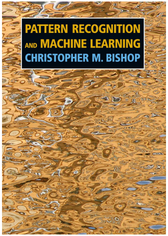
DONDERS INSTITUTE

Machine Learning I: basics

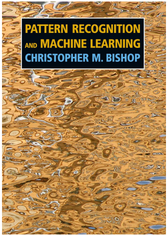
Charlotte Fraza
charlotte.fraza@donders.ru.nl



- "The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories."

Bishop (2006)

Machine Learning and Pattern Recognition



- "The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories."
- Pattern recognition has its origins in engineering, whereas machine learning grew out of computer science.
- **Automatically learn patterns**

Bishop (2006)

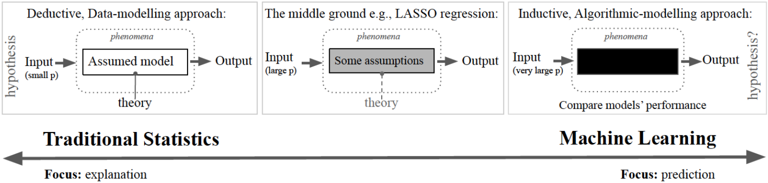


Figure 2: The continuum between traditional statistics and machine learning. p = set of possible predictors.

- "Traditional statistical approach; where the focus is on applying a theoretically conceived model to data and understanding the data generation mechanisms. ... Machine learning approach; where the focus is to make a prediction from the input to the output without assuming a theoretically motivated model.."

Lavelle-Hill et al. (2023)

What is pattern recognition used for?

Historically, has been applied in many application domains:

Example Applications

- Speech Recognition
- Automatic Character recognition / handwriting recognition
- Document classification (e.g. spam filters)
- Analysis of genetic microarray data
- Recommender systems / online shopping
- Natural language processing (NLP)
- ...

What is pattern recognition used for?

Increasingly used in computational psychiatry for:

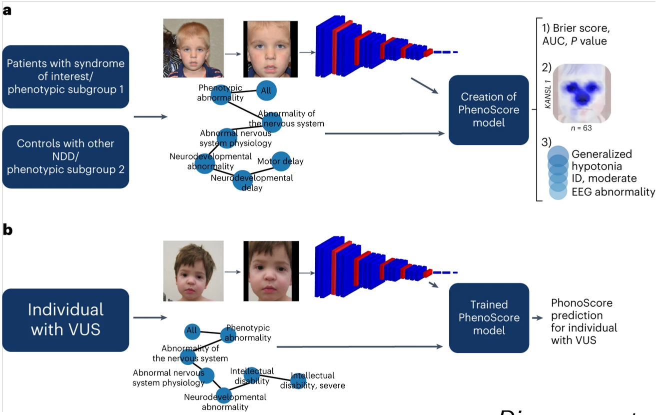
1. Predicting clinical variables (diagnosis or treatment response)
2. Stratifying psychiatric disorders
3. Learning mappings between behaviour and brain systems

Dingemans et al. (2023)

What is pattern recognition used for?

Increasingly used in computational psychiatry for:

1. Predicting clinical variables (diagnosis or treatment response)
2. Stratifying psychiatric disorders
3. Learning mappings between behaviour and brain systems



Dingemans et al. (2023)

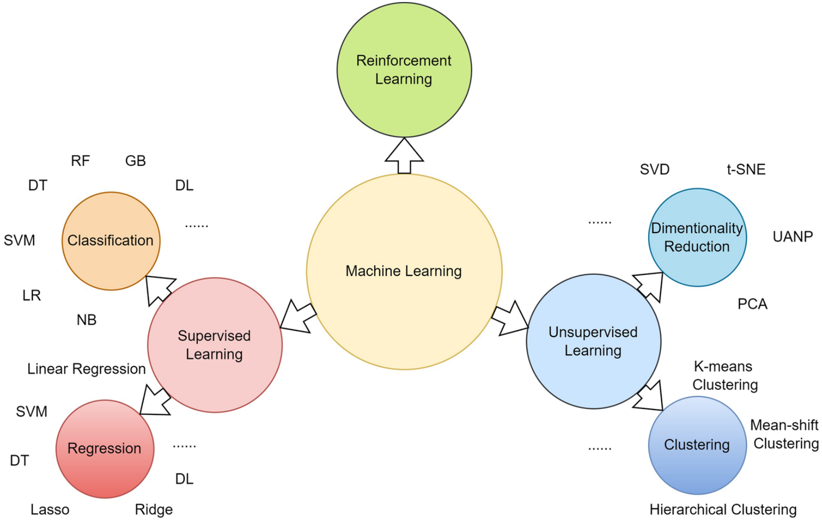
Outline

1. Introduction to Machine Learning
2. Basics of Pattern Recognition Analyses
3. Applications in Psychiatry
4. Conclusions

Outline

1. Introduction to Machine Learning
2. Basics of Pattern Recognition Analyses
3. Applications in Psychiatry
4. Conclusions

Types of pattern recognition



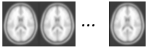
Wu et al. (2023)

Types of pattern recognition

Supervised learning involves learning a mapping between input and output:

Input

Training examples
$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$



No mathematical
Model available

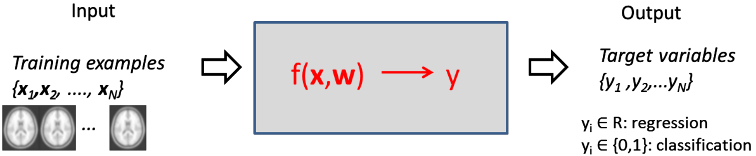
Output

Target variables
$\{y_1, y_2, \ldots y_N\}$

$y_i \in R$: regression
$y_i \in \{0,1\}$: classification

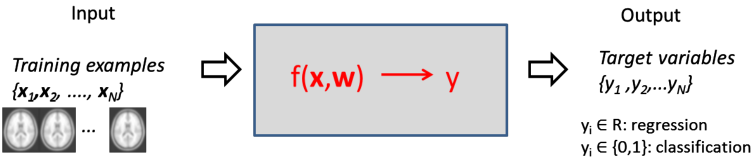
Types of pattern recognition

Supervised learning involves learning a mapping between input and output:

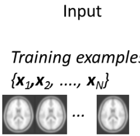


Types of pattern recognition

Supervised learning involves learning a mapping between input and output:

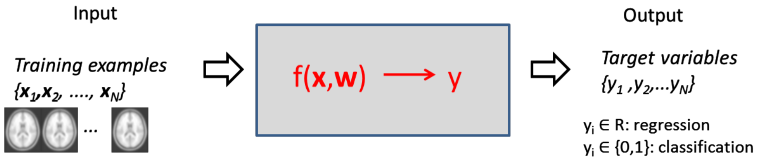


In **Unsupervised** learning, algorithms are not provided with output labels and must learn to structure the data by applying heuristics

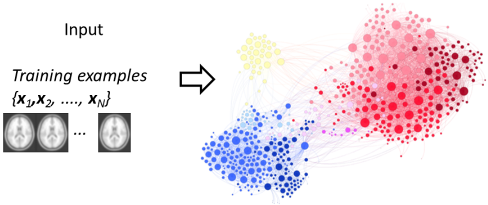


Types of pattern recognition

Supervised learning involves learning a mapping between input and output:



In **Unsupervised** learning, algorithms are not provided with output labels and must learn to structure the data by applying heuristics



Types of pattern recognition

Supervised learning involves learning a mapping between input and output:

Input

Training examples
{$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$}

$f(x,w) \longrightarrow y$

Output

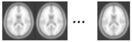
Target variables
{$y_1, y_2, \ldots y_N$}

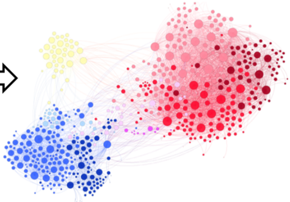
$y_i \in$ R: regression
$y_i \in \{0,1\}$: classification

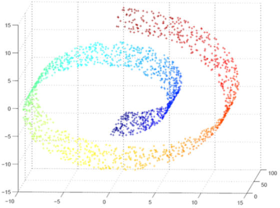
In **Unsupervised** learning, algorithms are not provided with output labels and must learn to structure the data by applying heuristics

Input

Training examples
{$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$}

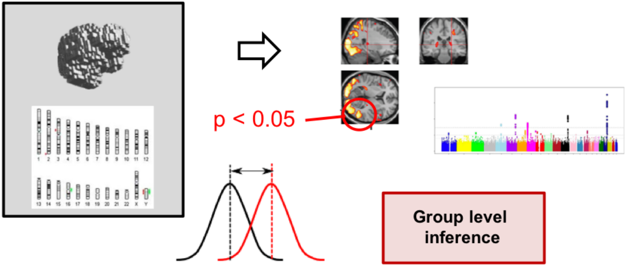




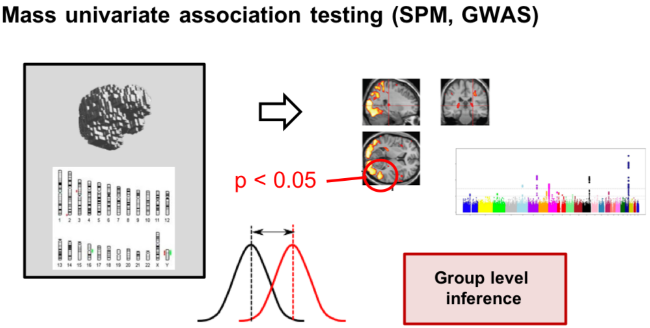


Biomarkers for prediction

Mass univariate association testing (SPM, GWAS)

