Experimentation users-group [Xug] Meetings

Beyond computational reproducibility: what does it mean for neuroimaging results to be irreproducible?

8 mars 2022

Camille Maumet

Univ Rennes, Inria, CNRS, Inserm

A crisis in **experimental research**

The **reproducibility crisis** has led to **reduced confidence in research findings**

Low reproduction rates in many fields :

Cancer research: <11%

Psychology: 36%

Medicine: 44%

(Begley & Ellis 2012 - Open Science Collab 2016 -Ioannidis 2005)

A crisis in **experimental research**

The **reproducibility crisis** has led to **reduced confidence in research findings**

Low reproduction rates in many fields :

Cancer research: <11%

Psychology: 36%

Medicine: 44%

(Begley & Ellis 2012 - Open Science Collab 2016 -Ioannidis 2005)

Wasted money & effort for research

Delayed translation into clinical practice

Reduced trust in science

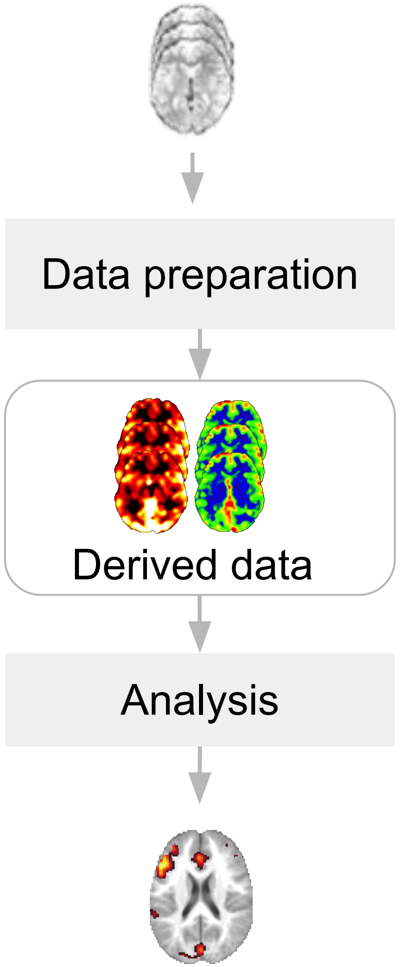
Reproducible evaluations ?

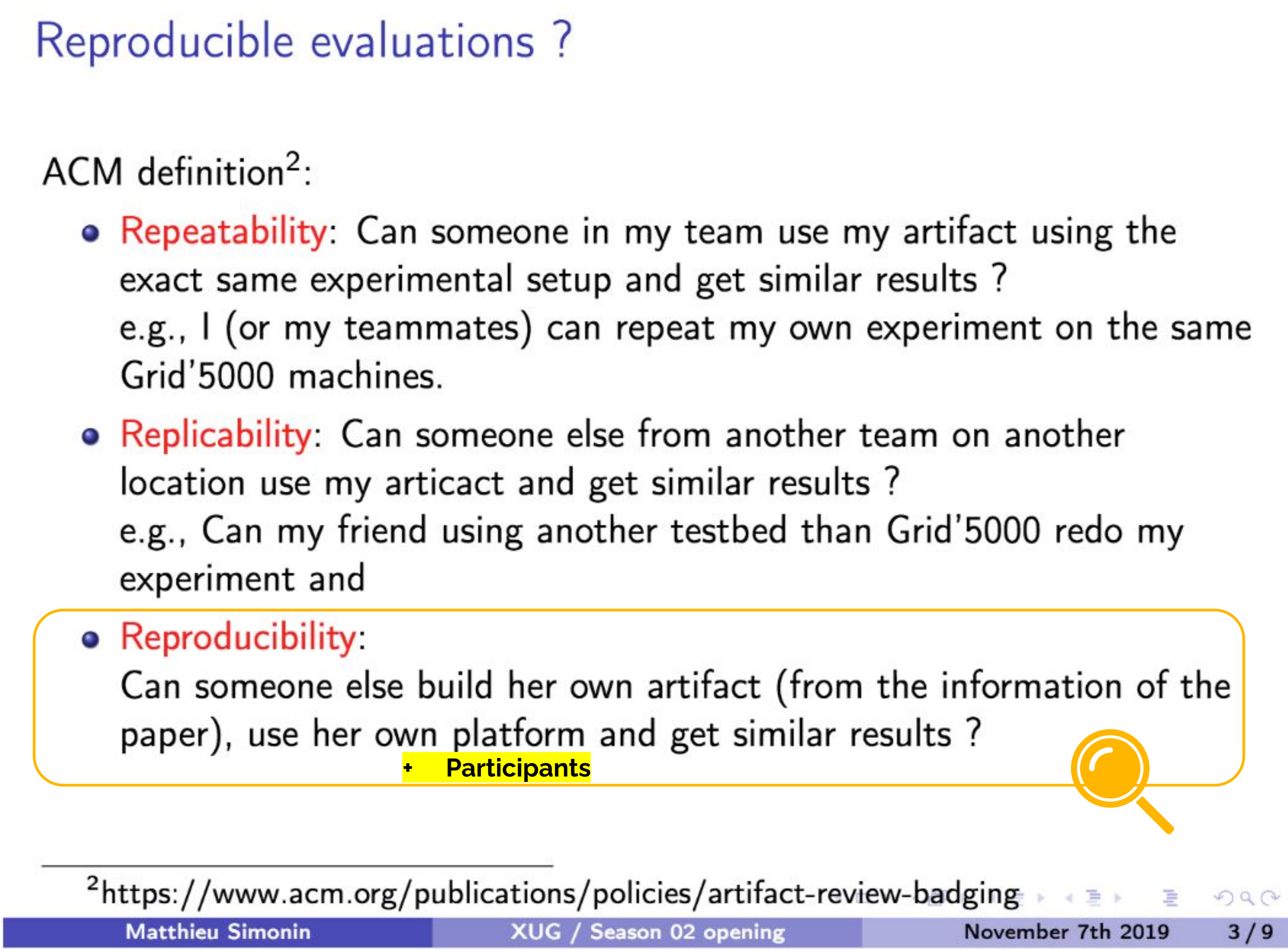
ACM definition[2]:

- Repeatability: Can someone in my team use my artifact using the exact same experimental setup and get similar results ?
  e.g., I (or my teammates) can repeat my own experiment on the same Grid'5000 machines.
- Replicability: Can someone else from another team on another location use my articact and get similar results ?
  e.g., Can my friend using another testbed than Grid'5000 redo my experiment and
- Reproducibility:
  Can someone else build her own artifact (from the information of the paper), use her own platform and get similar results ?

[2] https://www.acm.org/publications/policies/artifact-review-badging

A brain imaging study





Reproducible evaluations ?

ACM definition[2]:

- Repeatability: Can someone in my team use my artifact using the exact same experimental setup and get similar results ?
  e.g., I (or my teammates) can repeat my own experiment on the same Grid'5000 machines.
- Replicability: Can someone else from another team on another location use my articact and get similar results ?
  e.g., Can my friend using another testbed than Grid'5000 redo my experiment and
- Reproducibility:
  Can someone else build her own artifact (from the information of the paper), use her own platform and get similar results ?

+ **Participants**

[2] https://www.acm.org/publications/policies/artifact-review-badging

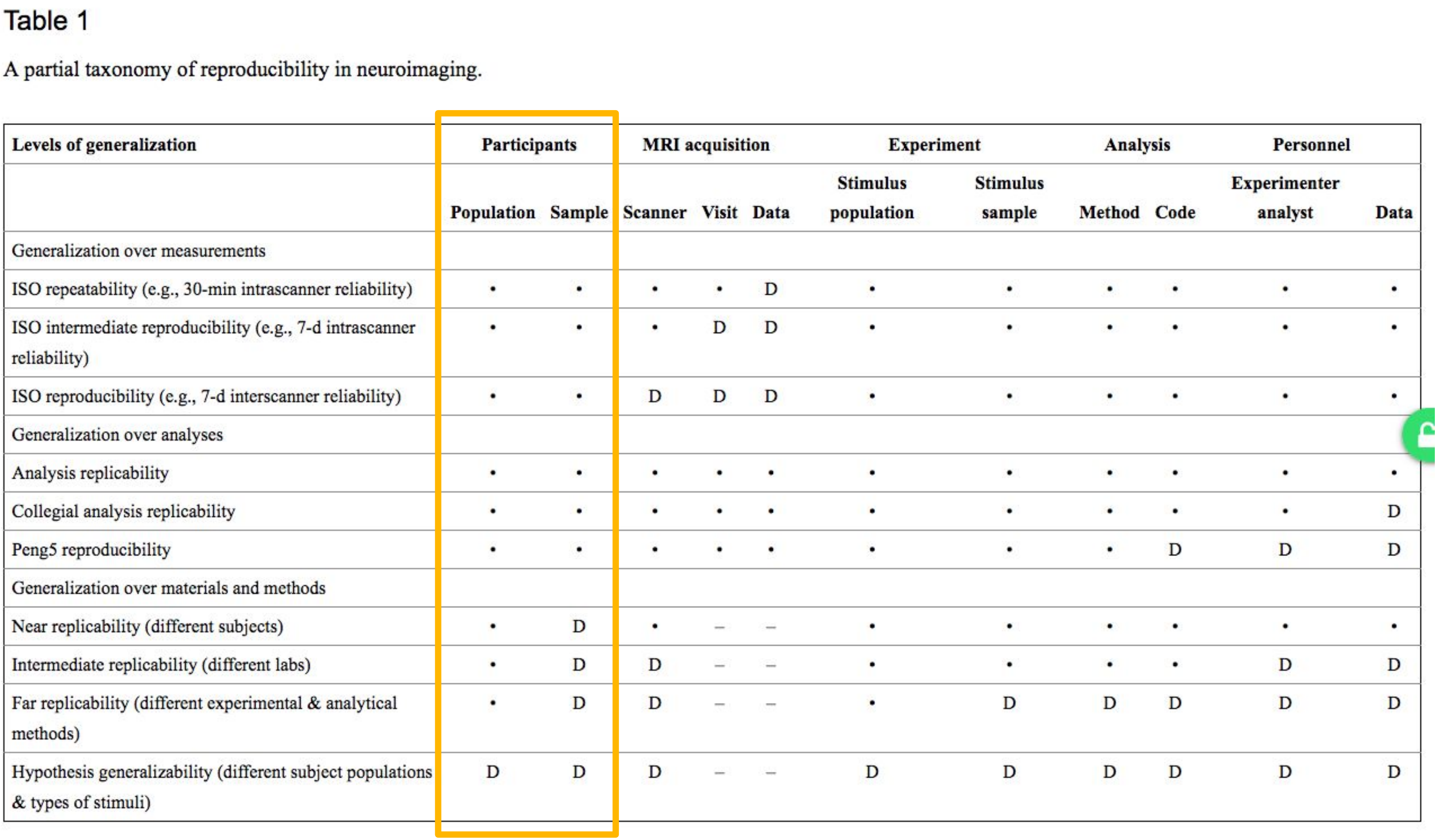


Table 1

A partial taxonomy of reproducibility in neuroimaging.

Levels of generalization	Participants		MRI acquisition			Experiment		Analysis		Personnel	
	Population	Sample	Scanner	Visit	Data	Stimulus population	Stimulus sample	Method	Code	Experimenter analyst	Data
Generalization over measurements											
ISO repeatability (e.g., 30-min intrascanner reliability)	•	•	•	•	D	•	•	•	•	•	•
ISO intermediate reproducibility (e.g., 7-d intrascanner reliability)	•	•	•	D	D	•	•	•	•	•	•
ISO reproducibility (e.g., 7-d interscanner reliability)	•	•	D	D	D	•	•	•	•	•	•
Generalization over analyses											
Analysis replicability	•	•	•	•	•	•	•	•	•	•	•
Collegial analysis replicability	•	•	•	•	•	•	•	•	•	•	D
Peng5 reproducibility	•	•	•	•	•	•	•	•	D	D	D
Generalization over materials and methods											
Near replicability (different subjects)	•	D	•	–	–	•	•	•	•	•	•
Intermediate replicability (different labs)	•	D	D	–	–	•	•	•	•	D	D
Far replicability (different experimental & analytical methods)	•	D	D	–	–	•	D	D	D	D	D
Hypothesis generalizability (different subject populations & types of stimuli)	D	D	D	–	–	D	D	D	D	D	D

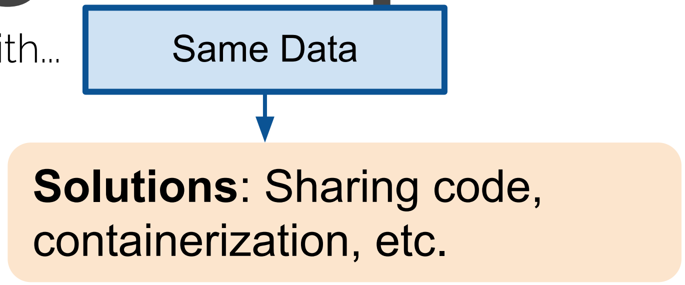
[Nichols et. al, Nature Neuroscience 2017]

Fixing the **reproducibility** issue

Irreproducible with... Same Data

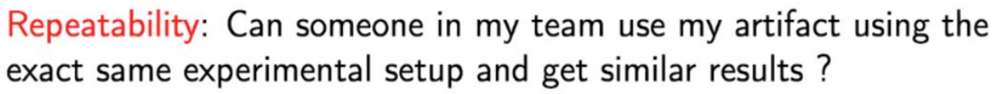
Fixing the **reproducibility** issue

Irreproducible with...



Same Data

Solutions: Sharing code, containerization, etc.



Repeatability: Can someone in my team use my artifact using the exact same experimental setup and get similar results ?

