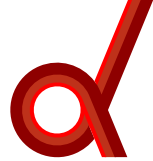
DONDERS INSTITUTE

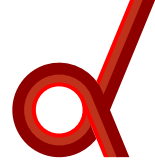
Machine Learning 2: advanced

Andre F. Marquand
a.marquand@donders.ru.nl

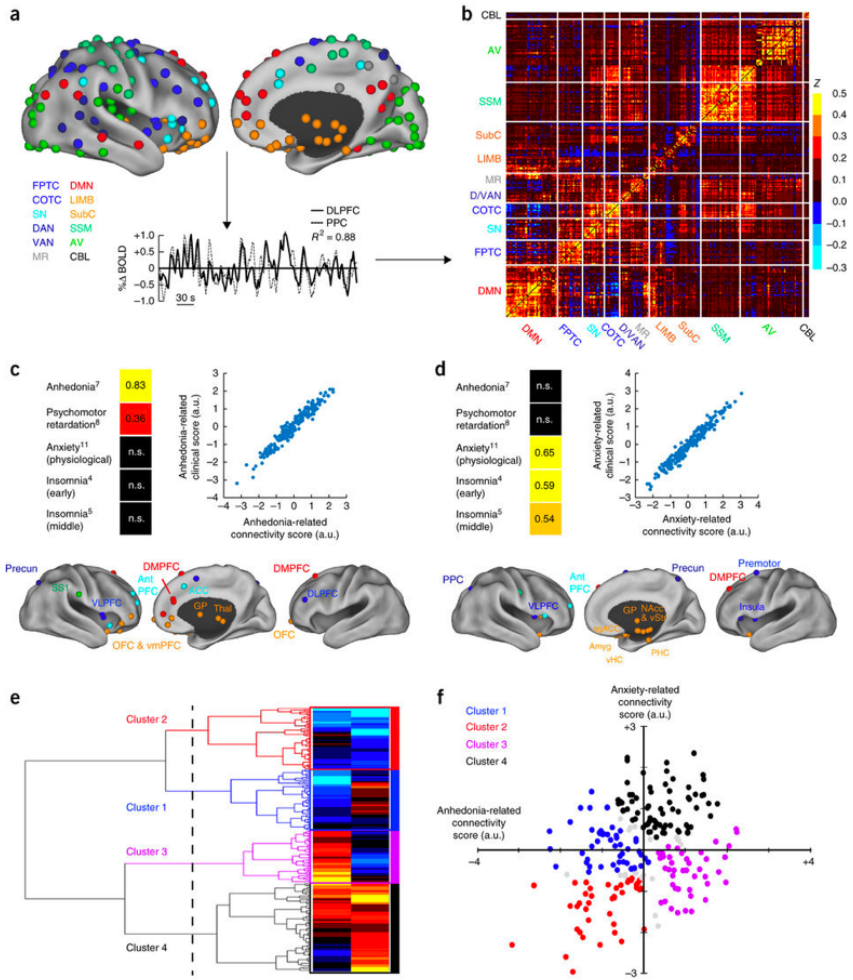
Outline



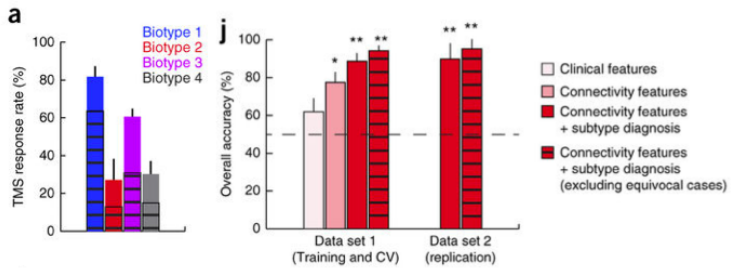
Outline



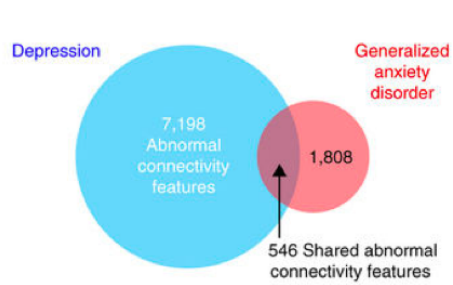
Stratification of major depression



- Extensive validation
- Predict treatment response (TMS)

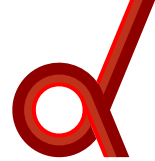


- Cut across diagnoses



Drysdale et al. (2017)

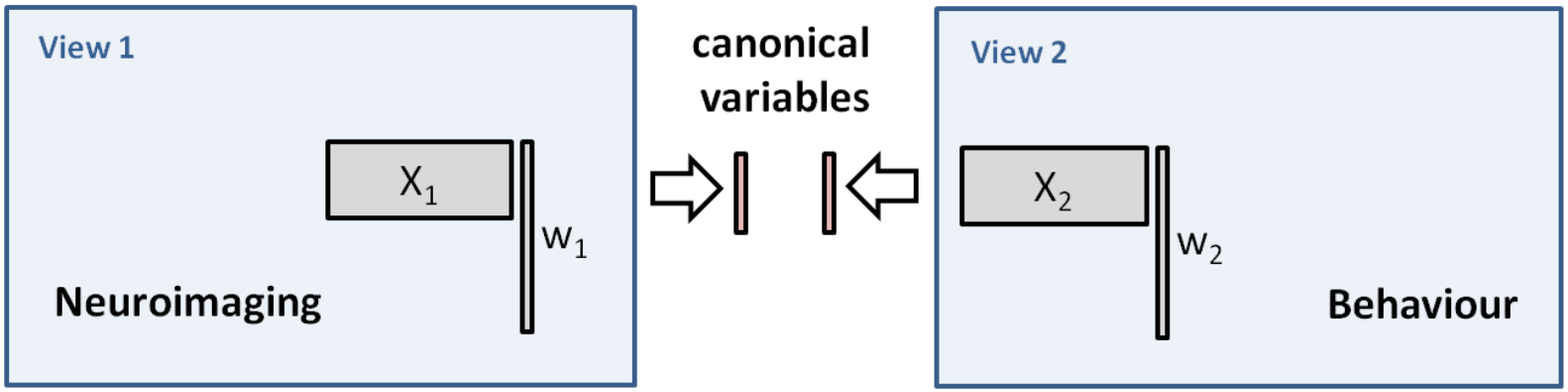
Finding mappings between brain and behaviour



- **Canonical Correlation Analysis** is a standard statistical tool for finding multivariate relationships between datasets
- Generalises Pearson correlation to multiple variables
- Finds projections of the data that maximise the correlation between "views" of the data

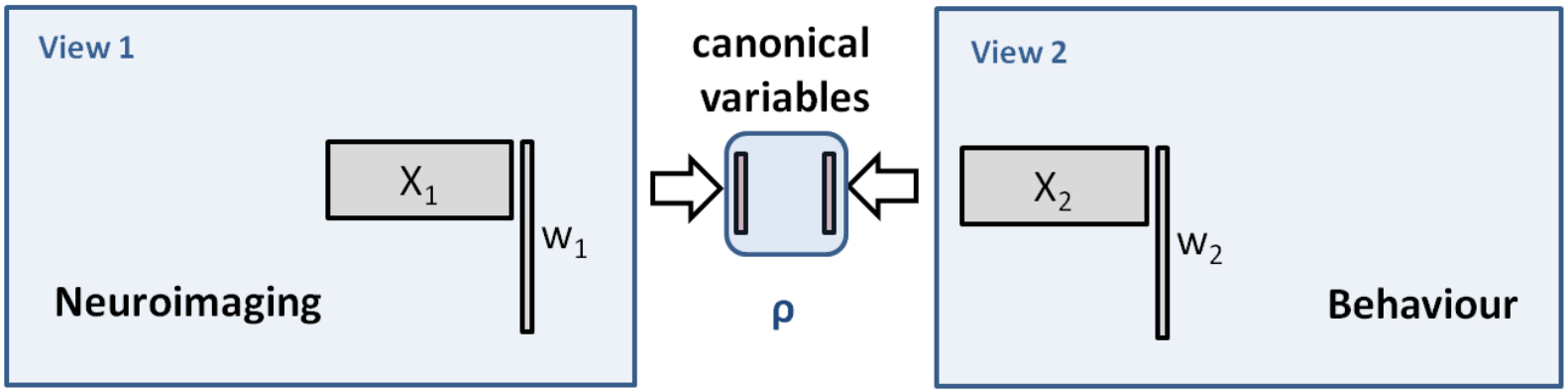
Finding mappings between brain and behaviour

- **Canonical Correlation Analysis** is a standard statistical tool for finding multivariate relationships between datasets
- Generalises Pearson correlation to multiple variables
- Finds projections of the data that maximise the correlation between "views" of the data



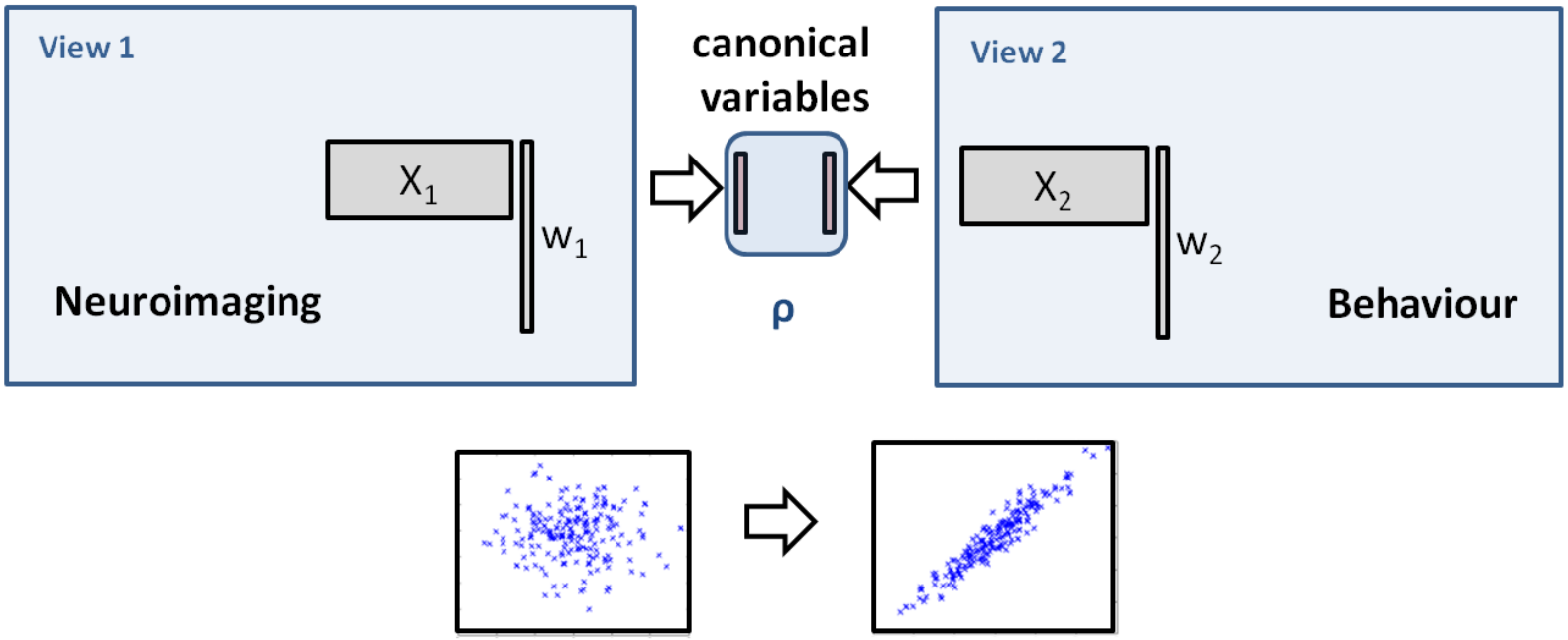
Finding mappings between brain and behaviour

- **Canonical Correlation Analysis** is a standard statistical tool for finding multivariate relationships between datasets
- Generalises Pearson correlation to multiple variables
- Finds projections of the data that maximise the correlation between "views" of the data



Finding mappings between brain and behaviour

- **Canonical Correlation Analysis** is a standard statistical tool for finding multivariate relationships between datasets
- Generalises Pearson correlation to multiple variables
- Finds projections of the data that maximise the correlation between "views" of the data



Canonical Correlation Analysis

- CCA is related to techniques such as partial least squares
- Formally, CCA solves the following objective function:

$$\max_{\mathbf{w}_1,\mathbf{w}_2} corr(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2)$$

$$\text{subject to } ||\mathbf{w}_1^T\mathbf{X}_1^T\mathbf{X}_1\mathbf{w}_1|| \leq 1 \text{ and } ||\mathbf{w}_2^T\mathbf{X}_2^T\mathbf{X}_2\mathbf{w}_2|| \leq 1$$

where the constraint is sometimes amended to:

$$\text{subject to } ||\mathbf{w}_1||^2 \leq 1 \text{ and } ||\mathbf{w}_2||^2 leq1$$

Canonical Correlation Analysis

- CCA is related to techniques such as partial least squares
- Formally, CCA solves the following objective function:

$$\max_{\mathbf{w}_1,\mathbf{w}_2} corr(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2)$$

$$\text{subject to } ||\mathbf{w}_1^T\mathbf{X}_1^T\mathbf{X}_1\mathbf{w}_1|| \leq 1 \text{ and } ||\mathbf{w}_2^T\mathbf{X}_2^T\mathbf{X}_2\mathbf{w}_2|| \leq 1$$

where the constraint is sometimes amended to:

$$\text{subject to } ||\mathbf{w}_1||^2 \leq 1 \text{ and } ||\mathbf{w}_2||^2 leq1$$

...and other constraints can be added (e.g. to promote sparsity)

$$P(\mathbf{w}_1) < c_1 \text{ and } P(\mathbf{w}_2) < c_2$$

Canonical Correlation Analysis

- CCA is related to techniques such as partial least squares
- Formally, CCA solves the following objective function:

$$\max_{\mathbf{w}_1,\mathbf{w}_2} corr(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2)$$

$$\text{subject to } ||\mathbf{w}_1^T\mathbf{X}_1^T\mathbf{X}_1\mathbf{w}_1|| \leq 1 \text{ and } ||\mathbf{w}_2^T\mathbf{X}_2^T\mathbf{X}_2\mathbf{w}_2|| \leq 1$$

where the constraint is sometimes amended to:

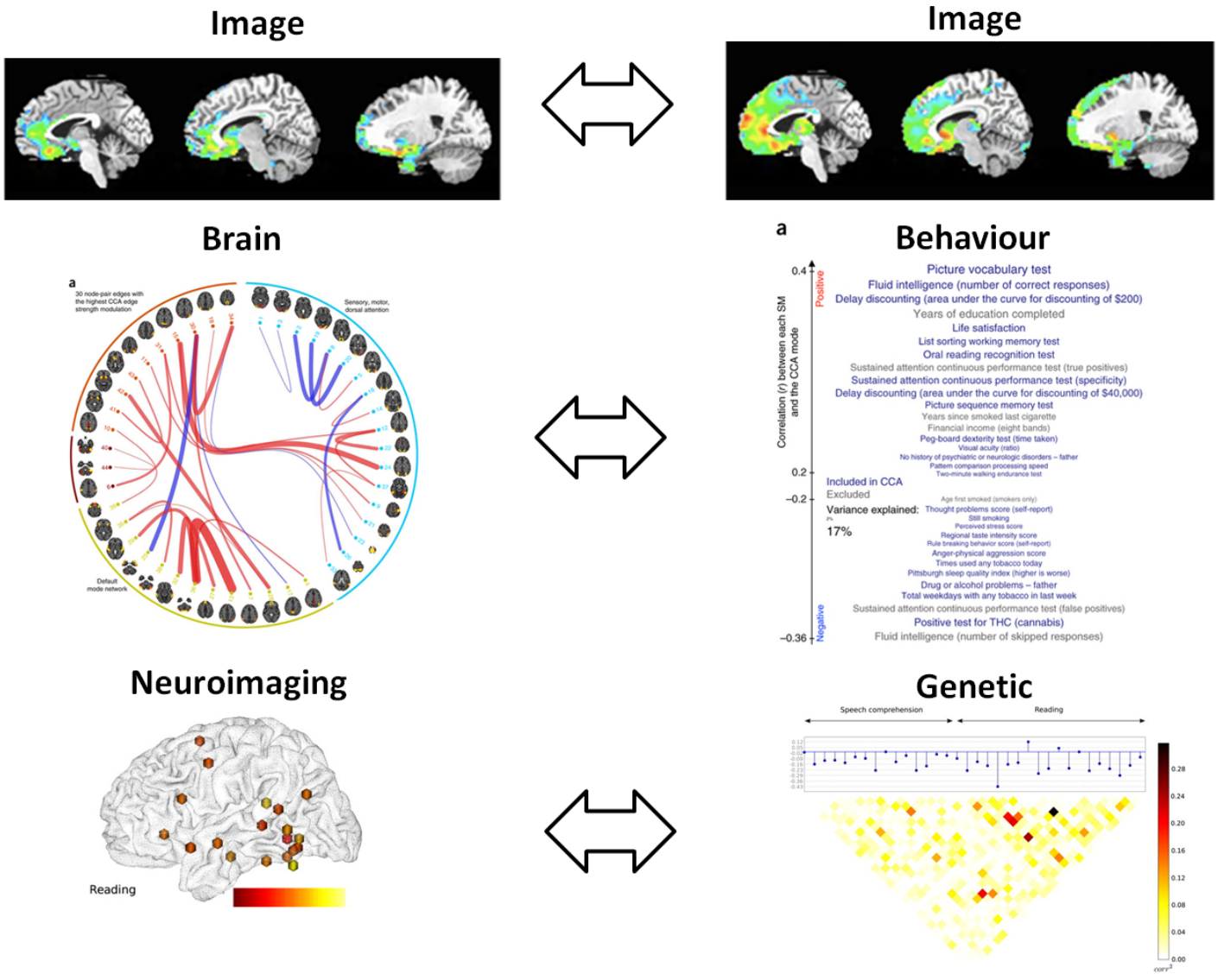
$$\text{subject to } ||\mathbf{w}_1||^2 \leq 1 \text{ and } ||\mathbf{w}_2||^2 leq1$$

...and other constraints can be added (e.g. to promote sparsity)

$$P(\mathbf{w}_1) < c_1 \text{ and } P(\mathbf{w}_2) < c_2$$

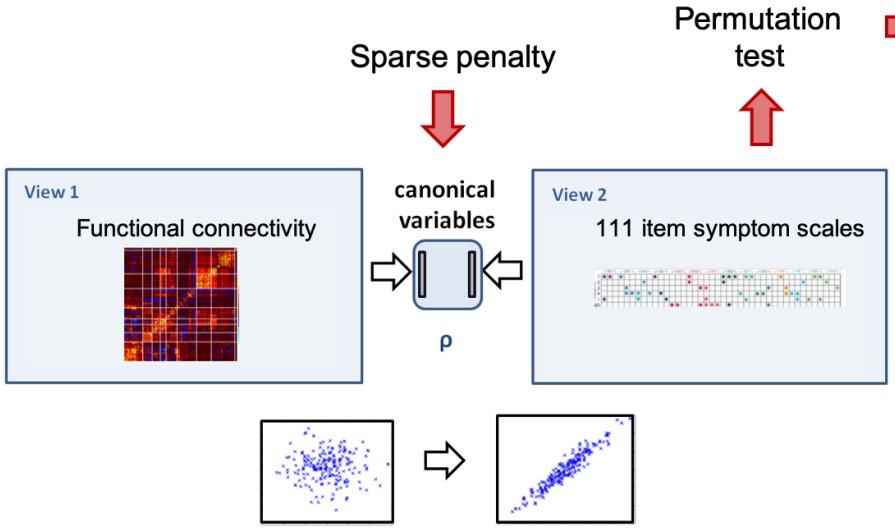
- if $n > p_1$ and p_2, an analytical solution is available
- There are many variants (kernel CCA, Bayesian CCA, deep CCA...)

