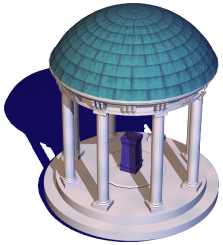
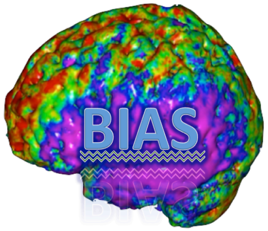
UNC SCHOOL OF PUBLIC HEALTH



Big Data Integration in Biomedical Studies

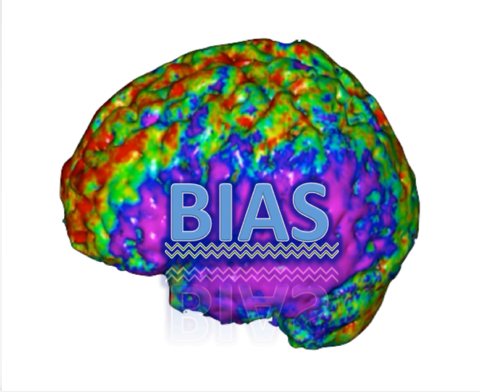
Hongtu Zhu, Ph.D
Department of Biostatistics[†] and Biomedical Research Imaging Center[‡]
The University of North Carolina at Chapel Hill,
Chapel Hill, NC 27599, USA



Outline

- Big Data Integration
- Statistical Challenges in Image Data
- Image-on-Scalar Models
- Image-on-Genetic Association Models
- Predictive Models

Big Data Integration



Big Data

What? *Wikipedia for Big data*

Big data refers data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

Big data is a set of techniques and techologies that require new forms of integration to uncover large hidden values from large daatsets that are diverse, complex, and of a massive scale

Size?

A few dozen terabytes to many petabytes of data.

Characteristics?

Volume, Variety, Velocity, Variability, Veracity, Complexity, ….

Big Data or Pig Data

Why?

Answer questions of personal or scientific interest.

What matters?

Ensuring accurate and appropriate data collection.
Correct variables, Collection methods (techniques and sampling),

Quality assurance and Quality control

Does it work?

Big data does not work in most cases, since we do not know
(i) which variables (information at which scale) are critical;
(ii) whether we have capability to collect such information.

Big Data Integration

Big data integration is to integrate multiple sources of data to improve knowledge discovery.

Data Sources Discovery:

Data Exploration (e.g., meta analysis):

(i) the use of prior knowledge,- and its efficient storage;

(ii) the development of statistical methods to analyze heterogeneous data sets;

(iii) the creation of data explorative tools that incorporate both useful summary statistics and new visualization tools.

Human Genome Project

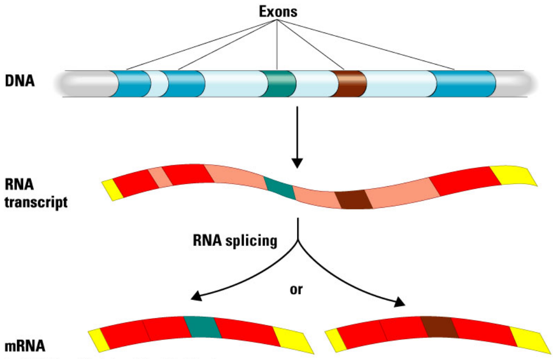
The HGP aims to determine the sequence of chemical base pairs which make up human DNA and identify and map all of the genes of the human genome.

1000 Genomes Project

Encyclopedia of DNA Elements Project (ENCODE)

The Cancer Genome Atlas Project (TGCA) is to generate insights into the heterogeneity of different cancer subtypes by creating a map of molecular alternations for every type of cancer at multiple levels.

Immunological Genome Project (ImmGen)



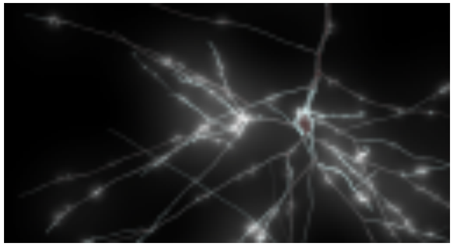
Copyright © 2004 Pearson Education, Inc., publishing as Benjamin Cummings.

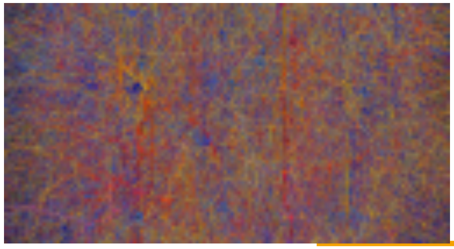
HBP and BRAIN

Human Brain Project

aims to simulate the complete human brain on Supercomputers to better understand how it functions.

The Brain Research through
Advancing Innovative Neurotechnologies or BRAIN,
aims to reconstruct the activity of every single neuron as they fire simultaneously in different brain circuits, or perhaps even whole brains.





Big Neuroimaging Data

NIH normal brain development

1000 Functional Connectome Project

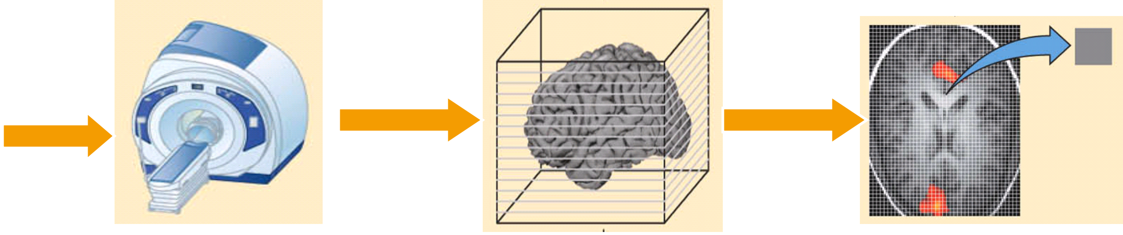
Alzheimer’s Disease Neuroimaging Initiative

National Database for Autism Research (NDAR)

Human Connectome Project

Philadelphia Neurodevelopmental Cohort

Genome superstruct Project



www.guysandstthomas.nhs.uk/.../T/Twins400.jpg

Big Data to Knowledge (BD2K)

NIH Big Data to Knowledge (BD2K)

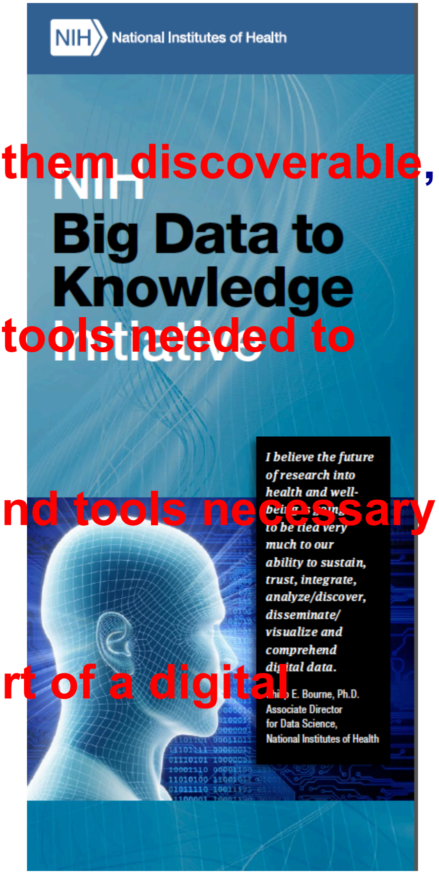
The four aims of BD2K are

To facilitate broad use of biomedical digital assets by **making them discoverable**, accessible, and citable

To conduct research and develop the methods, software, and **tools needed to** analyze biomedical data.

To enhance training in the development and use of methods **and tools necessary** for biomedical Big Data science

To support a data ecosystem that accelerates discovery **as part of a digital enterprise.**



NIH National Institutes of Health

NIH Big Data to Knowledge Initiative

I believe the future of research into health and well-being is going to be tied very much to our ability to sustain, trust, integrate, analyze/discover, disseminate/ visualize and comprehend digital data.

Philip E. Bourne, Ph.D.
Associate Director for Data Science,
National Institutes of Health

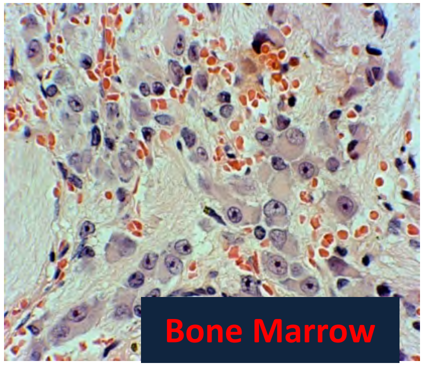
Precision Medicine

Precision medicine **(PM) is a medical model that proposes the customization of healthcare—with medical decisions, practices, and/or products being tailored to the individual patient.**

Precision Medicine refers to the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those diseases they may develop, or in their response to a specific treatment.

PM (wiki)

Cover Art: Nicolle Rager Fuller, Sayo-Art LLC
Photo: © Graham Bell/Corbis



Dream Challenges

http://dreamchallenges.org

Alzheimer's Disease Big Data DREAM Challenge

SANOFI
BrightFocus Foundation
Cure in Mind. Cure in Sight
Ray and Dagmar Dolby Family Fund
frontiers in NEUROSCIENCE
nature neuroscience
The AddNeuroMed Study
EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH
RUSH UNIVERSITY MEDICAL CENTER
Takeda
ADNI
Alzheimer's Disease Neuroimaging Initiative
Alzheimer's Research UK
Defeating Dementia
ROSENBERG ALZHEIMER'S PROJECT
Pfizer

Prostate Cancer DREAM Challenge

DREAM CHALLENGES
powered by Sage Bionetworks
Project Data Sphere
SANOFI
Celgene
AstraZeneca
Memorial Sloan Kettering Cancer Center
1884
PROSTATE CANCER FOUNDATION
SCHOOL OF MEDICINE
Department of Pharmacology
UNIVERSITY OF COLORADO ANSCHUTZ MEDICAL CAMPUS
Sage BIONETWORKS
UNC LINEBERGER
UT SOUTHWESTERN MEDICAL CENTER
COVANCE
SOLUTIONS MADE REAL
DANA-FARBER CANCER INSTITUTE
UCSF Helen Diller Family Comprehensive Cancer Center
IBM Research
Tulane University

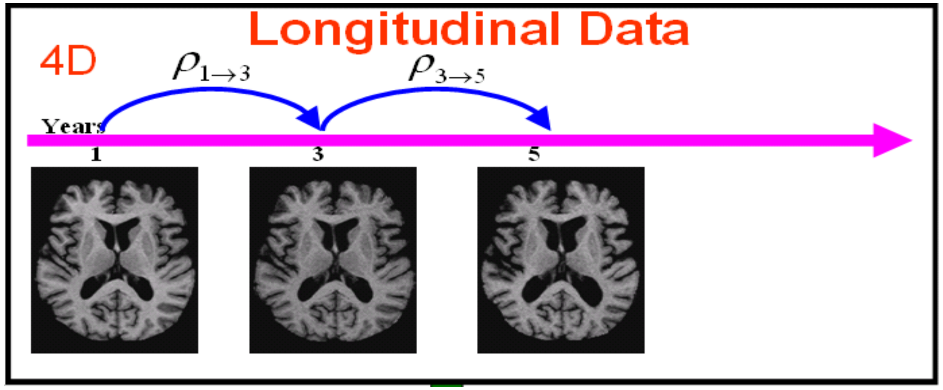
Study Design

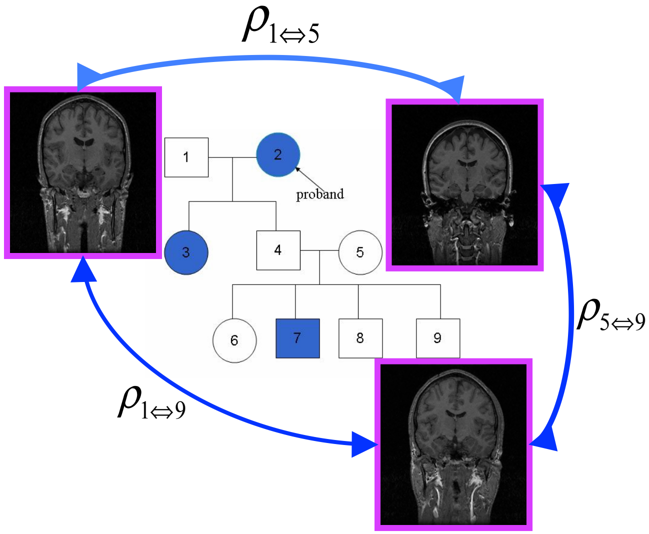
Scientific Questions

Design: cross-sectional studies;

clustered studies including

longitudinal and twin/familial studies;




$\rho_{1\Leftrightarrow5}$
1
2
proband
3
4
5
6
7
8
9
$\rho_{5\Leftrightarrow9}$
$\rho_{1\Leftrightarrow9}$

Imaging Data

Structural MRI

Functional MRI (task)

- Variety of acquisitions
- Measurement basics
- Limitations & artefacts
- Analysis principles
- Acquisition tips

Diffusion MRI

Functional MRI (resting)

PET

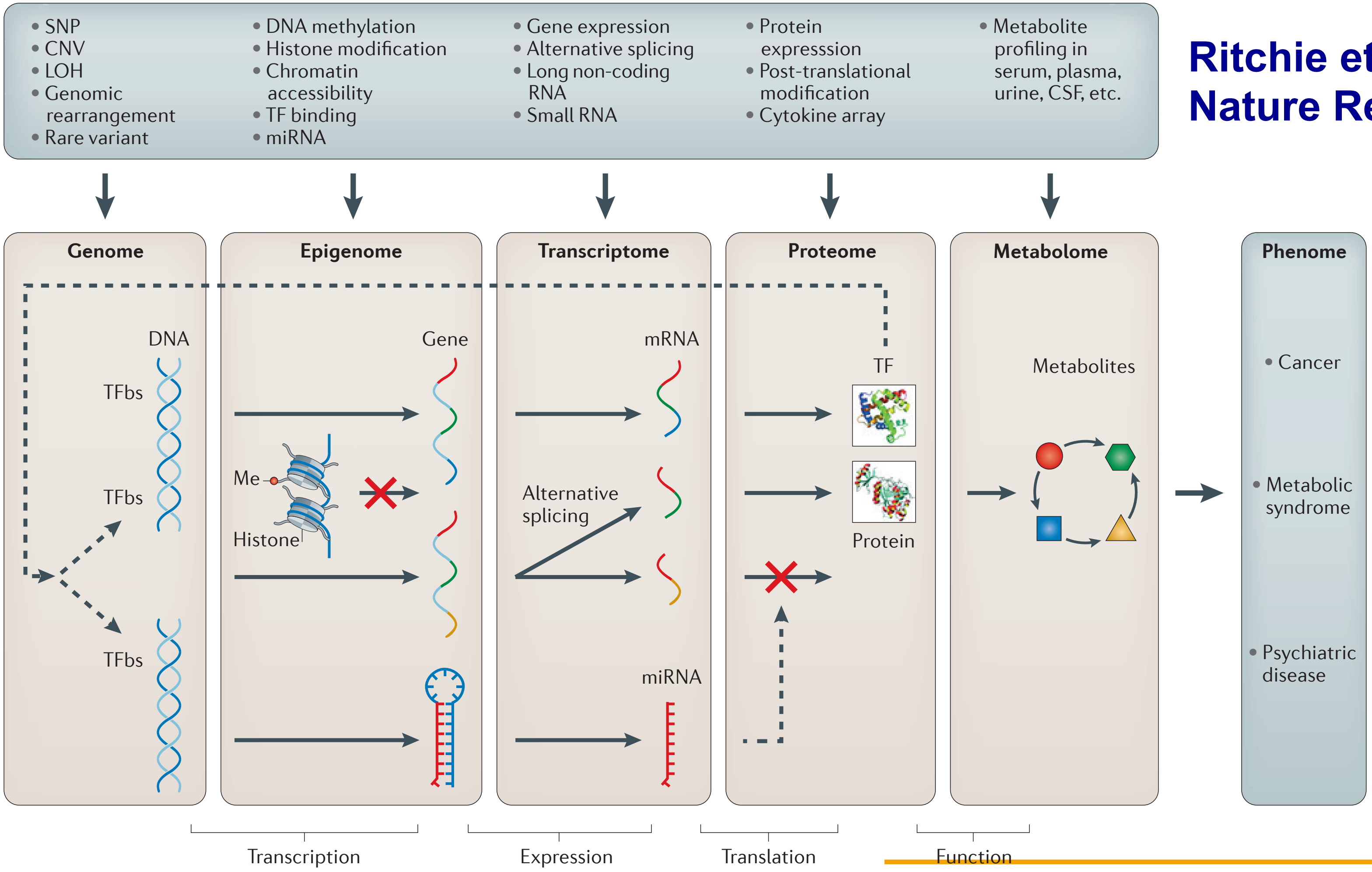
EEG/MEG

CT

Calcium

Multi-Omic Data

Ritchie et al. (2015). Nature Review Genetics



- SNP
- CNV
- LOH
- Genomic rearrangement
- Rare variant

- DNA methylation
- Histone modification
- Chromatin accessibility
- TF binding
- miRNA

- Gene expression
- Alternative splicing
- Long non-coding RNA
- Small RNA

- Protein expresssion
- Post-translational modification
- Cytokine array

- Metabolite profiling in serum, plasma, urine, CSF, etc.

Clinical Data and Acquisition

Clinical Data: a variety of clinical sources to present a unified view of a single patient.

clinical laboratory test results, patient demographics, pharmacy information, hospital admission, discharge and transfer date, progress report, etc.

Clinical Acquisition:

- Paper or electronic medical records
- Paper forms completed at a site
- Interactive voice response systems
- Local electronic data capture systems
- Central web based systems

Data Exploration

Data Analysis

- Single Level Data Analysis for imaging or omics data, e.g., denoise, segmentation, cluster, network,

- Multi-level Data Analysis for across imaging or omics data

- Data Integration Analysis for imaging, clinical, and omics data.

  Multi-staged analysis
  Meta-dimensional analysis
  Mediation/moderation analysis

Software/Computing Language/

Apache Spark

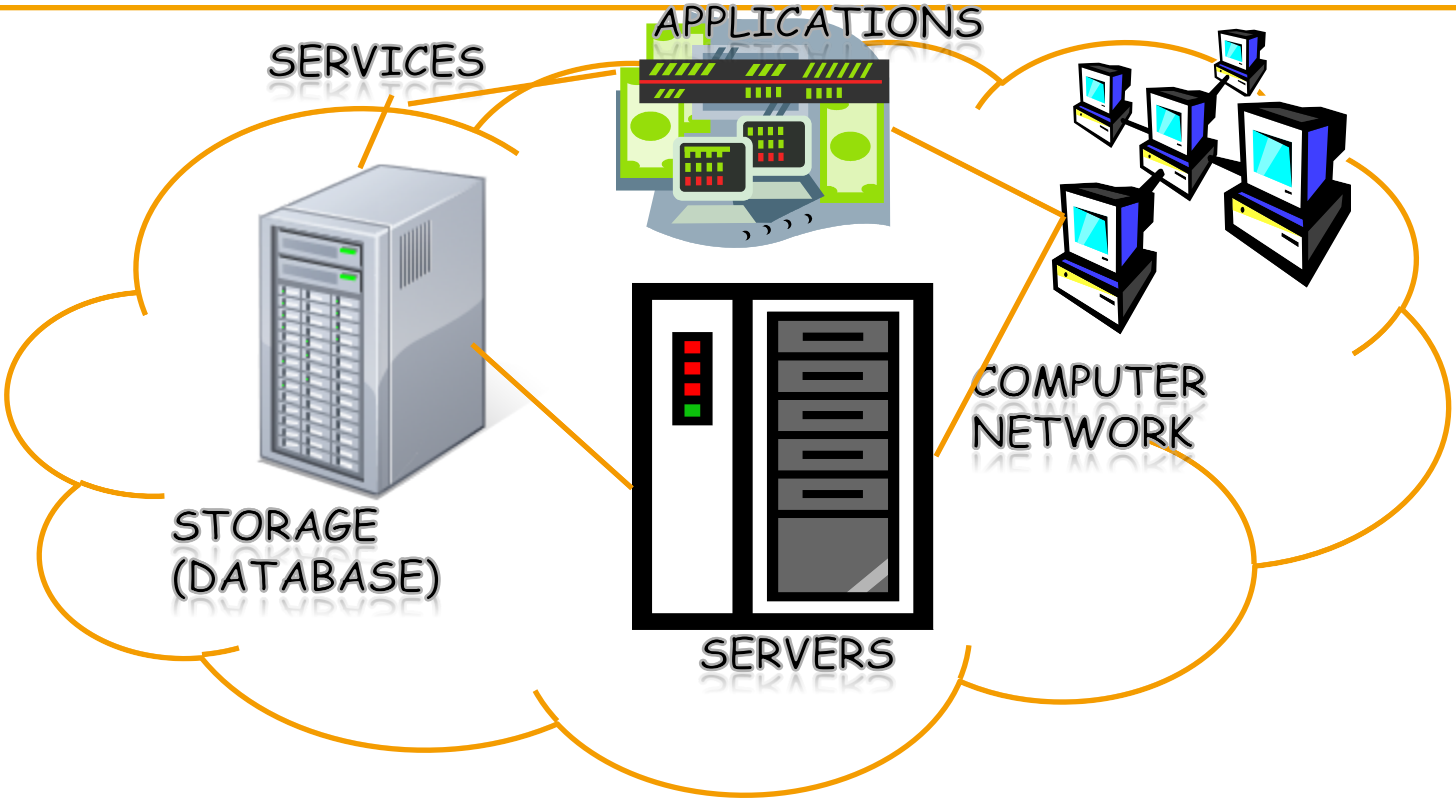
Data growing faster than processing speeds

Only solution is to parallelize on large clusters

» Wide use in both enterprises and web industry

How do we program these things?

Cloud Computing

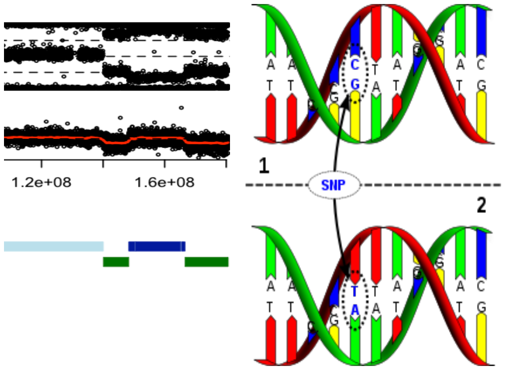


- Shared pool of configurable computing resources
- On-demand network access
- Provisioned by the Service Provider

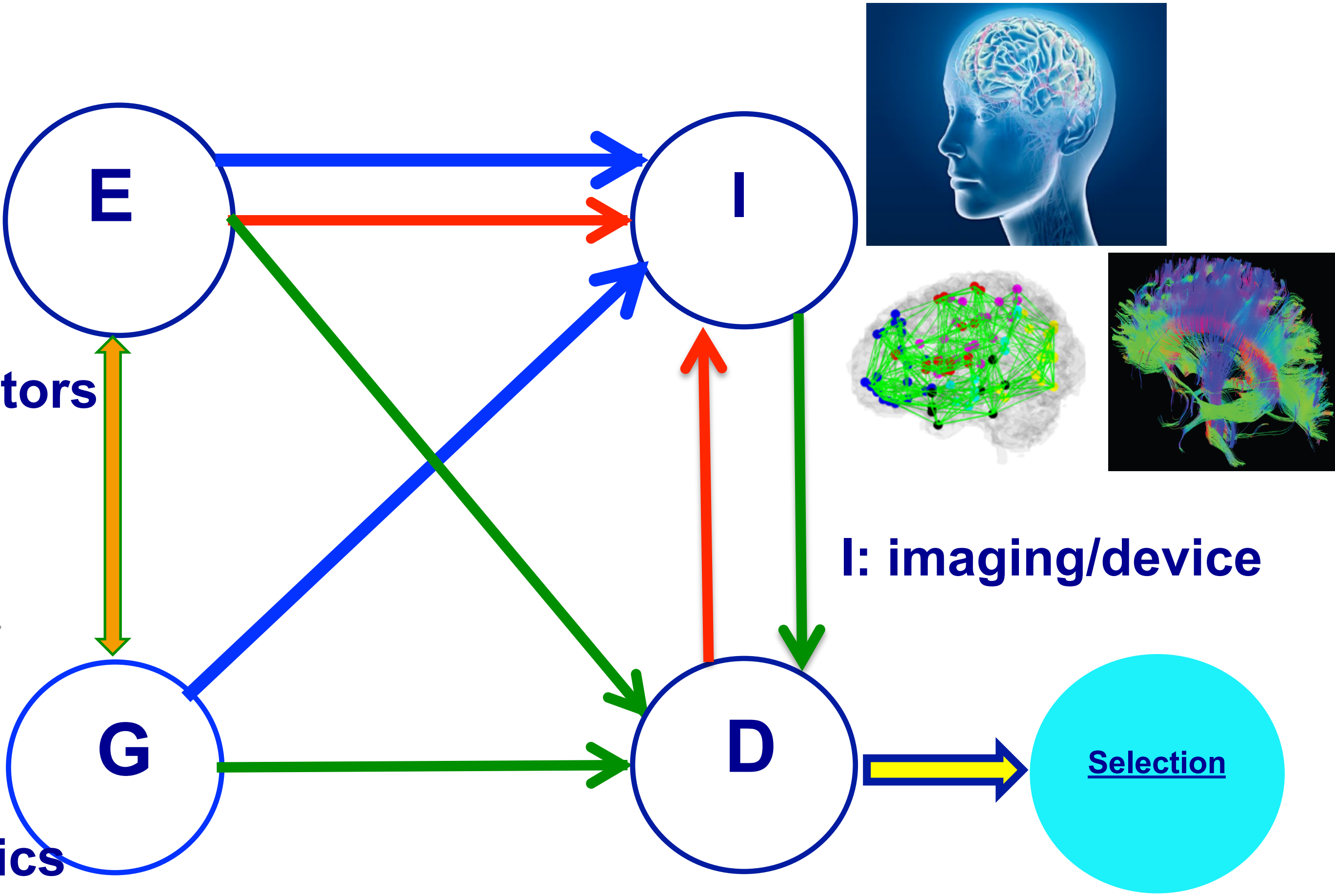
Adopted from: Effectively and Securely Using the Cloud Computing Paradigm by peter Mell, Tim Grance

Big Data Integration

E: environmental factors



G: genetic/genomics

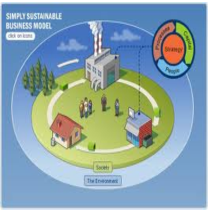


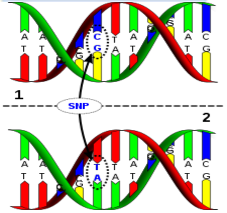
I: imaging/device

D: disease

http://en.wikipedia.org/wiki/DNA_sequence

Big Data Integration





Medical Informatics & Management

Disease

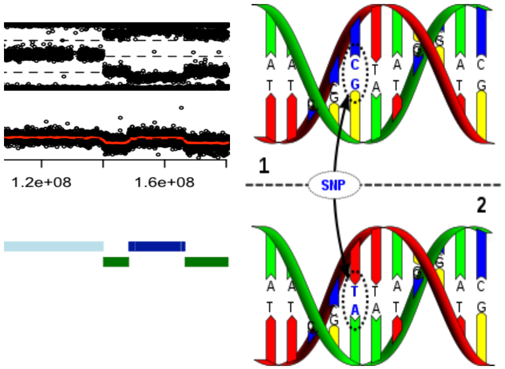
Etiology
Prevention
Treatment

Medical Industry

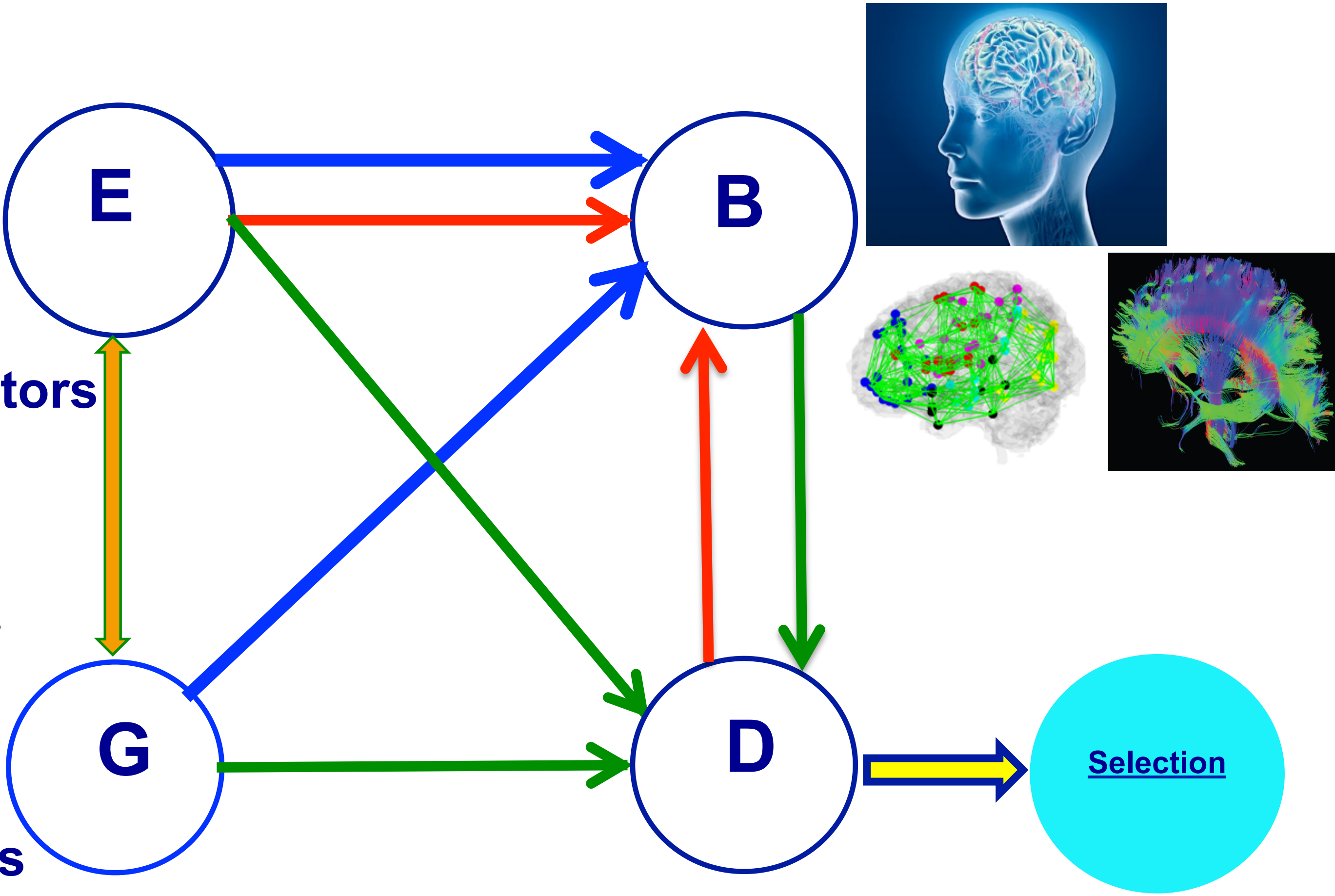
Care
Policy
System
Science
Insurance
Economics
Pharmaceutical

Big Data Integration

E: environmental factors



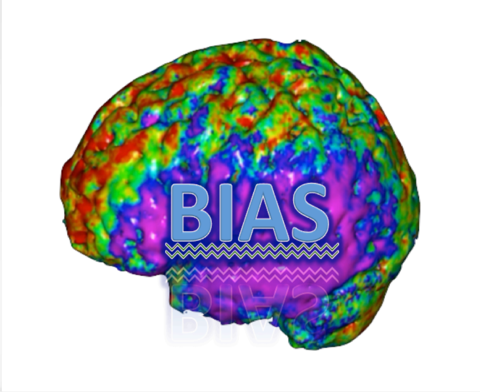
G: genetic markers



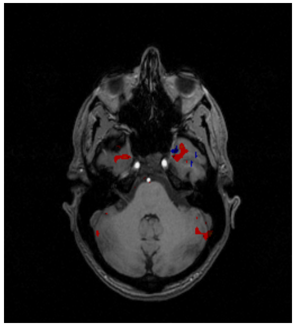
D: disease

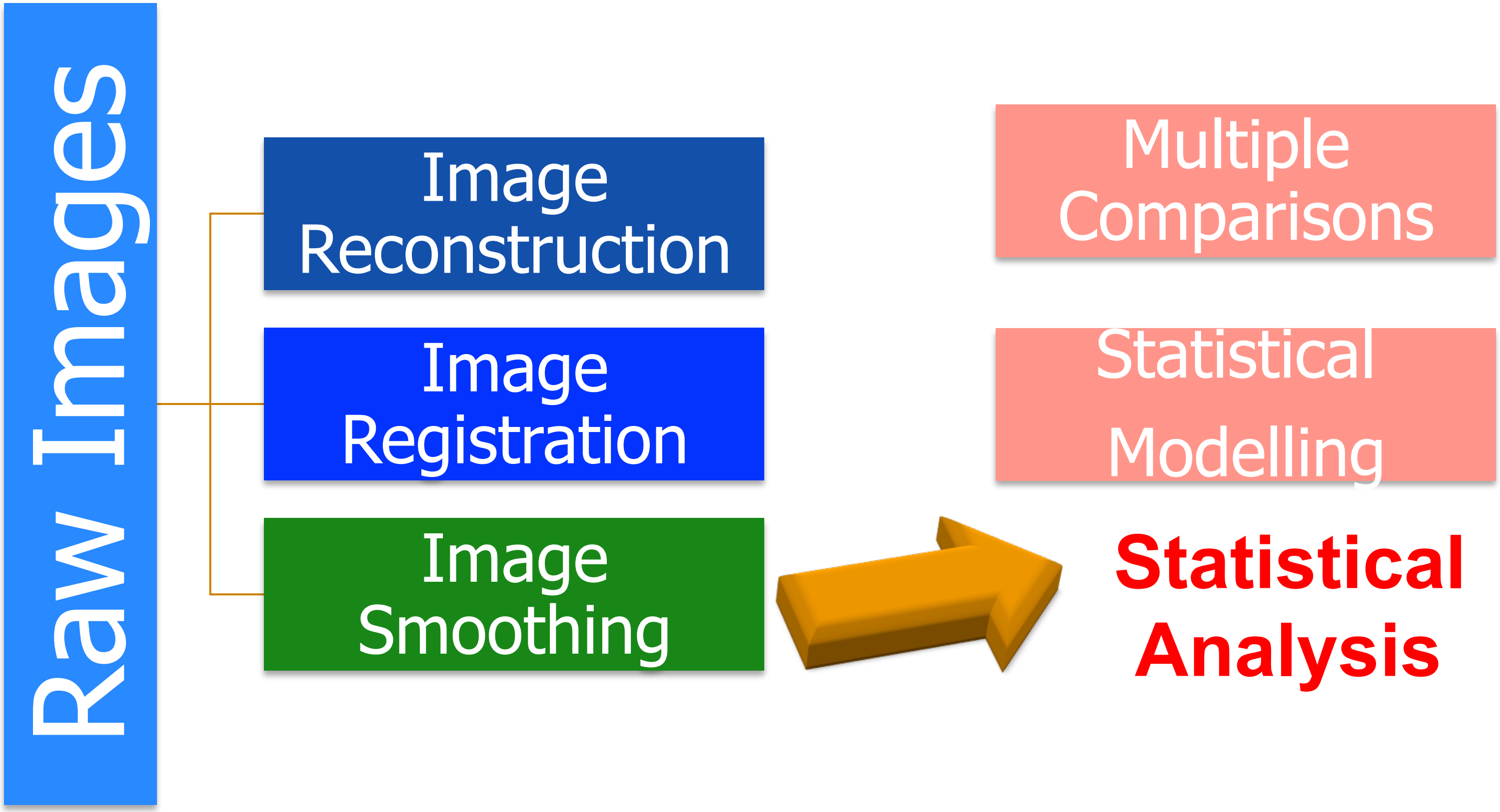
http://en.wikipedia.org/wiki/DNA_sequence

Statistical Challenges in Imaging Data



Imaging and Statistical Analysis





Raw Images

Image Reconstruction

Image Registration

Image Smoothing

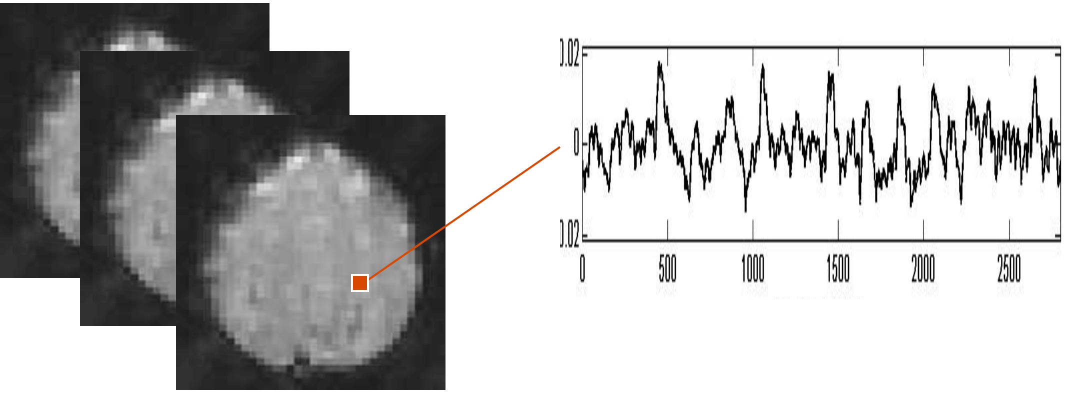
Multiple Comparisons

Statistical Modelling

Statistical Analysis

Individual Imaging Analysis

Imaging Construction



Multimodal Analysis

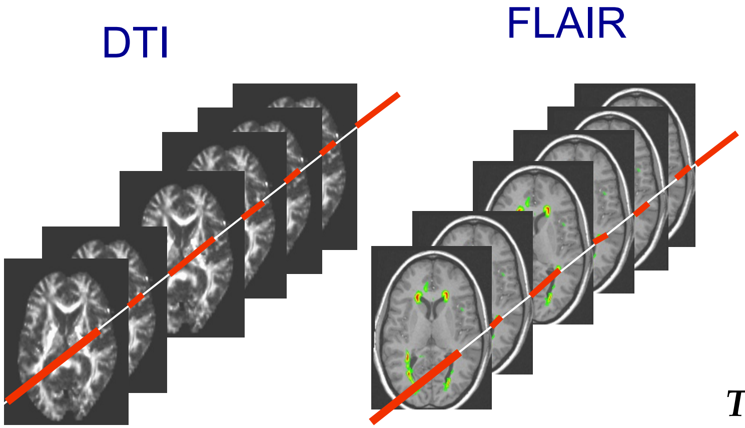
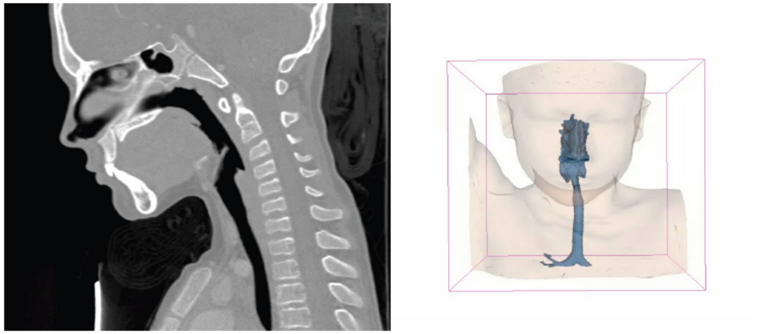
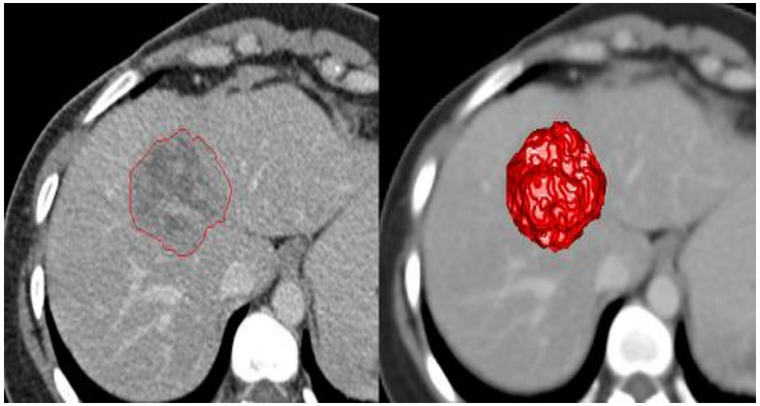


