Brain imaging, machine learning and imaging biomarkers

www.jussitohka.net

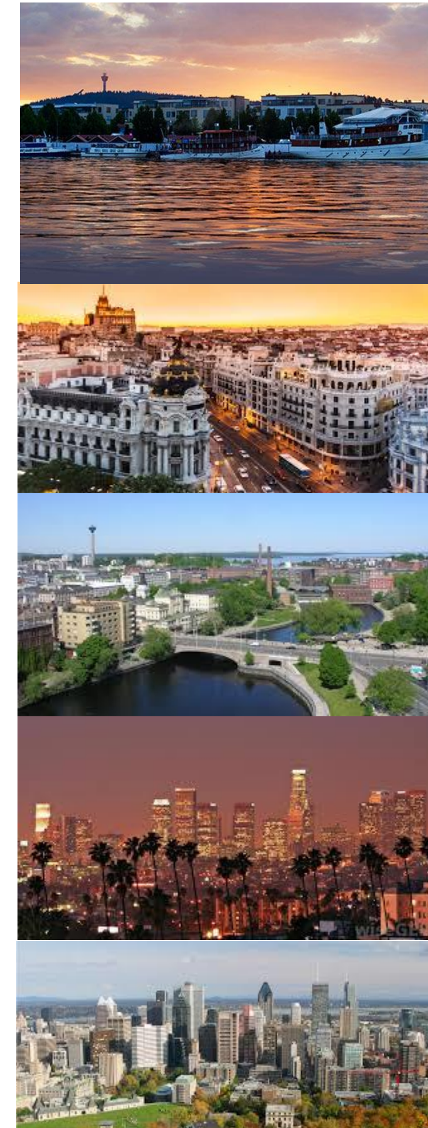
17th August 2017

Jussi Tohka

UEF // University of Eastern Finland

About me

Currently, Associate professor at AI Virtanen Institute, University of Eastern Finland
2015 – 2016, CONEX professor at Universidad Carlos III de Madrid, Spain
2009 – 2014, Academy research fellow, team leader, Tampere University of Technology
2005 – 2009, Senior researcher (Academy post-doc, Scientific coordinator of STATCORE research cluster of excellence) Tampere University of Technology
2004 – 2005, Post-doc, Laboratory of Neuro Imaging, UCLA, USA
1999 – 2003, PhD in signal processing, Tampere, including the first of several visits to Montreal Neurological Institute



Lecture plan

Session 1a brain imaging basics

Session 1b: Brain image analysis, or features from brain imaging

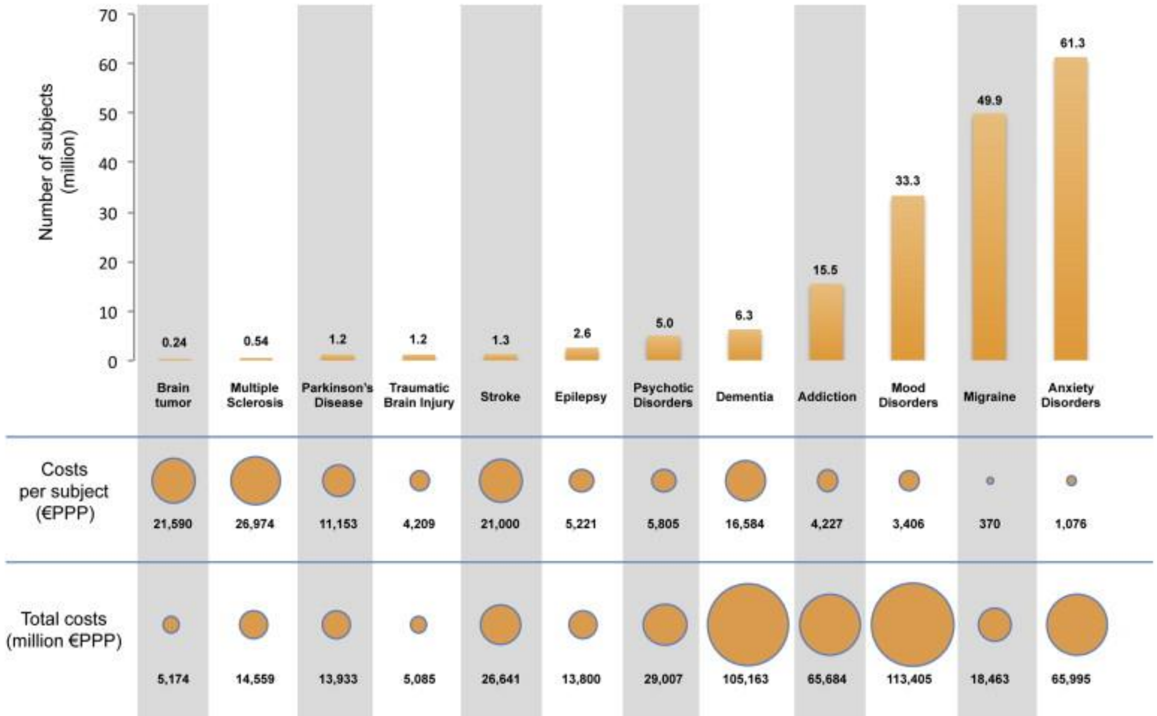
Session 2 a: Machine learning for brain imaging biomarkers (diagnosis, prognosis)

Session 2b: Small-sample considerations

Session 2c: Machine learning for image segmentation

Brain imaging basics

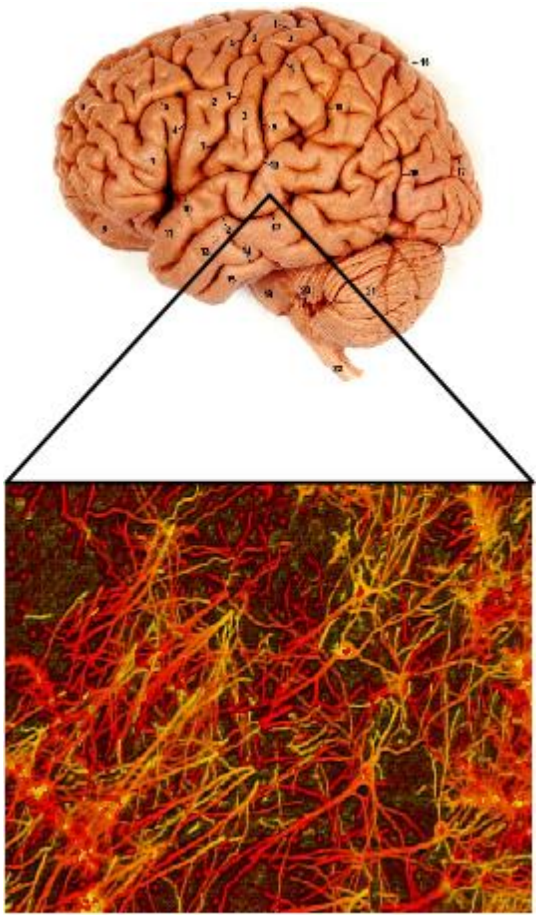
Prevalence and cost of brain disorders in Europe



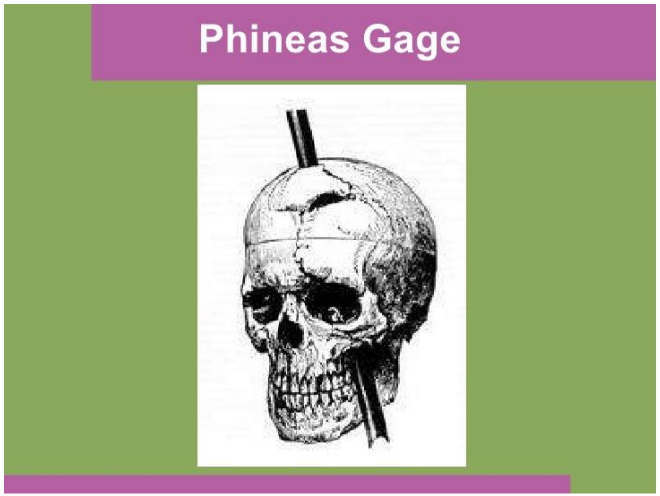
DiLuca, Olesen 2014

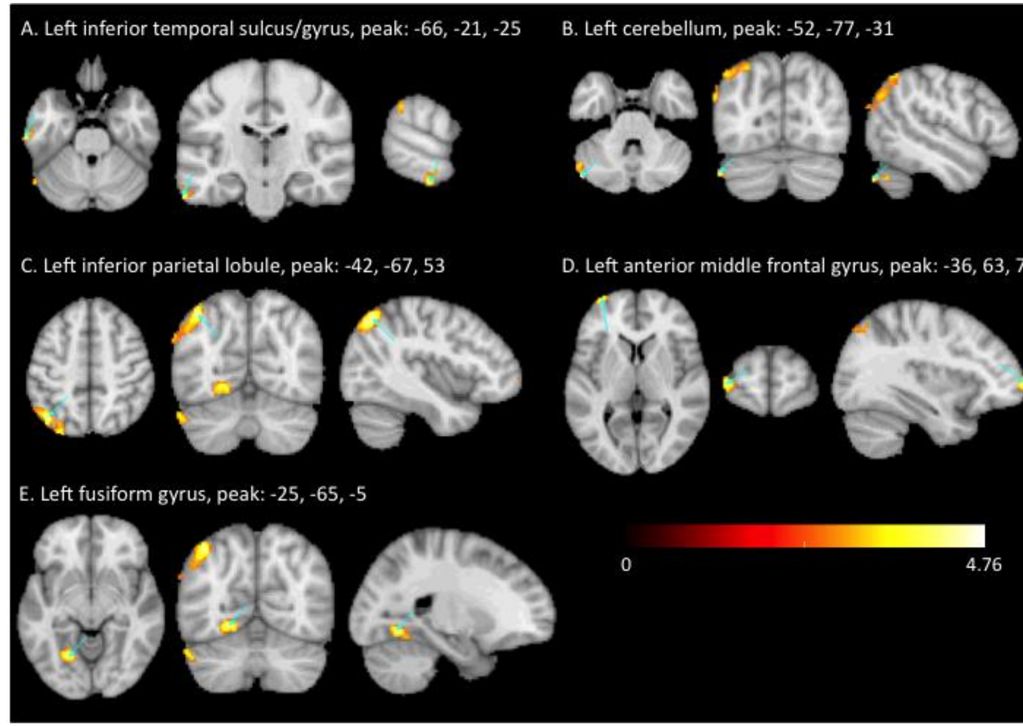
The brain

- The human brain is enormously complex in its spatial organization with thousands of anatomically distinct subdivisions and in terms of complex connections between subdivisions
- Brain structure and function are characterized by individual variation from one subject to another and variation in a single individual over time.
- 3-D imaging is the only technique to directly study the brain structure and function in living humans



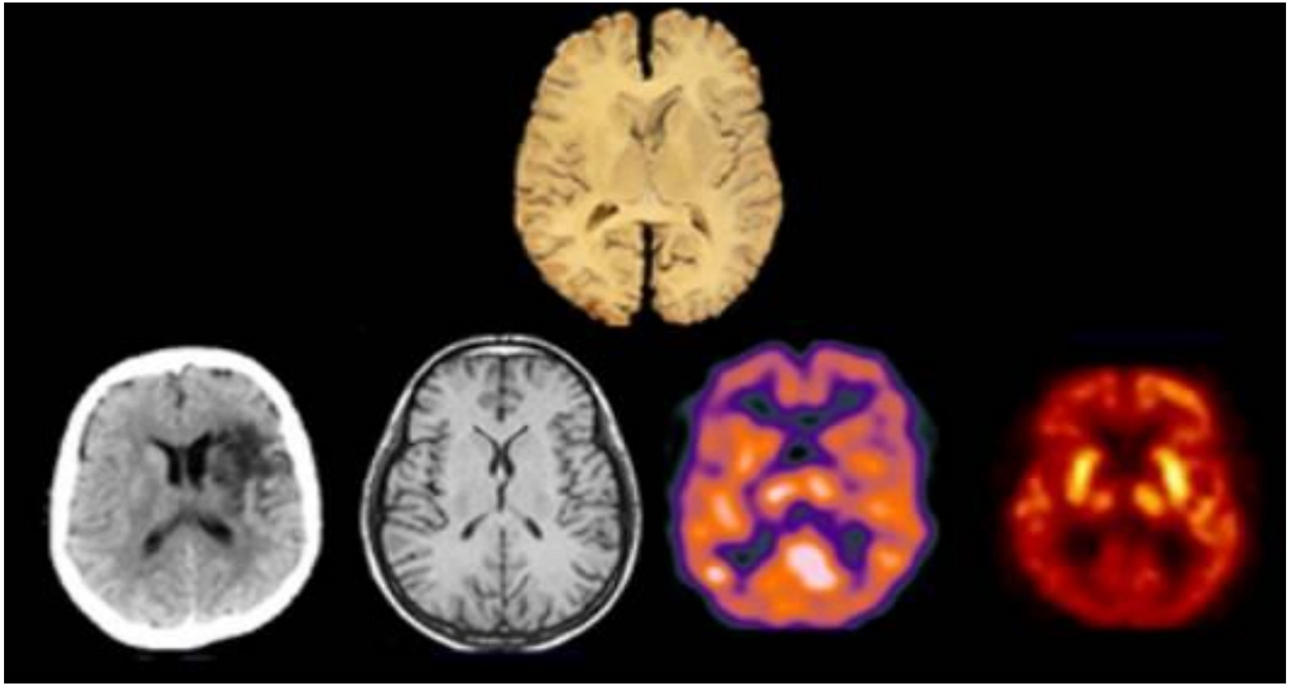
Brain research then and now



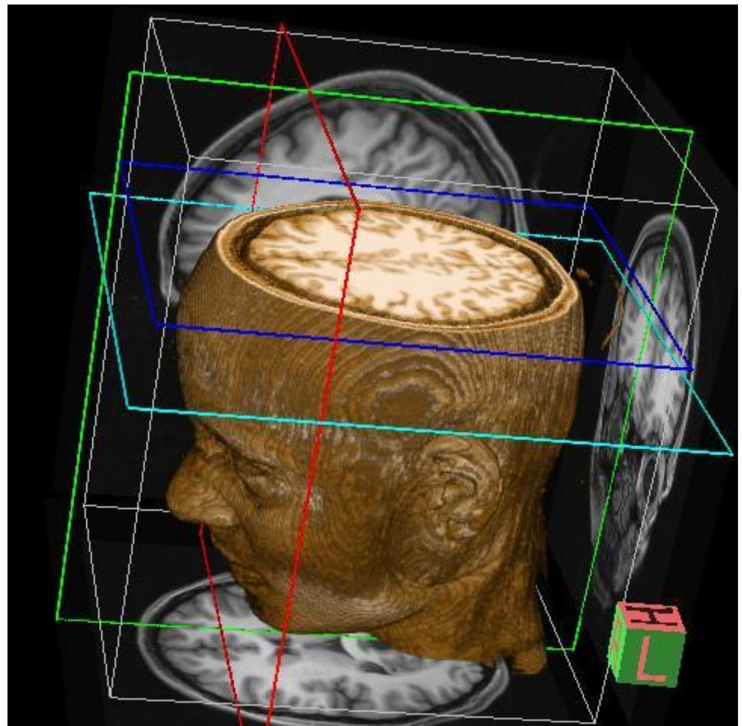


How self-rated humourness relates to brain activity?
Jääskeläinen et al, Sci Rep 2016.