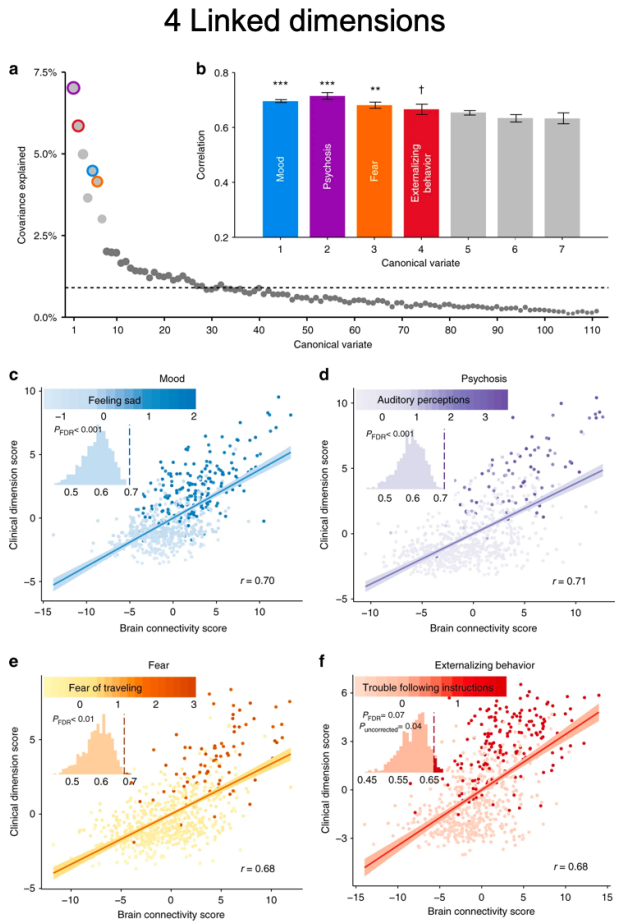
Applications of CCA



Smith et al. (2015); Floch et al. (2012)

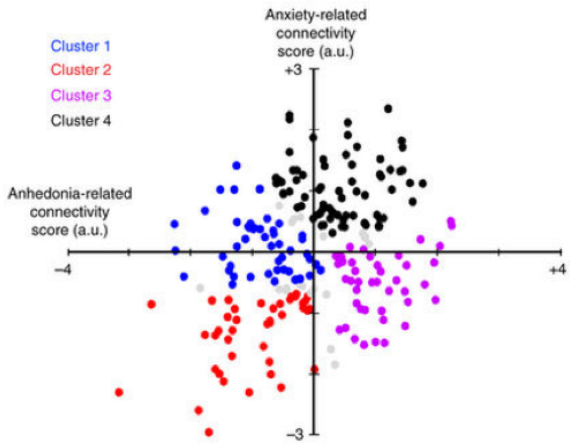
Applications of CCA





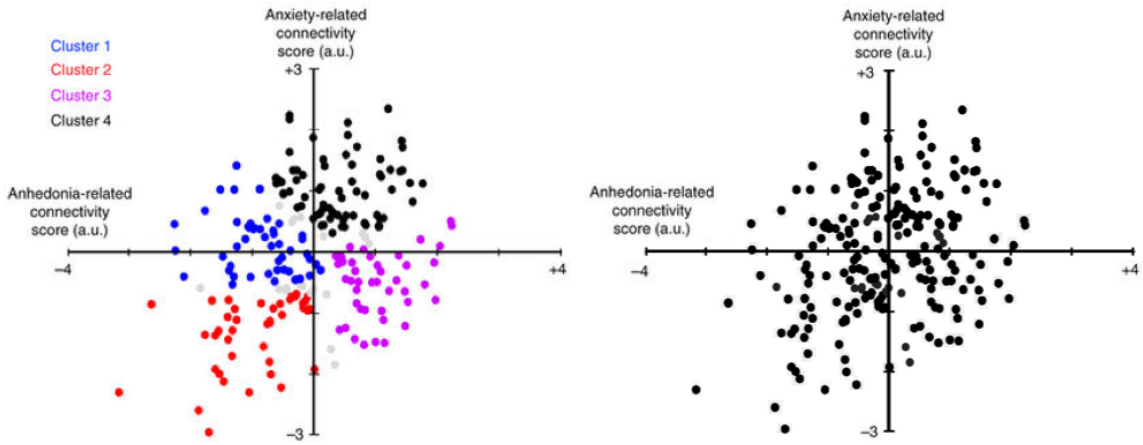
Xia et al. (2018)

Stratification of major depression, revisited



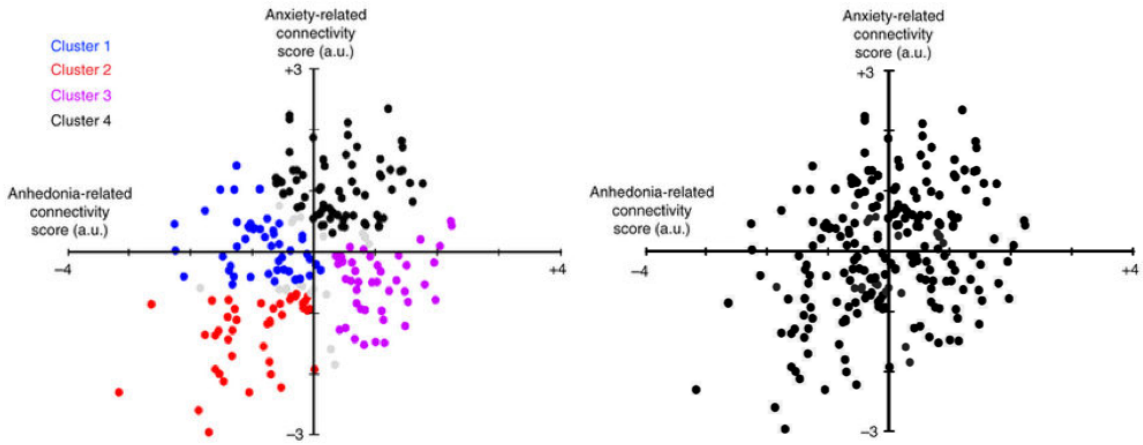
Dinga et al. (2019)

Stratification of major depression, revisited

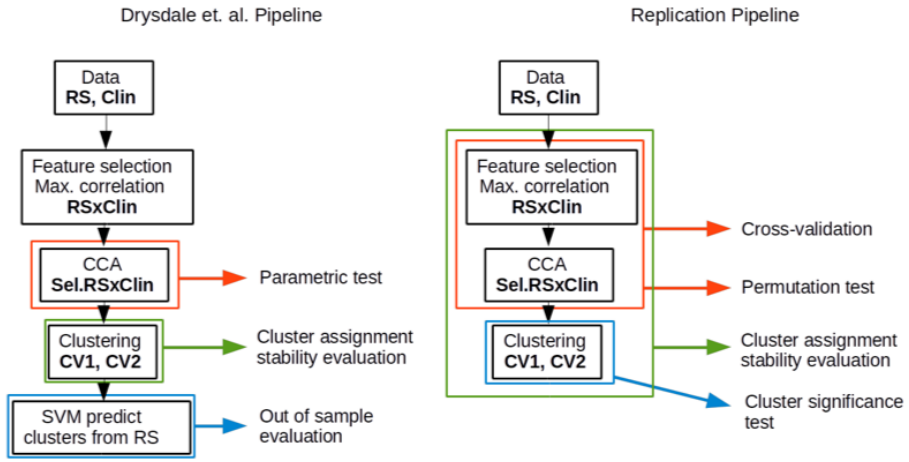


Dinga et al. (2019)

Stratification of major depression, revisited

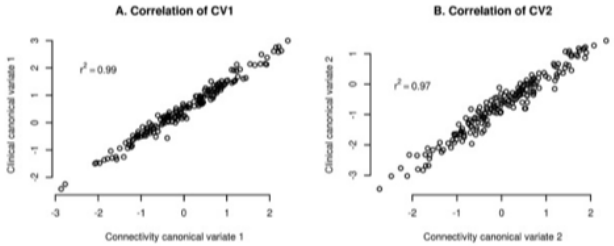


- All tests are in-sample
- No regularization
- Statistical tests are over-optimistic



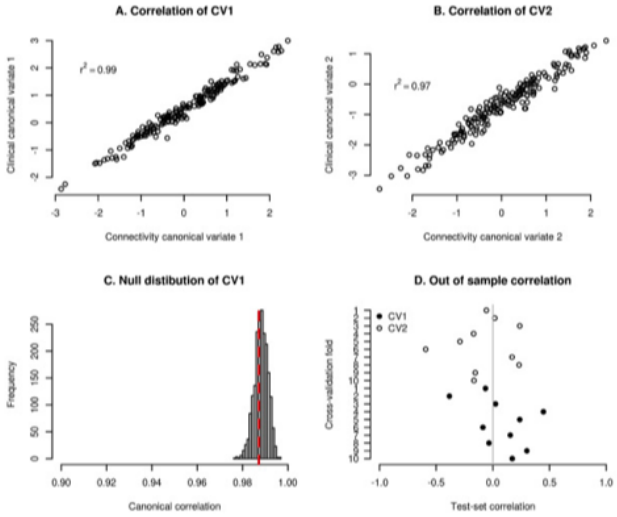
Dinga et al. (2019)

Stratification of major depression



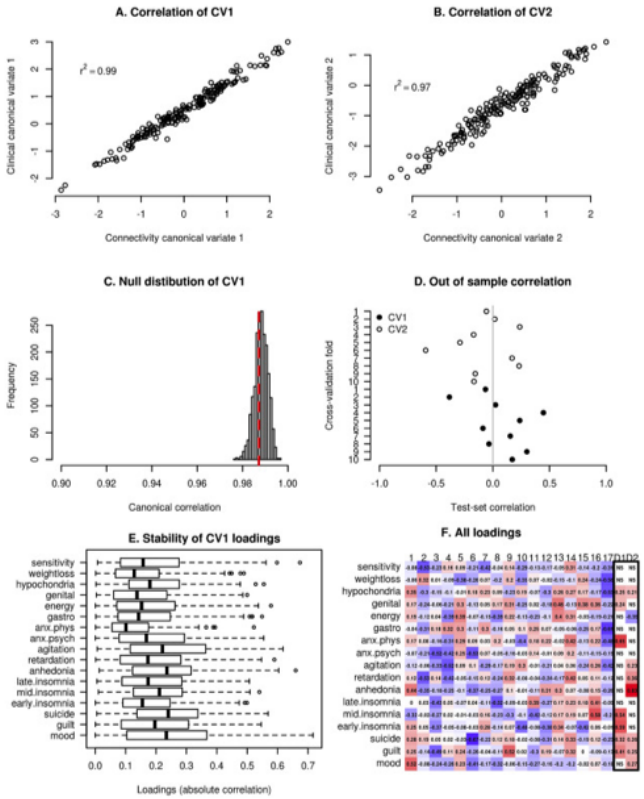
Dinga et al. (2019)

Stratification of major depression



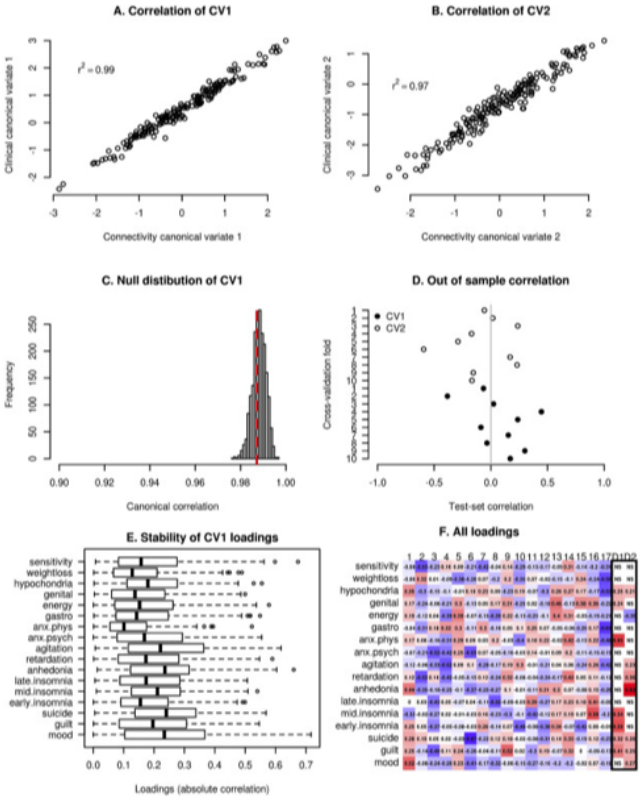
Dinga et al. (2019)

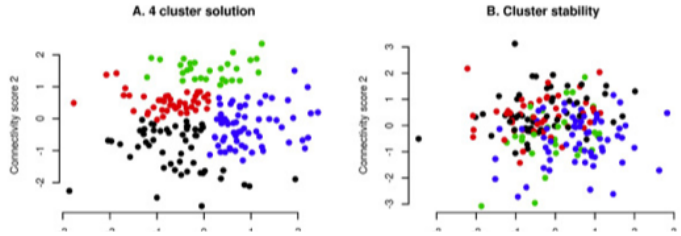
Stratification of major depression



Dinga et al. (2019)

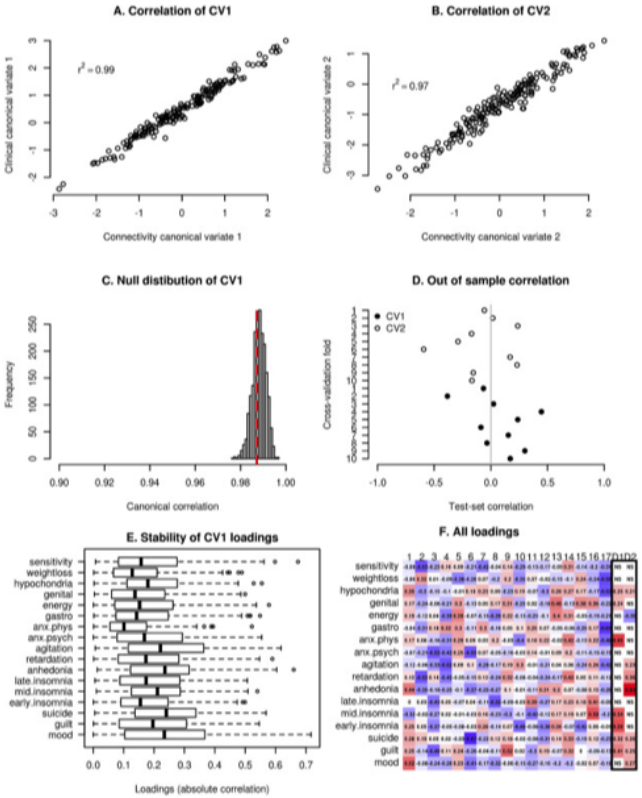
Stratification of major depression

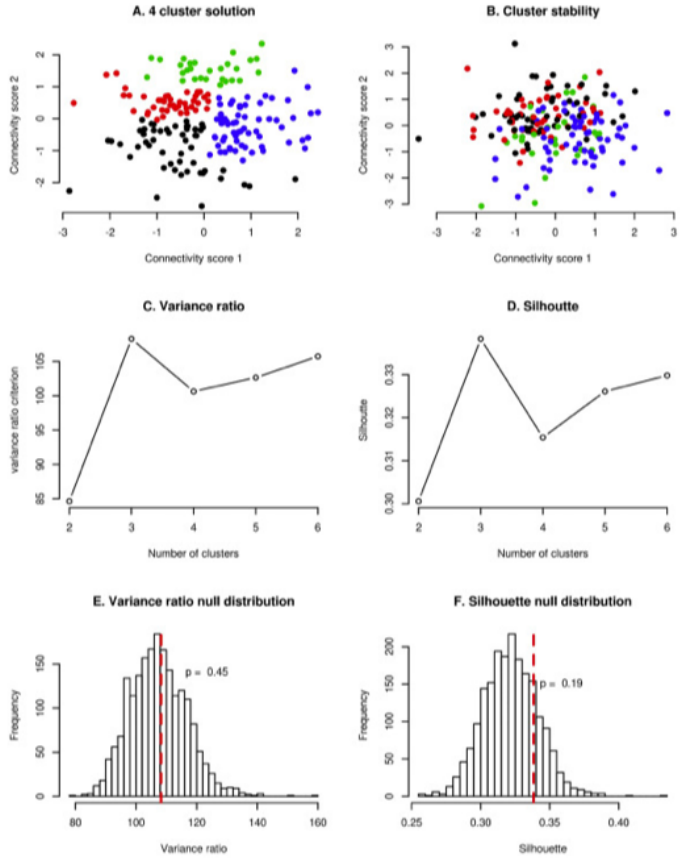




Dinga et al. (2019)

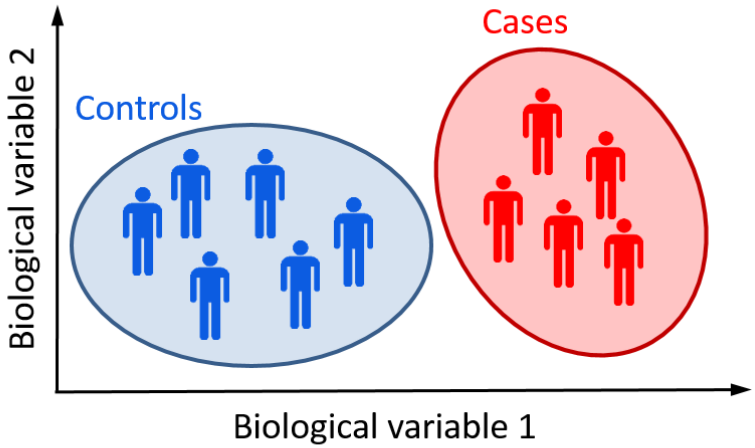
Stratification of major depression





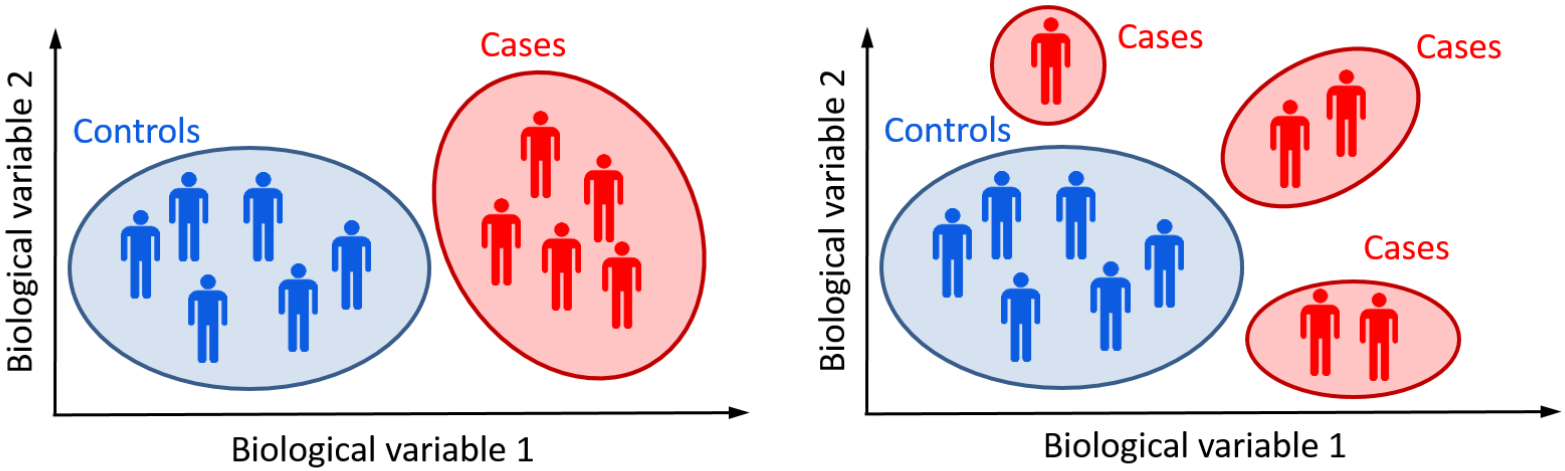
Dinga et al. (2019)

Many types of heterogeneity



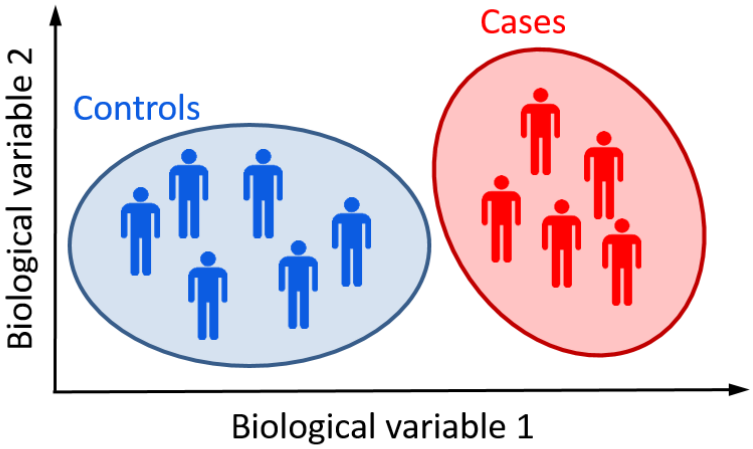
Marquand et al. (2016)

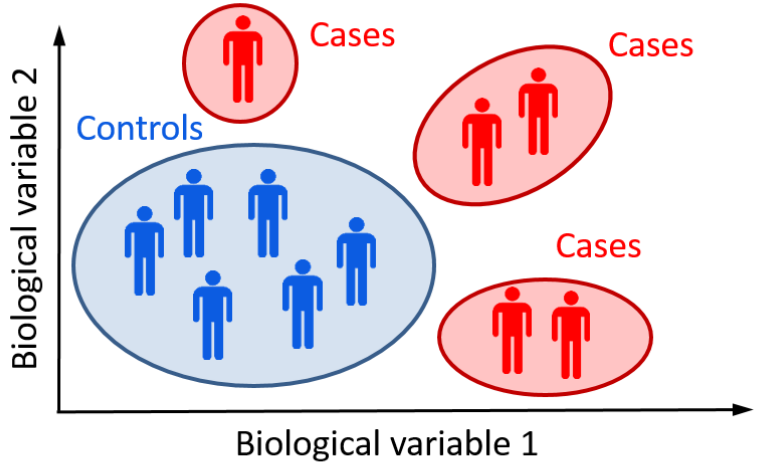
Many types of heterogeneity

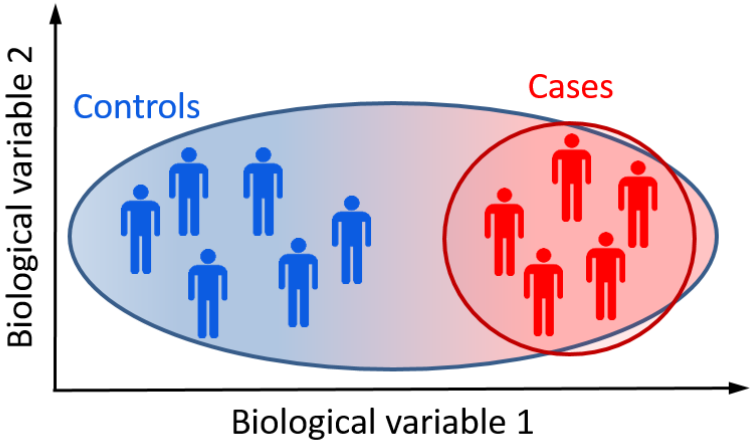


Marquand et al. (2016)

Many types of heterogeneity

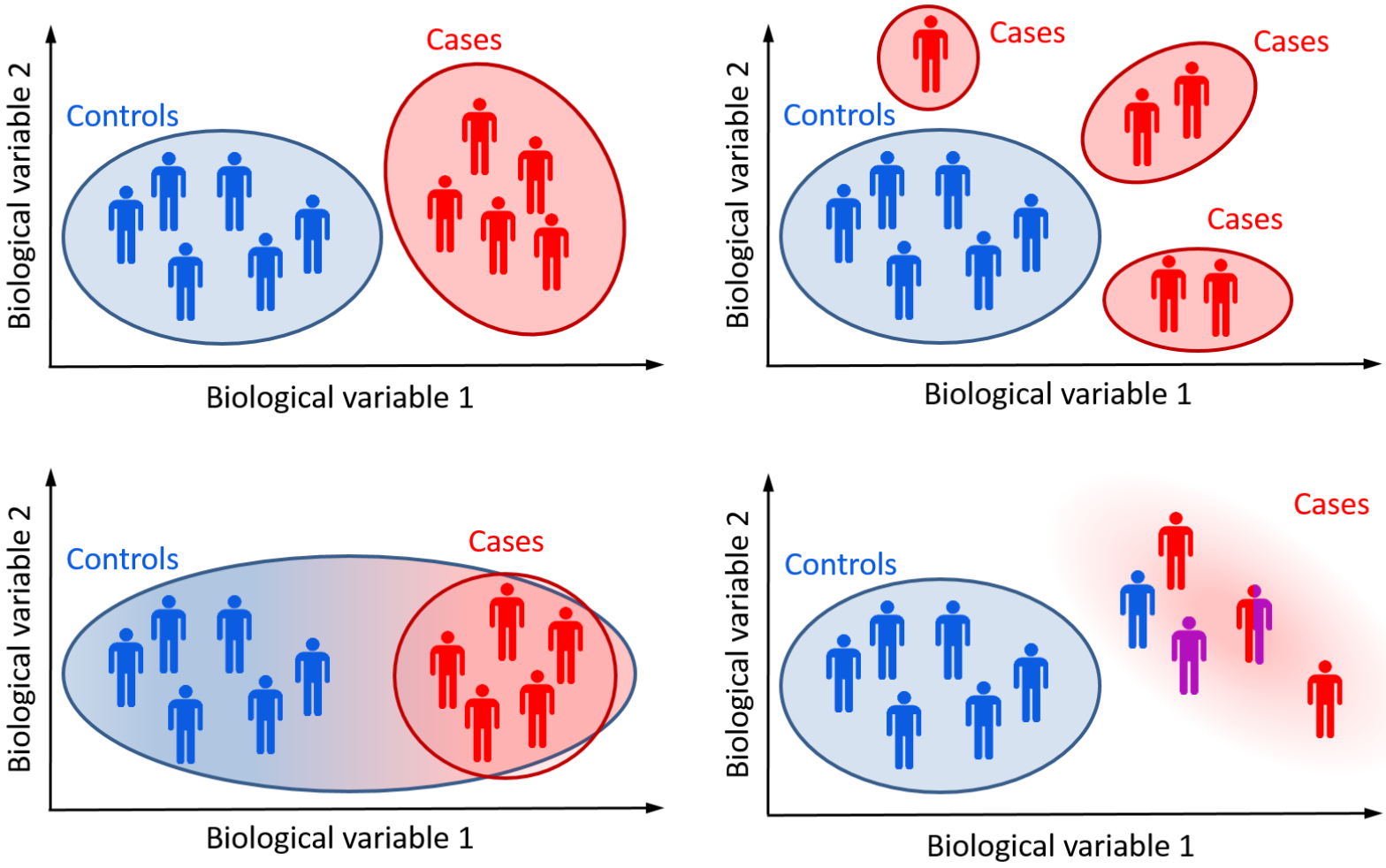






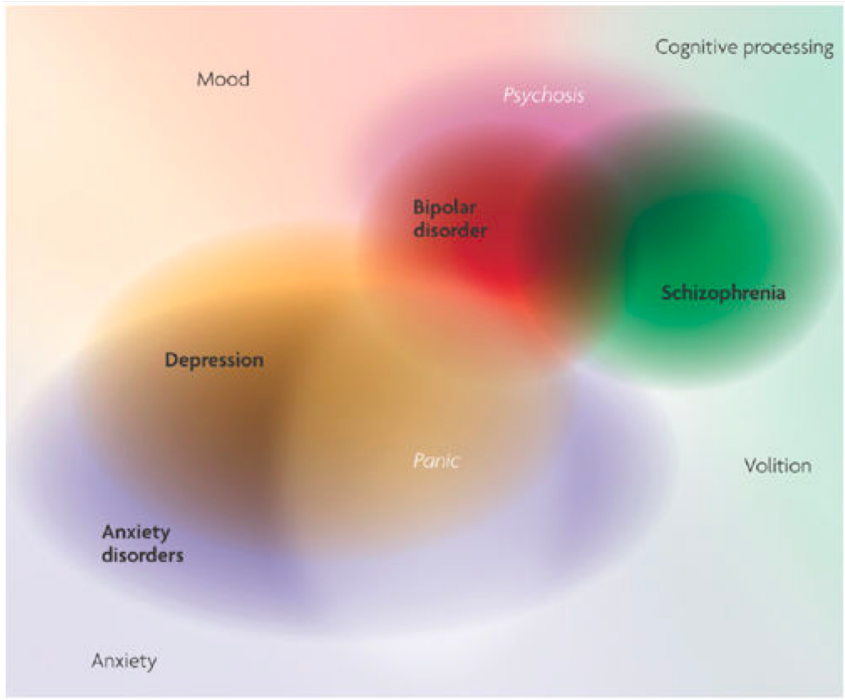
Marquand et al. (2016)

