DATA QUALITY

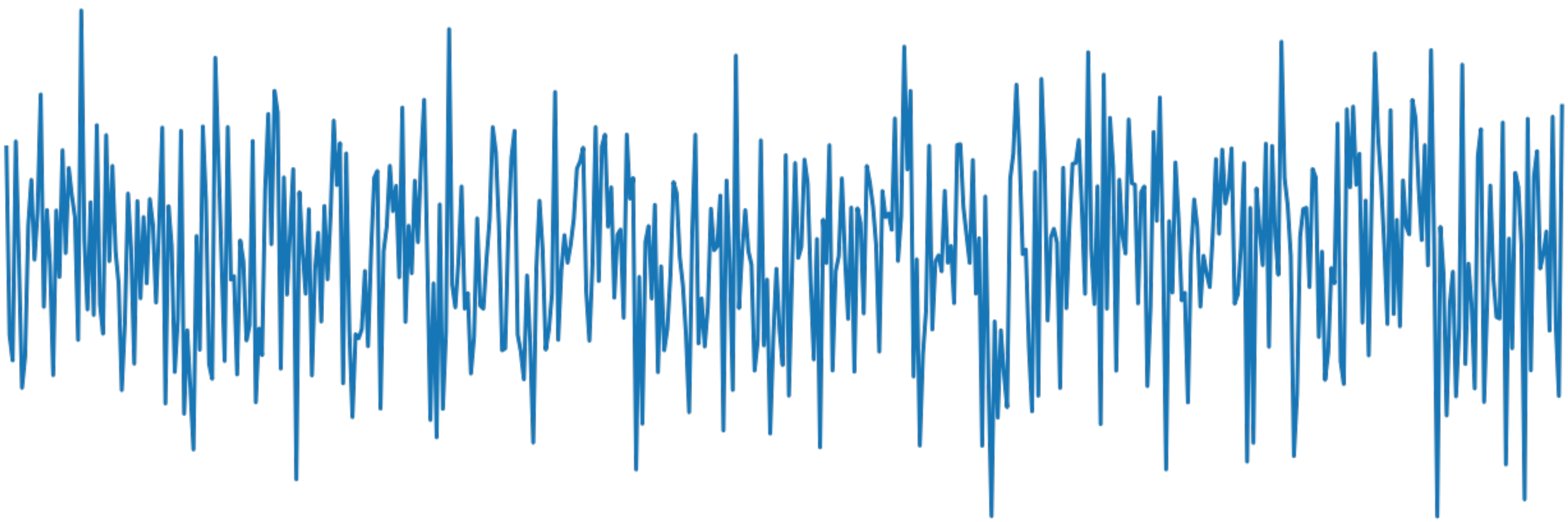
Prof. Alexander Huth

9/28/2017

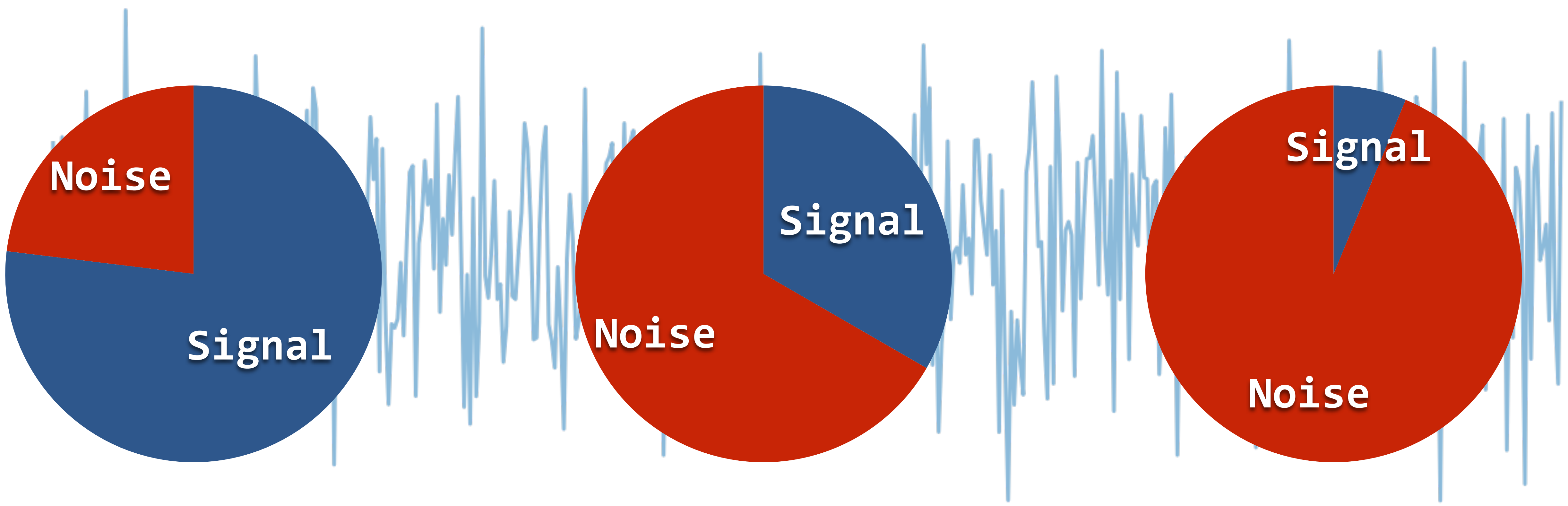
BEFORE YOU DO ANYTHING ELSE,

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

