Pattern Recognition Methods: basics

James Chapman

Slides adapted from Fabio Ferreira/Jessica Schrouff

Course 2021

TBC

UCL, London

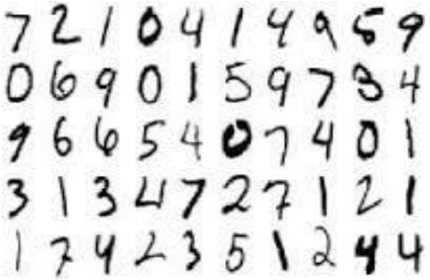
Outline

- **Pattern Recognition and Supervised Learning**
  - **Concepts & Advantages**
  - **Generalisation**
- **Preprocessing and Feature Extraction**
- **Linear Predictive Functions**
- **Challenges and Solutions in Neuroimaging**
  - **Regularisation**
  - **Kernel Methods**
- **Linear Models/Machines in PRoNTo**

Pattern Recognition

Pattern recognition aims to find patterns in the data which can be used to extract meaningful information to make predictions

Digit Recognition



Face Recognition

Finance

Advertising and Business Intelligence

Google Ads

Recommendation Engines

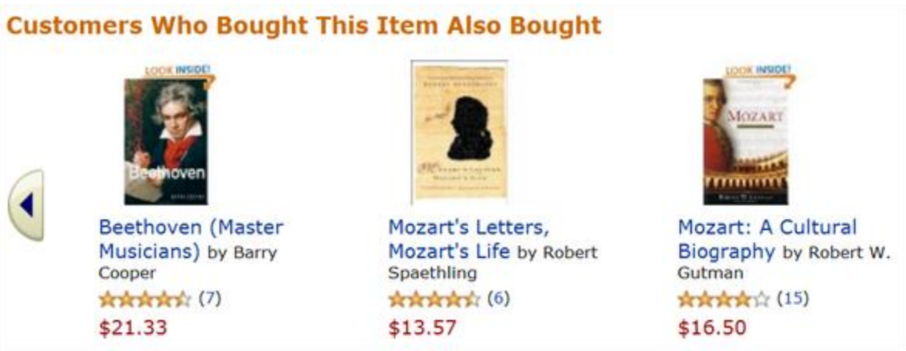
Types of Machine Learning

	Supervised	Unsupervised	Reinforcement Learning
Inputs	Features + Labels	Features	Environment/Actions
Goal	Prediction	Representation	Maximise Rewards

Labels can be discrete categories (Classification) or continuous (Regression)

Types of Machine Learning

	Supervised	Unsupervised	Reinforcement Learning
Inputs	Features + Labels	Features	Environment/Actions
Goal	Prediction	Representation	Maximise Rewards



Customers Who Bought This Item Also Bought

Beethoven (Master Musicians) by Barry Cooper
★★★★★ (7)
$21.33

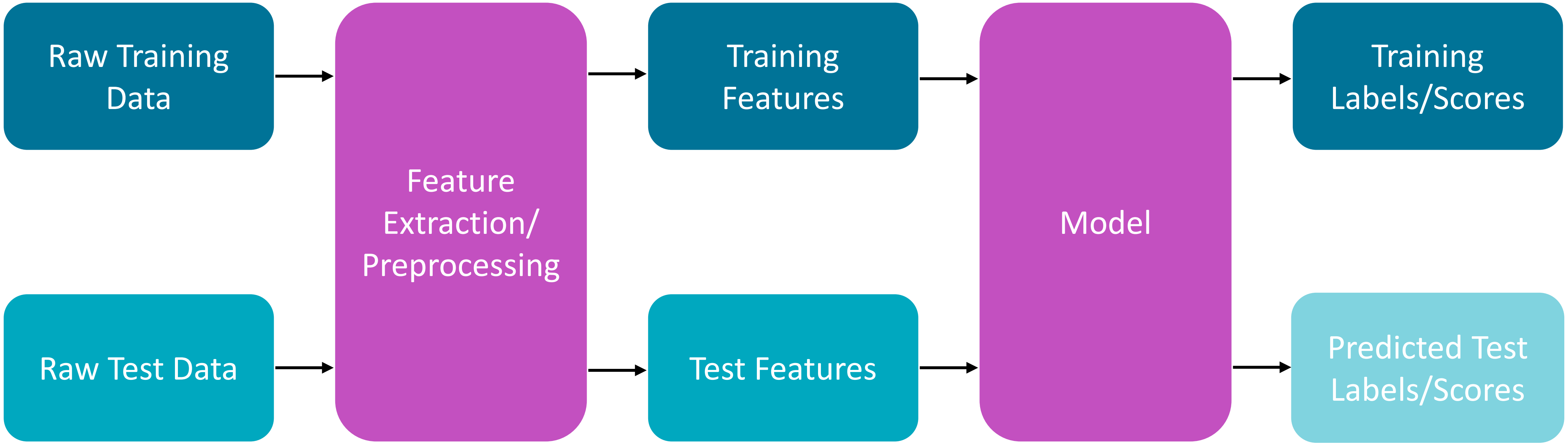
Mozart's Letters, Mozart's Life by Robert Spaethling
★★★★★ (6)
$13.57

Mozart: A Cultural Biography by Robert W. Gutman
★★★★☆ (15)
$16.50

Types of Machine Learning

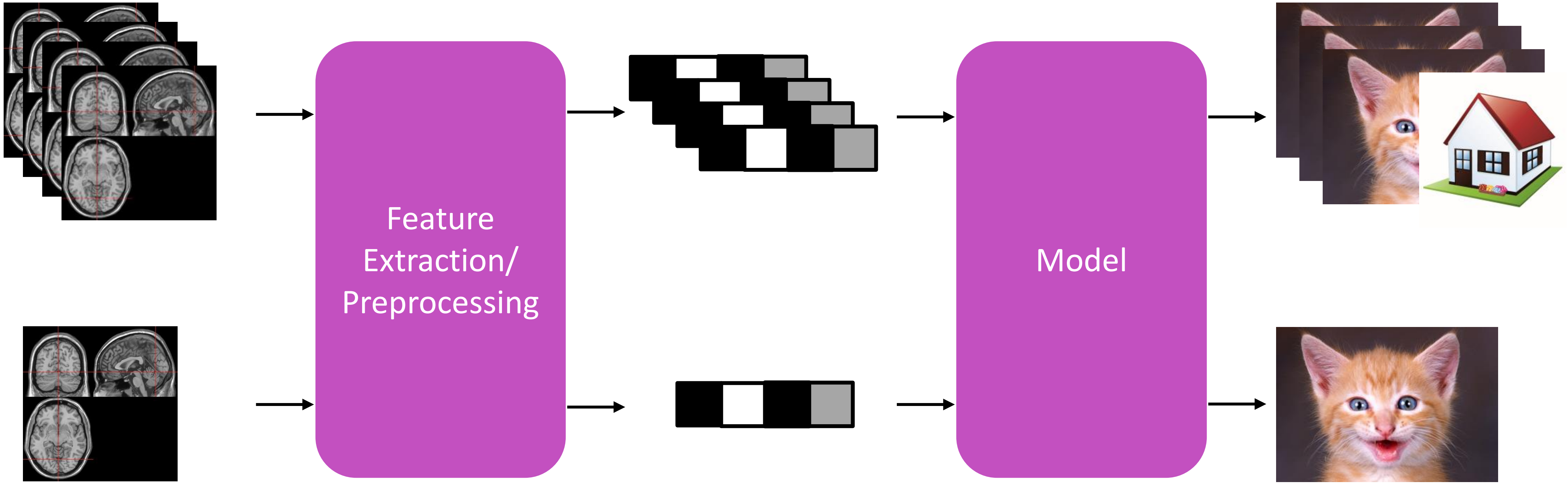
	Supervised	Unsupervised	Reinforcement Learning
Inputs	Features + Labels	Features	Environment/Actions
Goal	Prediction	Representation	Maximise Rewards