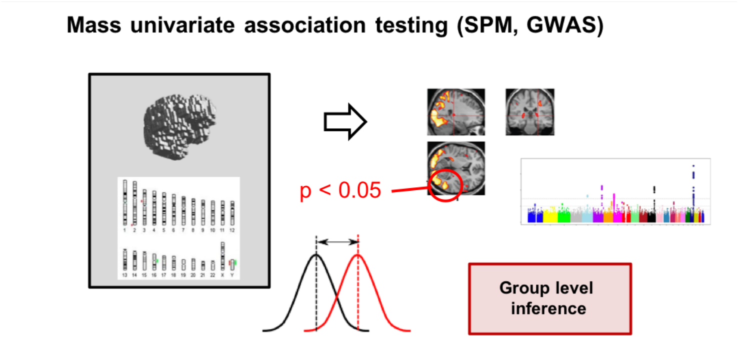


Biomarkers for prediction



- Useful for understanding mechanisms

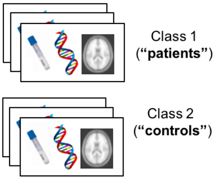
Biomarkers for prediction



- Useful for understanding mechanisms
- For clinical decision making this does not suffice. It is necessary to make predictions about individuals

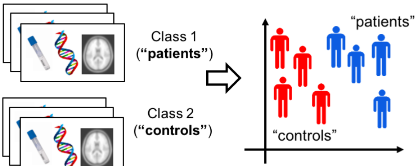
Predicting disease state

Making subject level predictions of diagnosis and outcome



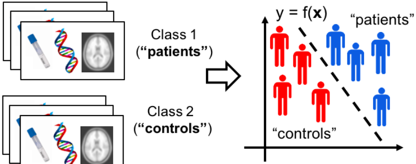
Predicting disease state

Making subject level predictions of diagnosis and outcome



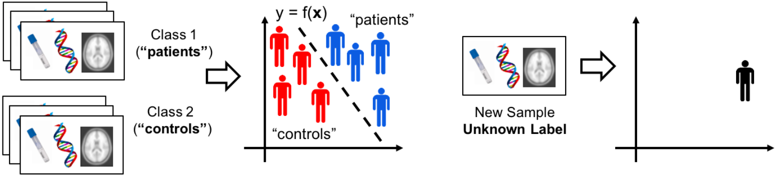
Predicting disease state

Making subject level predictions of diagnosis and outcome



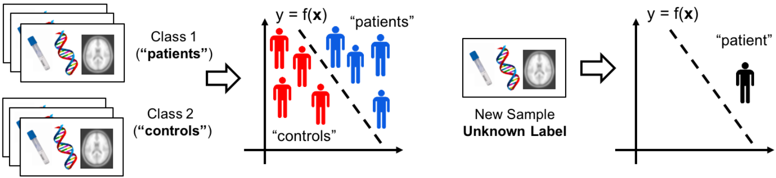
Predicting disease state

Making subject level predictions of diagnosis and outcome



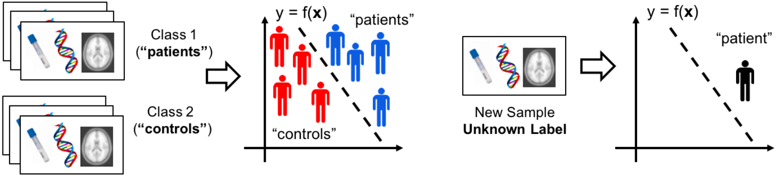
Predicting disease state

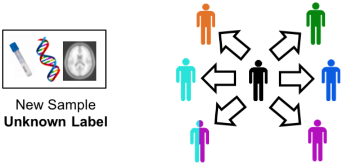
Making subject level predictions of diagnosis and outcome



Predicting disease state

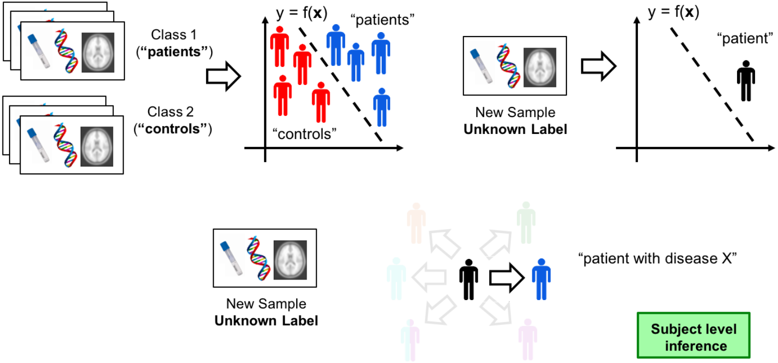
Making subject level predictions of diagnosis and outcome





Predicting disease state

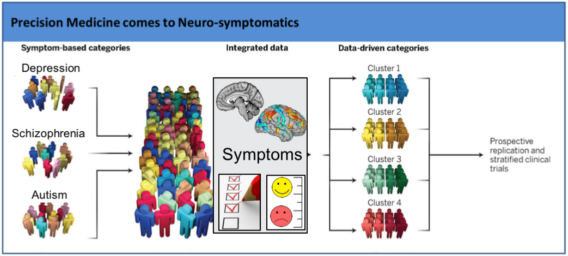
Making subject level predictions of diagnosis and outcome



Useful to find a measure of overall group separation

Machine Learning in Psychiatry: Stratification

Tackling the clinical and biological heterogeneity of psychiatric disorders

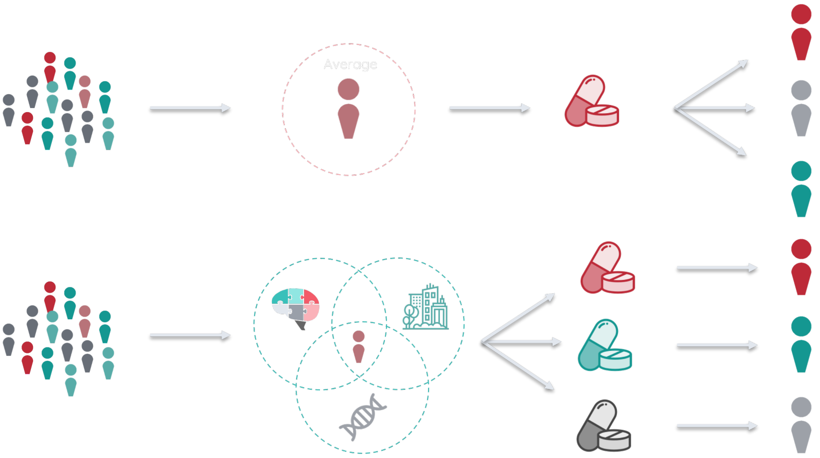


NIH National Institute of Mental Health

RDoC Research Domain Criteria Initiative

Insel and Cuthbert (2015)

Goal: Precision medicine



Outline

1 Introduction to Machine Learning

2 Basics of Pattern Recognition Analyses

3 Applications in Psychiatry

4 Conclusions

Encoding and Decoding

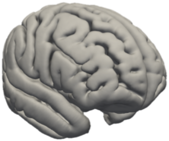
Cognitive state

- Response to stimulus
- Cognitive scores
- Diagnostic label
- Symptom scores
- ...

Encoding model

$$Y = f(X)$$

Brain activity (e.g. BOLD)



Naselaris et al. (2011)

Encoding and Decoding

Which brain regions are activated in a certain task condition?

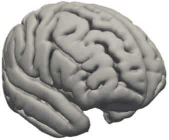
Cognitive state

- Response to stimulus
- Cognitive scores
- Diagnostic label
- Symptom scores
- ...

Encoding model

$$Y = f(X)$$

Brain activity (e.g. BOLD)



Naselaris et al. (2011)

Encoding and Decoding

Which brain regions are activated in a certain task condition?

Cognitive state

- Response to stimulus
- Cognitive scores
- Diagnostic label
- Symptom scores
- ...

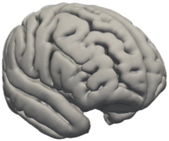
Encoding model

$$Y = f(X)$$

$$X = f(Y)$$

Decoding model

Brain activity (e.g. BOLD)



Naselaris et al. (2011)

Encoding and Decoding

Which brain regions are activated in a certain task condition?