Image Segmentation

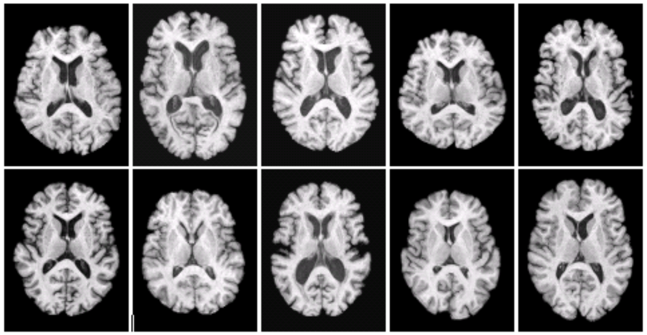
Example: Airway Segmentation from CT

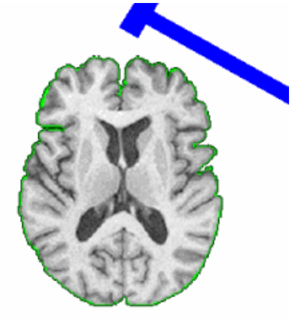




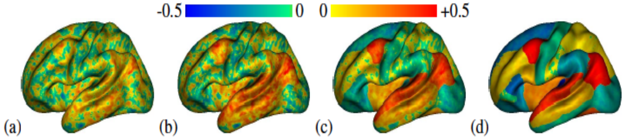
Group Imaging Analysis

Registration



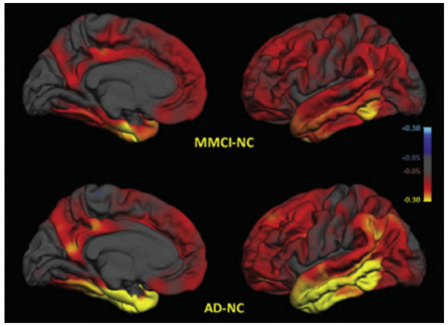


Prediction

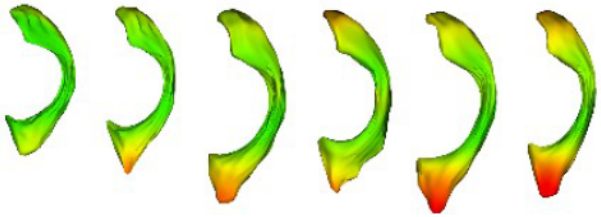


NC/Diseased

Group Differences

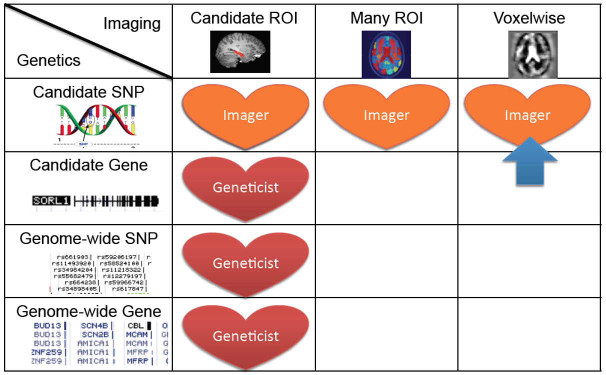


Longitudinal/Family Brain



Hibar, Dinggang, Martin

Imaging Genetics

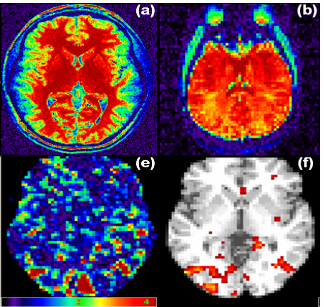


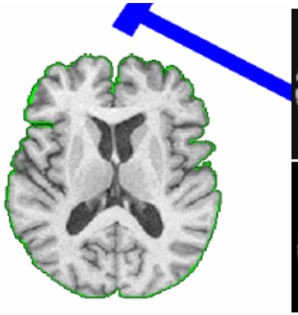
Imaging / Genetics	Candidate ROI	Many ROI	Voxelwise
Candidate SNP	Imager	Imager	Imager
Candidate Gene	Geneticist		
Genome-wide SNP	Geneticist		
Genome-wide Gene	Geneticist		

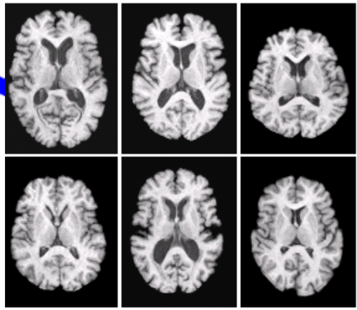
Noisy Imaging Data

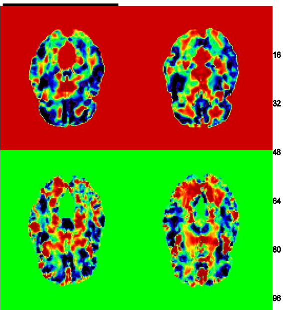
Key Features

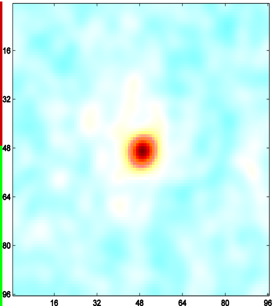
- Infinite Dimension
- Spatial Smoothness
- Spatial Correlation
- Spatial Heterogeneity

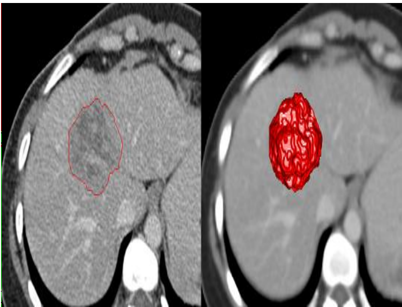












`Noisy' Spatial Maps

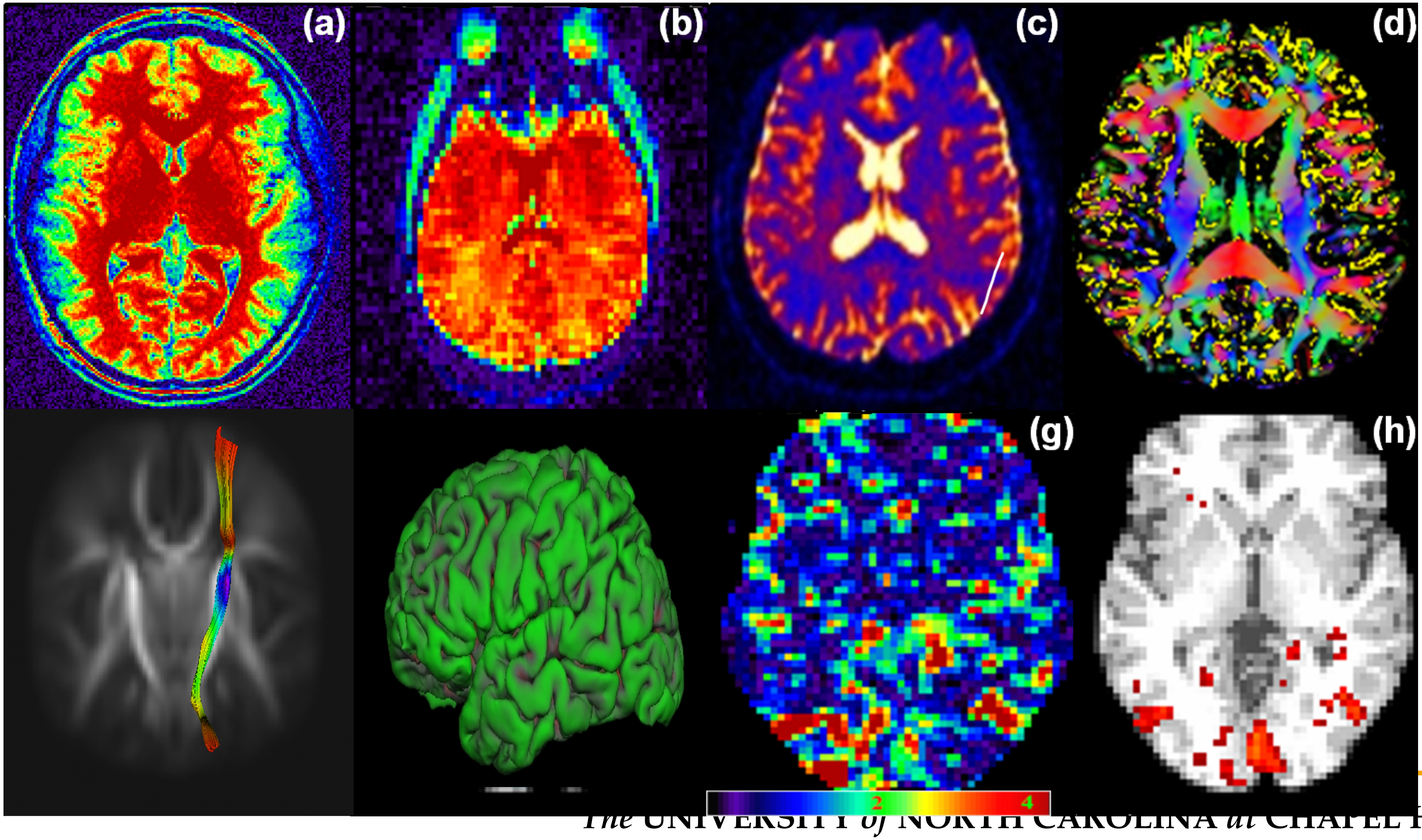
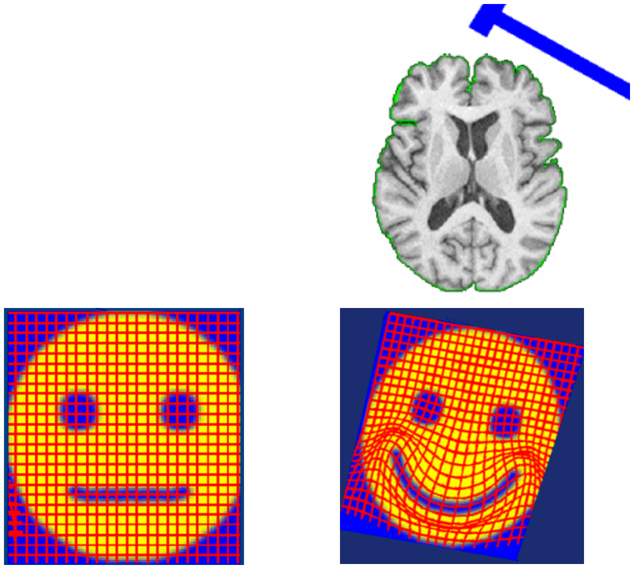


Image Registration

Image registration is the process of transforming different sets of data into ***one coordinate system***. Given a reference image R and a template image T, find a reasonable transformation Y, such that the transformed image T[Y] is similar to R.

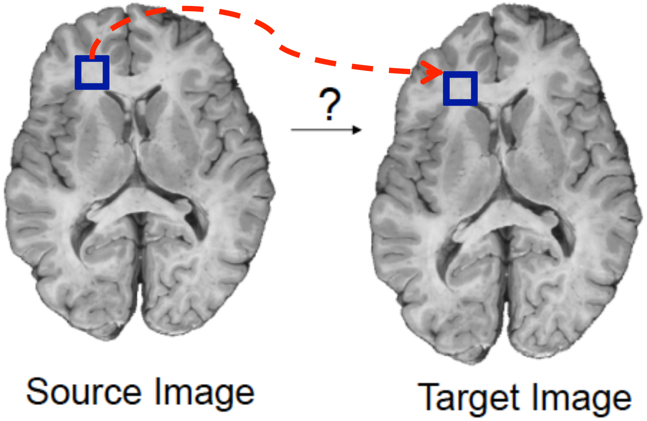


Dinggang

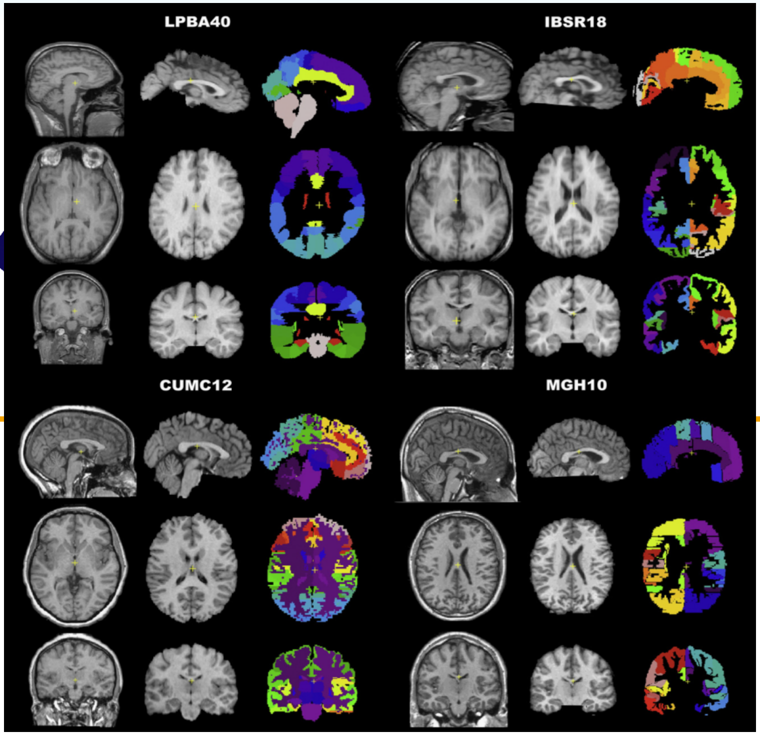
Establishing a geometric transformation

$$\underline{x}' = \underline{h}(\underline{x}) = \underline{x}' = \underline{x} + \underline{\Delta x}$$

relating points in one image to points in another.



Registration Errors



Brain image dataset with manually labeled ROIs

Method	LPBA40	IBSR18	CUMC12	MGH10
FLIRT	59.29±11.94	39.71±13.00	39.63±11.51	46.24±14.03
AIR	65.23±10.72	41.41±13.35	42.52±11.90	47.99±14.10
ANIMAL	66.20±10.17	46.31±13.51	42.78±11.95	50.40±15.21
ART	71.85±9.59	51.54±14.42	50.54±12.16	56.10±15.33
D. Demons	68.93±9.23	46.83±13.37	46.45±11.46	52.28±14.94
FNIRT	70.07±9.80	47.63±14.15	46.53±12.26	49.54±14.58
IRTK	70.02±10.26	52.09±14.97	51.75±12.45	54.90±15.70
JRD-fuild	70.02±9.83	48.95±13.87	46.37±12.06	52.33±14.81
ROMEO	68.49±10.12	46.48±13.91	44.49±13.04	51.23±14.55
SICLE	60.41±16.21	44.53±13.03	42.08±12.19	48.36±14.31
SyN	71.46±10.86	52.81±14.85	51.63±12.60	56.83±15.81
SPM_N[1]	66.97±10.14	42.10±13.25	36.70±12.43	49.77±14.54
SPM_N[2]	57.13±14.95	37.18±14.11	42.93±11.75	43.16±15.88
SPM_US[3]	68.62±9.00	45.29±12.60	44.81±11.35	49.61±14.08
SPM_D[4]	67.15±18.34	54.02±14.70	51.98±13.91	54.31±16.05
S-HAMMER	**72.48±8.46**	**55.47±11.27**	**53.74±9.82**	**58.20±15.03**

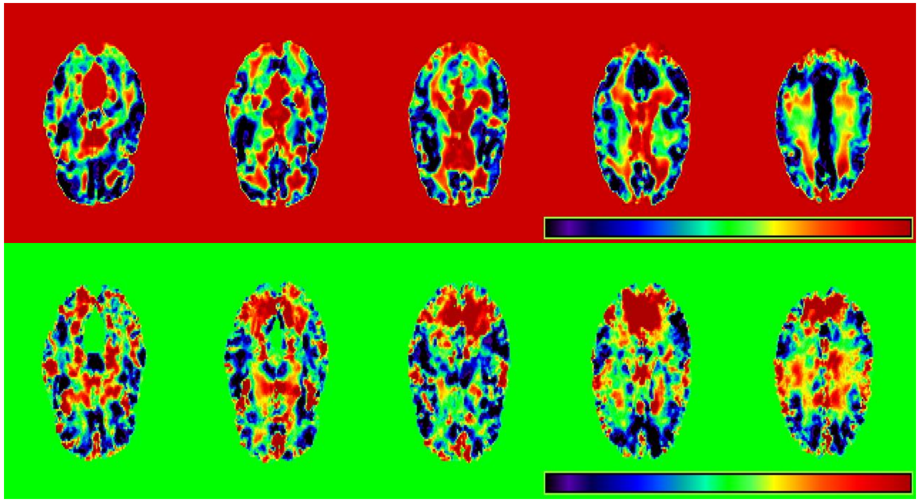
[1] SPM 5 (“SPM2-type” Normalization)
[2] SPM 5 (Normalization) [3] SPM 5 (Unified Segmentation) [4] SPM 5 (DARTEL Toolbox)

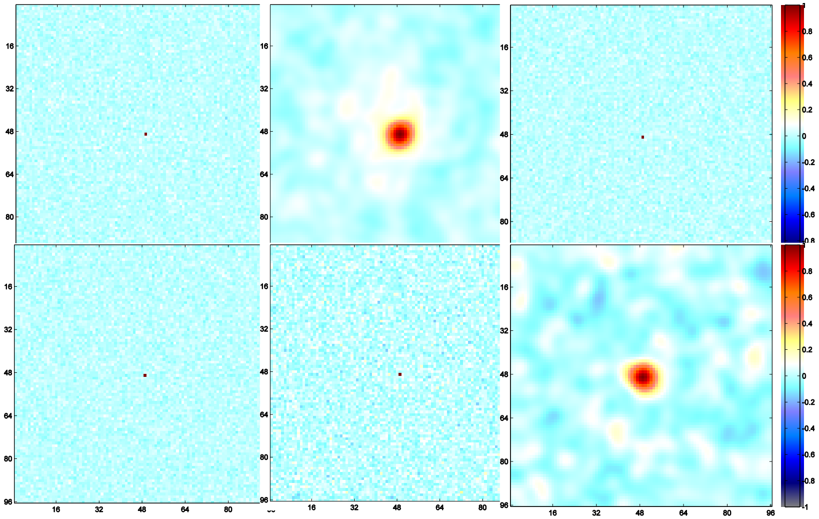
[1] Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J., Parsey, R.V., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46, 786-802.
[2] Wu, G., Kim, M., Wang, Q., Shen, D.: Hierarchical Attribute-Guided Symmetric Diffeomorphic Registration for MR Brain Images. MICCAI 2012, Nice, France (2012)

Noisy Spatial Correlation

Long-range Correlation

Short-range Correlation





“Unmodeled effects”

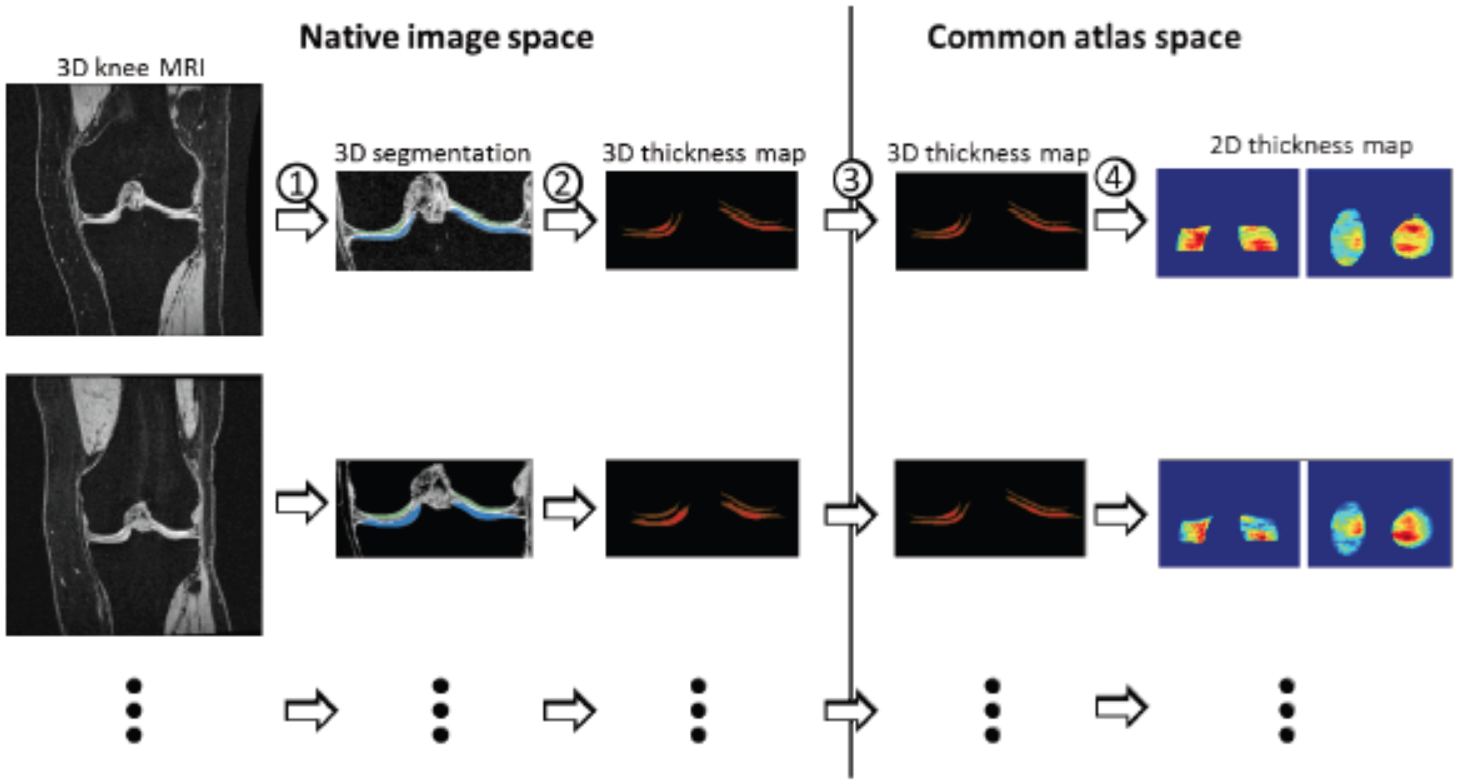
“Signal Processing”

Daniel

Noisy Spatial Heterogeneity

Osteoarthritis (OA)

Cartilage Loss



Marc

Complex Data Structure

Multivariate Imaging Measures

Smooth Functional Imaging Measures

Whole-brain Imaging Measures

4D-Time Series Imaging Measures

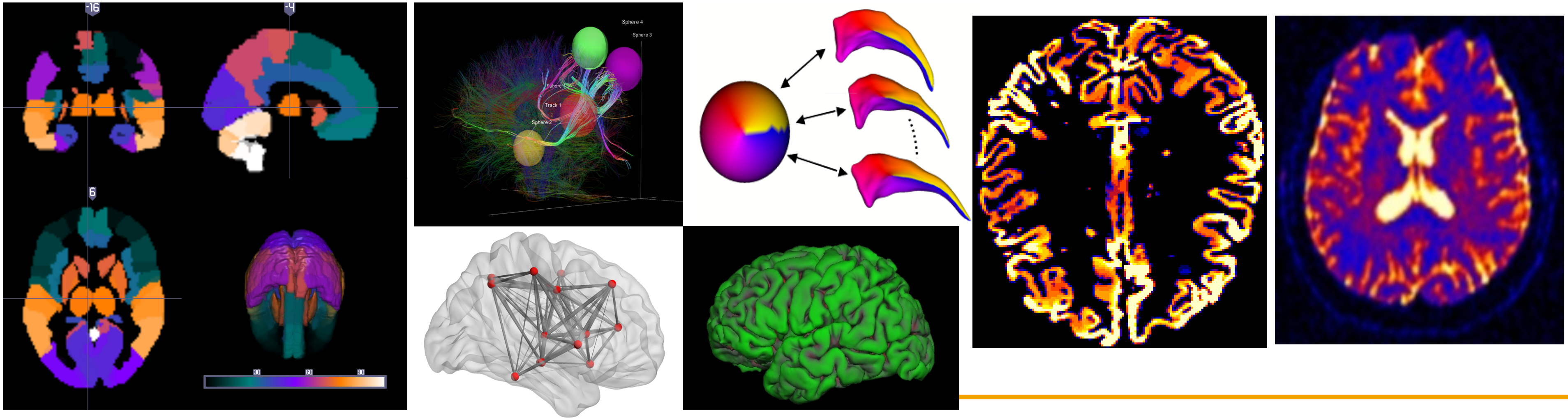
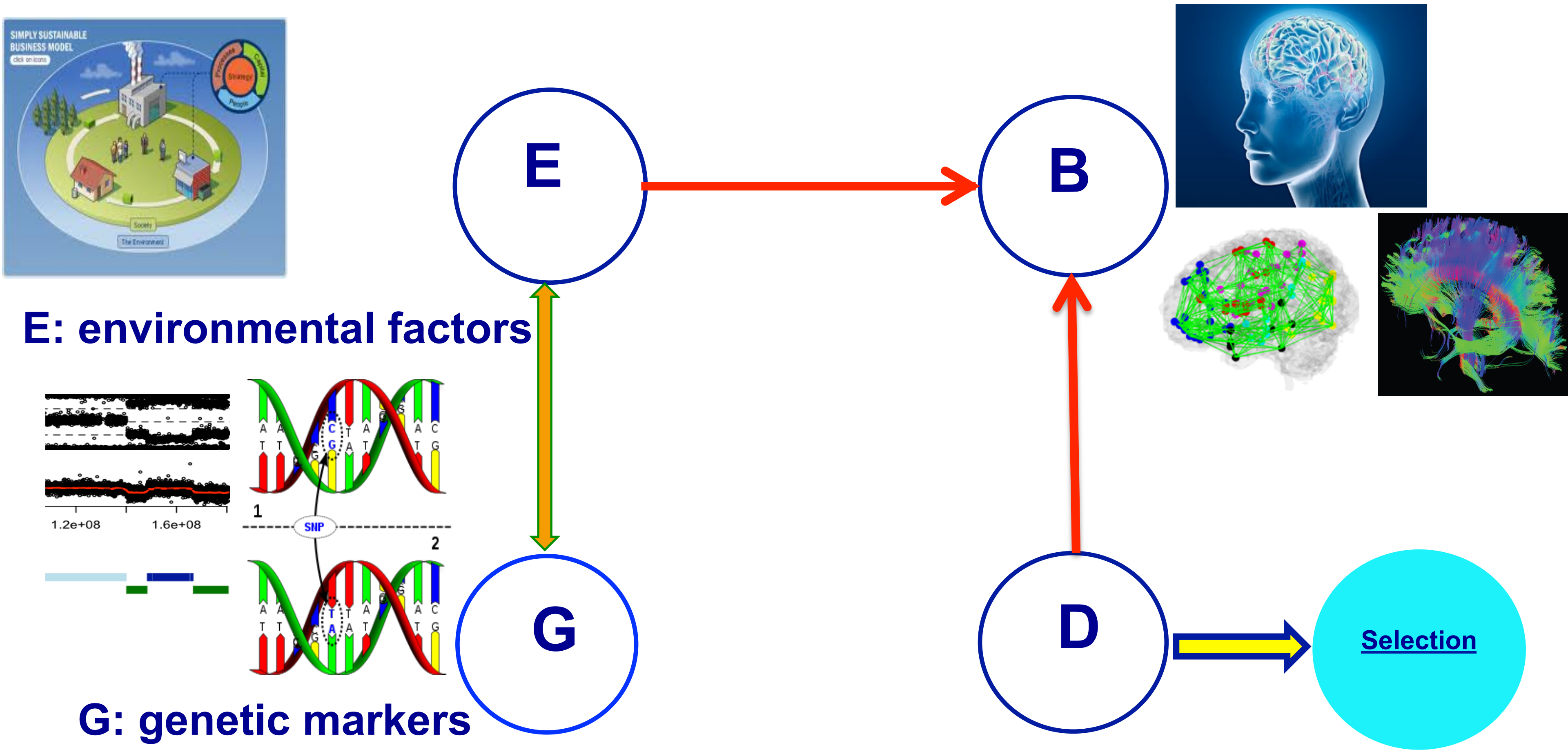


Image-on-Scalar Models

Big Data Integration



E: environmental factors

G: genetic markers

D: disease

http://en.wikipedia.org/wiki/DNA_sequence

Reading Materials

1. Zhu, H. T., Chen, K. H., Yuan, Y. and Wang, J. L. (2015). Functional Mixed Processes Models for Repeated Functional Data. In submission.
2. Luo, X. C., Zhu, L. X., Kong, L., *Zhu, H.T.* Functional Nonlinear Mixed Effects Models For Longitudinal Image Data. *Information Processing in Medical Imaging (IPMI)* 2015.
3. Liang, J. L., Huang, C., and Zhu, H.T. (2014). Functional single-index varying coefficient models. In submission.
4. Zhu, HT., Fan, J., and Kong, L. (2014). Spatial varying coefficient model and its applications in neuroimaging data with jump discontinuity. *JASA,* 109, 977-990, 2014*.*
5. J. W. Hyun, Li, Y. M., Gilmore, J., Lu, Z.H., Styner, M., and *Zhu, H.T.* SGPP: Spatial Gaussian Predictive Process Models for Neuroimaging Data. *NeuroImage*, 89, 70–80, 2014.
6. Yuan, Y., Gilmore, J., Geng, X. J., Styner, M., Chen, K. H., Wang, J. L., and *Zhu, H.T.* (2014). Fmem: Functional mixed effects modeling for the analysis of longitudinal white matter tract data. *NeuroImage* 84, 753–764.
7. Yuan, Y., Gilmore, J., Geng, X. J., Styner, M., Chen, K. H., Wang, J. L., and *Zhu, H.T.* (2013). A longitudinal functional analysis framework for analysis of white matter tract statistics. *NeuroImage*, 23:220-31, 2013.
8. Yuan, Y., Zhu, H.T., Styner, M., J. H. Gilmore., and Marron, J. S. (2013). Varying coefficient model for modeling diffusion tensors along white matter bundles. *Annals of Applied Statistics.* 7(1):102-125..
9. Zhu, H.T., Li, R. Z., Kong, L.L. (2012). Multivariate varying coefficient models for functional responses. *Ann. Stat.* 40, 2634-2666.
10. Hua, Z.W., Dunson, D., Gilmore, J.H., Styner, M., and *Zhu, HT*. (2012). Semiparametric Bayesian local functional models for diffusion tensor tract statistics. *NeuroImage,* 63, 460-674.
11. Zhu, HT., Kong, L., Li, R., Styner, M., Gerig, G., Lin, W. and Gilmore, J. H. (2011). FADTTS: Functional Analysis of Diffusion Tensor Tract Statistics, *NeuroImage*, 56, 1412-1425.
12. Zhu, H.T., Styner, M., Tang, N.S., Liu, Z.X., Lin, W.L., Gilmore, J.H. (2010). FRATS: functional regression analysis of DTI tract statistics. *IEEE Transactions on Medical Imaging*, 29, 1039-1049.

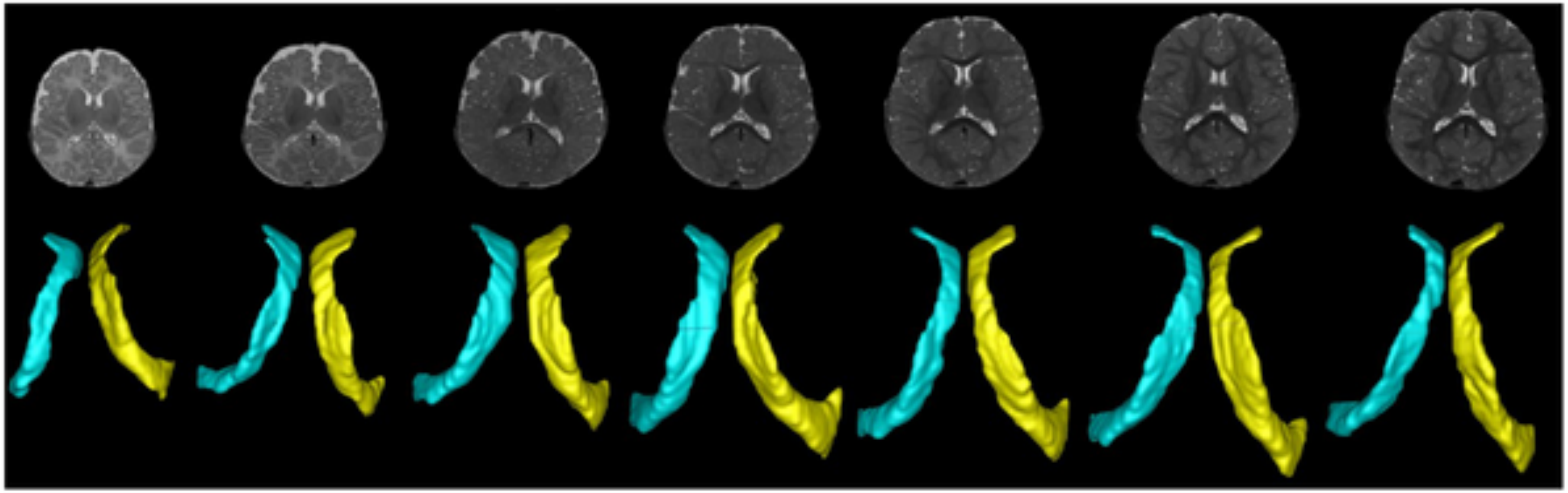
UNC Early Brain Development Studies

PIs: Drs. John H. Gilmore and Weili Lin

To track changes in behavior with brain structure, connectivity, and function, in order to characterize the progression from primary changes to subsequent clinical presentation, and to identify predictors of divergence from the typical trajectory.

- Singletons, twins, high risk
- A longitudinal prospective study
- 900 young children aged 0 to 6 years
- Recruited prenatally
  - Exclusion: ultrasound abnormality, significant fetal/ maternal medical problem, substance abuse
- 3TMRI (Seimens Allegra)
  - T1, T2, DTI, resting state fMRI
- Scanned during normal sleep(no meds)
- Ear protection, head in vac-fix device
- Success rate: 87% @ 2 weeks, 71% @ 1 year, 62% at 2 years

CS1: Longitudinal Analysis of Lateral Ventricles

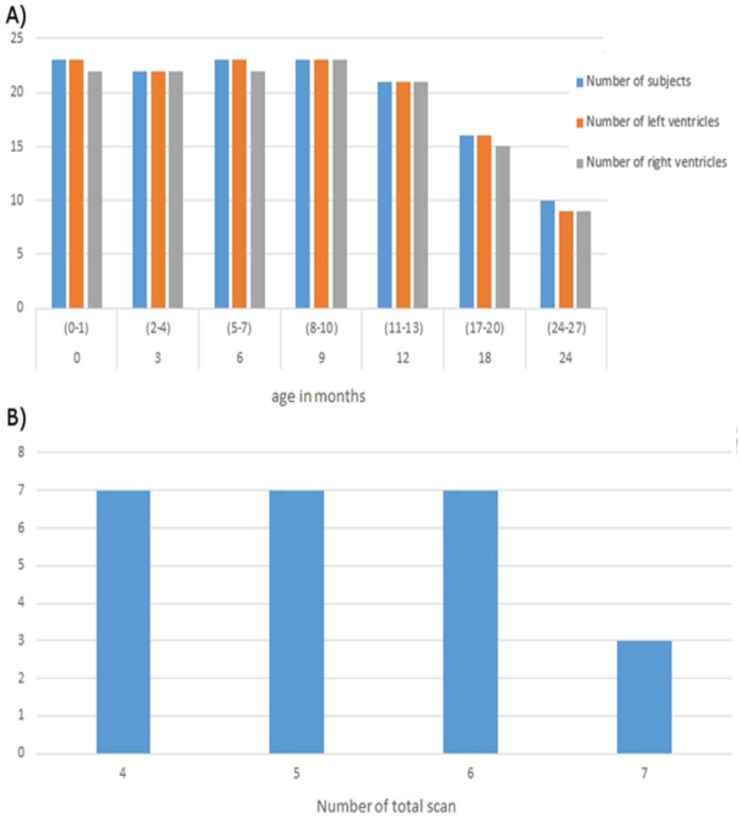


Representative T2-weighted images (upper row) from a subject imaged over the course of the first two years of life along with the segmented left and right ventricles (lower row) are shown.

