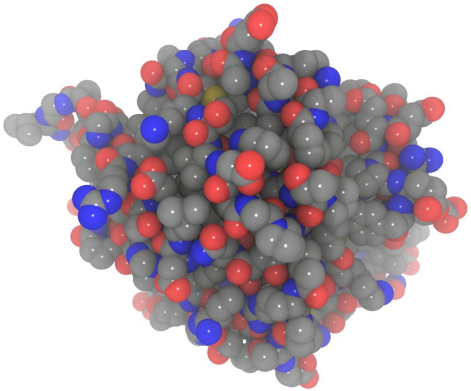
EEG signals – initial analysis

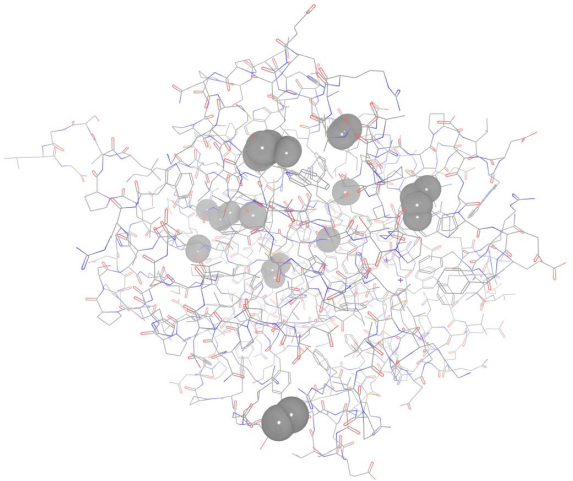
Martin Maňák
11.3.2025

Compression of molecular structures



atoms represented as balls
scale: 1A = 10^{-9}m

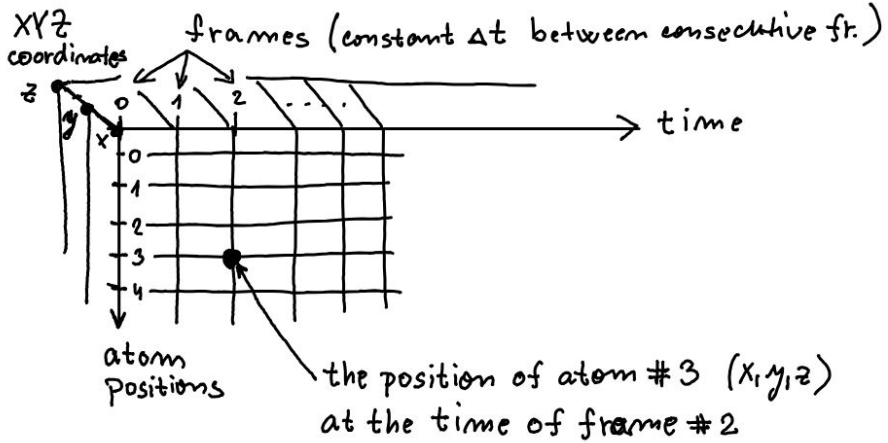
backbone (alpha-trace)



bonds between atoms

- Molecular dynamics: simulated motion of atoms.
- Constraints - bonds, forces, ...

Compression of molecular structures



N atoms, F frames, 32-bit float per coordinate

⇒ N·F·3·4 BYTES to store the trajectory

(real datasets ~ 5 GB (or even more))

Compact representations of trajectories

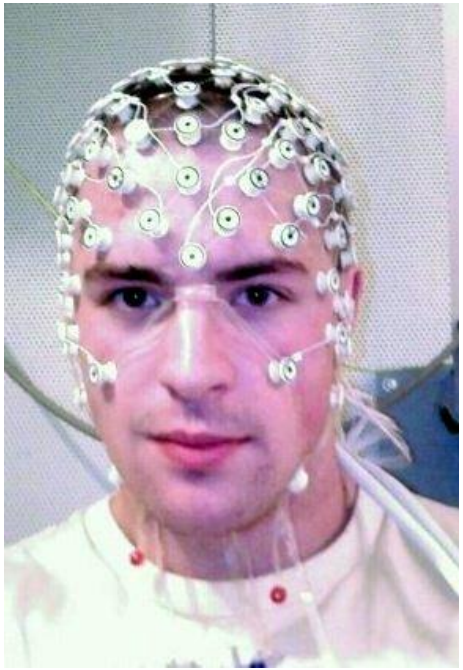
- NetCDF files (.nc)
  - zlib (deflate)
- Gromacs trajectory files (.xtc)
  - float to integer coordinates
  - store with reduced precision
- High-res. traj. compression (HRTC)
  - traj -> piecewise lin. func.
  - (tolerance corridor)
  - 1 bit per sample (particle simul.)
- Predictive compression (PMC)
  - uses the structure of bonds
  - next-frame positions prediction
  - error control (e.g., diff < 10^{-3}A)

