From machine learning to deep learning,

how do we ensure objective and reproducible evaluations?

ARAMIS LAB
BRAIN DATA SCIENCE

www.aramislab.fr

www.clinica.run

ninon.burgos@icm-institute.org

OHBM – ML4NI

2020

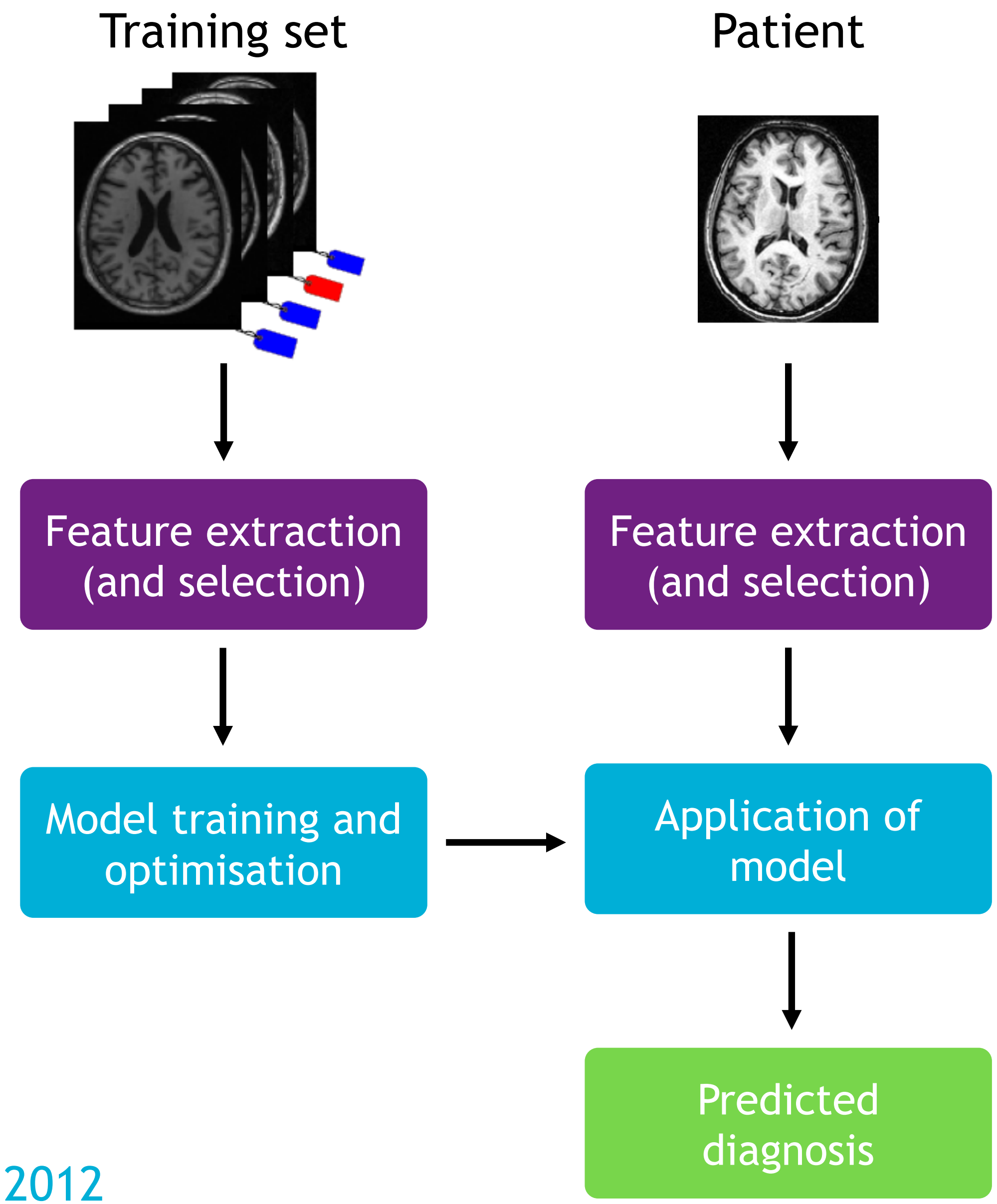
Ninon Burgos, CNRS Researcher

Aramis Lab, Paris Brain Institute, France

Clinical pattern recognition

Basic elements of a machine learning classification pipeline

- Training data set
- Feature extraction from raw data and dimensionality reduction
- Model training and optimization
- Application to test data