Brain imaging provides information about the structure and function of living brain

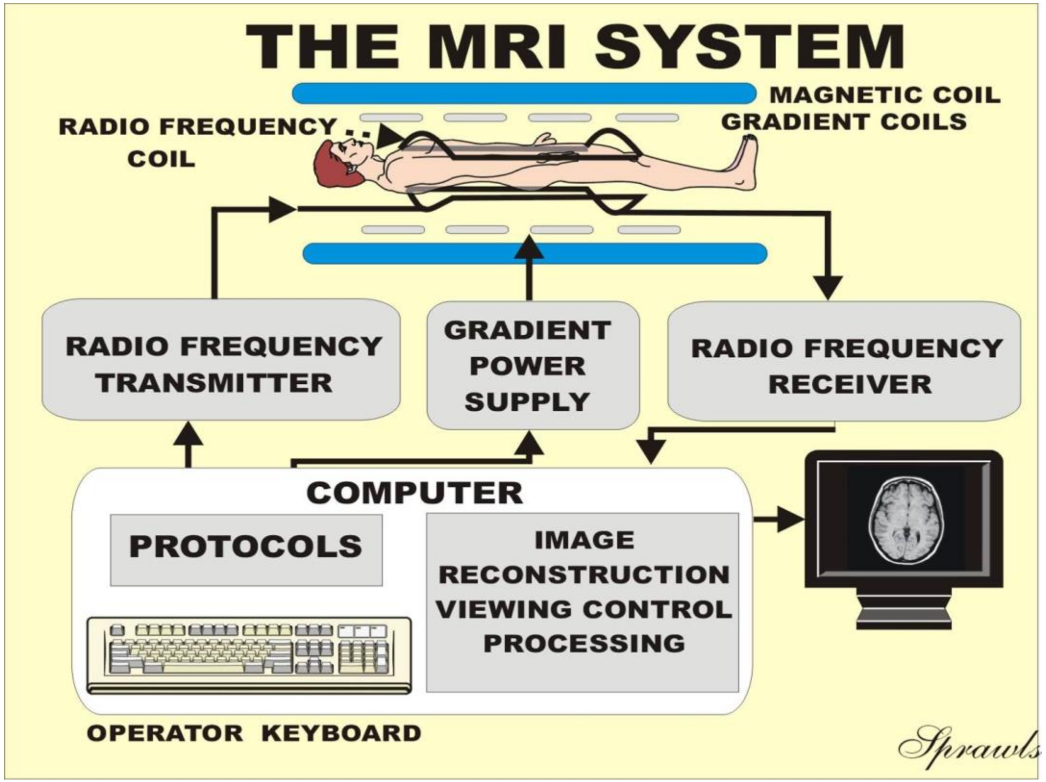


Most brain images are 3-dimensional



http://unlimitedcomputing.no/wp-content/uploads/2015/04/3Dbrain.jpg

Magnetic resonance imaging

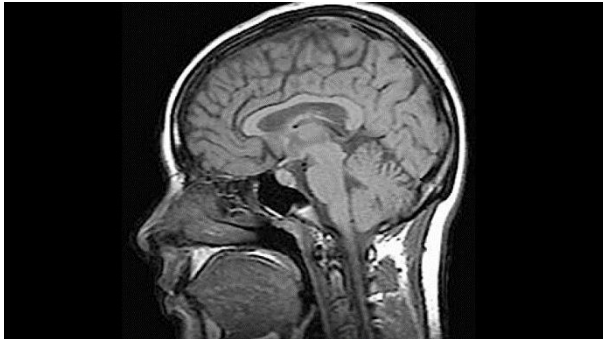


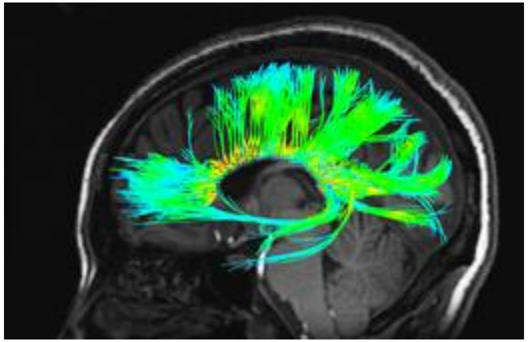
- The magnetic resonance image is a display of radio frequency signal intensities that are emitted by magnetized tissue during the imaging process.

http://www.sprawls.org/mripmt/MRI01/index.html

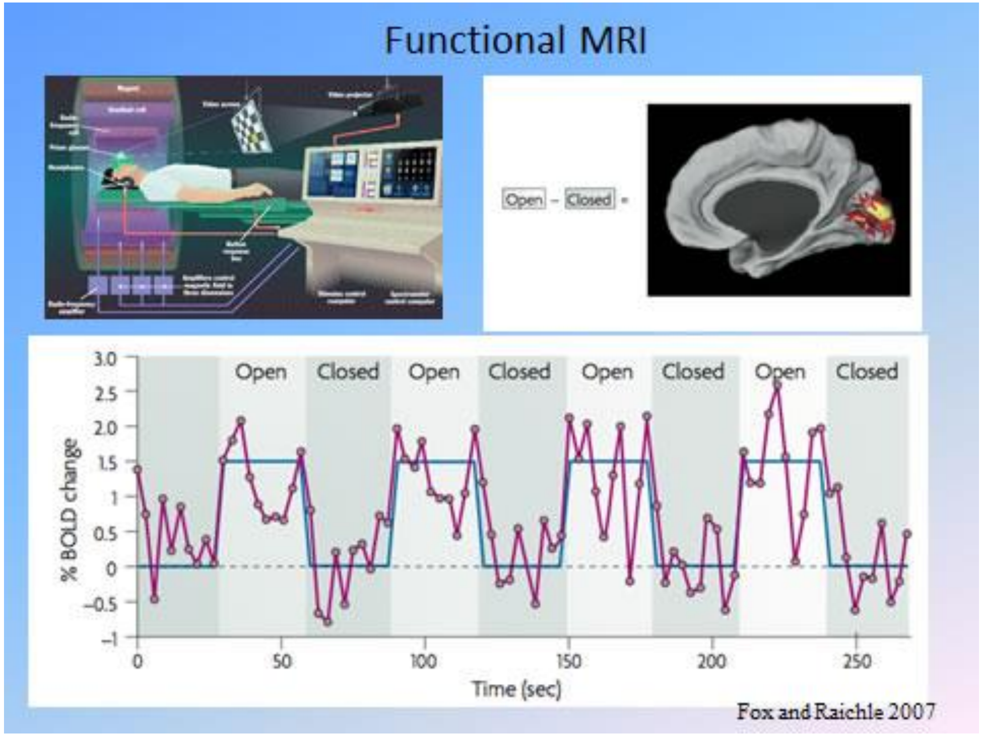
Types of MR images

- Structural MRI – measures different properties of brain tissue – typically not quantitative, good for differentiation of different tissue types
- Diffusion MRI - measures how water molecules diffuse through body tissues – can be used to map connectivity
- Functional MRI (BOLD) - measures changes in blood oxygeneation in different parts of the brain – proxy for brain activity as neurons use more oxygen when they are active

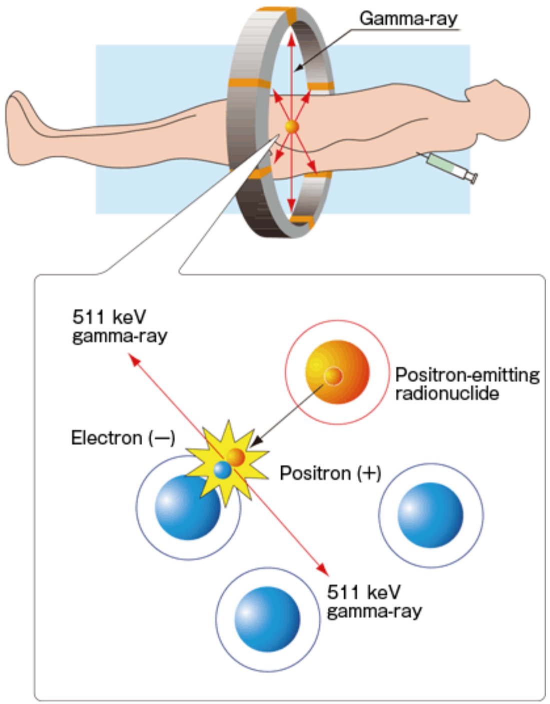




Functional MRI

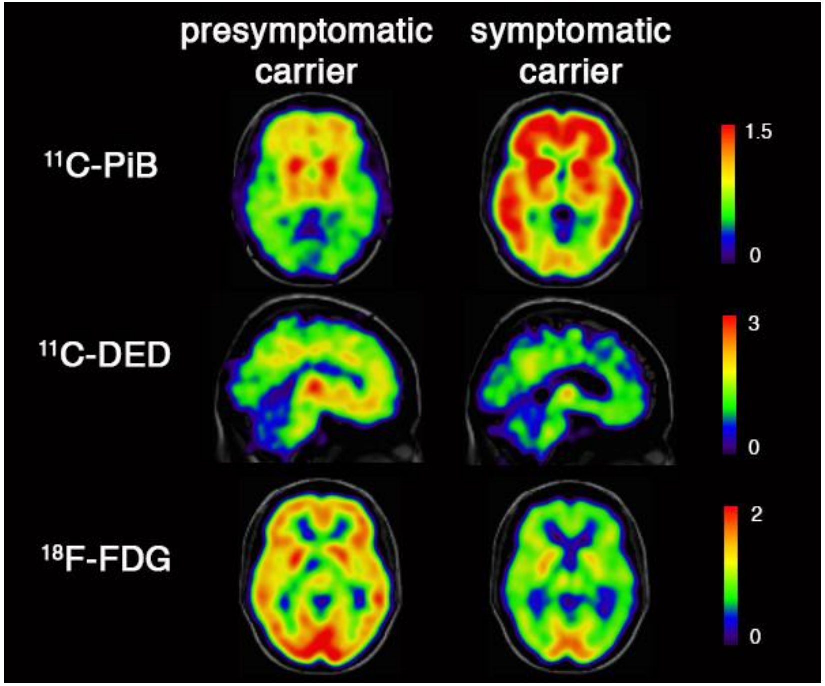


Positron emission tomography



PET can be used image various brain properties depending on the radiopharmaceutical (glucose consumption, dopamine receptor availability, Tau proteins, etc…)

Example: PET in very early Alzheimer's

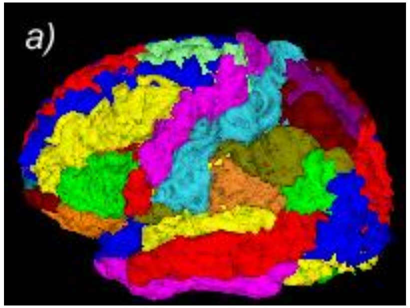


PET measures of amyloid, of glucose metabolism, but also of astrocytosis change over time in very early Alzheimer's disease

