Apprentissage automatique sur les bases de données biomédicales pour améliorer la compréhension des maladies neurodégénératives

C@fé-In
Sophia-Antipolis

Marco Lorenzi

Université Côte d'Azur
Inria Sophia Antipolis, Asclepios Research Project

William Utermolhen (1933-2007)

Self-portrais

1967

1996

1997

1998

1999

2000

1995: Alzheimer’s disease diagnosis

Alzheimer's disease: the most common form of dementia

Language problems

Memory loss

Functionality loss

Apraxia

Cognitive impairment

Mood alterations

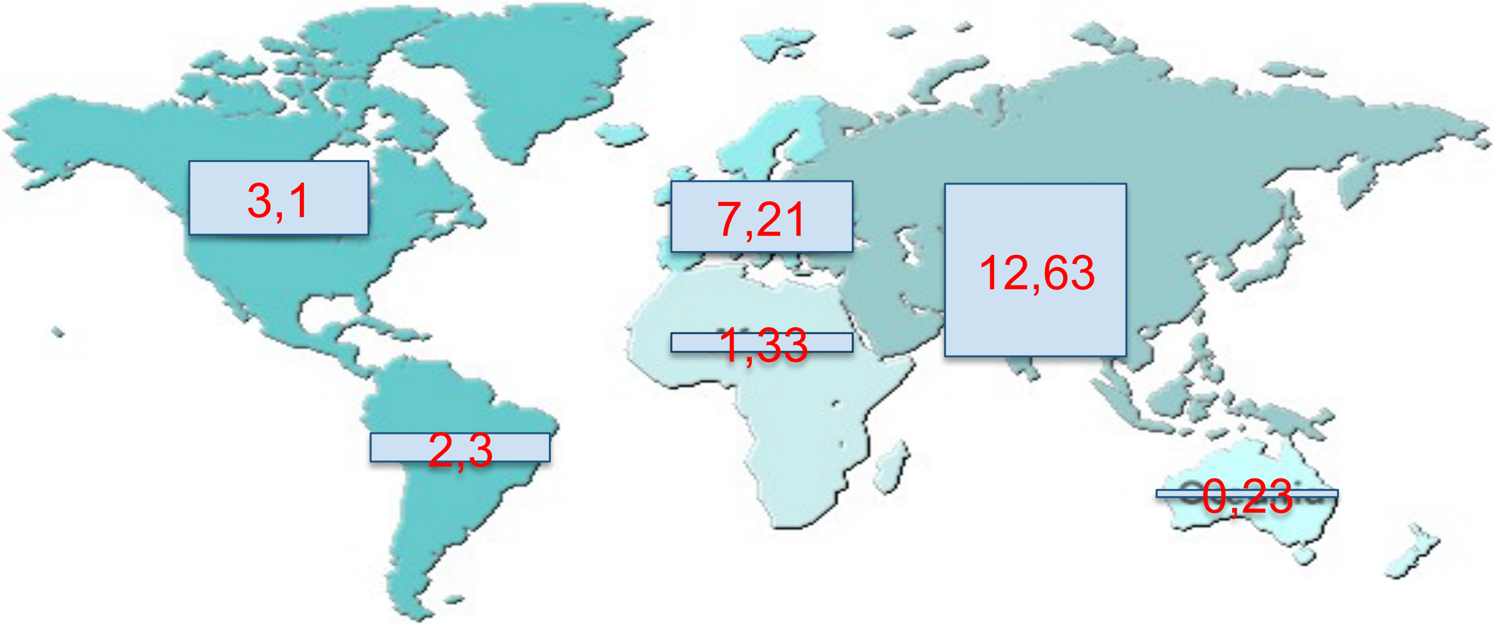
Enormous human and societal cost

The disease with the largest economic impact (Europe et US)

Impact on families
~20,000 $ every year in 1998
[Moore et al., J Gerontol B Psychol Sci Soc Sci 1998]

