Statistical Learning Methods for Neuroimaging Data Analysis with Applications

University of North Carolina at Chapel Hill

Hongtu Zhu

https://www.med.unc.edu/big-s2

CONTENTS

Part I

A Review of Neuroimaging Techniques

Part II

Imaging Processing Analysis Methods

Part III

Large scale Neuroimaging Related Studies

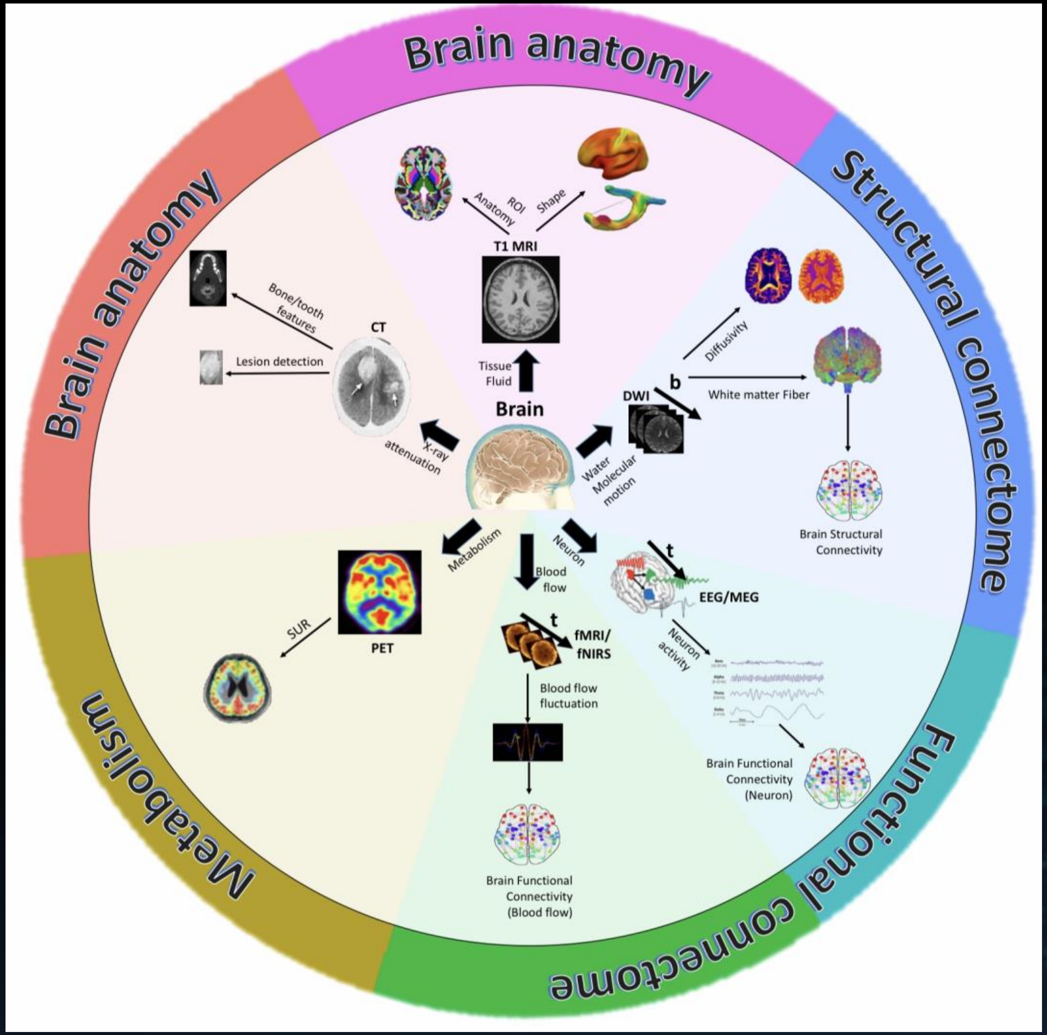
Part IV

Population-based Statistical Analysis Methods

Part I

A Review of Neuroimaging Techniques

Eight Popular Neuroimaging Techniques



- Structural magnetic resonance imaging (sMRI)

❖ Diffusion weighted MRI (DWI)

❖ Functional MRI (fMRI)

- Positron emission tomography (PET)

➢ Computerized tomography (CT)

❖ Electroencephalography (EEG)

❖ Magnetoencephalography (MEG)

➢ Functional near-infrared spectroscopy (fNIRS)

Each image modality has its tracer, data dimension, extracted features, and main clinical and research applications.

CT, PET, MEG, EEG, and fNIRS

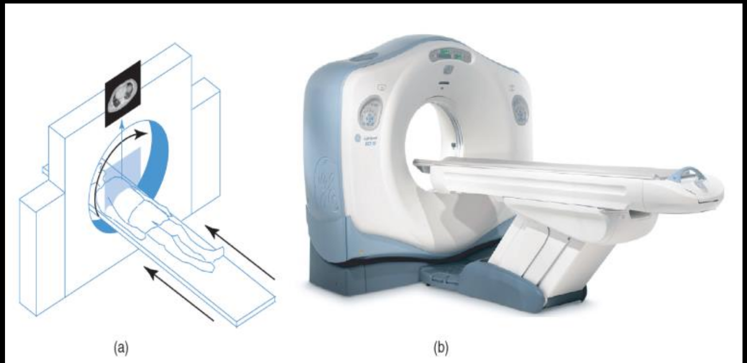
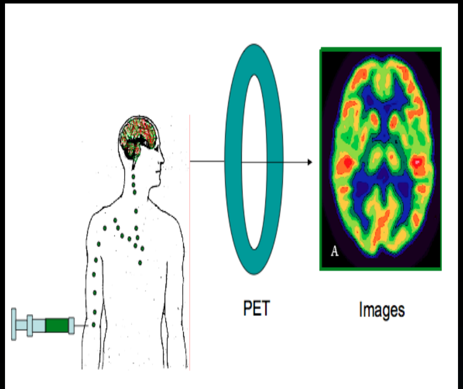
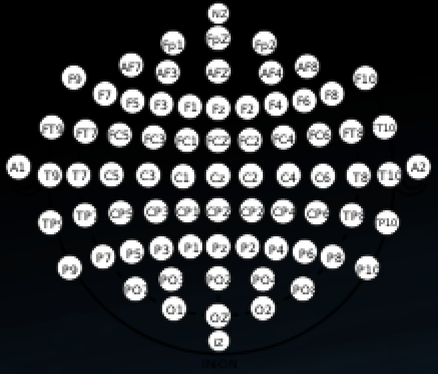
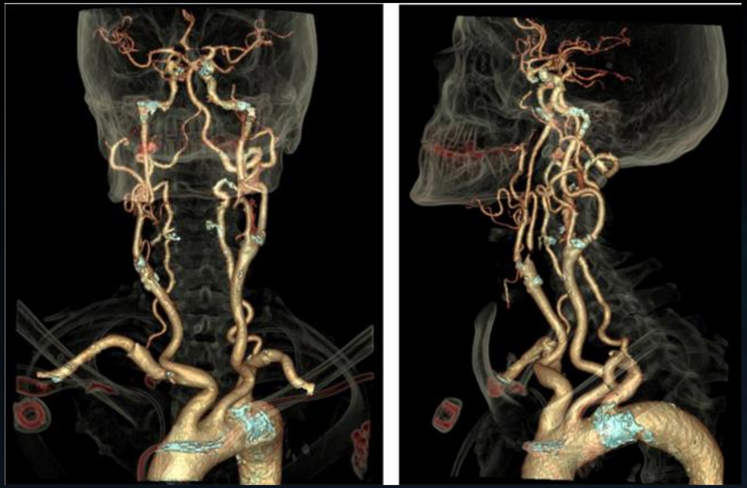
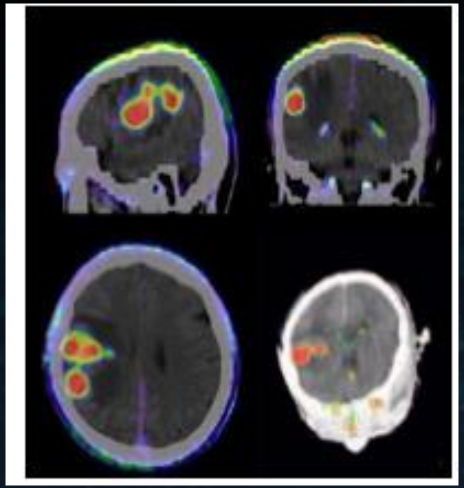


Figure 3.1 (a) Schematic representation, and (b) photograph of a CT scanner. (Courtesy of GE Healthcare.)




Nz
Fp1 Fpz Fp2
F9 AF7 AF3 AFz AF4 AF8 F10
F7 F5 F3 F1 Fz F2 F4 F6 F8
FT9 FT7 FC5 FC3 FC1 FCz FC2 FC4 FC6 FT8 T10
A1 T9 T7 C5 C3 C1 Cz C2 C4 C6 T8 T10 A2
TP9 TP7 CP5 CP3 CP1 CPz CP2 CP4 CP6 TP8 P10
P9 P7 P5 P3 P1 Pz P2 P4 P6 P8 P10
PO7 PO3 POz PO4 PO8
O1 Oz O2
Iz
INION

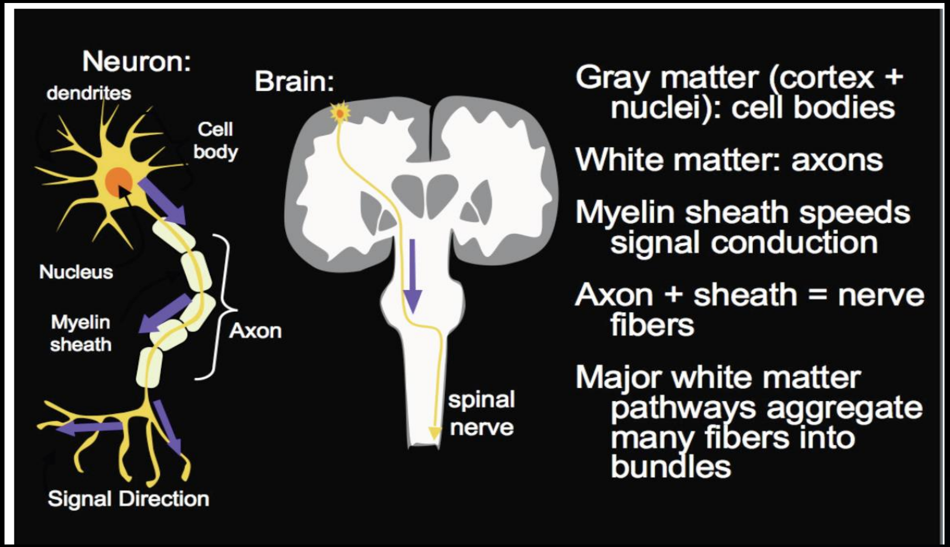




https://www.omegapds.com/ct-angiography-of-the-head-or-neck/

https://en.wikipedia.org/

sMRI, fMRI, and DWI



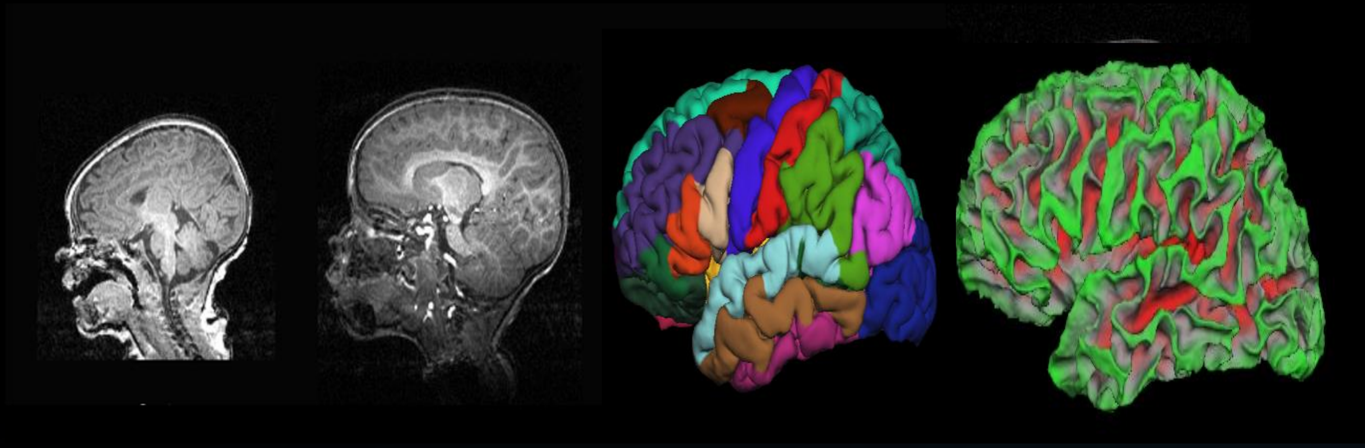
Gray matter (cortex + nuclei): cell bodies

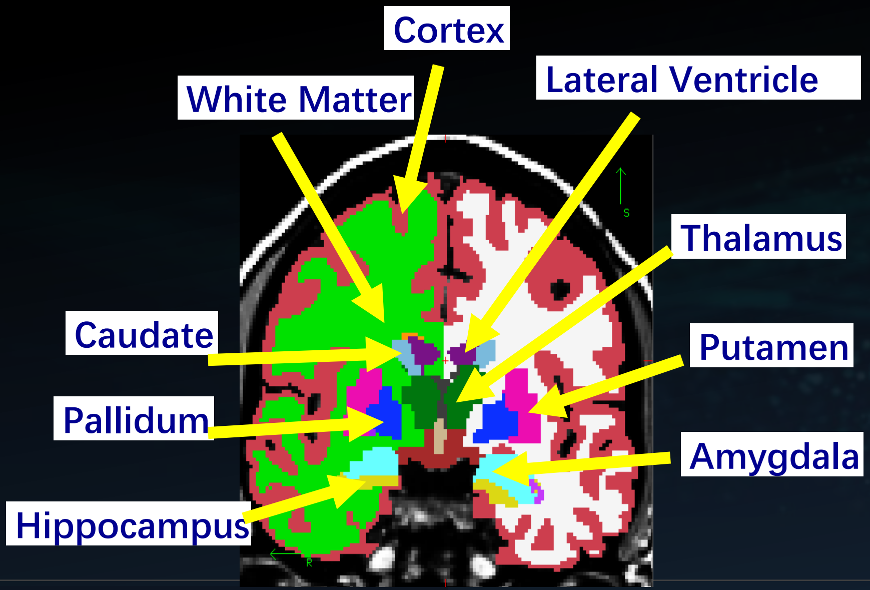
White matter: axons

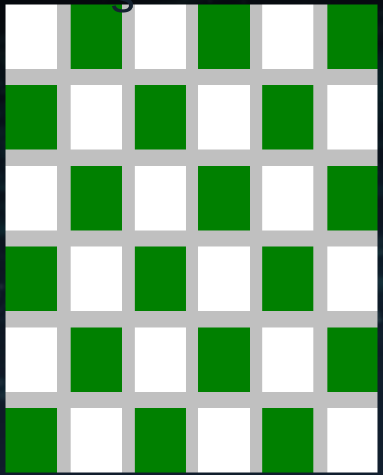
Myelin sheath speeds signal conduction

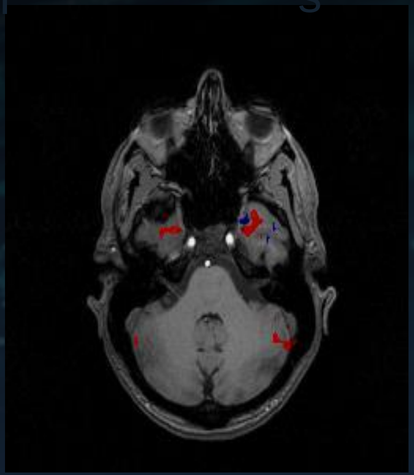
Axon + sheath = nerve fibers

Major white matter pathways aggregate many fibers into bundles









A Multi-model Approach

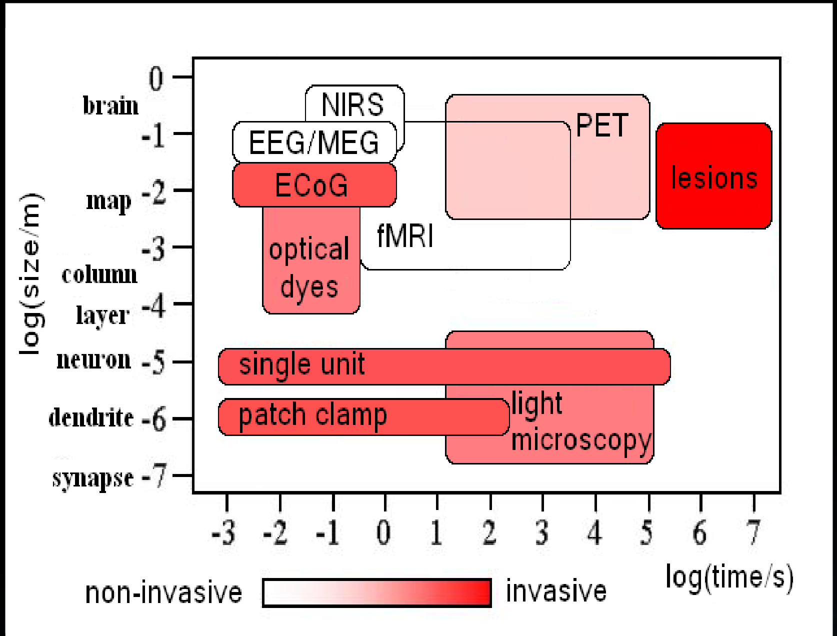
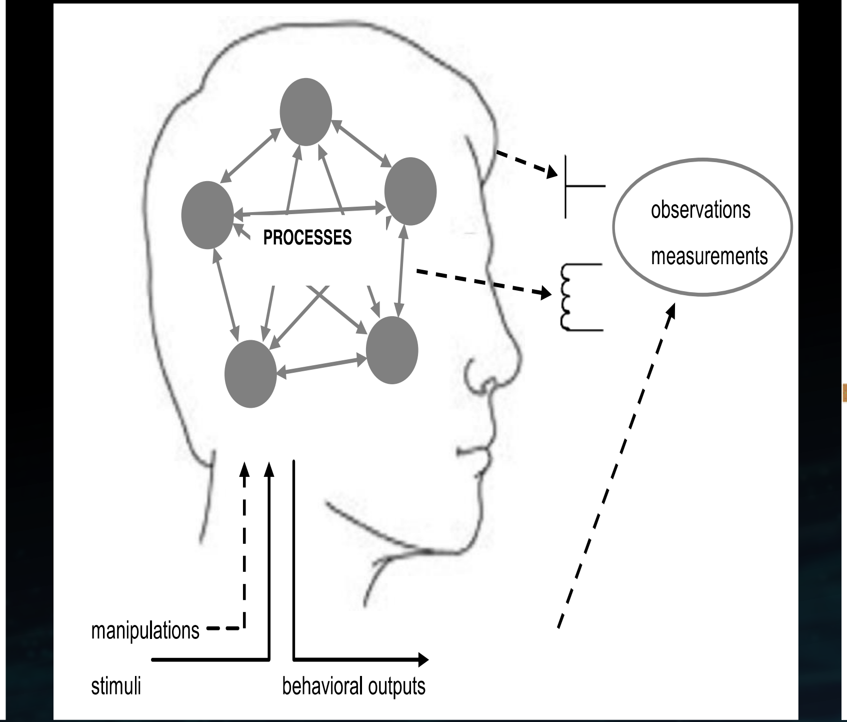
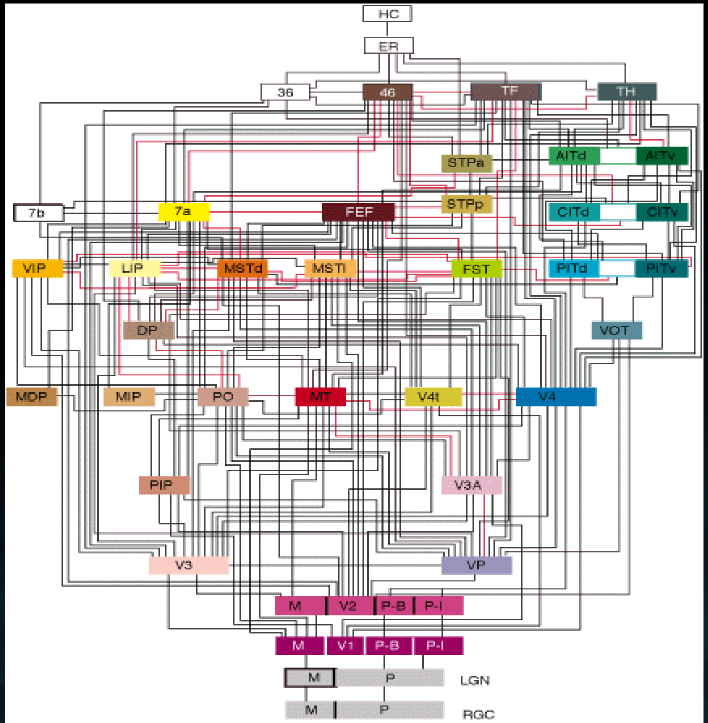


Image by A. Galka





- Different models at different scales.
- Ladder of overlapping models.
- Must be testable against multiple phenomena.

