Characterizing the Stability of Neuroimaging Analyses Through Perturbations in Experimental Design

Greg Kiar

2018-2021 NSERC Alexander Graham Bell Fellow,
2017 McGill University Healthy Brains for Healthy Lives Fellow,
McGill Centre for Integrative Neuroscience, Montreal Neurological Institute,
Ph.D. student McGill University,
M.S.E. Johns Hopkins University,
B.Eng Carleton University

Outline

- Background & Overview
- Chapter 1: Scalable and Provenance Rich Pipeline Deployment (Clowdr)
- Chapter 2: Evaluating the Stability of Neuroimaging Pipelines & Analyses
- Chapter 3: Exploring Sources of Instability & Dependence Within Pipelines
- Conclusion

Background & Overview

Reproducibility in Neuroscience

- Noisy data and incomplete statistics can lead to spurious results (Bennett et al., 2011) (fMRI)
- Operating system differences have led to different results (Glatard et al., 2015) (sMRI)
- Dominant software libraries have inflated false-positive rates (Eklund et al., 2016) (fMRI)
- 1-voxel perturbations to inputs result in significantly different outputs (Lewis et al., 2016) (sMRI)
- Similar tools performing similar operations give different results (Bowring et al., 2018) (fMRI)

Currently missing in neuroimaging:

1. Infrastructure for easily running and capturing “repro-analyses” at scale
   → I have created an infrastructure for this purpose

2. A consistent method for evaluating the stability of results and tools
   → I will develop a metric for evaluating stability of neuroimaging analyses

3. Methods for identifying sources of instability within pipelines
   → I will use the metric above to explore the impact of individual processes on pipeline stability

(and, applications of “repro-analyses” to diffusion neuroimaging, which I will focus on)

Chapter 1: Scalable and Provenance Rich Pipeline Deployment (Clowdr)

1 year; complete

Clowdr is…

- Server-less microtool for running pipelines at scale on HPC and cloud systems
- Captures system-level provenance information (i.e. CPU/RAM usage) and Reprozip
- Provides an interactive web-report for exploring and sharing experiments.

(Kiar, 2018; in review)

Analysis with Clowdr

```
$   # Installable on Python3…
$   pip install clowdr
$
$   # Run locally/on clusters, the cloud, and share results
$   clowdr local {tool} {invocation} {dataset} {output loc}
$   clowdr cloud {tool} {invocation} {dataset} {output loc} {cloud} {keys}
$   clowdr share {task loc} # {task loc} returned by any of the above
$
```

Chapter 2: Evaluating the Stability of Neuroimaging Pipelines & Analyses

1.5 years (total: 2.5 years)

In linear systems, this has been solved

Condition number of $\mathbf{Af} = \mathbf{x}$ can be evaluated as:

$$\kappa(\mathbf{A}) = \|A\|\|A^{-1}\| \geq \max_{x, f(x) \neq 0} \frac{|\delta f(x)| / |f(x)|}{|\delta x| / |x|}$$

$$\delta f(x) = f(x + \delta x) - f(x)$$

"maximum ratio of change in output, f, with respect to change in input, x."

(Davidson, 1981)

Applications in Diffusion Tensor Imaging

DWI tensor model: $S_i = S_0 e^{-b\mathbf{g}_i^T \cdot \mathbf{D} \cdot \mathbf{g}}$

We can rearrange this with a couple clever substitutions…

Applications in Diffusion Tensor Imaging

DWI tensor model: $S_i = S_0 e^{-b\mathbf{g}_i^T \cdot \mathbf{D} \cdot \mathbf{g}}$