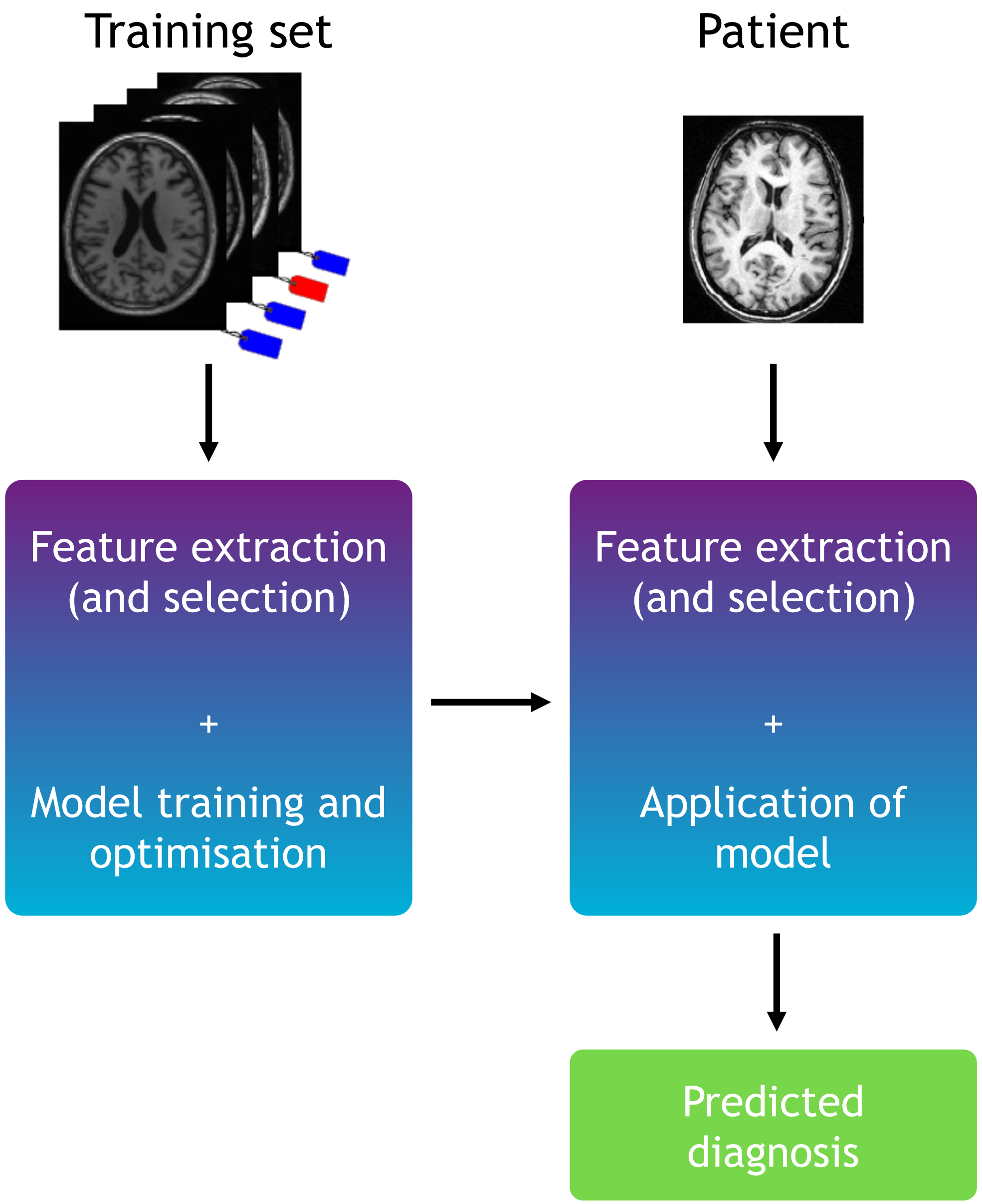


Klöppel et al., NeuroImage, 2012

Clinical pattern recognition

Basic elements of a deep learning classification pipeline

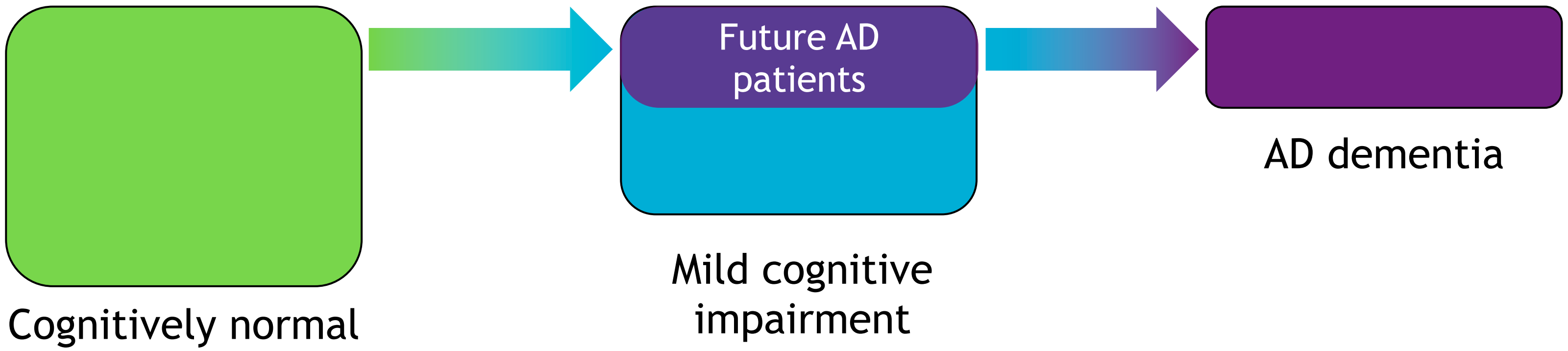
- Training data set
- Feature extraction from raw data and dimensionality reduction
- Model training and optimization
- Application to test data



Use case: Alzheimer’s disease (AD)

What is Alzheimer’s disease?

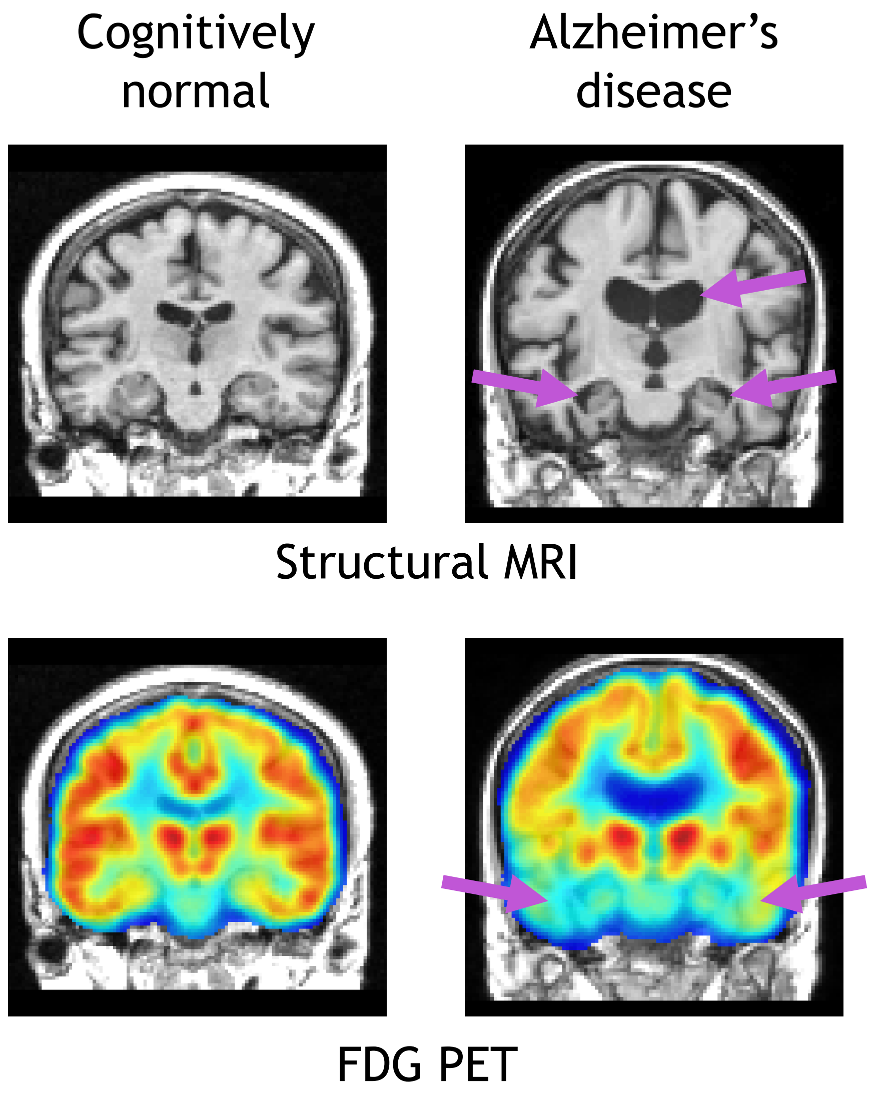
- Most common cause of dementia
- Disorder caused by abnormal brain changes
- Trigger decline in cognitive abilities, severe enough to impair daily life
- Affect behaviour, feelings and relationships
- Progressive disease



Use case: Alzheimer’s disease (AD)

AD-related biomarkers

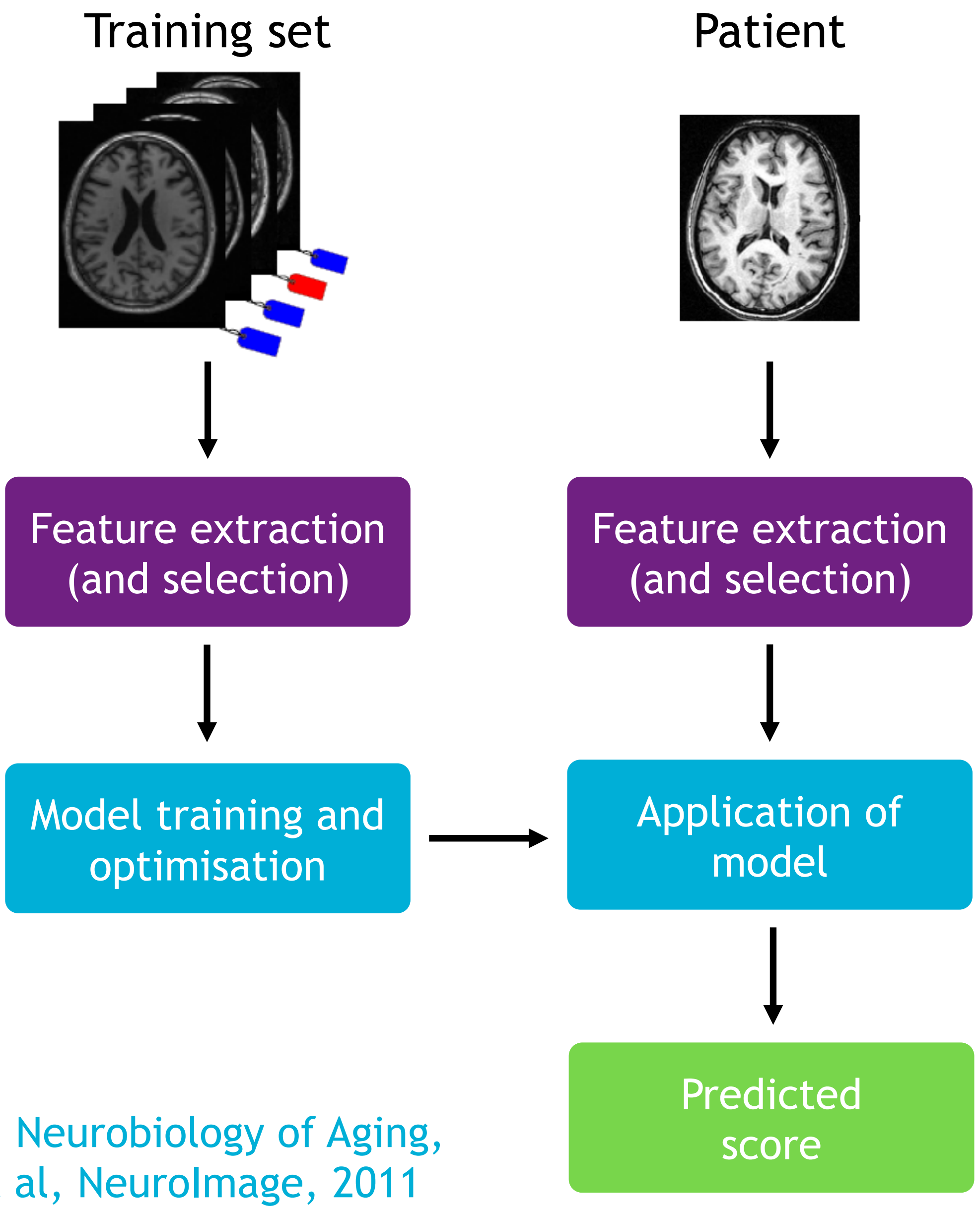
- **Clinical/cognitive tests**
  - Neuropsychological testing of cognitive functions (memory, language, etc.)
- **Structural MRI**
  - Atrophy
- **FDG PET**
  - Hypometabolism
- **CSF Aß42, CSF tau, amyloid PET, tau PET, diffusion MRI, etc.**