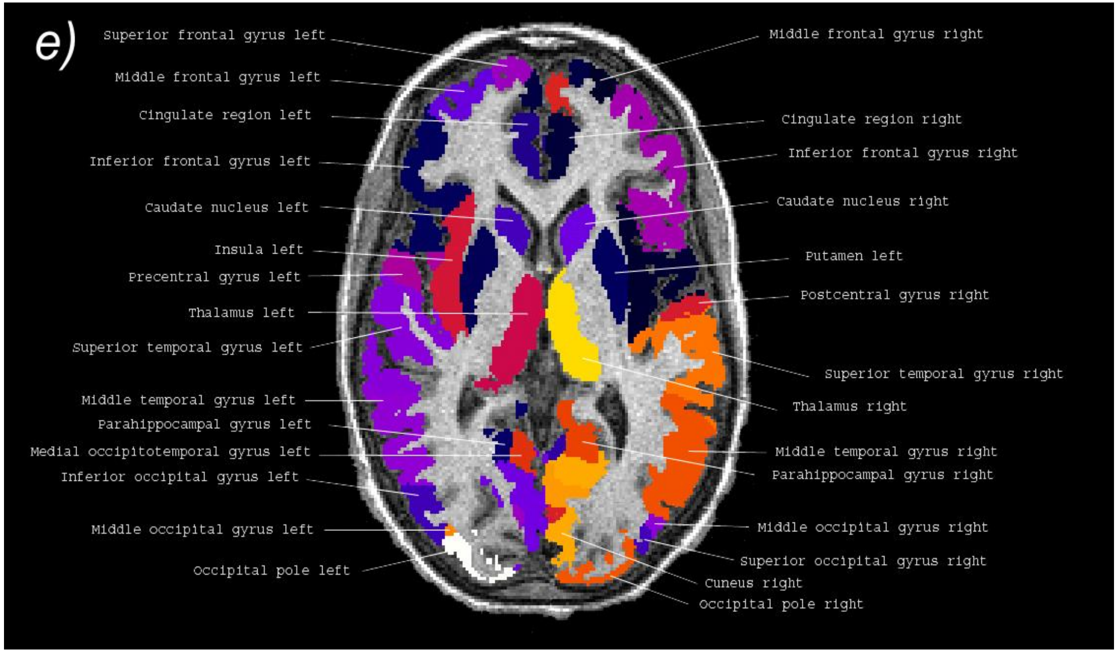
Rodriguez-Vieitez et al
Brain (2016)

Features from brain imaging – image analysis

ROI based analysis



- Brain can be divided into a number of specific brain regions
- Possible to study properties of these brain regions (volume, glucose consumption, etc..) and their relations to e.g. disease



http://www.thomaskoenig.ch/Lester/Files/Axial_Slice.jpg

Example: Hippocampal volume loss in Alzheimer's

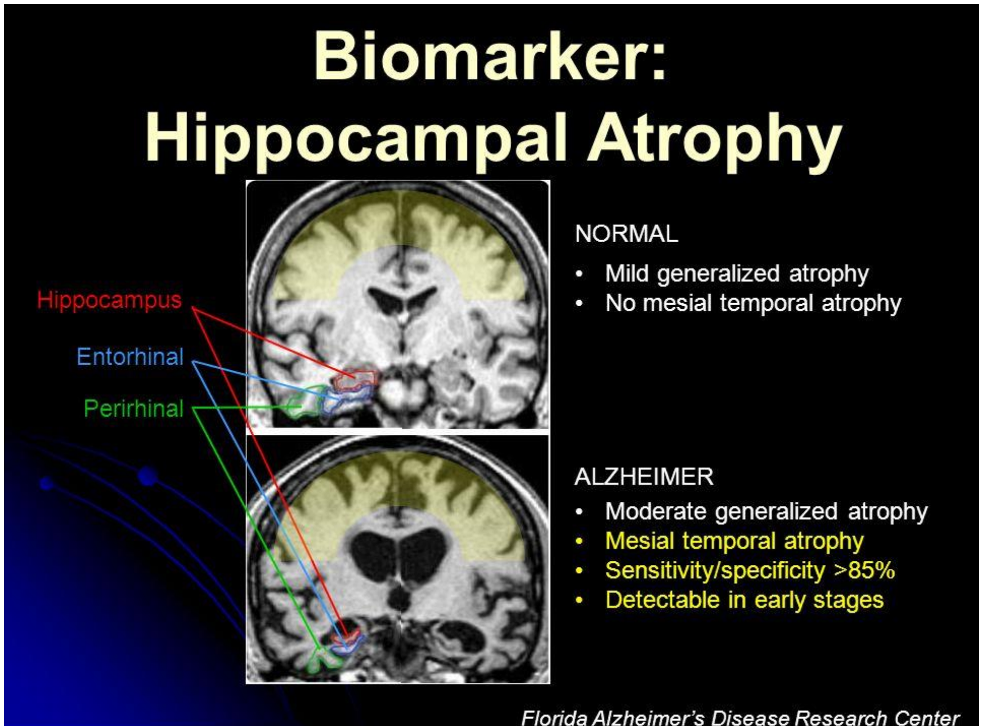
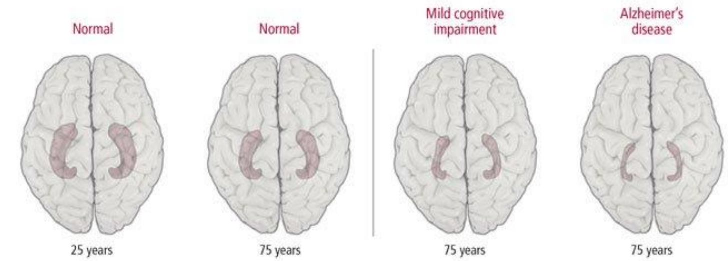
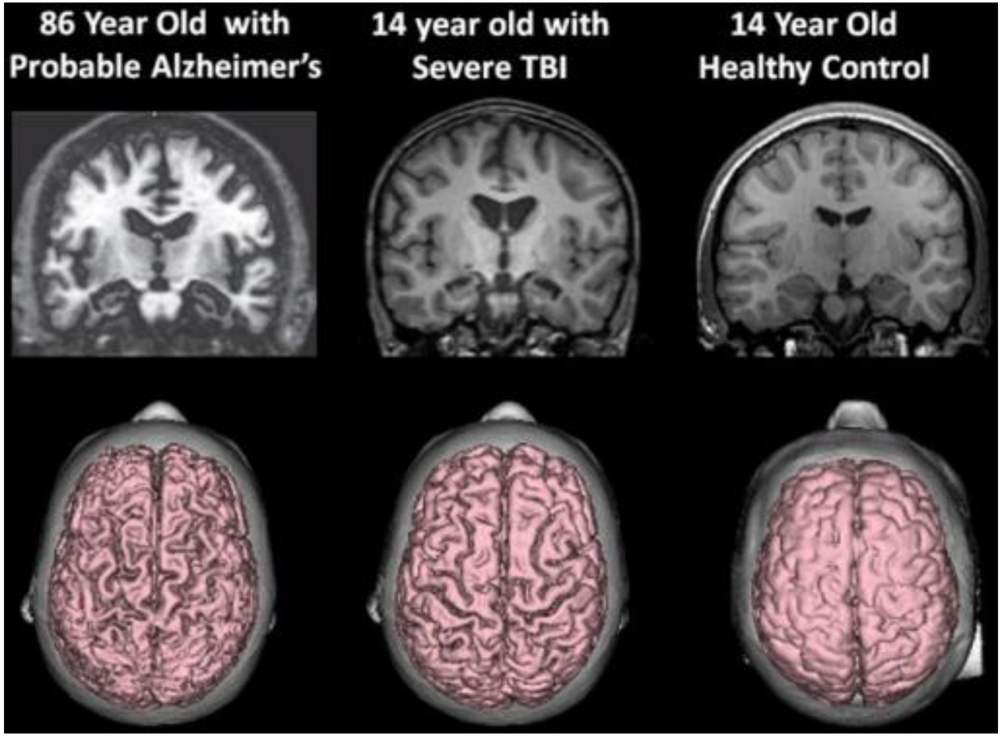


Figure 6 The shrinking hippocampus



A curved structure nestled deep within the brain, the hippocampus (from the Greek word for seahorse) plays a major role in forming, storing, and processing memories. The hippocampus becomes somewhat smaller as a part of normal aging, as shown by the comparison between the hippocampus in a healthy 25-year-old and a healthy 75-year-old. But the structure diminishes in size even more in a person with mild cognitive impairment and is markedly smaller than normal in a person with Alzheimer's disease.

However, hippocampal atrophy not specific to Alzheimer's



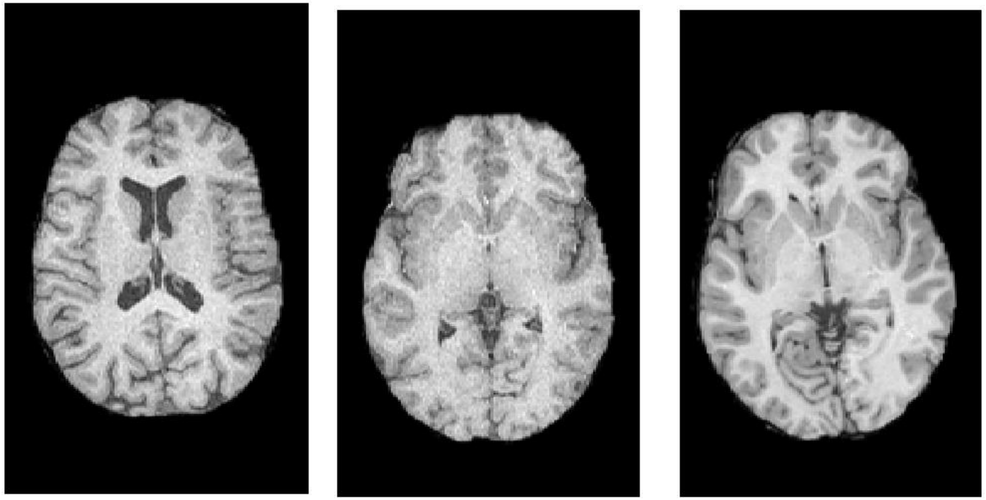
Bigler, Front Human Neuroscience 2013

Criticism towards the ROI based analysis

- Specific to fixed ROIs
- ROIs are usually large
- Not well adapted to cortical structures
- Manual segmentation burdensome and prone to intra and inter rater variability
- However, a lot of progress has recently been made towards automatic segmentation (multiatlas strategies, patch based data processing, deep learning)

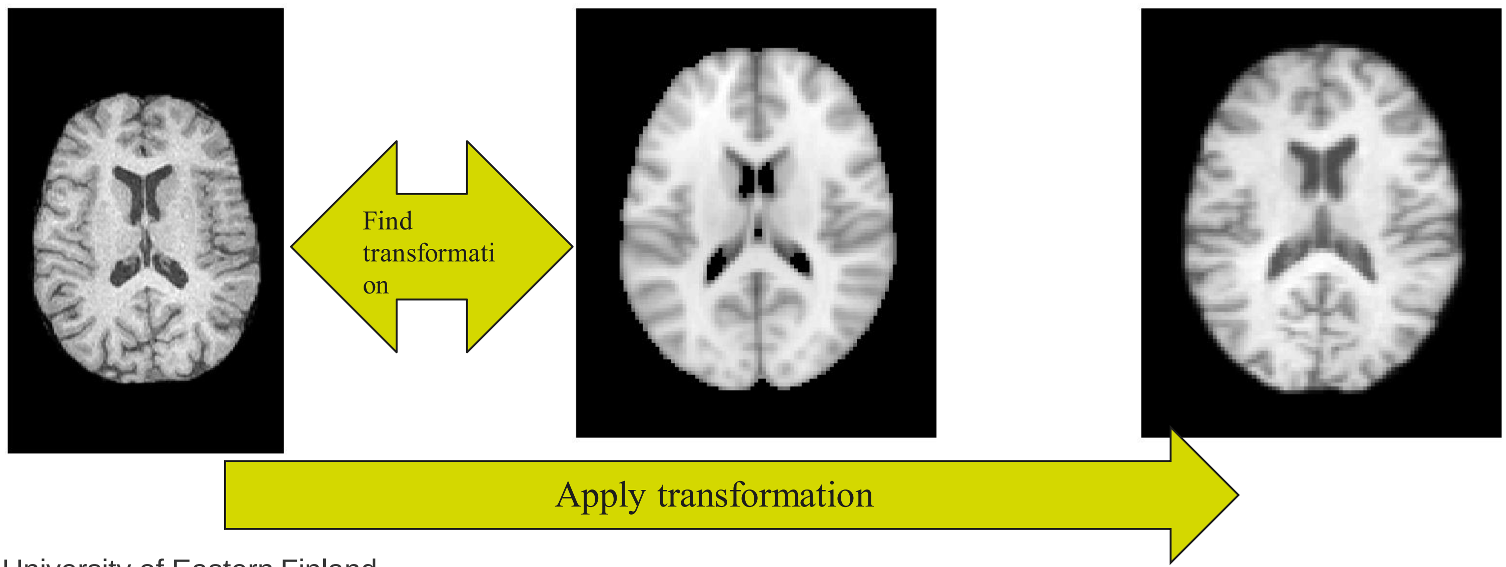
Stereotactic registration

- How to match the images of different subjects to each other?



