

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.



CONTENTS



Part I

Statistical Modeling: The Two Cultures



Part II

Opportunities for Statisticians



Part I

Statistical Modeling: The Two Cultures

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

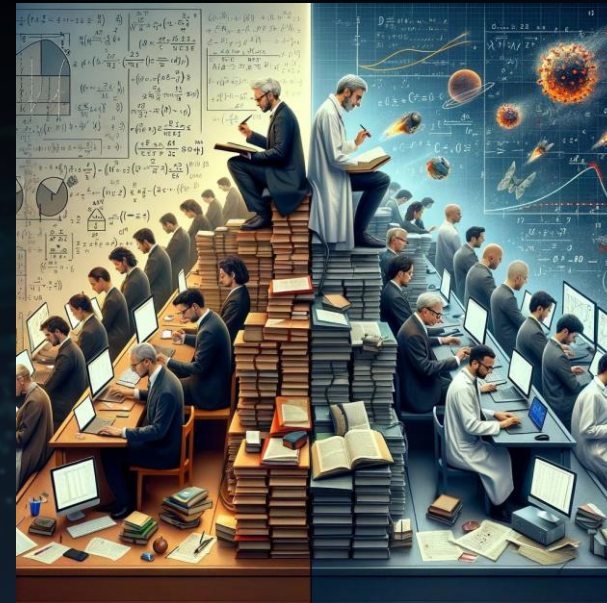
- John Tukey -

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



screen
esti: television

television
esti: television



screen
esti: television

television
esti: television



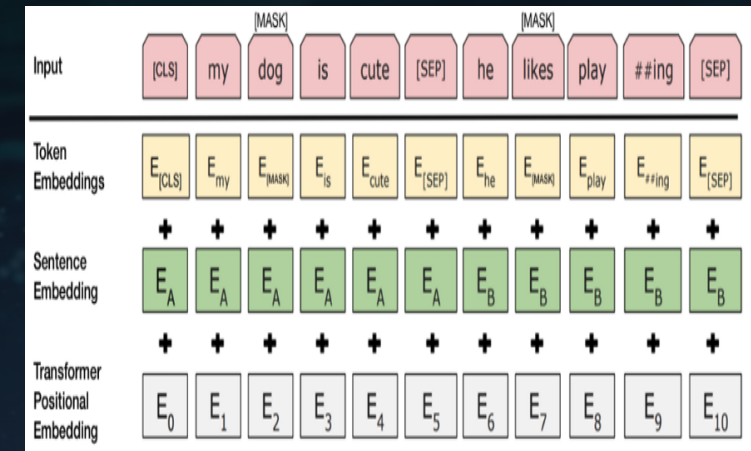
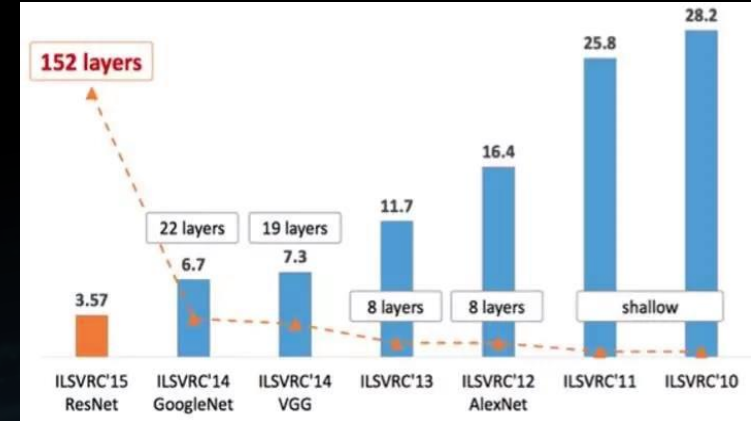
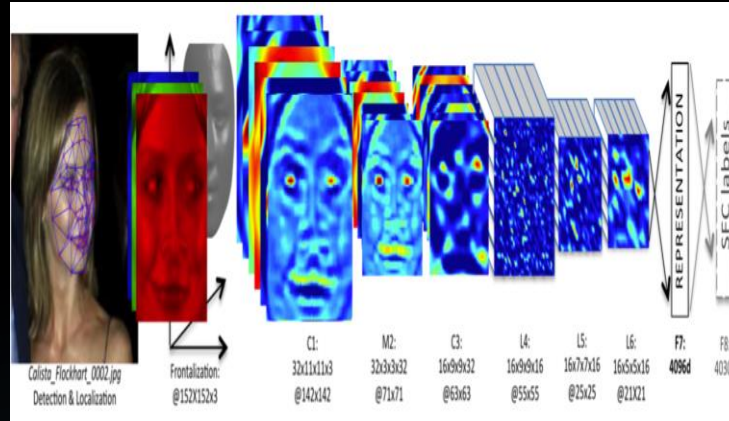
hair spray
esti: hair spray

hair spray
esti: web site



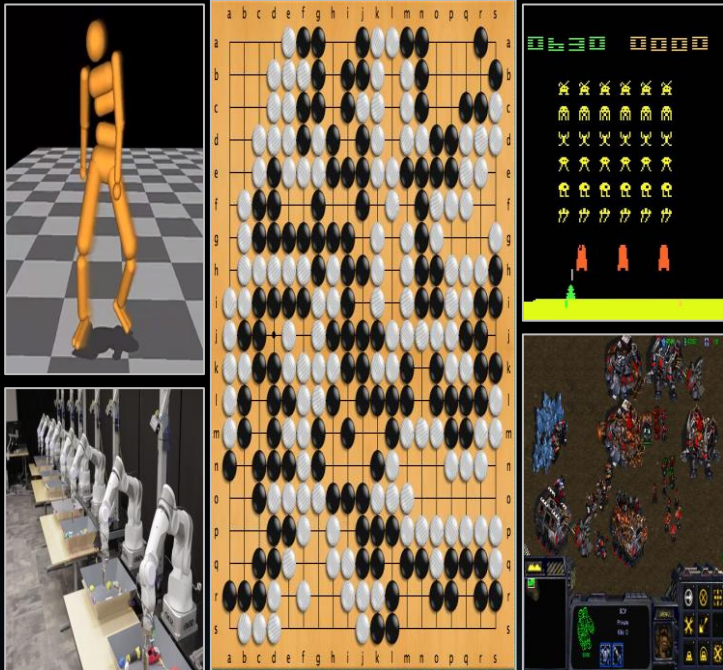
hair spray
esti: perfume

hair spray
esti: lighter



AI Milestones

Reinforcement Learning



AI Products

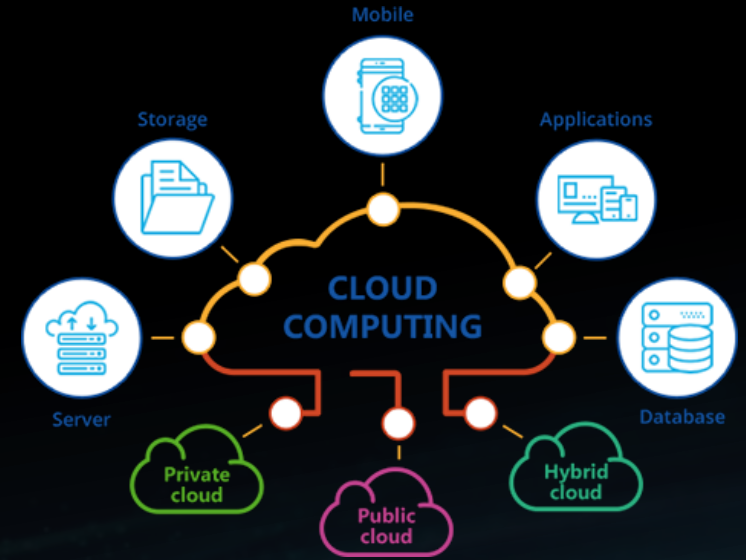
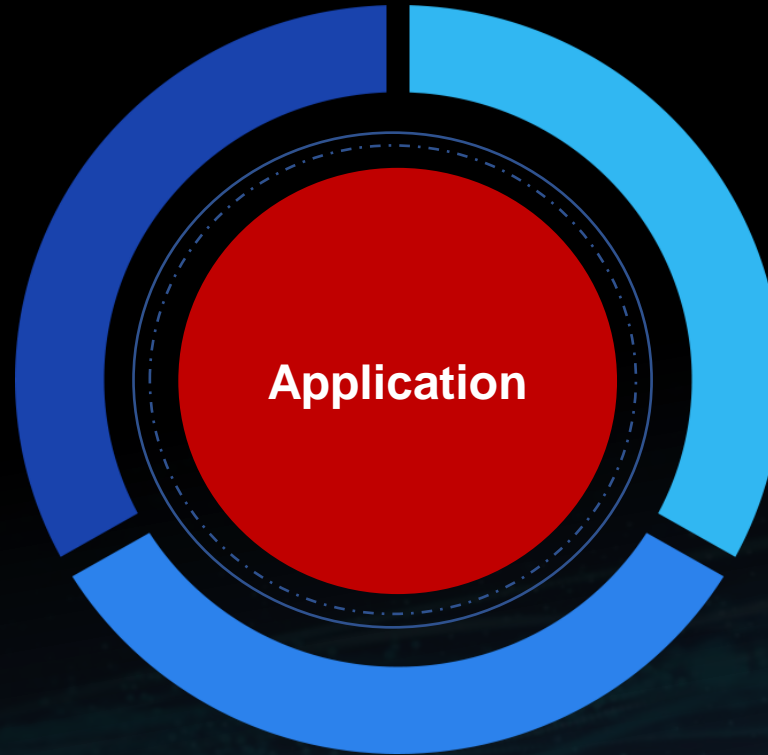