Clinical pattern recognition

Use case: Alzheimer's disease (AD)

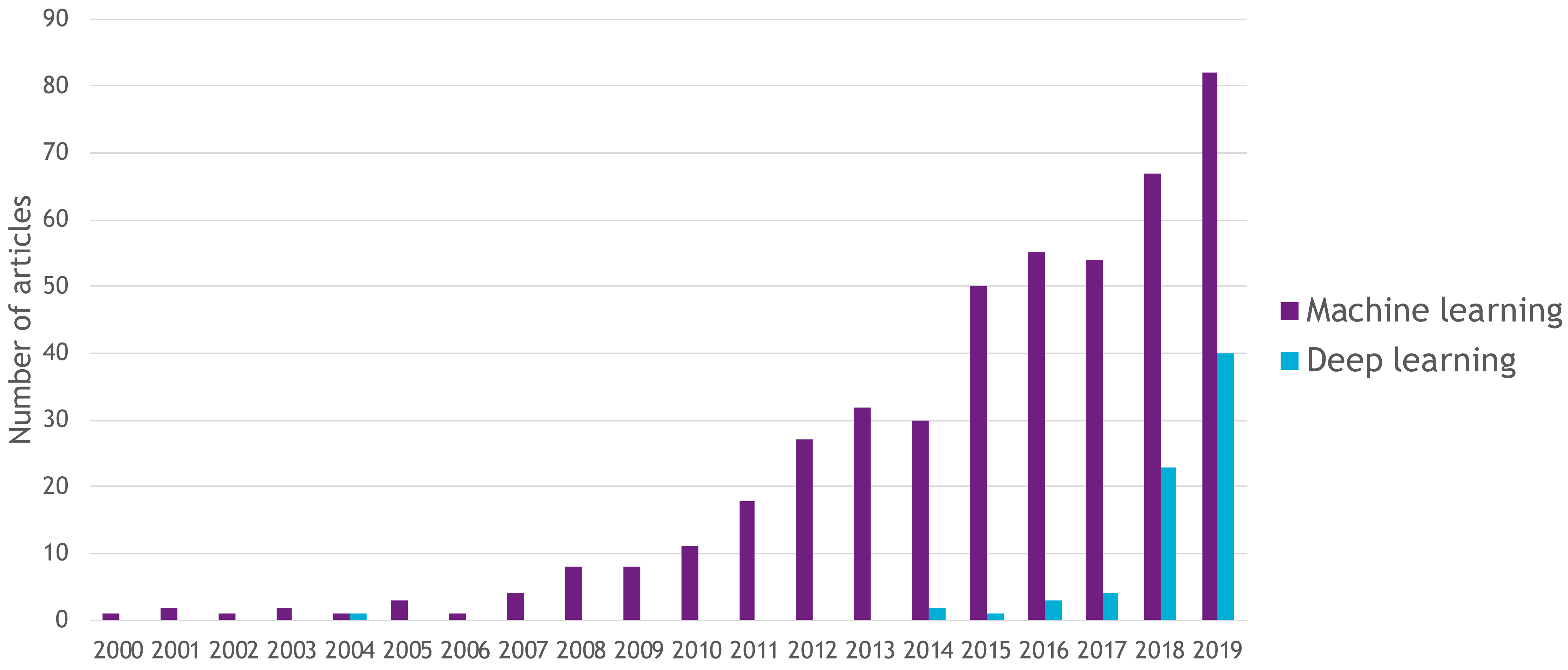
- **Classification**
  - Controls vs AD patients
  - Stable vs progressive mild cognitive impairment (MCI)
- **Regression**
  - Time of onset
  - Future clinical score



Klöppel et al., NeuroImage, 2008; Davatzikos et al, Neurobiology of Aging, 2008; Zhang et al, NeuroImage, 2011; Cuingnet et al, NeuroImage, 2011