We can rearrange this with a couple clever substitutions…

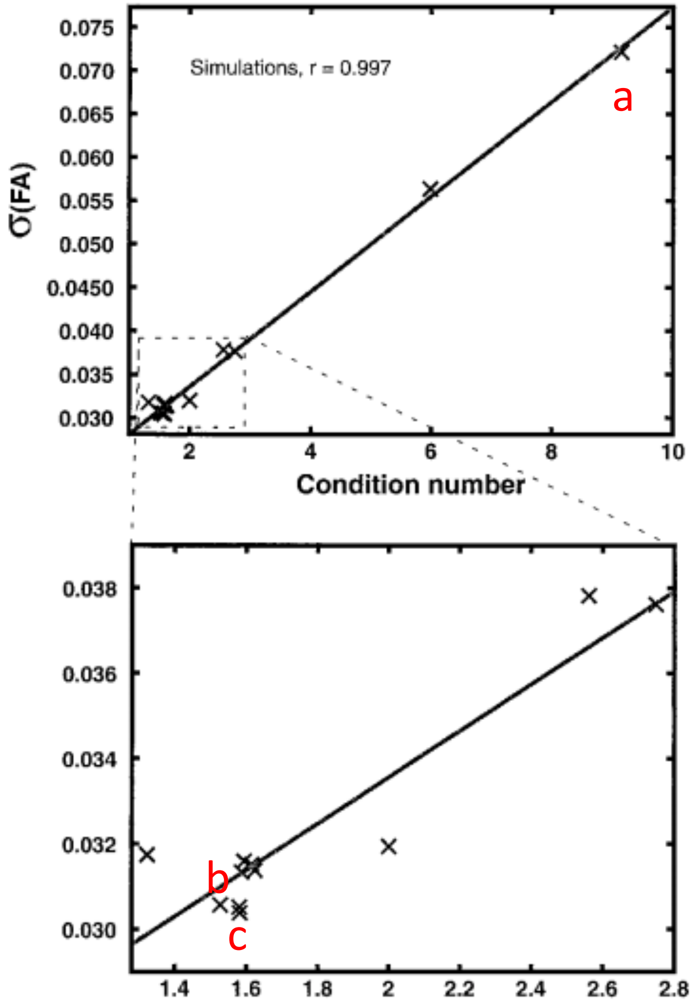
$$\mathbf{X} = (D_{xx}, D_{yy}, D_{zz}, D_{xy}, D_{xz}, D_{yz})^T$$
$$a_i = (g_{ix}^2, g_{iy}^2, g_{iz}^2, 2g_{ix}g_{iy}, 2g_{ix}g_{iz}, 2g_{iy}g_{iz})$$
$$a_i^T\mathbf{X} = ln(S_0/S_i)/b = ADC_i$$

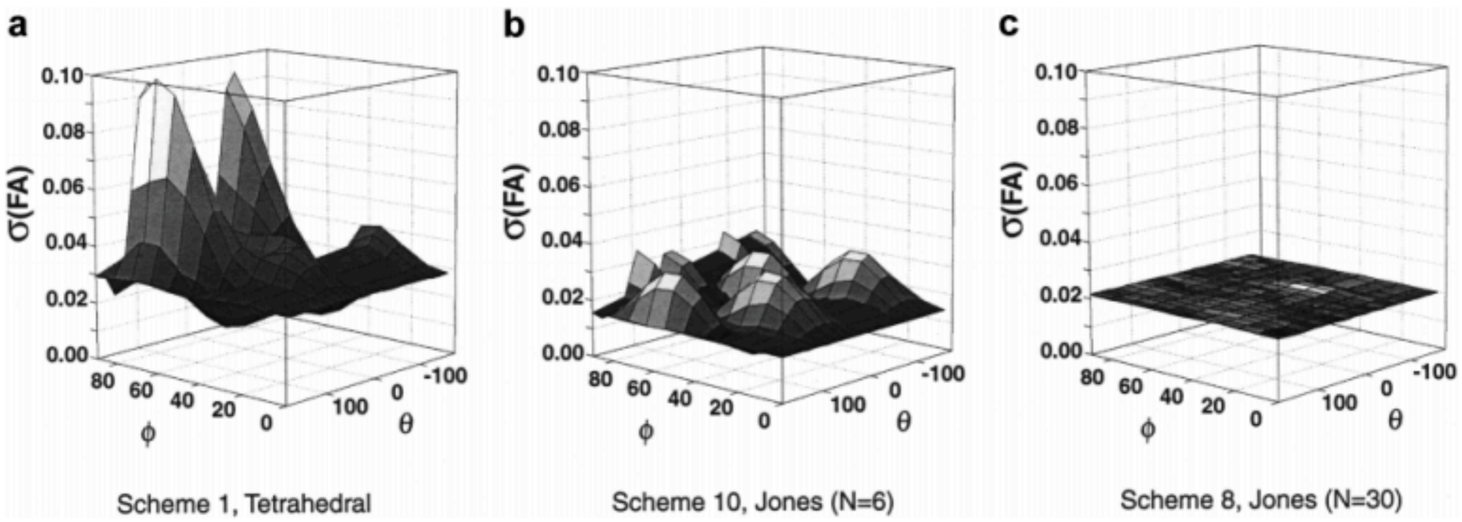
$$\mathbf{ADC} = (ADC_1, ADC_2, ..., ADC_N)^T \qquad \mathbf{A} = (a_1, a_2, ..., a_N)^T$$

$\underline{\mathbf{AX} = \mathbf{ADC}}$ … which is the same form as earlier

(Skare, 2000)

