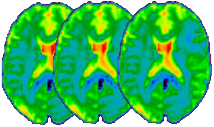
Learning Neural Representations of Human Cognition across Many fMRI Studies

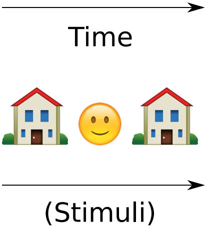
Arthur Mensch, Julien Mairal,
Danilo Bzdok, Gaël Varoquaux, Bertrand Thirion

Inria Parietal, Université Paris-Saclay

November 10, 2017

Introduction





Functional MRI data:

- **Undirect measure of brain activity**
- 200000 voxels ($8\,\mathrm{mm}^3$), every $2\,\mathrm{s}$.
- Value in each voxel $\sim$ activity in the area
- **Task fMRI**: timed psychological stimuli
- **Resting-state fMRI**: no stimulation

Objectives:

- Relate brain images to stimuli
- **Learn models of cognition from data**
- Validated by prediction on *unseen* subjects

Challenges and trends in fMRI

Challenges in functional MRI analysis:

- Very indirect measure of brain activity
- Low signal-to-noise ratio
- High inter-subject variability
- Gain statistical power by increasing data size

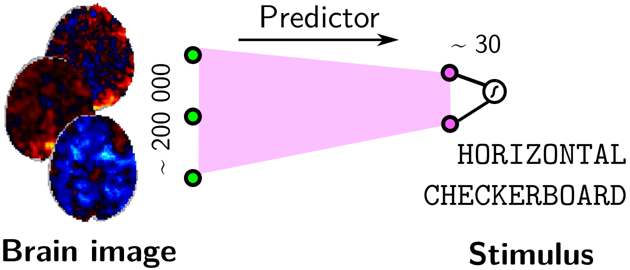
Community effort to provide larger datasets:

- From 15 subjects in 1997 to 100000 in 2020
- Costly studies on large cohort: HCP, UKBiobank
- Repositories of many different studies: OpenfMRI

Adapt algorithms and models to heavier data and richer labeling

Canonical learning problem for task fMRI

Prediction on a single dataset:



Decoding images from *unseen* subjects

What if we have several datasets at hand ?

Datasets in fMRI

Few big datasets

HCP

Resting-state
- 900 subjects
- 4 million samples
- 4 TB

Task
- 900 subjects
- 23 conditions
- 40,000 samples

...

UKBiobank 100,000 subjects

Many small datasets

Archi
- 30 cond.
- 80 subjects

Brainomics
- 18 cond.
- 100 subjects

Camcan
- 5 cond.
- 600 subjects

LA5c
- 10 cond.
- 200 subjects

...

OpenfMRI ~**150 studies**

Using all these datasets jointly to learn richer cognitive models?

Finding common aspects across fMRI studies

One of the oldest question in cognitive science: [Newell, 1973]

- Different cognitive stimuli for each task fMRI study

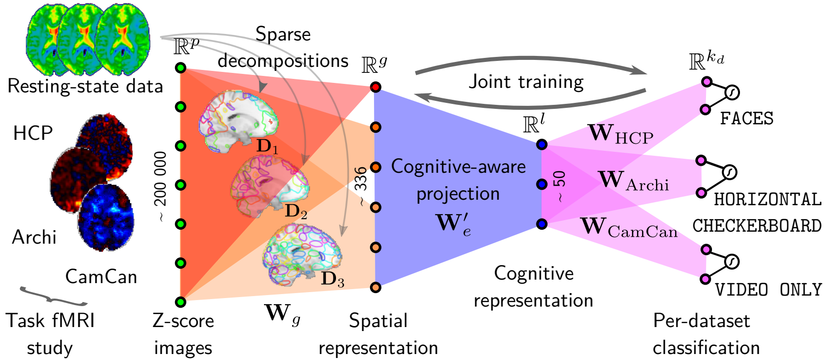
Challenges: Many-to-many mapping:

- Experimental stimulus $\rightarrow$ several brain regions
- Brain region $\rightarrow$ several brain functions

Previous approaches:

- Meta-analysis Noisy, coordinate-based
  - [Salimi-Khorshidi et al., 2009, Yarkoni et al., 2011]
- Decoding a cognitive ontology across studies Not scalable
  - [Schwartz et al., 2013, Wager et al., 2013]

Learn a common cognitive base where decoding is easy.



