Object Oriented Data Analysis in Large-scale Medical Studies

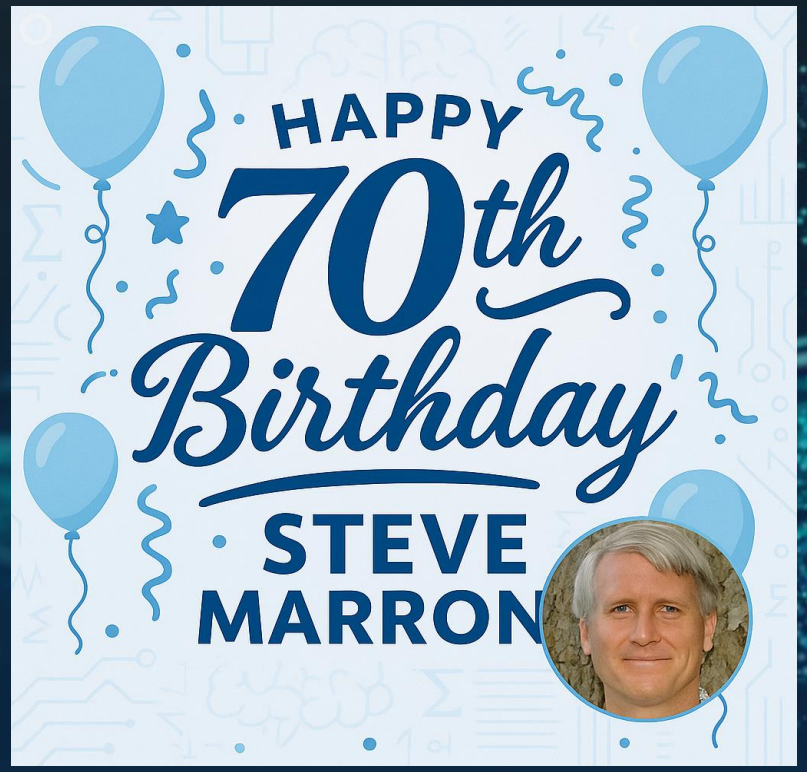
University of North Carolina at Chapel Hill

Hongtu Zhu

Joint works with Zhengwu Zhang, Hai Shu, Steve Marron, Di Xiong, Xueqing Wang, Wenliang Pan, Anuj Srivastava, and Joseph Ibrahim

https://www.med.unc.edu/bigs2/

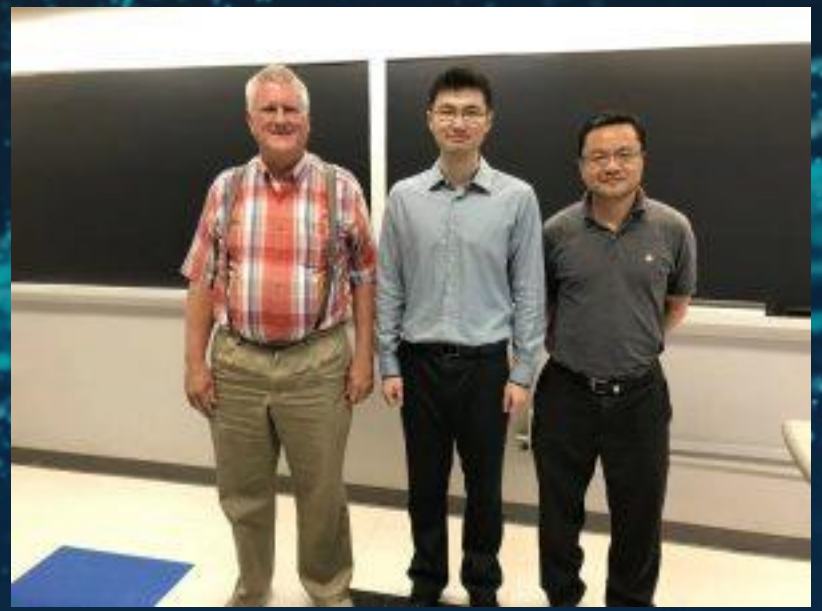
Happy 70th Birthday, Steve Marron! Wishing you a wonderful celebration and many more joyful years ahead!



Steve Marron has been an outstanding mentor and collaborator to me!

