

DATA QUALITY

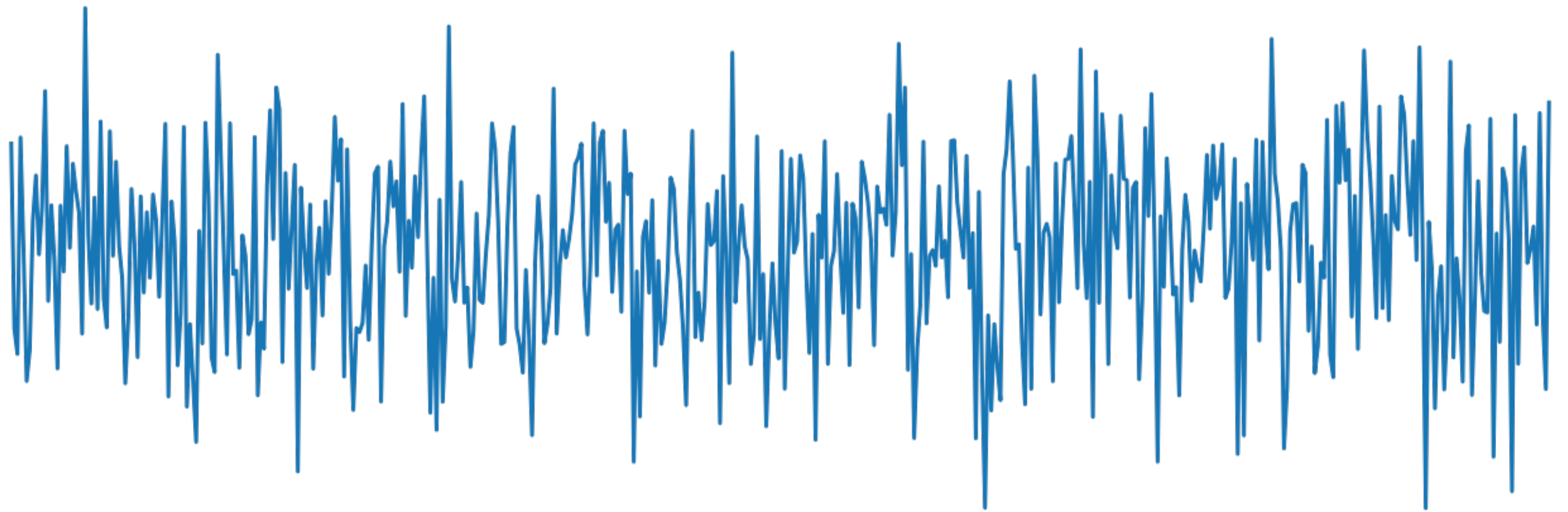
Alexander Huth
@ UBA ECI Course
2017-7-25

ONLINE MATERIALS

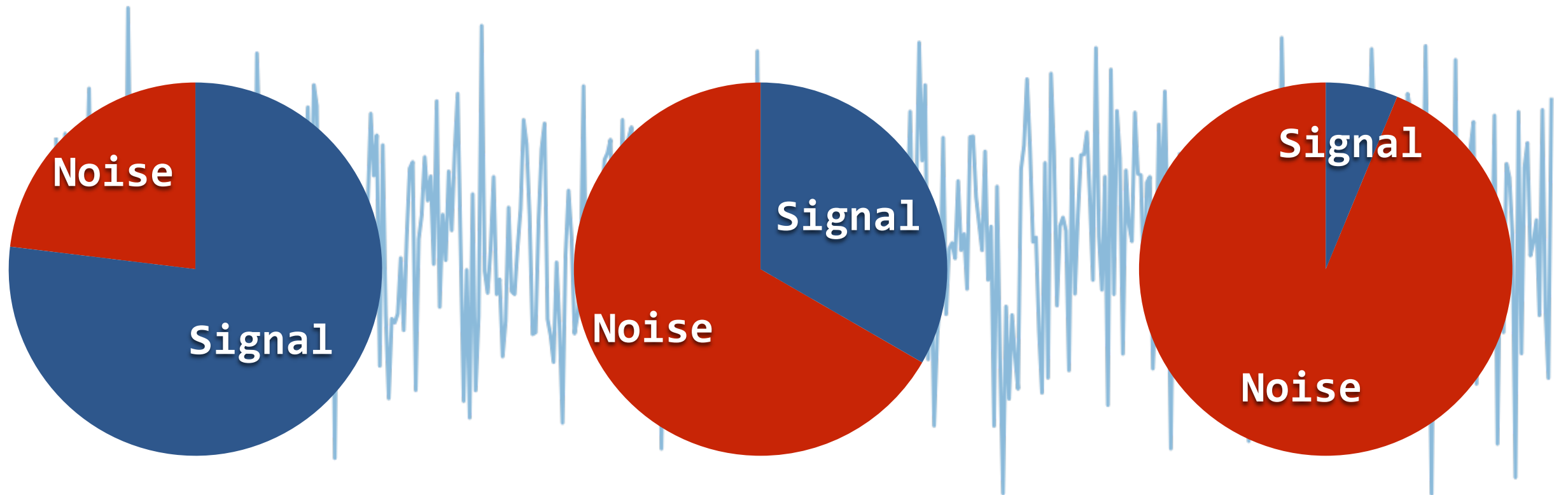
<http://github.com/alexhuth/eci-2017>

**BEFORE YOU DO
ANYTHING ELSE,
*MAKE SURE YOUR
DATA IS GOOD***

HOW GOOD IS YOUR (TIMESERIES) DATA?



HOW GOOD IS YOUR (TIMESERIES) DATA?



WHAT IS NOISE?

