WEEK 10: MACHINE LEARNING

Boris Bernhardt, PhD

Bratislav Misic, PhD

WHAT IS STATISTICAL LEARNING?

LINEAR REGRESION & LOGISTIC REGRESSION

SOME ALGROITHMS (LDA, KNN, TREES, SVMS)

RESAMPLING METHODS (CROSS-VALIDATION, BOOTSTRAP)

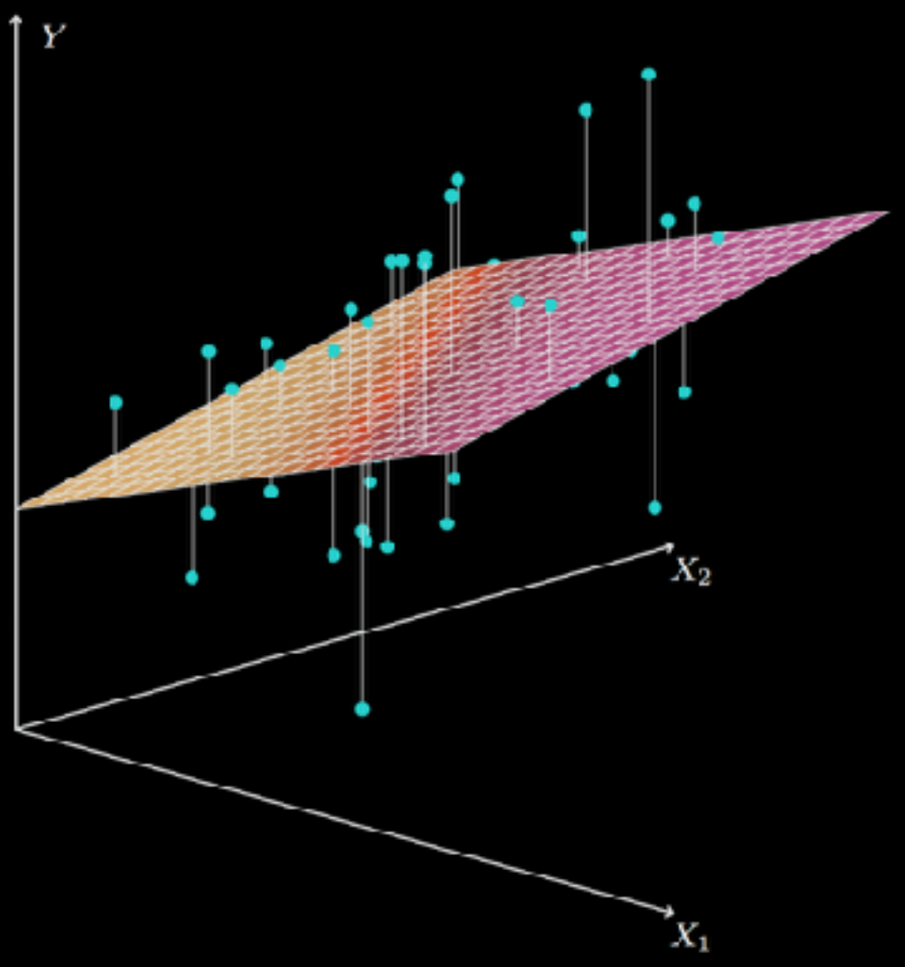
MODEL SELECTION AND REGULARIZATION

SUPERVISED LEARNING

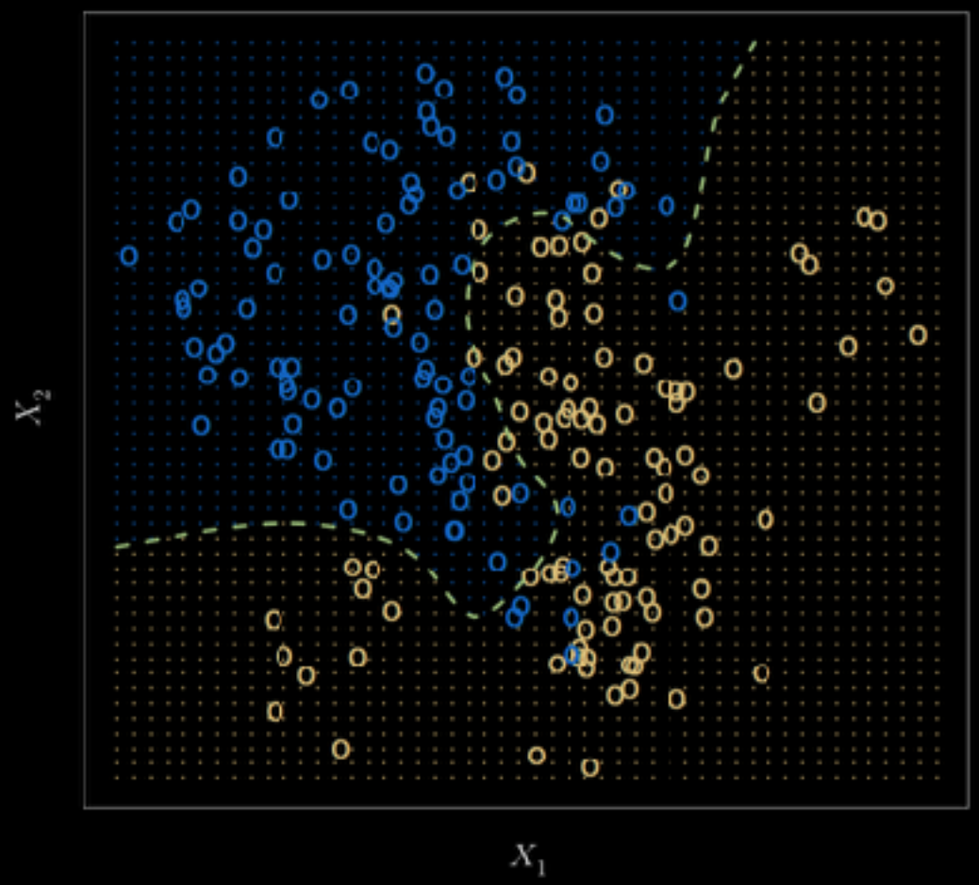
IDEA: LEARN A FUNCTION F
THAT MAPS DATA TO LABELS, BASED ON KNOWN DATA–LABEL PAIRS

$$Y = f(X) + \epsilon.$$

regression



classification

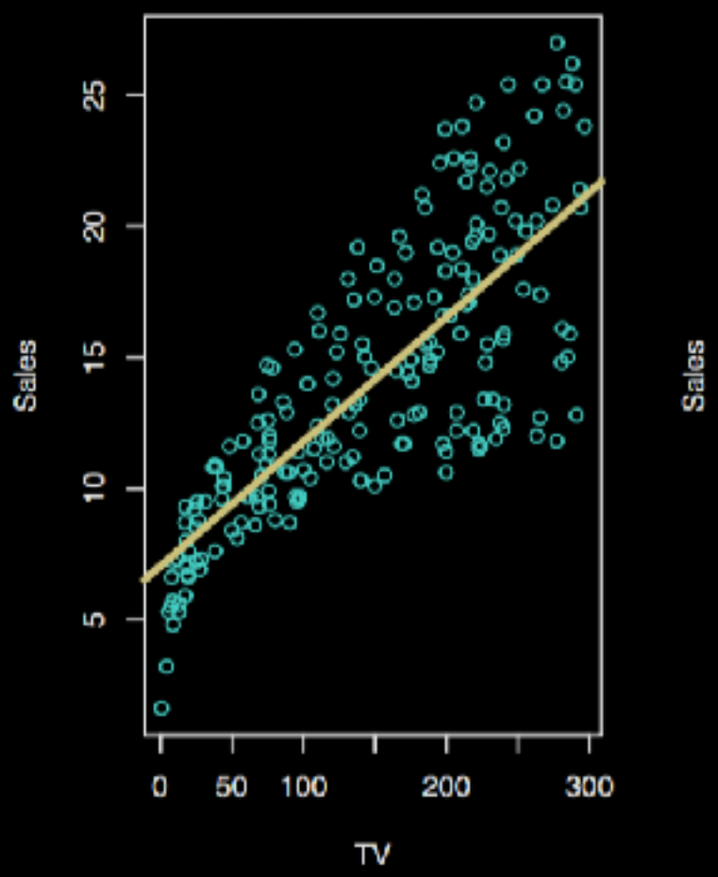


SUPERVISED LEARNING

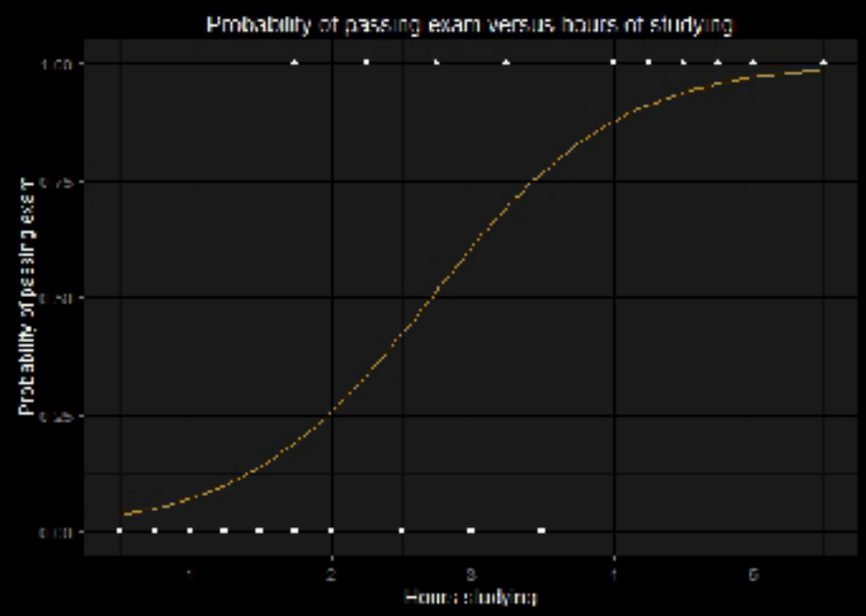
WELL-KNOWN CASE: GLMs

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

multiple linear regression



logistic regression

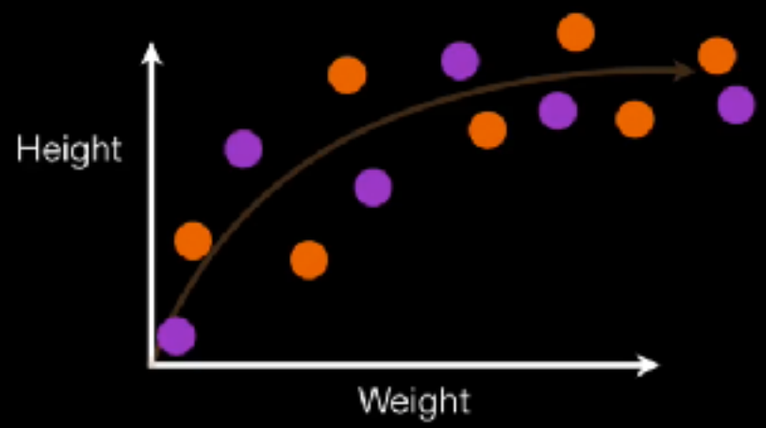


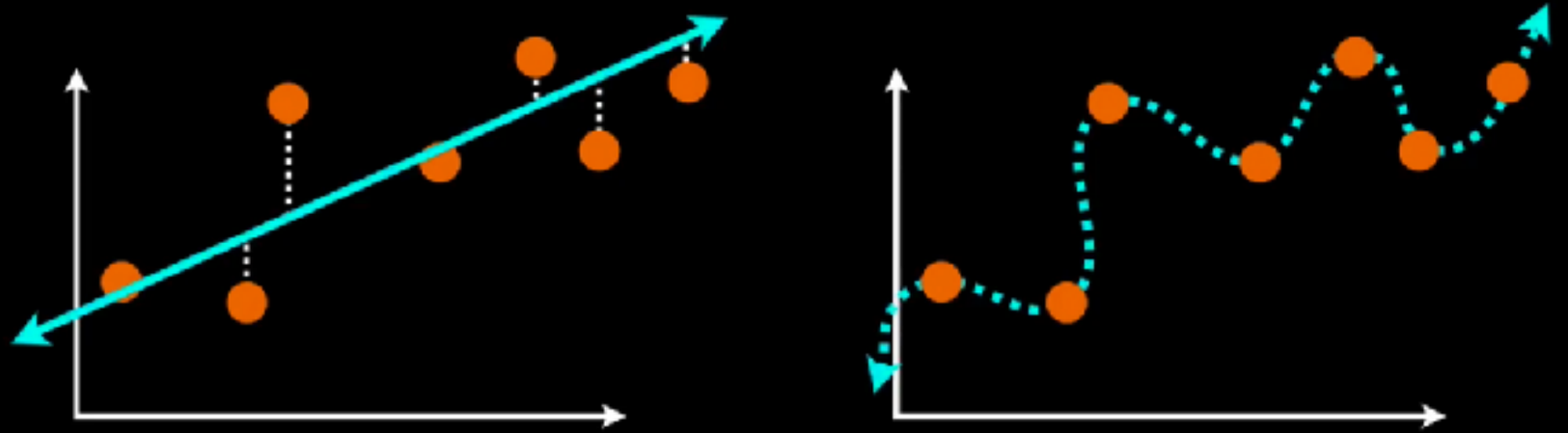
It's tough to make predictions, especially about the future

