Revolutionizing Medical Data Analysis: Uniting AI and Statistics for Breakthroughs and Challenges

University of North Carolina at Chapel Hill

Hongtu Zhu

Thanks to Drs. Mingxia Liu, Xin Wang, Lijuan Liu, Gang Li, Yukang Jiang, Shan Gao, Yue Yang, Hanchuan Peng, Wei Cheng, Marc Niethammer, Tengfei Li, and Bingxin Zhao for sharing their slides.

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

CONTENTS

Part I

Introduction to Medical Image Data Analysis

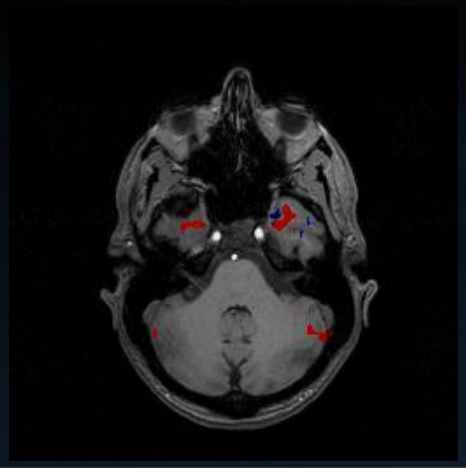
"Oddly, we are in a period where there has never been such a wealth of new statistical problems and sources of data. The danger is that if we define the boundaries of our field in terms of familiar tools and familiar problems, we will fail to grasp the new opportunities."

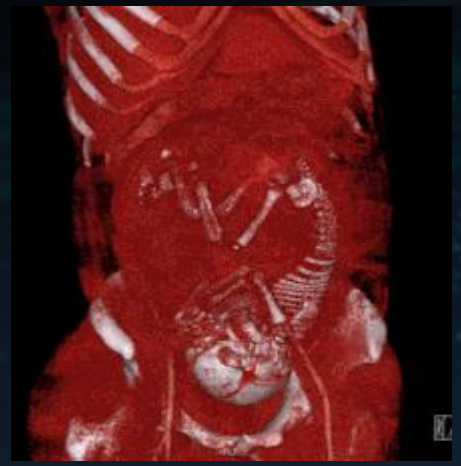
- Leo Breiman -

Medical Imaging

Medical imaging is the technique and process used to create images of the human body for clinical purposes or medical science. (https://en.wikipedia.org/)

- These imaging methods are essential for delineating the **structure and functionality of organs and tissues.** Each modality employs a distinct targeting agent, generates data in varying dimensions, extracts unique features, and serves specific purposes within clinical and research contexts.

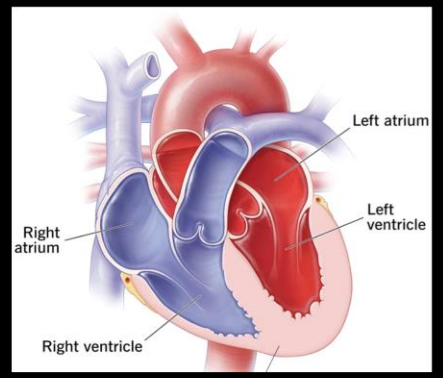


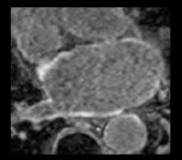


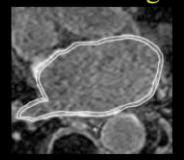
- X-ray radiography
- Computerized tomography (CT)
- Magnetic resonance imaging (MRI)
- Ultrasound
- Positron emission tomography (PET)
- Electroencephalography (EEG)
- Magnetoencephalography (MEG)
- Functional near-infrared spectroscopy (fNIRS)
- Mammography
- Light microscopy images
- Fluoroscopy
- Echocardiography

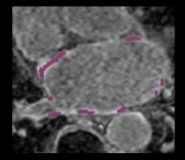
Cardiac Imaging

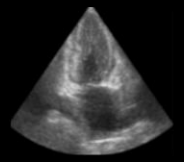
Heart	Target	Modality	Structure	Lesion/Function	Related Disease
Non-vessel	Atrium	LEG MRI	LA Wall Seg	LA fibrosis	Atrial Fibrillation
	Ventricle	Ultrasound	Ventricle Seg	Ventricle Function	Ejection Fraction Estimation
	Myocardium	Myocardial Perfusion MRI	Myocardium Seg	Myocardium Function	Ischemic Heart Disease
Vessel	Aorta	MRI	Aorta Seg	Aorta Flow	Aorta Stenosis
	Connery Arteries	CTA	Coronary Artery Seg	Fractional Flow Reserve	Coronary Artery Disease

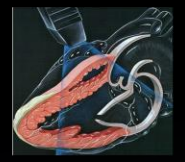


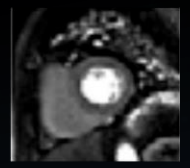


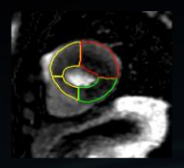


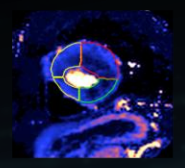


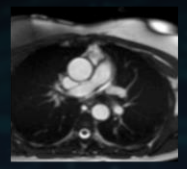


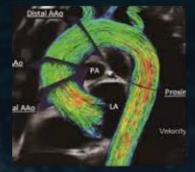


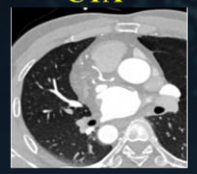


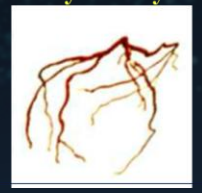


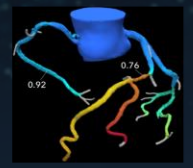


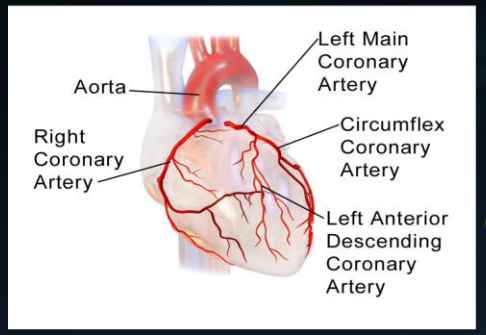












Wang, X. and Zhu, H (2024). Artificial Intelligence in Image-based Cardiovascular Disease Analysis: A Comprehensive Survey and Future Outlook

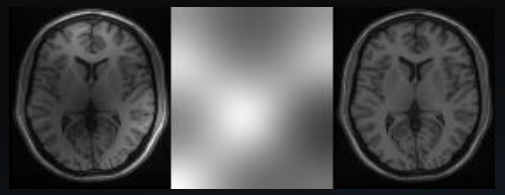
Image Processing Analysis Methods

How to enhance and extract signals of interest in imaging data?

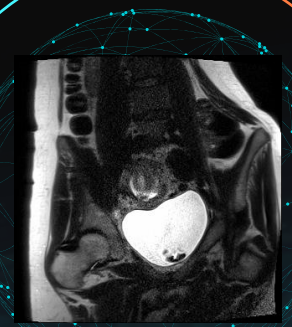
Deconvolution

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

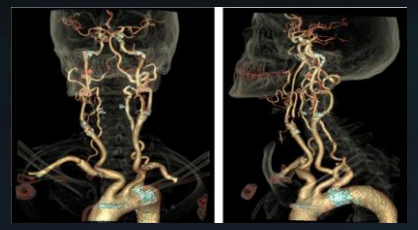
➢ Image Reconstruction



➢ Image Enhancement



❖ Image Segmentation



❖ Image Registration

Structural Learning

Engineering | Statistics | Machine Learning | Mathematics

Structural Learning

Image Segmentation

- Organ parcellation
- Localization of pathology
- Surgical planning
- Image-guided interventions
- Computer-aided diagnosis
- Quantification of organ change

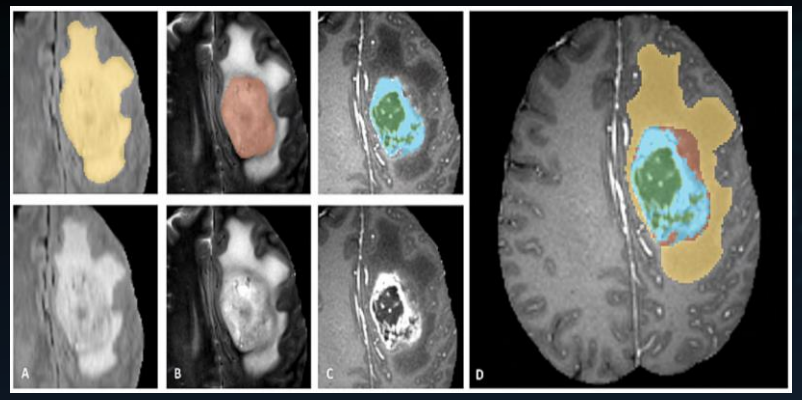
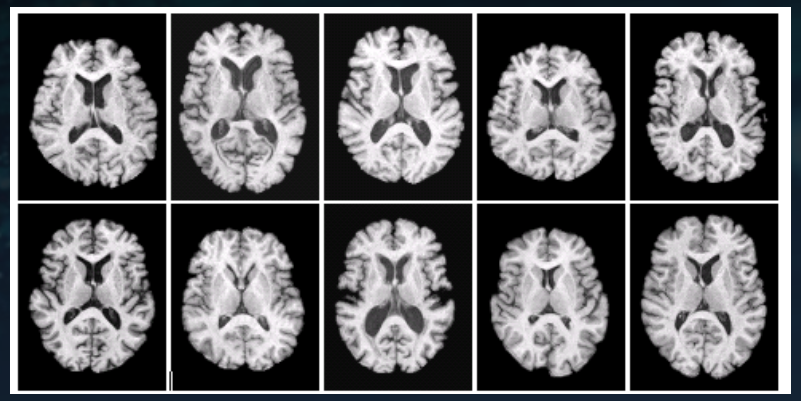
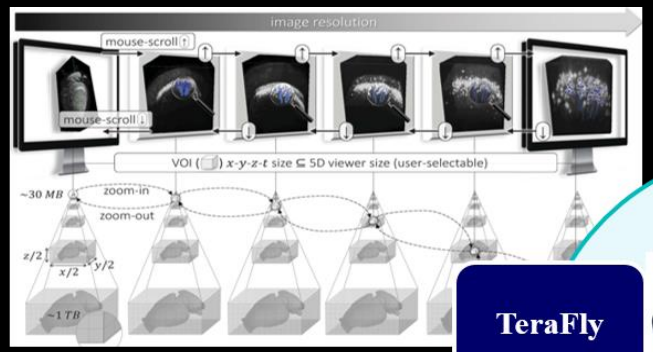


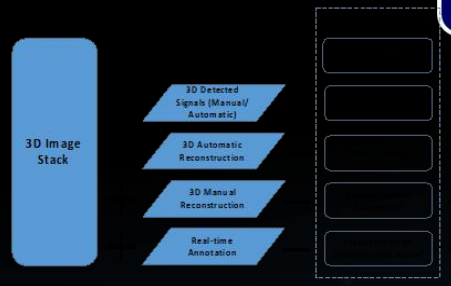
Image Registration

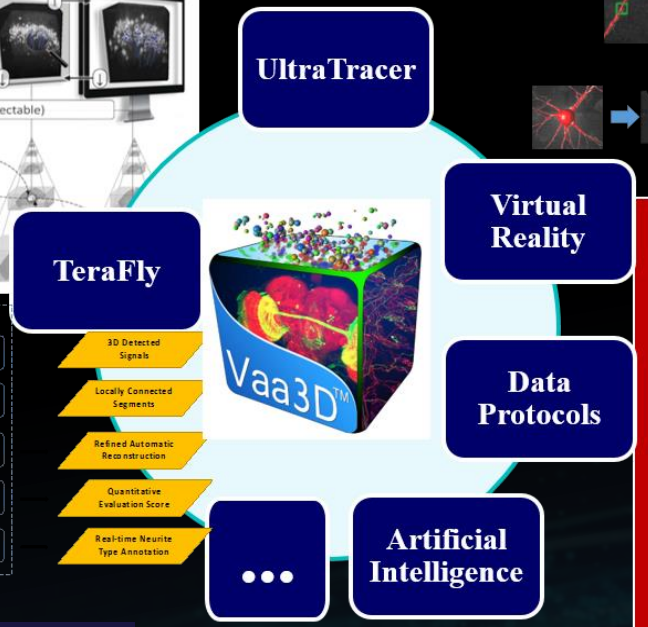
- Organ atlas
- Localization of pathology
- Automated image segmentation
- Multimodal fusion
- Population analysis
- Quantification of organ changes

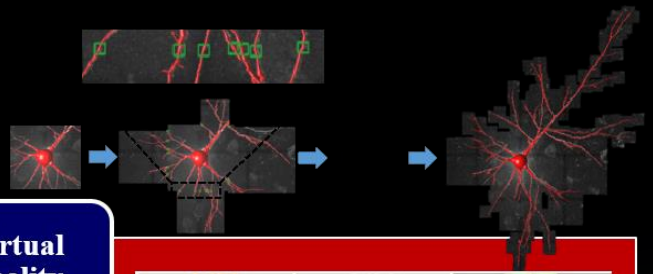


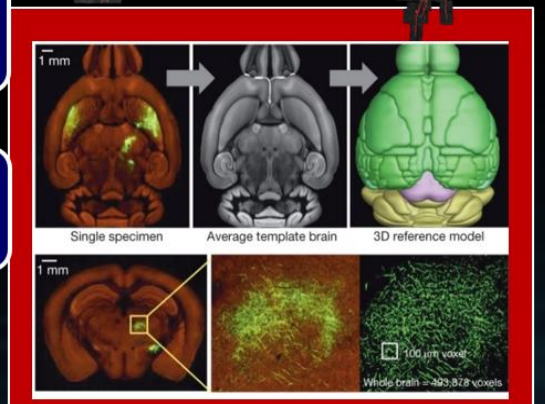
Light Microscopy Imaging at Single Cell

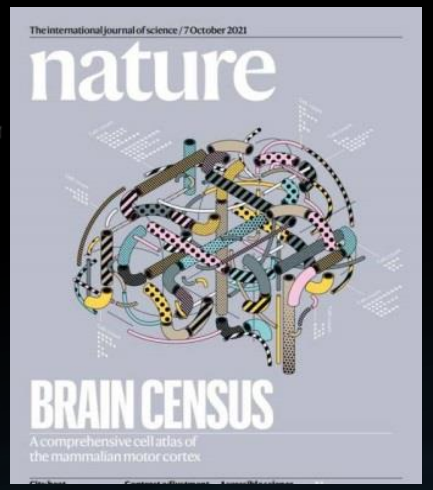


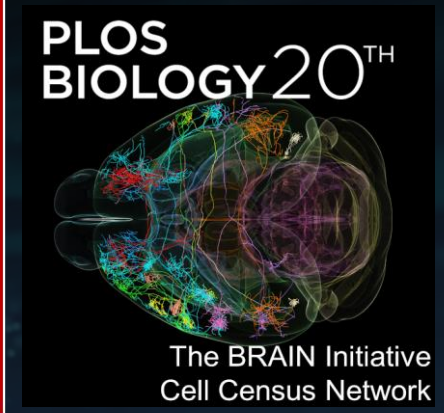












UltraTracer: Nature Methods 2017
TeraFly: Nature Methods, 2016
DeepNeuron: Brain Informatics 2018

Wang, et al. *Nature Commu* (2019)

Qu, et al. *Nature Methods*, (2022)

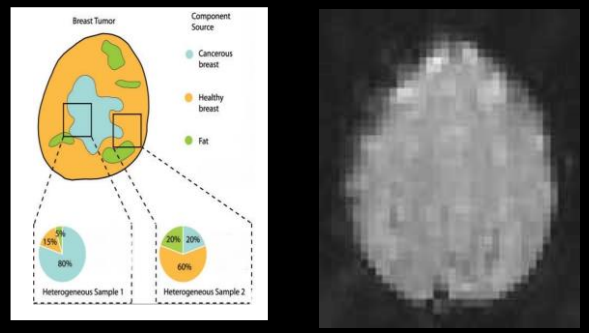
Han, X., et al. *Sci Adv.*, (2023)

"Top 10 life scientific advances of 2021" China

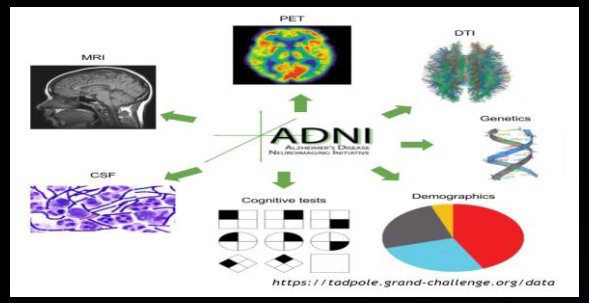
Morphological diversity of single neurons in molecularly defined cell types

Peng, H., et al. *Nature*, (2021)

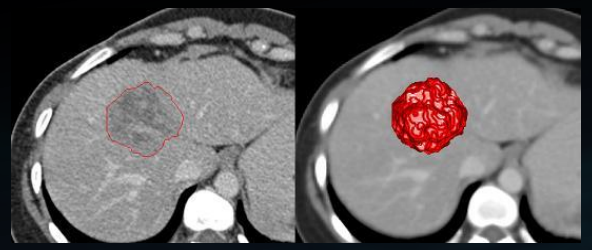
Ecological Layout for Imaging-based Analysis



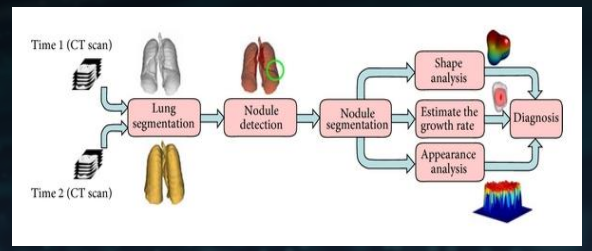
Deconvolution



Integration



Structural Learning



Prediction

Part II

State-of-the-Art AI Applications in Medical Imaging and Statistical Challenges

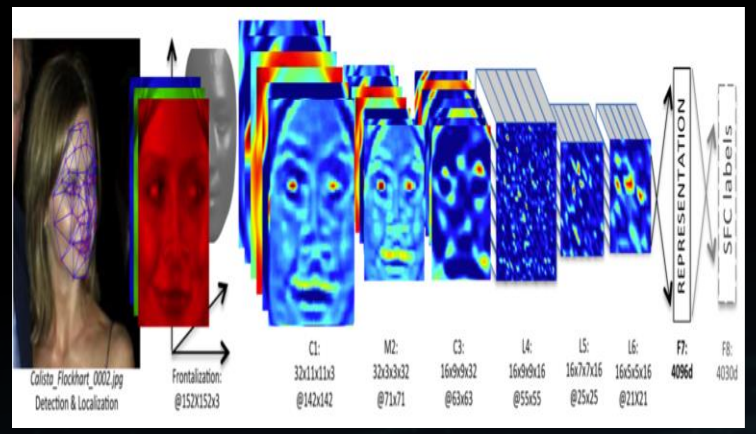
The best thing about being a statistician is that you get to play in everyone's backyard.
- John Tukey -

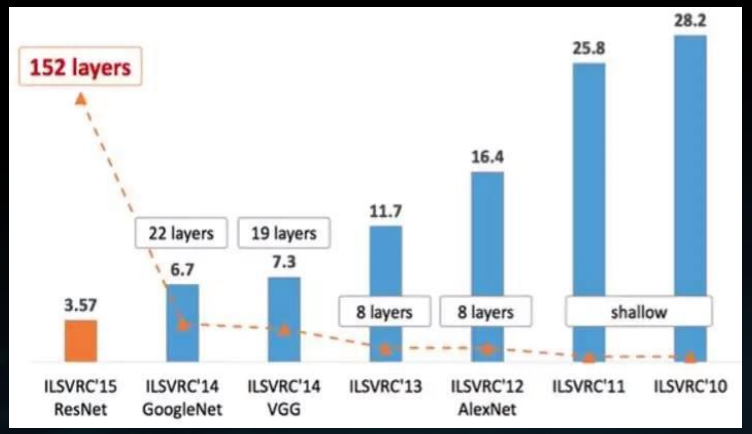
"If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools."
- Leo Breiman -