ML/DL for Alzheimer's diagnosis & prognosis

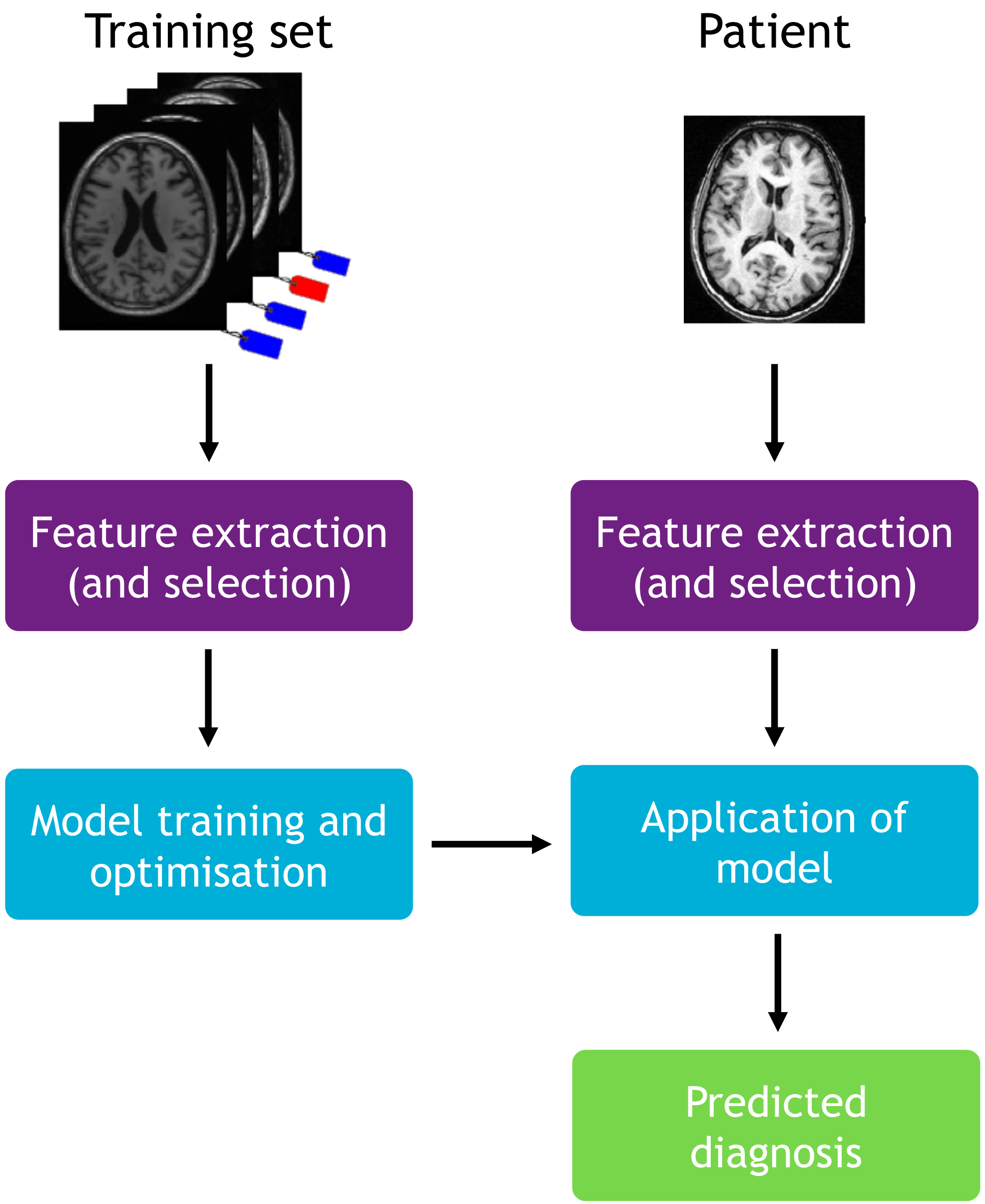
A very active field of research



ML/DL for Alzheimer's diagnosis & prognosis

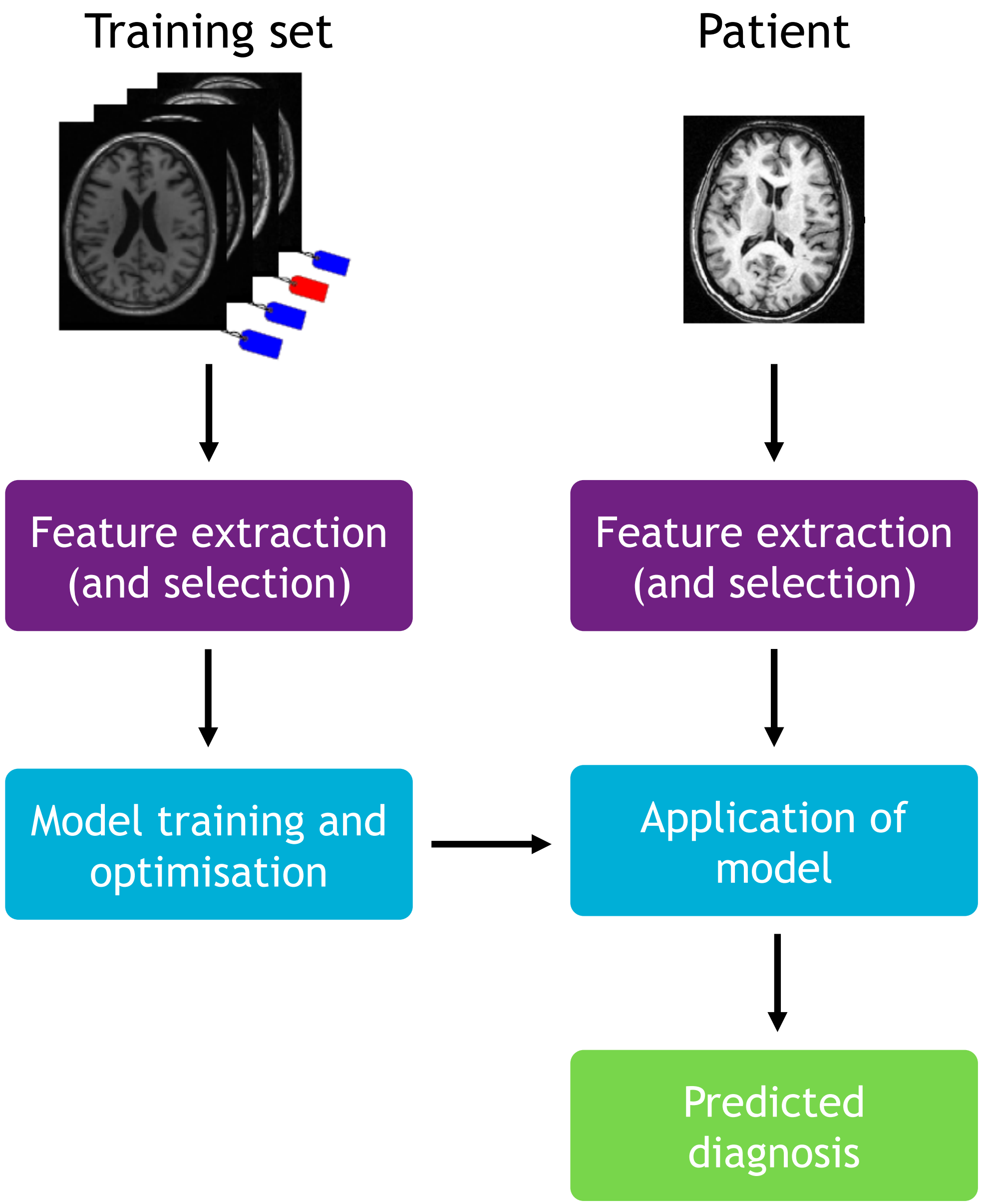
Elements that might differ between AD classification studies

- Training and test sets
- Imaging modality/ies
- Image preprocessing pipelines
- Features extracted
- Classification algorithms
- Cross-validation procedures
- Reported evaluation metrics



Clinical pattern recognition

- Where to find data?
- How to organise data?
- How to preprocess and extract features from images?
- Which classifiers can be selected?
- Which cross-validation strategy can be implemented?
- Which tasks may be of interest?
- What is the influence of these choices on the classification performance?



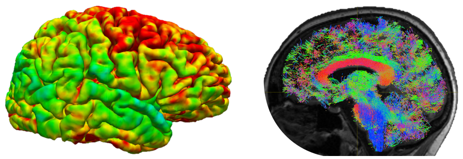
Software platform for clinical neuroimaging studies

Modalities

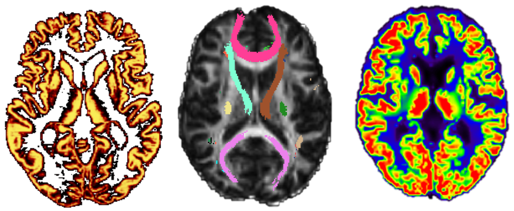
- Anatomical MRI
- Diffusion MRI
- PET

→

Preprocessing and Feature extraction



FreeSurfer, FSL, SPM,
MRtrix3, PETPVC, ANTS



→

Morphometric and statistical models

Machine learning / Deep learning

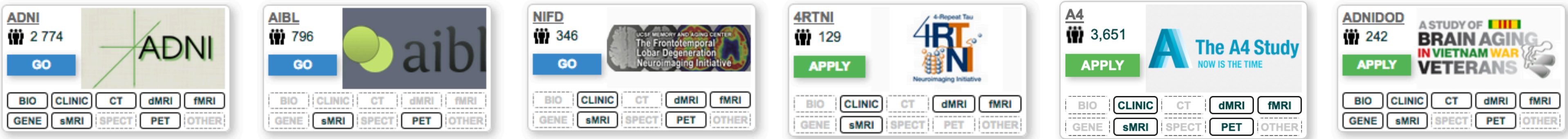
scikit learn
PyTorch

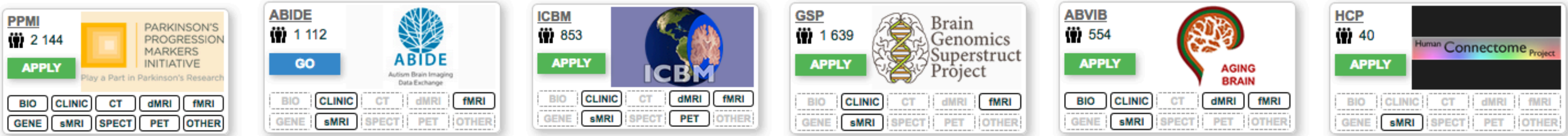
Standardized data structures: BIDS & CAPS

Finding data

Public datasets

- **Dementia**
  - Image and Data Archive (https://ida.loni.usc.edu)



  - Open Access Series of Imaging Studies (www.oasis-brains.org)
- **Other conditions**
  - 

  - BraTS (http://braintumorsegmentation.org)
  - IXI (https://brain-development.org/ixi-dataset)
  - etc.

Data organisation and curation


```
094_S_4089
...
|— Accelerated_SAG_IR-SPGR
|— AV45_Coreg,_Avg,_Standardized_Image_and_Voxel_Size
...
|— Average_DC
|— Axial_DTI
|— Axial_FLAIR
|— Axial_T2_Star
|— Calibration_Scan
|— Coreg,_Avg,_Standardized_Image_and_Voxel_Size
...
|— Eddy_current_corrected_image
|— EPI_current_corrected_image
|— Fractional_Ansio.
|— HarP_135_final_release_2015
|— HHP_6_DOF_AC-PC_registered_MPRAGE
|— MT1__GradWarp__N3m
|— Sag_IR-SPGR
|   |— 2011-06-29_14_37_16.0
|   |— 2011-10-18_12_15_56.0
|   |   └— S125692
|   |       |— ADNI_094_S_4089_MR_Sag_IR-SPGR__br_raw_20111019095510271_80_S125692_I261478.dcm
|   |       |— ADNI_094_S_4089_MR_Sag_IR-SPGR__br_raw_20111019095512256_62_S125692_I261478.dcm
|   |       |— ...
|   |— 2011-12-14_15_53_24.0
|   |— 2012-08-15_14_00_36.0
|   └— 2013-09-25_14_14_23.0
|— Sag_IR-SPGR_REPEAT
|— Spatially_Normalized,_Masked_and_N3_corrected_T1_image
└— T2-weighted_trace
```

```
sub-ADNI094S4089
|— ses-M00
|   |— anat
|   |   └— sub-ADNI094S4089_ses-M00_T1w.nii.gz
|   └— dwi
|       |— sub-ADNI094S4089_ses-M00_acq-axial_dwi.bval
|       |— sub-ADNI094S4089_ses-M00_acq-axial_dwi.bvec
|       └— sub-ADNI094S4089_ses-M00_acq-axial_dwi.nii.gz
|   |— pet
|   |   |— sub-ADNI094S4089_ses-M00_task-rest_acq-av45_pet.nii.gz
|   |   └— sub-ADNI094S4089_ses-M00_task-rest_acq-fdg_pet.nii.gz
|   └— sub-ADNI094S4089_ses-M00_scans.tsv
|— ses-M03
|— ses-M12
└— ses-M24
```

