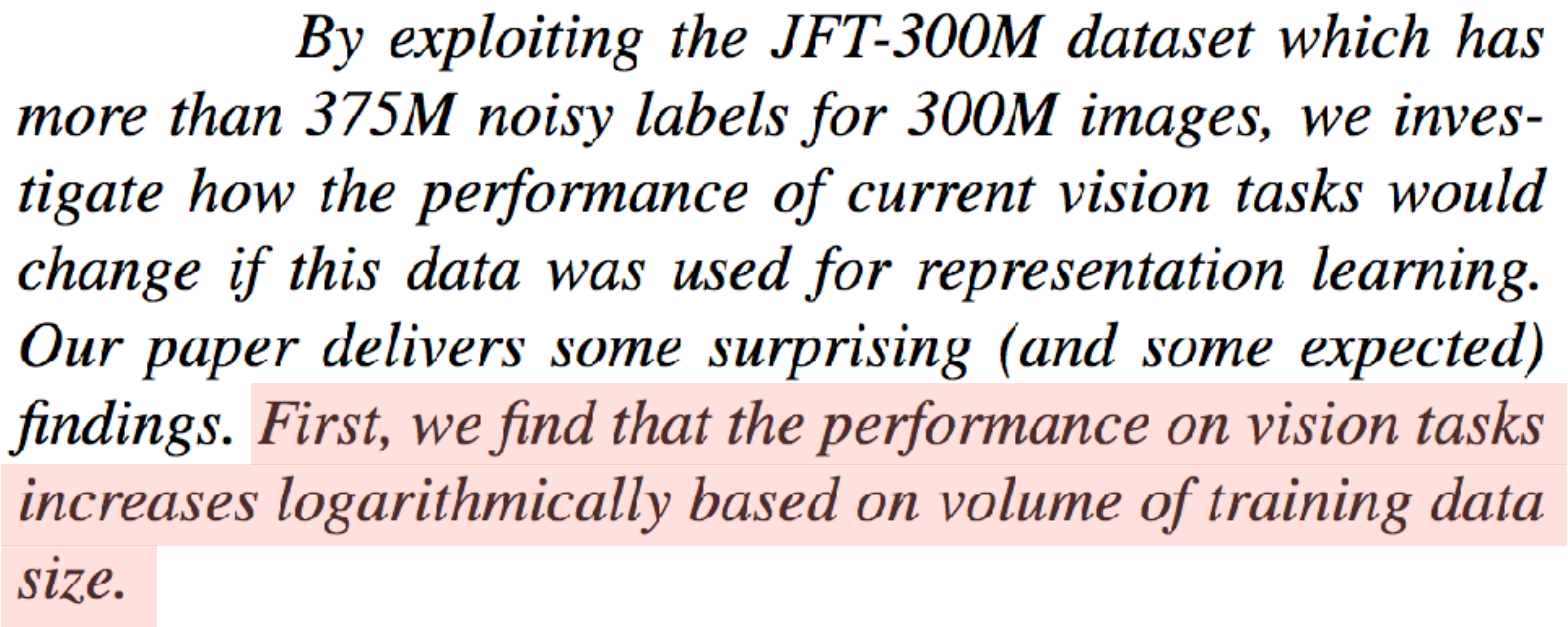
Sharing Data: Why, How, When?

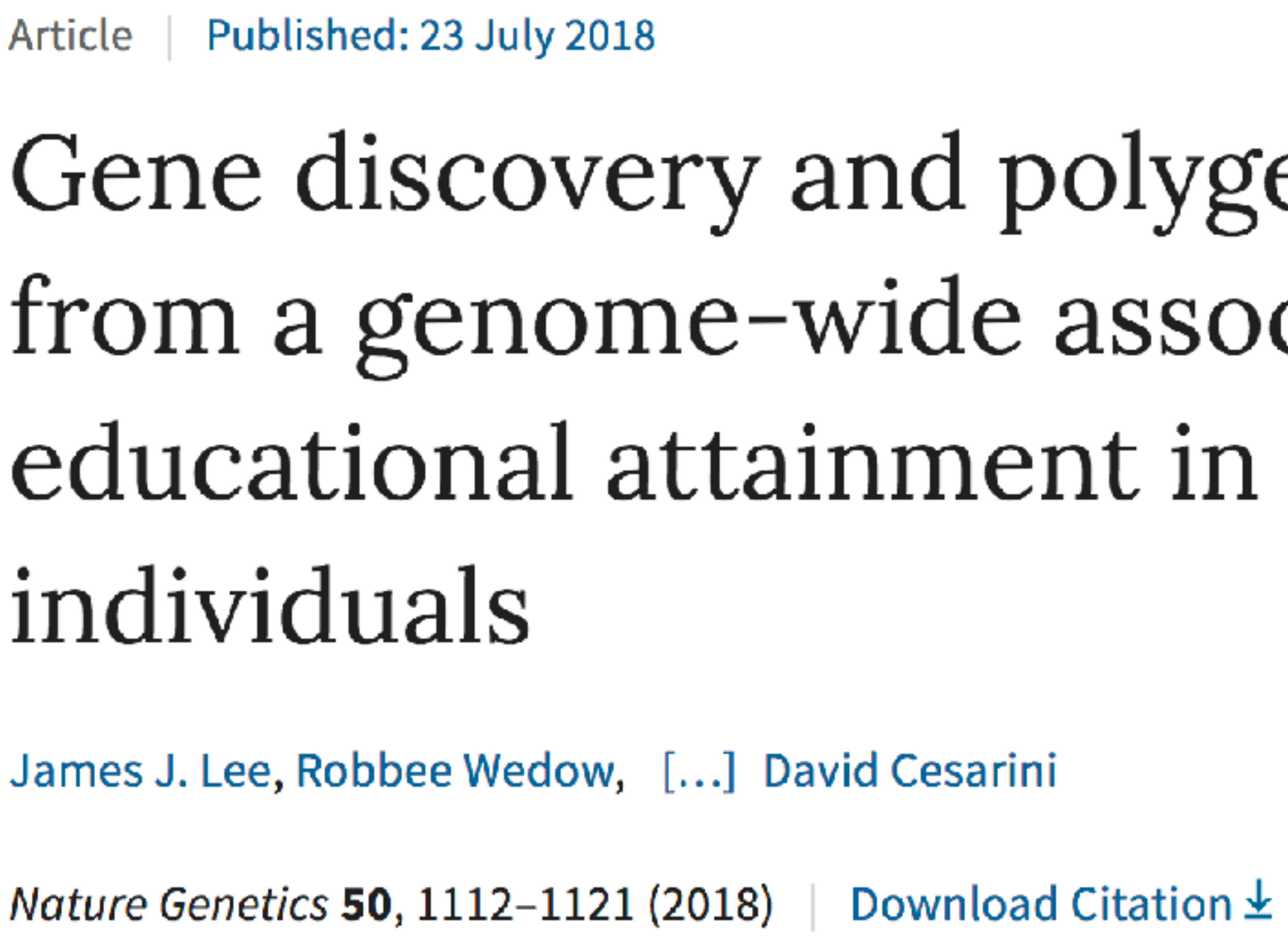
Bastian Greshake Tzovaras
(@gedankenstuecke)

Revisiting Unreasonable Effectiveness of Data in Deep Learning Era



By exploiting the JFT-300M dataset which has more than 375M noisy labels for 300M images, we investigate how the performance of current vision tasks would change if this data was used for representation learning. Our paper delivers some surprising (and some expected) findings. First, we find that the performance on vision tasks increases logarithmically based on volume of training data size.

It’s not only vision tasks…



Article | Published: 23 July 2018

Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals

James J. Lee, Robbee Wedow, […] David Cesarini

Nature Genetics **50**, 1112–1121 (2018) | Download Citation ⤓

tl;dr: data is pretty cool

So, data is pretty powerful, but how does it relate to reproducibility?

Which of these do we ultimately want?

	Same Data/Experimental System	**Different Data/Experiment System**
Same Methods	Reproducibility **Does this data/method combination do what someone said it does?**	Replicability **Was previous data just quirky or do we consistently find X?**
Different Methods	Robustness **Does the data actually say X if looking at it differently?**	Generalizibility **Multiple methods & data sets agree on X 🎉**

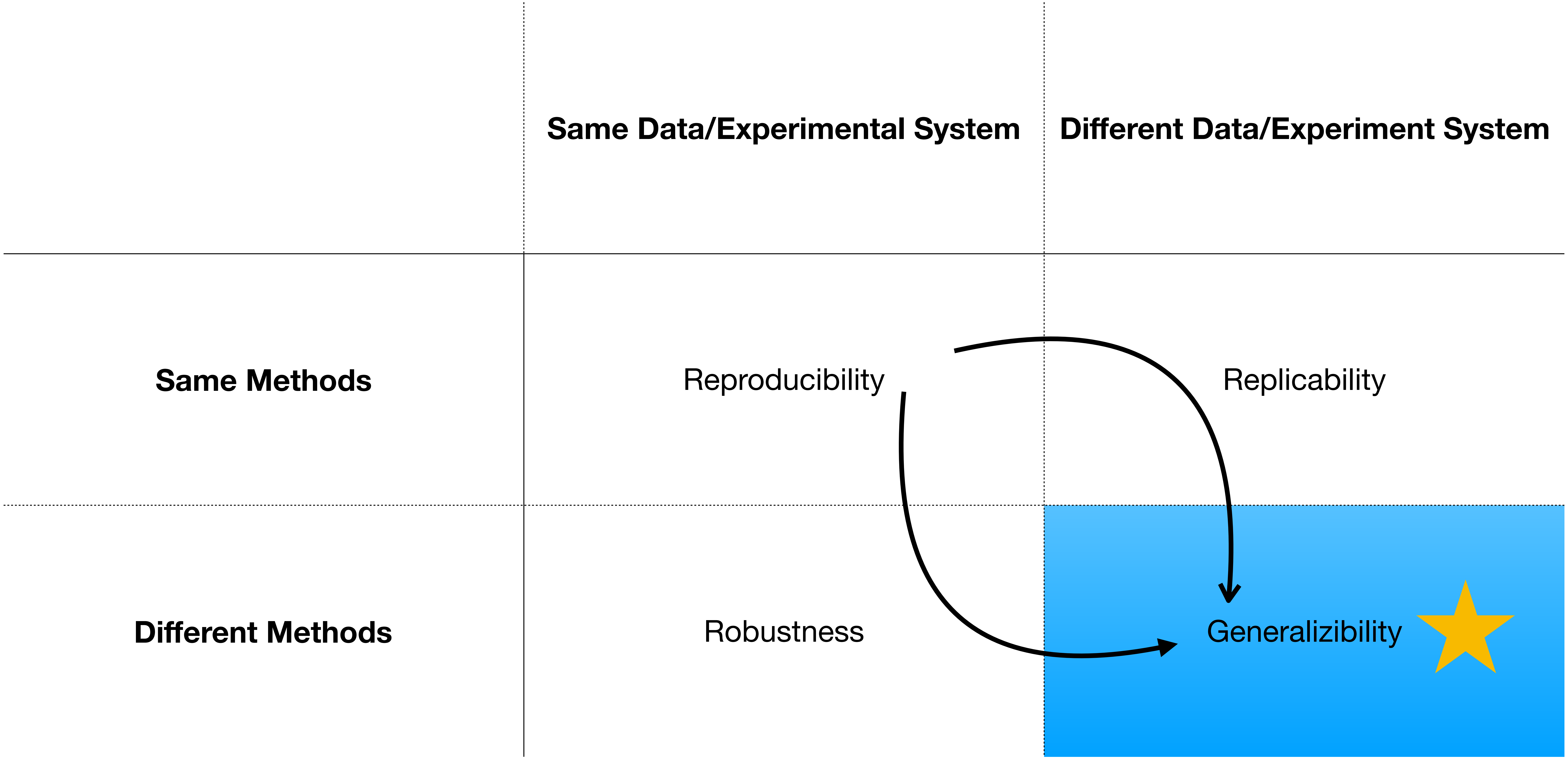
adapted from doi:10.6084/m9.figshare.5537101 & 10.1128/mBio.00525-18

Which of these do we ultimately want?

	Same Data/Experimental System	**Different Data/Experiment System**
Same Methods	Reproducibility	Replicability
Different Methods	Robustness	Generalizibility ★

How do we get there?

How do we get there?



	Same Data/Experimental System	Different Data/Experiment System
Same Methods	Reproducibility	Replicability
Different Methods	Robustness	Generalizability ★

Checking both methods & data are needed to generate generalizable knowledge*

***Or at least it speeds up building this knowledge**

Without the original data

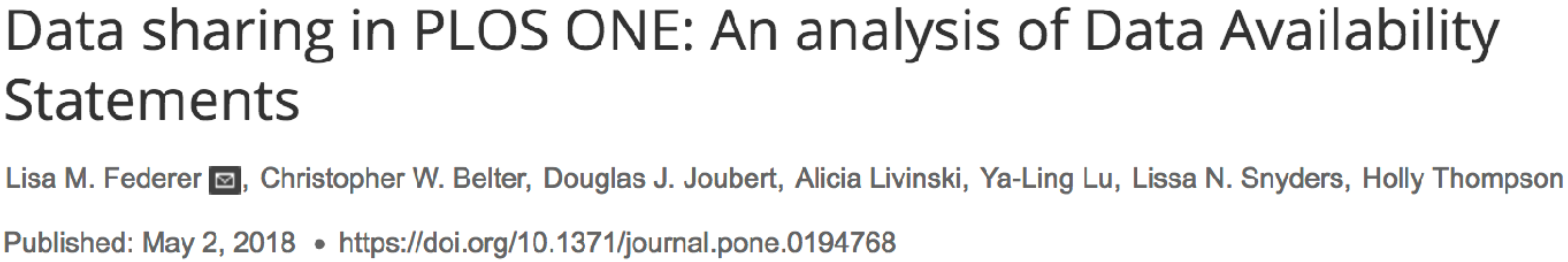
	Same Data/ Experimental System	**Different Data/ Experiment System**
Same Methods	~~Reproducibility~~	Replicability
Different Methods	~~Robustness~~	Generalizibility

- Secondary effect: Replication gets harder too, as no certainty whether original method even worked as expected 😭

- in the end: you need to collect twice as many data in order to get to replicability

The solution:
make your data available

The reality



Data sharing in PLOS ONE: An analysis of Data Availability Statements

Lisa M. Federer, Christopher W. Belter, Douglas J. Joubert, Alicia Livinski, Ya-Ling Lu, Lissa N. Snyders, Holly Thompson