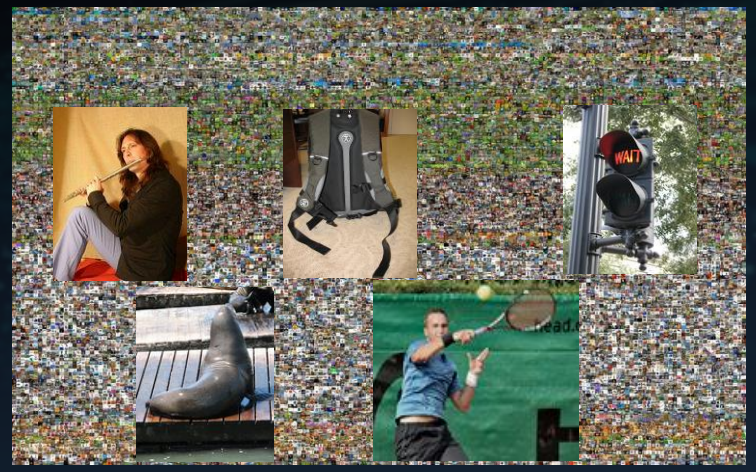
AI Milestones

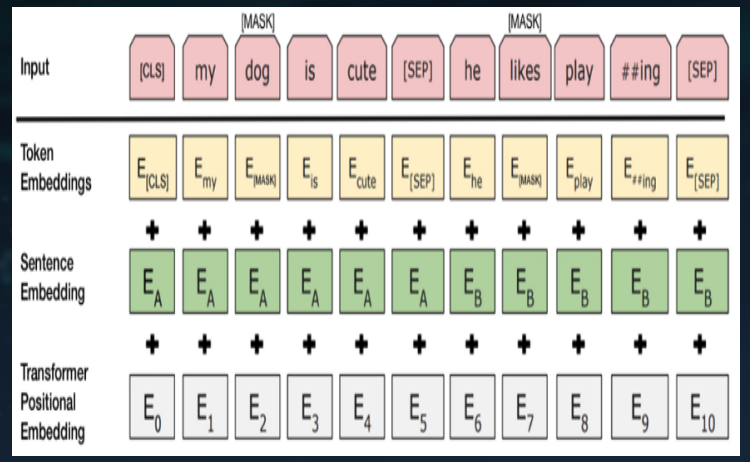
Annotated Datasets

Deep Learning



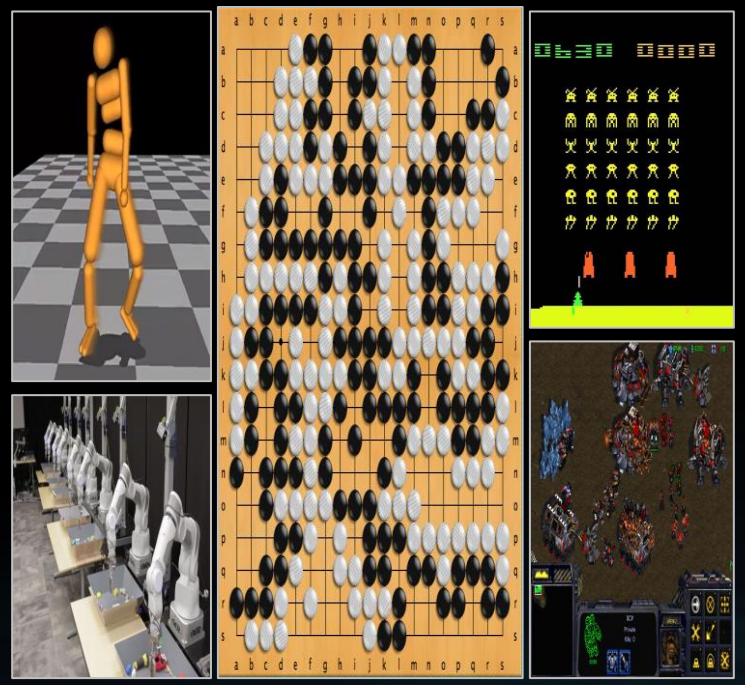




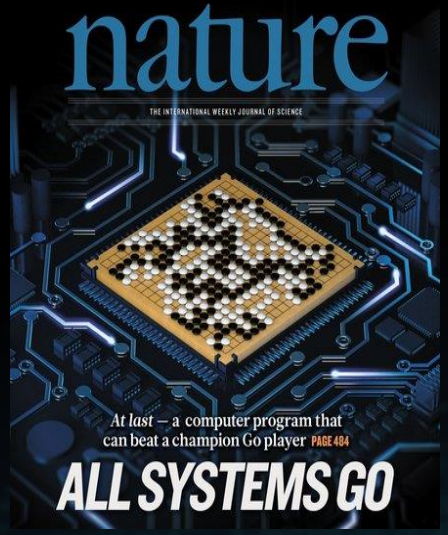


AI Milestones

Reinforcement Learning

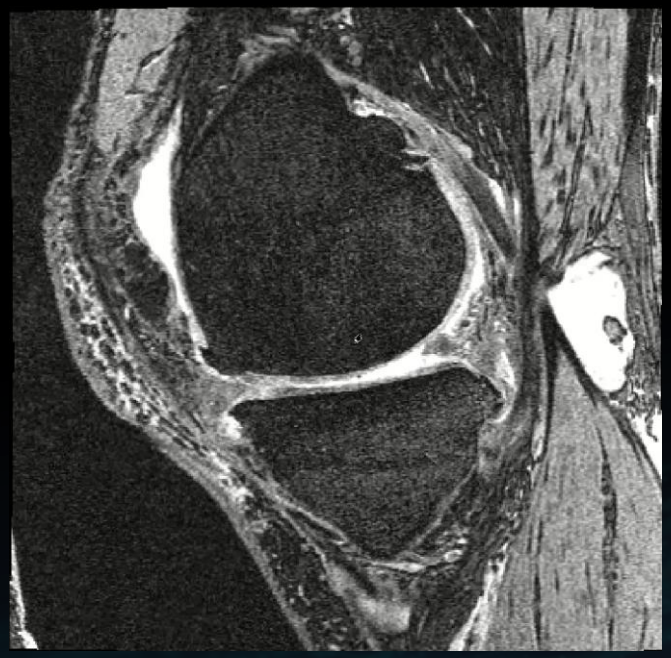


AI Products

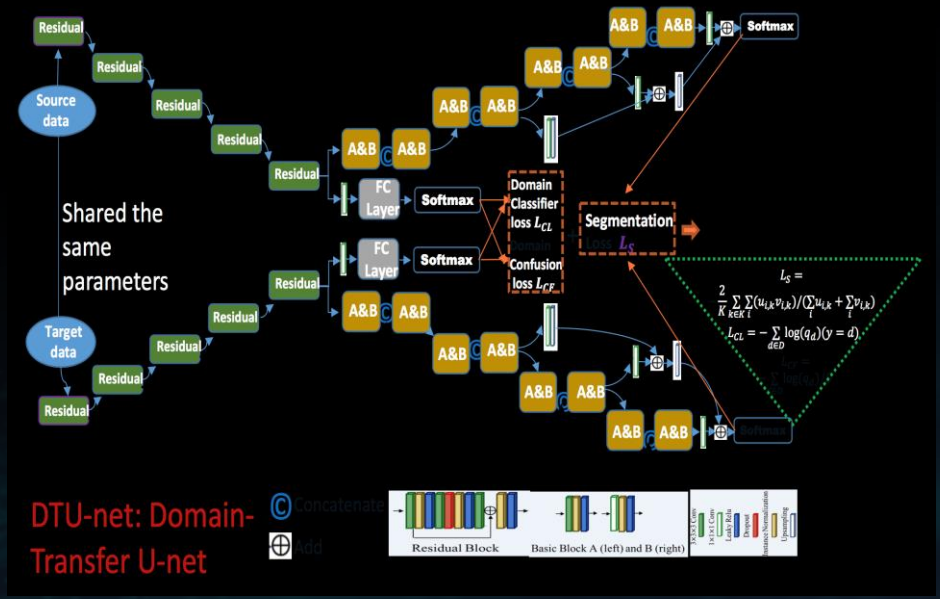


AI for Image Segmentation

Segmentation Annotation



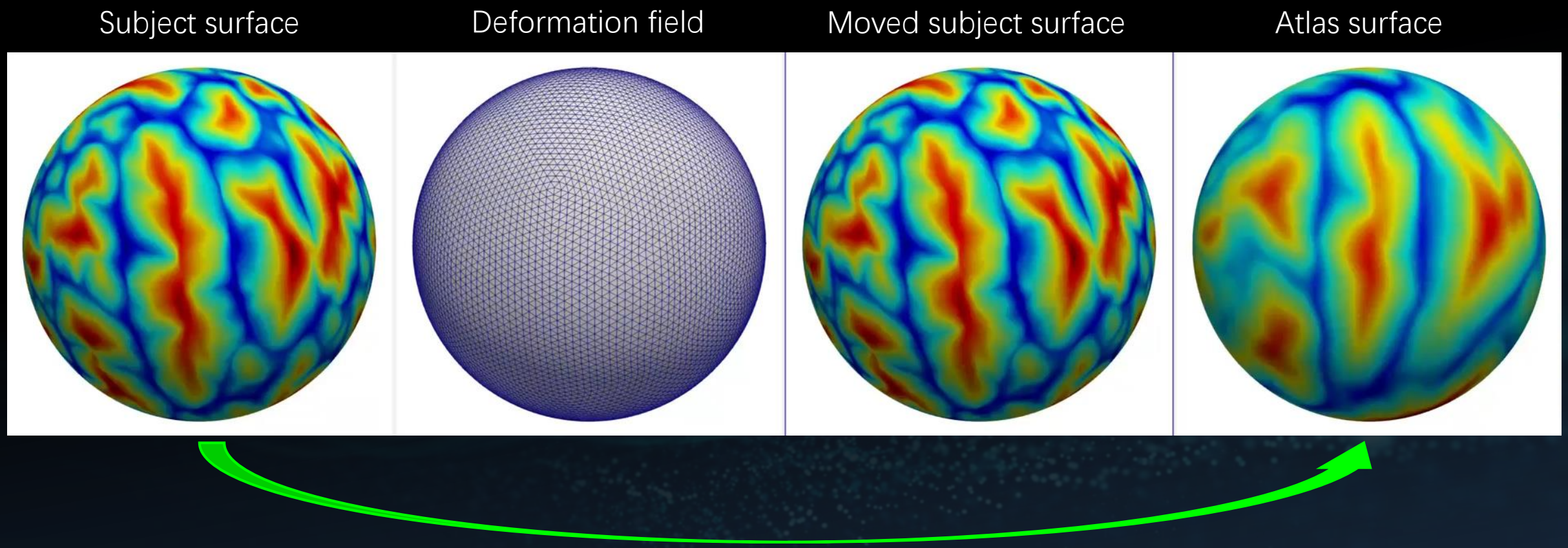
U-Nets



Liu, Q., Xu, Z., Bertasius, G., & Niethammer, M. (2023). SimpleClick: Interactive Image Segmentation with Simple Vision Transformers. ICCV., 22290-22300. 2023.

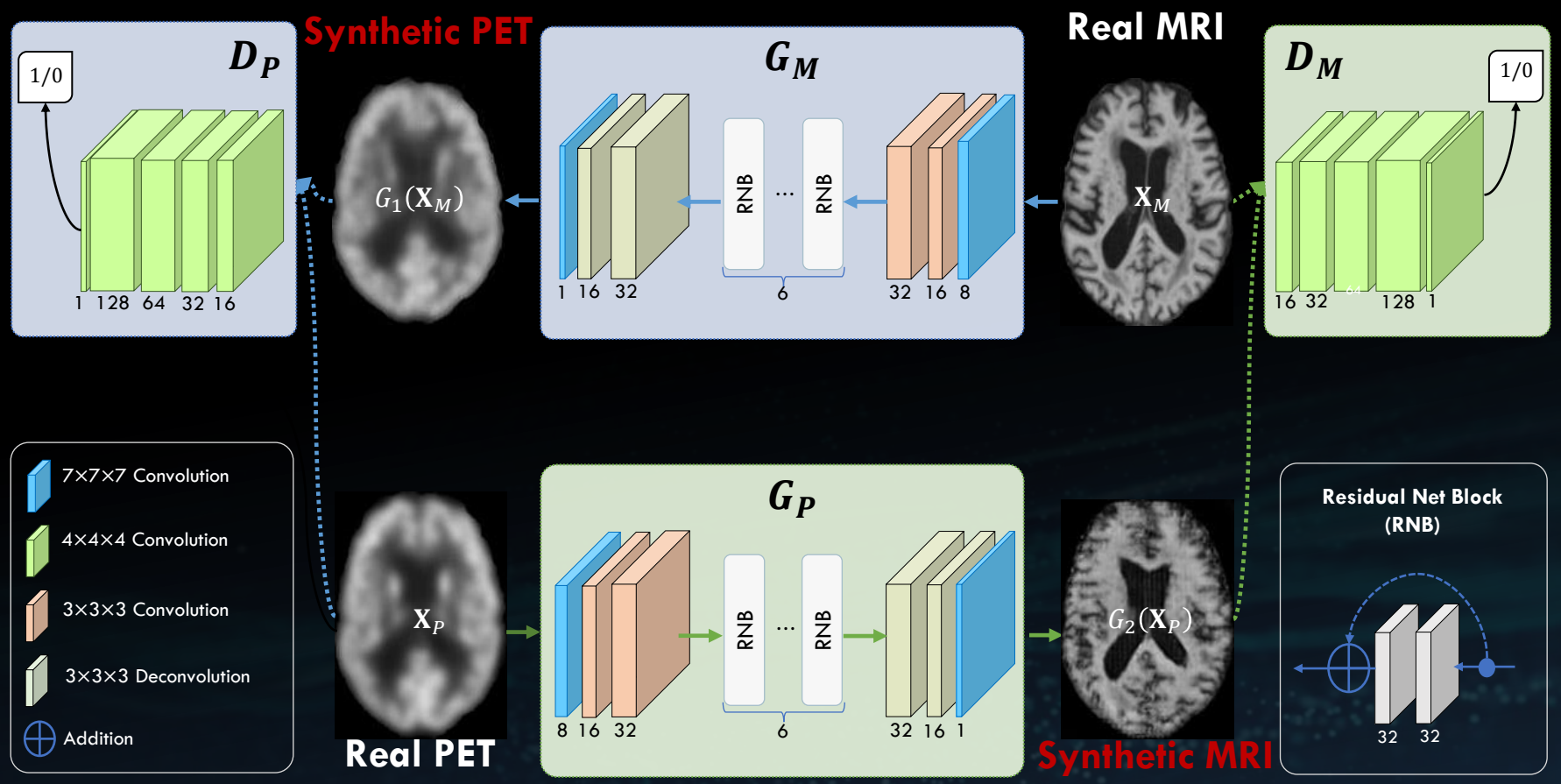
R. Azad *et al.*, “Medical Image Segmentation Review: The success of U-Net.” arXiv, Nov. 27, 2022.
Minaee, Shervin, et al. "Image segmentation using deep learning: A survey." *IEEE PAMI* 44.7 (2021): 3523-3542.

Superfast Spherical Surface Registration

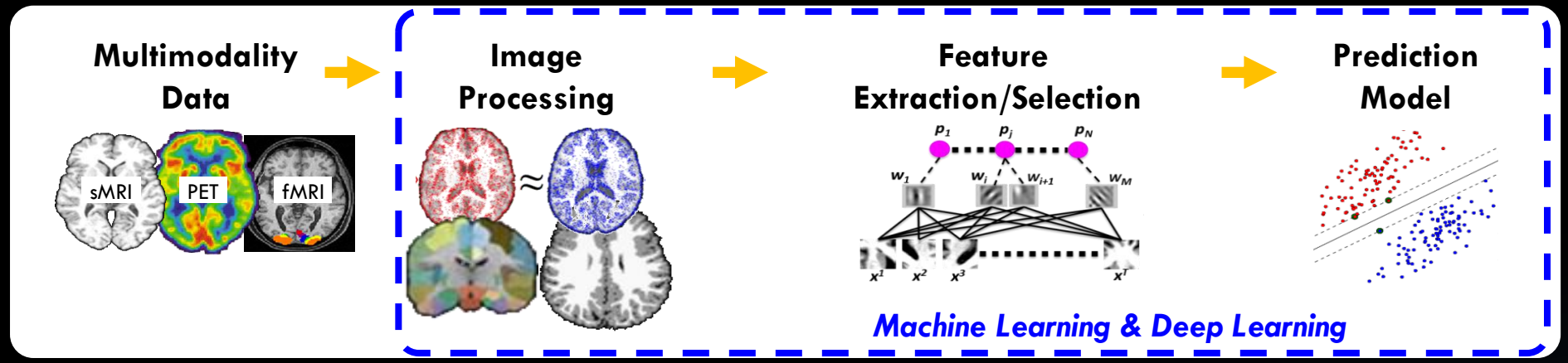


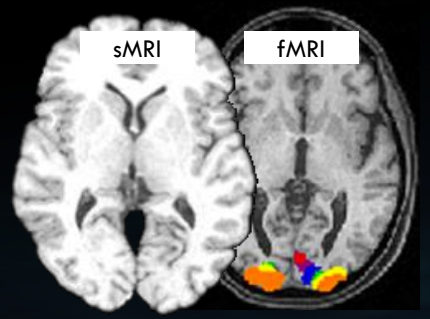
Zhao F, Wu Z, Wang F, Lin W, Xia S, Shen D, Wang L, Li G. S3Reg: Superfast Spherical Surface Registration Based on Deep Learning. IEEE Trans Med Imaging 2021; 40(8): 1964-1976.