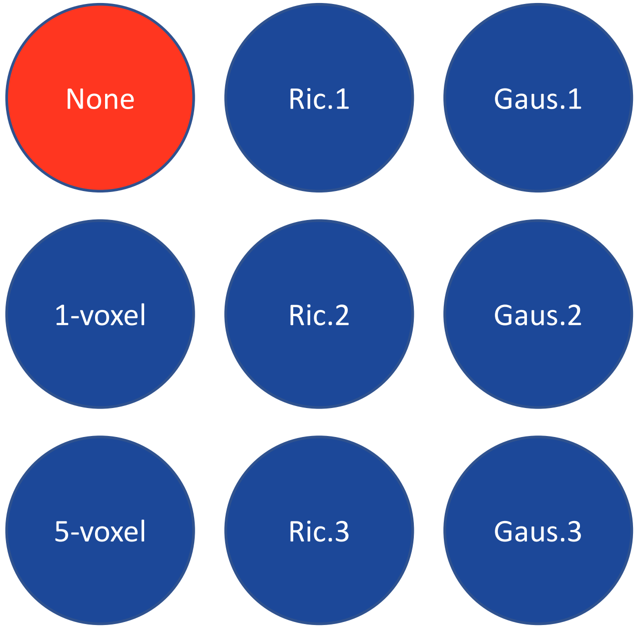
Stability of Tensor Estimation





(Skare, 2000)

Example comparison: noise effects



$\mathcal{A}$: None, N-voxel, Ric., Gaus.

$\mathcal{D}$: None

$\mathcal{A} - \mathcal{D}$: Rician, N-voxel, Gaussian

" How similar are connectomes with no noise to those with Rician-, N-voxel- and Gaussian-noise? "

$$\kappa(\mathbf{A}) \geq \max_{x, f(x) \neq 0} \frac{|\delta f(x)|/|f(x)|}{|\delta x|/|x|}$$

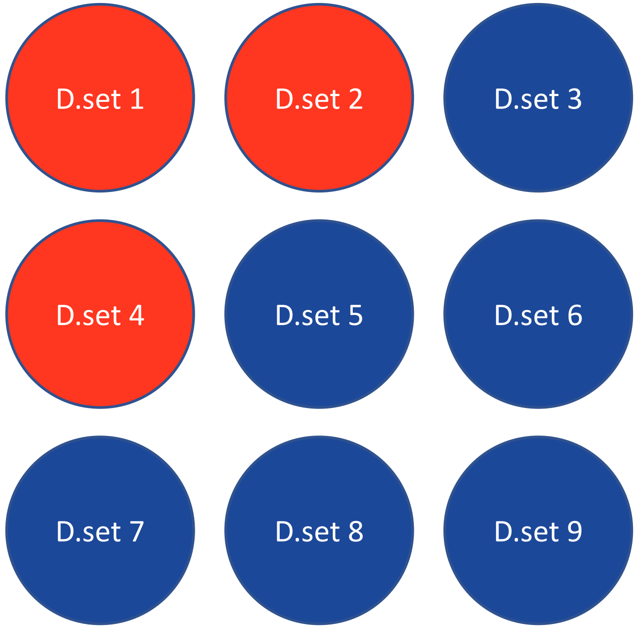
Evaluating stability with respect to data

In this case, $\mathcal{A}$ describes {datasets, subjects, noise, etc.}.

$$\hat{\kappa}(\mathcal{A}, \mathcal{D}) \geq \max_{x, x_d} \frac{\|f(x_d) - f(x)\|_R / \sigma_{f(x)}}{\|x_d - x\|_I / \sigma_x}$$

$$x_d \in \mathcal{D} \qquad\qquad x = x_d - \delta x \in \mathcal{A} - \mathcal{D}$$

Example: Evaluating dataset effects



$\mathcal{A}$: D.set {1,2,3,4,5,6,7,8,9}

$\mathcal{D}$: D.set {1,2,4}

$\mathcal{A} - \mathcal{D}$: D.set {3,5,6,7,8,9}

"*How similar are connectomes from datasets {1,2,4} to those from datasets {3,5,6,7,8,9}?*"

Example: Evaluating tool effects

$\mathcal{A}$: FSL, MRtrix, Dipy

$\mathcal{D}$: FSL

$\mathcal{A} - \mathcal{D}$: MRtrix, Dipy

"

How similar are connectomes from FSL to those from MRtrix/Dipy?

"

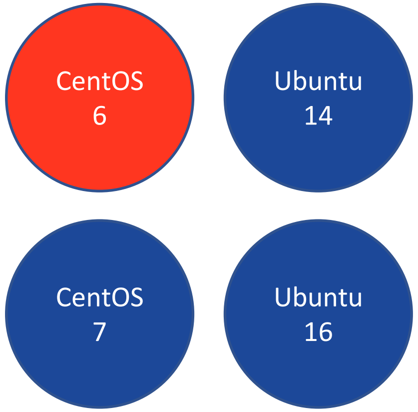
Evaluating stability with respect to tool

In this case, $\mathcal{A}$ describes {tool, operating system, etc.}.

