MODEL COMPARISON & VARIANCE PARTITIONING

Prof. Alexander Huth
3.5.2020

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

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

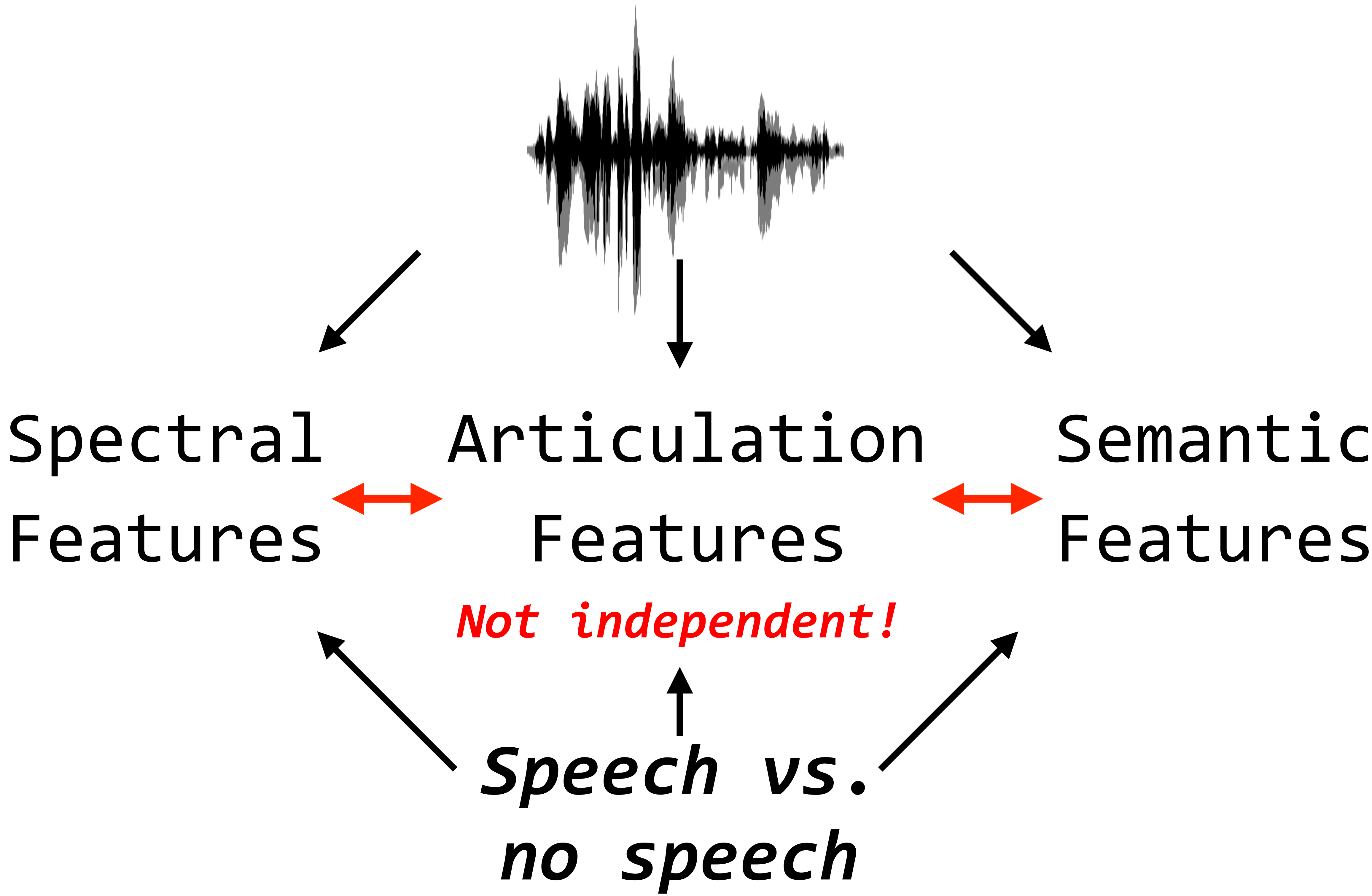
MANY LINEARIZED MODELS

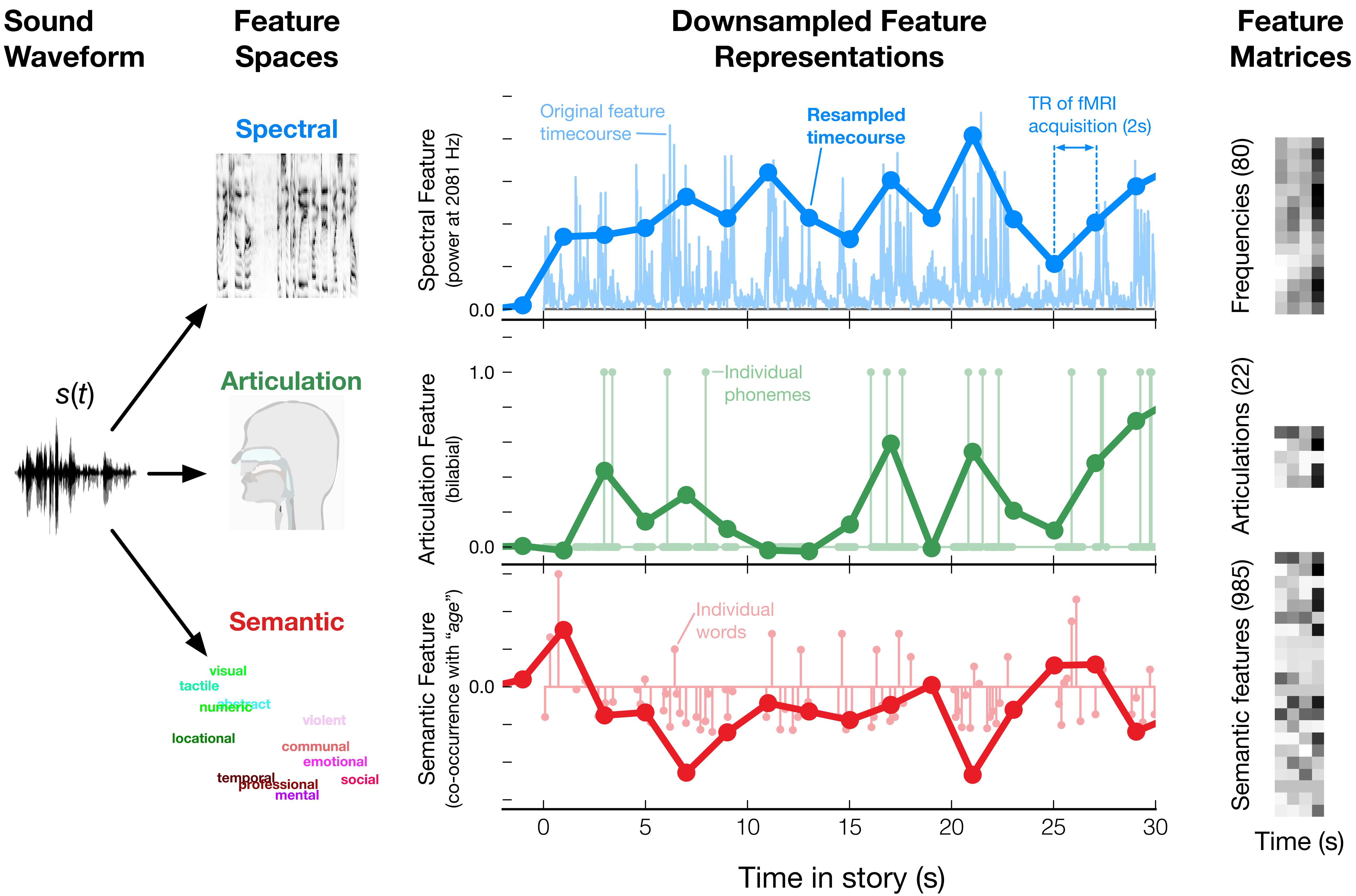
* Which feature space is **best**?

* Is that even the right question?

