Open for Research:
Challenges and opportunities in re-using publicly available datasets

Elizabeth DuPre

@emdupre_
Montreal Neurological Institute
McGill University

Overview

- Re-using open data drives ongoing tool development
- Re-using open data enables new scientific research
- Sufficiently available metadata is important for appropriate re-use

Open data: acknowledging the caveats

- Many researchers do not feel comfortable with data sharing best-practices (Borghi & Van Gulick, 2018)
- Further, not all data can be made openly available (White, Blok, & Calhoun, 2020)
- Despite these constraints, we can still examine the impacts of open data on scientific research and communities

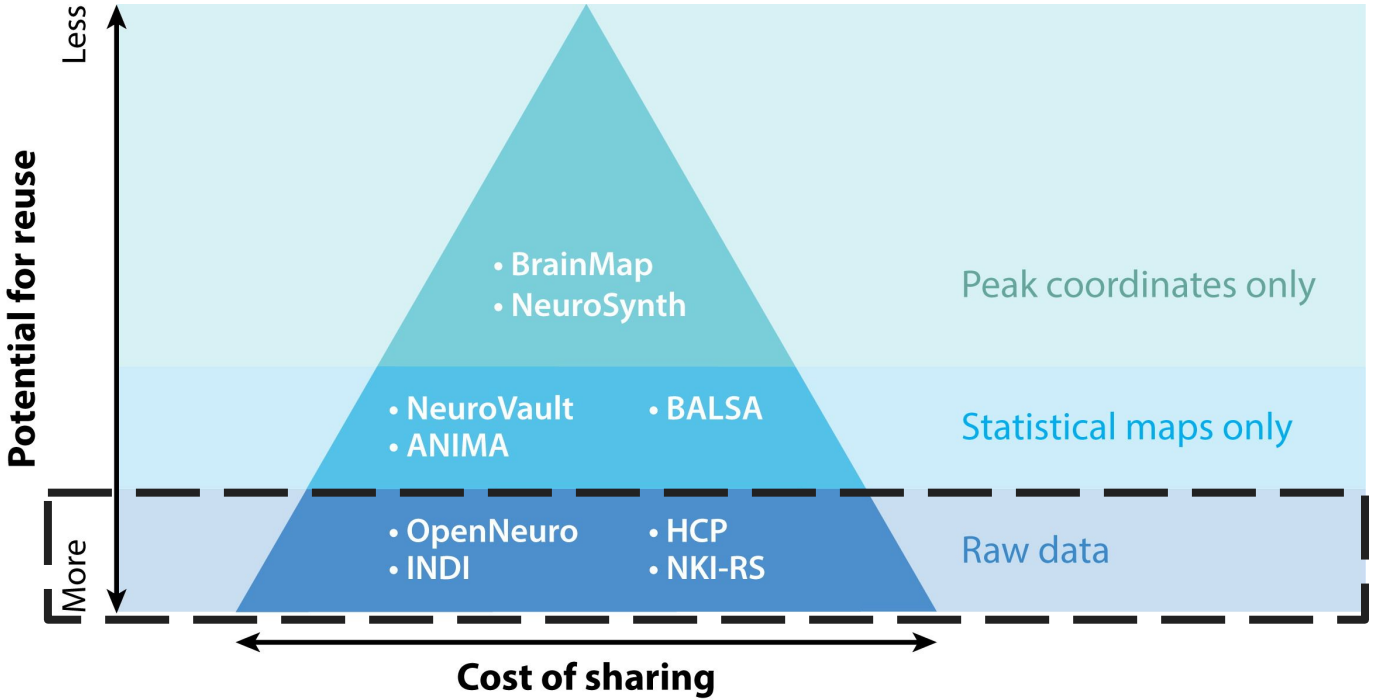


Figure adapted from
Poldrack, Gorgolewski, & Varoquaux (2019). *Annu Rev Biomed Data Sci*

Overview

- Re-using open data drives ongoing tool development
- Re-using open data enables new scientific research
- Sufficiently available metadata is important for appropriate re-use

Open data in scientific software

- The Python software library `nilearn` (http://nilearn.github.io; Abraham et al., 2014) heavily relies on open data for its example gallery
- These datasets are chosen for their didactic potential and are useful both:
  1. to ensure that the software is able to meet real-world scientific use cases
  2. to introduce community members to data analysis

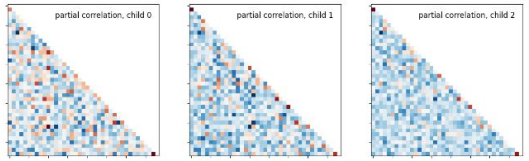
9.4.9.4. Studying partial correlations

We can also study **direct connections**, revealed by partial correlation coefficients. We just change the *ConnectivityMeasure* kind

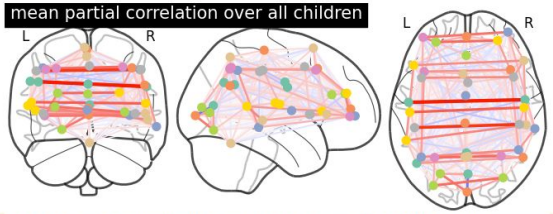
```
partial_correlation_measure = ConnectivityMeasure(kind='partial correlation')
partial_correlation_matrices = partial_correlation_measure.fit_transform(
    children)
```

Most of direct connections are weaker than full connections.

```
_, axes = plt.subplots(1, 3, figsize=(15, 5))
for i, (matrix, ax) in enumerate(zip(partial_correlation_matrices, axes)):
    plotting.plot_matrix(matrix, tri='lower', colorbar=False, axes=ax,
                         title='partial correlation, child {}'.format(i))
```




```
plotting.plot_connectome(
    partial_correlation_measure.mean_, msdl_coords,
    title='mean partial correlation over all children')
```



Out: `<nilearn.plotting.displays.OrthoProjector object at 0x7f49154e79d0>`

Data adapted from
Richardson, Lisandrelli, Riobueno-Naylor, & Saxe (2018). *Nat Comms*.

