Get to know your data preprocessing, segmentation and artifacts

Practical MEEG 3rd Edition

Robert Oostenveld

Donders Institute, Radboud University, Nijmegen, NL
Karolinska Institutet, Stockholm, SE

robert.oostenveld@donders.ru.nl

shared slides

Radboud University

DONDERS INSTITUTE

Get to know your data
artifacts, preprocessing, segmentation

Practical MEEG 3rd Edition

Robert Oostenveld

Donders Institute, Radboud University, Nijmegen, NL
Karolinska Institutet, Stockholm, SE

robert.oostenveld@donders.ru.nl

shared slides

Karolinska Institutet

NatMEG

Radbound University

DONDERS INSTITUTE

Get to know your data
artifacts, segmentation, preprocessing

Practical MEEG 3rd Edition

Robert Oostenveld

Donders Institute, Radboud University, Nijmegen, NL
Karolinska Institutet, Stockholm, SE

robert.oostenveld@donders.ru.nl

shared slides

Karolinska Institutet

NatMEG

Radbound University

DONDERS INSTITUTE

Get to know your data

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Practical MEEG 3rd Edition

Robert Oostenveld

Donders Institute, Radboud University, Nijmegen, NL
Karolinska Institutet, Stockholm, SE

robert.oostenveld@donders.ru.nl

shared slides

Karolinska Institutet

NatMEG

Get to know your data – learning goals

What are the characteristics of the data that we use in the analysis?

How to organize your raw data?

Quality assessment and control

What are the artifacts and why are they relevant?

Preprocessing and segmenting (or vice versa)

Selective averaging to get ERPs/ERFs

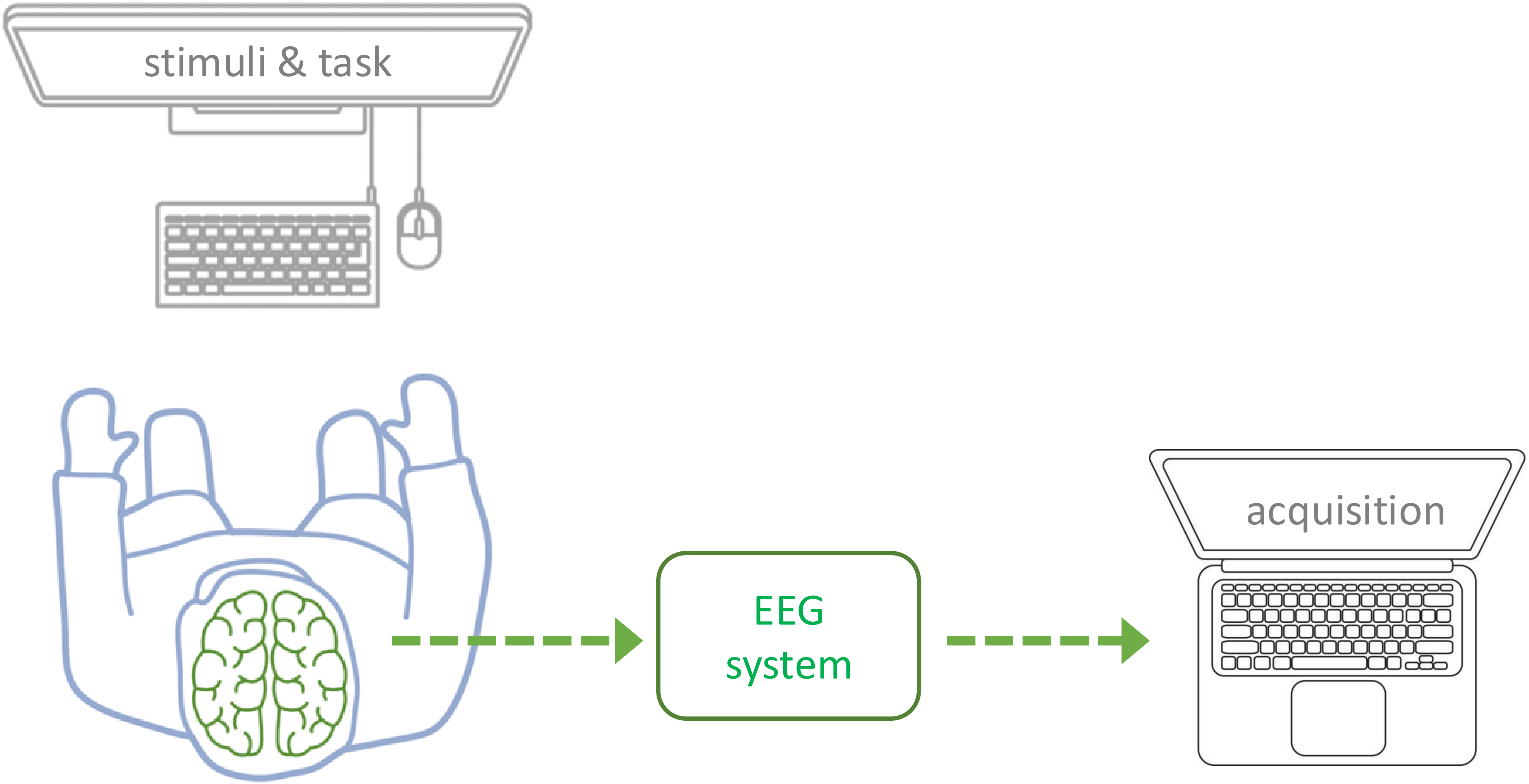
shared slides

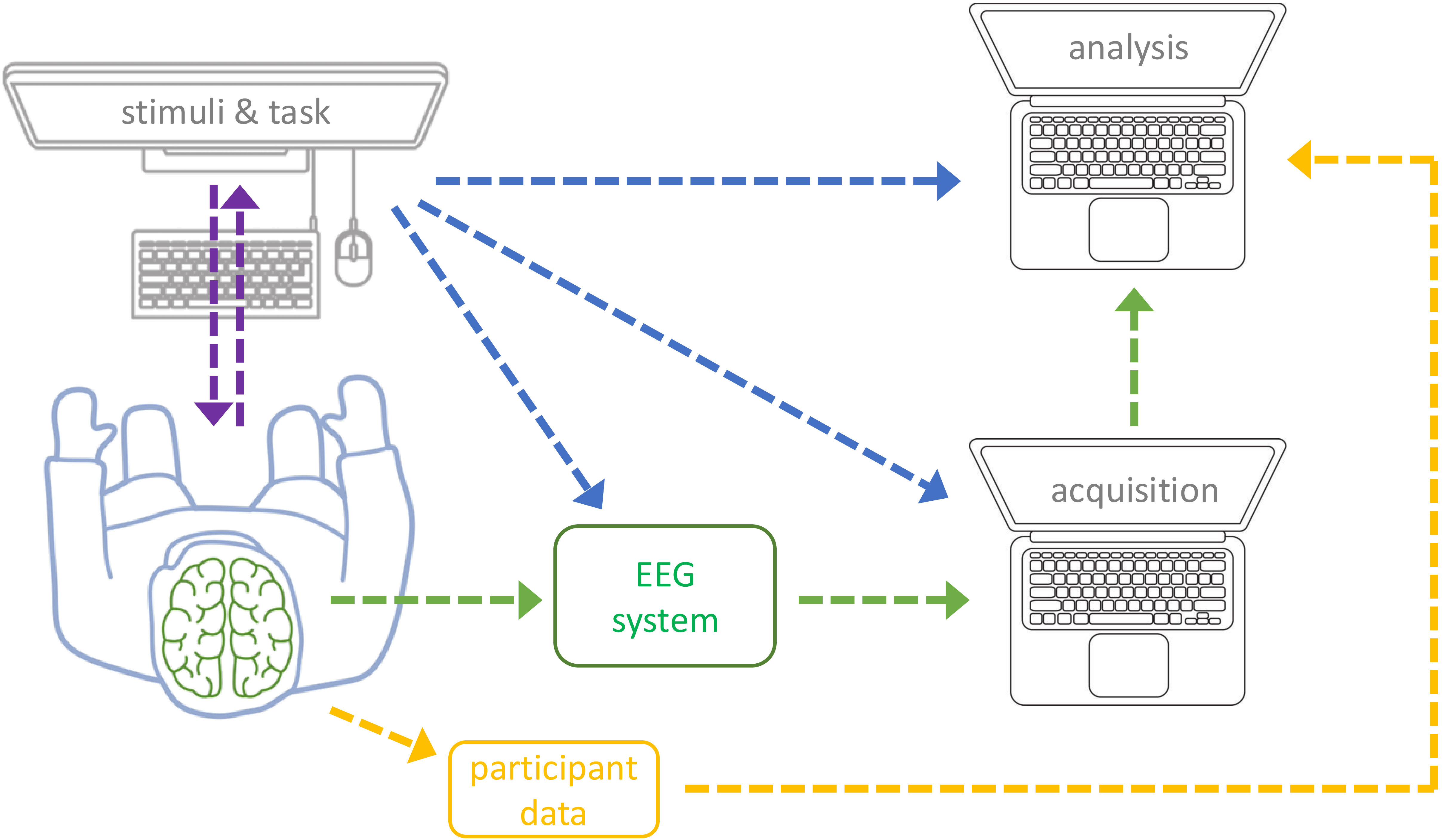
Collecting your data – in the lab or not

1. You have designed your own study, recruited your own participants, and collected your own data in the lab.

2. You have received data from a (former) colleague in the lab, or downloaded it from an online repository.

Either way: ***organize it!***


stimuli & task
EEG
system
acquisition


stimuli & task
analysis
acquisition
EEG
system
participant
data

Preprocessing, processing, analysis

Prior to preprocessing

Data curation: collecting all files, naming them consistently, etc.

EEG/MEG data is ***large, consists of many files, and is complex***

EEG data characteristics

64 channels, 500 Hz, 1 hour is approx. 500 MB

Typical study ~30 subjects, 15 GB of raw data

Many EEG companies, hence many file formats.

Analysis often done on laptops.

MEG data characteristics
SQUID-based systems

275 or 306 channels, 1000 Hz, 1 hour is approx. 4 GB

Typical study ~30 subjects, 120 GB of raw data

Few MEG companies, hence small number of formats

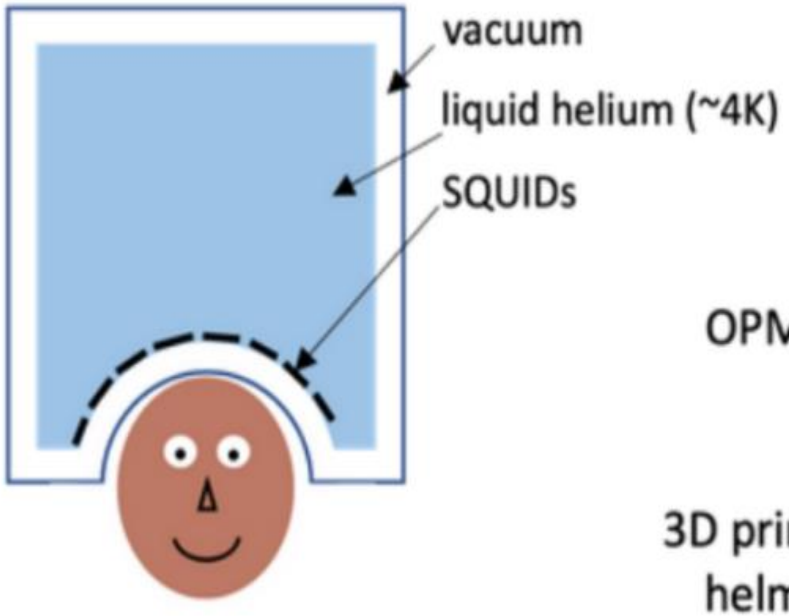
- Neuromag/Elekta/MEGIN: one recording is one *.fif file
- CTF: one recording is one *.ds directory with ~10 files

Analysis often done on “large” computers.