Applicability to digital EEG signals is limited

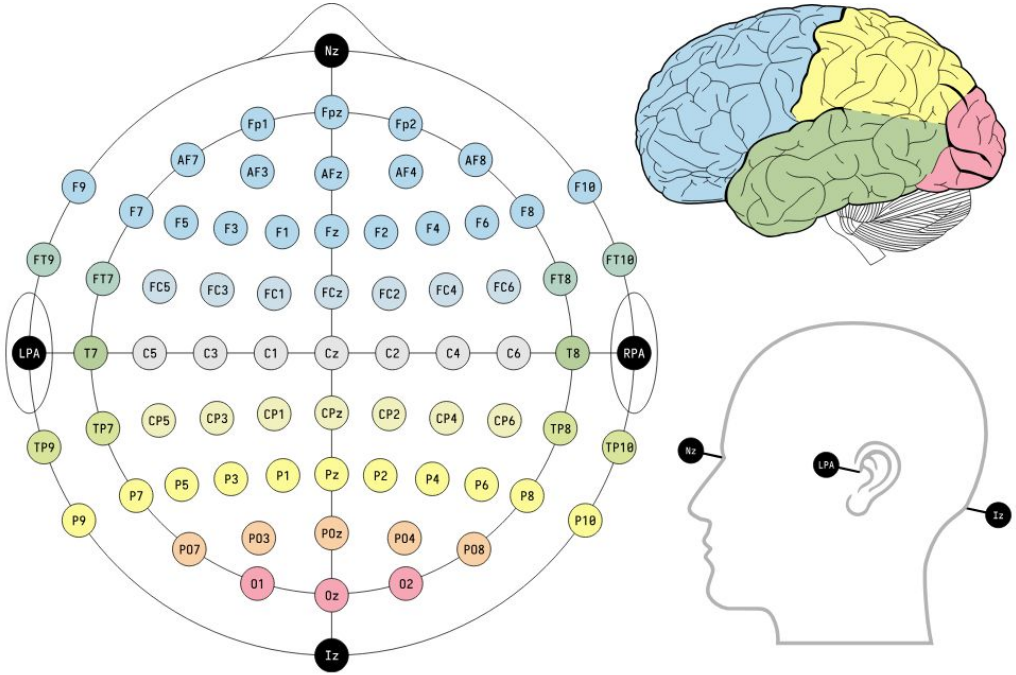
- different kind of data
- voltage measured at electrodes (up to 257)
- 1000 samples per second
- stored as 32-bit floats or 64-bit doubles
- reduction of precision is “problematic”
- no such connectivity as bonds in molecules

https://commons.wikimedia.org/wiki/File:ElectroEncephalogram.png

Data acquisition (wiki example)