AI=Application to ABC



Big Data

<http://medium.com>



Computing

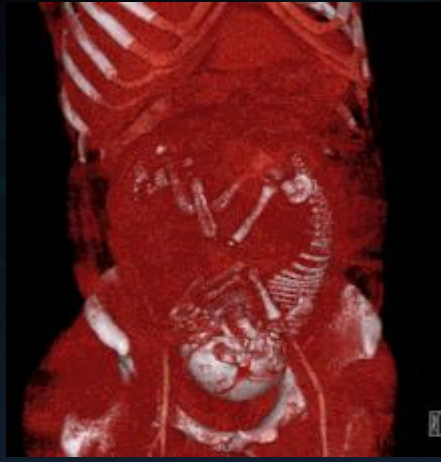
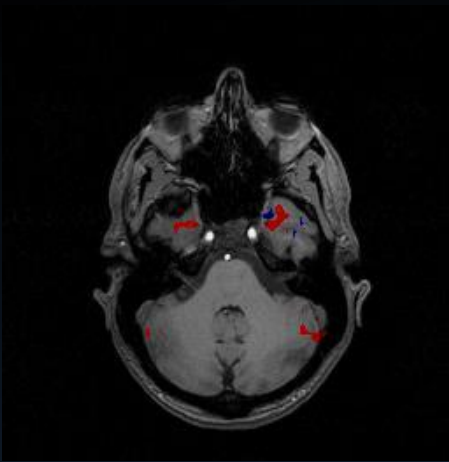
Analytical Tools

- 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/>)

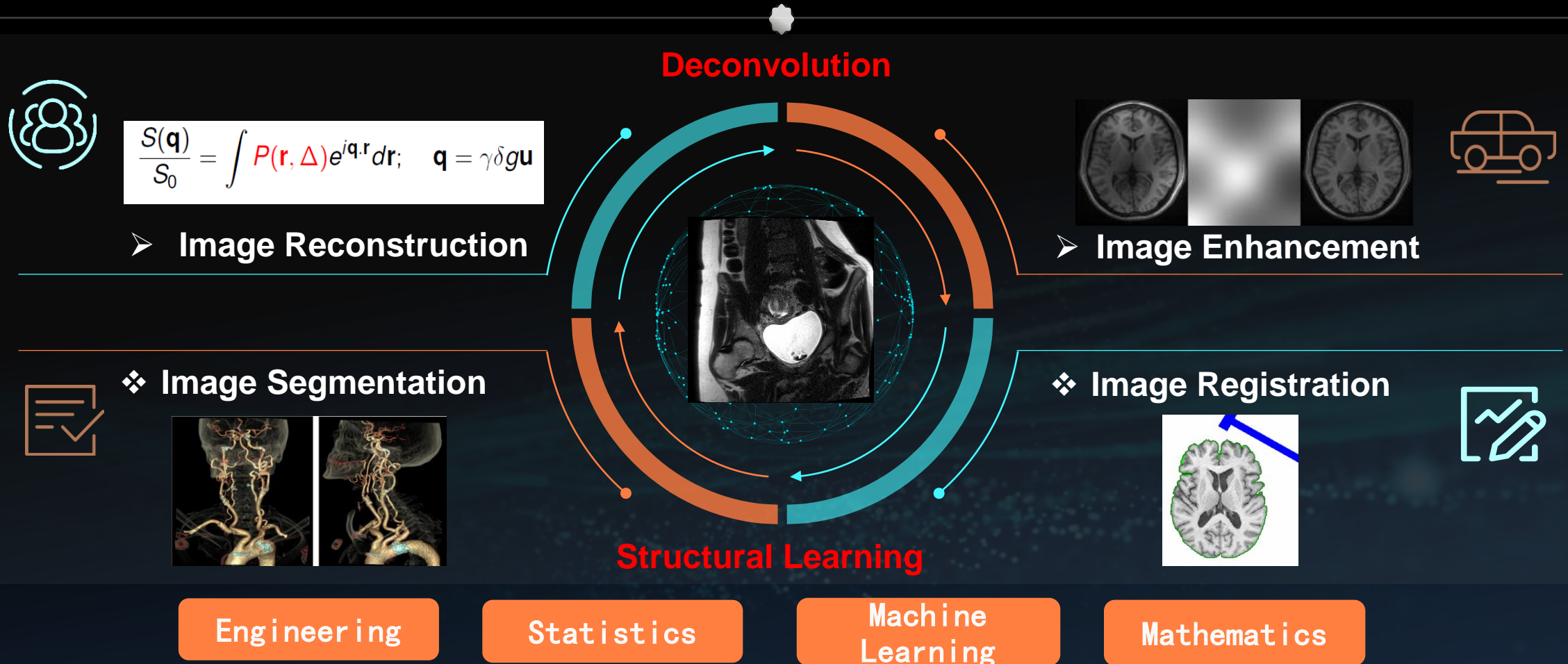
□ These imaging methods are essential for delineating the **structure and functionality of organs and tissues**. Each modality employs a distinct targeting agent, generates data in varying dimensions, extracts unique features, and serves specific purposes within clinical and research contexts.



- 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?

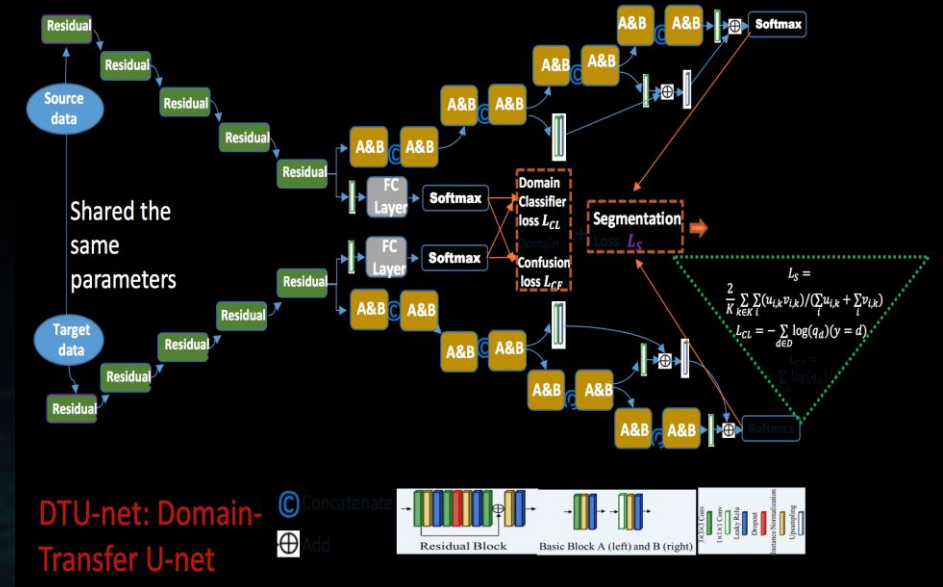


AI for Image Segmentation

Segmentation Annotation



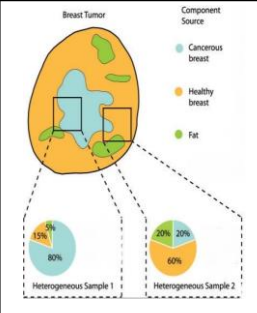
U-Nets



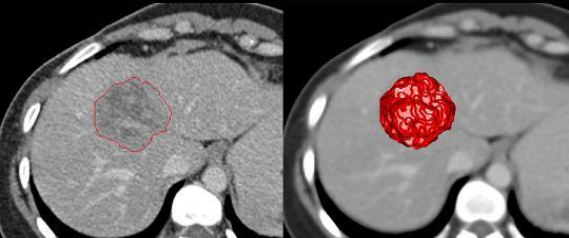
Liu, Q., Xu, Z., Bertasius, G., & Niethammer, M. (2023). SimpleClick: Interactive Image Segmentation with Simple Vision Transformers. ICCV., 22290-22300. 2023.

