Learning representations from functional MRI data

Arthur Mensch

CEA, Inria, Parietal team, Université Paris-Saclay

January 30, 2019

PhD advisors: Gaël Varoquaux *(CEA, Inria)*, Julien Mairal *(Inria, Univ. Grenoble Alpes)*, Bertrand Thirion *(CEA, Inria)*

Joint work with: Kamalakar Reddi *(CEA, Inria)*, Elvis Dohmatob *(CEA, Inria)*, Danilo Bzdok *(RWTH Aachen University)*

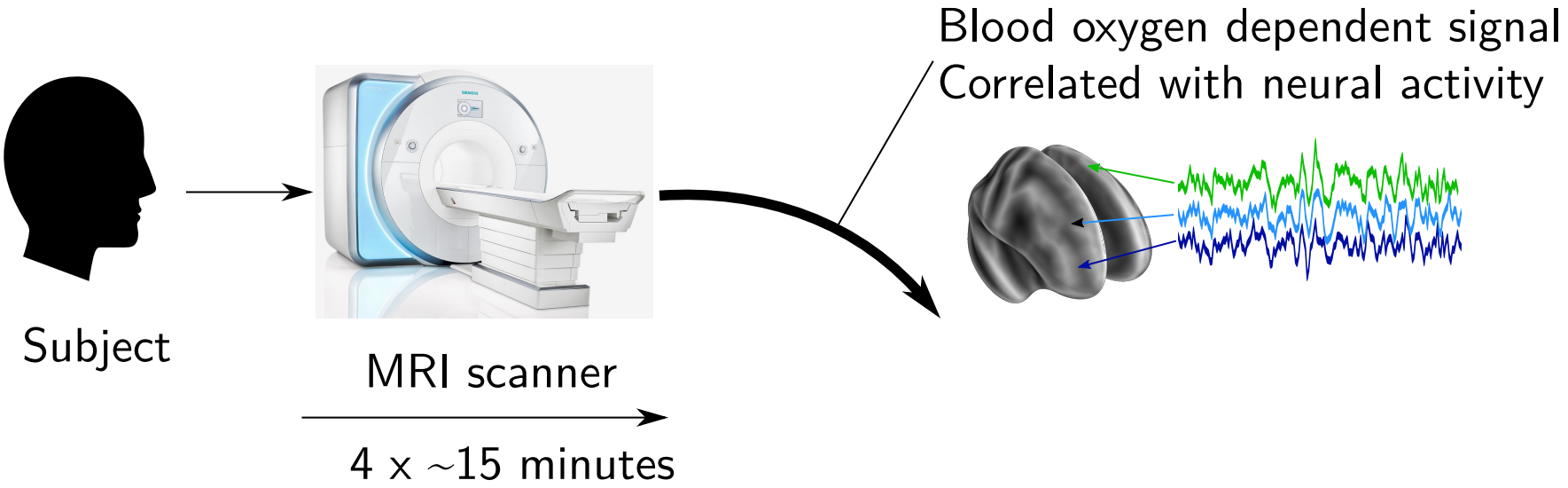
Functional Magnetic Resonance Imaging (fMRI)

Subject

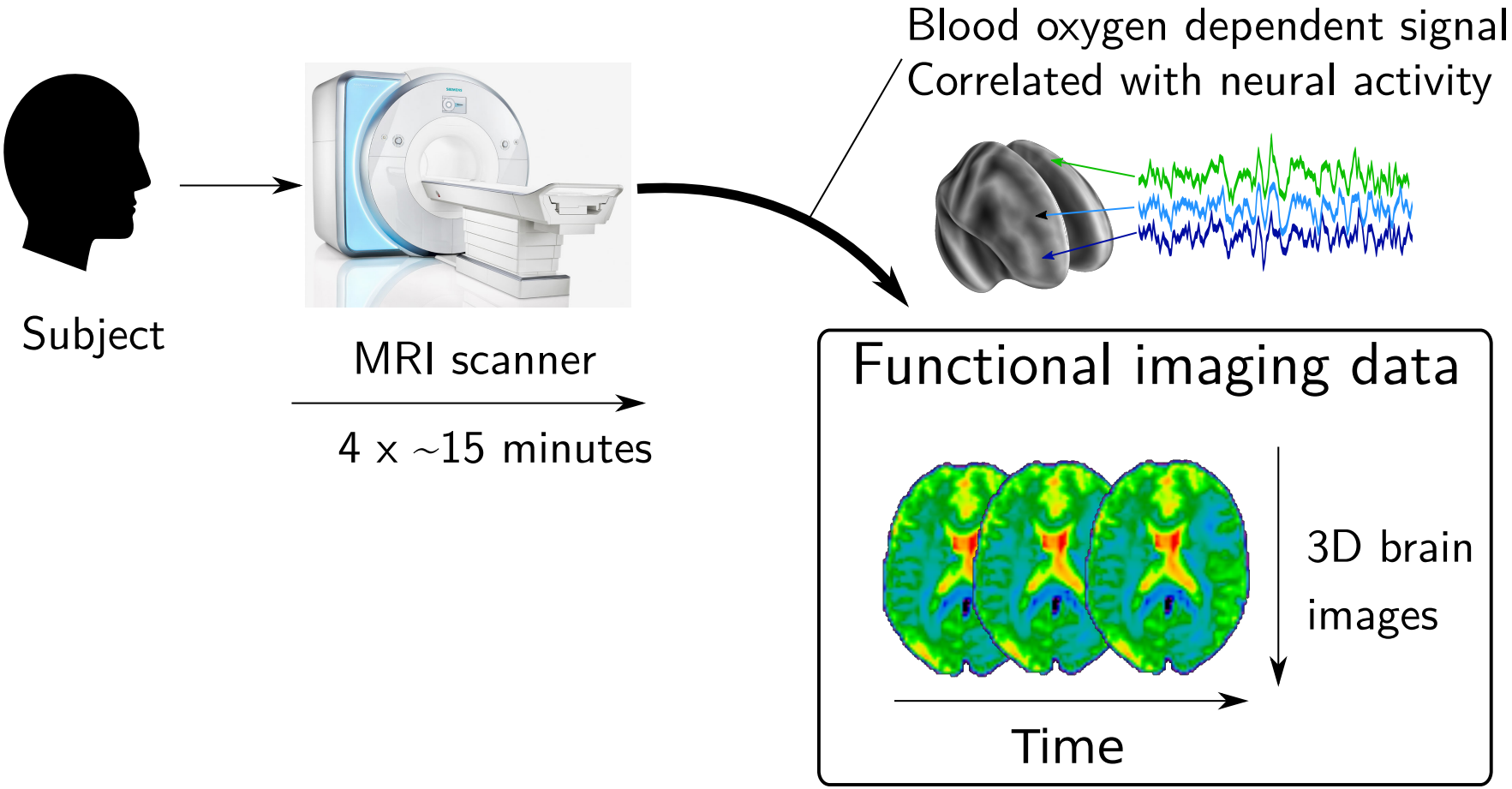
MRI scanner

4 x ~15 minutes

Functional Magnetic Resonance Imaging (fMRI)



Functional Magnetic Resonance Imaging (fMRI)



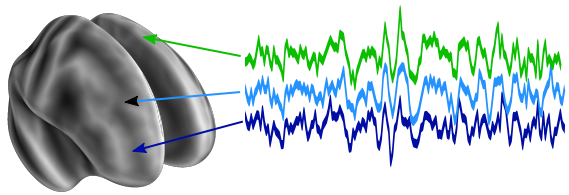
Functional Magnetic Resonance Imaging (fMRI)

Subject

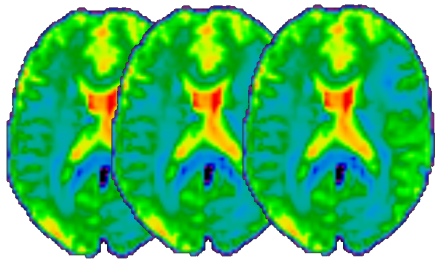
MRI scanner

4 x ~15 minutes

Blood oxygen dependent signal
Correlated with neural activity



Functional imaging data



3D brain images

Time

Task control

Stimuli

Marginal associations **Task fMRI**

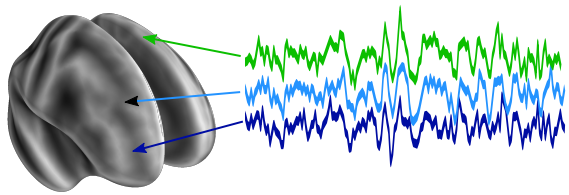
Functional Magnetic Resonance Imaging (fMRI)

Subject

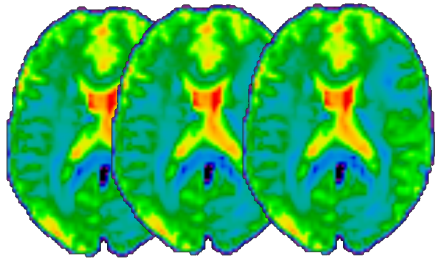
MRI scanner

4 x ~15 minutes

Blood oxygen dependent signal
Correlated with neural activity

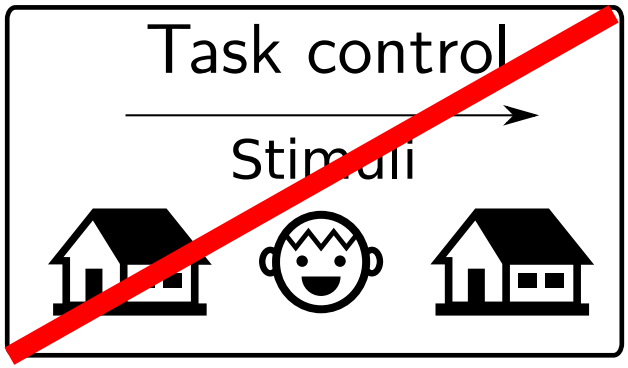


Functional imaging data



3D brain images

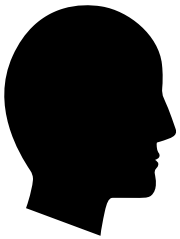
Time



Population study

Resting state fMRI

Functional Magnetic Resonance Imaging (fMRI)

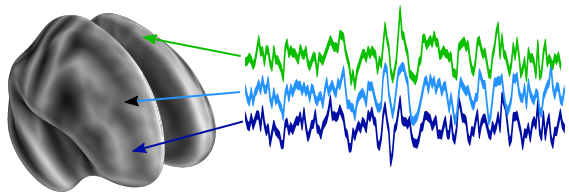


Subject

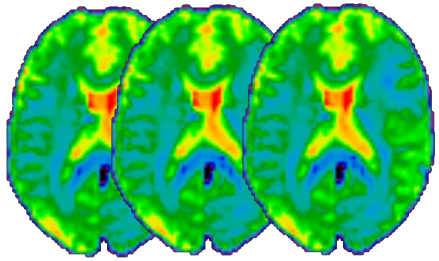
MRI scanner

4 x ~15 minutes

Blood oxygen dependent signal
Correlated with neural activity



Functional imaging data



3D brain images

Time

Task control

Stimuli

8 mm3/voxel

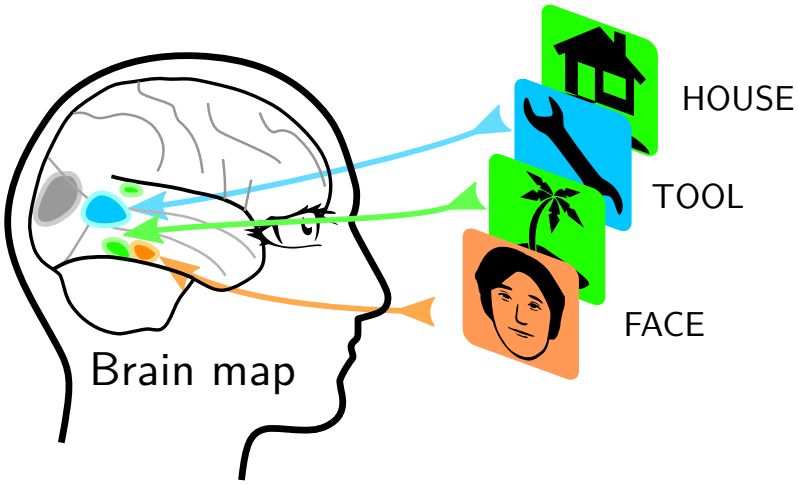
200,000 voxels

1 frame/s

1,000 frames/rec

Functional MRI data for cognitive science

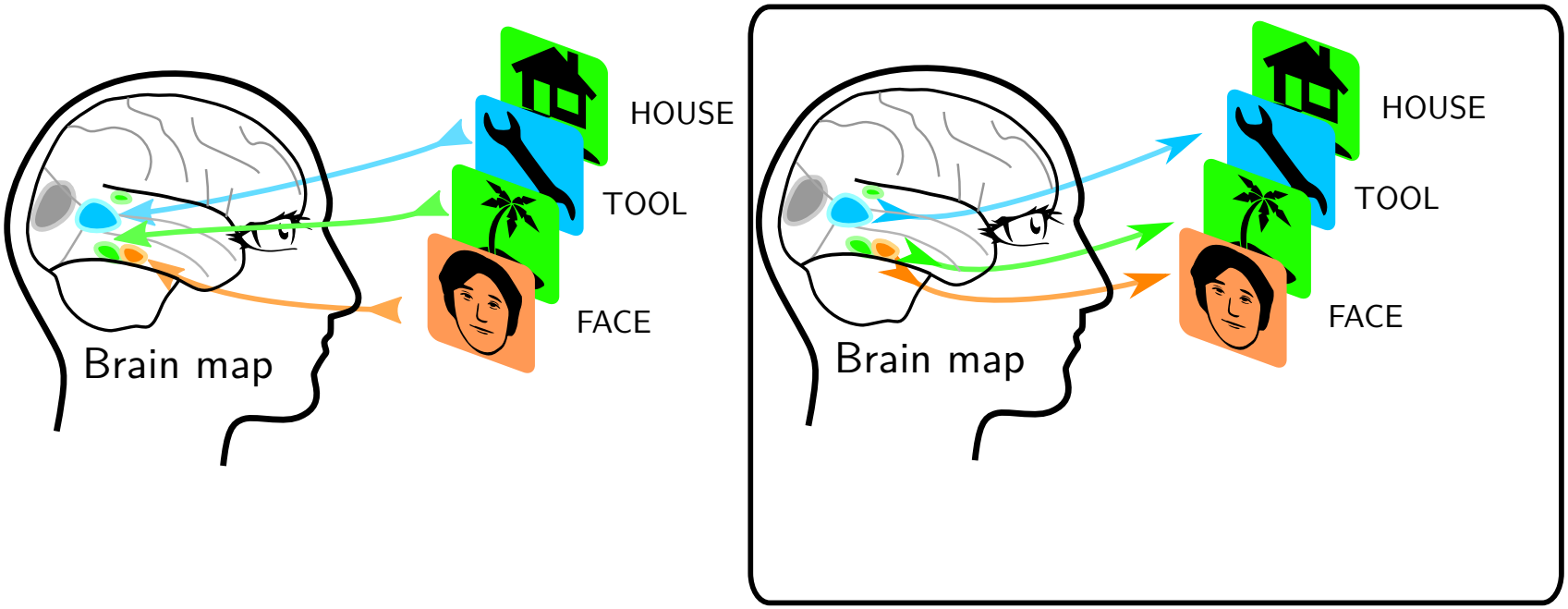
Relate **brain activation** to **cognitive activity**



Encoding models

Functional MRI data for cognitive science

Relate **brain activation** to **cognitive activity**

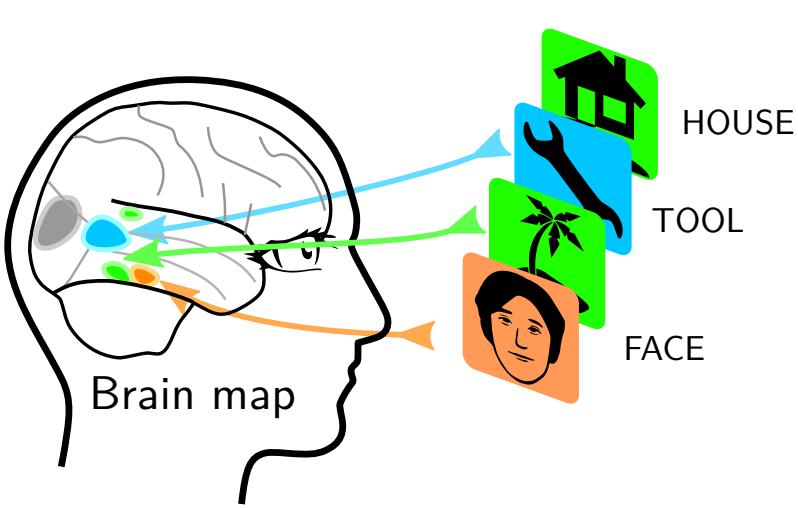


Encoding models

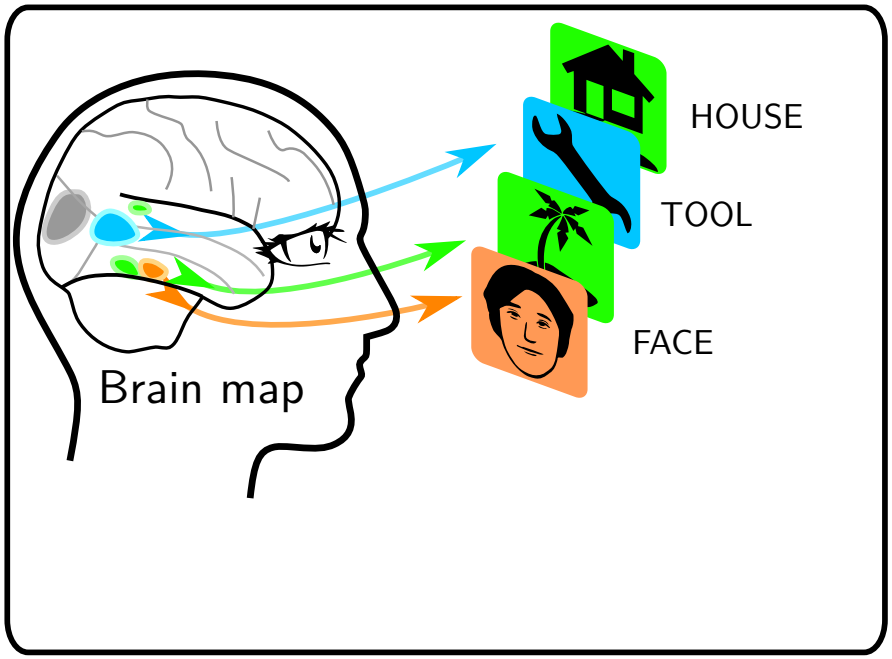
Decoding models

Functional MRI data for cognitive science

Relate **brain activation** to **cognitive activity**



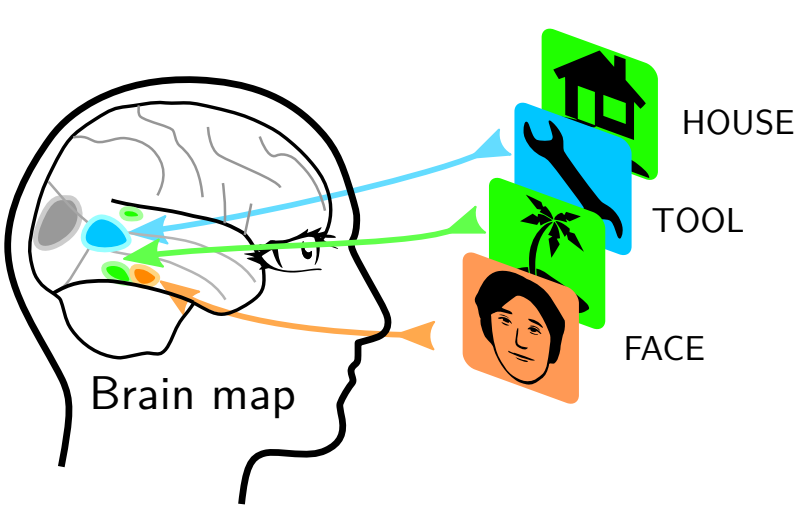
Encoding models
(voxel independent)



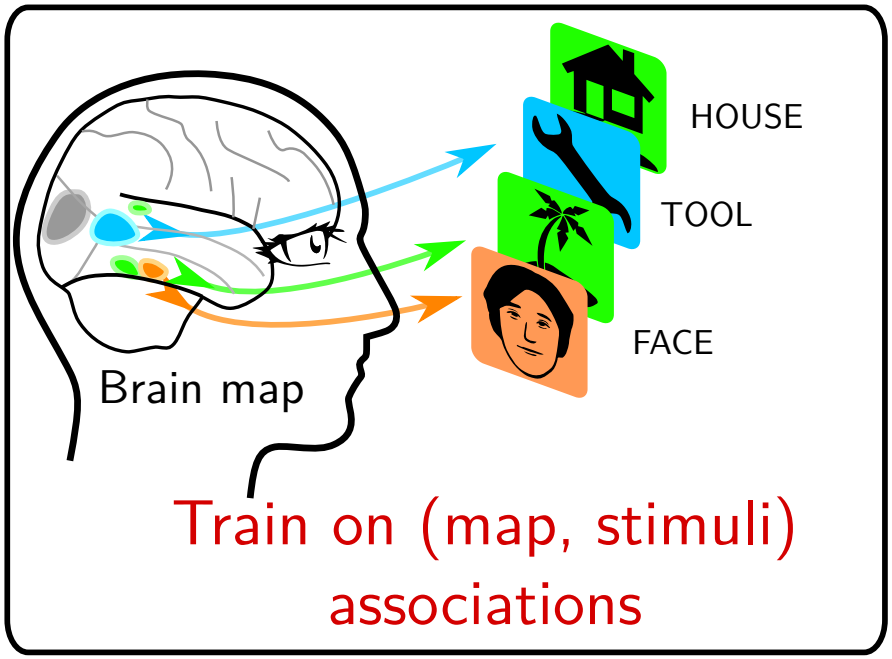
Decoding models
(whole brain)

Functional MRI data for cognitive science

Relate **brain activation** to **cognitive activity**



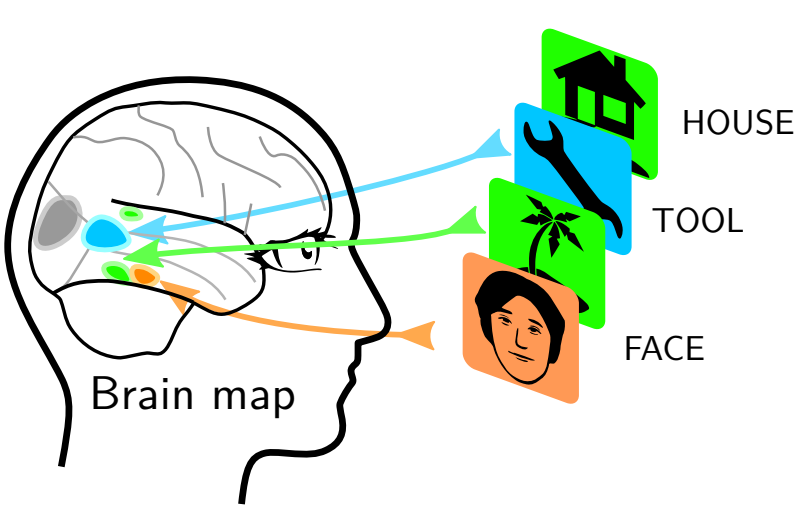
Encoding models
(voxel independent)



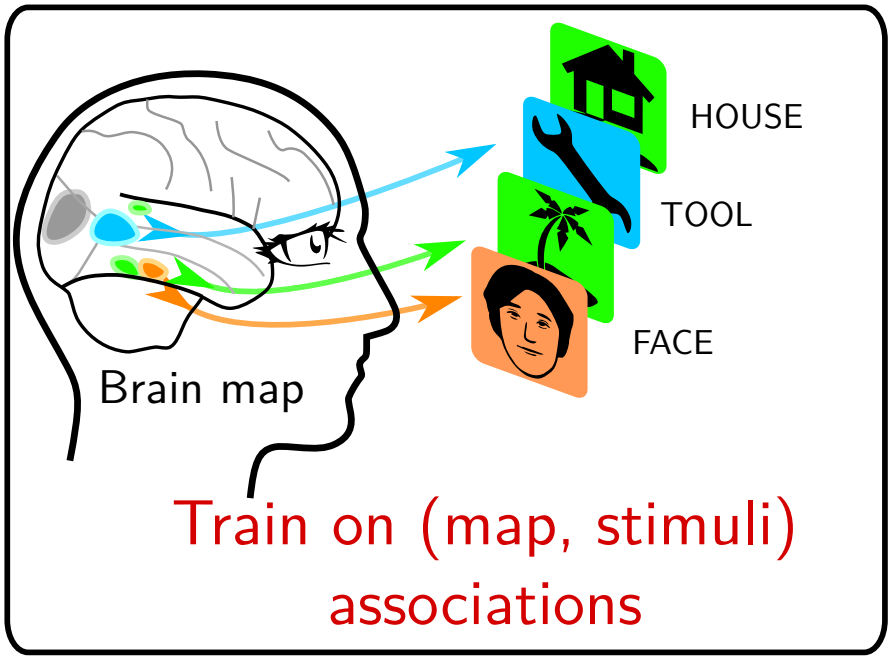
Decoding models
(whole brain)

Functional MRI data for cognitive science

Relate **brain activation** to **cognitive activity**



Encoding models
(voxel independent)



Decoding models
(whole brain)

Goal: Inspect the trained models for cognitive evidence

Training decoding models

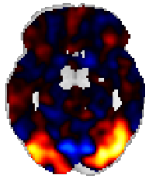
Subject

Psycho. condition

Training set

Encoding

Statistical map

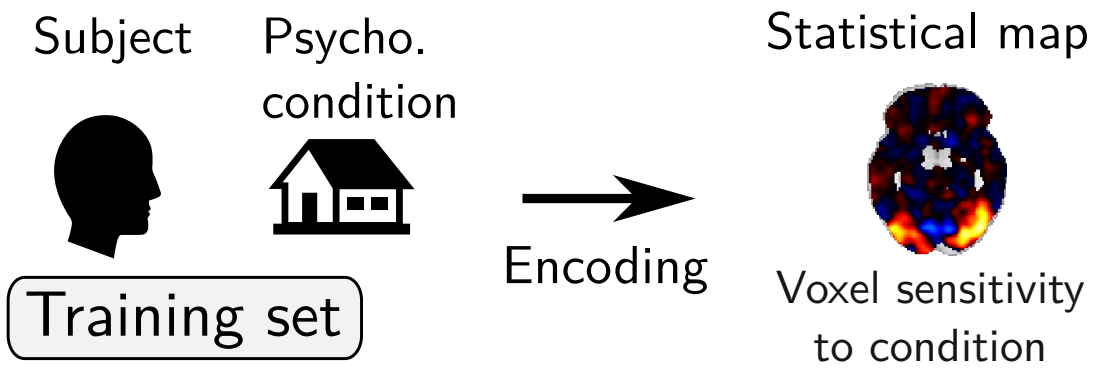


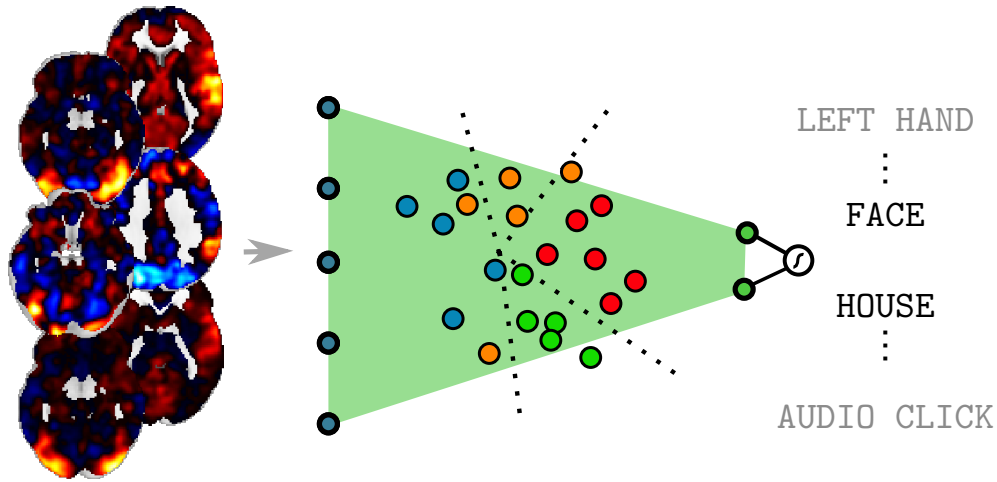
Voxel sensitivity to condition

Machine learning setting

- Input maps (flattened) $\boldsymbol{x}^{(i)} \in \mathcal{X} \subset \mathbb{R}^{200{,}000}$
- Output conditions $\boldsymbol{y}^{(i)} \in \mathcal{Y} = \Delta^c$