Many types of heterogeneity



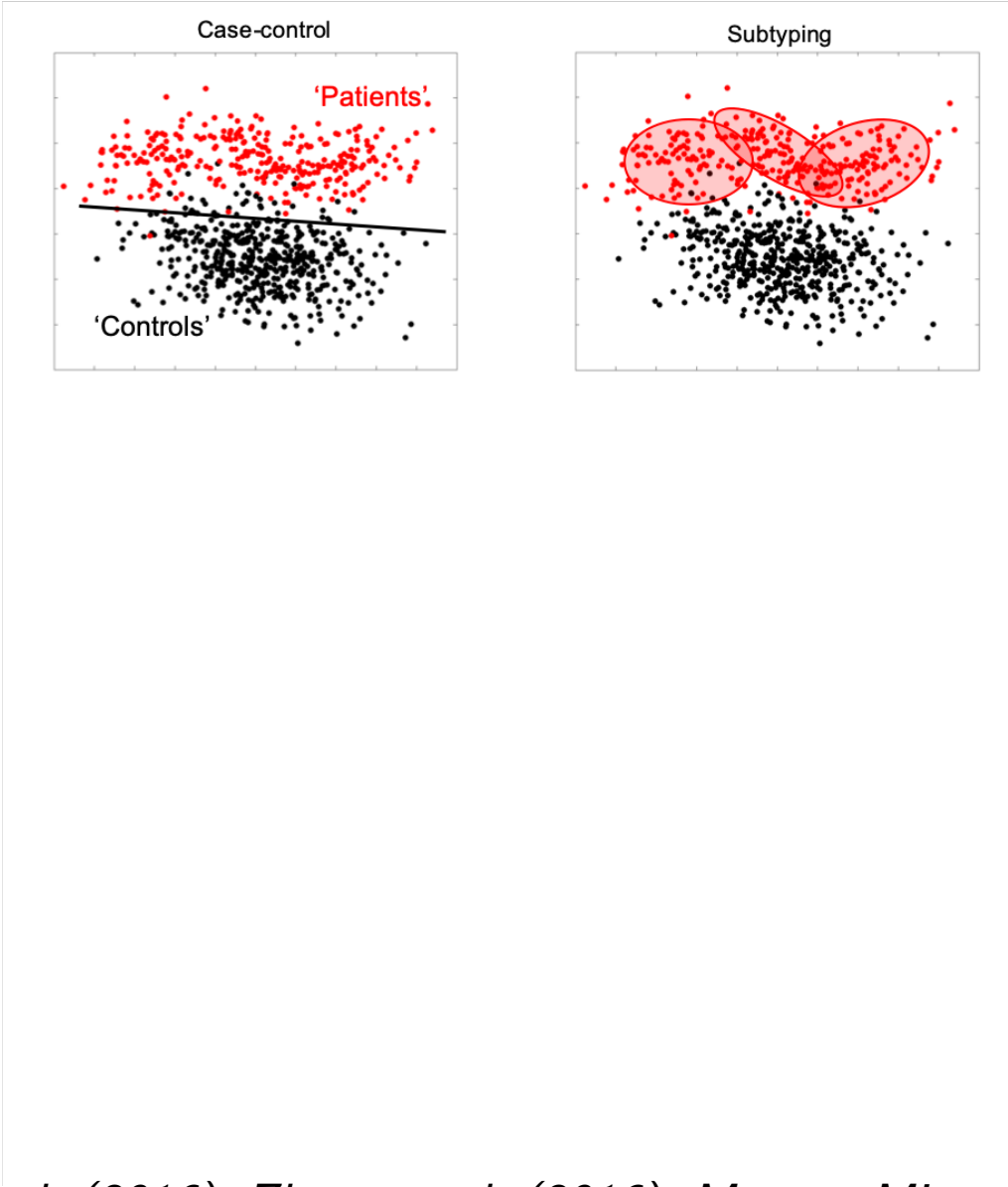
Marquand et al. (2016)

Many types of heterogeneity



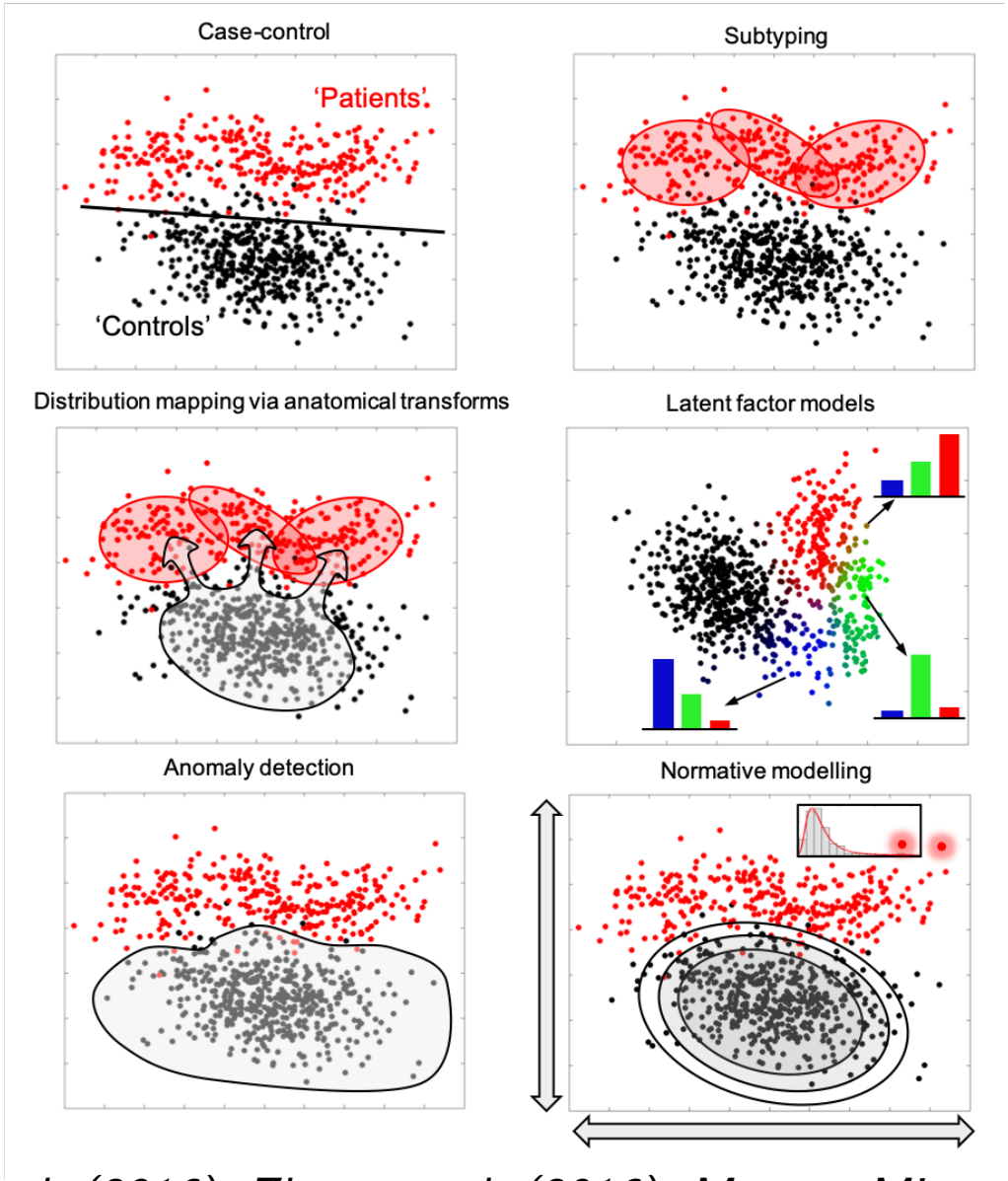
Burmeister et al. (2008)

Methods for addressing heterogeneity



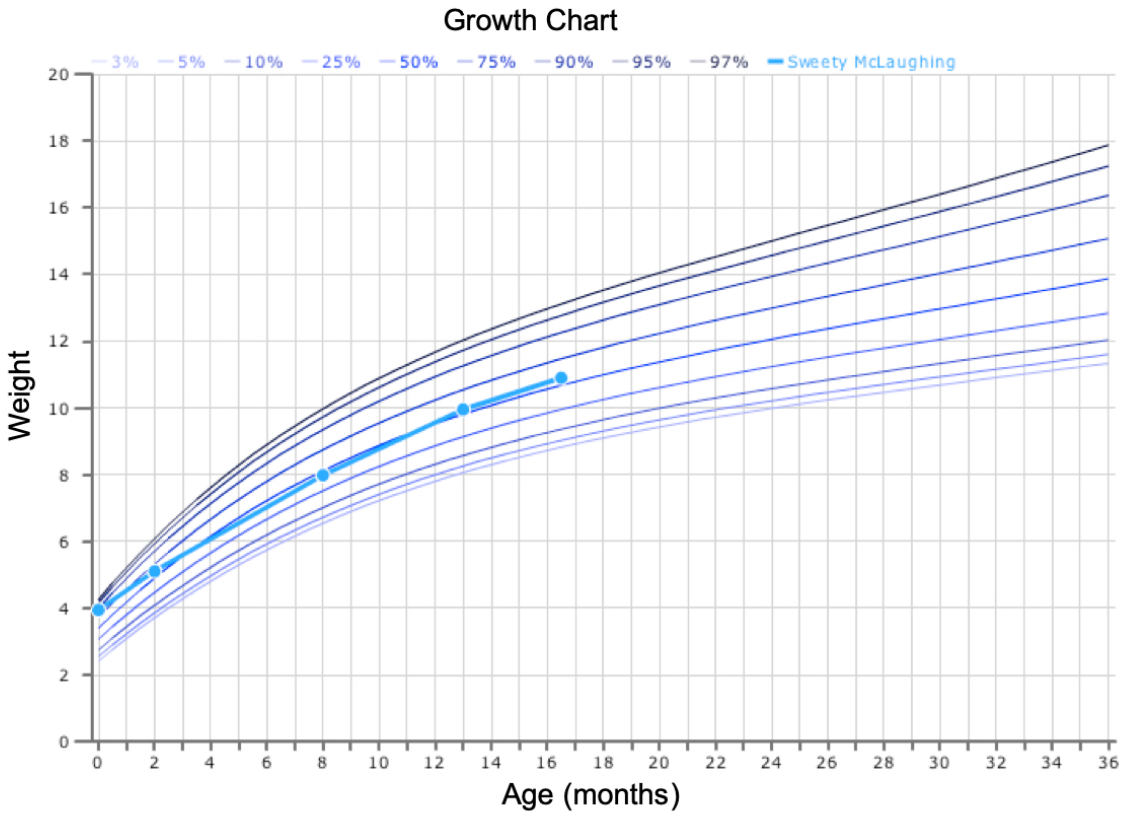
Dong et al. (2016); Zhang et al. (2016); Mourao-Miranda et al. (2011)

Methods for addressing heterogeneity



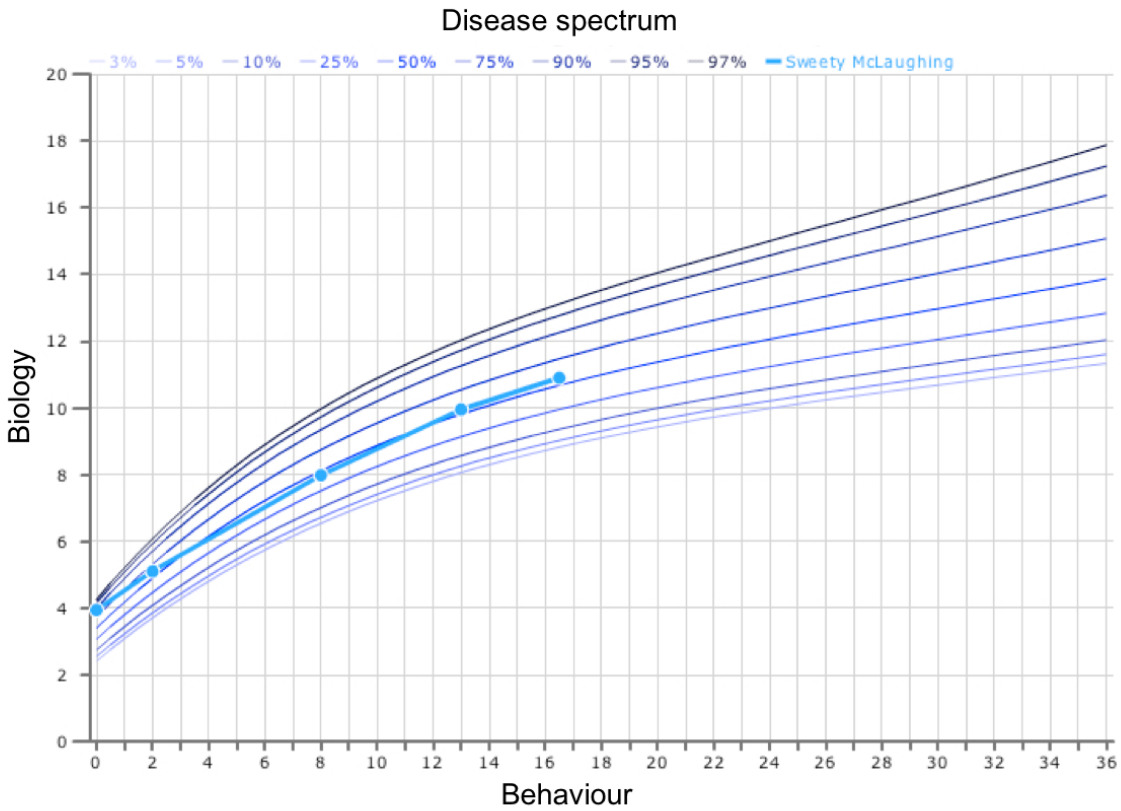
Dong et al. (2016); Zhang et al. (2016); Mourao-Miranda et al. (2011)

Normative modelling



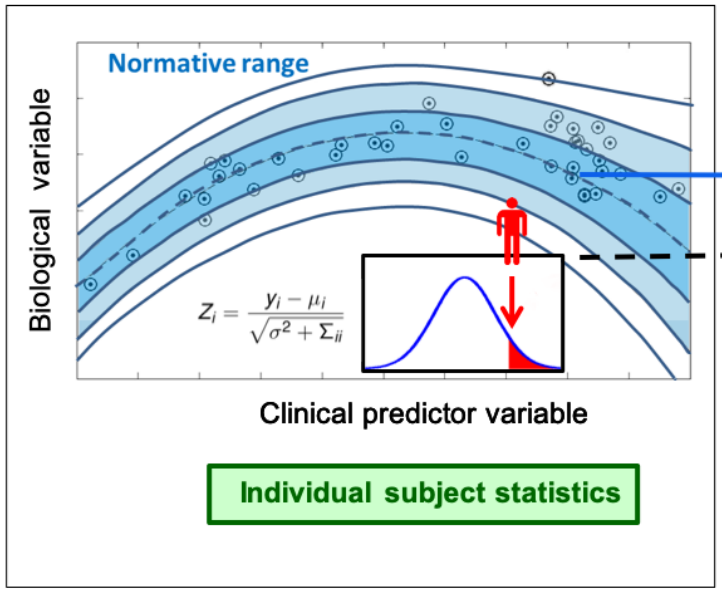
Marquand et al. (2016)

Normative modelling



Marquand et al. (2016)

Normative modelling



Uncertainty

Error

Gaussian process regression

$$p(y|\mathbf{x}, \sigma^2) = f(\mathbf{x}, \theta) + \epsilon$$

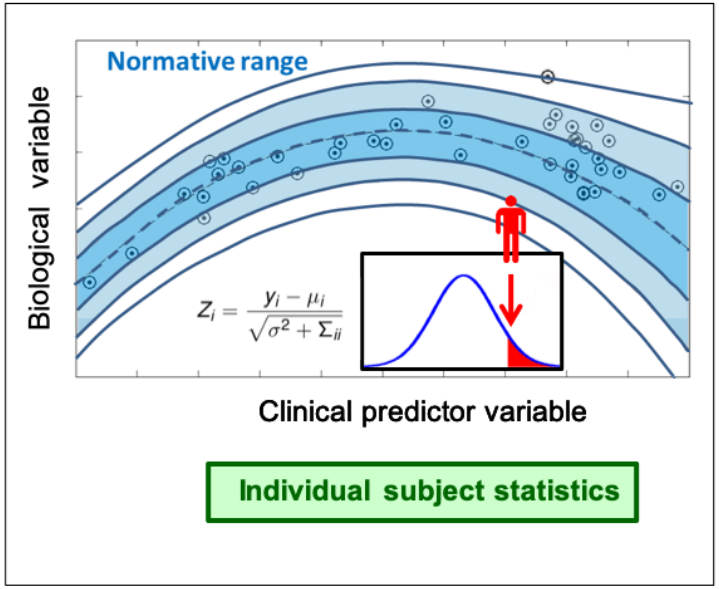
$$p(\epsilon) = \mathcal{N}(0, \sigma^2)$$

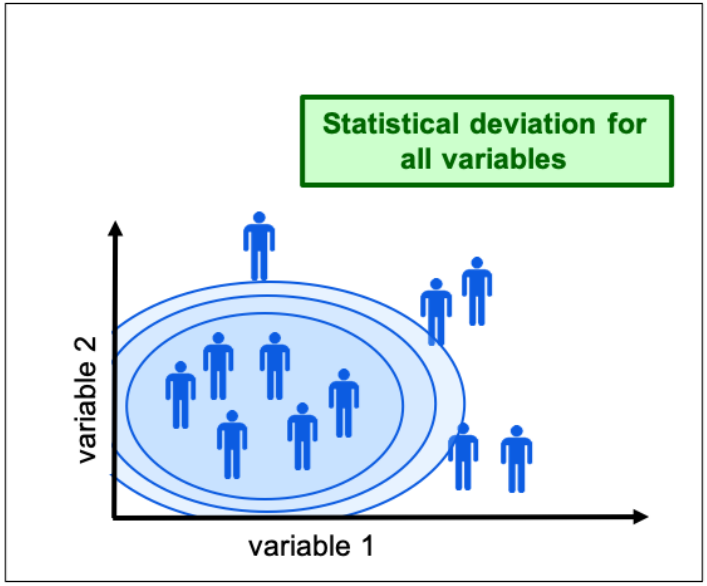
$$p(\mathbf{f}|\theta) = \mathcal{N}(m(\beta), \mathbf{K}(\theta))$$

$$p\left(\begin{bmatrix}\mathbf{y}\\ \mathbf{f}_*\end{bmatrix}\middle|\theta, \beta, \sigma^2\right) \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{m}\\ \mathbf{m}_*\end{bmatrix}, \begin{bmatrix}\mathbf{K}+\sigma^2\mathbf{I} & \mathbf{k}_*{}^T\\ \mathbf{k}_* & \mathbf{k}_{**}\end{bmatrix}\right)$$

Marquand et al. (2016)

