Psych 253
Advanced statistical modeling

Reliability (Part 1)

Daniel Yamins
Wu Tsai Neurosciences Institute
Departments of Psychology and Computer Science
Stanford University

Russ Poldrack
Department of Psychology
Stanford University

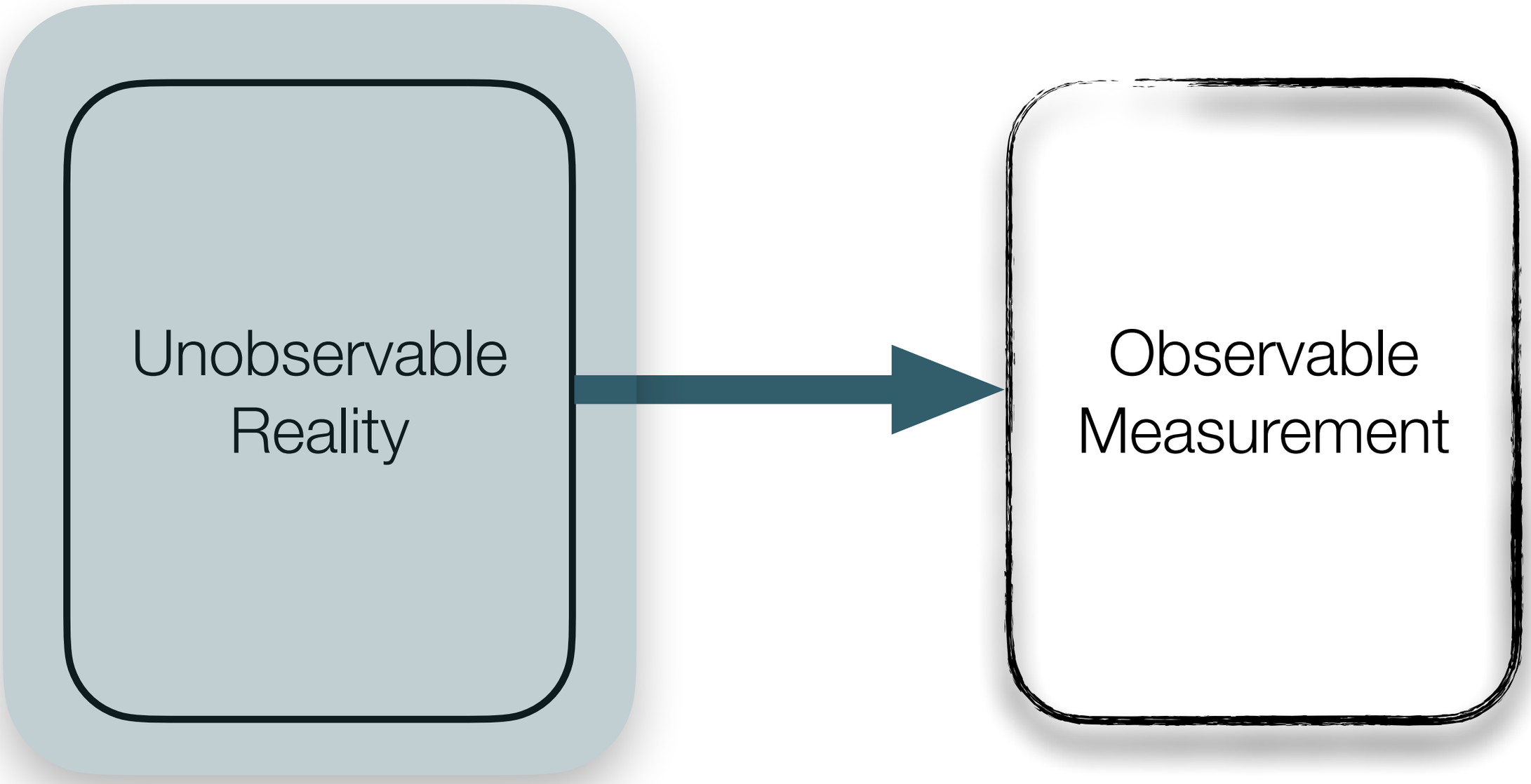
Science is ultimately about measurement and modeling

In the last session we asked what makes a model "good".

In this session, we turn to the question of what makes a measurement "good"?

Measurement vs reality

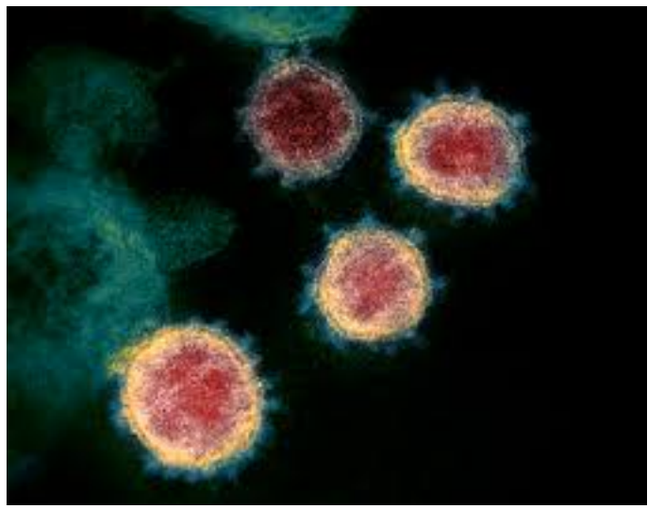
Rarely can we directly observe the underlying
structure of the world - we must use measurement to do so



Let's say a person presents to their physician with symptoms of COVID-19: dry cough, fever, and fatigue

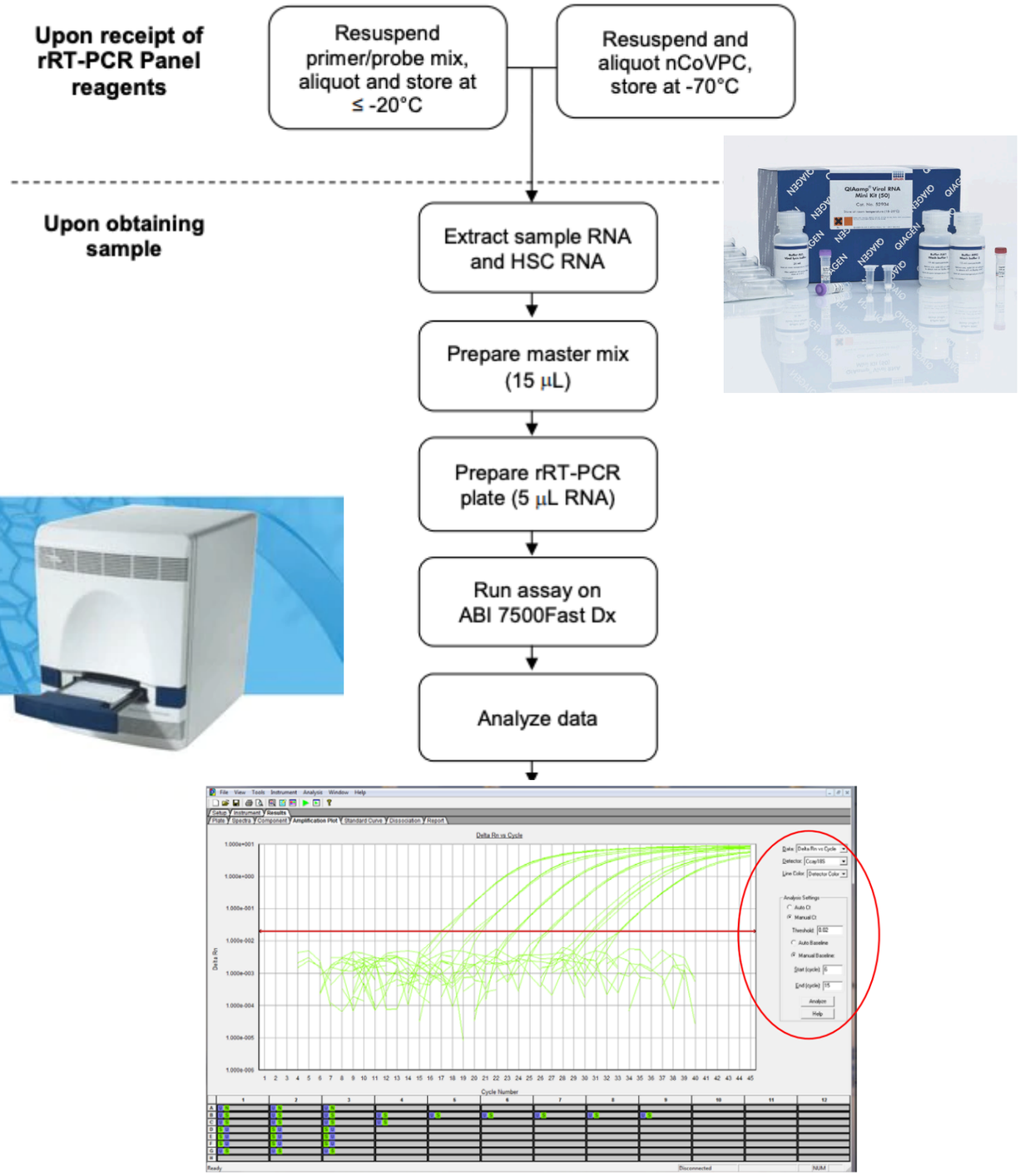
Reality

The person is either infected with SARS-CoV-2, or not



https://en.wikipedia.org/wiki/File:SARS-CoV-2_49534865371.jpg

Measurement



https://www.fda.gov/media/134922/download

"all measurement is befuddled by error" (McNemar, 1946)

There are two types of error:

Systematic error (or *bias*):
attributable to some particular causal factor

Example: differences in viral detection depending on the particular primers that are used for the test

Random error (or *noise*):
error due to any of a large possible number of unknown sources - often Gaussian (cf. central limit theorem)

Systematic error can (often) be eliminated - random error usually cannot

Reliability vs. Validity

- Validity:
  - Does the measurement measure the concept that we think it measures?

The Washington Post
Democracy Dies in Darkness

Inside the coronavirus testing failure: Alarm and dismay among the scientists who sought to help

On Feb. 8, when lab technicians for New York City's health department ran the test on samples that contained the virus, they saw on their computer screens a logarithmic curve sloping upward, indicating the virus was present. The problem was, they saw something similar when they ran the test on distilled water that contained no trace of the virus.

https://www-washingtonpost-com.stanford.idm.oclc.org/investigations/2020/04/03/coronavirus-cdc-test-kits-public-health-labs/?arc404=true

Reliability vs. Validity

- Reliability:
  - How precisely does the measurement quantify the thing that it measures?
- A measurement may be unreliable for several reasons
  - The underlying process could be fundamentally noisy
    - e.g. Reaction time measurements
  - The measurement procedure could be flawed
    - e.g. "guess my height"

Quantifying reliability: The data as a model of itself

If a measurement is reliable, then we should be able to use part of the data to accurately predict the values of the rest of the data

Example: Barratt Impulsiveness Scale (BIS-11)

DIRECTIONS: People differ in the ways they act and think in different situations. This is a test to measure some of the ways in which you act and think. Read each statement and put an X on the appropriate circle on the right side of this page. Do not spend too much time on any statement. Answer quickly and honestly.