Condition prediction for task fMRI

Common structure from every task fMRI dataset d:

- Cognitive stimulus $\underset{\text{GLM}}{\rightarrow}$ contrast map $\mathbf{x} \in \mathbb{R}^p$

Predictive task:

- *Input:* contrast $\mathbf{x} \in \mathbb{R}^p$ from dataset d
- *Output:* condition c from dataset d.

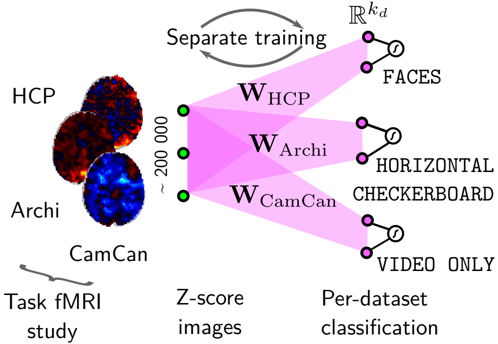
Multinomial classification

Parameters for dataset d: $(\mathbf{W}_d, \mathbf{b}_d) \in \mathbb{R}^{p \times k}, \mathbb{R}^k$

$$\mathbf{p}_d(\mathbf{x}, \mathbf{W}_d, \mathbf{b}_d) \in \mathbb{R}^c \triangleq \mathrm{softmax}(\mathbf{W}_d^\top \mathbf{x} + \mathbf{b}_d)$$

Condition prediction for task fMRI

Baseline: learn all multinomial regression separately



- **Few samples** per dataset ($n_d \sim 1000$)
- **High dimension** $W_d \in \mathbb{R}^{p \times k}$

$\rightarrow$ Aggregate datasets, **reduce** the dimension

Learning shared representations

Common structure for classification vectors $(\mathbf{W}_d)_d$:

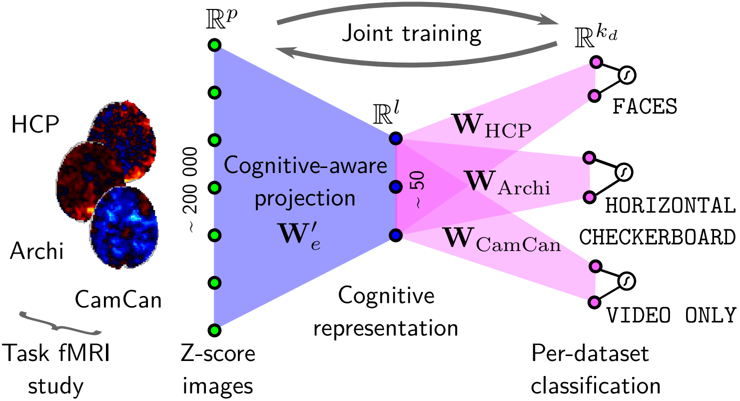
- Stimuli shares some common aspects
- Vector $\mathbf{W}_{\mathrm{Archi}}^{\mathrm{Faces}}$ should be close to $\mathbf{W}_{\mathrm{HCP}}^{\mathrm{Faces}}$
- Enforce *parameter sharing*

Shared layer model

- $\mathbf{W}_d = \mathbf{W}_e \mathbf{W}'_d$, where $\mathbf{W}_e \in \mathbb{R}^{p \times l}$ is shared across datasets.
- Joint minimization objective
- Regularization to enforce parameter sharing

Shared layer model

Joint training: Learn to reduce dimension trough $\mathbf{W}_e$ and to classify through $(\mathbf{W}'_d)_d$.



