RESEARCH ON RESEARCH INSTITUTE

# AFIRE Sprint on AI in Grantmaking

Workshop 2 2025-06-10

# Welcome From Tom Stafford

Professor of Cognitive Science
& University Research Practice Lead
University of Sheffield
https://tomstafford.github.io/

Senior Research Fellow,
Research on Research Institute
https://researchonresearch.org/

RoRI RESEARCH ON RESEARCH INSTITUTE

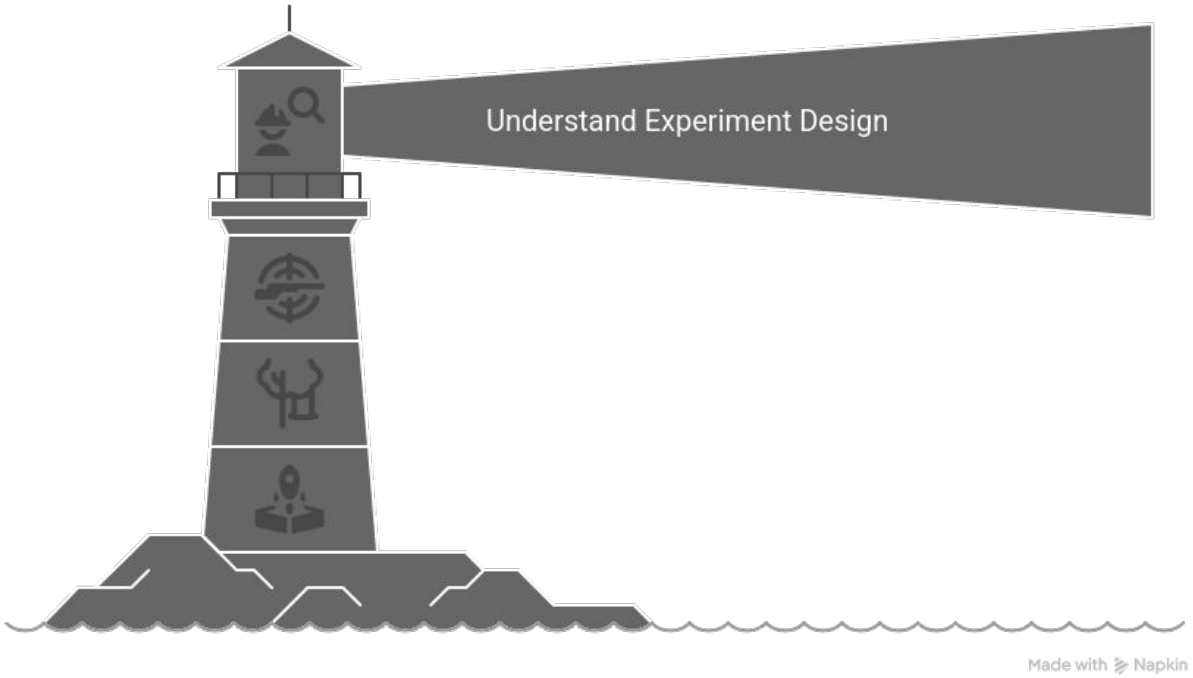
# You: many teams, many missions!

AQuAS
CDTI
CIFAR
CIHR
DSIT /
Metascience Unit
FFG
FNR
FWF
KBF
La Caixa
MRC / UKRI
NSERC
NWO
RCN
Research England
SNSF
SSHRC
Ukrainian Ministry of
Education and Science
Volkswagen Foundation
Wellcome
ZonMw

Introduce yourself in the chat!
+ add your organisation to your zoom name please

# Our Mission!

Together we will wrestle with the particular discipline that experiment design forces: how to plan an investigation that is both focussed enough to provide a clear outcome, but also general enough to provide relevant evidence for future action. The ambition is that everyone will finish the sessions some steps closer to the goal of launching new experiments in the space of evaluating the use of AI in research funding.



# Roadmap

| DATE | FOCUS |
|---|---|
| May 27 | Workshop 1: Why experiment with AI?<br>*Lessons from the GRAIL project, understanding the risks and benefits of experiments* |
| June 3 | - one to one slots : bespoke coaching on developing experiments |
| June 10 | Workshop 2: Examples of experiments with AI<br>*Discussion of case studies, experiment design* |
| June 17 | - one to one slots : bespoke coaching on developing experiments |
| June 24 | Workshop 3: Your experiment with AI<br>*Pitches for new experiments & feedback, advocating for experiments in your organisation* |

# Some logistics

**Email list**: you should be on this, let me know if not. Use for questions

**Recordings**: The talks (only) will be recorded

Discussions are under the Chatham House Rule:

> "anyone who comes to a meeting is free to use information from the discussion, but is not allowed to reveal who made any particular comment."