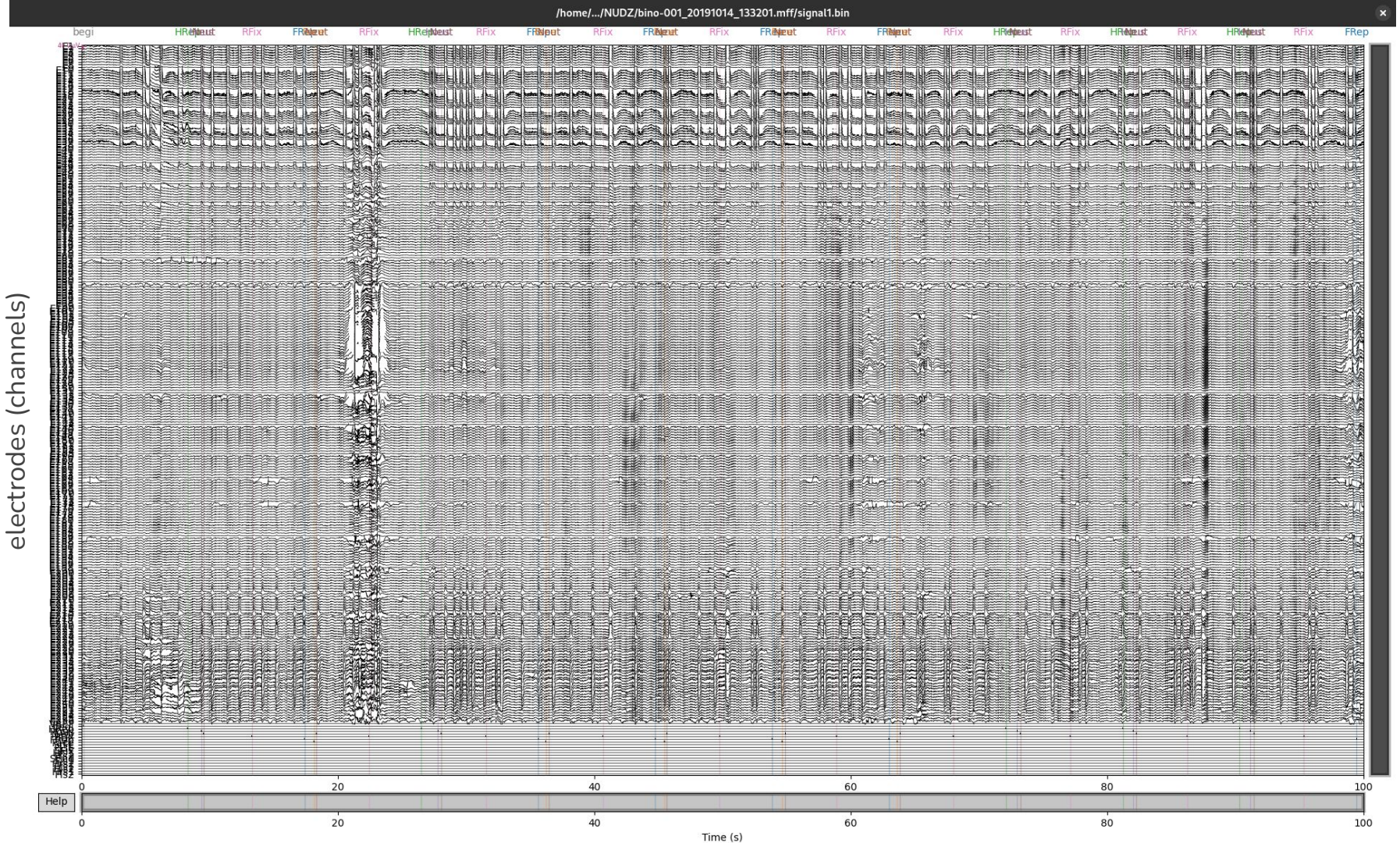


https://en.wikipedia.org/wiki/Electroencephalography#/media/File:EEG_cap.jpg



https://upload.wikimedia.org/wikipedia/commons/f/fb/EEG_10-10_system_with_additional_information.svg

NUDZ Binocular rivalry dataset


/home/.../NUDZ/bino-001_20191014_133201.mff/signal1.bin
begi
RFix
FRep
electrodes (channels)
0
20
40
60
80
100
Help
Time (s)

