A Statistician's Personal View on Uniting Statistics and AI for Real Applications

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, Mingyao Li, Marc Niethammer, Tengfei Li, and Bingxin Zhao for sharing their slides.

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

CONTENTS

Part I

Statistical Modeling: The Two Cultures

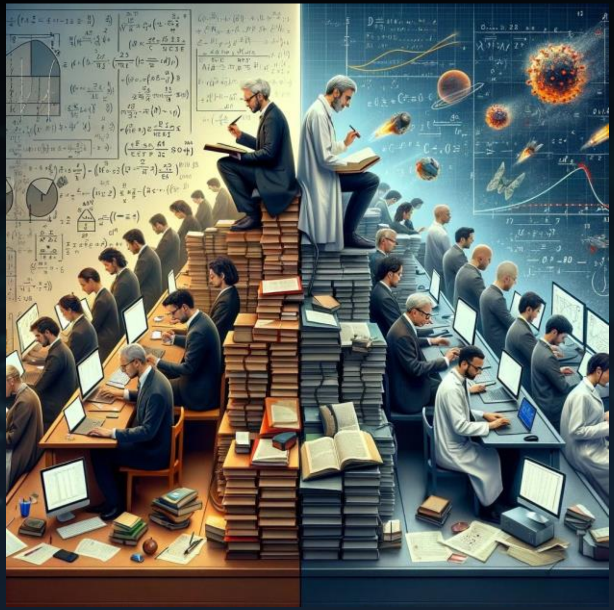
“The best thing about being a statistician is that you get to play in everyone's backyard.”
- John Tukey -

Statistical Modeling: The Two Cultures

Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data. **https://en.wikipedia.org/wiki/Statistics**

Leo Breiman (2001). Statistical Modeling: The Two Cultures. *Statistical Science*.

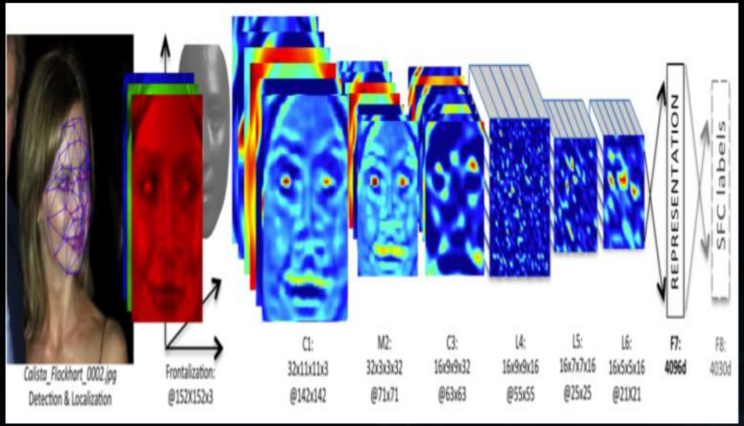
"There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of **data models**. This commitment has led to **irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems.** **Algorithmic modeling**, both in theory and practice, has developed rapidly in **fields outside statistics**. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. *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."*