  * Sometimes, ***NO!***

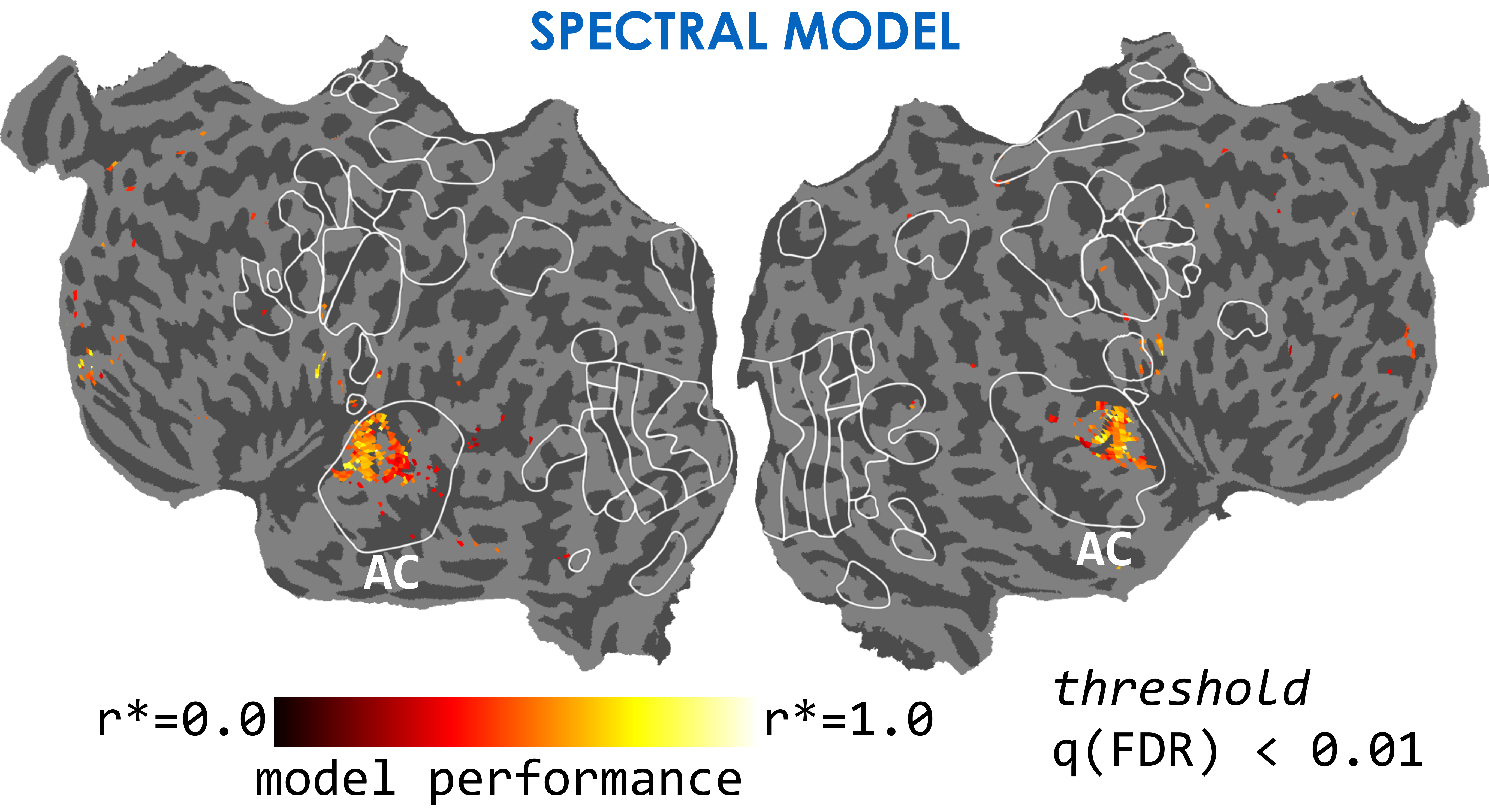
SPEECH MODELS





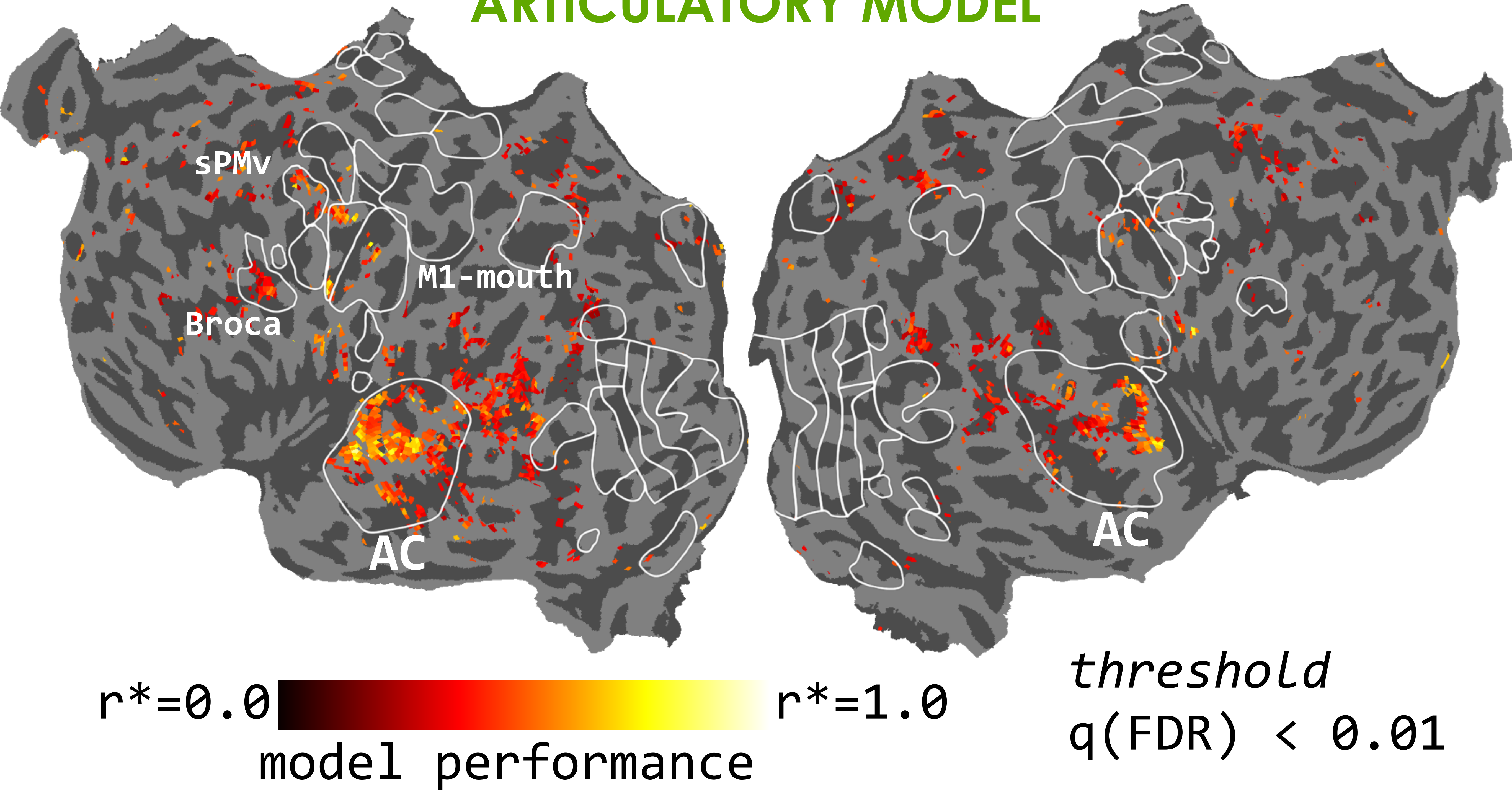
Sound Waveform
Feature Spaces
Downsampled Feature Representations
Feature Matrices
s(t)
Spectral
Articulation
Semantic
visual
tactile
numeric
abstract
violent
locational
communal
emotional
temporal
professional
social
mental
Spectral Feature
(power at 2081 Hz)
Original feature timecourse
Resampled timecourse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

