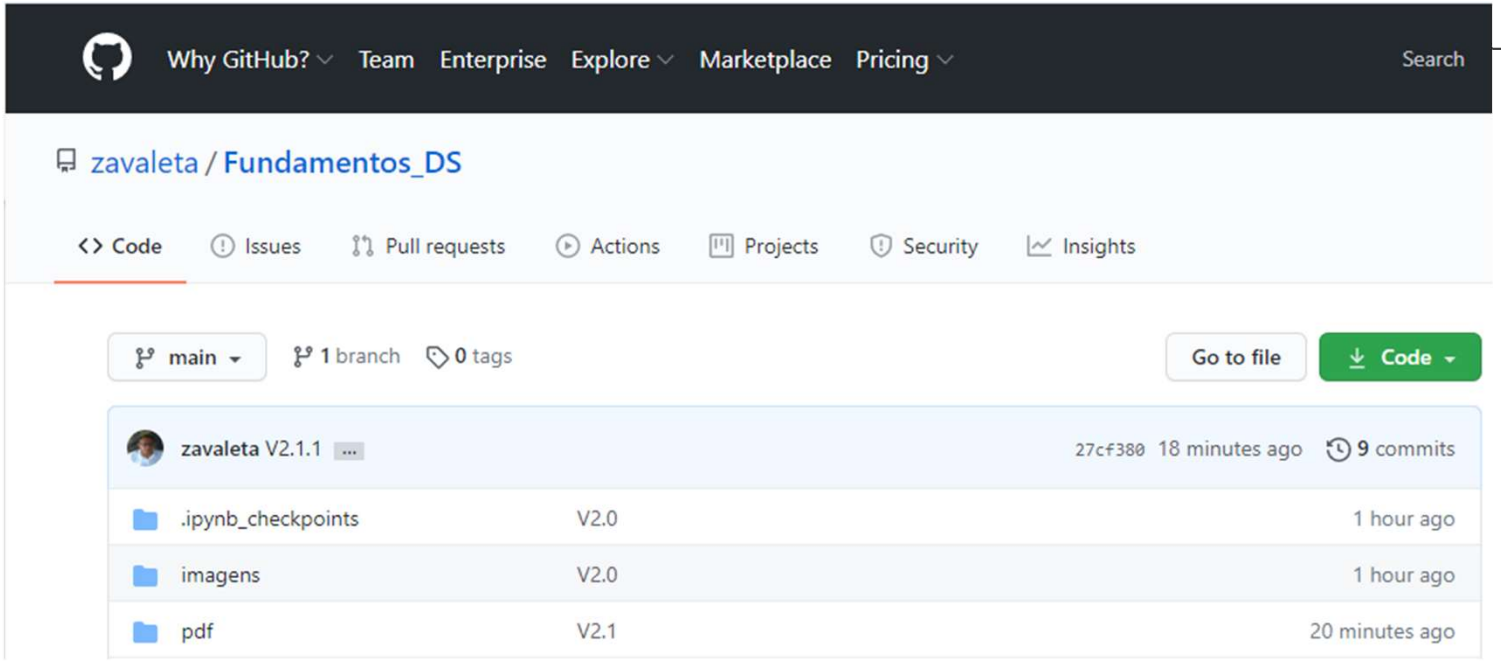
Summary

- 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





Fundamentos de Ciência de Dados

Professores:

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

Yano, J. et al (2022) The Case for Data Science in Experimental Chemistry: Examples and Recommendations. Nat Rev Chem 6, 357–370 (2022). https://doi.org/10.1038/s41570-022-00382-w