$$\hat{\kappa}_t(\mathcal{A}, \mathcal{D}, \mathbf{x}) \geq \max_{x \in \mathbf{x}} \|f_d(x) - f(x)\|_R / \sigma_{f_d}$$

$$f \in \mathcal{A} - \mathcal{D} \qquad\qquad f_d \in \mathcal{D}$$

Example: Evaluating OS effects



$\mathcal{A}$: CentOS, Ubuntu

$\mathcal{D}$: CentOS 6

$\mathcal{A} - \mathcal{D}$: CentOS 7, Ubuntu

" How similar are connectomes generated on CentOS 6 to those generated on CentOS 7 and Ubuntu? "

Experiment: Estimating Stability

Purpose	Characterizing the instability and variability of analyses		
Outcomes	❑ Metric for evaluating the stability of analyses with respect to dependent experimental variables ❑ Exploration of the variability introduced in Diffusion MRI experiments by dataset, noise, and tool selection		
Datasets	**Modality**	**Derivatives**	**Tools**
Consortium of Reproducibility and Reliability	Diffusion MRI, Structural MRI (Functional MRI)*	Structural Connectomes (Functional Activation Maps)*	Dipy, FSL, MRtrix (SPM, AFNI, FSL)*

Experiment	❑ Partial replication of [27] comparing conditioning to observed variance
	❑ Determine a space-independent proxy for conditioning
	❑ Process CoRR datasets using default Dipy, FSL, and MRtrix pipelines
	❑ Reprocess CoRR datasets with:
	    ❑ 1-voxel perturbation
	    ❑ Rician noise
	    ❑ Gaussian noise
	❑ Calculate conditioning across:
	    ❑ Noise (fixed tool and dataset)
	    ❑ Datasets (fixed tool)
	    ❑ Tool (fixed datasets)
	❑ Compare each setting, and identify axes and regions of instability		

Notes	❑ *Based on collaboration with Dr. Camille Maumet, this work may be extended to cover functional MRI applications. This will leverage her experience with fMRI evaluation, and will require the development of theory similar to that presented in [27] on algorithms used in fMRI.		

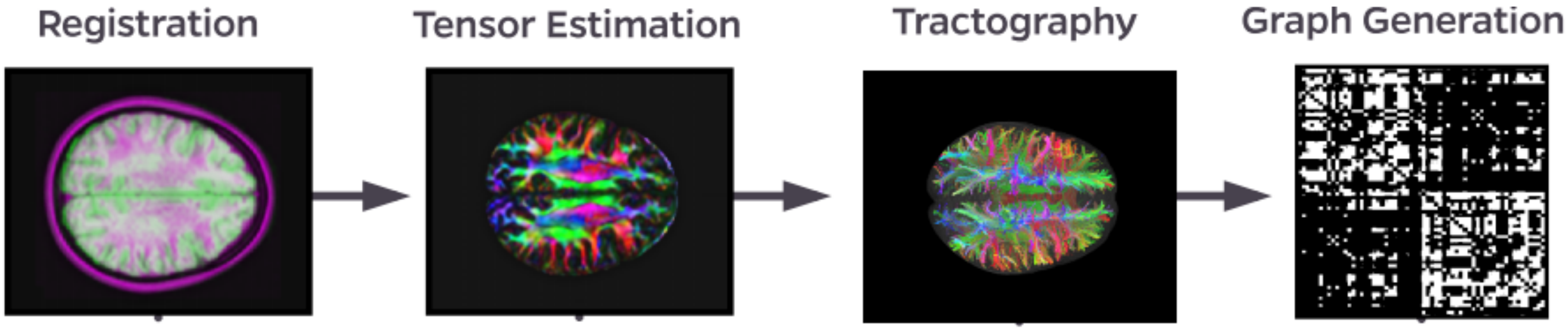
Paper 1

Paper 2

Chapter 3: Exploring Sources of Instability & Dependence Within Pipelines

1 year (total: 3.5 years)

Evaluating pipeline components



$$\hat{\kappa}_t = f(\ \hat{\kappa}_{t,1} \quad , \quad \hat{\kappa}_{t,2} \quad , \quad \hat{\kappa}_{t,3} \quad , \quad \hat{\kappa}_{t,4}\ )$$

(Kiar, 2018)

Experiment: Sources of Instability

Purpose	Explaining sources of instability and variability in tools
Outcomes	❑ Comparison of the stability of individual pipeline components with the overall tool stability ❑ Identification of sources of instability in pipelines ❑ Principled method for reconstructing pipelines with stable algorithms

Datasets	**Modality**	**Derivatives**	**Tools**
Consortium of Reproducibility and Reliability	Diffusion MRI, Structural MRI	Structural Connectomes	Dipy, FSL, MRtrix

Experiment

- ❑ Dissect structural pipelines into independently runnable components
- ❑ For each pipeline component, beginning with the first:
  - ❑ Process CoRR datasets (or derivatives of previous step) with:
    - ❑ 1-voxel perturbation
    - ❑ Rician noise*
    - ❑ Gaussian noise
  - ❑ Calculate conditioning for each component across:
    - ❑ Noise (fixed tool and dataset)
    - ❑ Datasets (fixed tool)
    - ❑ Tool (fixed datasets)
- ❑ Compare each algorithm for each setting

Notes

- ❑ *Rician noise will only be added to either raw MR images or minimally preprocessed MR images, since it is unexpected in other contexts.