SPEECH MODELS



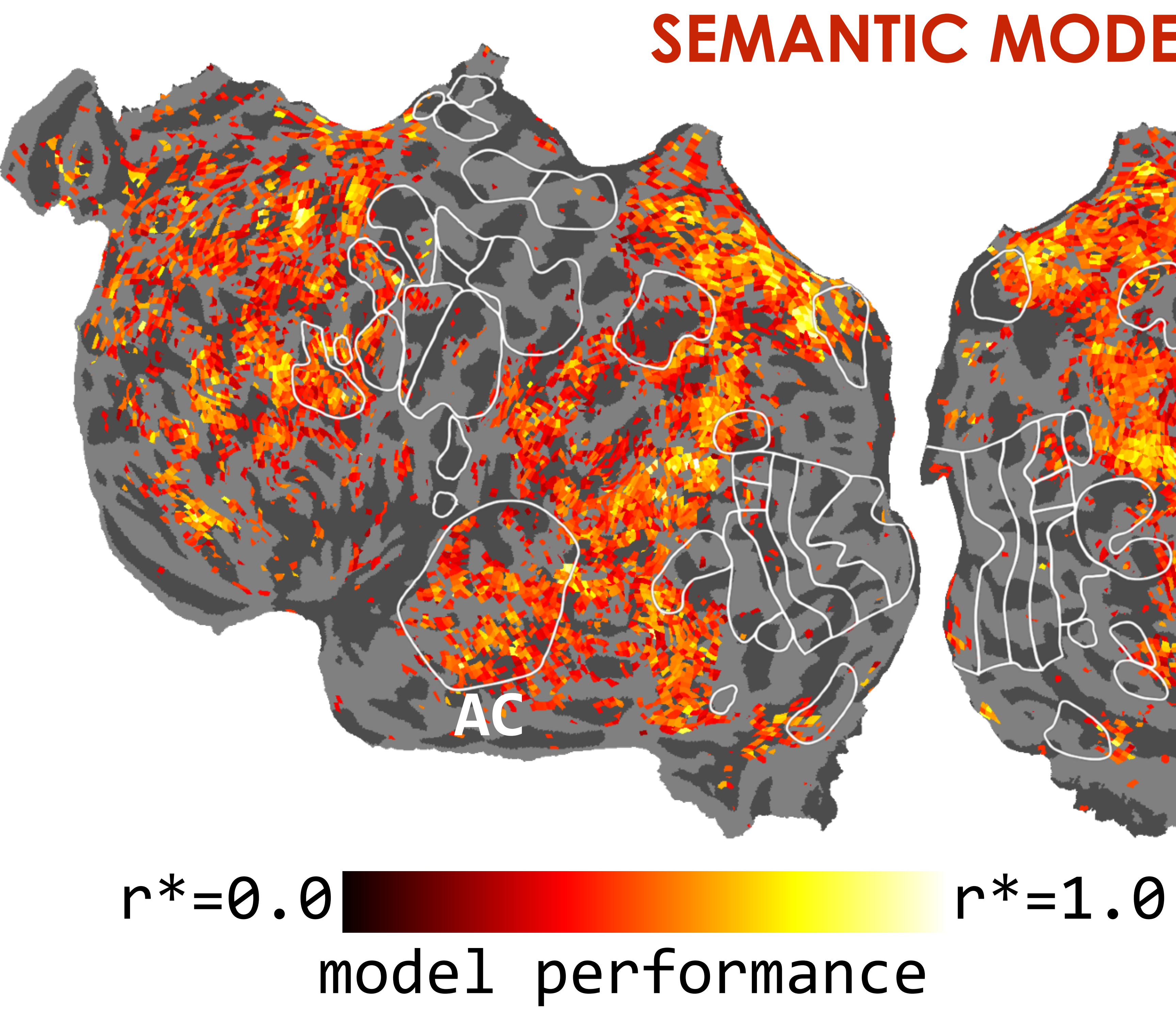
SPEECH MODELS

ARTICULATORY MODEL



threshold
q(FDR) < 0.01

SPEECH MODELS



threshold
q(FDR) < 0.01

MODEL COMPARISON

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

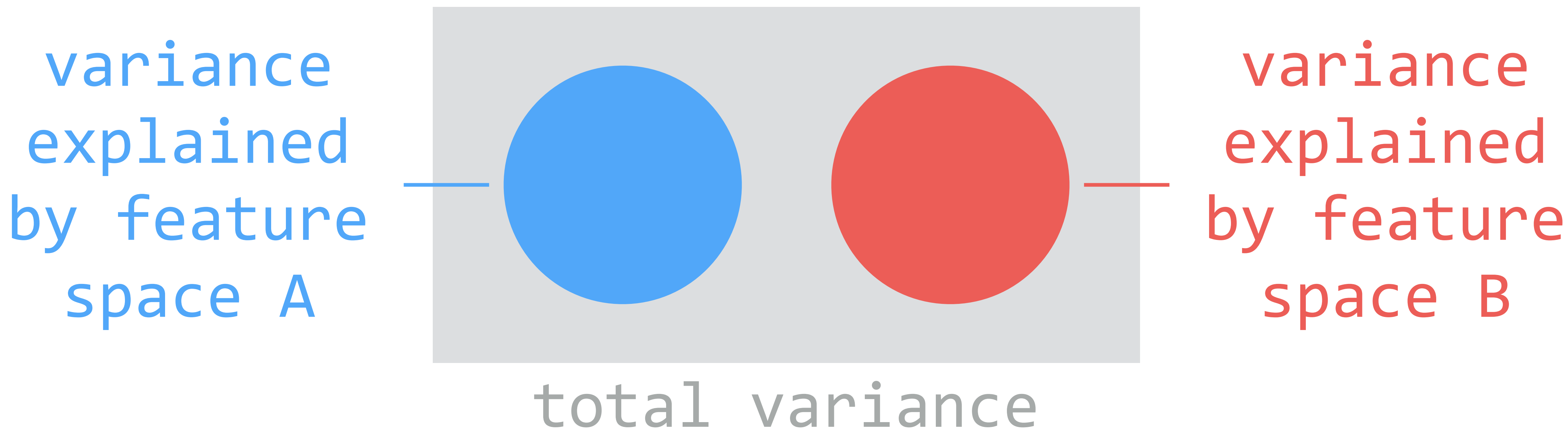
  * **Overall?** *Easy:* normal stats, parametric or non-parametric

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

VARIANCE PARTITIONING

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

  * Possibility 1: 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

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

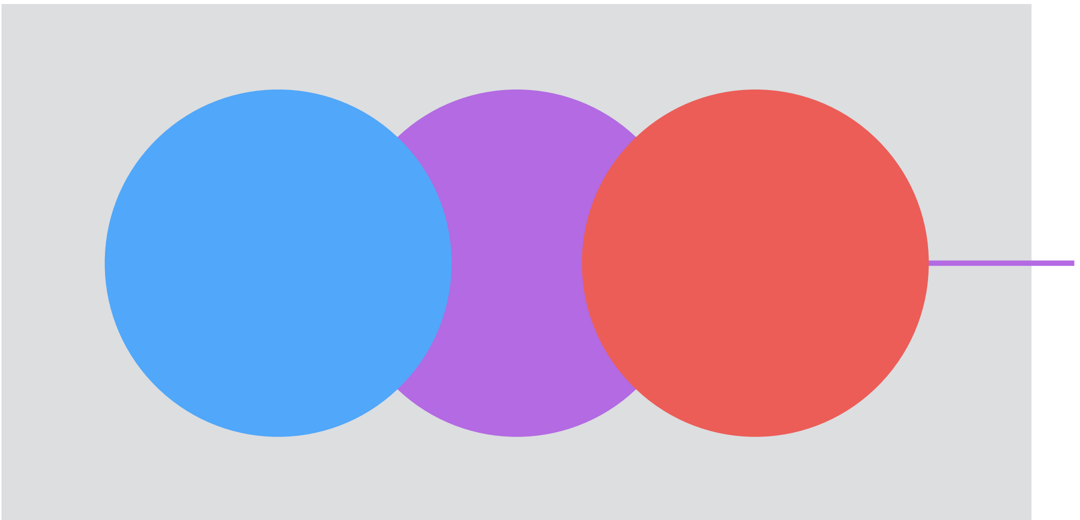
variance explained by feature space A (e.g. spectral)

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?
  * Possibility 2: both feature spaces explain **the same** variance

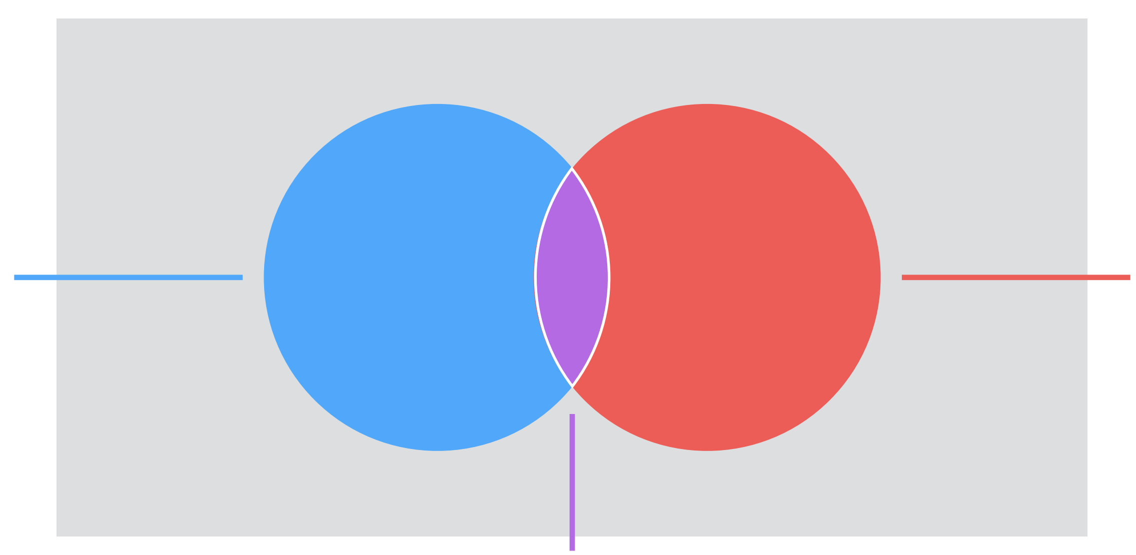
$$y = X_A\beta_A + \epsilon = X_B\beta_B + \epsilon$$