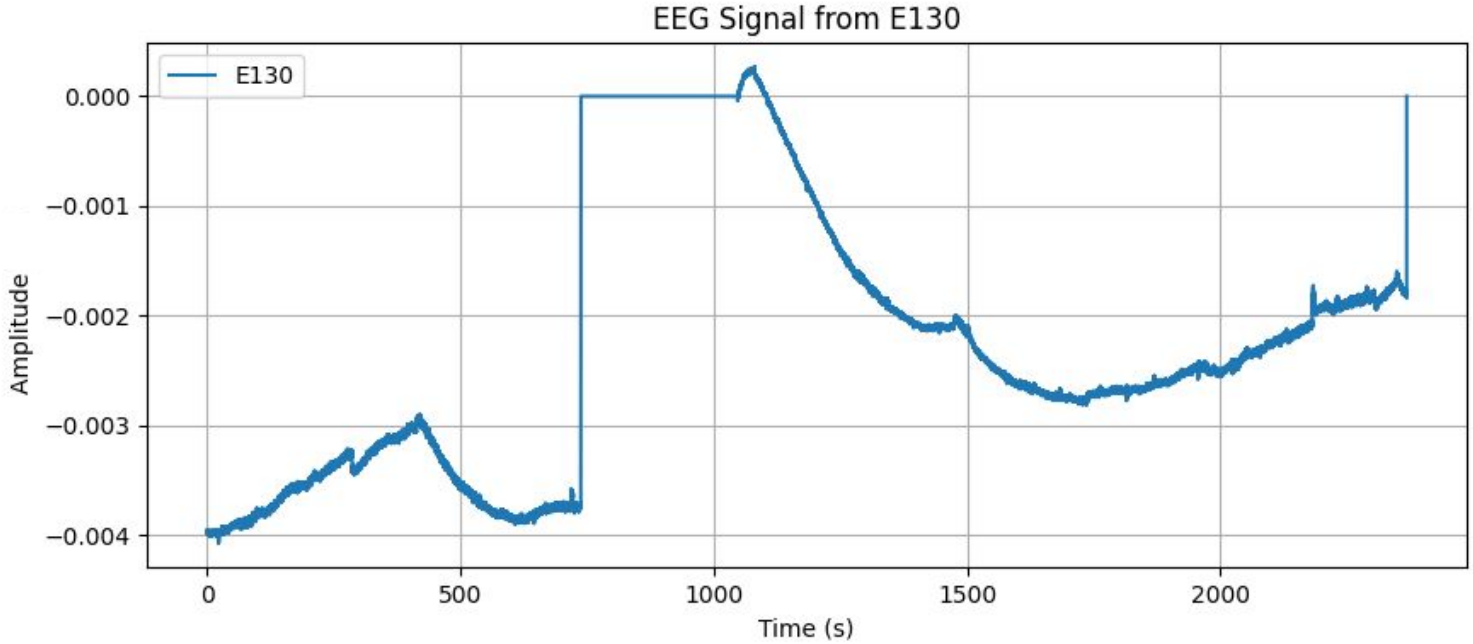
The size of EEG binary data (EGI MFF dataset)

- 257 EEG channels
- 2,3M samples/channel (1,000 samples/s, 40 minutes of recording)
- stored as 32-bit floats (probably), loaded as 64-bit doubles

- 257 * 2,367,324 * 4 bytes
  =>
  2,1 GB file (signal1.bin)
  + some 1 MB of meta-data

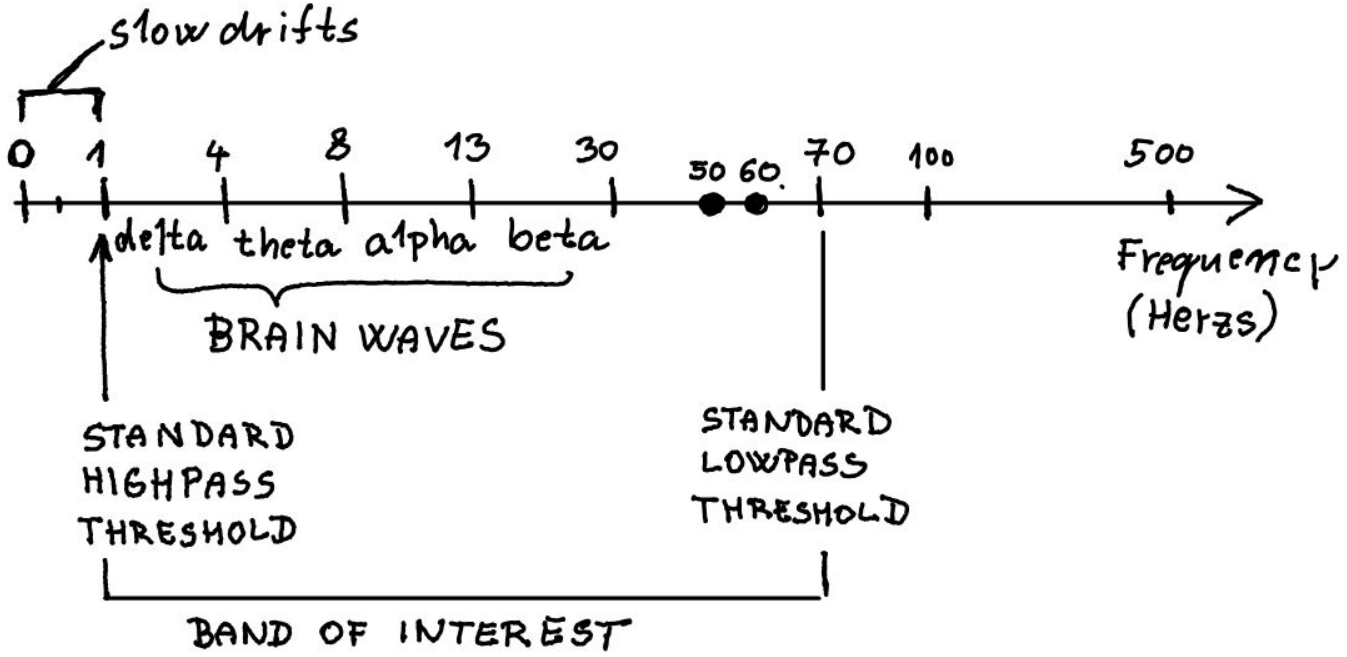
- Q: Could we compress the data somehow? Some correlation is visually apparent...