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.

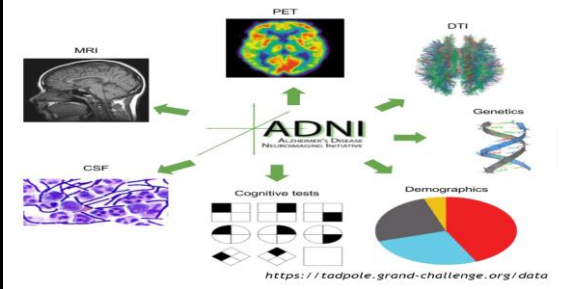
Ecological Layout for Imaging-based Analysis



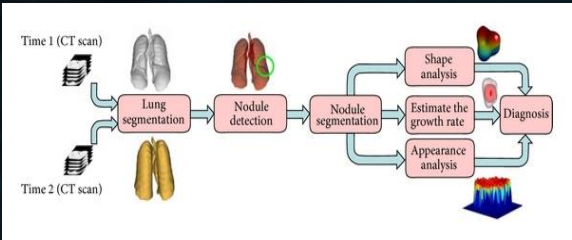
Deconvolution



Structural Learning



Integration



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

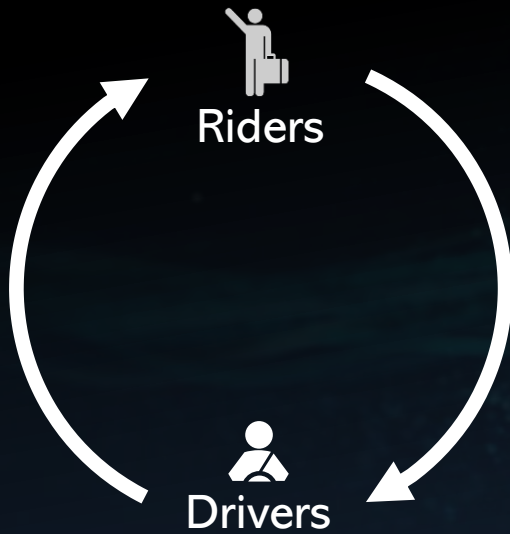


Uncertainty

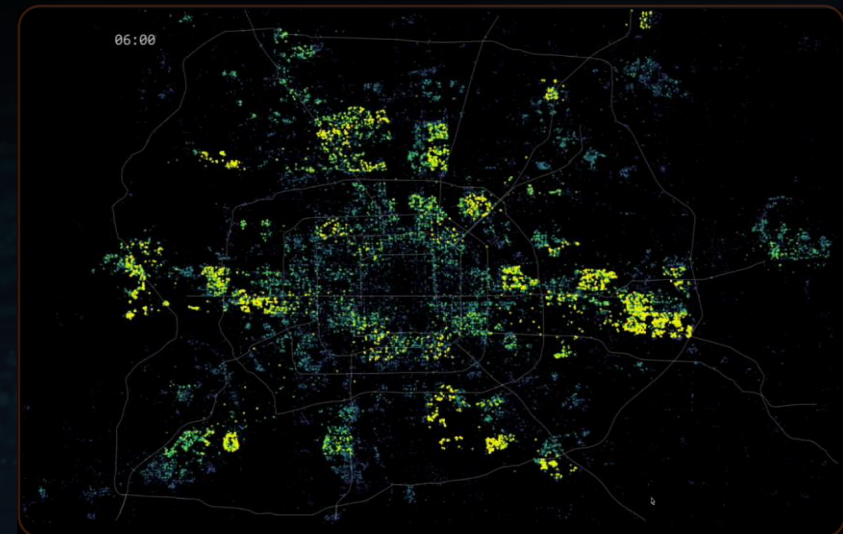


Causal

Two-sided Platform

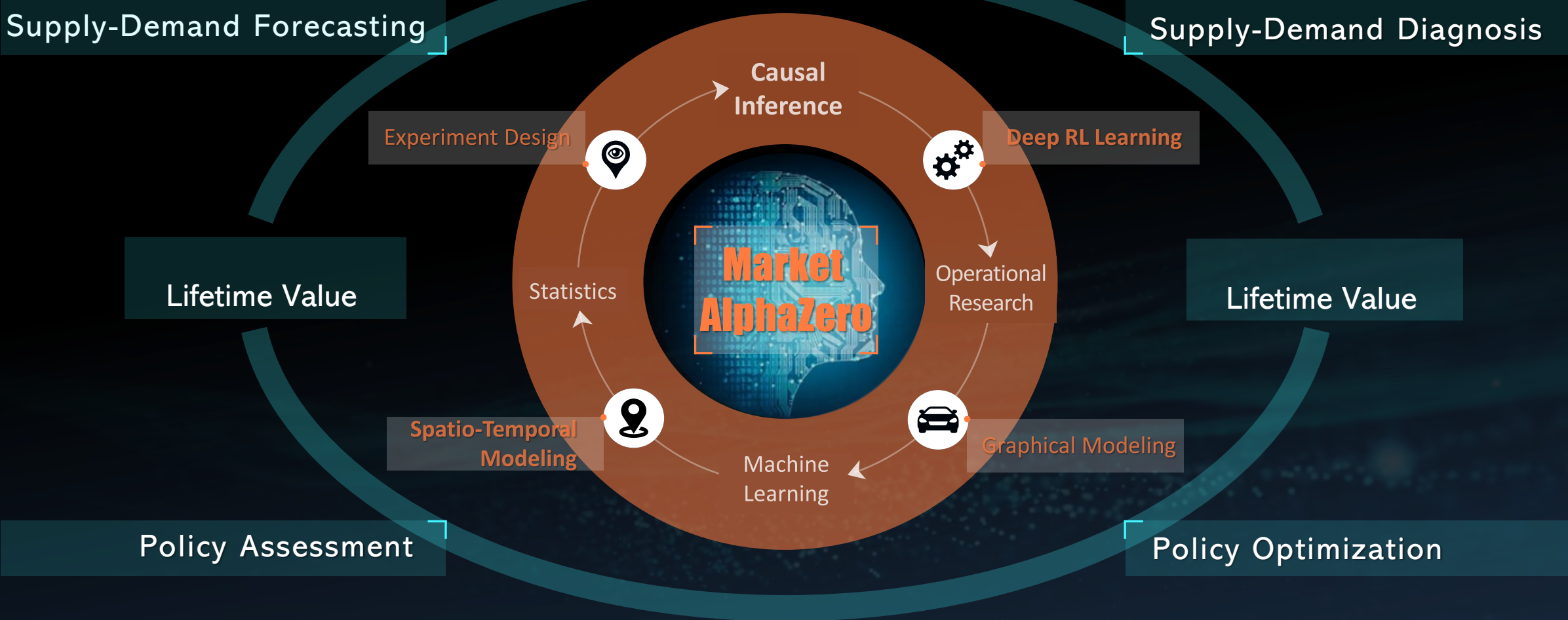


Complex Spatio-temporal System



Leverage Supply-Demand Network Effect

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



Supply-Demand Forecasting

The Problem



The Goal

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

Improve the service quality

Drivers



- Reduce empty driving

Riders



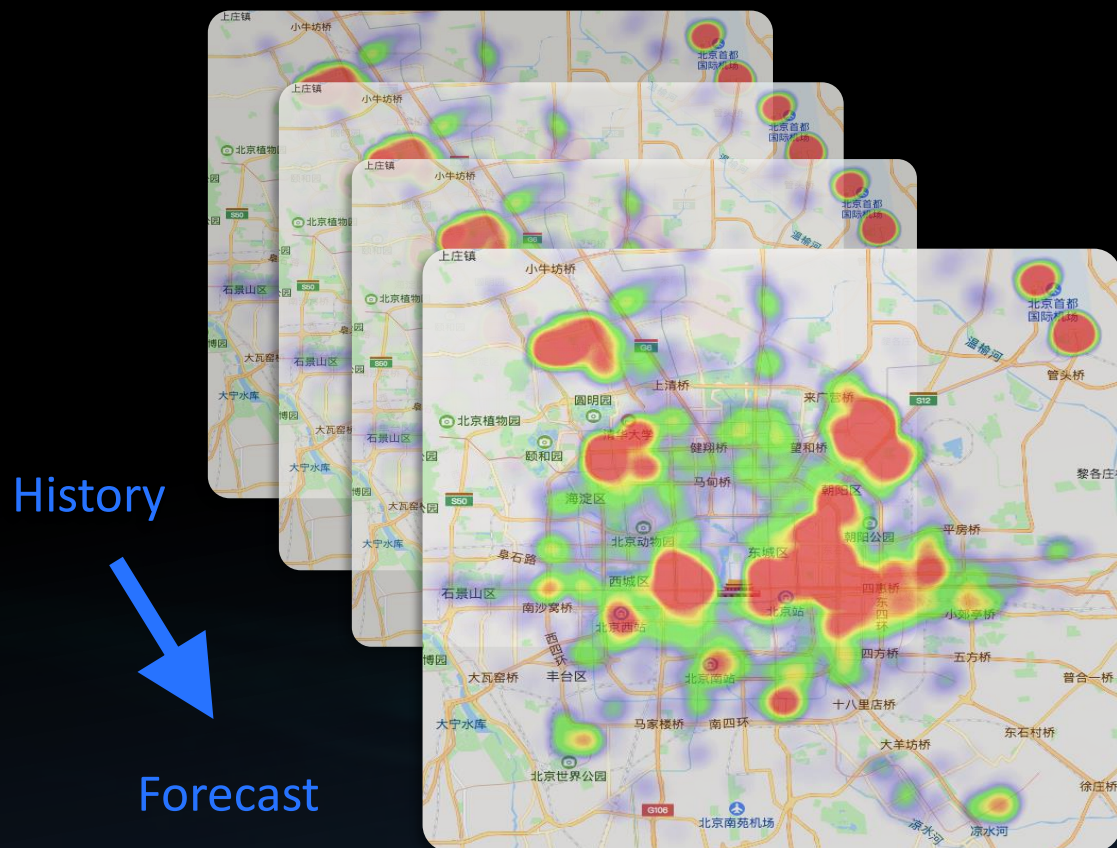
- Intelligent travel guidance
- Less queueing time

Platform

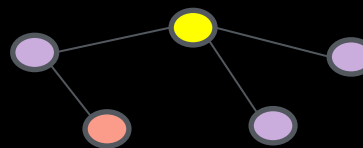


- Fill demand-supply gap
- Recognize the market
- Better dispatching and scheduling

A Deep S-T Forecasting Model



Graph Generation



Incorporating multi-relationship

Temporal: CGRNN

Contextual-gated temporal modeling



Spatio-MGCN



Non-euclidean spatial modeling

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: [f](#) [in](#) [t](#) [e](#)

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.



Synthesis Lectures on Learning, Networks, and Algorithms

Synthesis Lectures on Learning, Networks, and Algorithms

SYNTHESIS COLLECTION OF TECHNOLOGY

Series Editor: Lei Ying

Zhiwei (Tony) Qin · Xiaocheng Tang · Qingyang Li · Hongtu Zhu · Jieping Ye

Reinforcement Learning in the Ridesharing Marketplace

This book provides a comprehensive overview of reinforcement learning for ridesharing applications. The authors first lay out the fundamentals of the ridesharing system architectures and review the basics of reinforcement learning, including the major applicable algorithms. The book describes the research problems associated with the various aspects of a ridesharing system and discusses the existing reinforcement learning approaches for solving them. The authors survey the existing research on each problem, and then examine specific case studies. The book also includes a review of two of methods closely related to reinforcement learning: approximate dynamic programming and model-predictive control.