Conclusion

Expected collaborations

- Boutiques (Glatard)
- Enhancing data discovery and querying (Poline)

Tool Development

- Evaluating the stability of functional MRI software (Maumet)

Stability Analysis

- Mapping structural and functional connectivity (Suarez, Misic)
- Network evolution in development (Khundrakpam)
- Heritability of structural connectomes (Vogelstein, Priebe)

Connectomics

In summary

- Replicability can be difficult to achieve and assess in neuroimaging
- I have developed a tool increasing the ease with which scientists can perform repro-analyses
- I will develop a metric for evaluating the stability of results and identify their dependence on various variables such as tool, dataset, and noise

Acknowledgements

Fondation Brain Canada Foundation

HEALTHY BRAINS FOR HEALTHY LIVES

NSERC CRSNG

CANADA FIRST RESEARCH EXCELLENCE FUND

APOGÉE CANADA FONDS D'EXCELLENCE EN RECHERCHE

...

All code mentioned in this presentation is publicly available on GitHub.

Thanks!

Find me @

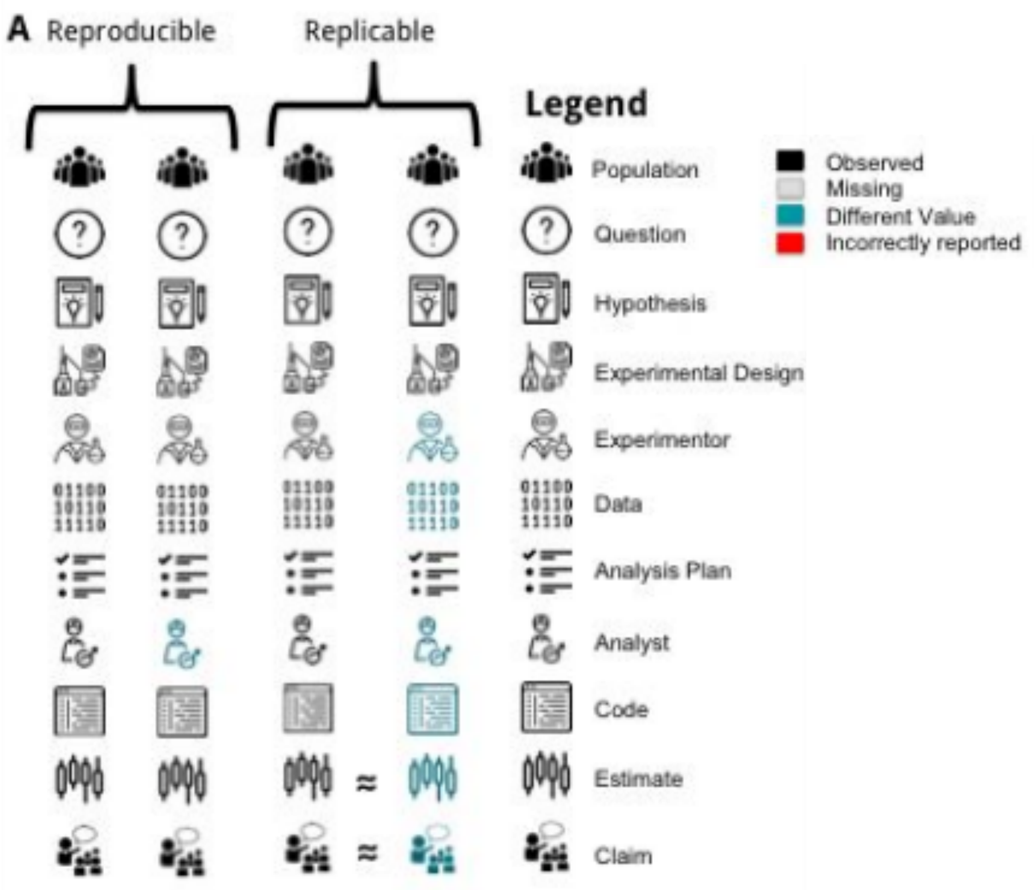
gkiar

g_kiar

greg.kiar@mcgill.ca

Extras

Reproducibility or Replicability?