Training decoding models

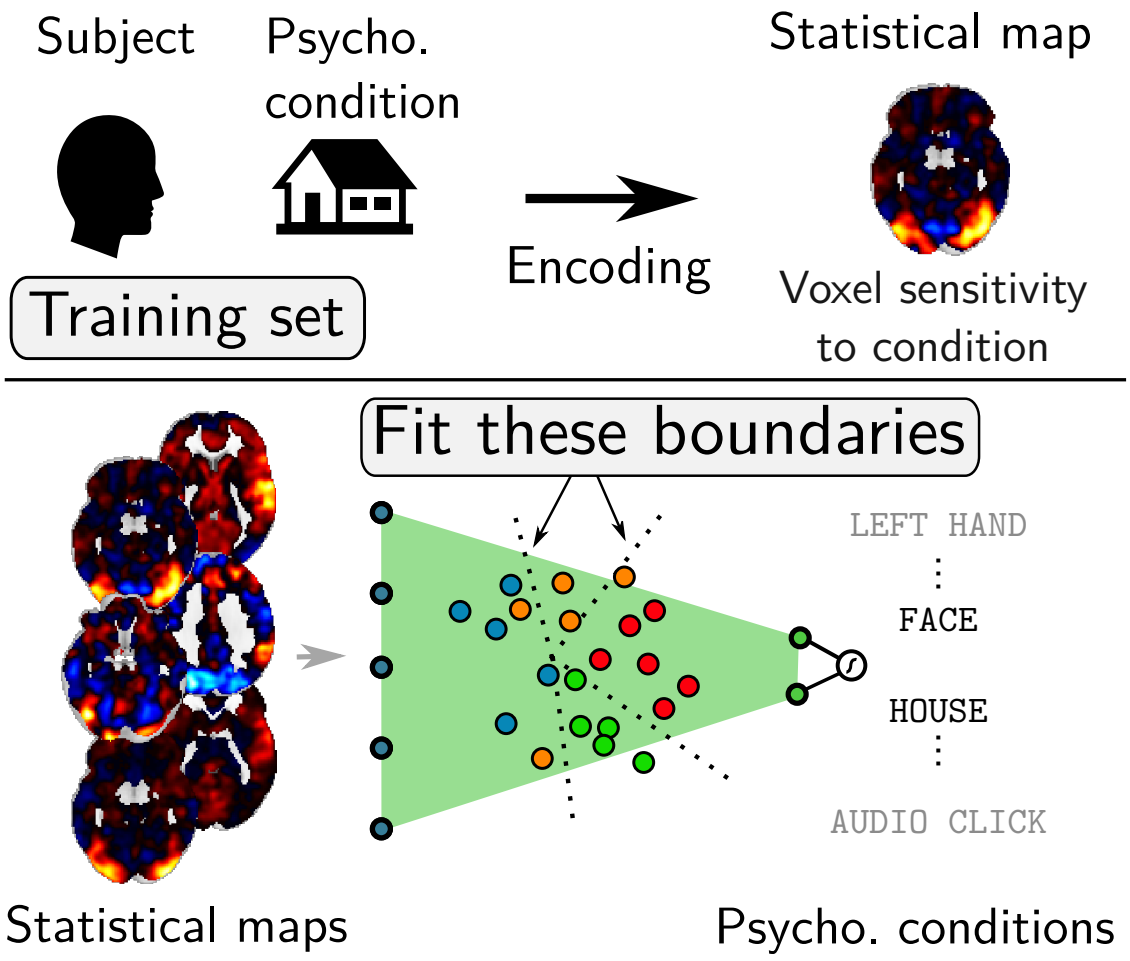




Machine learning setting

- Input maps (flattened) $\boldsymbol{x}^{(i)} \in \mathcal{X} \subset \mathbb{R}^{200{,}000}$
- Output conditions $\boldsymbol{y}^{(i)} \in \mathcal{Y} = \Delta^c$
- Predictive model $\boldsymbol{f}_\theta : \mathcal{X} \to \mathcal{Y}$

Training decoding models



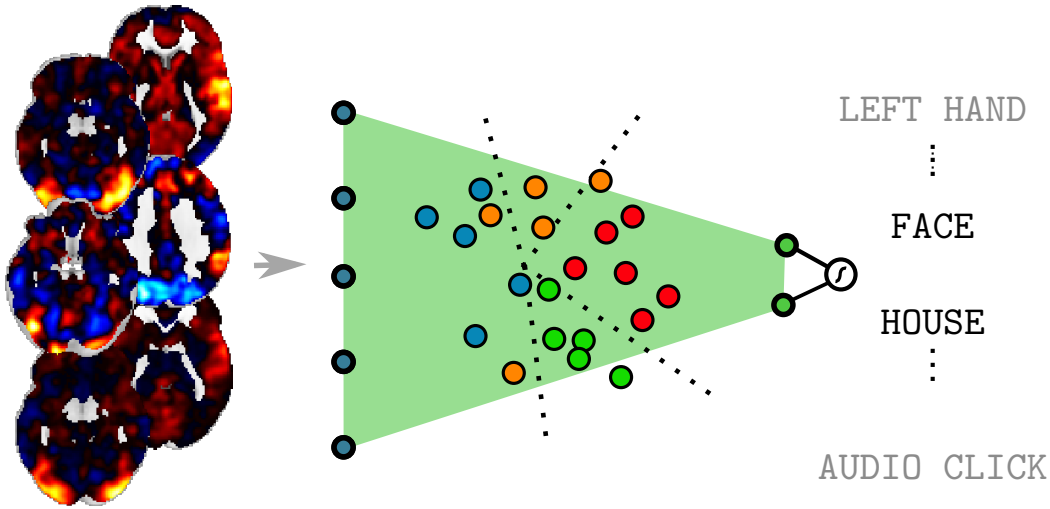
Machine learning setting

- Input maps (flattened) $\boldsymbol{x}^{(i)} \in \mathcal{X} \subset \mathbb{R}^{200{,}000}$
- Output conditions $\boldsymbol{y}^{(i)} \in \mathcal{Y} = \Delta^c$
- Predictive model $\boldsymbol{f}_\theta : \mathcal{X} \to \mathcal{Y}$

Training: tune $\theta \in \Theta$

- Prediction $\hat{\boldsymbol{y}}^{(i)} \triangleq f_\theta(\boldsymbol{x}^{(i)})$
- Should match $\boldsymbol{y}^{(i)}$

Generalizing beyond the train set

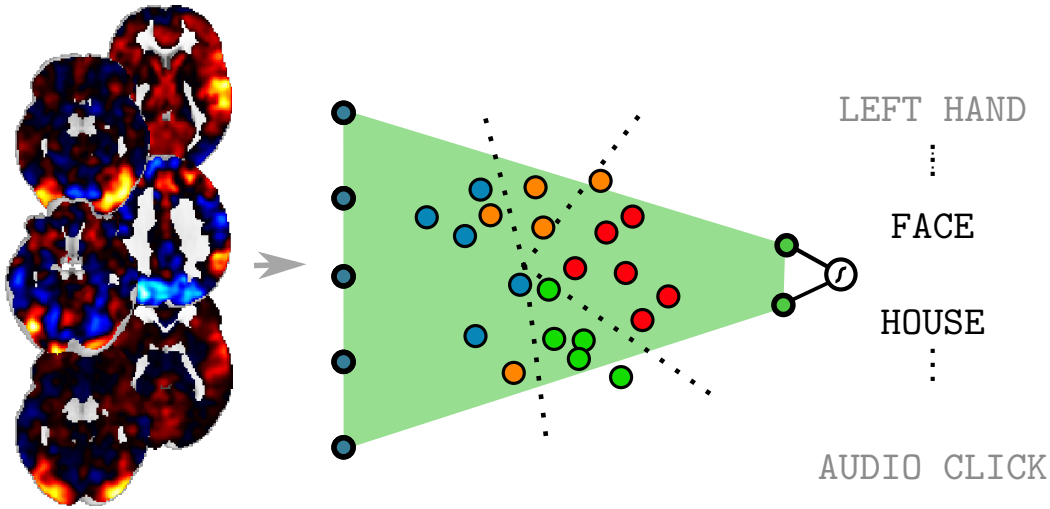


Statistical maps

Psycho. condition

A good model should handle:

Generalizing beyond the train set



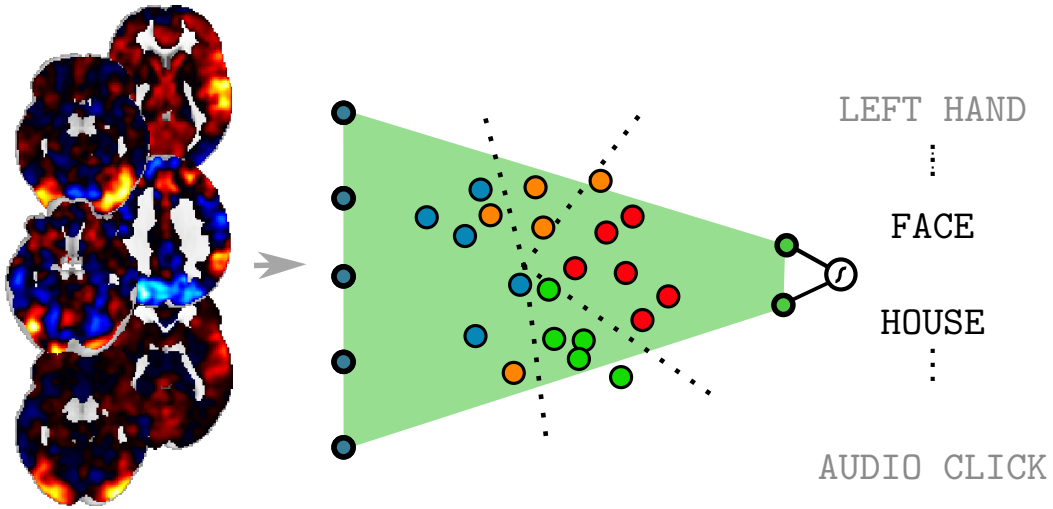
Statistical maps

Psycho. condition

A good model should handle:

New records
= Inter and intra-subject
variability

Generalizing beyond the train set



Statistical maps

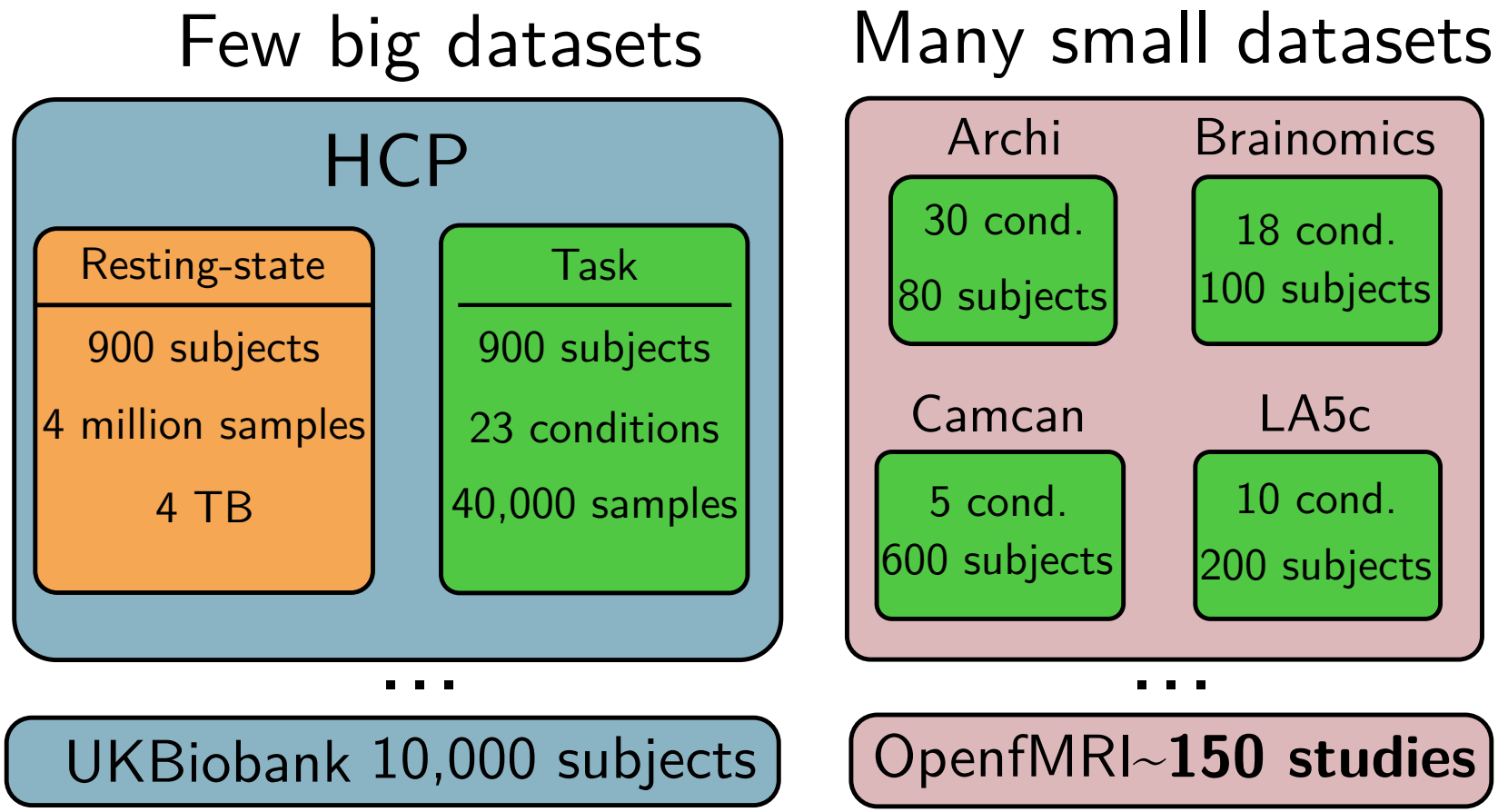
Psycho. condition

A good model should handle:

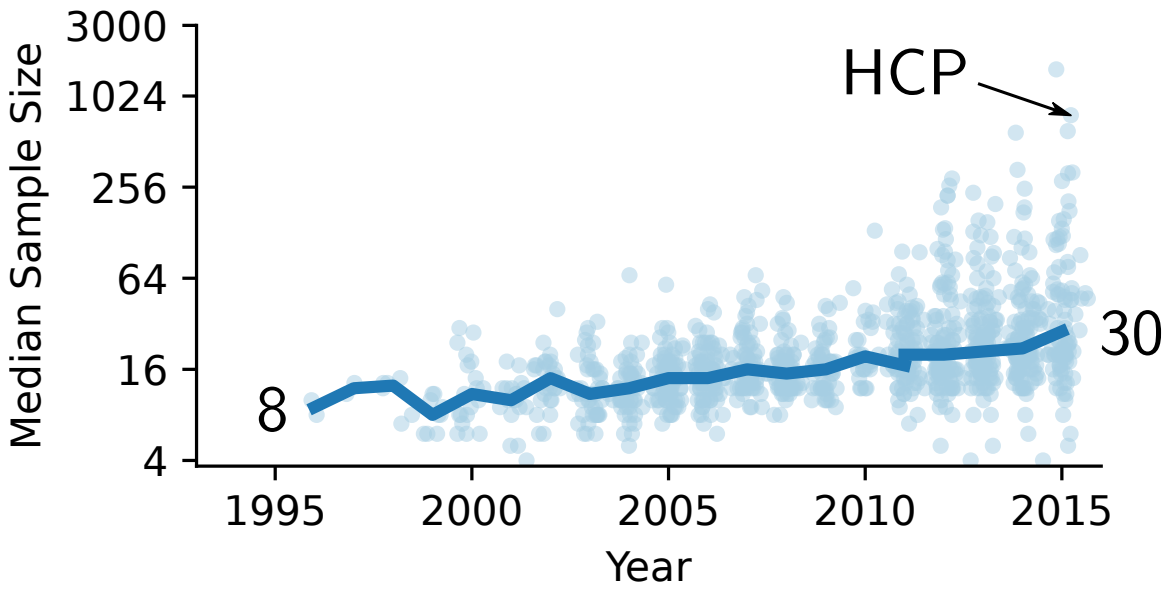
New records
= Inter and intra-subject variability

New but related psychological conditions

Cognitive neuroscience is becoming data-intensive

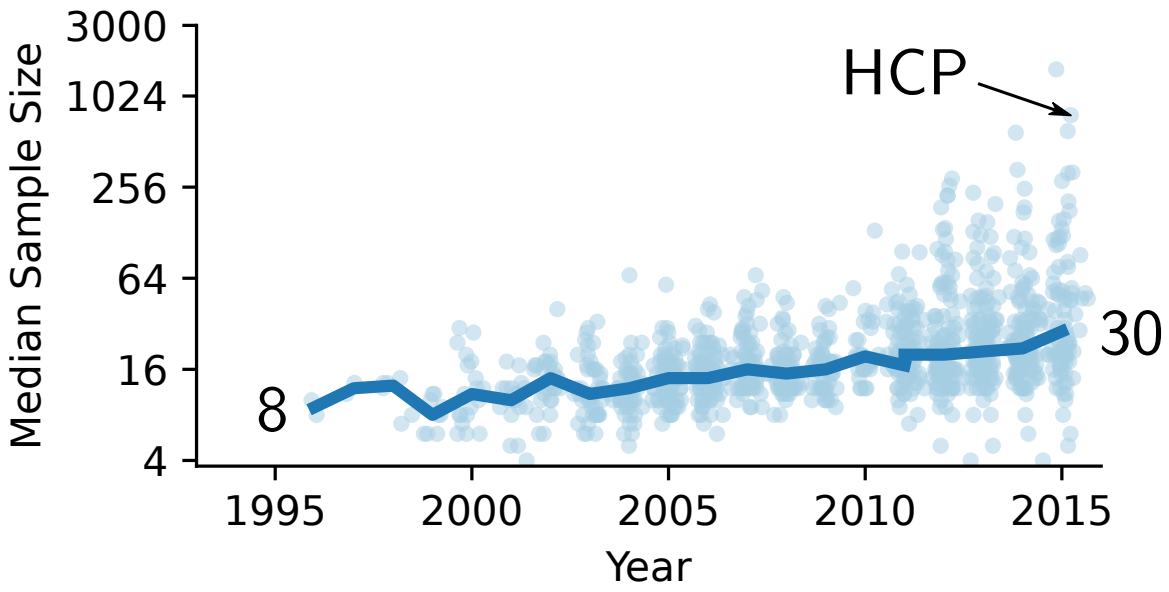


But is plagued by low sample-size issues



Adapted from Poldrack et al. (2016)

But is plagued by low sample-size issues

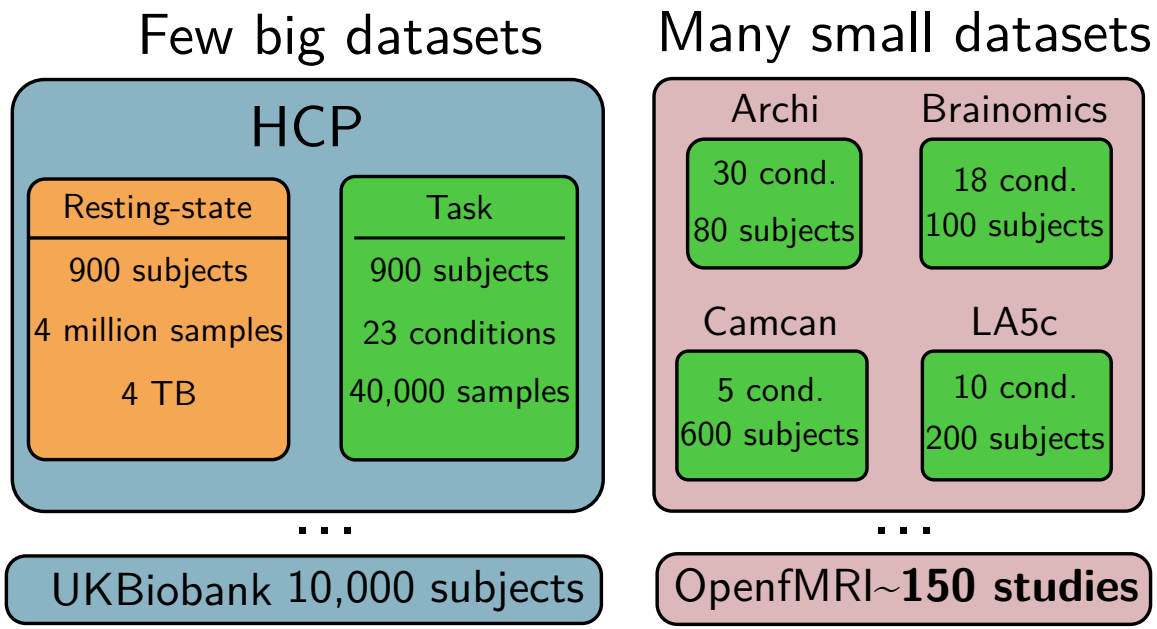


Adapted from Poldrack et al. (2016)

Button, K. S. *et al.* Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nature Reviews Neuroscience* **14,** 365–376 (2013)

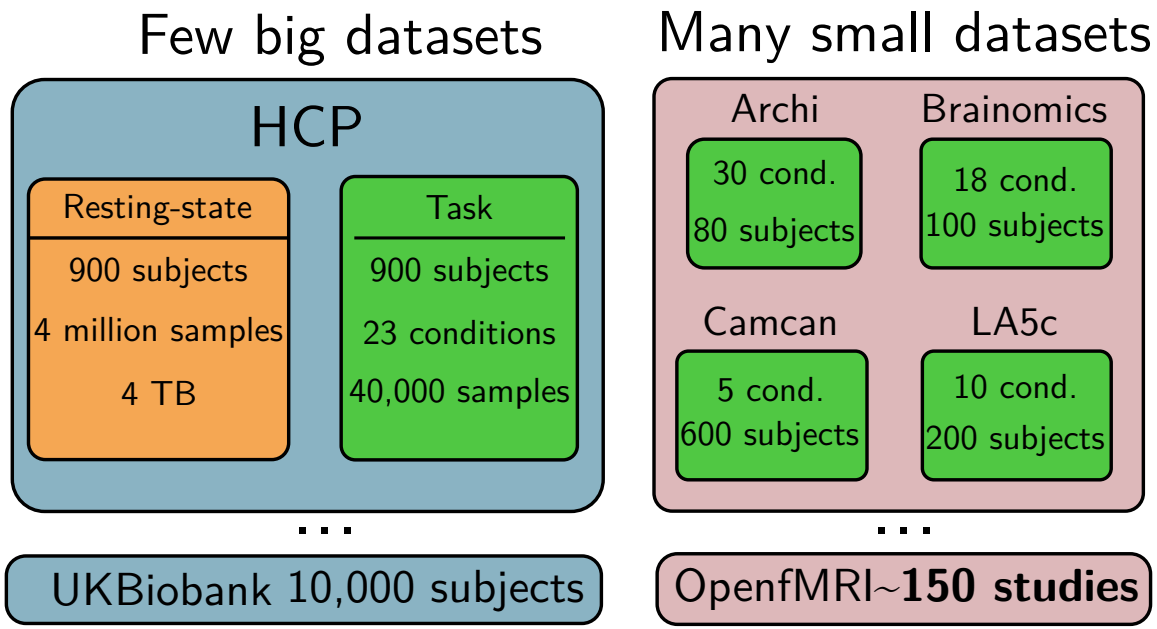
- Encoding small studies (< 30 subj.) $\rightarrow$ low statistical power
- Discoveries are hard to reproduce

New approaches to data intensive fMRI



- Single study analysis: small sample-size, restricted paradigms
- But many studies at hand

New approaches to data intensive fMRI



- Single study analysis: small sample-size, restricted paradigms
- But many studies at hand

Aggregate available data to learn more informative brain representations

Brain representation for predictive modelling

- Brain images are high dimensional $\mathcal{X} \subset \mathbb{R}^{200000}$ (curse)
- Noisy and confounded signal

$\rightarrow$ bad generalization performance $\hat{\boldsymbol{y}} = \boldsymbol{f}_\theta(\boldsymbol{x})$

Brain representation for predictive modelling

- Brain images are high dimensional $\mathcal{X} \subset \mathbb{R}^{200000}$ (curse)
- Noisy and confounded signal

$\rightarrow$ bad generalization performance $\hat{\boldsymbol{y}} = \boldsymbol{f}_\theta(\boldsymbol{x})$

Find lower-dimensional representations of brain images

- Learn $h : \mathcal{X} \rightarrow \mathcal{X}' \subset \mathbb{R}^k$ such that

$$\boldsymbol{x}_r = h(\boldsymbol{x}), \hat{\boldsymbol{y}} = f_\theta(h(\boldsymbol{x}))$$

- Predict from $\{\boldsymbol{x}_r\} \rightarrow$ better generalization

Large-scale representation learning

Presented contributions

1. Learn **localized spatial bases of brain signal** from rest data
   - Terabytes datasets
   - Unsupervised approach
   - **New stochastic algorithms for large matrix factorization**

Large-scale representation learning

Presented contributions

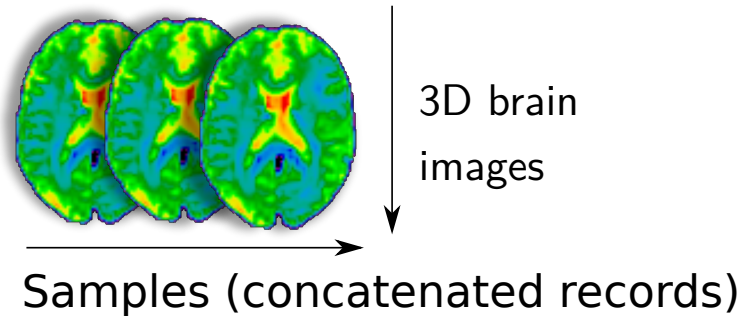
1. Learn **localized spatial bases of brain signal** from rest data
   - Terabytes datasets
   - Unsupervised approach
   - **New stochastic algorithms for large matrix factorization**

2. Learn **spatial bases** of brain signal from **many task studies**
   - Supervised approach: bases aware of psycho. conditions
   - **Deeper interpretable models for multi-study decoding**

Stochastic Subsampling for Massive Matrix Factorization

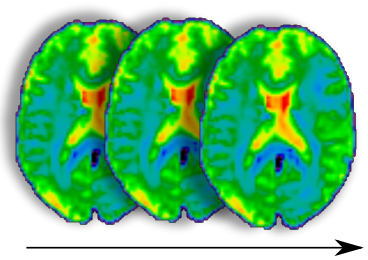
Analyzing resting-state fMRI

Unlabeled data



Analyzing resting-state fMRI

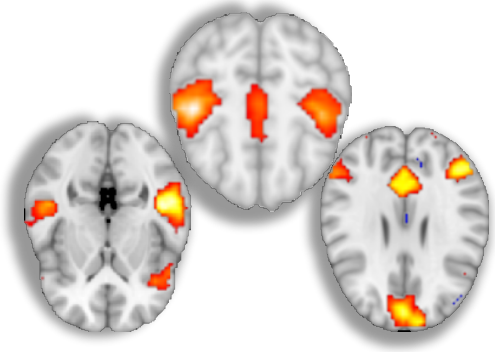
Unlabeled data



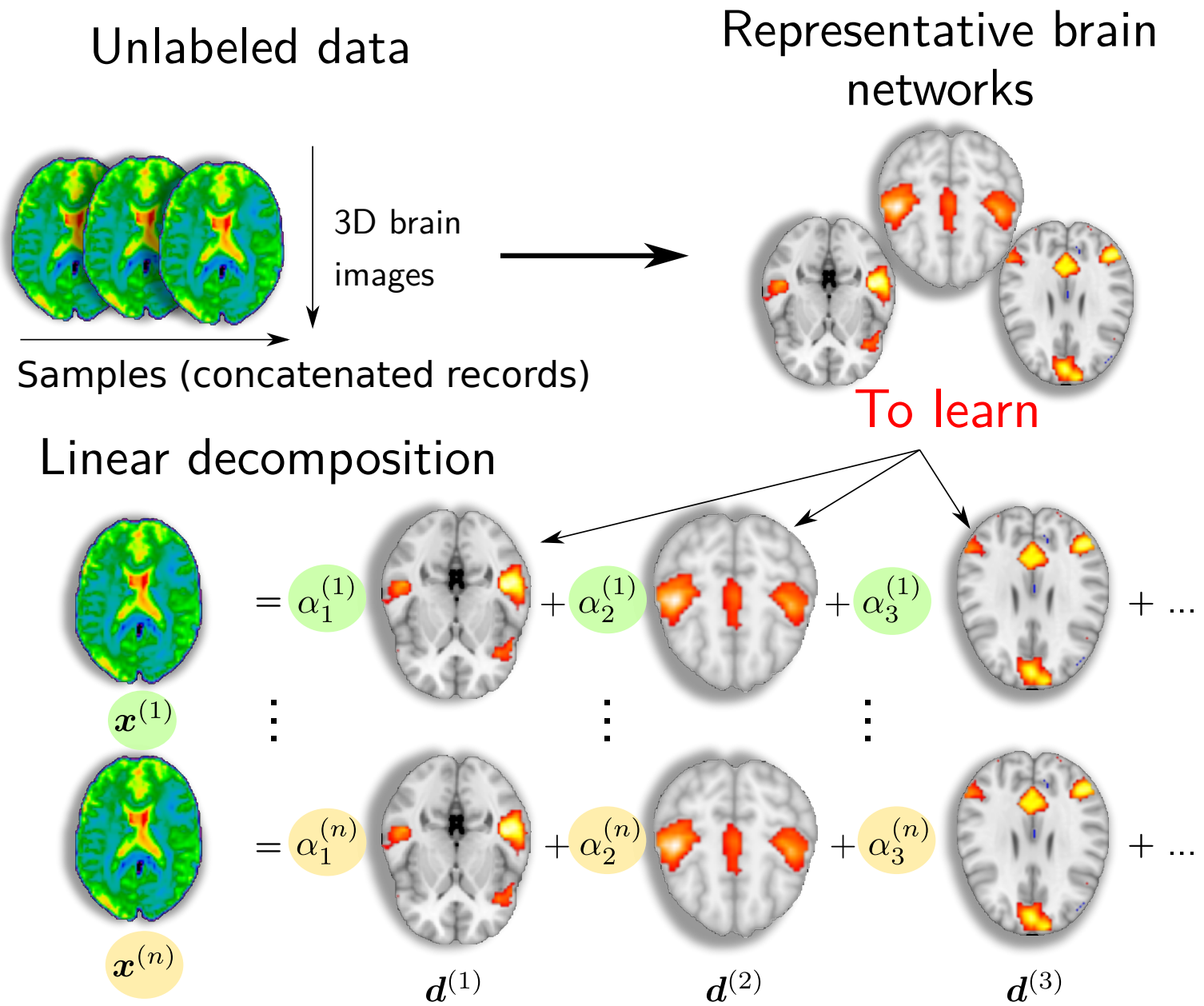
3D brain images

Samples (concatenated records)

