Applied Machine Learning Days

EPFL

AMLD

March 23-26
2024

Lausanne
Switzerland

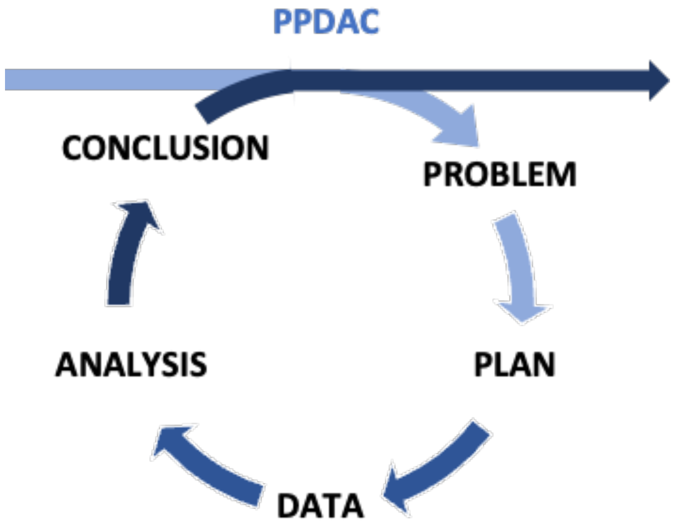
Data science thinking: making an impact

Kostas Sechidis & Mark Baillie
AMLD, Lausanne
March 24th, 2024

NOVARTIS | Reimagining Medicine

Today's agenda

Schedule	Activity and topics
35 mins	**Introduction**: motivating data science thinking
55 mins	**Problem** & **Plan**: getting questions right
30 mins	Break
55 mins	**Data** & **Analysis**: developing & executing a strategy
30 mins	**Conclusions:** effective visual communication



All the materials and more are online

https://datascience-thinking.github.io

Who we are …

Mark Baillie

Kostas Sechidis

Prashanti Goswami

Frank Bretz

Getting to know you!

What's your background: academia, industry, public sector, or other?

Do you consider yourself a data scientist?

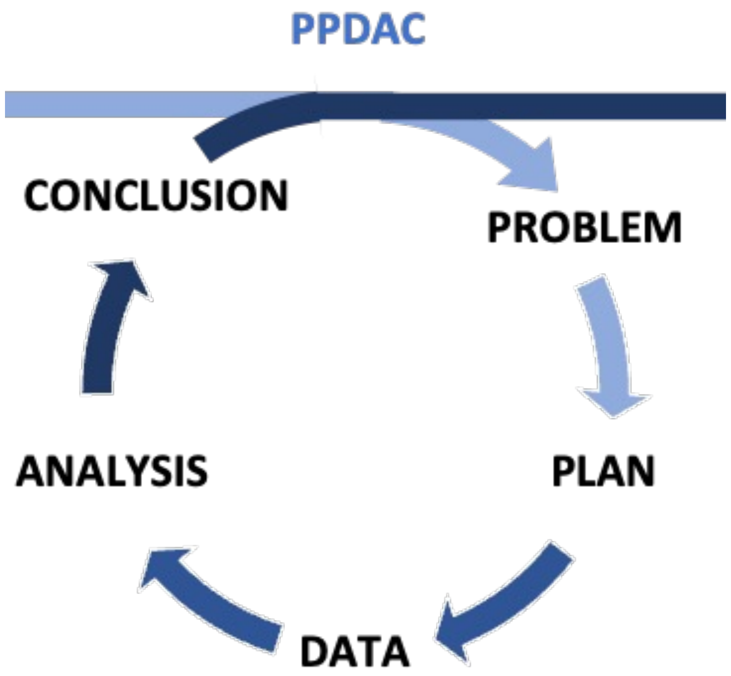
What do you hope to learn from this workshop?

Data science thinking process

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.

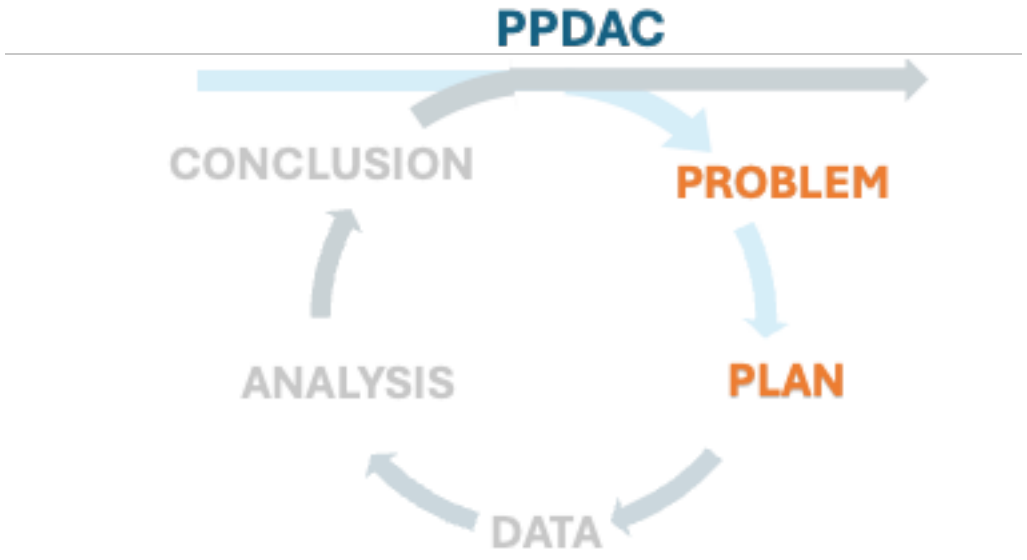
This workflow demonstrates the steps for abstracting and solving **a real problem**. An impactful solution requires a clear understanding of how things work.



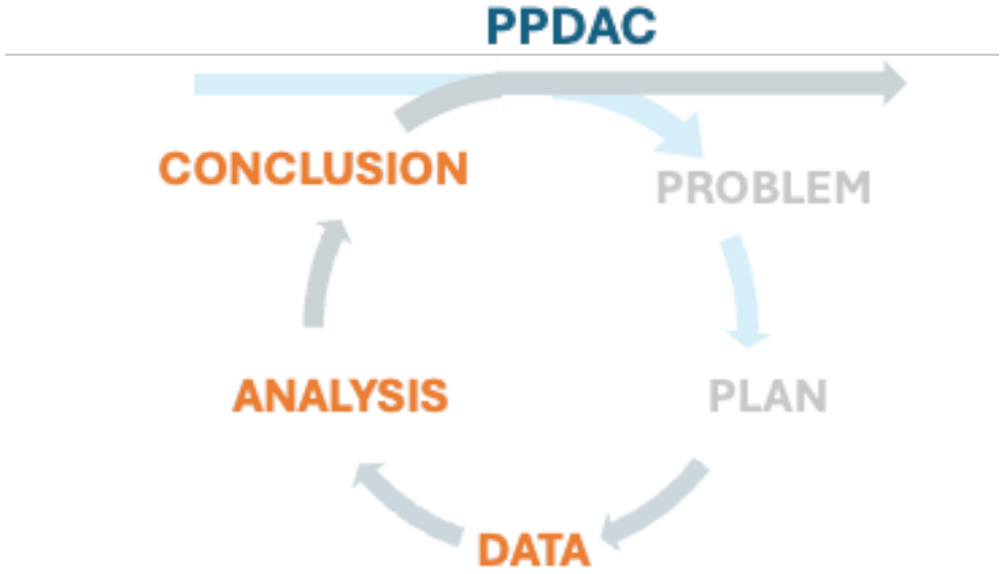
Source: MacKay, R.J. and Oldford, R.W., 2000. Scientific method, statistical method and the speed of light. *Statistical Science*, pp.254-278.

Data science thinking: making an impact

Part I



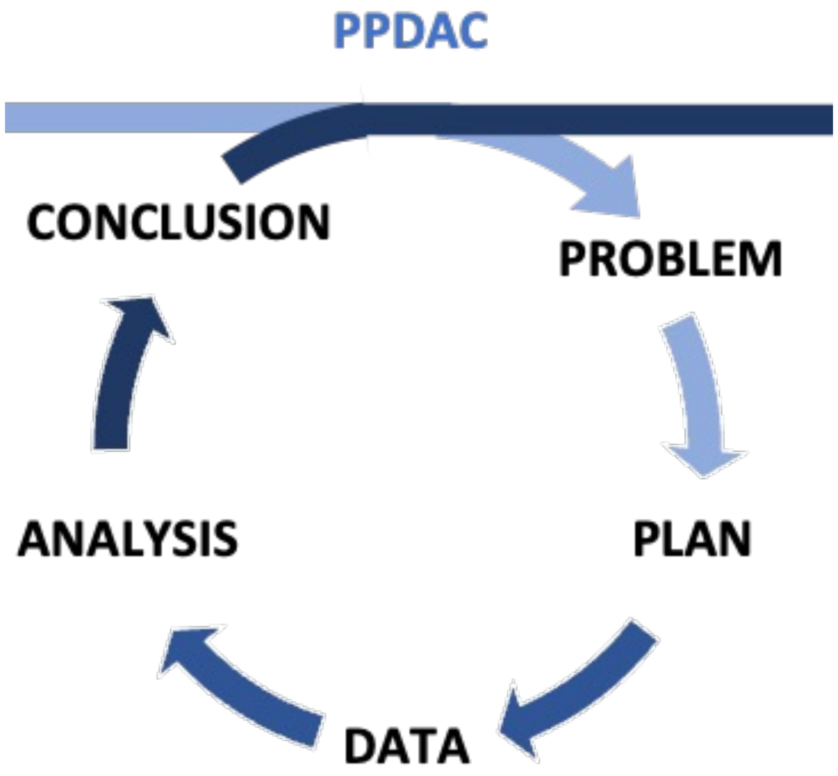
Part II



Learning objectives

By the end of the workshop, you will recognize:

- The integrated set of thinking skills and practices of data science thinking, refocused for answering questions with data.
- It is essential to have a clear understanding of the problem to provide an impactful solution.
- Strategies for problem formulation and how using these tactics is important before delving into solutions.
- The key skill of asking great questions in understanding the problem and its context, as well as in establishing trust and building relationships.
- How question framing helps in defining the broader problem, which has a domino effect on the subsequent workflow phases.



Introduction

Motivating the data science thinking process

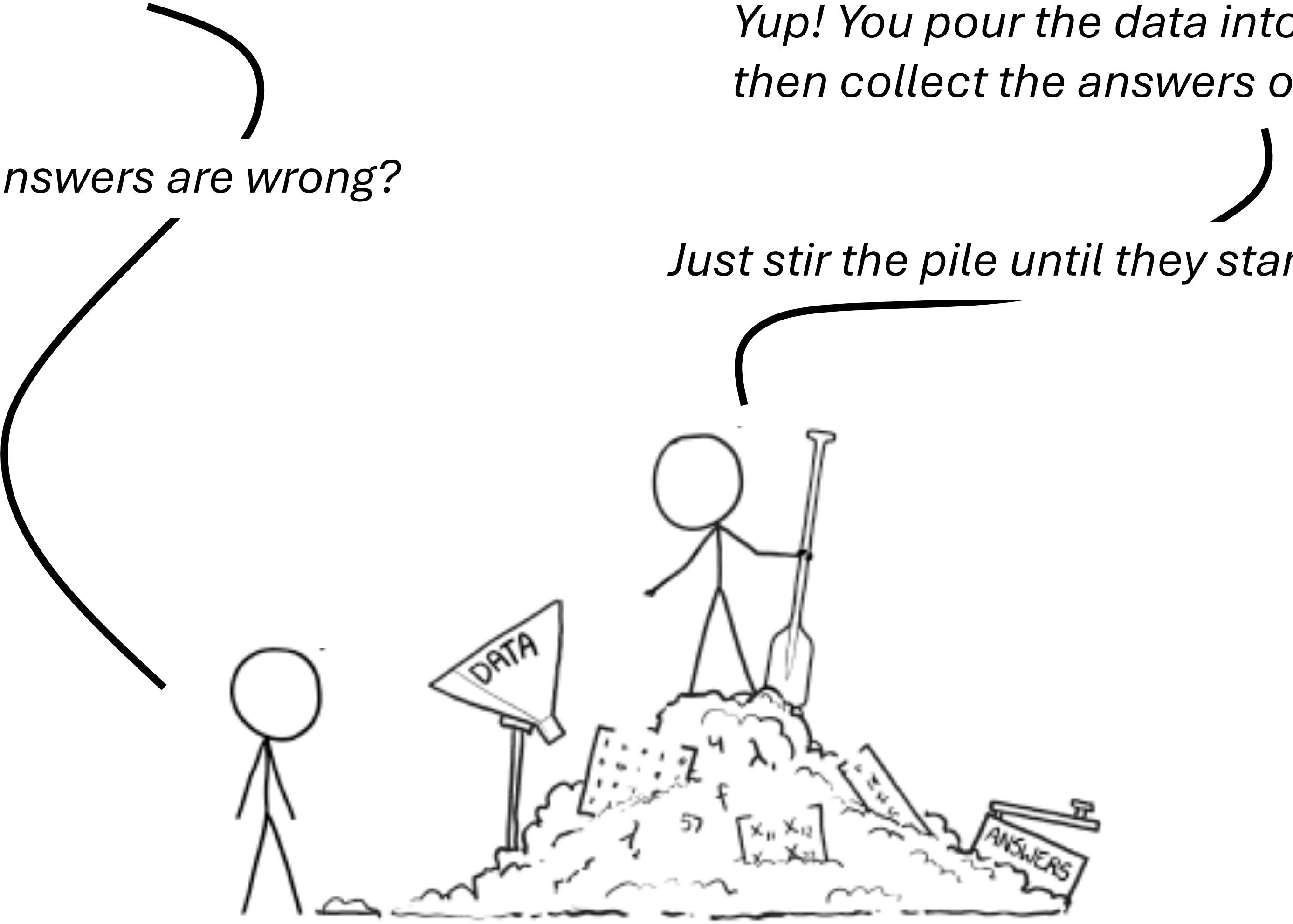
Why this course?

This is your Machine Learning System?

Yup! You pour the data into this big pile of math, then collect the answers on the other side.

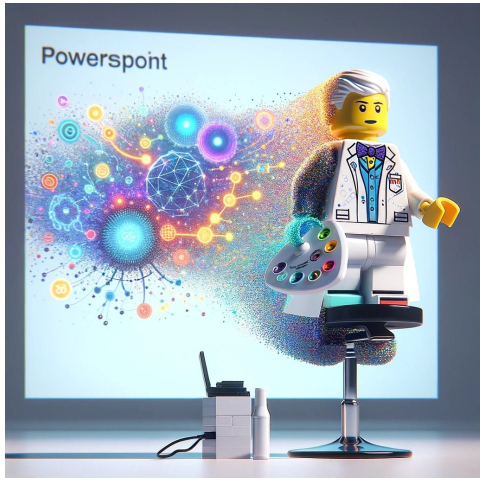
What if the answers are wrong?

Just stir the pile until they start looking right.



Credit: https://xkcd.com/1838/ (adapted)

Why this course? – ensuring impact



Domain expert

Numbers expert

!!!

Here are 100 brain images – can you run an algorithm and let me know the results?

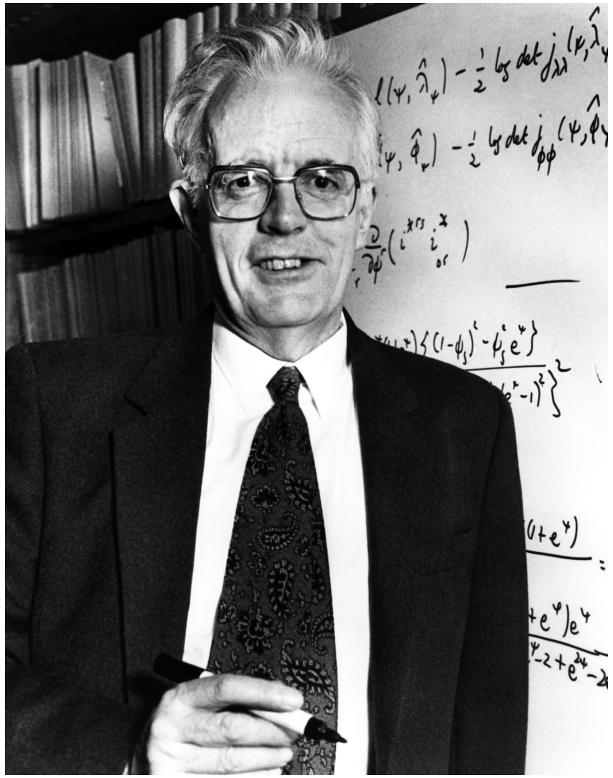
Happy to help! But what exactly are you are trying to answer? What's your aim?

Well, whether you can crack the disease.

What do you mean by 'crack the disease'? Also, how did you collect the data? How did you decide on having 100 images?

Good questions… not quite sure, to be honest. When can I see the results?

