MODEL COMPARISON

Alexander Huth
@ UBA ECI Course
2017-07-28

AGENDA

* How do you compare feature spaces?
  * Variance partitioning
* Encoding vs. decoding
* Data visualization

LINEARIZED MODELS

$$Y = \mathbb{L}(X)\beta$$

* **L** is some non-linear function of the stimulus ***X*** that gives us *features*

* ***Beta*** is a linear weighting of the *features* that gives us the response ***Y***

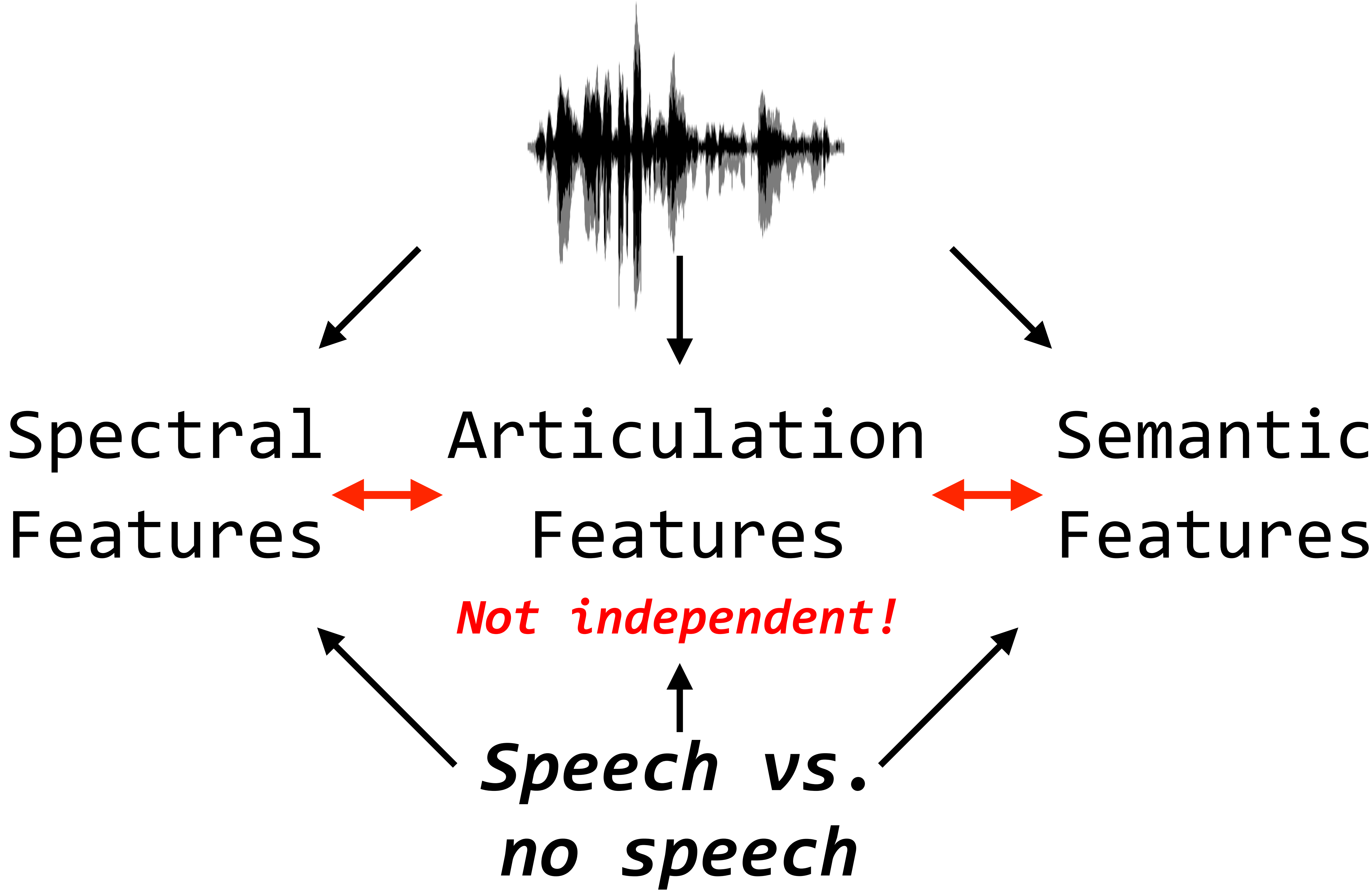
MANY LINEARIZED MODELS

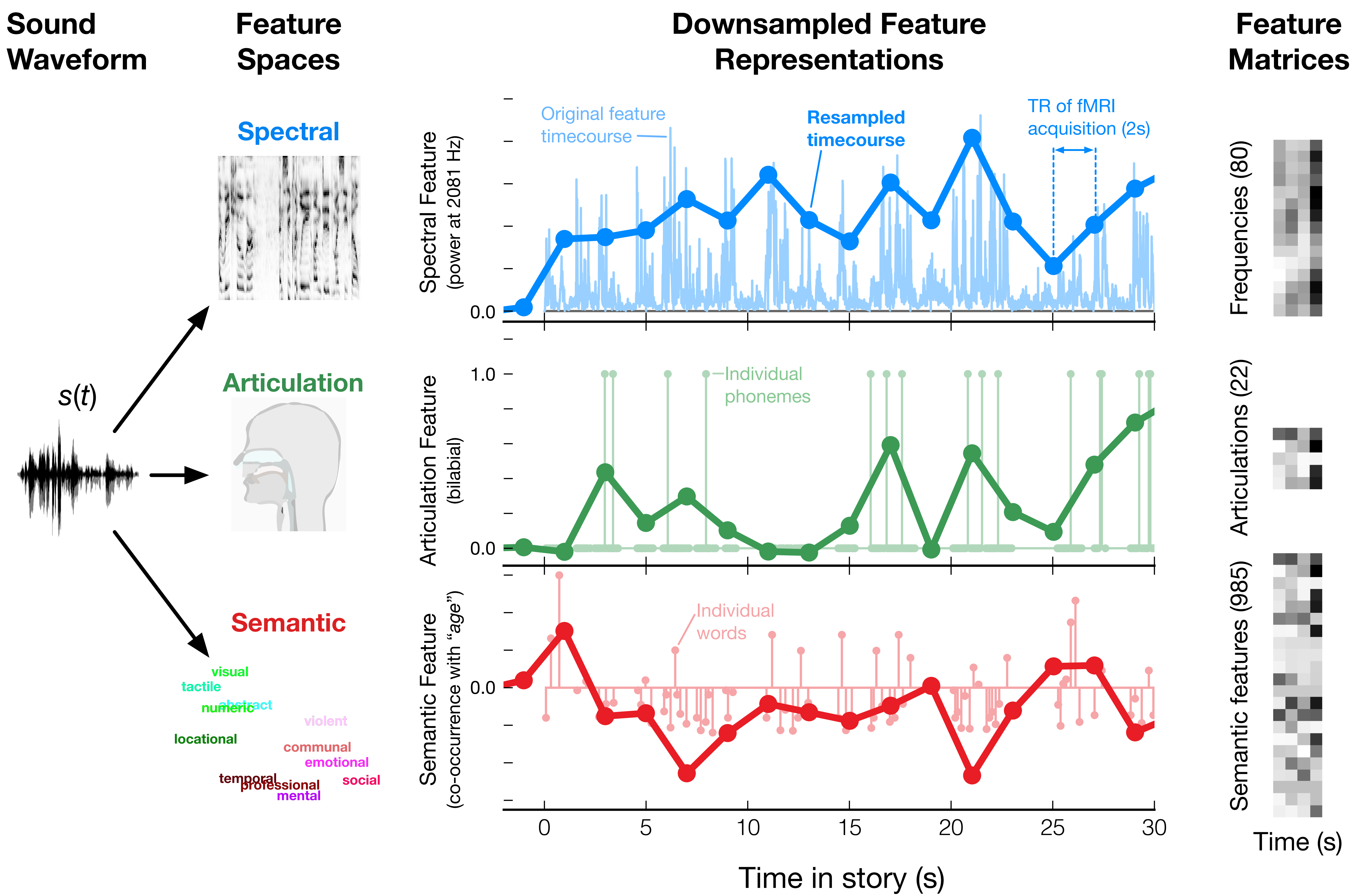
$$
\begin{aligned}
Y &= \mathbb{L}_1(X)\beta_1 \\
Y &= \mathbb{L}_2(X)\beta_2 \\
Y &= \mathbb{L}_3(X)\beta_3 \\
&\vdots
\end{aligned}
$$

MANY LINEARIZED MODELS

* Which feature space is **best**?
* Is that even the right question?
  * Sometimes, ***NO!***

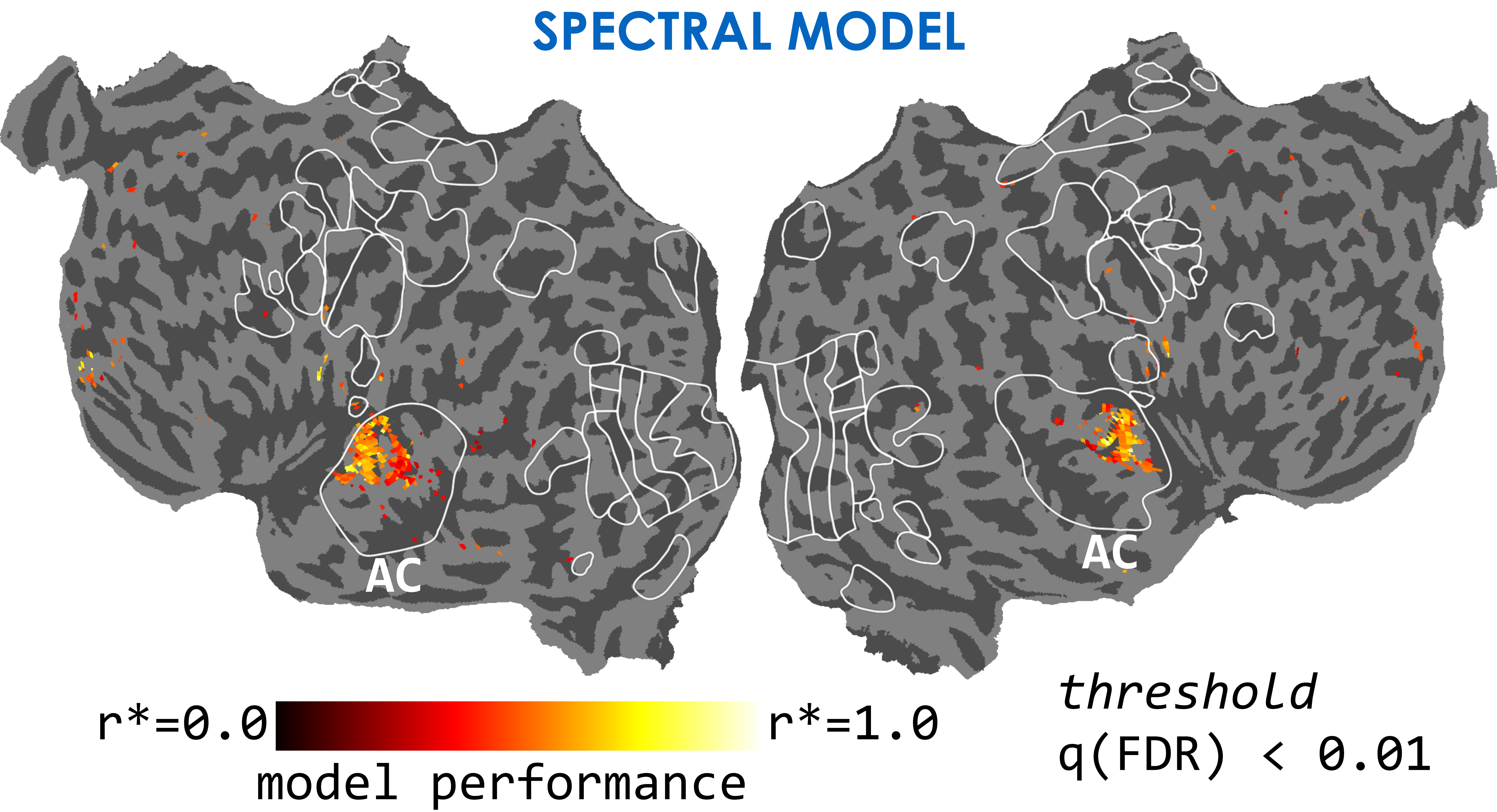
SPEECH MODELS





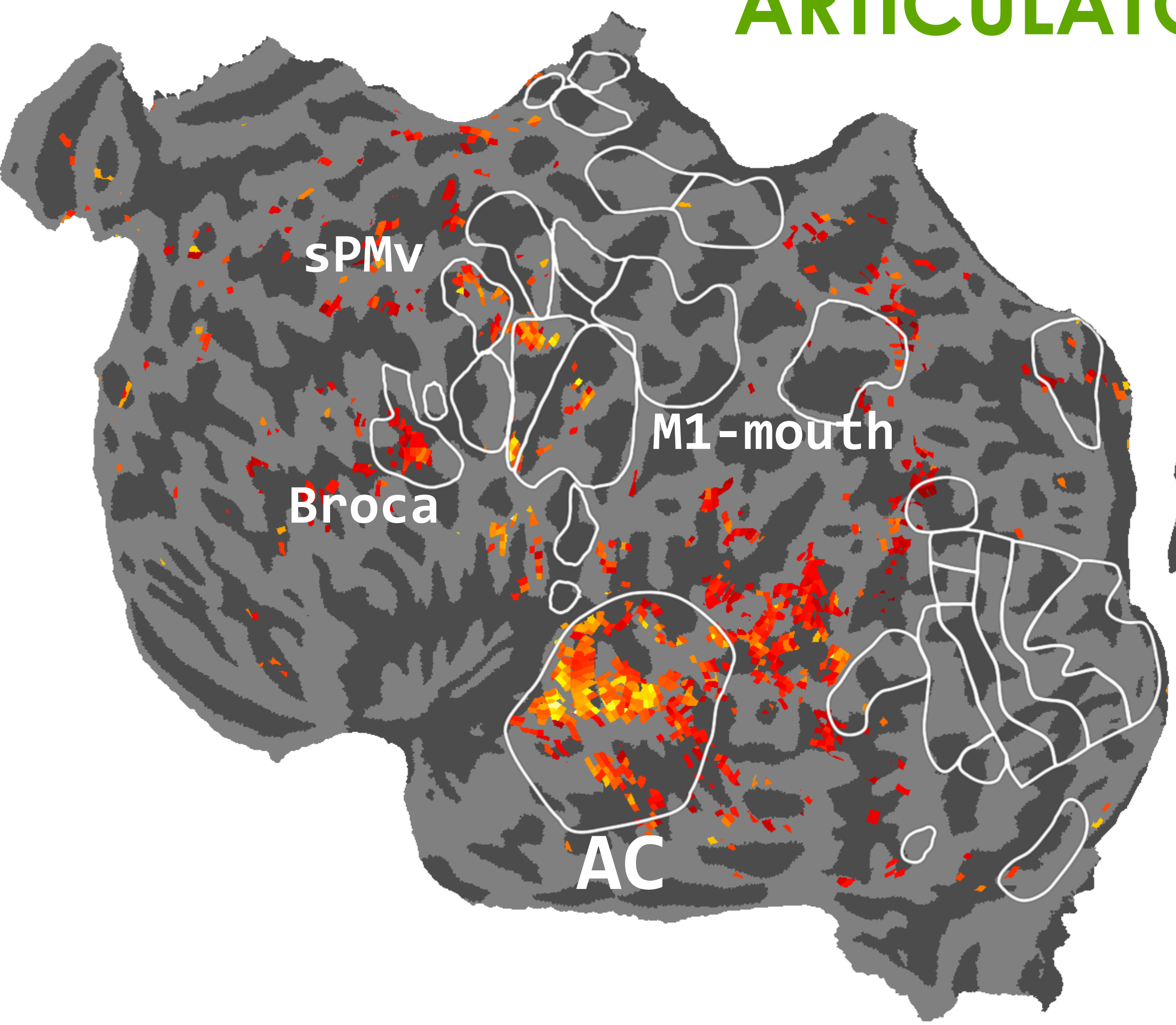
Sound Waveform
Feature Spaces
Downsampled Feature Representations
Feature Matrices
Spectral
Articulation
Semantic
s(t)
visual
tactile
abstract
numeric
violent
locational
communal
emotional
temporal
professional
social
mental
Original feature timecourse
Resampled timecourse
TR of fMRI acquisition (2s)
Spectral Feature
(power at 2081 Hz)
0.0
1.0
Individual phonemes
Articulation Feature
(bilabial)
Individual words
Semantic Feature
(co-occurrence with "age")
0
5
10
15
20
25
30
Time in story (s)
Frequencies (80)
Articulations (22)
Semantic features (985)
Time (s)

