Data standardization
A (very) short introduction

Jean-Baptiste Poline

MNI, Brain Imaging Centre, McGill, Montreal

Credit : S. Urchs

Part I: Motivations - if necessary !

Part II: Standardizing : meaning

Part III: How do we do it ?

Part IV: Conclusion

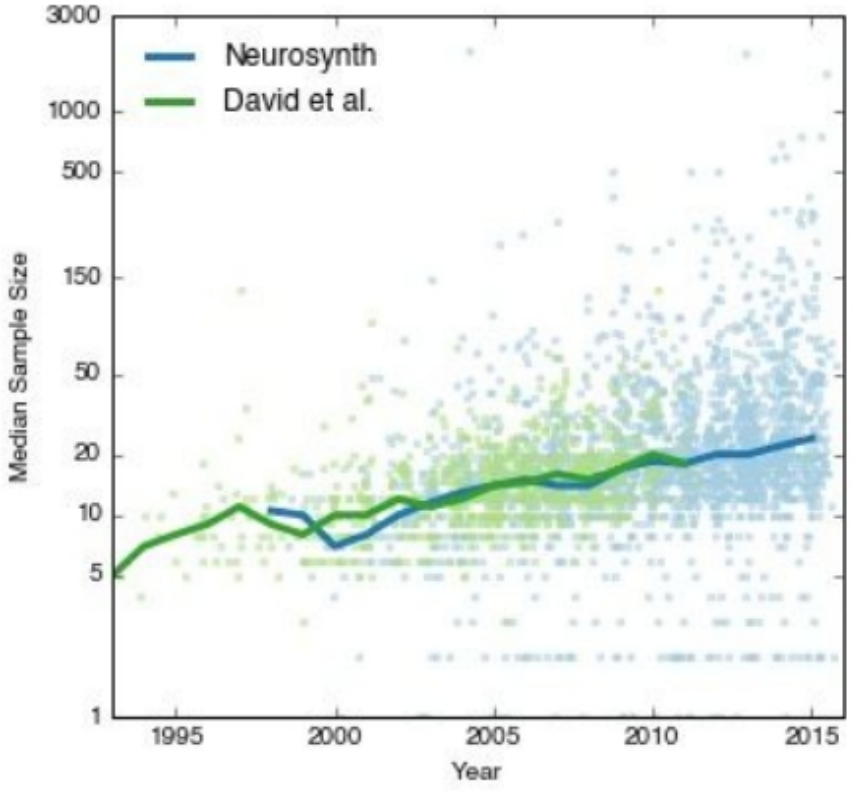
Part I: Motivations - if necessary !

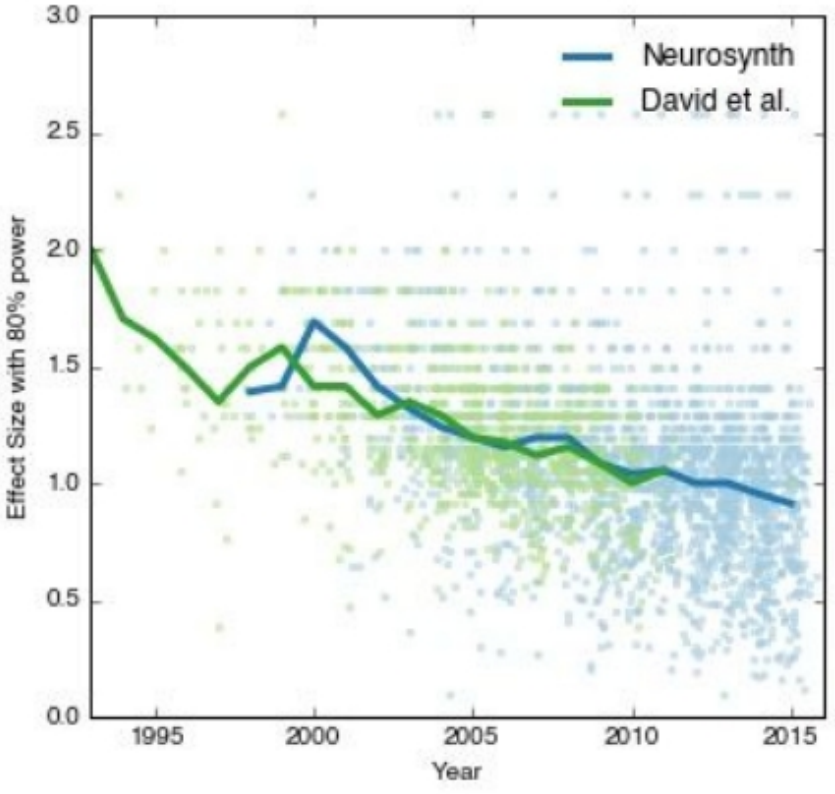
Part II: What does harmonization mean ?

Part III: How do we do it ?

Part IV: Conclusion

Feeling the Future





Poldrack et al., PNAS, 2016

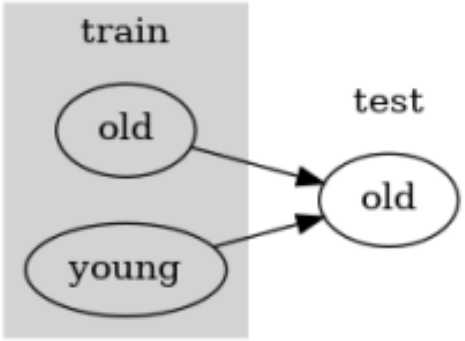
Feeling the Future

Paradigm	Intersection mask	mask size (vox)	Cohen D			BOLD		
			P10	median	P90	P10	median	P90
MOTOR	Bilateral Precentral Gyrus	12894	0.158	0.628	1.070	0.505	2.707	8.582
	Bilateral Supplementary motor cortex	3418	0.211	0.716	1.197	0.911	4.033	12.510
	Left putamen	1532	0.114	0.513	0.864	0.586	2.388	4.318
	Right putamen	1437	-0.008	0.369	0.749	-0.045	1.696	3.609
WM	Bilateral Middle frontal gyrus	7116	0.101	0.474	0.837	0.130	0.986	2.504
EMOTION	Left amygdala	1133	0.265	0.534	1.065	0.516	1.198	3.379
	Right amygdala	1082	0.308	0.645	1.140	0.581	1.350	3.557
GAMBLING	Left accumbens	455	0.138	0.310	0.461	0.369	0.849	1.440
	Right accumbens	417	0.141	0.332	0.488	0.373	0.981	1.618

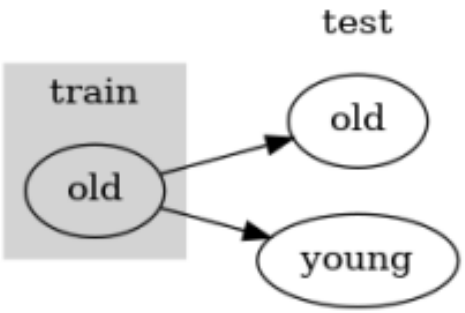
With effect size = 0.5 => Power ~ 30%

Machine learning needs data...

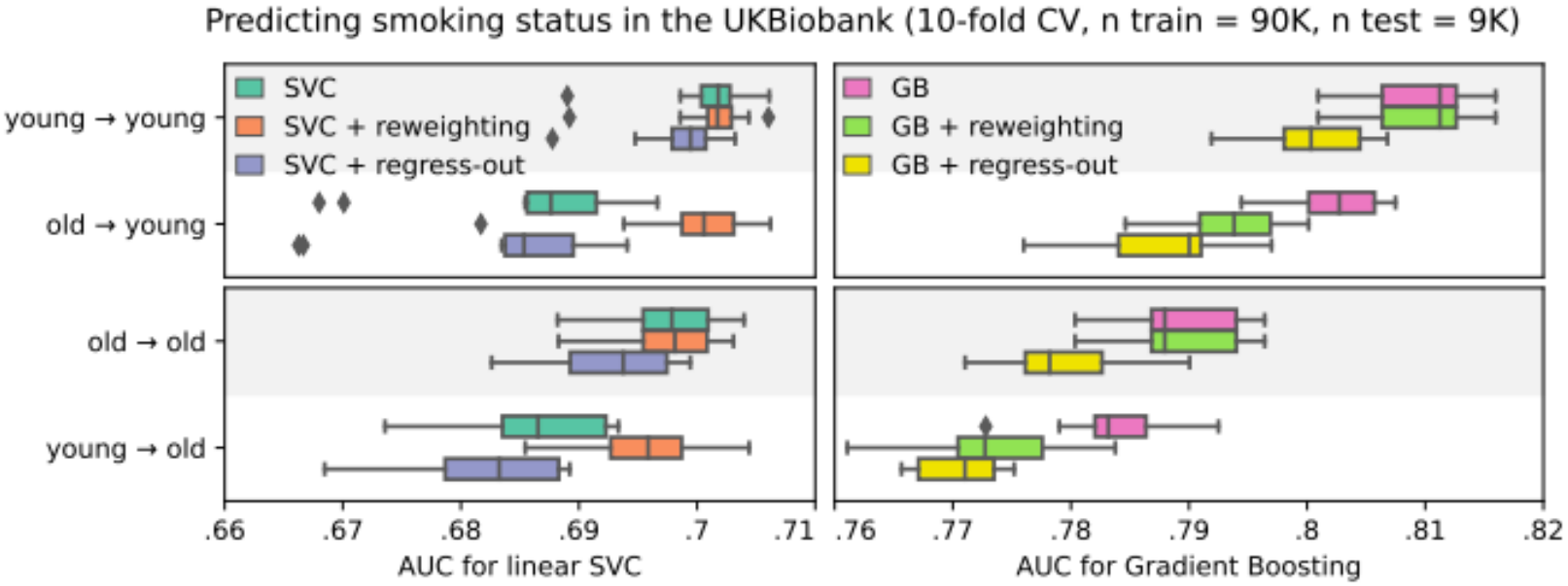
- What do we lose by training on the wrong population?