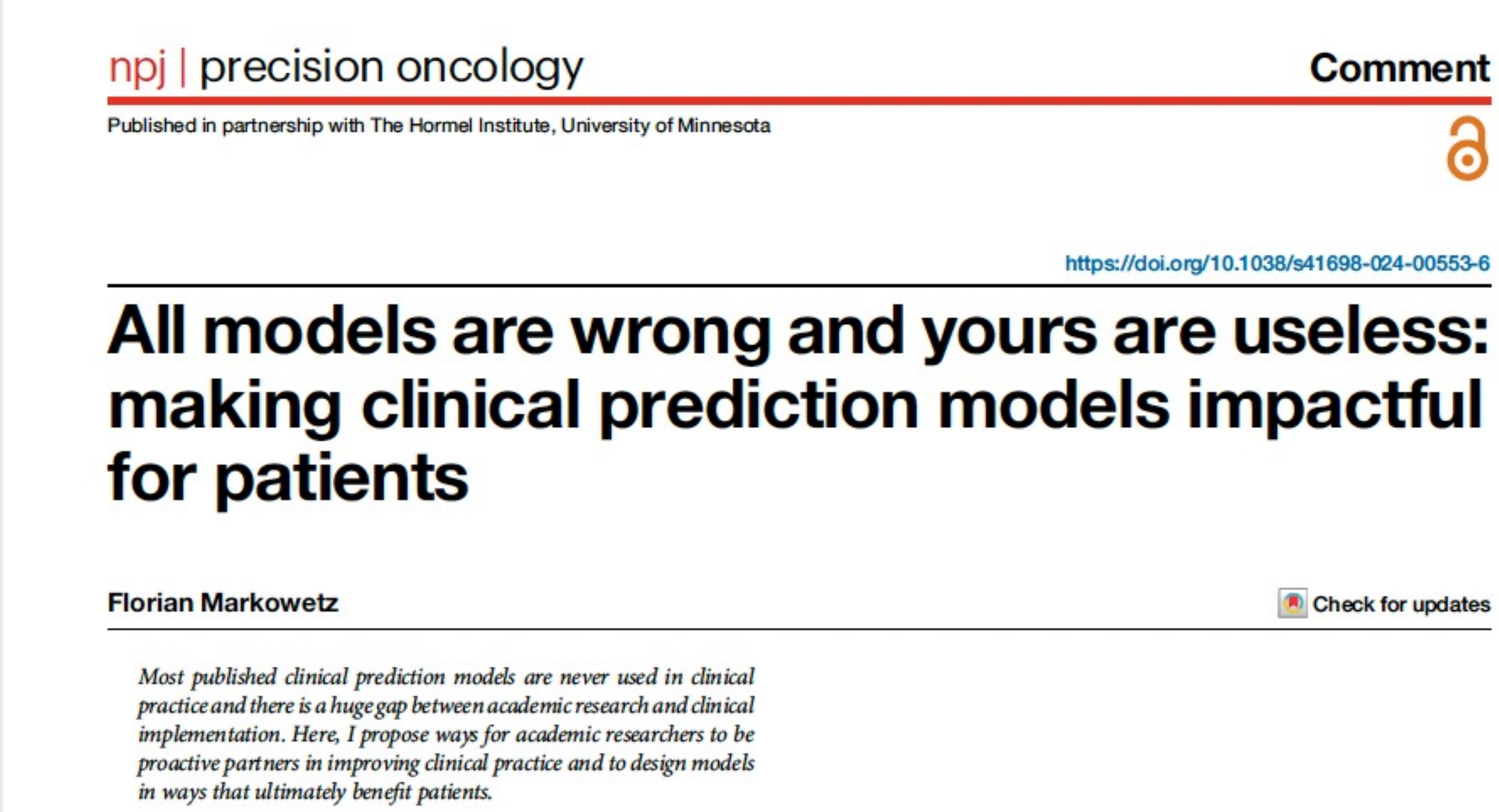
Why this course? – Avoiding Type III errors



> *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 this course? – 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

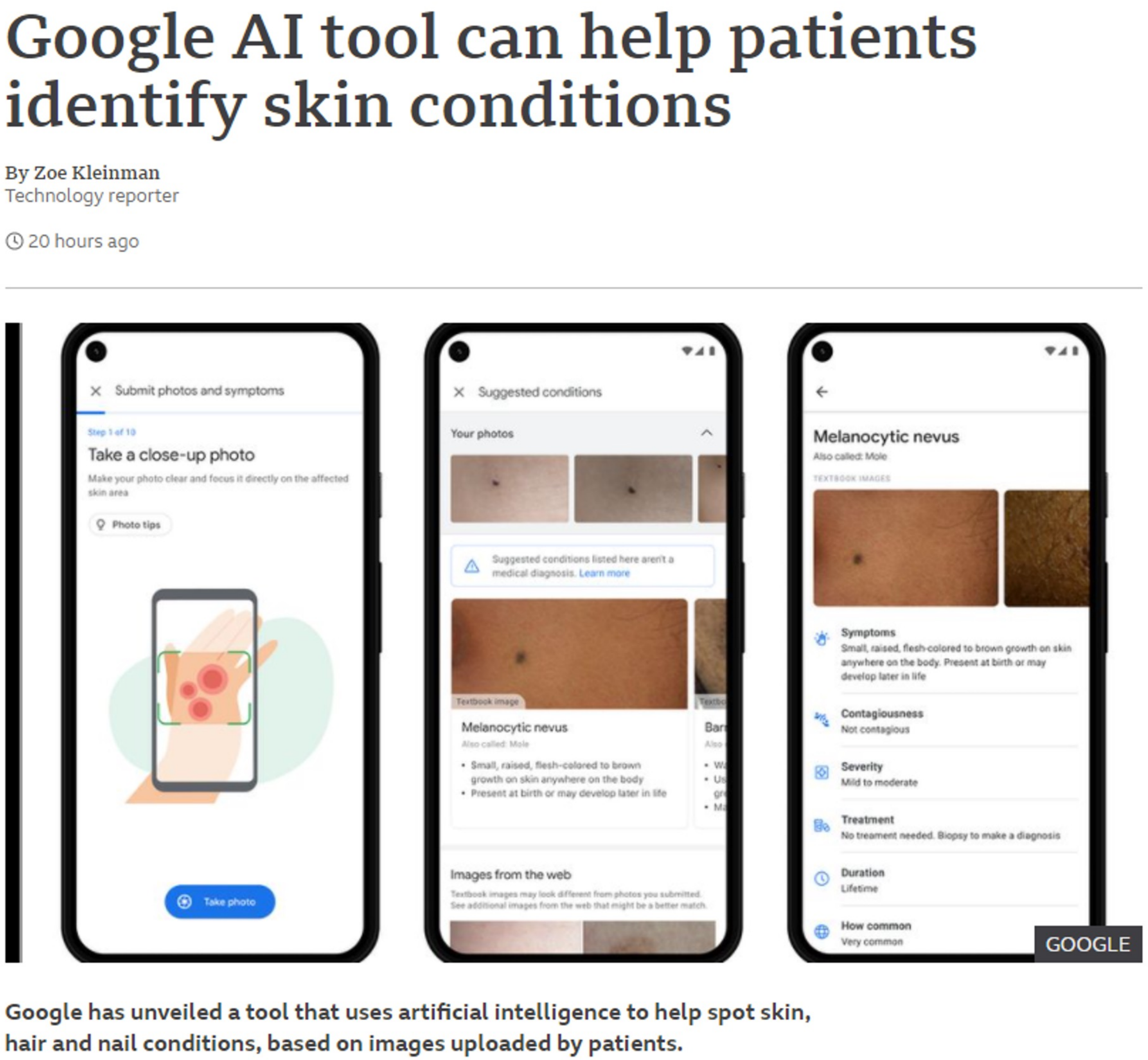
Check for updates

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.

Data Science

Motivating the data science thinking process

Advances in data science



Google AI tool can help patients identify skin conditions

By Zoe Kleinman
Technology reporter

20 hours ago

Submit photos and symptoms

Step 1 of 10

Take a close-up photo

Make your photo clear and focus it directly on the affected skin area

Photo tips

Take photo

Suggested conditions

Your photos

Suggested conditions listed here aren't a medical diagnosis. Learn more

Textbook image

Melanocytic nevus

Also called: Mole

- Small, raised, flesh-colored to brown growth on skin anywhere on the body
- Present at birth or may develop later in life

Images from the web

Textbook images may look different from photos you submitted. See additional images from the web that might be a better match.

Melanocytic nevus

Also called: Mole

TEXTBOOK IMAGES

Symptoms
Small, raised, flesh-colored to brown growth on skin anywhere on the body. Present at birth or may develop later in life

Contagiousness
Not contagious

Severity
Mild to moderate

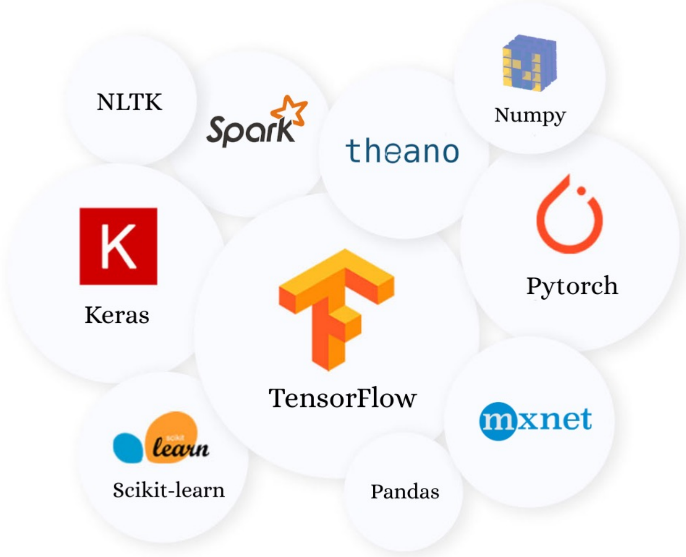
Treatment
No treament needed. Biopsy to make a diagnosis

Duration
Lifetime

How common
Very common

GOOGLE

Google has unveiled a tool that uses artificial intelligence to help spot skin, hair and nail conditions, based on images uploaded by patients.

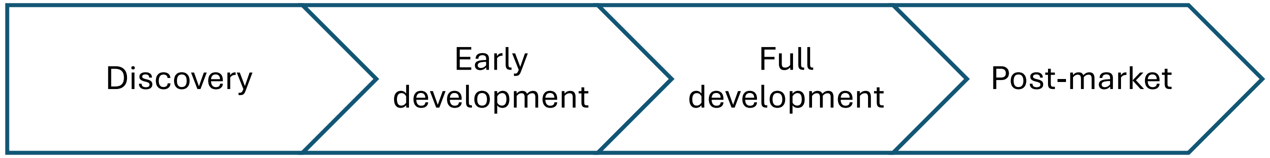


Source: https://www.bbc.com/news/technology-57157566

Can “data science” accelerate drug development?

By **learning from existing and future data** using advances in science, statistics, machine learning, computation, AI, etc. to:

- Increase understanding of drug, disease and patients,
- accelerate and improve development projects,
- inform decision making.

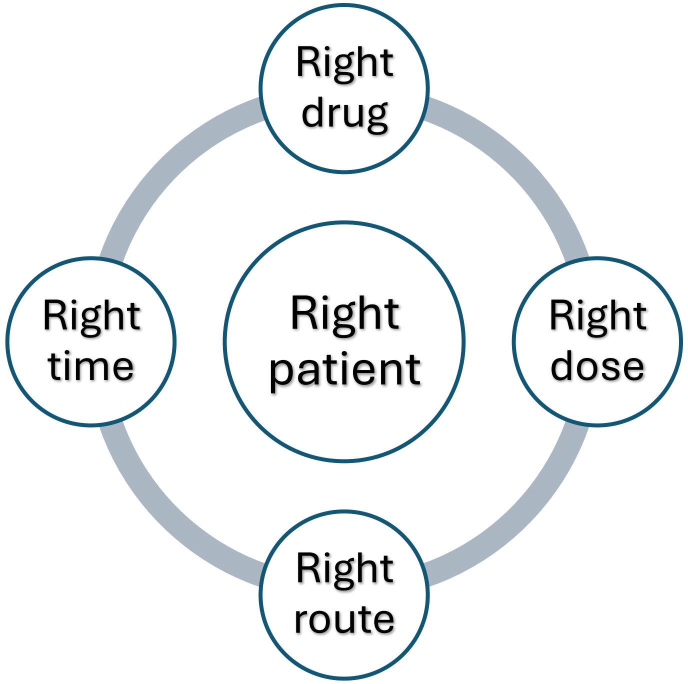


Discovery

Early development

Full development

Post-market



Right drug

Right time

Right patient

Right dose

Right route

Many questions in drug development

Leveraging Data Science in Drug Development

- **Treatment effect heterogeneity:** Who will respond better?
- **Disease Progression:** Better measures of disease/disability progression?
- **Clinical prediction:** What are the disease risk predictors?
- **Digital Sensors:** decentralised patient measurements?
- **Informative outcomes:** Developing and validating endpoints

Frontiers of Data Science:

- Causal Inference
- Multi-Modal Inference
- AI & medical devices
- Synthetic and Historical Controls
- Digital Twins

Can “data science” accelerate drug development?

By learning from existing and future data using advances in science, statistics, machine learning, computation, AI, etc. to:

- increase our understanding of drug, disease and patients,
- accelerate and improve our development projects, and
- inform our decision making.

What do we mean by data science?

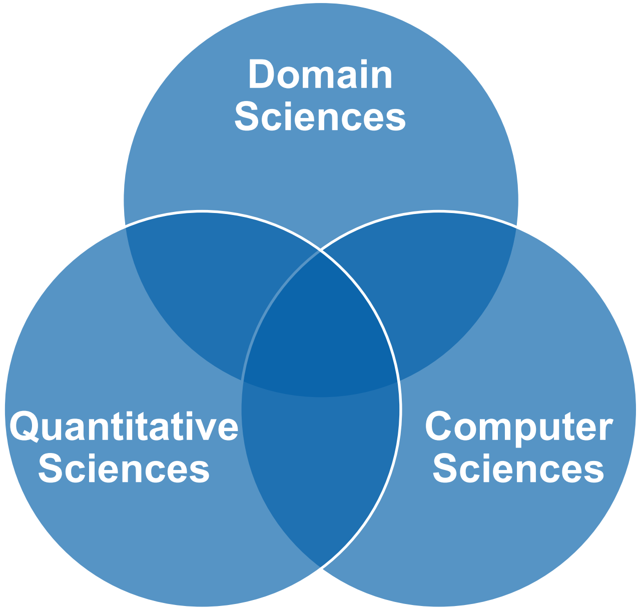
What are the practices of data science?

What do we mean by “good” practice in the context of drug development?

What is Data Science?

> “ *Data science is the study of extracing value from data.* ” – *Jeannette Wing*[1]

- 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



[1] Source: https://datascience.columbia.edu/news/2018/what-is-data-science

What is Data Science?