Cognitive state

- Response to stimulus
- Cognitive scores
- Diagnostic label
- Symptom scores
- ...

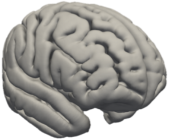
Encoding model

$Y = f(X)$

$X = f(Y)$

Decoding model

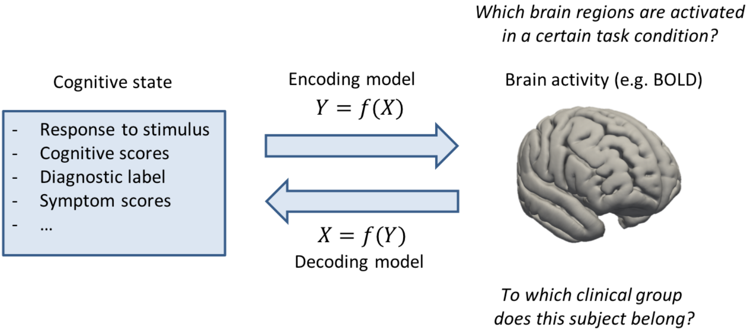
Brain activity (e.g. BOLD)



To which clinical group does this subject belong?

Naselaris et al. (2011)

Encoding and Decoding

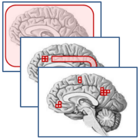


- Encoding models use stimuli to predict activity
- Decoding models use activity to predict information about the stimuli.

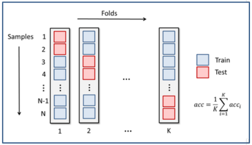
Naselaris et al. (2011)

Stages of supervised pattern recognition analysis

1. Feature extraction and/or feature selection

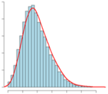


2. Classification / Regression using cross-validation



3. Performance evaluation

$$acc = \frac{1}{K}\sum_{i=1}^{K} acc_i$$



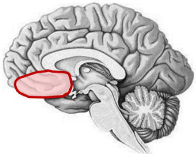
Feature selection and feature construction

Whole Brain

Feature selection and feature construction

Whole Brain

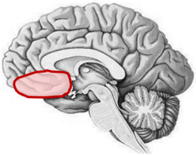
Region of Interest



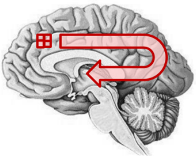
Feature selection and feature construction

Whole Brain

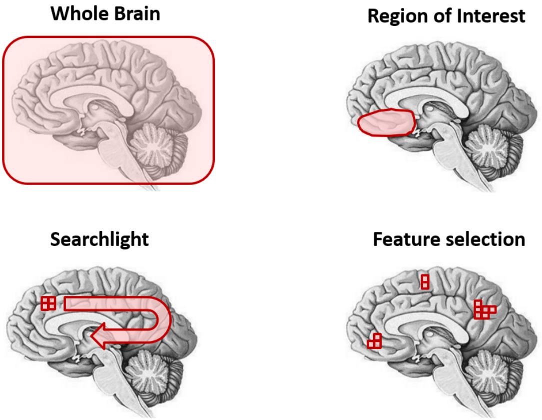
Region of Interest



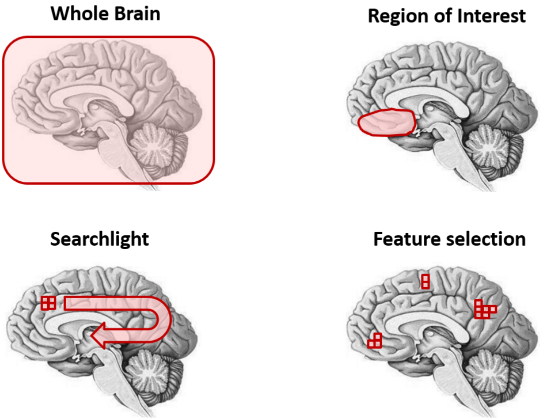
Searchlight



Feature selection and feature construction

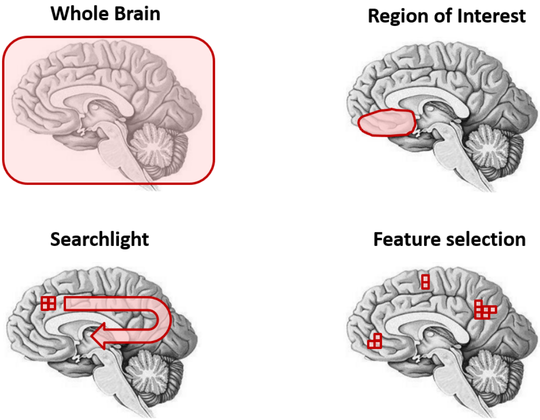


Feature selection and feature construction



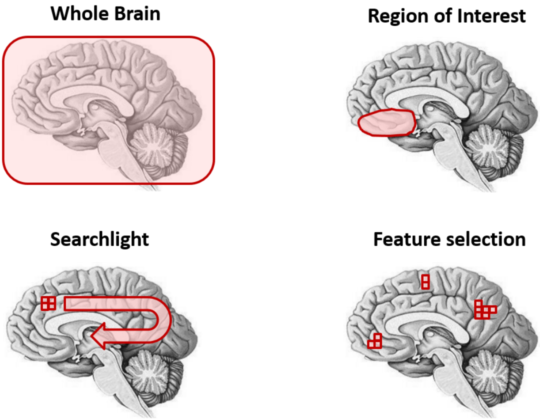
- Can also construct features (e.g. using ICA/ PCA,...)

Feature selection and feature construction



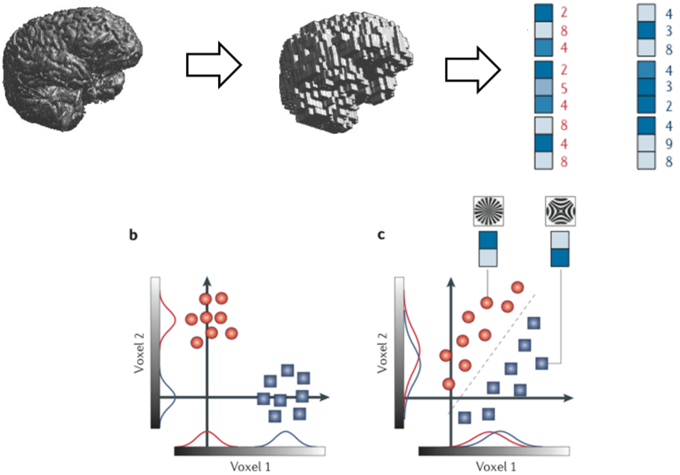
- Can also construct features (e.g. using ICA/ PCA,...)
- Or learn features from the data (e.g. deep learning)

Feature selection and feature construction



- Can also construct features (e.g. using ICA/ PCA,...)
- Or learn features from the data (e.g. deep learning)
- Feature selection should be performed on training data only!

Sensitivity for spatially distributed (or multivariate) effects:



Haynes and Rees (2006)

Notation

$$\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{y}\} \text{ or } \{\boldsymbol{X}, \boldsymbol{Y}\} \qquad \text{Dataset}$$

$$\boldsymbol{X}_{N\times D} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^T \qquad \text{N samples, D features}$$

$$\boldsymbol{y} = [y_1, ..., y_N]^T \qquad \text{Targets}$$

$$\boldsymbol{w} = [w_1, ..., w_D]^T \qquad \text{Weights}$$

Approaches to Supervised Pattern Recognition: Linear Regression

- Model Representation:

$$y_{\text{hat}} = w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_n \cdot x_n$$

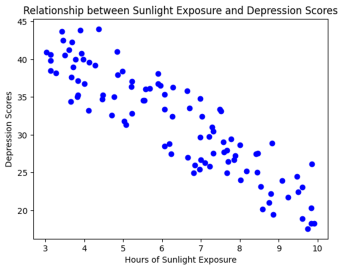
- Cost Function (Mean Squared Error):

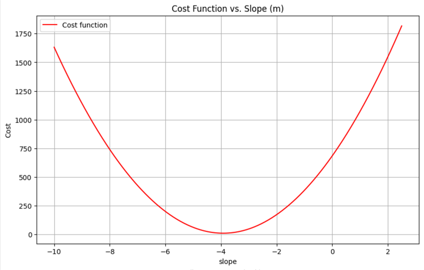
$$I(w) = \frac{1}{2m} \sum_{i=1}^{m} (y_{\text{hat}}^{(i)} - y^{(i)})^2$$

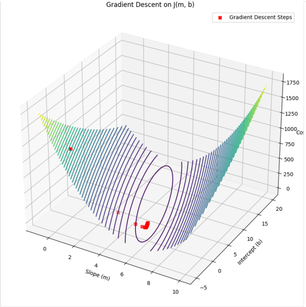
- Normal equation
- Optimization Update Rule (Gradient Descent):