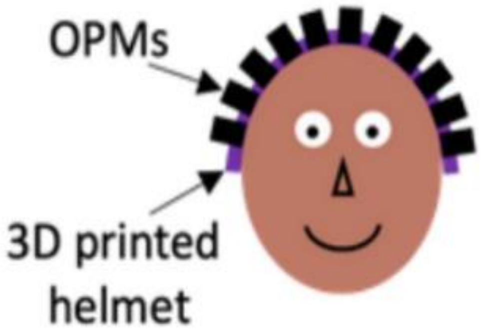
MEG data characteristics

OPM-based systems

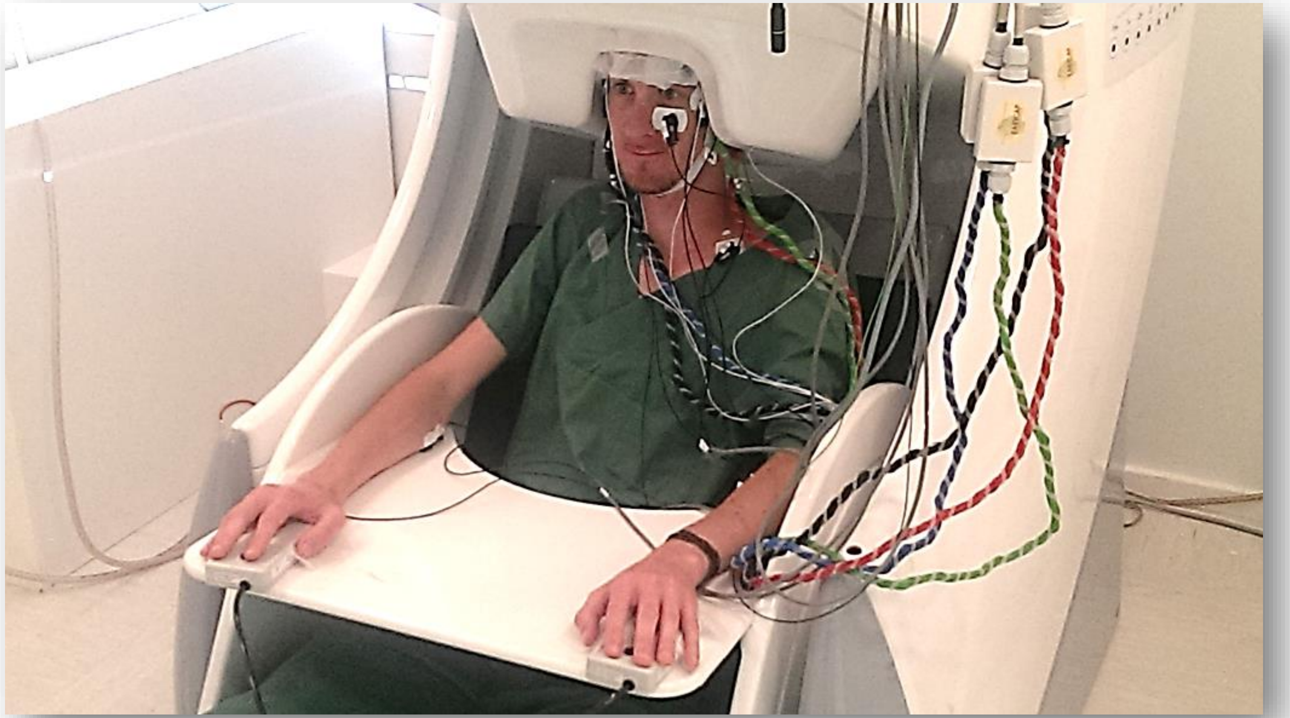
SQUIDs



OPMs

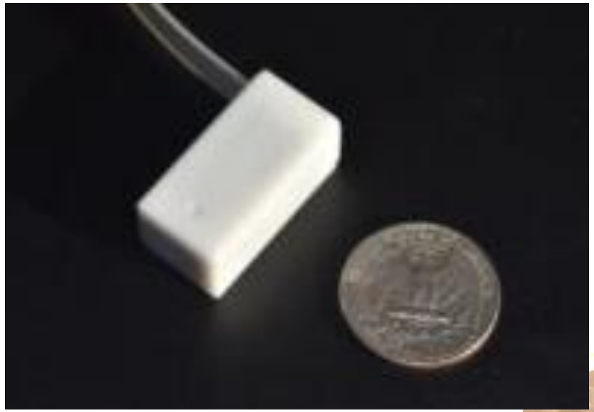


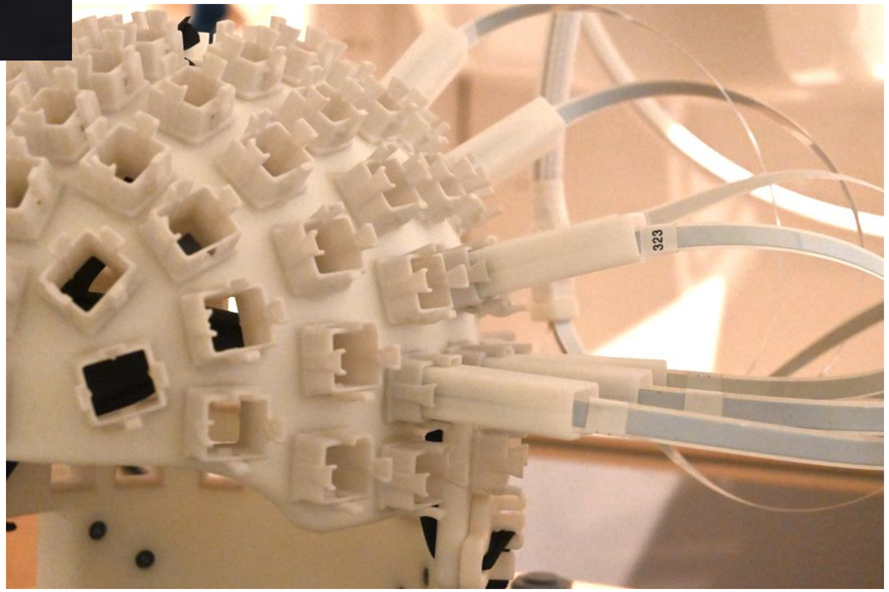
SQUIDs versus OPMs



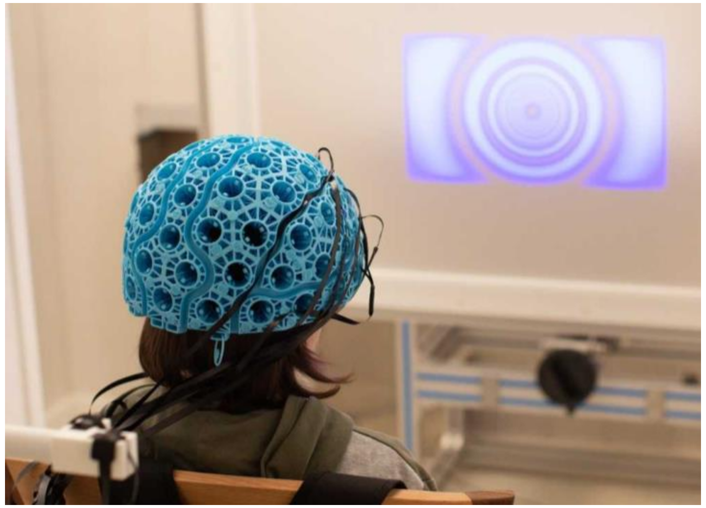
FieldLine OPM system at Karolinska/Stockholm

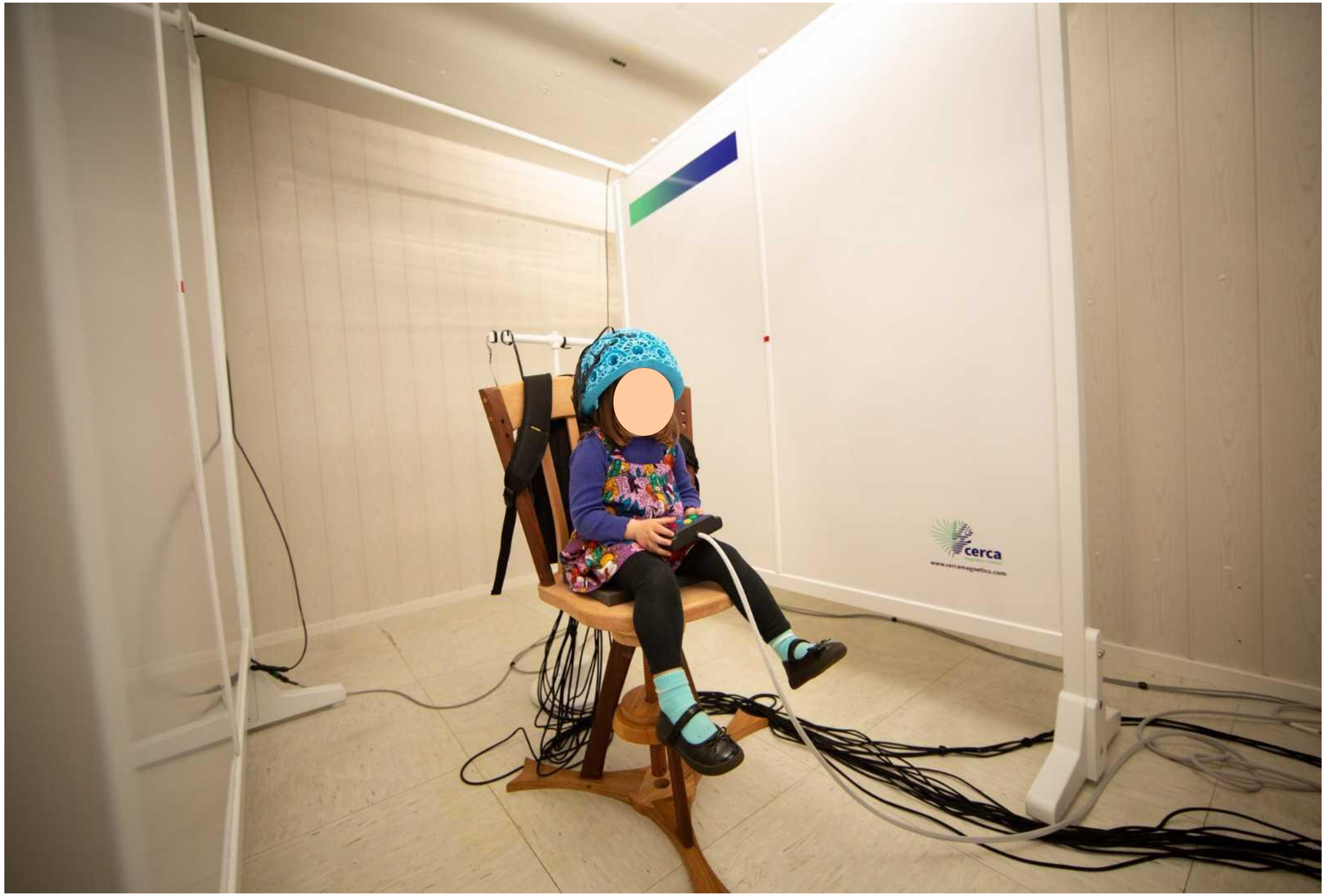
now upgraded to 128 sensors (256 channels) and a smart helmet





CercaMagnetics OPM system installed in Cardiff





MEG data characteristics
OPM-based systems

A single OPM sensor can have 1-3 channels (x, y, z)

Total system has anywhere between 1 to 384 channels

Typical study ~30 subjects, 120 GB of raw data

OPM sensors can be placed in a flexible cap, like EEG, or in a 3D printed helmet

Position of the sensors relative the cap/helmet and to the head.

3D scans of the head and sensors, Polhemus digitizer, etc.

Auxiliary data

Anatomical MRI data

- Directly from the scanner as ~200 DICOM files (*.ima, *.dcm) — ~ 50 MB
- Commonly converted to NIfTI format, one file (*.nii or *.nii.gz)

Behavioural data (time-resolved)

- Mostly encoded as “triggers” together with the MEG or EEG data stream — kB to GB
- Stimulus presentation log file
- Video and/or audio recordings (e.g., for verbal responses)
- Eye tracker for gaze and pupil diameter
- …

Other data (usually tabular, not time-resolved) — ~ 1 kB

- Handedness, gender, age, …
- Questionare outcomes

Organize your data FAIR

Findable

Make your data available in a catalog or repository
with a persistent identifier (DOI, handle) and metadata

Accessible

Be explicit about data usage terms (agreement with downloader)

Interoperable

Make your data human and machine readable, e.g. BIDS

Reusable

Make sure you document enough details, e.g. “data descriptor” paper
that can be cited, along with citing our data -> measurable impact!

Wilkinson et al. (2016) The FAIR Guiding Principles. *Sci Data.*

Organize your data FAIR

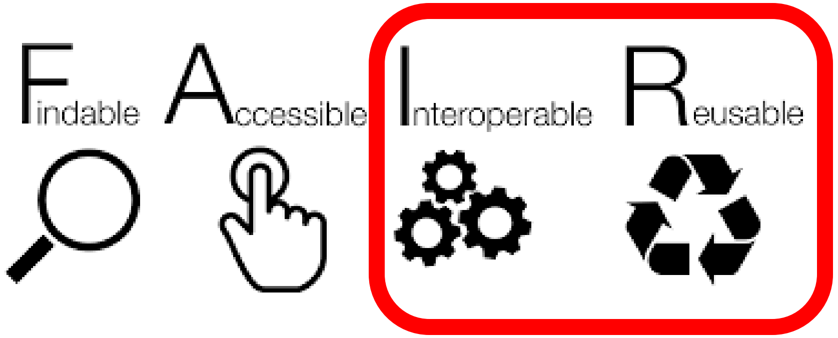
BIDS is a community initiative to make your data more FAIR

BIDS is a way to organize your existing raw data

- To improve consistent and complete documentation
- To facilitate re-use by your future self and others

BIDS is *not*

- A new file format
- A search engine
- A data sharing platform



http://www.bids-standard.org/

BIDS for EEG and MEG

also for iEEG, MRI, NIRS, PET, motion capture, …

Just a bunch of directories and files on disk.

No special software required (although tools are available).

