Federated Learning in Biomedical Data: Application to Brain Imaging and Genetics Analysis

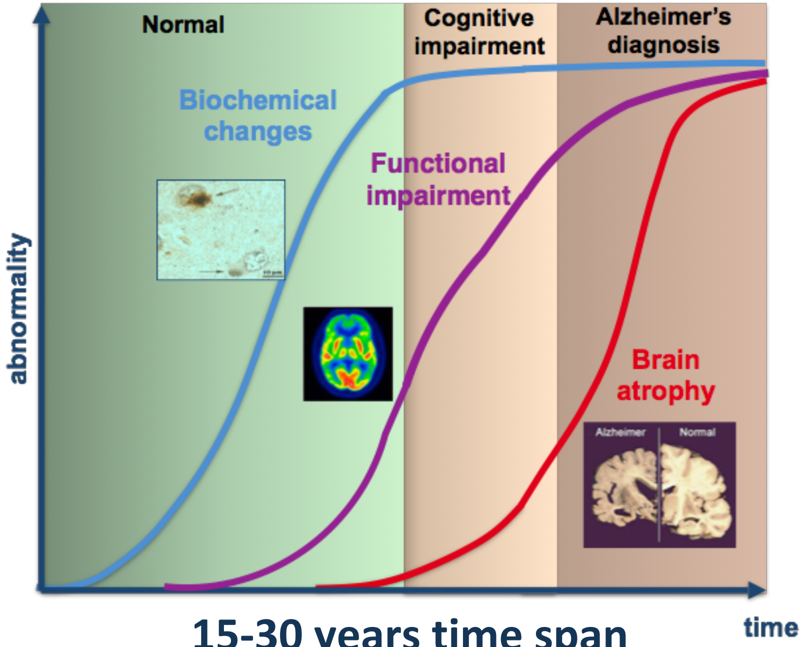
Marco Lorenzi

Université Côte d'Azur, Inria Sophia Antipolis

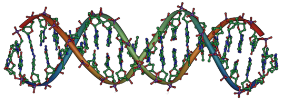
Epione Research Group

Dynamics of neurodegeneration

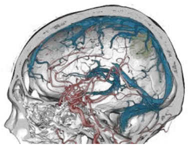
Jack et al, Lancet Neurol 2010;
Frisoni et al, Nature Rev Neurol 2010



15-30 years time span



Genetics



Vascularity

?

Sociodemographic

Microbiome

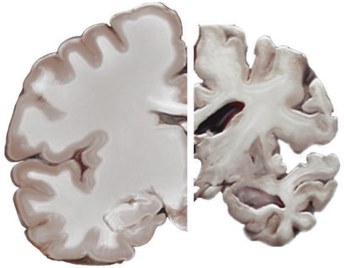
...

Data Science meets biomedical research

Statistical learning

\+

Neurology



Statistical learning

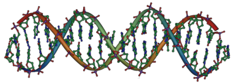
- formalizes hypothesis into models
- verifies models by means of data

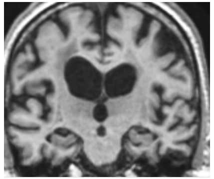
A discipline to answer challenging questions in medicine

Data Science meets biomedical research

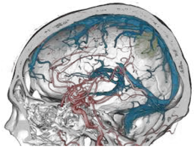
- Better understanding of the pathology

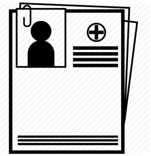
How to integrate heterogenous biomedical measures?





baseline





...

Data Science meets biomedical research

- Monitoring and studying diseases world-wide

How to analyse private healthcare data?

single hospital: 100s – 1’000s patients

Big Data

100s – 1’000s hospitals

1’000’000s patients

Challenges

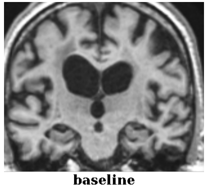
- How to integrate heterogeneous biomedical measures?

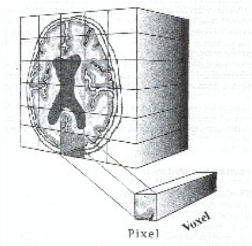
- How to analyse private healthcare data collected worldwide?

Imaging-genetics

Identifying genetic modulators of the brain phenotype

Brain imaging

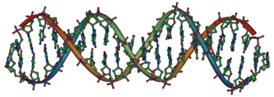


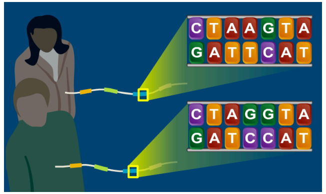


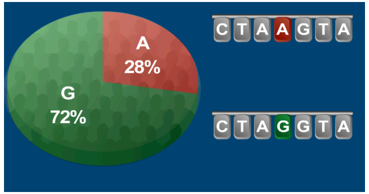
100’000 to 1’000’000 voxels

Genetics

Genetic variants (Single Nucleotide Polymorphism - SNP -)