- How does a model perform when deployed on a new population?



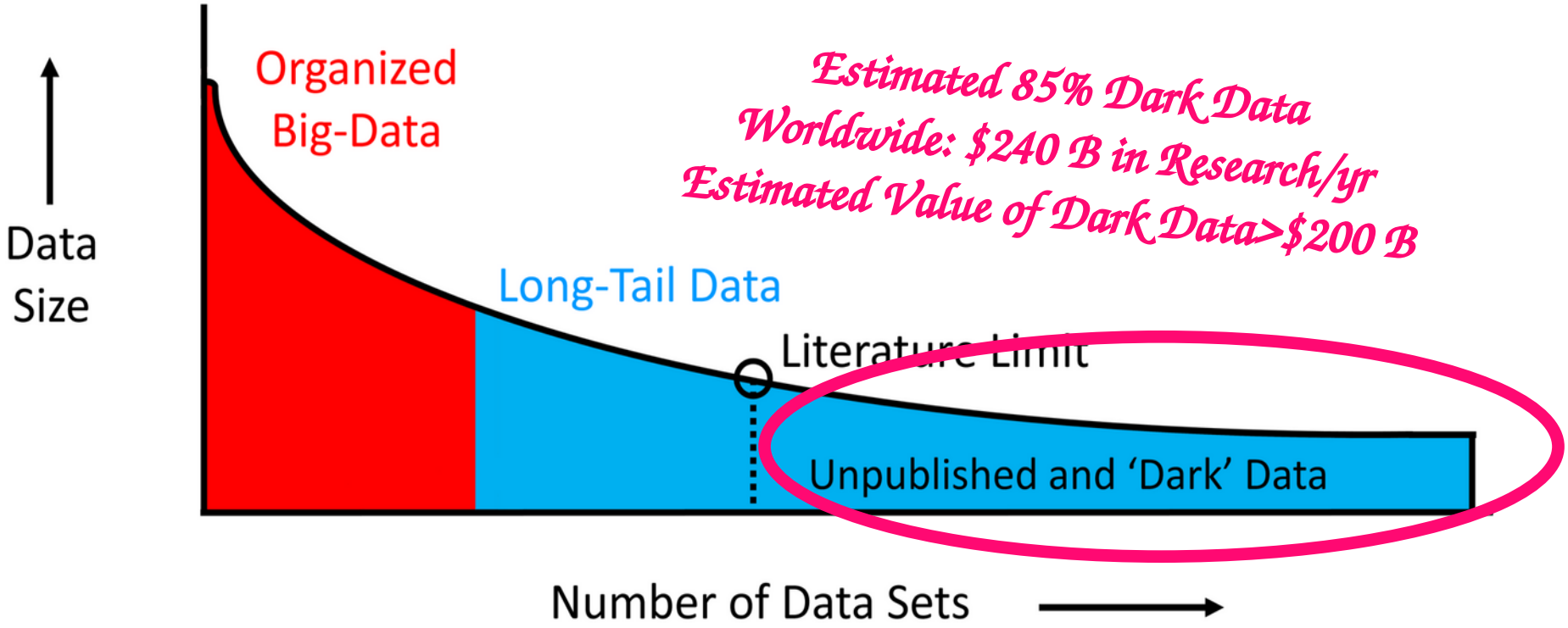
Dataset shift analysis in the UKB



- Dataset shifts threaten generalization
- Simpler models are not always more robust
- "Regressing out nuisance variables" does not help

Jerome Dockes et al., in prep.

Towards "datafication" of neuroscience



McLeod et al., Lancet, 2014
Ferguson et al., 2014, *Nature Neuroscience*

Cause: data issues

- A less rare case than usually thought !
- No license
- Database not containing what it describes
- Wrong QC – QC unreliable
- Headers of files are not correct (cf the Left/Right issue)
- Provenance of data is lost
- SAM1 SAM2 SAM3:
  https://www.youtube.com/watch?v=N2zK3sAtr-4

Issues with data: horror stories

From HCP:

"With the releases of FreeSurfer 7.X, there have been some regressions in surface placement performance when running FreeSurfer inside the HCP Pipelines. At this time, *I would recommend sticking with FreeSurfer 6.0* while we get these issues sorted out."

Part I: Motivations - if necessary !

Part II: What does standardisation mean - and provides ?

Part III: How do we do it ?

Part IV: Conclusion

Layers of data standardisation

- **Formats and units**
  - Information is exactly the same
  - Nifti / dicoms
  - Age in years or months, Farenheit or Celcius
- **Easily “mappable” measures, versions**
  - I measure motor control with UDPRS III in my population but I adapted the test to my clinical population
- **Related meaning**
  - I measure the concept of anxiety with TestA or TestB

Standardization benefits

individual researcher

open dataset consortium

clinical research institute / clinical consortium
Douglas Research Center / Enigma consortium

“**I need to expand my training sample to be more diverse (and larger)**”

“Where do I find data of PD patients that have completed the MOCA?”

Standardization benefits

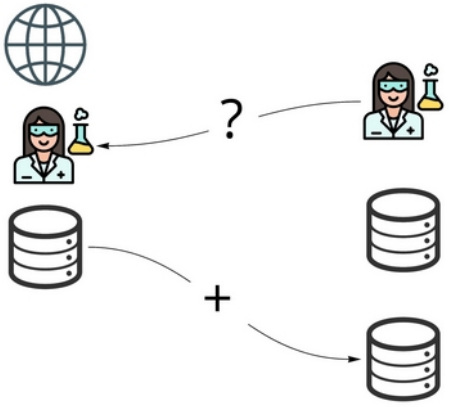
individual researcher

open dataset consortium

clinical research institute / clinical consortium
Douglas Research Center / Enigma consortium

"I need to expand my training sample to be more diverse (and larger)"

"**Where do I find data of PD patients that have completed the MOCA**?"



Standardization benefits

individual researcher

"I need to expand my training sample to be more diverse (and larger)"

open dataset consortium

"Where do I find data of PD patients that have completed the MOCA?"

clinical research institute / clinical consortium
Douglas Research Center / Enigma consortium

"**What data do our PIs have?**"

And

"**How can we advertise our data without sharing them**?"

Successful harmonizations

schema.org

- What if we had google power on datasets ?
- The briefest introduction ever to schema.org

Link data : 0.101

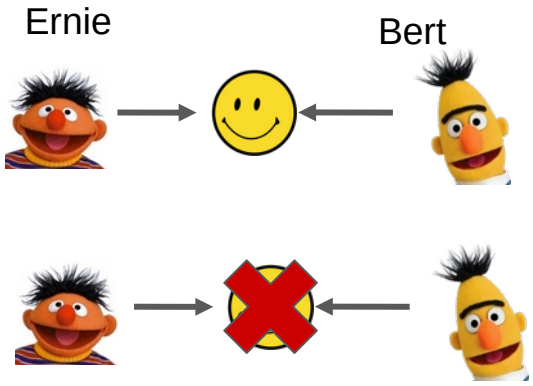
Documents

World Wide Web (WWW)

- Physical location is abstracted away
- **You just care about documents**
- **Hyperlinks**

Tim Berners-Lee, CERN
March 1989, May 1990

Services



"List all the friends of Ernie!"

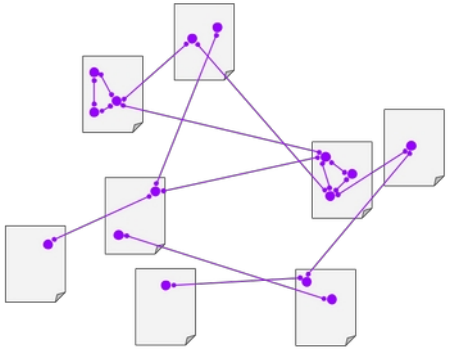
facebook

myspace.com
a place for friends

Create **information**

Meanings

"Link facts, not documents"



"Make the meaning of facts machine readable"

facebook

- Ernie is a friend of Bert

myspace.com
a place for friends

Berners-Lee, T., Hendler, J. **Publishing on the semantic web**. Nature 410, 1023–1024 (2001).