- If the same stimulus is repeated, the **NOISE** is different while **SIGNAL** is the same

$$x_i(t) = s(t) + \epsilon_i(t)$$

measured response on i 'th repetition

WHAT IS NOISE?

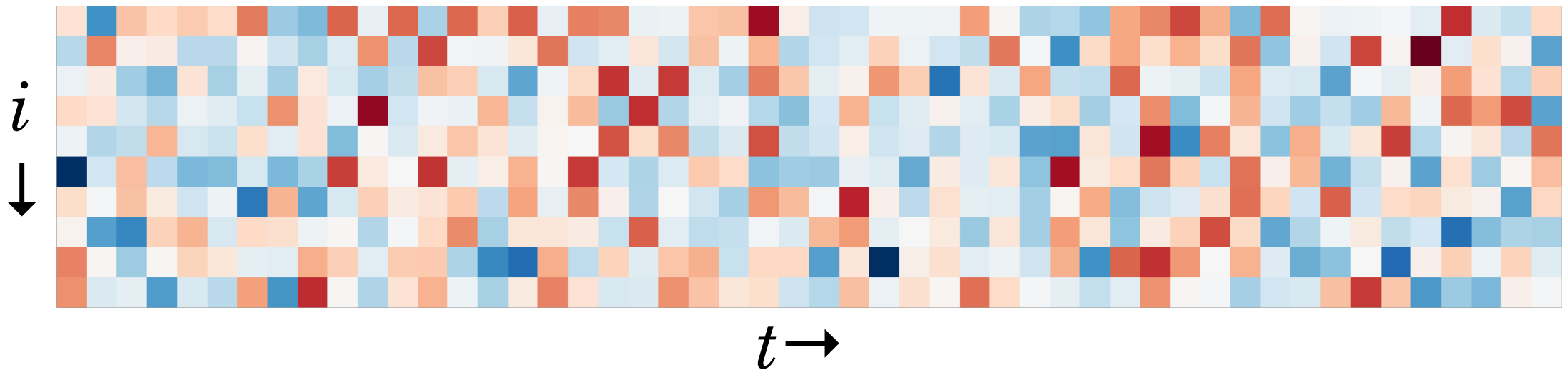
- (Assuming *stationarity* of the signal!)

HOW DO WE KNOW?

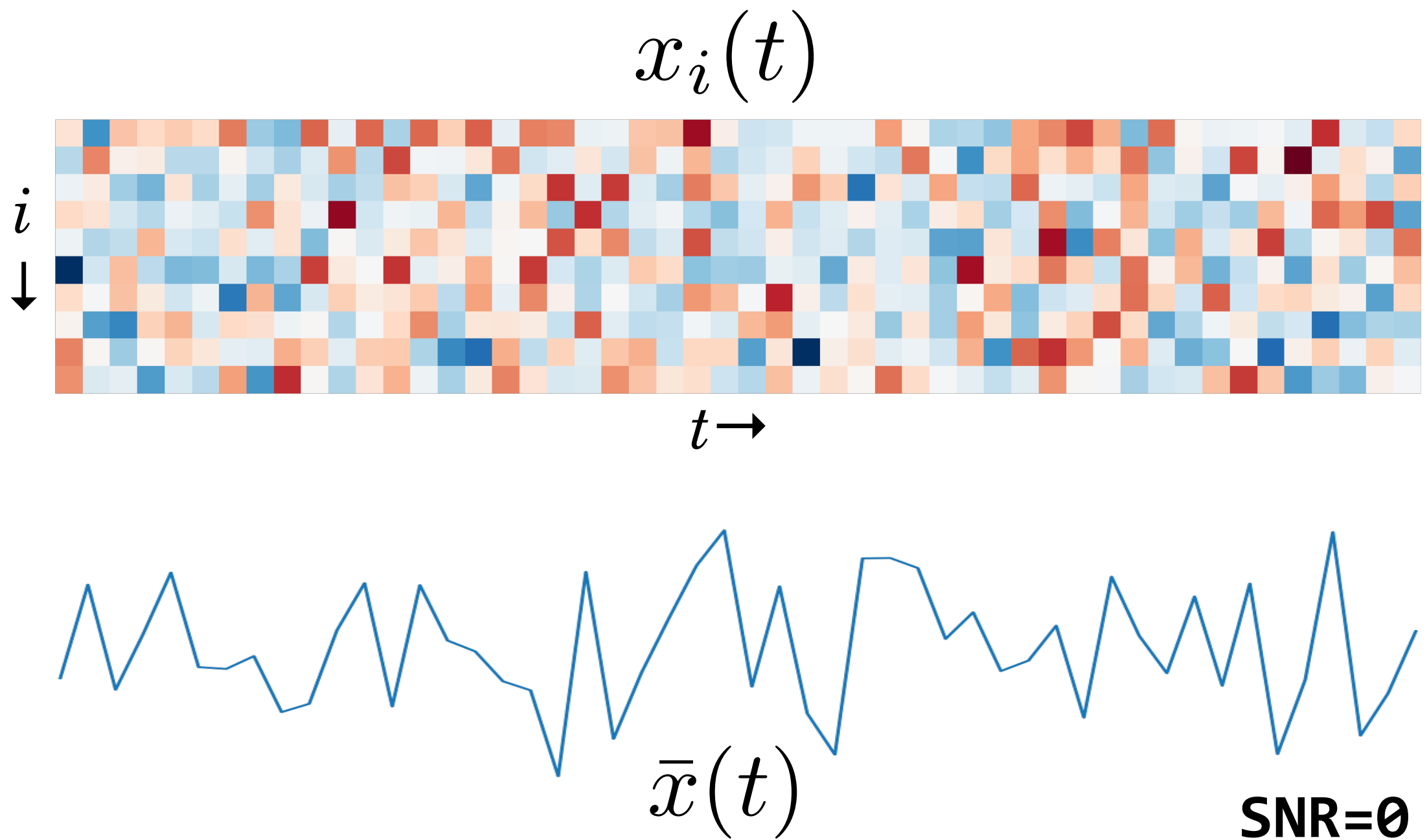
- Repeat the same experiment multiple times
- The component of the response that is the same across repetitions is the **SIGNAL**, the components that are different are **NOISE**

HOW DO WE KNOW?