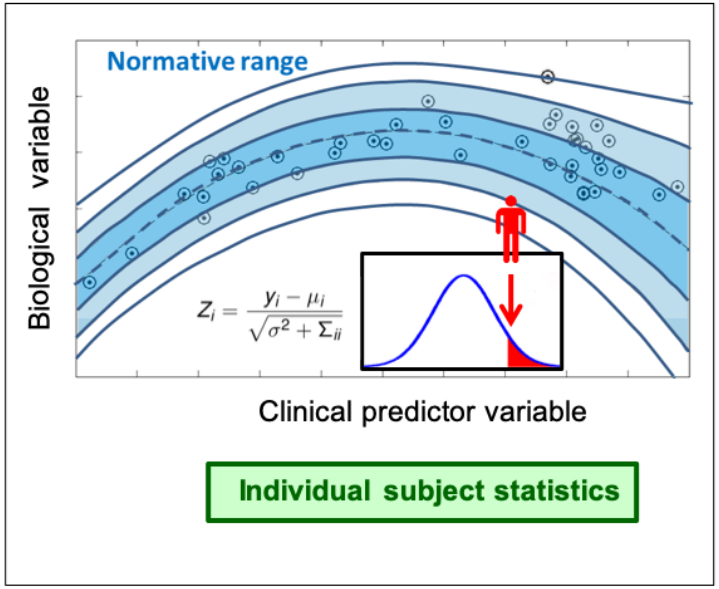
Normative modelling

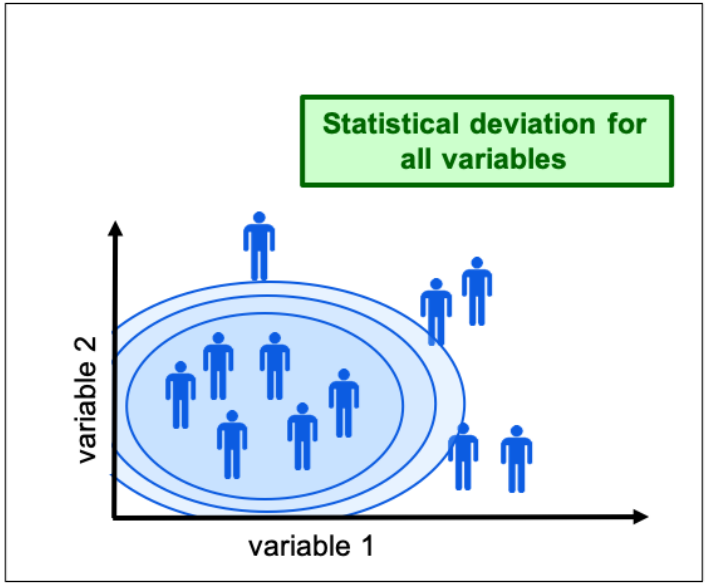


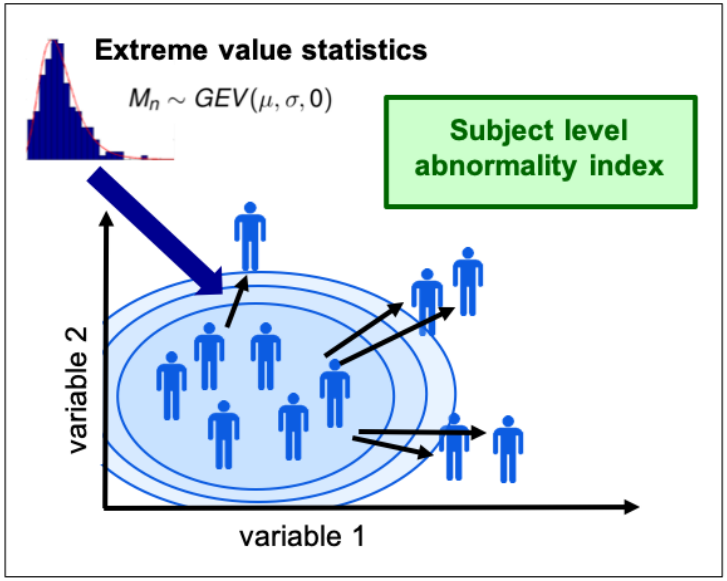


Marquand et al. (2016)

Normative modelling

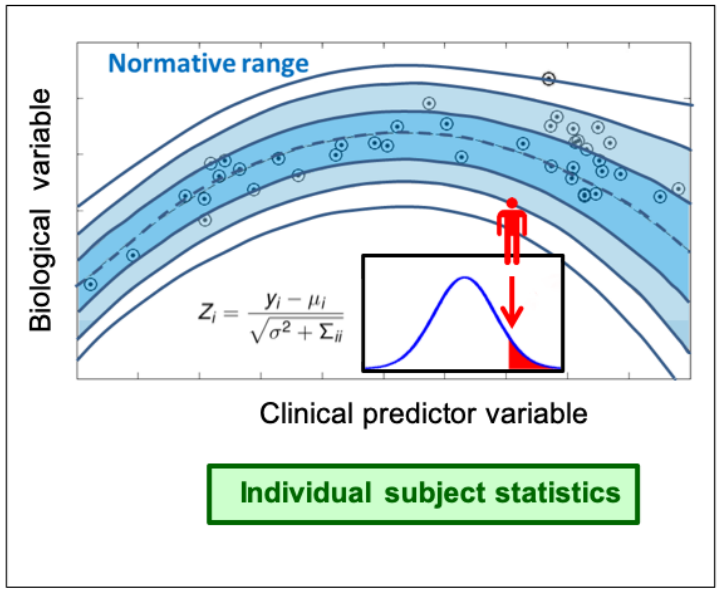


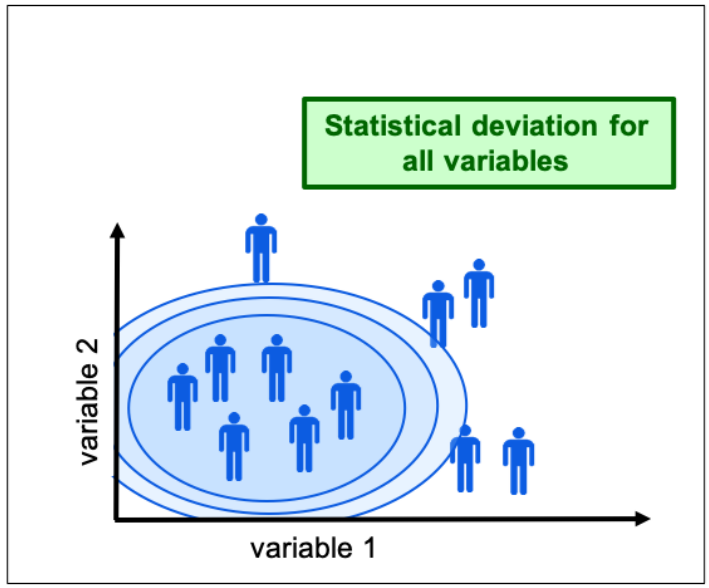


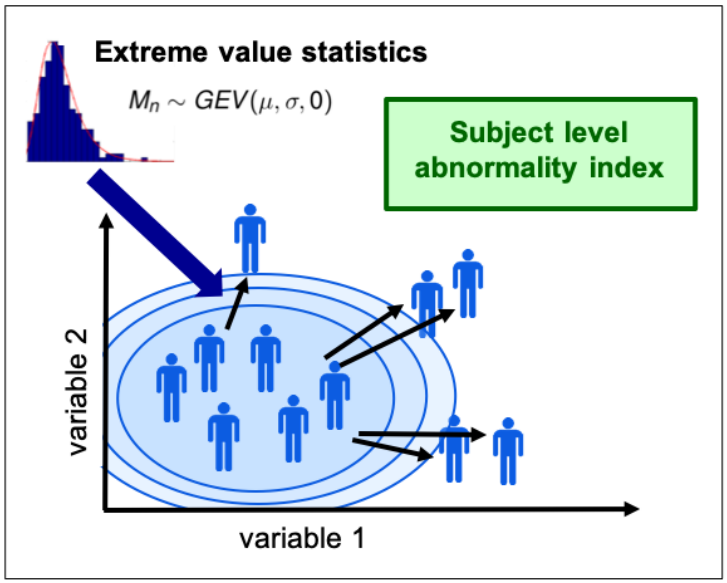


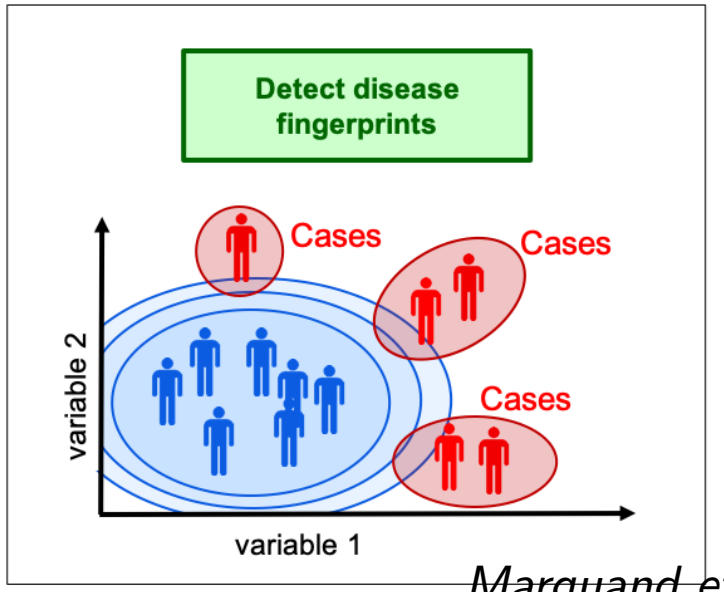
Marquand et al. (2016)

Normative modelling



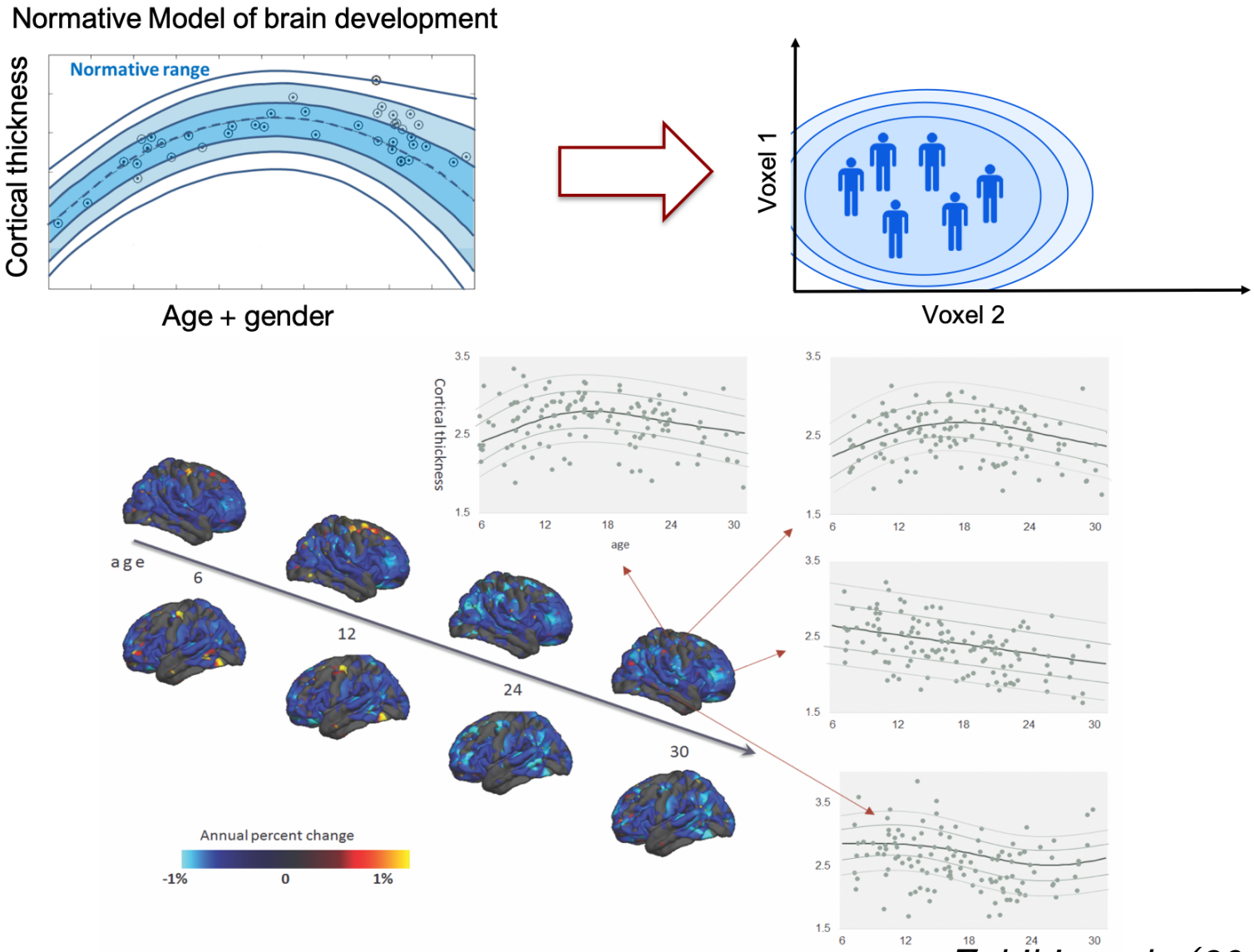






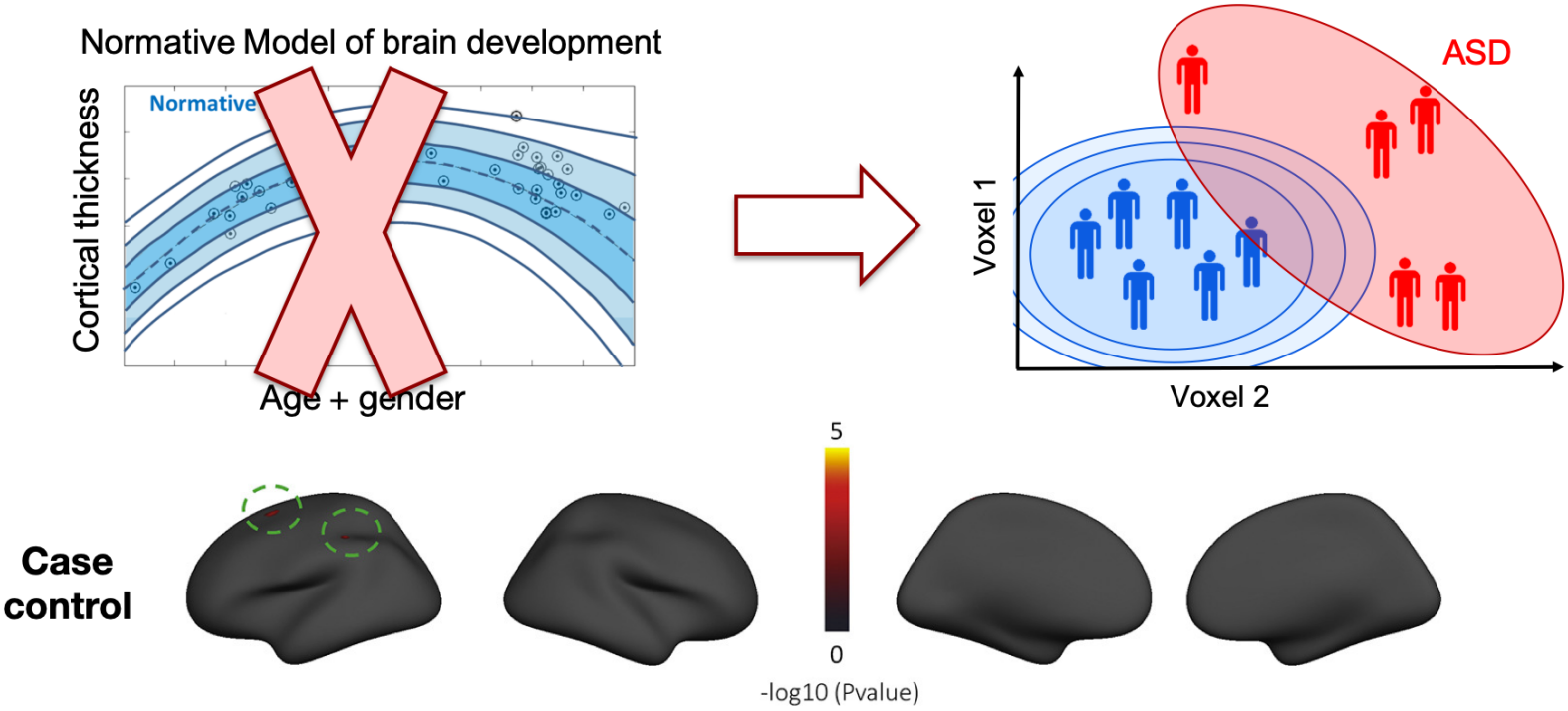
Marquand et al. (2016)

Normative modelling of autism



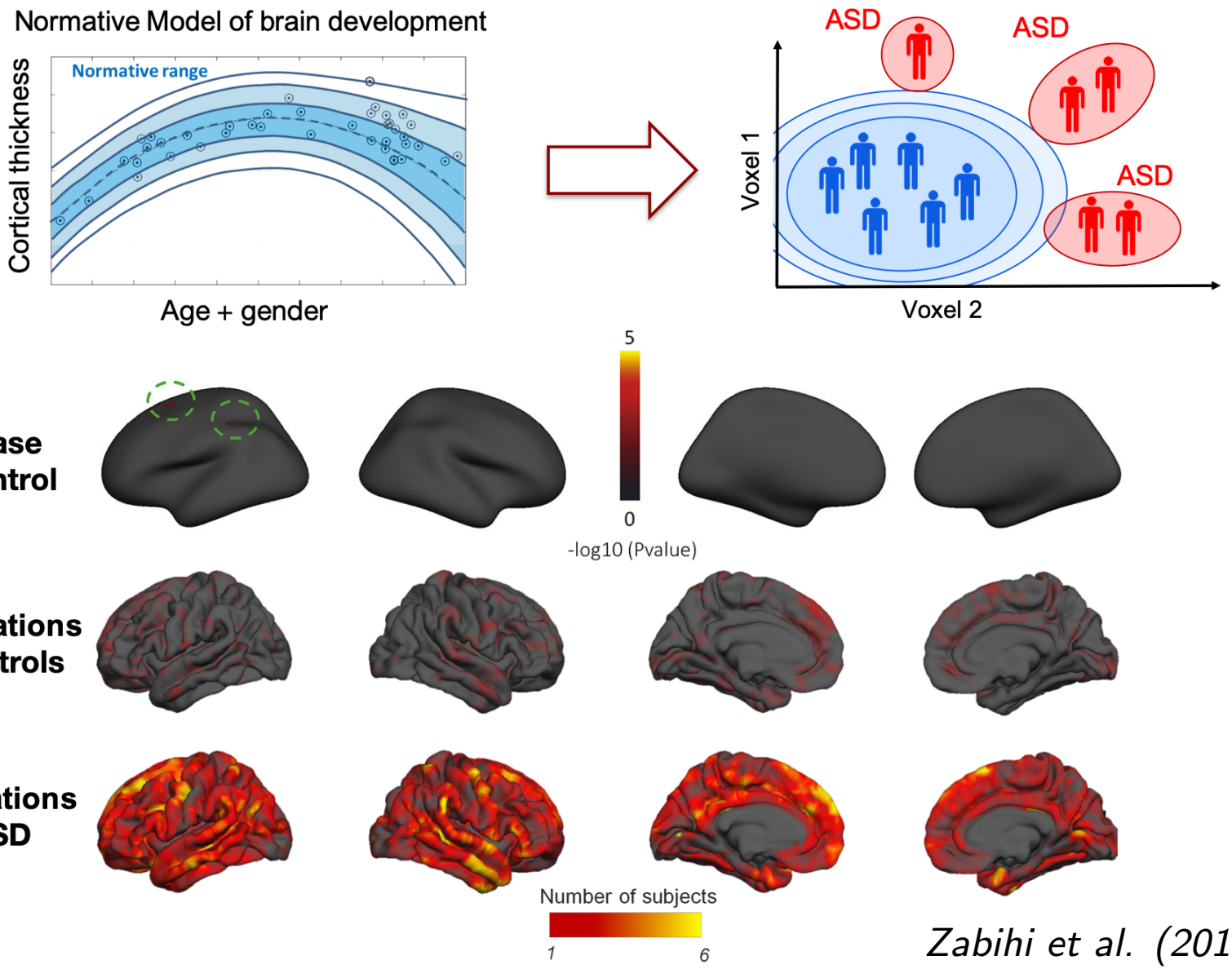
Zabihi et al. (2019)

Normative modelling of autism



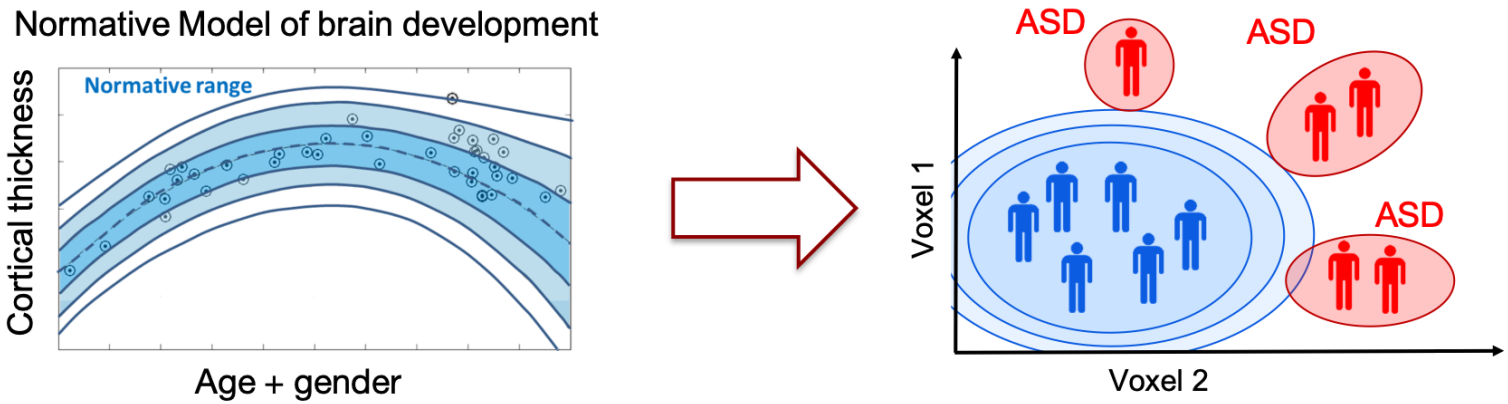
Zabihi et al. (2019)

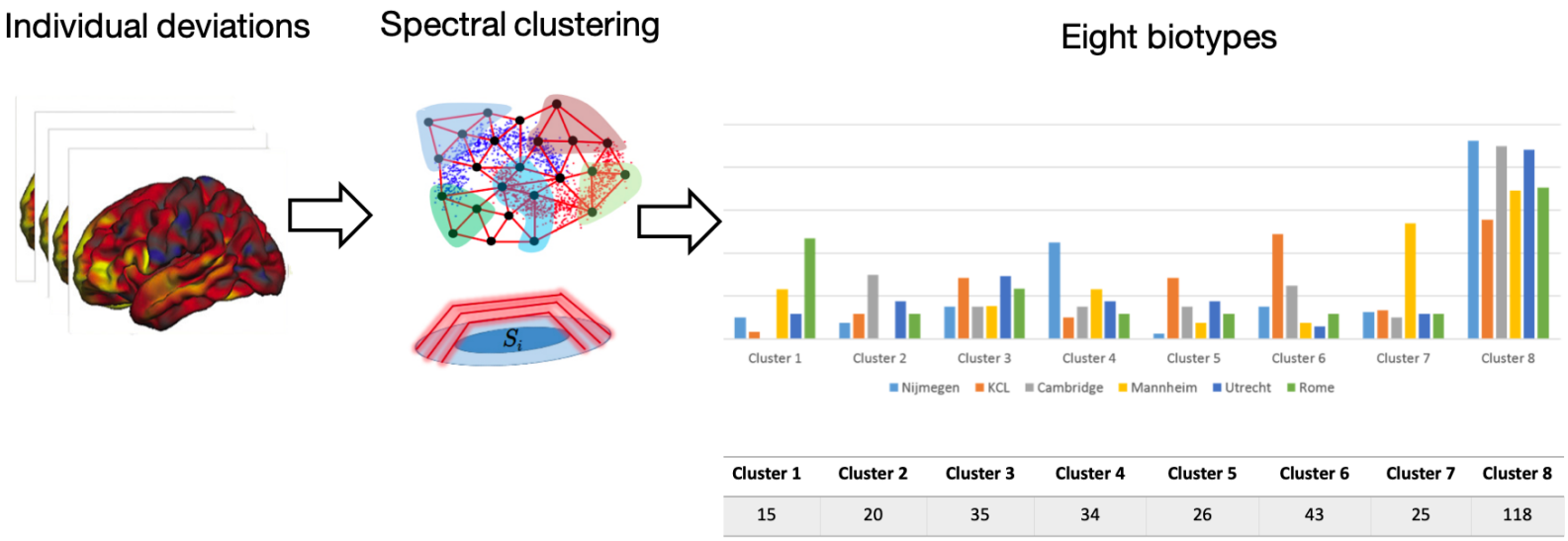
Normative modelling of autism



Zabihi et al. (2019)