Representative brain networks



Analyzing resting-state fMRI

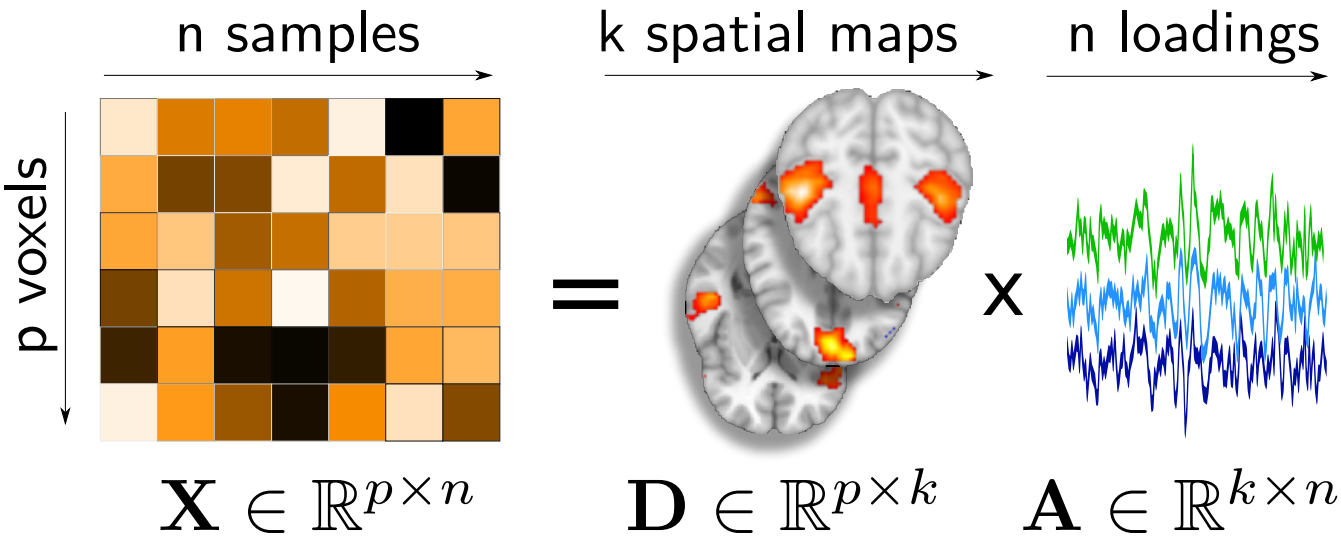


Functional network extraction

Functional networks $\boldsymbol{D}$ = set of non-negative sparse spatial maps

- Correlated brain activity in localized areas
- Known structure: e.g., part of auditory, visual, motor cortex

Matrix form: stack flat 3D brain images into matrix $\boldsymbol{X}$

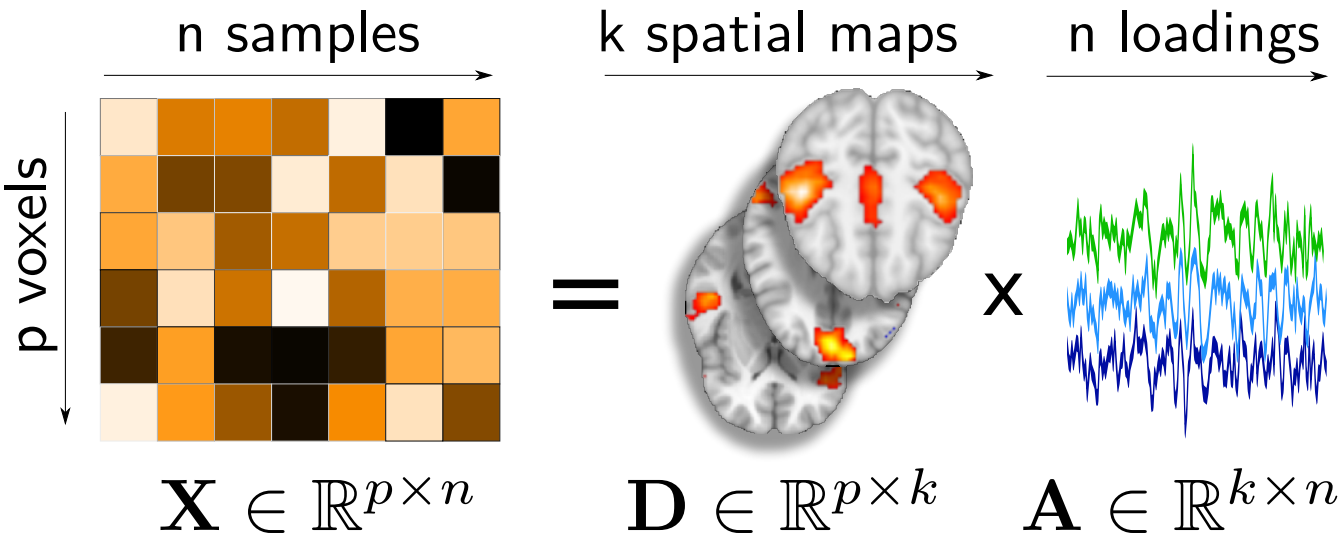


Functional network extraction

Functional networks $\boldsymbol{D}$ = set of non-negative sparse spatial maps

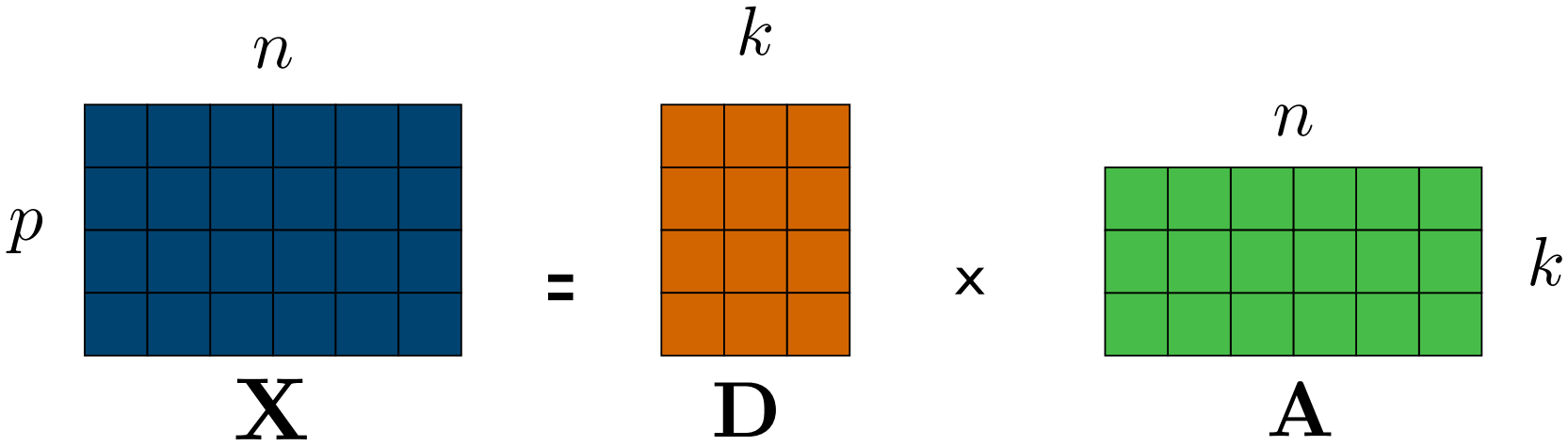
- Correlated brain activity in localized areas
- Known structure: e.g., part of auditory, visual, motor cortex

Matrix form: stack flat 3D brain images into matrix $\boldsymbol{X}$



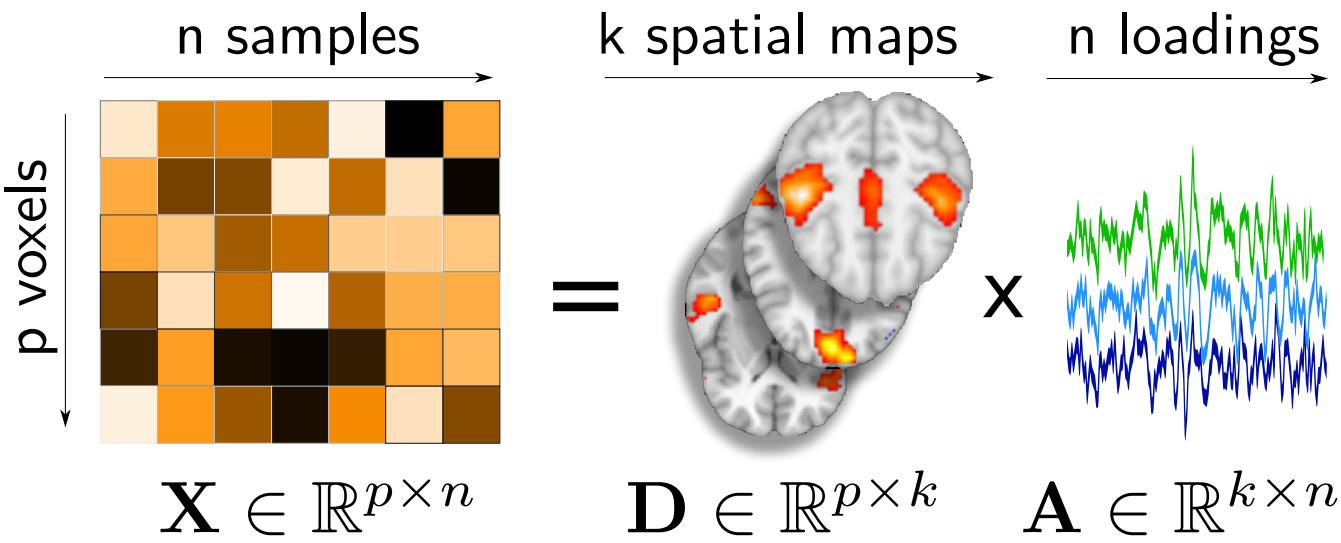
Column $\boldsymbol{\alpha}^{(i)}$ low-dimensional representation of column $\boldsymbol{x}^{(i)}$

Matrix factorization



- $\boldsymbol{X} \in \mathbb{R}^{p \times n} = \boldsymbol{DA} \in \mathbb{R}^{p \times k} \times \mathbb{R}^{k \times n}$
- Flexible tool for unsupervised data analysis
- Lower underlying complexity of $\boldsymbol{X}$

More efficient algorithms



$\boldsymbol{X}$ is large (**4TB**) in both number of samples $n = 3 \cdot 10^6$ and sample/parameter dimension $p = 2 \cdot 10^6$

New stochastic algorithms that scale in **both** directions

Formalism and methods

Matrix factorization: optimisation problem

$$\min_{\boldsymbol{D}\in\mathcal{C},\boldsymbol{A}\in\mathbb{R}^{k\times n}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{A}\|_F^2 + \lambda\Omega(\boldsymbol{A})$$

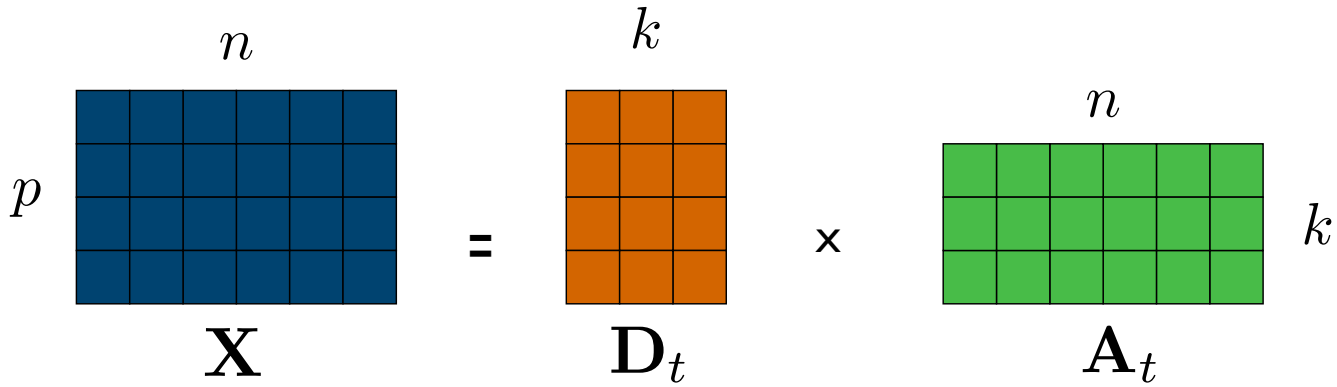
- Bounds on dictionary $\boldsymbol{D}$: sparsity/non-neg. fMRI: $\|\boldsymbol{d}^{(j)}\|_1 < 1$
- Penalty on loadings $\boldsymbol{A}$: sparsity/non-neg. fMRI: $\Omega(\boldsymbol{A}) = \|\boldsymbol{A}\|_F^2$

Formalism and methods

Matrix factorization: optimisation problem

$$\min_{\boldsymbol{D}\in\mathcal{C},\boldsymbol{A}\in\mathbb{R}^{k\times n}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{A}\|_F^2 + \lambda\Omega(\boldsymbol{A})$$

- Bounds on dictionary $\boldsymbol{D}$: sparsity/non-neg. fMRI: $\|\boldsymbol{d}^{(j)}\|_1 < 1$
- Penalty on loadings $\boldsymbol{A}$: sparsity/non-neg. fMRI: $\Omega(\boldsymbol{A}) = \|\boldsymbol{A}\|_F^2$

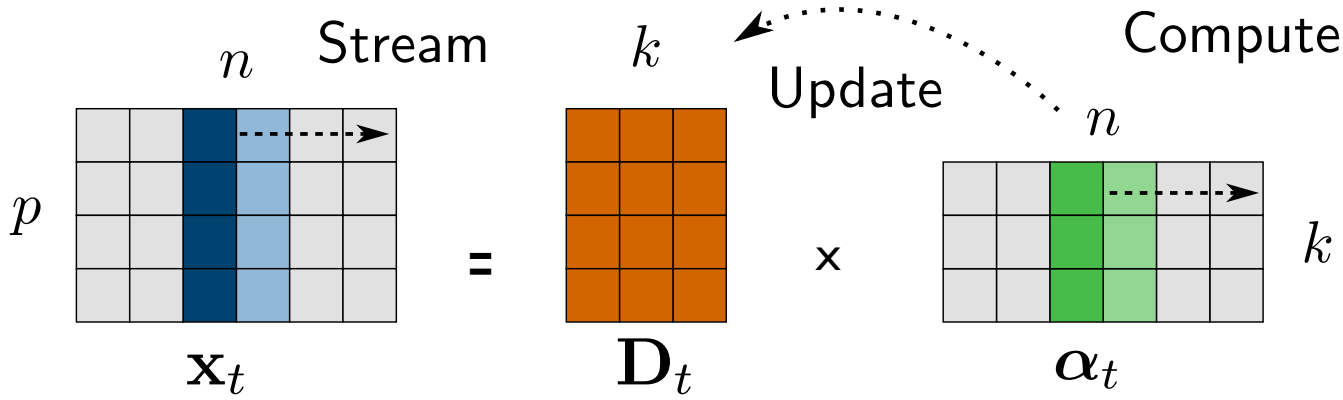


- Alternated minimization: use full $\boldsymbol{X}$ at each iteration
- **Slow:** single iteration cost in $\mathcal{O}(np)$ – min 1 hour to load 4TB

Online matrix factorization (Mairal *et al.* 2010)

Scaling in the number of sample n:

- Stream columns $(\boldsymbol{x}_t)_t$ and update $(\boldsymbol{D}_t)$ at each t
- Single iteration cost in $\mathcal{O}(p)$ — noisy update
- Convergence in a few epochs $\rightarrow$ large speed-up



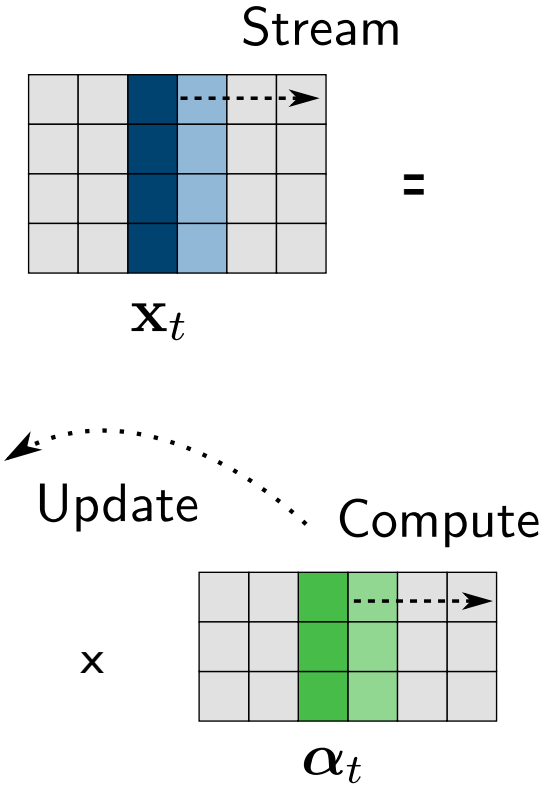
Use-cases:

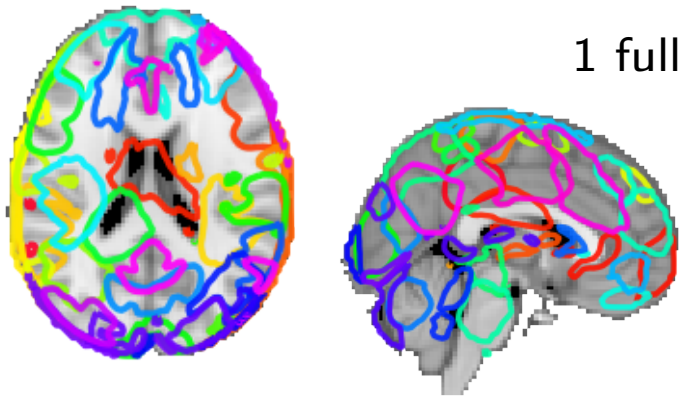
- Large n, regular p, *e.g.*, image patches:

$$p = 256 \qquad n \approx 10^6 \qquad \textbf{1GB}$$

Scaling-up for massive matrices

Out-of-the-box online algorithm ?



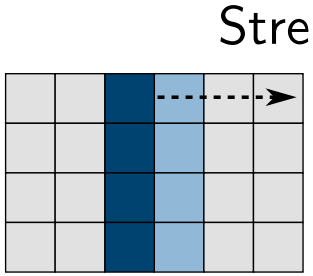


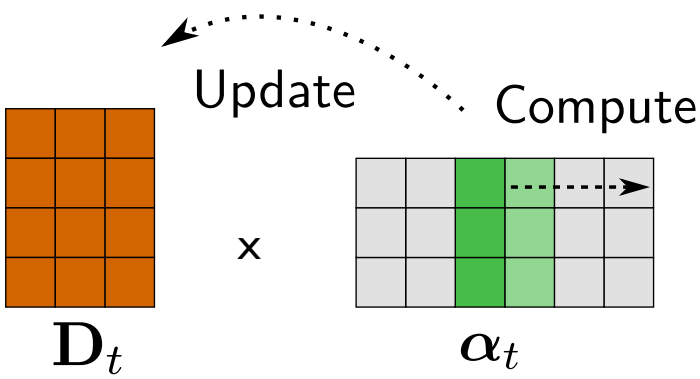
1 full epoch

235 h run time

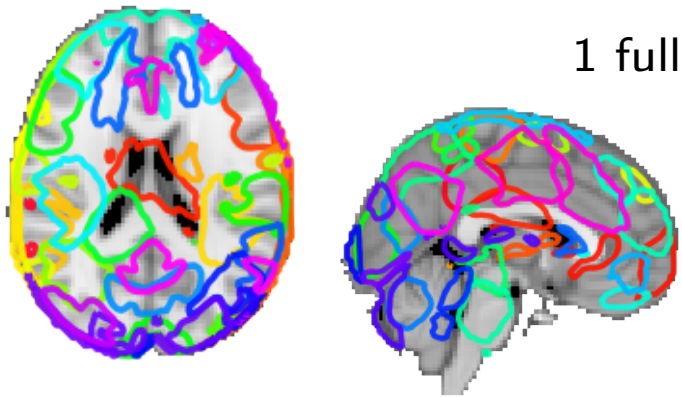
Scaling-up for massive matrices

Out-of-the-box online algorithm ?

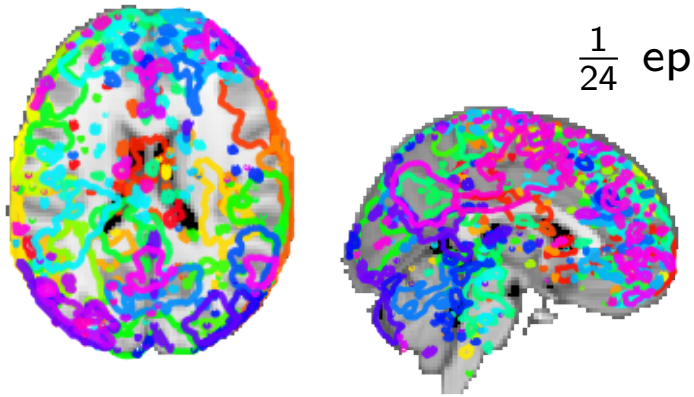




Limited time budget ?



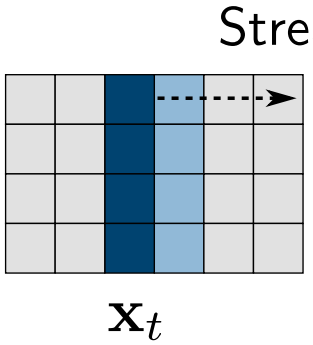
235 h run time

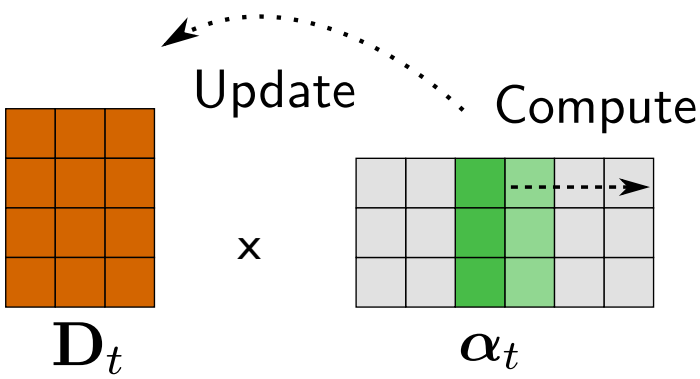


10 h run time

Scaling-up for massive matrices

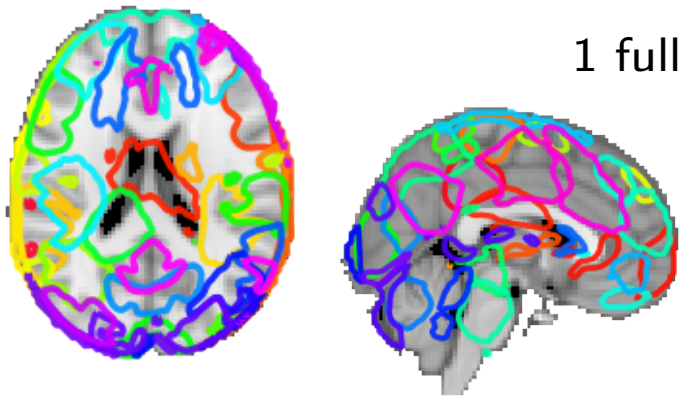
Out-of-the-box online algorithm ?



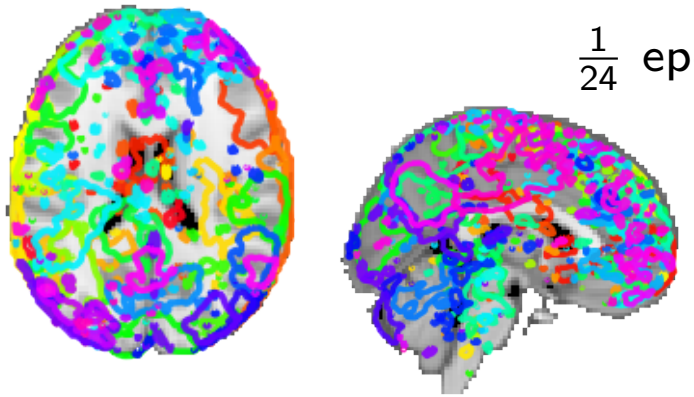


Limited time budget ?

Need to accomodate large p

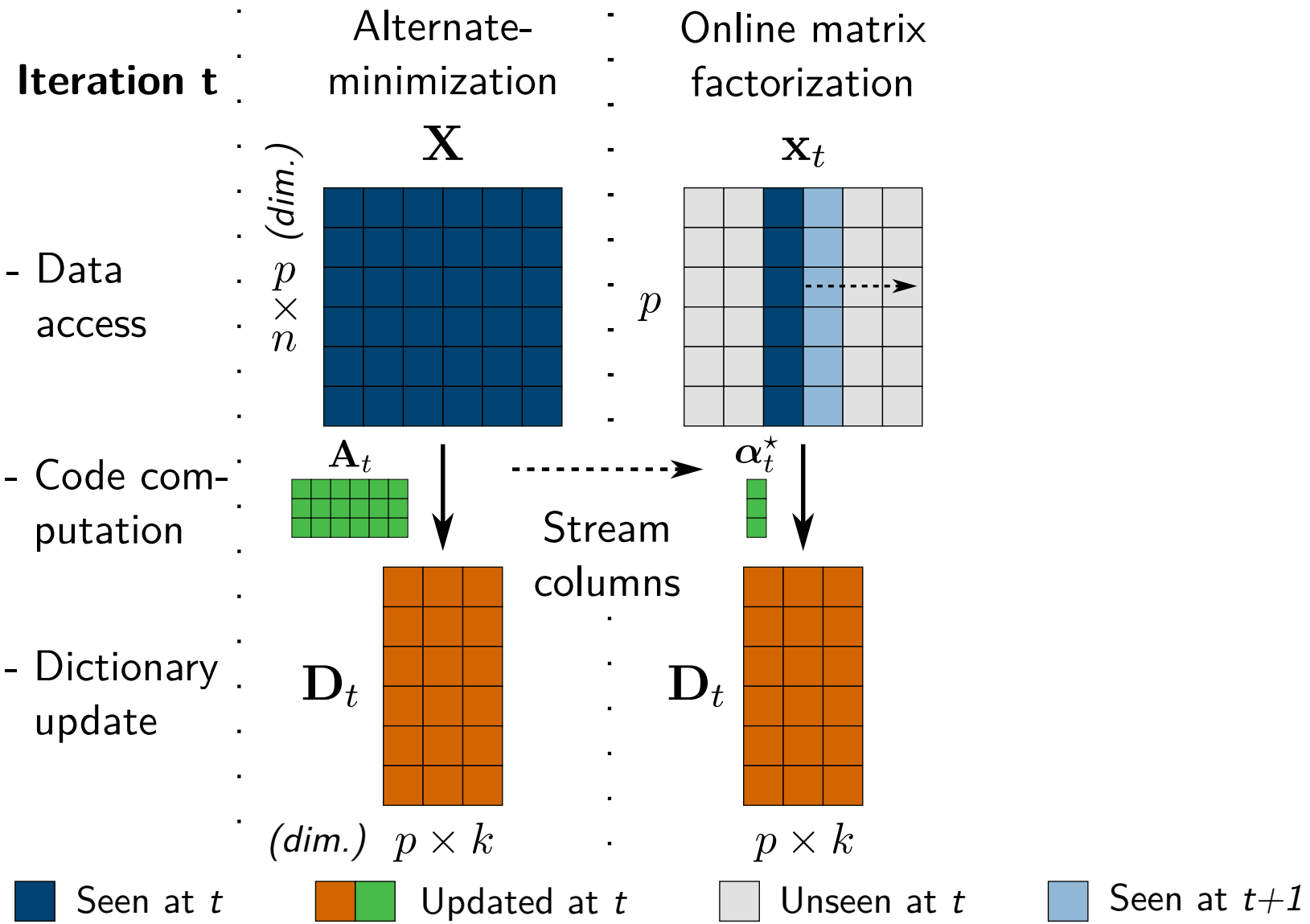


235 h run time

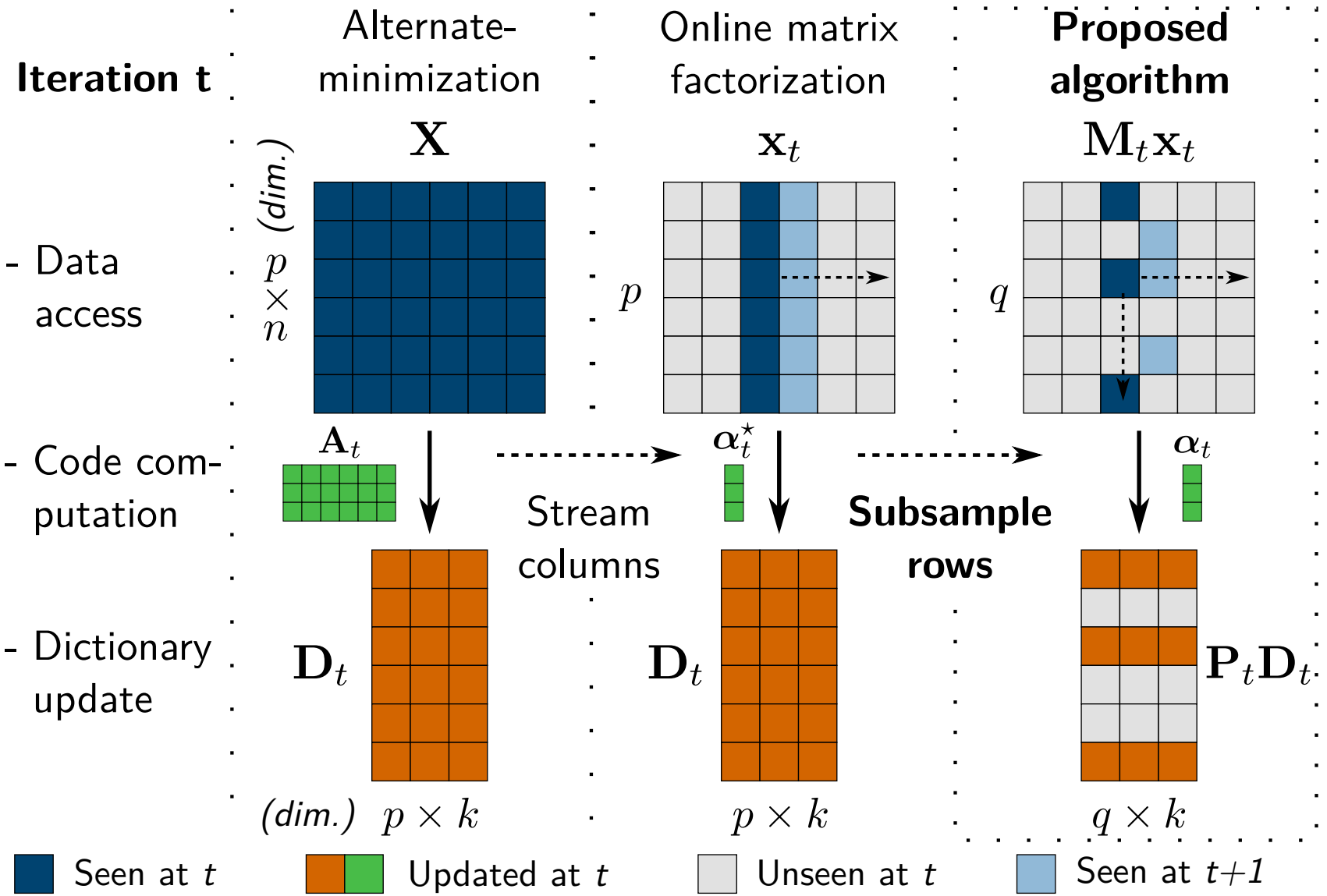


10 h run time

Scaling-up in both directions



Scaling-up in both directions



Online matrix factorization (Mairal *et al.* 2010)