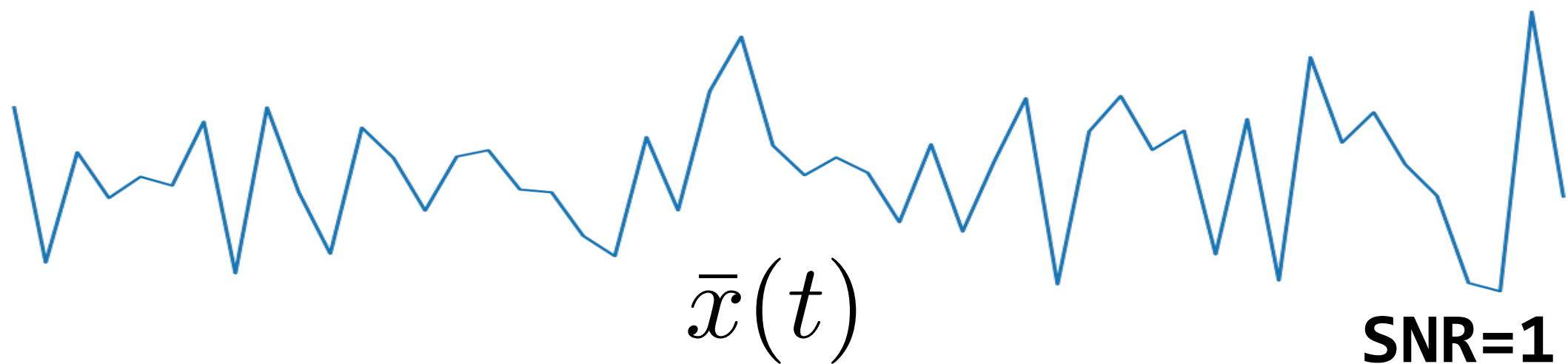
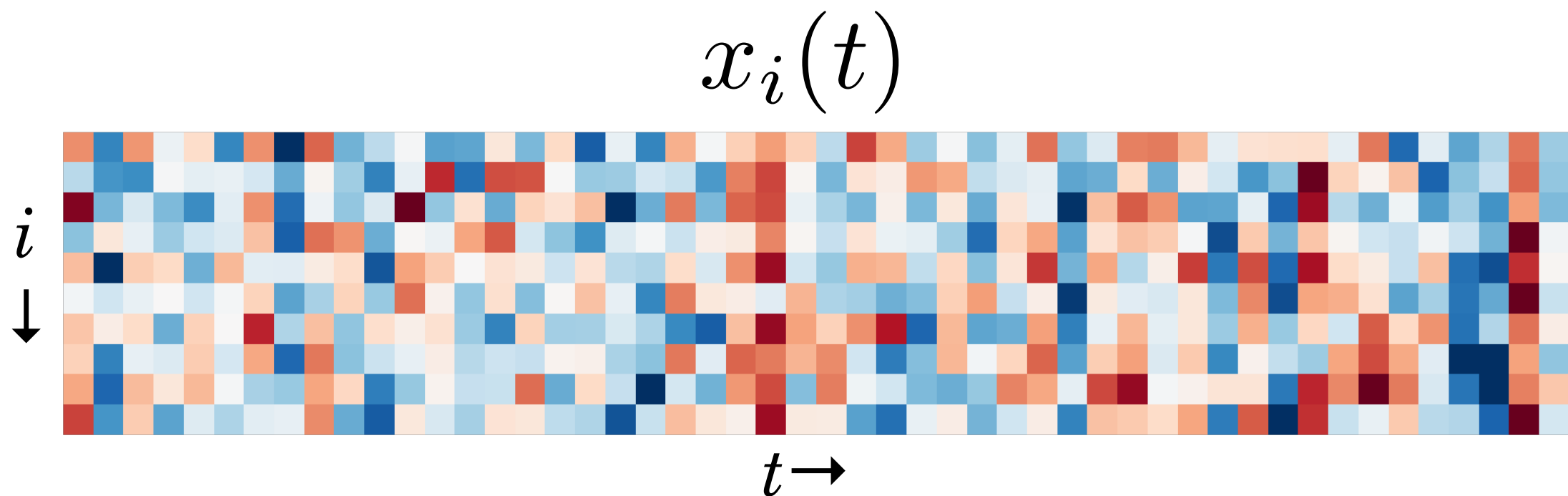
$$x_i(t)$$



HOW DO WE KNOW?



HOW DO WE KNOW?



PREGUNTA

What is noise? Is all trial-to-trial variability noise?

- **Please:**
 - Think about this question for 2 minutes
 - Turn to your neighbor and discuss what you thought of

METRICS FOR *REPEATABILITY*

- **SNR** (signal-to-noise ratio)
- **EV** (explainable variance)
- **MPWC** (mean pairwise correlation)
- **Coherence spectrum**

SIGNAL TO NOISE RATIO

- The signal-to-noise ratio (**SNR**) is defined as:

$$SNR = \frac{\text{var}(s(t))}{\text{var}(\epsilon(t))}$$

- But this is rarely used in practice (at least for neuroscience data)

$$SNR = \frac{\text{var}(s(t))}{\text{var}(\epsilon(t))}$$

SIGNAL TO NOISE RATIO

- In practice **SNR** must be computed using mean response:

$$\hat{SNR} = \frac{\text{var}(\bar{x}(t))}{\langle \text{var}(x_i(t) - \bar{x}(t)) \rangle_i}$$

$$\hat{SNR} = \frac{\text{var}(\bar{x}(t))}{\langle \text{var}(x_i(t) - \bar{x}(t)) \rangle_i}$$

SIGNAL TO NOISE RATIO

- **NB:** *Functional SNR* is not **tSNR** (temporal SNR) aka **SFNR** (signal to fluctuation noise ratio) commonly used in MRI & image processing
- **tSNR/SFNR** are usually defined as inverse of coefficient of variation:

$$tSNR = \frac{\text{mean}(x(t))}{\text{std}(x(t))}$$

EV (EXPLAINABLE VAR.)