Normative modelling of autism





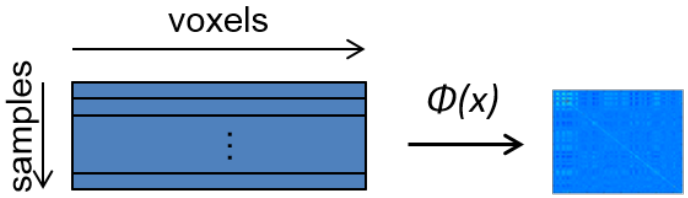
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
15	20	35	34	26	43	25	118

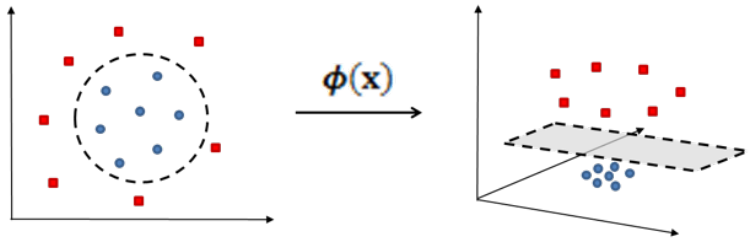
Zabihi et al. (2019)

Outline

Kernels

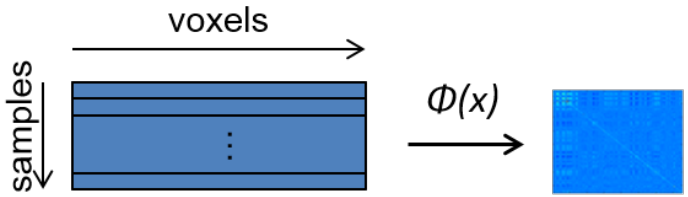
- Kernel methods (e.g. SVM, GPs) use the "kernel trick" to turn a linear model into a non-linear one

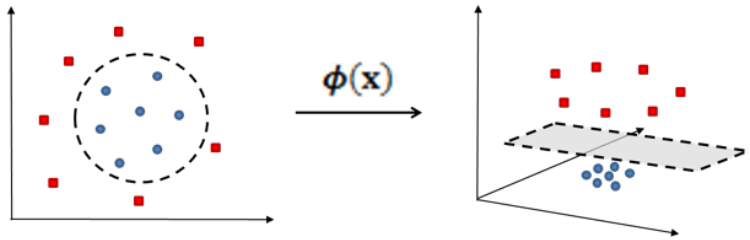




Kernels

- Kernel methods (e.g. SVM, GPs) use the "kernel trick" to turn a linear model into a non-linear one





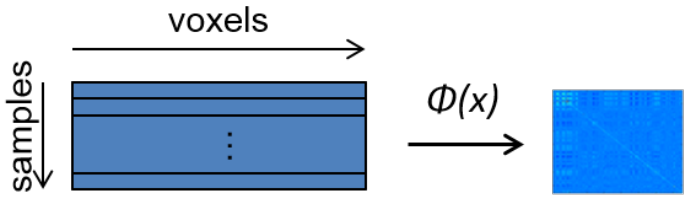
$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

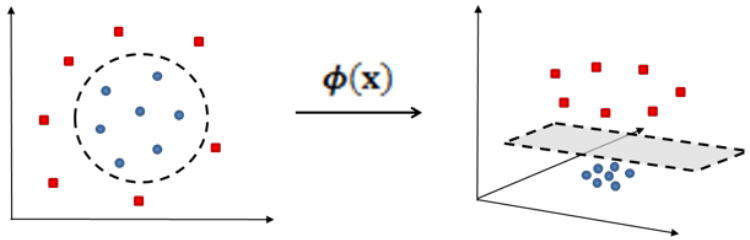
$$\text{s.t.:} \sum_i \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C$$

Kernels

- Kernel methods (e.g. SVM, GPs) use the "kernel trick" to turn a linear model into a non-linear one





$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{s.t.:} \sum_i \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C$$

- In the dual form, the data appear as an inner product, which can be substituted with a kernel function

$$\mathbf{x}_i^T \mathbf{x}_j \Rightarrow k(\mathbf{x}_i, \mathbf{x}_j) \Rightarrow \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$$

Kernels

- Any method that can be written in terms of inner products can be kernelized
- Kernel function must give rise to a positive definite matrix
- Many different functions are admissable

$$\begin{aligned}
\text{linear} &: k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}' \\
\text{RBF} &: k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma ||\mathbf{x} - \mathbf{x}'||^2) \\
\text{polynomial} &: k(\mathbf{x}, \mathbf{x}') = (\gamma \mathbf{x}^T \mathbf{x}' + c)^d \\
\text{sigmoid} &: k(\mathbf{x}, \mathbf{x}') = \tanh(\gamma \mathbf{x}^T \mathbf{x}' + c)
\end{aligned}$$

- linear operations on kernels also yield valid kernels, e.g.

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + 2k_2(\mathbf{x}, \mathbf{x}')k_3(\mathbf{x}, \mathbf{x}') + \ldots$$

- This is the basis for *multi-kernel learning*

Multi-Kernel Learning

- Kernels can represent different modalities, different views or different regions

$k(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{M} d_m k_m(\mathbf{x}, \mathbf{x}')$ with $d_m \geq 0$ and $\sum_m d_m = 1$

- Optimisation problem is a convex combination of kernels

$$\min_{\mathbf{w}} \frac{1}{2} \sum_{m=1}^{M} ||\mathbf{w}_m||^2 + C \sum_{i=1}^{n} \xi_i$$

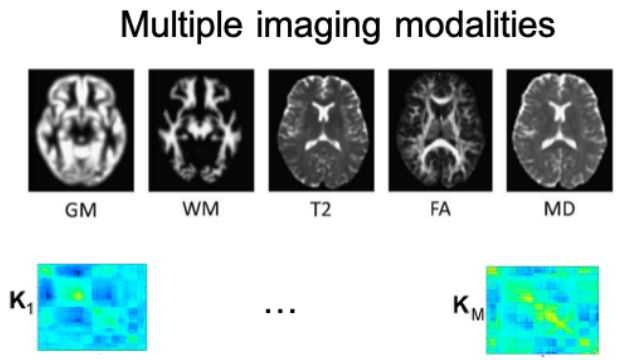
$$\text{subject to: } y_i(\sum_m (\mathbf{x}_i^T \mathbf{w}_m + b)) \geq 1 - \xi_i, \xi_i \geq 0 \; \forall i$$

$$\sum_m d_m = 1 \text{ and } d_m \geq 0$$

Applications of MKL in neuroscience

Multimodal data fusion for predicting brain disorders

$$p(y_{ic}|\mathbf{x}_1 \dots \mathbf{x}_M) = \frac{\exp f_{ic}}{\sum_d \exp f_{id}} \qquad p(\mathbf{y}|\mathbf{x}_1 \dots \mathbf{x}_M) = \mathcal{GP}(\mathbf{0}, \mathbf{K}(\theta))$$

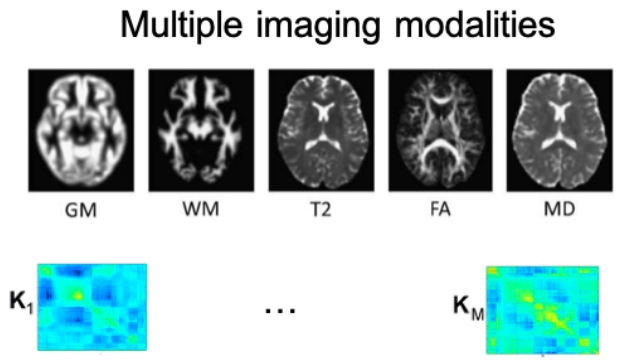


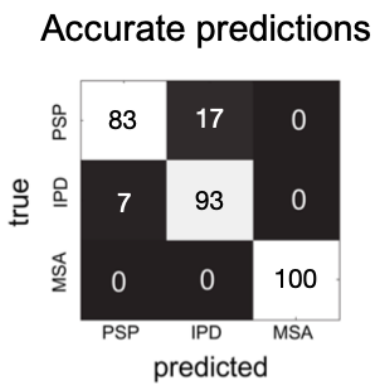
Filippone et al. (2012)

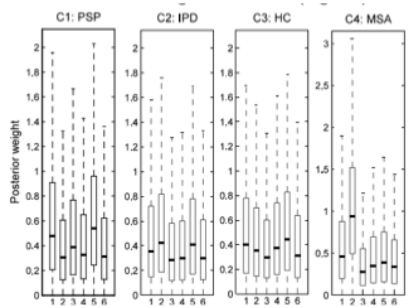
Applications of MKL in neuroscience

Multimodal data fusion for predicting brain disorders

$$p(y_{ic}|\mathbf{x}_1 \dots \mathbf{x}_M) = \frac{\exp f_{ic}}{\sum_d \exp f_{id}} \qquad p(\mathbf{y}|\mathbf{x}_1 \dots \mathbf{x}_M) = \mathcal{GP}(\mathbf{0}, \mathbf{K}(\theta))$$



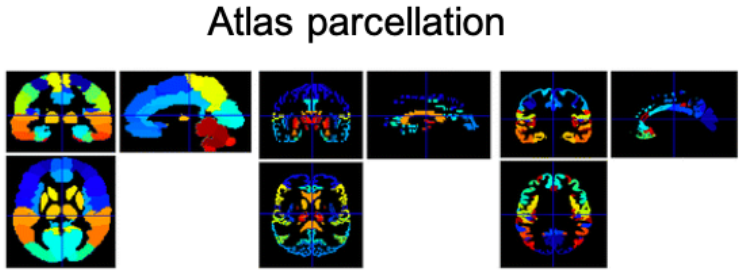


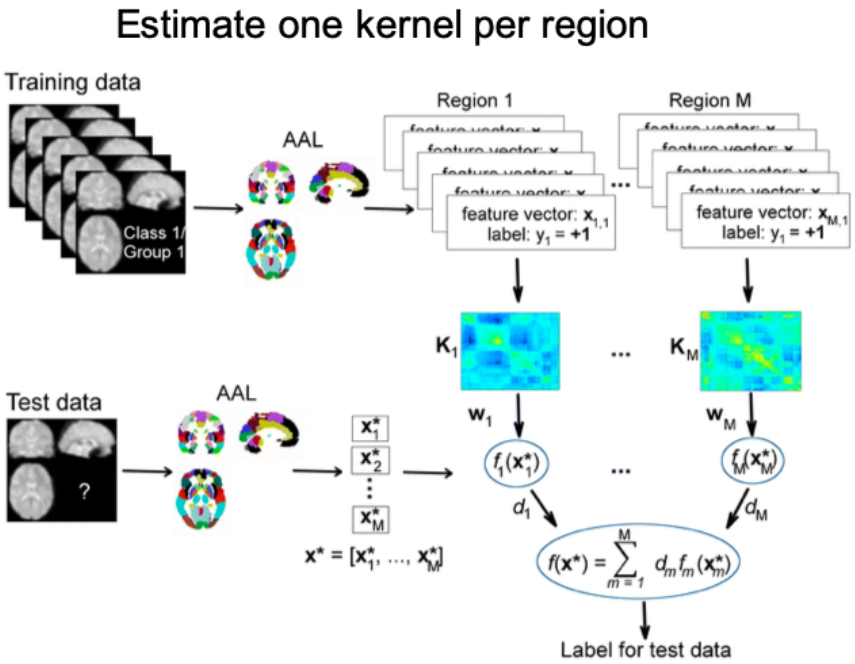


Filippone et al. (2012)

Applications of MKL in neuroscience

Better accommodating spatial variation

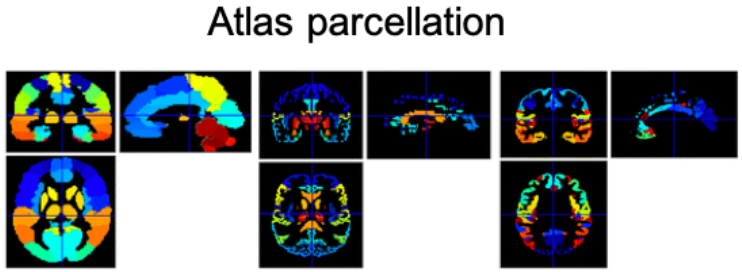


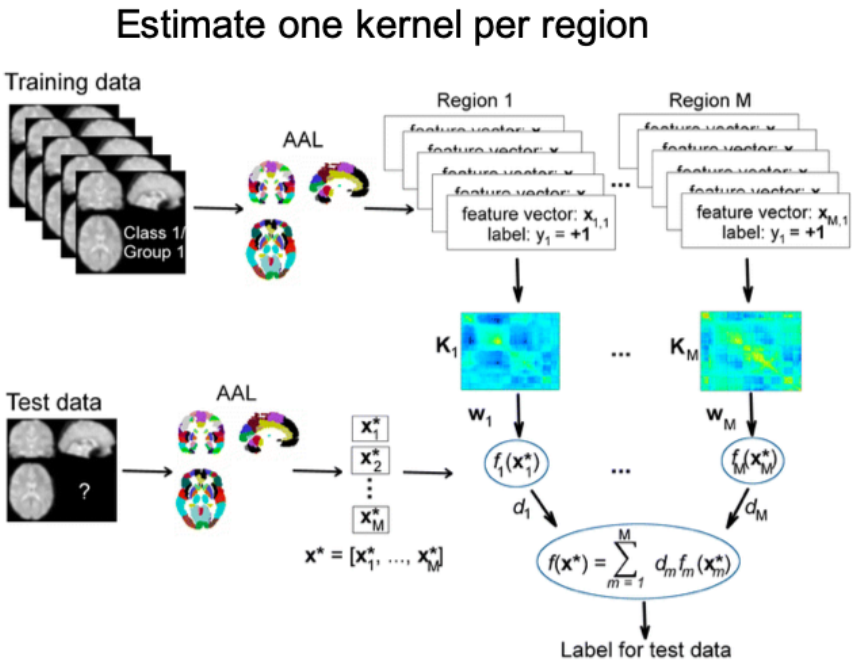


Schrouff et al. (2018)

Applications of MKL in neuroscience

Better accommodating spatial variation



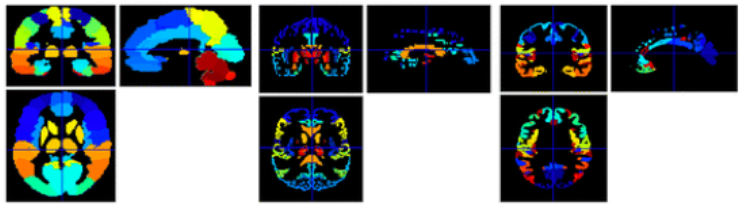


Schrouff et al. (2018)

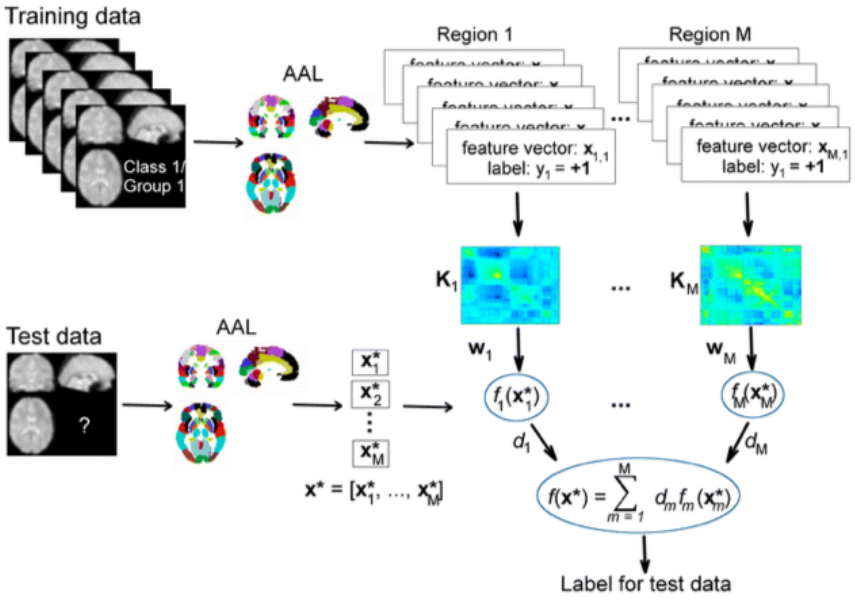
Applications of MKL in neuroscience

Better accommodating spatial variation

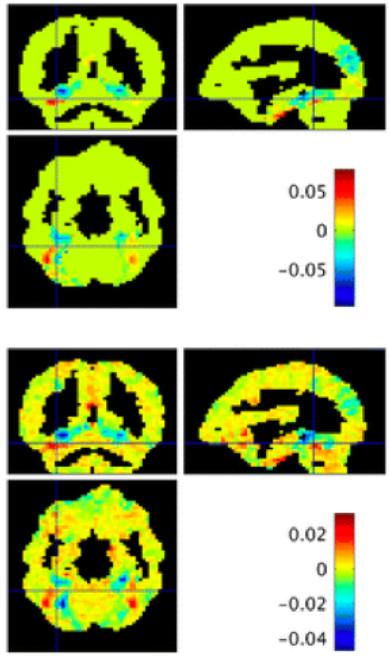
Atlas parcellation



Estimate one kernel per region



Improve SNR



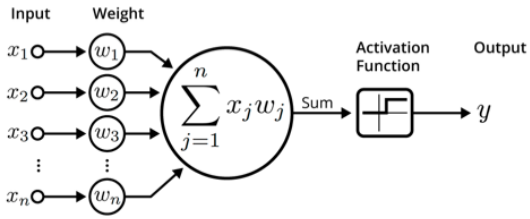
Schrouff et al. (2018)

Deep Learning

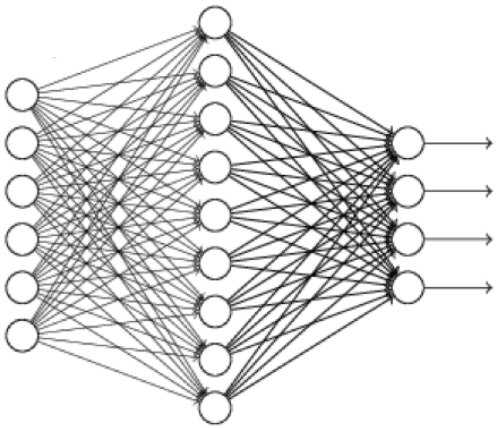
It is helpful to think of deep learning as combining matix products with point-wise linearity, e.g.:

$$f(\mathbf{x}) = \sigma\left(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1\right)\mathbf{W}_2$$

Artificial neuron
(pointwise non-linearity)



Fully connected

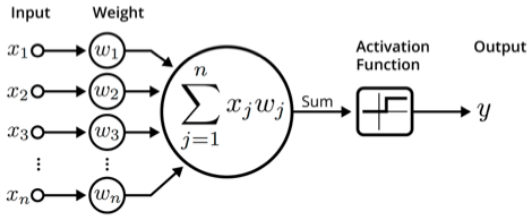


Deep Learning

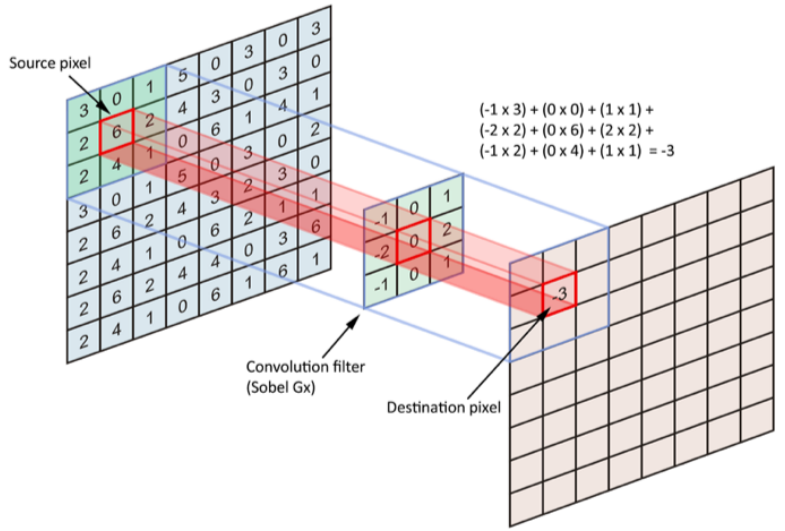
It is helpful to think of deep learning as combining matix products with point-wise linearity, e.g.:

$$f(\mathbf{x}) = \sigma\left(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1\right)\mathbf{W}_2$$

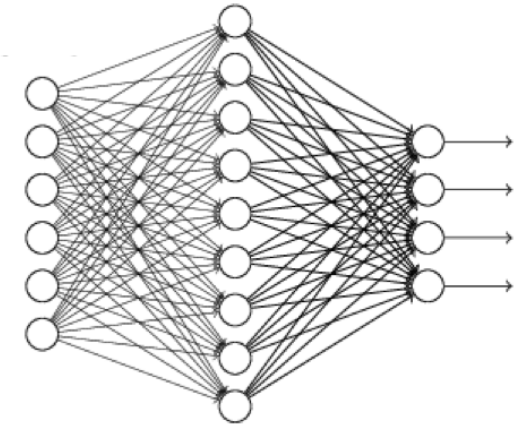
Artificial neuron
(pointwise non-linearity)