Dimension reduction using resting-state fMRI

Initial dimension reduction:

- Compute a dictionary of brain areas
- Represent a brain image from activations in these areas

How to learn a dictionary with good generic performances:

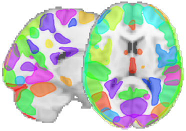
- Use resting-state data ! [Blumensath et al., 2013]
- Much more available than task data + not biased by stimuli

Dictionary learning on resting-state data

$$\mathbf{X}_{rs} \in \mathbb{R}^{p \times n} \sim \mathbf{D} \times \mathbf{A}$$

- $\mathbf{D} \in \mathbb{R}^{p \times g}, \mathbf{A} \in \mathbb{R}^{g \times n}$ with $\mathbf{D}$ sparse
- Contrast map $\mathbf{x} \to$ brain area loadings $\mathbf{x}_g$:

$$\mathbf{x}_g = \mathbf{W}_g^\top x = (\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top \mathbf{x}$$



$\mathbf{D}_2 \in \mathbb{R}^{p \times 64}$

Multi-scale dictionary

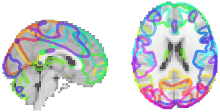
Unsupervised dictionary learning:

- A single **D** captures signal at the scale of its components
- Larger dictionaries → smaller components
- Unknown best number of components

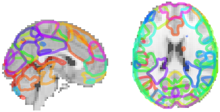
Solution:

→ Dictionaries with different scales

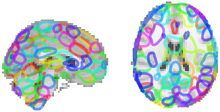
- *Overcomplete* unsupervised dimension reduction



$\mathbf{D}_1$ (16 components)



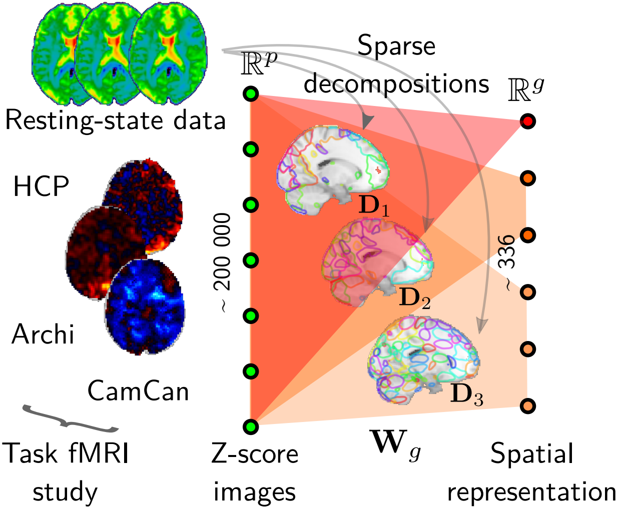
$\mathbf{D}_2$ (64 components)



$\mathbf{D}_3$ (256 components)

Unsupervised-learning driven dimension reduction

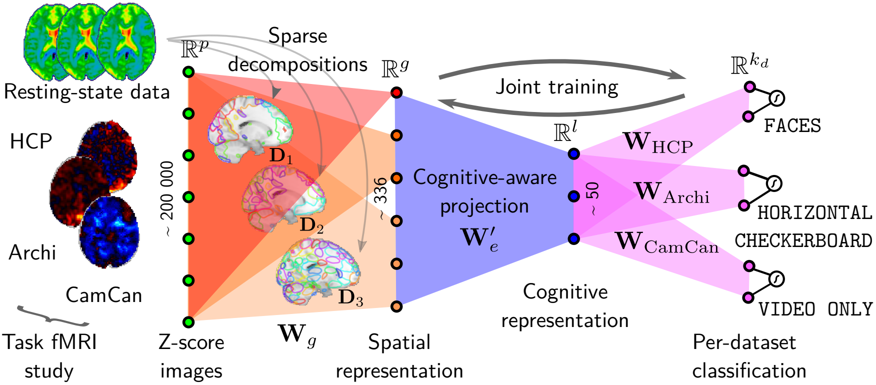
Concatenate loadings $\mathbf{x}_g(\mathbf{D})$, from various dictionaries:



