DATA QUALITY II

Prof. Alexander Huth

10/03/2017

LAST TIME

* Data = signal + noise

* How much is signal, how much is noise?

* What does it mean to be noise?

* Repeatability!

LAST TIME

* Methods for assessing repeatability
  * Signal-to-noise ratio (SNR)
  * Explainable variance (EV)
  * Mean pairwise correlation (MPWC)
  * Coherence spectrum

LAST TIME

* *Information trade-off problem*

  * You can sacrifice useful information to increase repeatability

TIMEPOINT CLASSIFICATION

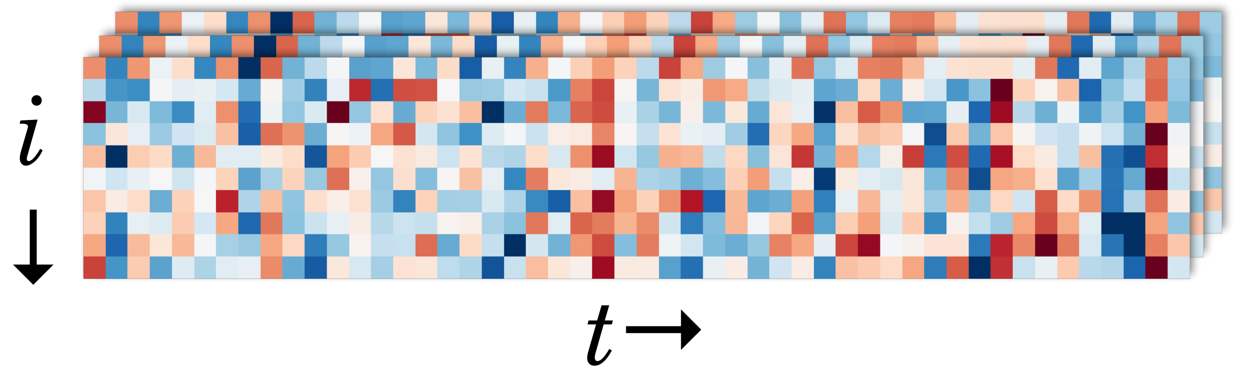
* Potential solution to information trade-off problem: test how much information is in the data

* Information about what?