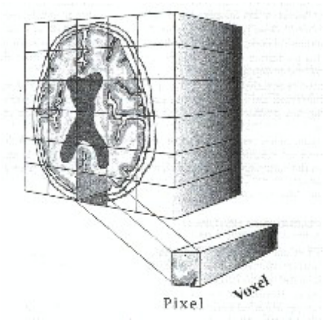
> 1’000’000 known SNPs

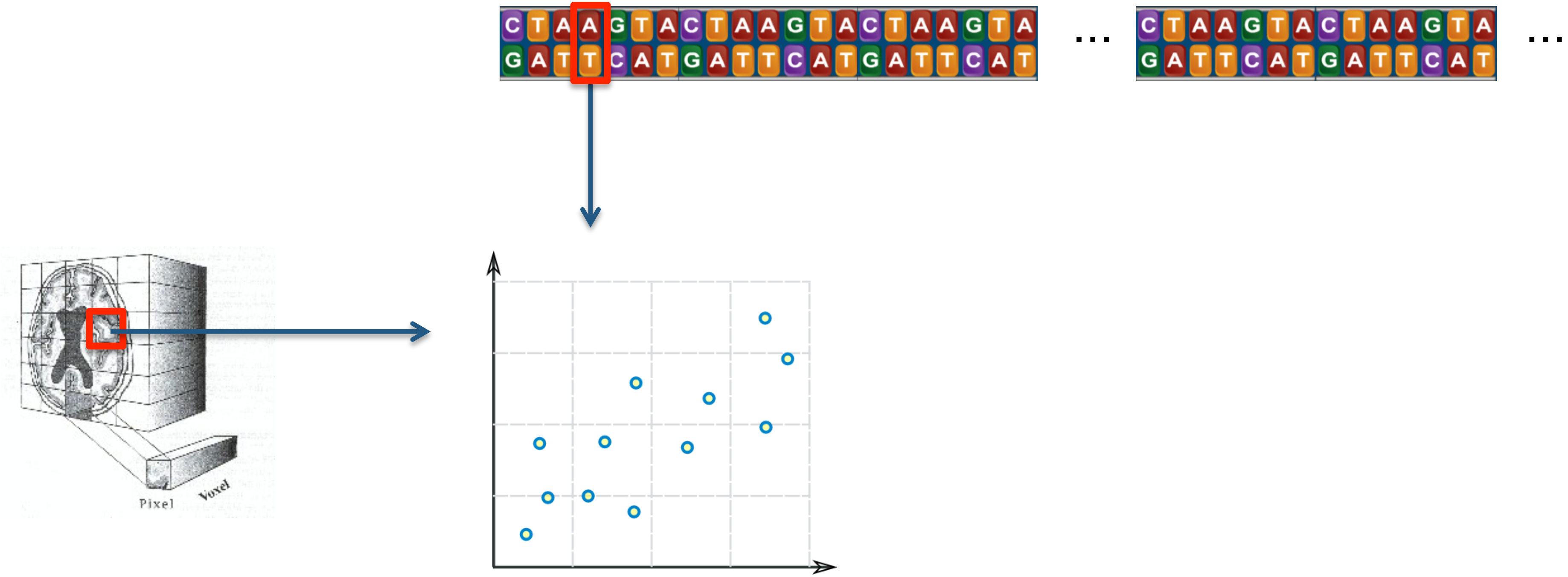

C T A A G T A C T A A G T A C T A A G T A
G A T T C A T G A T T C A T G A T T C A T

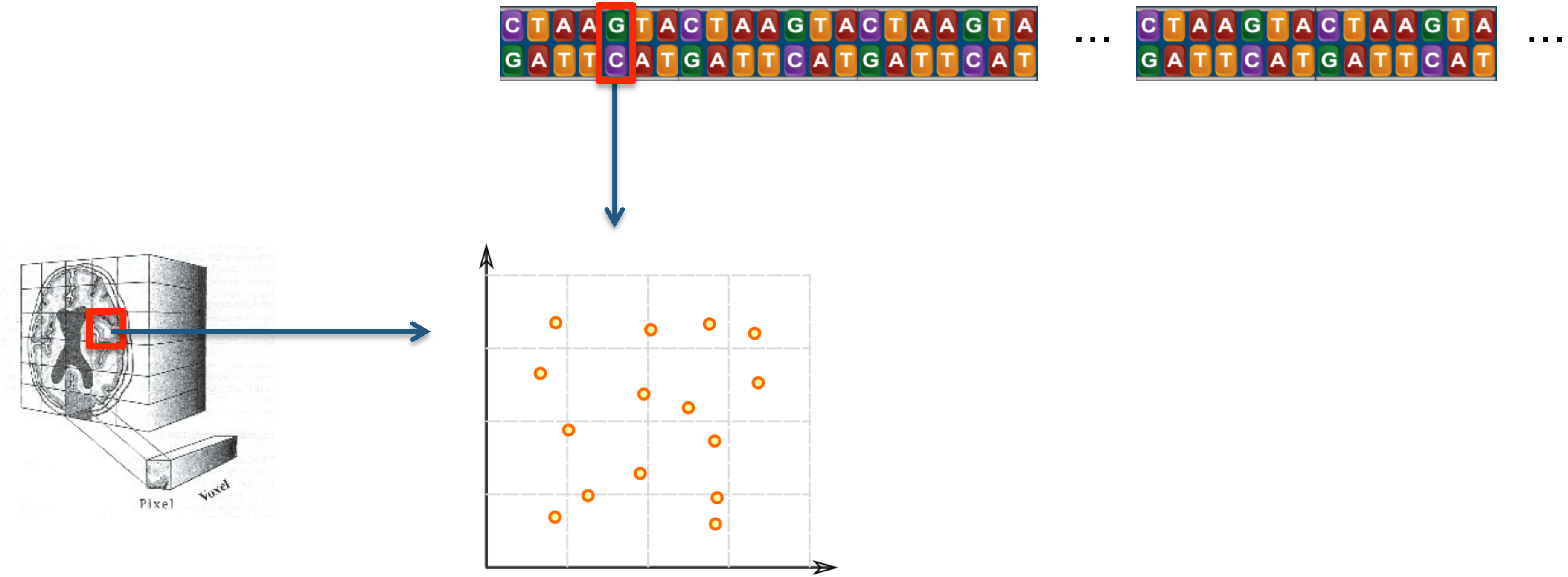
…

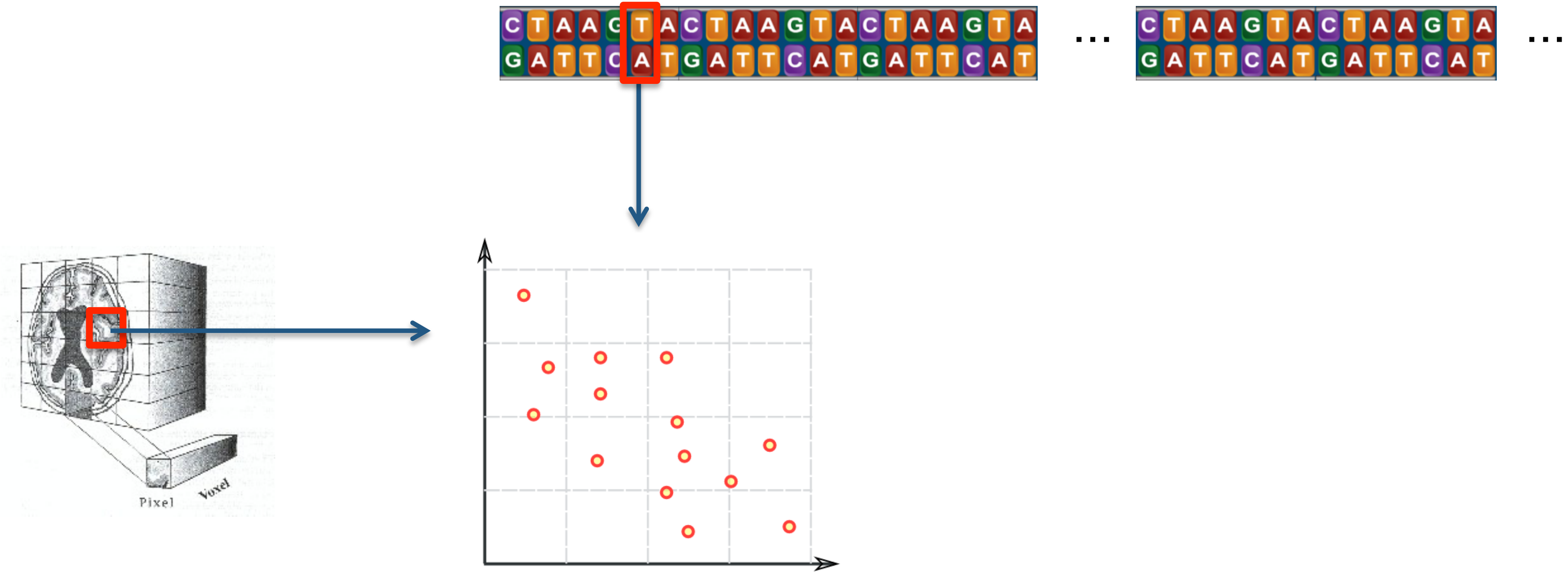

C T A A G T A C T A A G T A
G A T T C A T G A T T C A T

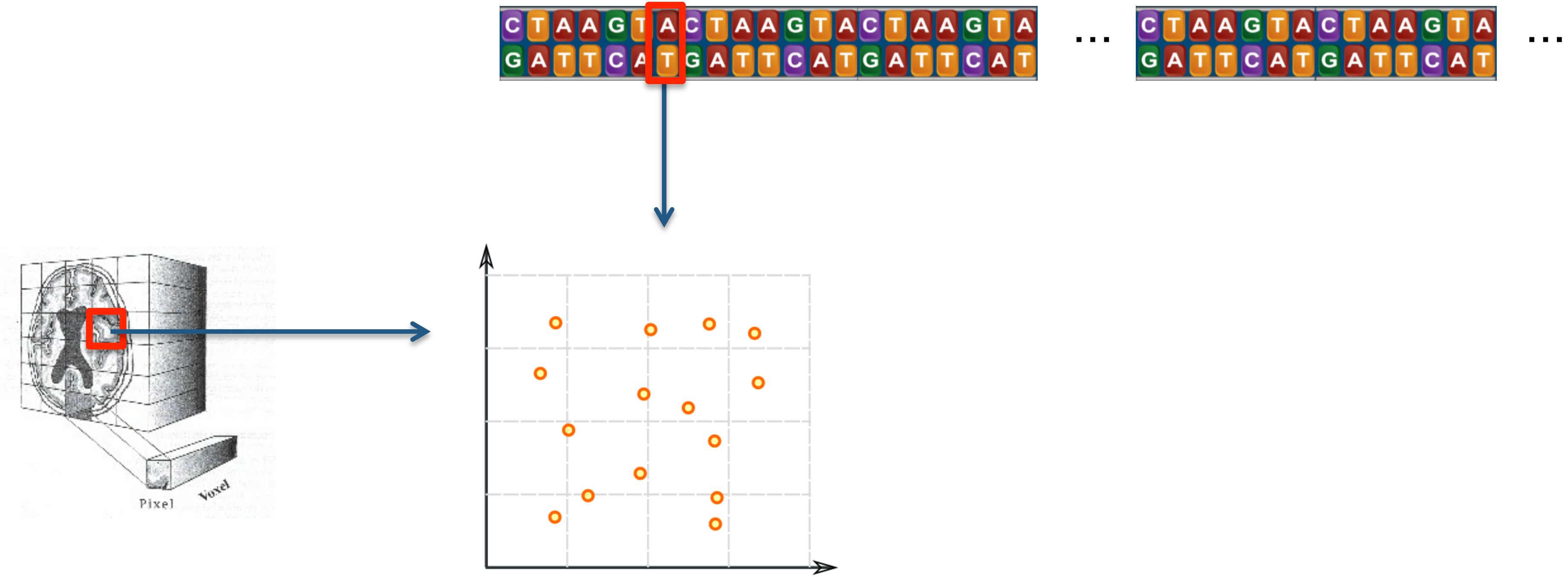
…


Pixel
Voxel

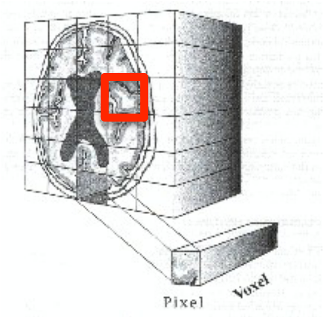

C T A A G T A C T A A G T A C T A A G T A
G A T T C A T G A T T C A T G A T T C A T
…
C T A A G T A C T A A G T A
G A T T C A T G A T T C A T
…
Pixel
Voxel