Ying Yuan	2011
Yang Yu	2017
Liuqing Yang	2019

5 joint papers (2 AOAS, 1 JRSSB, 2 Sinica)+



CONTENTS

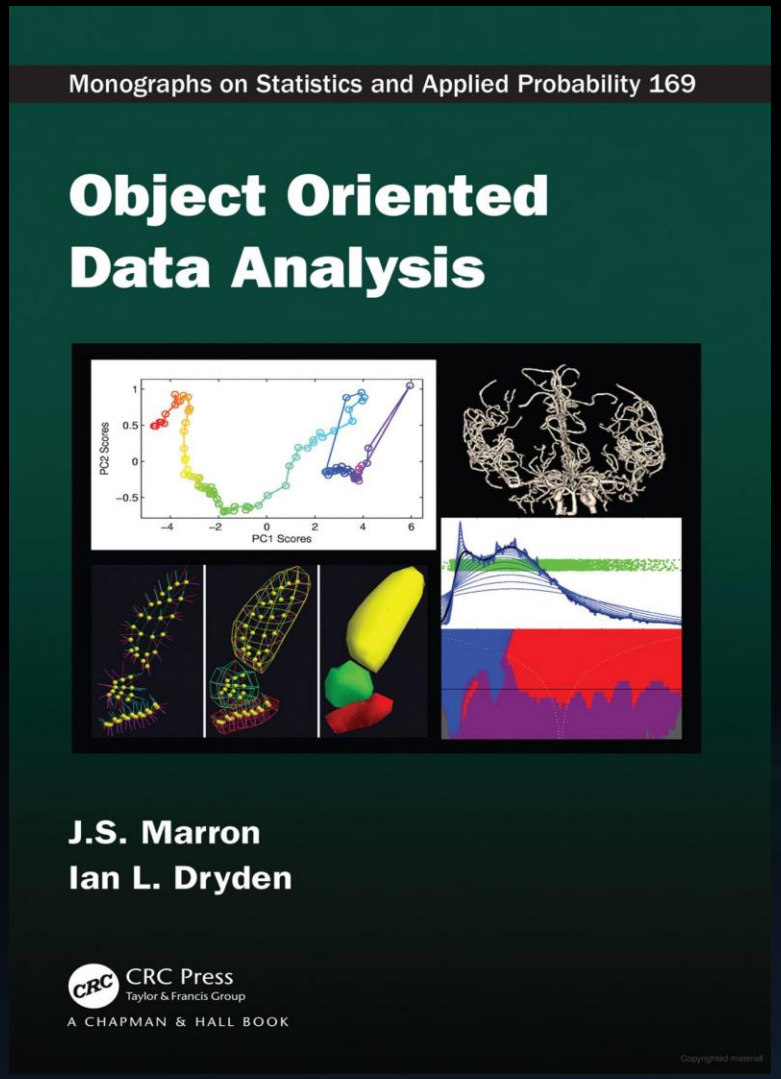
Part I

Object Oriented 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 -

Object Oriented Data Analysis



Marron's Teaching Material

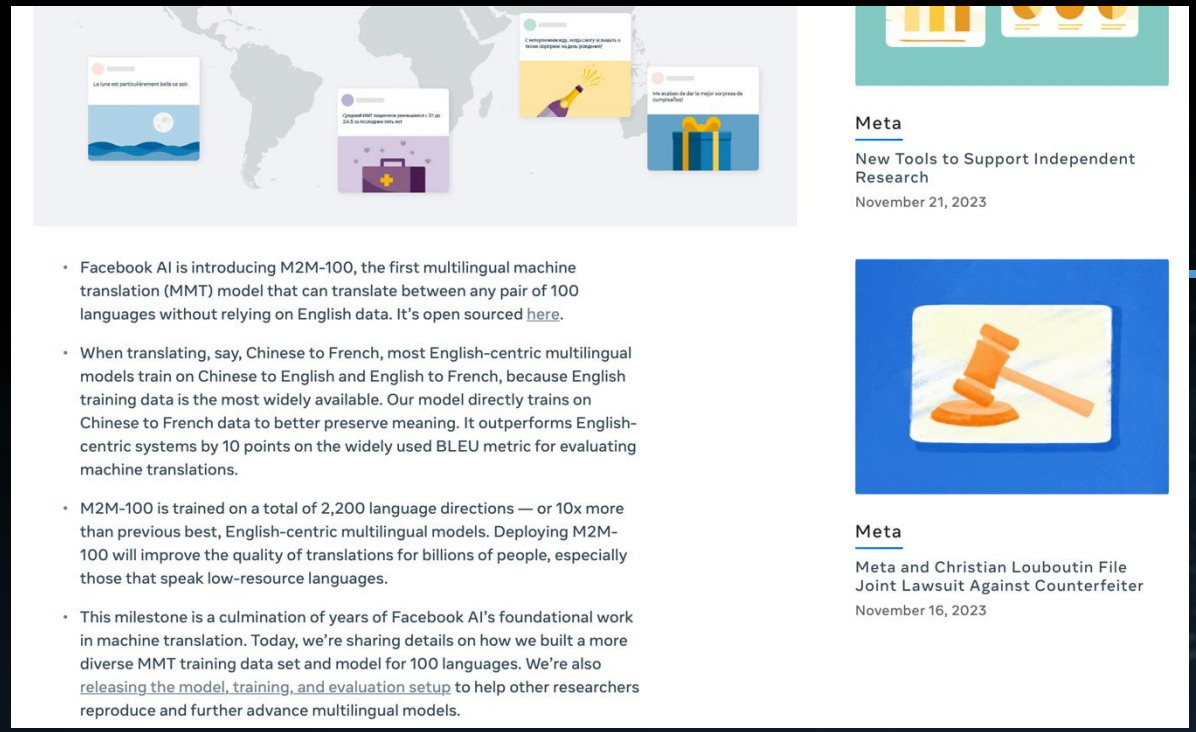
OODA:

- STOR 881 – Spring 2024
- STOR 881 – Spring 2022
- STOR 881 – Fall 2019
- STOR 881 – Fall 2017
- STOR 893 – Spring 2016
- STOR 892 – Fall 2014
- STOR 891 – Fall 2012
- STOR 891 – Fall 2007
- STAT 322 – Fall 2005
- Networks – Cornell – Fall 2001
- FDA – Cornell – Spring 2002
- FDA – Spring 2001

Objects are everywhere

Language Translation in Natural Language Processing

Deep learning enhances real-time, accurate translation of languages, as seen in tools like Google Translate. The following picture shows the translation of a webpage from English to Chinese.



- Facebook AI is introducing M2M-100, the first multilingual machine translation (MMT) model that can translate between any pair of 100 languages without relying on English data. It's open sourced here.
- When translating, say, Chinese to French, most English-centric multilingual models train on Chinese to English and English to French, because English training data is the most widely available. Our model directly trains on Chinese to French data to better preserve meaning. It outperforms English-centric systems by 10 points on the widely used BLEU metric for evaluating machine translations.
- M2M-100 is trained on a total of 2,200 language directions — or 10x more than previous best, English-centric multilingual models. Deploying M2M-100 will improve the quality of translations for billions of people, especially those that speak low-resource languages.
- This milestone is a culmination of years of Facebook AI's foundational work in machine translation. Today, we're sharing details on how we built a more diverse MMT training data set and model for 100 languages. We're also releasing the model, training, and evaluation setup to help other researchers reproduce and further advance multilingual models.

Meta

New Tools to Support Independent Research

November 21, 2023

Meta

Meta and Christian Louboutin File Joint Lawsuit Against Counterfeiter

November 16, 2023

- Facebook AI 正在推出 M2M-100，这是第一个多语言机器翻译 (MMT) 模型，可以在 100 种语言中的任意对之间进行翻译，而无需依赖英语数据。这里是开源的。
- 例如，在将中文翻译成法语时，大多数以英语为中心的多语言模型都会在中文到英语和英语到法语上进行训练，因为英语训练数据是最广泛可用的。我们的模型直接对中文到法语的数据进行训练，以更好地保留含义。在广泛使用的用于评估机器翻译的 BLEU 指标上，它比以英语为中心的系统高出 10 个点。
- M2M-100 接受了总共 2,200 种语言方向的训练，比以前最好的、以英语为中心的多语言模型多了 10 倍。部署 M2M-100 将为数十亿人提高翻译质量，尤其是那些使用资源匮乏语言的人。
- 这一里程碑是 Facebook AI 多年来在机器翻译领域基础工作的结晶。今天，我们将分享如何为 100 种语言构建更加多样化的 MMT 训练数据集和模型的详细信息。我们还发布了模型、训练和评估设置，以帮助其他研究人员重现和进一步推进多语言模型。

元

支持独立研究的新工具

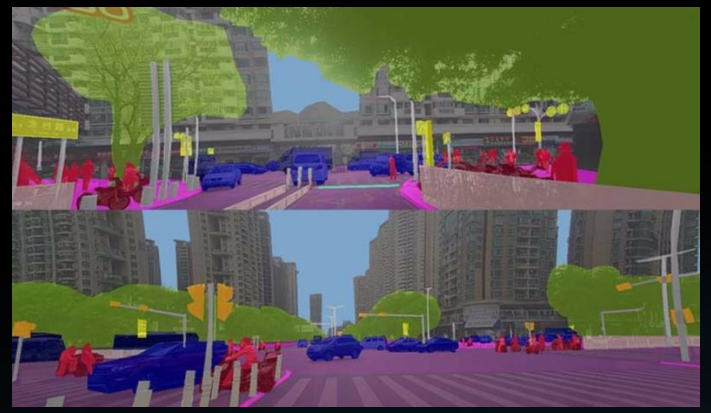
2023 年 11 月 21 日

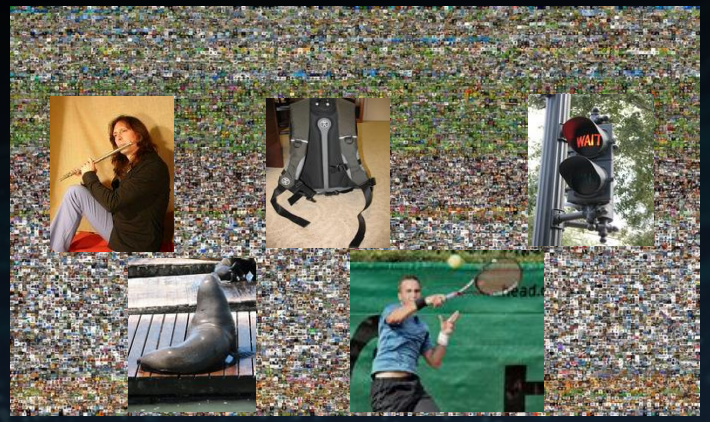
元

Meta 和 Christian Louboutin 对仿冒者提

Objects are everywhere

Disease Detection in Healthcare and Medicine





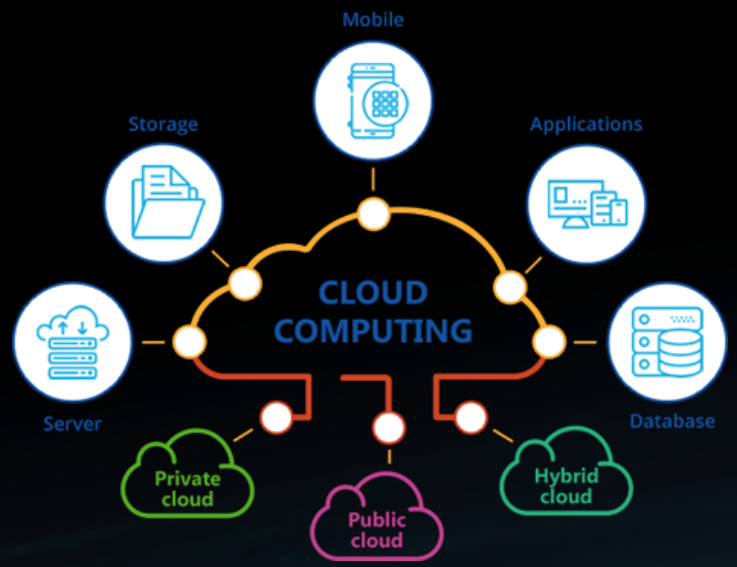
OODA/ Deep Learning Explosion

Downloaded from the NSF website and the medium.com

Deep Applications and Math/Stat

Big Data

http://medium.com



Computing

Analytical Tools

Applied Mathematics

Statistics

Machine Learning

Engineering

Part II

Case Studies for OODA

“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 -

Steve Marron's Contributions to SAMSI and My research

samsi
NSF Duke NCSU UNC

Statistical and Applied Mathematical Sciences Institute (SAMSI)

SAMSI summer workshop on Neuroimaging Data Analysis (NDA) 2013
Program Leader for SAMSI full-year program on Challenges in Computational Neuroscience (CCNS) with five workshops, one short course, and two regular courses 2015-2016