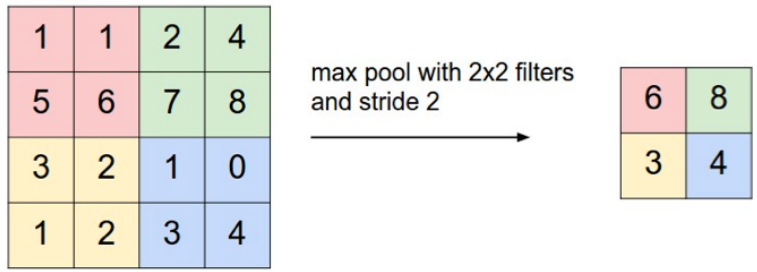
Convolution



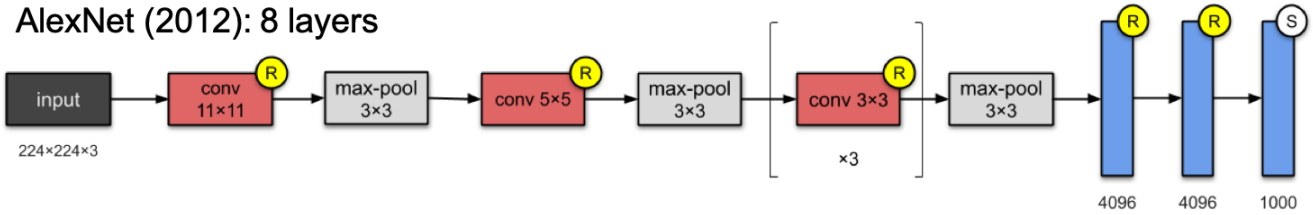
Fully connected



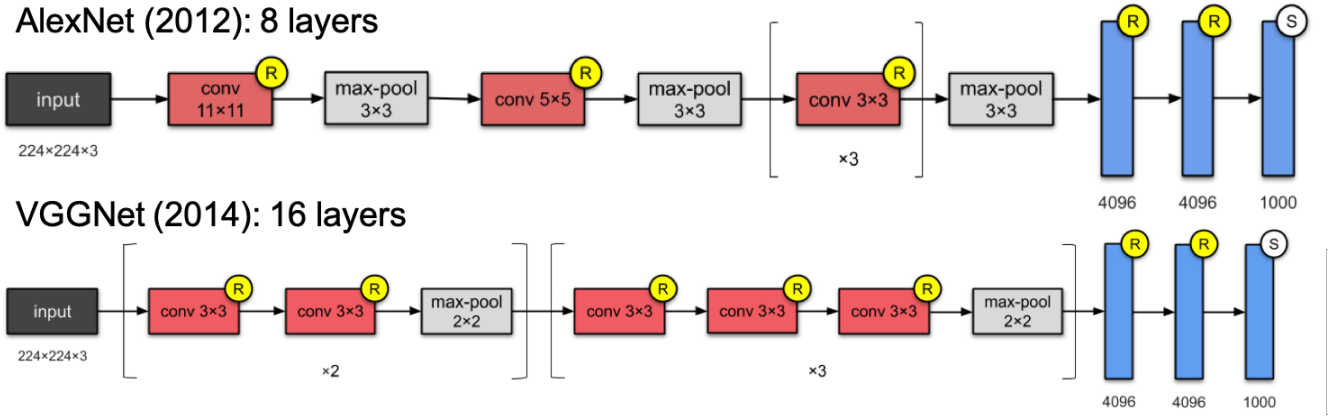
Max pooling



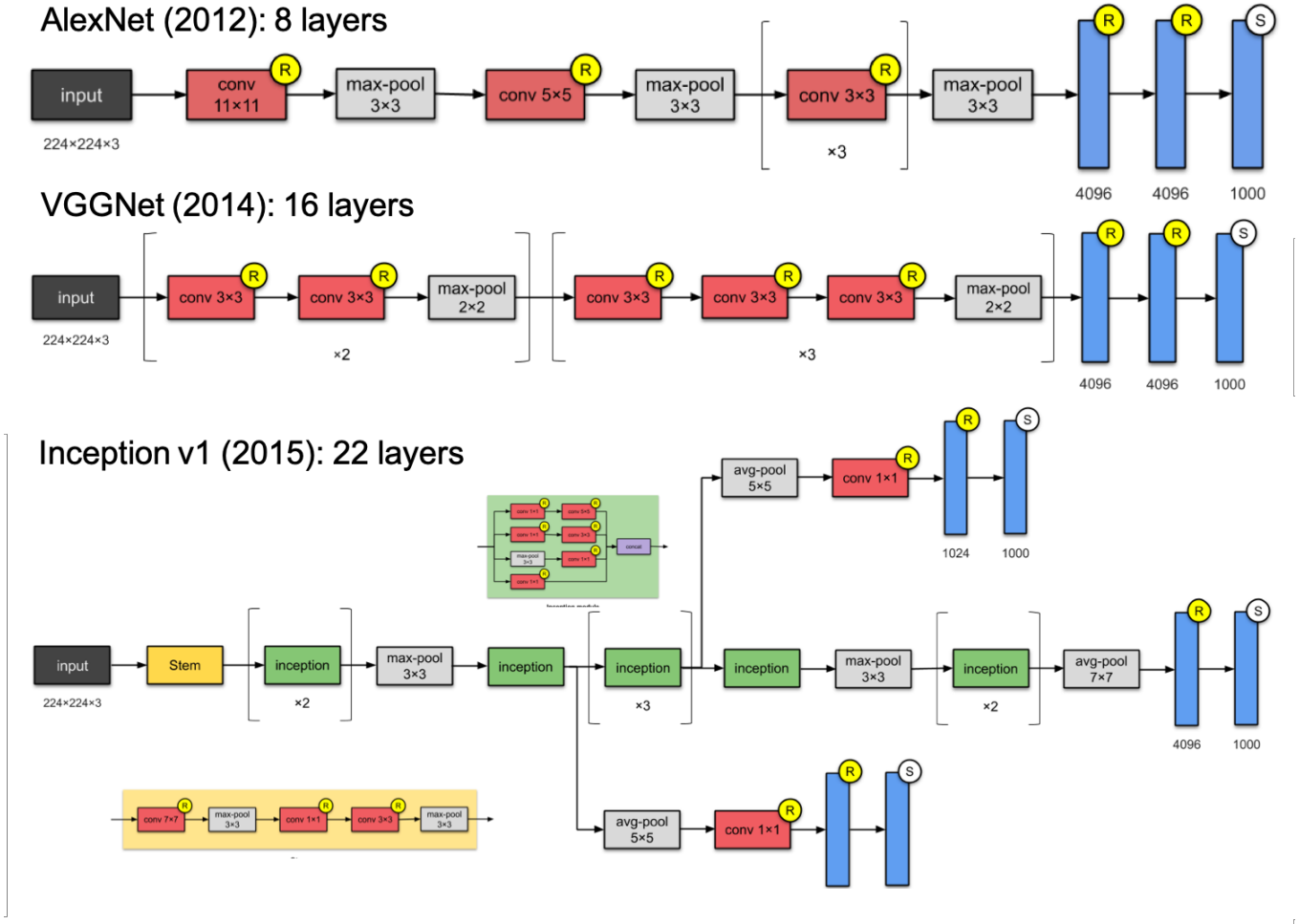
Convolutional Neural Networks



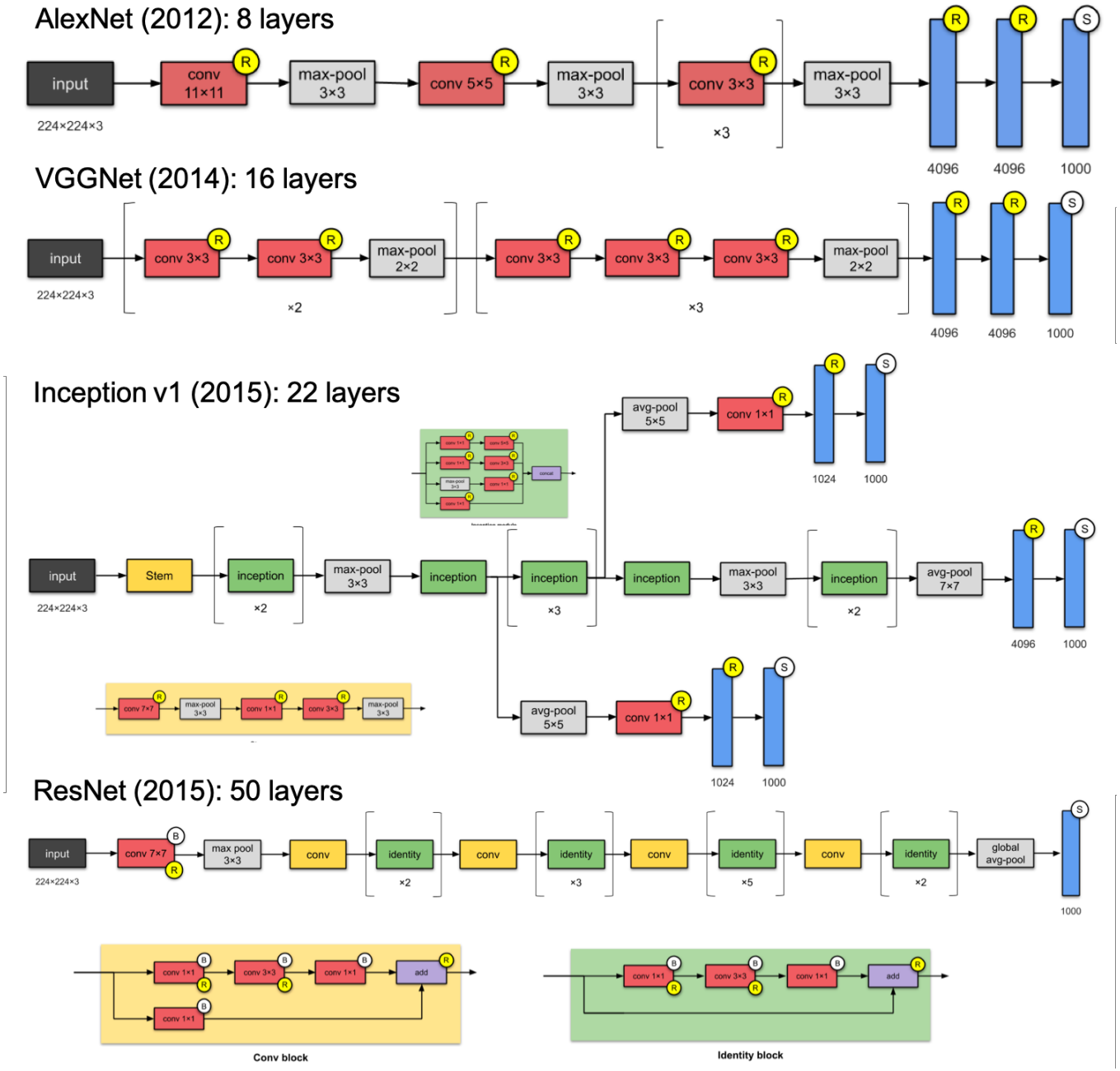
Convolutional Neural Networks



Convolutional Neural Networks

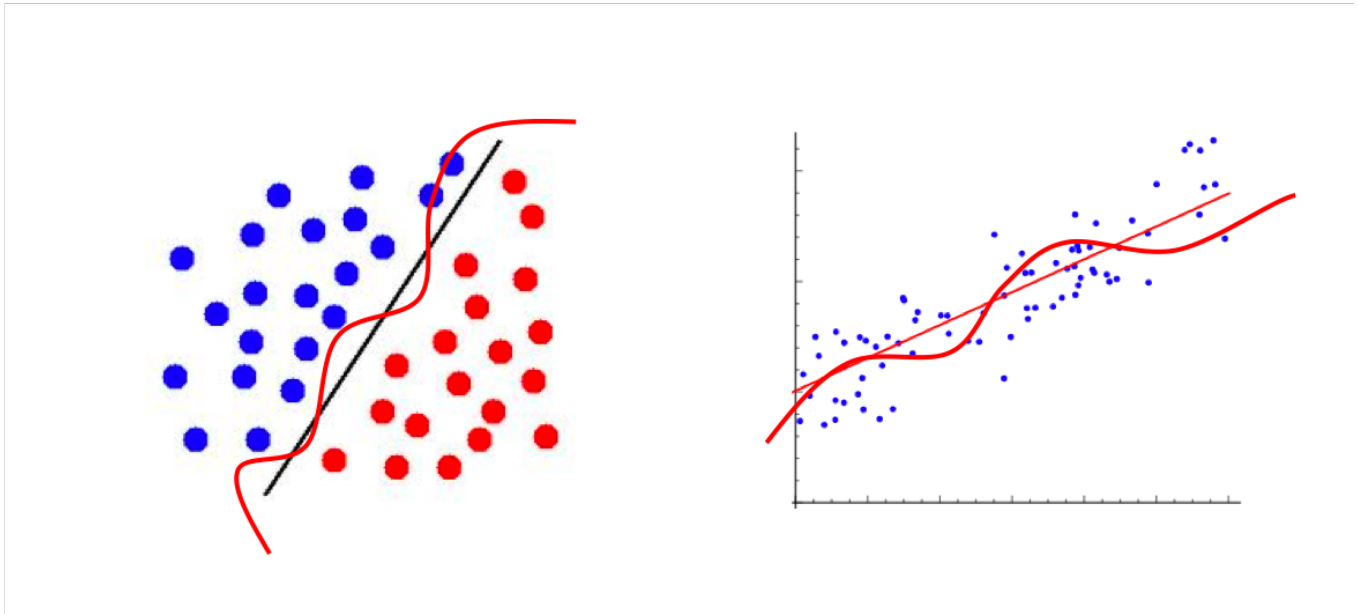


Convolutional Neural Networks



Overfitting (again)

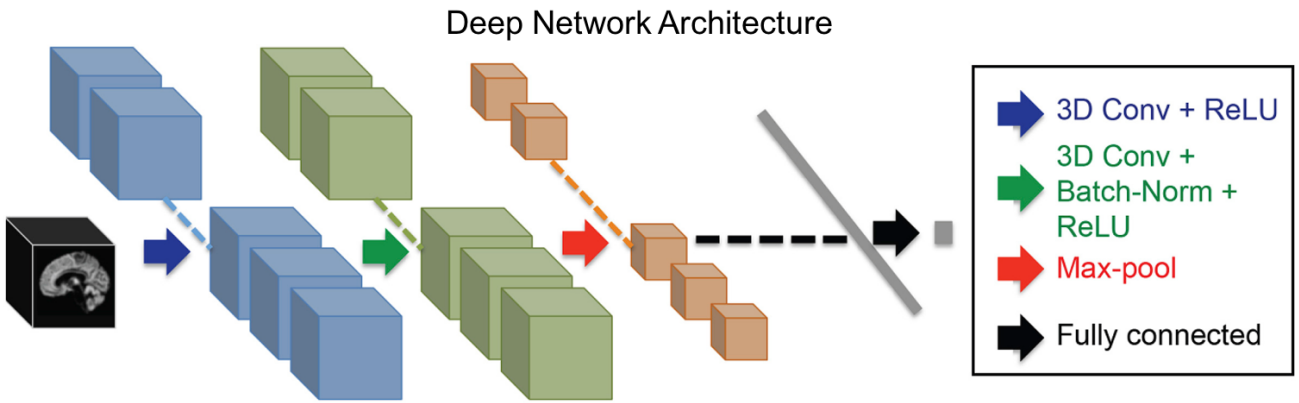
- But if your problem is linear, your fancy nonlinear algorithm will just overfit



- The more complex the model, the easier it is to overfit
- In complex (deep) models it is often not possible to properly optimise all parameters
- This makes validation extremely important!

Deep Learning in Psychiatry?

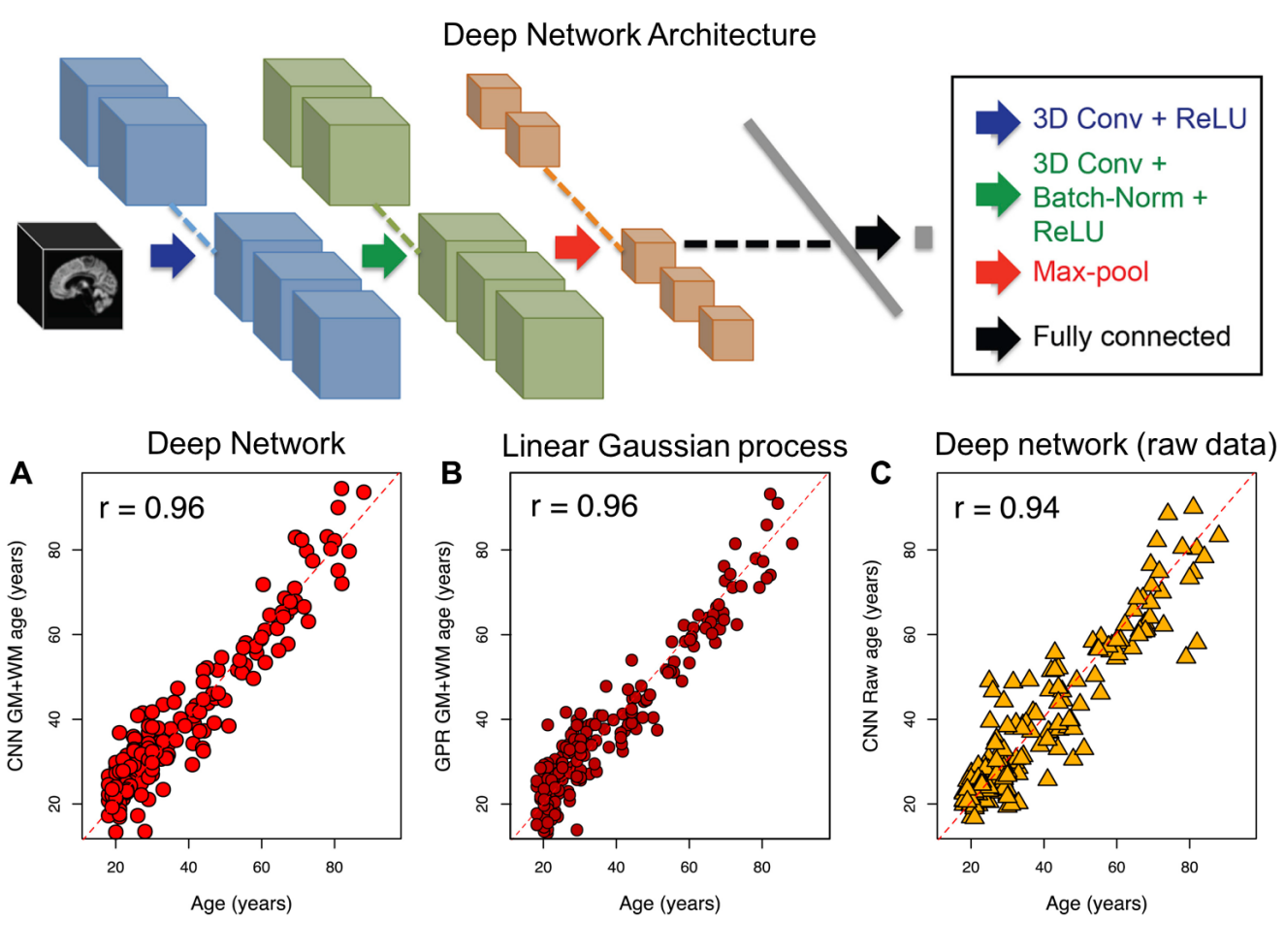
- Predict age from N= 2001 structural MRI images



Cole et al. (2017)

Deep Learning in Psychiatry?

- Predict age from N= 2001 structural MRI images



- Similar performance to a linear model on preprocessed data
- Better performance on minimally processed data *Cole et al. (2017)*

Deep Learning in Neuroscience?

Deep Neural Networks and Kernel Regression Achieve Comparable Accuracies for Functional Connectivity Prediction of Behavior and Demographics

Tong He[1,2], Ru Kong[1,2], Avram J. Holmes[3], Minh Nguyen[1,2], Mert R. Sabuncu[4], Simon B. Eickhoff[5,6], Danilo Bzdok[7,8,9], Jiashi Feng[2], B.T. Thomas Yeo[1,2,10,11,12]

He et al. (2019)

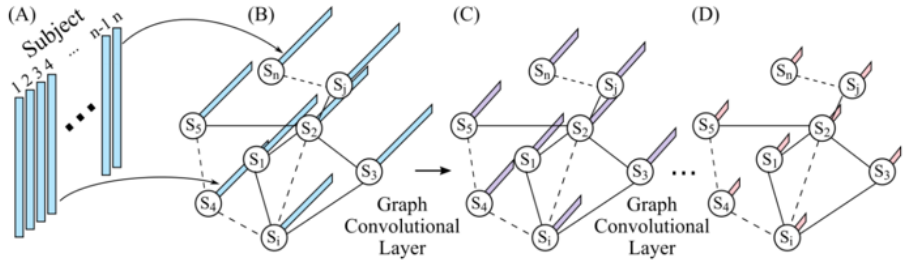
Deep Learning in Neuroscience?

Deep Neural Networks and Kernel Regression Achieve Comparable Accuracies for Functional Connectivity Prediction of Behavior and Demographics

Tong He[1,2], Ru Kong[1,2], Avram J. Holmes[3], Minh Nguyen[1,2], Mert R. Sabuncu[4], Simon B. Eickhoff[5,6], Danilo Bzdok[7,8,9], Jiashi Feng[2], B.T. Thomas Yeo[1,2,10,11,12]

BrainNet CNN

Graph CNN



HUMAN Connectome PROJECT

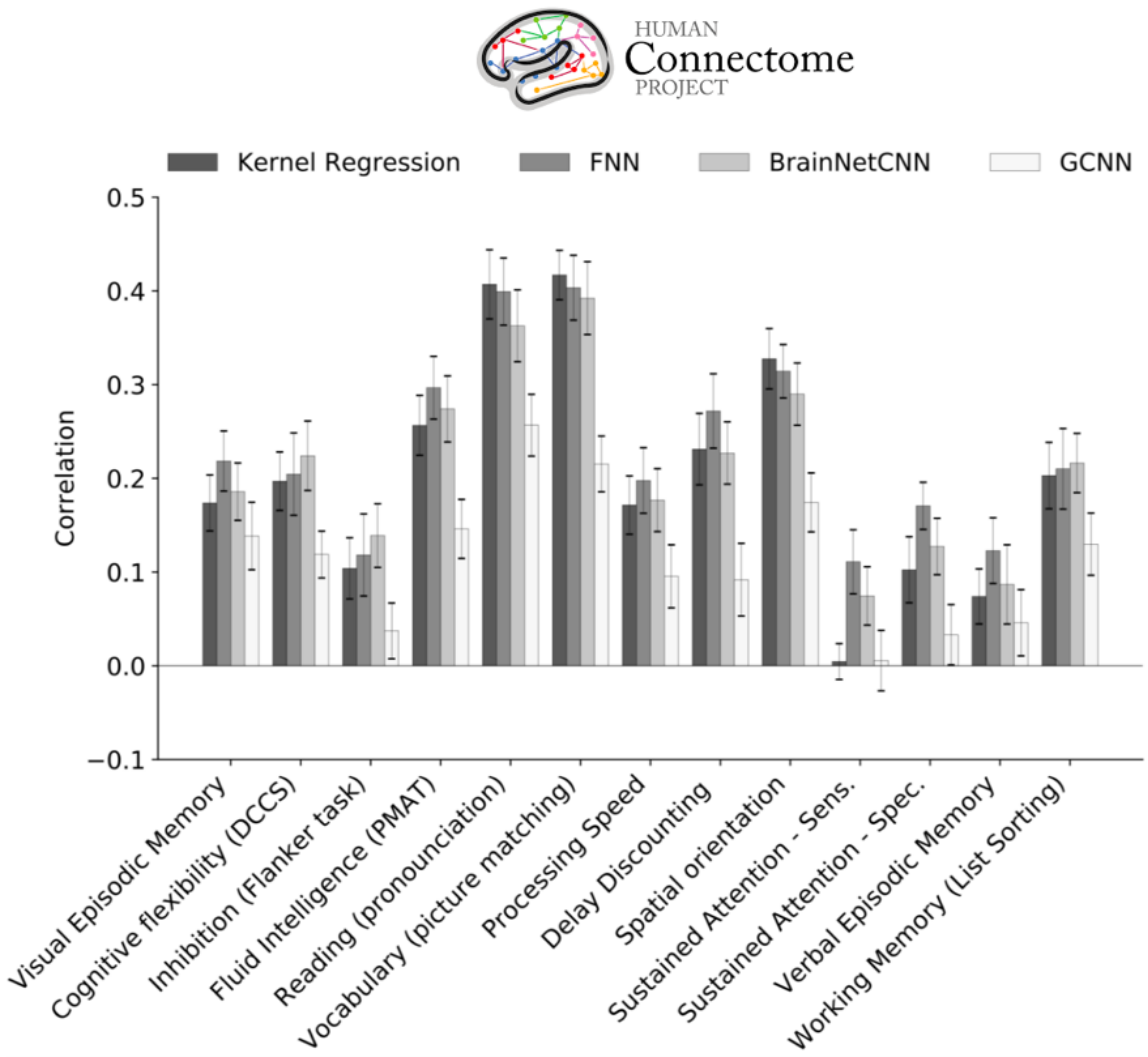
N = 1200

biobank uk

N = 10 000

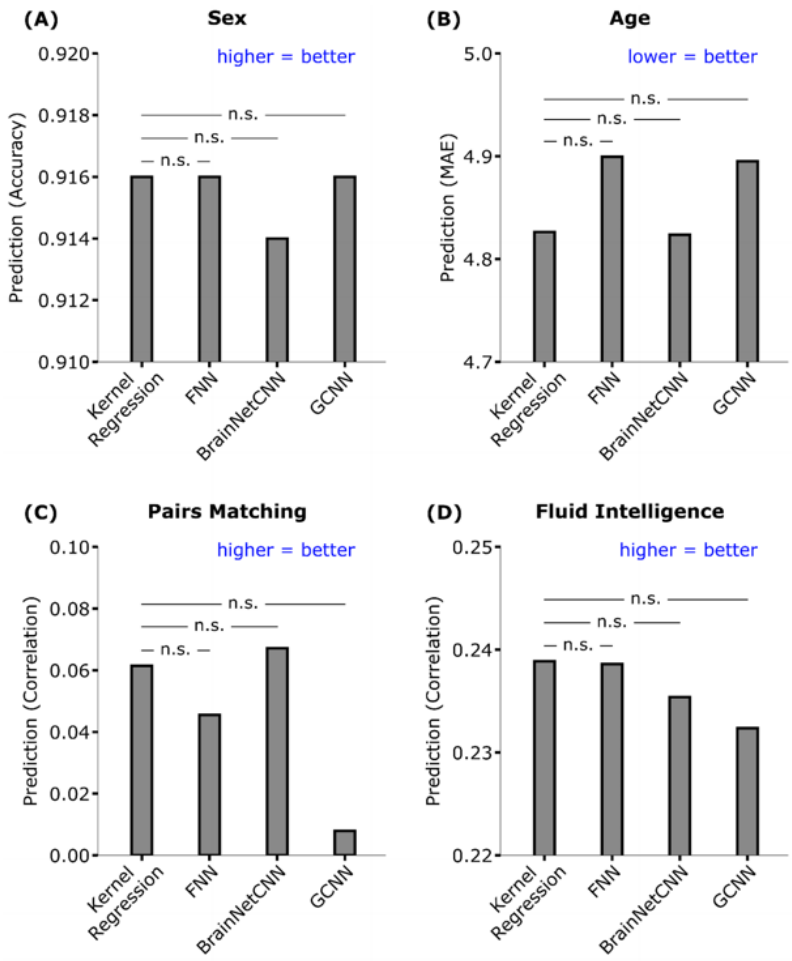
He et al. (2019)

Deep Learning in Neuroscience?