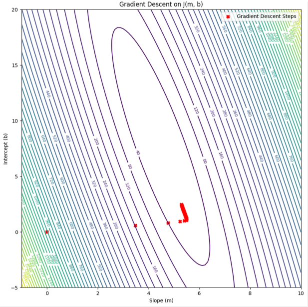
$$w_j := w_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (y_{\text{hat}}^{(i)} - y^{(i)}) \cdot x_j^{(i)}$$

Cost function intuition







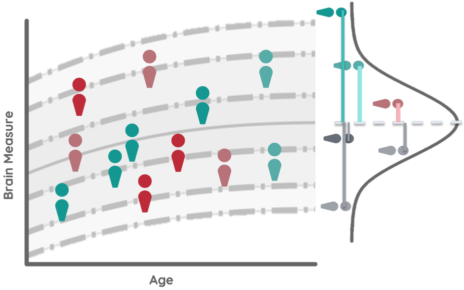


Bayesian Approach

- Bayesian models also depend on multiple variance/noise hyperparameters

$$p(\boldsymbol{w}|\boldsymbol{y},\theta,\sigma) = \frac{p(\boldsymbol{y}|\boldsymbol{w},\sigma)p(\boldsymbol{w}|\theta)}{p(\boldsymbol{y}|\theta,\sigma)}, \quad p(\boldsymbol{y}|\theta,\sigma) = \int p(\boldsymbol{y}|\boldsymbol{w},\sigma)p(\boldsymbol{w}|\theta)d\boldsymbol{w}$$

- Many approaches: nested CV, Empirical Bayes, MCMC ...



A smorgasbord of different approaches

- Regularisation methods (Penalized linear models, support vector machines, ...)
- Probabilistic approaches (Linear discriminant analysis, Gaussian processes, ...)
- Ensemble methods (Random forests, boosting, ...)
- Neural networks (multi-layer perceptrons, deep learning, ...)

A smorgasbord of different approaches

- Regularisation methods (Penalized linear models, support vector machines, ...)
- Probabilistic approaches (Linear discriminant analysis, Gaussian processes, ...)
- Ensemble methods (Random forests, boosting, ...)
- Neural networks (multi-layer perceptrons, deep learning, ...)

Most methods aim to trade-off data fit with complexity

$$f(\boldsymbol{x}_i, \boldsymbol{w}) = f_i = \boldsymbol{x}_i^T \boldsymbol{w}$$

A smorgasbord of different approaches

- Regularisation methods (Penalized linear models, support vector machines, ...)
- Probabilistic approaches (Linear discriminant analysis, Gaussian processes, ...)
- Ensemble methods (Random forests, boosting, ...)
- Neural networks (multi-layer perceptrons, deep learning, ...)

Most methods aim to trade-off data fit with complexity

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

A smorgasbord of different approaches

- Regularisation methods (Penalized linear models, support vector machines, ...)
- Probabilistic approaches (Linear discriminant analysis, Gaussian processes, ...)
- Ensemble methods (Random forests, boosting, ...)
- Neural networks (multi-layer perceptrons, deep learning, ...)

Most methods aim to trade-off data fit with complexity

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

$$\Rightarrow \hat{\boldsymbol{w}} = \min_{\mathbf{w}} - \ln p(\boldsymbol{w}|\boldsymbol{y})$$

Approaches to Supervised Pattern Recognition

A smorgasbord of different approaches

- Regularisation methods (Penalized linear models, support vector machines, ...)
- Probabilistic approaches (Linear discriminant analysis, Gaussian processes, ...)
- Ensemble methods (Random forests, boosting, ...)
- Neural networks (multi-layer perceptrons, deep learning, ...)

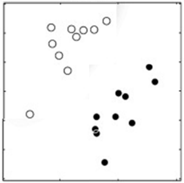
Most methods aim to trade-off data fit with complexity

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

$$\Rightarrow \hat{\boldsymbol{w}} = \min_{\mathbf{w}} - \ln p(\boldsymbol{w}|\boldsymbol{y})$$

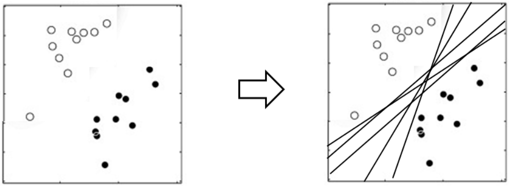
$$= \min_{\mathbf{w}} \sum_{i=1}^{n} \ln p(y_i|f_i) + \ln p(\boldsymbol{w}|\boldsymbol{\theta})$$

Choice of pattern recognition algorithm



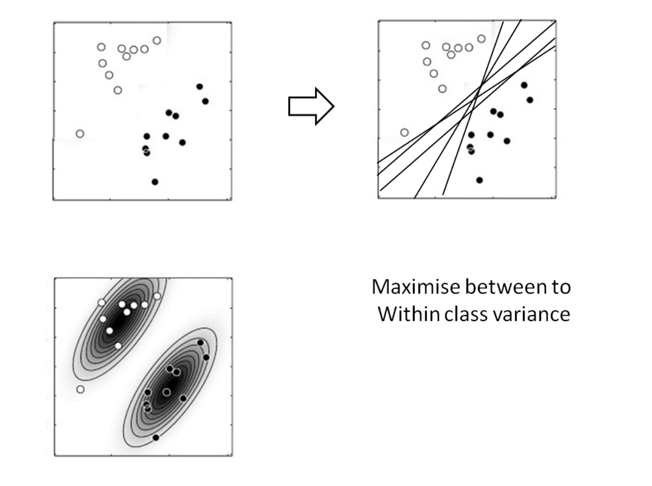
Ashburner and Klöppel (2011)

Choice of pattern recognition algorithm



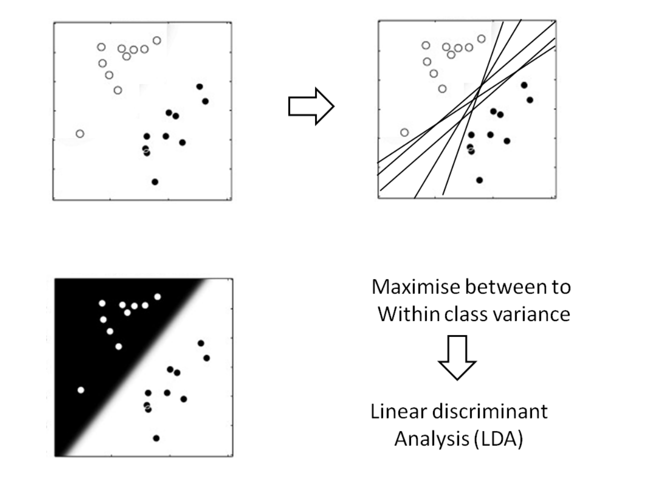
Ashburner and Klöppel (2011)

Choice of pattern recognition algorithm



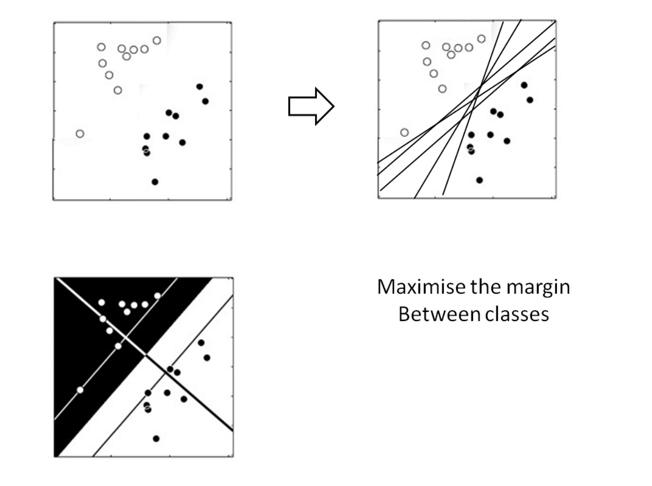
Ashburner and Klöppel (2011)

Choice of pattern recognition algorithm



Ashburner and Klöppel (2011)