$$x_i(t) = \boxed{s(t)} + \boxed{\epsilon_i(t)}$$

measured response on i'th repetition

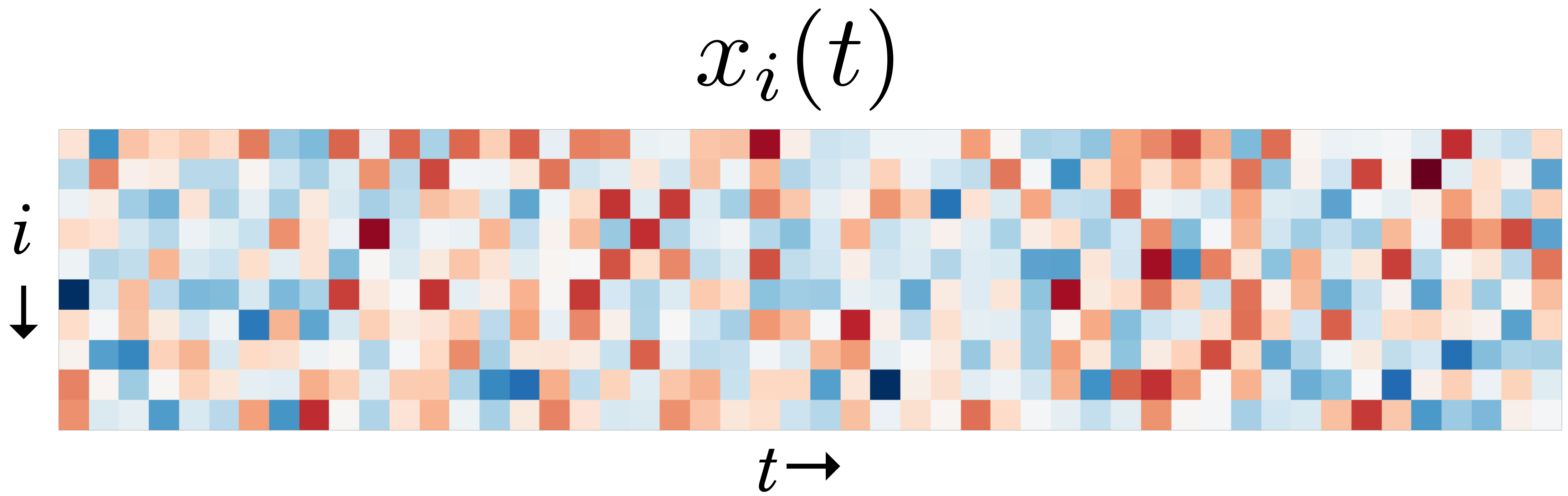
WHAT IS NOISE?

- (Assuming *stationarity* of the signal!)