Supervised Learning Framework



The goal of supervised learning is prediction

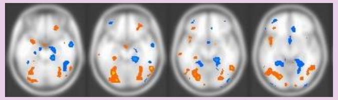
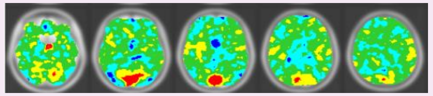
Supervised Learning in Neuroimaging



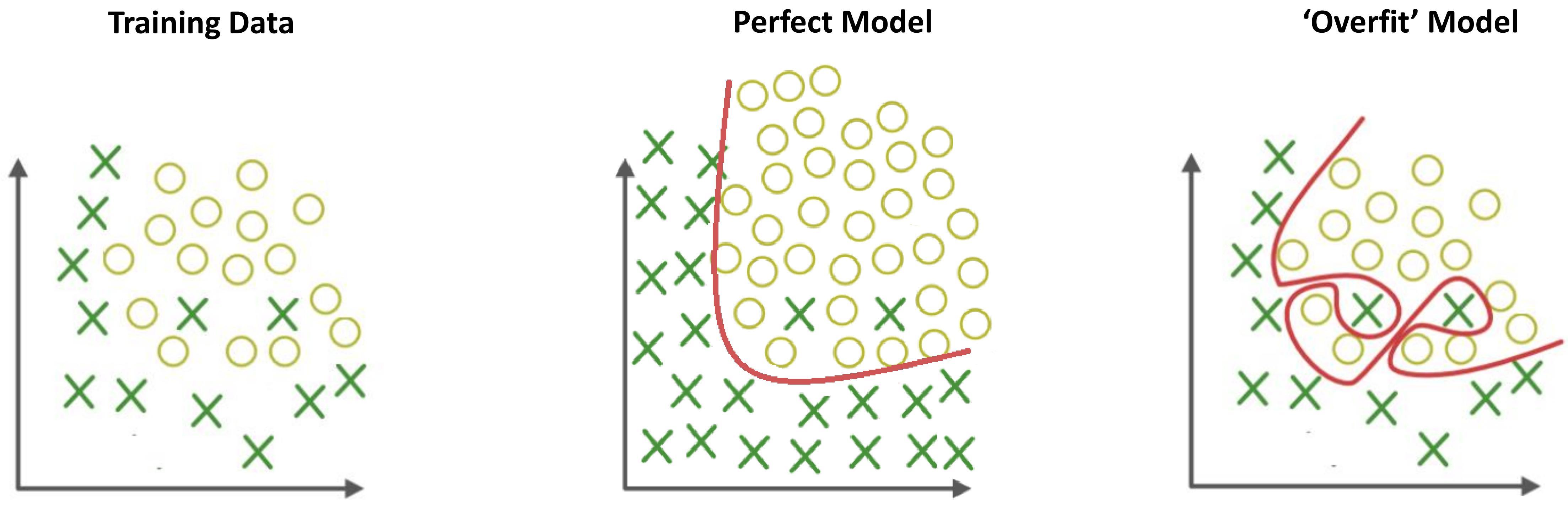
Feature Extraction/ Preprocessing

Model

Supervised Learning and Statistical Analysis

	Model Assumptions	Model Goal	Measure of Model	Output
Statistical Analysis	Independent Voxels	Inference	Statistical Significance	Univariate Map 
Supervised Learning	Voxels can be correlated	**Prediction**	Generalisation on **Test Data**	Multivariate Map 