Actual RAW data for a single electrode (randomly selected one)

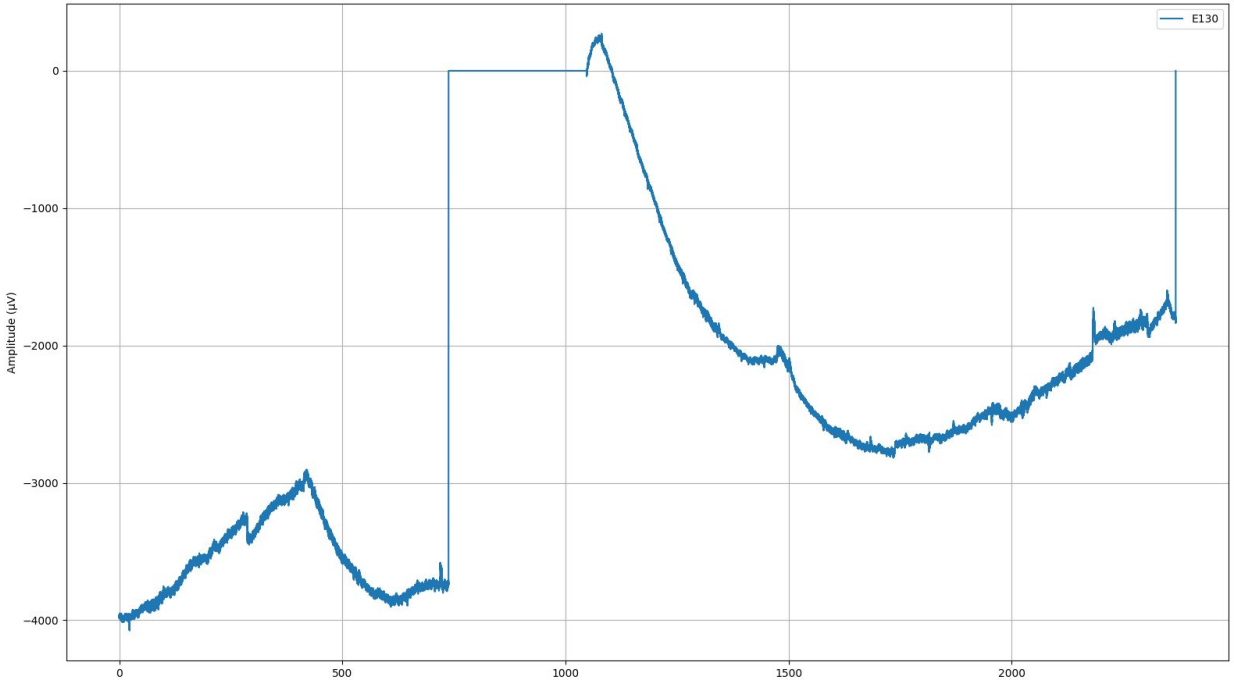


A spectral band (ai generated, DALL-E)