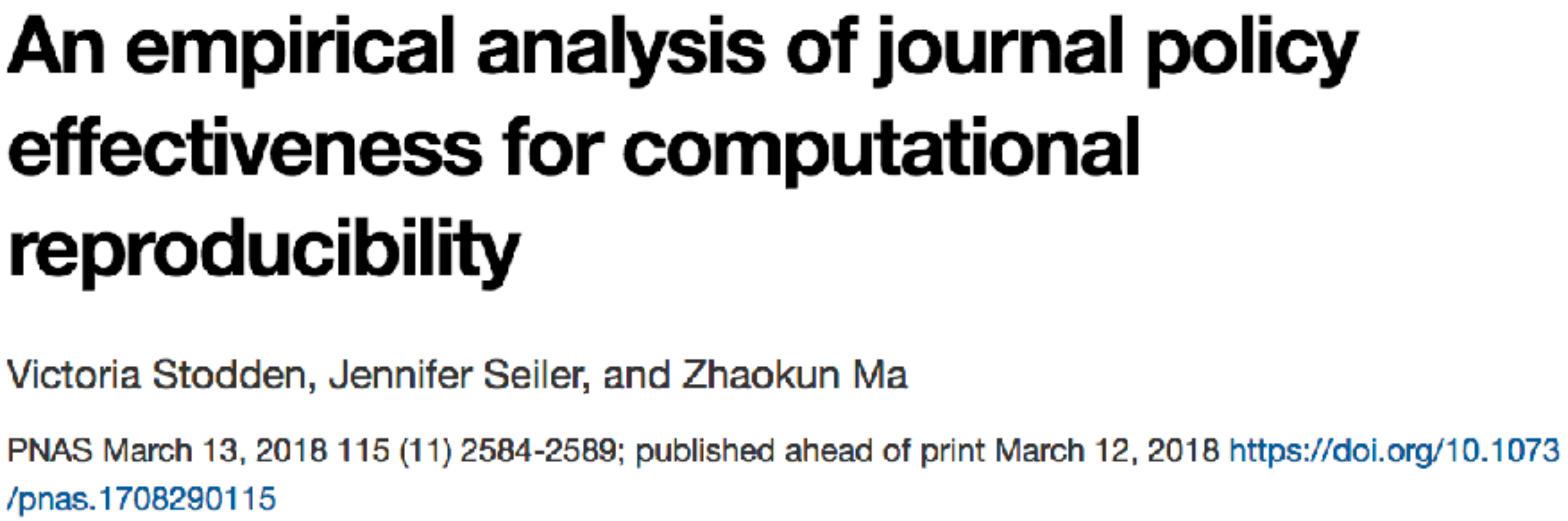
Published: May 2, 2018 • https://doi.org/10.1371/journal.pone.0194768

In this study, we evaluate the extent to which authors have complied with this policy by analyzing Data Availability Statements from 47,593 papers published in PLOS ONE between March 2014 (when the policy went into effect) and May 2016. Our analysis shows that compliance with the policy has increased, with a significant decline over time in papers that did not include a Data Availability Statement.

However, only about 20% of statements indicate that data are deposited in a repository, which the PLOS policy states is the preferred method.

make data available upon request?

The reality



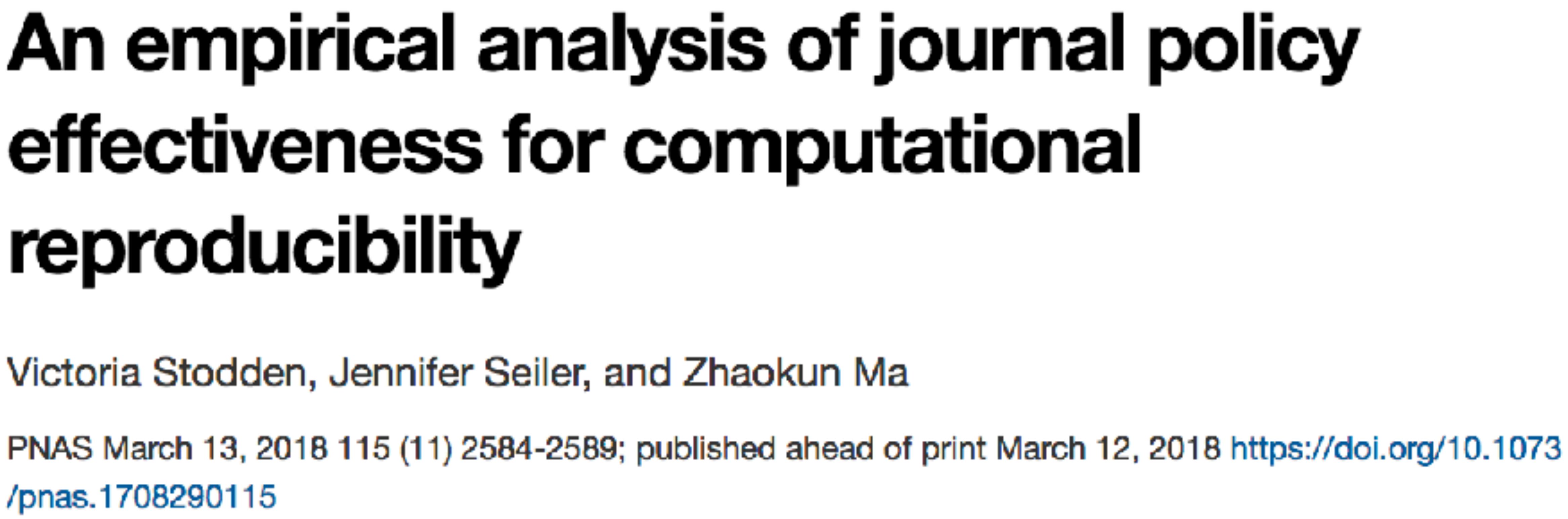
An empirical analysis of journal policy effectiveness for computational reproducibility

Victoria Stodden, Jennifer Seiler, and Zhaokun Ma

PNAS March 13, 2018 115 (11) 2584-2589; published ahead of print March 12, 2018 https://doi.org/10.1073/pnas.1708290115

This work evaluates the effectiveness of journal policy that requires the data and code necessary for reproducibility be made available post publication by the authors upon request. […] We chose a random sample of 204 scientific papers published in the journal Science after the implementation of their policy in February 2011. We found that we were able to obtain artifacts from 44% of our sample and were able to reproduce the findings for 26%.

The reality



An empirical analysis of journal policy effectiveness for computational reproducibility

Victoria Stodden, Jennifer Seiler, and Zhaokun Ma

PNAS March 13, 2018 115 (11) 2584-2589; published ahead of print March 12, 2018 https://doi.org/10.1073/pnas.1708290115

Responses when data/code was requested:

> *When you approach a PI for the source codes and raw data, you better explain who you are, whom you work for, why you need the data and what you are going to do with it.*

> *I have to say that this is a very unusual request without any explanation! Please ask your supervisor to send me an email with a detailed, and I mean detailed, explanation.*

> *The data files remains our property and are not deposited for free access. Please, let me know the purpose you want to get the file and we will see how we can help you.*

data available upon request

$\simeq$

data not available

Data should be FAIR

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**e-usable

https://www.force11.org/fairprinciples

figshare

zenodo

OSF

tl;dr: please make use of data repositories

NEUROVAULT

OpenNEURO BETA

RCSB PDB PROTEIN DATA BANK

PANGAEA.

How to choose a repository?

- it depends…
  - on data types,
  - data size,
  - and your academic field
- F1000Research has guidelines that are pretty comprehensive (for biosciences at large at least): https://f1000research.com/for-authors/data-guidelines

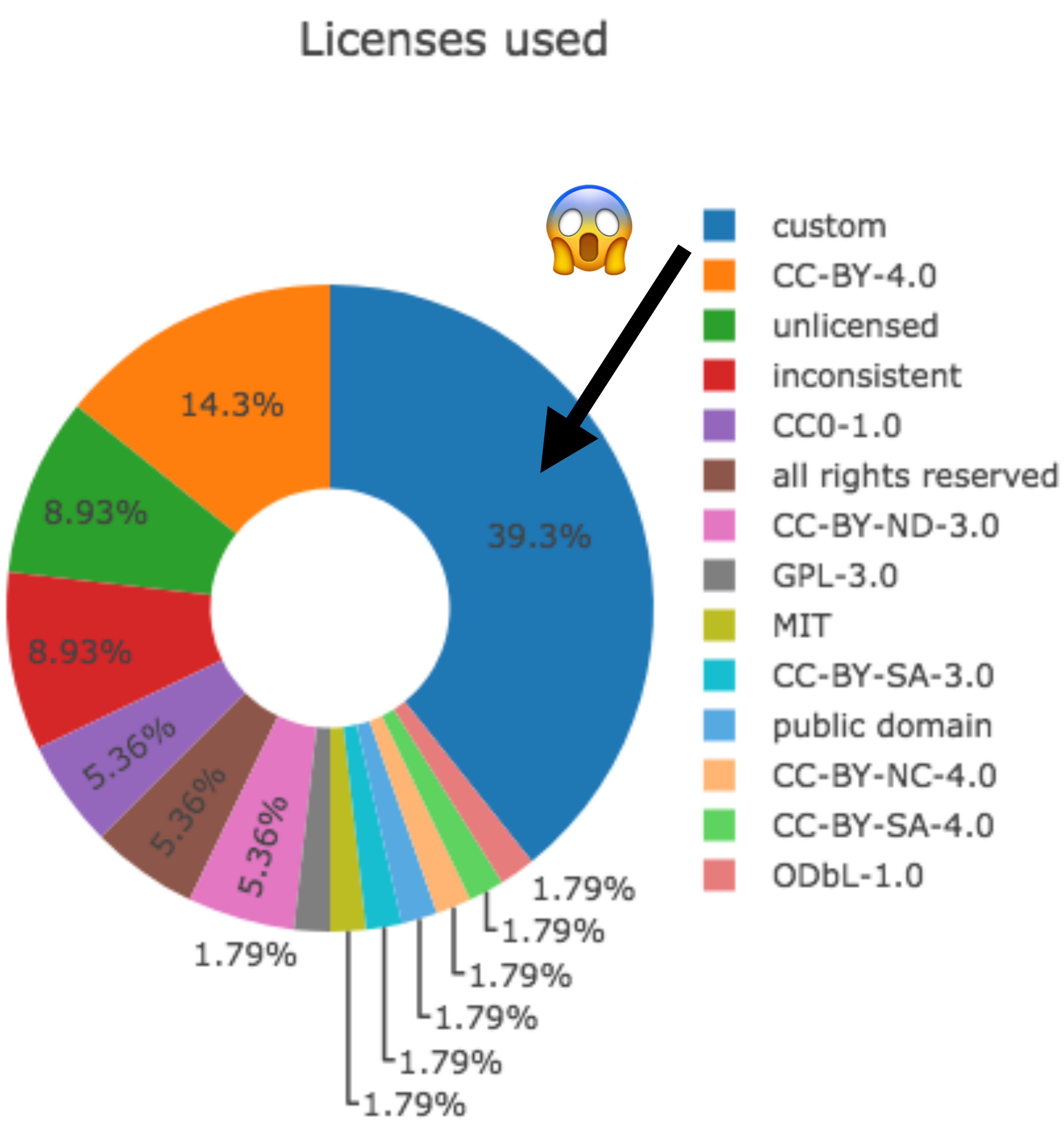
What if there’s no fitting repository?

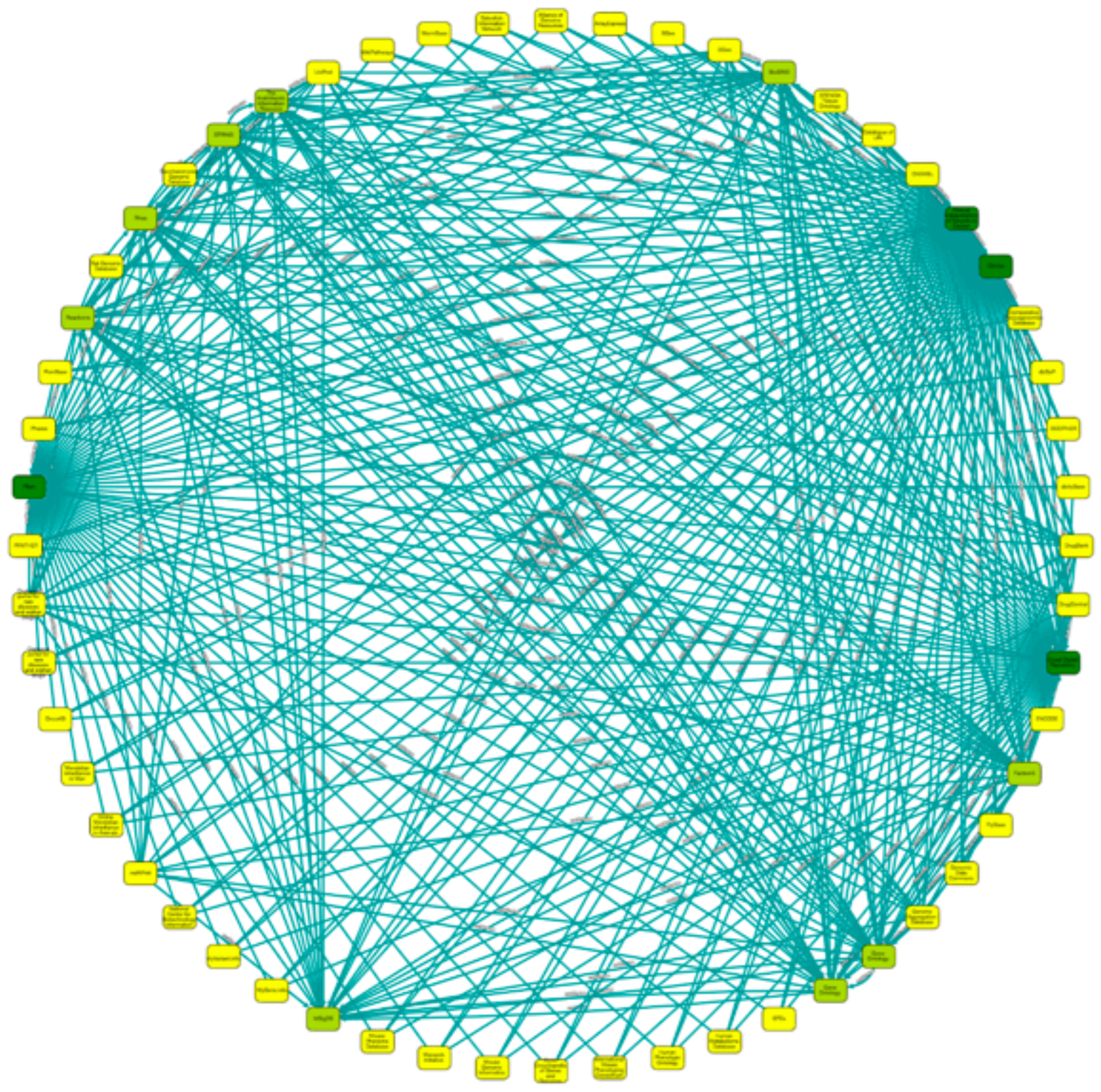
- your chance to learn how to make your own! 😉

how to (not) pick a license

- CC 0
- CC BY
- CC BY-SA
- CC BY-NC
- CC BY-NC-SA
- CC BY-ND
- CC BY-NC-ND
- CC BY-OMG-WTF
- making your own data license?