SPEECH MODELS



SPEECH MODELS

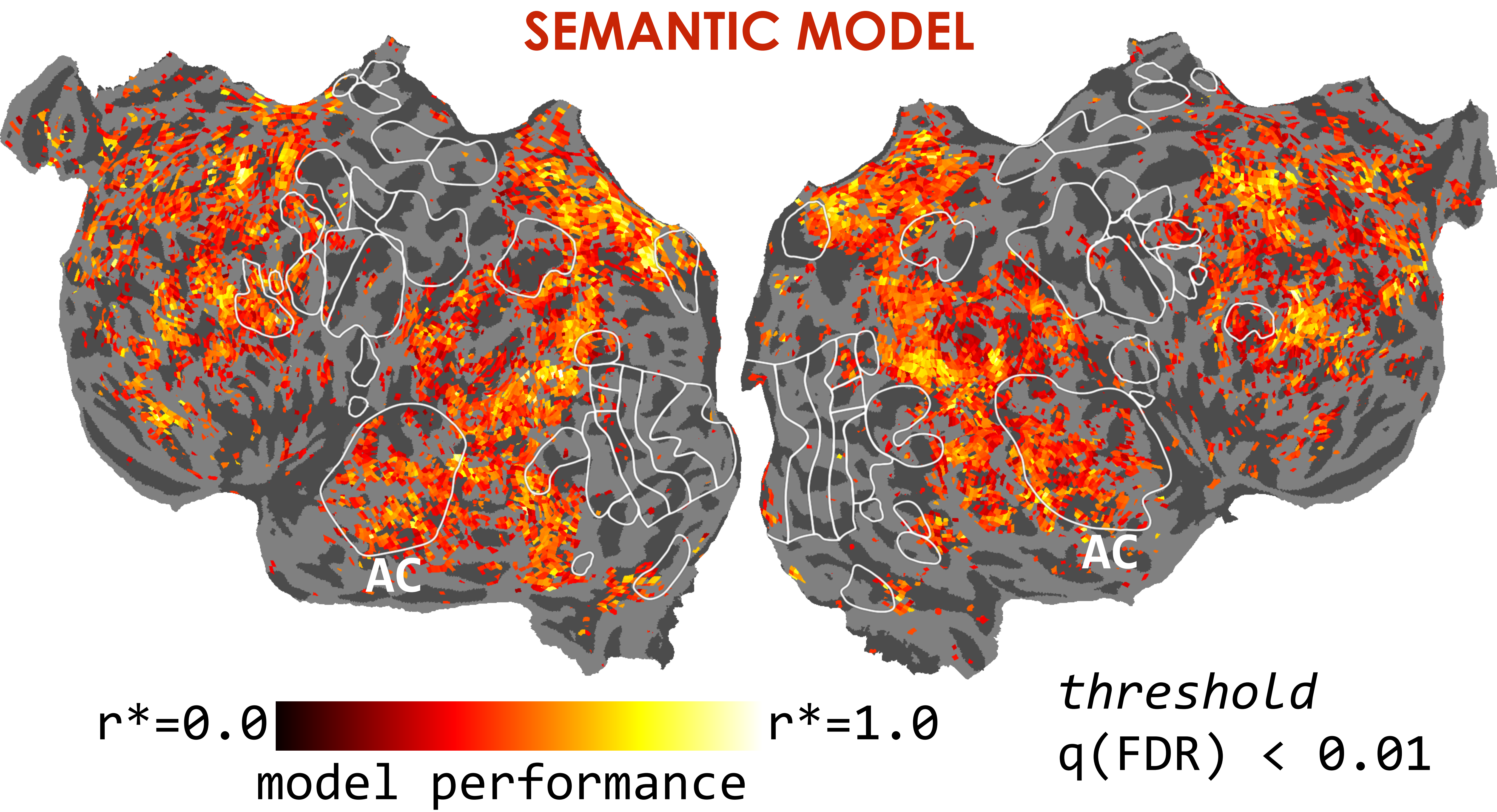
ARTICULATORY MODEL



r*=0.0 r*=1.0
model performance

threshold
q(FDR) < 0.01

SPEECH MODELS



MODEL COMPARISON

* Can we decide which feature space is **better/best**?

  * **Overall?** *Easy:* normal stats, parametric or non-parametric

  * **Per voxel?** *Harder:* low statistical power

VARIANCE PARTITIONING

* What about responses that are well-predicted by *multiple* feature spaces?

  * <u>Possibility 1:</u> each feature space explains **different** variance

VARIANCE PARTITIONING

* What about responses that are well-predicted by *multiple* feature spaces?

  * Possibility 1: each feature space explains **different** variance

$$y = \boxed{X_A\beta_A} + \boxed{X_B\beta_B} + \epsilon$$

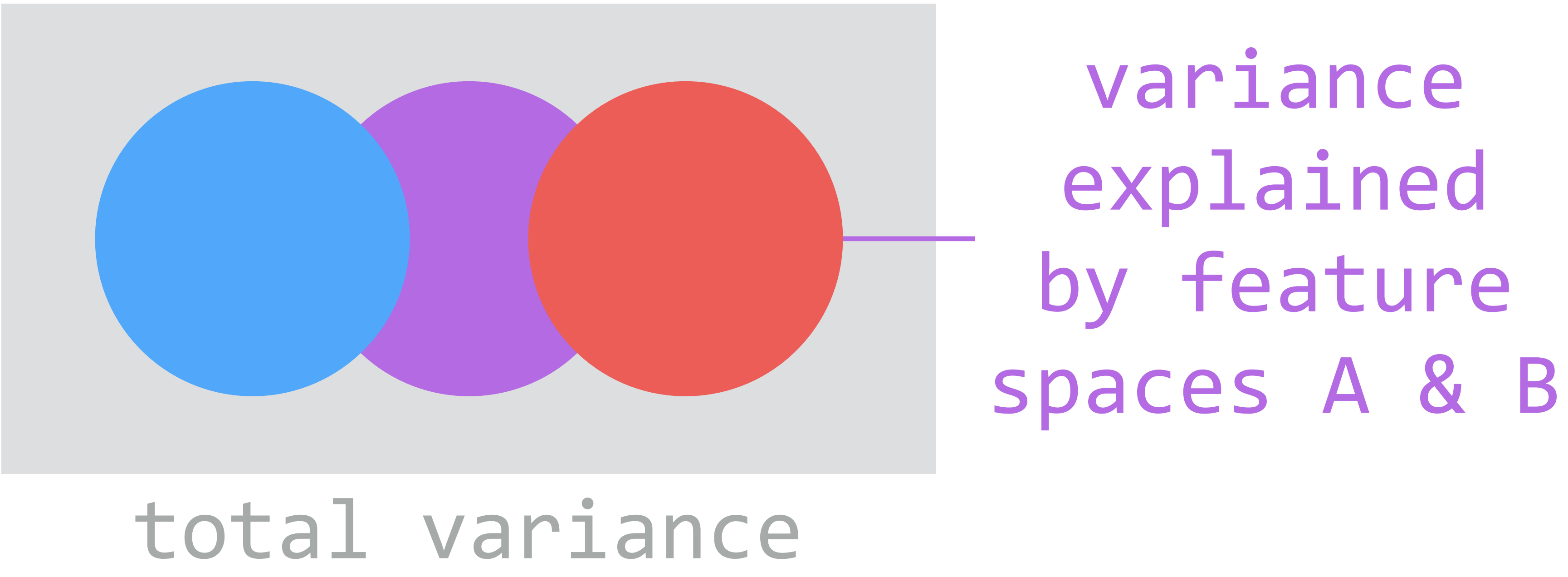
variance explained by feature space A (e.g. spectral)

variance explained by feature space B (e.g. semantic)

VARIANCE PARTITIONING

* What about responses that are well-predicted by *multiple* feature spaces?

  * <u>Possibility 2:</u> both feature spaces explain **the same** variance



VARIANCE PARTITIONING

* What about responses that are well-predicted by *multiple* feature spaces?

  * <u>Possibility 2:</u> both feature spaces explain **the same** variance

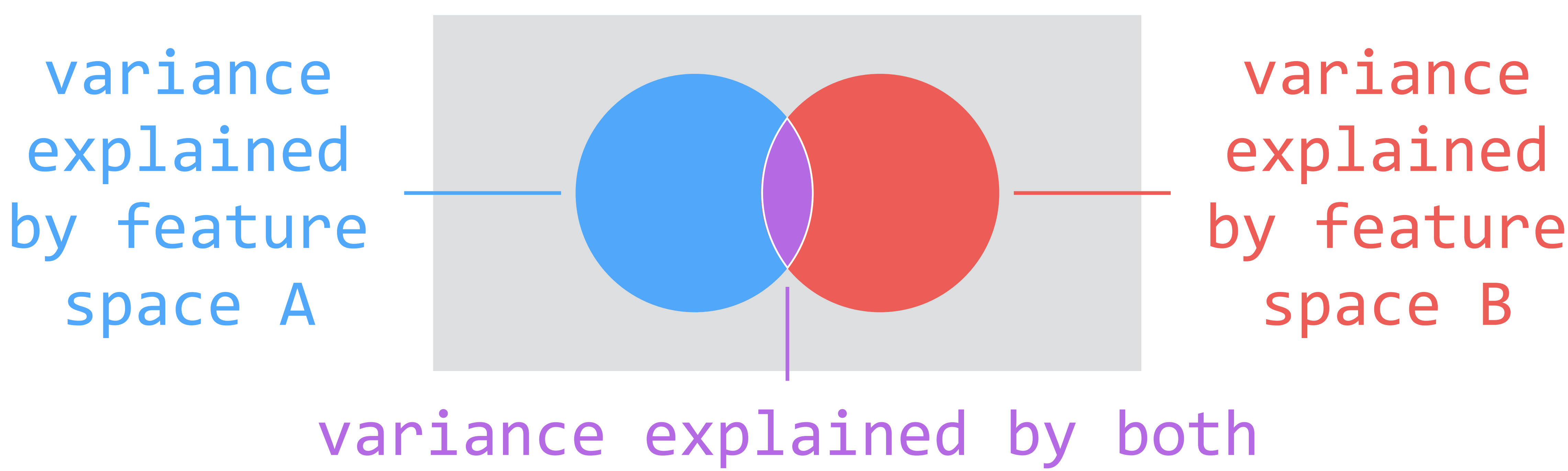
$$y = X_A\beta_A + \epsilon = X_B\beta_B + \epsilon$$

$$X_A\beta_A = X_B\beta_B$$

VARIANCE PARTITIONING

* What about responses that are well-predicted by *multiple* feature spaces?

    * <u>Possibility 3:</u> the feature spaces explain **some of the same** variance, and **some different**



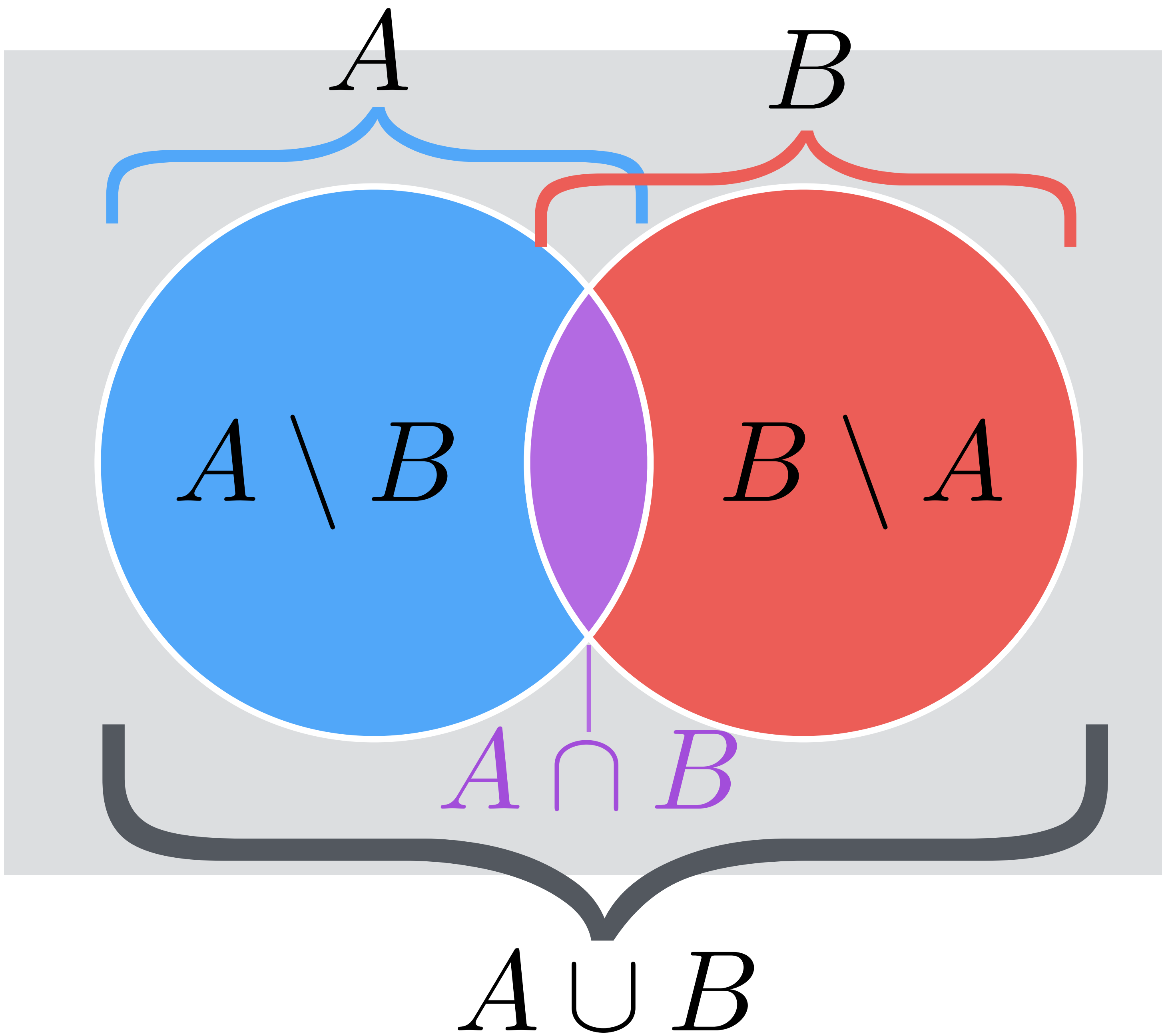
VARIANCE PARTITIONING

* What about responses that are well-predicted by *multiple* feature spaces?

