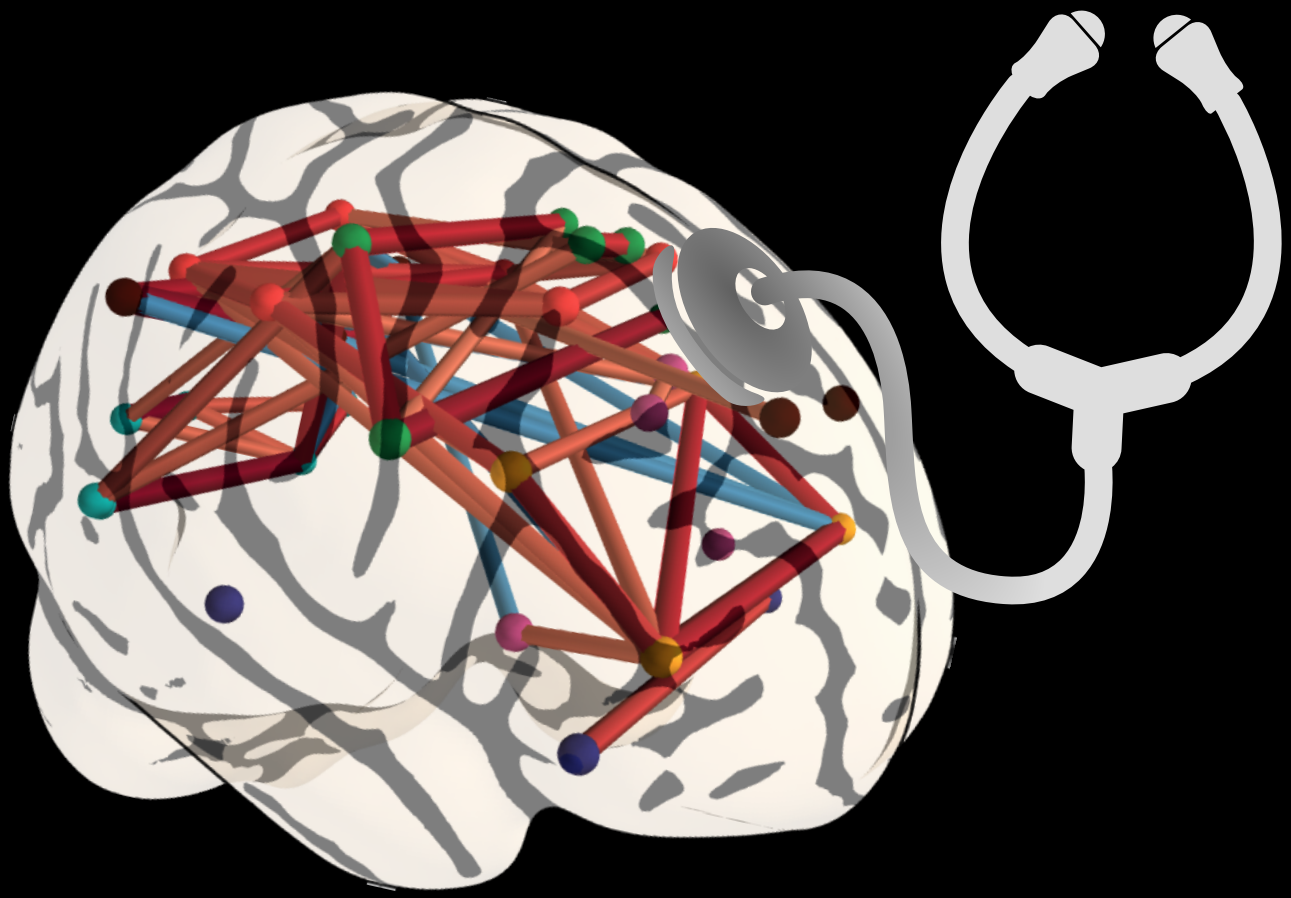
Detecting psychiatric disorders
with statistical learning tailored to brain activity

Gaël Varoquaux

PARIETAL

Inria



Psychiatric diseases are a major health challenge

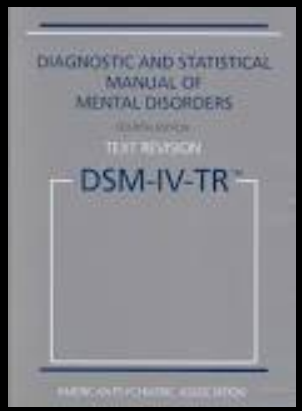
- $^1/_5$ American
- 2% are on the autistic spectrum
- 1% schizophrenia
- 7% severe depression
- Suicide is the leading causing of death in young people (15 to 29 years)

Psychiatric diseases are a major health challenge

- $^1/_5$ American
- 2% are on the autistic spectrum
- 1% schizophrenia
- 7% severe depression
- Suicide is the leading causing of death in young people (15 to 29 years)

Behavior, not biology, guides clinical practice

eg for diagnosis and prognosis



No progress — We need measurements

Psychiatric diseases are a major health challenge

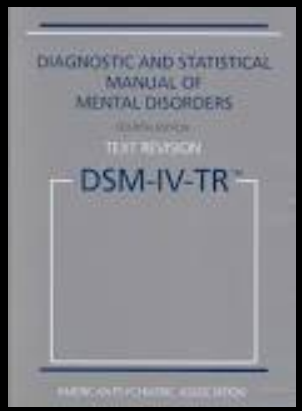
- $^1/_5$ American
- 2% are on the autistic spectrum
- 1%
- 7%
- Sui

Psychiatry biomarker extraction
From brain imaging
With supervised machine learning

eople
(19 to 29 years)

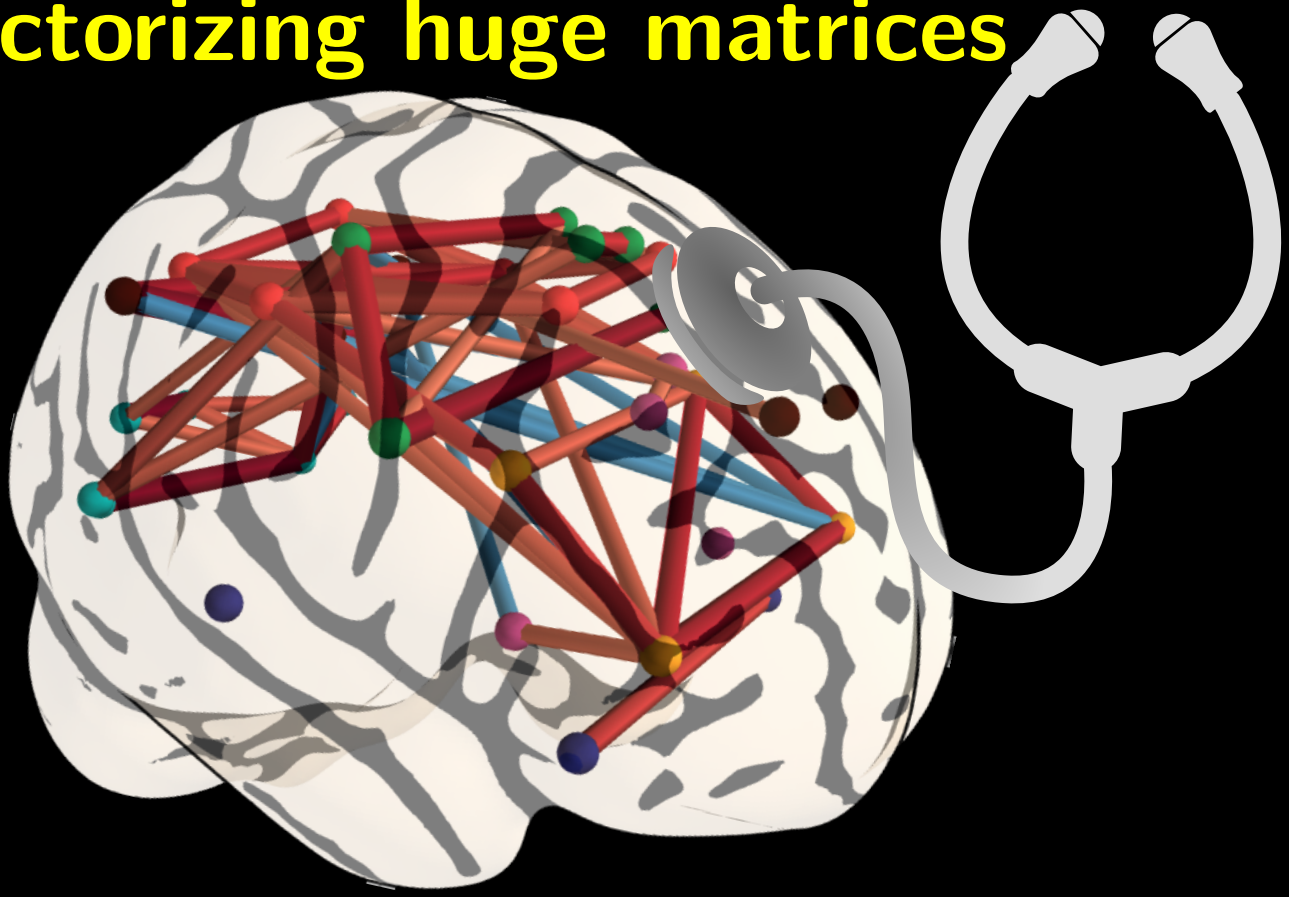
Behavior, not biology, guides clinical practice

eg for diagnosis and prognosis

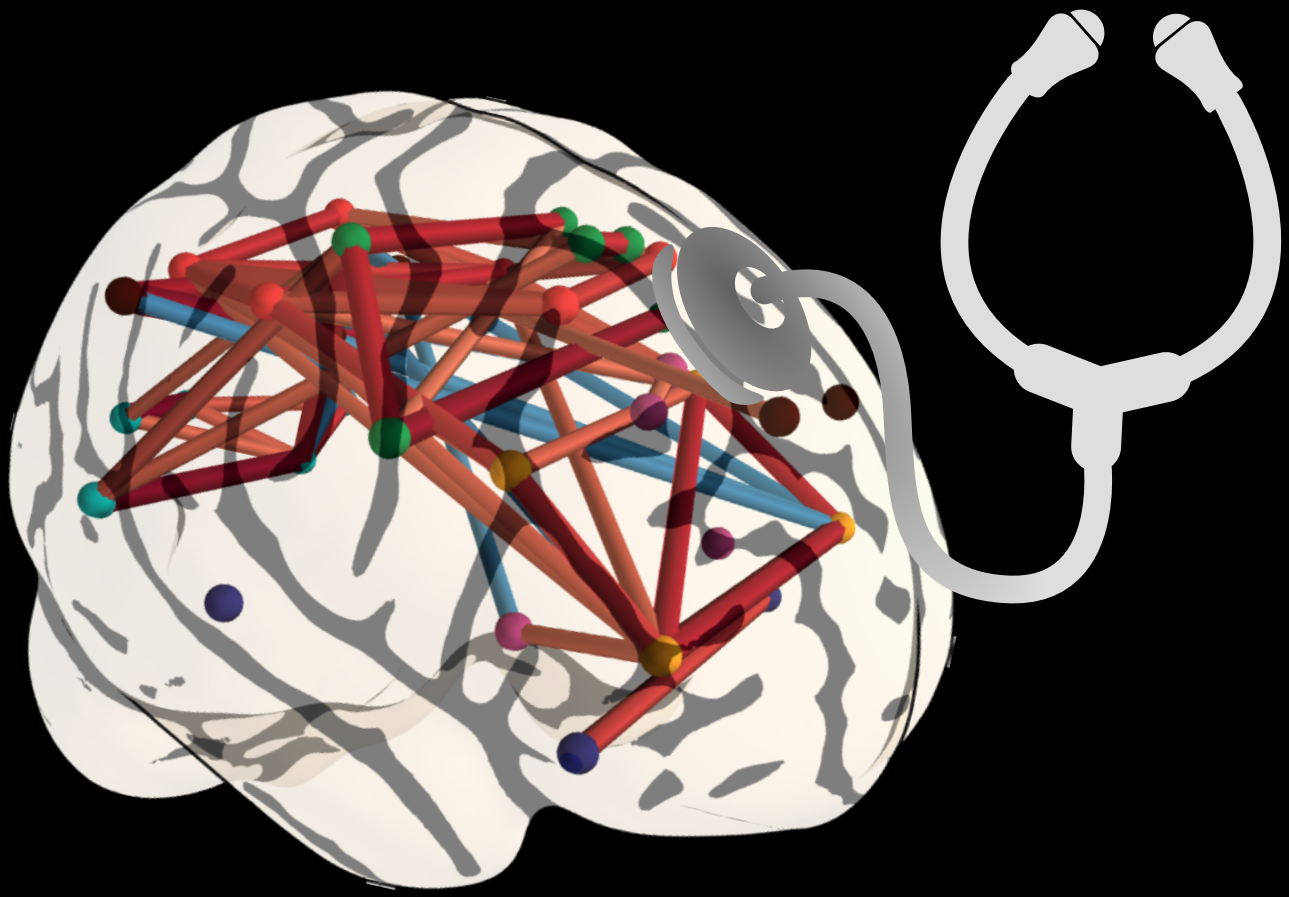


No progress — We need measurements

1. **Detecting psychiatric disorders**
2. **From activity at rest to biomarkers**
3. **Factorizing huge matrices**



1 Detecting psychiatric disorders



1 An open challenge on Autism prediction

Autism

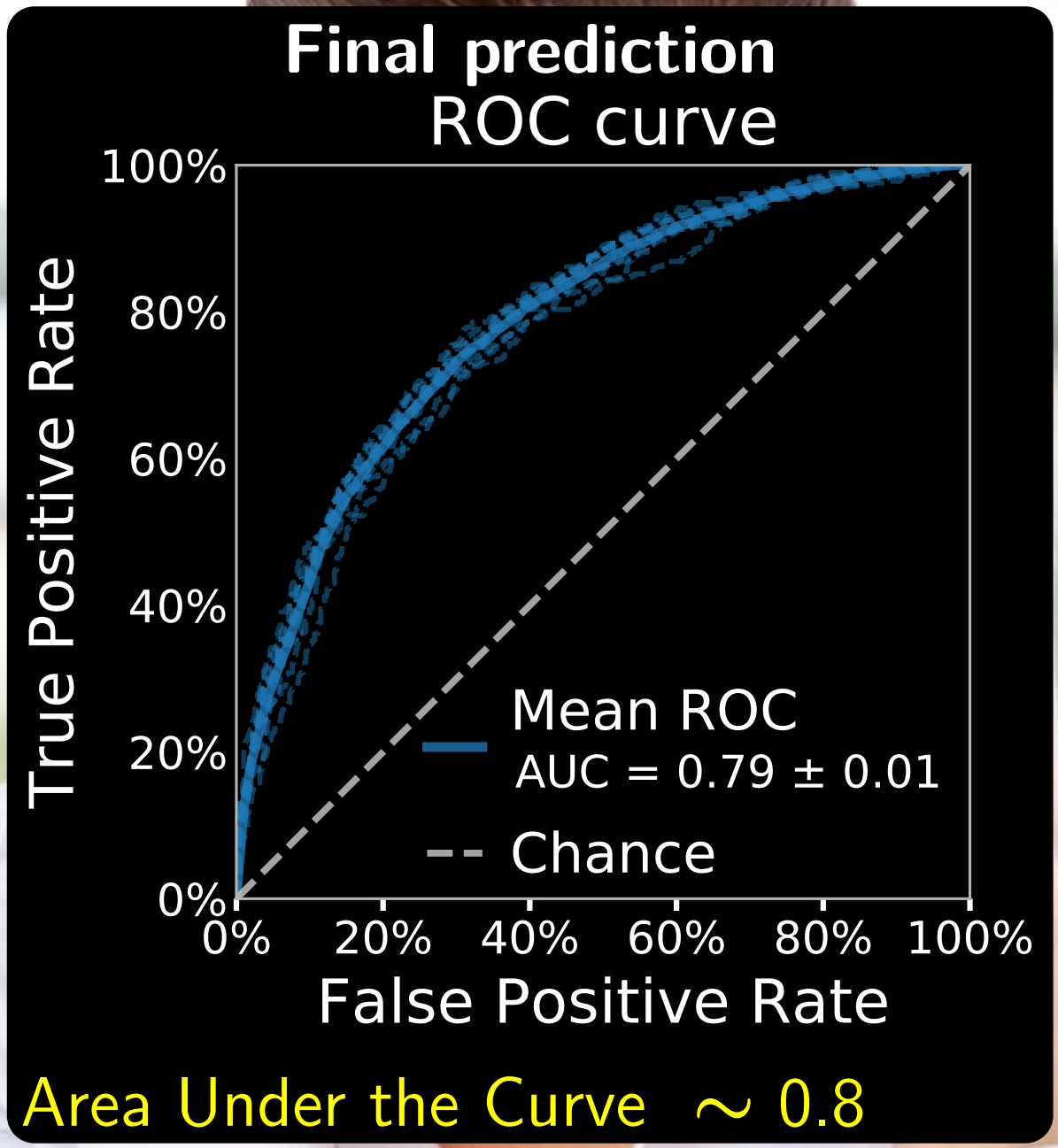
- A disease "of the mind"
- An imperfect diagnostic

The challenge

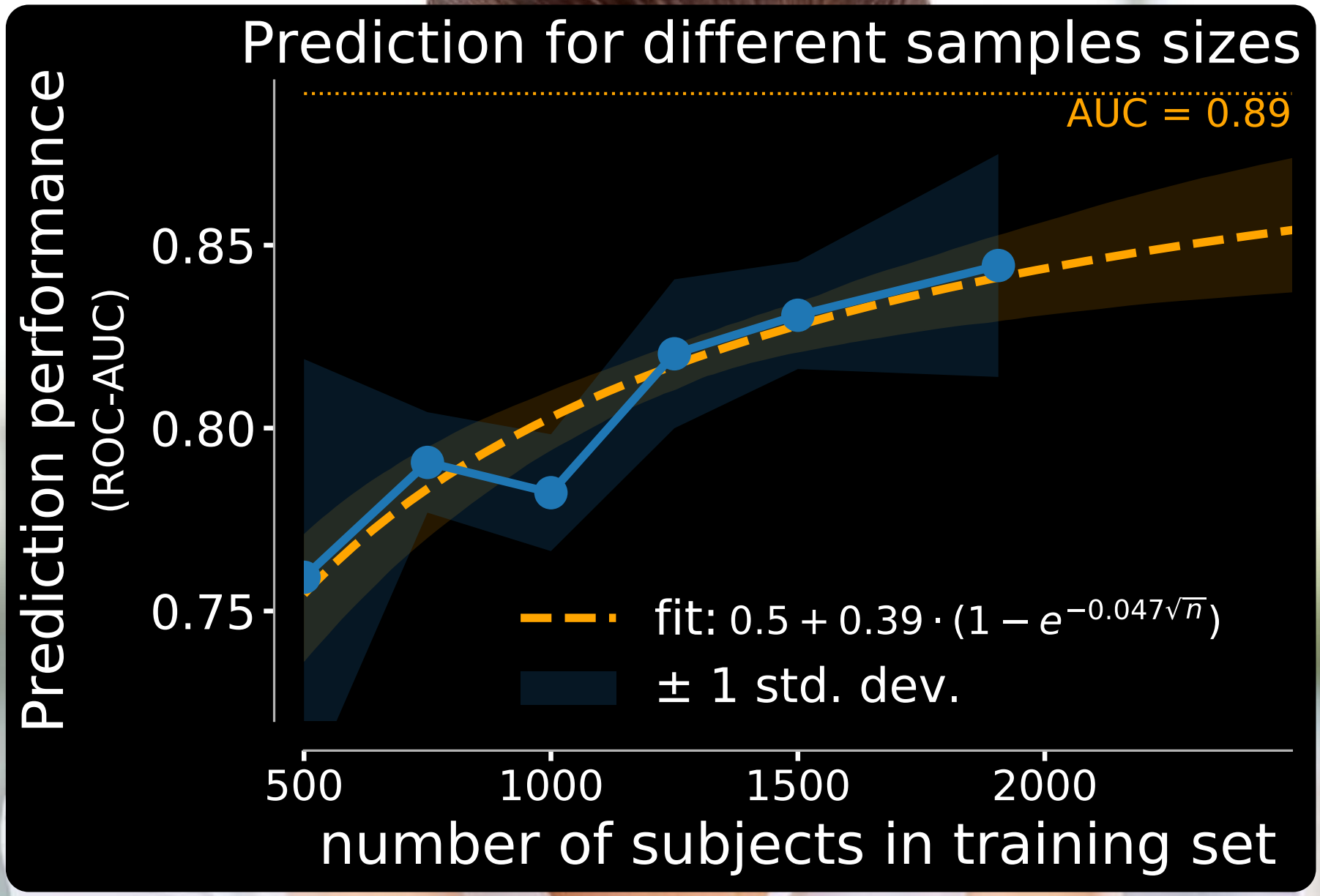
- Incentives: winner = 3000€
- Web-based: Participant submit code
- Hidden test set

Multimodal brain data:
Cortical thickness & brain activity at rest

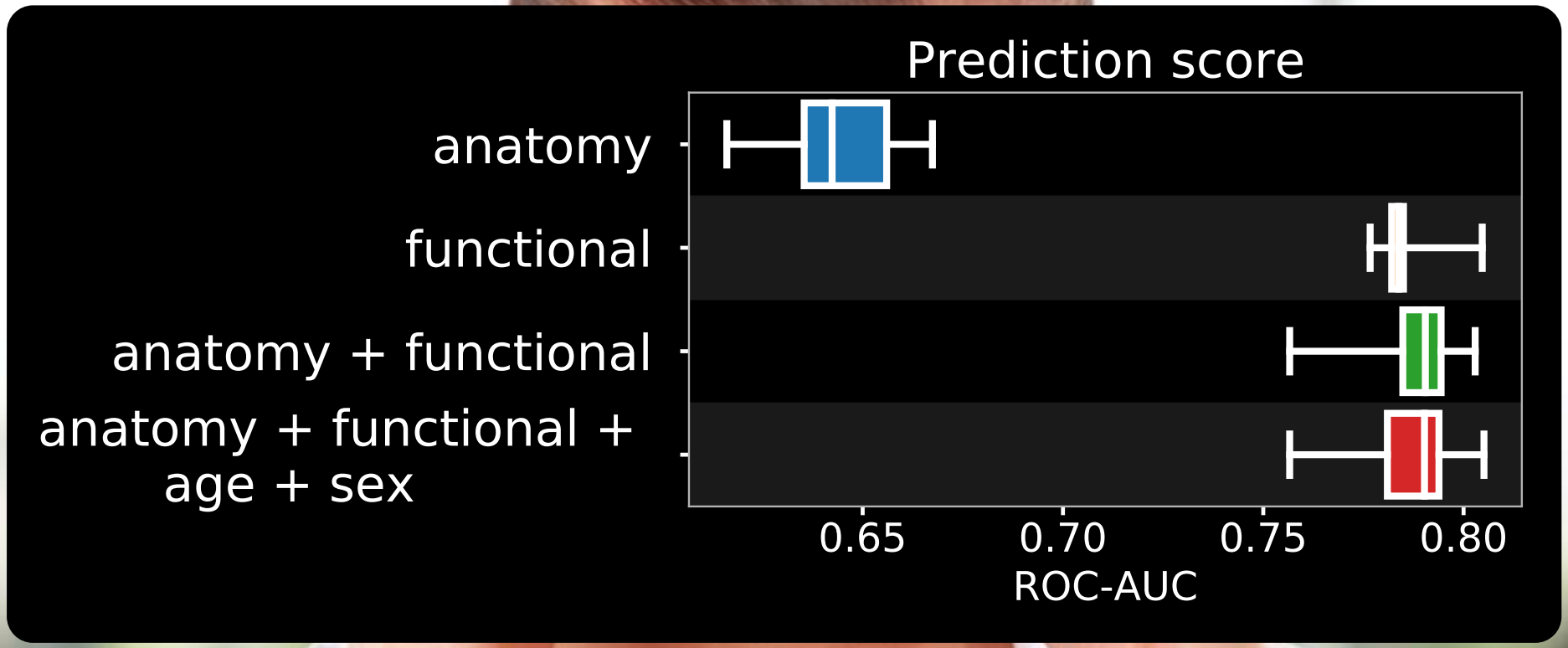
1 An open challenge on Autism prediction