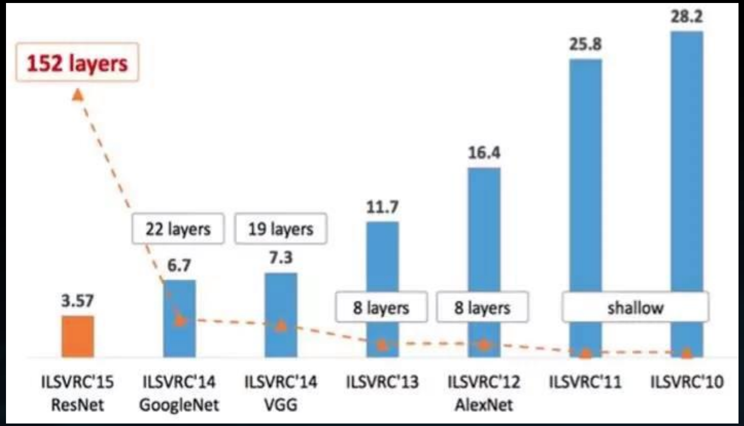


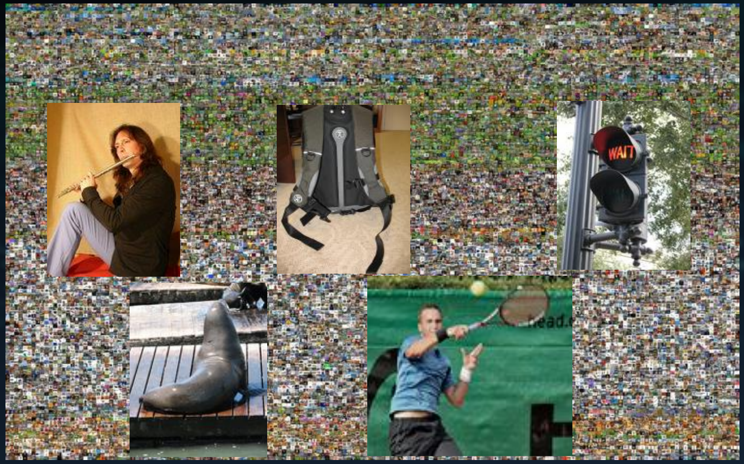
AI Milestones

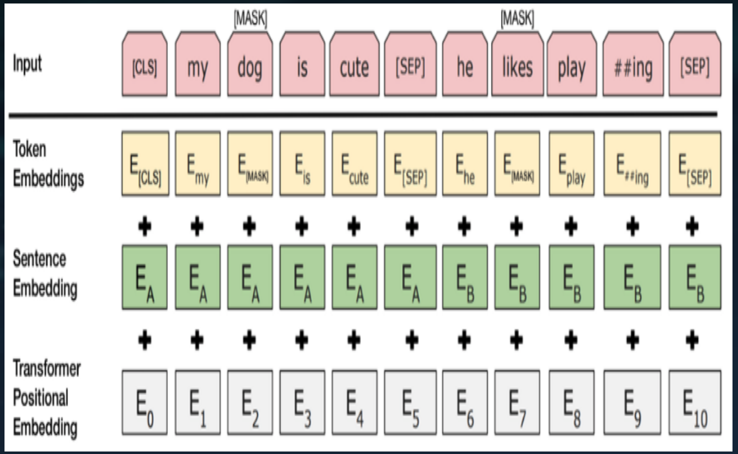
Annotated Datasets

Algorithmic modeling = Deep Learning



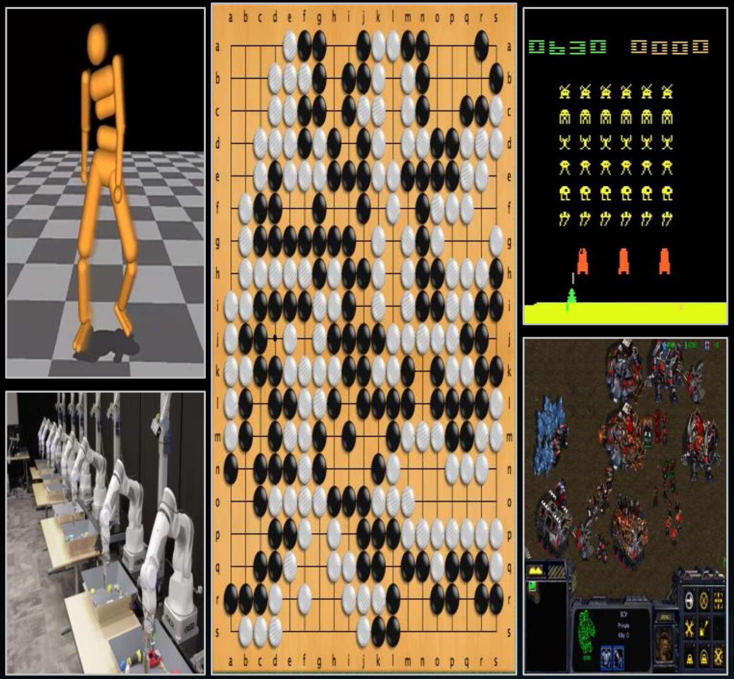




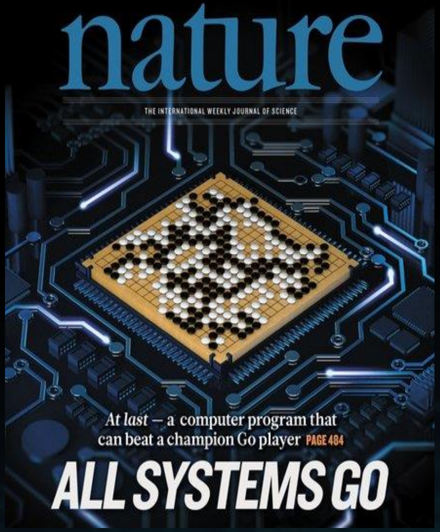


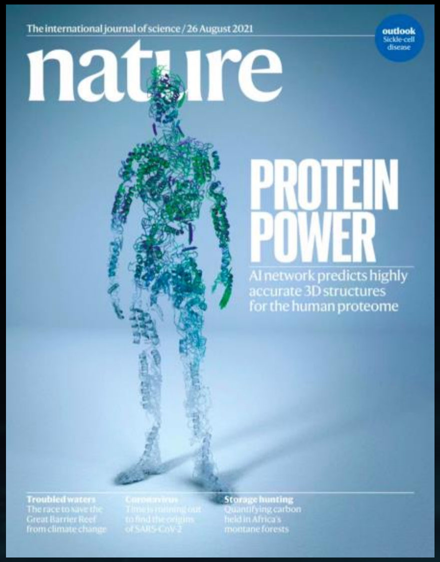
AI Milestones

Reinforcement Learning



AI Products

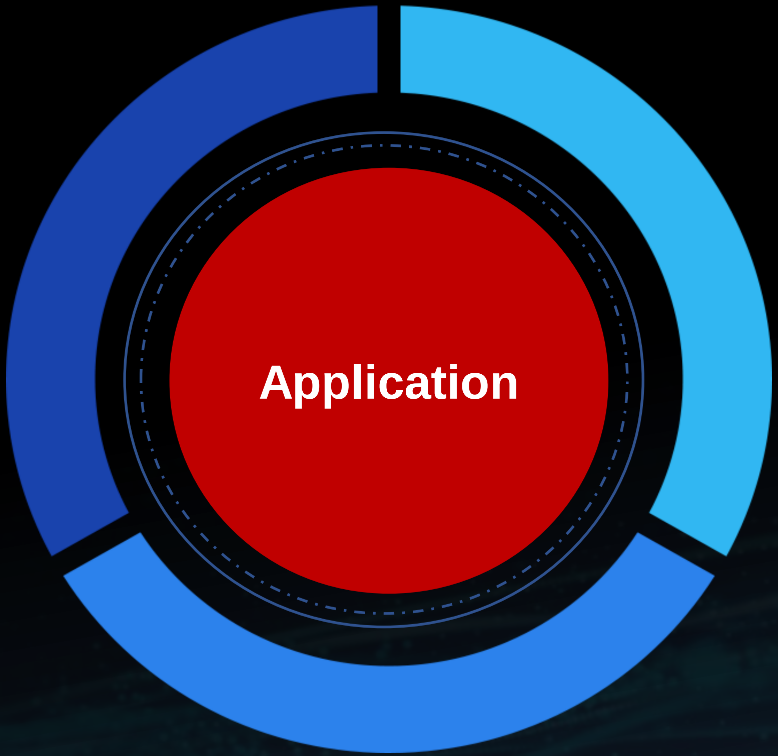


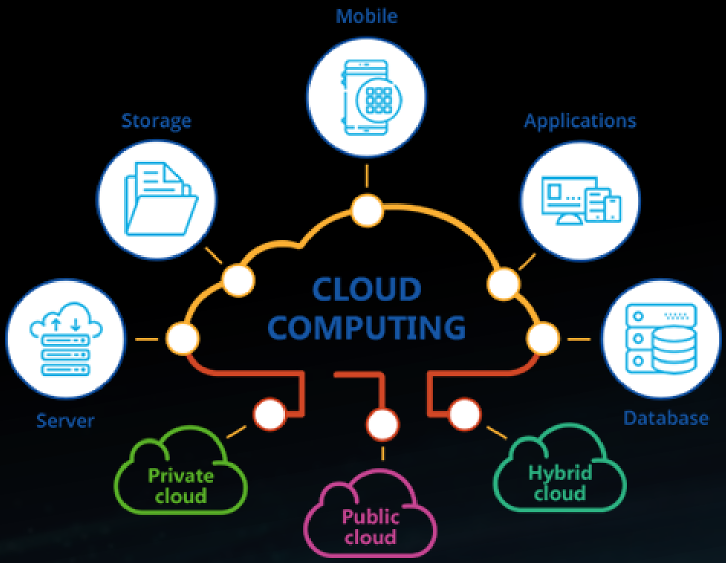


AI=Application to ABC

Big Data

http://medium.com





Computing

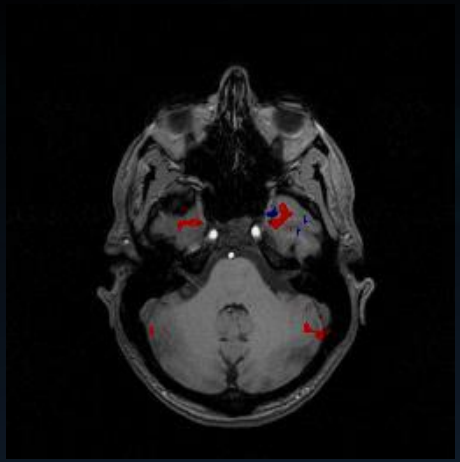
Analytical Tools

- Applied Mathematics
- Statistics
- Machine Learning
- Engineering

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

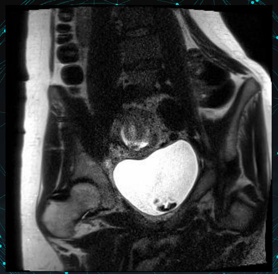
Image Processing Analysis Methods

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

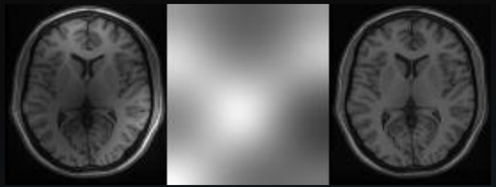
Deconvolution

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

➢ **Image Reconstruction**



➢ **Image Enhancement**



❖ **Image Segmentation**

❖ **Image Registration**

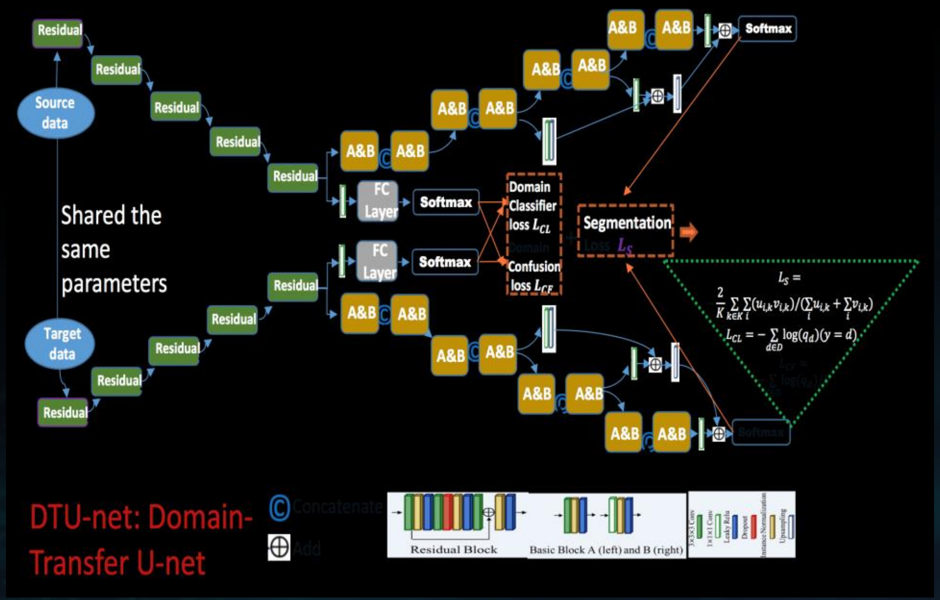
Structural Learning

Engineering | Statistics | Machine Learning | Mathematics

AI for Image Segmentation

Segmentation Annotation

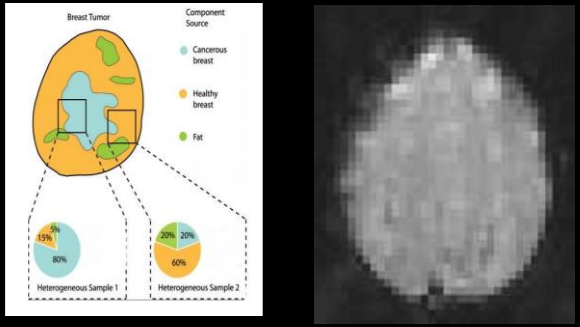
U-Nets



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

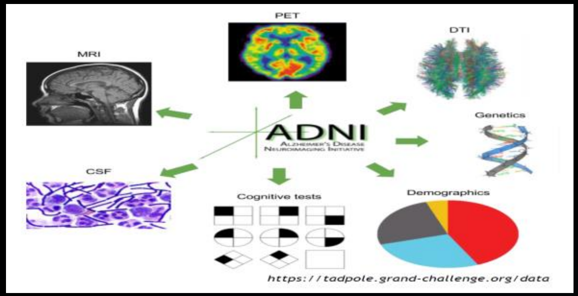
R. Azad *et al.*, “Medical Image Segmentation Review: The success of U-Net.” arXiv, Nov. 27, 2022.
Minaee, Shervin, et al. "Image segmentation using deep learning: A survey." *IEEE PAMI* 44.7 (2021): 3523-3542.

Ecological Layout for Imaging-based Analysis

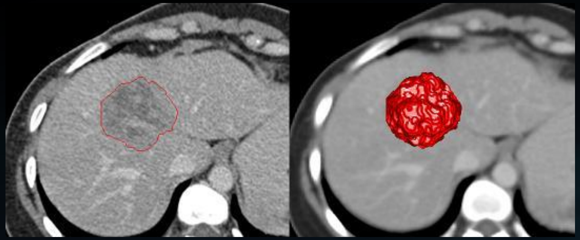


Deconvolution

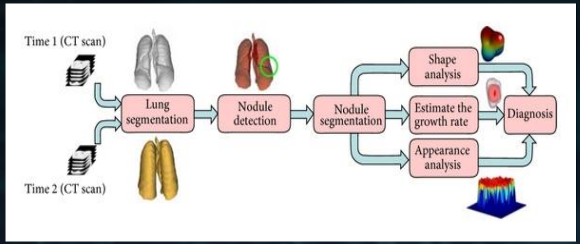
Imaging-related Database



Integration



Structural Learning



Prediction

Part II

Opportunities for Statisticians

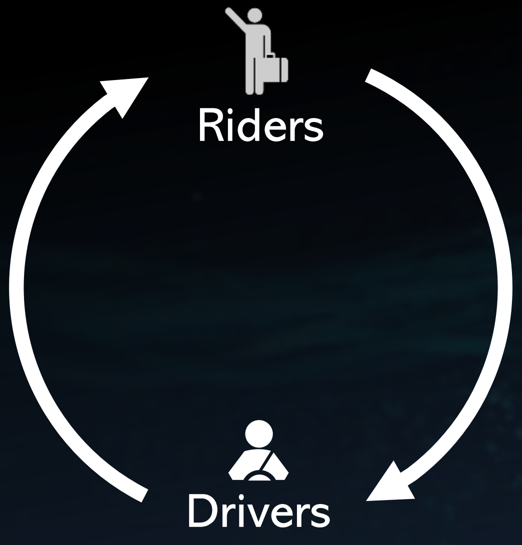
"If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools."

- Leo Breiman -

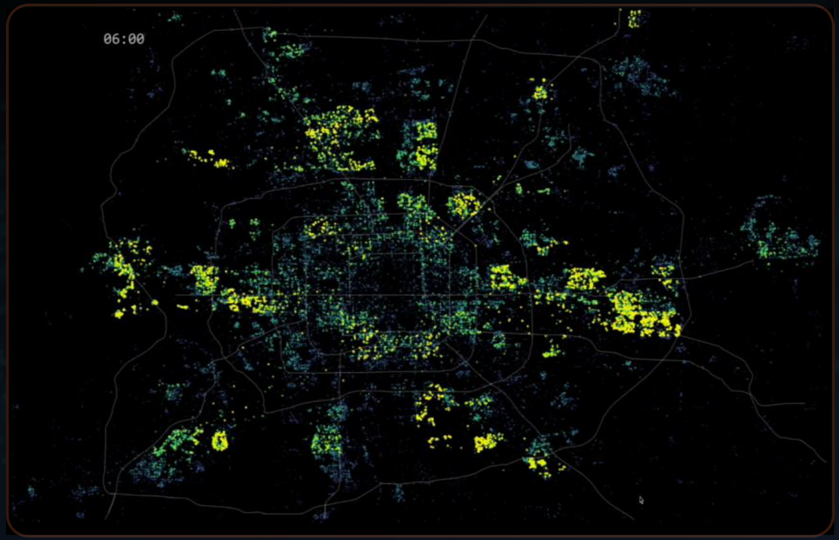
Ride-sharing Platform is a Complex Ecosystem

