

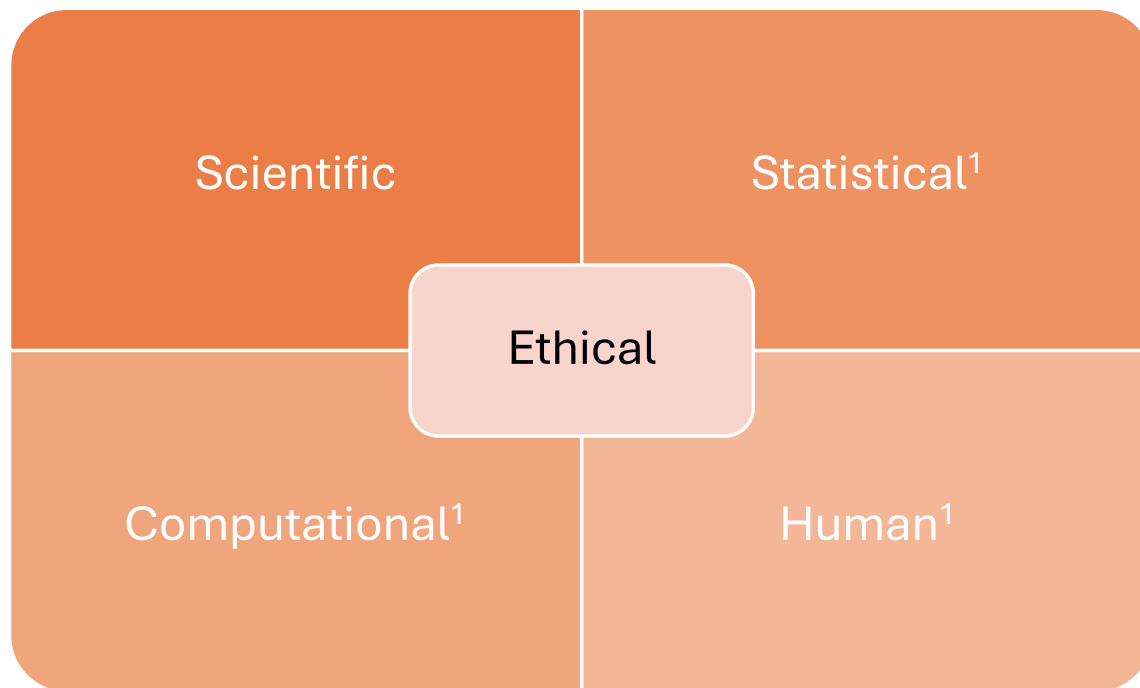
# **Good Data Science Practice: Moving Toward a Code of Practice for Drug Development**

Mark Baillie,  
BBS, Basel, 12<sup>th</sup> April 2024



# A view of data science for drug development



A set of integrated thinking skills and practices refocused for answering questions with data



STATISTICS IN BIOPHARMACEUTICAL RESEARCH  
2022, AHEAD-OF-PRINT, 1-12  
<https://doi.org/10.1080/19466315.2022.2063172>



## Good Data Science Practice: Moving Toward a Code of Practice for Drug Development

Mark Baillie <sup>a</sup>, Conor Moloney<sup>b</sup>, Carsten Philipp Mueller<sup>c</sup>, Jonas Dom <sup>d</sup>, Janice Branson<sup>a</sup>, and David Ohlssen<sup>e</sup>

<sup>a</sup> Clinical Development & Analytics, Novartis Pharma AG, Basel, Switzerland;; <sup>b</sup> Clinical Development & Analytics, Novartis Pharma AG, Dublin, Ireland;; <sup>c</sup> Data & Digital, Beyond Conception GmbH, Altendorf, Switzerland;; <sup>d</sup> Roche Pharma Research and Early Development, pRED Informatics, Roche Innovation Center, Basel, Switzerland;; <sup>e</sup> Clinical Development & Analytics, Novartis Pharma AG, East Hannover, NJ

## The Role of Statistical Thinking in Biopharmaceutical Research

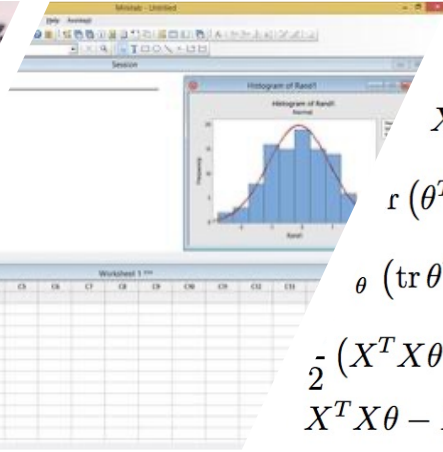
Frank Bretz  & Joel B. Greenhouse 

Pages 458-467 | Received 19 Dec 2021, Accepted 19 May 2023, Published online: 24 Jul 2023

<sup>1</sup> Blei & Smyth (2017) *Science and data science*. *PNAS* 114 (33):8689-8692

# A personal detour

- rote statistical education
- collaborative research
- cargo cult statistics
- statistical thinking



$$\bar{y}^T(X\theta - \bar{y})$$

$$X^T X \theta - \theta^T X^T \bar{y} - \bar{y}^T X \theta + \bar{y}^T \bar{y}$$

$$\theta^T X^T X \theta - \theta^T X^T \bar{y} - \bar{y}^T X \theta +$$

$$\theta (\text{tr } \theta^T X^T X \theta - 2\text{tr } \bar{y}^T X \theta)$$

$$\frac{1}{2} (X^T X \theta + X^T X \theta - 2X^T \bar{y})$$

$$X^T X \theta - X^T \bar{y}$$



“I have never let my schooling interfere with my education.”

**Mark Twain**

# Statistical rituals



## Paper Rejected ( $p > 0.05$ ): An Introduction to the Debate on Appropriateness of Null-Hypothesis Testing

Mark. D. Dunlop, Mark Baillie

Source Title: [International Journal of Mobile Human Computer Interaction \(IJMHCI\)](#) 1(3)

Copyright: © 2009 | Pages: 8

DOI: 10.4018/jmhci.2009070108

Journal of Consulting and Clinical Psychology  
1978, Vol. 46, 806-834.

#113

---

## The Earth Is Round ( $p < .05$ )

---

Jacob Cohen

Theoretical Risks and Tabular Asterisks:  
Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology

Paul E. Meehl  
University of Minnesota

# Statistics and research



Article |  Token Access

## Deconstructing Statistical Questions

David J. Hand 

First published: 1994 | <https://doi.org/10.2307/2983526> | Citations: 28

Donald A. Preece (University of Kent, Canterbury):

Professor Hand speaks of the questions that the researcher wishes to consider. These are often three in number:

1. how do I obtain a statistically significant result?;
2. how do I get my paper published?;
3. when will I be promoted?

Many researchers ask merely 'How do I get results?', meaning by 'results', not answers to questions, but things that are publishable in glossy reports.

This tends to confirm the statistician as a mere outside consultant whom people perhaps cannot afford until they are in a mess, by which time a statistician is needed to paint respectability over defective work.

No, as Box (1993) stated, '**the statistician must strive to earn the title of first class scientist**'.