Supervised Learning: Generalisation



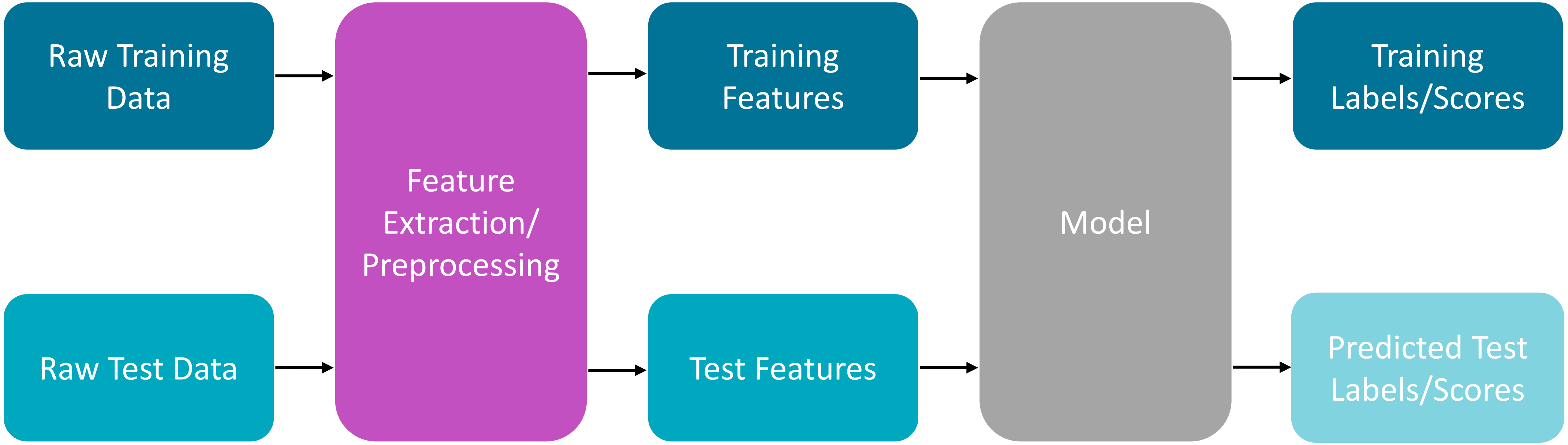
Training Data

Perfect Model

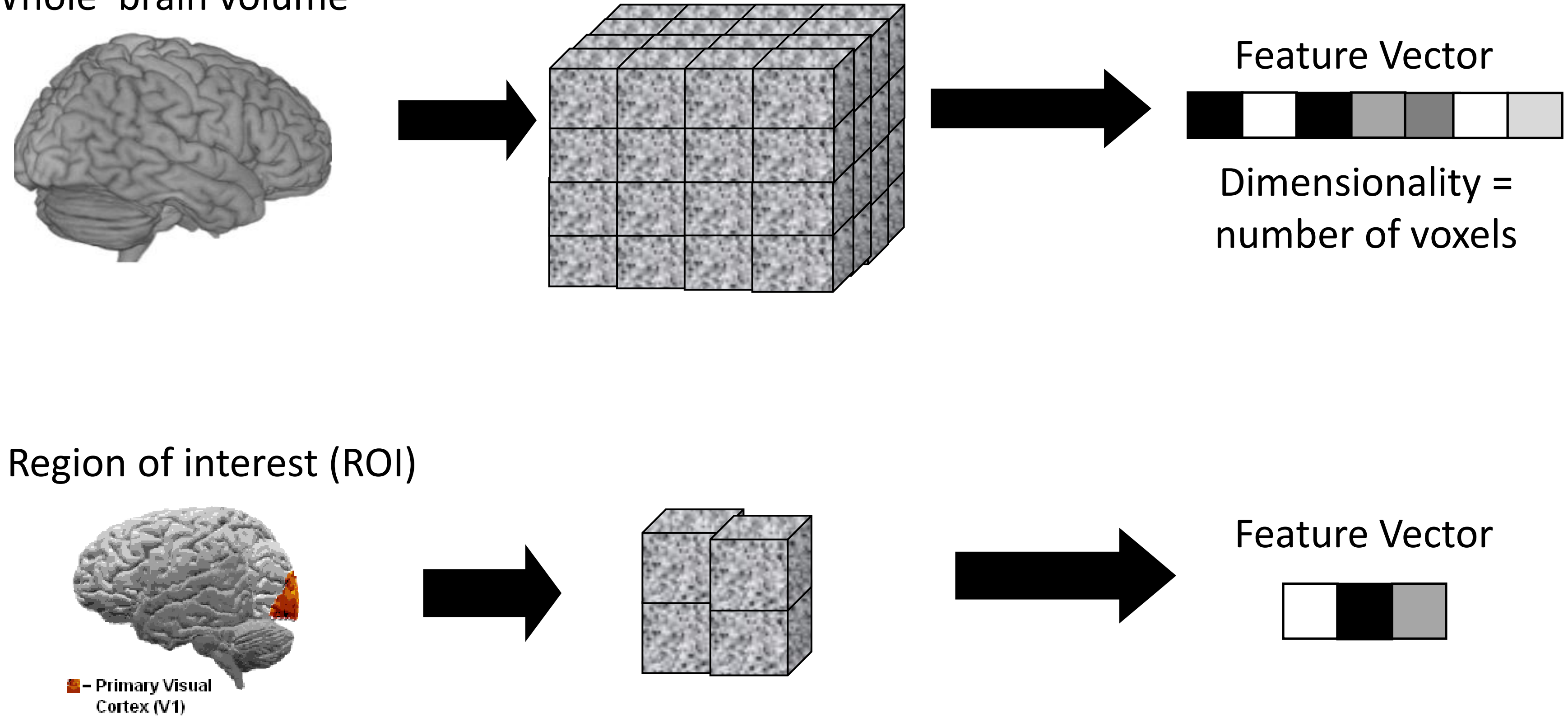
'Overfit' Model

A model that performs better on the training data might not perform better on unseen test data

Feature extraction and Preprocessing



Feature extraction and Preprocessing



Whole brain volume

fMRI/sMRI
3D matrix of voxels

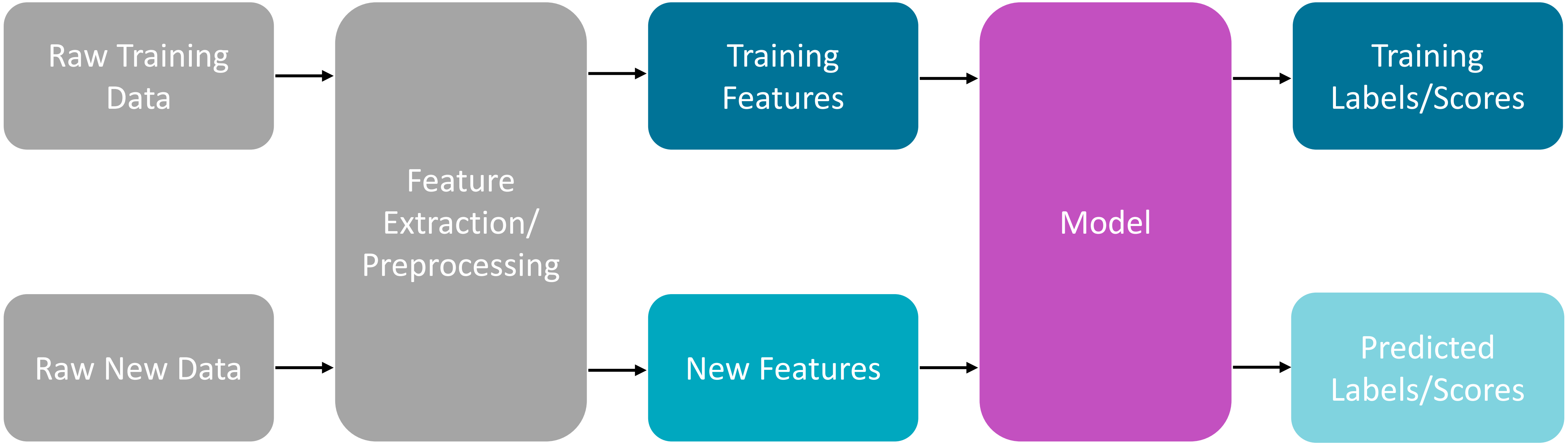
Feature Vector

Dimensionality =
number of voxels

Region of interest (ROI)

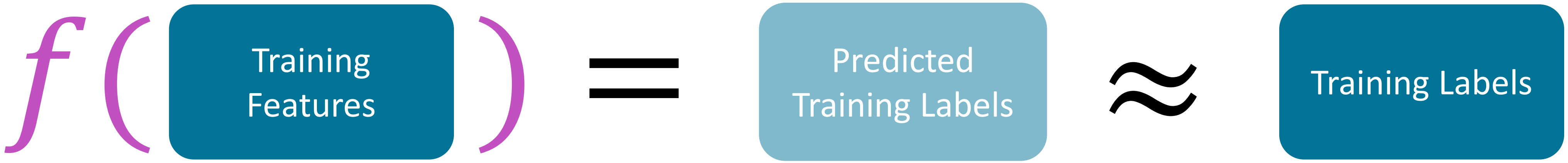
Feature Vector

Model

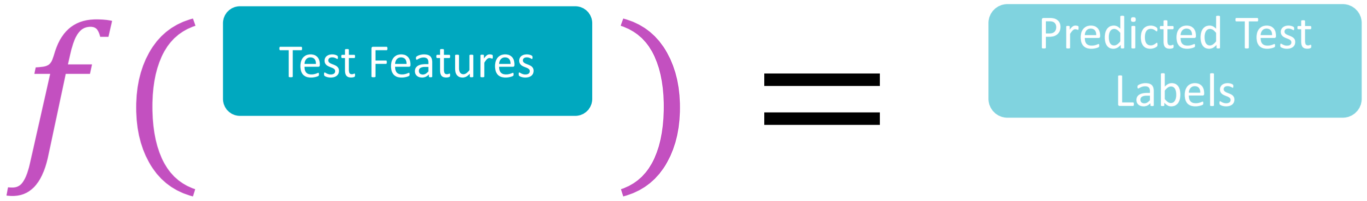

Raw Training Data
Raw New Data
Feature Extraction/ Preprocessing
Training Features
New Features
Model
Training Labels/Scores
Predicted Labels/Scores

Model

Our model can be understood as a function of the training features that best predicts the training labels

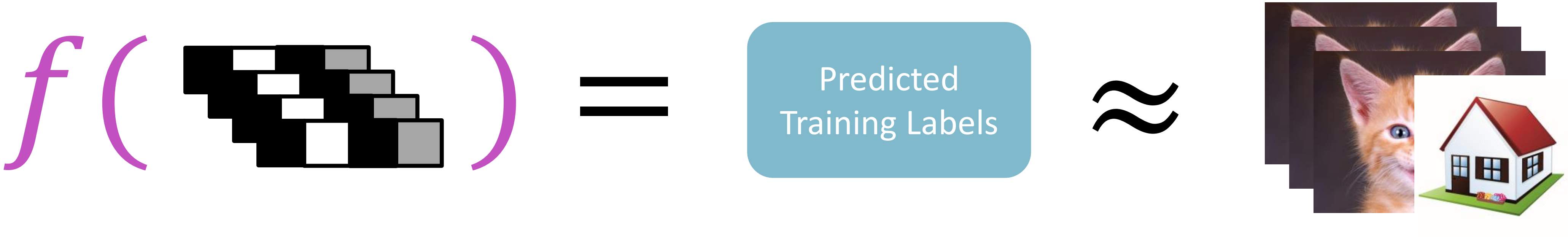


Model can then be applied unseen data



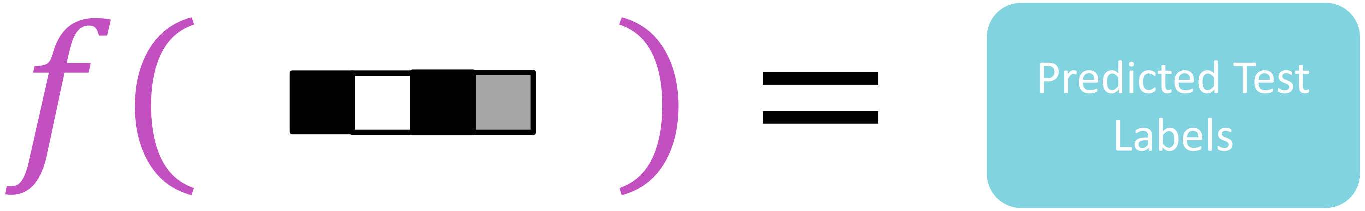
Model

Our model can be understood as a function of the training features that best predicts the training labels