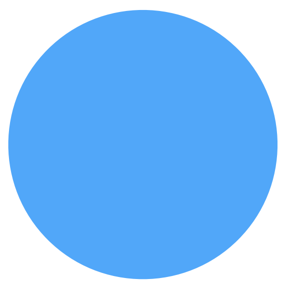
  * Possibility 3: the feature spaces explain **some of the same** variance, and **some different**

$$y = ?$$

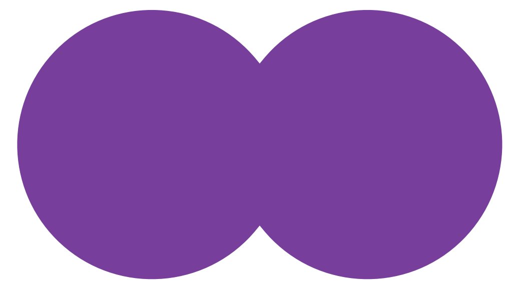
VARIANCE PARTITIONING



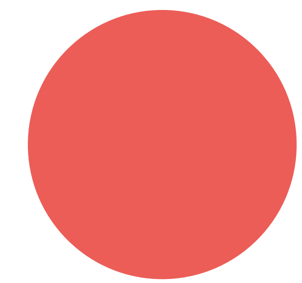
VARIANCE PARTITIONING

A


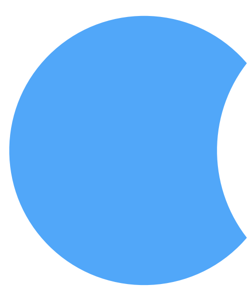
variance explained by feature space A

$A \cup B$


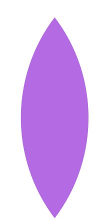
all variance explained by either A or B

B


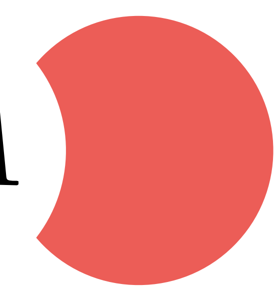
variance explained by feature space B

$A \setminus B$


variance explained by feature space A that isn’t explained by B

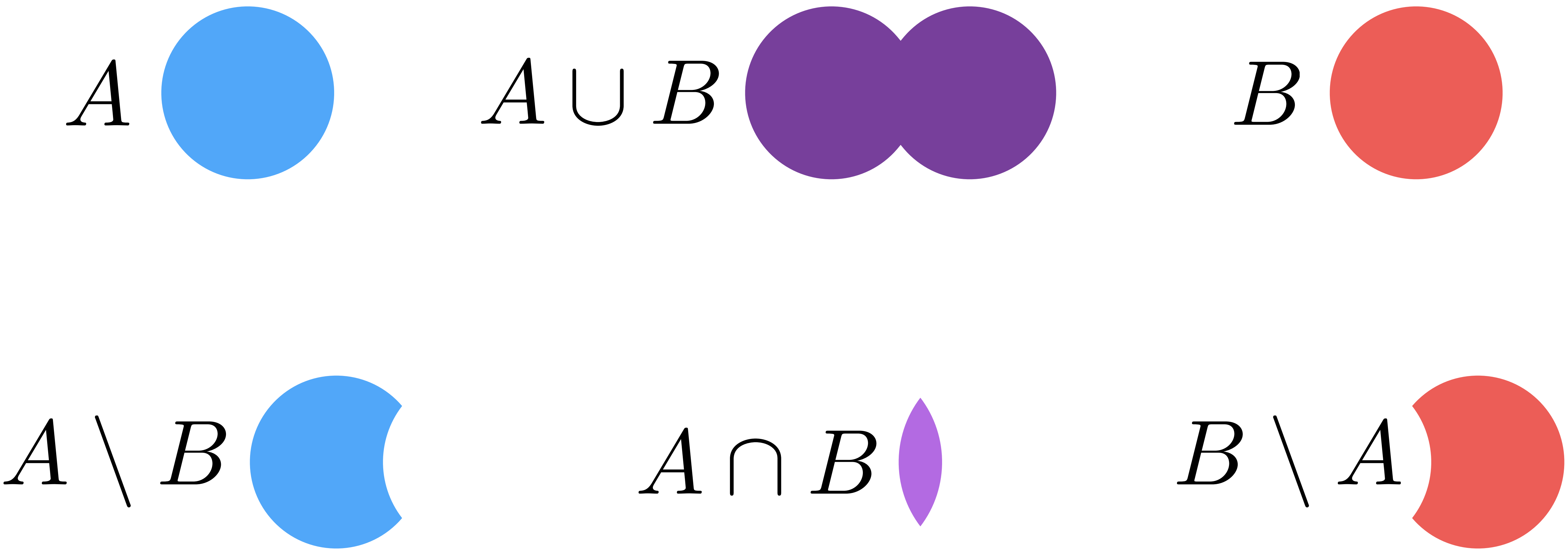
$A \cap B$


variance explained by both B and A

$B \setminus A$


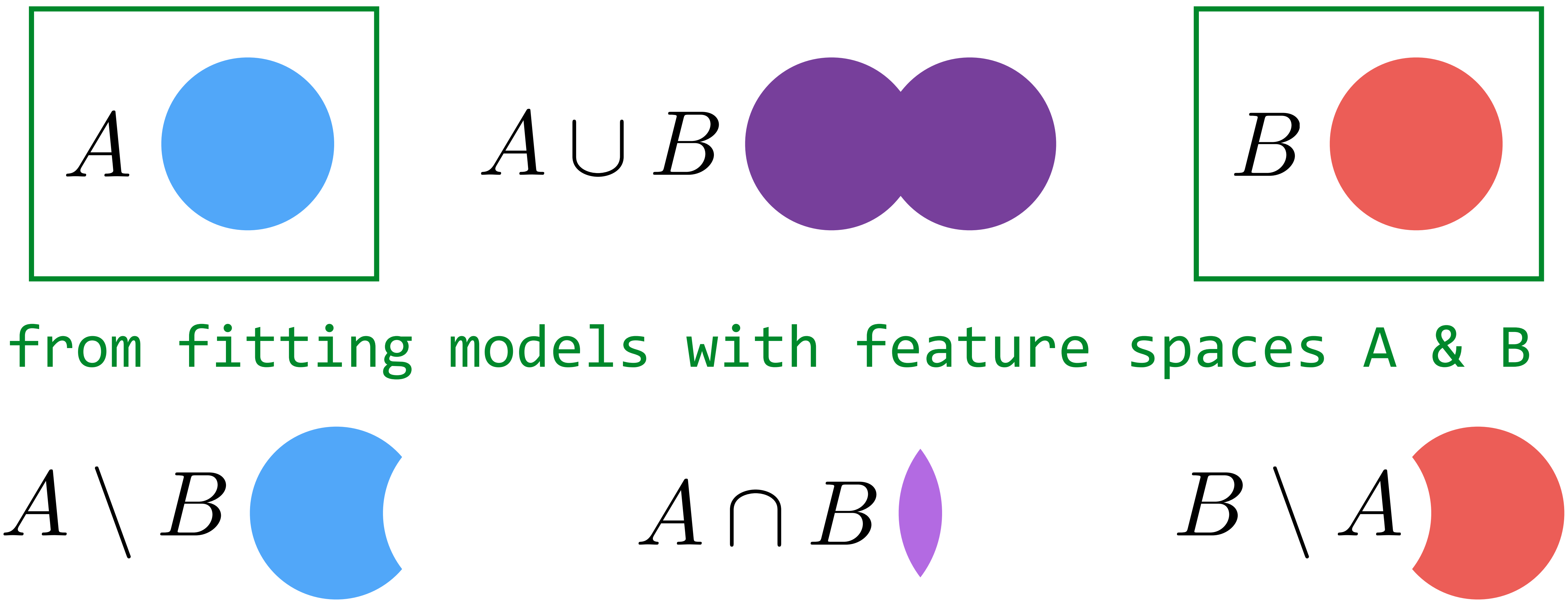
variance explained by feature space B that isn’t explained by A

VARIANCE PARTITIONING

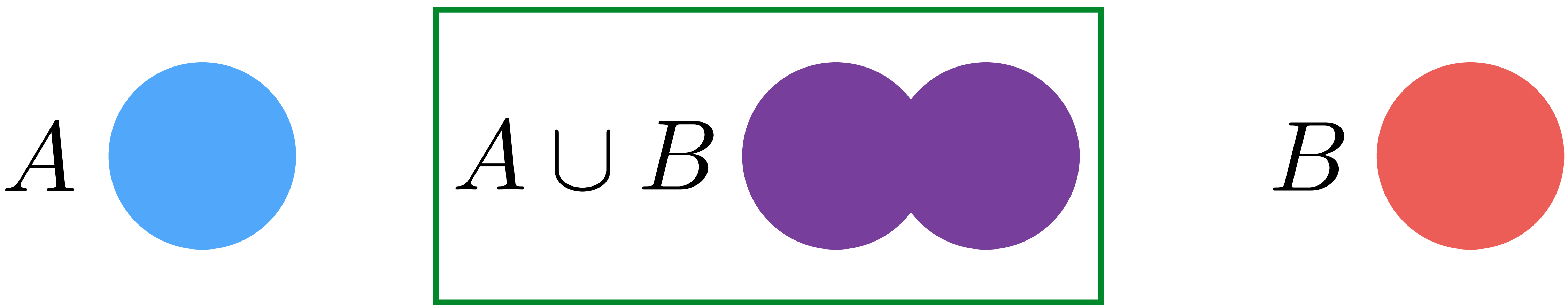


How do we measure the sizes of these partitions?

