Large Scale Analysis and Open Science

Robert Oostenveld

MEG/EEG toolkit 2018

Overview

Scaling up from pilot analysis to publication quality group analysis

Handling of data, scripts and results

Open Science

- BIDS for organizing your data
- Repositories for sharing your data
- Version control and publication of your analyses details

Legal issues and privacy of your subjects

Practical issues of data sharing

Single subject versus group analysis

https://humanconnectome.org/study/hcp-young-adult

https://github.com/Washington-University/megconnectome

Frontiers in Neuroscience - From raw MEG/EEG to publication: how to perform MEG/EEG group analysis with free academic software

https://github.com/robertoostenveld/Wakeman-and-Henson-2015

Small or large data
Small or large computers

Note: “big data” is complex data, “large data” is large in size but not per see complex

Managing the development of your pipeline

Start with version control

```
> git init
```

Write the pipeline for a single subject

```
> git commit
```

Manage subject differences

```
> git commit
```

Run for all subjects

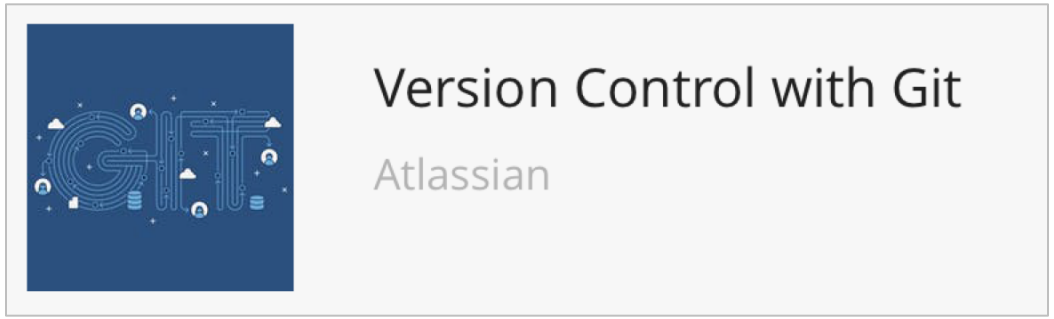
```
> git commit
```

Do group analysis

```
> git commit
```

Share your pipeline along with the paper and data

```
> git push
```



https://www.coursera.org/learn/version-control-with-git

https://software-carpentry.org/lessons/

Why *manage* research data?

Improve efficiency and quality of research

Researchers can use shared data to jump-start new projects

Research findings can be re-visited upon new insights

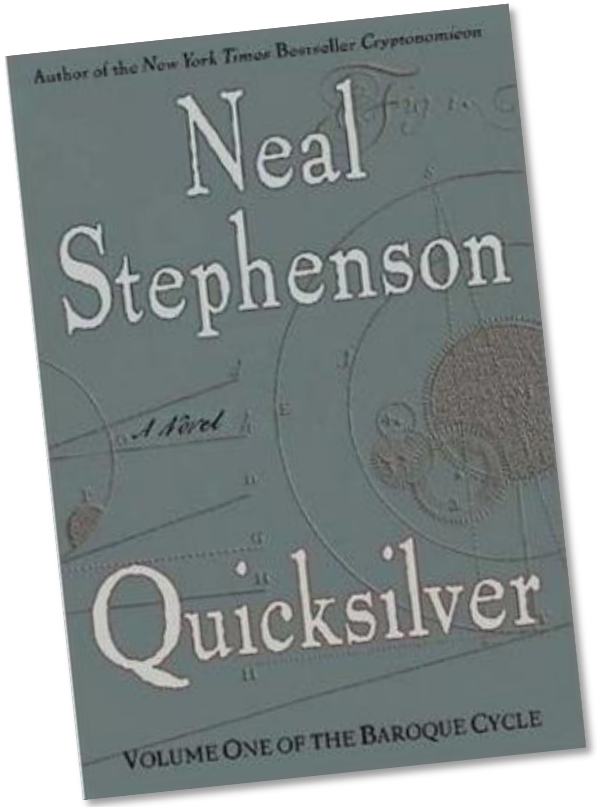
Why *share* data?