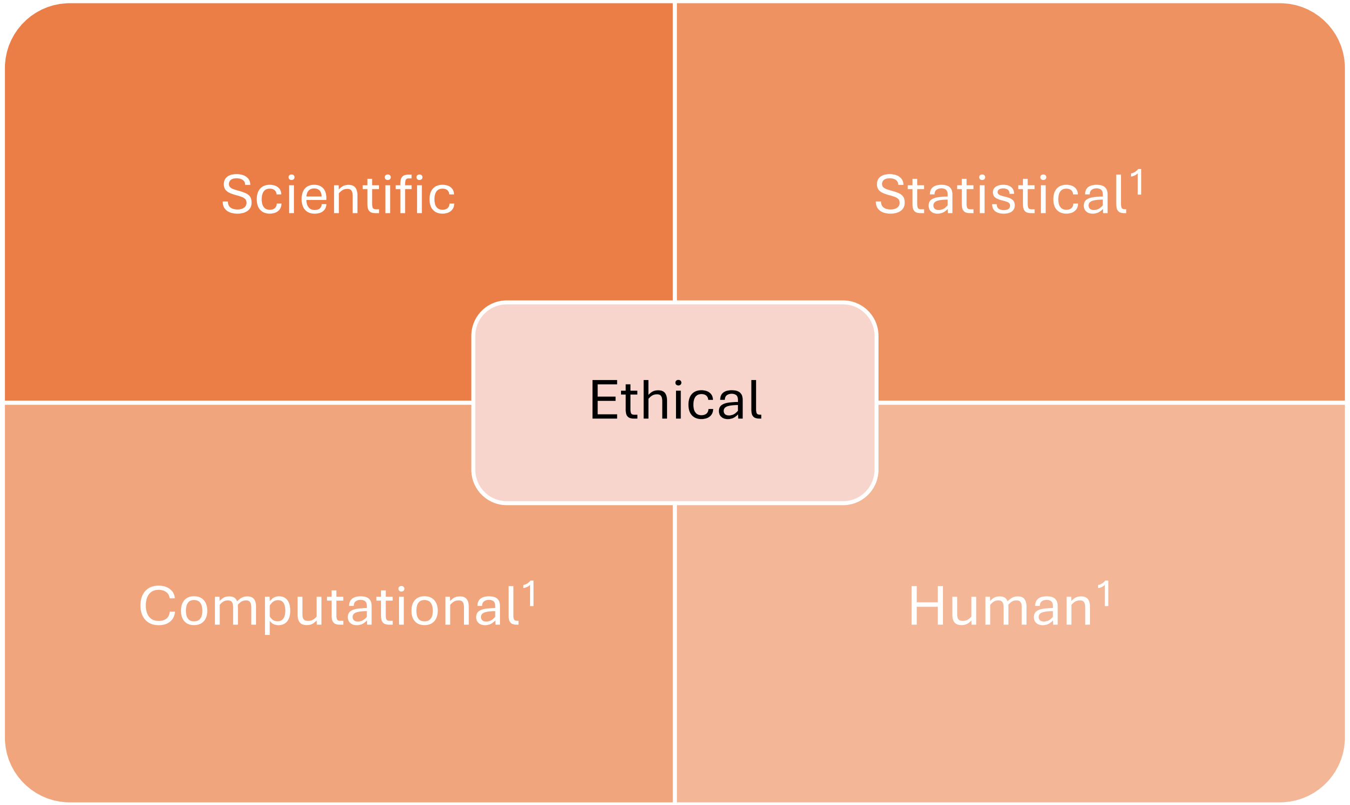
> “ *Data science is the study of extracing value from data.* ” *– Jeannette Wing*[1]

- 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

[1] Source: https://datascience.columbia.edu/news/2018/what-is-data-science

Data science thinking

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



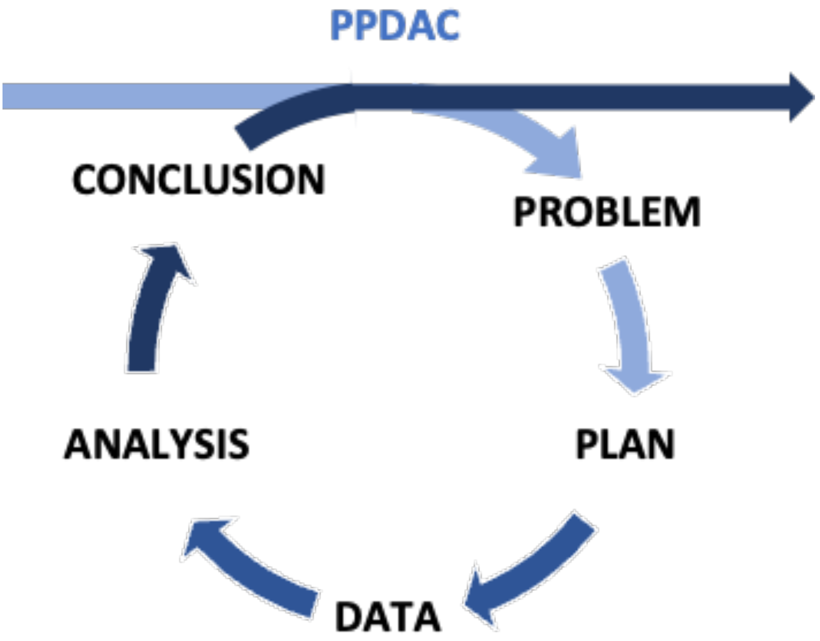
Scientific

Statistical[1]

Ethical

Computational[1]

Human[1]

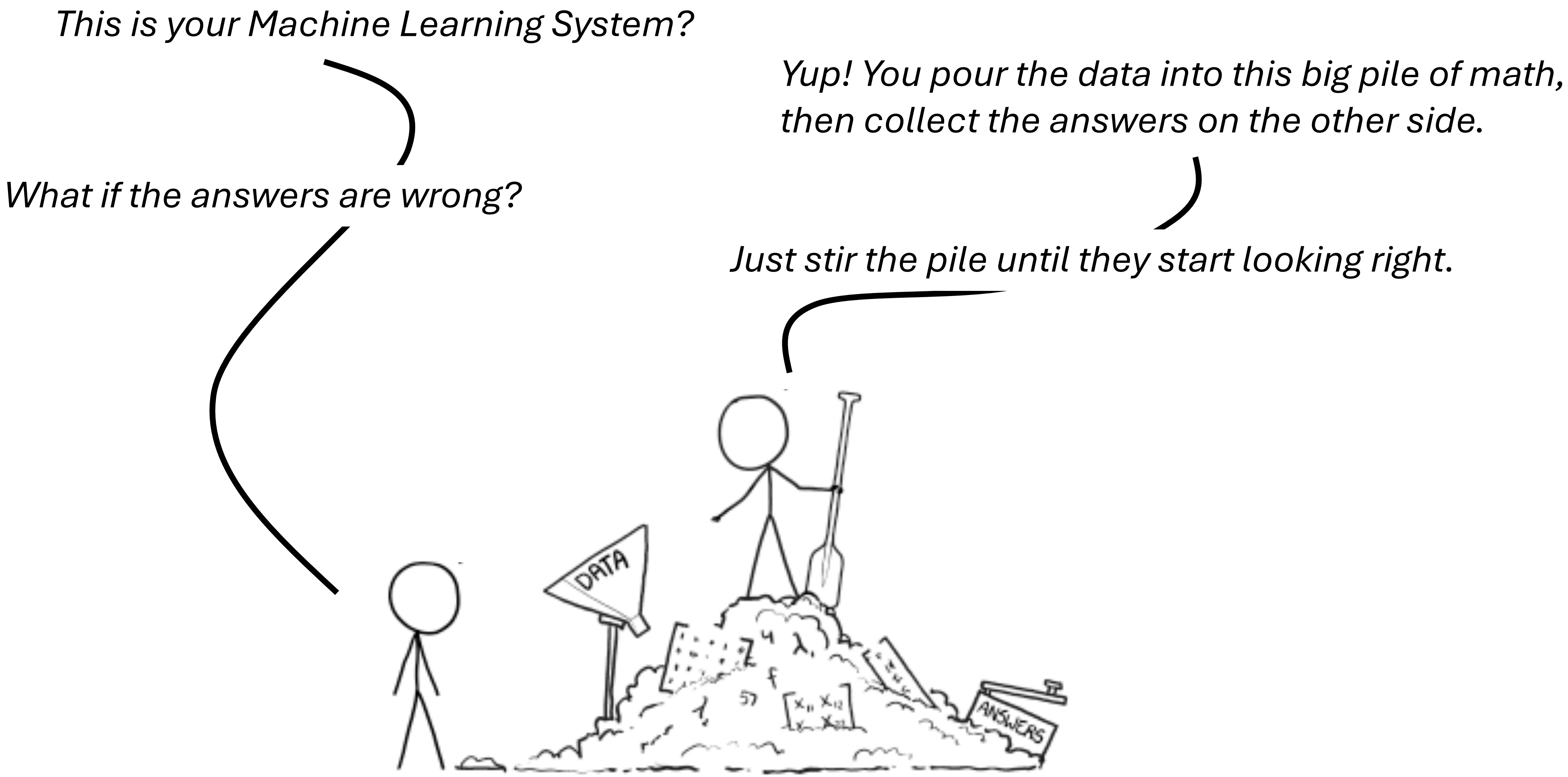


[1] *Blei & Smyth (2017) Science and data science. PNAS 114 (33):8689-8692*

Why good practice?

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.

Remember why this course?


This is your Machine Learning System?
Yup! You pour the data into this big pile of math, then collect the answers on the other side.
What if the answers are wrong?
Just stir the pile until they start looking right.
DATA
ANSWERS

Credit: https://xkcd.com/1838/ (adapted)

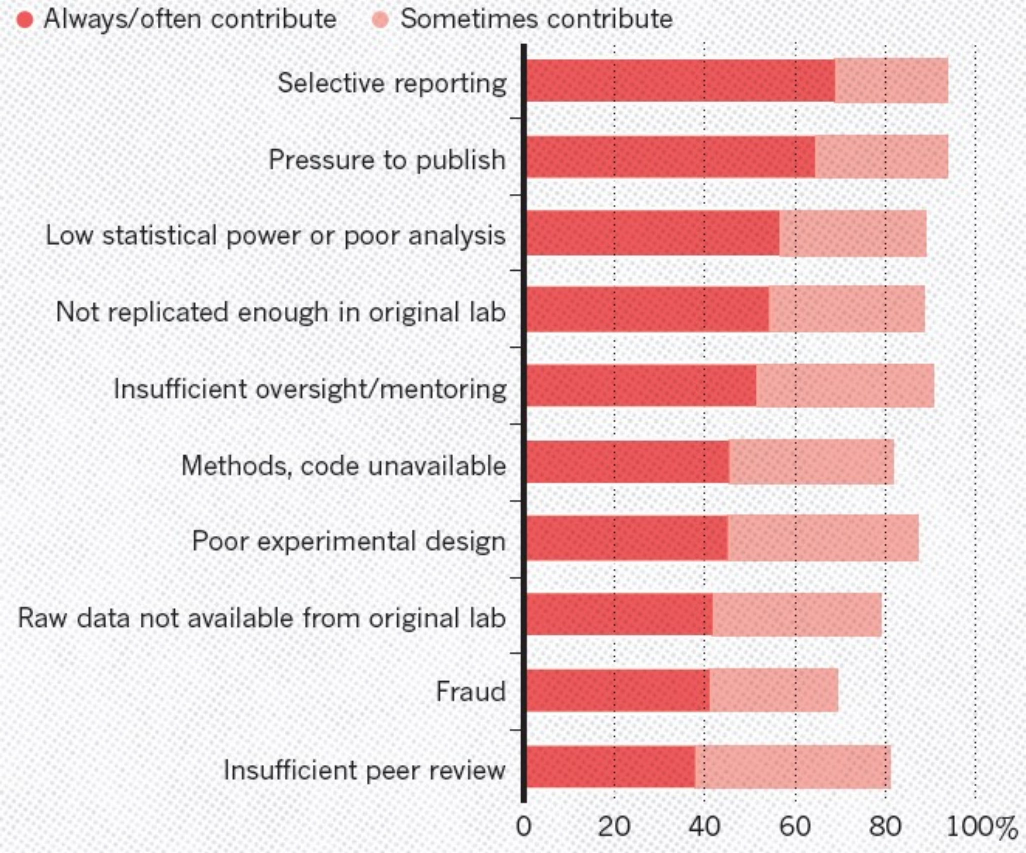
What are the practices of data science?

A crisis facing science?

1,500 scientists lift the lid on reproducibility

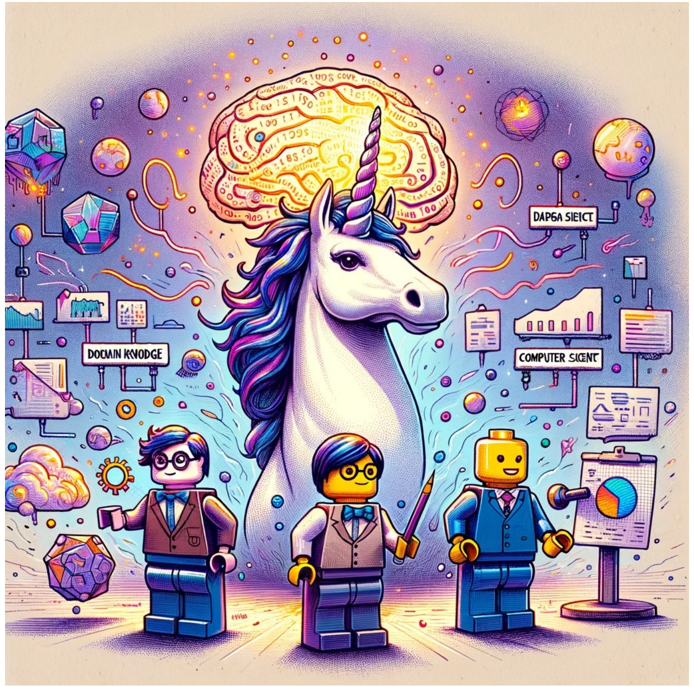
Monya Baker survey sheds light on the 'crisis' rocking research.

What factors contribute to irreproducible research?
(Nature's survey of 1,576 researchers)

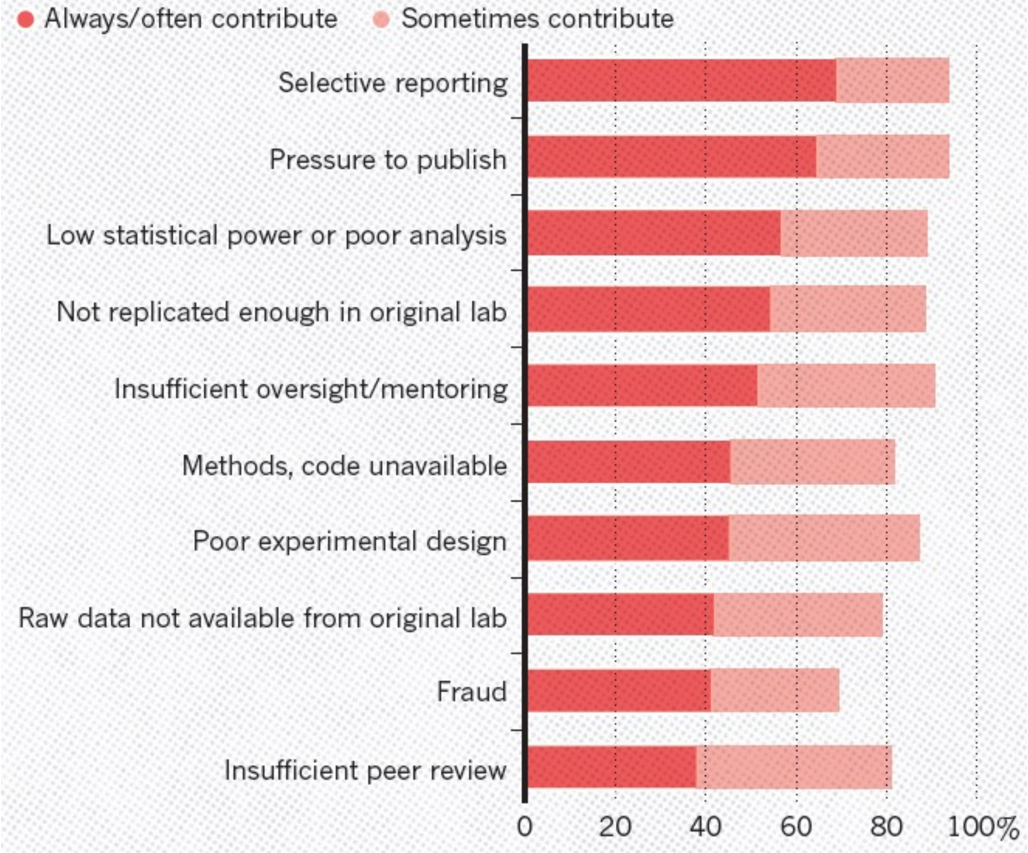


What are the practices of data science?

- Computational
- Statistical
- Scientific
- Ethical and legal
- Human



What factors contribute to irreproducible research?
(Nature's survey of 1,576 researchers)

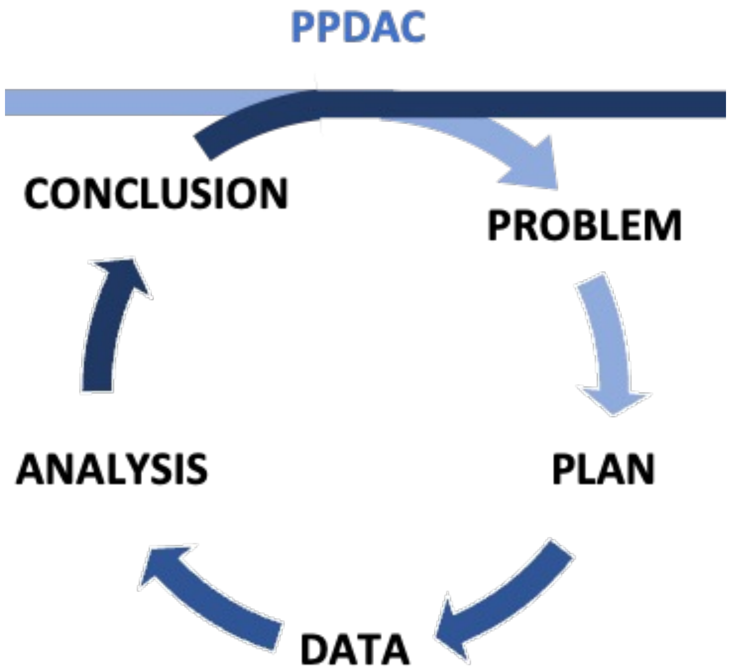


Data science thinking process

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.