CTAAGTACTAAGTACTAAGTA
GATTCATGATTCATGATTCAT
…
CTAAGTACTAAGTA
GATTCATGATTCAT
…
Pixel
Voxel

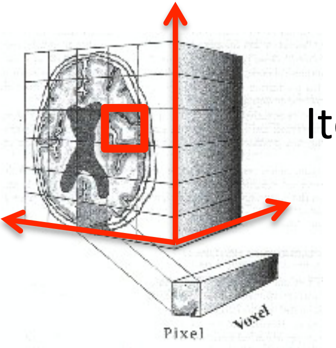

CTAAGTACTAAGTACTAAGTA
GATTCATGATTCATGATTCAT
…
CTAAGTACTAAGTA
GATTCATGATTCAT
…
Pixel
Voxel


CTAAGTACTAAGTACTAAGTA
GATTCATGATTCATGATTCAT
…
CTAAGTACTAAGTA
GATTCATGATTCAT
…
Pixel
Voxel

Iterate for > 1’000’000 variants

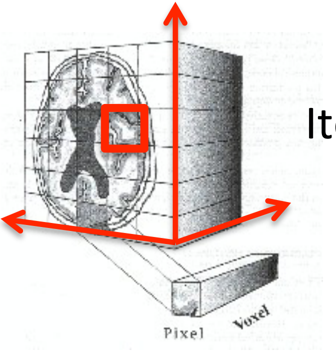


Iterate for > 1’000’000 variants



Iterate for > 1’000’000
image locations

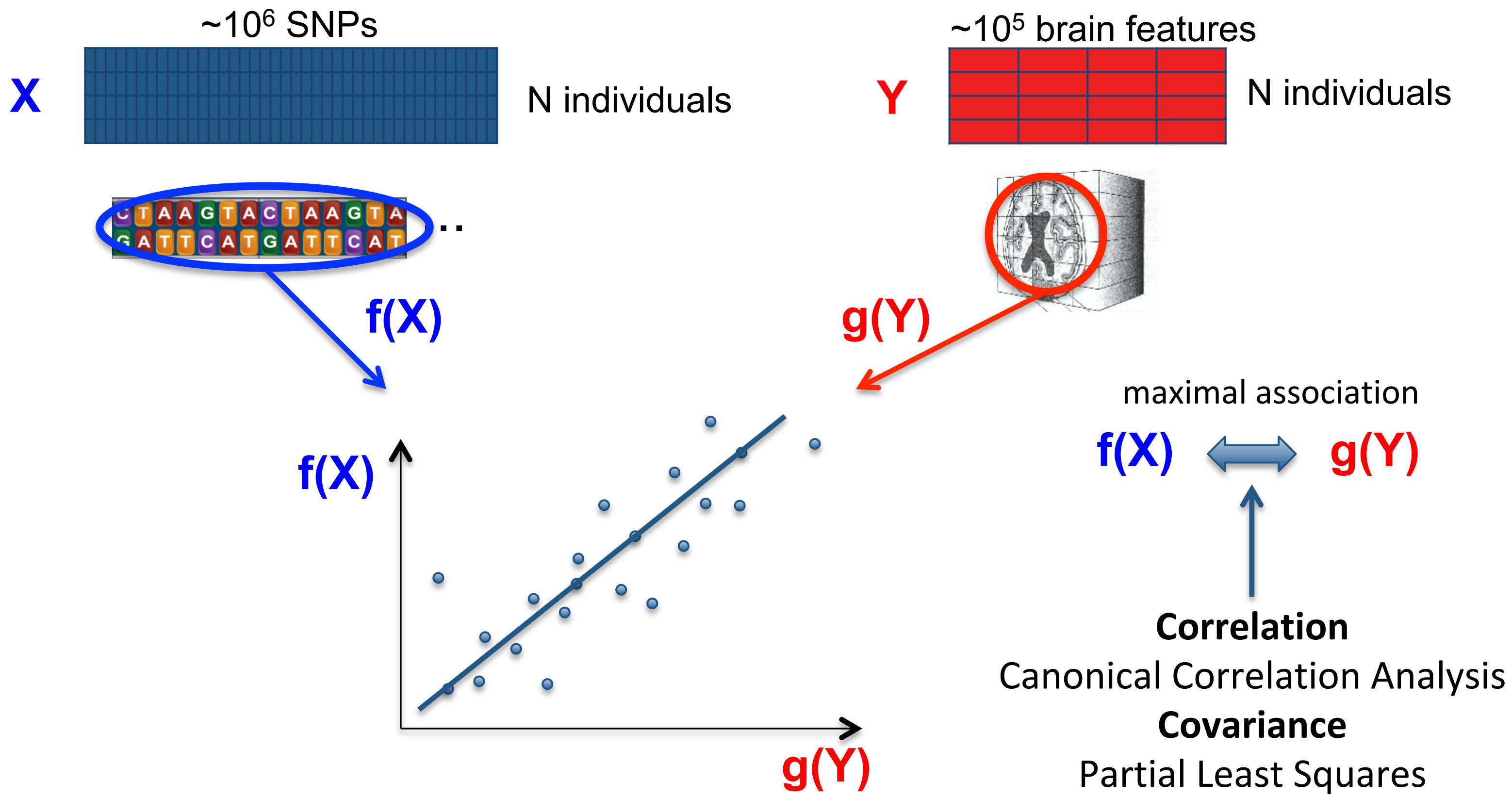
Iterate for > 1’000’000 variants



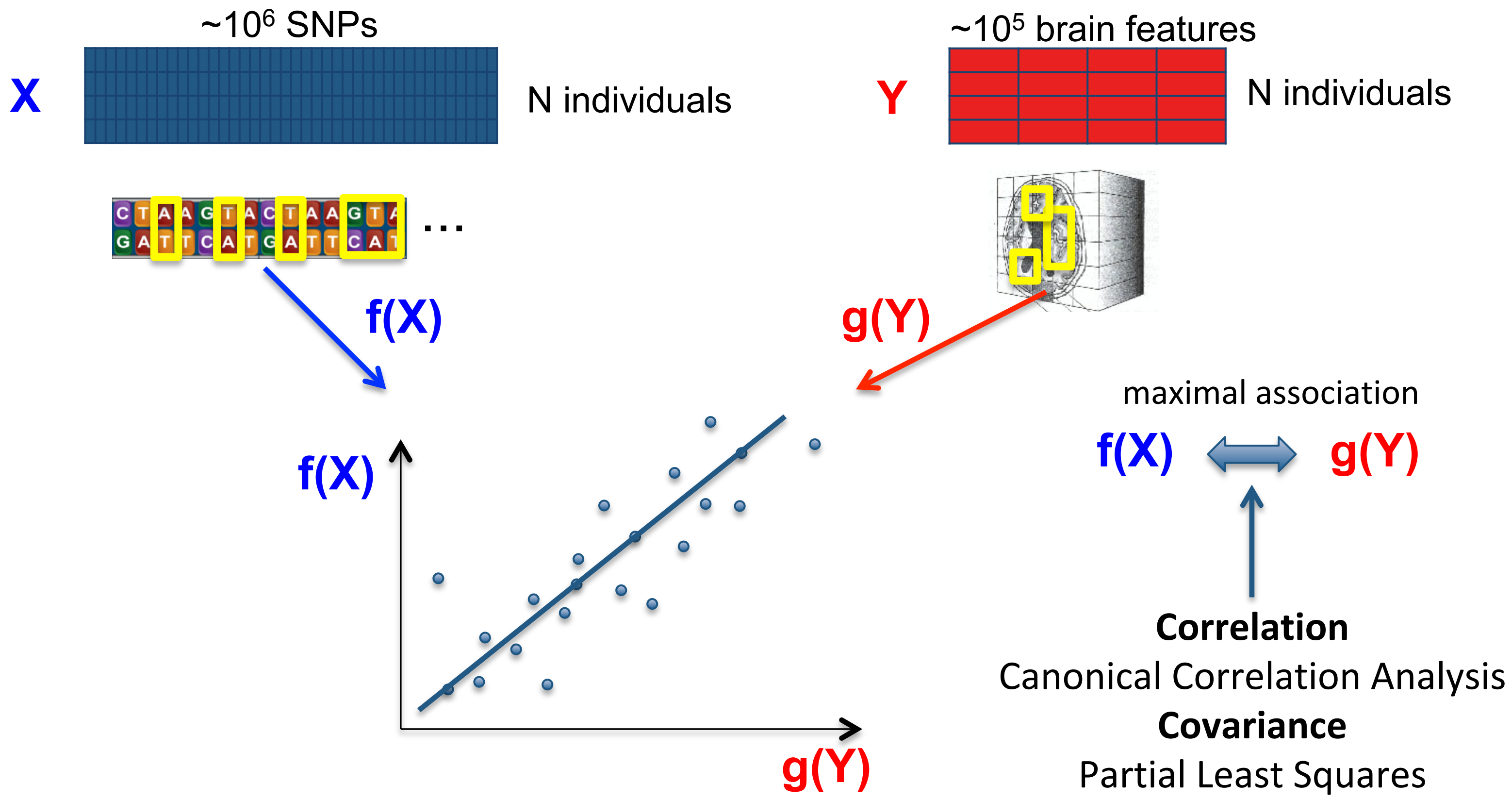
Iterate for > 1’000’000
image locations

- Hard interpretability
- False positive discoveries
- No interaction across brain and genetic areas

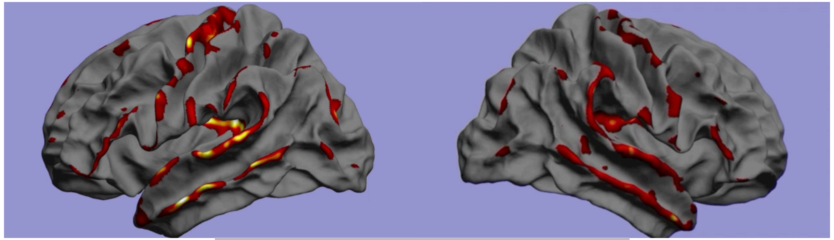
Multivariate modeling and dimensionality reduction

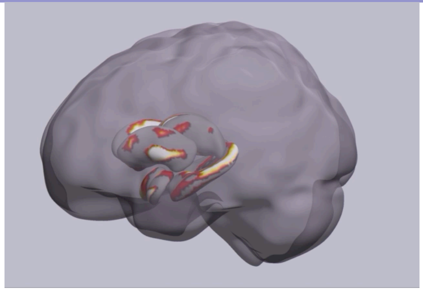


Multivariate modeling and dimensionality reduction

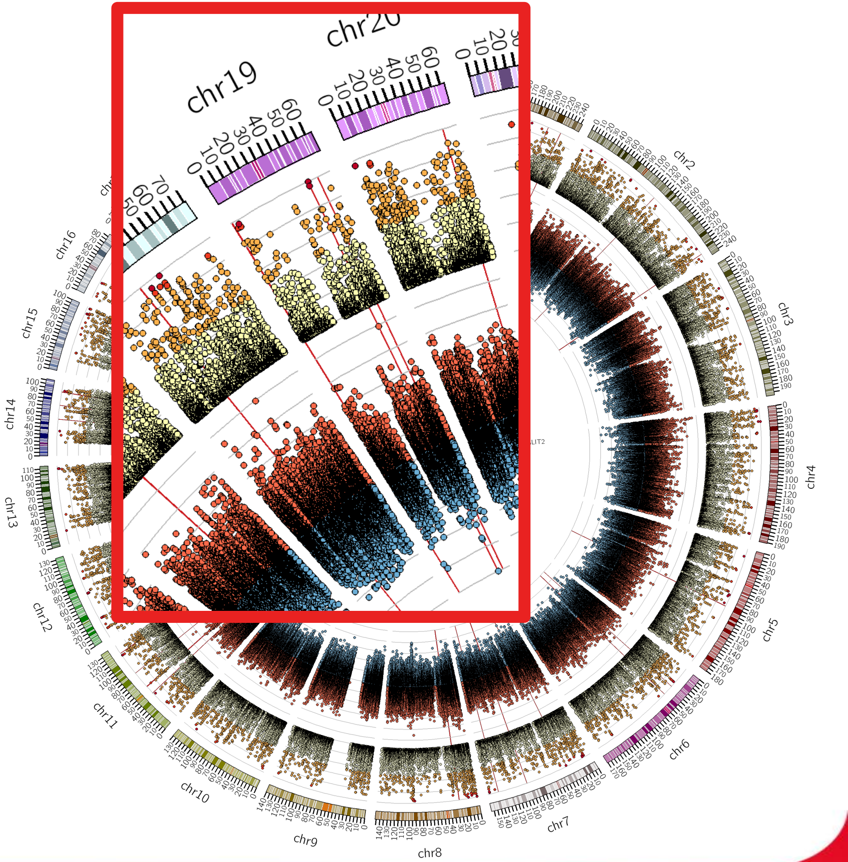


Multivariate Imaging-genetics modeling





Partial least squares association



639 individuals

401 healthy

238 Alzheimer’s

Challenges

How to integrate heterogenous biomedical measures?

Answer from Statistical Learning:

- In principle, data can tell how to combine 1’000’000s of measurements to extract meaningful information
- In practice, we need synergy with other experts: biologist and clinicians
- More data required to enhanced the discovery potential

Challenges

- How to integrate heterogenous biomedical measures

- **How to analyse private healthcare data collected worldwide?**

Large multicentric clinical studies

ENIGMA

Data for ~100’000 individuals

Big Data in medicine

Single hospital: 100s – 1’000s patients

Data from many hospitals needed

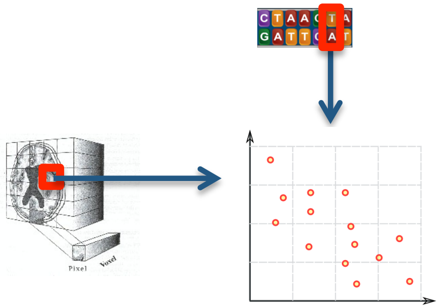
Access to multiple centers data
falls into General Data Protection Regulation (GDPR):
Privacy, confidentiality, security, ...