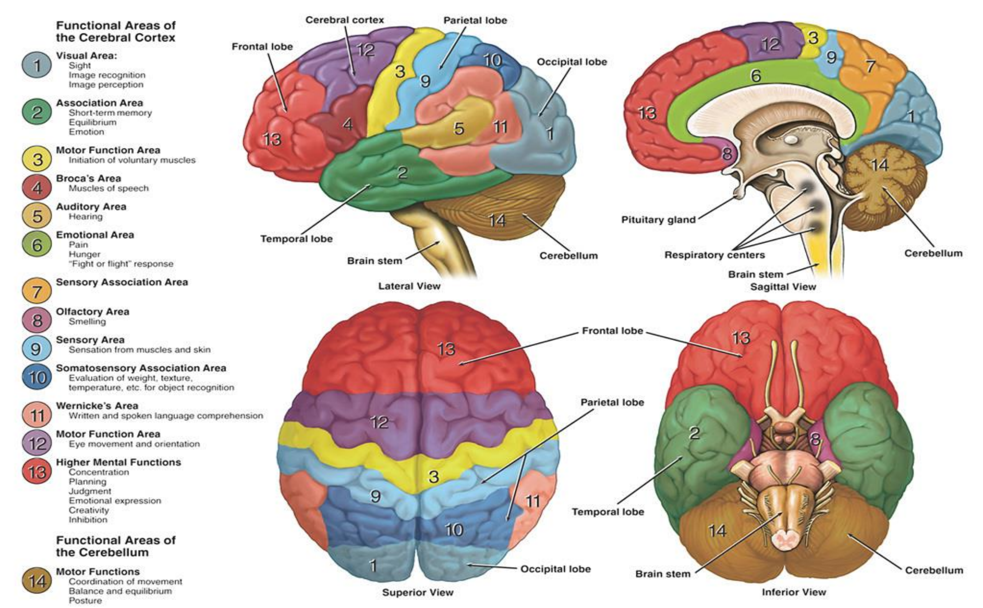
The van Essen diagram

Part II

Imaging Processing Analysis Methods

Four Common Themes (CT1)-(CT4)

(CT1) Complex Brain Objects

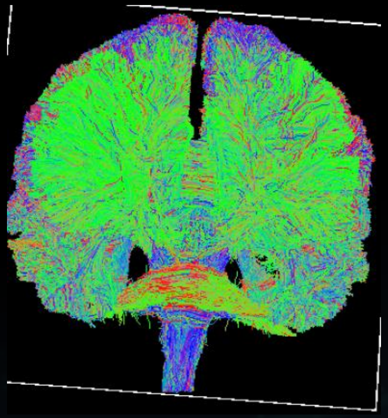

Functional Areas of the Cerebral Cortex
1 Visual Area: Sight Image recognition Image perception
2 Association Area Short-term memory Equilibrium Emotion
3 Motor Function Area Initiation of voluntary muscles
4 Broca's Area Muscles of speech
5 Auditory Area Hearing
6 Emotional Area Pain Hunger "Fight or flight" response
7 Sensory Association Area
8 Olfactory Area Smelling
9 Sensory Area Sensation from muscles and skin
10 Somatosensory Association Area Evaluation of weight, texture, temperature, etc. for object recognition
11 Wernicke's Area Written and spoken language comprehension
12 Motor Function Area Eye movement and orientation
13 Higher Mental Functions Concentration Planning Judgment Emotional expression Creativity Inhibition
Functional Areas of the Cerebellum
14 Motor Functions Coordination of movement Balance and equilibrium Posture
Cerebral cortex
Parietal lobe
Frontal lobe
Occipital lobe
Temporal lobe
Brain stem
Cerebellum
Lateral View
Pituitary gland
Respiratory centers
Brain stem
Cerebellum
Sagittal View
Frontal lobe
Parietal lobe
Occipital lobe
Superior View
Temporal lobe
Brain stem
Cerebellum
Inferior View

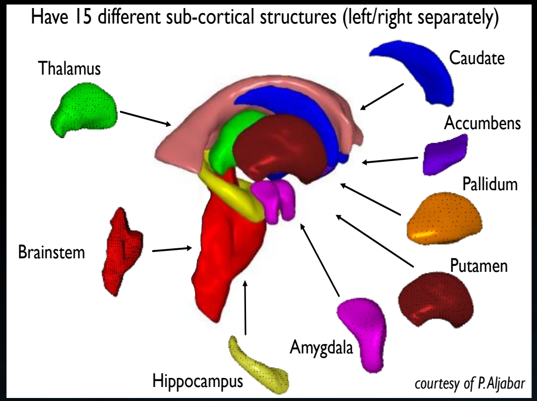
https://dana.org/article/neuroanatomy-the-basics/

(CT2) Complex Spatiotemporal Structures

- spatial and temporal resolutions
- spatio-temporal smoothness
- spatiotemporal correlation

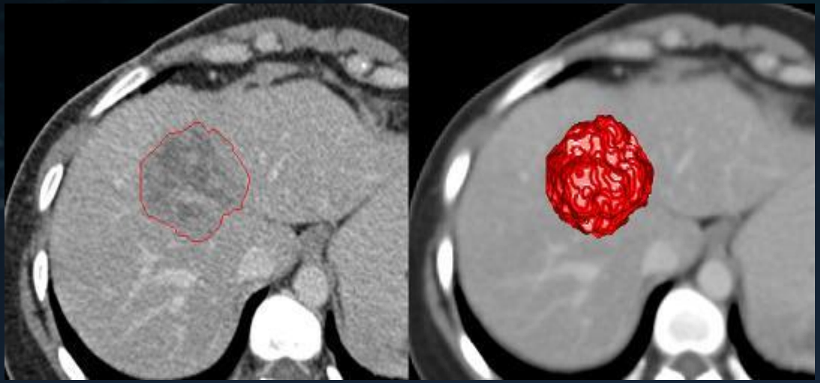
(CT3) Extremely High Dimensionality

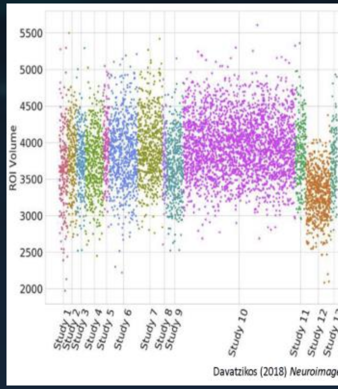



Have 15 different sub-cortical structures (left/right separately)

courtesy of P.Aljabar

(CT4) Heterogeneity within Individual Subjects and across Centers/Studies

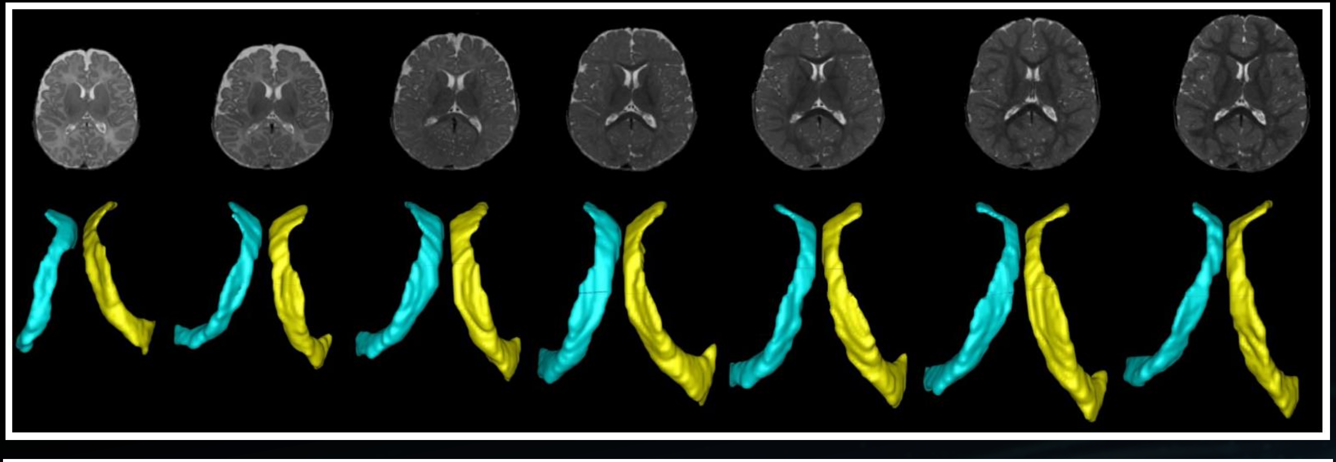


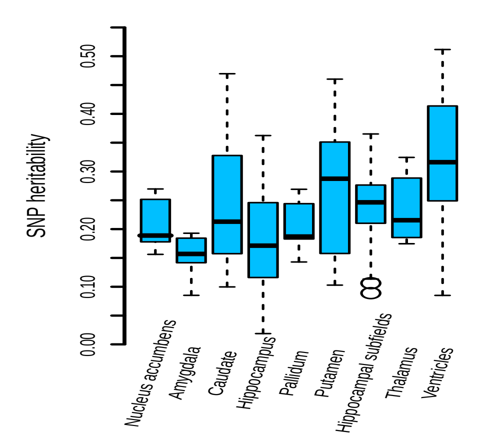


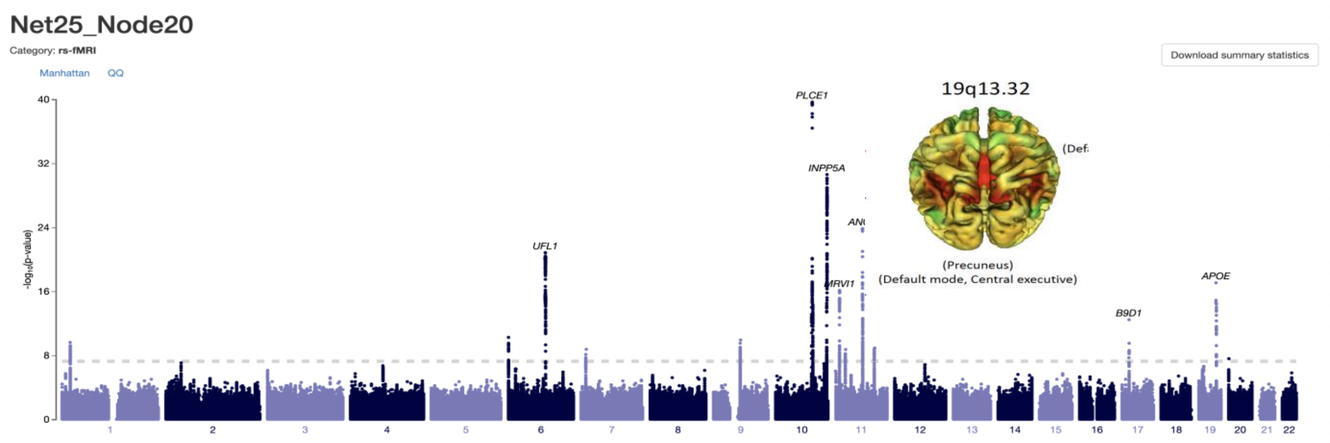
Davatzikos (2018) *Neuroimage*

Image Models

Image=*f*(*B(age, gene, race, disease, others)*, device, acquisition, noises)







Net25_Node20

Category: rs-fMRI

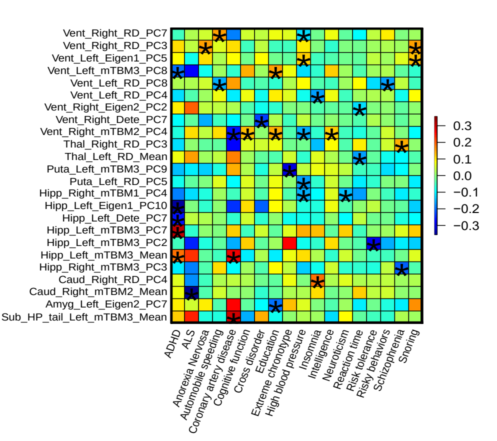
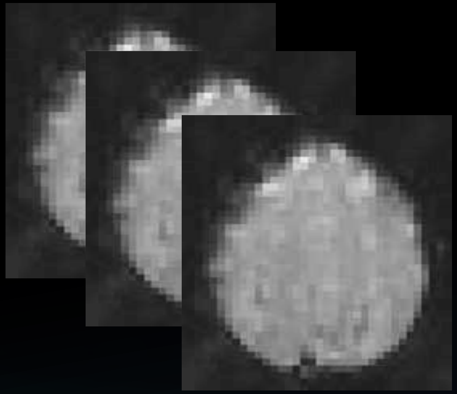
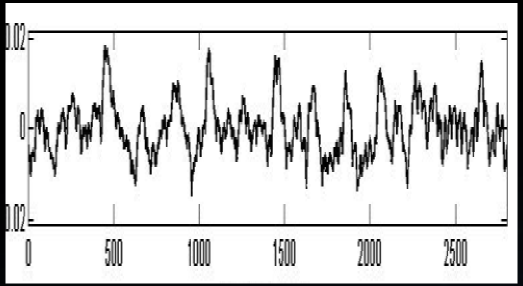


Image Processing Analysis Methods

IPA: Deconvolution

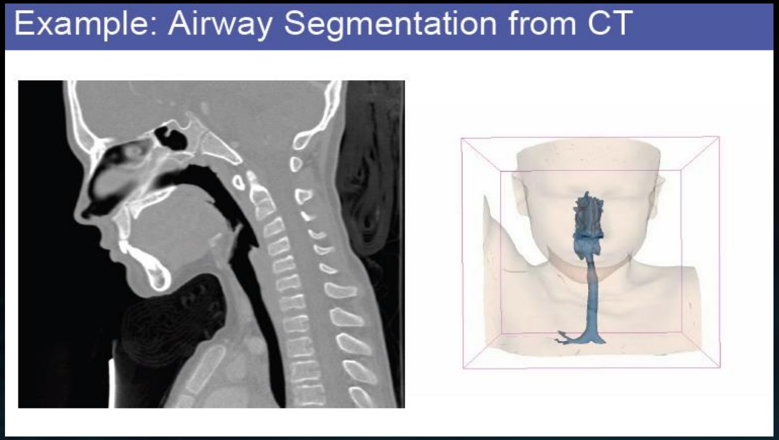
- Image Reconstruction Process
- Image Enhancement Process





IPA: Structural Learning

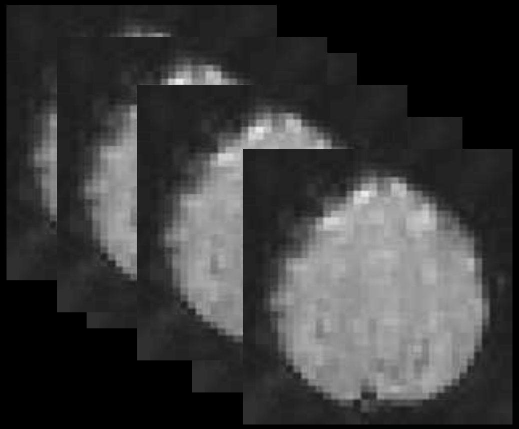
- Image Segmentation Process
- Image Registration Process



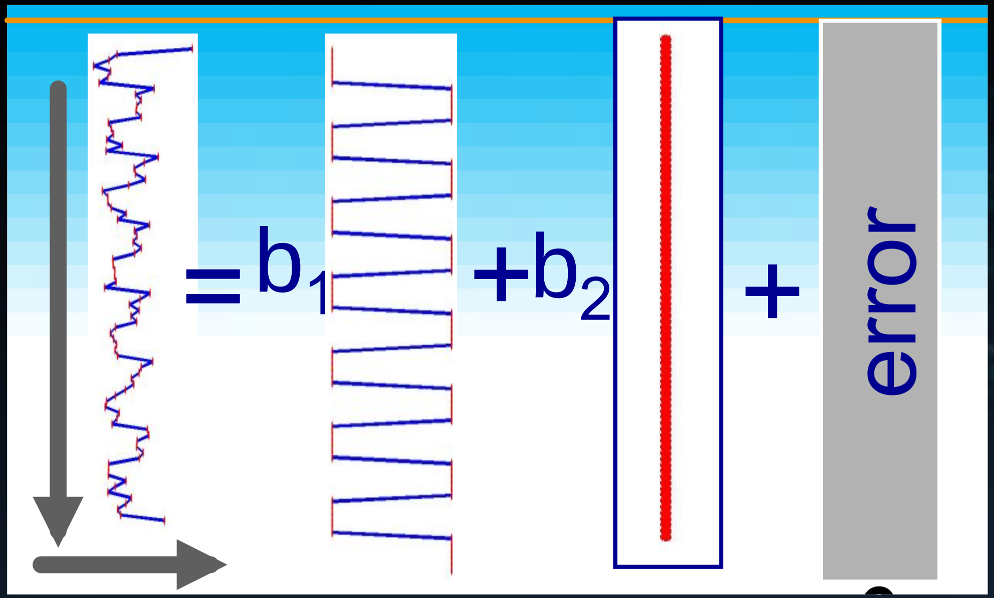
Example: Airway Segmentation from CT

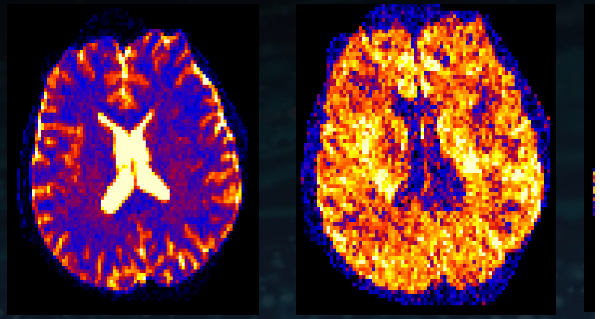
IPA: Deconvolution-IRP

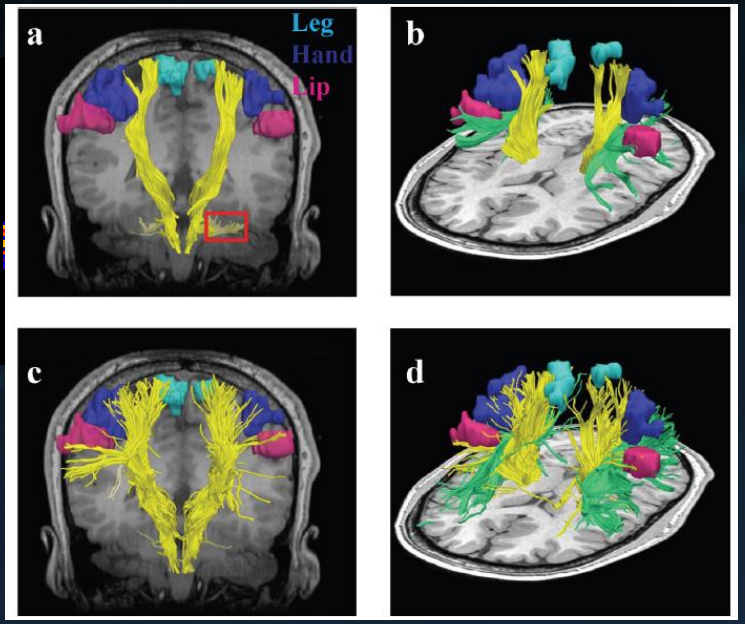
Image Reconstruction Process (IRP)



$$\frac{S(\mathbf{q})}{S_0} = \int P(\mathbf{r}, \Delta) e^{i\mathbf{q}.\mathbf{r}} d\mathbf{r}; \quad \mathbf{q} = \gamma \delta g \mathbf{u}$$



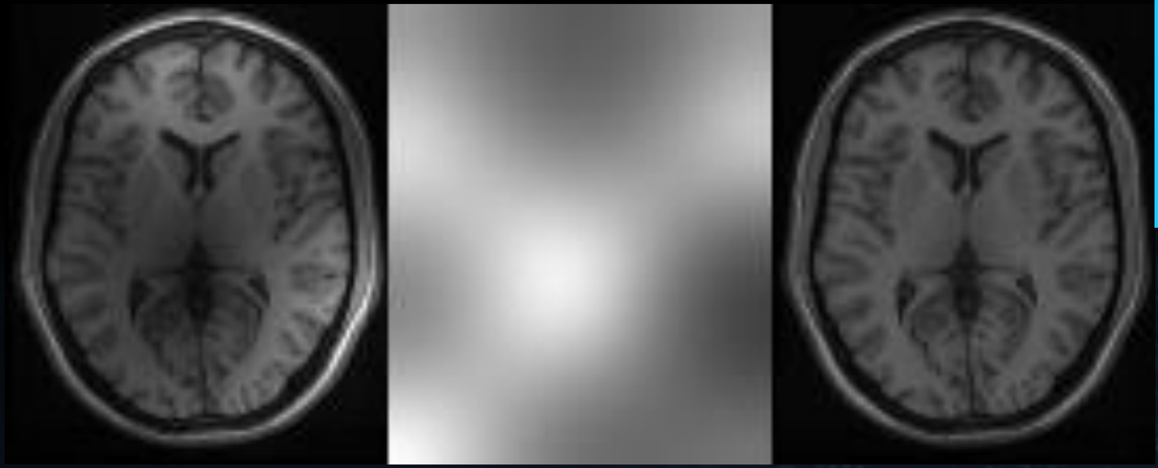


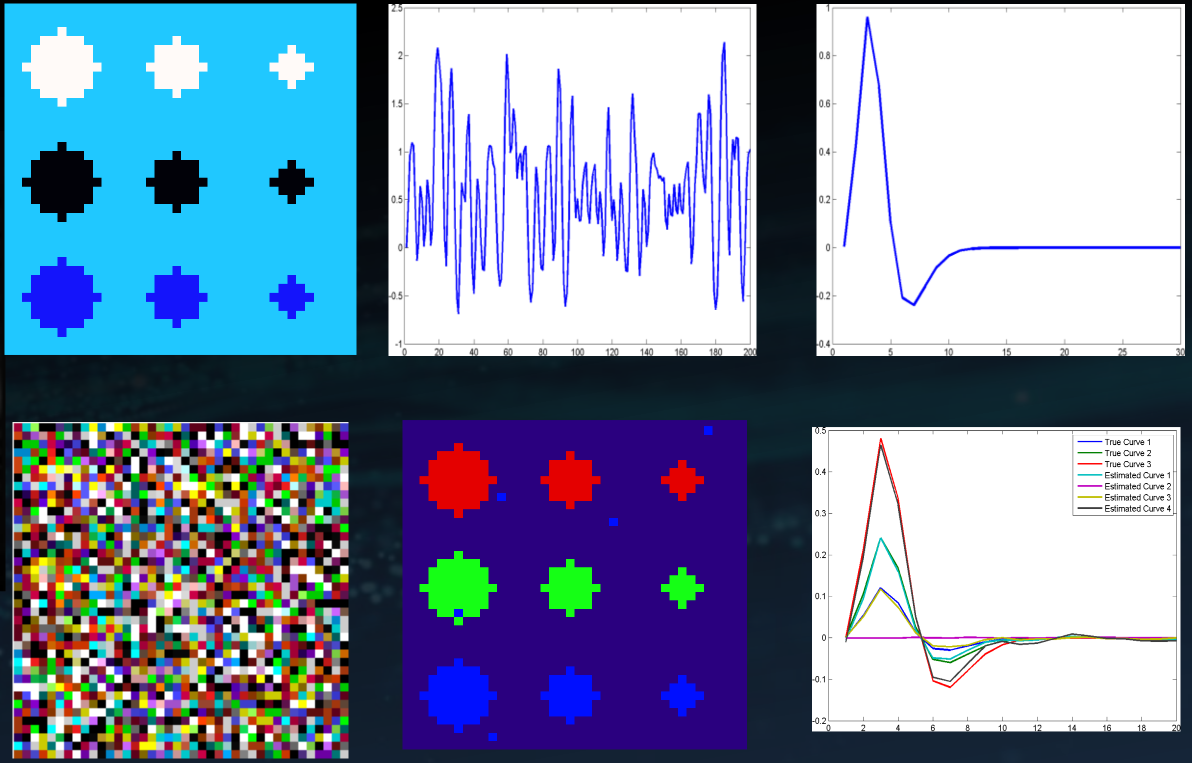


IPA: Deconvolution-IEP

Image Enhancement Process (IEP)