In addition, this book:

- Explains the benefits of taking a reinforcement learning approach to ridesharing optimization problems
- Analyzes a number of specific works that cover the optimization of ridesharing platforms using reinforcement learning
- Highlights the major challenges and opportunities that are crucial for advancing reinforcement learning for ridesharing

About the Authors

Zhiwei (Tony) Qin, Ph.D., is a Principal Scientist at Lyft Rideshare Labs.

Xiaocheng Tang, Ph.D., is an AI Research Scientist at Meta.

Qingyang Li, Ph.D., is a Senior Engineering Manager at Didi Autonomous Driving.

Jieping Ye, Ph.D. is affiliated with the Alibaba Group.

Hongtu Zhu, Ph.D. is a Professor in the Department of Biostatistics at The University of North Carolina at Chapel Hill.



springer.com

Qin · Tang · Li · Zhu · Ye



Reinforcement Learning in the Ridesharing Marketplace

Zhiwei (Tony) Qin · Xiaocheng Tang · Qingyang Li · Hongtu Zhu · Jieping Ye

Reinforcement Learning in the Ridesharing Marketplace

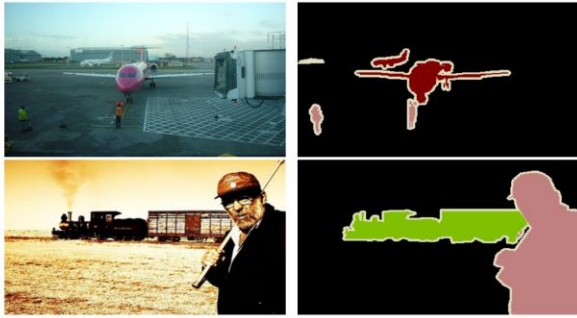
Springer

Generative Model

- Image data $X \in \mathcal{X}(\Omega, \mathbb{R}^3)$ on grid domain Ω . Given intensity $\lambda : \Omega \rightarrow \mathbb{R}^+$,

$$X = X_{obj}^1(u_1) \oplus \dots \oplus X_{obj}^K(u_K) \oplus \epsilon$$

- Objects: $X_{obj}^1, \dots, X_{obj}^K$; Background noise: ϵ .
- Number of objects: $K \sim \text{Poisson}(\Lambda(\Omega)), \Lambda(\Omega) = \int_{\Omega} \lambda(t) dt$
- Locations of objects: $u_k \sim P_{\lambda}, k = 1, \dots, K$.



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

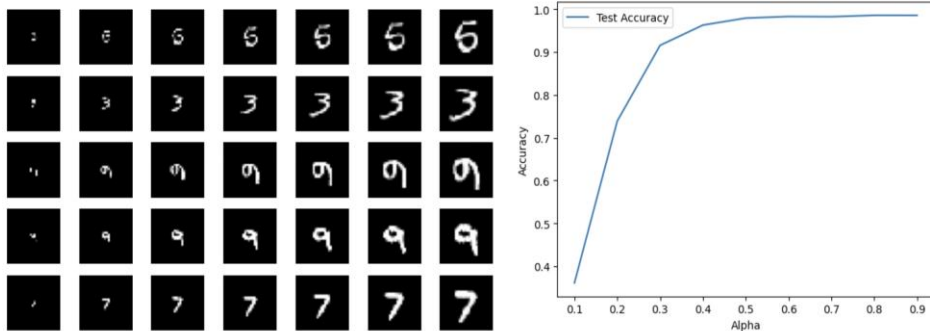


Figure: Test accuracy on MNIST dataset trained on a three-layer CNN.

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: $\lambda(\cdot) : \Omega \rightarrow \mathbb{R}^+$

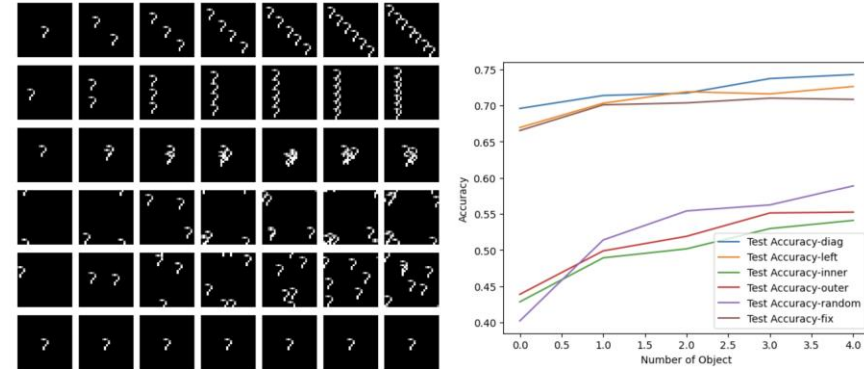


Figure: Test accuracy on MNIST dataset ($\alpha = 0.2$) trained on a three-layer CNN.