Steve Marron's Contributions to SAMSI and My research

My three SAMSI postdoctoral fellows

Zhengwu Zhang (UNC)

Yize Zhao (Yale)

Benjamin Risk (Emory)

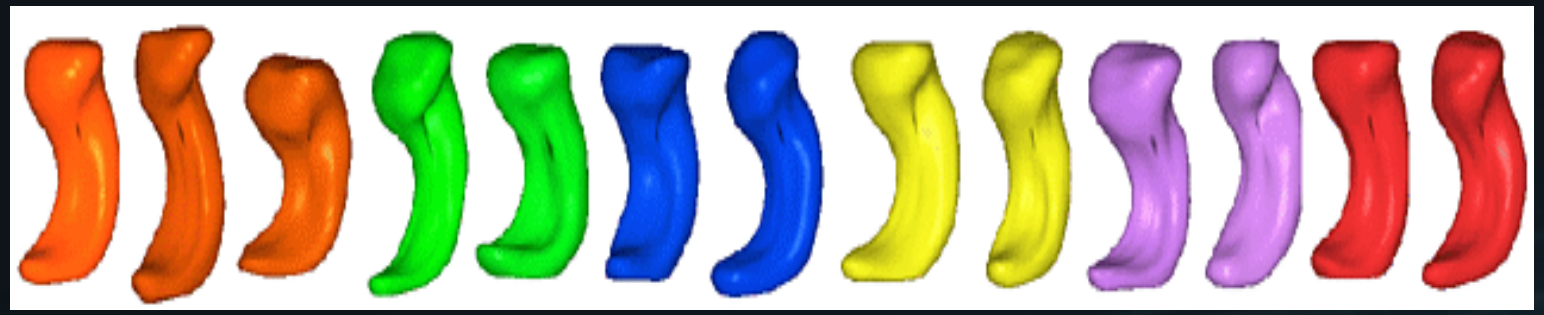
Case Study I

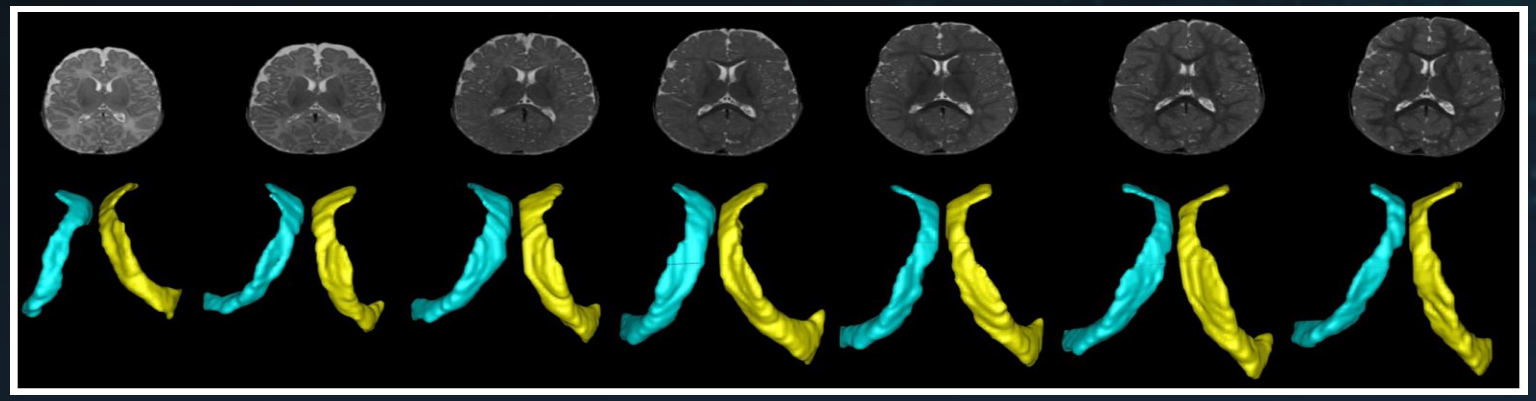
Regression Models for Manifold-value Data

1. **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.
2. **Shi, XY**, ***Zhu, HT***, Ibrahim, J.G., F. Liang, Styner M. Intrinsic regression models for median representation of subcortical structures. *Journal of American Statistical Association,* 107, 12-23, 2012.
3. **Yuan, Y., *Zhu, H.T.***, Lin, W. L., and Marron, J. S. Local polynomial regression for symmetric positive definitive matrices. *JRSS, Series B*, 74, 697-719, 2012.

Intrinsic Regression Models

Semiparametric and Nonparametric Regression for Manifold-valued Response from Cross-sectional, Longitudinal and Family-based Neuroimaging Studies





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

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

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

Intrinsic Regression Models

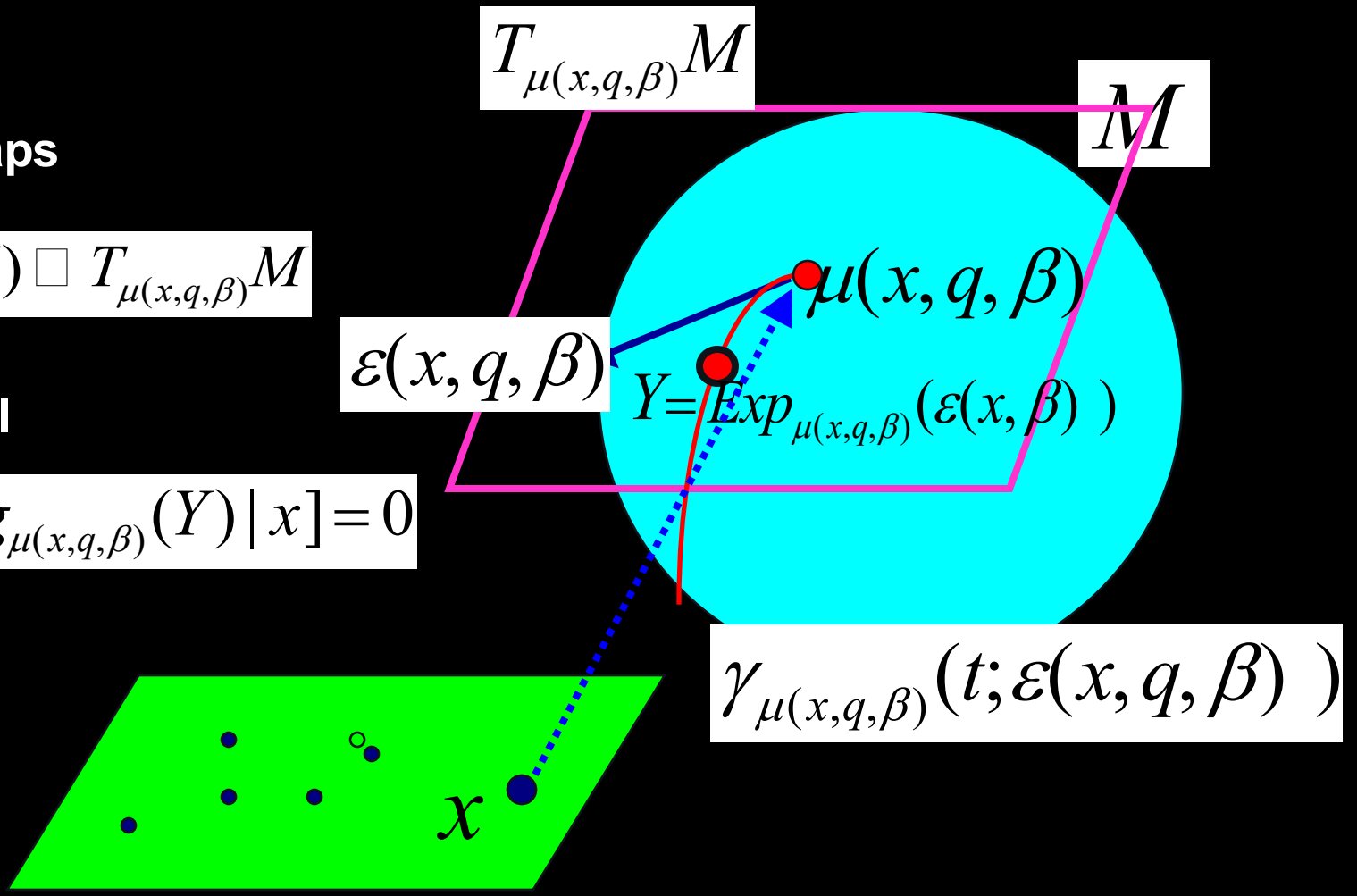
Conditional Mean

Riemannian logarithm maps

$$\varepsilon(x,q,\beta) = \mathrm{Log}_{\mu(x,q,\beta)}(Y) \in T_{\mu(x,q,\beta)}M$$