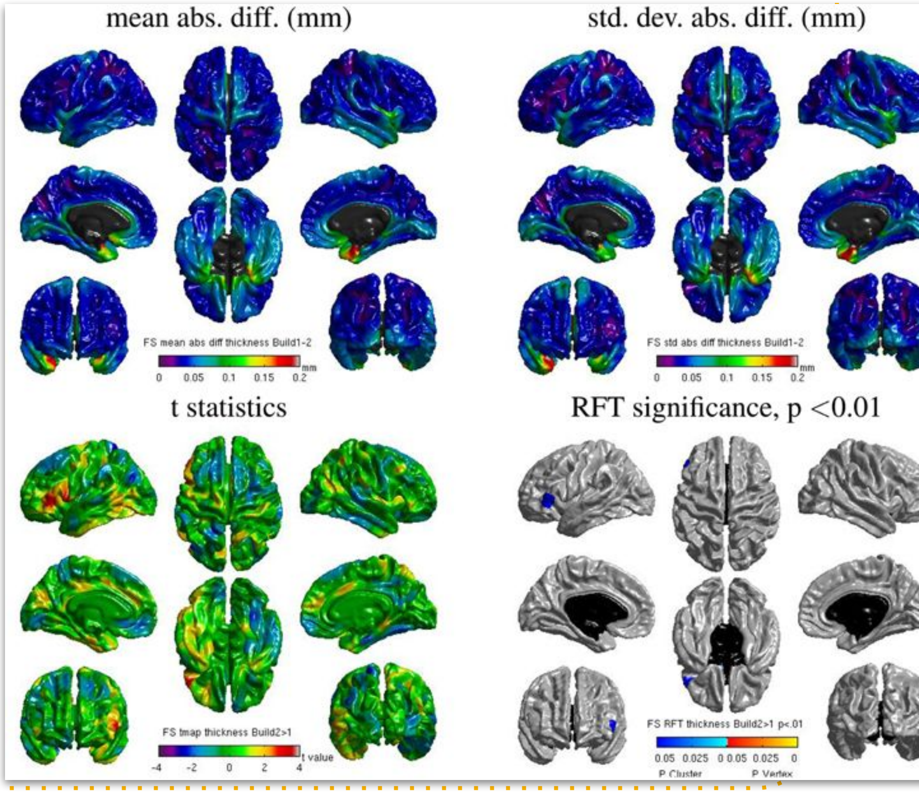
Fixing the **reproducibility** issue

Irreproducible with... Same Data

Solutions: Sharing code, containerization, etc.

Open question: impact of different software environments?

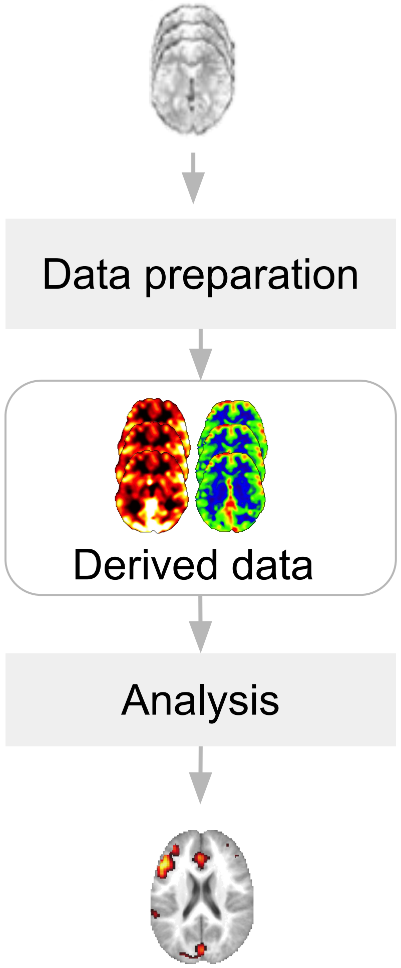
Replicability: Can someone else from another team on another location use my articact and get similar results ?


mean abs. diff. (mm)
std. dev. abs. diff. (mm)
FS mean abs diff thickness Build1-2
0 0.05 0.1 0.15 0.2 mm
FS std abs diff thickness Build1-2
0 0.05 0.1 0.15 0.2 mm
t statistics
RFT significance, p <0.01
FS tmap thickness Build2>1
-4 -2 0 2 4 t value
FS RFT thickness Build2>1 p<.01
0.05 0.025 0 0.05 0.025 0
P Cluster
P Vertex

[Glatard et. al, Neuroinformatics 2015]

Fixing the **reproducibility** issue

Irreproducible with... Different Data



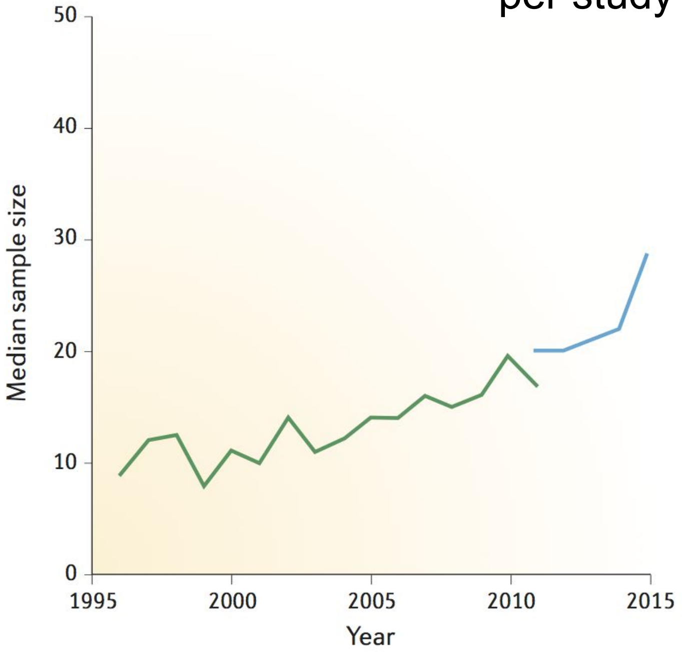
Data preparation

Derived data

Analysis

about 30 participants

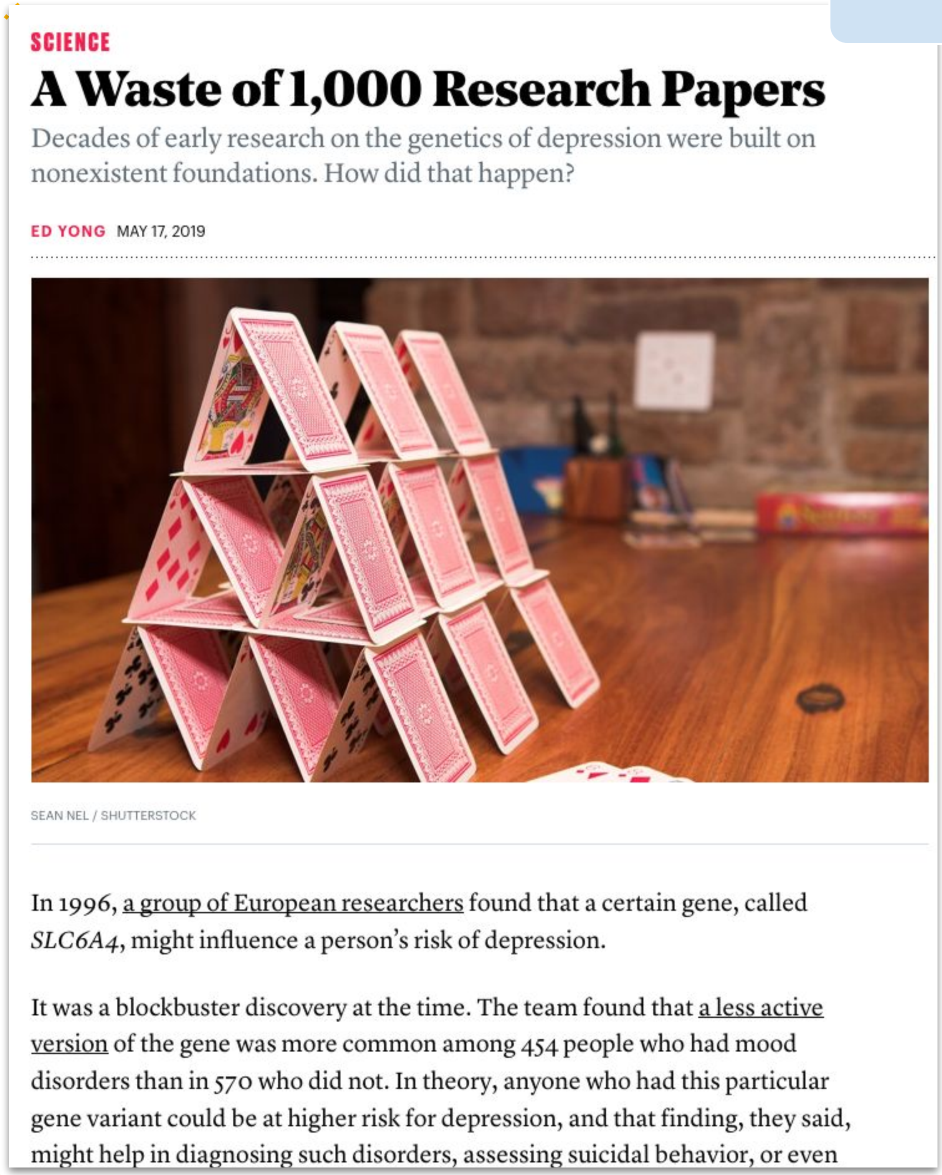
per study



[Poldrack et. al, Nature Neuroscience 2017]

Expl. 1: False positive finding

Low statistical power



SCIENCE

A Waste of 1,000 Research Papers

Decades of early research on the genetics of depression were built on nonexistent foundations. How did that happen?

ED YONG MAY 17, 2019

SEAN NEL / SHUTTERSTOCK

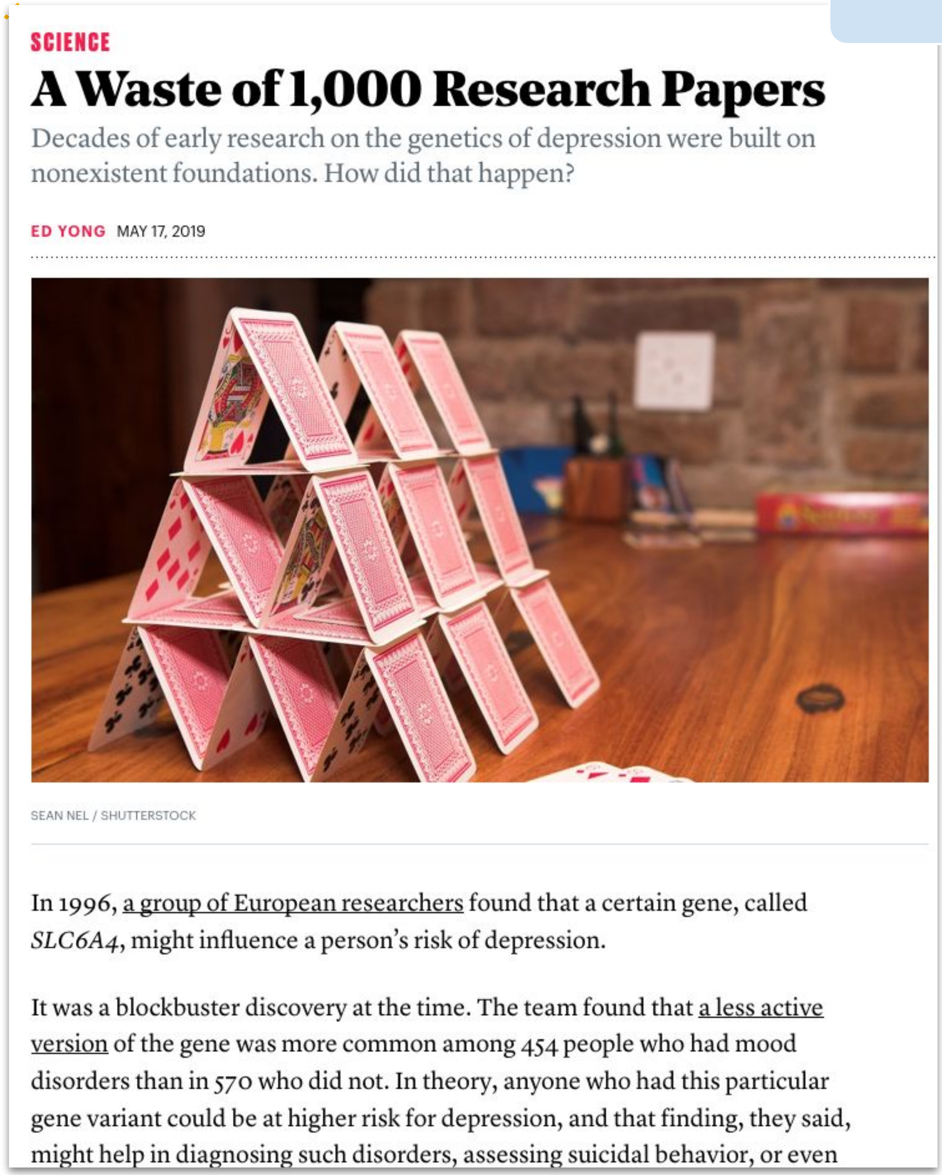
In 1996, a group of European researchers found that a certain gene, called *SLC6A4*, might influence a person's risk of depression.

It was a blockbuster discovery at the time. The team found that a less active version of the gene was more common among 454 people who had mood disorders than in 570 who did not. In theory, anyone who had this particular gene variant could be at higher risk for depression, and that finding, they said, might help in diagnosing such disorders, assessing suicidal behavior, or even

The Atlantic Science, "A waste of 1000 research papers", Ed Yong.

Expl. 1: False positive finding

Low statistical power



SCIENCE

A Waste of 1,000 Research Papers

Decades of early research on the genetics of depression were built on nonexistent foundations. How did that happen?

ED YONG MAY 17, 2019

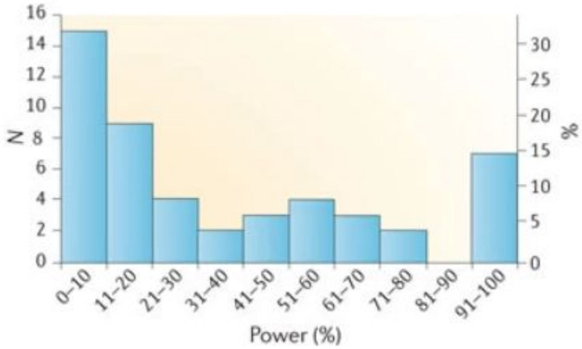
SEAN NEL / SHUTTERSTOCK

In 1996, a group of European researchers found that a certain gene, called *SLC6A4*, might influence a person's risk of depression.

It was a blockbuster discovery at the time. The team found that a less active version of the gene was more common among 454 people who had mood disorders than in 570 who did not. In theory, anyone who had this particular gene variant could be at higher risk for depression, and that finding, they said, might help in diagnosing such disorders, assessing suicidal behavior, or even

Power of neuroscience studies

Power = Prob. to correctly find a significant effect when a the alternative hypothesis is true.