**Coaching sessions:** Arranged via the lead for your group

**Notes file**: contains all essential info. Replicated at https://tomstafford.github.io/AIsprint/

# Timetable for today

**1400** Introductions, Logistics

**1410** Funder case studies: possible experiments

**1455** (break)

**1500** Basics of experimentation (Amanda Kvarven)

**1525** Breakout groups [PICO exercise]

**1555** END

# Feedback from session 1

Additional sessions: implementation/technical focus

Kind words: thank you

Usefulness of the IF THEN exercise: outcome measures

Breakout groups: today: new groups, future: matching by interest

# Case Studies

1. La Caixa

2. SNSF

3. AQuAS

Remember: presentation main contain speculations and unconfirmed plans - listen with a spirit of generosity

# *Innovative approaches to improve research assessment.*

*AI to pre-select proposals*

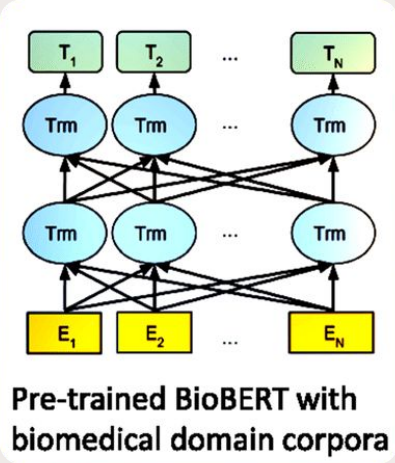
CaixaResearch

# 3 CONTEXT: AI for eligibility: *proposals with low probability of being selected*

**2020-2021**

## *Proof of concept*

- **Several AI models** based on natural language processing were **tested**
- **Data** for training and validation: **3.212 proposals** (HR18-HR22)
- Promising results




**Pre-trained BioBERT with biomedical domain corpora**

**2022**

## *Pilot*

- **546 proposals** (HR22)
- **Pre-rejected** proposals by AI are **evaluated by 2 experts**:
  - Proposals are rejected if both experts confirm it
  - Proposals are saved if at least 1 expert has doubts
- In **HR22** 86 proposals would have been rejected (AI + experts). Only 1 arrived to selection comittees and was funded (from reservation panel).

**2023**

## *Application*

- 493 proposals presented: **430 elegible and 63 non elegible (AI + experts).**
- **Results HR23**

# 3 New experiment: *Could we use AI to detect the best proposals?*

IF we could use **AI to pre-select the best proposals** THEN we could **skip remote evaluation** and move from eligibility to panel interviews, **reducing the evaluation process 4 months.**

## *Main Idea*

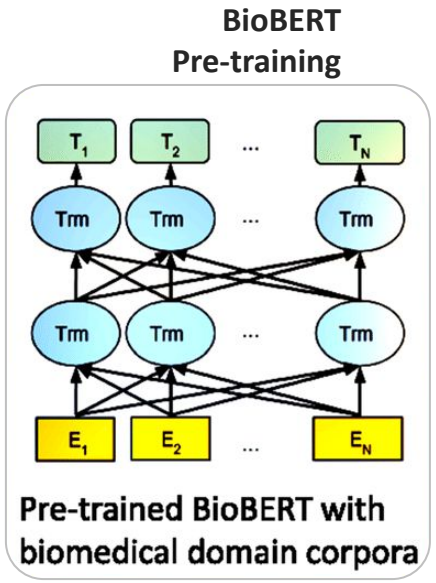
- **Same AI models** based on natural language NOW trained to detect the proposals with HIGH probability of being selected
- Rank this proposals per thematic area
- Pre-select the best X proposals for interview
- Pilot this process in parallel to HR26 ordinary evaluation pathway
- Compare AI results to actual results considering pre-selection and selection

## *For NOW*

- We have checked possible results with old data: HR24 using the AI algorithms trained to detect “bad” proposals but requesting to rank the best.
- Results were promising AI detected 70/90 pre-selected, and within this, 28/30 selected
- We have decided to train the algorithms accordingly and go for the experiment

The tool combines three AI biomedical research models based on natural language processing:

- BioBERT,
- BioELECTRA
- BioBERT with Adapter blocks.



# 3 New experiment: *doubts, limitations or weaknesses*

Our goal is to check if AI is good enough to pre-select proposals that will be finally funded.

/ Outcome expected in the experiment: AI pre-selected proposals MUST include the finally selected projects by the panel reviewers.

/ How can we perform the experiment with solid results and applicable to real the live process.

We see 2 possible interventions for this:

## *Same process*

/ **Strength:** we can do this pilot without extra work for our reviewers.

/ **Limitation:** little room of error for the AI. The AI should be able to detect within the pre-selected by AI, the selected proposals. That is around 30 out of 90.

/ **Doubt:** can we pilot a doble process to check the 2 pathways in the same experiment? → *More pre-selected proposals*

## *More pre-selected proposals*

/ **Strength:** if results are solid, it would be more likely that we actually apply this process, meaning a decrease on the reviewers' workload on the interviews and no remote evaluation.

/ **Limitation:** for this pilot we need to request extra work to our reviewers.

/ **Doubt**: could we do this pilot with only half of the panel testing the parallel process? Or better, different panel reviewers that do not participate that edition?

# Thank you *for your attention*

https://lacaixafoundation.org/

CaixaResearch

# Case Studies

2. SNSF

# Case Studies

3. AQuAS

# Developing a semi-automatic tool to support the assignment of reviewers to project proposals

Laura Puigcerver & Esther Vizcaino

Agency for Health Quality and Assessment of Catalonia (AQuAS)

Salut/ Agència de Qualitat i Avaluació Sanitàries de Catalunya

Generalitat de Catalunya

SIRIS A·C·A·D·E·M·I·C

# Context

- Annual call of **biomedical research**
- Different **topics** and **scope heterogenity**
- We receive around **180-250 projects** to evaluate
- Two step review process: a first remote evaluation of **3 reviews per proposal**

2025
Cancer

2024
Respiratory diseases

2023
Sexual and reproductive health

2022
Cardiovascular health

2021
Mental health

2020
Covid-19

2019
Rare diseases


2018
Cancer

2017
Infectious diseases

2016
Strokes and traumatic spinal cord and brain injury


2015
Diabetes and obesity

2014
Heart diseases

# The Problem

**Assigning manually peer-reviewers to grant proposals** is time-consuming, resource-intensive and sometimes inconsistent because the high volume of assignments.