Publishers require it

Funders require it

It is just the “right thing to do”

Royal Society



Our origins lie in a 1660 'invisible college' of natural philosophers and physicians. Today we are the UK's national science academy and a Fellowship of some 1,600 of the world's most eminent scientists.

Nullius in verba

The very first 'learned society' meeting on 28 November 1660 followed a lecture at Gresham College by Christopher Wren. Joined by other leading polymaths including Robert Boyle and John Wilkins, the group soon

Advancements and adventure

The early years of the Society saw revolutionary advancements in the conduct and communication of science. Hooke's Micrographia and the first issue of *Philosophical Transactions* were published in 1665 alone. *Philosophical Transactions*, which established the important concepts of scientific priority and peer review, is now the oldest continuously-published science journal in the world.

We published Isaac Newton's Principia Mathematica, and Benjamin Franklin's kite experiment demonstrating the electrical nature of lightning. We backed James Cook's journey to Tahiti, reaching Australia and New Zealand, to track the Transit of Venus. We published the first report in English of inoculation against disease, approved Charles Babbage's Difference Engine, documented the eruption of Krakatoa and published Chadwick's detection of the neutron that would lead to the unleashing of the atom.

https://royalsociety.org/about-us/history/

Open Science

Open educational resources

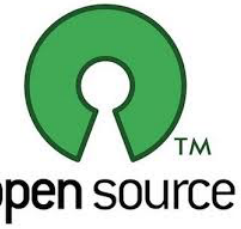
Open access publications

- Open peer review

Open methodology

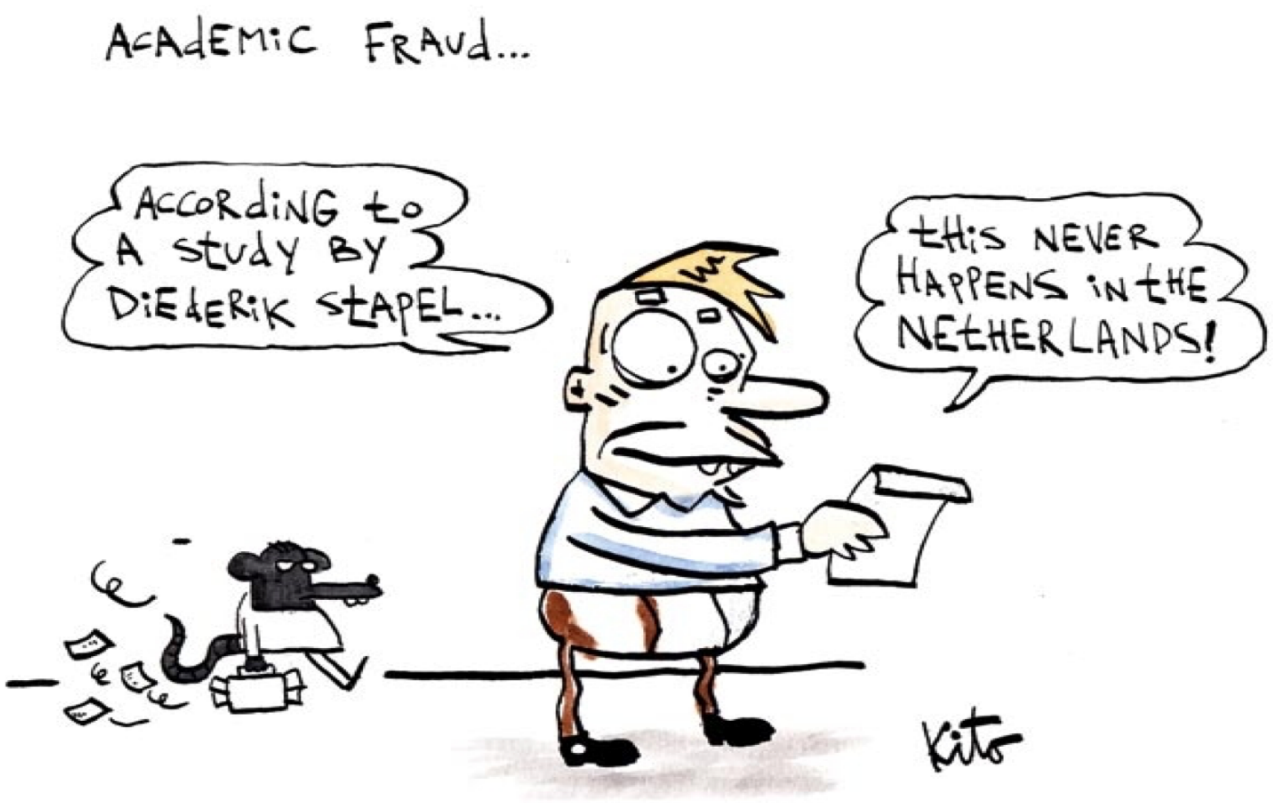
- Open source
- Open hardware

Open data



Markus Neuschäfer; https://www.slideshare.net/mneuschaefer/1504-open-knowledgefolien

Improve public trust in science



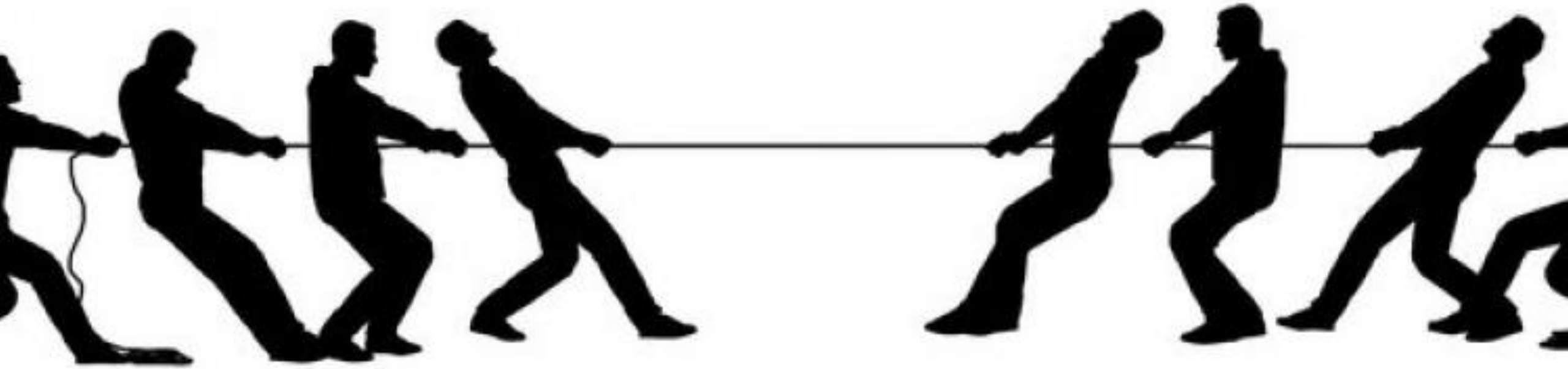
http://harrieverbon.blogspot.nl/2012/11/diederik-stapel-werd-ook-betaald-door.html

Open Definition

"Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**"

http://opendefinition.org

Open data versus privacy



Personal data

name

address

date of birth

phone number

license plate

IP address

...