Link data : 0.101

RDF triples form (linked) graphs

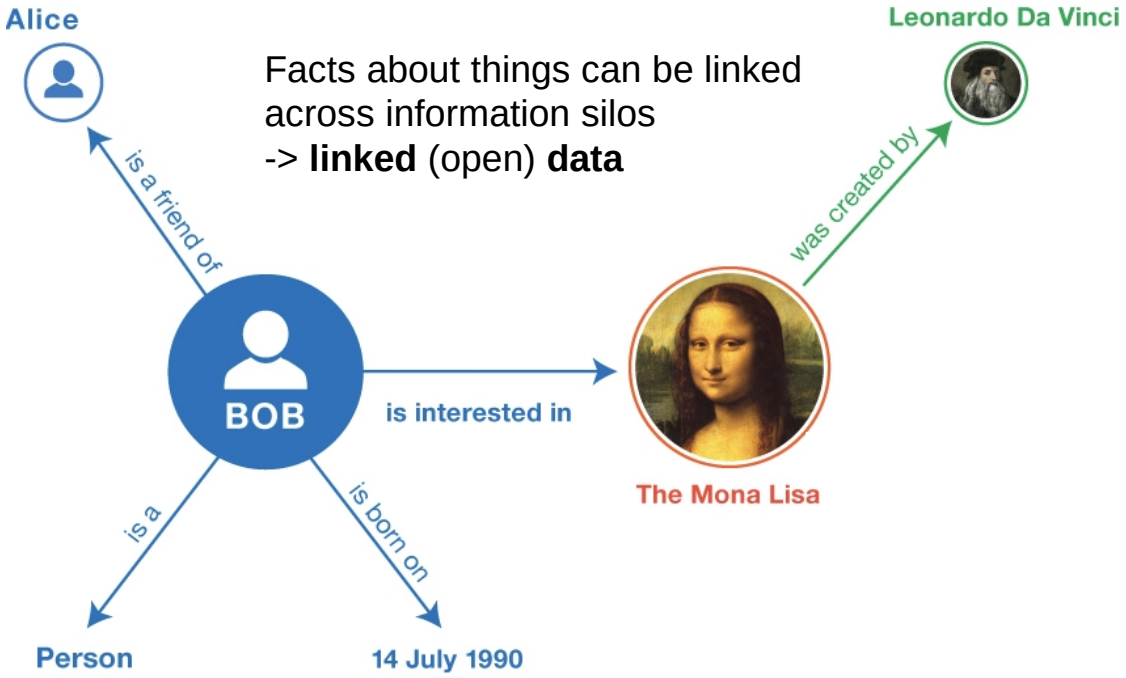
<Bob> <is a> <person>.

<Bob> <is a friend of> <Alice>.

<Bob> <is born on> <the 4th of July 1990>.

<Bob> <is interested in> <**the Mona Lisa**>.

<**the Mona Lisa**> <was created by> <Leonardo da Vinci>.



vocabularies? -> many!

Theory

Practice

https://www.w3.org/TR/rdf11-primer/

Clinical trials too small

OMOP origin stories: the 2004 VIOXX scandal

MERCK

FDA conducts regulatory action to withdraw approved painkiller because of **increased stroke risk**

Merck was alleged to have **known and suppressed** this information!

-> lawsuits
-> congress hearings

UNITED STATES CONGRESS

nature

Explore content ~ About the journal ~ Publish with us ~

nature > news > article

Published: 13 November 2007

Merck settles Vioxx lawsuits for $4.85 billion

Meredith Wadman

Nature (2007) | Cite this article

3893 Accesses | 3 Citations | 129 Altmetric | Metrics

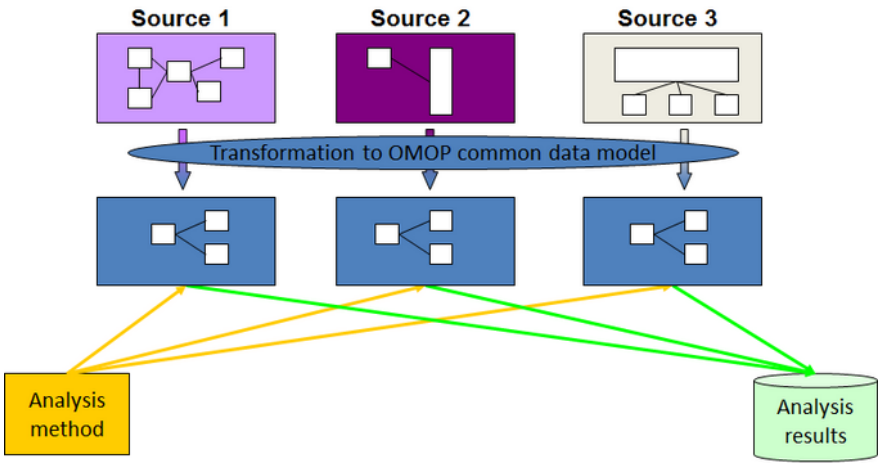
But drug firm maintains it was not at fault over arthritis drug.

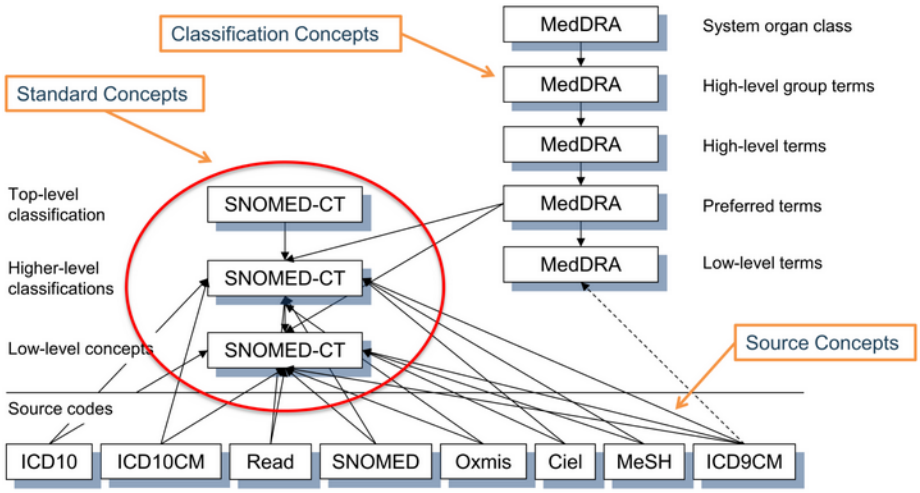
Three years after it pulled its blockbuster painkiller Vioxx from the market, Merck has agreed to pay $4.85 billion to settle nearly 27,000 lawsuits that claim the arthritis drug caused heart attacks and strokes.

Clinical trials too small + No system to integrate observational data

No system to integrate observational data

OMAP





* Transformation to **a common data model**

* generally only one vocabulary per domain (**SNOMED** for diagnosis, **RxNorm** for drugs …)

PARTICIPANTS

126 661 070 people observed for at least 365 days before 1 January 2017, 2018, or 2019 from 13 databases.

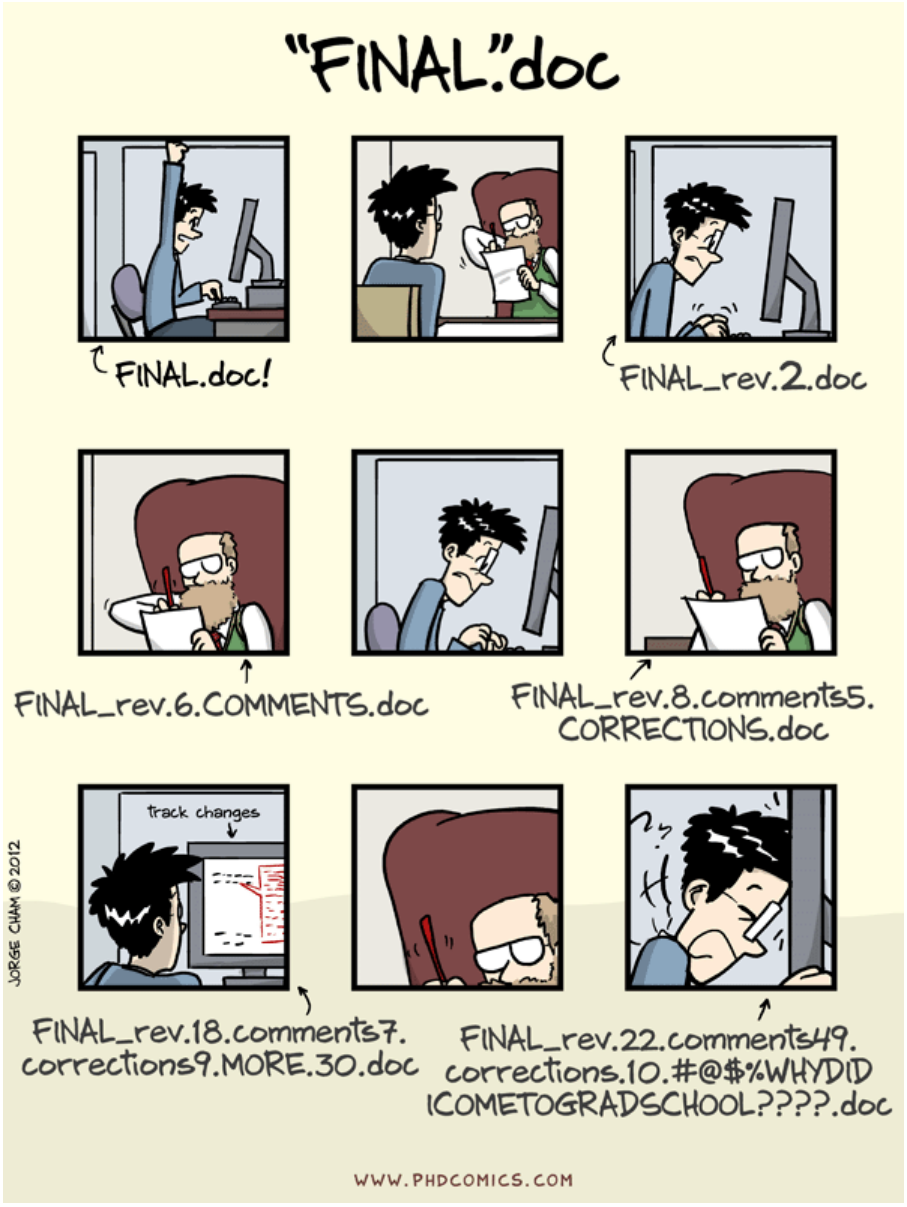
Part I: Motivations - if necessary !