BIDS
BRAIN IMAGING DATA STRUCTURE

http://bids.neuroimaging.io
Gorgolewski et al., Nature Scientific Data, 2016

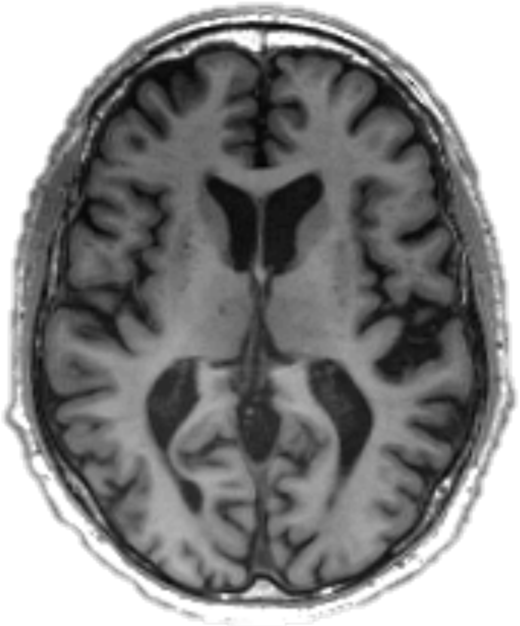
Data organisation and curation with Clinica

Converters available for:

- ADNI (Alzheimer's Disease Neuroimaging initiative)
- AIBL (Australian Imaging Biomarker & Lifestyle Flagship Study of Ageing)
- OASIS (Alzheimer's Disease and age-related dementia)
- NIFD (Neuroimaging in Frontotemporal Dementia)
- + internal studies to which we collaborate

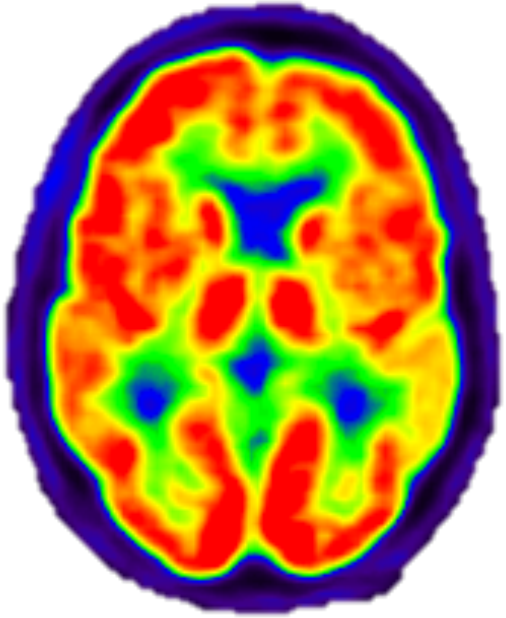
Imaging modalities

Example from the ADNI dataset:



Anatomical MRI
Atrophy

Diffusion MRI
White matter alterations



FDG PET
Hypometabolism

Amyloid PET
Protein aggregates

Image preprocessing

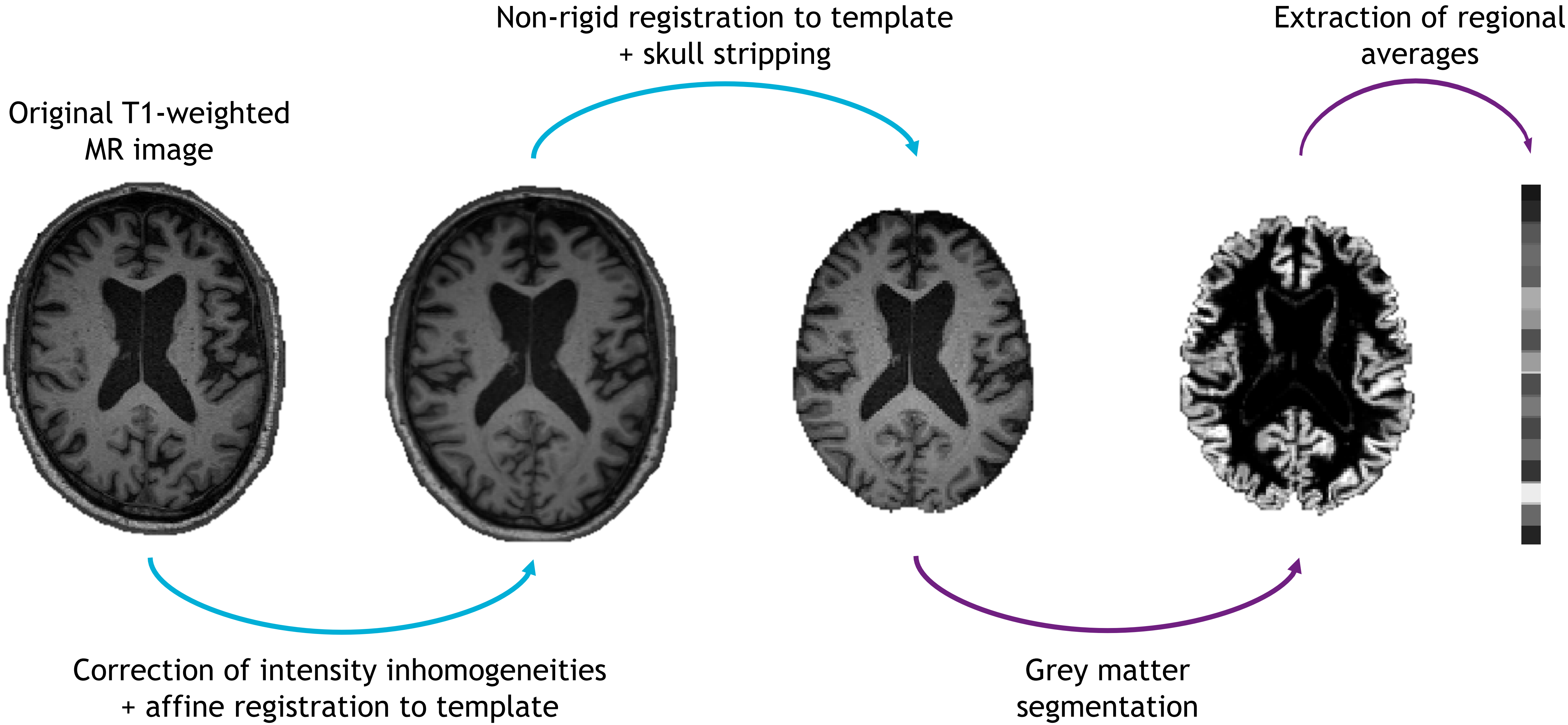


Image preprocessing

Statistical Parametric Mapping (SPM)

- www.fil.ion.ucl.ac.uk/spm
- **Modalities:** Structural and functional MRI, PET, SPECT, EEG, MEG
- **Features:** preprocessing, modelling, statistical inference, voxel-based morphometry, connectivity analysis

Frackowiak, Friston, Frith, Dolan, and Mazziotta, editors. Human Brain Function. Academic Press USA, 1997

FreeSurfer

- https://surfer.nmr.mgh.harvard.edu
- **Modalities:** Structural, functional, diffusion MRI
- **Features:** skullstripping, registration, cortical surface reconstruction, segmentation, longitudinal processing, fMRI analysis, tractography

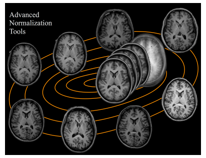
Fischl, NeuroImage, 2012

FMRIB Software Library (FSL)

