Columbia University Medical Center

The Art and Pitfalls of fMRI pre-processing:
Introduction and simple theoretical considerations

Christian Habeck, Ray Razlighi

Purpose of the course

- Educate fMRI practitioners about
  - **Standard pre-processing tools and software (main focus)**
  - Unresolved issues (raise awareness, but leave solution for later)

- Ideally, enable practitioners to engage in methodological research themselves

Why is pre-processing necessary?

The raw fMRI signal is influenced by several factors that are statistical or subject-dependent in nature:

- Temporal offset between slice acquisitions
- Subject motion
- Subject respiration and heart rate
- Subject's brain size and shape
- Scanner noise and field inhomogeneity

etc.....

making analysis across time or subjects impossible without pre-processing.

Typical within-subject pre-processing pipeline

- **Correction for motion (scrubbing/ residualization)**
- **Temporal band pass filtering**
- **Co-registration to structural**
- **Re-alignment**
- **Slice timing correction**

Successive discrete processing stages transforming raw signal **r(t)** into pure signal **s(t)**

$$s(t) = F5 \circ F4 \circ F3 \circ F2 \circ F1 \circ r(t)$$

Alternatively: $F5(F4(F3(F2(F1(r)))))$

Obvious optimization quest

- What are the “best” functions/algorithms *F*?
  - What are the best parameter settings?

➔ This nuance is not very important: for functions that acquire an additional setting of a parameter the difference in the parameter setting can be interpreted as leading to a *different* function

➔However, this is enough; the *order* of processing stages needs to be optimized too

Why is optimization of complete pipeline, including order of stages, necessary?

Successive discrete processing stages $\mathbf{s(t) = F5 \circ F4 \circ F3 \circ F2 \circ F1 \circ r(t)}$

Th**is cannot be true since:**

- Different artifacts interact with one another (e.g. slice timing and motion)
- Artifacts are happening in parallel at all times

➔ *Unless all functions f(n) are linear (=ordering irrelevant), **the ordering has to be optimized too***

➔ *There is no apriori optimal ordering of the tools in the pipeline*

Further: pre-processing pipelines for task-based fMRI are no guide for resting fMRI

Outcome of interest for

Task fMRI: **task activation**, i.e. correlation of voxel activity ***x*** with task design ***d*** (1st order)

$$\mathbf{x' \, d}$$

Resting fMRI: functional connectivity, i.e. correlation of between voxel activities **x** and **y** (2nd order)

$$\mathbf{x' \, y}$$

Pre-processing pipelines for task-based vs. resting fMRI

Simple pseudo notation for variational derivative with respect to pre-processing functions/parameters *u* for finding zero gradients

Task-fMRI:

$$\delta_u (\mathbf{x}' \mathbf{d}) = (\delta_u \mathbf{x})' \mathbf{d}$$

Task design not dependent on pre-processing!

Resting fMRI – product rule:

$$\delta_u (\mathbf{x}'\mathbf{y}) = (\delta_u\mathbf{x})' \mathbf{y} + \mathbf{x}' (\delta_u\mathbf{y})$$

➔ Different orders (1st vs. 2nd) invalidate generalization of optimized task-based pipelines to resting fMRI

Recall of these simple theoretical insights

Form the previous considerations, it follows that -

- Breaking out the pipeline into separate modules is strictly speaking incorrect
- No optimal apriori ordering of modules can be derived; optimization has be made empirically
- Optimization for first-order signal moments (=task-related activation) might give different results from higher-order moments (=connectivity)

How to proceed?

Optimization should give answers to the questions:

- Which tools to use?
- With what parameters?
- In what order?

➔ Before an informed search of prohibitively large search space can be attempted, there is an additional question:

What quality criterion should inform the judgment of good pipeline performance?

Simulated data (=gold-standard knowledge exists for judging performance)

- Residual sum of squares of signal in simulated data (minimize)
- False positives, false negatives (minimize)

Real-world data

- Replication of activation between data folds (maximize)
- Replication of functional connectivity between data folds (maximize)
- Correlation between replication quality and primary-interest measures (minimize)
- Prediction of subject identity (or task) in one data fold, after training a model in the other data fold (maximize)

➔ The questions touch on apriori value judgments and inform, but are not informed by, the tools to be used

Some questions to leave you with – some of which might not be answered the end of today

- What is the best pipeline for removing artefacts?
  - The pipelines constitutes of procedures/algorithms arranged in a certain order, with certain parameters for each algorithm

- What are the metrics used for judging good performance?
  - The decision of what constitutes good performance might use metrics such R^2 in simulations, and false positives/negatives in real-world data.