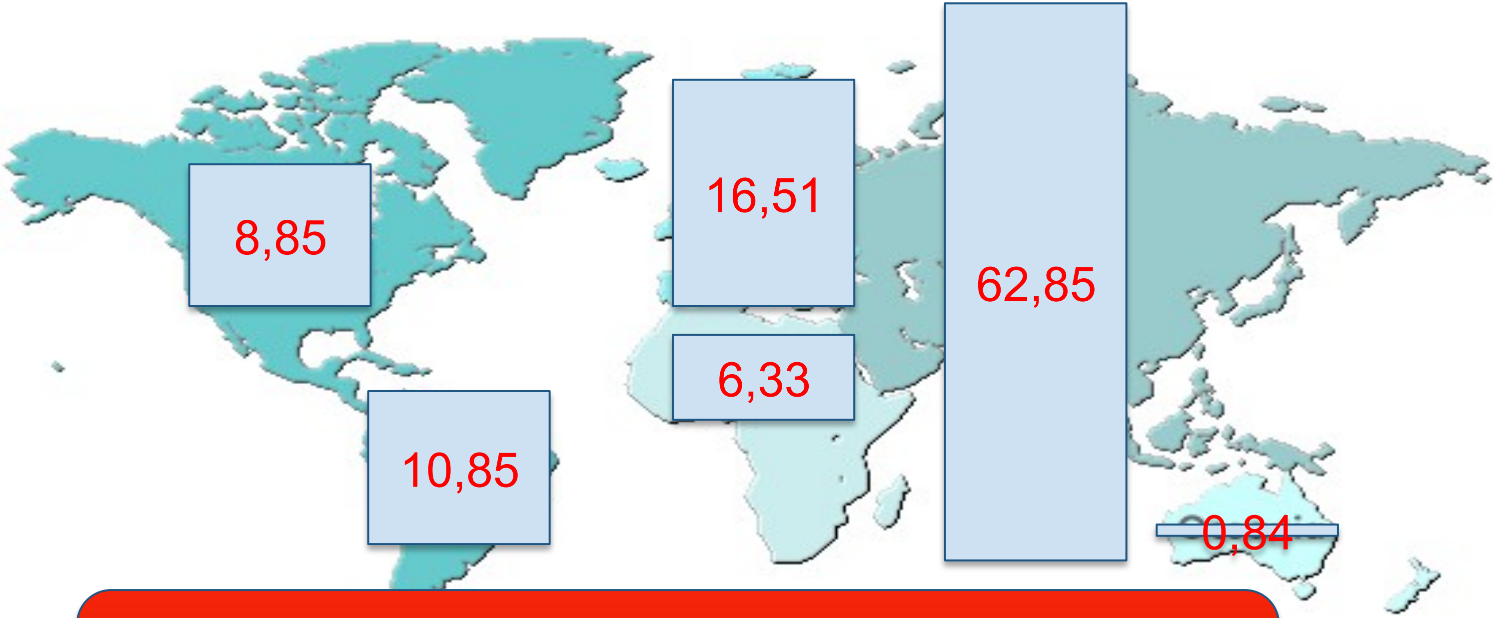
Health-care
160 billion $ every year worldwide
[Wimo et al., Dement Geriatr Cogn Disord 1998]

People affected in the world
26,6 millions in **2006**



[Brookmeyer et al., Alzheimers and Dementia 2007]

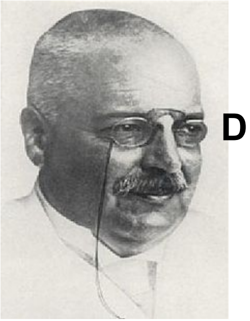
People affected in the world
106 millions en 2050



"Looming epidemy"
2017
No cures nor preventive measures

[Brookmeyer et al., Alzheimers and Dementia 2007]

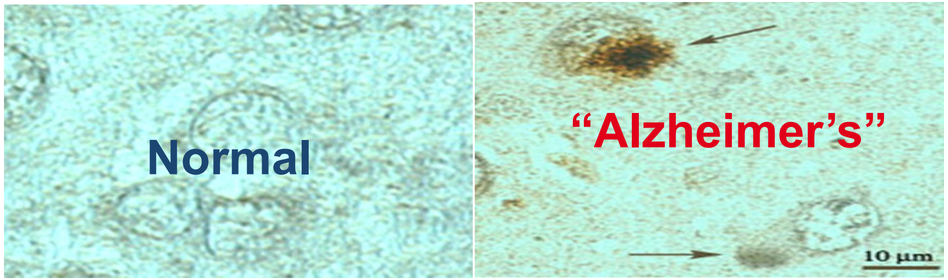
Urgent need: understanding the disease



Dr. Aloysius "Alois" Alzheimer (1864-1915)

Auguste Deter (1850-1906)

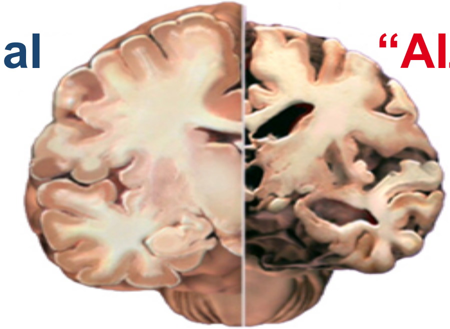
Amyloid plaques & neurofibrillary tangles



[Kahn et al, PNAS 2007]

Brain atrophy

Normal

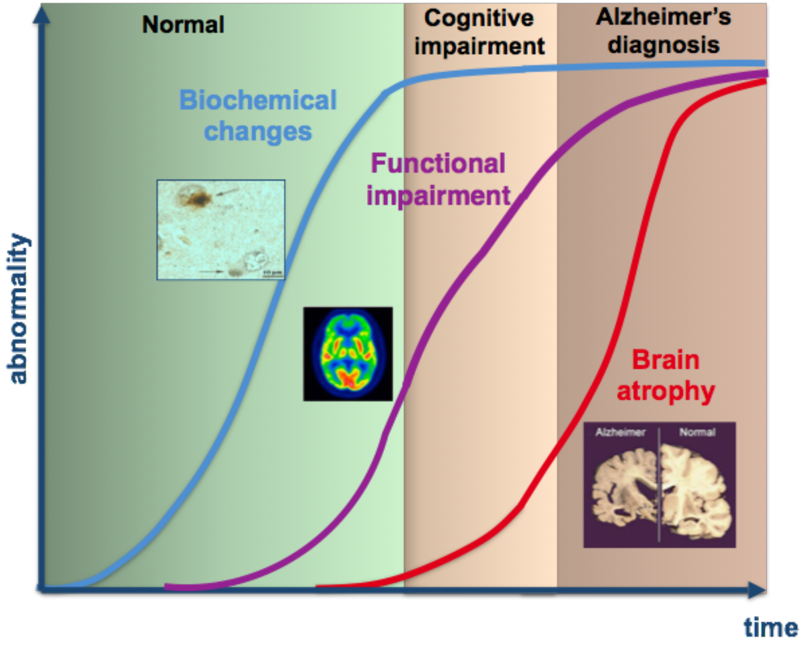


"Alzheimer's"