The full model

Three layer linear network:

- 1st layer: Unsupervised learning on resting-state
- 2nd/3rd layer: Joint training on many datasets



Training and regularization

Implementation ingredients:

- Computing $(\mathbf{D}_j)_j$: using [Mensch et al., 2017] (a few hours)
- Learn $\mathbf{W}'_e, (\mathbf{W}'_d)_d$ with stochastic gradient descent
- Use *Dropout* [Srivastava et al., 2014] as a regularizer

Dropout

- Randomly set coordinates to zero in $\mathbf{W}_g^\top \mathbf{x}$, $\mathbf{W}_e^\top \mathbf{W}_g^\top \mathbf{x}$.
- Prevent co-adaption of coordinates

$=$ Some columns of $\mathbf{W}_e$ used to classify a single dataset only

Datasets

Few big datasets

HCP

Resting-state
- 900 subjects
- 4 million samples
- 4 TB

Task
- 900 subjects
- 23 conditions
- 40,000 samples

...

UKBiobank 100,000 subjects

Many small datasets

Archi
- 30 cond.
- 80 subjects

Brainomics
- 18 cond.
- 100 subjects

Camcan
- 5 cond.
- 600 subjects

LA5c
- 10 cond.
- 200 subjects

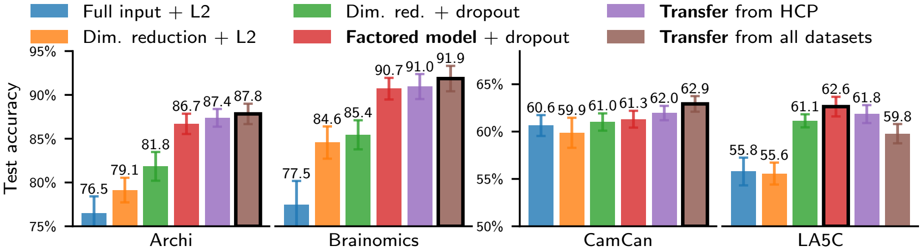
...

OpenfMRI $\sim$**150 studies**

Datasets

- **HCP**: gambling, working memory, motor, language, social and relational tasks. 800 subjects.
- **Archi** [Pinel et al., 2007]: localizer protocol, motor, social and relational task. 79 subjects.
- **Brainomics** [Papadopoulos Orfanos et al., 2017]: localizer protocol. 100 subjects.
- **Camcan** [Shafto et al., 2014]: audio-video task, with frequency variation. 600 subjects.
- **LA5c consortium** [Poldrack et al., 2016]: task-switching, balloon analog risk taking, stop-signal and spatial working memory capacity tasks — high-level tasks. 200 subjects.

Model performance

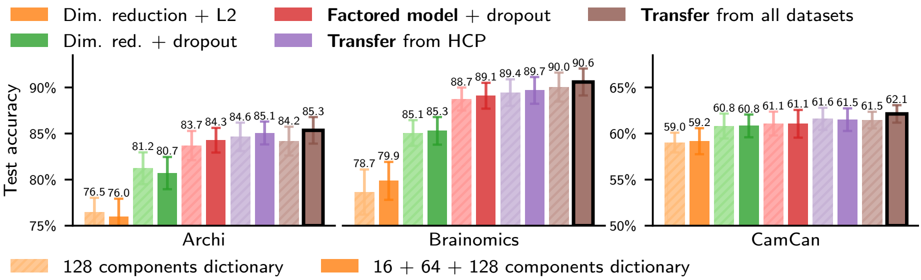


Multi-scale projection, **latent space learning** are useful.