VARIANCE PARTITIONING



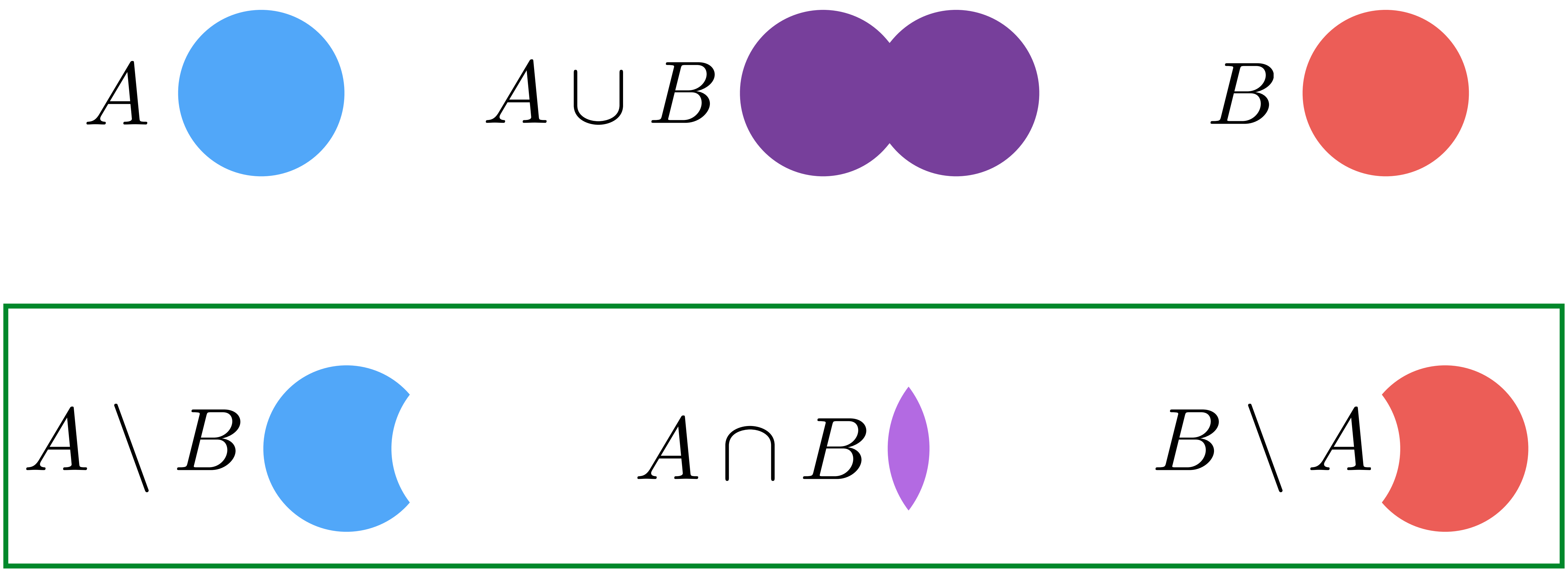
VARIANCE PARTITIONING



to get this we fit a model
with both feature spaces

$$y = \begin{bmatrix} X_A \\ X_B \end{bmatrix} \beta_{AB} + \epsilon$$

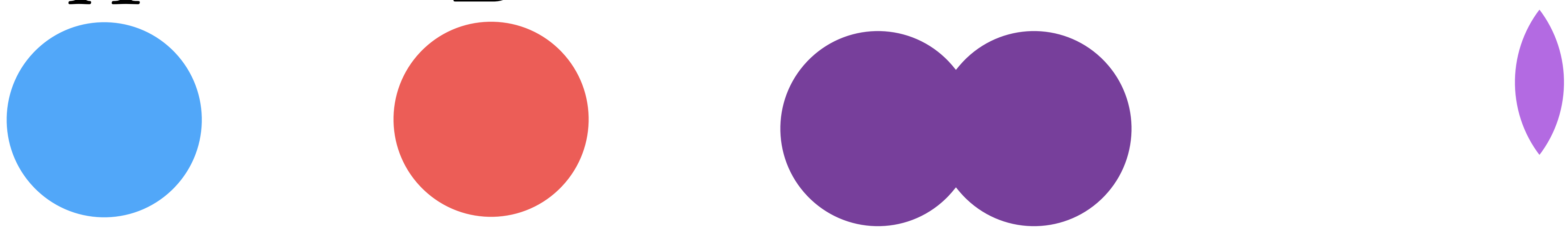
VARIANCE PARTITIONING



to get these we have to use set theory

VARIANCE PARTITIONING

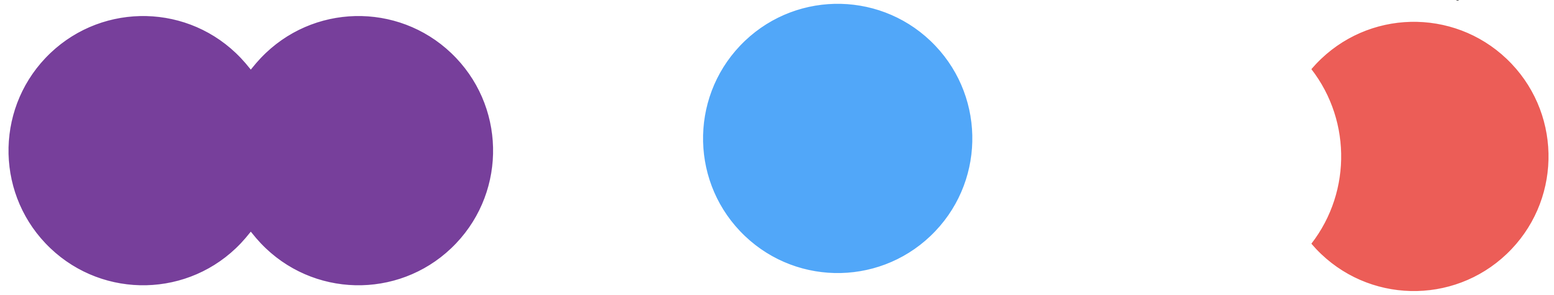
$$A \quad + \quad B \quad - \quad A \cup B \quad = \quad A \cap B$$



how much variance is explained by both A & B

VARIANCE PARTITIONING

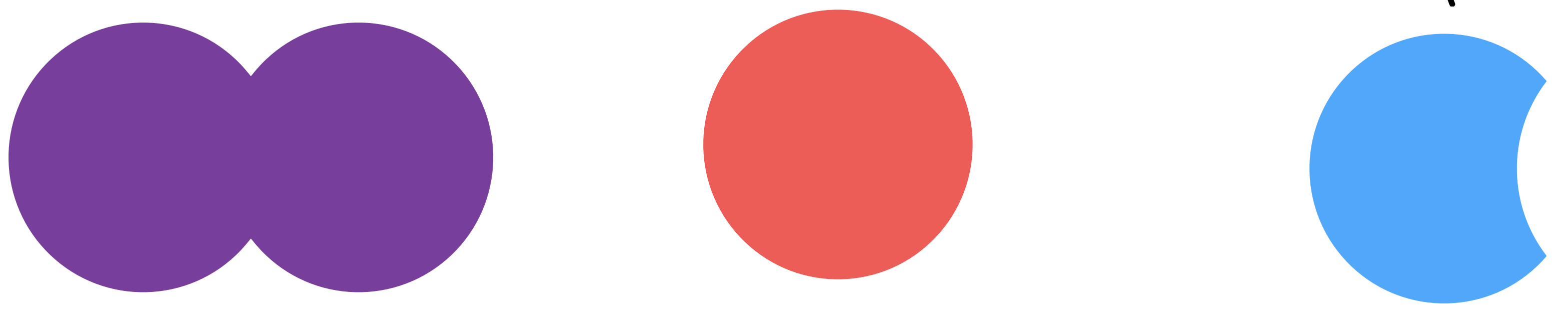
$$A \cup B \;-\; A \;=\; B \setminus A$$