Original problem:

$$\min_{\boldsymbol{D}\in\mathcal{C},\boldsymbol{A}\in\mathbb{R}^{k\times n}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{A}\|_F^2+\lambda\Omega(\boldsymbol{A})$$

We learn the **left side factor**: dictionary $\boldsymbol{D}^\star$ solution of

$$\min_{\boldsymbol{D}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}^{(i)}-\boldsymbol{D}\boldsymbol{\alpha}(\boldsymbol{x}^{(i)},\boldsymbol{D})\|_2^2$$

$$\boldsymbol{\alpha}(\boldsymbol{x}^{(i)},\boldsymbol{D})=\frac{1}{2}\operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^k}\|\boldsymbol{x}^{(i)}-\boldsymbol{D}\boldsymbol{\alpha}\|_2^2+\lambda\Omega(\boldsymbol{\alpha})$$

Online matrix factorization (Mairal *et al.* 2010)

Original problem:

$$\min_{\boldsymbol{D}\in\mathcal{C},\boldsymbol{A}\in\mathbb{R}^{k\times n}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{A}\|_F^2+\lambda\Omega(\boldsymbol{A})$$

We learn the **left side factor**: dictionary $\boldsymbol{D}^\star$ solution of

$$\min_{\boldsymbol{D}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}^{(i)}-\boldsymbol{D}\boldsymbol{\alpha}(\boldsymbol{x}^{(i)},\boldsymbol{D})\|_2^2$$

$$\boldsymbol{\alpha}(\boldsymbol{x}^{(i)},\boldsymbol{D})=\frac{1}{2}\operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^k}\|\boldsymbol{x}^{(i)}-\boldsymbol{D}\boldsymbol{\alpha}\|_2^2+\lambda\Omega(\boldsymbol{\alpha})$$

Expected risk minimization: $(i_t)_t\sim\mathcal{U}([1,n])\rightarrow\boldsymbol{x}_t=\boldsymbol{x}^{(i_t)}\sim\mathcal{D}$

$$\min_{\boldsymbol{D}\in\mathcal{C}}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[f(\boldsymbol{D},\boldsymbol{x})]\qquad f(\boldsymbol{D},\boldsymbol{x})\triangleq\|\boldsymbol{x}-\boldsymbol{D}\boldsymbol{\alpha}(\boldsymbol{x},\boldsymbol{D})\|_2^2$$

Online matrix factorization (Mairal *et al.* 2010)

Original problem:

$$\min_{\boldsymbol{D}\in\mathcal{C},\boldsymbol{A}\in\mathbb{R}^{k\times n}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{A}\|_F^2+\lambda\Omega(\boldsymbol{A})$$

We learn the **left side factor**: dictionary $\boldsymbol{D}^\star$ solution of

$$\min_{\boldsymbol{D}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}^{(i)}-\boldsymbol{D}\boldsymbol{\alpha}(\boldsymbol{x}^{(i)},\boldsymbol{D})\|_2^2$$

$$\boldsymbol{\alpha}(\boldsymbol{x}^{(i)},\boldsymbol{D})=\frac{1}{2}\operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^k}\|\boldsymbol{x}^{(i)}-\boldsymbol{D}\boldsymbol{\alpha}\|_2^2+\lambda\Omega(\boldsymbol{\alpha})$$

Expected risk minimization: $(i_t)_t\sim\mathcal{U}([1,n])\rightarrow\boldsymbol{x}_t=\boldsymbol{x}^{(i_t)}\sim\mathcal{D}$

$$\min_{\boldsymbol{D}\in\mathcal{C}}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[f(\boldsymbol{D},\boldsymbol{x})]\qquad f(\boldsymbol{D},\boldsymbol{x})\triangleq\|\boldsymbol{x}-\boldsymbol{D}\boldsymbol{\alpha}(\boldsymbol{x},\boldsymbol{D})\|_2^2$$

How to minimize exp. risk using stream $(x_t)_t$?

Online matrix factorization (Mairal *et al.* 2010)

Can we build an iterate sequence $(\boldsymbol{D}_t)$ from $(\boldsymbol{x}_t)_t$? Ideally:

$$\boldsymbol{D}_t = \underset{\boldsymbol{D}\in\mathcal{C}}{\operatorname{argmin}}\ \bar{f}_t(\boldsymbol{D}) = \frac{1}{t}\sum_{s=1}^{t} f_t(\boldsymbol{D}) \qquad f_t(\boldsymbol{D}) \triangleq \|\boldsymbol{x}_t - \boldsymbol{D}\boldsymbol{\alpha}(\boldsymbol{x}_t, \boldsymbol{D})\|_2^2$$

Online matrix factorization (Mairal *et al.* 2010)

Can we build an iterate sequence $(\boldsymbol{D}_t)$ from $(\boldsymbol{x}_t)_t$? Ideally:

$$\boldsymbol{D}_t = \underset{\boldsymbol{D}\in\mathcal{C}}{\operatorname{argmin}}\ \bar{f}_t(\boldsymbol{D}) = \frac{1}{t}\sum_{s=1}^{t} f_t(\boldsymbol{D}) \qquad f_t(\boldsymbol{D}) \triangleq \|\boldsymbol{x}_t - \boldsymbol{D}\boldsymbol{\alpha}(\boldsymbol{x}_t, \boldsymbol{D})\|_2^2$$

Replace $\boldsymbol{\alpha}(\boldsymbol{x}_t, \boldsymbol{D})$ by $\boldsymbol{\alpha}_t$ computed on previous dictionary $\boldsymbol{D}_{t-1}$

$$\boldsymbol{\alpha}(\boldsymbol{x}_t, \boldsymbol{D}) = \underset{\boldsymbol{\alpha}\in\mathbb{R}^k}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 + \lambda\Omega(\boldsymbol{\alpha})$$

$$\boldsymbol{\alpha}_t = \underset{\boldsymbol{\alpha}\in\mathbb{R}^k}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{D}_{t-1}\boldsymbol{\alpha}\|_2^2 + \lambda\Omega(\boldsymbol{\alpha})$$

Online matrix factorization (Mairal *et al.* 2010)

Can we build an iterate sequence $(\boldsymbol{D}_t)$ from $(\boldsymbol{x}_t)_t$? Ideally:

$$\boldsymbol{D}_t = \underset{\boldsymbol{D}\in\mathcal{C}}{\operatorname{argmin}}\ \bar{f}_t(\boldsymbol{D}) = \frac{1}{t}\sum_{s=1}^{t} f_t(\boldsymbol{D}) \qquad f_t(\boldsymbol{D}) \triangleq \|\boldsymbol{x}_t - \boldsymbol{D}\boldsymbol{\alpha}(\boldsymbol{x}_t, \boldsymbol{D})\|_2^2$$

Replace $\boldsymbol{\alpha}(\boldsymbol{x}_t, \boldsymbol{D})$ by $\boldsymbol{\alpha}_t$ computed on previous dictionary $\boldsymbol{D}_{t-1}$

$$\boldsymbol{\alpha}(\boldsymbol{x}_t, \boldsymbol{D}) = \underset{\boldsymbol{\alpha}\in\mathbb{R}^k}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 + \lambda\Omega(\boldsymbol{\alpha})$$

$$\boldsymbol{\alpha}_t = \underset{\boldsymbol{\alpha}\in\mathbb{R}^k}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{D}_{t-1}\boldsymbol{\alpha}\|_2^2 + \lambda\Omega(\boldsymbol{\alpha})$$

Replace f_t by a non-variational **surrogate** function g_t

$$g_t(\boldsymbol{D}) \triangleq \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{D}\boldsymbol{\alpha}_t\|_2^2 \qquad \boldsymbol{D}_t = \underset{\boldsymbol{D}\in\mathcal{C}}{\operatorname{argmin}}\ \bar{g}_t(\boldsymbol{D}) = \frac{1}{t}\sum_{s=1}^{t} g_t(\boldsymbol{D})$$

Computations at iteration t

1. **Compute code**

$$\boldsymbol{\alpha}_t = \underset{\boldsymbol{\alpha} \in \mathbb{R}^k}{\operatorname{argmin}} \left\| \boldsymbol{x}_t - \boldsymbol{D}_{t-1}\boldsymbol{\alpha} \right\|_2^2 + \lambda \Omega(\boldsymbol{\alpha}_t)$$

2. **Update surrogate**

$$\bar{g}_t(\boldsymbol{D}) = \frac{1}{t} \sum_{s=1}^{t} \left\| \boldsymbol{x}_s - \boldsymbol{D}\boldsymbol{\alpha}_s \right\|_2^2$$

3. **Minimize surrogate**

$$\boldsymbol{D}_t = \underset{\boldsymbol{D} \in \mathcal{C}}{\operatorname{argmin}}\, \bar{g}_t(\boldsymbol{D})$$

Computations at iteration t

1. **Compute code** – $\mathcal{O}(p)$

$$\boldsymbol{\alpha}_t = \underset{\boldsymbol{\alpha} \in \mathbb{R}^k}{\operatorname{argmin}} \left\| \boldsymbol{x}_t - \boldsymbol{D}_{t-1} \boldsymbol{\alpha} \right\|_2^2 + \lambda \Omega(\boldsymbol{\alpha}_t)$$

2. **Update surrogate** – $\mathcal{O}(p)$

$$\bar{g}_t(\boldsymbol{D}) = \frac{1}{t} \sum_{s=1}^{t} \left\| \boldsymbol{x}_s - \boldsymbol{D} \boldsymbol{\alpha}_s \right\|_2^2$$

3. **Minimize surrogate** – $\mathcal{O}(p)$

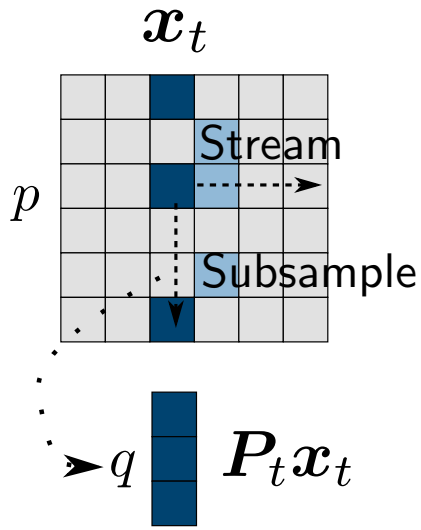
$$\boldsymbol{D}_t = \underset{\boldsymbol{D} \in \mathcal{C}}{\operatorname{argmin}} \, \bar{g}_t(\boldsymbol{D})$$

Access to $\boldsymbol{x}_t$ and $\boldsymbol{D} \rightarrow$ Algorithm in $\mathcal{O}(p)$

Stochastic subsampling

How to reduce single iteration cost $\mathcal{O}(p)$?

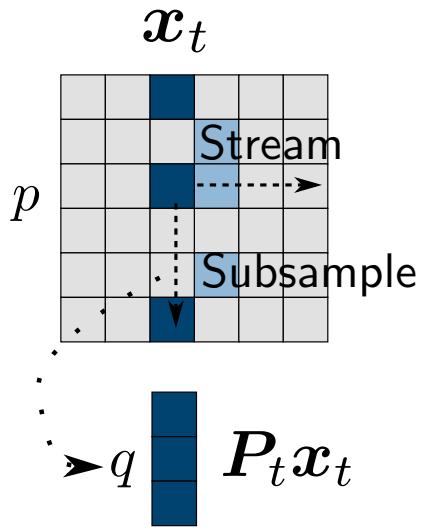
- Access a subsampled version of $\boldsymbol{x}_t$
- Subsampling projector $\boldsymbol{P}_t : \mathbb{R}^p \to \mathbb{R}^q$



Stochastic subsampling

How to reduce single iteration cost $\mathcal{O}(p)$?

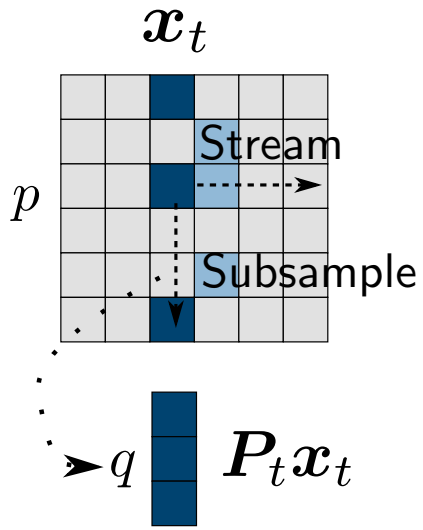
- Access a subsampled version of $\boldsymbol{x}_t$
- Subsampling projector $\boldsymbol{P}_t : \mathbb{R}^p \to \mathbb{R}^q$
- Use only $\boldsymbol{P}_t \boldsymbol{x}_t$ in computations
- $\mathcal{O}(q)$ iteration complexity



Stochastic subsampling

How to reduce single iteration cost $\mathcal{O}(p)$?

- Access a subsampled version of $\boldsymbol{x}_t$
- Subsampling projector $\boldsymbol{P}_t : \mathbb{R}^p \to \mathbb{R}^q$
- Use only $\boldsymbol{P}_t \boldsymbol{x}_t$ in computations
- $\mathcal{O}(q)$ iteration complexity



Subsampled Online Matrix Factorization (SOMF):

1. **Code computation**
2. **Surrogate update**
3. **Surrogate minimization**

Stochastic approximations

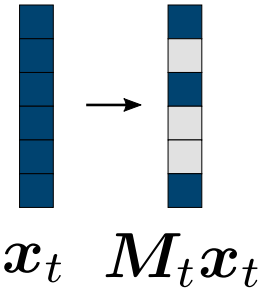
Approx. surrogate g_t: Linear regression with random sampling

$$\boldsymbol{\alpha}_t = \underset{\boldsymbol{\alpha} \in \mathbb{R}^k}{\operatorname{argmin}} \left\| \boldsymbol{x}_t - \boldsymbol{D}_{t-1}\boldsymbol{\alpha} \right\|_2^2 + \lambda \Omega(\boldsymbol{\alpha})$$

Stochastic approximations

Approx. surrogate g_t: Linear regression with random sampling

$$\boldsymbol{\alpha}_t = \underset{\boldsymbol{\alpha} \in \mathbb{R}^k}{\text{argmin}} \, \| \boldsymbol{M}_t (\boldsymbol{x}_t - \boldsymbol{D}_{t-1} \boldsymbol{\alpha}) \|_2^2 + \lambda \Omega(\boldsymbol{\alpha})$$



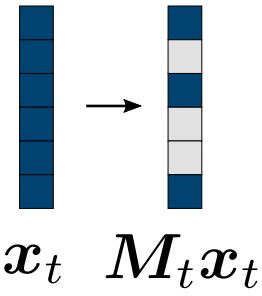
$\boldsymbol{M}_t$ diagonal masking matrix $\leftrightarrow \boldsymbol{P}_t$, $\mathbb{E}[\boldsymbol{M}_t] = \boldsymbol{I}_p$.

Defines $g_t(\boldsymbol{D}) = \| \boldsymbol{x}_t - \boldsymbol{D} \boldsymbol{\alpha}_t \|_2^2$ and averaged surrogate $\bar{g}_t$.

Stochastic approximations

Approx. surrogate g_t**:** Linear regression with random sampling

$$\boldsymbol{\alpha}_t = \underset{\boldsymbol{\alpha}\in\mathbb{R}^k}{\operatorname{argmin}} \|\boldsymbol{M}_t(\boldsymbol{x}_t - \boldsymbol{D}_{t-1}\boldsymbol{\alpha})\|_2^2 + \lambda\Omega(\boldsymbol{\alpha})$$

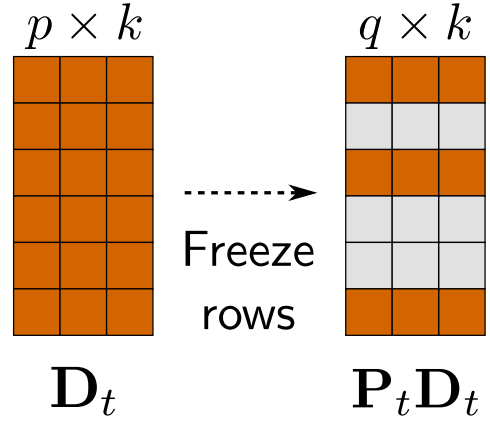


$\boldsymbol{M}_t$ diagonal masking matrix $\leftrightarrow$ $\boldsymbol{P}_t$, $\mathbb{E}[\boldsymbol{M}_t] = \boldsymbol{I}_p$.

Defines $g_t(\boldsymbol{D}) = \|\boldsymbol{x}_t - \boldsymbol{D}\boldsymbol{\alpha}_t\|_2^2$ and averaged surrogate $\bar{g}_t$.

Partial minimisation: Freeze the rows not selected by $\boldsymbol{M}_t$

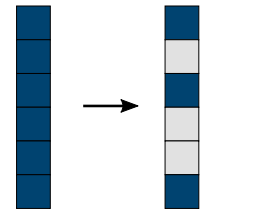
$$\boldsymbol{D}_t = \underset{\boldsymbol{D}\in\mathcal{C}}{\operatorname{argmin}}\, \bar{g}_t(\boldsymbol{D})$$



Stochastic approximations

Approx. surrogate g_t: Linear regression with random sampling

$$\boldsymbol{\alpha}_t = \underset{\boldsymbol{\alpha}\in\mathbb{R}^k}{\operatorname{argmin}} \|\boldsymbol{M}_t(\boldsymbol{x}_t - \boldsymbol{D}_{t-1}\boldsymbol{\alpha})\|_2^2 + \lambda\Omega(\boldsymbol{\alpha})$$



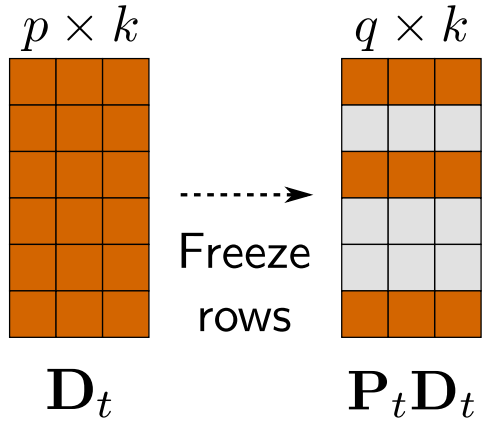
$\boldsymbol{M}_t$ diagonal masking matrix $\leftrightarrow \boldsymbol{P}_t$, $\mathbb{E}[\boldsymbol{M}_t] = \boldsymbol{I}_p$.

Defines $g_t(\boldsymbol{D}) = \|\boldsymbol{x}_t - \boldsymbol{D}\boldsymbol{\alpha}_t\|_2^2$ and averaged surrogate $\bar{g}_t$.

Partial minimisation: Freeze the rows not selected by $\boldsymbol{M}_t$

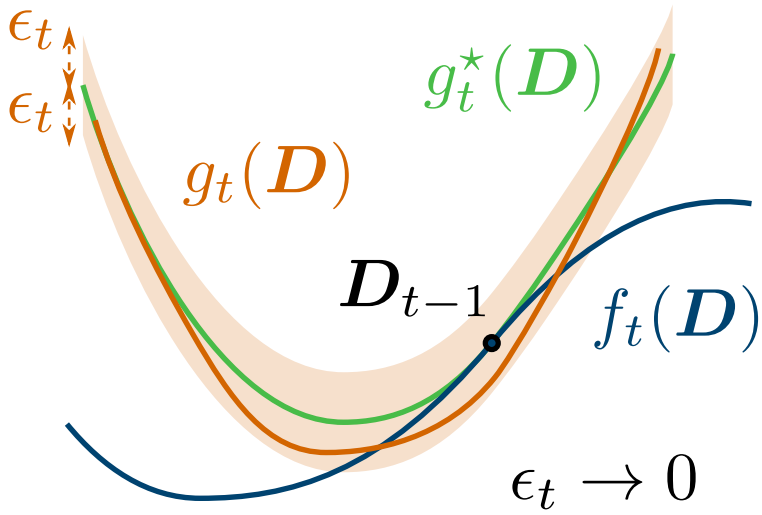
$$\boldsymbol{D}_t = \underset{\substack{\boldsymbol{D}\in\mathcal{C} \\ \boldsymbol{P}_t^{\perp}\boldsymbol{D}=\boldsymbol{P}_t^{\perp}\boldsymbol{D}_{t-1}}}{\operatorname{argmin}} \bar{g}_t(\boldsymbol{D})$$

Block coordinate descent in $\mathbb{R}^{q\times k}$



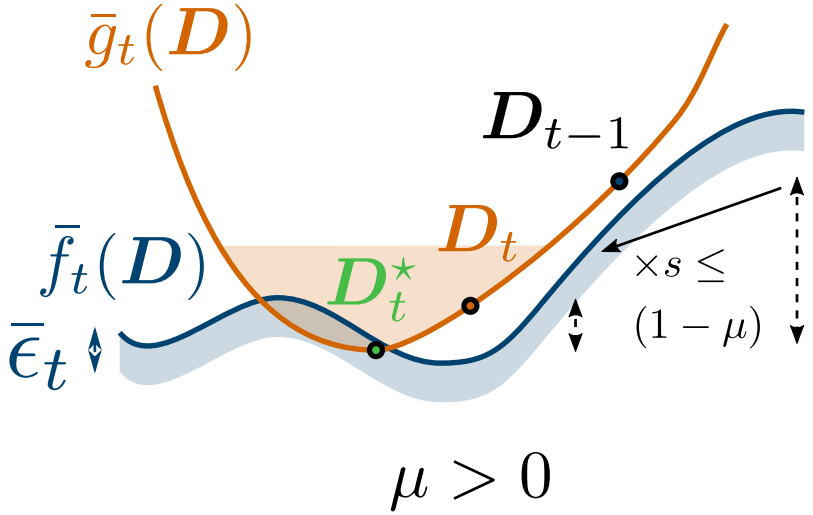
Theoretical analysis

Surrogate approximation



$g_t^\star$ majorizing surrogate of f_t
g_t approx. maj. surrogate

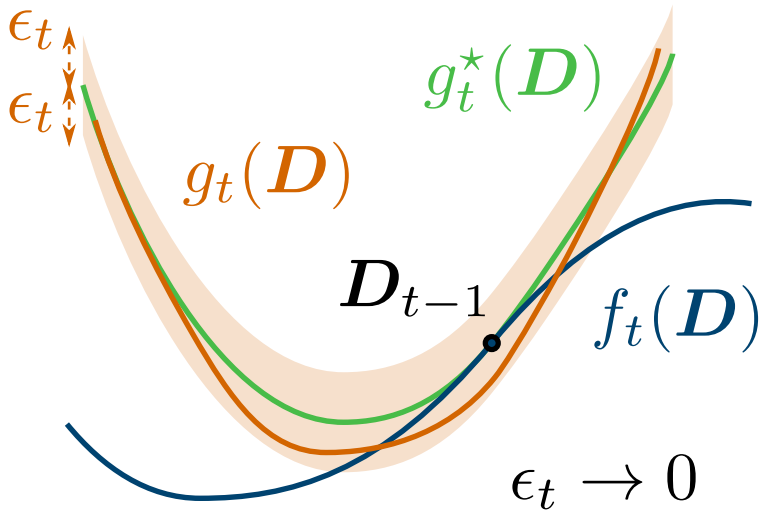
Partial minimization



Linear suboptimality decrease

Theoretical analysis

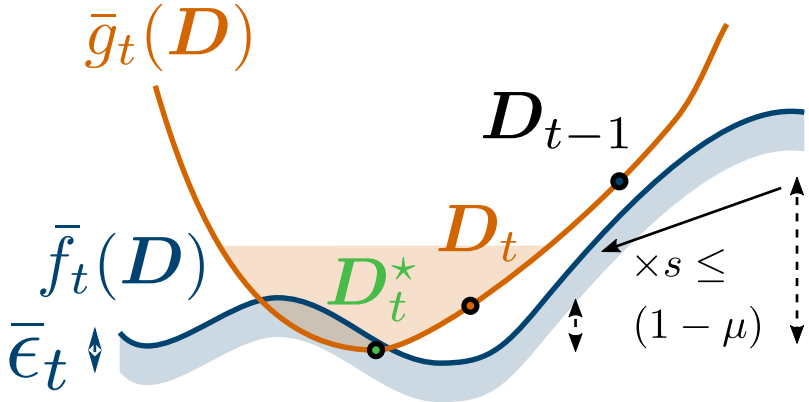
Surrogate approximation



$\epsilon_t \to 0$

$g_t^\star$ majorizing surrogate of f_t
g_t approx. maj. surrogate

Partial minimization



$\mu > 0$

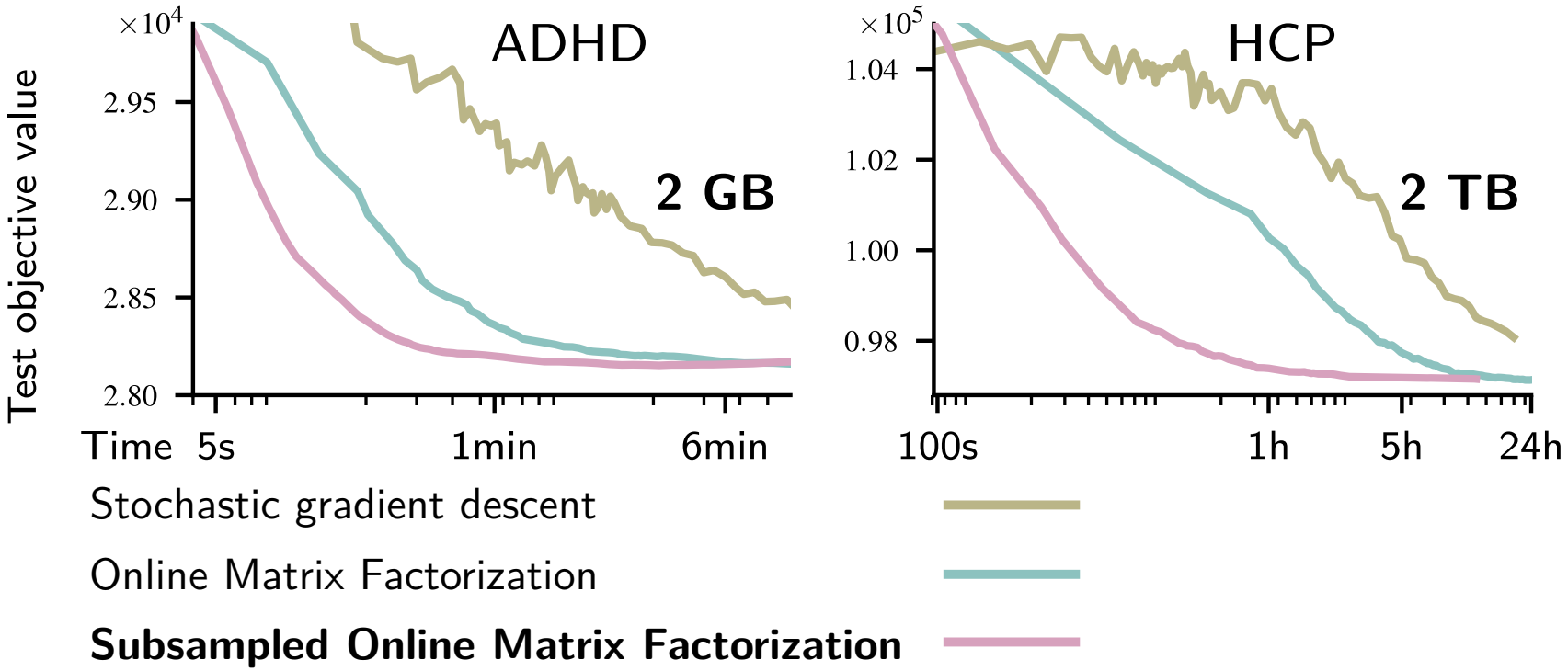
Linear suboptimality decrease

Convergence theorem (informal)