- It requires time, expert knowledge and it is hard to scale
- Matching errors can affect fairness and funding decisions

# The Intervention

Design a prototype of an AI tool to facilitate *–not to replace-* the manual assignment of reviewers to proposals incorporating:

- Large language models
- Information extraction methods
- Criteria (mandatory and desirable) and restrictions imposed by a provided set of guidelines

**IF** the AI tool could propose reviewers with at least 80% adequacy, **THEN** it could reduce manual burden and free up our staff for more strategic tasks

AQuAS did not have the in-house data science expertise to address this challenge, so we commissioned to a consultancy company

# DATA SOURCES

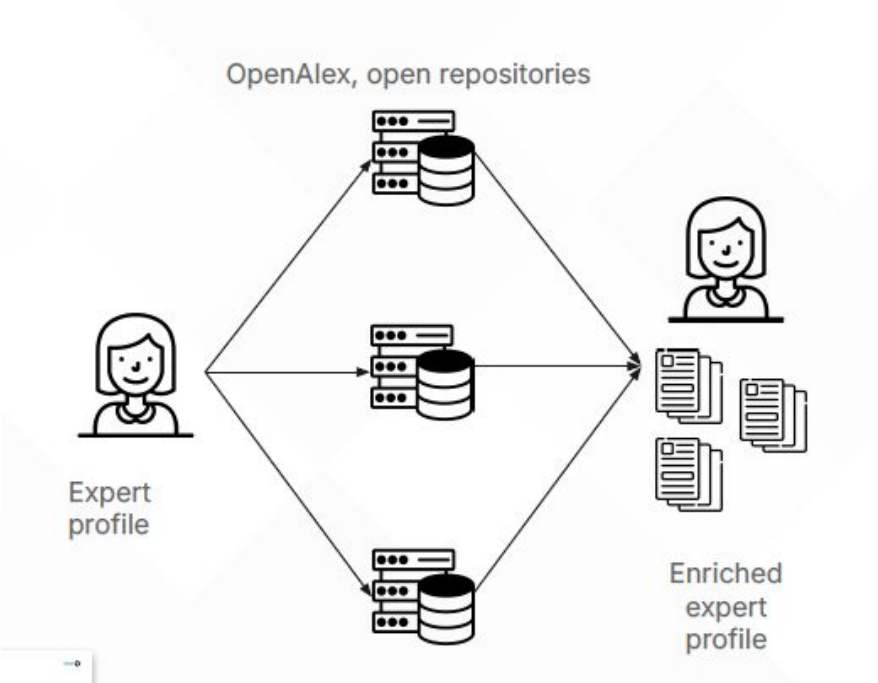
**Self-Reported data:** Reviewer's questionnaire including topics and methods expertise and **5 most relevant publications** in relation to the Call

Proposal abstracts and titles

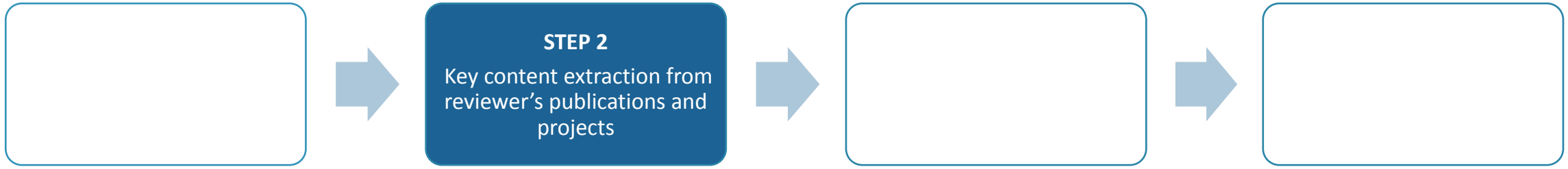
# The Reviewer-matcher methodology

**STEP 1**
Enrichment of expert profiles reviewers

Experts profile **are enriched with data automatically** extracted from their most relevant publications



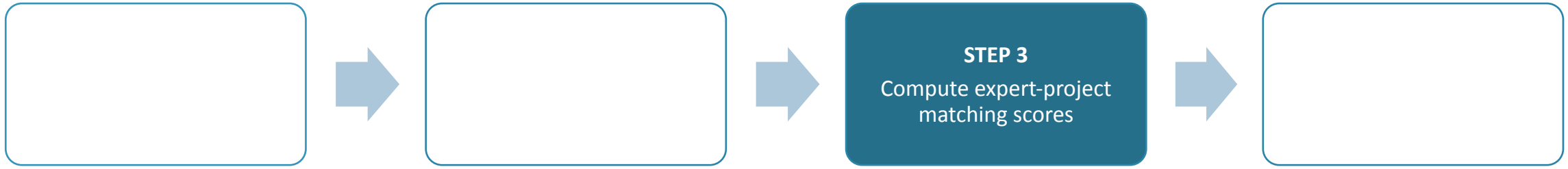
# The Reviewer-matcher methodology




Uses LLMs to extract structured information in **the reviewer's publications and projects** from **title and abstracts**

- Main research topic
- Objectives
- Methods
- MeSH terms

# The Reviewer-matcher methodology




The predicted model was trained and tested with historical data from **five previous calls**:

- *Manual assignments form previous calls (positive examples)*
- *Manual annotation by AQuAs experts ( negative examples*)

Training data 1463 assignments (2018-2022)

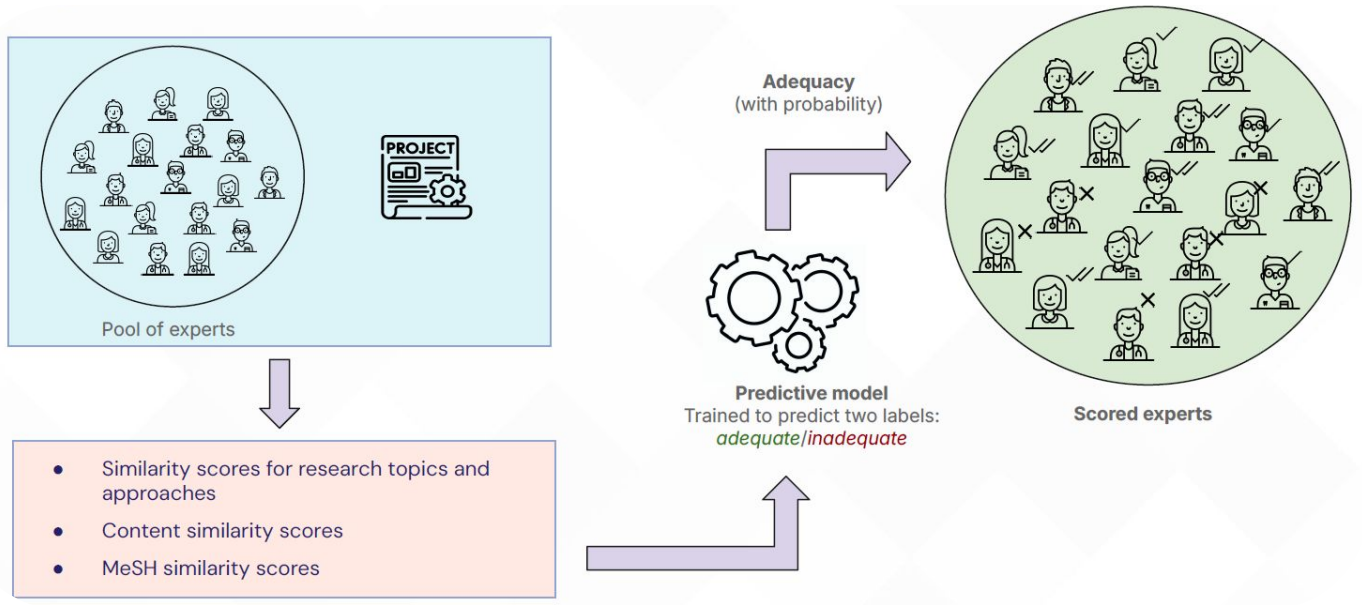
Test data: 363 assignments (2023)

| Year | Inadequate | Adequate | Total |
|---|---|---|---|
| 2018 | 166 | 203 | 369 |
| 2019 | 166 | 179 | 345 |
| 2021 | 173 | 197 | 370 |
| 2022 | 181 | 198 | 379 |
| 2023 | 178 | 185 | 363 |

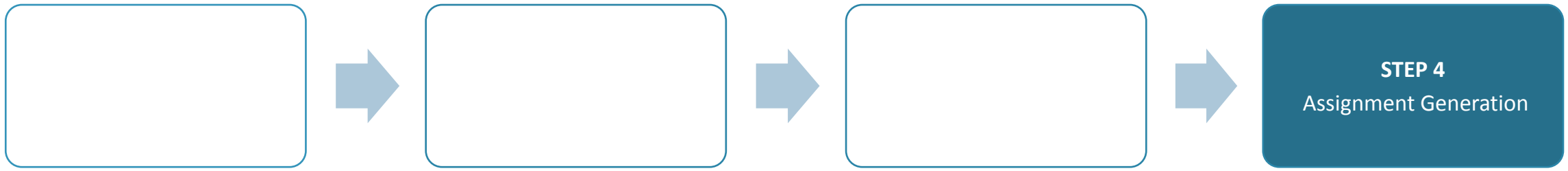
# The Reviewer-matcher methodology

**STEP 3**
Compute expert-project matching scores

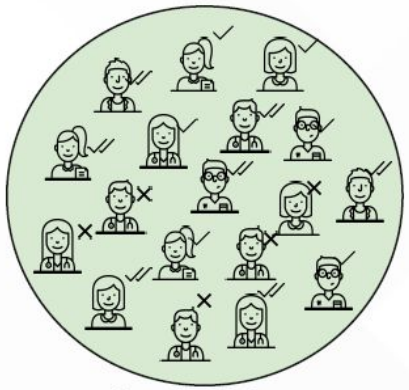
Computes multiple similarity scores producing a combined score per expert-proposal paired



# The Reviewer-matcher methodology




The model generates a subset of *3 proposed* and *8-15 alternative* candidates for each proposal



Scored evaluators


- Predicted Adequacy
- Workload
- Gender balanced

Experts satisfying project-level and global constraints (e.g., gender distribution, number of assignments) divided into *proposed* and *alternative*

# The Comparison and the outcomes

411 **expert-project pairs** generated by the model were manually reviewed by AQuAS, observing that:

- Only 18% of the **pairs reviewer-proposal** were adequate
- 31% could do the job although they were not the best match
- **45% of the 3 proposed reviewers were not adequate**