- How much of the total variance is explained by the mean across repeats?

$$EV = 1 - \frac{\sum_i \text{var}(x_i(t) - \bar{x}(t))}{\sum_i \text{var}(x_i(t))}$$

$$EV = 1 - \frac{\sum_i \text{var}(x_i(t) - \bar{x}(t))}{\sum_i \text{var}(x_i(t))}$$

EV (EXPLAINABLE VAR.)

- EV is between 0 and 1 (*nice!*)
- EV is related to noise ceiling (later!)

EV (EXPLAINABLE VAR.)

- EV is positive even for completely random datasets!

- EV is biased upwards!

- Bias correction: $EV^* = EV - \frac{1 - EV}{N - 1}$
|
number of repetitions

$$EV^* = EV - \frac{1 - EV}{N - 1}$$

MPWC (MEAN PAIRWISE CORR.)

- On average, how correlated are the responses from different repeats with each other?

$$MPWC = \langle \text{corr}(x_i(t), x_j(t)) \rangle_{i,j}$$

$$MPWC = \langle \text{corr}(x_i(t), x_j(t)) \rangle_{i,j}$$

MPWC

(MEAN PAIRWISE CORR.)

- MPWC is easy to explain!
- MPWC is unbiased
- MPWC is almost identical to bias-corrected EV (proof left as exercise...)

COHERENCE SPECTRUM

- First, **coherence** between two signals

cross-spectral density

$$C_{xy}(f) = \frac{|G_{xy}(f)|^2}{G_{xx}(f)G_{yy}(f)}$$

autospectral density

$$C_{xy}(f) = \frac{|G_{xy}(f)|^2}{G_{xx}(f)G_{yy}(f)}$$

COHERENCE SPECTRUM

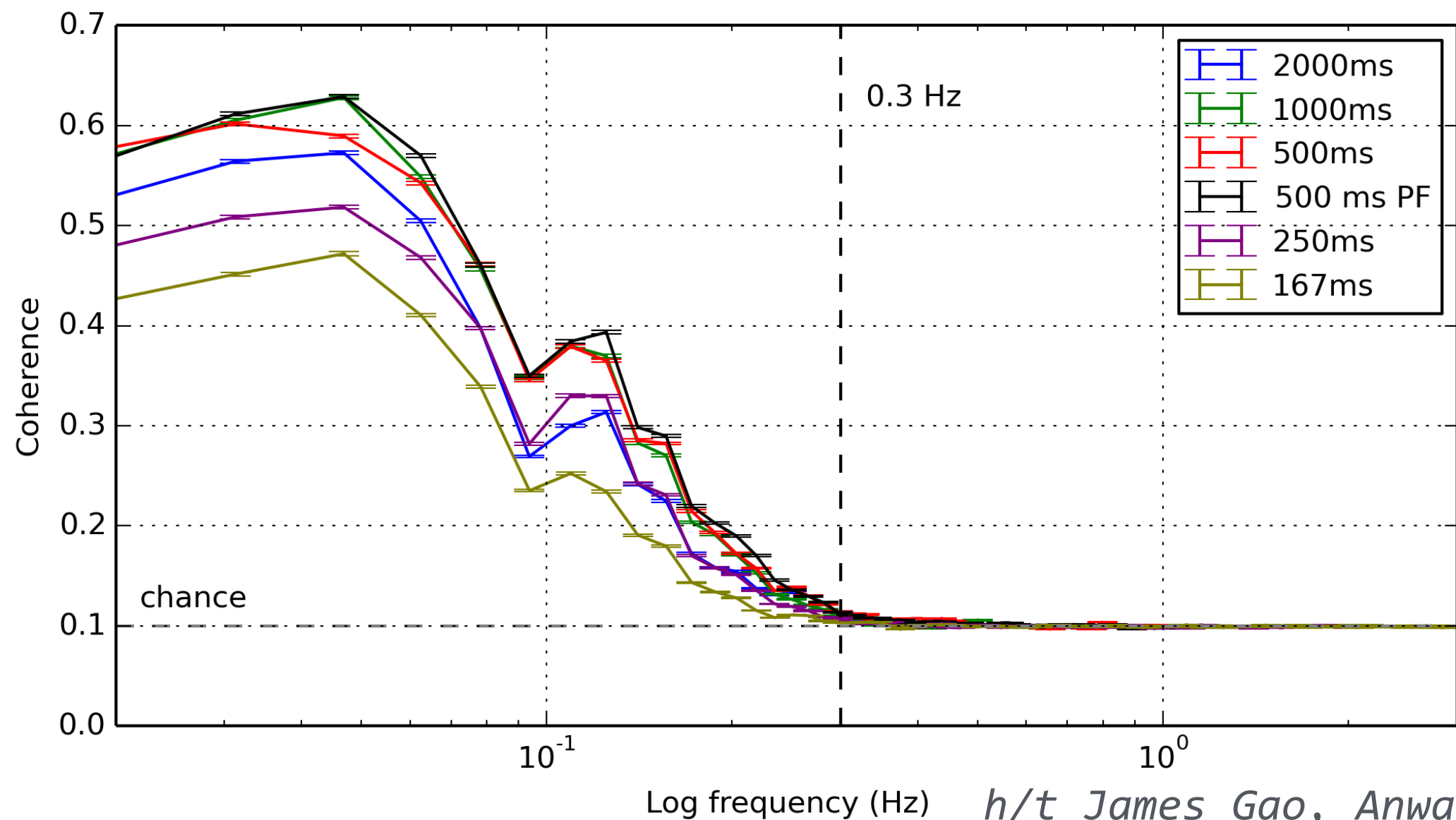
- When used to measure data quality, **coherence** gives repeatability at each frequency!

