Data Sharing and processing infrastructures :

A Brain Imaging journey

Jean-Baptiste Poline
Jean-baptiste.poline@mcgill.ca

ORIGAMI lab
MNI, Brain Imaging Centre,
McGill, Montreal

But why on earth do I work on infrastructures ?

Jean-Baptiste Poline
Jean-baptiste.poline@mcgill.ca

ORIGAMI lab
MNI, Brain Imaging Centre,
McGill, Montreal

Slowed canonical progress in large fields of science

Johan S. G. Chu[a,1] and James A. Evans[b,c,d]

PNAS 2021

Examining 1.8 billion citations among 90 million papers across 241 subjects, we find a deluge of papers does not lead to turnover of central ideas in a field, but rather to ossification of canon.

THE LANCET

COMMENT | VOLUME 385, ISSUE 9976, P1380, APRIL 11, 2015

Offline: What is medicine's 5 sigma?

Richard Horton

Published: April 11, 2015 . DOI: https://doi.org/10.1016/S0140-6736(15)60696-1

Most common causes :

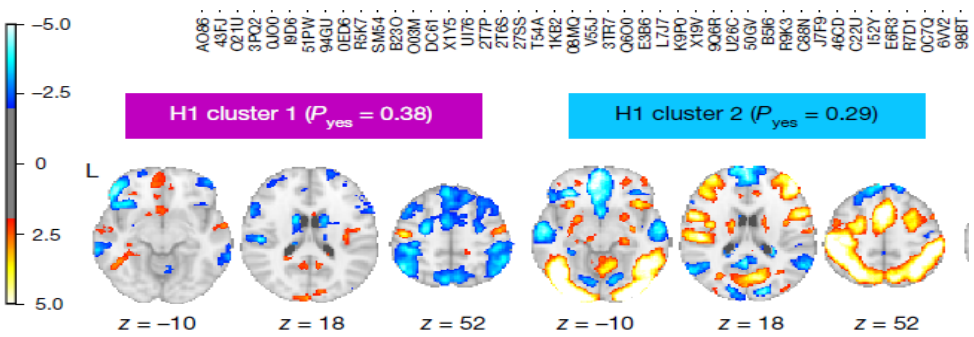
- Small sample size
- Analytical flexibility
- p-hacking
- Poor statistical training

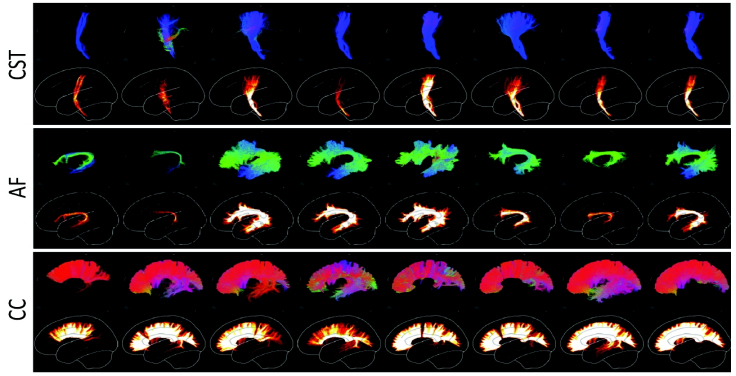
Many non reproducible papers

NIH plans to enhance reproducibility

Francis S. Collins and **Lawrence A. Tabak** discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

- Data sharing
- Statistical procedures
- Open access / open science
- ...





Many non reproducible papers : PD and MRI

Table 1: Reproduction and replication results (fractions of papers)

	Reproductions (n=3)	Replications (n=8)
Cohort Filtering		
C1 – Some cohort variable(s) not in PPMI	0.00	0.67
C2 – Groups missing subjects	0.33	0.27
C3 – Groups with demographic differences	0.42	0.44
C4 – Groups with clinical differences	0.50	0.39
Image Analysis Pipeline		
I1 – Main pipeline components available	0.67	0.83
I2 – Integrated pipeline available	0.67	0.00
I3 – Pipeline reproduced with high confidence	0.00	0.83
I4 – QC reproduced with high confidence	0.00	0.33
Statistical or ML Analysis Pipeline		
A1 – Original pipeline available	0.33	0.17
A2 – Pipeline reproduced with high confidence	1.00	0.67
Statistical Analyses		
S1 – Original paper reported statistical analyses	0.33	1.00
S2 – Analyses where significance was replicated	1.00	0.48
Machine Learning Models		
M1 – Original paper reported ML model(s)	1.00	0.00
M2 – Original ML metrics above chance-level	1.00	n/a
Fraction of obtained ML metrics:		
M3.1 – Comparable to or above original	0.29	n/a
M3.2 – Lower than original, above chance	0.38	n/a
M3.3 – At chance level	0.33	n/a
M4 – Original ML pipeline raised concerns	0.67	n/a

Part I: Reproducibility: background

Part II: Gap analysis

Part III: NeuroInformatics platforms

Part IV: Conclusion

Part I: Reproducibility: background

Part II: Gap analysis

Part III: NeuroInformatics platforms

Part IV: Conclusion

Amgen replication

- 53 papers examined at Amgen in preclinical cancer research
- Papers were selected that described something completely new and in very high impact factor journals
- **Scientific findings were confirmed in only 6 (11%)**

Begley and Ellis, Nature, 2012

JAMA Psychiatry 2017

Altered Brain Activity in Unipolar Depression Revisited Meta analyses of Neuroimaging Studies

Veronika I. Müller, PhD, Edna C. Cieslik, PhD, Ilinca Serbanescu, MSc, Angela R. Laird, PhD, Peter T. Fox, MD, and

RESULTS—In total, 57 studies with 99 individual neuroimaging experiments comprising in total 1058 patients were included; 34 of them tested cognitive and 65 emotional processing. Overall analyses across cognitive processing experiments ($P > .29$) and across emotional processing experiments ($P > .47$) revealed no significant results. Similarly, no convergence was found in analyses investigating positive (all $P > .15$), negative (all $P > .76$), or memory (all $P > .48$) processes. Analyses that restricted inclusion of confounds (eg, medication, comorbidity, age) did not change the results.

Followed the Genomic reproducibility crisis (GWAS) ?

Forensic Analysis

- Potti et al., Nat. Med. 2006, 2008 vs Baggerly and Coombes, “Forensic analysis”, Annals of applied Stat., 2009
- Choose cell lines that are most sensitive / resistant to a drug, use expression profiles to build a model that predicts patient response

Baggerly and Coombes Forensic:

*“**with poor documentation and irreproducibility even well meaning investigator may argue for drug that are contraindicated to some patients**”*

“**the most common errors are simple**
(e.g.,row or column offsets); conversely,
the most simple errors are common.”

In the general public

Essay

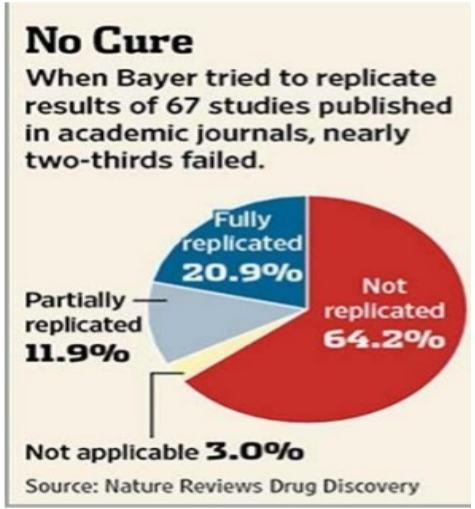
Why Most Published Research Findings Are False

John P. A. Ioannidis

2005. PLoS Medicine, 2(8), e124. doi: 10.1371/journal.pmed.0020124

"There is increasing concern about the reliability of biomedical research, with recent articles suggesting that up to 85% of research funding is wasted."

Bustin, S. A. (2015). The reproducibility of biomedical research: Sleepers awake! *Biomolecular Detection and Quantification*



Our love of "significance" pollutes the literature with many a statistical fairytale. We reject important confirmations. In their quest for telling a compelling story, scientists too often sculpt data to fit their preferred theory of the world.

Funding agencies reaction

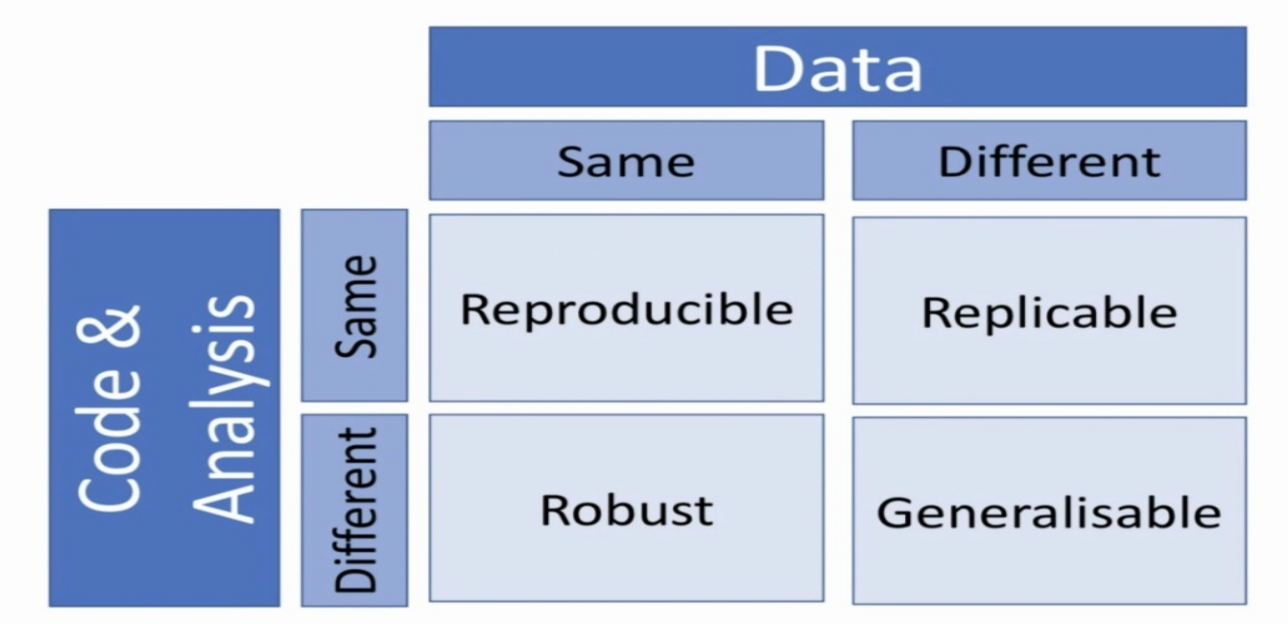
NIH plans to enhance reproducibility

Francis S. Collins and **Lawrence A. Tabak** discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

Collins and Tabak. 2014. Nature 505: 612–13.

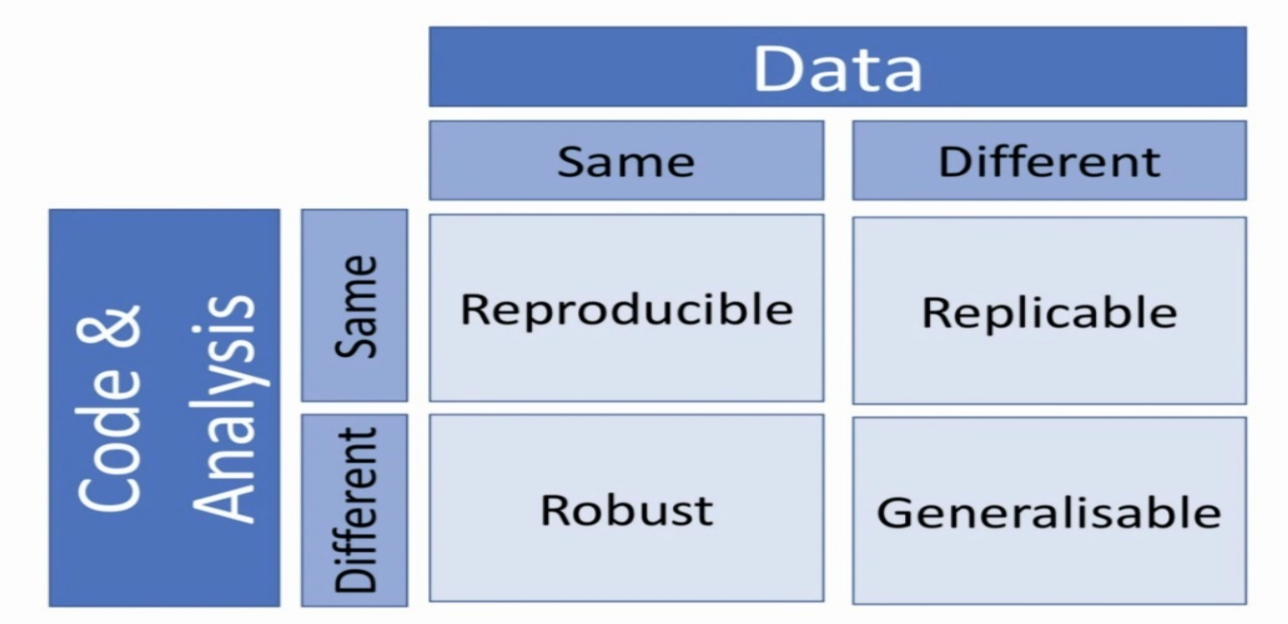
Know How To Address Rigor and Reproducibility in Your Next Application

Funding News Edition: **June 17, 2020**



Credit: UK software institute

Think **generalizability** across Data, Software, Time, Scanner, Stimuli, etc



The Turing institute

Think **generalizability** across Data, Software, Time, Population, Scanners, etc

Part I: Reproducibility: background

Part II: Gap analysis

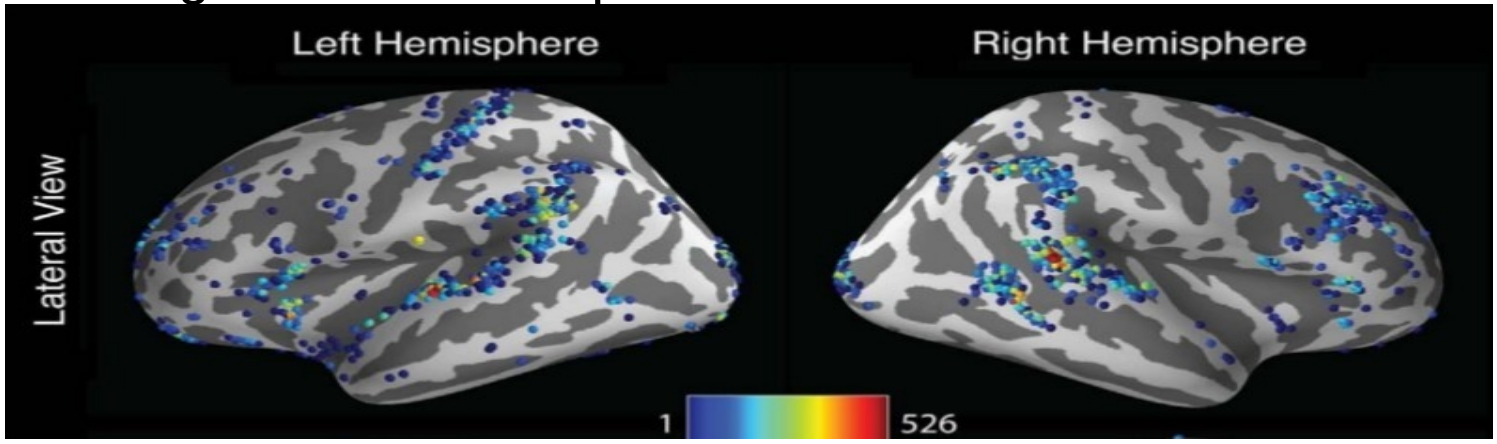
Part III: NeuroInformatics platforms

Part IV: Conclusion

Analytical flexibility

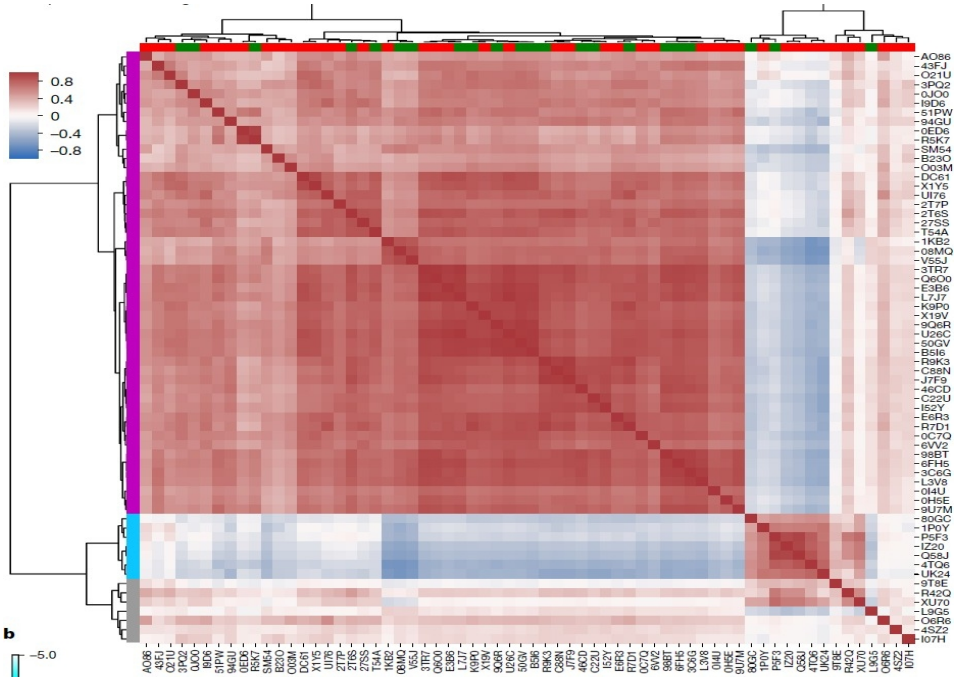
J. Carp, f. Neuroscience, 2012

- A **single** event-related fMRI experiment to a large number of unique analysis procedures
- Ten analysis steps for which multiple strategies appear in the literature : **6,912 pipelines**
- Plotting the maximum peak