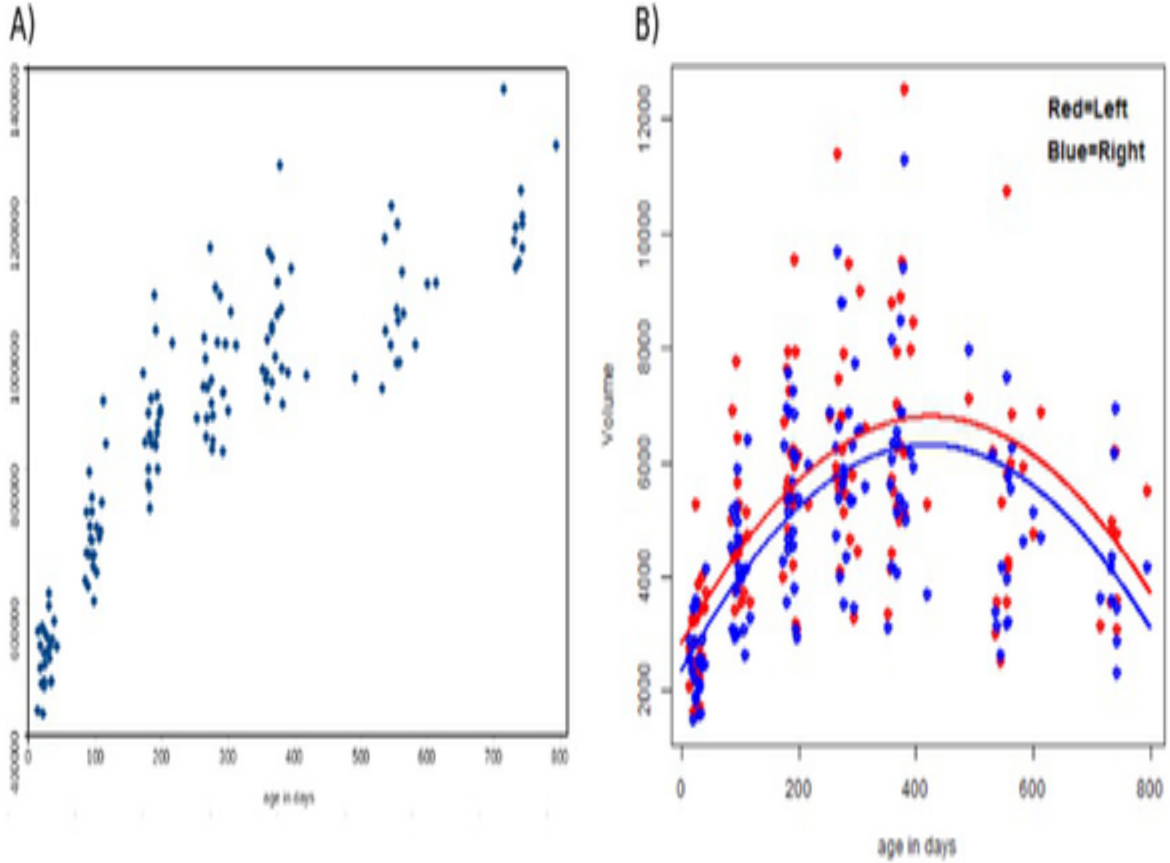
Objectives: Chart changes in brain structure

Bompard L, Xu S, Styner M, Paniagua B, et al. (2014) Multivariate Longitudinal Shape Analysis of Human Lateral Ventricles during the First Twenty-Four Months of Life. PLoS ONE 9(9):

CS1: Longitudinal Analysis of Lateral Ventricles

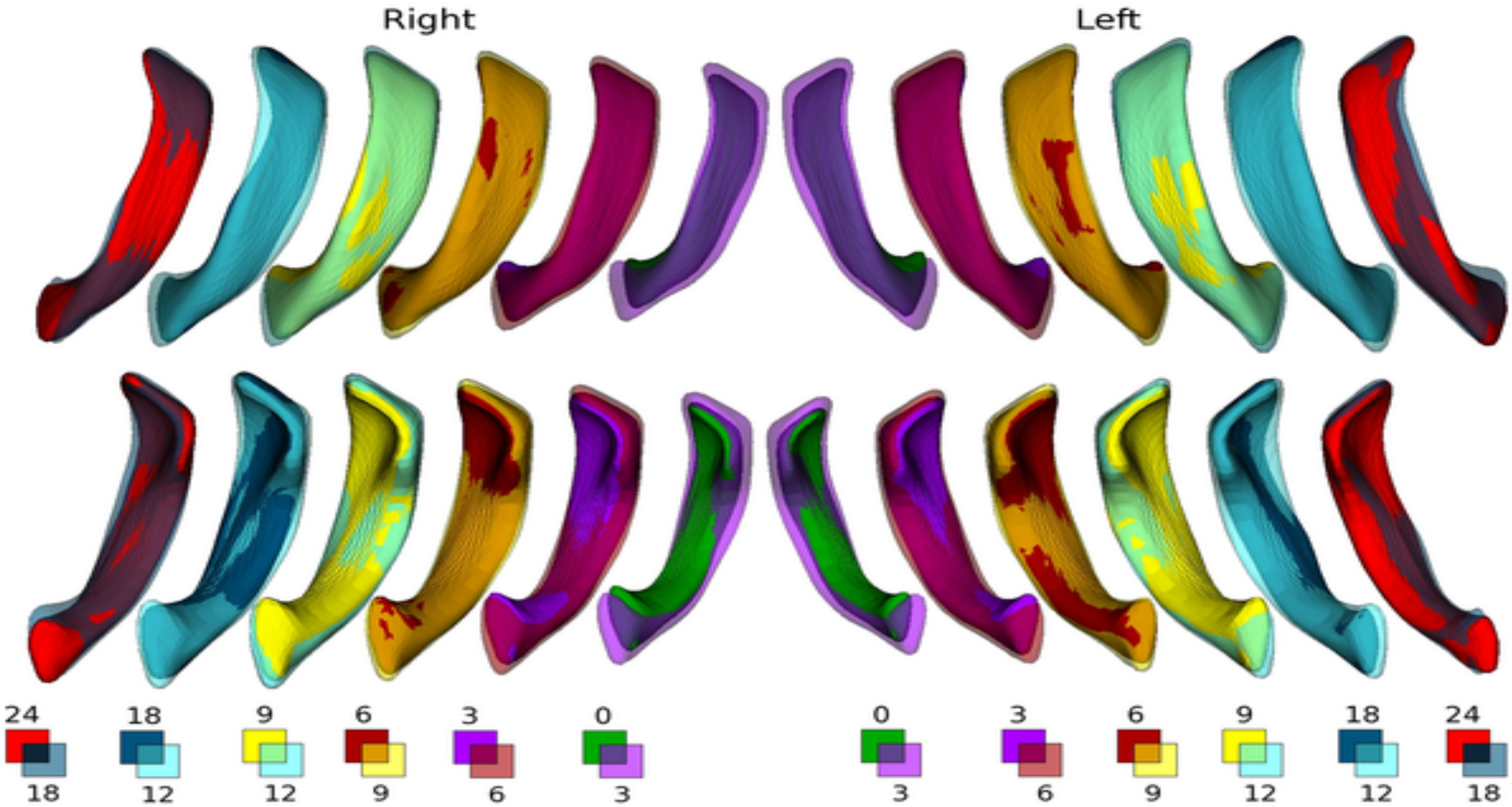


The number of subjects imaged and the number of right and left ventricles available for analysis at each age point



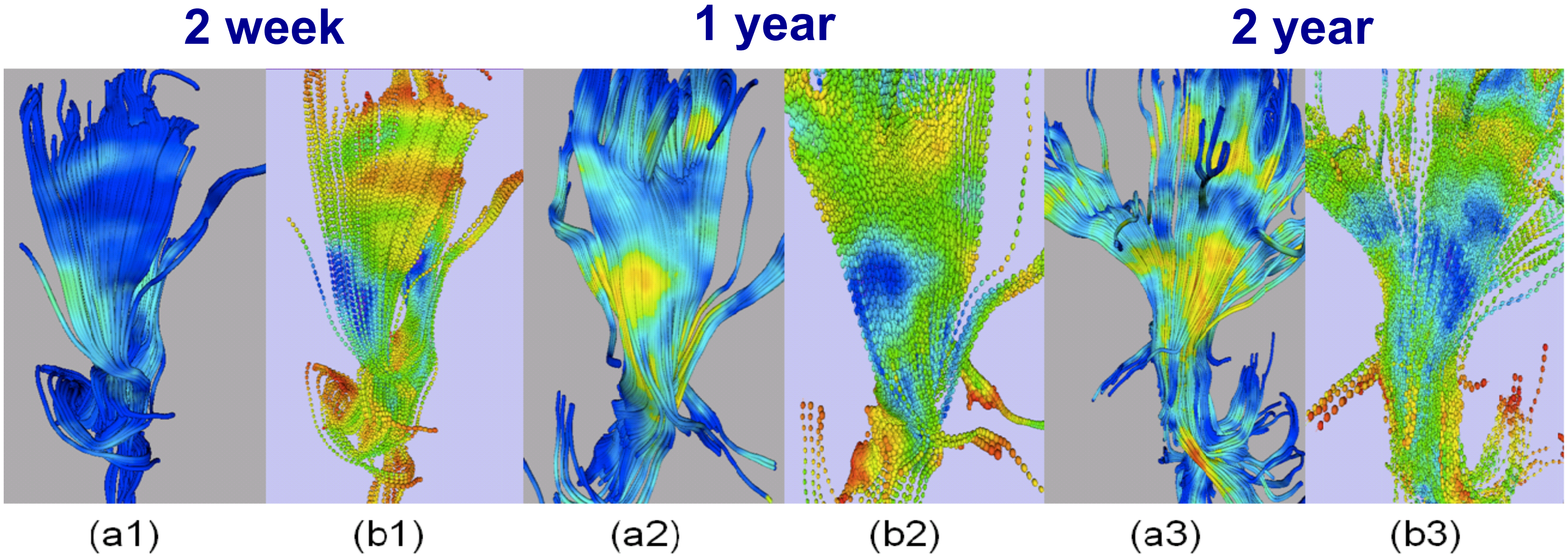
The total intracranial volume (ICV) and the left and right ventricular volumes with age are shown in A and B, respectively.

CS1: Longitudinal Analysis of Lateral Ventricles

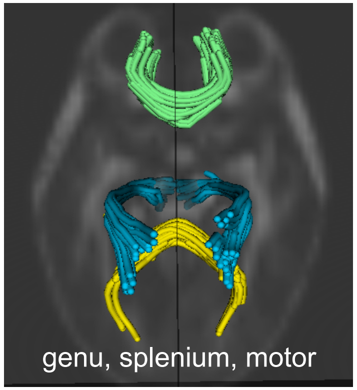


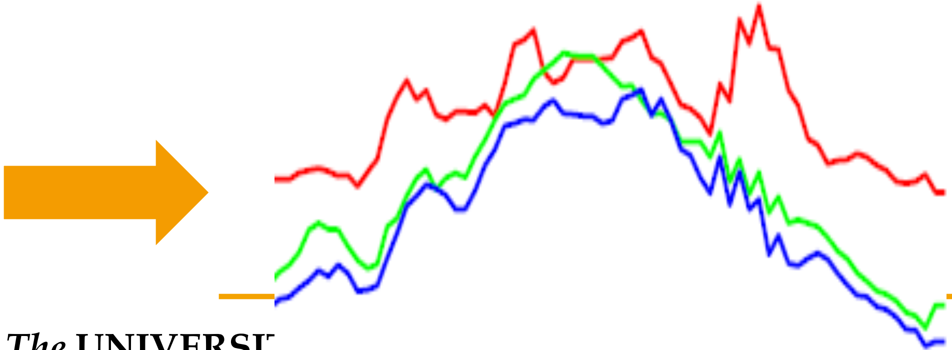
Qualitative comparisons of the shape changes of the right and left lateral ventricles between two contiguous imaging time points are shown.

CS2: White Matter Tract Development

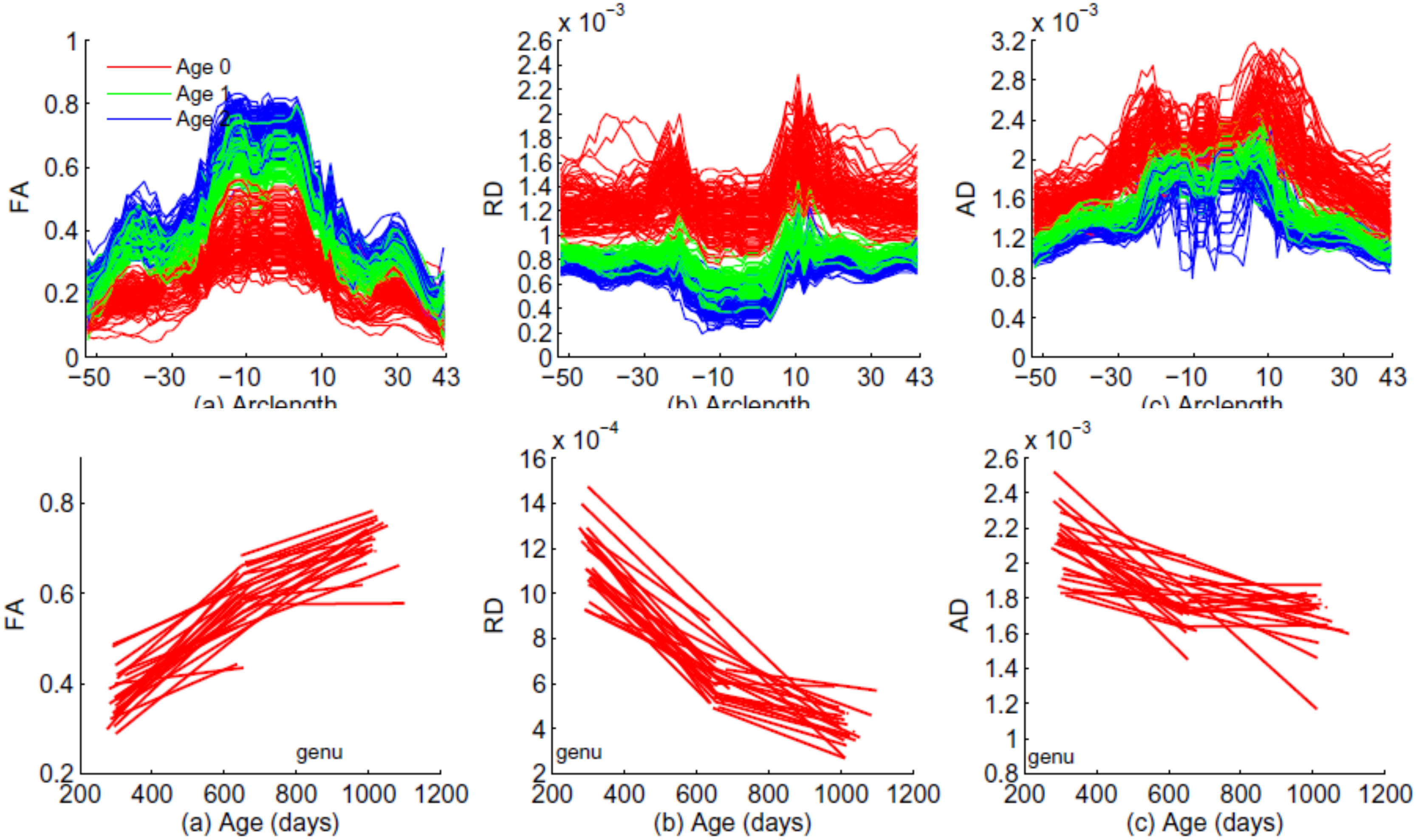


Objectives: Dynamic functional effects of covariates of interest on white matter tracts.

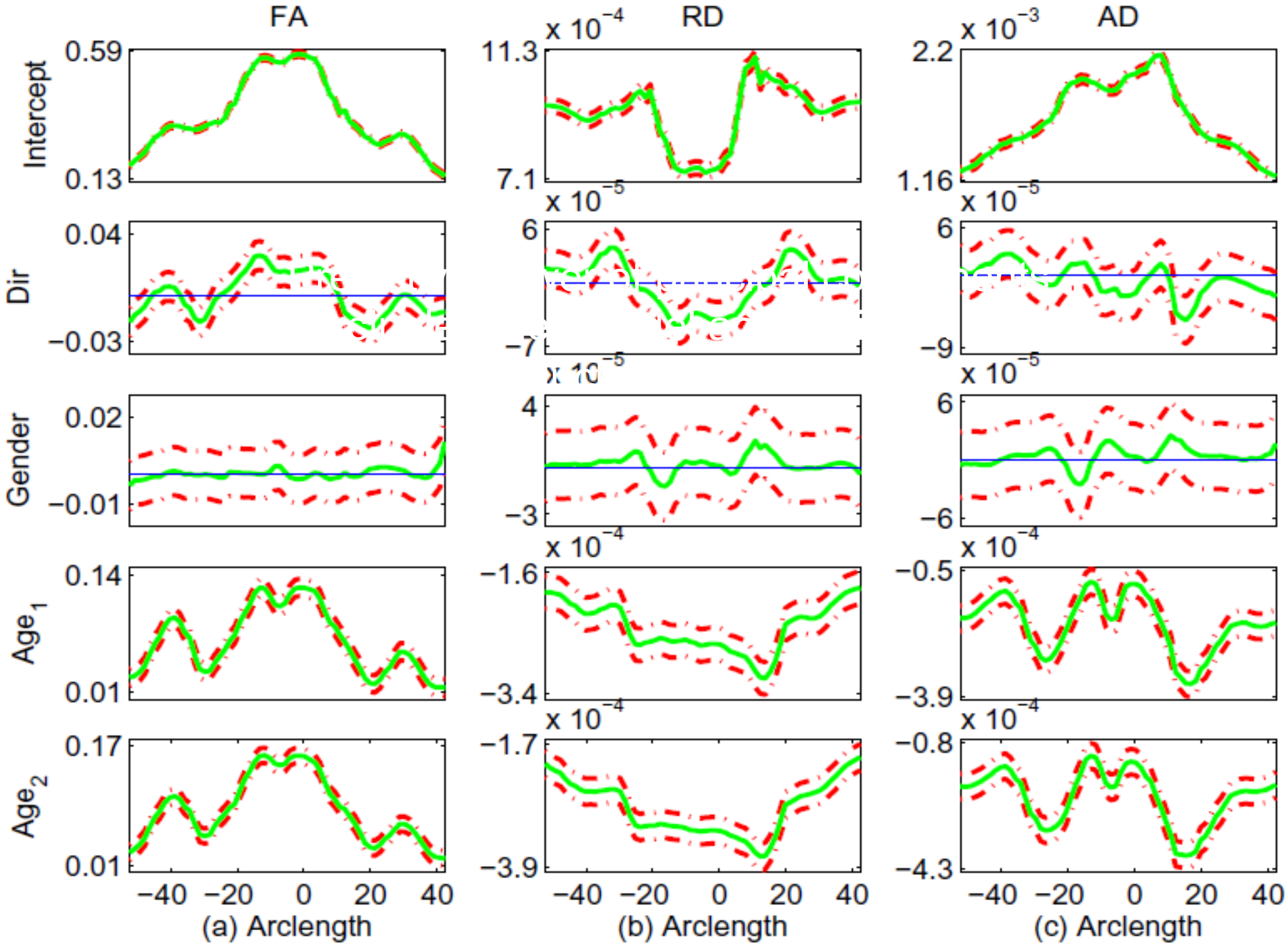




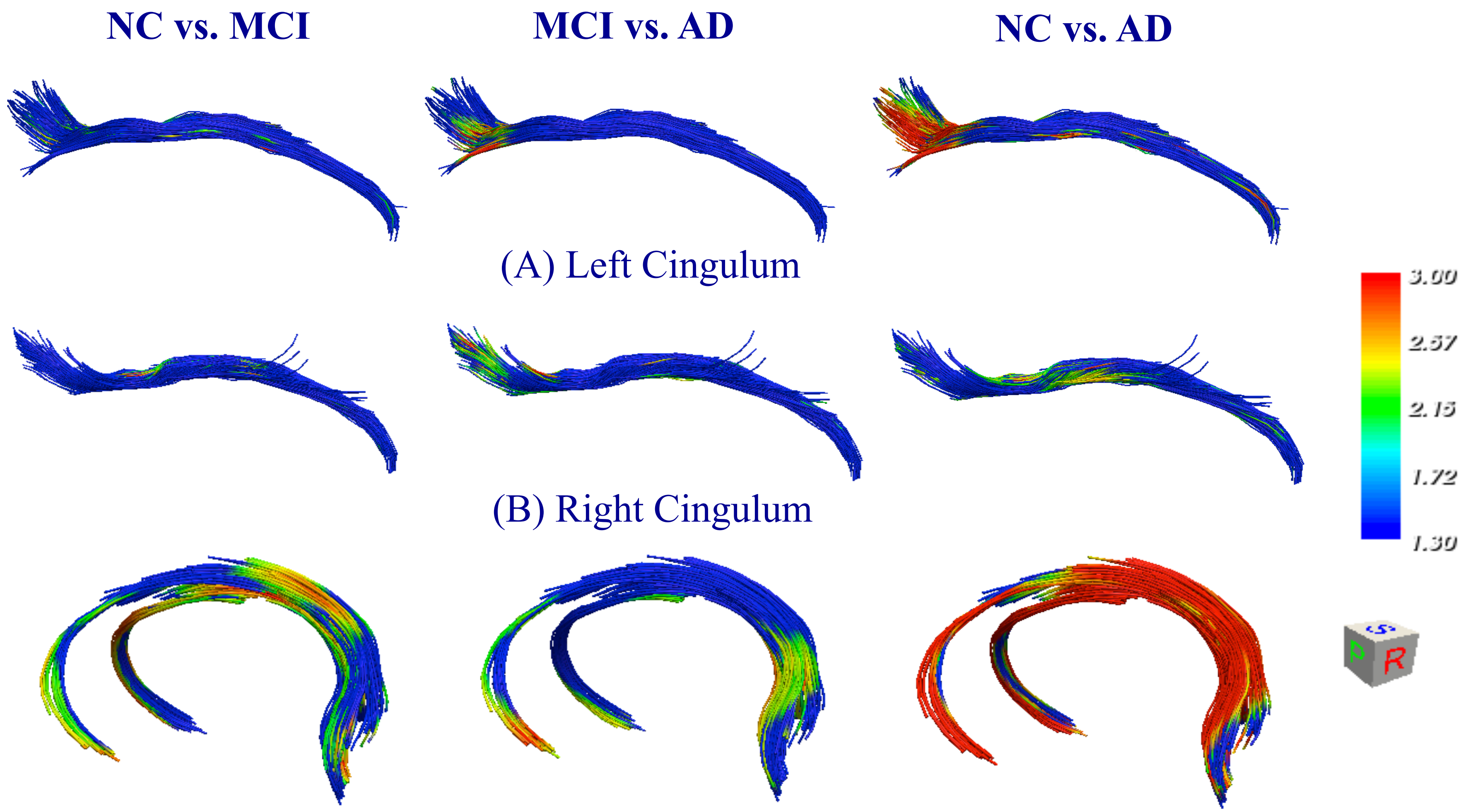
CS2: White Matter Tract Development



CS2: White Matter Tract Development



ANDI


NC vs. MCI

MCI vs. AD

NC vs. AD

(A) Left Cingulum

(B) Right Cingulum

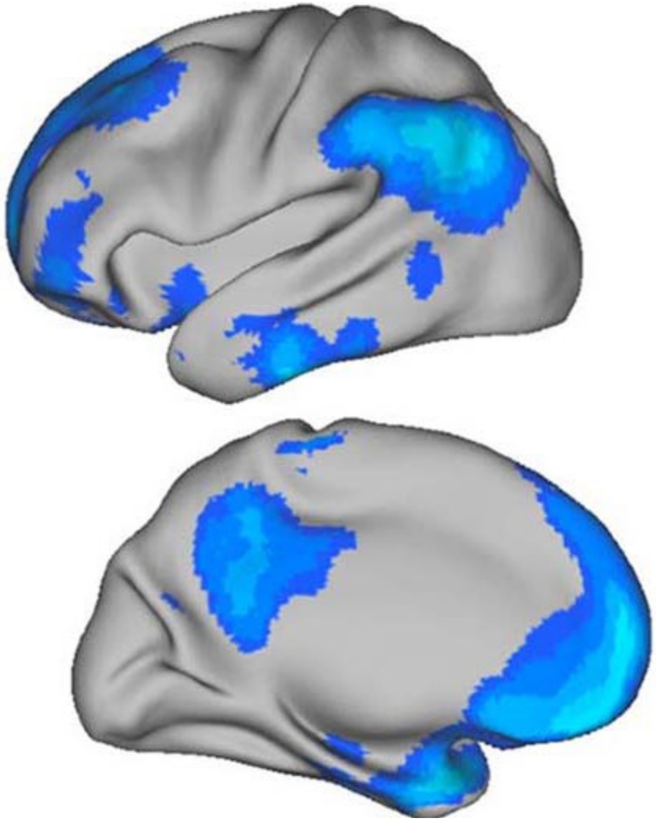
(C) Fornix

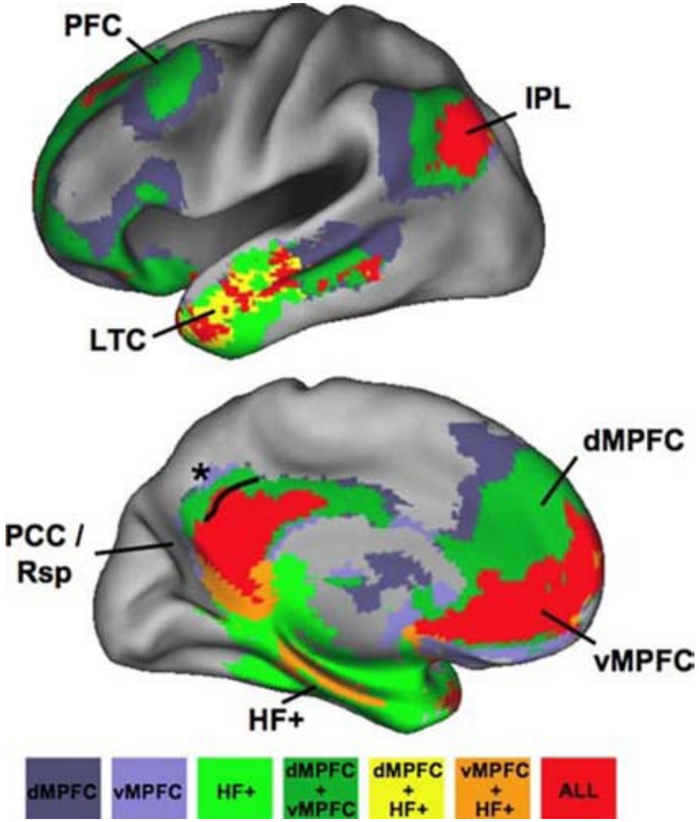
CS3: Development of Brain’s Default Network

- Purposes
  - ♦ To delineate the emergence and development of one of the most salient functional networks-the default network during the first two years of life.

- Subjects and imaging parameters
  - ♦ 71 normal subjects including 20 neonates (9M, 2412days (SD)); 24 1-year-olds (16M, 131mon) and 27 2-year-olds (17M, 251mon); 15 adut subjects (11M, 25~35 years) were also included for comparison.
  - ♦ For the rfcMRI studies, a T2*-weighted EPI sequence was used to acquire images. The imaging parameters were: TR=2sec, TE=32 ms; 33 slices; and voxel size =4x4x4 mm3. This sequence was repeated 150 times so as to provide time series images.

CS3: Development of Brain’s Default Network

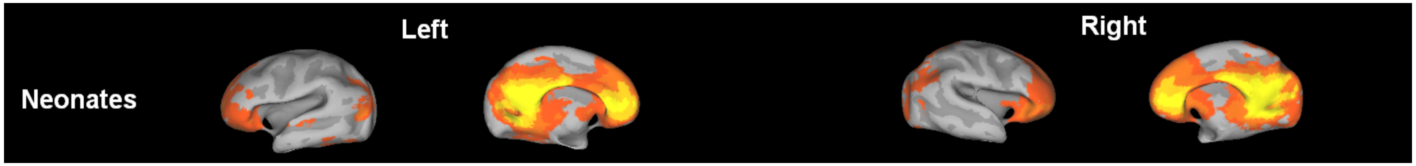




Buckner et al. (2008)

fcMRI

Results-the Emerging Default Network



1-year-old

2-year-old

Adult

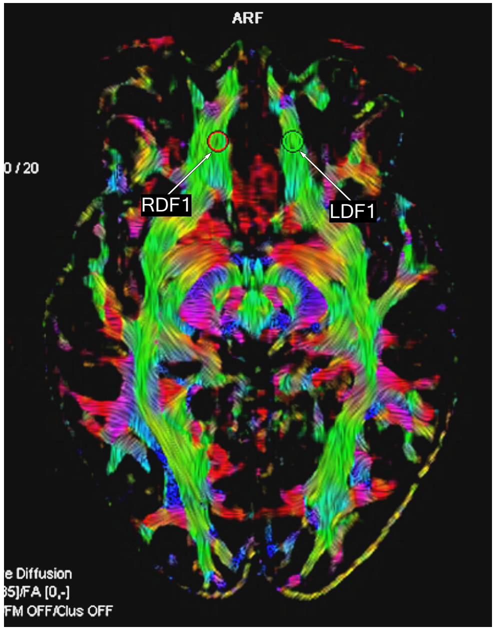
A primitive and incomplete default network is observed in 2wk olds, followed by a marked increase in the number of brain regions exhibiting functional connectivity and the percent of functional connection at 1yr olds, and finally becoming a similar network as that reported in adults at 2yr olds.

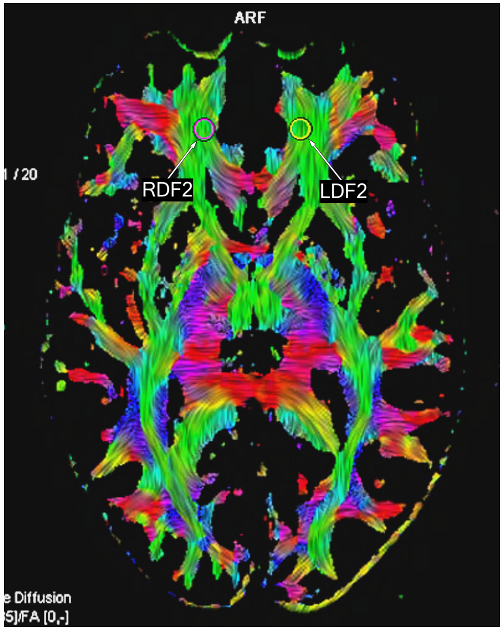
CS4: Detection of Traumatic Brain Injury

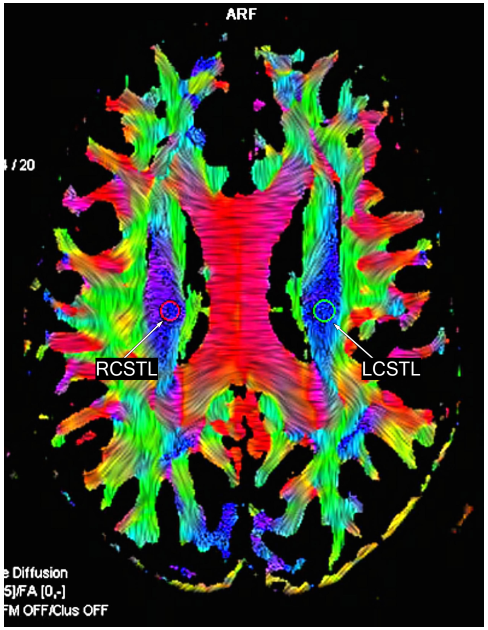
- **Purposes**

  Use DTI to detect traumatic axonal injury.

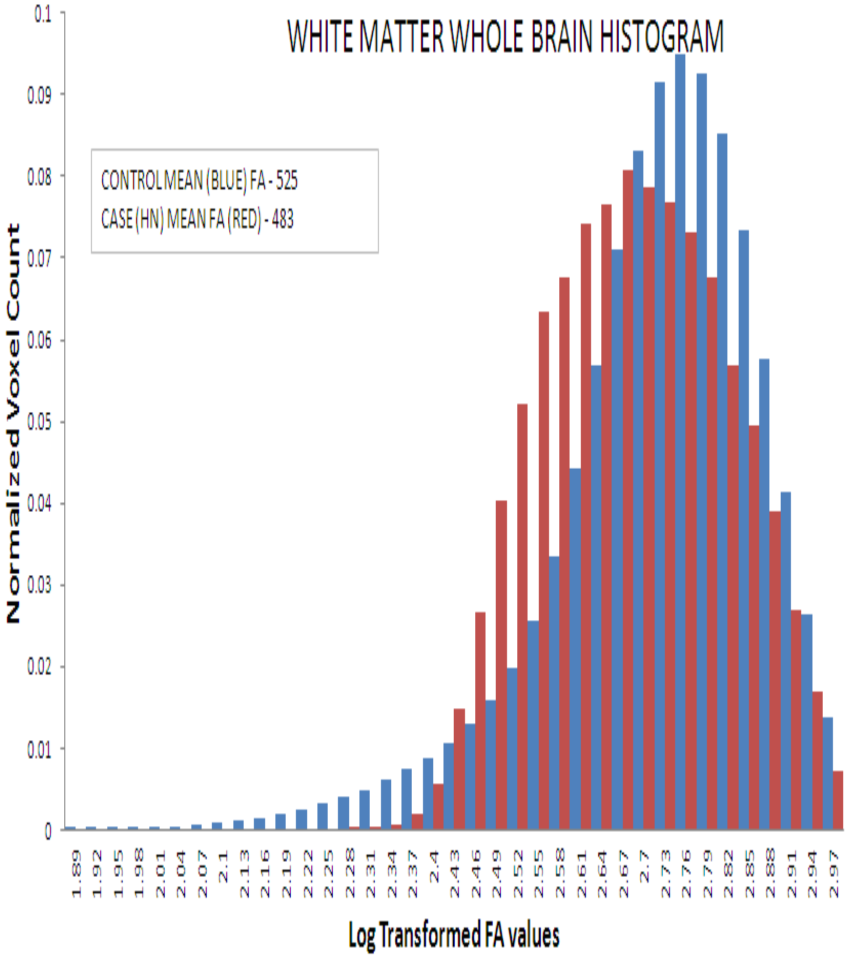
- **Subjects**
  - ♦ **235 normal subjects were also included for comparison.**
  - ♦ **Global measures (mean, median) and ROI measures**

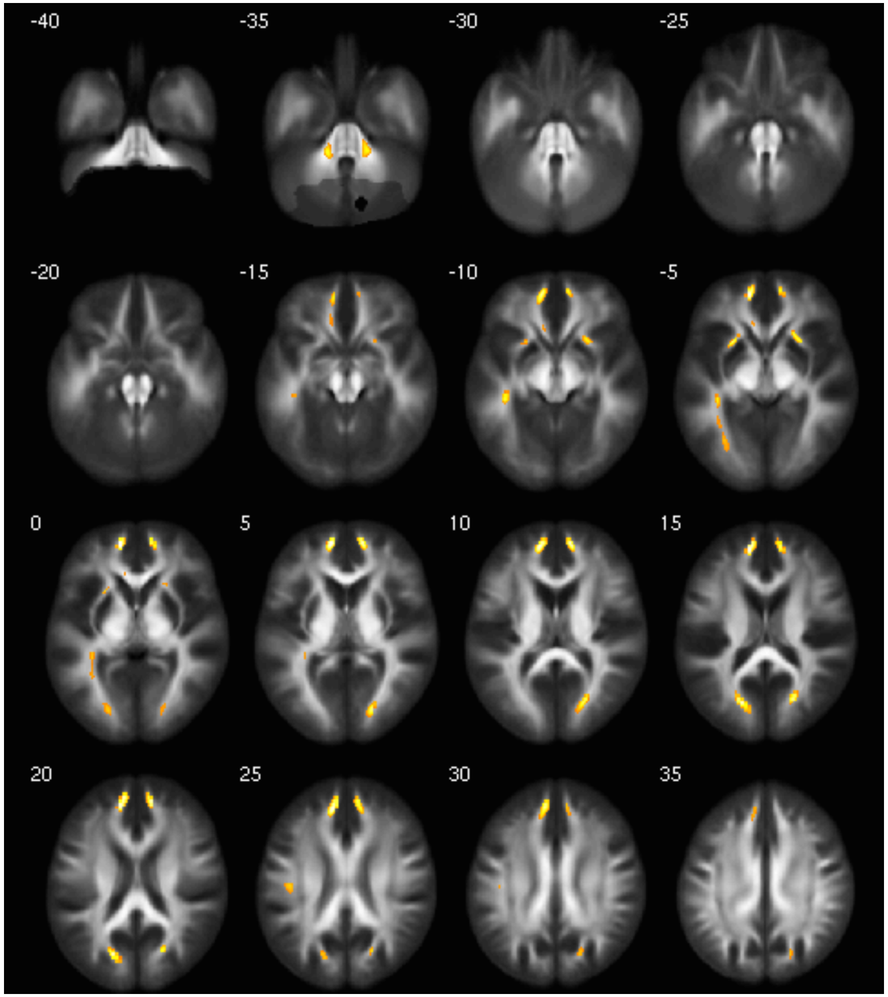






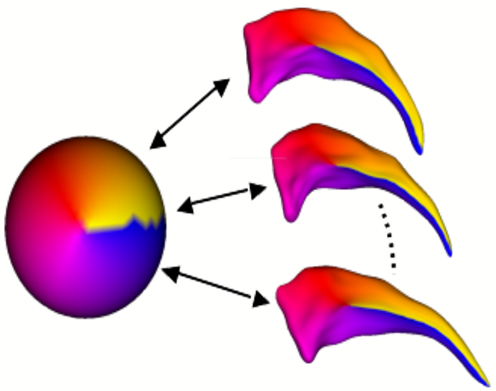
CS4: Detection of Traumatic Brain Injury

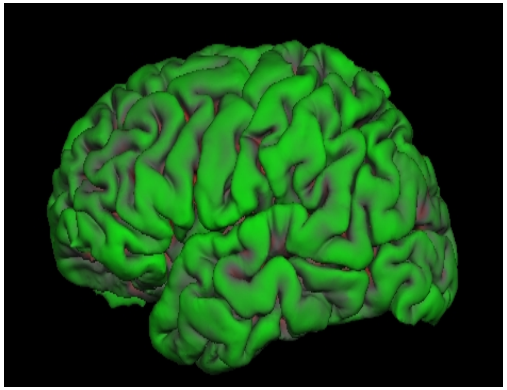


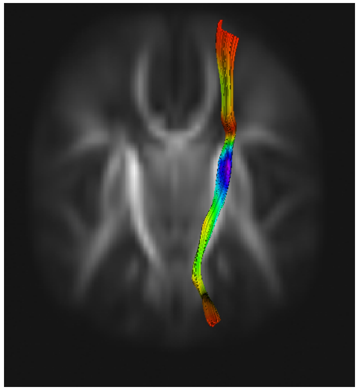


Data Structure

Smoothed Functional Data



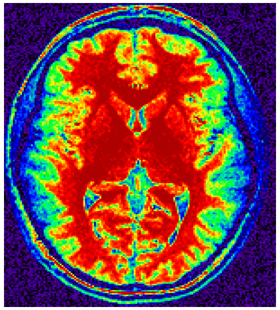


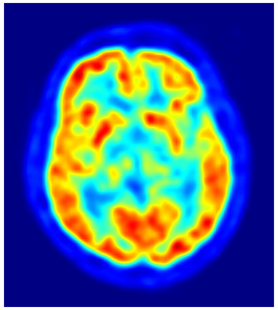


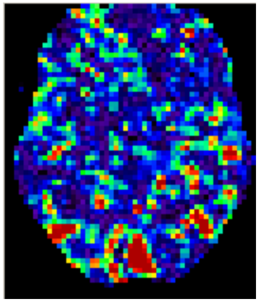
Covariates (e.g., age, gender, diagnostic)

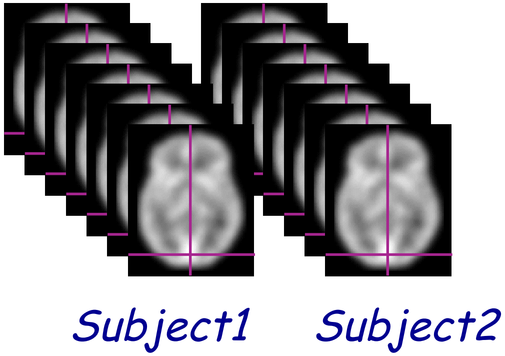
Neuroimaging Data with Discontinuity

Noisy Piecewise Smooth Function with Unknown Jumps and Edges









Covariates (e.g., age, gender, diagnostic, stimulus)

Challenging Issues

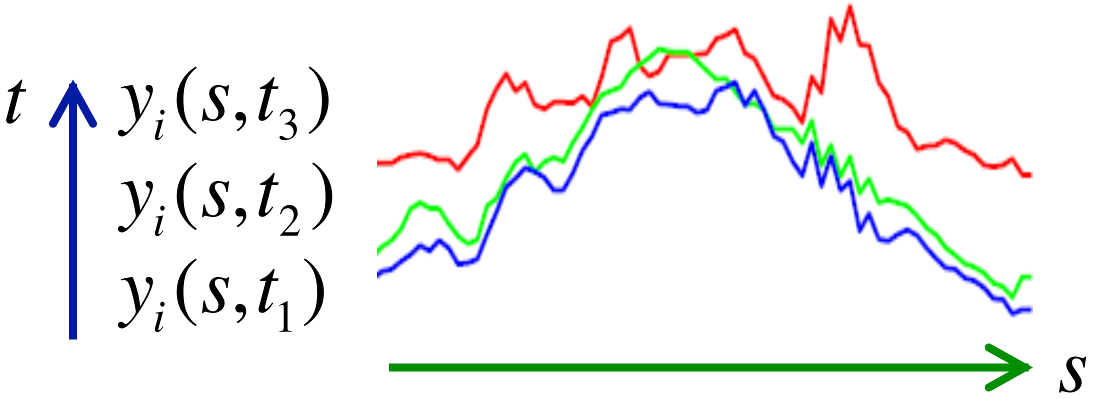
$$y_i(s) = f(x_i\ , B\ (s)) \oplus \varepsilon_i(x_i, s) \quad s \in S$$

- Complicated domains (e.g., surface mesh)
- Complicated objects (e.g., matrix response)
- Longitudinal and familial studies (e.g., heritability)
- Short-range to medium-to-long-range spatial correlations
- Asymptotic theory (e.g., simultaneous confidence bound, minimax theory)

Longitudinal Fiber Tracts

Longitudinal Data

Spatial-temporal Process



Functional Mixed Effect Models

$$y_i(s,t) = x_i\,(t)^T B\,(s) + z_i\,(t)^T \xi_i(s) + \eta_i(s,t) + \varepsilon_i(s,t)$$

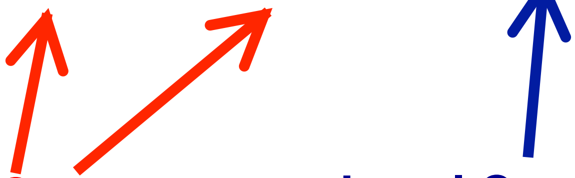
Objectives:

Dynamic functional effects of covariates of interest on functional response.