Cross-Modality Image Synthesis

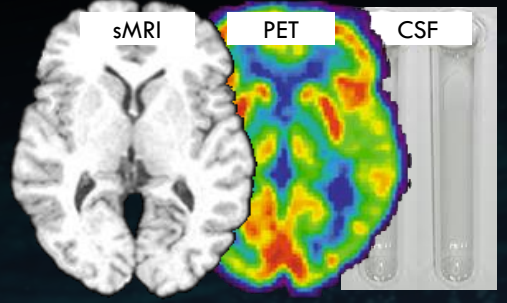


Computer-Aided Medical Data Analysis

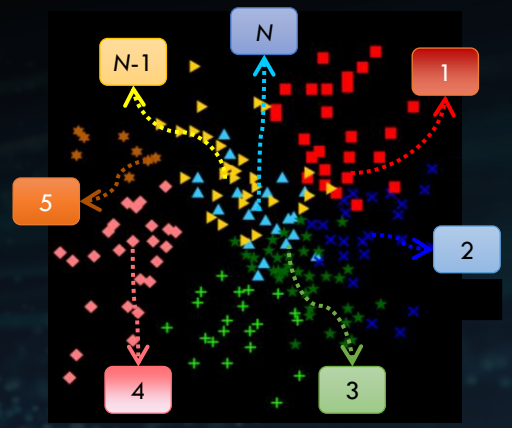




Neuroimage Representation Learning



Multimodality Data Fusion

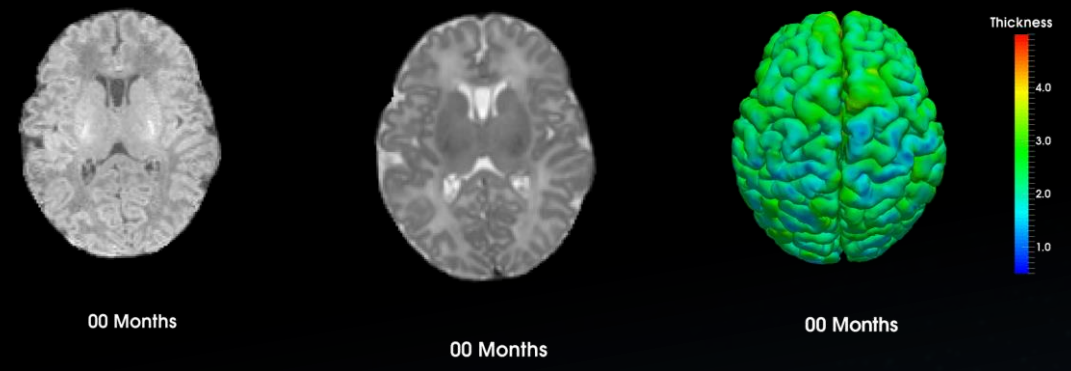


Multi-Site Data Adaptation

Major Challenges

Complex Organs and Tissues

Heterogeneity within Individual Subjects and across Centers/Studies



Image=

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

- **There is no publicly available, high-quality imaging datasets with detailed annotation information that cover a large spectrum of segmentation tasks in health care.**
- **How to quantify the uncertainty and generalizarability of organ atlas as well as deconvolution and structural learning models?**
- **How to develop DRL method for various segmentation and registration tasks?**

Part III

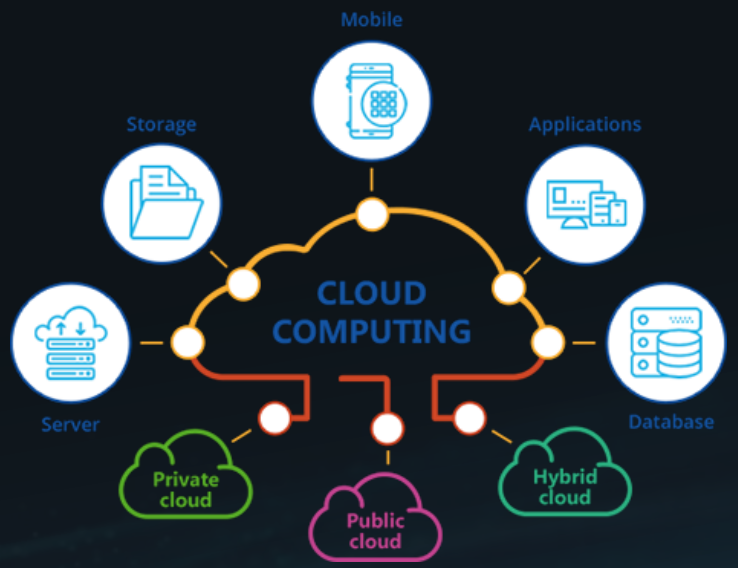
Opportunities for Statisticians in Advancing Medical Data Analysis

“The best thing about being a statistician is that you get to play in everyone's backyard.”
- John Tukey -

Application to ABC

Big Data

http://medium.com



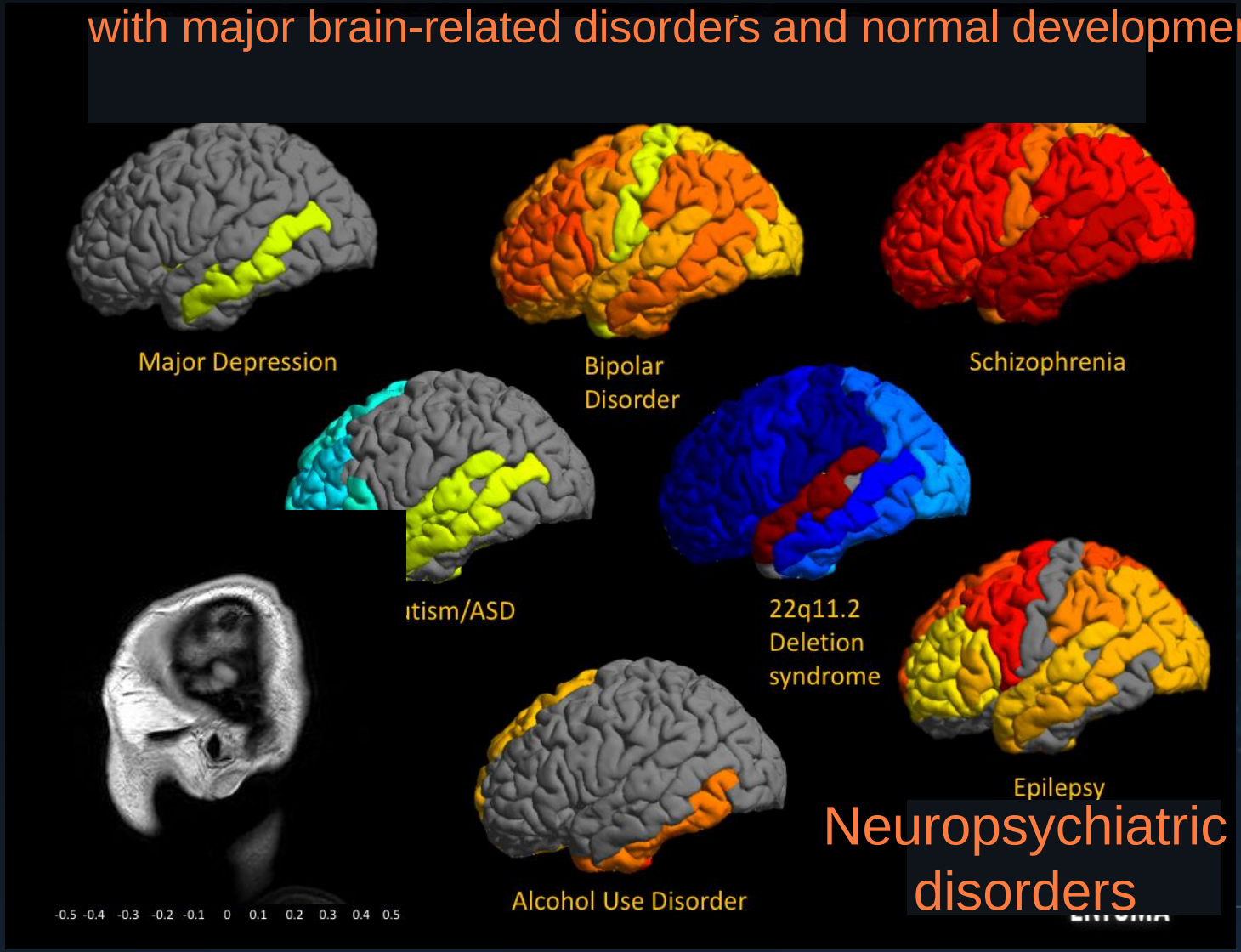
Computing

Analytical Tools

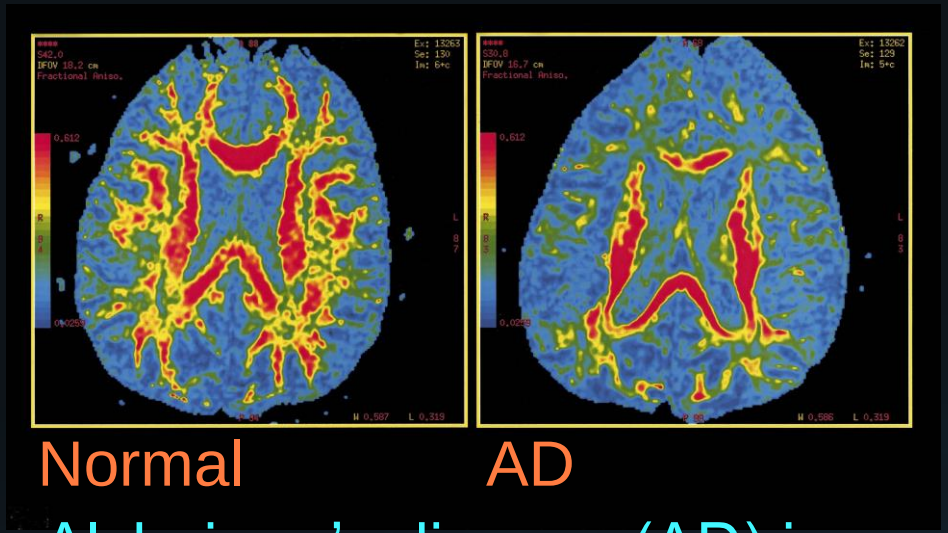
- Applied Mathematics
- Statistics
- Machine Learning
- Engineering

Brain Imaging for Brain Disorders

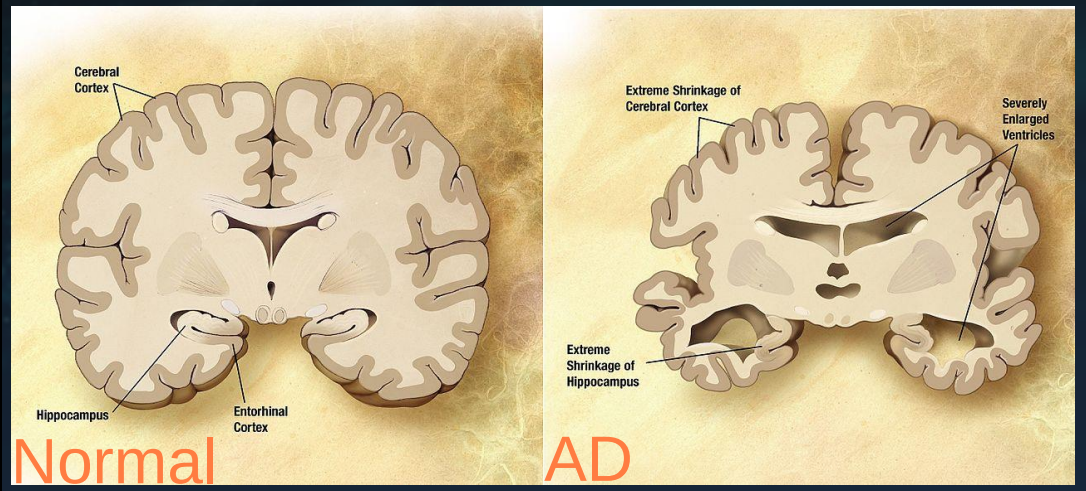
Capture the brain structure and function changes associated with major brain-related disorders and normal development



Neuropsychiatric disorders



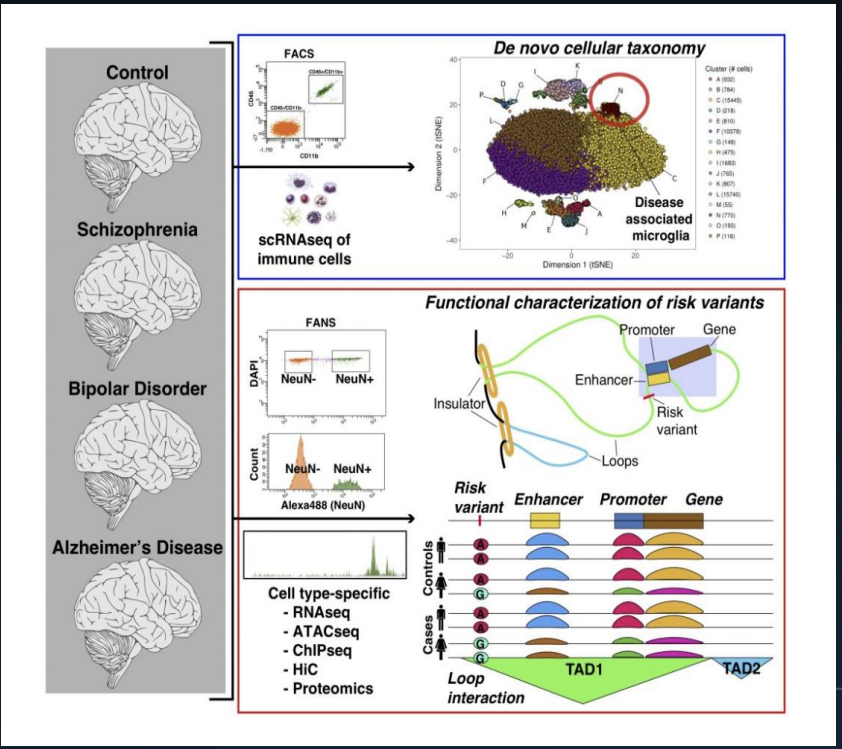
Alzheimer's disease (AD) is associated with brain shrinkage



Genetics of Brain Disorders

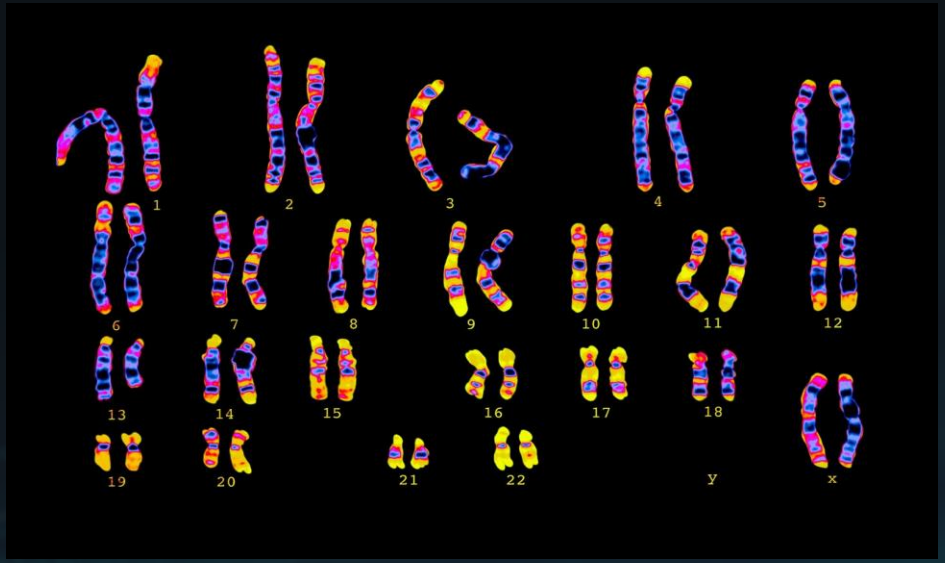
Most major brain disorders (like AD) are **heritable complex traits/diseases**

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



“Big Data” Imaging Cohorts

“Big data” Brain imaging datasets become available in recent few years

Systematically collect publicly available individual-level data for **> 120k** individuals