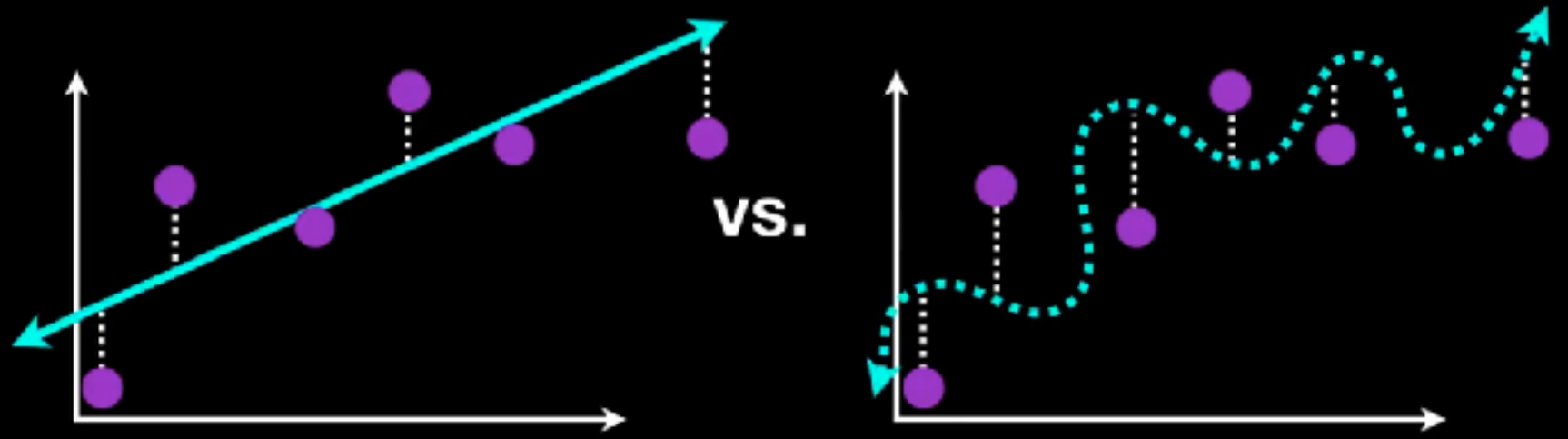
Yogi Berra

http://www-bcf.usc.edu/~gareth/ISL/

BIAS/VARIANCE







STAT QUEST ON YOUTUBE

NOTABLE LEARNERS

NAIVE BAYES

GIVEN TEST DATASET WITH DATA $X=x_1, \ldots x_n$,
ASSIGN CLASS LABEL C_k SO THAT

$$p(C_k \mid x_1, \ldots, x_n)$$

IS MAXIMAL

1) APPLY BAYES RULE

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

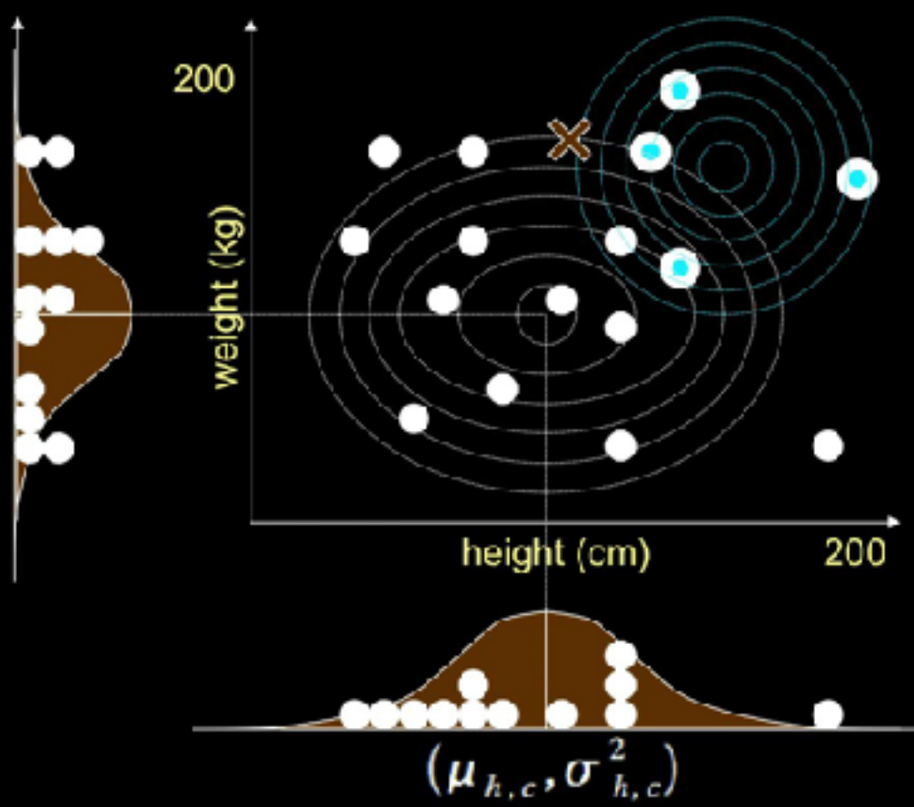
2) BE NAIVE

$$\begin{aligned} p(C_k \mid x_1, \ldots, x_n) &\propto p(C_k, x_1, \ldots, x_n) \\ &\propto p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k)\, \cdots \\ &\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,. \end{aligned}$$

NAIVE BAYES

3) DEFINE LIKELIHOOD AND PRIOR PROBABILITIES

$$P(a)=\frac{4}{4+12}=0.25\ ;\ P(c)=0.75 \qquad p(h_x|c)=\frac{1}{\sqrt{2\pi\,\sigma^2_{h,c}}}\exp-\frac{1}{2}\left(\frac{(h_x-\mu_{h,c})^2}{\sigma^2_{h,c}}\right)$$



$$p(w_x|c)=\frac{1}{\sqrt{2\pi\,\sigma^2_{w,c}}}\exp-\frac{1}{2}\left(\frac{(w_x-\mu_{w,c})^2}{\sigma^2_{w,c}}\right)$$

$$p(h_x|a)=\frac{1}{\sqrt{2\pi\,\sigma^2_{h,a}}}\exp-\frac{1}{2}\left(\frac{(h_x-\mu_{h,a})^2}{\sigma^2_{h,a}}\right)$$

$$p(w_x|a)=\frac{1}{\sqrt{2\pi\,\sigma^2_{w,a}}}\exp-\frac{1}{2}\left(\frac{(w_x-\mu_{w,a})^2}{\sigma^2_{w,a}}\right)$$

$$P(x|a)=p(h_x|a)\,p(w_x|a)$$

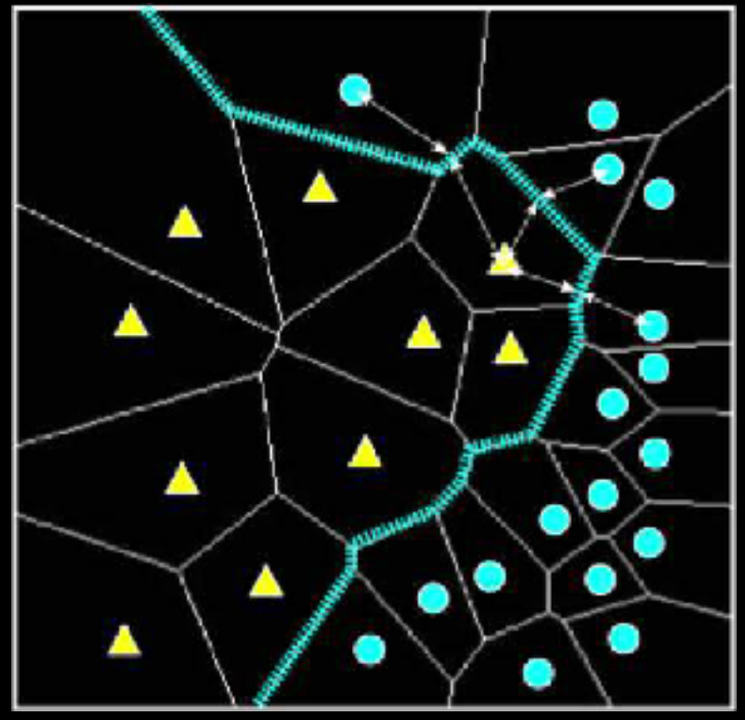
$$P(x|c)=p(h_x|c)\,p(w_x|c)$$

$$P(a|x)=\frac{P(x|a)P(a)}{P(x|a)P(a)+P(x|c)P(c)}$$

k-nearest neighbour

ASSIGN TEST CASE TO CLOSET TRAINING EXAMPLE

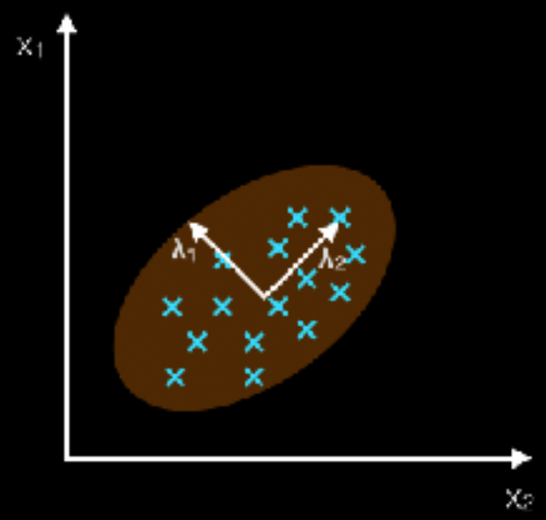
1) DEFINE DISTANCE



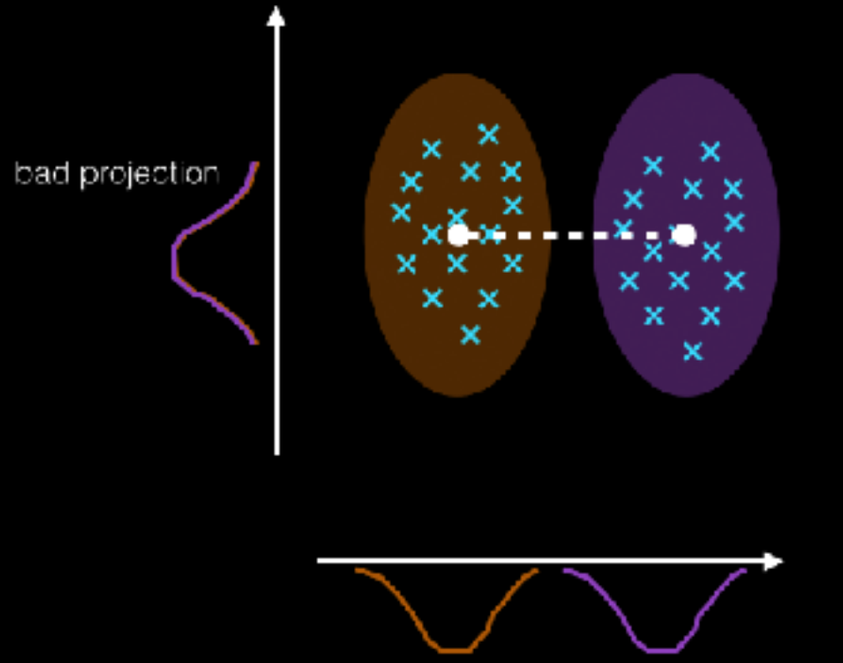
LDA

idea: rotation of a labeled dataset to identify vectors of maximum class separation