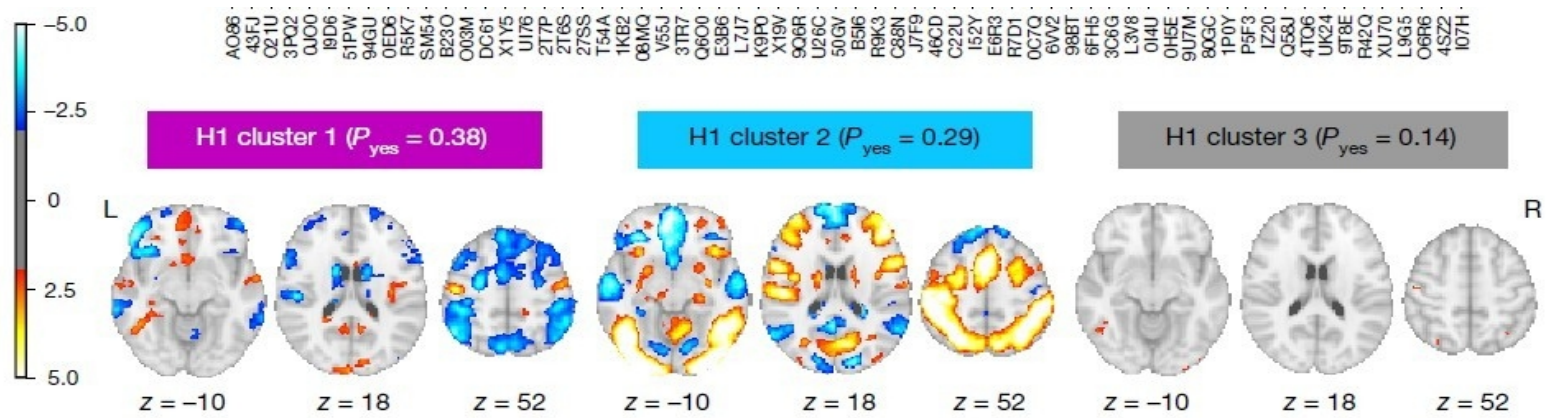
Analytical flexibility

- 70 independent teams analyzing *the same* fMRI dataset
- No team had the same pipeline
  - Results show three “clusters”
- Even within clusters decision to reject H0 varies



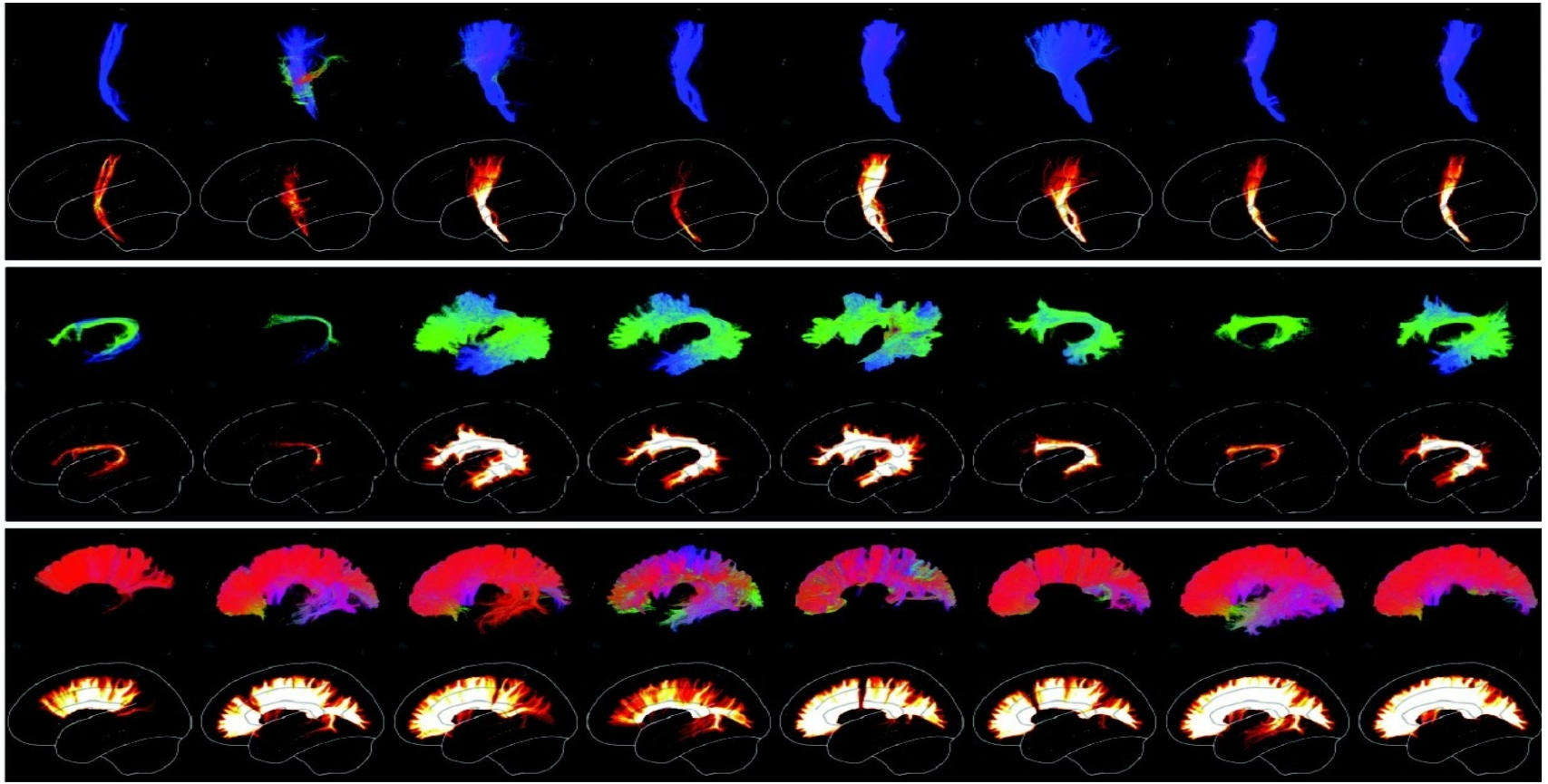
Botvinik-Nezer, 2020

Analytical flexibility



Botvinik-Nezer, 2020

Tractography: what happens when 42 groups dissect 14 white matter bundles on the same dataset?

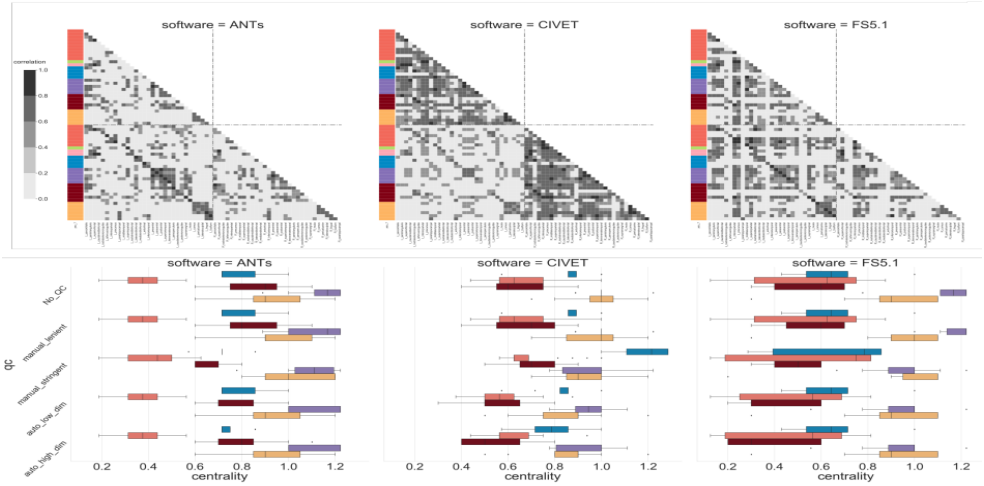


Kurt G. Schilling, … M. Descotaux, bioRxiv, 2020

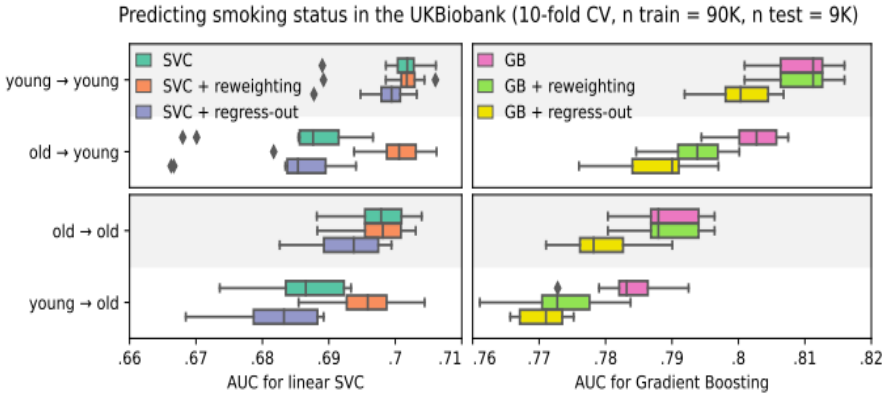
Analytical flexibility

- Open data with ABIDE
- Measuring brain region connectivity with different segmentation pipelines
- Poor correlation between pipelines and very different networks found

Nikhil Baghwat, Gigascience 2020



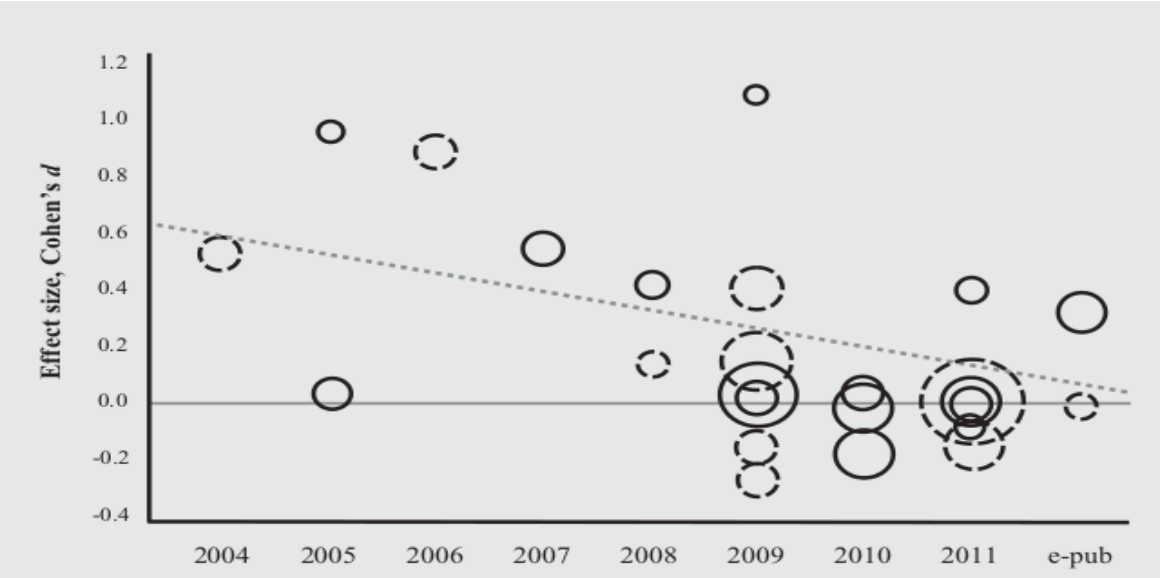
Dataset shift issues



Jerome Dockès, Gigascience 2021

- UK Biobank: N=500,000
- Machine Learning with simple or complex learners
- Change in population demographics and usual corrections methods can lead to significant loss in accuracy

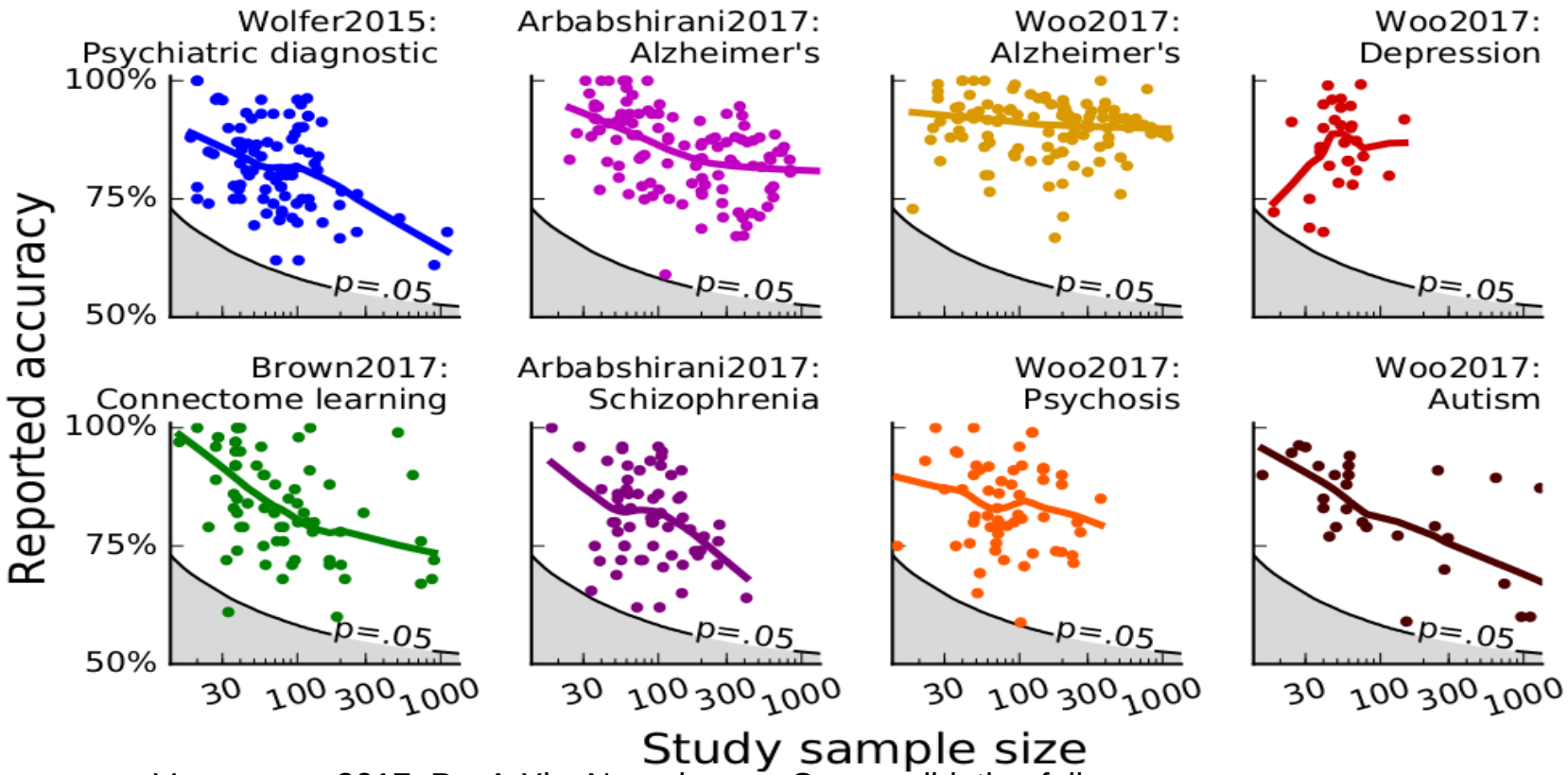
Sample size: One problem



Molendijk, 2012: BDNF and hippocampal volume

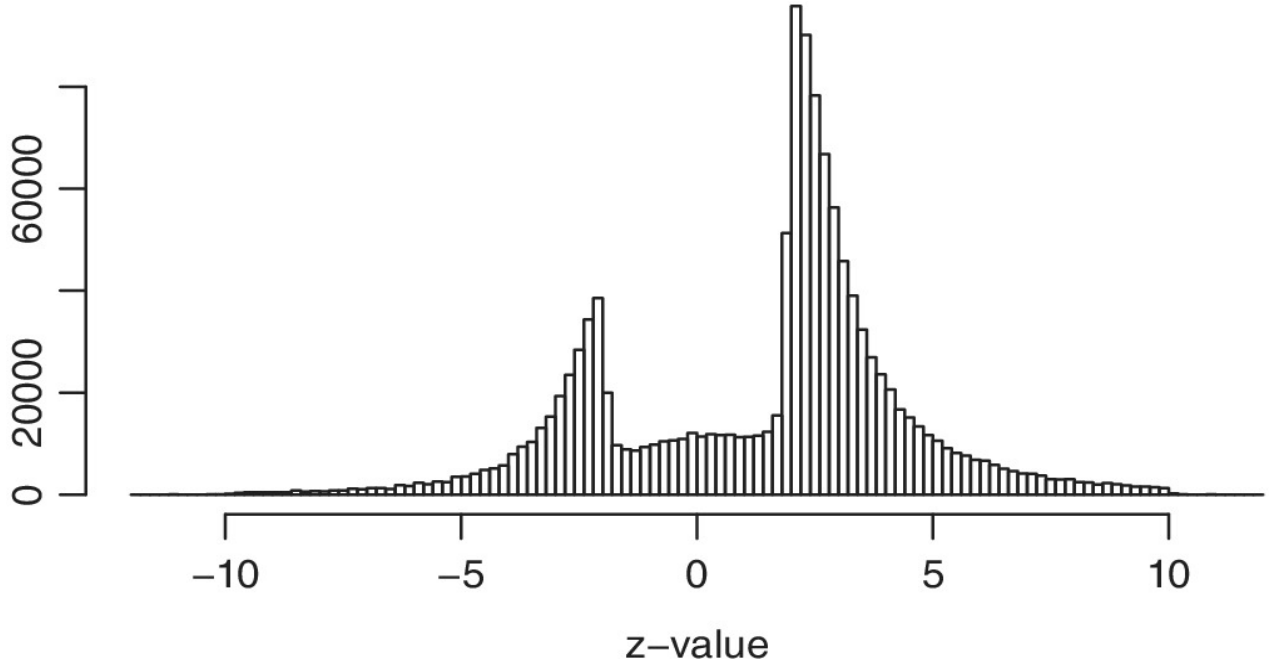
See also : Mier, 2009: COMT and DLPFC

Sample size: One problem



Varoquaux, 2017, PsyArXiv, Neuroimage, Cross validation failure

Z-values reported in Medline



The distribution of more than one million z-values from Medline (1976–2019)