Functional Mixed Process Models

Decomposition:

$$y_i(s,t) = x_i\,(t)^T B\,(s) + z_i\,(t)^T \xi_i(s) + \eta_i(s,t) + \varepsilon_i(s,t)$$



Global Noise Components

Local Correlated Noise

$$\eta_i(\bullet,\bullet) \sim SP(0,\Sigma_\eta), \quad \xi_i(\bullet) \sim SP(0,\Sigma_\xi) \qquad \varepsilon_i(\bullet) \sim SP(0,\Sigma_\varepsilon),$$

$$\sqrt{n}\{\mathrm{vec}(\hat{B}(s) - B(s) - 0.5O(H^2)) : s \in D\} \xrightarrow{L} G(0,\Sigma_B(s,s'))$$

Ying et al. (2014). NeuroImage.
Zhu, Chen, Yuan, and Wang (2014). Arxiv.

Functional Nonlinear Mixed Effects Model

Decomposition:

$$y_{i,j}(s) = f(\phi_i(s), x_{i,j}) + \varepsilon_{i,j}(s), \quad \phi_i(s) = \beta(s) + b_i(s)$$

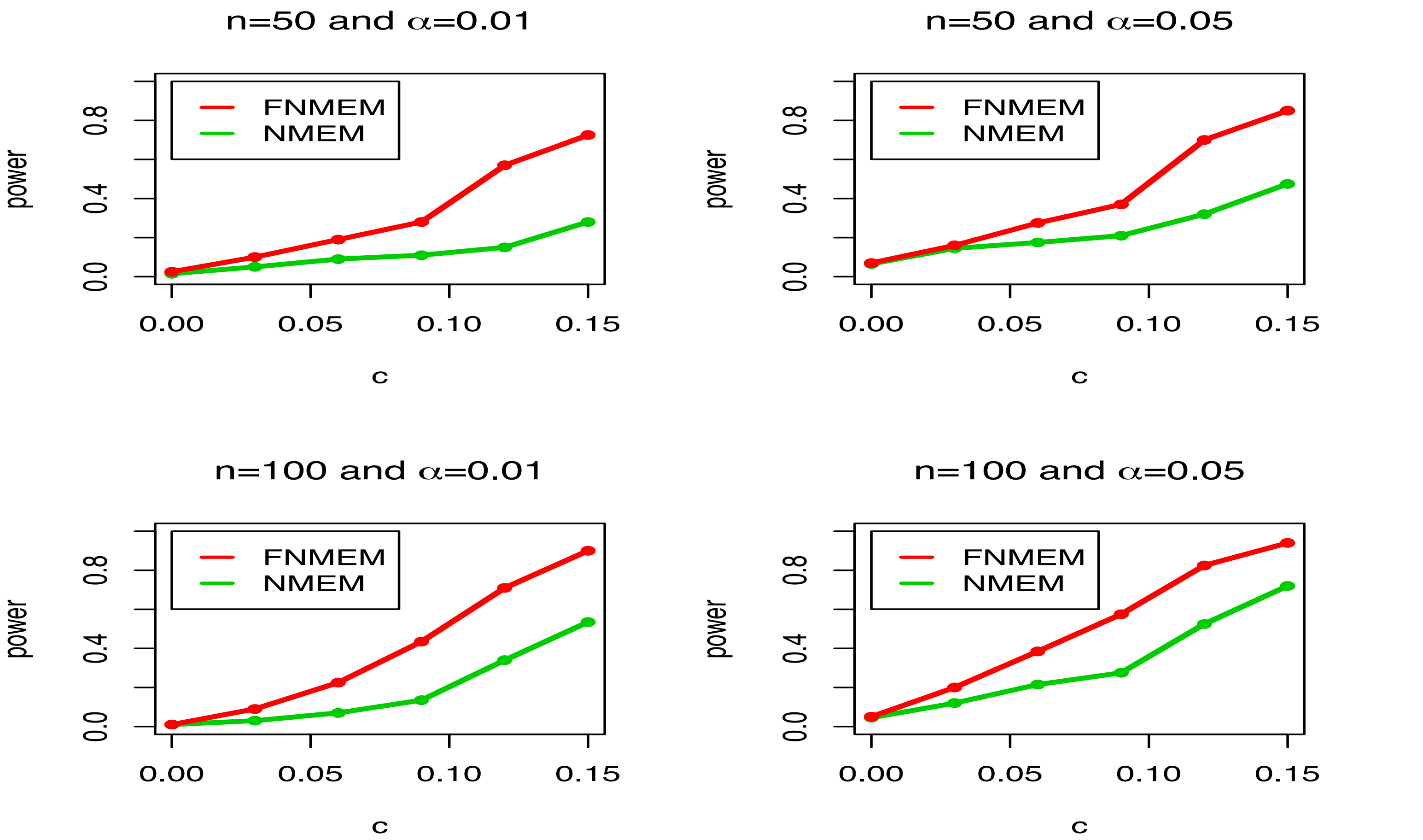
Nonlinear Function   Mixed Effect   Fixed Effect   Random Effect

Asymptotic Normality:

$$\sqrt{n}\{\mathrm{vec}(\tilde{\beta}(s) - \beta(s) - O(h^2)) : d \in D\} \xrightarrow{L} G(0, \Sigma_{\beta}(s, s'))$$

Luo, Zhu, Kong, and Zhu (2015). IPMI

Simulations



Plots of power curves. Rejection rates based on score bootstrap method are calculated using FNMEM and NMEM, with sample size 50 and 100 at significant levels 5% and 1% .

SVCM

Decomposition:

$$y_i(d) = x_i^T B(d) + \eta_i(d) + \varepsilon_i(d), d \in D$$

Piecewise Smooth Varying Coefficients

$$B(d) \in L^K$$

Long-range Correlation

$$\eta_{ij}(\bullet) \sim SP(0, \Sigma_\eta)$$

Short-range Correlation

$$\varepsilon_{ij}(\bullet) \sim SP(0, \Sigma_\varepsilon),$$

3D volume/ 2D surface

Covariance operator:

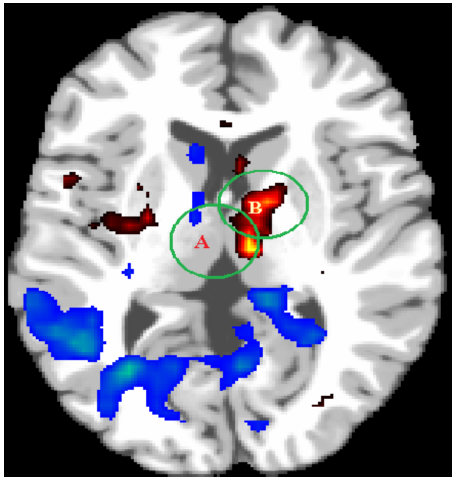
$$\Sigma_y(d,d') = \Sigma_\eta(d,d') + \Sigma_\varepsilon(d,d)$$

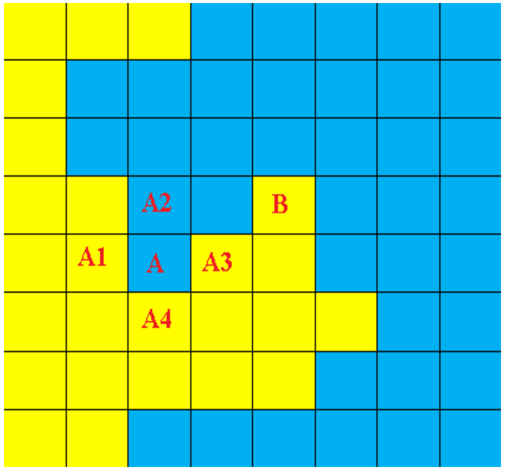
SVCM

Cartoon Model

$$B(d)=(\beta_1(d),\cdots,\beta_K(d))^T$$

- **Disjoint Partition** $D=\cup_{l=1}^{L}D_l \text{ and } D_l\cap D_{l'}=\phi$
- **Piecewise Smoothness: Lipschitz condition**
- **Smoothed Boundary**
- **Local Patch**
- **Degree of Jumps**





SVCM

Least Squares Estimates

$$\hat{B}(d;h_0) = (\sum_{i=1}^{n} x_i x_i^T)^{-1} \sum_{i=1}^{n} x_i y_i(d)$$

Smoothing residual images

$$\hat{\eta}_i(d) = S(y_i(d) - x_i^T \hat{B}(d;h_0))$$

Estimate covariance operator

$$\hat{\Sigma}_\eta(d,d') = \sum_{i=1}^{n} \hat{\eta}_i(d)\hat{\eta}_i(d')^T / n$$

$$\{(\hat{\lambda}_{kl}, \hat{\psi}_{kl}(d)) : l = 1,\ldots,\infty\}$$

Adaptively Smoothing LSEs

$$\hat{\beta}_j(d;h_s) = \sum_{d' \in B(d,h_s)} w_j(d,d';h_s)\hat{\beta}_j(d;h_0) \Big/ \sum_{d' \in B(d,h_s)} w_j(d,d';h_s)$$

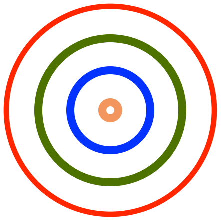
Calculate standard deviation

Propogation-Seperation Method
J. Polzehl and V. Spokoiny, (2000,2005)

Adaptive Smoothing Methods

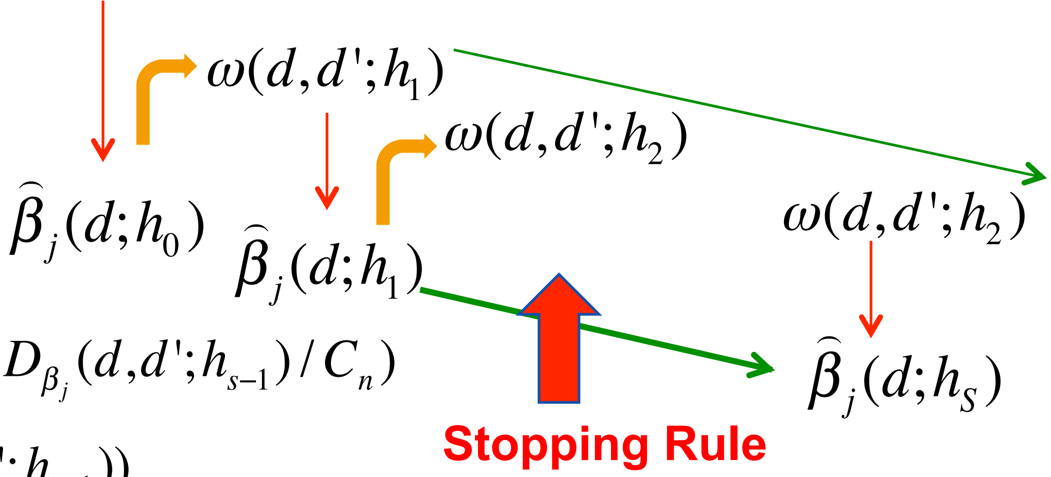
At each voxel d

$$\hat{\beta}_j(d;h_s)=\frac{\sum_{d'\in B(d,h_s)} w_j(d,d';h_s)\hat{\beta}_j(d;h_0)}{\sum_{d'\in B(d,h_s)} w_j(d,d';h_s)}$$



- **Increasing Bandwidth**
- **Adaptive Weights**
- **Adaptive Estimates**

$$0<h_0<h_1<\cdots<h_S=r_0$$

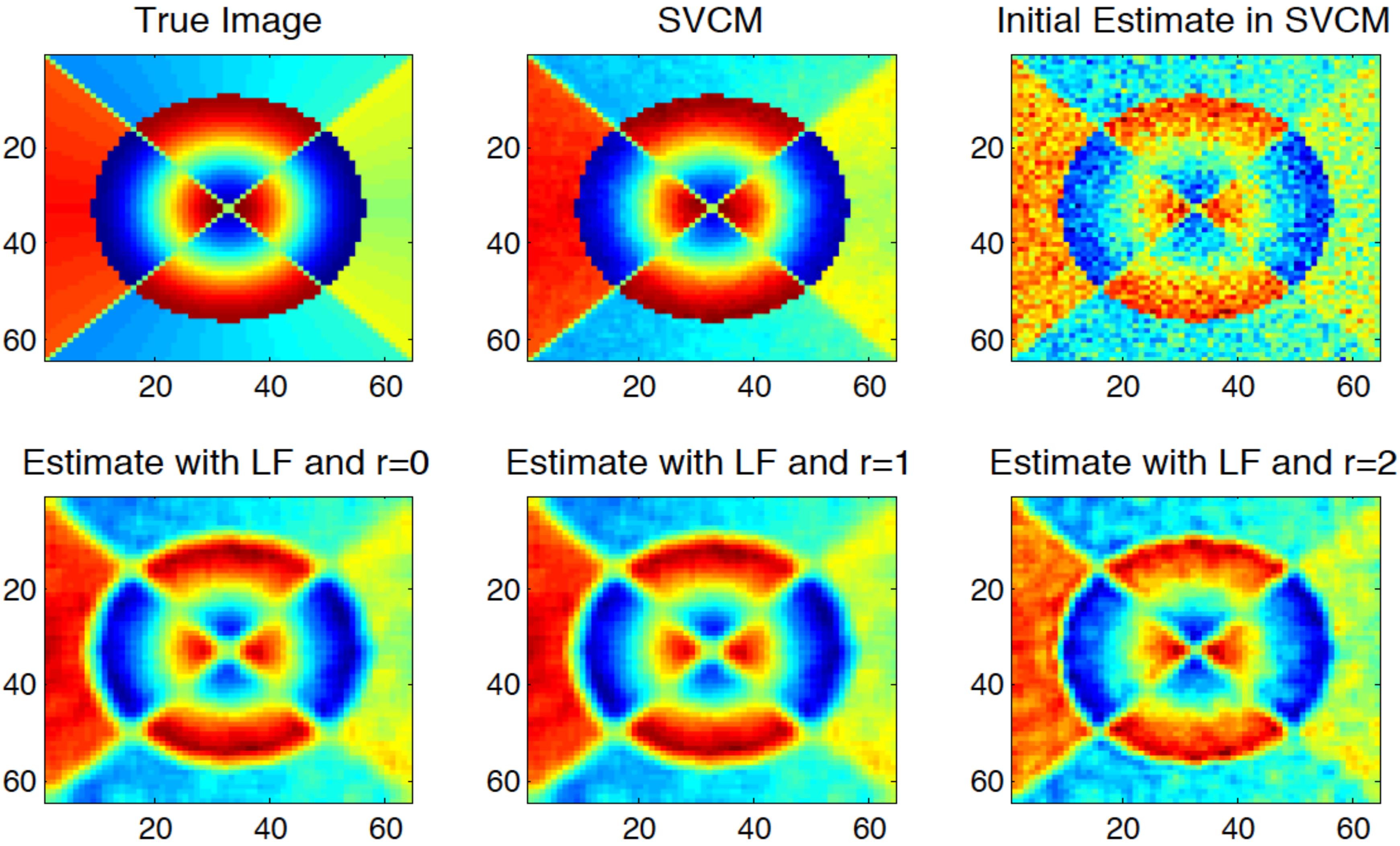


$\omega(d,d';h_1)$
$\omega(d,d';h_2)$
$\widehat{\beta}_j(d;h_0)$
$\widehat{\beta}_j(d;h_1)$
$\omega(d,d';h_2)$
$\widehat{\beta}_j(d;h_S)$
Stopping Rule

$$\omega(d,d';h_s)=K_{loc}(\|d-d'\|/h_s)K_{st}(D_{\beta_j}(d,d';h_{s-1})/C_n)$$

$$D_{\beta_j}(d,d';h_{s-1})=\rho(\widehat{\beta}_j(d;h_{s-1}),\widehat{\beta}_j(d';h_{s-1}))$$

Simulation



True Image

SVCM

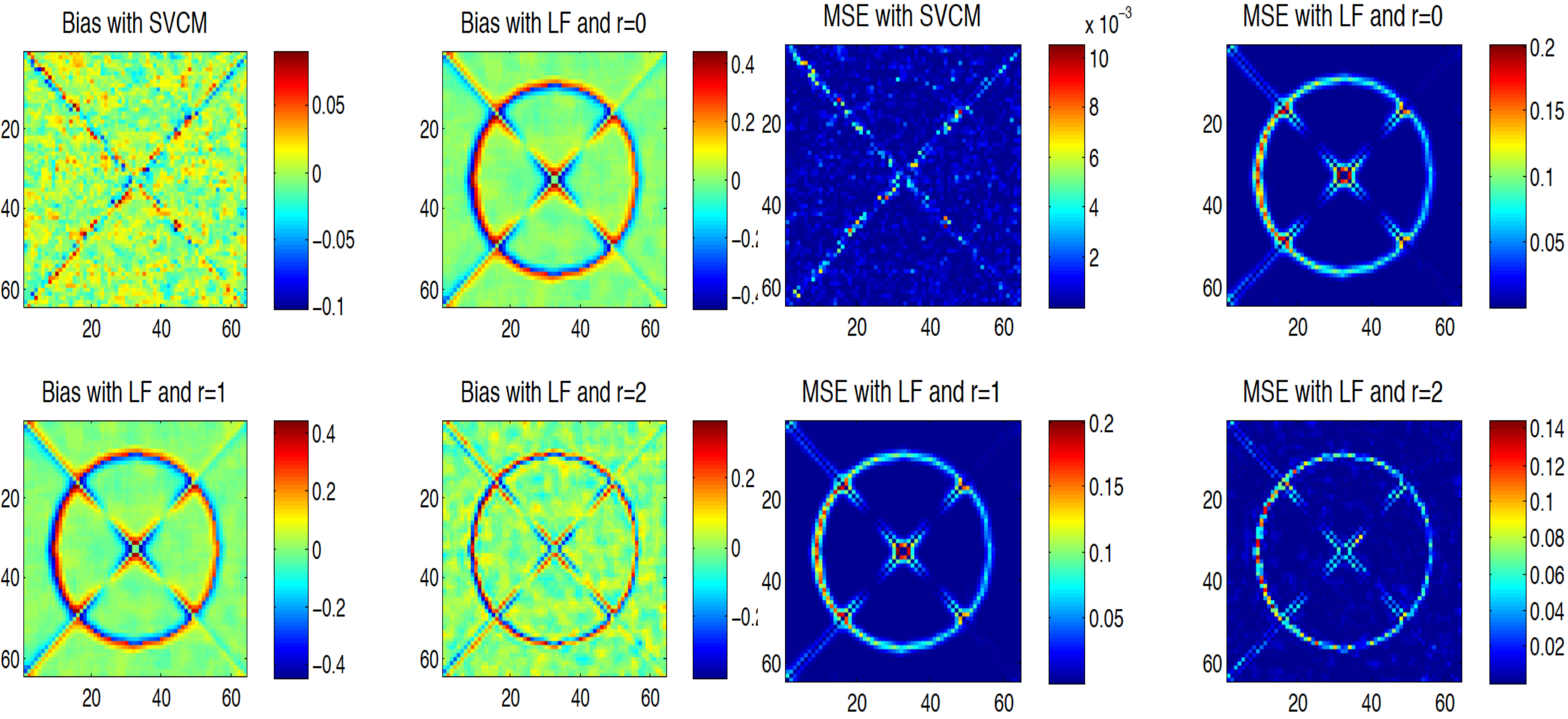
Initial Estimate in SVCM

Estimate with LF and r=0

Estimate with LF and r=1

Estimate with LF and r=2

Simulation


Bias with SVCM

Bias with LF and r=0

MSE with SVCM

MSE with LF and r=0

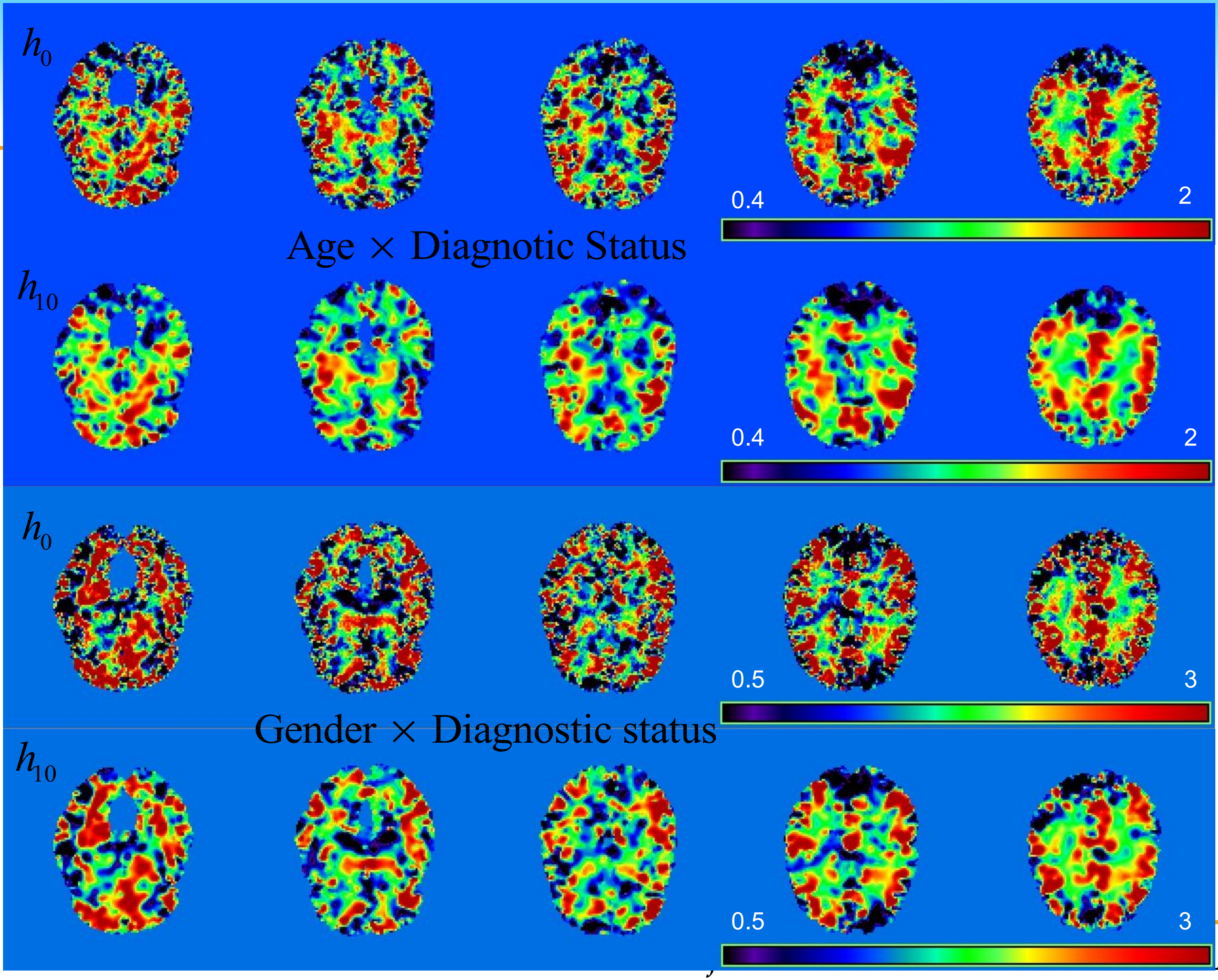
Bias with LF and r=1

Bias with LF and r=2

MSE with LF and r=1

MSE with LF and r=2

Interaction effect estimates



Longitudinal Neuroimaging Data

$$y_i(d,t) = \mu(d, \mathbf{x}_i(t)) + \eta_i(d,t) + \epsilon_i(d,t) \text{ for } i = 1, \ldots, n, \qquad (2)$$

where

Across subjects & time

$\mu(d, \mathbf{x}_i(t))$ is the fixed main effect, which depends semi-parametrically on the covariates $\mathbf{x}_i(t) = (x_{i,1}(t), \ldots, x_{i,p}(t))^T$,

Across Modality & time

$\eta_i(d,t)$ characterizes both individual image variations from $\mu(d, \mathbf{x}_i(t))$ and the medium-to-long-range dependence of imaging data between $y_i(d,t)$ and $y_i(d', t')$ for any $(d,t) \neq (d', t')$,

Local spatial-temporal smoothness

$\epsilon_i(d,t)$ are spatially and temporally correlated errors that capture the local (or short-range) dependence of imaging data,

$\eta_i(d,t)$ and $\epsilon_i(d,t)$ are, respectively, independent and identical copies of $\mathrm{GP}(\mathbf{0}, \Sigma_\eta)$ and $\mathrm{GP}(\mathbf{0}, \Sigma_\epsilon)$ and mutually independent.

Hyun,J.W., Li, Y. M., Wang, Y.P., H. Zhu (2014) LSGPP. In Submission.

ADNI PET Data

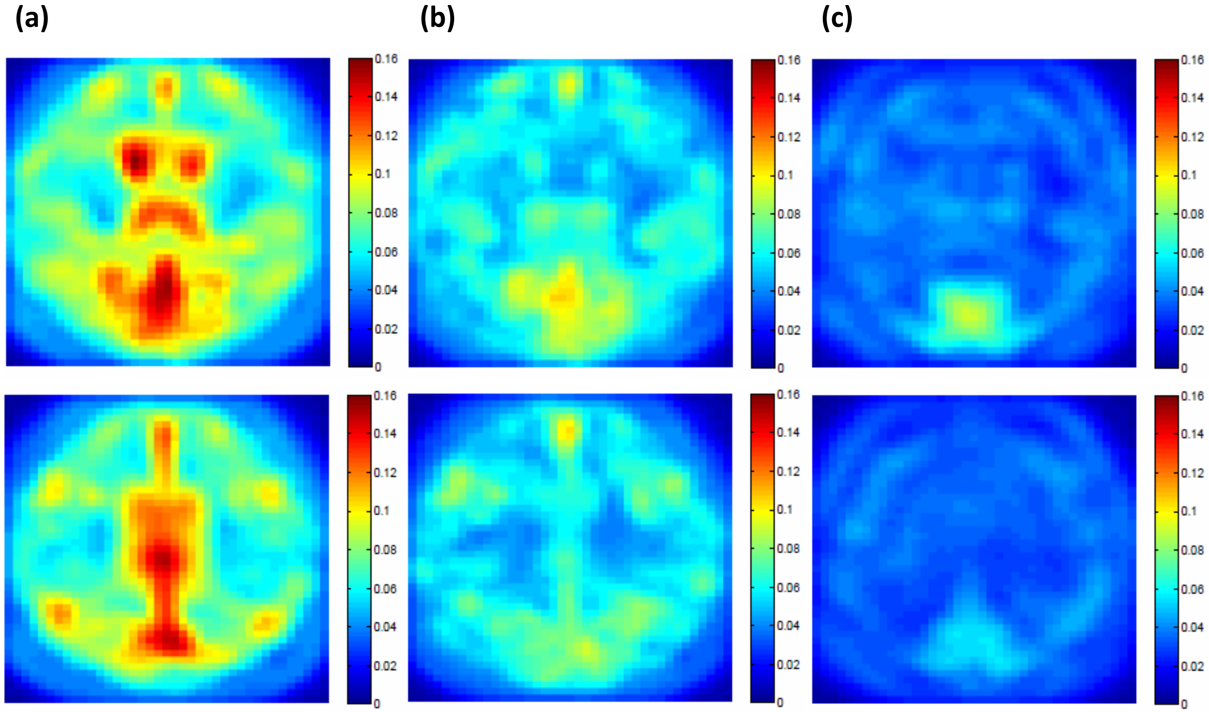
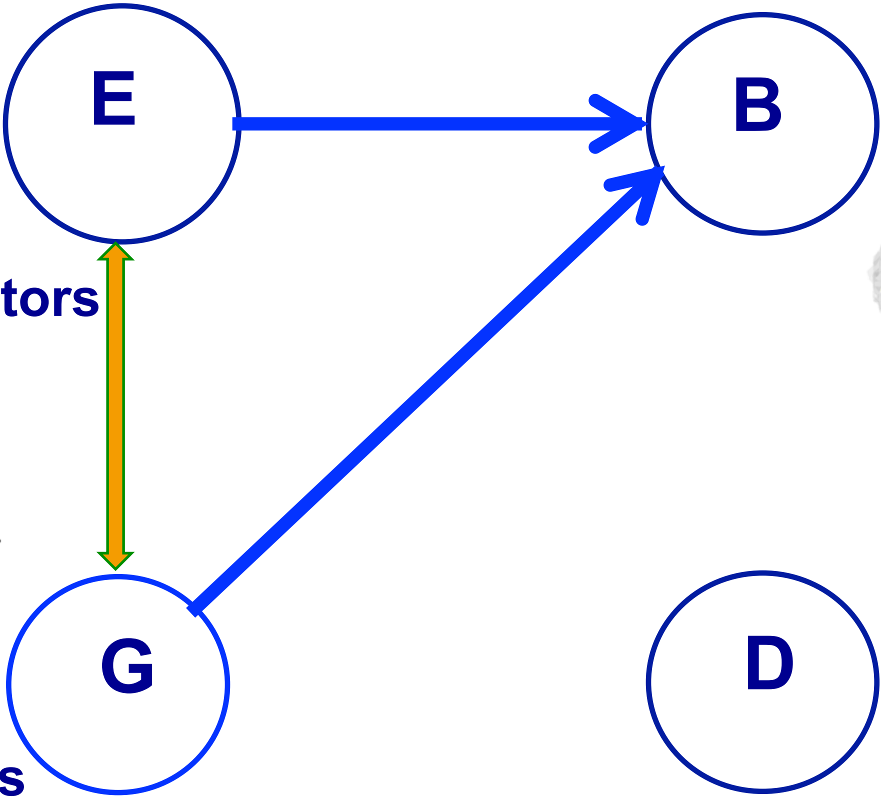
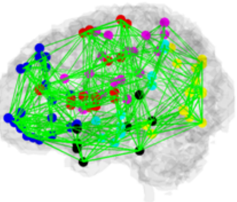


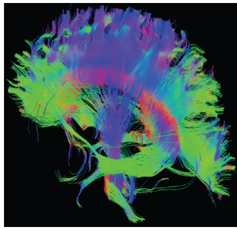
Figure : rtMSPE maps for prediction of ADNI PET images at month 12 for 79 test subjects. Selected slices are shown for (a) Semi-parametric model; (b) Semi-parametric model+FPCA; (c) Semi-parametric model+FPCA+Spatial-temporal model.

Image-on-Genetic Association Models

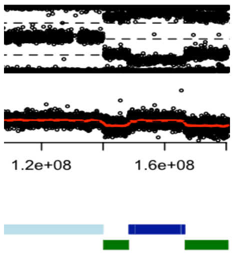
Big Data Integration

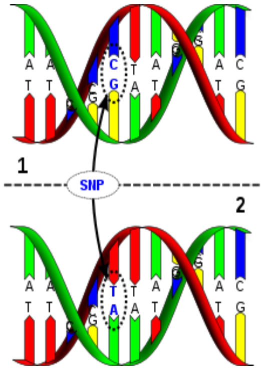


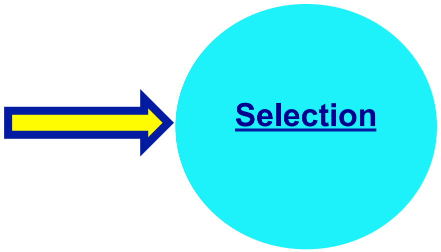




E: environmental factors







G: genetic markers

D: disease

http://en.wikipedia.org/wiki/DNA_sequence

References

Statistical Methodologies:

1. **Lin, J., Zhu, H.T.,** Knickmeyer, R., Styner, M., Gilmore, J. H. and Ibrahim, J.G. (2012). Projection Regression Models for Multivariate Imaging Phenotype. *Genetic Epidemiology*, 36, 631-641.
2. **Lin, J**., ***Zhu, H.T.,*** , Mihye, A., and Ibrahim, J.G. (2014). Functional Mixed Effects Models for Candidate Genetic Mapping in Imaging Genetic Studies. *Genetic Epidemiology,* 38(8):680-91.
3. **Zhu, H.T.,** Khondker, Z. S., Lu, Z.H., and Ibrahim, J. G. (2014). Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of American Statistical Association*, 507, 977-990.
4. **Zhu, HT**, Fan, J., and Kong, L. (2014). Spatial varying coefficient model and its applications in neuroimaging data with jump discontinuity. *Journal of American Statistical Association*, 109, 1084-1098.
5. **Sun, Q.**, ***Zhu, H.T.,*** Liu, Y. F., and Ibrahim, J.G. SPReM: Sparse Projection Regression Model for High-dimensional Linear Regression. *Journal of American Statistical Association*, in press, 2015.
6. Huang, M., Nichols, T., Huang, C., Yu, Y., Lu, Z., Knickmeyer, R. C., Feng, Q., and ***Zhu, H. T.*** (2015). FVGWAS: Fast Voxelwise Genome Wide Association Analysis of Large-scale Imaging Genetic Data, *NeuroImage*, in press.

Neuroscience/Psychiatry:

1. Bryant, C., Giovanello, K. S., Ibrahim, J. G., Shen, D. G., Peterson, B. S., and **Zhu, H.T.** (2013) Mapping the heritability of regional brain volumes explained by all common SNPs from the ADNI study. *PLOS ONE.*
2. Kai Xia, Yang Yu, Mihye Ahn, **H. Zhu**, Fei Zou, John Gilmore, Rebecca Christine Knickmeyer. Environmental and genetic contributors to salivary testosterone levels in infants. *Frontiers in Endocrinology.* 2014.
3. Wei Gao, Amanda Elton, **H. Zhu,** Sarael Alcauter, J. Smith, John H Gilmore, and Weili Lin. (2014). Inter-subject Variability of and Genetic Effects on the Brain's Functional Connectivity during Infancy. *Journal of Neuroscience*, 34: 11288-11296.
4. Knickmeyer, R. C., Wang, J. P., **Zhu, H.T.,** Geng, X., Woolson, S., Hamer, R. M., Konneker, T., Lin, W. L., Styner, M., and Gilmore, J. H. (2014). Common variants in psychiatric risk genes predict brain structure at birth. *Cerebra Cortex.* 24(5):1230-46.
5. S. J. Lee, R.J. Steiner; Shikai Luo; Michael C Neale; Martin Styner; Hongtu Zhu; John H. Gilmore. (2015). Quantitative tract-based white matter heritability in twin neonates. *NeuroImage,* 111:123-135.

Genome-wide Identification of Variants Affecting Early Human Brain Development

PI: Dr. Knickmeyer

The central objective of this project is to identify genetic factors which explain variation in neonatal brain structure, as assessed by magnetic resonance imaging (MRI) and diffusion tensor imaging (DTI).