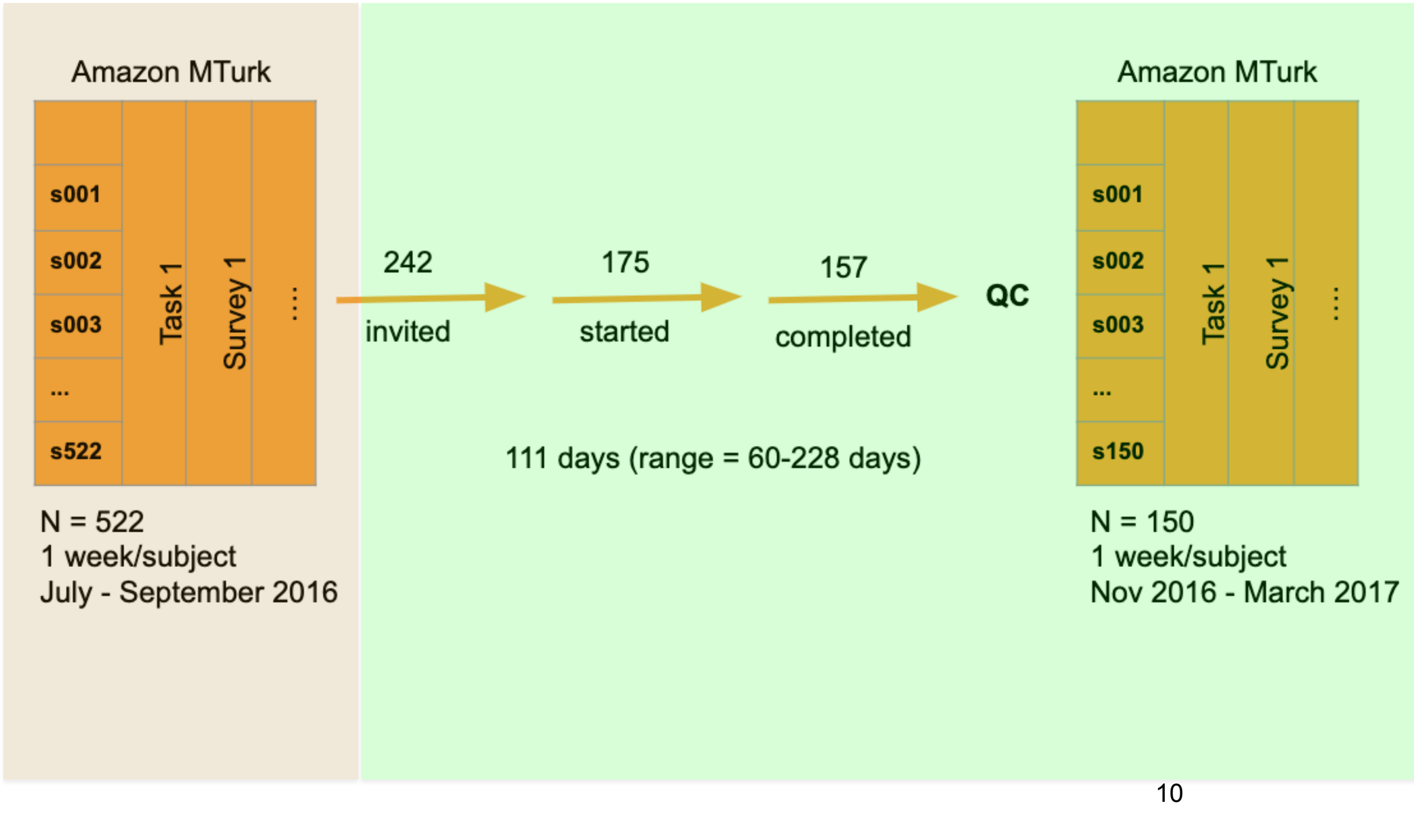
	① Rarely/Never	② Occasionally	③ Often	④ Almost Always/Always
1	I plan tasks carefully.			① ② ③ ④
2	I do things without thinking.			① ② ③ ④
3	I make-up my mind quickly.			① ② ③ ④
4	I am happy-go-lucky.			① ② ③ ④
5	I don't "pay attention."			① ② ③ ④
6	I have "racing" thoughts.			① ② ③ ④
7	I plan trips well ahead of time.			① ② ③ ④
8	I am self controlled.			① ② ③ ④
9	I concentrate easily.			① ② ③ ④
10	I save regularly.			① ② ③ ④
11	I "squirm" at plays or lectures.			① ② ③ ④

. . .

SRO project: Acquisition structure

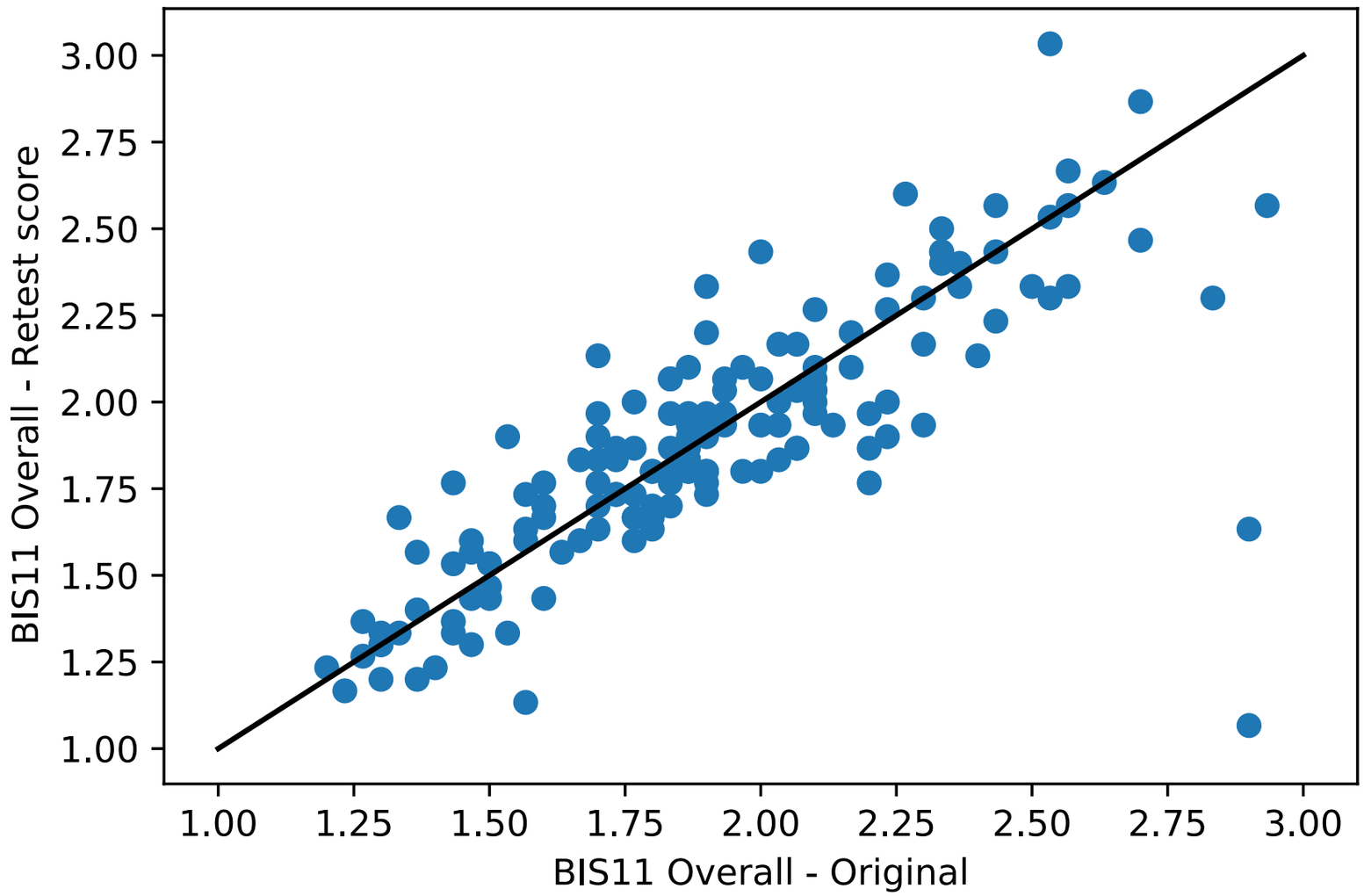
Initial test

Retest



N = 522
1 week/subject
July - September 2016

N = 150
1 week/subject
Nov 2016 - March 2017


BIS11 Overall - Retest score
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
BIS11 Overall - Original

$r = 0.83$

Classical test theory: the "true score"

Let's say we want to know the relationship between two psychological traits:

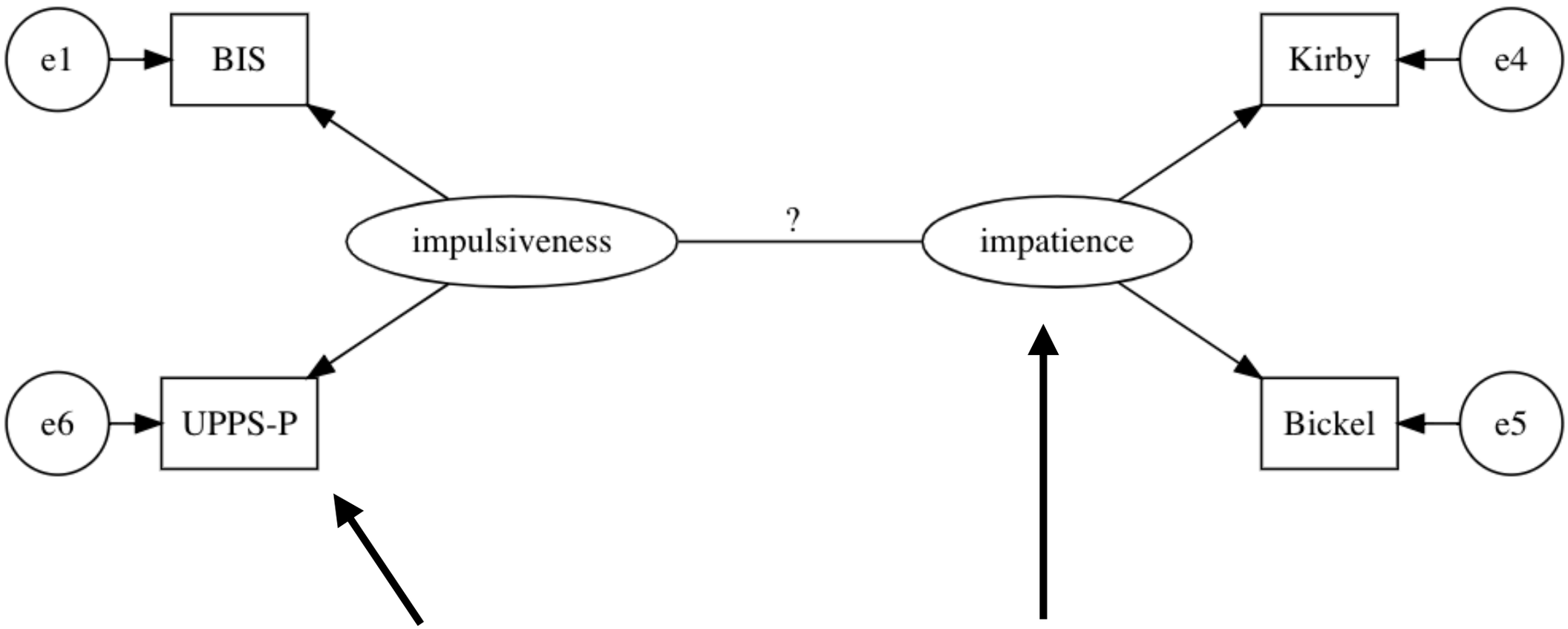
- *impulsiveness*
  - Defined as acting without attention to long-term goals or potential consequences
- *impatience*
  - Defined as a tendency to choose smaller immediate rewards over larger long-term rewards

Under classical test theory,
we assume that each person has an (unobservable) *true score* for each of these traits

Any measurement will be a noisy reflection of this true score

Modern test theory: Path diagrams

A graphical description of the relationships between observed and latent variables
- any observed variable reflects one or more latent "constructs" along with error



"Manifest variables" (measured)

"Latent variables" (unobservable)

Reliability

$$\sigma^2_{observed} = \sigma^2_{true} + \sigma^2_{error}$$

assumptions:

- observed measurement is only related to a single "true" latent variable
- error is uncorrelated with true value

Reliability refers to the proportion of variance in the measurement that is attributable to the true score:

$$\rho_Y = \frac{\sigma^2_{true}}{\sigma^2_Y} = \frac{\sigma^2_{true}}{\sigma^2_{true} + \sigma^2_{error}}$$

That is, how precisely does our measure reflect the underlying reality versus error variance?

Attenuation

The true correlation between estimates is attenuated by the geometric mean of their reliabilities

$$\beta, \theta : \text{true scores}$$

$$\text{estimates} : \hat{\beta} = \beta + \epsilon_\beta \qquad \hat{\theta} = \theta + \epsilon_\theta$$

$$\rho_{\hat{\beta},\hat{\theta}} = \rho_{\beta,\theta} * \sqrt{\rho_{\hat{\beta}} \rho_{\hat{\theta}}}$$

PERSPECTIVES ON PSYCHOLOGICAL SCIENCE

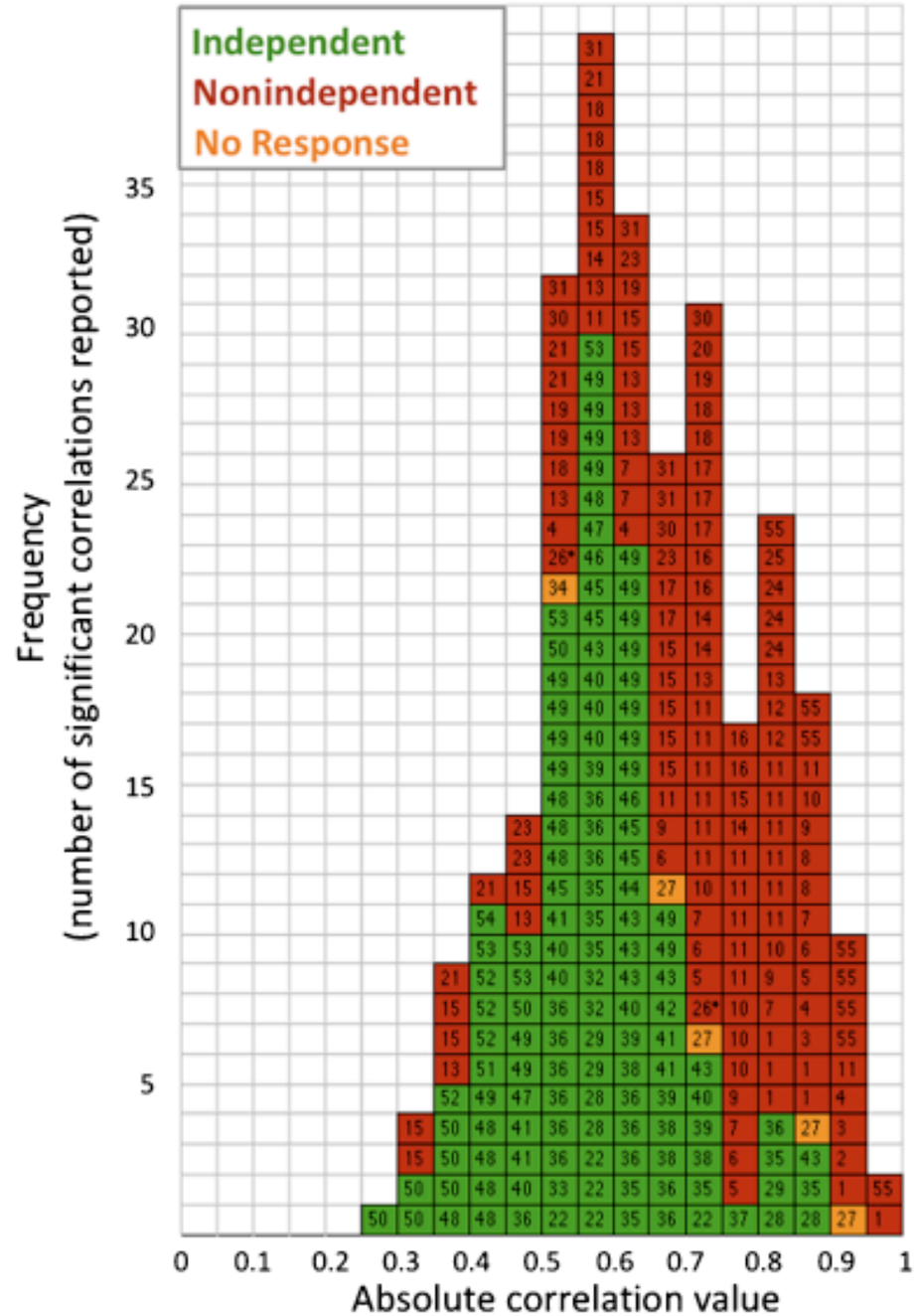
Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition[1]

Edward Vul,[1] Christine Harris,[2] Piotr Winkielman,[2] & Harold Pashler[2]

[1]Massachussetts Institute of Technology and [2]University of California, San Diego

[1]This article was formerly known as "Voodoo Correlations in Social Neuroscience."