$$Coh(f) = \langle C_{\bar{x}, x_i}(f) \rangle_i$$

$$Coh(f) = \langle C_{\{\bar{x}, x_i\}}(f) \rangle_i$$

COHERENCE SPECTRUM

Example: fMRI data collected at different sampling rates



10 MINUTE BREAK

WHY IS REPEATABILITY IMPORTANT?

- Models require signal - test of a good paradigm
- Explanation for Type II error (false negatives)
- Provides a ceiling on predictive model performance (**noise ceiling**)

NOISE CEILING

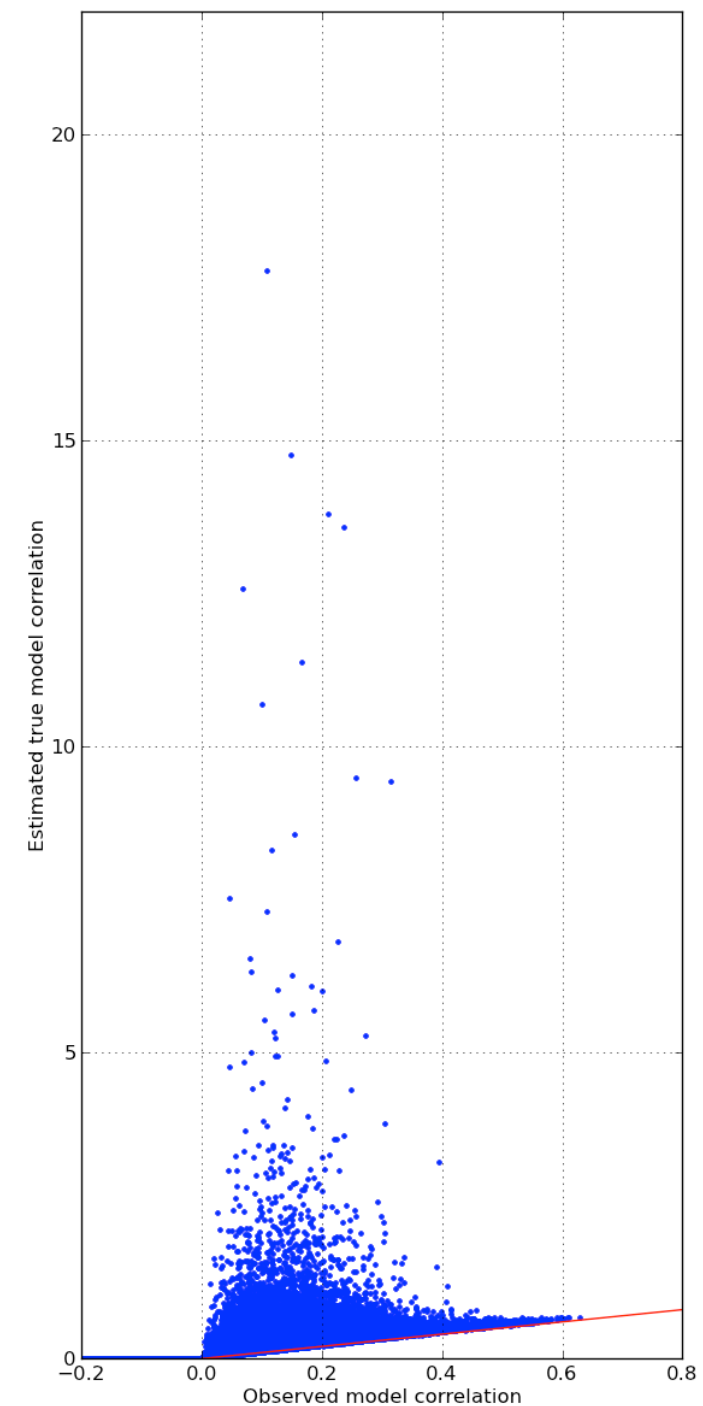
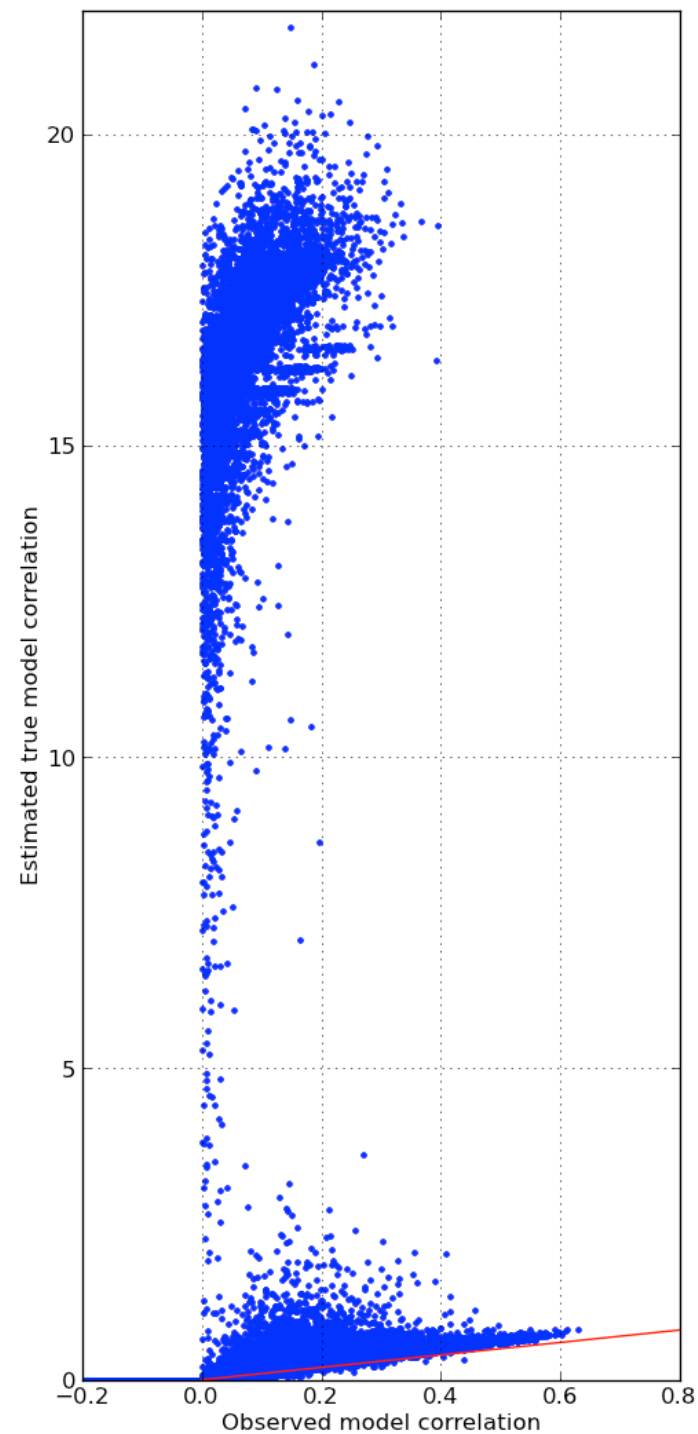
- Maximum predictive performance of a model is limited by repeatability of the data
- Even with n repeats, the average response is still noisy!