Spatio-temporal | Nonlinear | Interactive | Uncertainty | Causal

Two-sided Platform



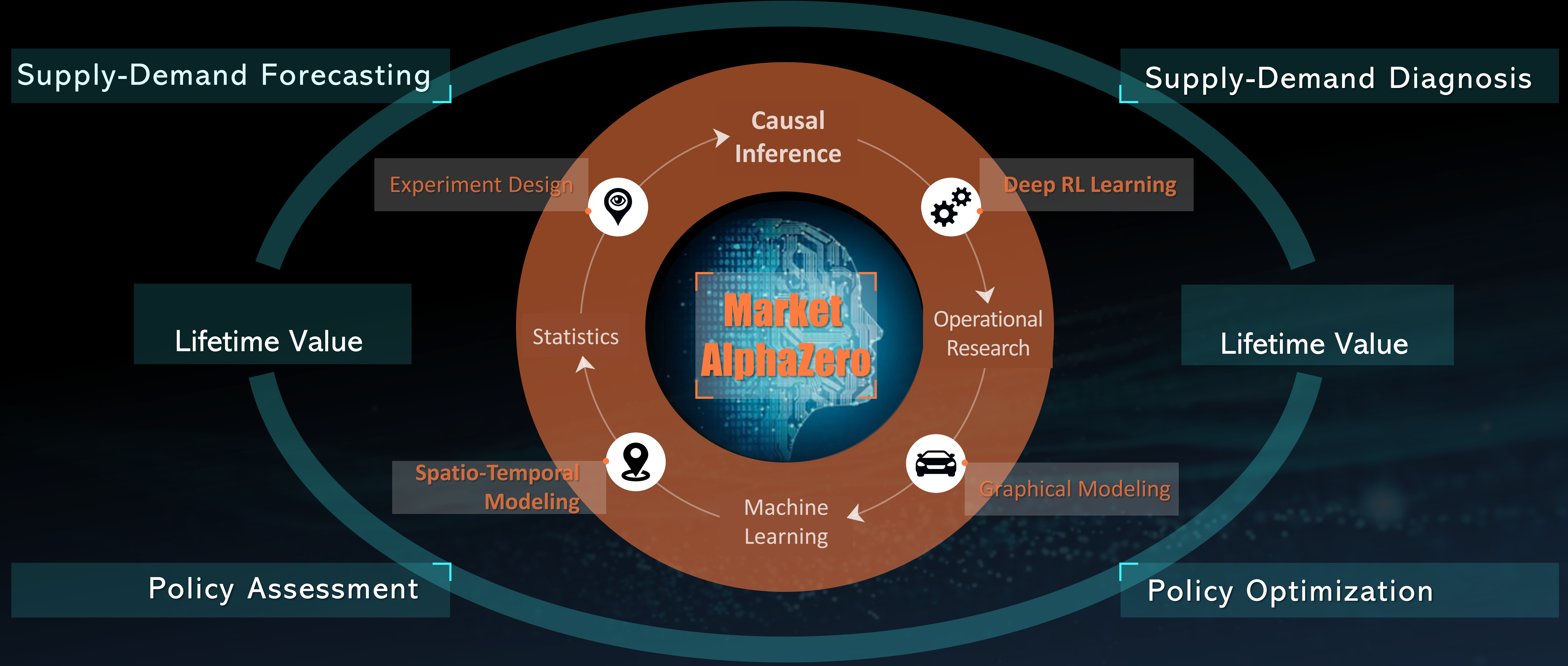
Complex Spatio-temporal System



UNC Biostatistics

Leverage Supply-Demand Network Effect

How to evaluate and improve the operational efficiency of ride-sharing platform?



UNC Biostatistics

Supply-Demand Forecasting

The Problem

Predicting the demand-supply distribution

Model

- Multi-modal data fusion
- Complex spatio-temporal patterns

Transfer

- Heterogeneous space among cities
- Heterogeneous feature among tasks

Recognition

- Causal inference
- Model interpretation
- Impact analysis

The Goal

Improve the service quality

Drivers

- Reduce empty driving

Riders

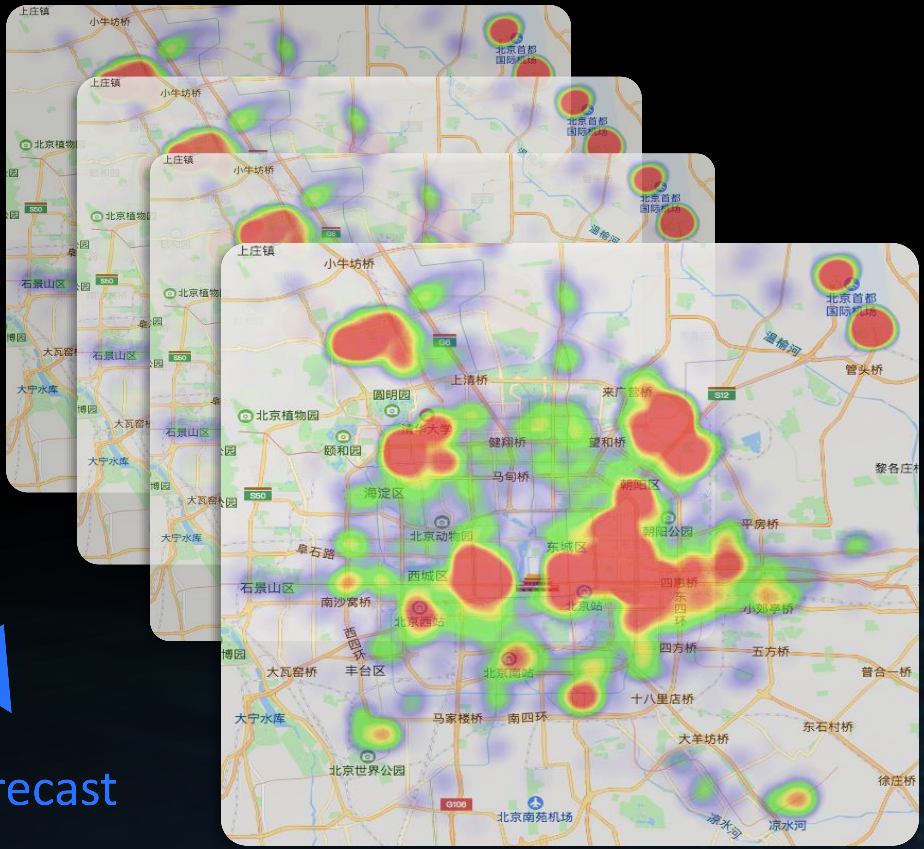
- Intelligent travel guidance
- Less queueing time

Platform

- Fill demand-supply gap
- Recognize the market
- Better dispatching and scheduling

UNC Biostatistics

A Deep S-T Forecasting Model

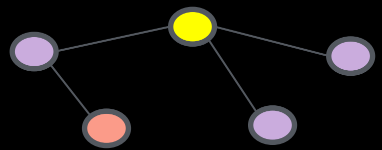


History

Forecast

Graph Generation

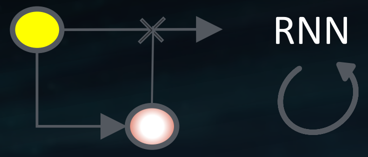
Spatio-temporal Forecasting



Incorporating
multi-relationship

Temporal: CGRNN

Contextual-gated
temporal modeling

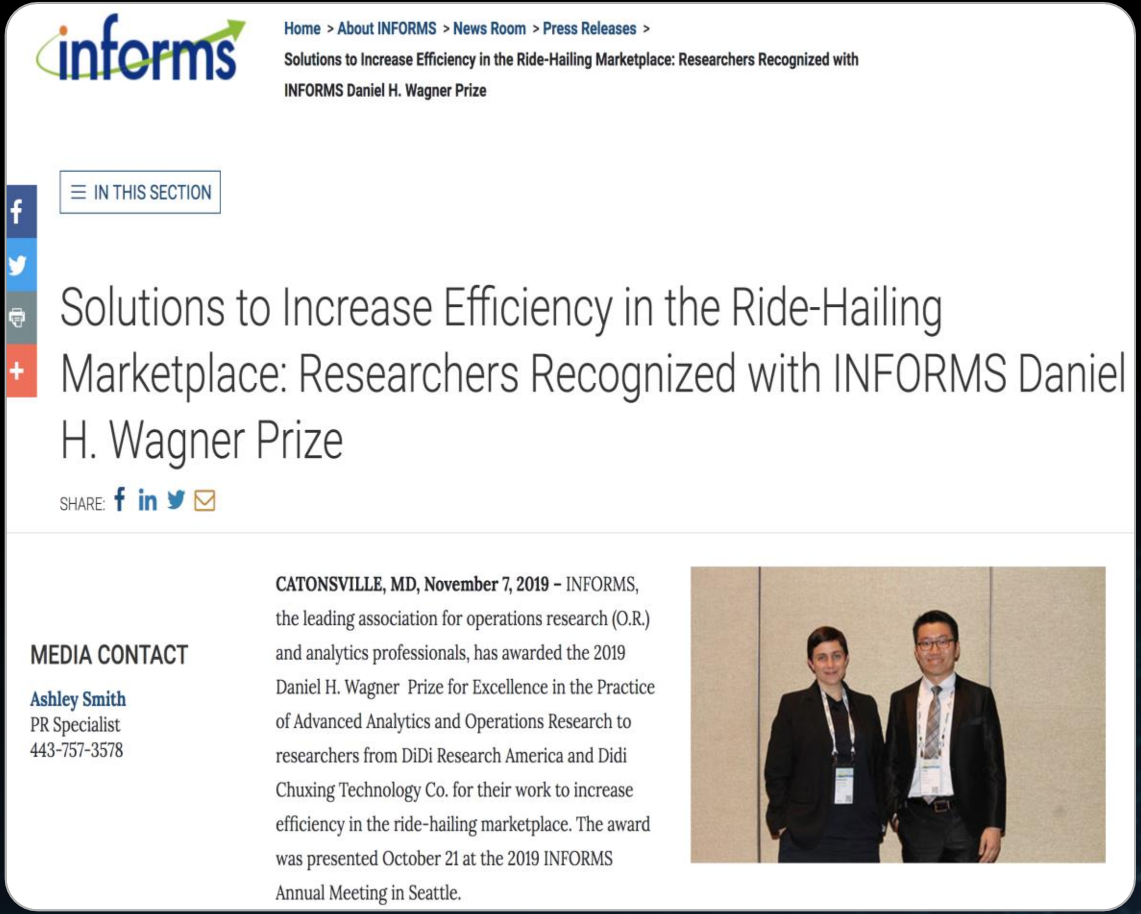


Spatio-MGCN

Non-euclidean
spatial modeling

UNC Biostatistics

Deep Reinforcement Learning


Home > About INFORMS > News Room > Press Releases >
Solutions to Increase Efficiency in the Ride-Hailing Marketplace: Researchers Recognized with INFORMS Daniel H. Wagner Prize

☰ IN THIS SECTION

Solutions to Increase Efficiency in the Ride-Hailing Marketplace: Researchers Recognized with INFORMS Daniel H. Wagner Prize

SHARE:

MEDIA CONTACT

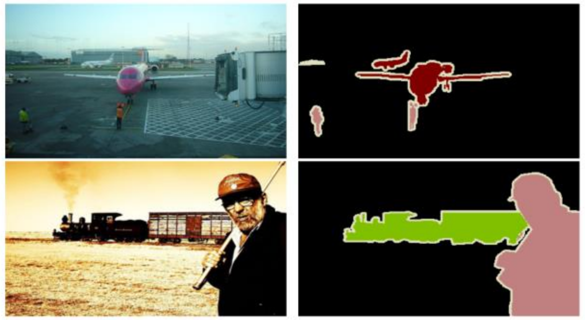
Ashley Smith
PR Specialist
443-757-3578

CATONSVILLE, MD, November 7, 2019 – INFORMS, the leading association for operations research (O.R.) and analytics professionals, has awarded the 2019 Daniel H. Wagner Prize for Excellence in the Practice of Advanced Analytics and Operations Research to researchers from DiDi Research America and Didi Chuxing Technology Co. for their work to increase efficiency in the ride-hailing marketplace. The award was presented October 21 at the 2019 INFORMS Annual Meeting in Seattle.