1 More data is the way forward



1 Brain activity at rest is a marker



- Autism is a “spectrum disorder”
  different causes under the same symptoms
- Diagnostic is imperfect
  Some labels are wrong
- We are probably only seeing the easy cases

1 Noisy labels are still useful

Predicting brain aging $\neq$ chronological age

- Predicts age with a mean absolute error of **4.3 years**

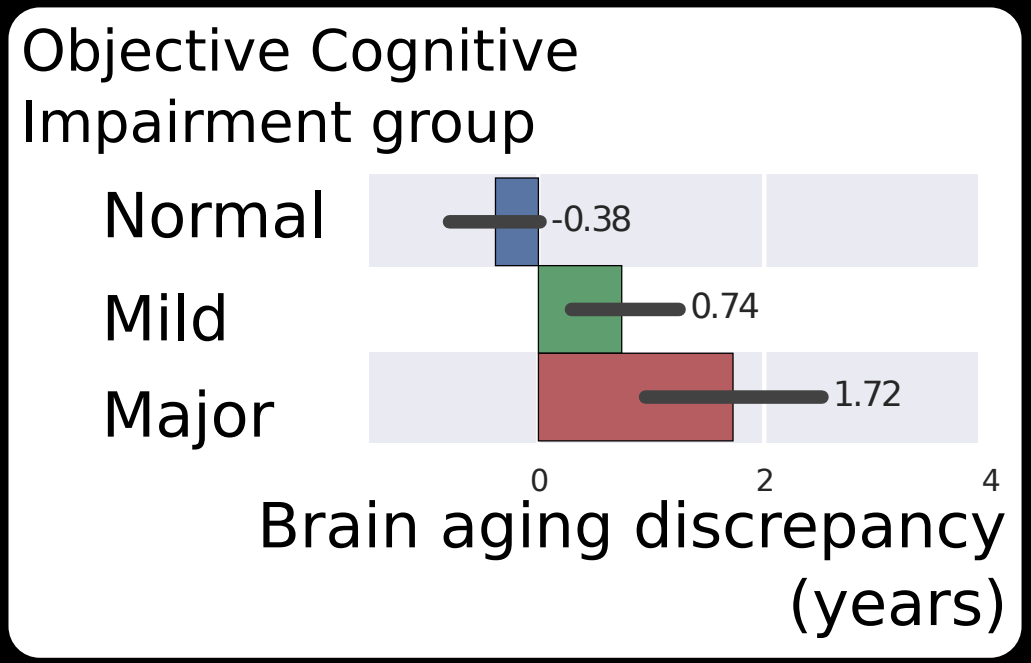
[Liem... 2016]

1 Noisy labels are still useful

Predicting brain aging $\neq$ chronological age

- Predicts age with a mean absolute error of **4.3 years**

- Discrepancy with chronological age
  correlates with cognitive impairment



[Liem... 2016]

Biomarker

surrogate,
but useful

1 Capturing subjects psychological traits

An individual should not be reduced to
a single diagnostic or behavioral quantity

[Rahim... 2017]

Multi-output prediction

Predict jointly multiple individual phenotypes

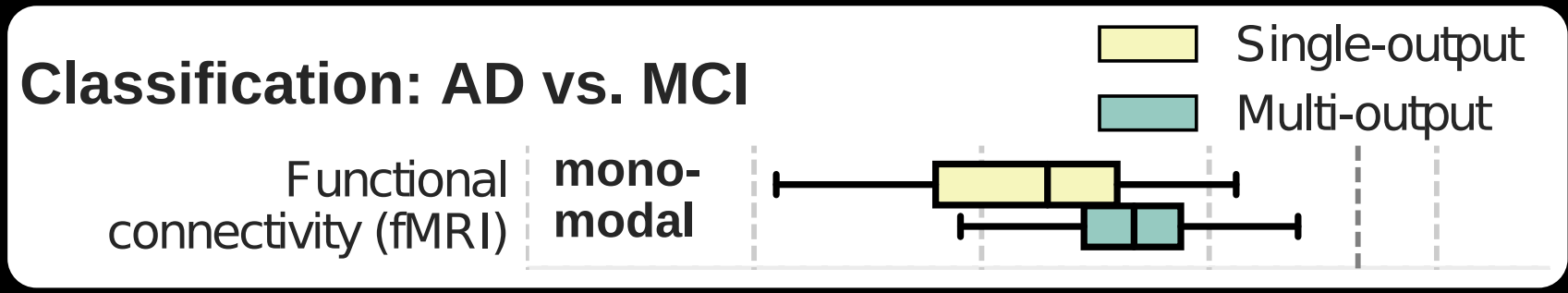
- behavioral scores
- diagnostic status

They improve eachother's prediction

MMSE: mini mental-state examination

A diagnostic exam for Alzheimer's Disease

Adding MMSE as a **target** improves AD prediction



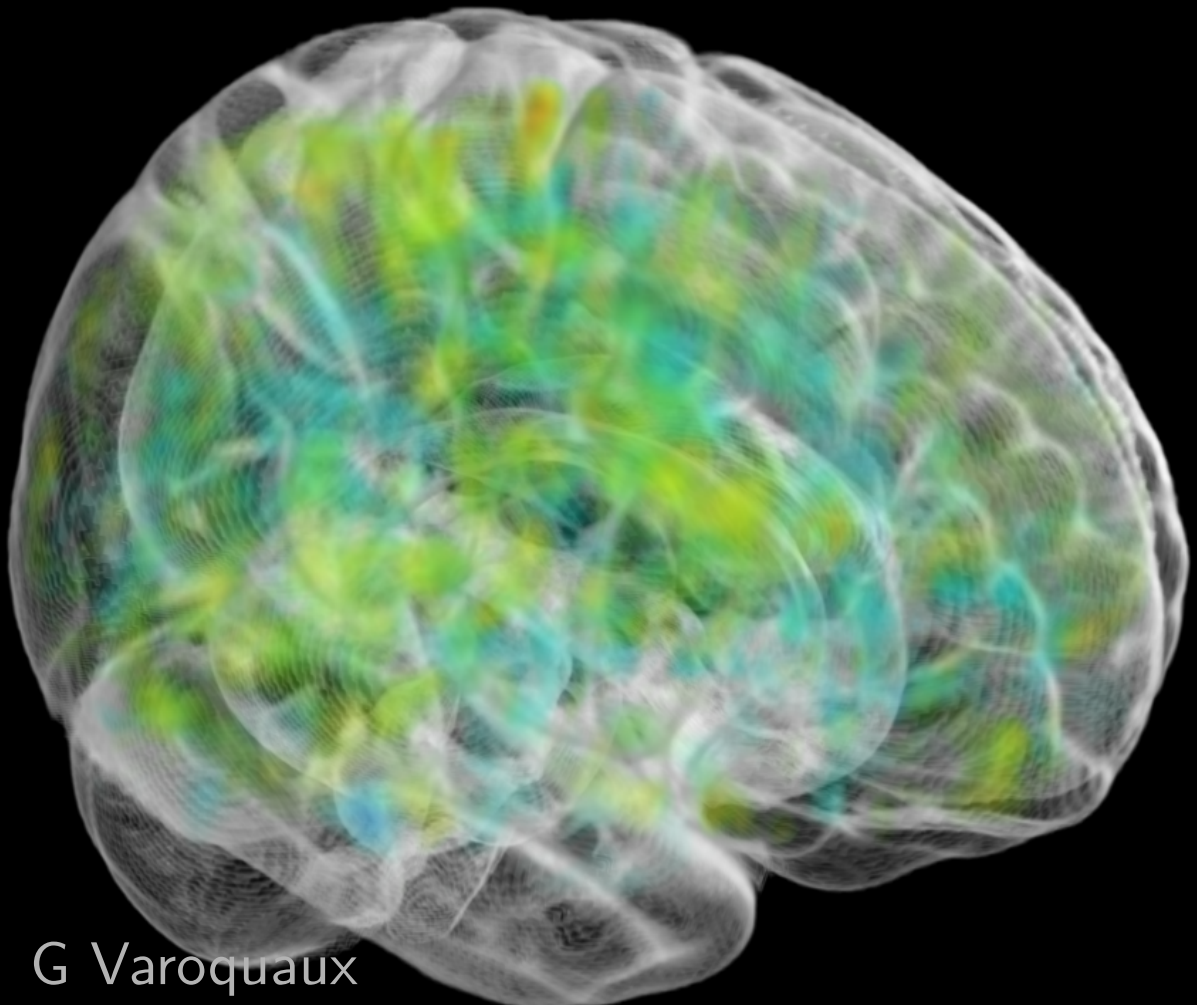
[Rahim... 2017]

Detecting psychiatric disorders

- Supervised learning on rest fMRI
- Across thousands of subjects
- Labels are wrong but useful
  to define *surrogate biomarkers*
- Predicting multiple dimensions of individual psychology

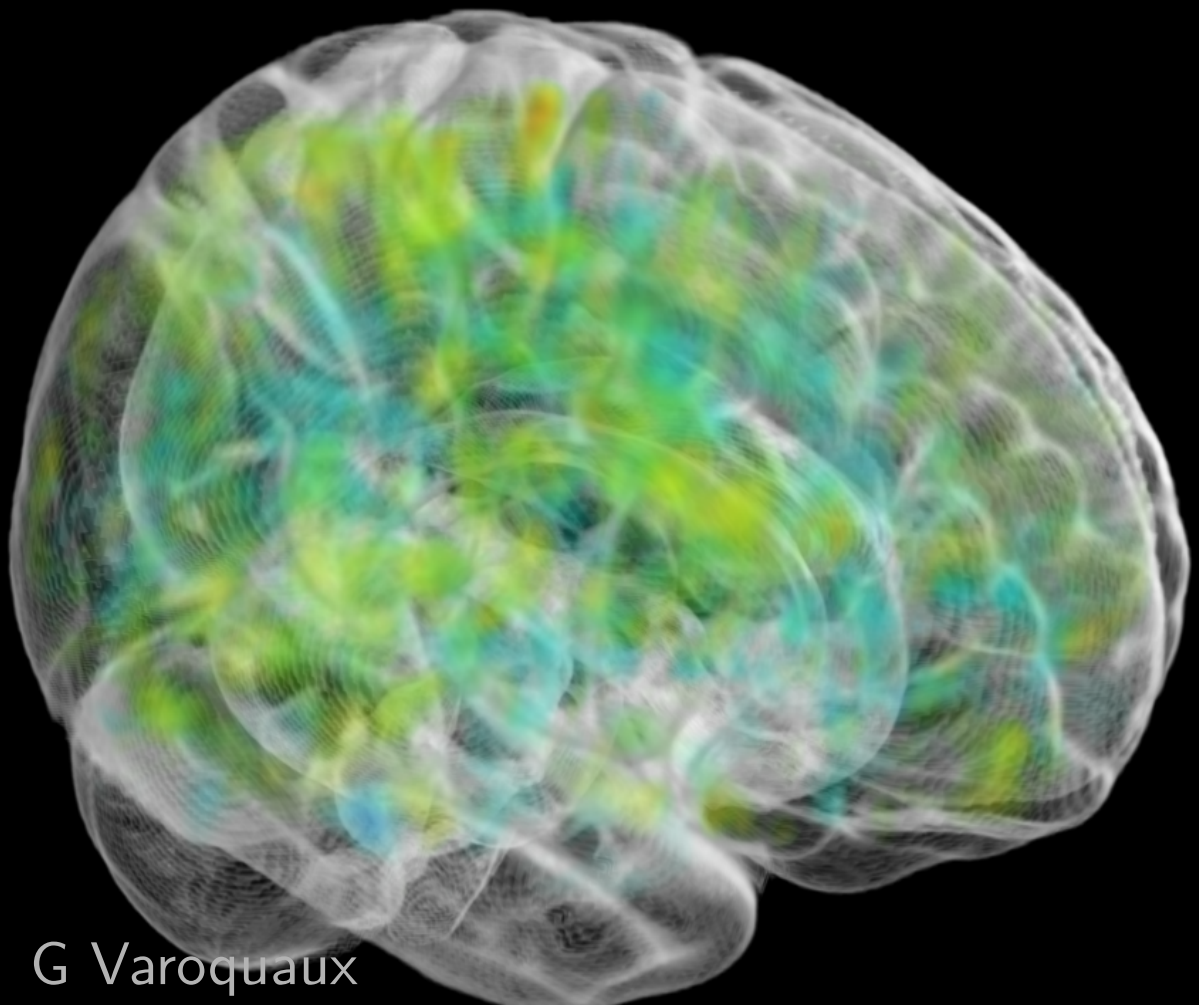
Define traits based on "biology"