Stereotactic registration/normalization

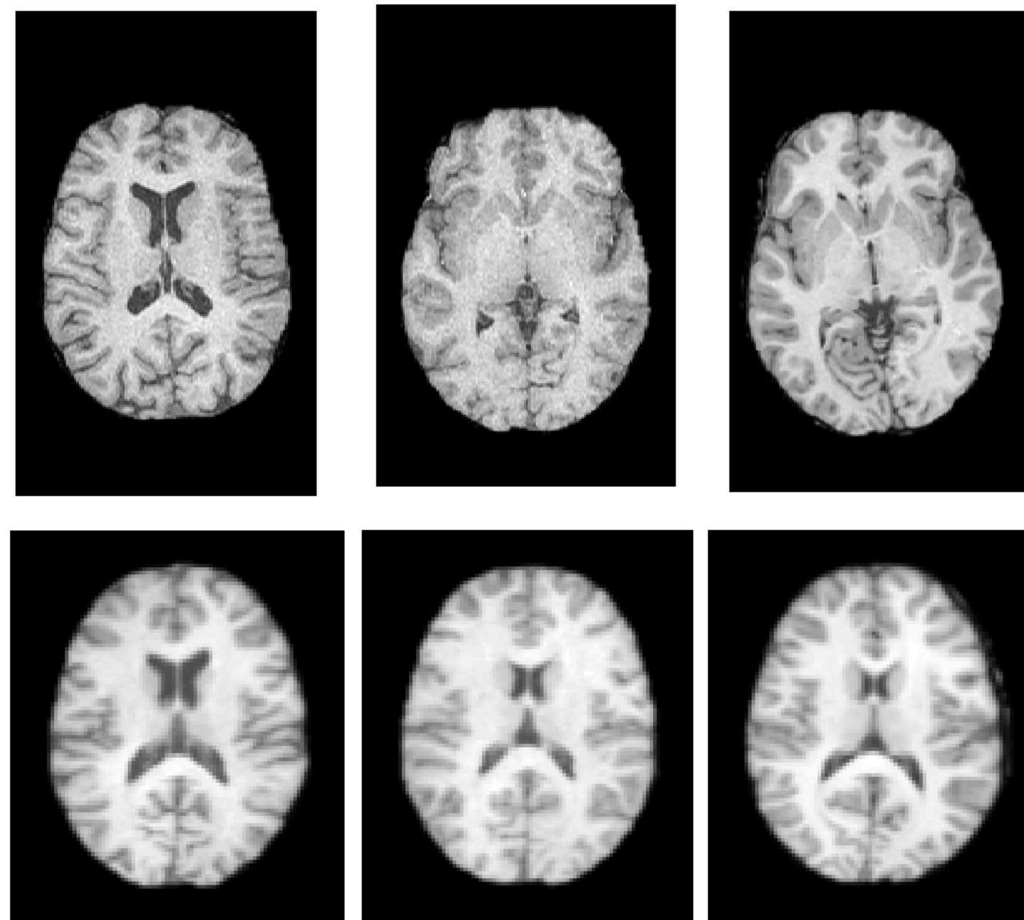
- Construct a template
- Find the spatial transformation that maximizes the similarity between template and the subject image
- Apply the transformation to the subject image



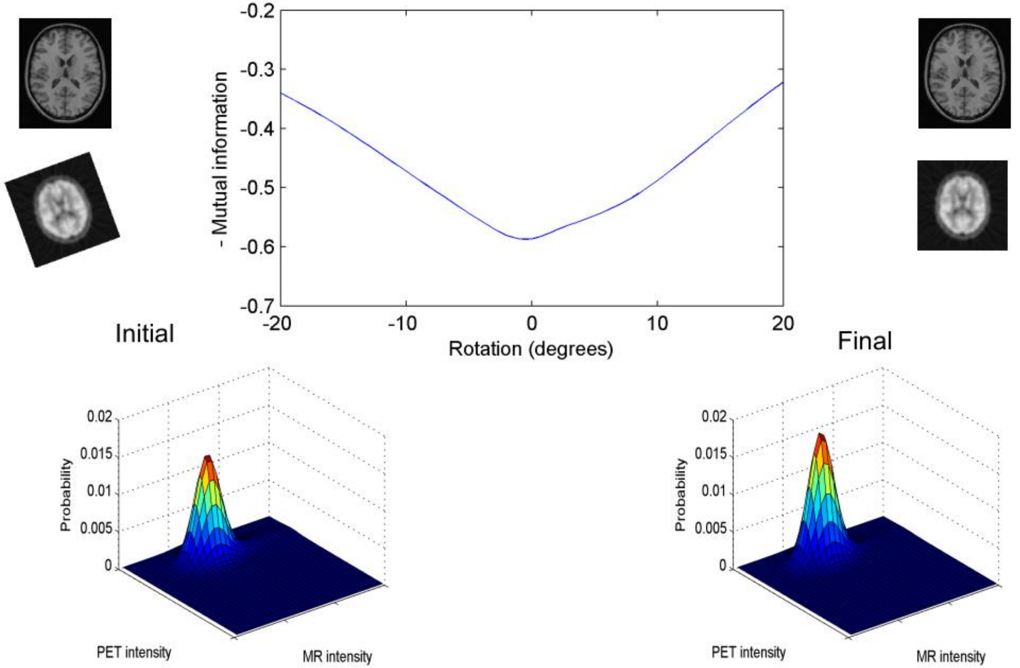
Stereotactic registration/normalization

Original

After stereotactic registration



Intermodality registration



Tohka, Brain imaging Encyclopedia, 2015

Voxel based analysis

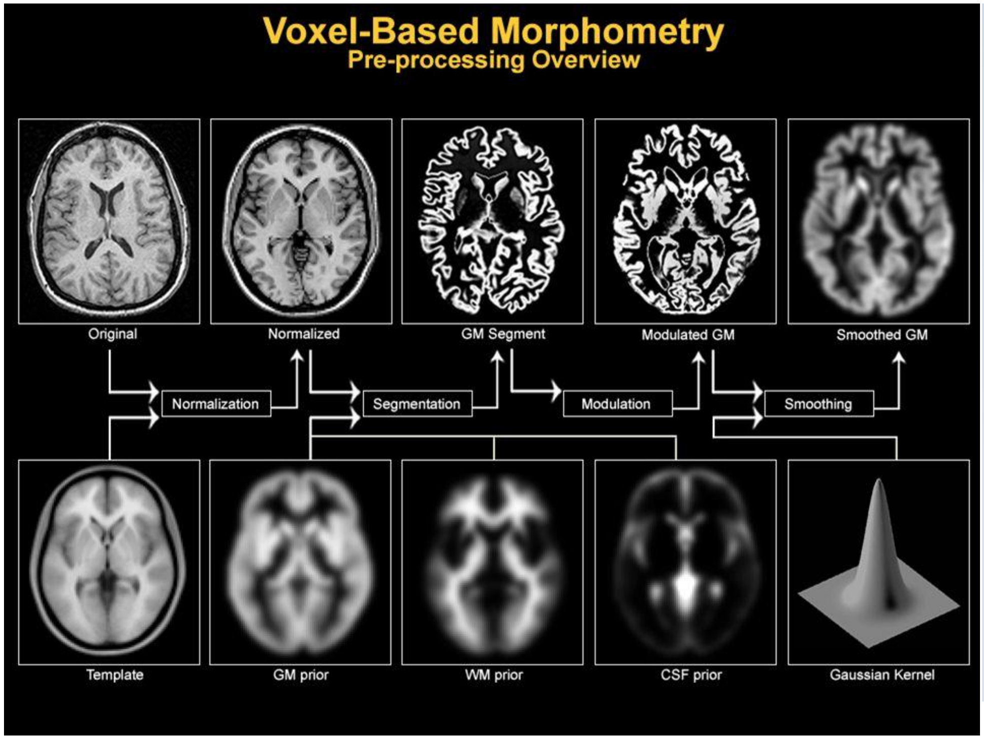
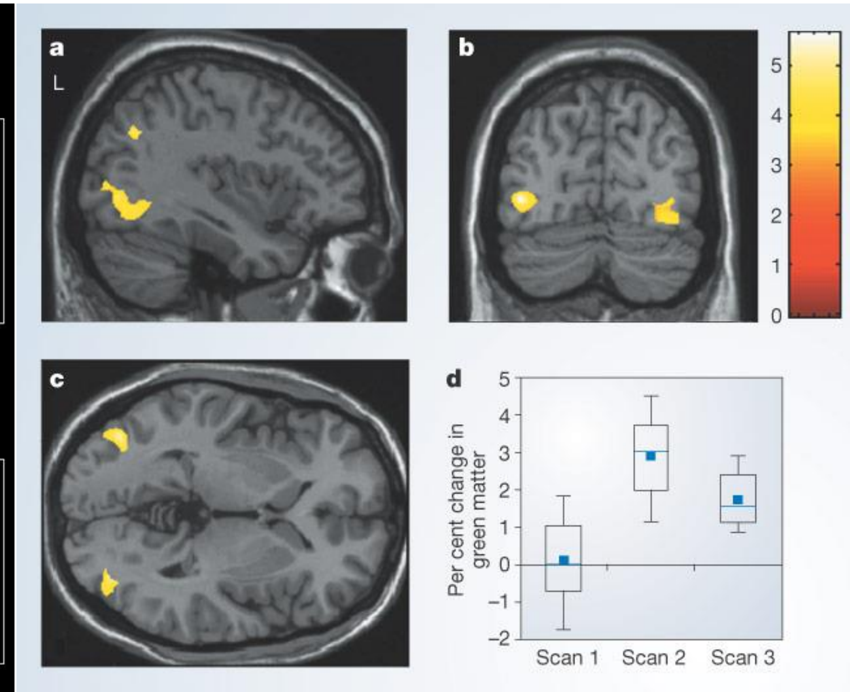


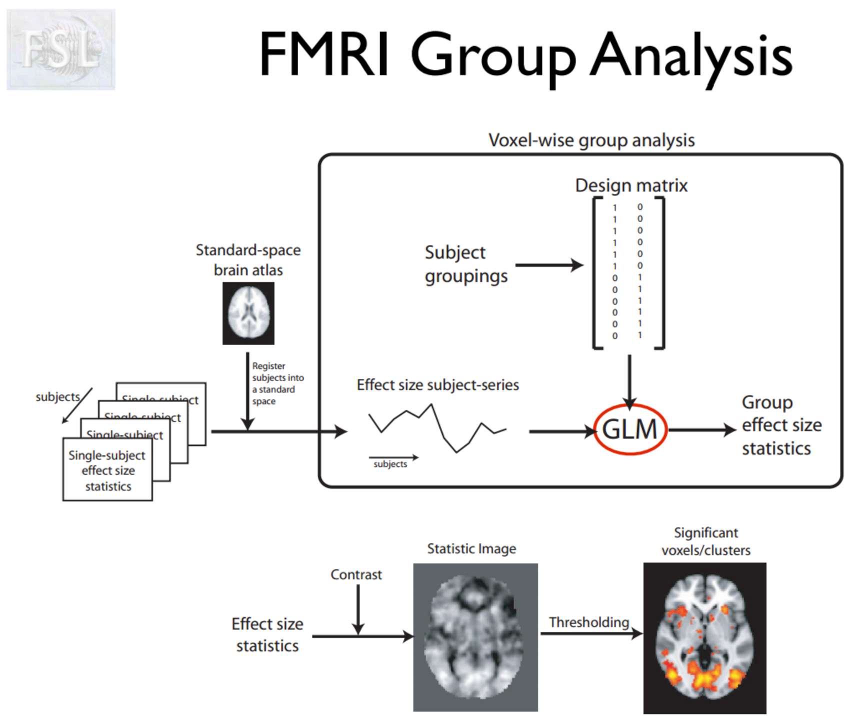
Figure by Suz Prejawa



- Jugglers vs. non-jugglers – brain structure changes due to training (Draginsky et al Nature 2004)

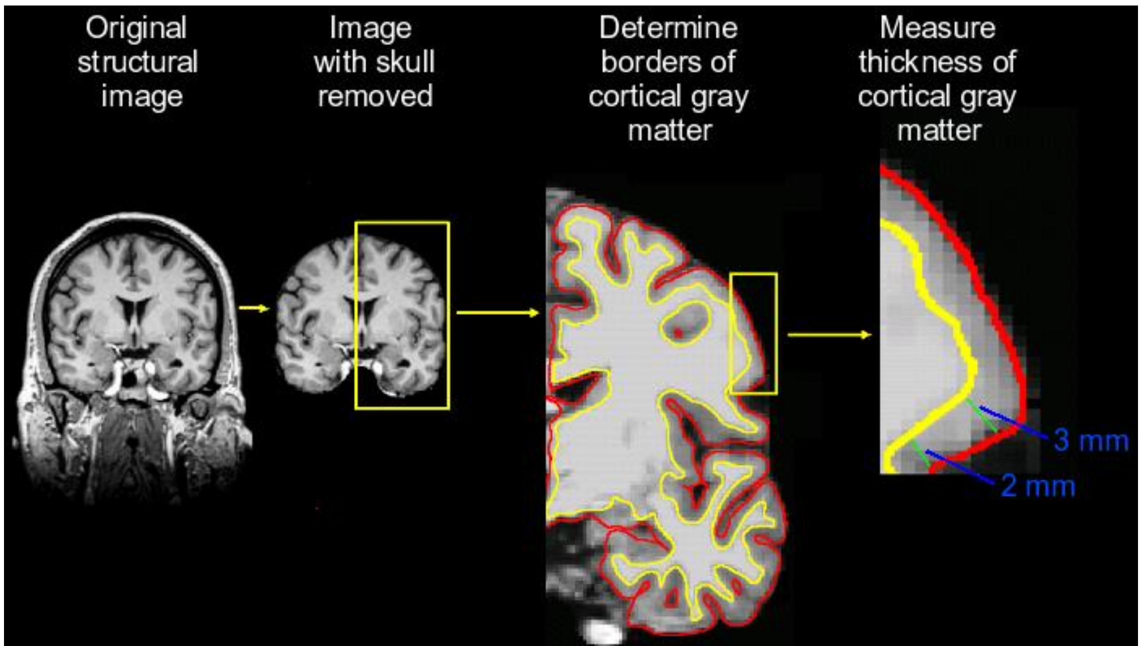
Voxel based fMRI analysis

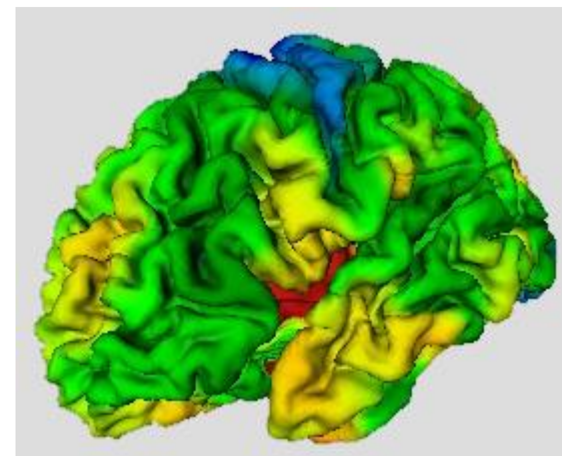
- http://fsl.fmrib.ox.ac.uk/fslcourse/lectures/feat2_part2.pdf