Thus, the reliabilities of two measures provide an upper bound on the possible correlation that can be observed between the two measures (Nunnally, 1970).[3]

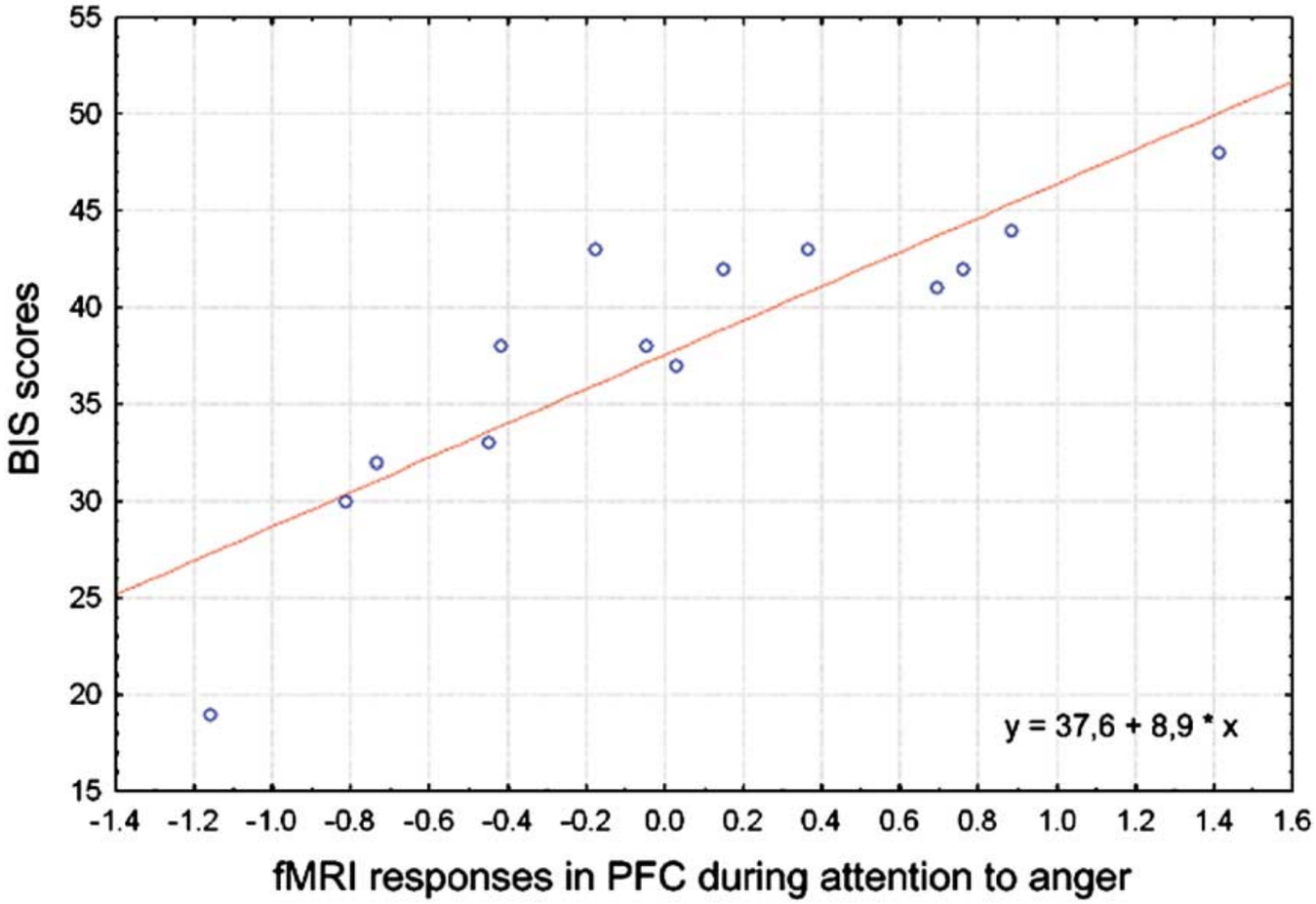


Example of "puzzlingly high correlations"

www.elsevier.com/locate/ynimg
NeuroImage 28 (2005) 848 – 858

Emotion and attention interactions in social cognition: Brain regions involved in processing anger prosody

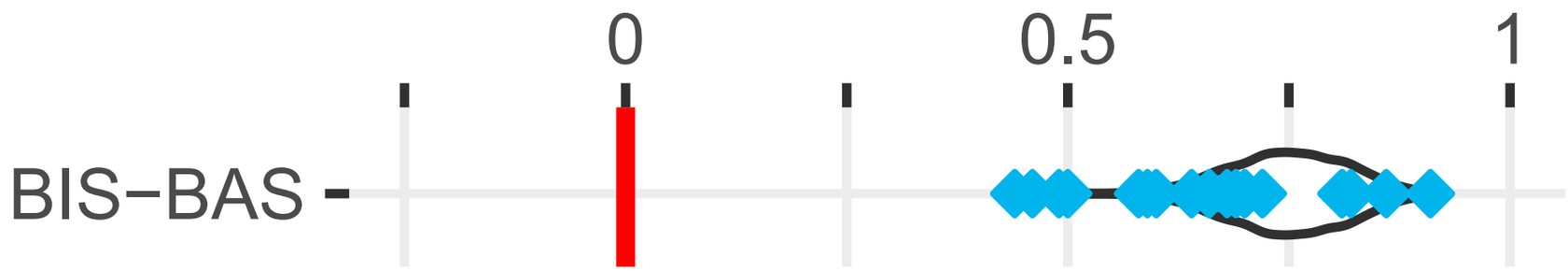
David Sander,[a,*,1] Didier Grandjean,[a,1] Gilles Pourtois,[b] Sophie Schwartz,[b]
Mohamed L. Seghier,[b,c] Klaus R. Scherer,[a] and Patrik Vuilleumier[b,d]



BIS: behavioral inhibition scale (from BIS/BAS)

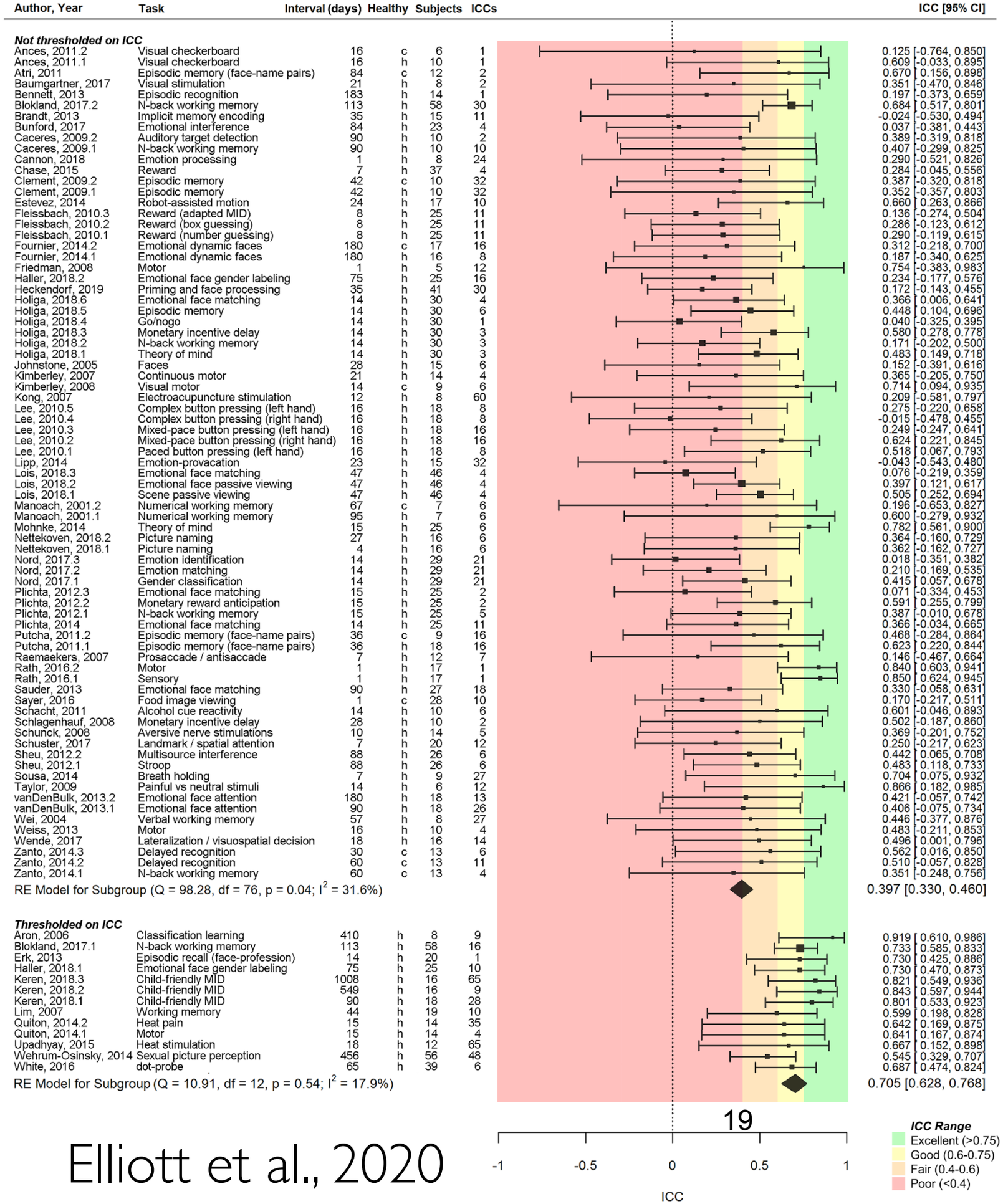
$(r = 0.87, P < 0.001).$

What is the reliability of BIS?



blue dots: estimates from literature
black violin: bootstrapped distribution of reliability estimates from Enkavi et al., 2019

What is the reliability of fMRI activation signals?