Part II: What does harmonization mean - and provides ?

Part III: How do we do it ?

Part IV: Conclusion

Use the simple tools first



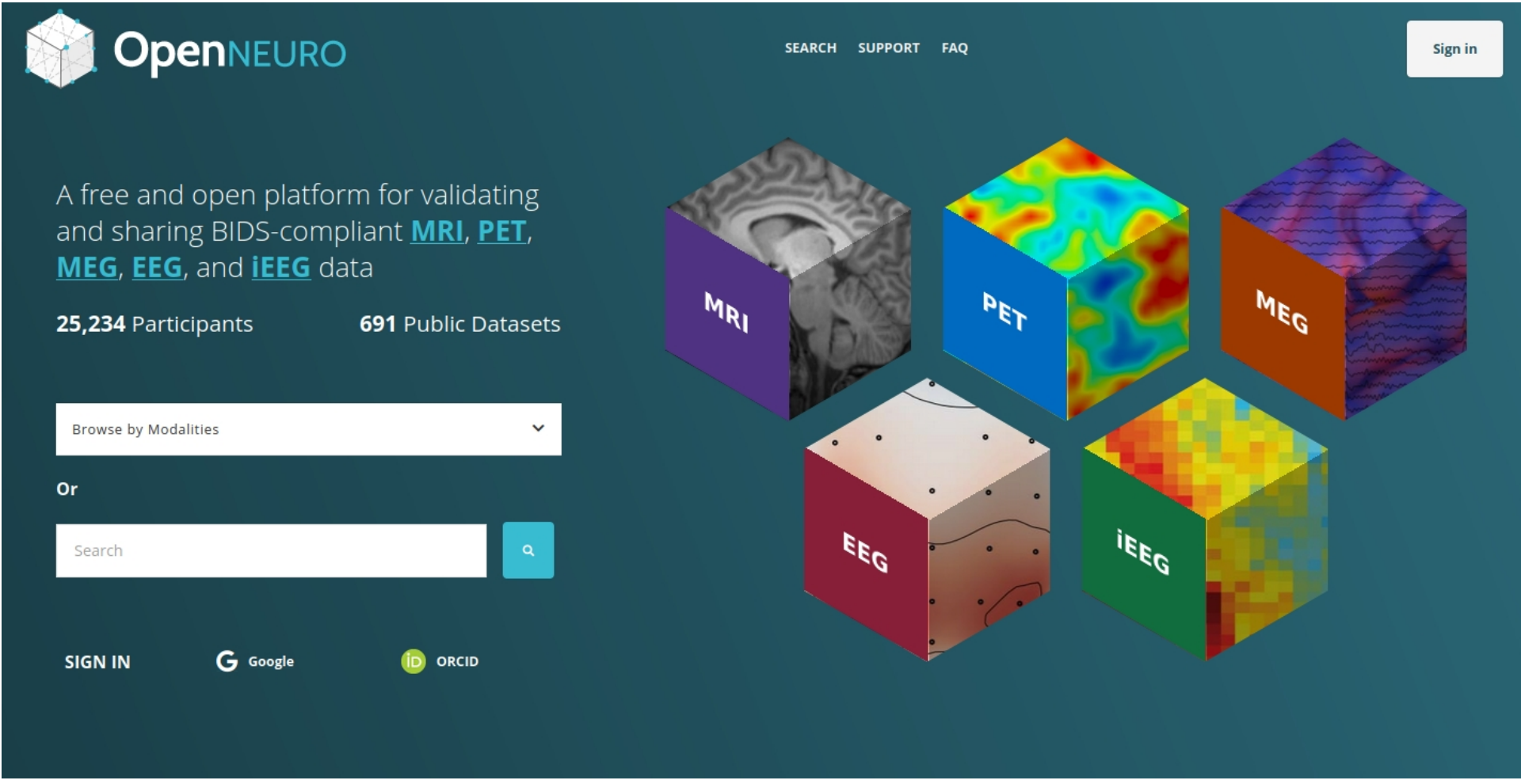
- File name <u>matters</u>
- Format **matters**
  - Excel or CSV ? Why?
  - Docx or markdown?
  - Propriatory ?
  - License ?
  - Copyright ? (quarto)
  - Supported ?
  - ...
- **Where** we store **<u>matters</u>**
  - Platform governance
  - Security / …
  - No free lunch: When will I pay ?
  - ...

Neuroimaging

- BIDS
- NIDM terms
- BIDS->NIDM
- Show annotation and search tools

- Why should I put my data in BIDS ?
    - And how much will I suffer in the process ?

BIDS Standard : OpenNeuro



BIDS is the “worst standard to the exclusion of all others”

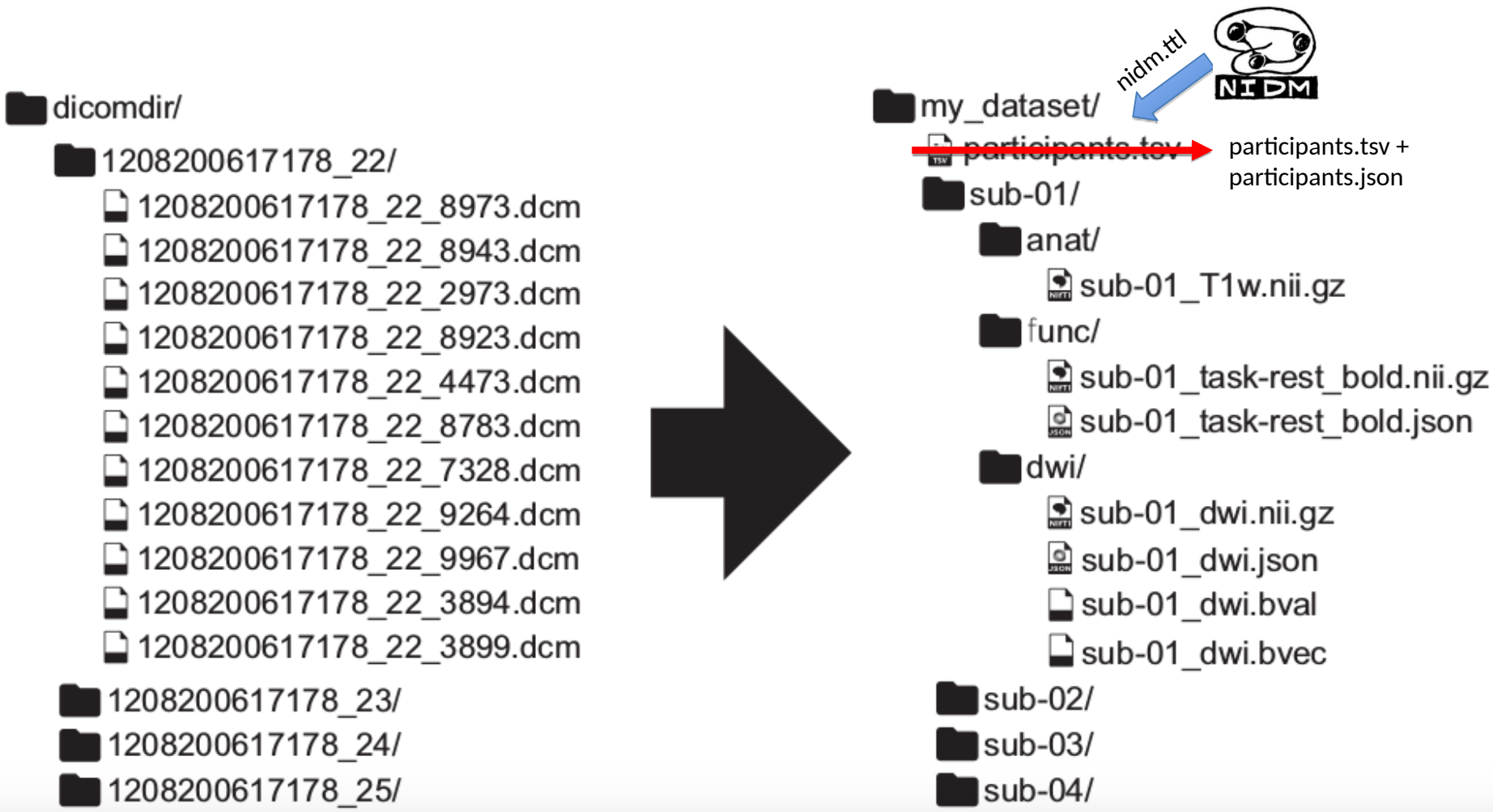
BIDS limitations

- *Participant.tsv* variables
  - Variables **column names** are not standardized
    - a much wider space than imaging !
  - Variable **values** (i.e. data element, cells) are not standardized