2 From activity at rest to biomarkers



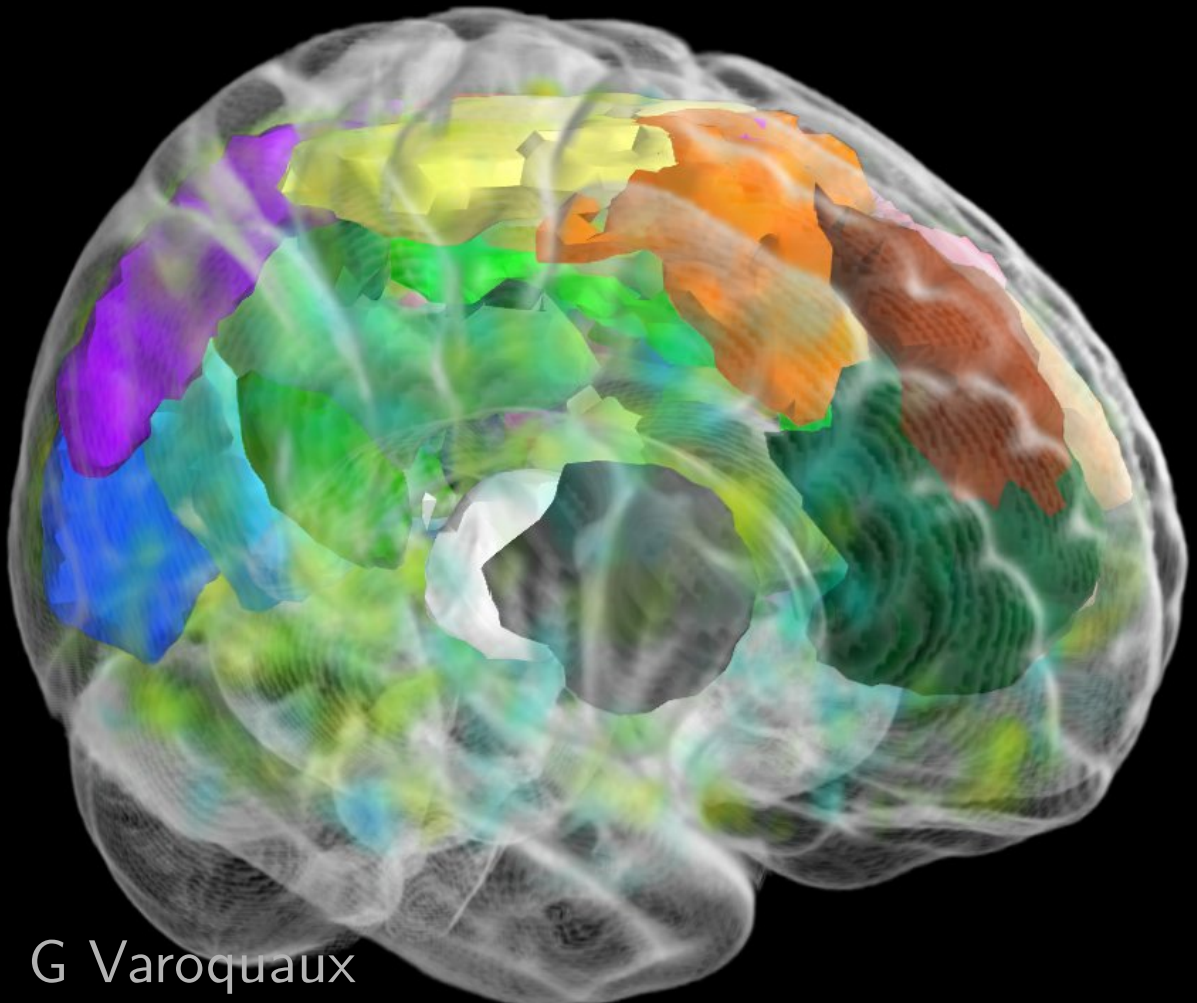
From rest-fMRI to biomarkers

No salient features in rest fMRI



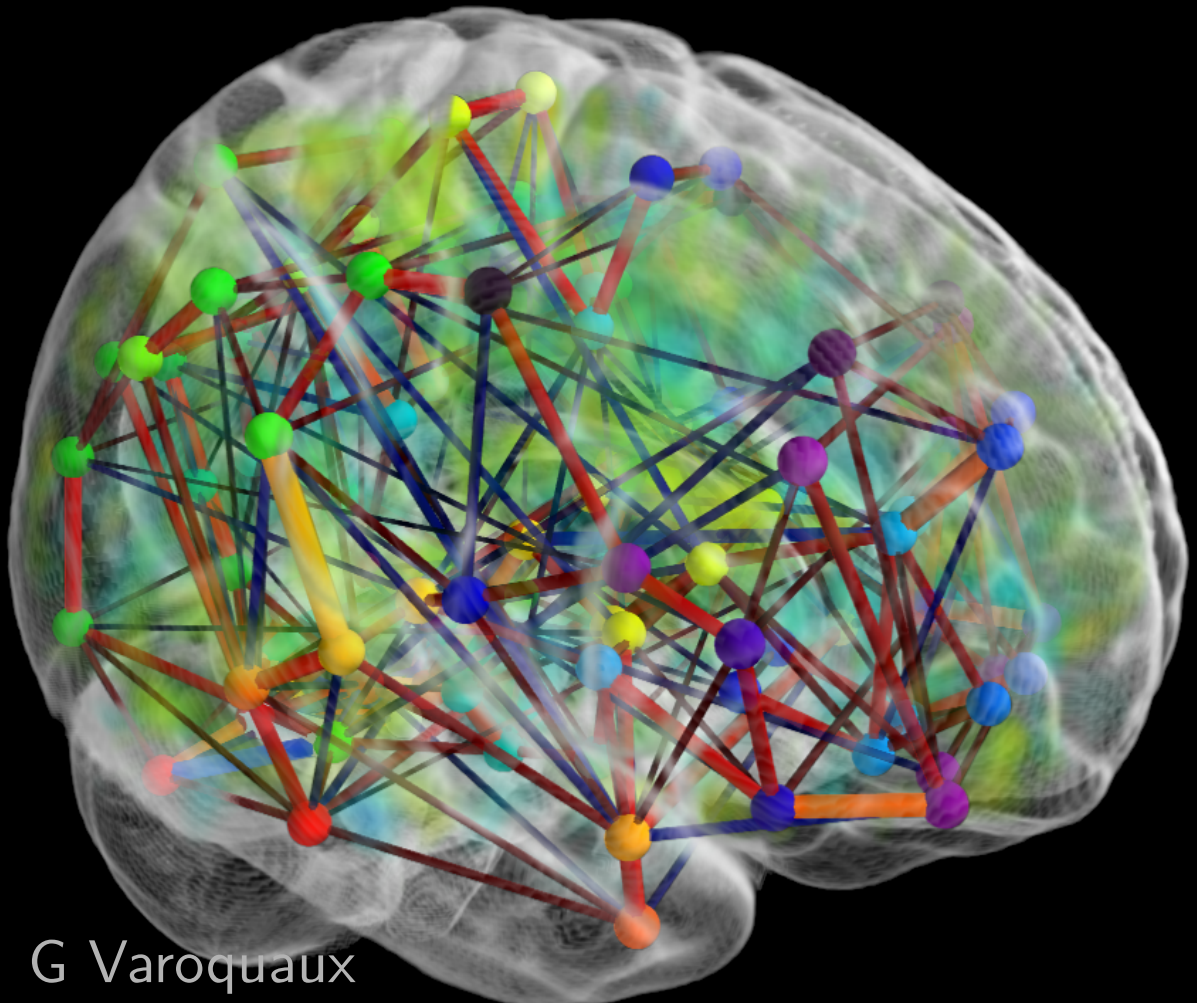
From rest-fMRI to biomarkers

- Define functional regions

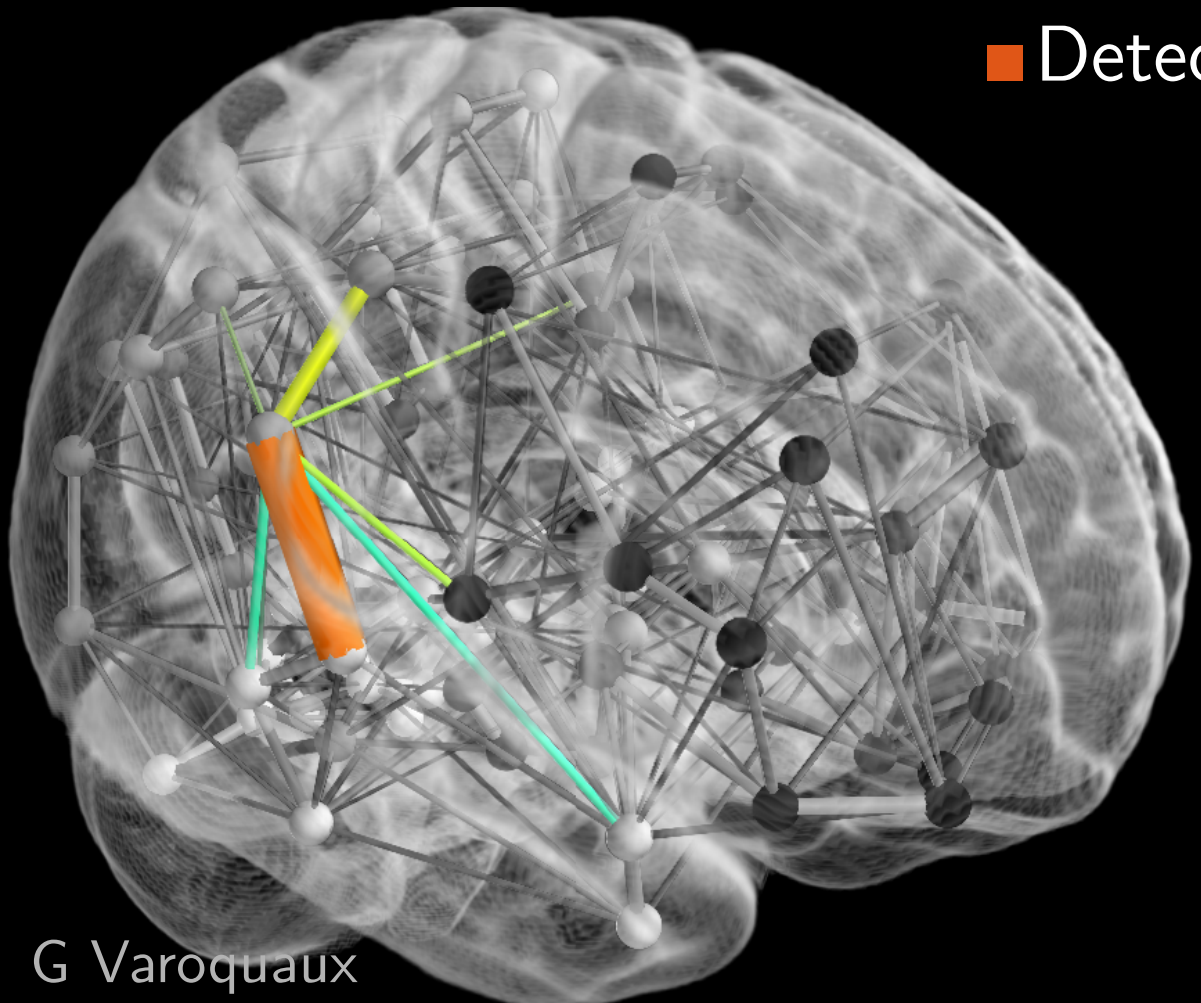


From rest-fMRI to biomarkers

- Define functional regions
- Learn interactions

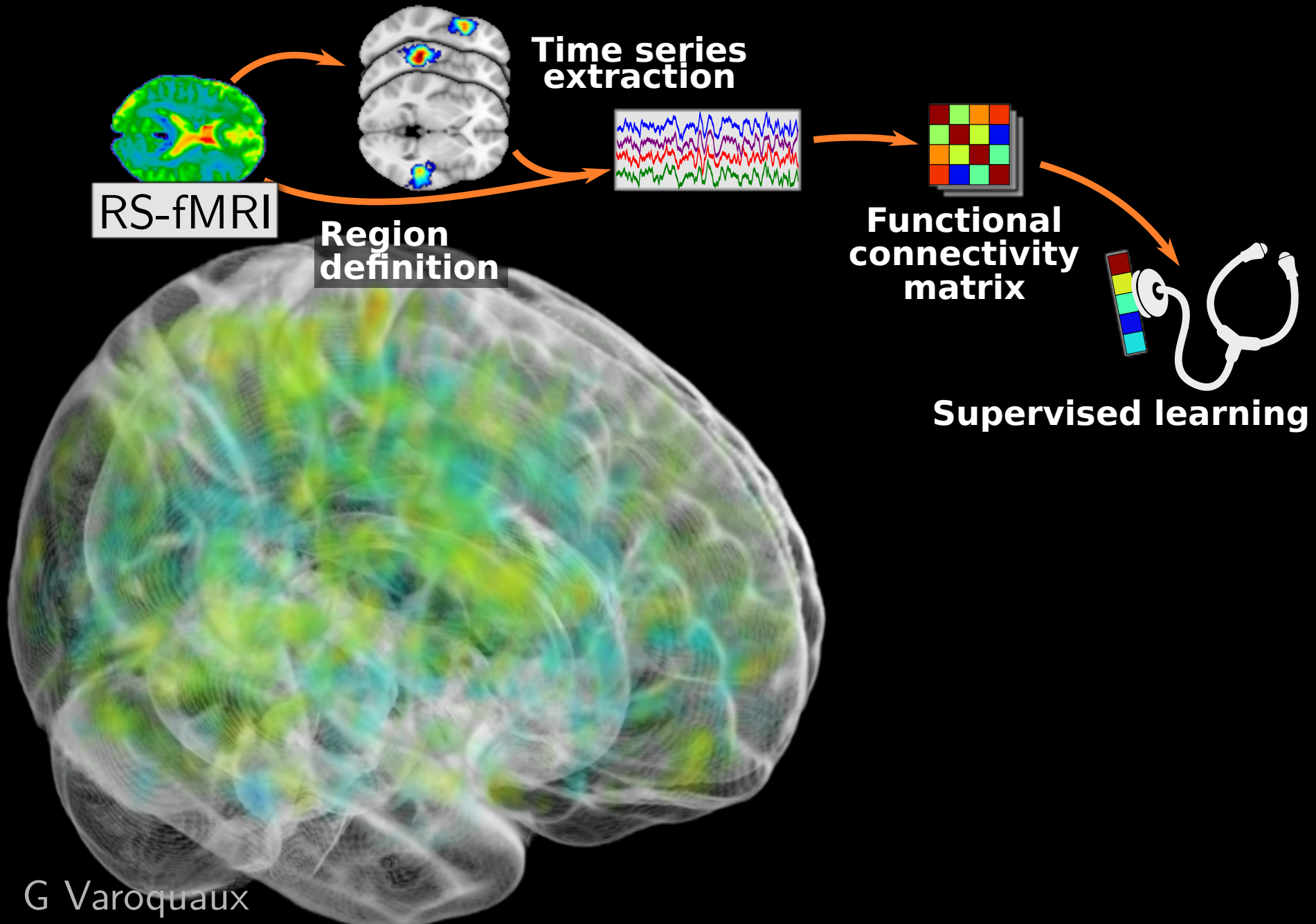


From rest-fMRI to biomarkers



- Define functional regions
- Learn interactions
- Detect differences

From rest-fMRI to biomarkers



Defining functional regions

Dividing the brain in regions

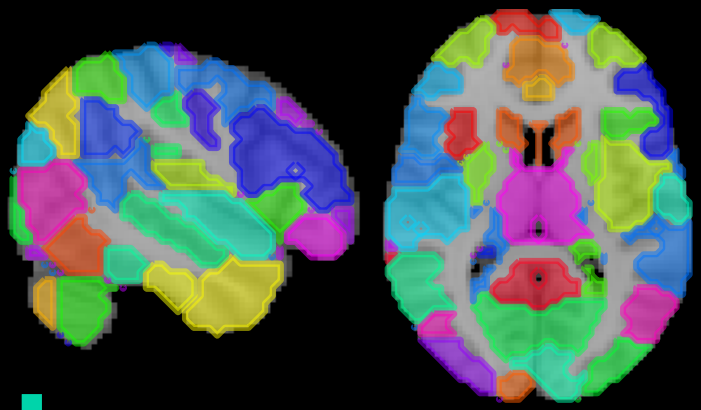
anatomical atlases, functional atlases, region extraction methods

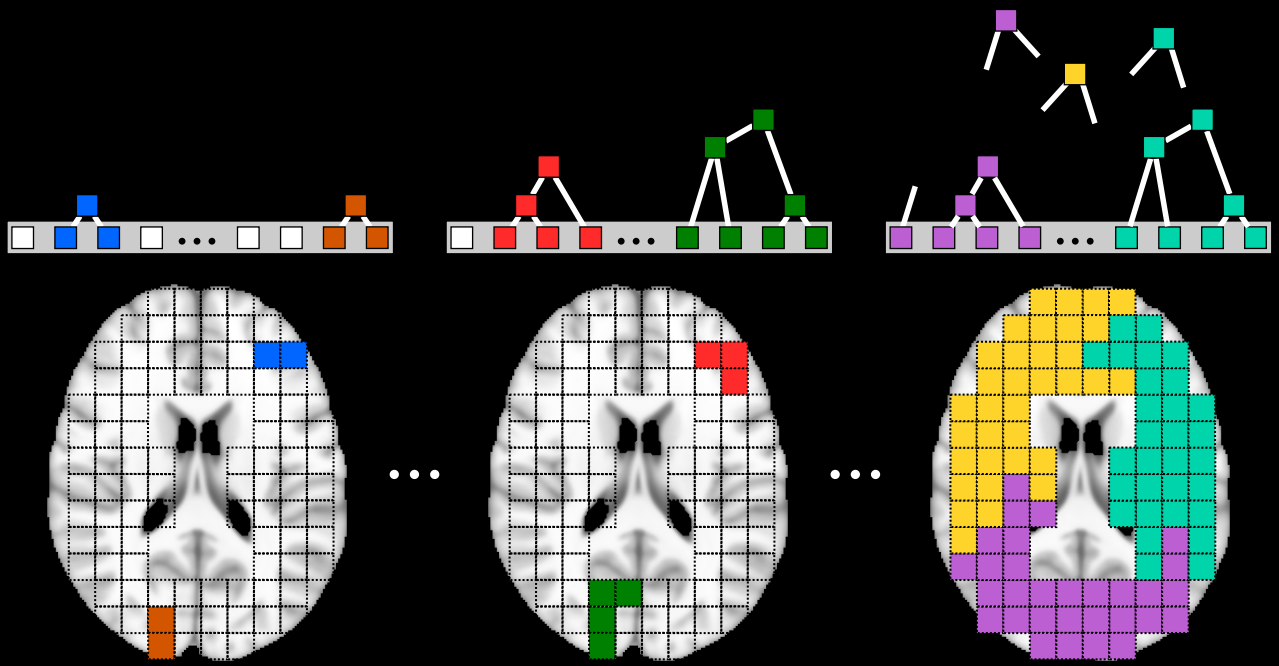
Some examples

Clustering

- k-means
- ward

[Thirion... 2014]

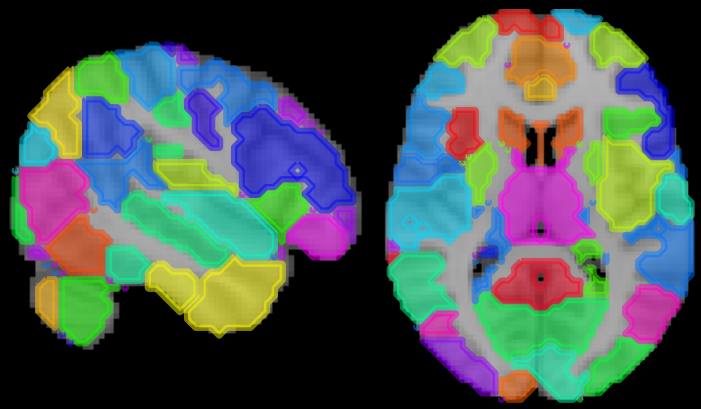




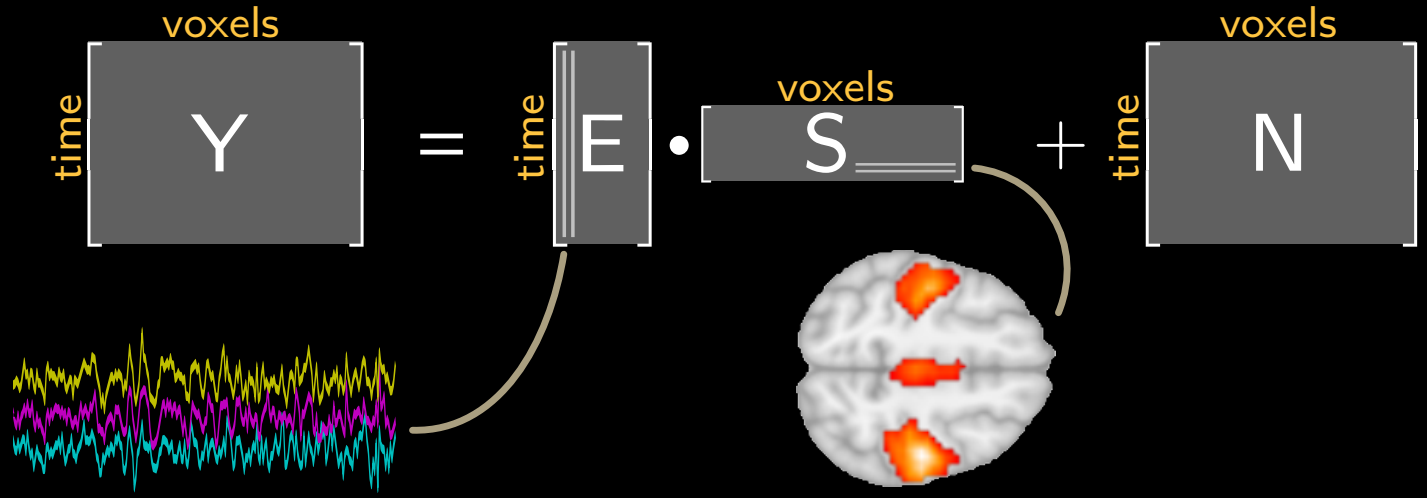
Clustering

- k-means
- ward

[Thirion... 2014]



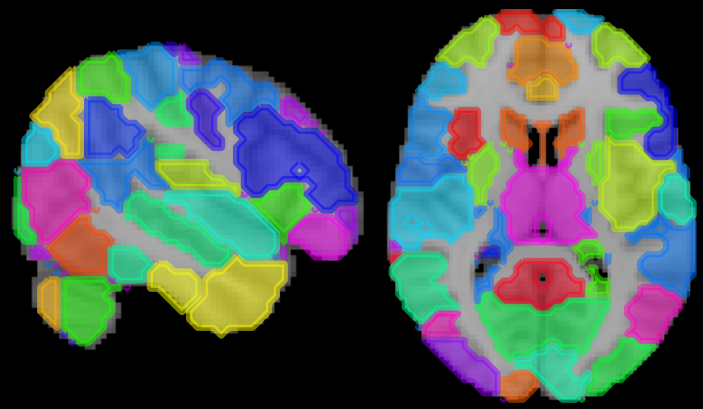
Decomposition models



Clustering

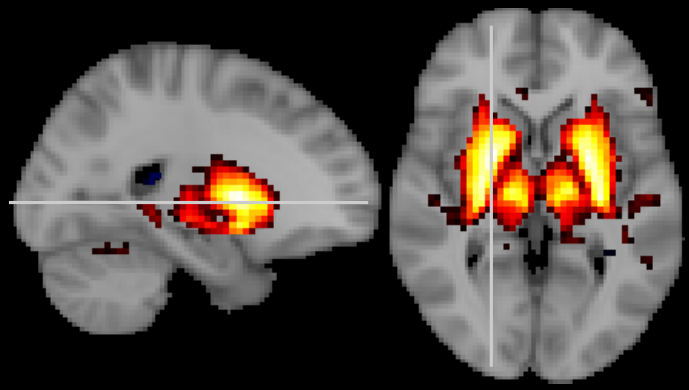
- k-means
- ward

[Thirion... 2014]

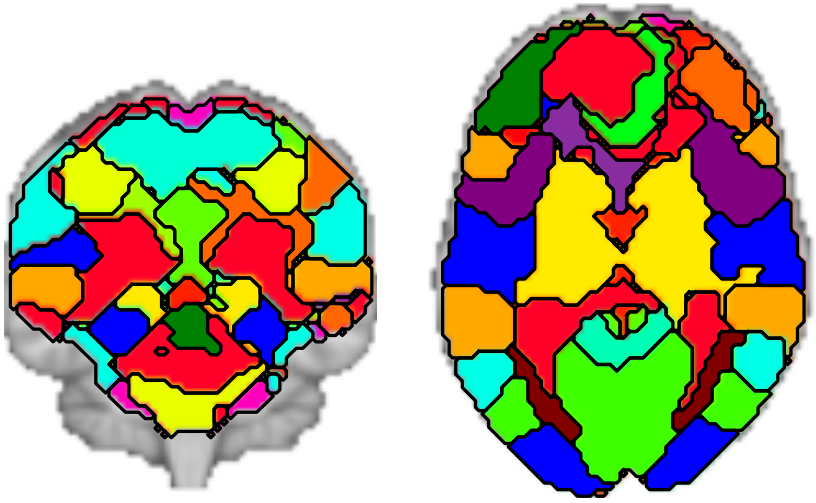


Decomposition models

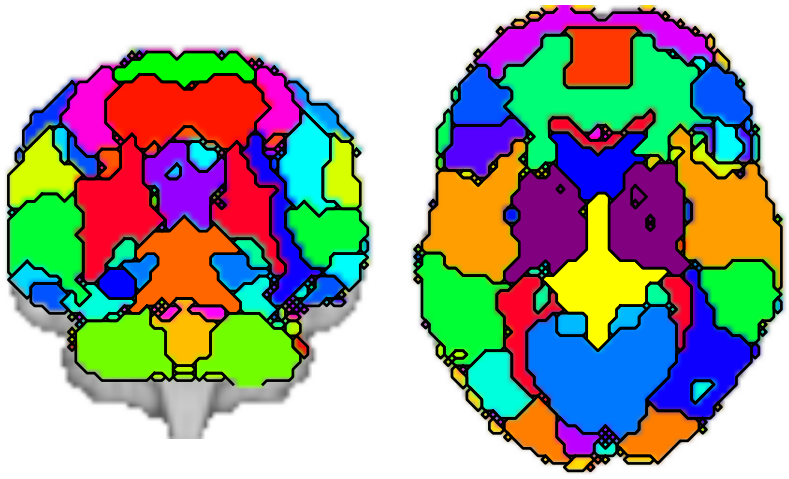
- ICA: seek independence of maps
- Sparse dictionary learning: seek sparse maps



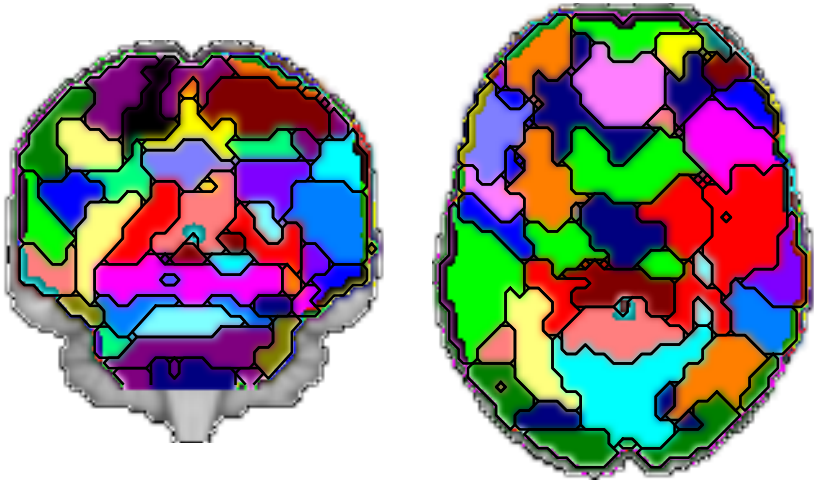
2 Region definition: resulting parcellations



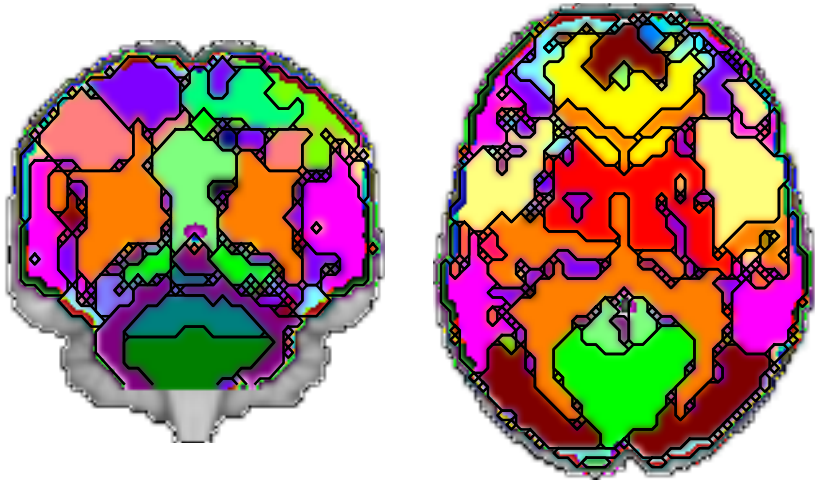
Dictionary learning



Group ICA

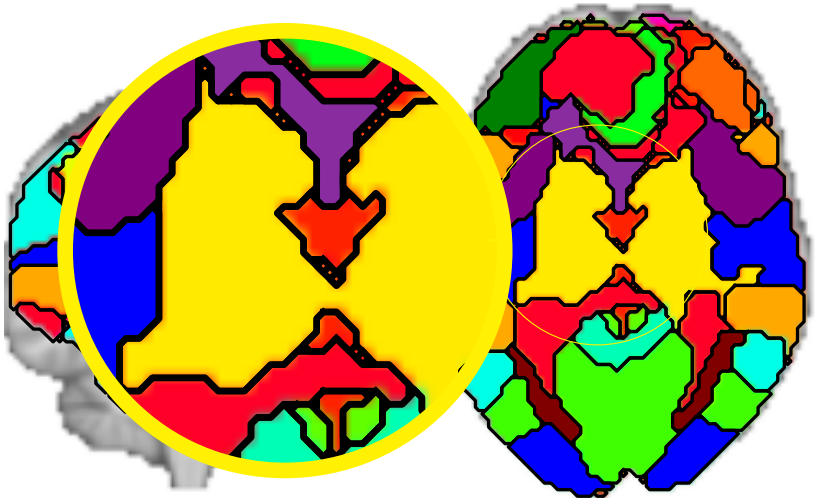


Ward clustering

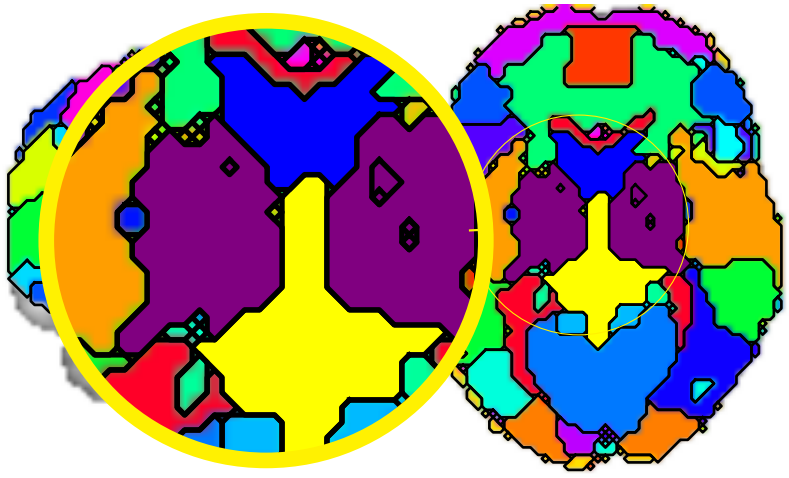


K-Means clustering

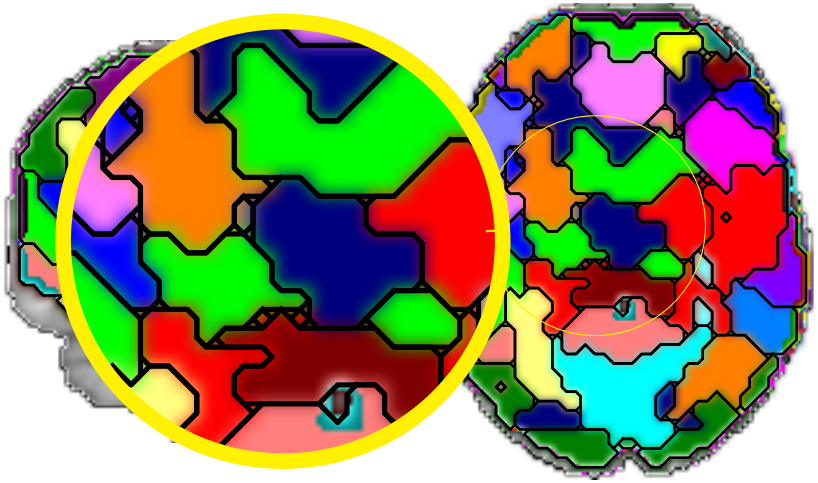
2 Region definition: resulting parcellations