Zwet, E.W. van, Cator, E.A., n.d. The significance filter, the winner's curse and the need to shrink. Statistica Neerlandica n/a. https://doi.org/10.1111/stan.12241

Part I: Reproducibility: background

Part II: Gap analysis

Part III: NeuroInformatics platforms

Part IV: Conclusion

Principles for the next generation: FAIR – and open infrastructures

- Implementation : good software dev practice + versatile and modular ecosystem
- Sustainability : transparent funding mechanism, distributed development
- Distributed governance (e.g BIDS governance)
- Building communities and standards:
  - GH4GH, INCF, W3C, BIDS++, etc

Proposed workflow and ecosystem

eLife

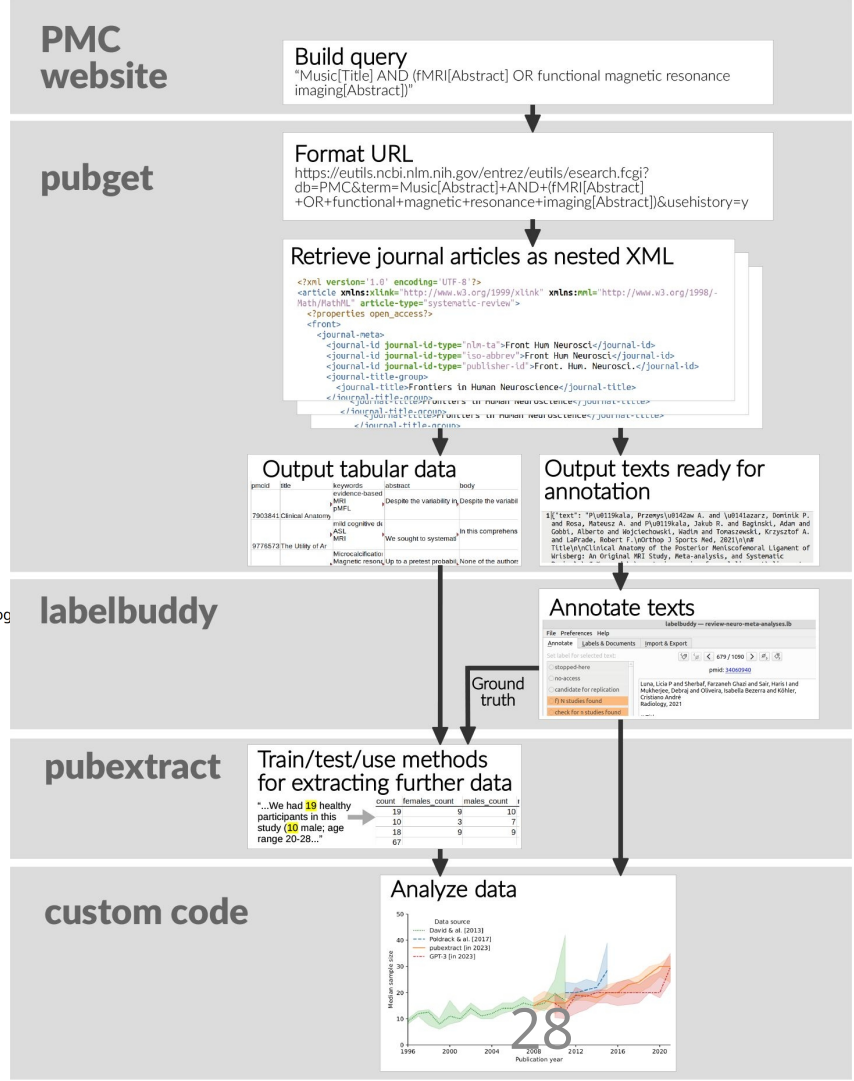
Neuroscience

Mining the neuroimaging literature

Jérôme Dockès, Kendra Oudyk, Mohammad Torabi, Alejandro I de la Vega, Jean-Baptiste Poline

National Institute for Research in Digital Science and Technology (INRIA), Paris, France • Montreal Neurolog
Institute, McGill University, Montreal, Canada • University of Texas at Austin, Austin, Texas, USA

https://doi.org/10.7554/eLife.94909.1



PMC website

pubget

Build query

"Music[Title] AND (fMRI[Abstract] OR functional magnetic resonance imaging[Abstract])"

Format URL

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=PMC&term=Music[Abstract]+AND+(fMRI[Abstract]+OR+functional+magnetic+resonance+imaging[Abstract])&usehistory=y

Retrieve journal articles as nested XML

```
<?xml version='1.0' encoding='UTF-8'?>
<article xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:mml="http://www.w3.org/1998/-
Math/MathML" article-type="systematic-review">
  <?properties open_access?>
  <front>
    <journal-meta>
      <journal-id journal-id-type="nlm-ta">Front Hum Neurosci</journal-id>
      <journal-id journal-id-type="iso-abbrev">Front Hum Neurosci</journal-id>
      <journal-id journal-id-type="publisher-id">Front. Hum. Neurosci.</journal-id>
      <journal-title-group>
        <journal-title>Frontiers in Human Neuroscience</journal-title>
      </journal-title-group>
```

Output tabular data

pmcid	title	keywords	abstract	body
		evidence-based MRI	Despite the variability in	Despite the variabil

Output texts ready for annotation

pubget

Format URL

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=PMC&term=Music[Abstract]+AND+(fMRI[Abstract]+OR+functional+magnetic+resonance+imaging[Abstract])&usehistory=y

Retrieve journal articles as nested XML

```
<?xml version='1.0' encoding='UTF-8'?>
<article xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:mml="http://www.w3.org/1998/-
Math/MathML" article-type="systematic-review">
  <?properties open_access?>
  <front>
    <journal-meta>
      <journal-id journal-id-type="nlm-ta">Front Hum Neurosci</journal-id>
      <journal-id journal-id-type="iso-abbrev">Front Hum Neurosci</journal-id>
      <journal-id journal-id-type="publisher-id">Front. Hum. Neurosci.</journal-id>
      <journal-title-group>
        <journal-title>Frontiers in Human Neuroscience</journal-title>
      </journal-title-group>
```

Output tabular data

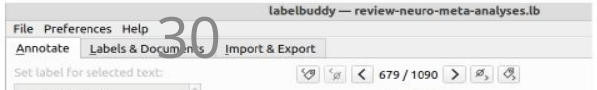
pmcid	title	keywords	abstract	body
7903841	Clinical Anatomy	evidence-based MRI pMFL	Despite the variability in	Despite the variabil
9776573	The Utility of Ar	mild cognitive de ASL MRI	We sought to systemati	In this comprehens
		Microcalcificatio Magnetic reson	Up to a pretest probabil	None of the authors

Output texts ready for annotation

```
1{"text": "P\u0119kala, Przemys\u0142aw A. and \u0141azarz, Dominik P.
 and Rosa, Mateusz A. and P\u0119kala, Jakub R. and Baginski, Adam and
 Gobbi, Alberto and Wojciechowski, Wadim and Tomaszewski, Krzysztof A.
 and LaPrade, Robert F.\nOrthop J Sports Med, 2021\n\n#
 Title\n\nClinical Anatomy of the Posterior Meniscofemoral Ligament of
 Wrisberg: An Original MRI Study, Meta-analysis, and Systematic
```

labelbuddy

Annotate texts



Output tabular data

pmcid	title	keywords	abstract	body
7903841	Clinical Anatomy	evidence-based MRI pMFL	Despite the variability in	Despite the variabil
9776573	The Utility of Ar	mild cognitive de ASL MRI	We sought to systemati	In this comprehens
		Microcalcificatio Magnetic reson	Up to a pretest probabil	None of the authors

Output texts ready for annotation


```
1{"text": "P\u0119kala, Przemys\u0142aw A. and \u0141azarz, Dominik P.
 and Rosa, Mateusz A. and P\u0119kala, Jakub R. and Baginski, Adam and
 Gobbi, Alberto and Wojciechowski, Wadim and Tomaszewski, Krzysztof A.
 and LaPrade, Robert F.\nOrthop J Sports Med, 2021\n\n#
 Title\n\nClinical Anatomy of the Posterior Meniscofemoral Ligament of
 Wrisberg: An Original MRI Study, Meta-analysis, and Systematic
```


labelbuddy

Annotate texts

Ground truth

pubextract

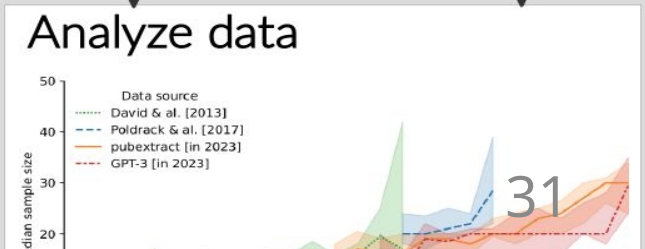
Train/test/use methods for extracting further data

"...We had 19 healthy participants in this study (10 male; age range 20-28..."

count	females_count	males_count
19	9	10
10	3	7
18	9	9
67		

custom code

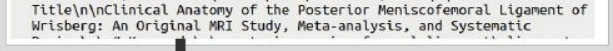
Analyze data



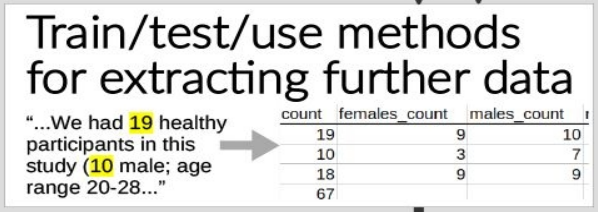
labelbuddy

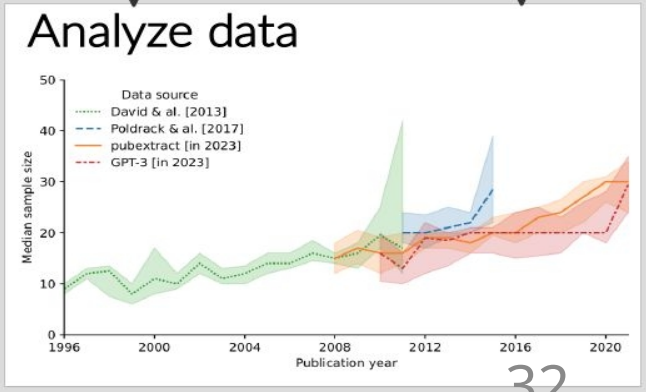
pubextract

custom code

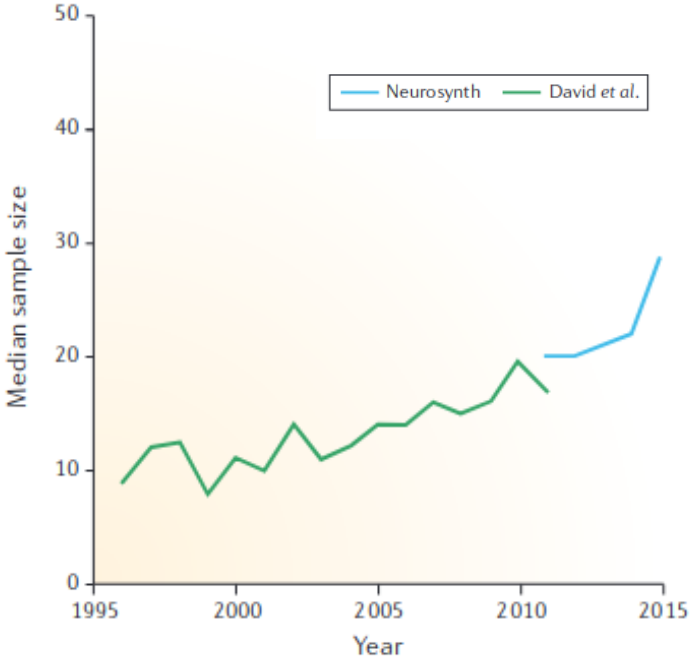


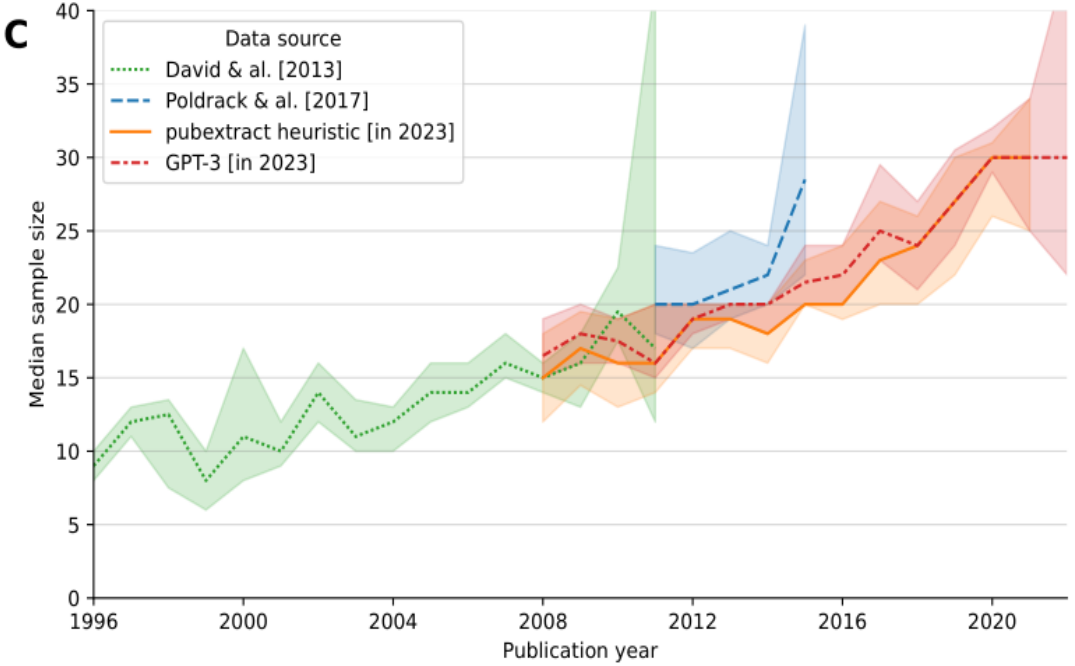
Ground truth





Example: Replicating and extending Poldrack et al. (2017)