- Should some standard tools be left out entirely since they cause more harm than help?

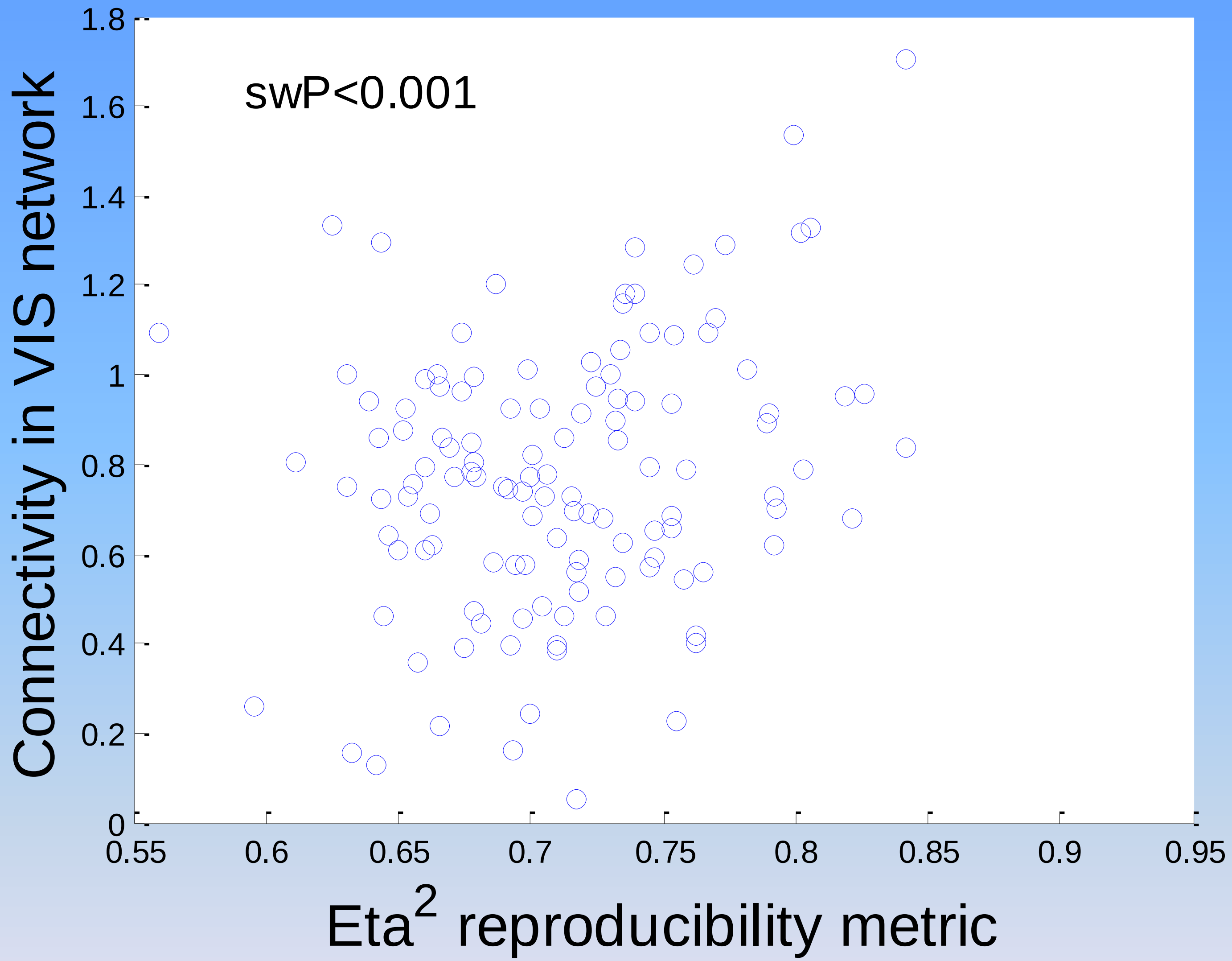
➔ Some of these questions have to be answered empirically, with simulations or real-world data, possibly on a case-by-case basis