Elliott et al., 2020

Correcting for attenuation

It is common in some subfields to correct for attenuation:

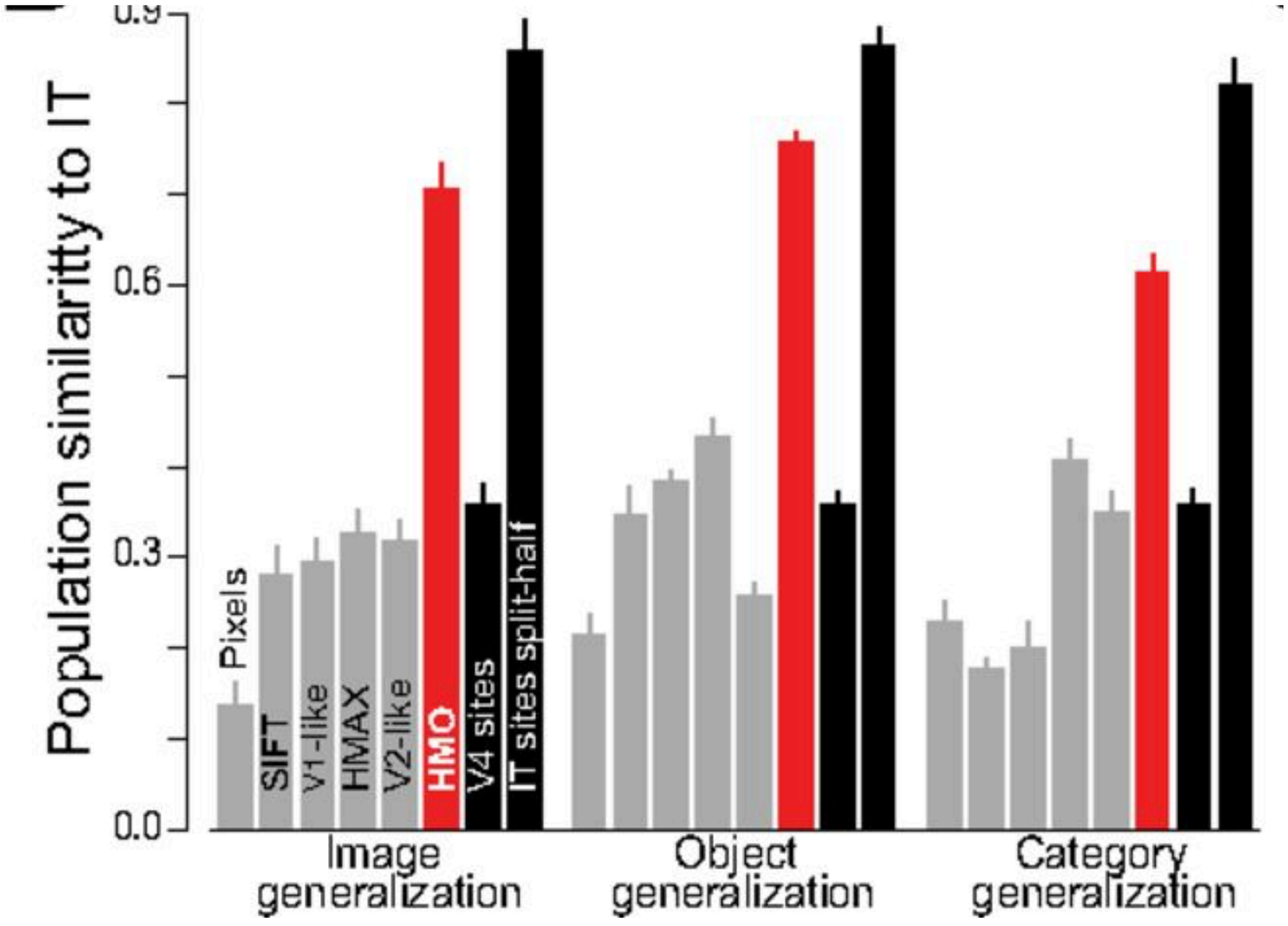
$$\rho_{\beta,\theta} = \frac{\rho_{\hat{\beta},\hat{\theta}}}{\sqrt{\rho_{\hat{\beta}}\rho_{\hat{\theta}}}}$$

Sometimes used to correct for differences in reliability in measures for SEM/factor analysis

Should be used carefully…
it can correct for unreliability, but doesn't necessarily approximate underlying "truth" (Boorsoom & Mellenberg, 2002)

Another view: Reliability as the "noise ceiling" for explained variance

The reliability provides a cap on the amount of "explainable variance"



"The IT bar represents the Spearman-Brown corrected consistency of the IT RDM for split-halves over the IT units, establishing a noise-limited upper bound."

Yamins et al., 2014, *PNAS*

Psych 253
Advanced statistical modeling

Reliability (Part 2): Neural Data

Daniel Yamins
Wu Tsai Neurosciences Institute
Departments of Psychology and Computer Science
Stanford University

Russ Poldrack
Department of Psychology
Stanford University

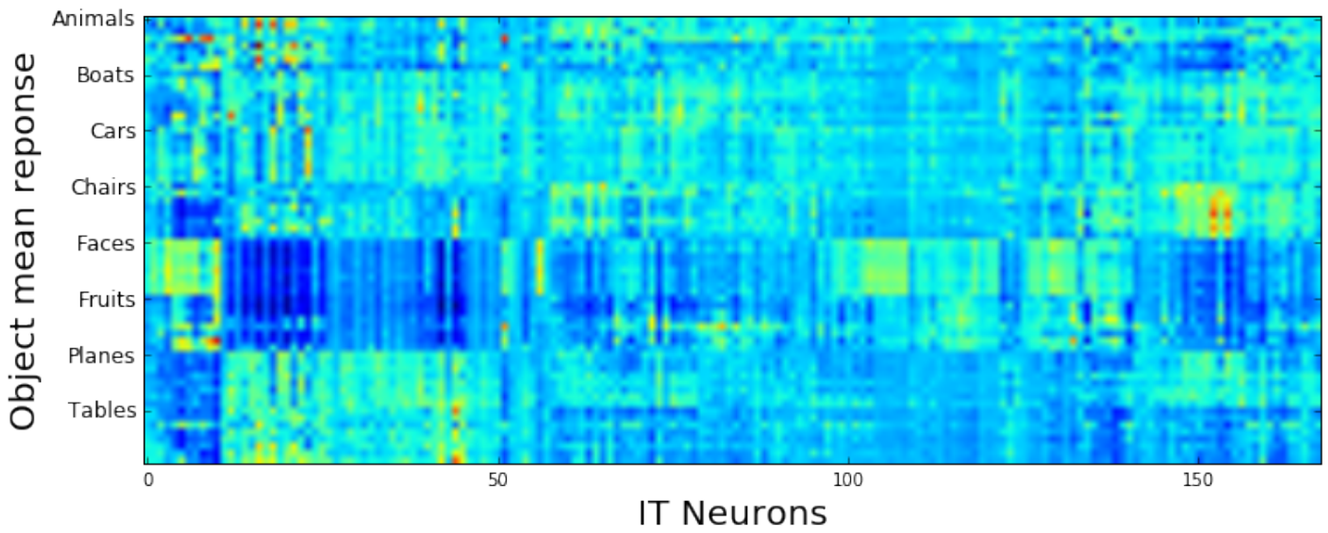
What is a Model?

$$\text{Data Slice 2} = F(\text{data\_slice\_1})$$

$$\text{Data Slice 2} = F_{\text{params}}(\text{data\_slice\_1})$$

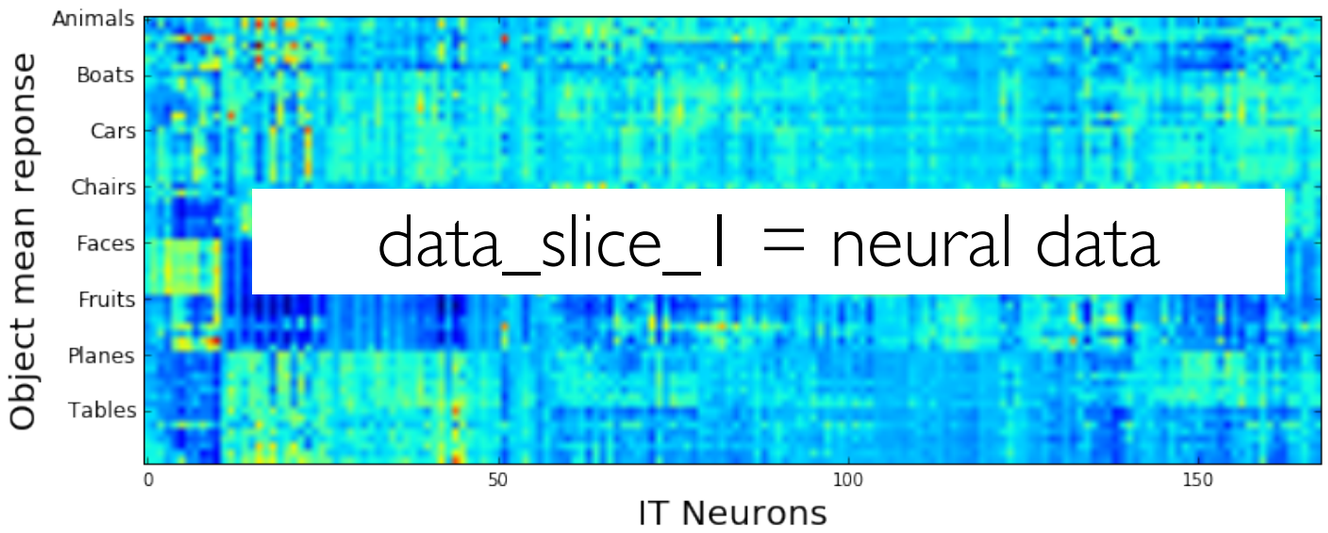
$$\text{Data Slice 2} = F_{\text{params}}(\text{data\_slice\_1}) + \text{Noise}$$

IT neural representation



Data Slice 2 = F_{params}(data_slice_1)

IT neural representation



data_slice_1 = neural data

1.
decoding

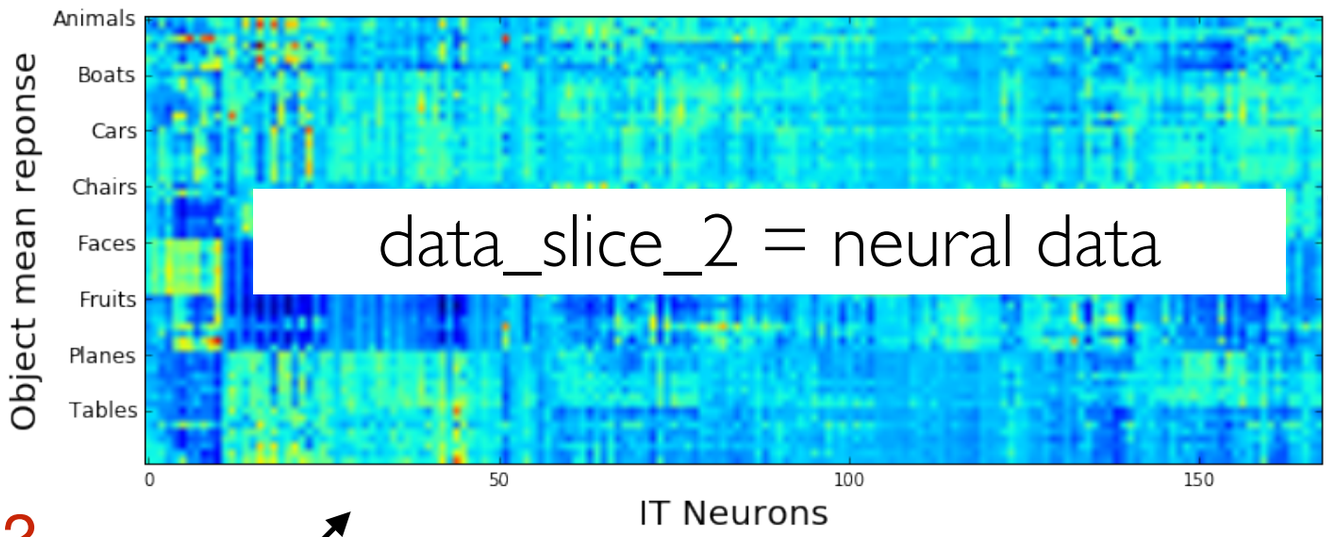
Behavioral patterns

"animal"
"boat"
"car"
.
.
.

data_slice_2 =
behavioral data

Data Slice 2 = F_{params}(data_slice_1)

IT neural representation



data_slice_2 = neural data

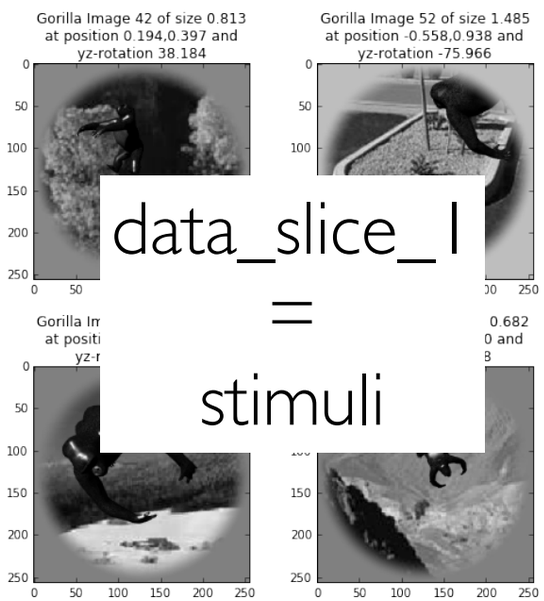
1. decoding

Behavioral patterns

"animal"
"boat"
"car"
.
.
.