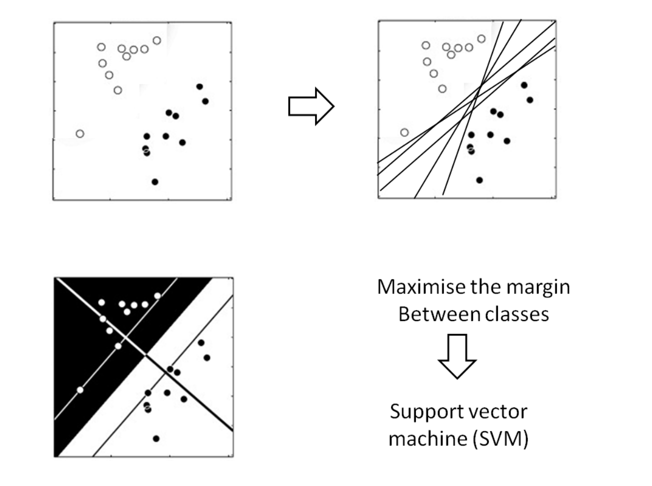
Choice of pattern recognition algorithm



Maximise the margin
Between classes

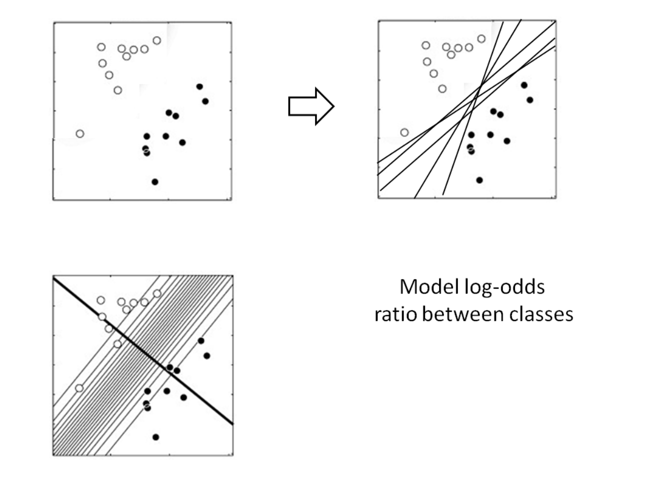
Ashburner and Klöppel (2011)

Choice of pattern recognition algorithm



Ashburner and Klöppel (2011)

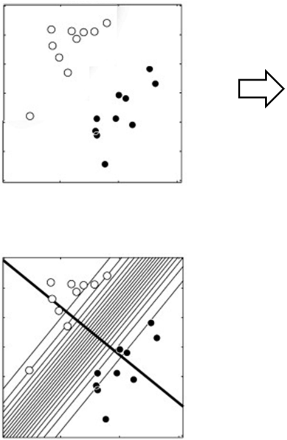
Choice of pattern recognition algorithm



Model log-odds
ratio between classes

Ashburner and Klöppel (2011)

Choice of pattern recognition algorithm

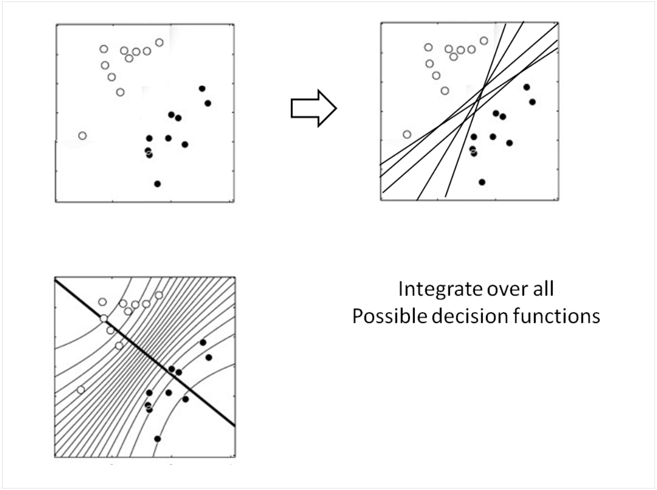


Model log-odds
ratio between classes

Logistic regression

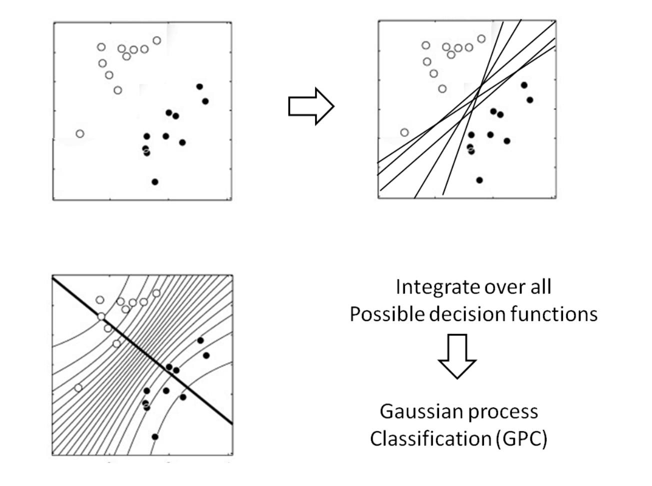
Ashburner and Klöppel (2011)

Choice of pattern recognition algorithm



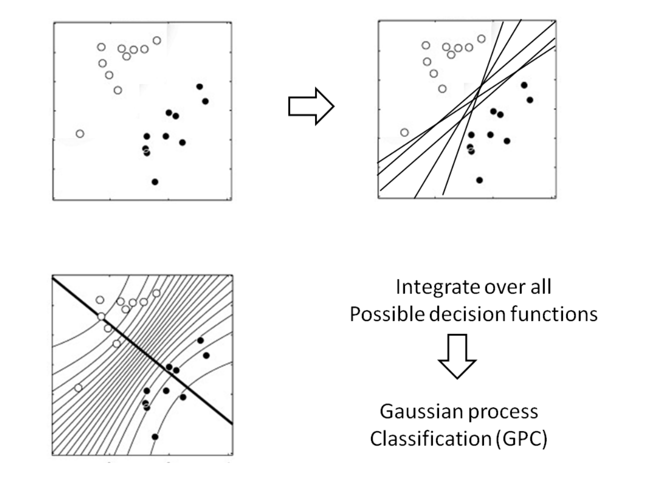
Ashburner and Klöppel (2011)

Choice of pattern recognition algorithm



Ashburner and Klöppel (2011)

Choice of pattern recognition algorithm

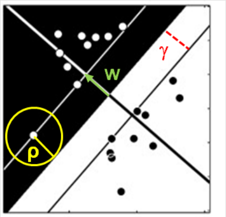


All methods make assumptions!

Ashburner and Klöppel (2011)

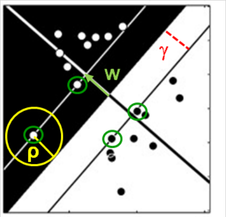
Support Vector Machines

- Finds a separating hyperplane that is "optimal" in that it leads to the largest margin between classes (γ)
- Based on the assumption that each point is bounded by unknown noise (ρ)
- New points will be well classified if $\gamma > \rho$
- The hyperplane is uniquely defined by a subset of the most ambiguous data points, making it sparse ("support vectors")

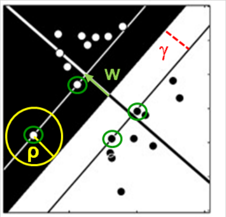


Support Vector Machines

- Finds a separating hyperplane that is "optimal" in that it leads to the largest margin between classes (γ)
- Based on the assumption that each point is bounded by unknown noise (ρ)
- New points will be well classified if $\gamma > \rho$
- The hyperplane is uniquely defined by a subset of the most ambiguous data points, making it sparse ("support vectors")



- Finds a separating hyperplane that is "optimal" in that it leads to the largest margin between classes (γ)
- Based on the assumption that each point is bounded by unknown noise (ρ)
- New points will be well classified if $\gamma > \rho$
- The hyperplane is uniquely defined by a subset of the most ambiguous data points, making it sparse ("support vectors")



$$\min_{w,\xi,b} \quad -\gamma + C\sum_{i=1}^{N} \xi_i$$

$$\text{s.t.:} \quad y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}) + b) > \gamma - \xi_i$$

$$\xi_i > 0$$

$$||\boldsymbol{w}||^2 = 1$$

Deep Learning

- ‘Deep’ neural networks have seen an enormous surge in popularity over the last few years
- Extend 1950s-era neural networks to have many hidden layers
- Now provide state of the art performance in many domains, e.g. computer vision, game playing and perception

LETTER

doi:10.1038/nature14236

Human-level control through deep reinforcement learning

Volodym... Martin I... Helen K...

ARTICLE

doi:10.1038/nature16961

Mastering the game of Go with deep neural networks and tree search

David Silv... Julian Sch... John Nhar... Thore Gra...