Spectral bands / EEG signals

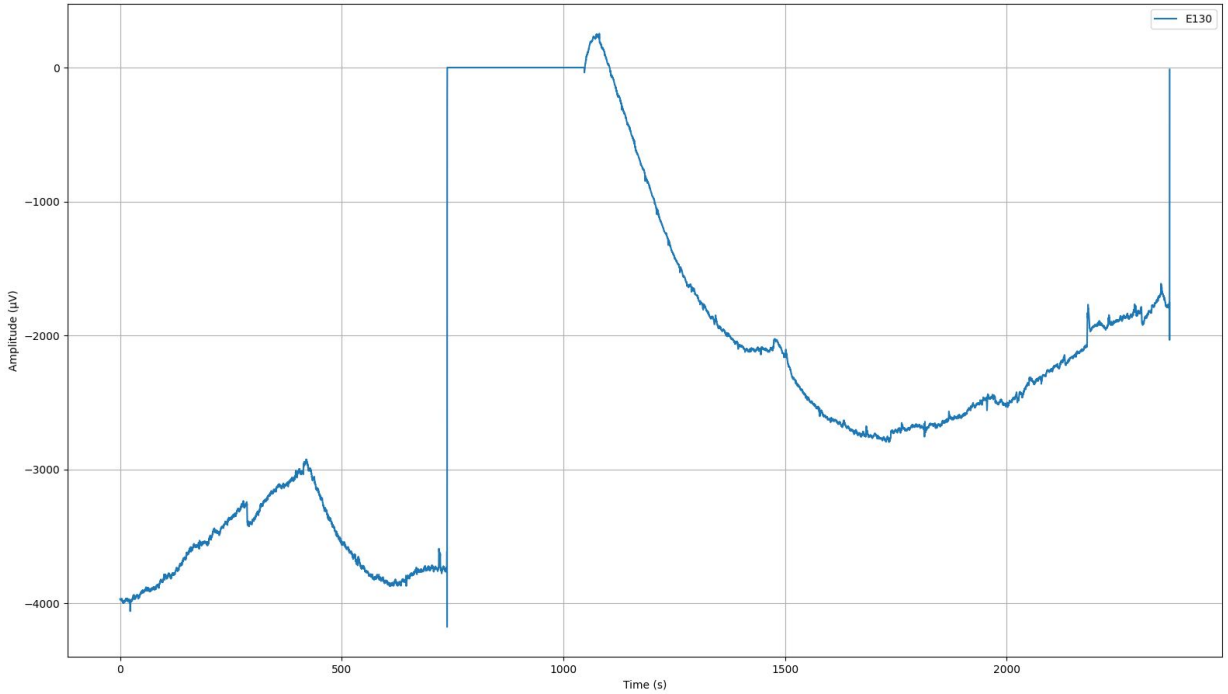


Channel E130 - RAW data



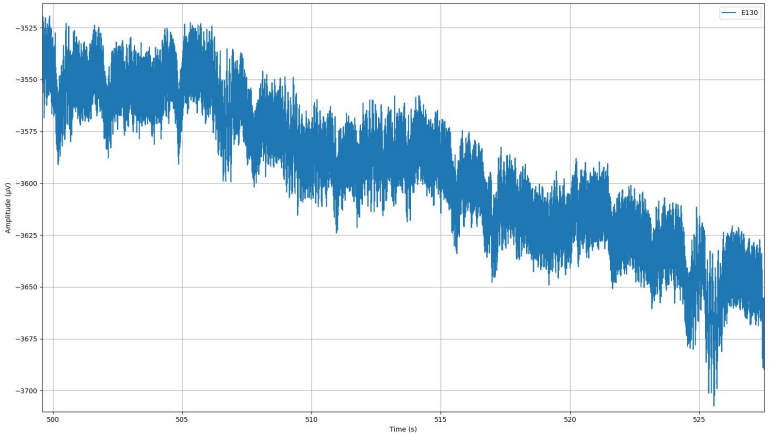
- Frequencies up to 500 Hz
- Nyquist:
  - need 1000 samples / s to represent the signal
  - means **2,1 GB** for all channels together