- *Sidecar json* files
  - Terms are not harmonized
  - Content not standardized

- Consequence:
  - You cannot easily recreate a dataset from several datasets / search across participants

BIDS to RDF (NIDM)



Example of sidecar json

```
{
  "MeasurementToolMetadata": {
    "Description": "Adult ADHD Clinical Diagnostic Scale V1.2",
    "TermURL": "http://www.cognitiveatlas.org/task/id/trm_5586ff878155d"
  },
  "adhd_b": {
    "Description": "B. CHILDHOOD ONSET OF ADHD (PRIOR TO AGE 7)",
    "Levels": {
      "1": "YES",
      "2": "NO"
    }
  },
  "adhd_c_dx": {
    "Description": "As child met A, B, C, D, E and F diagnostic criteria",
    "Levels": {
      "1": "YES",
      "2": "NO"
    }
  }
}
```

https://bids-specification.readthedocs.io/en/latest/03-modality-agnostic-files.html

Participants.tsv file:

ADHD200 - Brown

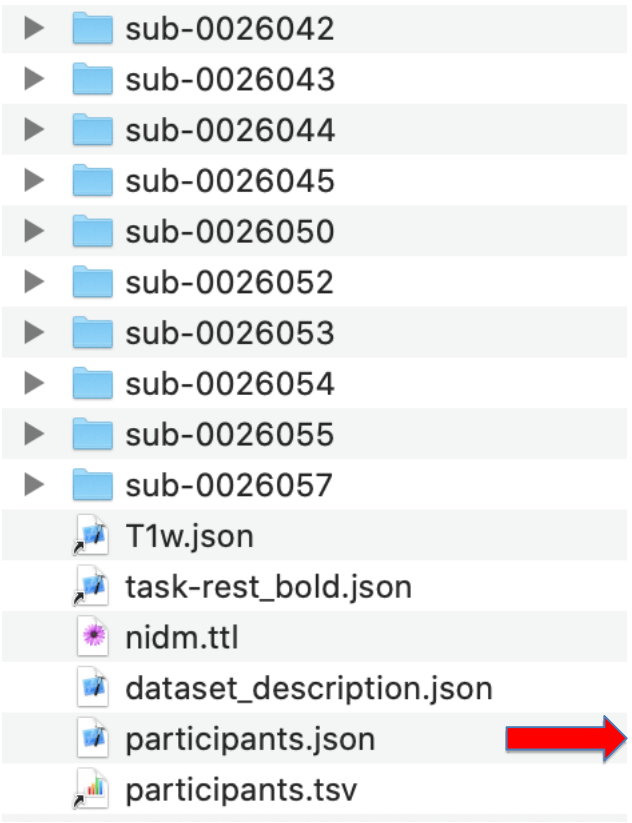
- sub-0026001
- sub-0026002
- sub-0026004
- sub-0026005

Participants.tsv file:

participant_id	gender	age	handedness	verbal_iq	performance_iq	full4_iq	qc_rest_1	qc_anatomical_1
26001	Male	16.92	Right	133	104	120	Pass	Pass
26002	Male	15.68	Right	106	106	107	Pass	Pass
26004	Female	14.99	Right	119	123	125	Pass	Pass
26005	Female	15.16	Right	116	131	126	Pass	Pass
26009	Male	16.91	Left	113	81	97	Pass	Pass
26014	Female	16.21	Right	101	102	102	Pass	Pass
26015	Female	15.2	Right	127	98	113	Pass	Pass
26016	Male	16.07	Right	120	96	109	Pass	Pass
26017	Female	14.56	Right	95	87	89	Pass	Pass
26022	Male	17.83	Right	105	111	109	Pass	Pass
26024	Female	17.77	Right	89	83	85	Pass	Pass
26027	Female	11.28	Right	108	103	106	Pass	Pass
26030	Female	14.51	Right	121	119	123	Pass	Pass
26039	Female	14.19	Right	125	117	124	Pass	Pass
26040	Female	13.67	Right	111	93	102	Pass	Pass
26041	Female	13.68	Right	129	120	128	Pass	Questionable
26042	Female	13.82	Right	106	110	109	Pass	Pass

Create nidm.ttl / data dictionary

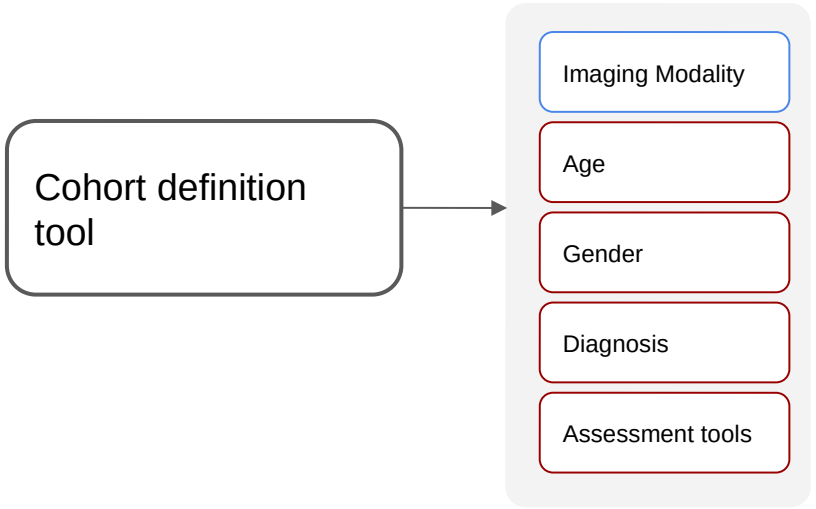
Brown Site – ADHD200




```
"handedness": {
    "minValue": "NA",
    "maxValue": "NA",
    "hasUnit": "NA",
    "label": "handedness",
    "description": "Simple handedness determined using the Edinburgh Handedness Inventory",
    "source_variable": "handedness",
    "valueType": "http://www.w3.org/2001/XMLSchema#complexType",
    "associatedWith": "NIDM",
    "levels": {
        "Left": "0",
        "Right": "1",
        "Ambidextrous": "3"
    },
    "isAbout": {
        "url": "http://uri.interlex.org/base/ilx_0104886",
        "label": "Handedness assessment"
    }
},
"verbal_iq": {
    "hasUnit": "",
    "minValue": "0",
    "maxValue": "200",
    "label": "verbal_iq",
    "description": "verbal IQ standard score measured with WISC-IV, WASI, WISCC-R, two-subt
    "source_variable": "verbal_iq",
    "valueType": "http://www.w3.org/2001/XMLSchema#integer",
    "associatedWith": "NIDM",
    "isAbout": {
        "url": "http://uri.interlex.org/base/ilx_0739359",
        "label": "verbal intelligence quotient"
    }
```

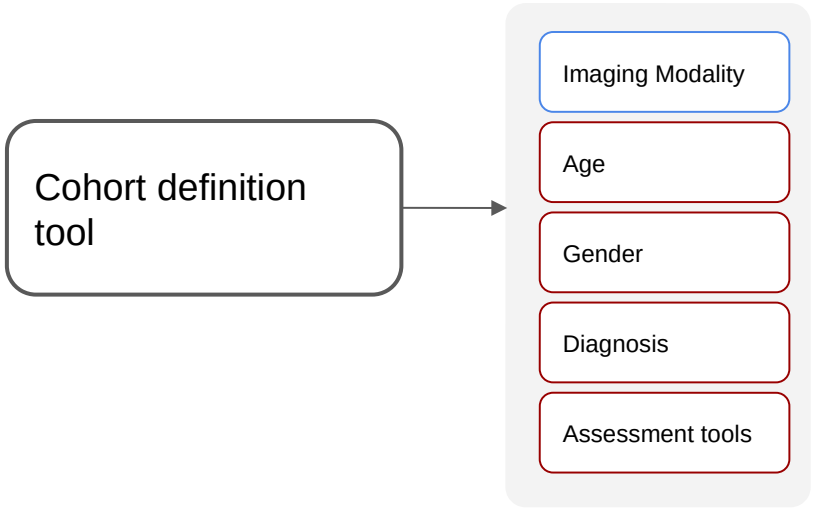
Use case: define a **clinical cohort** with multiple **neuroimaging data**

tools and services

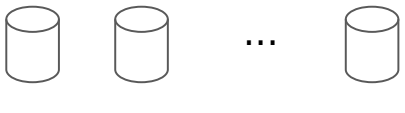


Use case: define a **clinical** **cohort** with multiple **neuroimaging** **data**