# What did not work (yet)

- **Missing self-declared research type:**

The final model did not incorporate the research type self-reported by reviewers and proposals

- **Mismatch in thematic expertise:**

Some proposed reviewers matched the methodology but lacked domain-specific knowledge

- **Mismatch in methodological expertise (less frequent):**

Some proposed viewers had relevant thematic expertise but lacked experience with required methods

# Next Steps

- Add **more past data and negative examples** of assignments
- Integrated **research type** (self-reported) as a constraint
- **Balance** topics/methods expertise in the model
- Use **keyword hierarchies** to reduce noise
- **Raise thresholds** for stronger matches

# Break

Start again on the hour (15.00 BST / 16.00 CEST / 10.00 ET)

# Welcome From Amanda Kvarven

Research Fellow,
Research on Research Institute
https://researchonresearch.org/

# The basics of experimentation

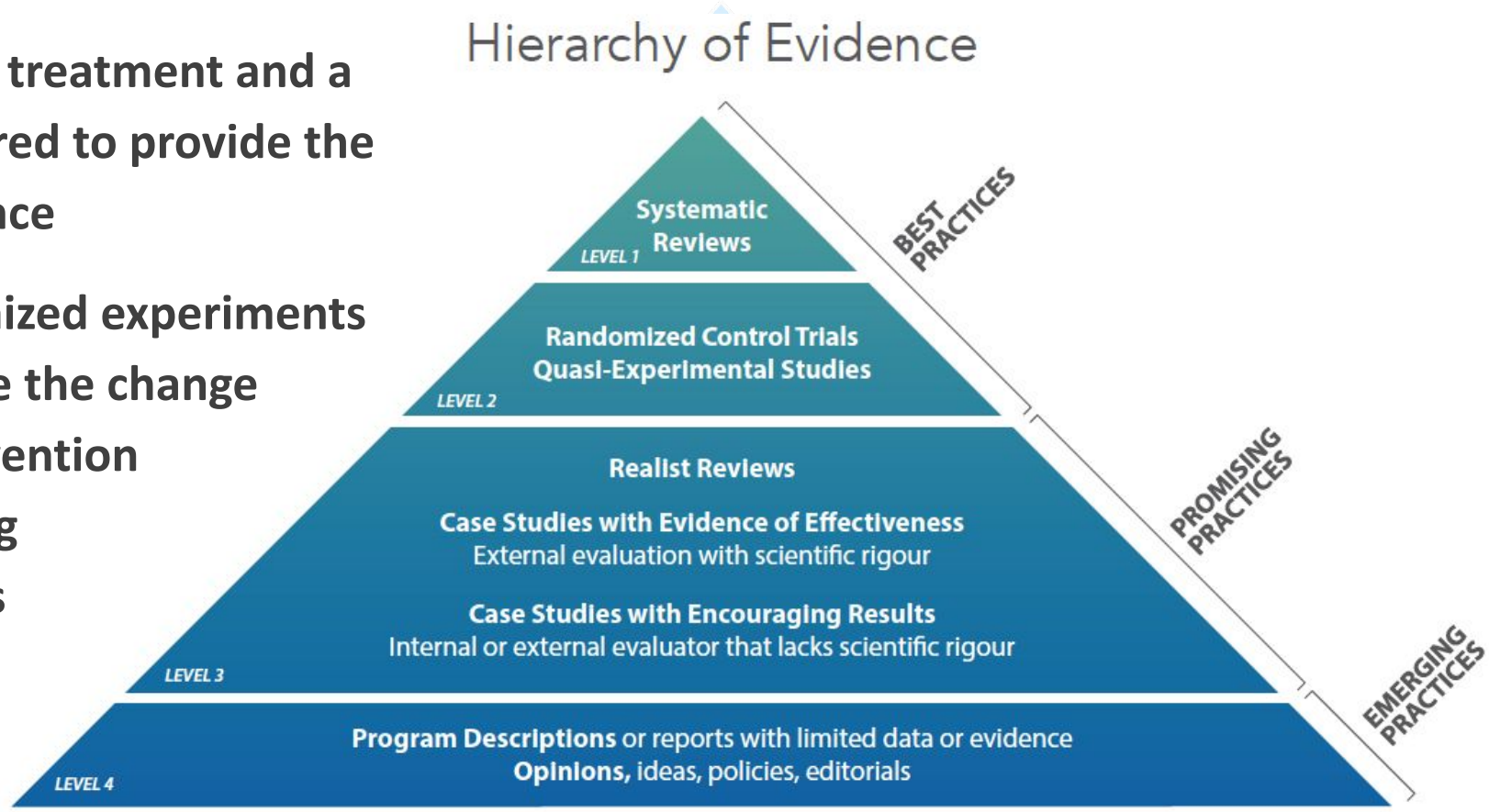
- **Intervention**
- **Randomization and comparison**
- **Sample size**
- **Outcomes**
- **Designing a research question**

# Intervention

- **Some change you want to measure the effect of**
- **Ideally something you can control, though doesn't need to be**