data licensing

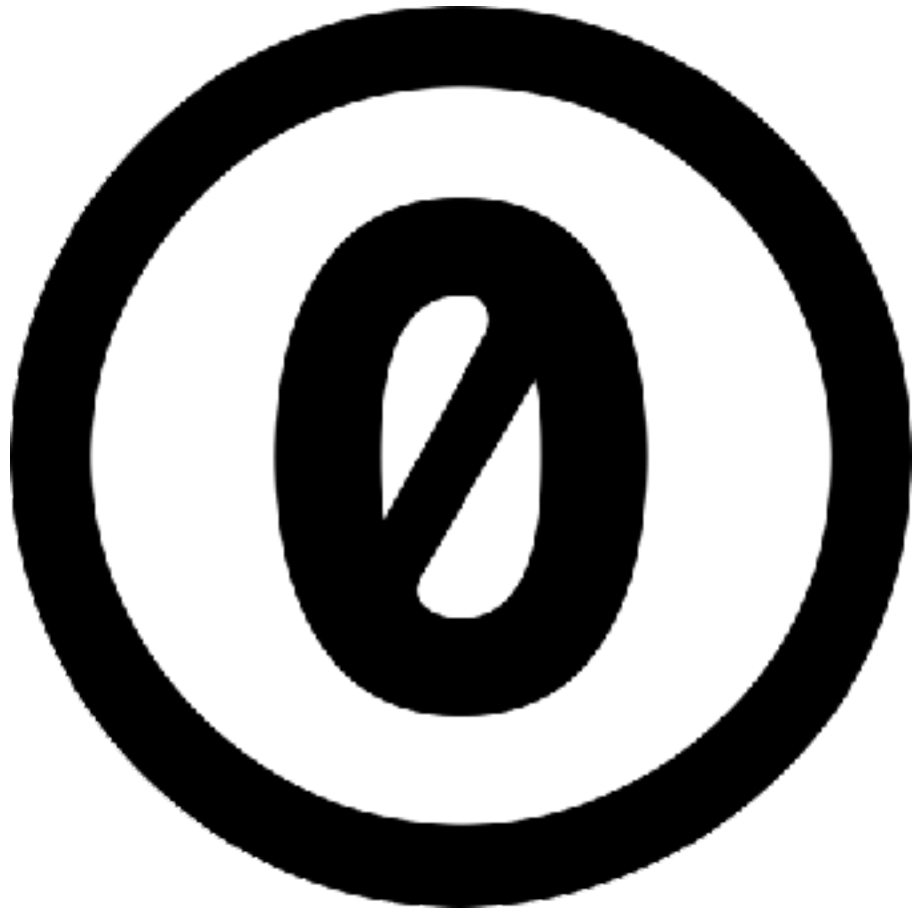

Licenses used



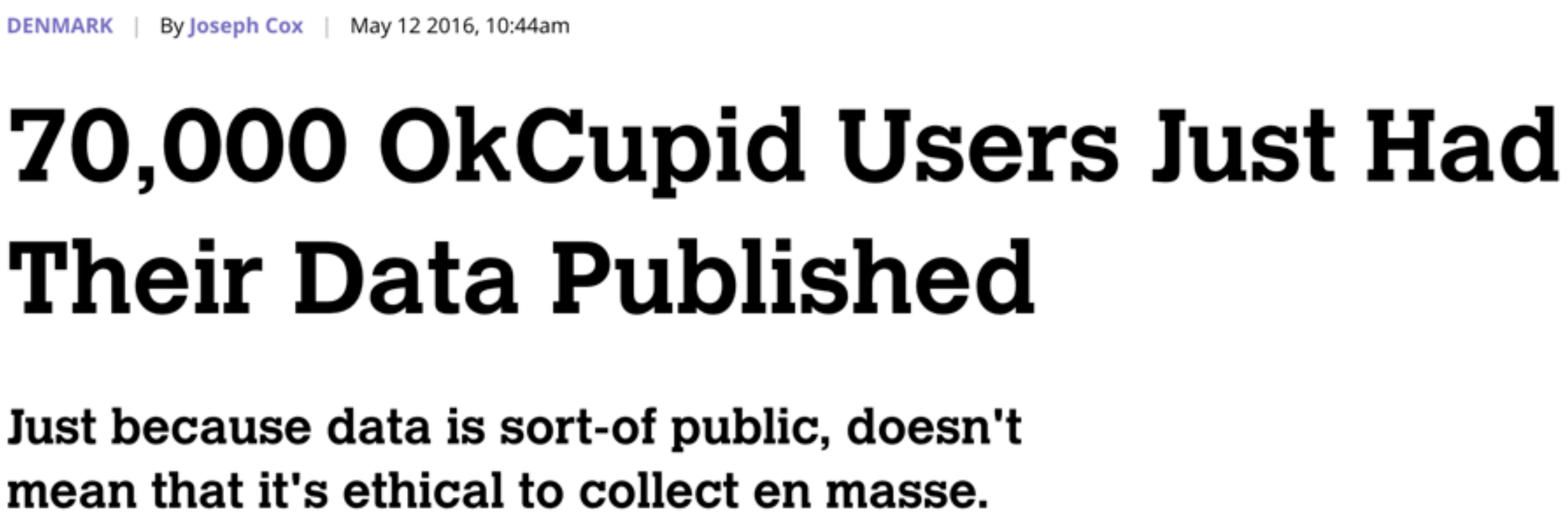
reusabledata.org

what to do?

- **as open as possible**
- as closed as necessary



how to not publish data

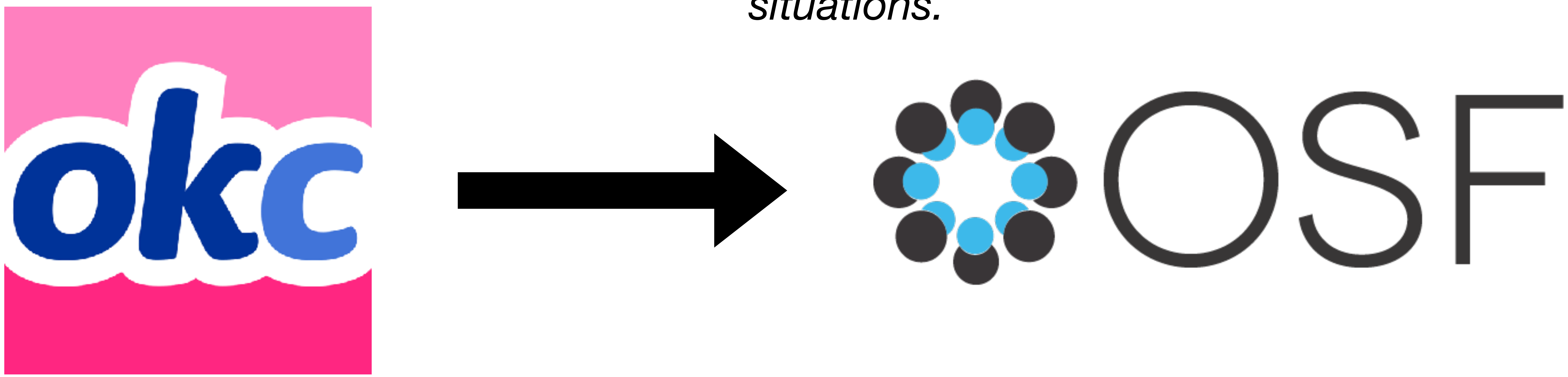


DENMARK | By Joseph Cox | May 12 2016, 10:44am

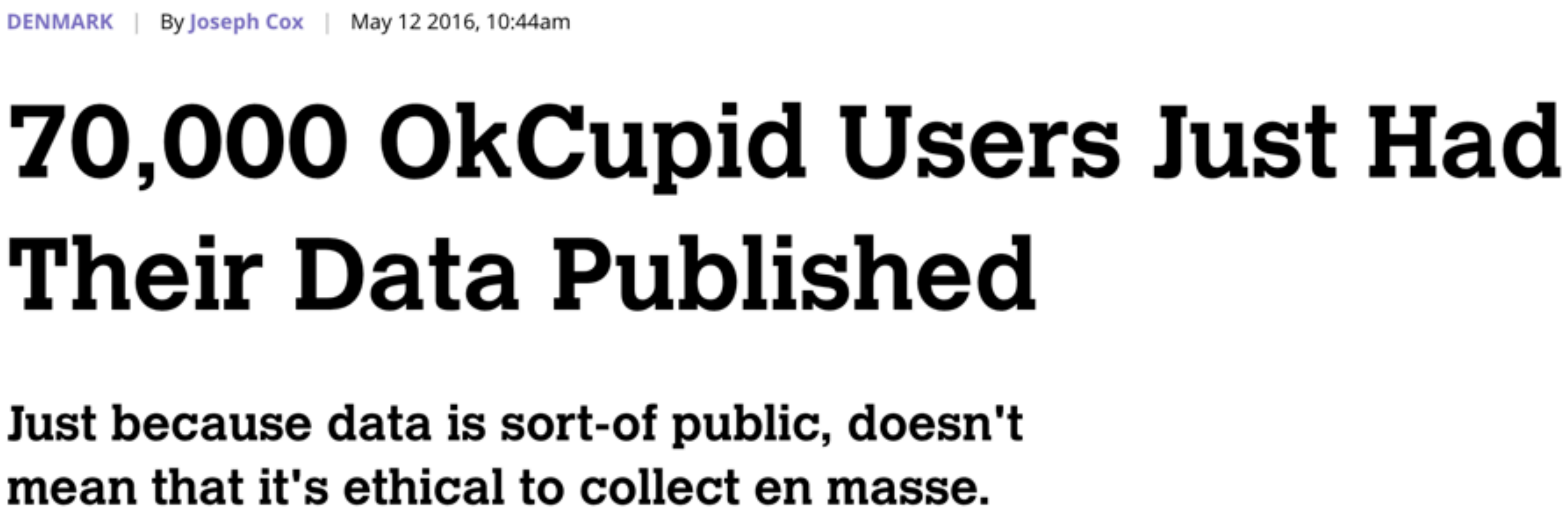
70,000 OkCupid Users Just Had Their Data Published

Just because data is sort-of public, doesn't mean that it's ethical to collect en masse.

[…] publicly released a dataset on nearly 70,000 users of the dating site OkCupid, including their sexual turn-ons, orientation, usernames and more. […] These include things like whether they ever do drugs, whether they'd like to be tied up during sex, or what's their favourite out of a series of romantic situations.



how to not publish data



DENMARK | By Joseph Cox | May 12 2016, 10:44am

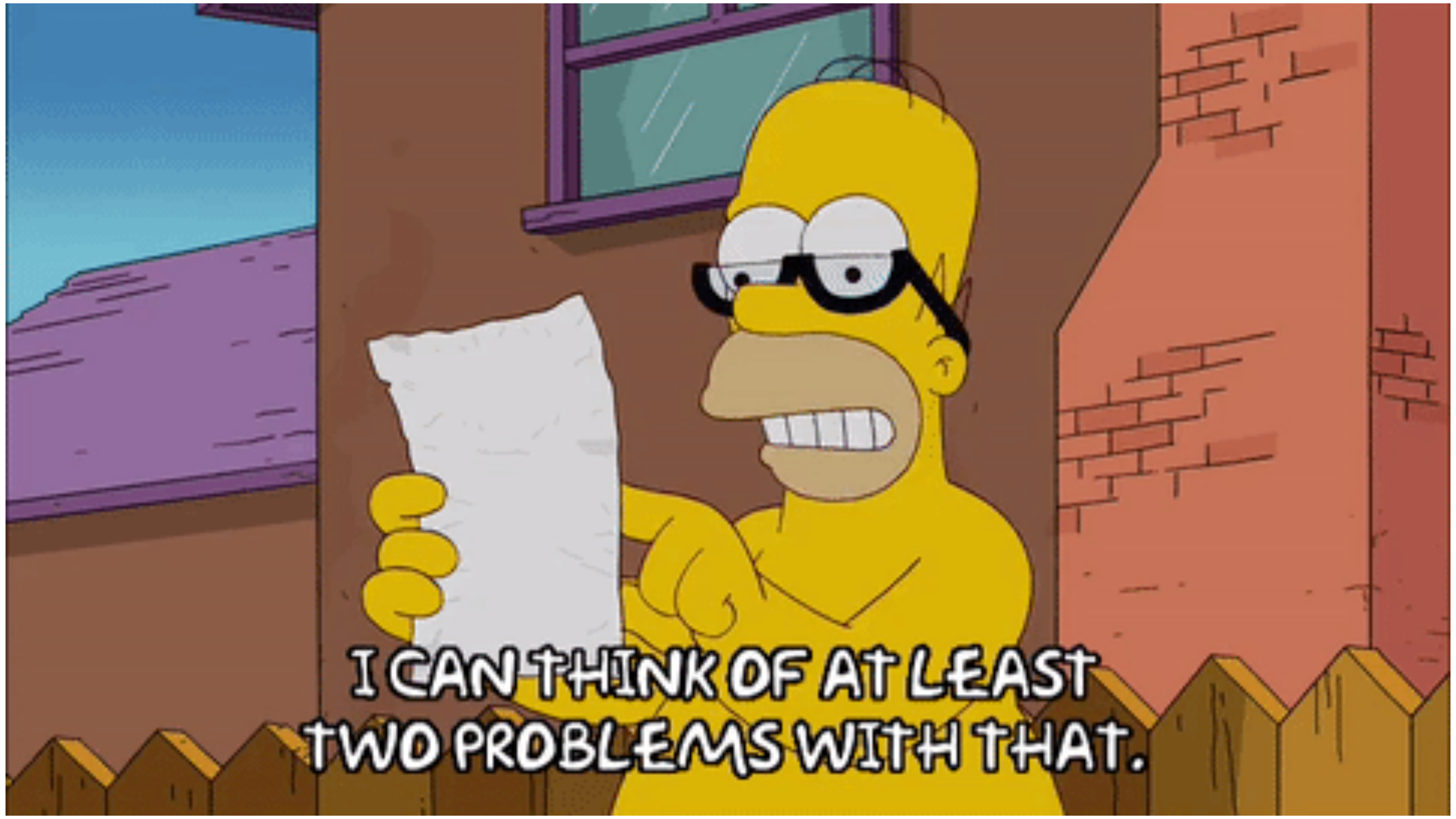
70,000 OkCupid Users Just Had Their Data Published

Just because data is sort-of public, doesn't mean that it's ethical to collect en masse.