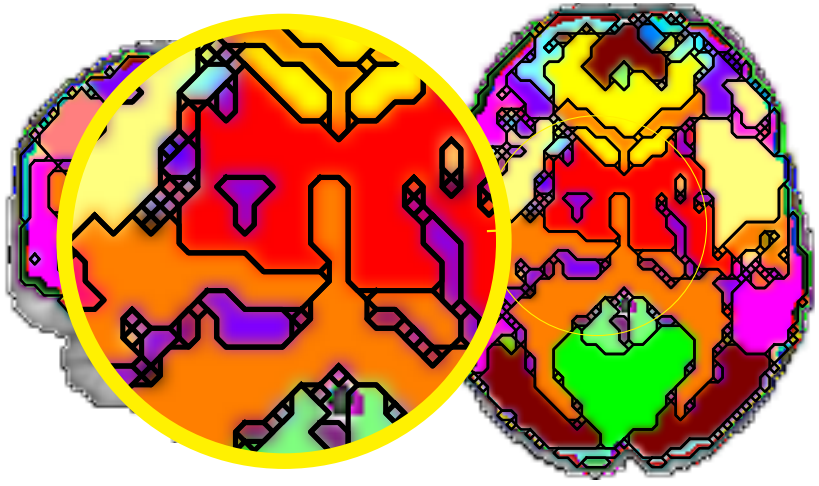
Dictionary learning



Group ICA

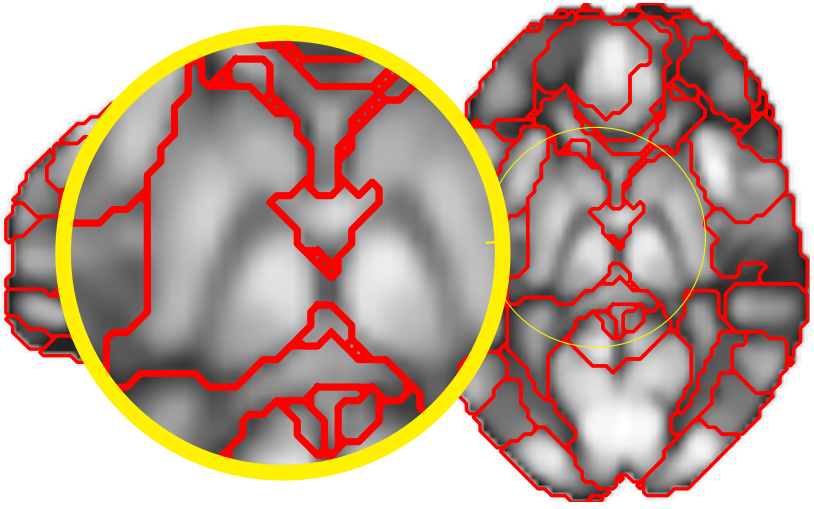


Ward clustering

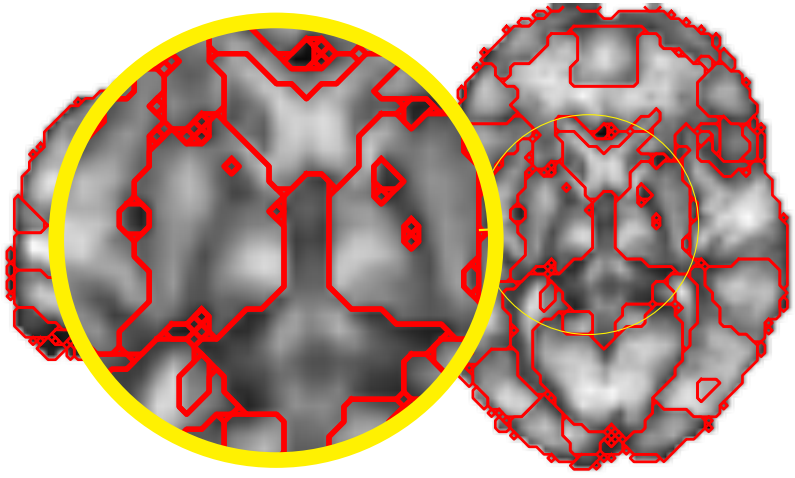


K-Means clustering

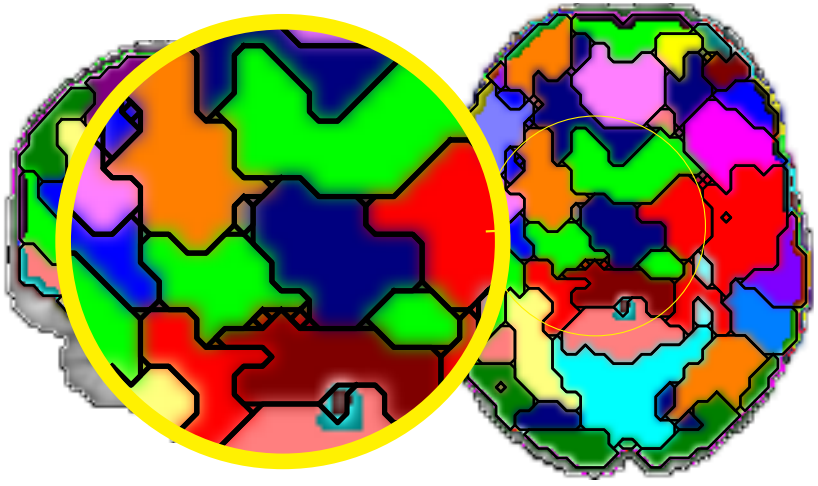
2 Region definition: resulting parcellations



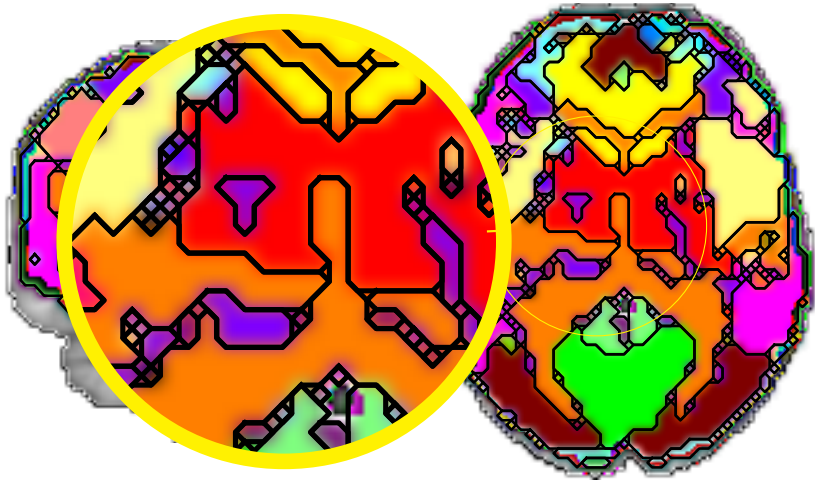
Dictionary learning



Group ICA

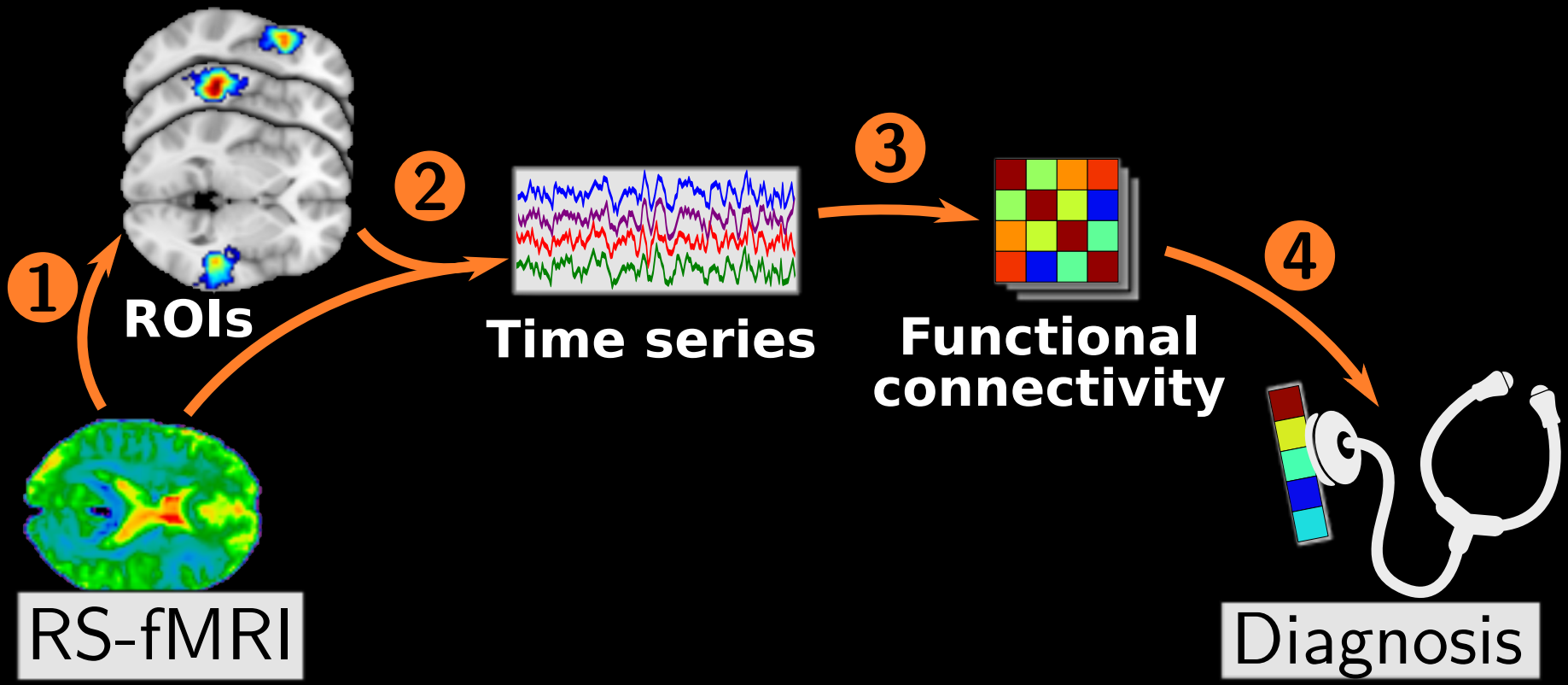


Ward clustering

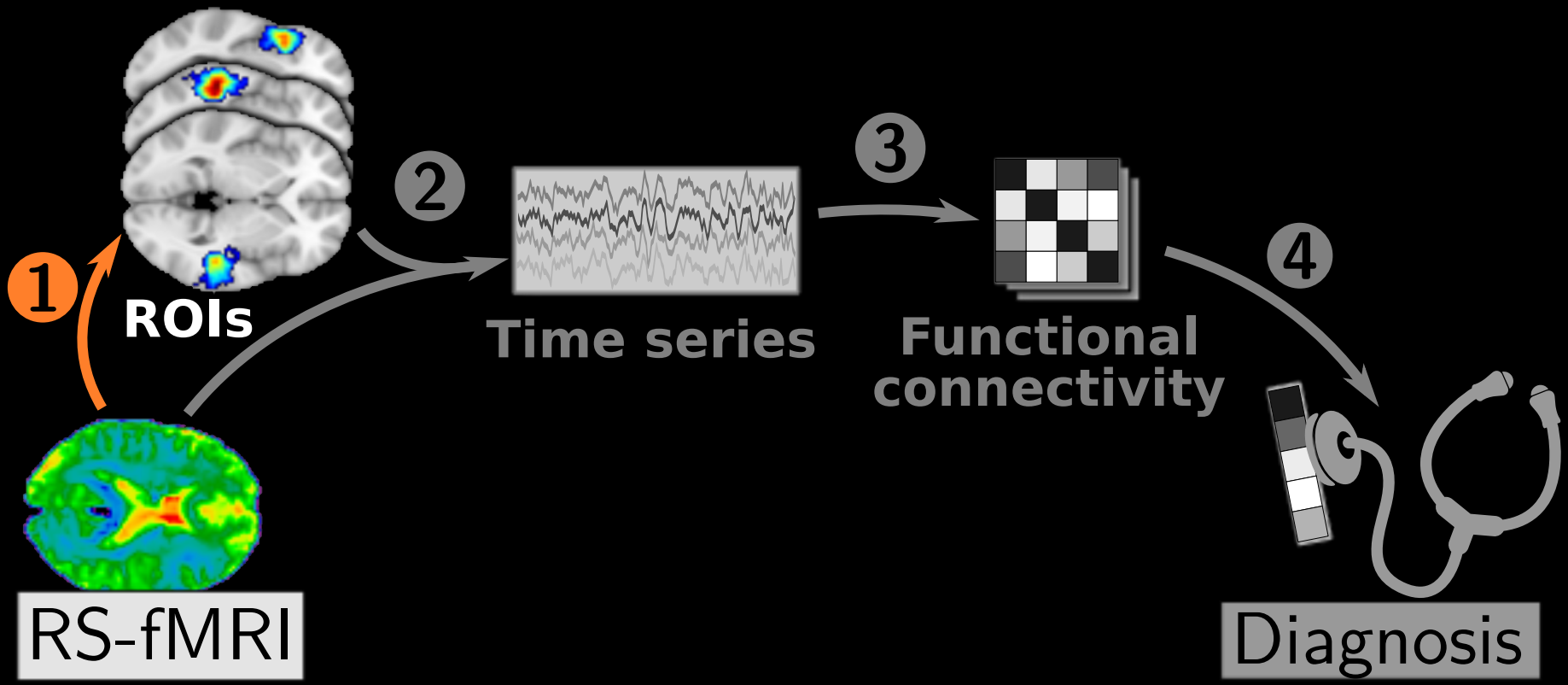


K-Means clustering

2 In connectome prediction settings



2 In connectome prediction settings



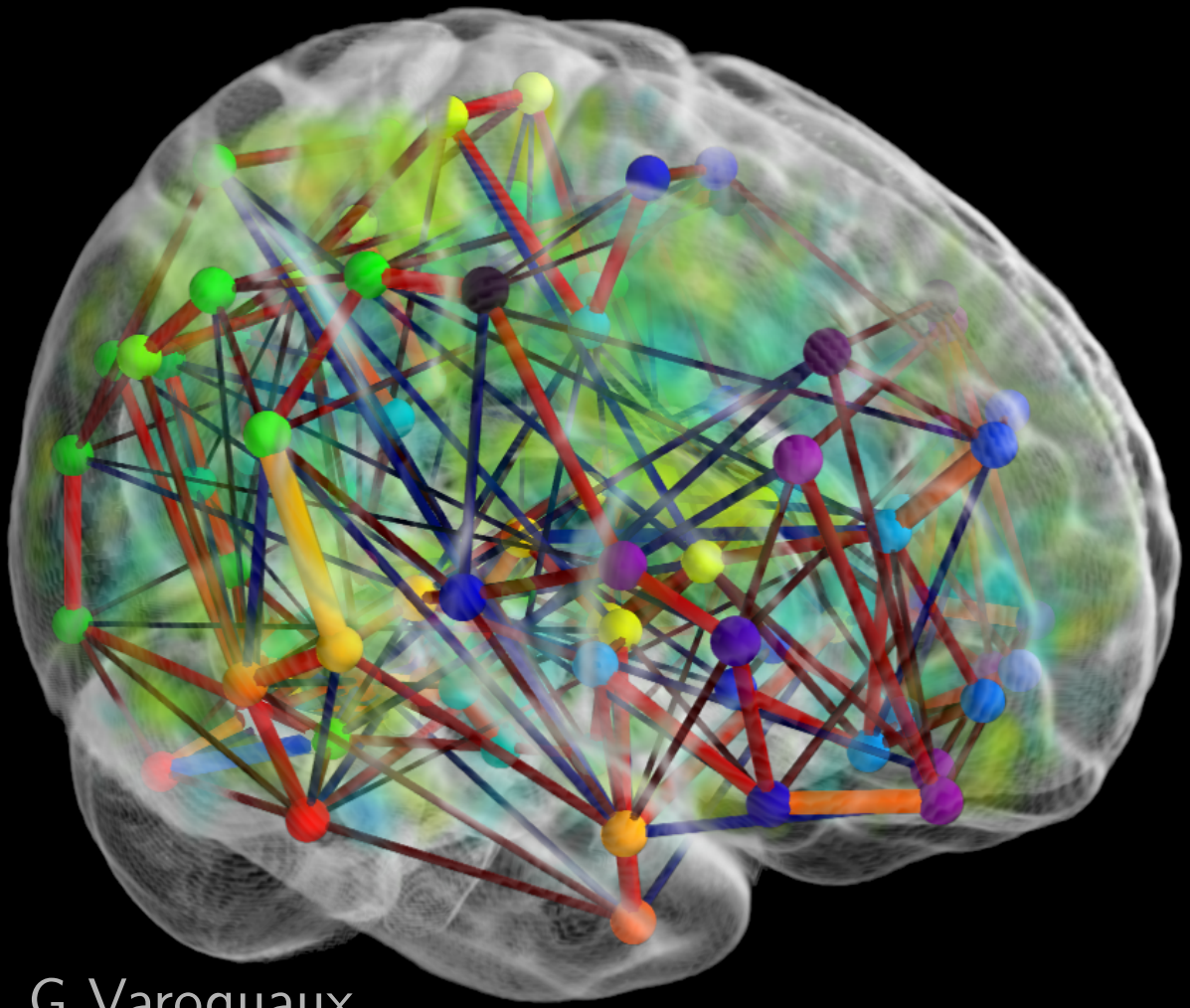
Best choice of regions for prediction

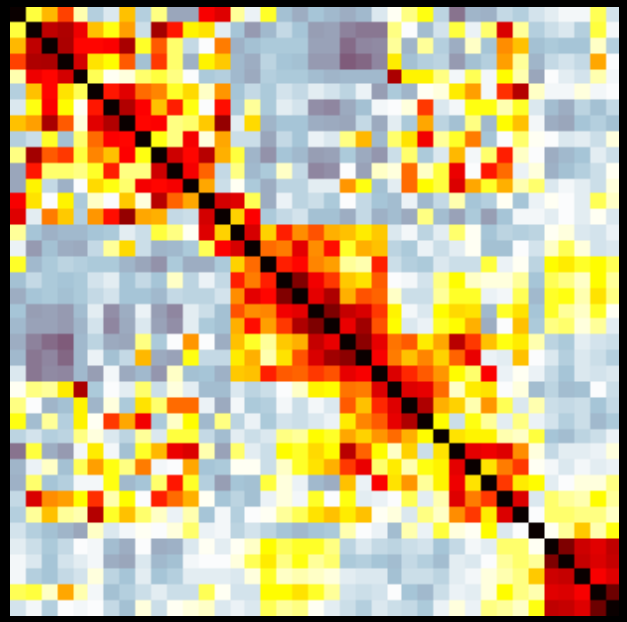
- Defining regions functionally is important
- Decomposition methods work best

[Reddy in rev, ArXiv]

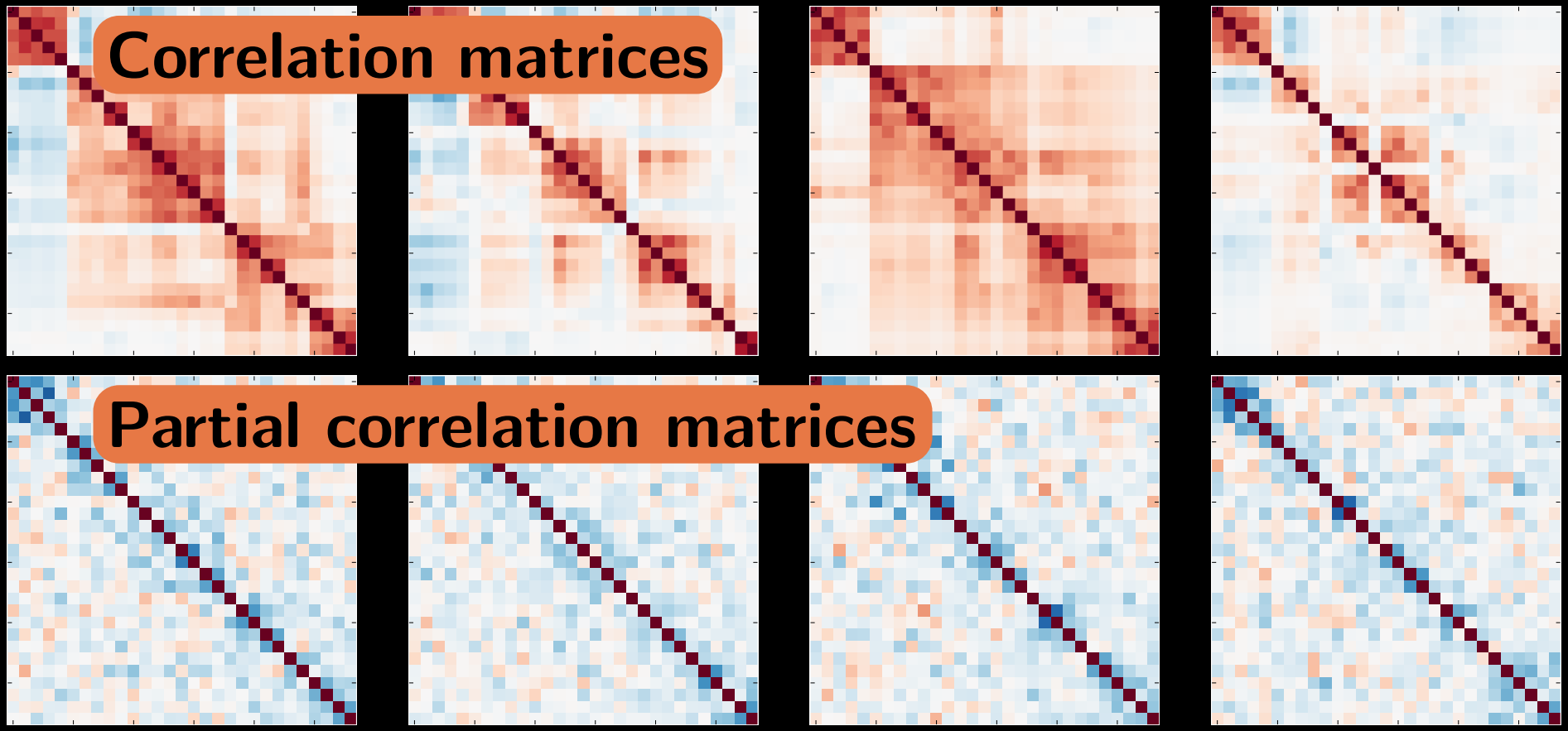
2 Connectome: building a connectivity matrix

How to capture and represent interactions?





