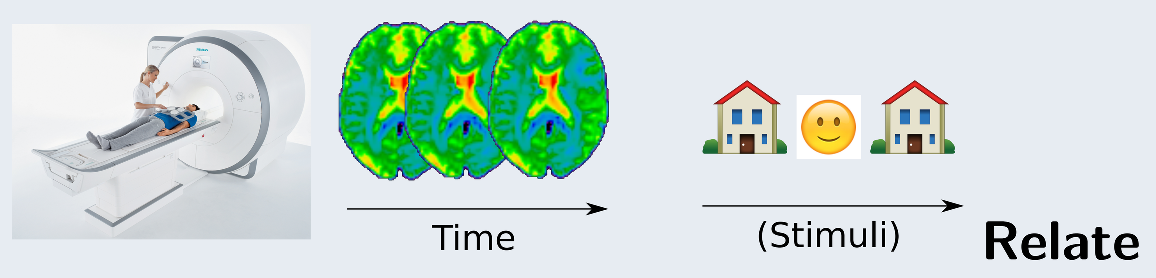
Learning Neural Representations of Human Cognition across Many fMRI Studies

Arthur Mensch[(1)] Julien Mairal[(2)] Danilo Bzdok[(3)] Bertrand Thirion[(1)] Gaël Varoquaux[(1)]

[(1)]Inria, CEA, Université Paris-Saclay, Gif-sur-Yvette, France [(2)]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, France [(3)]Department of Psychiatry, RWTH, Aachen, Germany

Functional MRI decoding



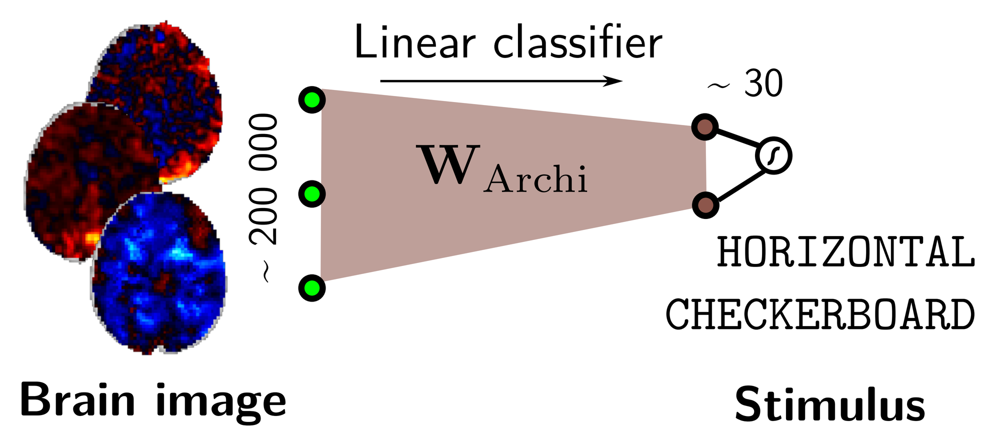
brain activation to cognitive stimuli

- Many small studies with ≠ psychological paradigms
- A few large scale studies (1000s subjects)