$\bar{f}(\boldsymbol{D}_t)$ converges with probability one and every limit point $\boldsymbol{D}_\infty$ of $(\boldsymbol{D}_t)_t$ is a stationary point of $\bar{f}$, *near local minimum* for all $\boldsymbol{D} \in \mathcal{C}$

$$\nabla \bar{f}(\boldsymbol{D}_\infty, \boldsymbol{D} - \boldsymbol{D}_\infty) \geq 0$$

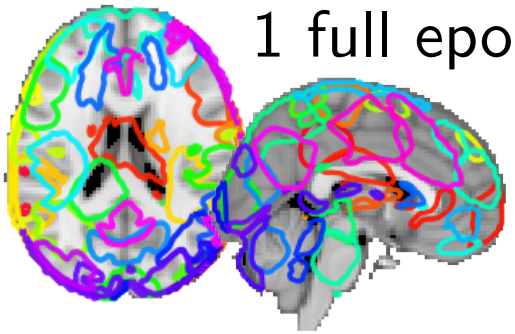
Results: up to 12x speed-up



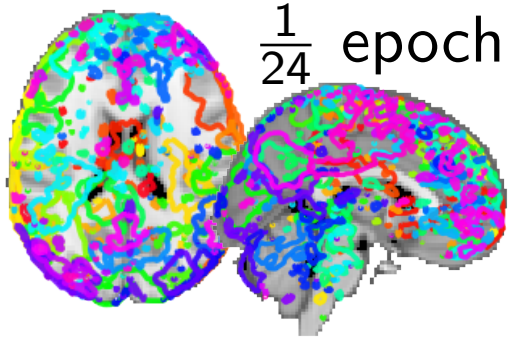
Very high gain on HCP, also useful on smaller problems

Resting-state fMRI

Online matrix factorization

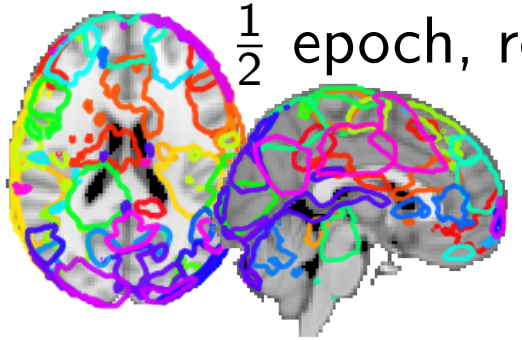


235 h run time



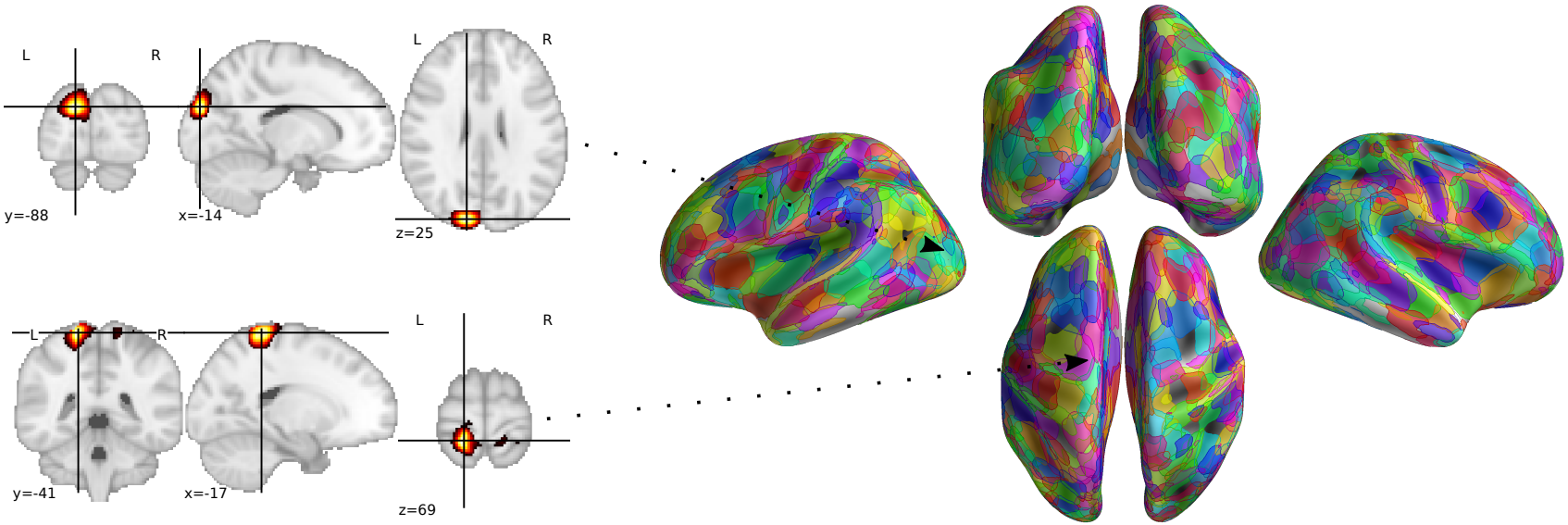
10 h run time

Subsampled Online Matrix Factorisation



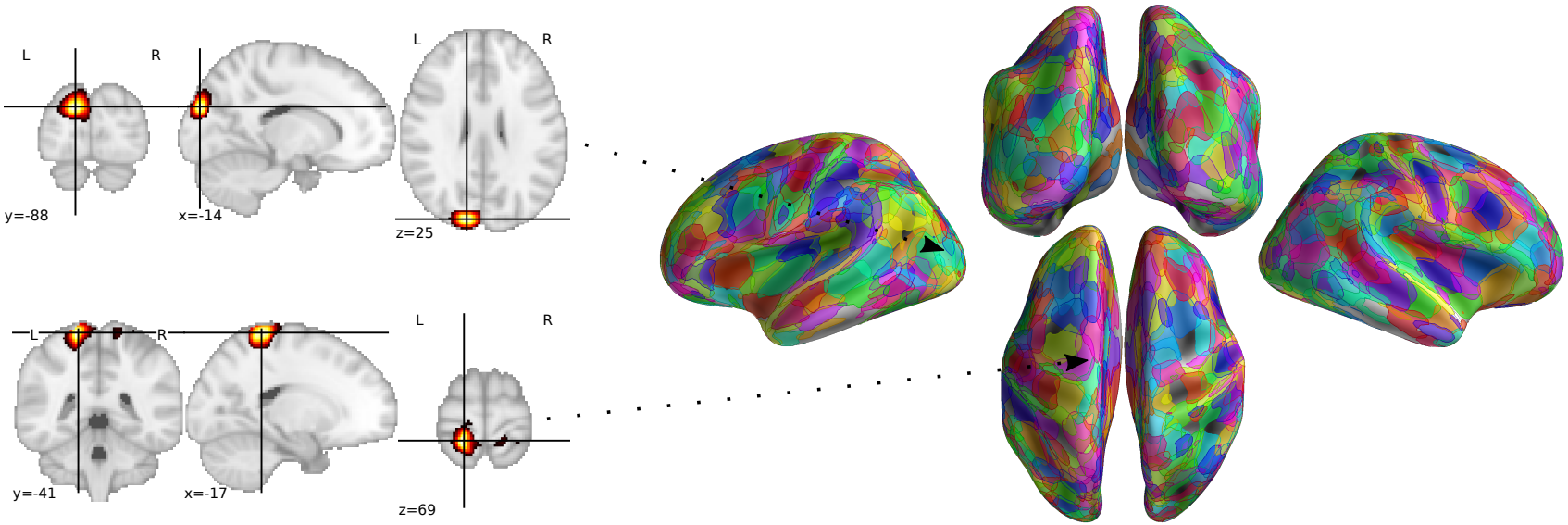
10 h run time

Finding bigger dictionaries



512 components dictionary learned on 3M samples of HCP

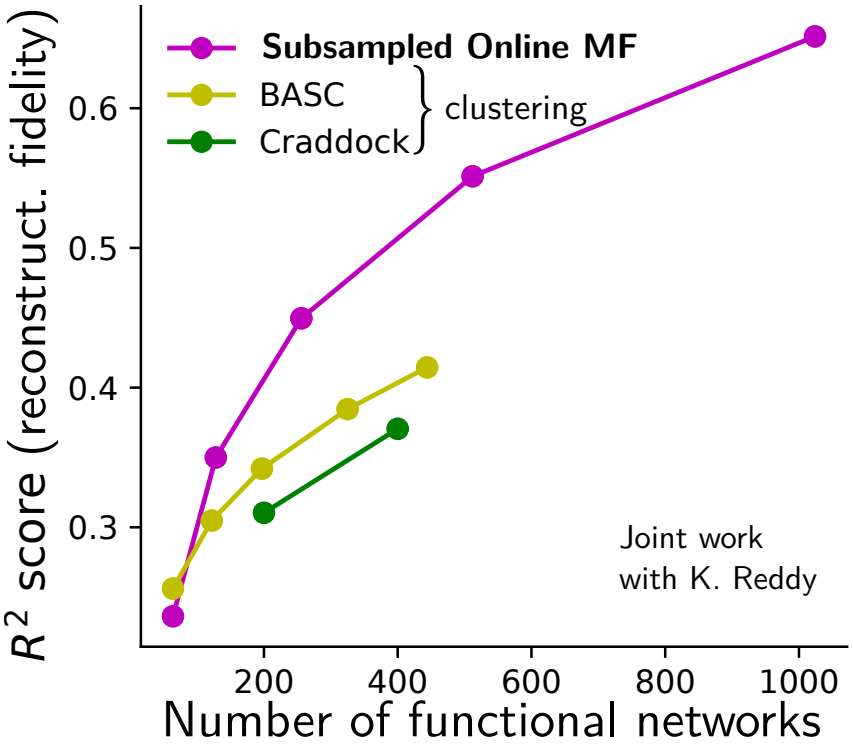
Finding bigger dictionaries



512 components dictionary learned on 3M samples of HCP

Performance of reducing $\boldsymbol{X}$ into loadings $\boldsymbol{A}$?

A compressing atlas: Neurovault (test repository)

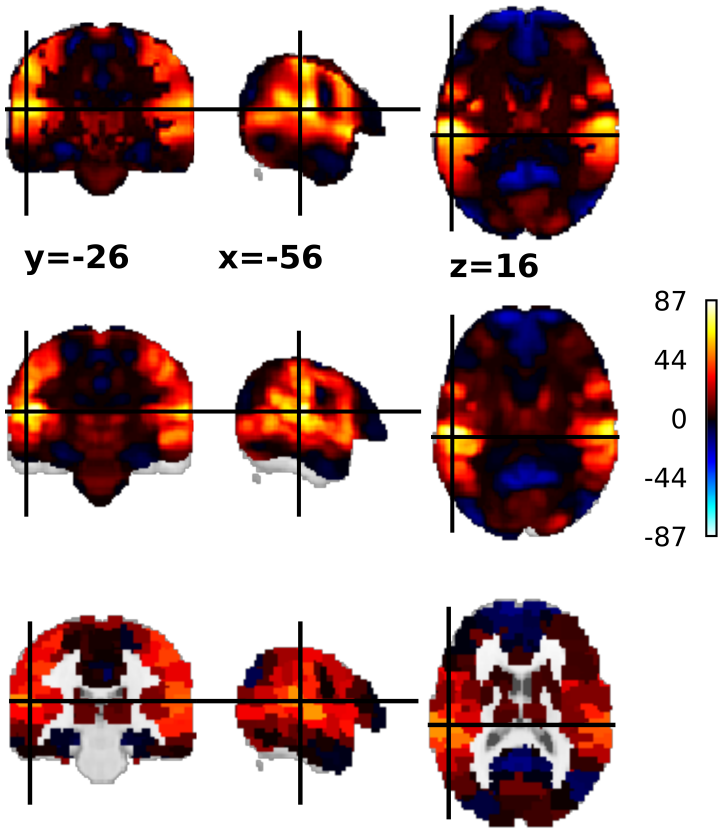


Replacing $\boldsymbol{x}$ with loadings $\boldsymbol{\alpha}$:

- From 200,000 voxels to 1024 components with little loss of information.

$$r^2 = 1 - \frac{\|\boldsymbol{X} - \boldsymbol{DA}\|_F^2}{\|\boldsymbol{X}\|_F^2}$$

Compression in practice



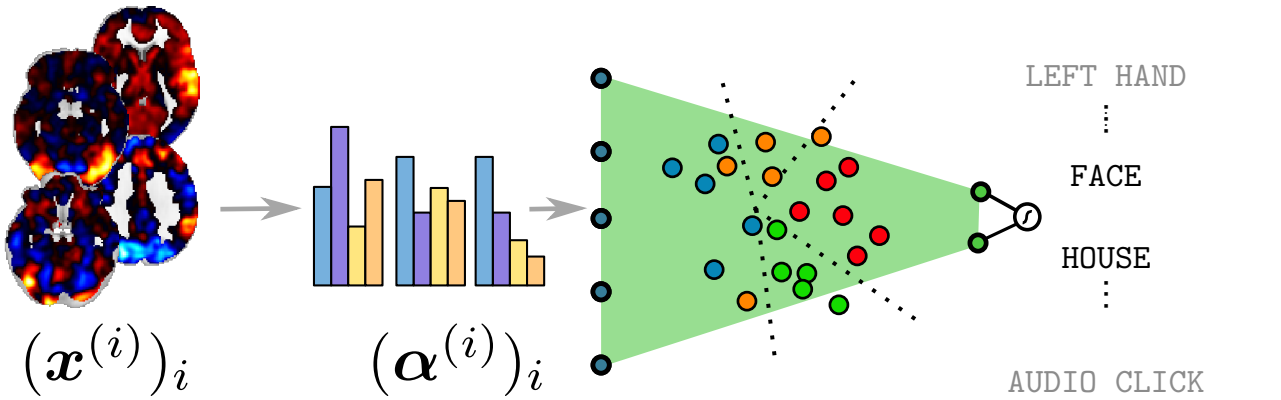
Motor statistical map from Neurovault

No reduction
200,000 voxels

Stochastic Online Matrix Factorization
512 atoms

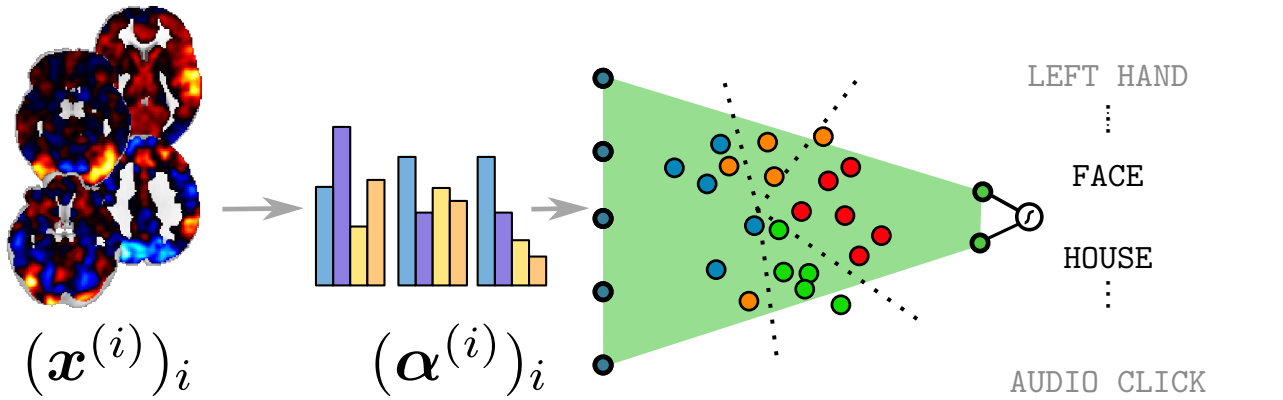
BASC (state-of-the art clustering method)
444 atoms

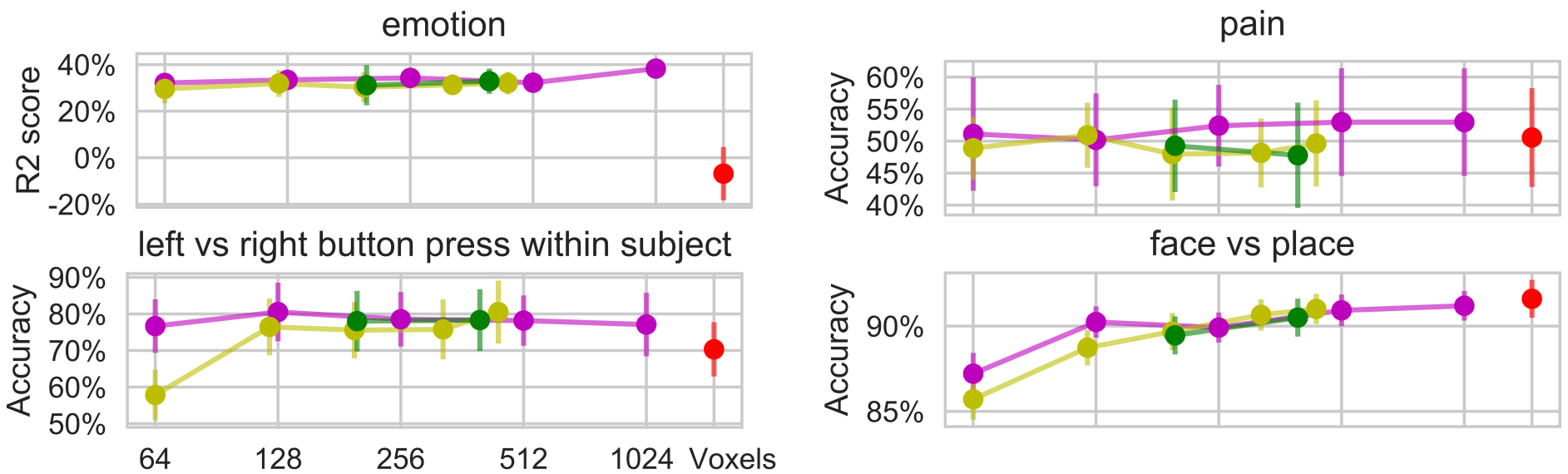
Powerful representation finder



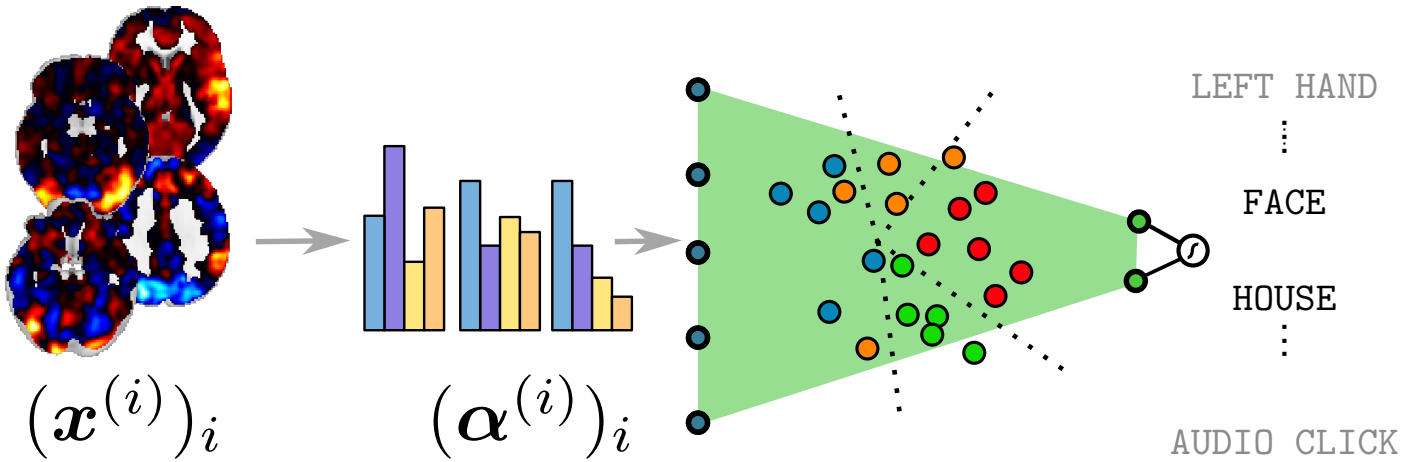
$(\boldsymbol{x}^{(i)})_i$ Statistical maps $(\boldsymbol{\alpha}^{(i)})_i$ Functional loadings Psycho. condition

Powerful representation finder





Going supervised



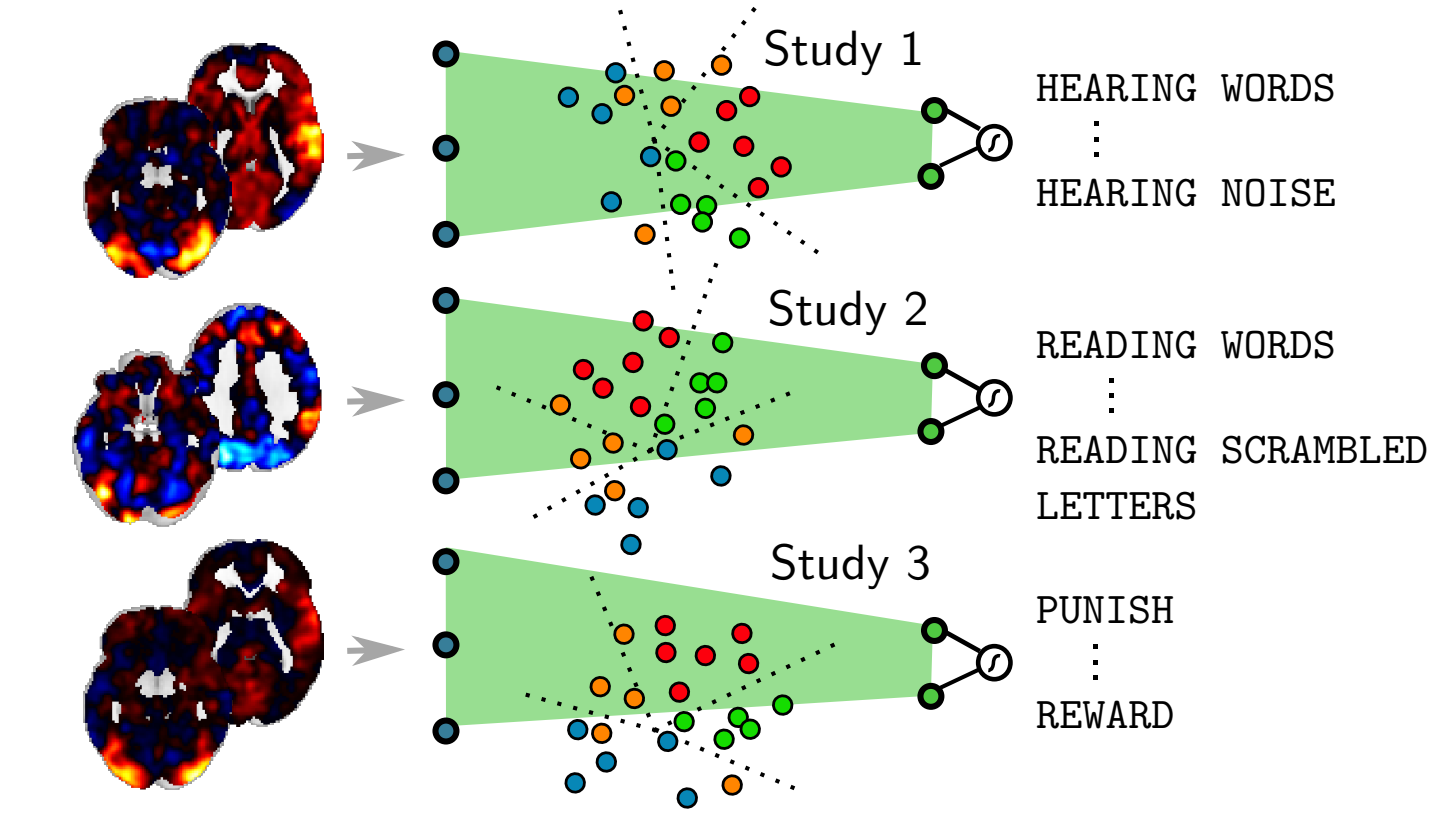
Statistical maps Functional loadings Psycho. condition

Functional networks not aware of labelling tasks.

Can we find better representations ?

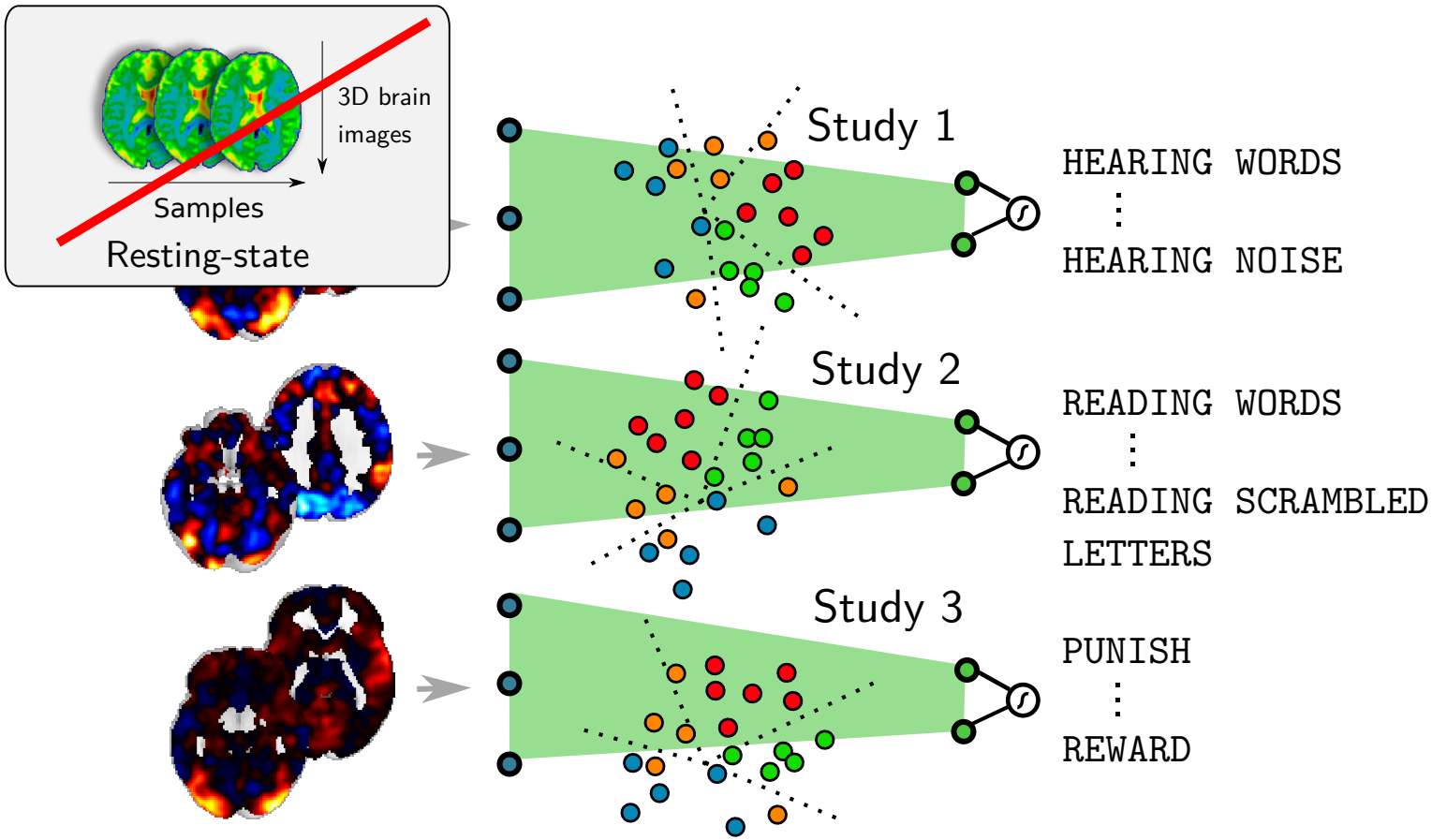
Extracting Supervised Representations of Cognition across Brain-Imaging Studies

Decoding different protocols



How to share information between studies ?

Decoding different protocols



How to share information between studies ?