UNC Biostatistics

Generative Model

- Image data $X \in \mathcal{X}(\Omega, \mathbb{R}^3)$ on grid domain Ω. Given intensity $\lambda : \Omega \to \mathbb{R}^+$,
$$X = X^1_{obj}(u_1) \oplus \cdots \oplus X^K_{obj}(u_K) \oplus \epsilon$$
  - ▷ Objects: $X^1_{obj}, \ldots, X^K_{obj}$; Background noise: ϵ.
  - ▷ Number of objects: $K \sim Poisson(\Lambda(\Omega)), \Lambda(\Omega) = \int_\Omega \lambda(t)dt$
  - ▷ Locations of objects: $u_k \sim P_\lambda, \quad k = 1, \ldots, K.$



- Size of the object: $\alpha := \frac{\mathrm{Card}(\mathrm{Box}(X_{obj}))}{\mathrm{Card}(\Omega)} \in (0, 1],$

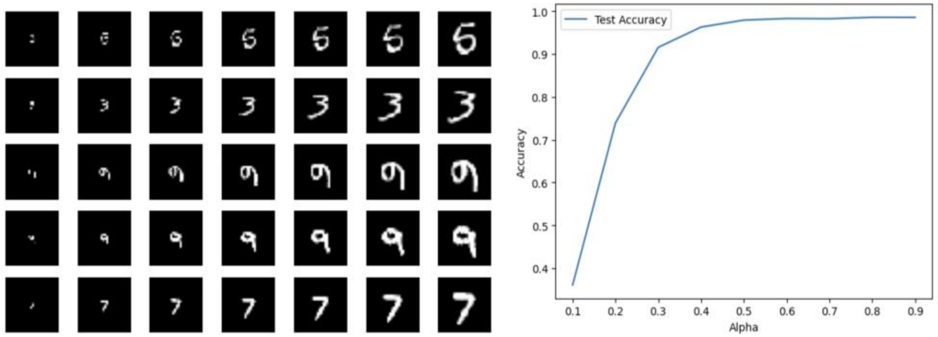


Figure: Test accuracy on MNIST dataset trained on a three-layer CNN.

CNN: Conv1(out_channels=16, kernel_size=3, padding=1); MaxPool(2); Conv2(out_channels=32, kernel_size=3, padding=1); MaxPool(2); Linear(out_dim=10).

- Entropy of the intensity: $\lambda(\cdot) : \Omega \to \mathbb{R}^+$

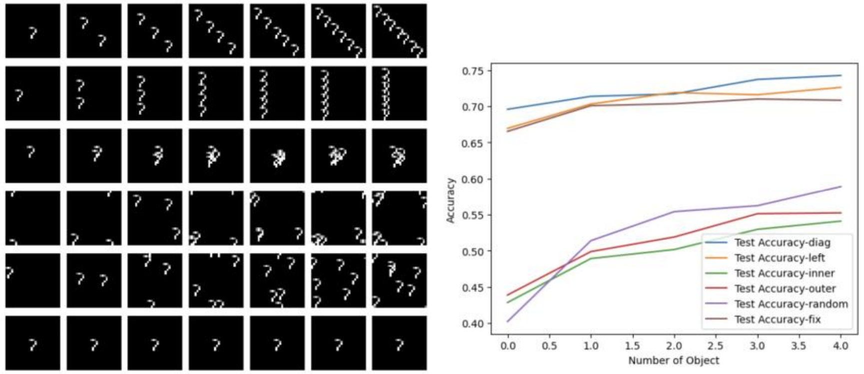


Figure: Test accuracy on MNIST dataset ($\alpha = 0.2$) trained on a three-layer CNN.

CNN: Conv1(out_channels=16, kernel_size=3, padding=1); MaxPool(2); Conv2(out_channels=32, kernel_size=3, padding=1);

- Number of objects: $K \sim Poisson(\Lambda(\Omega)), \Lambda(\Omega) = \int_\Omega \lambda(t)dt$

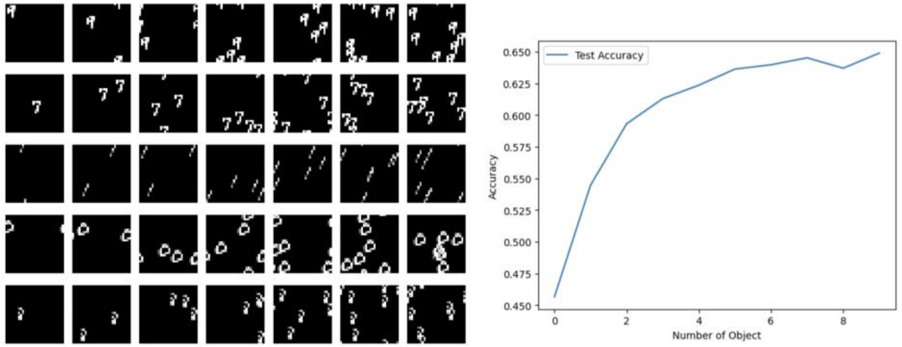
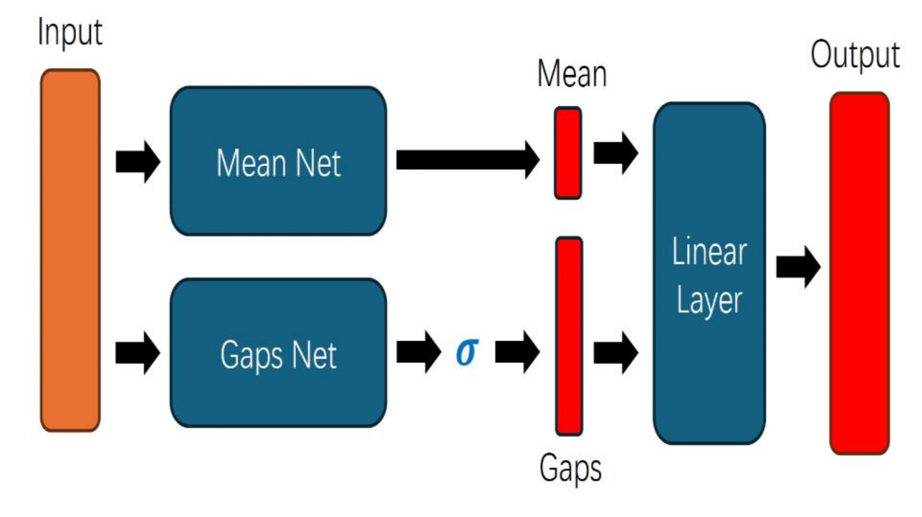
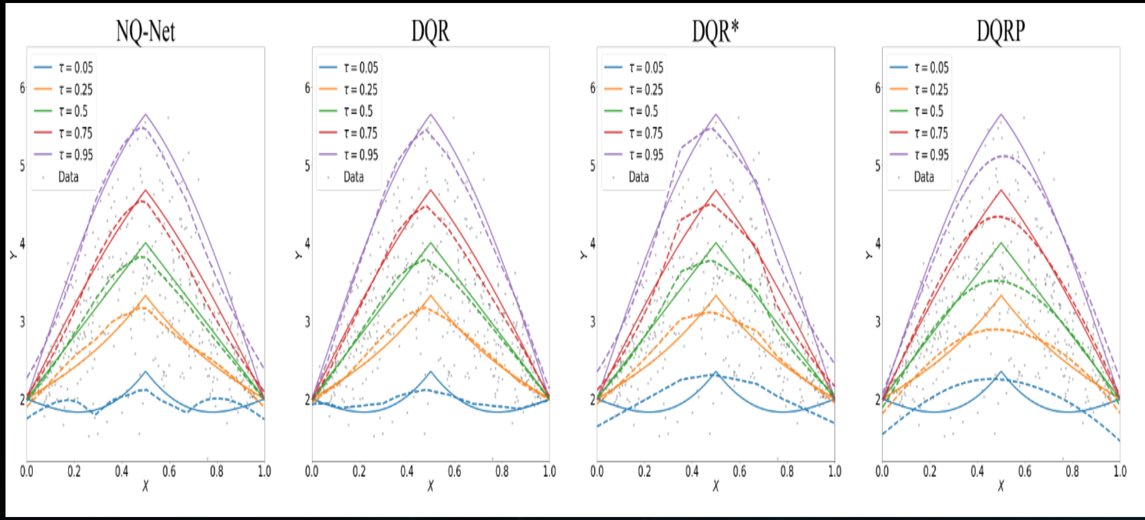


Figure: Test accuracy on MNIST dataset ($\alpha = 0.2$) trained on a three-layer CNN.

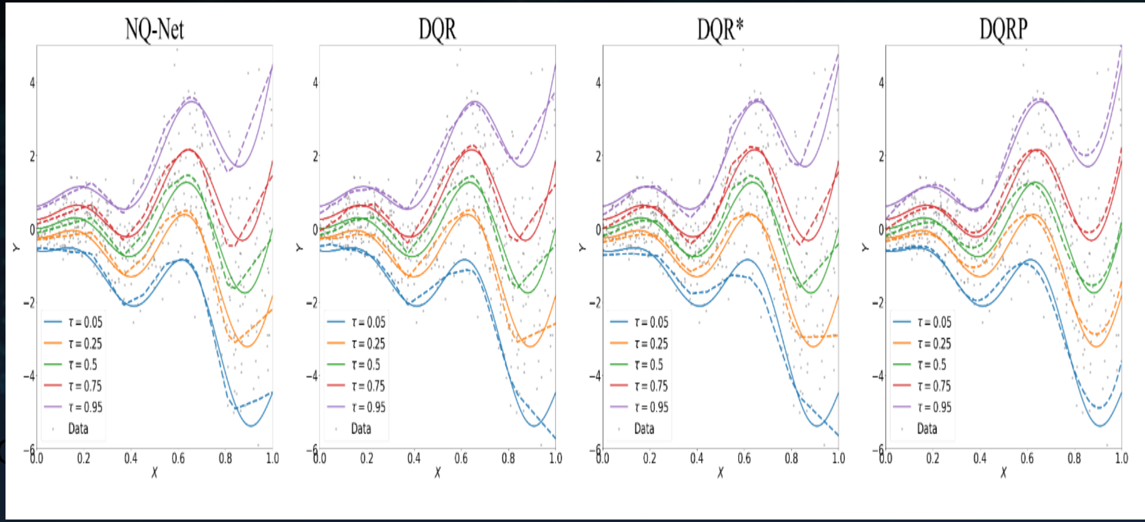
CNN: Conv1(out_channels=16, kernel_size=3, padding=1); MaxPool(2); Conv2(out_channels=32, kernel_size=3, padding=1); MaxPool(2); Linear(out_dim=10).