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

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

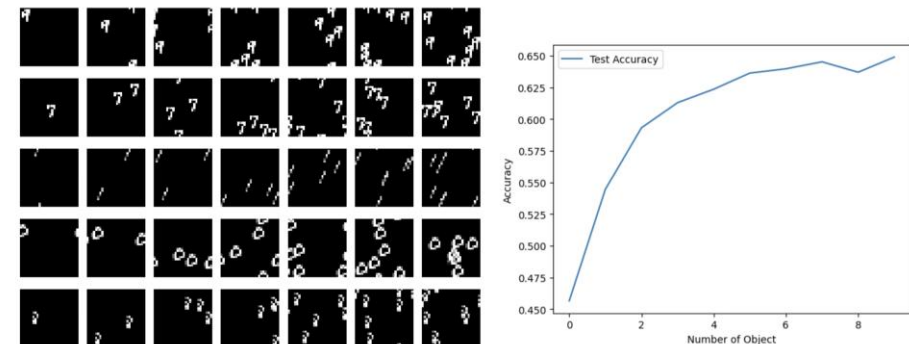
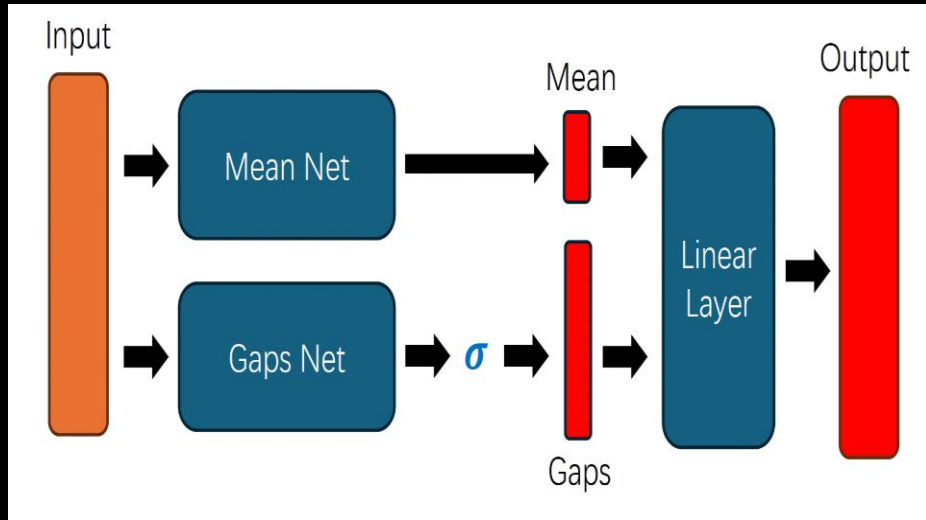


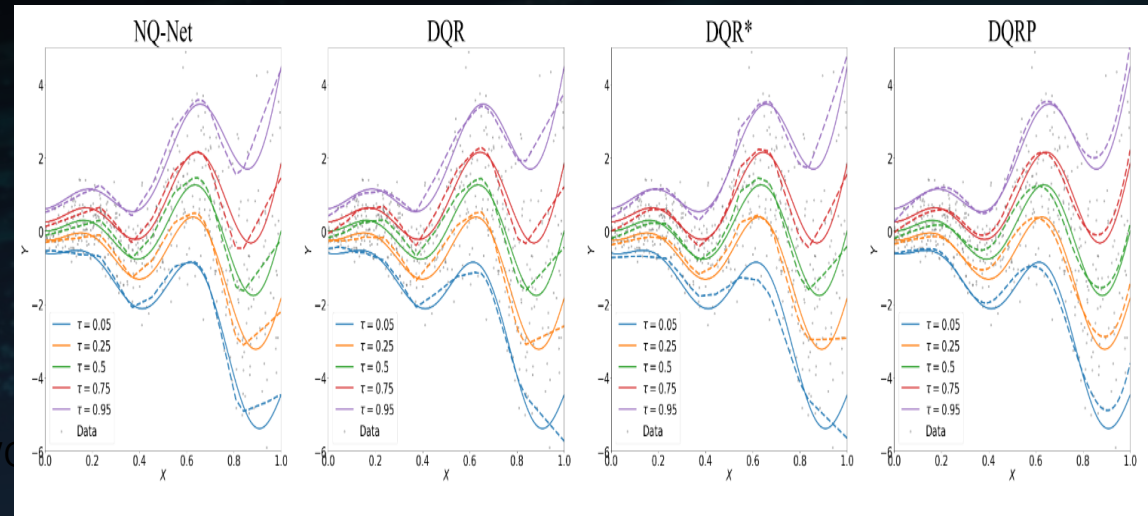
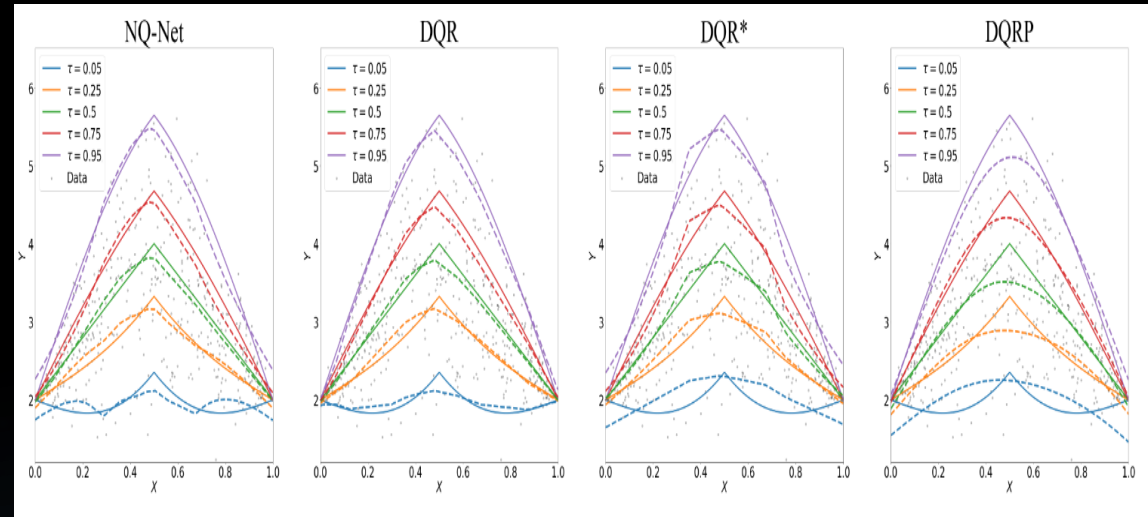
Figure: Test accuracy on MNIST dataset ($\alpha = 0.2$) trained on a three-layer CNN.