- Large data set full of sensitive, personal data
- No ethics approval for the data & paper
- No consent from the 70k participants
- Data not anonymized/de-identified

two problems here

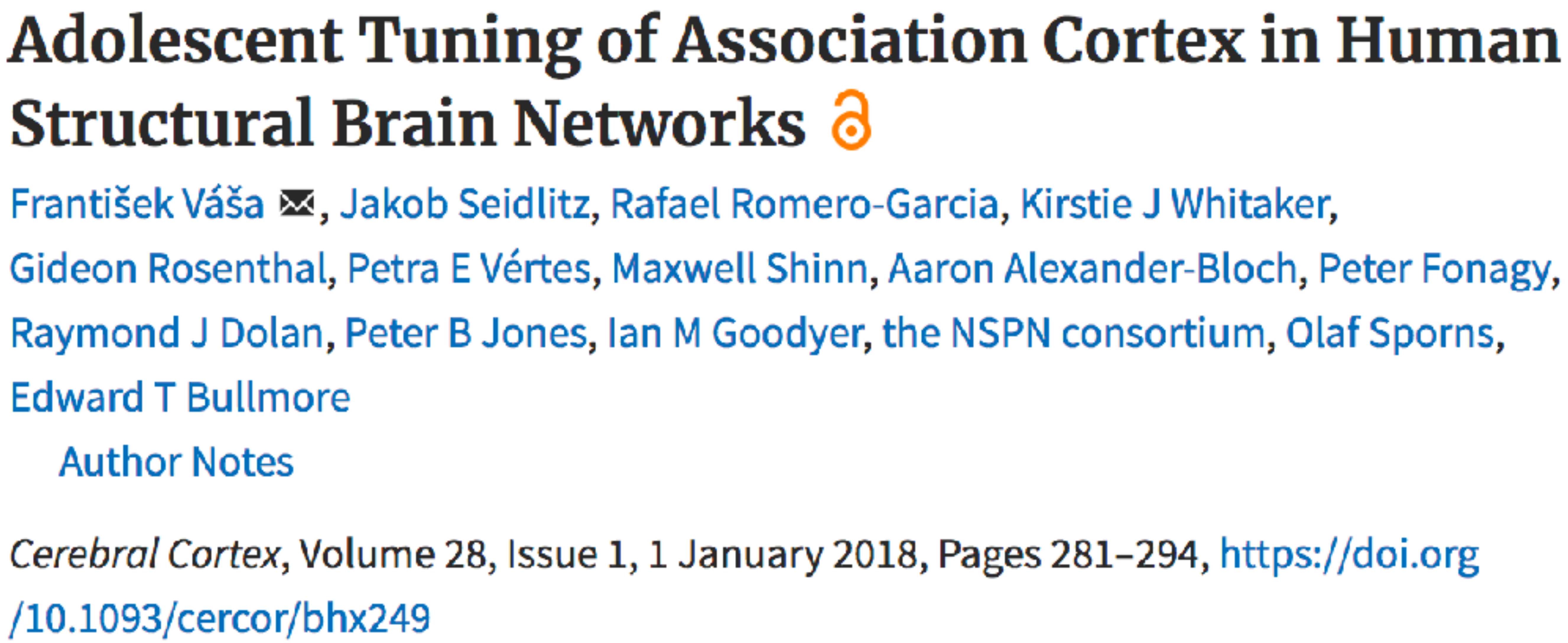
- how data was shared
- how data was acquired



unpublishable data

what to do?

- as open as possible
- **as closed as necessary**



Adolescent Tuning of Association Cortex in Human Structural Brain Networks

František Váša ✉, Jakob Seidlitz, Rafael Romero-Garcia, Kirstie J Whitaker, Gideon Rosenthal, Petra E Vértes, Maxwell Shinn, Aaron Alexander-Bloch, Peter Fonagy, Raymond J Dolan, Peter B Jones, Ian M Goodyer, the NSPN consortium, Olaf Sporns, Edward T Bullmore

Author Notes

Cerebral Cortex, Volume 28, Issue 1, 1 January 2018, Pages 281–294, https://doi.org/10.1093/cercor/bhx249

Data for this specific article has been uploaded to the Cambridge Data Repository (https://doi.org/10.17863/CAM.8856) and password protected.

Our participants did not give informed consent for their questionnaire measures to be made publicly available, and it is possible that they could be identified from this data set.

Access to the data supporting the analyses presented in this article will be made available to researchers with a reasonable request to NSPNdata@medschl.cam.ac.uk.

more formalized access restrictions?

Synapse

Awareness and Ethics Pledge

10. My organization offers free treatment fo a available on Synapse to find and re-contact individuals eligible to receive the free treatm

○ a) Yes. Data available on Synapse can be usec

○ b) No, the Synapse Terms and Conditions of U ts

○ c) Yes, all Data available on Synapse is unrest

○ d) Yes, but only if you ask the Synapse Access alf

Need help answering this question?

☐ Take full responsibility for my use of data, software

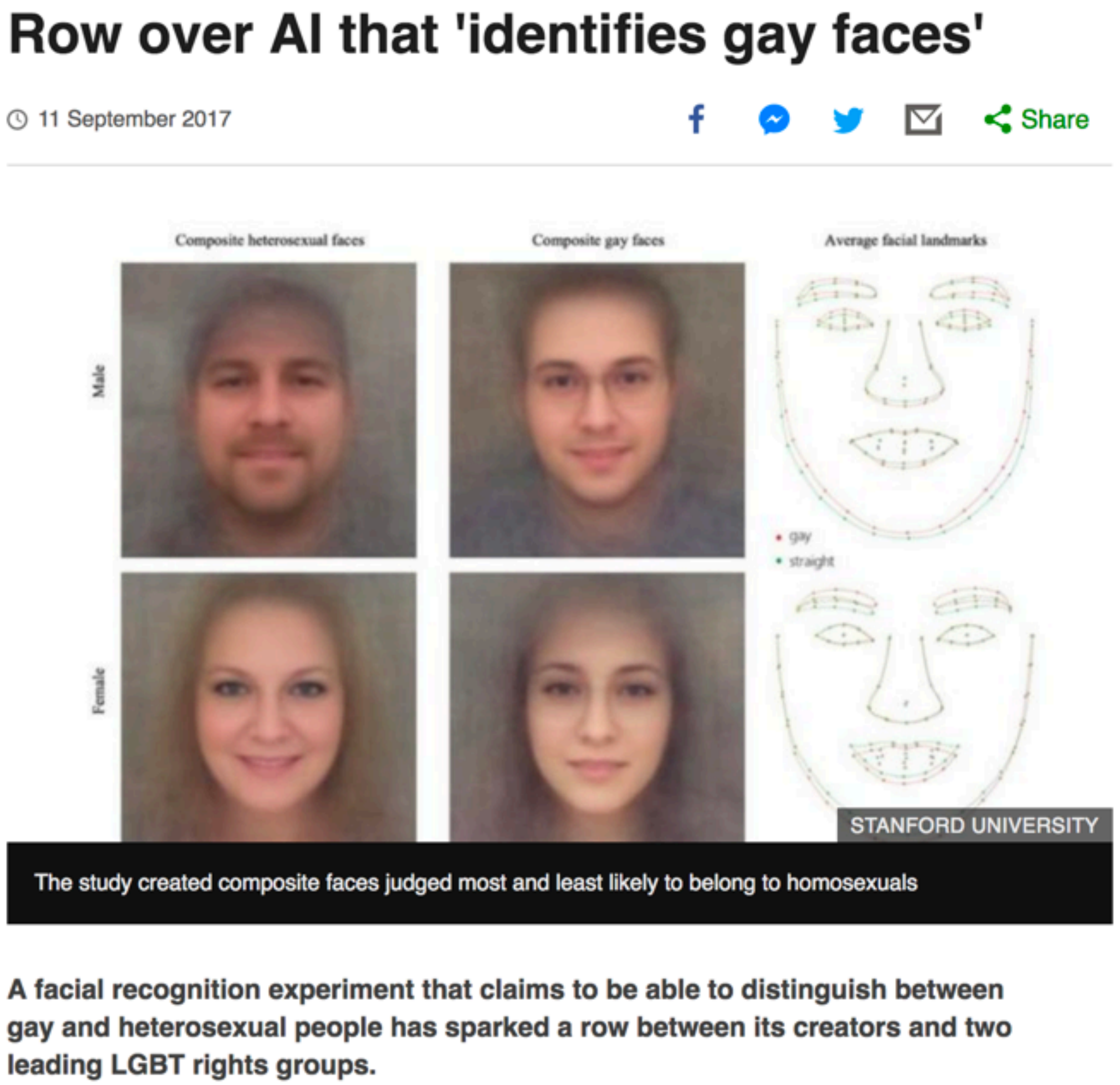
☐ Abide by all applicable laws and regulations as laic

I, ____________________ (your name), reaffirm my commitment to all Synapse Governance policies for responsible research and data handling, including:

___* I WILL NOT RE-IDENTIFY	___* I WILL NOT REDISTRIBUTE DATA	___* I WILL NOT USE FOR ADVERTISING
___* I WILL KEEP DATA SECURE	___* I WILL PROTECT PRIVACY	___* I WILL SUPPORT OPEN ACCESS
___* I WILL REPORT ANY BREACHES	___* I WILL CREDIT PARTICIPANTS	___* I WILL FOLLOW PRIVACY LAWS

acquiring data in the first place

data acquisition



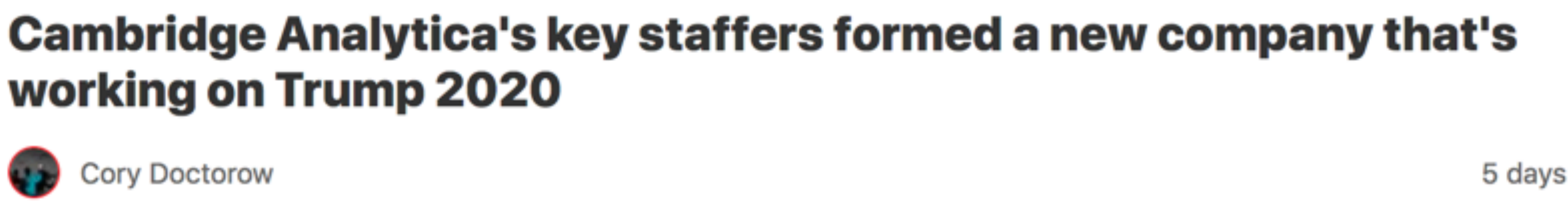
Row over AI that 'identifies gay faces'

11 September 2017

Share

The study created composite faces judged most and least likely to belong to homosexuals

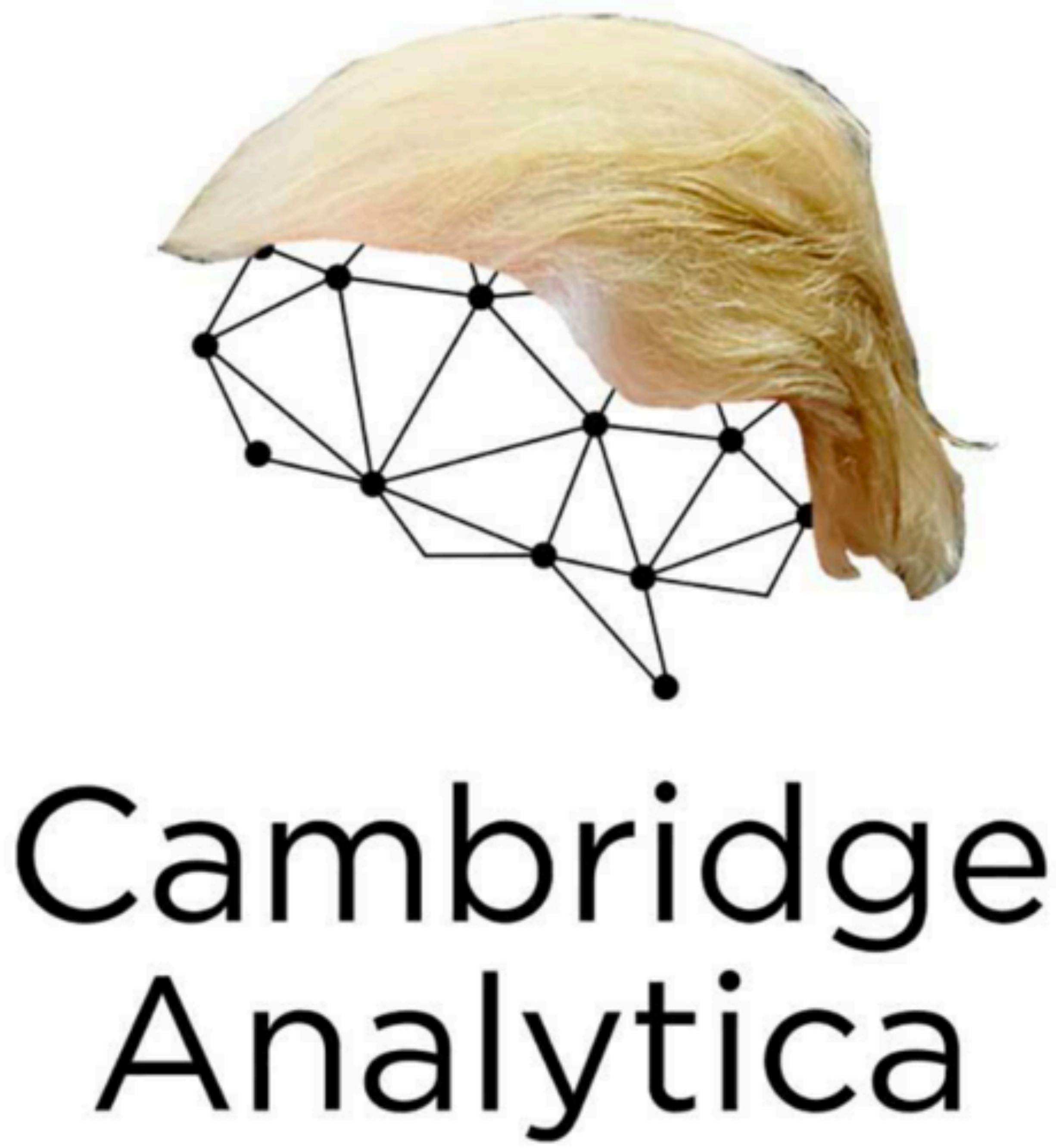
A facial recognition experiment that claims to be able to distinguish between gay and heterosexual people has sparked a row between its creators and two leading LGBT rights groups.