how much variance is explained by B that isn’t explained by A

VARIANCE PARTITIONING

$$A \cup B \quad - \quad B \quad = \quad A \setminus B$$



how much variance is explained by A that isn't explained by B

VARIANCE PARTITIONING

* The same theory can be applied for **any number** of feature spaces

* But the number of models fit increases quickly

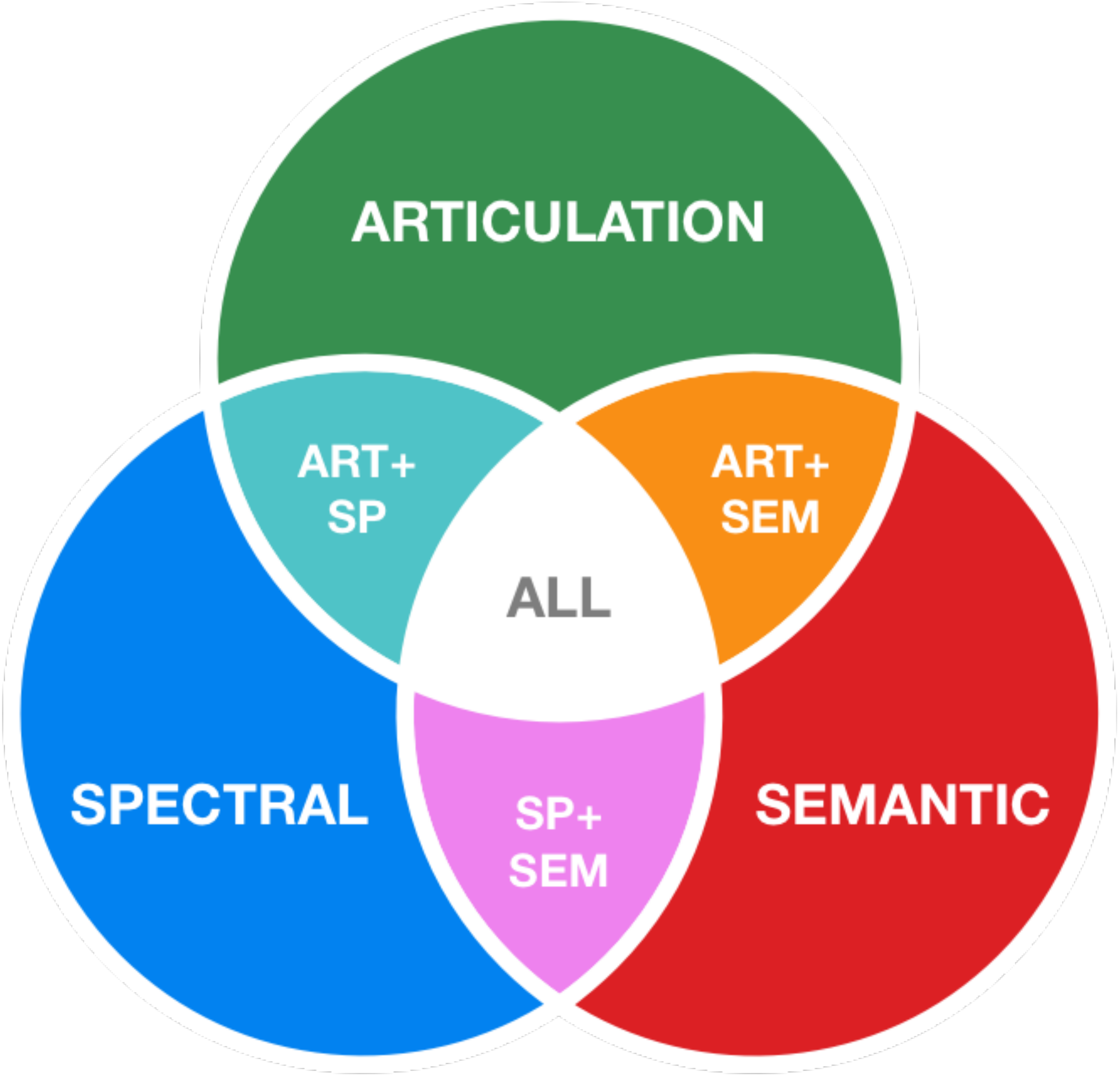
  * 2 feature spaces = 3 models

  * 3 feature spaces = 7 models

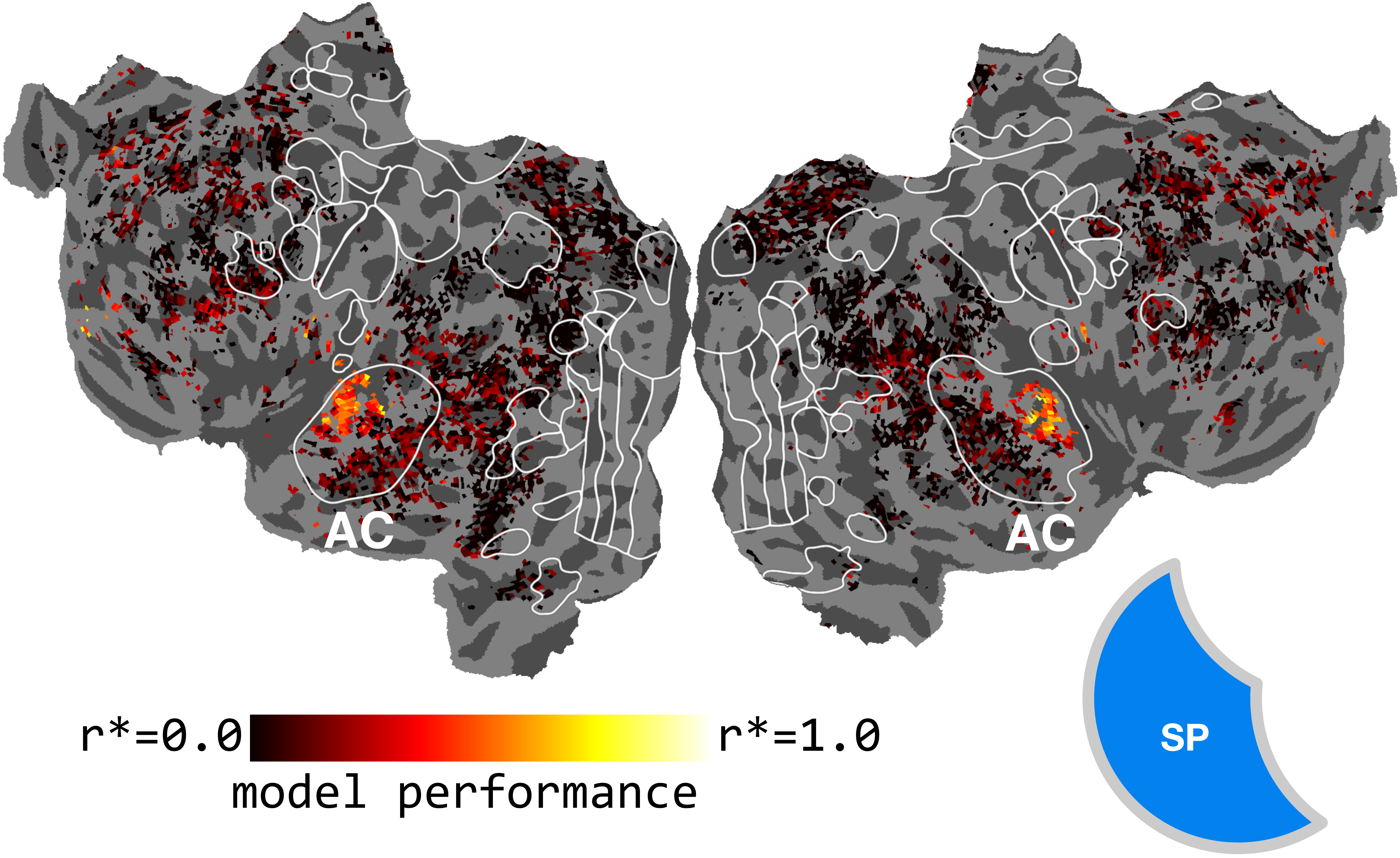
  * 4 feature spaces = 15 models

  * n feature spaces = 2^n-1 models

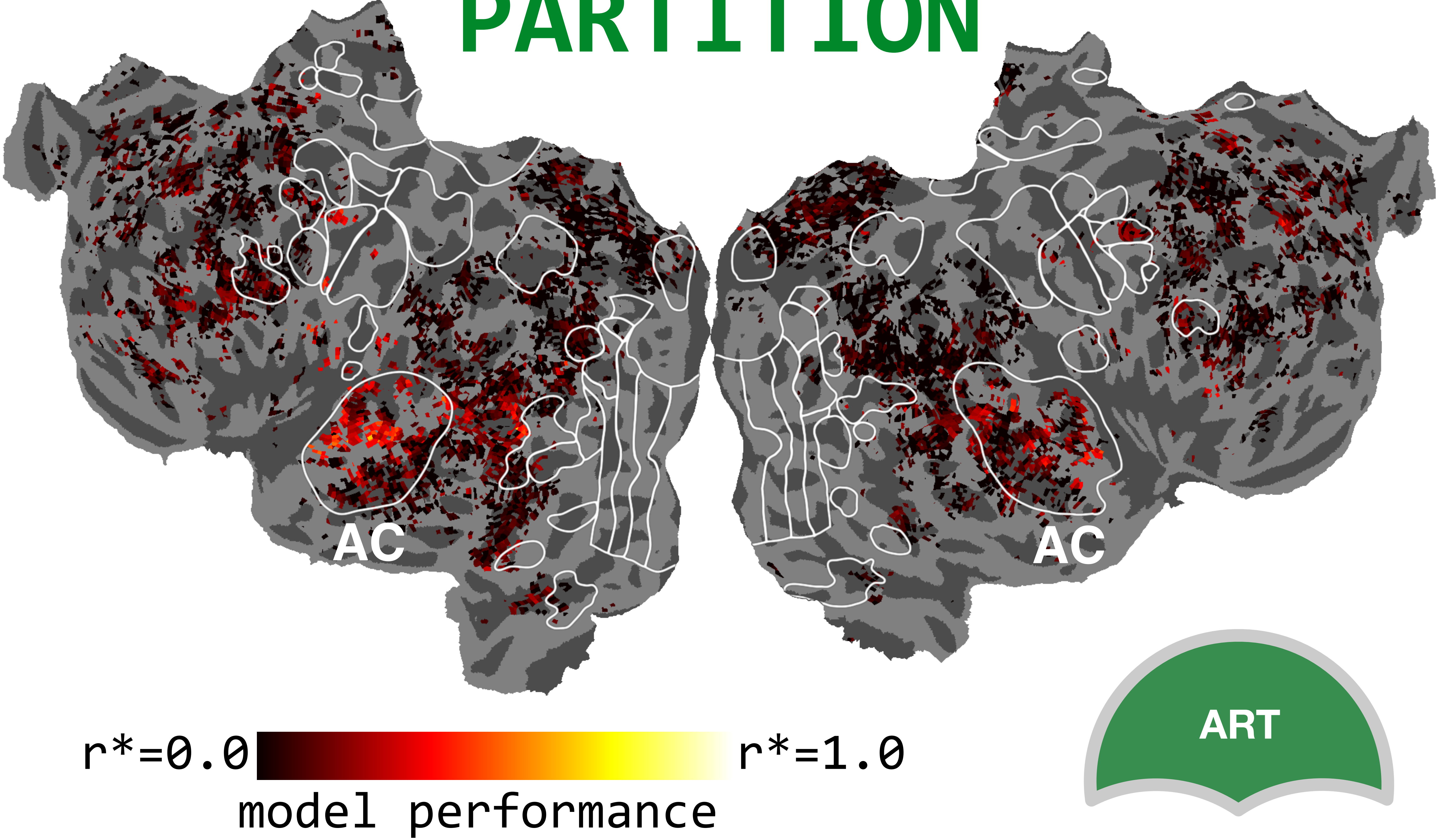
SPEECH MODELS

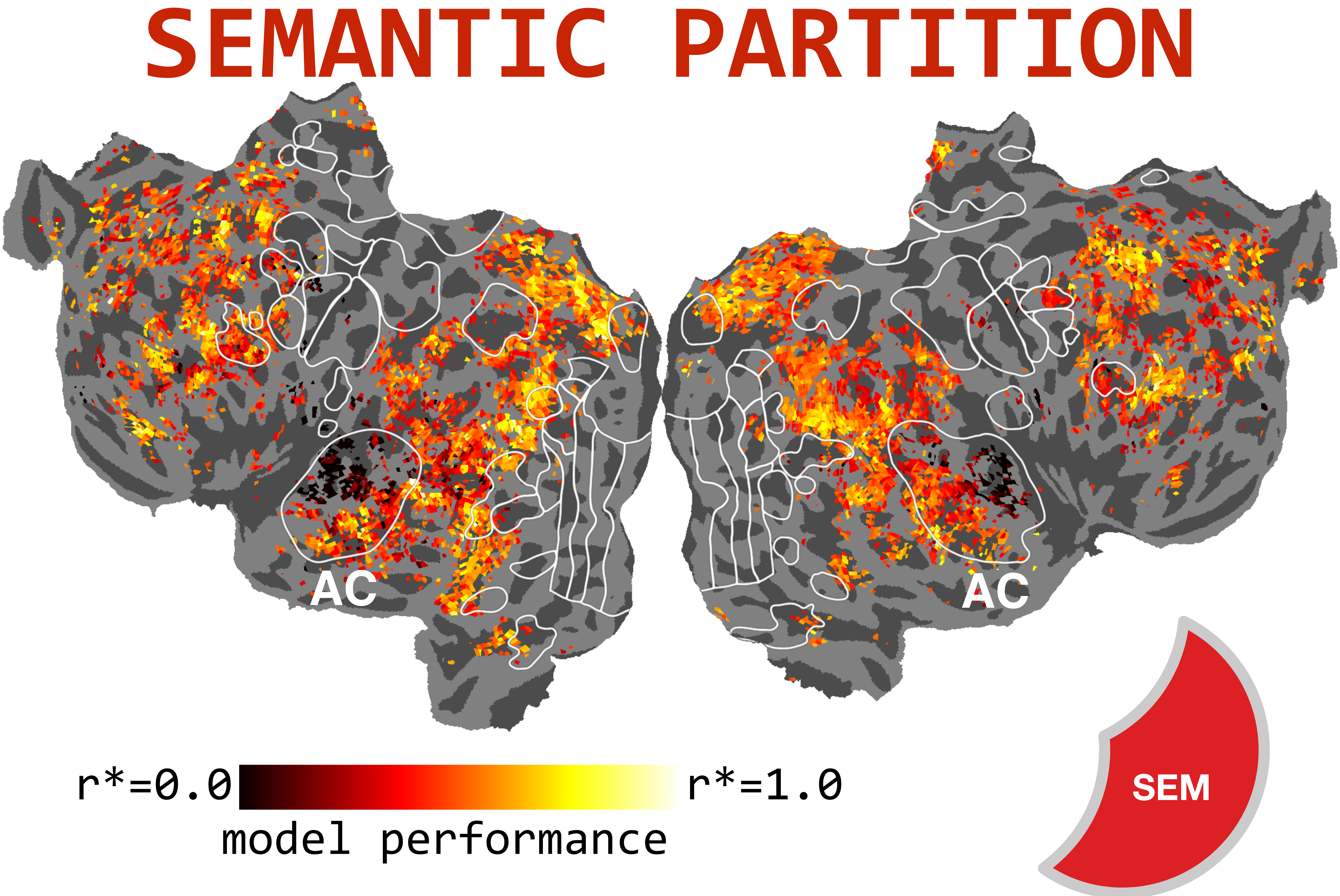


SPECTRAL PARTITION

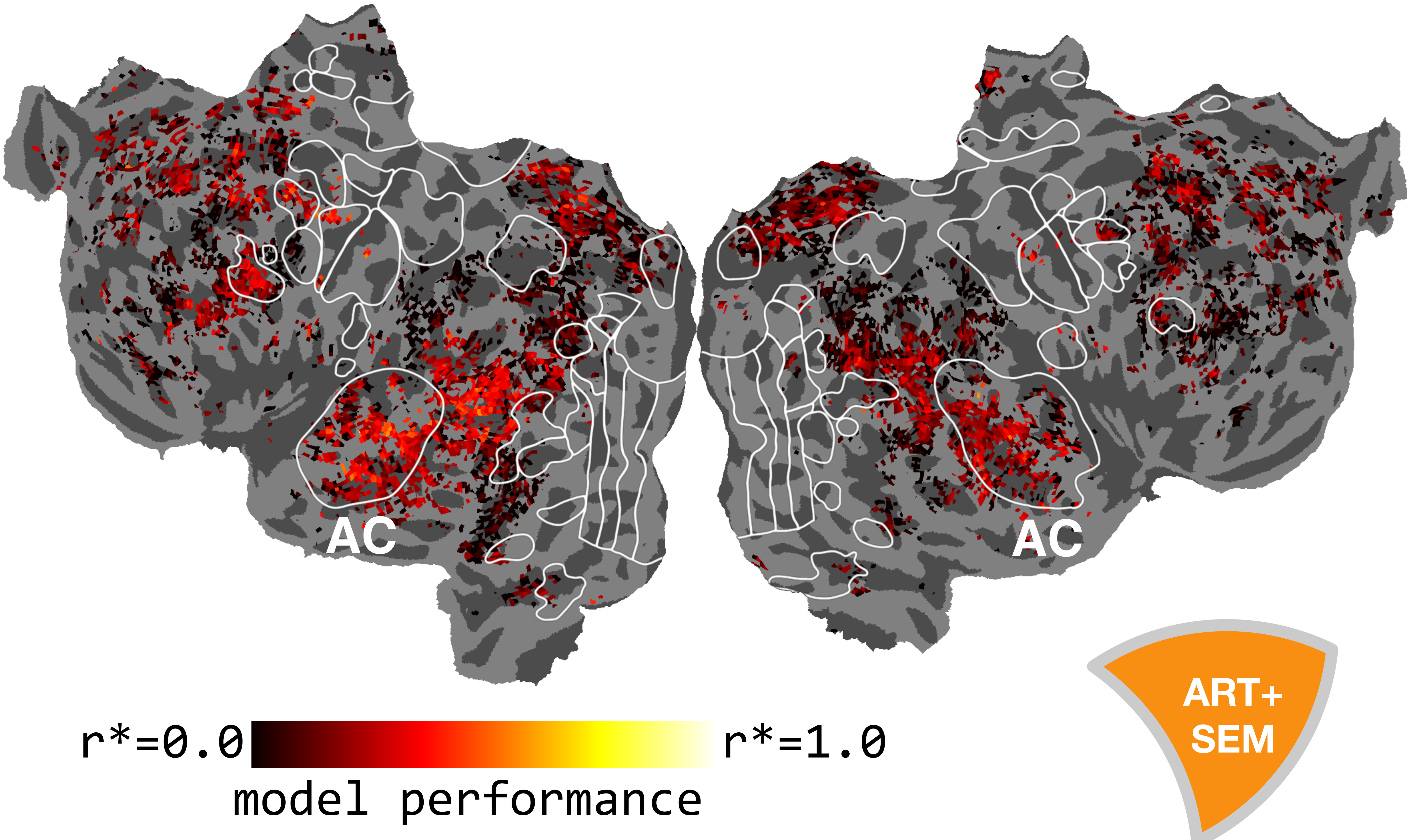


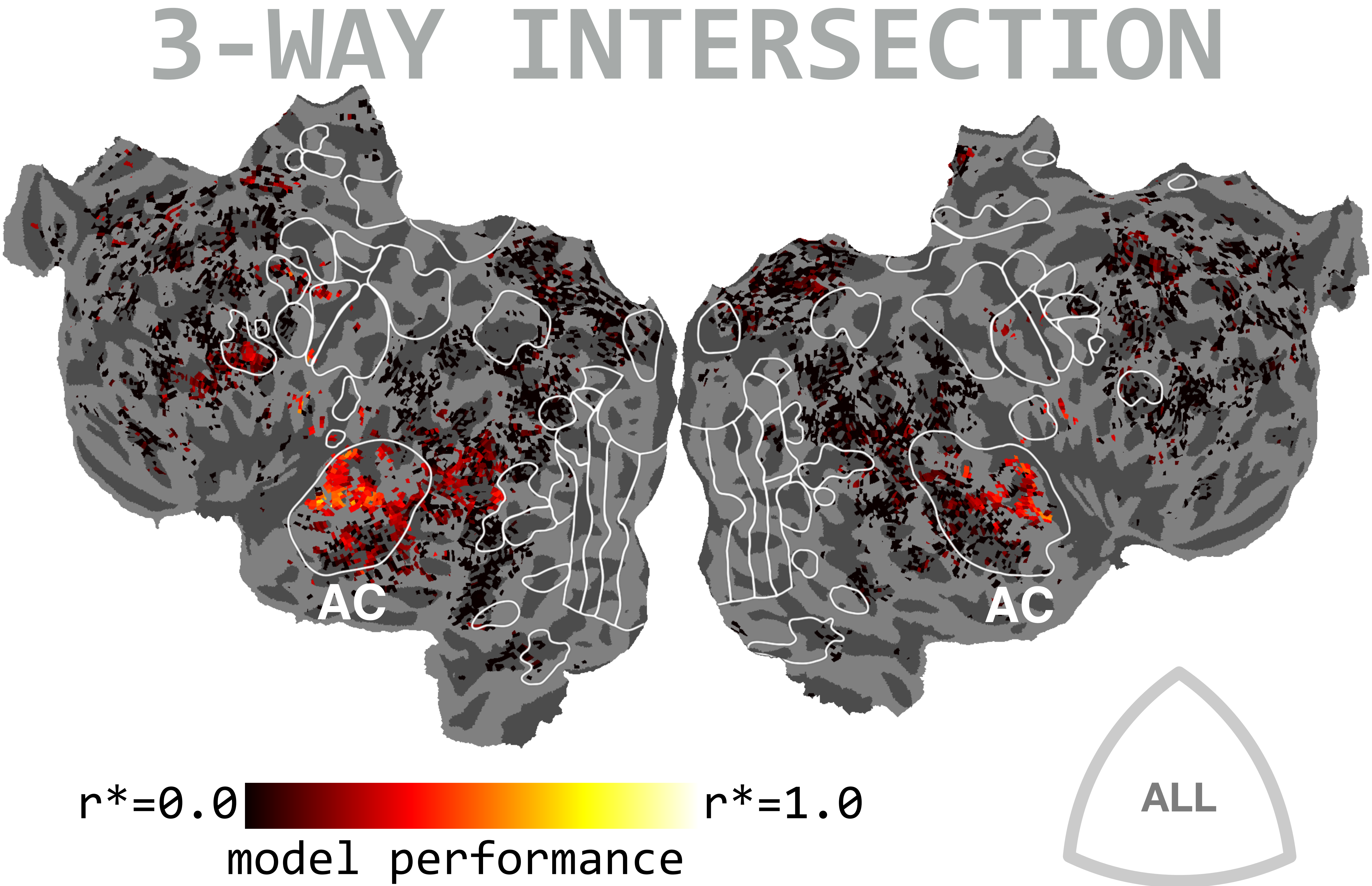
ARTICULATION PARTITION




SEMANTIC PARTITION
AC
AC
r*=0.0
r*=1.0
model performance
SEM

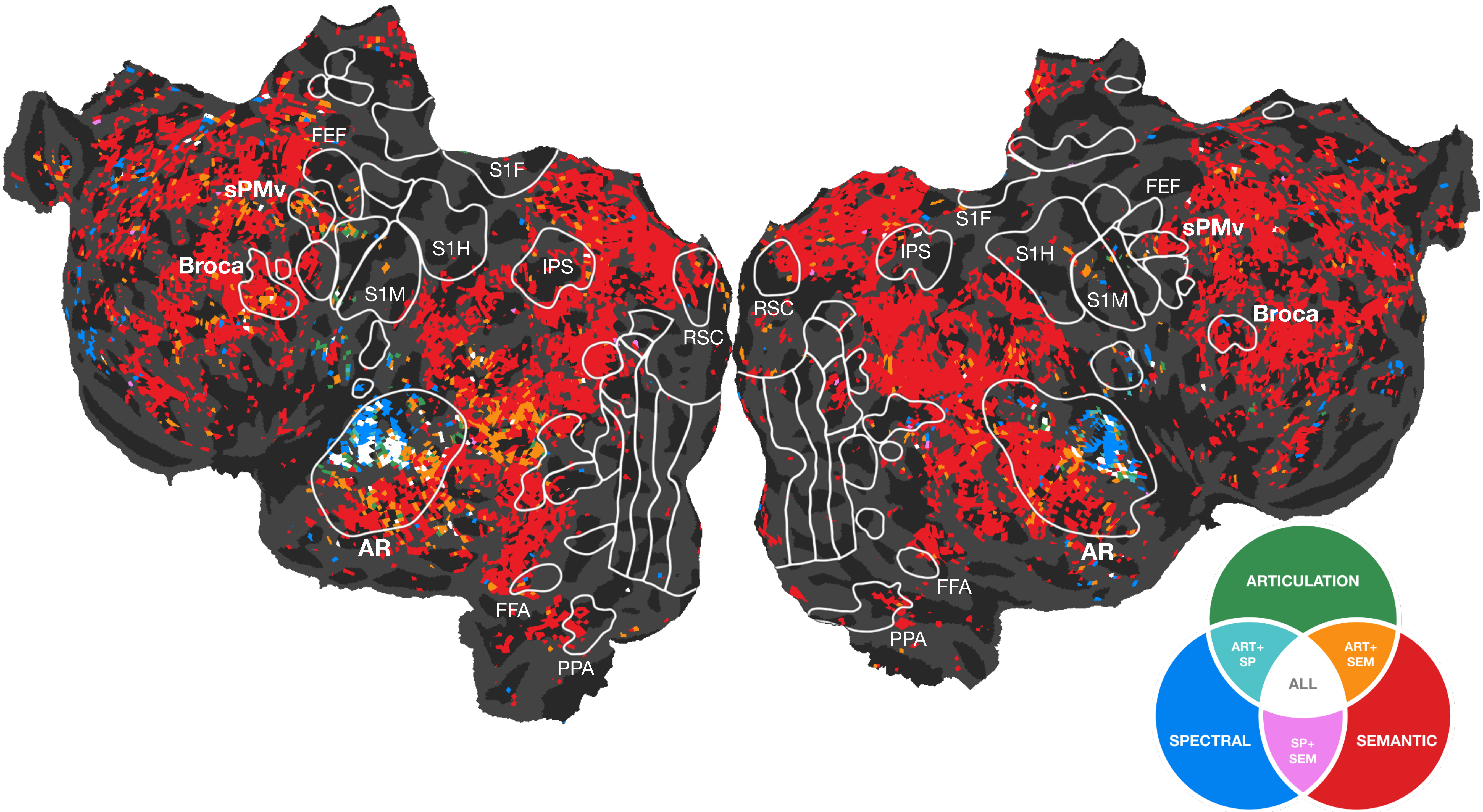
ART&SEM PARTITION





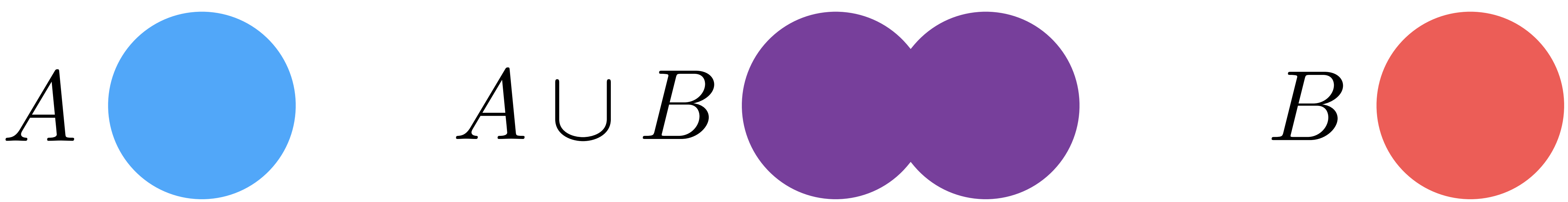
3-WAY INTERSECTION
AC
AC
r*=0.0
r*=1.0
model performance
ALL

LARGEST PARTITION PER VOXEL



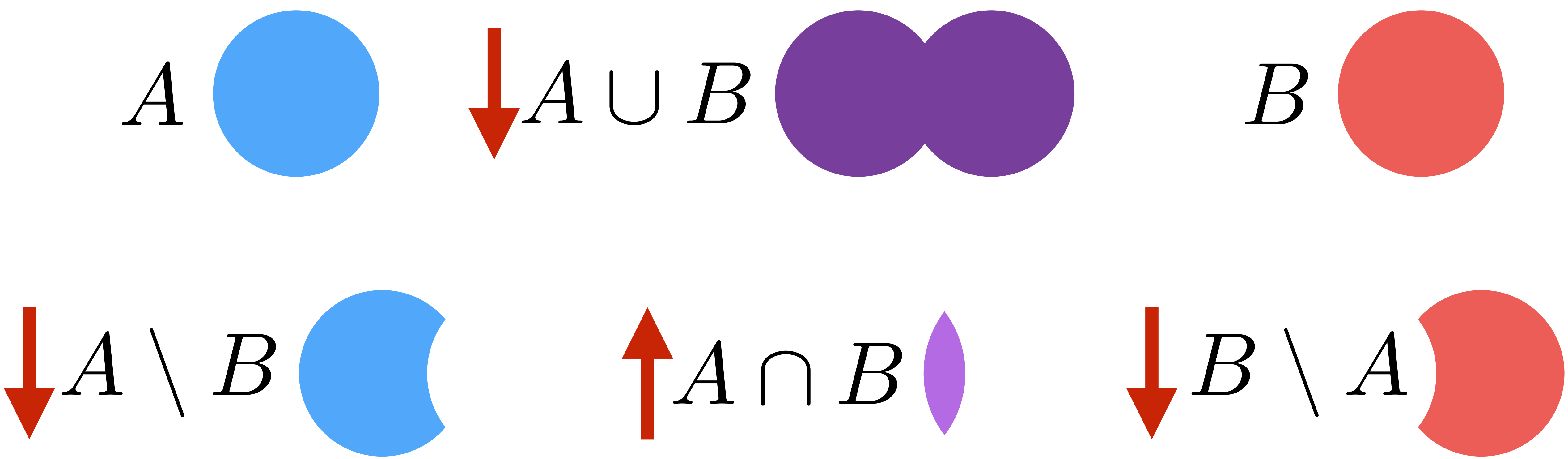
VARIANCE PARTITIONING

* Which of these models is most overfit?



VARIANCE PARTITIONING

* What effect does overfitting have on each partition?



VARIANCE PARTITIONING

* Best defense is to minimize overfitting: put a lot of effort into quality of AUB

