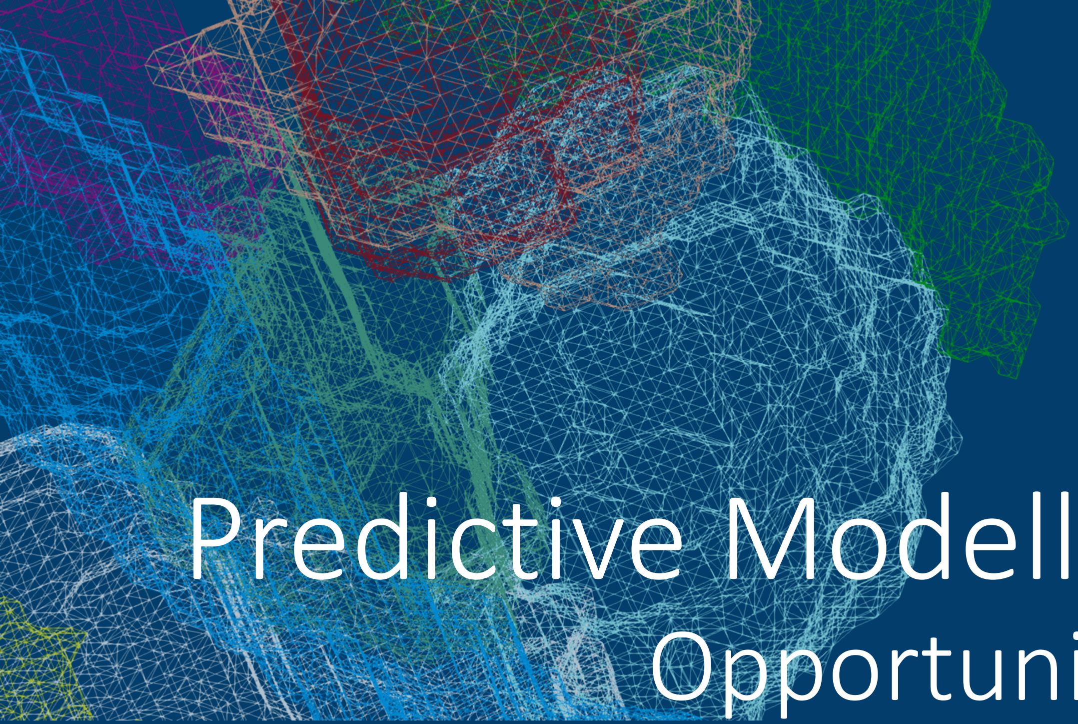


Predictive Modelling in Neuroimaging Data

Opportunities and Challenges

Kaustubh R. Patil <k.patil@fz-juelich.de>

Applied Machine Learning
Brain and Behavior (INM-7)
Forschungszentrum Jülich
Jülich, Germany

Institute of Systems Neuroscience
University Hospital
Henrich Heine University
Düsseldorf, Germany

Koita Center for Digital Health
IIT Bombay
Mumbai, India

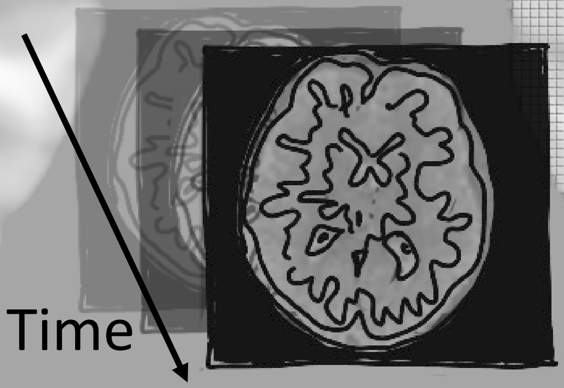
MRI takes 3D pictures of a brain with 100,000s of voxels

sMRI (T1w)
high resolution image

Structural images
(integrity)

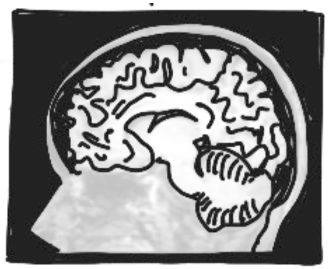
---

fMRI (T2w/EPI)
low resolution video

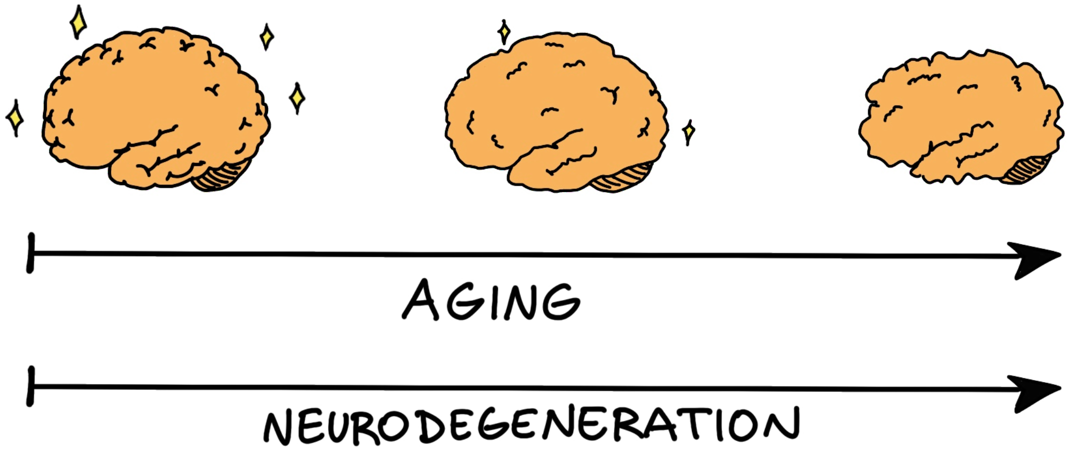


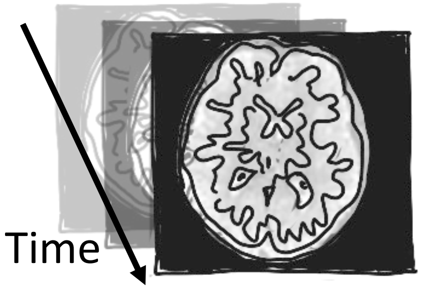
Functional images
(interaction)

MRI helps to study health and disease

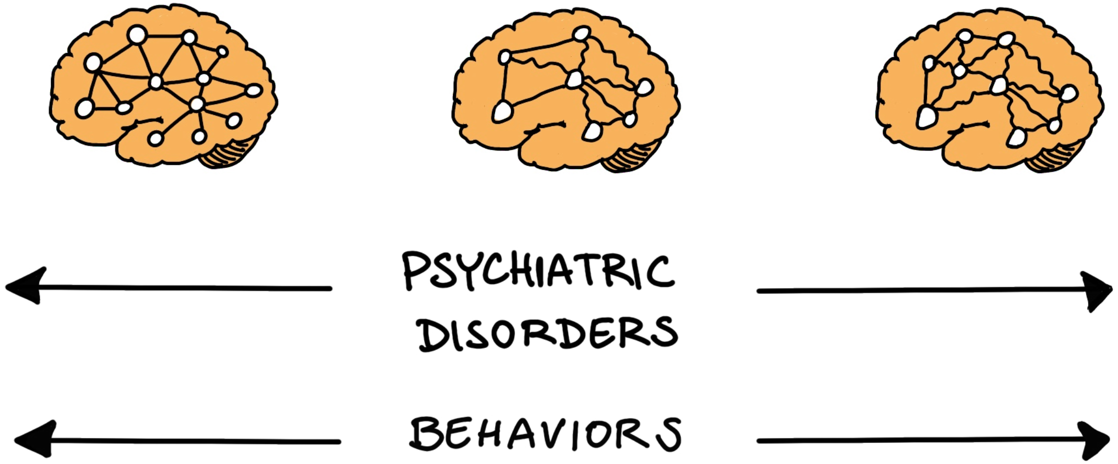


Structural images
(integrity)



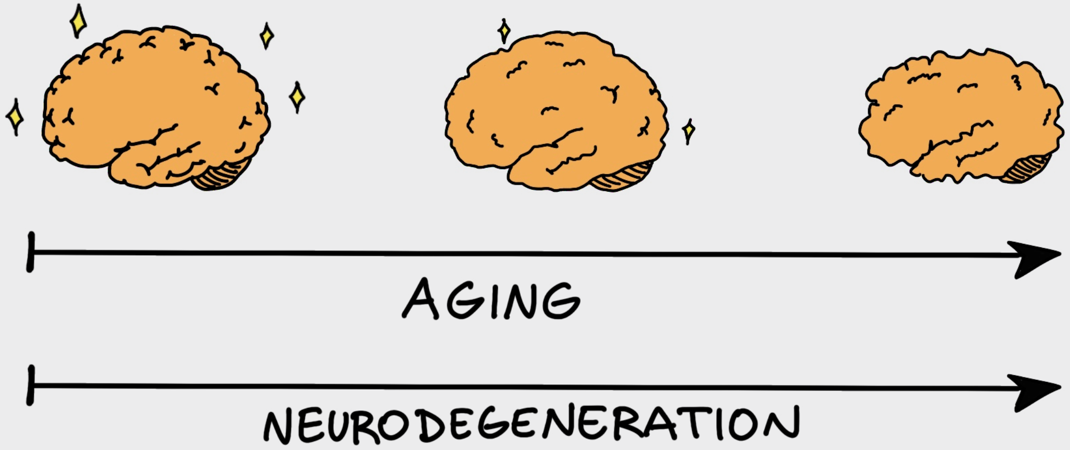


Functional images
(interaction)

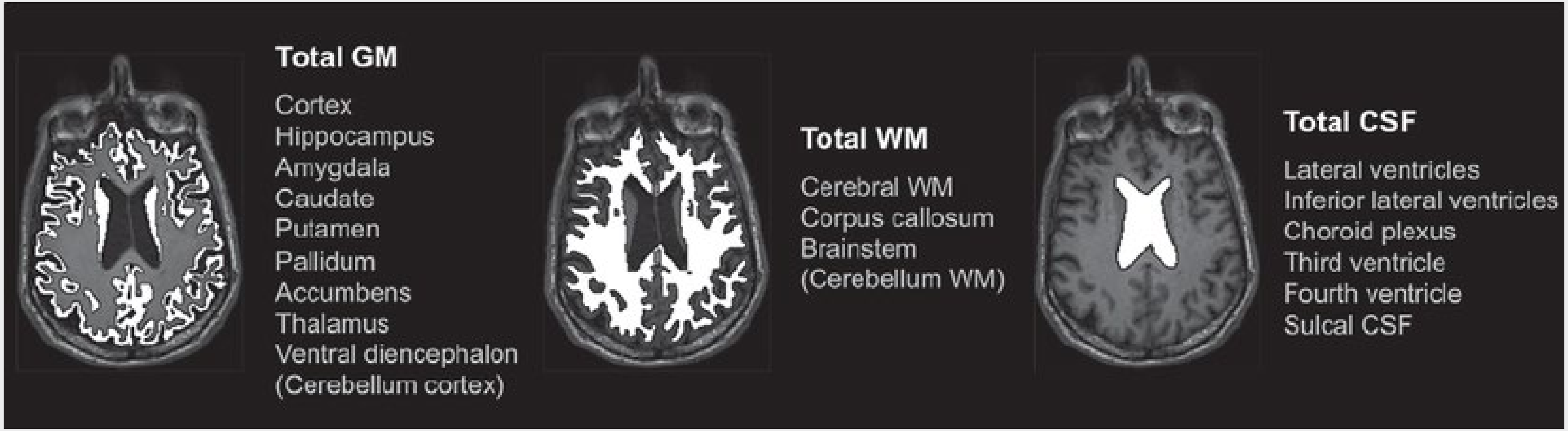


Structural MRI + Machine Learning

Structural images
(integrity)



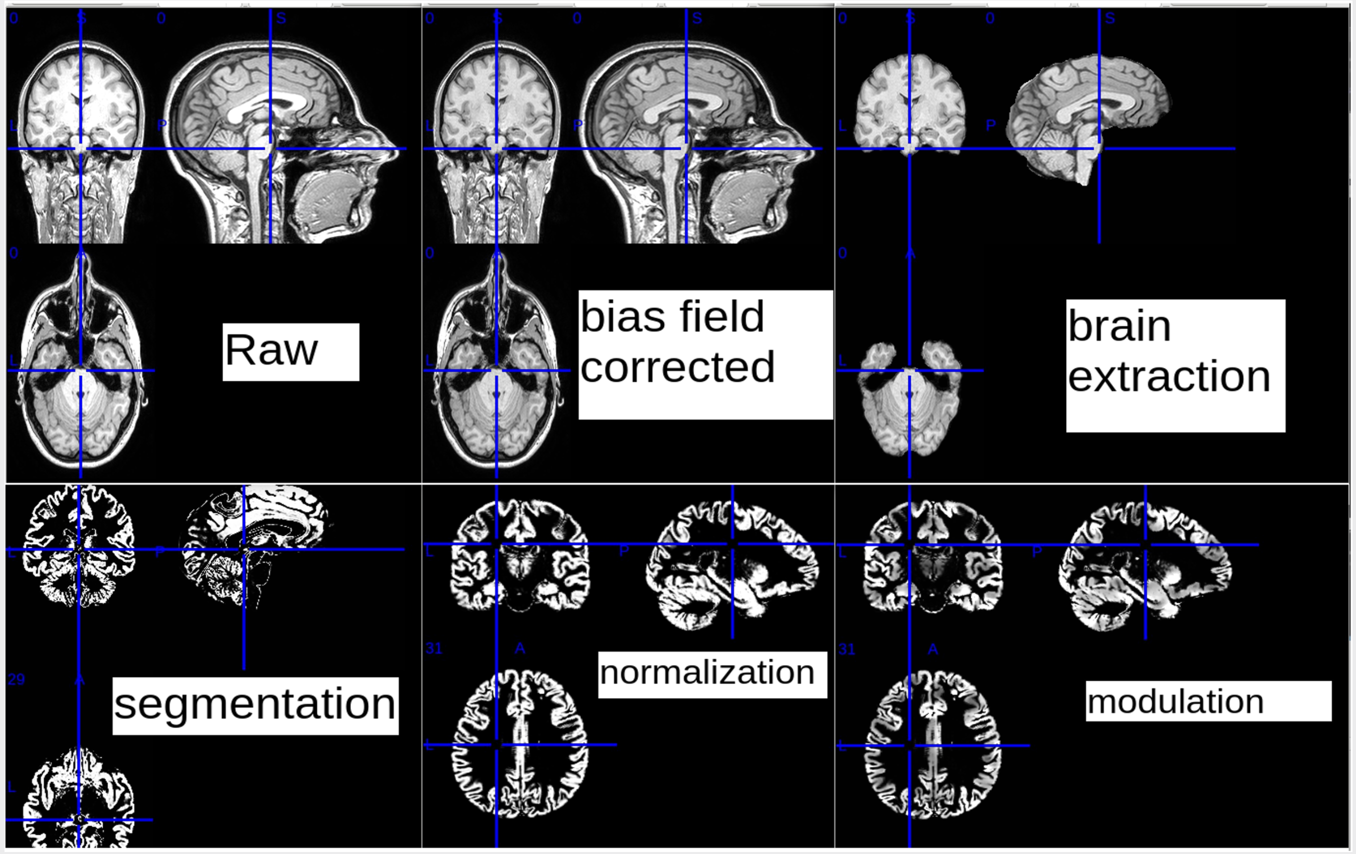
Structure of the Human Brain: T1w



- **Grey matter (GM)**: A thin layer surrounding the brain. Home to neural cell bodies, axon terminals, and dendrites, as well as all nerve synapses (cortical).
- **White matter (WM)**: Containing nerve fibres or axons, which are extensions of nerve cells or neurons. Found in the deeper tissues of the brain (subcortical).
- **Cerebro-Spinal Fluid (CSF)**: A clear and colourless fluid which surrounds the brain and spinal cord of all vertebrates.

Brain-age prediction: which features?

- sMRI to gray matter volume using voxel-based morphometry
- CAT toolbox
  - VBM tools comparison, Antonopoulos et al. (Under Review)



There are several ways to **extract features** from this voxel-wise data:

- Parcel-wise averages, different parcellation schemes
- Voxel-wise, different resampling and smoothing

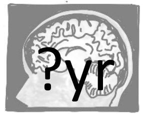
Which is the best choice?

And there are many **ML algorithms**

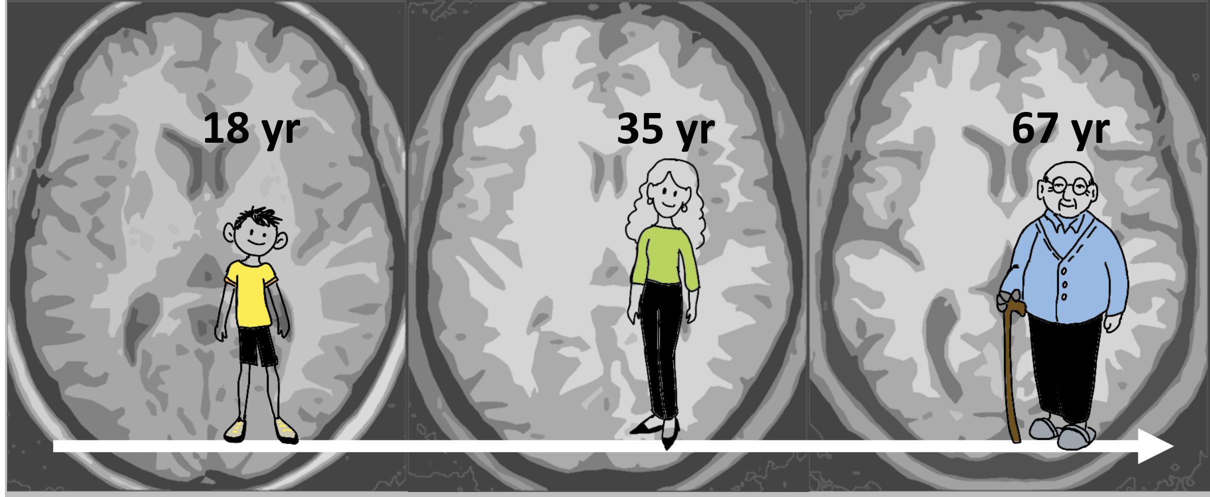
- Gaussian Process, Ridge, LASSO, Random Forests etc.

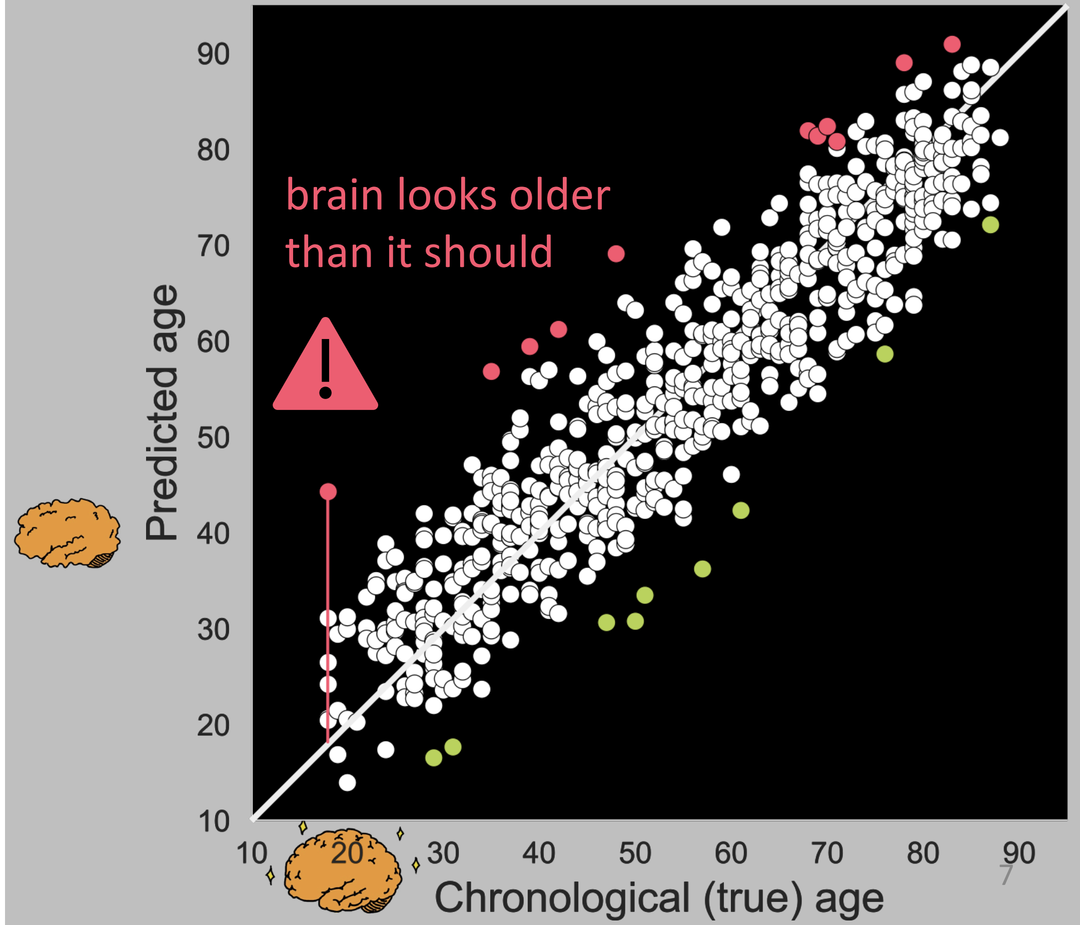
Can a 5-minute MRI scan reveal your age?

- Train a ML model using a database with sMRI images of many individuals
- Indeed, we can predict ~4-5 years!
- Predicted age > Actual age:
  - Brain-age delta = Predicted - Actual
  - Abnormal ageing: early warning system
  - In many cases brain changes happen (years) before external symptoms
  - Disorder-agnostic: Alzheimer's, Parkinson's, Schizophrenia



New indiv.