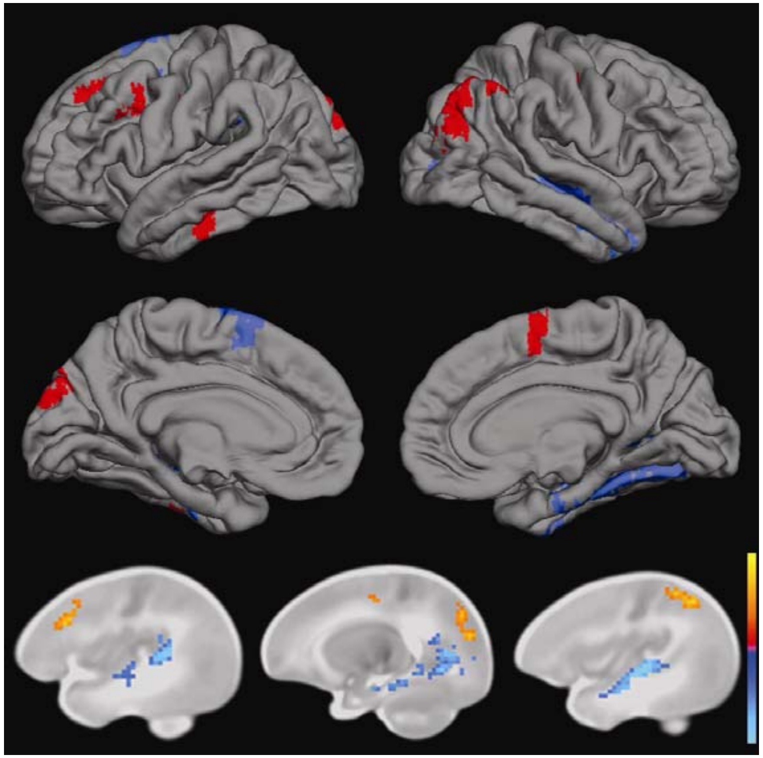
- Singletons, twins, high risk
- A longitudinal prospective study
- 900 young children aged 0 to 6 years
- 3TMRI (Seimens Allegra)
  - T1, T2, DTI, resting state fMRI
- Genotyping: the Illumina OMNI quad beadchip with 1,140,419 single nucleotide polymorphisms (SNPs) and more than 6,000 common and 5,000 rare CNV regions with 10-15 markers per region

CS5: Candidate Genes and Neonatal Gray Metter

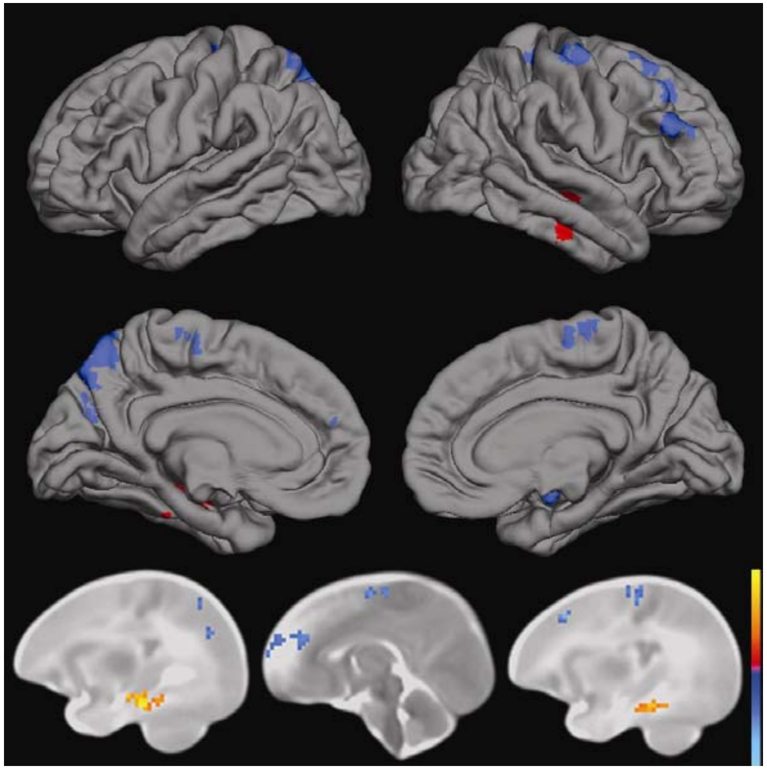
- **272 neonates**
  - 152 Male and 120 Female, 144 singletons, 128 twins
- **Tensor based morphometry**
- **Candidate Genes**

  – apolipoproteinE (APOE;ε3ε4 vs.ε3ε3)
  – catechol-O-methyltransferase (COMT, rs4680)
  – disrupted-in-schizophrenia-1(DISC1,rs821616andrs6675281)
  – neuregulin1 (NRG1,rs35753505andrs6994992)
  – estrogenreceptoralpha (ESR1,rs9340799andrs2234693)
  – brain-derivedneurotrophicfactor(BDNF,rs6265)
  – glutamatedecarboxylase1(GAD1akaGAD67,rs2270335)

CS5: Candidate Genes and Neonatal Gray Metter

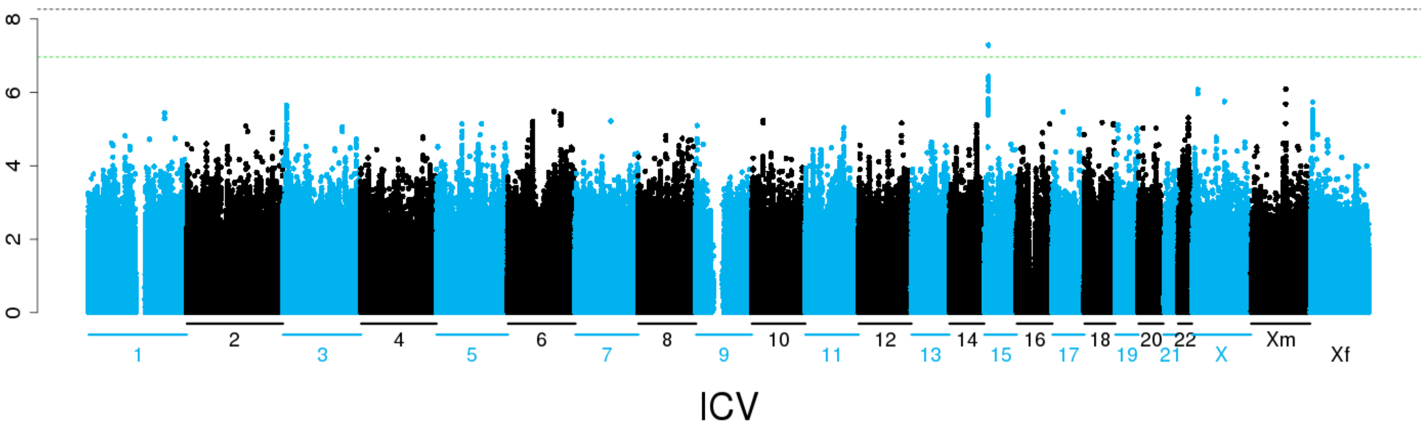


COMT (rs4680)

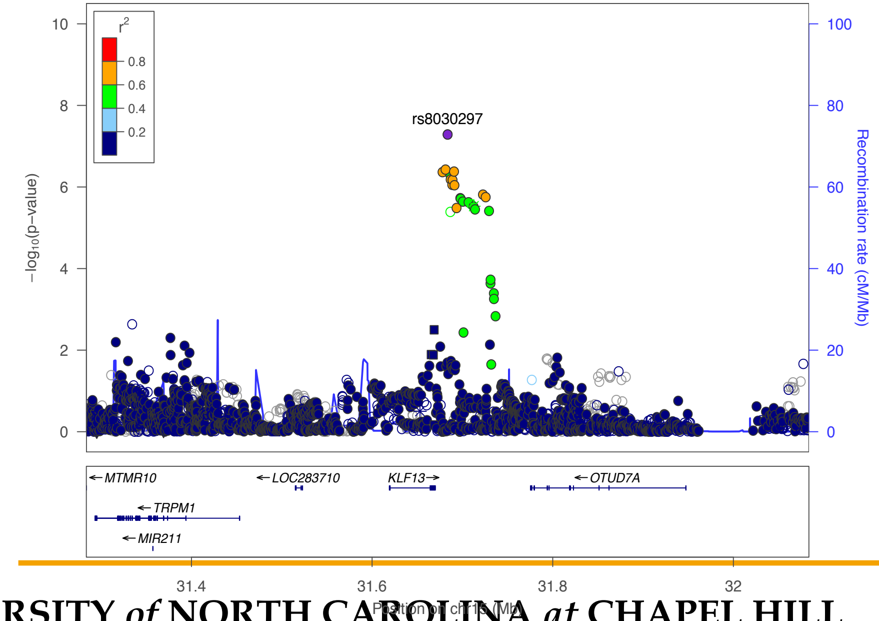


APOE

CS6: GWAS Neonatal ROIs



562 subjects (296 singletons and 246 twins) Buccal cells were genotyped with Affymetrix Axiom Genome-Wide LAT and Exome arrays. SNP imputation was performed using data from the 1000 Genomes project. An intergenic hotspot in 15q13.3 fell just short of genome-wide significance in relation to ICV itself (rs8030297; p=5.17 x 10^{-8}, nearest gene *KLF13*).



CS6: GWAS Neonatal ROIs

Table. Loci exceeding conventional GWAS threshold for ICV-adjusted brain volumes

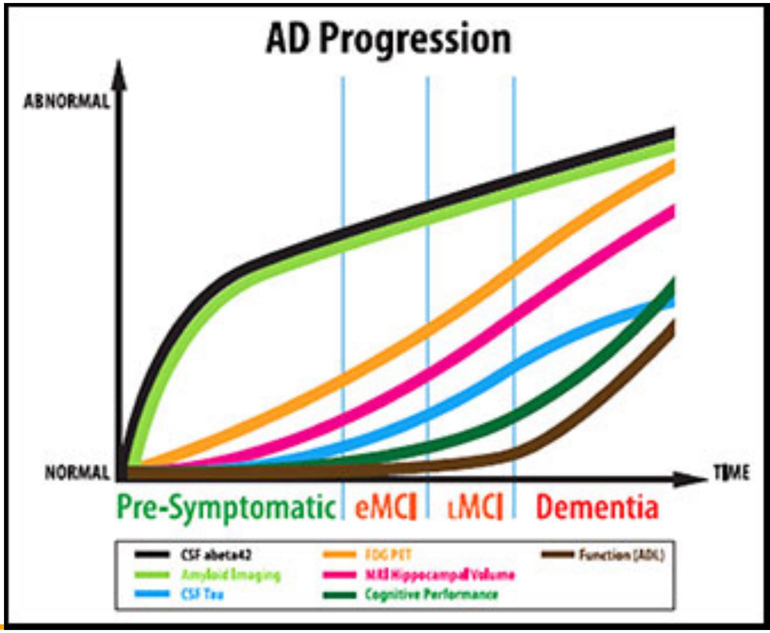
Tissue Volume	CHR	Best SNP	P-Value	Closest Gene*
WM	5	rs32892	3.95×10^{-9}	*MEF2C*
	17	rs78151819	2.33×10^{-8}	*c17orf112*
GM	4	rs114518130	1.59×10^{-9}	*IGFBP7*
	10	rs11012877	1.42×10^{-8}	*CACNB2*
	7	rs7786147	4.18×10^{-8}	*MPLKIP*
CSF	18	rs11875537	4.30×10^{-8}	*METTL4*
Cortical GM	NONE			
Cortical WM	5	rs76674566	7.65×10^{-10}	*DPYSL3*
	4	rs116957462	1.19×10^{-8}	*BANK1*
	14	rs80211808	3.86×10^{-8}	*CCDC88C*
	10	rs60689930	4.97×10^{-8}	*PPAPDC1A*

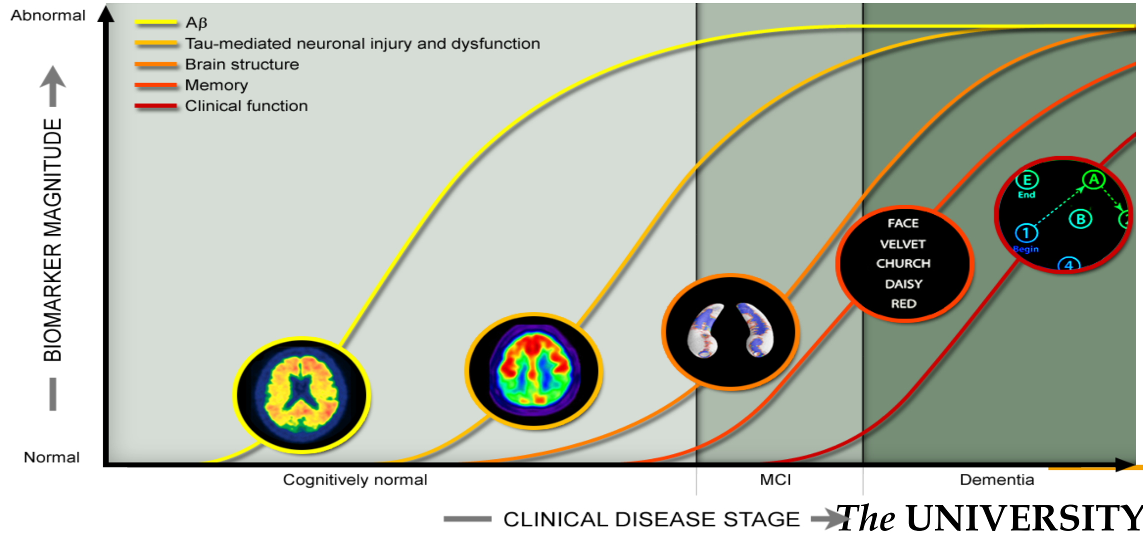
CS6: Imaging Genetics for ADNI

PI: Dr. Michael W. Weiner

- detecting AD at the earliest stage and marking its progress through biomarkers;
- developing new diagnostic methods for AD intervention, prevention, and treatment.

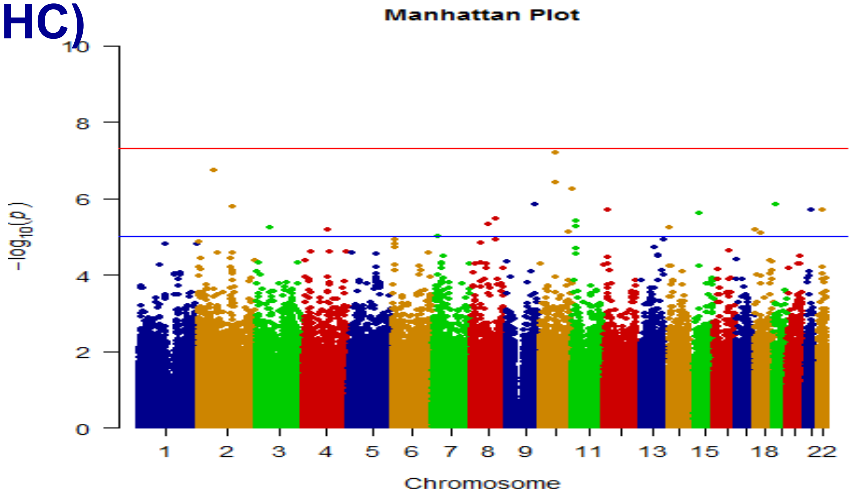
- A longitudinal prospective study with 1700 aged between 55 to 90 years
- Clinical Data including Clinical and Cognitive Assessments
- Genetic Data including Ilumina SNP genotyping and WGS
- MRI (fMRI, DTI, T1, T2)
- PET (PIB, Florbetapir PET and FDG-PET)
- Chemical Biomarker

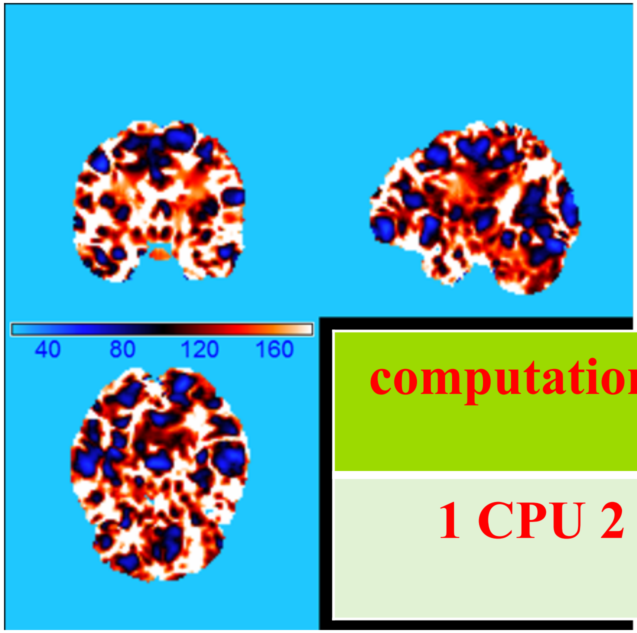




CS6: Fast Voxelwise Genome Wide Association analysiS

- 708 subjects (186 AD, 388 MCI, and 224 HC)
- 501,584 SNPs
- RAVEN Maps with 501,584 voxels

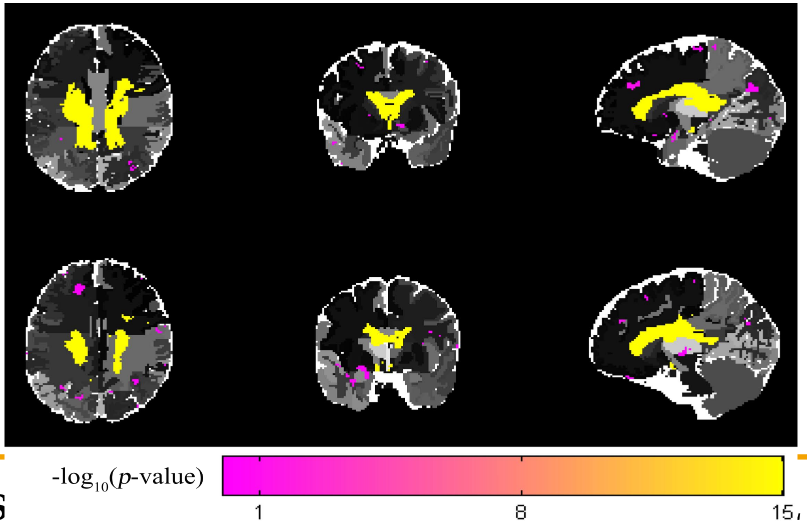




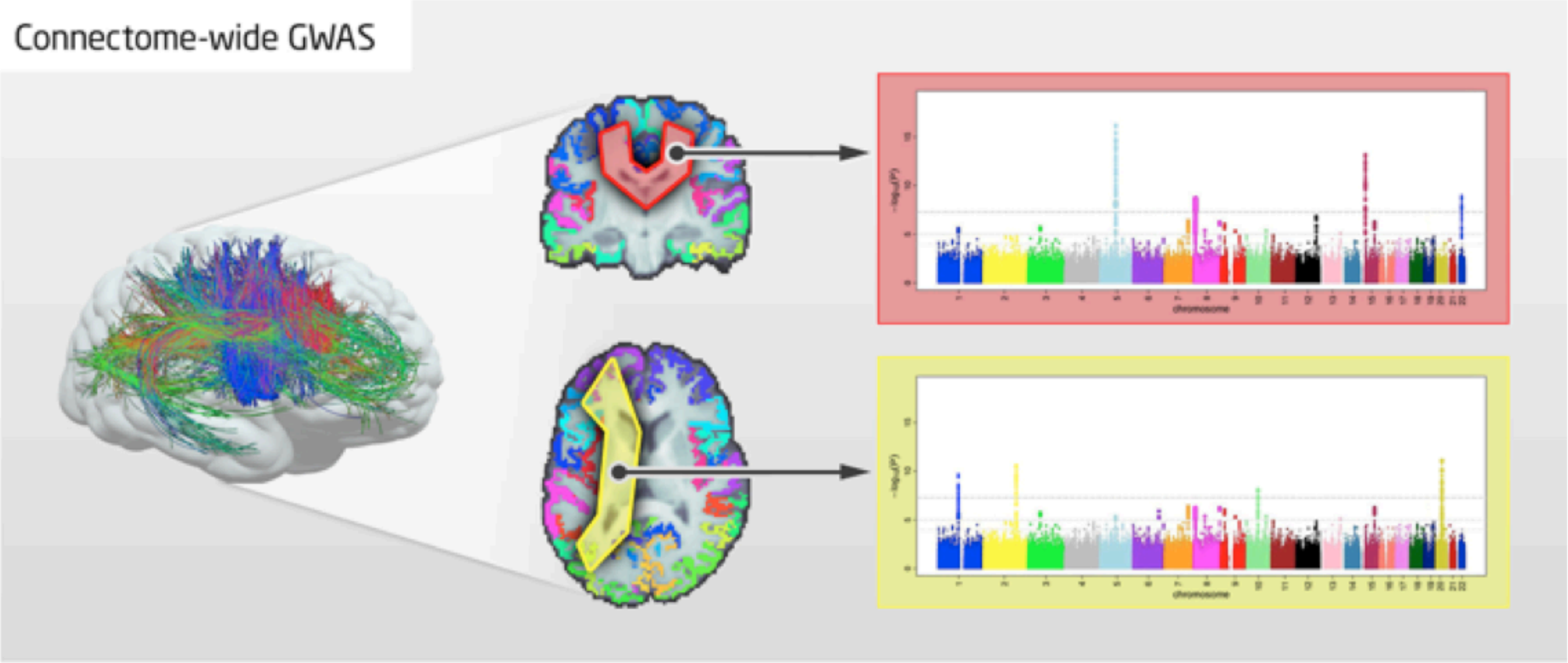
computational time

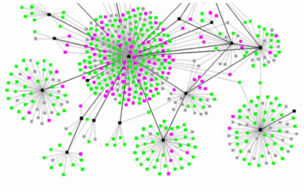
1 CPU 2 days

APOE



Connectome-Wide Genome-Wide Screen
Alzheimer risk gene

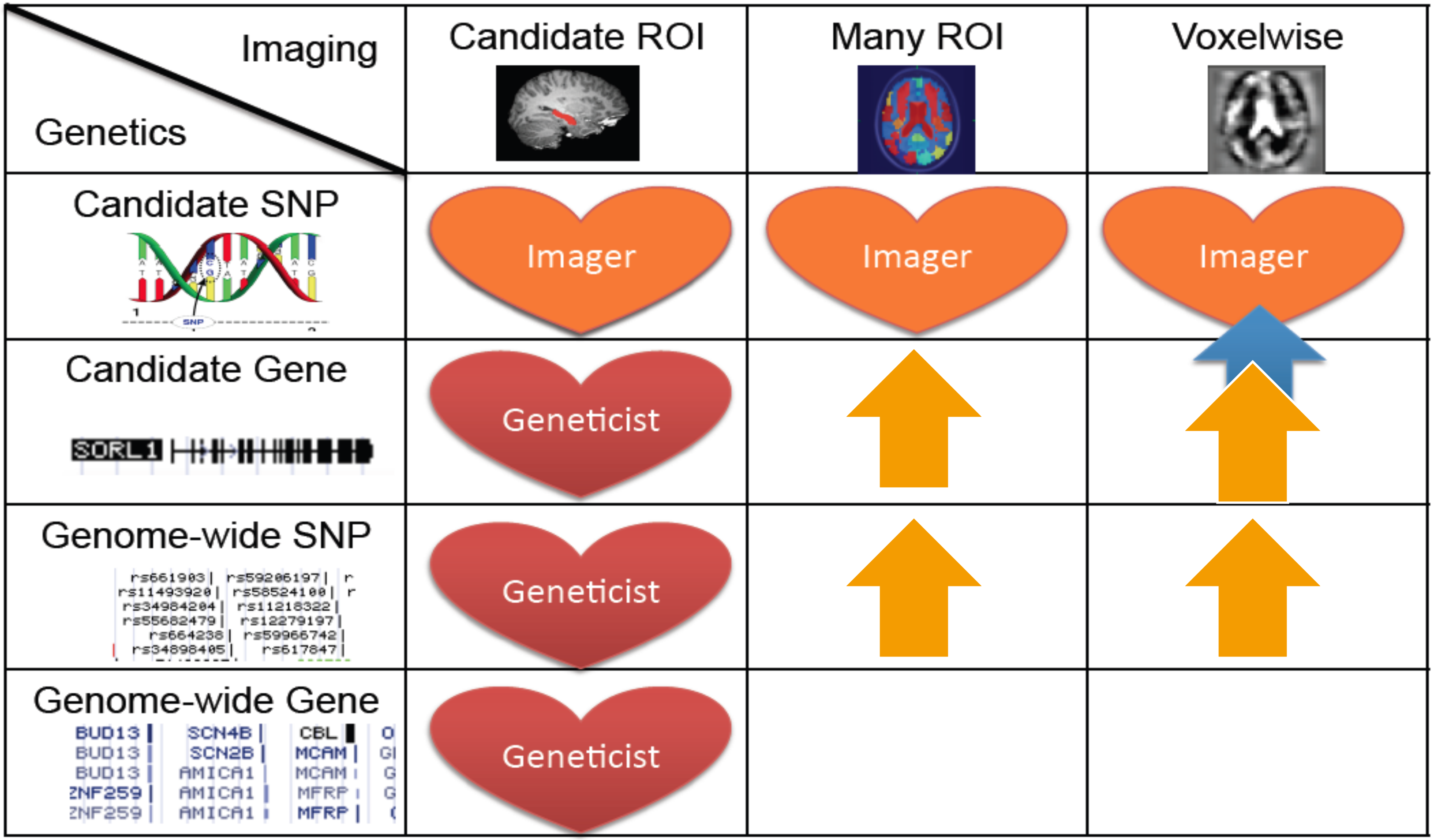




Discovery sample – Young Adults
Effect in ADNI
Within 2 weeks Sherva et al. published *SPON1*
Found in a cognitive GWAS in AD

Jahanshad et al., PNAS 2013

Statistical Methods

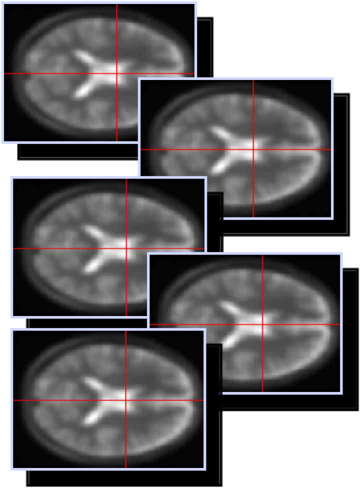


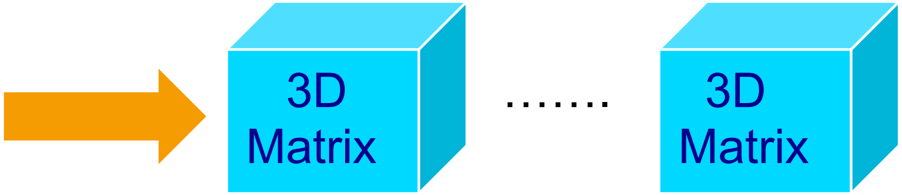
Imaging / Genetics	Candidate ROI	Many ROI	Voxelwise
Candidate SNP	Imager	Imager	Imager
Candidate Gene	Geneticist	↑	↑
Genome-wide SNP	Geneticist	↑	↑
Genome-wide Gene	Geneticist		

Hibar, et al. HBM 2012

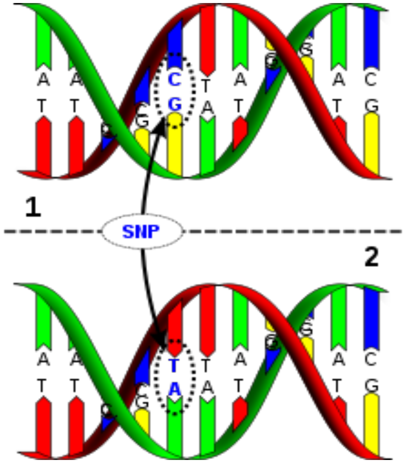
Data Structure

Imaging:





Genetic：



$$
\begin{array}{l|cccc|}
 & \text{SNP1} & \text{SNP2} & \text{.......} & \text{SNP} \\
\text{Person No. 1} & 1 & 2 & \cdots & 0 \\
\cdot & 0 & \ddots & & 1 \\
\cdot & \vdots & & \ddots & \vdots \\
\text{Person No. 100} & 1 & 0 & \cdots & 2
\end{array}
$$

Challenging Issues

$$y_i(\bullet) = f(x_i(\circ), B(\bullet,\circ)) \oplus \varepsilon_i(\bullet)$$

- Complicated domains (e.g., surface mesh, loci)
- Complicated objects (e.g., matrix response)
- Longitudinal and familial studies (e.g., heritability)
- Short-range to medium-to-long-range spatial/genetic correlations
- High-dimensional response and covariate
- Asymptotic theory (e.g., simultaneous confidence bound, minimax theory)

Big-Data Challenges

10^4

$X : p_x \times n$

10^6

10^6

$B : p_x \times p_y$

10^7

10^4

$Y : n \times p_y$

10^7

Memory:

$$O((p_x + p_y)n + p_x p_y)$$

Computational time: $O(p_x p_y n) = O(10^{17})$

A Heteroscedastic Linear Model

$$y_i(v) = \boldsymbol{x}_i^T \boldsymbol{\beta}(v) + \boldsymbol{z}_i(c)^T \boldsymbol{\gamma}(c,v) + e_i(v) \quad \text{for} \quad i = 1,\ldots,n$$

where $\boldsymbol{\beta}(v) = \left(\beta_1(v),\ldots,\beta_K(v)\right)^T$ is a $K\times 1$ vector associated with non-genetic predictors, and $\boldsymbol{\gamma}(c,v) = \left(\gamma_1(c,v),\ldots,\gamma_L(c,v)\right)^T$ is an $L\times 1$ vector of genetic fixed effects (e.g., additive or dominant).

Moreover, $e_i(v)$ are measurement errors with zero mean and $\boldsymbol{e}_i = \{e_i(v): v\in V\}$ are independent across *i*.

A Heteroscedastic Linear Model

We need to test:

$$H_0(c,v):\boldsymbol{\gamma}(c,v)=0 \quad \text{versus} \quad H_1(c,v):\boldsymbol{\gamma}(c,v)\neq 0 \quad \text{for} \quad \text{each} \quad (c,v)$$

We calculate a Wald-type statistic as:

$$W(c,v)=\tilde{\boldsymbol{\gamma}}(c,v)^T\left\{\mathrm{Cov}\left(\tilde{\boldsymbol{\gamma}}(c,v)\right)\right\}^{-1}\tilde{\boldsymbol{\gamma}}(c,v)$$

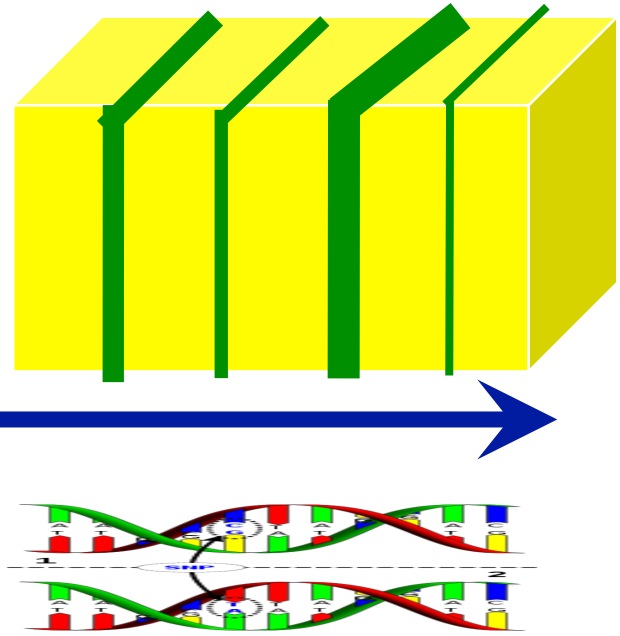
$$=\mathrm{tr}\left\{\left\{\boldsymbol{Z}_c^T(I_n-P_X)\boldsymbol{Z}_c\right\}^{-1}\boldsymbol{Z}_c^T(\boldsymbol{I}_n-P_X)\sigma_e^{-2}(c,v)\boldsymbol{Y}(v)\boldsymbol{Y}(v)^T(\boldsymbol{I}_n-P_X)\boldsymbol{Z}_c\right\}$$

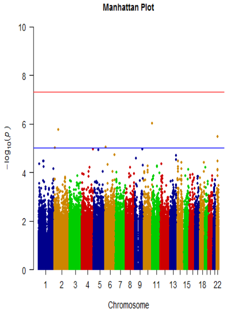
Fast Voxelwise Genome Wide Association analysiS

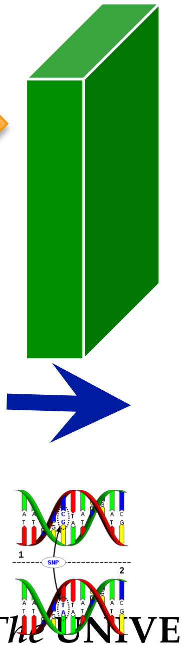
(I) Spatially Heteroscedastic Linear Model

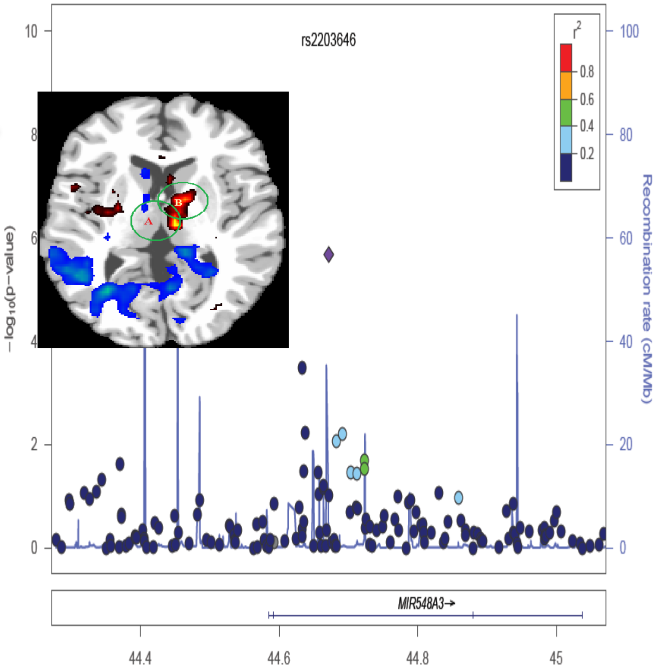
(II) Global Sure Independence Screening Procedure

(III) Detection Procedure

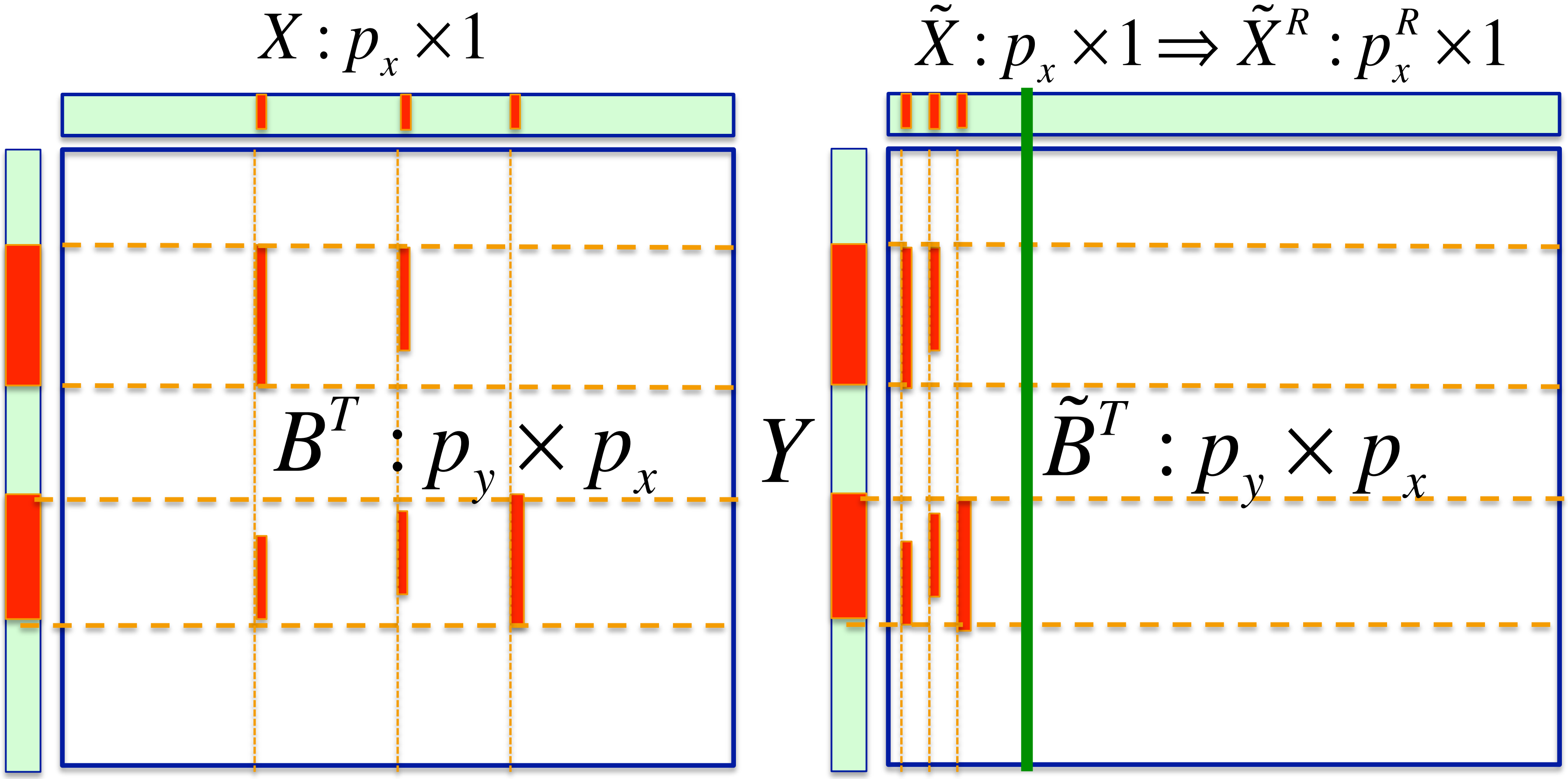






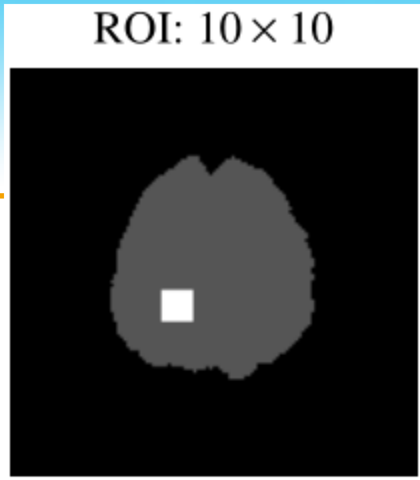


Key Features


$X : p_x \times 1$
$\tilde{X} : p_x \times 1 \Rightarrow \tilde{X}^R : p_x^R \times 1$
$B^T : p_y \times p_x$
Y
$\tilde{B}^T : p_y \times p_x$

X: Sparsity; Y|X: Clustered ROIs

Simulation Studies



Simulation settings: the dark, gray, and white regions in the figure, respectively, represent background, brain region, and the effected ROI associated with the causal SNPs.