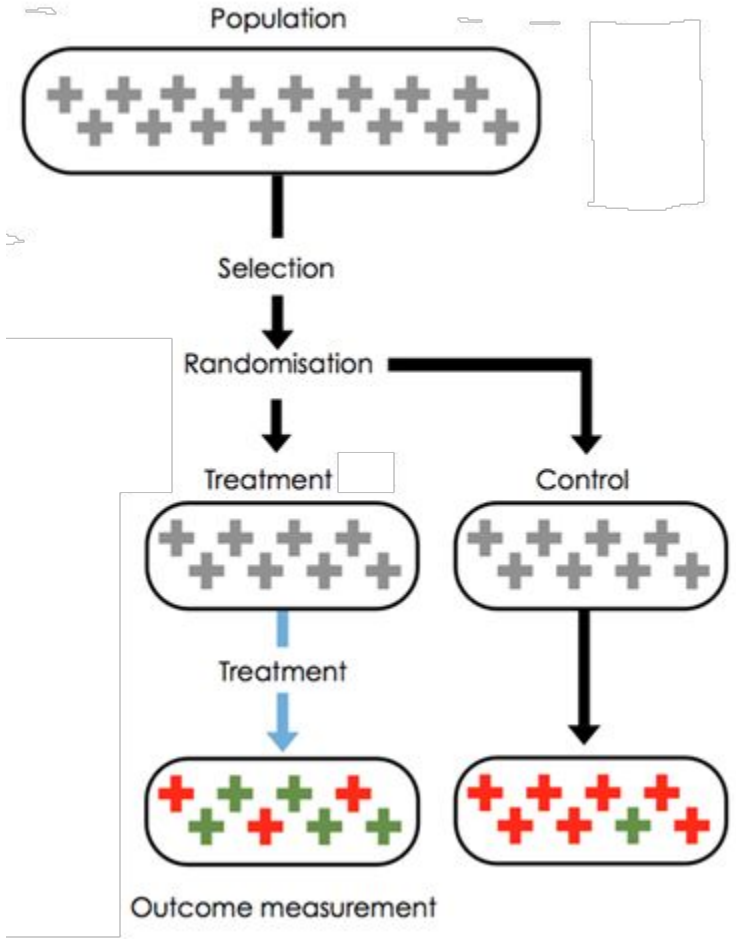
# Randomization and appropriate comparison

- **Randomizing between a treatment and a control group is considered to provide the highest quality of evidence**
- **The idea behind randomized experiments is that we can clearly see the change in outcome of our intervention without the results being affected by other factors**



# Basic RCT design

1. Select a sample
2. Randomize sample into treatment and control group
3. Apply intervention to treatment group
4. Compare outcome in treatment group to outcome in comparison group



# Why do we need a comparison?

- **A comparison is needed to ensure that the change in outcome is due to the intervention**
- **This is because sometimes changes in outcome might vary over time or due to other factors that change**
- **Therefore, it is important to find a comparison that is as identical as possible to the treatment group which can then be used to measure what the outcome would have been for the treatment group if there was no treatment**
- **If we have two groups, and the only difference between them is the intervention, we can be quite sure that the change in outcome is due to the intervention**

# Why randomize?

- One challenge when doing experiments is that subjects might self-select into groups if given the chance
- Ideally, in experimentation, we try to vary one thing at the time, to make sure that the change in outcome is caused by the intervention
- If we allow for selection into groups based on known traits of preference, we have two differences between the groups
  - The intervention, as this is only applied to the treatment group
  - The composition of the group
- This is called selection bias, and can be avoided by randomly allocating people into treatment and comparison

# What if we can’t randomize?

- **Shadow experiment**
  - **This would mean applying the intervention to the entire group of subjects, but to not implement it. This way, you can compare what the outcome would have been like if we applied treatment to the actual outcome (where the treatment is not applied)**
- **Quasi-experimental methods**
  - **Find groups that can be compared where the membership of each group is quasi-random**

# Sample size

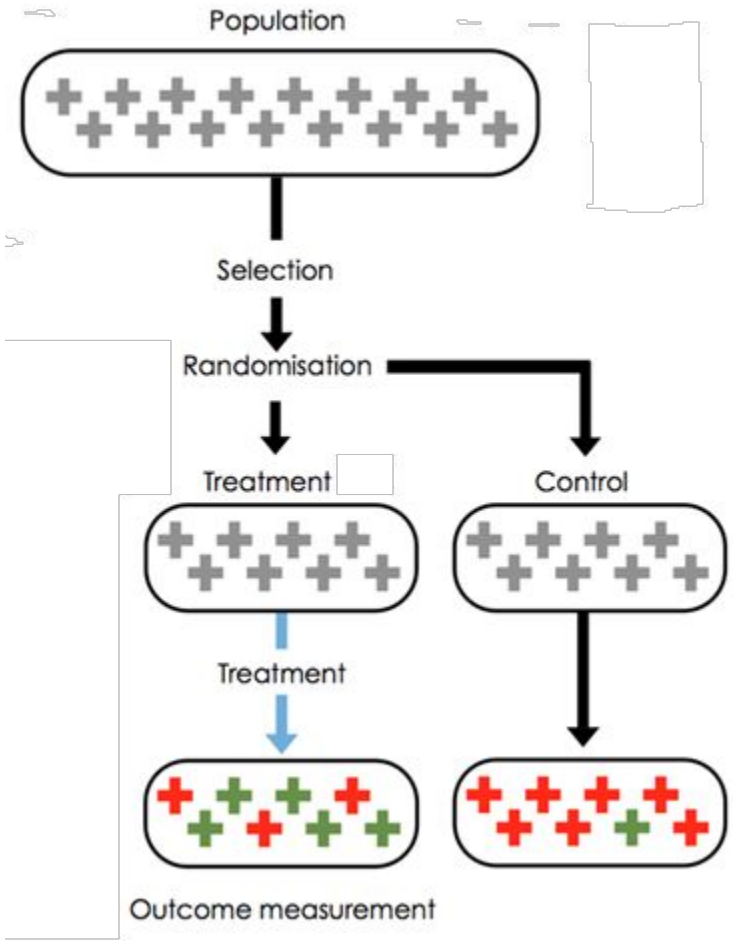
- **When planning an experiment it is important to factor in the sample size, because you can expect small random changes which can be confused with a real effect in small samples**
- **With a large sample size, one stray observation will have less to say for the overall, as it is a small percentage of the full sample**