- https://fsl.fmrib.ox.ac.uk
- **Modalities:** Structural, functional, diffusion MRI
- **Features:** brain extraction, segmentation, registration, tractography, longitudinal analysis, statistical analysis

Jenkinson et al., NeuroImage, 2012

Advanced Normalization Tools (ANTs)



- http://stnava.github.io/ANTs
- **Modalities:** Structural, functional, diffusion MRI, PET
- **Features:** bias field correction, registration, segmentation, cortical thickness estimation

Avants et al., Frontiers in Neuroinformatics, 2014

Clinica image (pre)processing pipelines

Anatomical MRI (T1-weighted)

- **t1-linear** Bias field correction and affine registration to standard space using ANTs
- **t1-volume** Tissue segmentation, spatial normalization and parcellation using SPM
- **t1-freesurfer** Cortical surface extraction, spatial normalization and parcellation using FreeSurfer

Diffusion MRI (DWI)

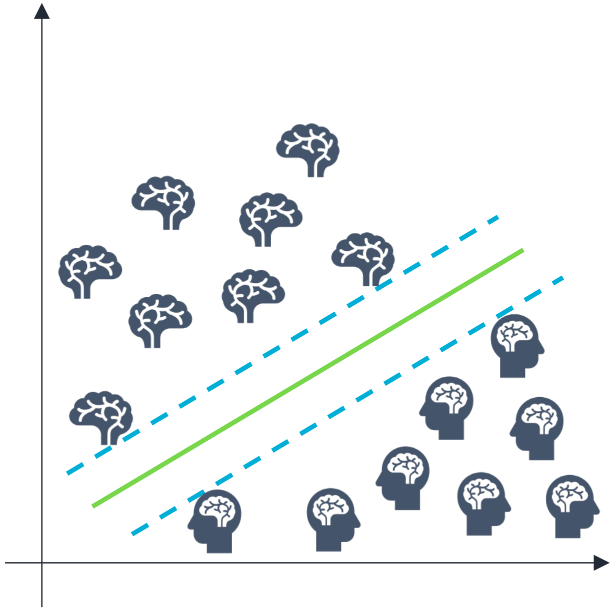
- **dwi-preprocessing-*** Correction of head motion, magnetic susceptibility, eddy current and bias field induced distortions
- **dwi-dti** Extraction of DTI-based measures and spatial normalization
- **dwi-connectome** Computation of fiber orientation distributions, tractogram and connectome

Positron emission tomography (PET)

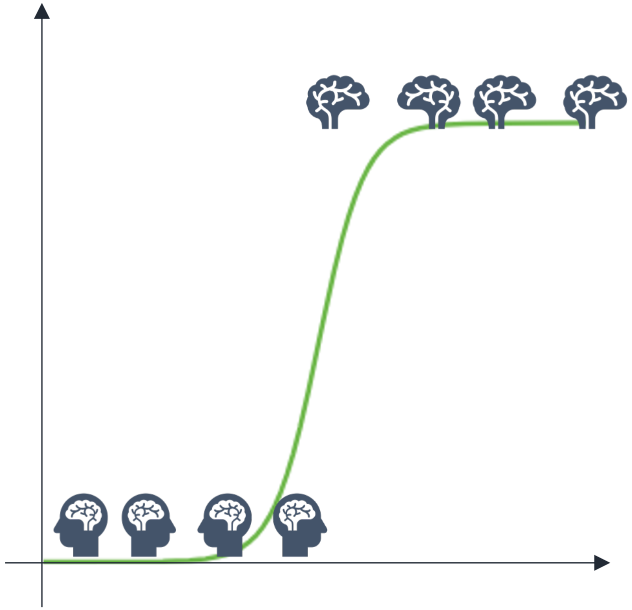
- **pet-volume** Registration to T1w MRI, intensity normalization, partial volume correction and spatial normalization
- **pet-surface** Registration to T1w MRI, intensity normalization, partial volume correction, projection of the PET signal onto the subject’s cortical surface and spatial normalization

Classifiers

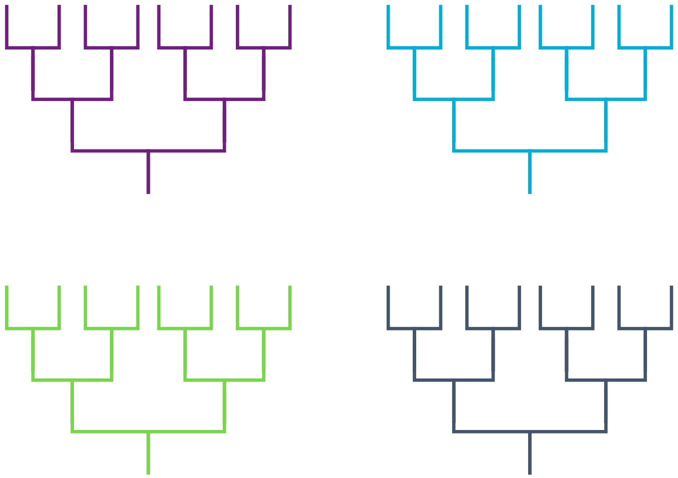
Linear SVM



$\ell 2$ logistic regression



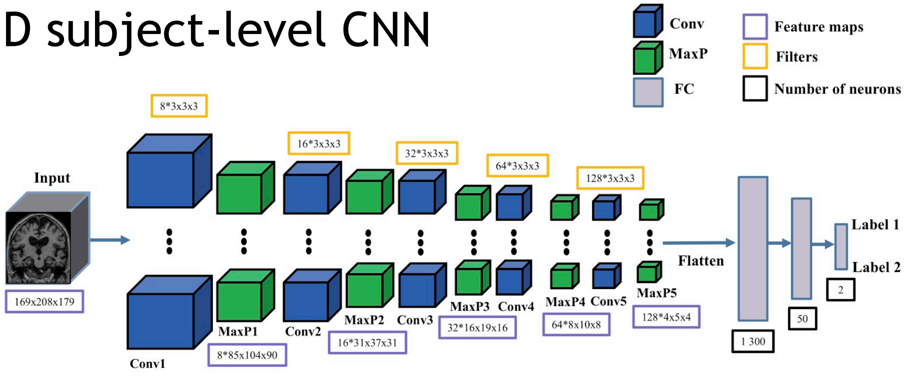
Random forest



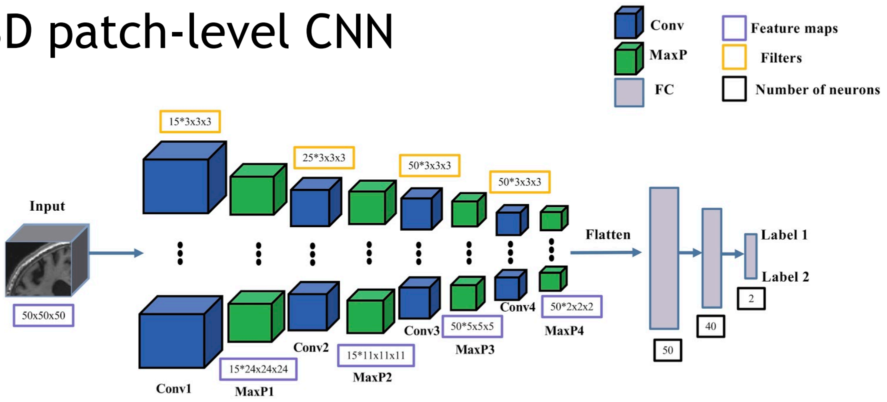
Pedregosa et al., JMLR, 2011; Samper-González et al., NeuroImage, 2018