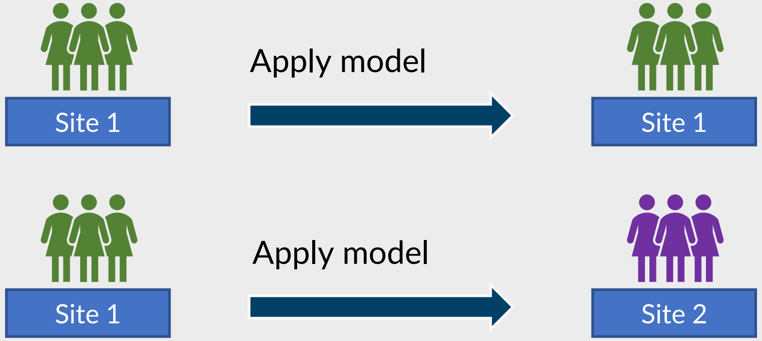




Shammi
More

Brain-age prediction: what do we need?

- **Generalizability:** work well on new data from the training site as well as on data from new sites

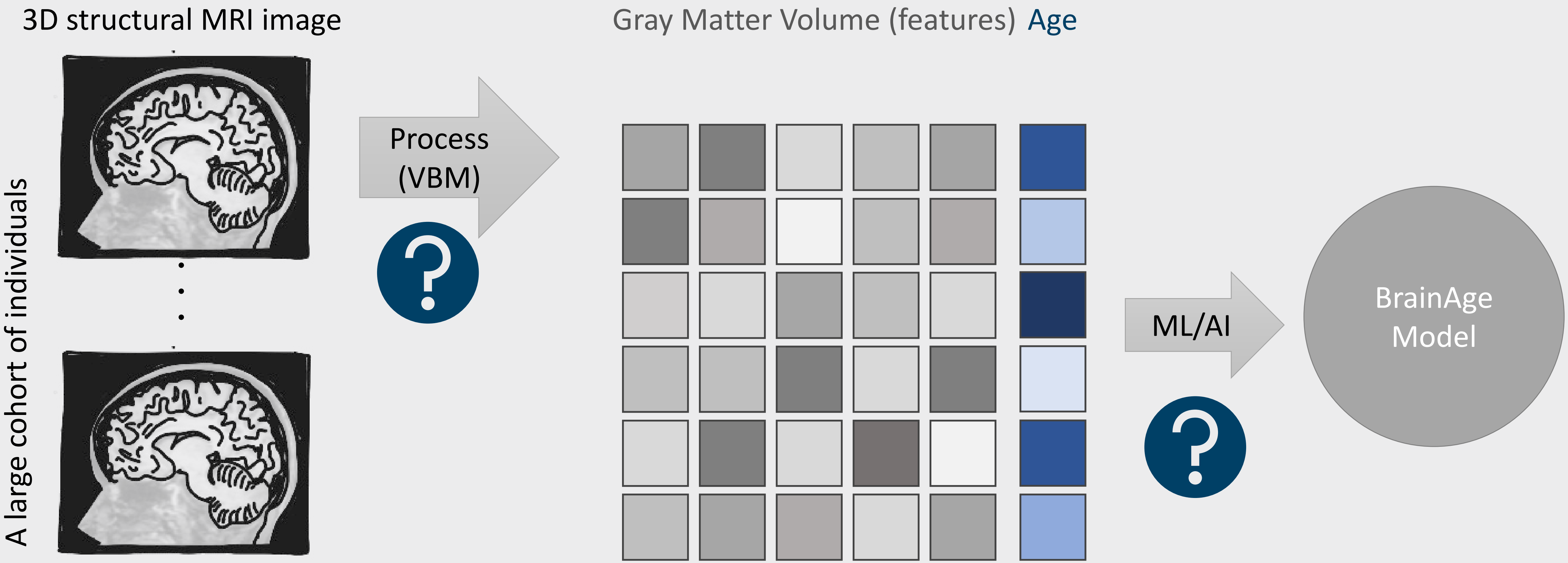


- **Reliability:** Estimated age must be reliable on repeated measurements

- **Longitudinal consistency:** the predicted age should be proportionally higher for later scans after a longer duration

More et al., NeuroImage 2023

How to build a brain-age model?



3D structural MRI image

A large cohort of individuals

Process (VBM)

Gray Matter Volume (features) Age

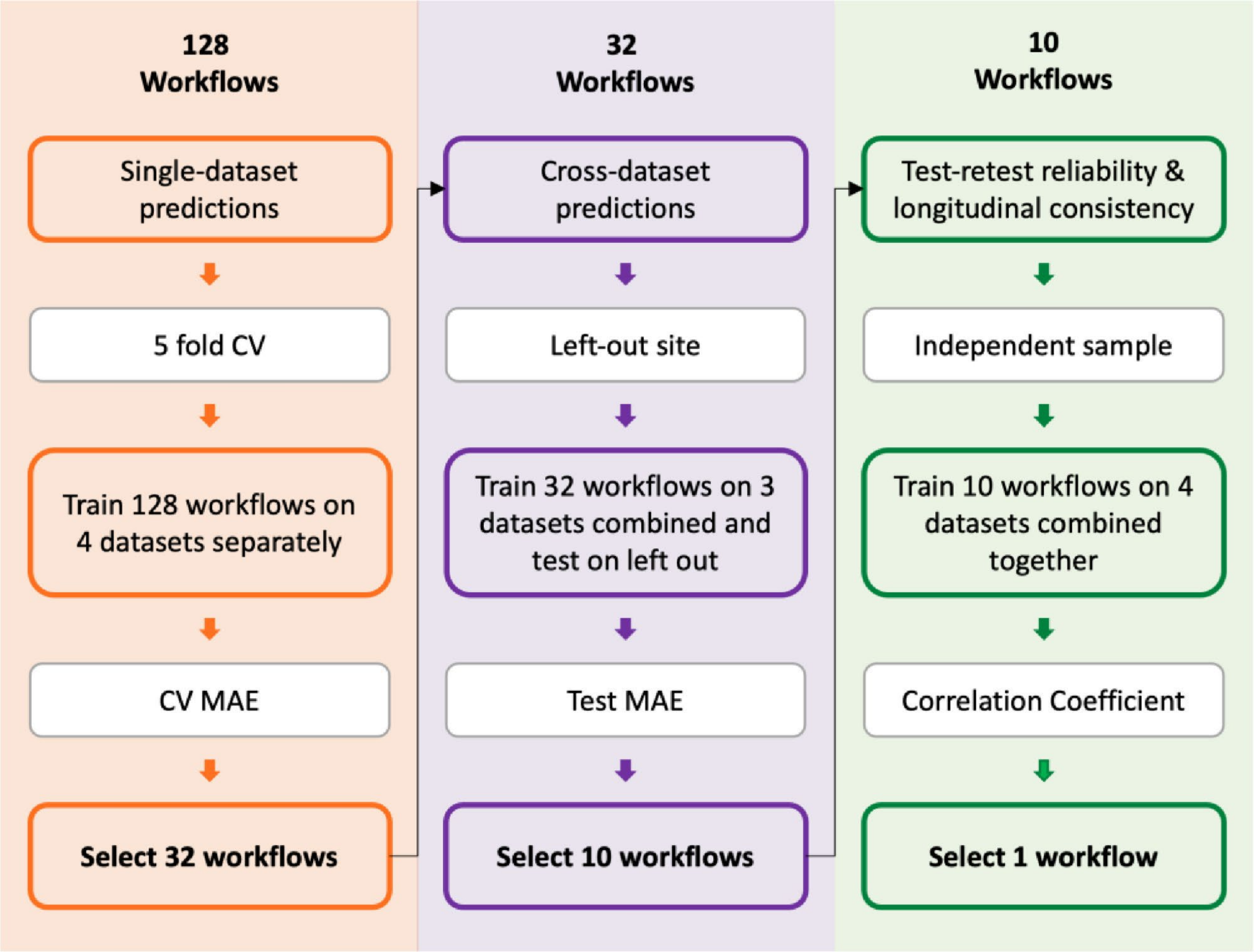
ML/AI

BrainAge Model

Brain-age prediction: which workflow?

Comparison of 128 workflows

- 16 feature spaces
  - Voxel- and parcel-wise
- 8 ML algorithms
  - LASSO, GPR, RF, RVRlin, RVRpoly, Ridge



128 Workflows

Single-dataset predictions

5 fold CV

Train 128 workflows on 4 datasets separately

CV MAE

Select 32 workflows

32 Workflows

Cross-dataset predictions

Left-out site

Train 32 workflows on 3 datasets combined and test on left out

Test MAE

Select 10 workflows

10 Workflows

Test-retest reliability & longitudinal consistency

Independent sample

Train 10 workflows on 4 datasets combined together

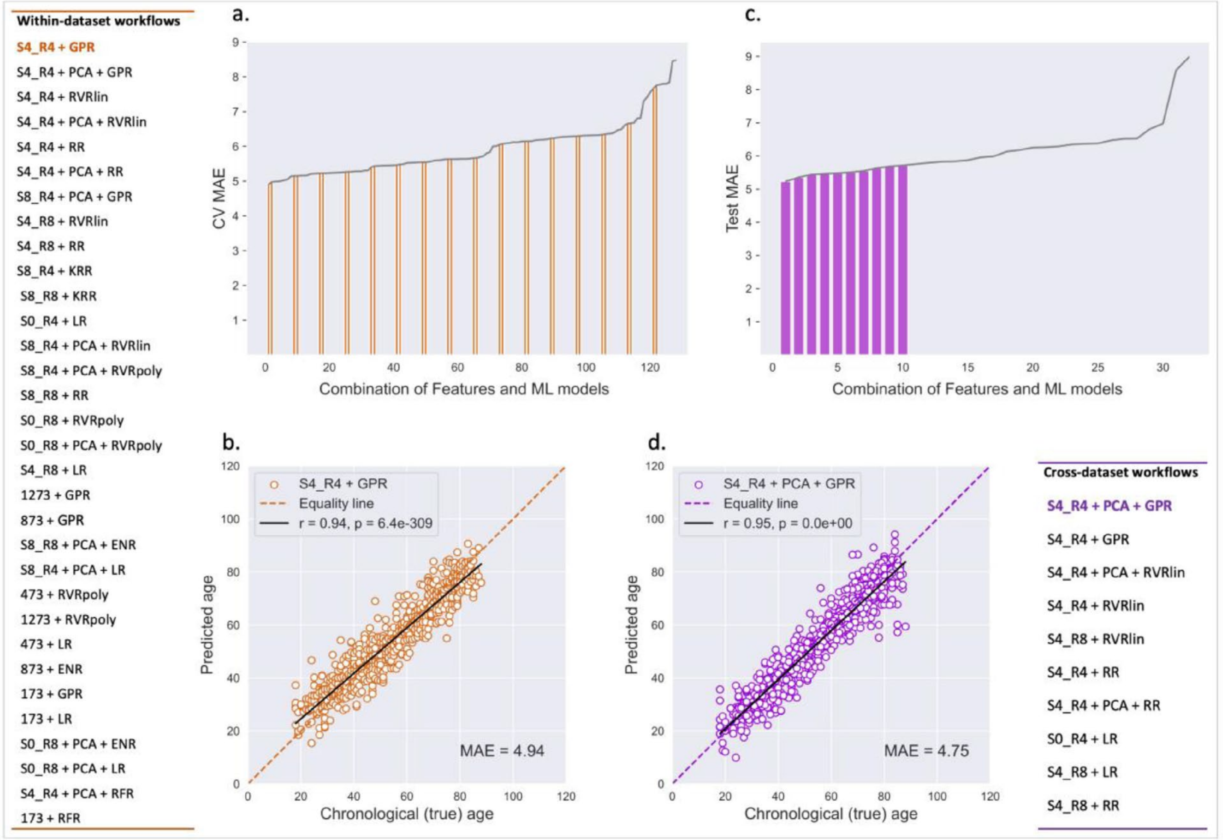
Correlation Coefficient

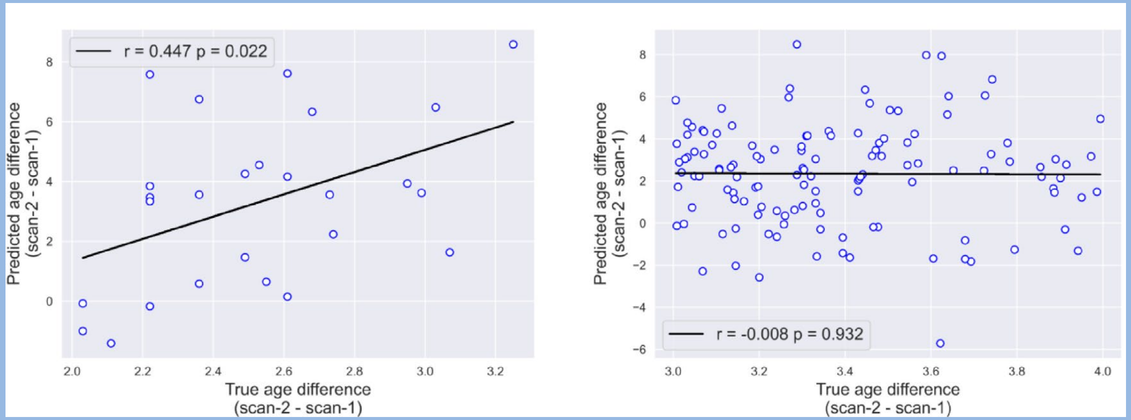
Select 1 workflow

Brain-age workflow selection

- S4_R4: smoothing with 4mm FWHM and resampling to 4x4x4 voxels
- GPR and RVR perform well
- It fulfills most of the desiderata
    - Within and cross-dataset generalization
    - Reliability

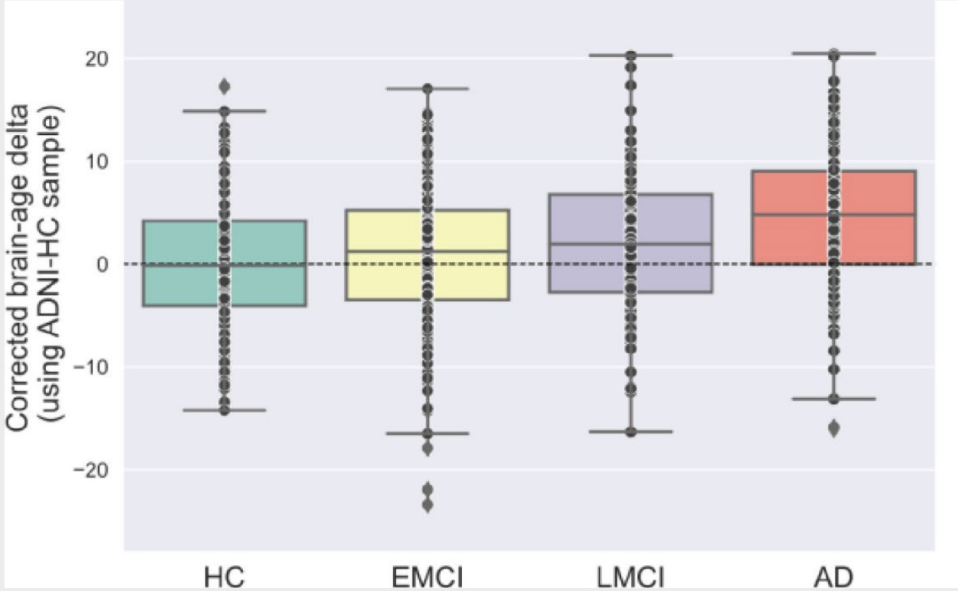
Difficult to achieve longitudinal consistency

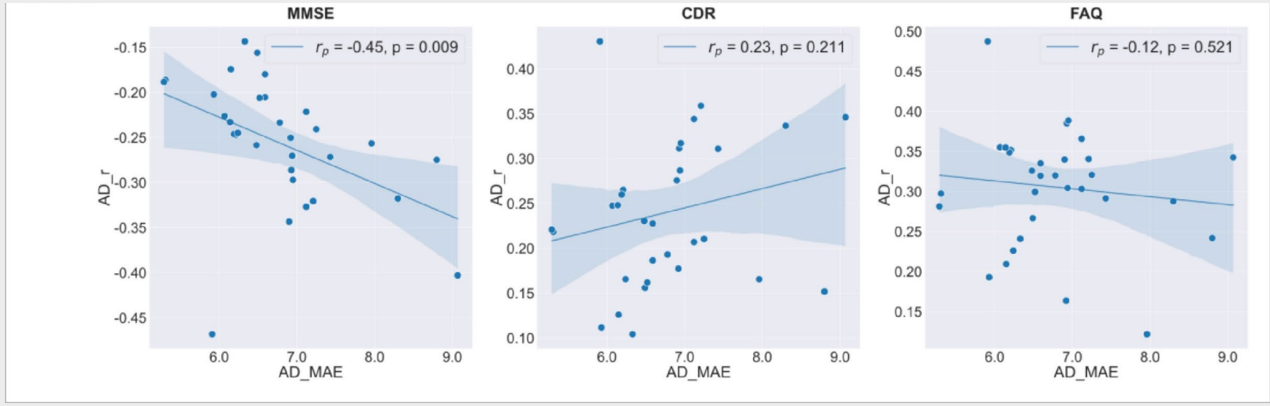




Brain-age application to Alzheimer’s disease

- ADNI dataset
  - Healthy control (HC), early and late mild cognitive impairment (EMCI, LMCI), and Alzheimer's disease (AD)
- Bias correction is needed
  - Models show systematic correlation with age
- AD indeed shows higher delta (deviation of predicted age from chronological age)
- The “brain-age delta” also correlates with cognitive scores

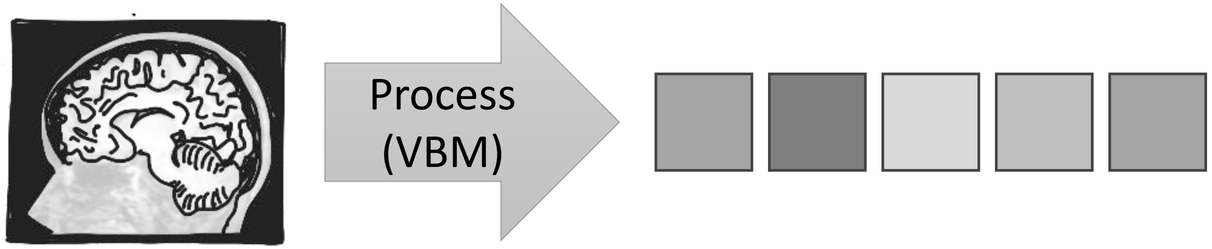




Mini-Mental State Examination (MMSE), Clinical Dementia Rating (CDR), Functional Activities Questionnaire (FAQ)

Geo.
Anto.

Which Voxel-based Morphometry Pipeline?