(Patil, 2016)

Replicability is a measurable problem

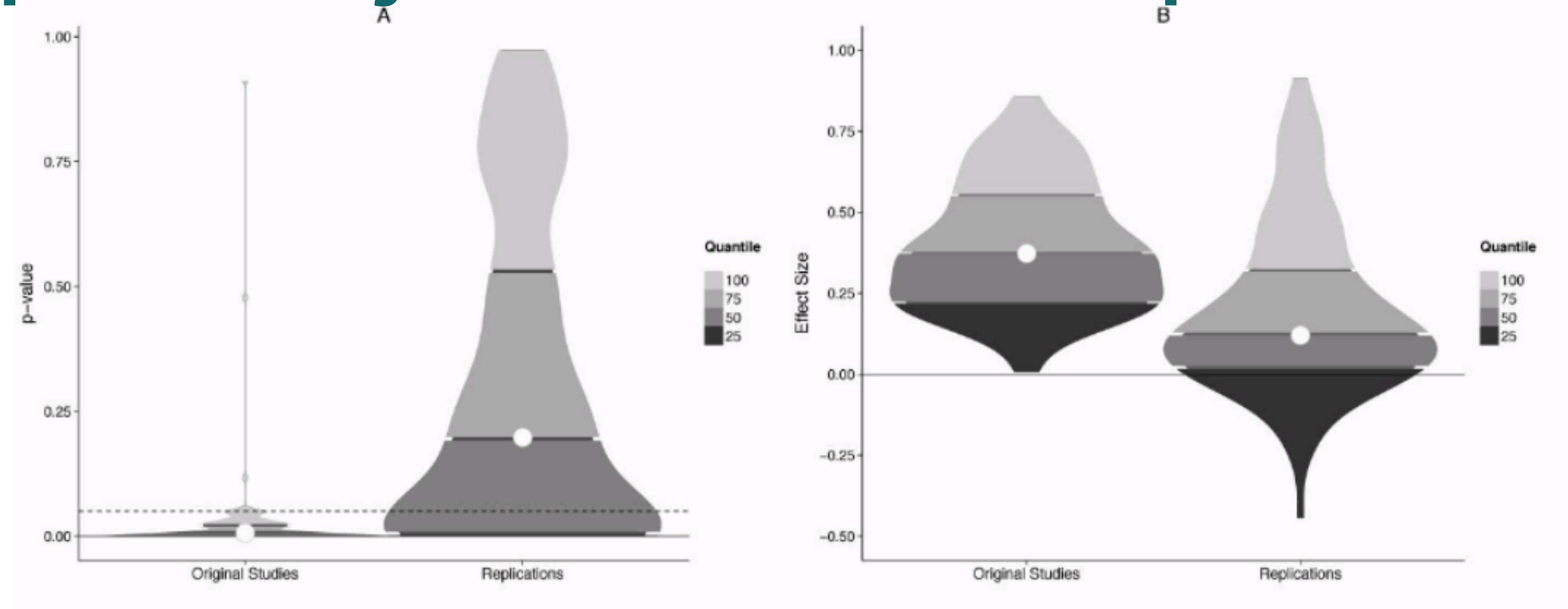


Fig. 1. Density plots of original and replication *P* values and effect sizes. (A) *P* values. (B) Effect sizes (correlation coefficients). Lowest quantiles for *P* values are not visible because they are clustered near zero.

(Open Science Collaboration, 2015)

Many tools make analyses accessible

- Standards for data and code interoperability increase re-usability
  … but, require in depth knowledge of the data/code in question

- Software virtualization allows for portable code deployment
  … but, different virtualizations are required for different systems