Alyssa Dai


Nikhil Bhagwat


Remi Gau


Arman Jahanpour


JB Poline


Sebastian Urchs


Michelle Wang

neurobagel
neurobagel.org

ORIGAMI Lab

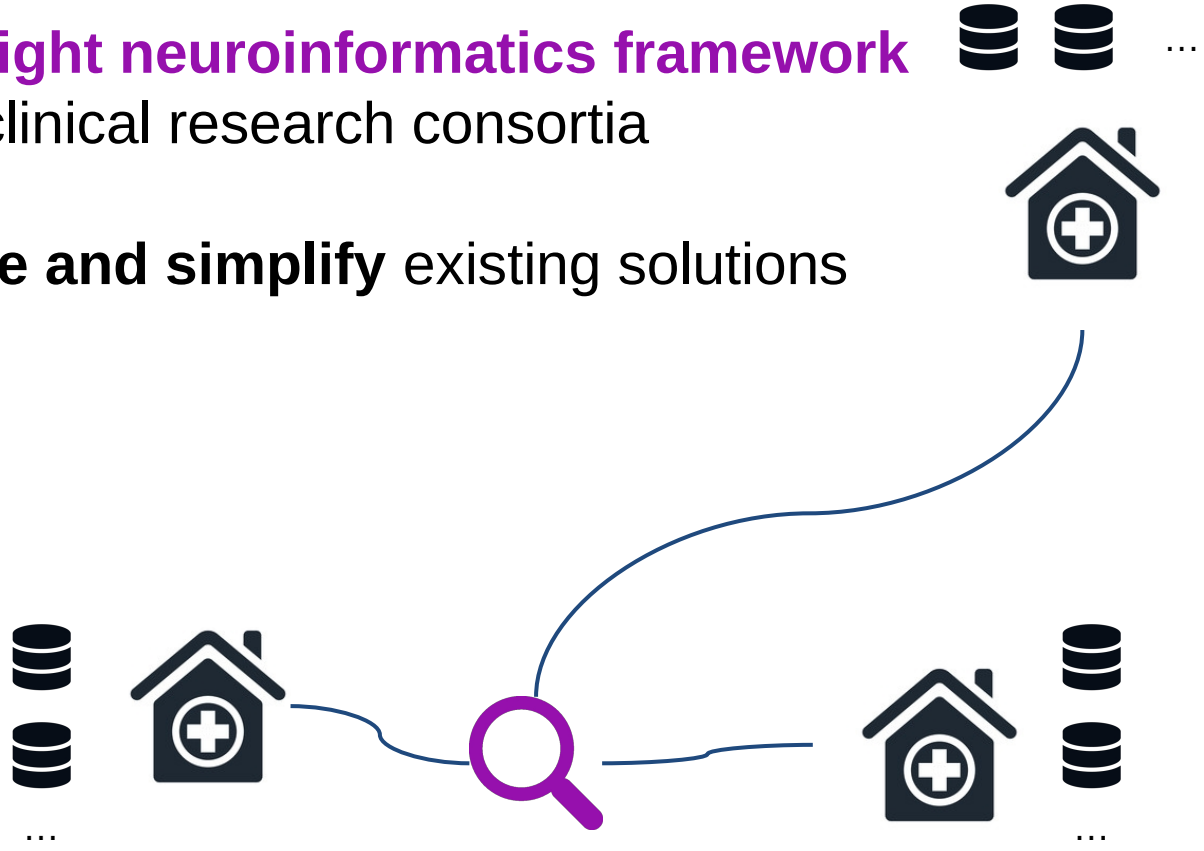
Neurobagel and NiPoppy

Data processing harmonization and discovery

HEALTHY BRAINS HEALTHY LIVES
neuro Montreal Neurological Institute-Hospital
CONP PCNO
NIH National Institutes of Health
Fondation Brain Canada Foundation
NIBIB

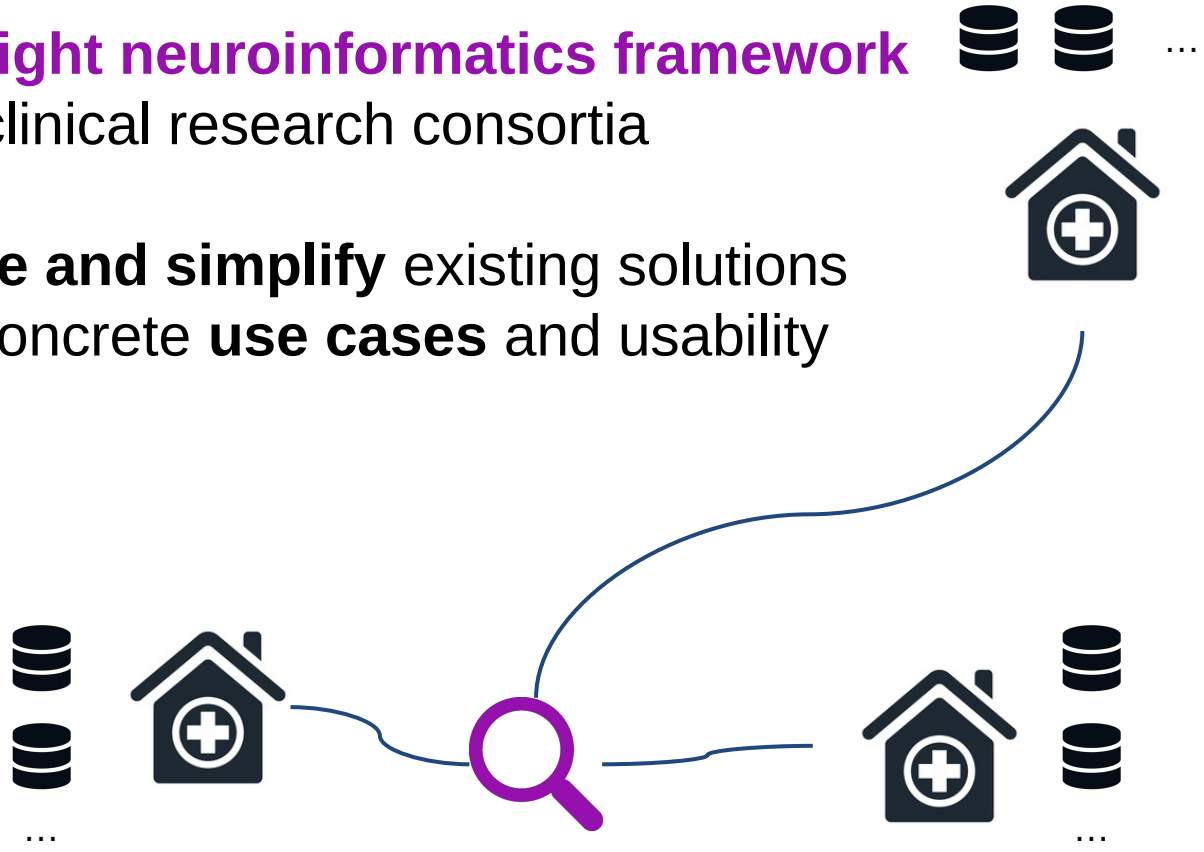
Adopt a lightweight neuroinformatics framework
for multi-centric clinical research consortia

What: **Streamline and simplify** existing solutions



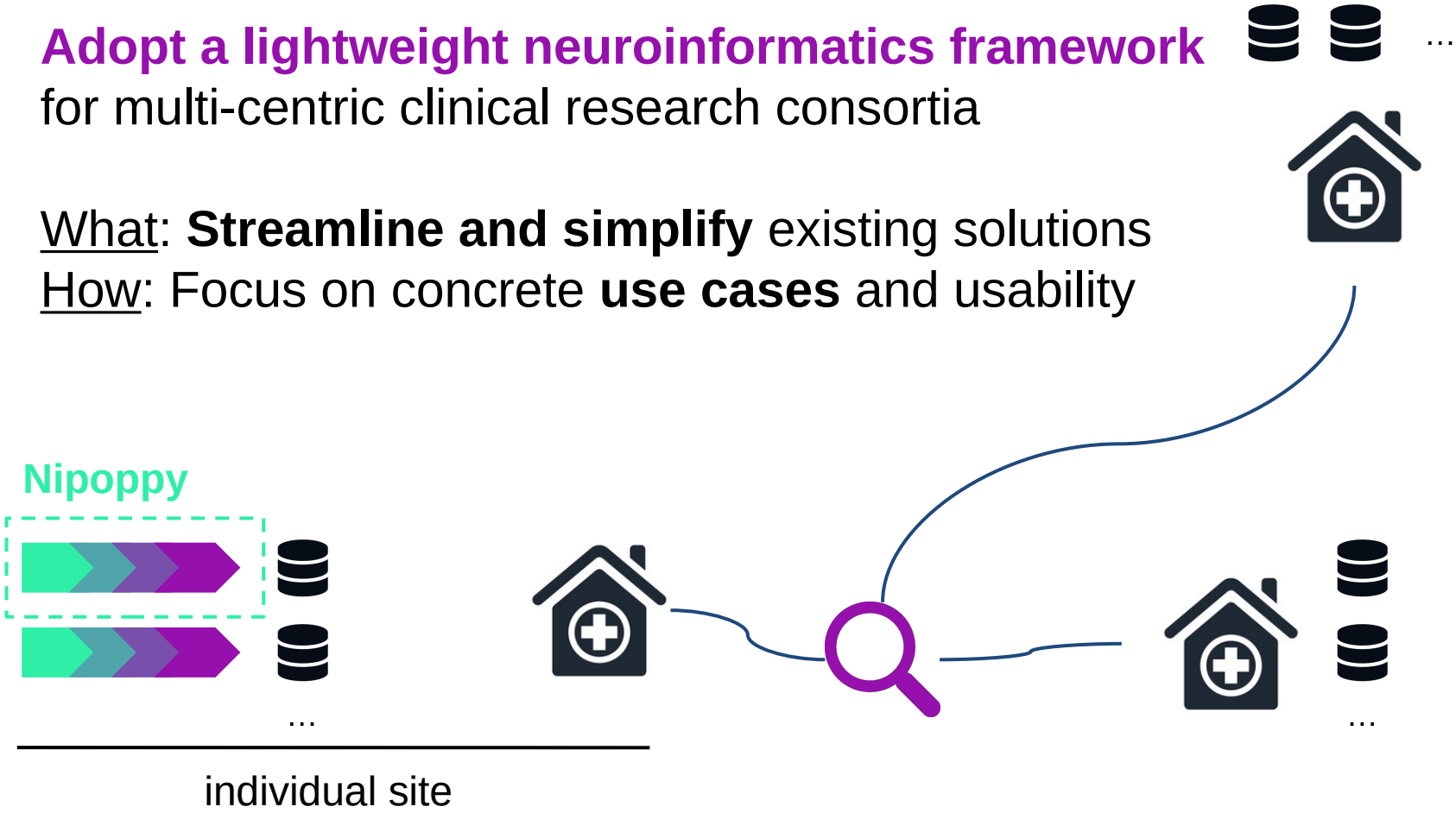
Adopt a lightweight neuroinformatics framework
for multi-centric clinical research consortia

What: **Streamline and simplify** existing solutions
How: Focus on concrete **use cases** and usability



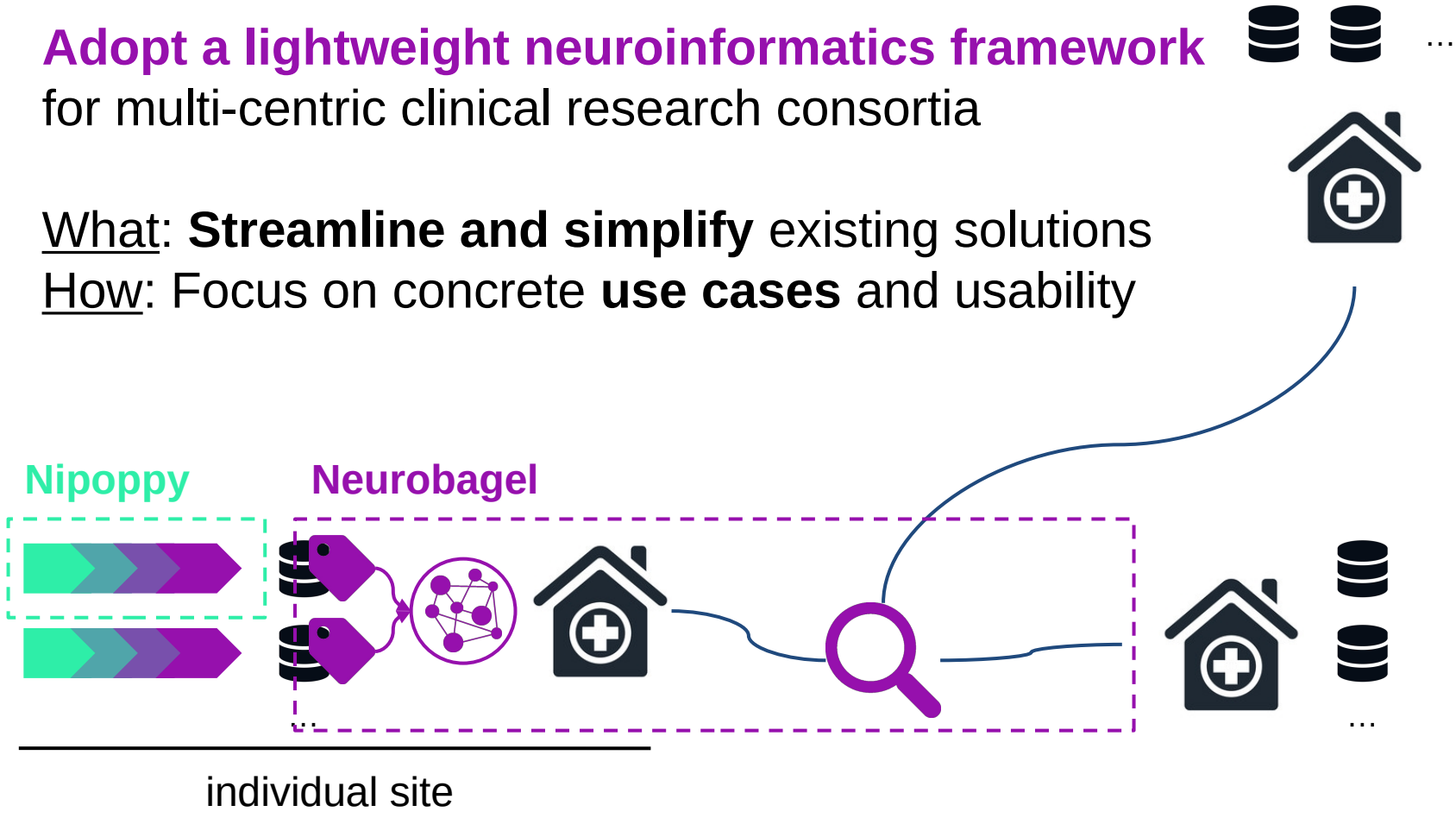
Adopt a lightweight neuroinformatics framework
for multi-centric clinical research consortia

What: **Streamline and simplify** existing solutions
How: Focus on concrete **use cases** and usability



Adopt a lightweight neuroinformatics framework
for multi-centric clinical research consortia

What: **Streamline and simplify** existing solutions
How: Focus on concrete **use cases** and usability



Nipoppy: protocol for the individual dataset

Capture | **Organize** | **Process** | **Track** | **Extract**

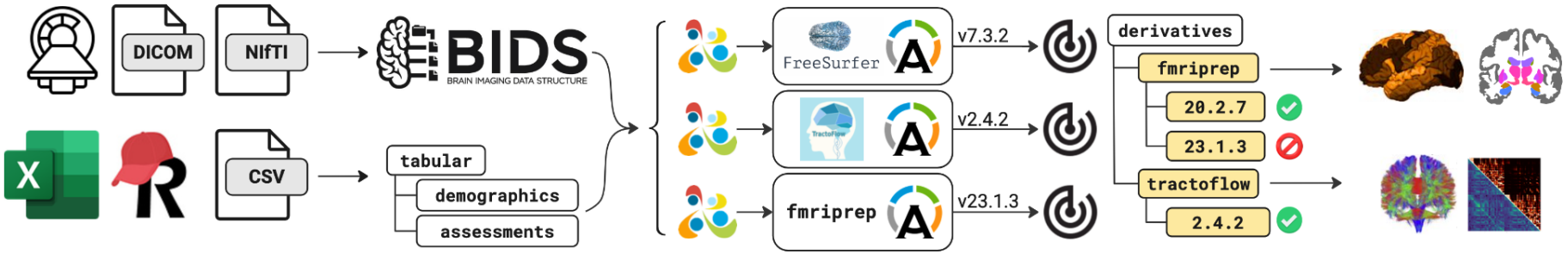
Take data **as they are**

Organize/curate **imaging** and **non-imaging** data

Use **reproducible** environments and parameters

Check data availability at the **participant level**

Get **analysis-ready** imaging-derived phenotypes (IDPs)



- **Extend BIDS** for phenotypic data and imaging derivatives

- **Containerized** pipelines
- **Provenance** information
- **Flexibility**: run many types of pipelines

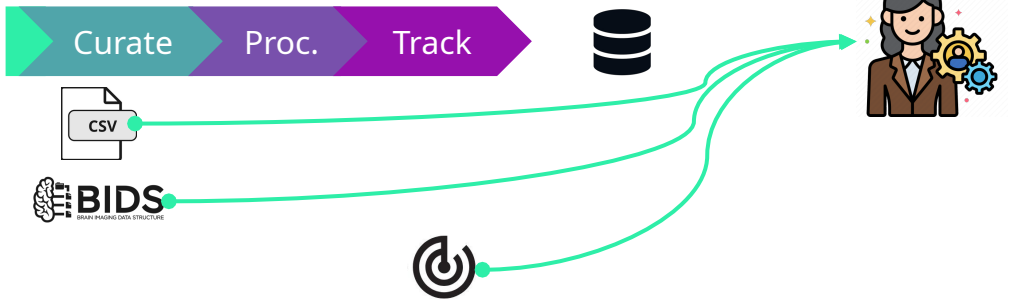
- **Output-aware** trackers
- **Participant-level** status
- Web **dashboard**

end to end framework for processing of **single dataset**

Nipoppy individual site

How many participants
- with PD diagnosis
- under 65
- 2 imaging sessions
- run with freesurfer v6

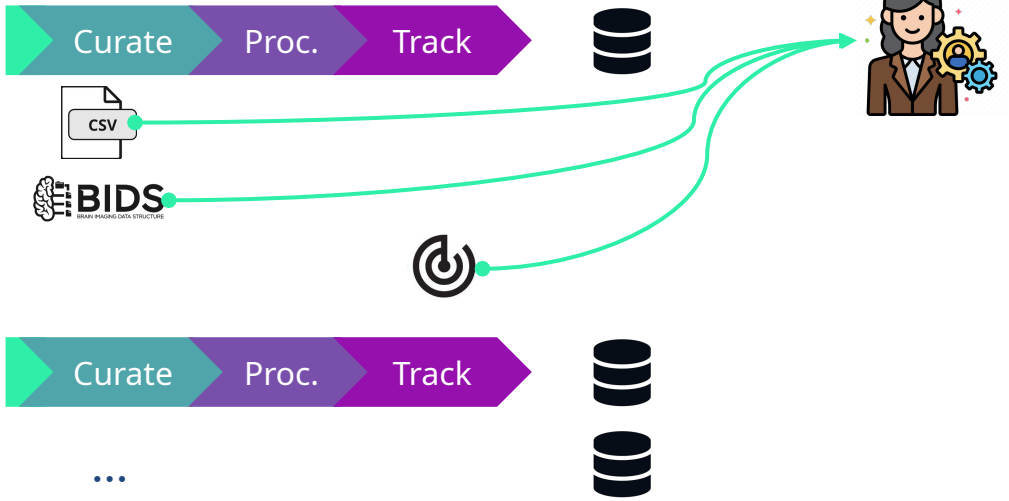
are in **this one dataset?**



Nipoppy individual site

How many participants
- with PD diagnosis
- under 65
- 2 imaging sessions
- run with freesurfer v6

do our PIs have?