- ❖ Denoising
- ❖ Super-resolution
- ❖ Bias-field correction
- ❖ Harmonization

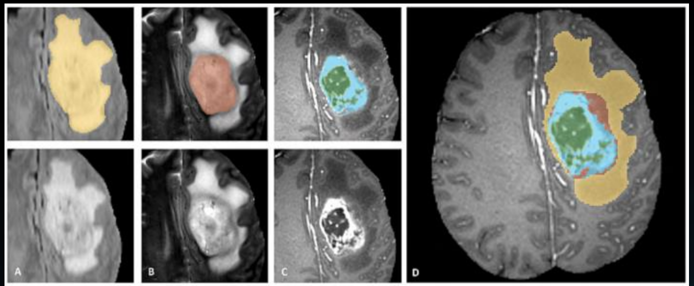


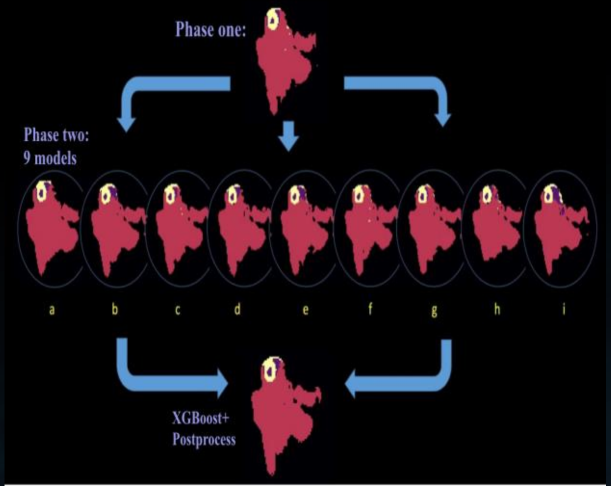


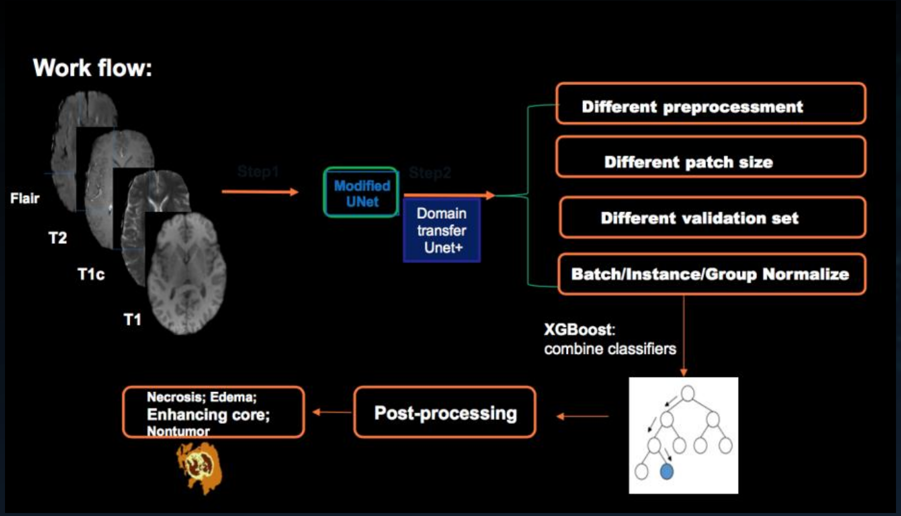
IPA: Structural Learning-ISP

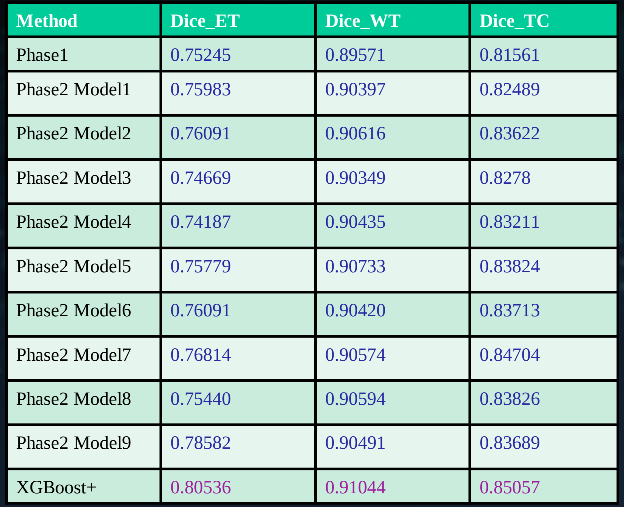
Image Segmentation Process (ISP)

- ❖ Quantification of brain development
- ❖ Localization of pathology
- ❖ Surgical planning
- ❖ Image-guided interventions
- ❖ Computer-aided detection and diagnosis
- ❖ Brain parcellation

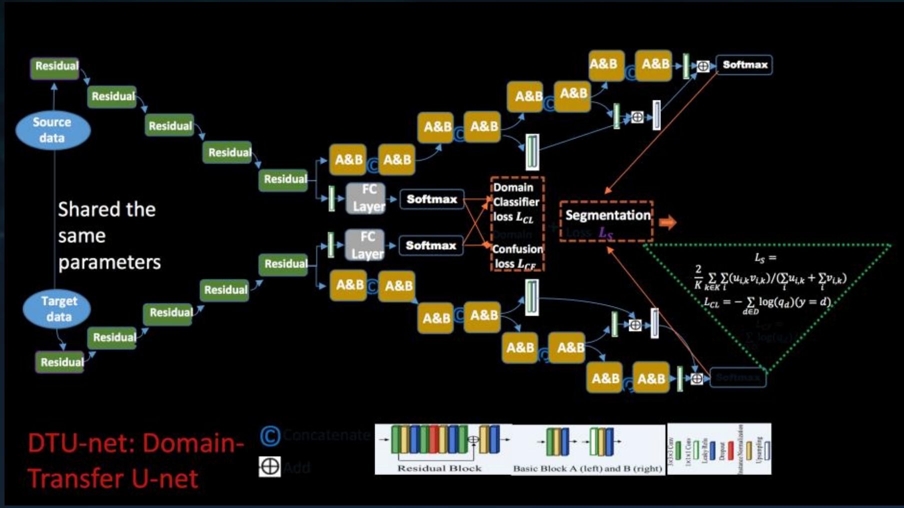








Method	Dice_ET	Dice_WT	Dice_TC
Phase1	0.75245	0.89571	0.81561
Phase2 Model1	0.75983	0.90397	0.82489
Phase2 Model2	0.76091	0.90616	0.83622
Phase2 Model3	0.74669	0.90349	0.8278
Phase2 Model4	0.74187	0.90435	0.83211
Phase2 Model5	0.75779	0.90733	0.83824
Phase2 Model6	0.76091	0.90420	0.83713
Phase2 Model7	0.76814	0.90574	0.84704
Phase2 Model8	0.75440	0.90594	0.83826
Phase2 Model9	0.78582	0.90491	0.83689
XGBoost+	0.80536	0.91044	0.85057

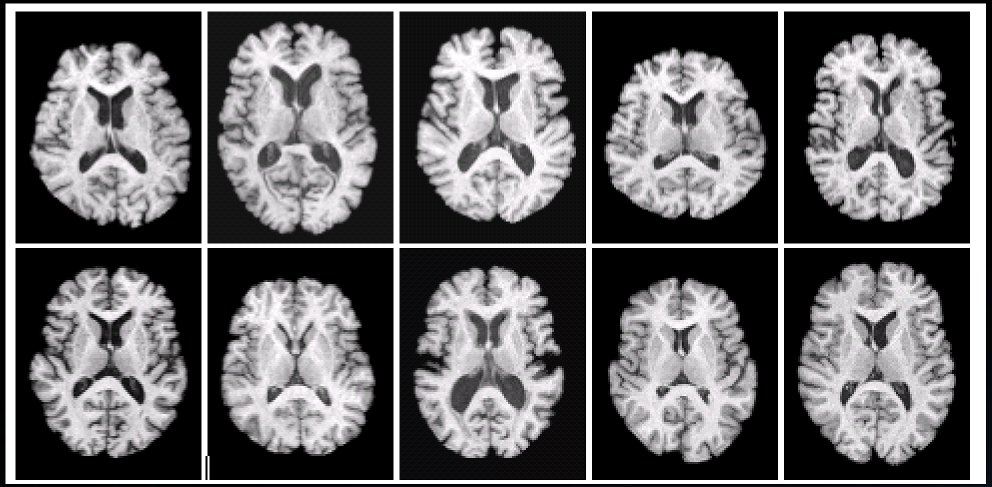

Source data
Target data
Residual
Shared the same parameters
A&B
FC Layer
Softmax
Domain Classifier loss L_{CL}
Domain Confusion loss L_{CF}
Segmentation Loss L_S
$L_S = \frac{2}{K}\sum_{k\in K}\sum_i (u_{i,k}v_{i,k})/(\sum_i u_{i,k}+\sum_i v_{i,k})$
$L_{CL} = -\sum_{d\in D}\log(q_d)(y=d)$
Concatenate
Add
Residual Block
Basic Block A (left) and B (right)
3×3×3 Conv
1×1×1 Conv
Leaky Relu
Dropout
Instance Normalization
Upsampling

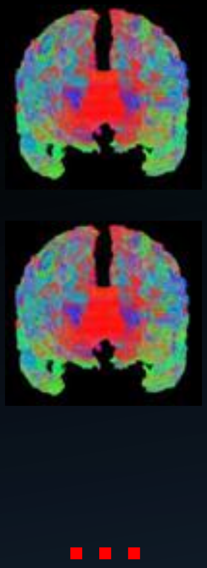
DTU-net: Domain-Transfer U-net

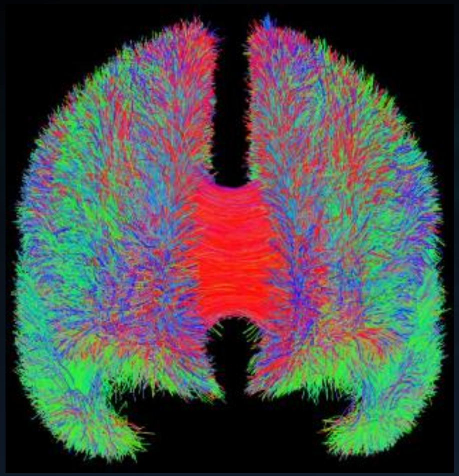
IPA: Structural Learning-IRP

Image Registration Process (IRP)

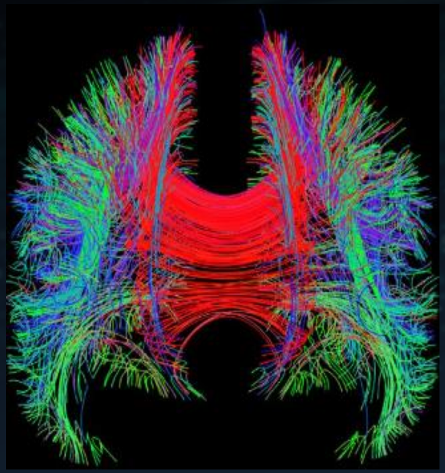
- Automated image segmentation
- Construction of brain atlas
- Localization of pathology
- Multimodal fusion
- Population analysis
- Quantification of brain development
- Shape analysis



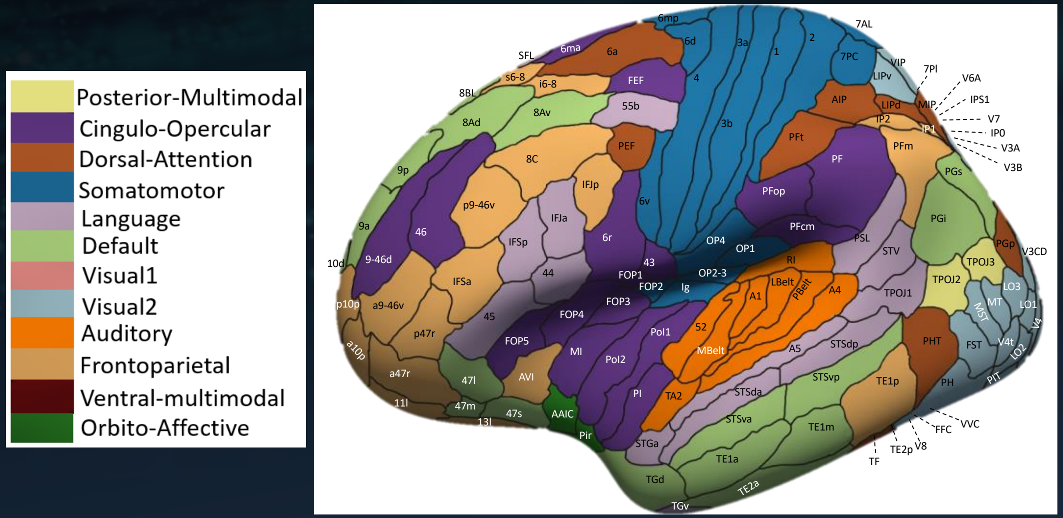




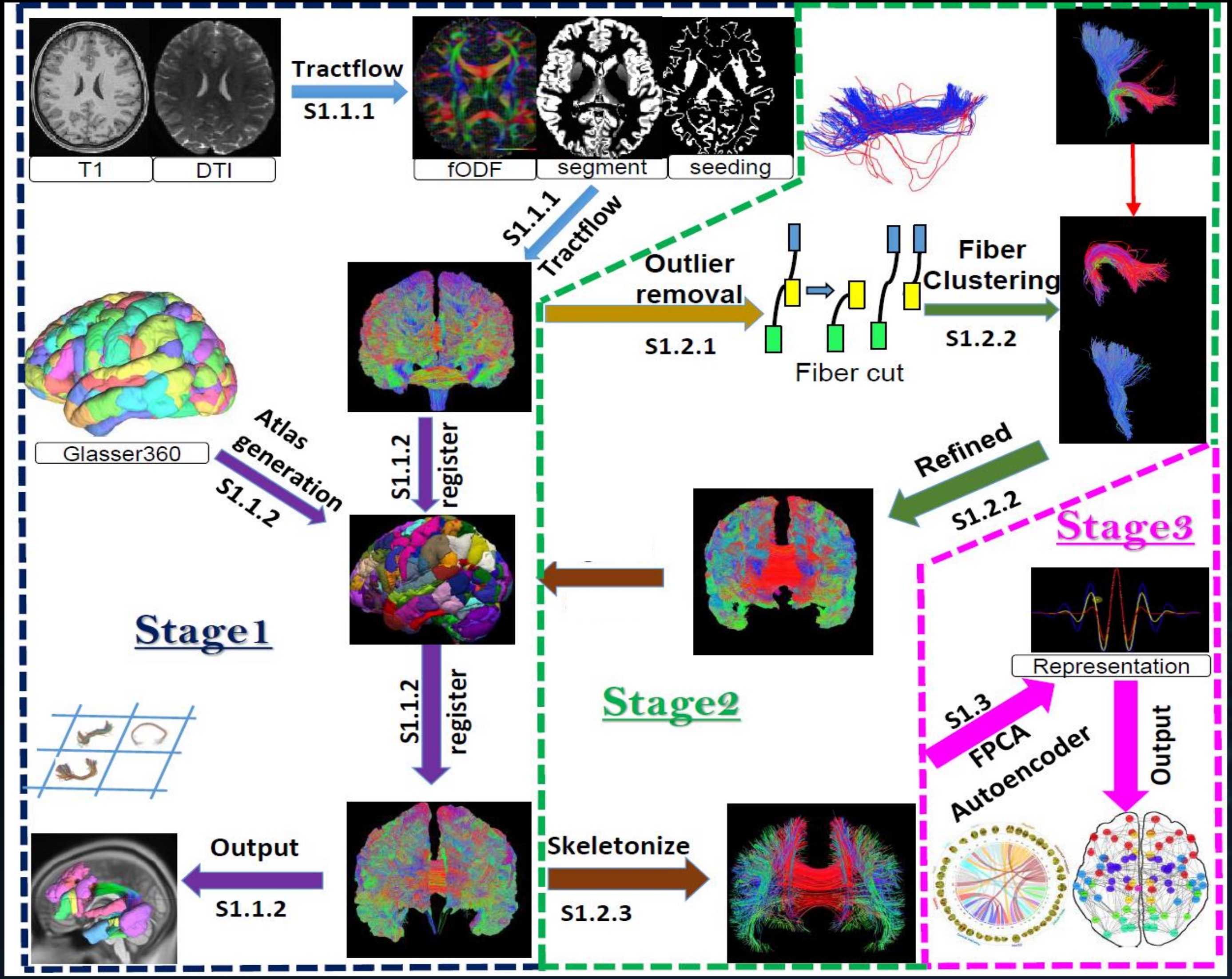
Fiber atlas



Fiber skeleton atlas



Brain Function-based Structural Connectome Atlas



Stage 1:

Whole-brain Structural Connectome

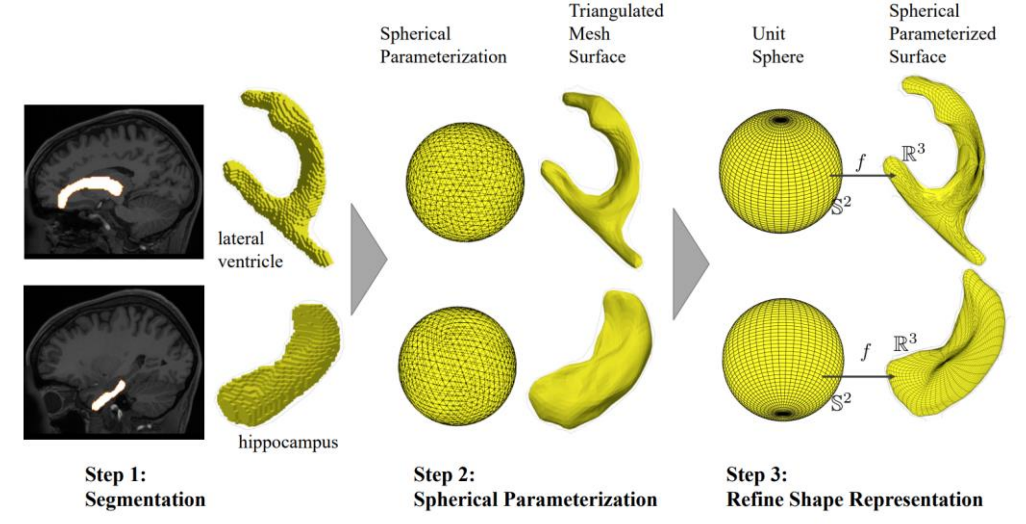
Stage 2:

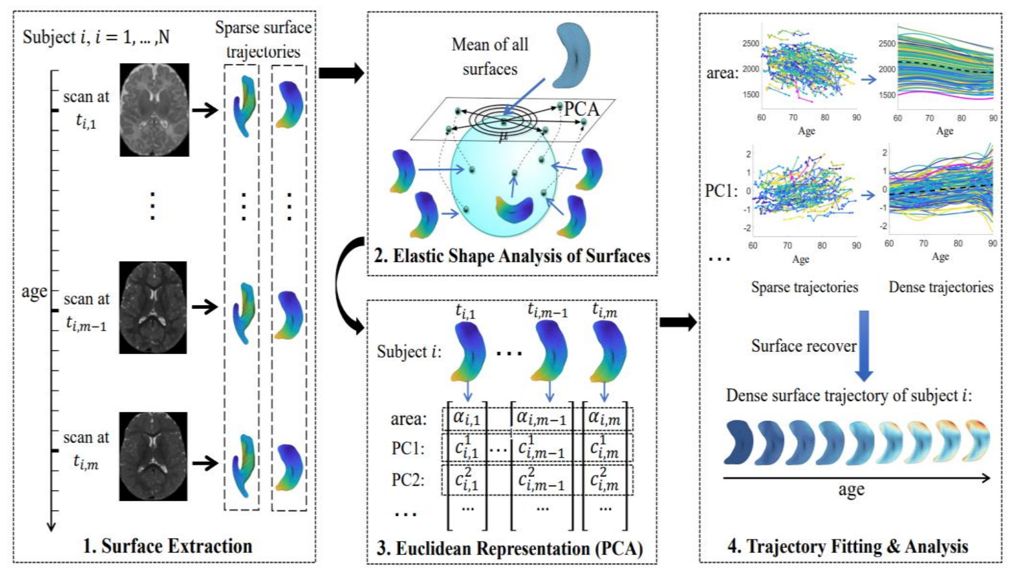
Creation of Fiber Skeleton

Stage 3:

Sparse Representation

Longitudinal Elastic Shape Data Analysis


Spherical Parameterization
Triangulated Mesh Surface
Unit Sphere
Spherical Parameterized Surface
lateral ventricle
hippocampus
f
$\mathbb{R}^3$
$\mathbb{S}^2$
Step 1: Segmentation
Step 2: Spherical Parameterization
Step 3: Refine Shape Representation