NOISE CEILING

- Many approaches:
 - Sahani & Linden (2003)
 - Hsu, Borst, & Theunissen (2004)
 - David & Gallant (2005)
 - **Schoppe et al. (2016)**
- All have problems with very noisy data (e.g. fMRI)

NOISE CEILING

David 2005 method HBT 2004 method



NOISE CEILING

- Recommended procedure given in:
Schoppe, Harper, Willmore, King, & Schnupp (2016)
- (But there's room to improve on this!)

IS REPEATABILITY EVERYTHING?

- ***No!***
- *Thought (fMRI) experiment: average together all the voxels in the brain.*

Does the resulting megavoxel have high repeatability? ***Yes!***

Is it useful? ***No!***

IS REPEATABILITY EVERYTHING?

- Repeatability is *GOOD* for comparing:
 - Across response channels (e.g. voxels) in same dataset
 - Across different types of stimuli
 - Across data acquisition methods where spatial and temporal resolution are preserved

IS REPEATABILITY EVERYTHING?

- Repeatability is *BAD* for comparing:
 - Across data acquisition methods where spatial or temporal resolution are *NOT* preserved

IS REPEATABILITY EVERYTHING?

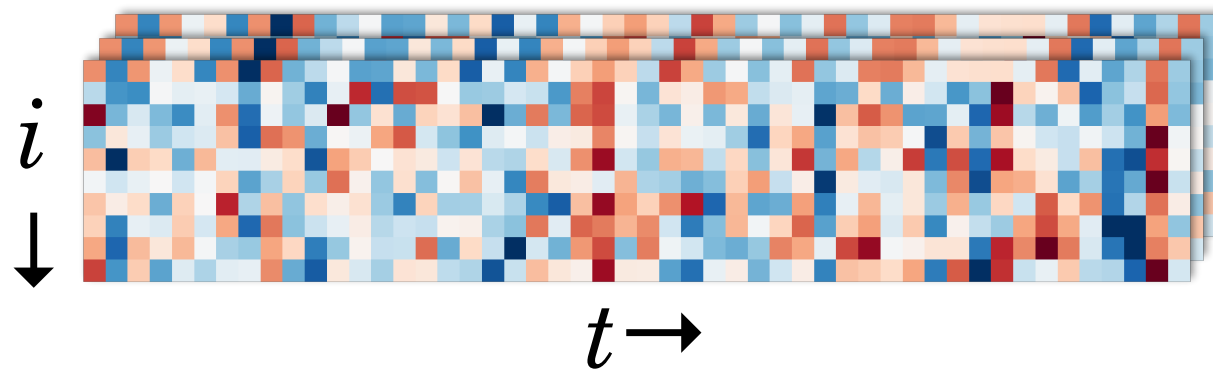
- Repeatability is susceptible to the *information trade-off problem*
- You can increase repeatability by sacrificing information
- Thus, repeatability can be “falsely” inflated

PREGUNTA

Can you think of a metric for data quality that could not be inflated by sacrificing information?