Cambridge Analytica's key staffers formed a new company that's working on Trump 2020

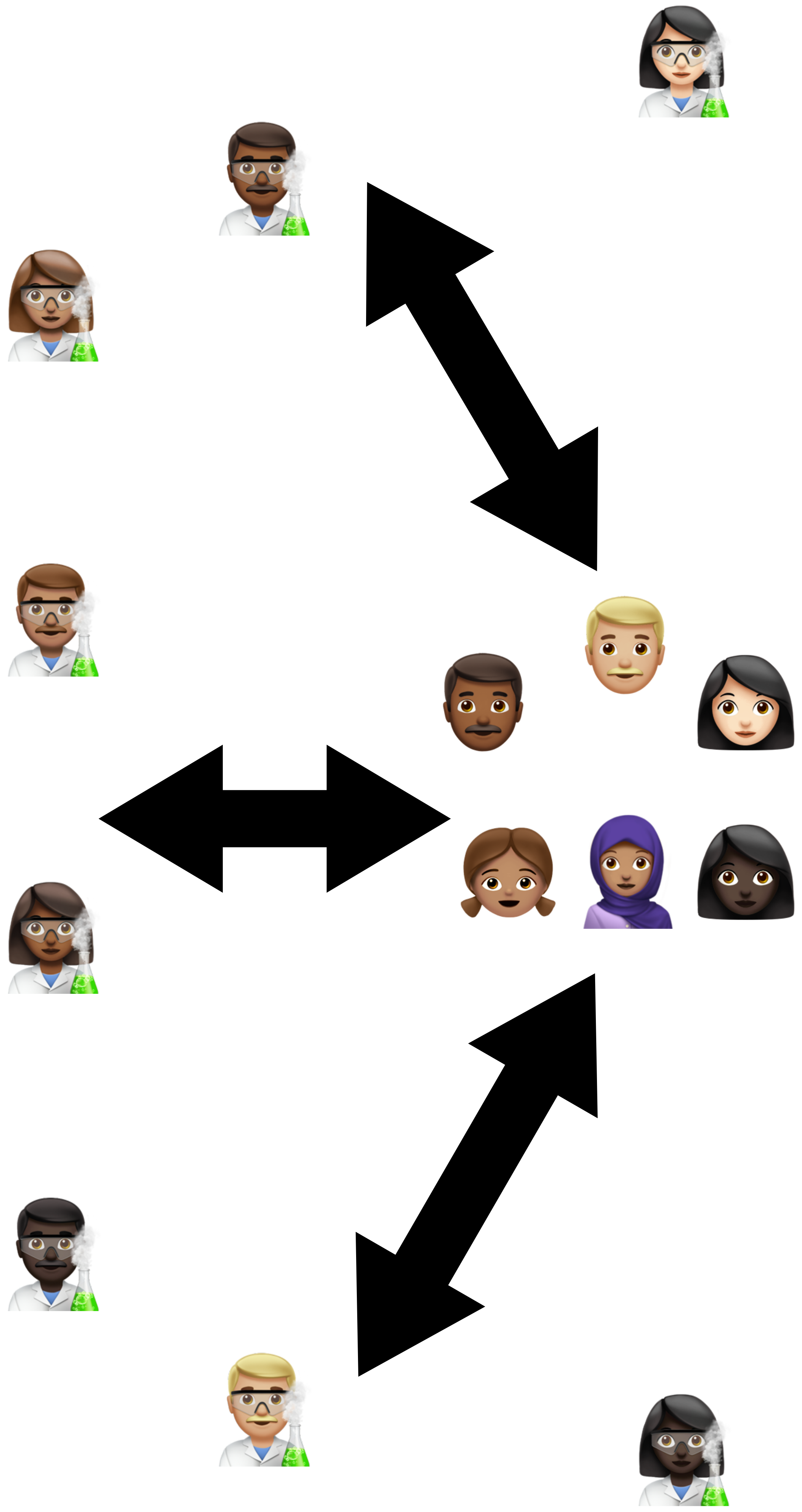
Cory Doctorow

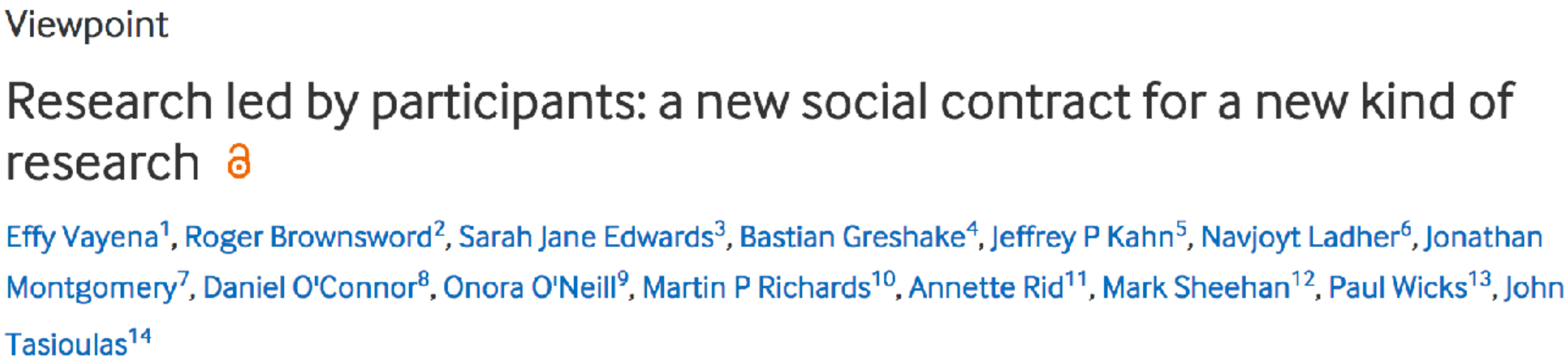
5 days



Cambridge Analytica

participant-centered/lead research





Viewpoint

Research led by participants: a new social contract for a new kind of research

Effy Vayena[1], Roger Brownsword[2], Sarah Jane Edwards[3], Bastian Greshake[4], Jeffrey P Kahn[5], Navjoyt Ladher[6], Jonathan Montgomery[7], Daniel O'Connor[8], Onora O'Neill[9], Martin P Richards[10], Annette Rid[11], Mark Sheehan[12], Paul Wicks[13], John Tasioulas[14]

- participants can be involved in designing a project / study

- participants can give much better informed consent as a consequence

openness & participant-centered research

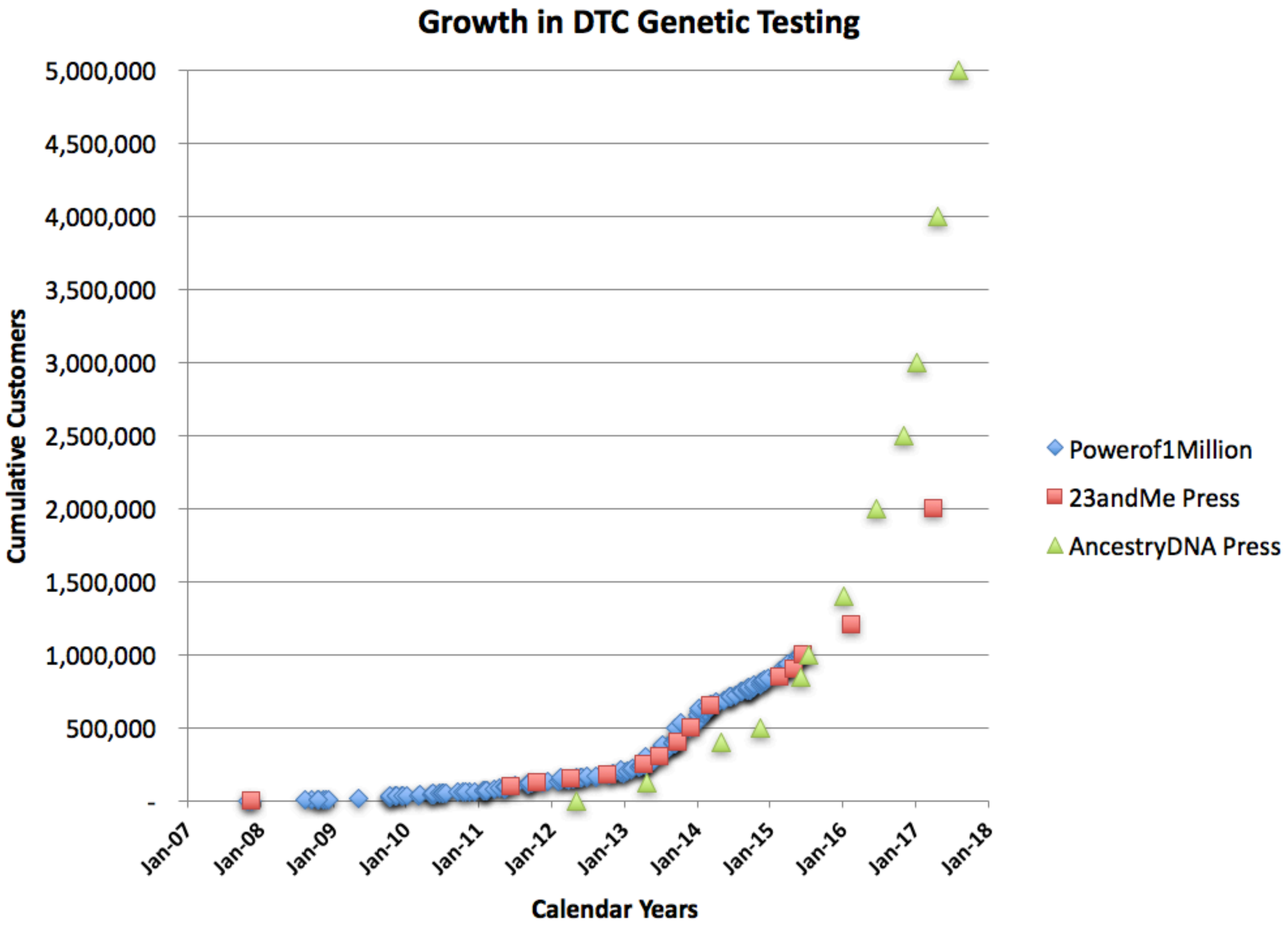
openSNP

OPEN HUMANS

openSNP

~10 - 17 **Mio** people already got their own genome analyzed through commercial entities!

But there was no data repository for it!



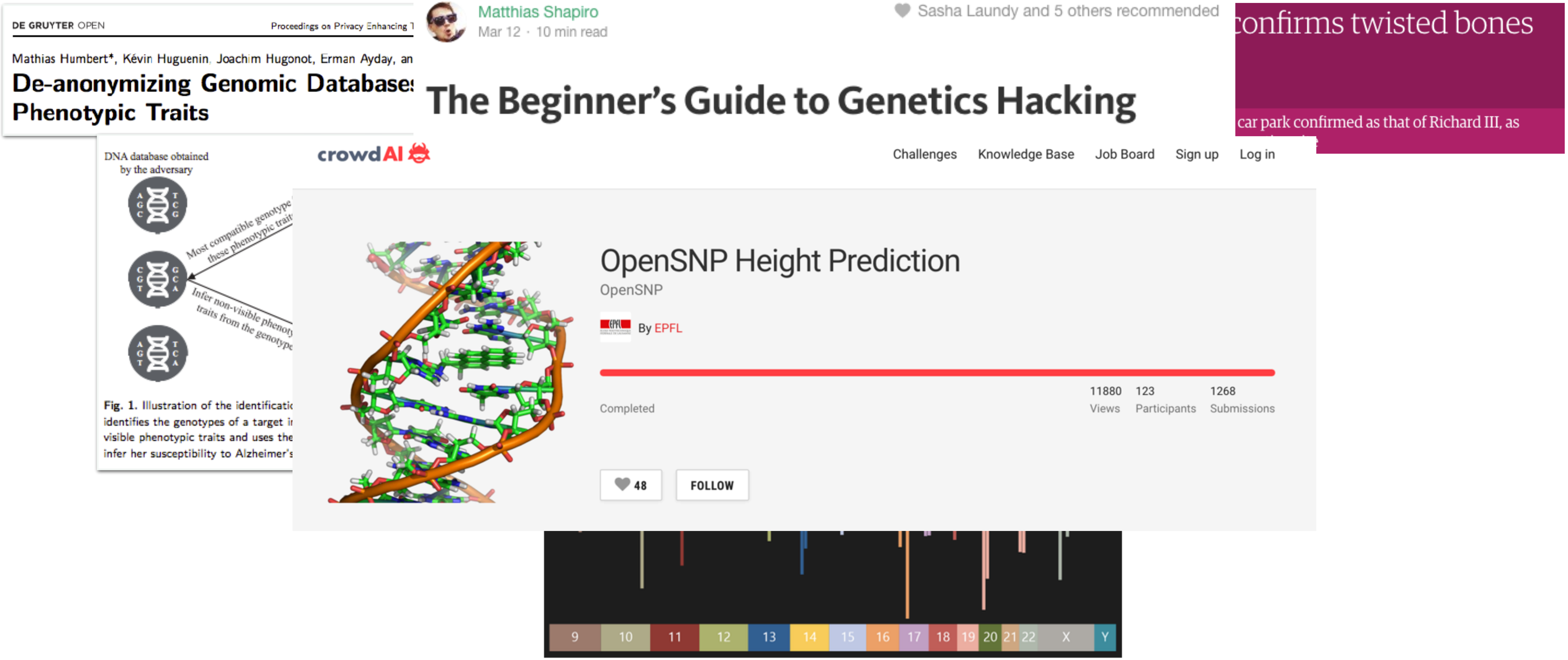
openSNP

- open data ✔
- open source ✔
- crowdsourced ✔
- crowdfunded ✔

🄍

- ~7,000 users
- >4,500 genetic data sets
- 543 phenotypes w/ total of 60,000 answers

Uses Cases



DE GRUYTER OPEN Proceedings on Privacy Enhancing T

Mathias Humbert*, Kévin Huguenin, Joachim Hugonot, Erman Ayday, an

De-anonymizing Genomic Database Phenotypic Traits

DNA database obtained by the adversary

A G C
T C G
C G T
G C A
A G T
T C A
Most compatible genotype these phenotypic trait
Infer non-visible phenoty traits from the genotype

Fig. 1. Illustration of the identificatio identifies the genotypes of a target i visible phenotypic traits and uses the infer her susceptibility to Alzheimer's