Dropout is efficient.

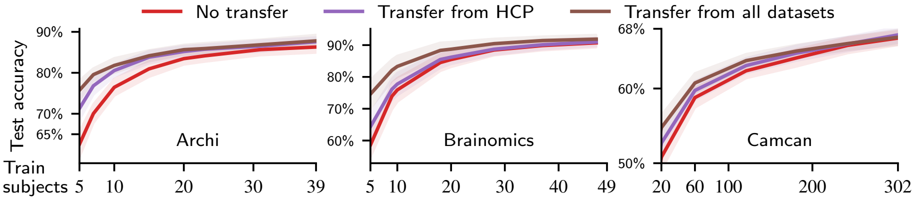
Transfer learning occurs

Model performance



Multi-scale projection is beneficial

Transfer on small datasets



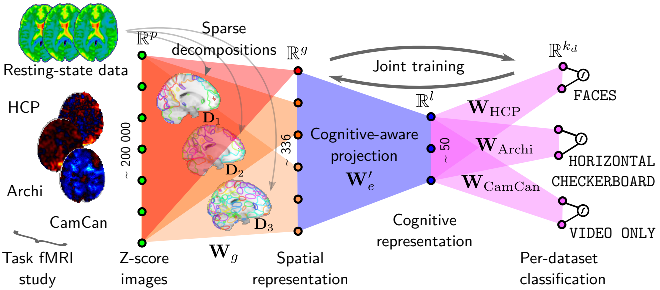
Joint training helps handling **small sample studies**

Better transfer effects using more datasets

Latent space exploration

Find representative elements of the projected data $(\mathbf{x}_l)$ in $\mathbb{R}^l$:

- K-means centroids $\{\mathbf{c}_k\}_k$ of projected data $(\mathbf{x}_l)_l$

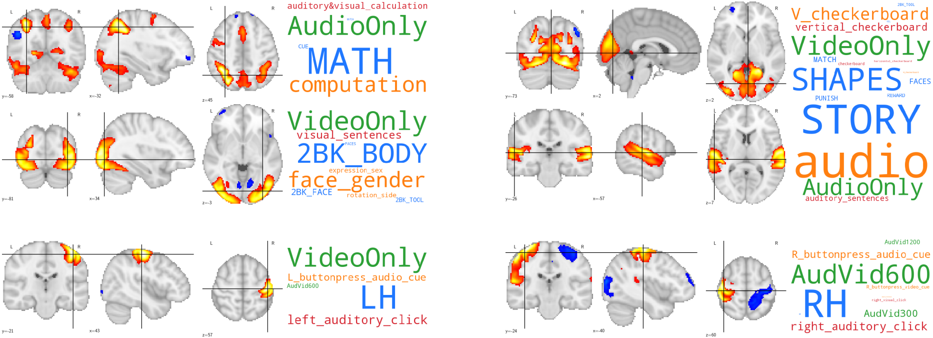


Backward and forward in the model:

$\rightarrow$ **Forward:** Classification vectors for each dataset $\{[\mathbf{W}'_{\mathrm{Archi}}, \ldots, \mathbf{W}'_{\mathrm{HCP}}]^\top \mathbf{c}\}$

$\leftarrow$ **Backward:** Brain maps (inverse images of $\{\mathbf{c}_k\}_k$)

Latent space exploration



Meaningful template images associated to multi-dataset predictions

Conclusion

A new model with better classification performance:

- Efficient use of unsupervised learning on resting-state data
- Good performance of regularization techniques and models from the deep learning (non-convex) world

Learning **representation of brain images** easy to label with **cognitive stimuli**

Evidence of transfer learning:

- Ongoing extended validation on larger data repositories (OpenfMRI)

Bibliography I

[Blumensath et al., 2013] Blumensath, T., Jbabdi, S., Glasser, M. F., Van Essen, D. C., Ugurbil, K., Behrens, T. E., and Smith, S. M. (2013).
Spatially constrained hierarchical parcellation of the brain with resting-state fMRI.
NeuroImage, 76:313–324.