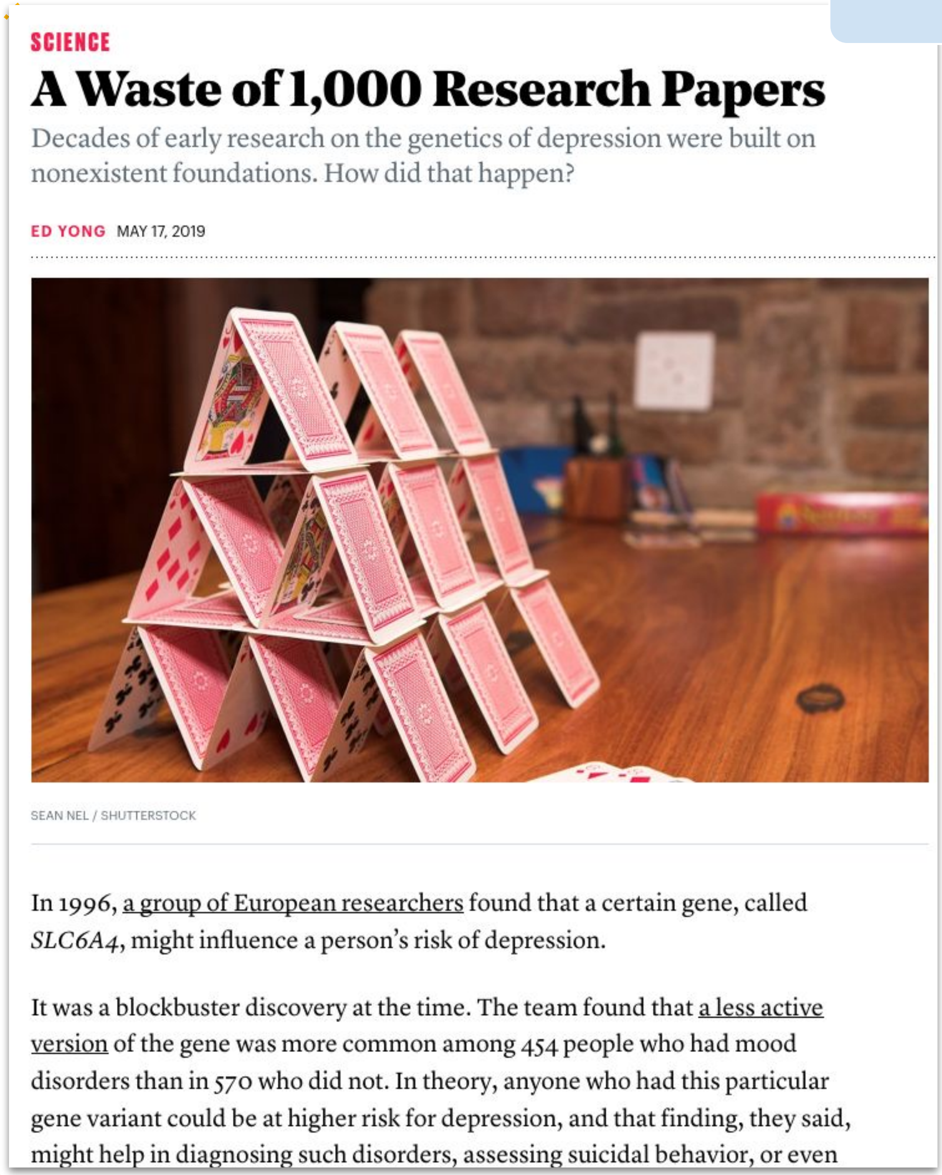


[Button et. al, Nat Rev Neurosci 2013]

The Atlantic Science, "A waste of 1000 research papers", Ed Yong.

Expl. 1: False positive finding

Low statistical power



SCIENCE

A Waste of 1,000 Research Papers

Decades of early research on the genetics of depression were built on nonexistent foundations. How did that happen?

ED YONG MAY 17, 2019

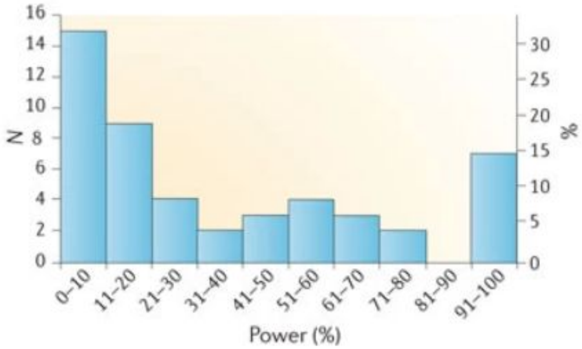
SEAN NEL / SHUTTERSTOCK

In 1996, a group of European researchers found that a certain gene, called *SLC6A4*, might influence a person's risk of depression.

It was a blockbuster discovery at the time. The team found that a less active version of the gene was more common among 454 people who had mood disorders than in 570 who did not. In theory, anyone who had this particular gene variant could be at higher risk for depression, and that finding, they said, might help in diagnosing such disorders, assessing suicidal behavior, or even

Power of neuroscience studies

Power = Prob. to correctly find a significant effect when a the alternative hypothesis is true.

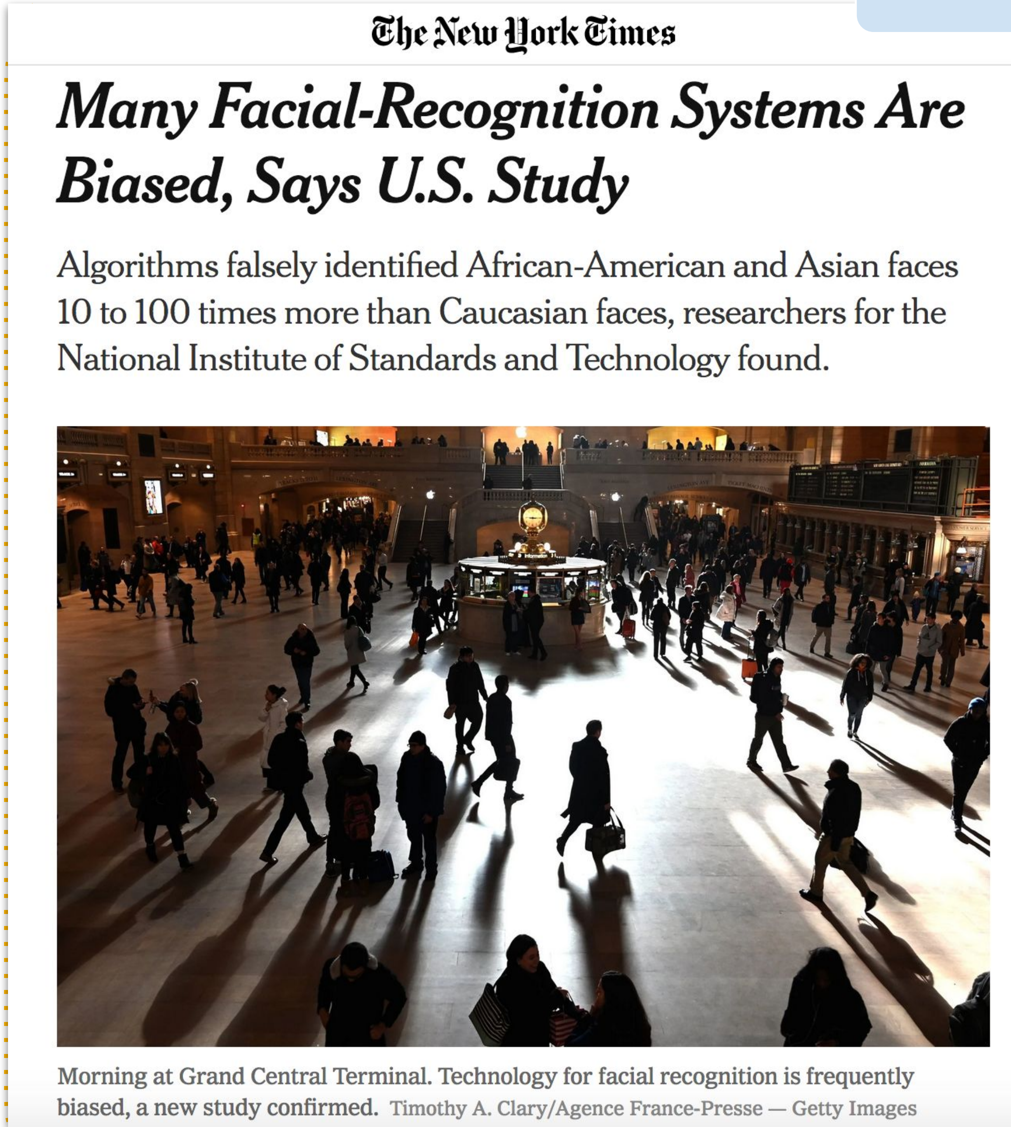


[Button et. al, Nat Rev Neurosci 2013]

Solutions: We need bigger datasets

The Atlantic Science, "A waste of 1000 research papers", Ed Yong.

Expl. 2: Lack of generalizability



The New York Times

Many Facial-Recognition Systems Are Biased, Says U.S. Study

Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

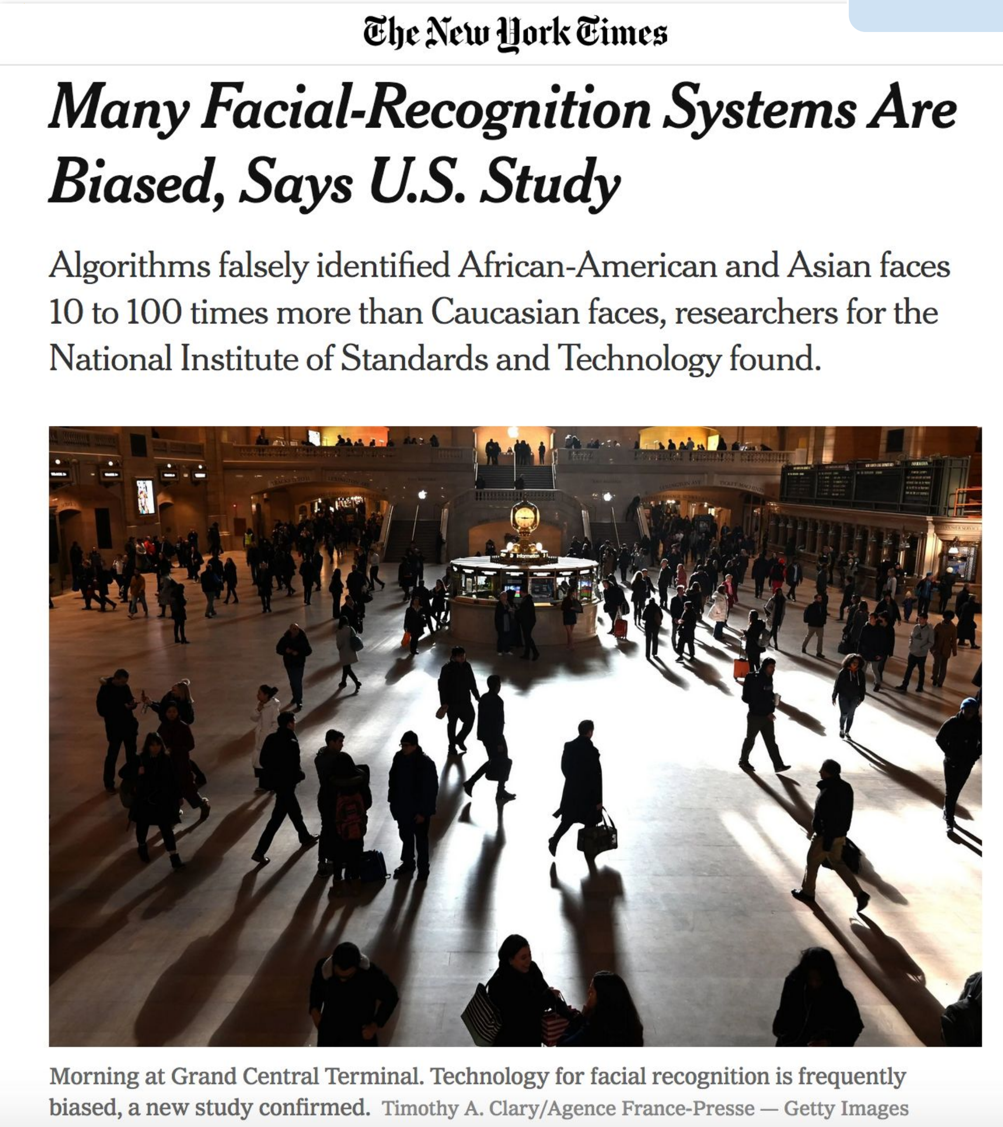
Morning at Grand Central Terminal. Technology for facial recognition is frequently biased, a new study confirmed. Timothy A. Clary/Agence France-Presse — Getty Images

Lack of representativity and diversity

The New York Times, [“Many Facial-Recognition Systems Are Biased, Says U.S. Study”](), By Natasha Singer & Cade Metz,. 2019

Expl. 2: Lack of generalizability

Lack of representativity and diversity



The New York Times

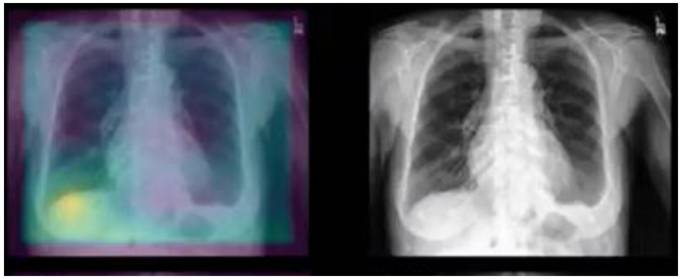
Many Facial-Recognition Systems Are Biased, Says U.S. Study

Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

Morning at Grand Central Terminal. Technology for facial recognition is frequently biased, a new study confirmed. Timothy A. Clary/Agence France-Presse — Getty Images