Finding commonalities between fMRI studies

Newell, A. You Can't Play 20 Questions with Nature and Win: Projective Comments on the Papers of This Symposium. *Visual Information Processing*, 1–26 (1973)

Old question in cognitive science:

- fMRI data holds various answers to specific cognitive questions
- To be fed into a **broader model of cognition**

Finding commonalities between fMRI studies

Newell, A. You Can't Play 20 Questions with Nature and Win: Projective Comments on the Papers of This Symposium. *Visual Information Processing*, 1–26 (1973)

Old question in cognitive science:

- fMRI data holds various answers to specific cognitive questions
- To be fed into a **broader model of cognition**

Many-to-many relationships

- Psychological conditions $\leftrightarrow$ brain regions
- Psychological conditions $\leftrightarrow$ experimental protocols

Existing multi-study approaches

Coordinate-based meta-analysis:

- Relabel tasks with common identifiers
- Brainmap (Laird *et al.* 2005), Neurosynth (Yarkoni *et al.* 2011)

Existing multi-study approaches

Coordinate-based meta-analysis: (Noisy)

- Relabel tasks with common identifiers (Costly or noisy)
- Brainmap (Laird *et al.* 2005), Neurosynth (Yarkoni *et al.* 2011)

Existing multi-study approaches

Coordinate-based meta-analysis: (Noisy)

- Relabel tasks with common identifiers (Costly or noisy)
- Brainmap (Laird *et al.* 2005), Neurosynth (Yarkoni *et al.* 2011)

Encoding cognitive ontologies:

- (Schwartz *et al.* 2013; Wager *et al.* 2013)
- Costly + limitations (Salimi-Khorshidi *et al.* 2009)

Existing multi-study approaches

Coordinate-based meta-analysis: (Noisy)

- Relabel tasks with common identifiers (Costly or noisy)
- Brainmap (Laird *et al.* 2005), Neurosynth (Yarkoni *et al.* 2011)

Encoding cognitive ontologies:

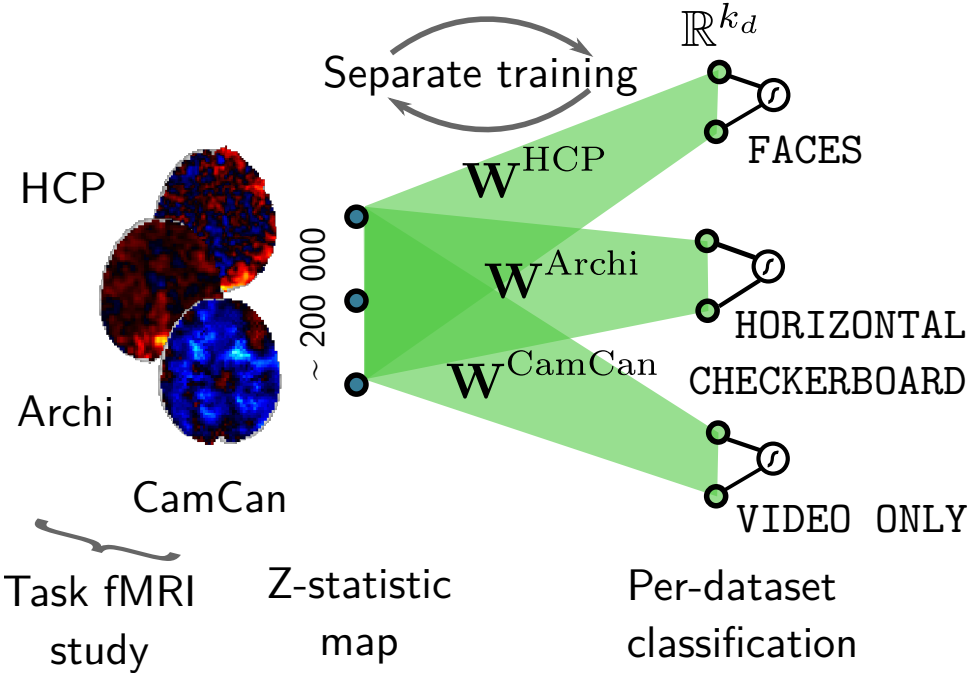
- (Schwartz *et al.* 2013; Wager *et al.* 2013)
- Costly + limitations (Salimi-Khorshidi *et al.* 2009)

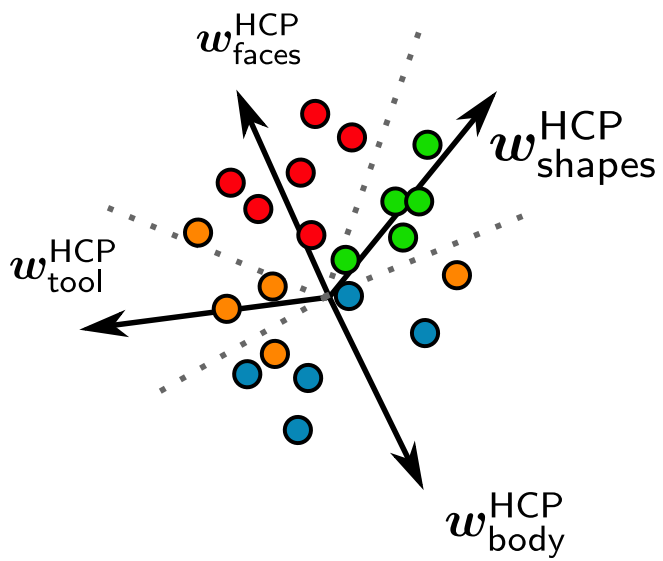
More automated approach to handle heterogenous protocols

Baseline: separated multinomial regressions

Model: Study j with c conditions: $\boldsymbol{W}^j \in \mathbb{R}^{c \times p}$, $\boldsymbol{b}^j \in \mathbb{R}^c$

$$\boldsymbol{p}(\boldsymbol{x}) = \text{softmax}(\boldsymbol{W}^j \boldsymbol{x} + \boldsymbol{b}^j) \qquad \hat{y} = \text{argmax}\, \boldsymbol{p}(\boldsymbol{x})$$

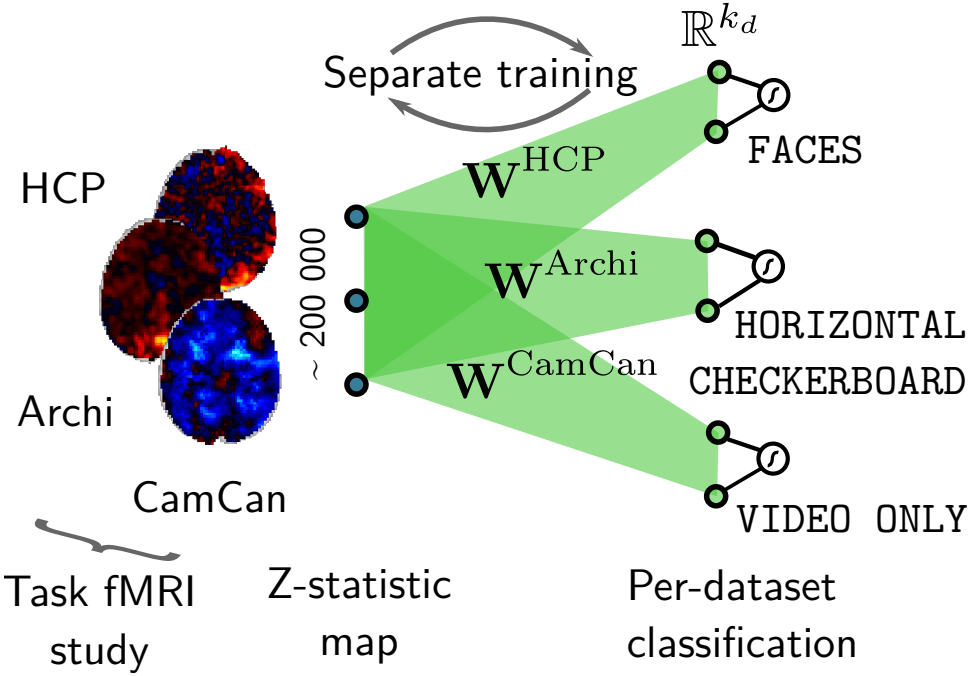


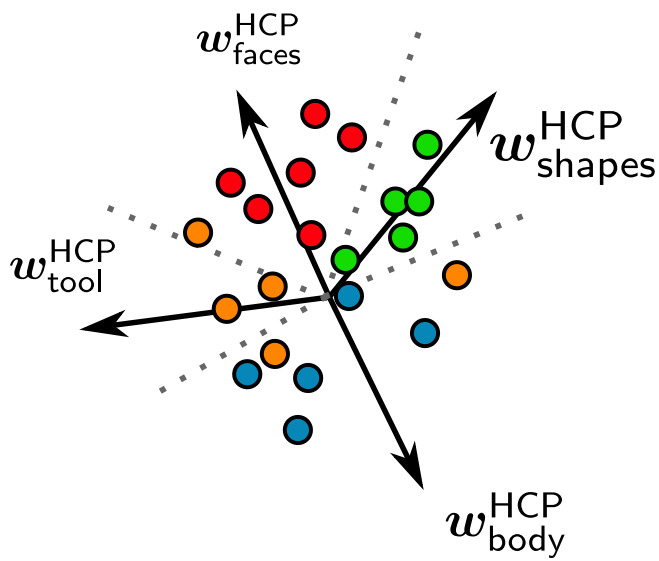


Baseline: separated multinomial regressions

Model: Study j with c conditions: $\boldsymbol{W}^j \in \mathbb{R}^{c \times p}$, $\boldsymbol{b}^j \in \mathbb{R}^c$

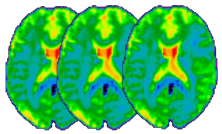
$$\boldsymbol{p}(\boldsymbol{x}) = \text{softmax}(\boldsymbol{W}^j \boldsymbol{x} + \boldsymbol{b}^j) \qquad \hat{y} = \text{argmax}\, \boldsymbol{p}(\boldsymbol{x})$$





Need to ↑ sample-size, ↓ sample dimension

Joint training of a common cognitive spatial basis



4TB resting-state data

HCP900

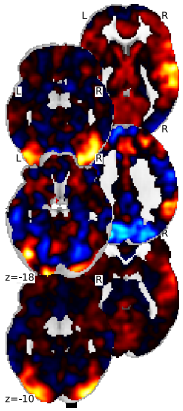
OpenfMRI

HCP

Camcan

Brainomics

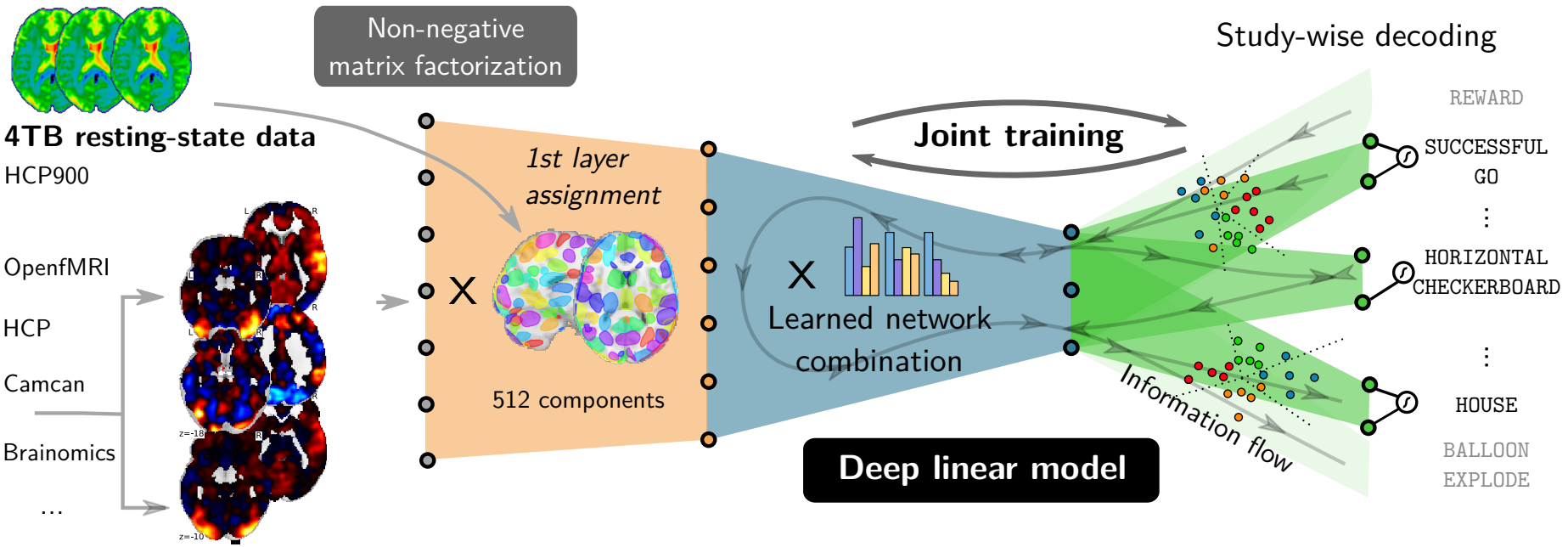
...



35 studies, 40000 task fMRI contrast maps into one model

Aggregation from many fMRI studies

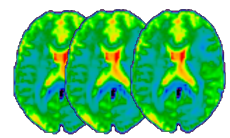
Joint training of a common cognitive spatial basis



35 studies, 40000 task fMRI contrast maps into one model

Aggregation from many fMRI studies

Joint training of a common cognitive spatial basis



4TB resting-state data

HCP900

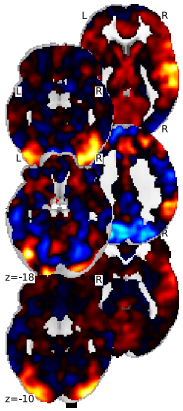
OpenfMRI

HCP

Camcan

Brainomics

...

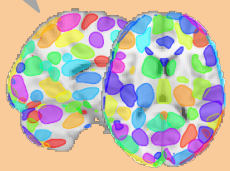


35 studies, 40000 task fMRI contrast maps into one model

Aggregation from many fMRI studies

Non-negative matrix factorization

1st layer assignment



512 components

Learned network combination

Joint training

Study-wise decoding

REWARD

SUCCESSFUL GO

HORIZONTAL CHECKERBOARD

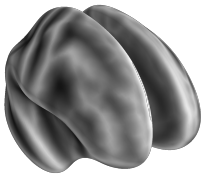
HOUSE

BALLOON EXPLODE

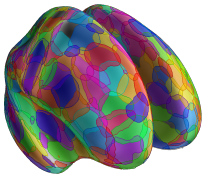
Information flow

Deep linear model

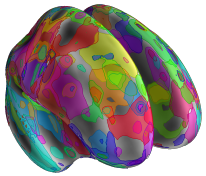
Contrast maps	Resting-state functional loadings	**Task functional loadings**	Contrast prediction



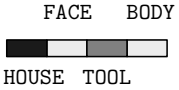
200,000 voxels



512 resting-state networks



128 task-optimized networks



Interpretable complexity reduction

Learning shared representations

Common structure for classification vectors $(\boldsymbol{W}^j)_{j\in[N]}$:

- Vector $\boldsymbol{w}_{\text{faces}}^{\text{Archi}}$ close to $\boldsymbol{w}_{\text{faces}}^{\text{HCP}}$

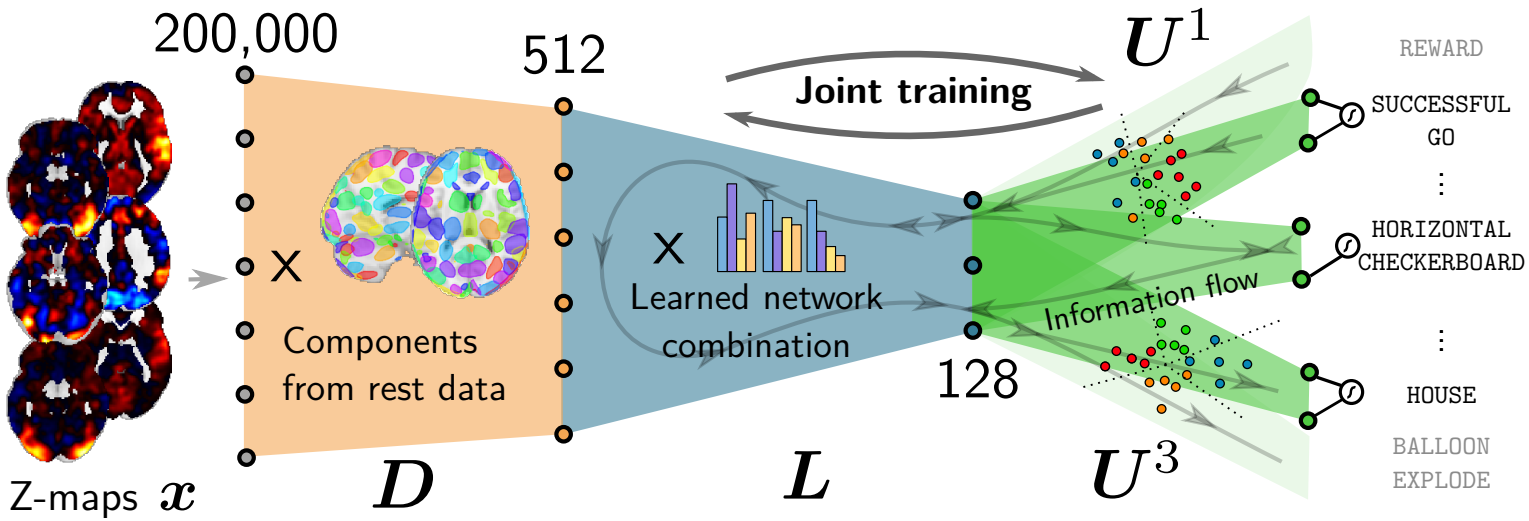
Learning shared representations

Common structure for classification vectors $(\boldsymbol{W}^j)_{j\in[N]}$:

- Vector $\boldsymbol{w}^{\text{Archi}}_{\text{faces}}$ close to $\boldsymbol{w}^{\text{HCP}}_{\text{faces}}$

Deep multi-task approach (Xue *et al.* 2007) $\rightarrow$ **transfer learning**

Three-layer model: $\boldsymbol{W}^j = \boldsymbol{U}^j\boldsymbol{L}\boldsymbol{D}$ for each study j



$$\min_{\substack{\boldsymbol{L}\in\mathbb{R}^{l\times k}\\(\boldsymbol{U}^j,\boldsymbol{b}^j)_j}} -\sum_{j=1}^{N}\frac{1}{n^j}\sum_{i=1}^{n^j}\left((\boldsymbol{U}^j\boldsymbol{L}\boldsymbol{D}\boldsymbol{x}_i^j+\boldsymbol{b}^j)_{y_i^j}-\log\Big(\sum_{c=1}^{c^j}\exp\left(\boldsymbol{U}^j\boldsymbol{L}\boldsymbol{D}\boldsymbol{x}_i^j+\boldsymbol{b}^j\right)_c\Big)\right)$$

Training and introspection

1. Learn $\boldsymbol{L}, (\boldsymbol{U}^j)_j$ with **stochastic gradient descent**

Training and introspection

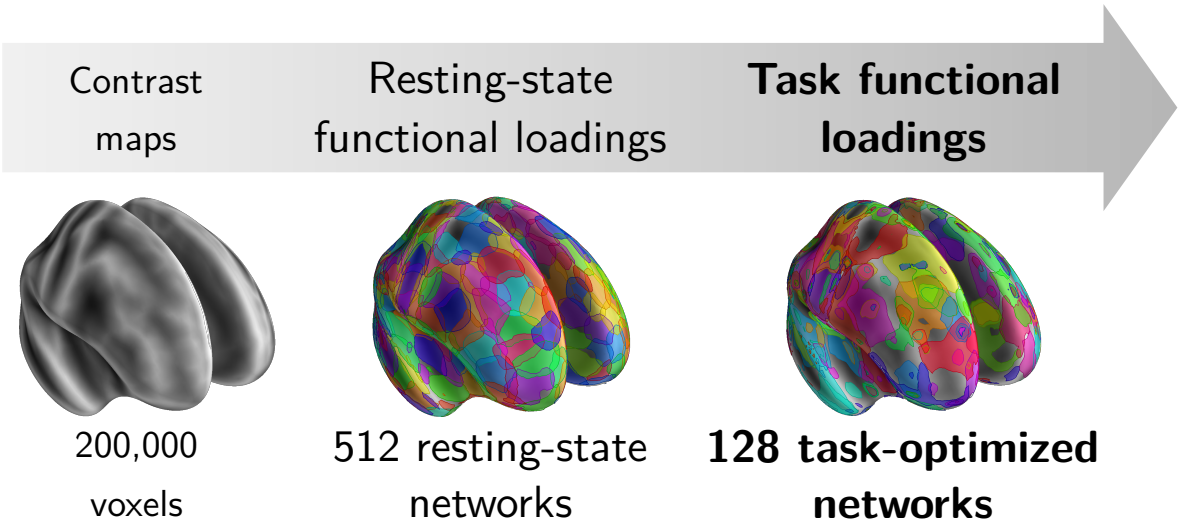
1. Learn $\boldsymbol{L}, (\boldsymbol{U}^j)_j$ with **stochastic gradient descent**
2. **Dropout** (Srivastava *et al.* 2014) to ensure transfer
   - Randomly set coordinates to zero in $\boldsymbol{Dx}$, $\boldsymbol{LDx}$ during training
   - All rows of $\boldsymbol{L}$ useful for prediction

Training and introspection

1. Learn $\boldsymbol{L}, (\boldsymbol{U}^j)_j$ with **stochastic gradient descent**
2. **Dropout** (Srivastava *et al.* 2014) to ensure transfer
   - Randomly set coordinates to zero in $\boldsymbol{Dx}$, $\boldsymbol{LDx}$ during training
   - All rows of $\boldsymbol{L}$ useful for prediction
3. **Ensemble** + matrix factorization $\rightarrow$ interpretable 2^{nd} layer $\boldsymbol{L}$

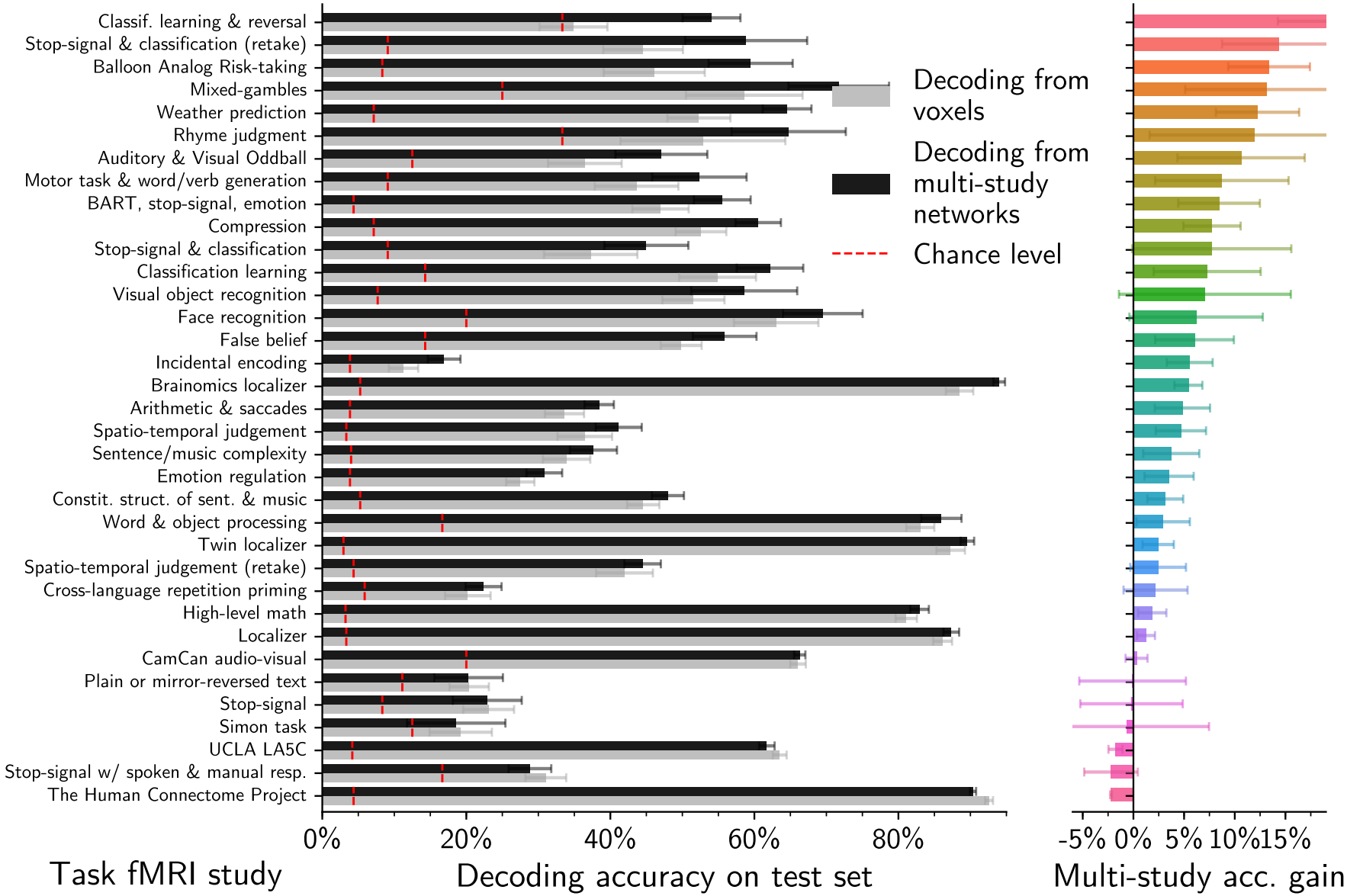
Training and introspection

1. Learn $\boldsymbol{L}, (\boldsymbol{U}^j)_j$ with **stochastic gradient descent**
2. **Dropout** (Srivastava *et al.* 2014) to ensure transfer
   - Randomly set coordinates to zero in $\boldsymbol{Dx}$, $\boldsymbol{LDx}$ during training
   - All rows of $\boldsymbol{L}$ useful for prediction
3. **Ensemble** + matrix factorization $\rightarrow$ interpretable 2nd layer $\boldsymbol{L}$



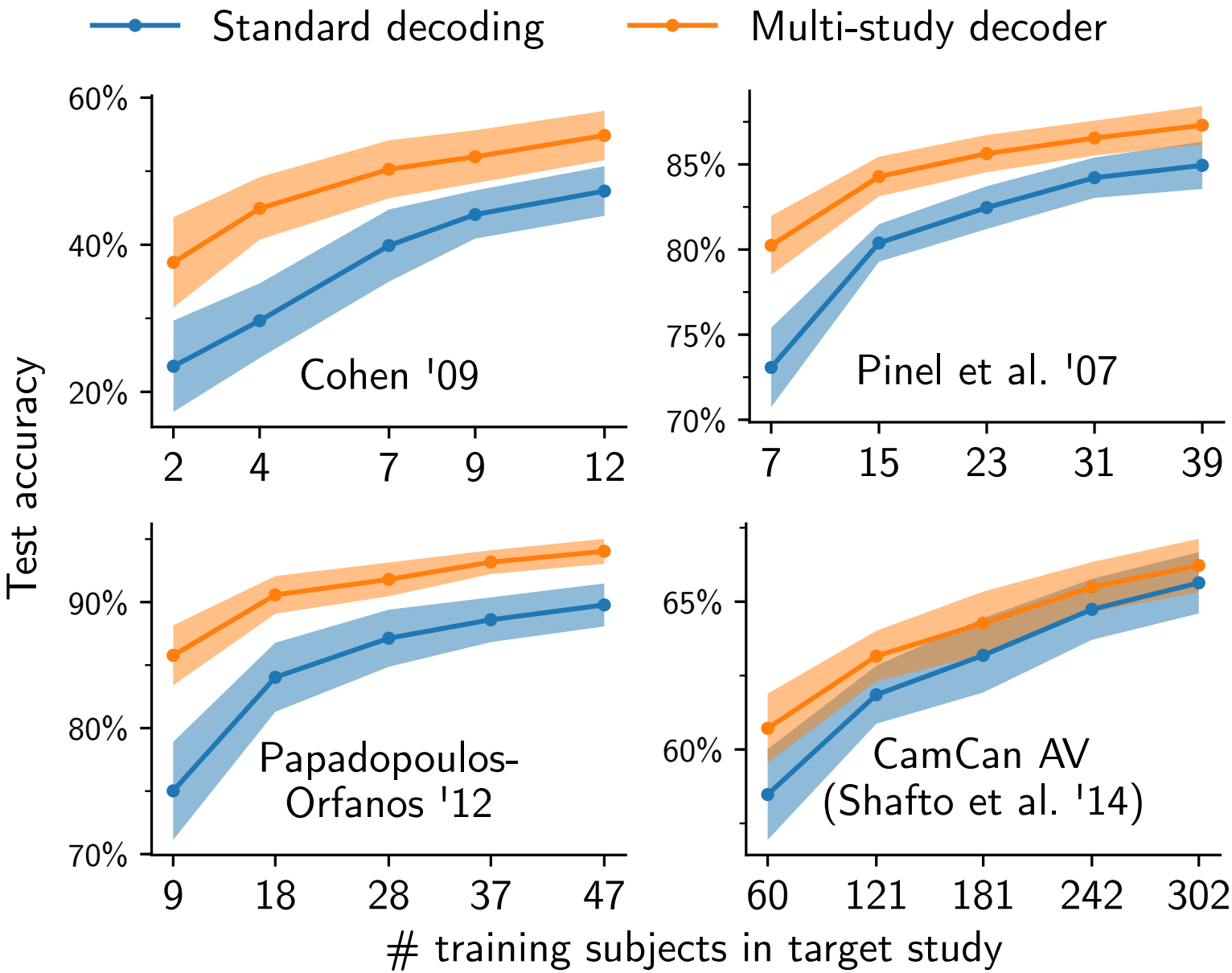
Representation on $\boldsymbol{LD}$: multi-study task-optimized networks

Quantitative improvement in decoding accuracy

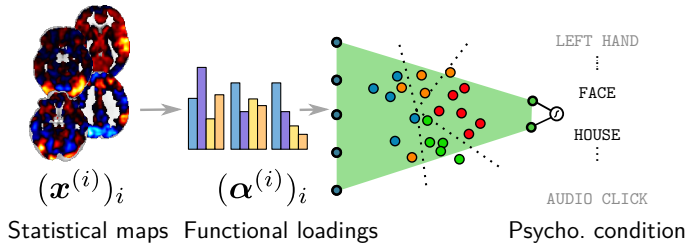


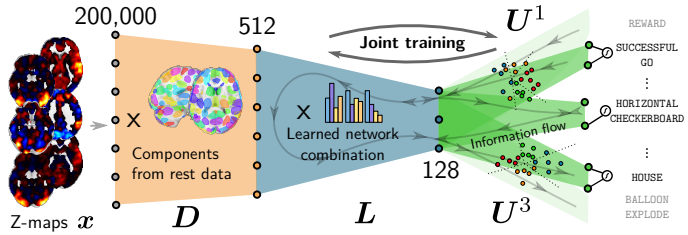
Test performance: random half-split of data for every study