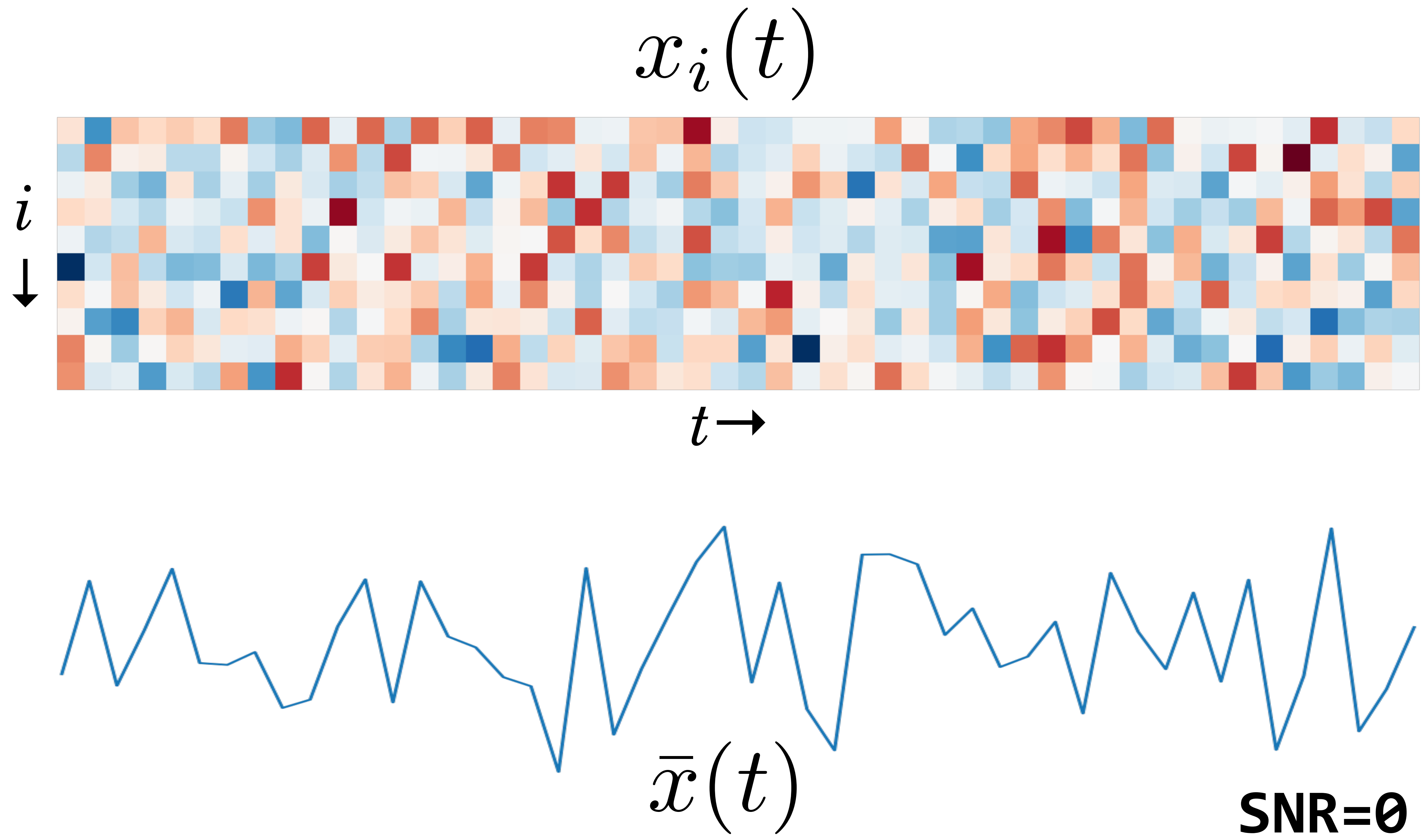
HOW DO WE KNOW?