## Cargo-cult statistics and scientific crisis

Written by Philip B. Stark and Andrea Saltelli on 05 July 2018. Posted in [Science](#)

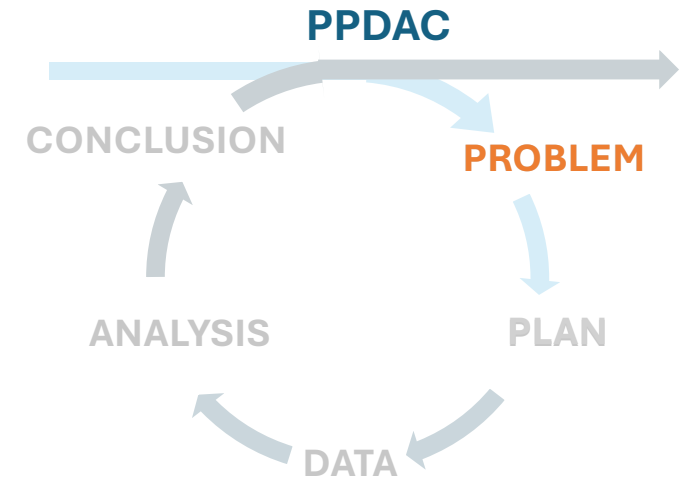


Poor practice is catching up with science,<sup>1-3</sup> manifesting in part in the failure of results to be reproducible and replicable.<sup>4-7</sup> Various causes have been posited,<sup>1, 8</sup> but we believe that poor statistical education and practice are symptoms of and contributors to problems in science as a whole.

The problem is one of cargo-cult statistics – the ritualistic miming of statistics rather than conscientious practice. This has become the norm in many disciplines, reinforced and abetted by statistical education, statistical software, and editorial policies.

At the risk of oversimplifying a complex historical process, we think the strongest force pushing science (and statistics) in the wrong direction is existential: science has become a career,

rather than a calling, while quality control mechanisms have not kept pace.<sup>9</sup>



# Good Data Science Practice

- big changes
- learn and confirm & research phases
- experimental vs found data
- the paradox of exploratory investigations
- what to do about data science?



# Background – Industry

---

- Contrary to **confirmatory** research, **exploratory** research in the Pharmaceutical industry is not subject to internationally accepted principles and practices such as ICH E8 or E9\*
- The inherent need of exploratory research for more flexibility in terms of data analysis must leave a **high degree of freedom for scientists** to organize their research
- At the same time this degree of freedom risks exploratory projects to be executed in an **unmethodical way**, which risks project results **not** to be reproducible or simply not valid
- In other words, the researcher's degree of freedom may negatively impact the failure rate of exploratory projects
- Currently, there is no consensus in the Pharmaceutical industry on how to set boundaries to this freedom; specifically, how to employ appropriate design and analysis methods in exploratory research projects
- Amplified by the promise of (big-)data, ML & AI, increased computation, and resulting society expectations, there is a need to set more rigorous principles and practices

We advocate to implement Good Data Science Practices (GDSP) for exploratory research projects in order to address these challenges

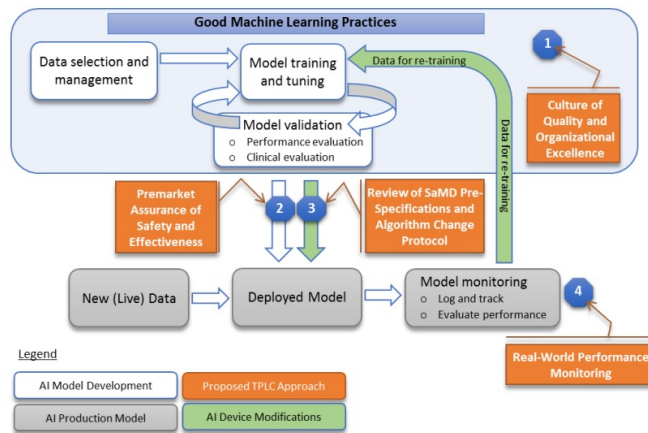


Figure 2: Overlay of FDA's TPLC approach on AI/ML workflow

Clinical Evaluation		
Valid Clinical Association	Analytical Validation	Clinical Validation
Is there a valid clinical association between your SaMD output and your SaMD's targeted clinical condition?	Does your SaMD correctly process input data to generate accurate, reliable, and precise output data?	Does use of your SaMD's accurate, reliable, and precise output data achieve your intended purpose in your target population in the context of clinical care?

Figure 3: IMDRF description of Clinical Evaluation components

AI/ML algorithm development involves learning from data and hence prompts unique considerations that embody GMLP. In this paper, GMLP are those AI/ML best practices (e.g., data management, feature extraction, training, and evaluation) that are akin to good software engineering practices or quality system practices. Examples of GMLP considerations as applied for SaMD include:

As envisioned in the Software Pre-Cert Program,<sup>16</sup> applying a TPLC approach to the regulation of software products is particularly important for AI/ML-based SaMD due to its ability to adapt and improve from real-world use. In the Pre-Cert TPLC approach, FDA will assess the culture of quality and organizational excellence of a particular company and have reasonable assurance of the high quality of their software development, testing, and performance monitoring of their products. This approach

<sup>16</sup> Developing a Software Precertification Program: A Working Model; v1.0 – January 2019: <https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/UCM629276.pdf>.

# 13 Principles of ICH Good Clinical Practice

1. Ethical Principles of Declaration of Helsinki	2. Benefit justifies risk	3. Rights, safety, wellbeing	4. Adequate information to support trial	5. Clear, scientifically sound protocol
6. IRB/EC approval prior to initiation	7. Medical care / decisions by qualified physician	<b>8. Researcher training, education and experience</b>	9. Freely given informed consent	10. Accurate data handling and storage
	11. Data Protection and confidentiality	12. Good Manufacturing Practice	13. Quality assurance systems	

# Problem Statement (1/2)

---

Exploratory research projects are not governed by a clear set of guiding principles nor a project methodology that provides standardized means for scientists to structure their project

The absence of such professional practices lead to increased risks, project failures and inefficiencies

Inconsistent project design lead to a **varying degree of project quality**

Poorly scoped and designed projects may **not** properly

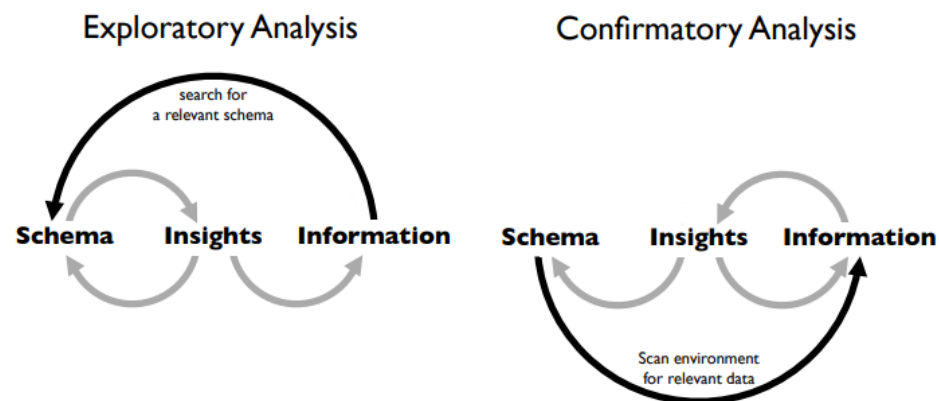
- define the research **purpose**
- formulate the **question(s)** of interest
- consider the relevant **context**
- specify the **intended use** of project outcomes

Therefore, inappropriate analytical strategies may be chosen increasing the risk of project failure

# Science has exploratory and confirmatory objectives

## Broadening the perspective

- Learn and confirm
- Analogous to clinical trial phases
- Learning covers most scientific work
- Many lines of inquiry and enquiry
- Pre-specified primary analysis, primary research, secondary research, meta-research
- Design of experiments vs data set(s) available for retrospective investigation



Grolemund, Garrett, and Hadley Wickham. "A Cognitive Interpretation of Data Analysis." *International Statistical Review / Revue Internationale De Statistique*, vol. 82, no. 2, 2014, pp. 184–204. *JSTOR*, [www.jstor.org/stable/43299753](http://www.jstor.org/stable/43299753). Accessed 26 May 2021.