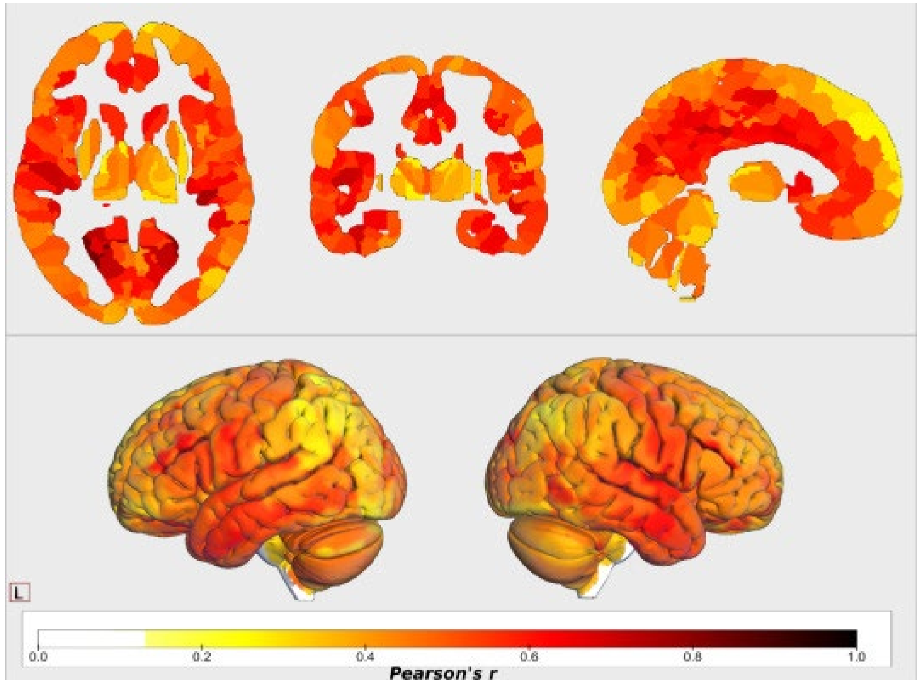


3D structural MRI image

Process (VBM)

Gray Matter Volume Features

- Process: Voxel-based Morphometry
- Several software tools are available
- They produce quite different GMV estimates!
- What to use for brain-age?
    - CAT 12.8 or fMRIPrep+FSL
    - Use a general template

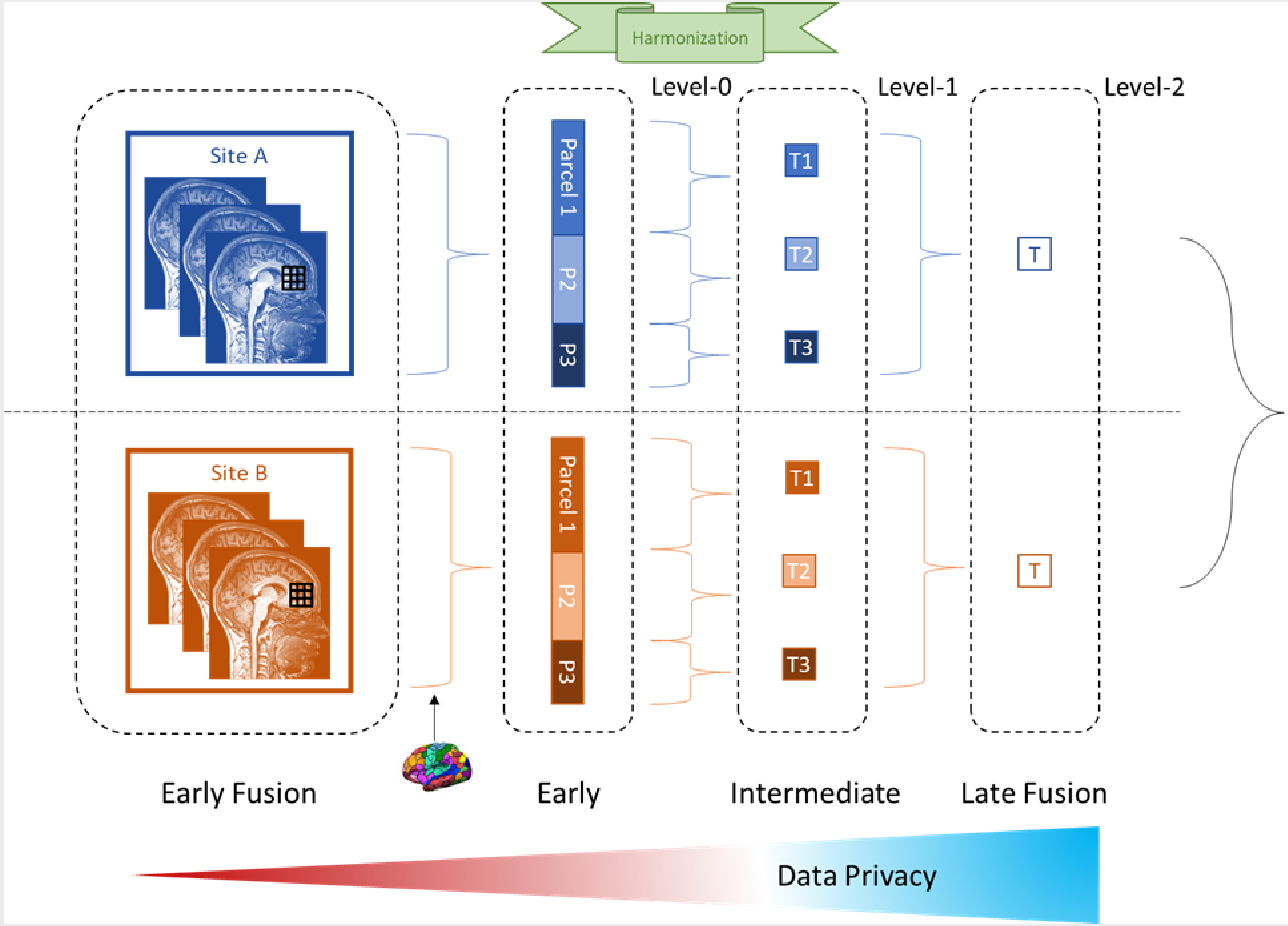


Median GMV-Age correlation across VBM pipelines is rather low!

Antonopoulos et al., NeuroImage 2023

Stacking for better accuracy and privacy

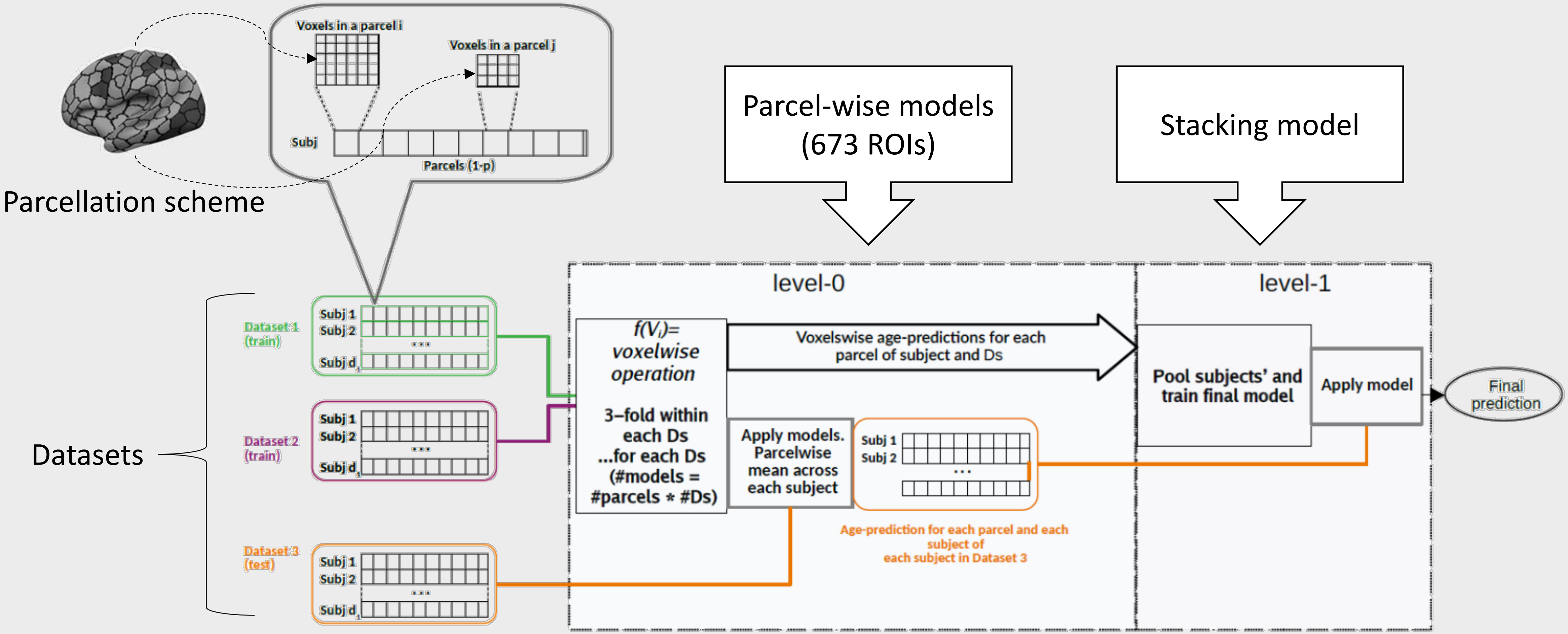
- Built-in harmonization
  - Age means the same across sites
- Better interpretation
- Improved data privacy
  - Controlled sharing of train/test data
  - Distributed learning



Patent filed, MS in preparation

A stacking model

Build a model for each brain region and stack them!



Parcellation scheme

Datasets

Parcel-wise models (673 ROIs)

Stacking model

Conventional parcel-wise model is **average**, e.g., hippocampal volume is used for tracking AD.

Stacking parcels

Improved accuracy

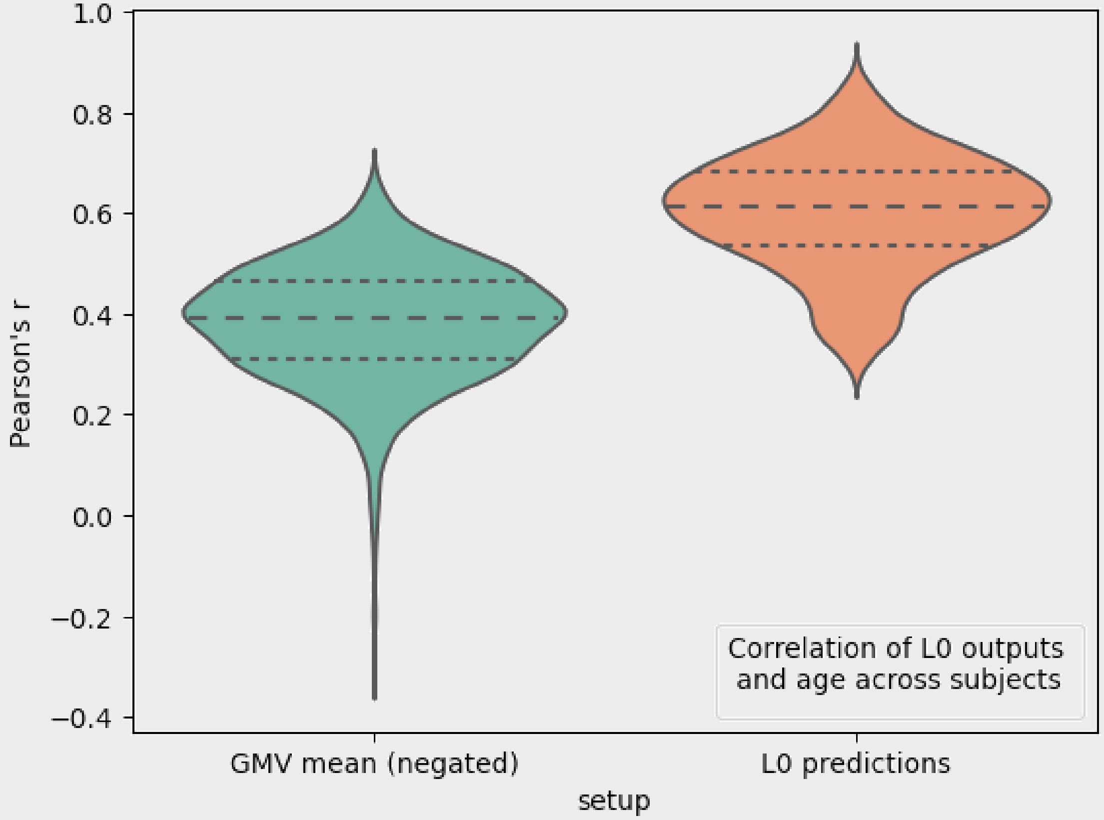
- Parcel-wise mean is least accurate (conventional method)
- Parcel-wise stacking is better
    - Best is pooled L1 models (MAE=4.69)
- L0 & L1 models at each training site, L0 from test site (MAE = 4.93)
    - Most private for both training and test sites

Set up	Pooling	MAE
Mean GMV l1 per site	None	6.70
l0 & l1 per site		5.12
l0 oos-test l0 & l1 per site		*4.93*
l0 per site l1 pooled	Average of site-wise l0 predictions	5.19
l0 oos-test **l1 pooled**		**4.69**
Mean GMV l1 pooled	Prior to l0 training	6.35
l0 pooled l1 pooled		4.97
l0 oos-test l1 pooled		4.76

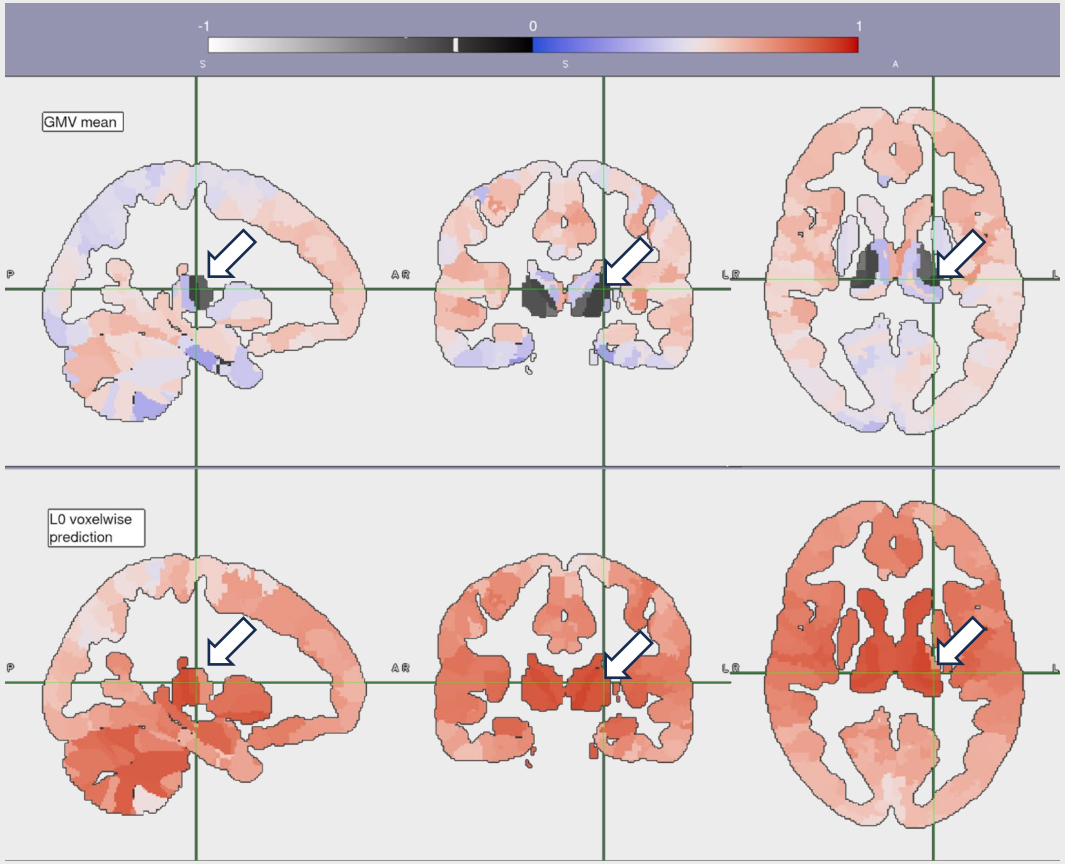
MAE = Mean Absolute Error

Stacking parcels

Improved interpretability



Higher correlation of each parcel with age



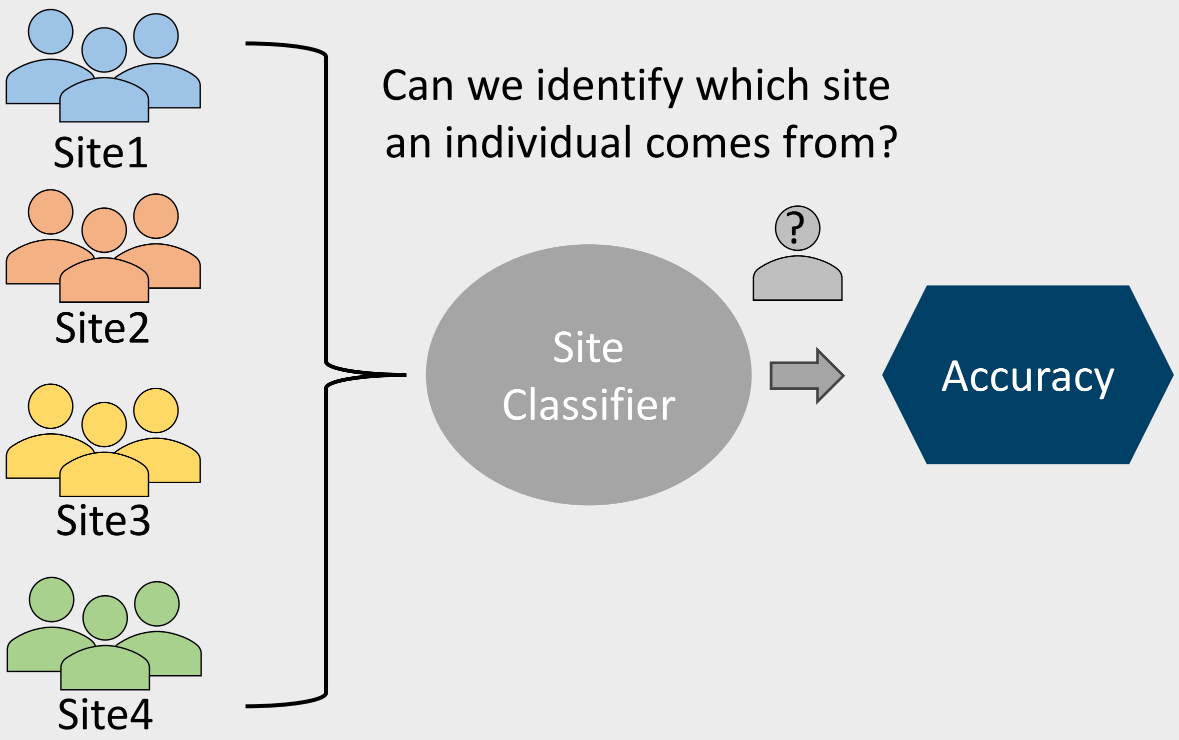
Subcortical regions show up as related to age

Stacking parcels

Improved privacy

- Data sharing, especially patient, raises privacy concerns
  - Privacy preserving methods are needed

✓L0-level predictions, i.e. age, provide a solution



Can we identify which site an individual comes from?

Lower Accuracy = Higher Privacy (*caveats)	
GMV mean	87%
L0 predictions	**63%** ✓

What about other modalities?

BrainAge with FDG-PET

- FDG-PET (Fluorodeoxyglucose Positron Emission Tomography) measures cellular metabolism
  - Reflecting the level of activity in different tissues

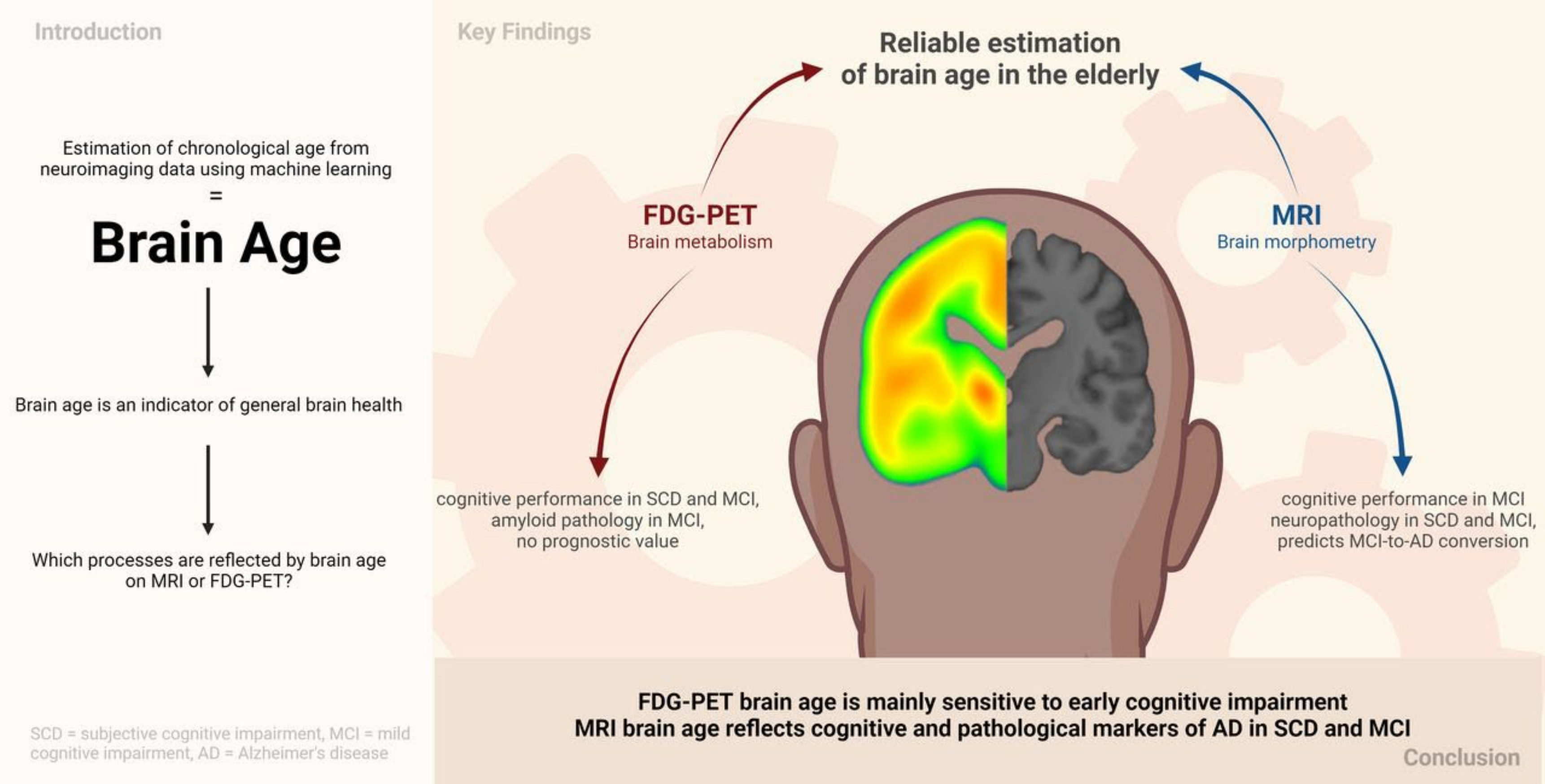
- Used in diagnosing and assessing the progression of neurodegenerative diseases like Alzheimer's
  - Affected brain areas show reduced glucose metabolism.

- PET data scarcer than MRI

Elena
Doering

Differing utility of FDG-PET and T1w MRI

- FDG PET: Early cognitive impairment
- T1w MRI: Subjective and mild cognitive impairment
- Brain-age gap predictive of MCI-to-AD conversion on par with clinical markers, e.g. P-tau/Aβ42 ratio



Estimation of chronological age from neuroimaging data using machine learning
=
Brain Age

Brain age is an indicator of general brain health

Which processes are reflected by brain age on MRI or FDG-PET?