- 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**

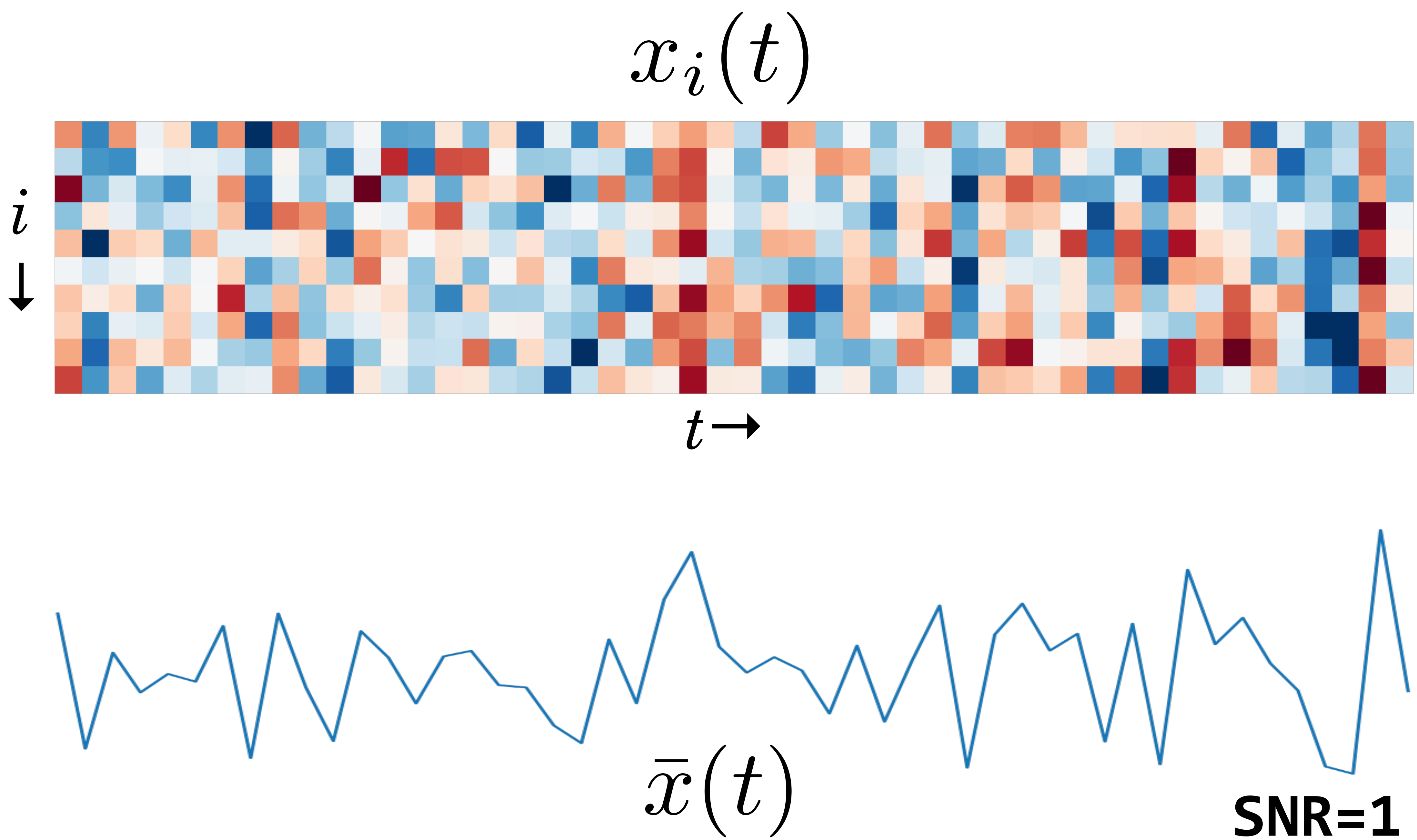
HOW DO WE KNOW?



HOW DO WE KNOW?



HOW DO WE KNOW?



QUESTION

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

METRICS FOR *REPEATABILITY*

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

add examples & images to this section!
compare the actual values

SIGNAL TO NOISE RATIO

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

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

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

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

SIGNAL TO NOISE RATIO

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

$$S\hat{N}R = \frac{\mathrm{var}(\bar{x}(t))}{\langle \mathrm{var}(x_i(t) - \bar{x}(t))\rangle_i}$$

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

SIGNAL TO NOISE RATIO

- **NB:** *Functional* **SNR** is <u>*not*</u> **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:

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

EV (EXPLAINABLE VAR.)