# Problem Statement (2/2)

---

Absence of a framework increases the risk of

- unstructured and undocumented project materials and outcomes affecting reproducibility and reusability, causing a loss or duplication of work
- generated project knowledge not translated into scientific or business value
- existing resources (e.g. guidance and cheat sheets) and initiatives not embedded into practice
- inefficient collaboration across teams due to no common ways of working

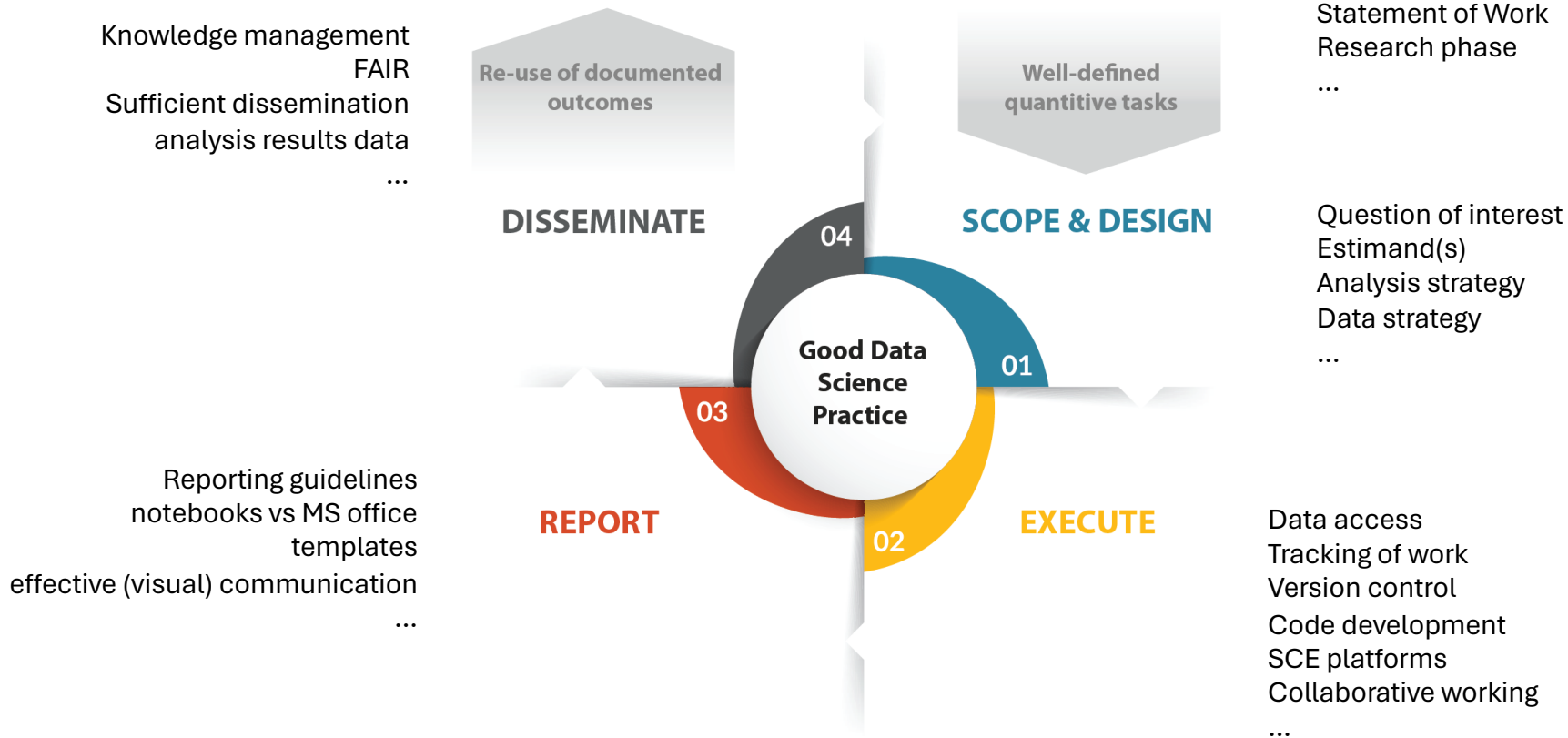
Exploratory research projects are not consistently evaluated in terms of **plausibility, reproducibility and strength of evidence**

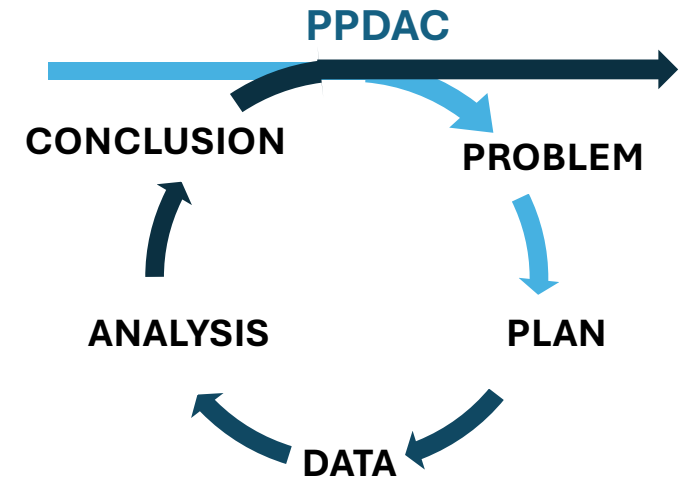
The identified issues above result in ineffective and **inefficient allocation of time, budget and resources!**

It is critical that we define a framework that promotes **good and robust scientific research practice adopted by data scientists**

It is also critical that we promote a **culture of self-discipline** to ensure a **balance between scientific flexibility and rigor**

# GDSP framework: big ideas





# Data Science and Statistics

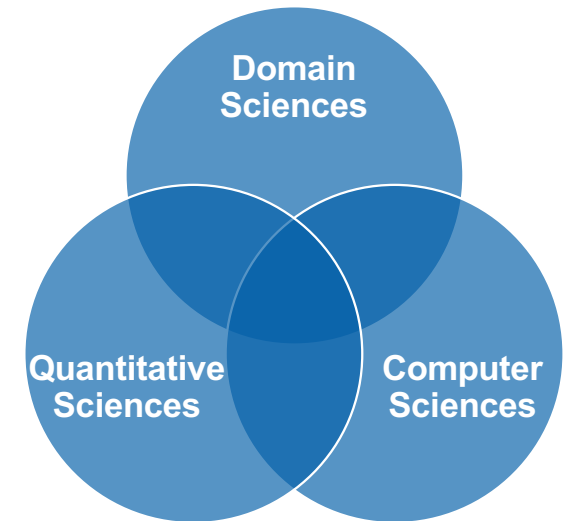
- What is data science?
- Integrated thinking skills and practices