BIDS for EEG and MEG

also for iEEG, MRI, NIRS, PET, motion capture, …

Directory structure

Metadata

Actual EEG data

```
data/README
     CHANGES
     dataset_description.json
     participants.tsv
    /sub-01/anat/…
    /sub-01/meg/…
    /sub-01/eeg/sub-01_task-auditory_eeg.edf
    /sub-01/eeg/sub-01_task-auditory_eeg.json
    /sub-01/eeg/sub-01_task-auditory_channels.tsv
    /sub-01/eeg/sub-01_task-auditory_events.tsv
    /sub-01/eeg/sub-01_electrodes.tsv
    /sub-01/eeg/sub-01_coordsystem.json
```

BIDS “sidecar” files for metadata

see also https://github.com/bids-standard/bids-examples

BIDS
BRAIN IMAGING DATA STRUCTURE

1) represent otherwise missing data
2) make it easier to query/search

As example for EEG and MEG:

_participants.tsv and json
_sessions.tsv and json
_scans.tsv and json

_meg.json
_events.tsv and json
_channels.tsv and json
_electrodes.tsv and json
_coordsystem.json
_photos.jpg


```
"TaskName": "facerecognition",
"SamplingFrequency": 1100,
"Manufacturer": "ElektaNeuromag",
"ManufacturersModelName": "ElektaVectorview",
"MEGChannelCount": 306,
-10% system",
d stimuli on a screen
```


onset	duration	onset_sample	stim_type	trigger	stim_file
23.93	0	26323	Unfamiliar	13	meg/u101.bmp
27.1709	0	29888	Unfamiliar	14	meg/u101.bmp
30.3782	0	33416	Famous	5	meg/f043.bmp
33.4355	0	36779	Famous	5	meg/f046.bmp
36.6091	0	40270	Unfamiliar	13	meg/u061.bmp
39.85	0	43835	Unfamiliar	14	meg/u061.bmp
43.0073	0	47308	Famous	5	meg/f050.bmp
46.1318	0	50745	Famous	6	meg/f050.bmp
49.3055	0	54236	Scrambled	17	meg/s082.bmp
52.3455	0	57580	Famous	5	meg/f147.bmp
55.67	0	61237	Famous	6	meg/f147.bmp
58.7273	0	64600	Famous	7	meg/f043.bmp
62.0682	0	68275	Famous	5	meg/f130.bmp
65.2591	0	71785	Famous	6	meg/f130.bmp
68.3836	0	75222	Famous	7	meg/f046.bmp
71.5909	0	78750	Unfamiliar	13	meg/u106.bmp
74.8309	0	82314	Unfamiliar	13	meg/u140.bmp
78.0545	0	85860	Unfamiliar	14	meg/u140.bmp
81.2118	0	89333	Scrambled	17	meg/s020.bmp
84.4527	0	92898	Scrambled	18	meg/s020.bmp
87.6936	0	96463	Scrambled	19	meg/s082.bmp
90.8682	0	99955	Famous	5	meg/f066.bmp
94.0582	0	103464	Unfamiliar	13	meg/u091.bmp

Get to know your data – learning goals

What are the characteristics of the data that we use in the analysis?

How to organize your raw data?

Quality assessment and control

What are the artifacts and why are they relevant?

Preprocessing and segmenting (or vice versa)

Selective averaging to get ERPs/ERFs

Preprocessing, processing, analysis

Prior to preprocessing
- Data curation: collecting all files, naming them consistently, etc.

First processing steps do not depend so much on the research question
- Quality assessment
- Artifact removal
- Filtering, baseline correction
- Aligning stimulus presentation and behavioral data with EEG/MEG
- Segmenting/epoching
- Aligning MRI with EEG/MEG sensors and anatomical processing

Later steps are more tightly linked to the research question
- Averaging ERPs in specific conditions
- Computing power spectra, time-frequency analysis, connectivity
- Source reconstruction
- Modelling (e.g., using GLM)
- Statistical inference

Preprocessing, processing, analysis

Prior to preprocessing

Data curation: collecting all files, naming them consistently, etc.

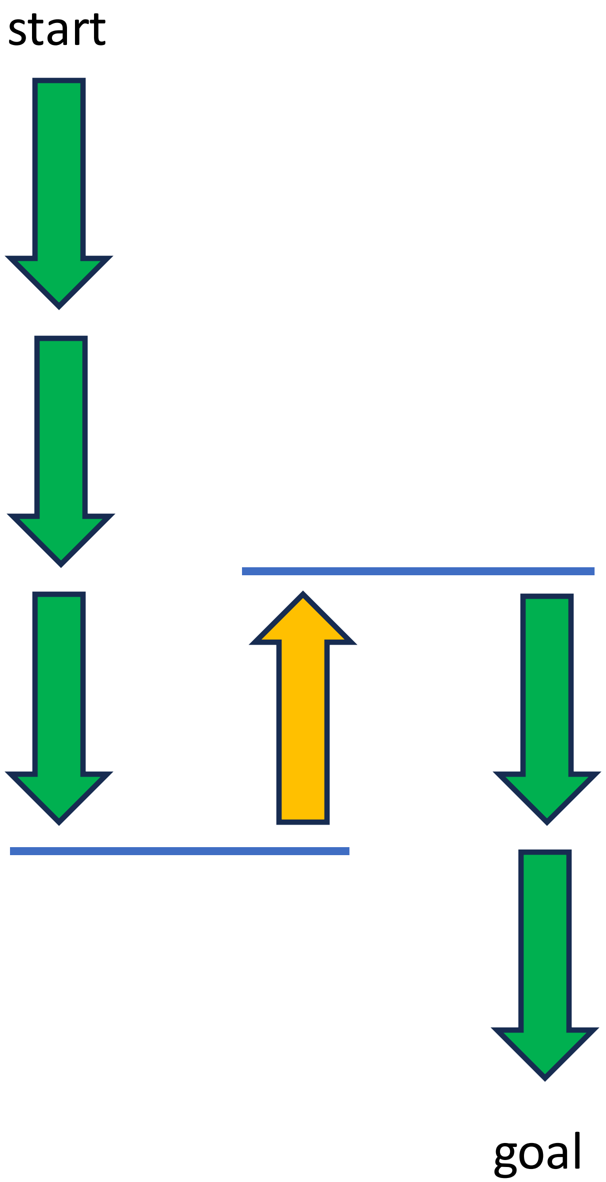
First processing steps do not depend so much on the research question

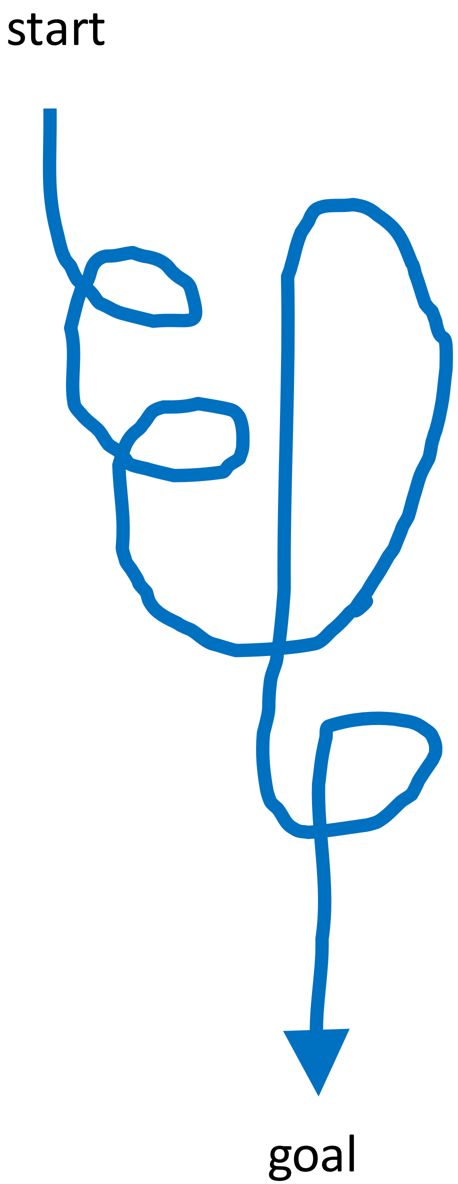
Quality assessment
Artifact removal
Filtering, baseline correction
Aligning stimulus presentation and behavioral data with EEG/MEG
Segmenting/epoching
Aligning MRI with EEG/MEG sensors and anatomical processing

Later steps are more tightly linked to the research question

Averaging ERPs in specific conditions
Computing power spectra, time-frequency analysis, connectivity
Source reconstruction
Modelling (e.g., using GLM)
Statistical inference

You should plan for multiple iterations of the preprocessing





Quality control and artifacts

EEG electrodes attached to subject’s head

- Bad attachment -> bad signals

MEG is not directly attached to subject

- Few bad channels (dependent on hardware tuning)

EEG artifacts

- Anything that causes potential differences

MEG artifacts

- Anything that causes magnetic fields

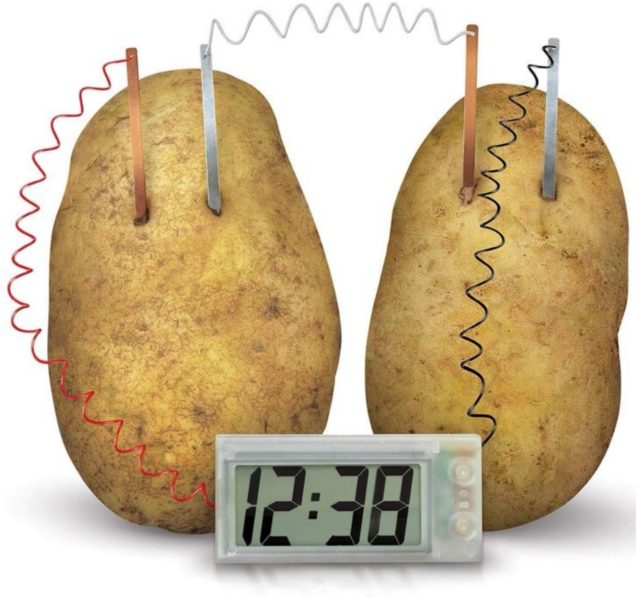
EEG artifacts

~1.000.000 Volt

1,5 Volt

1-10 micro Volt

EEG artifacts



Poor contact with the scalp

Electrochemical noise (sweating)

Electrostatic noise (e.g., rubbing feet over the carpet)

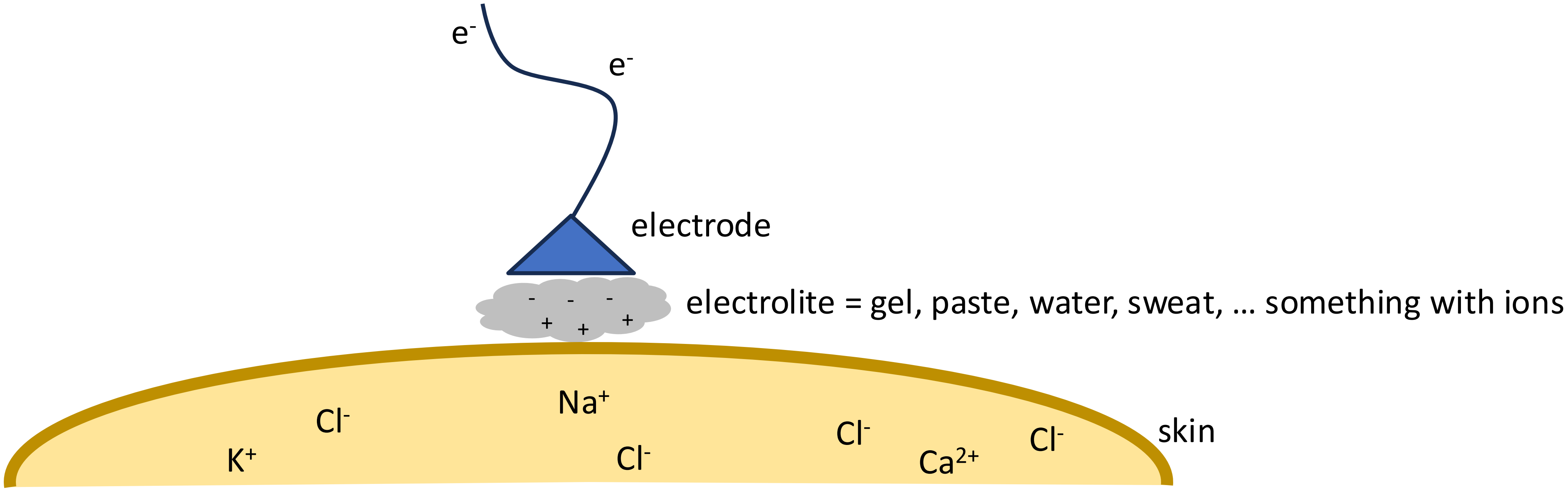
- Mostly common-mode, i.e., similar on all channels

Power line noise, 50Hz electrical equipment

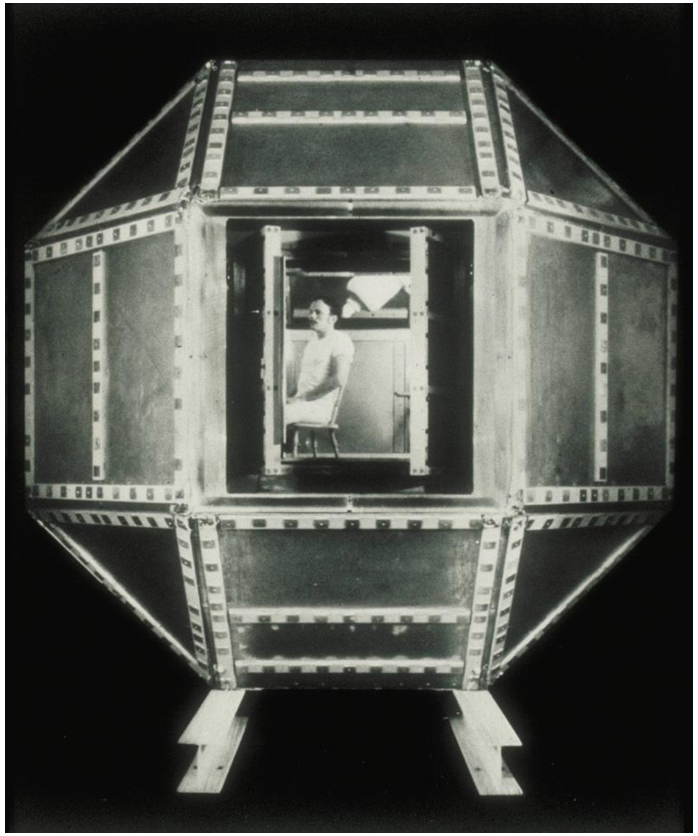
Other types of (physiological) bioelectricity