Data cannot be gathered in a single centre!
Standard learning algorithms cannot be used in multicentric data

Big Data in medicine

Circumventing the problem of data access

Federated-analysis (or meta-analysis)



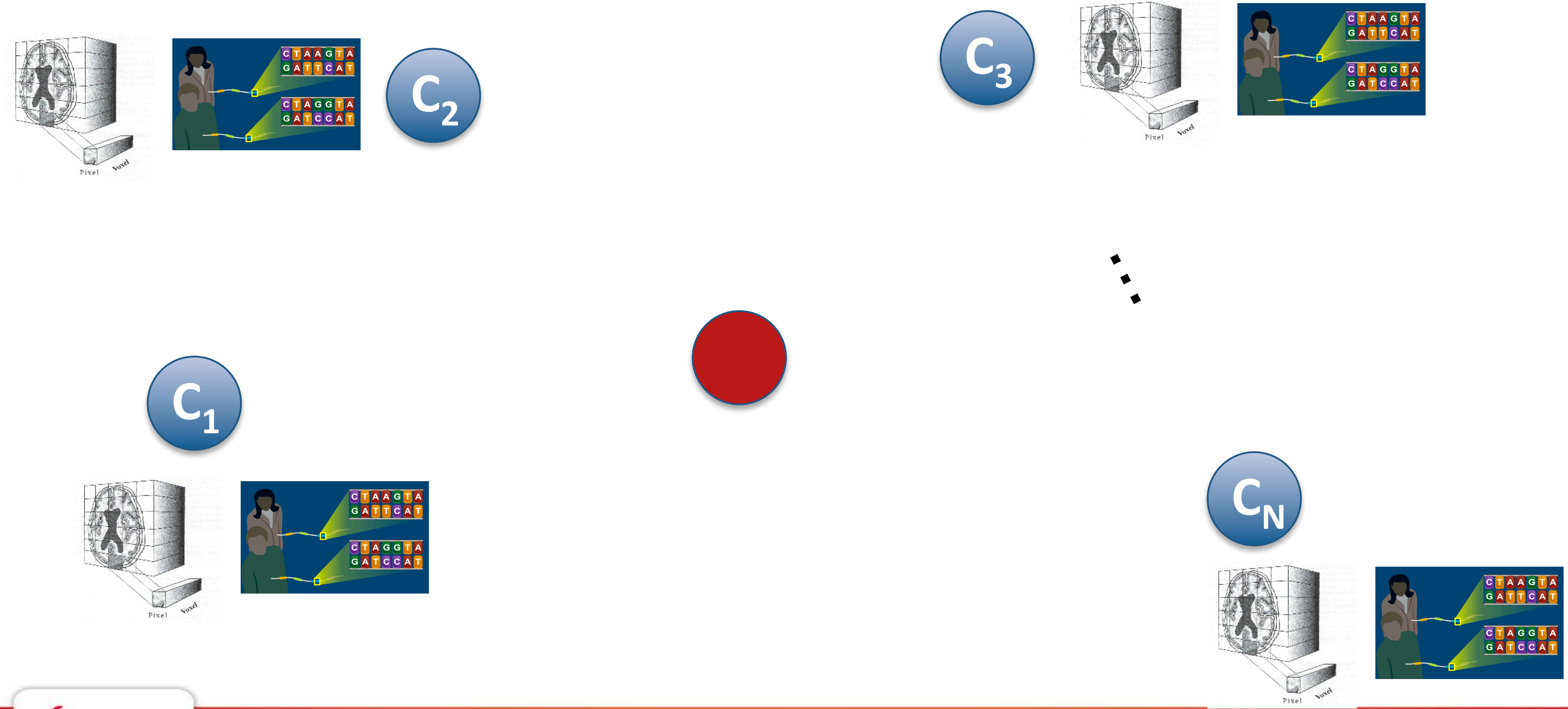
Is the association significant?

Hospital	1	2	3	4	…
Answer	yes	yes	no	yes	…

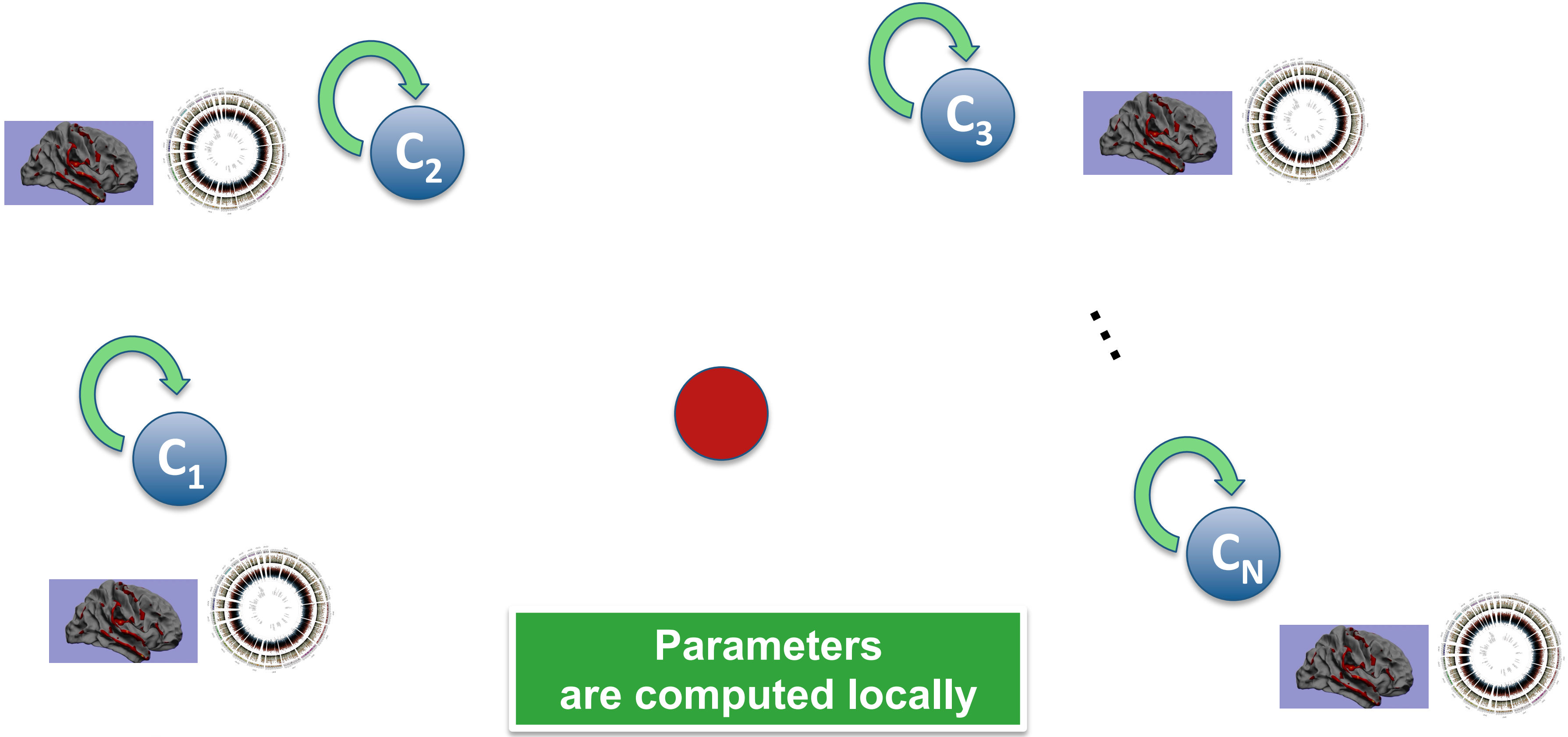
Meta-answer:
yes

- No data sharing
- Ok for standard statistical testing (p-values, effect size)
- No complex modeling possible

Advanced Data Science through federated learning

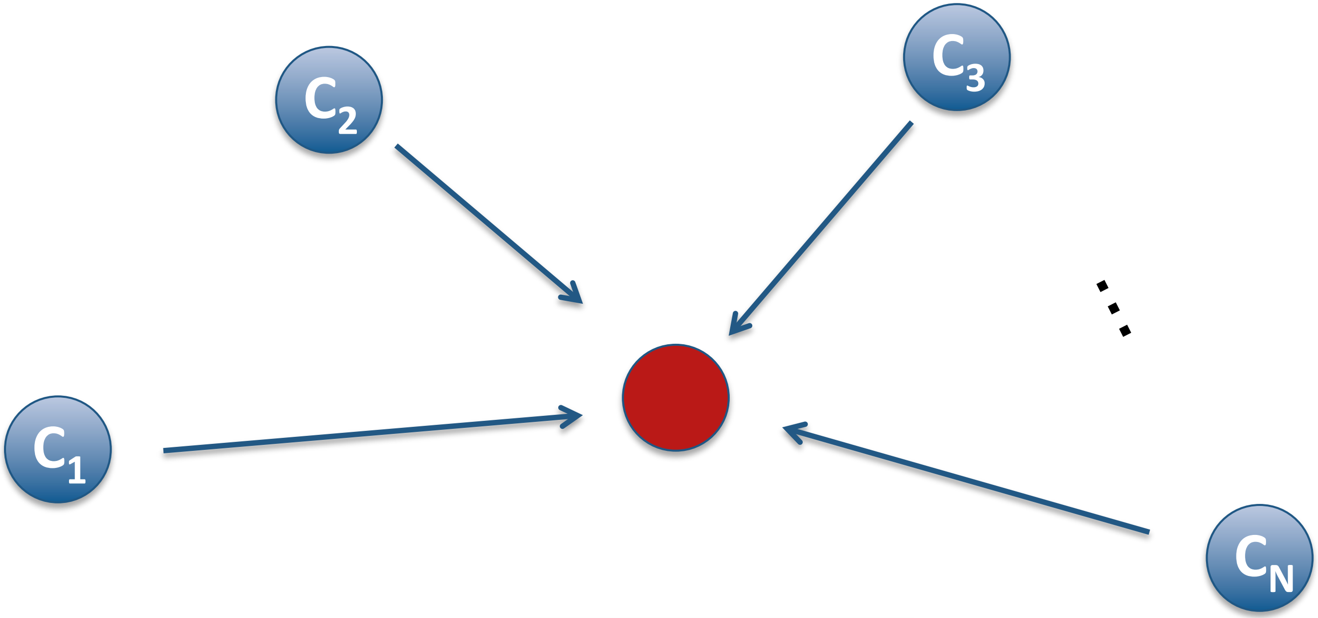


Advanced Data Science through federated learning



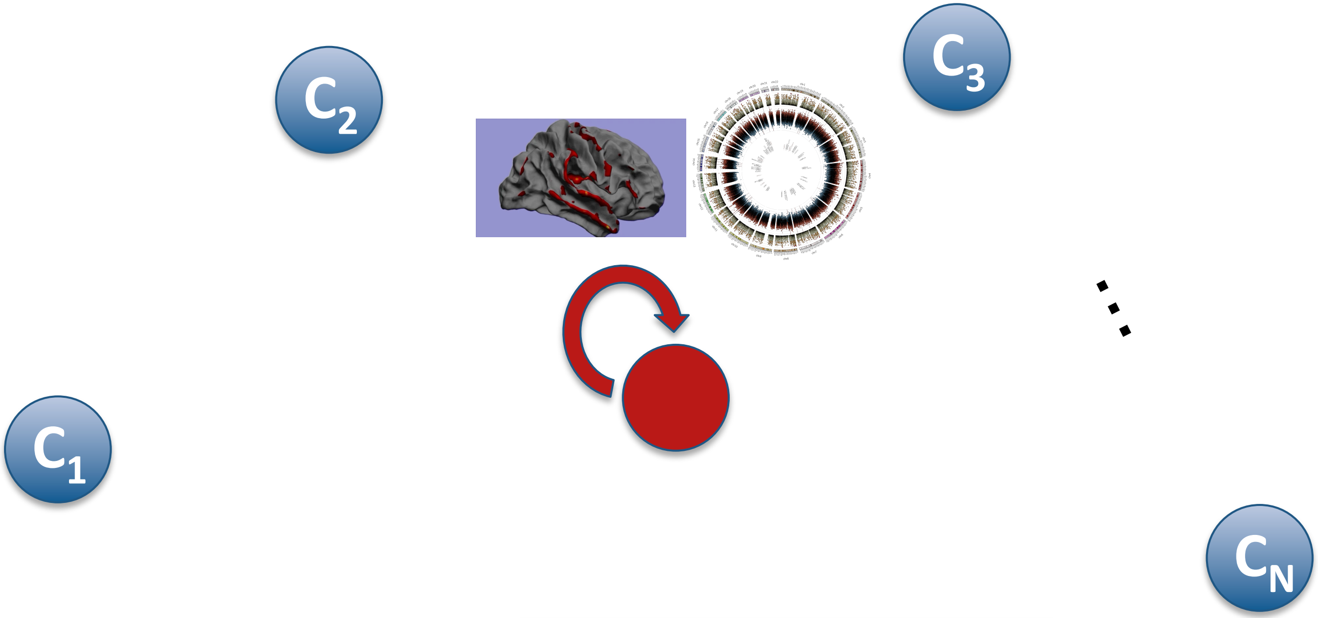
Silva, Gutman, Romero, Thompson, Altmann and Lorenzi. arXiv:1810.08553, 2018