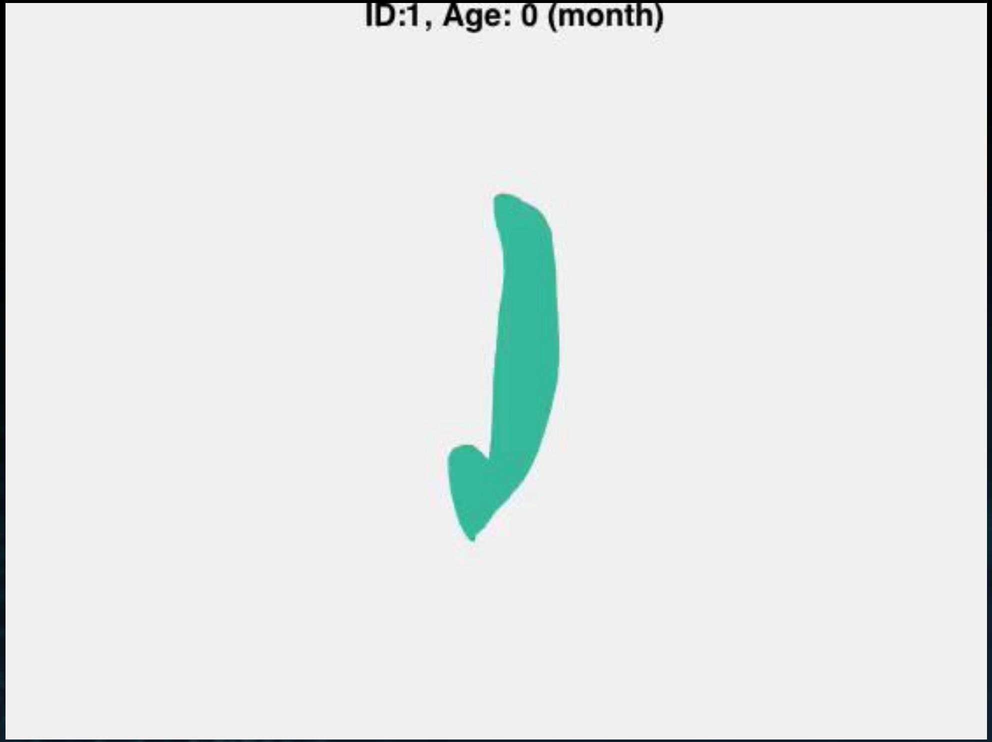


Figure 2: A schematic

Challenges

There is no publicly available, high-quality neuroimaging datasets with detailed annotation information that cover a large spectrum of segmentation tasks in neuroimaging research

Part III

Large-scale Neuroimaging related Studies

Large-scale Neuroimaging-related Studies

PING - 900 Pediatric Imaging, Neurocognition, and Genetics

BCP - 300 Baby Connectome Project

ADNI - 2000 Alzheimer's Disease Neuroimaging Initiative

PNC - 1400 Philadelphia Neurodevelopmental Cohort

HCP - 1200 Human Connectome Project

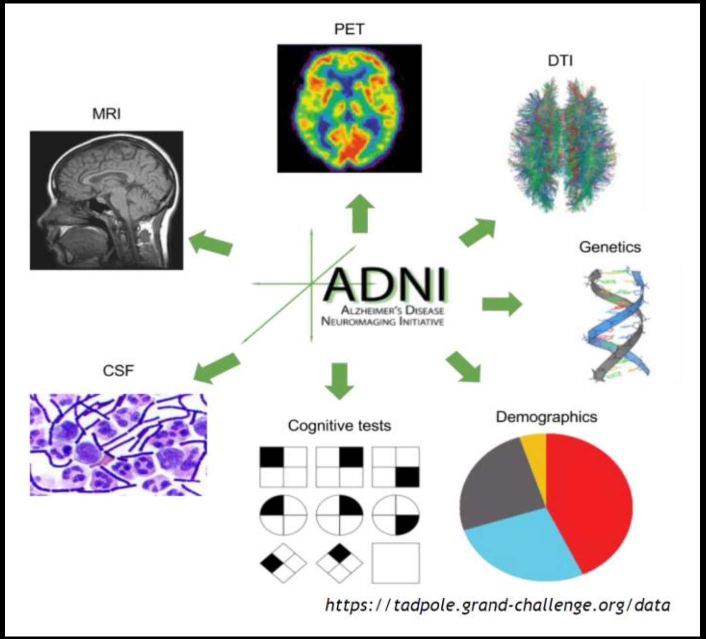
ABCD - 10000 Adolescent Brain Cognitive Development

UKB - 500,000 UK Biobank Project

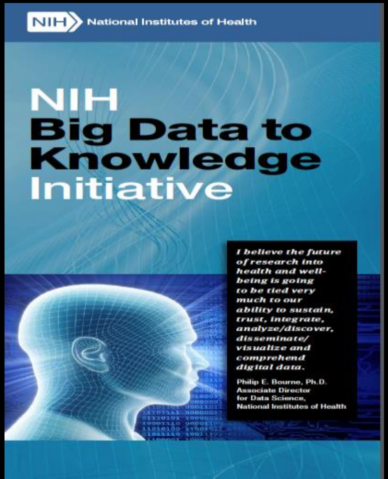
TCIA – 37,600 The Cancer Imaging Archive

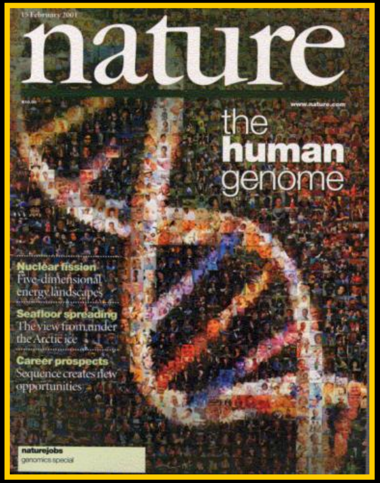
NLST - 19,000 National Lung Screening Trial

OAI – 4800 Osteoarthritis Initiative



https://tadpole.grand-challenge.org/data


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



biobank

Call Us On: 0800 0 276 276

Your feedback is important to us, tell us what you think

About | Participants | Resources | Scientists | Data Showcase | Register & Apply | Approved Research | Publications

UK Biobank: a global health resource

read more

UK Biobank: a global health resource

Genetic study targets smokers, lung disease

UK Biobank achievements so far

Participants

- Update your contact details
- BBC – challenge of saving lives with Big Data
- Find out how the resource is being used
- Participant Events

General Practice linkage →

Imaging study →

UK Biobank Annual Meeting

Scientists

- Imaging data on 5,000 now available
- UK Biobank's cardiovascular MR protocol
- Video: How to Register and Apply
- UK Biobank Annual Meeting 2016

News

Health and thinking skills linked to same genes, study shows

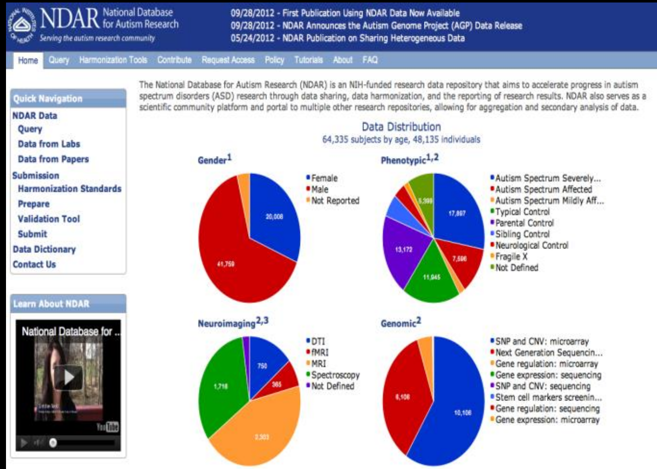
A broken bone may lead to widespread body pain

New appointments to the UK Biobank Ethics & Governance Council

View the video: Study finds red meat link to bowel cancer

Impact of pregnancy and birth on future health

Genetic study targets lung disease and smoking behaviour



NDAR National Database for Autism Research

Serving the autism research community

09/28/2012 - First Publication Using NDAR Data Now Available
09/28/2012 - NDAR Announces the Autism Genome Project (AGP) Data Release
05/24/2012 - NDAR Publication on Sharing Heterogeneous Data

Home | Query | Harmonization Tools | Contribute | Request Access | Policy | Tutorials | About | FAQ

Quick Navigation

- NDAR Data
  - Query
  - Data from Labs
  - Data from Papers
- Submission
  - Harmonization Standards
  - Prepare
  - Validation Tool
  - Submit
- Data Dictionary
- Contact Us

Learn About NDAR

The National Database for Autism Research (NDAR) is an NIH-funded research data repository that aims to accelerate progress in autism spectrum disorders (ASD) research through data sharing, data harmonization, and the reporting of research results. NDAR also serves as a scientific community platform and portal to multiple other research repositories, allowing for aggregation and secondary analysis of data.

Data Distribution

64,335 subjects by age, 48,135 individuals

Gender[1]
- Female
- Male
- Not Reported

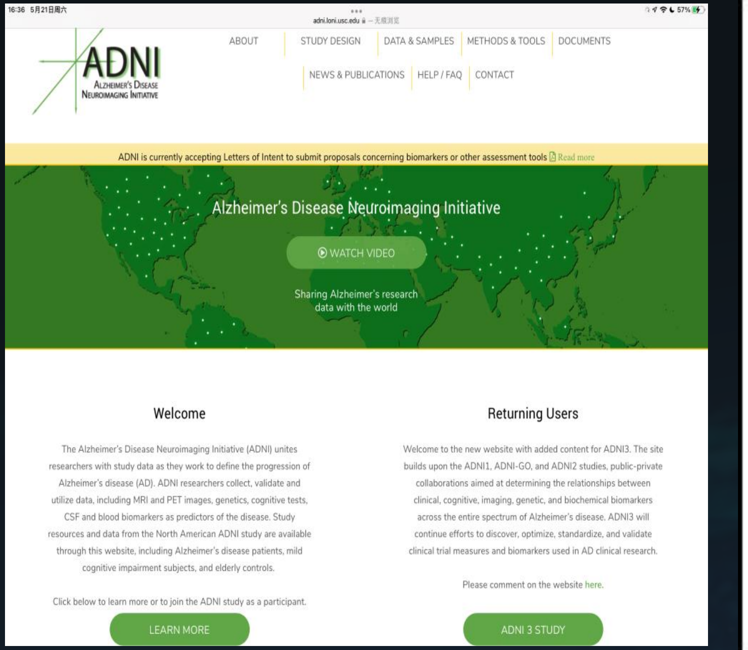
Phenotypic[1,2]
- Autism Spectrum Severely...
- Autism Spectrum Affected
- Autism Spectrum Mildly Aff...
- Typical Control
- Parental Control
- Sibling Control
- Neurological Control
- Fragile X
- Not Defined

Neuroimaging[2,3]
- DTI
- fMRI
- MRI
- Spectroscopy
- Not Defined

Genomic[2]
- SNP and CNV: microarray
- Next Generation Sequencin...
- Gene regulation: microarray
- Gene expression: sequencing
- SNP and CNV: sequencing
- Stem cell markers screenin...
- Gene regulation: sequencing
- Gene expression: microarray

Alzheimer’s Disease Neuroimaging Initiative

The overall goal of ADNI is to validate potentially useful biomarkers for AD clinical treatment trials. ADNI is a multisite, prospective clinical study and actively supports the investigation and development of treatments that may slow or stop the progression of AD https://adni.loni.usc.edu/study-design. Researchers across 63 sites in the US and Canada have been tracking the progression of AD through clinical, imaging, genetic and biospecimen biomarkers, starting from normal aging, early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI) to dementia or AD.



ADNI is currently accepting Letters of Intent to submit proposals concerning biomarkers or other assessment tools Read more

Alzheimer’s Disease Neuroimaging Initiative

Sharing Alzheimer’s research data with the world

Welcome

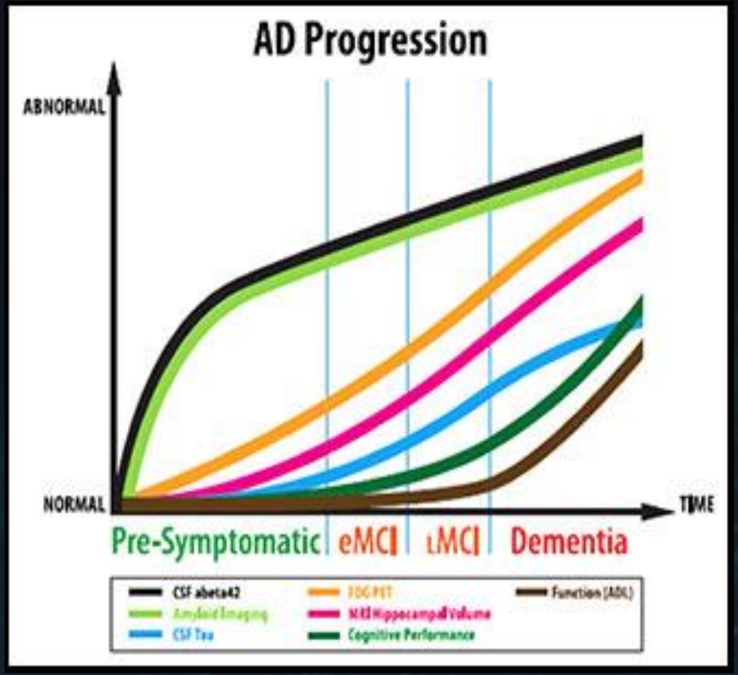
The Alzheimer’s Disease Neuroimaging Initiative (ADNI) unites researchers with study data as they work to define the progression of Alzheimer’s disease (AD). ADNI researchers collect, validate and utilize data, including MRI and PET images, genetics, cognitive tests, CSF and blood biomarkers as predictors of the disease. Study resources and data from the North American ADNI study are available through this website, including Alzheimer’s disease patients, mild cognitive impairment subjects, and elderly controls.

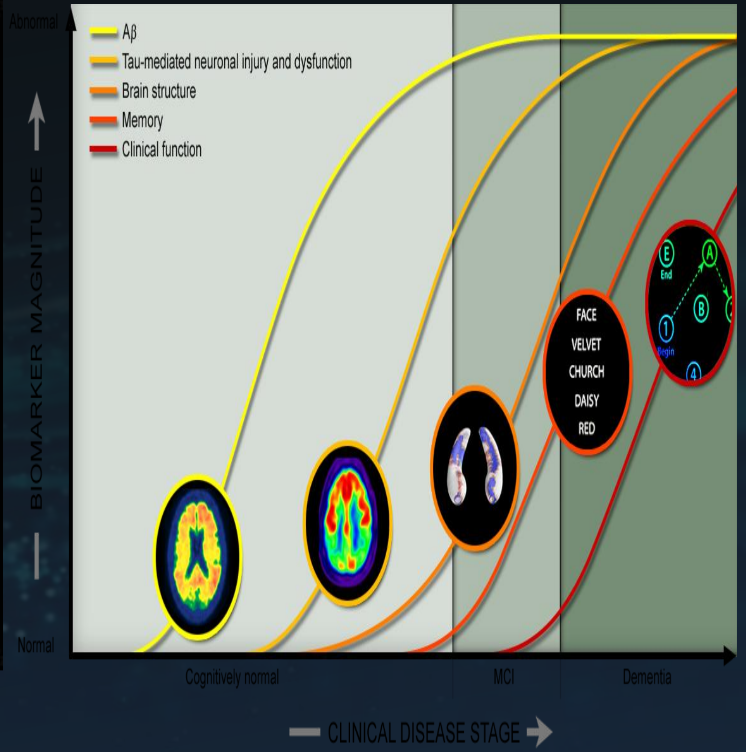
Click below to learn more or to join the ADNI study as a participant.

Returning Users

Welcome to the new website with added content for ADNI3. The site builds upon the ADNI1, ADNI-GO, and ADNI2 studies, public-private collaborations aimed at determining the relationships between clinical, cognitive, imaging, genetic, and biochemical biomarkers across the entire spectrum of Alzheimer’s disease. ADNI3 will continue efforts to discover, optimize, standardize, and validate clinical trial measures and biomarkers used in AD clinical research.

Please comment on the website here.





2004-now

The Human Connectome Project and Beyond

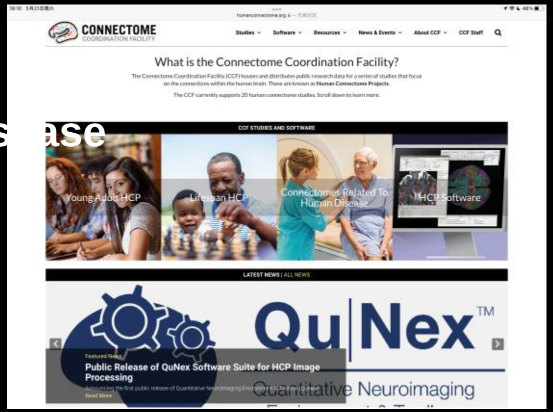
The primary goals of HCP include

- building a ``network map'' that will shed light on the anatomical and functional connectivity within the healthy human brain,
- promoting the understanding of inter-individual variability of brain circuits to behavior,
- facilitating research into brain disorders, such as autism, AD, and schizophrenia, and
- making all data freely available to the scientific community.

The Heavily Connected Brain
Peter Stern, "Connection, connection, connection…", Science, Nov. 1 2013: Vol. 342 no. 6158 P.577

- The NIH Human Connectome Project
  - The Harvard/MGH-UCLA project
  - The WU-Minn Project
- The EU's 7th Framework Programme for Research
  - Consortium Of Neuroimagers for the Non-Invasive Exploration of Brain Connectivity and Tracts

- Healthy Adult Connectome
- Lifespan Connectome Data
- Connectomes related to Human Disease



What is the Connectome Coordination Facility?

The Connectome Coordination Facility (CCF) houses and distributes public research data for a series of studies that focus on the connections within the human brain. These are known as **Human Connectome Projects.**

The CCF currently supports 20 human connectome studies. Scroll down to learn more.

Featured News

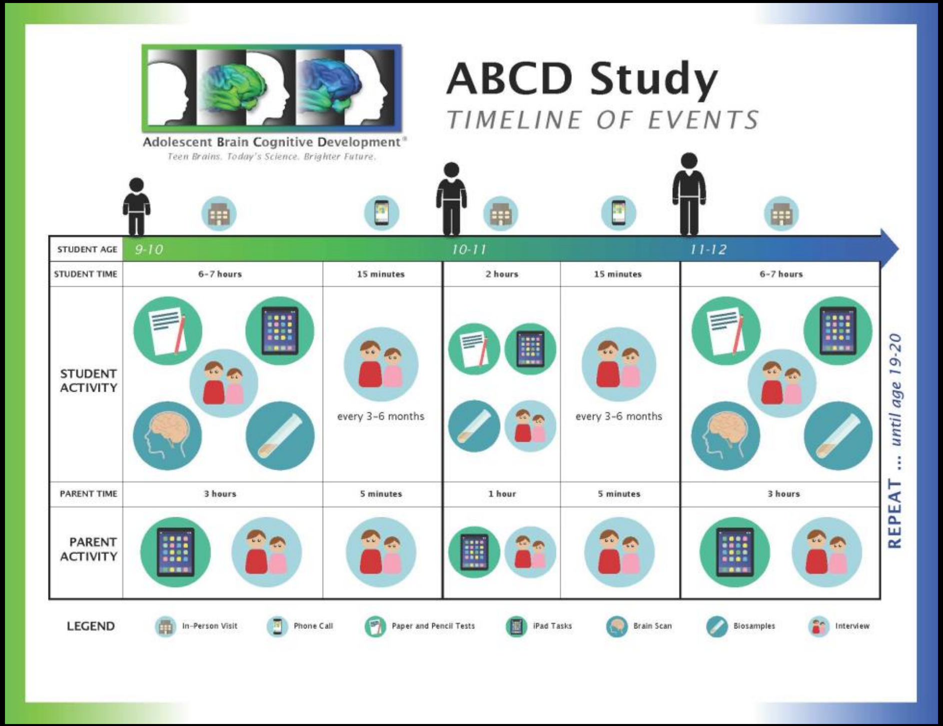
Public Release of QuNex Software Suite for HCP Image Processing

Read More

Adolescent Brain Cognitive Development

The ABCD study is the largest prospective longitudinal study of brain development and child health in the United States, which has recruited approximately 11,880 children aged 9-10 years old from 21 research sites and is following them for 10 years into early adulthood.

Its initial goal was to examine risk and resiliency factors associated with the development of substance use, and then expanded far beyond, into identifying the underlying biospecimens, neural alterations, and environmental factors, and their contributions to the development of behavior, brain function, and other mental and physical outcomes throughout adolescence.



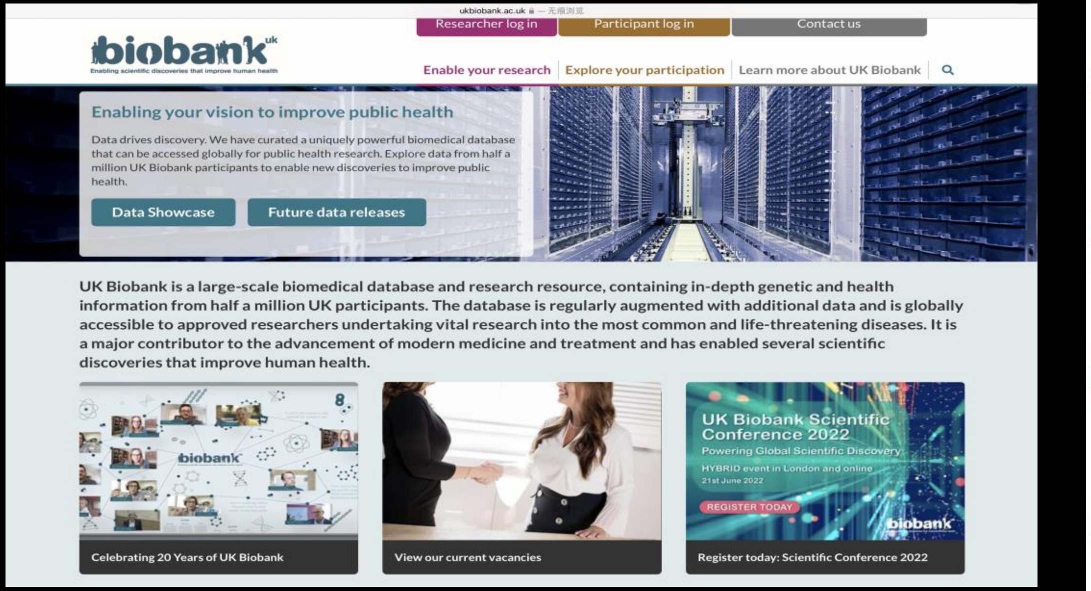
STUDENT AGE	9-10		10-11		11-12
STUDENT TIME	6–7 hours	15 minutes	2 hours	15 minutes	6–7 hours
PARENT TIME	3 hours	5 minutes	1 hour	5 minutes	3 hours

2015-now

https://abcdstudy.org/

The UK Biobank Study

UK Biobank has collected and continues to collect extensive environmental, lifestyle, and genetic data on half a million participants.


ukbiobank.ac.uk — 无痕浏览
Researcher log in
Participant log in
Contact us
biobank uk
Enabling scientific discoveries that improve human health
Enable your research
Explore your participation
Learn more about UK Biobank
Data Showcase
Future data releases
biobank
UK Biobank Scientific Conference 2022
Powering Global Scientific Discovery
HYBRID event in London and online
21st June 2022
REGISTER TODAY
biobank

Enabling your vision to improve public health

Data drives discovery. We have curated a uniquely powerful biomedical database that can be accessed globally for public health research. Explore data from half a million UK Biobank participants to enable new discoveries to improve public health.