Conditional Moment Model

$$E[\varepsilon(x,q,\beta)\,|\,x] = E[Log_{\mu(x,q,\beta)}(Y)\,|\,x] = 0$$



Intrinsic Regression Models

Rotated Residuals at the common tangent space

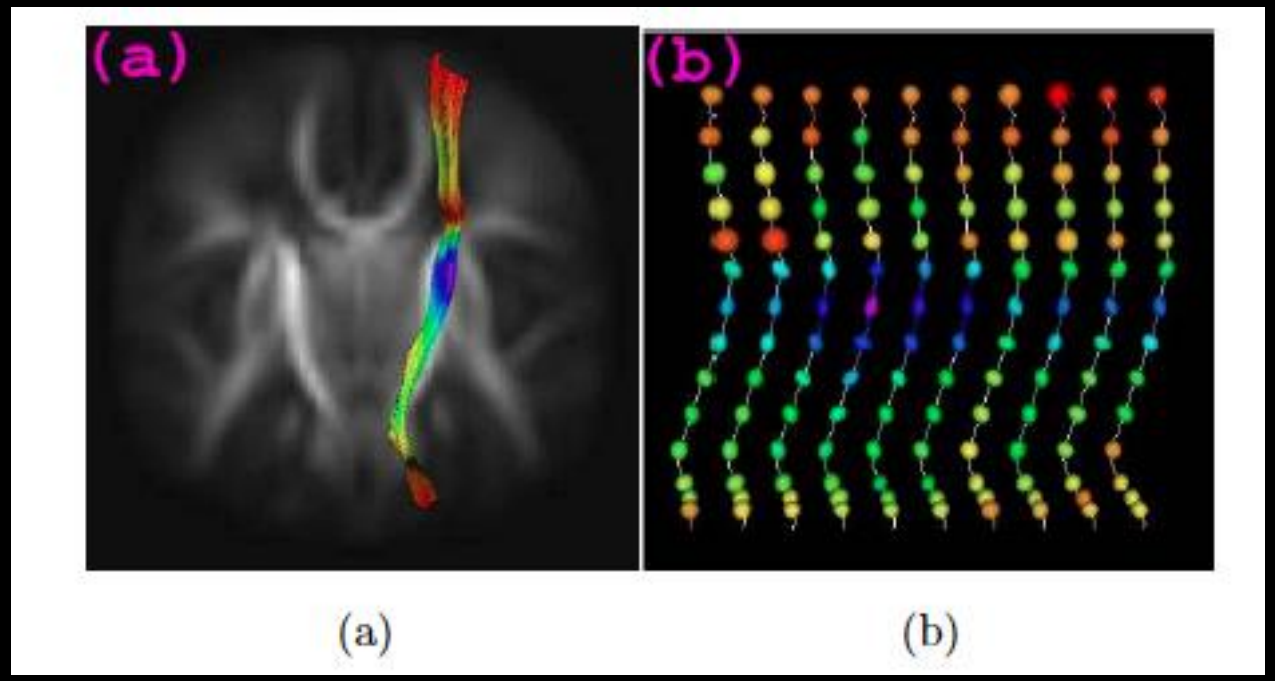
$$\mathcal{E}(y_i, \mathrm{x}_i; \mathrm{q}, \beta) = \mathcal{E}_i(\mathrm{q}, \beta) := \mathrm{Log}_{\mathrm{p}} \left(c(1; \mathrm{x}_i, \mathrm{q}, \beta)^{-1} \cdot y_i \right) \in T_{\mathrm{p}} M \quad \text{for} \quad i = 1, \ldots, n,$$

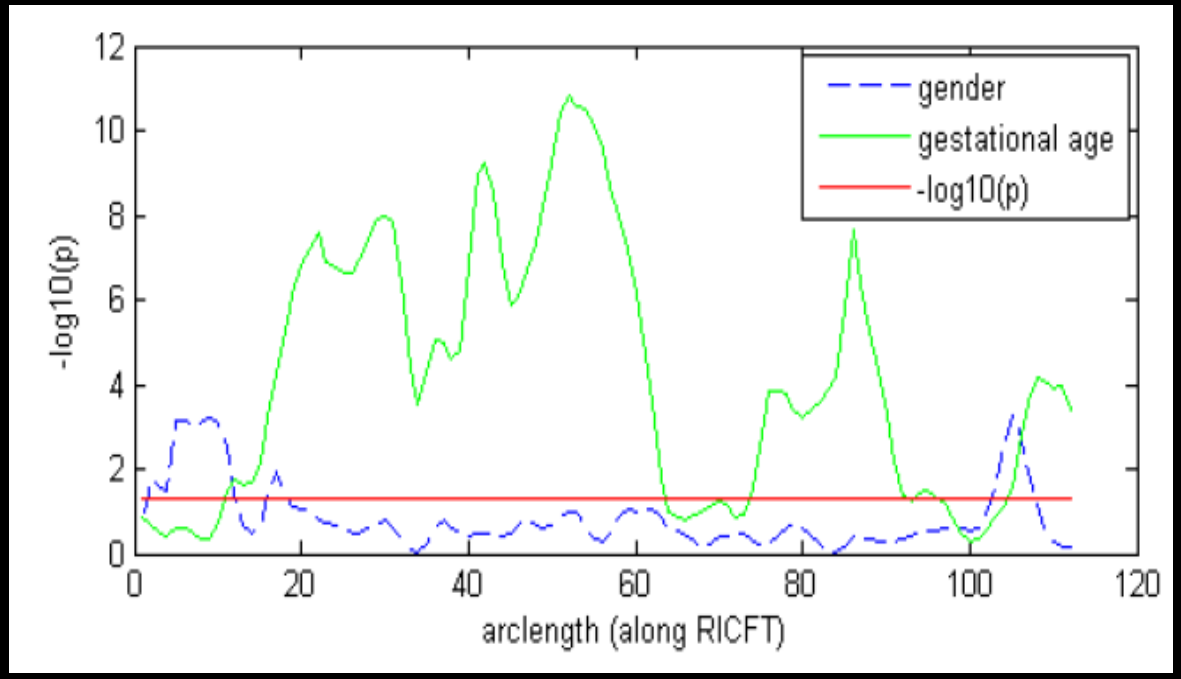
Geodesic: $\gamma(t; q, \beta) = c(t; q, \beta) \square p$

connecting base point: $p \square M$ and $\mu(x; q, \beta)$

$$L^{-1}_{c(1;\mathbf{x}_i,\mathrm{q},\boldsymbol{\beta})*} (\mathrm{Log}_{\mu(\mathbf{x}_i,\mathrm{q},\boldsymbol{\beta})}(y_i)) = L_{c(1;\mathbf{x}_i,\mathrm{q},\boldsymbol{\beta})^{-1}*} (\mathrm{Log}_{\mu(\mathbf{x}_i,\mathrm{q},\boldsymbol{\beta})}(y_i))$$
$$= \mathrm{Log}_{\mathrm{p}} \left(L_{c(1;\mathbf{x}_i,\mathrm{q},\boldsymbol{\beta})^{-1}}(y_i) \right) = \mathrm{Log}_{\mathrm{p}} \left(c(1; \mathrm{x}_i, \mathrm{q}, \beta)^{-1} \cdot y_i \right) \in T_{\mathrm{p}} M.$$

Real Data Analysis





Right internal capsule

48=18(m)+30(f)

Case Study II

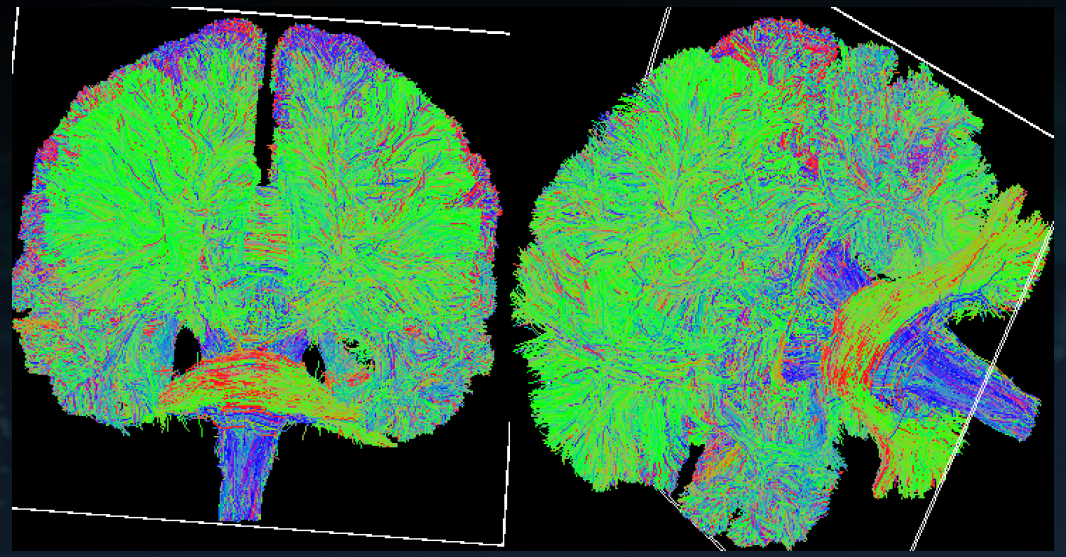
PSC: Population Structural Connectome Analysis