Advanced Data Science through federated learning



Parameters are shared

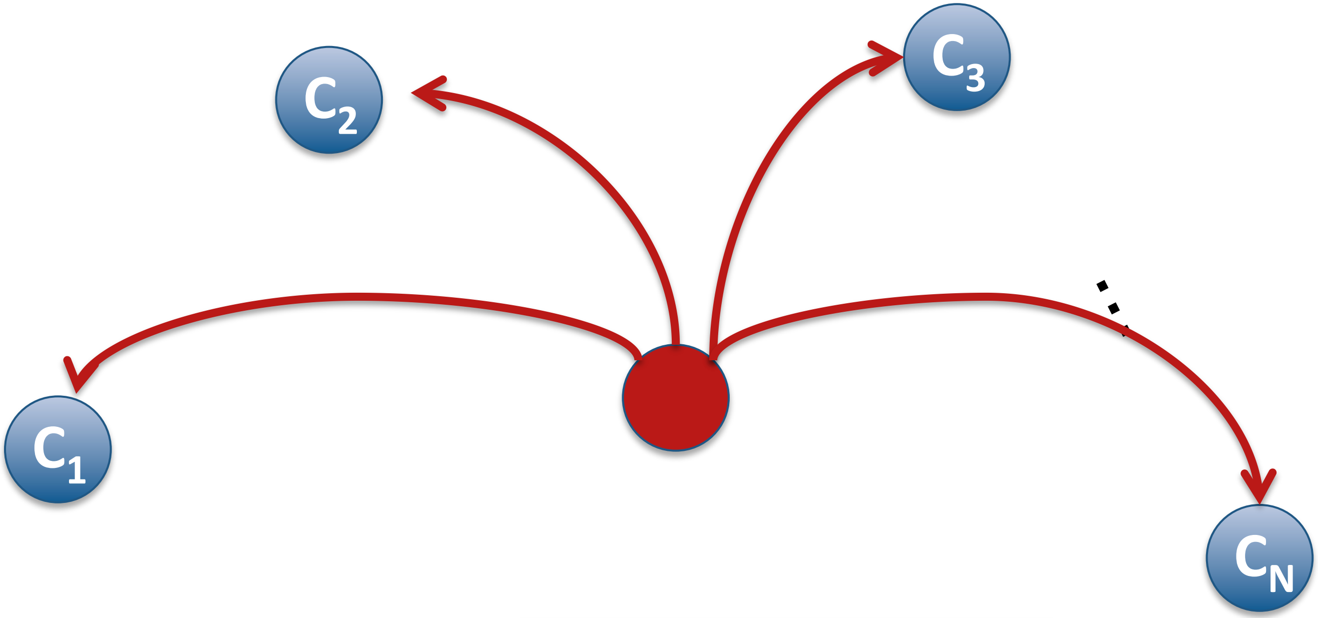
Advanced Data Science through federated learning



A federated model is learned

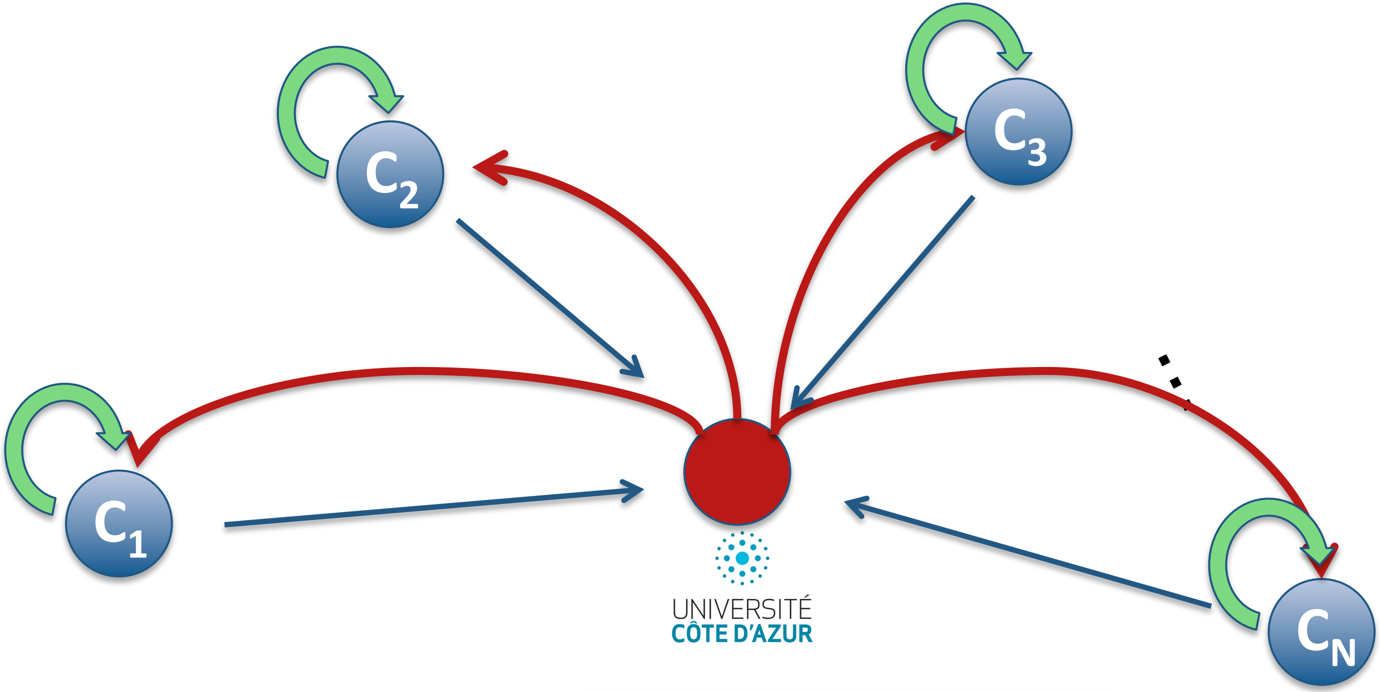
Silva, Gutman, Romero, Thompson, Altmann and Lorenzi. arXiv:1810.08553, 2018

Advanced Data Science through federated learning



The federated model is shared

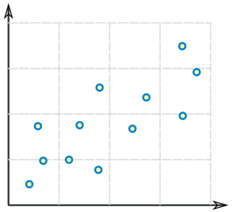
Advanced Data Science through federated learning



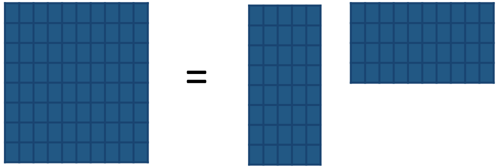
The procedure is iterated

Federated analysis toolkit

A methodology for distributed



Linear modeling



Matrix factorization

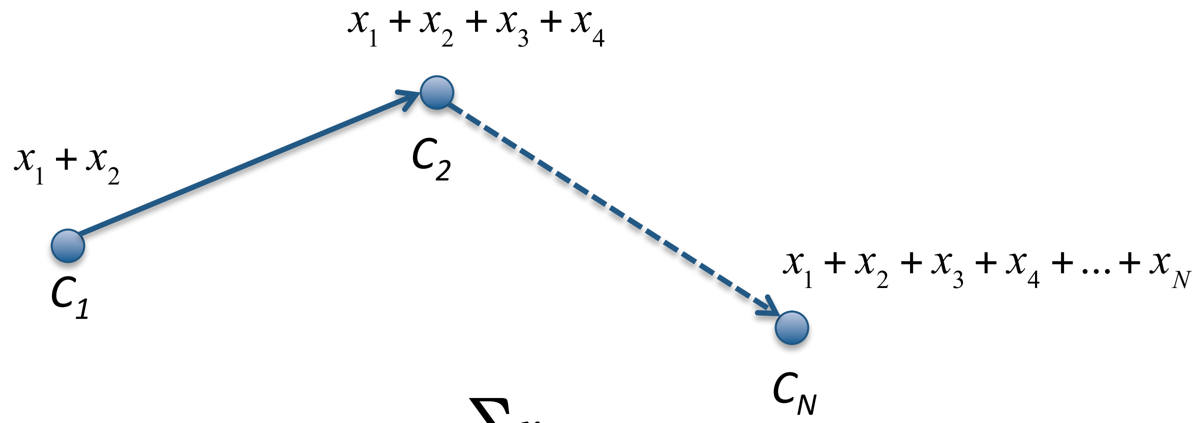
Allows a federated framework for several key statistical operations:
Data standardization, accounting for covariates, dimensionality reduction, ...