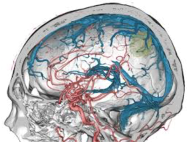
source http://www.alz.org

A story with several actors

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

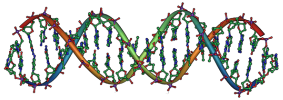


Sociodemographic



Vascularity

?

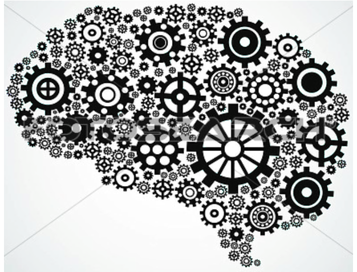


Genetics

Microbiome

…

Multifactorial processes



Disentangling the pathological mechanisms
drug discovery

Patient stratification (diagnostic)
effective clinical trials

Approaches

Forward

Models → Data

- Targeted
- Testing “mechanistic” hypothesis
- ☹ Difficult to account for several factors

Backward

Data → Models

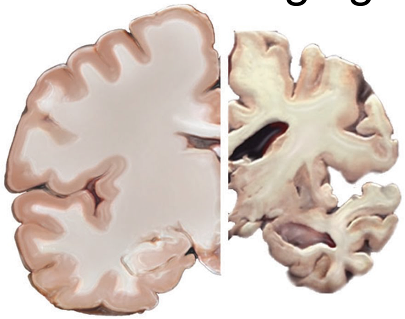
- Exploring unknown interactions
- Based on inferential methods
- ☹ Generalization and validation

A research challenge

Data science
Statistical learning

\+

Biomedical research
Neuroimaging



Combine heterogeneous data and observations for:

- **Improve the understanding of the disease**
- **Better treatment**
- **Better diagnostic**

Joint modeling of brain and genetic data in Alzheimer’s disease

- Ingredients -

- Data (disease markers)
- Algorithms
- Databases

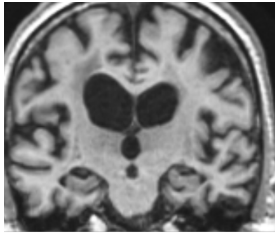
Joint modeling of brain and genetic data in Alzheimer's disease

- Ingredients -

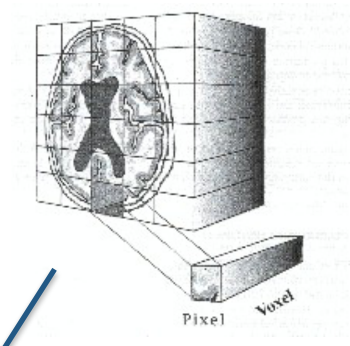
- **Data (disease markers)**
- Algorithms
- Databases

Brain imaging

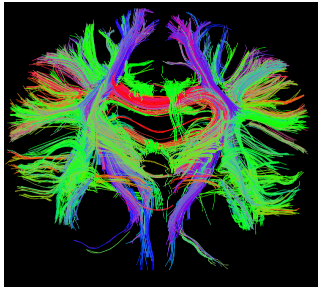
Quantify the **brain structure**



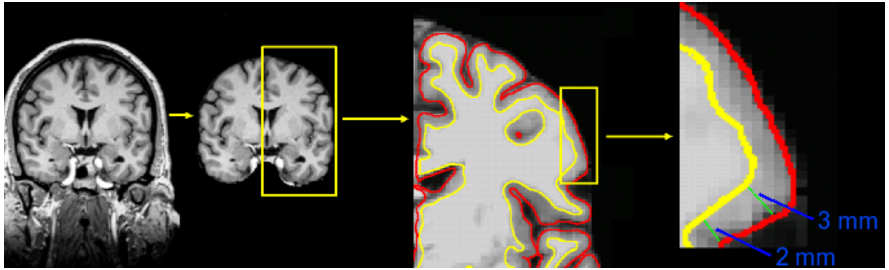
baseline



Grey matter



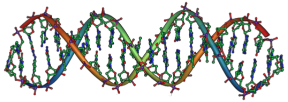
Connectivity

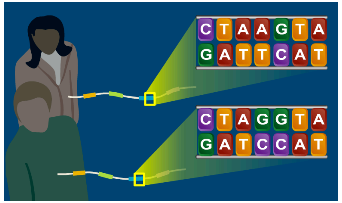


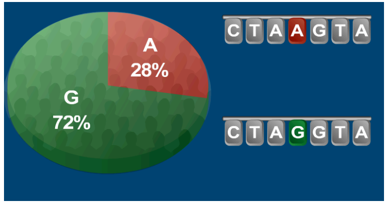
Brain cortical thickness

Genetics

Identifying meaningful **genetic variants**
(Single Nucleotide Polymorphism -SNP-) in a population