2. encoding

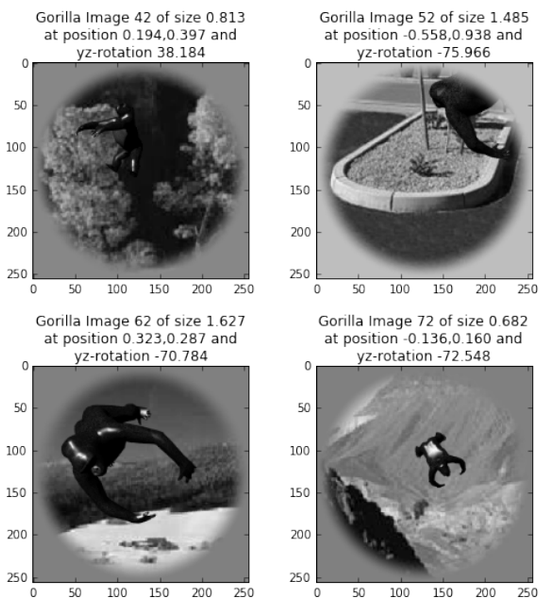
The stimulus



data_slice_1 = stimuli

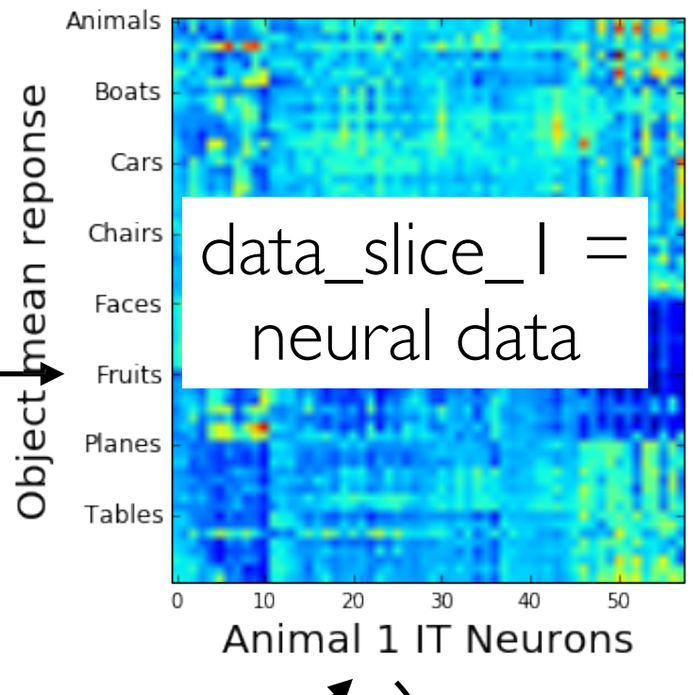
Data Slice 2 = F_{params}(data_slice_1)

The stimulus



2. encoding

Neurons



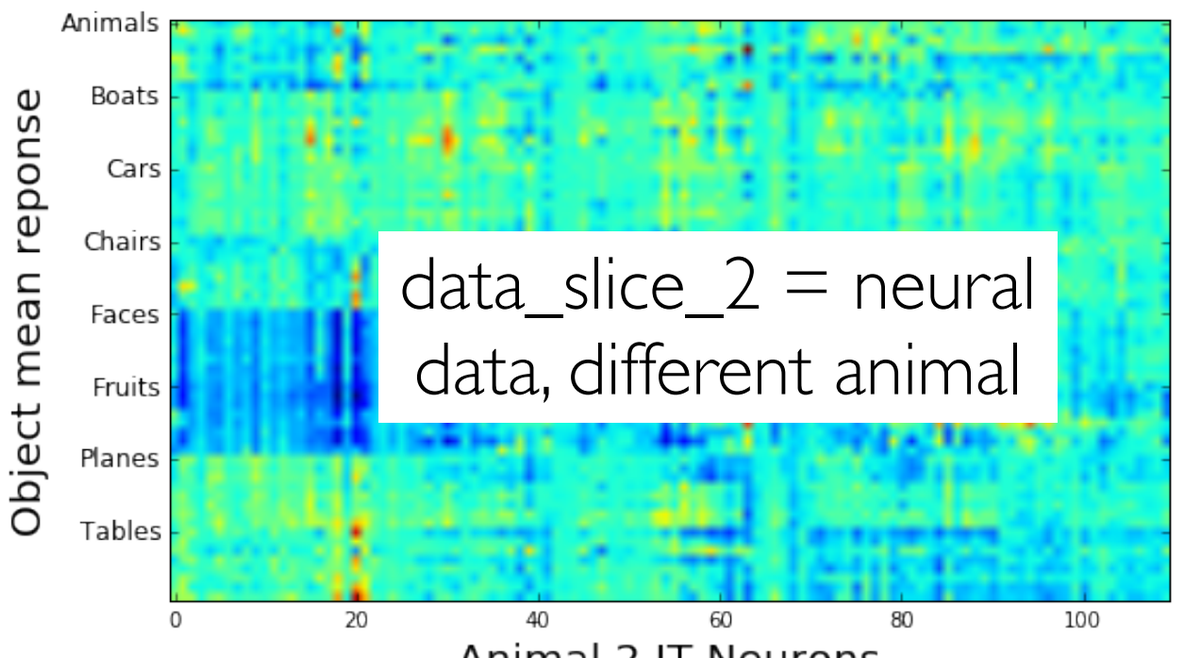
data_slice_1 = neural data

1. decoding

Behavioral patterns

"animal"
"boat"
"car"
.
.
.

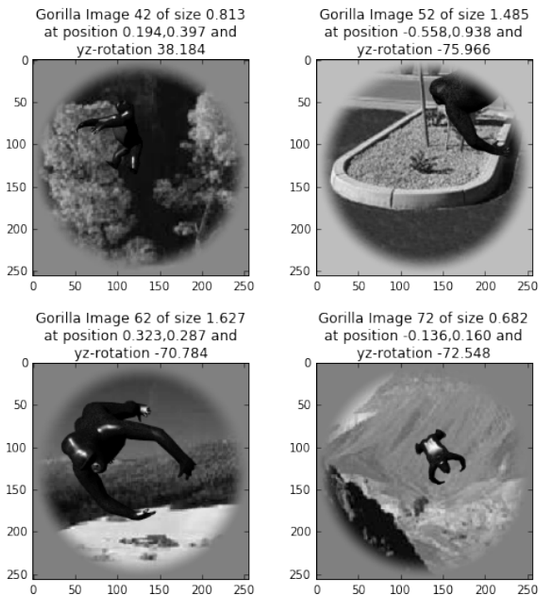
3. reliability



data_slice_2 = neural data, different animal

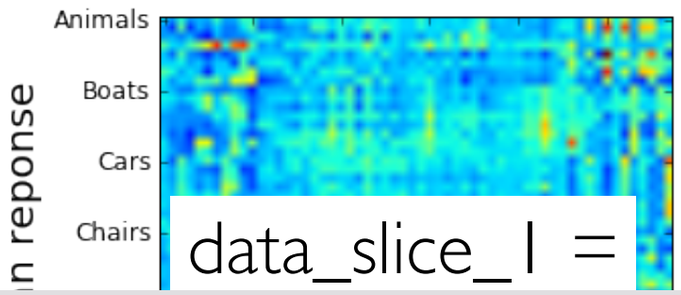
Data Slice 2 = F_{params}(data_slice_1)

The stimulus



2. encc

Neurons

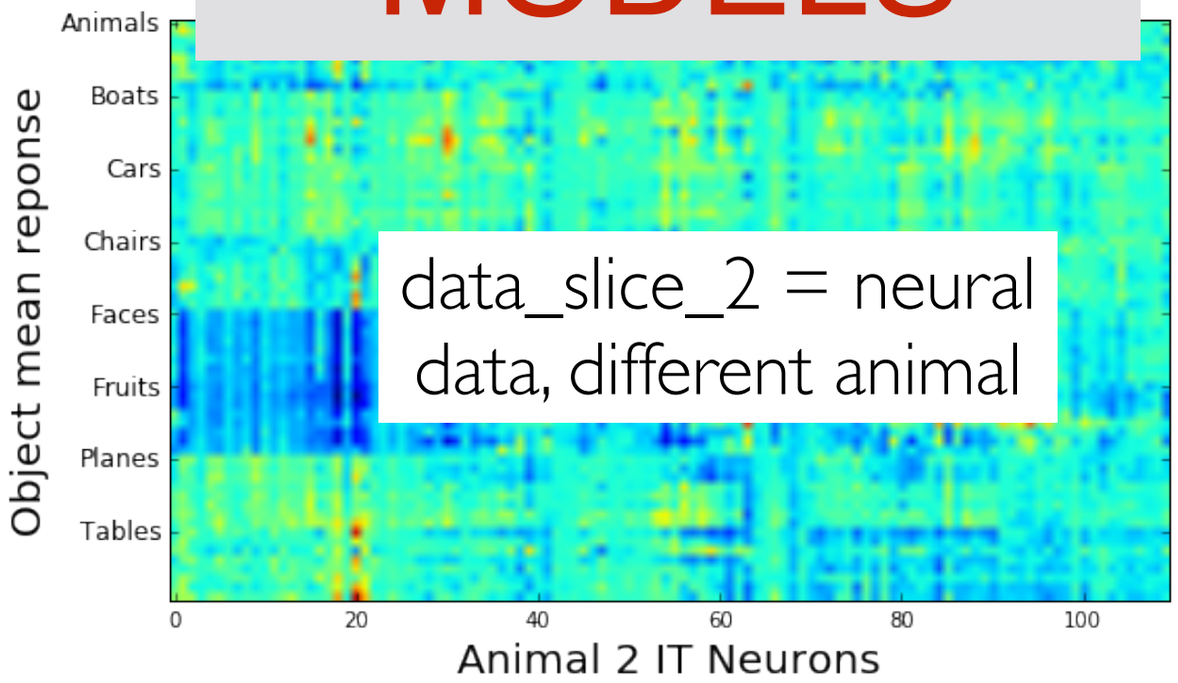


1. decoding

Behavioral patterns

"animal"
"boat"
"car"
.
.
.

ALL OF THESE THINGS ARE MODELS

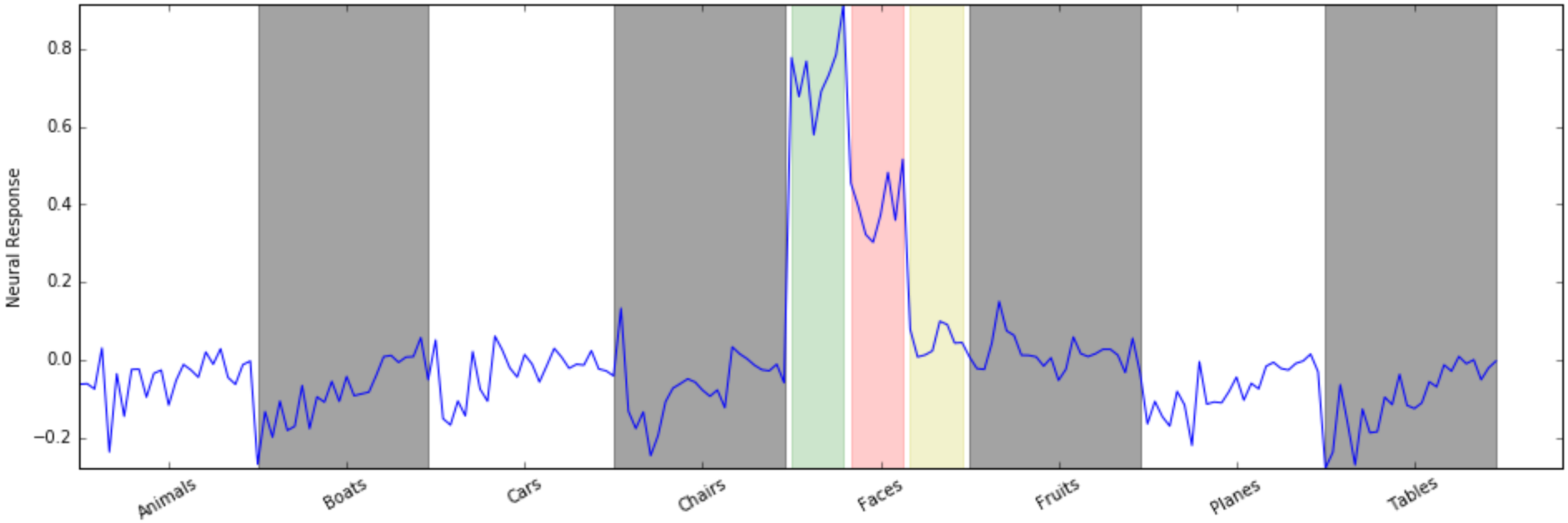


Reliability

[IPYNB: Finding some interesting neurons]

Reliability

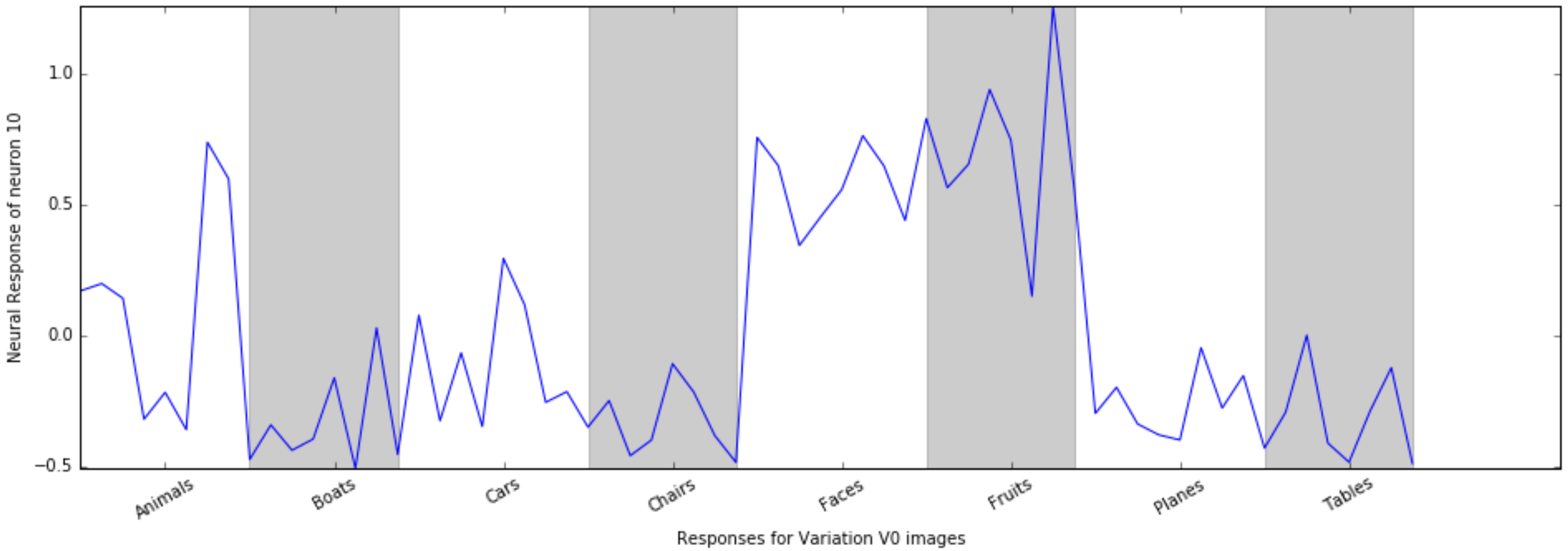
IT Unit 105



this is a face-selective unit — selectivity is high for face images at low and medium levels of variability, but not high variability images