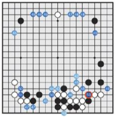
npj | Digital Medicine

www.nature.com/npjdigitalmed

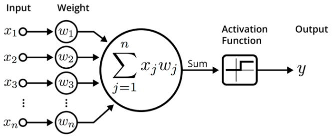
ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

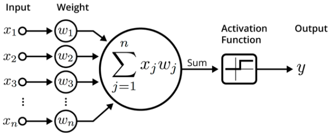
Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], 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]

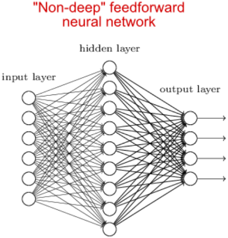


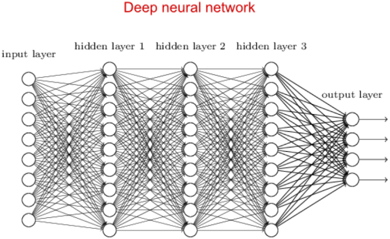
Deep Learning



Deep Learning







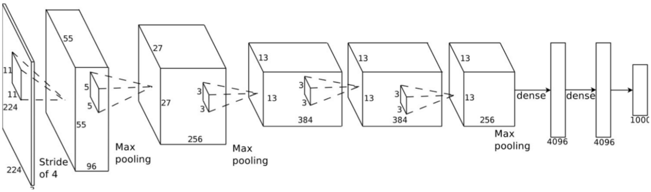
- Activation functions introduce non-linearity (Sigmoid (Logistic) Function)
- Predominantly supervised learning
- Many parameters to optimise (more in lecture 2)

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca



- 7 layer CNN network that won 2012 ImageNet large-scale visual recognition challenge by 10% (AlexNet)
- Trained the network on 15 million annotated images from over 22,000 categories

1. **Data Bias (Western, Educated, Industrialized, Rich, and Democratic (WEIRD))**
2. **Black Box: lack of transparency**
3. **Overfitting: perform poorly on unseen data**

Henrich et al. (2010)

1. **Data Bias (Western, Educated, Industrialized, Rich, and Democratic (WEIRD))**
2. **Black Box: lack of transparency**
3. **Overfitting: perform poorly on unseen data**
4. **High Complexity: resource-intensive and hard to interpret**
5. **Adversarial Attacks: small changes in data can cause large prediction errors**
6. **Ethical Concerns**

Henrich et al. (2010)

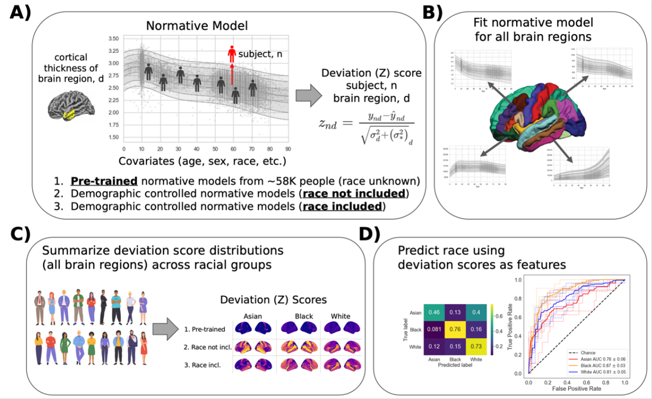
Dangers of Deep Learning

1. **Data Bias (Western, Educated, Industrialized, Rich, and Democratic (WEIRD))**
2. **Black Box: lack of transparency**
3. **Overfitting: perform poorly on unseen data**
4. **High Complexity: resource-intensive and hard to interpret**
5. **Adversarial Attacks: small changes in data can cause large prediction errors**
6. **Ethical Concerns**
7. **Environmental Impact**
8. **Data Privacy**

Henrich et al. (2010)

Data Bias

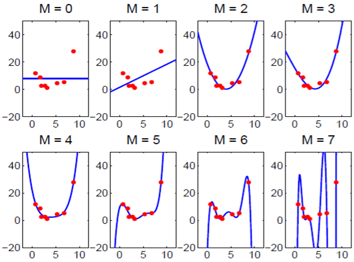
- Persistent racial biases are present in brain models, even when race is accounted for, suggesting that more flexible and representative modeling techniques are necessary.

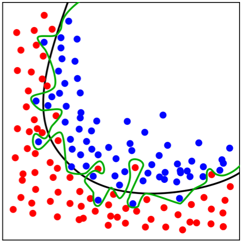


Rutherford et al. (2024)

Underfitting and Overfitting

- Occurs when an model performs well on the data that it was estimated or trained on, but poorly on new data
- Can arise in very many ways including improper parameter optimisation or feature selection

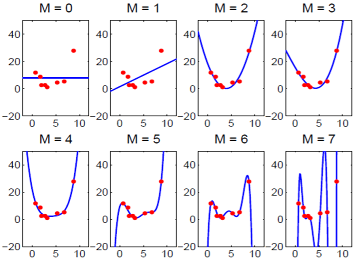


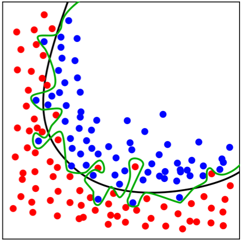


Bishop (2006)

Solutions Underfitting and Overfitting

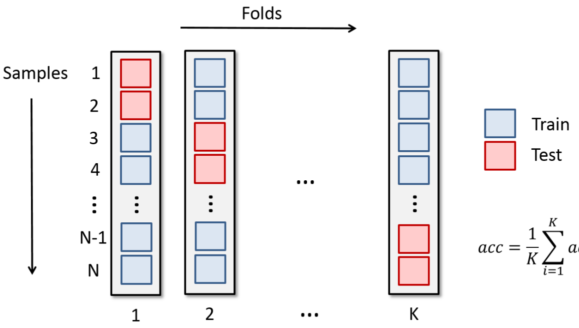
1. Increase/decrease model complexity
2. Feature selection
3. Regularization
4. Early stopping
5. Data augmentation
6. Cross-validation





Cross-validation

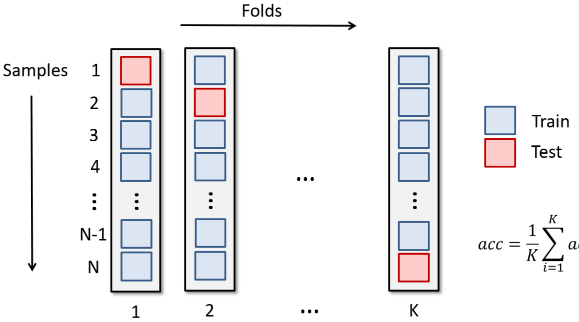
- Testing on unseen data is essential to assess generalizability (hold-out set)
- Cross-validation one popular way to do this
- 'K-fold CV': split the data into K approximately equal chunks
- 'Leave-one-out': one sample is left out at at a time ($K = N$)



$$acc = \frac{1}{K}\sum_{i=1}^{K} acc_i$$

Cross-validation

- Testing on unseen data is essential to assess generalizability (hold-out set)
- Cross-validation one popular way to do this
- 'K-fold CV': split the data into K approximately equal chunks
- 'Leave-one-out': one sample is left out at at a time ($K = N$)



$$acc = \frac{1}{K}\sum_{i=1}^{K} acc_i$$

Parameter optimisation

- Most approaches depend on multiple (hyper)parameters
- e.g. regularization parameters in penalized linear models

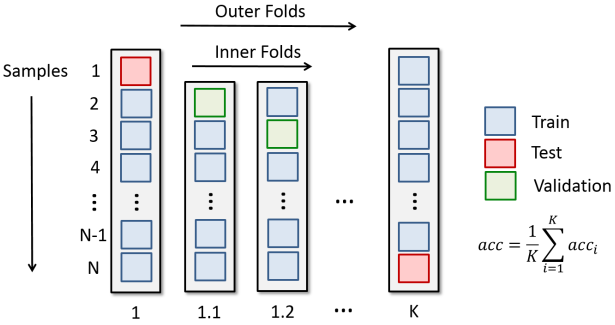
$$\hat{\boldsymbol{w}} = \min_w \sum_{i=1}^{n} \ell(y_i, f_i) + \lambda J(\boldsymbol{w})$$

Parameter optimisation

- Most approaches depend on multiple (hyper)parameters
- e.g. regularization parameters in penalized linear models

$$\hat{\boldsymbol{w}} = \min_w \sum_{i=1}^{n} \ell(y_i, f_i) + \lambda J(\boldsymbol{w})$$

- Standard approach is *nested* cross-validation with a grid search



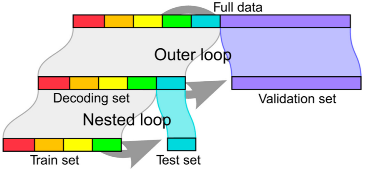
Multi-stage validation

- CV can be overly optimistic
- Data Leakage during Preprocessing; Hyperparameter Tuning Bias; Data Imbalance;

Varoquaux et al. (2017)

Multi-stage validation

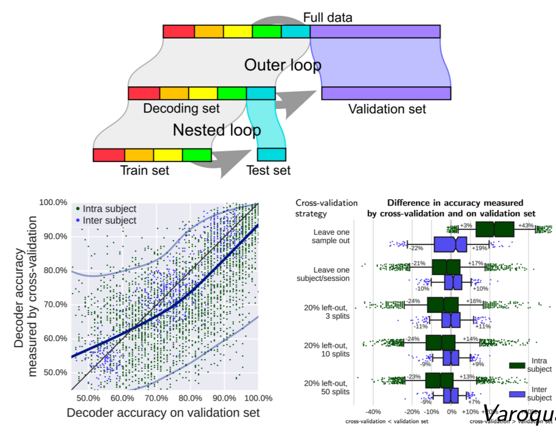
- CV can be overly optimistic
- Data Leakage during Preprocessing; Hyperparameter Tuning Bias; Data Imbalance;
- A multistage validation approach protects against this
- CV also invalidates parametric tests (more later)



Varoquaux et al. (2017)