Neurobagel  individual site

How many participants
- with PD diagnosis
- under 65
- 2 imaging sessions
- run with freesurfer v6

do our PIs have?

Curate Proc. Track

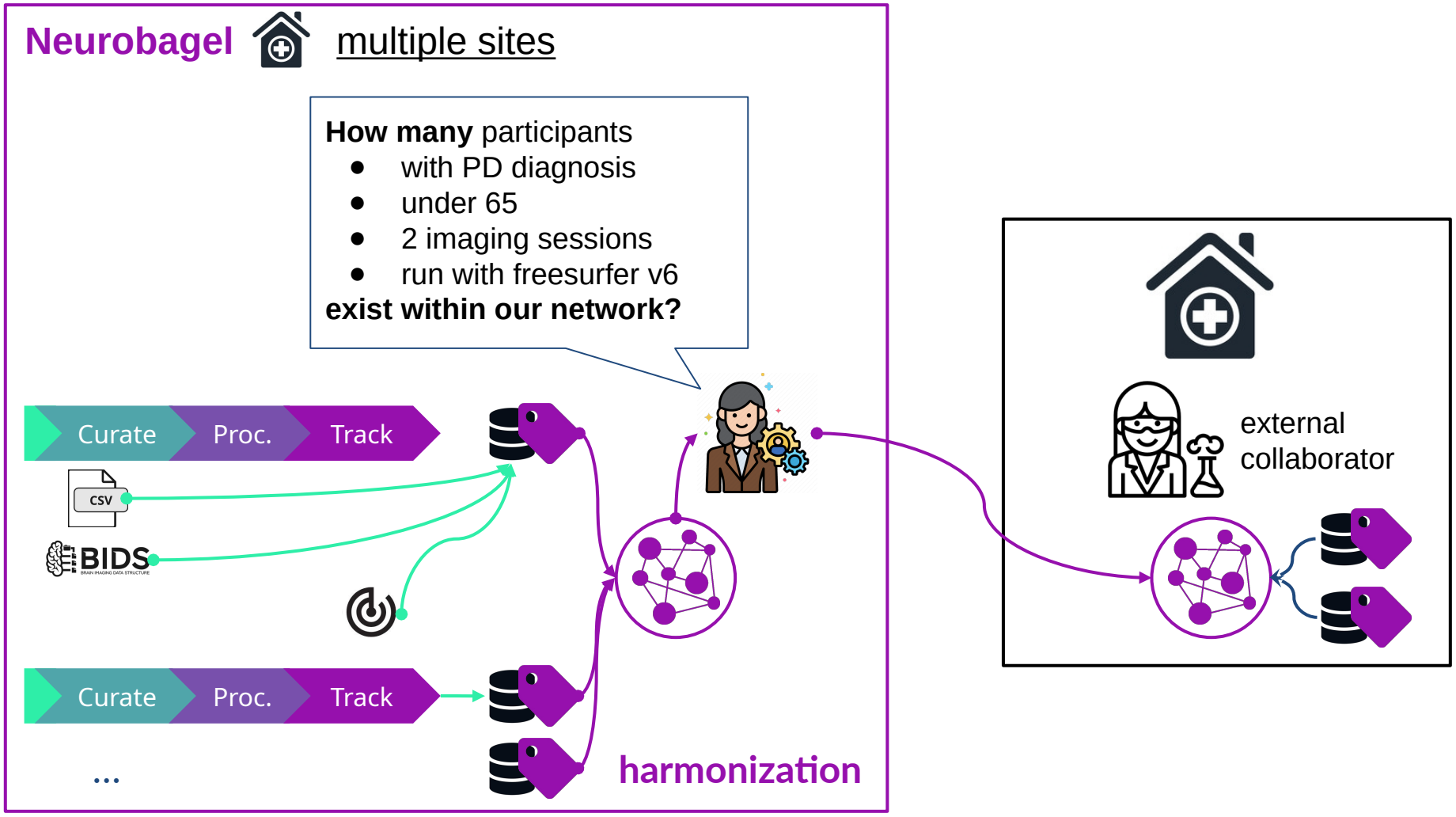
CSV

BIDS

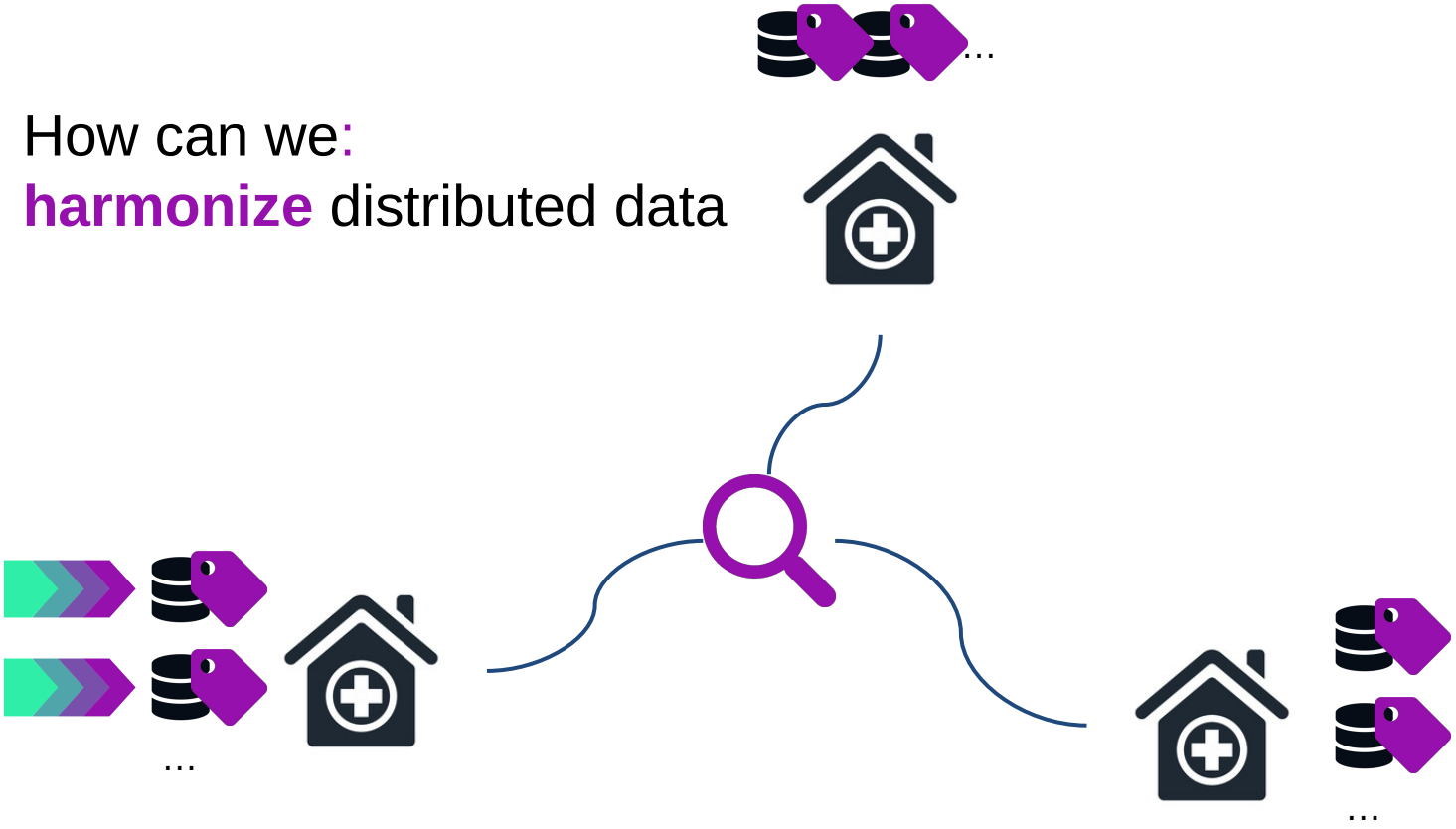
Curate Proc. Track

...

harmonization

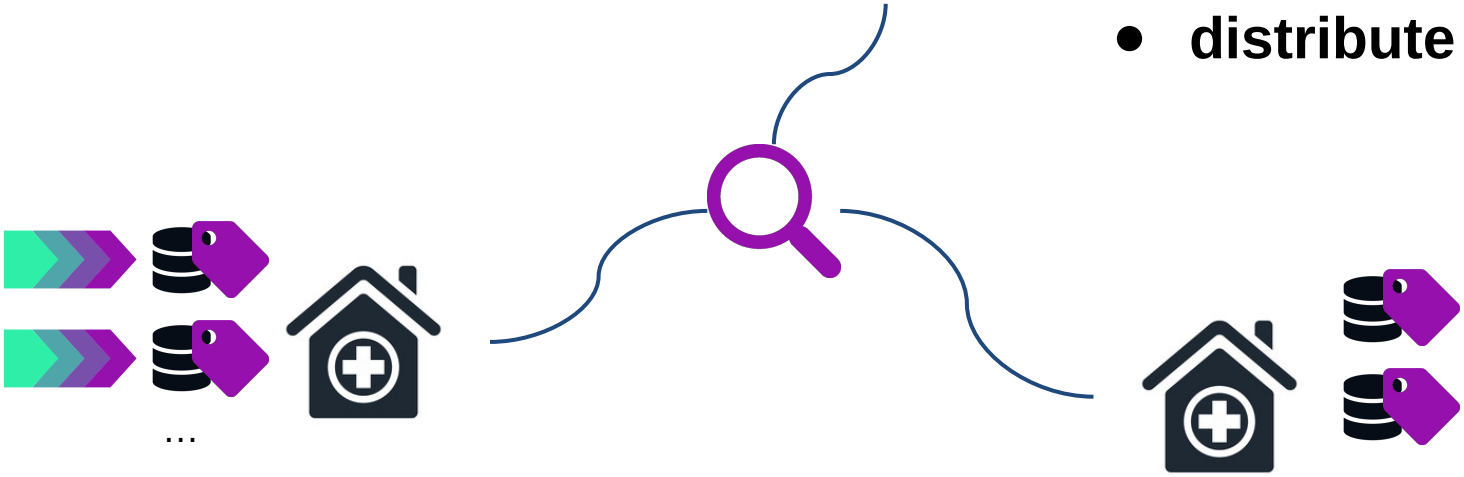

Neurobagel
multiple sites
How many participants
with PD diagnosis
under 65
2 imaging sessions
run with freesurfer v6
exist within our network?
Curate
Proc.
Track
CSV
BIDS
Curate
Proc.
Track
...
harmonization
external
collaborator