X-ray: Lung opacity detection

Model trained on male images, tested on female images

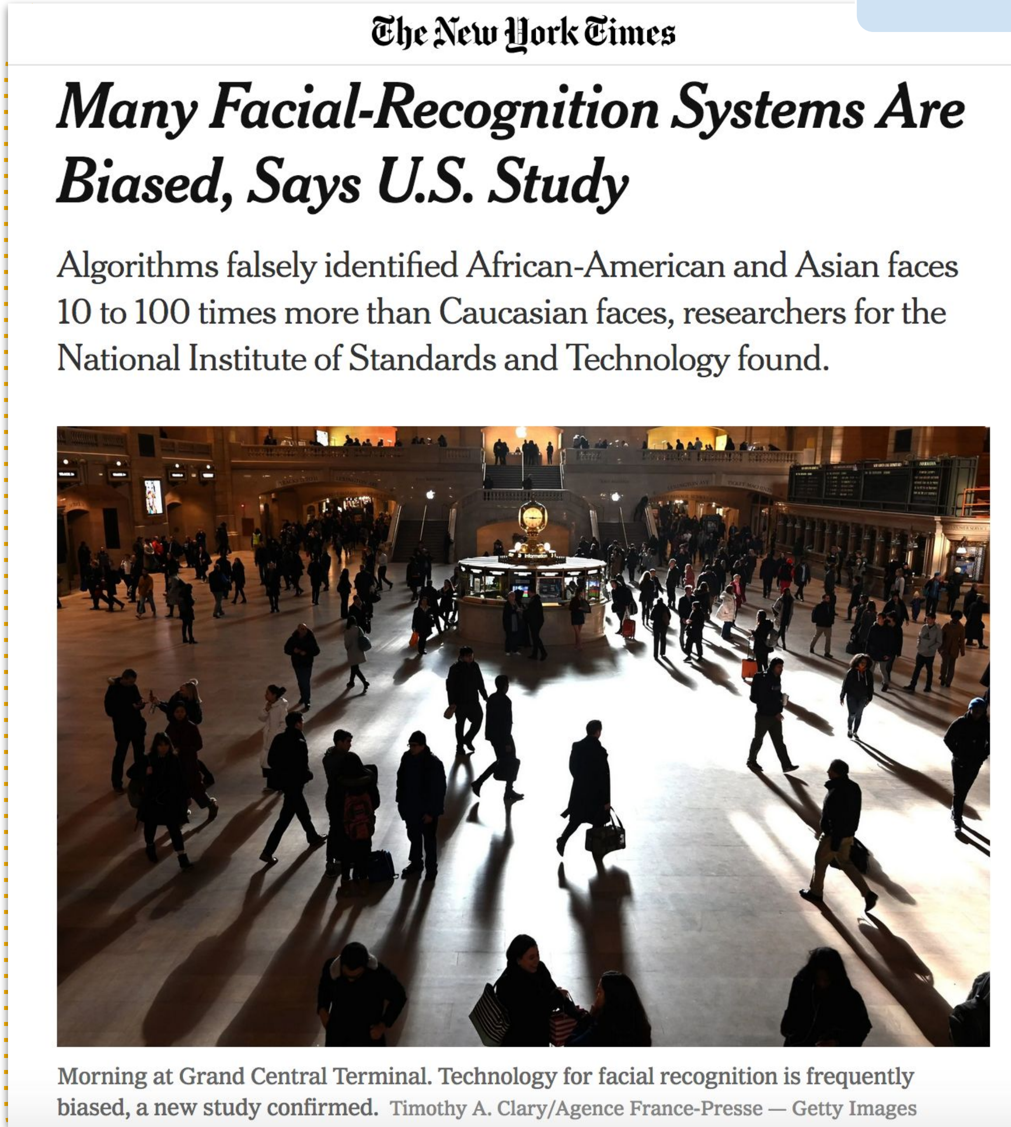


[Larrazabal et. al, PNAS 2020]

The New York Times, "Many Facial-Recognition Systems Are Biased, Says U.S. Study", By Natasha Singer & Cade Metz,. 2019

Expl. 2: Lack of generalizability

Lack of representativity and diversity



The New York Times

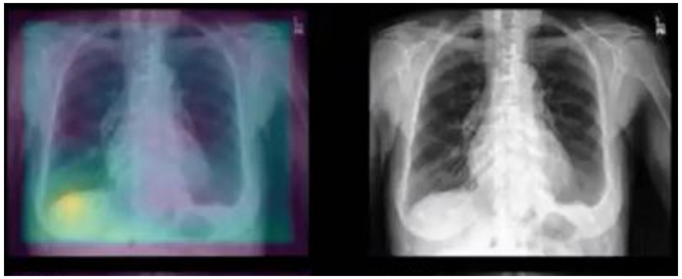
Many Facial-Recognition Systems Are Biased, Says U.S. Study

Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

Morning at Grand Central Terminal. Technology for facial recognition is frequently biased, a new study confirmed. Timothy A. Clary/Agence France-Presse — Getty Images

X-ray: Lung opacity detection

Model trained on male images, tested on female images



[Larrazabal et. al, PNAS 2020]

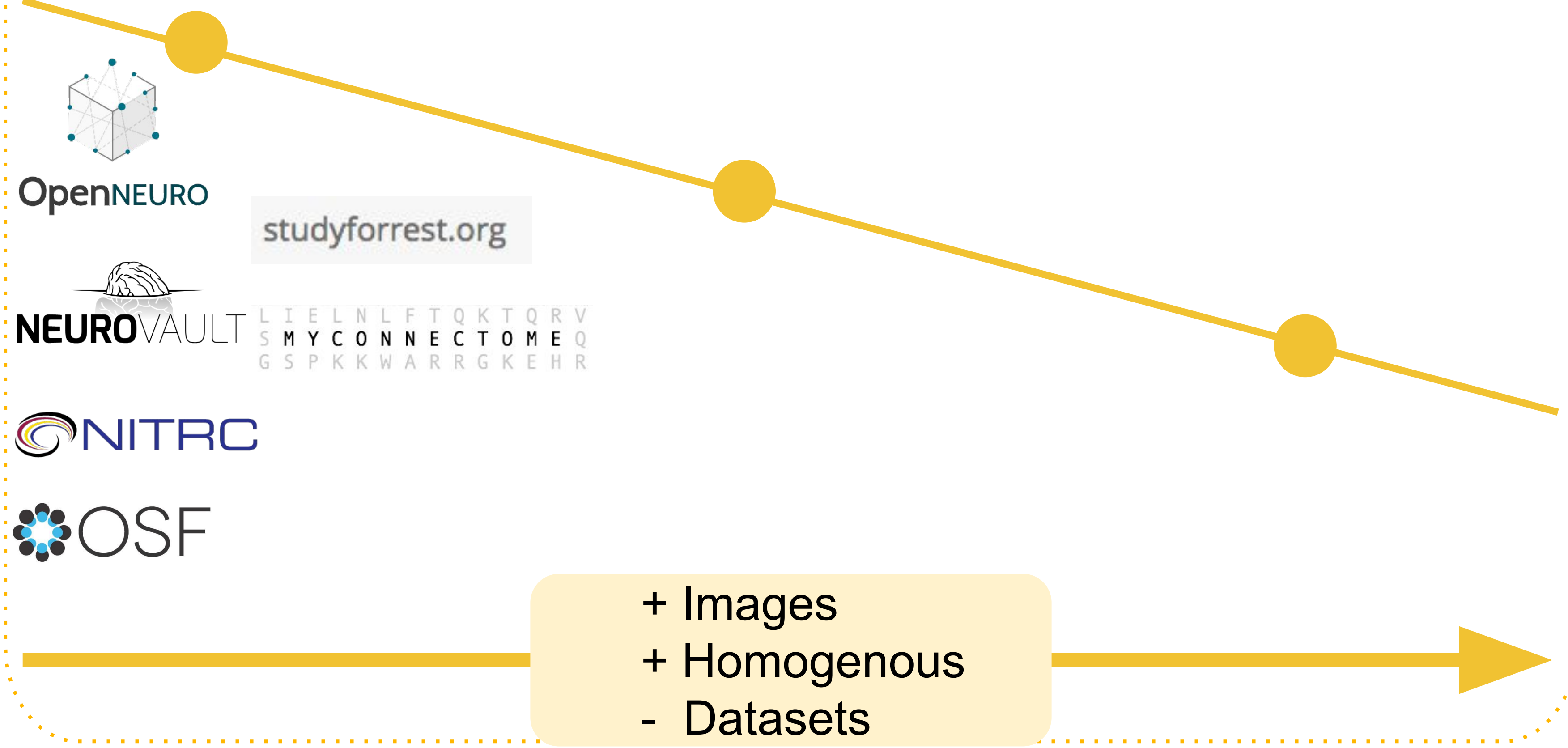
Solutions: We need representative and diverse datasets

The New York Times, "Many Facial-Recognition Systems Are Biased, Says U.S. Study", By Natasha Singer & Cade Metz,. 2019

Open data

Unique study

30 participants



+ Images
+ Homogenous
- Datasets

Credits : Brains, Neil Conway, Flickr (CC BY 2.0)

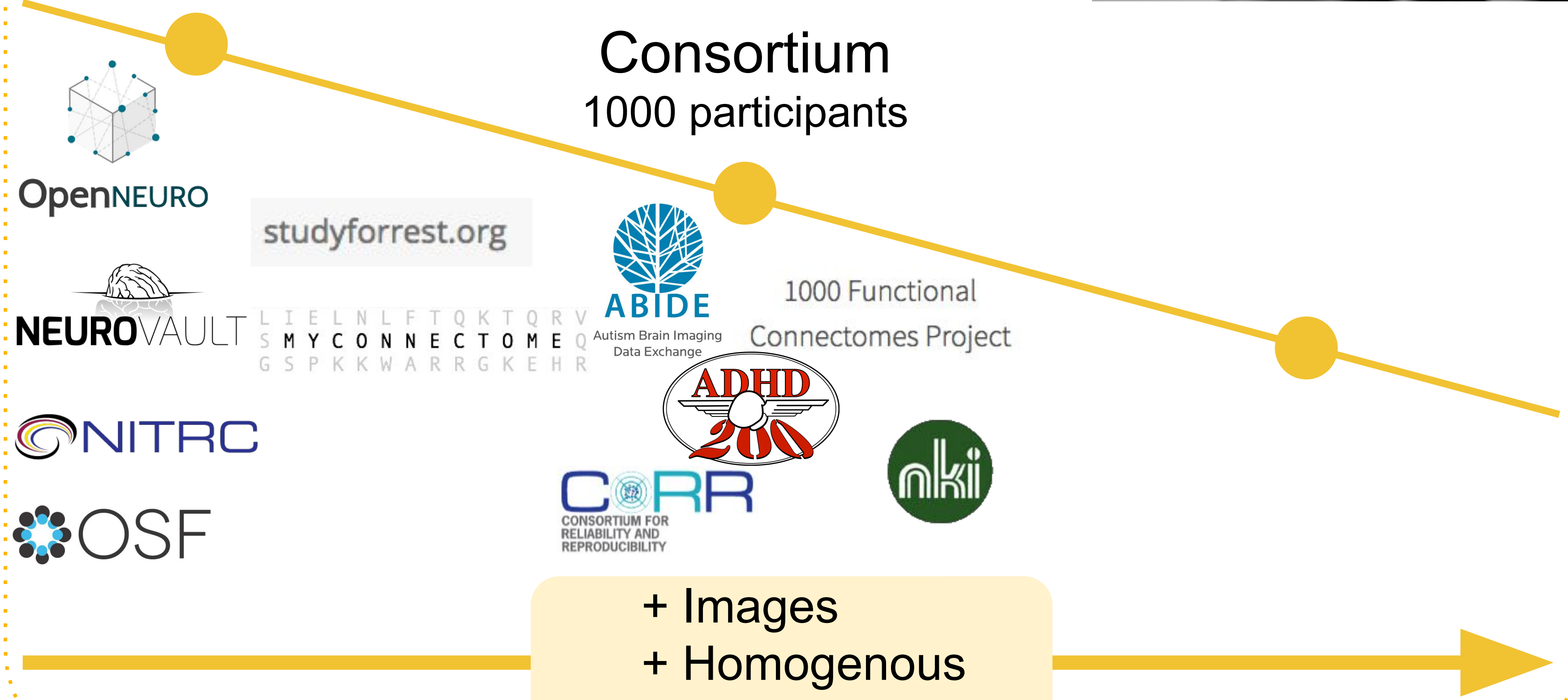
Open data

Unique study

30 participants

Consortium

1000 participants


OpenNEURO
studyforrest.org
NEUROVAULT
L I E L N L F T Q K T Q R V
S M Y C O N N E C T O M E Q
G S P K K W A R R G K E H R
NITRC
OSF
ABIDE
Autism Brain Imaging Data Exchange
1000 Functional Connectomes Project
ADHD 200
CoRR
CONSORTIUM FOR RELIABILITY AND REPRODUCIBILITY
nki

+ Images
+ Homogenous
- Datasets

Credits : Brains, Neil Conway, Flickr (CC BY 2.0)

Open data

Unique study
30 participants

Consortium
1000 participants

Cohort
100 000 participants

OpenNEURO

studyforrest.org

NEUROVAULT

L I E L N L F T Q K T Q R V
S M Y C O N N E C T O M E Q
G S P K K W A R R G K E H R

NITRC

OSF

ABIDE
Autism Brain Imaging
Data Exchange

1000 Functional
Connectomes Project

ADHD 200

CoRR
CONSORTIUM FOR
RELIABILITY AND
REPRODUCIBILITY

nki

HUMAN
Connectome
PROJECT

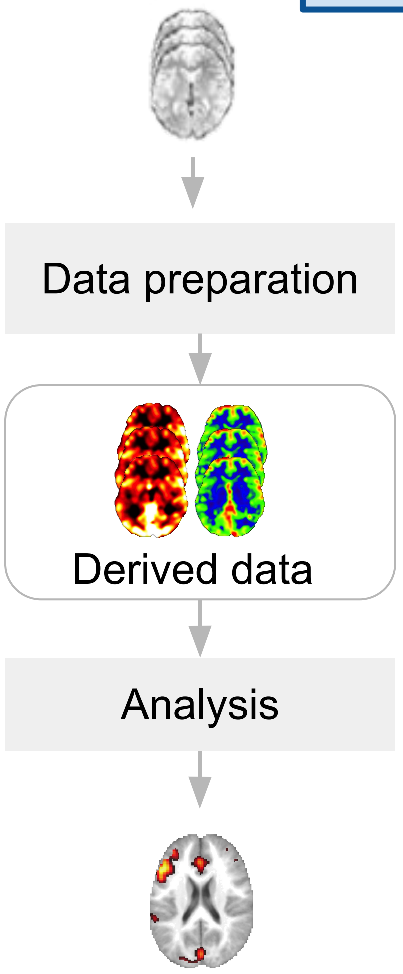
biobank uk
Improving the health of future generations