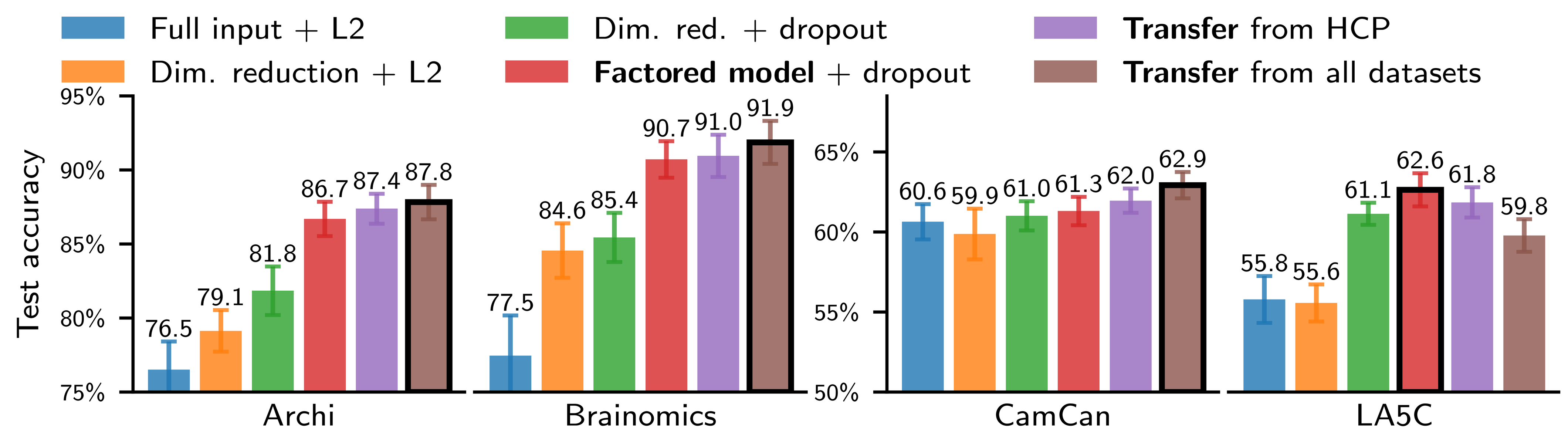
How to aggregate heterogeneous fMRI data into a common cognitive model ?

Learning setting

- First level GLM → **z-score maps**
- One map / record / base-condition
- Condition prediction on **new subjects**
- Baseline: **Multinomial regression**