* Information about *when each datapoint comes from in the stimulus*

h/t James Gao, Anwar Nuñez

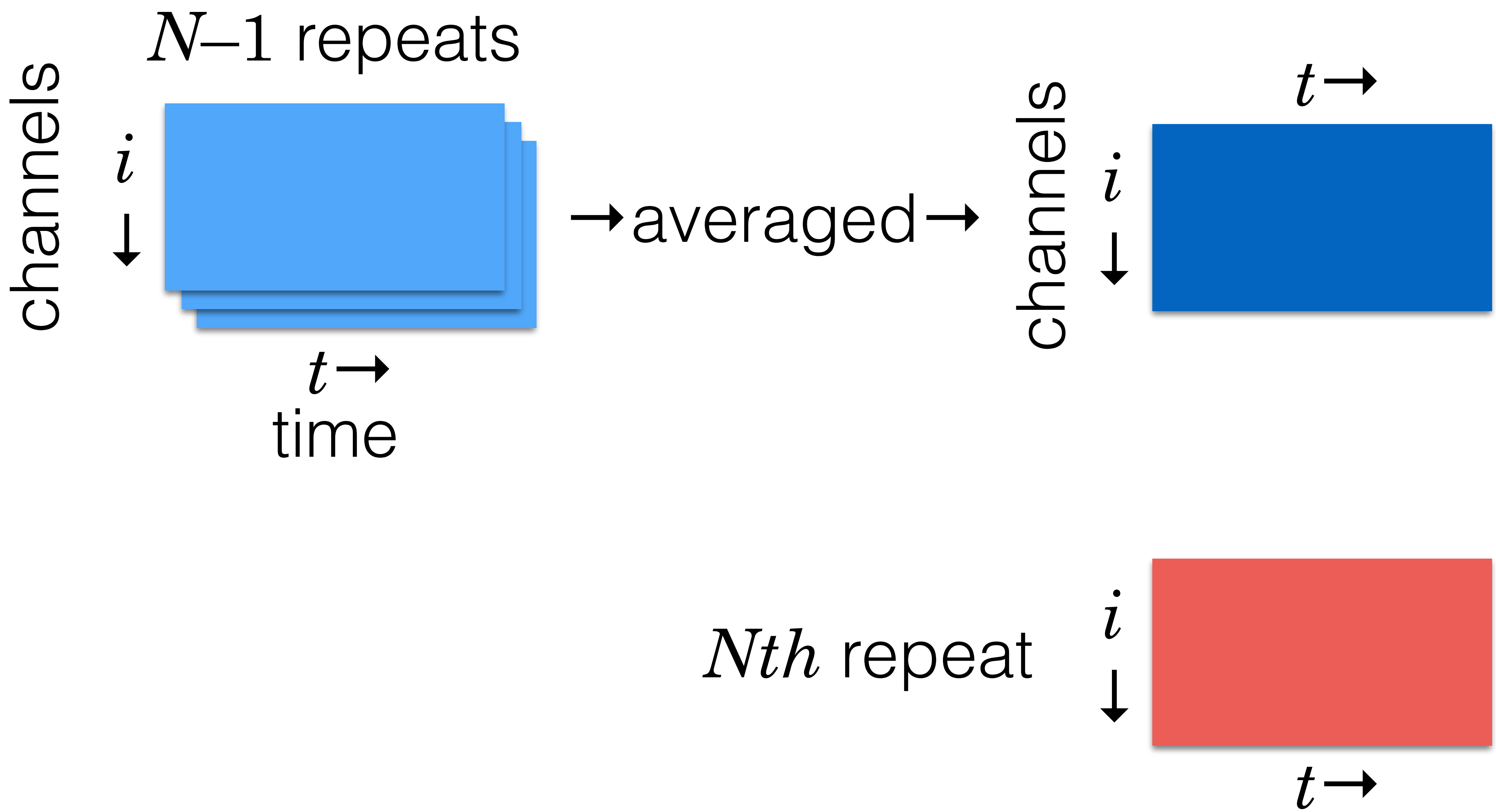
TIMEPOINT CLASSIFICATION



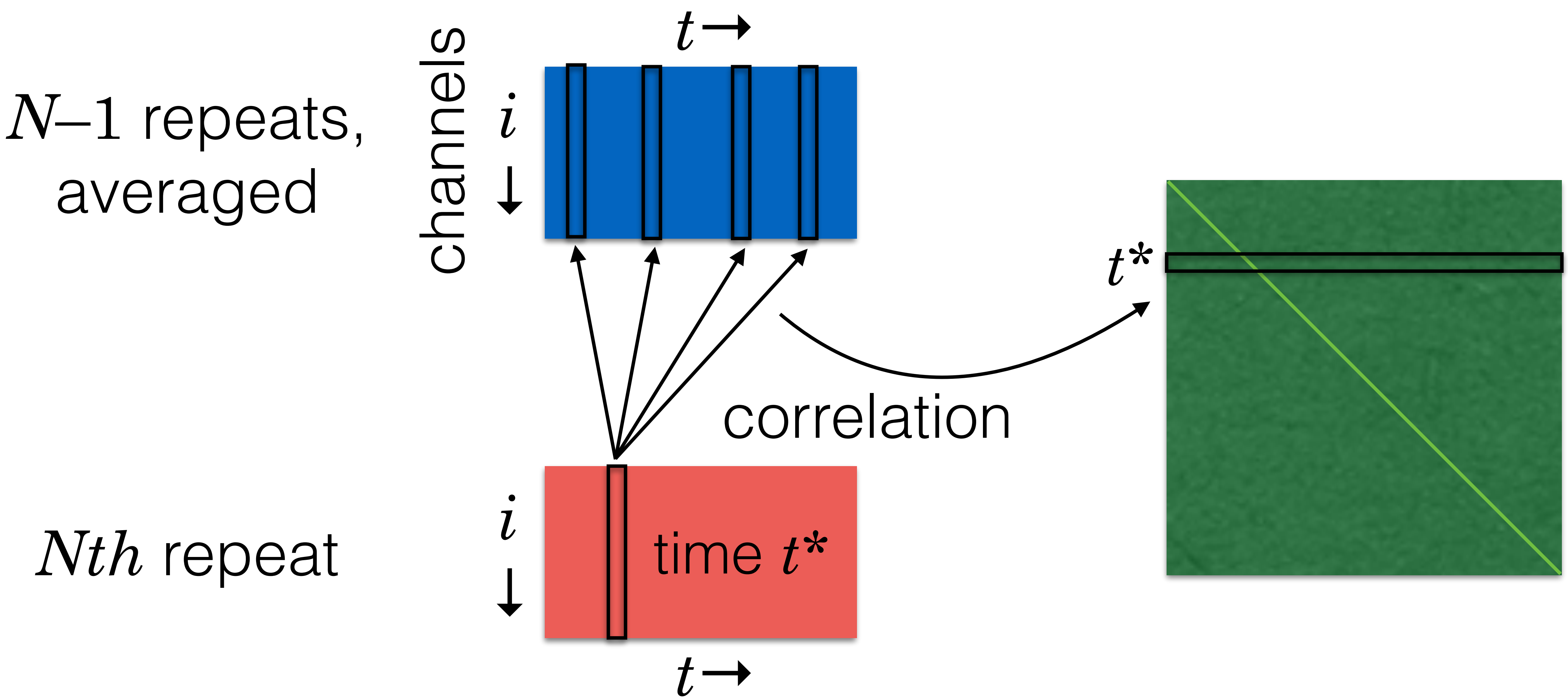
1. *(Optional: temporally up-sample data)*
2. Average $n-1$ repeats
3. In n'th repeat, take timepoint t^*
4. Decide which of T timepoints in average response best matches t^* (by correlation)