# What is Data Science?

“ *Data science is the study of extracting value from data.* ” – Jeannette Wing<sup>1</sup>

- Data science is an interdisciplinary field to facilitate learning from data
- Impactful data science projects are **cross-functional** efforts
- The technical foundations of data science draw on quantitative and computer sciences, used in conjunction with profound domain expertise



<sup>1</sup>Source: <https://datascience.columbia.edu/news/2018/what-is-data-science>

# What is Data Science?

“ *Data science is the study of extracting value from data.* ” – Jeannette Wing<sup>1</sup>

- Data science is an interdisciplinary field to facilitate learning from data
- Impactful data science projects are **cross-functional** efforts
- The technical foundations of data science draw on quantitative and computer sciences, used in conjunction with profound domain expertise

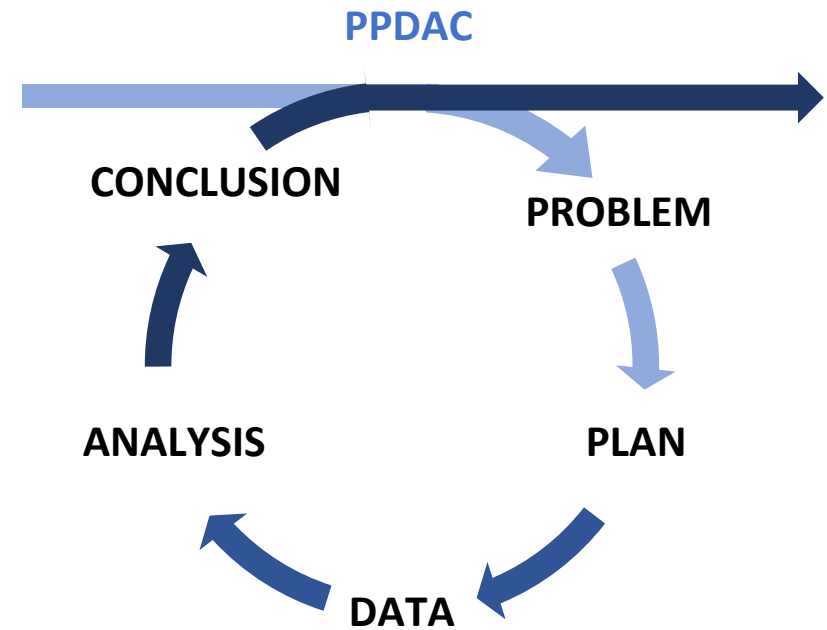


<sup>1</sup>Source: <https://datascience.columbia.edu/news/2018/what-is-data-science>

# The data science recipe

The recipe to ensure impact:

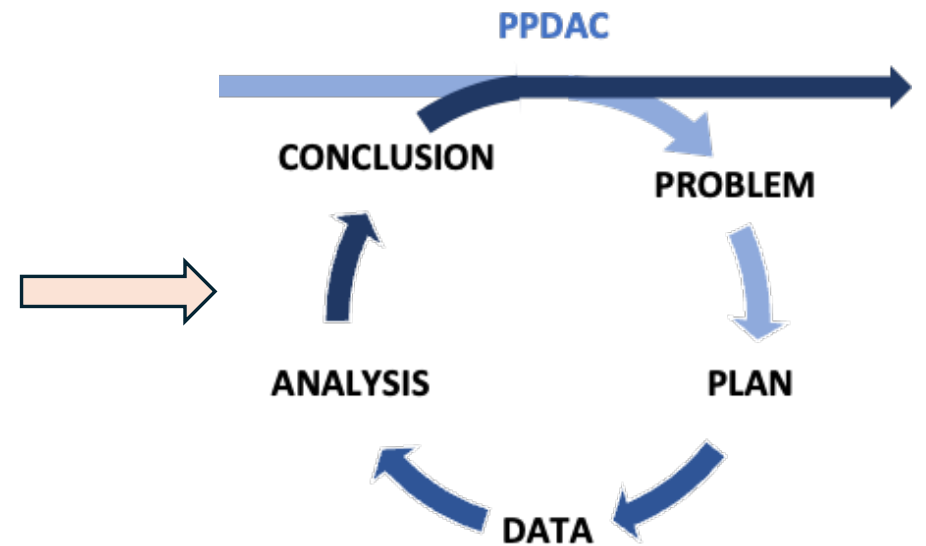
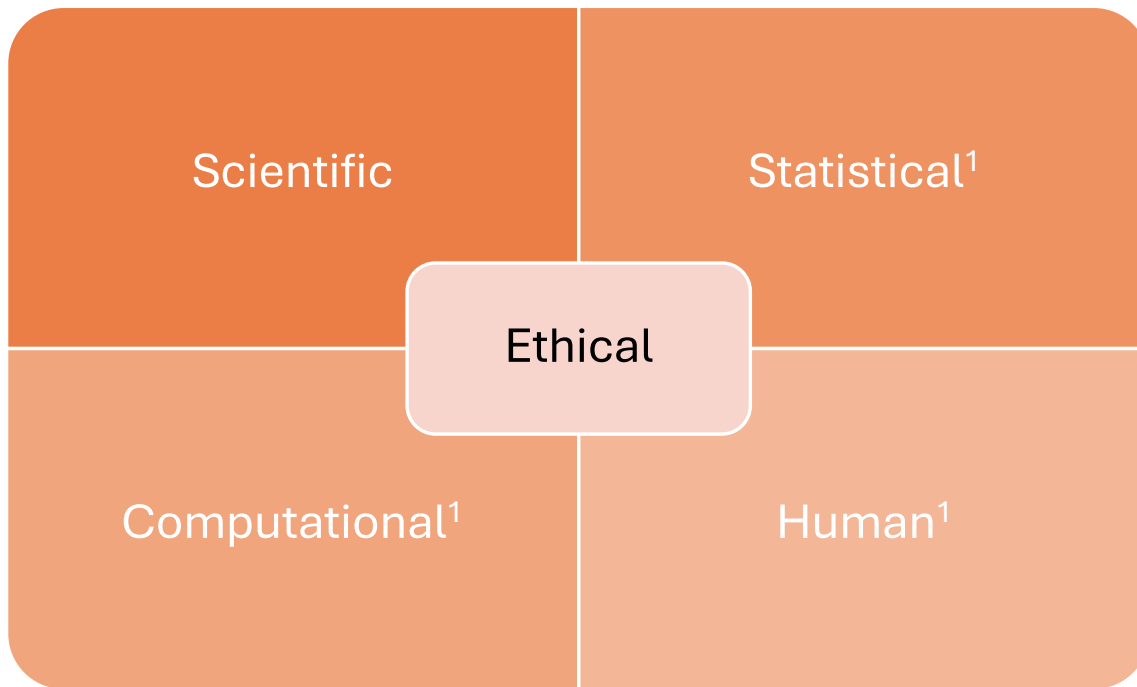
- A clearly motivated problem
- A well-defined question(s)
- Quality data and valid analysis strategy
- Smooth execution
- Appropriate evaluation, and
- Effective communication of outcomes



Source: MacKay, R.J. and Oldford, R.W., 2000. Scientific method, statistical method and the speed of light. *Statistical Science*, pp.254-278.