+ Images
+ Homogenous
- Datasets

Credits : Brains, Neil Conway, Flickr (CC BY 2.0)

Fixing the **reproducibility** issue

Irreproducible with... Different Methods



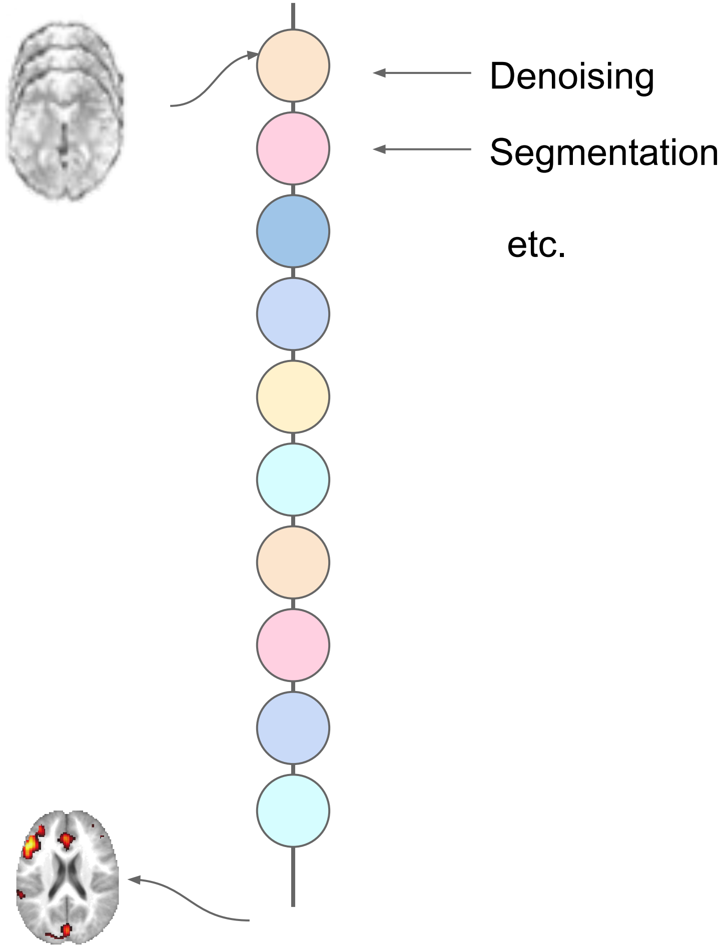
Data preparation

Derived data

Analysis

Fixing the **reproducibility** issue

Irreproducible with... Different Methods



Denoising

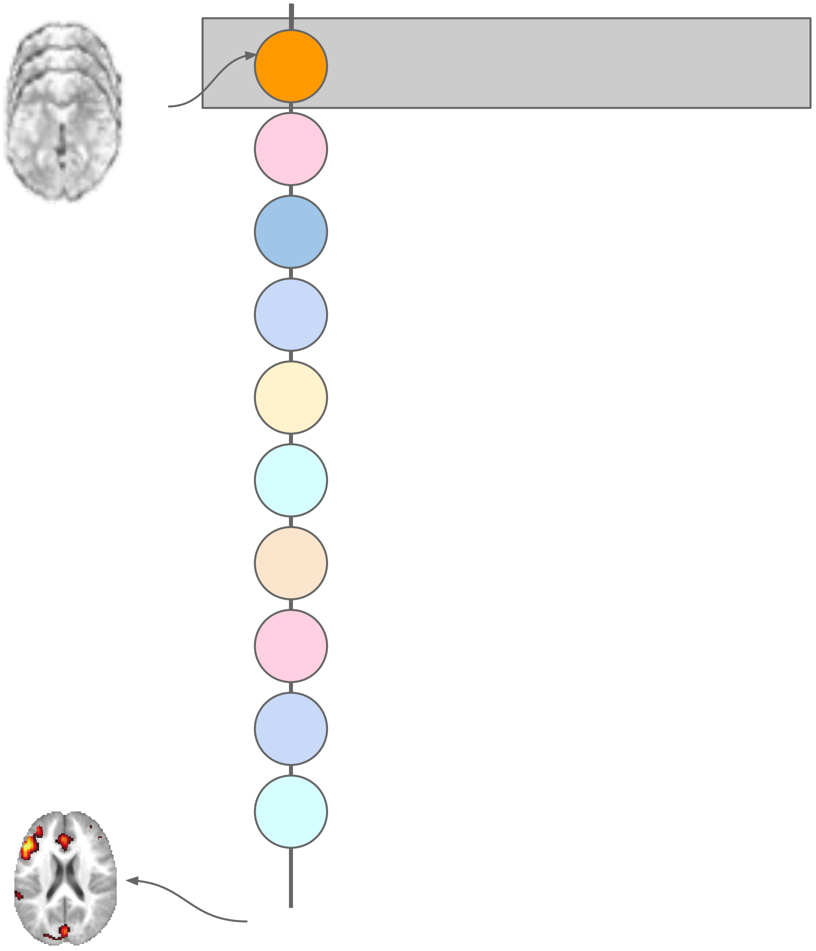
Segmentation

etc.

Fixing the **reproducibility** issue

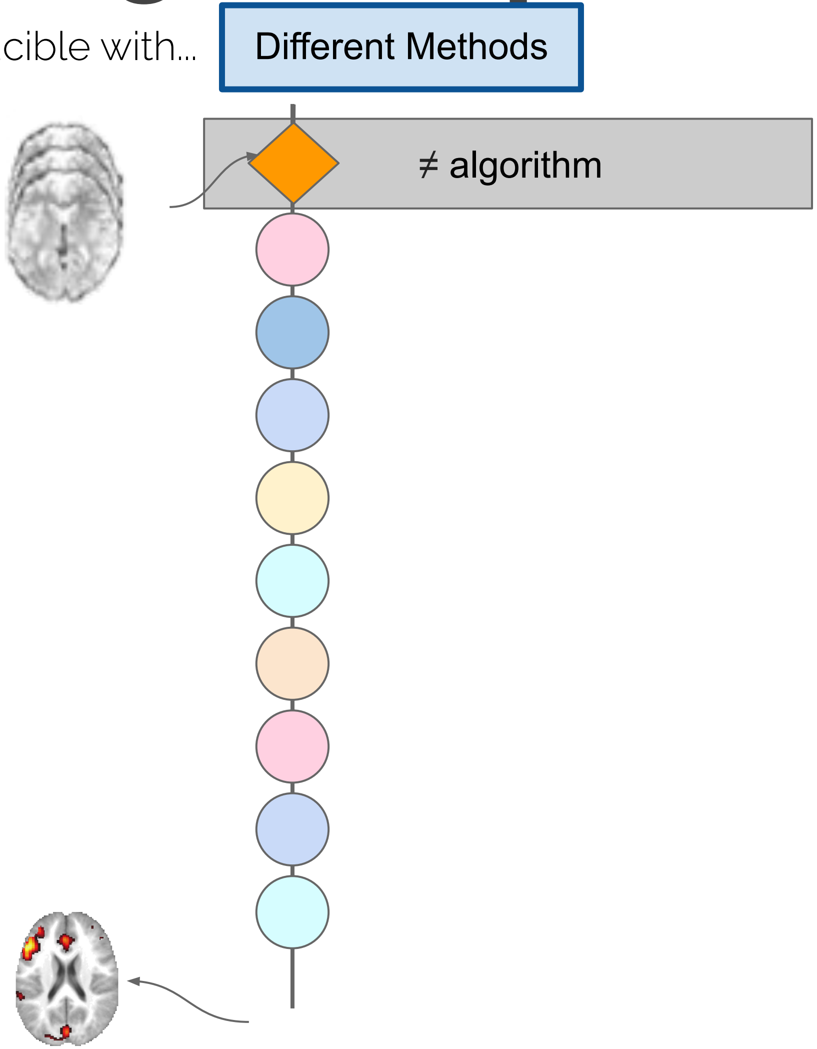
Irreproducible with...

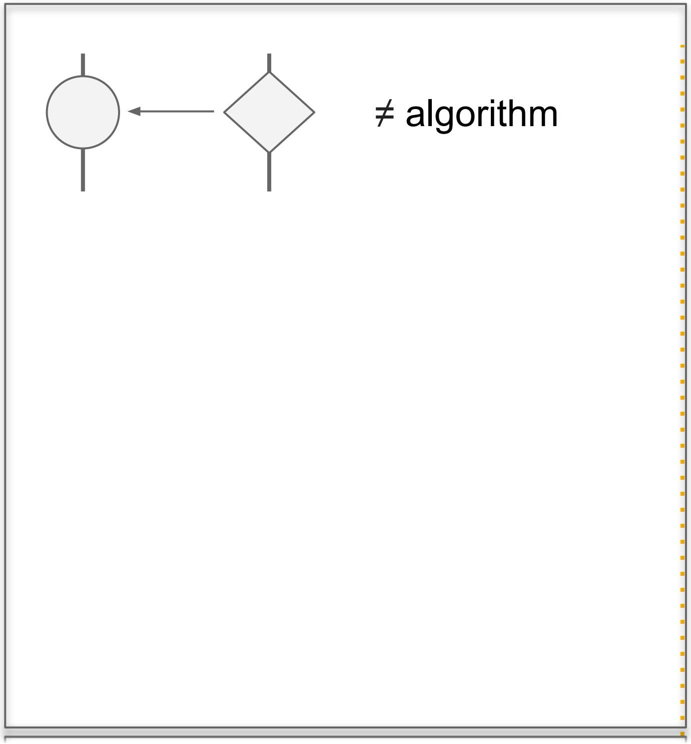
Different Methods



Fixing the **reproducibility** issue

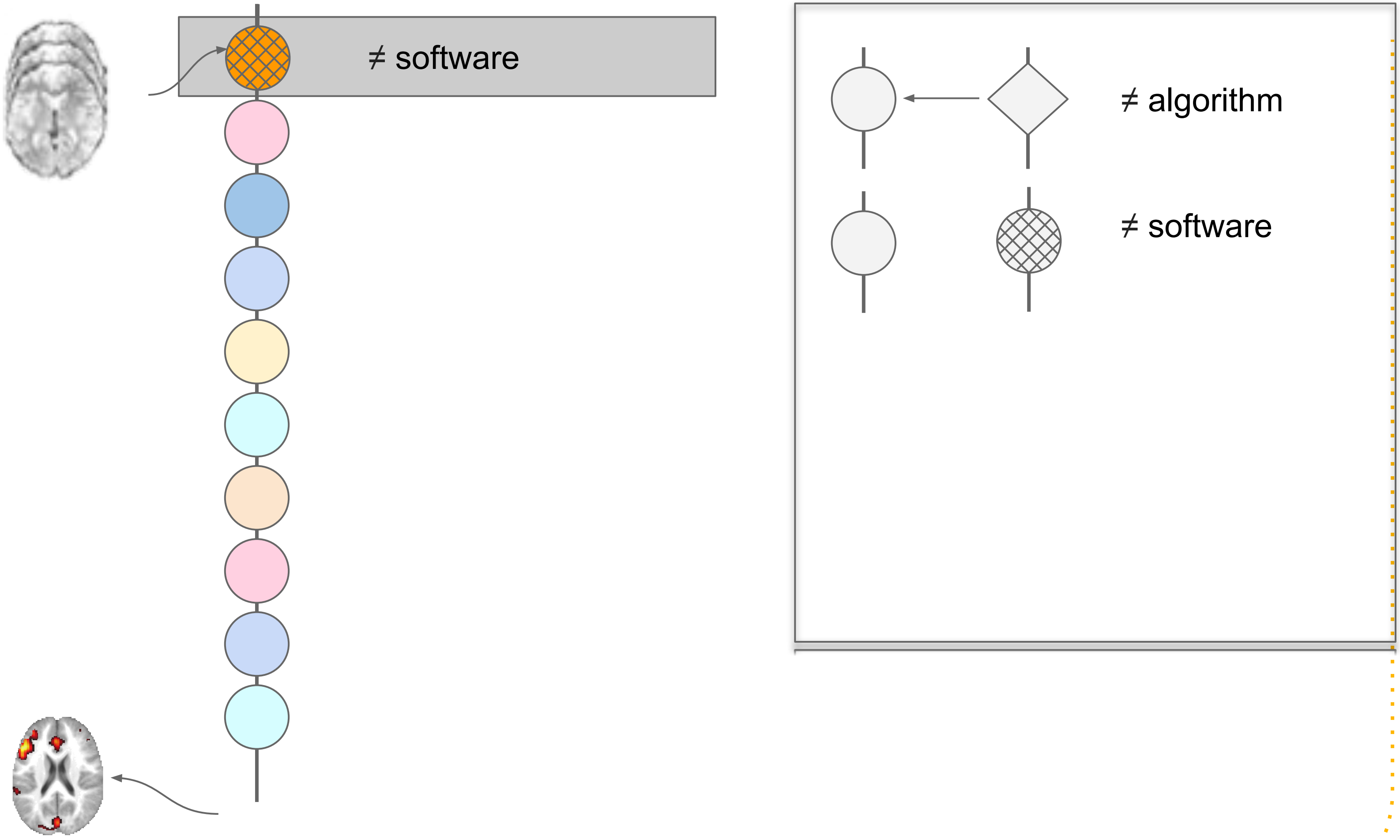
Irreproducible with...