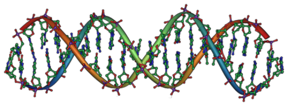


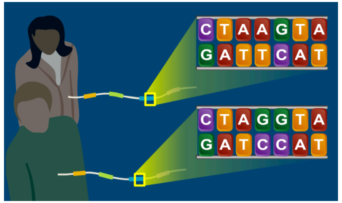


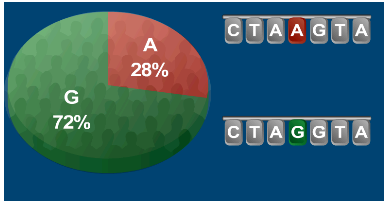


Genetics

Identifying meaningful **genetic variants**
(Single Nucleotide Polymorphism -SNP-) in a population

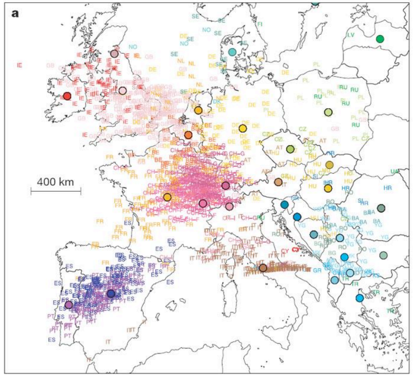






Discovering the encoded information

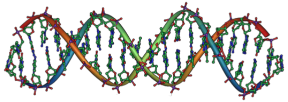
Heritability

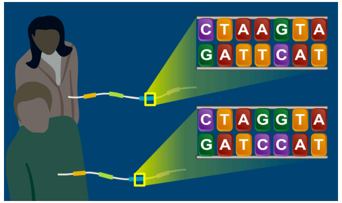


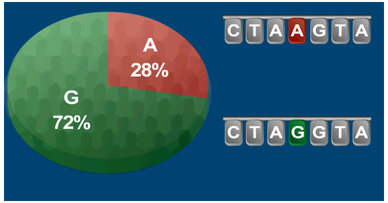
Novembre et al, Nature, 2008

Genetics

Identifying meaningful **genetic variants**
(Single Nucleotide Polymorphism -SNP-) in a population