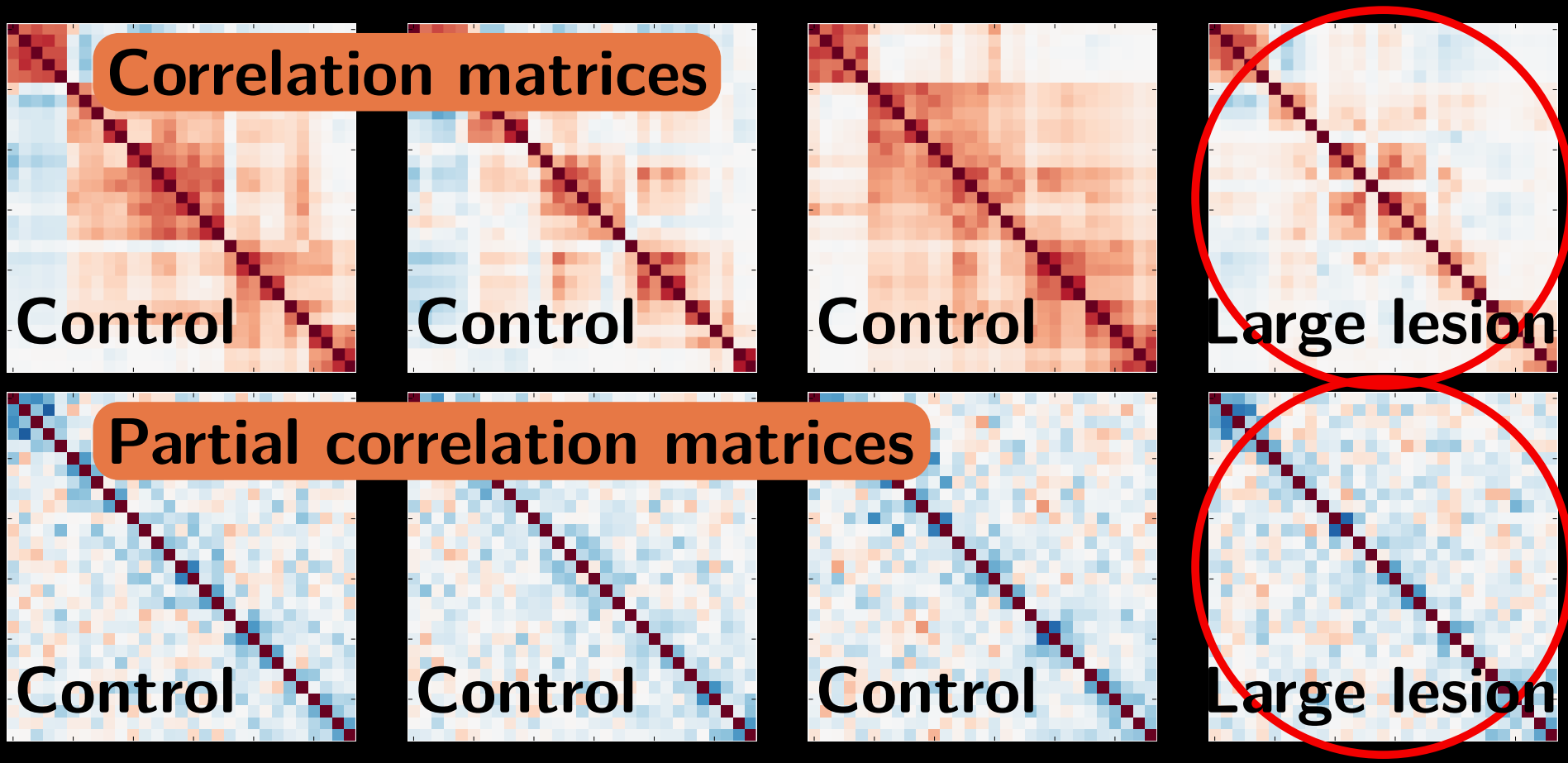
2 Connectome: differences across subjects



3 controls, 1 severe stroke patient

Which is which?

2 Connectome: differences across subjects



- Spread-out variability in correlation matrices
- Noise in partial-correlations

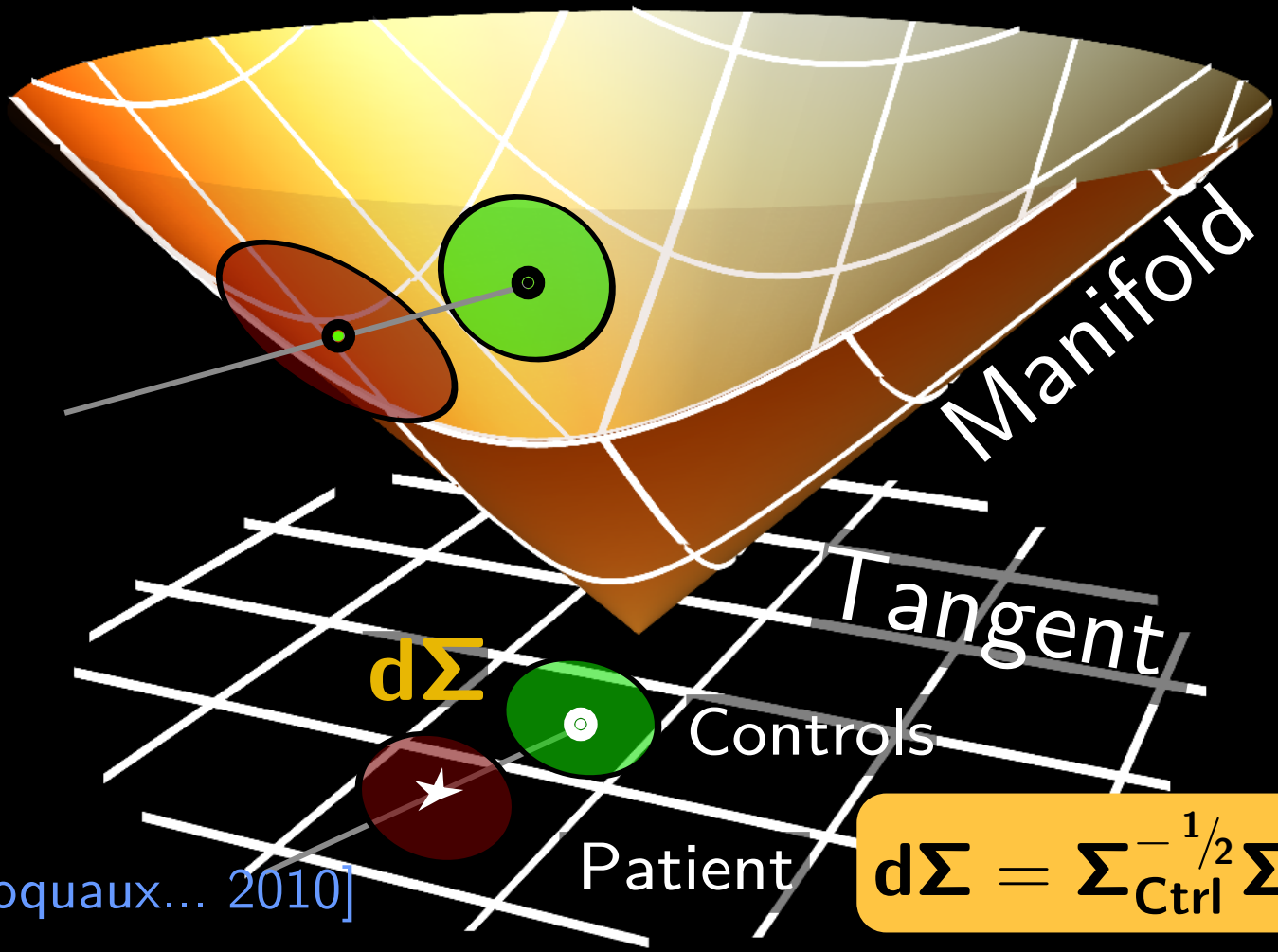
Strong dependence between coefficients

[Varoquaux... 2010]

2 Information geometry: uniform-error parametrization

- Estimation errors given by Fisher Information matrix
- Covariance matrices form a manifold

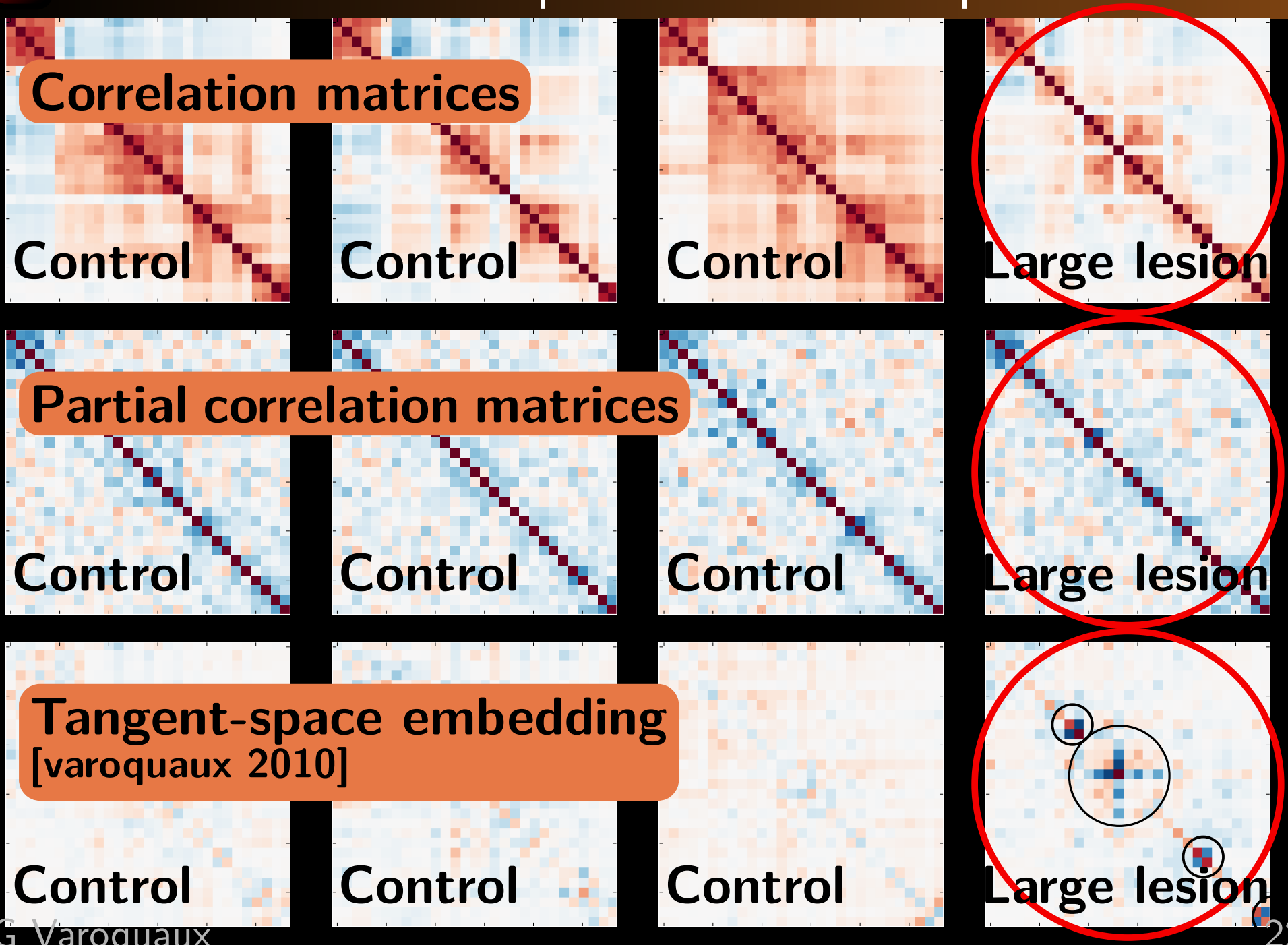
$\Rightarrow$ project to tangent space



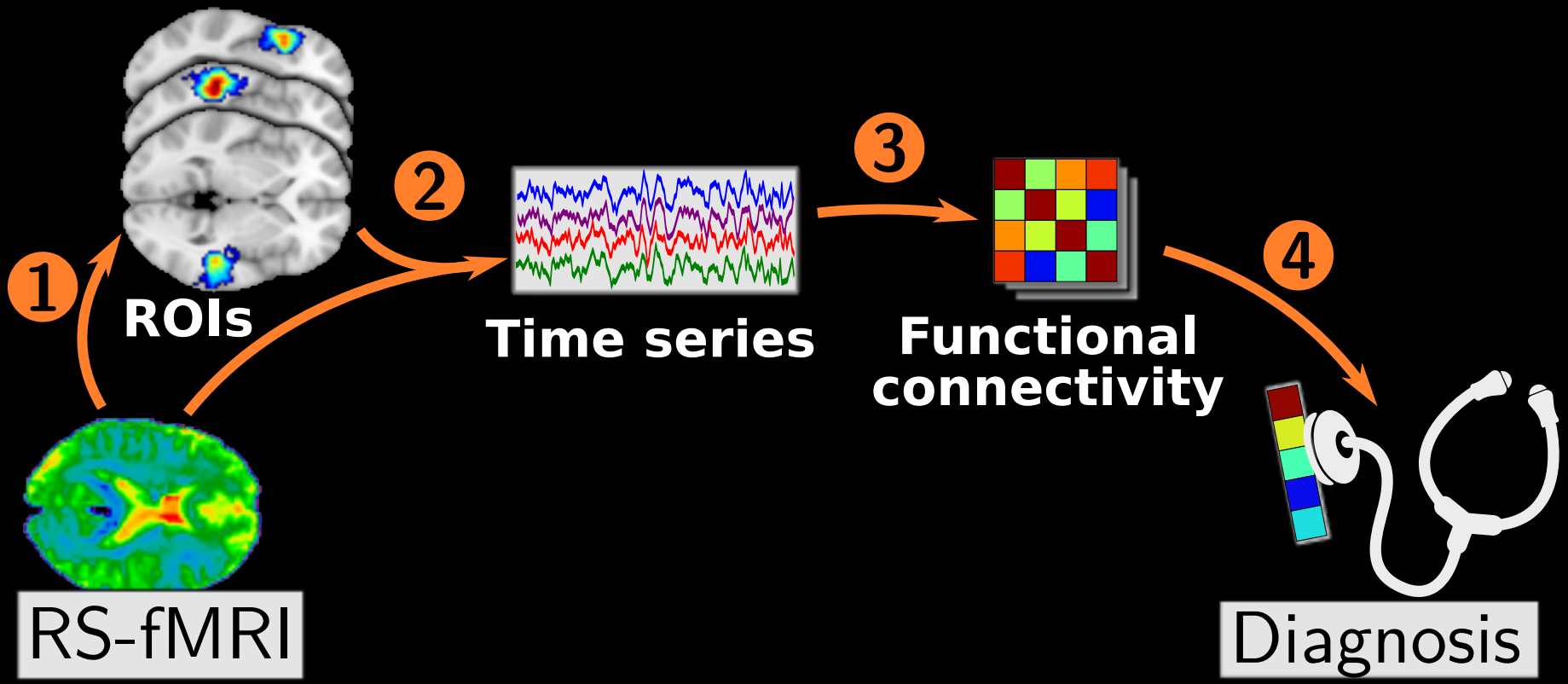
$$d\Sigma = \Sigma_{Ctrl}^{-1/2} \Sigma_{Patient} \Sigma_{Ctrl}^{-1/2}$$

[Varoquaux... 2010]

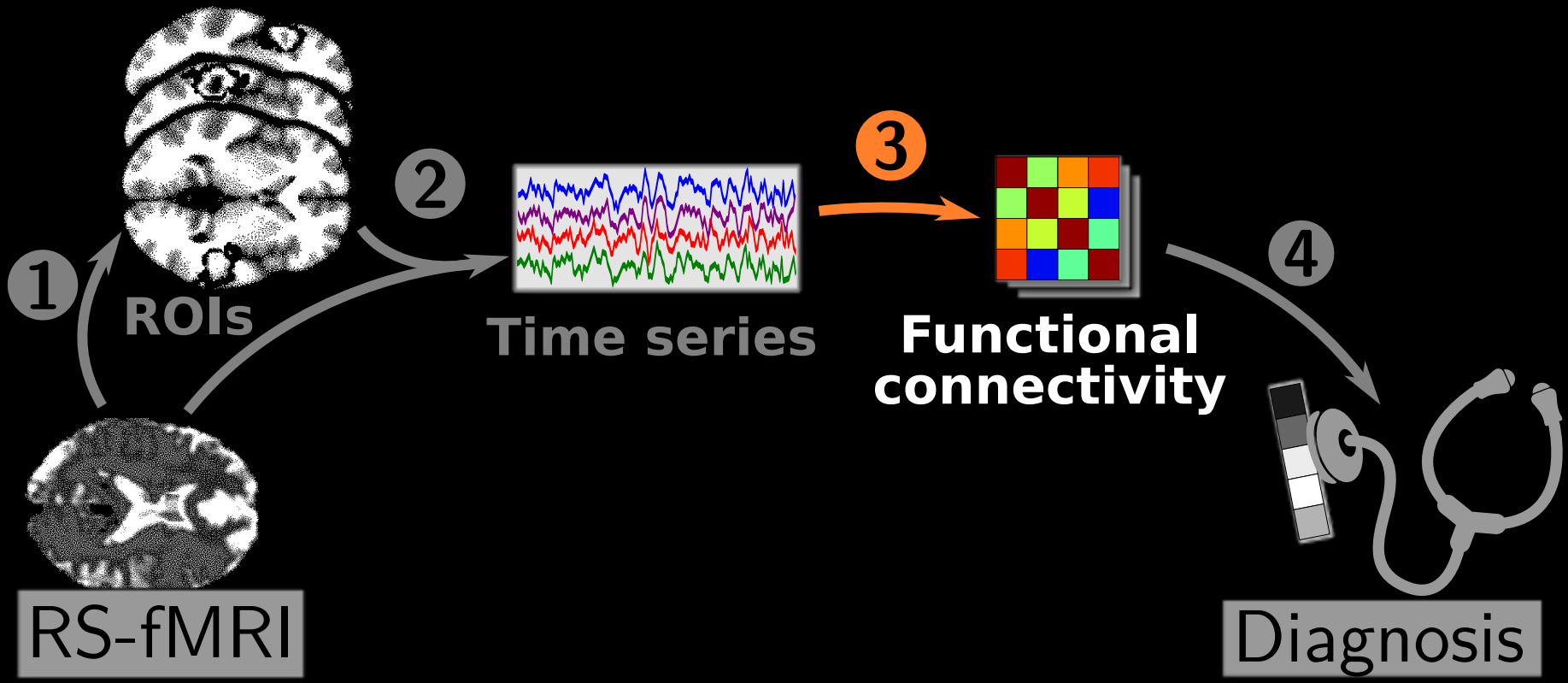
2 Connectome: which parametrization maps differences?