One simple example of deficient pre-processing in resting BOLD

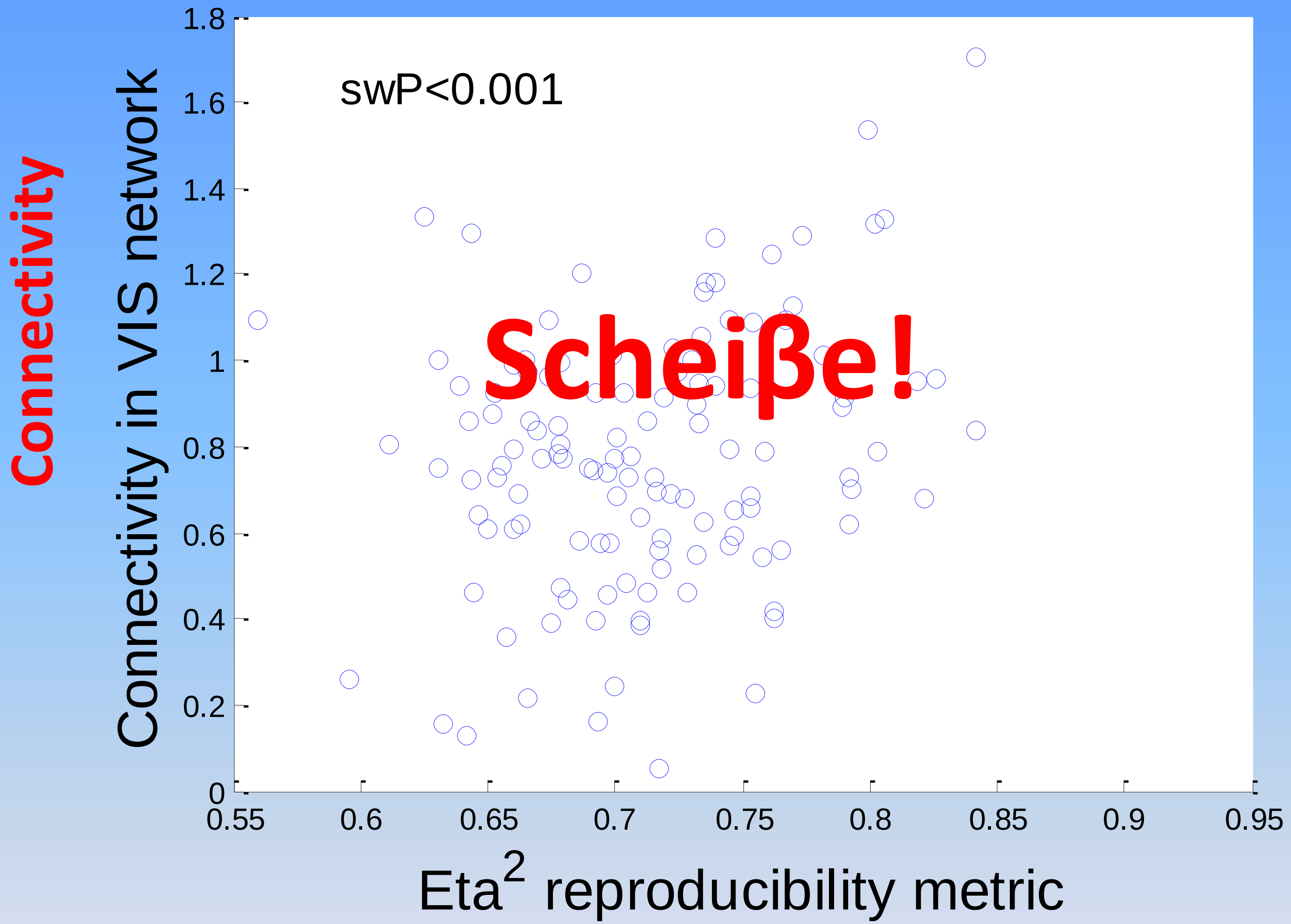
Look at influence of quality of replication of global connectivity between split halves on selected connectivity

- Replication of global connectivity between data folds captured with mean η^2
- Outcome of interest: connectivity within visual resting-state network

➔ Computation details are not very important here



➔ Data robustness influences correlation strength, measurement bias



Replication between data folds

Enjoy the course!

Take it away speakers!