1. **Z. Zhang**, M. Descoteaux, **J. Zhang**, G. Girard, M. Chamberland, D. Dunson, A. Srivastava, and ***H. Zhu***. (2018). Mapping Population based Structural Connectomes. *NeuroImage*, 172， 130-145.
2. Wang, X. F. (2021). *Statistical Learning Methods for Diffusion Magnetic Resonance Imaging. Ph.D. Dissertation.*

Brain Imaging Genetics Paradigm

- **Extract the connectome using dMRI and T1 image**
- **Use tractography algorithm proposed in** *Girard et al. 2015*
  - **Step1. Construct HARDI**
  - **Step 2. Fiber tracking (incorporate anatomical info):**
  - **Step 3. Final output**:

1. **More than a million streamlines**
2. **Each streamline has hundreds 3D points**
3. **Each subject takes > 3 GBs**



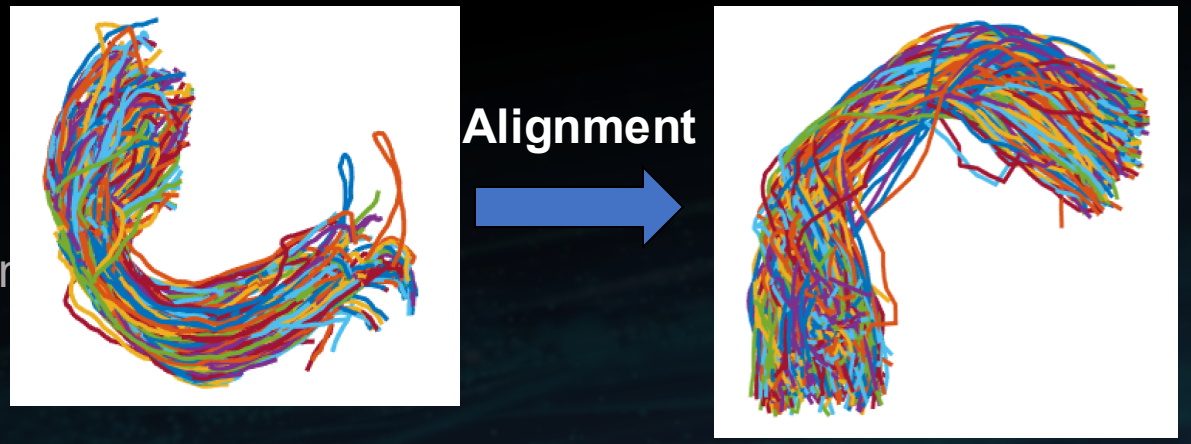
Efficient Representation of Streamlines

- **Represent the streamlines through basis and coefficients**
- **Basis can be learnt from data to increase its representing power**

Step 1. Generate atlas for streamlines connecting each pair of regions

Step 2. Alignment using the **Elastic Shapes Analysis** framework (Srivastava et al. 2012)