Multi-stage validation

- CV can be overly optimistic
- Data Leakage during Preprocessing; Hyperparameter Tuning Bias; Data Imbalance;
- A multistage validation approach protects against this
- CV also invalidates parametric tests (more later)



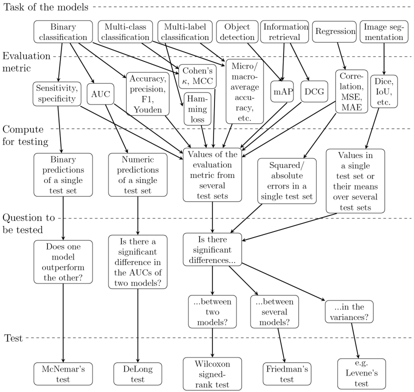
Varoquaux et al. (2017)

1. Choice of error measure

Rainio et al. (2024)

Statistical assessment of model performance

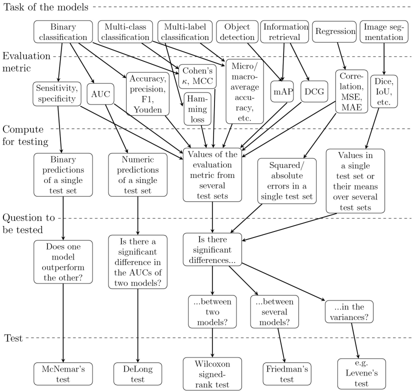
1 Choice of error measure



Rainio et al. (2024)

Statistical assessment of model performance

1. Choice of error measure



- Different error metrics are sensitive to different aspects (e.g. MSE depends on the scale of the data)
- **Accuracy ≠ Clinical utility**

Rainio et al. (2024)

Choice of error metric

2. Statistical testing framework. There are various options:
   - Parametric tests (e.g. binomial test, t-test)
   - Randomization tests (permutation, bootstrapping)

Stelzer et al. (2013); Winkler et al. (2015)

Choice of error metric

2. Statistical testing framework. There are various options:
   - Parametric tests (e.g. binomial test, t-test)
   - Randomization tests (permutation, bootstrapping)

- Cross-validation induces dependency between the folds invalidating parametric statistics

Stelzer et al. (2013); Winkler et al. (2015)

Choice of error metric

2. Statistical testing framework. There are various options:
    - Parametric tests (e.g. binomial test, t-test)
    - Randomization tests (permutation, bootstrapping)

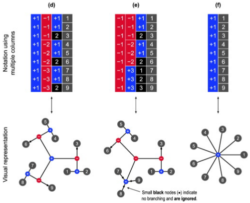
- Cross-validation induces dependency between the folds invalidating parametric statistics
- Parametric assumptions may not be met (e.g. interval data)

Stelzer et al. (2013); Winkler et al. (2015)

Choice of error metric

2. Statistical testing framework. There are various options:
    - Parametric tests (e.g. binomial test, t-test)
    - Randomization tests (permutation, bootstrapping)

- Cross-validation induces dependency between the folds invalidating parametric statistics
- Parametric assumptions may not be met (e.g. interval data)
- Permutation tests must respect *exchangeability*, e.g. site effects, family structure ...



Stelzer et al. (2013); Winkler et al. (2015)

Outline

1. Introduction to Machine Learning
2. Basics of Pattern Recognition Analyses
3. Applications in Psychiatry
4. Conclusions

Supervised learning for automated diagnosis and prognosis

Neuroscience and Biobehavioral Reviews 57 (2015) 328–349

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

ELSEVIER

Review

From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics

Thomas Wolfers[a,b,*], Jan K. Buitelaar[c,d], Christian F. Beckmann[b,c,e], Barbara Franke[a,f], Andre F. Marquand[b,g]

NeuroImage 145 (2017) 137–165

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

ELSEVIER

Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls

Mohammad R. Arbabshirani[a,b,*], Sergey Plis[a], Jing Sui[a,c], Vince D. Calhoun[a,d]

nature neuroscience

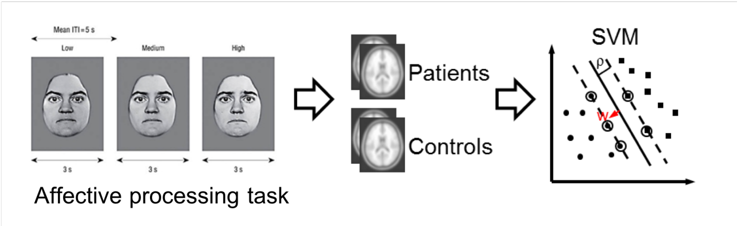
Building better biomarkers: brain models in translational neuroimaging

Choong-Wan Woo[1–4], Luke J Chang[5], Martin A Lindquist[6] & Tor D Wager[3,4]

Wolfers et al. (2015); Arbabshirani et al. (2017); Woo et al. (2017)

Supervised learning for depression diagnosis

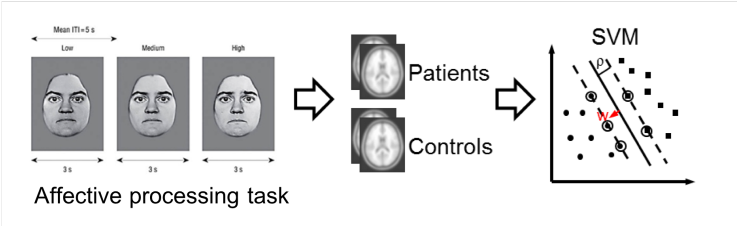
This study was an early application of Pattern recognition to predict disease state in major depression



Fu et al. (2008)

Supervised learning for depression diagnosis

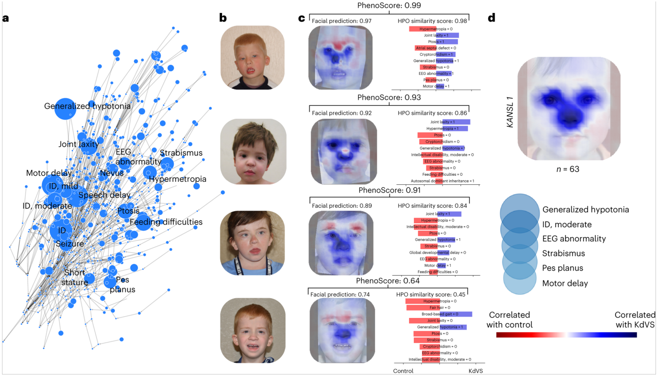
This study was an early application of Pattern recognition to predict disease state in major depression



- Patients could be discriminated from controls with 87% accuracy
- Patients who responded well to fluoxetine could be discriminated from non-responders with 67% accuracy

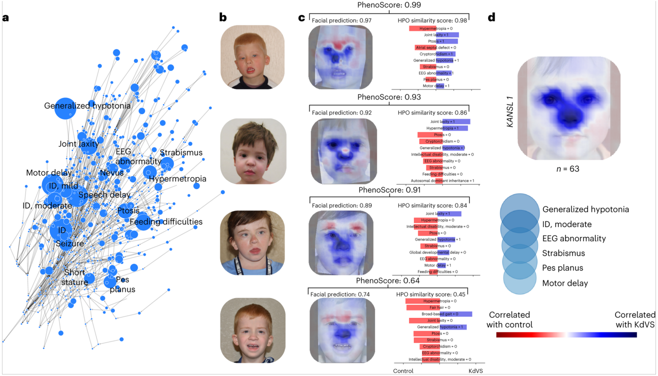
Fu et al. (2008)

Supervised learning for rare genetic diseases



Dingemans et al. (2023)

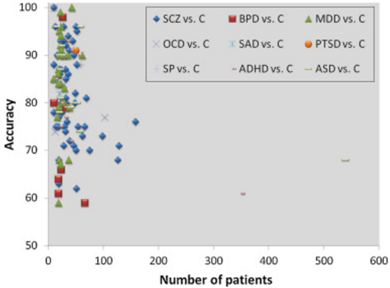
Supervised learning for rare genetic diseases



- Personalized medicine by automatically recognizing distinct phenotypic subtypes leading to more tailored clinical prognosis

Dingemans et al. (2023)

Where are we now?