$$X_A\beta_A = X_B\beta_B$$

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**



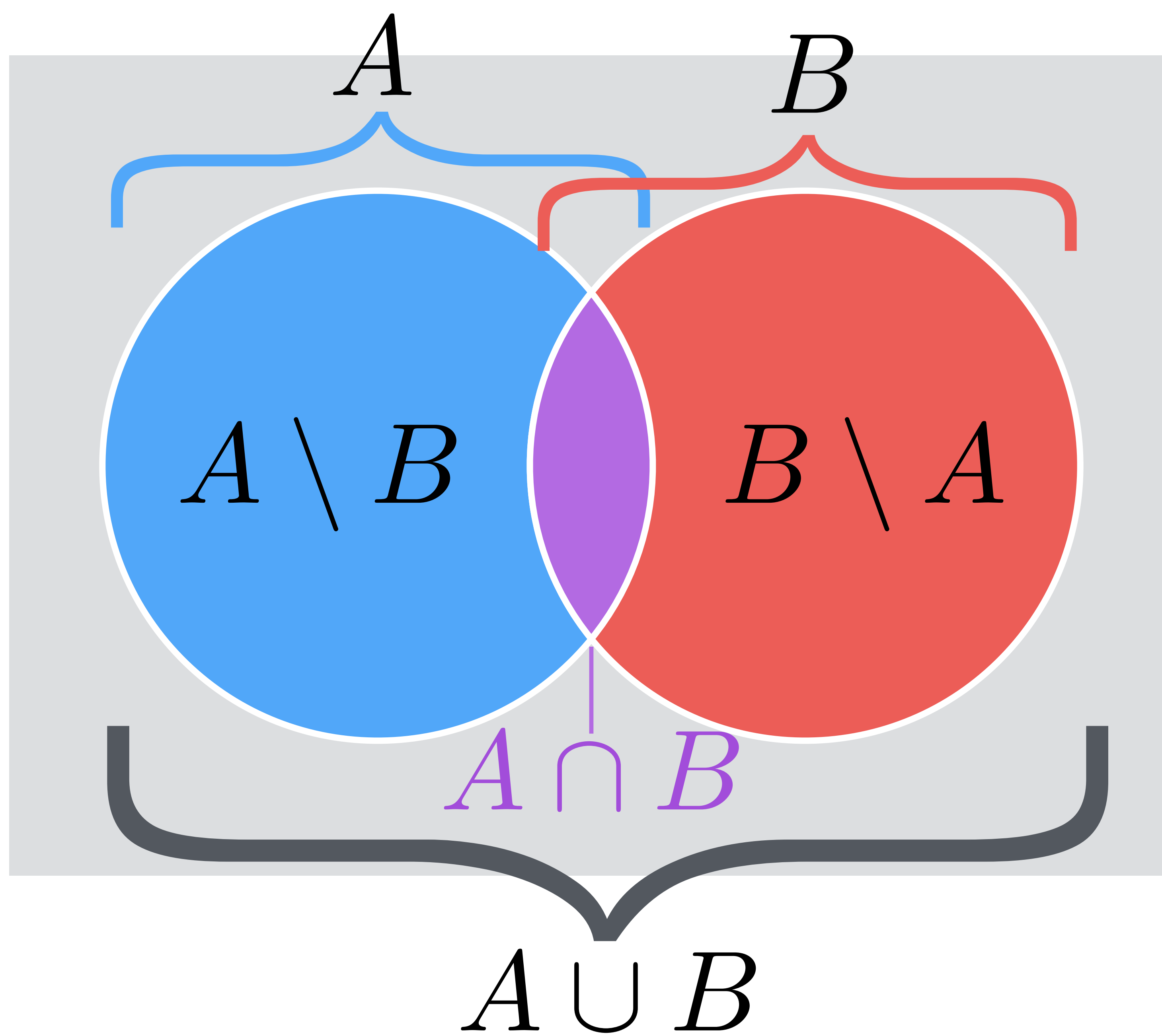
VARIANCE PARTITIONING

* 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**

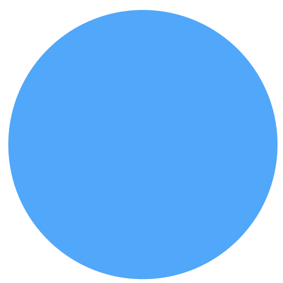
$$y = ?$$

VARIANCE PARTITIONING



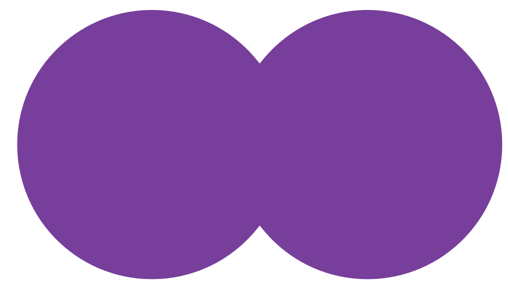
VARIANCE PARTITIONING

A



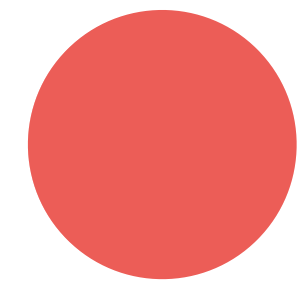
variance explained by feature space A

$A \cup B$



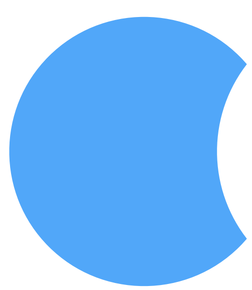
all variance explained by either A or B

B



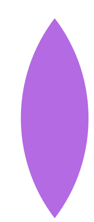
variance explained by feature space B

$A \setminus B$



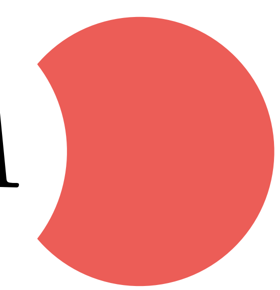
variance explained by feature space A that isn’t explained by B

$A \cap B$



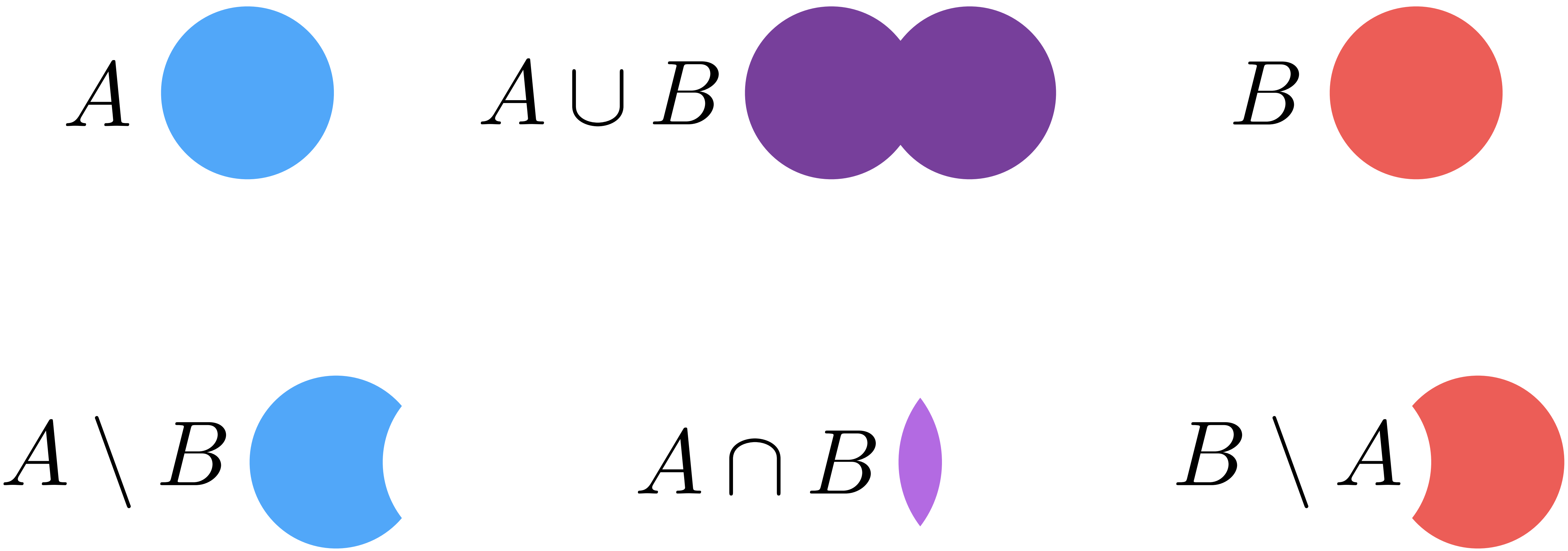
variance explained by both B and A

$B \setminus A$



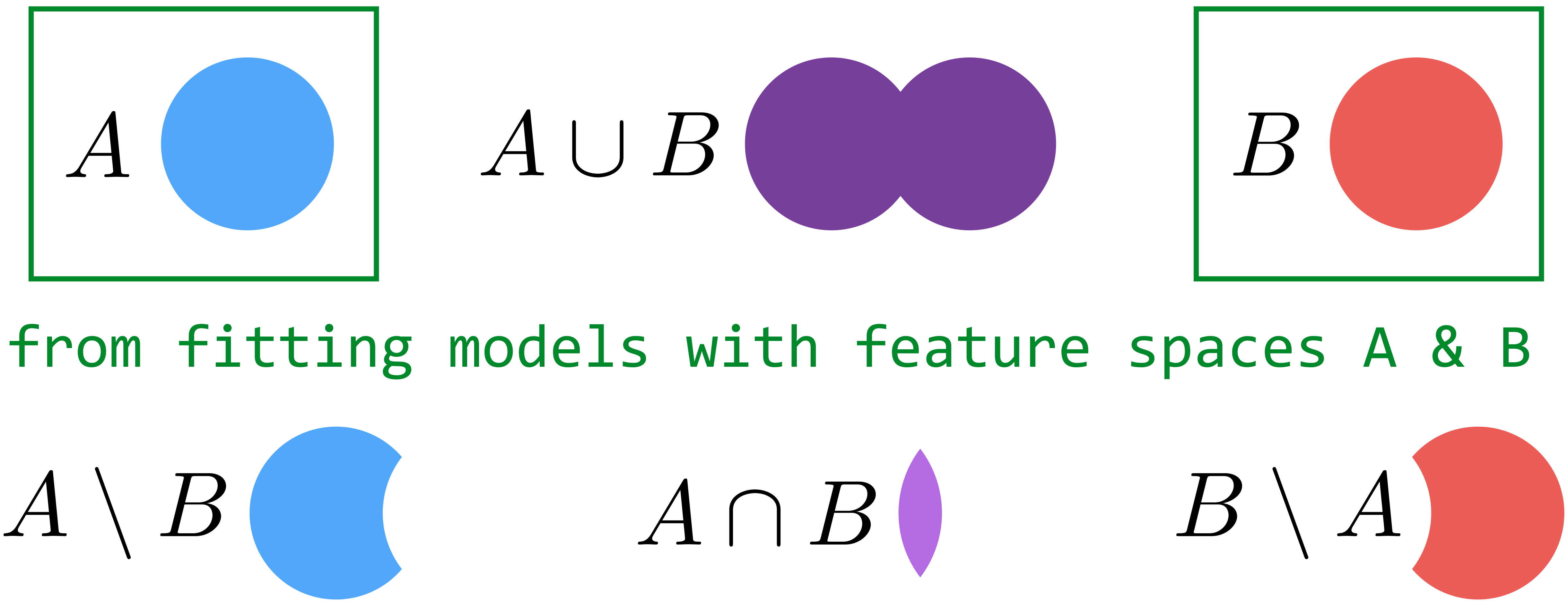
variance explained by feature space B that isn’t explained by A

VARIANCE PARTITIONING

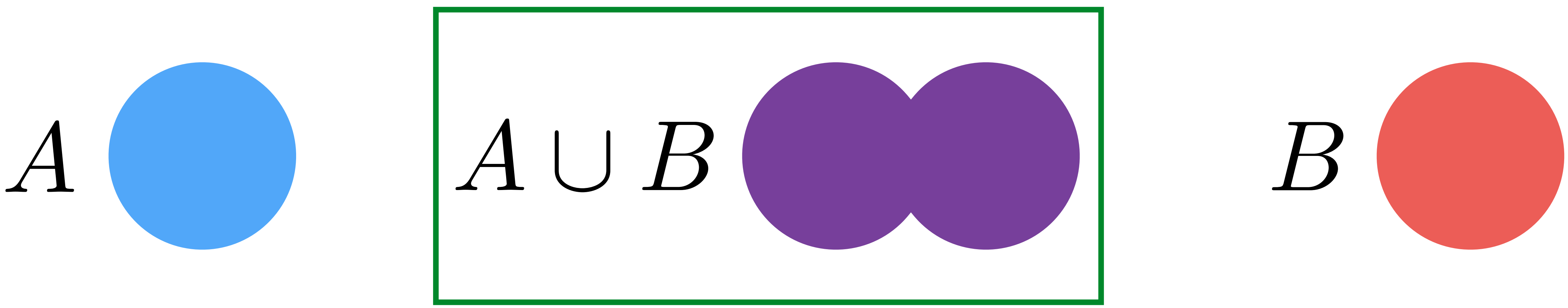


How do we measure the sizes of these partitions?