Deep Distributional Learning



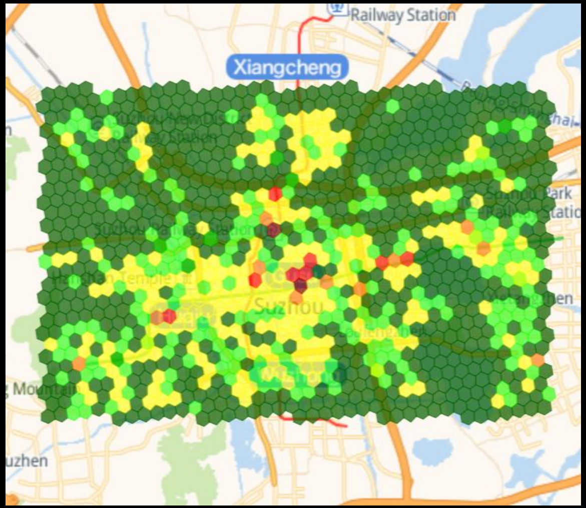


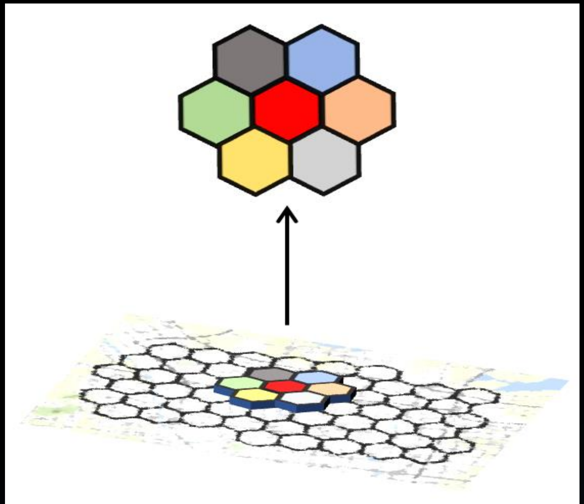
Non-crossing Quantile Network

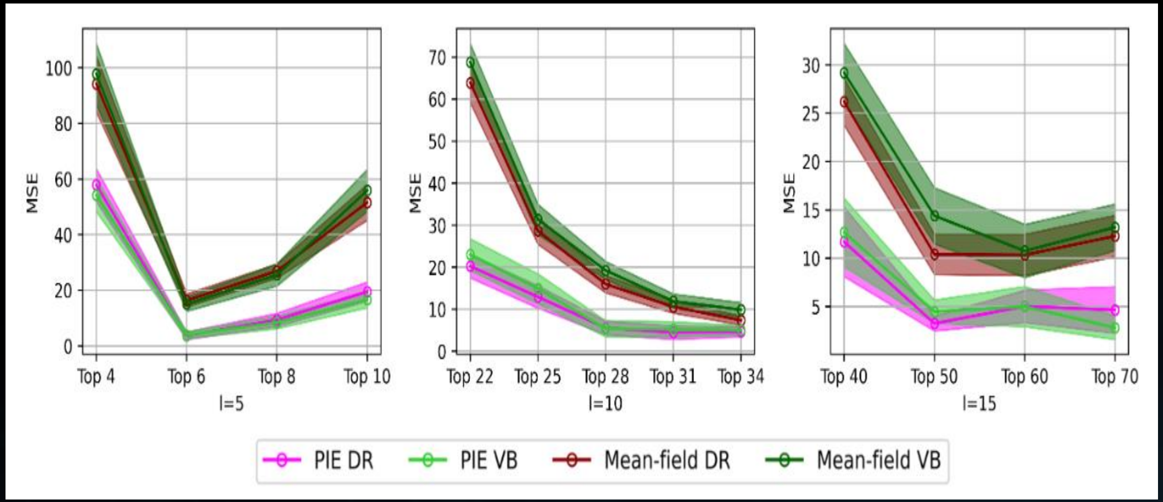


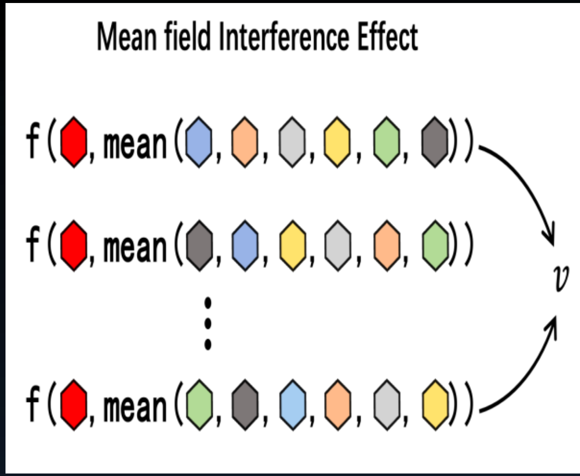
NQ netw
structure

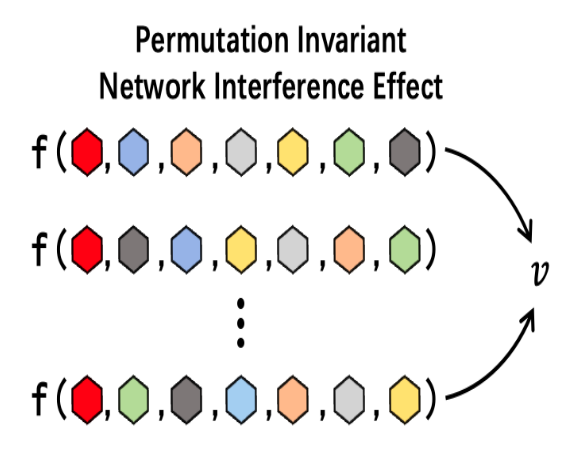
Causal Deepset for Offline Policy Learning