… this is not surprising, since the faces at high variability are at quite weird angles

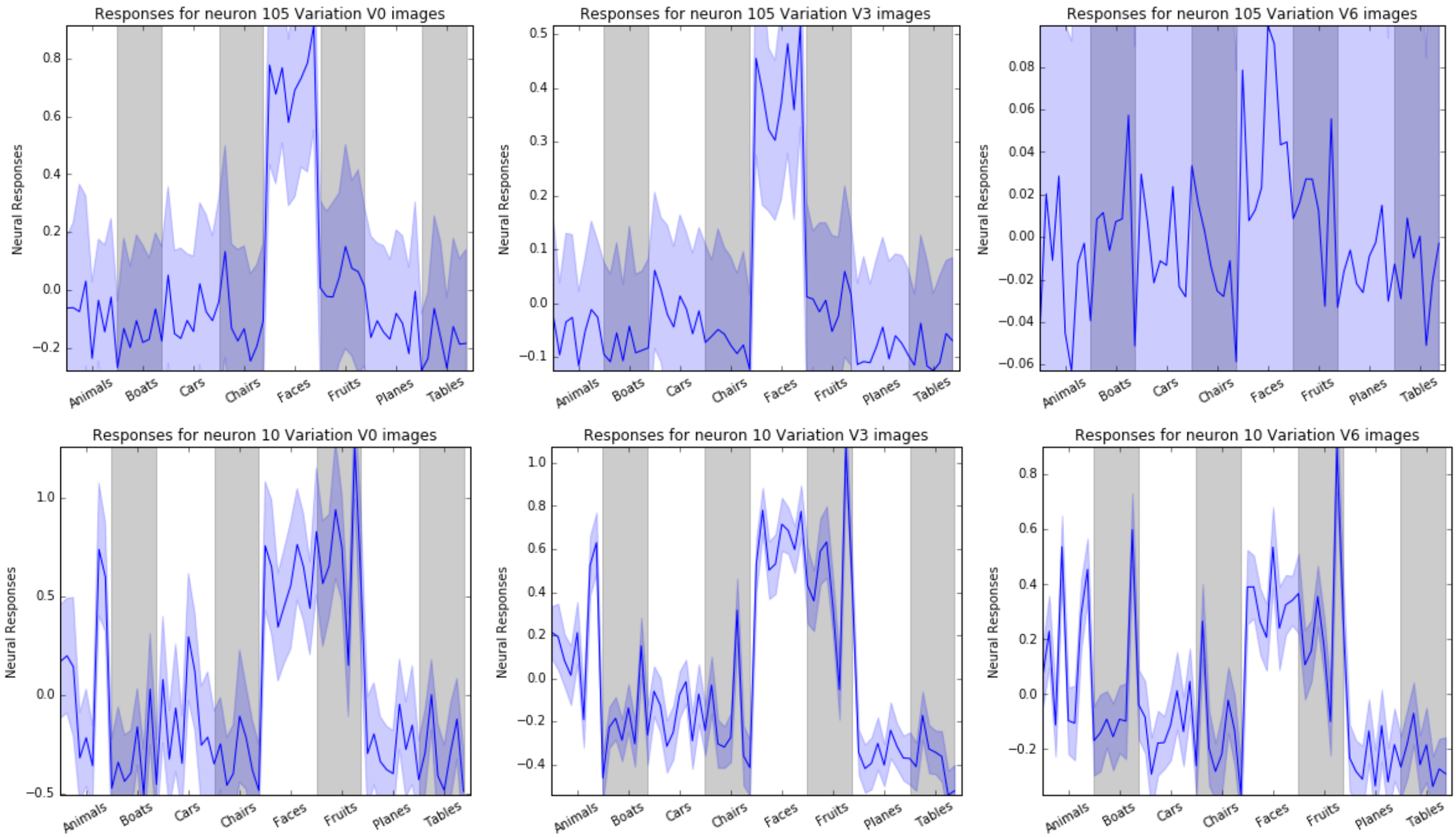
Reliability



perhaps this is a "round" object selective unit? :)

Reliability

the "round object" detector (IT unit 10) is much more reliable at variation the the putative face unit (IT unit 105) … (error bars = SEM of trial-average value)



… which **would** be true if the interpretation was real.

Reliability

Building up a pair-wise reliability analysis tool:

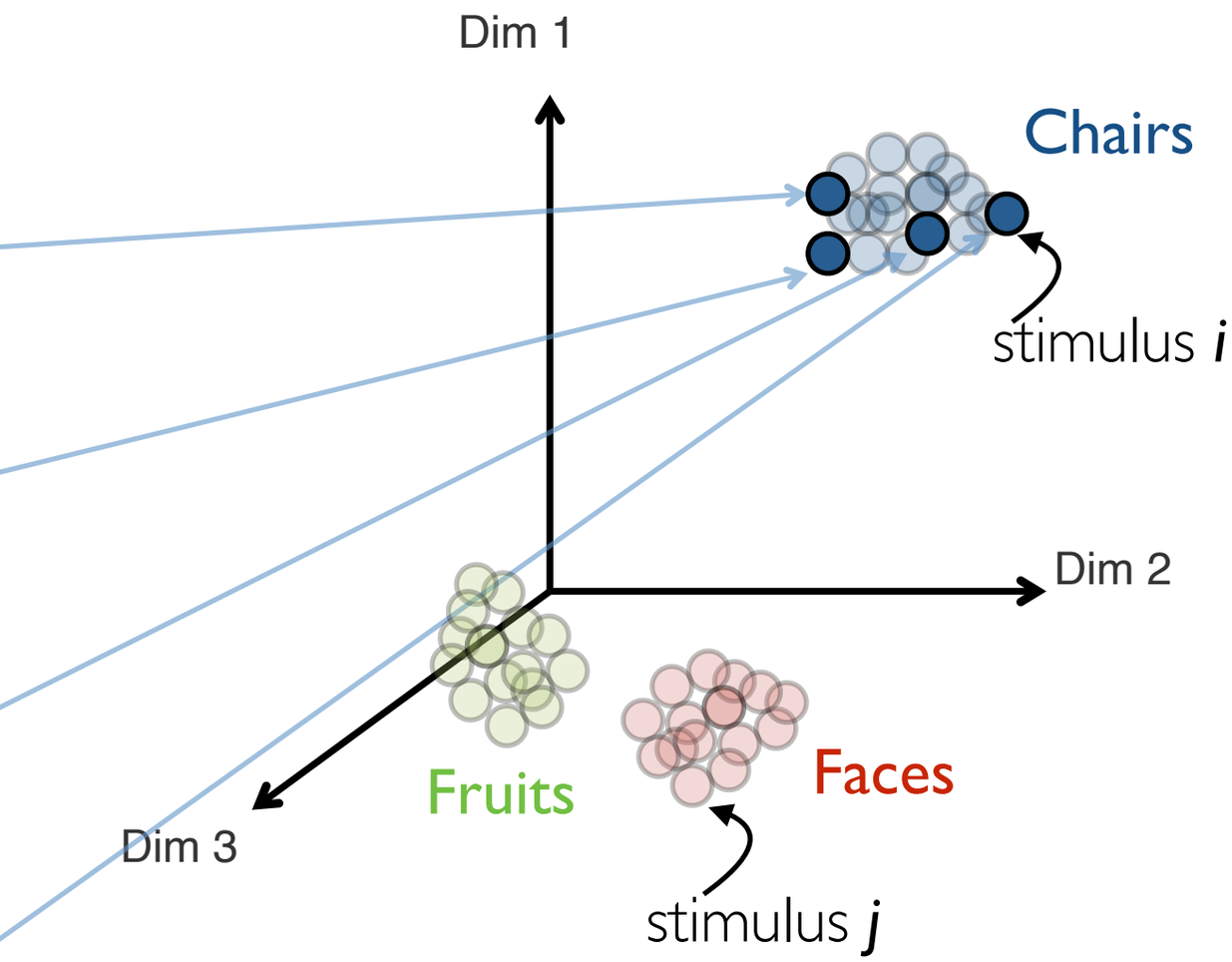
1) Look at the trial-averaged data on a per-object per-variation ("per condition") basis