Convolutional neural networks

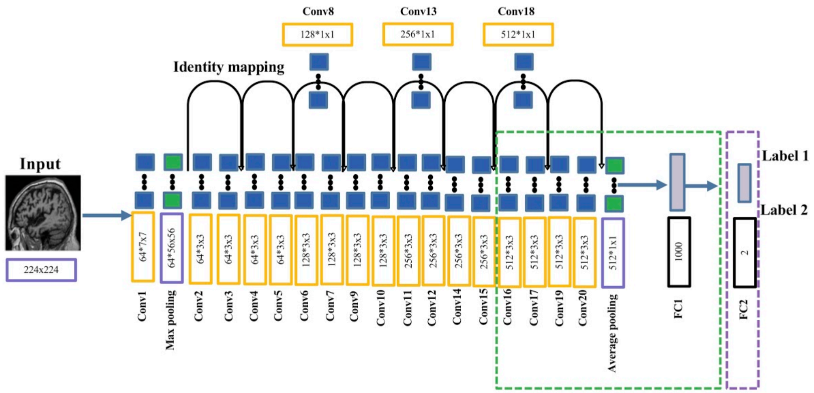
3D subject-level CNN



3D patch-level CNN



2D slice-level CNN

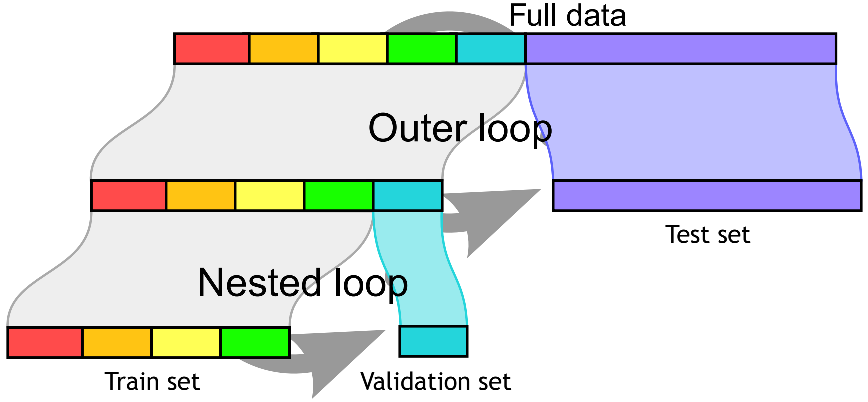


Wen, Thibeau-Sutre et al., Medical Image Analysis, 2020

Cross-validation procedures

Nested cross-validation

- Outer loop
- Nested loop



Adapted from Varoquaux et al., NeuroImage, 2017

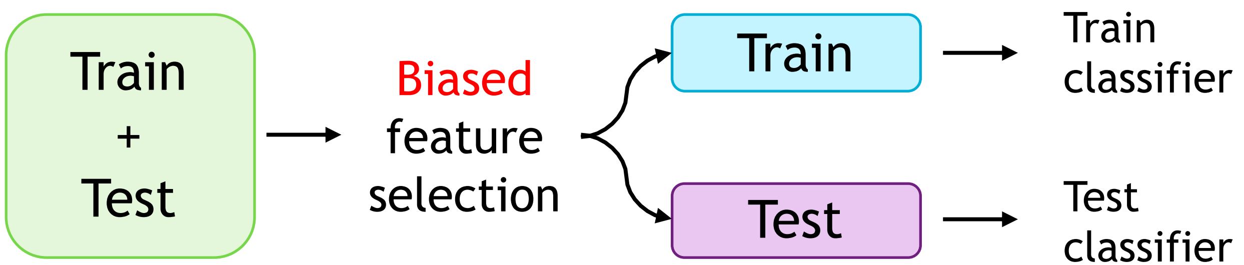
(Repeated) k-fold

Repeated hold-out

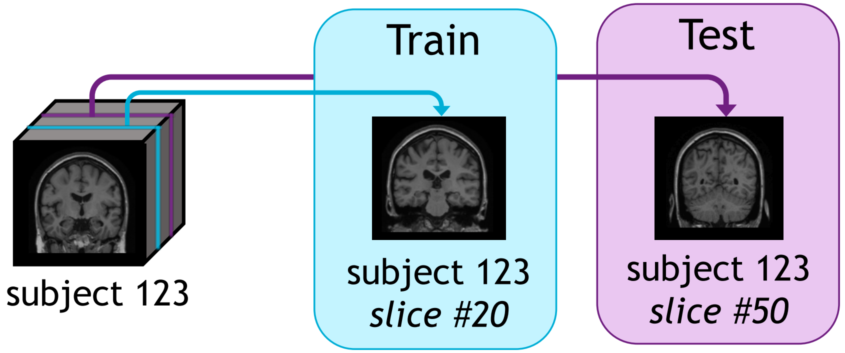
Samper-González et al., NeuroImage, 2018

Main causes of data leakage in DL scenarios

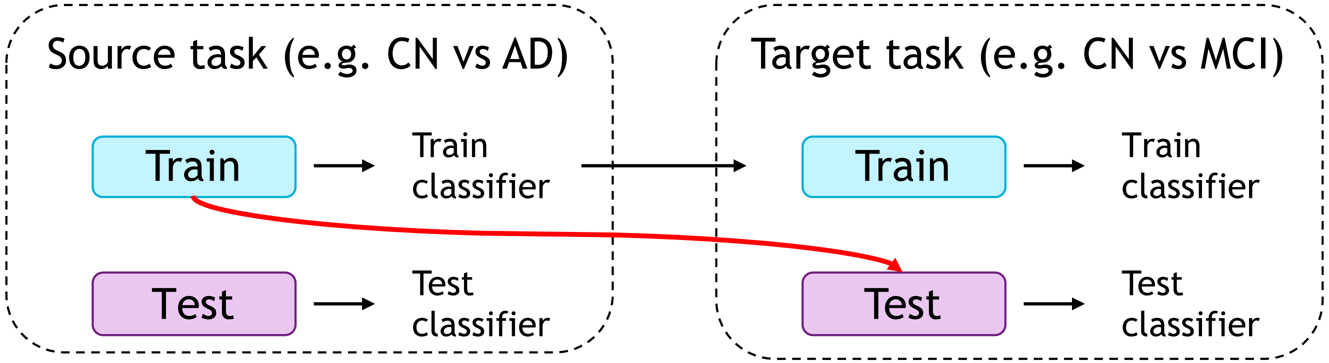
- Late split



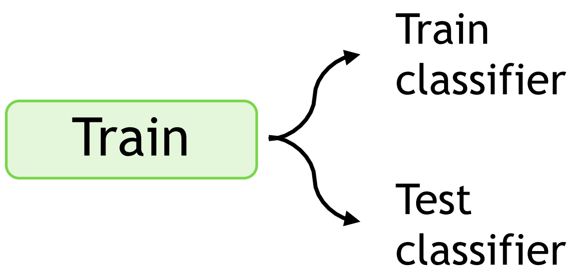
- Biased within-subject split



- Biased transfer learning



- No independent test set



Wen, Thibeau-Sutre et al., Medical Image Analysis, 2020

Task selection

ADNI dataset

	N	Age	Gender	MMSE	CDR
CN	282	74. 3 ± 5.9 [56.2, 89.0]	147 M / 135 F	29.0 ± 1.2 [24, 30]	0: 281; 0.5: 1
MCI	640	72.7 ± 7.5 [55.0, 91.4]	378 M / 262 F	27.8 ± 1.8 [23, 30]	0: 1; 0.5: 638; 1:1
AD	237	74.9 ± 7.8 [55.1, 90.3]	137 M / 100 F	23.2 ± 2.1 [18, 27]	0.5: 99; 1: 137; 2: 1
AD	237	74.9 ± 7.8 [55.1, 90.3]	137 M / 100 F	23.2 ± 2.1 [18, 27]	0.5: 99; 1: 137; 2: 1

Values are presented as mean ± SD [range].
M: male, F: female, MMSE: mini-mental state examination, CDR: global clinical dementia rating

- **Tasks selected:**
  - CN vs AD clinical diagnosis classification tasks
  - sMCI vs pMCI “predictive” task of the evolution