# Outcomes

- **When planning an experiment, there are usually several possible outcomes to choose from**
- **A good outcome measure will be**
  - **Easy to define and measure**
  - **Contain enough observations to provide a high sample size**
  - **Tell us something about the effect of the intervention**

# Designing a research question - PICO

- **Population**
  - **Describe your population**
    - **What kind of grant call is it?**
    - **What are the relevant requirements?**
    - **How large is your population**
- **Intervention/treatment**
  - **Describe your intervention**
- **Comparison**
  - **What is your comparison group?**
- **Outcome**
  - **Use IF THEN statement to find good outcome measures**



# Breakout group discussion

**Topic:** Research question
**Duration**: 30 minutes
**Structure:** New groups x 3

Going around each funder:

- Using the PICO framework, design a research question that your organization is interested in

For each funder, the aim is to have **a clearly defined research question that you want answered**.

# Plenary

Facilitator feedback + opportunity for questions

# Next time

**Homework**: submit PICO + volunteer to pitch in Workshop 3

**Session 2 feedback & homework survey** https://forms.gle/UfkJqDYJ6VEqD1u16

- Your coach will pick up on your answers next week (as well as informing Workshop 3)

- If you want to develop two ideas, please fill out the survey twice

**Pre-session activity** for session 3: coming soon!

**One-to-one coaching slot**: 17th of June (or anytime that week)

**Discuss**: via the google group, share resources, form interest groups

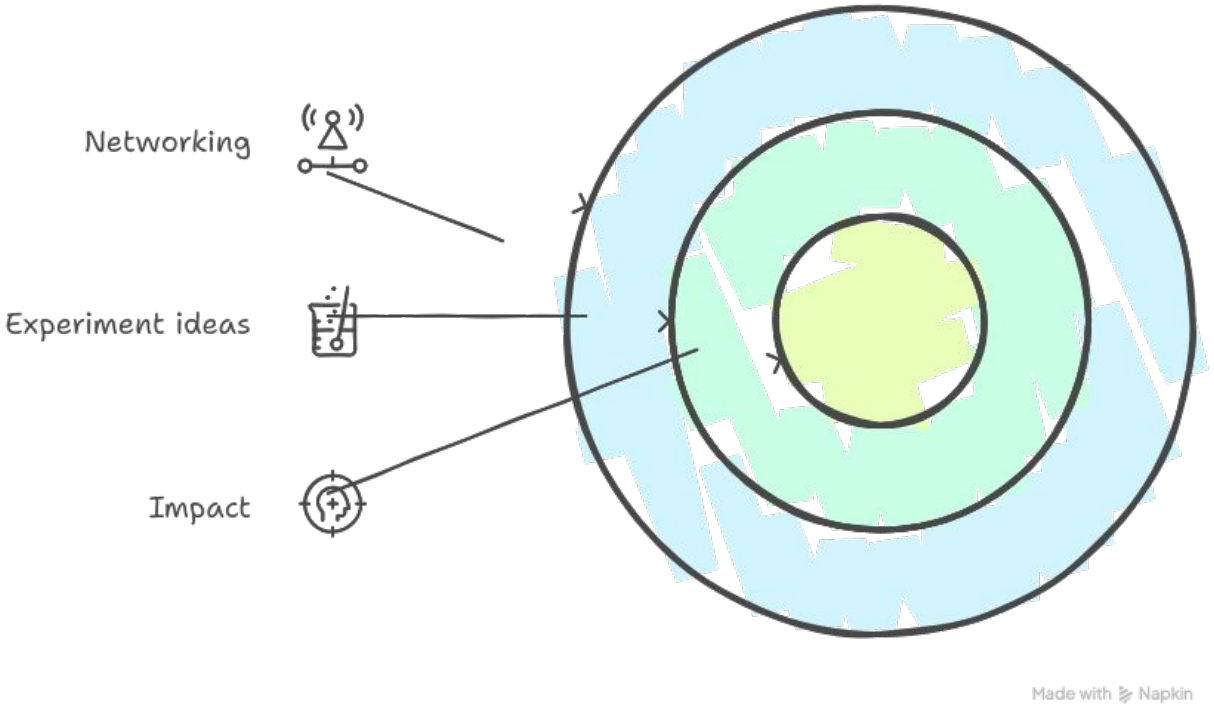
**Next workshop**: 24th of June: pitching + advocating for experiments

# METASCIENCE 2025 CONFERENCE

Save the Date

## June 30 – July 2, 2025
## University College London

# Metascience Lab @ MS2025




- in partnership with Open Philanthropy and RoRI's AFiRE programme

- three linked sessions will facilitate matchmaking and networking for experimentation

- all areas of metascience, with a focus on interventions to support higher quality, lower cost and more impactful research.

- Each session will showcase metascience principles, methods or examples of experimentation, as well as providing a platform for co-developing new project ideas by participants. Researchers, funders, universities, publishers and other actors in the research ecosystem are invited to propose experiments and matchmake with potential collaborators.