Fixing the **reproducibility** issue

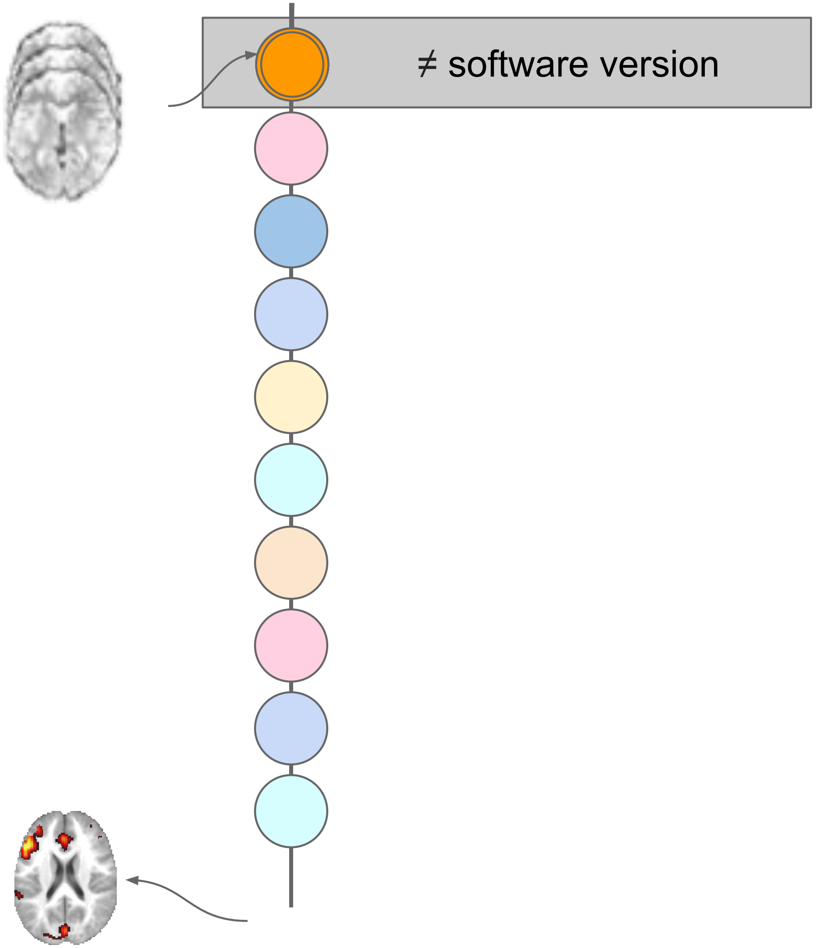
Irreproducible with... Different Methods

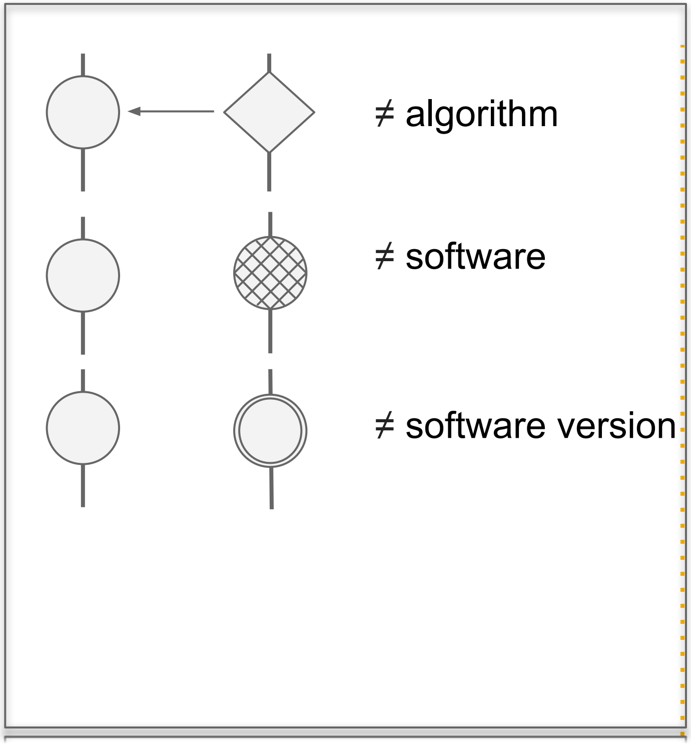


Fixing the **reproducibility** issue

Irreproducible with...