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



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

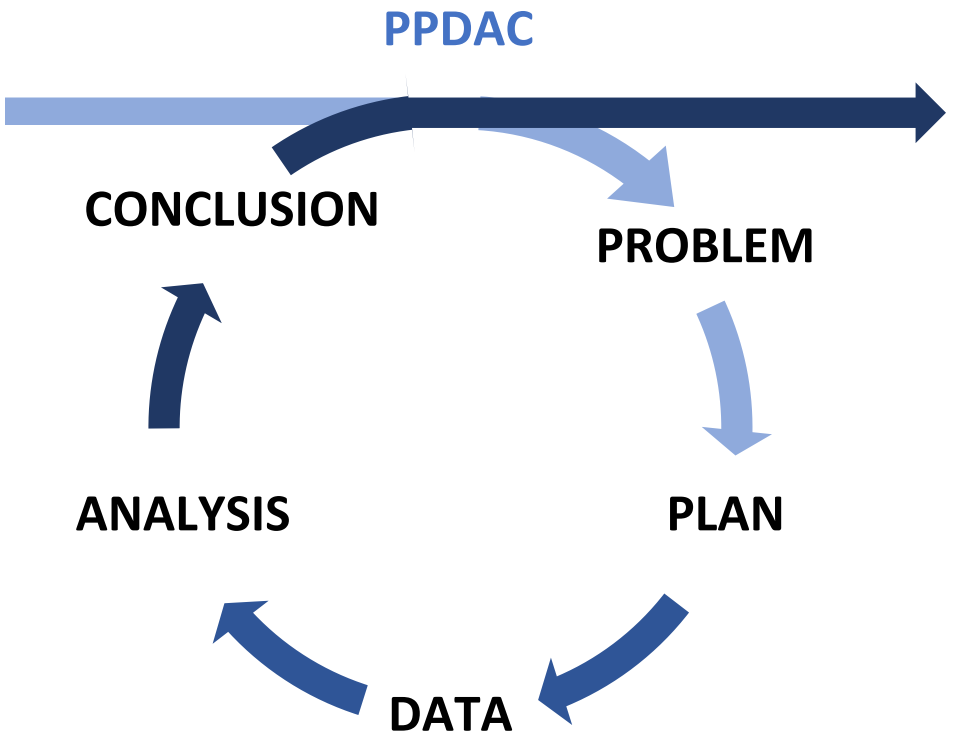
Data Science Thinking Process

How to make an impact

Data science thinking process

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

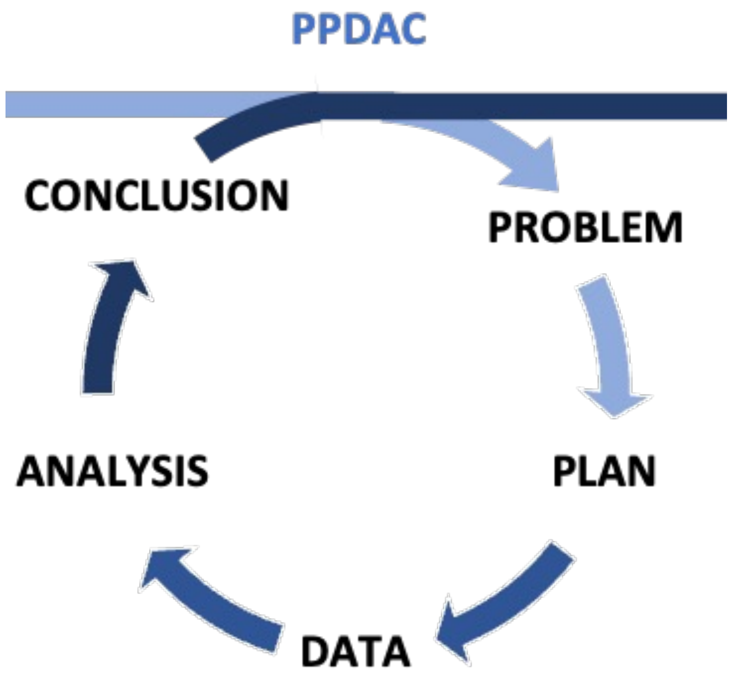


Data science thinking process

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

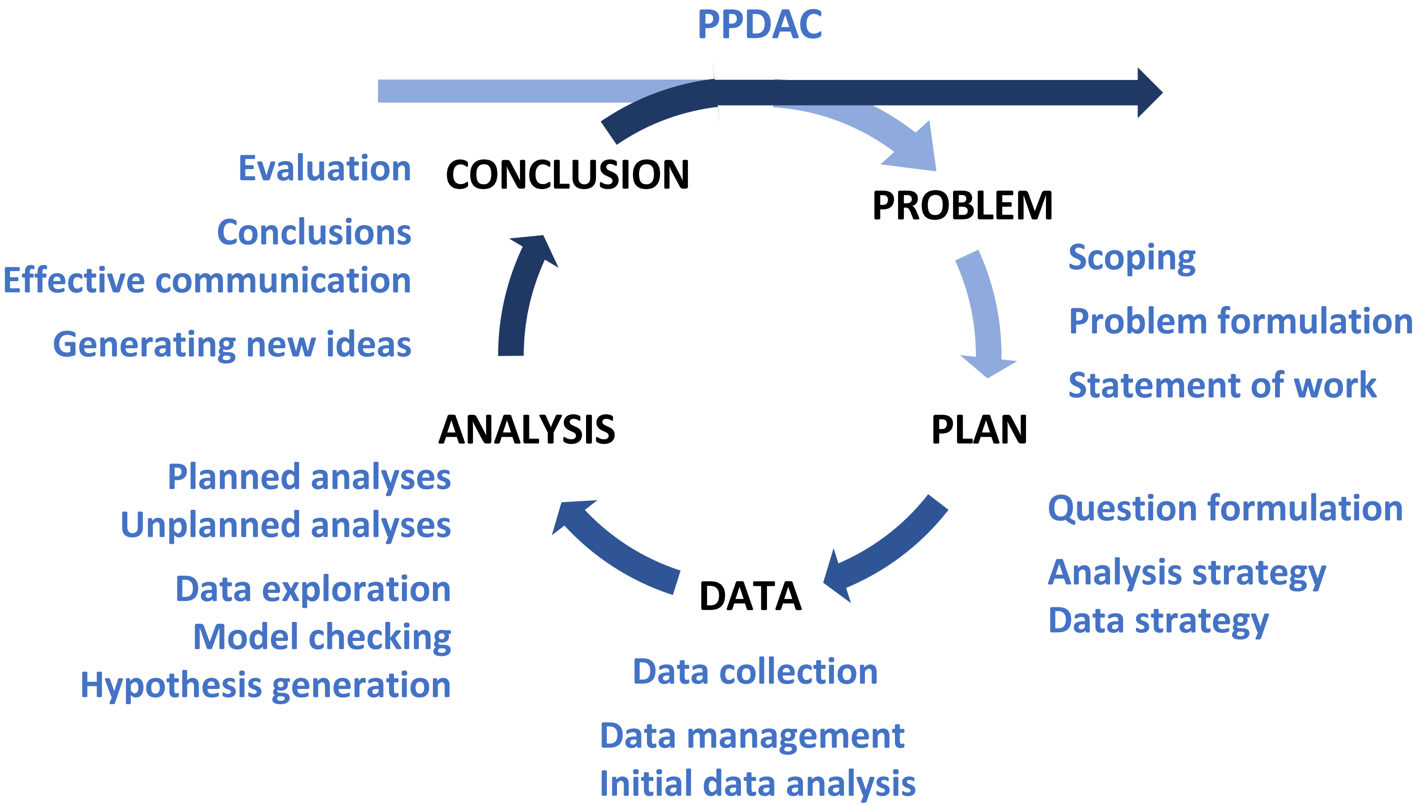
A good **workflow** is an established set of habits that help drive you forward towards your goal. They enable complexity to scale in the right areas.

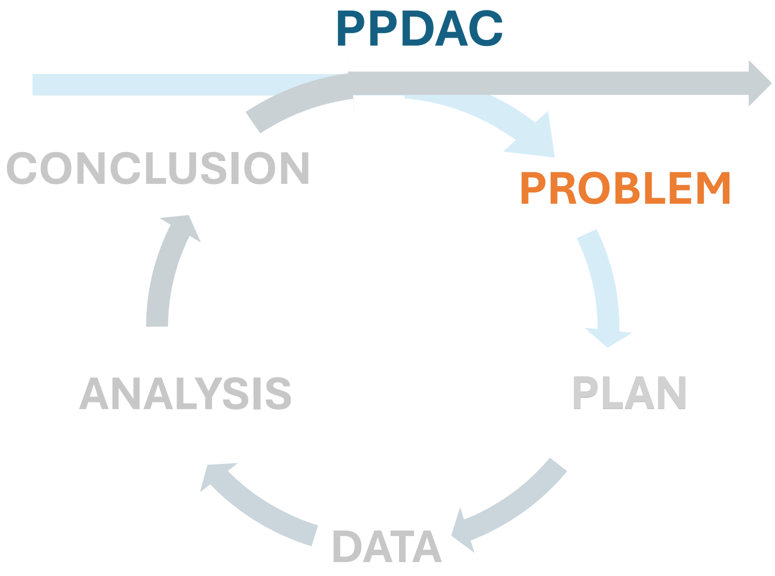
This workflow demonstrates the steps for abstracting and solving **a real problem**. An impactful solution requires a clear understanding of how things work.



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 process





Problem

- elicit and understand the problem
- asking great qustions
- translating into answerable questions

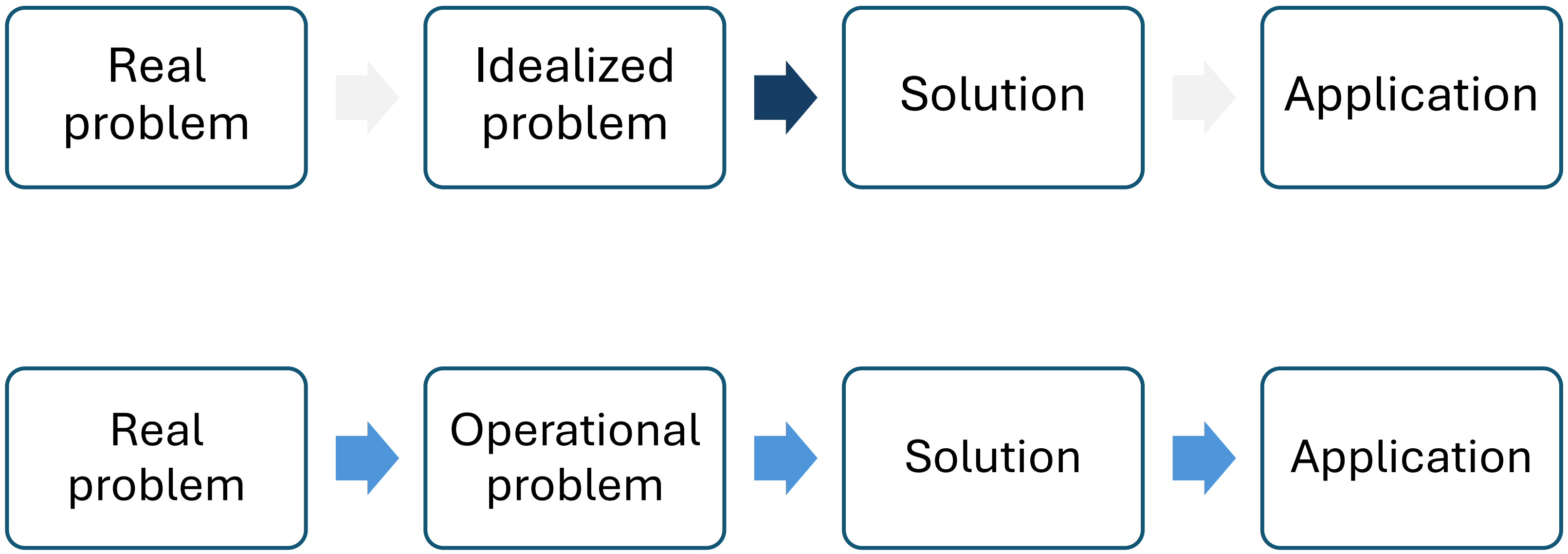
The importance of problem formulation

“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)

- **From Aim to Question:** Translate high-level aims into tangible, answerable questions.
- **Context Comprehension:** Strive to grasp the full scope, aims, objectives, and constraints
- **Objectives vs Questions:** Differentiate broad research aims from specific, data-driven (research) questions
- **Initial Scoping:** Investigate the context thoroughly to inform objectives and outcomes
- **Critical Questioning:** Ask expansive questions to preclude Type III errors and gain essential domain understanding to help formulate answerable questions.
- **Multiple Pathways:** At this phase explore various approaches to achieving project aims.

Why? - impact and problem formulation

"I have never let my schooling interfere with my education." **Mark Twain**



Real problem → Idealized problem → Solution → Application

Real problem → Operational problem → Solution → Application

https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/1467-9884.00130

Philosophy of language

- A young married couple move into a new apartment and decide to repaper the dining room.
- They call on a neighbor who has a dining room the same size and ask:
- **“How many rolls of wallpaper did you buy when you papered your dining room?”**
- “Seven,” he says.
- So the couple buys seven rolls of expensive paper, and they start papering.
- When they get to the end of the fourth roll, the dining room is finished.

h/t Miguel Hernán

Asking the wrong question

- Annoyed, they go back to the neighbor and say,
- “We followed your advice, but we ended up with three extra rolls!”
- “So,” he says, “that happened to you too.”
- Oops!

“How many rolls of wallpaper **did you** buy when you papered your dining room?”

Problem formulation: tactics

- **Objective:** Gain a thorough understanding of the problem, its context, and potential impact.
- **Engage with experts:** Identify and clarify crucial details with domain experts:
  - Purpose, objectives, and expected outcomes.
  - Investigation phases and types.
  - Attributes vital for crafting focused questions.
  - Project boundaries and constraints.
- **Strategize information gathering:**
  - Formulate effective tactics to collect information efficiently.
  - “asking great questions” (Vance et al. 2022)
- **Questioning techniques:** Employ probing questions to convert domain expertise into precise research questions.
- **Communicate effectively:** Use clear, direct language to foster understanding and build trust.

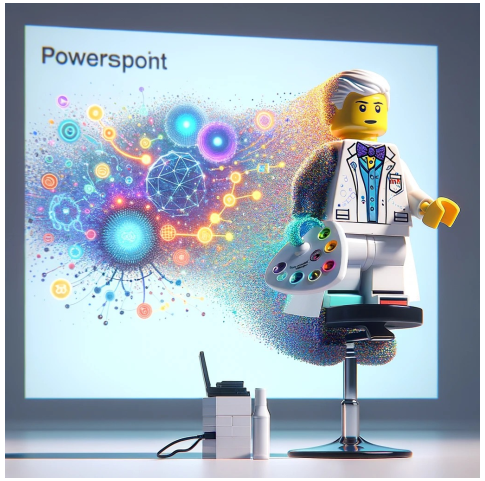
Bad, good and great questions

1. So that we do not waste any more time, what is the exact statistics question you need my help answering?
2. With respect to the scientific area to which these ideas refer, just what are they about?
3. My goal for this initial meeting is to understand your research problem better, which will help me think about the specific statistical issues. What would you like to accomplish in this meeting?

Asking great questions

- Great questions (Vance et al., 2022)
  - elicit information useful for accomplishing project tasks
  - Strengthen team relationships and understanding
    - i.e. better understanding between ”domain” and ”number” experts
- Great questions have 3 parts:
  - The question
  - The answer
  - Paraphrasing the answer to create shared understanding
- Strategies
  - Preface questions with intent behind asking the question

Asking good questions



Domain expert

Numbers expert

!!!

Here are 100 brain images – can you run an algorithm and let me know the results?

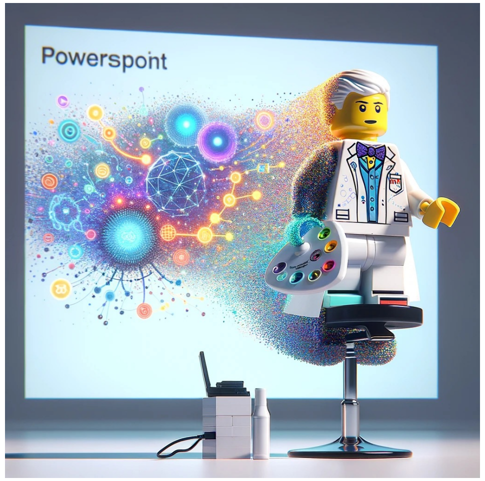
Happy to help! But what exactly are you are trying to answer? What's your aim?

Well, whether you can crack the disease.

What do you mean by 'crack the disease'? Also, how did you collect the data? How did you decide on having 100 images?

Good questions… not quite sure, to be honest. When can I see the results?

Asking great questions



Domain expert

Numbers expert

Here are 100 brain images – can you run an algorithm and let me know the results?

Happy to help! Tell me more about how the results would push forward the project?