Crime Scene Investigation
http://www.abc.net.au/news/2017-09-19/csi/8960590

(Biometric) data

facial details

dental record

fingerprint

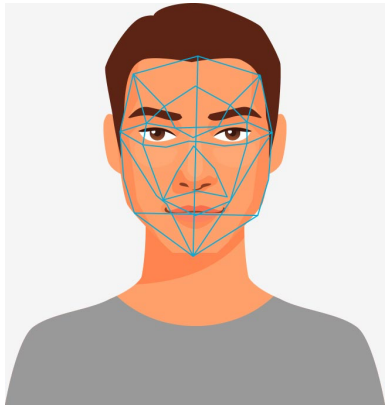
genetics

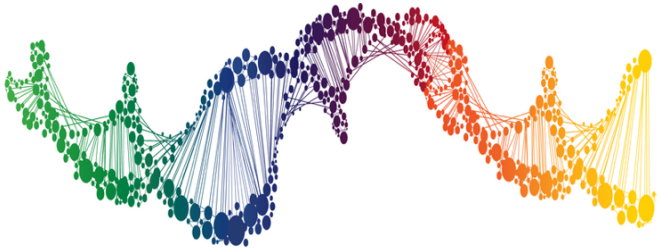
cortical folding pattern

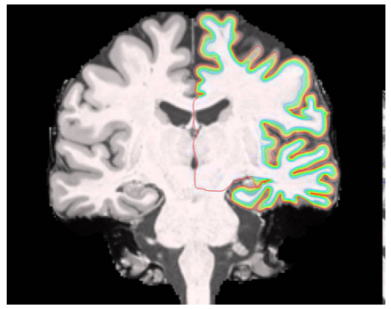
clinical data

gait/movement pattern

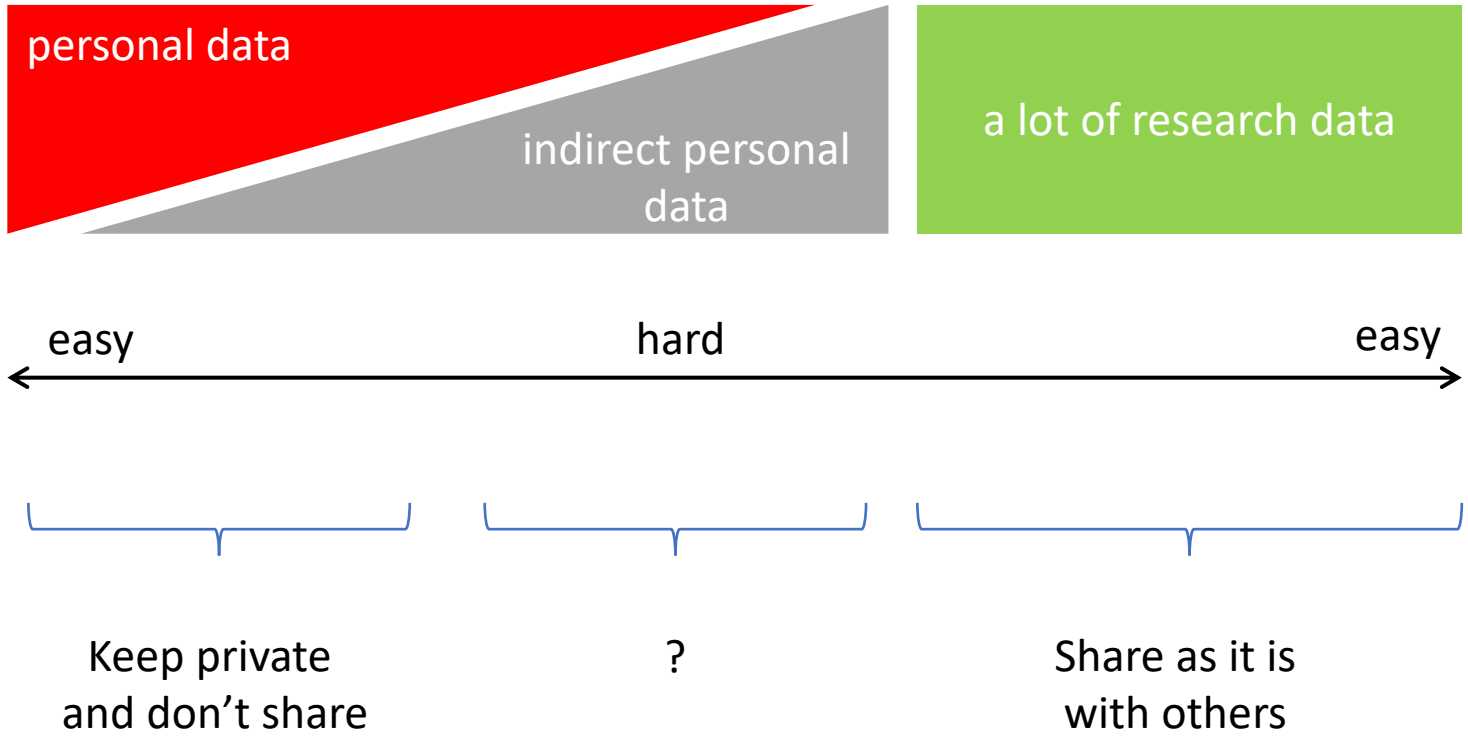
responses on questionnaires







Gradient between personal and research data



Limit possible identification

Personal data

restrict access to personal data
protect the key that maps between the pseudonym and the identity

Biometric data

data minimization only acquire, store and share data that is needed
acquire ***anonymous*** data
acquire data using a ***pseudonym***
use ***de-identification*** techniques

Legal constraints

collaboration: access only for specific authorized researchers
sharing: access for everyone but only following data use agreement

Limit possible identification

Anonymous

You never knew the subjects identity to start with

Pseudonymization

Use a code instead of the subjects name

De-identification

Remove (indirectly) identifying features

Blur the indirect personal data

- Deface anatomical MRI
- Age at the time of acquisition instead of date of birth
- Use age bins instead of years
- Questionnaire outcomes rather than individual item scores
- …