Matthias Shapiro
Mar 12 · 10 min read

Sasha Laundy and 5 others recommended

The Beginner's Guide to Genetics Hacking

confirms twisted bones

car park confirmed as that of Richard III, as

crowdAI

Challenges Knowledge Base Job Board Sign up Log in

OpenSNP Height Prediction

OpenSNP

By EPFL

Completed

11880	123	1268
Views	Participants	Submissions

48 FOLLOW

openSNP

👍

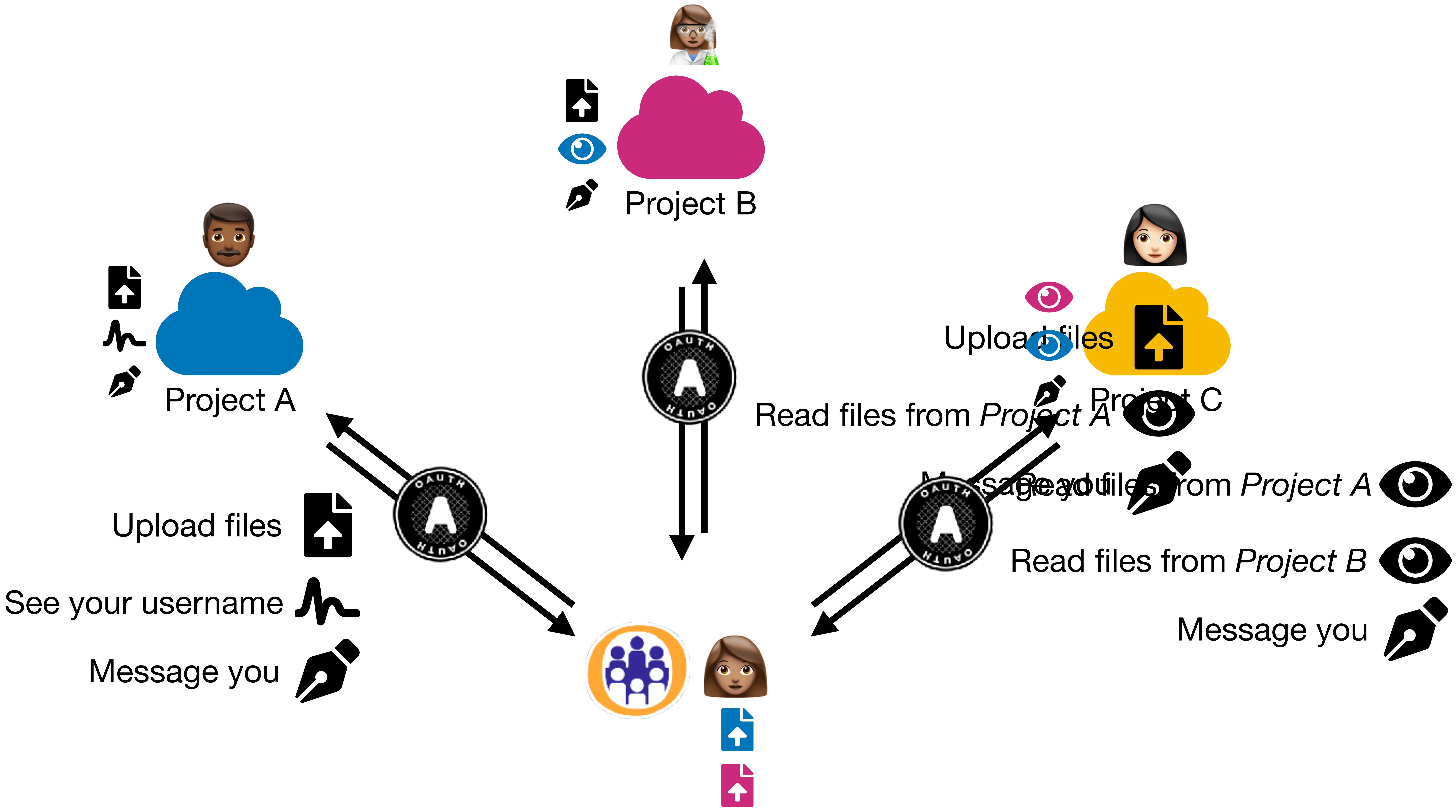
- people give consent to data sharing
- people are willing to donate personal data
- data is already being used

👎

- no grades of sharing
- makes consent a one-time decision
- complete loss of control over data

granular consent & OPEN HUMANS

OPEN HUMANS



Project B

Project A

Upload files

See your username

Message you

Upload files

Read files from Project A

Project C

Message you

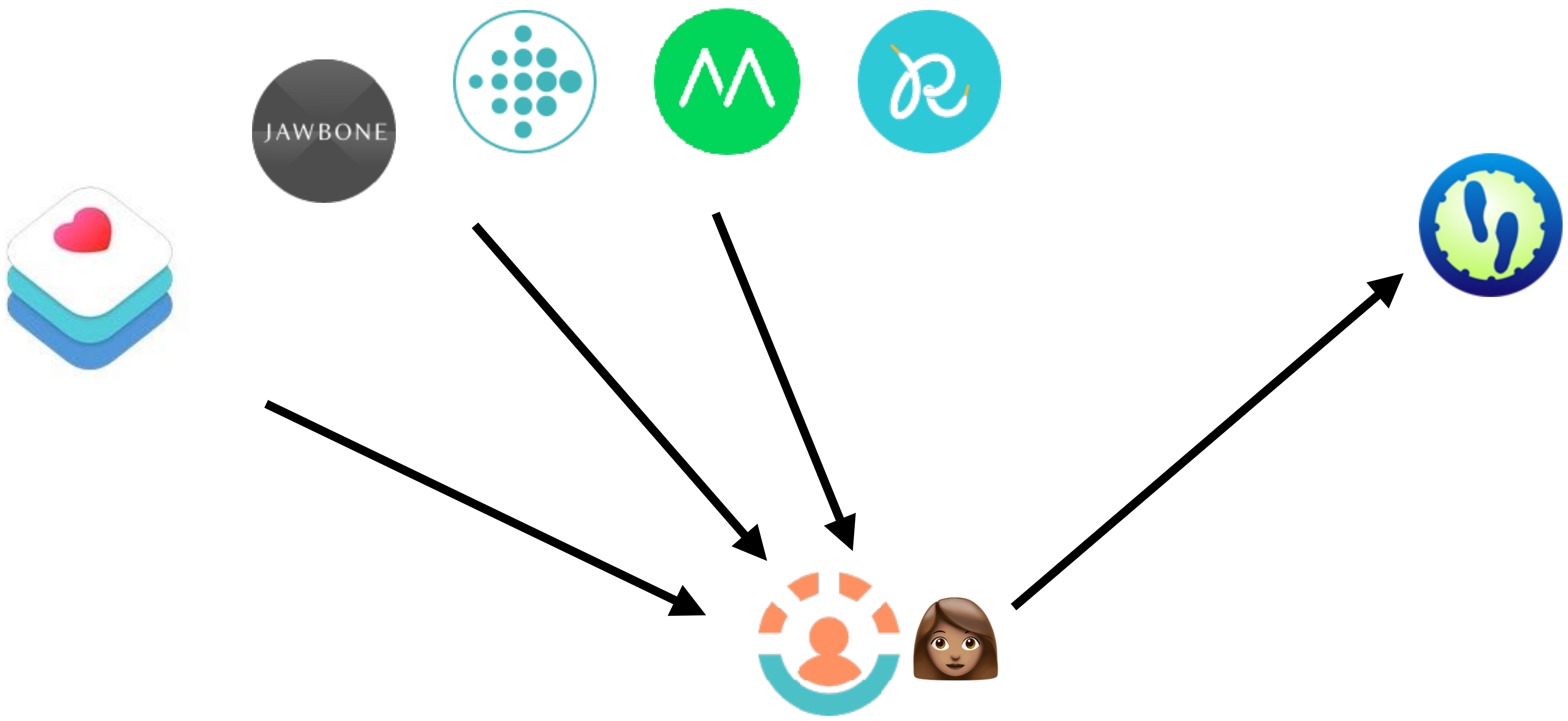
Read files from Project A

Read files from Project B

Message you

how does that work in practice?

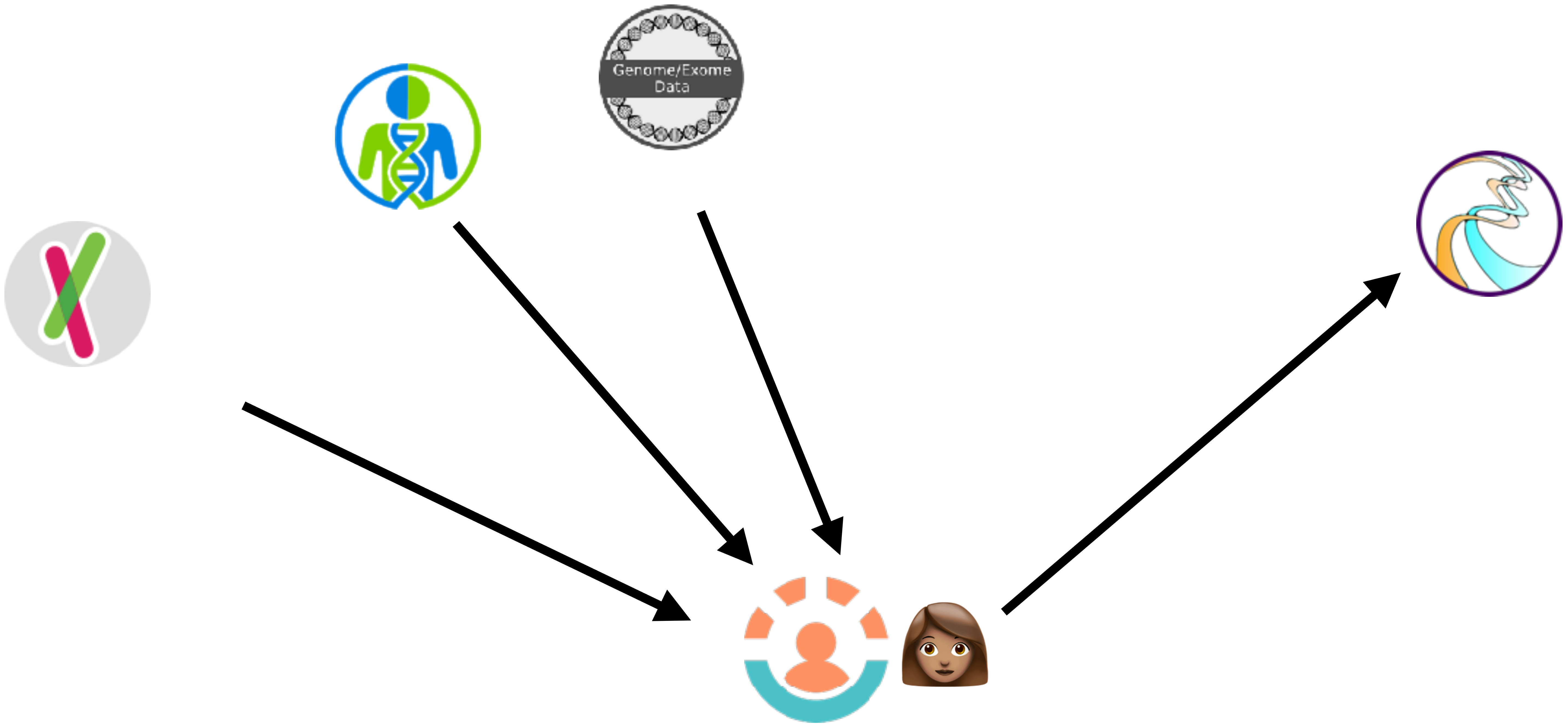
Academic research using activity trackers



Activity tracking apps / devices

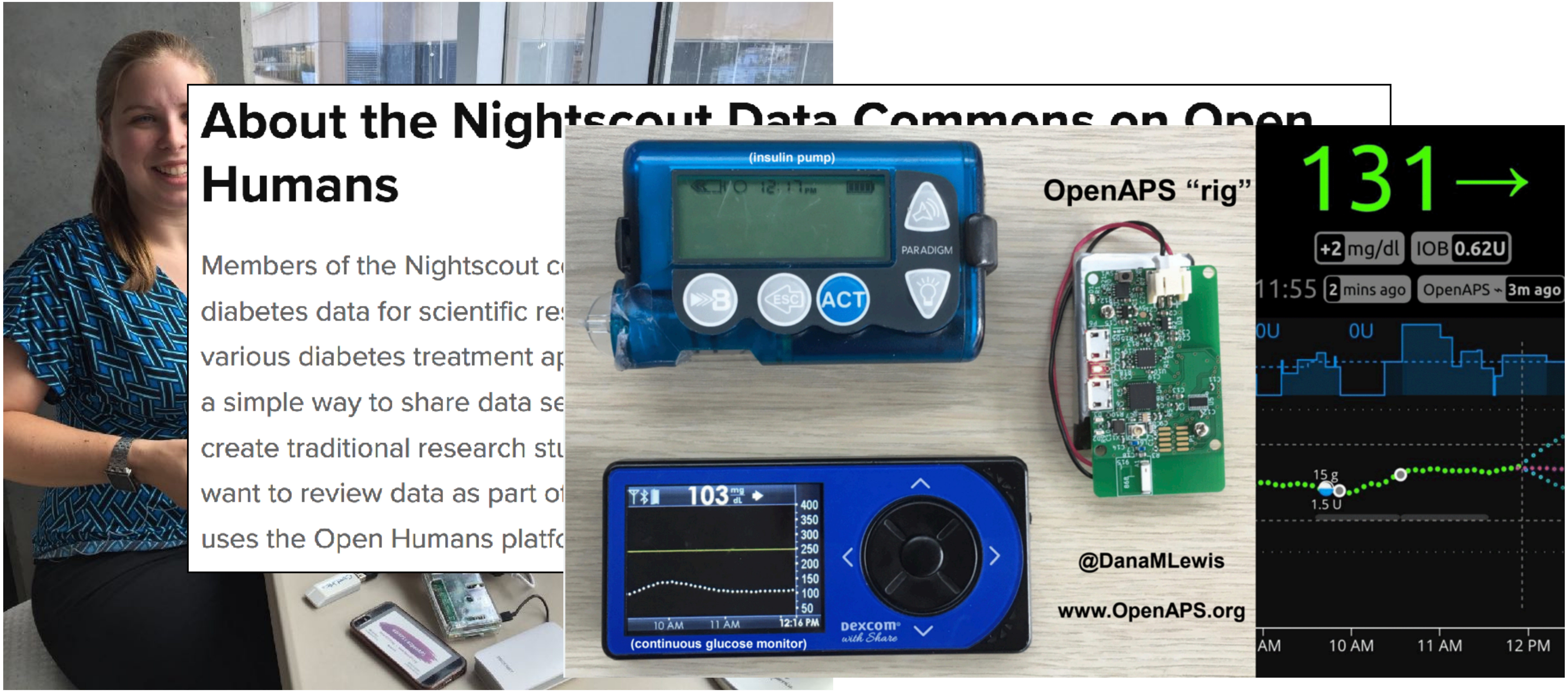
Keeping Pace
(Rumi Chunara, NYU)

Genome Exploration



Genevieve
Personal
Genome Annotation
with *Clinvar*

Community research in Diabetes



About the Nightscout Data Commons on Open Humans

Members of the Nightscout c
diabetes data for scientific re
various diabetes treatment ap
a simple way to share data se
create traditional research stu
want to review data as part o
uses the Open Humans platfo

some numbers

- 5,691 registered users (2,251 w/ at least 1 data set)
- 15,119 data files available in total
- 26 projects currently running

current data sources

23andMe Upload
Open Humans
Open Humans Foundation
Connected by **1023** members
23andMe is a direct-to-consumer genetic testing company that tests about one million genetic locations.
Learn more

Personal Data Notebooks
Bastian Greshake Tzovaras
Open Humans Foundation
Connected by **296** members
Play with code for analyzing personal data! Your data stays private on a personal virtual machine.
Learn more

Nightscout Data Transfer
James Wedding, P.E.
The Nightscout Foundation
Connected by **216** members
A tool to easily enable upload of data from individual Nightscout databases to the Open Humans platform

AncestryDNA Upload
Open Humans
Open Humans Foundation
Connected by **361** members
Ancestry.com's AncestryDNA is a direct-to-consumer genetic testing product that tests about 700,000 genetic locations.
Learn more

Gencove
Joseph K. Pickrell
Gencove, Inc.
Connected by **283** members
Your genome app - get your ancestry, microbiome, and more! Contribute your data to OpenHumans.
Learn more

Runkeeper connection
Open Humans
Open Humans Foundation
Connected by **207** members
RunKeeper is a free smartphone app for GPS fitness-tracking. You can use it to record GPS timepoint data for runs, walks, bicycling, and oth

Fitbit Connection
Open Humans
Open Humans Foundation
Connected by **354** members
Connect your Fitbit account to add data from your Fitbit activity trackers and other Fitbit devices.
Learn more

Twitter Archive Analyzer
Bastian Greshake Tzovaras
Connected by **293** members
The TwArχiv is a Twitter Archive Analyzer. Upload your Twitter archive and get new insights.
Learn more

openSNP
Bastian Greshake Tzovaras
openSNP
Connected by **240** members
openSNP allows you to put your genetic and phenotypic data into the public domain. Connect your openSNP account to Open Humans.