tools and services



distributed data repositories



OpenNEURO

Parkinson's Progression Markers Initiative

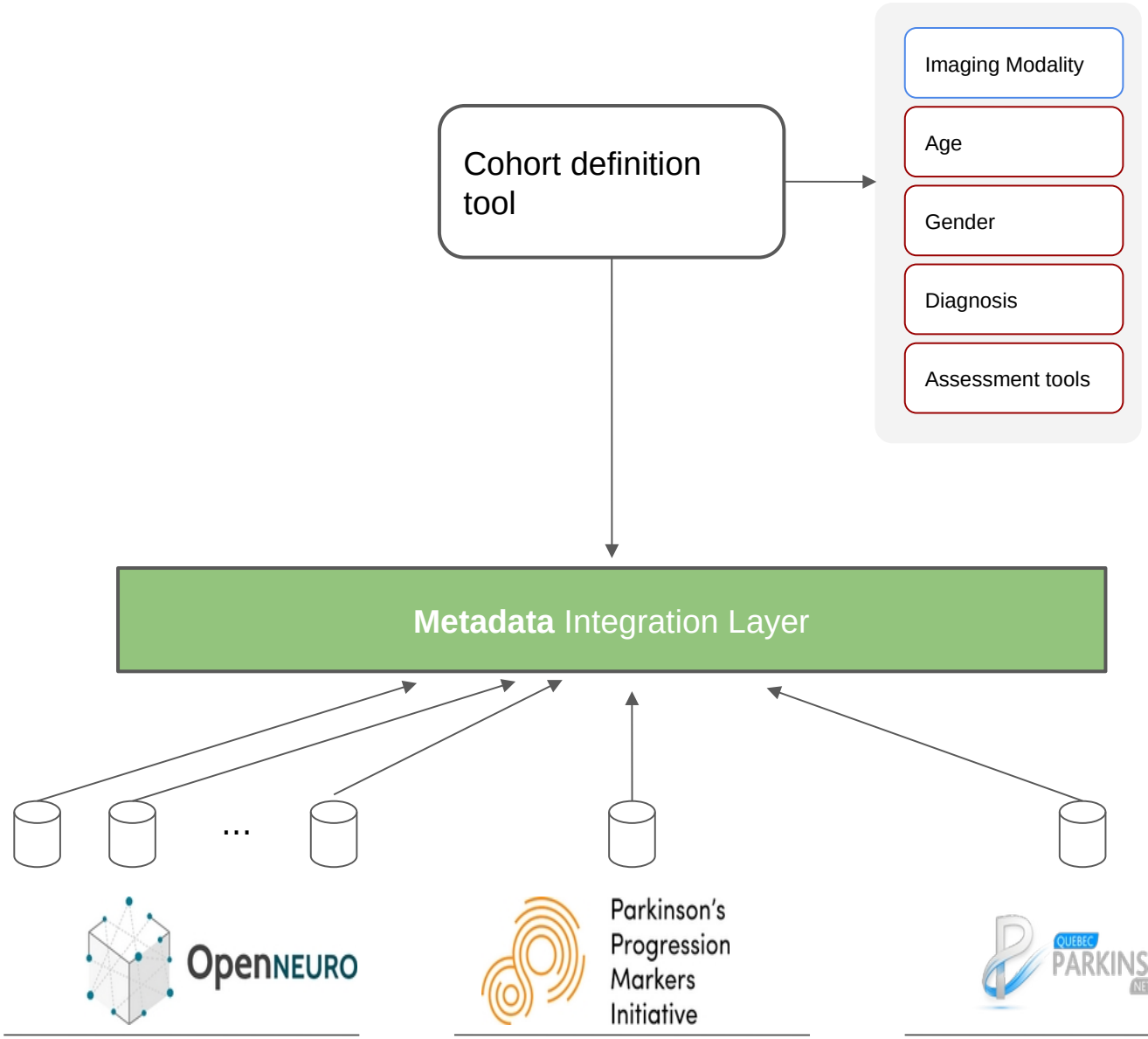
QUEBEC PARKINSON NETWORK

Use case: define a **clinical** **cohort** with multiple **neuroimaging** **data**

tools and services

integrated metadata

distributed data repositories



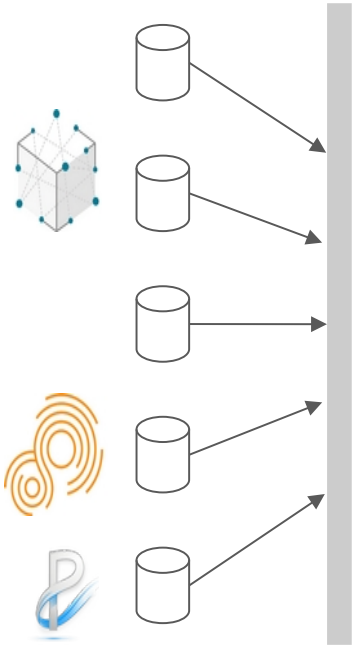
Metadata integration workflow

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



local
metadata

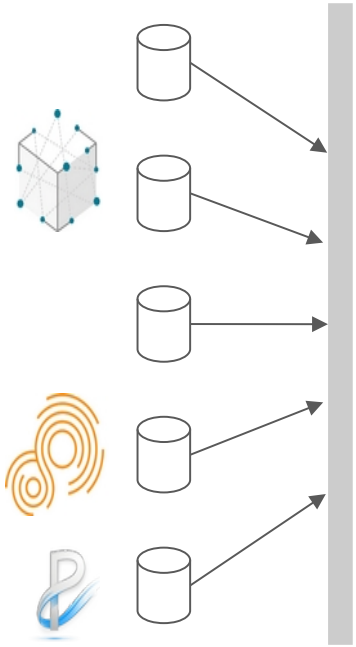
Metadata integration workflow

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



local **metadata**

1: Integration

Imaging Modality	BIDS
Age	“019Y”
Gender	“M”
Diagnosis	“PD”
Assessment tools	“MCTOT”

gathering the metadata

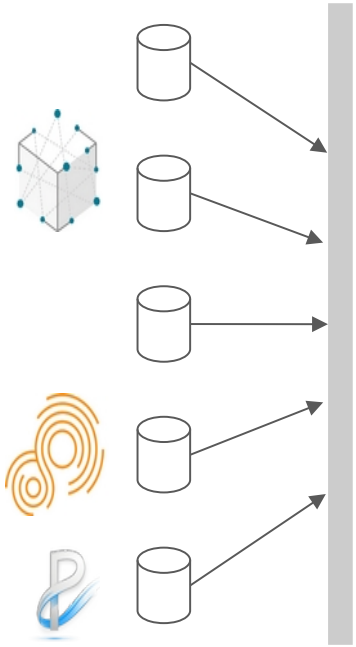
Metadata integration workflow

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



local **metadata**

1: Integration

2: Standardisation

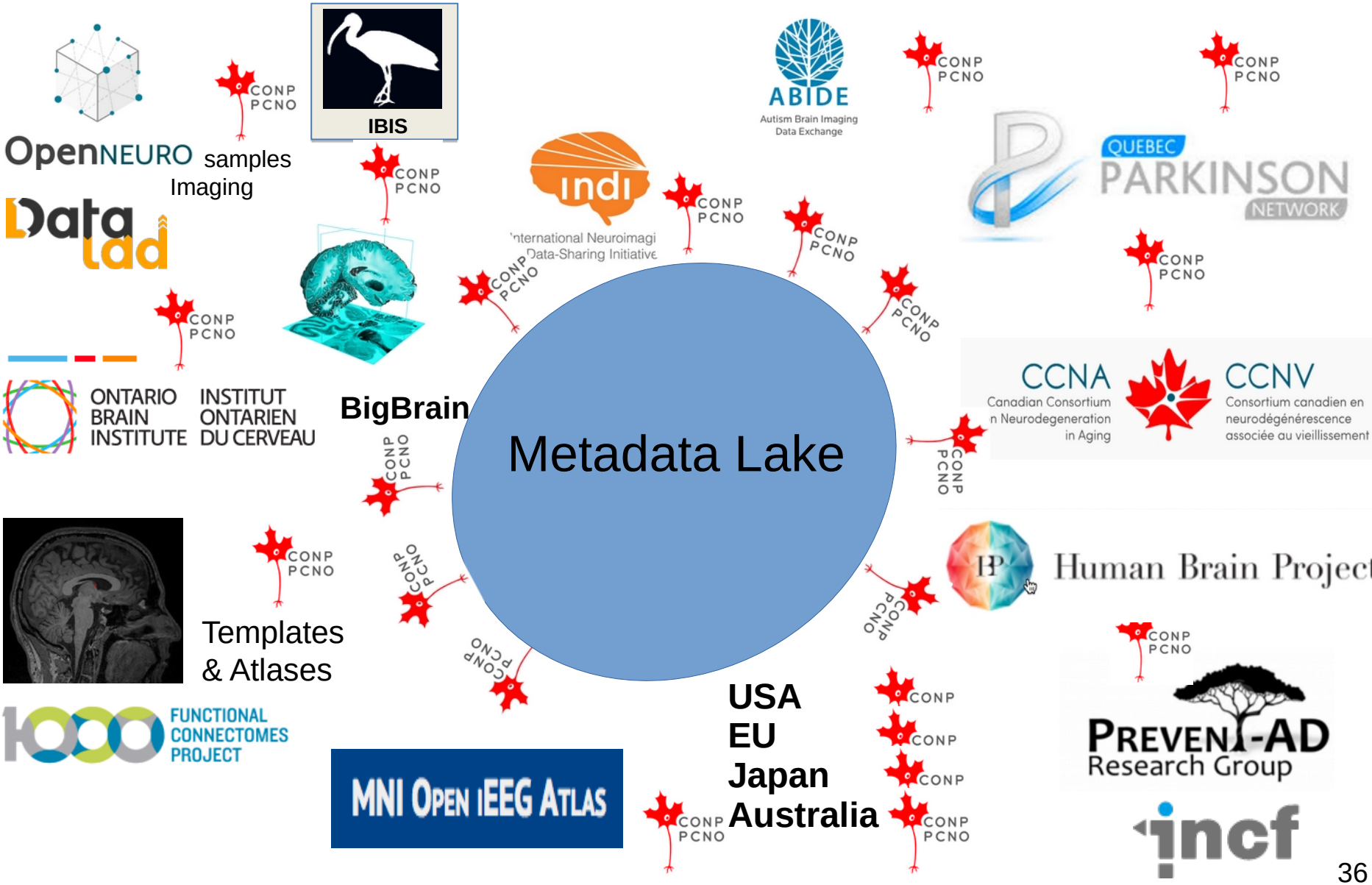
Imaging Modality — BIDS — NIDM

Age	“019Y”	19.0	years (float)
Gender	“M”	“male”	vocab term (string)
Diagnosis	“PD”	“Parkinson’s disease”	SNOMED-CT (**URI**)
Assessment tools	“MCTOT”	Montreal Cognitive Assessment	reproschema (**URI**)