# Data science thinking

A set of integrated thinking skills and practices refocused for answering questions with data



<sup>1</sup> Blei & Smyth (2017) *Science and data science*. PNAS 114 (33):8689-8692

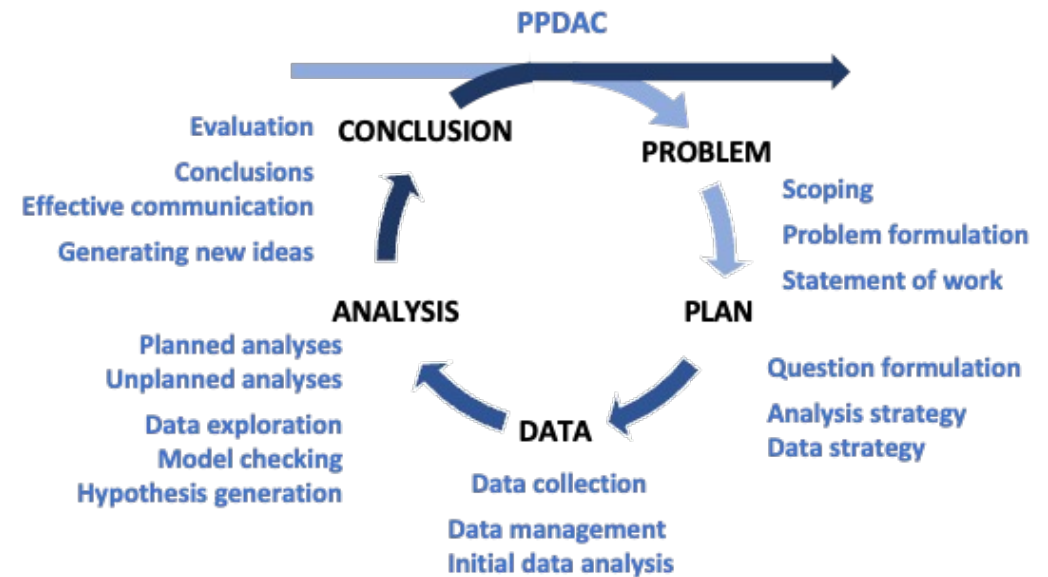
# Data science thinking embedded in a workflow

---

A set of integrated **thinking skills** and practices refocused for answering questions with data

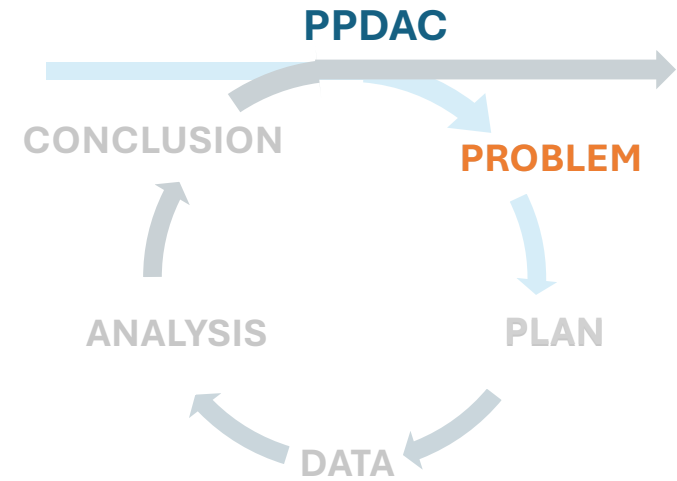
A good **workflow** is an established set of habits that help drive you forward towards your goal. They enable complexity to scale in the right areas.

PPDAC demonstrates the steps for abstracting and solving a **real problem**. An impactful solution requires a clear understanding of how things work.

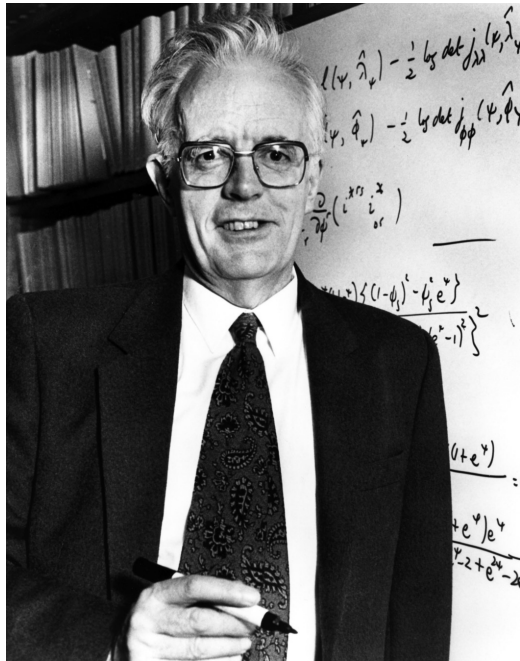


# Problem formulation

- elicit and understand the problem
- understanding the context
- Scoping
- translating into answerable questions



# Why? – avoiding Type III errors



*In all fields of work, even in pure mathematics, the formulation of issues or questions for investigation is central. Better a rough answer to an **important issue** than a beautiful study of a **topic of no real concern**. Statistical considerations enter in at least two ways. The first is to ensure that the questions are **reasonably defined** and **capable of being addressed**. Then, do we have or can we collect **data capable of giving a reasonable answer?***

**Sir David Cox (2017)**

# Why? – ensuring impact

npj | precision oncology

Comment

Published in partnership with The Hormel Institute, University of Minnesota



<https://doi.org/10.1038/s41698-024-00553-6>

## All models are wrong and yours are useless: making clinical prediction models impactful for patients

Florian Markowetz

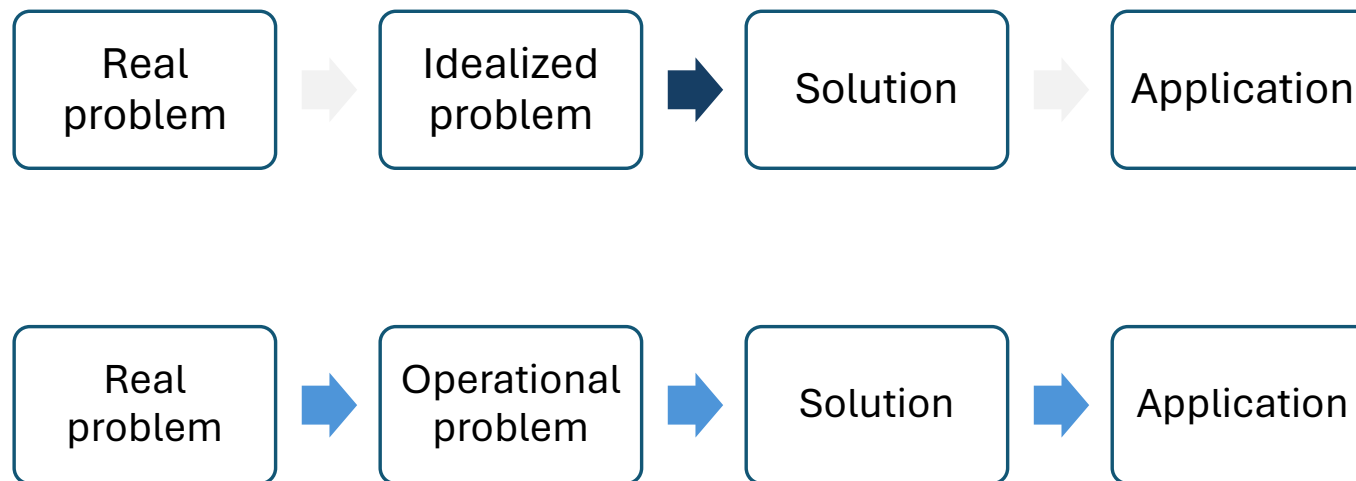
 Check for updates

*Most published clinical prediction models are never used in clinical practice and there is a huge gap between academic research and clinical implementation. Here, I propose ways for academic researchers to be proactive partners in improving clinical practice and to design models in ways that ultimately benefit patients.*