Build the largest database in this field

Aging Brain

BCP *(Age [0,5])*

PING *(Age [3,21])*

ABCD ***(n ~ 10k, Age [9,11])***

PNC *(Age [14,29])*

HCP *(Age [22,35])*

UK Biobank ***(n ~ 100k [Ongoing], Age [40,69]))***

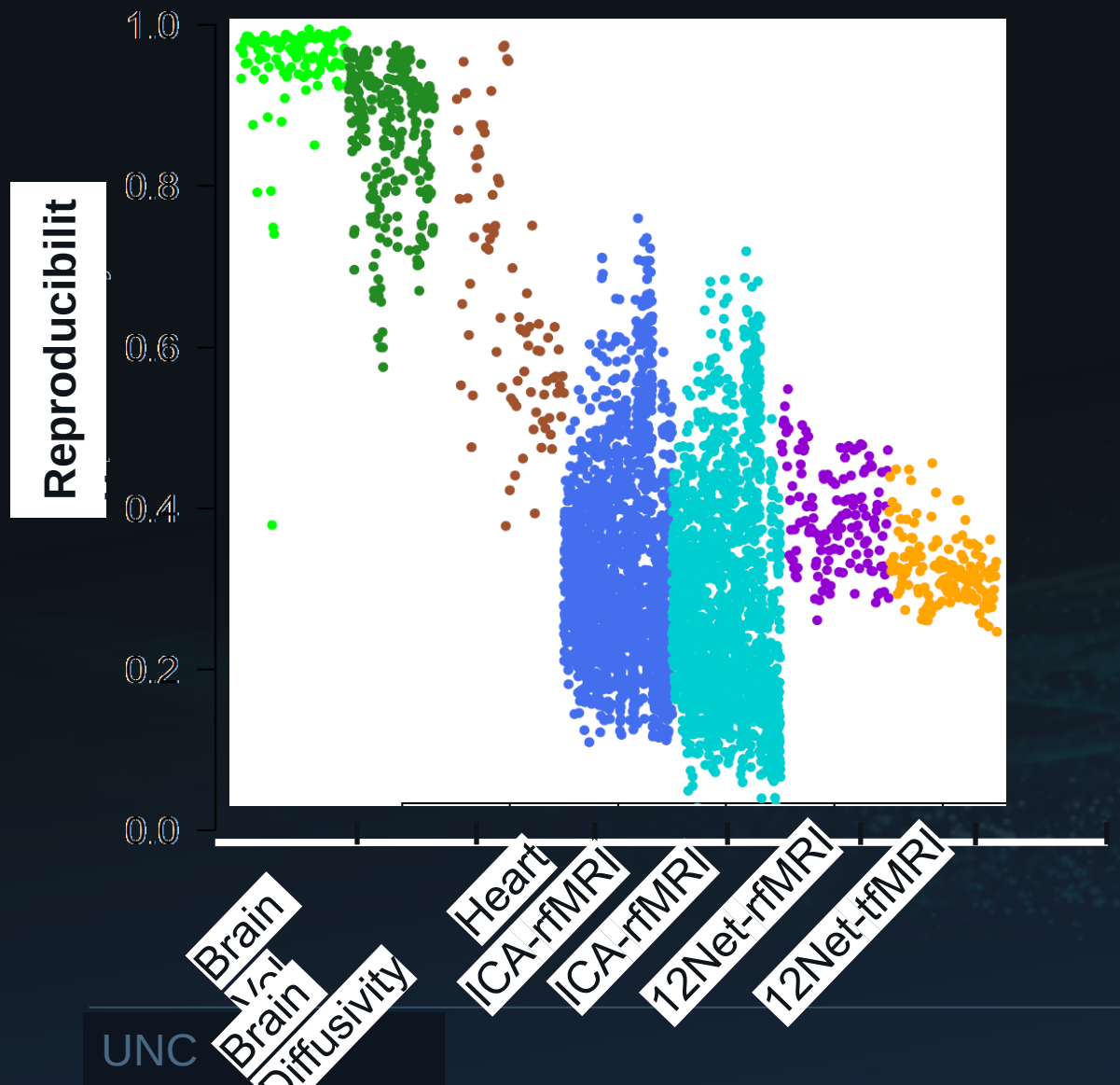
RADC *(Age > 65)*

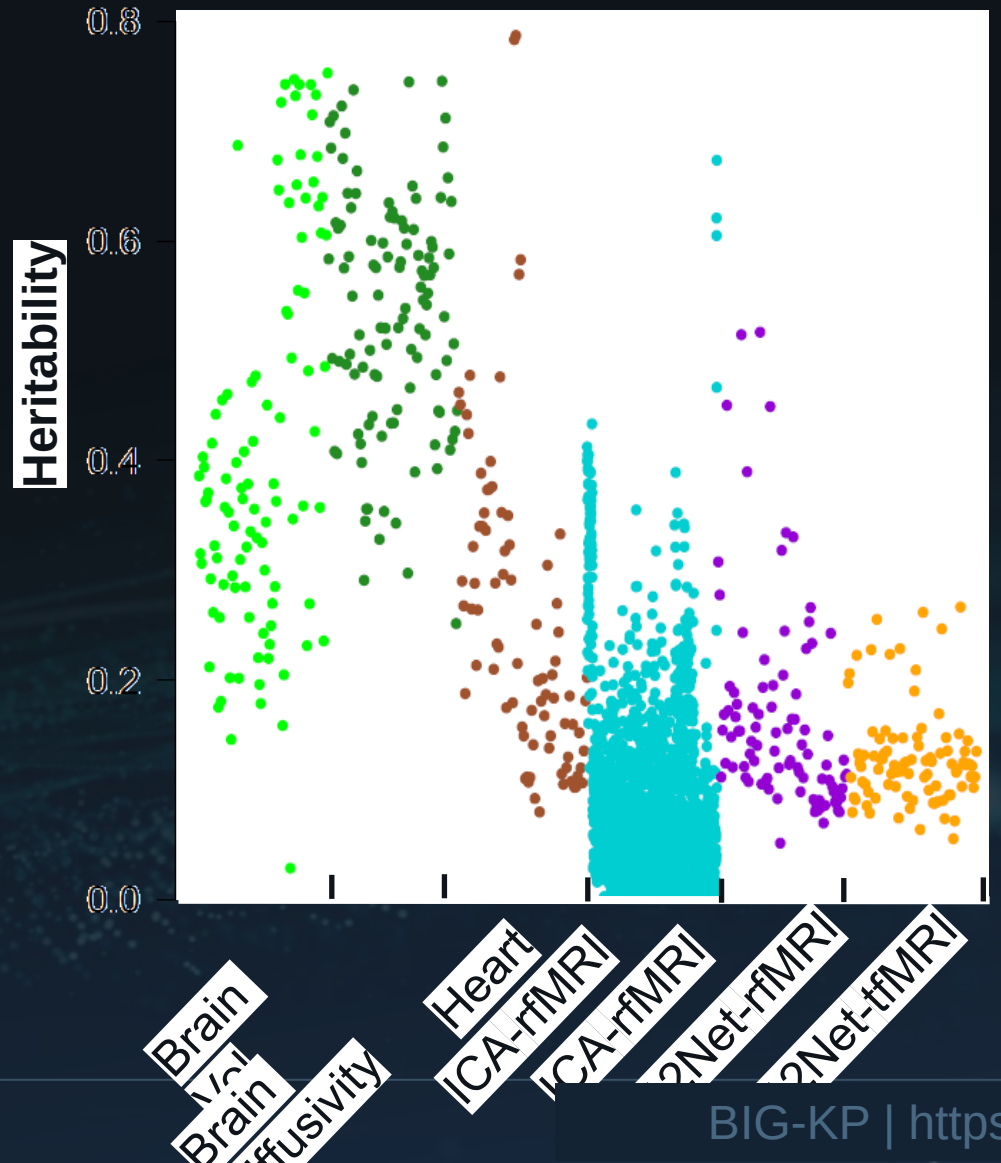
ADNI *(Age [55,92])*

Brain Development

IMAGEN *(Age [14,22])*

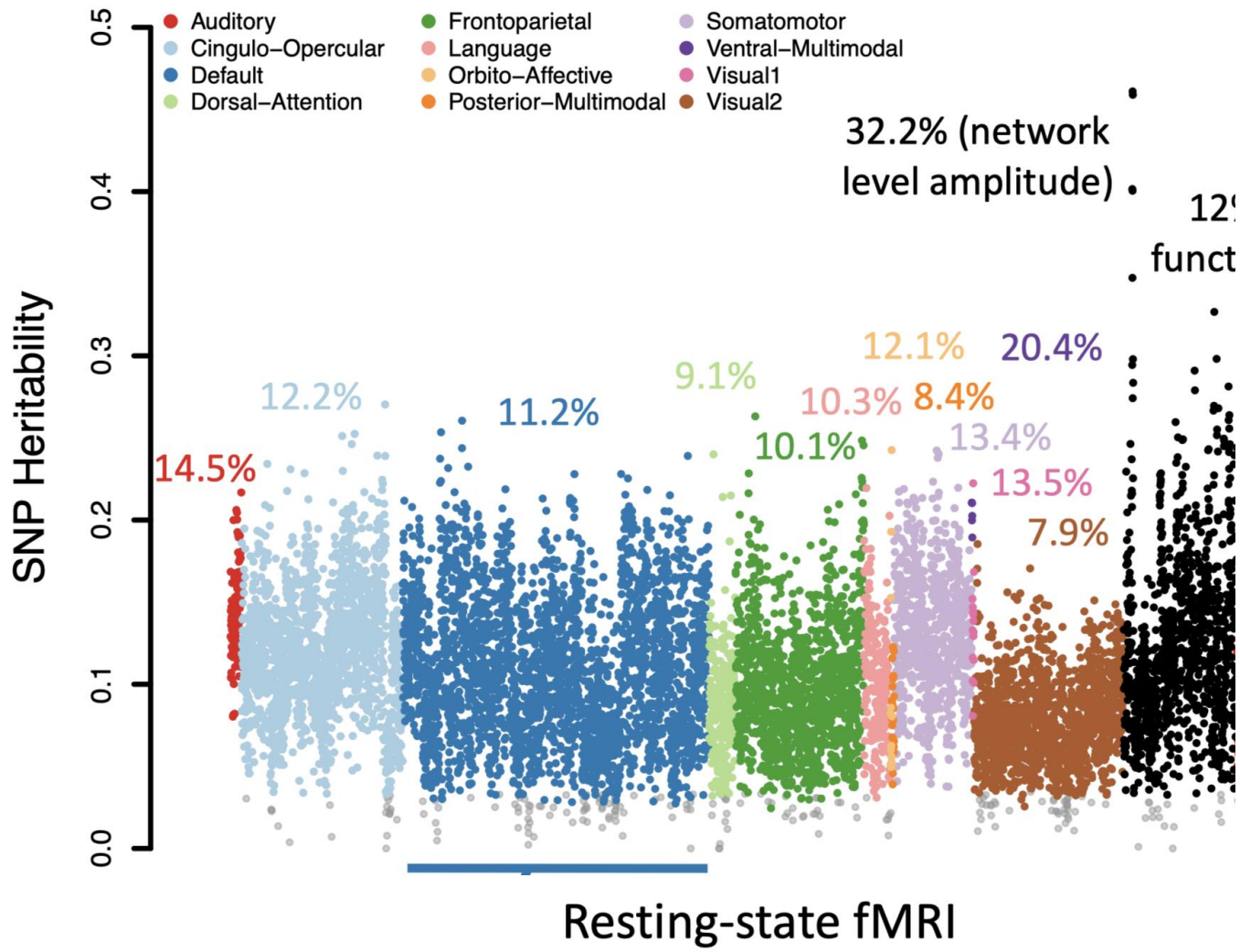
IG: Reproducibility and Heritability

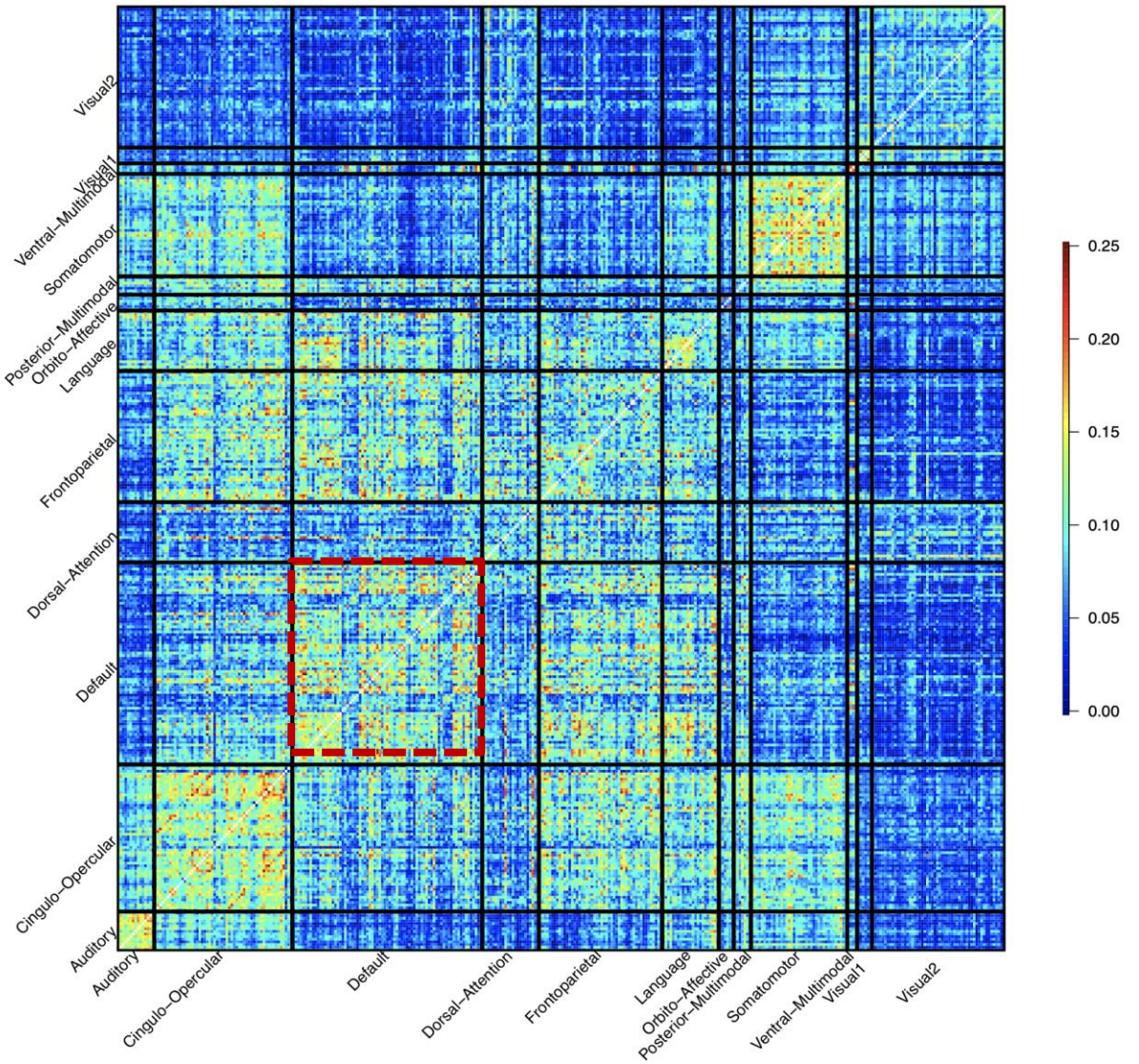




Area-level Heritability Pattern of Functional Brain

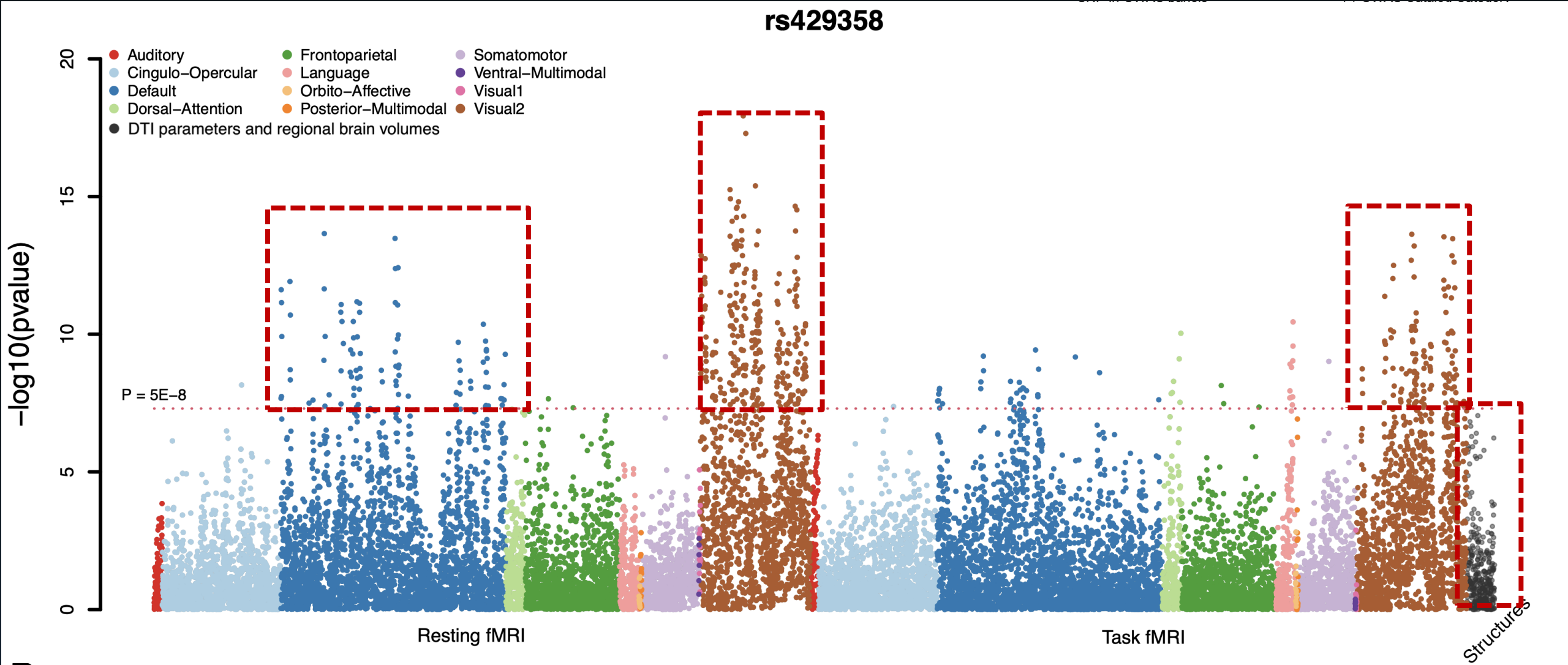
Fine details about the heritability pattern (> 64k fMRI connectivity traits among 360 regions)





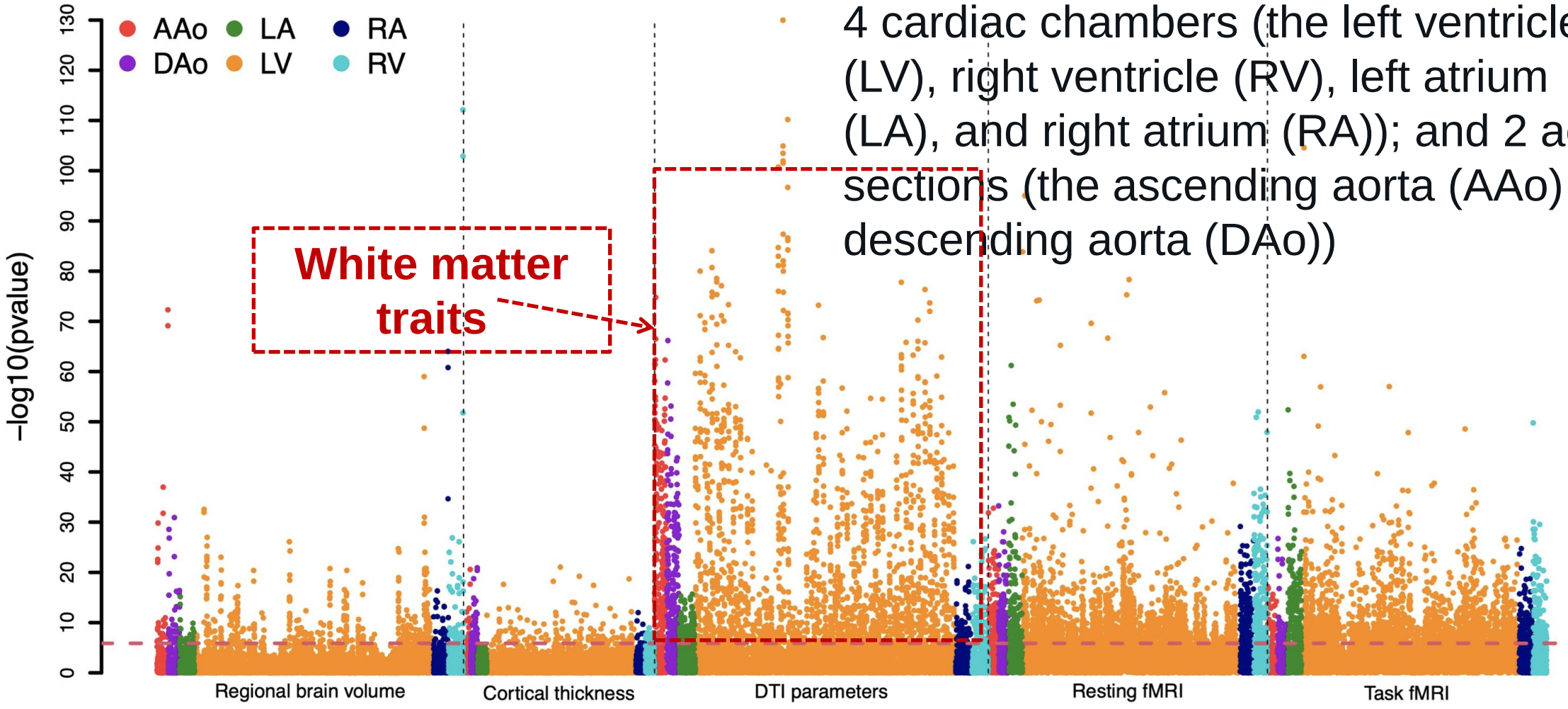
APOE-associations across functional networks

observations: 1) Enriched in the secondary visual and default mode networks;
2) Stronger connections in fMRI than in structural MRI.



Phenotypic Heart-Brain Connections

Heart imaging traits are widely associated with regional brain volumes, cortical thickness, white matter microstructures, and fMRI traits.



4 cardiac chambers (the left ventricle (LV), right ventricle (RV), left atrium (LA), and right atrium (RA)); and 2 aortic sections (the ascending aorta (AAo) and descending aorta (DAo))

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**,

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 v -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 with cognitive and mental health 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

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

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. 309 parcellation-based functional MRI (task/rs-fMRI(Glasser360))

Genetics discovery in human brain by big data integration

GWAS Summary Statistics

The full set of GWAS summary statistics have been made freely available to the research community

Resources with the largest sample size (>11156 page views since Sep 2019)

GWAS Summary Statistics for Brain Imaging Phenotypes

Involved datasets: UK Biobank (UKB), Adolescent Brain Cognitive Development (ABCD) Study, Human Connectome Project (HCP), Philadelphia Neurodevelopmental Cohort (PNC), Alzheimer's Disease` Neuroimaging Initiative (ADNI), Pediatric Imaging, Neurocognition, and Genetics (PING)

Terms of Use:

- By downloading these data, you acknowledge that they will be used for research purposes and that you are in compliance with applicable rules, policies and regulations.
- When reporting results of research that utilizes these data we request that you cite the original publication.