- The Abundance and Growth Fund at Open Philanthropy is happy to consider proposals that emerge from this process

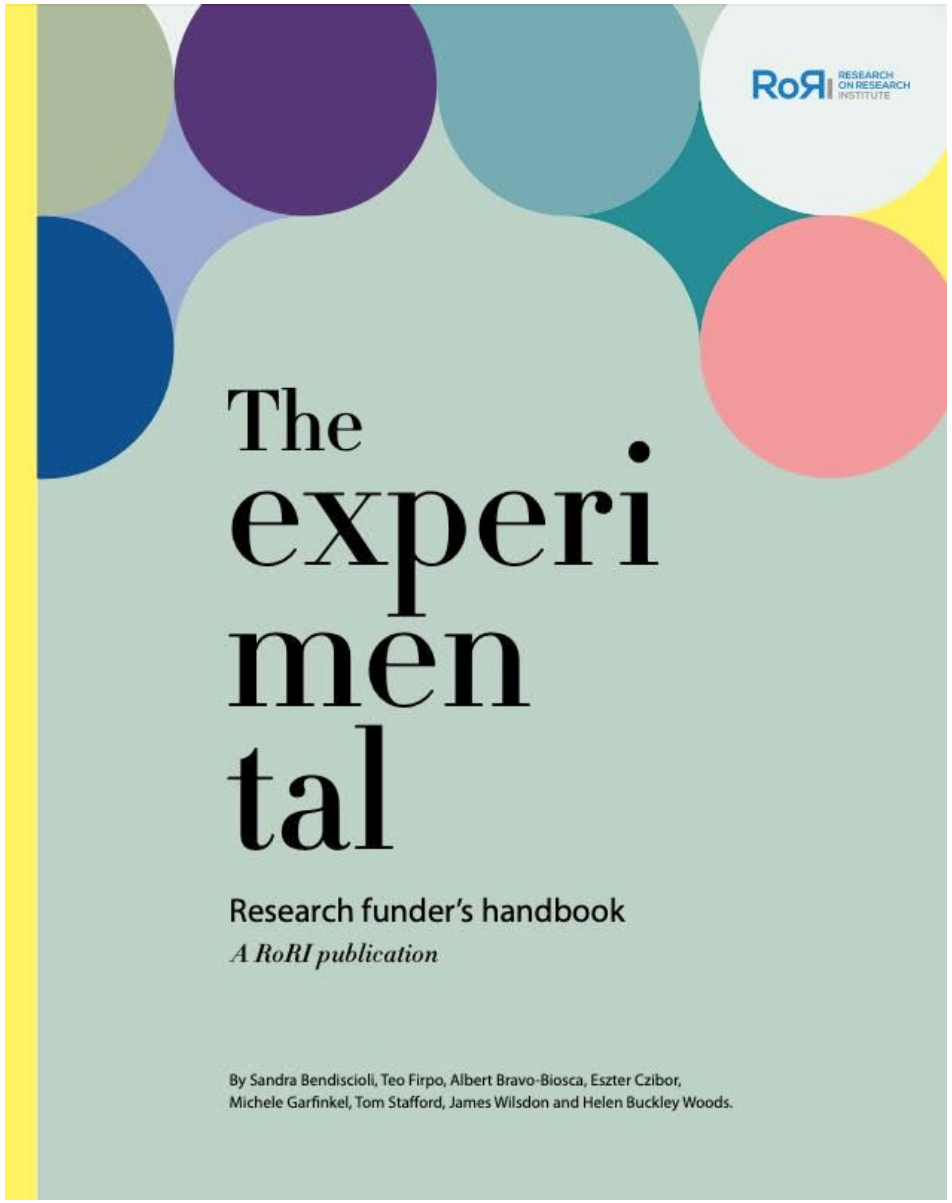
- Topics you'd like considered? Please get in touch

# Art of Funding @ MS2025

Come join a small group of funders to discuss the “Art of Funding”. Topics may include advancing new ideas within your organization, overcoming bottlenecks, efficiencies, and logistics of making and monitoring awards.

Bring your questions and ideas to share. **Drinks will be served**.

Please RSVP https://forms.gle/aHhpVWWc2gs7Fpux8. Discussions will continue afterwards at a restaurant of your choice.




Bendiscioli, Sandra; Firpo, Teo; Bravo-Biosca, Albert; Czibor, Eszter; Garfinkel, Michele; Stafford, Tom; et al. (2022):

The experimental research funder's handbook (Revised edition, June 2022, ISBN 978-1-7397102-0-0).

Research on Research Institute. Report. https://doi.org/10.6084/m9.figshare.19459328.v2

These slides: http://bit.ly/tom-talks

# Further resources on experimentation

Bendiscioli, Sandra; Firpo, Teo; Bravo-Biosca, Albert; Czibor, Eszter; Garfinkel, Michele; Stafford, Tom; et al. (2022): The Experimental Research Funder's Handbook (Revised edition, June 2022, ISBN 978-1-7397102-0-0). Research on Research Institute.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., ... & Ioannidis, J. (2017). A manifesto for reproducible science. Nature human behaviour, 1(1), 1-9. https://doi.org/10.1038/s41562-016-0021

Why observation is not enough, even if you have sophisticated analysis:

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. Marketing Science, 38(2), 193-225.

Westfall, J., & Yarkoni, T. (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. PLOS ONE, 11(3), e0152719. https://doi.org/10.1371/journal.pone.0152719

# Join the conversation

researchonresearch.org

@RoRInstitute

RoRI RESEARCH ON RESEARCH INSTITUTE