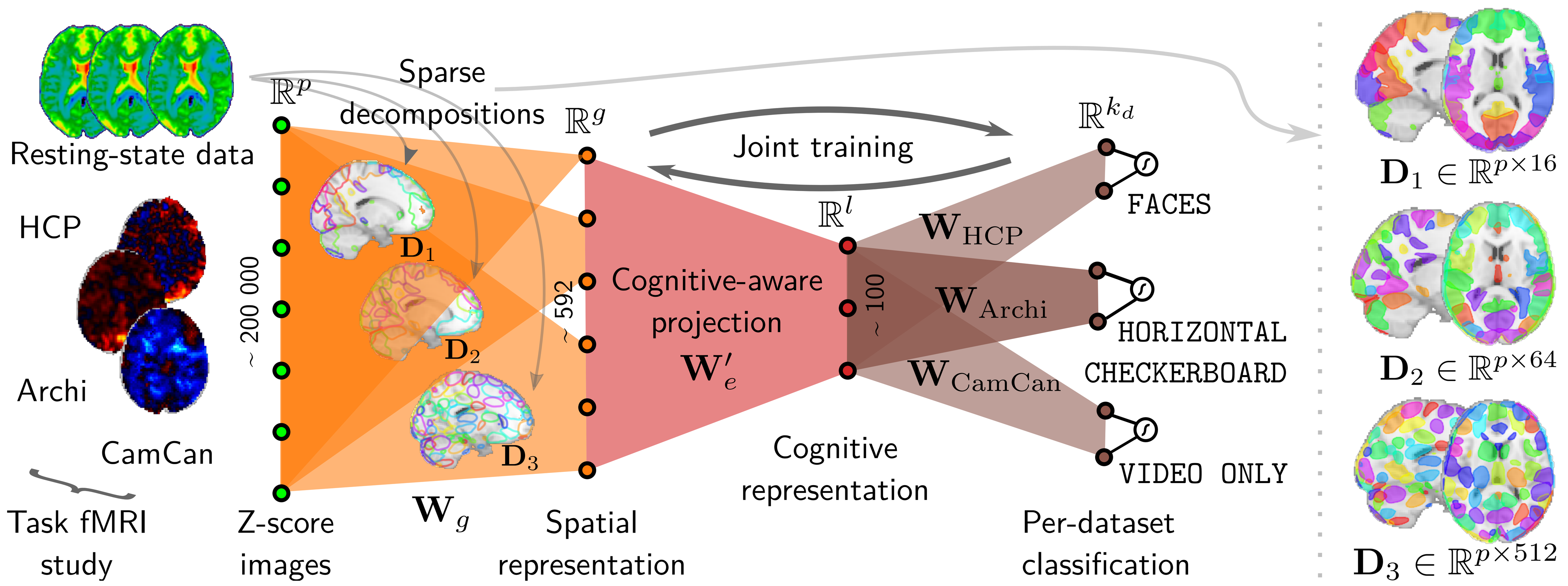


Performance



- **Dimension reduction** using resting state data is efficient regularization (and ↓ train cost)
- Extra efficient **latent layer + Dropout** (explains green to red improvement)
- **Transfer learning** occurs, and is stronger with more datasets (red → purple → brown)

Model



Dimension reduction $\mathbf{W}_g$

- **Sparse dictionaries** from HCP resting-state [2]

$$\mathbf{X}_{rs} = \mathbf{D}\mathbf{A} \in \mathbb{R}^{p \times g} \times g \times n, \qquad \mathbf{D} \text{ sparse}$$

- Orthogonal projection → one time-serie / component
- Unsupervised setting →: number of components ?
- → **Multi-scale** dictionaries

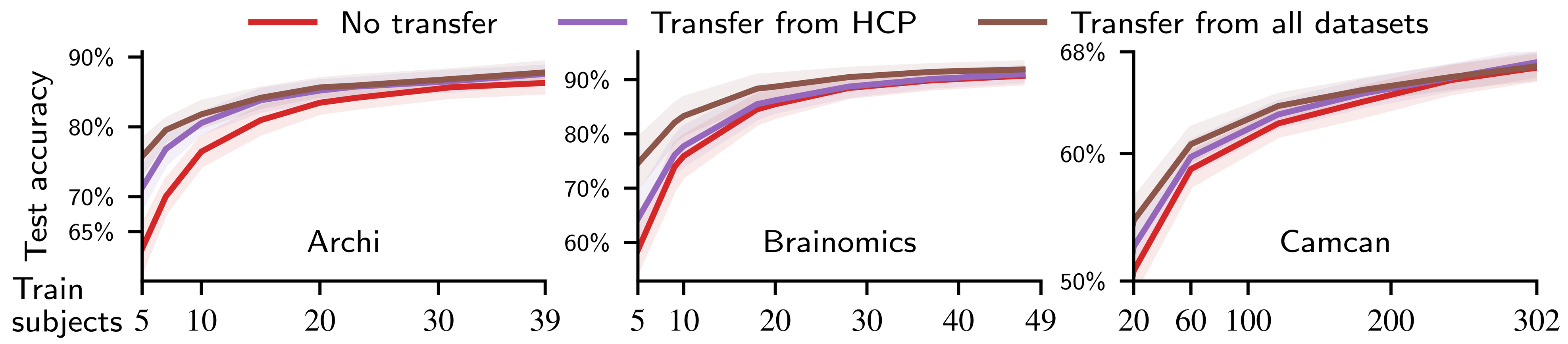
Latent space embedding

- Finding a common representation of brain images
- Easy to classify for **multiple datasets**
- Factorized linear model (with 2 layers, $l < g$)

$$\forall d \in D,\ \mathbf{W}_d = \mathbf{W}_e \mathbf{W}'_d \in \mathbb{R}^{g \times l} \times \mathbb{R}^{l \times k_d}$$