* ***But!*** It is also possible to de-bias, based on the assumption that no variance partition should be negative

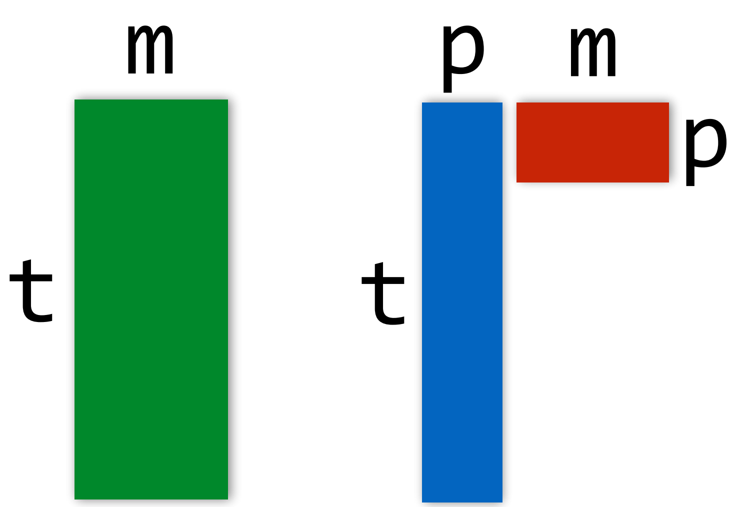
VARIANCE PARTITIONING

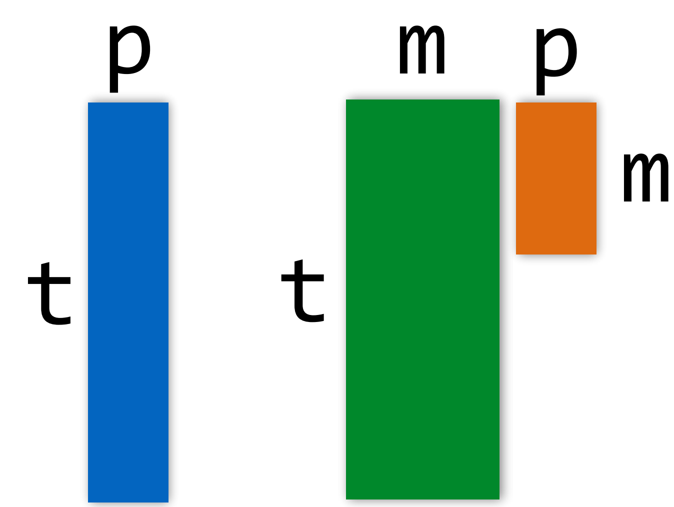
* Extremely valuable technique

* Tells you how well your hypotheses (feature space) can be distinguished using one particular dataset

ENCODING V DECODING

$$Y = X\beta + \epsilon \qquad X = Y\gamma + \nu$$





* **Encoding:** predict *Y* from *X*

* **Decoding:** predict *X* from *Y*

ENCODING V DECODING

$$Y = X\beta + \epsilon \qquad\qquad X = Y\gamma + \nu$$

* Both tell you **whether** there is a relationship between *X* and *Y*

* But interpretation can be very different!

ENCODING V DECODING

$$Y = X\beta + \epsilon \qquad\qquad X = Y\gamma + \nu$$

* **Our goal:** find what kind of information is represented/processed by each part of the brain

* Which is more useful, **beta** or **gamma**?

ENCODING V DECODING

* For one voxel
    * If that voxel has a large encoding weight for a given feature, that voxel probably responds to that feature
    * If that voxel has a large decoding weight for a given feature, that voxel doesn’t necessarily respond to that feature

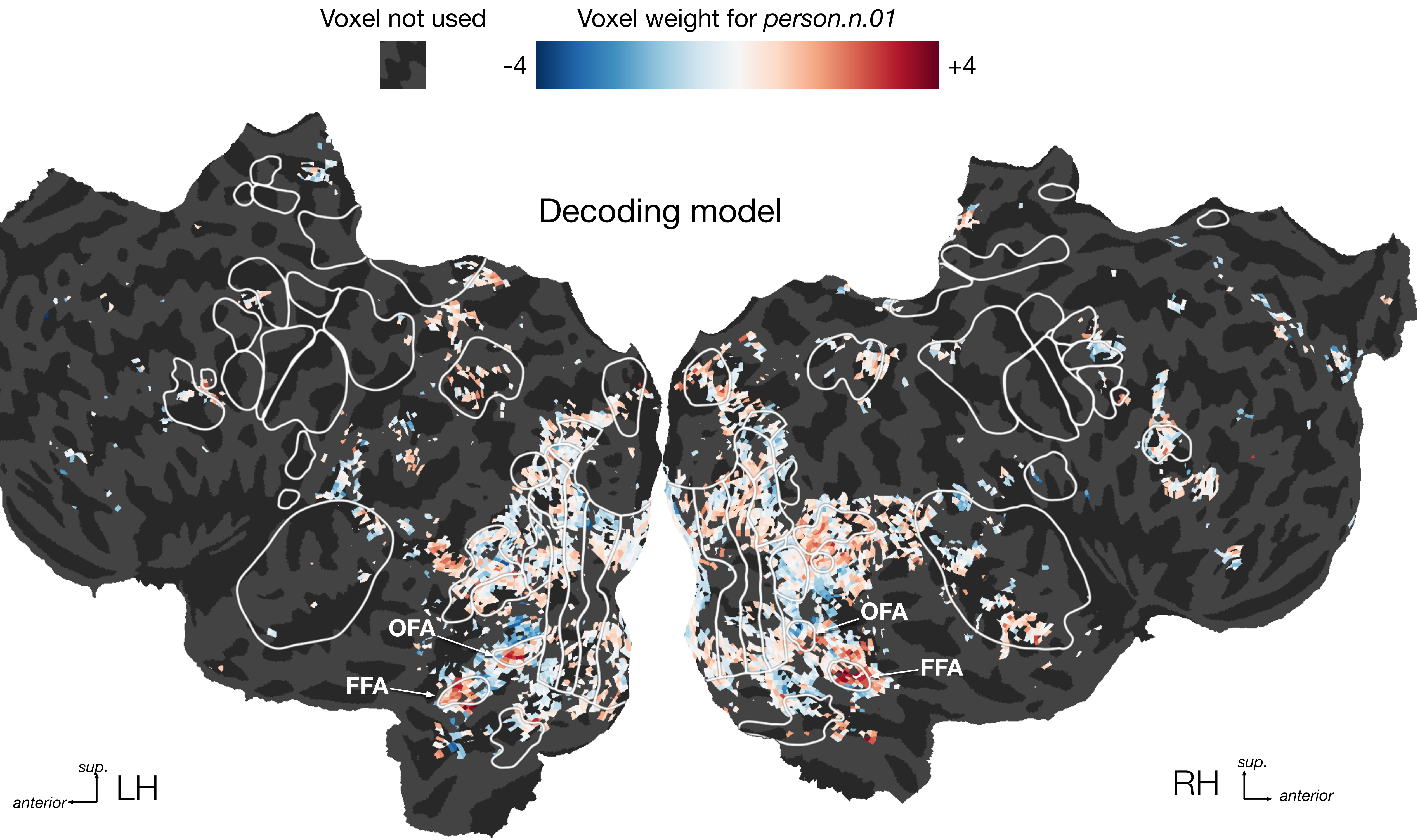
ENCODING V DECODING

* Voxel can have large weight for a feature **in decoding** if:
  * The voxel actually responds to the feature
  * The voxel is (anti)correlated with noise in voxels that ***do*** respond to the feature