PCA:
component axes that maximize the variance



LDA:
maximizing the component axes for class-separation



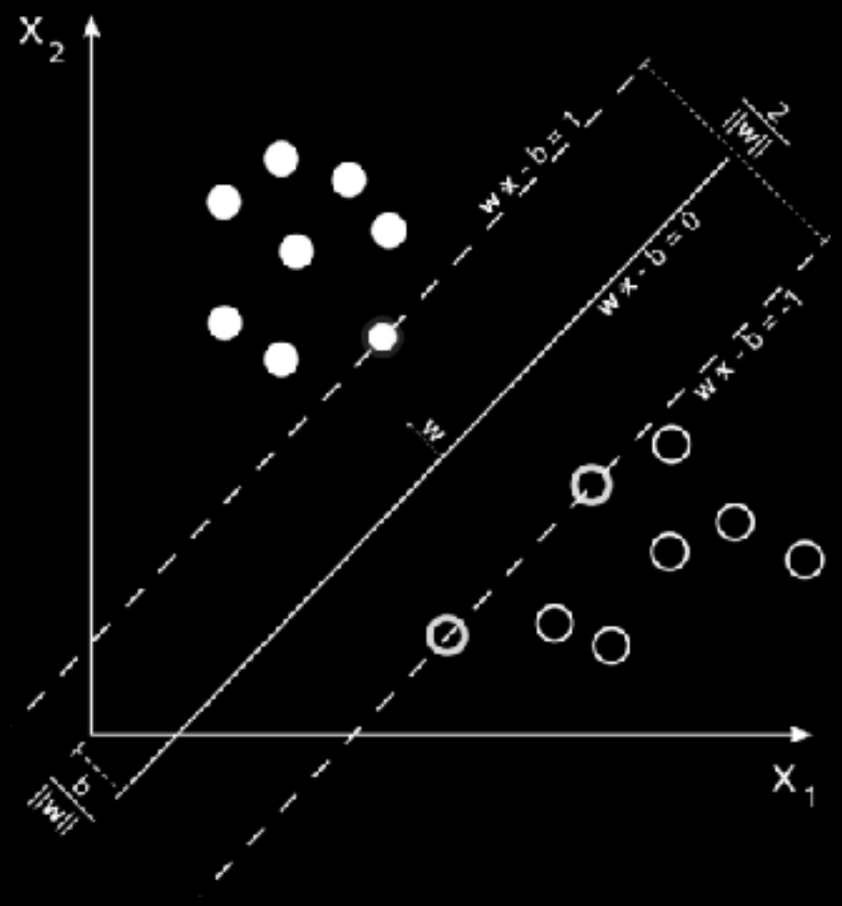
LDA

Algorithm steps

1. Compute the d-dimensional mean vectors for the different classes from the dataset.
2. Compute the scatter matrices (in-between-class and within-class scatter matrix).
3. Compute the eigenvectors ($\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$) and corresponding eigenvalues ($\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_d$) for the scatter matrices.
4. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a $d \times k$ dimensional matrix $\boldsymbol{W}$ (where every column represents an eigenvector).
5. Use this $d \times k$ eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: $\boldsymbol{Y} = \boldsymbol{X} \times \boldsymbol{W}$ (where $\boldsymbol{X}$ is a $n \times d$-dimensional matrix representing the n samples, and $\boldsymbol{y}$ are the transformed $n \times k$-dimensional samples in the new subspace).

SUPPORT VECTOR MACHINES

idea: find hyperplane that maximizes margin between classes, not necessarily between all elements in class



grounded in statistical learning theory — very good generalization abilities

Maps data into a high dimensional space where linear separability is possible

often combined with ‘kernel trick’ to avoid curse of dimensionality

OTHER LEARNERS

DECISION TREES

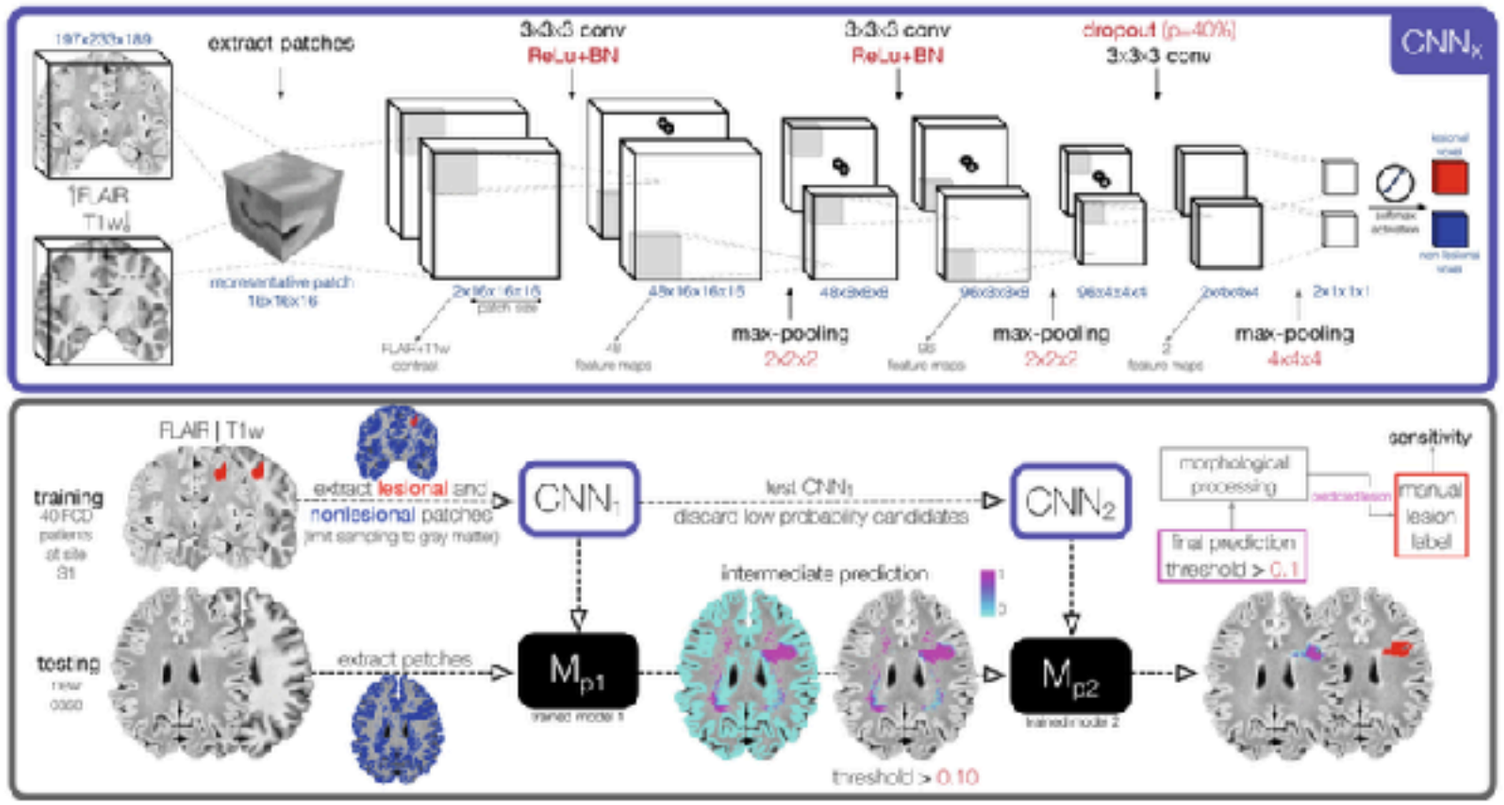
RANDOM FORESTS

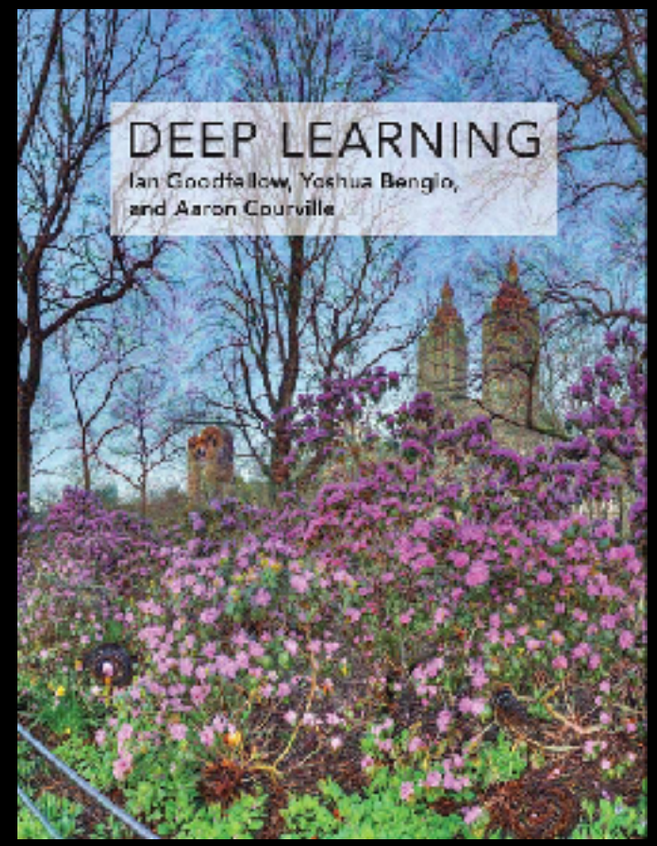
NEURAL NETWORKS

DEEP NN

"THERE IS NO FREE LUNCH" (WOLPERT & MACREADY, 1997)

DNN


extract patches
3x3x3 conv
ReLu+BN
3x3x3 conv
ReLu+BN
dropout (p=40%)
3x3x3 conv
CNN$_X$
FLAIR
T1w
max-pooling
2x2x2
max-pooling
2x2x2
max-pooling
4x4x4
FLAIR | T1w
training
extract lesional and
nonlesional patches
CNN$_1$
test CNN$_1$
discard low probability candidates
CNN$_2$
morphological processing
final prediction
threshold > 0.1
sensitivity
manual lesion label
intermediate prediction
testing
extract patches
M$_{p1}$
M$_{p2}$
threshold > 0.10


DEEP LEARNING
Ian Goodfellow, Yoshua Bengio,
and Aaron Courville

RESAMPLING METHODS

LEAVE ONE OUT
K-FOLD
SPLIT-HALF

IDEA:
TRAIN LEARNER ON A SUBSET OF THE DATA
AND TEST ON A PREVIOUSLY UNSEEN DATASET

CONSERVATIVE HOLD-OUT PROCEDURES MAY REDUCE ‘OPTIMISM’

MODEL SELECTION AND REGULARIZATION

FEATURE SELECTION ROUTINES CAN BE
EMBEDDING IN THE TRAINING STEP

DIMENSIONALITY REDUCTION TECHNIQUES MAY
IMPROVE GENERALIZATION PERFORMANCE
AND REDUCE OVERFITTING
"OCCAM'S RAZOR"