Predicted Training Labels

Model can then be applied unseen data



Predicted Test Labels

Linear predictive function

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$$

- Linear predictive functions (classifier or regression) are parameterized by a weight vector **w** and a bias term *b*

- We can optimise these parameters so that the difference between:

$$f(\boldsymbol{x}_{train}) = y_{predicted} \text{ and } y_{train}$$

- We can apply this function to test examples as:

$$f(\boldsymbol{x}_{test}) = y_{predicted}$$

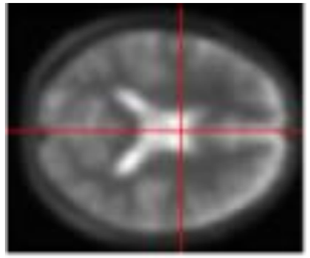
Linear predictive function: prediction

Estimated (**w**, *b*) →

🟥 5	🟨 -6	🟦 -1	🟩 2

× × ×

New example ($\boldsymbol{x}^*$)



→

1	2	-2

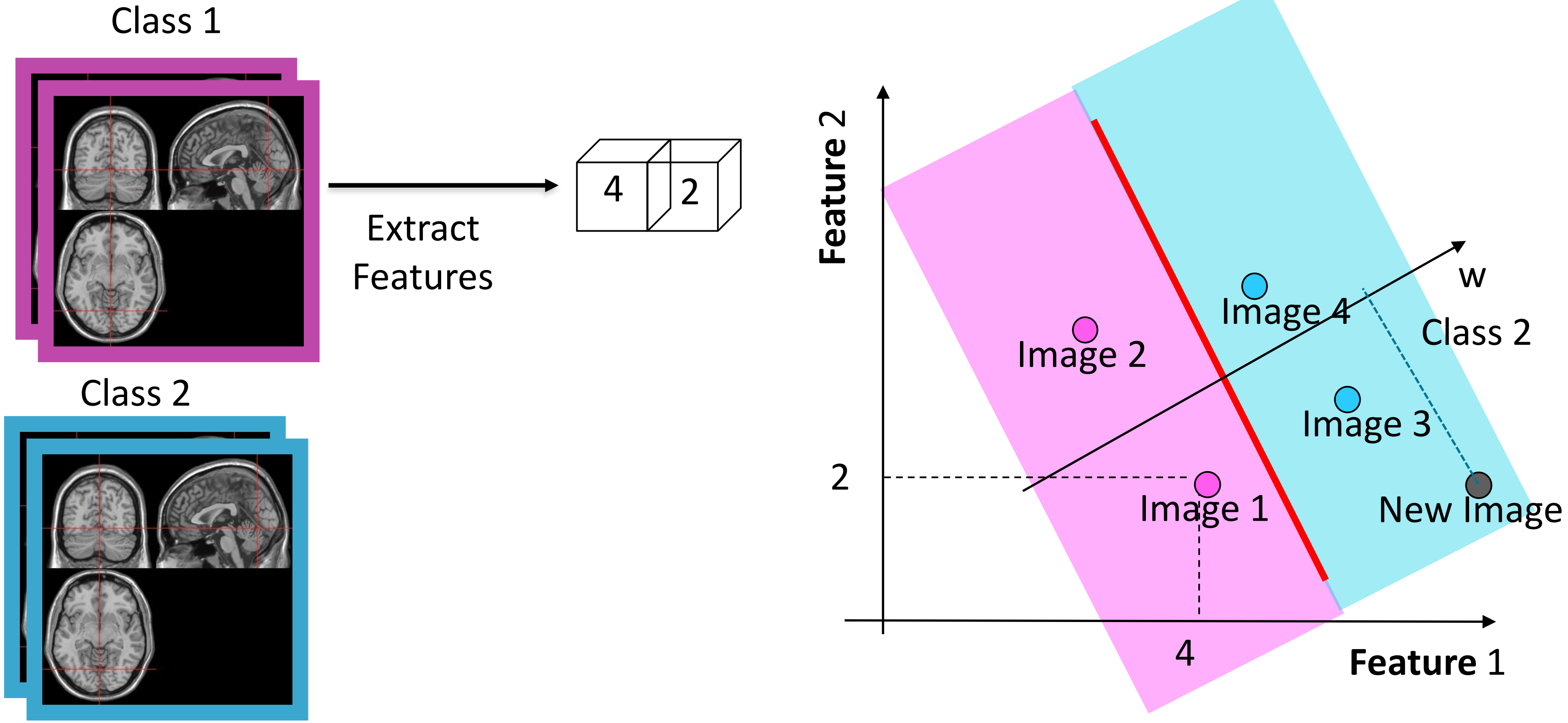
Predictive function:

$$f(\boldsymbol{x}^*) = \boldsymbol{w} \cdot \boldsymbol{x}^* + b$$

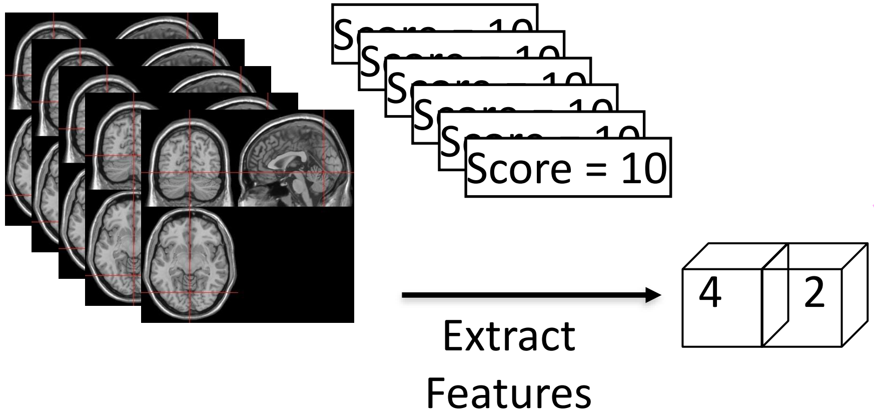
$$f(\boldsymbol{x}^*) = (5 \times 1) + (-6 \times 2) + (-1 \times -2) + 2$$
$$f(\boldsymbol{x}^*) = -3$$