SCD = subjective cognitive impairment, MCI = mild cognitive impairment, AD = Alzheimer's disease

Reliable estimation of brain age in the elderly

cognitive performance in SCD and MCI,
amyloid pathology in MCI,
no prognostic value

cognitive performance in MCI
neuropathology in SCD and MCI,
predicts MCI-to-AD conversion

FDG-PET brain age is mainly sensitive to early cognitive impairment
MRI brain age reflects cognitive and pathological markers of AD in SCD and MCI

Doering et al., J Nucl Med 2024

Brain-age prediction: ongoing work

Applications

- Alzheimer’s: is accuracy the best metric for model selection?
- Schizophrenia: brain ageing and behavioral interventions
- Astronauts: effect of space travel

Methodological

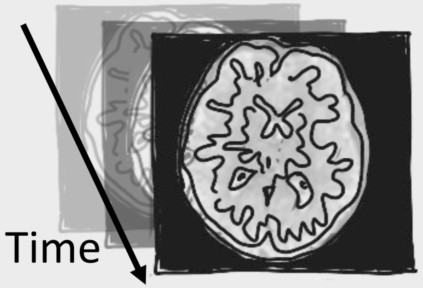
- Stacking: integrate information within brain regions and across clinical cohorts.
- Clinical standards validation
- Deep neural network for rank consistent prediction.

Summary: Structural Imaging

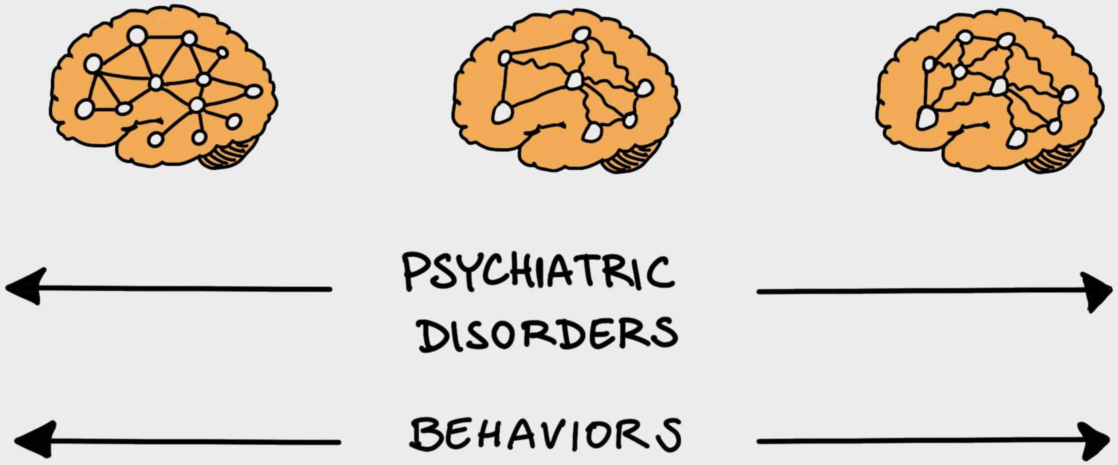
- sMRI can uncover structural organization of the human brain.
- Data analysis & ML can help understand brain structure organization and how it changes in health and disease.

- Several techniques can be used
  - Voxel-based morphometry (VBM)
  - Shape analysis
  - Surface and thickness
  - … and more

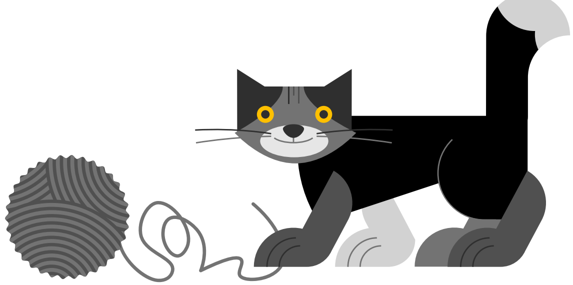
functional MRI + Machine Learning



Functional images
(interaction)



Task versus Resting-state fMRI

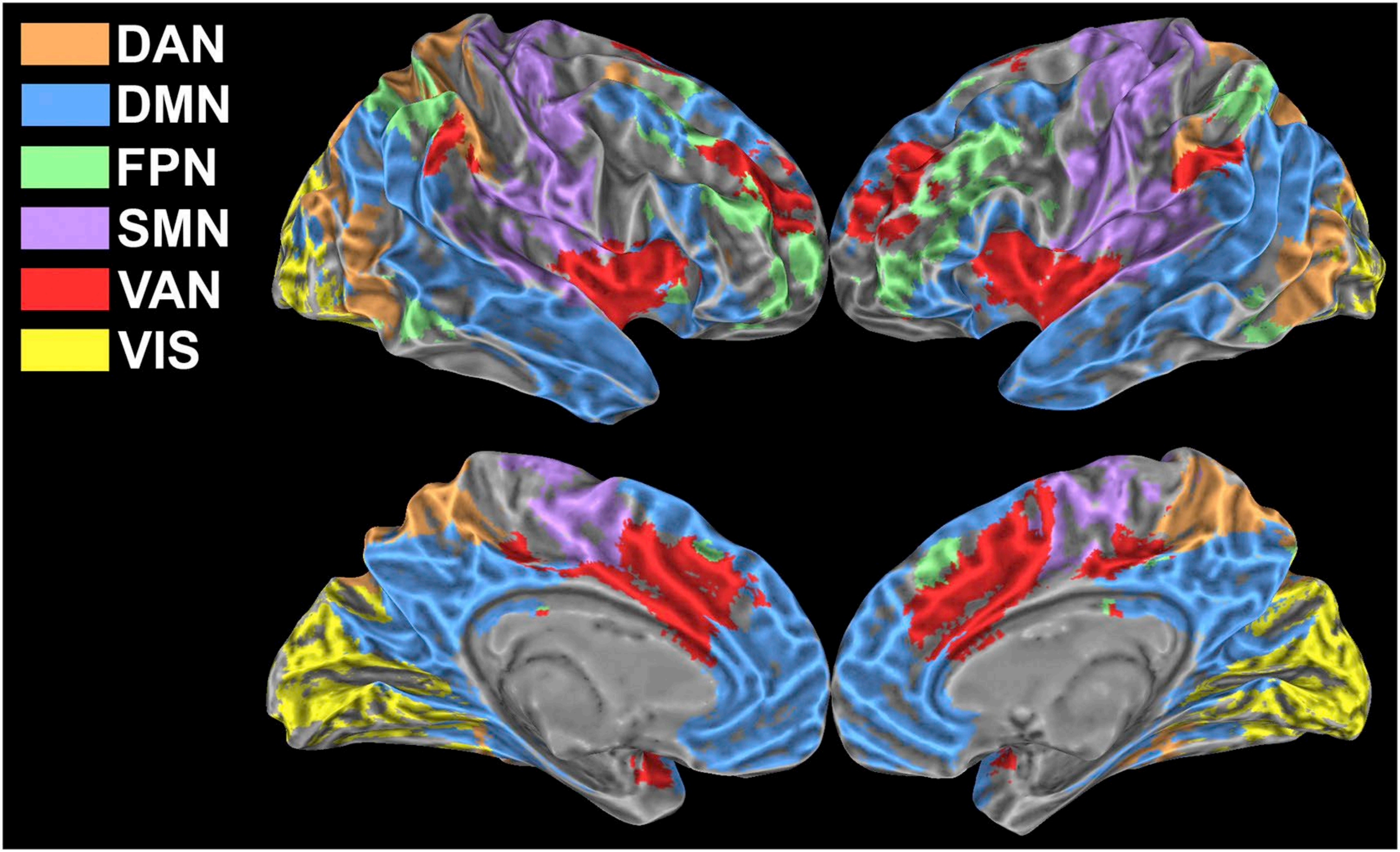


- **Task fMRI (tfMRI)**
  - Explicit task: e.g., finger tapping
  - Specific questions with a hypothesis

- **Resting-state fMRI (rsfMRI)**
  - No explicit task, no stimulation
  - Eyes open or close, rarely covered
  - Instruction:
    - *“Think of nothing in particular and try not to sleep!”*
    - *“Watch movie and Think of nothing in particular!”* (Naturalistic fMRI)
    - *“Try to sleep!”* (EEG-fMRI studies of Epilepsy)
  - Duration 5-15 minutes

The Brain at Rest

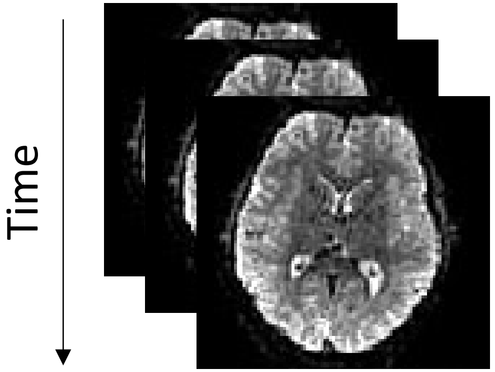
- The brain is always active
- Even when we are not doing anything actively
  - The brain regions are communicating with each other
- This is called as the "intrinsic connectivity"
- High utility in clinical settings where task engagement is tricky



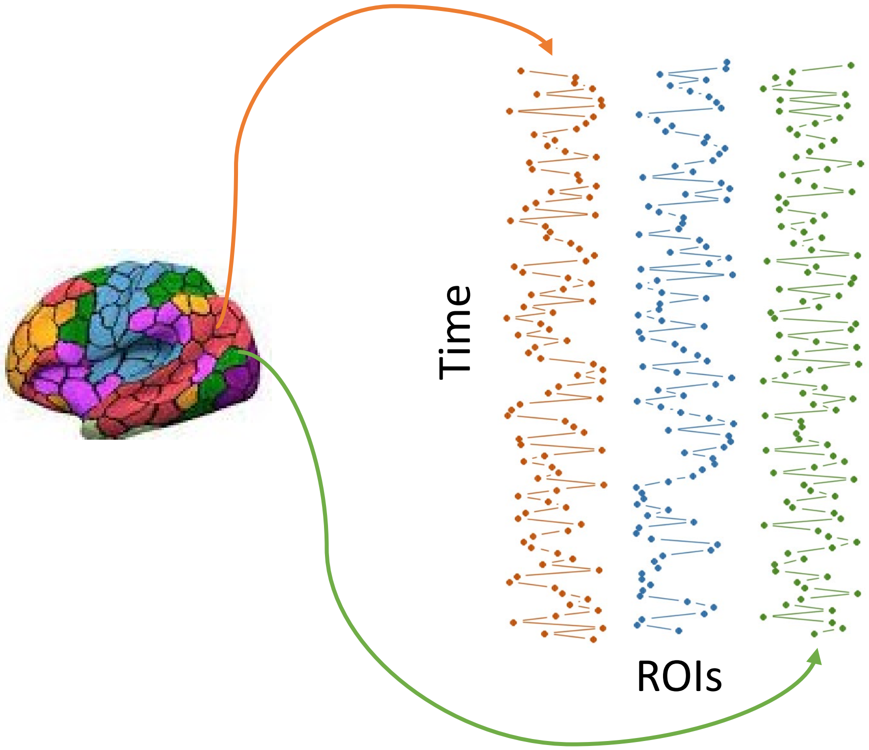
http://marbilab.eu/publications-menu-en/papers-menu-en/tommasin2018

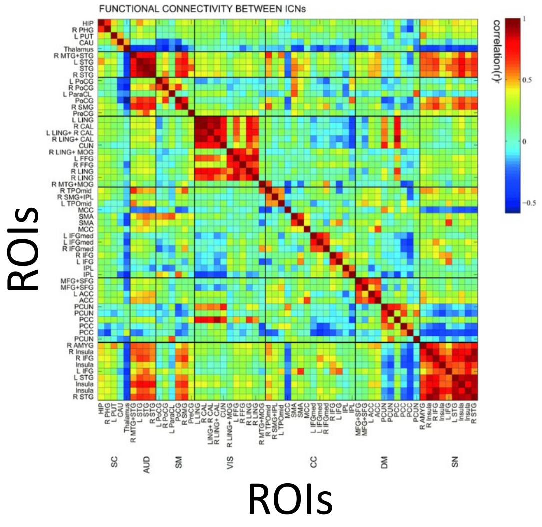
fMRI to connectome

fMRI volumes (preprocessing) → **Parcellation** → **Time series** → **Connectome (Pearson correlation)**



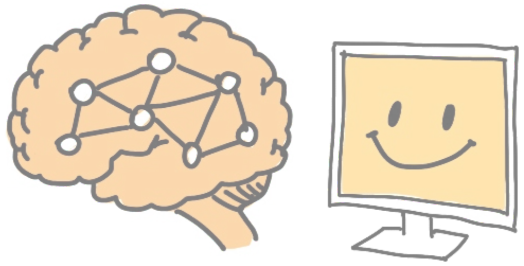
MR imaging with BOLD-sensitive measures
- EPI-sequences (usually)
- Repetition time (TR)

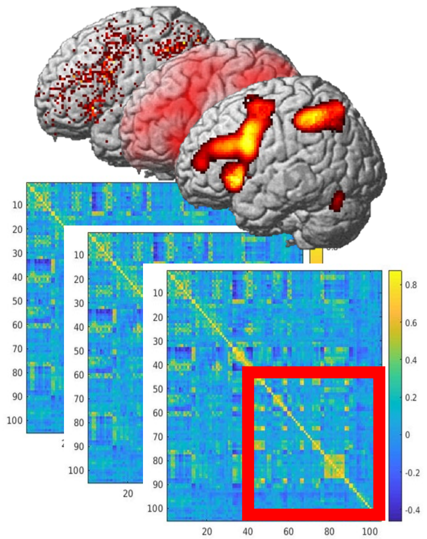




- Connectomes are high dimensional
- 200 nodes ~ 20k connections
- **Curse of Dimensionality** → ML suffers

Prediction of schizophrenia symptoms: Biologically meaningful priors



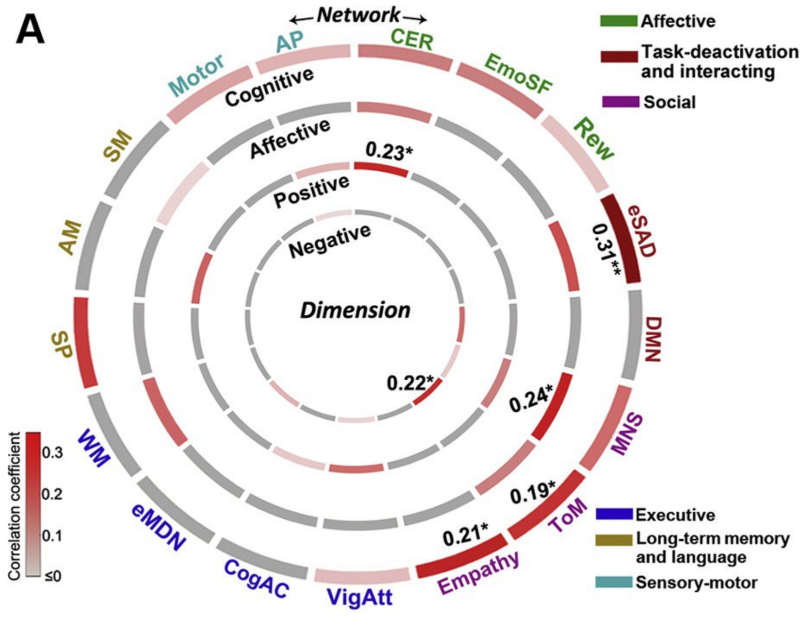


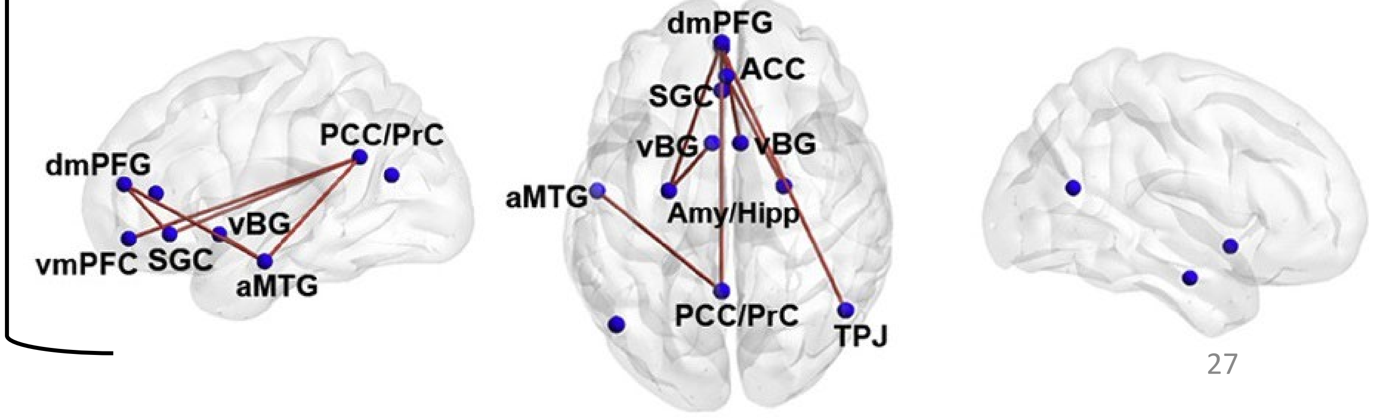
Connectomes

+

- Reuse brain mapping knowledge
- Lower dimensionality
- Better interpretation

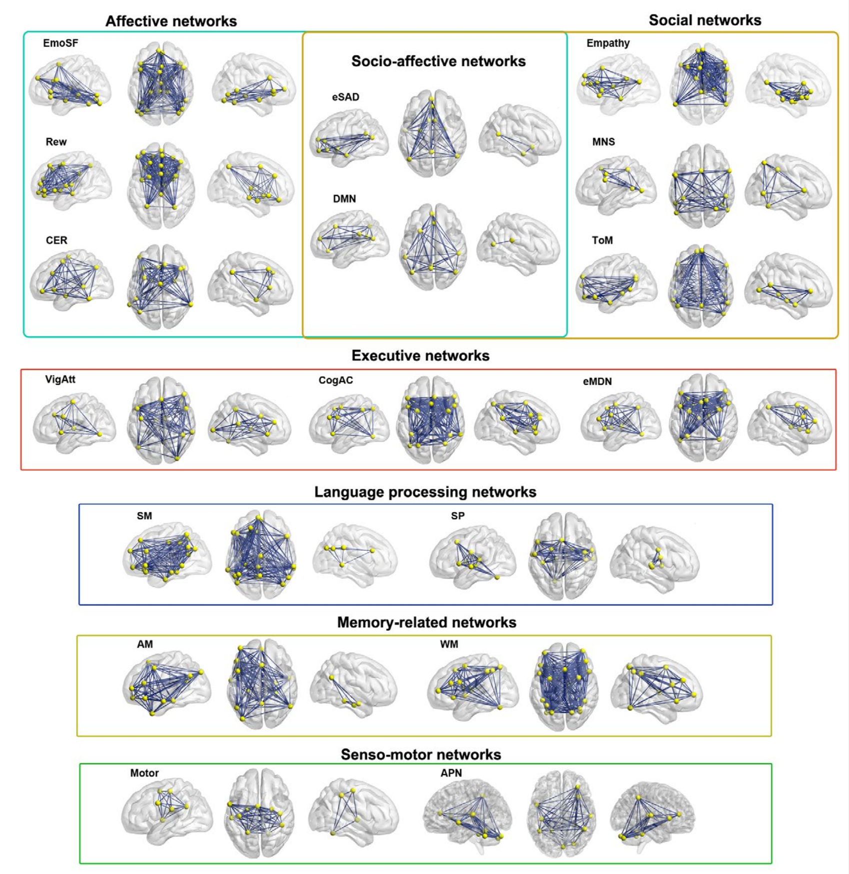
Cognitive dimension predicted by *social and affective* network





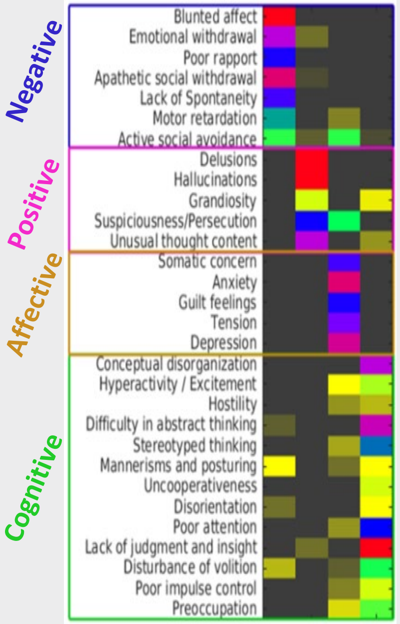
Schizophrenia: Dimensional Psychopathology

Prediction of Symptom scales using Meta-analytic Networks

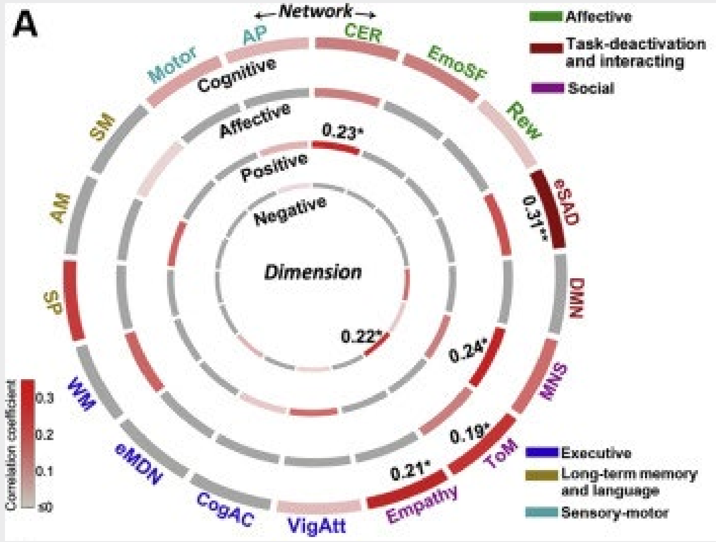


Predict

Chen et al., Biol. Psych. 2020



Significance



Connectivity within meta-analytic networks as biological priors

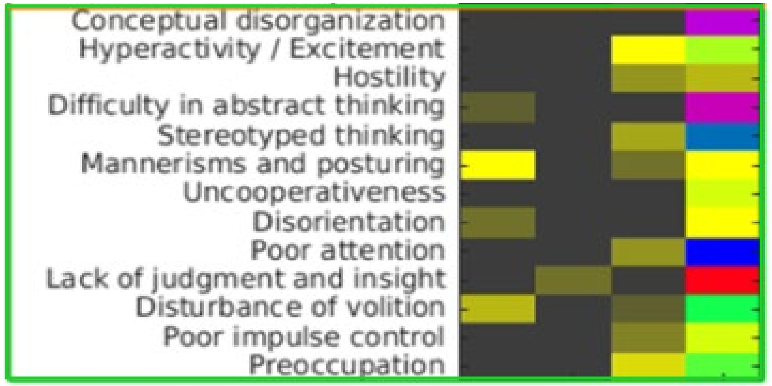
Psychopathology dimensions derived using OPNMF factorization of PANSS score (Chen et.al, Biol Psych 2020)