Channel E130 - lowpass 1 Hz

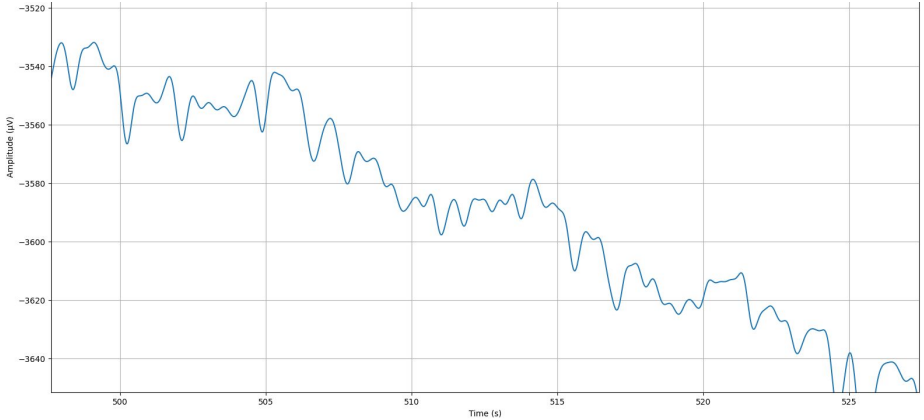


- Frequencies up to 1 Hz
- Nyquist
  - need 2 samples / s to represent the signal
- We have: 1000 samples/s
- Downsampling (after lowpass filtering)
  - keep every 500th sample (0.2 %)
  - means **4,2 MB** for all channels together
- Upsampling (to get it back)

Channel E130 - a closer look on details



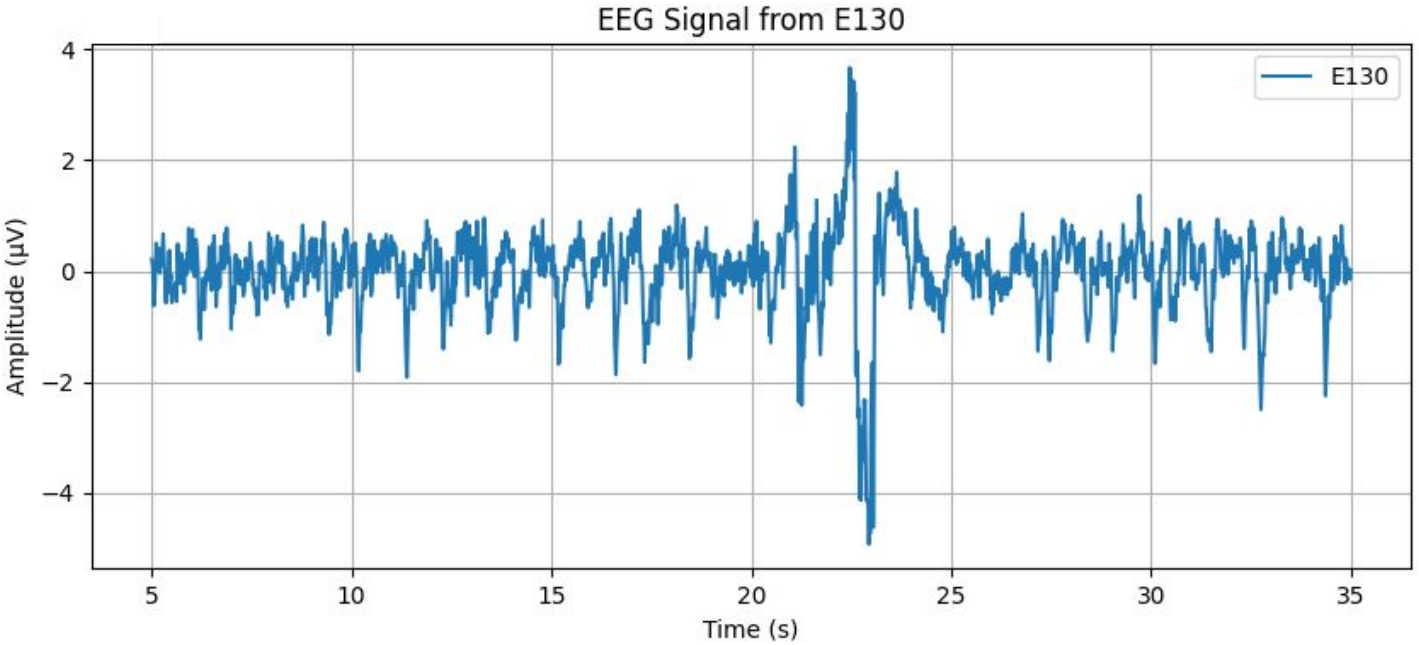
25 seconds, RAW data, 1000 samples/s



25 seconds, 1Hz lowpass (yet still 1000 samples/s)