- rotation
- translation
- scaling
- re-parameterization

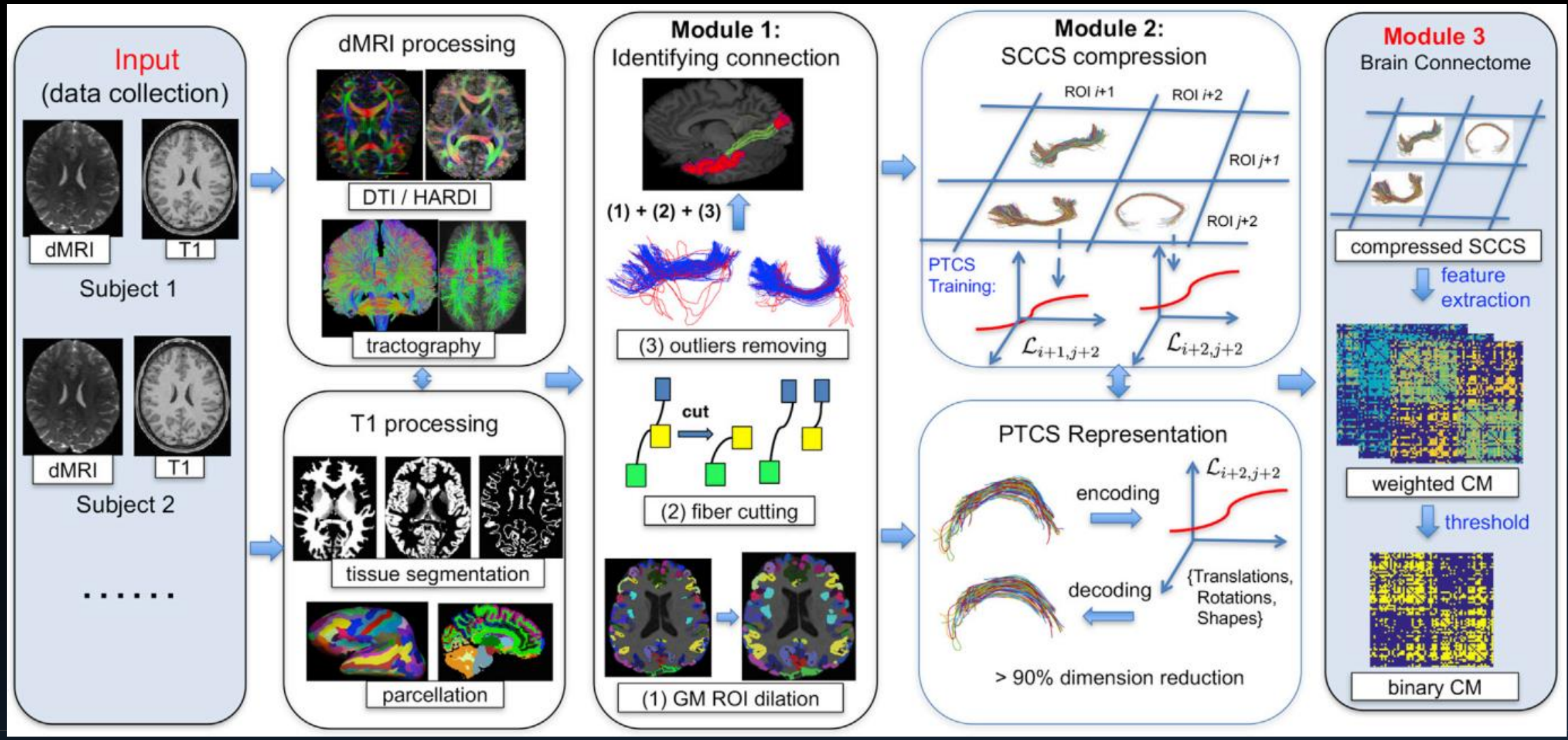


K-means clustering may be used if these streamlines have different shapes

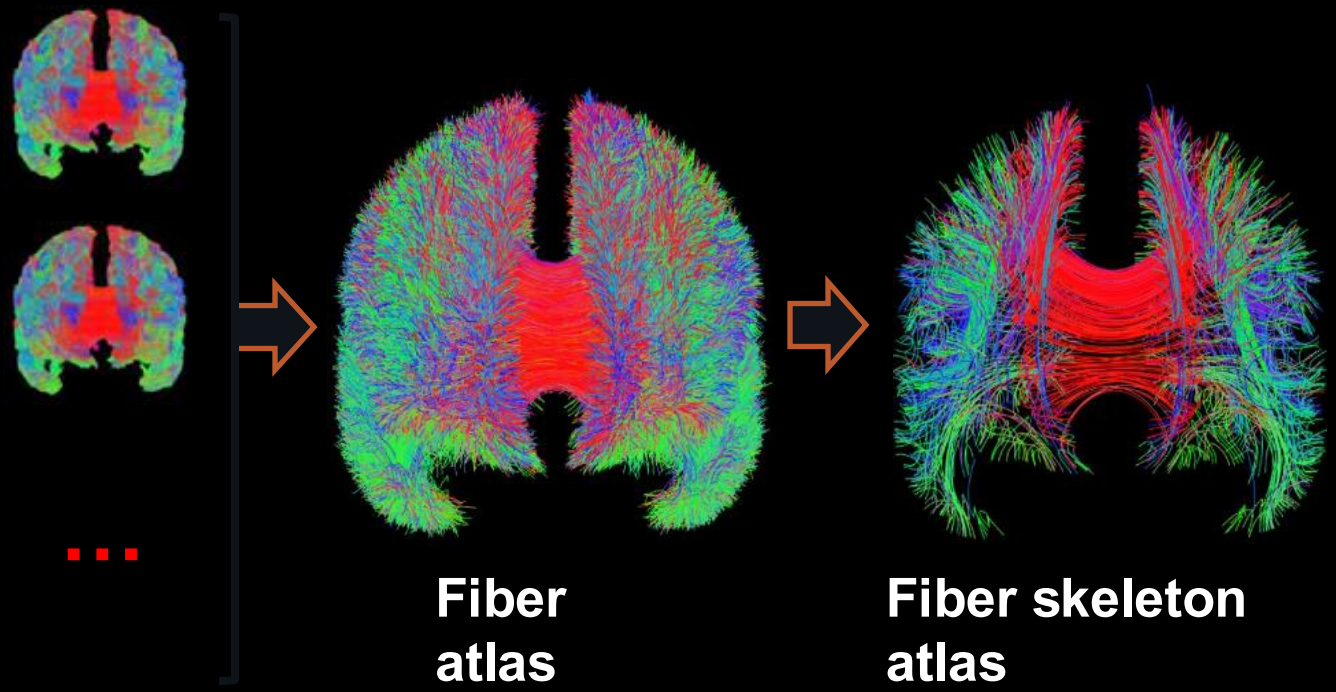
Square-root velocity function (SRVF) and Fisher-Rao metric

Srivastava, A. and Klassen, E. P. (2017) Functional and Shape Data Analysis. Springer.

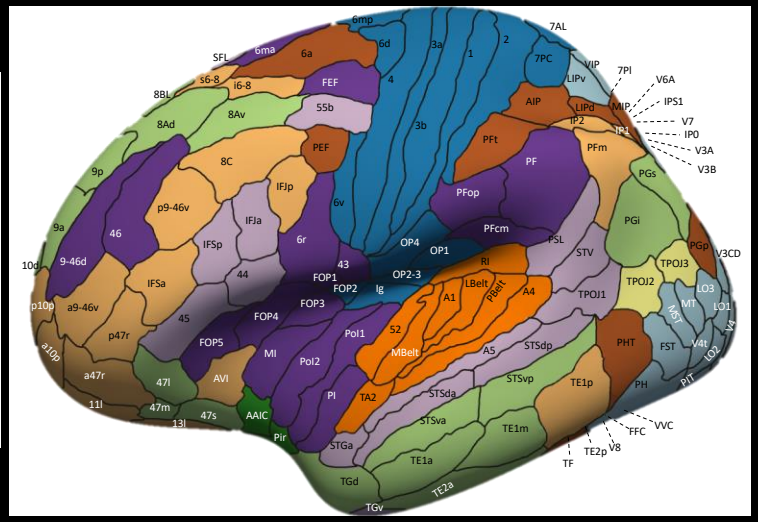
Population Structural Connectome



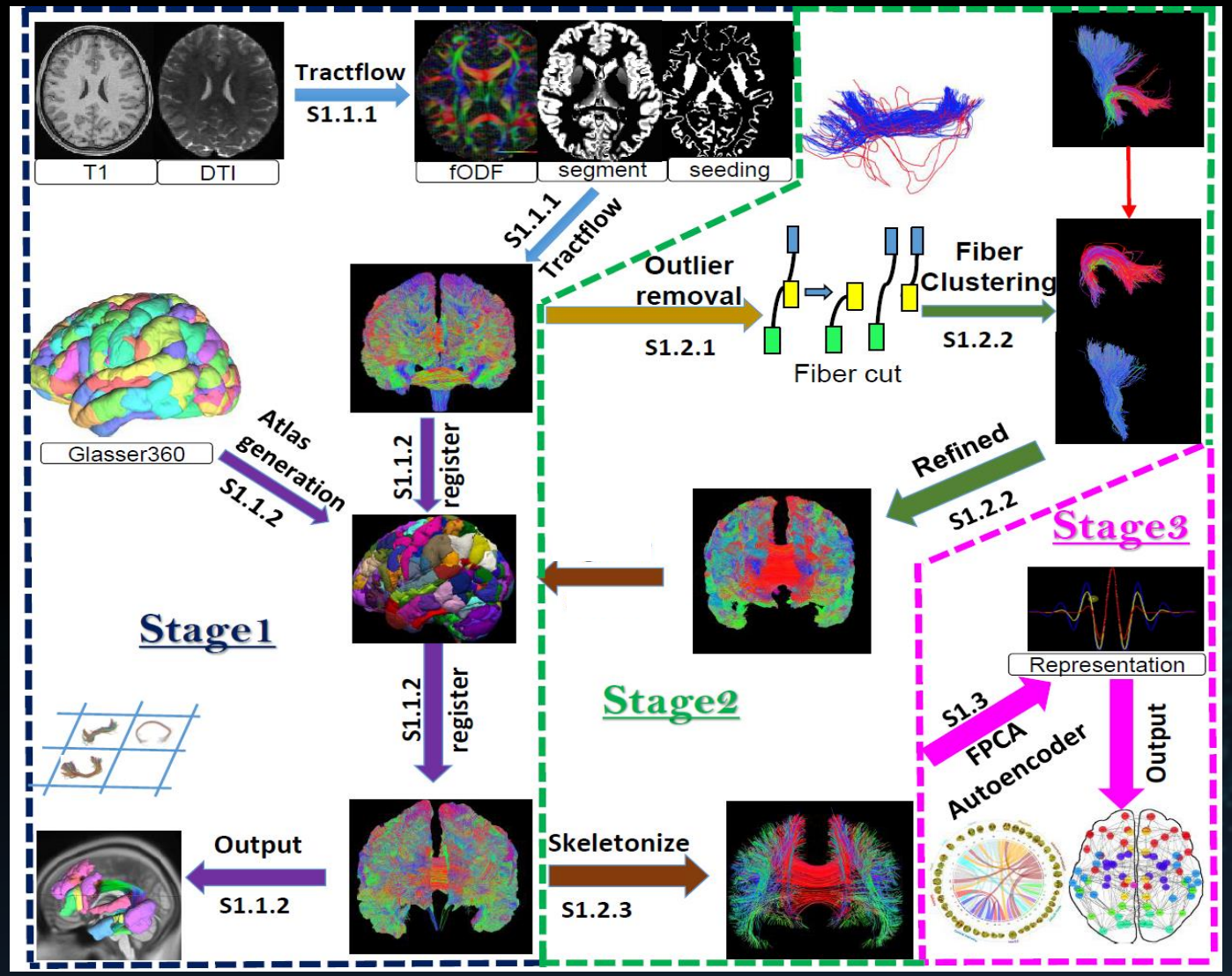
Brain Function-based Structural Connectome Atlas



(B)



Brain Function-based Structural Connectome Atlas



Stage 1:

Whole-brain Structural Connectome

Stage 2:

Creation of Fiber Skeleton

Stage 3:

Sparse Representation

Case Study III

LESA: Longitudinal Elastic Shape Analysis

Zhang, Z. W., Yuexuan Wu, Di Xiong, Joseph G. Ibrahim, Anuj Srivastava, and Hongtu Zhu. LESA: Longitudinal elastic shape analysis of brain subcortical structures (with discussion). *Journal of the American Statistical Association, Application and Case Studies*.