2) Look at pairs of trials — esp. scatter between trials

3) Make a matrix of pairwise pearson correlation

Correlation

IT feature space

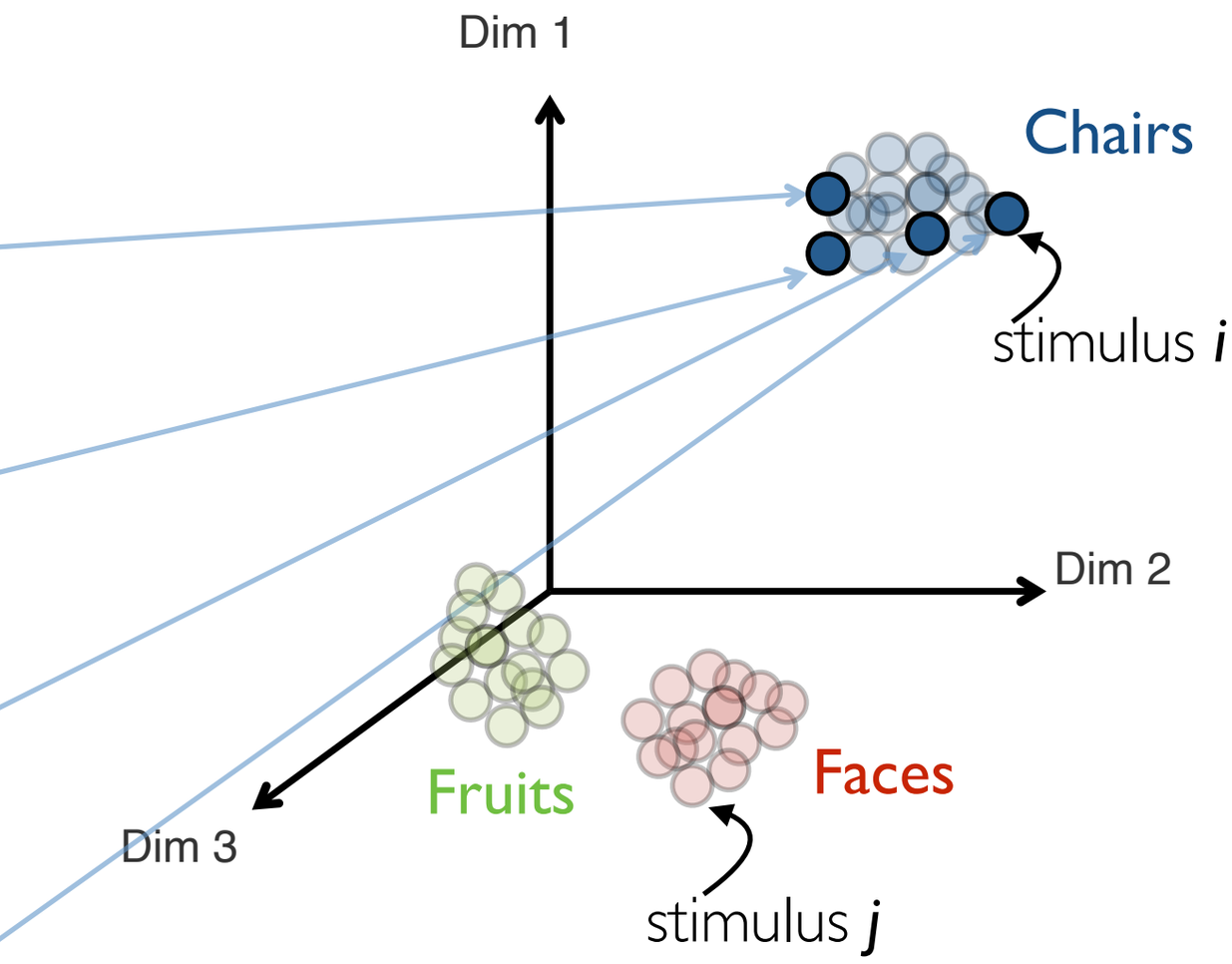


r_i = vector of neural responses to stimulus ***i***

r_j = vector of neural responses to stimulus ***j***

Correlation

IT feature space



r_i = vector of neural responses to stimulus ***i***

r_j = vector of neural responses to stimulus ***j***

$$\text{pearson's r}(r_i, r_j)$$

$$= \frac{cov(r_i, r_j)}{\sqrt{var(r_i) \cdot var(r_j)}}$$

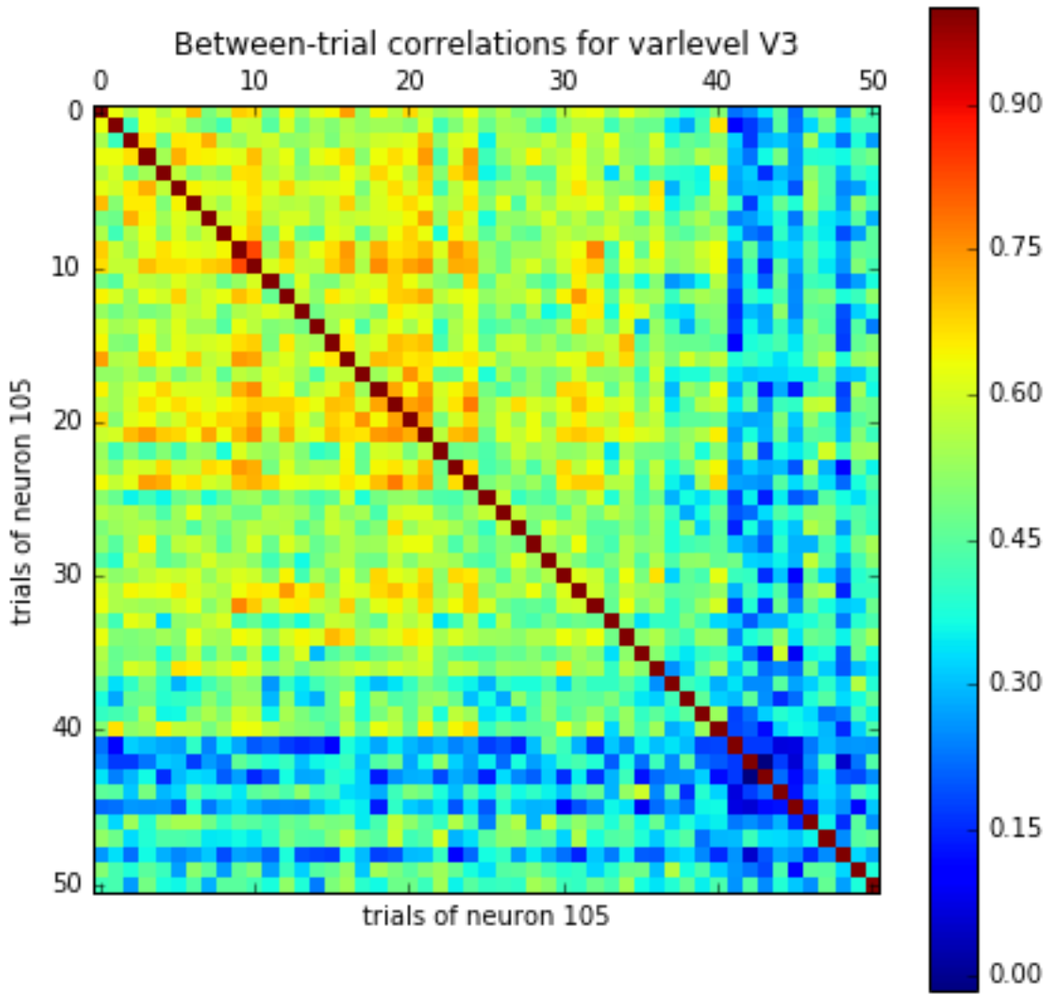
$$\mathbf{E}_k[r_i r_j] - \mathbf{E}_k[r_i]\mathbf{E}_k[r_j]$$

$$\mathbf{E}_k[r_i^2] - \mathbf{E}_k[r_i]^2$$

expectations over neurons

Correlation

Making matrices like these:



ADDITIONAL NOTES ON RELIABILITY

[IPYNB: Simple Reliability Analysis]

Split-Half Reliability

```
func get_splithalf:
```

Split-Half Reliability

```
func get_splithalf:
  for various splits of trials:
```

Split-Half Reliability

```
func get_splithalf:
  for various splits of trials:

    m1 = mean of first half of trials
    m2 = mean of second half of trials
```

Split-Half Reliability

```
func get_splithalf:
  for various splits of trials:

    m1 = mean of first half of trials
    m2 = mean of second half of trials

    c = pearson(m1, m2)
```

Split-Half Reliability

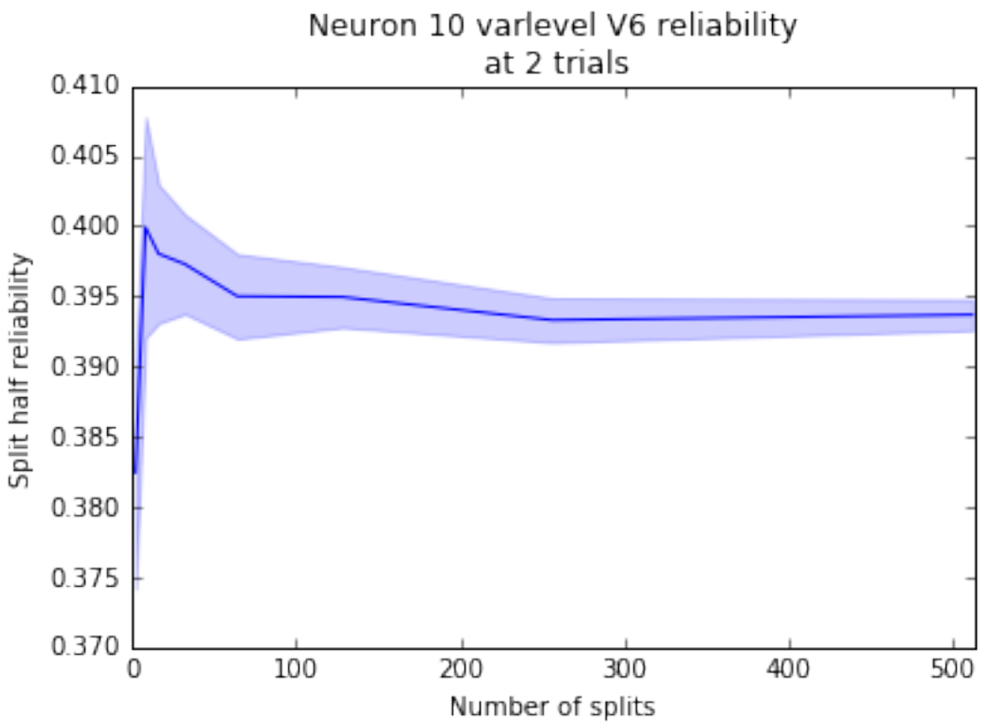
```
func get_splithalf:
  for various splits of trials:

    m1 = mean of first half of trials
    m2 = mean of second half of trials

    c = pearson(m1, m2)

return average of c's
```

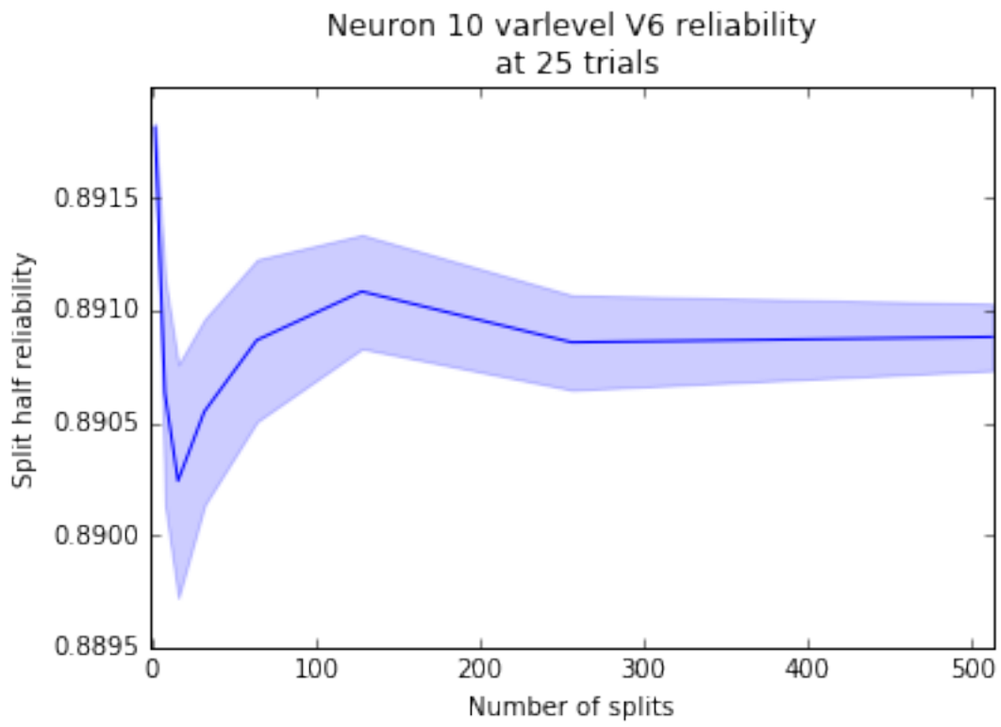
Reliability

[IPYNB: Split-Half Reliability]

Reliability

num_splits ~ 10 * num_trials is good enough to get reasonable estimate for reliability:





error bars = SEM of reliability estimate

The Prophecy Formula

for normally distributed independent measure, correlation between averages of two length-k samples is:

$$\rho_k = \frac{\sigma_T^2}{\sigma_T^2 + a^2/k}$$

σ_T = true variability of data

a = error std for 1 sample

$\rho_k = corr(X_k/k, X'_k/k)$

The Prophecy Formula

$$
\begin{aligned}
\rho_k &= \frac{\sigma_T^2}{\sigma_T^2 + a^2/k} = \frac{k\sigma_T^2}{k\sigma_T^2 + a^2} \\
&= \frac{k\sigma_T^2}{k\sigma_T^2 - \sigma_T^2 + (\sigma_T^2 + a^2)} \\
&= \frac{k\sigma_T^2}{(k-1)\sigma_T^2 + (\sigma_T^2 + a^2)} \\
&= \frac{k\frac{\sigma_T^2}{\sigma_T^2+a^2}}{(k-1)\frac{\sigma_T^2}{\sigma_T^2+a^2} + 1} \\
&= \frac{k\rho_1}{(k-1)\rho_1 + 1}
\end{aligned}
$$

The Prophecy Formula

$$\text{spearman-brown}(\rho, k) = \frac{k \cdot \rho}{1 + (k-1)\rho}$$

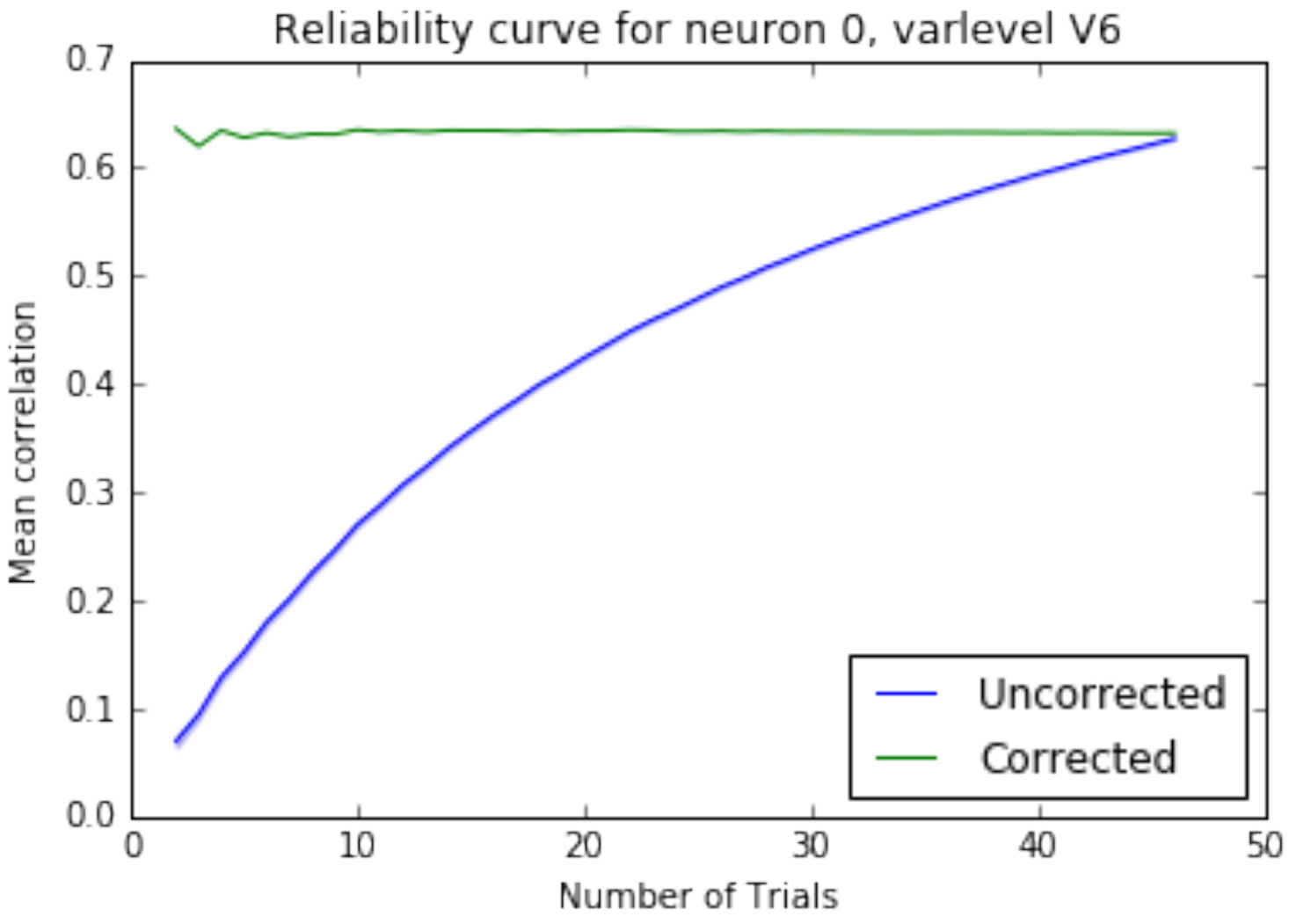
ρ = correlation for some number of trials

k = multiple of original number of trials for which you want to estimate correlation