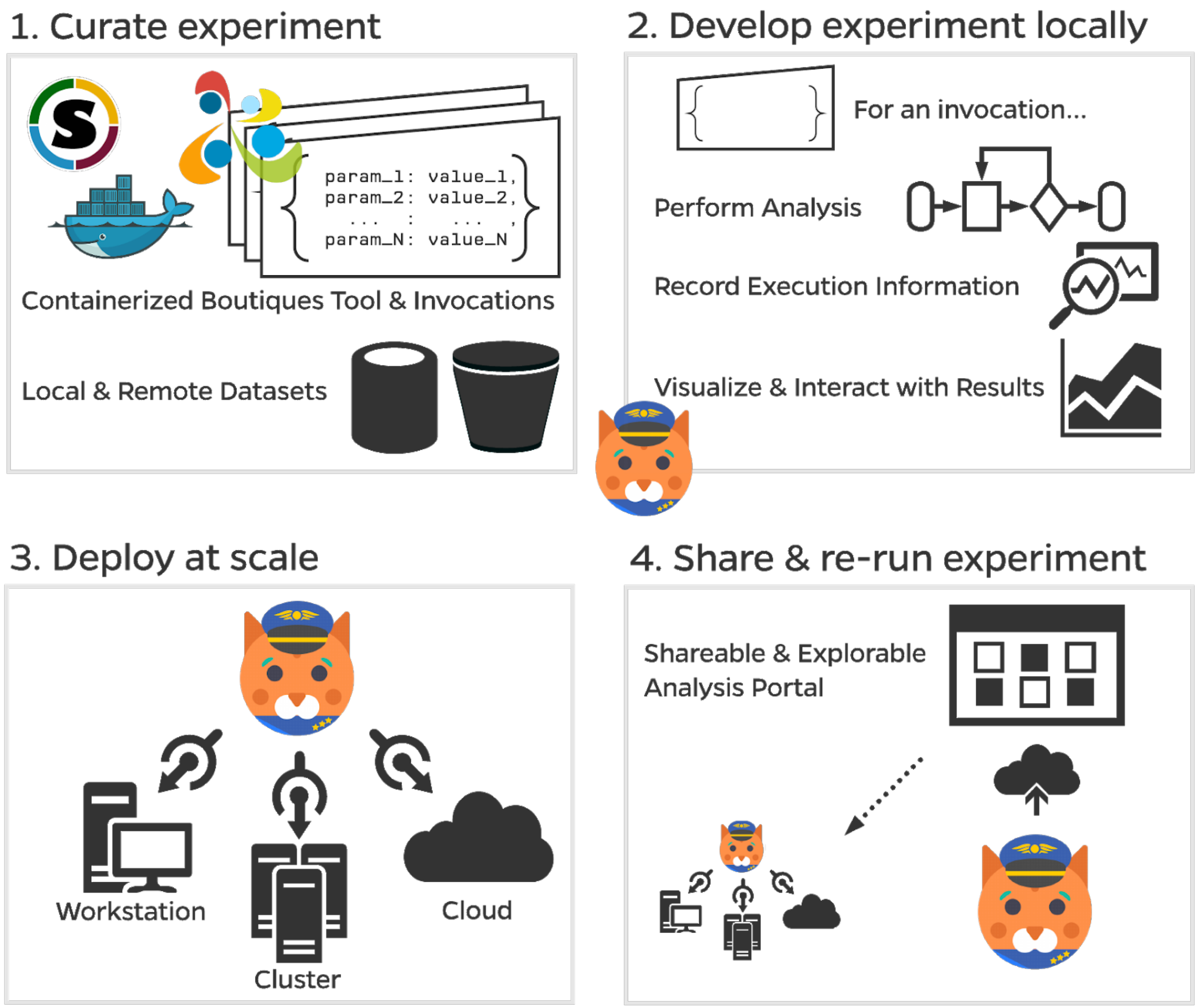
- Workflow engines enable constructing graphs between processing steps
  … but, are tied to specific programming languages and constructs

- Capturing provenance records informs analysis and future experiments
  … but, provenance tools and standards are typically complex and unintuitive

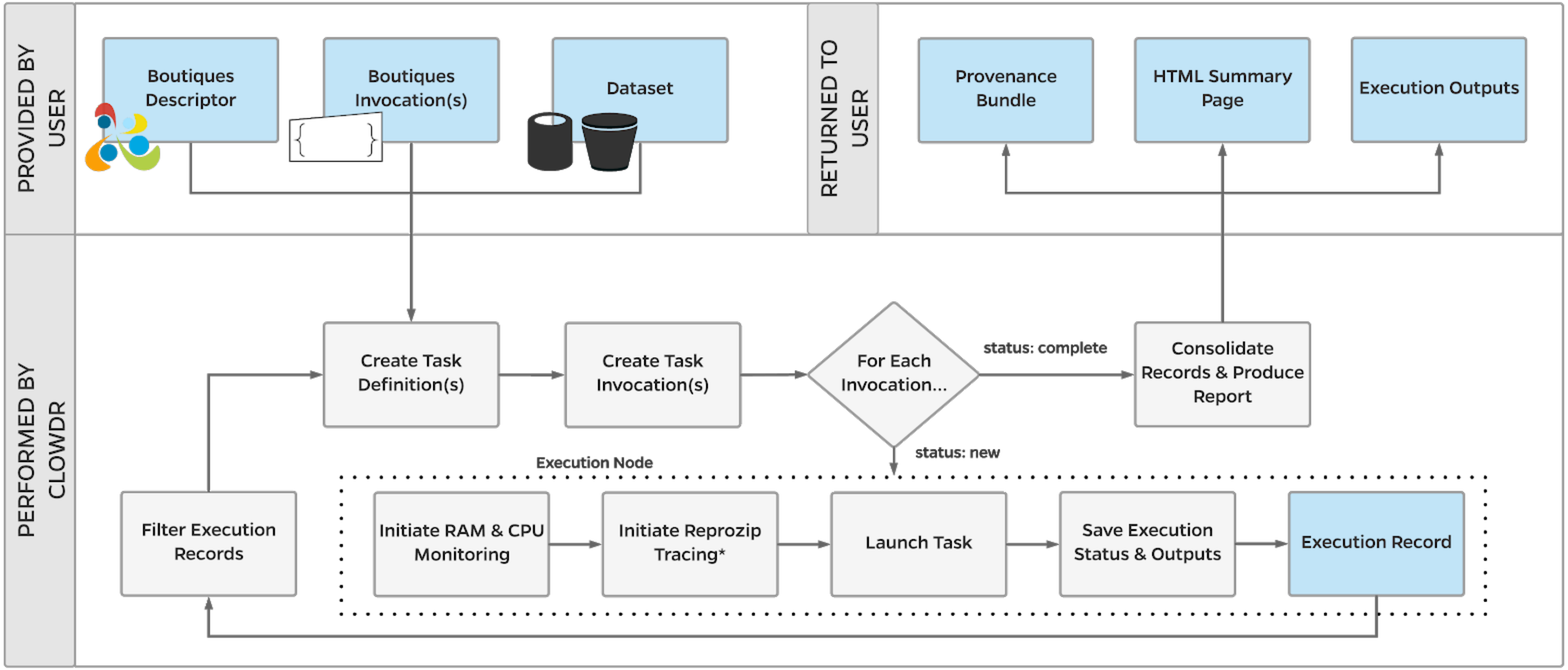
- Navigating through web platforms is user-friendly and intuitive
  … but, they are bulky and don't allow for the development or prototyping of tools and analyses