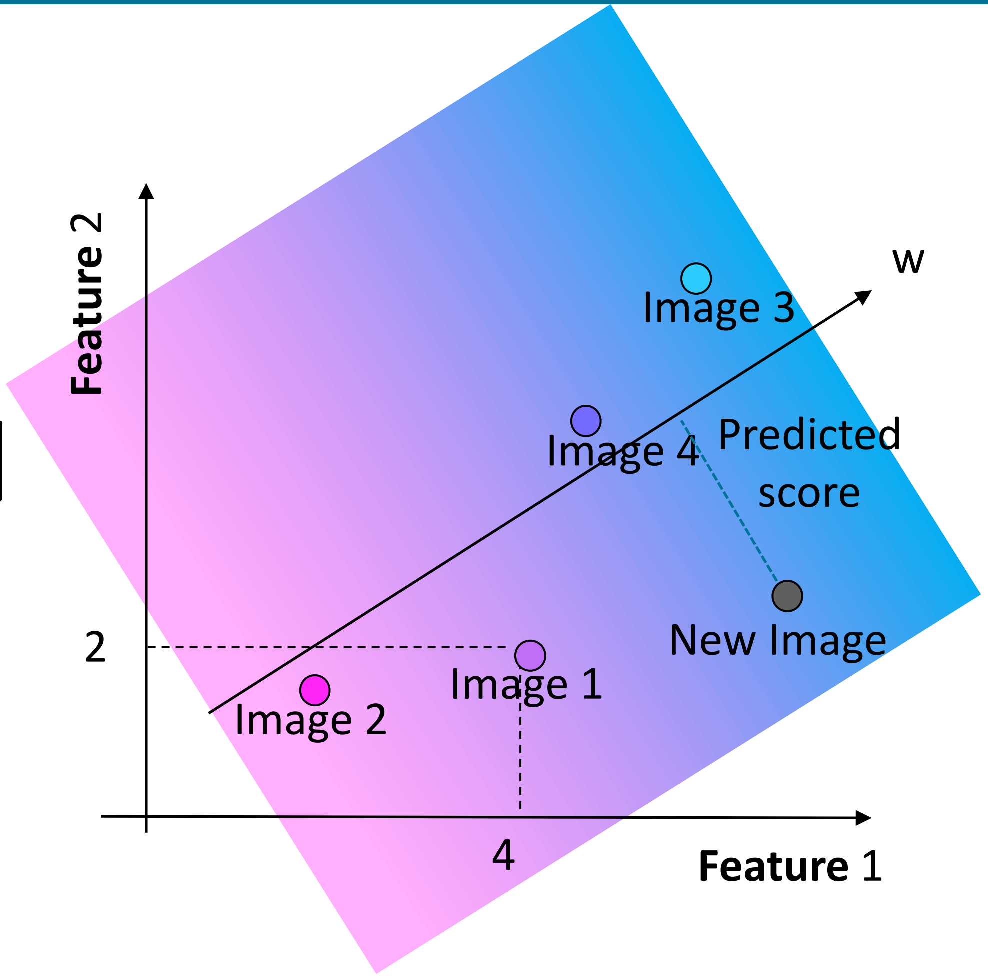
$f(\boldsymbol{x}^*)$ is the predicted score for regression or the distance to the decision boundary for classification models.

Classification Model in 2D



Regression Model in 2D





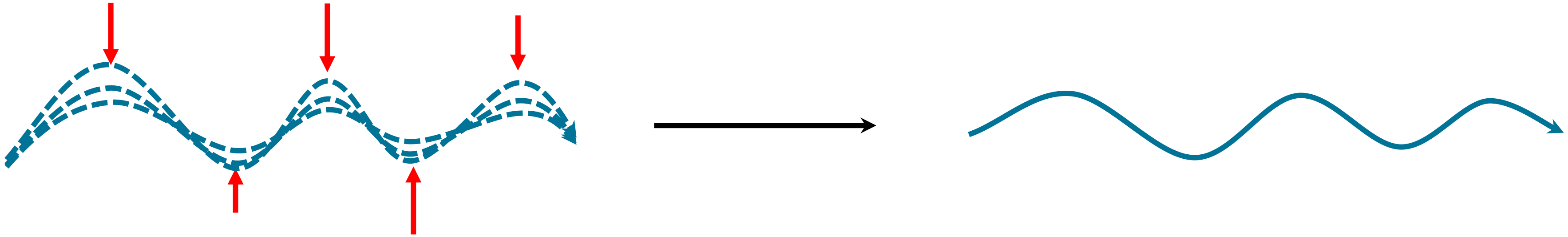
Challenges in Neuroimaging

Problem	Manual Solution	**Data Driven Solution**
Overfitting (poor generalization)	Manual feature selection strategies	**Regularisation**
Slow Computation	Only consider smaller regions of interest (ROIs)	**Kernel Methods**

Regularisation

- **Regularisation** is the technique of adding information to **solve ill-posed problems** and/or to **prevent overfitting** in statistical/machine learning models.

- Ridge regularization encourages weights to be smaller and therefore more plausible



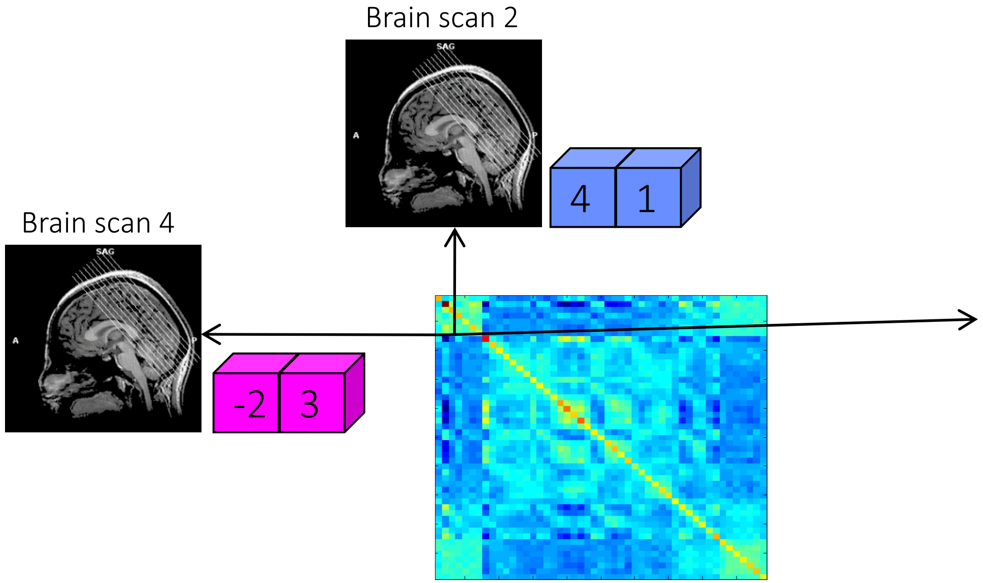
- (Lasso regularization encourages some weights to be zero which helps interpretability and can prevent overfitting if the underlying relationship is sparse)

Kernel Methods

Kernel Function ("similarity" measure)

- Kernel is a function that, given **x** and **x**$_*$, returns a real number characterizing their similarity
- A simple type of similarity measure between two vectors is a dot product (**linear kernel**)