Shape Changes of Brain Subcortical Structures

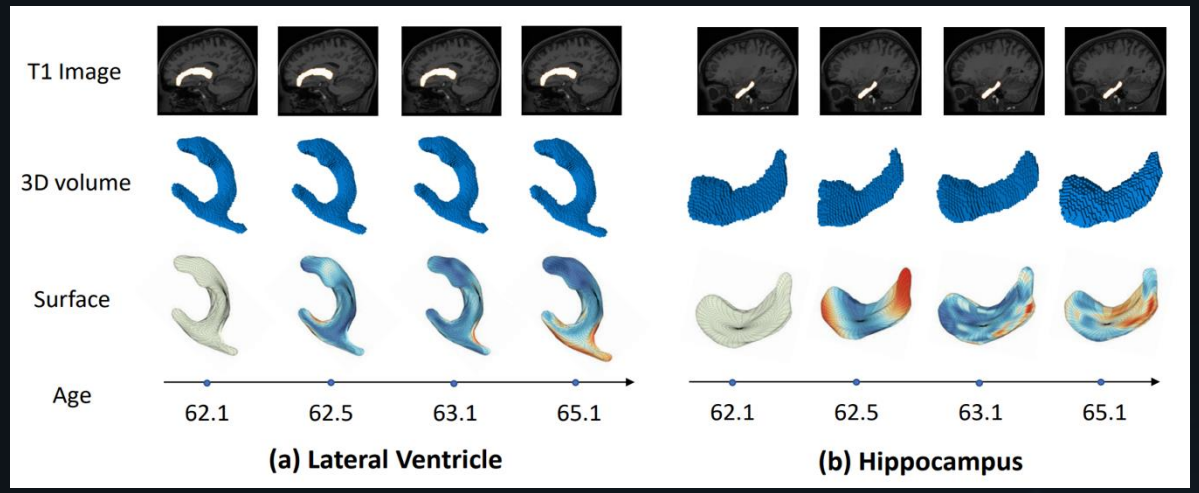


Figure 1: An example of three different representations of lateral ventricle and hippocampus across four time points for a randomly selected subject.

(Q1) How to measure developmental changes in the shape of subcortical regions?

(Q2) How to quantify the effect of disease or other covariates on subcortical shape changes?

Longitudinal Elastic Shape Analysis

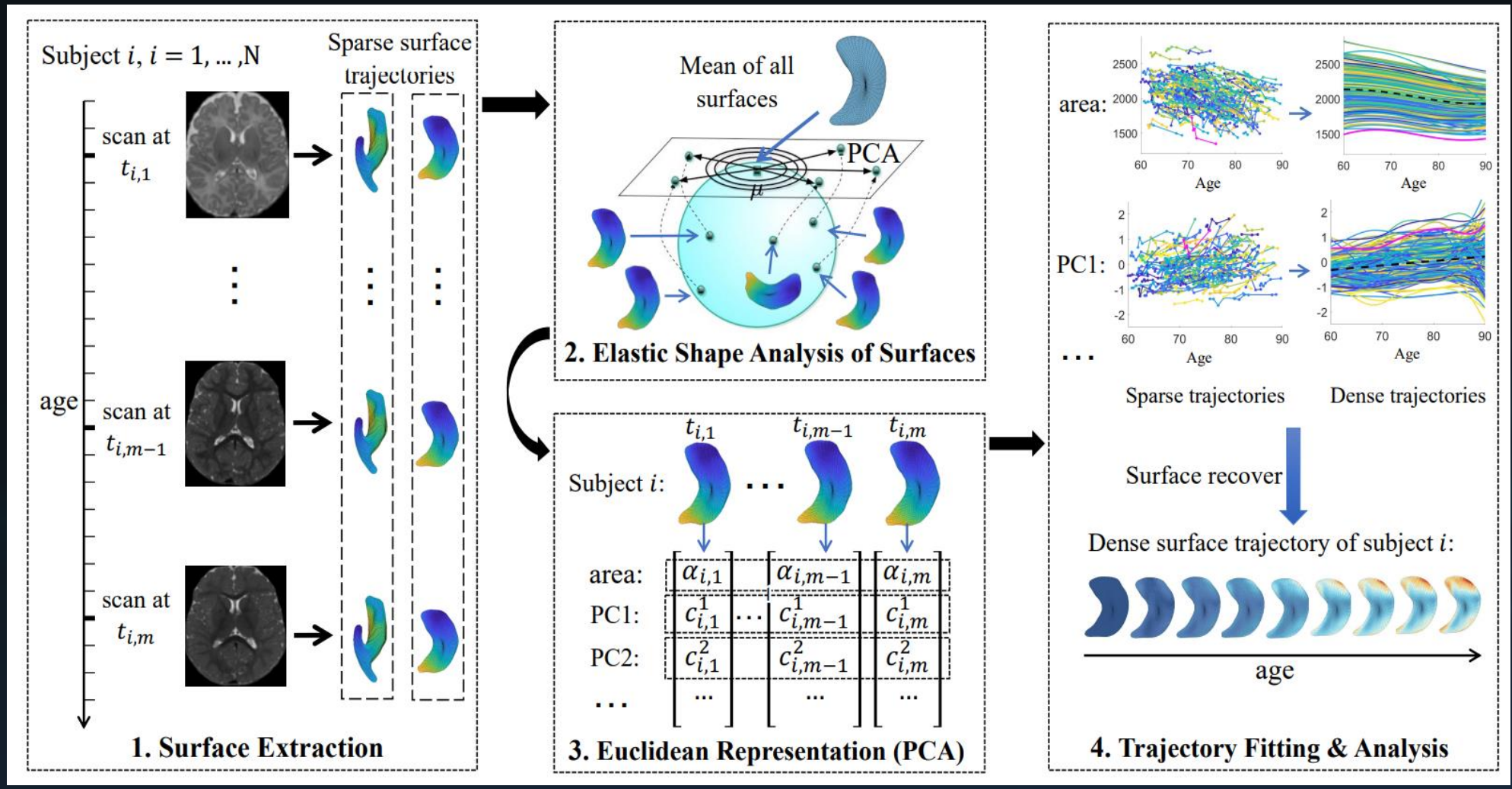
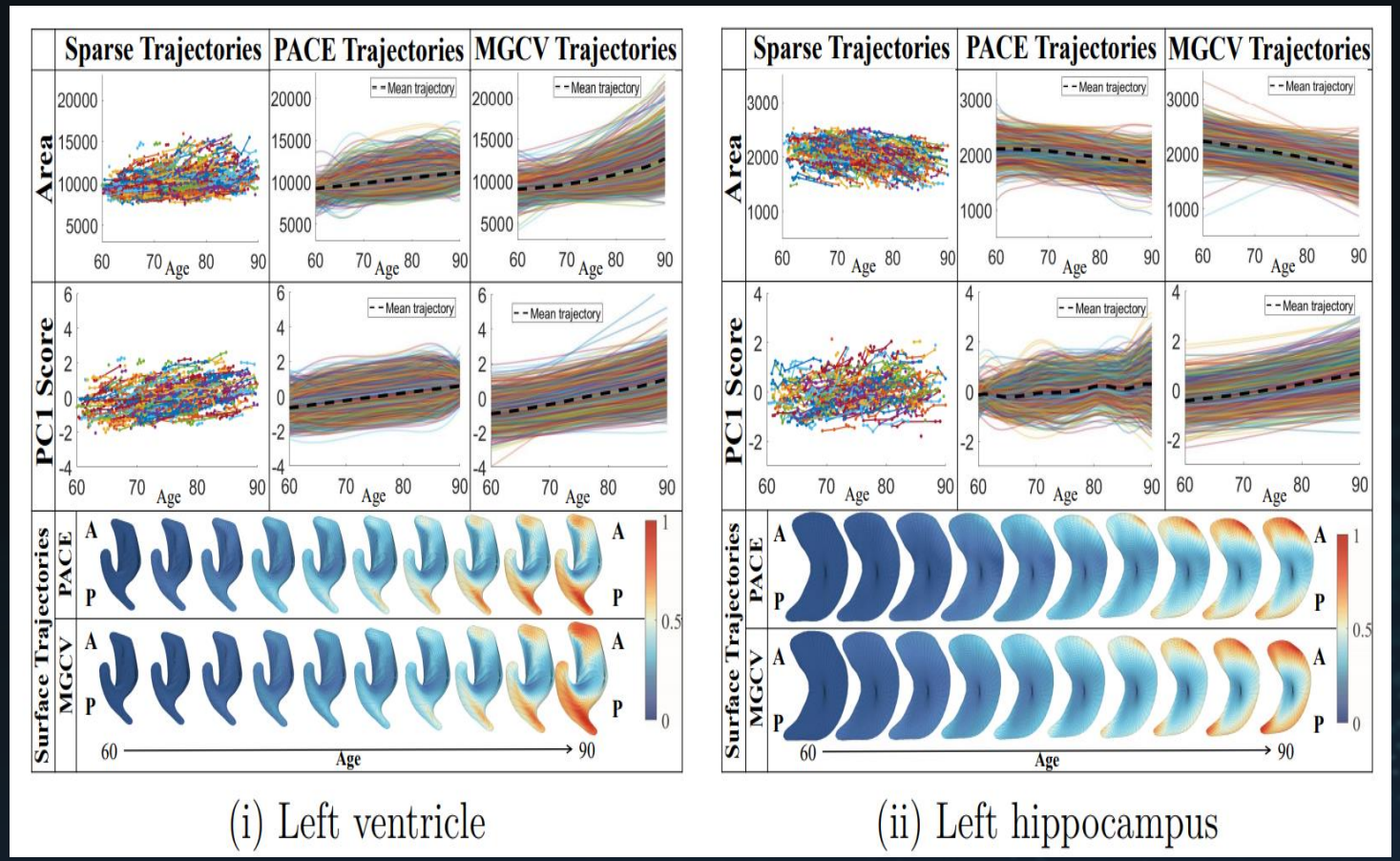


Figure 2: **A schematic overview of LESA**.