Open Humans Healthkit Integration
James Turner
self
Connected by **176** members
Install this third-party app on your iPhone or iPad to upload HealthKit data to Open Humans.
Learn more

OpenAPS Data Commons
Dana Lewis
OpenAPS
Joined by **94** members
The OpenAPS Data Commons collects data from DIY closed loopers and facilitates research in partnership with the DIY closed loop community.
Learn more

uBiome Upload
Open Humans
Open Humans Foundation
Connected by **57** members
uBiome is a company based in San Francisco that performs microbiome sequencing
Learn more

Data sharing and ethical oversight
Prof. Dr. Effy Vayena
ETH Zurich
Joined by **133** members
Through a quick survey we aim at understanding what the Open Humans community thinks about ethical oversight.
Learn more

Nokia Health (Withings) Connection
Open Humans
Open Humans Foundation
Connected by **91** members
Add your Nokia Health (Withings) data to Open Humans
Learn more

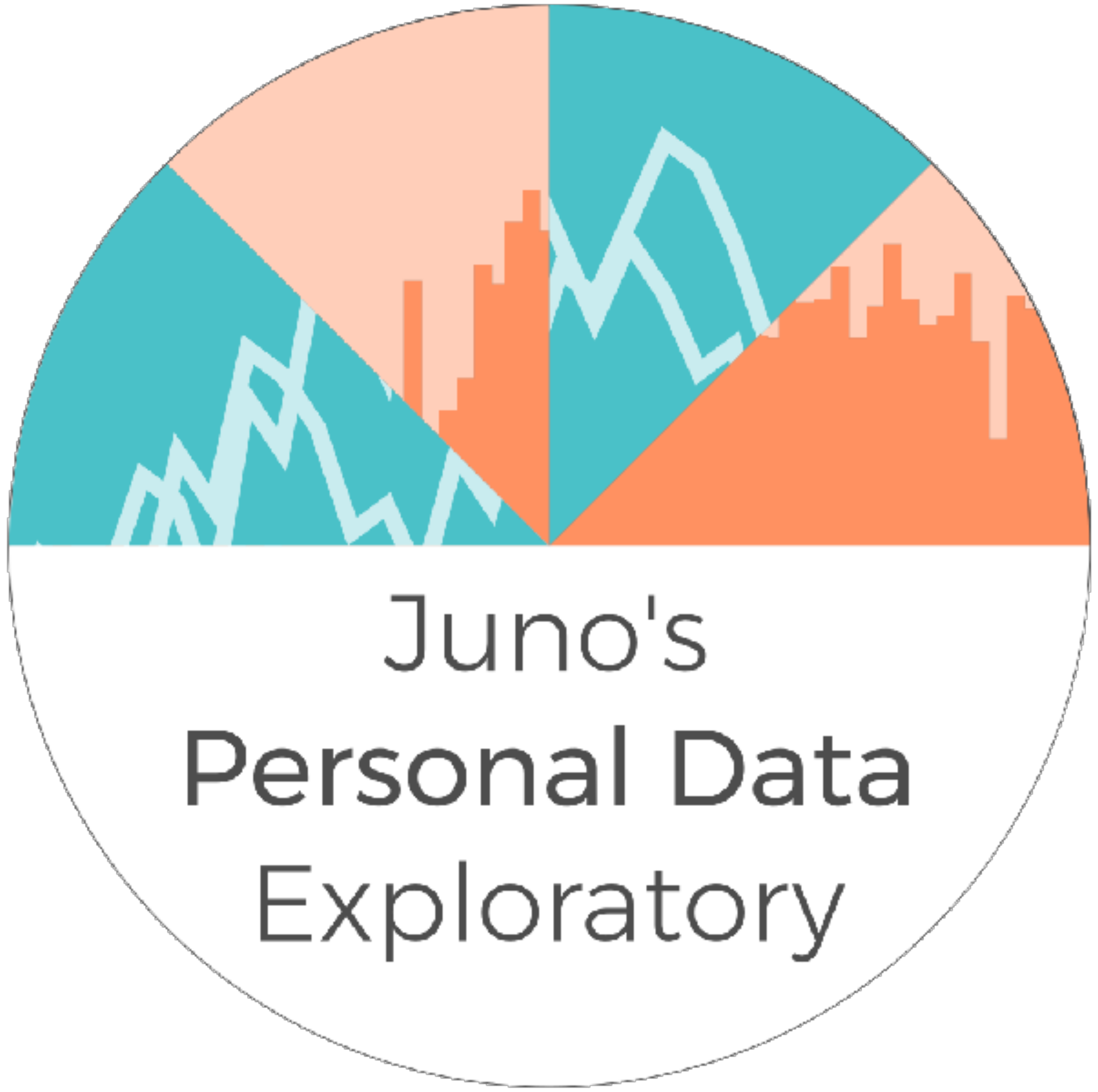
FamilyTreeDNA integration
Bastian Greshake Tzovaras
Connected by **48** members
Upload your FamilyTreeDNA data to Open Humans
Learn more

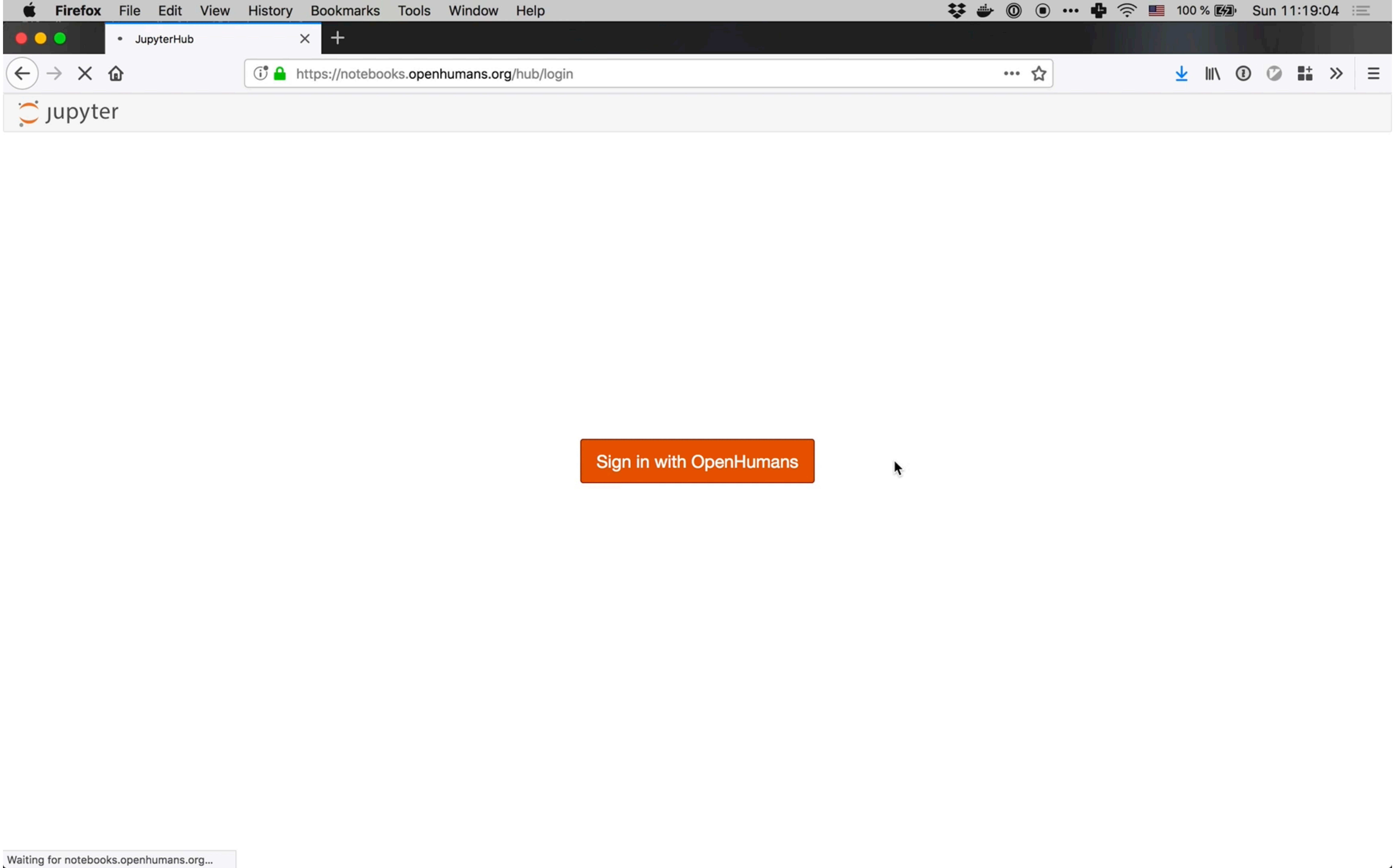
Genome/Exome Upload
Open Humans
Open Humans Foundation
Connected by **152** members
Do you have genome or exome data? You can upload genome, exome, and genotyping data in VCF format.
Learn more

Data Selfies
Open Humans
Open Humans Foundation
Connected by **106** members
Upload Data Selfies
Learn more

Nightscout Data Commons
The Nightscout Data Commons Committee
The Nightscout Foundation
Joined by **75** members
The Nightscout Data Commons collects data from Nightscout users and facilitates research in partnership with the DIY diabetes community.
Learn more

personal data
notebooks

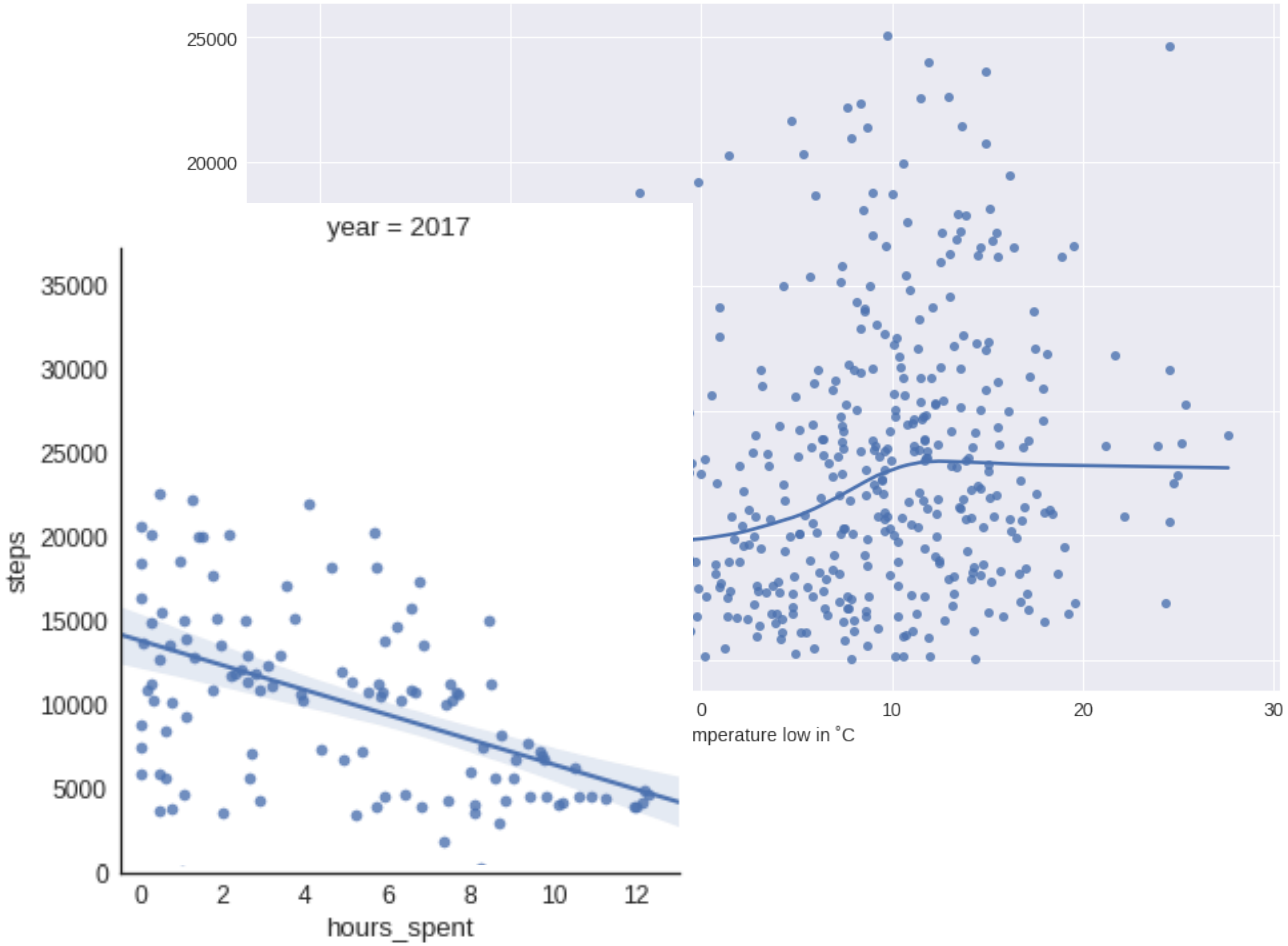

Juno's
Personal Data
Exploratory


Firefox
File
Edit
View
History
Bookmarks
Tools
Window
Help
100 %
Sun 11:19:04
JupyterHub
https://notebooks.openhumans.org/hub/login
jupyter
Sign in with OpenHumans
Waiting for notebooks.openhumans.org...