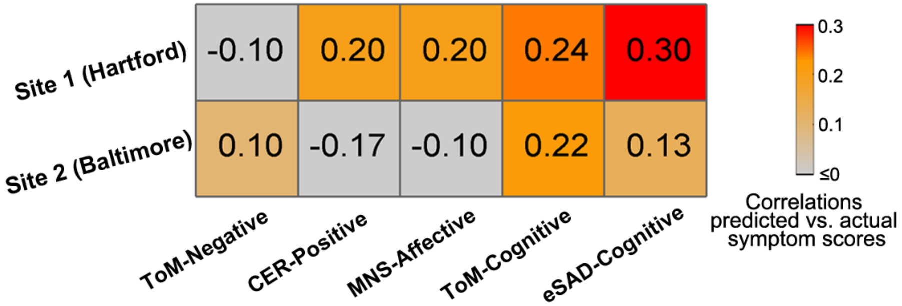
Predictive networks

Chen et al., Biol. Psych. 2021

Schizophrenia: from networks to receptors

The extended social-affective default mode network (eSAD) predicts cognitive symptoms



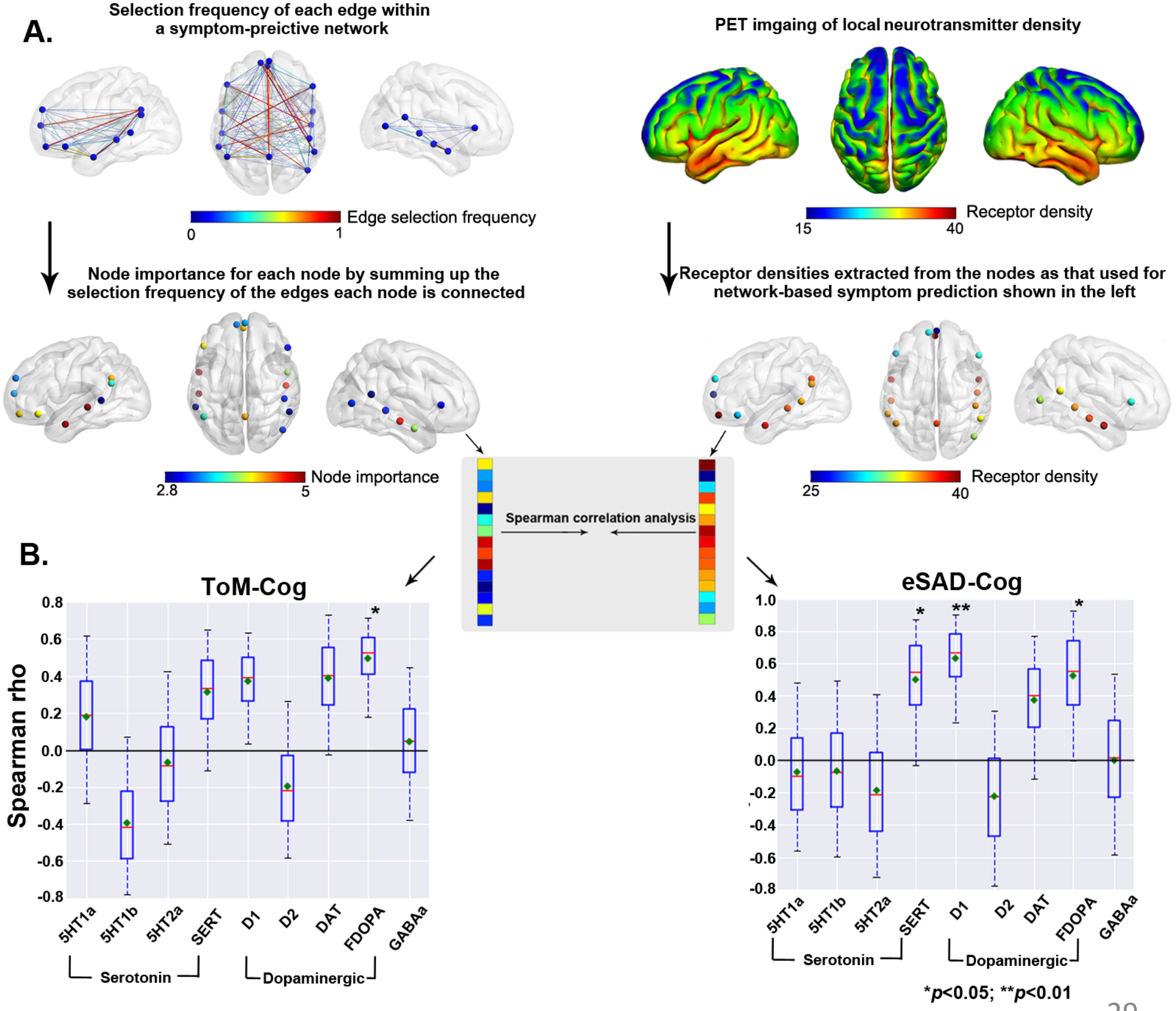


	ToM-Negative	CER-Positive	MNS-Affective	ToM-Cognitive	eSAD-Cognitive
Site 1 (Hartford)	-0.10	0.20	0.20	0.24	0.30
Site 2 (Baltimore)	0.10	-0.17	-0.10	0.22	0.13

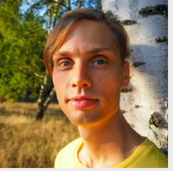
Correlations predicted vs. actual symptom scores

Replication in two cohorts
→ The result is robust

Network node importance correlates with receptor densities (FDOPA, SERT)



Chen et al., Biol. Psych. 2021



Leonard Sasse

Prediction of behavioral scores

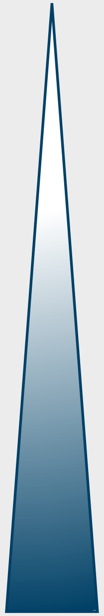
Do fewer timepoints provide similar or better information regarding behavior?

Region-wise time series

Edgewise timeseries and RSS

Esfahlani et al., 2020 PNAS

Ordering of timepoints



Low amplitude co-fluctuation

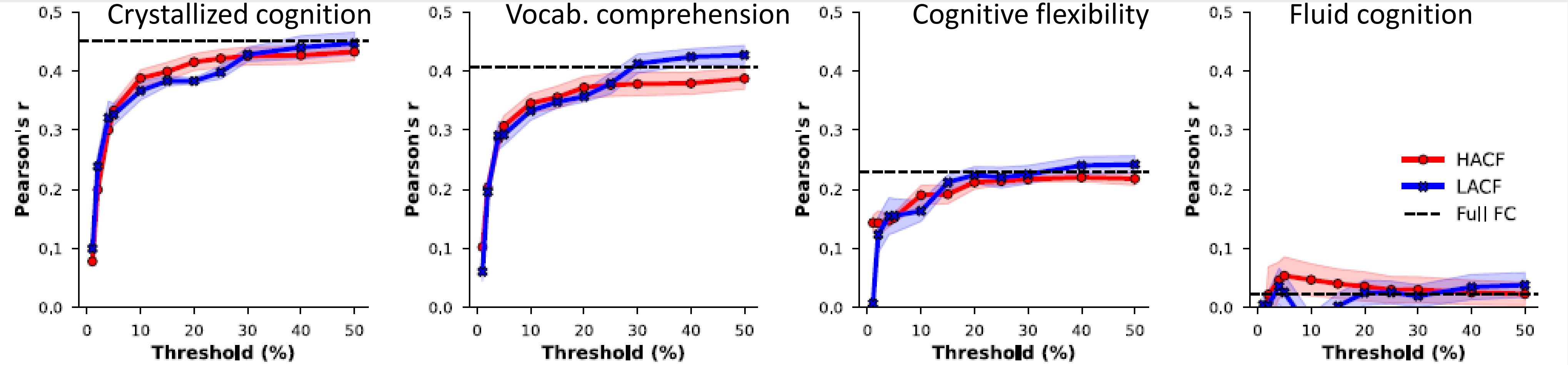
High amplitude co-fluctuation

- HACF timepoints contain idiosyncratic information
- Do they also contribute towards behavior?

Sasse et al., Comms. Biol. 2023

Prediction of behavioral scores

Sequentially adding HACF or LACF timepoints (Human Connectome Project-Ageing)

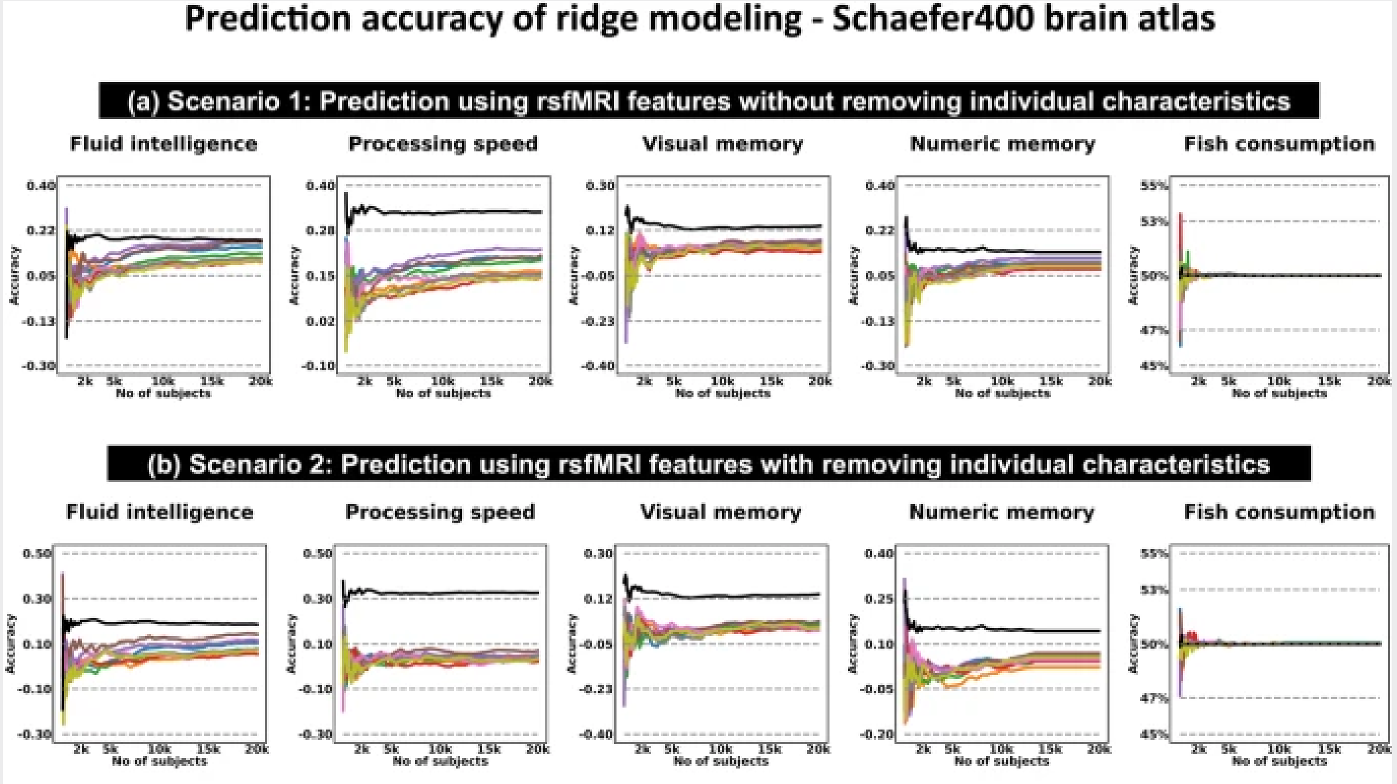


Sasse et al., biorXiv 2022, Under revision

- HACF and LACF do not seem to provide different information
- Intermediate bins contain more information, counter to the original hypothesis
- Different scores show different predictability

How useful is RS?

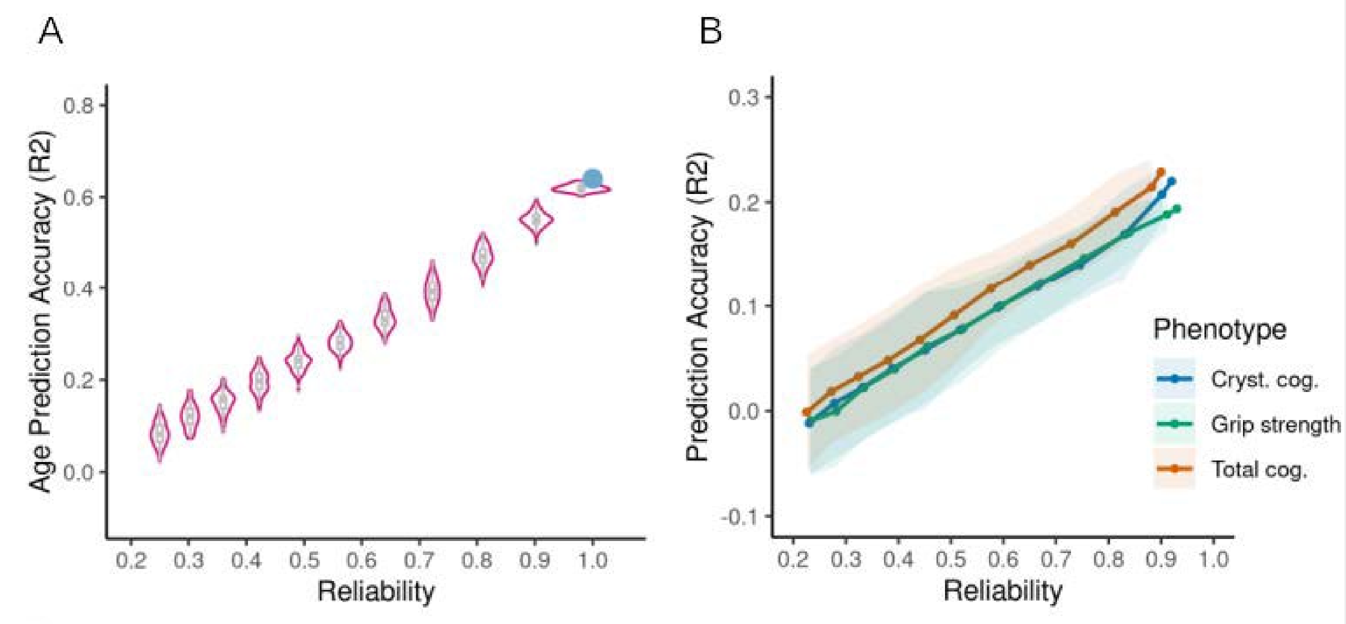
- Various RS properties
  - Local and global connectivity
  - Entropy measures
- Individual characteristics
  - age, gender, and total intracranial volume
- Characteristics > RS



Omidvarnia et al., Comms. Biol. 2024

Impact of target reliability

- Target reliability impacts prediction performance
- Lower reliability means worse prediction
- Many results could be because of this



Gell et al., Nat. Comm. accepted

Challenges in ML/AI

Outline

- Biased models
  - TIV bias in male/female classification

- Confound leakage
  - Increased accuracy after confound removal?

- Data harmonization
  - Leakage and site-target dependence

Photo by Dainis Graveris on Unsplash

female & male brains

Are there organizational differences between female & male brains?


Lisa Weirsch

- Clinical prevalence of many diseases differ
- Pharmacological differences
  - e.g., anesthetics
- ML models can uncover organizational differences
- Naturally female-male brains are different in size
  - The body sizes are different
- ML models likely learn this “simpler” signal while ignoring organizational differences

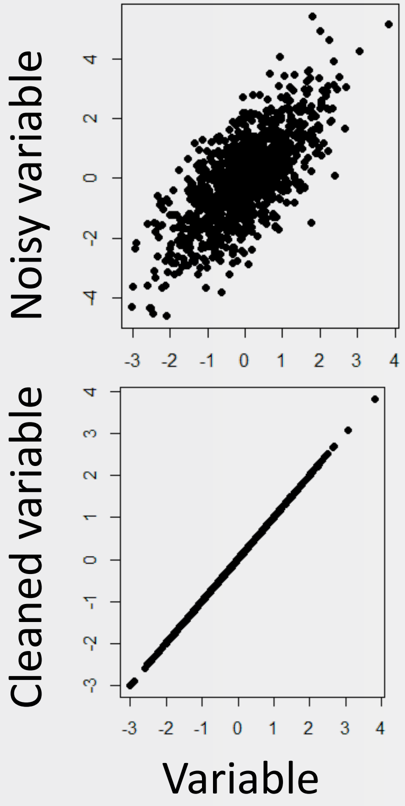
- Measure brain size using MRI: Total Intracranial Volume (TIV)
- Train a ML model using VBM features while ignoring this information
  - Confound removal: from each feature (voxel-wise GMV) linearly regression out TIV signal
  - Matching/stratification: Sample males and females within same TIV range

Weirsch et al., Sci. Reports 2023

Shammi More

Confounding

Simulation with **known noise**



What is it?

- “Nuisance” variables bias the data and in turn the model
  - Older people are more likely to be diagnosed with Parkinson’s
  - Male bodies (and brains) tend to be larger on average
- We want a de-confounded model

How to deal with it?

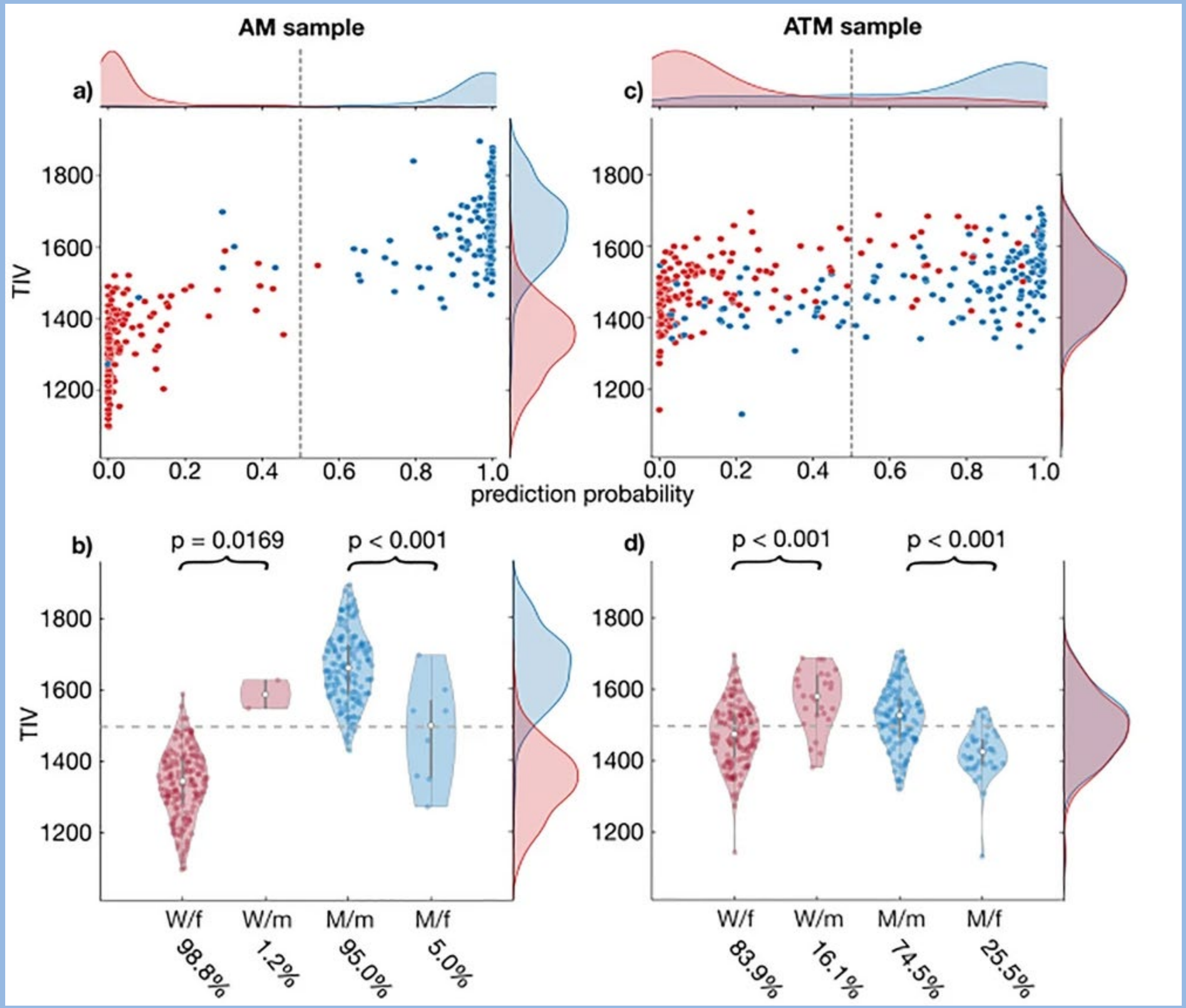
- Featurewise confound removal in a CV-consistent manner
  - Avoid data leakage
  - More et al., 2021 ECML

1. $\mathbf{f}\colon\ \mathbf{x} \sim \mathbf{conf}$
2. $\hat{\mathbf{x}} = \mathbf{f}(\mathbf{conf})$
3. $\mathbf{x}_{\mathbf{CR}} = \mathbf{x} - \hat{\mathbf{x}}$

There are other ways, e.g. stratifying w.r.t. confounding variance, with their own pros and cons.