Overview

- Re-using open data drives ongoing tool development
- Re-using open data enables new scientific research
- Sufficiently available metadata is important for appropriate re-use

Open data in research studies

- Recently, we wanted to compare the performance of several functional alignment methods (Bazeille, DuPre, et al., 2020)
- Although we had in-house data available for testing, benchmarking across multiple datasets provides more reasonable estimates of performance
  - Even small variations in data characteristics can cause significant changes in performance (Recht et al., 2018)

Included datasets in Bazeille, DuPre et al. (2020)

BOLD5000	https://bold5000.github.io
Courtois NeuroMod	https://docs.cneuromod.ca
Individual Brain Charting	https://project.inria.fr/IBC
Study Forrest	https://www.studyforrest.org

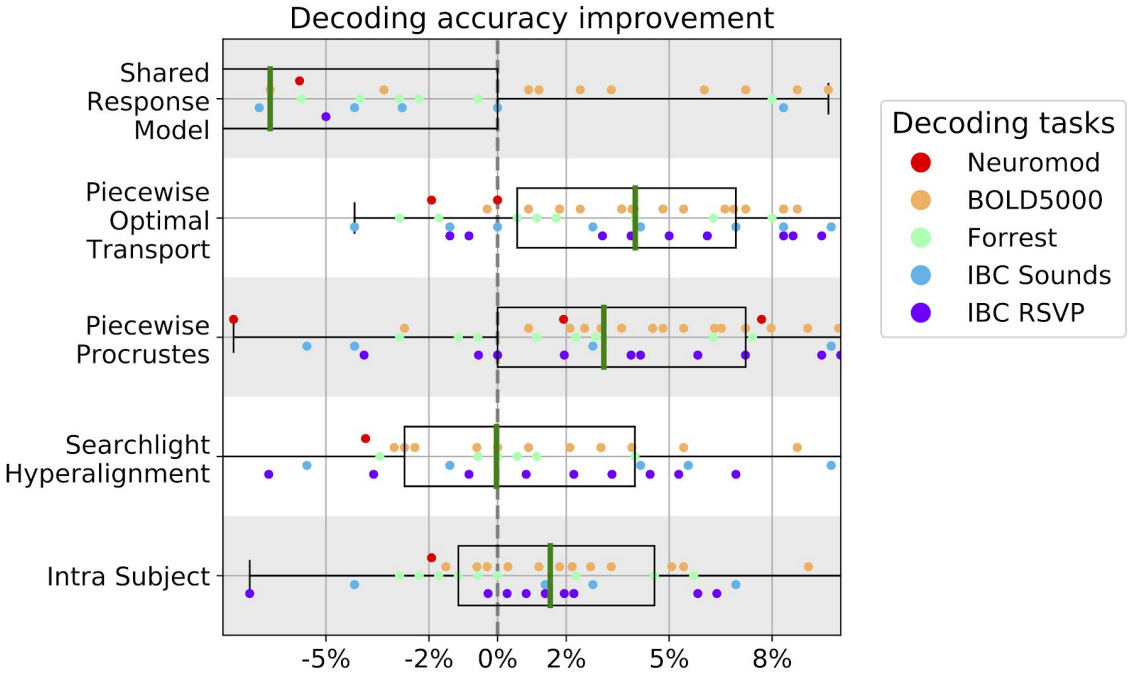


Figure adapted from [Bazeille, DuPre, Poline, & Thirion (2020). *BioRxiv*]

Overview

- Re-using open data drives ongoing tool development
- Re-using open data enables new scientific research
- Sufficiently available metadata is important for appropriate re-use

Challenges in reusing open datasets

- Finding appropriate, well-described datasets remains a challenge
  - Standards like BIDS (Gorgolewski et al., 2016) and NIDM (Keator et al., 2013) help to address this
- For naturalistic neuroscience in particular, finding datasets with fully available stimulus information is difficult due to copyright restrictions (DuPre, Hanke, and Poline, 2019)

Recommendations for sharing naturalistic stimuli

- Directly share stimuli if your local copyright law allows
- Consider using public domain stimuli, or re-using stimuli that have already been publicly shared
- If you cannot release the stimuli, provide sufficient information such that another researcher could recreate your materials; for example using reporting guidelines from Vanderwal, Eilbott, & Castellanos (2018)

Thank you

Jean-Baptiste **Poline**

The **ORIGAMI** Lab

Thomas **Bazeille**

Bertrand **Thirion**

Michael **Hanke**

Individual Brain Charting Project

Courtois-NeuroMod

… and you for your attention !

Take-home ideas

- Openly sharing data creates new opportunities in scientific tool development, research
- Challenges in describing and sharing part or all of datasets makes re-use more difficult