- Muscle (EMG)
- Heart (ECG)
- Eye movements (EOG)

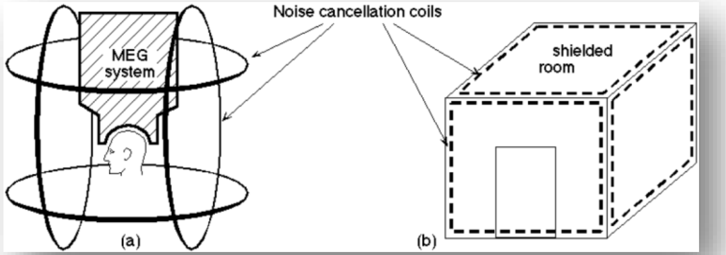
EEG electrode movement



MEG artifacts (and shielding)



magnetically shielded room (MSR)
built by David Cohen at MIT in 1969



MRI magnet

3 T

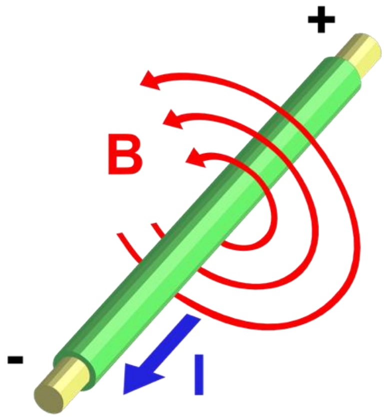
Earth field

10^{-5} T

Human brain

10^{-12} T

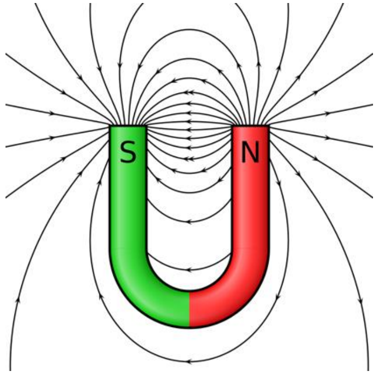
Common MEG artifacts



Power line noise, 50Hz equipment

Large metal objects moving outside the MSR

- Car, trolly, elevator, the fan of airconditioning

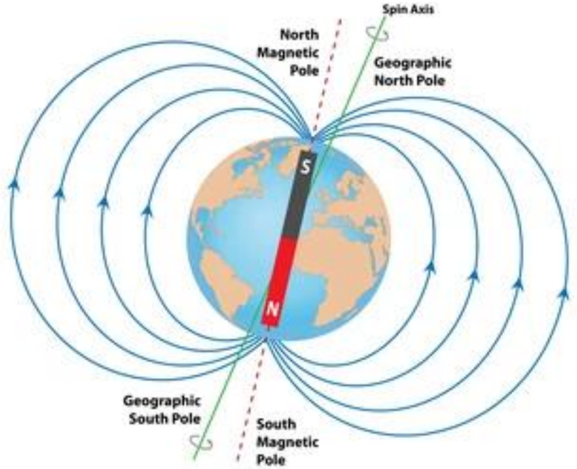


Residual field of the earth

- Building vibrations cause movements of the MSR walls and dewar

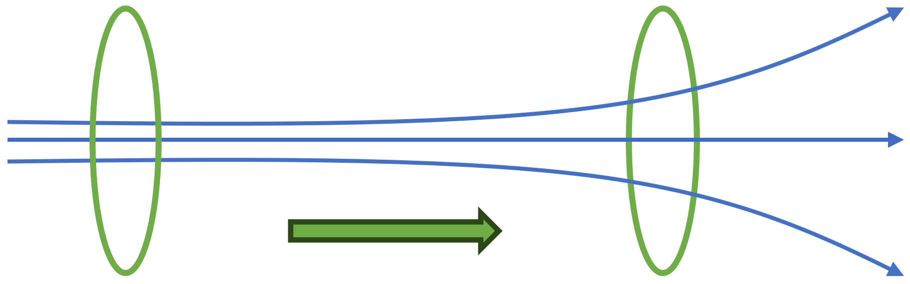
Other types of (physiological) bioelectricity

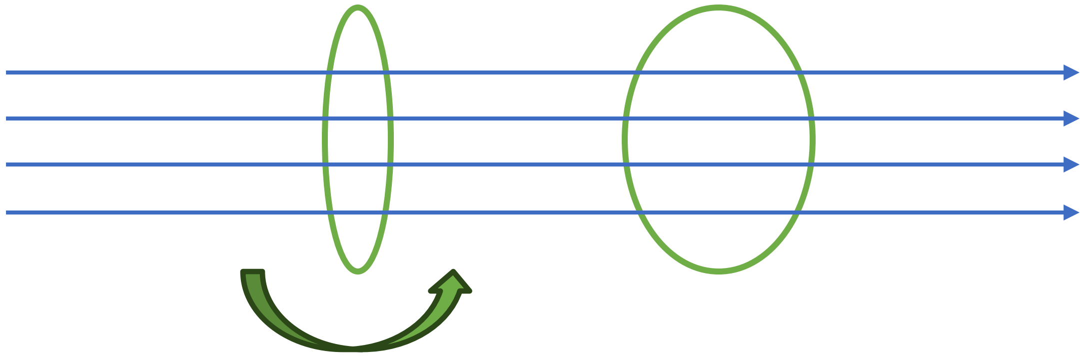
- Muscle (EMG)
- Heart (ECG)
- Eye movements (EOG)

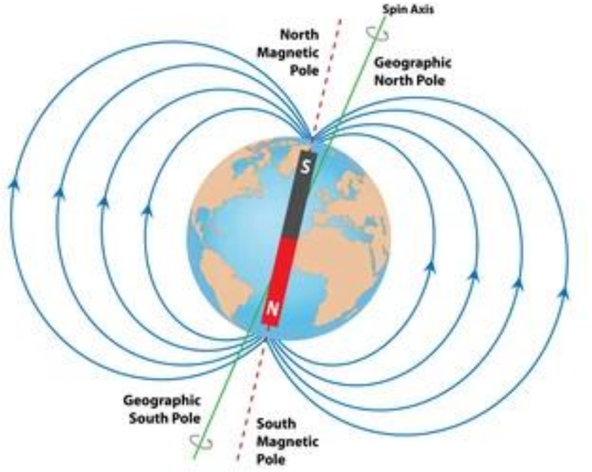


Movements in mobile MEG with OPMs

moving the sensor in the residual gradient or rotating the sensor in the residual field







EEG/MEG artifact removal

Identify and remove bad channels (or interpolate)

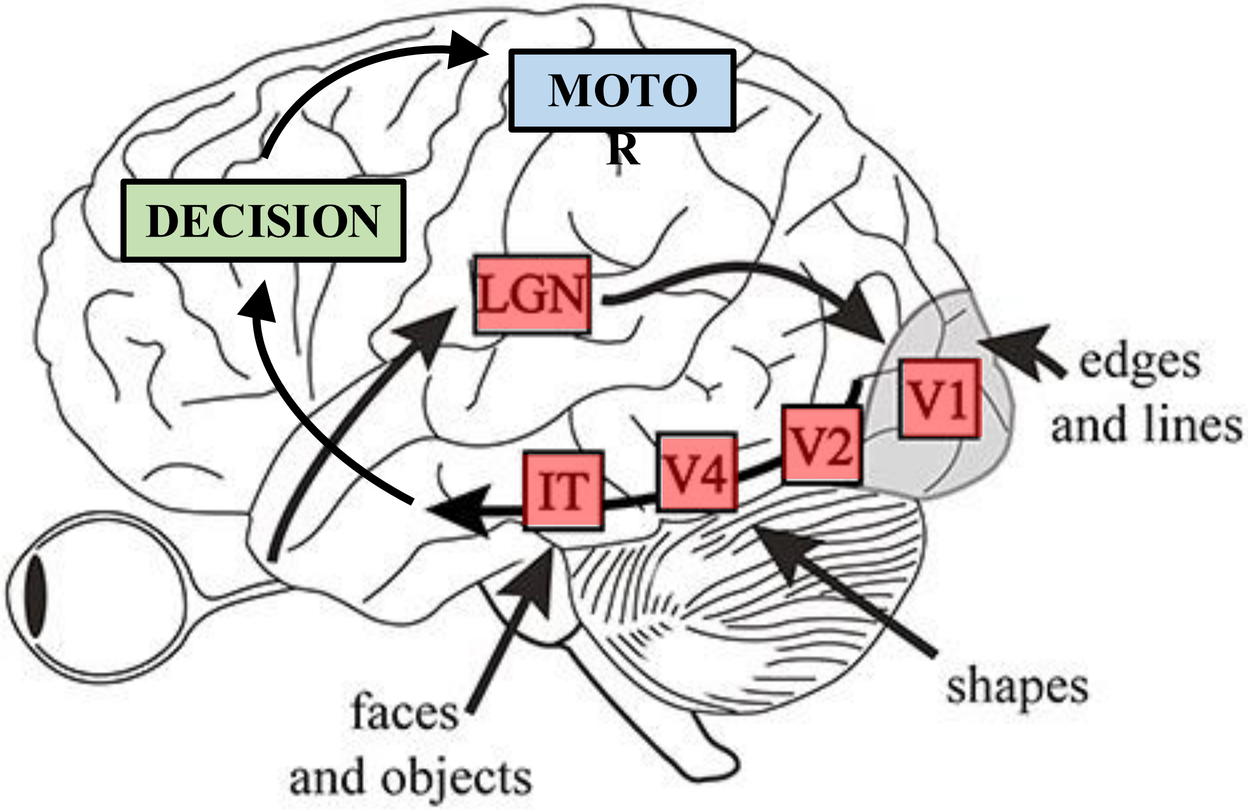
Identify and remove bad segments

- Continuous data
- Segmented data, only the pieces of interest

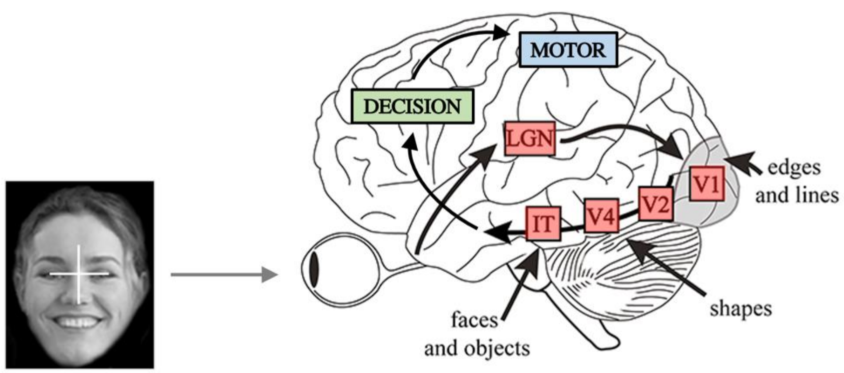
Identify trials in which the ***behavior was incorrect***
or in which the data ***cannot be recovered****.*

The brain is a hierarchical functional network

both sequential and parallel processing, ff and fb



EEG/MEG to study perception, cognition and behavior



Our experimental task and behavioral readouts ensure that we are tapping in to the desired cognitive processes.

Infant EEG, baby looking away -> they did not see the stimulus

Participant blinks at the stimulus -> they did not see the stimulus

No response in stimulus-response task -> the stimulus was probably processed differently

Participant responds too slow -> a different cognitive process was interfering

The experimental task often involves attention monitoring, includes catch trials, or an extra condition with responses, these **behavioral responses (or artifacts)** need to be analyzed.

EEG/MEG data cleaning

*Only **after** rejecting data corresponding to bad behavior and broken data, we proceed to clean the remainder.*

The data is a spatio-temporal mixing of different sources.

Spatio-temporal models can separate brain and noise sources.

For EEG: ICA, PCA, IClabel, ASR, MARA, GEDAI

For MEG: 3^{rd} order gradients, SSS/tSSS (Maxfilter), SSP, HFC, AMM, ICA

These are based on **data-driven** or **biophysical** models of the spatial distribution of the brain activity or the noise.

Get to know your data – learning goals

What are the characteristics of the data that we use in the analysis?

How to organize your raw data?

Quality assessment and control

What are the artifacts and why are they relevant?