2 In connectome prediction settings



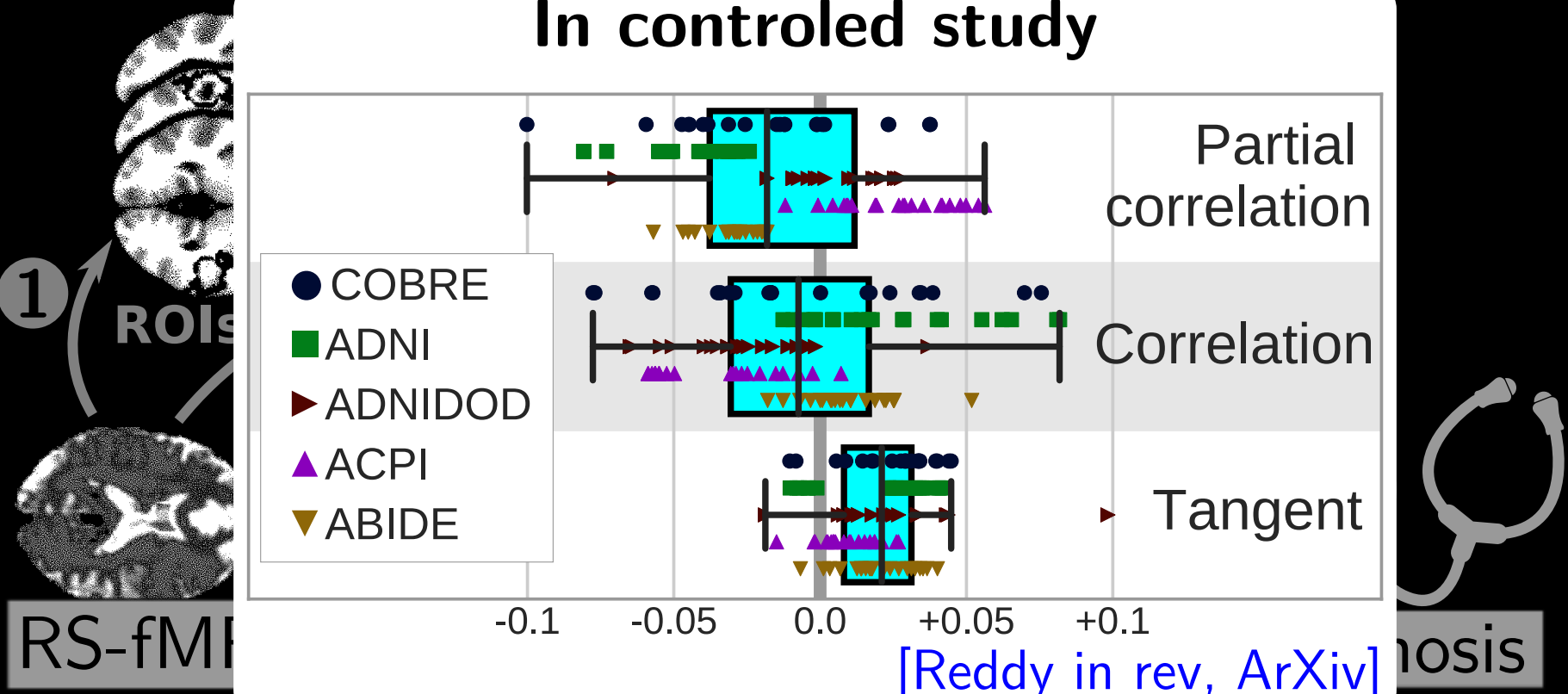
2 In connectome prediction settings



Connectivity matrix

- Correlation
- Partial correlations
- Tangent space

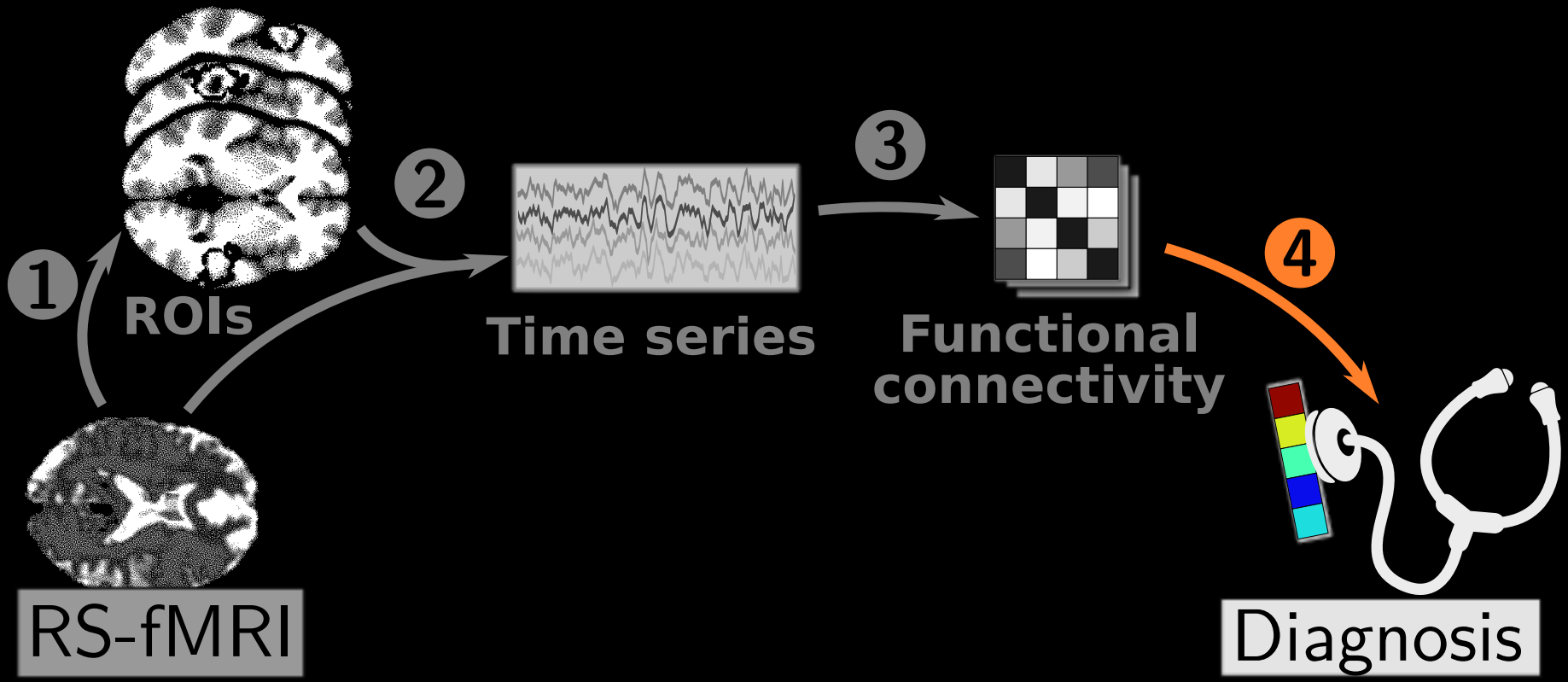
2 In connectome prediction settings



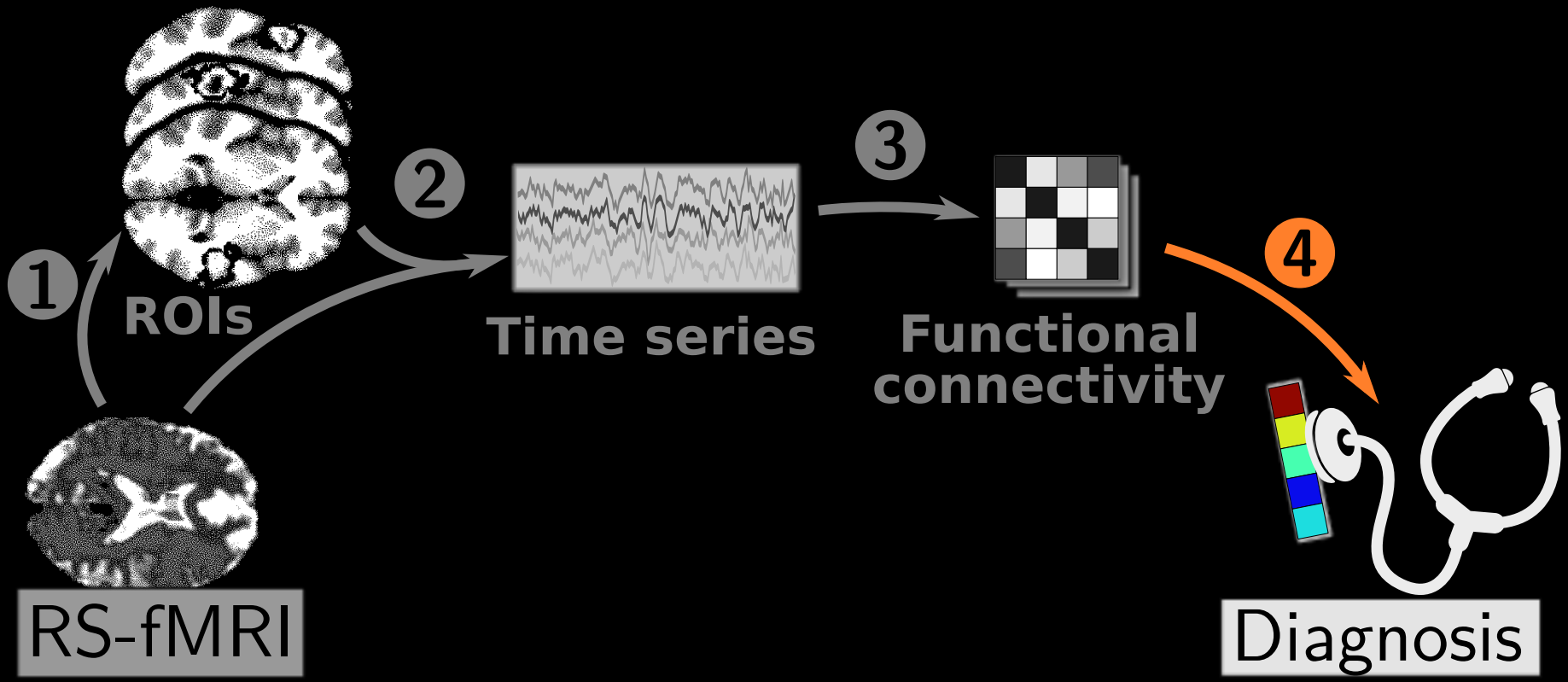
Connectivity matrix

- Correlation
- Partial correlations
- Tangent space

2 Machine learning for connectome prediction



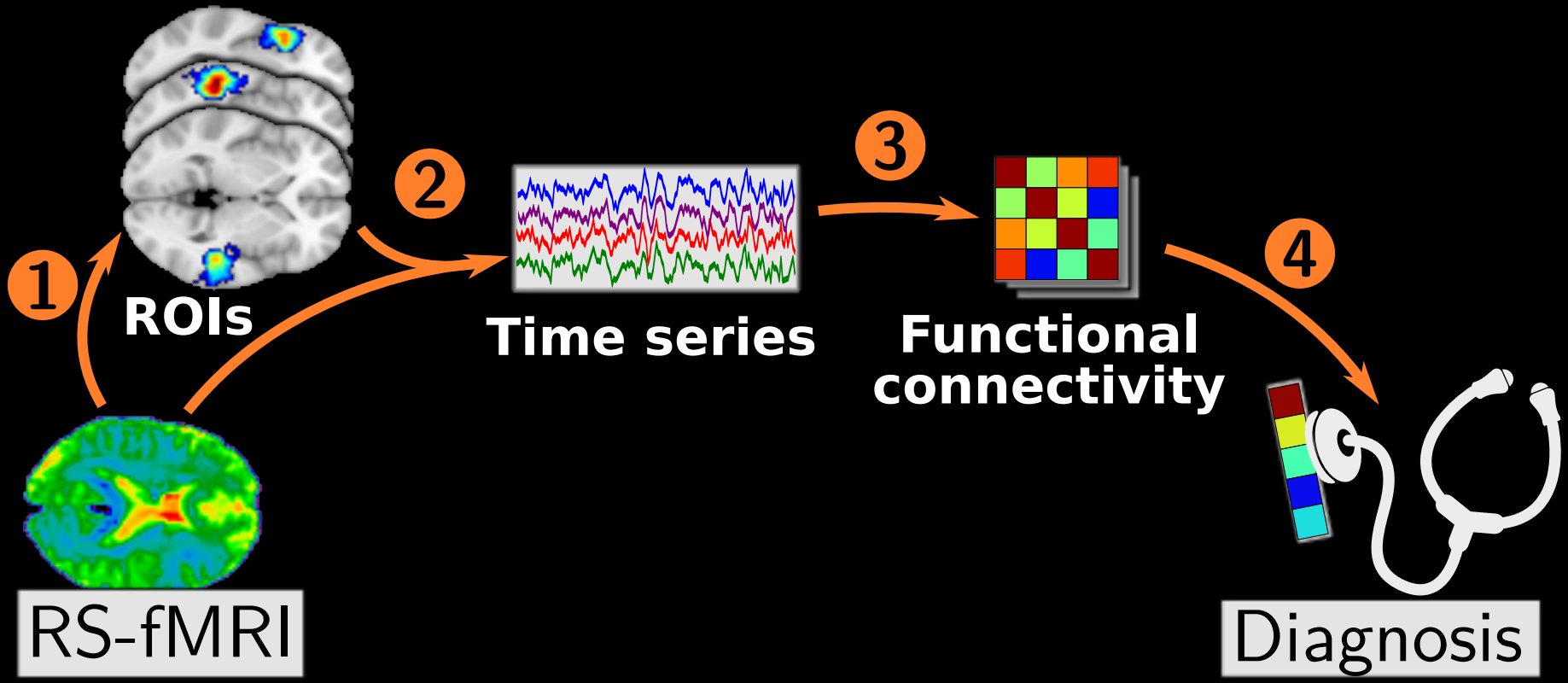
2 Machine learning for connectome prediction



Supervised learning

- Linear models

Predicting from brain activity at rest



1. Functional regions via linear decompositions
   [Abraham... 2013]
2. Tangent-space reparametrization
   [Varoquaux... 2010]
3. Supervised linear models (logistic regression)

3 Factorizing huge matrices

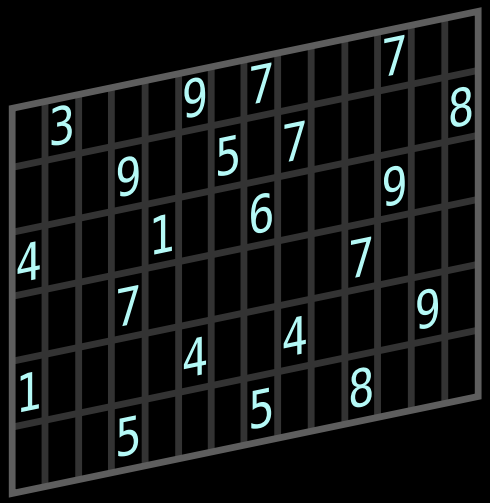
with A. Mensch, J. Mairal, B. Thirion

[Mensch... 2016, 2017]

Challenge: scalability

1. Intuitions
2. Experiments
3. Algorithms
4. Proof

3 Huge matrices: recommender systems

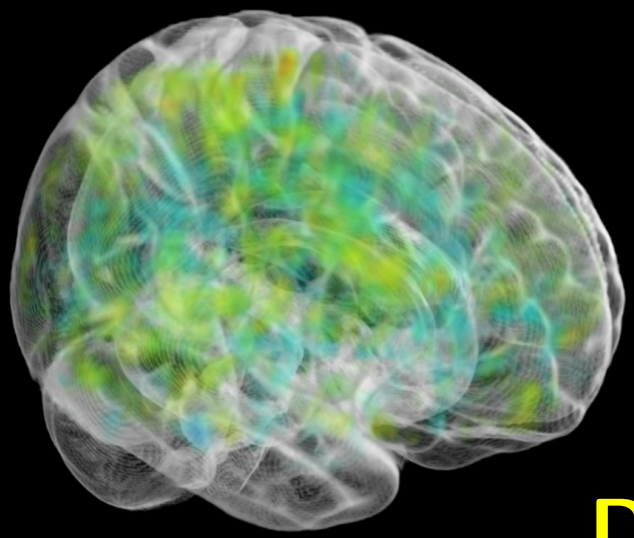


Product ratings

- Millions of entries
- Hundreds of thousands of products and users

Large sparse matrix

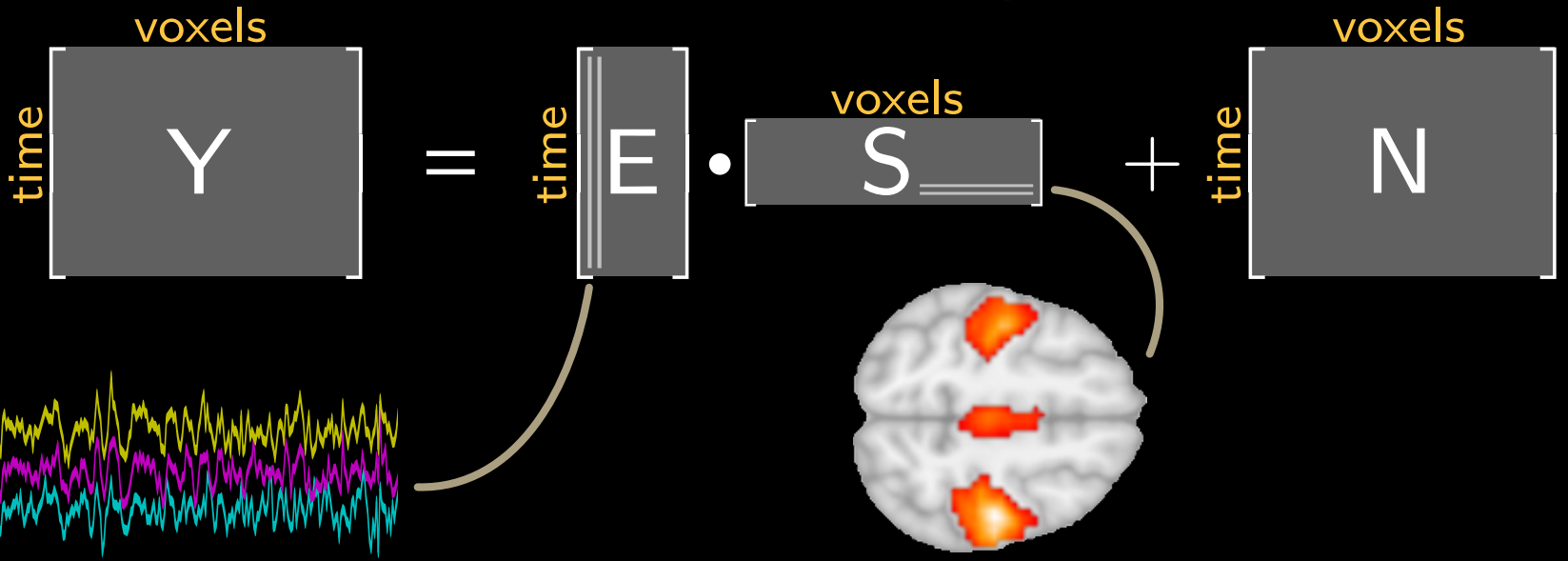
3 Huge matrices: brain imaging



Brain activity at rest

- 1000 subjects with $\sim$ 100–10 000 samples
- Images of dimensionality $> 100\,000$

Dense matrix, large both ways



$$\min_{\mathbf{w}} \sum_i l(\mathbf{x}_i\, \mathbf{w})$$

Many samples Stochastic gradient descent

$$\min_{\mathbf{w}} \mathbb{E}[l(y, \mathbf{x}\, \mathbf{w})]$$

Gradient descent: $\mathbf{w} \leftarrow \mathbf{w} + \alpha \nabla_{\mathbf{w}} l$

Stochastic gradient descent $\mathbf{w} \leftarrow \mathbf{w} + \alpha \mathbb{E}[\nabla_{\mathbf{w}} l]$

Use a cheap estimate of $\mathbb{E}[\nabla_{\mathbf{w}} l]$ (*e.g.* subsampling)

Large matrices

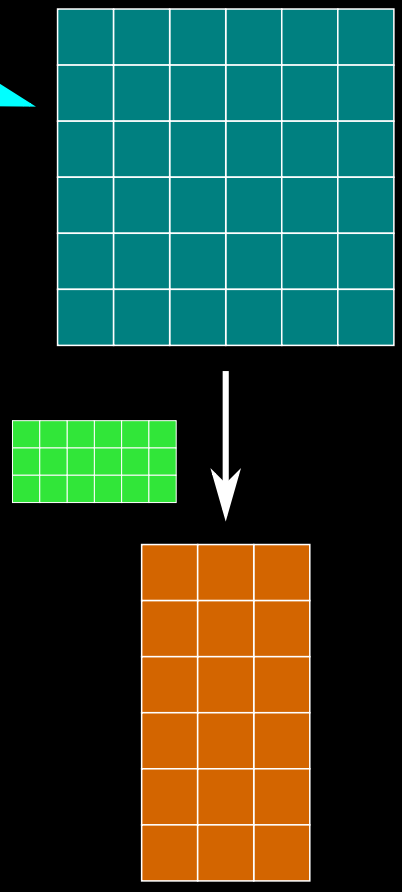
= terabytes of data

$$\underset{\mathbf{E},\mathbf{S}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{E}\,\mathbf{S}^T\|_{\text{Fro}}^2 + \lambda\Omega(\mathbf{S})$$

3 Scalable solvers for matrix factorizations

Data matrix

Alternating minimization



- Data access
- Code computation
- Dictionary update

Large matrices
= terabytes of data

$$\underset{\mathbf{E},\mathbf{S}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{E}\,\mathbf{S}^T\|_{\text{Fro}}^2 + \lambda\Omega(S)$$

Seen at t | Seen at $t+1$ | Unseen at t

Large matrices
= terabytes of data

$$\underset{\mathbf{E},\mathbf{S}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{E}\,\mathbf{S}^T\|_{\text{Fro}}^2 + \lambda\Omega(\mathbf{S})$$

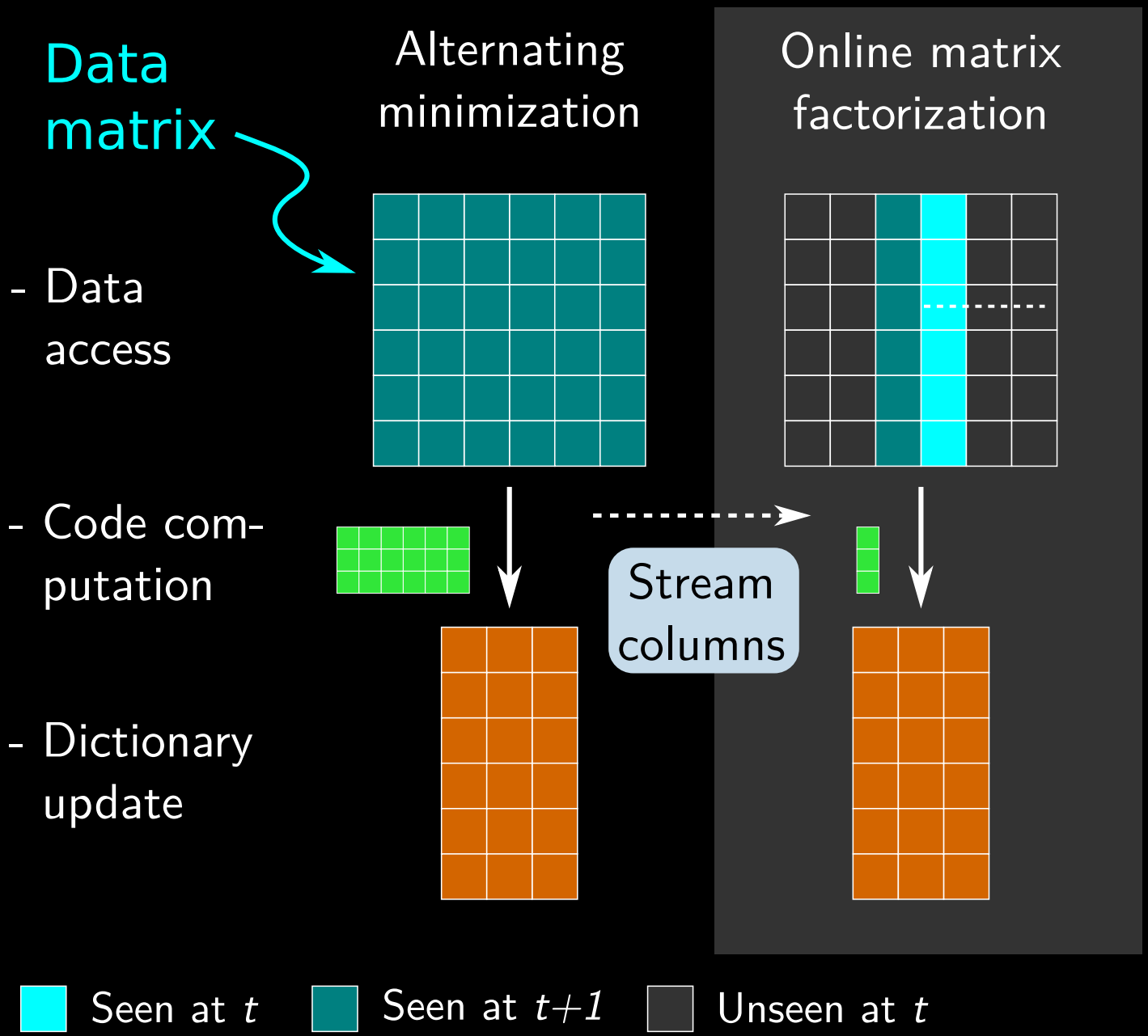
Rewrite as an expectation: [Mairal... 2010]

$$\underset{\mathbf{E}}{\operatorname{argmin}} \sum_i \left(\min_{\mathbf{s}} \|\mathbf{Y}_i - \mathbf{E}\,\mathbf{s}^T\|_{\text{Fro}}^2 + \lambda\Omega(\mathbf{s}) \right)$$

$$\underset{\mathbf{E}}{\operatorname{argmin}}\, \mathbb{E}\big[f(\mathbf{E})\big]$$

$\Rightarrow$ Optimize on approximations (sub-samples)