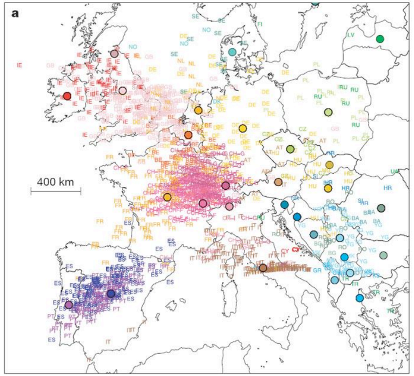






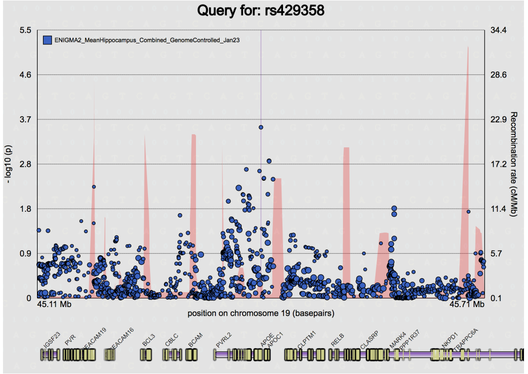
Discovering the encoded information

Heritability



Novembre et al, Nature, 2008

Association with a disease

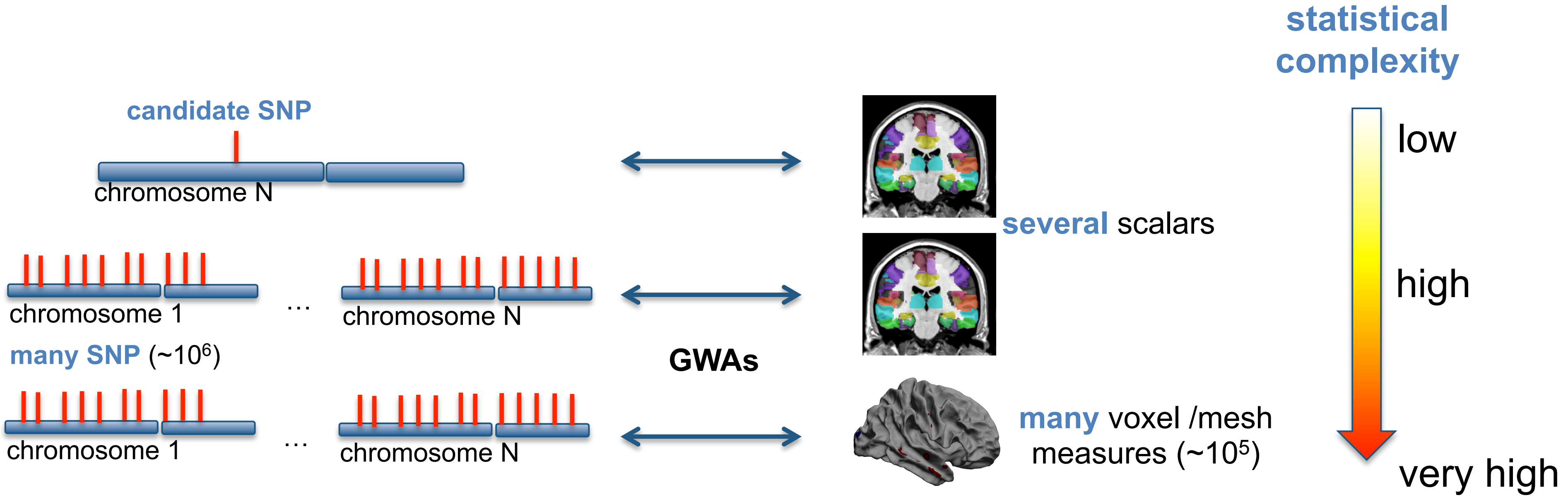


Joint modeling of brain and genetic data in Alzheimer’s disease

- Ingredients -

- Data (disease markers)
- **Algorithms**
- Databases

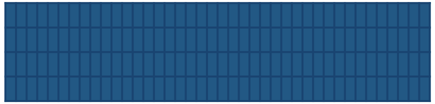
Association between SNP and brain features



Multivariate Association studies

Maximizing the joint relationship between genetic variants and brain features

~10^6 SNPs

$\mathbf{X}$ =  N individuals

~10^5 brain features

$\mathbf{Y}$ =  N individuals

Partial least squares (PLS)

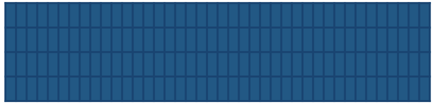
$$\max_{p,q} \mathrm{Cov}(\mathbf{X} \cdot p, \mathbf{Y} \cdot q)$$

Liu et al, Front in Neuroinformatics, 2014; Silver et al, NeuroImage 2012; Szymczak et al, Genetic Epidemiology 2009; …

Multivariate Association studies

Maximizing the joint relationship between genetic variants and brain features

~10^6 SNPs

X =  N individuals

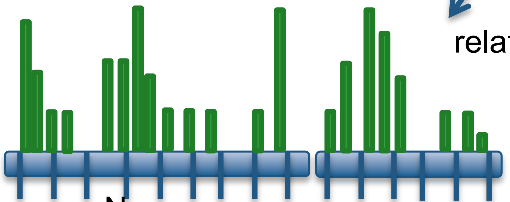
~10^5 brain features

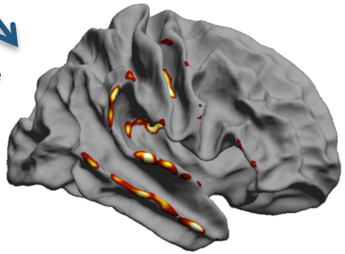
Y =  N individuals

Partial least squares (PLS)

$$\max_{p,q} \mathrm{Cov}(\, X \cdot p,\ Y \cdot q \,)$$

PLS weights
=
relative importance



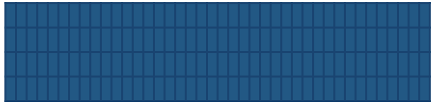


Liu et al, Front in Neuroinformatics, 2014; Silver et al, NeuroImage 2012; Szymczak et al, Genetic Epidemiology 2009; …

Multivariate Association studies

Maximizing the joint relationship between genetic variants and brain features

~10^6 SNPs

X =  N individuals

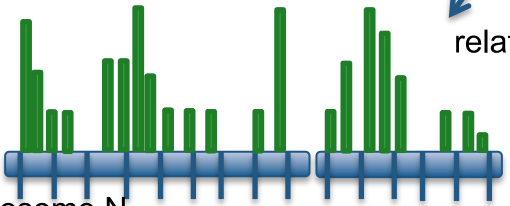
~10^5 brain features

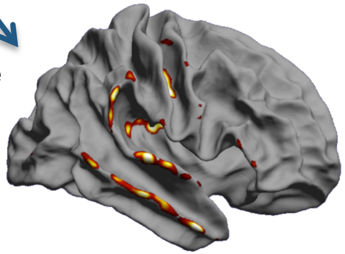
Y =  N individuals

Partial least squares (PLS)

$$\max_{p,q} \mathrm{Cov}(X \cdot p, Y \cdot q)$$

PLS weights = relative importance





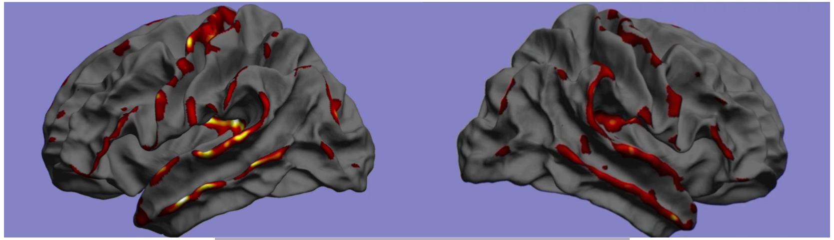
Pros. Overcomes issues of mass univariate analysis
- Avoiding independent **multiple testing**
- Exploring **SNP-SNP interaction** (epistatic effects)