VARIANCE PARTITIONING



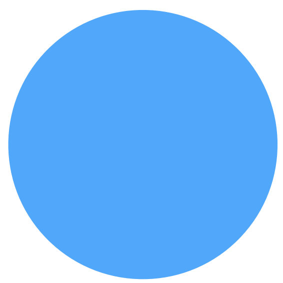
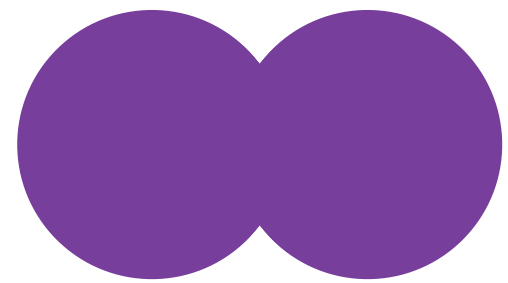
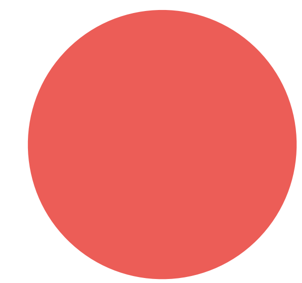
VARIANCE PARTITIONING

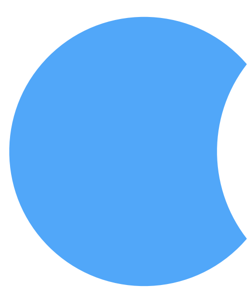
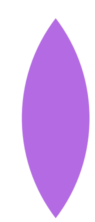
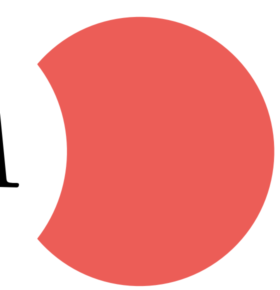


to get this we fit a model
with both feature spaces

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

VARIANCE PARTITIONING

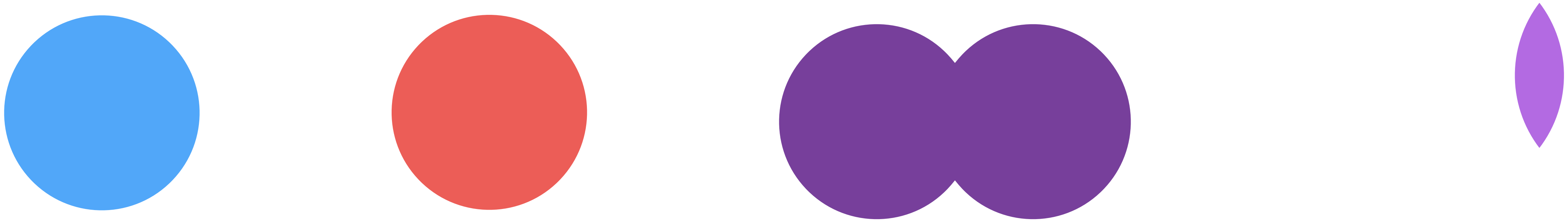
A  $A \cup B$  B 

$A \setminus B$  $A \cap B$  $B \setminus A$ 

to get these we have to use set theory

VARIANCE PARTITIONING

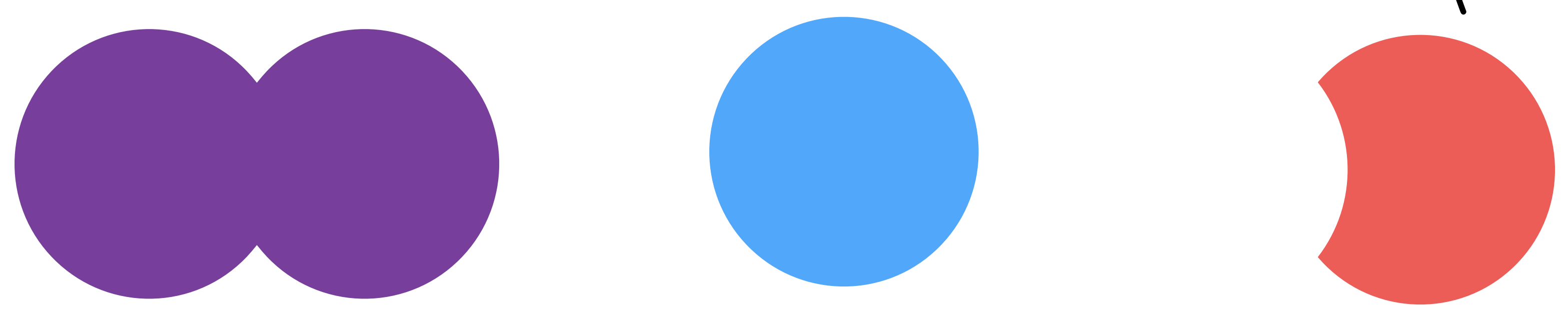
$$A \quad + \quad B \quad - \quad A \cup B \quad = \quad A \cap B$$



how much variance is explained by both A & B

VARIANCE PARTITIONING

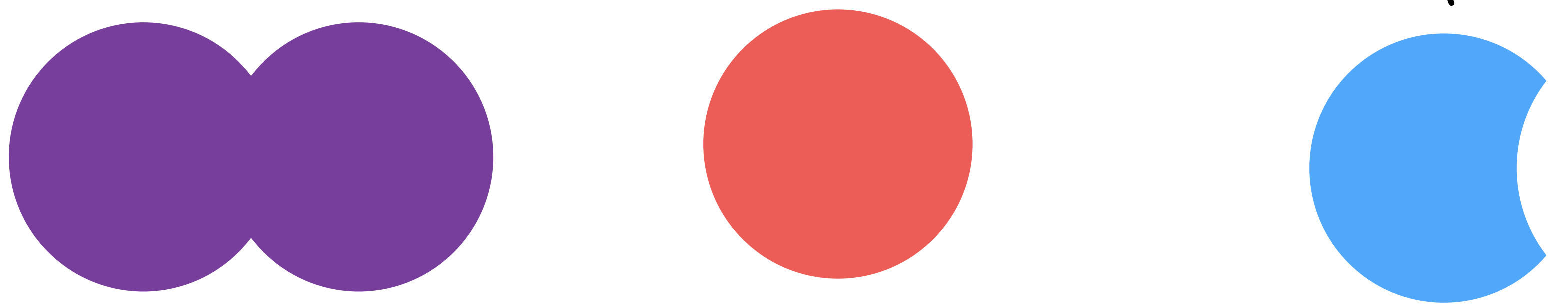
$$A \cup B \; - \; A \; = \; B \setminus A$$