Preprocessing and segmenting (or vice versa)

Selective averaging to get ERPs/ERFs

Analyzing the brain activity in an event-related task

Signal-to-Noise-Ratio (SNR) is not sufficient to directly observe the brain responses

Stimulus or task is repeated many times (i.e. trials)

- For example: one trial every 4 seconds, ~900 trials in one hour
- Experimental manipulation is usually a subtle difference between trials
- EEG/MEG response of interest is only about 1 second around the stimulus
- So 1 hour recording results in only ~900 seconds of useable data

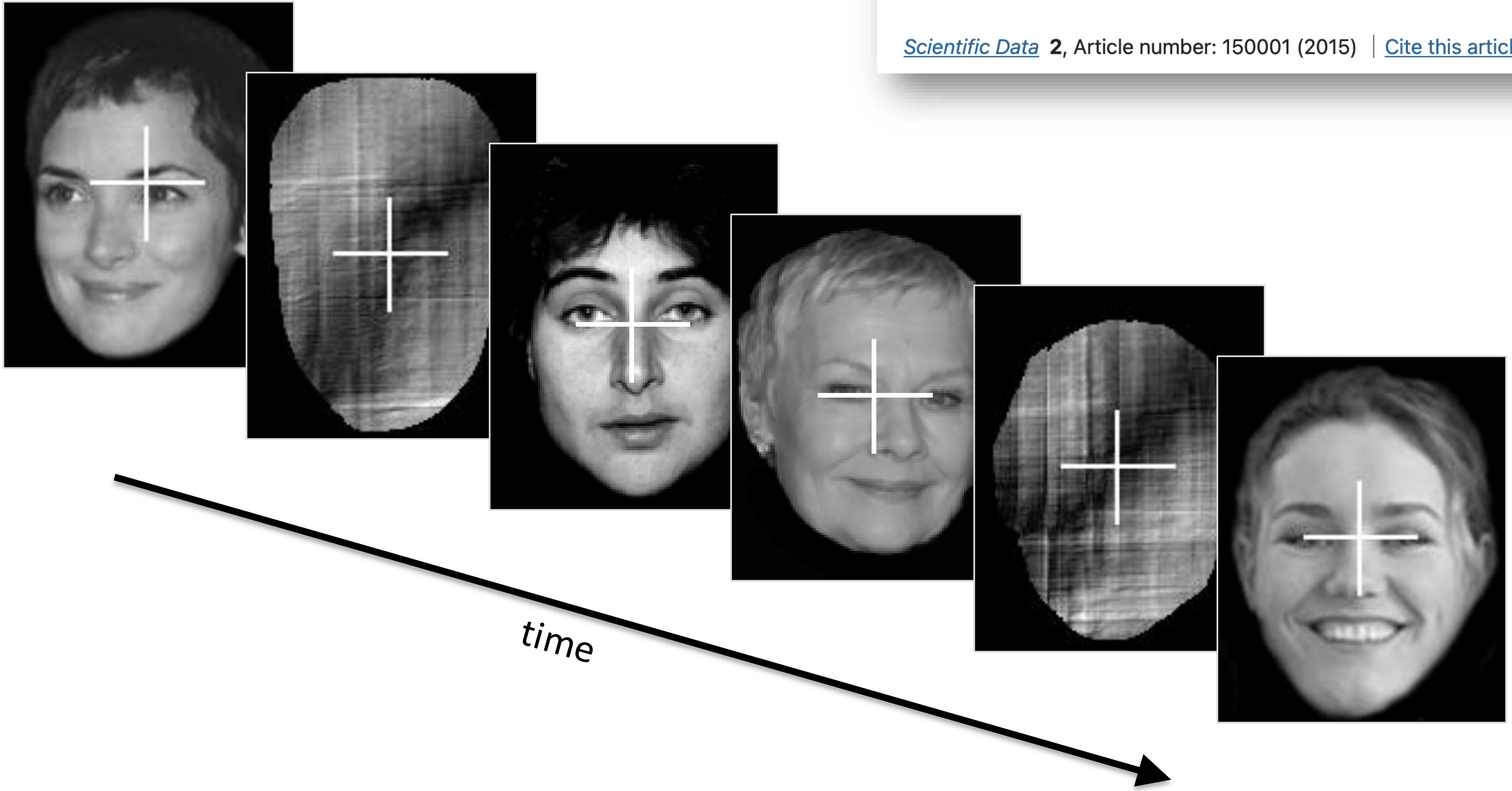
Analyzing the brain activity in an event-related task

Data Descriptor | Open access | Published: 20 January 2015

A multi-subject, multi-modal human neuroimaging dataset

Daniel G Wakeman & Richard N Henson

Scientific Data **2**, Article number: 150001 (2015) | Cite this article



http://doi.org/10.1038/sdata.2015.1

Analyzing the brain activity in an event-related task

900 trials

- 900 times a picture of "something"

900 trials

- 600 faces
  - 300 unfamiliar
  - 300 familiar (i.e., celebrities)
- 300 scrambled faces

900 trials

- 150 unfamilar faces 1st time, 150 unfamilar faces 2nd time
- 150 familar faces 1st time, 150 familar faces 2nd time
- 150 scrambled faces 1st, 150 scrambled 2nd time

	1st presentation	*2nd presentation*
unfamiliar	**150**	**150**
familiar	**150**	**150**
scrambled	**150**	**150**

Also other dimensions: gender, emotional expression, gaze direction, …

http://doi.org/10.1038/sdata.2015.1

Analyzing the brain activity in an event-related task

Multi-factorial design:

stimulus × presentation × gender × emotional expression × gaze direction × …

Analyzing the brain activity in an event-related task

	1st	2nd
unfamiliar	150	150
familiar	150	150
scrambled	150	150

faces vs scrambled

150	150
150	150
150	150

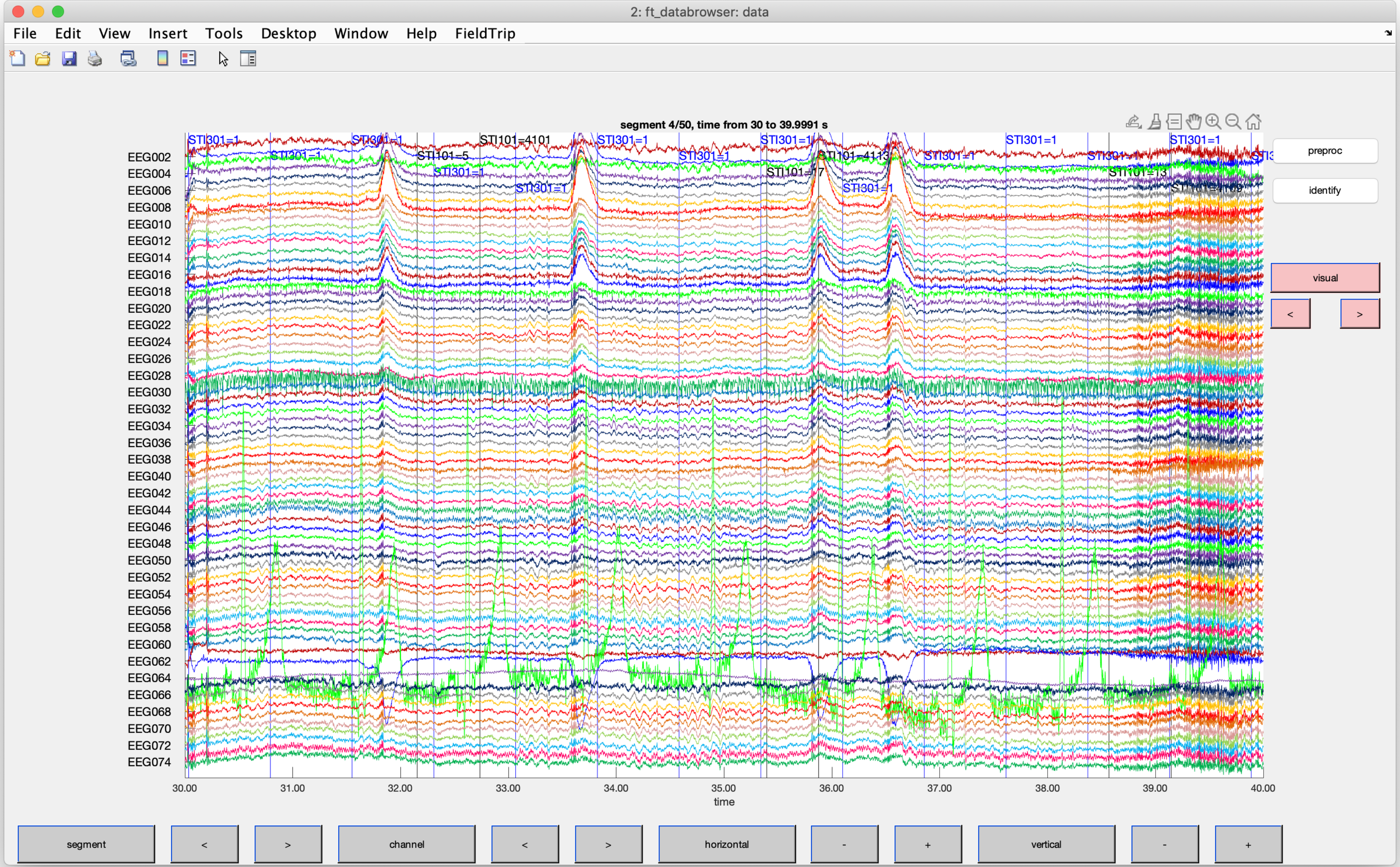
unfamiliar vs familiar

150	150
150	150
150	150

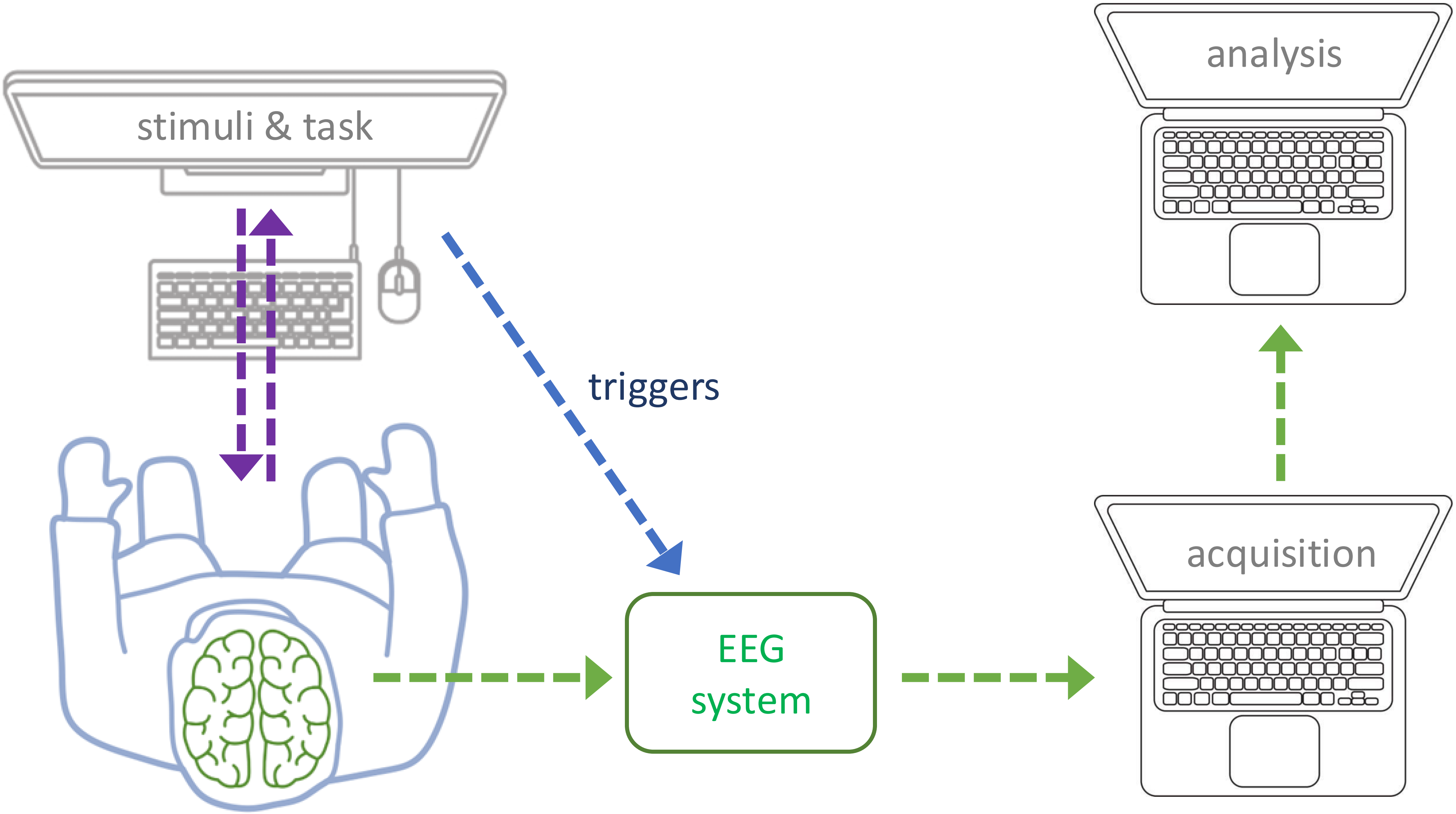
1st vs 2nd presentation

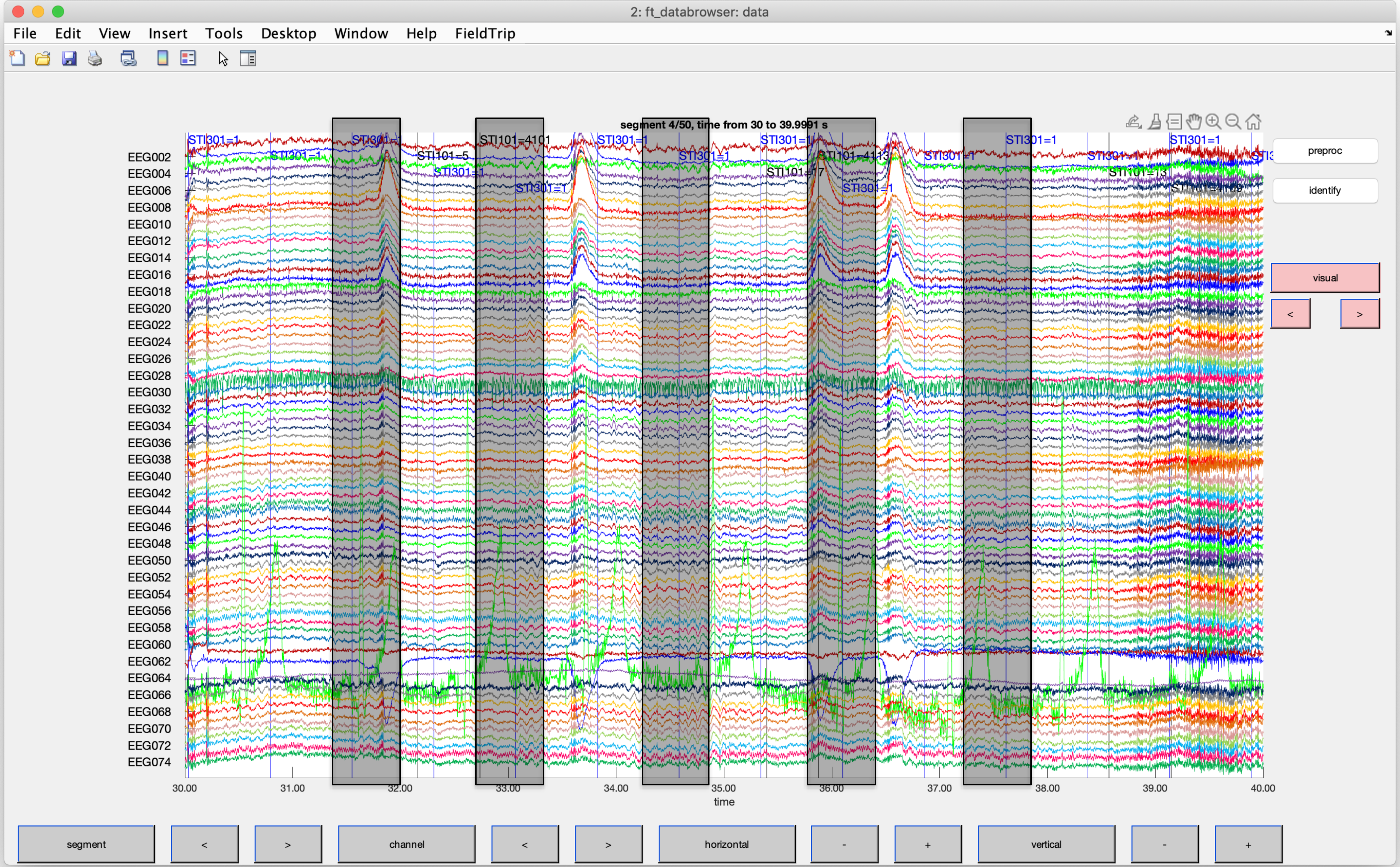
150	150
150	150
150	150

interaction (ANOVA)

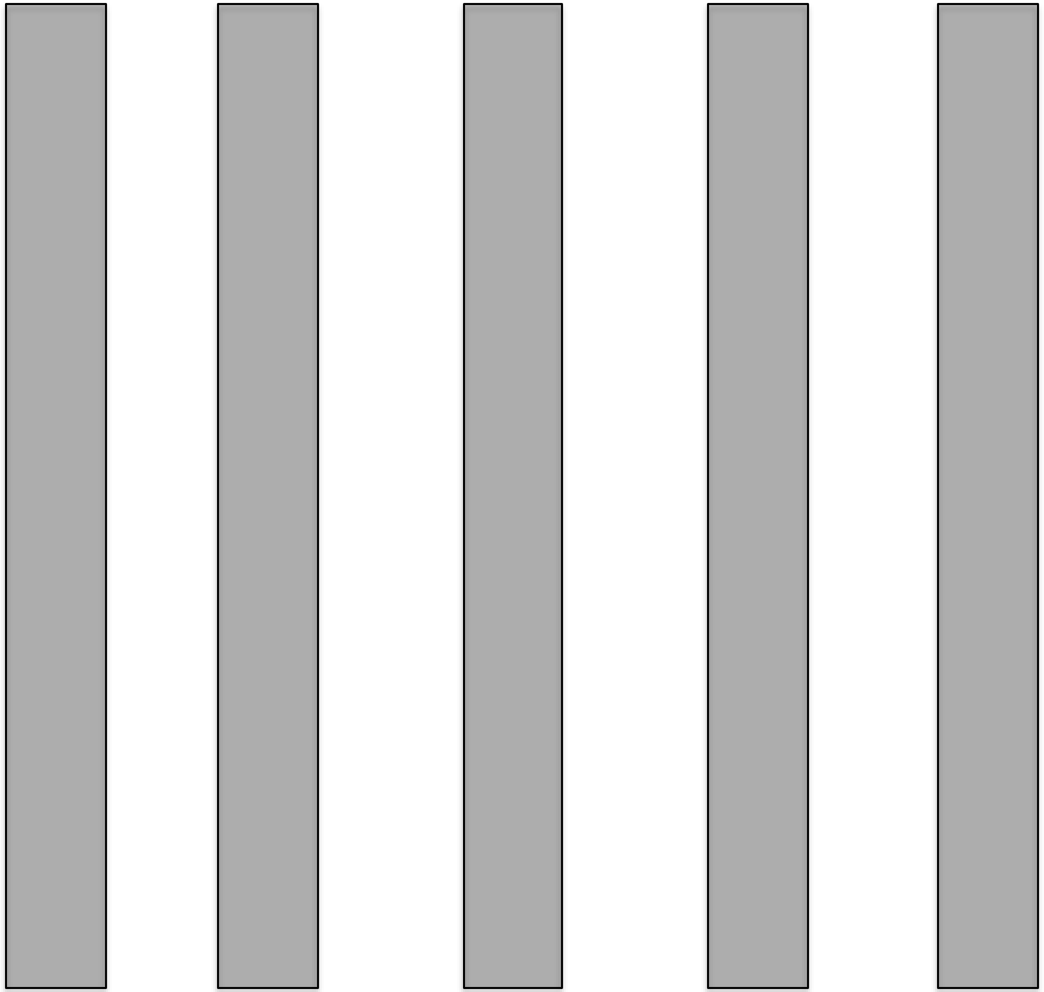


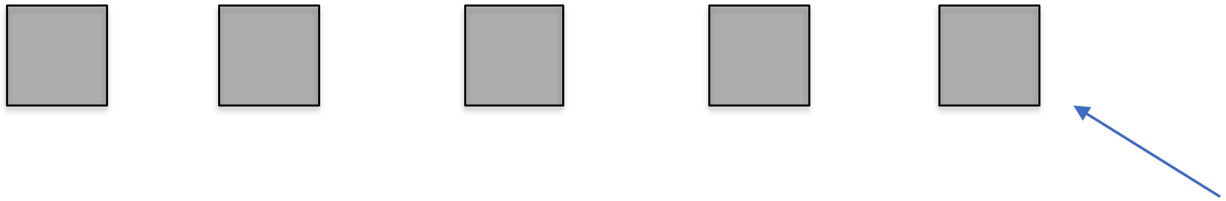
2: ft_databrowser: data
File
Edit
View
Insert
Tools
Desktop
Window
Help
FieldTrip
segment 4/50, time from 30 to 39.9991 s
STI301=1
STI101=4101
STI101=5
STI101=7
STI101=4113
STI101=13
EEG002
EEG004
EEG006
EEG008
EEG010
EEG012
EEG014
EEG016
EEG018
EEG020
EEG022
EEG024
EEG026
EEG028
EEG030
EEG032
EEG034
EEG036
EEG038
EEG040
EEG042
EEG044
EEG046
EEG048
EEG050
EEG052
EEG054
EEG056
EEG058
EEG060
EEG062
EEG064
EEG066
EEG068
EEG070
EEG072
EEG074
30.00
31.00
32.00
33.00
34.00
35.00
36.00
37.00
38.00
39.00
40.00
time
preproc
identify
visual
<
>
segment
<
>
channel
<
>
horizontal
-
+
vertical
-
+


stimuli & task
triggers
analysis
acquisition
EEG
system



2: ft_databrowser: data
File
Edit
View
Insert
Tools
Desktop
Window
Help
FieldTrip
segment 4/50, time from 30 to 39.9991 s
EEG002
EEG004
EEG006
EEG008
EEG010
EEG012
EEG014
EEG016
EEG018
EEG020
EEG022
EEG024
EEG026
EEG028
EEG030
EEG032
EEG034
EEG036
EEG038
EEG040
EEG042
EEG044
EEG046
EEG048
EEG050
EEG052
EEG054
EEG056
EEG058
EEG060
EEG062
EEG064
EEG066
EEG068
EEG070
EEG072
EEG074
30.00
31.00
32.00
33.00
34.00
35.00
36.00
37.00
38.00
39.00
40.00
time
preproc
identify
visual
<
>
segment
<
>
channel
<
>
horizontal
-
+
vertical
-
+