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

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

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

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

MPWC = \langle \mbox{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)}

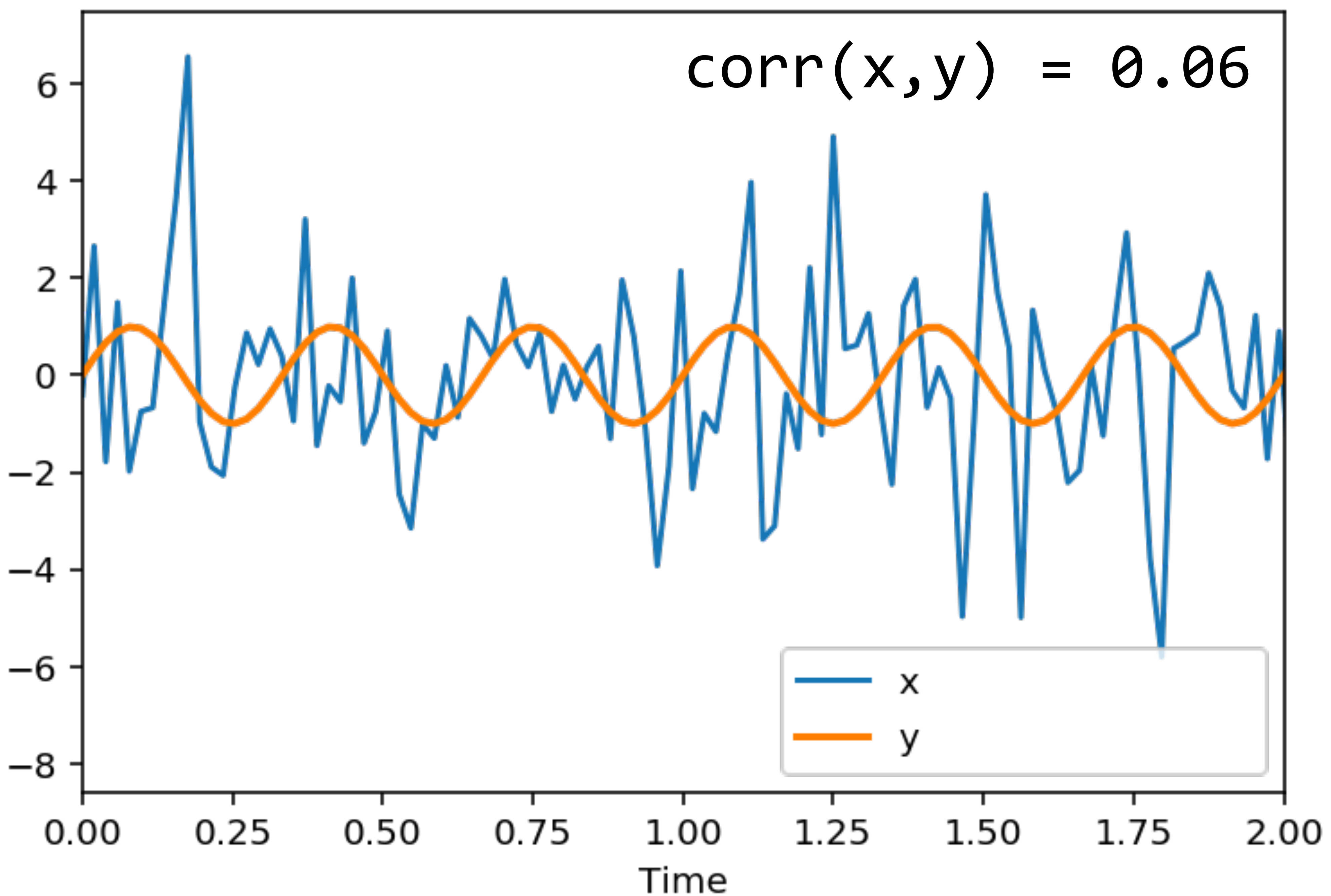
COHERENCE SPECTRUM

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

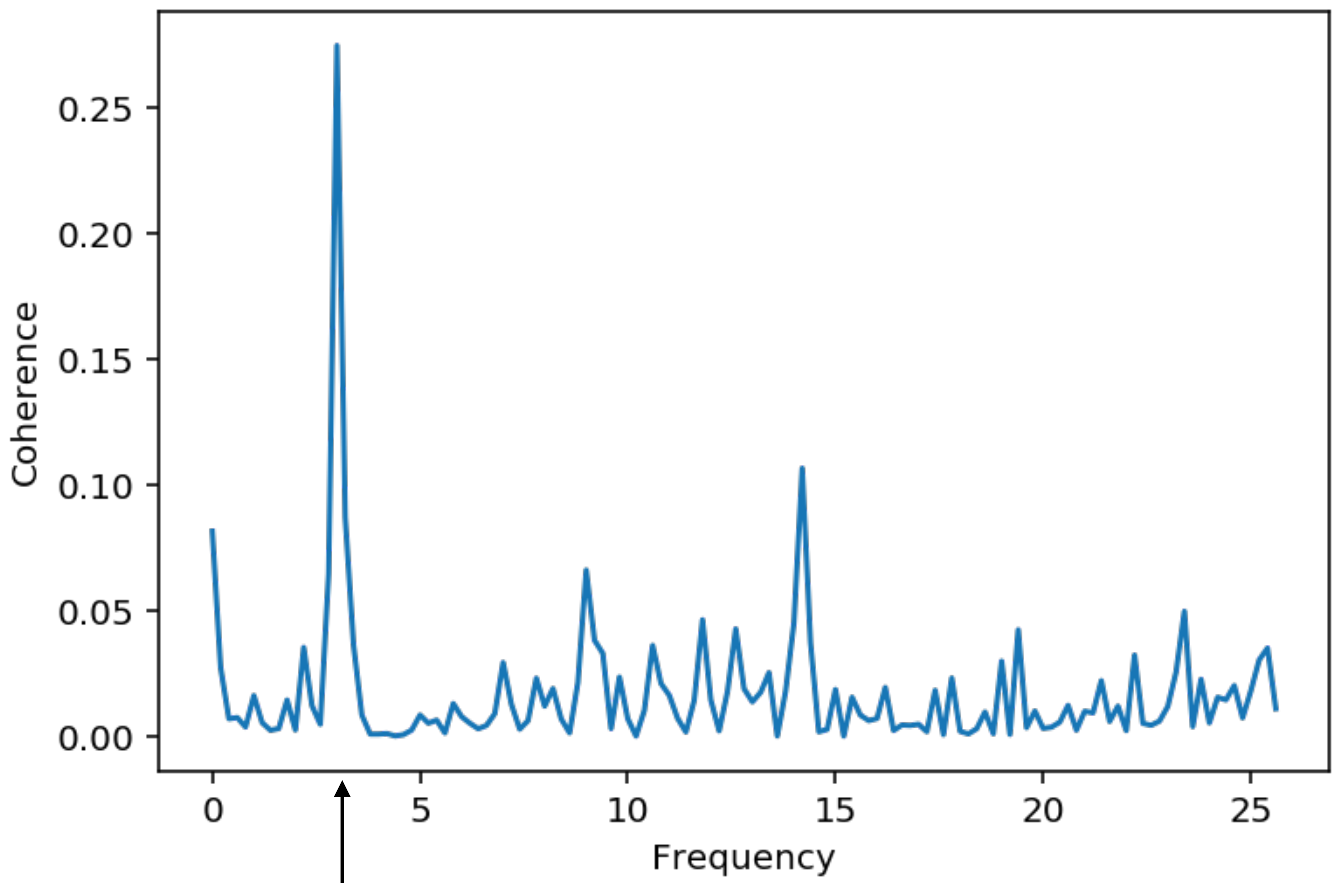
$$Coh(f) = \left\langle C_{\bar{x}, x_i}(f) \right\rangle_i$$