GWAS summary statistics for 200 resting-state functional MRI (rs-fMRI) traits

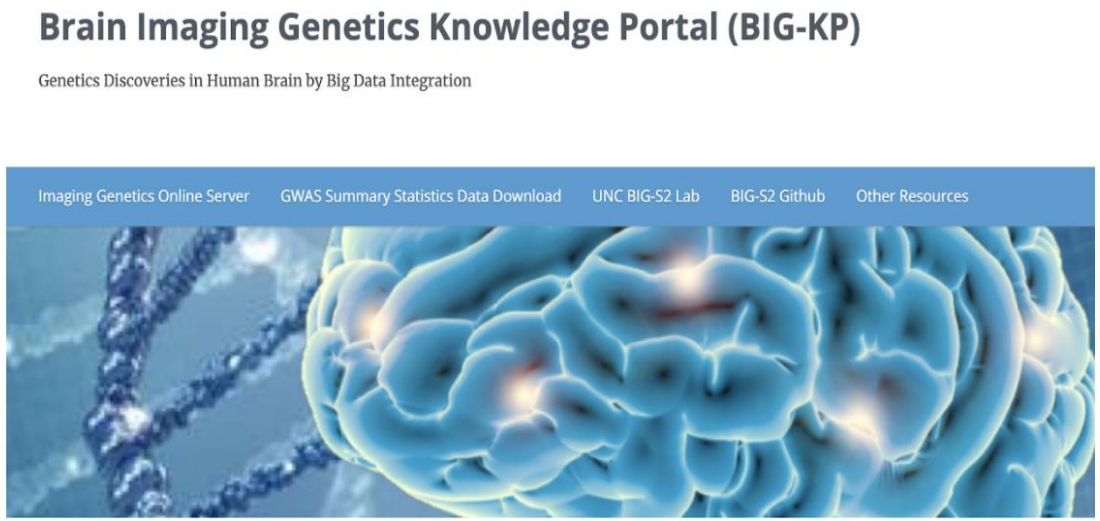
- **Sample size:** n=34,691
- **Version**: July 15, 2020
- Download Summary Statistics:

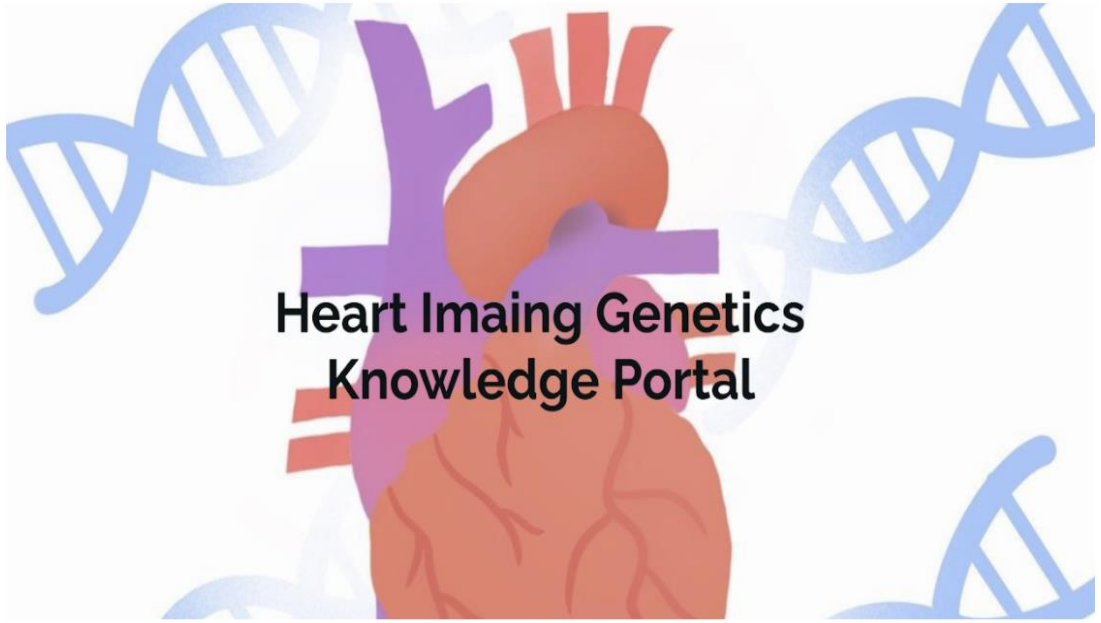
```
wget --no-check-certificate --content-disposition https://raw.
githubusercontent.com/stat-yyang/sumstats/master/fMRI.list
wget -i fMRI.list
```

- **Description**: readme
- **Citation**: Zhao et al (2020) Common variants contribute to intrinsic

Contents [hide]

Brain- Heart Imaging Genetics Knowledge Portal





Brain Imaging Genetics Knowledge Portal (BIG-KP)

Heart Imaging Genetics Knowledge Portal (Heart-KP)

Aim to build the best knowledge database of neuroimaging genetics

Important Statistical Topics

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

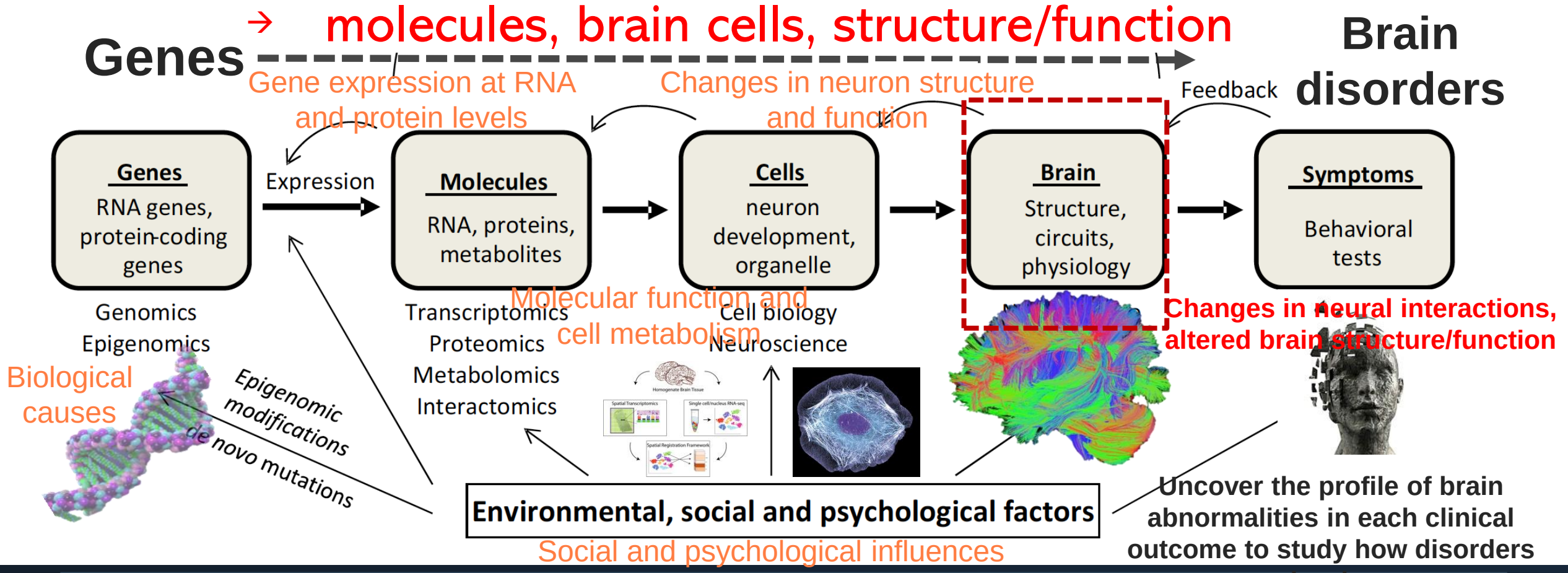
- ➢ Dimension Reduction Methods
- ➢ Image Genetics
- ➢ Causality Research
- ➢ Predictive Analysis
- ➢ Knowledge-based Methods
- ➢ Reinforcement Learning

Zhu, H., Li, T., & Zhao, B. Statistical learning methods for neuroimaging data analysis with applications. *Annual Review of Biomedical Data Science, Volume 6, Issue 1, 2023.*

Other Important Topics

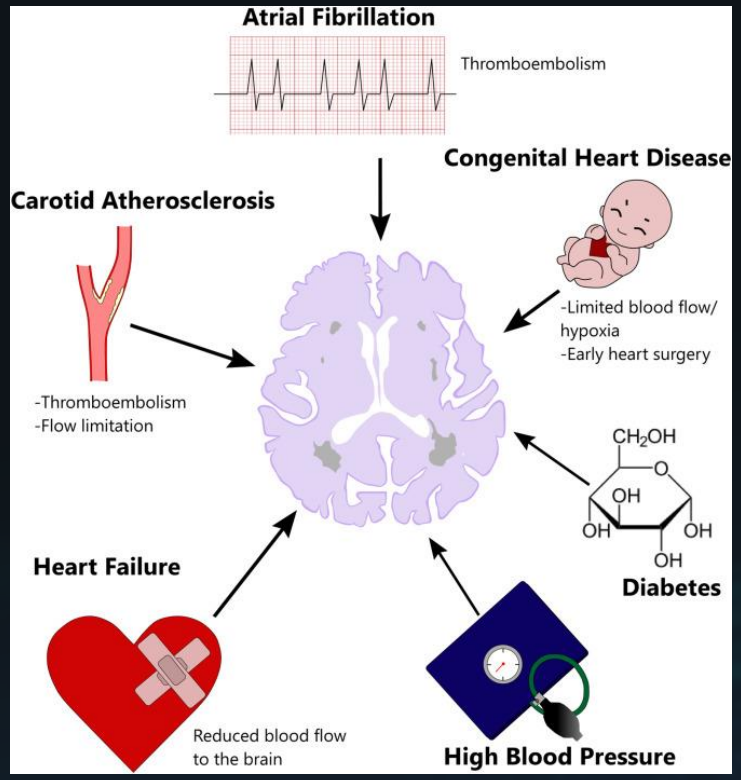
Brain Imaging Genetics Paradigm

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

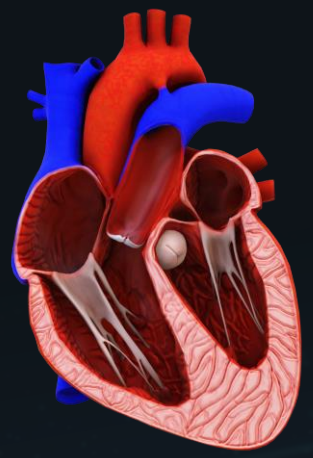


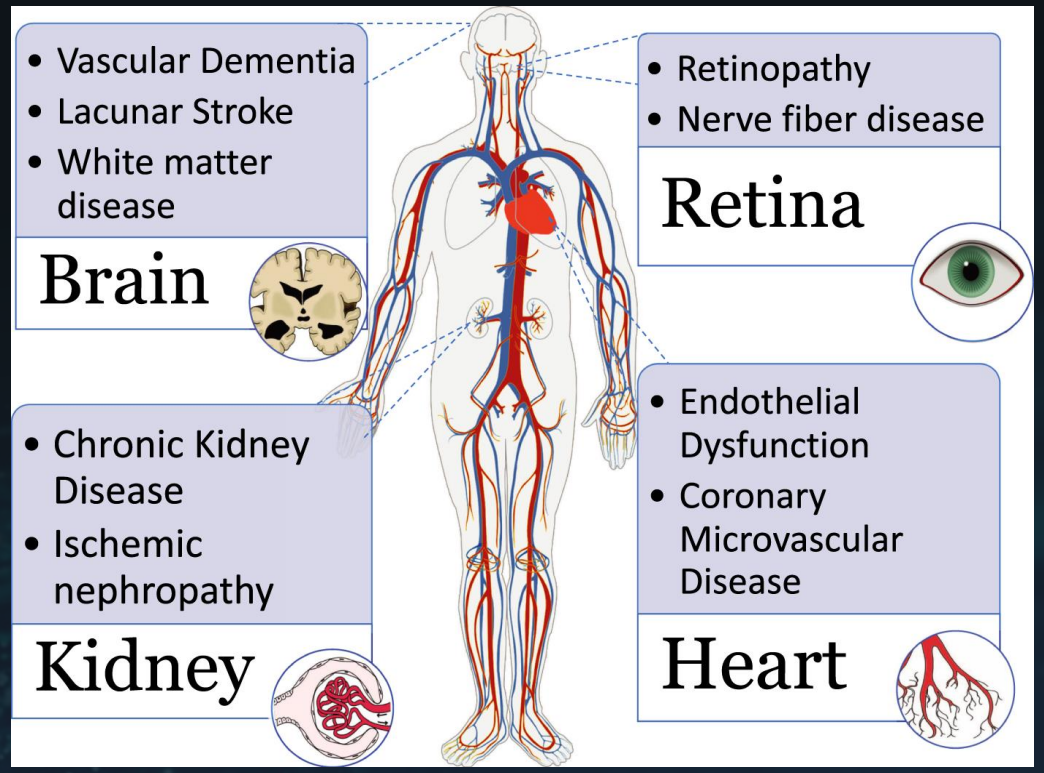
Cardiovascular Disease & Brain Health

(Neuro)imaging: help understand the complex interplay between brain and other human organs and their underlying genetic overlaps



Possible causal factors of brain structure changes, resulting in brain disorders like stroke, dementia and cognitive impairment

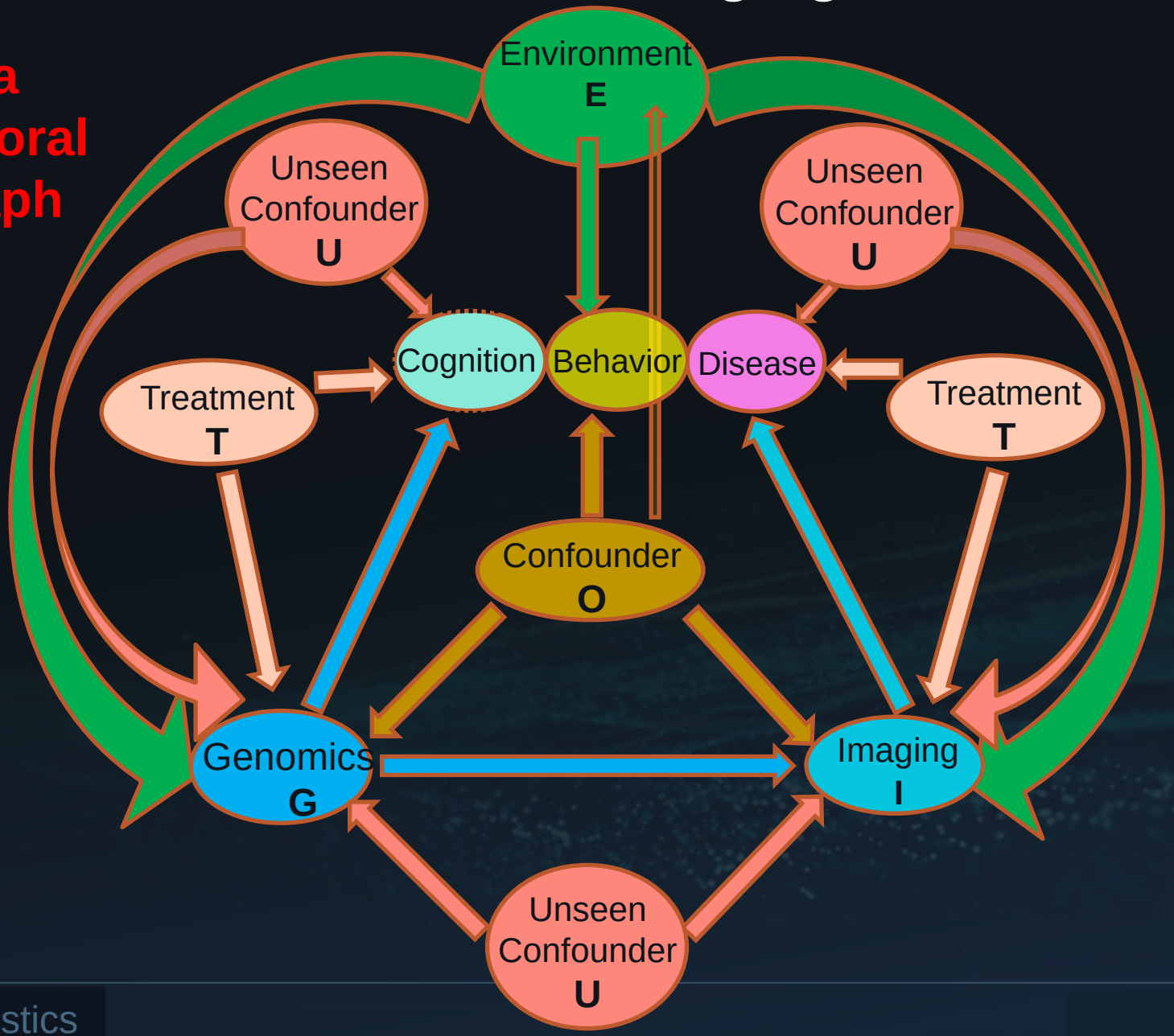


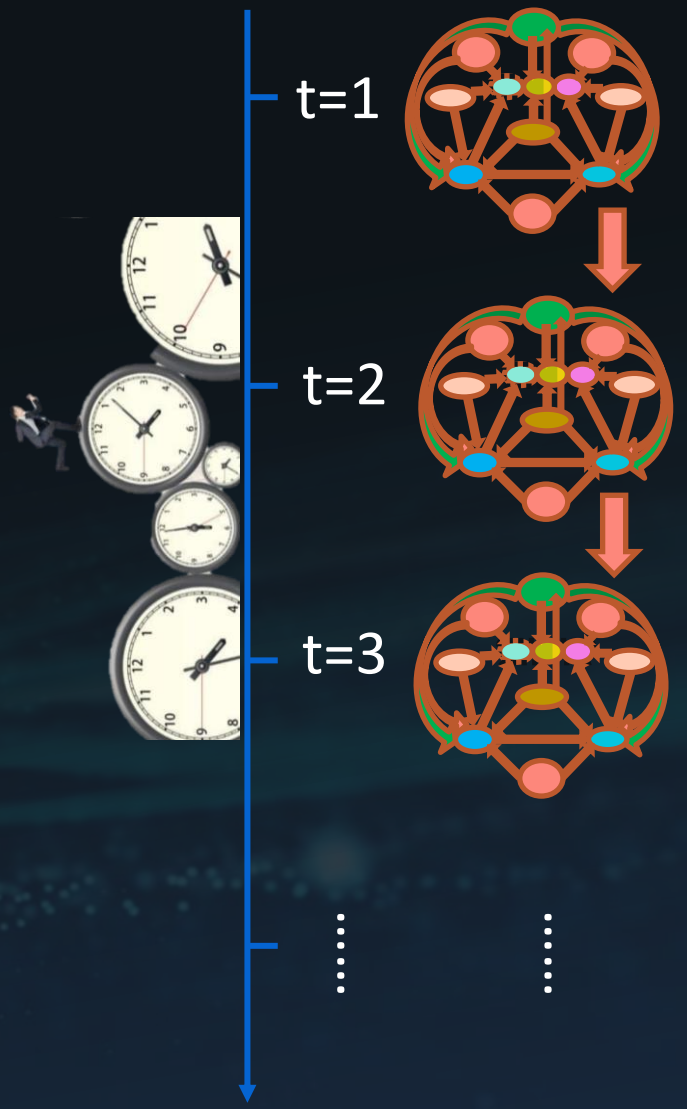


Many diseases (e.g., microvascular disease, high blood pressure) are multisystem disorders