LASSO + RIDGE REGRESSION

IN ADDITION TO OPTIMIZING BASED ON TRAINING ERROR,
ONE CAN ALSO PENALIZE PREDICTORS WITH MANY FEATURES

LASSO TECHNIQUES SHRINK
ITERATIVELY THE WEIGHTS OF THE FEATURES

CHANG

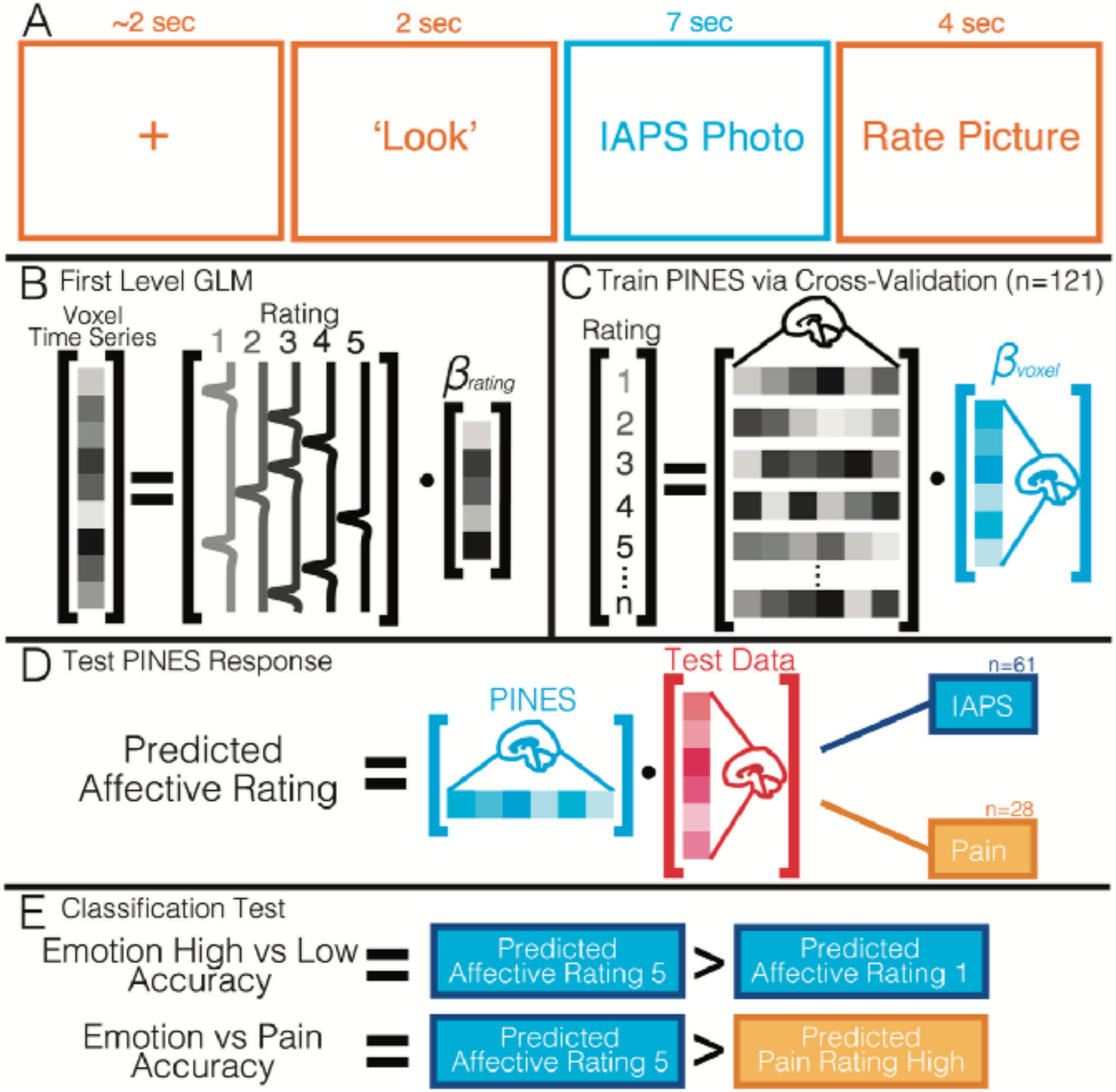
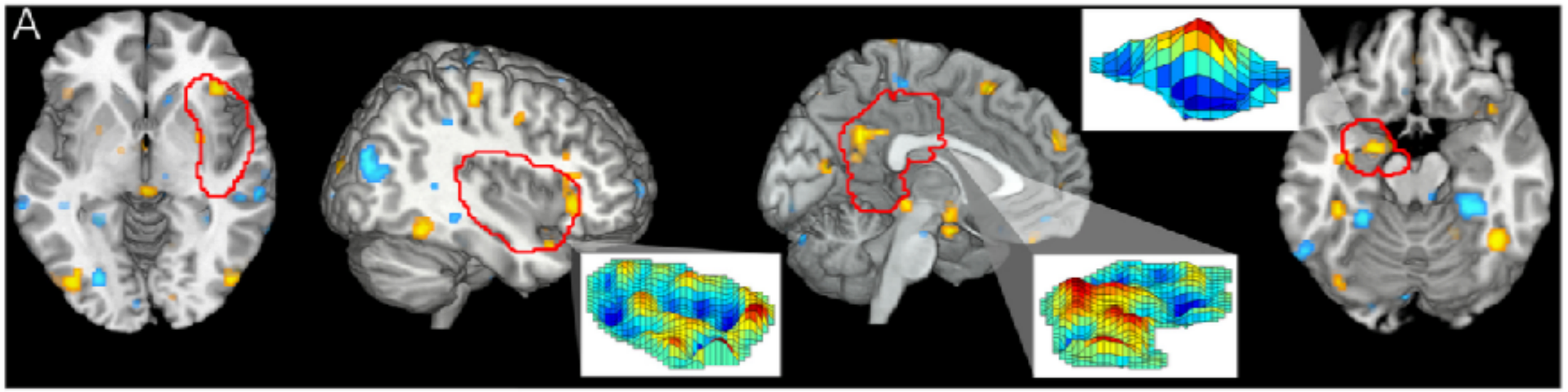


Fig 1. Experimental paradigm and analysis overview. Panel A depicts the sequence of events for a given trial. Participants view an initial fixation cross and then are instructed to look at the picture (compared to reappraise). Participants then see a photo and are asked to rate how negative they feel on a likert scale of 1–5. Panel B illustrates the temporal data reduction for each rating level using voxel-wise univariate analysis and an assumed hemodynamic response function. Panel C: these voxels are then treated as features and trained to predict ratings using LASSO-PCR with leave-one-subject-out cross validation. Subject's data for each rating is concatenated across participants. Panel D: this multivoxel weight map pattern can be tested on new data using matrix multiplication to produce a scalar affective rating prediction. Panel E: we calculated two different types of classification accuracy: (a) the ability to discriminate between high (rating = 5) and low (rating = 1) affective ratings and (b) the ability to discriminate between high affective and high pain data.



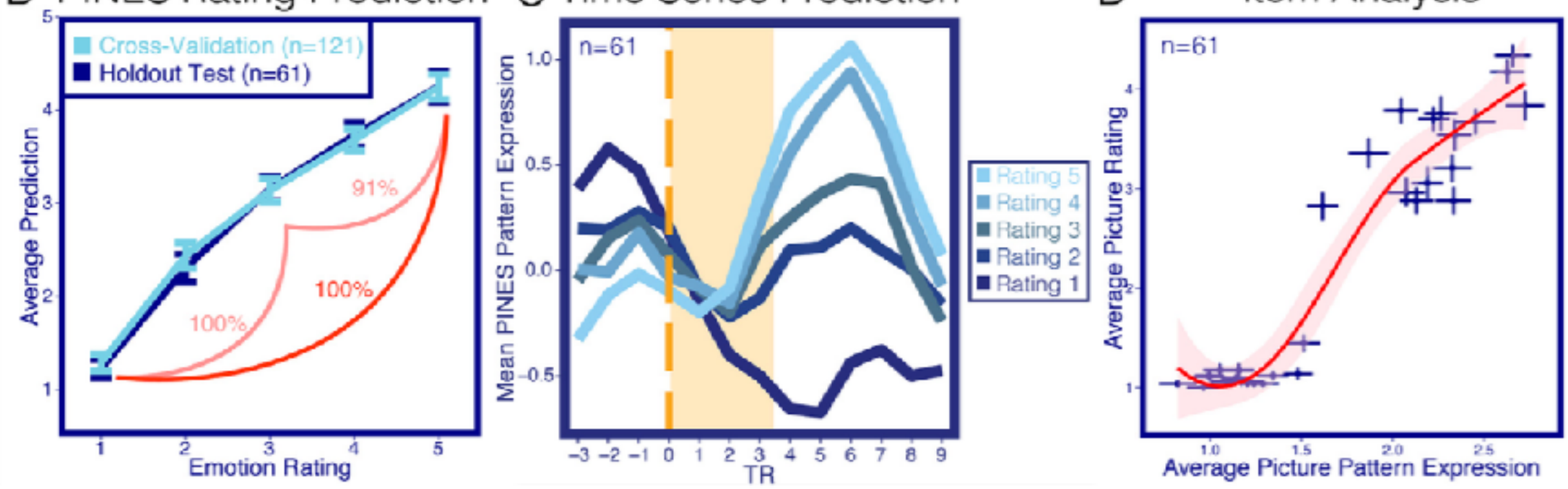
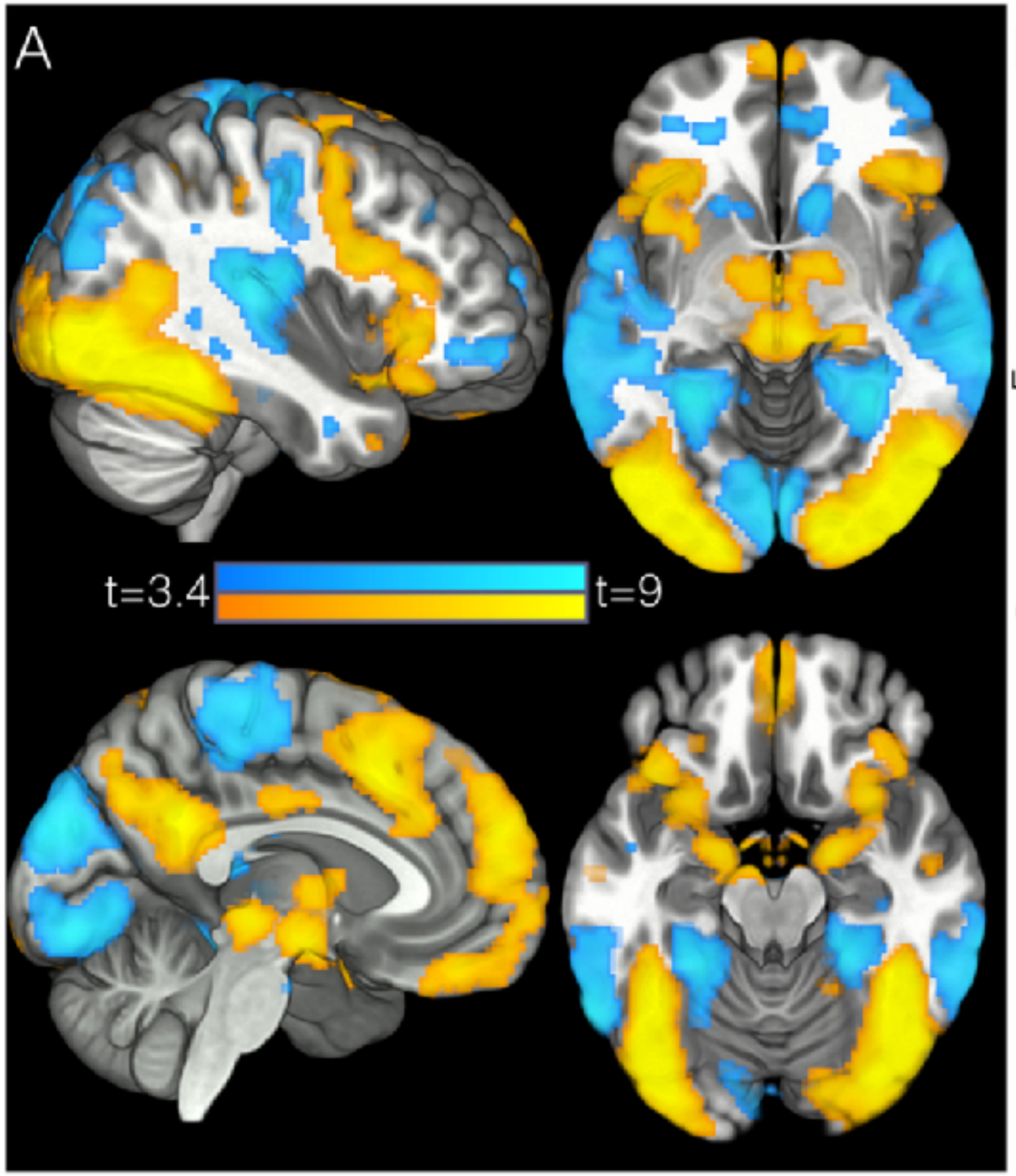
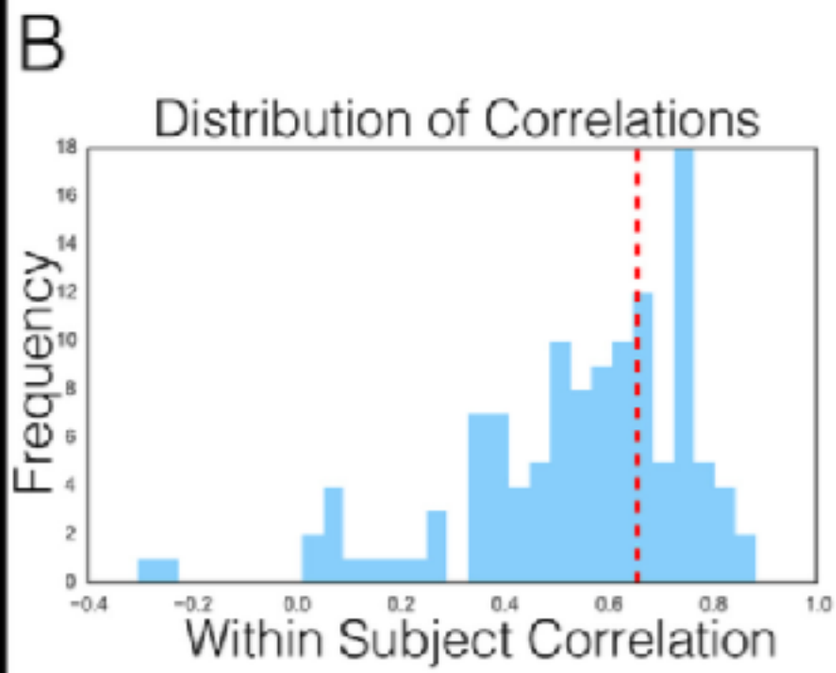