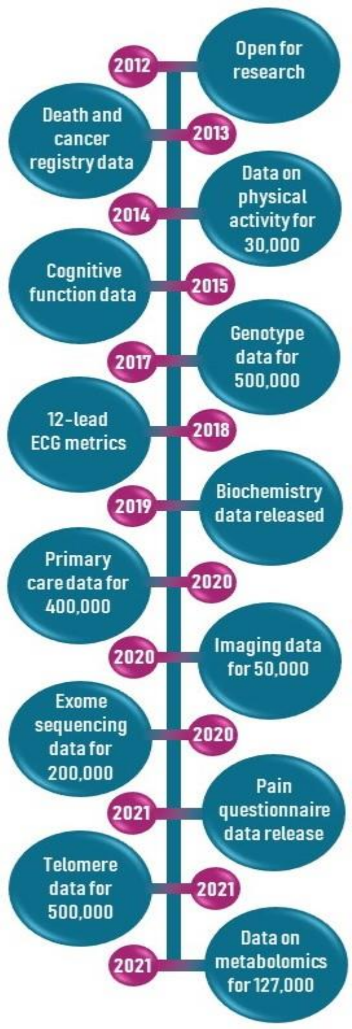
UK Biobank is a large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK participants. The database is regularly augmented with additional data and is globally accessible to approved researchers undertaking vital research into the most common and life-threatening diseases. It is a major contributor to the advancement of modern medicine and treatment and has enabled several scientific discoveries that improve human health.

Celebrating 20 Years of UK Biobank

View our current vacancies

Register today: Scientific Conference 2022

2006-now



2012 Open for research

2013 Death and cancer registry data

2014 Data on physical activity for 30,000

2015 Cognitive function data

2017 Genotype data for 500,000

2018 12-lead ECG metrics

2019 Biochemistry data released

2020 Primary care data for 400,000

2020 Imaging data for 50,000

2020 Exome sequencing data for 200,000

2021 Pain questionnaire data release

2021 Telomere data for 500,000

2021 Data on metabolomics for 127,000

- **Imaging:** Brain, heart and full body MR imaging, plus full body DEXA scan of the bones and joints and an ultrasound of the carotid arteries. The goal is to image 100,000 participants, and to invite participants back for a repeat scan some years later.
- **Genetics:** Genotyping, whole exome sequencing & whole genome sequencing for all participants.
- **Health linkages:** Linkage to a wide range of electronic health-related records, including death, cancer, hospital admissions and primary care records.
- **Biomarkers:** Data on more than 30 key biochemistry markers from all participants, taken from samples collected at recruitment and the first repeat assessment.
- **Activity monitor:** Physical activity data over a 7-day period collected via a wrist-worn activity monitor for 100,000 participants plus a seasonal follow-up on a subset.
- **Online questionnaires:** Data on a range of exposures and health outcomes that are difficult to assess via routine health records, including diet, food preferences, work history, pain, cognitive function, digestive health and mental health.
- **Repeat baseline assessments:** A full baseline assessment is undertaken during the imaging assessment of 100,000 participants.
- **Samples:** Blood & urine was collected from all participants, and saliva for 100,000.

ENIGMA

The major goals of ENIGMA include

- pushing forward the field of imaging genetics,

❖ ensuring promising and reproducible findings,

➢ sharing data, ideas, methods, algorithms and other information, and

- training new investigators.

https://enigma.ini.usc.edu/

2009-now

The Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) Consortium is a global alliance of over 1,400 scientists across 43 countries in the fields of imaging genomics, neurology, and psychiatry, studying a range of large-scale human brain studies that integrate data based on sMRI, DWI, fMRI, genetic data and many patient populations from over 70 institutions worldwide

Part IV

Population-based Statistical Analysis Methods

Four Common Themes (CT5)-(CT8)

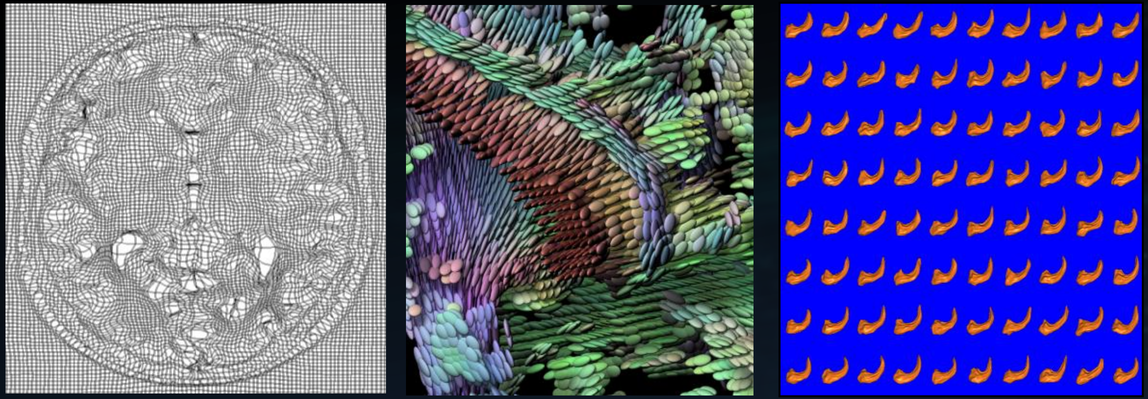
(CT5) Sampling Bias

- undercoverage
- observer bias,
- voluntary response bias
- survivorship bias
- recall bias
- exclusion bias

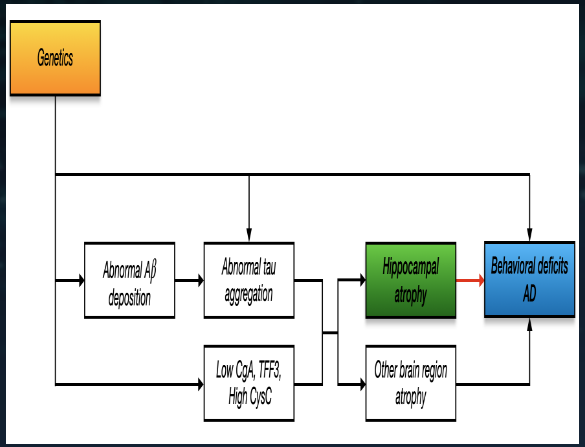
(CT6) Complex Missing Data Patterns

- missing by design
- faulty scanning
- attrition in longitudinal studies
- mis-entry
- non-responses in surveys

(CT7) Complex Data Objects

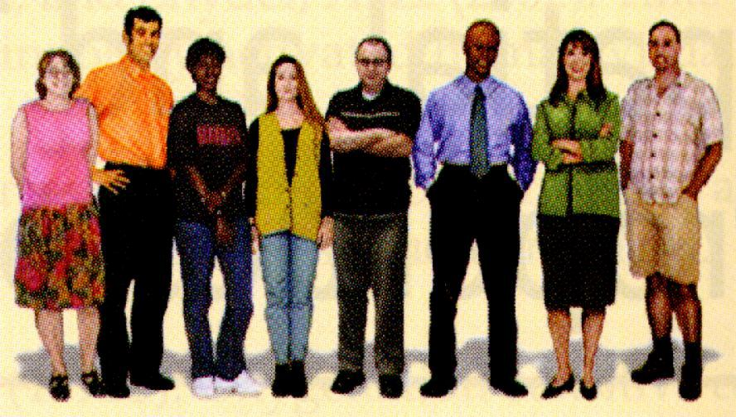


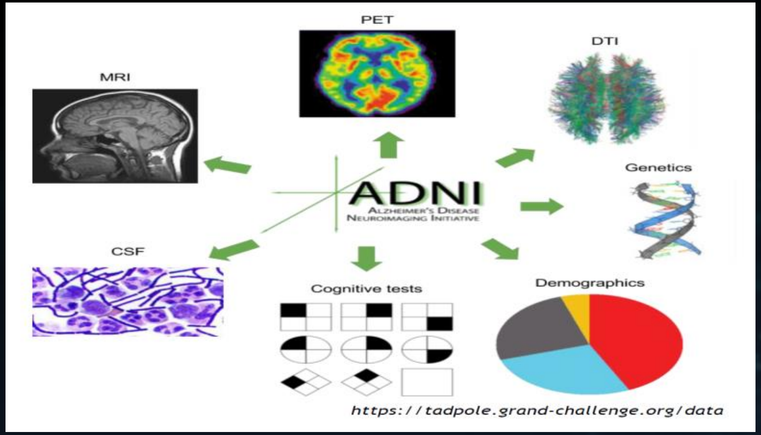
(CT8) Complicated Causal Pathways in Brain-related Disorders



Population-based Statistical Analysis (PSA)

- Study Design
- Statistical Parametric Mapping
- Object Oriented Data (OOD) Analysis
- Imputation Methods
- Data Integration Methods





https://tadpole.grand-challenge.org/data

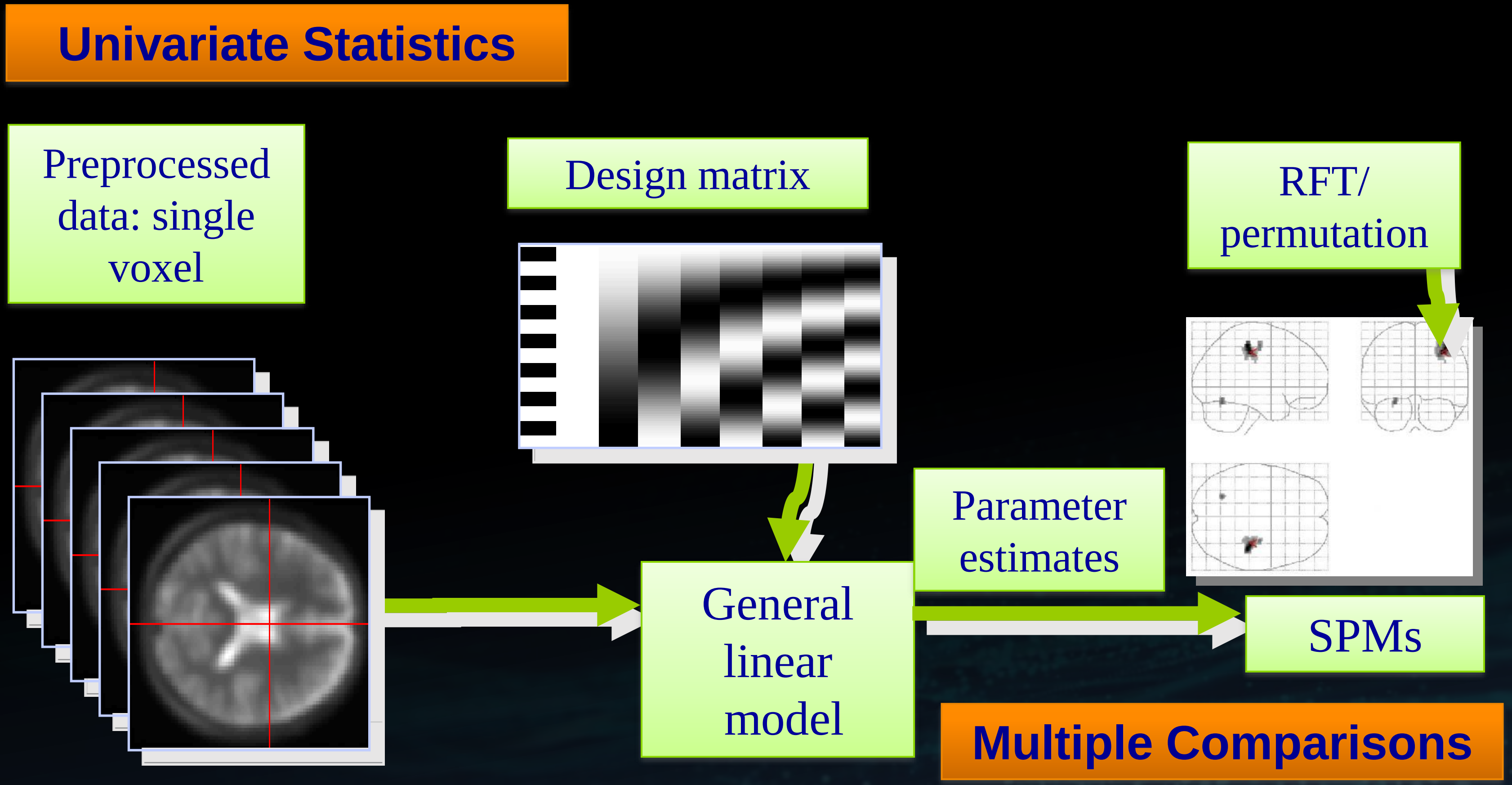
- Dimension Reduction Methods
- Image Genetics
- Causality Research
- Predictive Analysis
- Knowledge-based Methods

Study Design

- **Case control study**
- **Cross-sectional study**
- **Cohort study**
- **Experimental study**
- **Descriptive Study: case reports, case series, Descriptive surveys.**

- The UKB is a large, population-based cohort study, and many cross-sectional analyses have been conducted based on baseline data from UKB.
- The UKB is well known for its ”healthy volunteer” selection bias, and may not be a true representation of the general population.
- Neuroimaging biomarkers are usually secondary outcome.

Statistical Parametric Mapping



Univariate Statistics

Preprocessed data: single voxel

Design matrix

General linear model

Parameter estimates

RFT/ permutation

SPMs

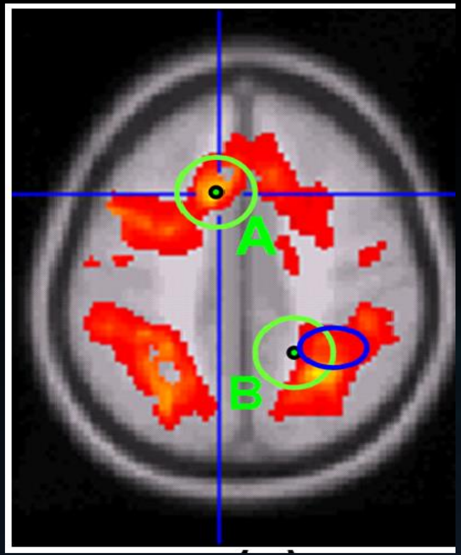
Multiple Comparisons

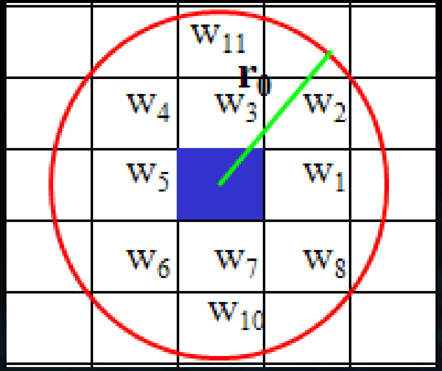
- More complex models
- Multiple comparisons

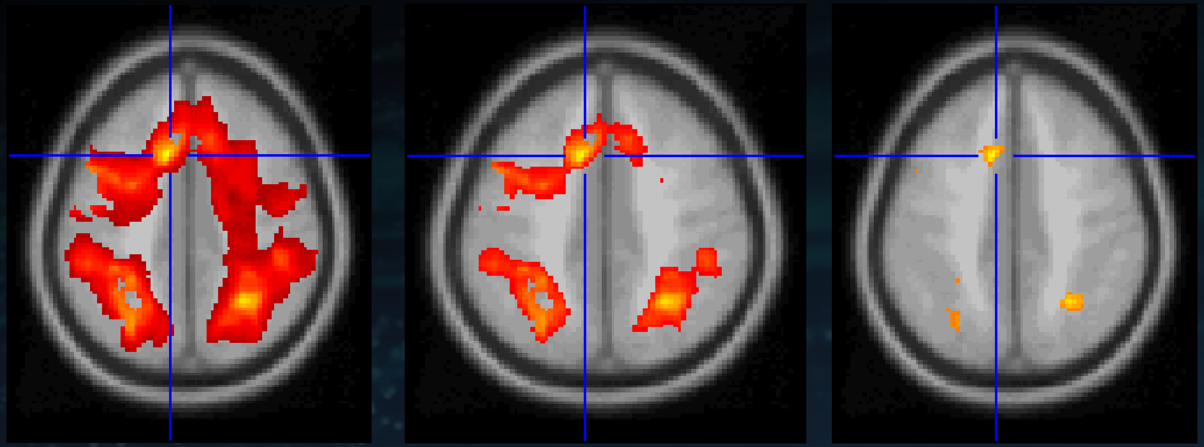
Statistical Parametric Mapping

- From voxel-wise models to functional models
- Multiscale-adaptive estimation and inference procedures
- Wild-bootstrap methods to correct for multiple comparisons

Image=***f***(*B(age, gene, race, disease, others)*, device, acquisition, noises)

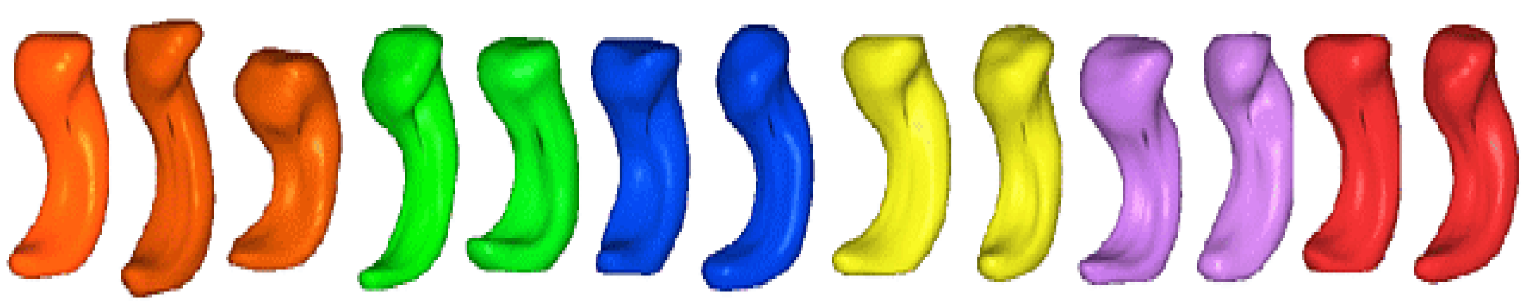


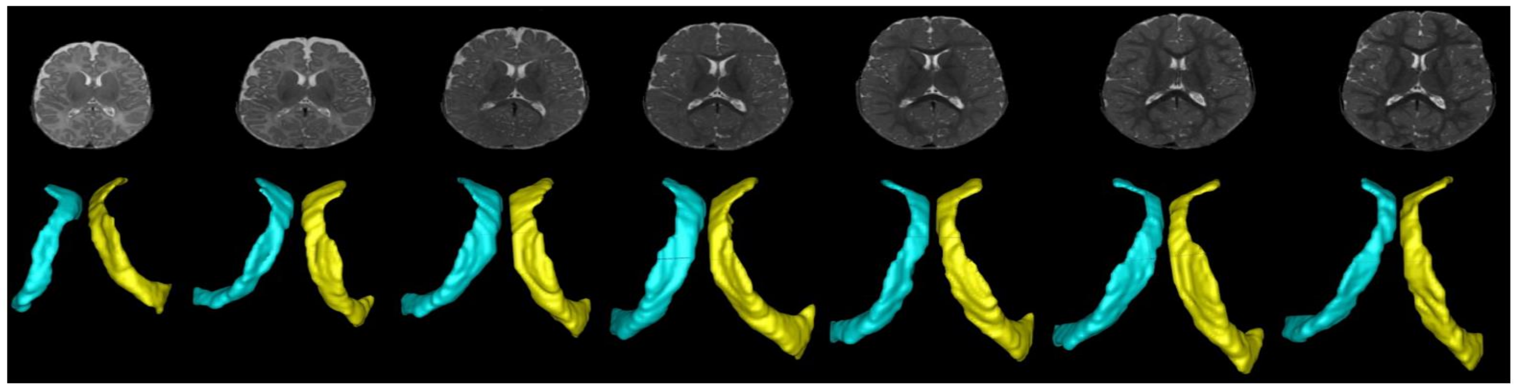




OOD analyses

Parametric, Semiparametric and Nonparametric Models for OOD analyses





$$= g(x, \theta, f) \oplus \varepsilon$$

$$x \in R^k, \theta \in \Theta \subset R^p, f \in F$$

$$g : R^k \times R^p \times F \to M$$

Intrinsic Regression Models

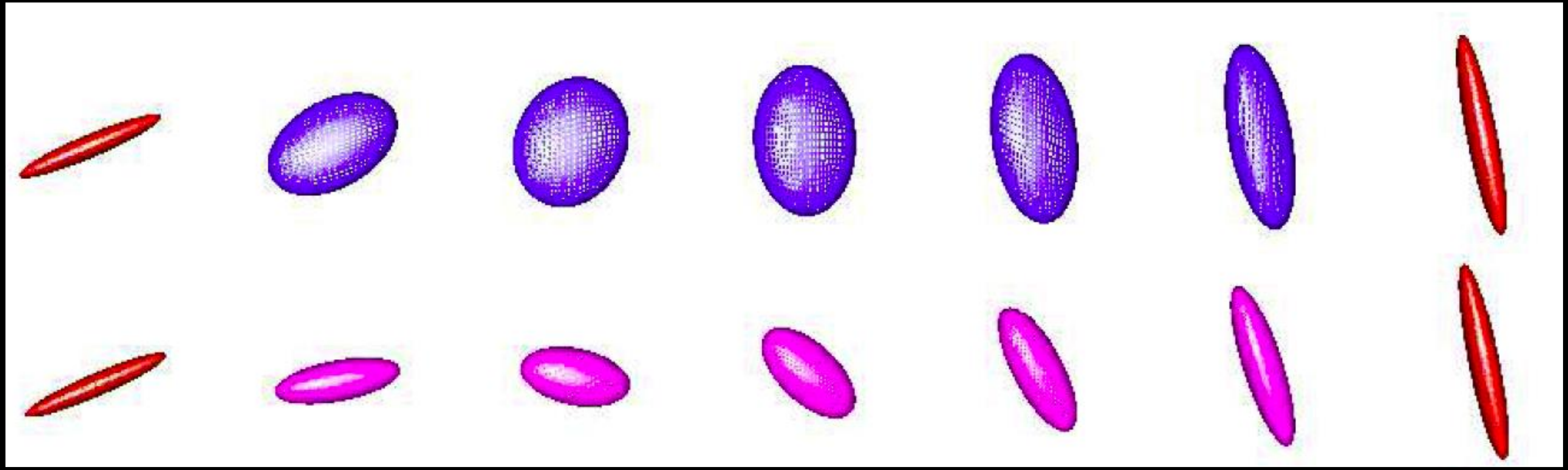
- **Feature Methods:** Use some feature extraction functions to project random objects to Euclidean-valued variables.