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

VARIANCE PARTITIONING

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



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

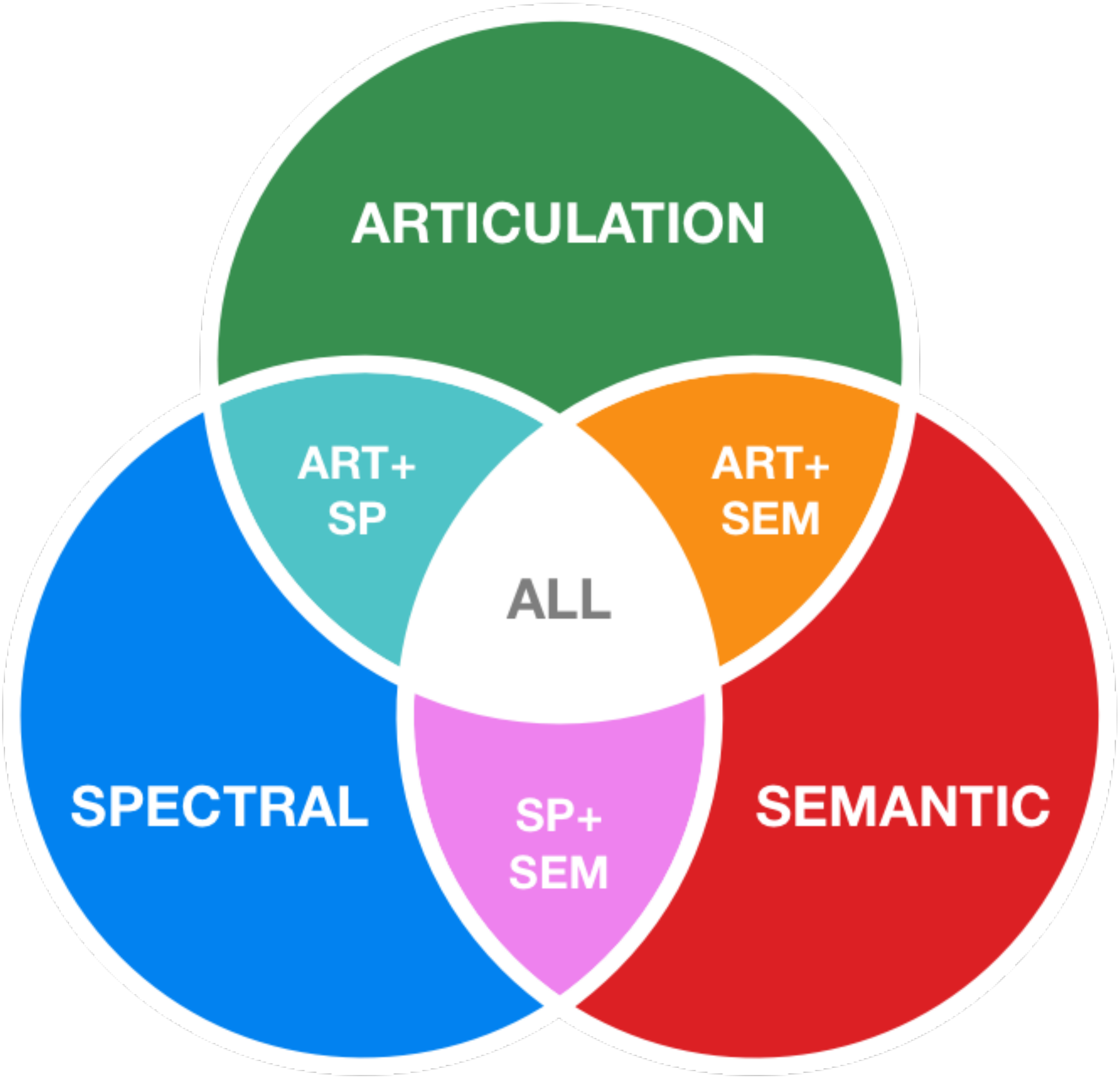
  * 2 feature spaces = 3 models

  * 3 feature spaces = 7 models

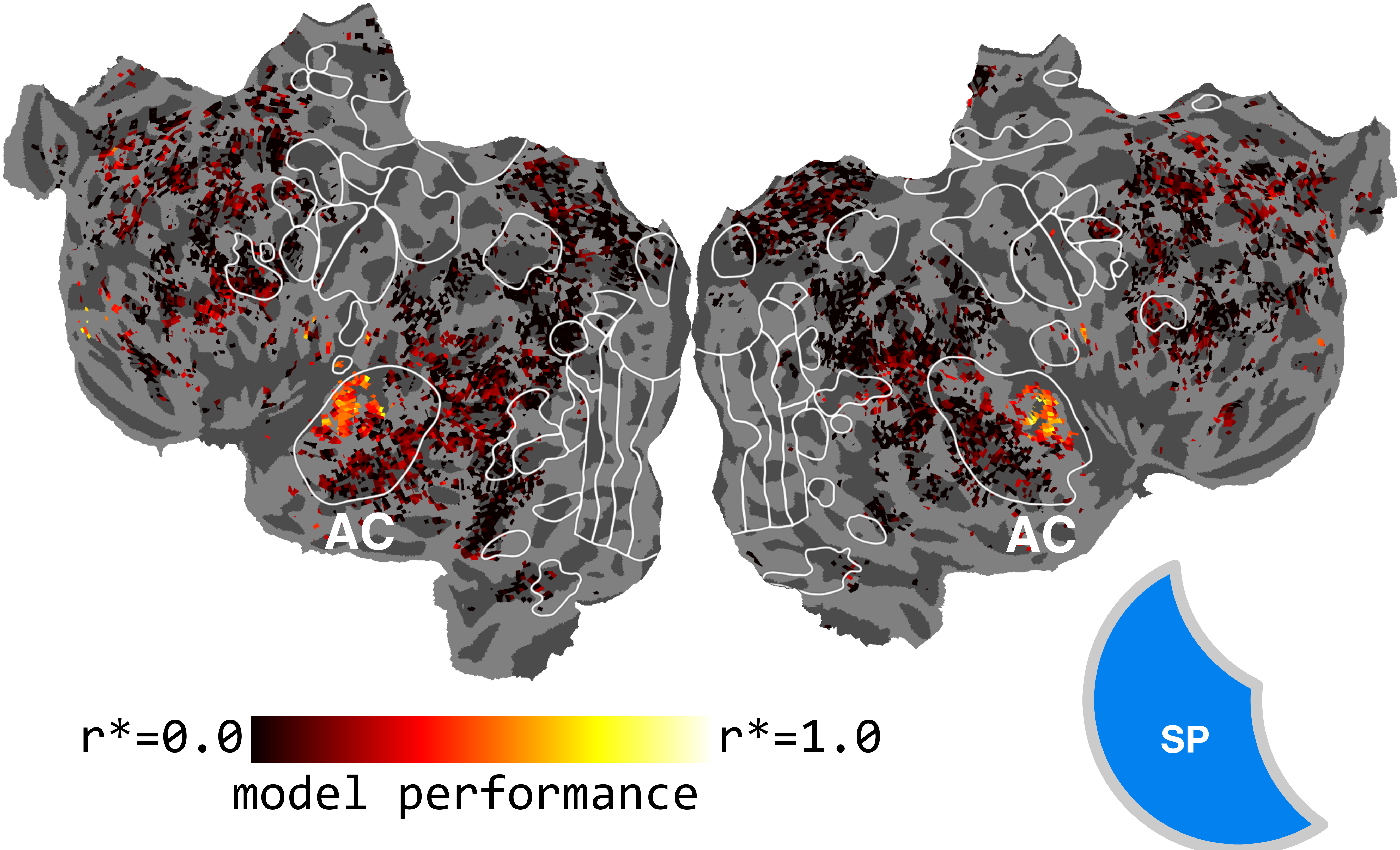
  * 4 feature spaces = 15 models

  * n feature spaces = $2^n - 1$ models