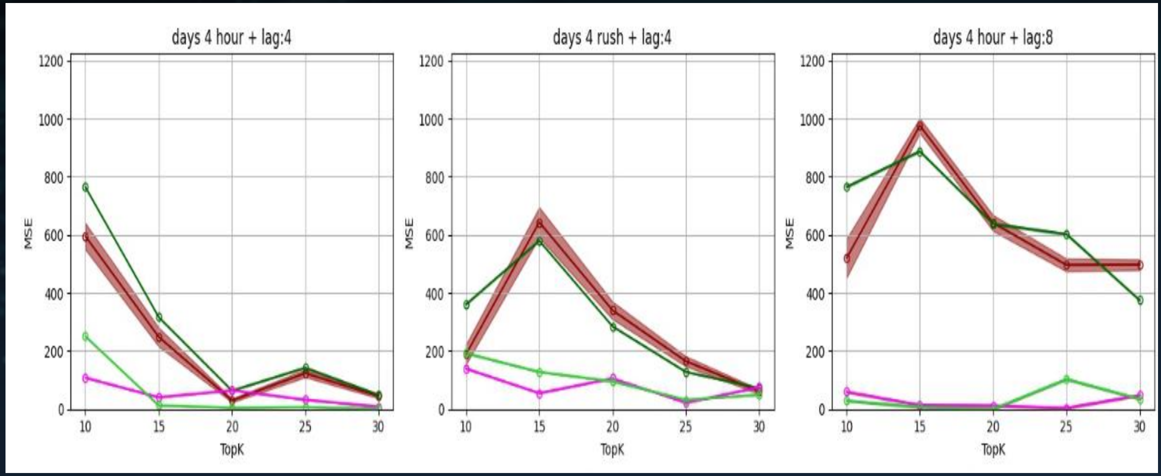






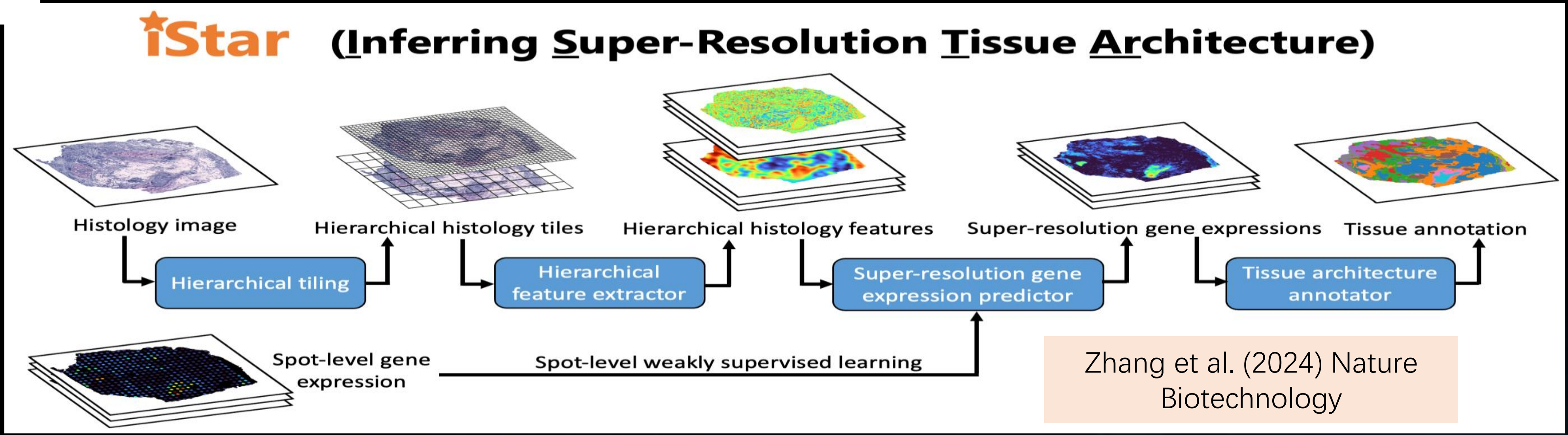






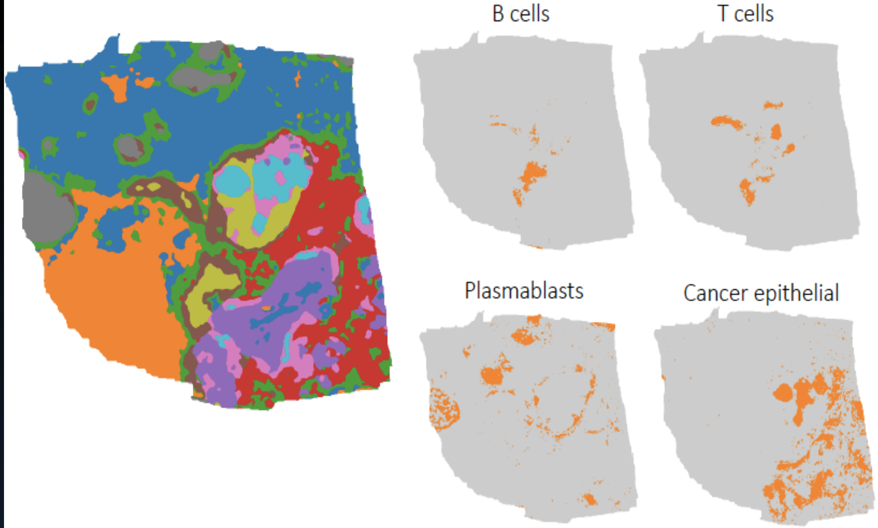
structure

IStar



Zhang et al. (2024) Nature Biotechnology

iStar can automatically annotate cell types



iStar can automatically detect positive surgical margin

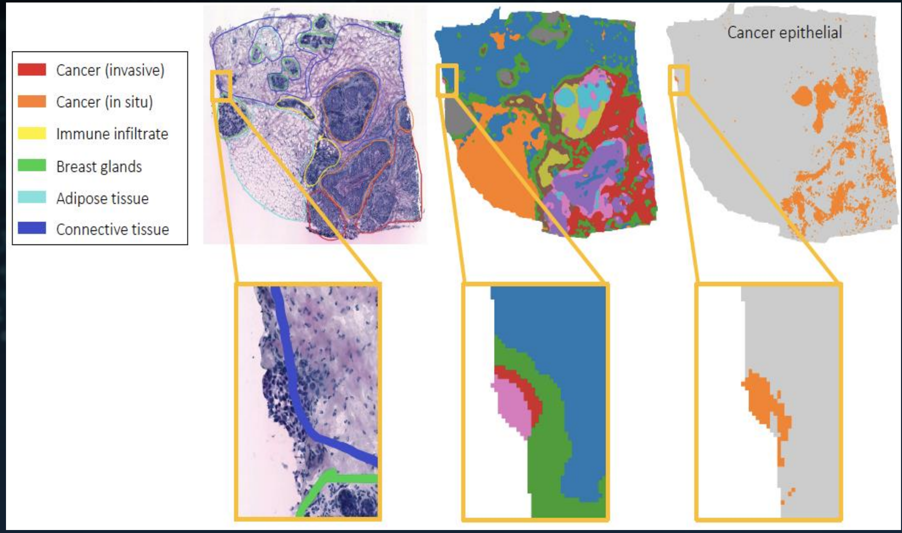
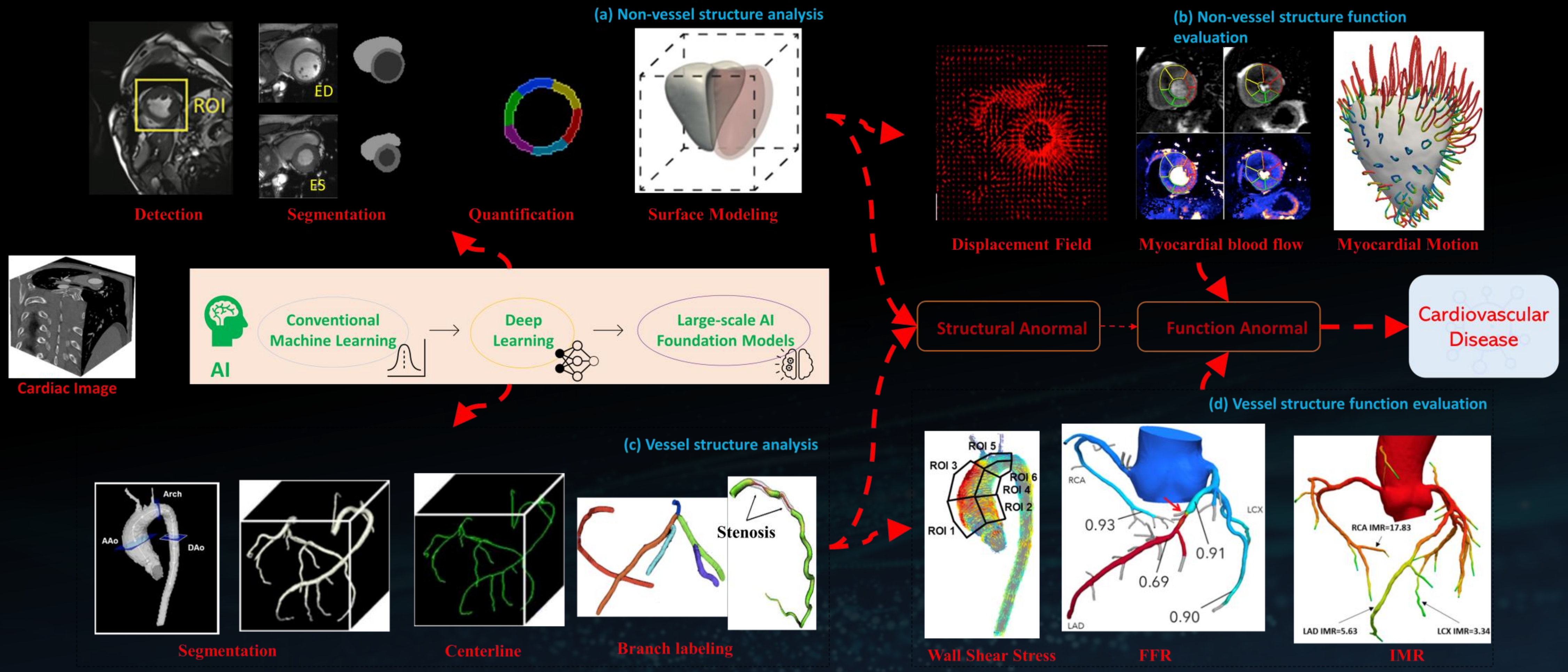
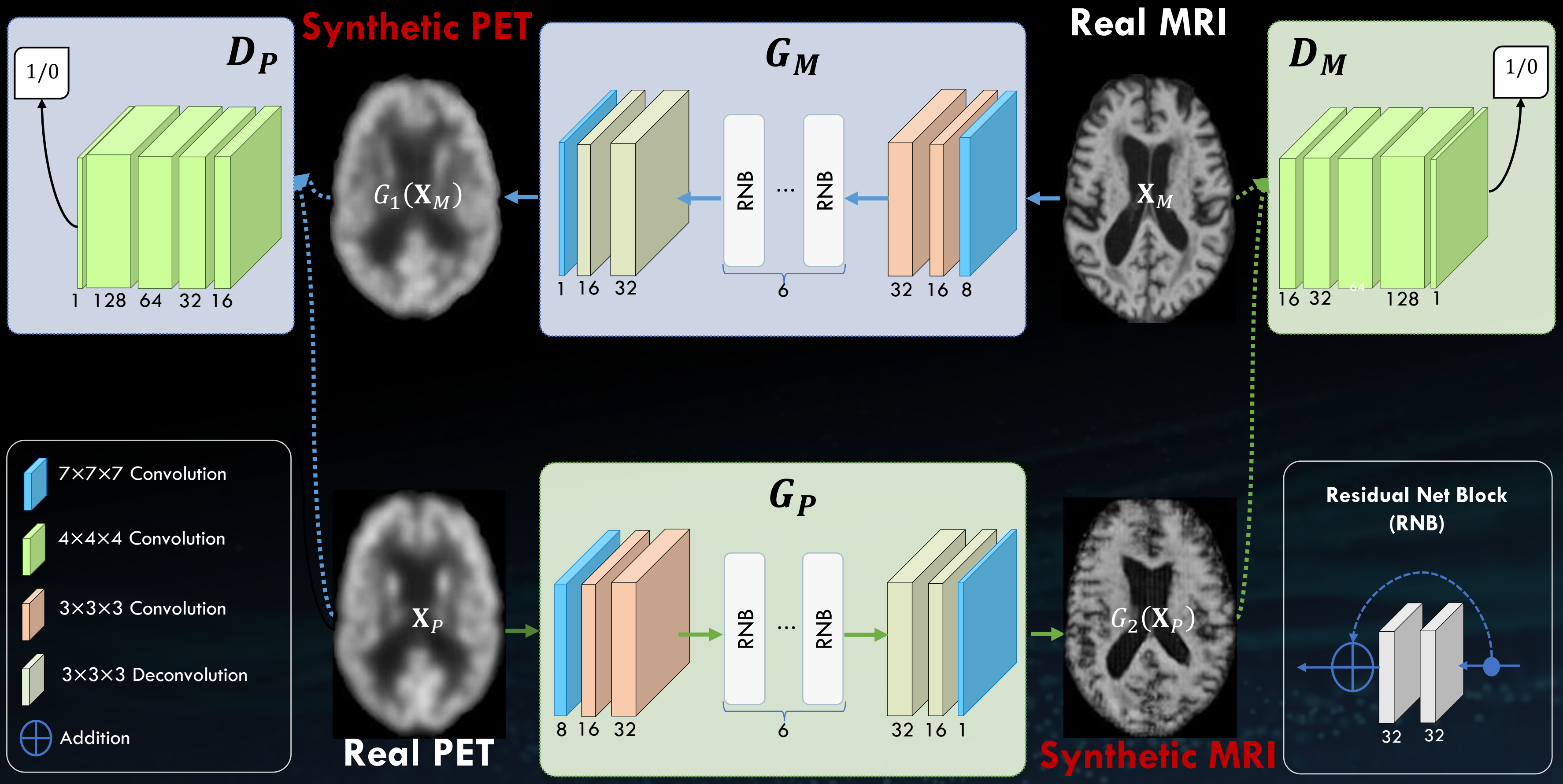


Image Analysis Pipeline

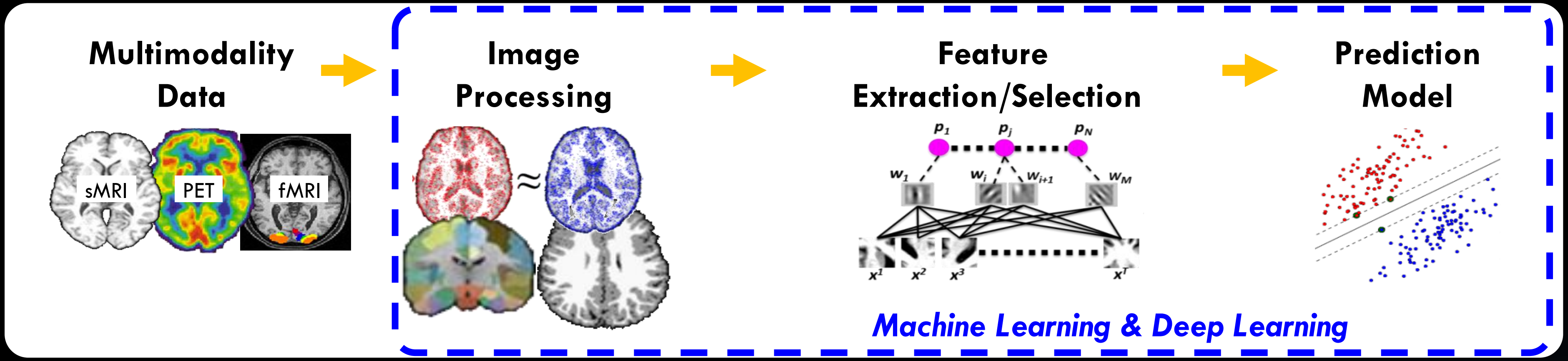


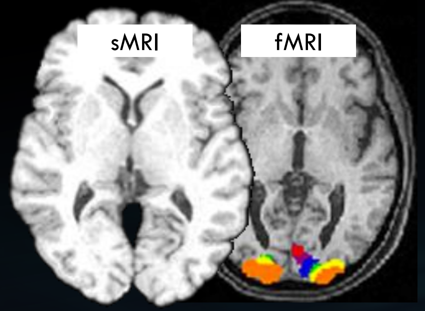
Wang, X. and Zhu, H (2024). Artificial Intelligence in Image-based Cardiovascular Disease Analysis: A Comprehensive Survey and Future Outlook