Causal Genetics Imaging Clinical Pathway

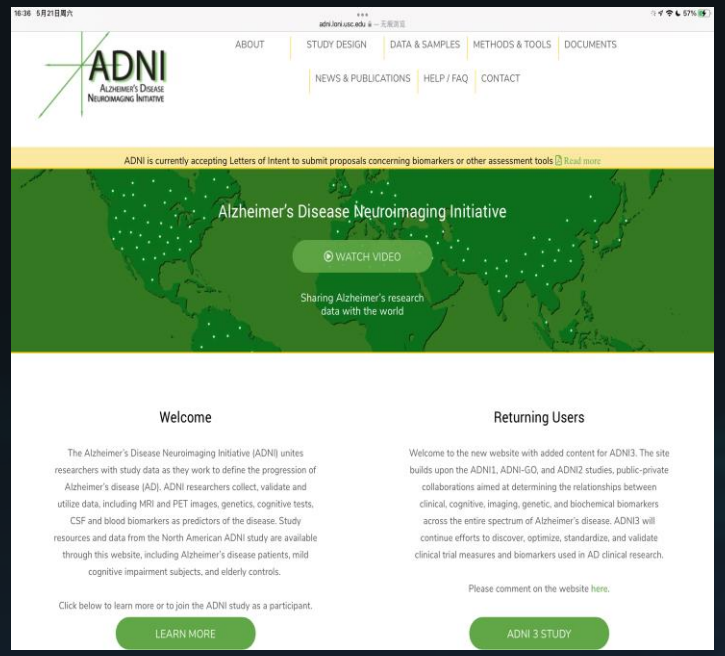
CGIC is a spatiotemporal causal graph

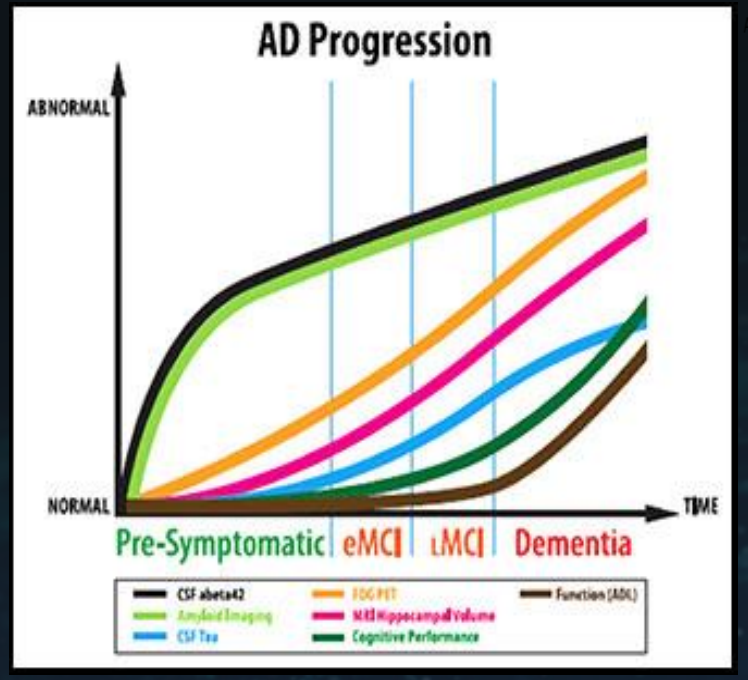


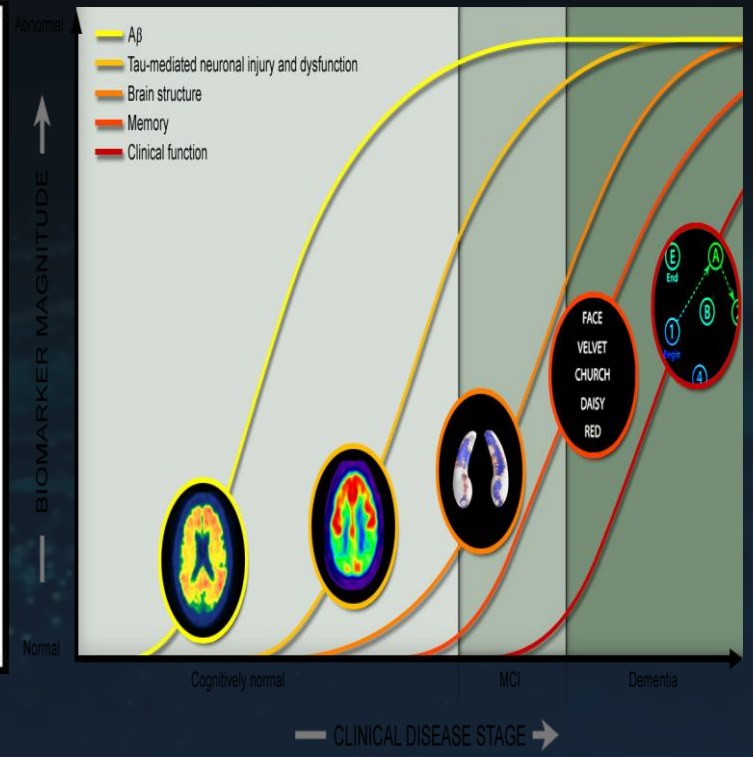


Alzheimer’s Disease Neuroimaging Initiative

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.



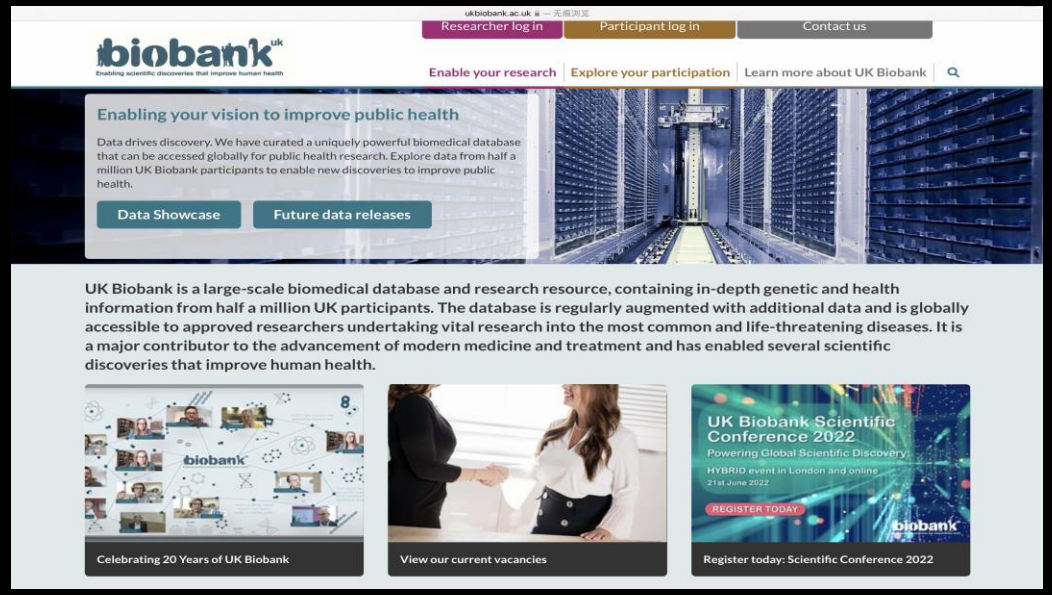




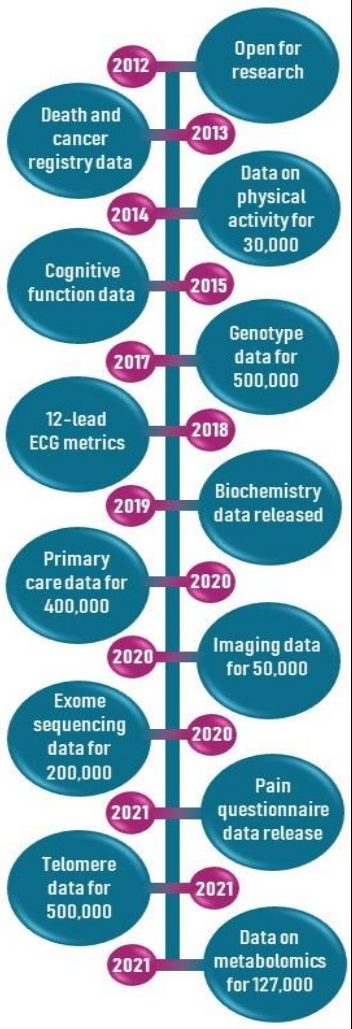
2004-now

The UK Biobank Study

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

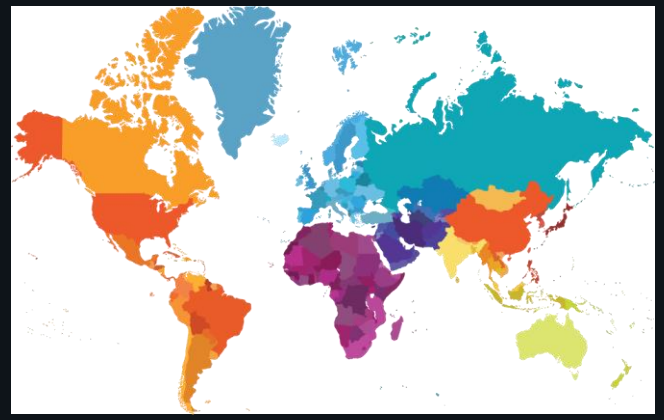


2006-now

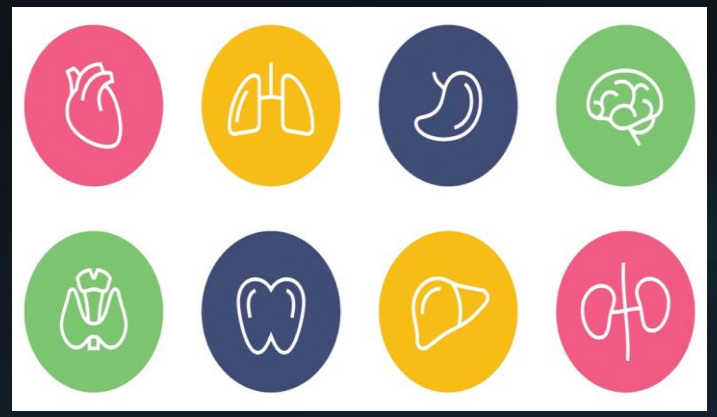


- **Imaging:** Brain, heart and full body MR imaging, plus full body DEXA scan of the bones and joints and an ultrasound of the carotid arteries. The goal is to image 100,000 participants, and to invite participants back for a repeat scan some years later.
- **Genetics:** Genotyping, whole exome sequencing & 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.

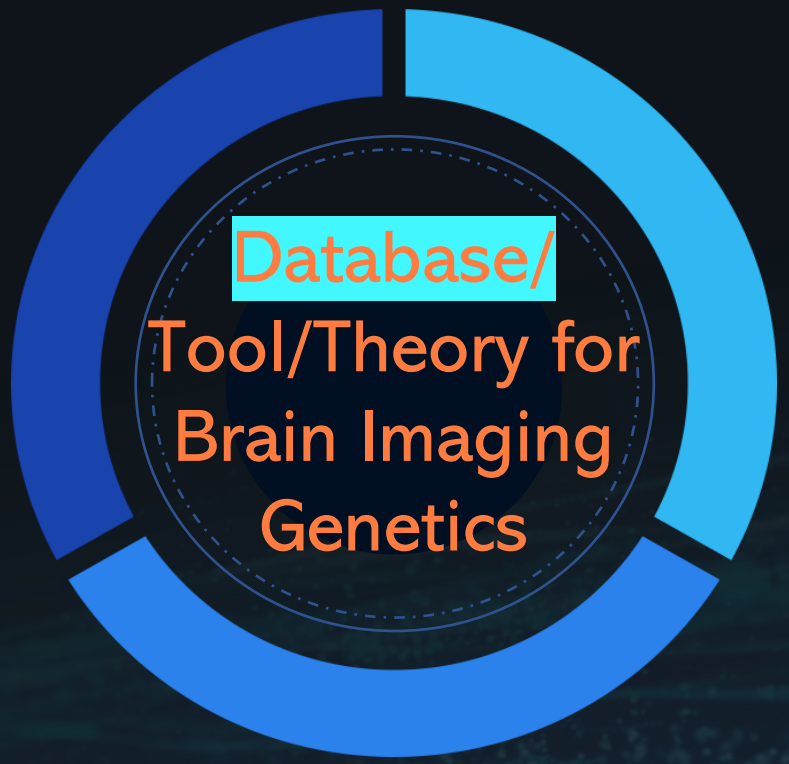
Methodological Challenges

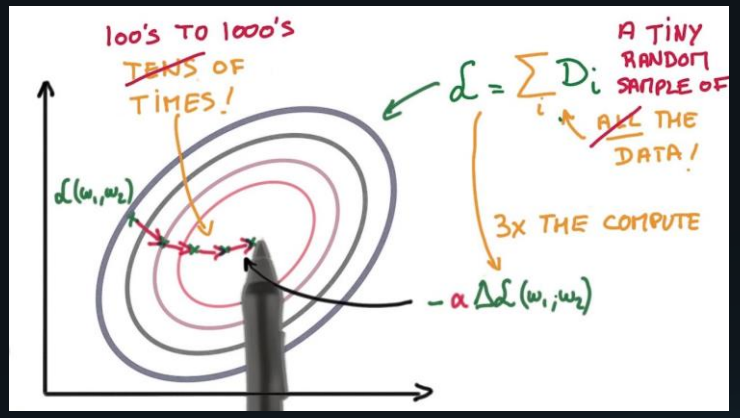


Multiple Biobanks/Trials Integration
(e.g., Heterogeneity in global populations)

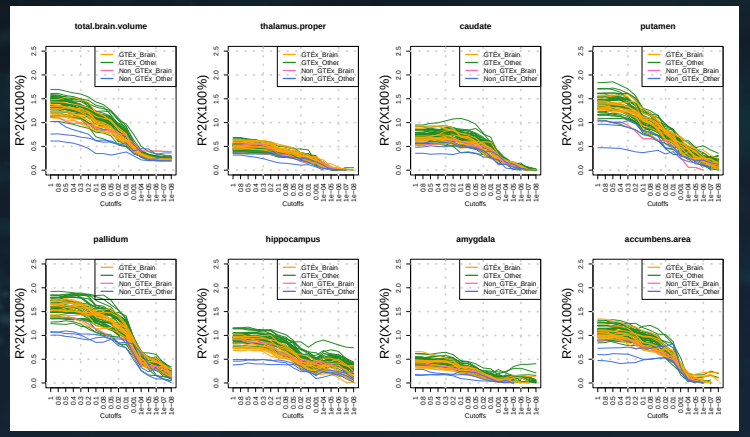


Omics Data Integration
(e.g., new tech, biological pathway)



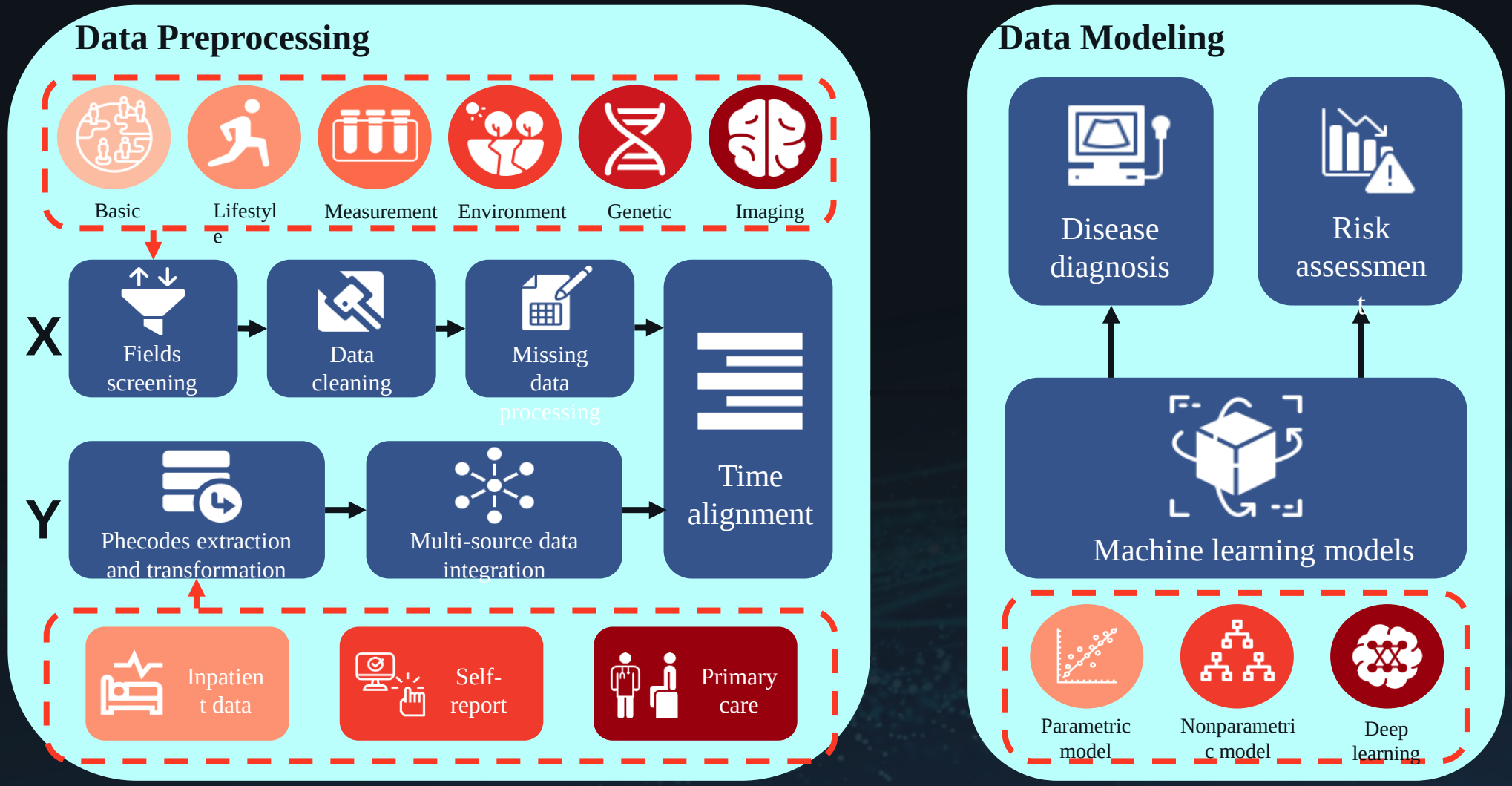


New Computational Tools
(e.g., challenge of dense signal in biobank-scale database)



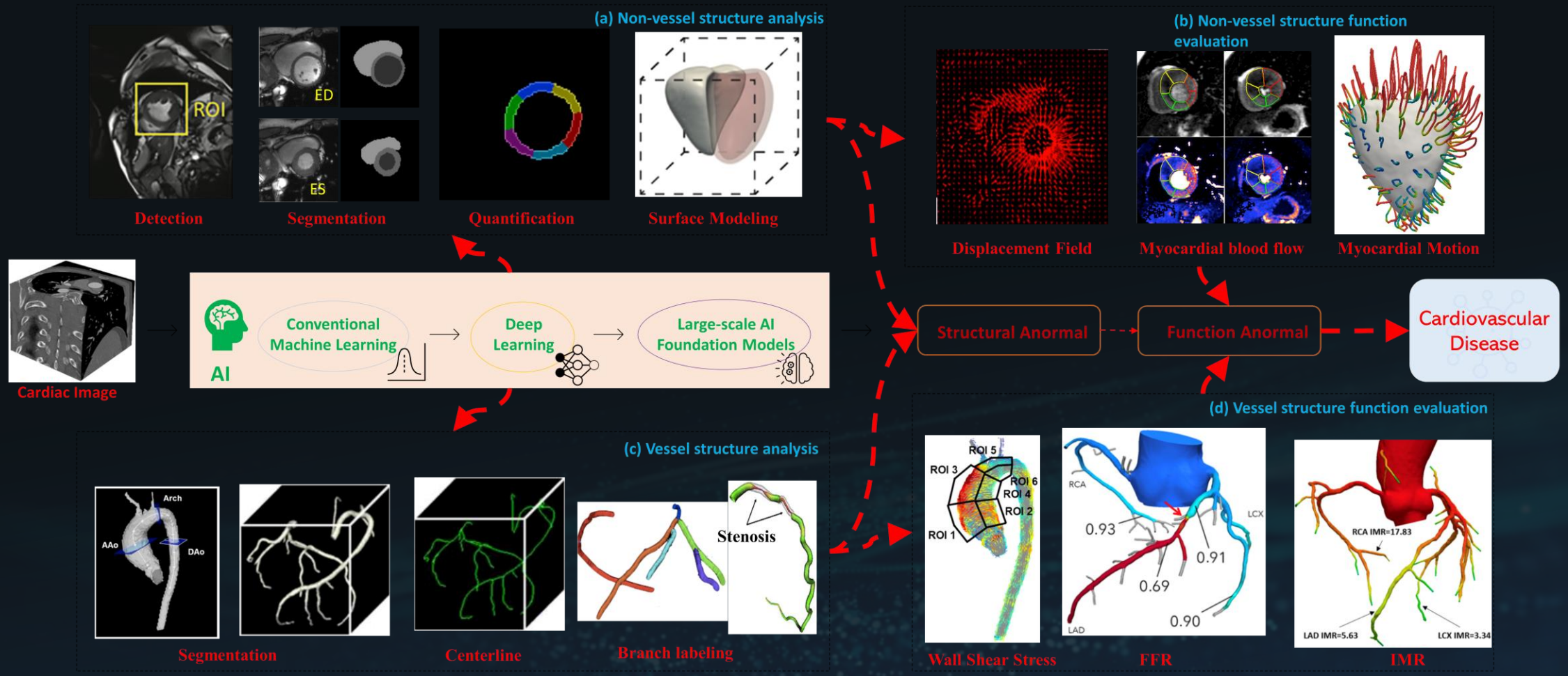
Advanced Methods for Dense Signals
(e.g., deep learning)

Data Preprocessing and Data Modeling



Jiang.et al. (2024). UKBFound: A Foundation Model for Multi-Disease Prediction and Individual Risk Assessment Based on UK Biobank Data

Image Analysis Pipeline



Prediction Models

Preprocess

Baseline | Lifestyle | Measurement

Baseline | Lifestyle | Measurement

Baseline | Lifestyle | Measurement

Purpose determined data input

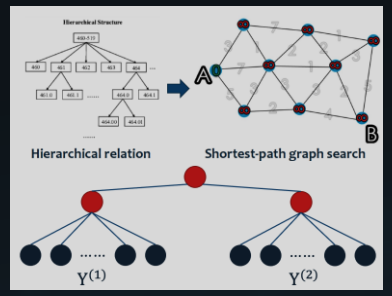
Baseline | Imaging | Environment
Genetic | Lifestyle | Disease Diagnosed | Measurement

Baseline | Imaging | Environment
Genetic | Lifestyle | Disease Diagnosed | Measurement

Model

Boosting | POPDX | Neural network | Logistic regression | ...