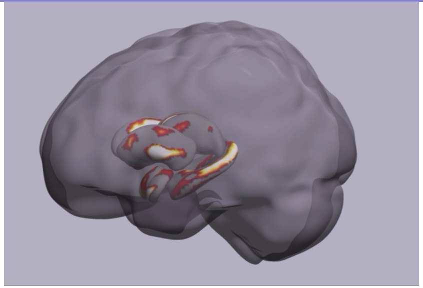
Cons.
- **Overfitting** and reproducibility
- Computational **complexity**

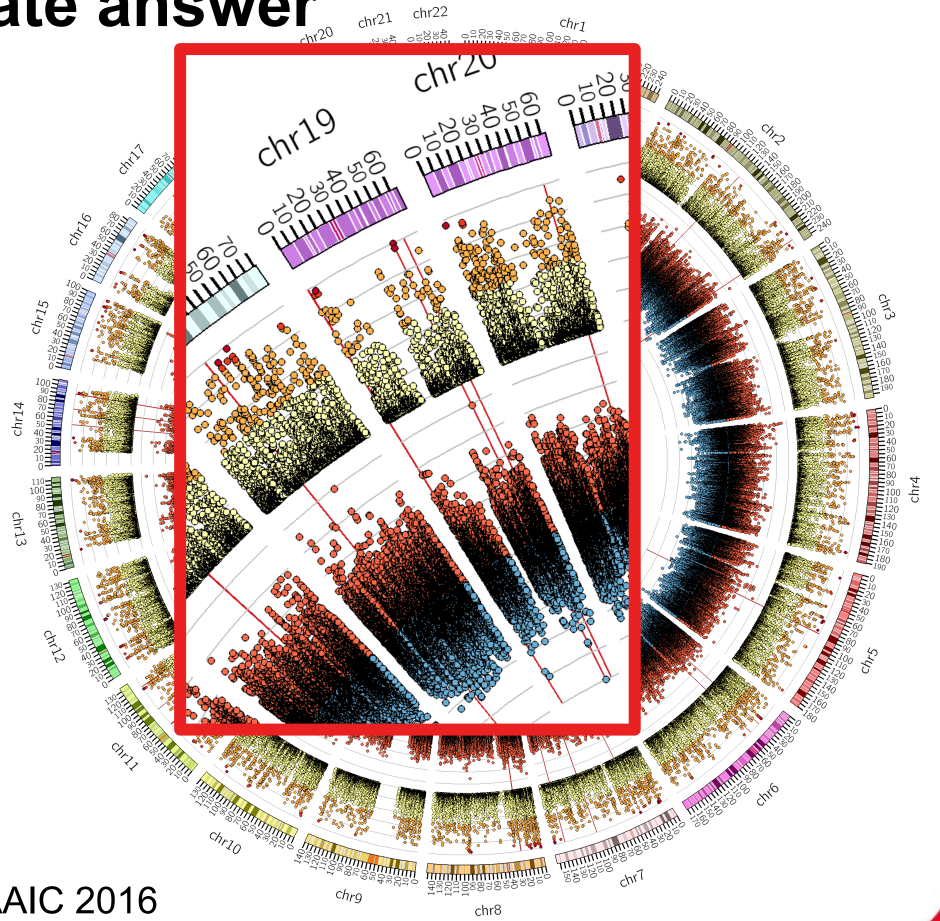
Liu et al, Front in Neuroinformatics, 2014; Silver et al, NeuroImage 2012; Szymczak et al, Genetic Epidemiology 2009;

A. Altmann

A multivariate answer



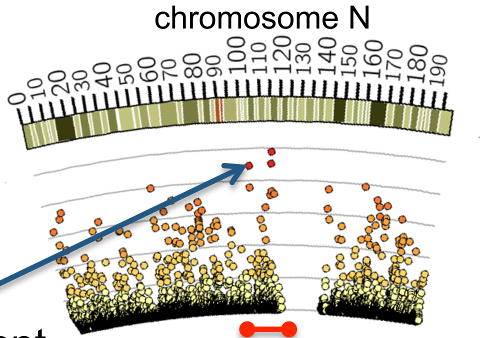




Lorenzi et al. AAIC 2016

Investigating biological mechanisms through Meta-analysis

PLS statistical result



p relevant locus

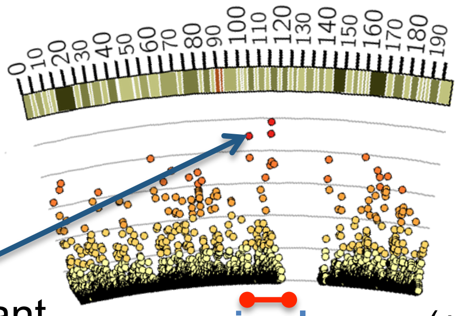
proximal areas (+/- 5kbp)