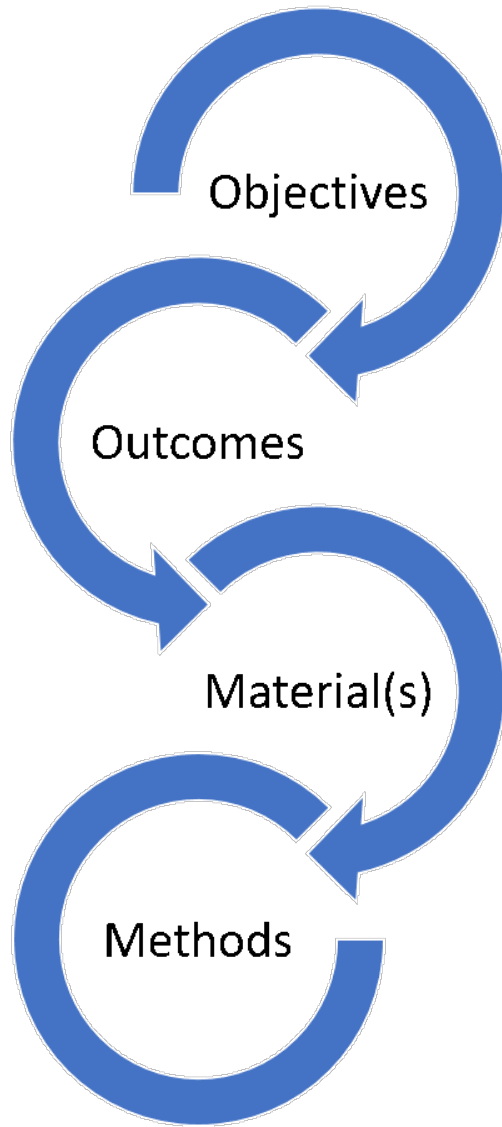


# Why? - impact and problem formulation

“we can't stress this enough – you simply must understand the real problem if you hope to help solve it.” (Ron Kenett and Thomas C. Redman)



# The iterative steps of scoping



1. Objective(s) – capture, define and refine the goal and objective(s) of the project;
2. Outcome(s) – capture what actions, decisions or interventions will the outcome of the project inform;
3. Materials – what data and other resources are required to achieve these goals;
4. Methods –
  - What analyses need to be performed?
  - What analysis type(s) and strategies are required (describe, detection, prediction, intervention, explanation? (Hernán et al. 2019)).

# SoW template

<https://github.com/datascience-thinking/SoW>

## GDSP statement of work (SOW)

**Document goal:** Ensure the right business and scientific questions are formulated, and the right analyses are designed to address these questions, and assess necessary resources identified to plan and execute plans.

**Output:** a brief written description of the questions to be addressed, the activities to address them and who was involved in this assessment. GDSP SOW should be stored on a **knowledge management** system, with the location of the document captured in a **tracker**.

**Revision tracking:** The GDSP SOW should be maintained to capture major changes in project scope during execution and completion. A change log is available to capture this information.

### PROJECT INFORMATION

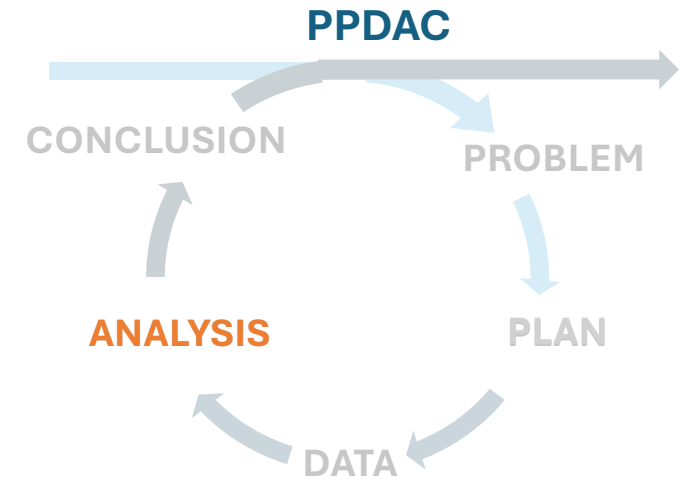
Project Title	<i>Provide a descriptive project title.</i>
Project code / identifier (if applicable)	<i>Add the project code or identifier in here if one exists. This will help with retrieval of the project materials.</i>
Project requestor / sponsor (if applicable)	<i>Add details of the requestor i.e. principal investigator, business unit, etc.</i>
GxP applicability?	<i>Indicate if this work is purely exploratory (and for internal purposes only) or if the project outcomes could be subject to regulatory interactions, part of a submission to a health authority, to a scientific publication, etc. The purpose is to help discussion and planning around potential verification and validation activities, and especially to avoid rework later.</i>
Project keywords	<i>Add keywords to help with retrieval of the project.</i>

### PURPOSE & BACKGROUND

*Provide an informal summary of the scientific/business context, and what is known about the situation at the beginning of the project. Point out the value added, including the scientific and business impact for your organization, with a clear business justification for why the project is needed? Also, provide a rationale in terms of what is already known about the problem and what gaps exist (i.e. why this project is required). It may be helpful to answer the following questions when filling out this section:*

- What problem is this project solving?*
- How do we know this is a real problem and worth solving?*

*Capture the details of any background research performed such as project identifiers or links to useful resources. It is useful to capture references to previous relevant projects, or similar work performed, to ensure existing materials and resources are utilized, as well as connecting projects to support future knowledge management and discovery.*



# Common task framework

- Learning to sail in a safe harbour
- Learning by doing
- Fostering collaboration

# Common task framework

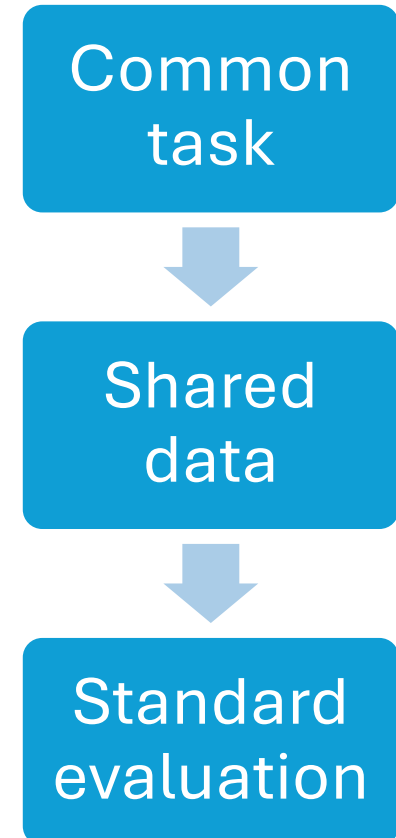
Discussion

## 50 Years of Data Science

David Donoho 

Pages 745-766 | Received 01 Aug 2017, Published online: 19 Dec 2017

 Download citation  <https://doi.org/10.1080/10618600.2017.1384734>



<https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1384734>

# Common task framework

## Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.