CoxPH | Deep Surv | Deep Hit | POPDx Surv | ...

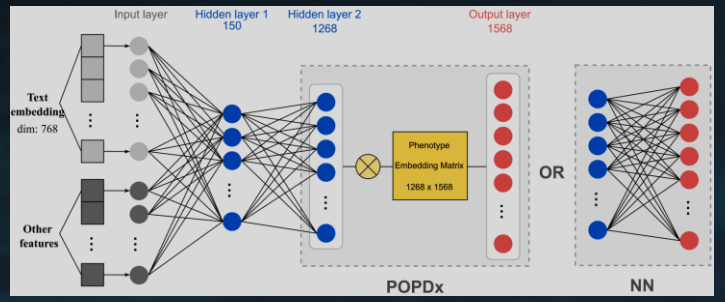


Disease diagnosis & Risk assessment

Individual disease

Categorized disease

Joint diagnosis and evaluation



Post analysis

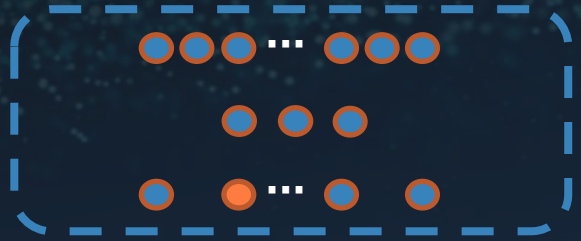
Feature importance extraction

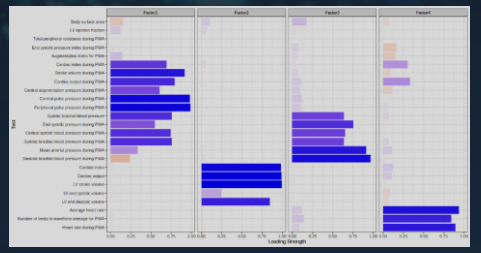
Deep Lift

SHAP

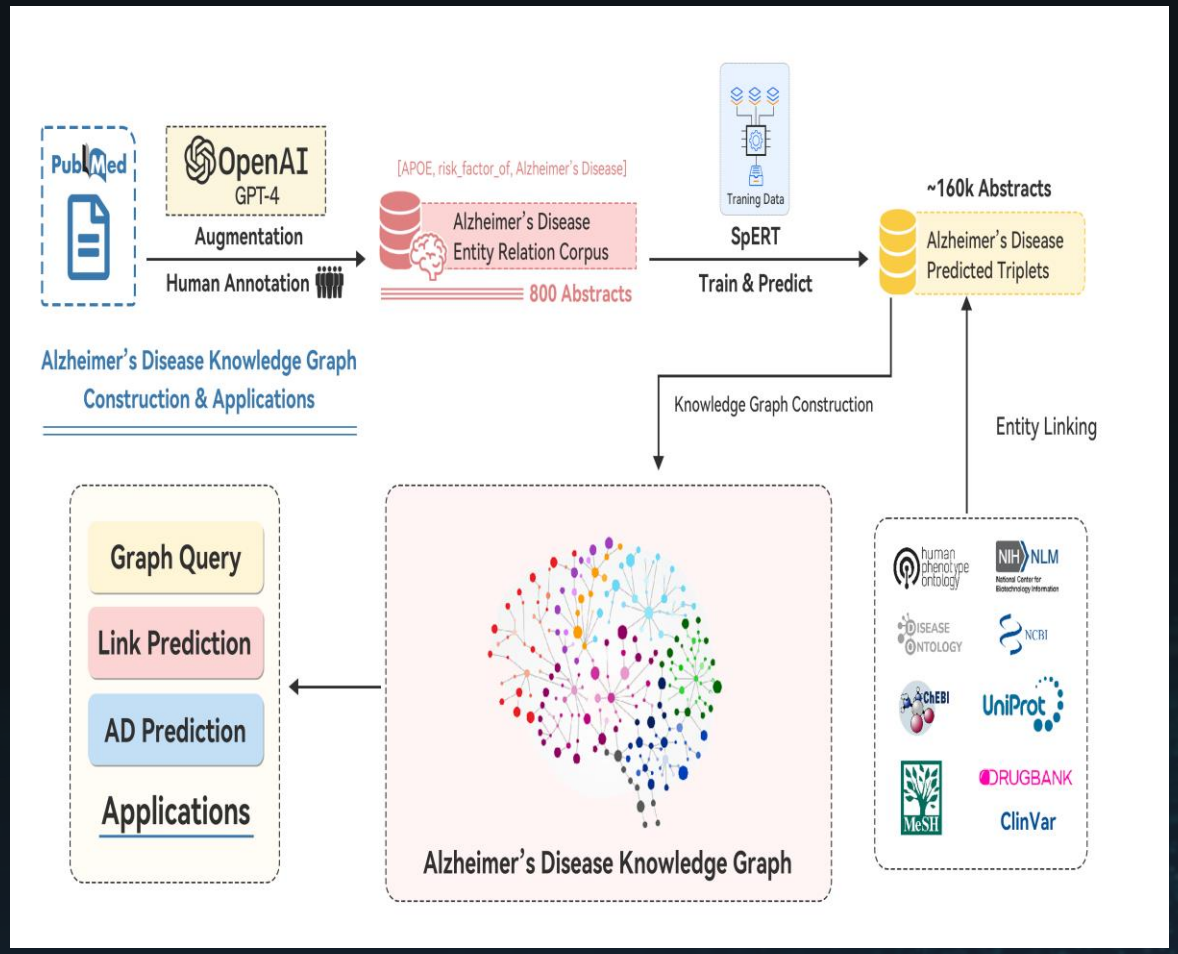
Integrated Gradients

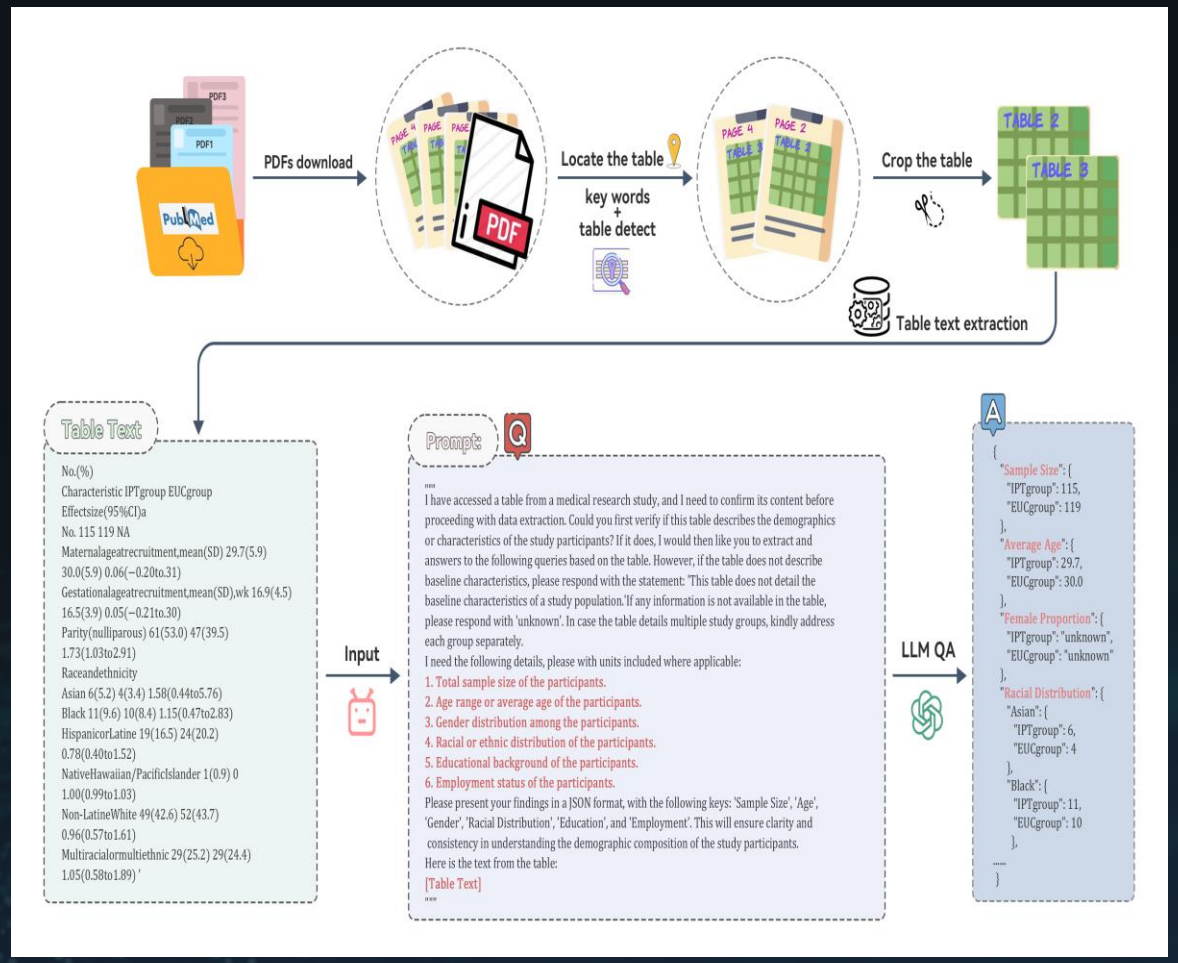
Data-driven dimension reduction





Knowledge Graph Construction





Yang et al., Alzheimer's Disease Knowledge Graph Enhances Knowledge Discovery and Disease Prediction.
Gao et al., Empowering Mental Health Insights: The Synergy of Knowledge Graphs and Large Language Models

Foundation Models for GMAI and Pan Biobank

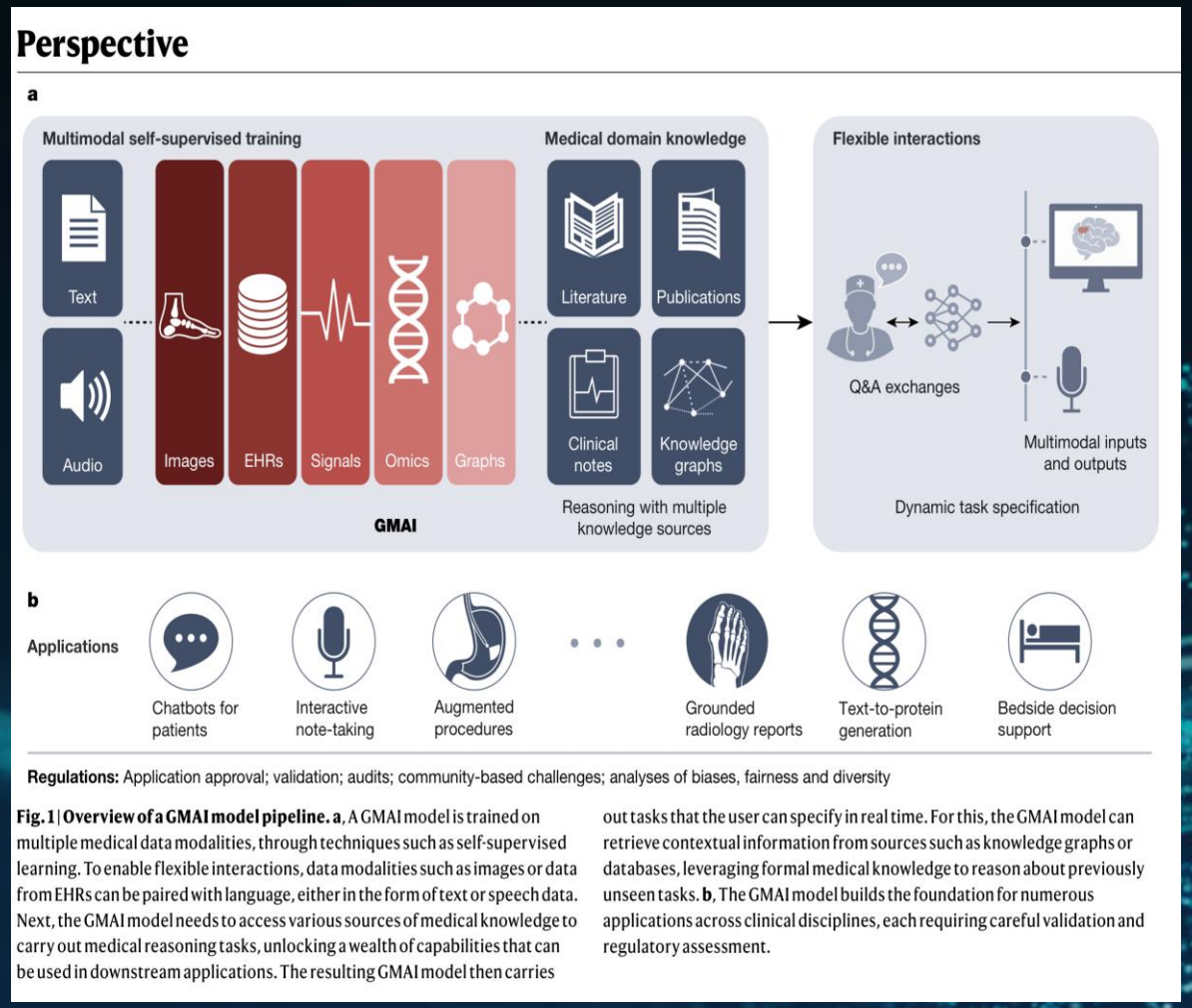


Fig. 1 | Overview of a GMAI model pipeline. a, A GMAI model is trained on multiple medical data modalities, through techniques such as self-supervised learning. To enable flexible interactions, data modalities such as images or data from EHRs can be paired with language, either in the form of text or speech data. Next, the GMAI model needs to access various sources of medical knowledge to carry out medical reasoning tasks, unlocking a wealth of capabilities that can be used in downstream applications. The resulting GMAI model then carries out tasks that the user can specify in real time. For this, the GMAI model can retrieve contextual information from sources such as knowledge graphs or databases, leveraging formal medical knowledge to reason about previously unseen tasks. **b**, The GMAI model builds the foundation for numerous applications across clinical disciplines, each requiring careful validation and regulatory assessment.

Moor, M., … ., Rajpurkar, P. (2023) Foundation models for generalist medical artificial intelligence. *Nature*.

Pan-biobank studies

Part IV

Statistical Causal Models

“Causation is not merely a useful concept, it is fundamental to our understanding of the world. Without causal inference, we are merely describing patterns, not explaining them.”

-Judea Pearl-

PFLM

- Consider a high-dimensional Partially Functional Linear Model (PFLM)

$$Y_i = \alpha + X_i^\top \beta + \int_{\mathcal{T}} Z_i(t)\xi(t)dt + \epsilon_i, \qquad i = 1, \dots, n$$

- Estimation

$$\min_{\beta \in \mathbb{R}^p, \xi \in \mathcal{H}} \left\{ (2n)^{-1} \sum_{i=1}^{n} \left[Y_i - \left(X_i^\top \beta + \int_{\mathcal{T}} Z_i(t)\xi(t)dt \right) \right]^2 + \tau ||\beta||_0 + 0.5\lambda ||\xi||_{\mathcal{H}} \right\}$$

- Representer Theorem:

$$\hat{\xi}(\beta) = \sum_{i=1}^{n} c_i(\beta) \left(\int_{\mathcal{T}} K(s,t) Z_i(s) ds \right) \longrightarrow$$

$$c = (\Sigma + n\lambda I)^{-1}(Y - X\beta)$$

$$\Sigma_{ii'} = \int\int_{\mathcal{T}\times\mathcal{T}} Z_i(s)K(s,t)Z_{i'}(t)dsdt$$

- The minimization problem becomes

$$\min_{\beta} \{ (2n)^{-1} (\boldsymbol{Y} - \boldsymbol{X}\beta)^\top P_\lambda (\boldsymbol{Y} - \boldsymbol{X}\beta) + \tau ||\beta||_0 \}$$

$$P_\lambda = n\lambda(\Sigma + n\lambda I)^{-1}$$

Estimation Algorithm

- Modify the support detection and root finding algorithm in Huang et al. (2018)

Step-1: profile out the functional part by using the Representer Theorem

Step-2: simultaneously identify the important features and obtain scalar estimates

Step-3: plug the scalar estimates into the loss function to derive the functional estimate

Algorithm 1 Functional support detection and root finding (FSDAR)

Input: An initial β^0 and the sparsity level J; set $k = 0$.

select λ^0 by minimizing the GCV criterion $\mathrm{GCV}_\lambda^0 = n\|\mathbf{P}_\lambda(\mathbf{Y} - \mathbf{X}\beta^0)\|_2^2/[\mathrm{tr}(\mathbf{P}_\lambda)]^2$ and calculate $d^0 = \mathbf{X}^T\mathbf{P}_{\lambda^0}(\mathbf{Y} - \mathbf{X}\beta^0)/n$;

for $k = 0, 1, 2, \ldots$ **do**

  $A^k = \{i : |\beta_i^k + d_i^k| \geq \|\beta^k + d^k\|_{J,\infty}\}$, $I^k = (A^k)^c$;

  $\lambda^k = \arg\min_\lambda \left\{n\|\mathbf{P}_\lambda^k(\mathbf{Y} - \mathbf{X}_{A^k}\beta_{A^k}^k)\|_2^2/[\mathrm{tr}(\mathbf{P}_\lambda^k)]^2\right\}$, $\mathbf{P}_{\lambda^k} = n\lambda^k(\mathbf{\Sigma} + n\lambda^k\mathbf{I})^{-1}$;

  $\beta_{A^k}^{k+1} = (\mathbf{X}_{A^k}^T\mathbf{P}_{\lambda^k}\mathbf{X}_{A^k})^{-1}\mathbf{X}_{A^k}^T\mathbf{P}_{\lambda^k}\mathbf{Y}$, $\beta_{I^k}^{k+1} = \mathbf{0}$;

  $d_{A^k}^{k+1} = \mathbf{0}$, $d_{I^k}^{k+1} = \mathbf{X}_{I^k}^T\mathbf{P}_{\lambda^k}(\mathbf{Y} - \mathbf{X}_{A^k}\beta_{A^k}^{k+1})/n$;

  if $A^{k+1} = A^k$ **then**

    Stop and denote $\widehat{\beta} = (\widehat{\beta}_{A^k}^T, \widehat{\beta}_{I^k}^T)^T$.

  else

    $k = k + 1$;

  end if

end for

Output: $\widehat{\beta}$, $\widehat{\mathbf{c}} = (\mathbf{\Sigma} + n\lambda^k\mathbf{I})^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\beta})$, and $\widehat{\xi} = \sum_{i=1}^n \widehat{\mathbf{c}}_i(KZ_i)$.