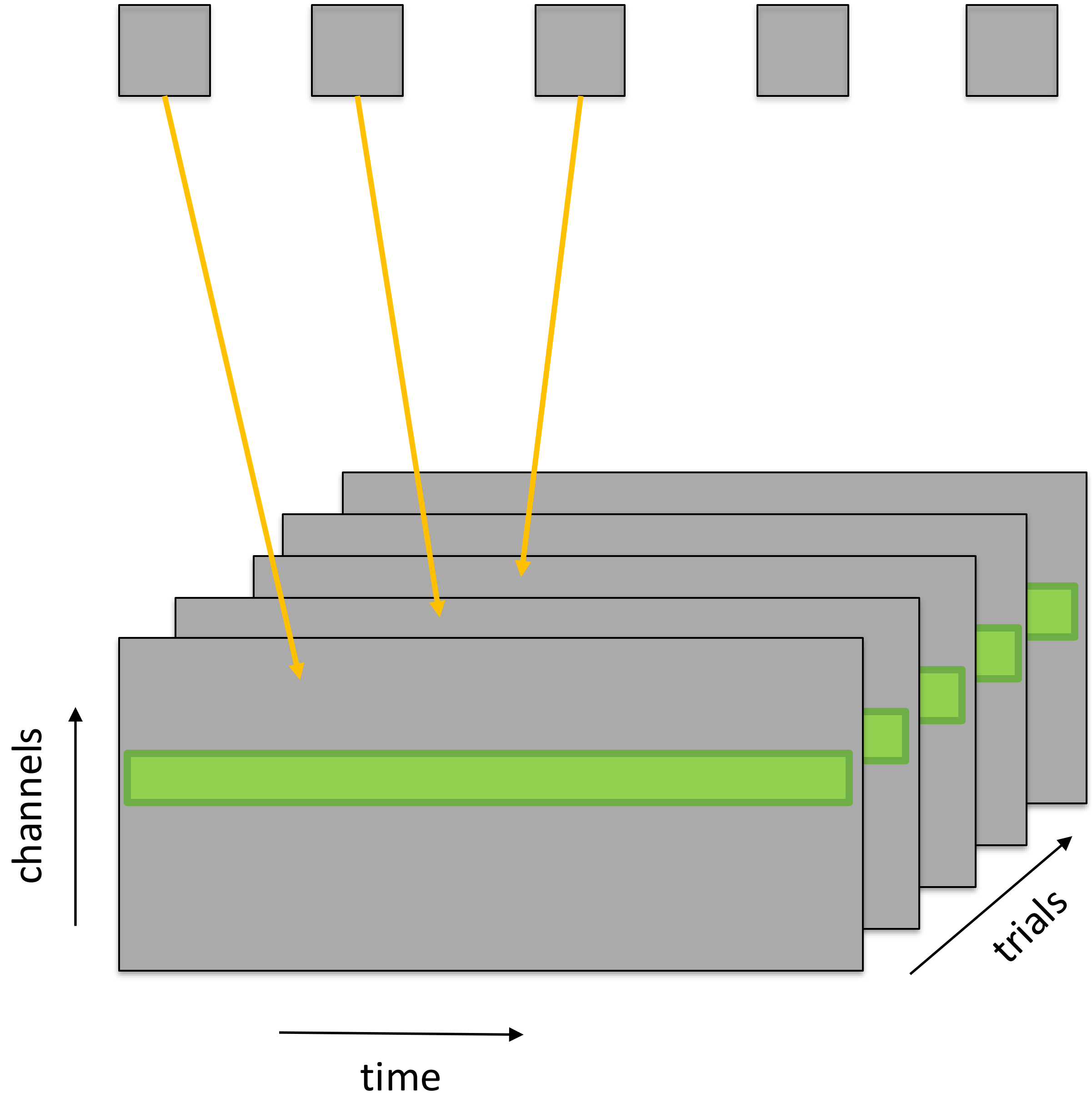




~900 trials
~64 channels
~ 1 second = 1100 samples

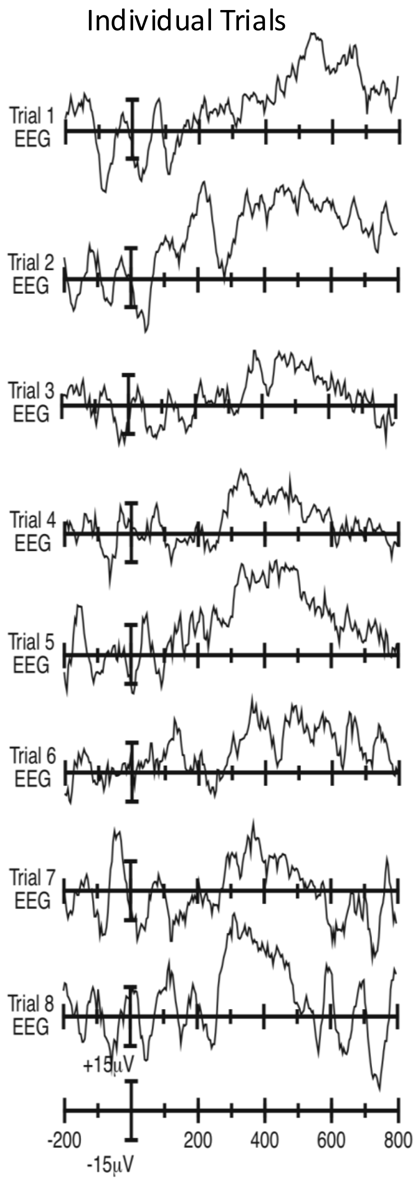
So 900x64x1100 = 63.000.000 numbers

… times 16 subjects

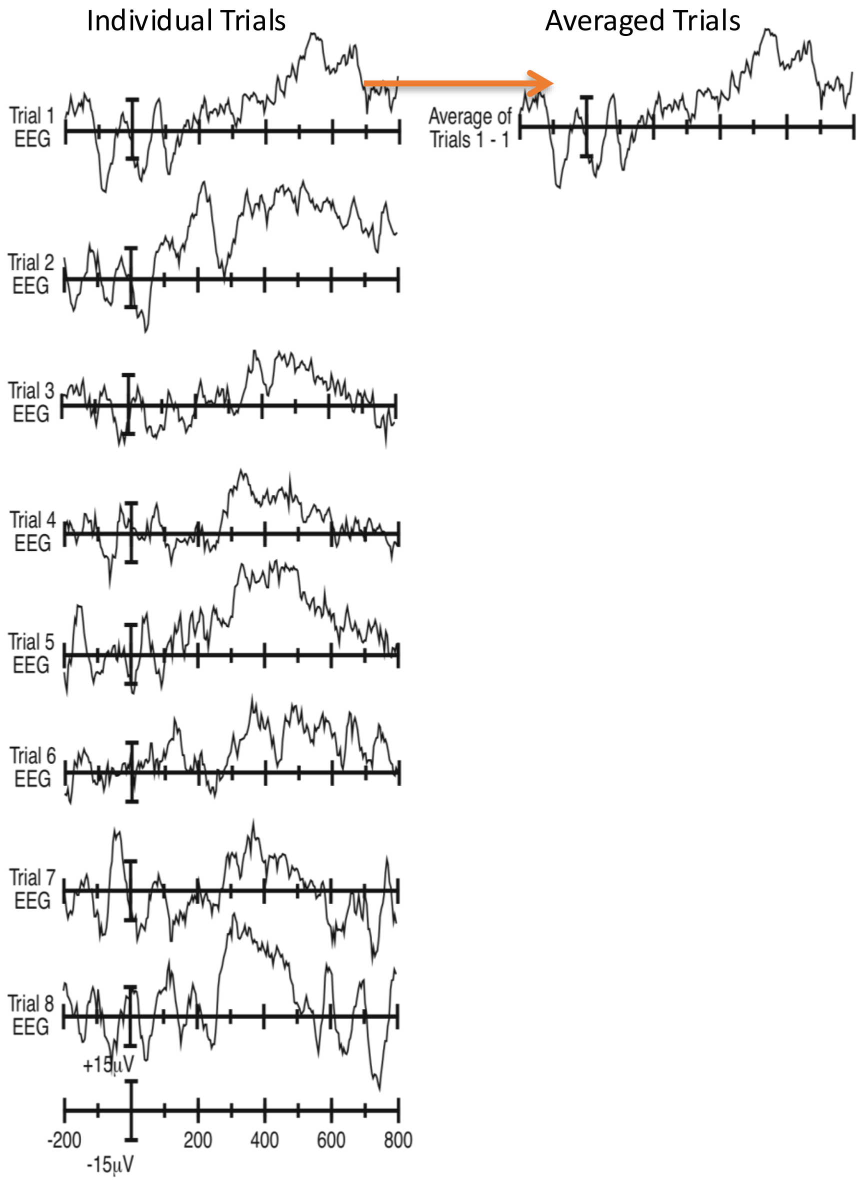


channels
time
trials

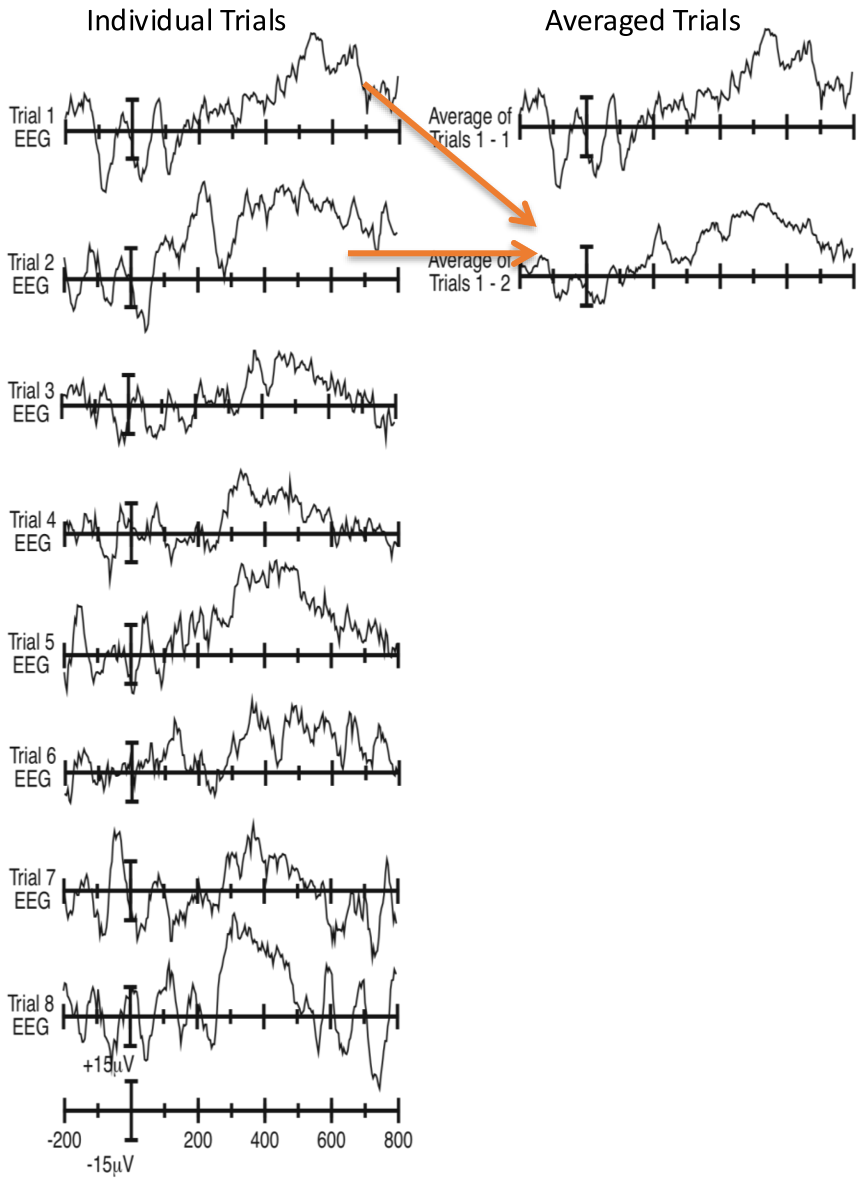
Event-Related Potential
ERP



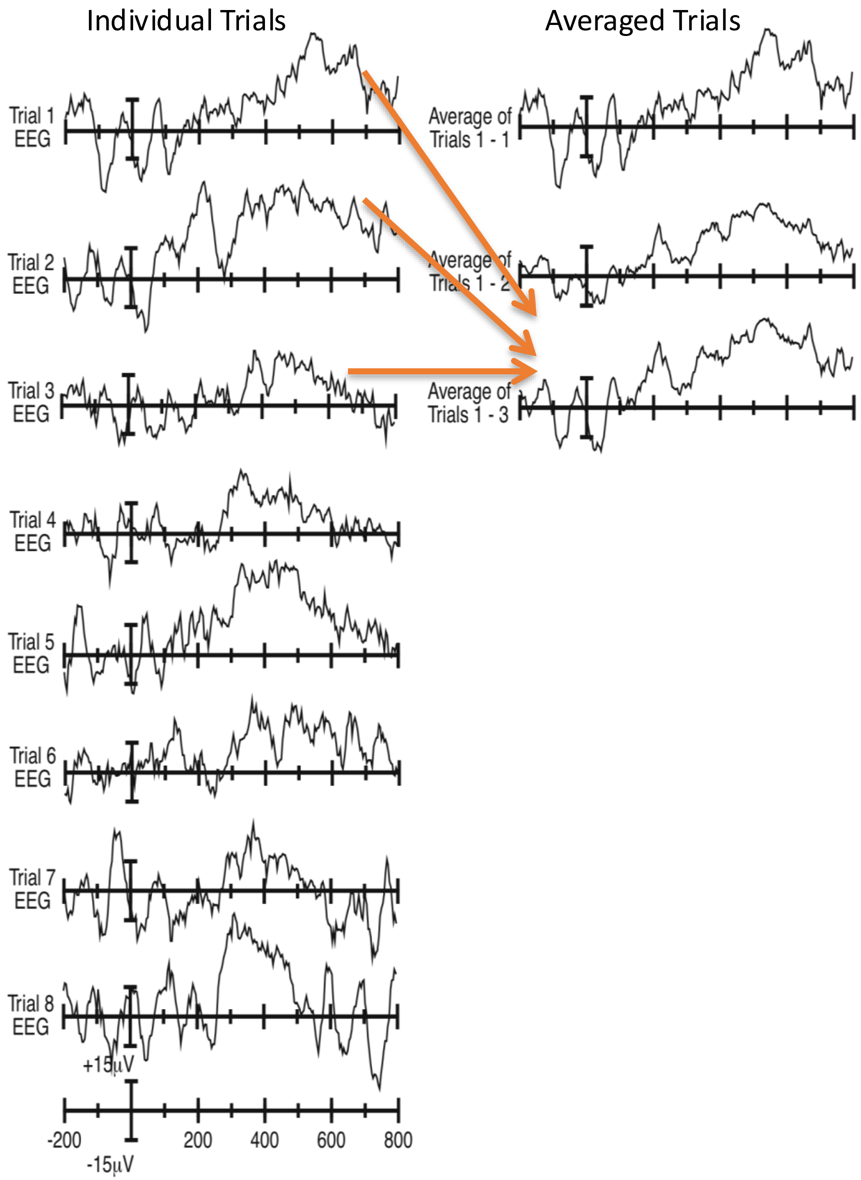
Event-Related Potential
ERP



Event-Related Potential
ERP

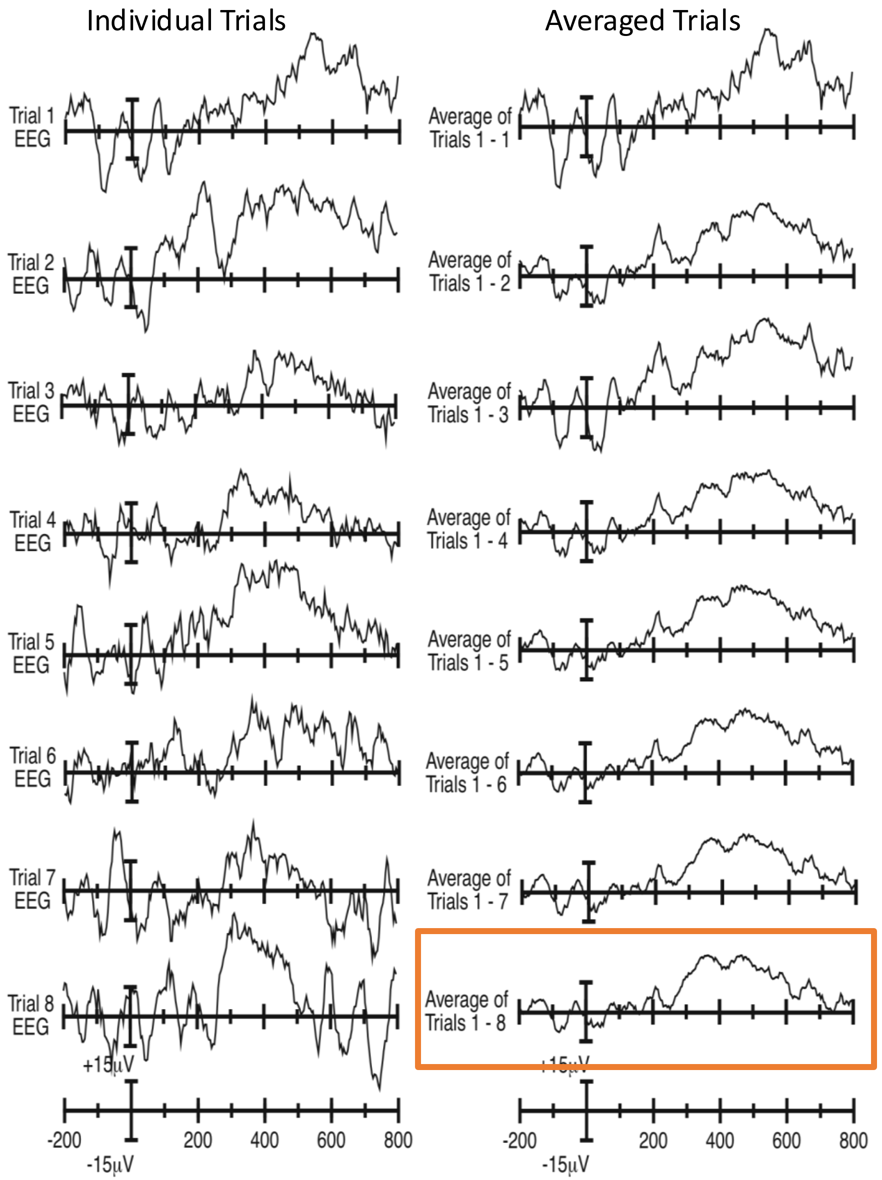


Event-Related Potential
ERP



Event-Related Potential

ERP



Event-Related Potential
ERP

The brain ***signal*** of interest is assumed to be constant over all trials.