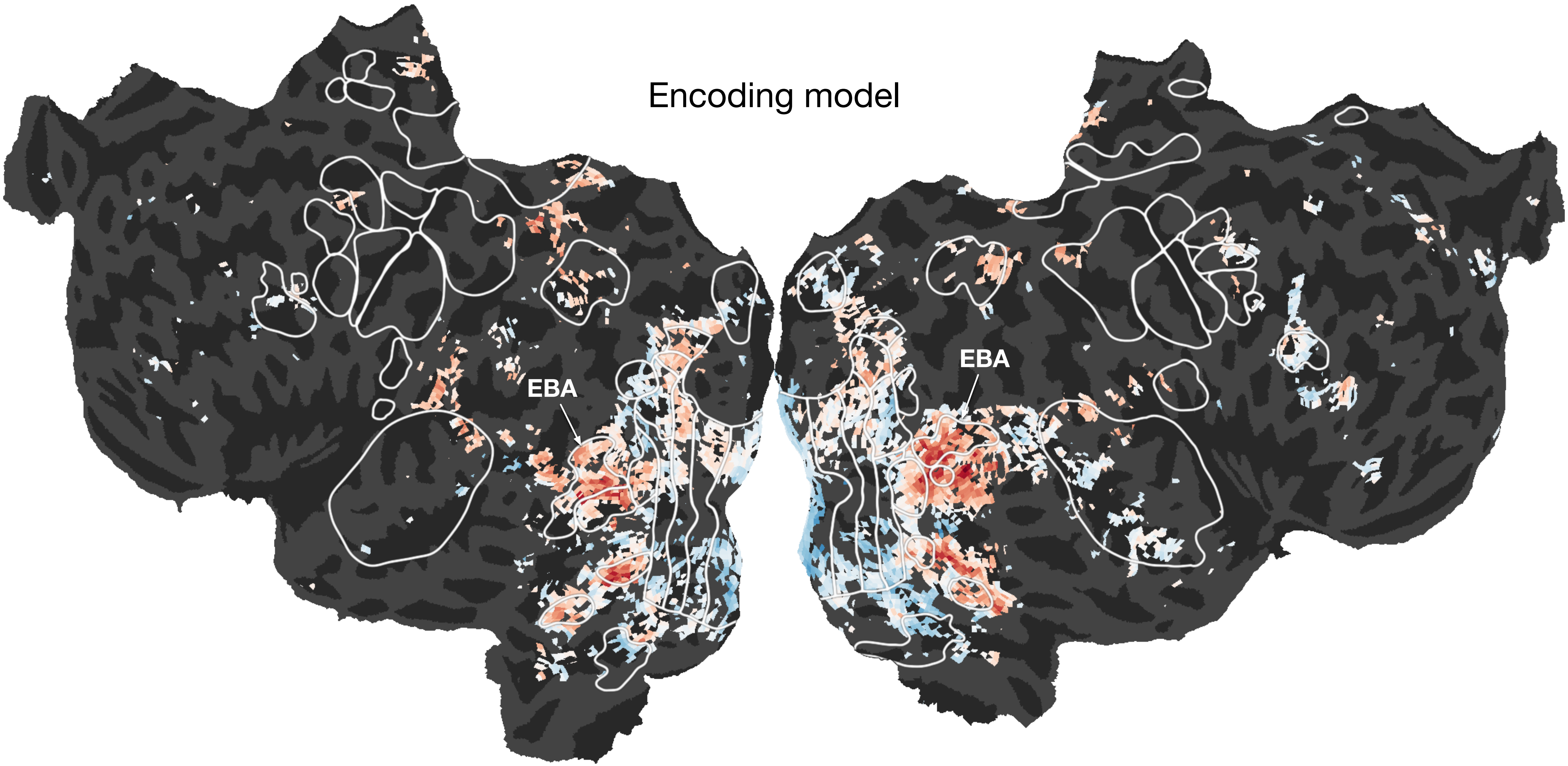
ENCODING V DECODING

* Voxel can have small weight **in decoding** for a feature that it *actually responds to* if:
  * The voxel also responds to other features, so it isn’t useful for decoding that feature

ENCODING V DECODING



ENCODING V DECODING



ENCODING V DECODING

* Encoding assumes that all Y are independent

* Decoding assumes that all X are independent

* There are some cases where assuming Y are independent is bad (especially MEG, EEG)

DATA VISUALIZATION

“Above all else, *show the data*”

—Edward Tufte

EDWARD TUFTE

* *The Visual Display of Quantitative Information* (1983)

DATA VISUALIZATION

* The goal is to **communicate data**

* Avoid *distraction*

* Avoid *deception*

AVOID DISTRACTION

* Tufte: **“data-ink ratio”**

* Maximize the amount of data communicated per drop of ink used

Remove to improve
the **data tables** edition

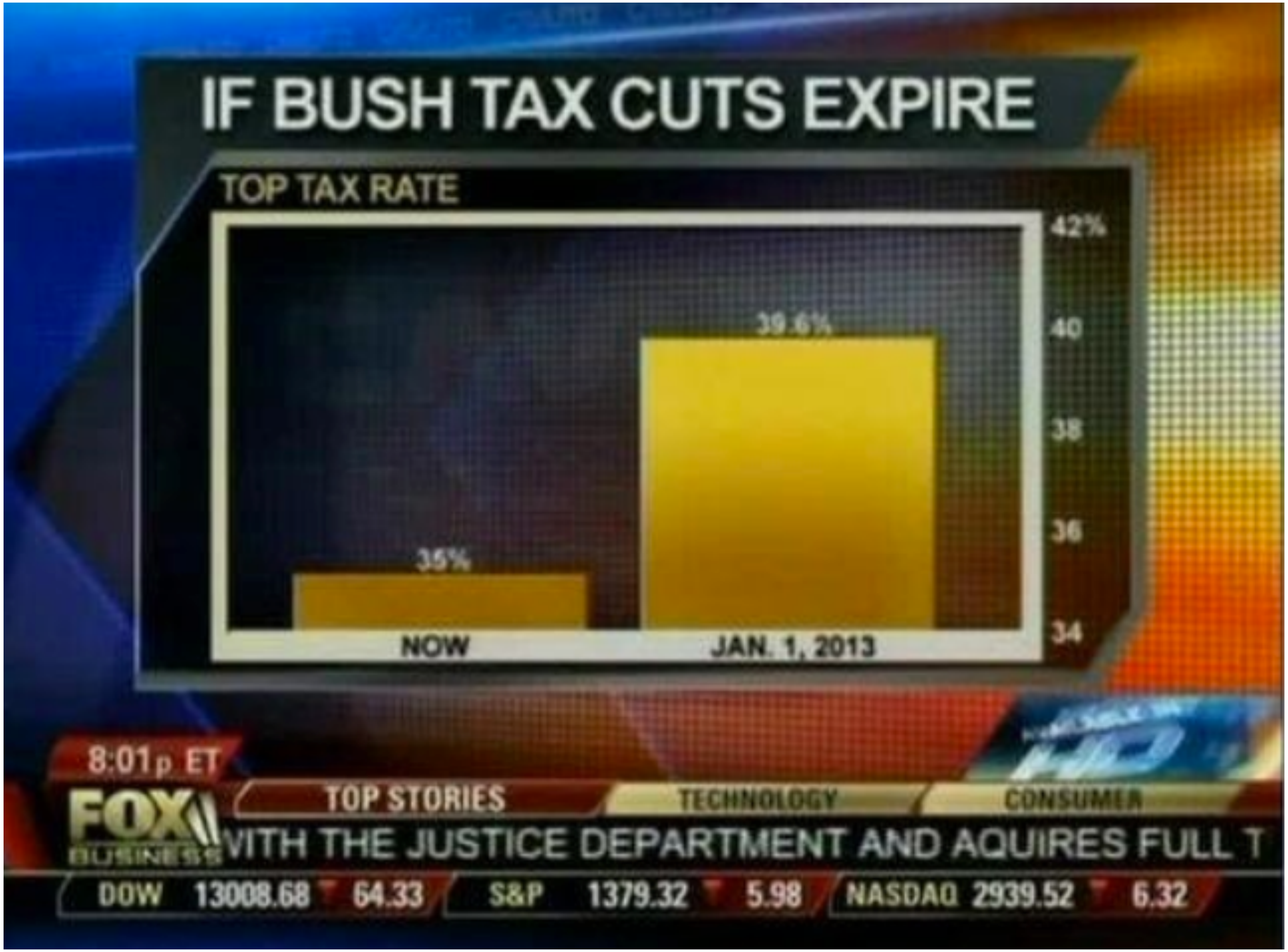
Created by **Darkhorse Analytics**   www.**darkhorseanalytics**.com

AVOID DISTRACTION

* … but don’t take it too far!

* Don’t sacrifice ***understandability*** for ***beauty***

AVOID DECEPTION



KENNETH MORELAND

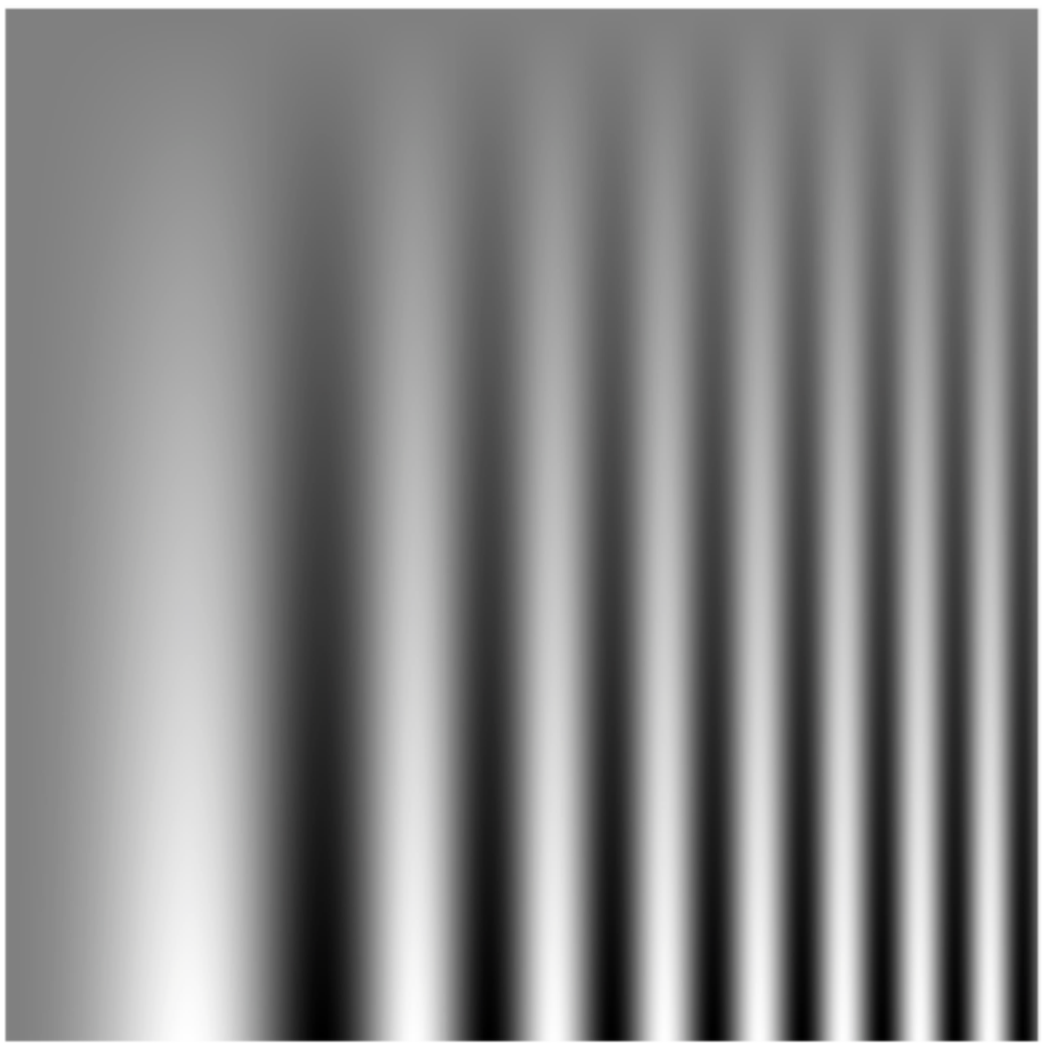
* *Diverging Color Maps for Scientific Visualization*

Fig. 1. The rainbow color map. Know thy enemy.