Cross-Modality Image Synthesis

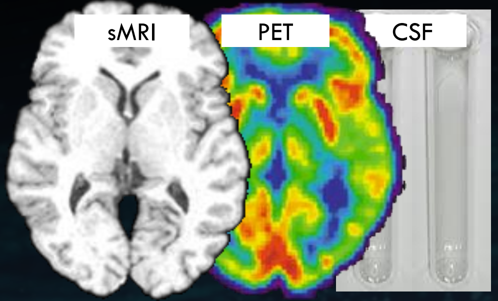


Computer-Aided Medical Data Analysis

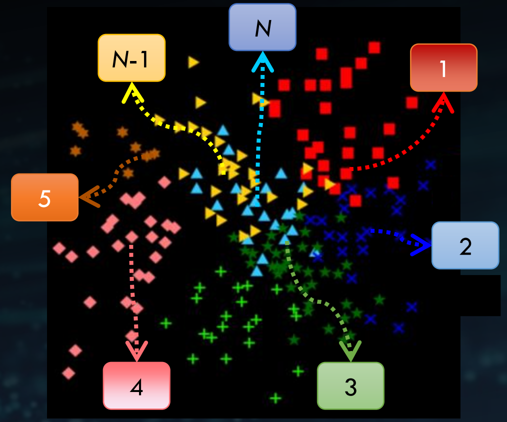




Neuroimage Representation Learning



Multimodality Data Fusion



Multi-Site Data Adaptation

Prediction Models

Preprocess

Baseline	Baseline	Baseline
Lifestyle	Lifestyle	Lifestyle
Measurement	Measurement	Measurement

Purpose determined data input

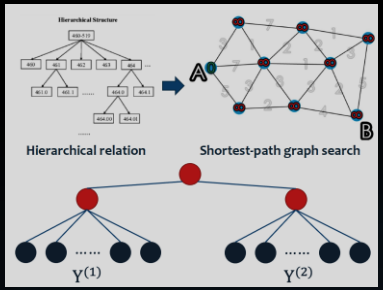
Baseline Imaging Environment
Genetic Lifestyle Disease Diagnosed Measurement

Baseline Imaging Environment
Genetic Lifestyle Disease Diagnosed Measurement

Model

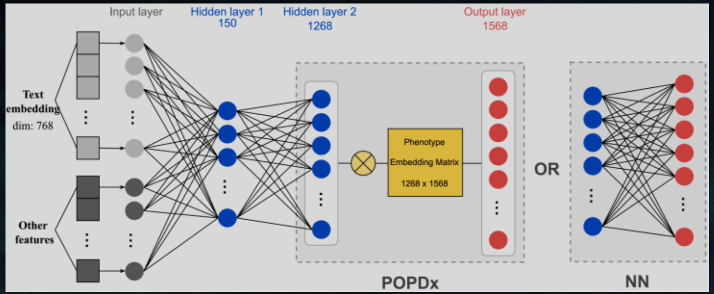
Boosting | POPDX | Neural network | Logistic regression | …

CoxPH | Deep Surv | Deep Hit | POPDx Surv | …



Disease diagnosis & Risk assessment

- Individual disease
- Categorized disease
- Joint diagnosis and evaluation

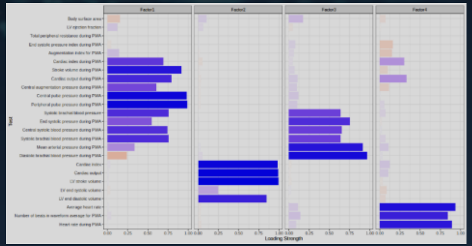


Post analysis

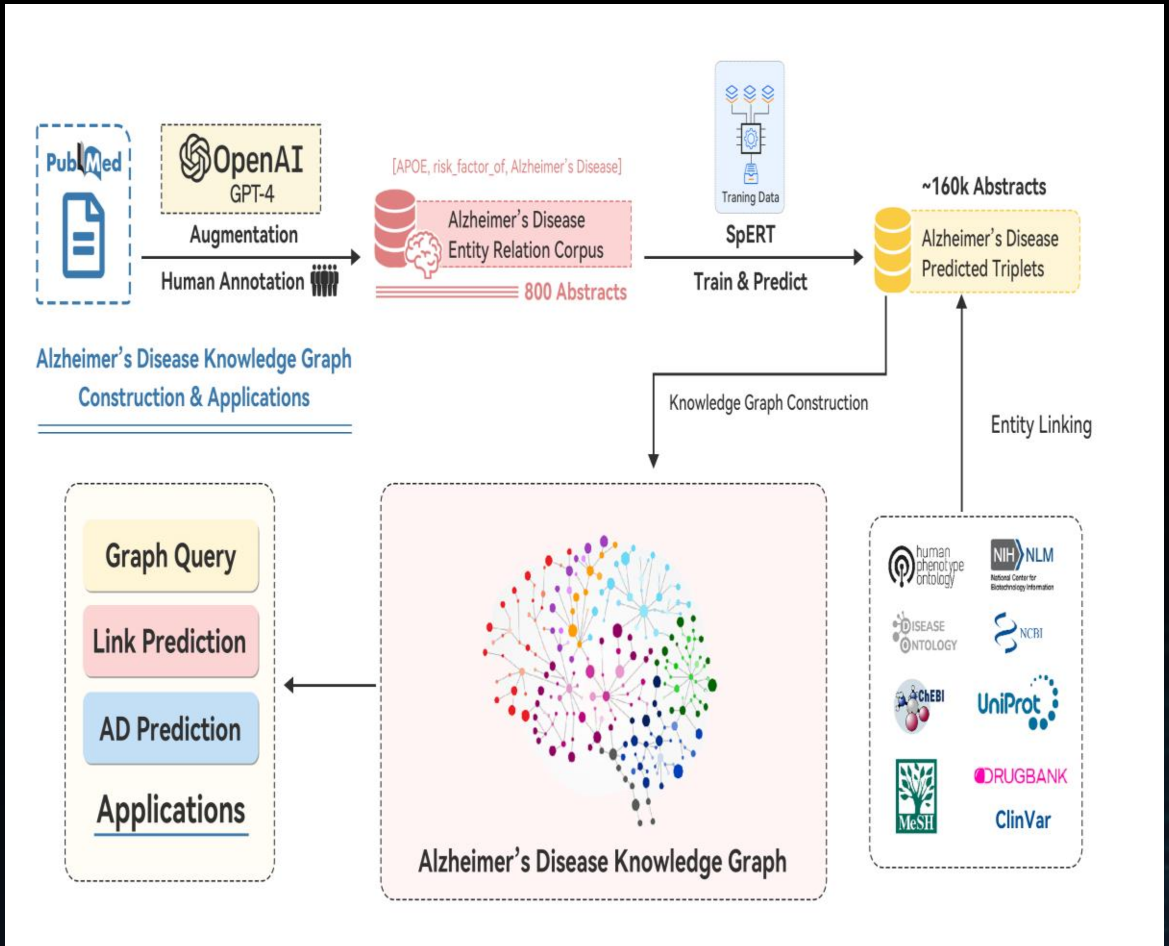
Feature importance extraction

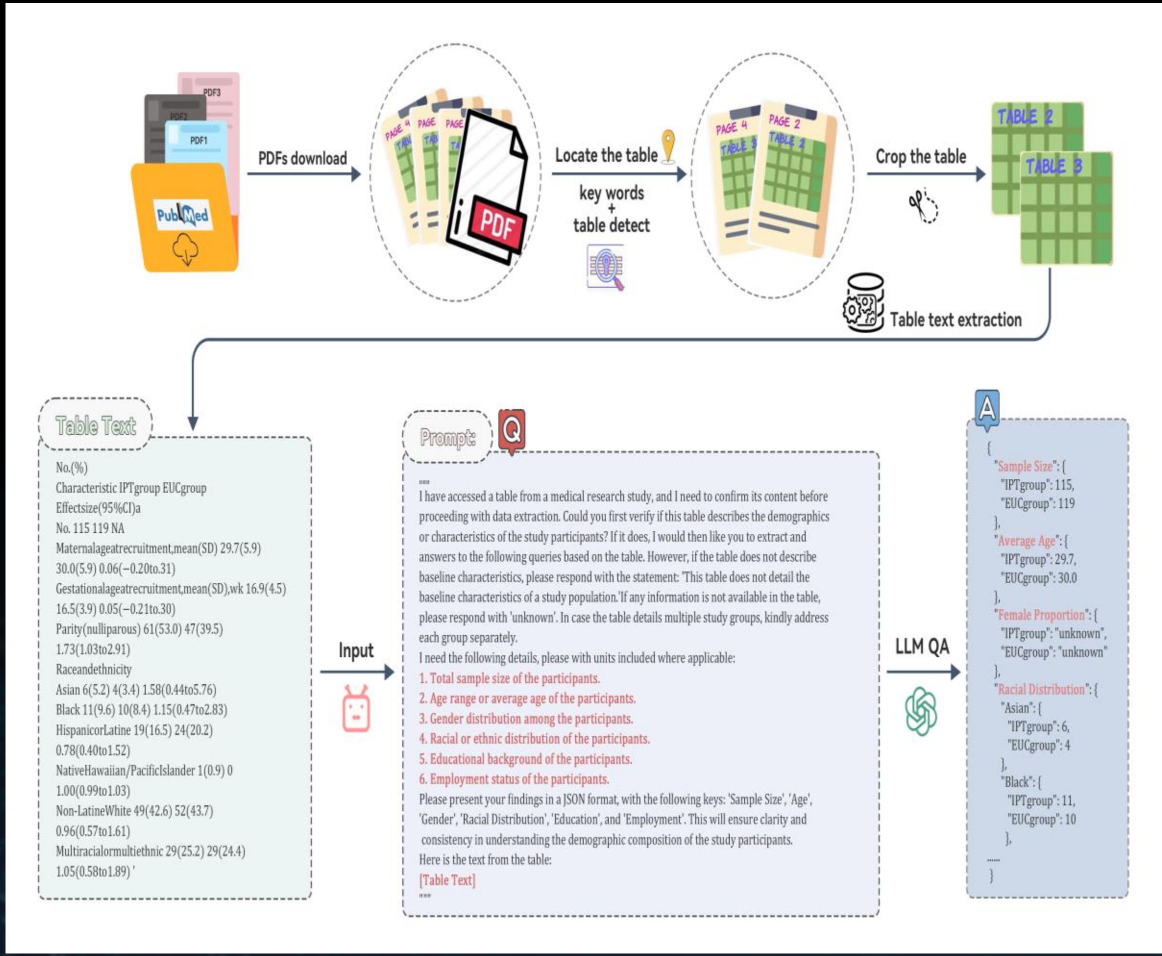
- Deep Lift
- SHAP
- Integrated Gradients