3 Scalable solvers for matrix factorizations



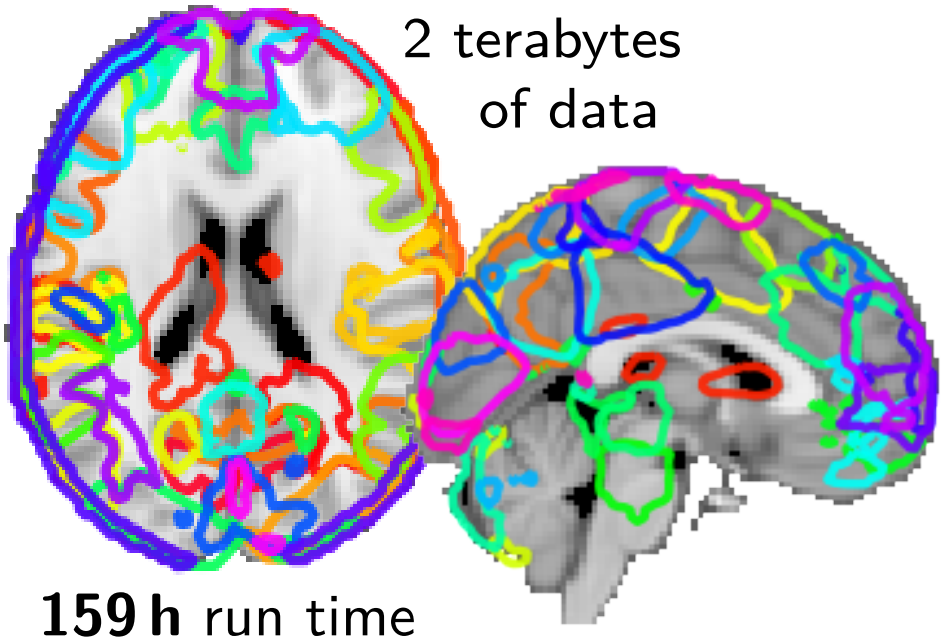
3 Scalable solvers for matrix factorizations

Data matrix

- Data

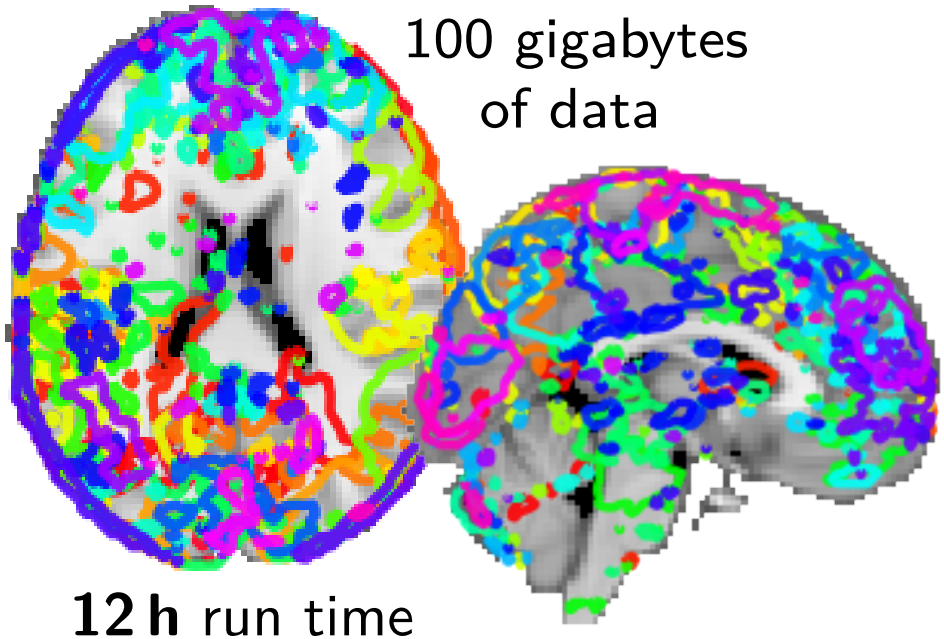
Alternating minimization

Online matrix factorization



2 terabytes of data

159 h run time

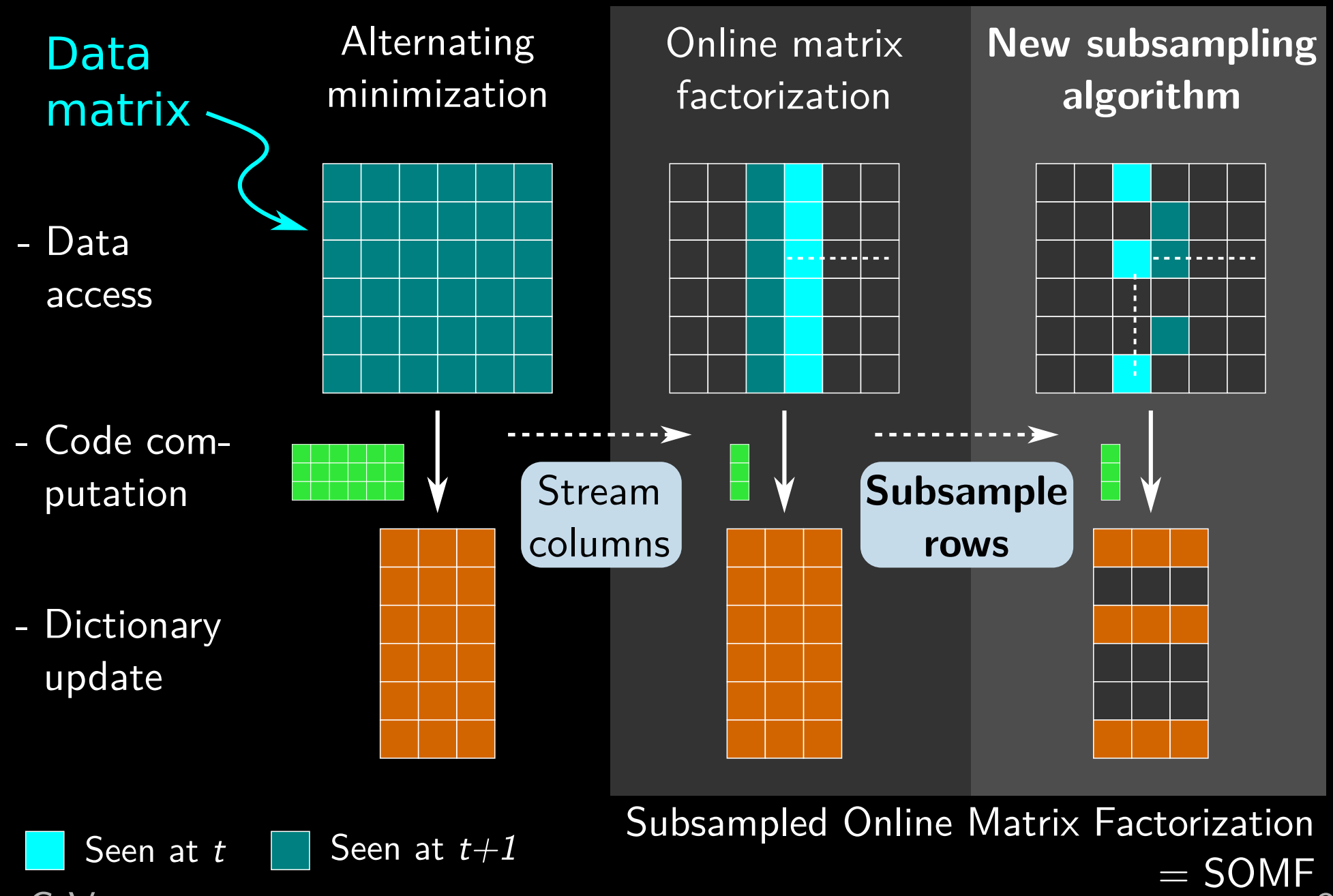


100 gigabytes of data

12 h run time

Online matrix factorization [Mairal... 2010]

3 Scalable solvers for matrix factorizations – SOMF



3 Scalable solvers for matrix factorizations – SOMF

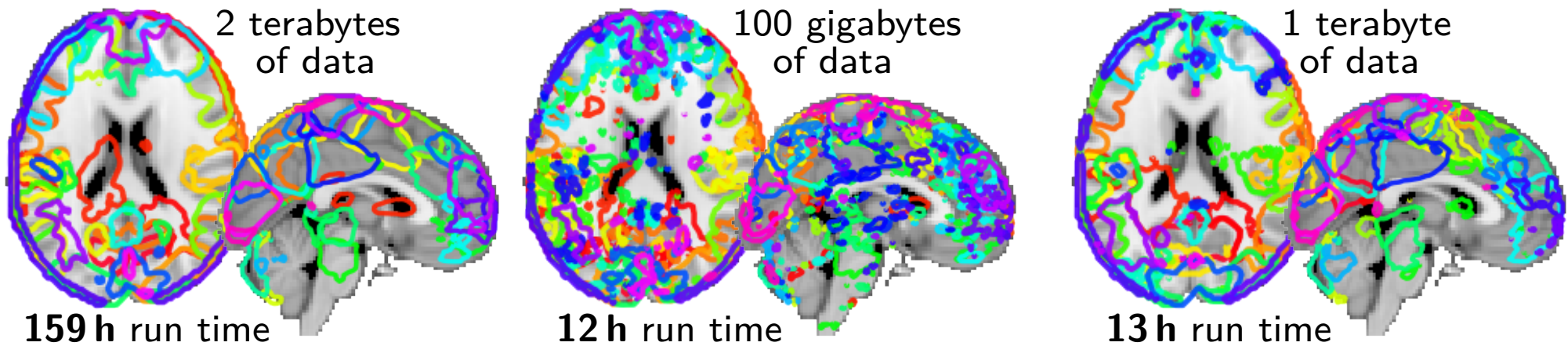
Data matrix

- Data access

Alternating minimization

Online matrix factorization

New subsampling algorithm

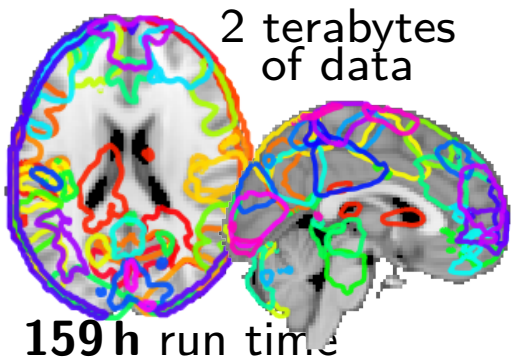


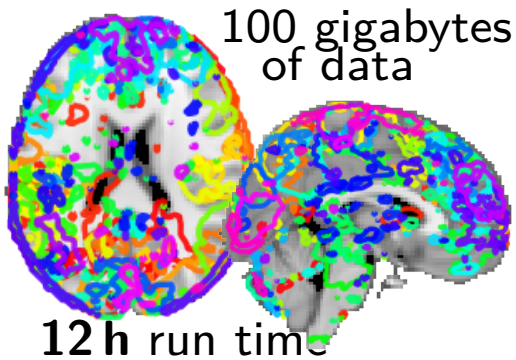
Online matrix factorization [Mairal... 2010]

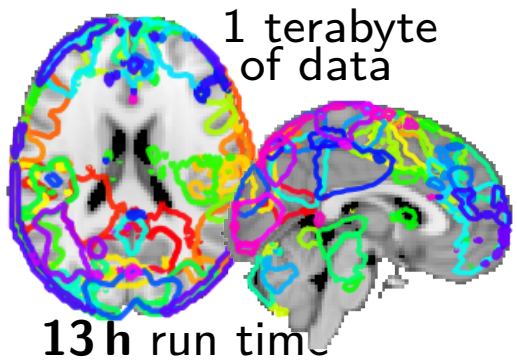
[Mensch... 2017]

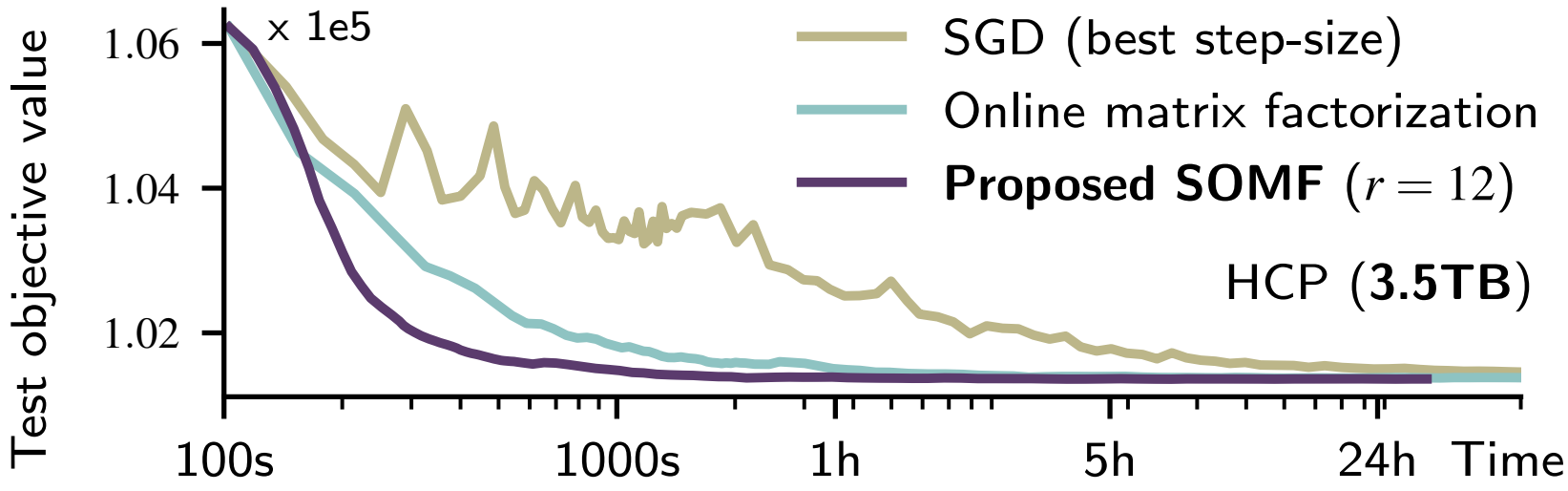
×10 speed up

3 Experimental results: resting-state fMRI



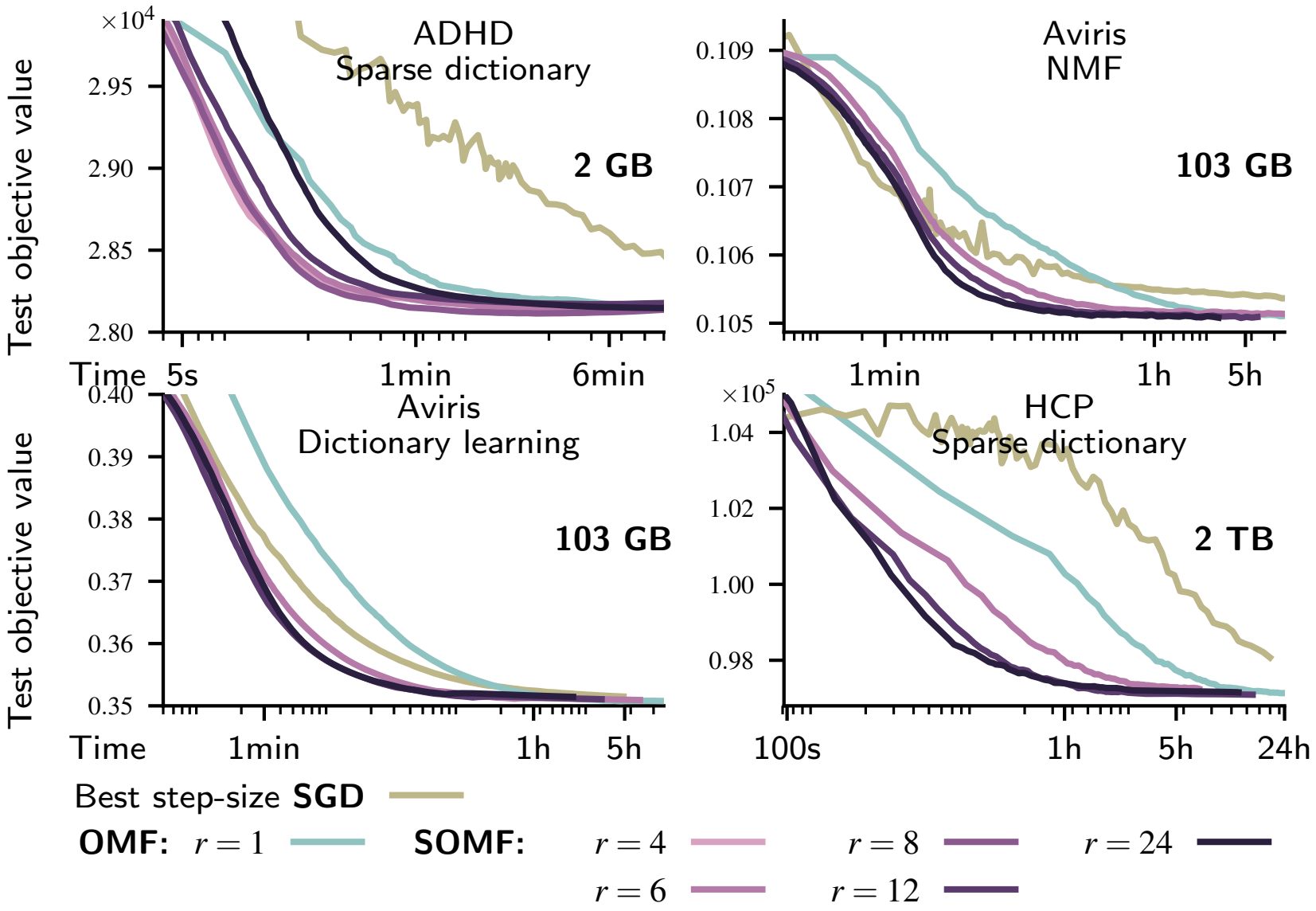






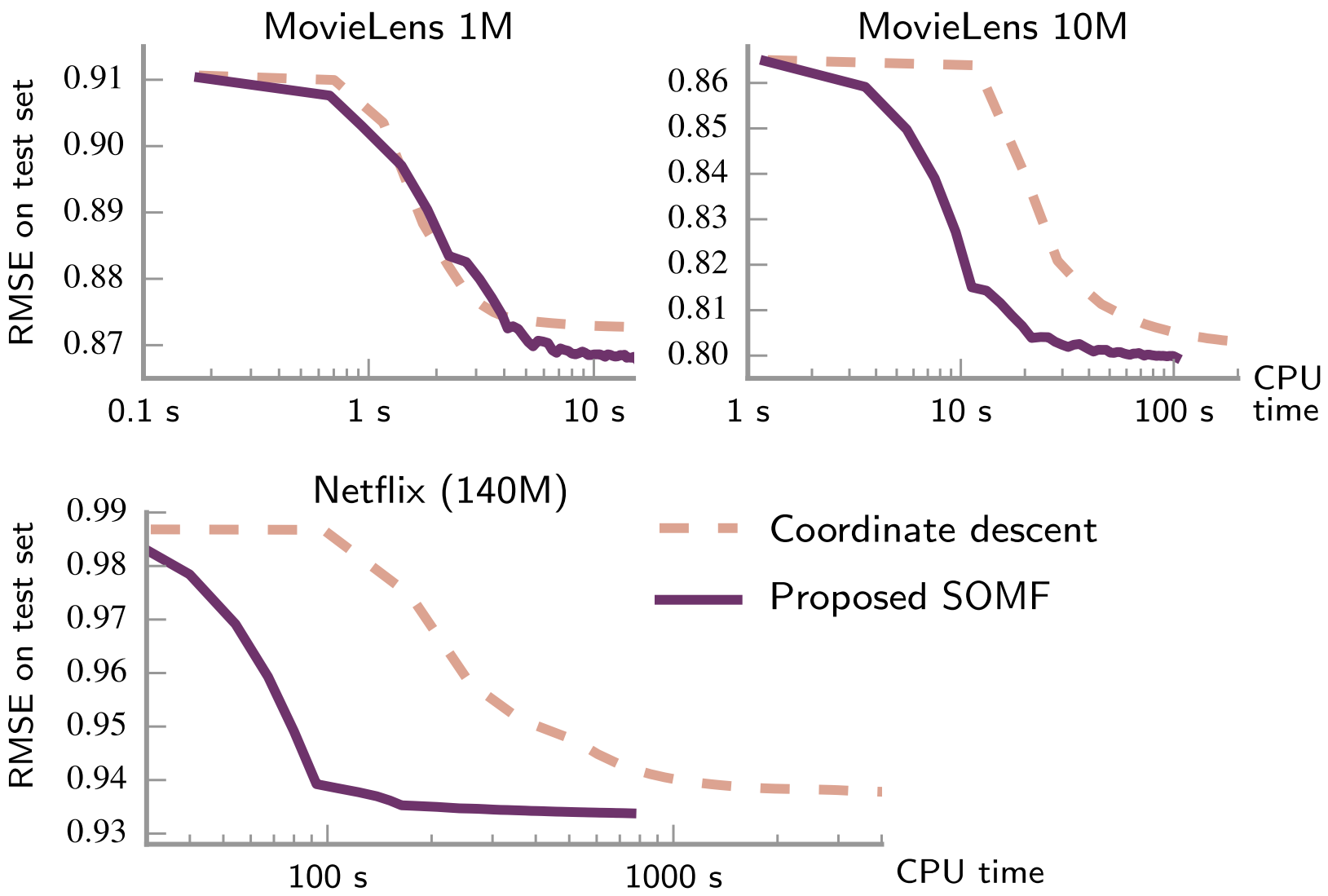
SOMF = Subsambled Online Matrix Factorization

3 Experimental results: large images



SOMF = Subsampled Online Matrix Factorization

3 Experimental results: recommender system



SOMF = Subsampled Online Matrix Factorization

Stream samples $\mathbf{x}_t$: [Mairal... 2010]

1. **Compute code**

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

2. **Update the surrogate function**

$$g_t(\mathbf{D}) = \frac{1}{t}\sum_{i=1}^{t} \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 = \operatorname{trace}\left(\frac{1}{2}\mathbf{D}^\top\mathbf{D}\mathbf{A}_t - \mathbf{D}^\top\mathbf{B}_t\right)$$

$$\mathbf{A}_t = (1-\frac{1}{t})\mathbf{A}_{t-1} + \frac{1}{t}\alpha_t\alpha_t^\top \qquad \mathbf{B}_t = (1-\frac{1}{t})\mathbf{B}_{t-1} + \frac{1}{t}\mathbf{x}_t\alpha_t^\top$$

3. **Minimize surrogate**

$$\mathbf{D}_t = \underset{\mathbf{D}\in\mathcal{C}}{\operatorname{argmin}}\, g_t(\mathbf{D}) \qquad \nabla g_t = \mathbf{D}\mathbf{A}_t - \mathbf{B}_t$$

Stream samples $\mathbf{x}_t$: [Mairal... 2010]