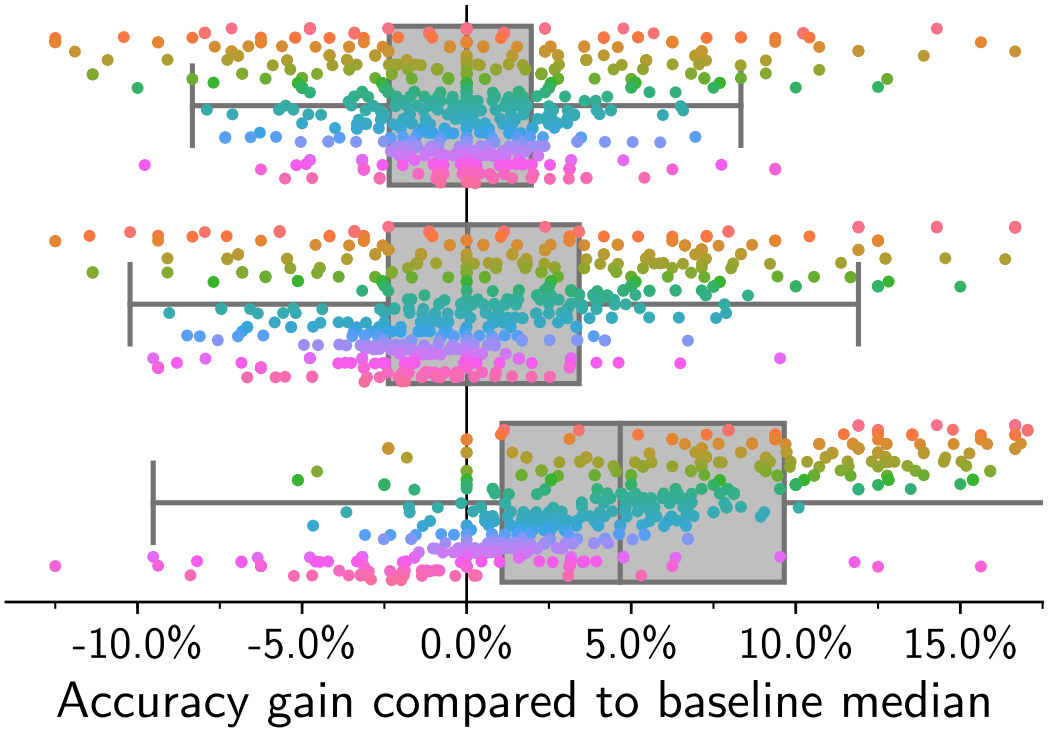
Transfer works best for small studies



Importance of the supervised second-layer





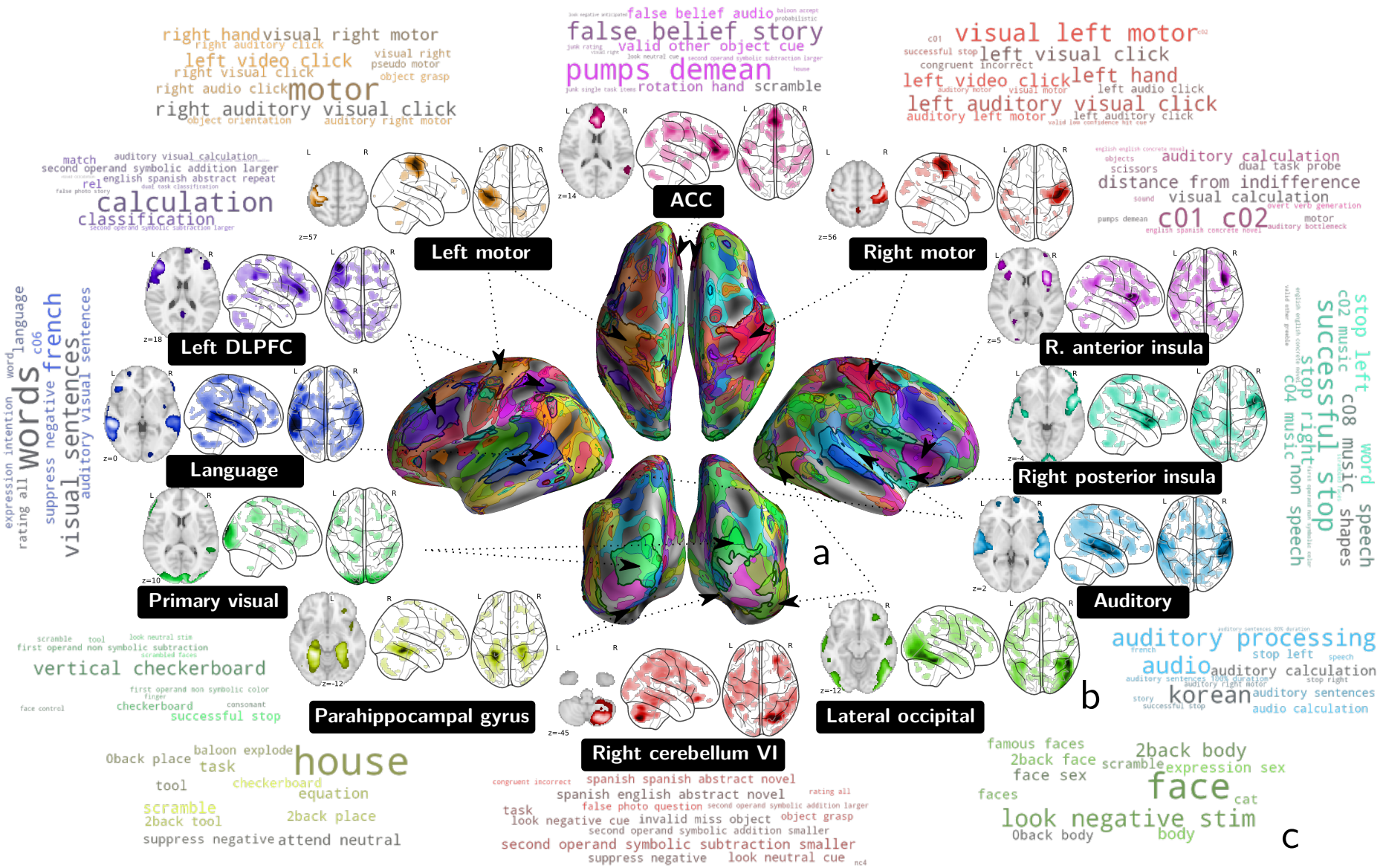


Standard decoding from voxels

Decoding from functional networks

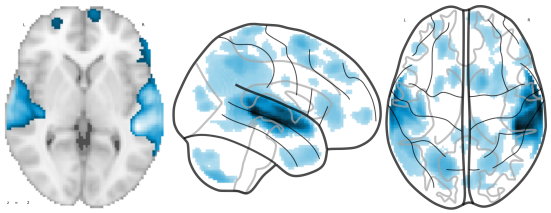
Decoding from multi-study task-optimized networks

Task-optimized networks (*LD*)

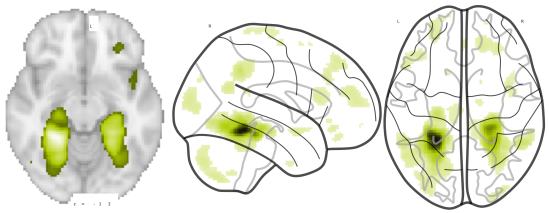


Meaningful cognitive dimensions

Auditory



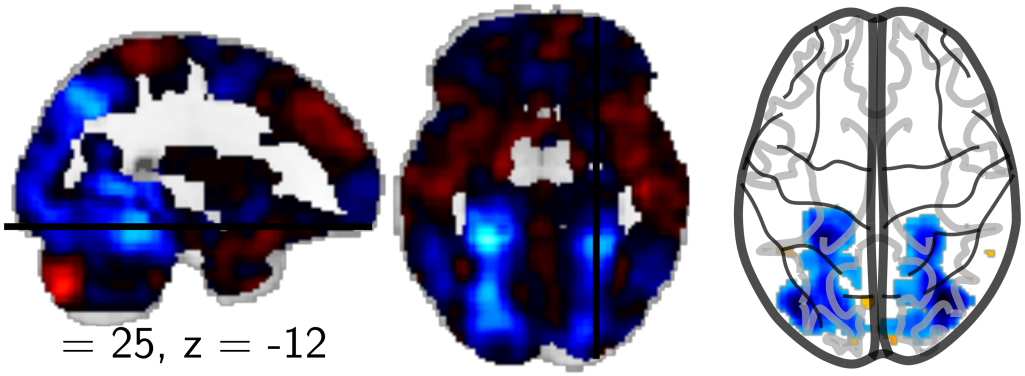
Parahippocampal gyrus



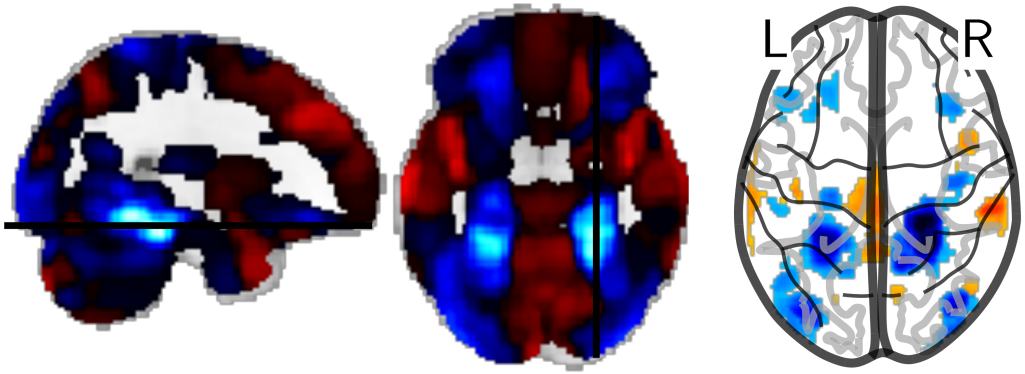
Better defined classification maps

Face vs house
Haxby *et al.* (2001)
B-accuracy: 99.4%
B-acc. gain: +6.9%

Voxelwise decoder



Task-network decoder



Parahippocampal Place Area (PPA)

Discussion

Algorithmic dev. (Part 1) allowed model dev. (Part 2)

Subsampled Online Matrix Factorization:

- Algorithm useful in other matrix factorization setting
- Collaborative filtering, computer vision
- **Perturbation approach** may be extended

Pooling resting-state + 35 task-fMRI studies into shared models

$\rightarrow$ Powerful cognitive representation

- Use deep learning tools but **linear model**
- Need regularization and interpretability transform.

Perspectives and reproducibility

Future development:

- fMRI data standards → larger corpora → complex models
- More structured representations (beyond linear)
- Perturbating stochastic algorithms to prone information faster

Software + networks: (Python)

- `github.com/arthurmensch/modl`: massive matrix factorizat.
- `github.com/cogspaces/cogspaces`: multi-study training

M., A., Mairal, J., Thirion, B. & Varoquaux, G. Extracting Universal Representations of Cognition across Brain-Imaging Studies. *arXiv:1809.06035 [stat.ML], journal article under review* (2018)

M., A. & Blondel, M. *Differentiable Dynamic Programming for Structured Prediction and Attention.* in *Proceedings of the International Conference on Machine Learning (ICML)* (2018)

M., A., Mairal, J., Thirion, B. & Varoquaux, G. Stochastic Subsampling for Factorizing Huge Matrices. *IEEE Transactions on Signal Processing* **66,** 113–128 (2018)

M., A., Mairal, J., Bzdok, D., Thirion, B. & Varoquaux, G. *Learning Neural Representations of Human Cognition Across Many fMRI Studies.* in *Advances in Neural Information Processing Systems (NIPS)* (2017)

Dohmatob, E., M., A., Varoquaux, G. & Thirion, B. *Learning Brain Regions via Large-Scale Online Structured Sparse Dictionary Learning.* in *Advances in Neural Information Processing Systems (NIPS)* (2016)

M., A., Mairal, J., Thirion, B. & Varoquaux, G. *Dictionary Learning for Massive Matrix Factorization.* in *Proceedings of the International Conference on Machine Learning (ICML)* (2016)

M., A., Varoquaux, G. & Thirion, B. *Compressed Online Dictionary Learning for Fast fMRI Decomposition.* in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)* (2016)

Advisors
Bertrand Thirion
Gaël Varoquaux
Julien Mairal

Joint projects
Olivier Grisel
Elvis Dohmatob
Kamalakar Reddy
Mathieu Blondel
Danilo Bzdok

Team
Jérôme Dockès
Thomas Moreau
Mehdi Rahim
Alexandre Gramfort
Philippe Ciuciu
Alberto Bietti
Pierre Ablin
Mathurin Massias
Patricio Cerda
Carole Lazarus
Loubna El Gueddari
Joseph Salmon
Denis Engelmann
and many others !

Team
Guillaume Lemaitre
Joris V. den Bossche
Joan Massich
Ana Luísa Pinho
Darya Chyzhyk
Daria La Rocca
Hamza Cherkaoui
Jérôme-A. Chevalier
A. Machlouzarides
Hicham Janati
Alexandre Abraham
Mickael Eickenberg
Andrés Hoyos-Idrobo
Loïc Estève

Stochastic Approximated Majorization Minimization

Input: Initial iterate θ_0, sample stream $(\boldsymbol{x}_t)_{t>0}$
for t from 1 to T **do**
Draw $x_t \sim \mathcal{P}$, get $f_t : \theta \in \Theta \to f(\boldsymbol{x}_t, \theta)$.
Construct a surrogate of f_t near θ_{t-1}, that meets

$$g_t - f_t \leq \epsilon_t, \quad g_t(\theta_{t-1}) - f_t(\theta_{t-1}) \geq -\epsilon_t, \quad \nabla(g_t - f_t) \; L\text{-Lipschitz}$$

Update the aggregated surrogate:

$$\bar{g}_t = (1 - w_t)\bar{g}_{t-1} + w_t g_t.$$

Compute

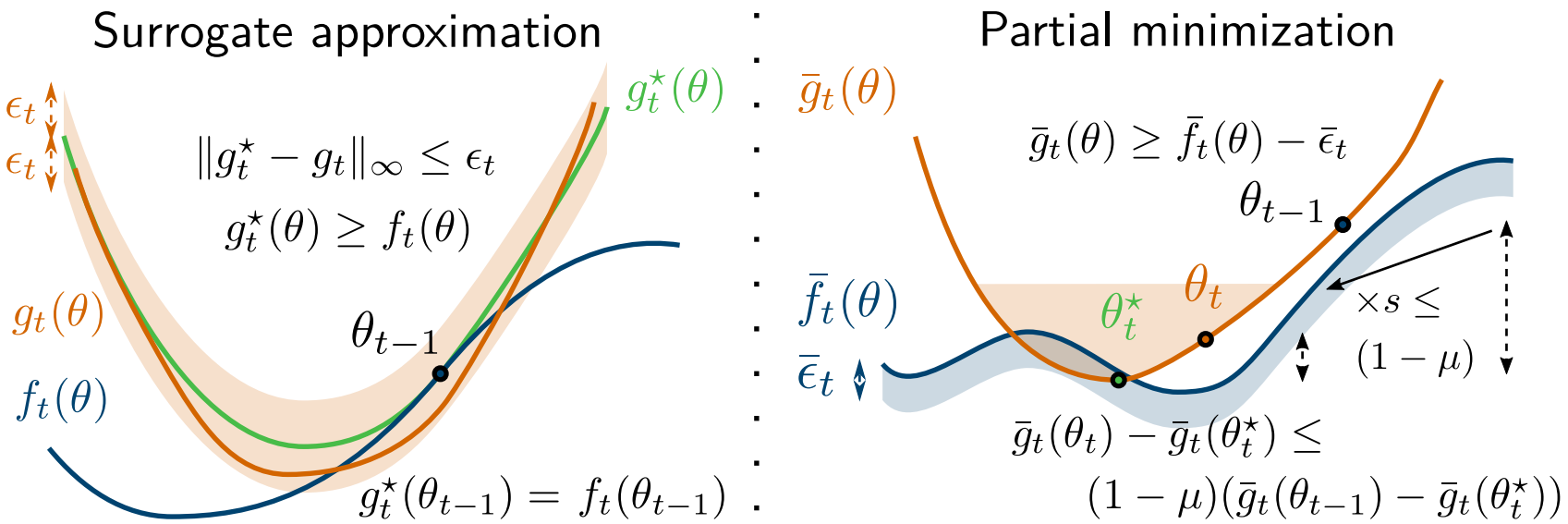
$$\theta_t \approx \underset{\theta \in \Theta}{\operatorname{argmin}}\, \bar{g}_t(\theta)$$

such that

$$\mathbb{E}\left[\bar{g}_t(\theta_t) - \bar{g}_t(\theta_t^\star) | \mathcal{F}_{t-\frac{1}{2}}\right] \leq (1 - \mu)(\bar{g}_t(\theta_{t-1}) - \bar{g}_t(\theta_t^\star)).$$

Output: Final iterate θ_T.

Stochastic Approximated Majorization Minimization



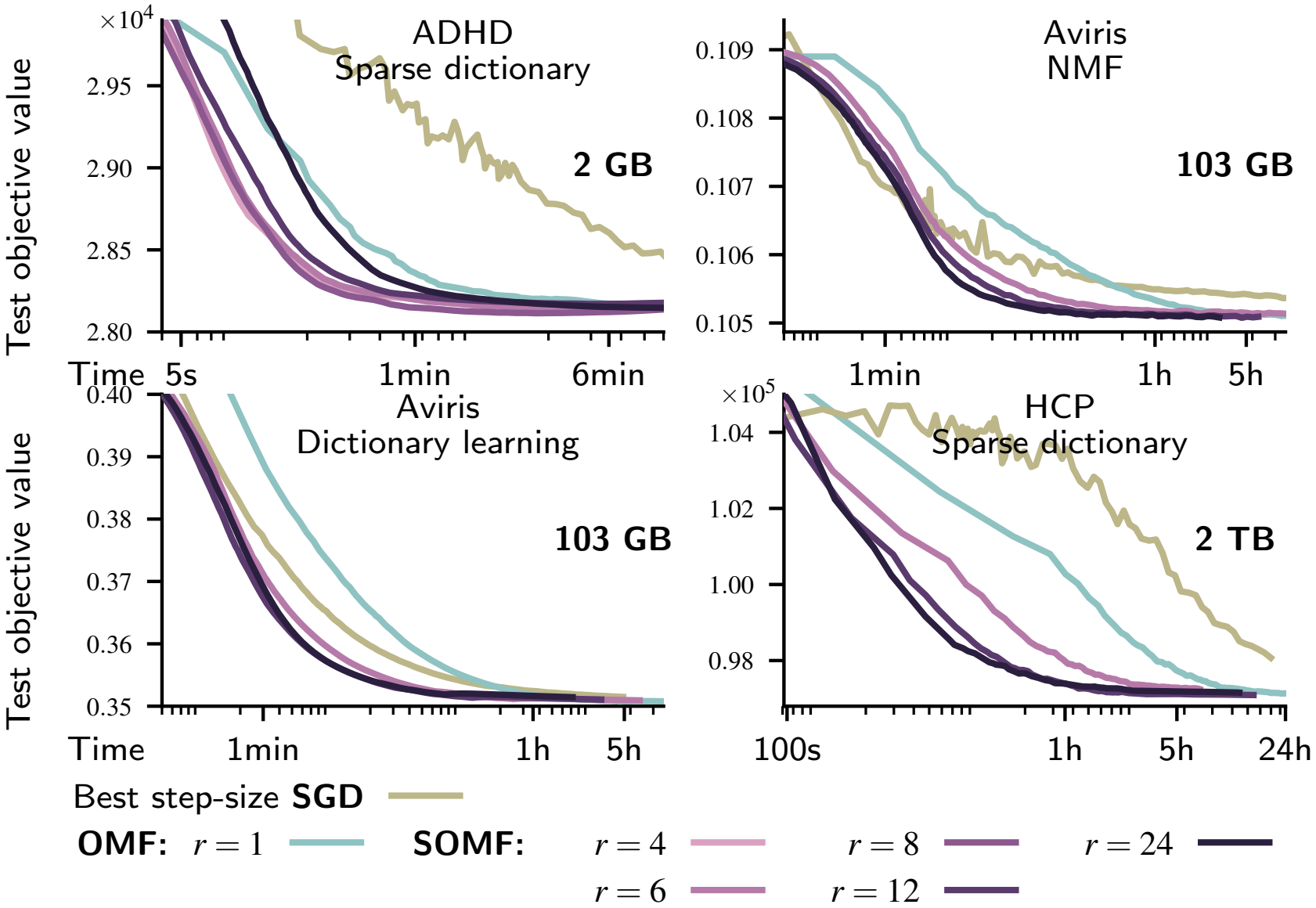
- Learning weights $w_t = t^{-u}$, $u \in]\frac{3}{4}, 1]$
- $\epsilon_t \to 0$ a.s, $\exists\, \eta > 0$ s.t. $\mathbb{E}[\epsilon_t] \in \mathcal{O}(t^{2(u-1)-\eta})$

$=$ sufficient approximation decrease

- $\mu_t > \mu > 0$ for all t = sufficient step-wise progress

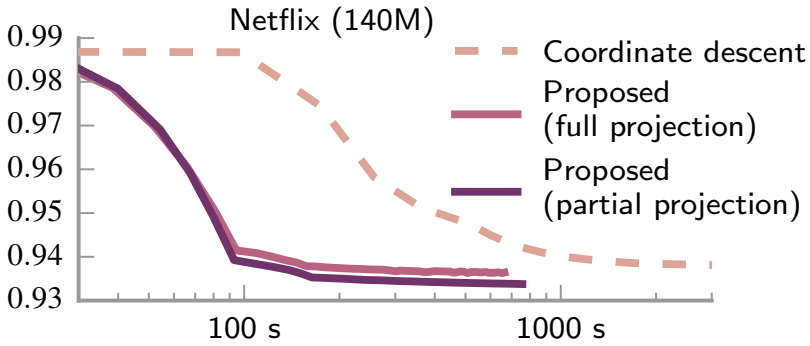
$(\theta_t)_t$ acc. points = local minimizers of $\bar{f}(\theta) = \mathbb{E}_{\boldsymbol{x}}[f(\boldsymbol{x}, \theta)]$.

Other SOMF applications



Collaborative filtering

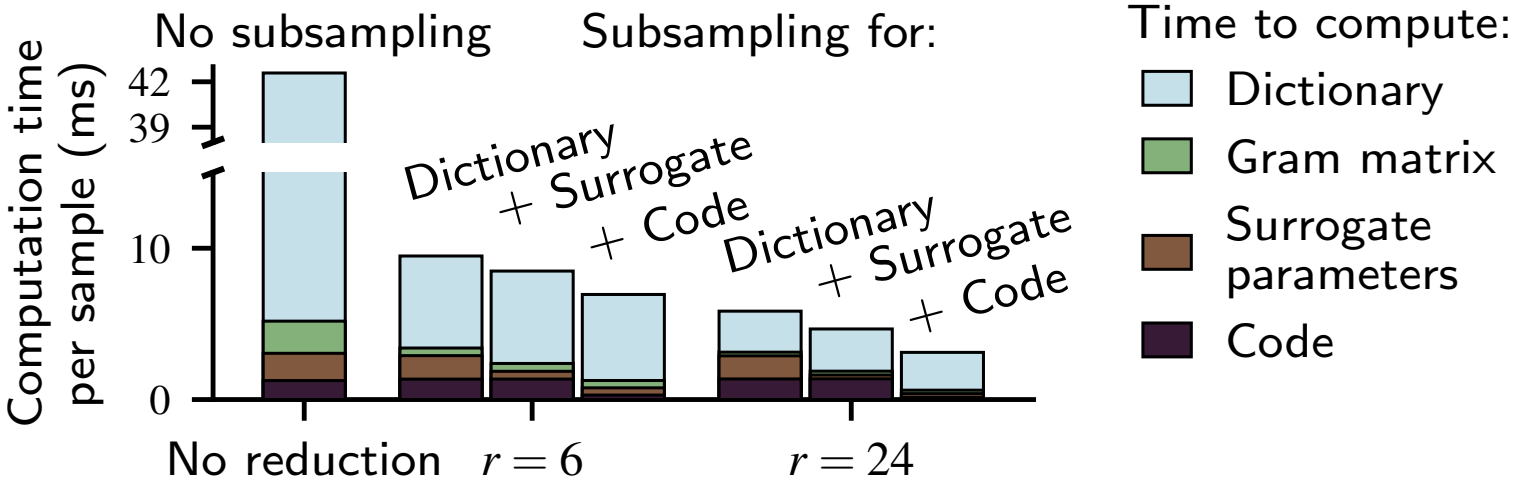
- $\boldsymbol{M}_t \boldsymbol{x}_t$ movie ratings from user t
- vs. coordinate descent for maximum-margin-matrix-factorization loss (no hyperparameters)



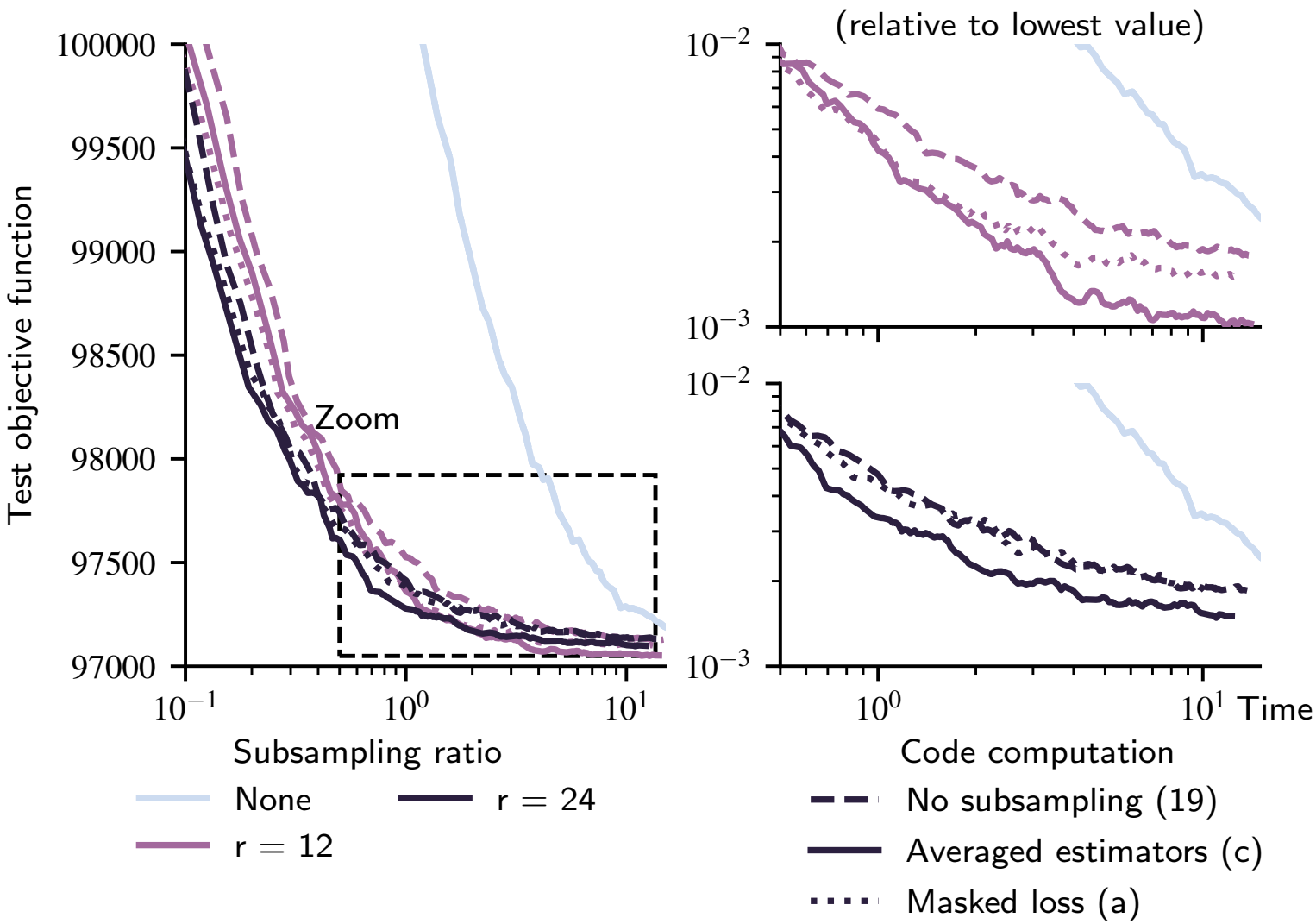
Dataset	Test RMSE		Speed
	CD	**SOMF**	-up
ML 1M	0.872	**0.866**	$\times$0.75
ML 10M	0.802	**0.799**	$\times$**3.7**
NF (140M)	0.938	**0.934**	$\times$**6.8**

- Outperform coordinate descent beyond 10M ratings
- Same prediction performance
- Speed-up 6.8$\times$ on Netflix