Fig 2. PINES. Panel A depicts the PINES pattern thresholded using a 5,000 sample bootstrap procedure at $p < 0.001$ uncorrected. Blowout sections show the spatial topography of the pattern in the left amygdala, right insula, and posterior cingulate cortex. Panel B shows the predicted affective rating compared to the actual ratings for the cross validated participants ($n = 121$) and the separate holdout test data set ($n = 61$). Accuracies reflect forced-choice comparisons between high and low and high, medium, and low ratings. Panel C depicts an average peristimulus plot of the PINES response to the holdout test dataset ($n = 61$). This reflects the average PINES response at every repetition time (TR) in the timeseries separated by the rating. Panel D illustrates an item analysis which shows the average PINES response to each photo by the average ratings to the photos in the separate test dataset ($n = 61$). Error bars reflect ±1 standard error.





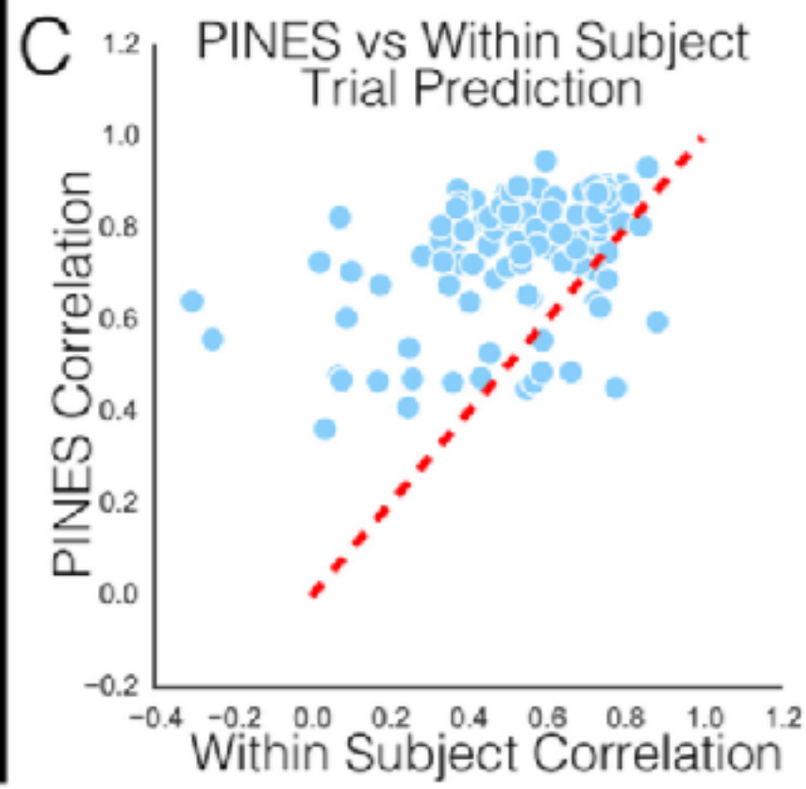


Fig 3. Within participant emotion prediction. This figure depicts results from our within-participant analysis, in which the PINES was retrained separately for each participant to predict ratings to individual photos. Panel A shows the voxels in the weight map that are consistently different from zero across participants using a one sample *t* test thresholded at $p < 0.001$ uncorrected. Panel B shows a histogram of standardized emotion predictions (correlation) for each participant. The dotted red line reflects the average cross validated PINES correlation for predicting each photo's rating. Panel C depicts how well each participant's ratings were predicted by the PINES (y-axis) versus an idiographically trained, cross-validated map using their individual brain data (x-axis). Each point on the graph reflects one participant. The dotted red line reflects the identity line. Any data point above the identity line indicates that the participant was better fit by the PINES than their own weight map.

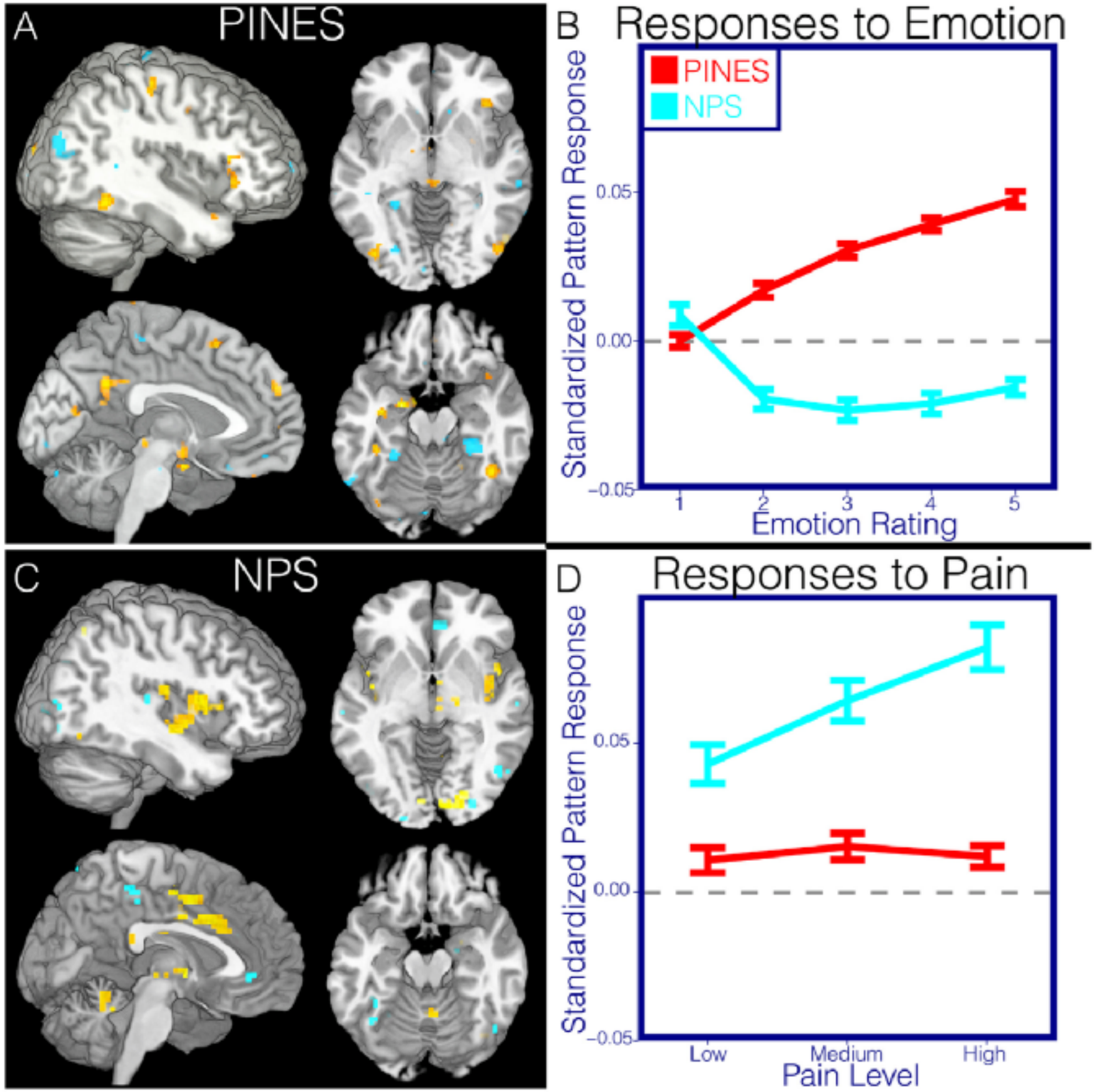


Fig 4. Affective and pain responses to PINES and NPS. This figure illustrates differences in the spatial topography in the thresholded PINES and NPS patterns and their predictions in independent emotion ($n = 61$) and pain ($n = 28$) test data. Panel A depicts the PINES thresholded at $p < 0.001$ uncorrected (see Fig 2). Panel B depicts the average standardized PINES and NPS pattern responses at each level of emotion calculated using a spatial correlation. Error bars reflect ±1 standard error. Panel C depicts the NPS thresholded at false discovery rate (FDR) $q < 0.05$ whole-brain corrected. Panel D depicts the average standardized PINES and NPS pattern responses at each pain level calculated using a spatial correlation. Error bars reflect ±1 standard error.