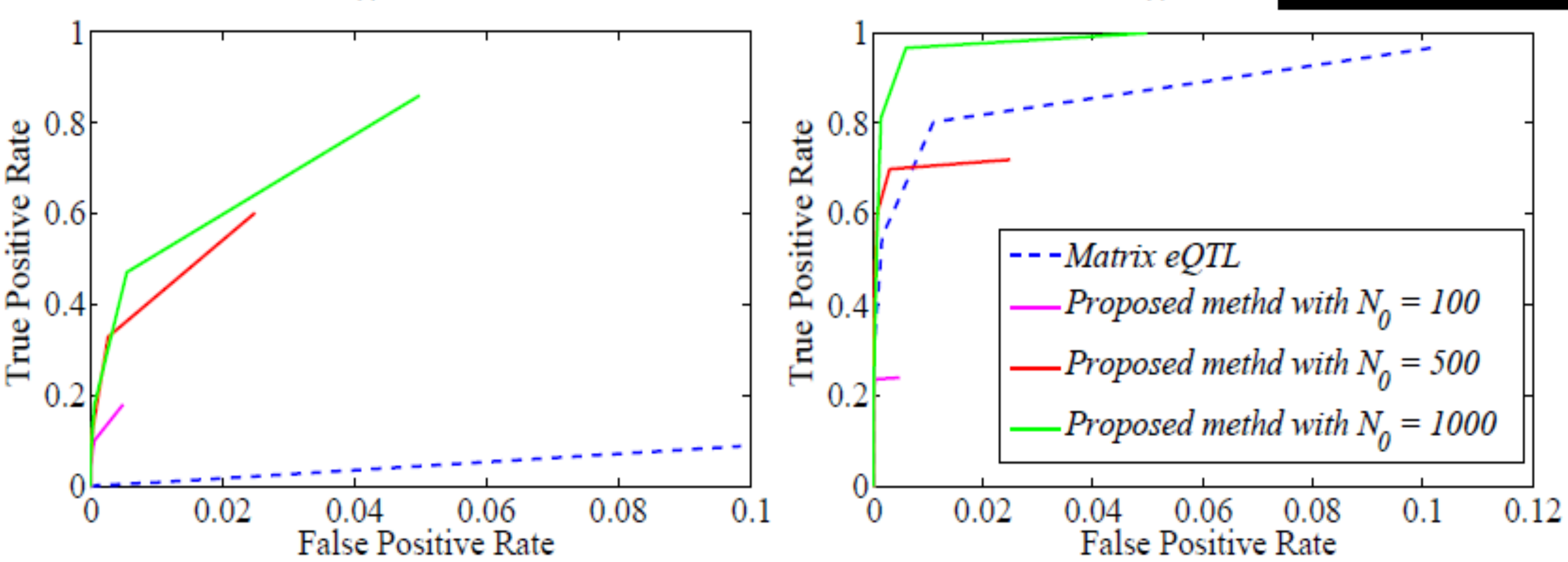
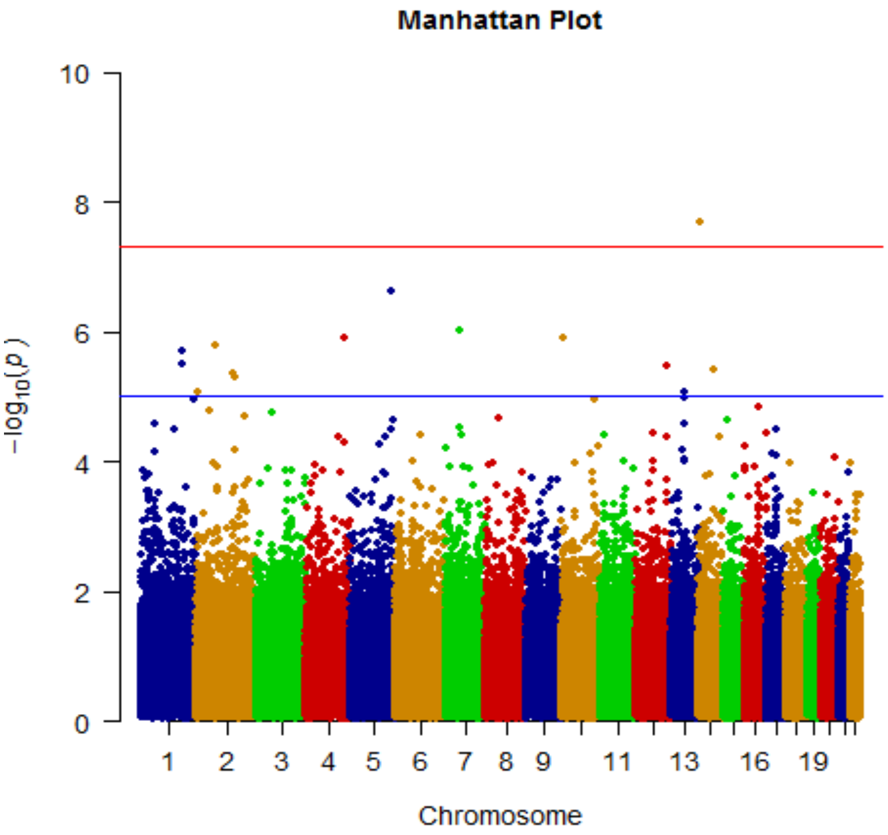
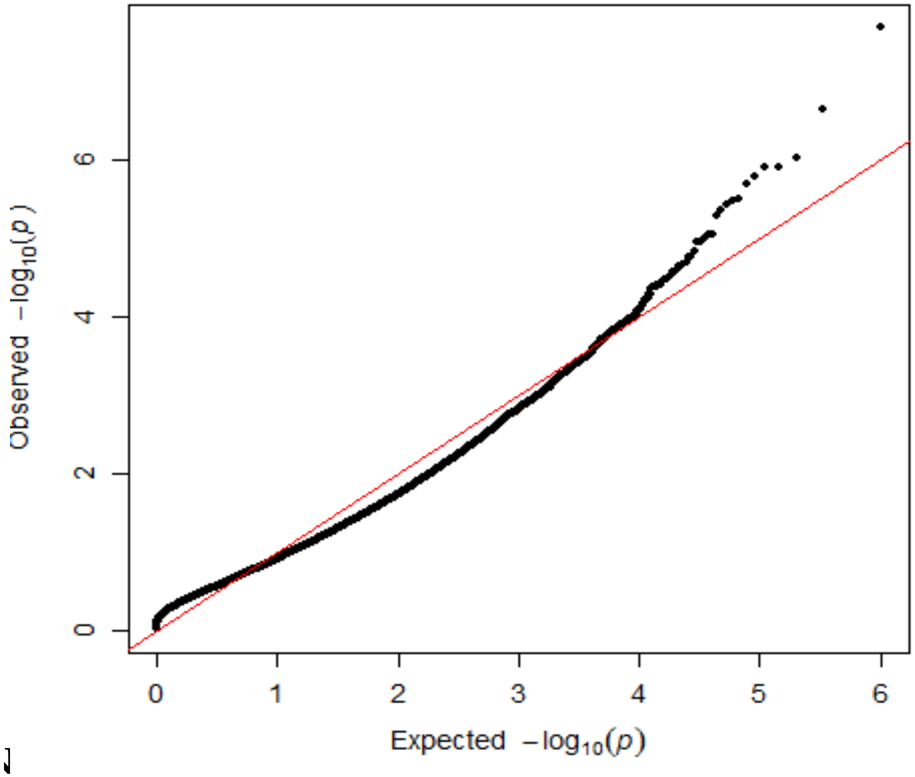


Fig. Simulation results for comparisons between FVGWAS and the Matrix eQTL in identifying significant voxel-SNP pairs.

Results

Our computational time
About 33,800 s

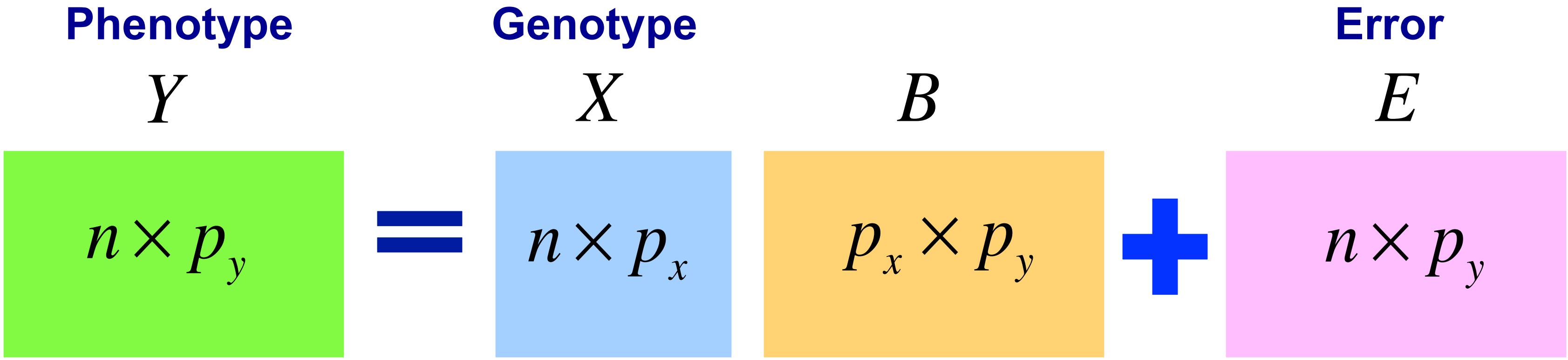




High Dimensional Regression Model

Data $\{(Y_i, X_i) : i = 1, \cdots, n\}$

$$Y_i = \{y_i(v) : v \in V_0\} \qquad X_i = \{X_i(g) : g \in G_0\}$$



$$\underset{n \times p_y}{Y} = \underset{n \times p_x}{X}\ \underset{p_x \times p_y}{B} + \underset{n \times p_y}{E}$$

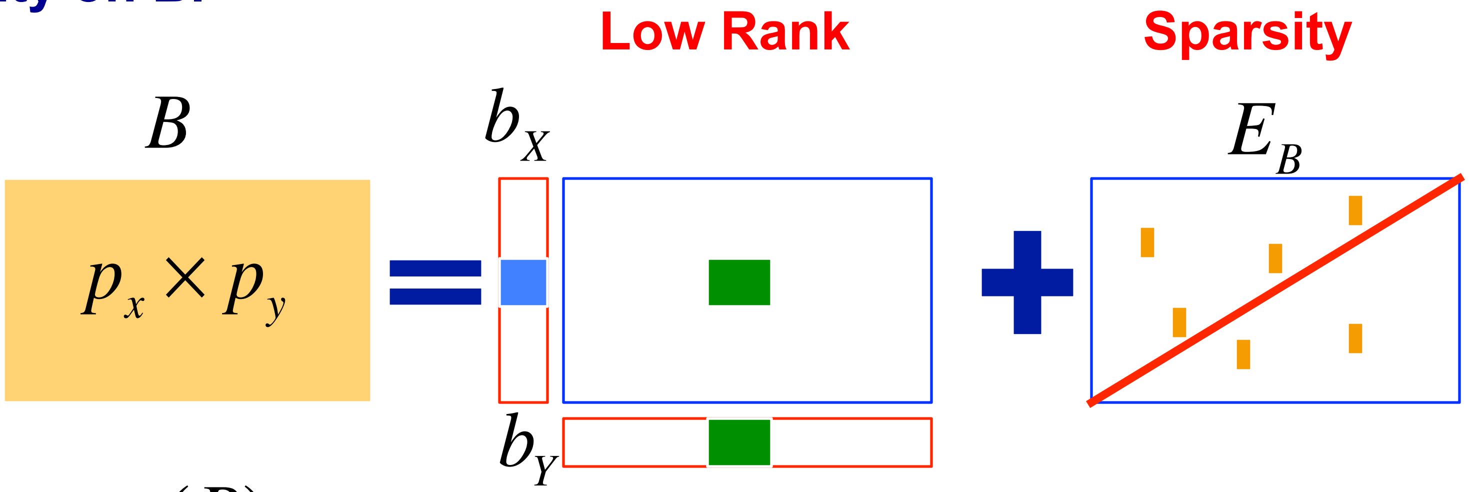
Key Conditions:

$$\max(p_x, p_y) \sim n$$

- Sparsity of B
- Restricted null-space property for design matrix X

Sparse and Low-rank Representation

Sparsity on B.



Low Rank

Sparsity

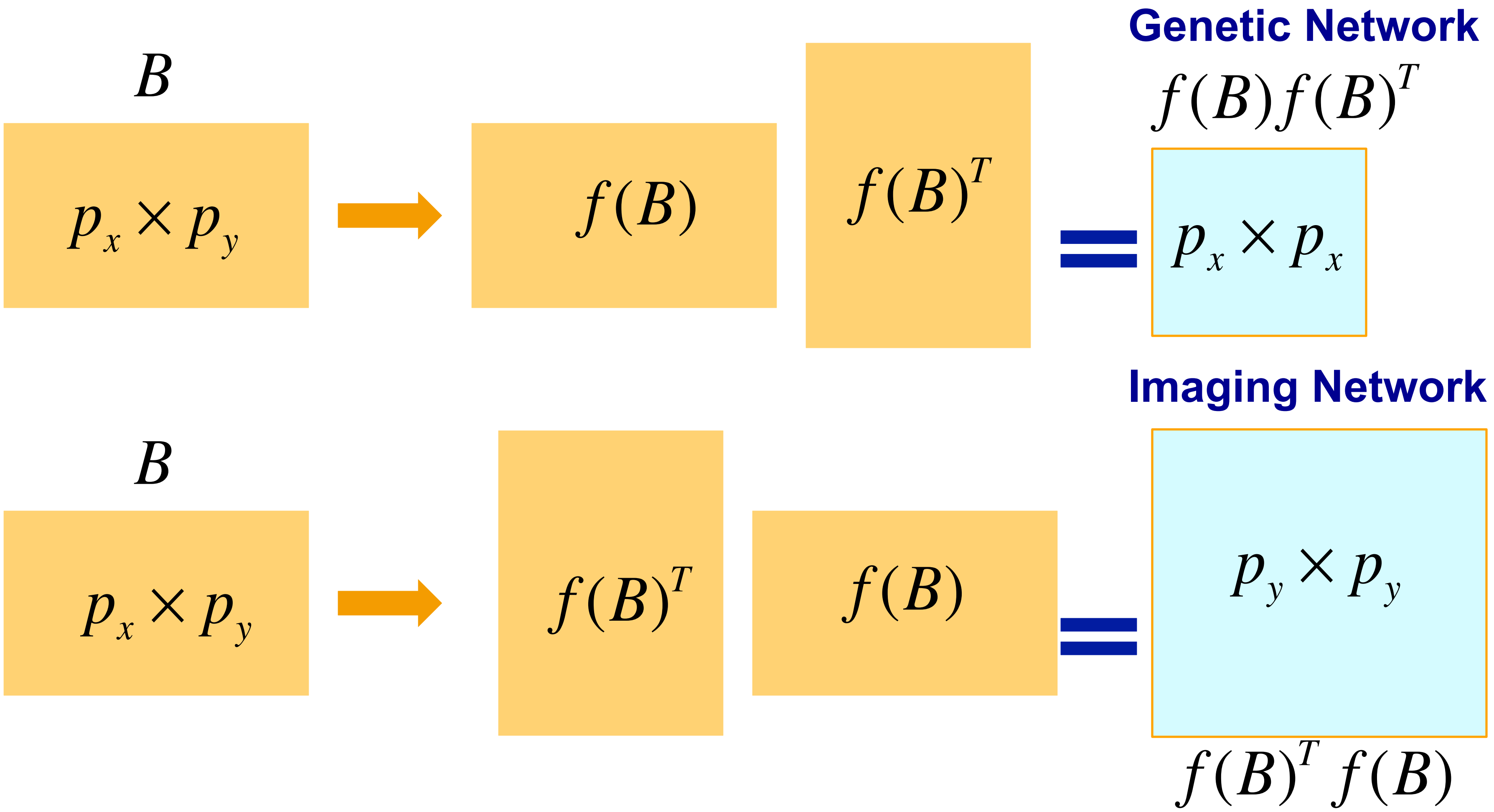
$$p_\lambda(B) \rightarrow p_\lambda(b_X) + p_\lambda(b_Y) + p_\lambda(E_B)$$

Regularization Methods

- Lasso 1, 2, 3, ….
- SCAD, MCP, …..

$$\hat{\theta} \in \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda_n \sum_{j=1}^{p} |\theta_j|$$

Genetic and Imaging Networks



Genetic Network

$f(B)f(B)^T$

B

$p_x \times p_y$ → $f(B)$ $f(B)^T$ = $p_x \times p_x$

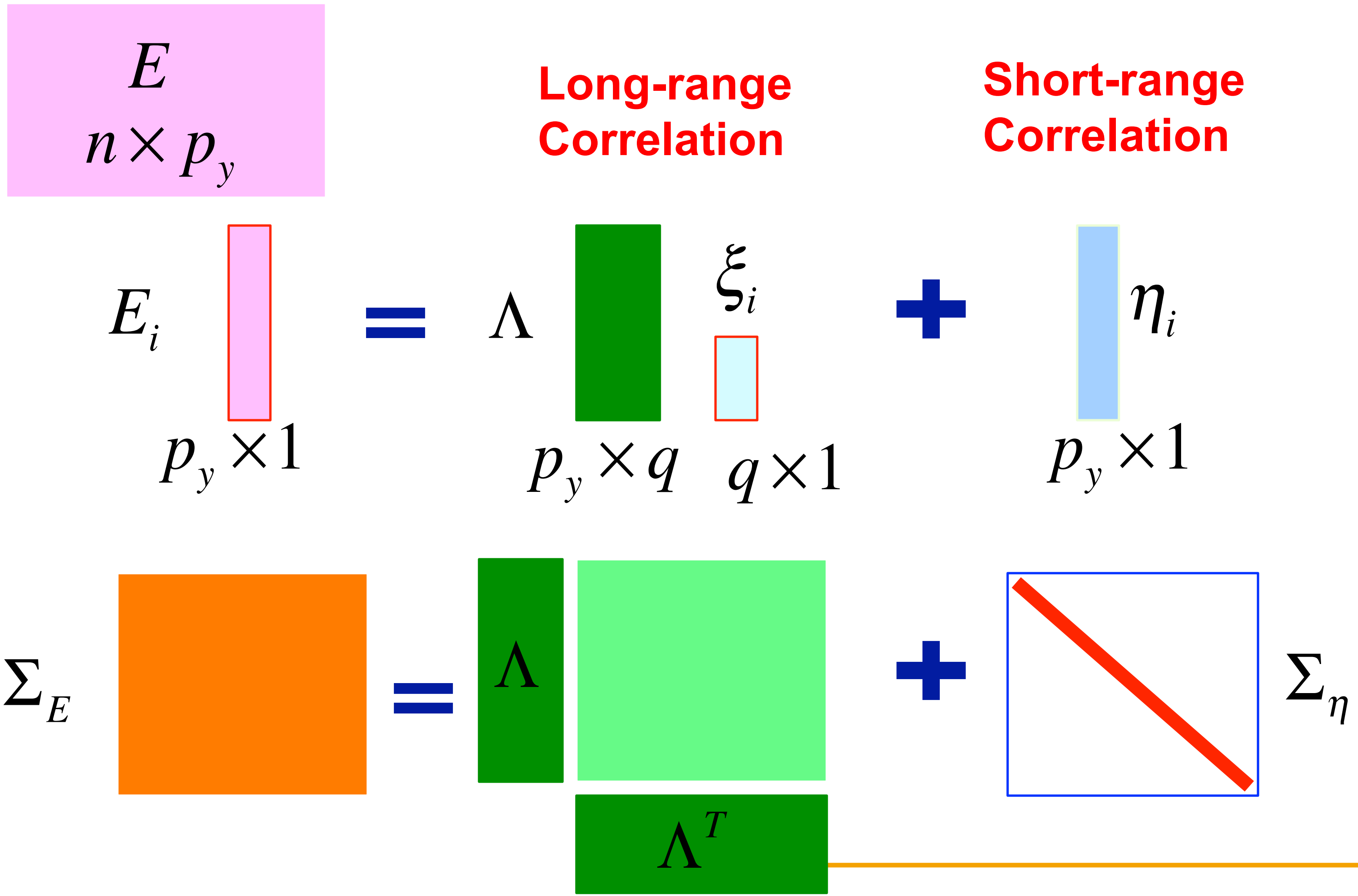
Imaging Network

B

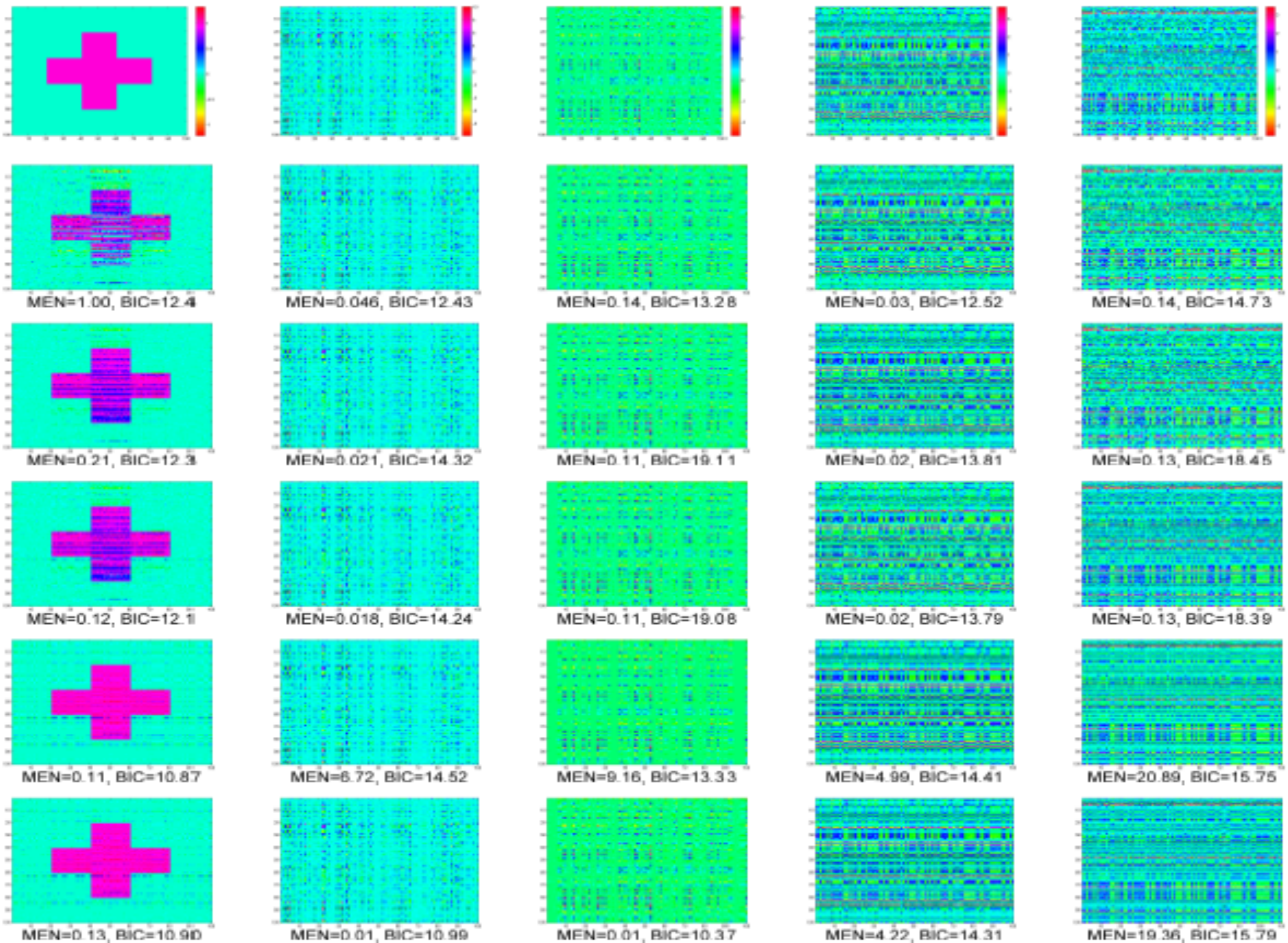
$p_x \times p_y$ → $f(B)^T$ $f(B)$ = $p_y \times p_y$

$f(B)^T f(B)$

Factor Model



Simulation

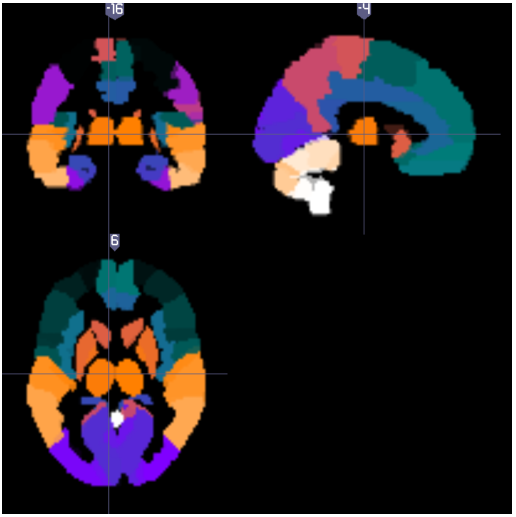


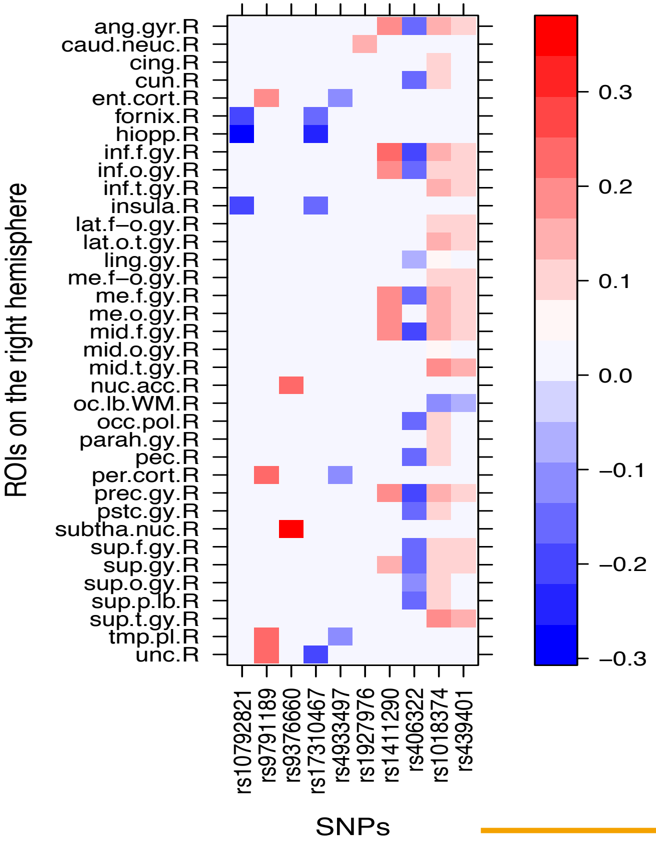
Patterns	Plus	SVD	SVD	UN	UN
True B					
LASSO	MEN=1.00, BIC=12.4	MEN=0.046, BIC=12.43	MEN=0.14, BIC=13.28	MEN=0.03, BIC=12.52	MEN=0.14, BIC=14.73
BLASSO	MEN=0.21, BIC=12.3	MEN=0.021, BIC=14.32	MEN=0.11, BIC=19.11	MEN=0.02, BIC=13.81	MEN=0.13, BIC=18.45
G-SMuRFS	MEN=0.12, BIC=12.1	MEN=0.018, BIC=14.24	MEN=0.11, BIC=19.08	MEN=0.02, BIC=13.79	MEN=0.13, BIC=18.39
GLRR3	MEN=0.11, BIC=10.87	MEN=6.72, BIC=14.52	MEN=9.16, BIC=13.33	MEN=4.99, BIC=14.41	MEN=20.89, BIC=15.75
GLRR5	MEN=0.13, BIC=10.90	MEN=0.01, BIC=10.99	MEN=0.01, BIC=10.37	MEN=4.22, BIC=14.31	MEN=19.36, BIC=15.79

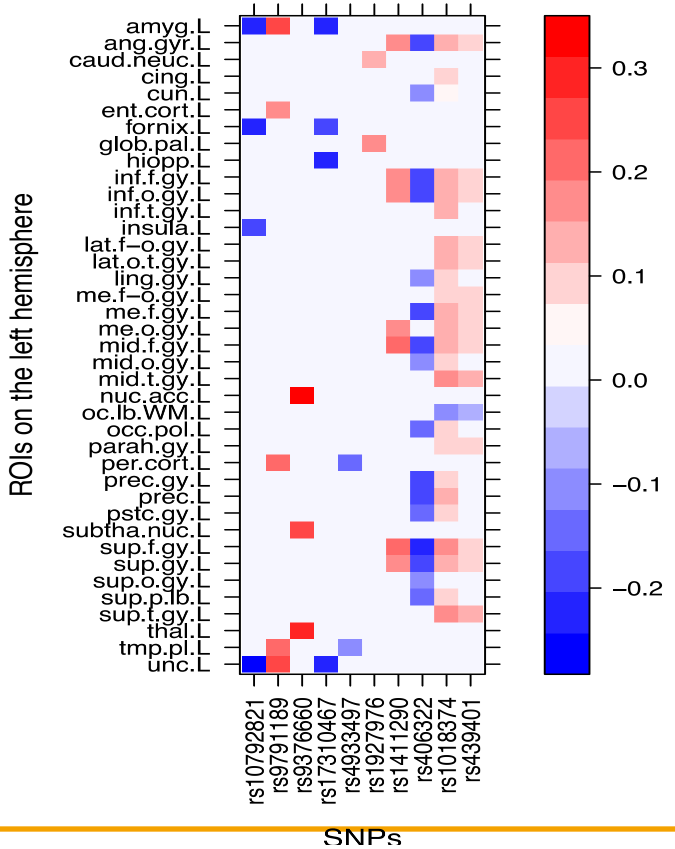
ADNI

749 AD/MCI/NC subjects, 93 ROIs

40 AD candidate genes on the AlzGene web

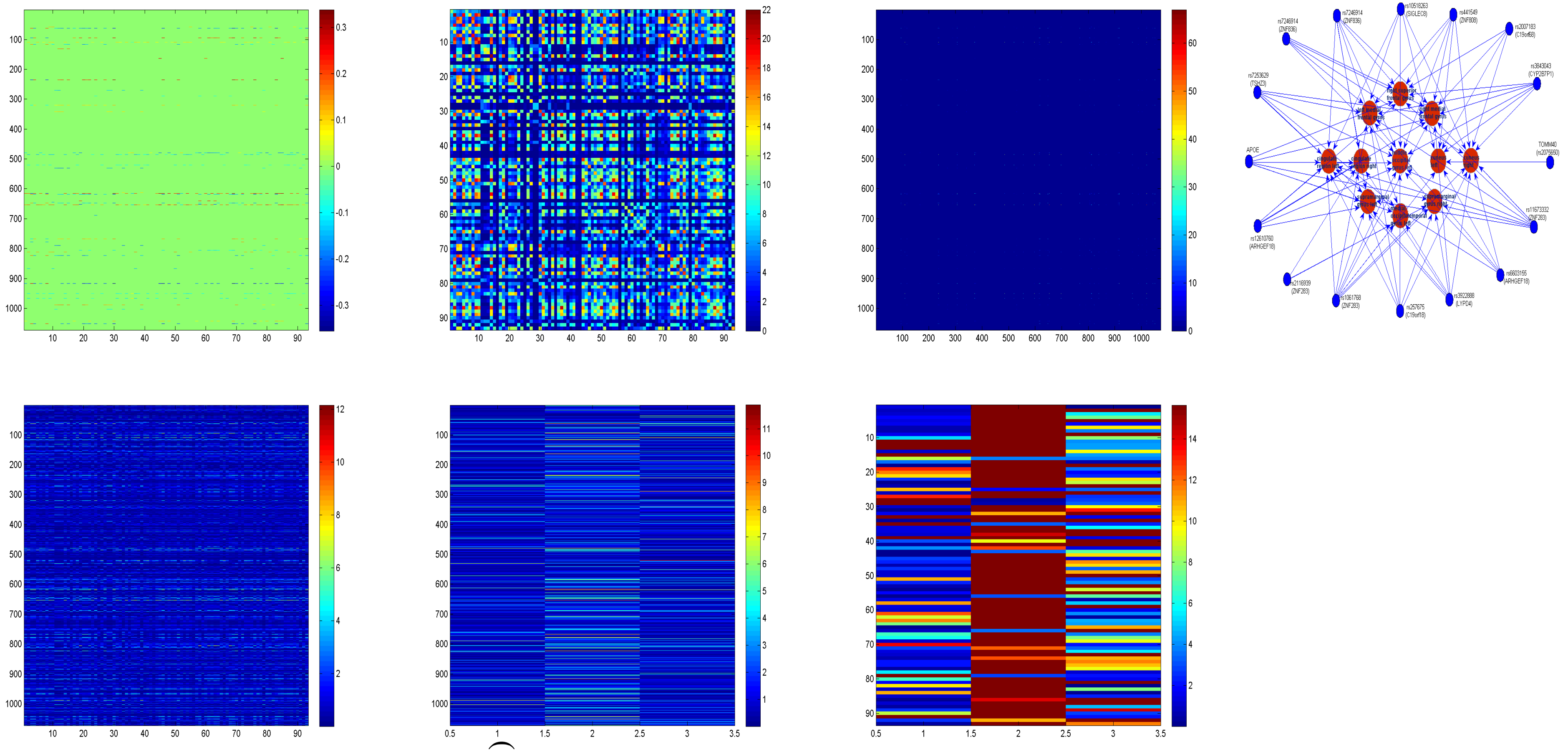






ADNI

$\widehat{B}$ | **ROI network** | **Genetic network**



$-\log_{10}(p)$ for $\widehat{B}$

Sparse Projection Regression Model

- Multivariate regression with a high-dimensional responses and a multivariate covariate of interest
- Consider a Multivariate Linear Model (MLM):

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \quad or \;\; \mathbf{y}_i = \mathbf{B}^T\mathbf{x}_i + \mathbf{e}_i$$

- We are interested in the hypothesis testing problem:

$$H_0 : \mathbf{C}\mathbf{B} = \mathbf{B}_0 \quad v.s. \quad H_1 : \mathbf{C}\mathbf{B} \neq \mathbf{B}_0$$

- Diverging q , fixed p case
  - High-dimension two sample test
  - Imaging genetics association study

Sparse Projection Regression Model

- Let $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_k]$, then a projection regression model is given by:

$$\mathbf{W}^T y_i = (\mathbf{BW})^T \mathbf{x}_i + \mathbf{W}^T \mathbf{e}_i = \boldsymbol{\beta}_{\mathbf{w}}^T \mathbf{x}_i + \varepsilon_i$$

- Hypothesis problem reduces to:

$$H_{0W} : \mathbf{C}\boldsymbol{\beta}_{\mathbf{w}} = \mathbf{b}_0 \quad v.s. \quad H_{1W} : \mathbf{C}\boldsymbol{\beta}_{\mathbf{w}} \neq \mathbf{b}_0$$

$$where\ \mathbf{C}\boldsymbol{\beta}_{\mathbf{w}} = \mathbf{CBW}\ and\ \mathbf{b}_0 = \mathbf{B}_0\mathbf{W}$$

- How to determine an 'optimal' $\mathbf{W}$?

Sparse Projection Regression Model

- We show that this is achieved by optimizing the following generalized heritability ratio (GHR):

$$\mathrm{GHR}(\mathbf{w};\mathbf{C}) = \frac{\mathbf{w}^T(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)^T S_{\tilde{X}_1}(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}} = \frac{\mathbf{w}^T\Sigma_C\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}}$$

- High Dimensional Setting
- noise accumulation
  - ill-conditioned sample covariance estimator: $\hat{\Sigma}_R$
- Sparse Projection Regression Model is proposed as following:

$$\mathrm{argmax}\{\frac{\mathbf{w}^T\hat{\Sigma}_C\mathbf{w}}{\mathbf{w}^T\tilde{\Sigma}_R\mathbf{w}}\} \quad \text{s.t.} \quad ||\mathbf{w}||_1 \leq t$$

Sparse Projection Regression Model

- After estimating $\mathbf{W}$, we can calculate a $k \times k$ matrix as:

$$T_n = (\mathbf{C}\hat{\beta}_{\mathbf{w}} - \mathbf{b}_0)^T \Sigma_{\tilde{\Omega}}^{-1} (\mathbf{C}\hat{\beta}_{\mathbf{w}} - \mathbf{b}_0)$$

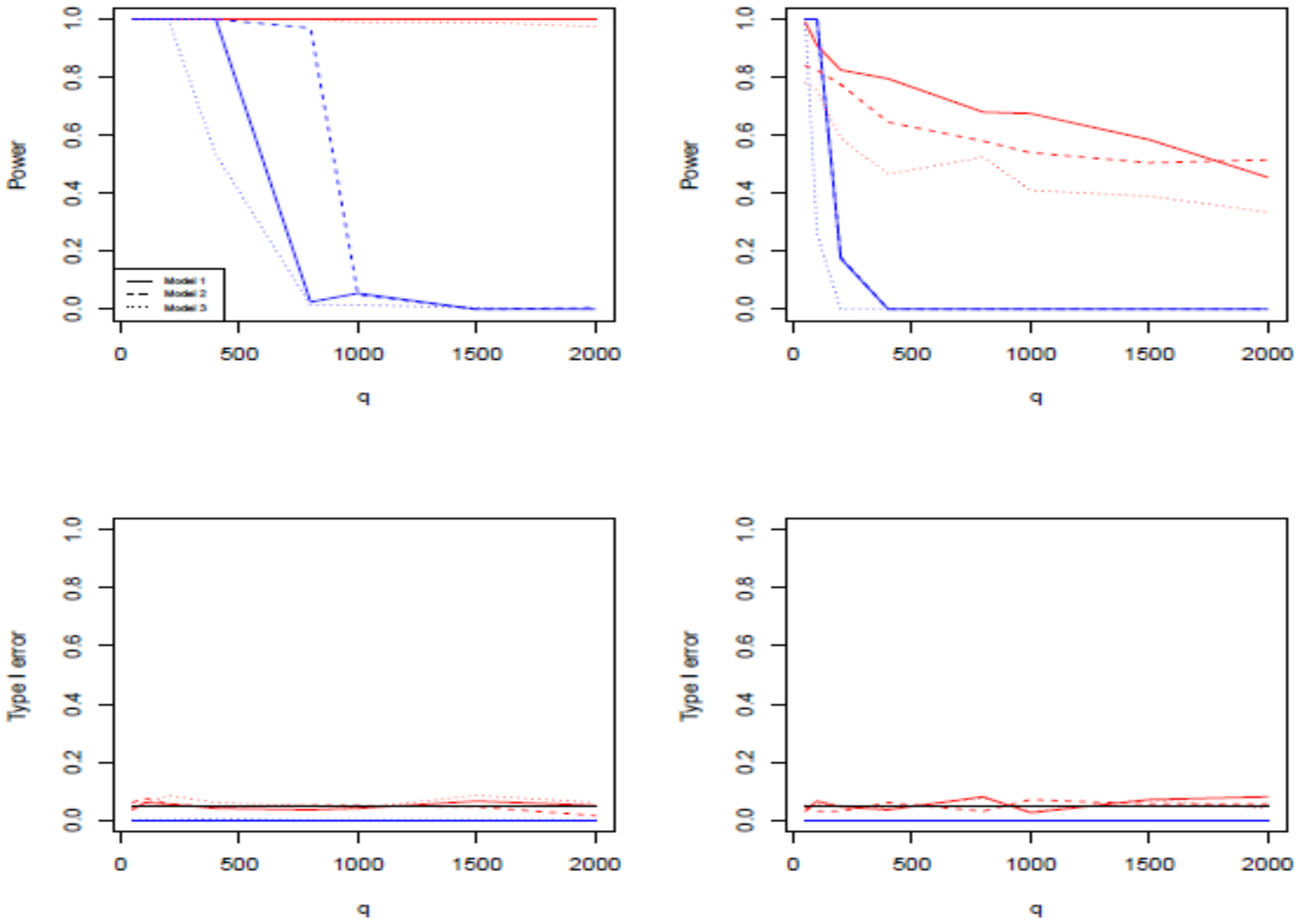
- Test statisitic: $\text{Tr}_n = \text{trace}(T_n)$
- Wild bootstrap
  - Fit MLM under the null hypothesis to calculate the estimated multivariate regression coefficient, denoted by $\widehat{\mathbf{B}}_{\mathbf{0}}$, residuals $\hat{\mathbf{e}}_i = \mathbf{y}_i - \widehat{\mathbf{B}}_0^T \mathbf{x}_i$.
  - Generate G bootstrap samples $\mathbf{z}_i^{(g)} = (\widehat{\mathbf{B}}_0)^T \mathbf{x}_i + \eta_i^{(g)} \hat{\mathbf{e}}_i$.
  - Repeat the estimation procedure for estimating the optimal weights and the calculation of the test statistic $\text{Tr}_n^{(g)}$.
  - p-value of Tr_n is computed as $\frac{1}{G} \sum_{g=1}^{G} \mathbf{1}(\text{Tr}_n^{(g)} \geq \text{Tr}_n)$.