- **Extrinsic Methods:** Ignore the fact that manifold-valued data are in a nonlinear space and then directly apply classical multivariate regression.

- **Intrinsic Methods:** few parametric models for manifold-valued data .

Extrinsic Methods

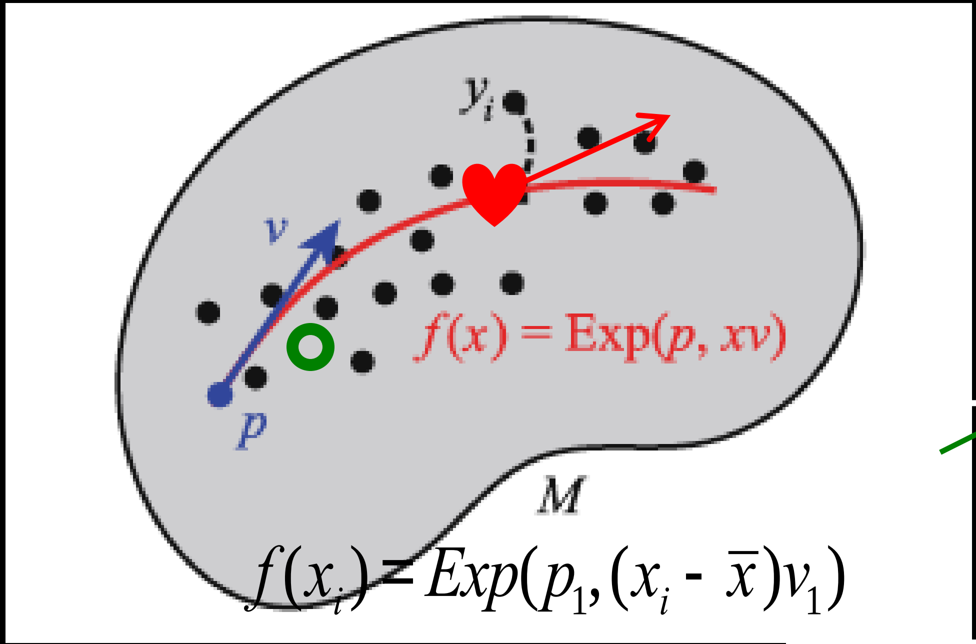
Intrinsic Methods



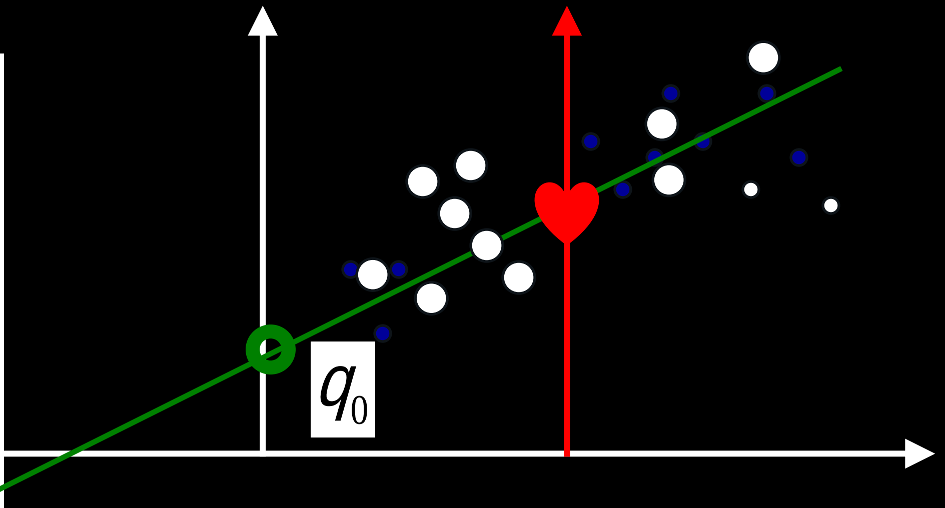
Dryden, I.L., Koloydenko, A. and Zhou, D. (2008).

Intrinsic Regression Models

Geodesic Link Function



$$f(x_i) = Exp(p_1, (x_i - \bar{x})v_1)$$



$$g(x_i, q) = q_0 + x_i q_1 = q_0 + \bar{x} q_1 + (x_i - \bar{x}) q_1$$

$$f(x_i) = Exp(p_1, (x_i - \bar{x})v_1)$$

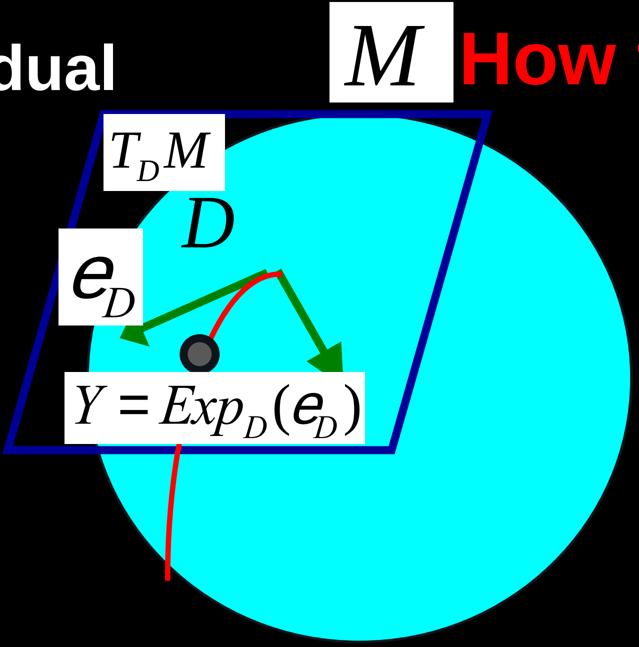
$$f(x_i) = Exp(p_1, \underset{k}{\text{å}} (x_{ik} - \bar{x}_k) v_k)$$

Single-center

Fletcher (2013)
Maxwell et al. (2014)

Intrinsic Regression Models

Residual



How to define residual?

Inner product $T_D M$

$$<< e_D, \tilde{e}_D >>$$

Geodesic $g_D(t; e_D)$

Riemannian exponential maps

$$Y = Exp_D(e_D) = g_D(1; e_D)$$

Riemannian logarithm maps

$$e_D = \mathrm{Log}_D(Y) \in B(0, r) \subset T_D M$$

radius of injectivity

Intrinsic Regression Models

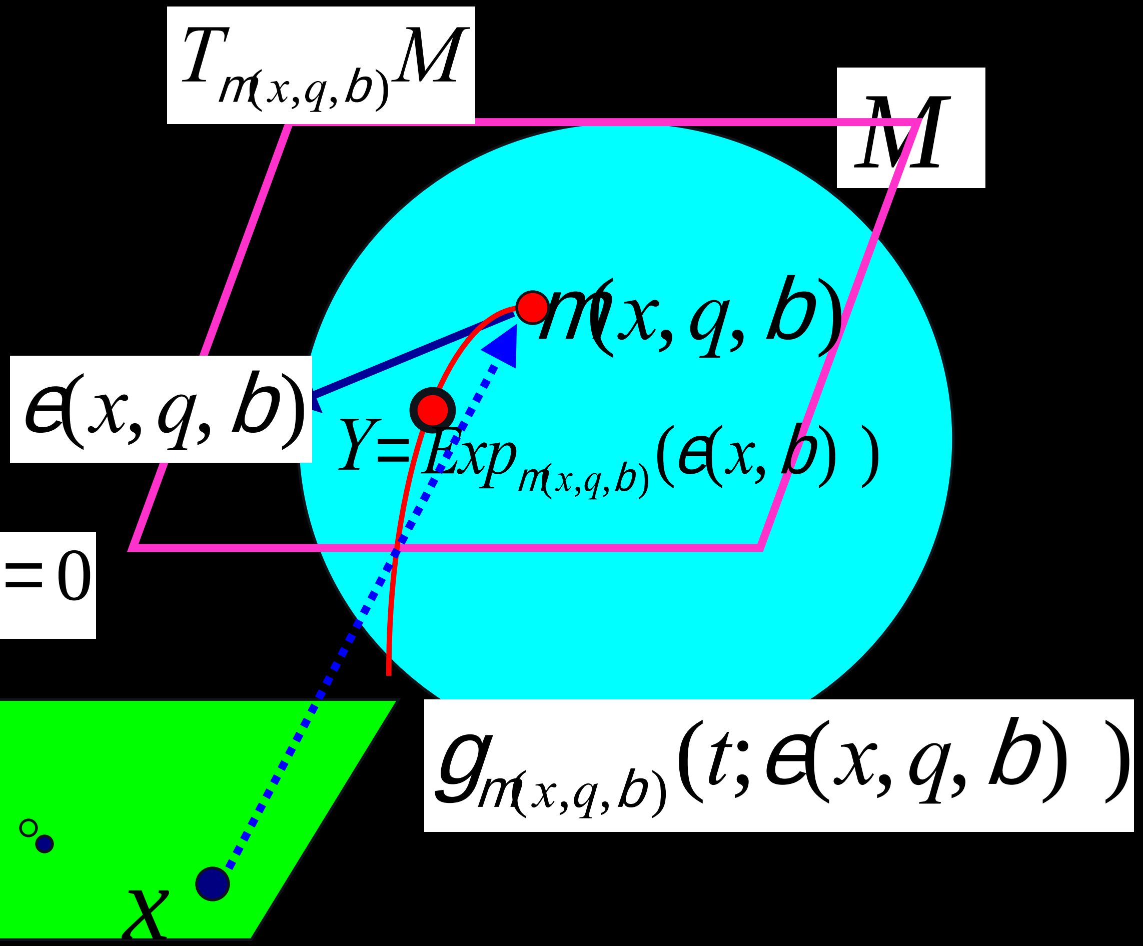
Conditional Mean

Riemannian logarithm maps

$$\epsilon(x, q, b) \ = \mathrm{Log}_{m(x,q,b)}(Y)\,\hat{\mathrm{I}}\ \ T_{m(x,q,b)}M$$

Conditional Moment Model

$$E[\epsilon(x, q, b)\,|\,x] = E[Log_{m(x,q,b)}(Y)\,|\,x] = 0$$



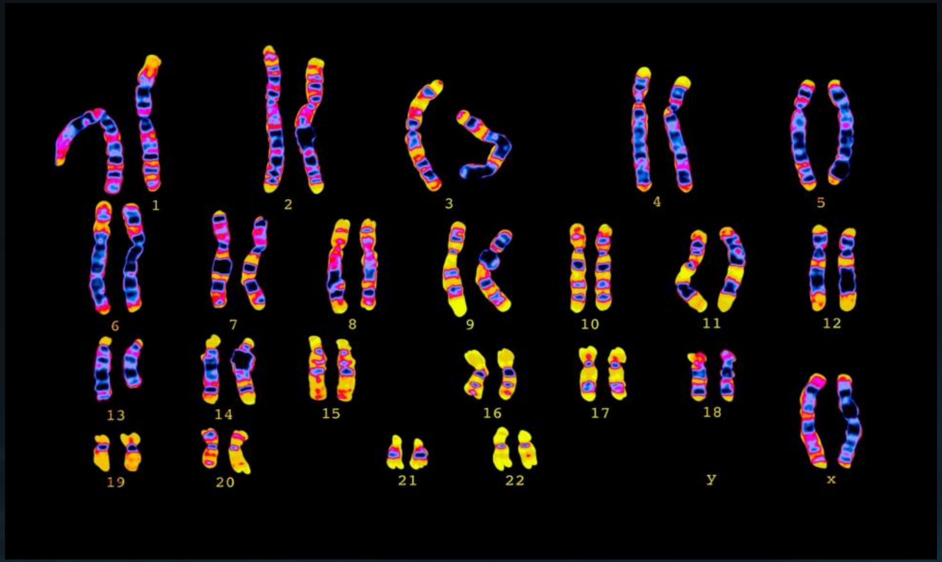
$T_{m(x,q,b)}M$
M
$m(x, q, b)$
$\epsilon(x, q, b)$
$Y = Exp_{m(x,q,b)}(\epsilon(x, b)\ )$
$g_{m(x,q,b)}(t; \epsilon(x, q, b)\ \ )$
X

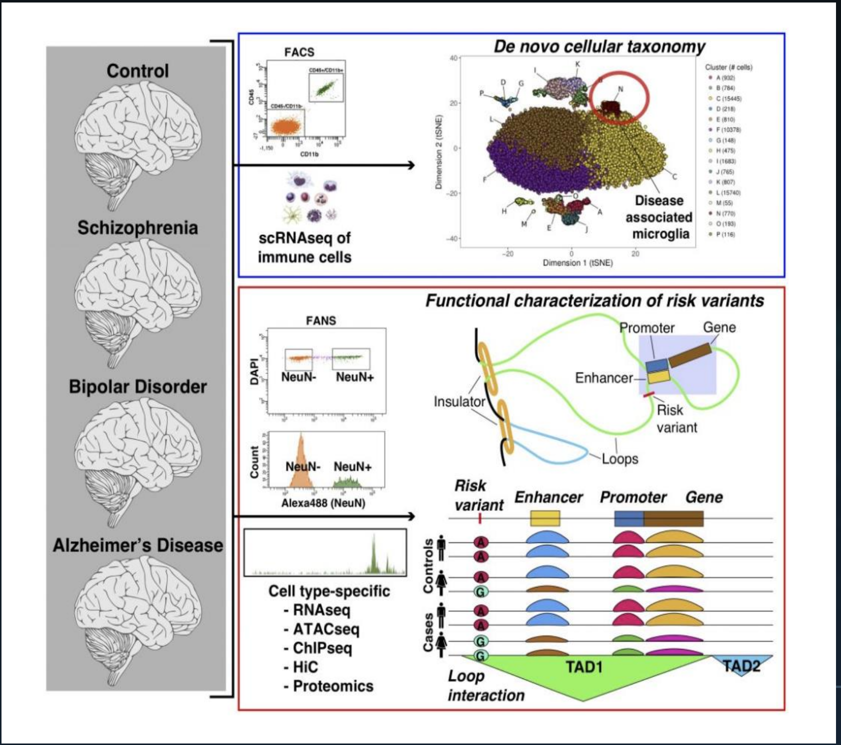
Cornea, E., ***Zhu, H.T.,*** Kim, P. and Ibrahim, J. G. Intrinsic regression model for data in Riemannian symmetric space. *JRSS, Series B*, 79, 463-482, 2017.

Imaging Genetics of Brain Disorders

Most major brain disorders (like AD) are **heritable complex traits/diseases**

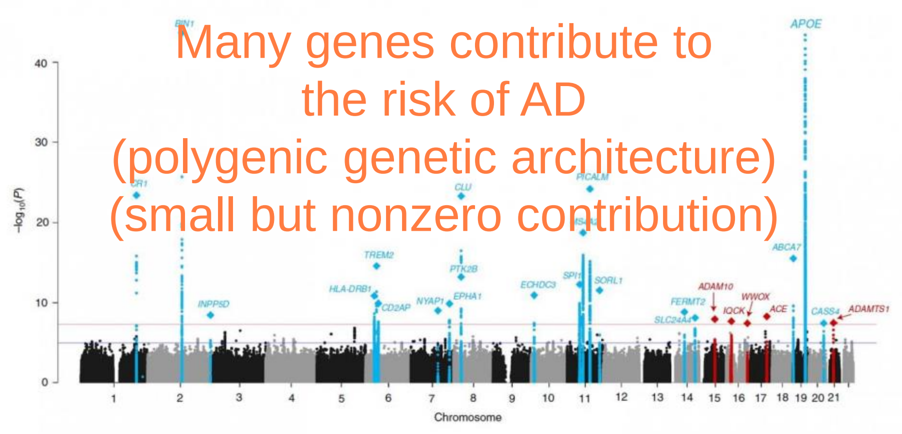
Together 50%-70% of AD risk
75%-90% of ADHD risk
60%-85% of Schizophrenia risk
~80% of Autism Spectrum Disorder (ASD) risk





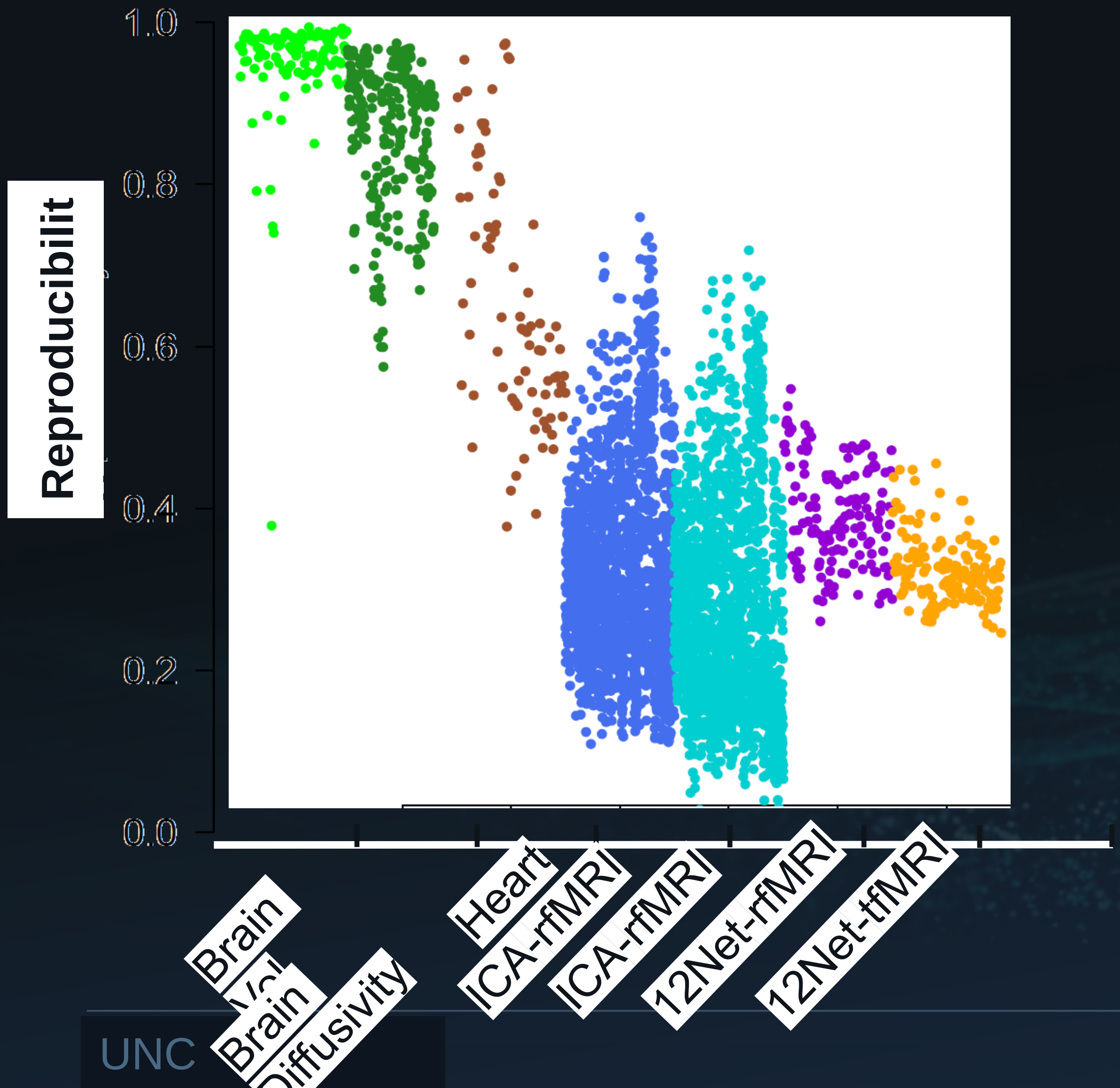
Complex traits/diseases
(many genes,
environmental factors,
complex functional
mechanism)