h/t James Gao, Anwar Nuñez

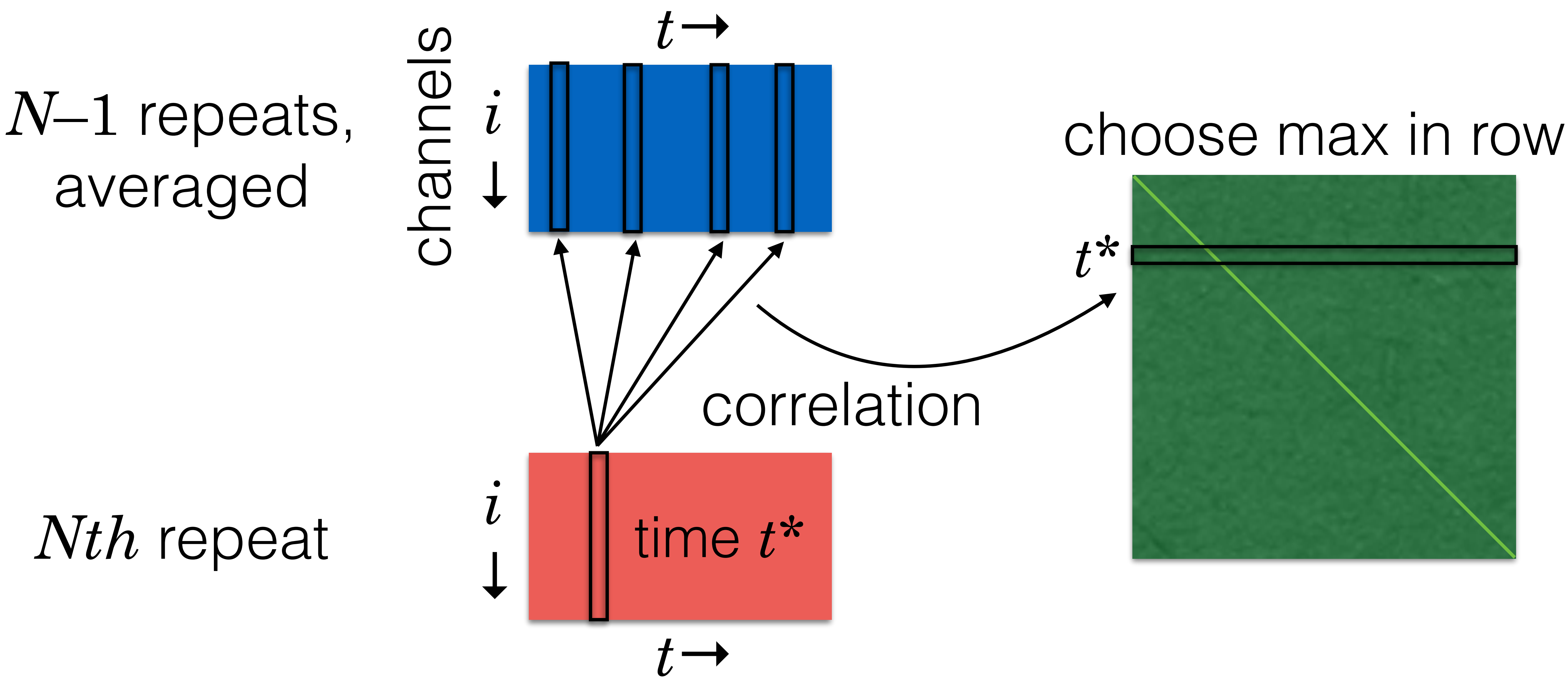
TIMEPOINT CLASSIFICATION



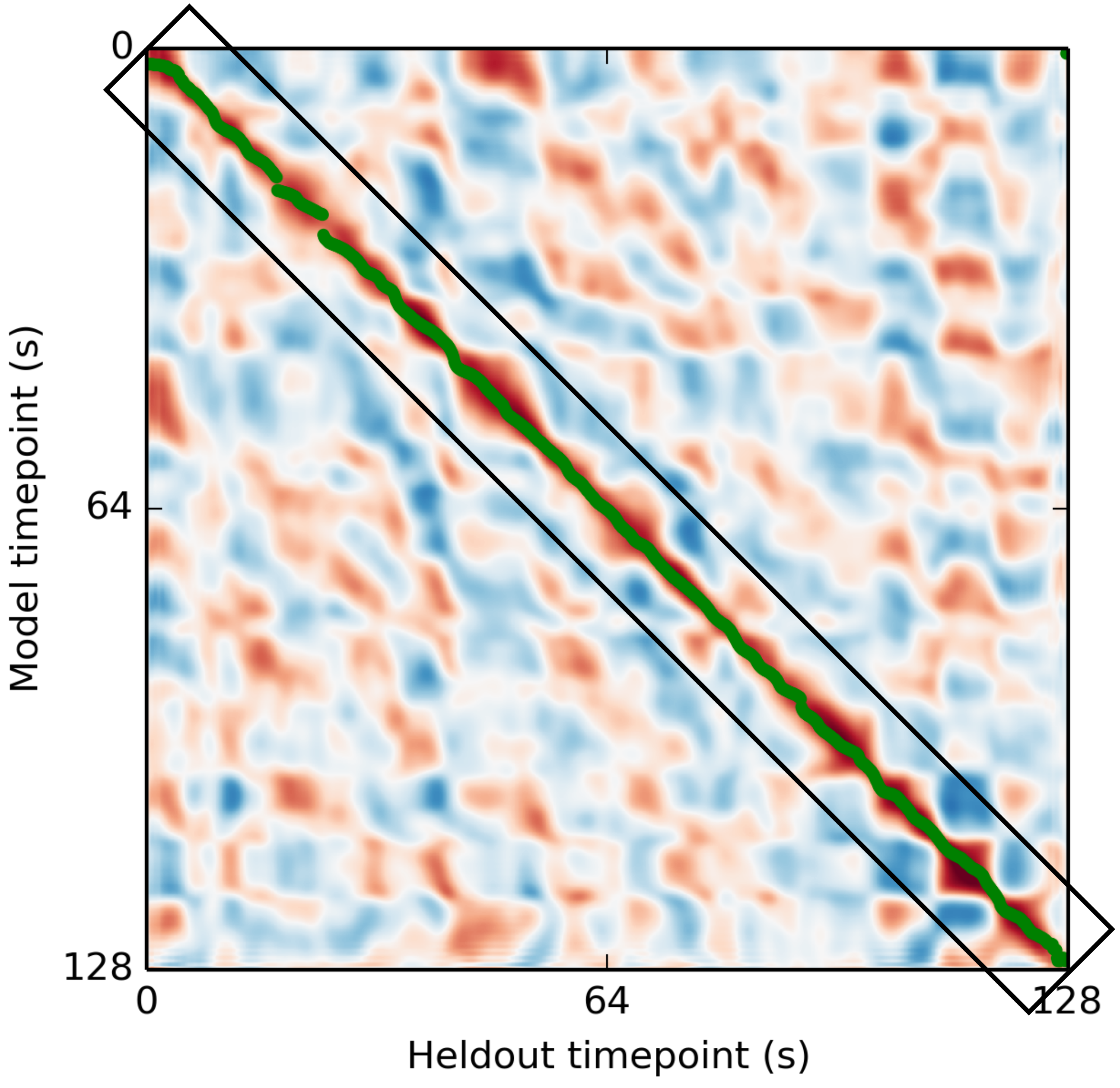
TIMEPOINT CLASSIFICATION

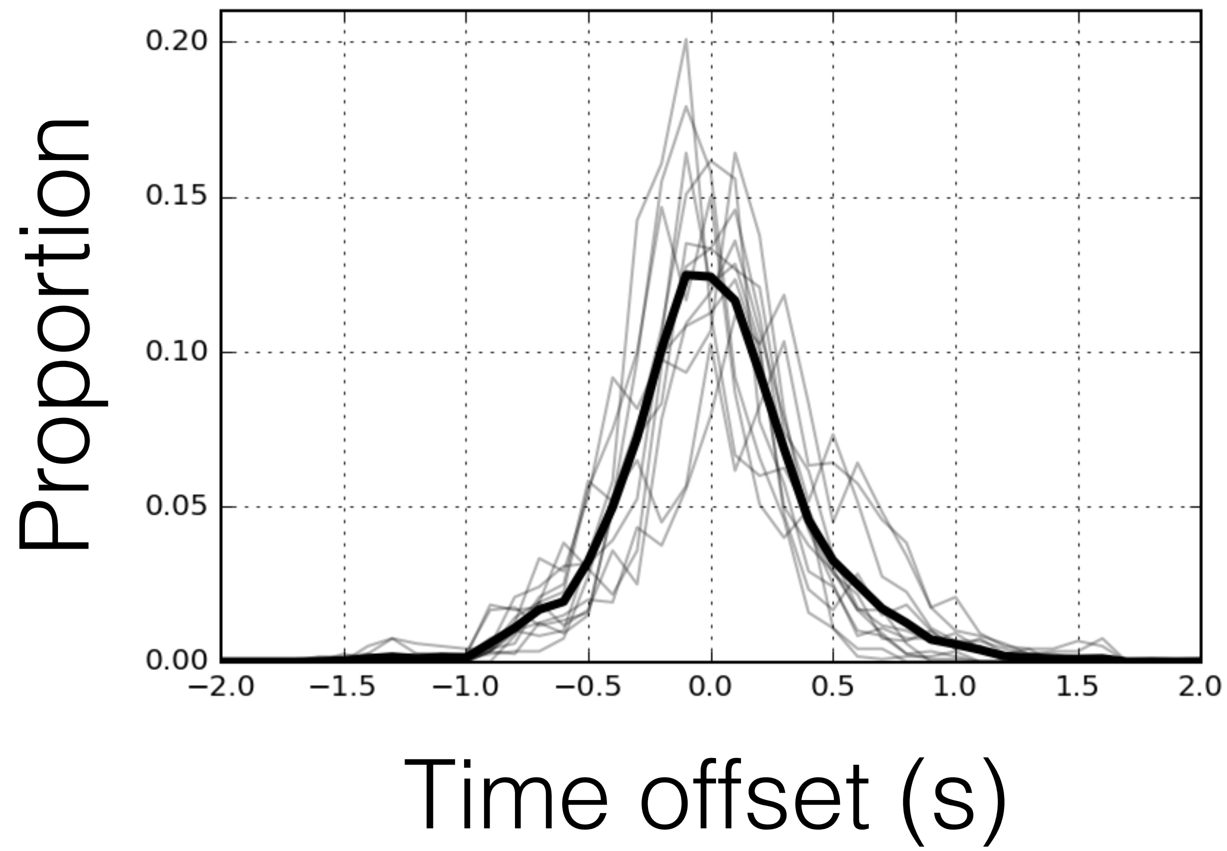


TIMEPOINT CLASSIFICATION



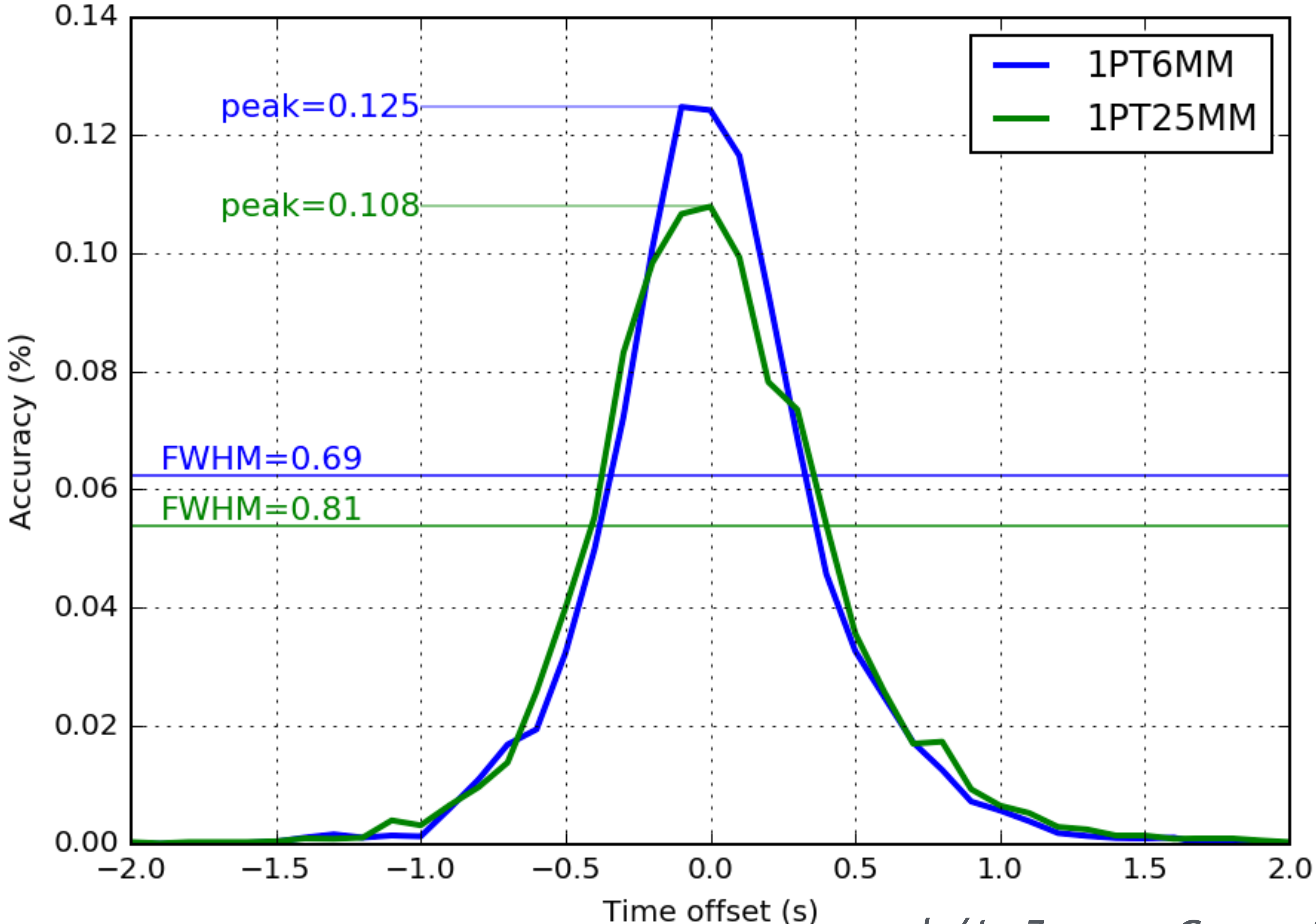
TIMEPOINT CLASSIFICATION





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

h/t James Gao, Anwar Nuñez

NOISE CEILING

* Assume that our data was generated by a linear process with Gaussian noise:

$$y = X\beta_{true} + \epsilon$$

y = X \beta_{true} + \eps

* What's the best we could possibly do at predicting new data?

NOISE CEILING

* Even if beta_true is known *exactly*, our best prediction would still be wrong

$$y = X\beta_{true} + \epsilon$$

$$\hat{y} = X\beta_{true}$$

$$y - \hat{y} = \epsilon$$

NOISE CEILING

* We can reduce the effect of the noise by averaging multiple trials together

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i = X\beta_{true} + \epsilon_N$$

\bar{y} = \frac{1}{N} \s
X \beta_{true} + \epsilo

* But that only reduces noise, does not eliminate it

NOISE CEILING

* We can reduce the effect of the noise by averaging multiple trials together

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = X\beta_{true} + \epsilon_N$$

athcal{N}(0, \sigma^2)
mathcal{N}(0,
N})

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \qquad \epsilon_N \sim \mathcal{N}(0, \frac{\sigma^2}{N})$$

Gaussian, 0 mean, variance sigma squared   *Gaussian, 0 mean, variance sigma squared / N*

NOISE CEILING

* Maximum predictive performance of a model is limited by size of noise, and thus by repeatability of the data