Simulation

Numerical Example: High Dimensional Two Sample Test

- $\{\mathbf{y}_1, \ldots, \mathbf{y}_{n_1}\}$ and $\{\mathbf{y}_{n_1+1}, \ldots, \mathbf{y}_n\} \subset R^q$ from $N(\beta_1, \Sigma_R)$ and $N(\beta_2, \Sigma_R)$, respectively.
- We set: $n = 2n_1 = 100$ and q is 50, 100, 200, 400, 800, 1000, 1500, and 2000, respectively.
- $H_0 : \beta_1 = \beta_2$ against $H_1 : \beta_1 \neq \beta_2$
- Can be formulated by a regression model with $\mathbf{B}^T = [\beta_1, \beta_2]$ and $\mathbf{C} = (1, -1)$.
- Error covariance matrix $\Sigma_R = \sigma^2(\rho_{j,j'})$:
  - Model 1: is an independent covariance matrix with $(\rho_{jj'}) = \text{diag}(1, \cdots, 1)$.
  - Model 2: is a weak correlation matrix with $\rho_{jj'} = \mathbf{1}(j' = j) + 0.3 \times \mathbf{1}(j' \neq j)$.
  - Model 3: is a strong correlation covariance matrix with $\rho_{jj'} = 0.8^{|j'-j|}$.

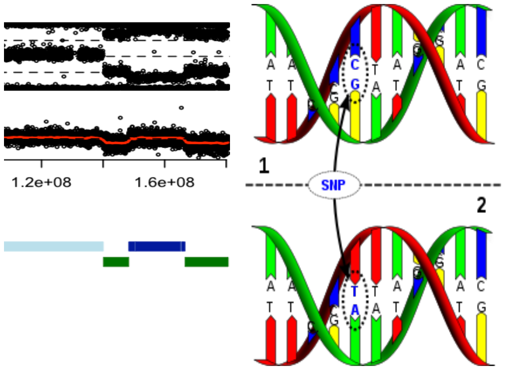
Simulation


Power
Power
Type I error
Type I error
q
q
q
q
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Model 1
Model 2
Model 3

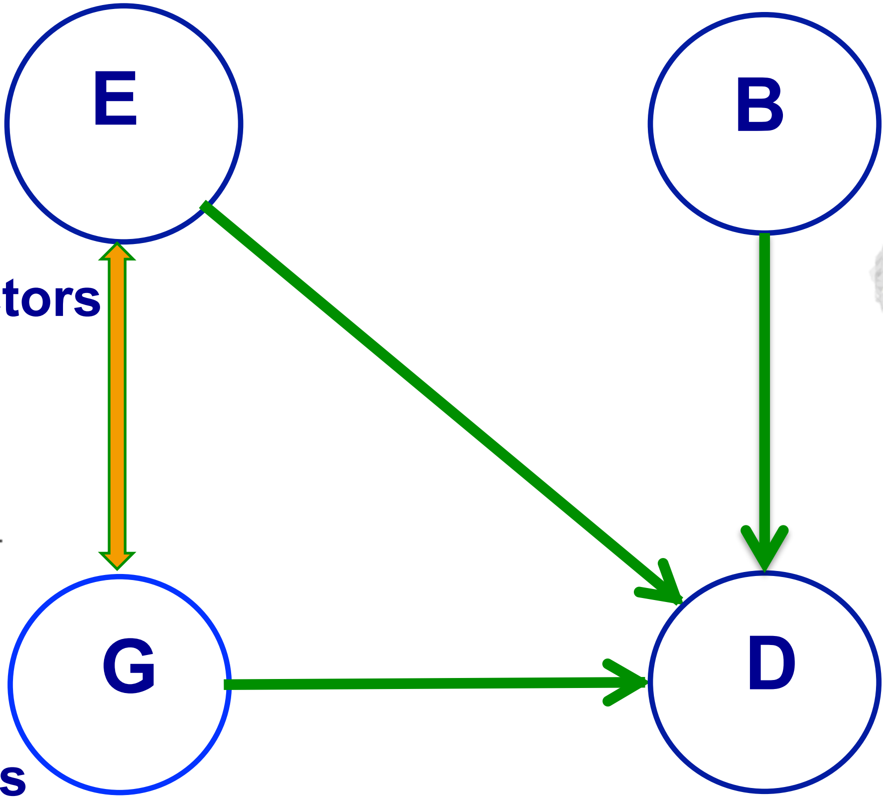
Predictive Models

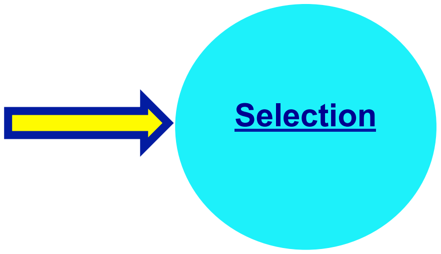
Big Data Integration

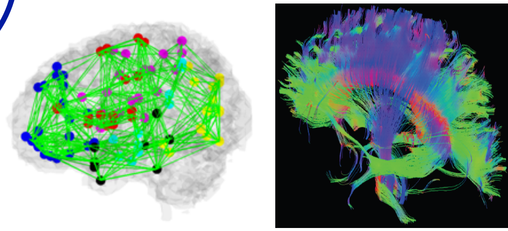
E: environmental factors



G: genetic markers







D: disease

http://en.wikipedia.org/wiki/DNA_sequence

References

1. D. Kong, J. G. Ibrahim, E. Lee and H. Zhu (2015). FLCRM: Functional Linear Cox Regression Model. In submission.
2. Yang, H., Zhu, H.T., and Ibrahim, J. G. (2015). SILFM: Single Index Latent Factor Model Based on High-dimensional Features. In submission.
3. Miranda, M., Zhu, H.T., and Ibrahim, J. G. (2015). TPRM: Tensor partition regression models with applications in imaging biomarker detection. In submission.
4. Shen, D. and Zhu, H.T. (2015). MWPCR: Multiscale weighted PCR for high-dimensional prediction. *Information Processing in Medical Imaging 2015.*
5. D. Kong, K. S. Giovanello, Y.L. Wang, W.L. Lin, E. Lee, Yong Fan, M. Doraiswamy, and *H.T. Zhu* and ADNI. (2015). Predicting Alzheimer's disease using combined imaging-whole genome SNP data. *Journal of Alzheimer's Disease.* In press.
6. Zhang, C., Liu, Y.F., Wang, J. H., and Zhu, H.T. (2015). Reinforced Angle-based Multicategory Support Vector Machines. *Journal of Computational and Graphical Statistics.* In press.
7. Lee, S., Zhu, H. T., Kong, D., Wang, Y., Giovanello, K. S., and Ibrahim, J. G. (2015). A Bayesian functional linear Cox regression model for predicting time to conversion to Alzheimer's disease. *Annals of Applied Statistics,* Under revision.
8. Wang, X. and Zhu, H.T. (2015) Generalized Scalar-to-Image Regression Models via Total Variation. *Journal of American Statistical Association.* Under revision.
9. Guo, R.X., Ahye M., and Zhu, H. (2015). Spatially weighted PCA for imaging classification. *Journal of Computational and Graphical Statistics.* 24, 274-296 .
10. Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in Neuorimaging data analysis. *Journal of American Statistical Association.* 108(502), 540-552.

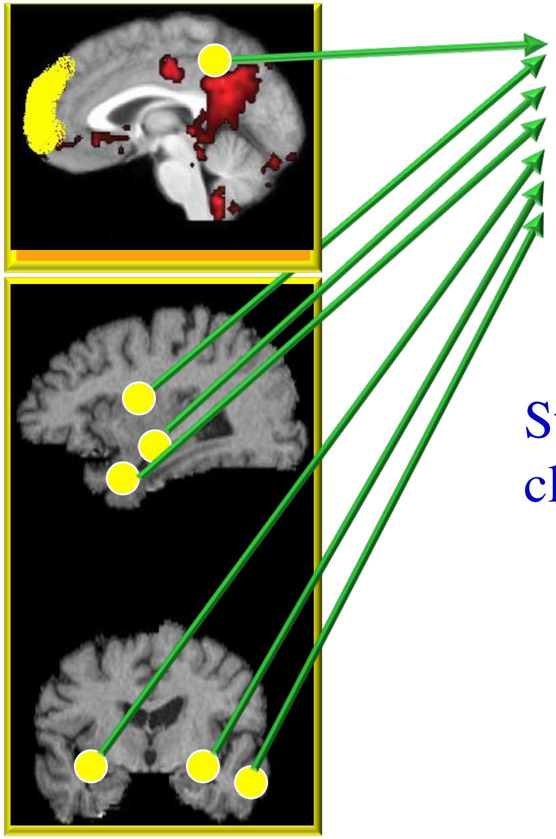
Predictive Modeling

Predictive models can either be used directly to estimate a response (output) given a defined set of features (input), or indirectly to drive the choice of decision rules.

- **Determining the ‘correct’ features**
- **Fitting the predictive model**
- **Performance assessment**

CS8: Pattern classification of neuroimages

Functional information

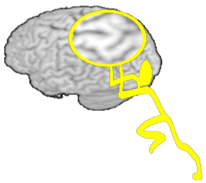


Morphological information

Pattern Classification → Quantitative Diagnosis

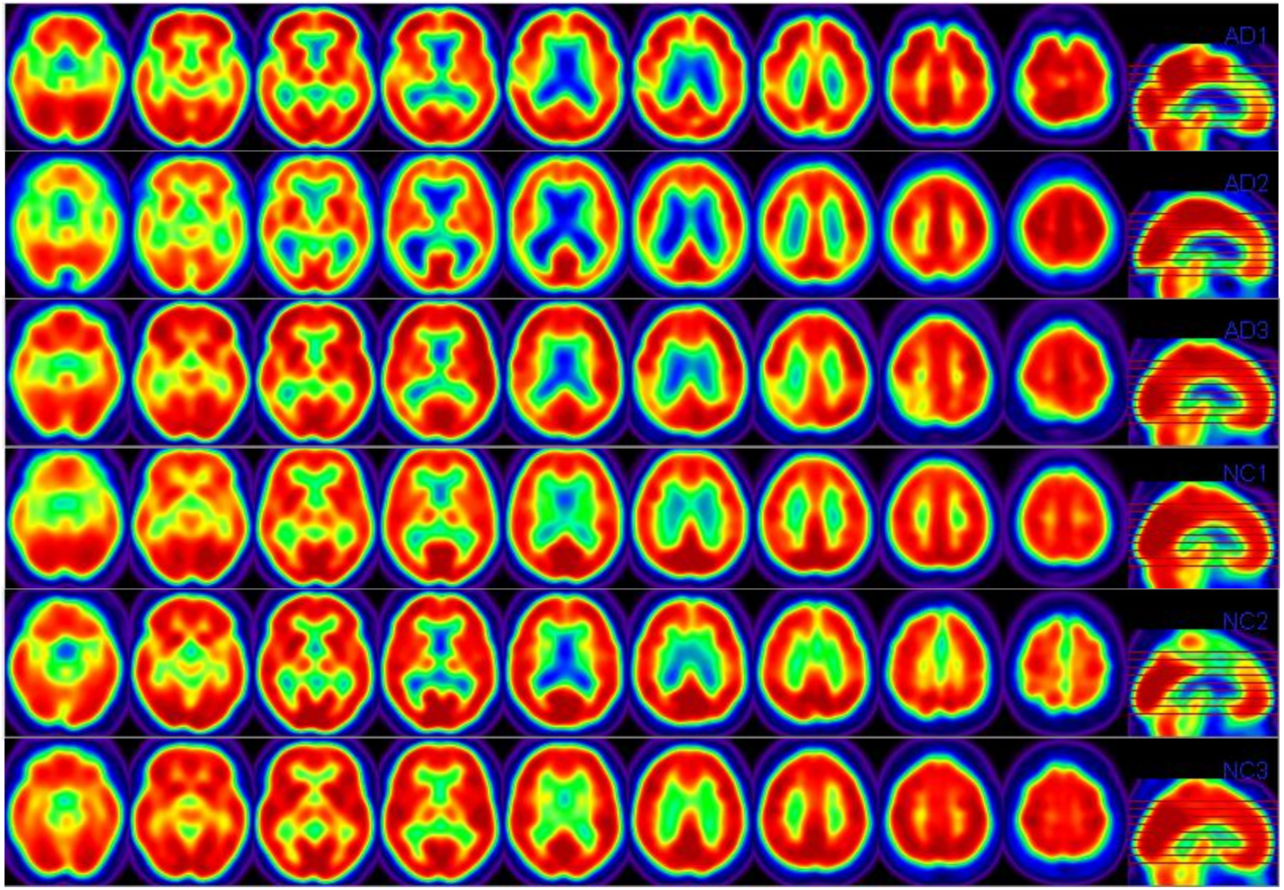
Structural, functional, and multimodality image classification

- Diagnosis of Schizophrenia
- Diagnosis of Alzheimer’s disease (AD)
- Clinical outcomes



ADNI

PET

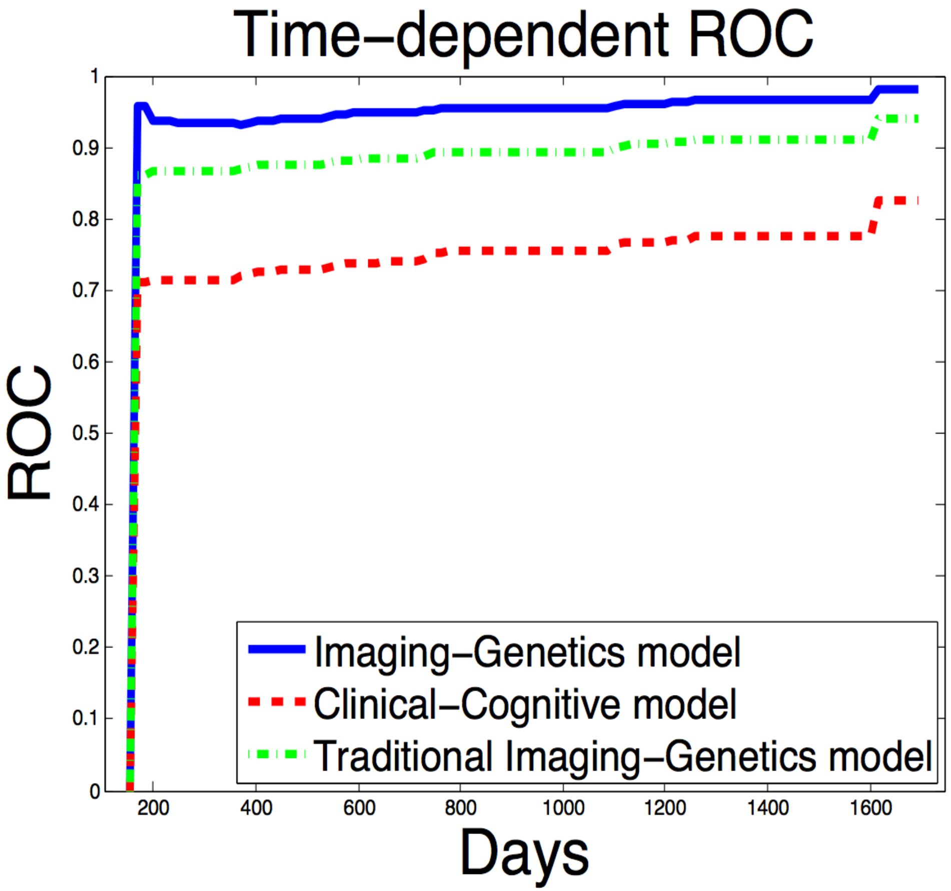


CS9: Predicting Conversion Time MCI-AD

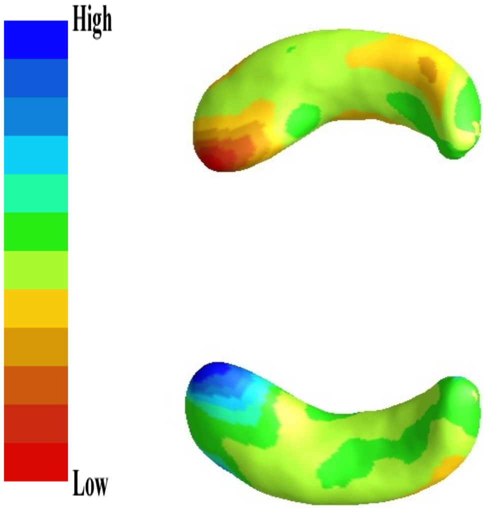
343 MCI patients were then followed over 48 months, with 150 participants progressing to AD.

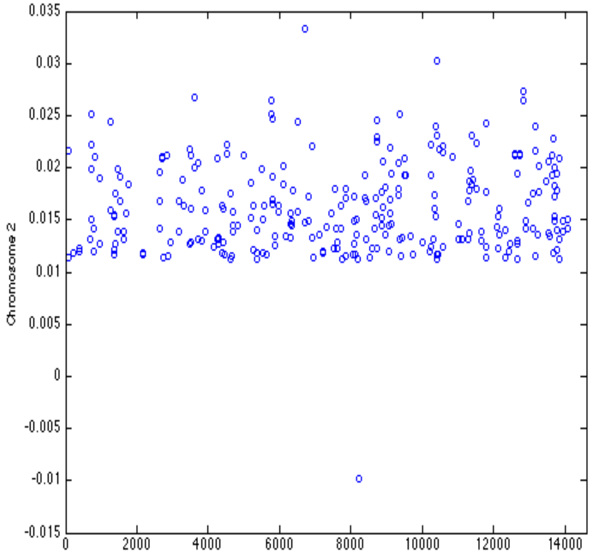
We extracted high dimensional MR imaging (volumetric data on 93 brain regions plus a hippocampal surface data), and whole genome data (504,095 SNPs from GWAS), as well as routine neurocognitive and clinical data at baseline.

Conversion time from MCI to AD.

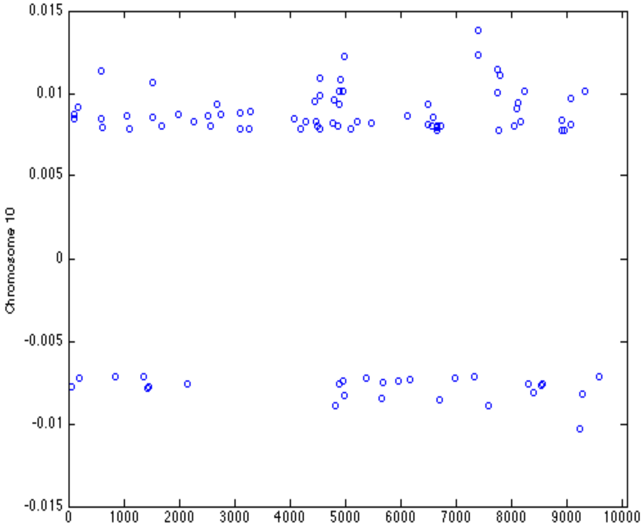


CS9: Predicting MCI-AD





Ch 2

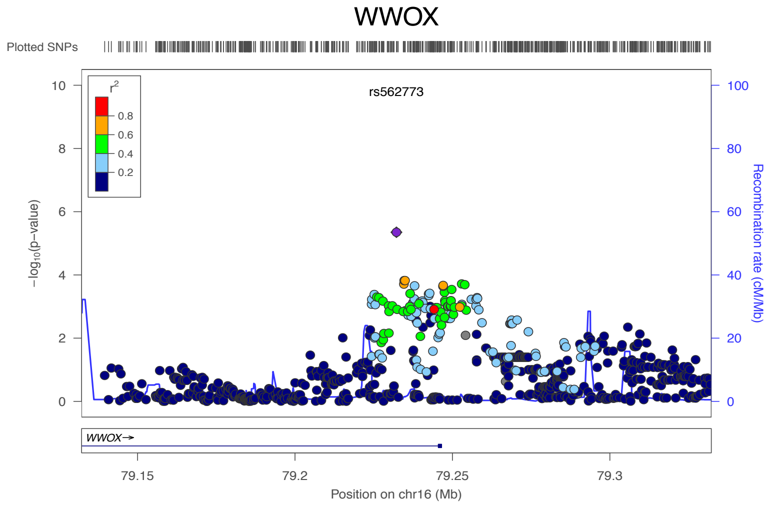


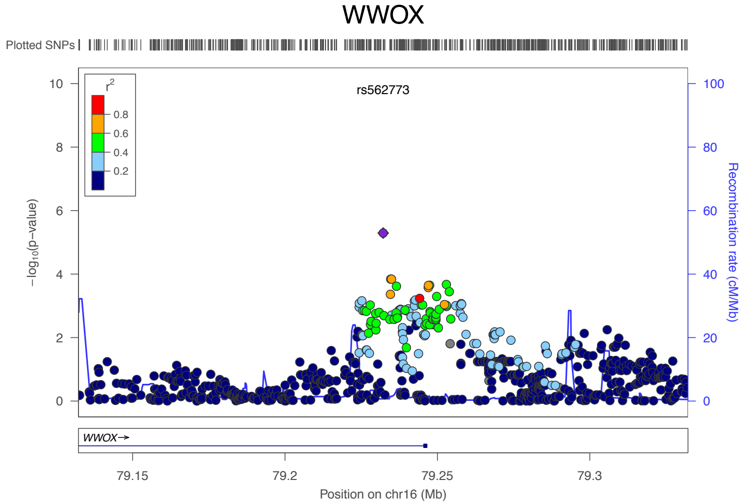
Ch 10

CS9: GWAS for Conversion Time MCI-AD

APOE4 effects were not adjusted

SNP	Chromosome	Position	P-value	Gene
rs62514059	8	128638024	1.5×10^{-7}	
rs78908045	1	78720788	1.4×10^{-6}	MGC27382
rs2694974	12	19954322	2.1×10^{-6}	
rs7278371	21	44025176	4.0×10^{-6}	
rs562773	16	79232220	4.5×10^{-6}	WWOX
rs74712657	22	50834181	4.8×10^{-6}	PPP6R2
ATAG	7		6.4×10^{-6}	NPSR1
rs7810386	7	1952031	1.0×10^{-5}	MAD1L1





APOE4 effects were adjusted

SNP	Chromosome	Position	P-value	Gene
rs62514059	8	128638024	1.2×10^{-6}	
rs74712657	22	50834181	1.3×10^{-6}	PPP6R2
rs562773	16	79232220	2.6×10^{-6}	WWOX
rs11044865	12	19954488	3.7×10^{-6}	
rs3856926	3	189082792	4.0×10^{-6}	
rs12683859	9	4727444	5.5×10^{-6}	AK3
rs7278371	21	44025176	6.6×10^{-6}	LOC101928233

Lee,

C10. Alzheimer’s Disease DREAM Challenge 1

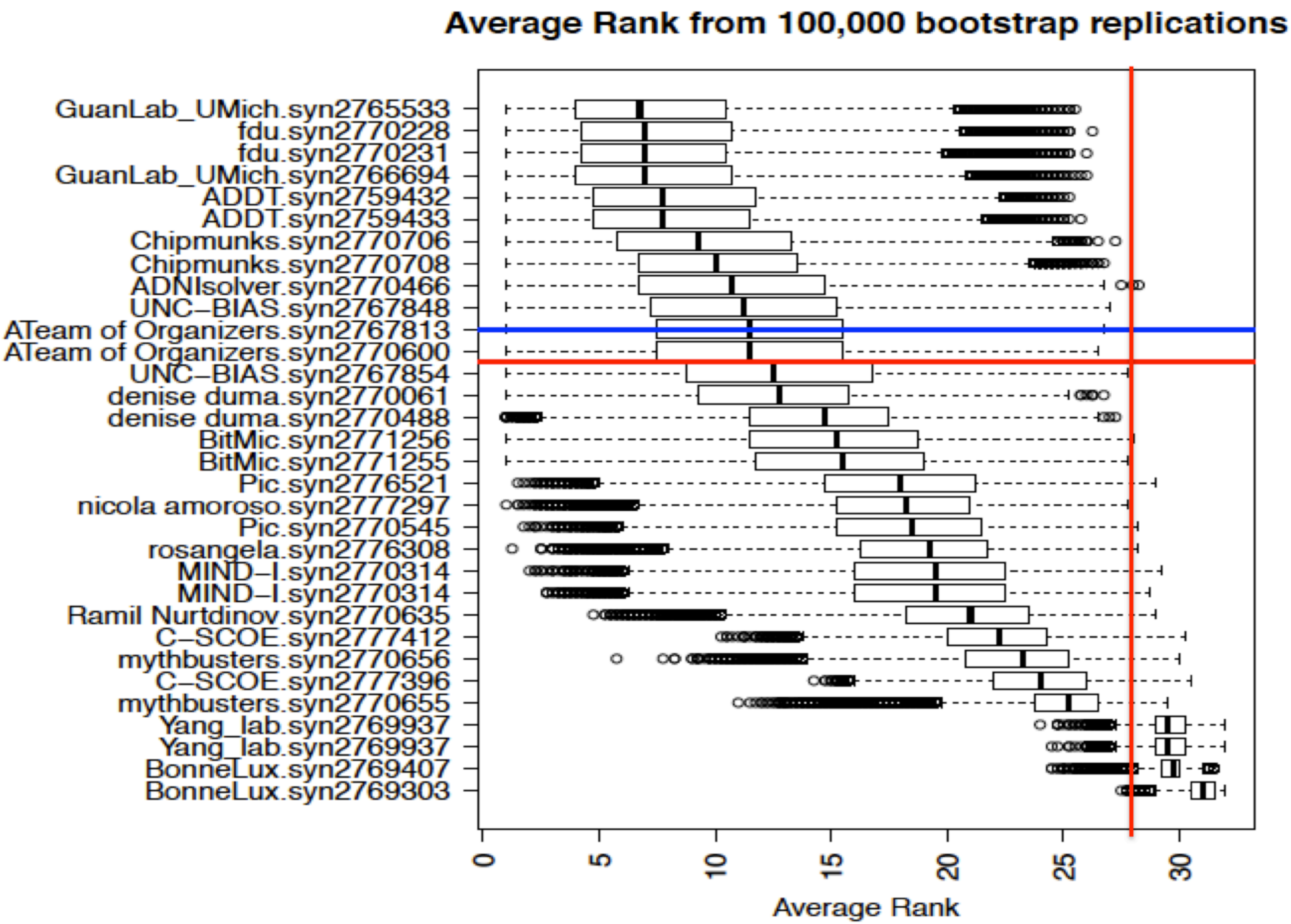
Its goal is to apply an open science approach to rapidly identify **accurate predictive AD biomarkers** that can be used by the scientific, industrial and regulatory communities to improve AD diagnosis and treatment.

Sub 1: Predict the change in cognitive scores 24 months after initial assessment.

Sub 2: Predict the set of cognitively normal individuals whose biomarkers are suggestive of amyloid perturbation.

Sub 3: Classify individuals into diagnostic groups using MR imaging.

Alzheimers Disease Big Data DREAM Challenge 1

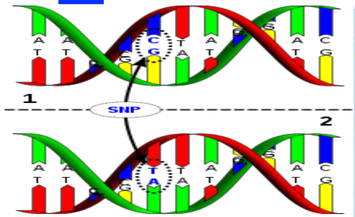


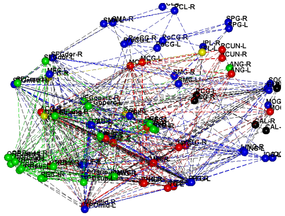
Formulation

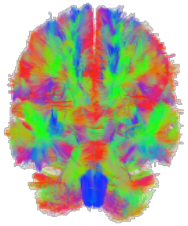
Data $\{(y_i, X_i) : i = 1, \cdots, n\}$ $\quad X_i = \{X_i(d) : d \in D\}$

$$y_i = f(X_i) + \varepsilon_i$$

Disease Status, Survival Time, Treatment, Trajectories







Interesting scientific questions include

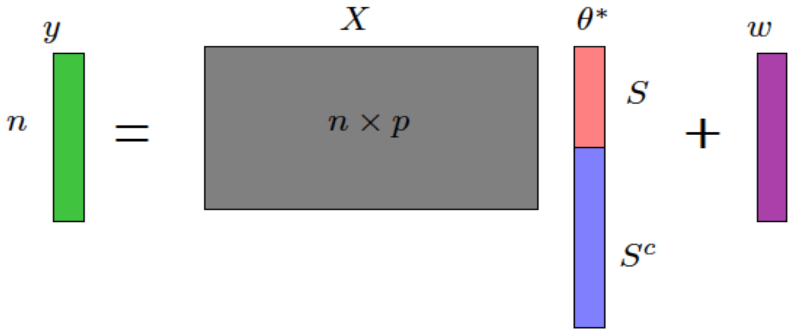
- Determine disease status
- Identify earlier biomarker
- Predict disease trajectories
- Predict survival time (e.g., time-to-event)

HRM versus FRM

Data $\{(y_i, X_i) : i = 1, \cdots, n\}$ $X_i = \{X_i(d) : d \in D\}$

$$y_i =< X_i, \theta > + \varepsilon_i$$

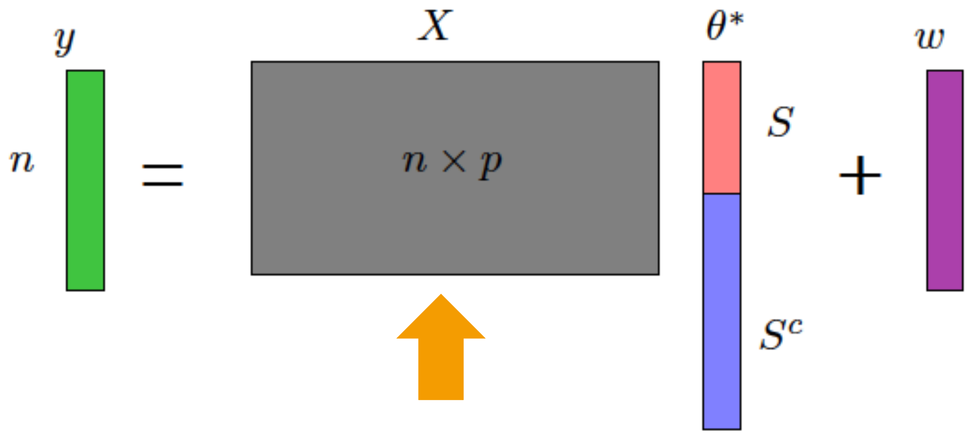
Strategy 1: Discrete Approach
(High-dimension Regression Model (HRM))

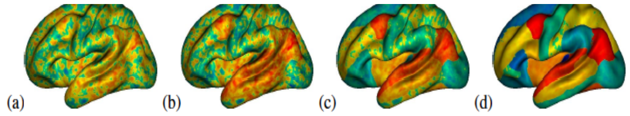


Strategy 2: Functional Regression Model (FRM)

$$y_i = \theta_0 + \int_D \theta(d) X_i(d) m(d) + \varepsilon_i$$

HRM





$$\widehat{\theta} \in \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \lambda_n \sum_{j=1}^{p} |\theta_j|$$

Key Conditions:

$$S = \{j : \beta_j \neq 0\}$$

- Sparsity of S
- Restricted Isometry Property (RIP) for design matrix X

FRM

Strategy 2: Functional Approach

$$y_i = \theta_0 + \int_D \theta(d) X_i(d) m(d) + \varepsilon_i$$

$$\theta(d) = \sum_{k=1}^{\infty} \theta_k \psi_k(d)$$

$$y_i = \theta_0 + \sum_{k=1}^{\infty} \theta_k \int_D \psi_k(d) X_i(d) m(d) + \varepsilon_i$$

Basis Methods: fixed and data-driven basis functions

$$K_\theta = \{\theta(d) = \sum_{k=1}^{\infty} \theta_k \psi_k(d) : (\theta_1, \cdots) \in \ell^2\} \Longleftrightarrow C(d,d') = Cov(X(d), X(d')) = \sum_{k=1}^{\infty} \lambda_k \zeta_k(d) \zeta_k(d')$$

Key Conditions

Key Conditions: an excellent set of basis functions

$$\theta(d) \approx \sum_{k=1}^{K} \theta_k \psi_k(d) \qquad K << n$$

$$K_\theta = \{\theta(.)\} \quad \underline{\textbf{Alignment}} \quad K_X = \{X(.)\}$$

- **Sparsity of basis representation** $\{\theta_k : k = 1, \cdots\}$
- **Decay rate of spectral of** C **or** $K^{1/2}CK^{1/2}$

HRM

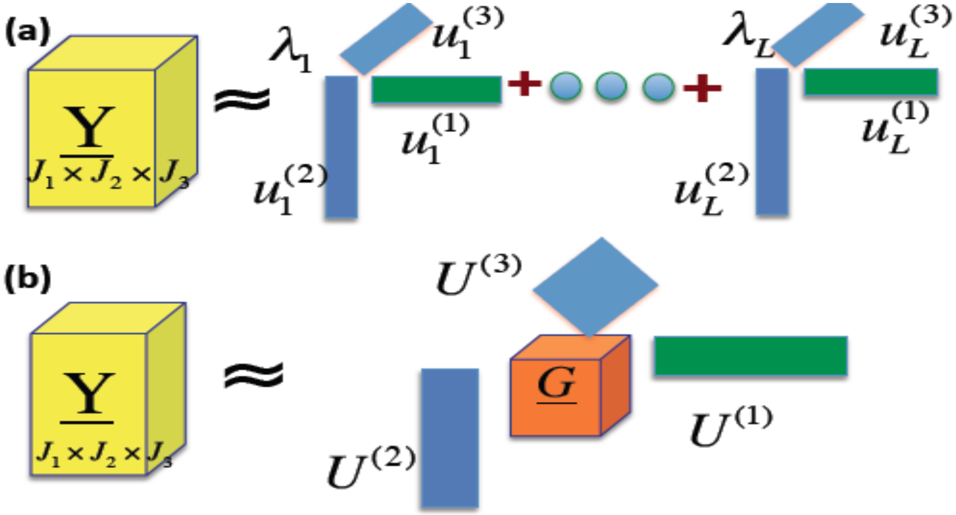
$$Y \mid X \sim \text{Exponential Family}(\mu, \phi)$$

$$g(\mu) = \theta_0^T Z + < X, \beta_0 >$$

CP decomposition

Tucker decomposition

$$\beta_0$$



Total Variation Penalty:

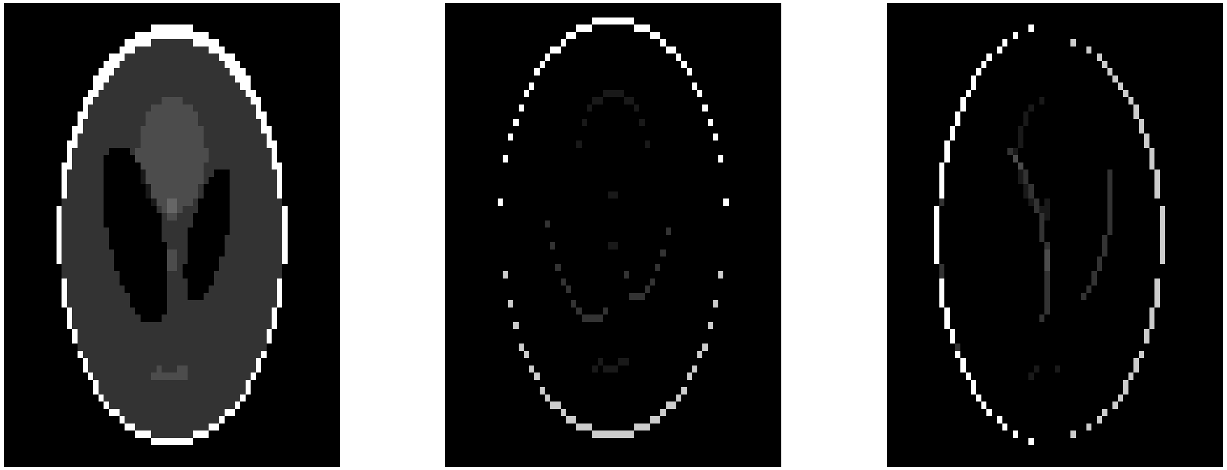
$$||\beta_0||_{TV} = \sup\left\{ \int_\Omega \beta_0(u,v)\text{div } f(u,v)dudv : f \in C_c^\infty(\Omega; R^2), |f|_\infty \leq 1 \right\}$$

Total Variation

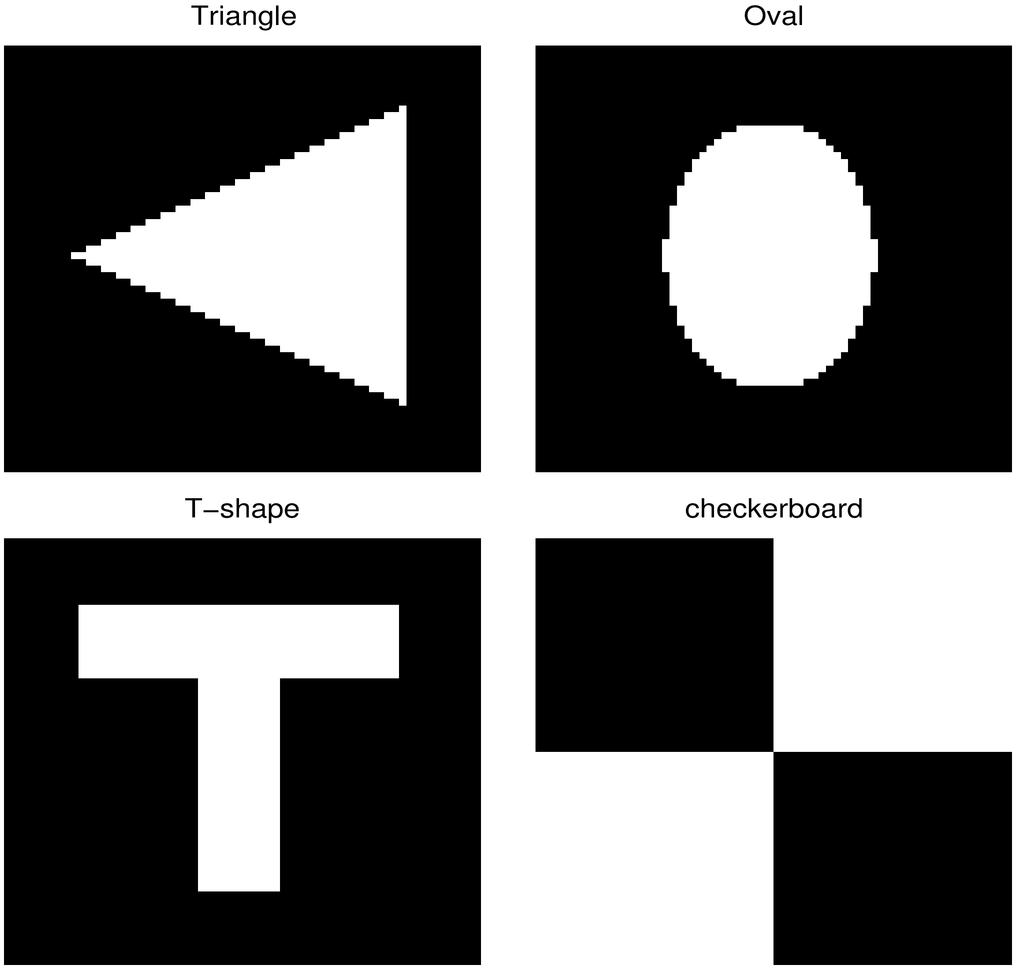
The total variation has been introduced in Computer Vision first by Rudin, Osher and Fatemi, 1992.