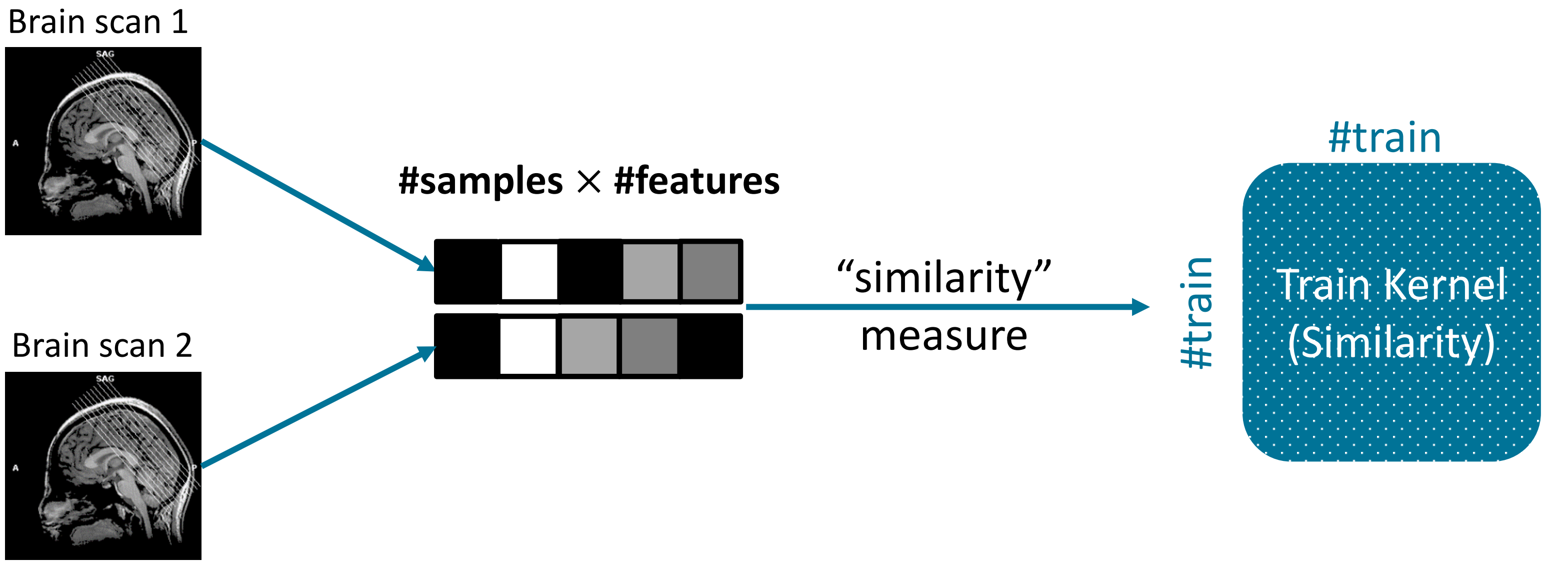
$$\kappa(\boldsymbol{x}, \boldsymbol{x}_*) = \langle \boldsymbol{x} \cdot \boldsymbol{x}_* \rangle$$



Dot product = (4 × -2)+(1 × 3) = -5

Kernel Methods

How can we solve the high-dimensional problem efficiently?



Brain scan 1

Brain scan 2

#samples × #features

"similarity" measure

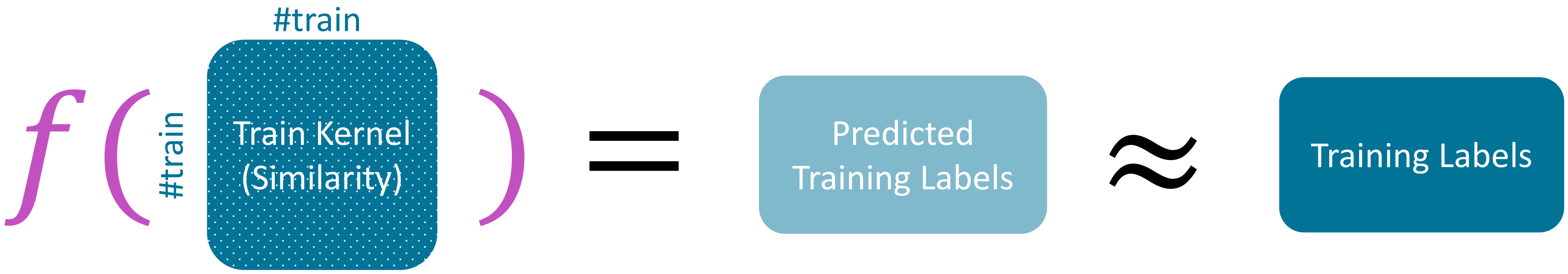
#train

#train

Train Kernel (Similarity)

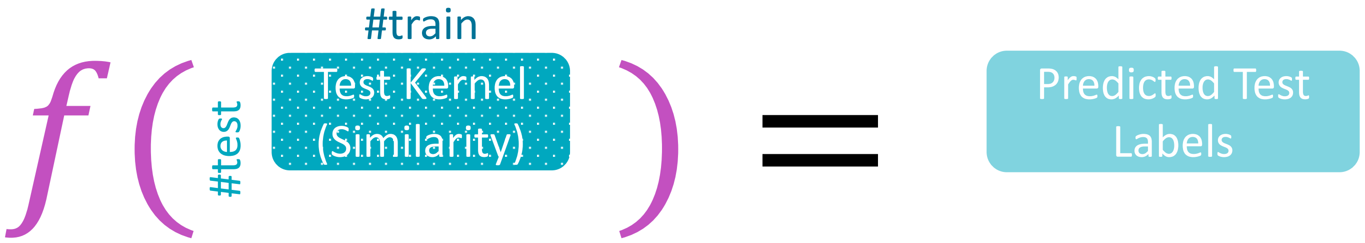
Model in kernel space

Our model can be understood as a function of the *similarity with the training samples* that best predicts the training labels



$f($ #train × #train Train Kernel (Similarity) $)$ = Predicted Training Labels ≈ Training Labels

Model can then be applied unseen data by computing similarities with the training data for each test point



$f($ #test × #train Test Kernel (Similarity) $)$ = Predicted Test Labels

Models in Pronto

Deterministic (hard classifications)	Probabilistic (soft classifications)
Kernel Ridge Regression	Gaussian Process Classifier
Support Vector Machine	Relevance Vector Machine
Multiple Kernel Learning	

Kernel Ridge Regression (KRR)

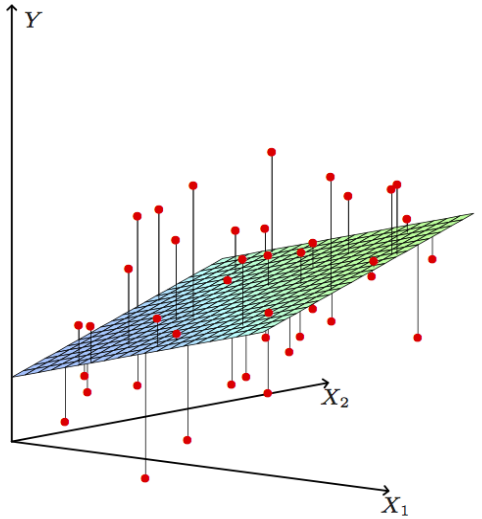
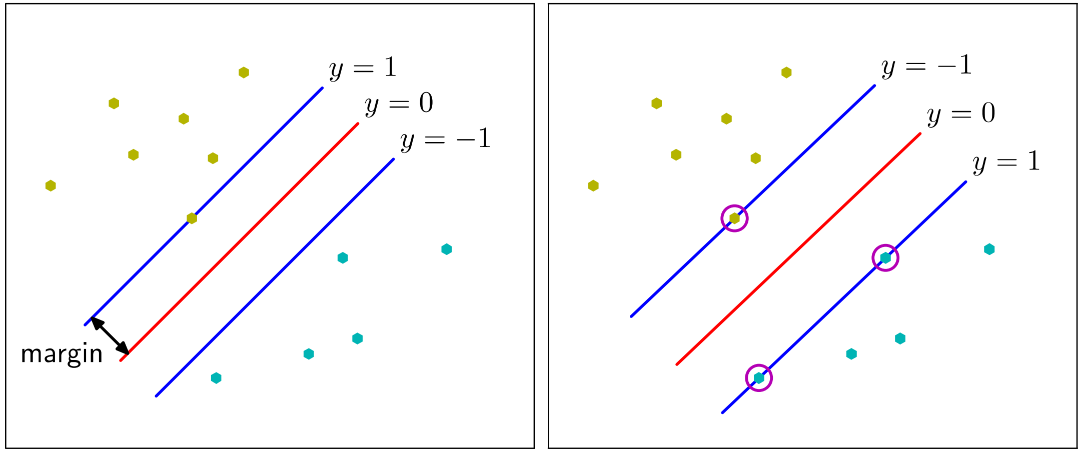


Illustration of a linear least squares fitting with X ∈ IR^2. We seek the linear function of X that minimizes the sum of squared residuals from Y.