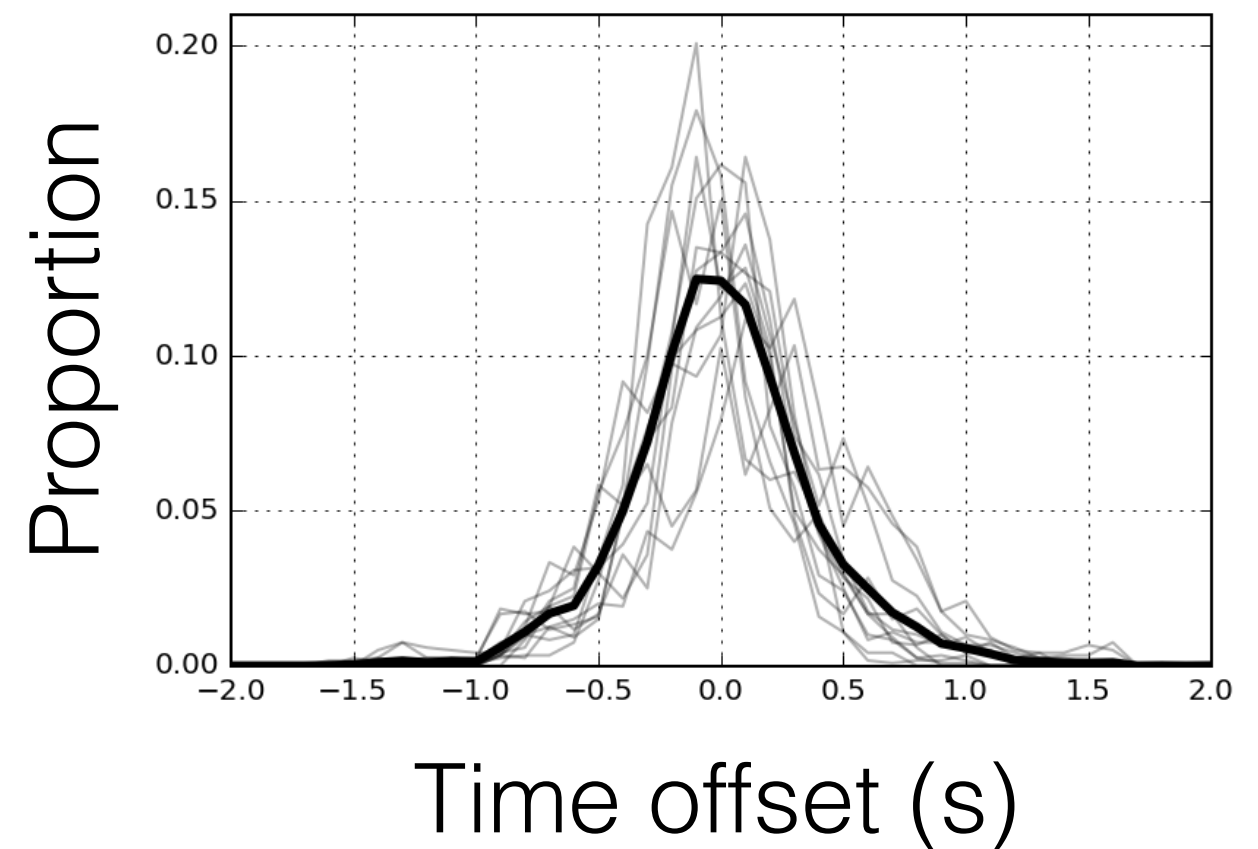
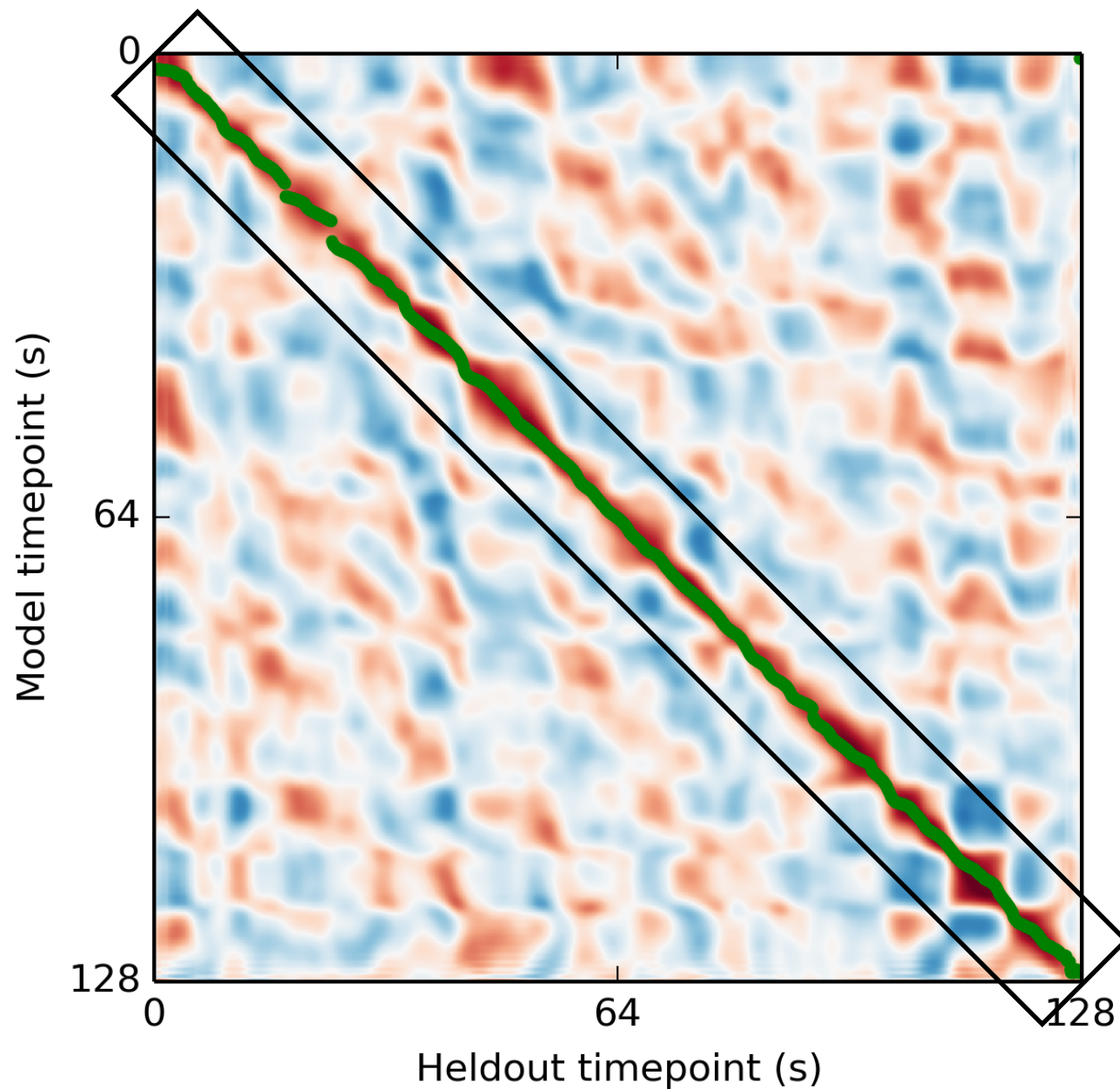
- **Please:**
 - Think about this question for 2 minutes
 - Turn to your neighbor and discuss what you thought of