## IMAGENET

14,197,122 Images, 21841 synsets Indexed

[Explore](#) [Download](#) [Challenges](#) [Publications](#) [Updates](#) [About](#)

Not logged in. [Login](#) | [Signup](#)

ImageNet is an image database organized according to the **WordNet** hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.

[Click here](#) to learn more about ImageNet, [Click here](#) to join the ImageNet mailing list.

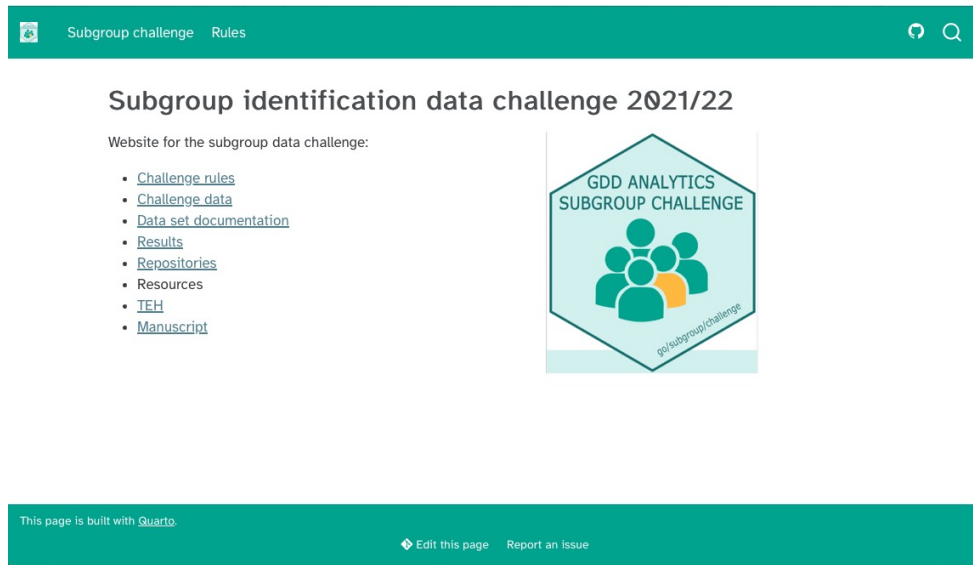


What do these images have in common? *Find out!*

[Research updates on improving ImageNet data](#)

<https://trec.nist.gov/> & <http://www.image-net.org/>

# Tiny changes



The screenshot shows the 'Subgroup challenge' website. The header is green with the text 'Subgroup challenge Rules' and navigation icons. The main content area is white and features the title 'Subgroup identification data challenge 2021/22'. Below the title, it says 'Website for the subgroup data challenge:' followed by a list of links: 'Challenge rules', 'Challenge data', 'Data set documentation', 'Results', 'Repositories', 'Resources', 'TEH', and 'Manuscript'. To the right of the list is a hexagonal logo with the text 'GDD ANALYTICS SUBGROUP CHALLENGE' and a graphic of people icons. The footer is green and contains the text 'This page is built with Quarto.' and links for 'Edit this page' and 'Report an issue'.



The screenshot shows the 'GDD Covariate Adjustment Challenge' website. The header is blue with the text 'go/covadj' and a search bar. The main content area is white and features the title 'GDD Covariate Adjustment Challenge' in large green letters. Above the title is a hexagonal logo with the text 'GDD COVARIATE ADJUSTMENT CHALLENGE' and a scatter plot. Below the title is a grid of buttons: 'Challenge Results', 'Provide Feedback', 'Getting Started', 'Challenge Rule Book', 'Data Dictionary', 'Study Protocols', 'Technical Help / Tutorials', 'Additional Resources', and 'FAQ'. The footer is white.

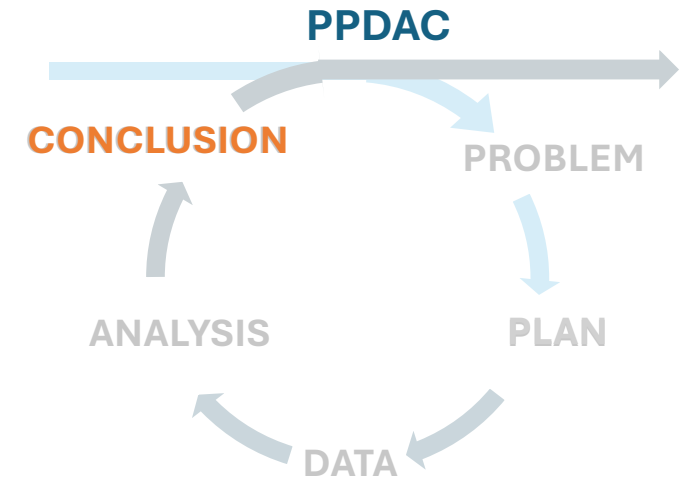
- Challenge protocols - common task framework (task, data, scoring)
- Kick off and learning meetings @MS stream
- Team repositories (methods, R code, etc.) @gitlab
- Pooled challenge data (anonymized)

# The consequences of being open

"According to Pearl and Bareinboim (2014), assumptions are **“self-destructive in their honesty.”** Such a “curse of transparency” can also occur in other situations in which researchers aim for honesty. For example, a preregistration may alert reviewers to discrepancies that would have gone unnoticed otherwise; open code may invite critical scrutiny in which reviewers would have simply assumed that no errors occurred.”

Rohrer JM, Hünernmund P, Arslan RC, Elson M. That’s a Lot to Process! Pitfalls of Popular Path Models. *Advances in Methods and Practices in Psychological Science*. 2022;5(2). doi:[10.1177/25152459221095827](https://doi.org/10.1177/25152459221095827)





# Wrap up

- Positive signs
- Effective projects involve the integration of different skills
- Practice



Welcome

Guide for Reproducible Research

**Guide for Project Design**

Overview of Project Design

**Project Design Checklist**

Creating Project Repositories

Personas and Pathways

File Naming Convention

Code Styling and Linting



☰ Contents

Aims & Values

Timeline & Milestones

Methodology

Operations

Stakeholders

Community

Outputs

Communications

Maintenance & Archiving

## Getting Started Checklist

We can begin the project design process by identifying different parts of our research, such as main research questions, methods and materials, code and data requirements, workflow, communication channels, ways of working, collaborative practices, and so on. This process allows us to be intentional from the start to ensure that our research is reproducible, well-communicated, and inclusive of all stakeholders where decisions are collaboratively made. We can explore and select the right tools and methods for reproducibility in our research and promote good practices such as documentation, version control, peer-review processes, testing, workflow, archiving, and data management plans from the beginning. Finally, we can plan for publishing and sharing research components before, during, and after the project.