He et al. (2019)

Deep Learning in Neuroscience?



He et al. (2019)

Deep Learning in Medicine?

- Predict mortality from electronic health records

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[2], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]

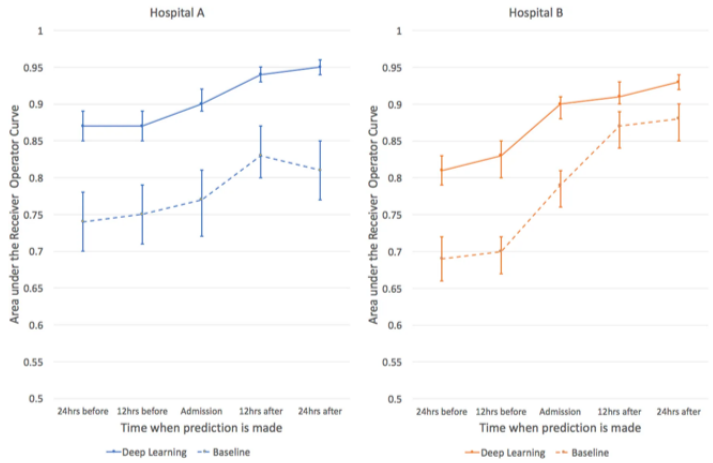
Deep Learning in Medicine?

- Predict mortality from electronic health records

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[2], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]



Deep Learning in Medicine?

- Predict mortality from electronic health records

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[2], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]

	Hospital A	Hospital B
Inpatient Mortality, AUROC[1](95% CI)		
Deep learning 24 hours after admission	**0.95**(0.94-0.96)	**0.93**(0.92-0.94)
Full feature enhanced baseline at 24 hours after admission	0.93 (0.92-0.95)	0.91 (0.89-0.92)
Full feature simple baseline at 24 hours after admission	0.93 (0.91-0.94)	0.90 (0.88-0.92)
Baseline (aEWS[2]) at 24 hours after admission	0.85 (0.81-0.89)	0.86 (0.83-0.88)
30-day Readmission, AUROC (95% CI)		
Deep learning at discharge	**0.77**(0.75-0.78)	**0.76**(0.75-0.77)
Full feature enhanced baseline at discharge	0.75 (0.73-0.76)	0.75 (0.74-0.76)
Full feature simple baseline at discharge	0.74 (0.73-0.76)	0.73 (0.72-0.74)
Baseline (mHOSPITAL[3]) at discharge	0.70 (0.68-0.72)	0.68 (0.67-0.69)
Length of Stay at least 7 days AUROC (95% CI)		
Deep learning 24 hours after admission	**0.86**(0.86-0.87)	**0.85**(0.85-0.86)
Full feature enhanced baseline at 24 hours after admission	0.85 (0.84-0.85)	0.83 (0.83-0.84)
Full feature simple baseline at 24 hours after admission	0.83 (0.82-0.84)	0.81 (0.80-0.82)
Baseline (mLiu[4]) at 24 hours after admission	0.76 (0.75-0.77)	0.74 (0.73-0.75)

Deep Learning in Medicine?

- Predict mortality from electronic health records

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[2], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]

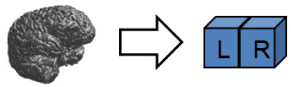
	Hospital A	Hospital B
Inpatient Mortality, AUROC[1](95% CI)		
Deep learning 24 hours after admission	**0.95**(0.94-0.96)	**0.93**(0.92-0.94)
Full feature enhanced baseline at 24 hours after admission	0.93 (0.92-0.95)	0.91 (0.89-0.92)
Full feature simple baseline at 24 hours after admission	0.93 (0.91-0.94)	0.90 (0.88-0.92)
Baseline (aEWS[2]) at 24 hours after admission	0.85 (0.81-0.89)	0.86 (0.83-0.88)
30-day Readmission, AUROC (95% CI)		
Deep learning at discharge	**0.77**(0.75-0.78)	**0.76**(0.75-0.77)
Full feature enhanced baseline at discharge	0.75 (0.73-0.76)	0.75 (0.74-0.76)
Full feature simple baseline at discharge	0.74 (0.73-0.76)	0.73 (0.72-0.74)
Baseline (mHOSPITAL[3]) at discharge	0.70 (0.68-0.72)	0.68 (0.67-0.69)
Length of Stay at least 7 days AUROC (95% CI)		
Deep learning 24 hours after admission	**0.86**(0.86-0.87)	**0.85**(0.85-0.86)
Full feature enhanced baseline at 24 hours after admission	0.85 (0.84-0.85)	0.83 (0.83-0.84)
Full feature simple baseline at 24 hours after admission	0.83 (0.82-0.84)	0.81 (0.80-0.82)
Baseline (mLiu[4]) at 24 hours after admission	0.76 (0.75-0.77)	0.74 (0.73-0.75)

Outline

Mapping the discriminative pattern

For clinical applications it is crucial to infer which variables drive the predictions. There are multiple options:

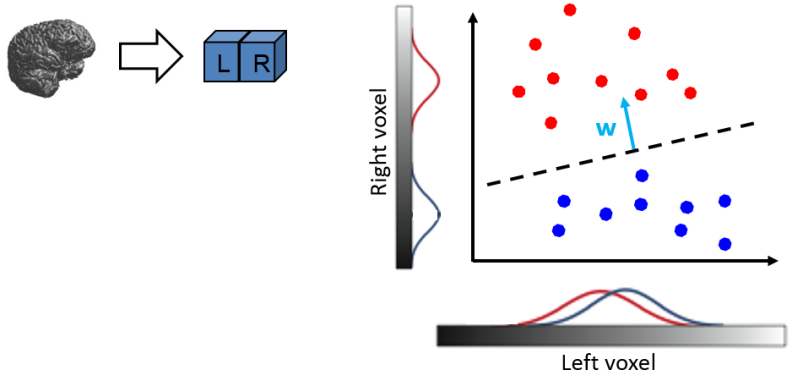
- Regional classification accuracy as a proxy (searchlight)
- For linear models, the weights can be directly visualised ("discriminative mapping")



Mapping the discriminative pattern

For clinical applications it is crucial to infer which variables drive the predictions. There are multiple options:

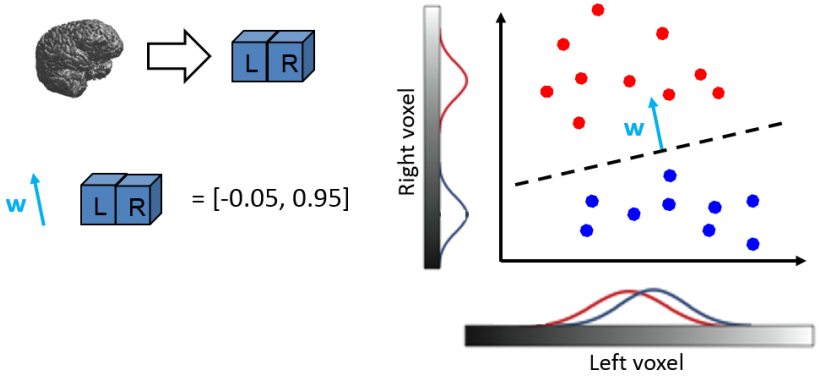
- Regional classification accuracy as a proxy (searchlight)
- For linear models, the weights can be directly visualised ("discriminative mapping")



Mapping the discriminative pattern

For clinical applications it is crucial to infer which variables drive the predictions. There are multiple options:

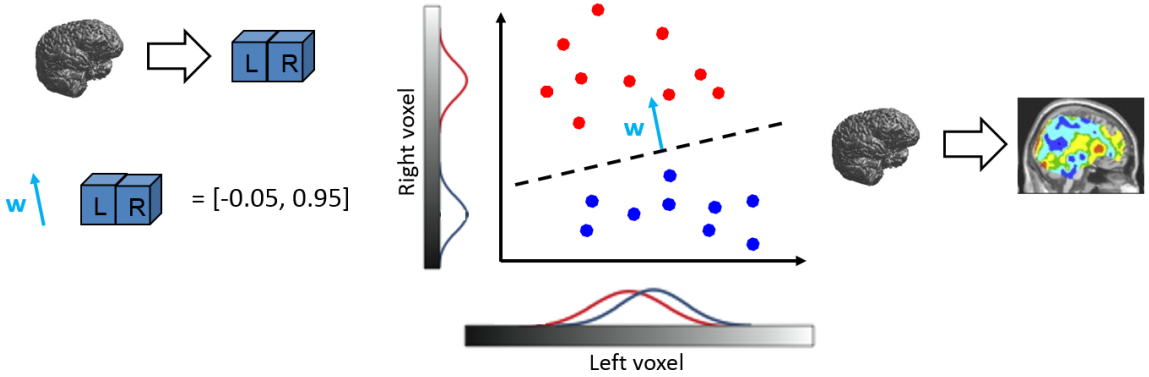
- Regional classification accuracy as a proxy (searchlight)
- For linear models, the weights can be directly visualised ("discriminative mapping")



Mapping the discriminative pattern

For clinical applications it is crucial to infer which variables drive the predictions. There are multiple options:

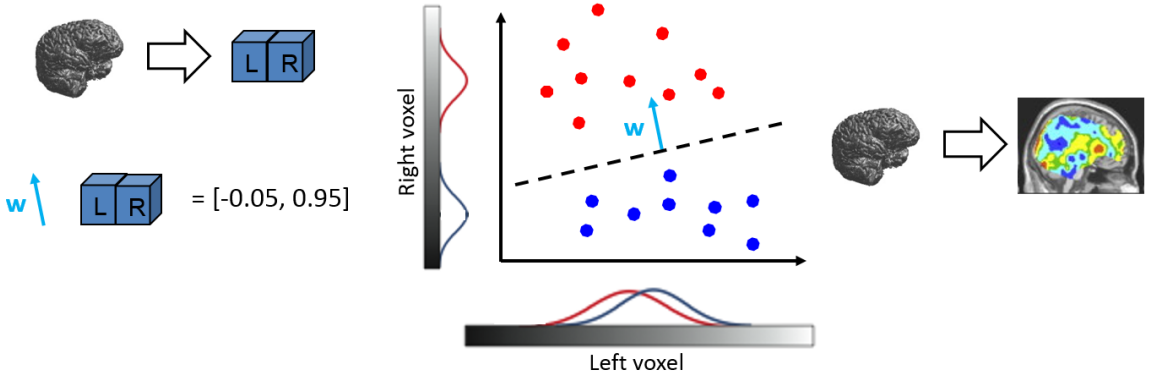
- Regional classification accuracy as a proxy (searchlight)
- For linear models, the weights can be directly visualised ("discriminative mapping")



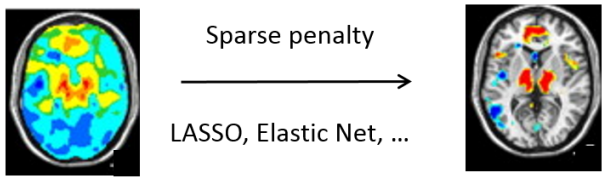
Mapping the discriminative pattern

For clinical applications it is crucial to infer which variables drive the predictions. There are multiple options:

- Regional classification accuracy as a proxy (searchlight)
- For linear models, the weights can be directly visualised ("discriminative mapping")



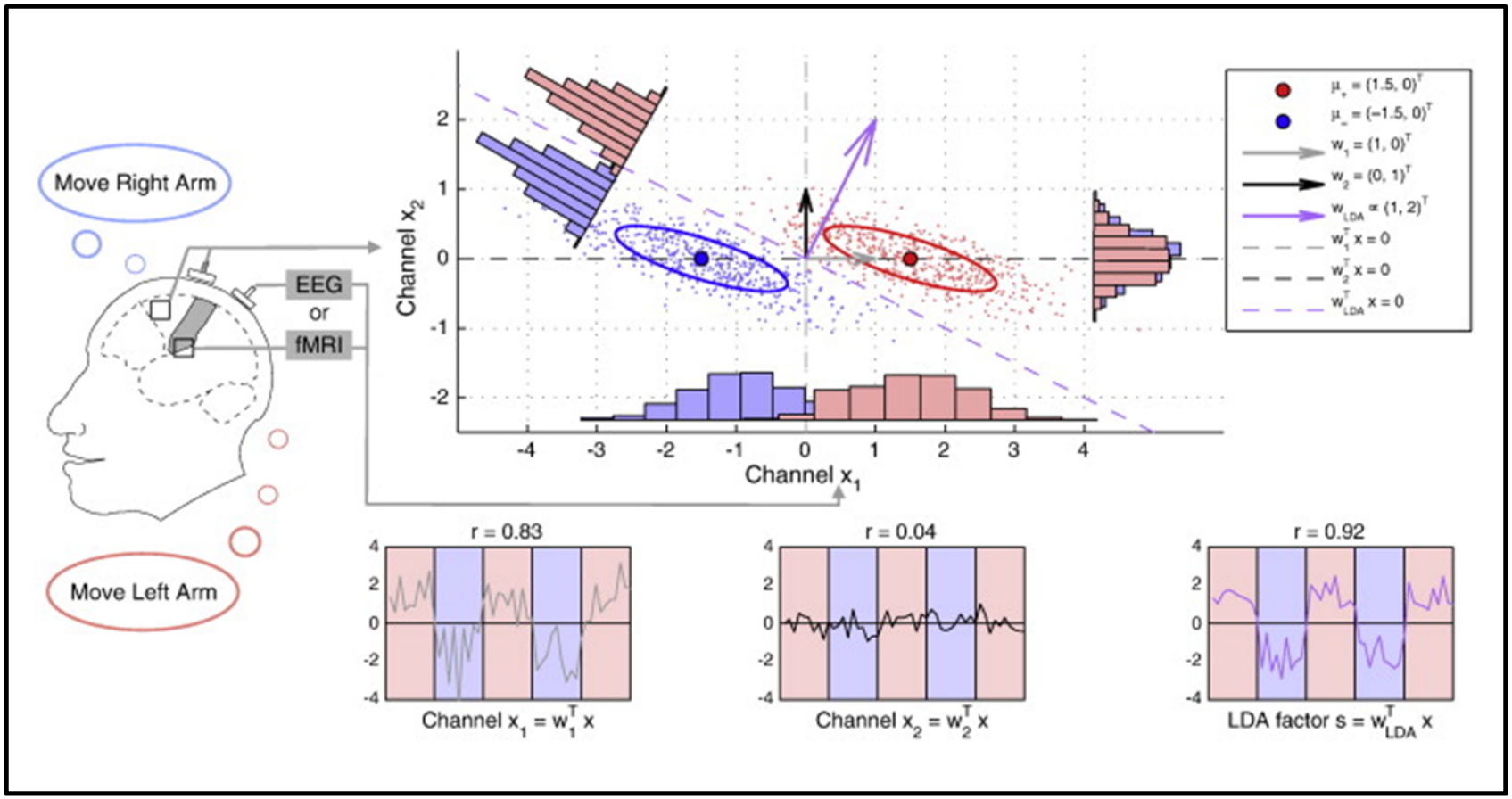
- Can also use regularization to enforce sparse weights



- Weights are often regarded as being difficult to interpret, but this is not always true

Weights do not reflect univariate differences

One proposal is to consider the the weights from a forward model



Construct 'forward maps' by premultiplying by the data covariance

$$\mathbf{a} = \frac{1}{\sigma_y^2}\mathbf{\Sigma}_x\mathbf{w}$$

Haufe et al. (2014)

Understanding weights of discriminative models

- The correct interpretation of the weights is the contribution of each feature to the predictions. This is the same as in a GLM
- Difficulty arises only due to multicollinearity between predictor variables which inflates the variance of the weights
- A variable can have a high weight because:
  1. It is associated with the response variable
  2. It acts as a 'supressor' variable that helps to cancel out noise or mismatch in other covariates

Kraha et al. (2012)

Understanding weights of discriminative models

- The correct interpretation of the weights is the contribution of each feature to the predictions. This is the same as in a GLM
- Difficulty arises only due to multicollinearity between predictor variables which inflates the variance of the weights
- A variable can have a high weight because:
  1. It is associated with the response variable
  2. It acts as a 'supressor' variable that helps to cancel out noise or mismatch in other covariates
- To distinguish between these possibilities, we can do the following (assumes standardized data)

$$\mathbf{a} \propto \mathbf{\Sigma}_x \mathbf{w} = \text{cov}\left[\mathbf{X}, \hat{\mathbf{y}}\right] = \text{corr}\left[\mathbf{X}, \hat{\mathbf{y}}\right]$$

Kraha et al. (2012)

Understanding weights of discriminative models

- The correct interpretation of the weights is the contribution of each feature to the predictions. This is the same as in a GLM
- Difficulty arises only due to multicollinearity between predictor variables which inflates the variance of the weights
- A variable can have a high weight because:
  1. It is associated with the response variable
  2. It acts as a 'supressor' variable that helps to cancel out noise or mismatch in other covariates
- To distinguish between these possibilities, we can do the following (assumes standardized data)

$$\mathbf{a} \propto \mathbf{\Sigma}_x \mathbf{w} = \text{cov}\left[\mathbf{X}, \hat{\mathbf{y}}\right] = \text{corr}\left[\mathbf{X}, \hat{\mathbf{y}}\right]$$

- These are **structure coefficients** from multivariate statistics
- The univariate association between covariate p and the predictions is given simply by:

Kraha et al. (2012)

$$\rho(\mathbf{x}_p, \hat{\mathbf{y}})$$

But what about the penalty?

- Collinearity is well-known in classical GLM settings, where advice is often given to avoid collinearity
- It is true that collinearity impacts on efficiency, but models with collinear predictors are still interpretable
- Collinearity also impacts penalised regression. Recall that:

$$f(\mathbf{x}_i, \mathbf{w}) = \mathbf{x}_i^T \mathbf{w} \quad \Rightarrow \hat{\mathbf{w}} = \min_{\mathbf{w}} \sum_{i=1}^{n} \ell(y_i, f_i) + \lambda J(\mathbf{w})$$

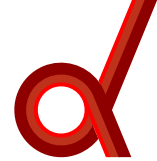
- Considering ridge regression, where the objective function is:

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} \sum_{i=1}^{n} \mathcal{N}(\mathbf{w}^T \mathbf{x}_i, \sigma^2) + \frac{\lambda}{2} ||\mathbf{w}||_2^2$$

- This is eqivalent to maximising the following

$$\min_{\mathbf{w}} -\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \mathbf{w})^2 + \lambda \mathbf{w}^T \mathbf{w} \right)$$

But what about the penalty?



- This is exactly equivalent to finding the MAP estimate of a posterior distribution over $\mathbf{w}$, with prior:

$$\mathcal{N}(\mathbf{0}, \sigma^2/\lambda\mathbf{I})$$

But what about the penalty?

- This is exactly equivalent to finding the MAP estimate of a posterior distribution over $\mathbf{w}$, with prior:

$$\mathcal{N}(\mathbf{0}, \sigma^2/\lambda\mathbf{I})$$

So what does this mean?

But what about the penalty?