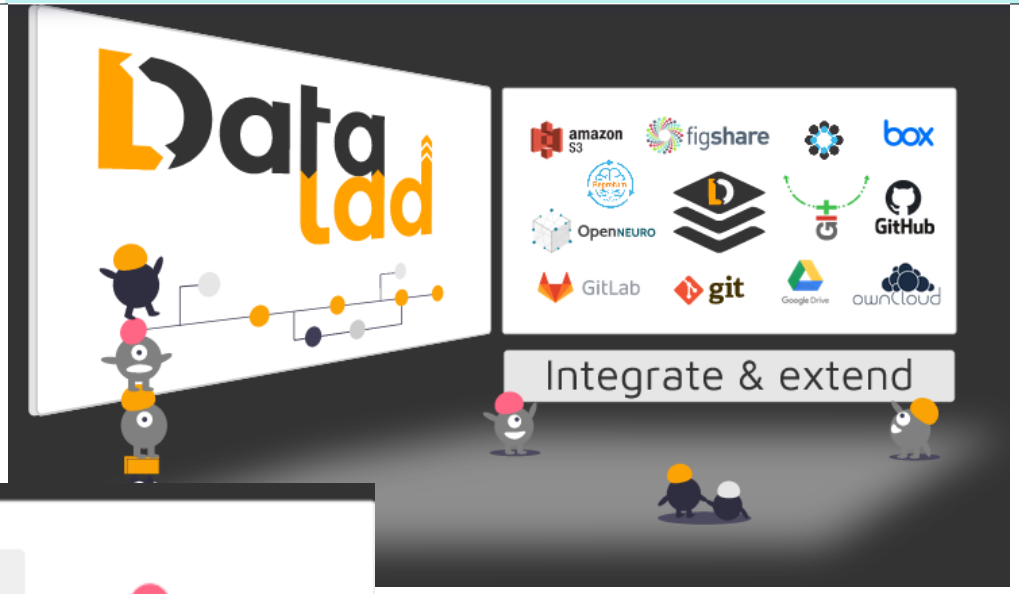
gathering the metadata

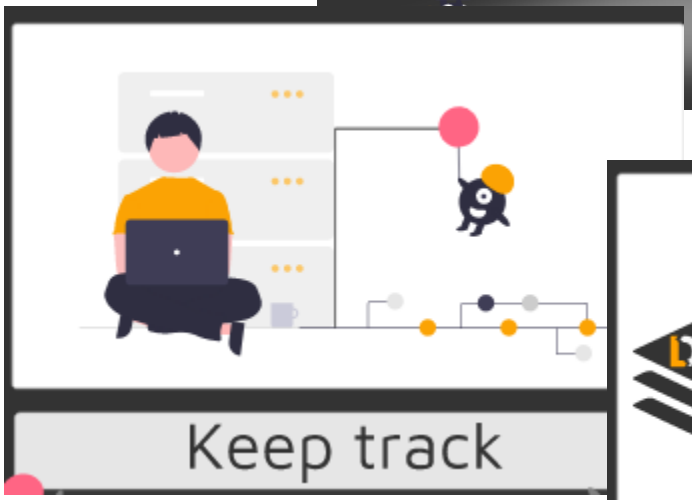
what do your metadata **mean**

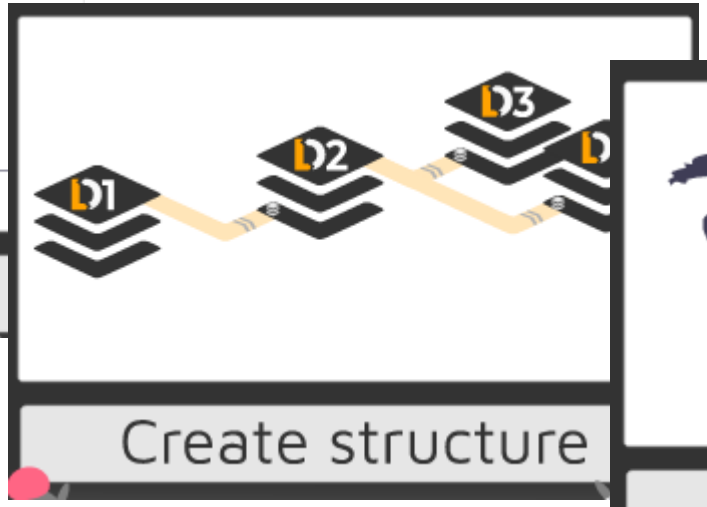
Datasets : distributed or centralized?



The mighty DataLad








Satrajit Ghosh


Seb. Urchs


Jean-Baptiste Poline


J. Armoza


Dorota Jarecka


Yaroslav Halchenko


Jeffrey Grethe


Sanu Abraham


Nazek Queder


Troy Sincomb


Karl Helmer


Tom Gillespie


David Kennedy


Theo VanErp

NIH National Institute of Mental Health

Datalad

NIDM

NeuroDebian

NITRC

NIBIB

NIF

NIPYPE

ReproNim

incf

Three causes

1. Poor statistical procedures

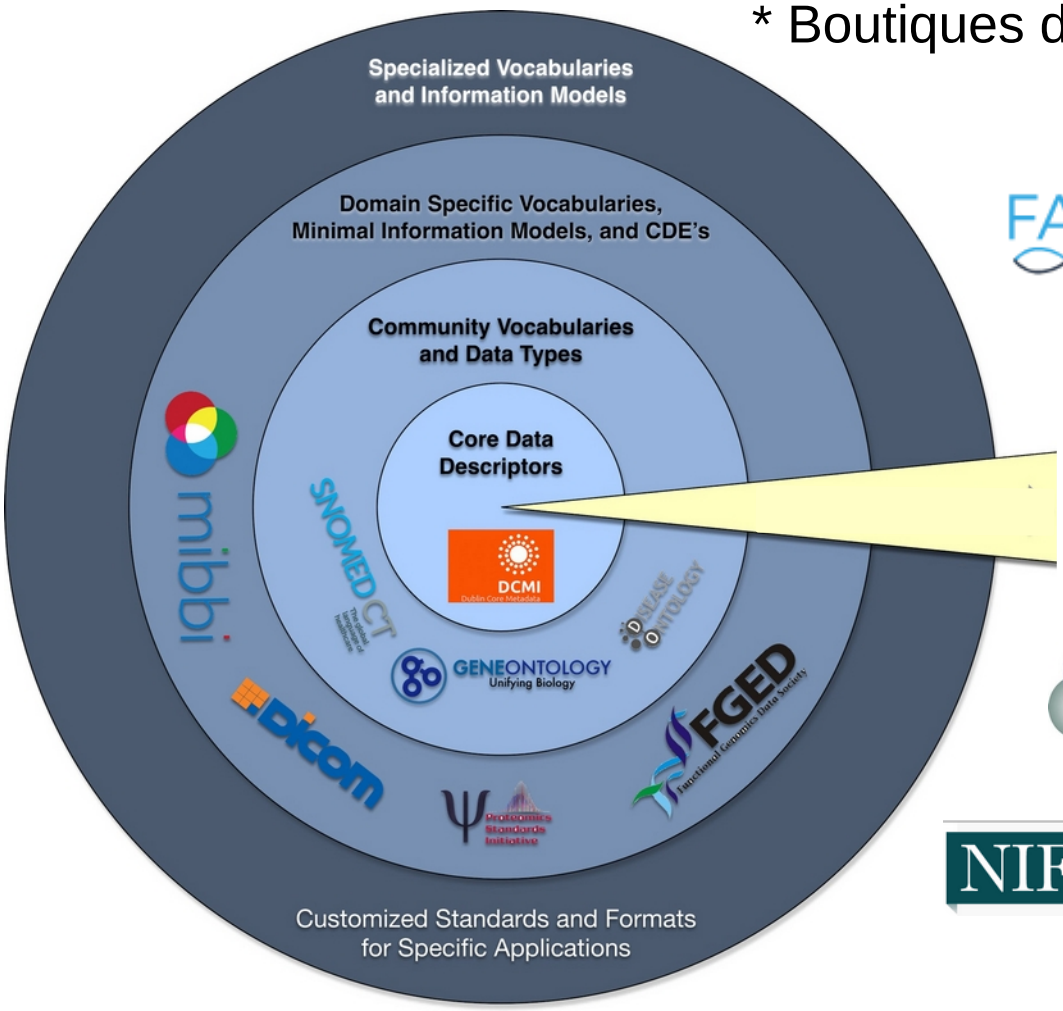
2. Issues in data and software

3. A cultural issue: Publication practices and research incentives

Neuroinformatics platforms may partly bring answers to these three gaps

Relevant Community Standards

* Reuse existing SDT (DATS, schema.org)
* Develop metadata SDT (INCF, FDE)
* Develop Landing page STD (Monii)
* Boutiques descriptors STD