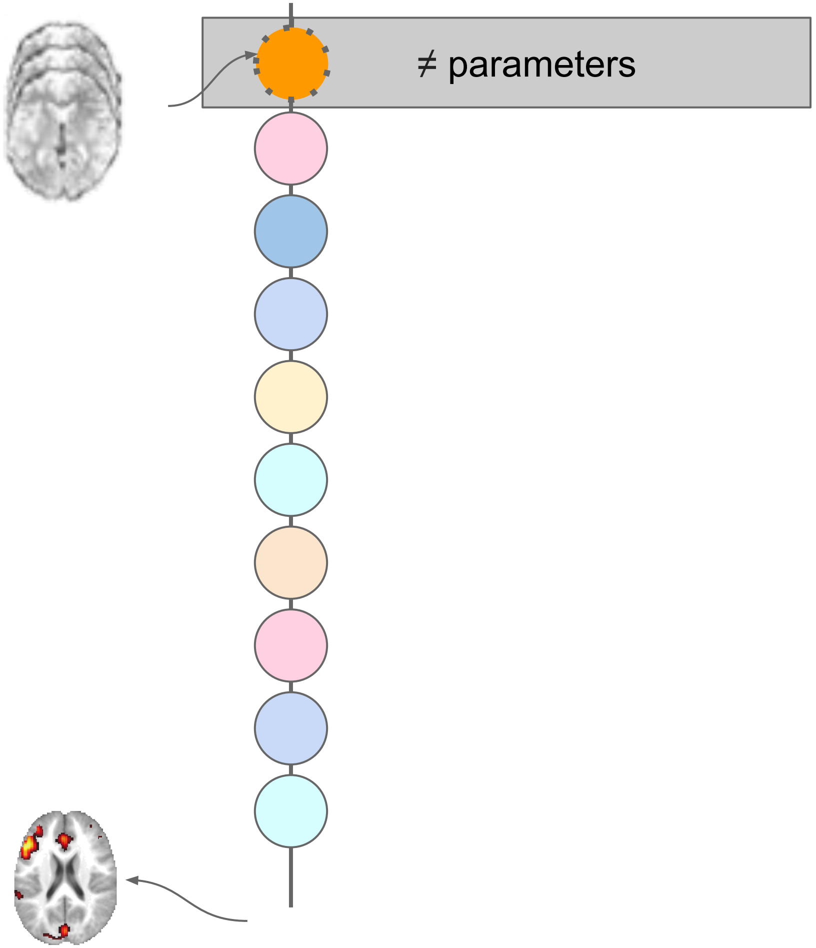
Different Methods



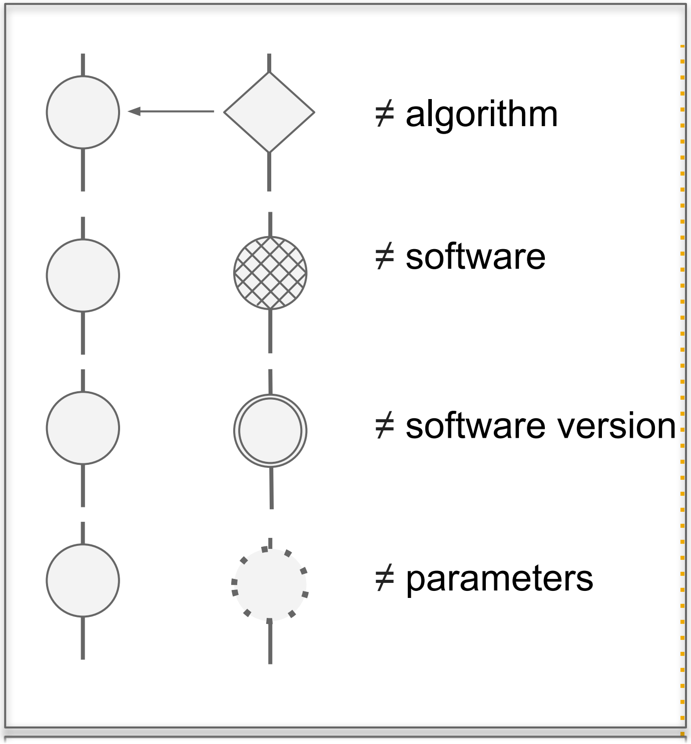


Fixing the **reproducibility** issue

Irreproducible with... **Different Methods**



≠ parameters



≠ algorithm

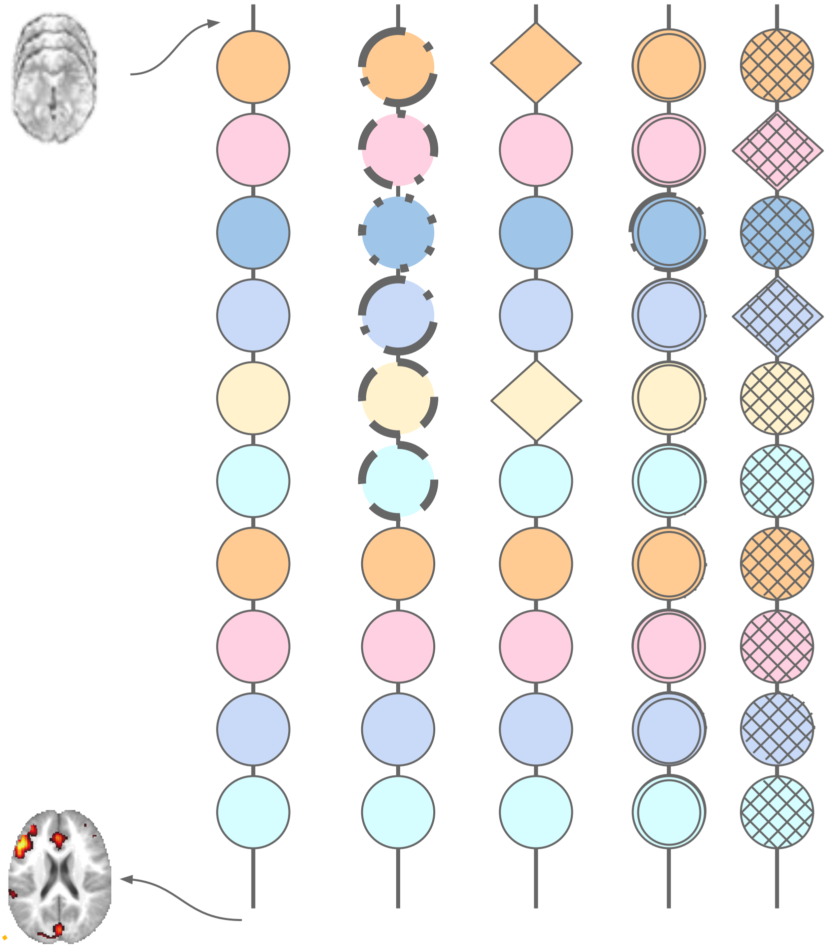
≠ software

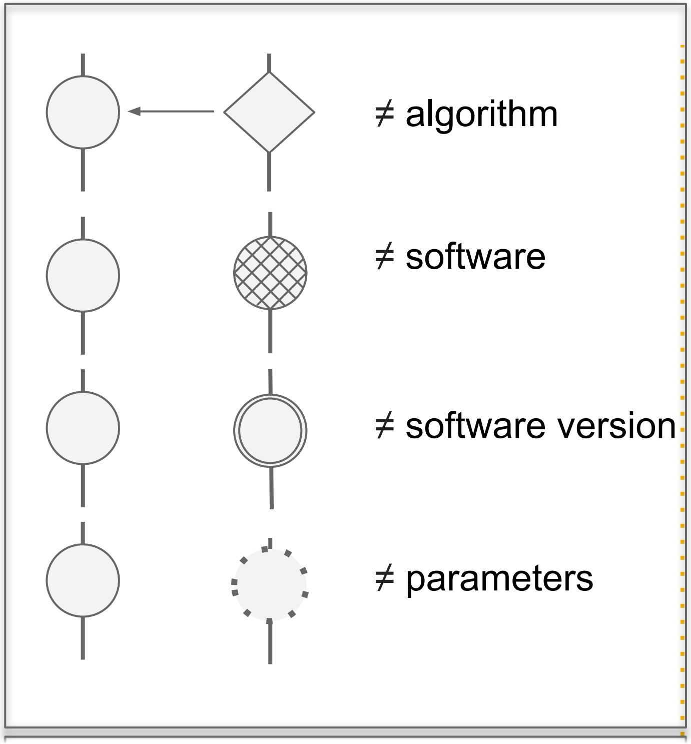
≠ software version

≠ parameters

Fixing the **reproducibility** issue

Irreproducible with... Different Methods





≠ algorithm

≠ software

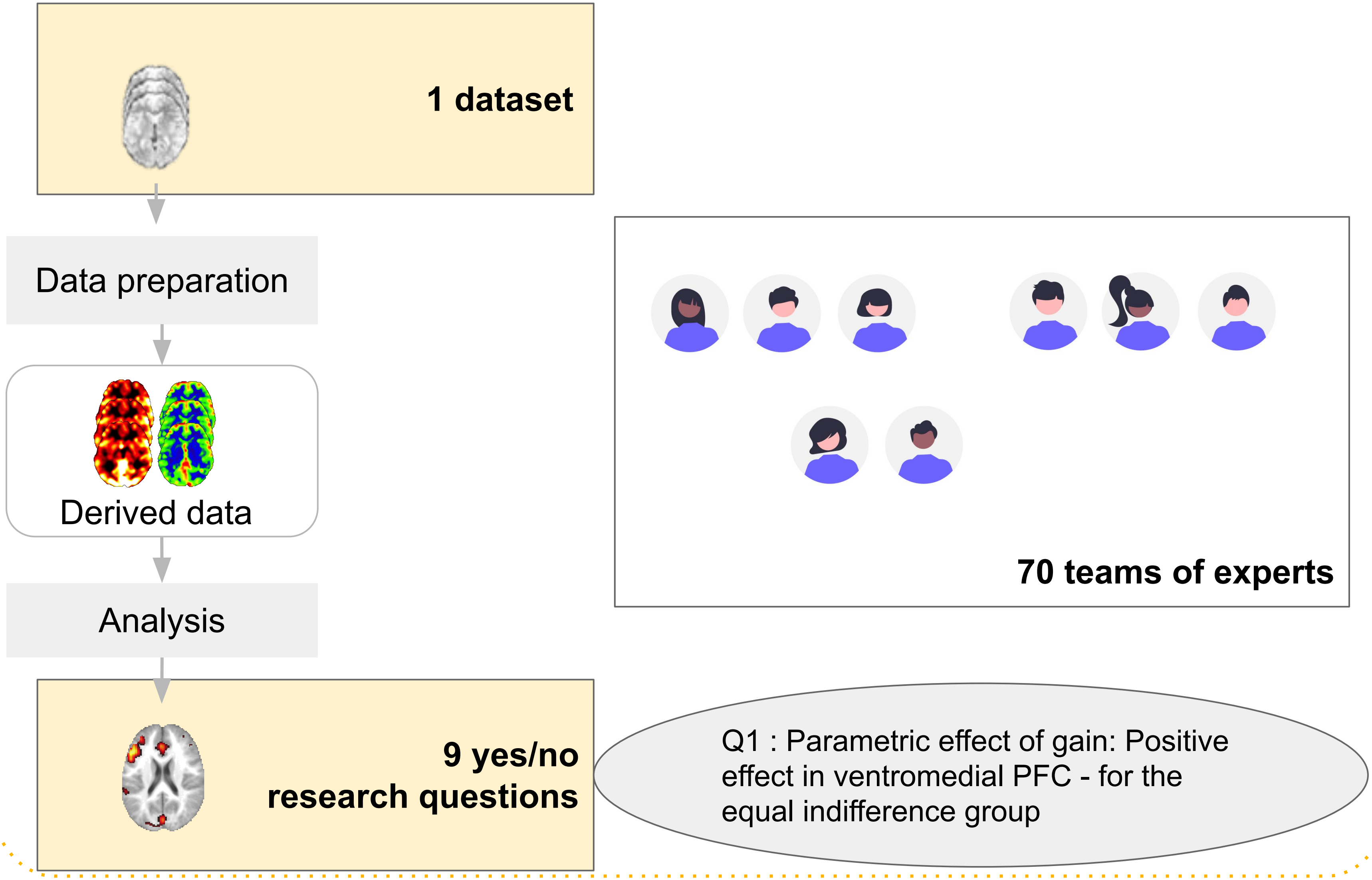
≠ software version

≠ parameters

A family of acceptable pipelines

100 000+ combinations

Many analysts project : NARPS



1 dataset

Data preparation

Derived data

Analysis

9 yes/no research questions

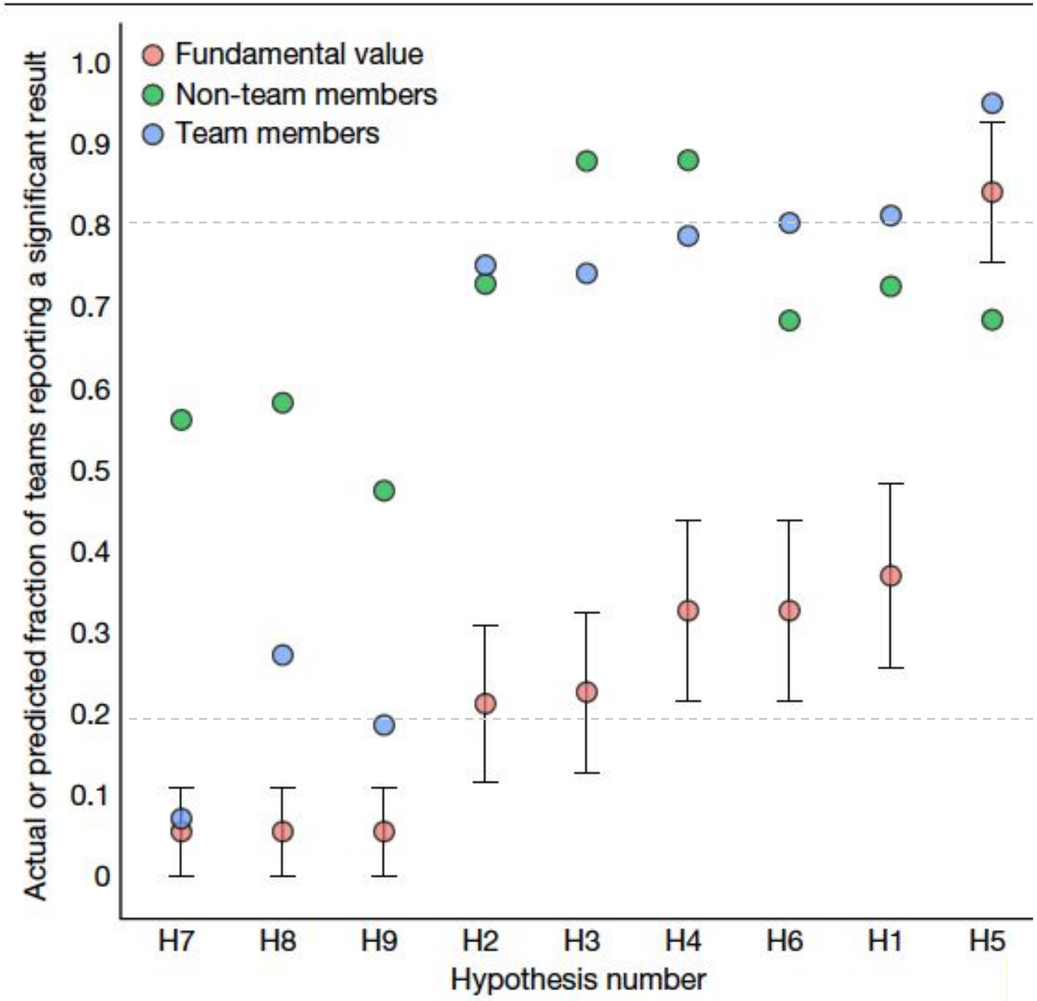
70 teams of experts

Q1 : Parametric effect of gain: Positive effect in ventromedial PFC - for the equal indifference group

Image credits : Teams, undraw.co.

[Botvinik-Nezer et. al, Nature 2020]

Many analysts project : NARPS



80% of teams agree on
Presence of significant finding

Contradictory results

(Teams disagree)

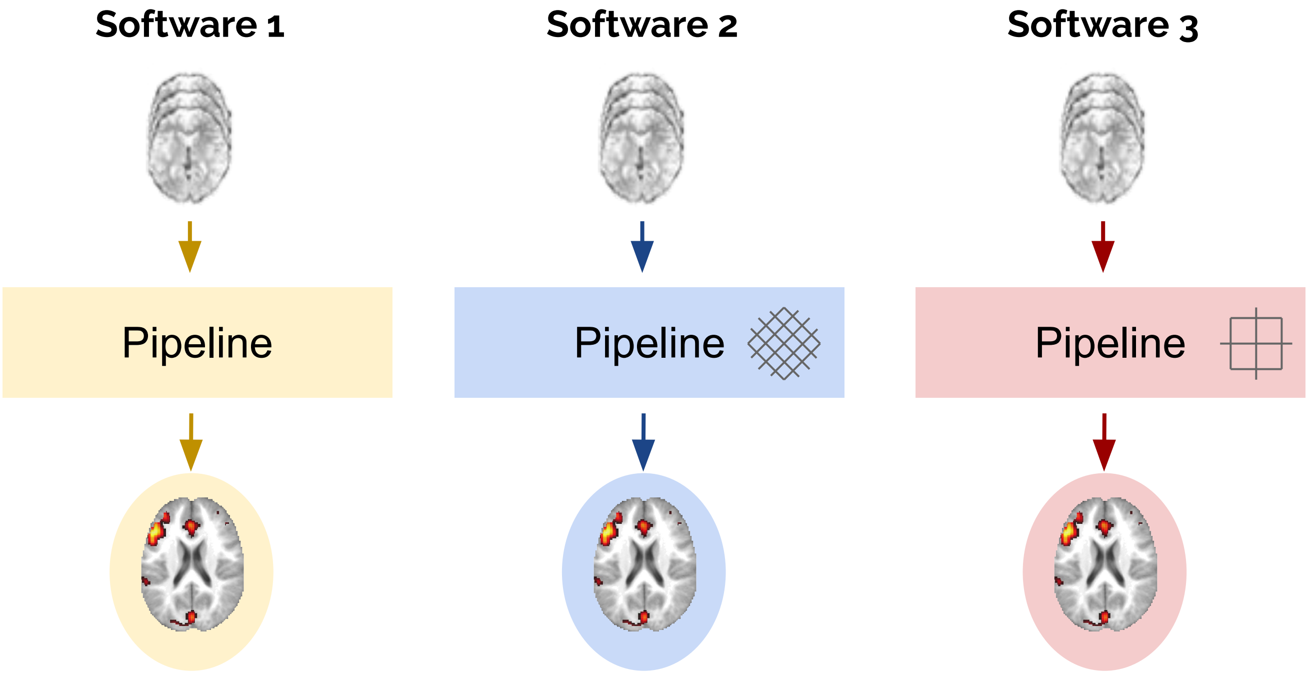
80% of teams agree on
No significant finding

[Botvinik-Nezer et. al, Nature 2020]

Variability across software

- Reproduced 3 published functional MRI studies
- Using 3 different software

Alex Bowring Tom Nichols



Software 1

Pipeline

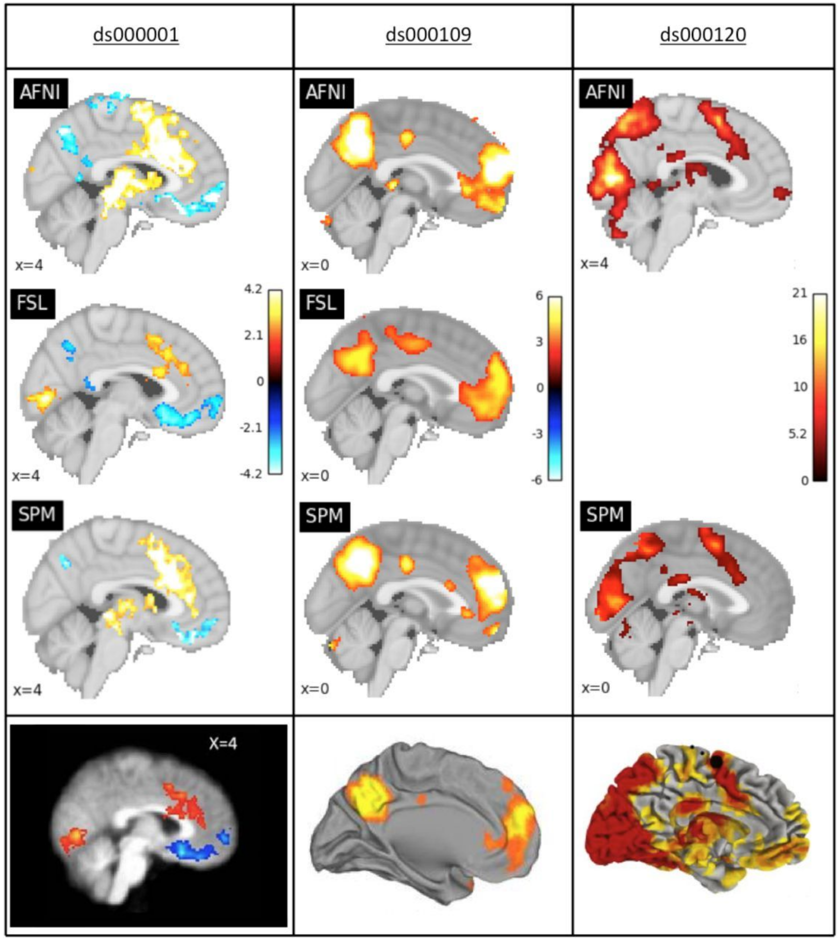
Software 2

Pipeline

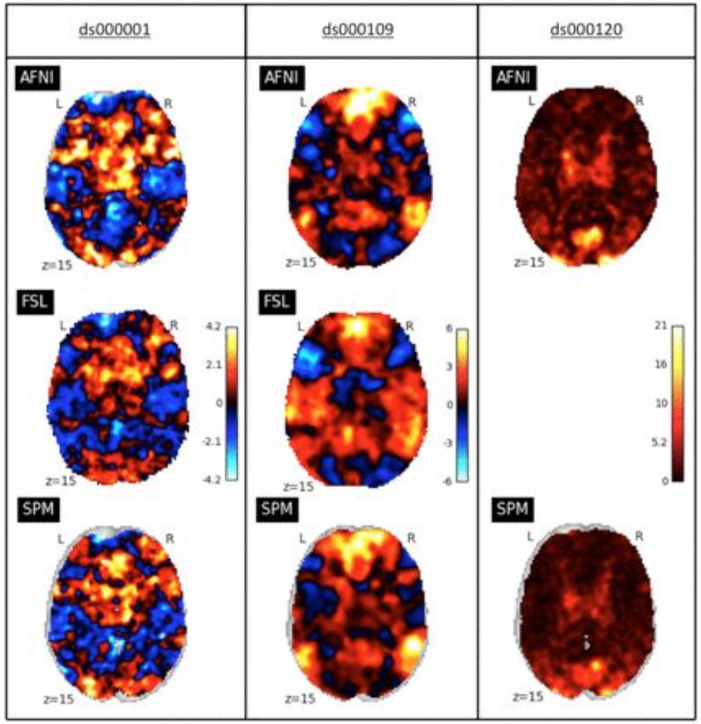
Software 3

Pipeline

Software Comparison Project



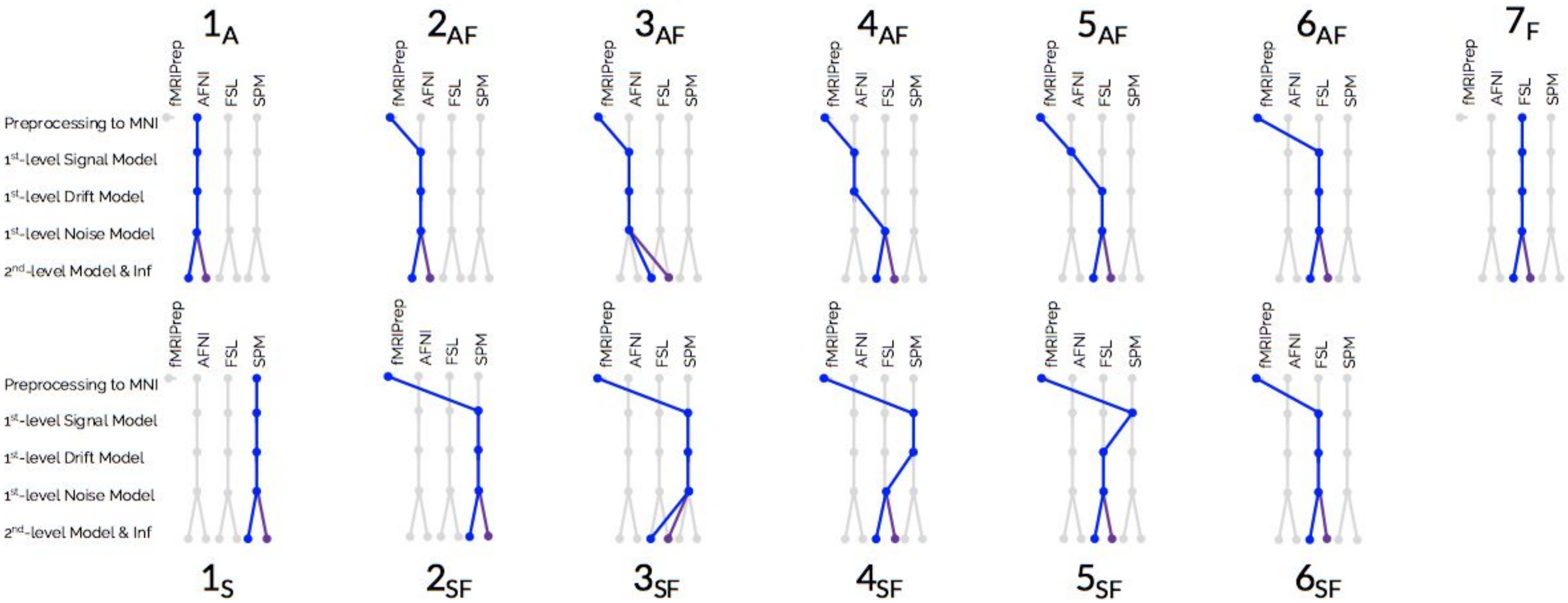
Comparison of the final results



Comparison of the statistic maps

[Bowring et. al, HBM 2019]