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

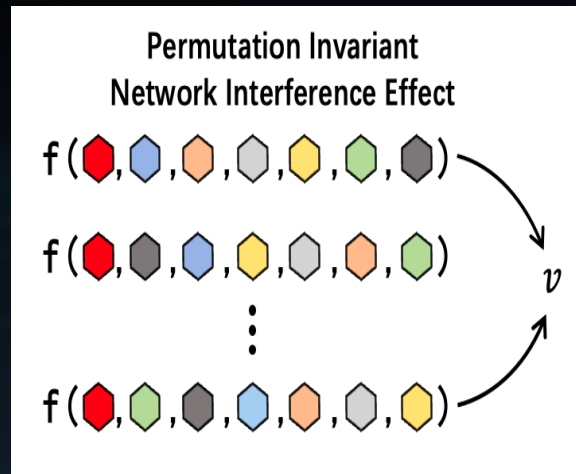
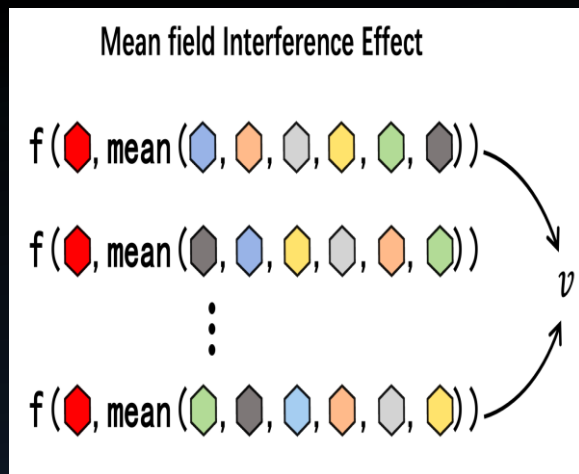
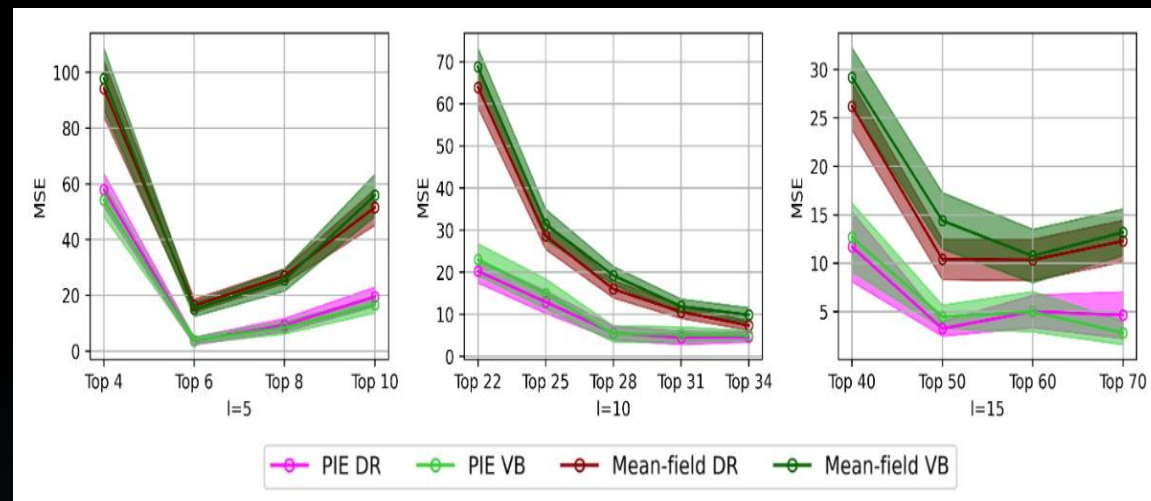
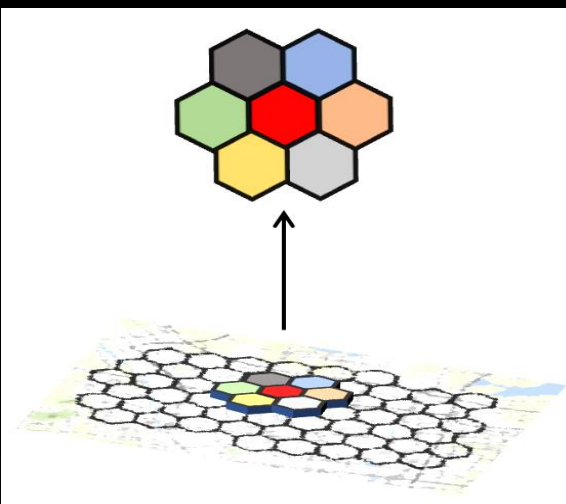
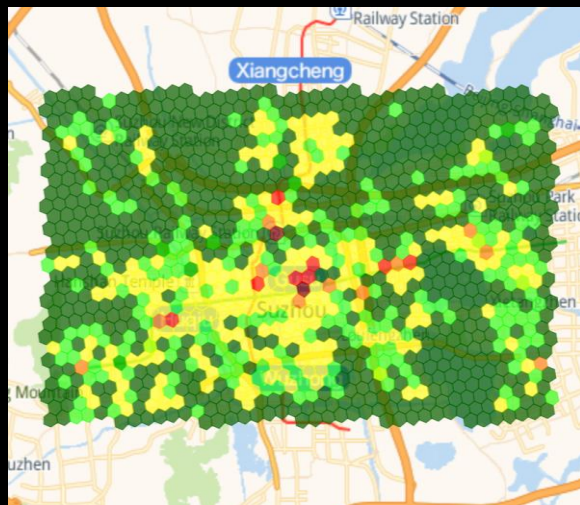


Non-crossing Quantile Network

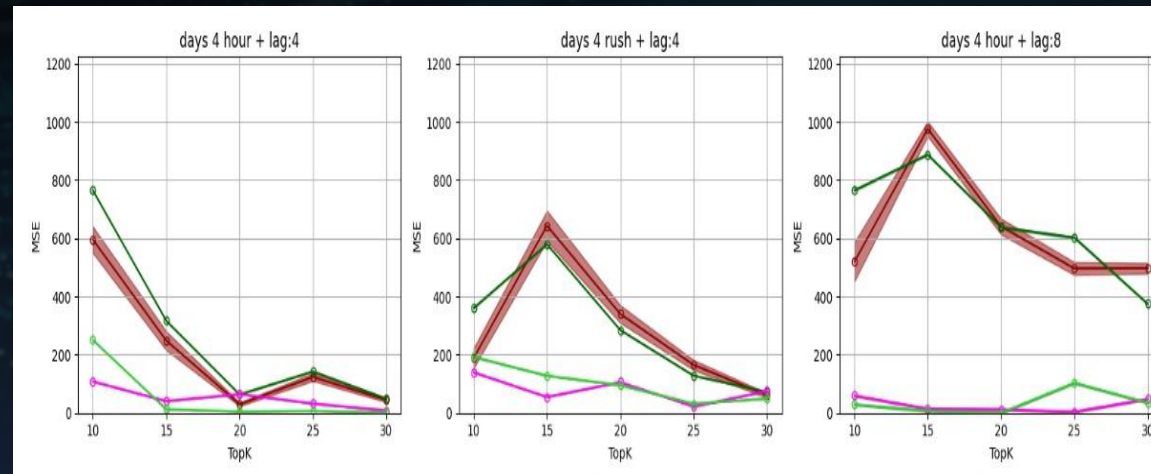


NQ network
structure

Causal Deepset for Offline Policy Learning

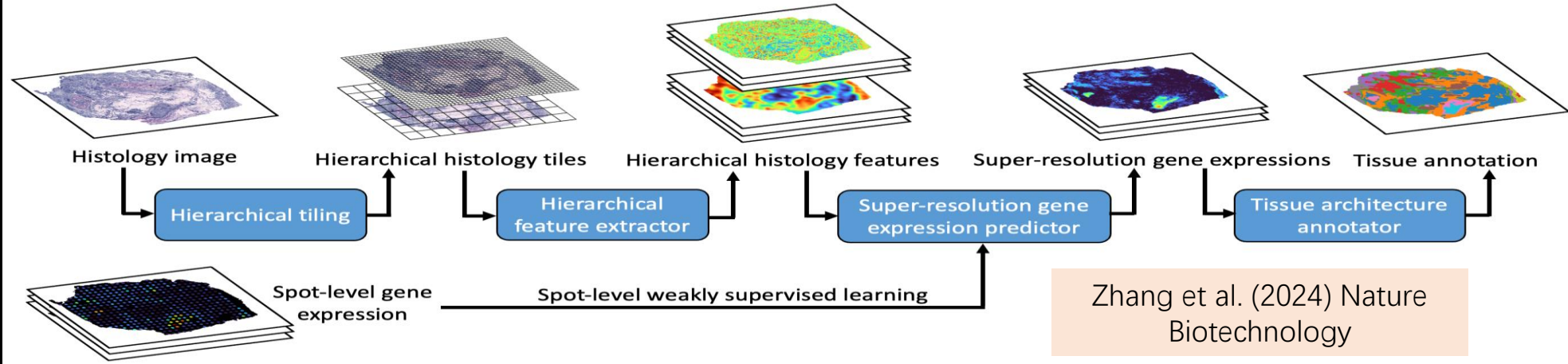


structure

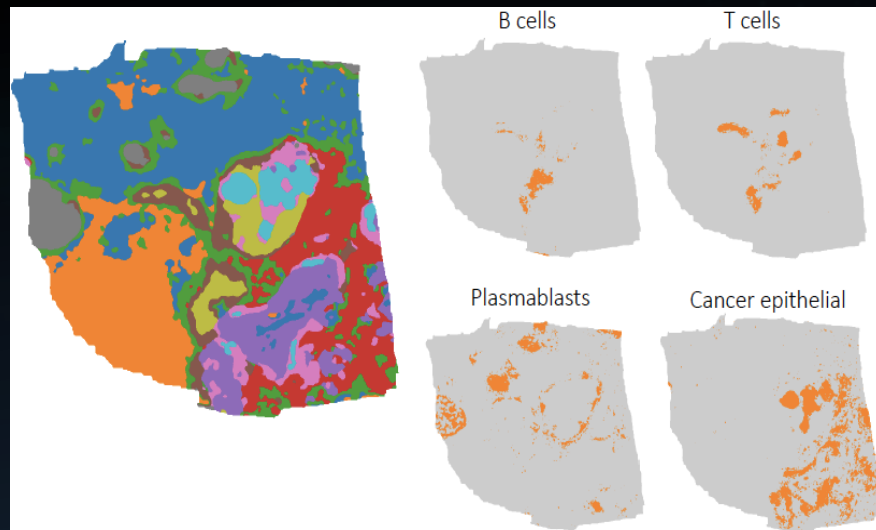


IStar

iStar (Inferring Super-Resolution Tissue Architecture)