Data-driven dimension reduction



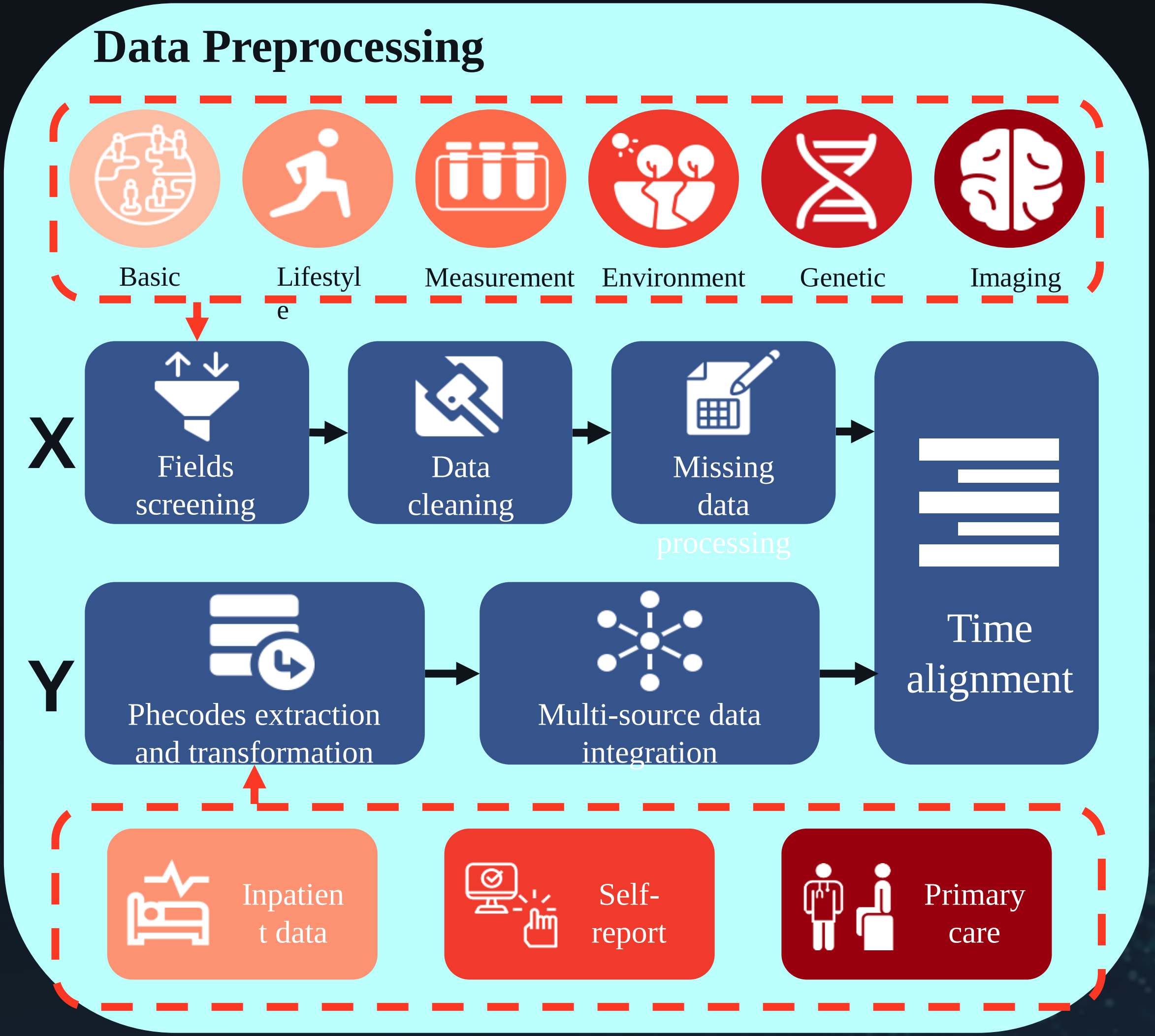
Knowledge Graph Construction

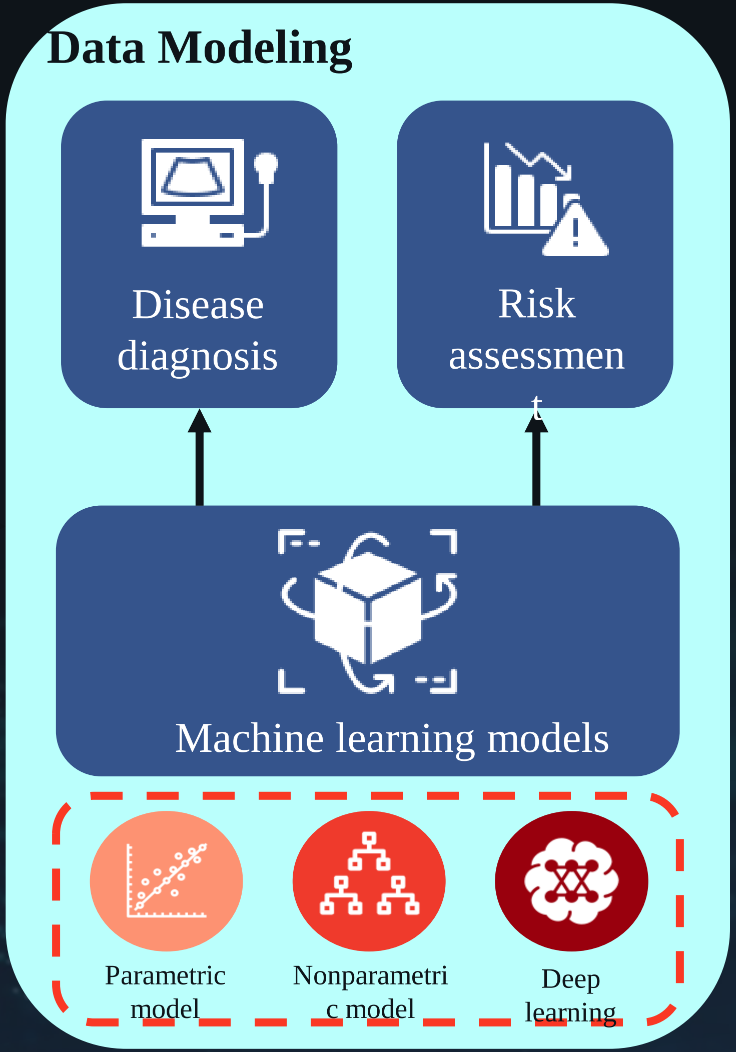




Yang et al., Alzheimer's Disease Knowledge Graph Enhances Knowledge Discovery and Disease Prediction.

Data Preprocessing and Data Modeling





Foundation Models for GMAI

Perspective

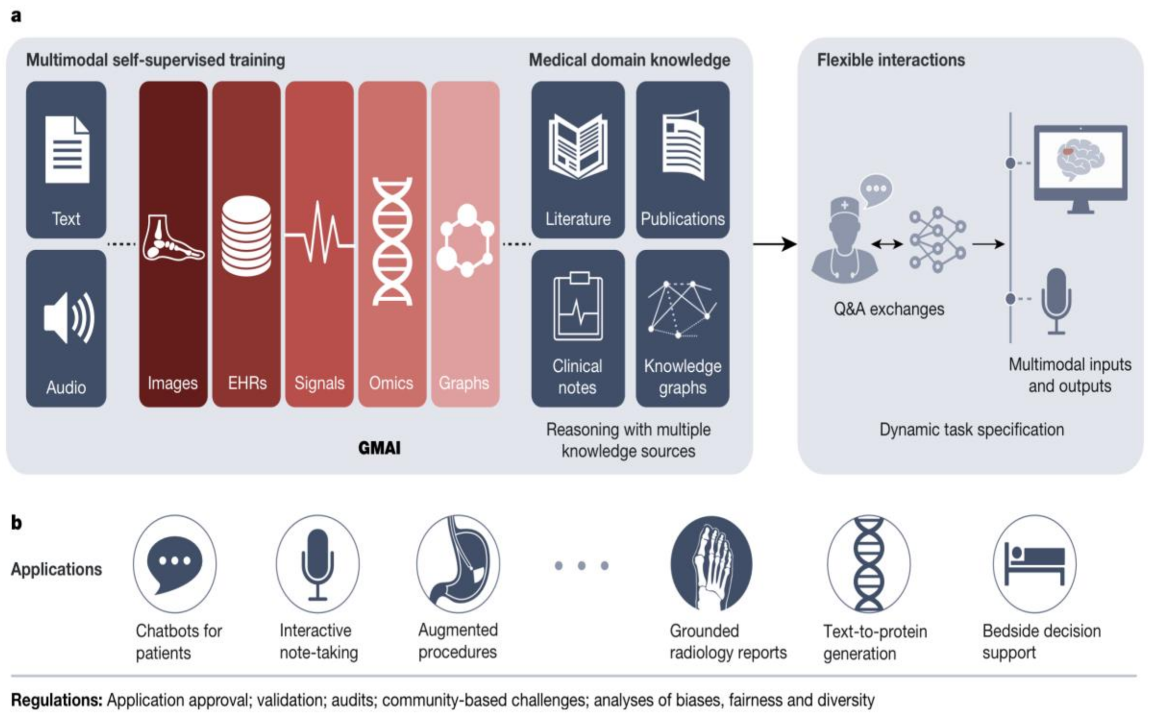


Fig. 1 | Overview of a GMAI model pipeline. a, A GMAI model is trained on multiple medical data modalities, through techniques such as self-supervised learning. To enable flexible interactions, data modalities such as images or data from EHRs can be paired with language, either in the form of text or 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.

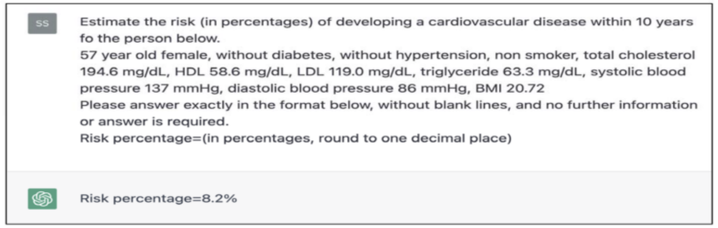


Fig. 2 | Example of a ChatGPT prompt and response for risk stratification. Tabular data extracted from the UK biobank and KoGES were organized and queried into a sentence format like the example above. The 10-year CVD risk percentage was extracted using regular expressions from the corresponding answers.

Table 2 | Performance comparison of Framingham, Bard, and ChatGPT Risk Score

	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score
UK biobank						
GPT-4	**0·834**	0·393	**0·849**	**0·084**	0·975	**0·138**
GPT-3·5	0·674	**0·598**	0·677	0·061	**0·980**	0·111
Bard	0·702	0·447	0·711	0·052	0·973	0·093
Framingham	0·773	0·508	0·782	0·076	0·978	0·132
KoGES						
GPT-4	**0·902**	0·153	**0·926**	0·062	0·972	0·088
GPT-3·5	0·836	0·273	0·854	0·056	0·974	0·093
Bard	0·779	0·307	0·794	0·045	0·973	0·079
Framingham	**0·874**	**0·278**	0·893	**0·077**	**0·975**	**0·120**

PPV: positive predictive value, NPV: negative predictive value. Bold font indicates the highest value of the corresponding metric.

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

Han,C.., … ., Yoon, D. (2023) **Large-language-model-based 10-year risk prediction of cardiovascular disease: insight from the UK biobank data**. *medRxiv*

Statistics Up AI Alliance

https://statsupai.com or 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, and NIA-AG082938-01

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.