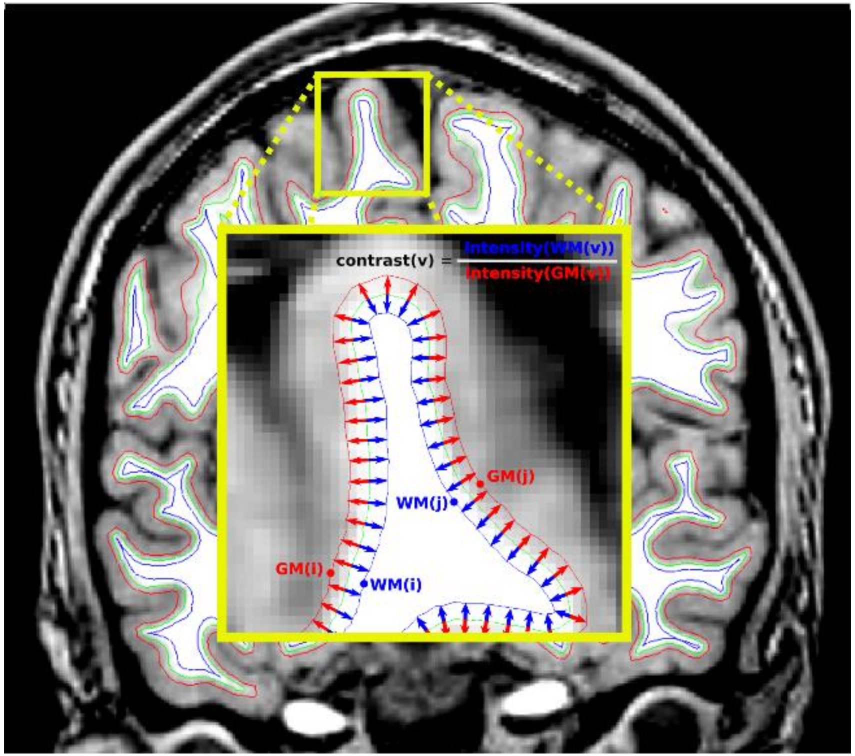
Surface based analysis

- Cortical thickness



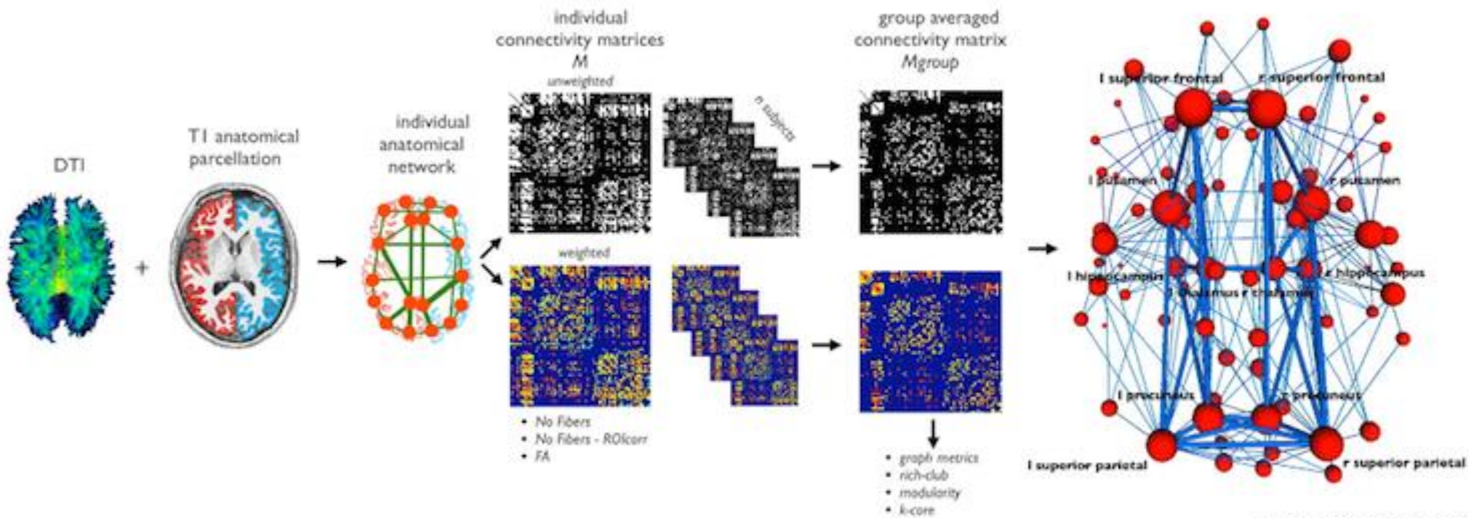


WM/GM contrast ratio



- Lewis, Evans, Tohka, biorxiv 2017

Networks



Heuvel & Sporns (2011)

Machine learning in brain imaging

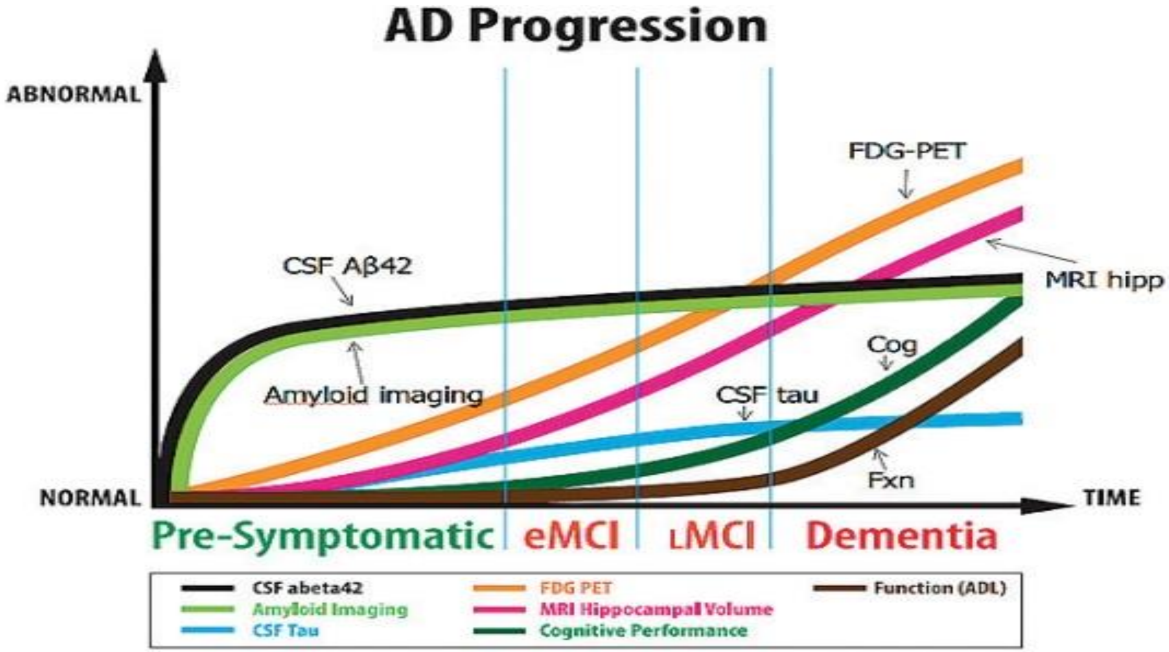
Variable selection, variable importance, error estimation

Machine learning in brain imaging

- Many uses in brain imaging
  - Diagnosis/prognosis/biomarkers (I will concentrate on this)
  - Segmentation
  - Brain computer interfaces
  - Functional imaging data analysis (multivoxel pattern analysis, searchlights)

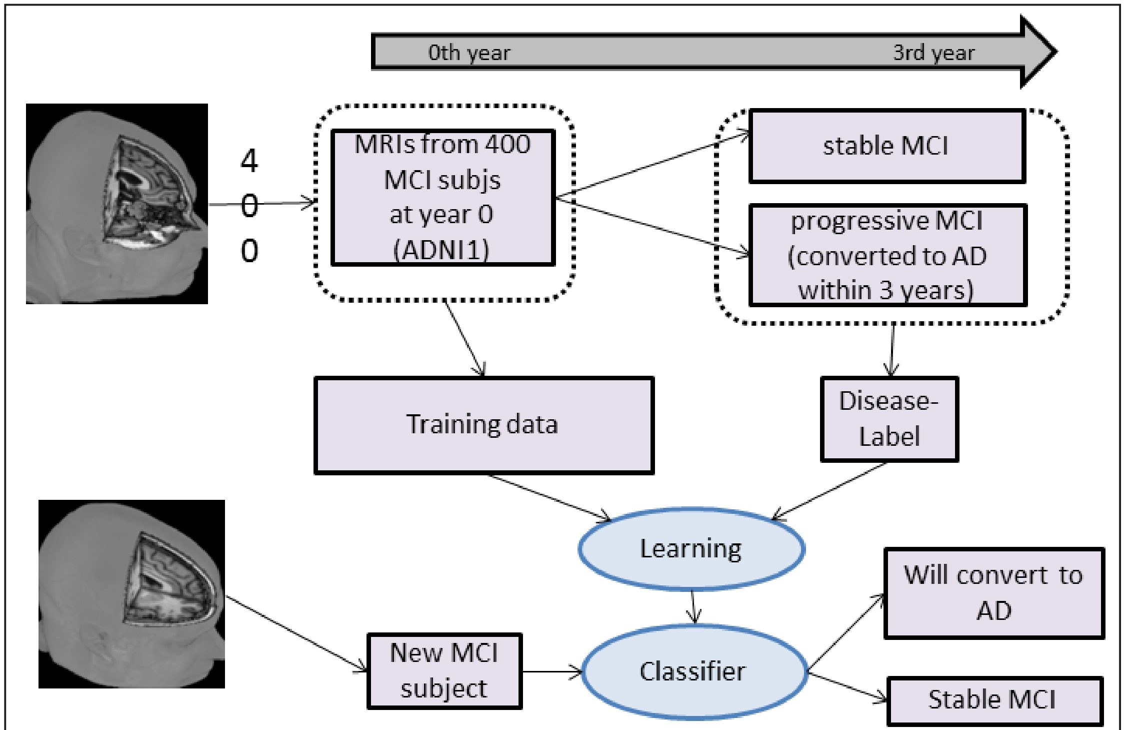
Imaging biomarkers and machine learning

Changes in the brain precede visible symptoms of diseases



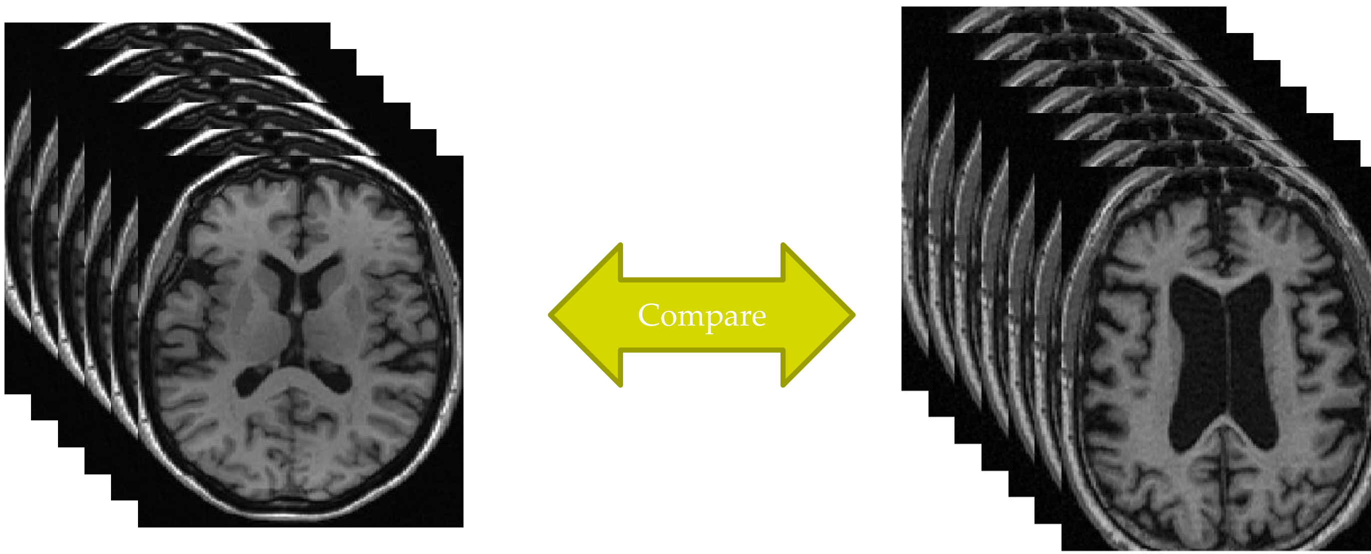
Early diagnosis possible using brain imaging