How can we:
harmonize distributed data



How can we:
harmonize distributed data

- **use common data model**
- **distribute intuitive GUI**



SNOMED CT
The global language of healthcare

RDF

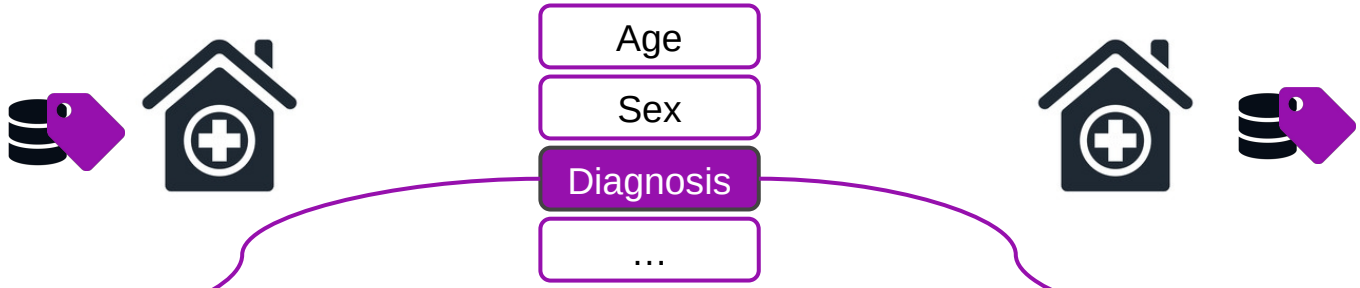
W3C
Word Wide Web Consortium

SPARQL

BIDS
BRAIN IMAGING DATA STRUCTURE

Findable Accessible Interoperable Reusable

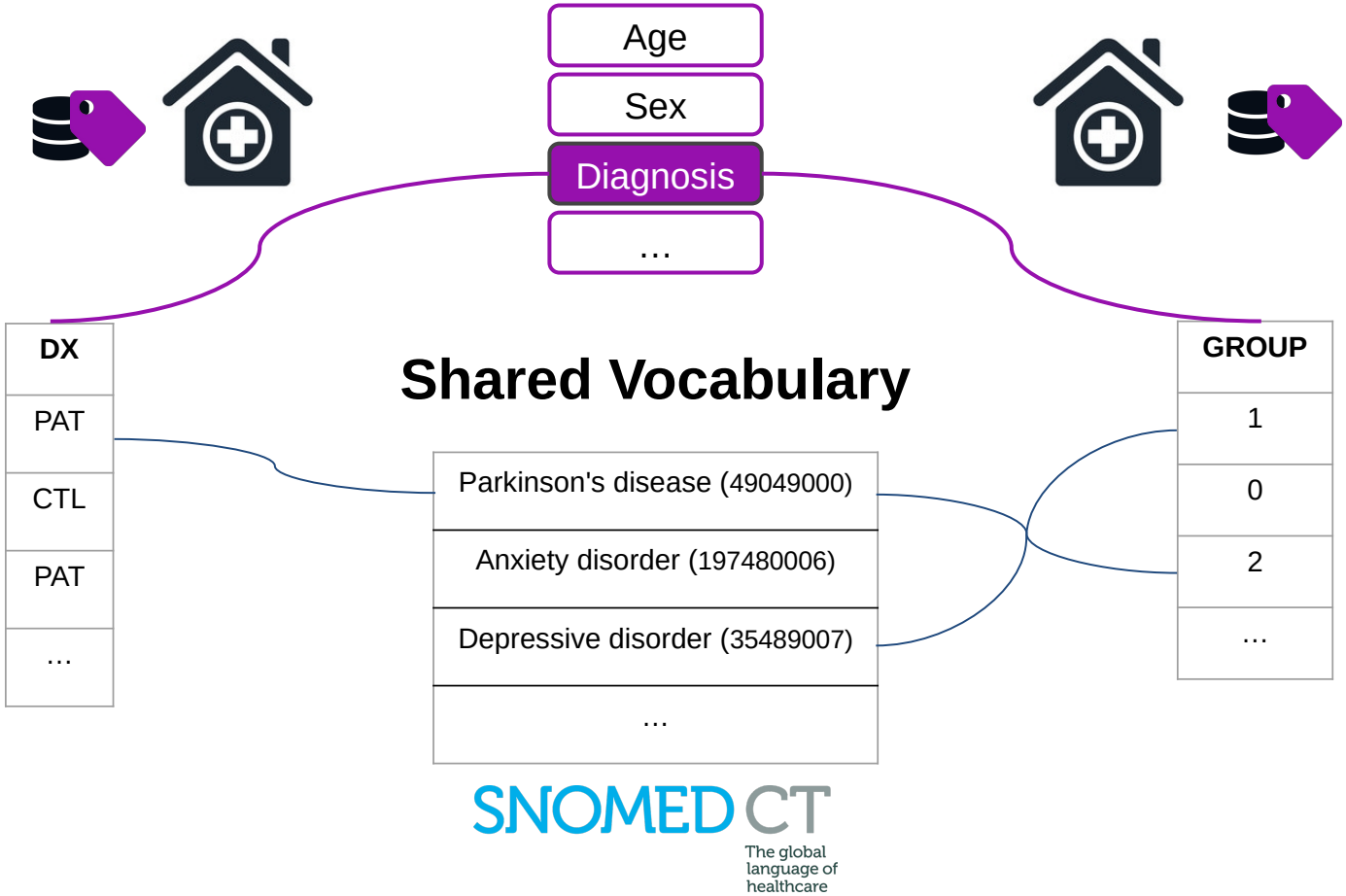
Common participant level data model



DX
PAT
CTL
PAT
…

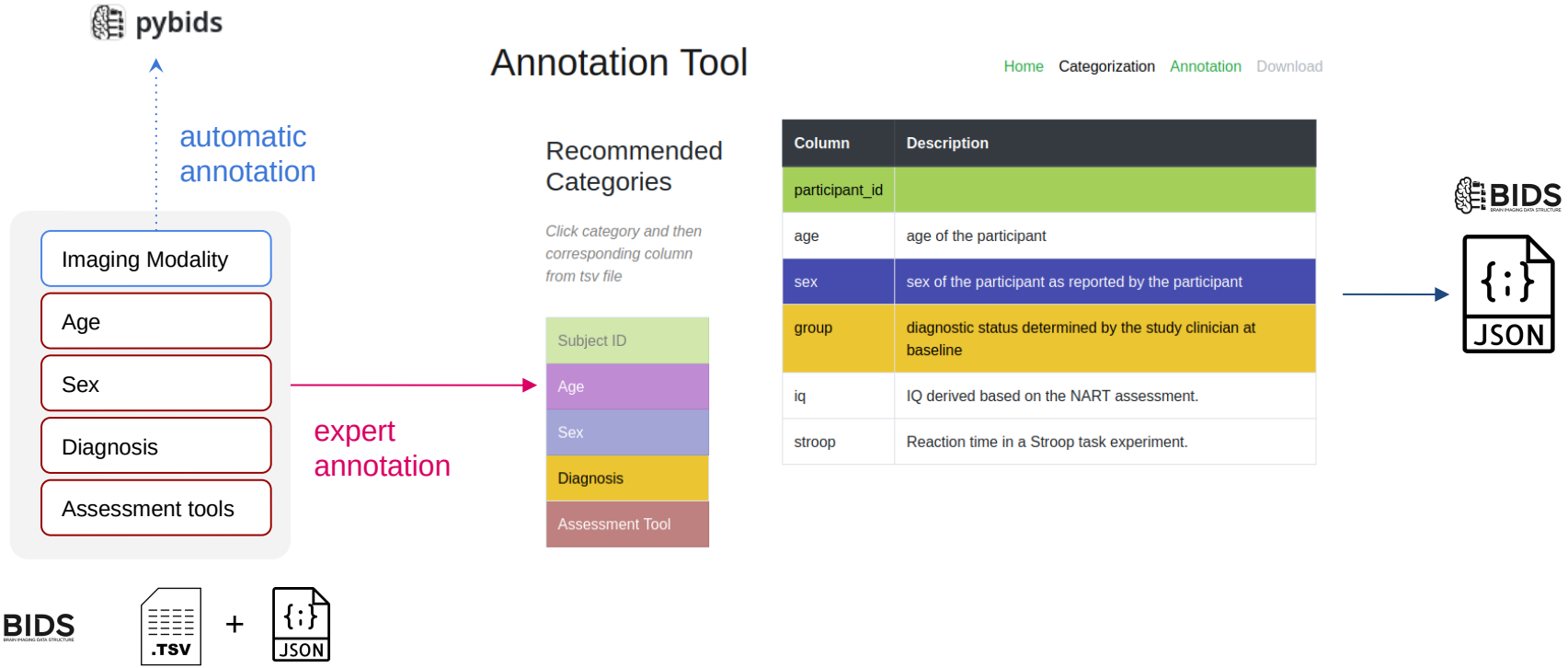
GROUP
1
0
2
…

Common participant level data model



Distribute model as graphical annotation tool

(that makes / augments BIDS data dictionaries)



Use existing controlled vocabularies

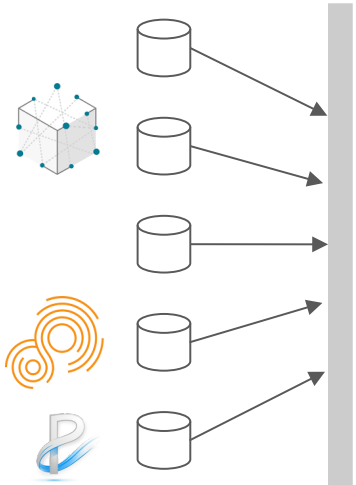
Limited and specified terms for query

imaging metadata

```
sub-control01/
  anat/
    sub-control01_T1w.nii.gz
```

observational metadata

age	gender	group
44	f	depr_no_treatment
21	f	depr_cbt
28	m	depr_no_treatment

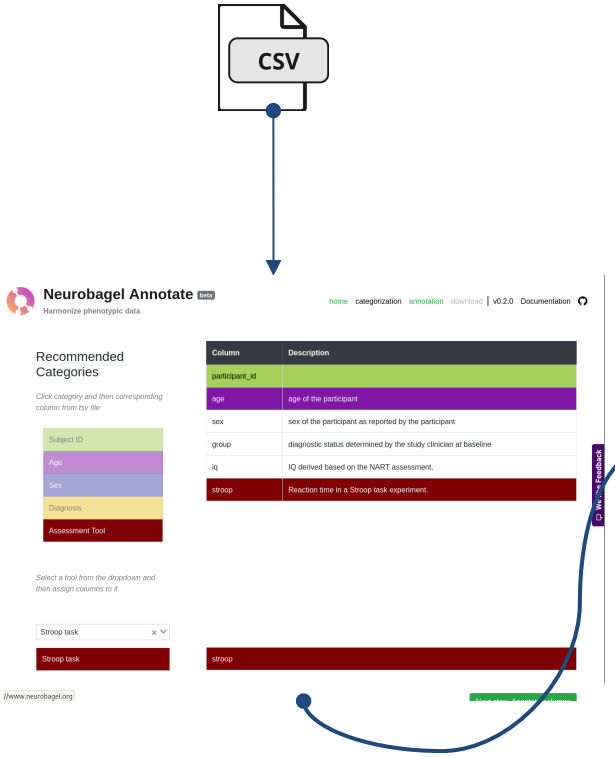


1: Integration		2: Harmonization	
Imaging Modality	BIDS		NIDM
Age	“019Y”	19.0	years (int, float) BIDS
Gender	“M”	“male”	vocab term (string)
Diagnosis	“PD”	“Parkinson’s disease”	SNOMED-CT (**URI**) SNOMED International
Assessment tools	“MCTOT”	Montreal Cognitive Assessment	cognitiveAtlas (**URI**)

local **metadata**

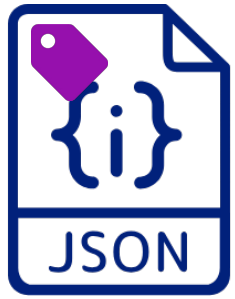
gathering the metadata (where)

what do your metadata **mean**




```
{
  "age": {
    "Description": "age of the participant",
    "Units": "years",
    "Annotations": {
      "IsAbout": {
        "TermURL": "http://purl.obolibrary.org/obo/NCIT_C25150",
        "Label": "Age"
      },
      "Unit": {
        "TermURL": "xsd:integer",
        "Label": "Integer"
      }
    },
  "sex": {
    "Description": "sex of the participant as reported by the participant",
    "Levels": {
      "M": "male",
      "F": "female"
    },
    "Annotations": {
      "IsAbout": {
        "TermURL": "http://purl.obolibrary.org/obo/NCIT_C28421",
        "Label": "Sex"
      },
      "Levels": {
        "M": {
          "TermURL": "http://purl.obolibrary.org/obo/NCIT_C20197",
          "Label": "Male"
        },
        "F": {
          "TermURL": "http://purl.obolibrary.org/obo/NCIT_C16576",
          "Label": "Female"
        }
      },
      "MissingValues": [
        "",
        " "
      ]
    }
  }
}
```

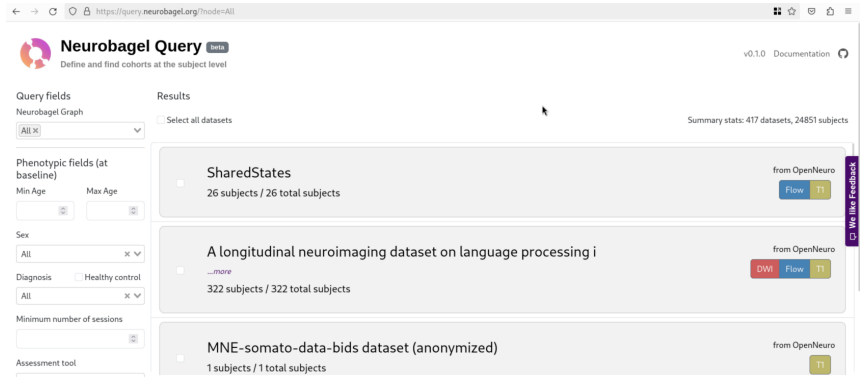
controlled terms



semantic
data dictionary

How can we:
search distributed data

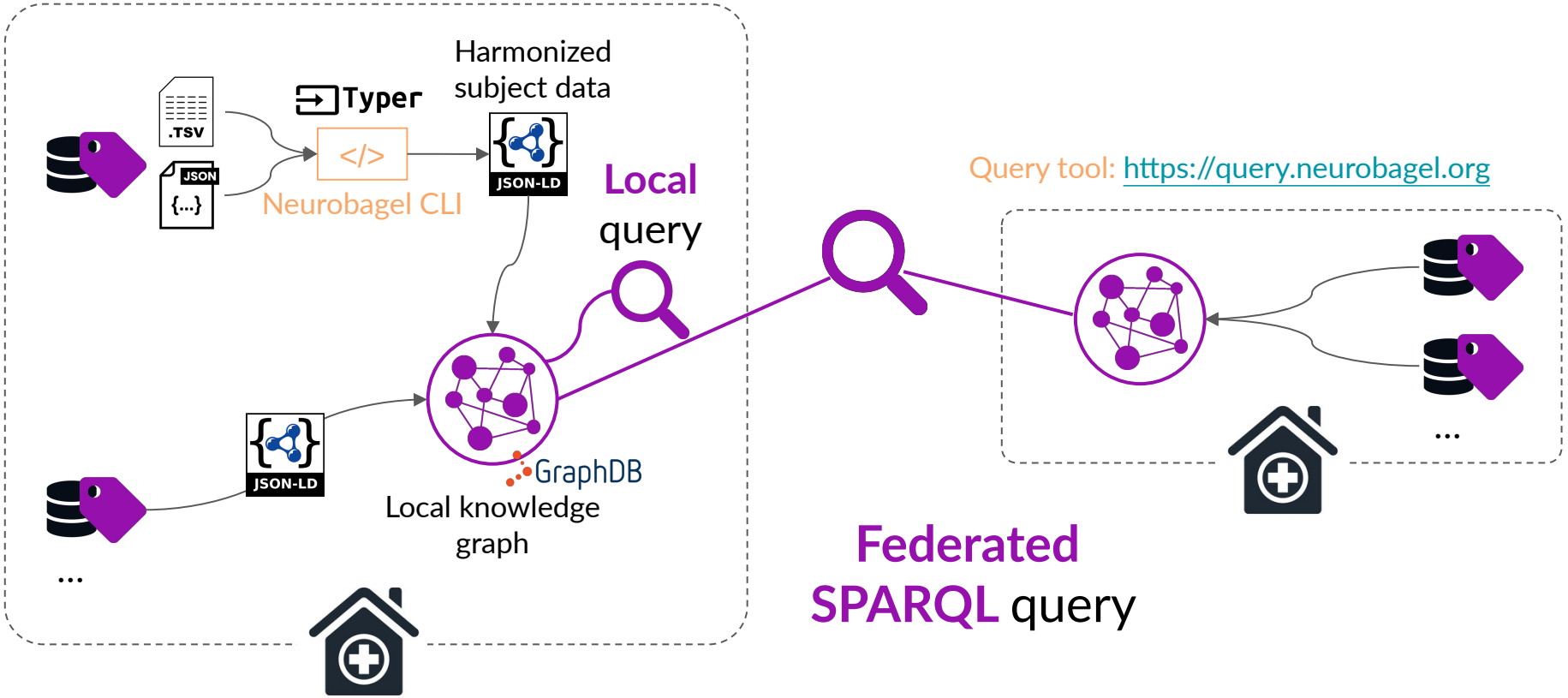
How can we:
search distributed data



- **create query federation**
- **intuitive search interface**

query.neurobagel.org/

Connect harmonized dataset representations via knowledge graph stores



Neurobagel Query beta

Define and find cohorts at the subject level

query.neurobagel.org

Query fields

Neurobagel Graph

OpenNeuro ×

Parkinson's Progression Markers Initiative ×

Quebec Parkinson Network ×

Phenotypic fields (at baseline)