https://cfwebprod.sandia.gov/cfdocs/CompResearch/docs/ColorMapsExpanded.pdf

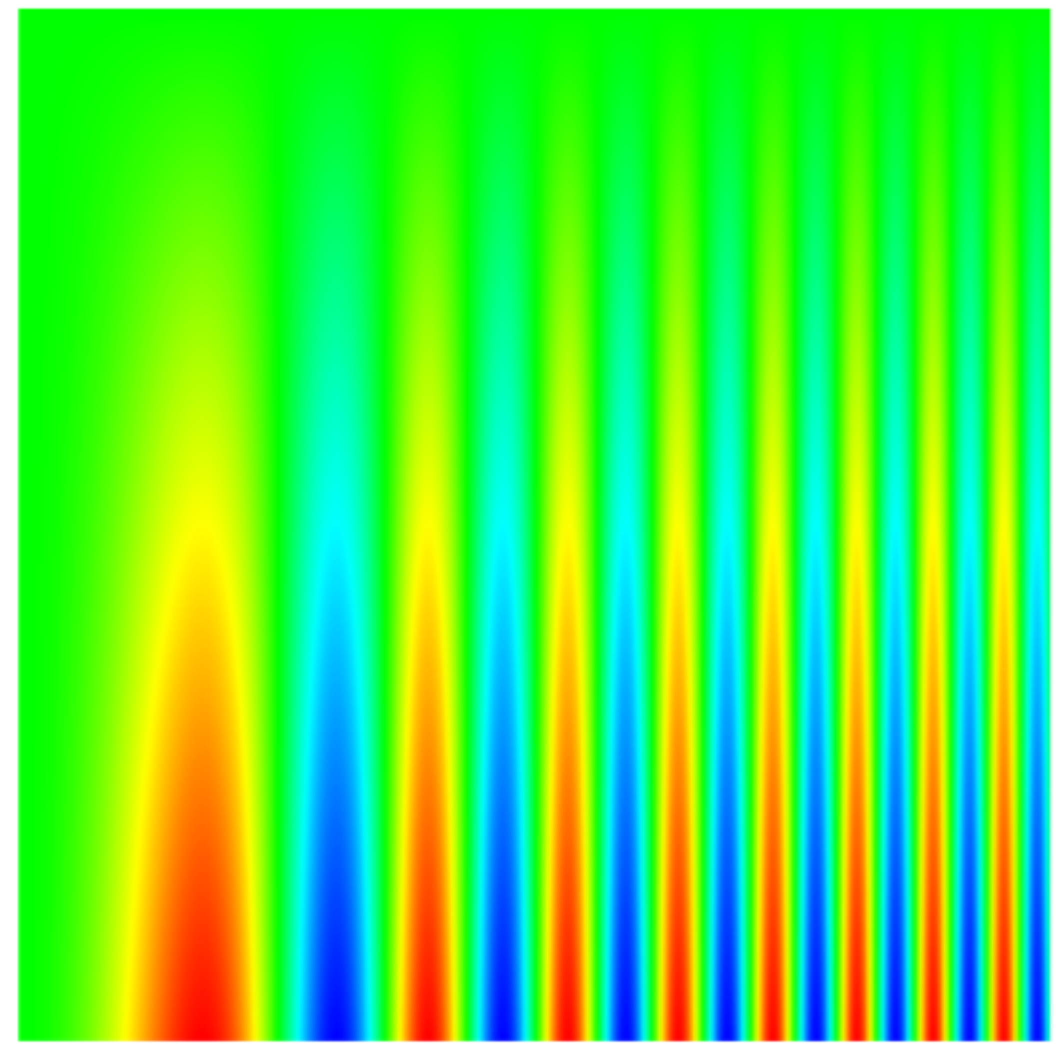
COLORMAPS

grayscale

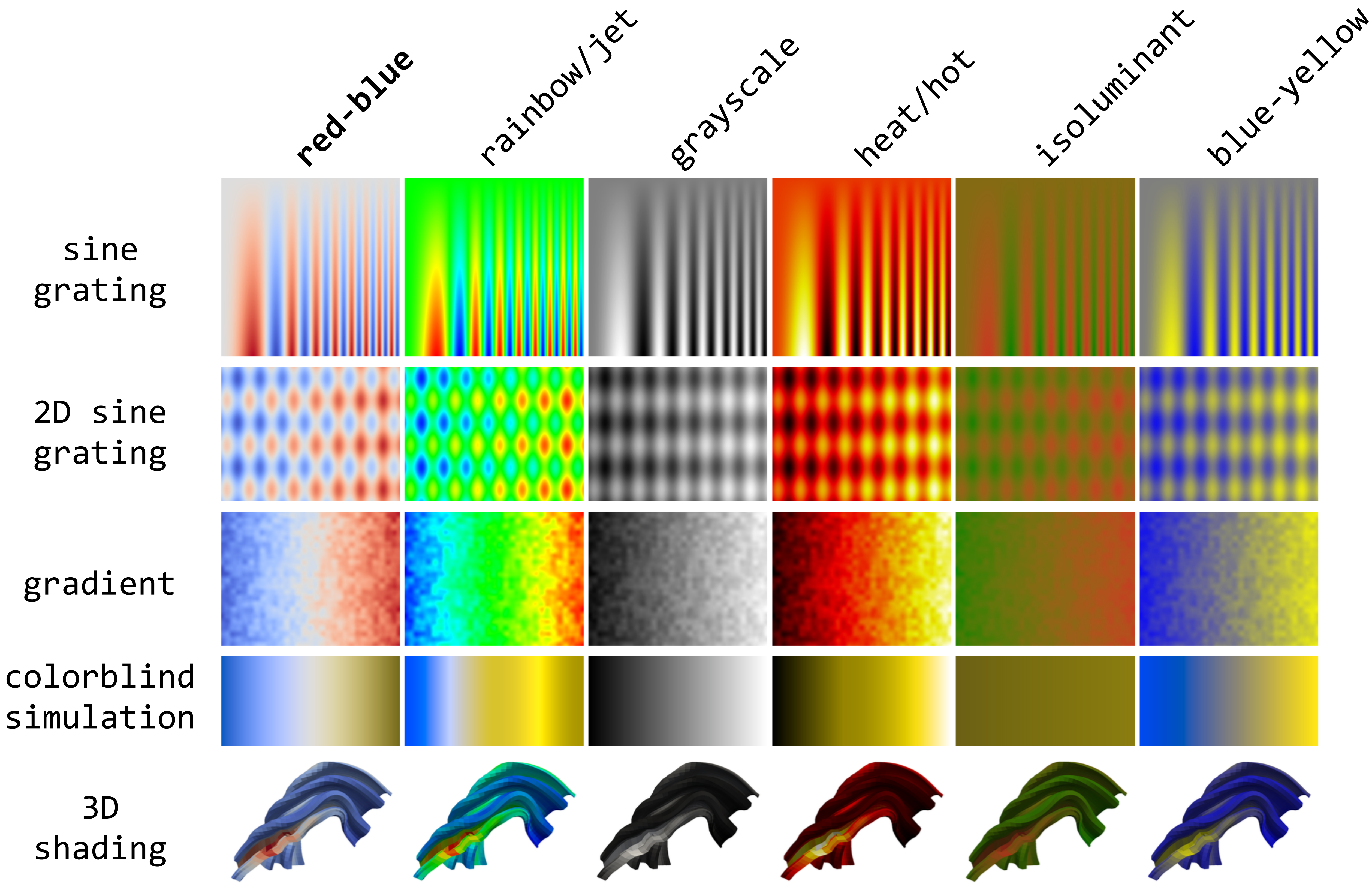


👍

rainbow/jet



👎


red-blue
rainbow/jet
grayscale
heat/hot
isoluminant
blue-yellow
sine grating
2D sine grating
gradient
colorblind simulation
3D shading

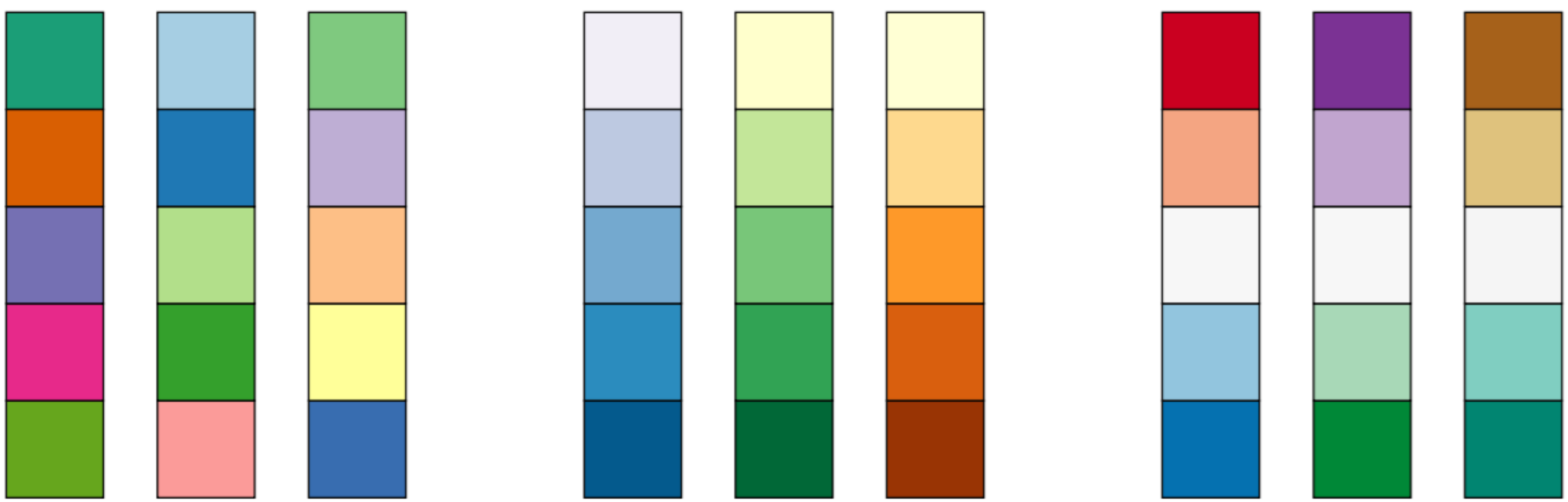
CYNTHIA BREWER

* *ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps*

* Colormaps available in matplotlib (python) and as package for MATLAB

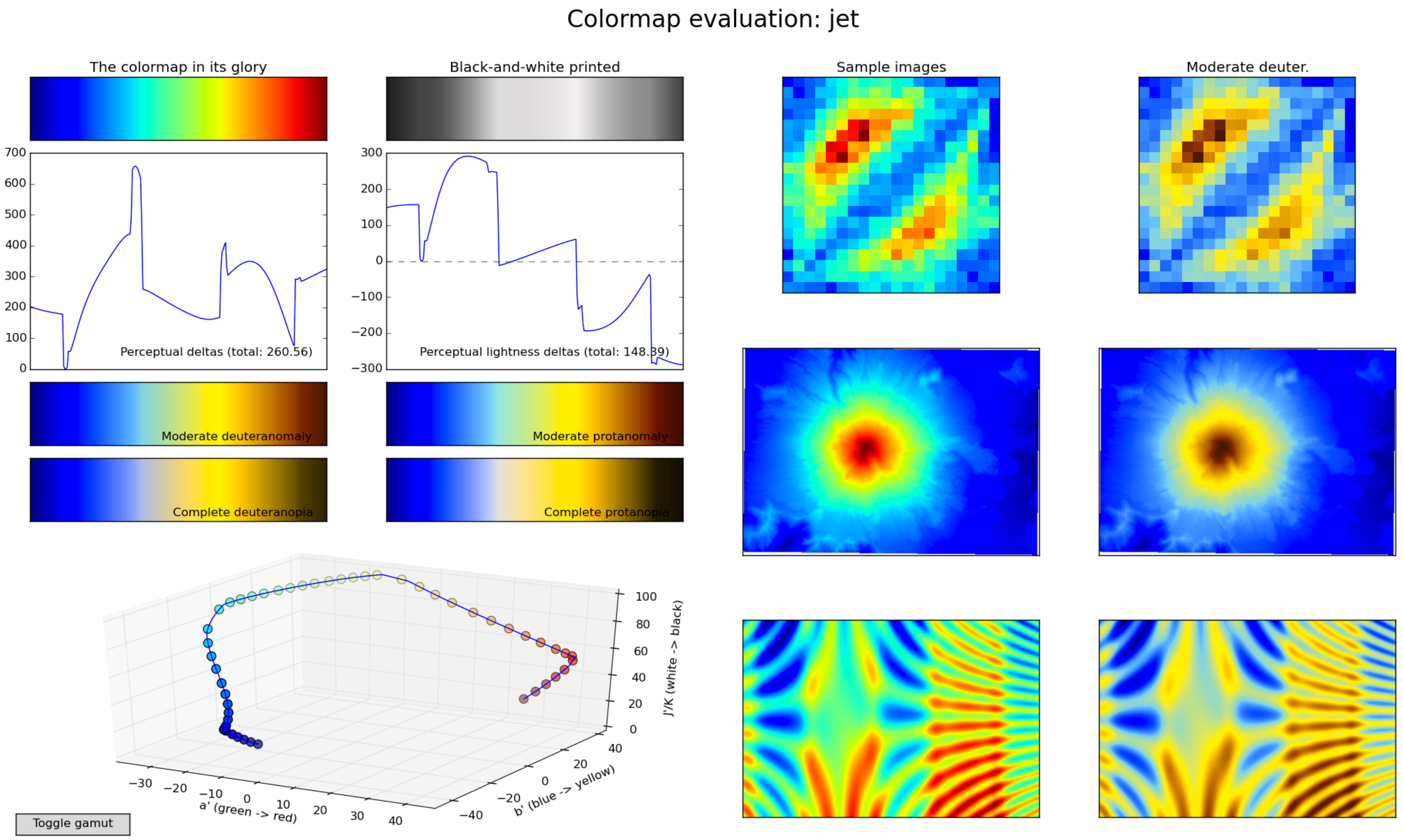
https://pdfs.semanticscholar.org/8c9a/f5bd12b36e450ba564f644008b871dba5cdf.pdf

CHOOSE THE RIGHT TYPE OF COLORMAP



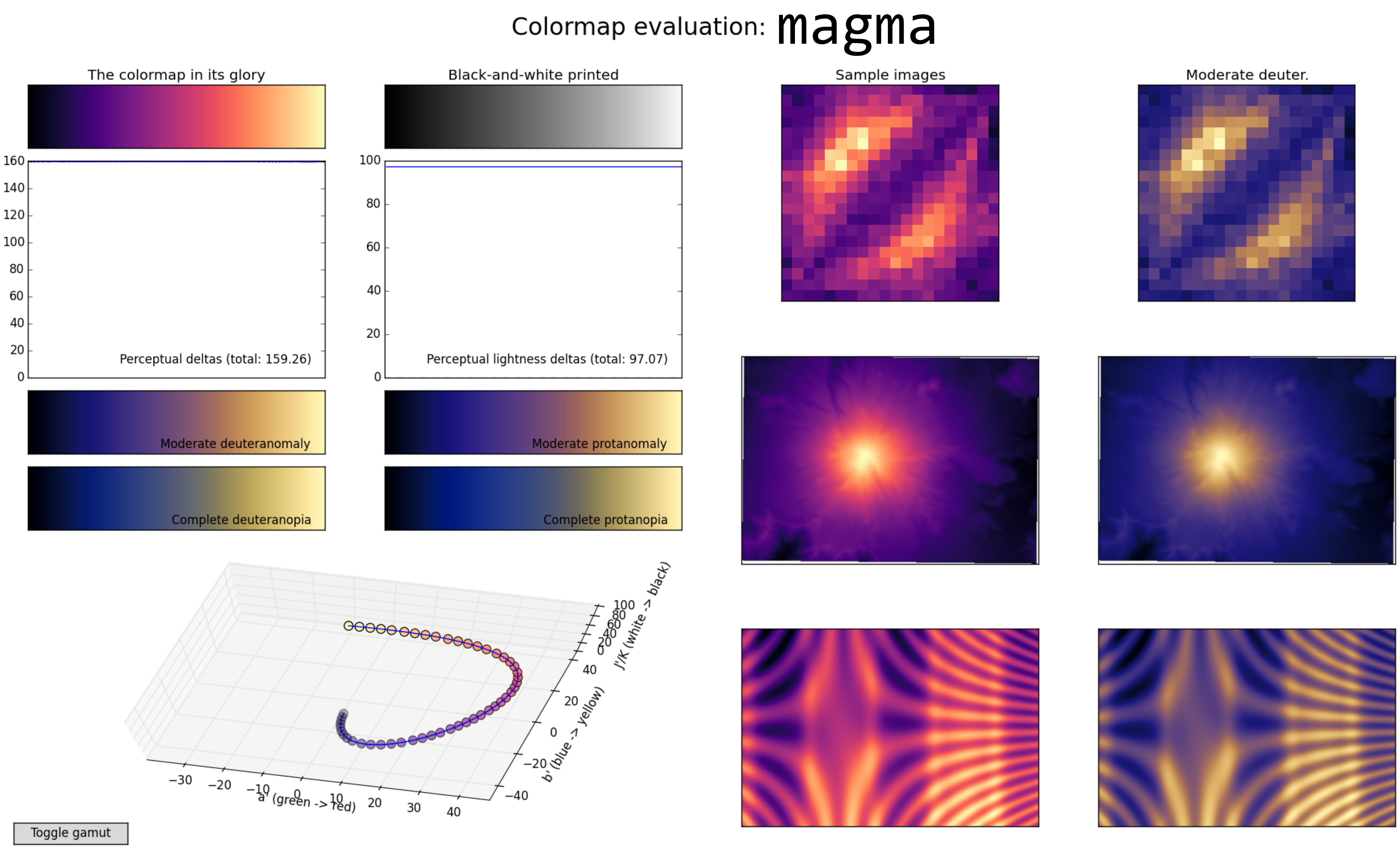
(a) Qualitative (b) Sequential (c) Diverging

PERCEPTUAL UNIFORMITY



https://bids.github.io/colormap/ S. Van der Walt, N. Smith

PERCEPTUAL UNIFORMITY



https://bids.github.io/colormap/

S. Van der Walt, N. Smith

THERE'S NO EXCUSE TO USE A BAD COLORMAP

THAT'S IT!

PREGUNTAS?

GRACIAS, CHICOS!