Hastie, Tibshirani & Friedman, 2009

- Fit a linear model that minimizes the squared prediction error
- Ridge regularization encourages small weights
- Kernel method makes computation fast

Support Vector Machine (SVM)

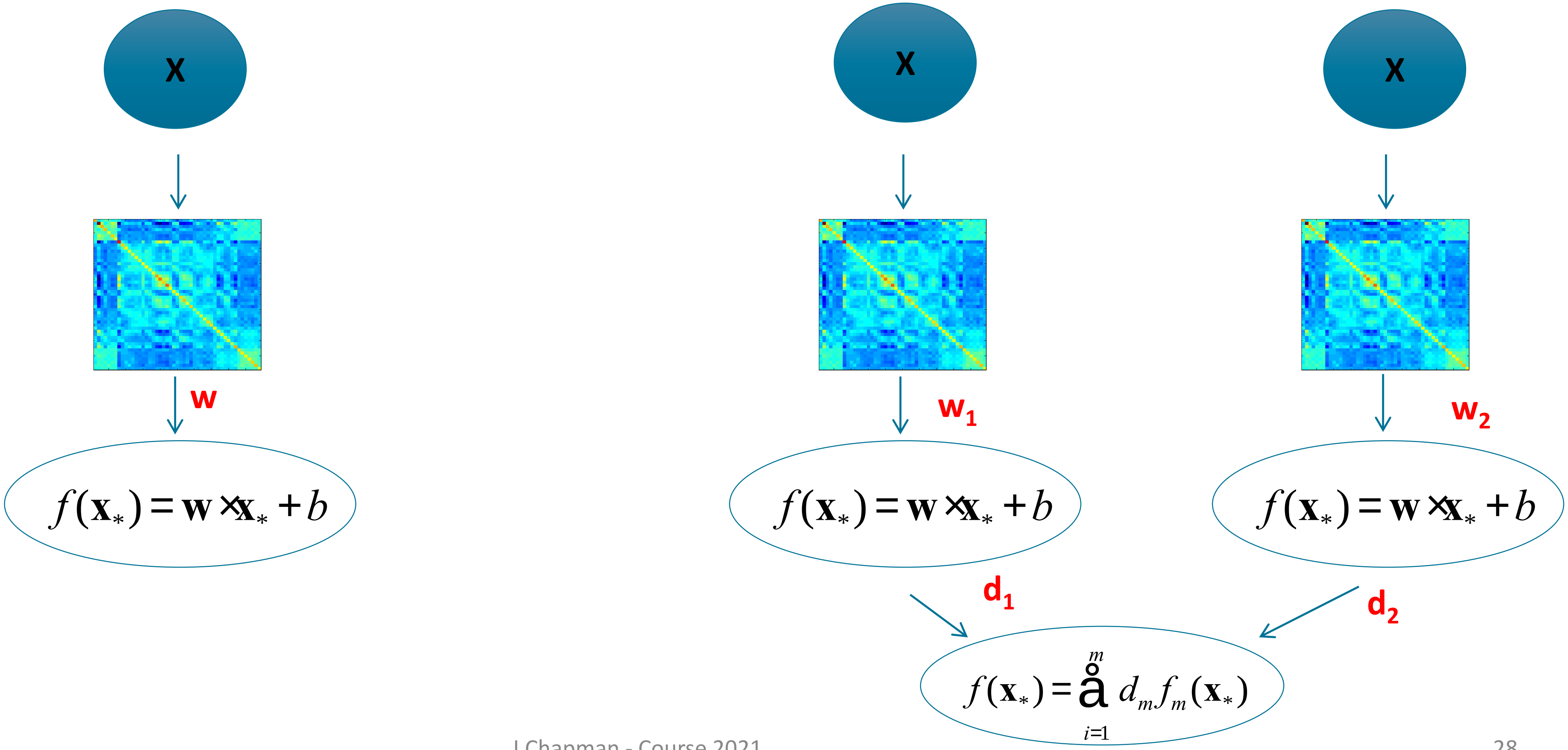


- Maximise the margin between two classes
- Solution only in terms of points on the boundary
- Can be adapted for regression

Multiple Kernel Learning (MKL)

Single kernel SVM

Multiple kernel SVM



$$f(\mathbf{x}_*) = \mathbf{w} \cdot \mathbf{x}_* + b$$

$$f(\mathbf{x}_*) = \mathbf{w} \cdot \mathbf{x}_* + b$$

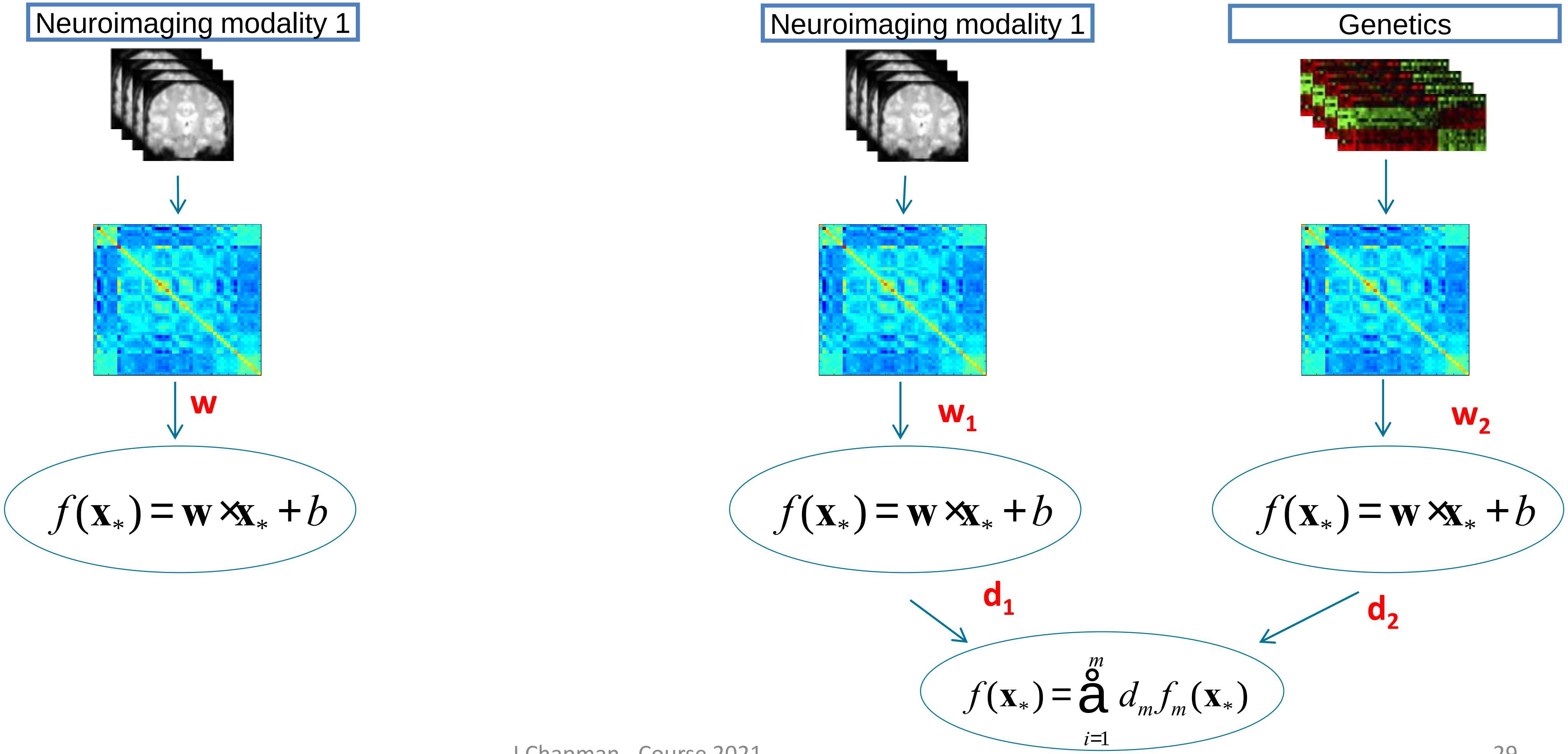
$$f(\mathbf{x}_*) = \mathbf{w} \cdot \mathbf{x}_* + b$$

$$f(\mathbf{x}_*) = \sum_{i=1}^{m} d_m f_m(\mathbf{x}_*)$$

Multiple Kernel Learning (MKL)