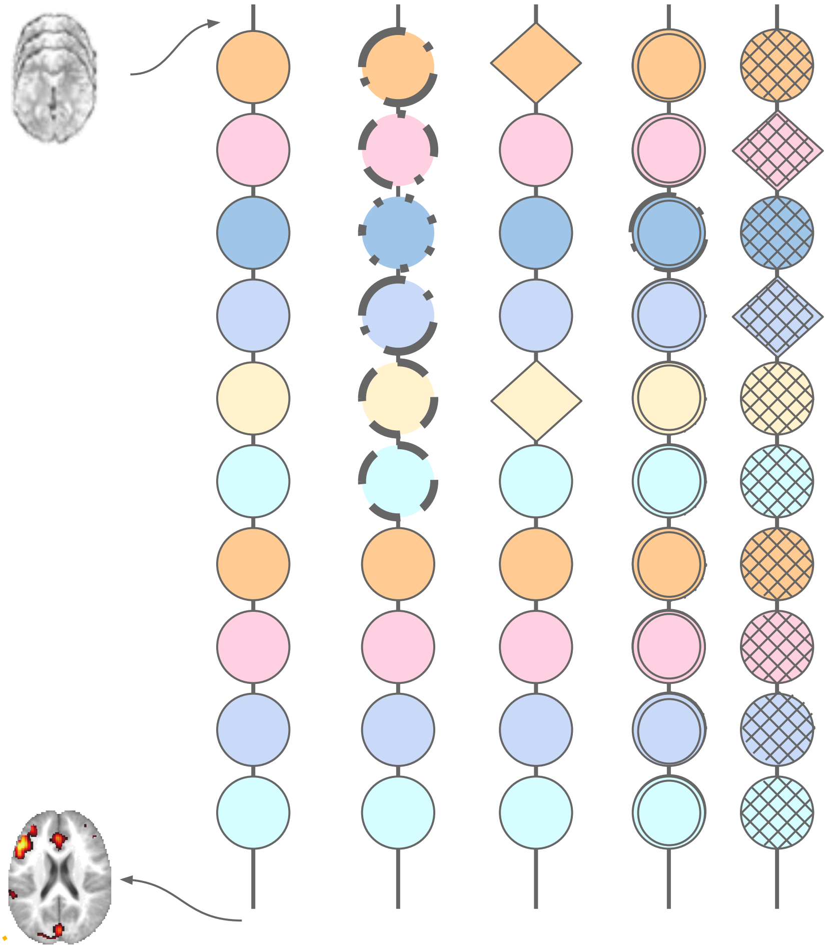
Software Comparison Project 2

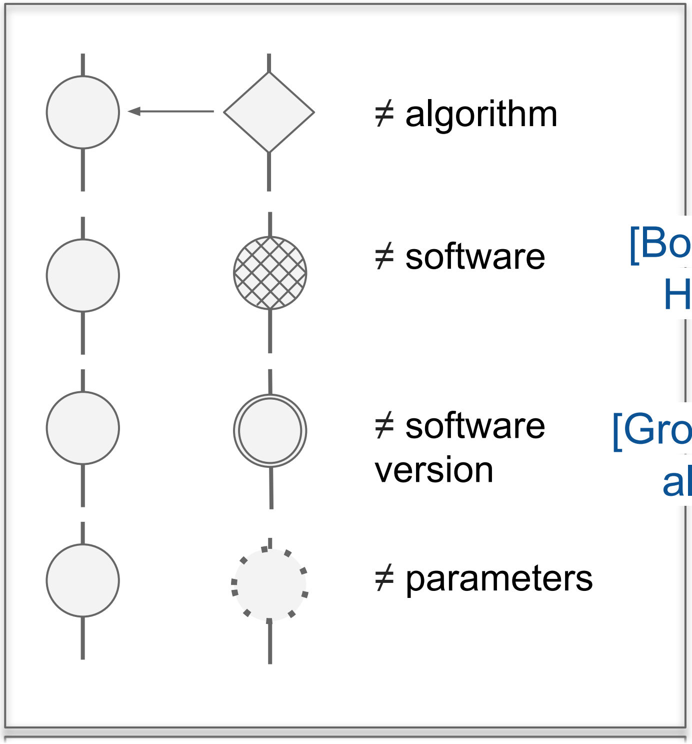


[Bowring et. al, HBM 2021]

Fixing the **reproducibility** issue

Irreproducible with... Different Methods





[Bowring et. al, HBM 2021]

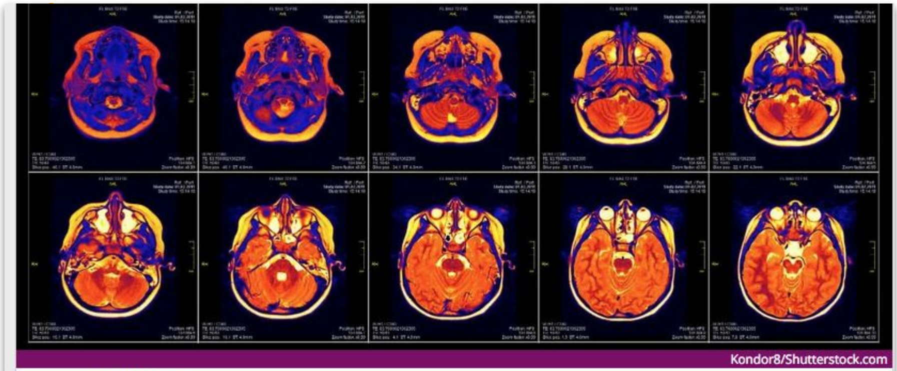
[Gronenschild et. al, PlosOne 2012]

Explanations ???

Fixing the **reproducibility** issue

Irreproducible with... Different Methods

Explanation 1: There is a bug!

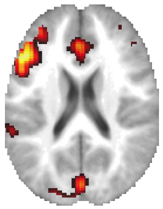


HUMANS

A Bug in FMRI Software Could Invalidate 15 Years of Brain Research

BEC CREW 6 JULY 2016

There could be a very serious problem with the past 15 years of researc human brain activity, with a new study suggesting that a bug in fMRI sof could invalidate the results of some 40,000 papers.

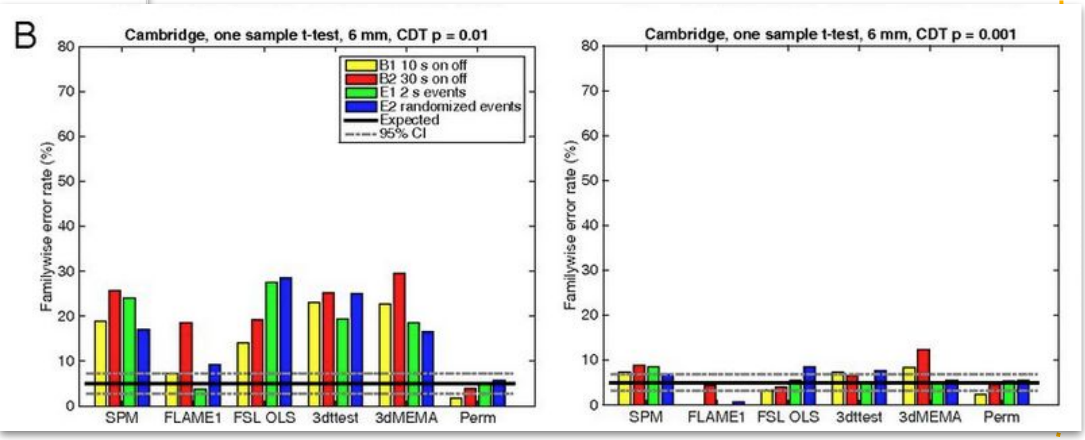


No ground truth to most neuroimaging problems.

Validation is a challenge

Multiple levels:

- **Inadequate methodology** (assumption violations)
- **Boggus implementation**



Science Alert, "A Bug in FMRI Software Could Invalidate 15 Years of Brain Research", By BEC CREW. 2016.

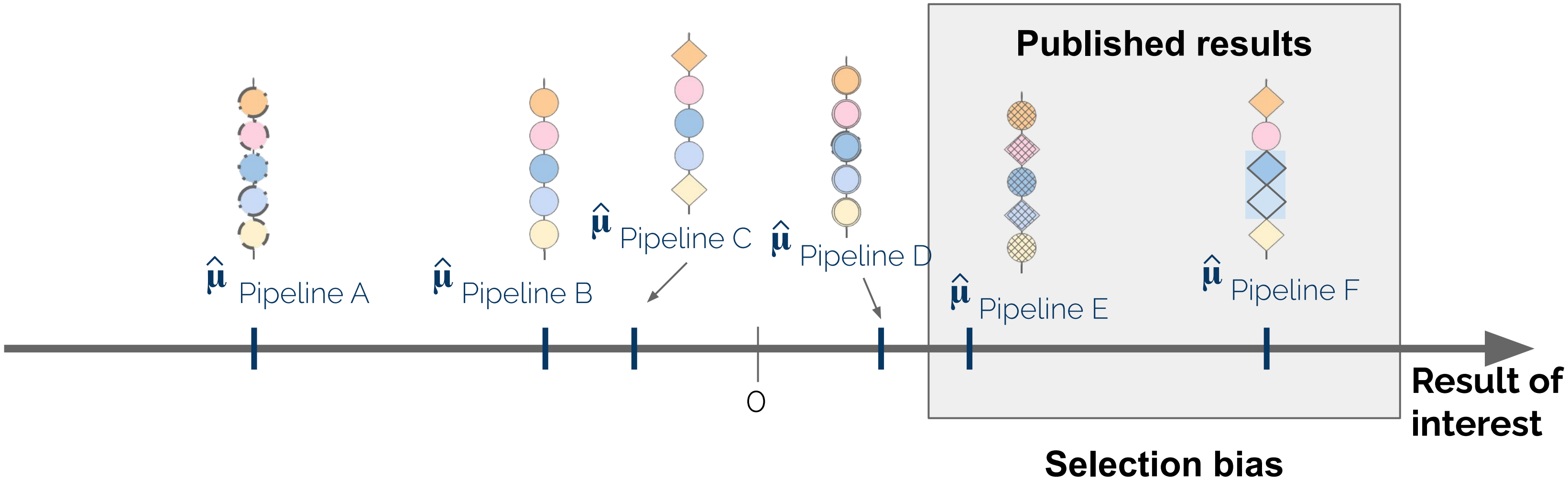
[Eklund et. al, PNAS 2016]

Fixing the **reproducibility** issue

Irreproducible with... Different Methods

Explanation 2: False positive finding

Vibration of effects

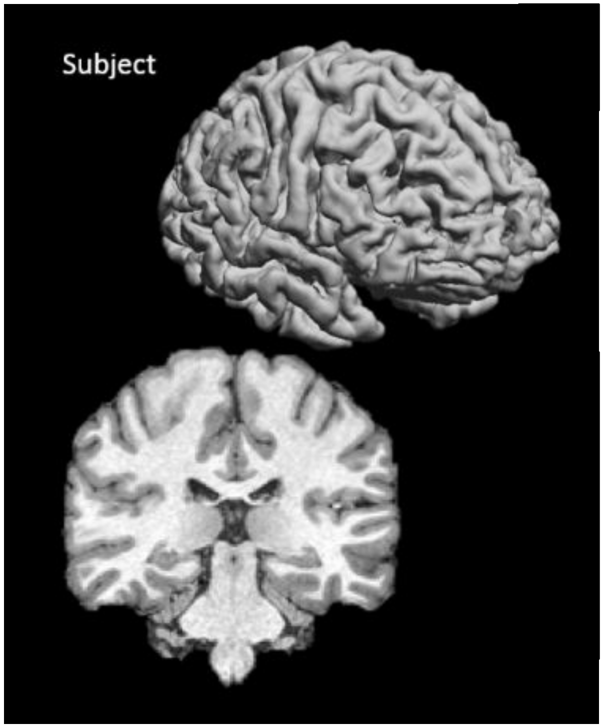


Emerging solutions: Multiverse analyses…

Fixing the **reproducibility** issue

Irreproducible with... Different Methods

Explanation 3: Different pipelines inform us in different ways



Image

Solutions: Finding common ground for comparisons…

Image credit : Brain suite.

On our way to study the ***"pipeline space"***

- **Huge pipeline space** : 100 000+ combinaisons
- Which pipelines are **suitable to answer a given problem**?
  - Expert knowledge
  - But also dependent on characteristics of the dataset under study…
- Which pipelines are **used in the community**? Lack of transparency.
  - Very coarse descriptions in scientific papers, and still limited code sharing.
- Even when code is shared, it is difficult to **compare pipeline**.
  - Which pipelines are "equivalent"?
    - Implementation of the same method in two different software packages might "hide" crucial implementation details.
- And many more…

Experimentation users-group [Xug] Meetings

March 8, 2022

Beyond computational reproducibility: what does it **mean** for **neuroimaging** results to be **irreproducible**?

Camille Maumet

Thank you!

Credit: Sculpture by Malin Bjornsdotter “Cerebia”, OHBM Brain Art SIG; Presentation template by SlidesCarnival, adapted

@cmaumet