Profiling

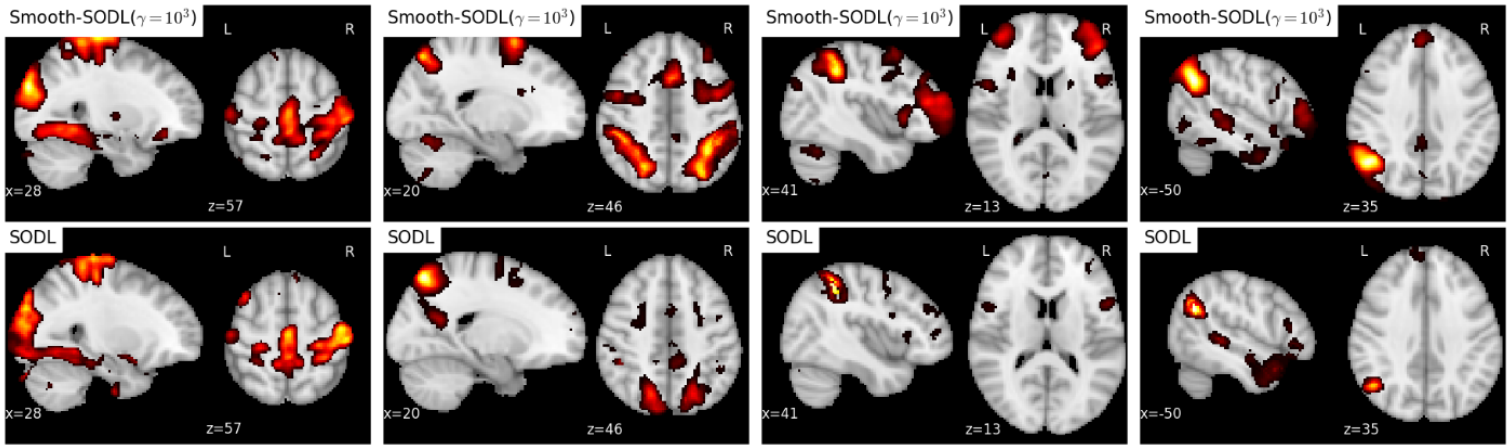

No subsampling
Subsampling for:
Computation time per sample (ms)
42
39
10
0
Dictionary
+ Surrogate
+ Code
Dictionary
+ Surrogate
+ Code
No reduction
$r = 6$
$r = 24$
Time to compute:
Dictionary
Gram matrix
Surrogate parameters
Code

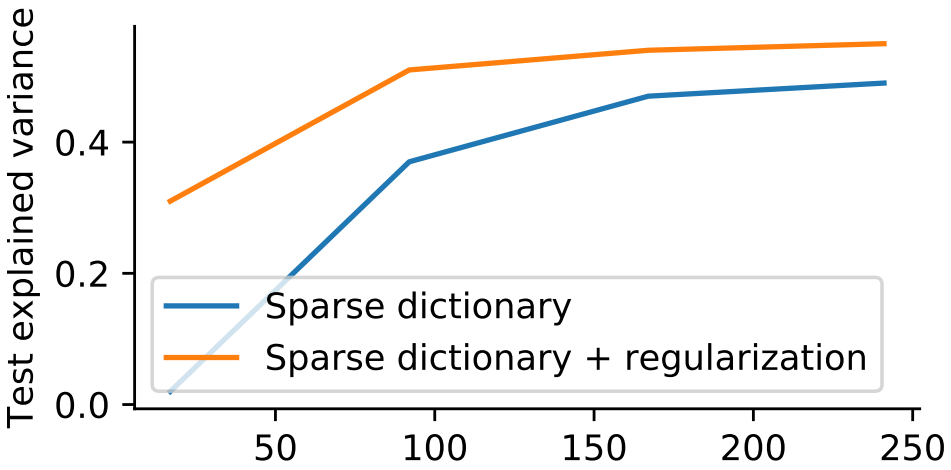
Method comparison



Adding complex regularization to the dictionary

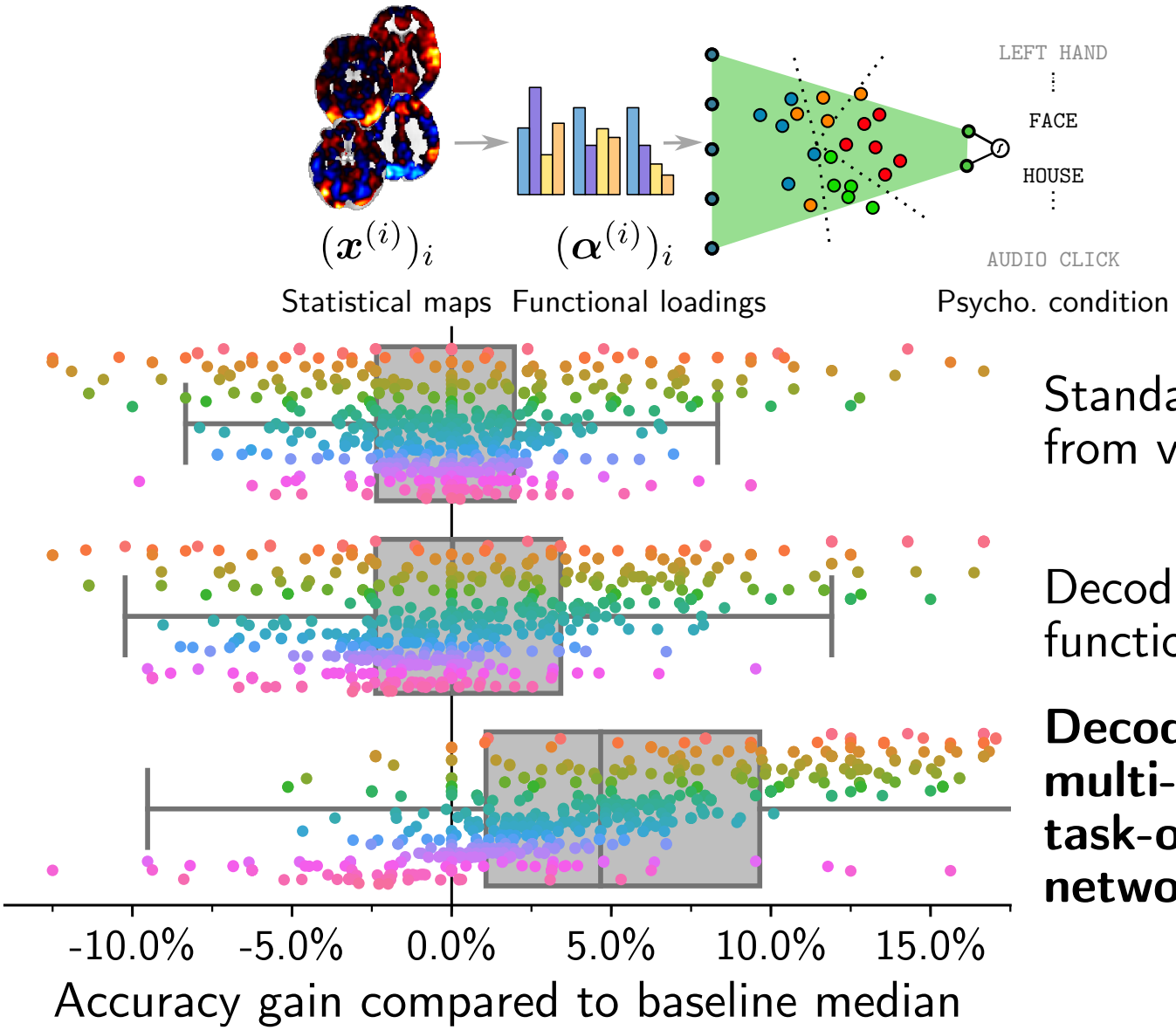
$$\min_{\boldsymbol{D}\in\mathcal{C},\boldsymbol{A}\in\mathbb{R}^{k\times n}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{A}\|_F^2 + \lambda\Omega(\boldsymbol{A}) + \mu\|\triangle\boldsymbol{D}\|_F^2$$



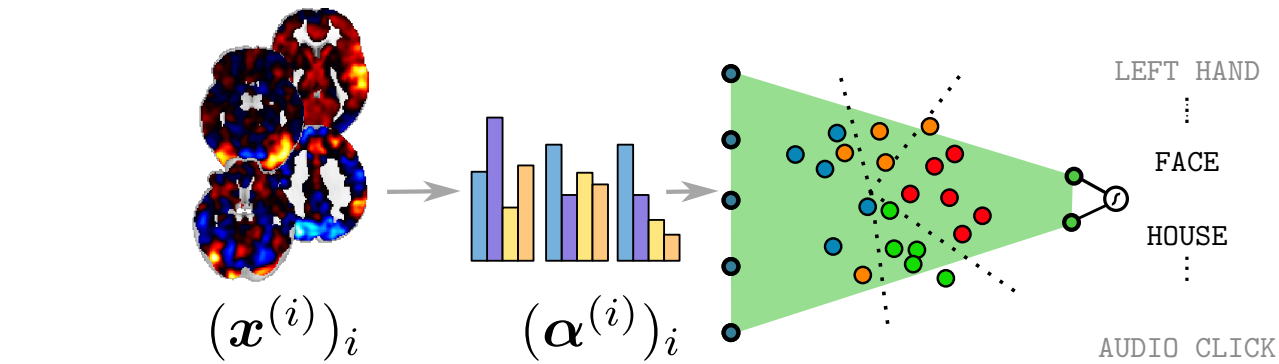


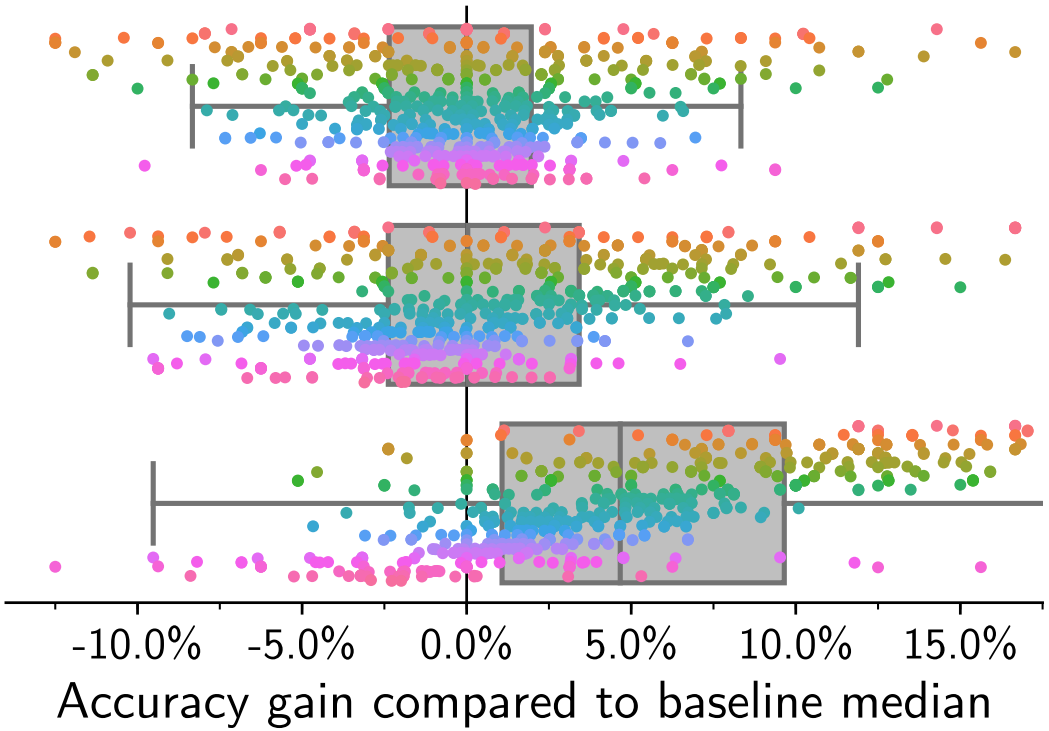
Dohmatob *et al.* (2016)

Further comparison of multi-study decoding



A two layer model ?



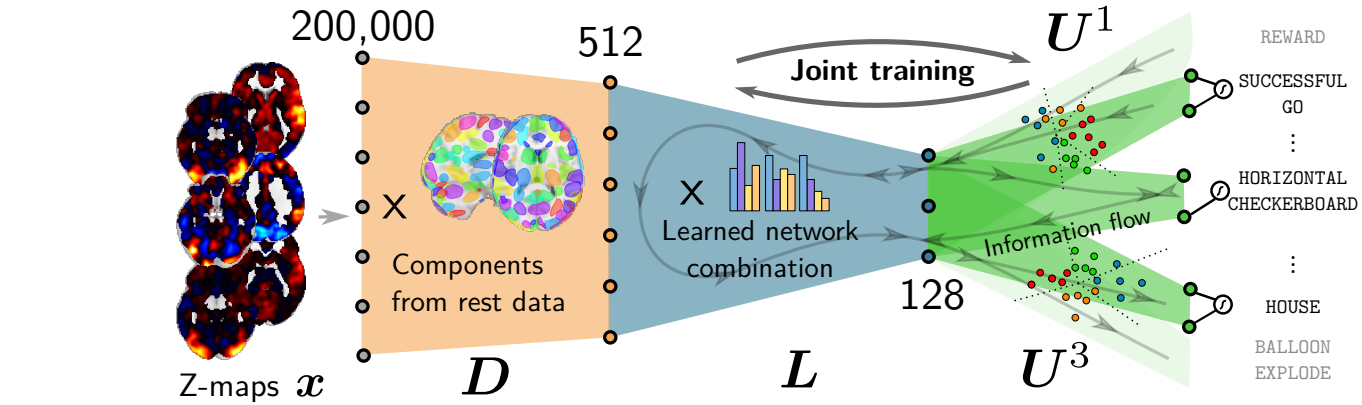


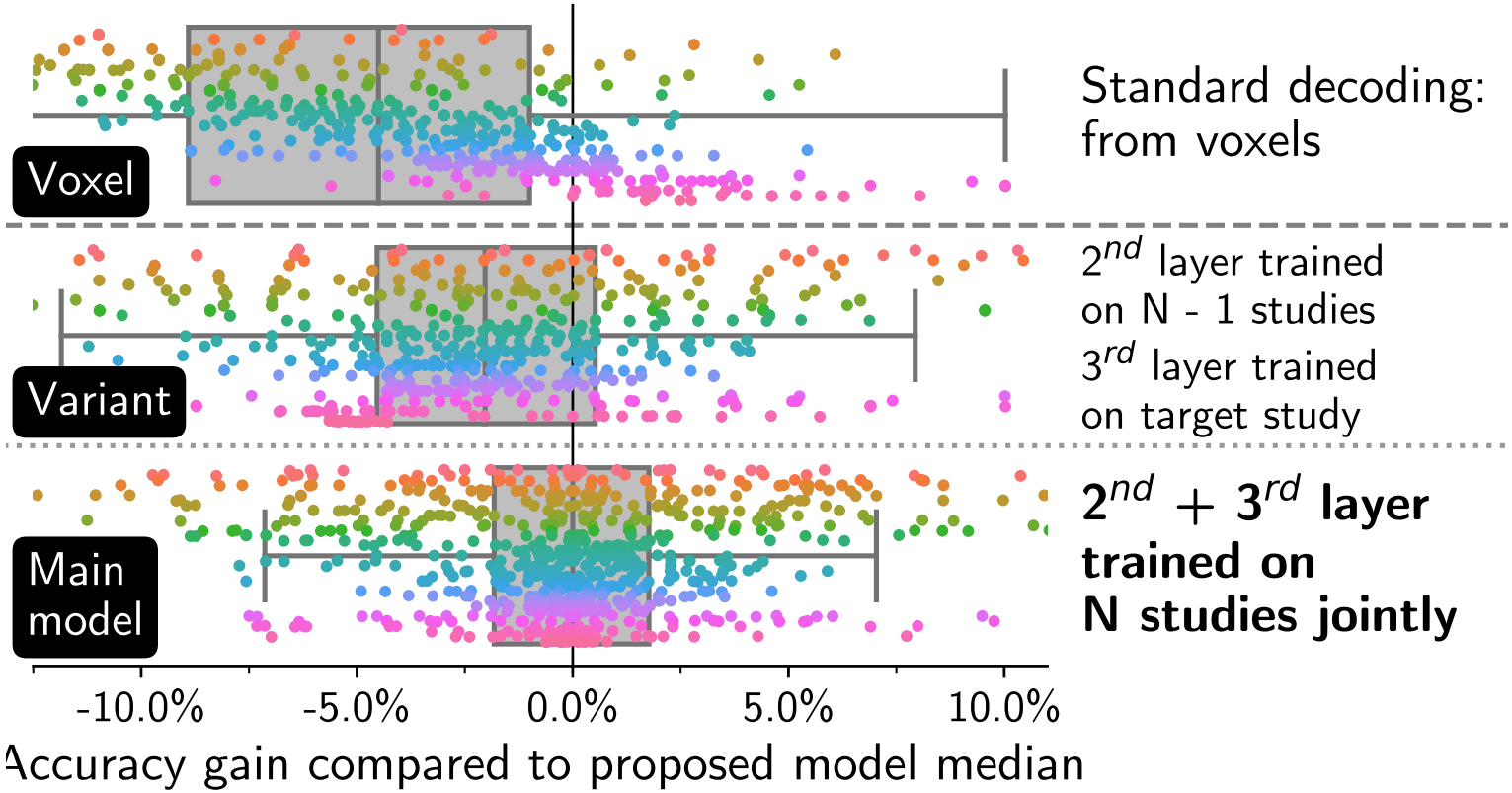
Standard decoding from voxels

Decoding from functional networks

Decoding from multi-study task-optimized networks

Using MSTON over a new study





Standard decoding: from voxels

2^{nd} layer trained on N - 1 studies
3^{rd} layer trained on target study

2^{nd} + 3^{rd} layer trained on N studies jointly

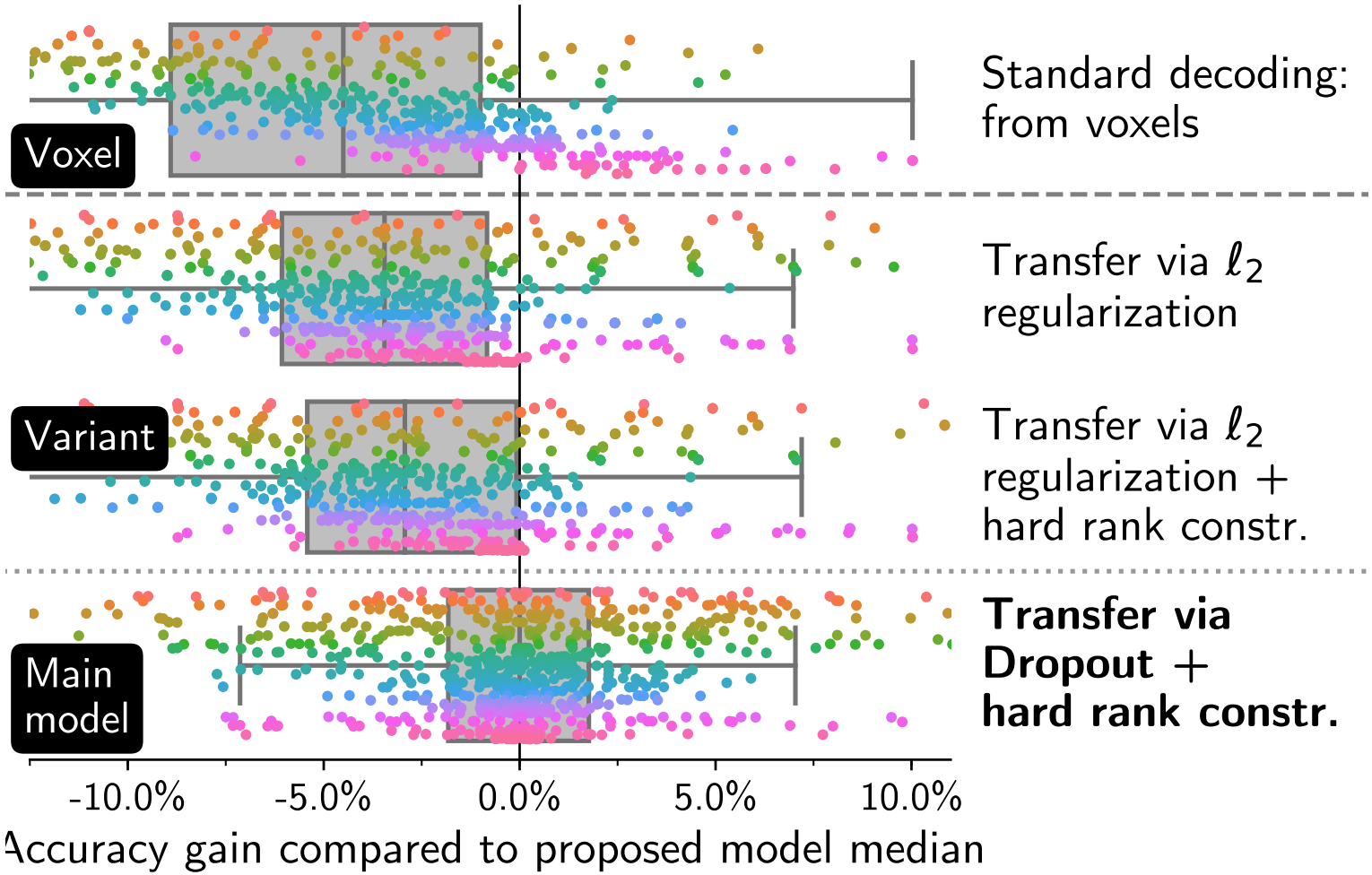
Role of dropout

Without modification nor constraint on the second layer output size l, we cannot expect to observe any transfer learning by solving the joint objective (4). Indeed, in the general case where we allow $l \geq c \triangleq \sum_{j=1}^{N} c^j$, we let $(\tilde{\boldsymbol{V}}^j, \boldsymbol{b}^j)_j$ be the unique solutions of the N non-regularized convex problems (2). We let $\tilde{\boldsymbol{V}} \in \mathbb{R}^{c \times k}$ be the vertical concatenation of $(\boldsymbol{V}^j)_j$. We then form the matrices

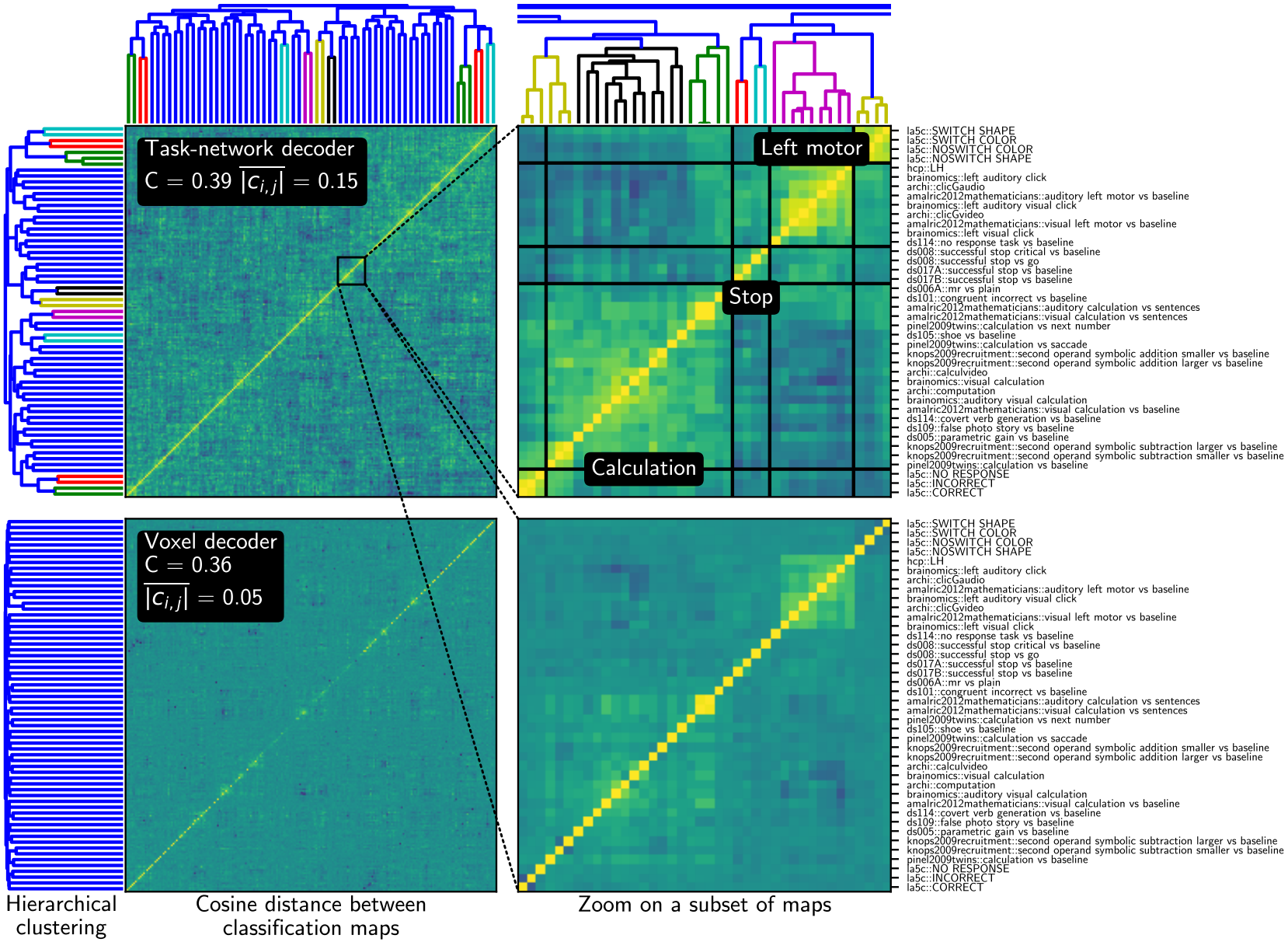
$$\boldsymbol{L} = \begin{bmatrix} \tilde{\boldsymbol{V}} \\ \mathbf{o} \in \mathbb{R}^{l-c \times k} \end{bmatrix} \in \mathbb{R}^{l \times k} \text{ and } \begin{bmatrix} \boldsymbol{U}^1 \\ \vdots \\ \boldsymbol{U}^N \end{bmatrix} \triangleq \left[\boldsymbol{I}_c \in \mathbb{R}^{c \times c}, \mathbf{o} \in \mathbb{R}^{l-c \times l} \right], \tag{8}$$

where $\boldsymbol{I}_c$ is the identity matrix of $\mathbb{R}^{c \times c}$. $\boldsymbol{L}$ is thus split into row-blocks $(\tilde{\boldsymbol{V}}^j)_j$, dedicated to and learned on *single studies*. It follows from elementary considerations that the matrices $(\boldsymbol{L}, (\boldsymbol{U}^j, \boldsymbol{b}^j)_j)$ form a global minimizer of (4), that is

Role of dropout



Correlation between maps



Bibliography I

1. Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. & Munafò, M. R. Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nature Reviews Neuroscience* **14,** 365–376 (2013).
2. Mairal, J., Bach, F., Ponce, J. & Sapiro, G. Online Learning for Matrix Factorization and Sparse Coding. *Journal of Machine Learning Research* **11,** 19–60 (2010).
3. Newell, A. You Can't Play 20 Questions with Nature and Win: Projective Comments on the Papers of This Symposium. *Visual Information Processing,* 1–26 (1973).
4. Laird, A. R., Lancaster, J. J. & Fox, P. T. Brainmap. *Neuroinformatics* **3,** 65–77 (2005).

Bibliography II

5. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-Scale Automated Synthesis of Human Functional Neuroimaging Data. *Nature Methods* **8,** 665–670 (2011).

6. Schwartz, Y., Thirion, B. & Varoquaux, G. *Mapping Paradigm Ontologies to and from the Brain.* in *Advances in Neural Information Processing Systems* (2013), 1673–1681.

7. Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W. & Kross, E. An fMRI-Based Neurologic Signature of Physical Pain. *New England Journal of Medicine* **368,** 1388–1397 (2013).

Bibliography III

8. Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D. & Nichols, T. E. Meta-Analysis of Neuroimaging Data: A Comparison of Image-Based and Coordinate-Based Pooling of Studies. *NeuroImage* **45,** 810–823 (2009).

9. Xue, Y., Liao, X., Carin, L. & Krishnapuram, B. Multi-Task Learning for Classification with Dirichlet Process Priors. *Journal of Machine Learning Research* **8,** 35–63 (2007).

10. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15,** 1929–1958 (2014).

Bibliography IV

11. M., A., Mairal, J., Thirion, B. & Varoquaux, G. Extracting Universal Representations of Cognition across Brain-Imaging Studies. *arXiv:1809.06035 [stat.ML], journal article under review* (2018).

12. M., A. & Blondel, M. *Differentiable Dynamic Programming for Structured Prediction and Attention.* in *Proceedings of the International Conference on Machine Learning (ICML)* (2018).

13. M., A., Mairal, J., Thirion, B. & Varoquaux, G. Stochastic Subsampling for Factorizing Huge Matrices. *IEEE Transactions on Signal Processing* **66,** 113–128 (2018).

14. M., A., Mairal, J., Bzdok, D., Thirion, B. & Varoquaux, G. *Learning Neural Representations of Human Cognition Across Many fMRI Studies.* in *Advances in Neural Information Processing Systems (NIPS)* (2017).

Bibliography V

15. Dohmatob, E., M., A., Varoquaux, G. & Thirion, B. *Learning Brain Regions via Large-Scale Online Structured Sparse Dictionary Learning.* in *Advances in Neural Information Processing Systems (NIPS)* (2016).

16. M., A., Mairal, J., Thirion, B. & Varoquaux, G. *Dictionary Learning for Massive Matrix Factorization.* in *Proceedings of the International Conference on Machine Learning (ICML)* (2016).

17. M., A., Varoquaux, G. & Thirion, B. *Compressed Online Dictionary Learning for Fast fMRI Decomposition.* in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)* (2016).