- 116 studies
- Most disorders
- Many data modalities

- Moderate acccuracy, highly variable across studies
- Mostly small samples, minimal validation across cohorts
- Accuracy in small samples is extremely variable
- **Heterogeneity** is a major challenge in clinical cohorts

Wolfers et al. (2015)

Subtyping psychiatric disorders

PNAS

Distinct neuropsychological subgroups in typically developing youth inform heterogeneity in children with ADHD

Damien A. Fair[a,b,c,1], Deep

Departments of [a]Behavioral Neur
97239; [d]Department of Computer
University of Iowa, Iowa City, IA 5

The American Journal of Psychiatry

Current Issue | Archive | About | Residents' Journal | AJP In Advance | Podcast | CME | Author Resources

Back to table of contents

Previous Next

Articles

Full Access

Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers

Brett A. Clementz, Ph.D.,
Godfrey D. Pearlson, M.D., M

Published Online: 7 Dec 2

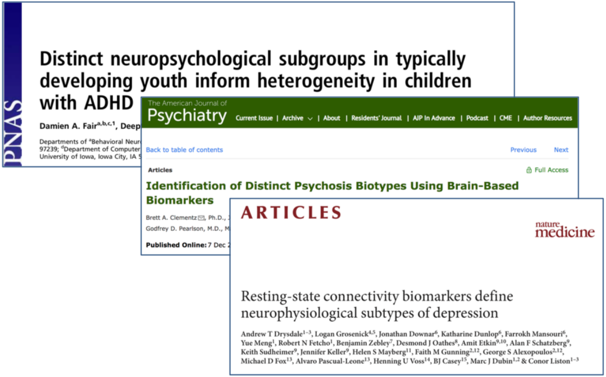
ARTICLES

nature medicine

Resting-state connectivity biomarkers define neurophysiological subtypes of depression

Andrew T Drysdale[1–3], Logan Grosenick[4,5], Jonathan Downar[6], Katharine Dunlop[6], Farrokh Mansouri[6], Yue Meng[1], Robert N Fetcho[1], Benjamin Zebley[7], Desmond J Oathes[8], Amit Etkin[9,10], Alan F Schatzberg[9], Keith Sudheimer[9], Jennifer Keller[9], Helen S Mayberg[11], Faith M Gunning[2,12], George S Alexopoulos[2,12], Michael D Fox[13], Alvaro Pascual-Leone[13], Henning U Voss[14], BJ Casey[15], Marc J Dubin[1,2] & Conor Liston[1–3]

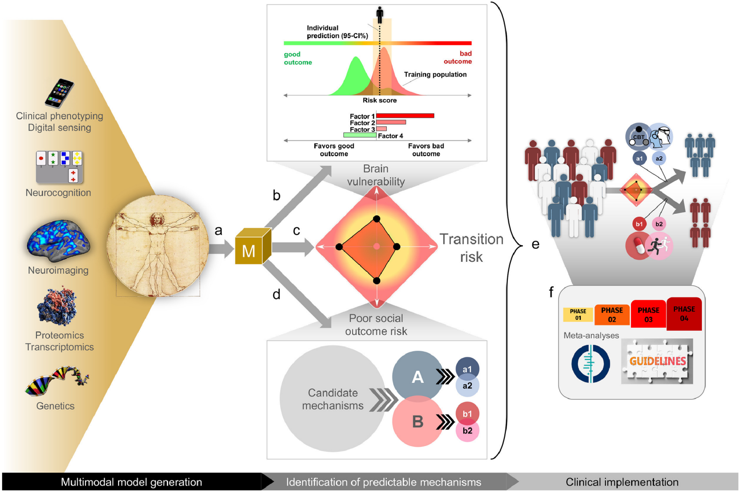
Subtyping psychiatric disorders



Validation of clusters is difficult:

- Clustering always gives a result and there is no clear measure of success (e.g. stability? separability? predictive ability?)
- Rarely test against the 'null' hypothesis that there are no clusters in the data
- Clustering using symptoms may not map onto biology

Ethical considerations for precision psychiatry

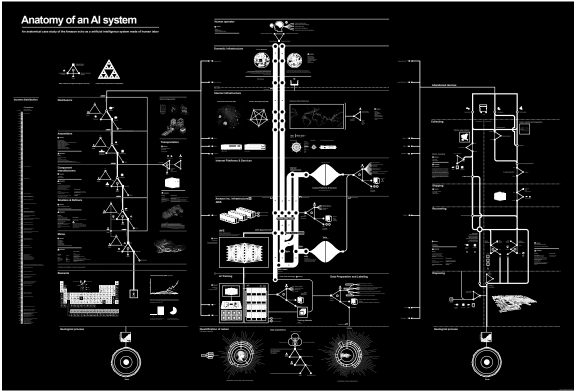


Fusar-Poli et al. (2022)

Ethical Challenges

- **Interpretability**: Balancing model complexity with interpretability for clinical decision-making.
- **Overreliance**: Avoid overreliance on algorithmic predictions.
- **Long-term Impact**: Consider potential psychological and societal impacts of diagnosis and treatment based on machine learning.
- **Equity**: Ensure equal access to computational psychiatry tools.

Ethical Challenges



- 'Put simply: each small moment of convenience – be it answering a question, turning on a light, or playing a song – requires a vast planetary network, fueled by the extraction of non-renewable materials, labor, and data.'

https://anatomyof.ai/

Ethical Challenges

‘While ‘off the shelf’ machine learning tools, like TensorFlow, are becoming more accessible from the point of view of setting up your own system, the underlying logics of those systems, and the datasets for training them are accessible to and controlled by very few entities. In the dynamic of dataset collection through platforms like Facebook, users are feeding and training the neural networks with behavioral data, voice, tagged pictures and videos or medical data. In an era of extractivism, the real value of that data is controlled and exploited by the very few at the top of the pyramid.’

https://anatomyof.ai/

Outline

1 Introduction to Machine Learning

2 Basics of Pattern Recognition Analyses

3 Applications in Psychiatry

4 Conclusions

Interim summary

- PR is a powerful tool to perform single subject inference and detect spatially distributed effects
- Useful in clinical neuroscience for:
  1. Making predictions at the subject level (e.g. prognosis)
  2. Stratifying psychiatric disorders
  3. Estimating mappings between brain and behaviour
- **Accuracy $\neq$ Clinical utility**

References I

Mohammad R. Arbabshirani, Sergey Plis, Jing Sui, and Vince D. Calhoun. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145:137 – 165, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2016.02.079. Individual Subject Prediction.

J. Ashburner and S. Klöppel. Multivariate models of inter-subject anatomical variability. *NeuroImage*, 56(2): 422–439, 2011.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition, August 2006. ISBN 0387310738. URL `http://www.worldcat.org/isbn/0387310738`.

Alexander JM Dingemans, Max Hinne, Kim MG Truijen, Lia Goltstein, Jeroen Van Reeuwijk, Nicole De Leeuw, Janneke Schuurs-Hoeijmakers, Rolph Pfundt, Illja J Diets, Joery Den Hoed, et al. Phenoscore quantifies phenotypic variation for rare genetic diseases by combining facial analysis with other clinical features using a machine-learning framework. *Nature Genetics*, 55(9):1598–1607, 2023.

C. H. Fu, J. Mourao-Miranda, S. G. Costafreda, A. Khanna, A. F. Marquand, S. C. Williams, and M. J. Brammer. Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. *Biol Psychiatry*, 63(7):656–62, 2008.

Paolo Fusar-Poli, Mirko Manchia, Nikolaos Koutsouleris, David Leslie, Christiane Woopen, Monica E Calkins, Michael Dunn, Christophe Le Tourneau, Miia Mannikko, Tineke Mollema, et al. Ethical considerations for precision psychiatry: a roadmap for research and clinical practice. *European Neuropsychopharmacology*, 63: 17–34, 2022.

J. D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nat Rev Neurosci*, 7(7):523–34, 2006.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

Thomas R Insel and Bruce N Cuthbert. Brain disorders? precisely. *Science*, 348(6234):499–500, 2015.

R. Lavelle-Hill, G. Smith, and K. Murayama. Machine learning meets psychological data: Challenges and future directions. October 30 2023. URL `https://doi.org/10.31219/osf.io/6xt82`. Preprint on OSF.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Oona Rainio, Jarmo Teuho, and Riku Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086, 2024.

References II

Saige Rutherford, Thomas Wolfers, Charlotte Fraza, Nathaniel G Harrnet, Christian F Beckmann, Henricus G Ruhe, and Andre F Marquand. To which reference class do you belong? measuring racial fairness of reference classes with normative modeling. *arXiv preprint arXiv:2407.19114*, 2024.

Johannes Stelzer, Yi Chen, and Robert Turner. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *NeuroImage*, 65:69–82, 2013.

Gael Varoquaux, Pradeep Reddy Raamana, Denis A. Engemann, Andres Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145:166 – 179, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2016.10.038. Individual Subject Prediction.

Anderson M. Winkler, Matthew A. Webster, Diego Vidaurre, Thomas E. Nichols, and Stephen M. Smith. Multi-level block permutation. *NeuroImage*, 123:253 – 268, 2015. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2015.05.092.

T. Wolfers, J. K. Buitelaar, C. F. Beckmann, B. Franke, and A. F. Marquand. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience and Biobehavioral Reviews*, in press, 2015.

Choong-Wan Woo, Luke J Chang, Martin A Lindquist, and Tor D Wager. Building better biomakers: brain models in translational neuroimaging. *Nature Neuroscience*, (20):365–377, 2017.

Yuqi Wu, Kaining Mao, Liz Dennett, Yanbo Zhang, and Jie Chen. Systematic review of machine learning in ptsd studies for automated diagnosis evaluation. *npj Mental Health Research*, 2(1):16, 2023.