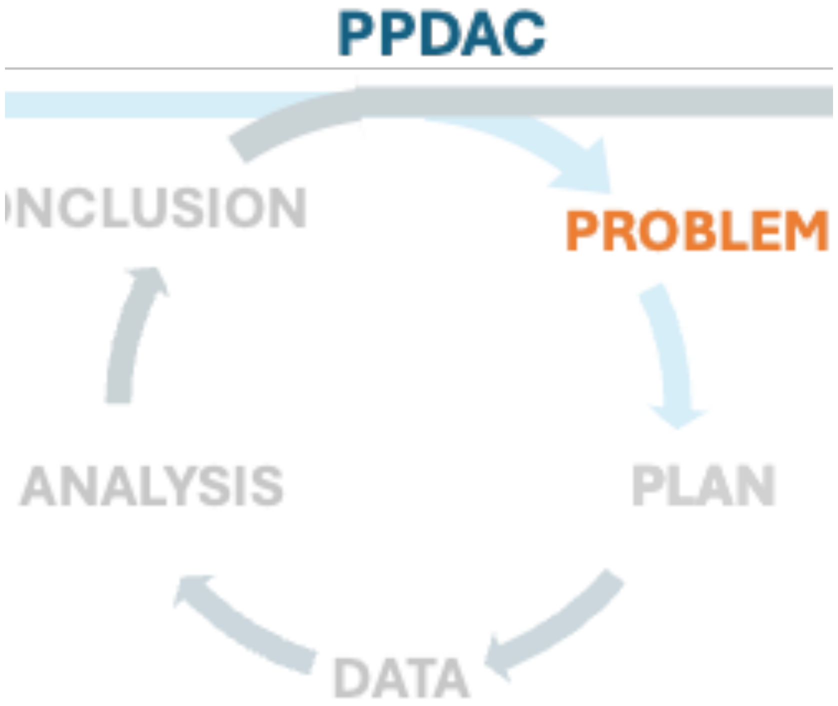
Well, if we can determine if the scans measure disease progression we could use these measurements in the upcoming clinical trial for drug X to account for treatment heterogeneity

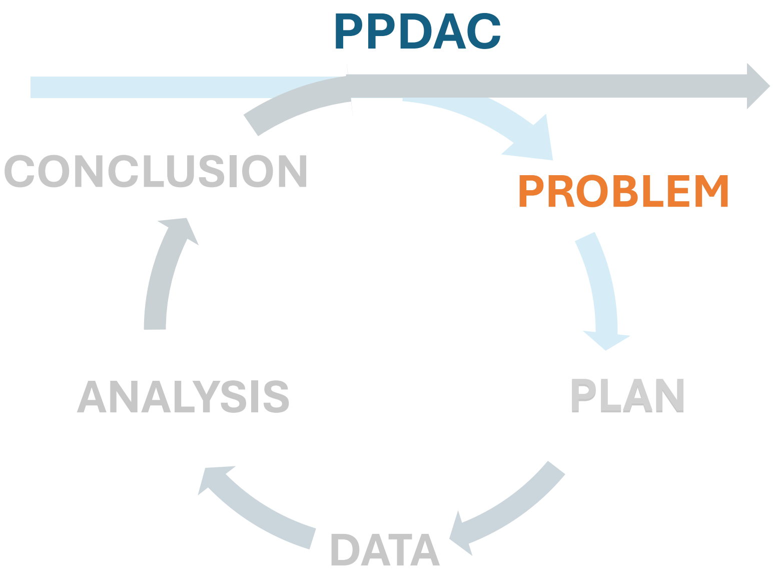
Really interesting. If I understand correctly, first we want to demonstrate that image features correlate with patient prognosis. If there is evidence of this, we could apply these learnings in a new study?

Good questions. Lets meet and discuss our options so we can plan effectively.

Summary

- 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)
- Ask probing questions for understanding context and elicit key information
- Asking effective questions is a key competency during the problem phase
- Next step – we will outline a framework to help formulate which questions to ask to define “answerable” questions





Problem

-scoping

-statement of work

-clear thinking, transparency, reproducibility & knowledge transfer

Which information do we need to clarify?

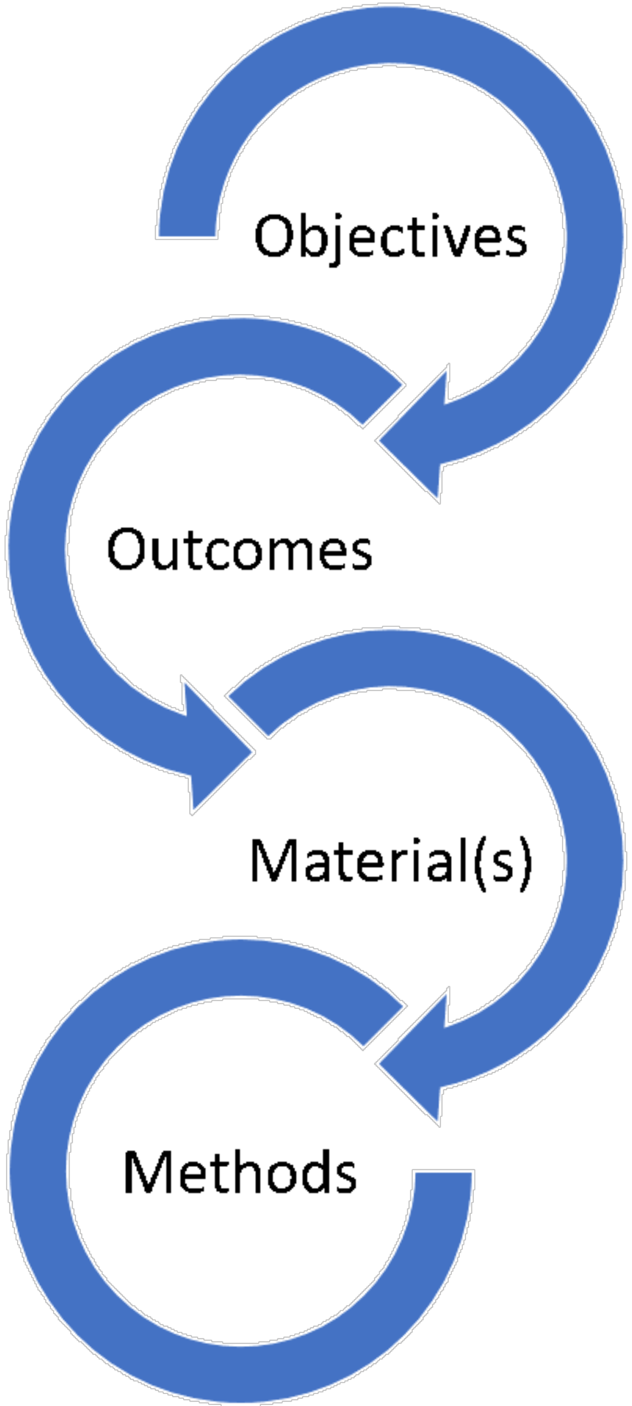
Last section we highlighted the inportance of problem formulation and how “great questions” can help with understanding the context.

The aim was to formulate a problem clearly

- But what questions to ask?
- What information is important to find out?

Now we provide a structured framework for obtaining this key information. This is often called the process of “scoping”

We also provide a statement of work (SoW) template to help collect this information systematically



The iterative steps of scoping

1. Objectives(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)).

Type of question (the zeroth problem)

Descriptive

Provide insight into the past and answer: ‘What has happened?’

- How many females under the age of 50 were enrolled into the study?

Predictive

Understand the future and answer: ‘What could happen?’

- What is the probability of having a cardiovascular event in the next 3 years for females with specific characteristics?

Prescriptive

Describe causal relationships and answer: ‘What should we do?’

- Will prescribing Super Drug X reduce the risk of a cardiovascular event, on average, compared to staying on the current standard of care?

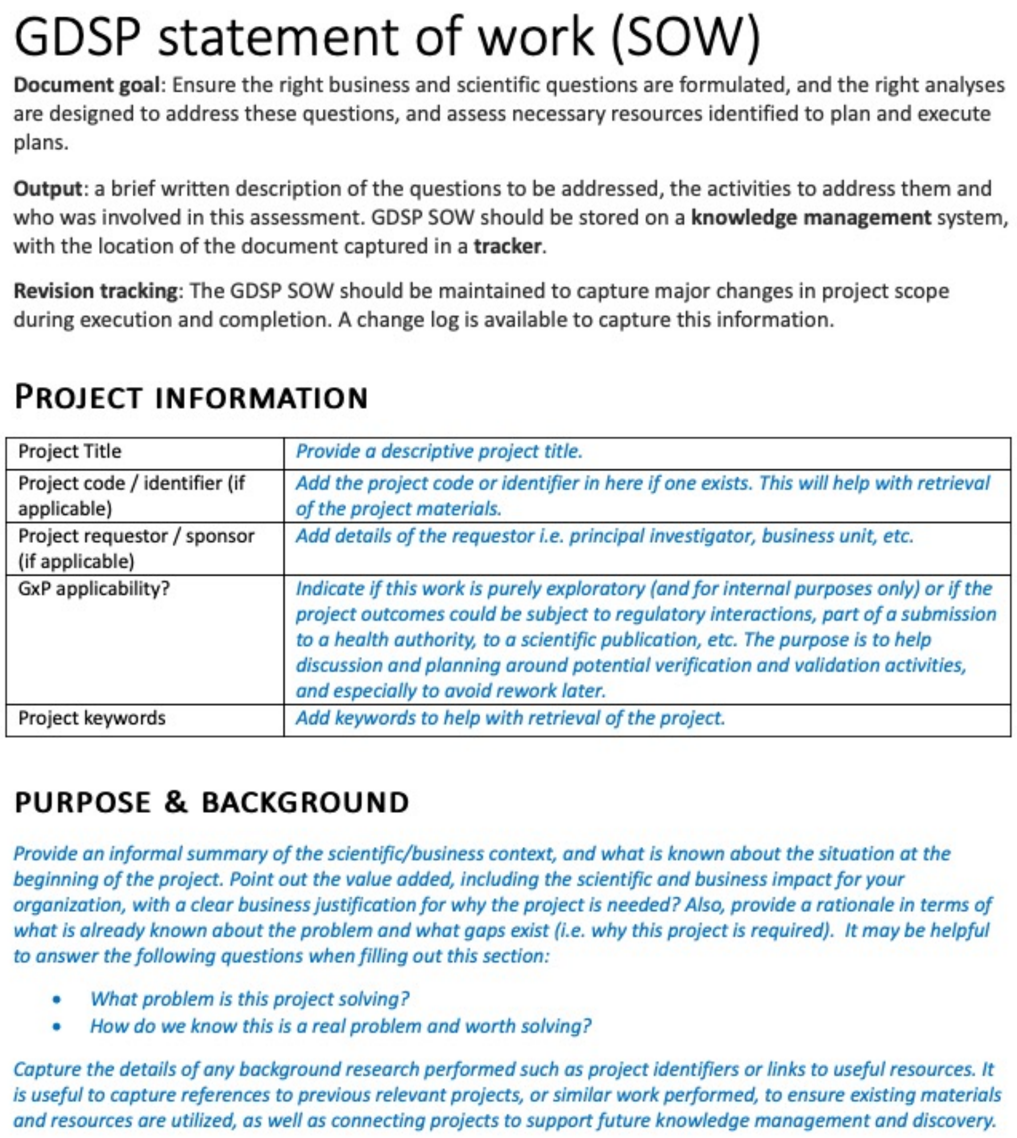
Hernan et al. (2019), Mallows (1998), Shmueli (2010)

Enhancing Impact with a Statement of Work

- **SoW Objectives:**
  - Ensure alignment of business and scientific inquiries with actionable plans
  - Facilitate knowledge transfer and enhance project oversight
- **Document essentials:**
  - Project purpose, background, and detailed objectives.
  - Expected outcomes and methodologies.
  - Governance structure, team roles, and contribution details.
  - Defined timelines with critical milestones.
- **Knowledge Management:**
  - Capture and store project details for easy retrieval.
  - Enable efficient knowledge sharing and continuity.
- **Benefits:**
  - Provides a framework for reproducibility and systematic execution
  - Accessible SoW allows for comprehensive project tracking and transparency

SoW – available online

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	*Provide a descriptive project title.*
Project code / identifier (if applicable)	*Add the project code or identifier in here if one exists. This will help with retrieval of the project materials.*
Project requestor / sponsor (if applicable)	*Add details of the requestor i.e. principal investigator, business unit, etc.*
GxP applicability?	*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.*
Project keywords	*Add keywords to help with retrieval of the project.*

Purpose & Background

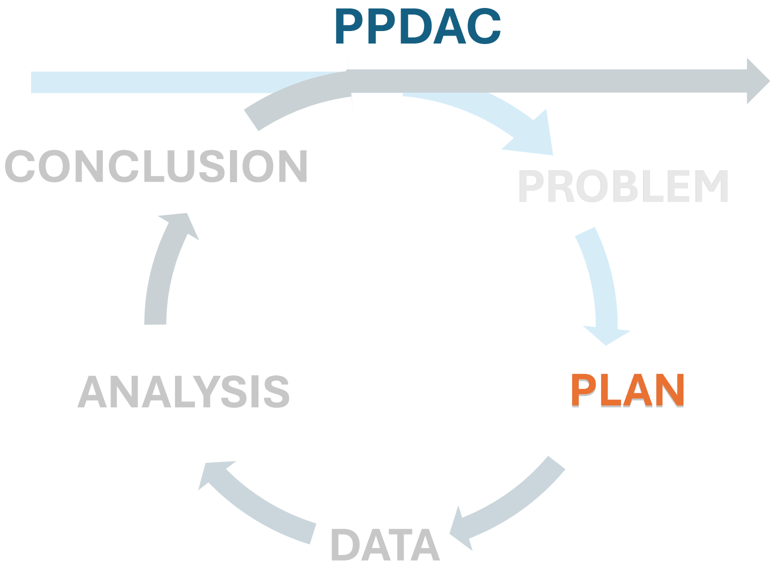
Provide an informal summary of the scientific/business context, and what is known about the situation at the beginning of the project. Point out the value added, including the scientific and business impact for your organization, with a clear business justification for why the project is needed? Also, 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.

Summary, scoping & statement of work

- **Scoping:**
  - Provides a structured approach for eliciting crucial problem-formulating information.
  - What are the areas to focus on for information gathering
- **Iterative Process:** Encourages regular refinement and documentation for clarity and direction.
- **Documentation benefits:**
  - Enhances reproducibility and transparency within the team.
  - Assists in identifying and mitigating assumptions and biases.
- **Best Practices:** considered essential and ethical across various sectors for robust project management.



Plan

- getting the question(s) right

Plan: getting the questions right

- Moving from **problem** formulation into analysis strategy – how to solve problems
- The **problem** phase helps with framing the questions and defining the broader problem
- Why is this important? An ‘open-ended’ or “vague” question can be a reason for a failed project
- Getting the questions right has a domino effect on the subsequent steps, such as **data** strategy, **analysis** strategy, **conclusions** & communication
- In practice, it will involve an **iterative** process
  - Do not let the **data** strategy or **analysis** strategy dictate the question!
  - Data, question, analysis all tightly linked and influence each other
  - Questions that can only be answered if data can be collected
  - Analyses may have to adjust for biases in the data

Principles for formulating a question

Clarify

Ask focused questions in order to clarify the question sufficiently? Develop a strategy to obtain the necassry information (population, comparison, etc.)

Answerable

Need to have an answerable question by applying an iterative, hierarchical approach to start from a high-level question and end with a reasonably granular, answerable question

Question before solution

Distinction between descriptive / predictive / prescriptive questions; recognize that the same data can answer different questions

Factorial principle

May have multiple questions; need to prioritize and **document** them

Thought experiment[1]

Assume we are interested in solving an immediate **health crisis in 1872**:

Does exposure to water have an effect on risk of death among London residents?

Or, for brevity, **‘Does water kill?’**

Is this question precise enough?

Source: [1]Hernan (2016): https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5207342/pdf/nihms836995.pdf

Version #1

Does water kill?

Do we mean:

- Immersion (death by drowning)
- Flash flood (death by trauma), or
- Drinking water (death by poisoning)?

Version #2

Does drinking water kill?

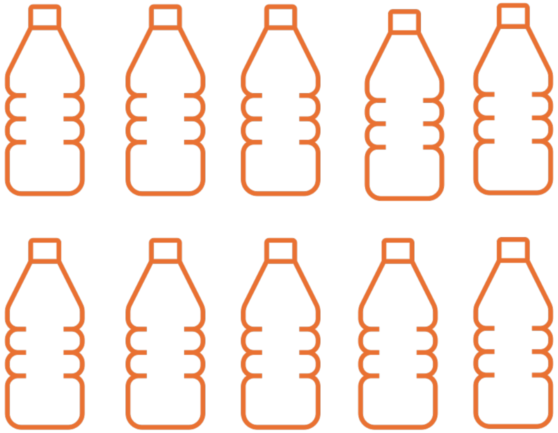
We mean fresh water, not salty water.