TIMEPOINT CLASSIFICATION



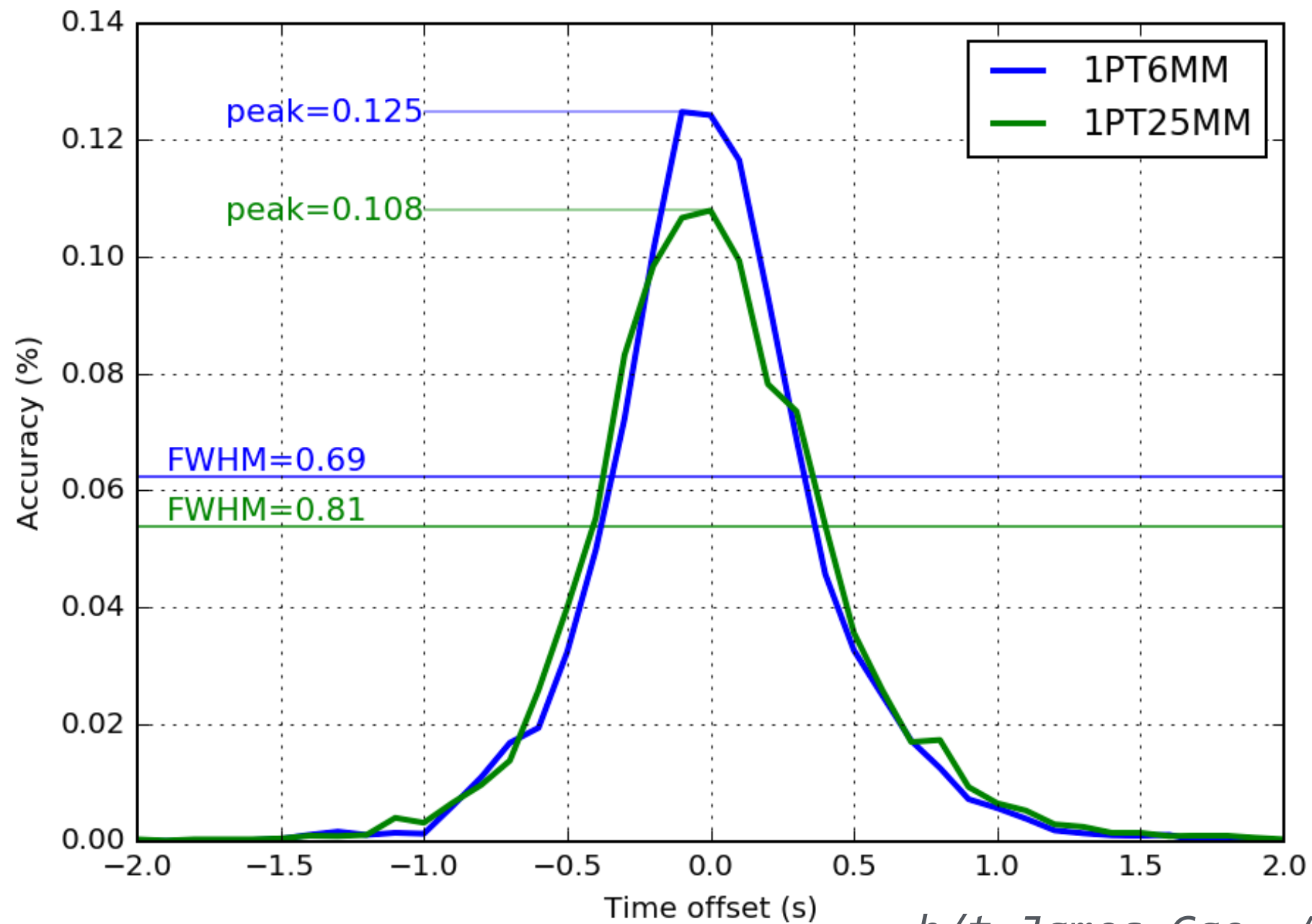
1. (Optional: temporally up-sample data)
2. Average $n - 1$ repeats
3. In n 'th repeat, take timepoint t^*
4. Decide which of T timepoints in average response best matches t^* (by correlation)

TIMEPOINT CLASSIFICATION



h/t James Gao, Anwar Nuñez

TIMEPOINT CLASSIFICATION



h/t James Gao, Anwar Nuñez

TIMEPOINT CLASSIFICATION

- A measure of how much information there is about the stimulus in the measured responses
- Perhaps a more absolute (& thus comparable) measure than repeatability

TAKEAWAYS

- Data quality is important
- Learn how to measure data quality
- Before doing an experiment, cycle:
 - (1) select data collection parameters,
 - (2) collect test data,
 - (3) assess data quality

THANKS