iStar can automatically annotate cell types



iStar can automatically detect positive surgical margin

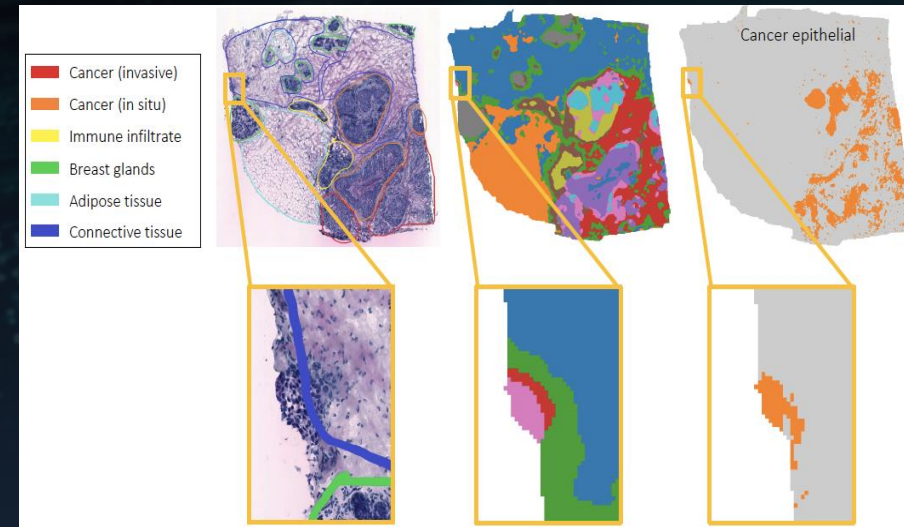
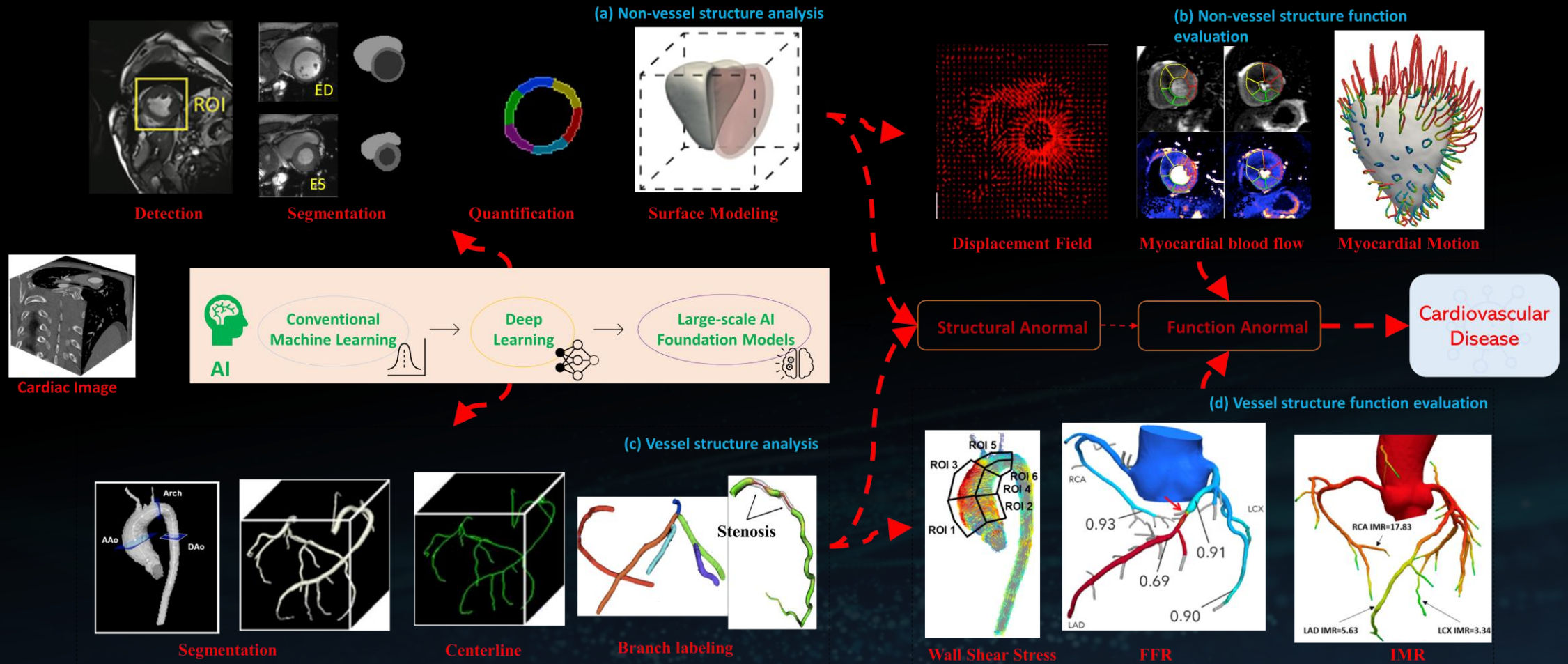
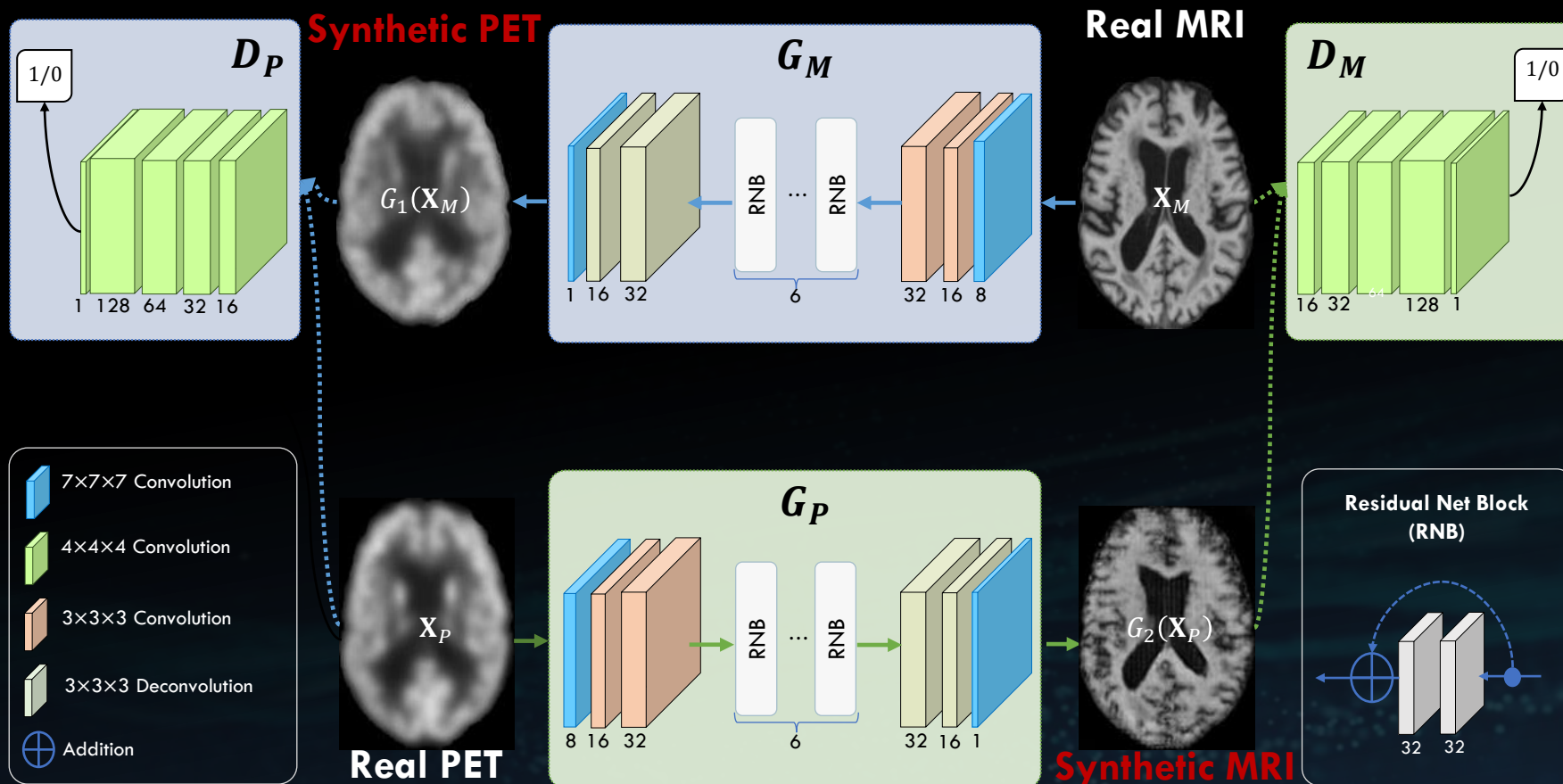