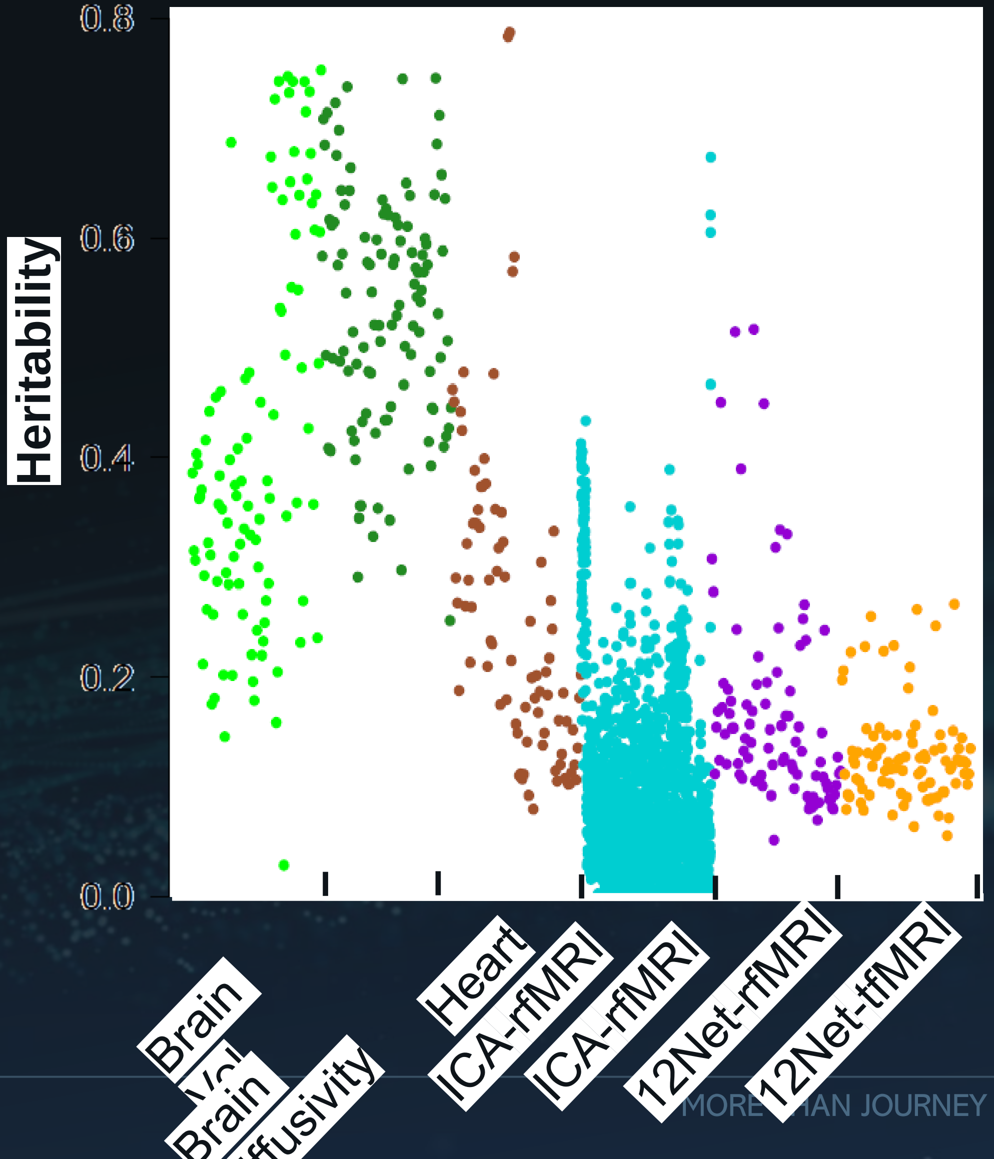
Genetic signals are non-spare and weak:
Need large sample size to detect weak signals



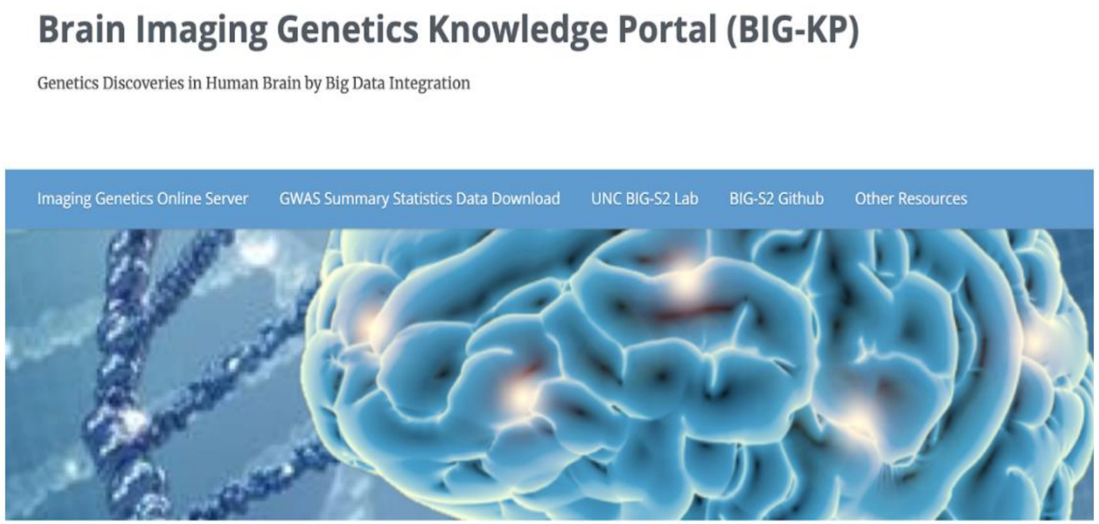
Many genes contribute to
the risk of AD
(polygenic genetic architecture)
(small but nonzero contribution)

IG: Reproducibility and Heritability



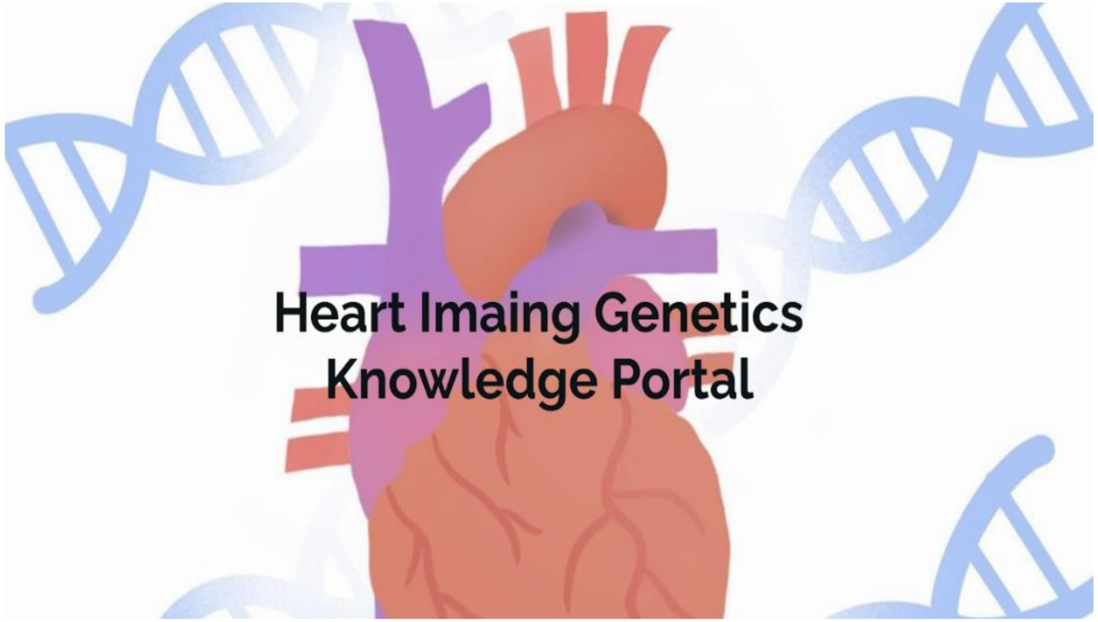


Brain- Heart Imaging Genetics Knowledge Portal


Brain Imaging Genetics Knowledge Portal (BIG-KP)

Genetics Discoveries in Human Brain by Big Data Integration

Imaging Genetics Online Server
GWAS Summary Statistics Data Download
UNC BIG-S2 Lab
BIG-S2 Github
Other Resources

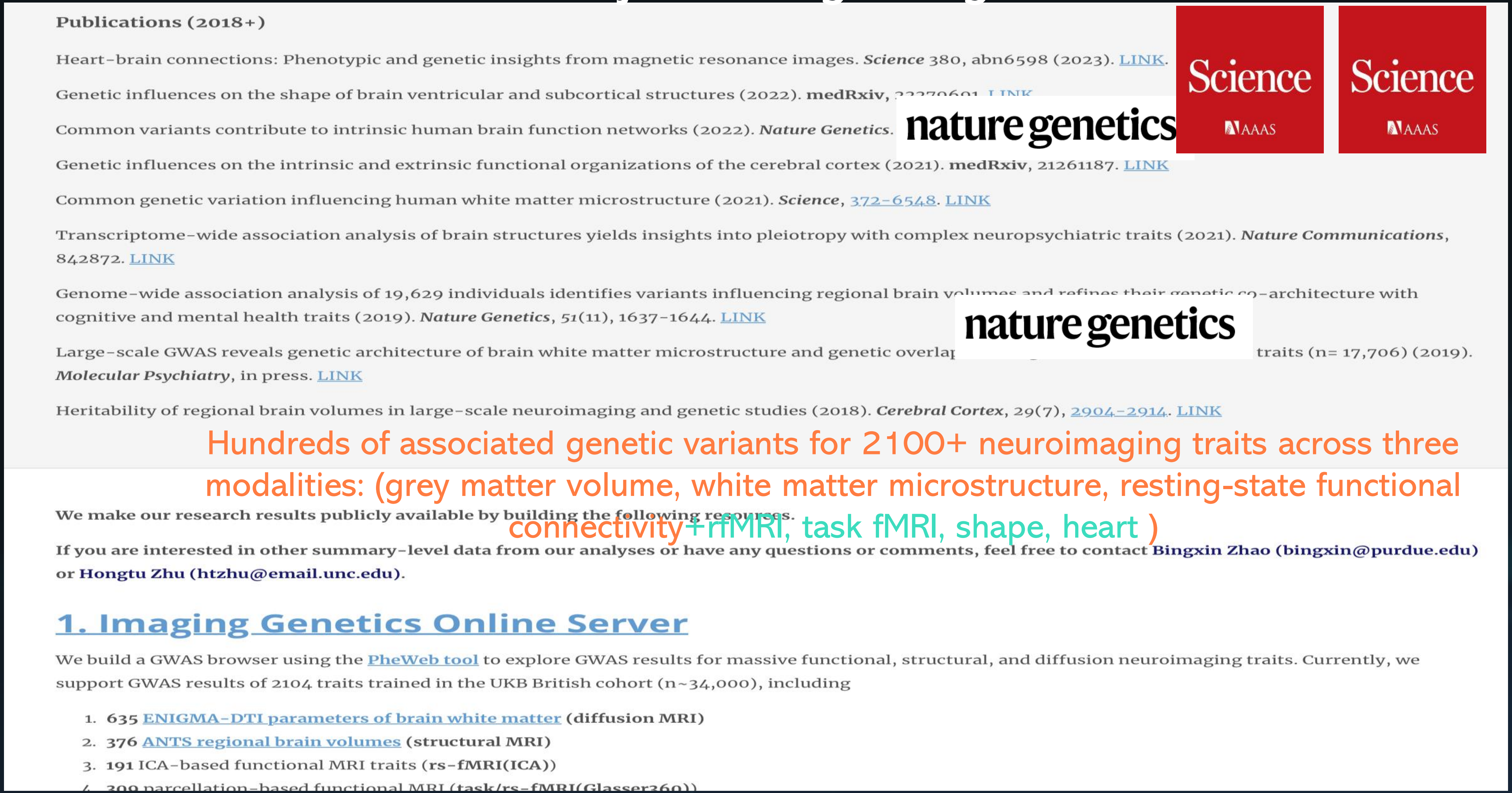


Brain Imaging Genetics Knowledge Portal (BIG-KP)

Heart Imaging Genetics Knowledge Portal (Heart-KP)

Aim to build the best knowledge database of neuroimaging genetics

It's just a beginning



Publications (2018+)

Heart-brain connections: Phenotypic and genetic insights from magnetic resonance images. ***Science*** 380, abn6598 (2023). LINK.

Genetic influences on the shape of brain ventricular and subcortical structures (2022). **medRxiv**, 22279601. LINK

Common variants contribute to intrinsic human brain function networks (2022). ***Nature Genetics***.

Genetic influences on the intrinsic and extrinsic functional organizations of the cerebral cortex (2021). **medRxiv**, 21261187. LINK

Common genetic variation influencing human white matter microstructure (2021). ***Science***, 372-6548. LINK

Transcriptome-wide association analysis of brain structures yields insights into pleiotropy with complex neuropsychiatric traits (2021). ***Nature Communications***, 842872. LINK

Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits (2019). ***Nature Genetics***, 51(11), 1637-1644. LINK

Large-scale GWAS reveals genetic architecture of brain white matter microstructure and genetic overlap traits (n= 17,706) (2019). ***Molecular Psychiatry***, in press. LINK

Heritability of regional brain volumes in large-scale neuroimaging and genetic studies (2018). ***Cerebral Cortex***, 29(7), 2904-2914. LINK

We make our research results publicly available by building the following resources.

If you are interested in other summary-level data from our analyses or have any questions or comments, feel free to contact **Bingxin Zhao (bingxin@purdue.edu)** or **Hongtu Zhu (htzhu@email.unc.edu)**.

1. Imaging Genetics Online Server

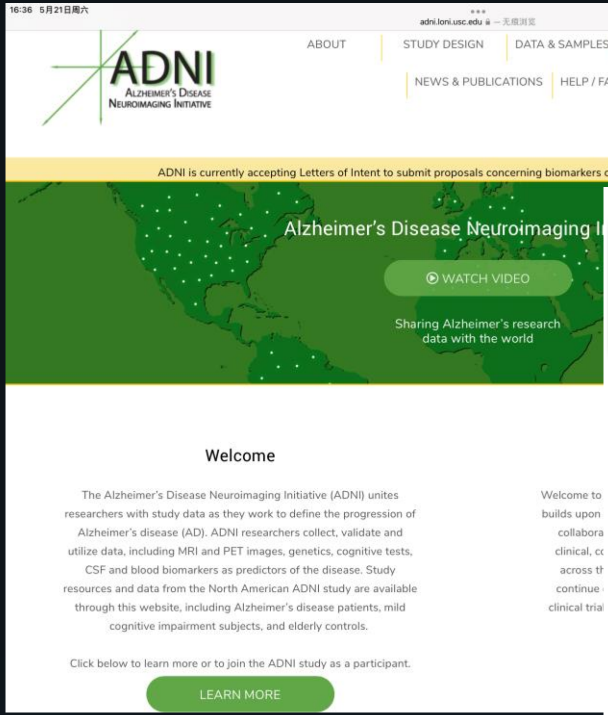
We build a GWAS browser using the PheWeb tool to explore GWAS results for massive functional, structural, and diffusion neuroimaging traits. Currently, we support GWAS results of 2104 traits trained in the UKB British cohort (n~34,000), including

1. 635 ENIGMA-DTI parameters of brain white matter (diffusion MRI)
2. 376 ANTS regional brain volumes (structural MRI)
3. 191 ICA-based functional MRI traits (rs-fMRI(ICA))
4. 299 parcellation-based functional MRI (task/rs-fMRI(Glasser360))

Hundreds of associated genetic variants for 2100+ neuroimaging traits across three modalities: (grey matter volume, white matter microstructure, resting-state functional connectivity+rfMRI, task fMRI, shape, heart)

Genetics discovery in human brain by big data integration

Alzheimer’s Disease Neuroimaging Initiative



16:36 5月21日周六

adni.loni.usc.edu — 无痕浏览

57%

ADNI
ALZHEIMER'S DISEASE
NEUROIMAGING INITIATIVE

ABOUT | STUDY DESIGN | DATA & SAMPLES | METHODS & TOOLS | DOCUMENTS

NEWS & PUBLICATIONS | HELP / FAQ | CONTACT

ADNI is currently accepting Letters of Intent to submit proposals concerning biomarkers or other assessment tools Read more

Alzheimer’s Disease Neuroimaging I

WATCH VIDEO

Sharing Alzheimer’s research data with the world

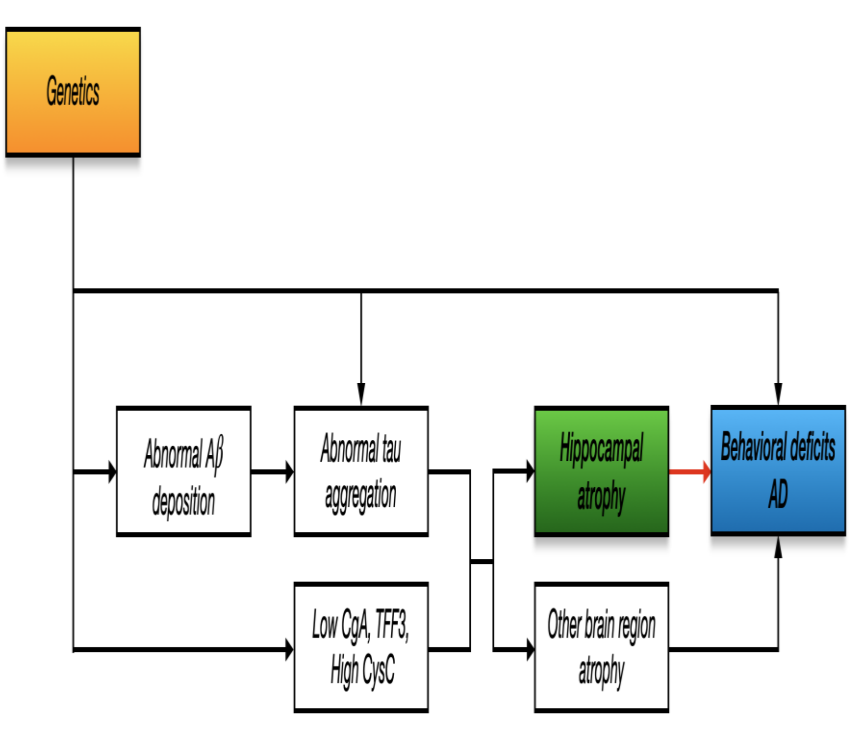
Welcome

The Alzheimer’s Disease Neuroimaging Initiative (ADNI) unites researchers with study data as they work to define the progression of Alzheimer’s disease (AD). ADNI researchers collect, validate and utilize data, including MRI and PET images, genetics, cognitive tests, CSF and blood biomarkers as predictors of the disease. Study resources and data from the North American ADNI study are available through this website, including Alzheimer’s disease patients, mild cognitive impairment subjects, and elderly controls.

Click below to learn more or to join the ADNI study as a participant.

LEARN MORE

Welcome to
builds upon
collabora
clinical, c
across th
continue
clinical trial



Genetics

Abnormal Aβ deposition

Abnormal tau aggregation

Low CgA, TFF3, High CysC

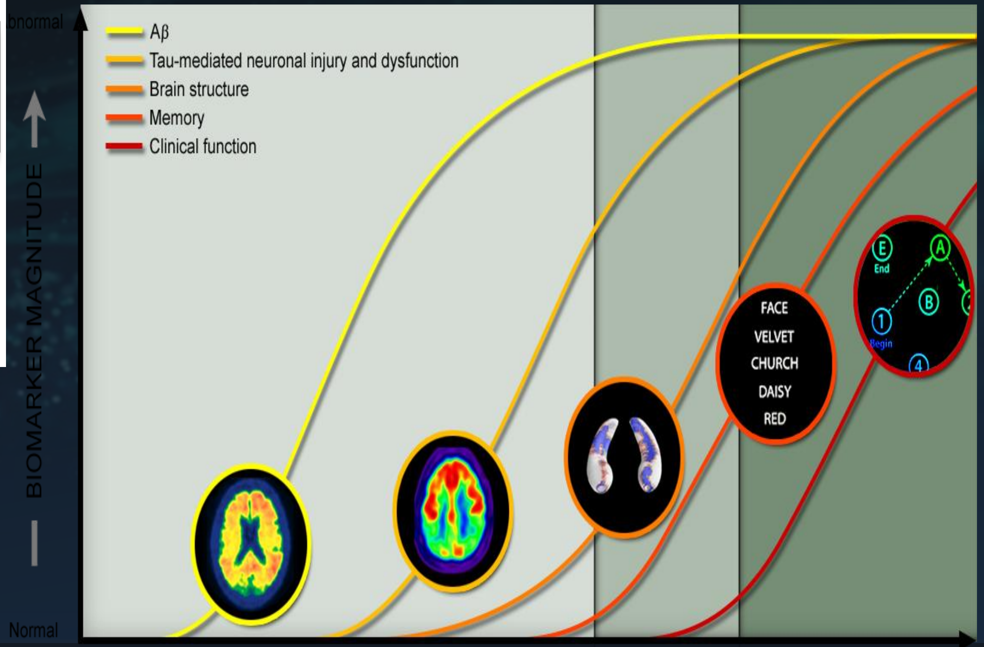
Hippocampal atrophy

Other brain region atrophy

Behavioral deficits AD

For a heterogeneous, clinically defined disorder, the endophenotype is ‘closer to the underlying biology,’

- **Increasing the power of genetic search**
- **Being informative about disorder risk.**
- **Providing mechanistic connections linking genetic variation to behavioral measures.**



2004-now

Model Setup

Outcome generating model

$$Y_i \ = \textstyle\sum_{l=1}^{S} x_{il}\, \beta_l + < \ \boldsymbol{Z}_i, \boldsymbol{B} \ > \ + \epsilon_i$$

Exposure generating model

$$\boldsymbol{Z}_i = \sum_{\mathrm{l=1}}^{\mathrm{S}} \mathrm{x_{il}} * \boldsymbol{C}_l + \boldsymbol{E}_i$$

$\boldsymbol{B}$ is the main parameter of interest, representing the association between the 2D imaging exposure $\mathrm{Z_i}$ and the behavioral outcome Y_i , β_l represents the association between the l−th observed covariate $x_{il}$and the behavioral outcome Y_i, and ϵ_i and $\mathrm{E_i}$are random errors that may be correlated. The symbol "$*$" denotes element-wise multiplication.

Ye, Wang, Kong, and Zhu (2022). Mapping the Genetic-Imaging-Clinical Pathway with Applications to Alzheimer's Disease. JASA, in press.

True Confounders, Precision, Instrumental and Irrelevant Variables

Outcome generating model

$$Y_i = \sum_{l=1}^{s} x_{il}\, \beta_l + < \boldsymbol{Z_i}, \boldsymbol{B} > + \epsilon_i$$

Exposure generating model

$$\boldsymbol{Z_i} = \sum_{\mathrm{l=1}}^{\mathrm{s}} \mathrm{x_{il}} * \boldsymbol{C_l} + \boldsymbol{E_i}$$

True Confounders $\mathcal{C} = \{l \in \mathcal{A} \mid \beta_l \neq 0 \text{ and } \boldsymbol{C_l} \neq 0\},$

Precision Variables $\mathcal{P} = \{l \in \mathcal{A} \mid \beta_l \neq 0 \text{ and } \boldsymbol{C_l} = 0\},$

Instrumental Variables $\mathcal{I} = \{l \in \mathcal{A} \mid \beta_l = 0 \text{ and } \boldsymbol{C_l} \neq 0\},$

Irrelevant Variables $\mathcal{S} = \{l \in \mathcal{A} \mid \beta_l = 0 \text{ and } \boldsymbol{C_l} = 0\}.$

Aim (to correctly estimate $\boldsymbol{B}$): retain all covariates from $\mathcal{M}_1 = \mathcal{C} \cup \mathcal{P} = \{l \in \mathcal{A} \mid \beta_l \neq 0\}$, while excluding covariates from $I \cup S = \{l \in \mathcal{A} \mid \beta_l = 0\}$.