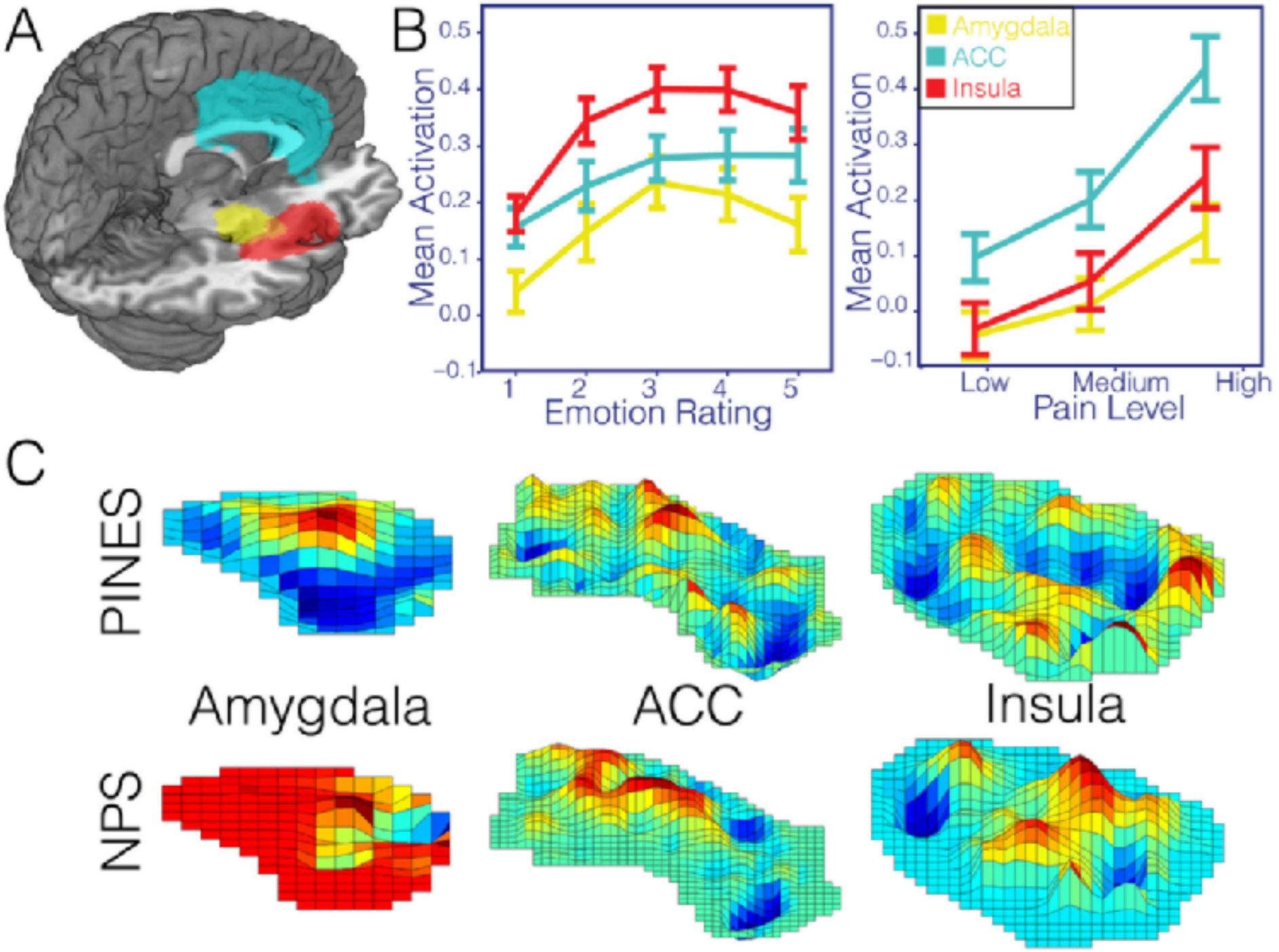


Fig 5. Region of interest analysis. Panel A illustrates the spatial distribution of the three anatomical ROIs used in all analyses (amygdala = yellow, insula = red, ACC = cyan). Panel B depicts the average activation within each ROI across participants for each level of emotion and pain in the emotion hold out ($n = 61$) and pain test datasets ($n = 28$). Error bars reflect ±1 standard error. Panel C illustrates the spatial topography of the PINES and NPS patterns within each of these anatomical ROIs. While these plots show one region, correlations reported in the text reflect bilateral patterns.

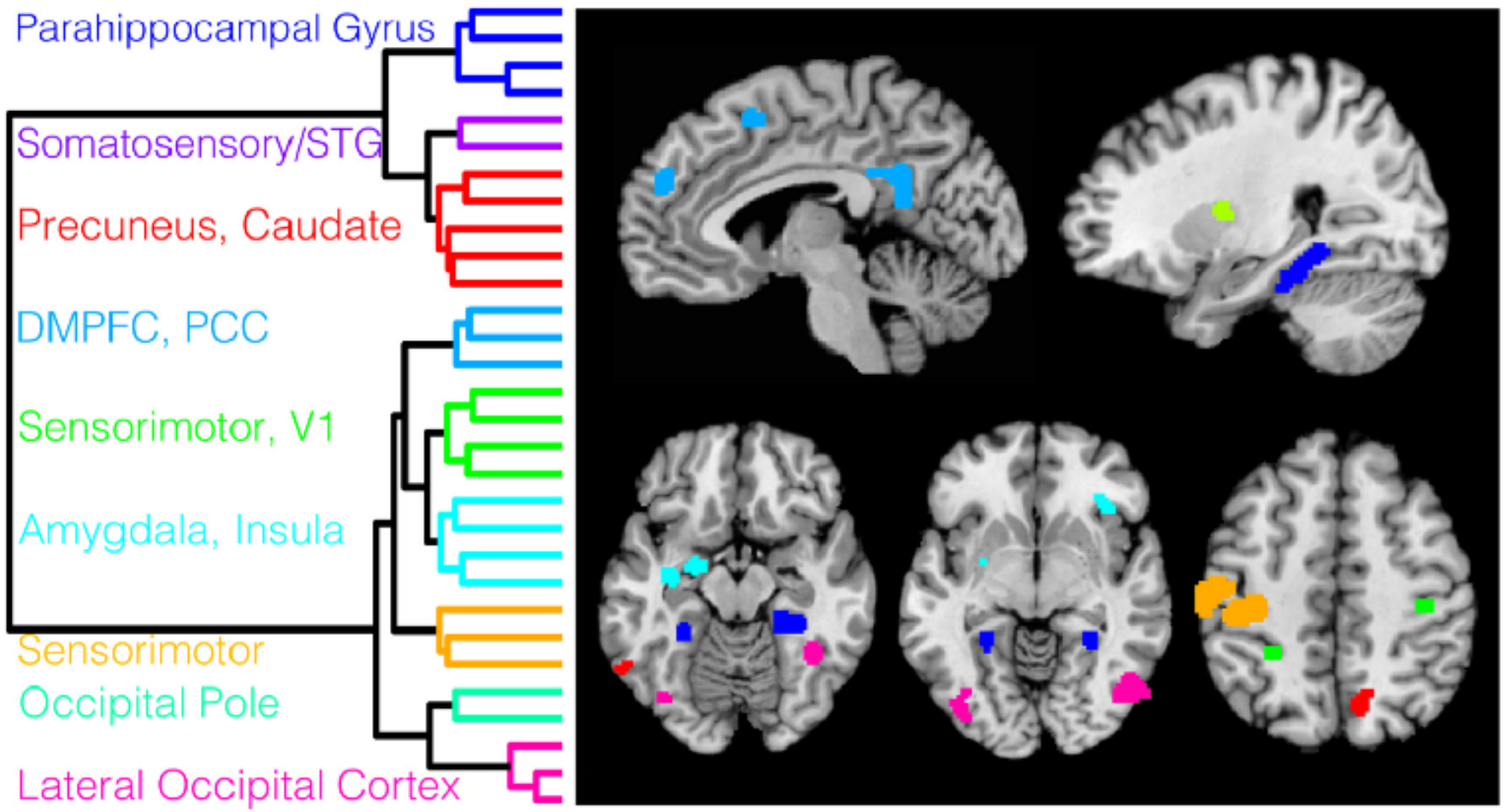


Fig 6. PINES clustering based on shared patterns of connectivity. This figure depicts the results of the hierarchical clustering analysis of the functional connectivity of the largest regions from the $p < 0.001$ thresholded PINES pattern. Clusters were defined by performing hierarchical agglomerative clustering with ward linkage on the trial-by-trial local pattern responses for each region using Euclidean distance. Data were ranked and normalized within each participant and then aggregated by concatenating all 61 subjects' trial x region data matrices. Panel A depicts the dendrogram separated by each functional network. Panel B depicts the spatial distribution of the networks. Colors correspond to the dendrogram labels.

Table 1. Patten sensitivity and specificity.

	Map	Emotion 5 versus 1 (SE)	Pain High versus Low (SE)	Emotion versus Pain (SE)[†]	Emotion Correlation (SE)	Pain Correlation (SE)
Pattern						
	PINES	93.6 (2.6%)[+]	60.7 (8%)	93.2 (2.9%)[+]	0.92 (0.01)	0.64 (0.11)
	Neurologic Pain Signature (NPS)	27.7 (5%)[+*]	82.1 (5.6%)[+]	10.7 (3.6%)[+*]	-0.35 (0.06)	0.91 (0.04)
Average Region of Interest (ROI)						
	Amygdala	55.3 (6%)*	64.3 (8%)	50.5 (5.8%)*	0.31 (0.07)	0.62 (0.09)
	Anterior Cingulate (ACC)	55.3 (5.6%)*	75 (6.7%)[+]	50.5 (5.8%)*	0.26 (0.07)	0.9 (0.02)
	Insula	55.3 (6%)*	78.6 (6.2%)[+]	45.6 (5.7%)*	0.32 (0.07)	0.92 (0.02)
Network						
	Visual	50 (6.5%)*	57.1 (8%)	78.6 (4.7%)[+*]	-0.01 (0.08)	0.22 (0.13)
	Somatomotor	36.2 (6.2%)[+*]	71.4 (7.1%)[+]	28.1 (5.2%)[+*]	-0.38 (0.06)	0.78 (0.09)
	Dorsal Attention	57.4 (6.4%)*	71.4 (6.2%)[+]	61.2 (5.6%)*	0.34 (0.07)	0.57 (0.12)
	Ventral Attention (Salience)	51.1 (6%)*	71.4 (6.2%)[+]	13.5 (3.9%)[+*]	0.14 (0.07)	0.56 (0.13)
	Limbic	57.4 (6%)*	35.7 (8%)	53.4 (5.8%)*	0.28 (0.06)	-0.5 (0.13)
	Frontoparietal	51.1 (5.8%)*	60.7 (7.6%)	42.7 (5.7%)*	0.29 (0.07)	0.34 (0.13)
	Default	63.8 (5.4%)[+*]	57.1 (7.6%)	70.8 (5.3%)[+*]	0.34 (0.06)	-0.03 (0.15)

All balanced accuracies reported in this table result from single-interval classification on the test dataset (*n* = 47; see S1 Table for forced-choice test). Analyses involving Level 5 and/or Level 1 comparisons exclude participants that did not rate any stimuli with that label. Accuracy values reflect the ability to discriminate the conditions compared, but are signed, so that values >50% indicate the proportion of participants for which high intensity was classified as greater than low intensity, for high vs. low analyses, or emotion was greater than pain, for Emotion vs. Pain analyses. Values < 50% indicate the proportion of participants for which low intensity was classified as greater than high intensity or pain was classified as greater than emotion. For example, the 10.7% emotion classification of the NPS in the Emotion vs. Pain analysis should be interpreted as a 89.3% hit rate in discriminating pain from emotion. Correlations reflect Pearson correlations between participant's pattern responses to levels of affective intensity and self-reported ratings averaged across participants.

[†]Please note that this column does not reflect accuracy but rather percent classified as emotion.

[+]Indicates that accuracy is significantly different from chance (50%), using a two-tailed dependent binomial test.

*Indicates accuracy significantly different from PINES performance using a two-sample two-tailed z-test for proportions (only tested on Emotion 5 versus 1 and Emotion versus Pain columns).

Table 2. Single-cluster and "virtual lesion" analysis.