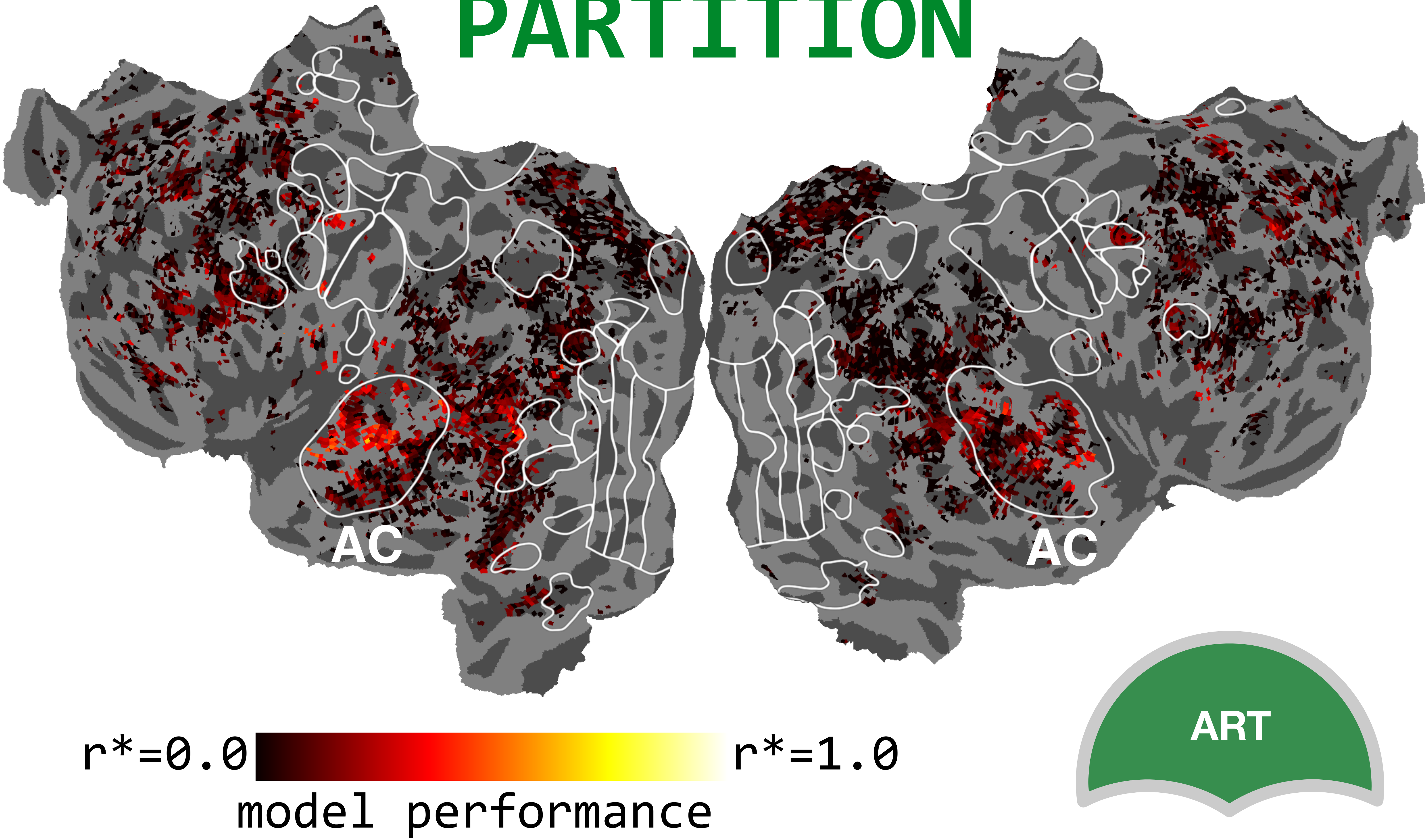
SPEECH MODELS

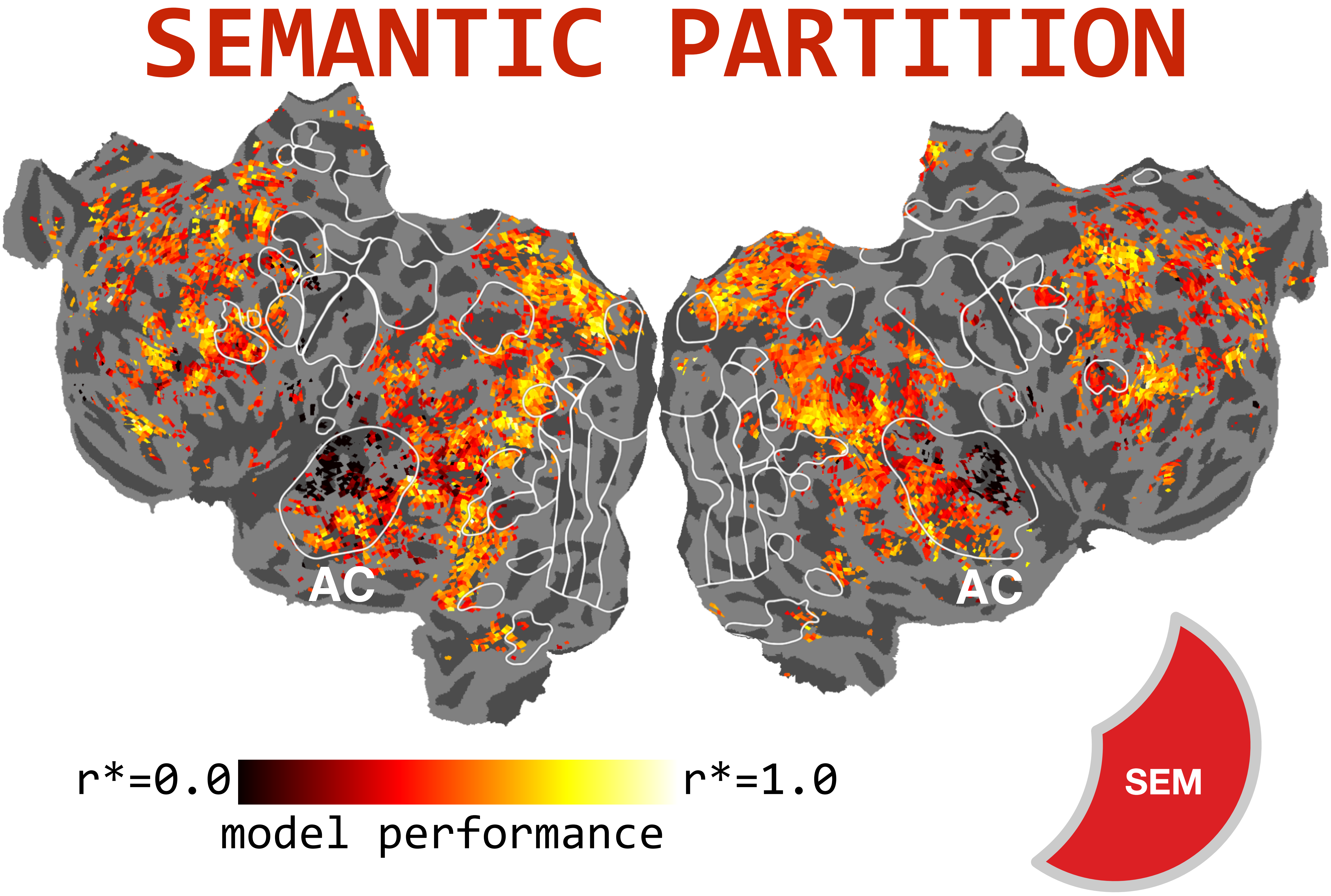


SPECTRAL PARTITION

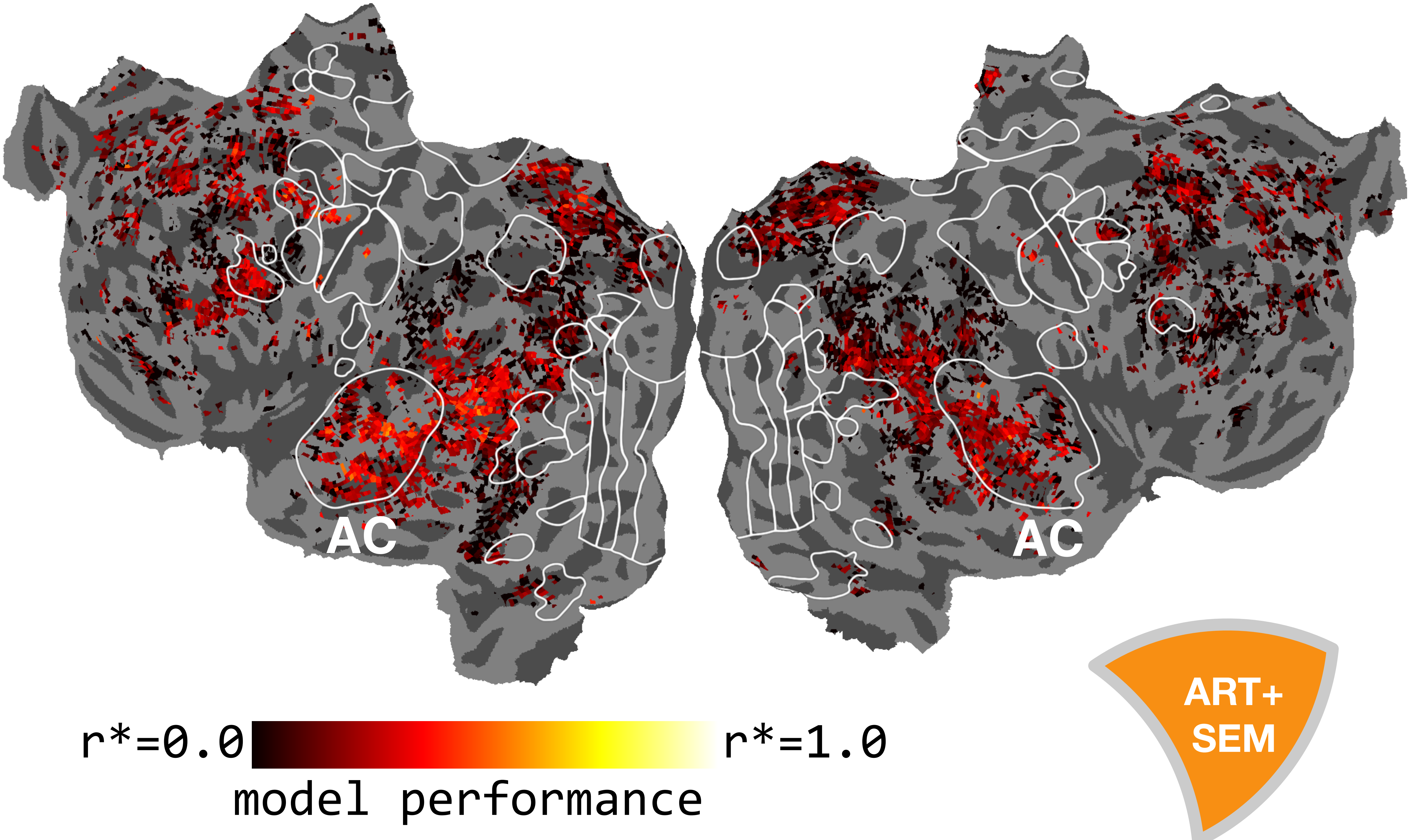


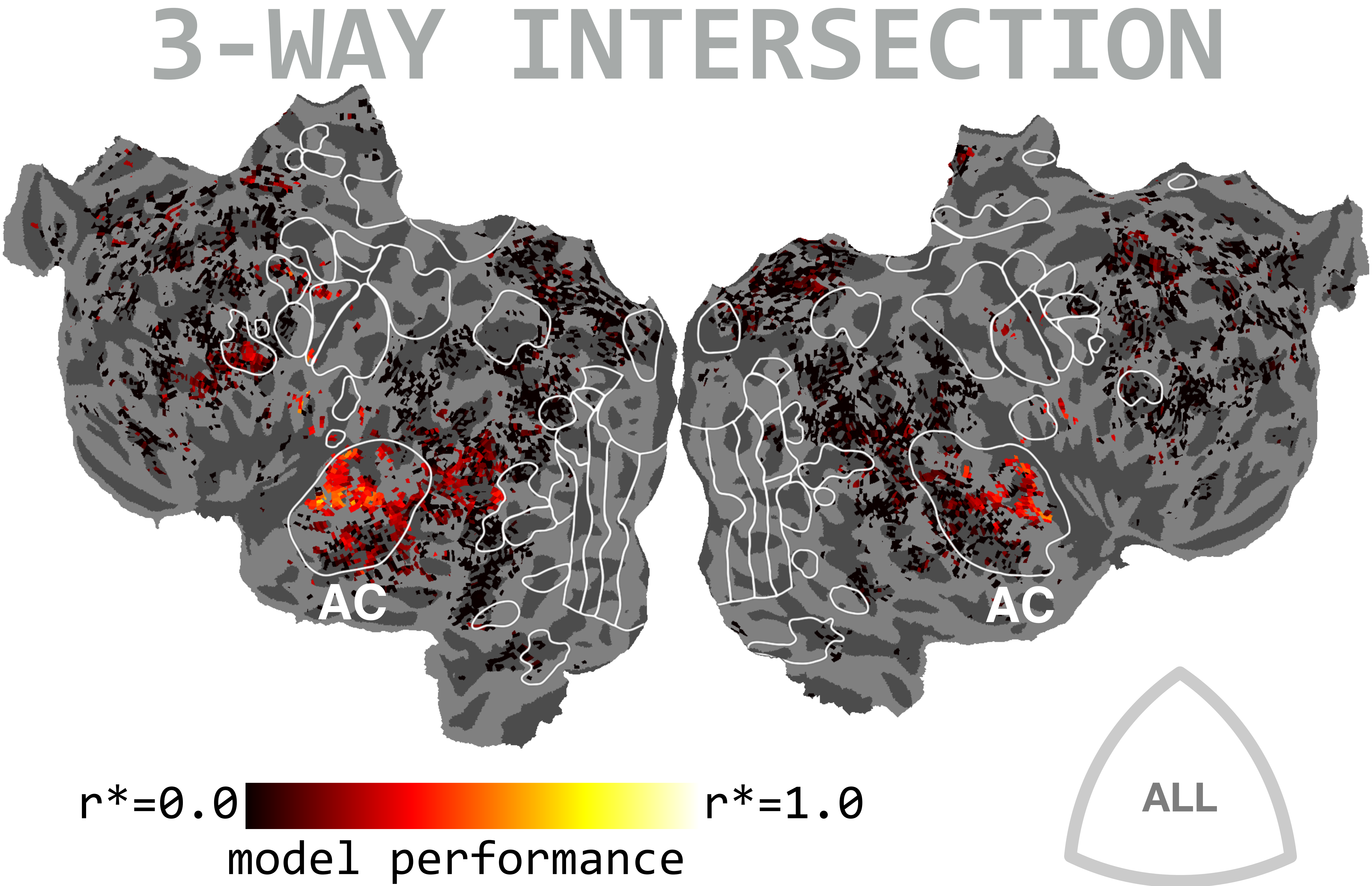
ARTICULATION PARTITION



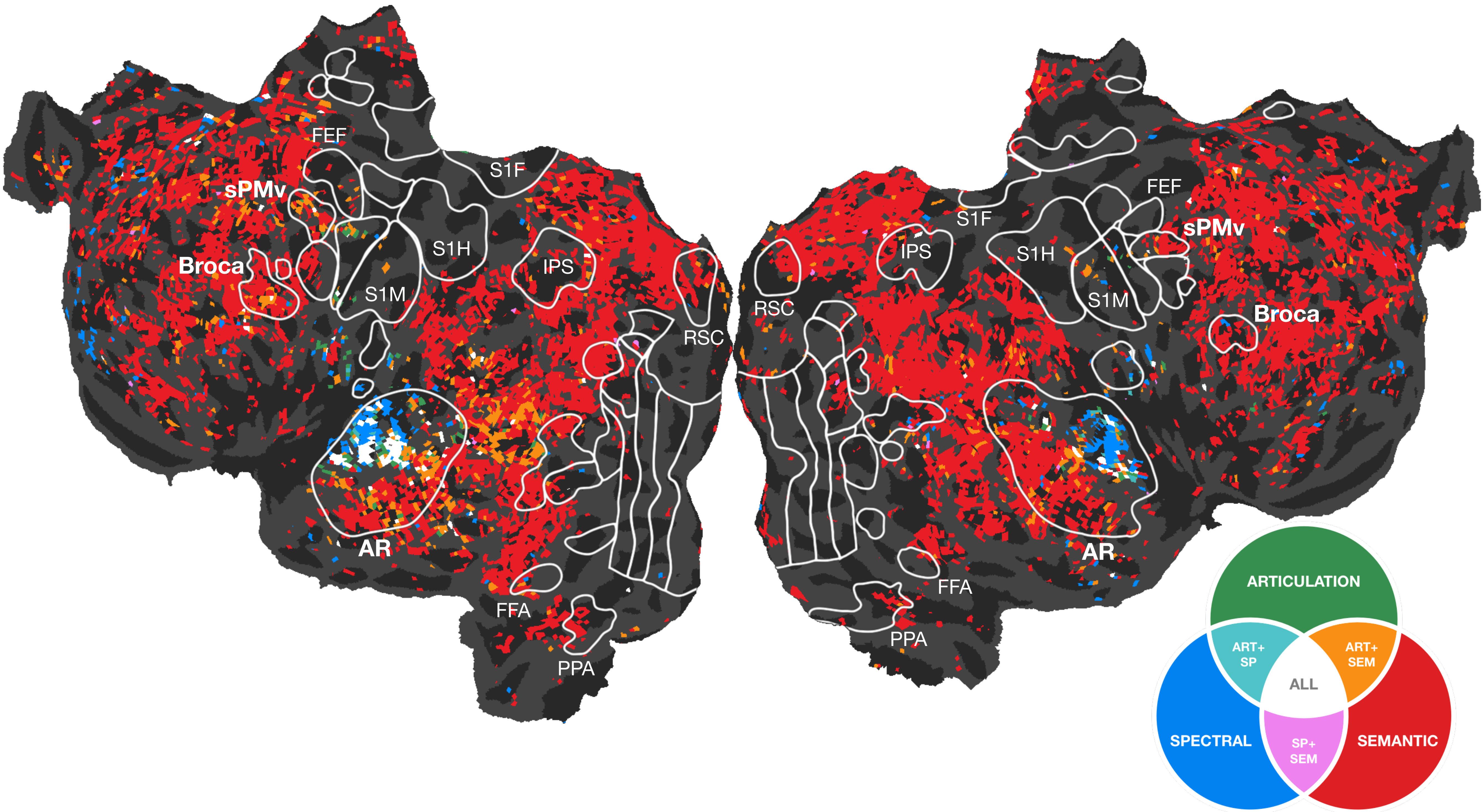

SEMANTIC PARTITION
AC
AC
r*=0.0
r*=1.0
model performance
SEM

ART&SEM PARTITION