FAIRsharing.org
standards, databases, policies

FORCE11
The Future of Research Communications and e-Scholarship

incf

RDA
RESEARCH DATA ALLIANCE

BETA

ReproNim

BRAIN IMAGING DATA STRUCTURE

NIF

NEURODATA WITHOUT BORDERS

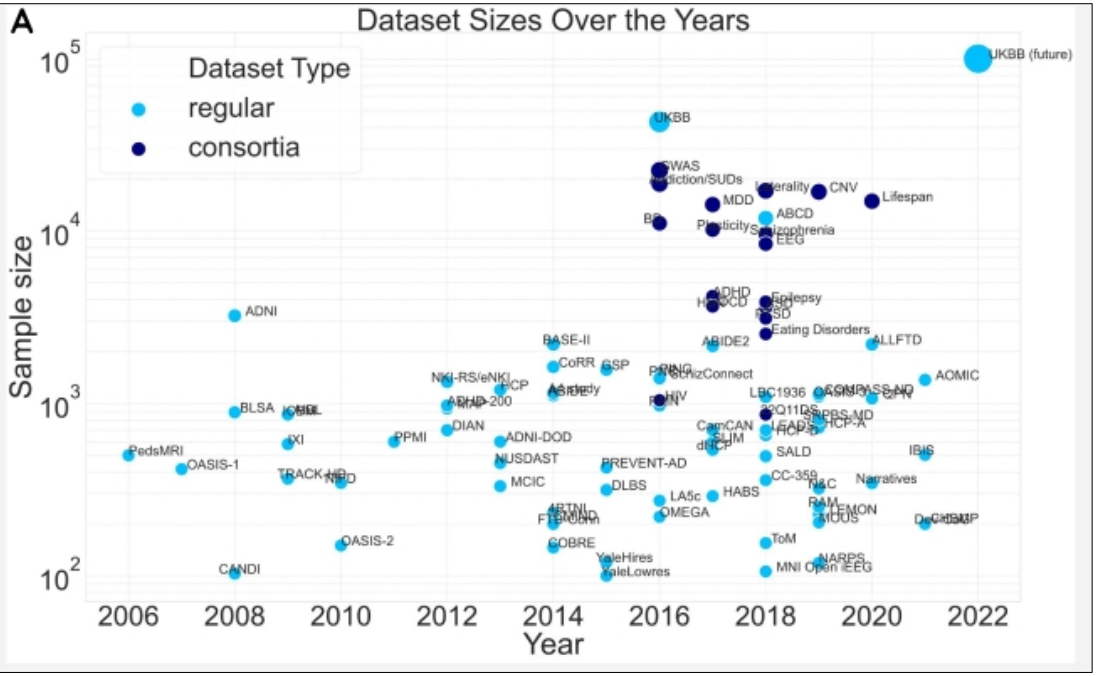
Part I: Reproducibility: background

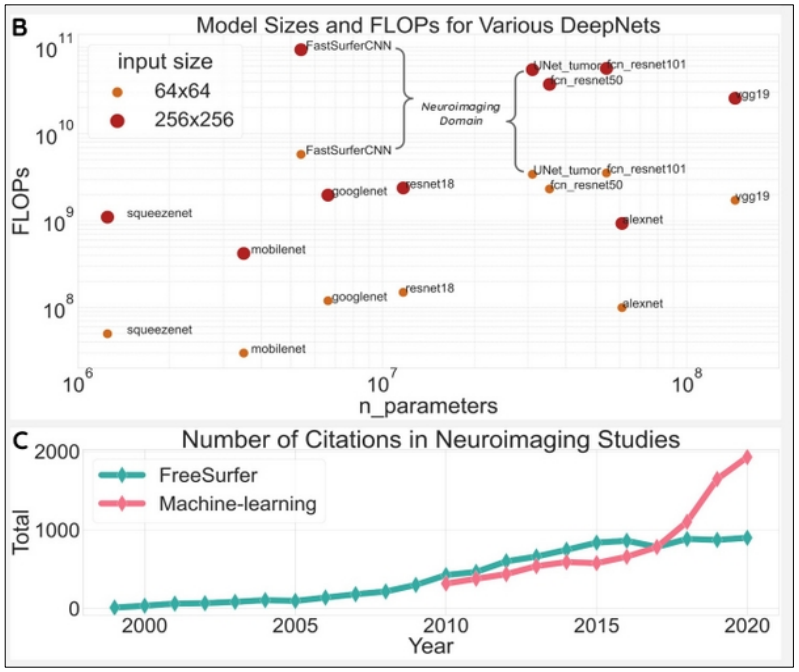
Part II : Gap analysis

Part III : NeuroInformatics platforms

Part IV : Conclusion

Is this sustainable ?





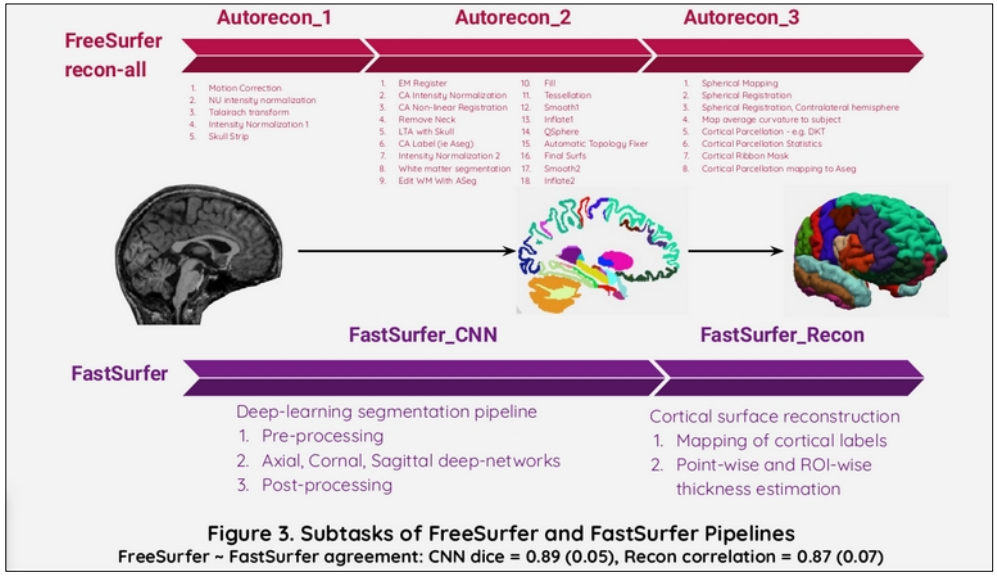


Figure 3. Subtasks of FreeSurfer and FastSurfer Pipelines
FreeSurfer ~ FastSurfer agreement: CNN dice = 0.89 (0.05), Recon correlation = 0.87 (0.07)

Nikhil Bhagwat et al., 2021

Actions for our academic system

- Sustainability : We need to recalibrate our research objectives at individual and infrastructural levels with computing “cost” in mind
- Promote journals not handled by large companies when you can
  - When asked for review or editing
  - When publishing if possible
- Promote / use / contribute to tools, standards and best practices that are
  - Community developed and governed
  - Helping the larger vision of data and code sharing
- Choose your research project wisely
  - Will it be in extreme competition ?
  - Will it help move the scientific community in a better place ?
- Is it my job ?

Thank you

- Lab@McGill**: https://neurodatascience.github.io/**
- **McGill** colleagues: S. Brown, T. Glatard, G. Kiar, A. Evans, C. Greenwood, A. DeGuise and others
- **ReproNim** colleagues: D. Kennedy, D. Keator, S. Ghosh, M. Martone, J. Grethe, M. Hanke, Y. Halchenko
- **Berkeley** colleagues: S. Van der Walt, M. Brett, J. Millman, Dan Lurie, M. D'Esposito, et al
- **Pasteur** colleagues: G. Dumas, R. Toro, T. Bourgeron, and others
- **Paris** colleagues: B. Thirion, G. Varoquaux, V. Frouin, et al
- **Funders**: McGill HBHL, HBHL NeuroHub, NIH, NIMH

NeuroDataScience: ORIGAMI lab

neurodatascience.github.io


Jean-Baptiste Poline
Principal Investigator


Kendra Oudyk
PhD student


Ting Zhang
PhD student


Gaël Varoquaux
Visiting Professor (Research Director, Inria)


Elizabeth DuPre
PhD student


Jacob Sanz-Robinson
MSc student

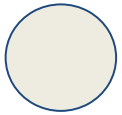

Peer Herholz
Postdoctoral researcher


Qing Wang (Vincent)
Postdoctoral researcher


Jérôme Dockès
Postdoctoral researcher


Alexandre Hutton
Research Assistant


Adam Trefonides
Research Software


Kate Kim

Data Analysis projects: ⟺	Methods development ⟺	Infrastructures and Tools
- QW	- KO	- AH
- JD	- ED	- JSR
- PH	- TZ	- AT
- KK	- GV	- JB