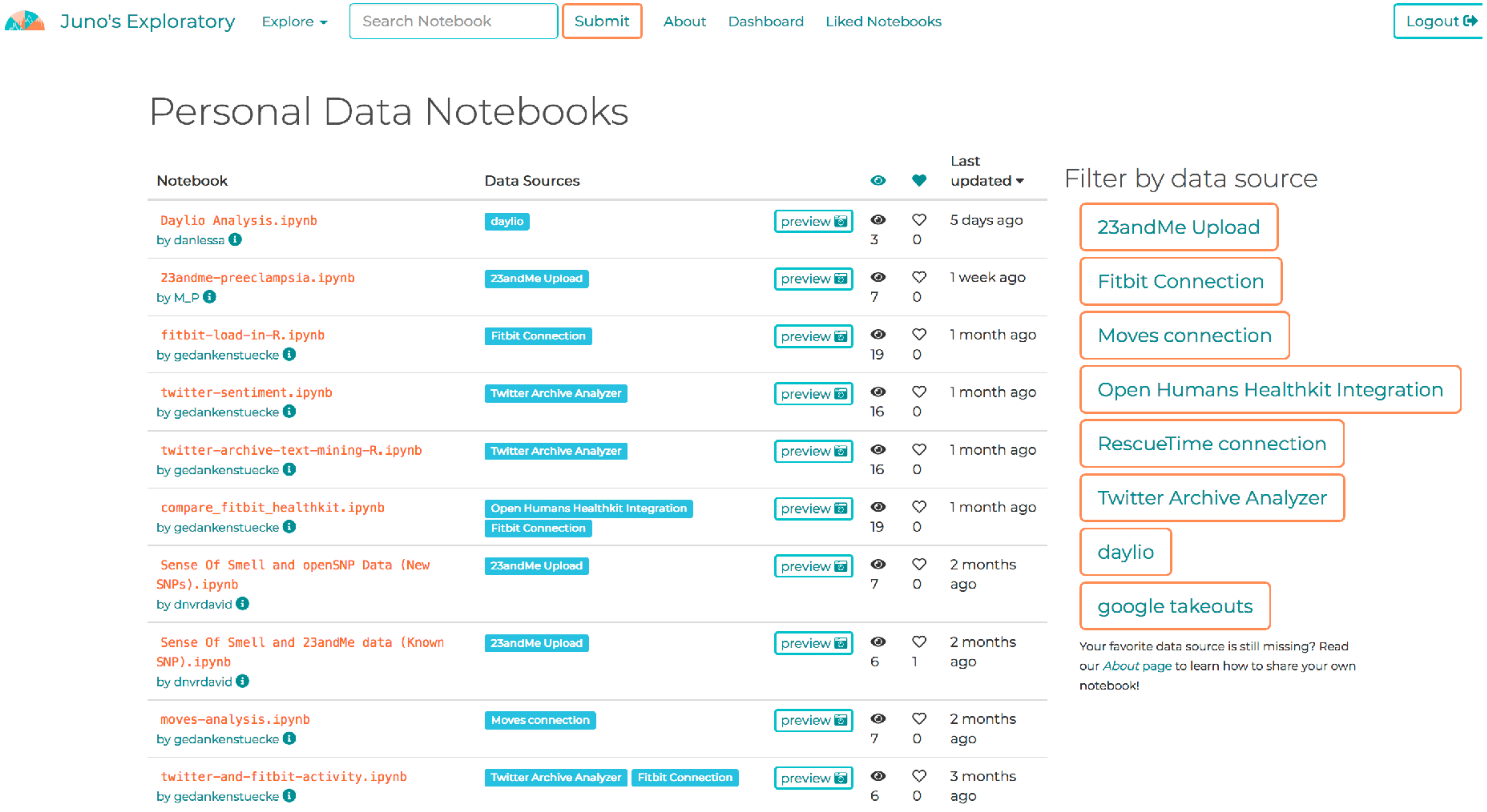
we can now correlate data sources!

Do my step counts drop
when it’s cold/rainy outside?

Does my daily workload influence
how much I walk?



benefits of data aggregation: reproducible n=1 experiments



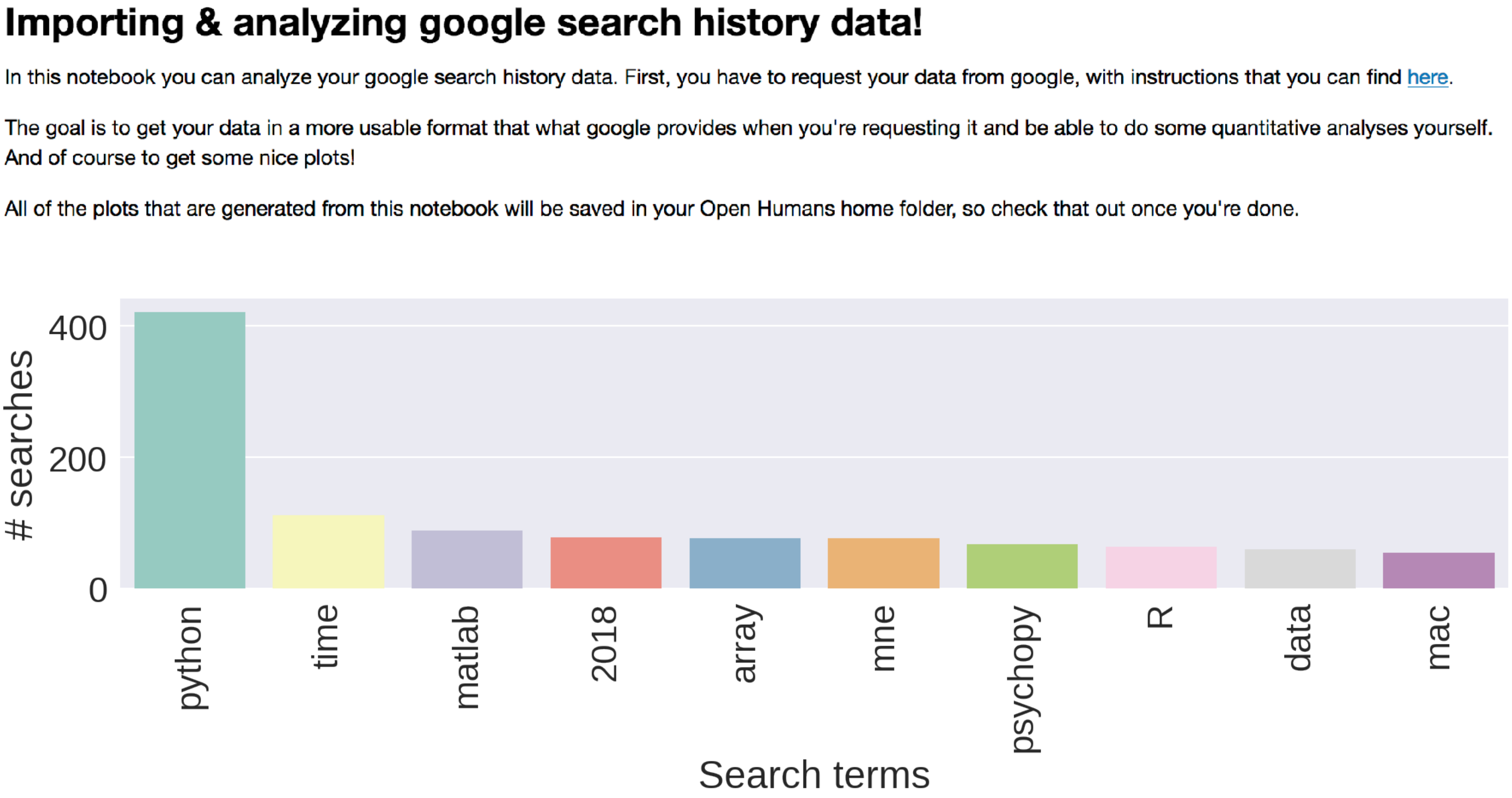
Personal Data Notebooks

Notebook	Data Sources		👁	♥	Last updated ▾
Daylio Analysis.ipynb by danlessa ℹ	daylio	preview	3	0	5 days ago
23andme-preeclampsia.ipynb by M_P ℹ	23andMe Upload	preview	7	0	1 week ago
fitbit-load-in-R.ipynb by gedankenstuecke ℹ	Fitbit Connection	preview	19	0	1 month ago
twitter-sentiment.ipynb by gedankenstuecke ℹ	Twitter Archive Analyzer	preview	16	0	1 month ago
twitter-archive-text-mining-R.ipynb by gedankenstuecke ℹ	Twitter Archive Analyzer	preview	16	0	1 month ago
compare_fitbit_healthkit.ipynb by gedankenstuecke ℹ	Open Humans Healthkit Integration Fitbit Connection	preview	19	0	1 month ago
Sense Of Smell and openSNP Data (New SNPs).ipynb by dnvrdavid ℹ	23andMe Upload	preview	7	0	2 months ago
Sense Of Smell and 23andMe data (Known SNP).ipynb by dnvrdavid ℹ	23andMe Upload	preview	6	1	2 months ago
moves-analysis.ipynb by gedankenstuecke ℹ	Moves connection	preview	7	0	2 months ago
twitter-and-fitbit-activity.ipynb by gedankenstuecke ℹ	Twitter Archive Analyzer Fitbit Connection	preview	6	0	3 months ago

Filter by data source

- 23andMe Upload
- Fitbit Connection
- Moves connection
- Open Humans Healthkit Integration
- RescueTime connection
- Twitter Archive Analyzer
- daylio
- google takeouts

Your favorite data source is still missing? Read our *About* page to learn how to share your own notebook!

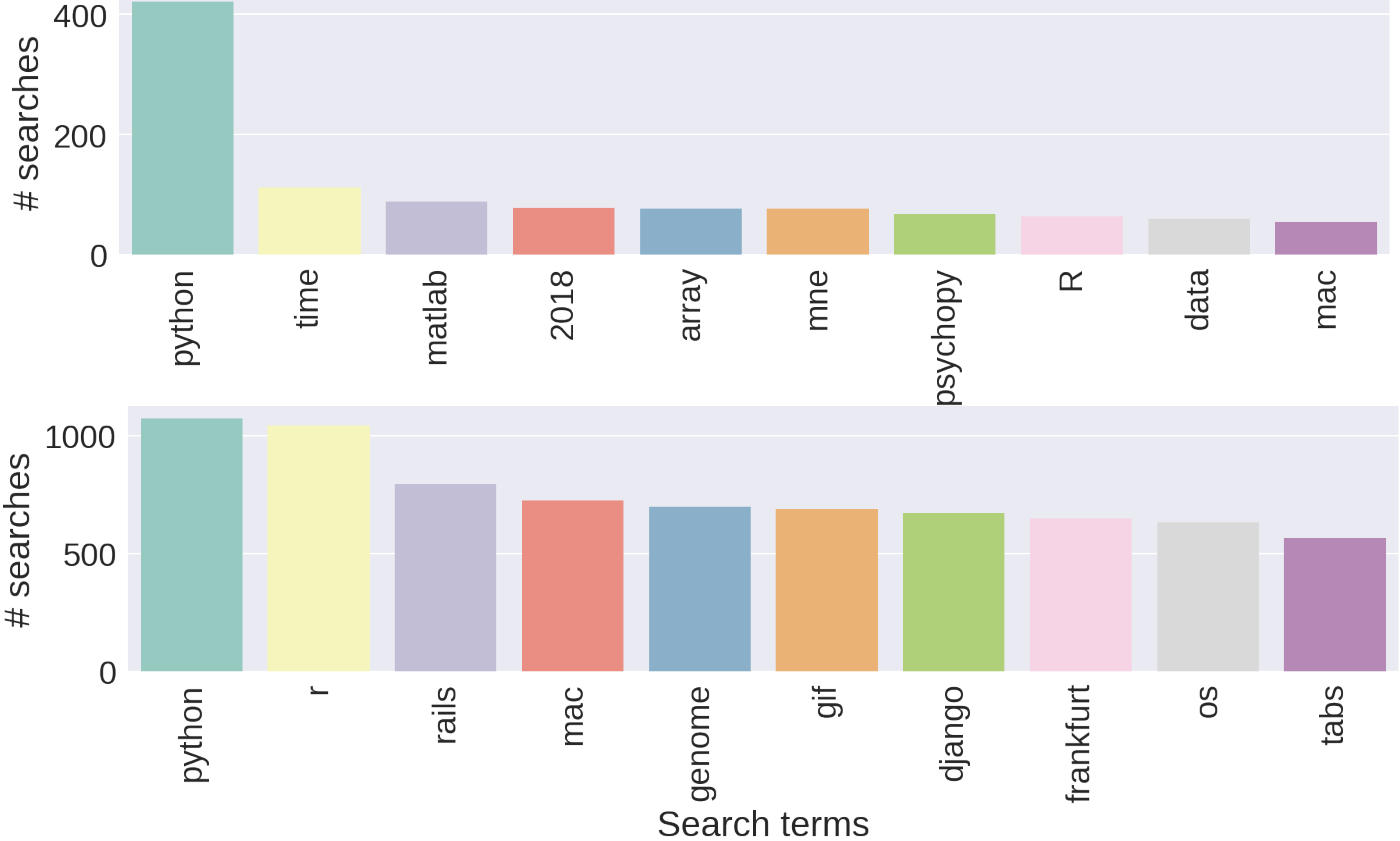


Importing & analyzing google search history data!

In this notebook you can analyze your google search history data. First, you have to request your data from google, with instructions that you can find here.

The goal is to get your data in a more usable format that what google provides when you're requesting it and be able to do some quantitative analyses yourself. And of course to get some nice plots!

All of the plots that are generated from this notebook will be saved in your Open Humans home folder, so check that out once you're done.


400
200
0
# searches
python
time
matlab
2018
array
mne
psychopy
R
data
mac
1000
500
0
# searches
python
r
rails
mac
genome
gif
django
frankfurt
os
tabs
Search terms

summary

- generalizable findings rely on sharing data
- there's plenty of databases for sharing your data (your discipline is missing amongst these? maybe start your own 😉)
- share data with a license that allows re-use and remixing, but…
- data sharing can be tricky when dealing with humans
- be as open as possible, as closed as necessary
- think about how you collect and distribute data (consent-less scraping is out)

thanks!

Tim Head

John Wilbanks

Philipp Bayer

Kirstie Whitaker

Mad Ball

Bastian Greshake Tzovaras
(@gedankenstuecke)