Reliability

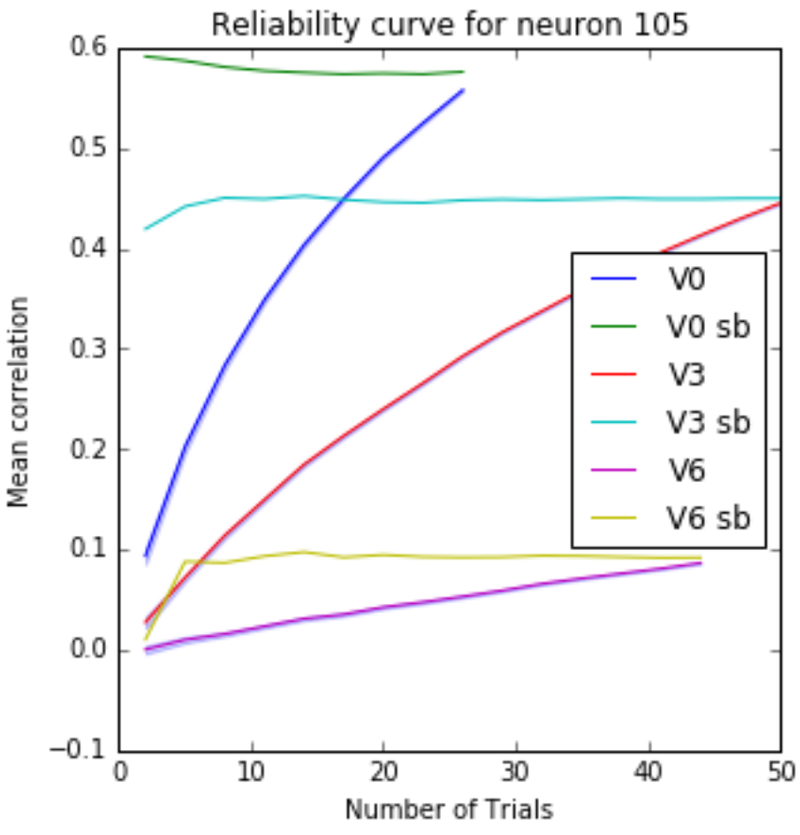
[IPYNB: Spearman-Brown]

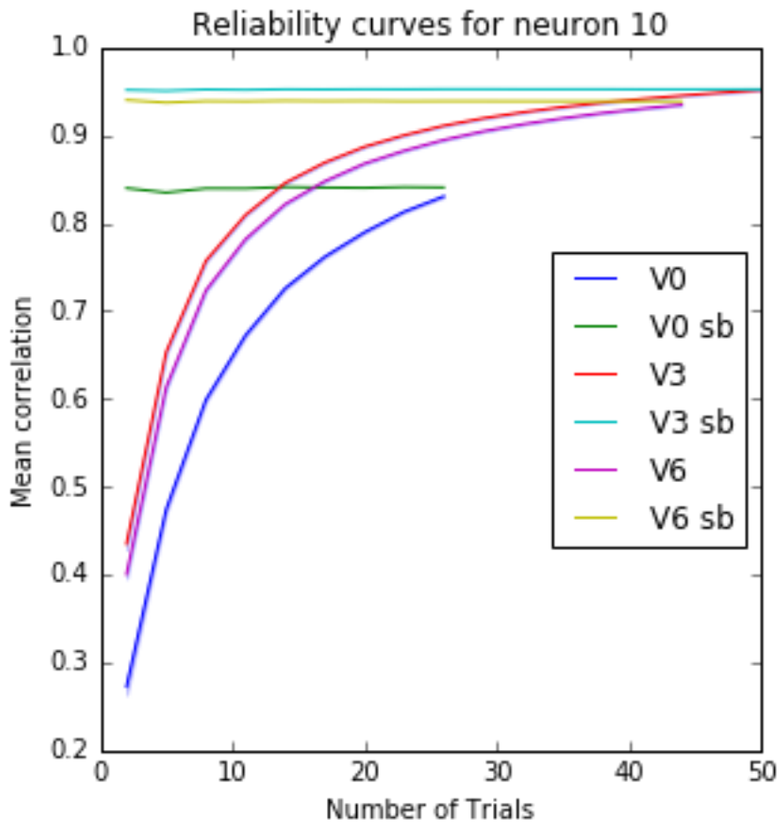
Reliability

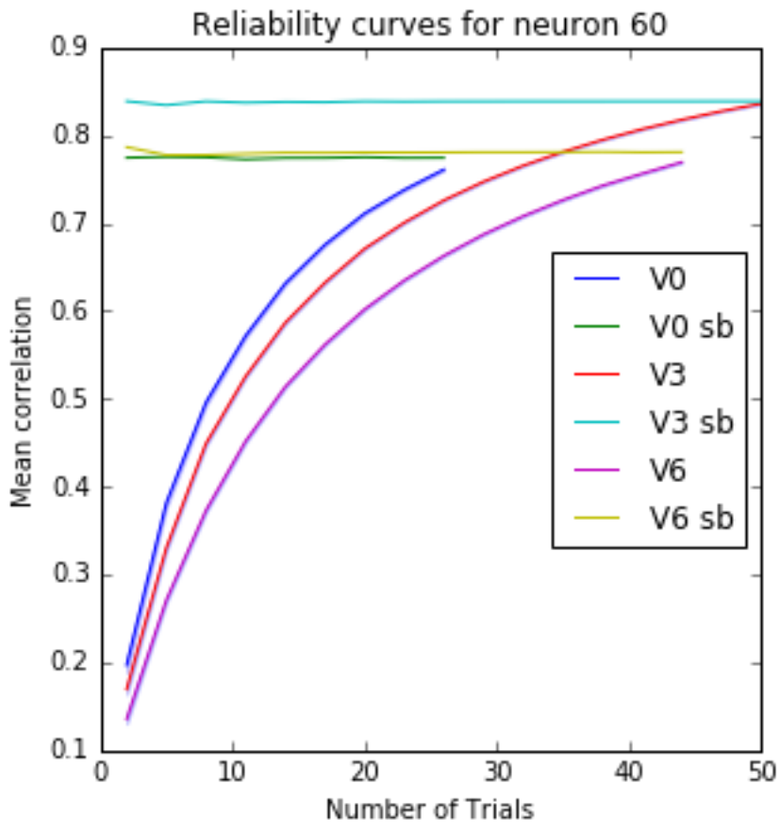


Reliability

Spearman-brown estimates are pretty good …







… so neural noise (in these whitened signals) must be pretty gaussian

Bootstrapping

The split-half reliability calculation relies on sampling ***without*** replacement

```
func get_splithalf:
  for various splits of trials:

    m1 = mean of first half of trials
    m2 = mean of second half of trials

    c = pearson(m1, m2)

return average of c's
```

non-overlapping trials

In theory, this could be a problem, especially for small sample-sized datasets — hard to get a good estimate of what larger sample might be like.

Bootstrapping

Alternative: "boostrapping" — create synthetic trials by resampling ***with*** replacement

```
func get_bootstrap_sample (data_by_trial):
```

Bootstrapping

Alternative: "boostrapping" — create synthetic trials by resampling ***with*** replacement

```
func get_bootstrap_sample (data_by_trial):
  do num_trials times:
```

Bootstrapping

Alternative: "boostrapping" — create synthetic trials by resampling ***with*** replacement

```
func get_bootstrap_sample (data_by_trial):
  do num_trials times:
    for each stimulus:
      randomly select trial to pick value from
      average over these "internal samples"
```

Bootstrapping

Alternative: "boostrapping" — create synthetic trials by resampling ***with*** replacement

```
func get_bootstrap_sample (data_by_trial):
  do num_trials times:
    for each stimulus:
      randomly select trial to pick value from
    average over these "internal samples"

do num_iter times:
  call get_bootstrap_sample
```

Bootstrapping

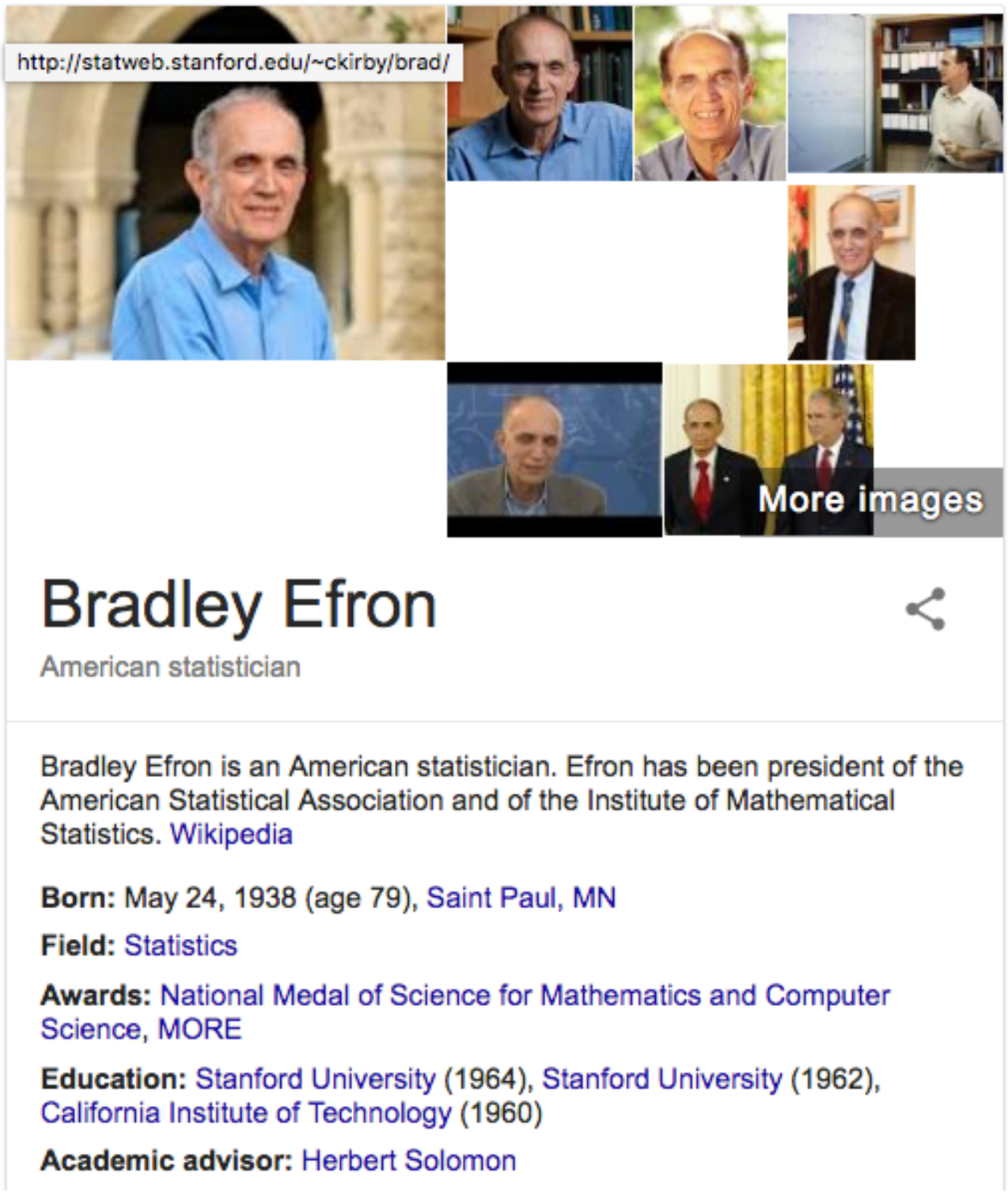
Alternative: "boostrapping" — create synthetic trials by resampling ***with*** replacement

```
func get_bootstrap_sample (data_by_trial):
  do num_trials times:
    for each stimulus:
      randomly select trial to pick value from
    average over these "internal samples"

do num_iter times:
  call get_bootstrap_sample

compute mean over correlations between
bootstrap samples
```

Bootstrapping

Alternative: "boos ampling ***with*** replacement

```
func get_                                    ial):
  do num_
    for e
      ranc                                   lue from
  average                                    s"

do num_it
  call ge

compute m                                    n
bootstrap
```



Bootstrapping

Alternative: "boostrapping" — create synthetic trials by resampling ***with*** replacement

```
func get_bootstrap_sample (data_by_trial):
  do num_trials times:
    for each stimulus:
      randomly select trial to pick value from
      average over these "internal samples"

do num_iter times:
  call get_bootstrap_sample

compute mean over correlations between
bootstrap samples
```

Actually, debatable whether better est. of true reliability at given (large) number of trials than split-half + Spearman-Brown …

Bootstrapping

[IPYNB: Bootstrapping]