Standard statistical pipeline in multivariate analysis

- Data standardization
- Confounding effect correction
- Multivariate analysis

Federated moment estimation: Mean

$$\frac{x-\mu}{\sigma}$$



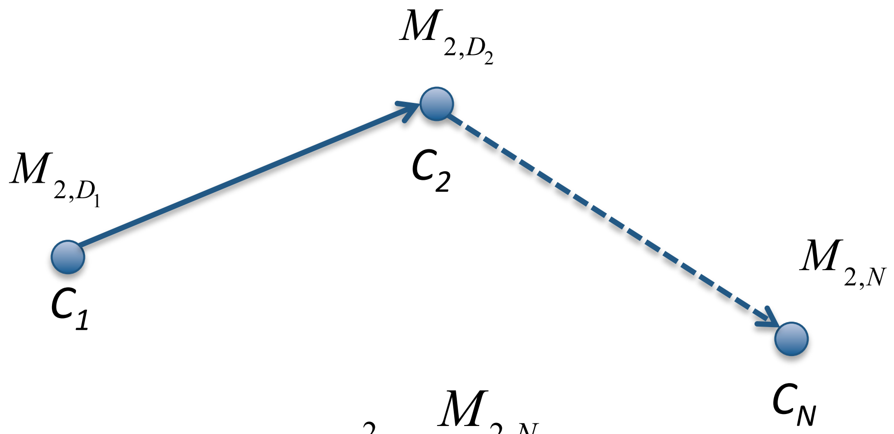
$$\mu = \frac{\sum x_i}{N}$$

Federated moment estimation: SD

$$\frac{x-\mu}{\sigma}$$

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N-1}$$

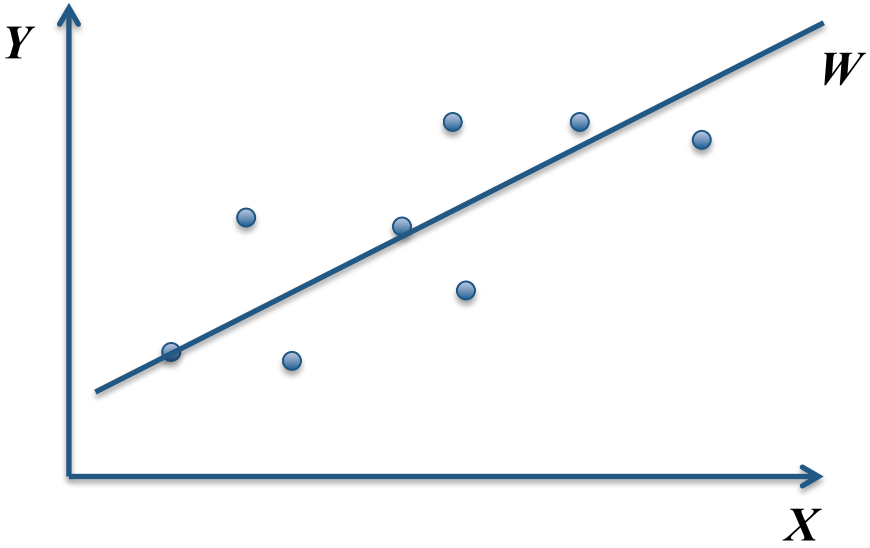
$$M_{2,n} = M_{2,n-1} + (x_n - \overline{x}_{n-1})(x_n - \overline{x}_n)$$



$$\sigma^2 = \frac{M_{2,N}}{N}$$

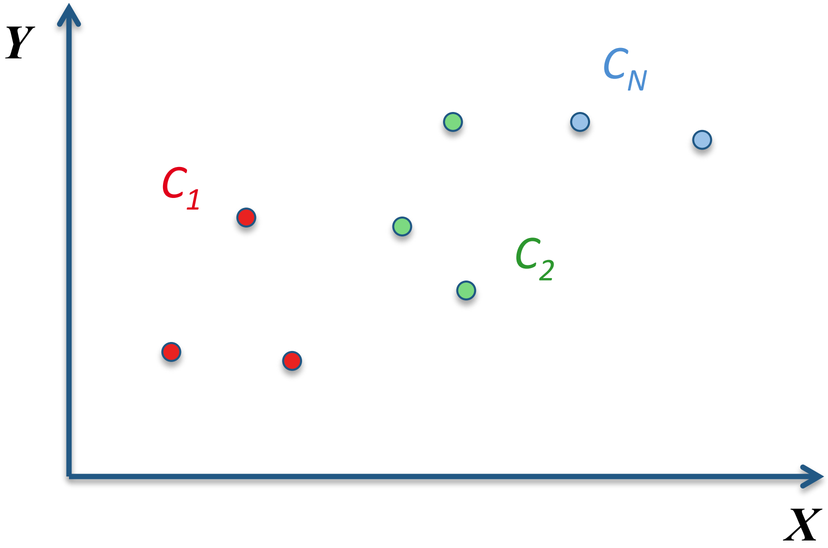
Federated linear modeling

$$L(Y \mid X, W) = \left\| Y - XW^T \right\|^2 \qquad W = (X^T X)^{-1} X^T Y$$



Federated linear modeling

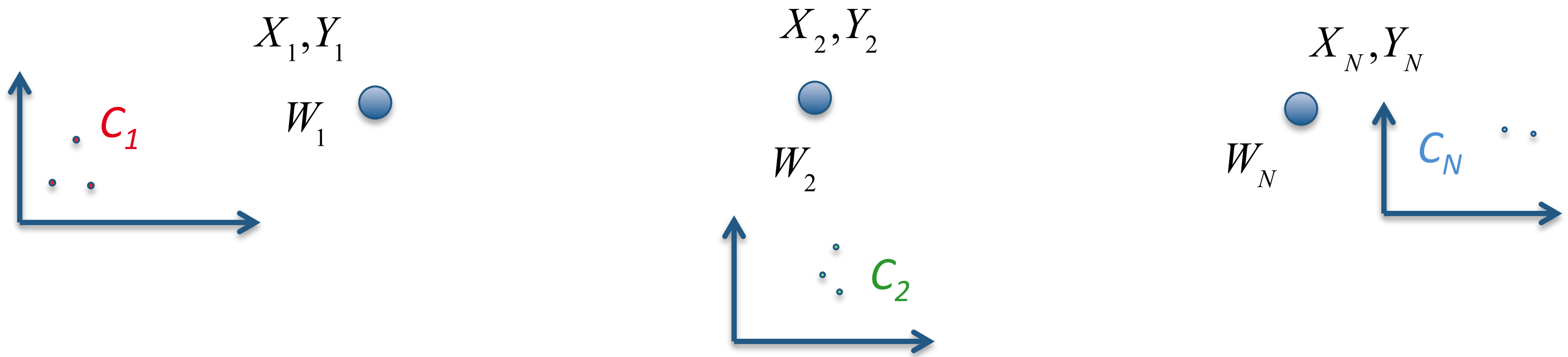
$$L(Y \mid X, W) = \left\| Y - XW^T \right\|^2 \qquad W = (X^T X)^{-1} X^T Y$$



Federated linear modeling

$$L(Y \mid X, W) = \left\| Y - XW^T \right\|^2$$

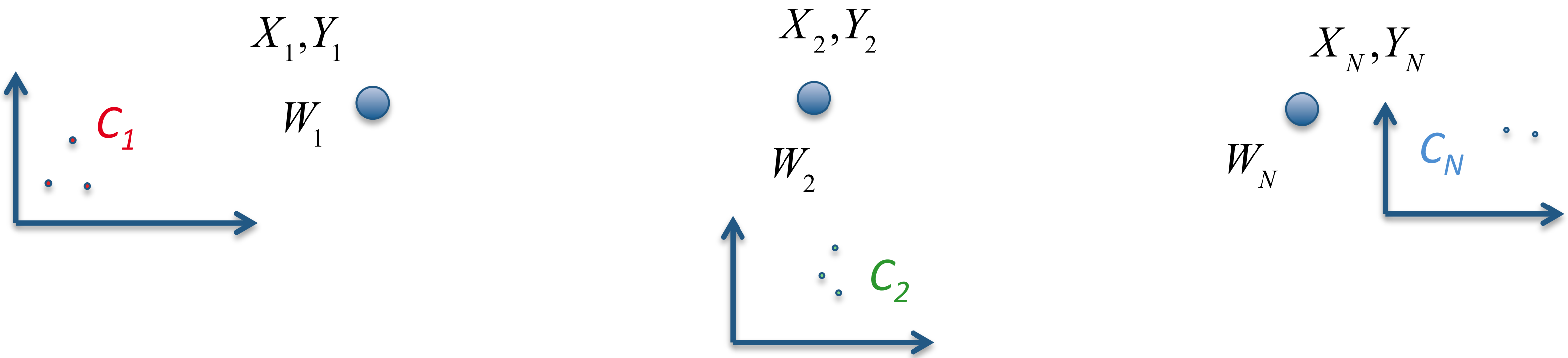
$$W = (X^T X)^{-1} X^T Y$$



Federated linear modeling

$$L(Y \mid X, W) = \left\| Y - XW^T \right\|^2$$

$$L(Y_c \mid X_c, W_c) = \left\| Y_c - X_c W_c^T \right\|^2$$

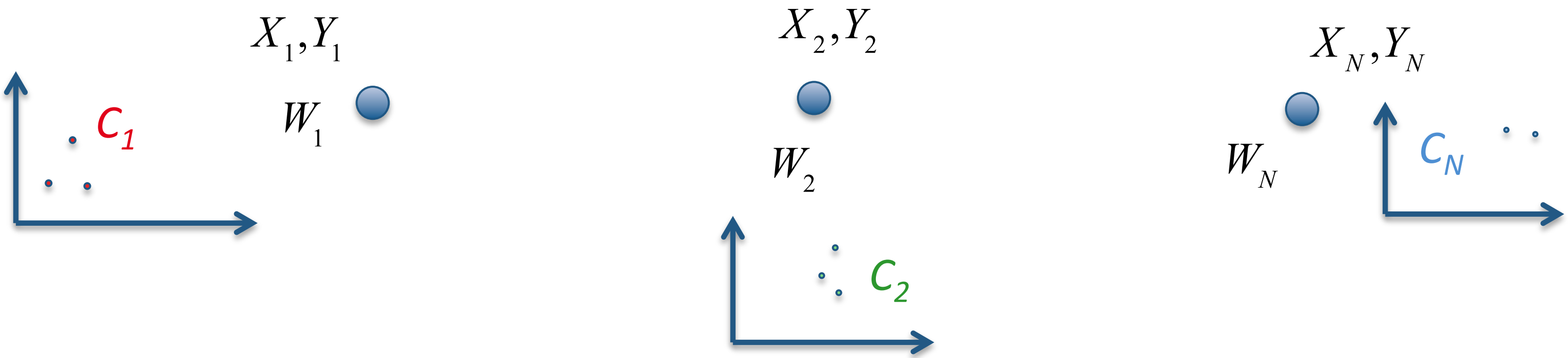


Federated linear modeling

$$L(Y \mid X, W) = \left\| Y - XW^T \right\|^2$$

$$L(Y_c \mid X_c, W_c) = \left\| Y_c - X_c W_c^T \right\|^2$$

$$\boxed{W_c = W}$$



Alternating direction method of multipliers

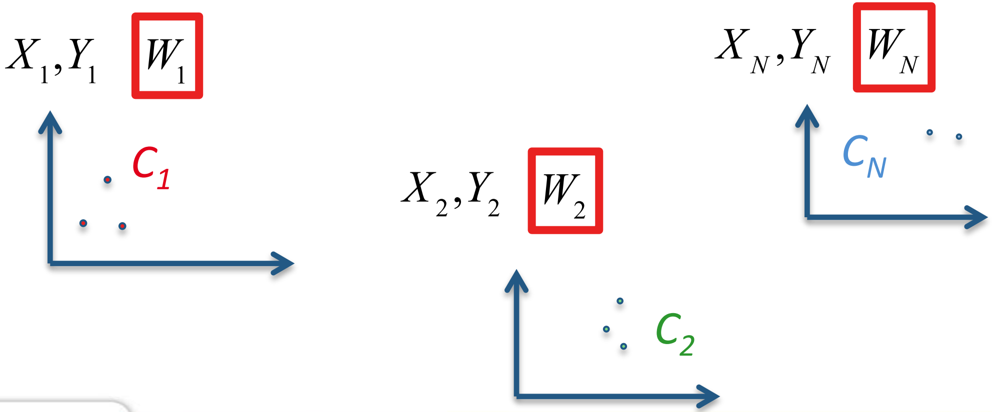
$$L_\rho(W_c, \tilde{W}, \alpha) = \sum_c L(Y_c \mid X_c, W_c) + \left\langle \alpha_c, W_c - \tilde{W} \right\rangle + \frac{\rho}{2} \left\| W_c - \tilde{W} \right\|_2^2$$