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

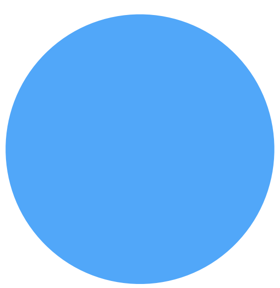
LARGEST PARTITION PER VOXEL



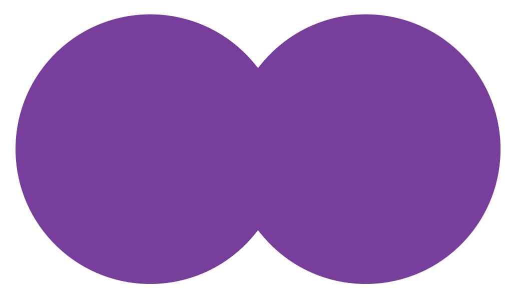
VARIANCE PARTITIONING

* Which of these models is most overfit?

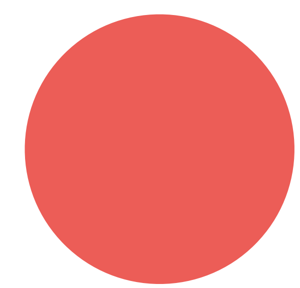
A



$A \cup B$

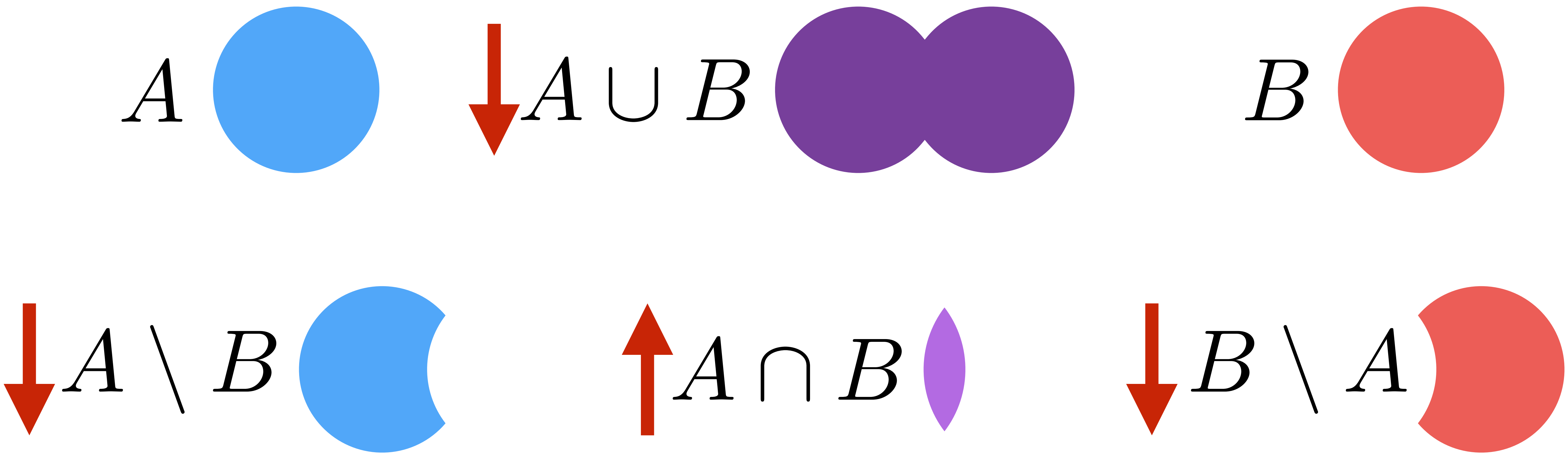


B



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

NEXT TIME

* Nonlinear models