More et al., ECML-PKDD 2021

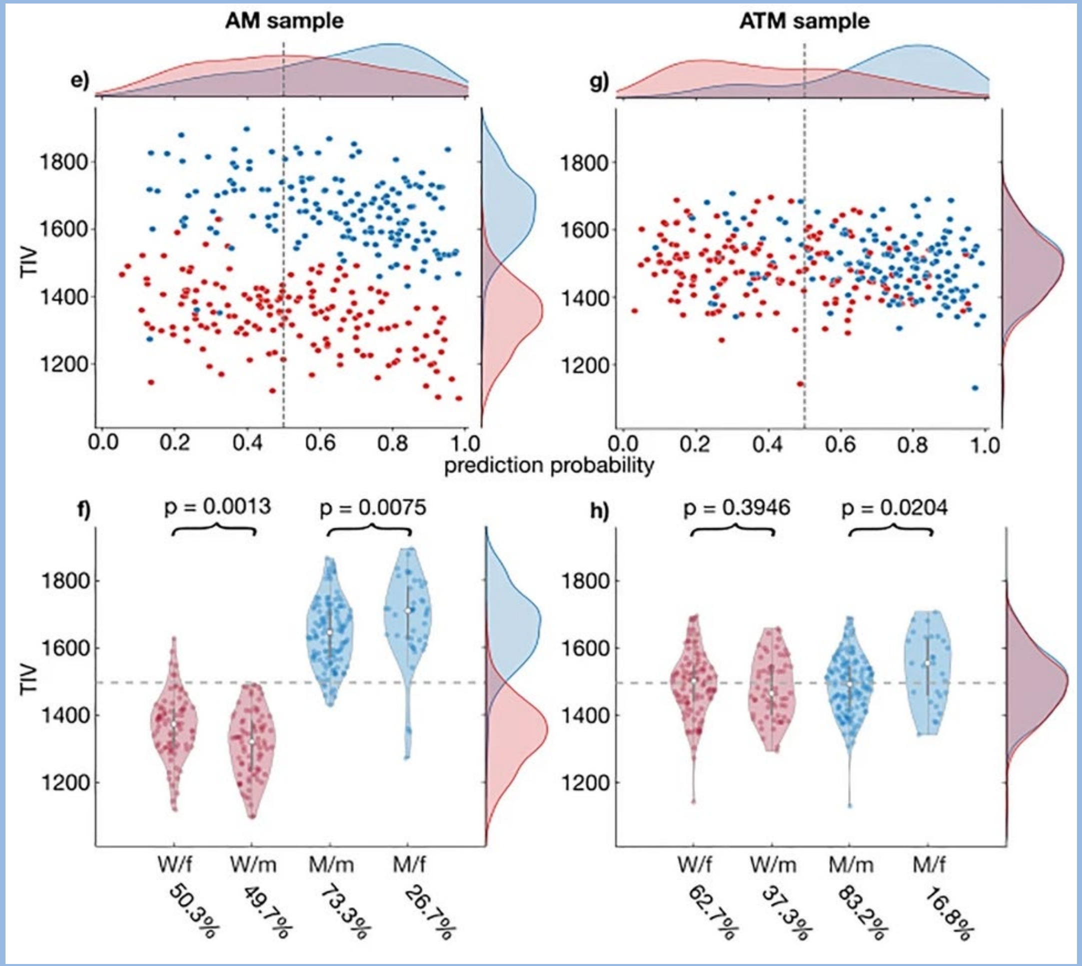
Is there a TIV bias?



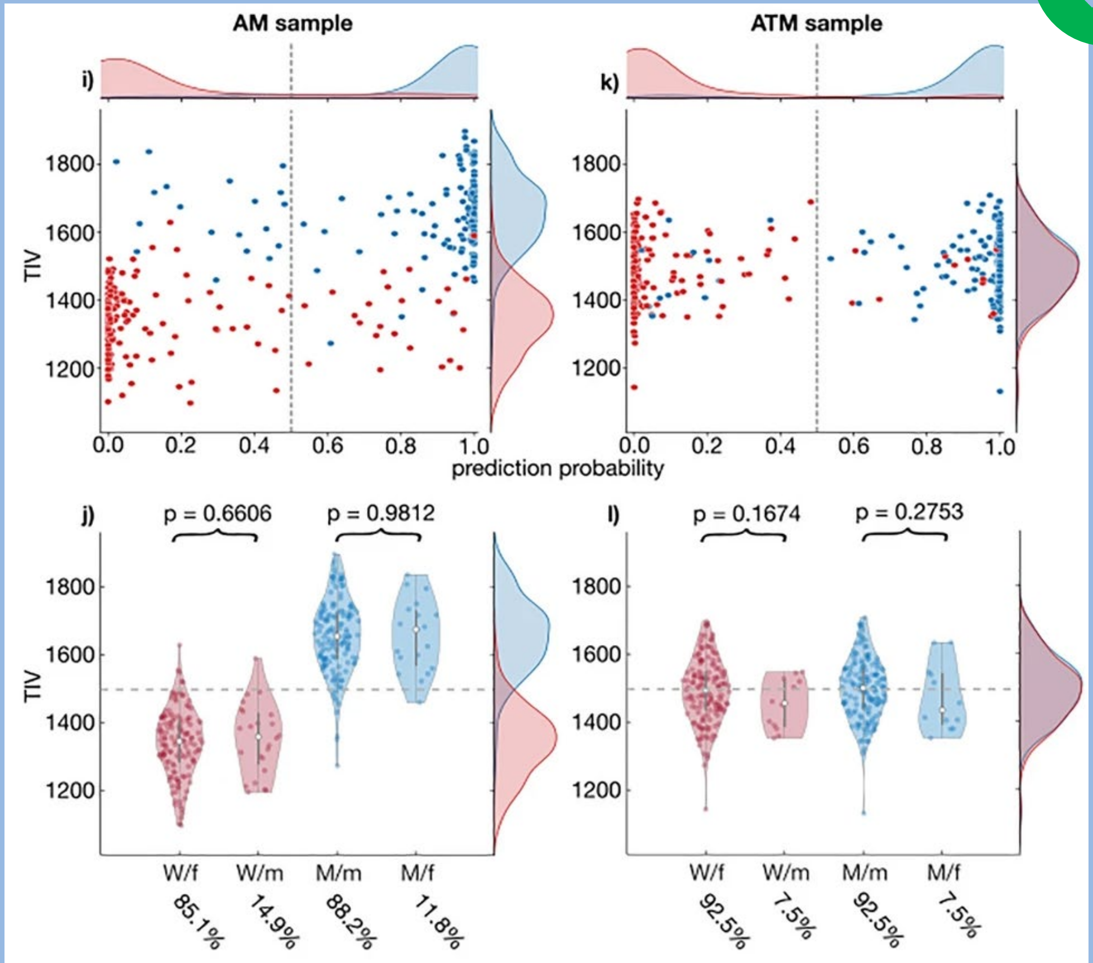
Errors

Different ways to a bias-free models

Confound removal



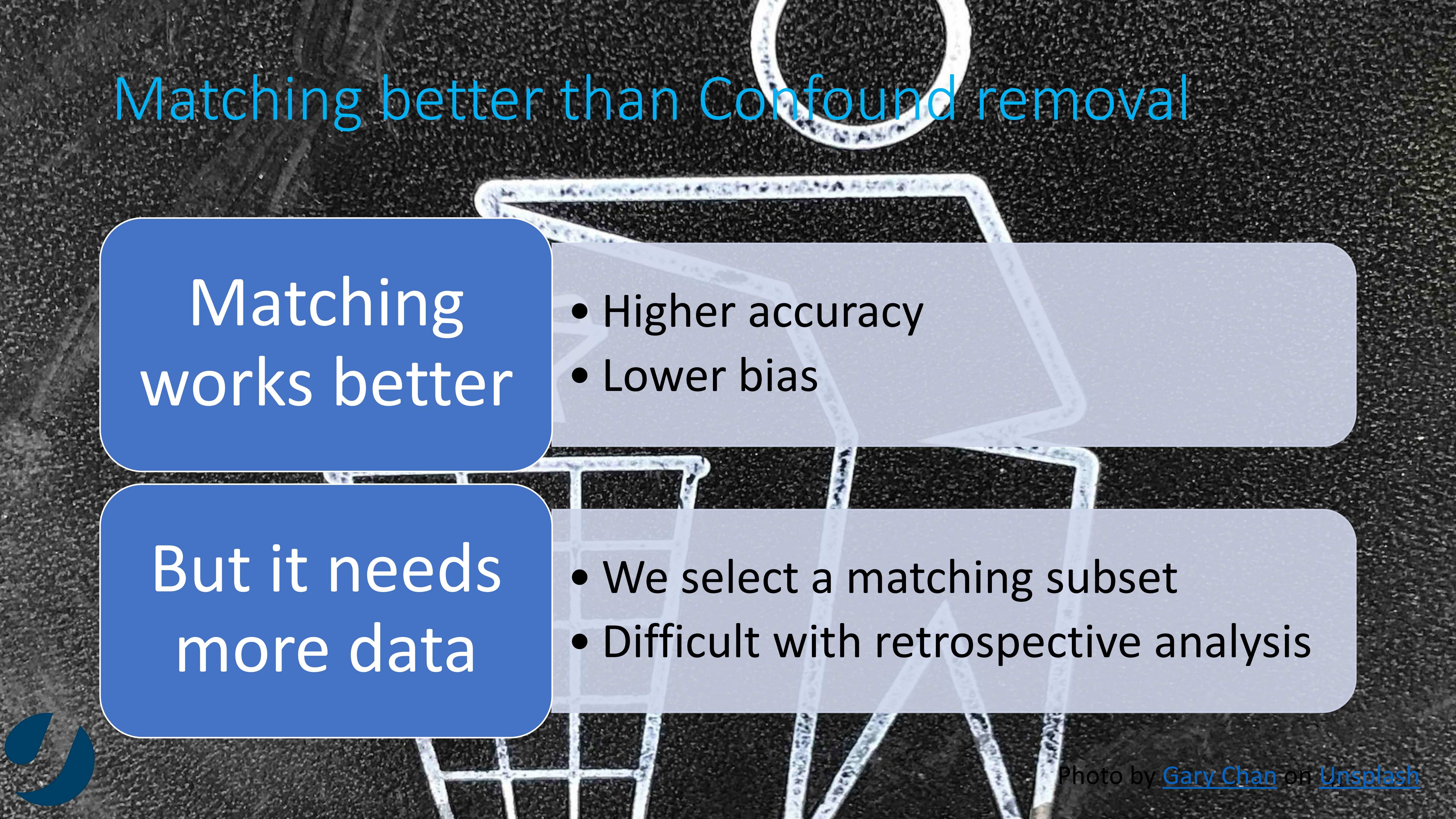
Matching



⬆ ⬆ **Errors** ⬆ ⬆

⬆ ⬆ **Errors** ⬆ ⬆

Matching better than Confound removal



Matching works better

- Higher accuracy
- Lower bias

But it needs more data

- We select a matching subset
- Difficult with retrospective analysis

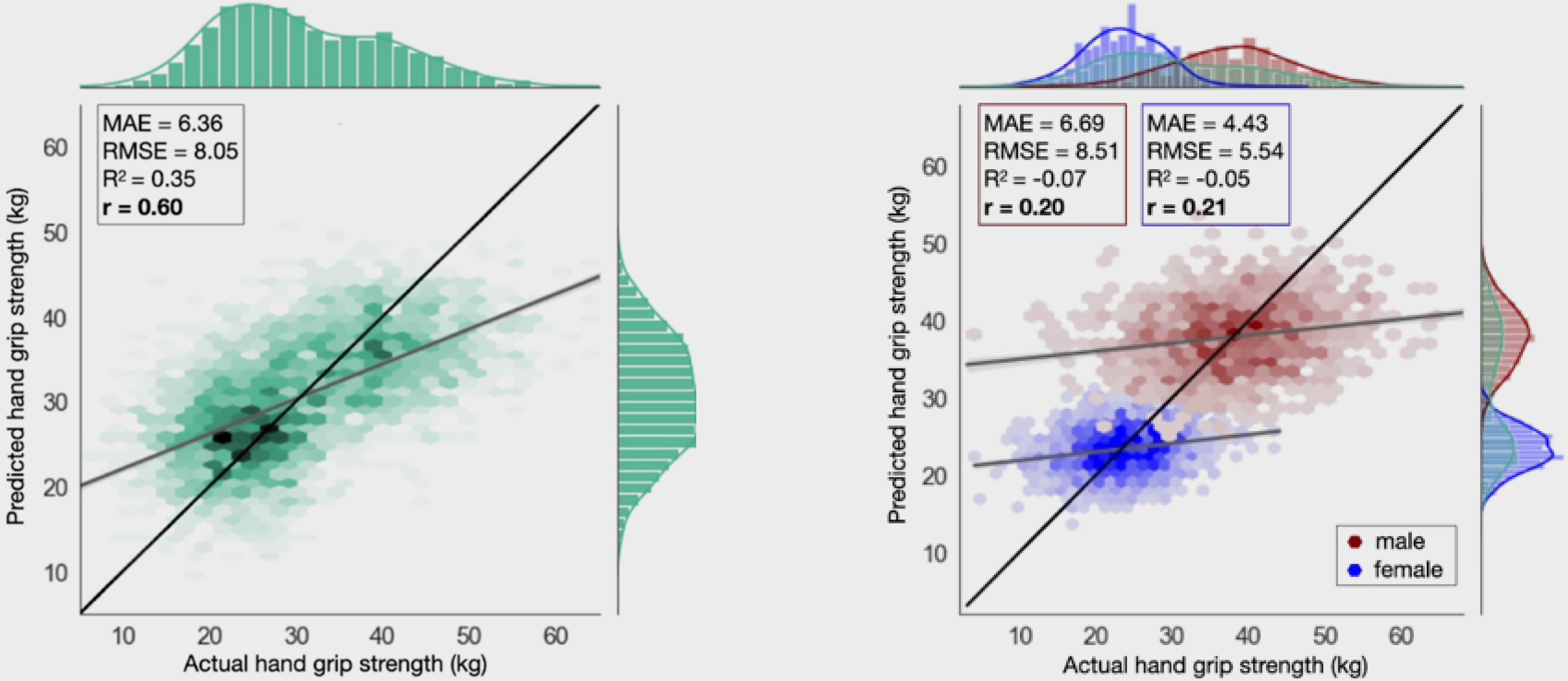
Photo by Gary Chan on Unsplash

Photo by Sangga Rima Roman Selia on Unsplash

Confound removal and leakage

Vera Komeyer

Confounding: Predicting hand-grip-strength (HGS) using brain structure (VBM)

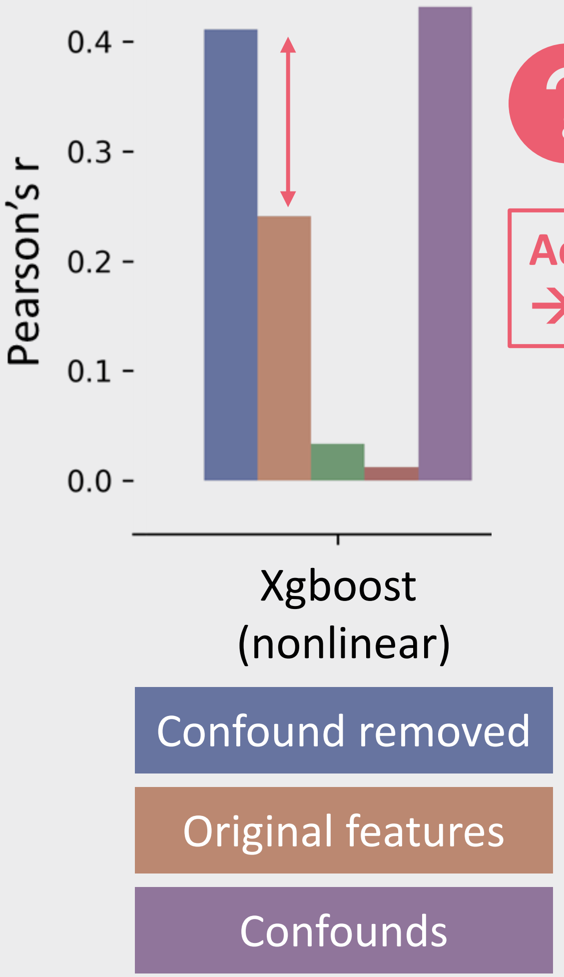


Predictions are driven by the sex of the subjects

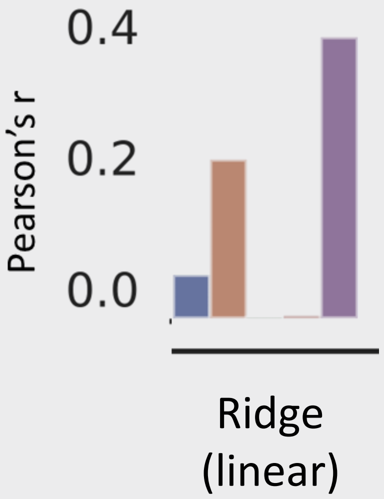
Accuracy decrease after confound removal
→ Signal was confounded

What if accuracy increases after CR?

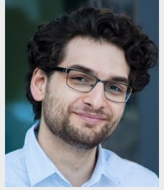
- 400x400 DWI connectomes (UK Biobank)
- Masked using white matter hyperintensity lesion maps
- Calculate the “disconnectivity” matrices, perform UMAP (3D)
- Age and sex as confound
- Predict: Cognition symbol digit substitution correct matches



Accuracy increase after confound removal
→ Where is the signal coming from?



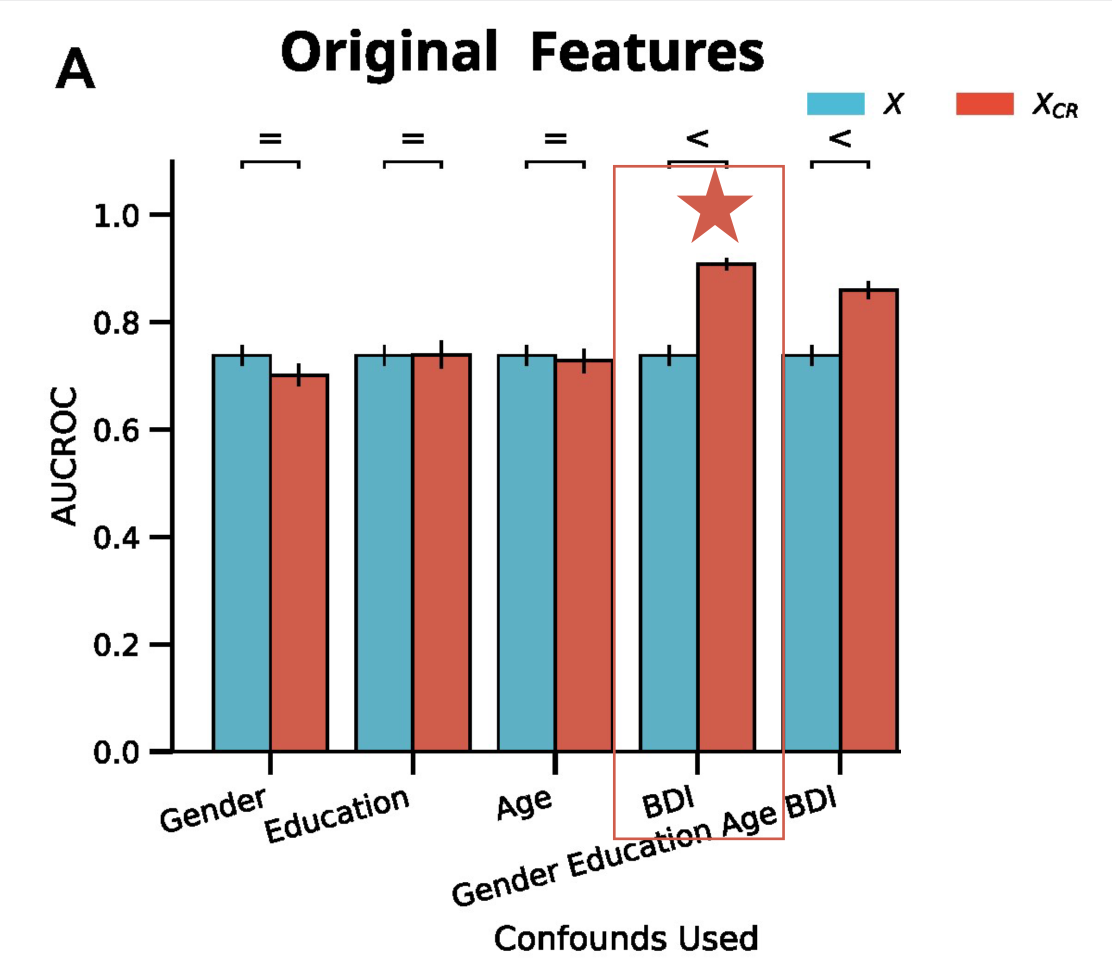
We do not see the same pattern with a linear model



Sami
Hamdan

Confound leakage: ADHD prediction

- Voice-derived features
  - Can aid in objective diagnosis
- Depression is a comorbidity
  - We want the model to learn “ADHD” and not depression.
  - Measured using BDI
- Featurewise confound removal
- BDI removal gives high accuracy: we solved an important clinical problem?
  - AUC ~ 0.9 → diagnostic tool!
  - Wait! Is this real?