Clowdr …

- is based on Boutiques and is BIDS-aware
- runs bare-metal and Docker/Singularity virtualized tools on HPC systems and clouds
- supports the batch deployment of pipelines constructed with workflow-engines
- captures system-level provenance information (i.e. CPU and RAM usage) and Reprozip
- supports both development- and production-level tools without an active server, and provides a web-report for exploring and sharing experiments.



(Kiar, 2018; in review)

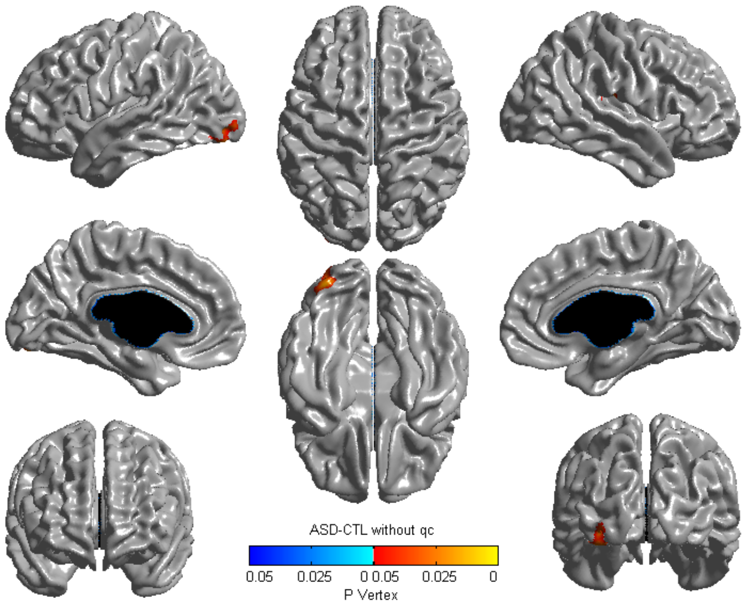


(Kiar, 2018; in review)

Differences in ABIDE nulled with motion

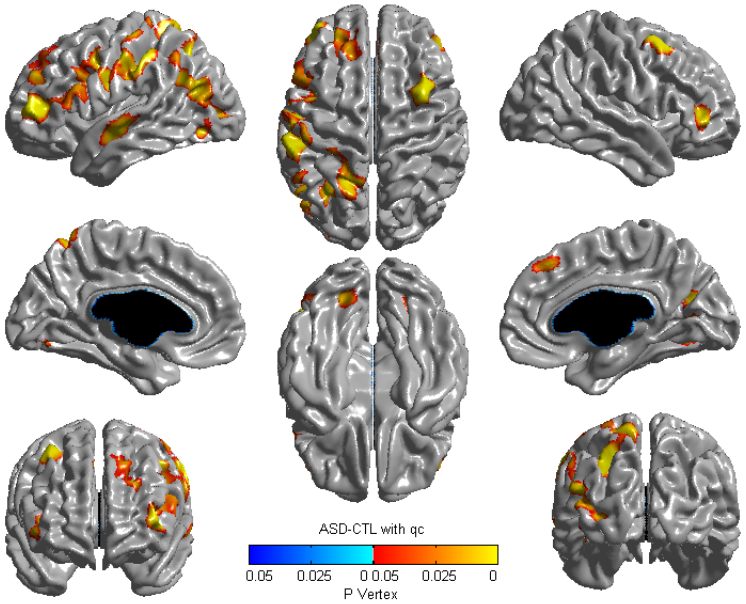
N ~ 1100

Data including subjects with motion



N ~ 400

Data with quality control



Stability of processing and strictness of quality control can meaningfully change resulting scientific claims (Khundrakpam et al., 2017)

Expected Contributions to knowledge

- Accessible and portable tool for reproducible experiments
- Method for evaluating stability and tool-dependence in neuroimaging
- Method for identifying the sources of instability within pipelines