Min Age

Max Age

Sex

All

Diagnosis

Healthy control

Parkinson's disease

Minimum number of sessions

Assessment tool

All

Imaging fields

Results

Select all datasets

PD De Novo: Resting State fMRI and Physiological Signals

14 subjects / 28 total subjects

PPMI

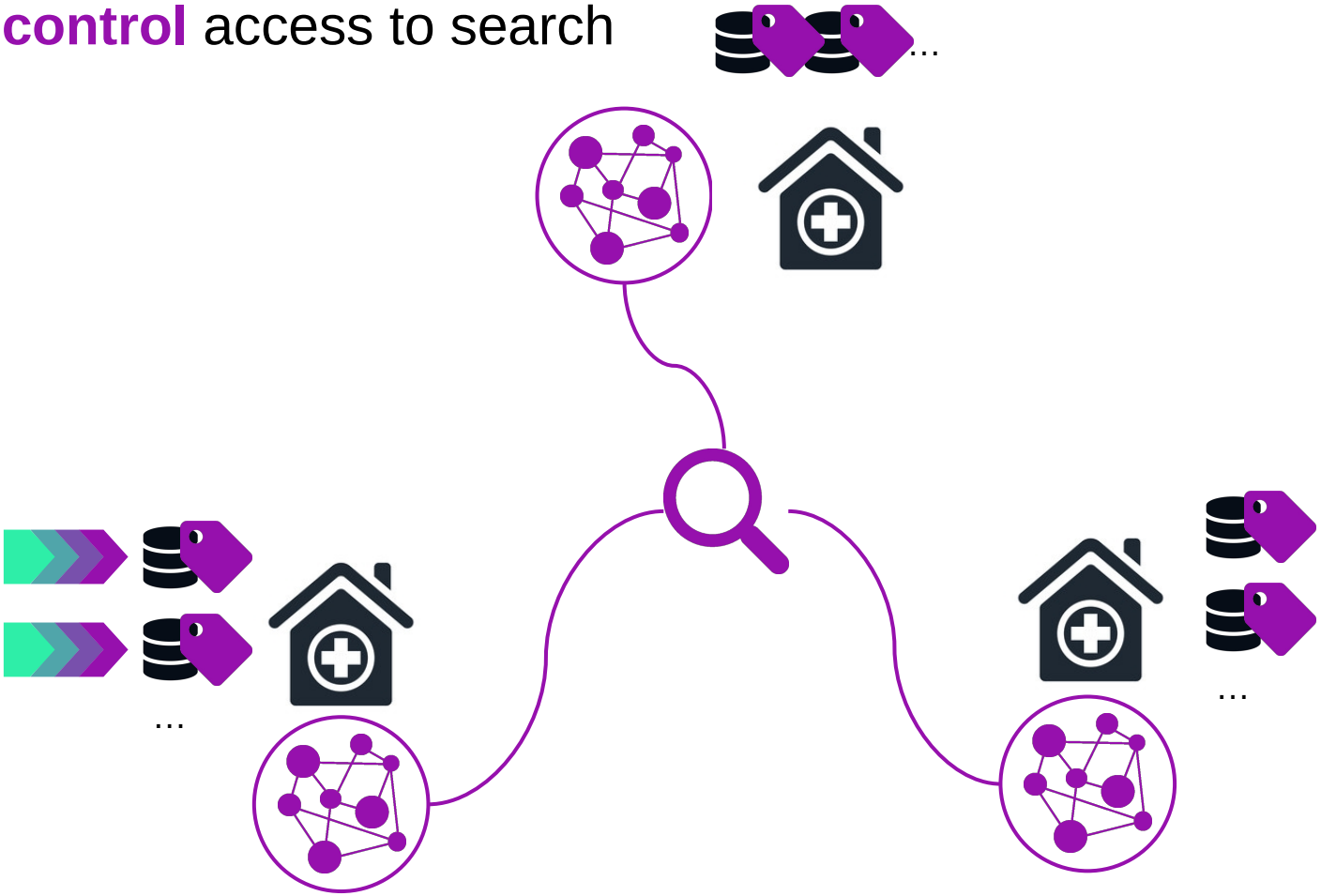
1491 subjects / 3060 total subjects

Quebec Parkinson Network

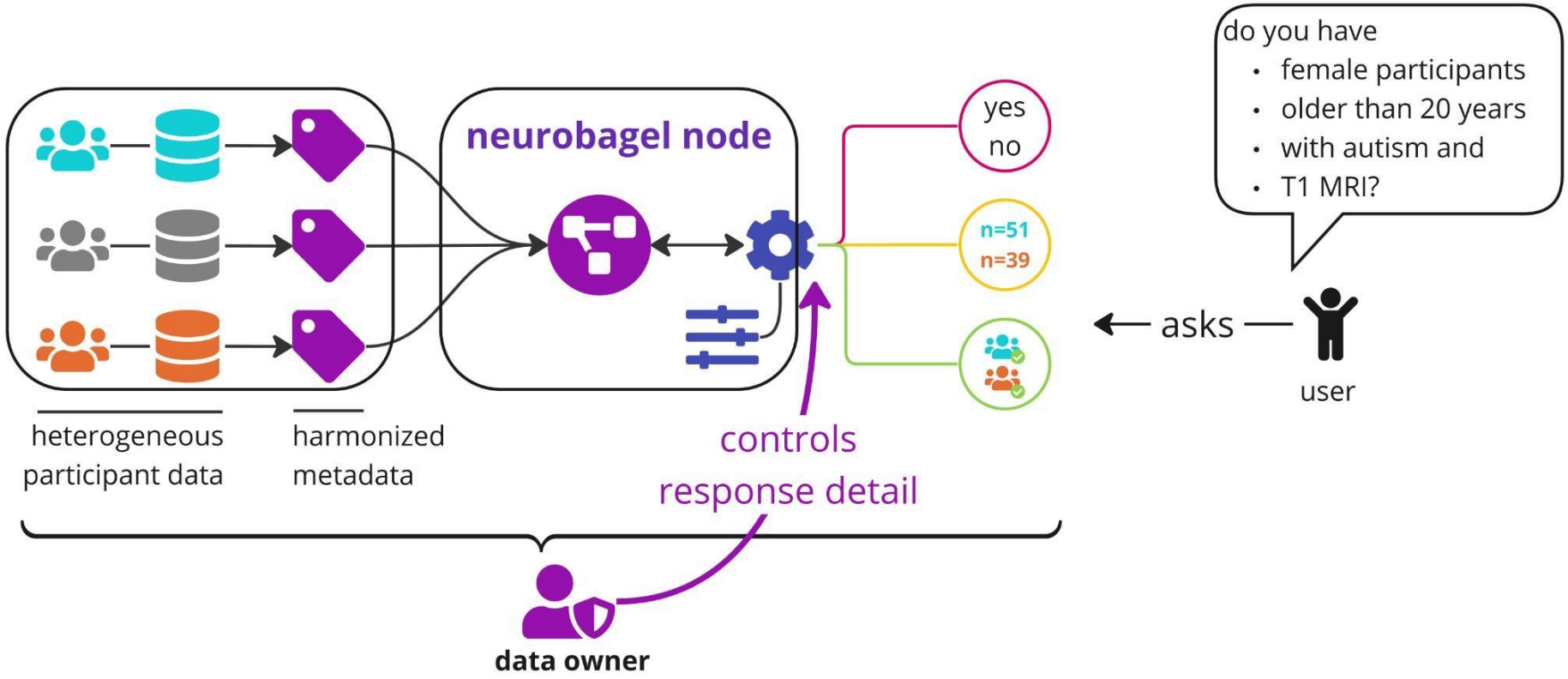
140 subjects / 265 total subjects

How can we:

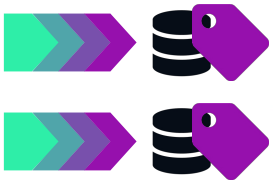
control access to search

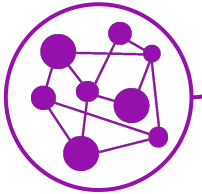


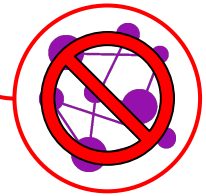
One tunable neurobagel node for each data owner

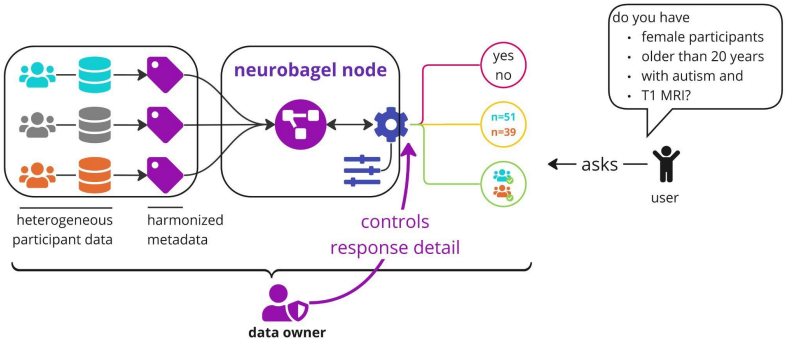


How can we:
control access to search



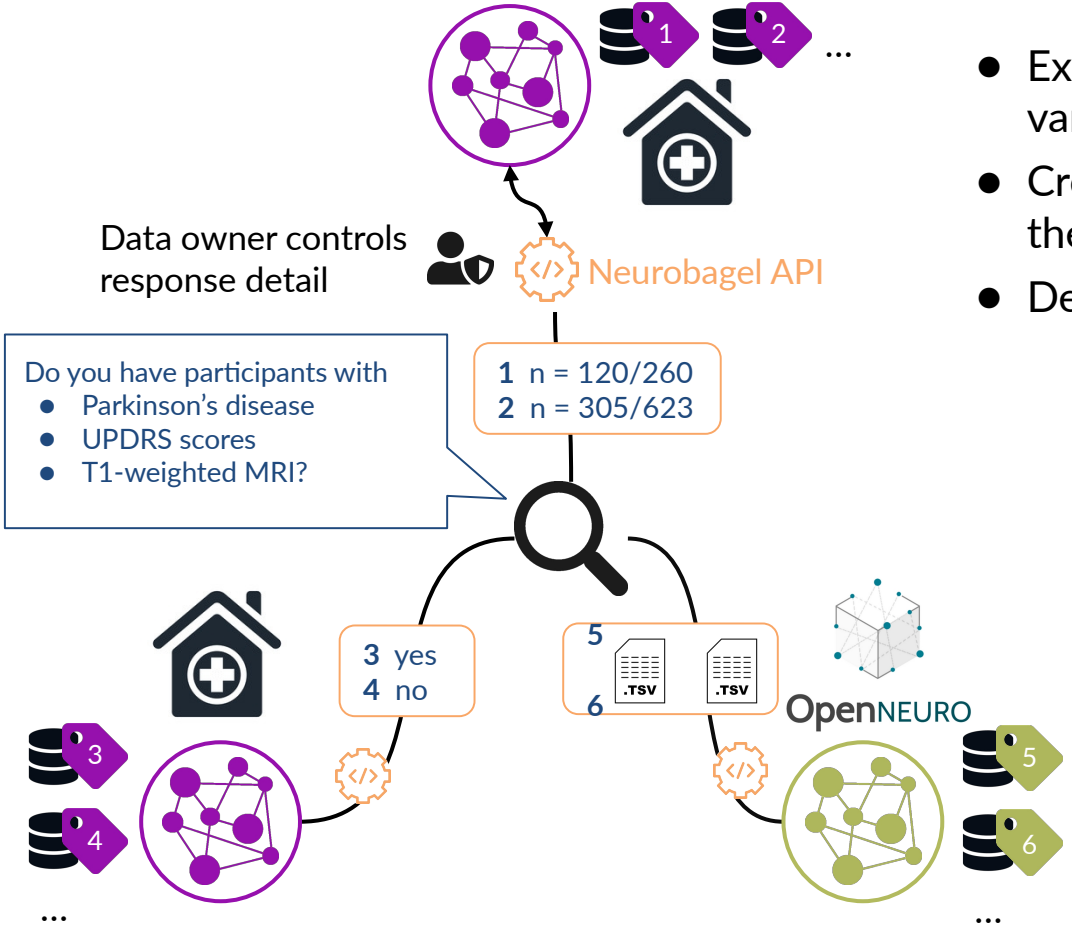






- **local control of visibility**
- **data curated by owner**

Federated dataset search, local data access/visibility control



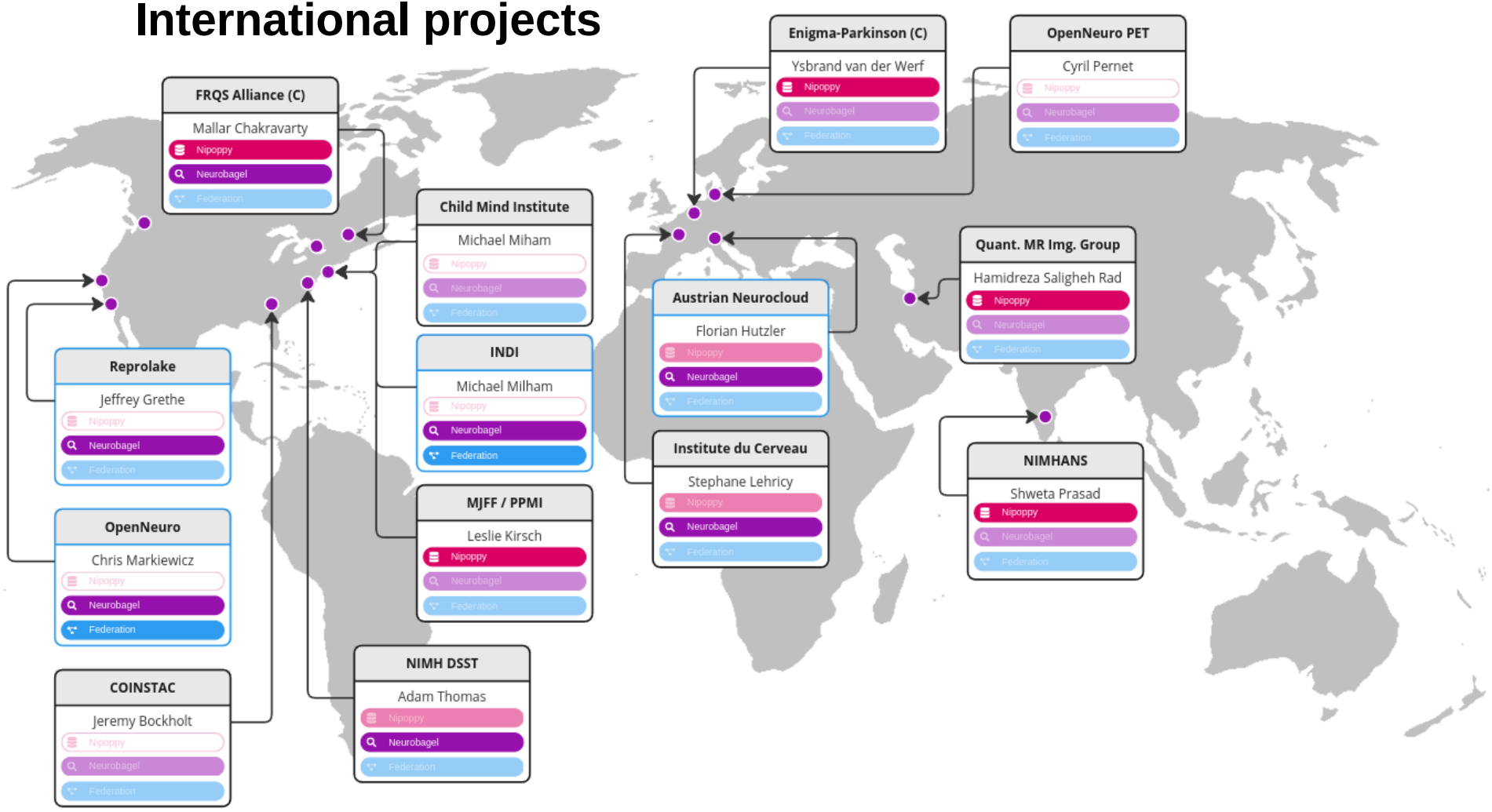
- Extendable model for harmonizing dataset variables
- Cross-dataset, cross-site data findability at the **subject level**
- Decentralized data storage & governance

Neurobagel is developed in active collaboration with:

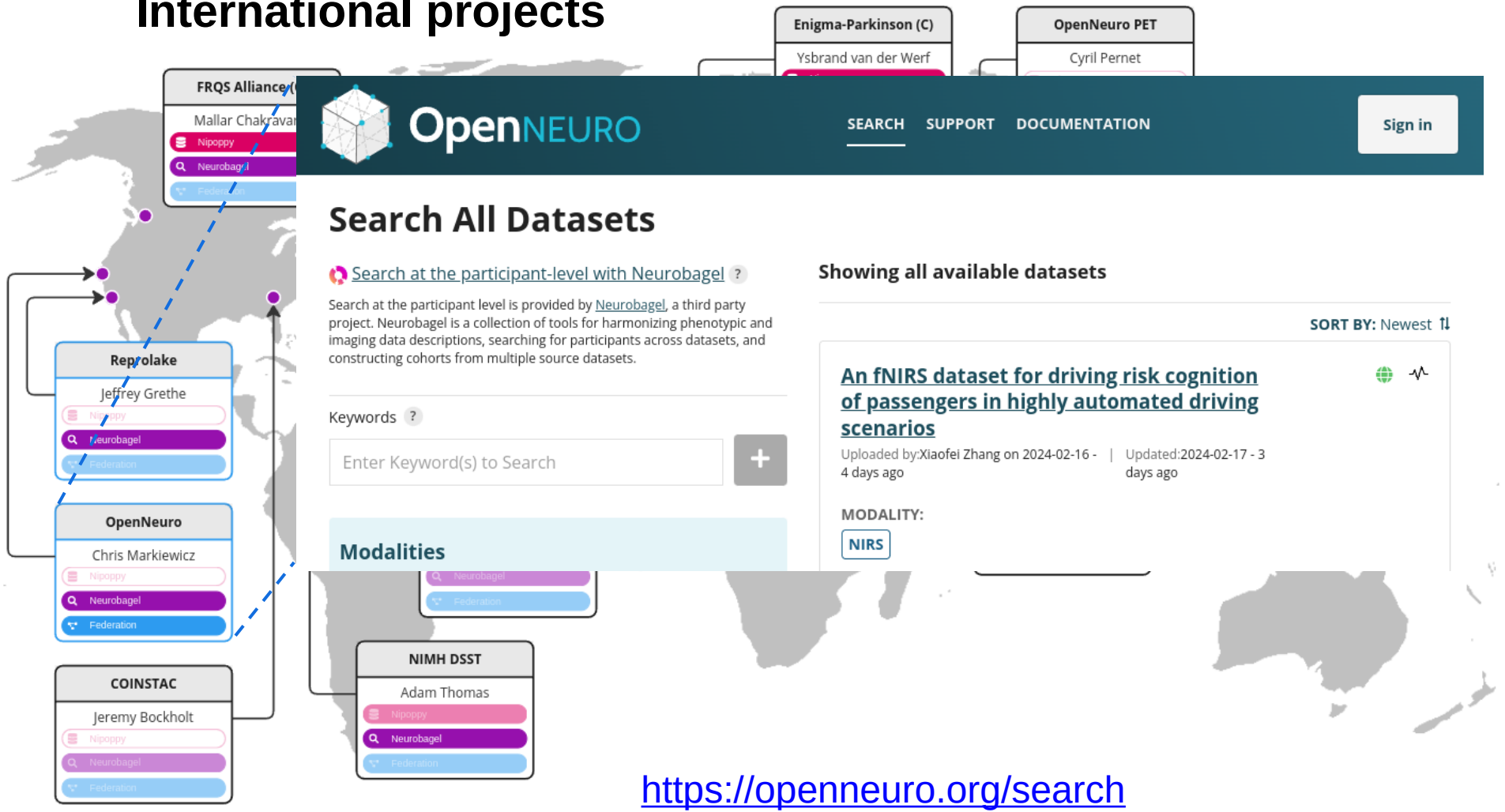
BIDS BRAIN IMAGING DATA STRUCTURE · ReproNim · DataLad · OpenNEURO · ONTARIO BRAIN INSTITUTE · Douglas CENTRE DE RECHERCHE RESEARCH CENTRE

For more info: https://neurobagel.org

International projects



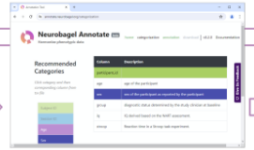
International projects



https://openneuro.org/search

How do I get my data into a **Neurobagel** node

Annotate



participants.tsv

Neurobagel supports imaging data in BIDS and text based tabular data. Our sister project Nipoppy provides tools for data curation.