[Dohmatob et al., 2016] Dohmatob, E., Mensch, A., Varoquaux, G., and Thirion, B. (2016).
Learning brain regions via large-scale online structured sparse dictionary-learning.
In *30th Conference on Neural Information Processing Systems (NIPS)*.

[Mensch et al., 2016a] Mensch, A., Mairal, J., Thirion, B., and Varoquaux, G. (2016a).
Dictionary learning for massive matrix factorization.
In *33rd International Conference on Machine Learning (ICML)*.

Bibliography II

[Mensch et al., 2016b] Mensch, A., Mairal, J., Thirion, B., and Varoquaux, G. (2016b).
Subsampled online matrix factorization with convergence guarantees.
In *NIPS Workshop on Optimization for Machine Learning (OPT)*.

[Mensch et al., 2017] Mensch, A., Mairal, J., Thirion, B., and Varoquaux, G. (2017).
Stochastic Subsampling for Factorizing Huge Matrices.
arXiv:1701.05363 [cs, math, q-bio, stat].

[Mensch et al., 2016c] Mensch, A., Varoquaux, G., and Thirion, B. (2016c).
Compressed Online Dictionary Learning for Fast fMRI Decomposition.
In *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*.

[Newell, 1973] Newell, A. (1973).
You can't play 20 questions with nature and win: Projective comments on the papers of this symposium.

Bibliography III

[Papadopoulos Orfanos et al., 2017] Papadopoulos Orfanos, D., Michel, V., Schwartz, Y., Pinel, P., Moreno, A., Le Bihan, D., and Frouin, V. (2017).
The Brainomics/Localizer database.
NeuroImage, 144:309–314.

[Pinel et al., 2007] Pinel, P., Thirion, B., Meriaux, S., Jobert, A., Serres, J., Le Bihan, D., Poline, J.-B., and Dehaene, S. (2007).
Fast reproducible identification and large-scale databasing of individual functional cognitive networks.
BMC neuroscience, 8(1):91.

[Poldrack et al., 2016] Poldrack, R., Congdon, E., Triplett, W., Gorgolewski, K., Karlsgodt, K., Mumford, J., Sabb, F., Freimer, N., London, E., Cannon, T., and Bilder, R. (2016).
A phenome-wide examination of neural and cognitive function.
bioRxiv, page 059733.

Bibliography IV

[Salimi-Khorshidi et al., 2009] Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D., and Nichols, T. E. (2009).
Meta-analysis of neuroimaging data: A comparison of image-based and coordinate-based pooling of studies.
NeuroImage, 45(3):810–823.

[Schwartz et al., 2013] Schwartz, Y., Thirion, B., and Varoquaux, G. (2013).
Mapping paradigm ontologies to and from the brain.
In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 1673–1681. Curran Associates, Inc.

[Shafto et al., 2014] Shafto, M. A., Tyler, L. K., Dixon, M., Taylor, J. R., Rowe, J. B., Cusack, R., Calder, A. J., Marslen-Wilson, W. D., Duncan, J., Dalgleish, T., Henson, R. N., Brayne, C., and Matthews, F. E. (2014).
The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: A cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing.
BMC Neurology, 14.

Bibliography V

[Srivastava et al., 2014] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014).
Dropout: A simple way to prevent neural networks from overfitting.
Journal of Machine Learning Research, 15(1):1929–1958.

[Wager et al., 2013] Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., and Kross, E. (2013).
An fMRI-Based Neurologic Signature of Physical Pain.
New England Journal of Medicine, 368(15):1388–1397.

[Yarkoni et al., 2011] Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011).
Large-scale automated synthesis of human functional neuroimaging data.
Nature methods, 8(8):665–670.