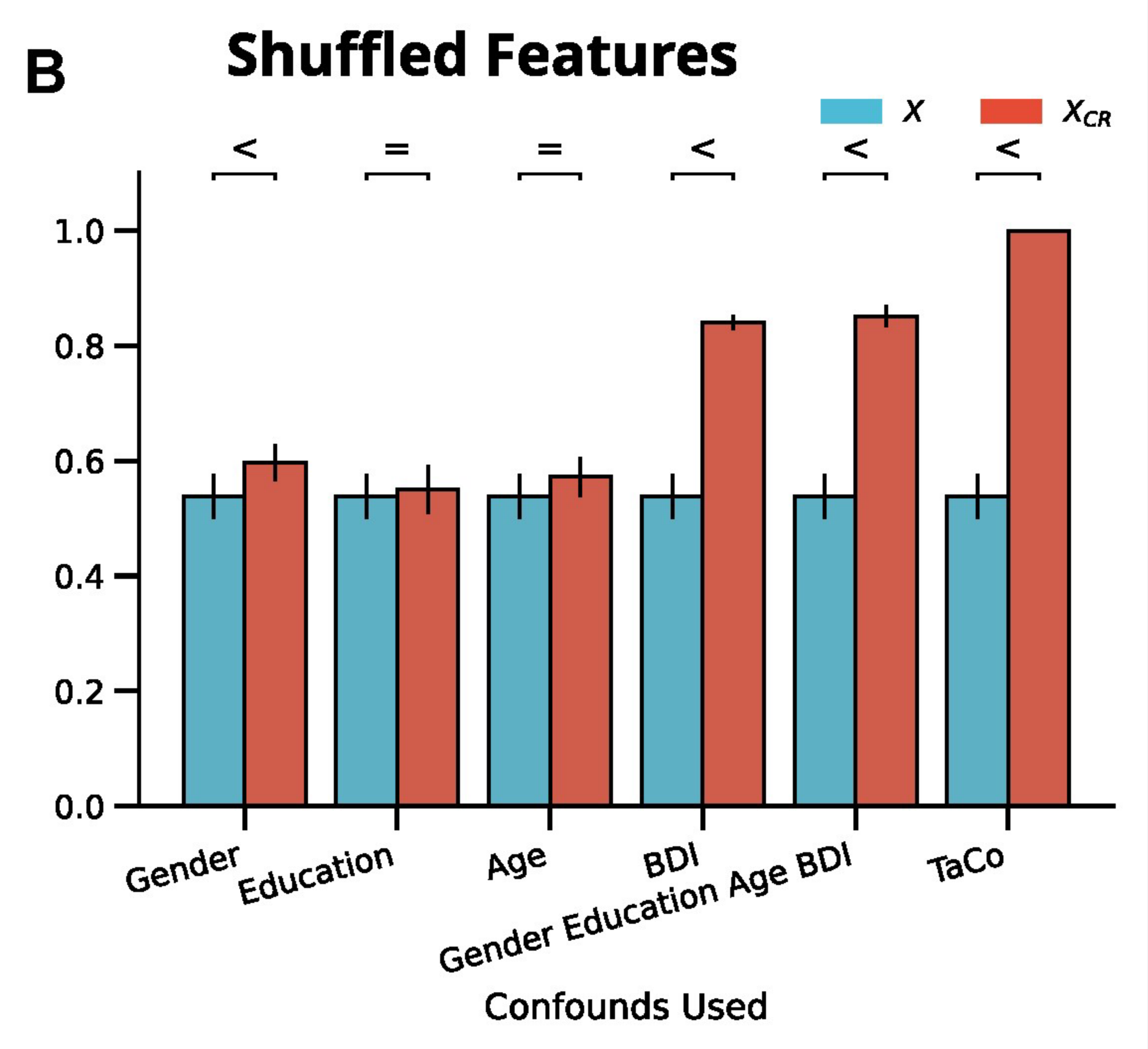


The comparison signs are from the Bayesian ROPE

Hamdan et al., GigaScience 2023

Confound leakage: ADHD prediction

- Is it really leakage?
- Let's shuffle the features
  - Destroys features-confounds relationship
  - Keeps confounds-target relationship
- High AUC with BDI and TaCo (Target as Confound) indicative of leakage

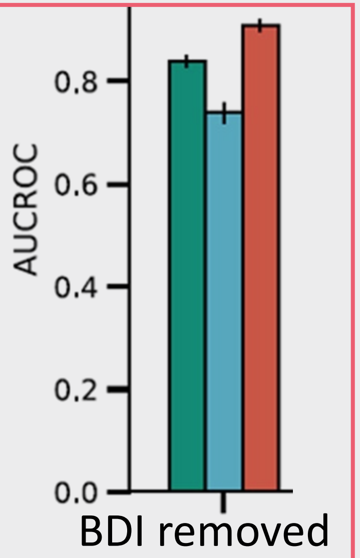


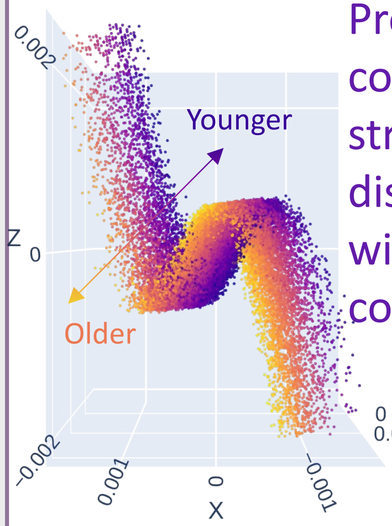
Confound leakage: Summary

- The confounding variance can leak into the features
- New features (residuals) are not confound-free
- Nonlinear models (RF, MLP) are more likely to pick up the leaked signal
- Misleading models and predictions
- Report results with and without confound removal

Hamdan et al., 2022 arXiv, Under revision

Prediction of ADHD using voice features with BDI as a confound





Prediction of cognition using structural disconnectome with age as a confound

Prediction of executive functioning using voice features with age as a confound

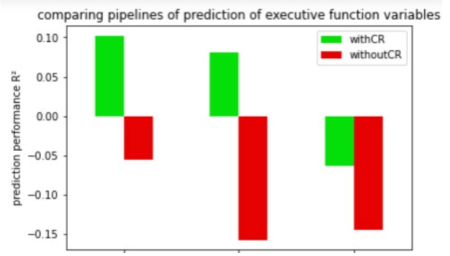


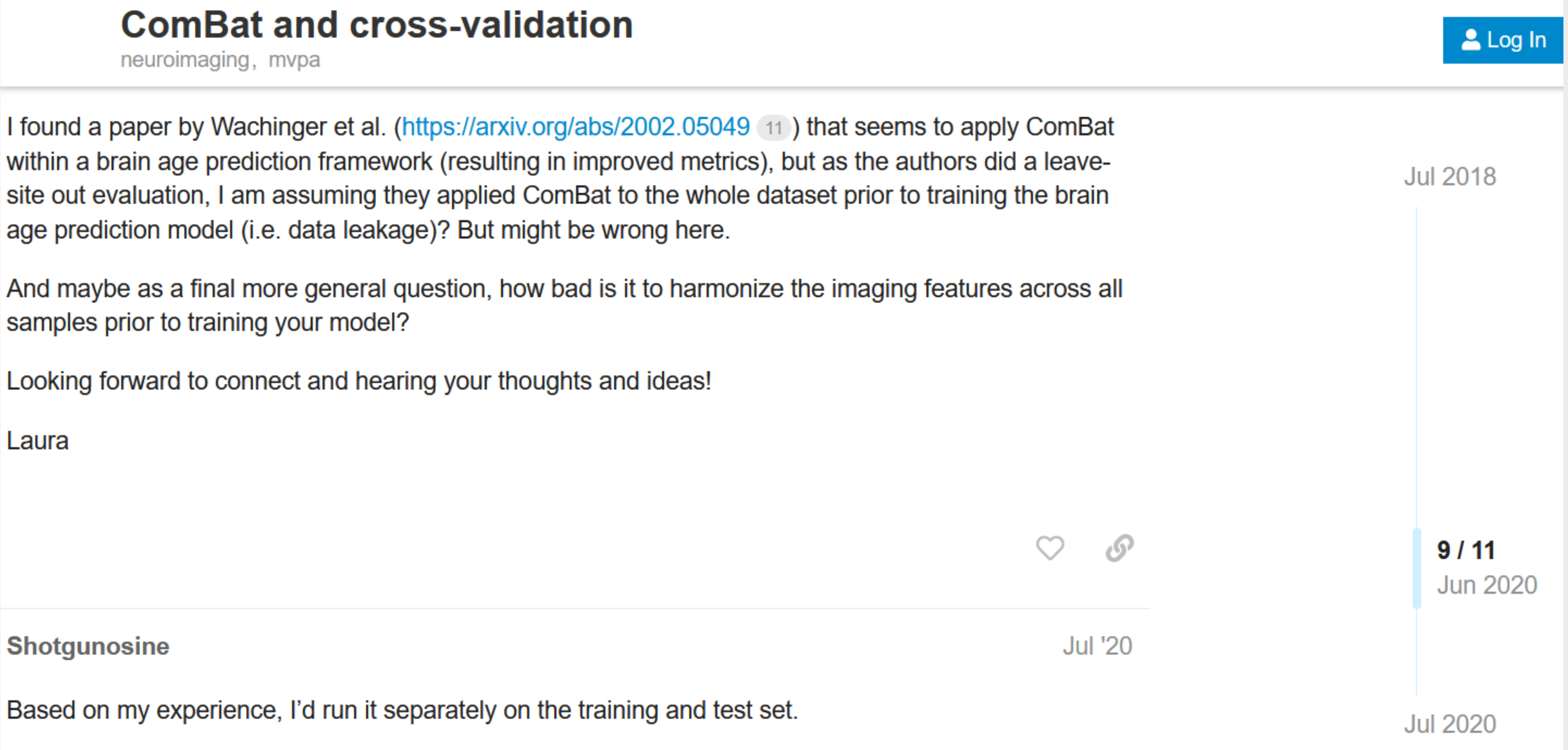
Photo by Vinicius "amnx"" Amano on Unsplash

Data Harmonization

So, what is the problem?

Need to know the biological variability to preserve it.

We need to tell ComBat the labels. **Also, on the test data!**



ComBat and cross-validation

neuroimaging, mvpa

I found a paper by Wachinger et al. (https://arxiv.org/abs/2002.05049 11) that seems to apply ComBat within a brain age prediction framework (resulting in improved metrics), but as the authors did a leave-site out evaluation, I am assuming they applied ComBat to the whole dataset prior to training the brain age prediction model (i.e. data leakage)? But might be wrong here.

And maybe as a final more general question, how bad is it to harmonize the imaging features across all samples prior to training your model?

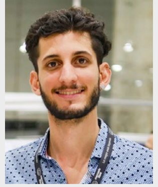
Looking forward to connect and hearing your thoughts and ideas!

Laura

Shotgunosine Jul '20

Based on my experience, I'd run it separately on the training and test set.

https://neurostars.org/t/combat-and-cross-validation/2055/10

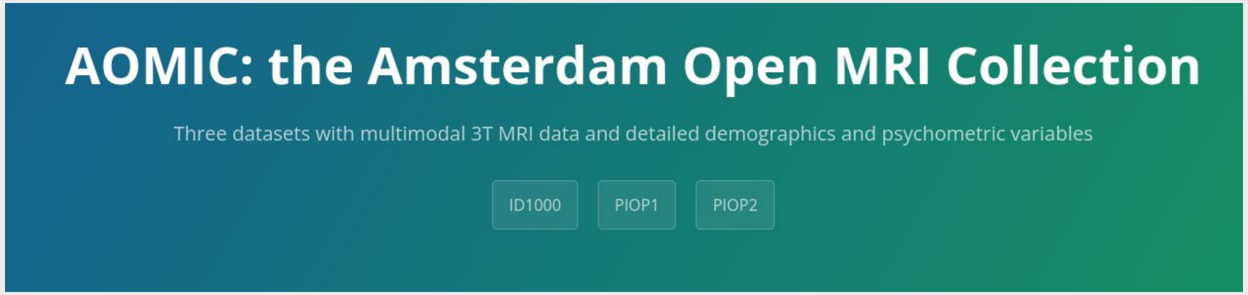


Nicolas
Nieto

Harmonization

- Real-world data is acquired from different sources or sites.
- ML can benefit from large datasets → combining datasets is appealing
- Sites present intrinsic variability
  - observer effect, scanner effect, batch effect

1000BRAINS 1000



AOMIC: the Amsterdam Open MRI Collection

Three datasets with multimodal 3T MRI data and detailed demographics and psychometric variables

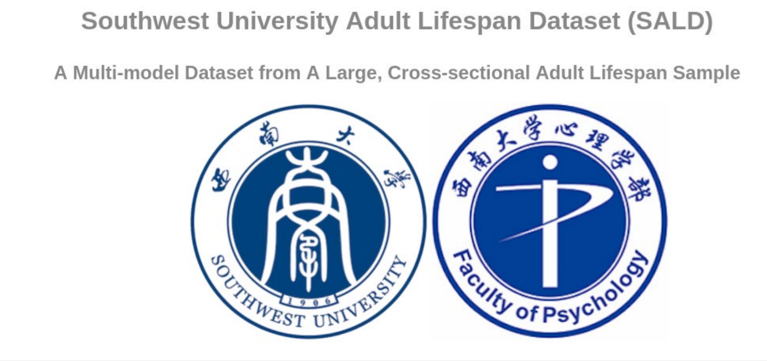
ID1000 PIOP1 PIOP2

CamCAN
Cambridge
Centre for Ageing
and Neuroscience

The Enhanced Nathan Kline Institute-Rockland Sample (NKI-RS)

Multiband Imaging Test-Retest Pilot Dataset

indi
International Neuroimaging
Data-Sharing Initiative



Southwest University Adult Lifespan Dataset (SALD)

A Multi-model Dataset from A Large, Cross-sectional Adult Lifespan Sample

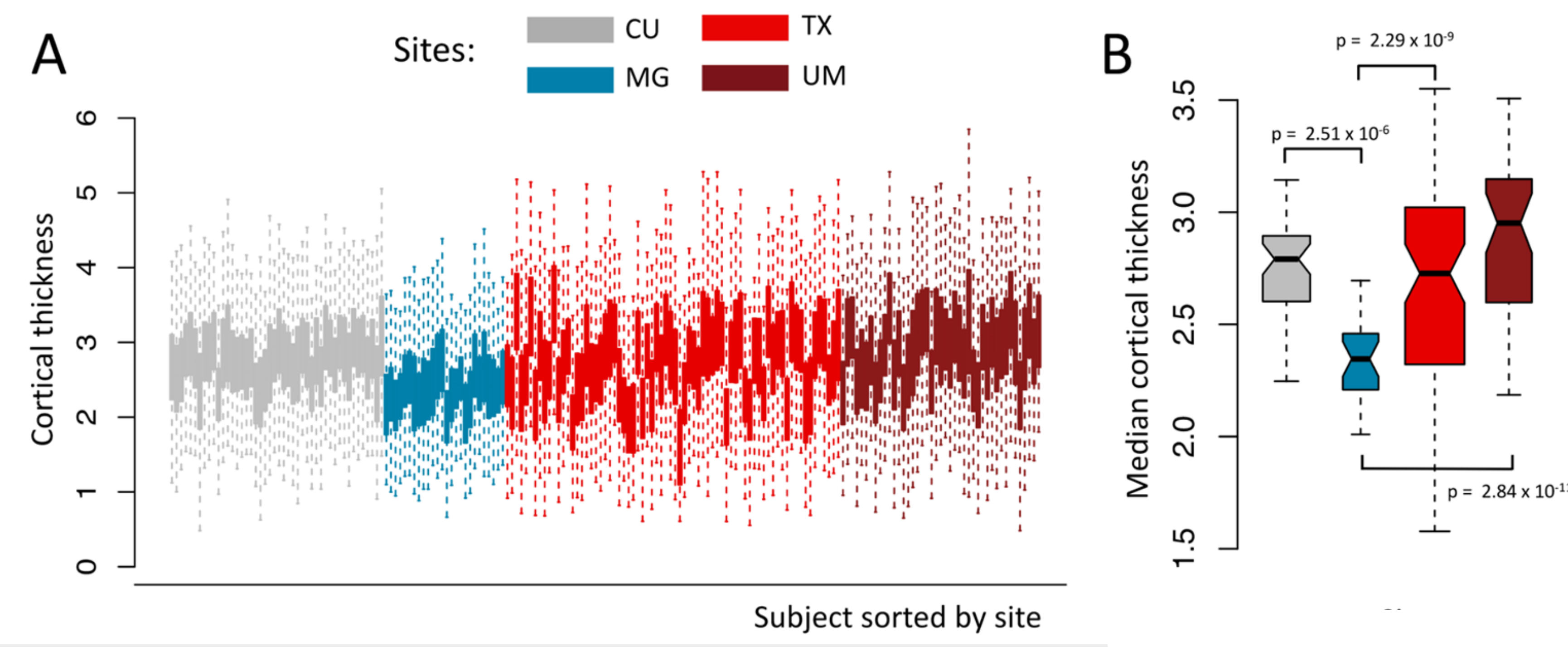
PARK
AGING MIND
LABORATORY
INNOVATION. INTELLECT. INTEGRITY.

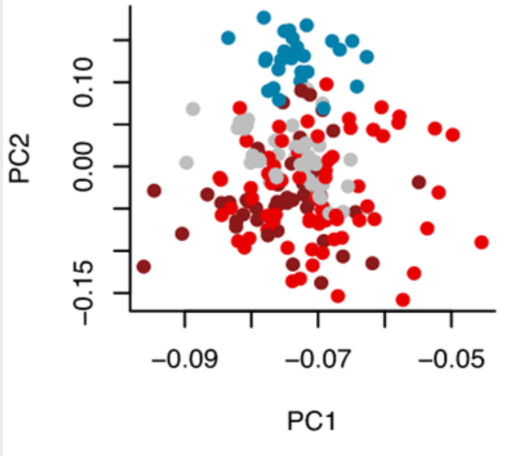
Different scanners → different data

Datasets

- CU = Columbia University
- TX = University of Texas Southwestern
- MG = Massachusetts General Hospital
- UM = University of Michigan

They further demonstrate that this differences are not due another covariate (age, gender, etc)





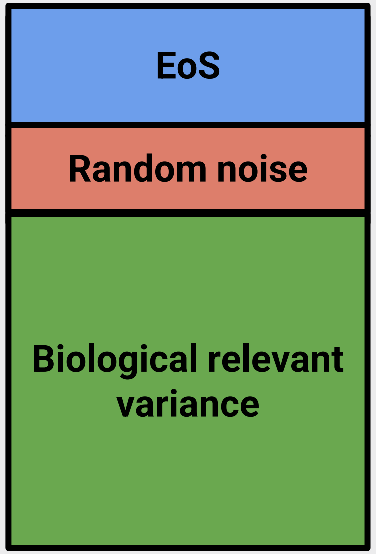
Figures from: Fortin et al. 2018 NeuroImage
Harmonization of cortical thickness measurements across scanners and sites

Remove Effect of Site (EoS)

ComBat

- Estimate a feature-wise **location** and **scale** correction for each site.
  - Empirical Bayes

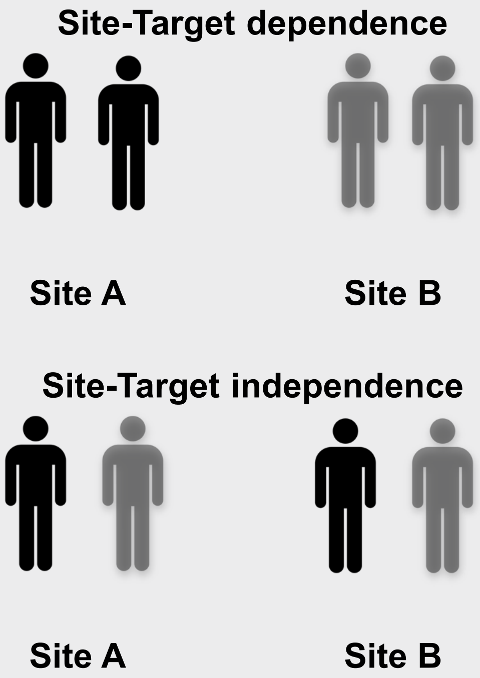
- ComBat cannot differentiate between biologically relevant variance and site-effect when the *site and target are dependant*.
  - Class proportion differs across sites
  - In the extreme cases all control and all patients acquired at different sites



EoS

Random noise

Biological relevant variance



Site-Target dependence

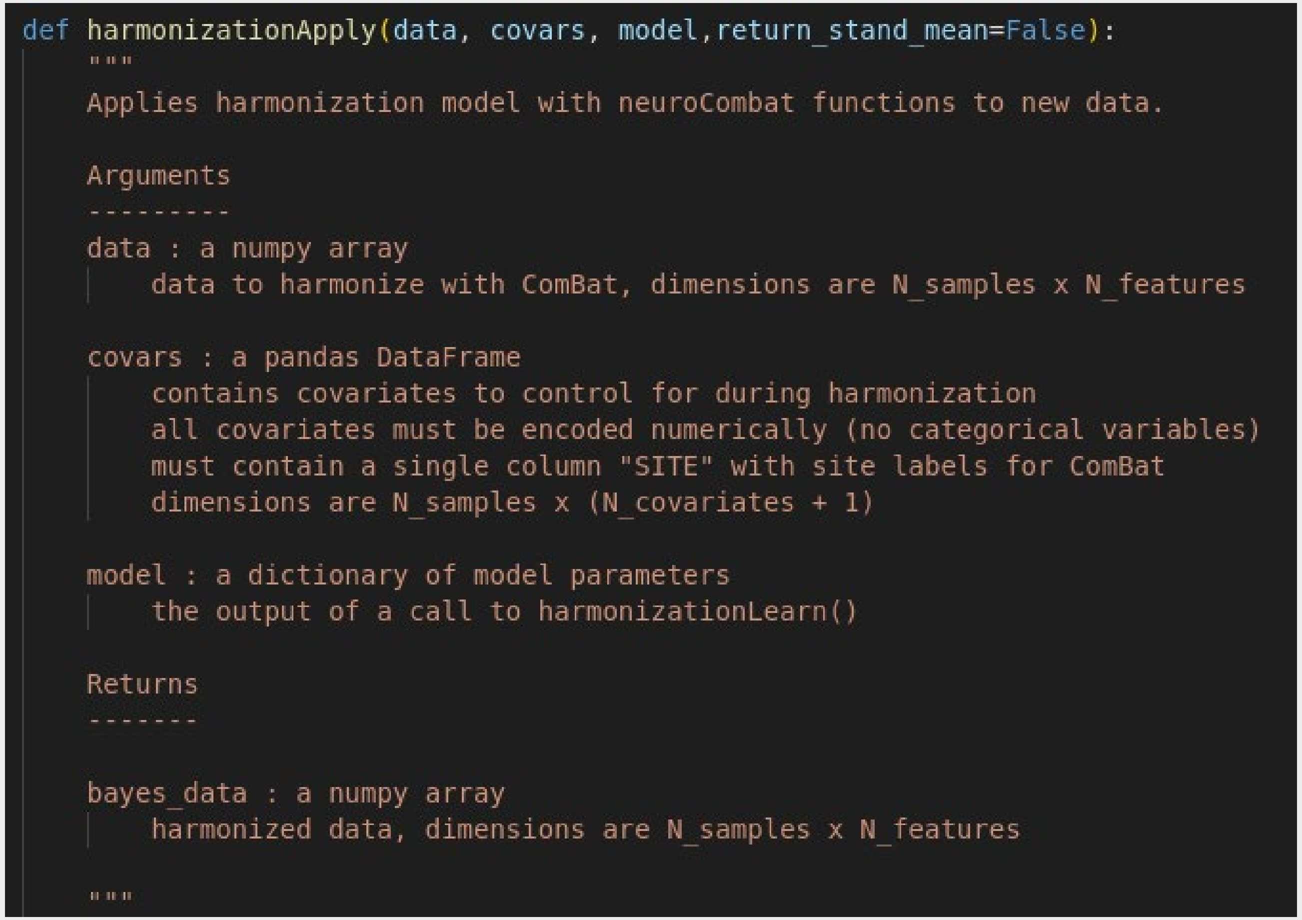
Site A **Site B**

Site-Target independence

Site A **Site B**

Data leakage, how?