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

COHERENCE SPECTRUM

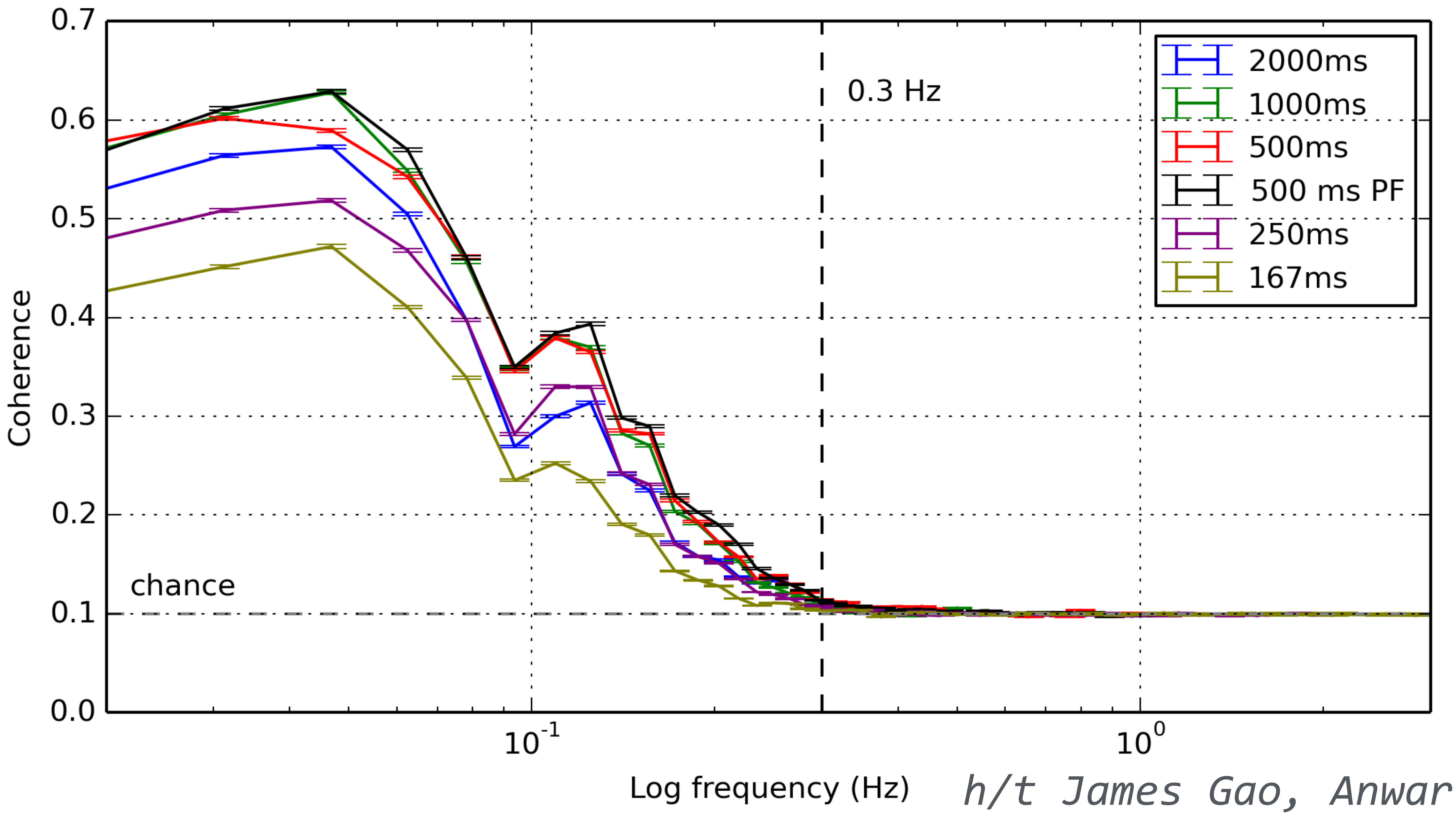


COHERENCE SPECTRUM



COHERENCE SPECTRUM

Example: fMRI data collected at different sampling rates



h/t James Gao, Anwar Nuñez

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**)

IS REPEATABILITY EVERYTHING?

- ***No!***

- *Thought (fMRI) experiment:* average together all the voxels in the brain.

  Does the resulting megavoxel have high repeatability?

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

QUESTION

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

NEXT TIME

- More about data quality:
  - Timepoint classification
  - Noise ceilings!