Alternating direction method of multipliers

$$L_\rho(W_c,\tilde{W},\alpha)=\sum_c L(Y_c\mid X_c,W_c)+\left\langle \alpha_c,W_c-\tilde{W}\right\rangle+\frac{\rho}{2}\left\|W_c-\tilde{W}\right\|_2^2$$

Iteratively:

$$W_c^{(k+1)}=\operatorname{argmin}_{W_c} L_\rho(W_c,\tilde{W}^{(k)},\alpha_c^{(k)})=(X_c^T X_c+\frac{\rho}{2}Id)^{-1}(X_c^T Y_c-\frac{1}{2}\alpha_c^{(k)}+\frac{\rho}{2}\tilde{W}^{(k)})$$

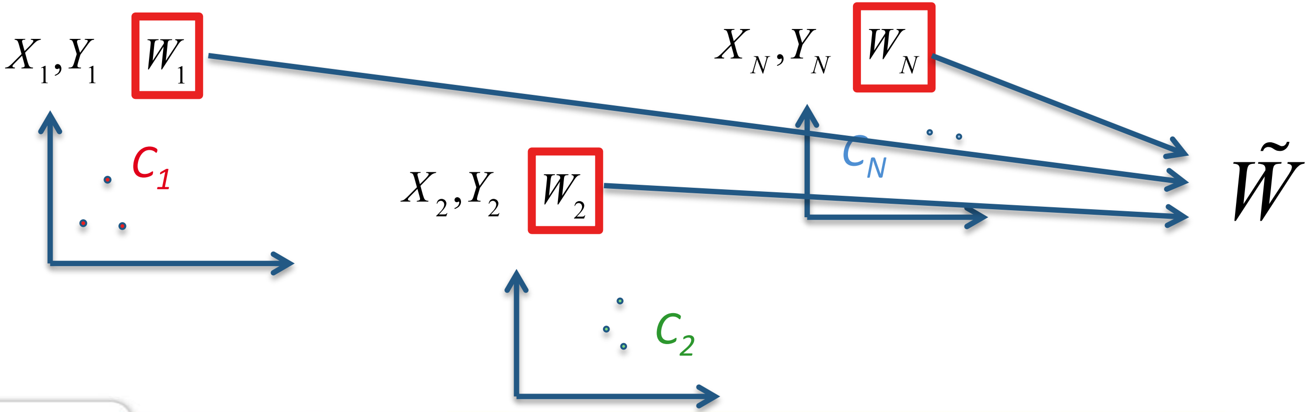


Alternating direction method of multipliers

$$L_\rho(W_c,\tilde{W},\alpha) = \sum_c L(Y_c \mid X_c, W_c) + \left\langle \alpha_c, W_c - \tilde{W} \right\rangle + \frac{\rho}{2}\left\| W_c - \tilde{W} \right\|_2^2$$

Iteratively:

$$\tilde{W}^{(k+1)} = \operatorname{argmin}_{\tilde{W}} \; L_\rho(W_c^{(k+1)}, \tilde{W}, \alpha_c^{(k)}) = \frac{1}{C}\sum \frac{\alpha_c^{(k)}}{\rho} + W_c^{(k+1)}$$



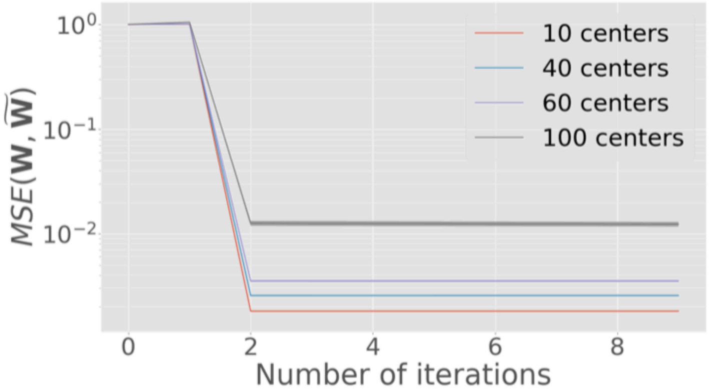
Alternating direction method of multipliers

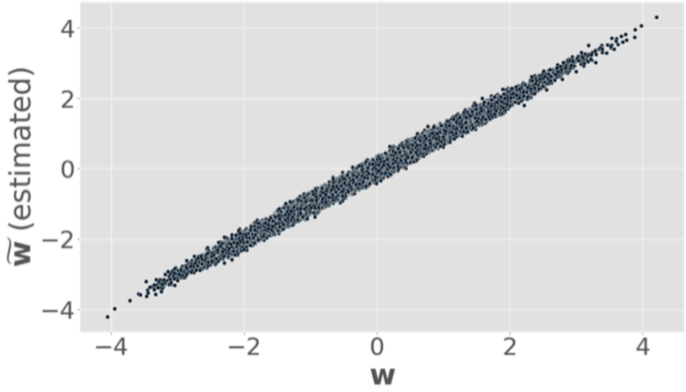
$$L_\rho(W_c,\tilde{W},\alpha) = \sum_c L(Y_c \mid X_c, W_c) + \left\langle \alpha_c, W_c - \tilde{W} \right\rangle + \frac{\rho}{2}\left\| W_c - \tilde{W} \right\|_2^2$$

Iteratively:

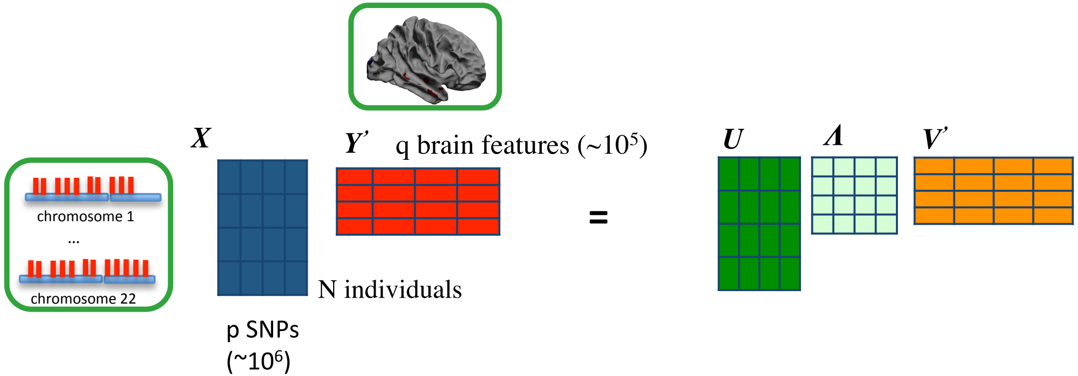
$$\alpha_c^{(k+1)} = \alpha_c^{(k)} + \rho(W_c^{(k+1)} - \tilde{W}^{(k+1)})$$

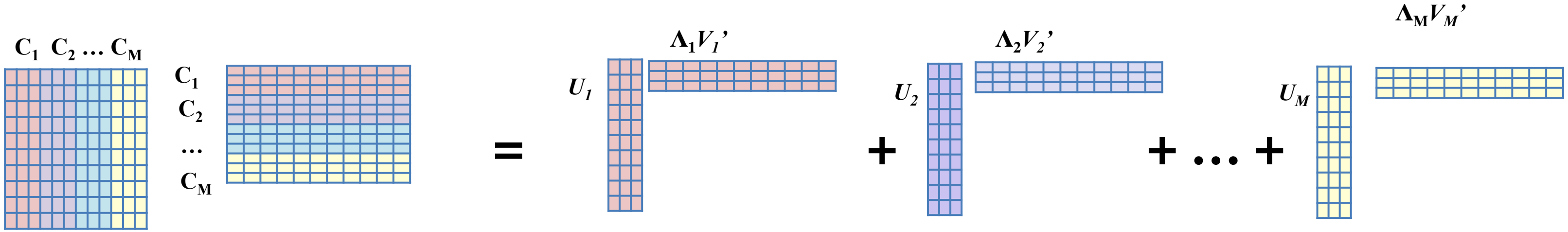
Results on synthetic tests





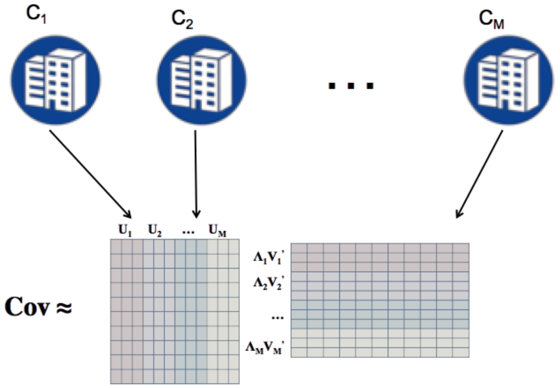
Covariance estimation and eigen-decomposition