Samper-González et al., NeuroImage, 2018

Machine learning for AD classification

Inputs

Feature types:
- Voxels
- Regions

↓

Classifier

- Linear SVM
- L2 Logistic regression
- Random forest

↓

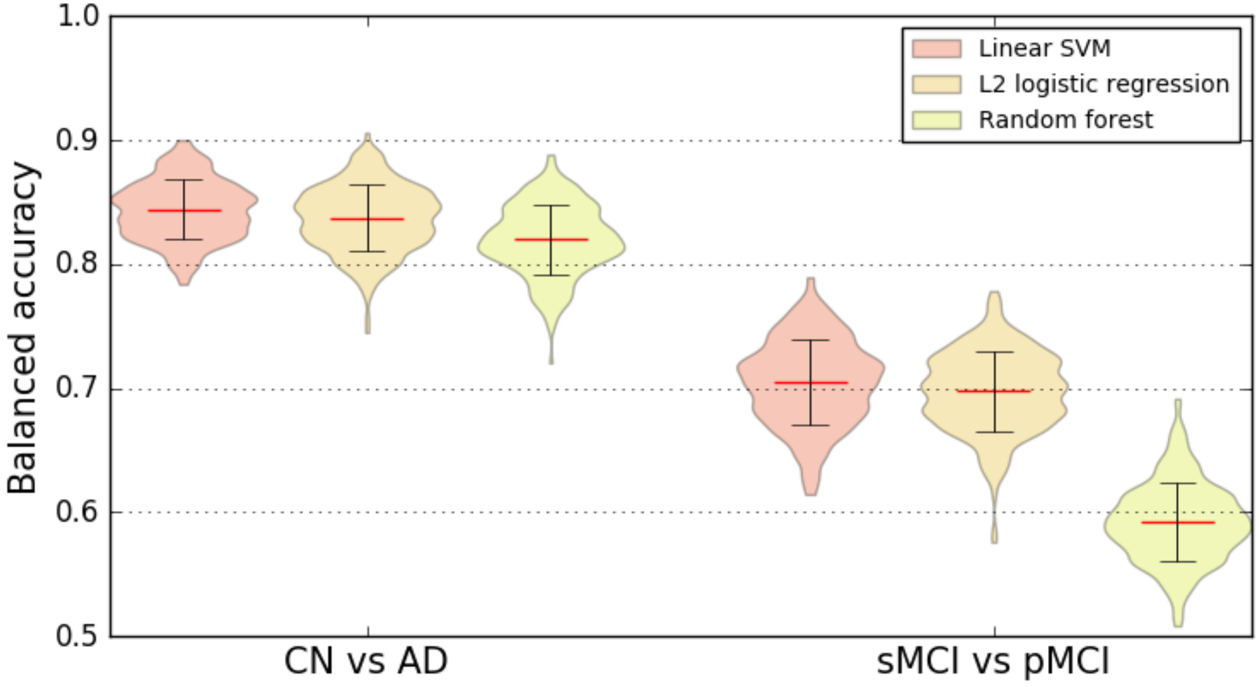
CV

- K-fold
- Repeated k-fold
- Repeated hold-out

Samper-González et al., NeuroImage, 2018;
https://gitlab.icm-institute.org/aramislab/AD-ML

Machine learning for AD classification

Influence of the classifier



- Linear SVM and logistic regression with L2 regularization: similar balanced accuracies
- Random forest: consistently lower balanced accuracy

Machine learning for AD classification

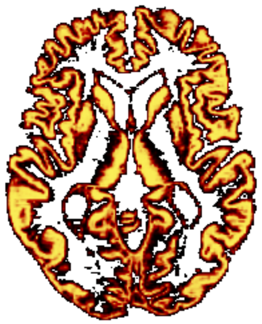
Influence of the type of features

	Linear SVM	
	Voxel-based	Region-based
CN vs AD	87% ± 2.6%	84% ± 2.4%
sMCI vs pMCI	66% ± 4.0%	70% ± 3.4%

Balanced accuracy — Values are presented as mean ± SD.

➢ No systematic effect

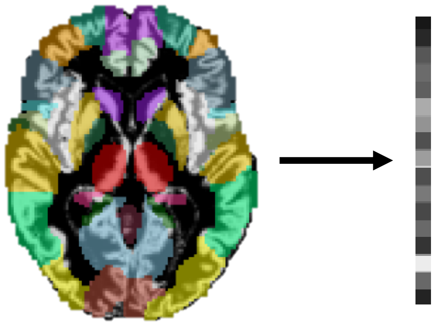
- Voxel-based features



$x \in \mathbb{R}^p$

where p is the number of voxels

- Region-based features



$x \in \mathbb{R}^p$

where p is the number of regions

Machine learning for AD classification

Generalisation across datasets

- Task: CN vs AD
- Subsets of equal size for each dataset (CN: 70, AD: 70)

Testing dataset	Training dataset	Voxel-based	Region-based
ADNI	ADNI	85% ± 4.8%	81% ± 6.0%
AIBL	AIBL	86% ± 4.8%	85% ± 5.8%
	ADNI	86%	87%
OASIS	OASIS	67% ± 6.3%	64% ± 7.2%
	ADNI	67%	70%

Balanced accuracy — Values are presented as mean ± SD.

➢ The classifiers trained on ADNI data generalise well

Deep learning for AD classification

Inputs

Preprocessing types:
- Minimal
- Extensive

↓

Classifier

- 3D subject-level CNN
- 3D patch-level CNN
- 2D slice-level CNN

↓

CV

- K-fold

Wen, Thibeau-Sutre et al., Medical Image Analysis, 2020;
https://gitlab.icm-institute.org/aramislab/AD-DL; https://zenodo.org/record/3491003

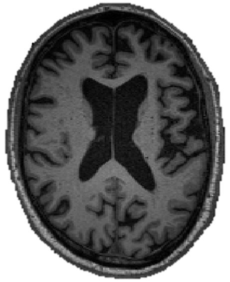
Deep learning for AD classification

Influence of the type of preprocessing

	3D subject-level CNN	
	Minimal	**Extensive**
CN vs AD	85% ± 4%	86% ± 6%

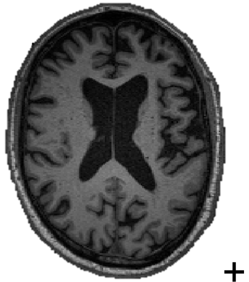
Balanced accuracy on the validation set
Values are presented as mean ± SD.

- Minimal preprocessing



Bias field correction
+ affine registration

- Extensive preprocessing



Bias field correction
+ non-rigid registration
+ skull stripping

➢ No systematic effect

Deep learning for AD classification

Influence of the type of network architecture

	3D subject-level	3D patch-level	3D patch-level (hippocampi only)	2D slice-level
CN vs AD	85%	86%	85%	74%
sMCI vs pMCI	73%	70%	74%	-

Balanced accuracy on the test set — Values are presented as mean ± SD.

- 3D subject-level and 3D patch-level approaches: similar balanced accuracies
- 2D-slice approach: consistently lower balanced accuracy

Deep learning for AD classification

Generalisation across datasets

- Training: ADNI

	3D subject-level		
	ADNI	**AIBL**	**OASIS**
CN vs AD	85%	86%	68%
sMCI vs pMCI	73%	50%	-

Balanced accuracy on the test set –
Values are presented as mean ± SD.

➢ **The models trained on ADNI data do not always generalise well**

Deep/machine learning for AD classification

Comparison deep learning / machine learning

- Training: ADNI

	3D subject-level			Linear SVM		
	ADNI	AIBL	OASIS	ADNI	AIBL	OASIS
CN vs AD	85%	86%	68%	87%	87%	71%
sMCI vs pMCI	73%	50%	-	76%	68%	-

Balanced accuracy on the test set — Values are presented as mean ± SD.

➢ Machine learning at least as good as deep learning

Reproducible evaluation of AD classification

Evaluation of machine learning and deep learning approaches in Alzheimer's disease

- **More reproducible**
  - Data sharing
  - Storing of data using community standards
  - Fully automatic data manipulation
  - Code sharing
- **More objective**
  - Baseline approaches against which new methods can easily be compared
  - Rigorous validation
    - Large number of repeated random split to extensively assess the performance variability
    - Reporting of full distribution of metrics
    - Adequate nested CV for hyperparameter tuning

Varoquaux et al., NeuroImage, 2017;
Samper-González et al., NeuroImage, 2018; Wen, Thibeau-Sutre et al., Medical Image Analysis, 2020

Reproducible evaluation of AD classification with Clinica

- Clinica
  - www.clinica.run
  - Preprint: https://hal.inria.fr/hal-02308126
- Reproducible evaluation of AD classification
  - Machine learning
    - https://github.com/aramis-lab/AD-ML
    - Samper-González et al., NeuroImage, 2018
    - Wen et al., Neuroinformatics, 2020
  - Deep learning
    - https://gitlab.icm-institute.org/aramislab/AD-DL
    - https://zenodo.org/record/3491003
    - Wen, Thibeau-Sutre et al., Medical Image Analysis, 2020

ARAMIS LAB
BRAIN DATA SCIENCE

Thank you!

ARAMIS LAB
BRAIN DATA SCIENCE

Olivier Colliot
Ninon Burgos
Stanley Durrleman

Michael Bacci
Simona Bottani
Mauricio Diaz
Sabrina Fontanella
Jérémy Guillon

Alexis Guyot
Thomas Jacquemont
Pascal Lu
Arnaud Marcoux
Tristan Moreau
Alexandre Routier
Jorge Samper-Gonzalez
Elina Thibeau--Sutre
Junhao Wen
Adam Wild

www.clinica.run

Software platform for clinical neuroimaging studies

Paris Brain Institute ICM · Inserm · ARAMIS LAB BRAIN DATA SCIENCE · cnrs · SORBONNE UNIVERSITÉ · Inria