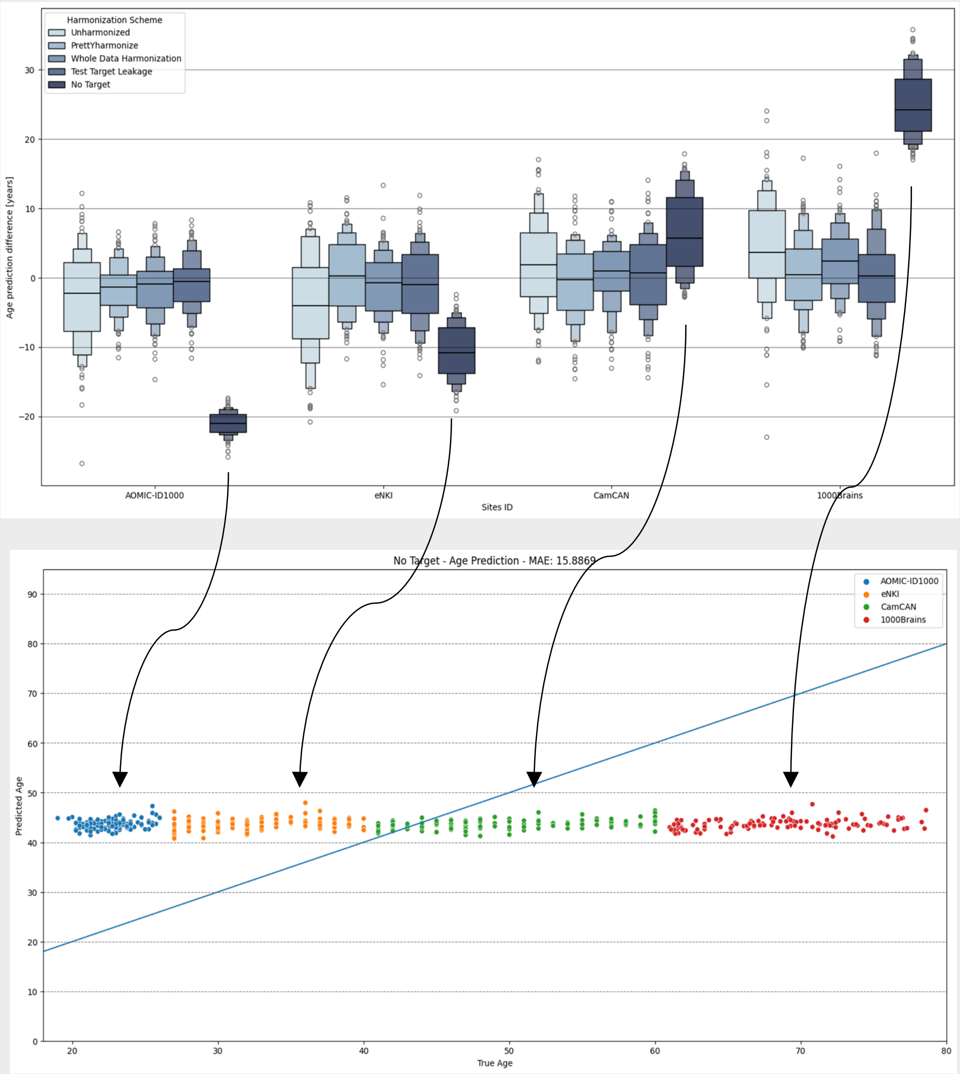
- neuroHarmonize (based on ComBat) needs the **train and test labels** as “covars” to preserve the associated variance
- Prevents real-word applications, as the **test labels are not know!**



```
def harmonizationApply(data, covars, model,return_stand_mean=False):
    """
    Applies harmonization model with neuroCombat functions to new data.

    Arguments
    ---------
    data : a numpy array
        data to harmonize with ComBat, dimensions are N_samples x N_features

    covars : a pandas DataFrame
        contains covariates to control for during harmonization
        all covariates must be encoded numerically (no categorical variables)
        must contain a single column "SITE" with site labels for ComBat
        dimensions are N_samples x (N_covariates + 1)

    model : a dictionary of model parameters
        the output of a call to harmonizationLearn()

    Returns
    -------

    bayes_data : a numpy array
        harmonized data, dimensions are N_samples x N_features

    """
```

Empirical evaluation

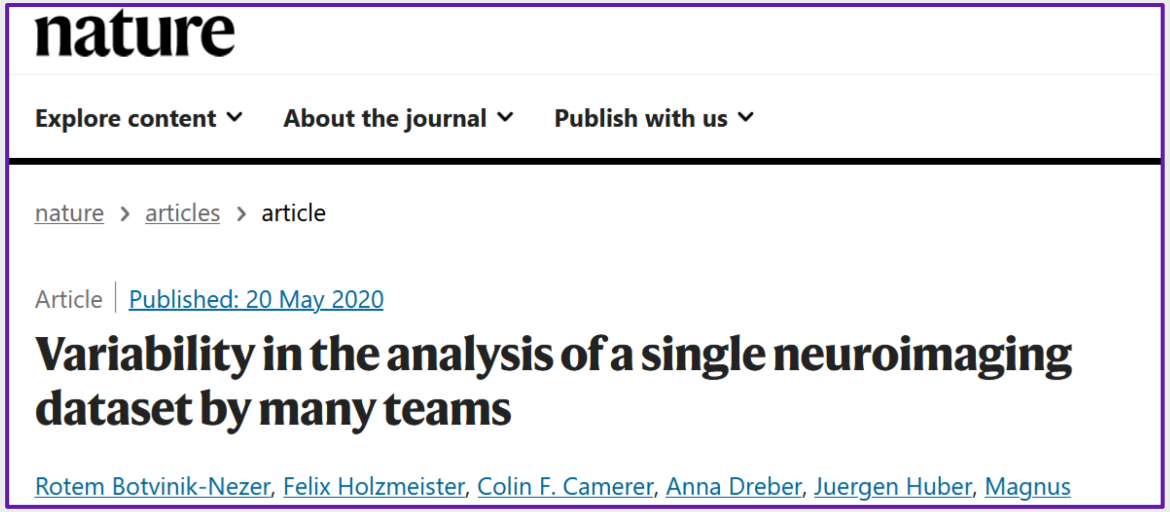
- “No Target” removed the biological signal → Worst performance
- WDH and TTL better than Unharmonized → Leakage
- *PrettYharmonize* was the same or slightly better without data leakage
- **None** of the harmonization methods showed an improvement when site-target were independent.
    - MRI: age, sex, dementia
    - ICU: mortality



Nieto et al., arXiv 2024

Further conceptual/data challenges

1. Low prediction accuracy often driven by demographics
2. Reliability issues
3. Data biases, e.g. ethnicity
4. Replicability and analysis freedom



nature

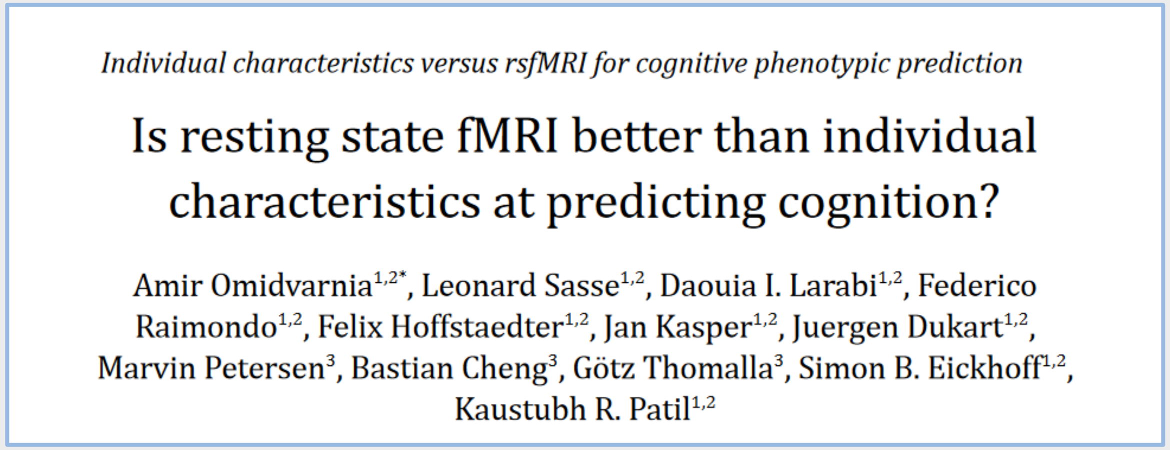
Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Published: 20 May 2020

Variability in the analysis of a single neuroimaging dataset by many teams

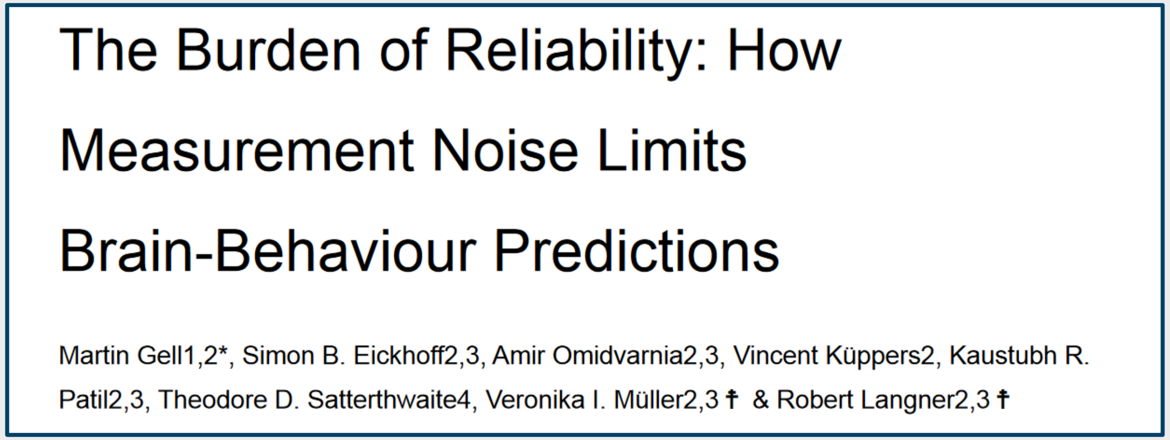
Rotem Botvinik-Nezer, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus



Individual characteristics versus rsfMRI for cognitive phenotypic prediction

Is resting state fMRI better than individual characteristics at predicting cognition?

Amir Omidvarnia[1,2*], Leonard Sasse[1,2], Daouia I. Larabi[1,2], Federico Raimondo[1,2], Felix Hoffstaedter[1,2], Jan Kasper[1,2], Juergen Dukart[1,2], Marvin Petersen[3], Bastian Cheng[3], Götz Thomalla[3], Simon B. Eickhoff[1,2], Kaustubh R. Patil[1,2]



The Burden of Reliability: How Measurement Noise Limits Brain-Behaviour Predictions

Martin Gell1,2*, Simon B. Eickhoff2,3, Amir Omidvarnia2,3, Vincent Küppers2, Kaustubh R. Patil2,3, Theodore D. Satterthwaite4, Veronika I. Müller2,3 ✝ & Robert Langner2,3 ✝

HOME › SCIENCE ADVANCES › VOL. 8, NO. 11 › CROSS-ETHNICITY/RACE GENERALIZATION FAILURE OF BEHAVIORAL PREDICTION FROM RESTING-STATE FUNCTIONAL CONNEC-...

RESEARCH ARTICLE | NEUROSCIENCE

Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity

JINGWEI LI, DANILO BZDOK, JIANZHONG CHEN, ANGELA TAM, LEON QI RONG OOI, AVRAM J. HOLMES, TIAN GE, KAUSTUBH R. PATIL, MBEMBA JABBI, SIMON B. EICKHOFF, B. T. THOMAS YEO, AND SARAH GENON

ML mistakes are Expensive

Article | Published: 30 October 2017

RETRACTED ARTICLE: Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth

Nature Human Behaviour **1**, 911–919 (2017) | Cite this article

8595 Accesses | **159** Citations | **1353** Altmetric | Metrics

This article was retracted on 06 April 2023

The authors are retracting this article after concerns were raised about the **validity of their machine learning method** in a Matters Arising[1]. While revising their response to these concerns, the authors confirmed that their method was indeed flawed, which affects the conclusions of the article. Specifically, the stepwise classification method used in the article **overestimated the classification accuracy of who is a suicidal ideator because the features of the classifier were tuned to that particular dataset**.

Considerations when Building a ML Pipeline

Nested cross-validation

- Avoid overfitted generalization estimates

Data transformations in CV-consistent manner

- Confound removal
- Principal Components Analysis (PCA)

Rapidly evolving field

- New methods proposed regularly
- For feature engineering and learning

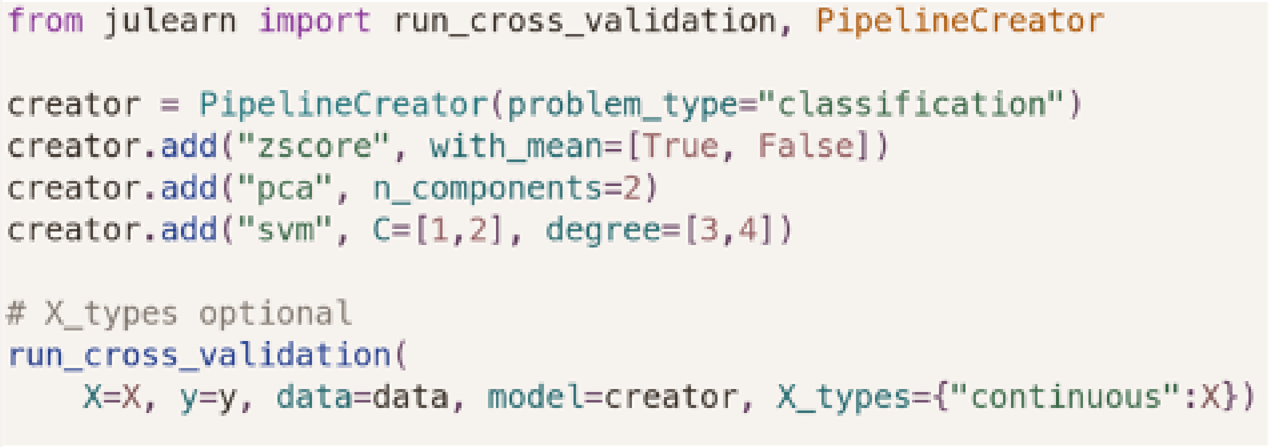
Expertise and overhead

- Programming and replicability
- Conceptual understanding

JULEARN: An easy-to-use ML library

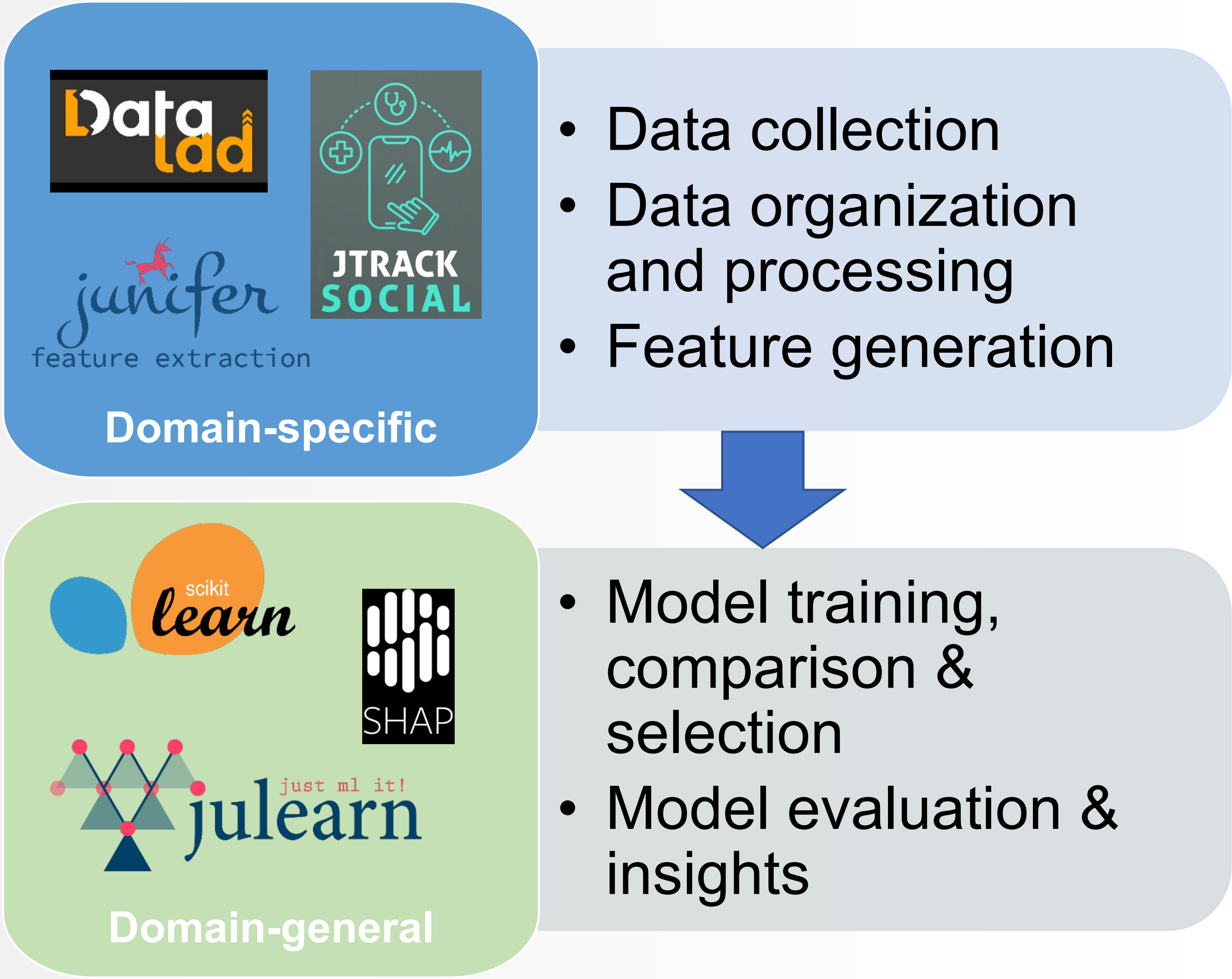
- One-line nested-CV pipelines
- Built-in CV-consistent data transformations
- Modular: plug-and-play scikit-learn transformers
- Data type support
- Specific models: CPM, CPMEX
- Built for non CS/ENG/ML
  - But suitable for them too!

keep calm and
julearn
run_cross_validation



```
from julearn import run_cross_validation, PipelineCreator

creator = PipelineCreator(problem_type="classification")
creator.add("zscore", with_mean=[True, False])
creator.add("pca", n_components=2)
creator.add("svm", C=[1,2], degree=[3,4])

# X_types optional
run_cross_validation(
    X=X, y=y, data=data, model=creator, X_types={"continuous":X})
```

From Data to ML Results



Domain-specific

- Data collection
- Data organization and processing
- Feature generation

Domain-general

- Model training, comparison & selection
- Model evaluation & insights

Thank you for your attention!

The AML group

Collaborators (FZJ)
Simon Eickhoff
Susanne Weis
Robert Langner
Felix Hoffstaedter
Masoud Thahmasian
… (and more)

Collaborators (external)
Julian Caspers, UKD
Christian Jung, UKD
Konrad Oexle, HMGU
Martin Reuter, DZNE
Götz Thomalla, UKE
Bastian Cheng, UKE
Bradley Love, UCL
Kshitij Jadhav, IITB
Nivethida T, IITB
… (and more)

Funding support

HELMHOLTZ IMAGING
HELMHOLTZAI
DFG Deutsche Forschungsgemeinschaft
Helmholtz Joint Lab SMHB Supercomputing and Modeling for the Human Brain
TVG