Channel E130, bandpass 1-30Hz

The filter what a domain-expert might set up to analyze the signal (may further stretch the signal horizontally/vertically according to standards - sensitivity / reading speed)

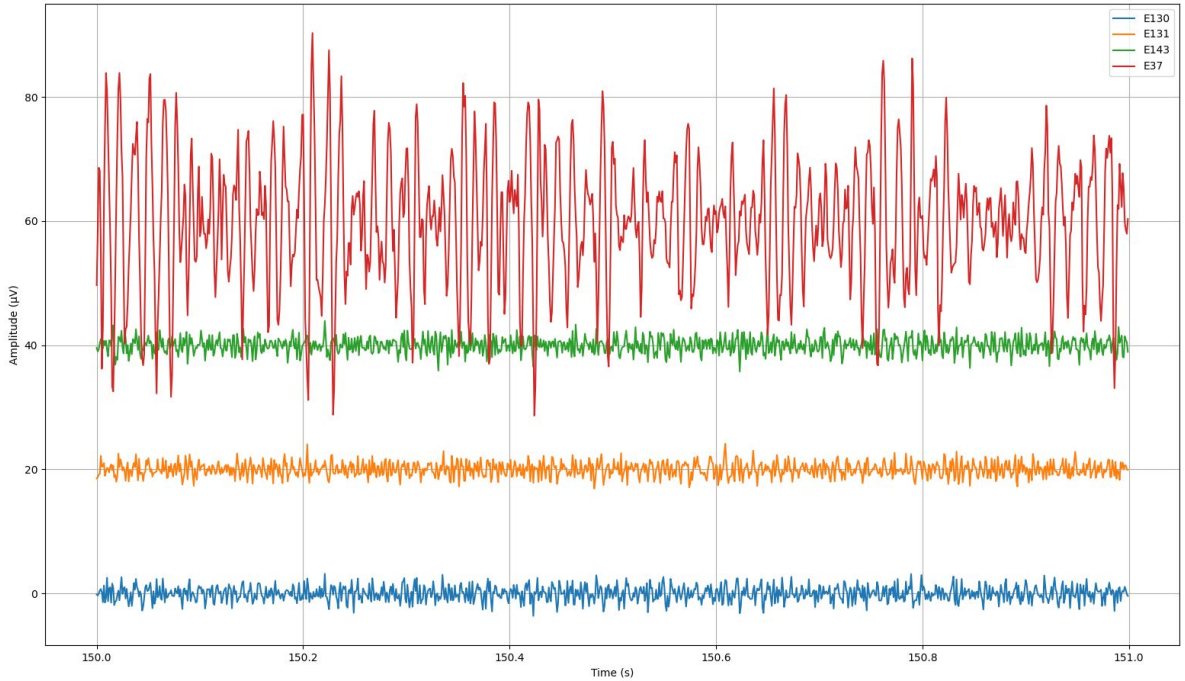


Spectral bands - some conclusions

worth to keep

- 0-1Hz (slow drifts)
  - domain-experts filter-out these frequencies explicitly
  - **0.2 %** of data (**4-5 MB** out of **2.1 GB**)
- 1-30-100 Hz (interesting brain waves, spikes, artifacts)
  - sampling rate 60-200 Hz
  - **5,8 - 19,8 %** of data (**126-420 MB** out of **2.1 GB** - if we are 32 bits/sample generous)
- 100-500 Hz (mostly noise)
  - also epileptic seizure spikes, some drug-effects, high temporal resolution, etc.
  - **94,2 - 80,2 %** of data (**1.68-1.974 GB** out of **2.1 GB**)

...mostly noise / residual content



70 Hz highpass filter

E130, E131, E143 are close to each other (on the head-mount). E37 is far.

“low” amplitude (a few uV)

Final thoughts

Is it worth to further compress the interesting 10% of data if we have to keep the remaining 90% of noise? (lossless compression)

- Probably not so much.

How to define information loss threshold?

- Perhaps as the ability to restore data to 1 uV resolution (basic unit)

How to compactly represent high frequency noise (the residual content)?

- An appropriate number-coding scheme could give us 1 + $\log_2$(std_dev) bits / sample