NOISE CEILING

CC = \mbox{corr}(y, \hat{y
\frac{\mbox{cov}(y, \hat{y}
{\sqrt{\mbox{var}(y) \mbox

* Suppose we quantify *predictive performance* as the correlation between predicted and actual (averaged) responses

$$CC = \mathrm{corr}(y, \hat{y}) = \frac{\mathrm{cov}(y, \hat{y})}{\sqrt{\mathrm{var}(y)\mathrm{var}(\hat{y})}}$$

NOISE CEILING

* Can we find *CCmax*, the maximum possible performance we should be able to get with our noisy data?

$$CC = \mathrm{corr}(y, \hat{y}) = \frac{\mathrm{cov}(y, \hat{y})}{\sqrt{\mathrm{var}(y)\mathrm{var}(\hat{y})}}$$

NOISE CEILING

* Can we find *CCmax*, the maximum possible performance we should be able to get with our noisy data?

$$CC_{max} = \frac{1}{\sqrt{1 + \frac{1}{N} SNR^{-1}}}$$

```
CC_{max} = \frac{1}{\sc
{N}SNR^{-1}}}
```

For derivation see Hsu et al. 2004, Schoppe et al., 2016

NOISE CEILING

* Using CCmax, we can define the “normalized” correlation coefficient:

$$CC_{norm} = \frac{CC}{CC_{max}}$$

CC_{norm} = \frac{CC}{CC_{

NOISE CEILING

* Using CCmax, we can define the "normalized" correlation coefficient:

$$CC_{norm} = \frac{\mathrm{cov}(y, \hat{y})}{\sqrt{\mathrm{var}(\hat{y}) SP}}$$

$$SP = \frac{1}{N-1}\left(N\mathrm{var}(y) - \frac{1}{N}\sum_{i=1}^{N}\mathrm{var}(y_i)\right)$$

"signal power": bias-corrected version of signal variance

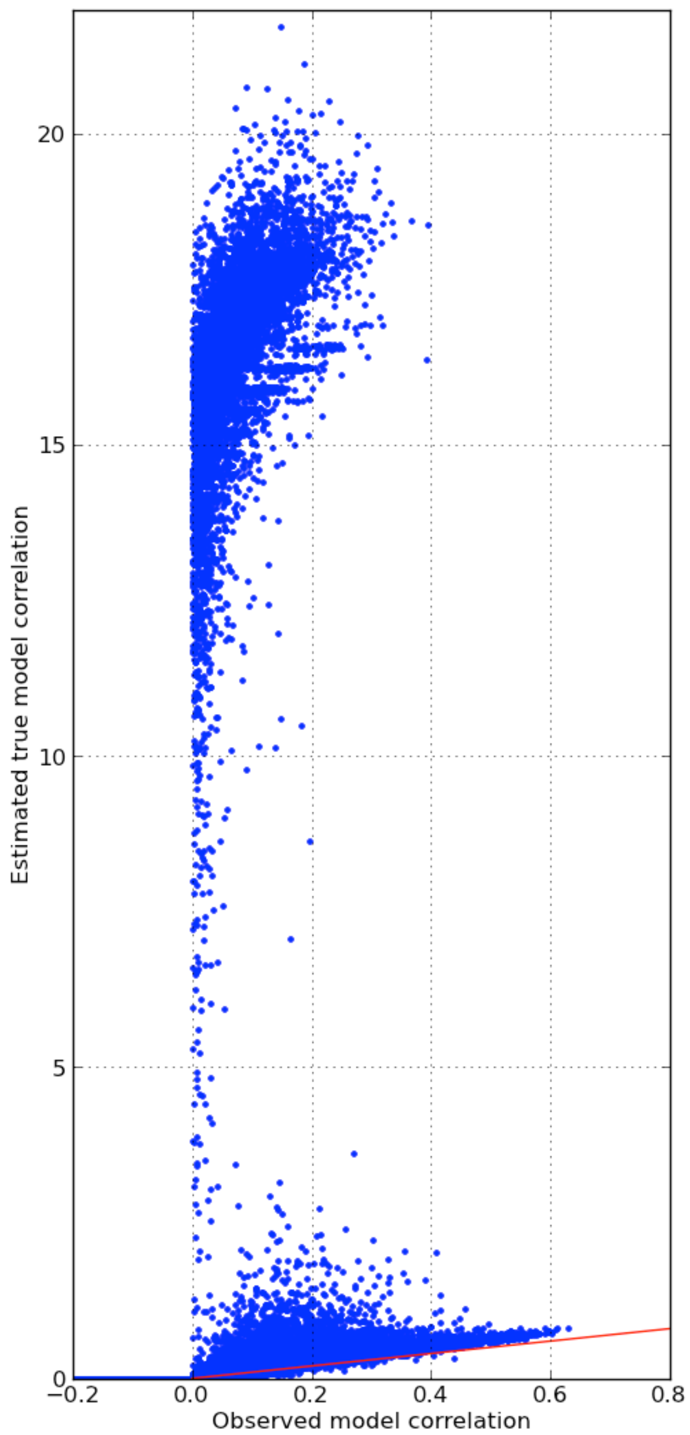
CC_{norm} = \frac{\mbox{
\hat{y})}{\sqrt{\mbox{var}(