Appropriate blurring depends on the situation

… for example blurring the age of the subject

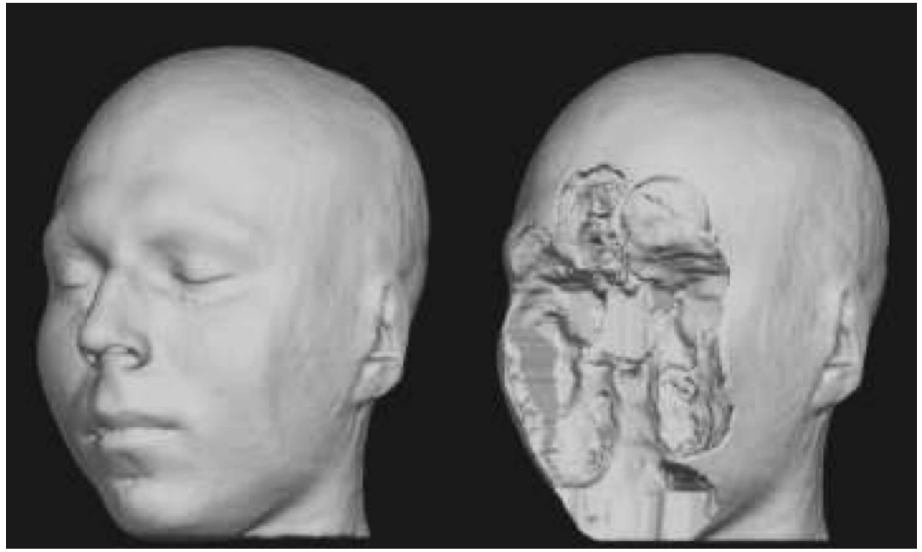
1 month bins

5 or 10 year bins

Personal and research data

Personal and research data

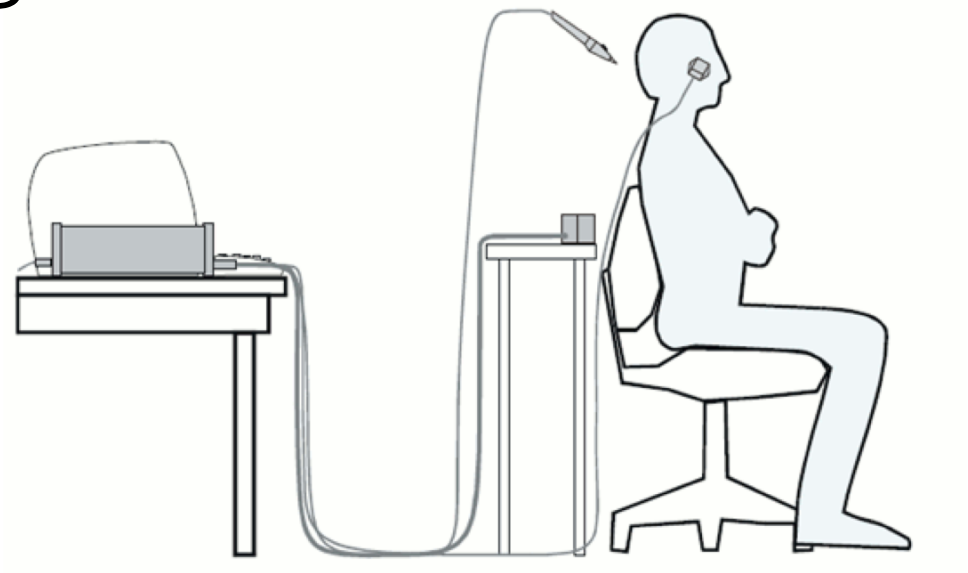
Sharing deidentified imaging data



facial details have been removed, e.g. using ft_defacevolume or ft_defacemesh.

nasion is missing, the outline of the nose is missing, sometimes also the ears are missing.

Coregistration between MEG/EEG and anatomy

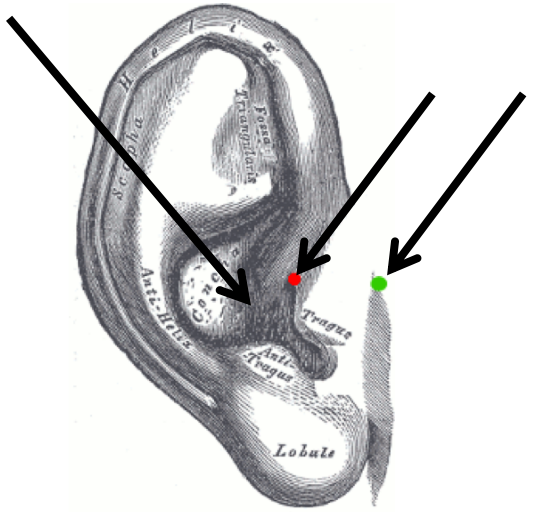


1) anatomical landmarks (lpa, rpa, nas)
2) HPI/HCL coil locations
3) scalp surface points