Huang, J., Jiao, Y., Liu, Y., & Lu, X. (2018). A constructive approach to ℓ_0 penalized regression. *The Journal of Machine Learning Research*, *19*(1), 403-439.

Theoretical Properties

Theorem 1: Under suitable conditions, as $n \to \infty$, the following inequalities hold with probability approaching one,

$$||\beta^*|_{A^*\setminus A^{k+1}}||_2 \le \gamma^{k+1}||\beta^*||_2 + \frac{\gamma}{(1-\gamma)C_,}h(J), \qquad ||\beta^{k+1} - \beta^*||_2 \le C\gamma^{k+1}||\beta^*||_2 + bh(J),$$

where β^* is the true value of the scalar coefficients, A^* is the true index set of nonzero variables, C , b are constants and $h(J) = \max_{A\subset S:|A|\le J}\left(\frac{||X_A^\top P_\lambda\langle Z,\delta^*\rangle}{n} + \frac{||X_A^\top P_\lambda \epsilon||_2}{n}\right)$

- $||\beta^*|_{A^*\setminus A^{k+1}}||_2$: estimation error of false zero elements, $||\beta^{k+1} - \beta^*||_2$: estimation error of the scalar estimators

Theorem 2: Under suitable conditions, if $K^{-1/2}\xi^* \in Ran(T^r)$ with $r \in [0, 1/2]$ and if the eigenvalues of the operator $T = K^{1/2}E\{Z(t)Z(s)\}K^{1/2}$ satisfy $s_j \asymp j^{-2\alpha}$, by choosing $\lambda \asymp n^{-2\alpha/(2\alpha+1+4\alpha r)}$, we can have

$$E^*\langle\hat{\xi} - \xi^*, Z^*\rangle^2 = O\left(n^{-\frac{2\alpha+4\alpha r}{2\alpha+1+4\alpha r}} + J^2\log(p)\,n^{-1}\right),$$

$$||\hat{\xi} - \xi^*||_{\mathcal{H}}^2 = O\left(n^{-\frac{4\alpha r}{2\alpha+1+4\alpha r}} + J^2\log(p)\,n^{-1}\right).$$

Model Setup

Outcome generating model

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

Exposure generating model

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

$\boldsymbol{B}$ is the main parameter of interest, representing the association between the 2D imaging exposure 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 $E_i$are random errors that may be correlated. The symbol "$*$" denotes element-wise multiplication.

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

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

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:

$$\mathbf{Z}_{\mathrm{i}} = \sum_{\mathrm{l}=1}^{\mathrm{s}} \mathrm{x}_{\mathrm{il}} * \mathbf{C}_l + \mathrm{E}_{\mathrm{i}}$$

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

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

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

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

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

Select submodel: $\widehat{\mathcal{M}} = \widehat{\mathcal{M}}_1^* \cup \widehat{\mathcal{M}}_2$. (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 $\parallel \mathrm{B} \parallel_* = \sum_k \sigma_k(\mathrm{B})$.

L1 penalty, exclude instrumental and irrelevant variables.

Nuclear penalty, low-rank estimation of B.

Estimated effect size of imaging exposure z,

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

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

Theoretical Properties

Theorem 3: Under suitable conditions, let $\gamma_{1,n} = \alpha D_1 n^{-\kappa}, \gamma_{2,n} = \alpha D_1 (pq)^{\frac{1}{2}} n^{-\kappa}$ with $0 < \alpha < 1$, then $P(\mathcal{M}_1 \subset \widehat{\mathcal{M}}) \to 1$ and $P(|\widehat{\mathcal{M}}| = O(n^{2\kappa+\tau})) \to 1$ as $n \to \infty$.

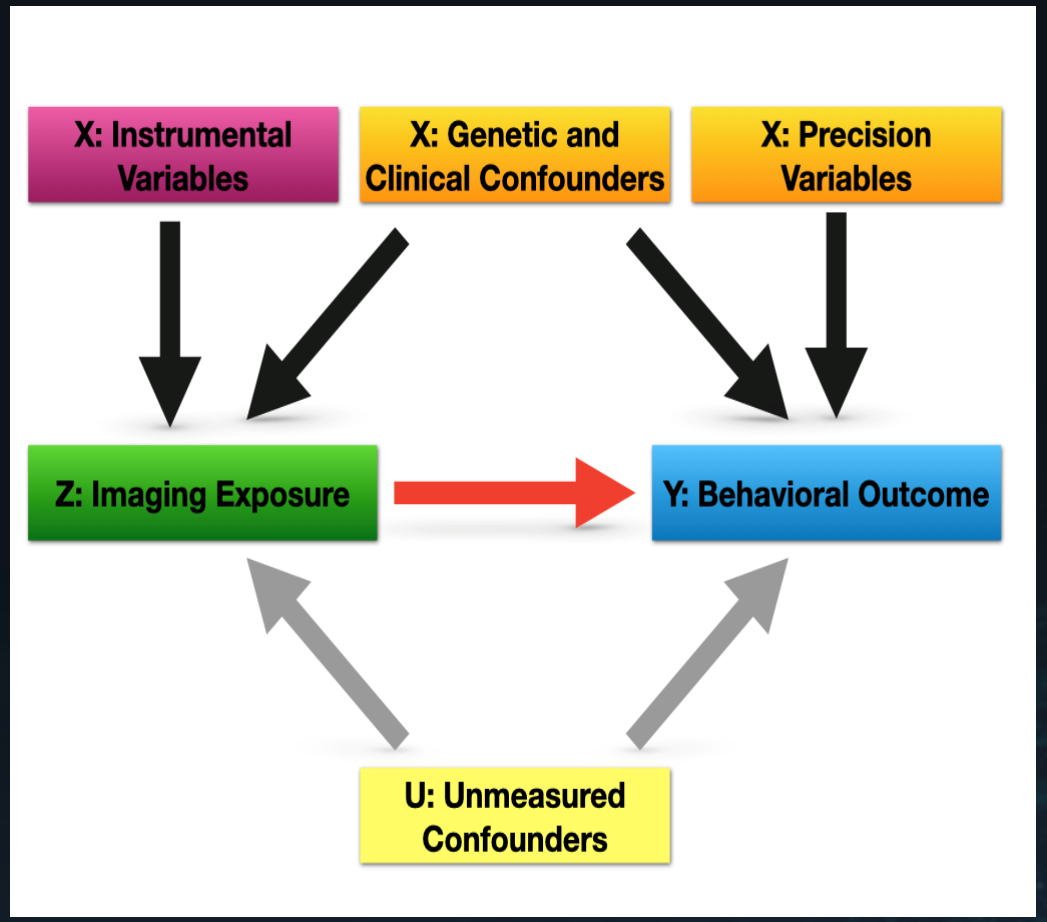
- With properly chosen $\gamma_{1,n}$ and $\gamma_{2,n}$, the joint screening set includes the confounders and precision variables with high probability
- The size of selected model from the screening is only a polynomial order of n.

Theorem 4: Let $\hat{\theta}_\lambda = \left(\hat{\beta}^\top, vec(\widehat{\mathrm{B}})^\top\right)^\top$, under suitable conditions, as $n \to \infty$,

$$||\hat{\theta}_\lambda - \theta^*|| = O_p(\max\{n^{2\kappa+\tau-1}, n^{1-2\tau})$$

- The convergence rate is controlled by κ and τ
- κ controls the exponential rate of model complexity that can diverge
- τ controls the rate of largest eigenvalue of population covariance matrix that can grow

DAG and Mandelian Randomization



Our DAG is closely related with the causal path diagram of multiple instrumental variables in the Mandelian Randomization (MR) literature.

Imaging measures can be regarded as an exposure function.

If there are no unmeasured confounders, then we ca make the causal inference on the effect of Z on Y.

Including more confounders or generalizing MR methods for functional exposure.

Unobserved Confounder

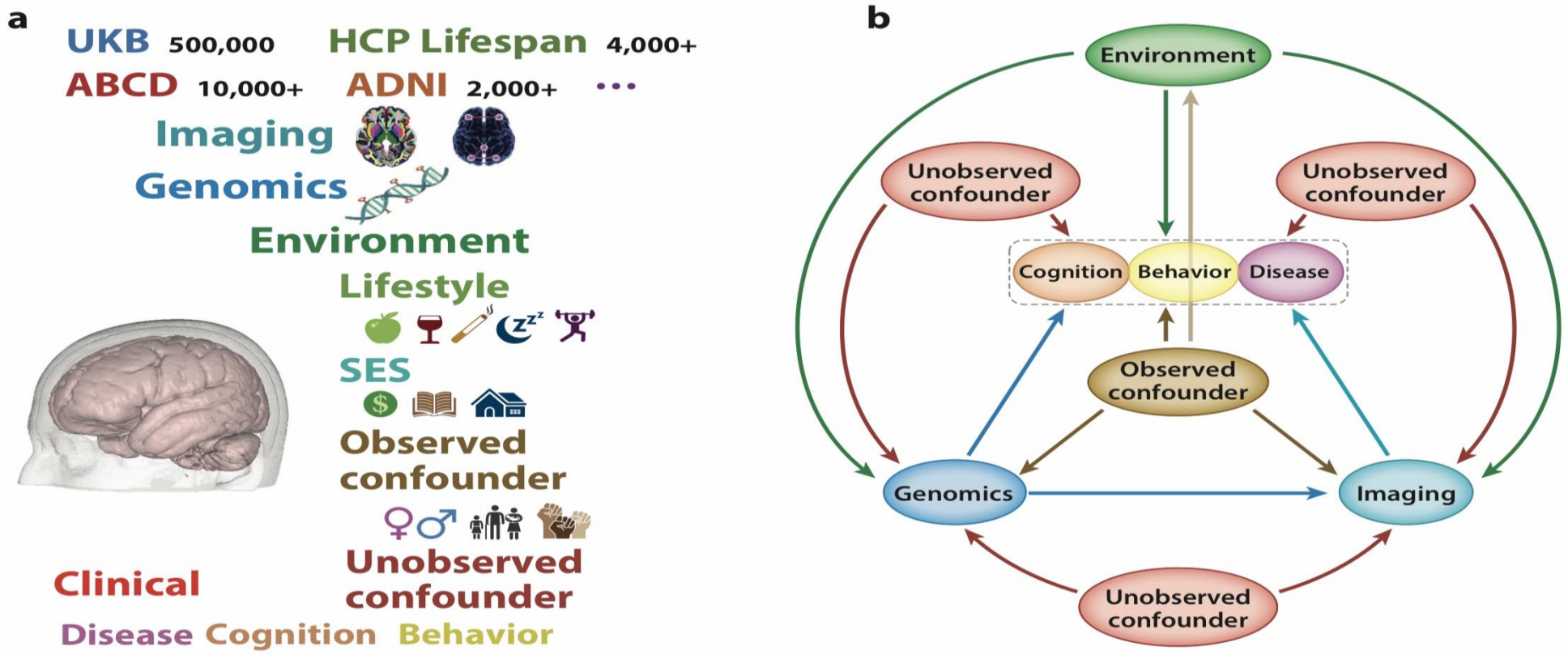


Figure 1

(*a*) Major data types from different domains in several representative large-scale biomedical studies. The number after each dataset represents the sample size. (*b*) A dynamic causal model for delineating the CGIC pathway confounded with environmental factors and unobserved confounders. An arrow from a factor X to a factor Y represents the direct effect of X on Y. Abbreviations: ABCD, Adolescent Brain Cognitive Development; ADNI, Alzheimer's Disease Neuroimaging Initiative; CGIC, causal genetic-imaging-clinical; HCP, Human Connectome Project; SES, socioeconomic status; UKB, UK Biobank.

Exclude the Effect of the Unobserved Confounders

- Consider the high-dimensional functional structure equation Models with endogeneity

$$Y_i = \alpha + \sum_{\ell=1}^{p_n} X_{i\ell}\,\beta_\ell + \int_{\mathcal{T}} Z_i\,(t)B(t)dt + \epsilon_i, \qquad (1)$$

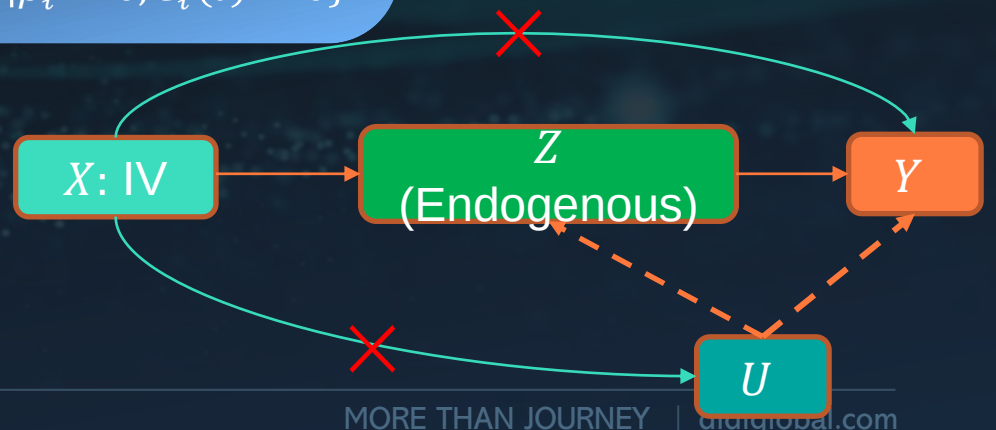
$$Z_i(t) = \sum_{\ell=1}^{p_n} X_{i\ell} C_\ell(t) + E_i(t), \qquad (2)$$

U_i

- ✓ The error process $E_i(t)$ is allowed to be correlated with the error term ϵ_i.
- ✓ The common unobserved confounders cause the correlation.
- ✓ Four types of genes: $\mathcal{C} = \{\ell \in \mathcal{A} | \beta_\ell \neq 0, C_\ell(t) \neq 0\}$, $\mathcal{P} = \{\ell \in \mathcal{A} | \beta_\ell \neq 0, C_\ell(t) = 0\}$

  $\mathcal{I} = \{\ell \in \mathcal{A} | \beta_\ell = 0, C_\ell(t) \neq 0\}$, $\mathcal{S} = \{\ell \in \mathcal{A} | \beta_\ell = 0, C_\ell(t) = 0\}$

- Challenges
  - ✓ Infinite dimensional endogenous variable with scalar instruments
  - ✓ A mixed set of instruments and control variables
  - ✓ Some invalid instruments such that $\beta_l \neq 0$
  - ✓ High-dimensional covariates



Identification Problem

- Consider one valid instrumental variable

$$Y_i = \alpha + \int_{\mathcal{T}} Z_i(t)B(t)dt + \epsilon_i, \qquad Z_i(t) = X_{i\ell}C_\ell(t) + E_i(t)$$

- Plugging $Z_i(t)$ into Y_i

$$Y_i = \alpha + X_{i\ell}\int_{\mathcal{T}} C_\ell(t)\,B(t)dt + \tilde{\epsilon}_i, \qquad (3)$$

$\xi(t)$ is not identifiable if the space $\mathcal{N} = \left\{\xi: \int_{\mathcal{T}} C_\ell(t)\,B(t)dt = 0\right\} \neq \{0\}$.

✓ Existing works use functional instruments

- Using the fact

$$E\left[X_i\left(Y_i - \int_{\mathcal{T}} Z_i(t)B(t)dt\right)\right] = 0, \qquad E\left[X_i\left(Z_i(t) - X_i^\top C(t)\right)\right] = 0$$

Identify unique leading coefficients $\{b_k\}_{k=1}^K$ (p equations with $p + K$ parameters)

$$E\left(X_iX_i^\top\right)^{-1}E(X_iY_i) = \Gamma^* = \beta + \int_{\mathcal{T}} E\left(X_iX_i^\top\right)^{-1}E\left(X_iZ_i(t)\right)B(t)dt = \beta + \int_{\mathcal{T}} C(t)B(t)dt \approx \beta + \sum_{k=1}^K b_kc_k$$

Corollary: Suppose that $\int_{\mathcal{T}} C(t)B(t)dt$ can be approximated by $\sum_{k=1}^K b_kc_k$ with c_k being a vector, and for any K of the relevant instruments identifies a unique $\{b_k\}_{k=1}^K$. If the number of invalid instruments is less than $(p - K + 1)/2$, there is a unique solution to the above equation.

✓ If $K = 1$, it reduces to the majority rule.

Simulation Studies of FLSEM

Table 1: Monte Carlo averages with standard errors in parentheses for $n = 400, p = 20$ for two-dimensional functional exposure

ρ_1	ρ_2		FZ_Z	FN_Z	FZ_Y	FN_Y	MSE_B	MSE_β
0.3	0	FLSEM	0.000(0.000)	7.120(1.996)	0.000(0.000)	0.080(0.274)	0.049(0.014)	0.027(0.019)
		PFLM	-	-	0.000(0.000)	2.300(1.216)	0.053(0.021)	0.063(0.036)
	0.2	FLSEM	0.000(0.000)	6.960(2.194)	0.000(0.000)	0.040(0.198)	0.051(0.016)	0.033(0.016)
		PFLM	-	-	0.000(0.000)	3.380(0.830)	0.235(0.074)	0.579(0.238)
	0.5	FLSEM	0.000(0.000)	7.380(2.108)	0.000(0.000)	0.120(0.385)	0.047(0.015)	0.037(0.026)
		PFLM	-	-	0.000(0.000)	3.960(0.198)	0.644(0.123)	3.296(0.524)
	0.7	FLSEM	0.000(0.000)	7.260(2.068)	0.000(0.000)	0.020(0.141)	0.048(0.013)	0.034(0.024)
		PFLM	-	-	0.000(0.000)	4.000(0.000)	0.945(0.015)	7.226(0.169)
0.5	0	FLSEM	0.000(0.000)	6.280(2.176)	0.000(0.000)	0.080(0.274)	0.046(0.012)	0.032(0.022)
		Plugged-In	-	-	0.000(0.000)	1.360(1.241)	0.065(0.101)	0.180(0.862)
		PFLM	-	-	0.000(0.000)	2.600(1.355)	0.059(0.028)	0.084(0.075)
	0.2	FLSEM	0.000(0.000)	7.220(2.122)	0.000(0.000)	0.060(0.314)	0.047(0.013)	0.034(0.031)
		PFLM	-	-	0.000(0.000)	3.560(0.733)	0.206(0.097)	0.552(0.288)
	0.5	FLSEM	0.000(0.000)	6.400(2.356)	0.000(0.000)	0.080(0.274)	0.042(0.010)	0.039(0.027)
		PFLM	-	-	0.000(0.000)	3.960(0.198)	0.616(0.089)	3.503(0.522)
	0.7	FLSEM	0.000(0.000)	7.180(2.077)	0.000(0.000)	0.080(0.274)	0.047(0.014)	0.037(0.026)
		PFLM	-	-	0.000(0.000)	4.000(0.000)	0.944(0.014)	7.217(0.172)

ρ_1: control the correlation within the scalar variables

ρ_2: control the correlation between error terms

- **Estimation Procedure**
  - ✓ Estimate the function-on-scalar model (2) under RKHS with L_0 penalty
  - ✓ Obtain the fitted value of $Z_i(t)$, $\hat{Z}_i(t)$ is not correlated to ϵ_i
  - ✓ Estimate the linear model with L_0 penalty after projection of $\hat{Z}_i(t)$
  - ✓ Plug $\hat{Z}_i(t)$ into Model (1) and estimate the PFLM using the selected variable

- FZ: number of false zero scalar predictors
- FN: number of false nonzero scalar predictors
- MSE_β: scalar mean squared error
- FLSEM: functional linear structure equation model
- PFLM: the partial functional linear model that ignores endogeneity

Statistics Up AI Alliance

https://statsupai.org

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 MH116527, NIA-AG082938-01, U01AG079847, and R01AR082684.

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.