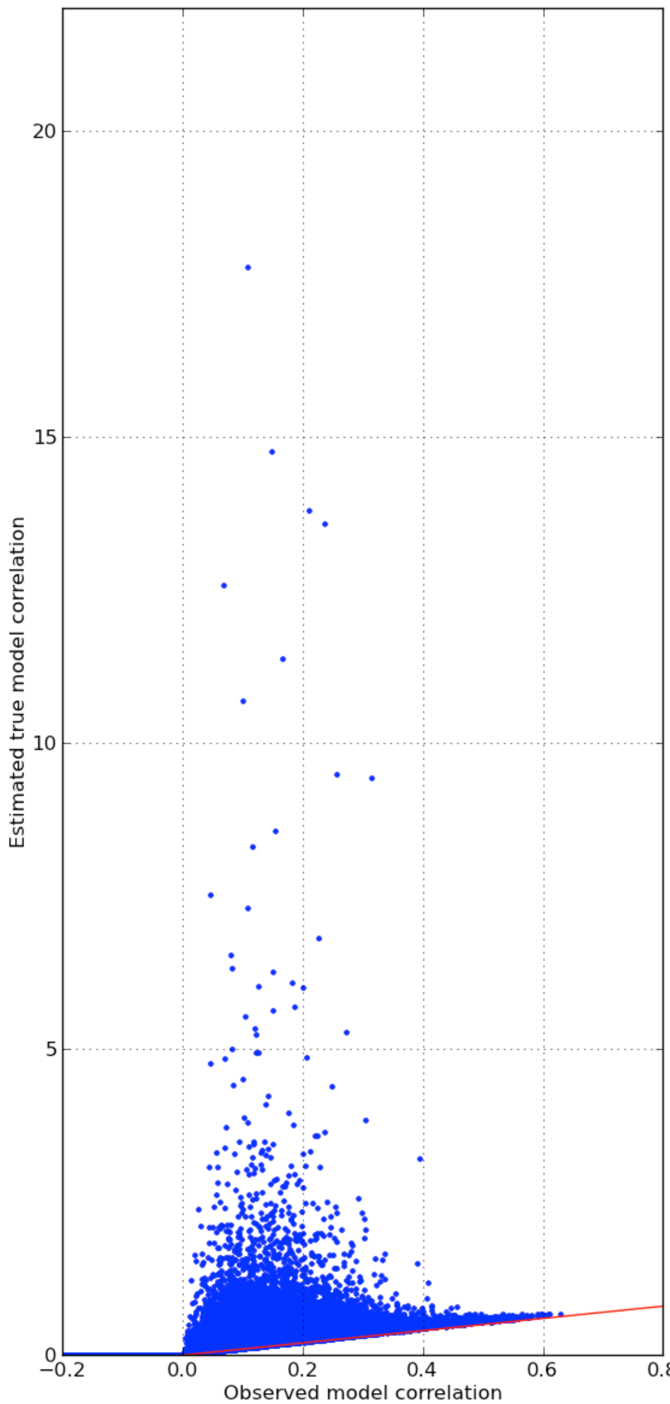
NOISE CEILING

- 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!)

NOISE CEILING

* Knowing the noise ceiling is important

* Because it can save you from being overly pessimistic

HOMEWORK

* Posted **tonight** on github:
  https://github.com/alexhuth/n4cs-fa2017

* Due **10/17**

NEXT TIME

* Thursday: **GUEST LECTURE by Rishi Chaudhuri**

* Next Tuesday: Feature spaces!