Lab specific conventions for landmarks and markers

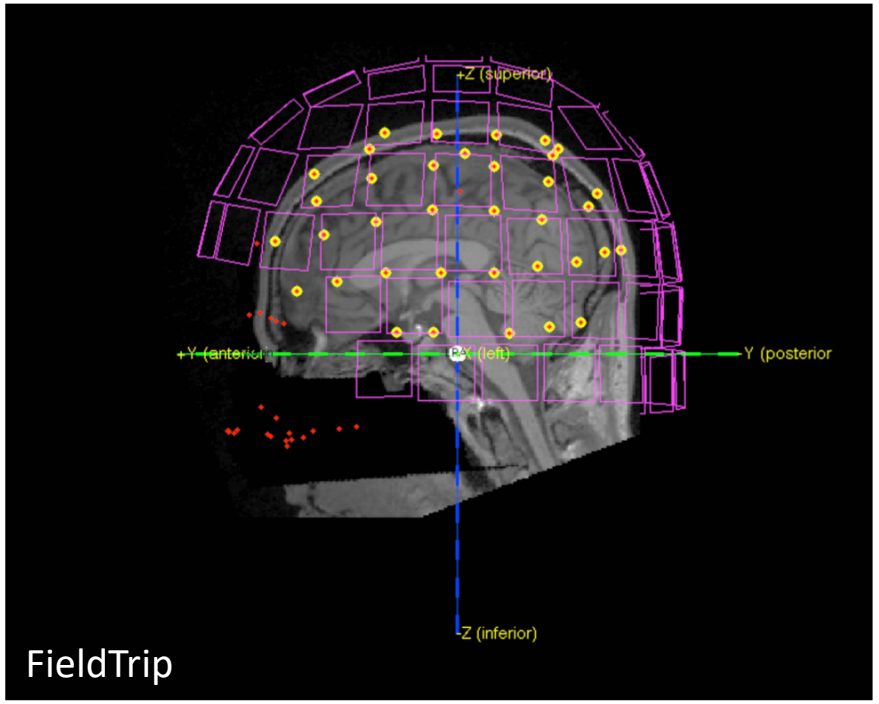
Landmarks:
anatomically recognizable points on a head

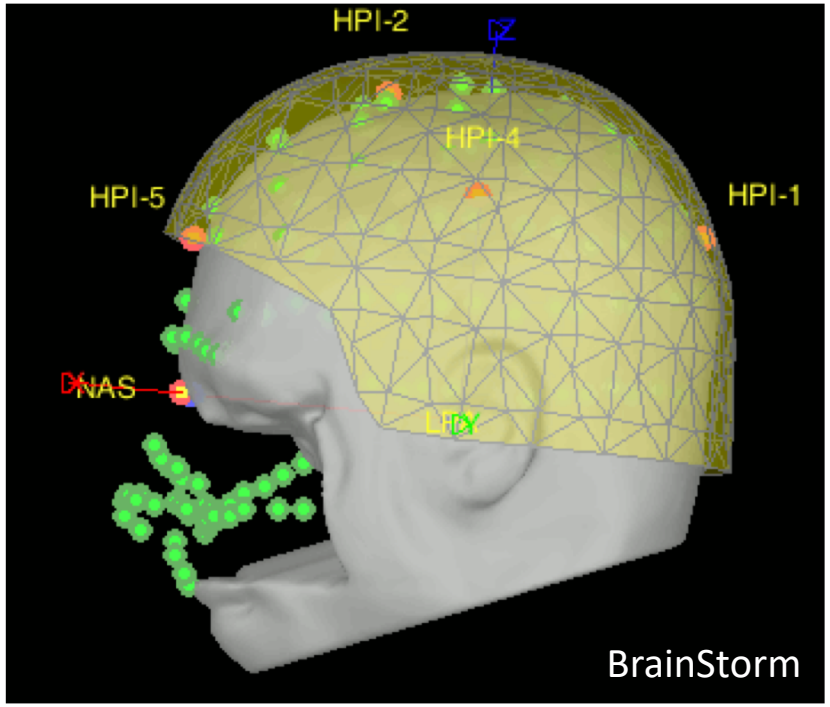
Markers (or fiducials):
points that are visible in multiple modalities, e.g. HPI coils or vitamin E capsules



Coregistration

Redo the coregistration using limited data, or trust the coregistration that was provided.





Repositories for data sharing

Institutional Repository

Donders Repository

Generic repositories (note the DUA)

Zenodo, Harvard DataVerse, DataDryad, …

Specific repositories

Genetics, astromomy, openfmri, …

Re3data - repository of data repositories

Narcis – scholarly information (and data) in NL

Elsevier - datasearch