Example: Early diagnosis of Alzheimer's Disease



Moradi, Pepe, Gaser, Huttunen, Tohka,
Neuroimage, 2015

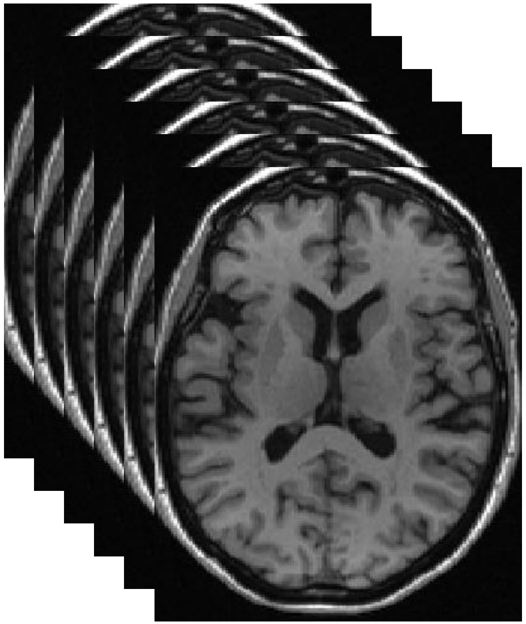
Traditional data analysis in brain imaging do not support making predictions about indivuals



Normal controls

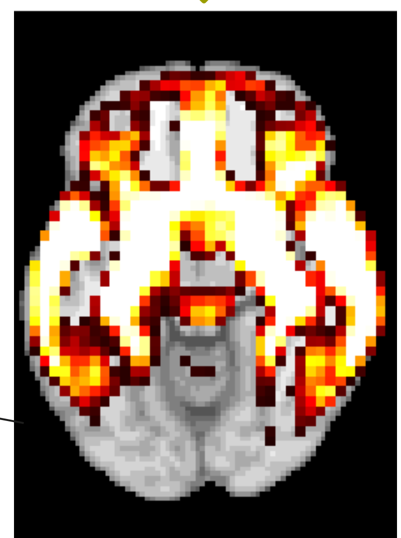
AD patients

Traditional data analysis - voxel based morphometry



Normal controls

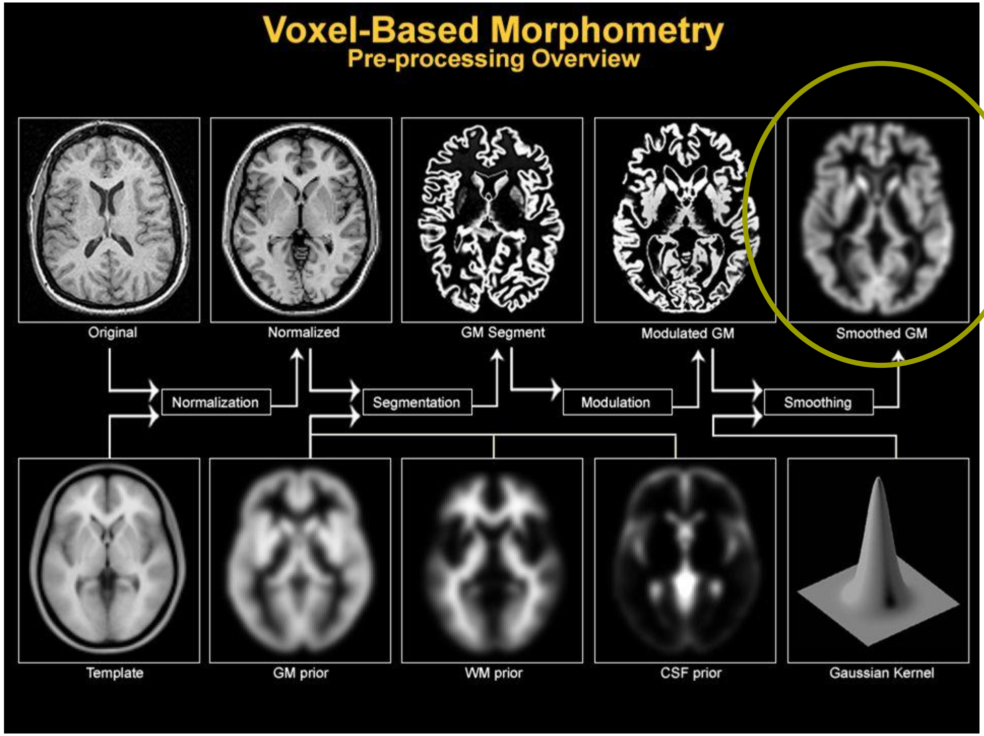
Voxel-wise maps of statistical differences of the brain structure between the two groups



Differences between two groups

AD patients

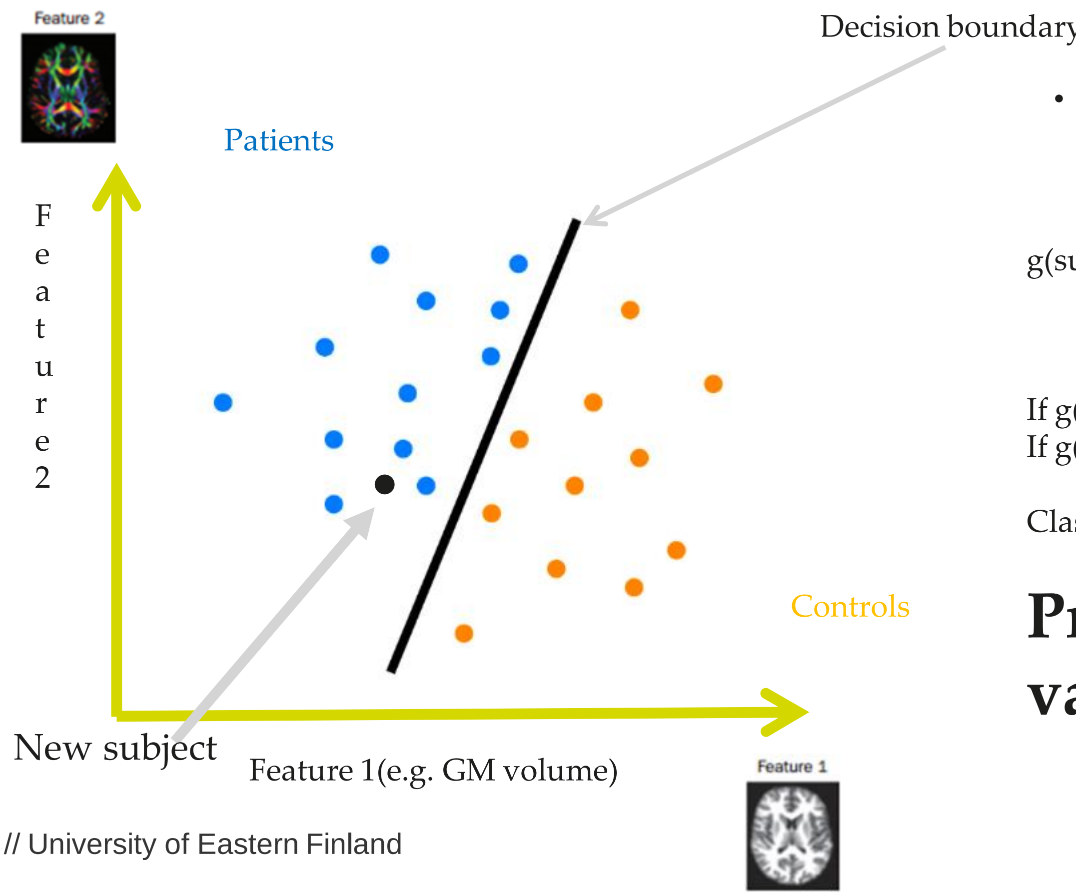
Preprocessing in traditional data analysis is useful for predictions



These are our features for each subject

Figure by Suz Prejawa

Machine learning is used to make predictions at the individual level



- Finding the decision boundary is in most cases formulated as a cost function optimization

$$g(\text{subject}) = b_1 * \text{Feature1}(\text{subject}) + b_2 * \text{Feature2}(\text{subject}) + \text{constant}$$

If g(subject) > 0 then subject = patient
If g(subject) < 0 then subject = control

Classifier training: find b1, b2, constant

Problem: Many more variables than subjects

MCI to AD conversion prediction algorithm: useful to combine imaging with other info

MRI-based, achieves cross-validated AUC of 0.76

1. Image preprocessing. Produces ~30K features representing voxel level gray matter density.
2. Age Confound Removal. Linear regression using **normal** subjects.
3. Feature selection using **normal** and **AD** subjects. Elastic-net (Zou, Hastie JRSS 2005) with re-iterated CV.
4. Classifier training. Low density separation (Chapelle AISTATS 2005)
5. Combining MRI with additional data via random forest (Breiman 2001)
   - LDS-MRI feature, age, cognitive test results (RAVLT, FAQ, MMSE, ADAS)

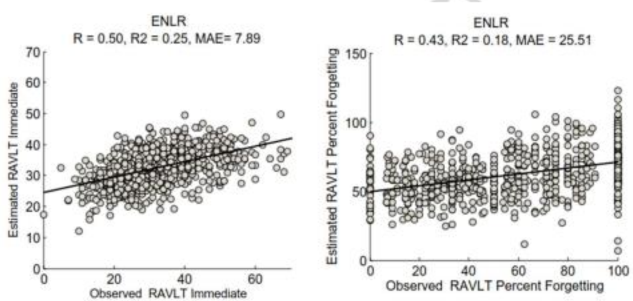
Aggregate biomarker, achieves cross-validated AUC of 0.90

Moradi, Pepe, Gaser, Huttunen, Tohka, Neuroimage, 2015

Machine learning is not limited to classification

Rey's Auditory Verbal Learning Test scores can be predicted from whole brain MRI in Alzheimer's disease

Elaheh Moradi[1a,h], Ilona Hallikainen[b], Tuomo Hänninen[c], Jussi Tohka[d,e,f],
Alzheimer's Disease Neuroimaging Initiative[g]



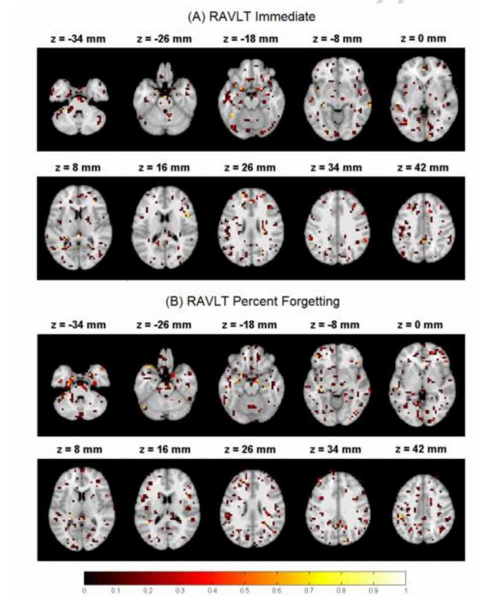
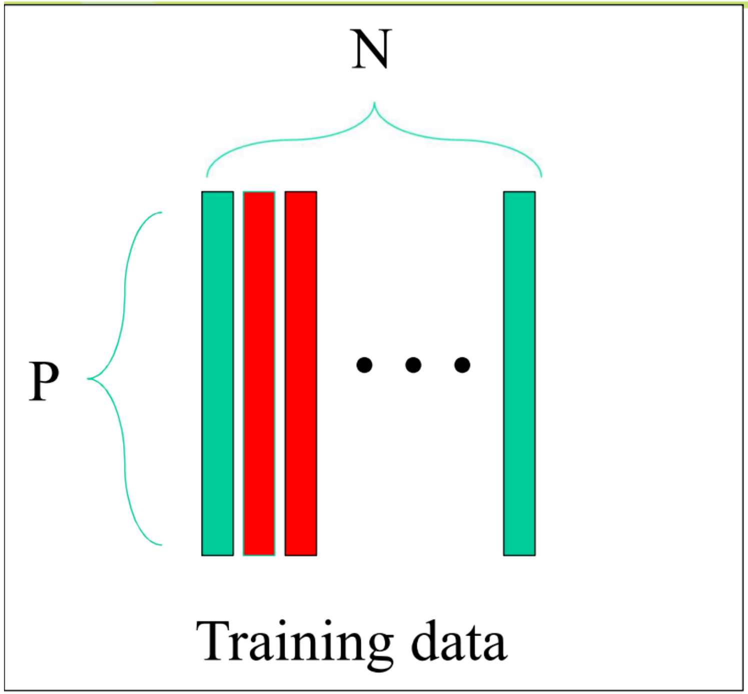
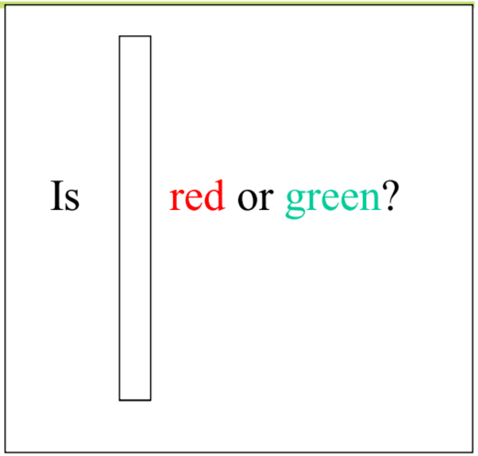


Figure 2: The selection probability of voxels in the estimation RAVLT Immediate (A) and RAVLT Percent Forgetting (B) across 100 different 10-fold CV iterations. The images are displayed according to the neurological convention.