	Map	nVoxels	Emotion 5 versus 1 (SE)	Pain H versus L (SE)	Emotion versus Pain (SE)	Emotion Correlation (SE)	Pain Correlation (SE)
Pattern							
	PINES	328796	93.5 (2.4%)	60.7 (6.5%)	93.2 (2.9%)	0.92 (0.01)	0.64 (0.11)
	PINES ($p < .001$)	5303	91.5 (3%)$^{+}$	67.9 (7.6%)$^{-}$	97.2 (1.9%)$^{+}$	0.89 (0.01)	0.51 (0.13)
Single Cluster							
	Visual (LOC)	981	83 (4.3%)$^{+*}$	64.3 (7.1%)$^{-}$	85.4 (4.1%)$^{+}$	0.73 (0.03)	0.56 (0.12)
	Somatosensory and superior temporal gyrus (STG)	308	59.6 (5.8%)*	32.1 (7.1%)$^{-}$	61.2 (5.6%)*	0.12 (0.07)	-0.66 (0.11)
	Sensorimotor and V1	335	57.4 (6.2%)*	67.9 (7.6%)$^{-}$	57.3 (5.7%)*	0.23 (0.07)	0.8 (0.07)
	DMPFC and PCC	318	70.2 (5.4%)$^{+*}$	60.7 (7.6%)	70.8 (5.3%)$^{+*}$	0.47 (0.06)	0.61 (0.1)
	Sensorimotor and Cerebellum	1227	78.7 (4.5%)$^{+*}$	60.7 (7.6%)	93.2 (2.9%)$^{+}$	0.72 (0.04)	0.39 (0.14)
	Parahippocampal Gyrus	1025	51.1 (6.4%)*	39.3 (7.1%)	39.9 (5.7%)*	-0.05 (0.07)	-0.43 (0.13)
	Occipital Pole	118	55.3 (6.7%)*	53.6 (8%)	85.4 (4.1%)$^{+}$	0.29 (0.08)	0.22 (0.14)
	Precuneus and Caudate	537	48.9 (6.2%)*	28.6 (7.1%)$^{-}$	53.4 (5.8%)*	-0.15 (0.07)	-0.82 (0.06)
	Amygdala and Insula	454	59.6 (6%)*	75 (6.7%)$^{+}$	54.4 (5.7%)*	0.39 (0.06)	0.76 (0.08)
Virtual Lesion-Cluster Removed	Visual (LOC)	4322	85.1 (4%)$^{+}$	46.4 (8.4%)	96.1 (2.3%)$^{+}$	0.72 (0.05)	-0.17 (0.13)
	Somatosensory and STG	4995	91.5 (3%)$^{+}$	64.3 (8%)	93.2 (2.9%)$^{+}$	0.87 (0.01)	0.67 (0.11)
	Sensorimotor and V1	4968	95.7 (2.1%)$^{+}$	50 (8%)	97.2 (1.9%)$^{+}$	0.9 (0.01)	0.08 (0.15)
	DMPFC and PCC	4985	89.4 (3.4%)$^{+}$	57.1 (8.7%)	97.2 (1.9%)$^{+}$	0.9 (0.01)	0.37 (0.14)
	Sensorimotor and Cerebellum	4076	91.5 (3%)$^{+}$	60.7 (8.4%)	96.1 (2.3%)$^{+}$	0.84 (0.02)	0.56 (0.11)
	Parahippocampal Gyrus	4278	85.1 (4%)$^{+}$	67.9 (7.1%)$^{-}$	96.1 (2.3%)$^{+}$	0.83 (0.02)	0.62 (0.11)
	Occipital Pole	5185	93.6 (2.6%)$^{+}$	64.3 (7.6%)	97.2 (1.9%)$^{+}$	0.89 (0.01)	0.46 (0.14)
	Precuneus and Caudate	4766	89.4(3.4%)$^{+}$	66.1(7.8%)$^{+}$	96.1(2.3%)$^{+}$	0.85(0.02)	0.76(0.07)
	Amygdala and Insula	4849	91.5(3%)$^{+}$	57.1(8.4%)	97.2(1.9%)$^{+}$	0.9(0.01)	0.25(0.15)

All balanced accuracies reported in this table result from single interval classification on the test sample ($n = 47$; see S2 Table for forced-choice test). Analyses involving Level 5 and/or Level 1 comparisons exclude participants that did not rate any stimuli with that label. Accuracy values reflect the ability to discriminate the conditions compared, but are signed so that values >50% indicate the proportion of participants for which high intensity was classified as greater than low intensity for high vs. low analyses, or emotion was greater than pain for Emotion vs. Pain analyses. Values < 50% indicate the proportion of participants for which low intensity was classified as greater than high intensity or pain was classified as greater than emotion. For example, the 10.7% emotion classification of the NPS in the Emotion vs. Pain analysis should be interpreted as a 89.3% hit rate in discriminating pain from emotion. Correlations reflect Pearson correlations between participant's pattern responses to levels of affective intensity and self-reported ratings averaged across participants.

$^{+}$Indicates that accuracy is significantly different from chance (50%) using a two-tailed binomial test.

*Indicates accuracy is significantly different from PINES performance using a two-sample, two-tailed z-test for proportions (only tested on Emotion 5 versus 1 and Emotion versus Pain columns).

BERMAN

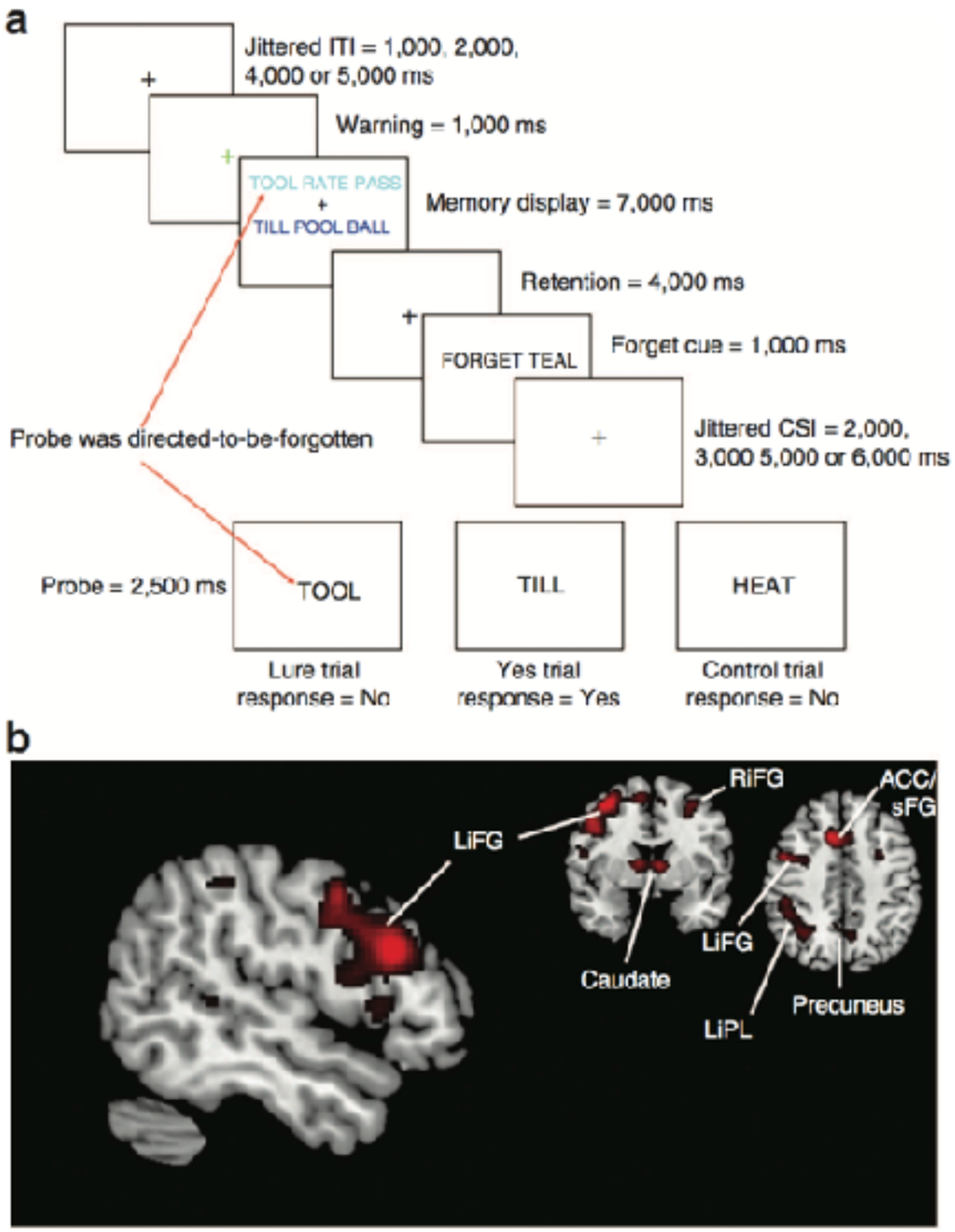


Figure 1 | Directed-forgetting task and activations. (**a**) Schematic of the working memory directed-forgetting task. The task is composed of three trial types: lure, yes and control trials. Of most interest is the comparison of accuracy and RT for lure versus control trials. (**b**) Activation patterns for the lure—control contrast across all participants. Significant activation is seen in the left inferior frontal gyrus (LiFG), the right inferior frontal gyrus (RiFG), the anterior cingulate cortex (ACC)/superior frontral gyrus (sFG), the caudate, the precuneus and the left inferior parietal lobule (LiPL). These images are thresholded at $P<0.005$ uncorrected for ten contiguous voxels.

Table 1 | RT and ACC data by group with s.d. in parentheses.

	Trial type		
	Control	Yes	Lure
ACC data (% correct)			
High	93.3% (11.4)	90.0% (6.0)	84.0% (12.2)
Low	91.7% (12.2)	87.3% (7.7)	81.7% (9.7)
RT data (ms)			
High	876.6 (181.5)	867.0 (158.2)	1,036.1 (214.0)
Low	920.3 (138.6)	900.7 (166.2)	1,116.1 (182.2)

ACC, accuracy; RT, reaction time

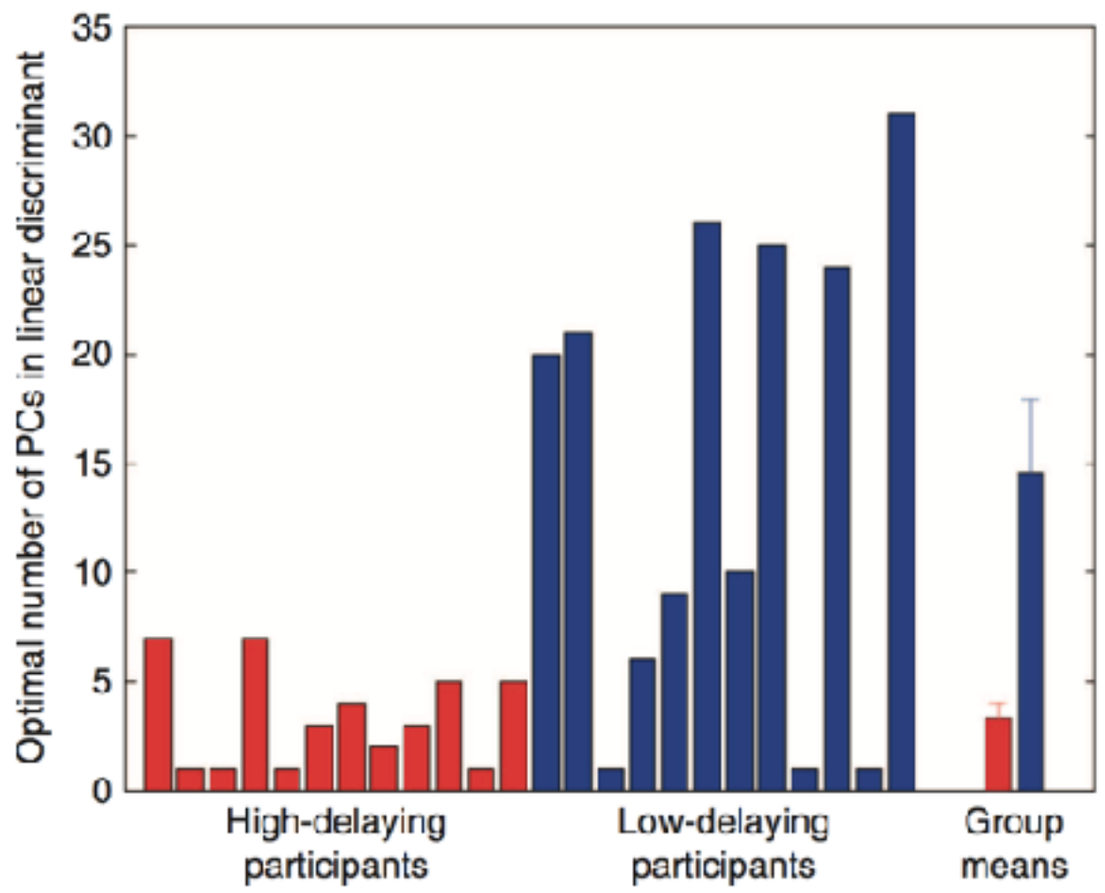


Figure 2 | The optimal number of dimensions to maximize classification accuracy. The number of LD dimensions/components that were required to achieve maximum classification between lure and control trials for each participant with the group averaged data to the far right. High-delay group = red; low-delay group = blue. Error bars represent s.e.m.