**Below is a checklist you can use to help identify areas of project planning you might want to look at.**

### Aims & Values

- Define the main research questions and objectives.
- Identify the core values and principles that guide your project.
- Useful documentation: [project canvas](#), [values document](#), [project 1-pagers](#).



**OHDSI**  
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

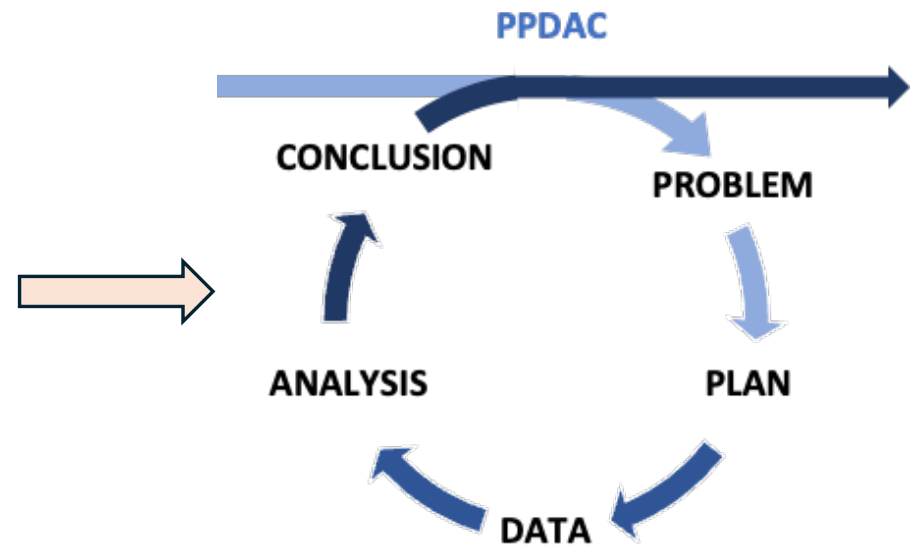
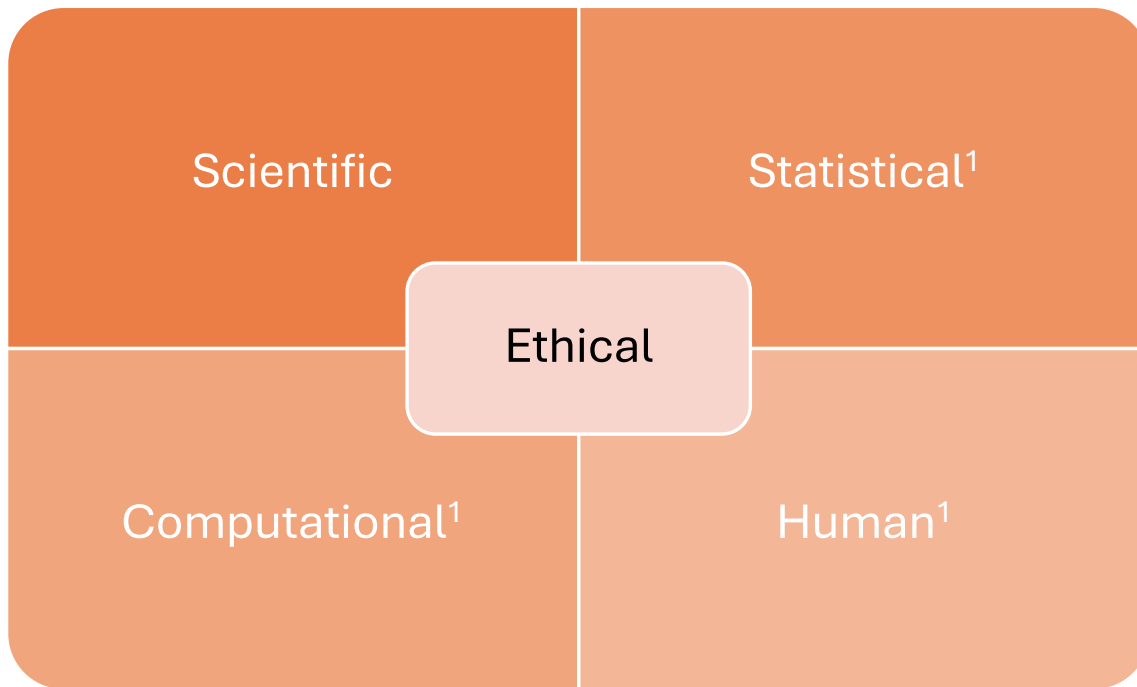
## LEGEND Guiding Principles

1. Evidence will be generated at **large-scale**.
2. **Dissemination** of the evidence will not depend on the estimated effects.
3. Evidence will be generated by consistently applying a **systematic approach** across all research questions.
4. The evidence will be generated using a **pre-specified** analysis design.
5. The evidence will be generated using **open source** software that is freely available to all.
6. The evidence generation process will be **empirically evaluated** by including control research questions where the true effect size is known.
7. The evidence will be generated using **best-practices**.
8. LEGEND will **not** be used to **evaluate methods**.
9. The evidence will be **updated** on a regular basis.
10. **No patient-level data** will be shared between sites in the network, only aggregated data.

Martijn J Schuemie, Patrick B Ryan, Nicole Pratt, RuiJun Chen, Seng Chan You, Harlan M Krumholz, David Madigan, George Hripcsak, Marc A Suchard, Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND), *Journal of the American Medical Informatics Association*, Volume 27, Issue 8, August 2020, Pages 1331–1337, <https://doi.org/10.1093/jamia/ocaa103>

# Data science thinking

A set of integrated thinking skills and practices refocused for answering questions with data



<sup>1</sup> Blei & Smyth (2017) *Science and data science*. PNAS 114 (33):8689-8692

# Good data science practice

---

Data science is only the current label that represents a trend towards the integration of computational and statistical practises.

We should also keep in mind that Science itself has problems, and critique alone will not help address these issues.

It is easy for statisticians to be cynical of the current data science trend, but critique from the outside looking in is not enough to address the wider problems of poor scientific practice.

As statisticians we should be prepared to engage early with other disciplines who possess different skill sets and perspectives that balance our own.

This engagement should start early to enable statistical thinking to influence the direction from the outset, and not towards the end in firefighting mode.

*“What would data science look like if its key critics were engaged to help improve it, and how might critiques of data science improve with an approach that considers the day-to-day practises of data science? (Neff et al. 2017)”*

# Acknowledgments

- Conor Moloney
- Carsten Philipp Mueller
- Lukas Widmer
- Jonas Dorn
- Peter Krusche
- Jelena Cuklina
- Kostas Sechidis
- Frank Bretz
- Janice Branson
- David Ohlssen
- Prashanti Goswami
- + others

# Definitions

Desired attribute	Question	Researcher	Data	Analysis	Result
<b>Repeatable</b>	Identical	Identical	Identical	Identical	= Identical
<b>Reproducible</b>	Identical	Different	Identical	Identical	= Identical
<b>Replicable</b>	Identical	Same or different	Similar	Identical	= Similar
<b>Generalizable</b>	Identical	Same or different	Different	Identical	= Similar
<b>Robust</b>	Identical	Same or different	Same or different	Different	= Similar
<b>Calibrated</b>	Similar (controls)	Identical	Identical	Identical	= Statistically consistent

Source: <https://ohdsi.github.io/TheBookOfOhdsi/EvidenceQuality.html>