Version #3

- **Does drinking fresh water kill?**

How much water?

10 liters per day will kill you...



Version #4

- **Does drinking a large sip of fresh water kill?**

What is the source of the water?

Tap, fountain, directly from the river…

Version #5

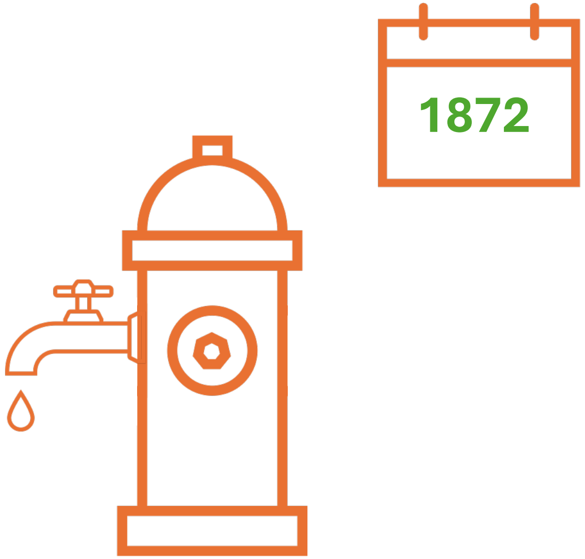
- **Does drinking a large sip of fresh water from the**
- **Broad Street pump kill?**
-

  Are we talking under **1871** conditions, or **1872**
  when drinking water regulations looked very different?

Version #6

- **Does drinking a large sip of fresh water from the Broad Street pump in 1872 kill?**

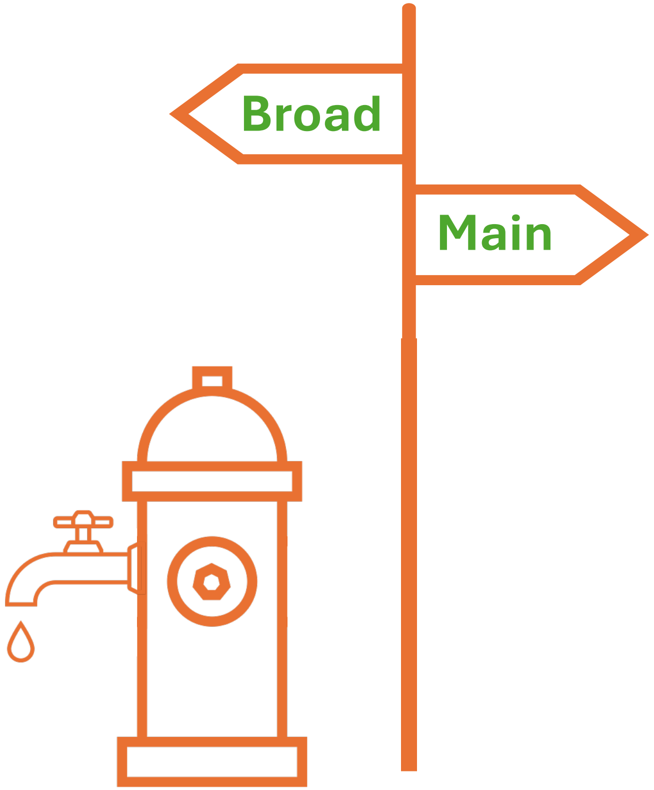
- Ok, but compared with what? With drinking 3 liters of beer?



Version #7

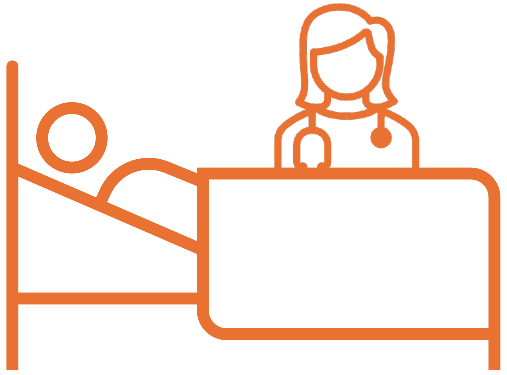
- **Does drinking a large sip of fresh water from the**
- **Broad Street pump in 1872, compared to drinking all your water from the Main Street pump, kill?**

- What about other concomitant factors that may affect your outcome of interest?



Version #8

- **Does drinking a large sip of fresh water from the Broad Street pump in 1872 and not initiating a rehydration treatment if diarrhea occurs, compared to drinking all your water from the Main Street pump, kill?**



- And so on ...

Version #1’000’000

- **[*very long text goes here*]**
- Is Version #1’000’000 precise enough?
  - ✓ No version of the question is perfectly well-defined
  - ✓ Fortunately, absolute precision is not necessary
  - ✓ Further specification is required only while meaningful vagueness remains
    - For example, we may not believe the speed of the water from the pump would lead to different conclusions ➔ no need to specify the water speed

Summary 1/2

The original question was hopelessly vague because there are many possible versions of 'water' and 'kill’

Certain degree of vagueness remains despite our efforts to refine the question. However, we can start to design how to answer version #8 than with versions #1 and #2

Many versions of ‘water’ are irrelevant (e.g., the speed at which water flowed) and therefore do not bother to specify them

Summary 2/2

Likewise, the question

‘Does obesity kill?’

is hopelessly vague as well because there are many possible versions of ‘obesity' and 'kill.’

Is more **appropriate question**:

Would the 20-year death risk differ if the human population of interest were assigned to gaining weight **by force feeding** 5000 vs. 2500 calories per day between ages 30 and 50?

Where are we?

What we have learned so far:

Questions are often too vague and need to be refined iteratively to make them precise enough.

What comes next:

Recognize that

…different types of questions are possible

Type of question (revisited)

Descriptive

Provide insight into the past and answer: ‘What has happened?’

- How many females under the age of 50 were enrolled into the study?

Predictive

Understand the future and answer: ‘What could happen?’

- What is the probability of having a cardiovascular event in the next 3 years for females with specific characteristics?

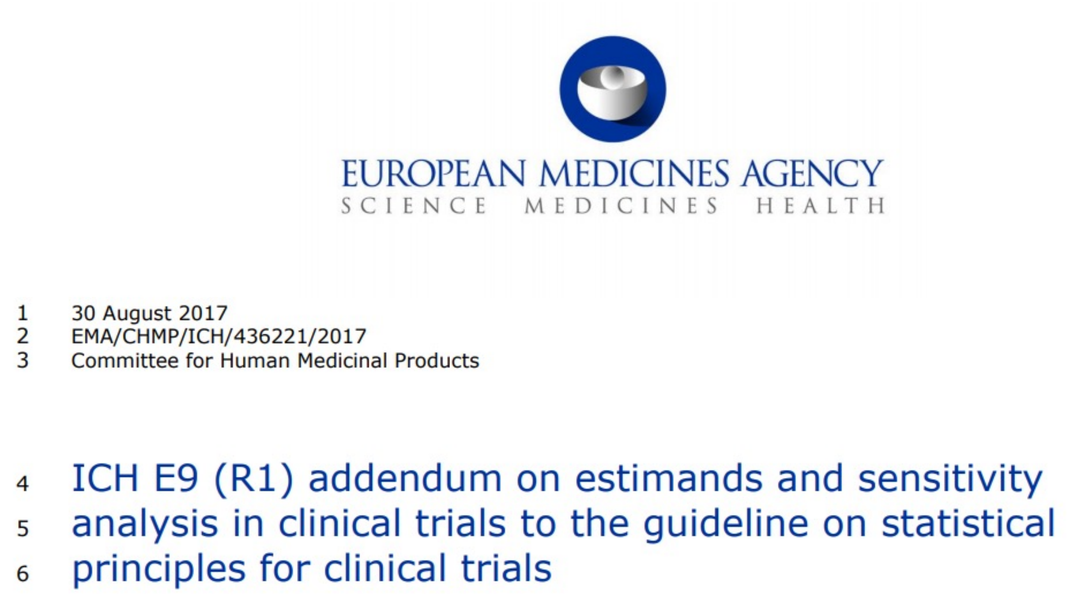
Prescriptive

Describe causal relationships and answer: ‘What should we do?’

- Will prescribing Super Drug X reduce the risk of a cardiovascular event, on average, compared to staying on the current standard of care?

Hernan et al. (2019), Mallows (1998), Shmueli (2010)

Example thinking frameworks (estimands)



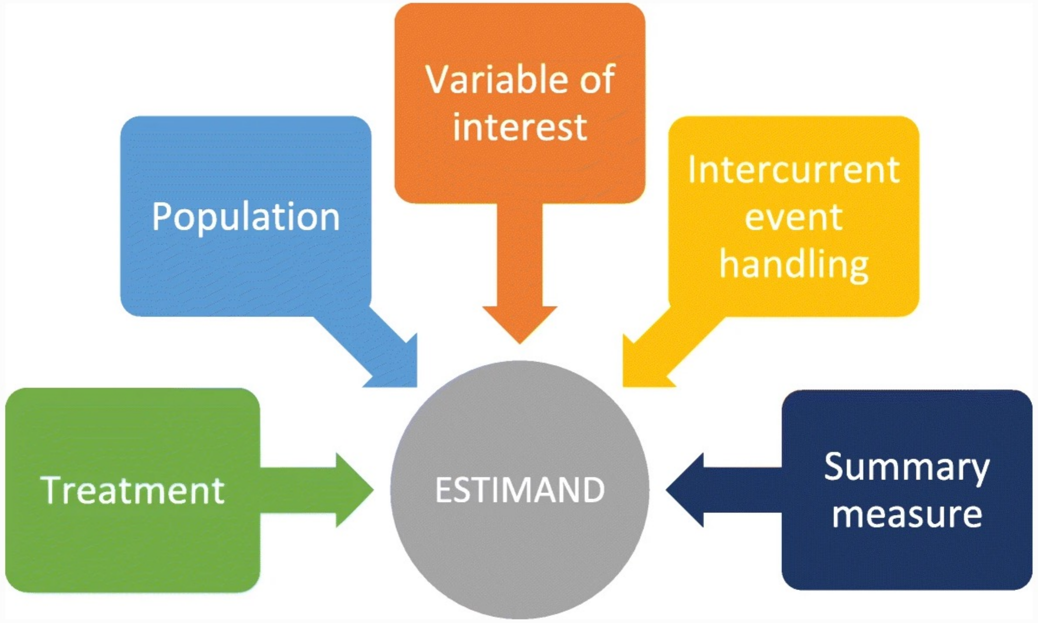
EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

30 August 2017
EMA/CHMP/ICH/436221/2017
Committee for Human Medicinal Products

ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials

ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials - Step 2b (europa.eu)

From: What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials?

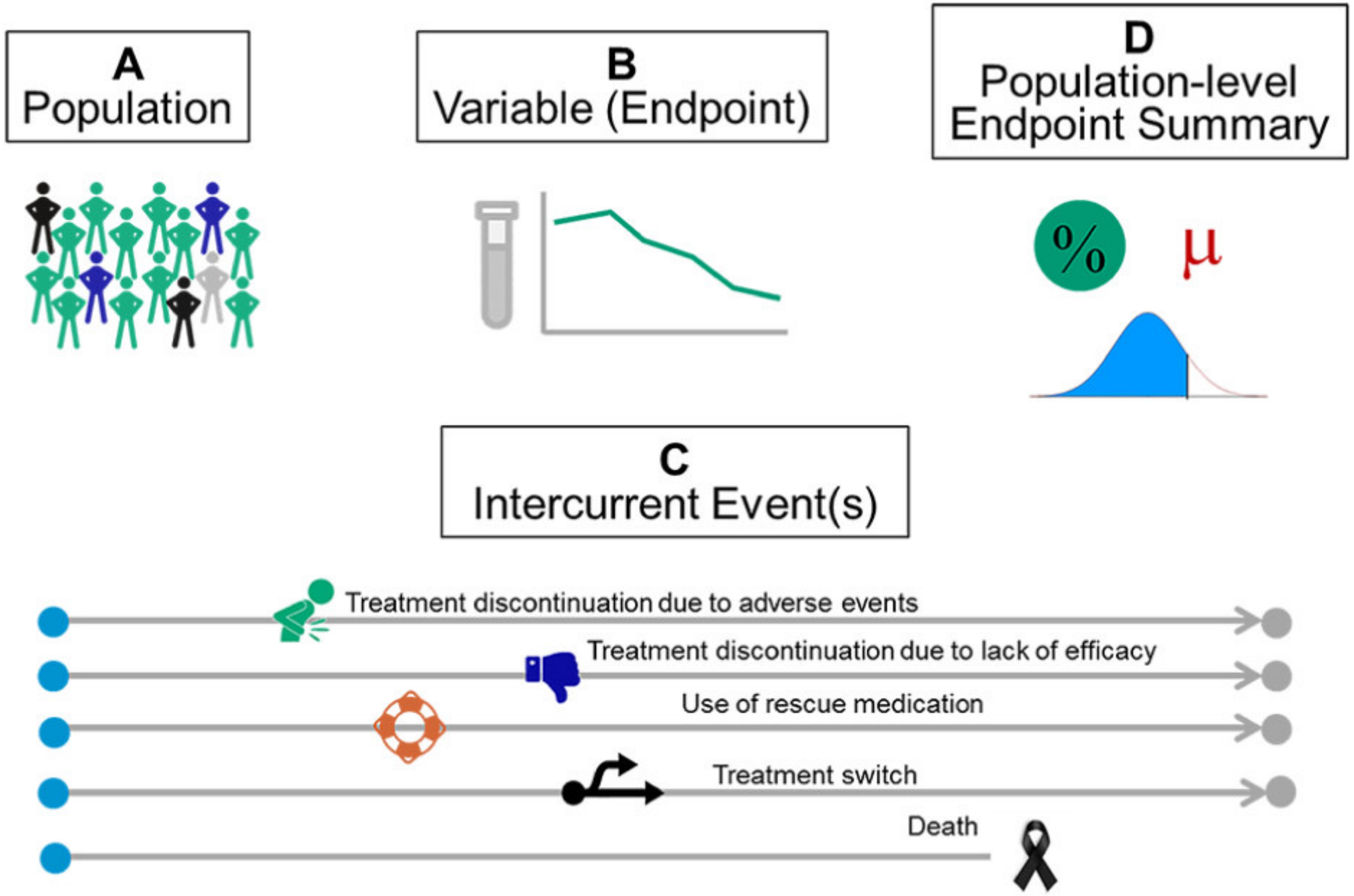


The five attributes of an estimand according to the ICH E9 (R1) addendum

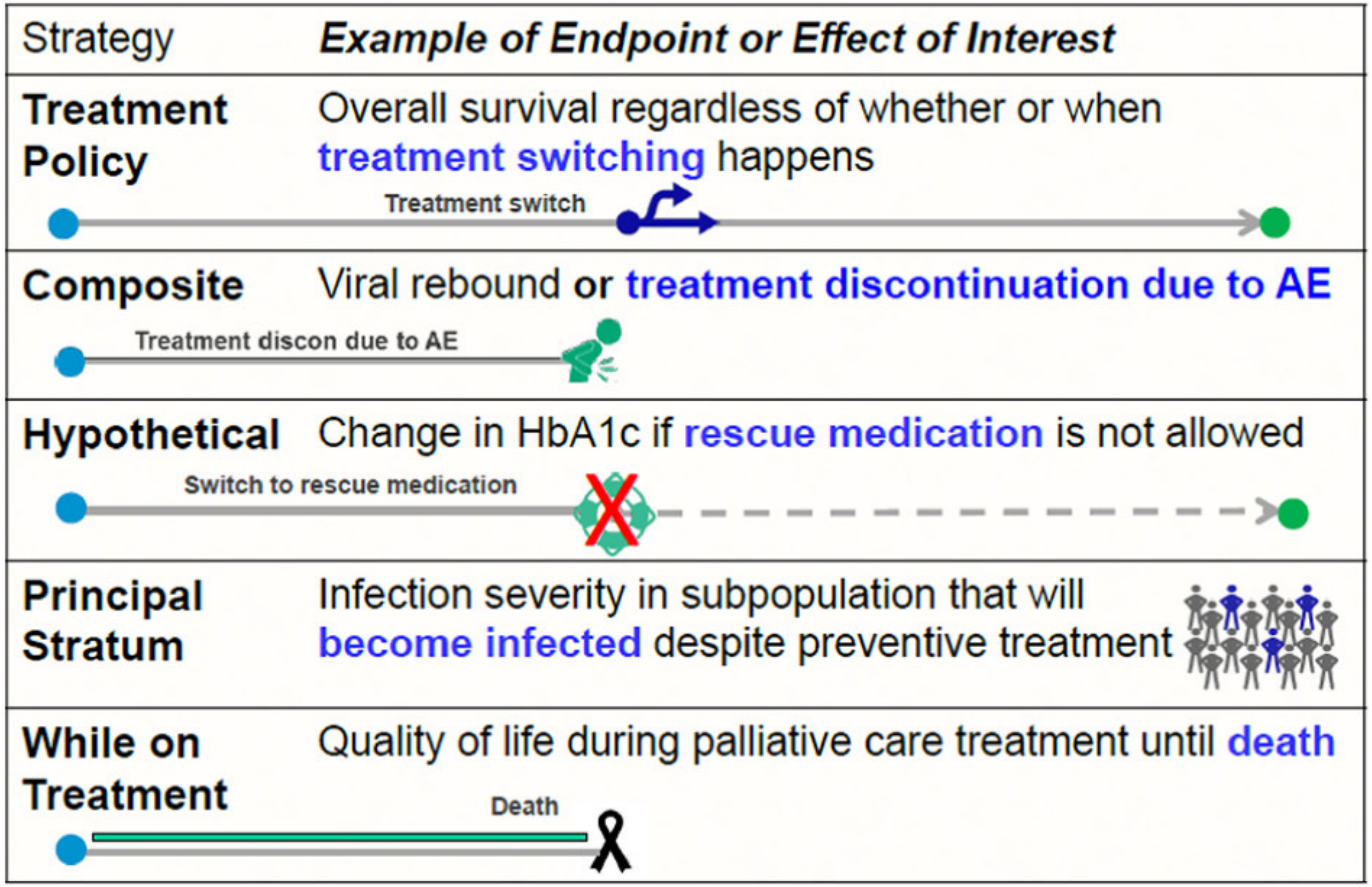
Lawrance, R., Degtyarev, E., Griffiths, P. et al.