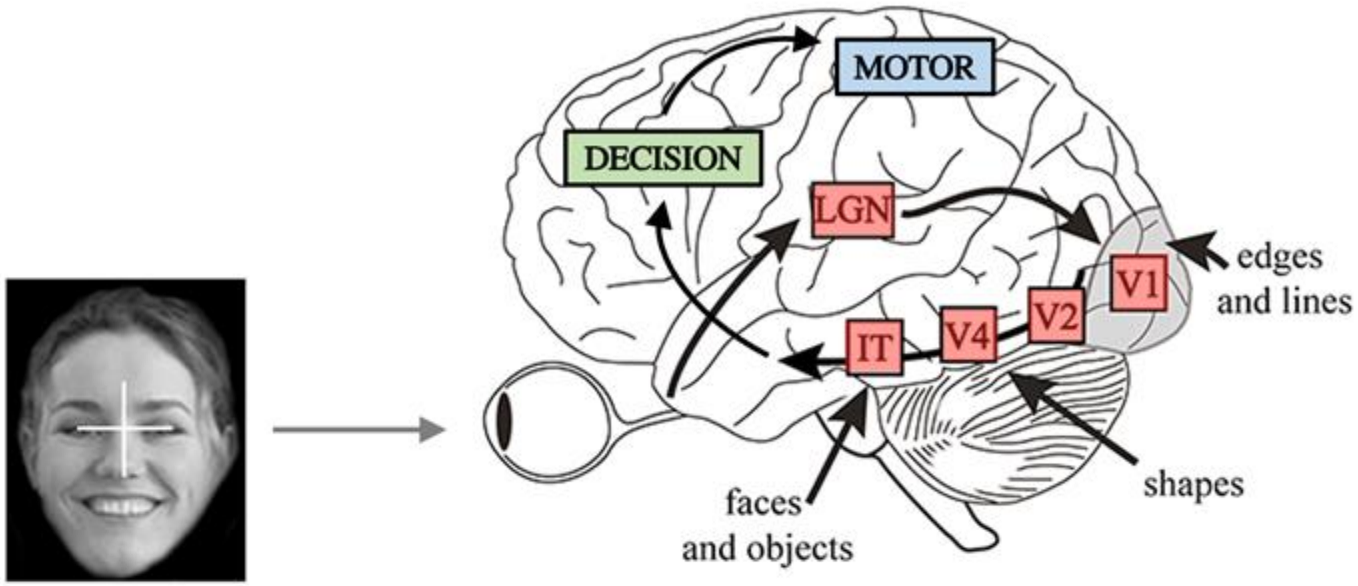
The ***noise*** is independent over trials.

After averaging the noise is proportional to 1/sqrt(Ntrials).

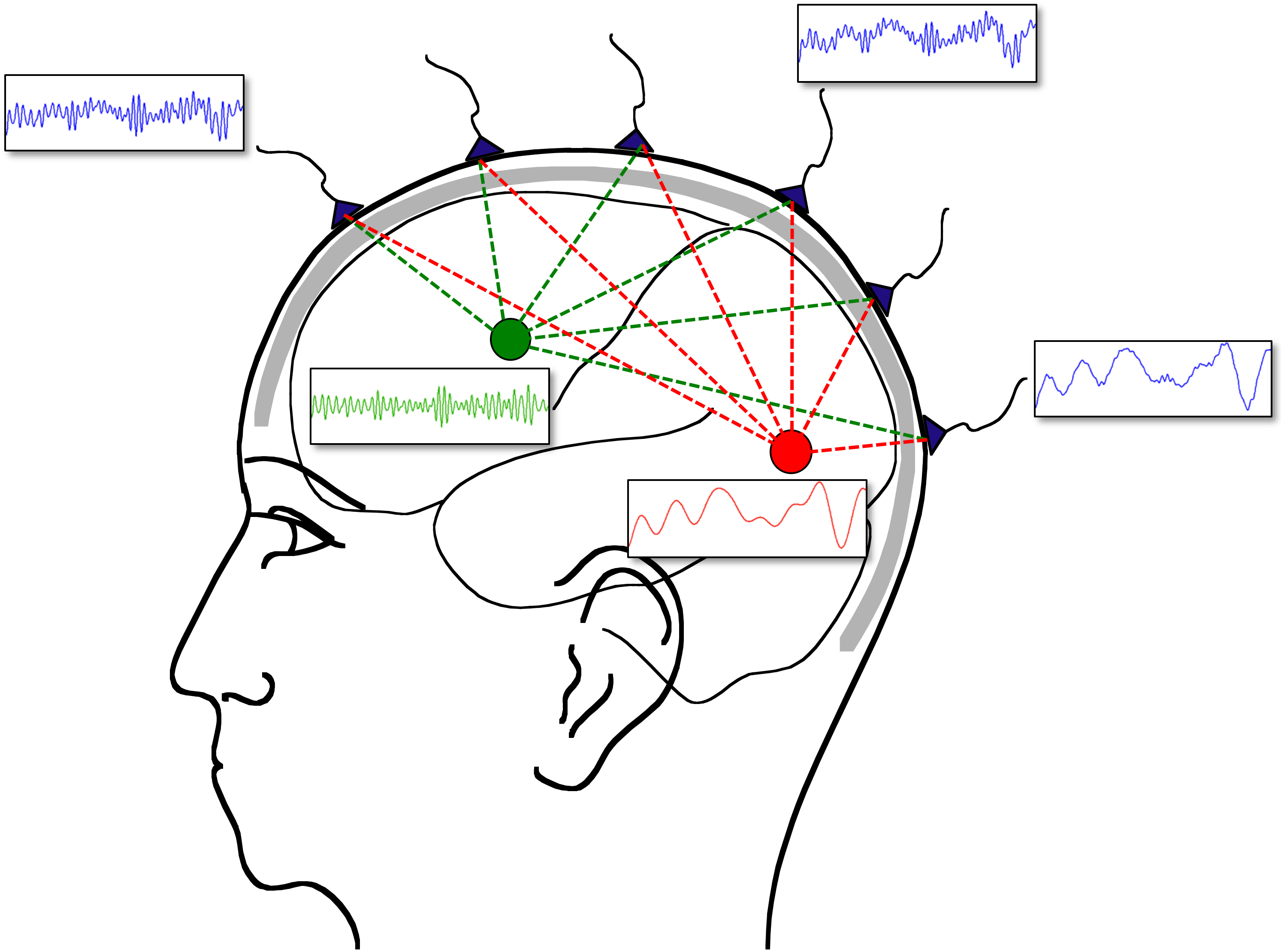
Averaging over trials improves the signal-to-noise ratio (SNR).

10 trials	-> SNR is sqrt(10)	≅ 3x better
100 trials	-> SNR is sqrt(100)	≅ 10x better

Linear model for ERP superposition



Linear model for ERP superposition



Linear model for ERP superposition

Related to the task, constant over trials

Not related to the task

Data = signal + noise

Data in condition 1 = signal_1 + noise

Data in condition 2 = signal_2 + noise

ERP_1 = average(Data in condition 1) = signal_1 + noise

ERP_2 = average(Data in condition 2) = signal_2 + noise

$\text{ERP}_1 - \text{ERP}_2 = \text{signal}_1 - \text{signal}_2$ + noise

Linear model for ERP superposition

Data = signal + noise

Data in condition 1 = visual + reconition_1 + noise

Data in condition 2 = visual + recognition_2 + noise

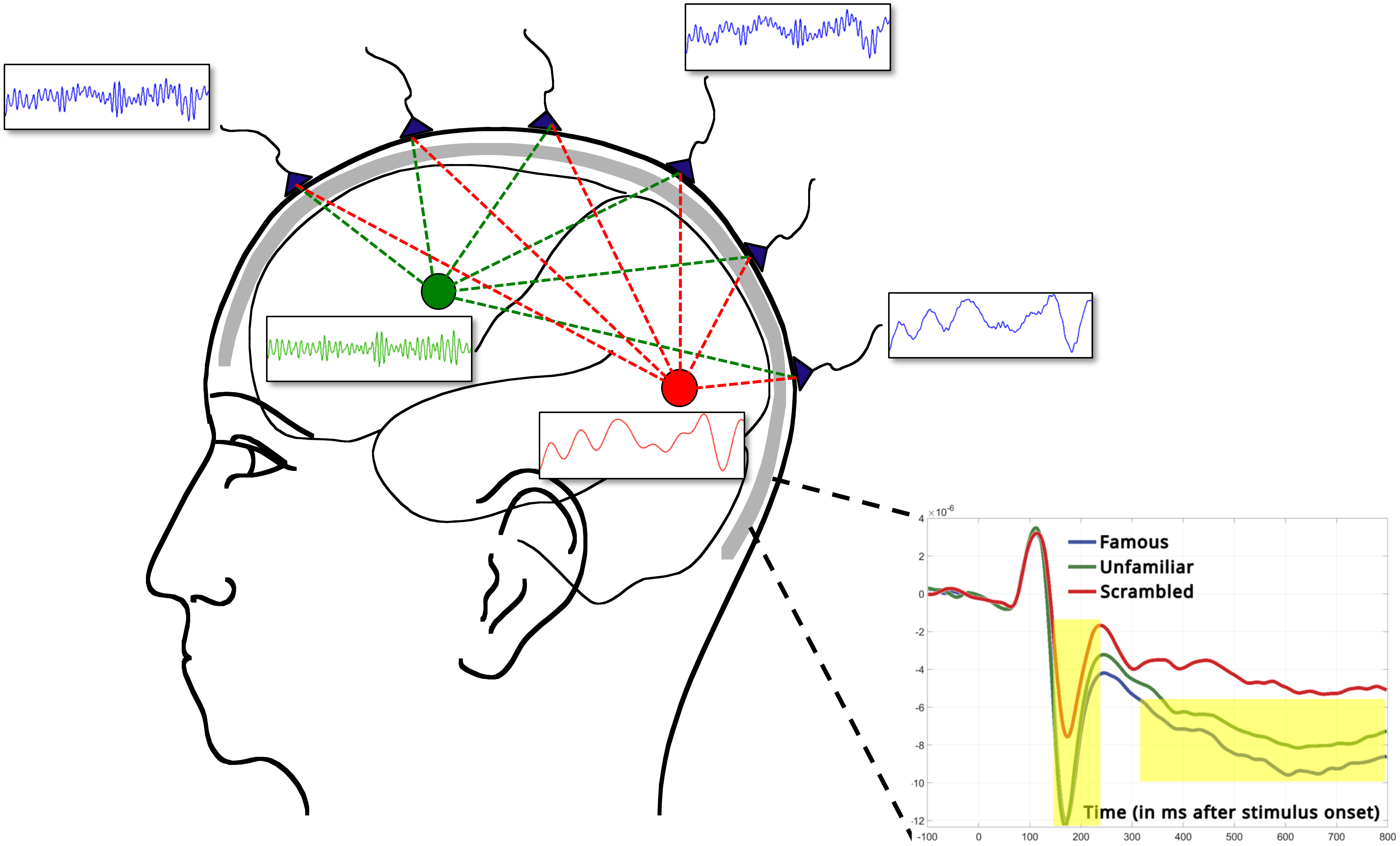
ERP_1 = average(Data in condition 1) = visual + recognition_1 + noise

ERP_2 = average(Data in condition 2) = visual + recognition_2 + noise

$\text{ERP}_1 - \text{ERP}_2 = \text{reconition}_1 - \text{recognition}_2 + \text{noise}$

ERP difference to tap into specific cognitive process

followed by statistics, etc.



Get to know your data – learning goals

What are the characteristics of the data that we use in the analysis?

How to organize your raw data?

Quality assessment and control

What are the artifacts and why are they relevant?

Preprocessing and segmenting (or vice versa)

Selective averaging to get ERPs/ERFs

Radboud University

DONDERS INSTITUTE

Get to know your data artifacts, segmentation, preprocessing

Practical MEEG
3rd Edition

Robert Oostenveld

Donders Institute, Radboud University, Nijmegen, NL
Karolinska Institutet, Stockholm, SE

robert.oostenveld@donders.ru.nl

shared slides

Karolinska Institutet

NatMEG