Analysis of genomic areas

Investigating biological mechanisms through Meta-analysis

PLS statistical result

chromosome N

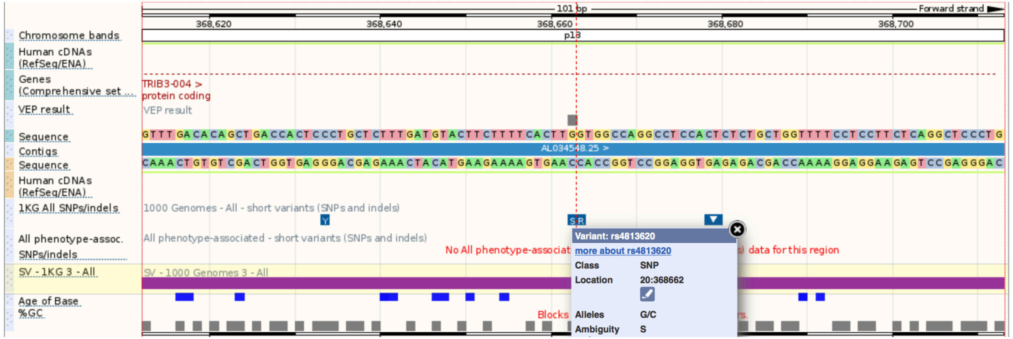


p relevant locus

proximal areas (+/- 5kbp)

Analysis
of genomic areas

Querying gene annotation databases



McLaren et al. The Ensembl Variant Effect Predictor. Genome Biology, 20

Investigating biological mechanisms through Meta-analysis


S. Wray

148 SNP-gene combinations

6 tested tissues
hippocampus, whole blood,
Adipose subcutaneous, artery tibia, nerve tibial,
treated fibroblast

14 Significantly expressed genes

TM2D1 (amyloid-beta binding protein),
IL10RA (increase in hippo in mouse model),
TRIB3
(neuronal cell death, modulates PSEN1 stability, interacts with APP)

	Significance (p-value) training	Significance (p-value) testing
TM2D1	0.005	0.053
IL10RA	0.107	0.620
TRIB3	0.003	0.003
ZBTB7A	0.036	0.913
LYSMD4	0.000	0.206
CRYL1	0.621	0.118
FAM135B	0.000	0.559
IP6K3	0.000	0.465
ITGA1	0.099	0.731
KIN	0.001	0.206
LAMC1	0.002	0.062
LINC00941	0.000	0.690
RBPMS2	0.000	0.215
RP11-181K3.4	0.002	0.053

Joint modeling of brain and genetic data in Alzheimer’s disease

- Ingredients -

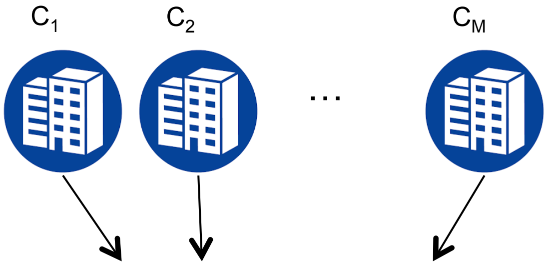
- Data (disease markers)
- Algorithms
- **Databases**

Large multicentric clinical studies

Data for ~100'000 individuals

Challenge: Meta-study

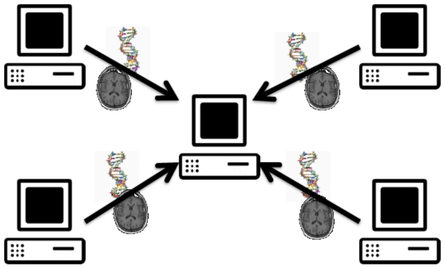
Meta-analysis in genetic studies



State-of-art:

analysis of univariate outcome

(p-value, effect size, standard error, …)



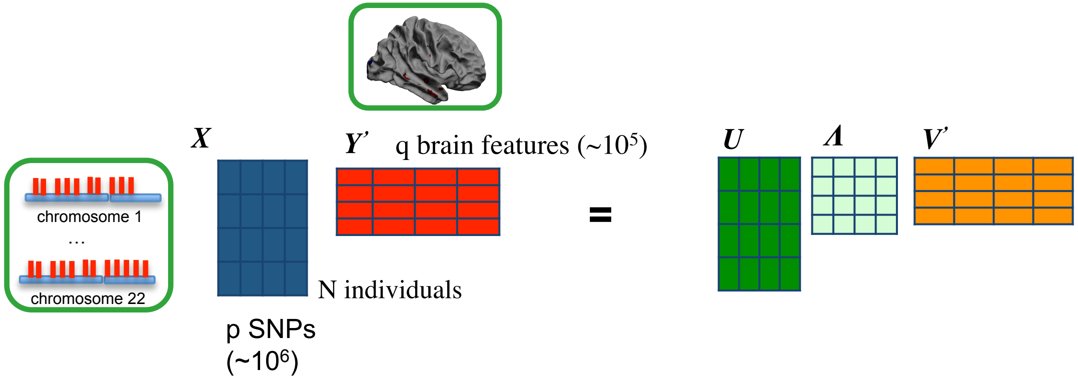
Cons.