What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials?. J Patient Rep Outcomes 4, 68 (2020). https://doi.org/10.1186/s41687-020-00218-5

Attributes for formulating the estimand



Strategies for addressing intercurrent events in the scientific question of interest



Strategy	*Example of Endpoint or Effect of Interest*
Treatment Policy	Overall survival regardless of whether or when **treatment switching** happens Treatment switch
Composite	Viral rebound **or treatment discontinuation due to AE** Treatment discon due to AE
Hypothetical	Change in HbA1c if **rescue medication** is not allowed Switch to rescue medication
Principal Stratum	Infection severity in subpopulation that will **become infected** despite preventive treatment
While on Treatment	Quality of life during palliative care treatment until **death** Death

Example thinking frameworks: The PICO question statement

Question Type	Patient or Problem	Intervention or Exposure	Outcome	Comparison
Therapy	In patients with hypertension and at least one additional cardiovascular disease risk factor	Does tight systolic blood pressure control	Lead to lower rates of myocardial infarction, stroke, heart failure, and cardiovascular mortality	Compared to conservative control?
Diagnosis	Among asymptomatic adults at low risk of colon cancer	Is fecal immunochemical testing (FIT)	As sensitive and specific for diagnosing colon cancer	As colonoscopy?
Prognosis	Among adults with pneumonia	Do those with chronic kidney disease (CKD)	Have a higher mortality rate	Than those without CKD?
Etiology or Harm	Are women	With a history of pelvic inflammatory disease (PID)	At higher risk for gynecological cancers	Than women with no history of PID?
Prevention	Among adults with a history of myocardial infarction	Does adherence to a mediterranean diet	Lower risk of a second myocardial infarction	Compared to those who do not adopt a mediterranean diet?

Target trial emulation

https://academic.oup.com/aje/article/183/8/758/1739860

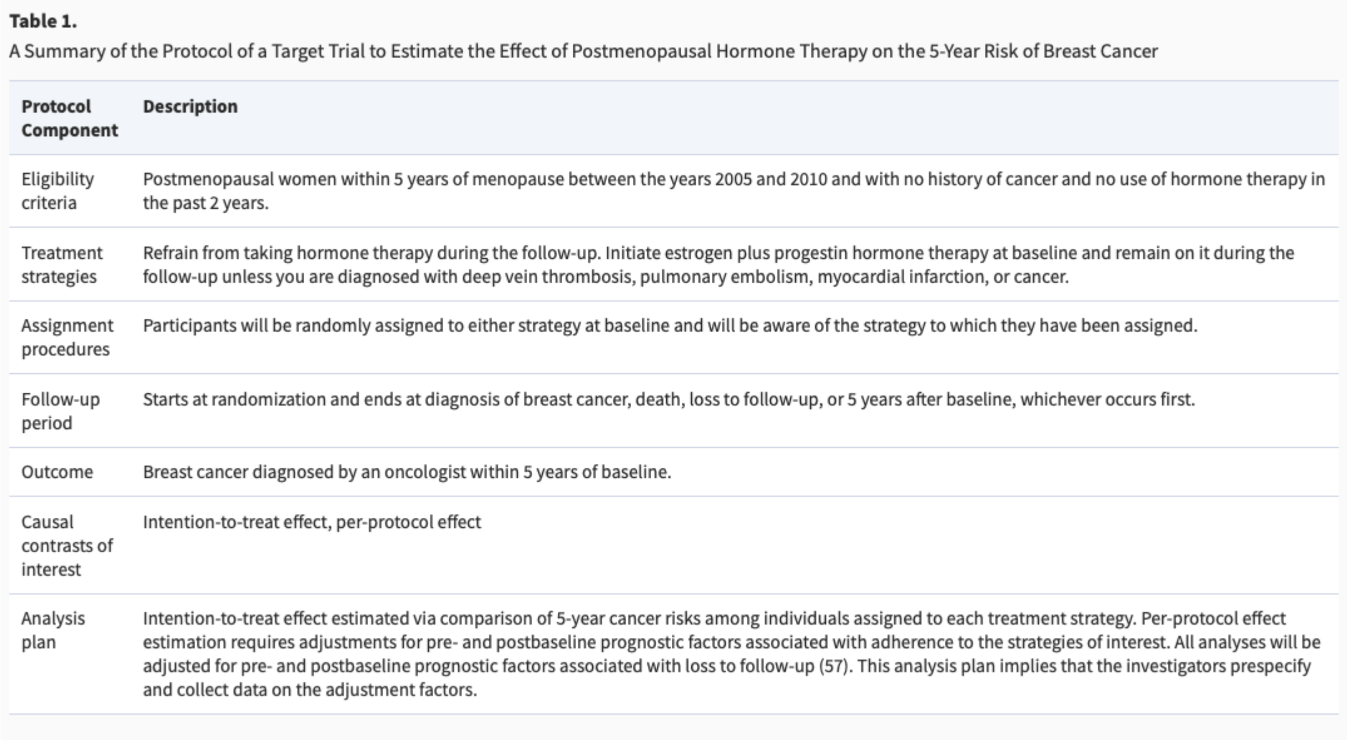
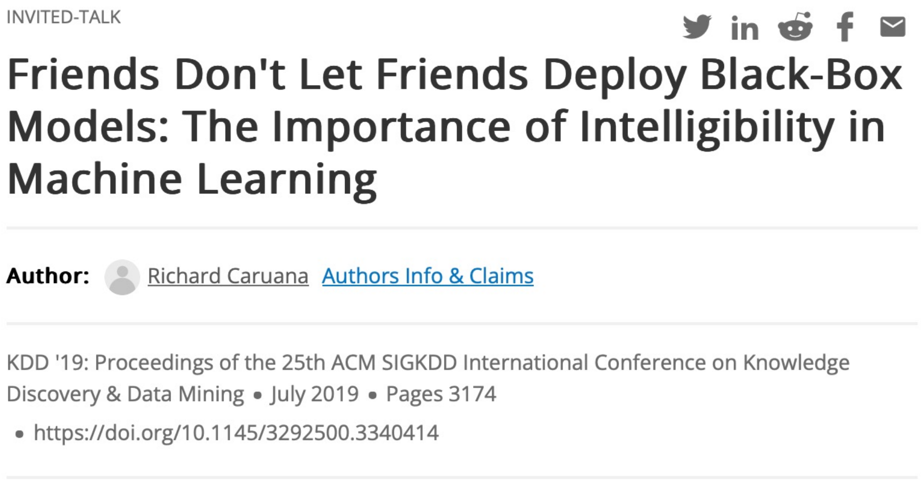


Table 1.

A Summary of the Protocol of a Target Trial to Estimate the Effect of Postmenopausal Hormone Therapy on the 5-Year Risk of Breast Cancer

Protocol Component	**Description**
Eligibility criteria	Postmenopausal women within 5 years of menopause between the years 2005 and 2010 and with no history of cancer and no use of hormone therapy in the past 2 years.
Treatment strategies	Refrain from taking hormone therapy during the follow-up. Initiate estrogen plus progestin hormone therapy at baseline and remain on it during the follow-up unless you are diagnosed with deep vein thrombosis, pulmonary embolism, myocardial infarction, or cancer.
Assignment procedures	Participants will be randomly assigned to either strategy at baseline and will be aware of the strategy to which they have been assigned.
Follow-up period	Starts at randomization and ends at diagnosis of breast cancer, death, loss to follow-up, or 5 years after baseline, whichever occurs first.
Outcome	Breast cancer diagnosed by an oncologist within 5 years of baseline.
Causal contrasts of interest	Intention-to-treat effect, per-protocol effect
Analysis plan	Intention-to-treat effect estimated via comparison of 5-year cancer risks among individuals assigned to each treatment strategy. Per-protocol effect estimation requires adjustments for pre- and postbaseline prognostic factors associated with adherence to the strategies of interest. All analyses will be adjusted for pre- and postbaseline prognostic factors associated with loss to follow-up (57). This analysis plan implies that the investigators prespecify and collect data on the adjustment factors.

Prediction models



INVITED-TALK

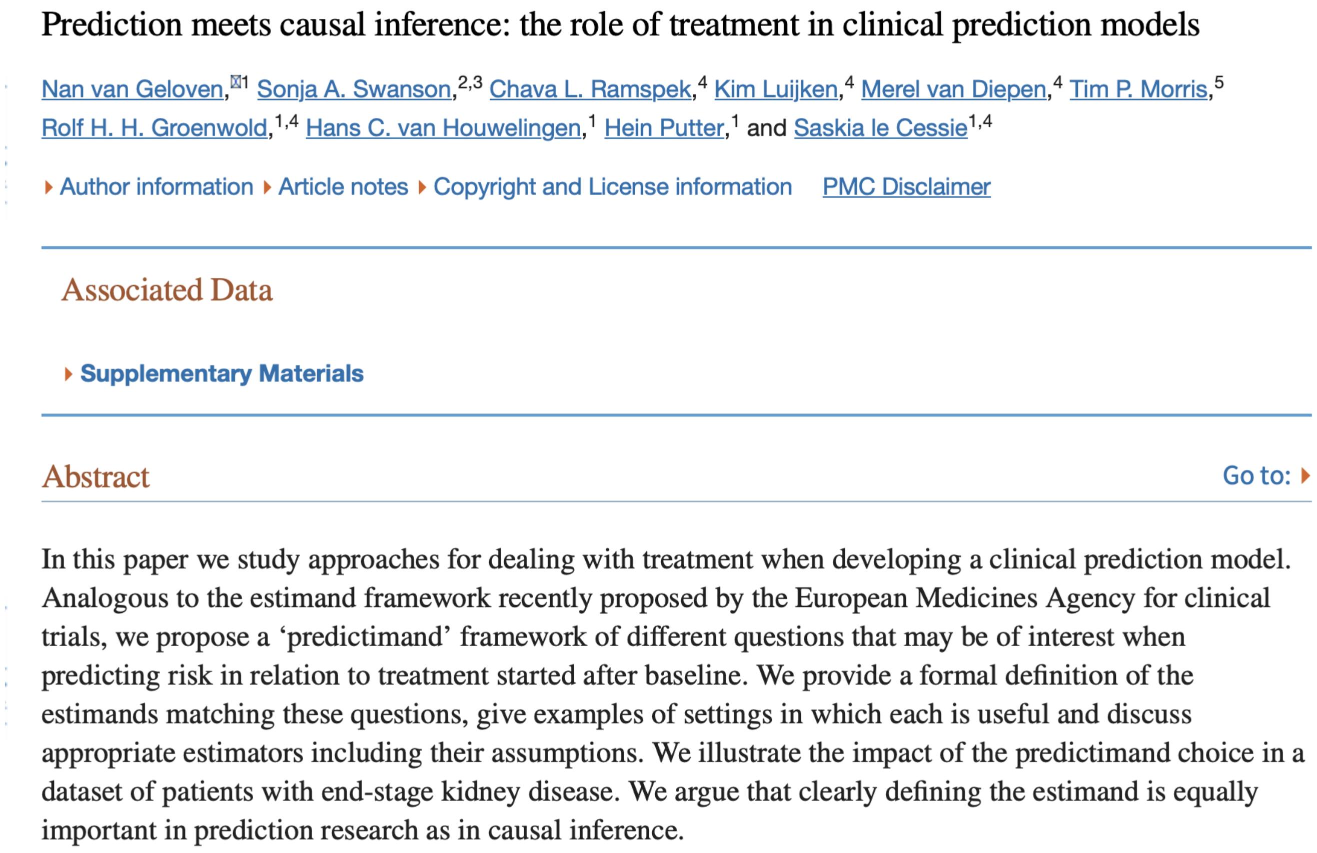
Friends Don't Let Friends Deploy Black-Box Models: The Importance of Intelligibility in Machine Learning

Author: Richard Caruana Authors Info & Claims

KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining • July 2019 • Pages 3174

- https://doi.org/10.1145/3292500.3340414

https://dl.acm.org/doi/10.1145/3292500.3340414



Prediction meets causal inference: the role of treatment in clinical prediction models

Nan van Geloven,[1] Sonja A. Swanson,[2,3] Chava L. Ramspek,[4] Kim Luijken,[4] Merel van Diepen,[4] Tim P. Morris,[5] Rolf H. H. Groenwold,[1,4] Hans C. van Houwelingen,[1] Hein Putter,[1] and Saskia le Cessie[1,4]

▸ Author information ▸ Article notes ▸ Copyright and License information PMC Disclaimer

Associated Data

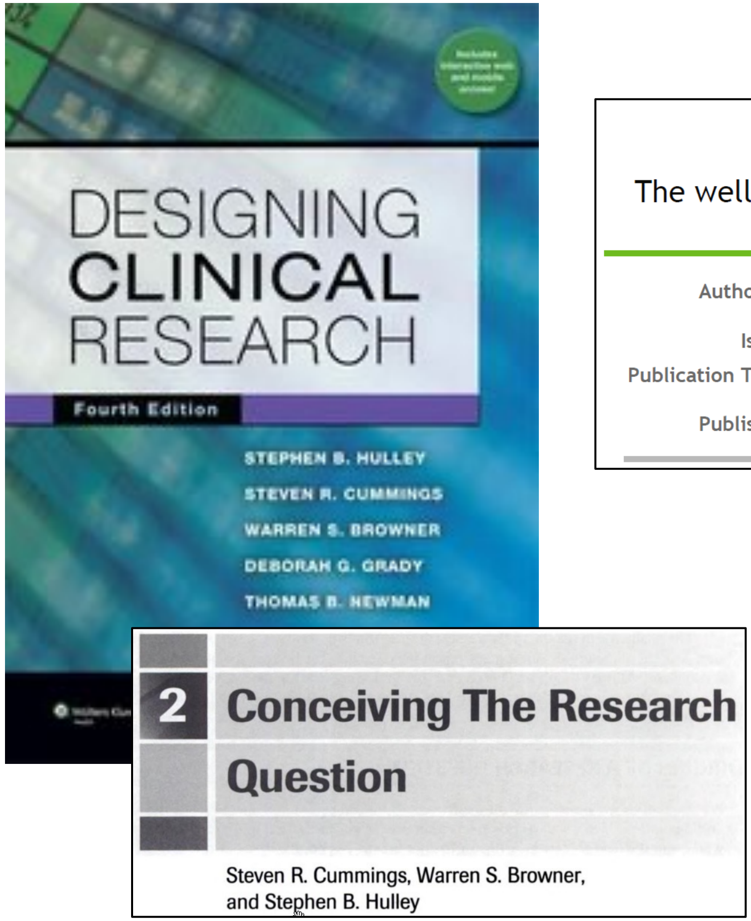
▸ **Supplementary Materials**

Abstract

In this paper we study approaches for dealing with treatment when developing a clinical prediction model. Analogous to the estimand framework recently proposed by the European Medicines Agency for clinical trials, we propose a ‘predictimand’ framework of different questions that may be of interest when predicting risk in relation to treatment started after baseline. We provide a formal definition of the estimands matching these questions, give examples of settings in which each is useful and discuss appropriate estimators including their assumptions. We illustrate the impact of the predictimand choice in a dataset of patients with end-stage kidney disease. We argue that clearly defining the estimand is equally important in prediction research as in causal inference.