Methodological aspects in small sample machine learning

Small sample considerations: variable/feature selection





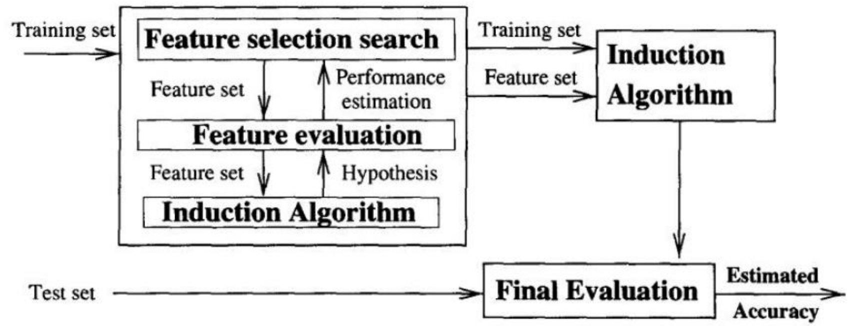
- What to do if

  P >> N?

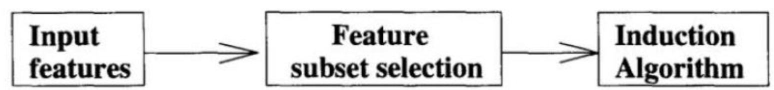
Even linear classifiers not stable

Variable/feature selection: filters and wrappers

Wrappers



Filters

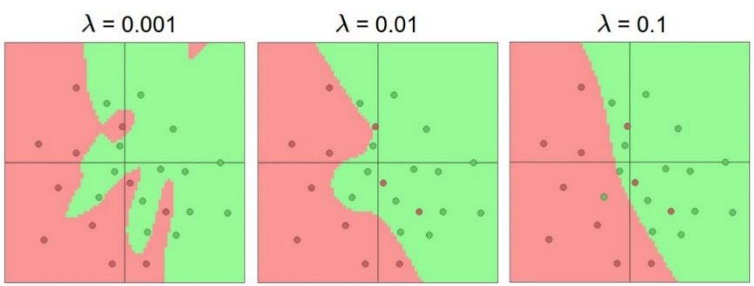


Kohavi and John Artif Intell 97

Overview of various methods:
https://github.com/jussitohka/brainimagingFS/blob/master/YTP_Huttunen_final.pdf

Regularization

- Many machine learning algorithms formulated as optimization problems
- These can be regularized

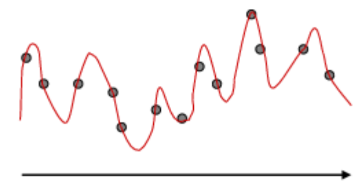


Regularization

– The minimization

$$\min_f |Y_i - f(X_i)|^2$$

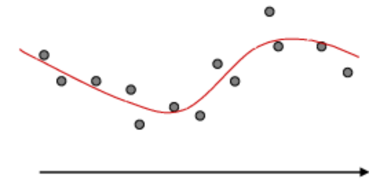
may be attained with zero errors.
But the function may not be unique.



– Regularization

$$\min_{f \in H} \sum_{i=1}^{n} |Y_i - f(X_i)|^2 + \lambda \|f\|_H^2$$

- Regularization with smoothness penalty is preferred for uniqueness and smoothness.
- Link with some RKHS norm and smoothness is discussed in Sec. IV.



II-26

Silde by Prof. Kenji Fukumizu

LASSO/elastic net for variable selection and classification

- Embedded variable selection: Minimize for joint classification/regression and feature selection:

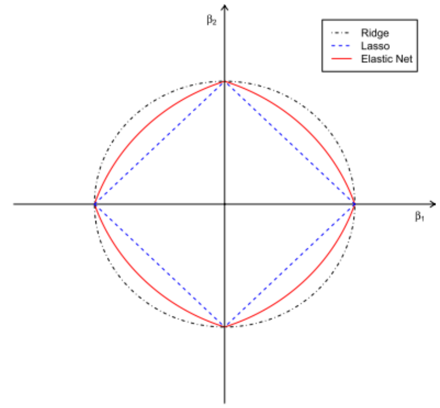
$$D(\mathbf{b} \mid \mathbf{X}, \mathbf{y}) + \lambda(1-\alpha) \| \mathbf{b} \|_1 + \lambda\alpha \| \mathbf{b} \|^2$$

Data term that relates independent (predictor) variables **X** to dependent variables **y** via unknown parameters **b**.
Here: **X** is imaging data, **y** is whatever we wish to predict, and **b** is what we want to estimate.

An example: least squares regression

Penalties for unknown parameters: the first one favors sparse solutions and the second one shrinks the solutions

2-dimensional illustration $\alpha = 0.5$



Friedmann et al JSS 2010, Zou and Hastie JRSS-B 2005
In brain imaging: Huttunen et al, MVAA 2013

Graphnet for variable selection and classification

- Embedded variable selection: Minimize for joint classification/regression and feature selection:

$$D(\mathbf{b}|\mathbf{X},\mathbf{y}) + \lambda(\alpha_1||\mathbf{b}||_1 + \alpha_2||\mathbf{b}||^2 + \alpha_3||\Delta\mathbf{b}||^2)$$

Data term that relates independent (predictor) variables **X** to dependent variables **y** via unknown parameters **b**.
Here: **X** is imaging data, **y** is whatever we wish to predict, and **b** is what we want to estimate.

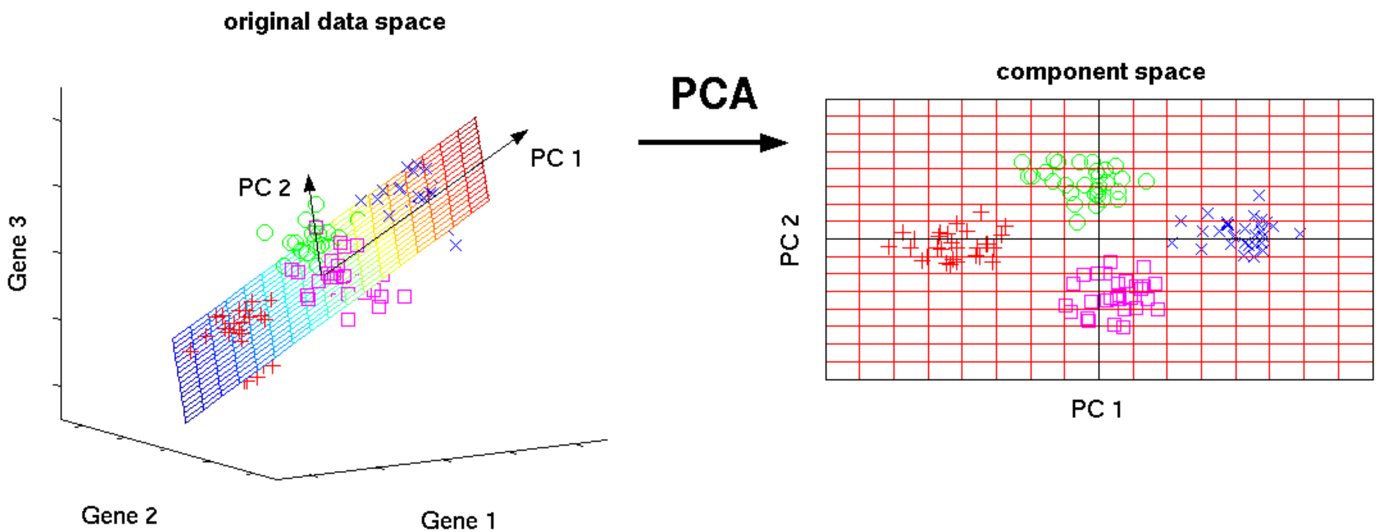
An example: least squares regression

Penalties for unknown parameters: the first one favors sparse solutions and the second one shrinks the solutions **and third one says that b-map should be smooth**

Grosenick et al. NeuroImage 2013
Tohka, Moradi, Huttunen, Neuroinformatics, 2016

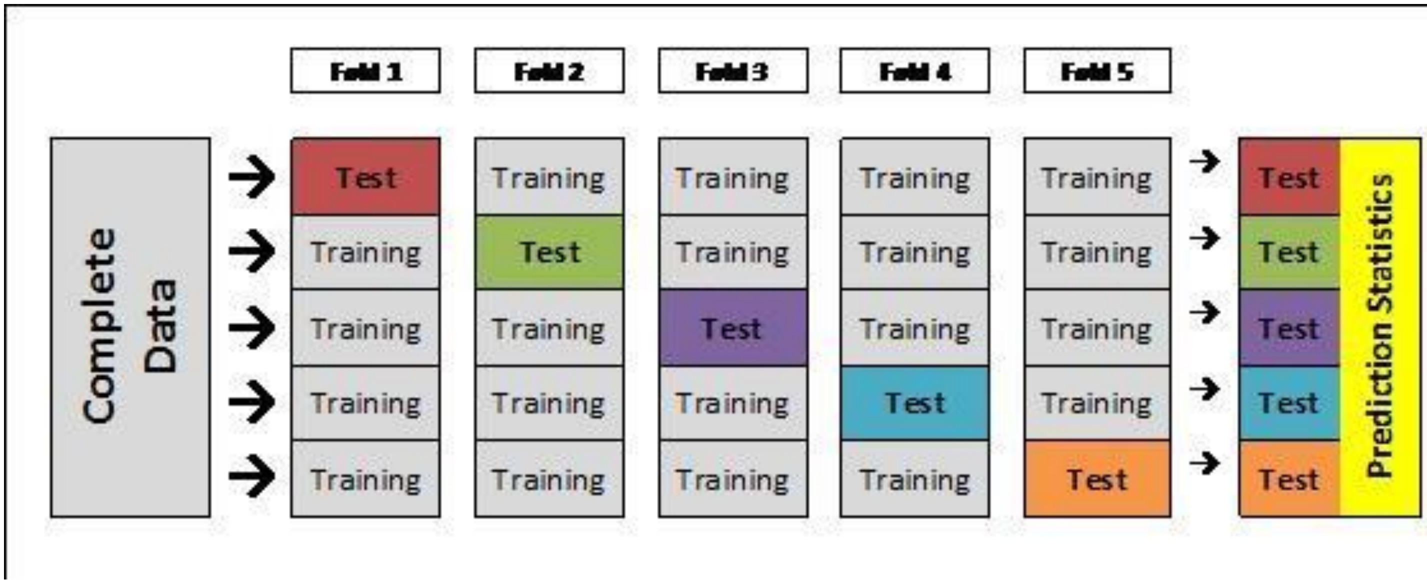
Component analyses for feature extraction

- Principal component analysis (PCA): For this, a data projection technique that helps to get rid of unnecessary dimensions
- Also, independent component analysis (ICA), etc…



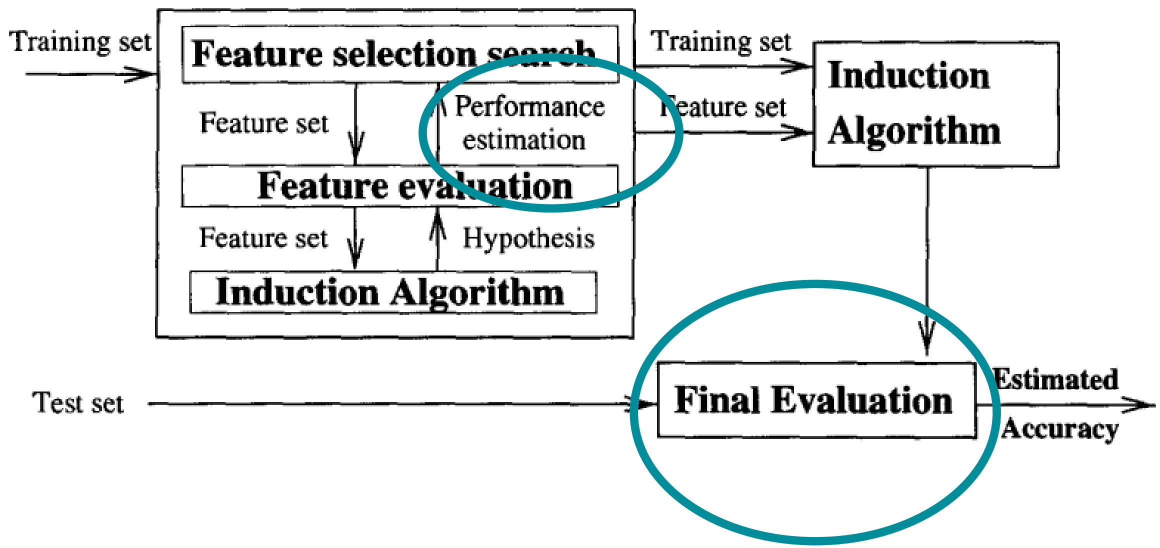
https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/

Evaluation of machine learning algorithms is important: cross-validation



http://blog.goldenhelix.com/bchristensen/cross-validation-for-genomic-prediction-in-svs/

Cross validation and variable selection

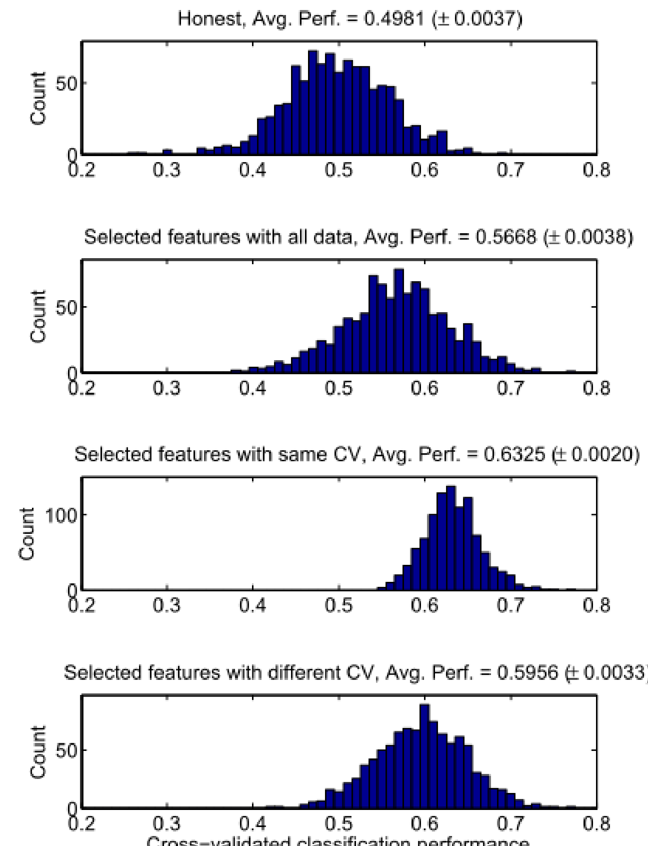


Modified from Kohavi and John 1997

- The test data should be kept in vault before the final evaluation

Cross-validation is easy to misuse

- Evaluation data has to be kept in the vault while selecting variables or other algorithm parameters
- Nested cross-validation
- Multiple cross-validation runs recommended



Huttunen, Manninen, Tohka, FCSE 2012

CV-based error estimates have large variance

Pattern Recognition Vol. 10, pp. 211-222.
Pergamon Press Ltd. 1978. Printed in Great Britain.