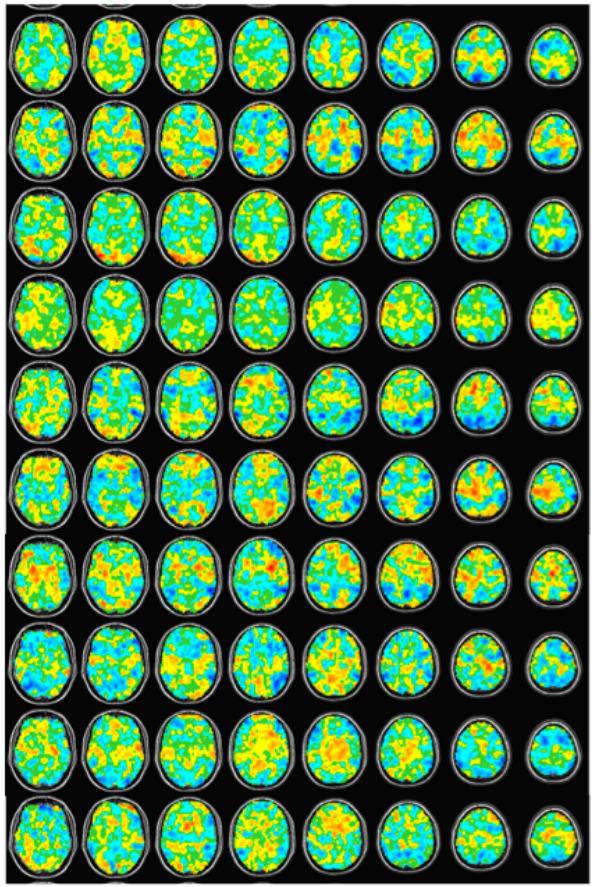
- This is exactly equivalent to finding the MAP estimate of a posterior distribution over $\mathbf{w}$, with prior:

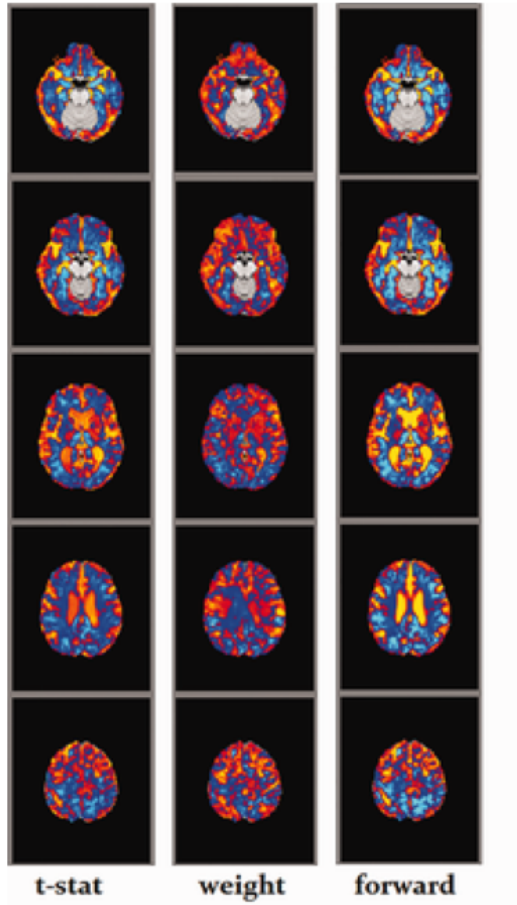
$$\mathcal{N}(\mathbf{0}, \sigma^2/\lambda\mathbf{I})$$

So what does this mean?

- The magnitude and sign of weights are influenced by collinearity
- Including and excluding variables variables can change the magnitude and sign of variables in the model
- It is obvious that when $p > n$, the problem is ill-posed in that there are many ways the same prediction can be achieved.
- Regularisation helps to stabilise coefficients, but this does not eliminate the problem

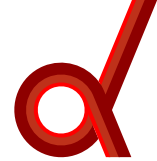
Examples of weights





Aksman et al. (2016)

Stability selection



- It is important to recognise that coefficient stability is largely independent from accuracy

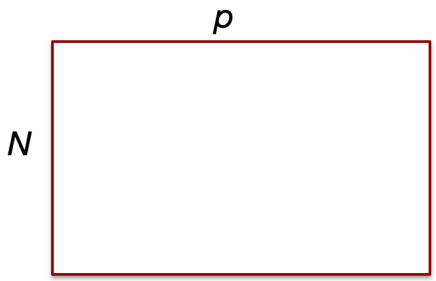
Meinshausen and Buehlmann (2010)

Stability selection

- It is important to recognise that coefficient stability is largely independent from accuracy
- **stability selection** is a method that aims to identify truly important variables by examining stability under perturbation to the data

Lasso

$$\hat{\beta} = \arg\min\left(\|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} |\beta_k| \right)$$



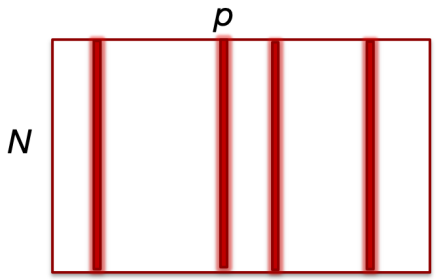
Meinshausen and Buehlmann (2010)

Stability selection

- It is important to recognise that coefficient stability is largely independent from accuracy
- **stability selection** is a method that aims to identify truly important variables by examining stability under perturbation to the data

Lasso

$$\hat{\beta} = \arg\min\left(\|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} |\beta_k| \right)$$



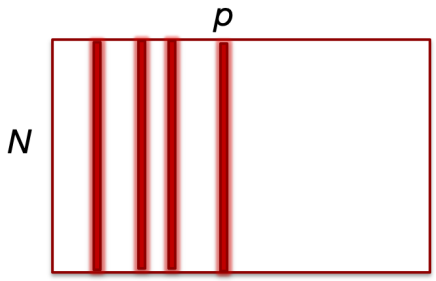
Meinshausen and Buehlmann (2010)

Stability selection

- It is important to recognise that coefficient stability is largely independent from accuracy
- **stability selection** is a method that aims to identify truly important variables by examining stability under perturbation to the data

Lasso

$$\hat{\beta} = \arg\min\left(\|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} |\beta_k| \right)$$



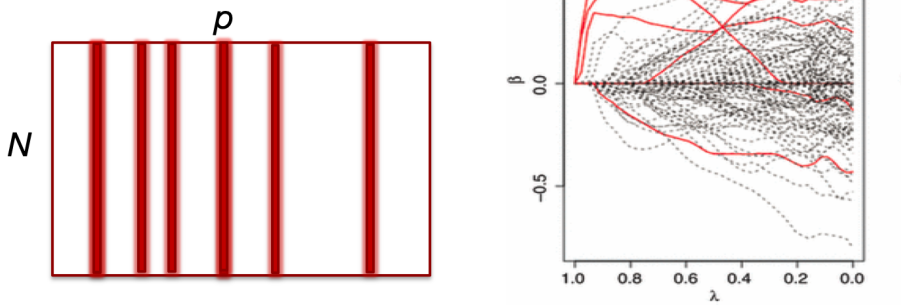
Meinshausen and Buehlmann (2010)

Stability selection

- It is important to recognise that coefficient stability is largely independent from accuracy
- **stability selection** is a method that aims to identify truly important variables by examining stability under perturbation to the data

Lasso

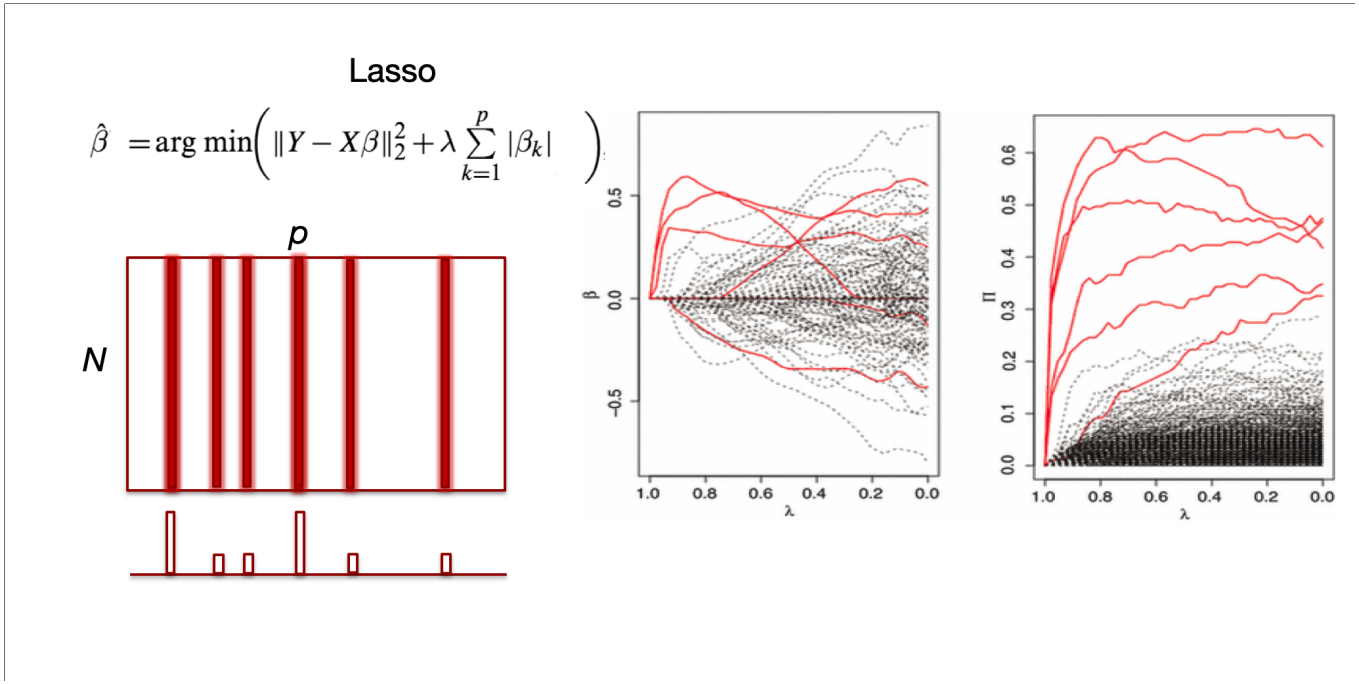
$$\hat{\beta}^{\lambda} = \arg\min_{\beta \in \mathbb{R}^p} \left(\|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} |\beta_k| \right)$$



Meinshausen and Buehlmann (2010)

Stability selection

- It is important to recognise that coefficient stability is largely independent from accuracy
- **stability selection** is a method that aims to identify truly important variables by examining stability under perturbation to the data



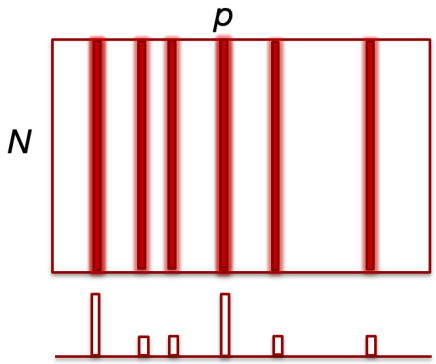
Meinshausen and Buehlmann (2010)

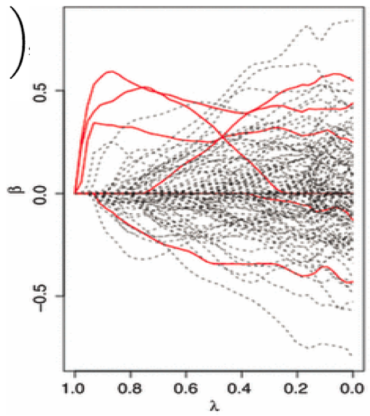
Stability selection

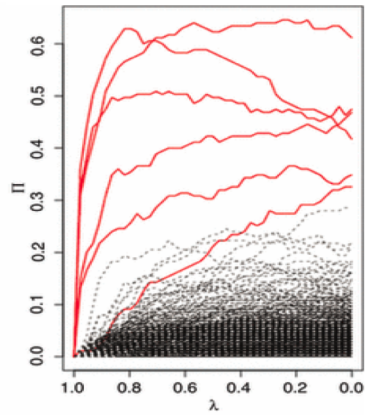
- It is important to recognise that coefficient stability is largely independent from accuracy
- **stability selection** is a method that aims to identify truly important variables by examining stability under perturbation to the data

Lasso

$$\hat{\beta} = \arg\min\left(\|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} |\beta_k| \right)$$





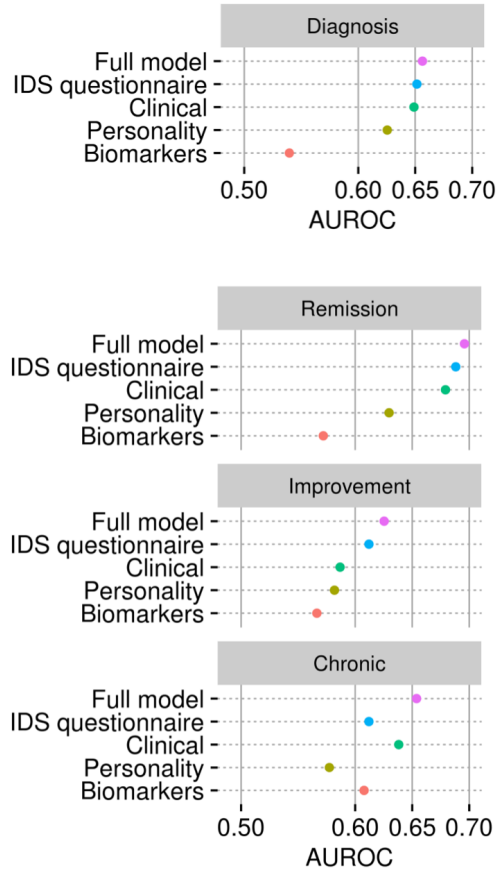


Statistical guarantees against false positives
(controls family-wise error)

Meinshausen and Buehlmann (2010)

Stability selection in major depression

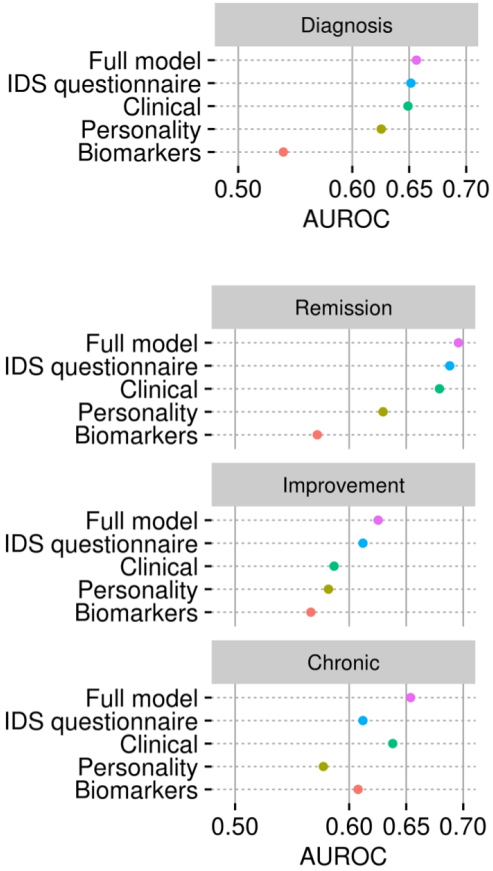
Aim: to identify prognostic markers for depression from an extensive panel (cognition, symptoms, biomarkers...)

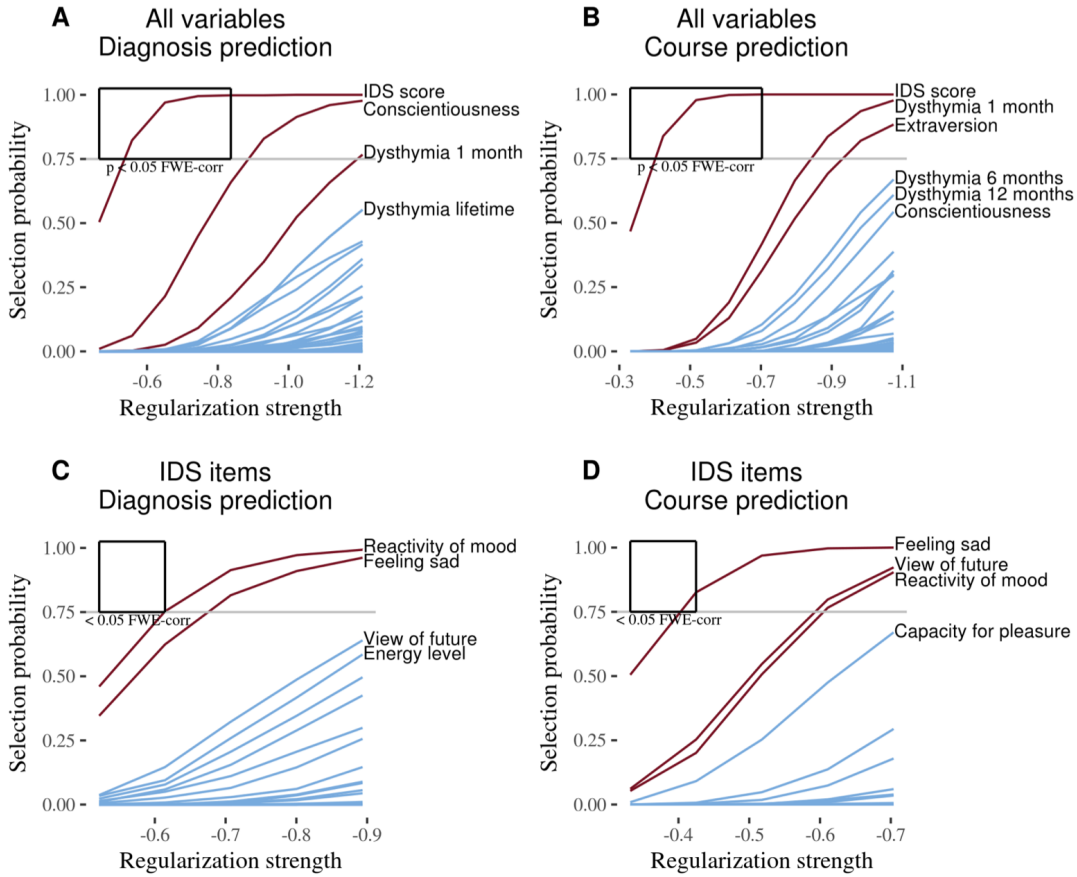


Dinga et al. (2018)

Stability selection in major depression

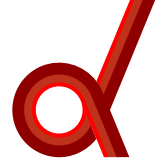
Aim: to identify prognostic markers for depression from an extensive panel (cognition, symptoms, biomarkers...)



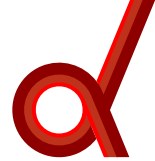


Dinga et al. (2018)

Outline



Take home messages



- Machine learning provides powerful tools for single subject inference and detect spatially distributed effects
- Many different approaches beyond simple notions such as ‘classification’ or ‘clustering’

Take home messages

- Machine learning provides powerful tools for single subject inference and detect spatially distributed effects
- Many different approaches beyond simple notions such as 'classification' or 'clustering'

Recommendations

- Linear models are often sufficient: they are fast, interpretable and often perform as well as non-linear methods
- Careful validation is extremely important for all methods to guard agains overfitting
- Machine learning can be easily integrated with neurocognitive models (e.g. to assess candidate models)

References I

Leon Aksman, David J. Lythgoe, Steven C.R. Williams, Martha Jokisch, Christoph Moenninghoff, Johannes Streffer, Karl Heinz Joeckel, Christian Weimar, and Andre F. Marquand. Making use of longitudinal information in pattern recognition. *Human Brain Mapping*, 37(12):4385–4404, 2016. doi: 10.1002/hbm.23317. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23317.

James Cole, Rudra Poudel, Dimosthenis Tsagkasoullis, Matthan Caan, Claire Steves, Tim Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *Neuroimage*, (In Press), 2017.

Richard Dinga, Lianne Schmaal, Brenda W.J.H. Penninx, Marie Jose van Tol, Dick J. Veltman, Laura van Velzen, Maarten Mennes, Nic J.A. van der Wee, and Andre F. Marquand. Evaluating the evidence for biotypes of depression: Methodological replication and extension of drysdale et al. (2017). *NeuroImage: Clinical*, 22: 101796, 2019. ISSN 2213-1582. doi: https://doi.org/10.1016/j.nicl.2019.101796. URL http://www.sciencedirect.com/science/article/pii/S2213158219301469.

A. Dong, N. Honnorat, B. Gaonkar, and C. Davatzikos. Chimera: Clustering of heterogeneous disease effects via distribution matching of imaging patterns. *IEEE Transactions on Medical Imaging*, 35(2):612–621, Feb 2016. ISSN 0278-0062. doi: 10.1109/TMI.2015.2487423.

Andrew T Drysdale, Logan Grosenick, Jonathan Downar, Katharine Dunlop, Farrokh Mansouri, Yue Meng, Robert N Fetcho, Benjamin Zebley, Desmond J Oathes, Amit Etkin, Alan F Schatzberg, Keith Sudheimer, Jennifer Keller, Helen S Mayberg, Faith M Gunning, George S Alexopoulos, Michael D Fox, Alvaro Pascual-Leone, Henning U Voss, BJ Casey, Marc J Dubin, and Conor Liston. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, (23):28–38, 2017.

Maurizio Filippone, Andre Marquand, Camilla Blain, Steven Williams, Janaina Mourao-Miranda, and Mark Girolami. Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities. *Annals of Applied Statistics*, 6:1883–1905, 2012.

Edith Le Floch, Vincent Guillemot, Vincent Frouin, Philippe Pinel, Christophe Lalanne, Laura Trinchera, Arthur Tenenhaus, Antonio Moreno, Monica Zilbovicius, Thomas Bourgeron, Stanislas Dehaene, Bertrand Thirion, Jean-Baptiste Poline, and Edouard Duchesnay. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. *NeuroImage*, 63 (1):11 – 24, 2012. ISSN 1053-8119.

Stefan Haufe, Frank Meinecke, Kai Grgen, Sven Dhne, John-Dylan Haynes, Benjamin Blankertz, and Felix Biemann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87(0):96 – 110, 2014.

References II

Tong He, Ru Kong, Avram Holmes, Minh Nguyen, Mert Sabuncu, Simon Eickhoff, Danilo Bzdok, Jiashi Feng, and B. T. Thomas Yeo. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *BioRxiv*, 2019.

Amanda Kraha, Heather Turner, Kim Nimon, Linda Zientek, and Robin Henson. Tools to support interpreting multiple regression in the face of multicollinearity. *Frontiers in Psychology*, 3:44, 2012. ISSN 1664-1078. doi: 10.3389/fpsyg.2012.00044. URL https://www.frontiersin.org/article/10.3389/fpsyg.2012.00044.

Andre F. Marquand, Iead Rezek, Jan Buitelaar, and Christian F. Beckmann. Understanding heterogeneity in clinical cohorts using normative models: beyond case control studies. In Press, 2016.

Janaina Mourao-Miranda, David R. Hardoon, Tim Hahn, Andre F. Marquand, Steve C.R. Williams, John Shawe-Taylor, and Michael Brammer. Patient classification as an outlier detection problem: An application of the one-class support vector machine. *NeuroImage*, 58(3):793 – 804, 2011. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2011.06.042. URL http://www.sciencedirect.com/science/article/pii/S1053811911006872.

Stephen M Smith, Thomas E Nichols, Diego Vidaurre, Anderson M Winkler, Timothy E J Behrens, Matthew F Glasser, Kamil Ugurbil, Deanna M Barch, David C Van Essen, and Karla L Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. 18:1565–1567, 2015.

Cedric Xia, Zongming Ma, Rastko Ciric, Shi Gu, Richard F. Betzel, Antonia N. Kaczkurkin, Monica E. Calkins, Philip A. Cook, Angel Garcia de la Garza, Simon N. Vandekar, Zaixu Cui, Tyler M. Moore, David R. Roalf, Kosha Ruparel, Daniel H. Wolf, Christos Davatzikos, Ruben C. Gur, Raquel E. Gur, Russell T. Shinohara, Danielle S. Bassett, and Theodore D. Satterthwaite. Linked dimensions of psychopathology and connectivity in functional brain networks. *Nature Communications*, 9, 2018.

Xiuming Zhang, Elizabeth C. Mormino, Nanbo Sun, Reisa A. Sperling, Mert R. Sabuncu, and B. T. Thomas Yeo. Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in alzheimer's disease. *Proceedings of the National Academy of Sciences*, 113(42):E6535–E6544, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1611073113. URL https://www.pnas.org/content/113/42/E6535.