Many real images with edges have small total variation since image edges usually reside in a low-dimensional subset of pixe

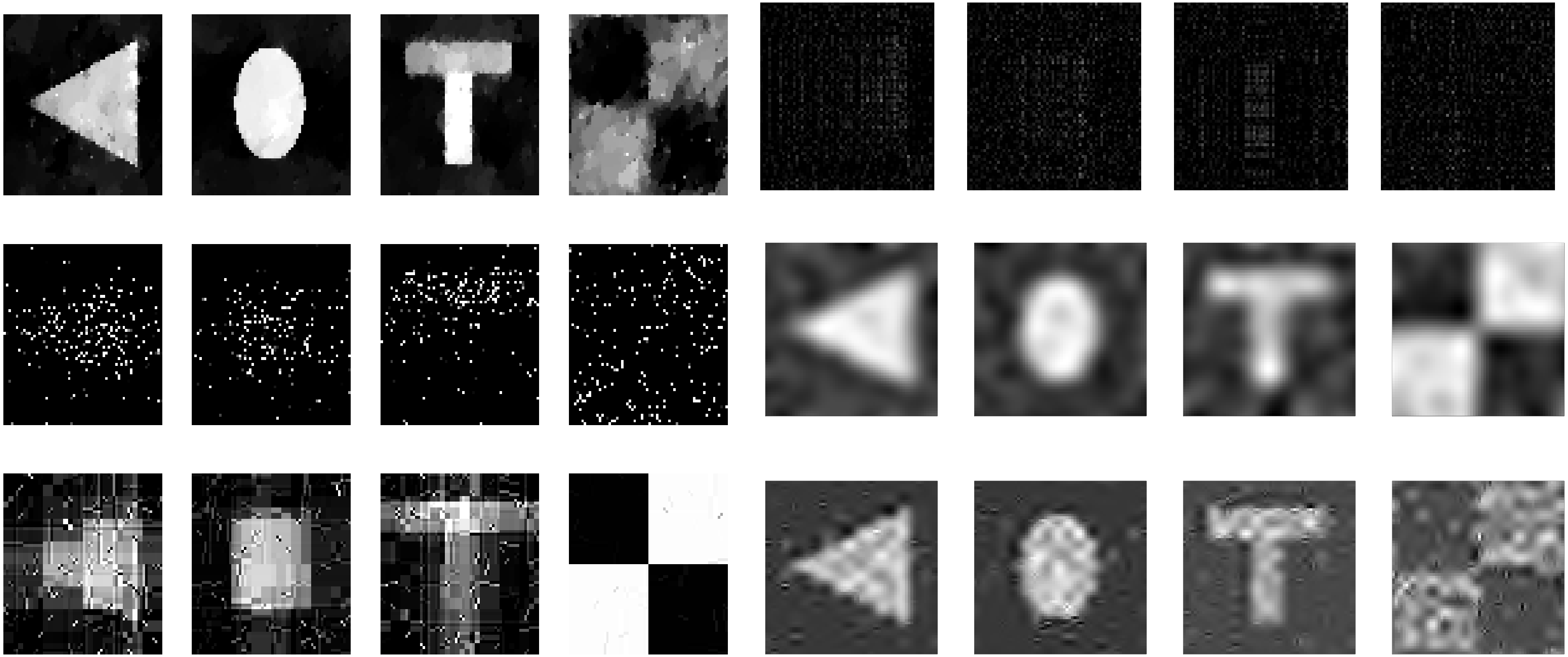
It has proved to be quite efficient for regularizing images without smoothing the boundaries of the objects.



True Image



Results



TV (Top row); Lasso (Second row); Lasso-Haar (Third row); Matrix regression (fourth row); FPCR (Fifth row); and WNET(Sixth row).

ADNI

- The sample in our investigation includes $n = 403$ subjects: 223 healthy controls (HC) (107 females and 116 males) and 180 individuals with AD (87 females and 93 males).
- The image predictor X_i is the 2D representation of left hippocampus. The covariate vector Z_i includes constant(=1), gender (Female=0 and Male = 1), age (55—92), and behavior score (1—36).
- Given (X_i, Z_i), Y_i is assumed to follow a Bernoulli distribution with the success probability p_i satisfying

$$\text{logit}(p_i) = \langle X_i, \beta_0 \rangle + \theta_0^T Z_i \quad \text{for} \quad i = 1, \ldots, n.$$

Estimated Coefficient Maps

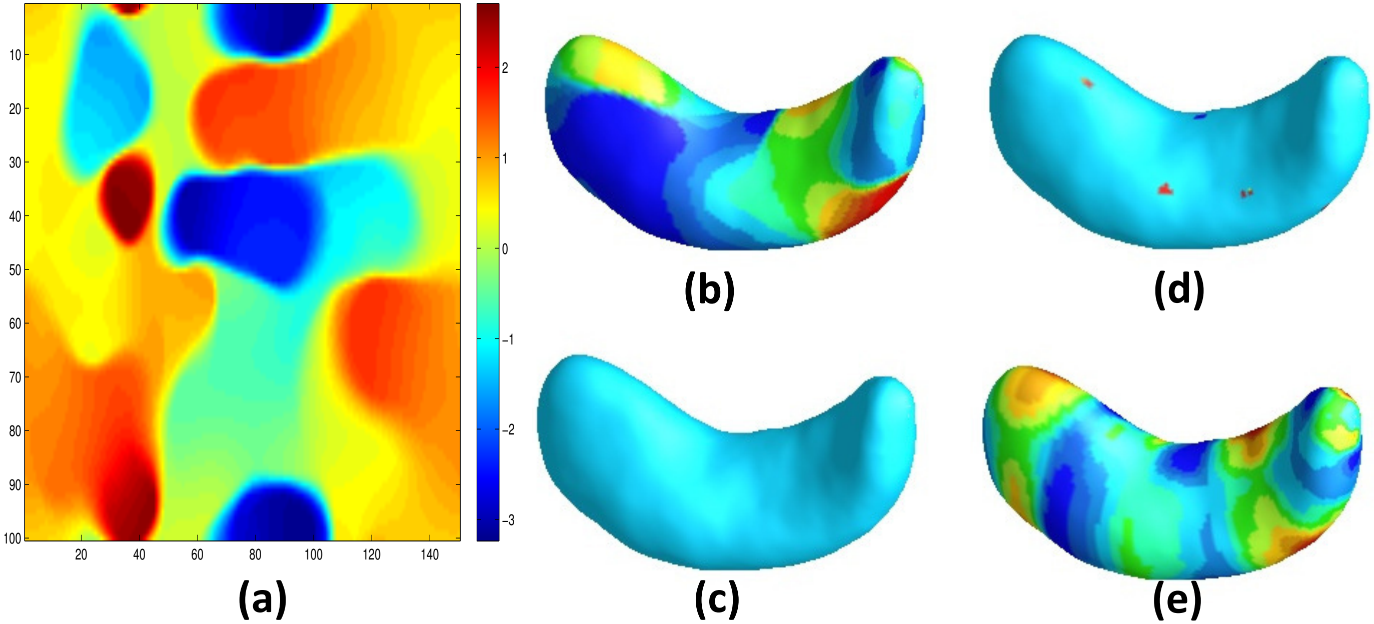


Figure : Estimated coefficient images for hippocampus data based four methods: the 2d-representation of TV estimator (a) and the surface representation of TV estimator (b), Lasso estimator (c), Lasso-wavelet estimator (d), and matrix regression estimator (e).

Functional Linear Cox Regression Model

- $h_i(t)$, the *i*-th hazard function, is defined as the event rate at time *t* conditional on survival until time t or later.

- The covariates are multiplicatively related to the hazard.

- $X_i(s)$, denotes the image data, z_{ik} denotes the scalar covariates

- The hazard function of the i-th subject under Cox regression is

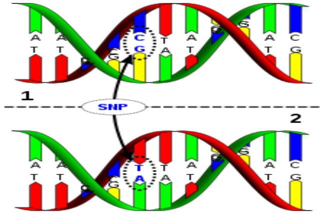
$$h_i(t) = h_0(t)\exp(\sum_{k=1}^{p} z_{ik}\gamma_k + \int_S X_i(s)\beta(s)ds)$$

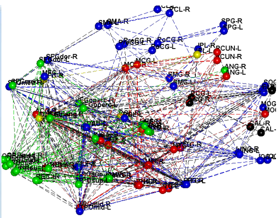
Formulation

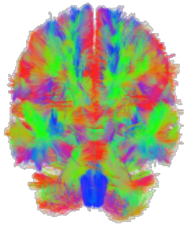
Data $\{(y_i, X_i) : i = 1, \cdots, n\}$ $X_i = \{X_i(d) : d \in D\}$

$$y_i = f(X_i) + \varepsilon_i$$

Disease Status
Survival Time
Treatment
Trajectories





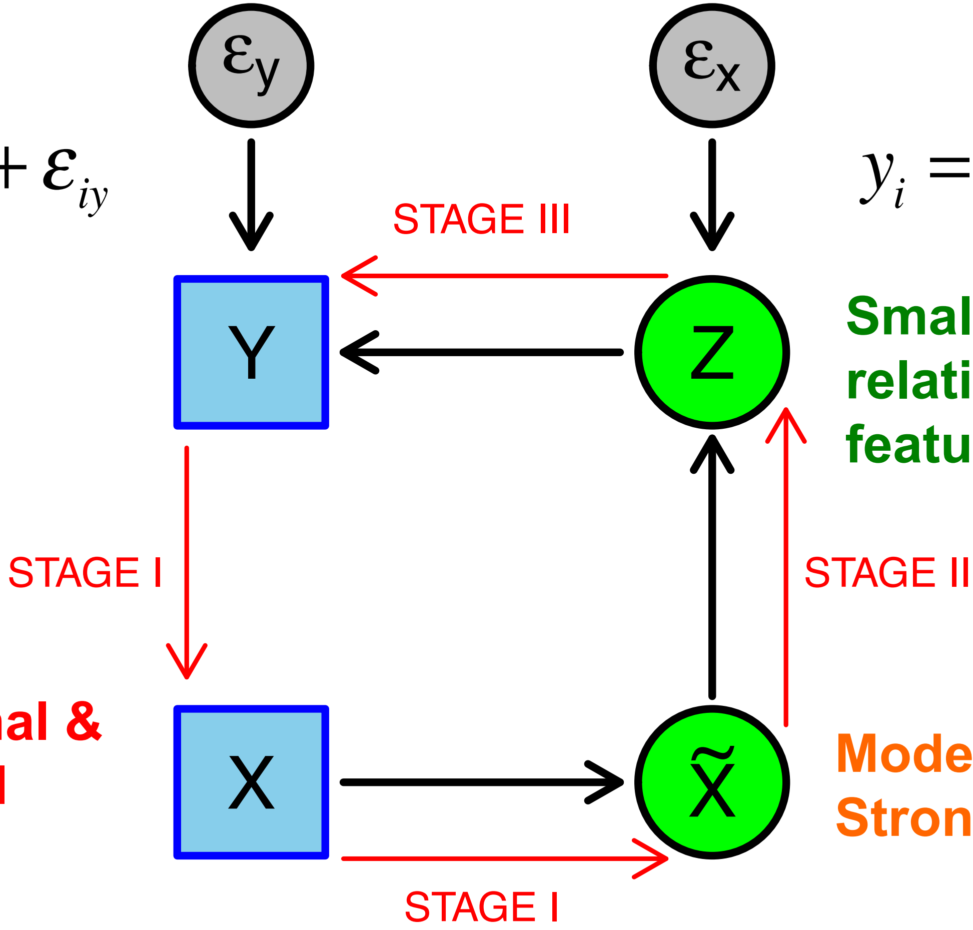


- Is this the right X space for prediction?
- How to deal with the curse of dimensionality?
- How to choose the loss function?

Path Diagram

$$y_i = f_0(X_i) + \varepsilon_{iy}$$

$$y_i = f(z_i) + \varepsilon_i$$



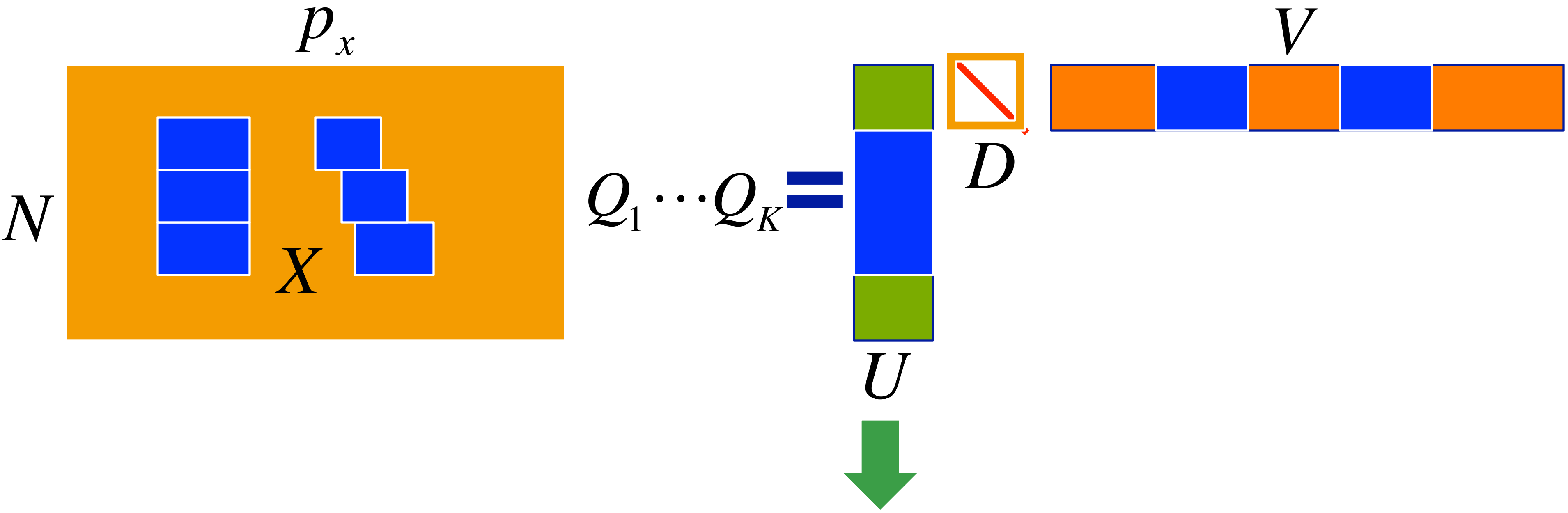
Small dimensional & relatively independent features

High-dimensional & Strongly Spatial features

Moderate dimensional & Strong Spatial features

MWPCR

Model $$(X - 1_N \mu_x^T) Q_1 \cdots Q_K = UDV + E$$



$$y_i = f(X_i) + \varepsilon_i = g(u_i) + \varepsilon_i$$

MWPCR

Prewhiten

$$\tilde{X}_R = (X - 1_N \widehat{\mu}_x^T) Q_1 \cdots Q_K$$

GPCA

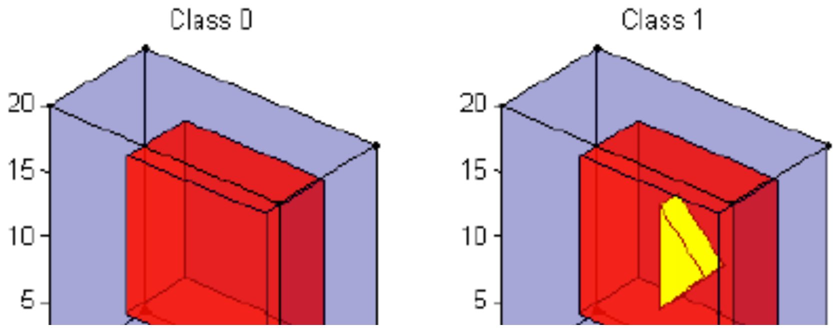
$$Q_K \cdots Q_1 (x_i - \mu_x) = \Lambda u_i + e_i$$

$$\|\tilde{X}_{R,\ell} - \sum_{k=1}^{K} d_{k,\ell} \boldsymbol{u}_{k,\ell} \boldsymbol{v}_{k,\ell}^T\|^2 + \lambda_u \sum_{k=1}^{K} P_1(d_{k,\ell} \boldsymbol{u}_{k,\ell}) + \lambda_v \sum_{k=1}^{K} P_2(d_{k,\ell} \boldsymbol{v}_{k,\ell})$$

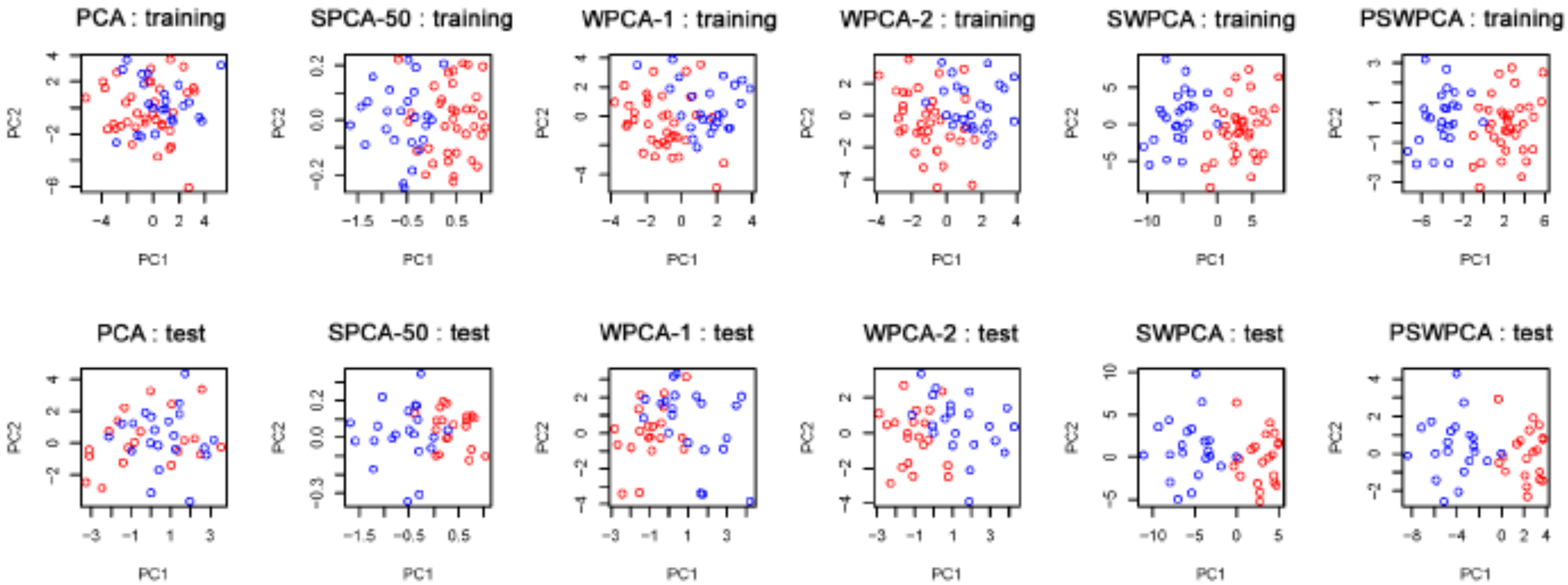
Regression

$$y_i = f(X_i) + \varepsilon_i = g(u_i) + \varepsilon_i$$

Spatially Weighted PCA



Guo, Ahn, and Zhu (2014) JCGS



Spatially Weighted PCA

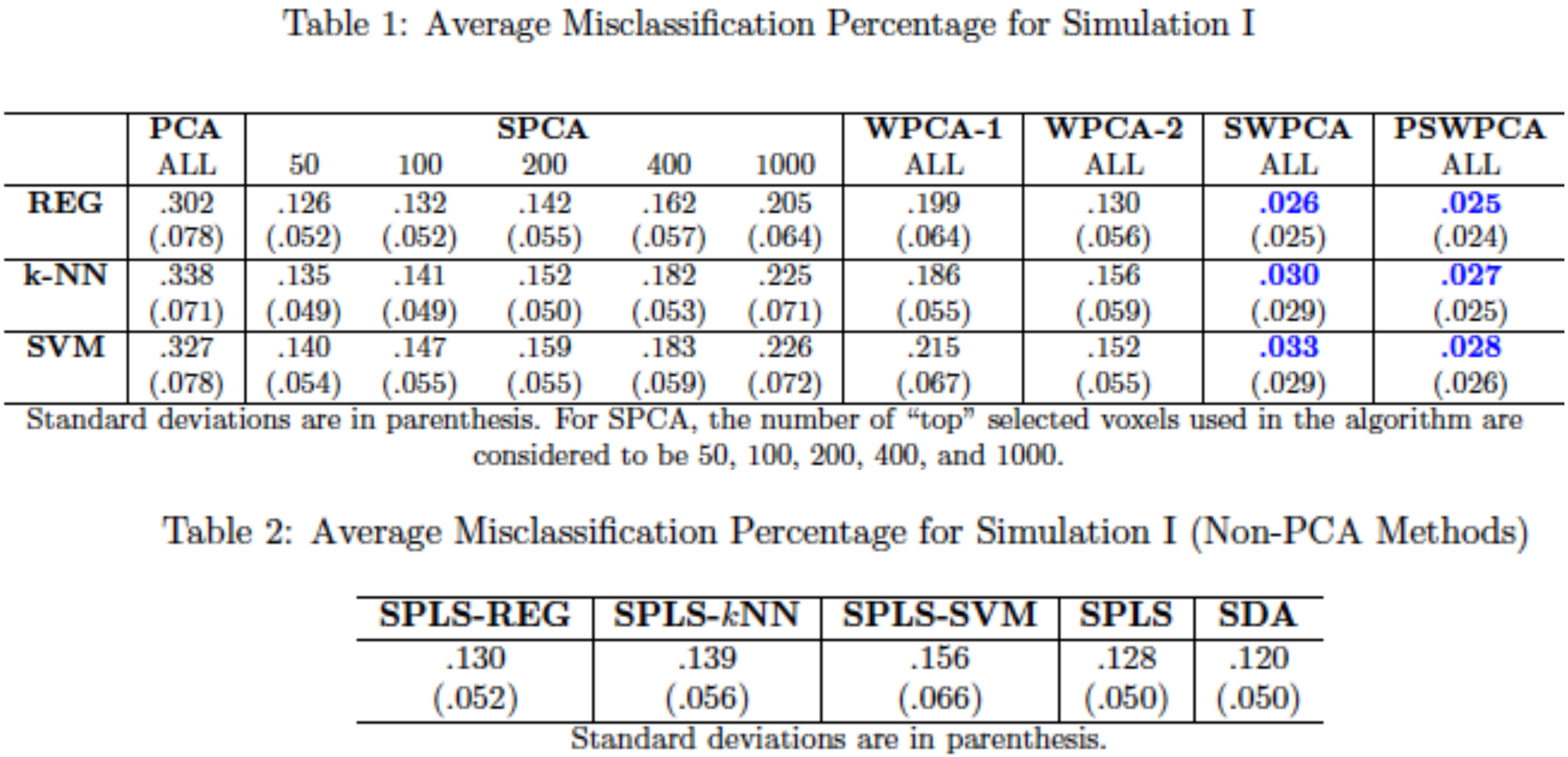


Table 1: Average Misclassification Percentage for Simulation I

	PCA	SPCA					WPCA-1	WPCA-2	SWPCA	PSWPCA
	ALL	50	100	200	400	1000	ALL	ALL	ALL	ALL
REG	.302 (.078)	.126 (.052)	.132 (.052)	.142 (.055)	.162 (.057)	.205 (.064)	.199 (.064)	.130 (.056)	**.026** (.025)	**.025** (.024)
k-NN	.338 (.071)	.135 (.049)	.141 (.049)	.152 (.050)	.182 (.053)	.225 (.071)	.186 (.055)	.156 (.059)	**.030** (.029)	**.027** (.025)
SVM	.327 (.078)	.140 (.054)	.147 (.055)	.159 (.055)	.183 (.059)	.226 (.072)	.215 (.067)	.152 (.055)	**.033** (.029)	**.028** (.026)

Standard deviations are in parenthesis. For SPCA, the number of "top" selected voxels used in the algorithm are considered to be 50, 100, 200, 400, and 1000.

Table 2: Average Misclassification Percentage for Simulation I (Non-PCA Methods)

SPLS-REG	**SPLS-*k*NN**	**SPLS-SVM**	**SPLS**	**SDA**
.130 (.052)	.139 (.056)	.156 (.066)	.128 (.050)	.120 (.050)

Standard deviations are in parenthesis.

Spatially Weighted PCA

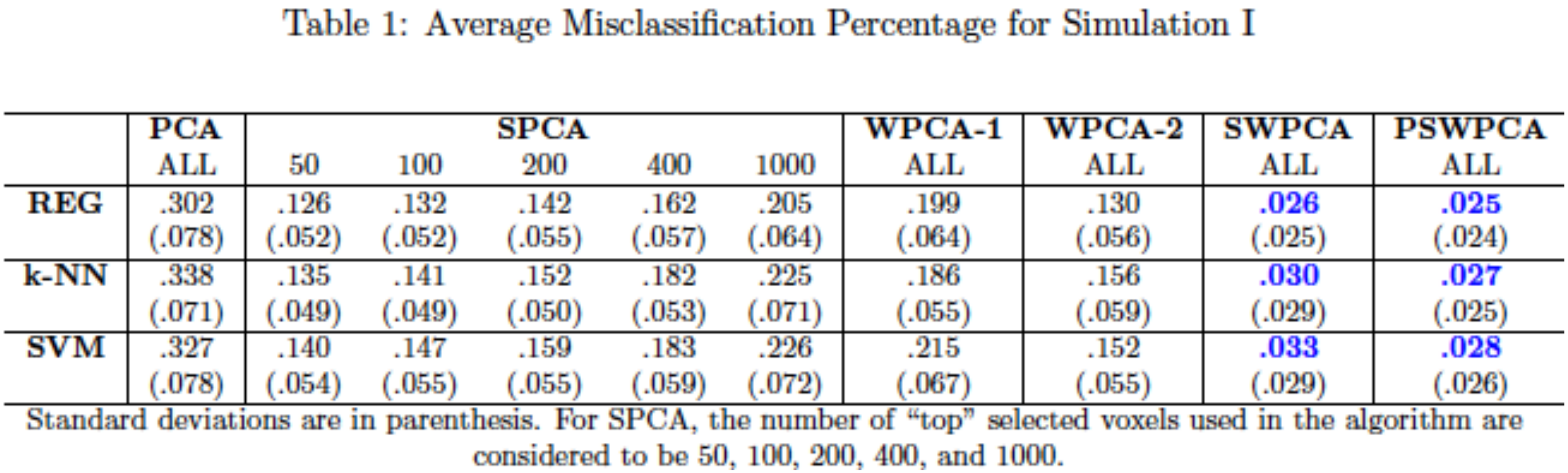


Table 1: Average Misclassification Percentage for Simulation I

	PCA	SPCA					WPCA-1	WPCA-2	SWPCA	PSWPCA
	ALL	50	100	200	400	1000	ALL	ALL	ALL	ALL
REG	.302 (.078)	.126 (.052)	.132 (.052)	.142 (.055)	.162 (.057)	.205 (.064)	.199 (.064)	.130 (.056)	**.026** (.025)	**.025** (.024)
k-NN	.338 (.071)	.135 (.049)	.141 (.049)	.152 (.050)	.182 (.053)	.225 (.071)	.186 (.055)	.156 (.059)	**.030** (.029)	**.027** (.025)
SVM	.327 (.078)	.140 (.054)	.147 (.055)	.159 (.055)	.183 (.059)	.226 (.072)	.215 (.067)	.152 (.055)	**.033** (.029)	**.028** (.026)

Standard deviations are in parenthesis. For SPCA, the number of "top" selected voxels used in the algorithm are considered to be 50, 100, 200, 400, and 1000.

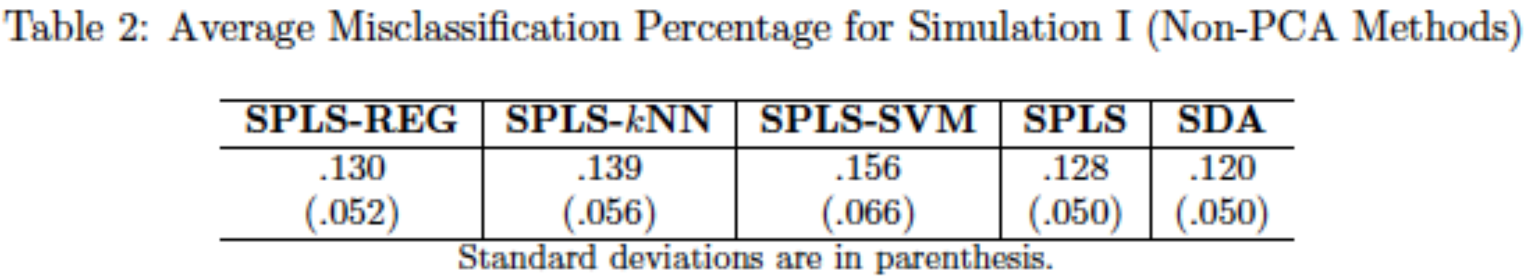
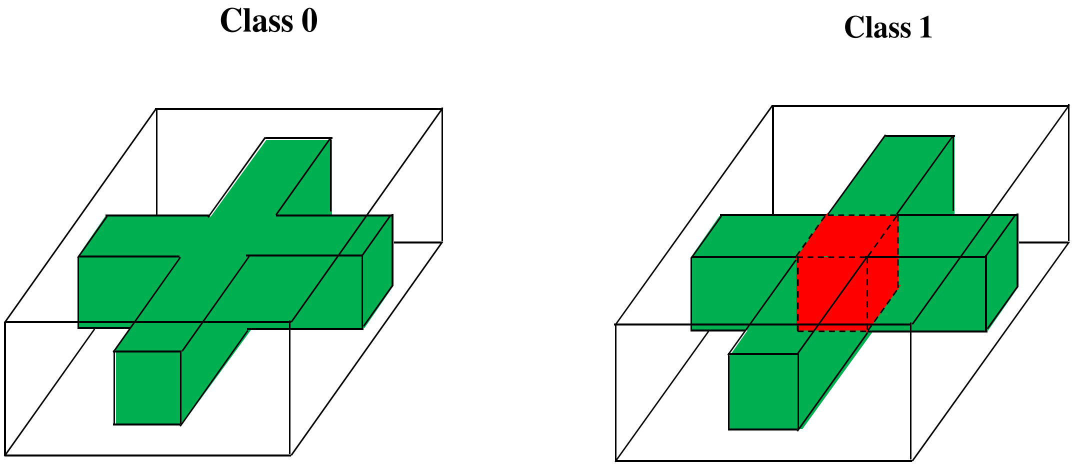


Table 2: Average Misclassification Percentage for Simulation I (Non-PCA Methods)

SPLS-REG	SPLS-*k*NN	SPLS-SVM	SPLS	SDA
.130 (.052)	.139 (.056)	.156 (.066)	.128 (.050)	.120 (.050)

Standard deviations are in parenthesis.

Simulation I: Classification



0	White
1	Green
2	Red

$$X_i(d)=\beta_0(d)+\beta_1(d)y_i+\varepsilon_i(d)$$

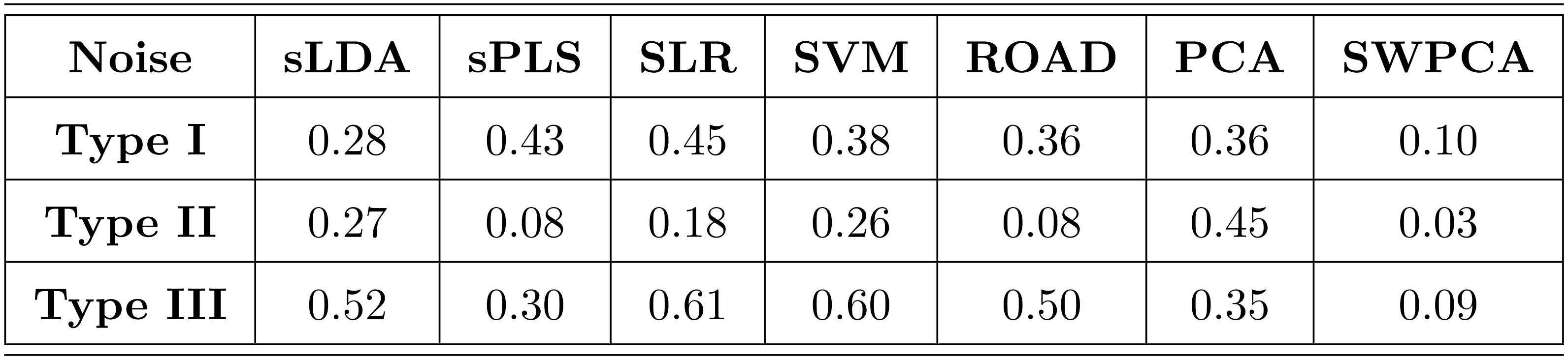
Type I	Type II	Type III
$N(0,4)$	Short-range correlation	Long-range correlation

Simulation I: Classification

Table 1: Misclassification rates for PCA and SWPCA under the different number of PCs.

Noise	Number of PCs	PCA	SWPCA1	SWPCA2	SWPCA3
Type I	**5**	0.40	0.11	0.09	0.10
	7	0.40	0.13	0.11	0.10
	10	0.40	0.13	0.11	0.10
Type II	**5**	0.40	0.04	0.08	0.03
	7	0.39	0.03	0.09	0.04
	10	0.38	0.03	0.07	0.04
Type III	**5**	0.40	0.13	0.10	0.09
	7	0.41	0.13	0.10	0.10
	10	0.41	0.13	0.10	0.10

Simulation I: Classification

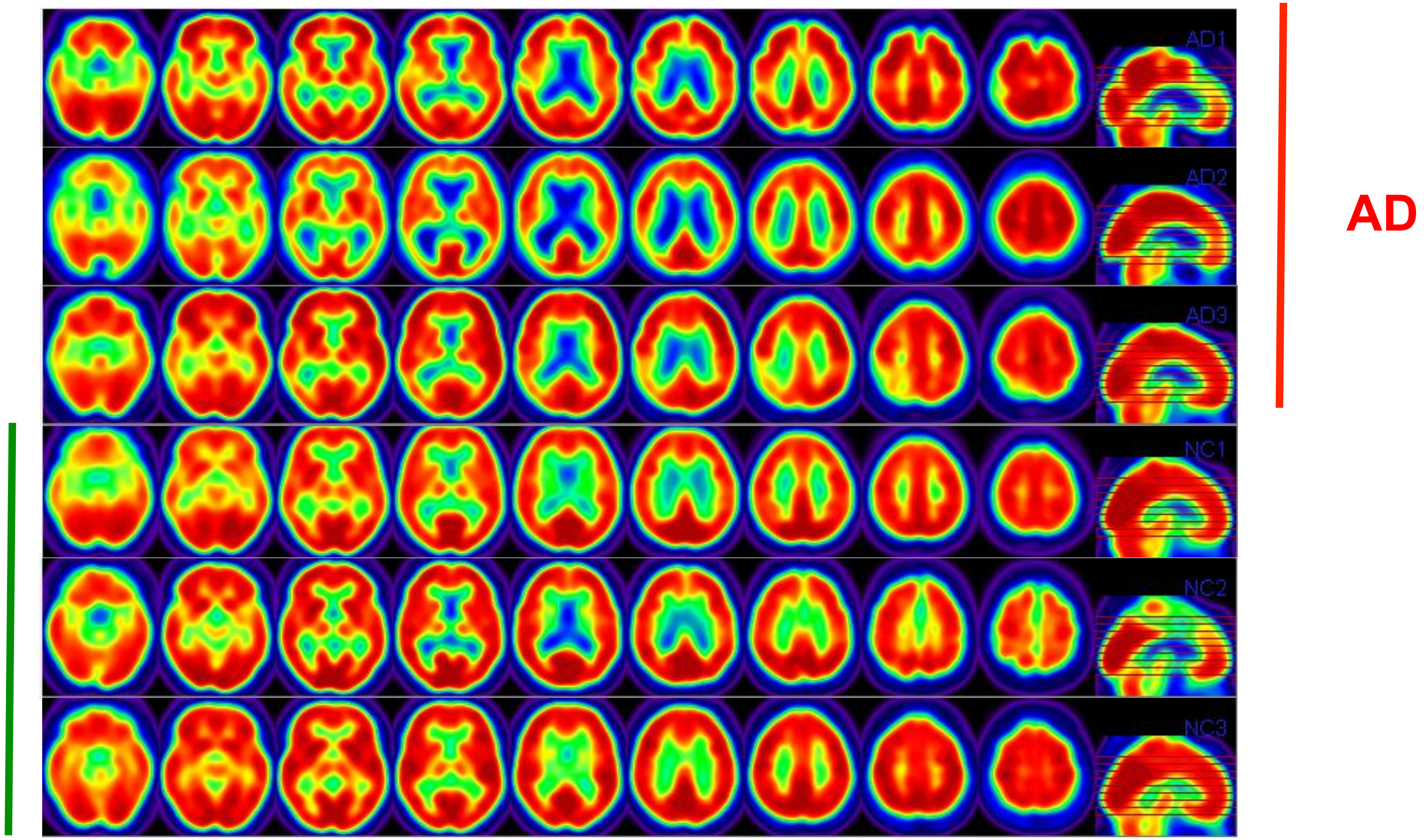


Noise	sLDA	sPLS	SLR	SVM	ROAD	PCA	SWPCA
Type I	0.28	0.43	0.45	0.38	0.36	0.36	0.10
Type II	0.27	0.08	0.18	0.26	0.08	0.45	0.03
Type III	0.52	0.30	0.61	0.60	0.50	0.35	0.09

sLDA: sparse discriminant analysis
sPLS: sparse partial least squares analysis
SLR: sparse logistic regression
SVM: support vector machine
ROAD:

ADNI

PET



AD

NC

ADNI

94 AD subjects and 104 NC subjects

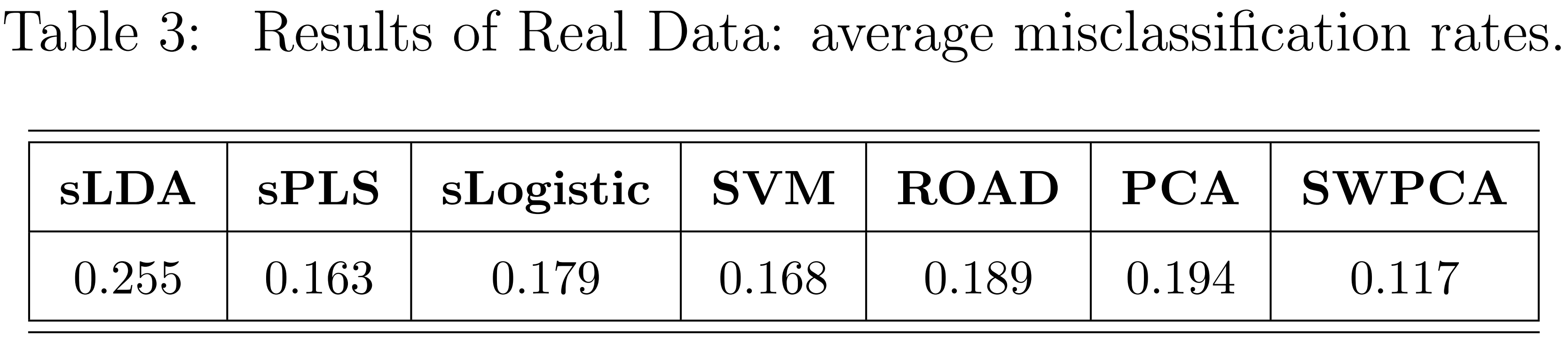


Table 3: Results of Real Data: average misclassification rates.

sLDA	sPLS	sLogistic	SVM	ROAD	PCA	SWPCA
0.255	0.163	0.179	0.168	0.189	0.194	0.117

Take-home Message

fPCA may not work in many cases.

Modified fPCA may work in some of these cases.

ASA: Statistics in Imaging Section

SAMSI

- 2013 Neuroimaging Data Analysis
- 2015-2016 Challenges in Computational Neuroscience

Thank You!!