https://pubmed.ncbi.nlm.nih.gov/32445007/

Many existing frameworks



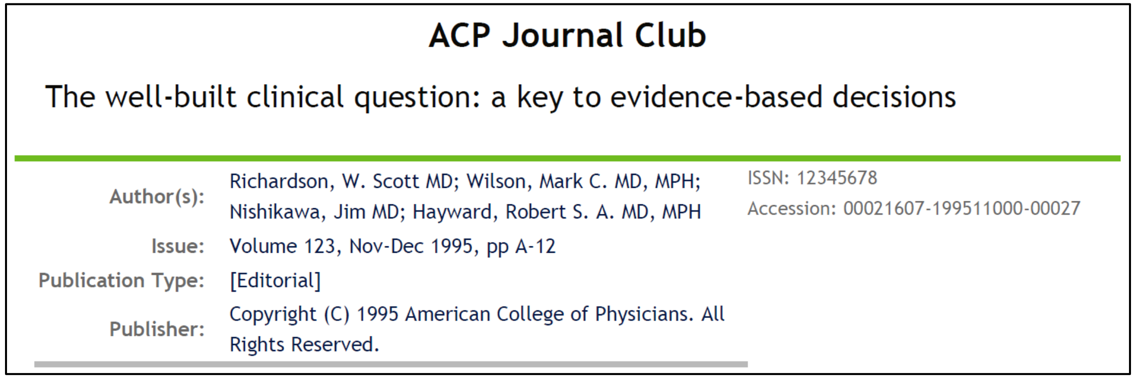
DESIGNING CLINICAL RESEARCH

Fourth Edition

STEPHEN B. HULLEY
STEVEN R. CUMMINGS
WARREN S. BROWNER
DEBORAH G. GRADY
THOMAS B. NEWMAN

2 Conceiving The Research Question

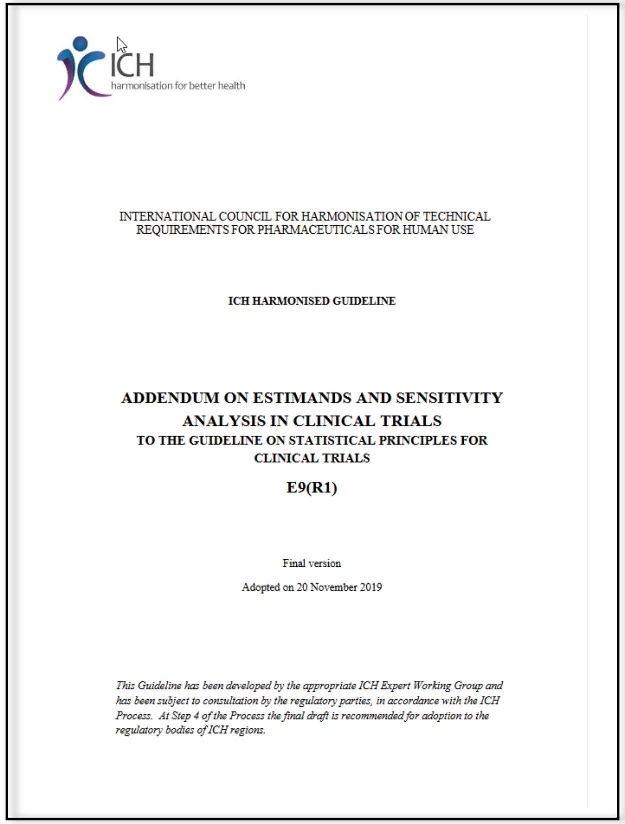
Steven R. Cummings, Warren S. Browner,
and Stephen B. Hulley



ACP Journal Club

The well-built clinical question: a key to evidence-based decisions

Author(s):	Richardson, W. Scott MD; Wilson, Mark C. MD, MPH; Nishikawa, Jim MD; Hayward, Robert S. A. MD, MPH	ISSN: 12345678 Accession: 00021607-199511000-00027
Issue:	Volume 123, Nov-Dec 1995, pp A-12	
Publication Type:	[Editorial]	
Publisher:	Copyright (C) 1995 American College of Physicians. All Rights Reserved.	



ICH
harmonisation for better health

INTERNATIONAL COUNCIL FOR HARMONISATION OF TECHNICAL
REQUIREMENTS FOR PHARMACEUTICALS FOR HUMAN USE

ICH HARMONISED GUIDELINE

ADDENDUM ON ESTIMANDS AND SENSITIVITY
ANALYSIS IN CLINICAL TRIALS
TO THE GUIDELINE ON STATISTICAL PRINCIPLES FOR
CLINICAL TRIALS

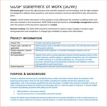
E9(R1)

Final version

Adopted on 20 November 2019

This Guideline has been developed by the appropriate ICH Expert Working Group and has been subject to consultation by the regulatory parties, in accordance with the ICH Process. At Step 4 of the Process the final draft is recommended for adoption to the regulatory bodies of ICH regions.

Principles for formulating a question



GDSP statement of work (SOW)

PROJECT INFORMATION

PURPOSE & BACKGROUND

Clarify

Ask focused questions in order to clarify the question sufficiently? Develop a strategy to obtain the necassry information (population, comparison, etc.)

Answerable

Need to have an answerable question by applying an iterative, hierarchical approach to start from a high-level question and end with a reasonably granular, answerable question

Question before solution

Distinction between descriptive / predictive / prescriptive questions; recognize that the same data can answer different questions

Factorial principle

May have multiple questions; need to prioritize and document them

Plan summary

The primary importance of framing the questions is that it narrows down a real problem of interest into a specific answerable task

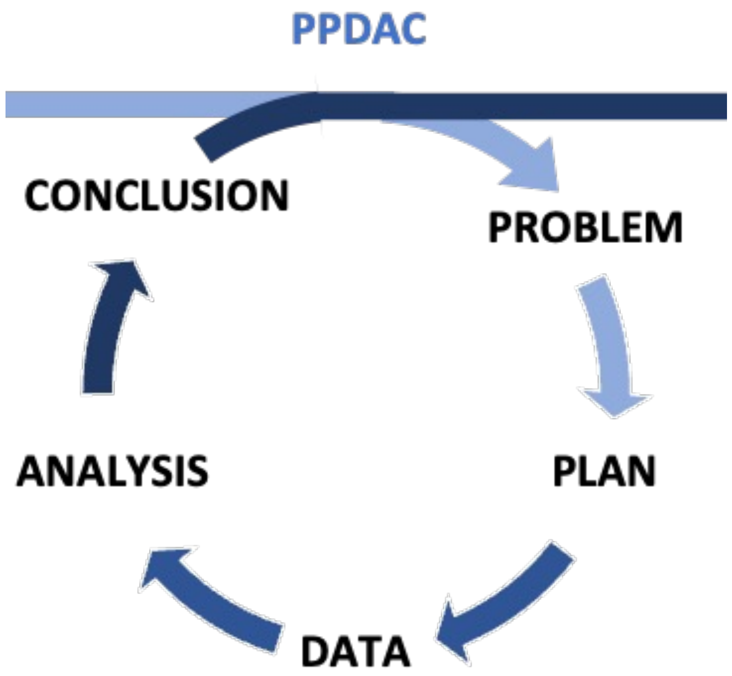
- Framing the questions helps defining the broader problem
- Getting the questions right has a domino effect on the subsequent steps highlighted in the workflow
- A vague ‘open-ended’ question can be a reason for a failed project
- For many problems, it is wise to ask a series of interrelated questions rather than a single question

Data science thinking process

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.

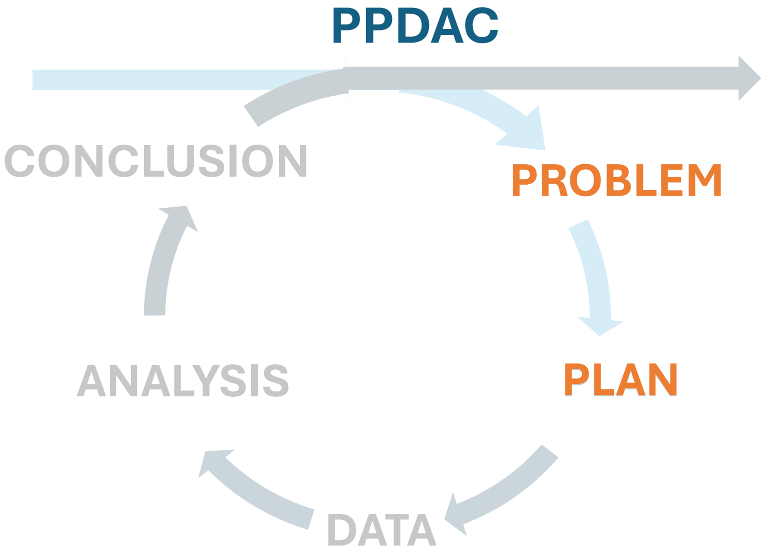
This workflow demonstrates the steps for abstracting and solving **a real problem**. An impactful solution requires a clear understanding of how things work.



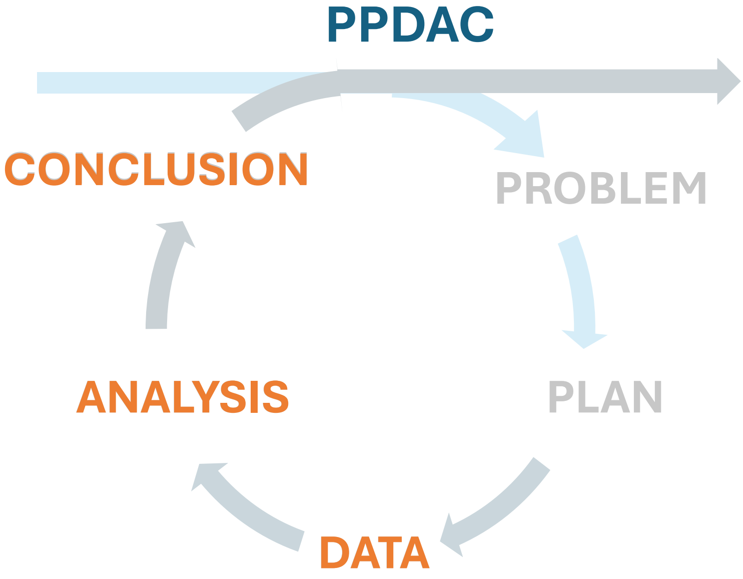
Source: MacKay, R.J. and Oldford, R.W., 2000. Scientific method, statistical method and the speed of light. *Statistical Science*, pp.254-278.

Data science thinking: making an impact

Part I



Part II



References & Further Reading Material

References:

Cox (2017) Statistical Science: A Grammar for Research. European Journal of Epidemiology 32: 465–71. https://doi.org/10.1007/s10654-017-0288-1.

Cummings et al. (2007) Conceiving The Research Question In: Hulley et al (editors) Designing clinical research (link)

International Council for Harmonisation (2019) ICH E9(R1): Estimands and sensitivity analysis in clinical trials. (link)

Richardson et al. (1995). The well-built clinical question: A key to evidence-based decisions. ACP Journal Club, 123, A12-13. https://doi.org/10.7326/ACPJC-1995-123-3-A12

Vance et al. (2022) Asking great questions. Stat, 11(1), p.e471.

Kenett, R.S. and Redman, T.C., 2019. *The real work of data science: turning data into information, better decisions, and stronger organizations*. John Wiley & Sons.

Hernán, M.A., Hsu, J. and Healy, B., 2019. A second chance to get causal inference right: a classification of data science tasks. *Chance*, *32*(1), pp.42-49

Shmueli, Galit. "To explain or to predict?." (2010): 289-310.

Wild, Chris J, and Maxine Pfannkuch. 1999. "Statistical Thinking in Empirical Enquiry. "*International Statistical Review* 67 (3): 223–48.

MacKay, R Jock, and R Wayne Oldford. 2000. "Scientific Method, Statistical Method and the Speed of Light." *Statistical Science*, 254–78.

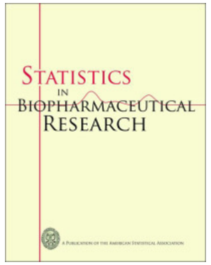
Further reading material:

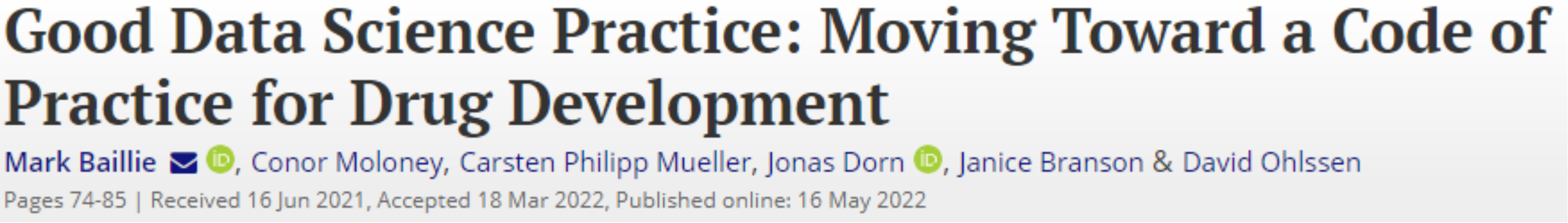
Bouchrika (2023) How to write a research question: Types, steps, and examples. https://research.com/research/how-to-write-a-research-question

Haynes (2006) Forming research questions. Journal of Clinical Epidemiology 59: 881–86. https://doi.org/10.1016/j.jclinepi.2006.06.006

Mallows (1998) The Zeroth Problem. The American Statistician 52: 1–9. https://doi.org/10.1080/00031305.1998.10480528.

Build upon these works ...

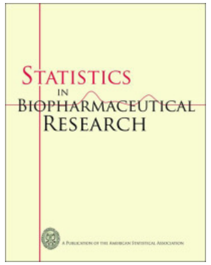


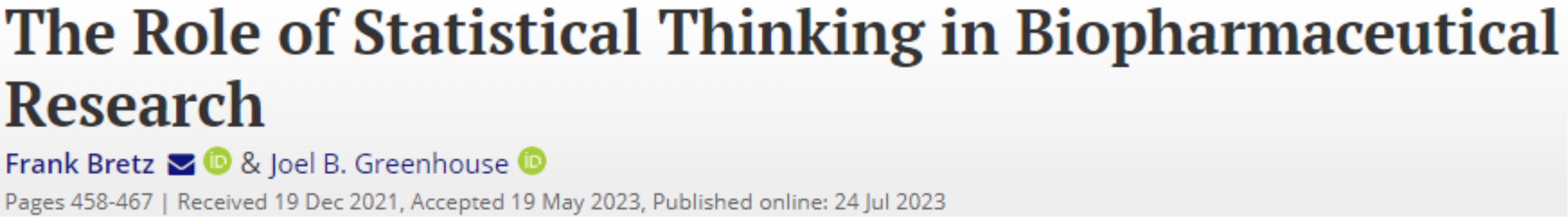


Good Data Science Practice: Moving Toward a Code of Practice for Drug Development

Mark Baillie, Conor Moloney, Carsten Philipp Mueller, Jonas Dorn, Janice Branson & David Ohlssen

Pages 74-85 | Received 16 Jun 2021, Accepted 18 Mar 2022, Published online: 16 May 2022

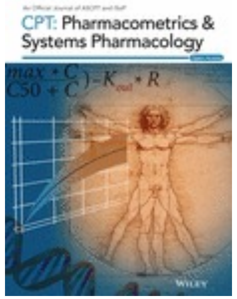


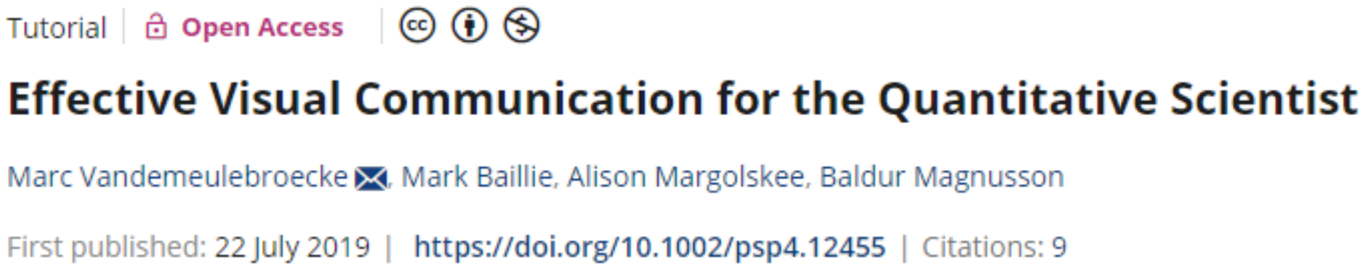


The Role of Statistical Thinking in Biopharmaceutical Research

Frank Bretz & Joel B. Greenhouse

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





Tutorial | Open Access | 🅭 🅯 🄏

Effective Visual Communication for the Quantitative Scientist

Marc Vandemeulebroecke, Mark Baillie, Alison Margolskee, Baldur Magnusson

First published: 22 July 2019 | https://doi.org/10.1002/psp4.12455 | Citations: 9

Any Questions?

Applied Machine Learning Days

EPFL

AMLD

March 23-26
2024

Lausanne
Switzerland