1. **Compute code**

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

2. **Update the surrogate function**

$$g_t(\mathbf{D}) = \frac{1}{t}\sum_{i=1}^{t} \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 = \text{trace}\left(\frac{1}{2}\mathbf{D}^\top\mathbf{D}\mathbf{A}_t - \mathbf{D}^\top\mathbf{B}_t\right)$$

$$\mathbf{A}_t = (1-\frac{1}{t})\mathbf{A}_{t-1} + \frac{1}{t}\alpha_t\alpha_t^\top \qquad \mathbf{B}_t = (1-\frac{1}{t})\mathbf{B}_{t-1} + \frac{1}{t}\mathbf{x}_t\alpha_t^\top$$

$$g_t(\mathbf{D}) \overset{\text{surrogate}}{=} \sum_{\mathbf{x}} l(\mathbf{x}, \mathbf{D}) \qquad \alpha_i \text{ is used, and not } \alpha$$

$\Rightarrow$ Stochastic Majorization-Minimization

No nasty hyper-parameters

Stream samples $\mathbf{x}_t$: [Mairal... 2010]

1. Compute code complexity depends on p

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

2. Update the surrogate function $\mathcal{O}(p)$

$$g_t(\mathbf{D}) = \frac{1}{t}\sum_{i=1}^{t} \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 = \operatorname{trace}\left(\frac{1}{2}\mathbf{D}^\top\mathbf{D}\mathbf{A}_t - \mathbf{D}^\top\mathbf{B}_t\right)$$

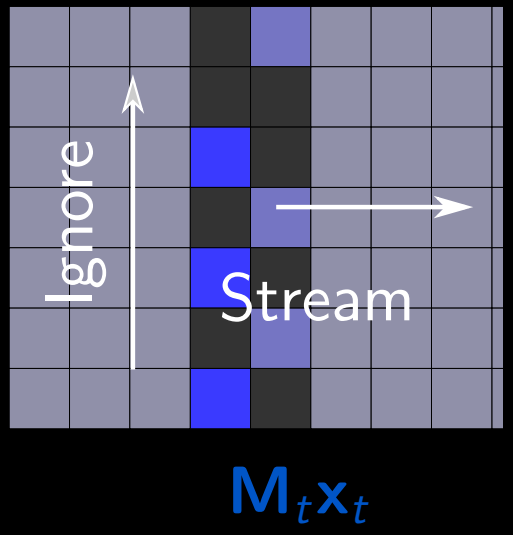
$$\mathbf{A}_t = (1-\frac{1}{t})\mathbf{A}_{t-1} + \frac{1}{t}\alpha_t\alpha_t^\top \qquad \mathbf{B}_t = (1-\frac{1}{t})\mathbf{B}_{t-1} + \frac{1}{t}\mathbf{x}_t\alpha_t^\top$$

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

$$\mathbf{D}_t = \underset{\mathbf{D}\in\mathcal{C}}{\operatorname{argmin}}\, g_t(\mathbf{D}) \qquad \nabla g_t = \mathbf{D}\mathbf{A}_t - \mathbf{B}_t$$

3 Sub-sample features

- **Data stream**: $(\mathbf{x}_t)_t \rightarrow$ masked $(\mathbf{M}_t\mathbf{x}_t)_t$
- **Dimension**: $p \rightarrow s$
- Use only $\mathbf{M}_t\mathbf{x}_t$ in computation
  $\rightarrow$ **complexity in** $\mathcal{O}(s)$



Modify all steps to work on s features

Code computation	**Surrogate update**	**Surrogate minimization**

3 Sub-sample features

Original online MF

1. Code computation

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

2. Surrogate aggregation

$$\mathbf{A}_t = \frac{1}{t}\sum_{i=1}^{t} \alpha_i \alpha_i^\top$$

$$\mathbf{B}_t = \mathbf{B}_{t-1} + \frac{1}{t}(\mathbf{x}_t \alpha_t^\top - \mathbf{B}_{t-1})$$

3. Surrogate minimization

$$\mathbf{D}^j \leftarrow p_{\mathcal{C}_j^r}^{\perp}(\mathbf{D}^j - \frac{1}{(\mathbf{A}_t)_{j,j}}(\mathbf{D}\mathbf{A}_t^j - \mathbf{B}_t^j))$$

Our algorithm

1. Approximate code computation: masked

$$\beta_t^{(i)} \leftarrow (1-\gamma)\mathbf{G}_{t-1}^{(i)} + \gamma \mathbf{D}_{t-1}^\top \mathbf{M}_t \mathbf{x}^{(i)}$$

$$\mathbf{G}_t^{(i)} \leftarrow (1-\gamma)\mathbf{G}_{t-1}^{(i)} + \gamma \mathbf{D}_{t-1}^\top \mathbf{M}_t \mathbf{D}_{t-1}$$

$$\alpha_t \leftarrow \underset{\alpha \in \mathbb{R}^k}{\text{argmin}} \frac{1}{2}\alpha^\top \mathbf{G}_t \alpha - \alpha^\top \beta_t + \lambda\,\Omega(\alpha).$$

2. Surrogate aggregation, averaging

$$\mathbf{A}_t = \frac{1}{w_t}\alpha_t \alpha_t^\top + (1 - \frac{1}{w_t})\mathbf{A}_{t-1}$$

$$\mathbf{P}_t \bar{\mathbf{B}}_t \leftarrow (1-w_t)\mathbf{P}_t \bar{\mathbf{B}}_{t-1} + w_t \mathbf{P}_t \mathbf{x}_t \alpha_t^\top$$

3. Surrogate minimization

$$\mathbf{P}_t \mathbf{D}_t \leftarrow \underset{\mathbf{D}^r \in \mathcal{C}^r}{\text{argmin}} \frac{1}{2}\text{tr}(\mathbf{D}^{r\top}\mathbf{D}^r \bar{\mathbf{A}}_t) - \text{tr}(\mathbf{D}^{r\top}\mathbf{P}_t \bar{\mathbf{B}}$$

$$\mathbf{P}_t{}^{\perp}\bar{\mathbf{B}}_t \leftarrow (1-w_t)\mathbf{P}_t{}^{\perp}\bar{\mathbf{B}}_{t-1} + w_t \mathbf{P}_t{}^{\perp}\mathbf{x}_t \alpha_t^\top.$$

3 Sub-sample features – variance reduction

Original online MF | Our algorithm

1.

α_t

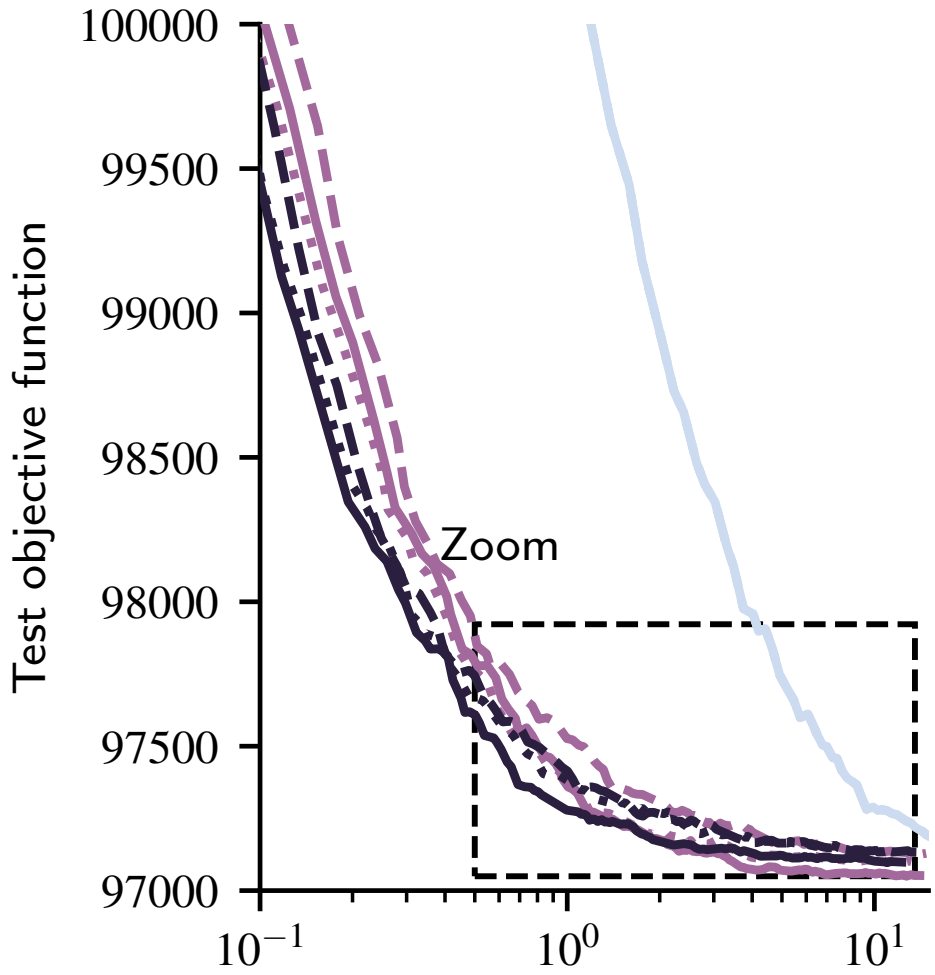
2.

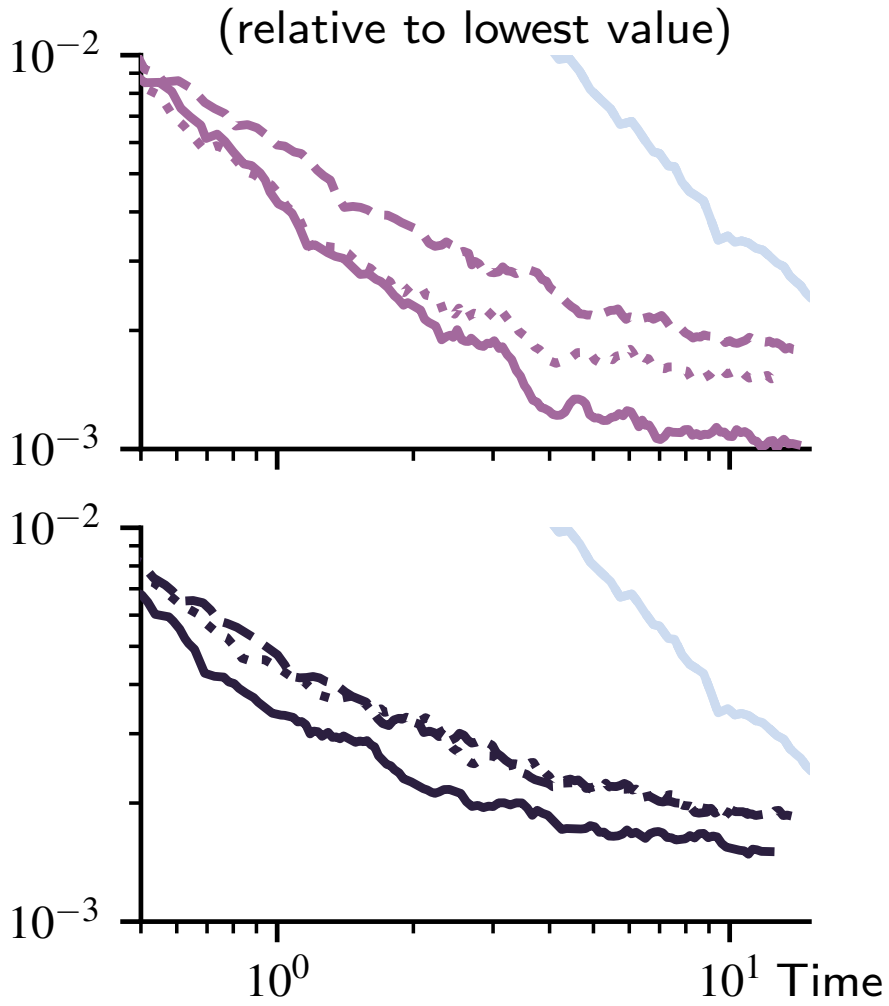
$\mathbf{A}_t$

$\mathbf{B}_t$

3.

$\mathbf{D}^j$





3 Why does it work?

Objective:

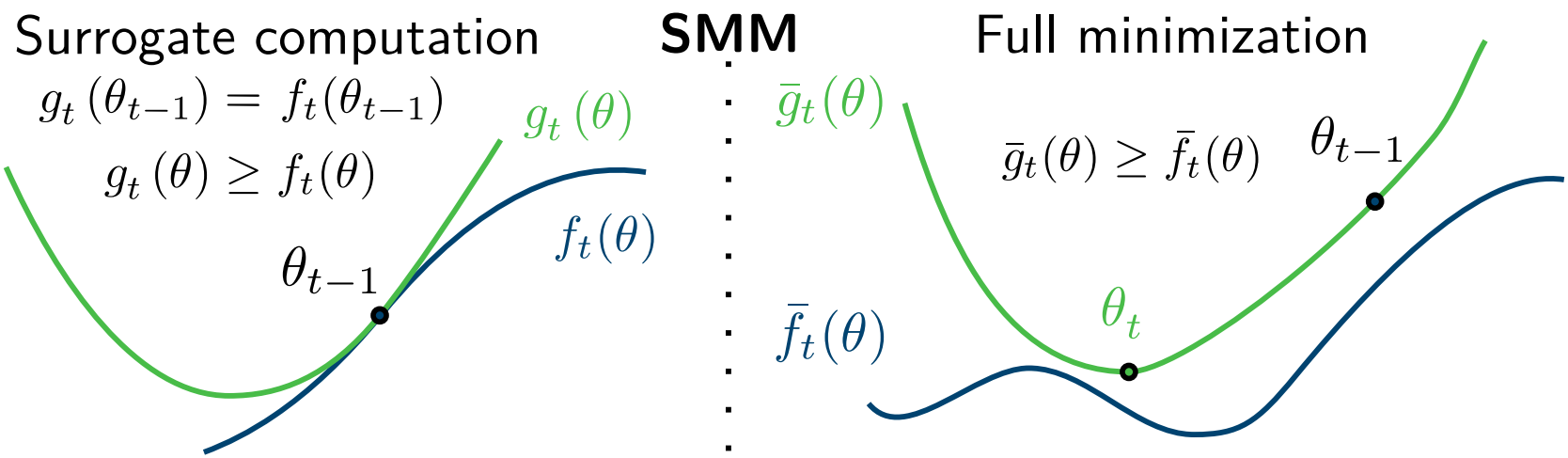
$$\mathbf{D} = \underset{\mathbf{D}\in\mathcal{C}}{\operatorname{argmin}} \sum_{\mathbf{x}} l(\mathbf{x}, \mathbf{D}) \quad \text{where } l(\mathbf{x}, \mathbf{D}) = \min_{\alpha} f(\mathbf{x}, \mathbf{D}, \alpha)$$

Algorithm (online matrix factorization)

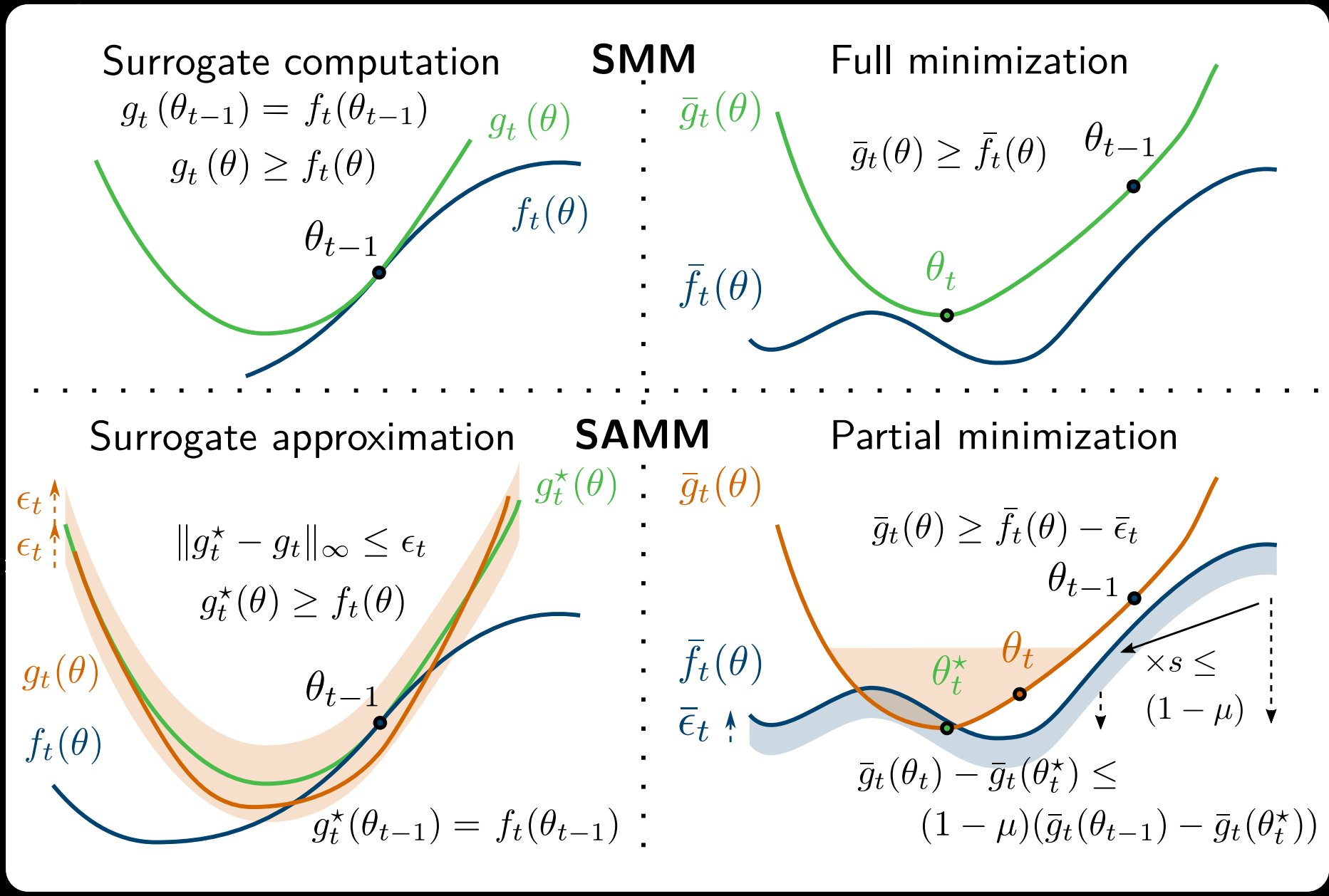
$$g_t(\mathbf{D}) \overset{\text{majorant}}{=} \sum_{\mathbf{x}} l(\mathbf{x}, \mathbf{D}) \quad \alpha_i \text{ is used, and not } \alpha^\star$$

$\Rightarrow$ Stochastic Majorization-Minimization [Mairal 2013]

3 Why does it work?



$\Rightarrow$ Stochastic Majorization-Minimization [Mairal 2013]


Surrogate computation
SMM
Full minimization
$g_t(\theta_{t-1}) = f_t(\theta_{t-1})$
$g_t(\theta) \geq f_t(\theta)$
$g_t(\theta)$
$f_t(\theta)$
θ_{t-1}
$\bar{g}_t(\theta)$
$\bar{g}_t(\theta) \geq \bar{f}_t(\theta)$
θ_{t-1}
θ_t
$\bar{f}_t(\theta)$
Surrogate approximation
SAMM
Partial minimization
$g_t^\star(\theta)$
ϵ_t
ϵ_t
$\|g_t^\star - g_t\|_\infty \leq \epsilon_t$
$g_t^\star(\theta) \geq f_t(\theta)$
$g_t(\theta)$
$f_t(\theta)$
θ_{t-1}
$g_t^\star(\theta_{t-1}) = f_t(\theta_{t-1})$
$\bar{g}_t(\theta)$
$\bar{g}_t(\theta) \geq \bar{f}_t(\theta) - \bar{\epsilon}_t$
θ_{t-1}
$\bar{f}_t(\theta)$
$\theta_t^\star$
θ_t
$\times s \leq$
$(1-\mu)$
$\bar{\epsilon}_t$
$\bar{g}_t(\theta_t) - \bar{g}_t(\theta_t^\star) \leq$
$(1-\mu)(\bar{g}_t(\theta_{t-1}) - \bar{g}_t(\theta_t^\star))$

Massive matrix factorization via subsampling

- Subsampling features $\Rightarrow$ doubly stochastic
- 10x speed ups on a fast algorithm
- Analysis via stochastic approximate majorization-minization
- Conclusive on various high-dimensional problems

Detecting psychiatric disorders with statistical learning tailored to brain activity

Improving psychiatry

Not a formal problem

Learning to map brain to behavior

More data is better

@GaelVaroquaux

Detecting psychiatric disorders with statistical learning tailored to brain activity

Improving psychiatry epistomology & sociology

Huge data stochastic computation

Factorization: costly in large-p, large-n

Sub-sampling p gives huge speed ups

Stochastic Approximate Majorization-Minimization

https://github.com/arthurmensch/modl

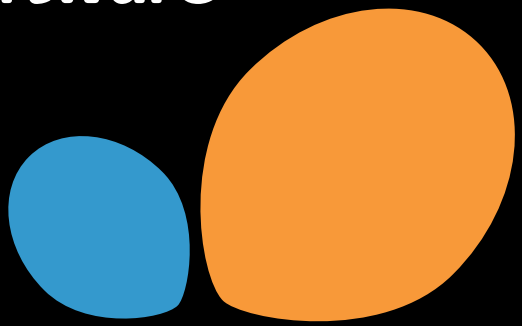
@GaelVaroquaux

Detecting psychiatric disorders with statistical learning tailored to brain activity

Improving psychiatry epistomology & sociology

Huge data stochastic computation

Software



Scikit-learn

nilearn

Putting new methods in the hands of everybody

@GaelVaroquaux

References I

A. Abraham, E. Dohmatob, B. Thirion, D. Samaras, and G. Varoquaux. Extracting brain regions from rest fMRI with total-variation constrained dictionary learning. In *MICCAI*, page 607. 2013.

F. Liem, G. Varoquaux, J. Kynast, F. Beyer, S. K. Masouleh, J. M. Huntenburg, L. Lampe, M. Rahim, A. Abraham, R. C. Craddock, ... Predicting brain-age from multimodal imaging data captures cognitive impairment. *NeuroImage*, 2016.

J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems*, 2013.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19, 2010.

A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux. Dictionary learning for massive matrix factorization. In *ICML*, 2016.

References II

A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux. Stochastic subsampling for factorizing huge matrices. *IEEE Transactions on Signal Processing*, 66(1):113–128, 2017.

M. Rahim, B. Thirion, D. Bzdok, I. Buvat, and G. Varoquaux. Joint prediction of multiple scores captures better individual traits from brain images. *Neuroimage*, in rev, 2017.

B. Thirion, G. Varoquaux, E. Dohmatob, and J. Poline. Which fMRI clustering gives good brain parcellations? *Name: Frontiers in Neuroscience*, 8:167, 2014.

G. Varoquaux, F. Baronnet, A. Kleinschmidt, P. Fillard, and B. Thirion. Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. In *MICCAI*. 2010.