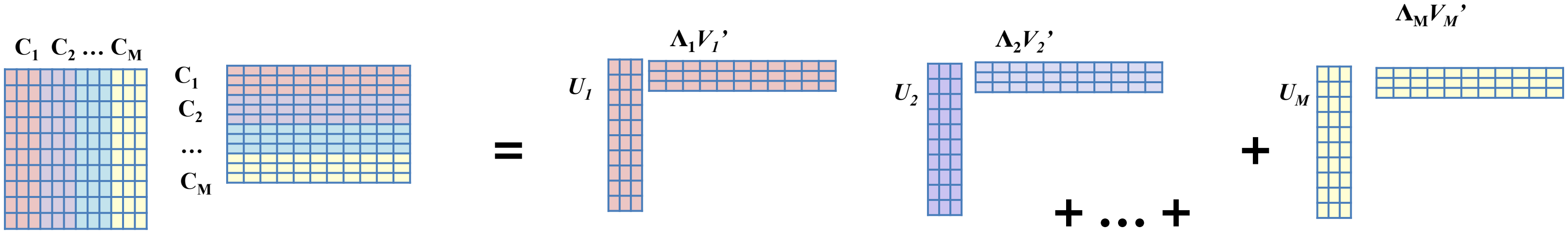
- **Multiple testing** → low statistical power
- No **SNP-SNP interaction**
- Limited interpretability

Problem.

How to develop **multivariate imaging-genetics** modeling approaches within a **meta-analysis** context?

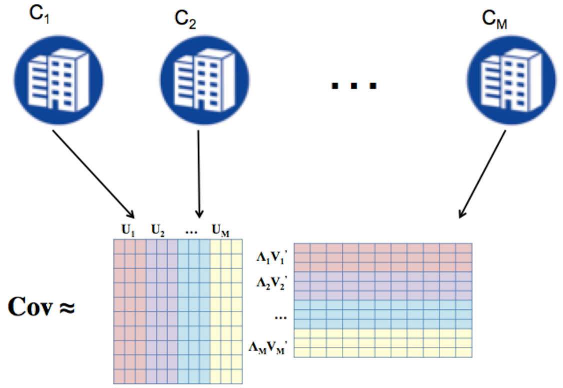
Extending meta-analysis for multivariate models



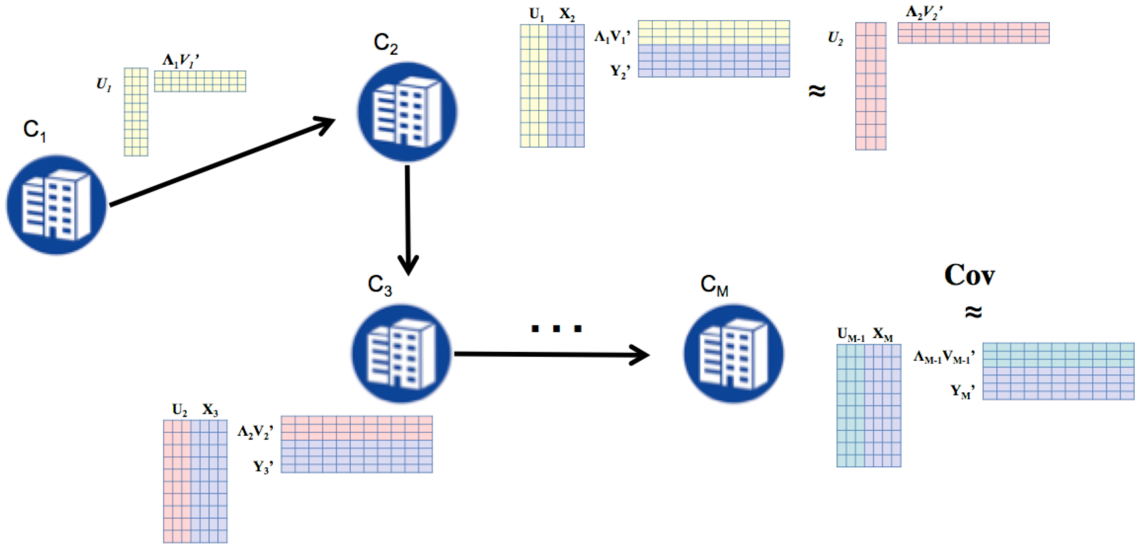


Extending meta-analysis for multivariate models

Meta PLS



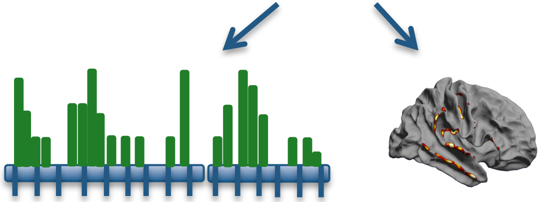
Sequential PLS



Conclusions

Linking **brain atrophy to biological functions** through
Multivariate analysis of **genotype-phenotype relationship**
+
thorough cross-validation for stability assessment

$\max_{p,q} \mathrm{Cov}(\mathbf{X} \cdot \boldsymbol{p}, \mathbf{Y} \cdot \boldsymbol{q})$



Warnings

- Often required to process large datasets with standard hardware
- Need of processing large datasets across different sites

Thank you!