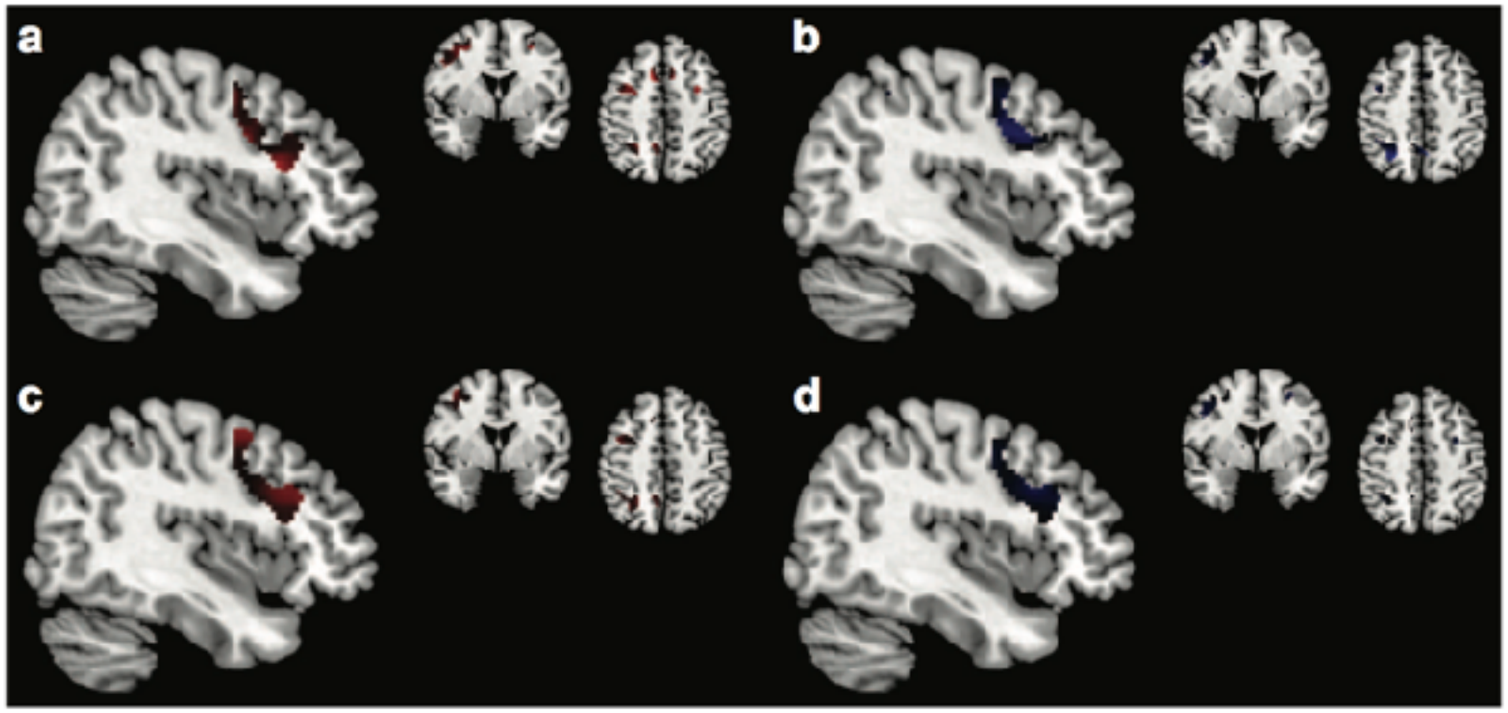


Figure 3 | Averaged first and second PC for each group. (**a**) Average PC 1 for high delayers (**b**) Average PC 1 for low delayers (**c**) Average PC 2 for high delayers (**d**) Average PC 2 for low delayers.

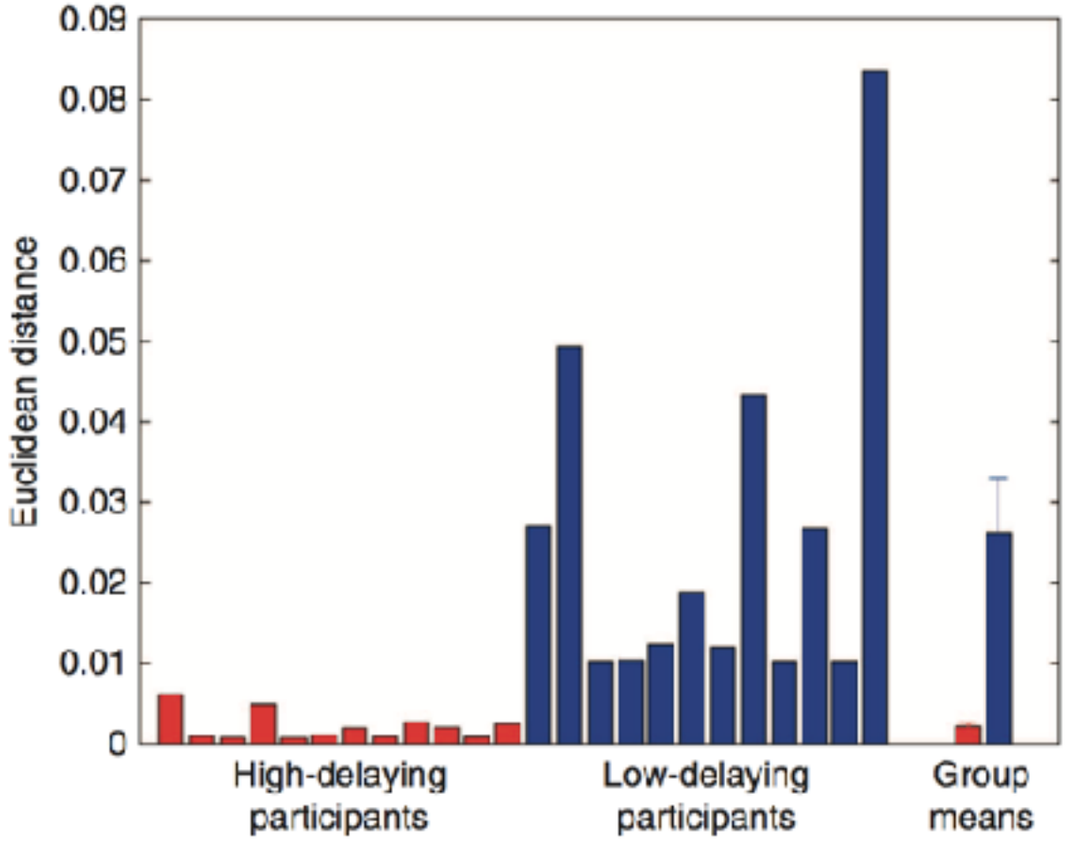


Figure 4 | The Euclidean distances for each individual participant's LD map from their group mean LD map. High-delay group = red; low-delay group = blue. Error bars represent s.e.m.

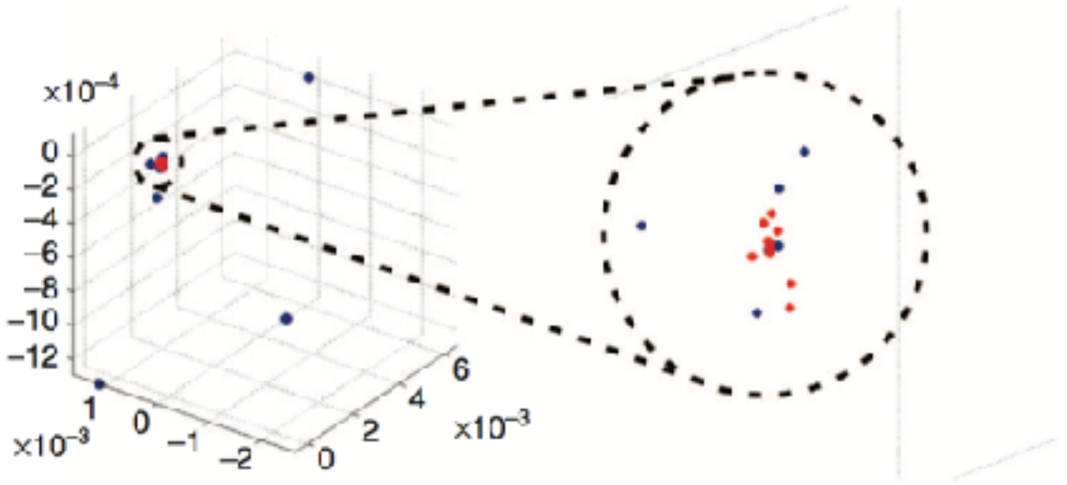


Figure 5 | MDS results for the first three dimensions of the LD map distance matrix. The high (red) delayers are grouped together more closely than the low delayers (blue).

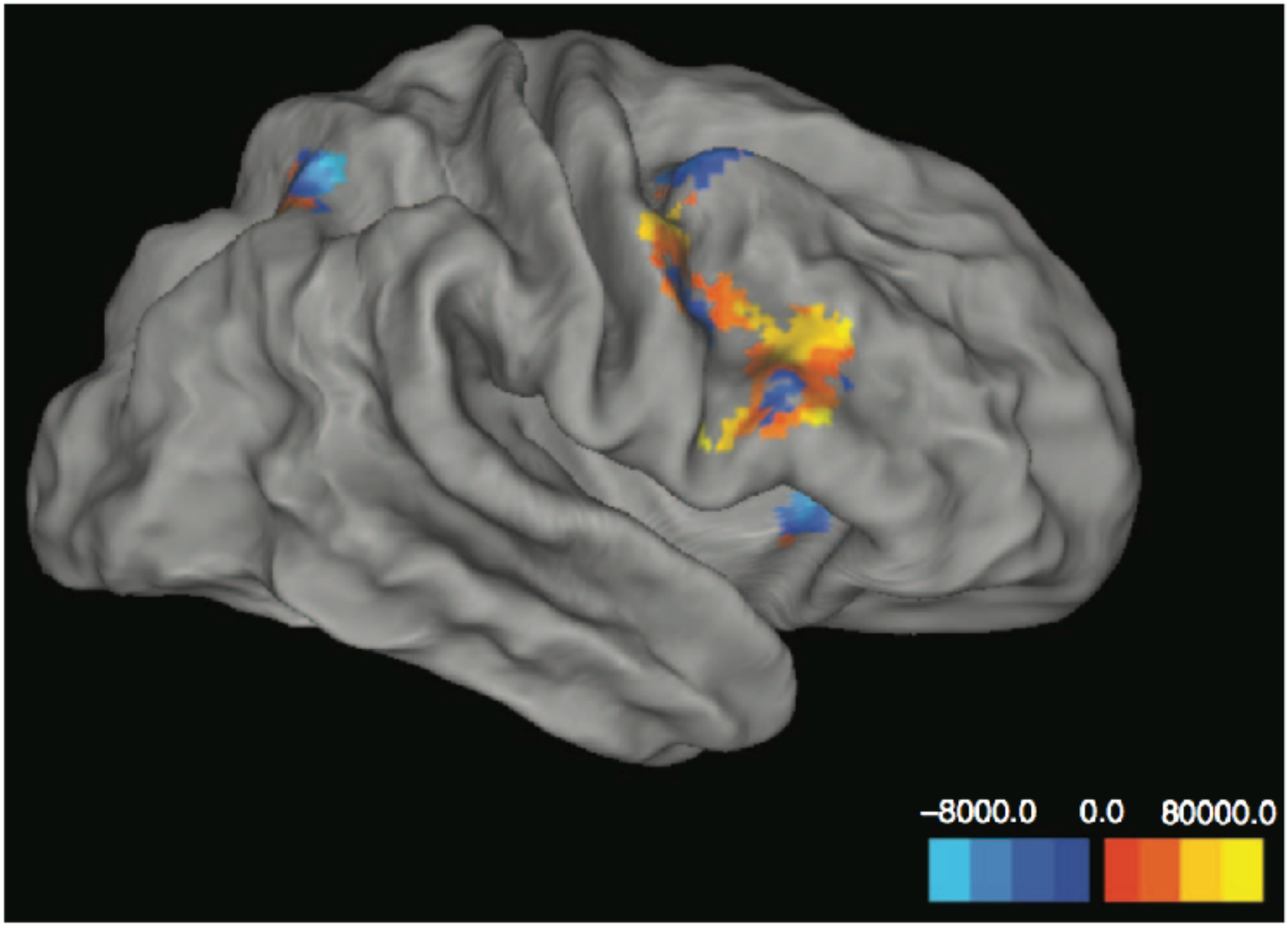


Figure 6 | QD sensitivity map for classifying high- versus low delayers' LD maps. Areas in blue represent voxels that are higher in low delayers' maps. Areas in orange/yellow represent voxels that are higher in high delayers' maps. The left hemisphere is shown.