0031-3203/78/0601-0211 $02.00/0

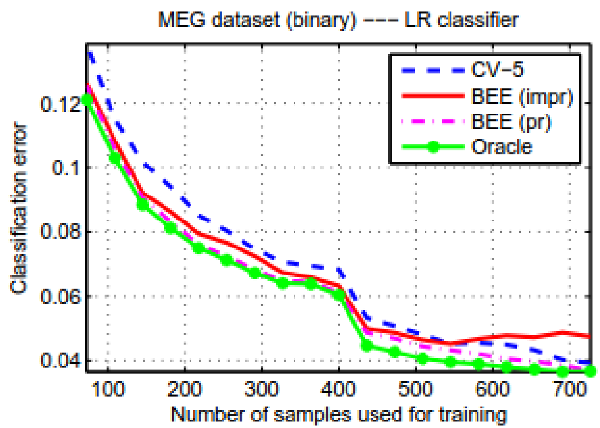
ADDITIVE ESTIMATORS FOR PROBABILITIES OF CORRECT CLASSIFICATION*†

NED GLICK
Department of Mathematics and Department of Health Care and Epidemiology,
University of British Columbia, Vancouver, B.C., Canada

(*Received* 6 *July* 1977; *in revised form* 17 *November* 1977; *received for publication* 3 *January* 1978)

What’s wrong with CV when used correctly?

- Nothing **if** you have loads of subjects and unlimited computational facilities
- **Otherwise:** CV based error estimates tend to have (too) high variance and nested CV takes ages to compute
- **Take home message**: Don’t trust machine learning results if N < 50
- **Parametric error estimates** can improve/speed up the model selection



Huttunen and Tohka, Pattern recognition, 2015
Huttunen, Manninen, Tohka, IEEE-MLSP-2013

Bayesian Error Estimate for Linear Classifiers

Assume that

1) The data is multivariate Gaussian
2) Model the prior of Gaussian parameters by inverse-Wishart distribution

Closed form expression for the MMS-estimate of the classification error:

These hyperparameters lead to a closed form solution [19]:

$$E[\varepsilon_c \mid \mathbf{X}, \mathbf{y}] = \frac{1}{2} + \frac{\operatorname{sign}(A_c(\lambda))}{2} I\left(\frac{A_c(\lambda)^2}{A_c(\lambda)^2 + \boldsymbol{\beta}(\lambda)^T \mathbf{S}_c \boldsymbol{\beta}(\lambda)}; \frac{1}{2}, \frac{N_c + 3}{2}\right),$$

where

$$A_c(\lambda) = -y g_\lambda(\mathbf{m}_c(\lambda)) \sqrt{(0.5 + N_c)/(1.5 + N_c)}$$

and

$$\mathbf{m}_c(\lambda) = \frac{\hat{\boldsymbol{\mu}}_c(\lambda) N_c}{N_c + 0.5} \quad \text{and} \quad \mathbf{S}_c(\lambda) = (N_c - 1)\hat{\boldsymbol{\Sigma}}_c(\lambda) + \mathbf{I}_{N_c} + \frac{0.5 N_c}{N_c + 0.5} \hat{\boldsymbol{\mu}}_c \hat{\boldsymbol{\mu}}_c^T.$$

Code available (Matlab, Python):
https://sites.google.com/site/bayesianerrorestimate/

Dalton and Dougherty IEEE-TSP 2011
Huttunen and Tohka, PR 2015

Is machine learning stable?

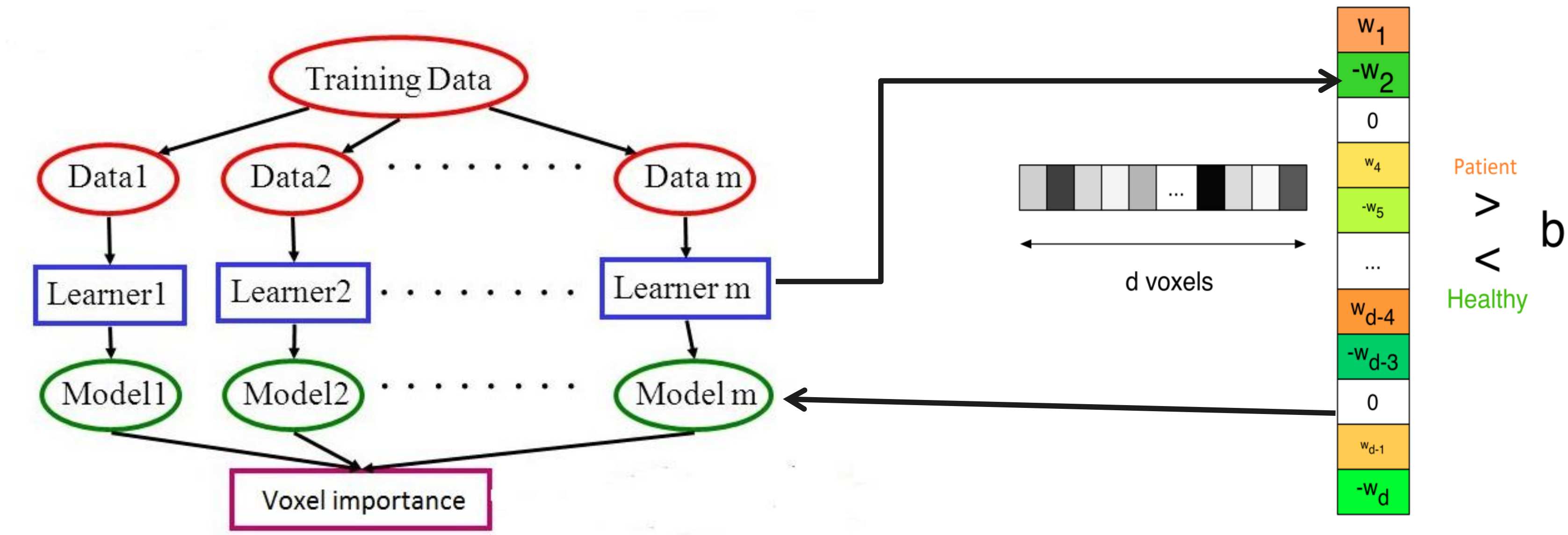
- What happens if we change the training subjects?
  - Do we get the same accuracy?
  - Do we get similar models?

Is machine learning stable?

- What happens if we change the training subjects?
  - Do we get the same accuracy? Approximately
  - Do we get similar models? No, but BEE helps

Tohka, Moradi, Huttunen, Neuroinformatics 2016

Feature selection instability: sign consistency bagging could be the answer



Verdejo-Gomez, Parrado, Tohka, 2016, 2017

Conclusions: Why machine learning?

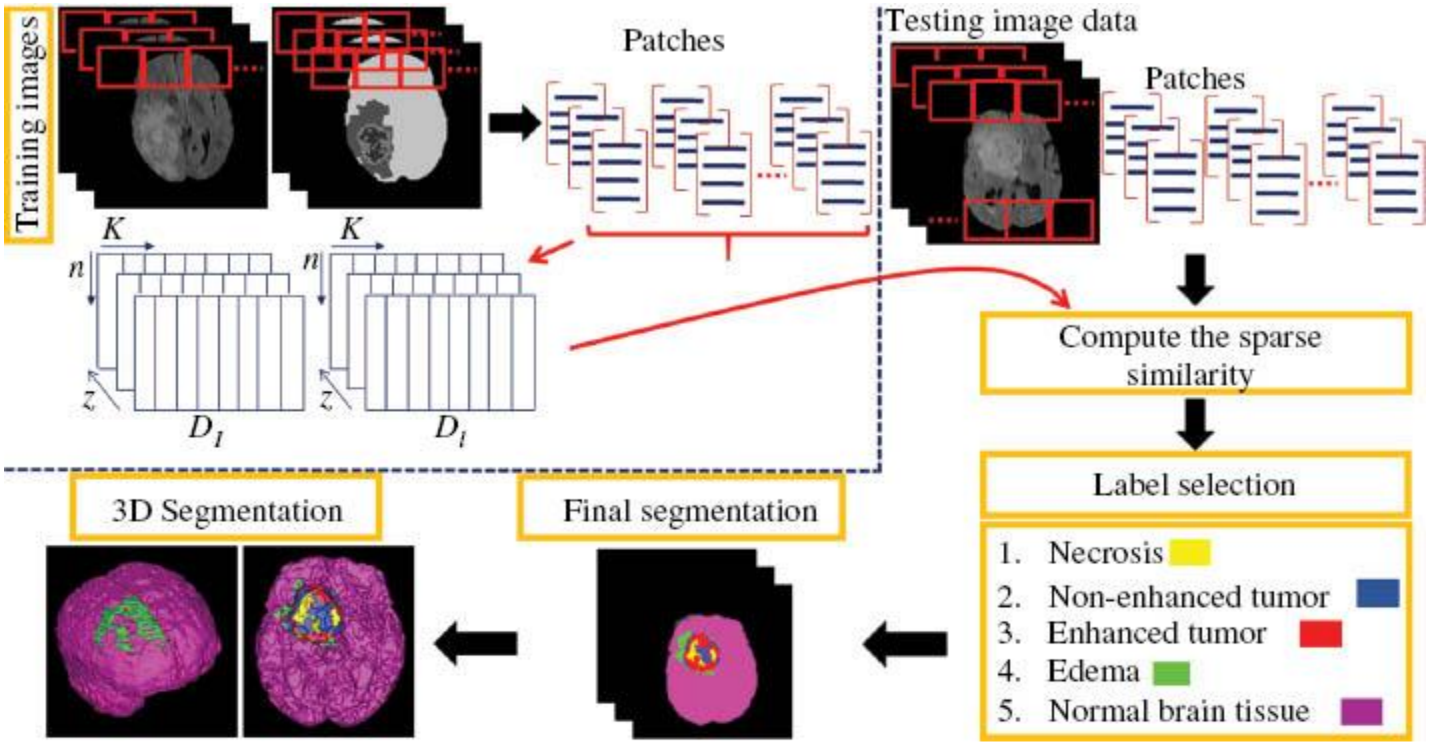
- Close to actual applications (learn from group, apply to individual)
- Non-trivially integrate other kinds of data to neuroimaging data
- Data-driven, no alpha thresholds
- Multivariate

Conclusions: Take home messages

- Advanced machine learning methods adapted to imaging data exist, but little validation has been done
- Combining imaging data to behavioural data likely to be necessary for translational potential

Machine learning for brain image segmentation

Patch-based segmentation



- Salman Al-Shaikhli, et al *Biomedical Engineering / Biomedizinische Technik*, 61(4), pp. 413-429, 2015.

Patch-based segmentation

- Main interest: the segmentation of anatomical MRI
- Patch based segmentation not very popular until recently because
  - Typical MRI not quantitative, exact tissue intensities vary between subjects and (especially) between machines
  - In MRI high level spatial context important
- Label probagation and multi atlas label probagation widely used
- Good results from patch-based segmentation with expert priors (Coupe et al 2011, Rousset et al 2011); Mix of patch based segmentation and multiatlas label probagation; Preprocessing with clustering based segmentation
  - Not very robust to gross pathologies

Patch-based deep learning with spatial context

- DeepNAT: Wachinger et al 2016
- 23 x 23 x 23 patches + spatial coordinates + conditional random field based pruning
- Very good results
- Not clear whether mainly due to DNN or clever spatial coordinate coding scheme or both

Machine learning for segmentation vs. imaging biomarkers

- Very different problems in terms of machine learning
- Biomarkers: Small number of samples and small to large dimensionality
- Segmentation: large number of samples (due to patch based processing) and large dimensionality