Marginal Screening

Fit:

$$Y_i = x_{il}\beta_l + \epsilon_i$$

Obtain:

$$\hat{\beta}_l^M = n^{-1}\sum_{i=1}^{n} x_{il}Y_i$$

Problem!!! (plugging exposure model into outcome model)

Outcome generating model $Y_i = \sum_{l=1}^{s} x_{il}\,\beta_l + < \boldsymbol{Z_i}, \boldsymbol{B} > + \epsilon_i$

Exposure generating model $\boldsymbol{Z_i} = \sum_{l=1}^{s} x_{il} * \boldsymbol{C}_l + \boldsymbol{E_i}$

Obtain:

$$Y_i = \sum_{l=1}^{s} x_{il}\,(\beta_l + < \boldsymbol{C_l}, \boldsymbol{B} >) + < \boldsymbol{E_i}, \boldsymbol{B} > + \epsilon_i$$

Miss a portion of confounders when β_l and $< \boldsymbol{C_l}, \boldsymbol{B} >$ are of similar magnitude but opposite sign.

Joint Screening (proposed)

Marginal screening:

$$\mathrm{\mathbf{Z}_i} = \sum_{\mathrm{l=1}}^{\mathrm{s}} \mathrm{x_{il}} * \mathbf{C}_l + \mathrm{E_i}$$

Obtain (*Kong, An, Zhang and Zhu, 2020*):

$$\widehat{\boldsymbol{C}}_l^M = n^{-1} \sum_{i=1}^{n} x_{il} * \boldsymbol{Z}_i \in \mathbb{R}^{p\times q}$$

$$\widehat{\mathcal{M}}_1^* = \left\{1 \le I \le s: \left|\widehat{\beta_l^M}\right| \ge \gamma_{1,n}\right\}$$

$$\widehat{\mathcal{M}}_2 = \left\{1 \le I \le s: \| \widehat{\boldsymbol{C}}_l^M \|_{op} \ge \gamma_{2,n}\right\}$$

$\mathcal{C} = \{l \in \mathcal{A} \mid \beta_l \neq 0$ and $\boldsymbol{C}_l \neq 0\},$ ☹ ▲

$\mathcal{P} = \{l \in \mathcal{A} \mid \beta_l \neq 0$ and $\boldsymbol{C}_l = 0\},$ ☹

$\mathcal{I} = \{l \in \mathcal{A} \mid \beta_l = 0$ and $\boldsymbol{C}_l \neq 0\},$ ▲

$\mathcal{S} = \{l \in \mathcal{A} \mid \beta_l = 0$ and $\boldsymbol{C}_l = 0\}.$

Select submodel: $\widehat{\mathcal{M}} = \widehat{\mathcal{M}}_1^* \cup \widehat{\mathcal{M}}_2.\ (\mathrm{Union})$

Alternative choices (both worse): $\widehat{\mathcal{M}}_1^*$ (outcome) or $\widehat{\mathcal{M}}_1^* \cap \widehat{\mathcal{M}}_2$(Outcome).

Estimation (proposed)

Minimize:

$$\frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \langle \boldsymbol{Z}_i, \mathrm{B}\rangle - \sum_{l\in\widehat{\mathcal{M}}} X_{il}\beta_l\right)^2 + \lambda_{1,n}\sum_{l\in\widehat{\mathcal{M}}} |\beta_l| + \lambda_{2,n} \parallel \mathrm{B} \parallel_*$$

where $\parallel \mathrm{B} \parallel_* = \sum_k \sigma_k(\mathrm{B})$.

$$\begin{aligned}
\mathcal{C} &= \{l \in \mathcal{A} \mid \beta_l \neq 0 \text{ and } \boldsymbol{C}_l \neq 0\},\\
\mathcal{P} &= \{l \in \mathcal{A} \mid \beta_l \neq 0 \text{ and } \boldsymbol{C}_l = 0\},\\
\mathcal{I} &= \{l \in \mathcal{A} \mid \beta_l = 0 \text{ and } \boldsymbol{C}_l \neq 0\},\\
\mathcal{S} &= \{l \in \mathcal{A} \mid \beta_l = 0 \text{ and } \boldsymbol{C}_l = 0\}.
\end{aligned}$$

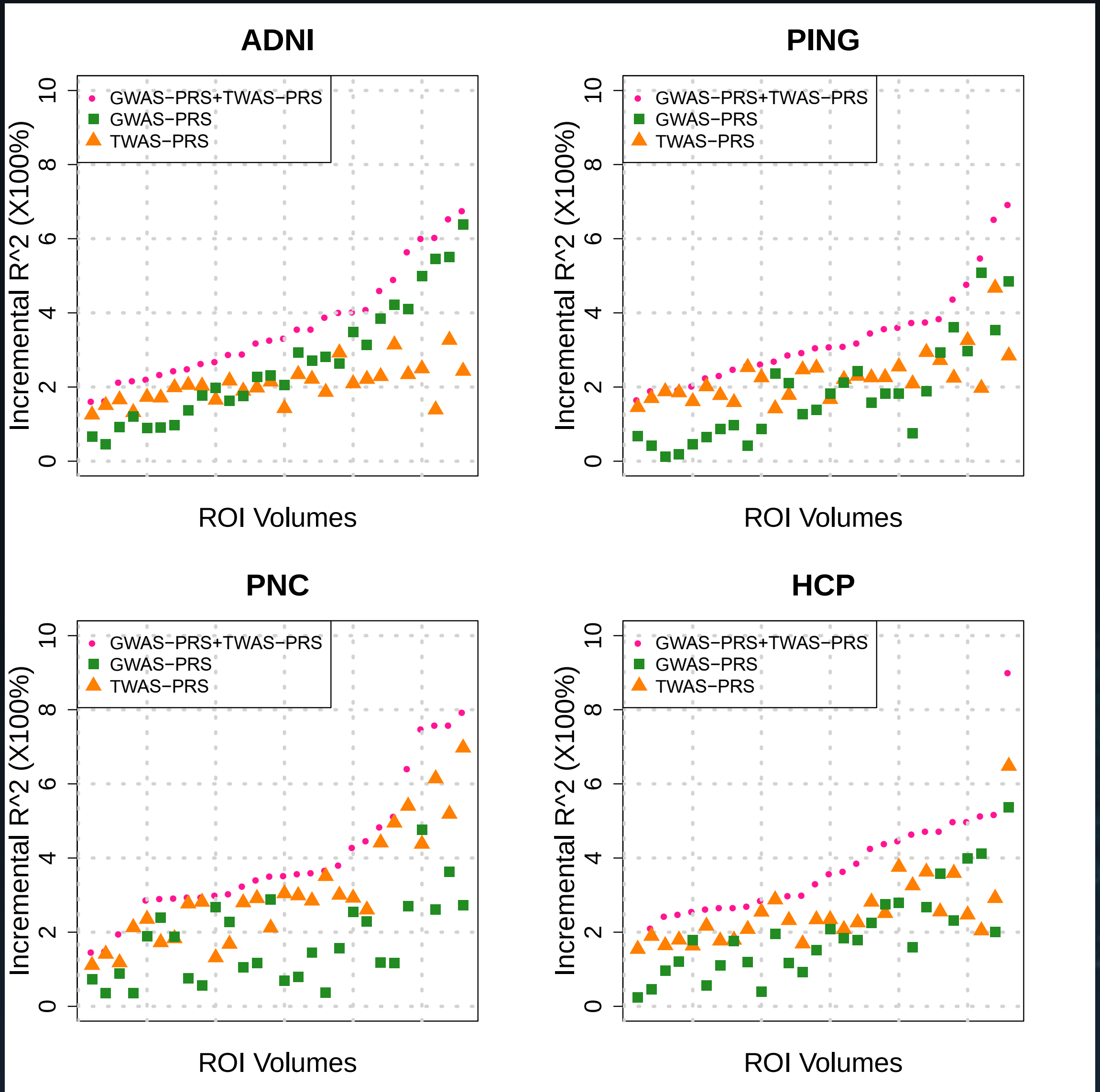
L1 penalty, exclude instrumental and irrelevant variables.

Nuclear penalty, low-rank estimation of B.

Estimated effect size of imaging exposure z,

$$\hat{\mu}(z) = \langle z, \hat{B}\rangle$$

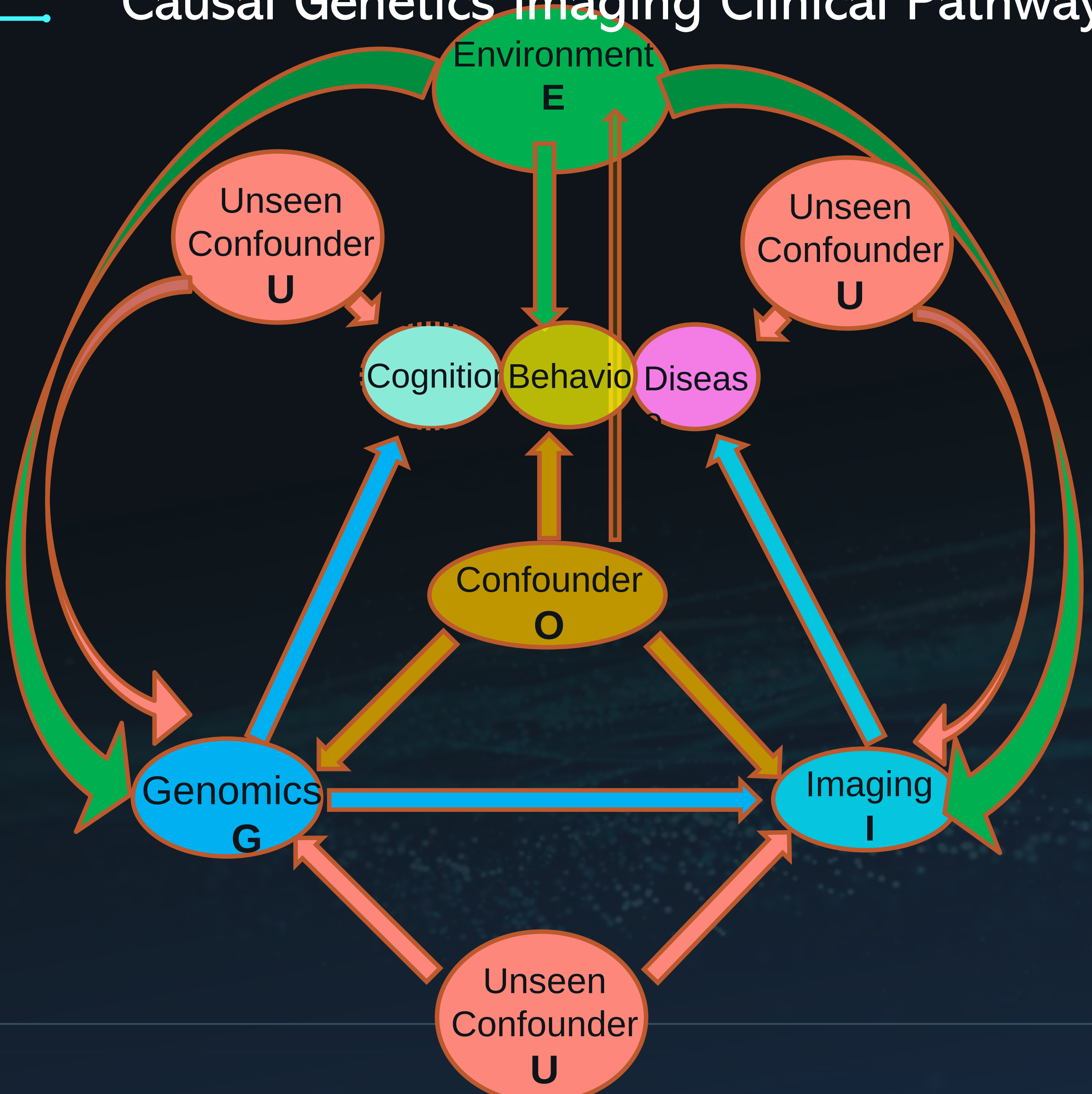
Predictive Analysis



Gene expression-informed gene-level PRS + GWAS PRS has higher prediction accuracy

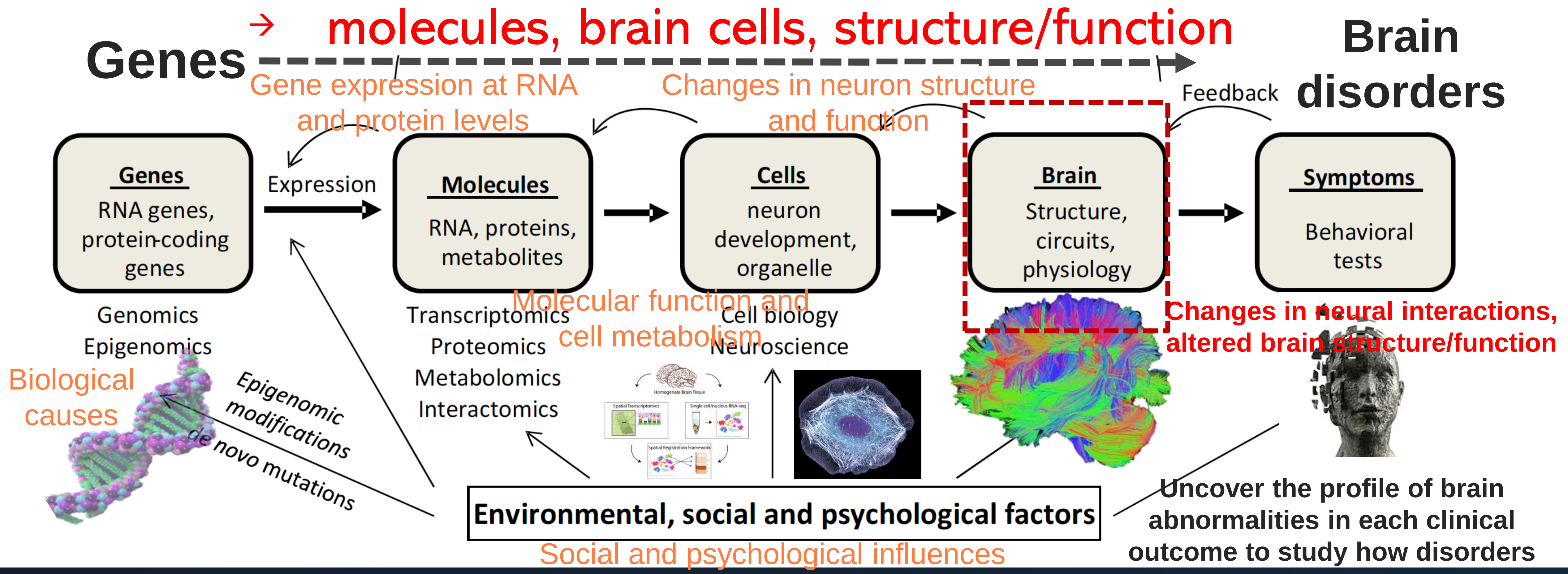
Construct gene-level PRS (polygenic risk scores) by leveraging gene expression reference panels
(e.g., GTEx) in TWAS

Causal Genetics Imaging Clinical Pathway



Brain Imaging Genetics Paradigm

Neuroimaging: an important component to help understand the complex biological pathways of brain disorders



Genes → molecules, brain cells, structure/function Brain disorders

Gene expression at RNA and protein levels

Changes in neuron structure and function

Feedback

Genes
RNA genes, protein-coding genes

Expression

Molecules
RNA, proteins, metabolites

Cells
neuron development, organelle

Brain
Structure, circuits, physiology

Symptoms
Behavioral tests

Genomics
Epigenomics

Transcriptomics
Proteomics
Metabolomics
Interactomics

Molecular function and cell metabolism

Cell biology
Neuroscience

Changes in neural interactions, altered brain structure/function

Biological causes

Epigenomic modifications

de novo mutations

Environmental, social and psychological factors

Social and psychological influences

Uncover the profile of brain abnormalities in each clinical outcome to study how disorders develop

Challenges

- The complexity of those large-scale neuroimaging-related data sets is too high for most research teams in both academia and industry.
- It is very difficulty to appropriately process data across different domains with high quality, while controlling for potential bias introduced during the preprocessing stage.
- It remains uncertain as to how to appropriately integrate data across different domains obtained from different studies and cohorts with possible different study designs for unbiased data integration.
- It remains unclear how to appropriately and efficiently analyze neuroimaging related data sets with multiple Vs (e.g., Volume, Velocity, Variety and Veracity), while ensuring algorithmic fairness.

Statistical Learning Methods for NDA

Statistical Learning Methods for Neuroimaging Data Analysis with Applications

Hongtu Zhu[1], Tengfei Li[2], and Bingxin Zhao[3]

[1] Departments of Biostatistics, Statistics, Genetics, and Computer Science and Biomedical Research Imaging Center, University of North Carolina, Chapel Hill, NC, 27514; email: htzhu@email.unc.edu
[2] Departments of Radiology and Biomedical Research Imaging Center, University of North Carolina, Chapel Hill, NC,27514
[3] Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA, 19104

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–37

https://doi.org/10.1146/((please add article doi))

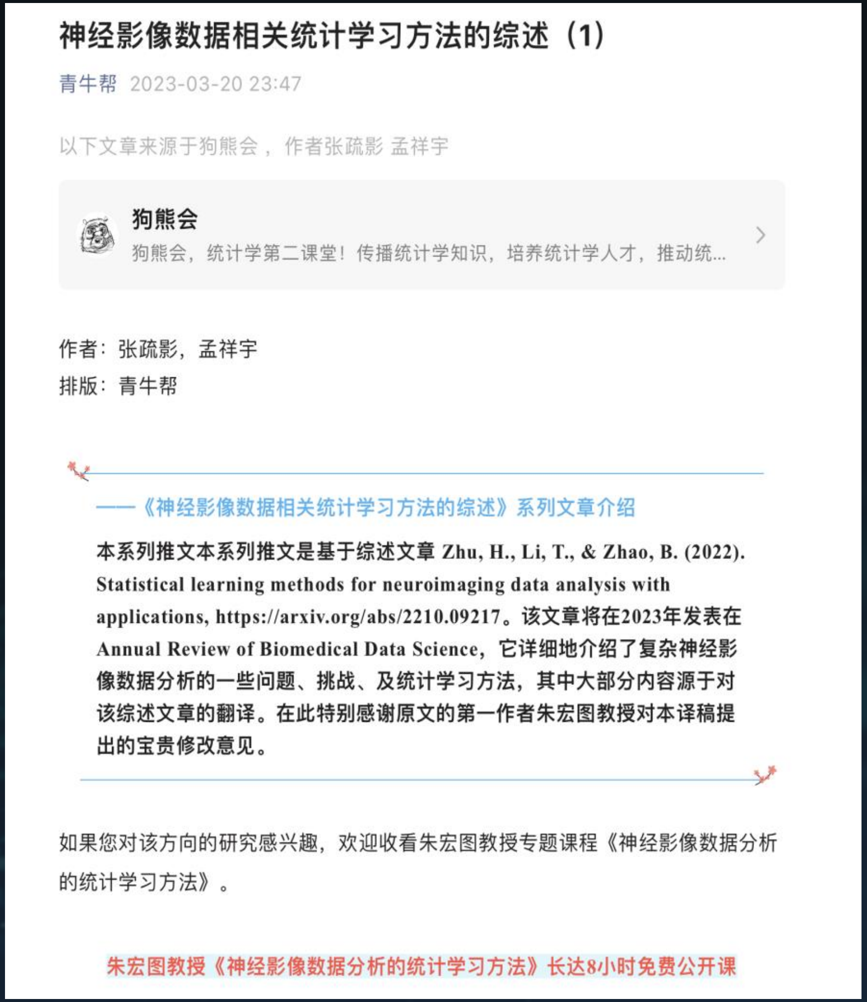
Keywords

causal pathway, heterogeneity, image processing analysis, neuroimaging techniques, population-based statistical analysis, study design

Abstract

The aim of this paper is to provide a comprehensive review of statistical challenges in neuroimaging data analysis from neuroimaging techniques to large-scale neuroimaging studies to statistical learning methods. We briefly review eight popular neuroimaging techniques and their potential applications in neuroscience research and clinical translation. We delineate the four common themes of neuroimaging data and review major image processing analysis methods for processing neuroimaging data at the individual level. We briefly review four large-scale neuroimaging-related studies and a consortium on imaging genomics and discuss four common themes of neuroimaging data analysis at the population level. We review nine major population-based statistical analysis methods and their associated statistical challenges and present recent progress in statistical methodology to address these challenges.

1



神经影像数据相关统计学习方法的综述（1）

青牛帮 2023-03-20 23:47

以下文章来源于狗熊会，作者张疏影 孟祥宇

狗熊会
狗熊会，统计学第二课堂！传播统计学知识，培养统计学人才，推动统...

作者：张疏影，孟祥宇
排版：青牛帮

——《神经影像数据相关统计学习方法的综述》系列文章介绍

本系列推文本系列推文是基于综述文章 Zhu, H., Li, T., & Zhao, B. (2022). Statistical learning methods for neuroimaging data analysis with applications, https://arxiv.org/abs/2210.09217。该文章将在2023年发表在Annual Review of Biomedical Data Science，它详细地介绍了复杂神经影像数据分析的一些问题、挑战、及统计学习方法，其中大部分内容源于对该综述文章的翻译。在此特别感谢原文的第一作者朱宏图教授对本译稿提出的宝贵修改意见。

如果您对该方向的研究感兴趣，欢迎收看朱宏图教授专题课程《神经影像数据分析的统计学习方法》。

朱宏图教授《神经影像数据分析的统计学习方法》长达8小时免费公开课

(Zhu, Li & Zhao, 2023)

Acknowledgement

UNC | GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Brain Imaging Genetics Knowledge Portal (BIG-KP)

Genetics Discoveries in Human Brain by Big Data Integration

bigkp.org

Funding: U.S. NIH Grants MH086633 and MH116527

Pictures: Copyrights belong to their own authors and/or holders.

Data: We thank Bingxin Zhao, Tengfei Li and other members of the UNC BIG-S2 lab (https://med.unc.edu/bigs2/) for processing the neuroimaging data.

UK Biobank resource application number: 22783.