Single kernel SVM

Multiple kernel SVM



Neuroimaging modality 1

w

$$f(\mathbf{x}_*) = \mathbf{w} \cdot \mathbf{x}_* + b$$

Neuroimaging modality 1

Genetics

$\mathbf{w}_1$

$\mathbf{w}_2$

$$f(\mathbf{x}_*) = \mathbf{w} \cdot \mathbf{x}_* + b$$

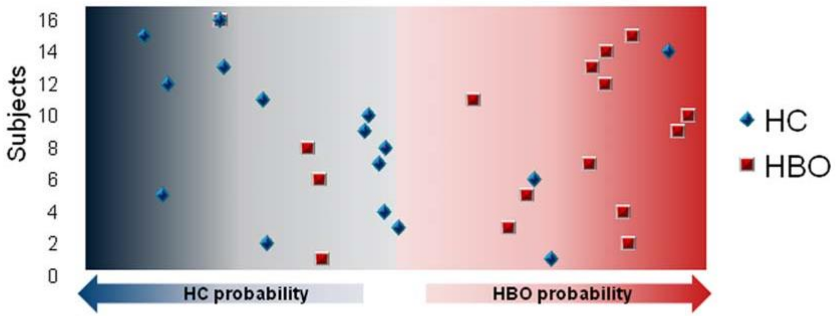
$$f(\mathbf{x}_*) = \mathbf{w} \cdot \mathbf{x}_* + b$$

$\mathbf{d}_1$

$\mathbf{d}_2$

$$f(\mathbf{x}_*) = \sum_{i=1}^{m} d_m f_m(\mathbf{x}_*)$$

Gaussian Process Classifier – Binary/Multiclass



- Provides probabilistic class predictions
- Natural extension to direct multi-class classification
- It does not find sparse solutions

Relevance Vector Machine

- A probabilistic version of SVM
- It finds sparser solutions (relevance vectors) than SVM
- For large datasets, the training times can be longer than SVM

Conclusion

- Pattern Recognition can be used to model multivariate relationships in data

- Supervised machine learning allows us to learn functions that predict outcomes out of sample

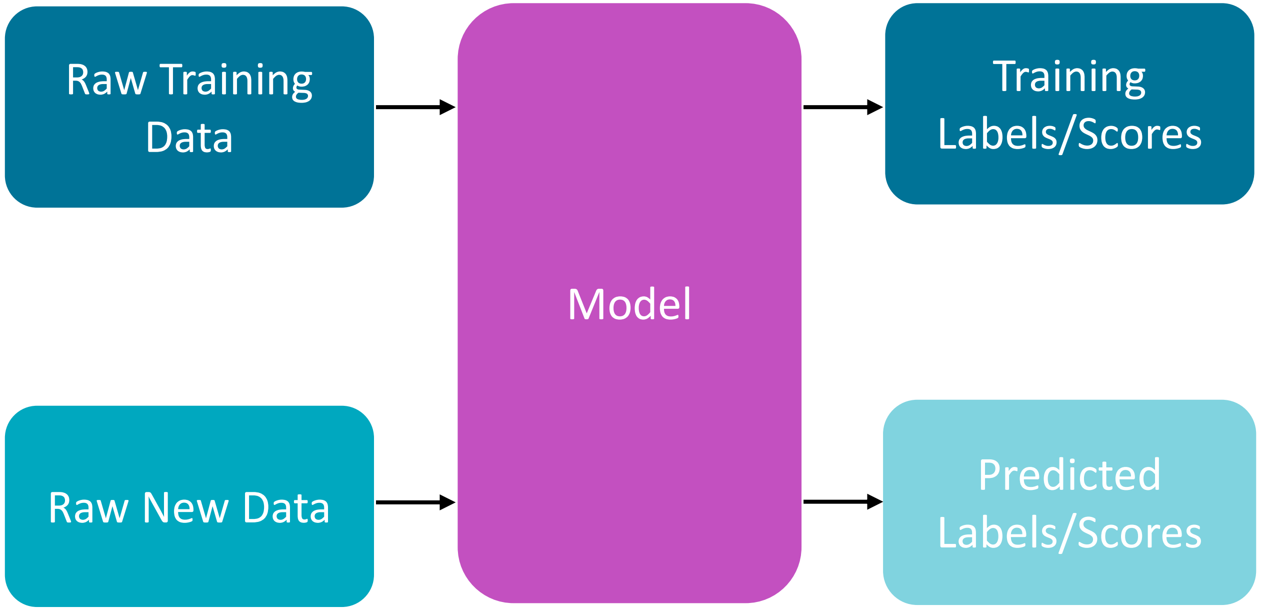
- Linear models can be used as both regression models and classifiers

- Regularisation allows us to learn these models in a data driven way, even when the data is high dimensional

Thank you!

Questions?

Deep Learning works on raw data



References

PRoNTo papers:

- Schrouff J*, Rosa MJ*, Rondina J, Marquand A, Chu C, Ashburner J, Phillips C, Richiardi J, Mourao-Miranda J. PRoNTo: Pattern Recognition for Neuroimaging Toolbox, Neuroinformatics, February 2013. *co-first authors
- Schrouff J*, Monteiro, JM*, Portugal L, Rosa MJ, Phillips C, Mourao-Miranda J. Embedding Anatomical or Functional Knowledge in Whole-Brain Multiple Kernel Learning Models Neuroinformatics, 2018

Reviews:

- Pereira, Mitchell, Botnivik (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage,*45, S199-S209
- Haynes (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, 87(2), 257-270

Books:

- Hastie , Tibishirani, Friedman (2009). Elements of Statistical Learning. *Springer*
- Bishop, Jordan, Kleinberg, Schölkopf (2006). Pattern Recognition and Machine learning. *Springer*
- Shawe-Taylor and Cristianini (2004). Kernel Methods for Pattern Analysis. *Cambridge University Press.*
- Schölkopf and Smola (2001). Learning with Kernels. *The MIT Press.*

References

Machines/Models:

- Burges (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167.
- Rasmussen, Williams (2006) Gaussian Processes for Machine Learning. *The MIT Press*
- Tipping (2001) Sparse Bayesian Learning and the Relevance Vector Machine *Journal of Machine Learning Research,* 1, 211-244
- Breiman (1996) Bagging Predictors Machine Learning, 24, 123-140
- Dietterich, Bakiri (1995) Solving multiclass learning problem via error-correcting output codes. Journal of Artificial Intelligence Research, 2: 263-286
- Rakotomamonjy, A., Bach, F., Canu, S., & Grandvalet, Y. (2008). SimpleMKL. Journal of Machine Learning Research, 9, 2491-2521
- Marquand (2010) Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. Neuroimage, 49(3), 2178-2189