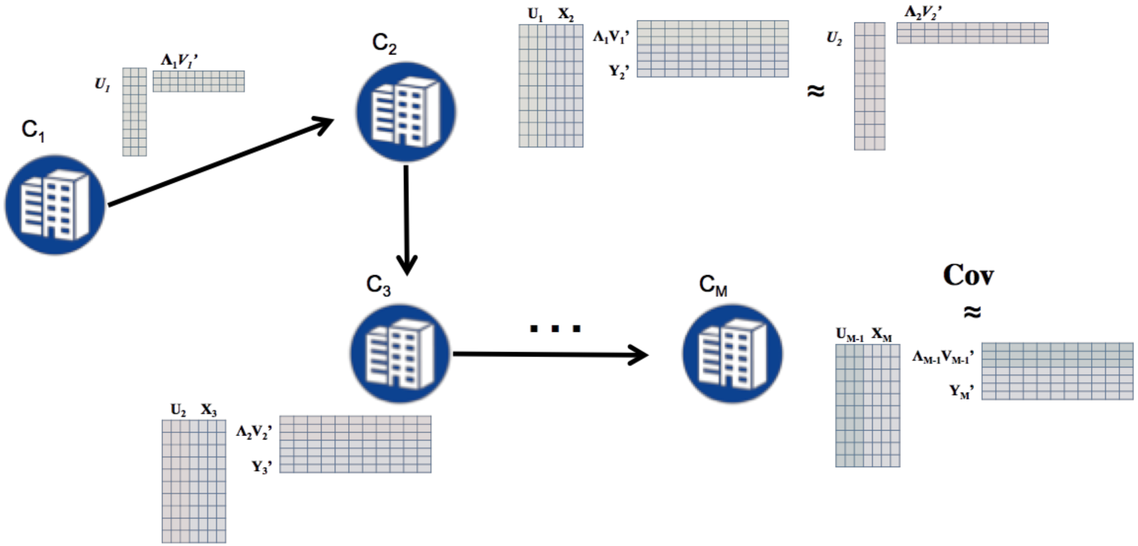


Covariance estimation and eigen-decomposition

Meta PLS

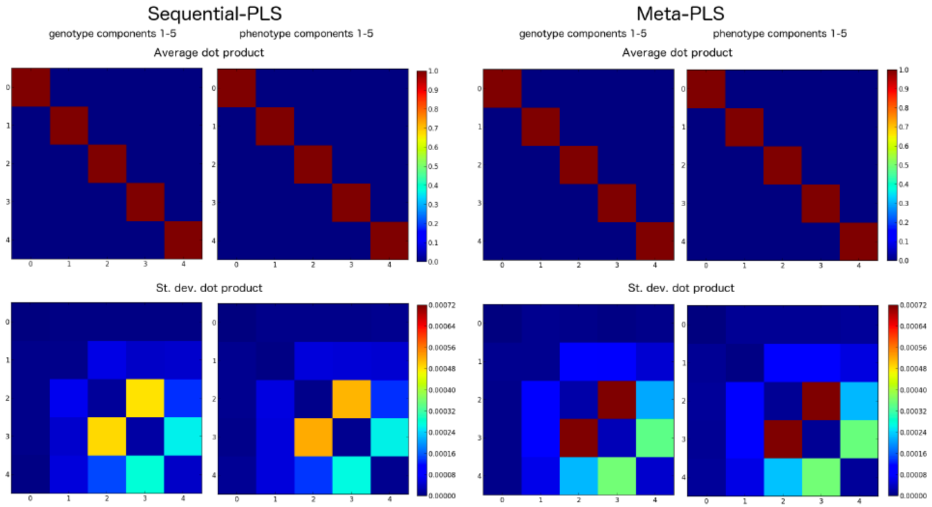


Sequential PLS

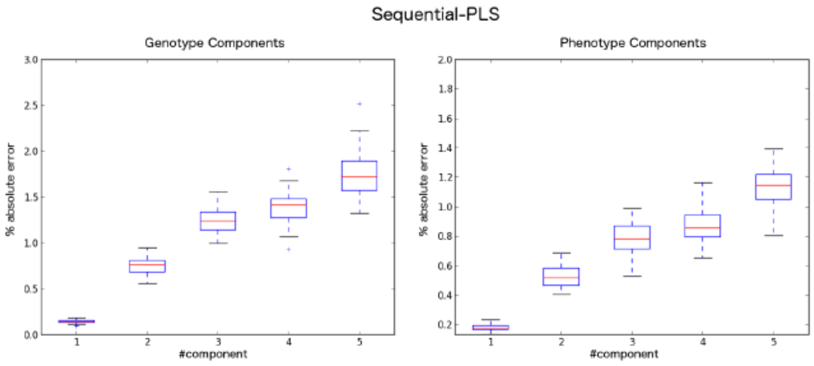


Testing

Mean and sd of dot product



Absolute feature-wise error

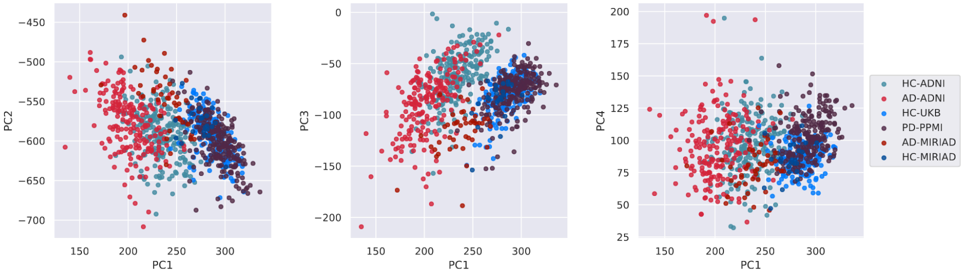


Lorenzi et al. MASAMB 2016

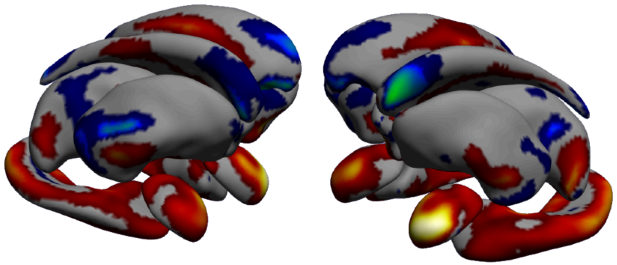
Federated analysis of subcortical brain regions in dementia

ADNI	PPMI	UK Biobank	Miriad
Alzheimer's	Parkinson's	Healthy	Alzheimer's
802	232	208	68

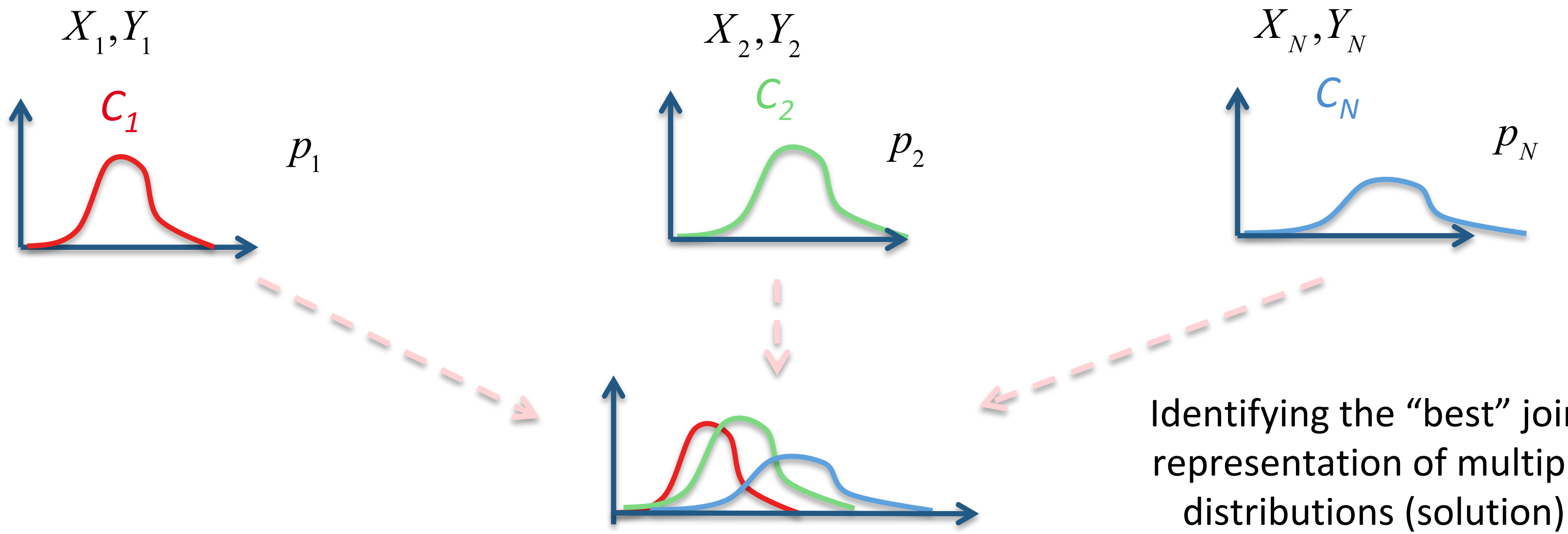
Projection on latent components



Brain subcortical components



Next steps: federated learning of parameters distribution



Identifying the "best" joint representation of multiple distributions (solution)

- Extension to Bayesian learning framework
- Adaptation of Gaussian processes and Bayesian Neural Nets
- Modeling **bias**

Internship/PhD offer

Bias of Federate Learning with Heterogeneous and Missing Data

During the project the candidate will:

- Develop learning methods for federated analysis for private and distributed data;
- Develop a formalism for bias identification and correction in federated learning;
- Gather knowledge in advanced statistical learning methods - Bayesian learning, Kernel methods, non-parametric learning, variational inference -;
- Develop and deploy algorithms in several context, with special focus in biomedical and clinical application;
- Participate to the activity of Accenture Labs, interact with the research and engineering personnel;
- Interact with INRIA students and researchers, and participate to scientific life of the team.

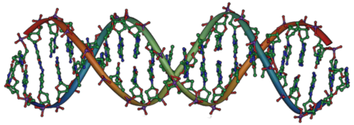
Challenges

How to analyse private healthcare data collected worldwide?

Answer from Statistical Learning:

- Advanced statistical modeling can be compatible with Data privacy and anonymity
- Federated learning applies for simple linear models as well as for more complex ones (e.g. neural networks or Gaussian processes)
- Future research needed for improving data security, data transfer bottlenecks, modeling flexibility

Conclusions



+

+

- Bridging domains through statistical learning
- Integrating biomedical hypothesis is fundamental:
  plausibility, interpretability, no Big Data (yet)

Inria

UNIVERSITÉ CÔTE D'AZUR

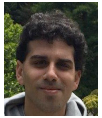
Epione
e-patient / e-medicine

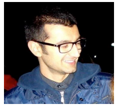
mNC³


N. Ayache


P. Robert


V. Manera


C. Abi Nader


L. Antelmi

Meta-ImaGen


S.S. Silva

UCA MDLab

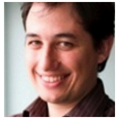

M. Milanesio

UCA-Ville de Nice Young Researcher award


S. Garbarino

Brain-Heart


J. Banus


M. Sermesant

EURECOM
Sophia Antipolis


M. Filippone

UCL


A. Altmann

ILLINOIS INSTITUTE OF TECHNOLOGY


B. Gutman

HUG Hôpitaux Universitaires Genève


G.B Frisoni


F. Ribaldi

Inria

UNIVERSITÉ CÔTE D'AZUR

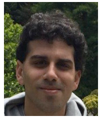
Epione
e-patient / e-medicine

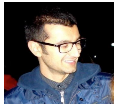
mNC3


N. Ayache


P. Robert


V. Manera


C. Abi Nader


L. Antelmi

Meta-ImaGen


S.S. Silva

UCA MDLab

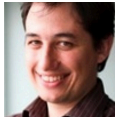

M. Milanesio

UCA-Ville de Nice
Young Researcher award


S. Garbarino

Brain-Heart


J. Banus


M. Sermesant

EURECOM
Sophia Antipolis


M. Filippone

UCL


A. Altmann

ILLINOIS INSTITUTE OF TECHNOLOGY


B. Gutman

HUG Hôpitaux Universitaires Genève


G.B Frisoni


F. Ribaldi

Thank you!