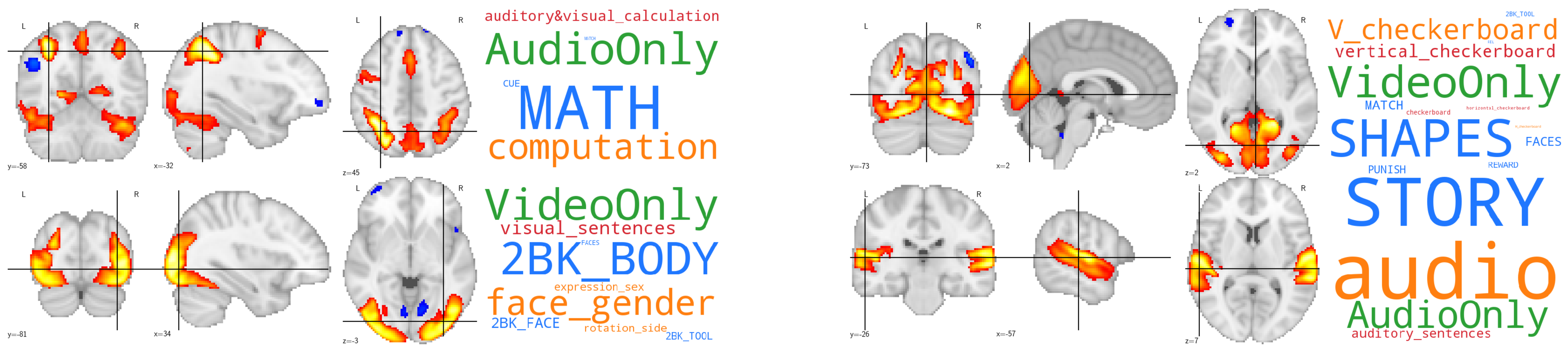
- **Share parameter for tranfer learning**

Transfer improves learning curves



Transfer learning becomes more useful as train-set size reduces

Cognitive space visualization



Meaningful template images associated to multi-dataset predictions (one color per dataset)

References

[1] A. Mensch, J. Mairal, D. Bzdok, B. Thirion, and G. Varoquaux. Learning neural representations of human cognition across many fmri studies. In *Advances in Neural Information Processing Systems*.

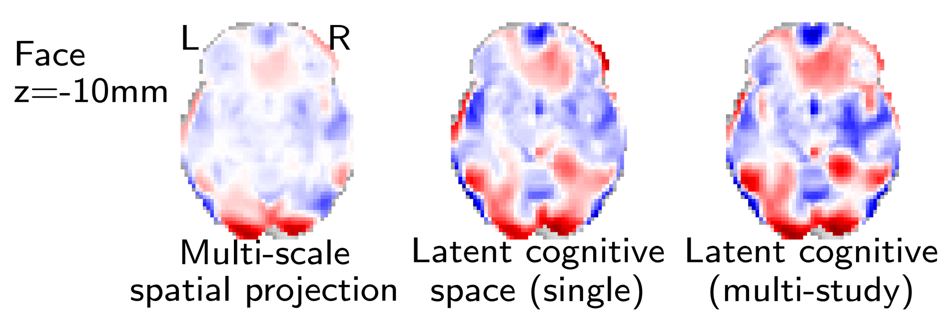
[2] A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux. Stochastic Subsampling for Factorizing Huge Matrices. *IEEE Transactions on Signal Processing*, 99(to appear), 2017.

Regularization

- We allow $l > k$: not a reduced rank regression
- Trivial with ℓ_2 regularization: no transfer
- **Dropout** allows transfer despite $\mathbf{W}_d$ full rank:

$$\text{(training)}\ \mathbf{y} = \mathbf{x}\mathbf{W}_e\mathbf{M}\mathbf{W}'_d, \qquad \mathbf{M} \text{ masking matrix}$$

Interpretability



Model → Higher-level regions (*e.g.* FFA)

Conclusion and future work

- **Scalable** and **paradigm-agnostic** model
- With evidence of **transfer learning**
- Dictionaries + code available

`github.com/arthurmensch/cogspaces`

- Model to be tested on the whole **openFMRI** repository