A **domain expert** maps the local naming of tabular data to standard terms from controlled vocabularies using the graphical annotation tool.

Imaging data (BIDS)

Tabular data + annotations

participants.json

The mapping is stored as additional information in a BIDS data dictionary. This is the only additional file, all local data are kept unchanged as they are.

Harmonise

A **system administrator** uses the Neurobagel CLI to create a harmonised metadata view of the local data for the graph.

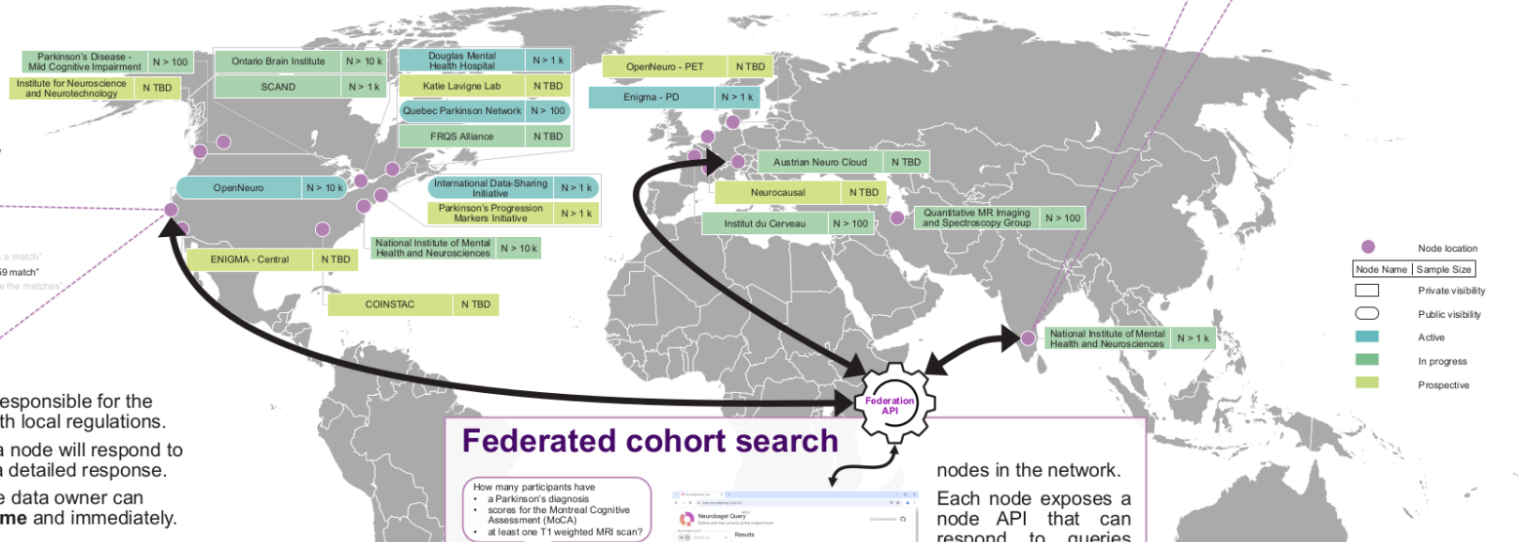
Deploy

Harmonised metadata

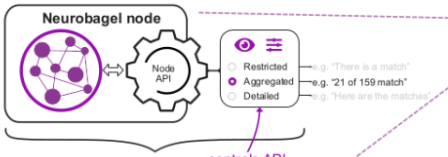
A **system administrator** deploys a local Neurobagel node to store metadata views of one or more local datasets for querying.

A **Neurobagel node** contains standard metadata about local datasets for querying. By default a node is private, i.e. only accessible to users within the institute.

Data described in a node remains local.



Local control of visibility



controls API response detail

Every node has a **data owner** who is responsible for the data in the node and ensures compliance with local regulations.

The data owner can **choose how detailed** a node will respond to an external query, from simple yes or no to a detailed response.

Because the node and data remain local, the data owner can **change and apply these settings at any time** and immediately.

Federated cohort search

How many participants have
- a Parkinson's diagnosis
- scores for the Montreal Cognitive Assessment (MoCA)
- at least one T1 weighted MRI scan?

nodes in the network.

Each node exposes a node API that can respond to queries

Ongoing work

- Authentication / Authorization

- Search for preprocessed data / derivatives

- Expand what phenotypic variables can be annotated / searched

Google
Summer of Code

- Ease of use, helper features

Computer assisted data annotation

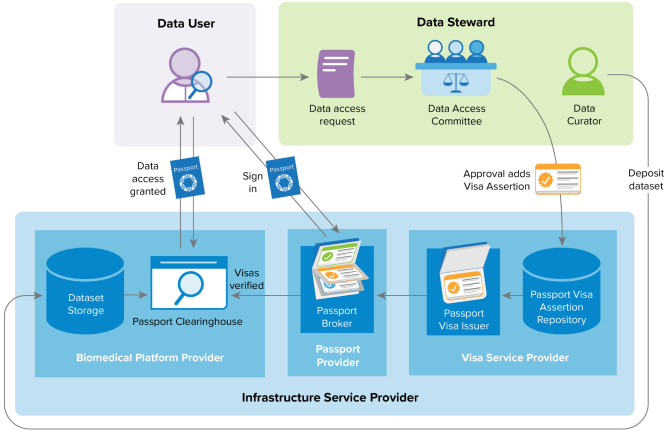
Google
Summer of Code

An LLM-assisted service for annotating research data with machine-understandable, semantic data dictionaries

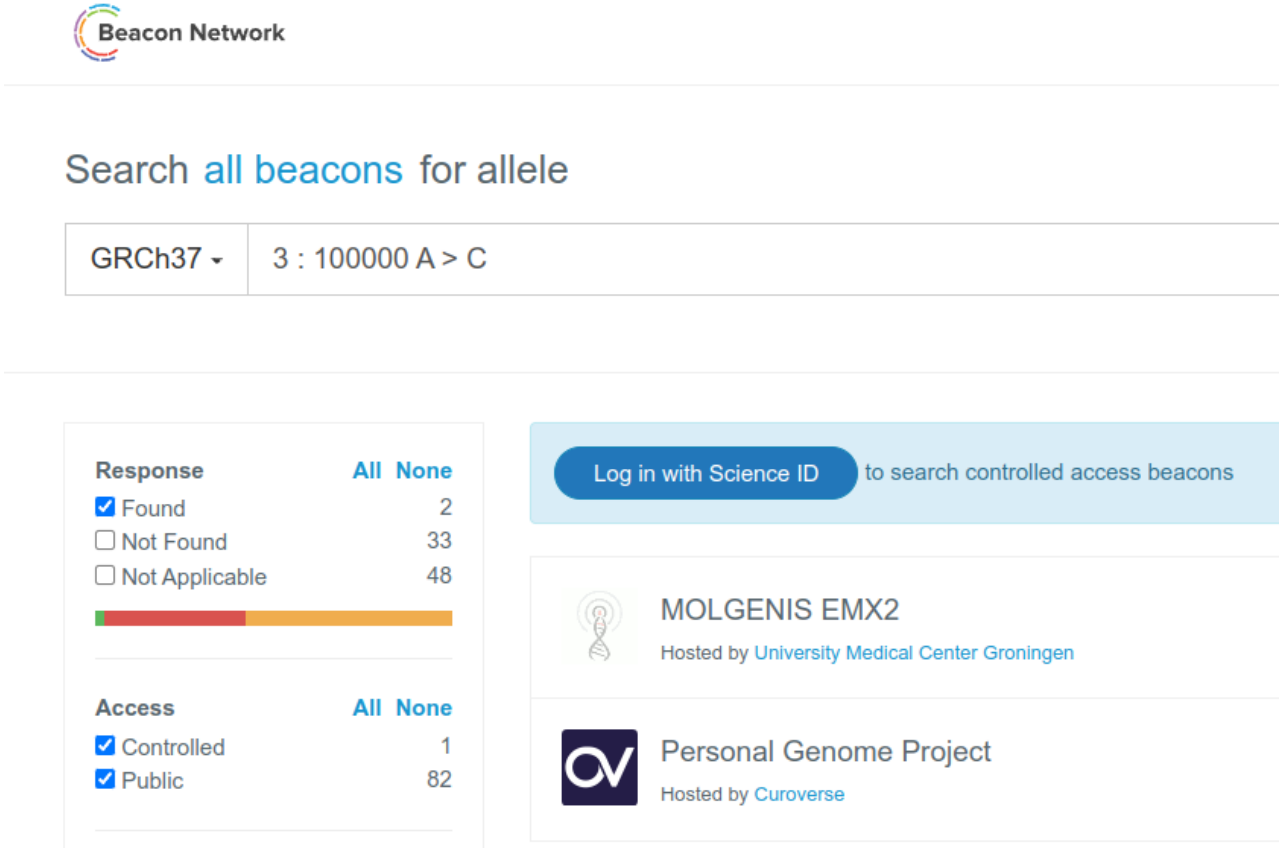
Neurobagel builds tools to provide a way for researchers and other data users to define and find cohorts of individuals across a federated ecosystem of data nodes. These tools are developed with the

Federated identity + local control

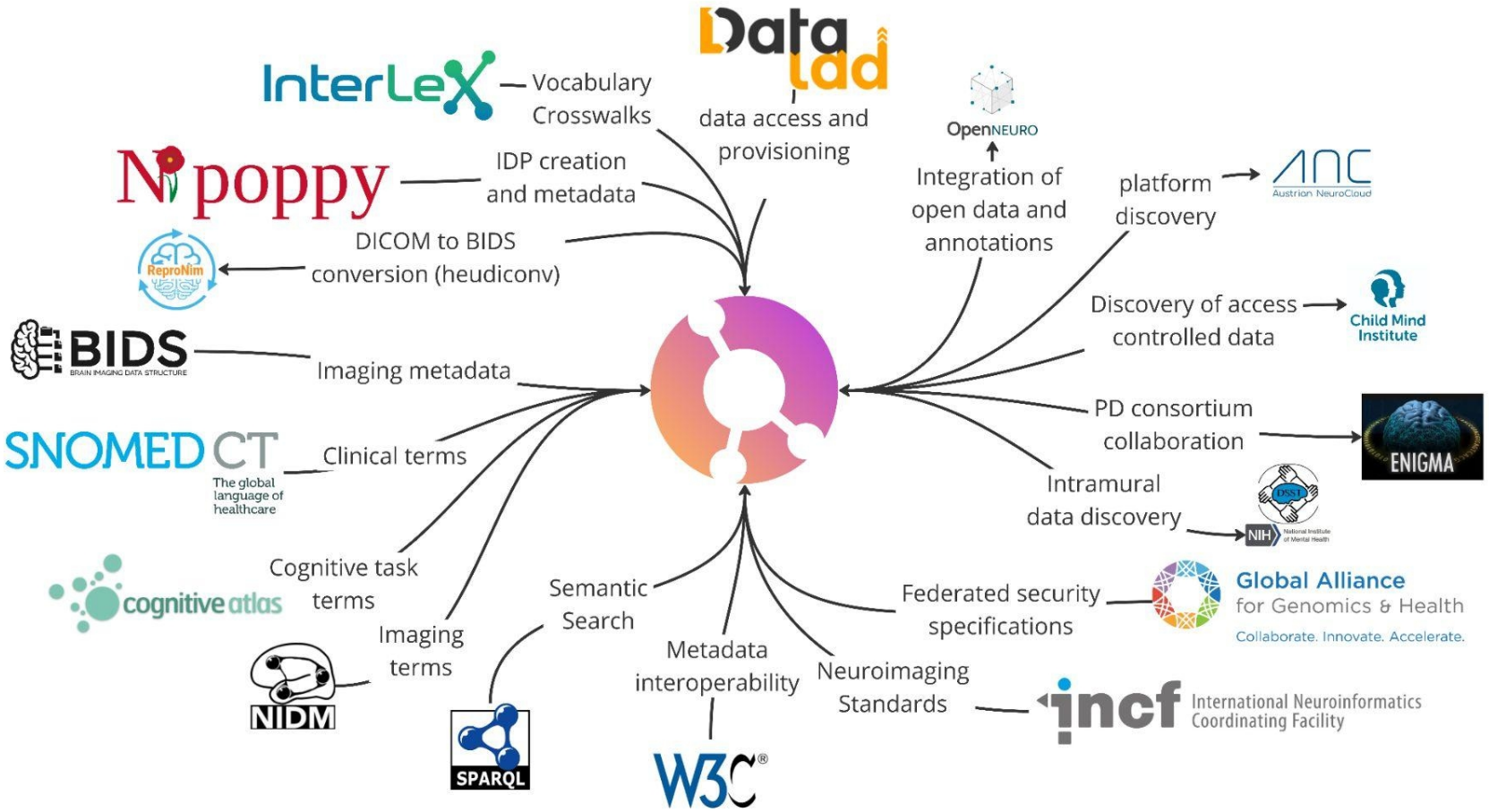
Global Alliance
for Genomics & Health



Interoperability with Beacon

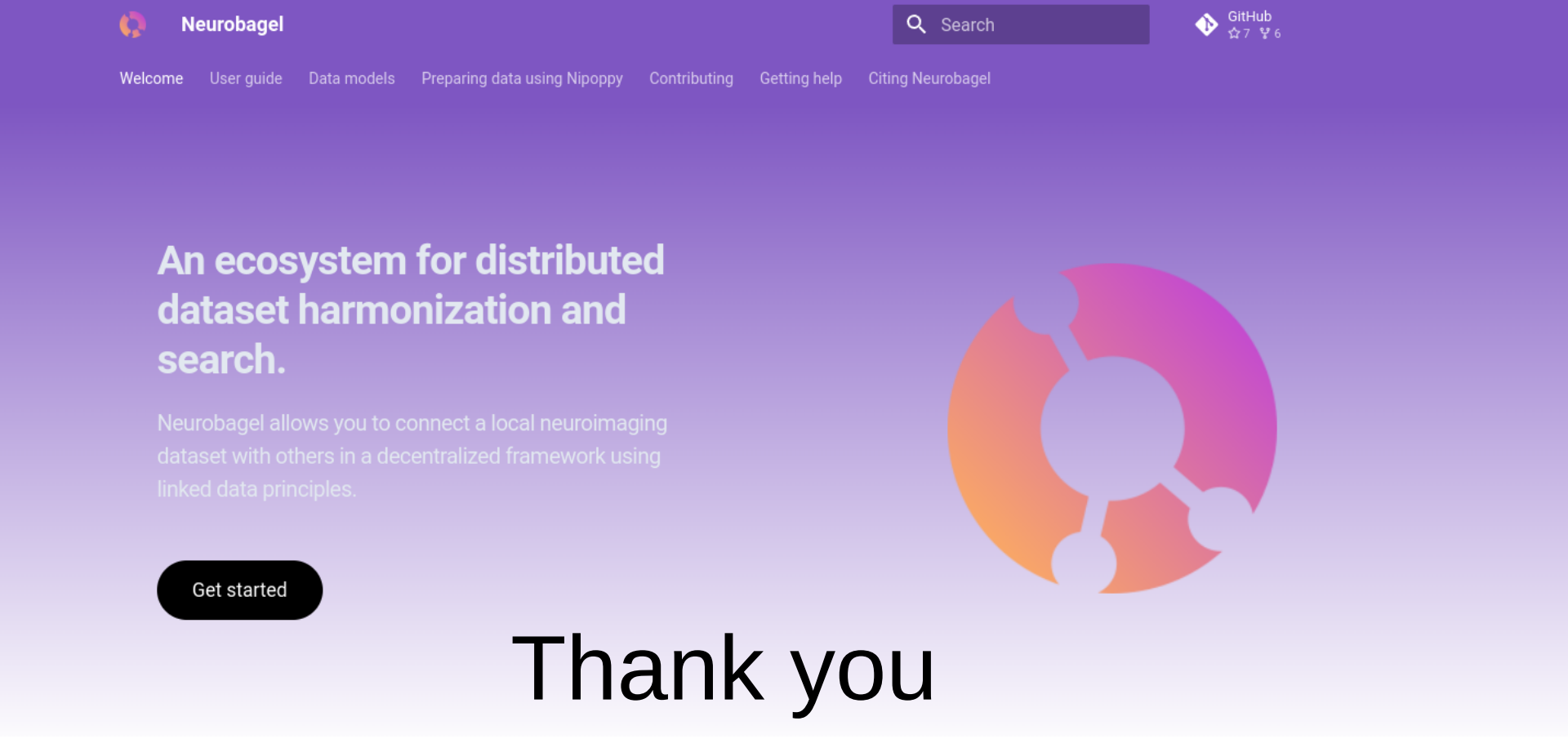


Neurobagel and the community



Take home messages

- Build standards and communities of practices
- Be sustainable: cheap, >3 lab bus factor, can scale, can update
- Distributed governance for distributed infrastructures / ecosystems
- **Love, Fame, Money, and Rules**



Thank you

https://neurobagel.org