Longitudinal Shape Data Analysis



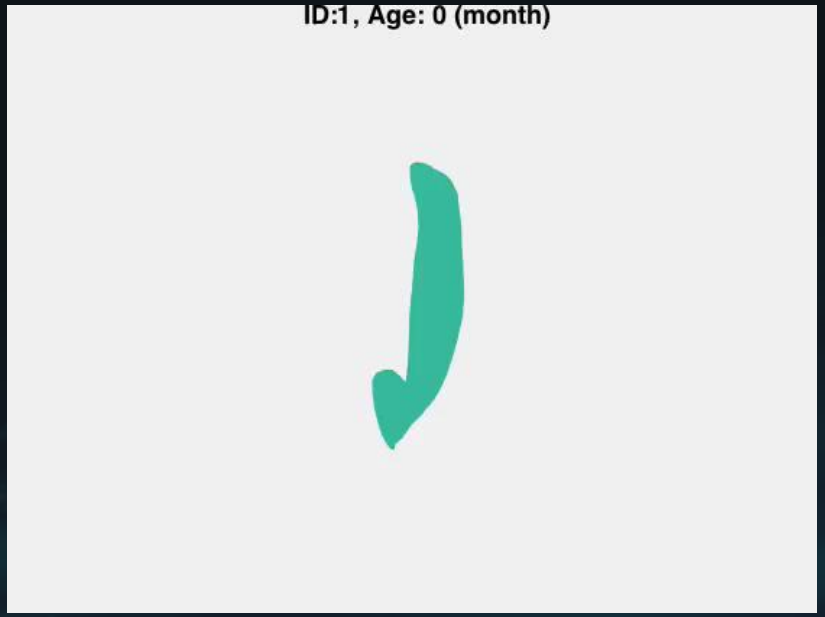


Figure 4: Trajectory fitting results of LESA from the observed sparse data in ADNIGO2

Case Study IV

Data Integration via D-CCA

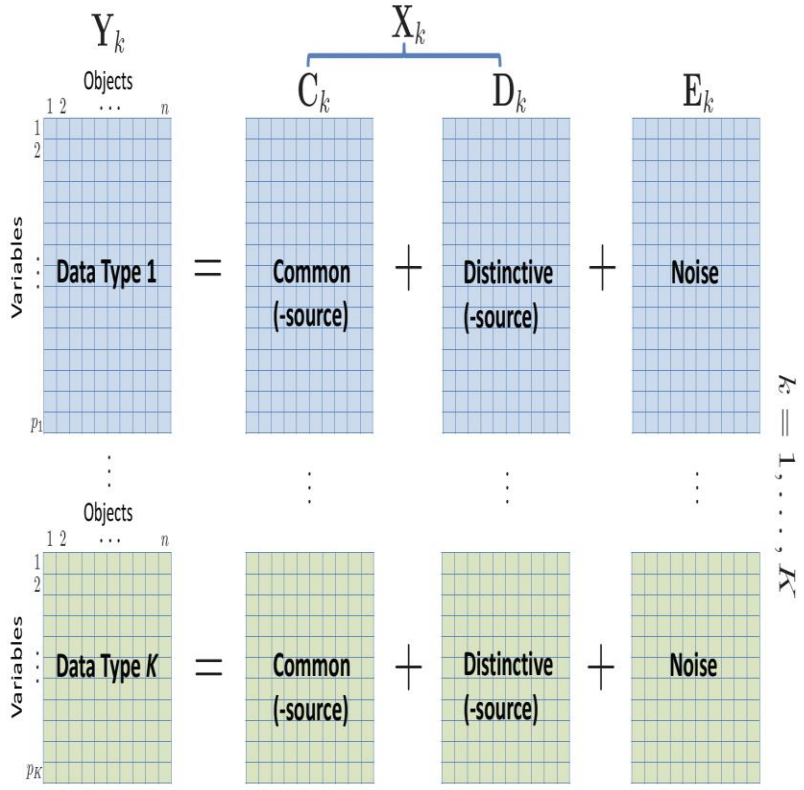
1. H. Shu, & H. Zhu (2025). D-CDLF: Decomposition of Common and Distinctive Latent Factors for Multi-view High-dimensional Data. Draft note available on arXiv:2407.00730v2

2. H. Shu, & H. Zhu (2025). Comments on: Data integration via analysis of subspaces (DIVAS), Test.

3. H. Shu, Z. Qu, & H. Zhu (2022). D-GCCA: Decomposition-based generalized canonical correlation analysis for multi-view high-dimensional data. Journal of Machine Learning Research. 23(169):1–64.

4. H. Shu, X. Wang, & H. Zhu (2020). D-CCA: A decomposition-based canonical correlation analysis for high-dimensional datasets. Journal of the American Statistical Association,115(529): 292-306.

Low-rank Plus Noise Models



- Only $\mathbf{Y}_k \in \mathbb{R}^{p_k \times n}$, $k = 1, \ldots, K$, are observable.
- $\mathbf{Y}_k = \mathbf{X}_k + \mathbf{E}_k$ is low-rank + noise.
- Decompose $\mathbf{X}_k = \mathbf{C}_k + \mathbf{D}_k$.

"Low-rank plus noise" model:

$$\mathbf{Y}_k = \mathbf{X}_k + \mathbf{E}_k = \mathbf{C}_k + \mathbf{D}_k + \mathbf{E}_k, \quad k = 1, \ldots, K$$

All common-source matrices

$\{\mathbf{C}_k\}_{k=1}^K$: from the common latent factors of the K data views.

k-th distinctive-source matrix

$\mathbf{D}_k$: from the distinctive latent factors of the k-th data view.

Steve's Contributions

Assume the columns of $\mathbf{Y}_k = \mathbf{X}_k + \mathbf{E}_k = \mathbf{C}_k + \mathbf{D}_k + \mathbf{E}_k \in \mathbb{R}^{p_k \times n}$

are n i.i.d. samples of mean-zero $\boldsymbol{y}_k = \boldsymbol{x}_k + \boldsymbol{e}_k = \boldsymbol{c}_k + \boldsymbol{d}_k + \boldsymbol{e}_k \in \mathbb{R}^{p_k}$.

Existing methods: different in defining the common & distinctive latent factors.

JIVE
Lock et al. (2013)

R.JIVE
O'Connell & Lock (2016)

AJIVE
Feng et al. (2018)

OnPLS
Löfstedt & Trygg (2011)

DISCO-SCA
Schouteden et al. (2014)

COBE
Zhou et al. (2016)

SLIDE
Gaynanova & Li (2019)

D-GCCA
Shu et al. (2022)

DIVAS
Prothero et al. (2024)

Drawback

Assume the columns of $\mathbf{Y}_k = \mathbf{X}_k + \mathbf{E}_k = \mathbf{C}_k + \mathbf{D}_k + \mathbf{E}_k \in \mathbb{R}^{p_k \times n}$

are n i.i.d. samples of mean-zero $\boldsymbol{y}_k = \boldsymbol{x}_k + \boldsymbol{e}_k = \boldsymbol{c}_k + \boldsymbol{d}_k + \boldsymbol{e}_k \in \mathbb{R}^{p_k}$.

In the (L^2, E) space of mean-zero real-valued random variables, $x \perp y \Leftrightarrow \operatorname{cov}(x, y) = 0$

Drawback: Fail to well consider the orthogonality (i.e., uncorrelatedness) among the common latent factors (CLFs) and distinctive latent factors (DLFs).

- Some only consider the orthogonality between CLFs and DLFs (OnPLS, DISCO-SCA, COBE, JIVE, AJIVE, DIVAS).
- Some only consider the orthogonality btwn DLFs from different data views (D-GCCA).
- Some consider both the two types of orthogonality, but either sacrifice unexplained signal as noise (R.JIVE) or offer an asymmetrical decomp. for identically distributed signals (SLIDE).

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 R01AG082938, 1R01AG085581, 1R01MH136055, 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.