Image Analysis Pipeline

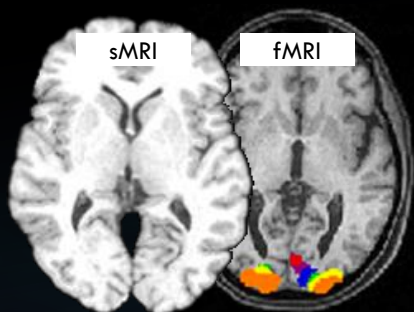
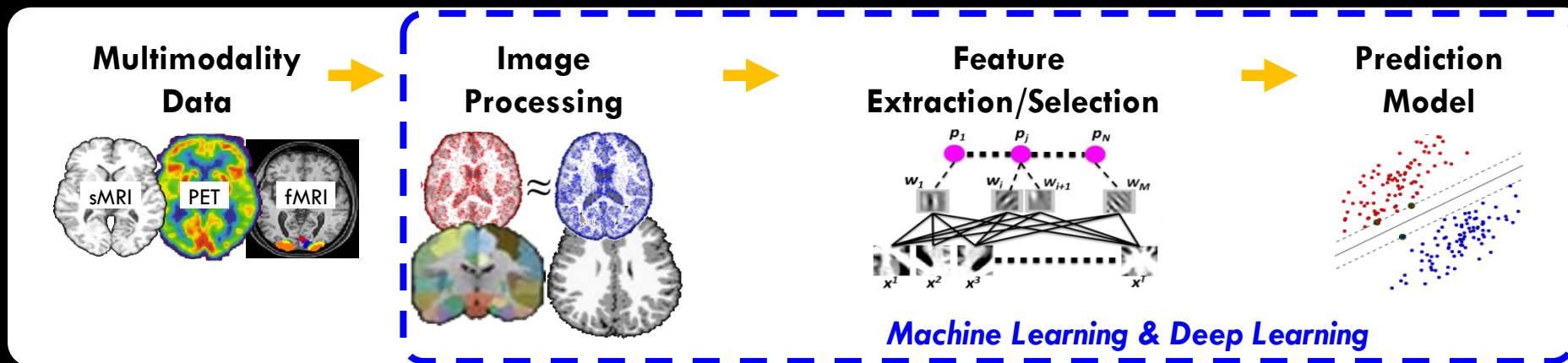


Wang, X. and Zhu, H (2024). Artificial Intelligence in Image-based Cardiovascular Disease Analysis: A Comprehensive Survey and Future Outlook

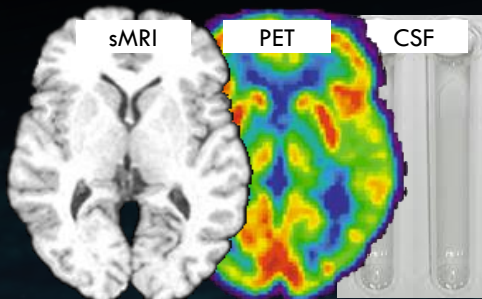
Cross-Modality Image Synthesis



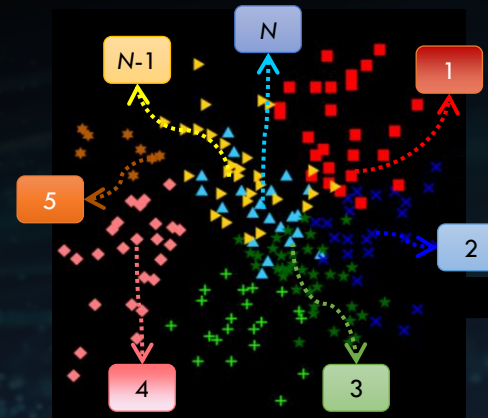
Computer-Aided Medical Data Analysis



Neuroimage
Representation Learning

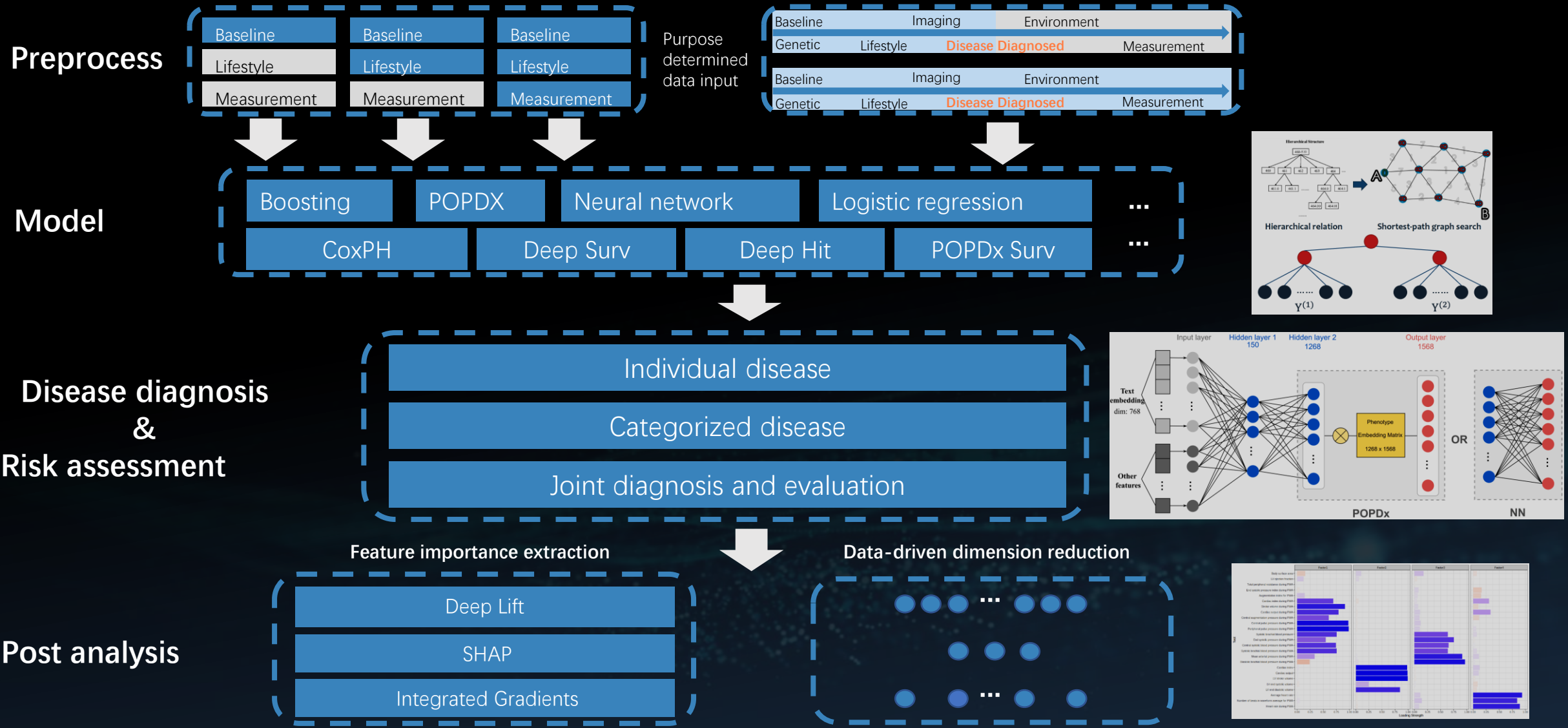


Multimodality
Data Fusion

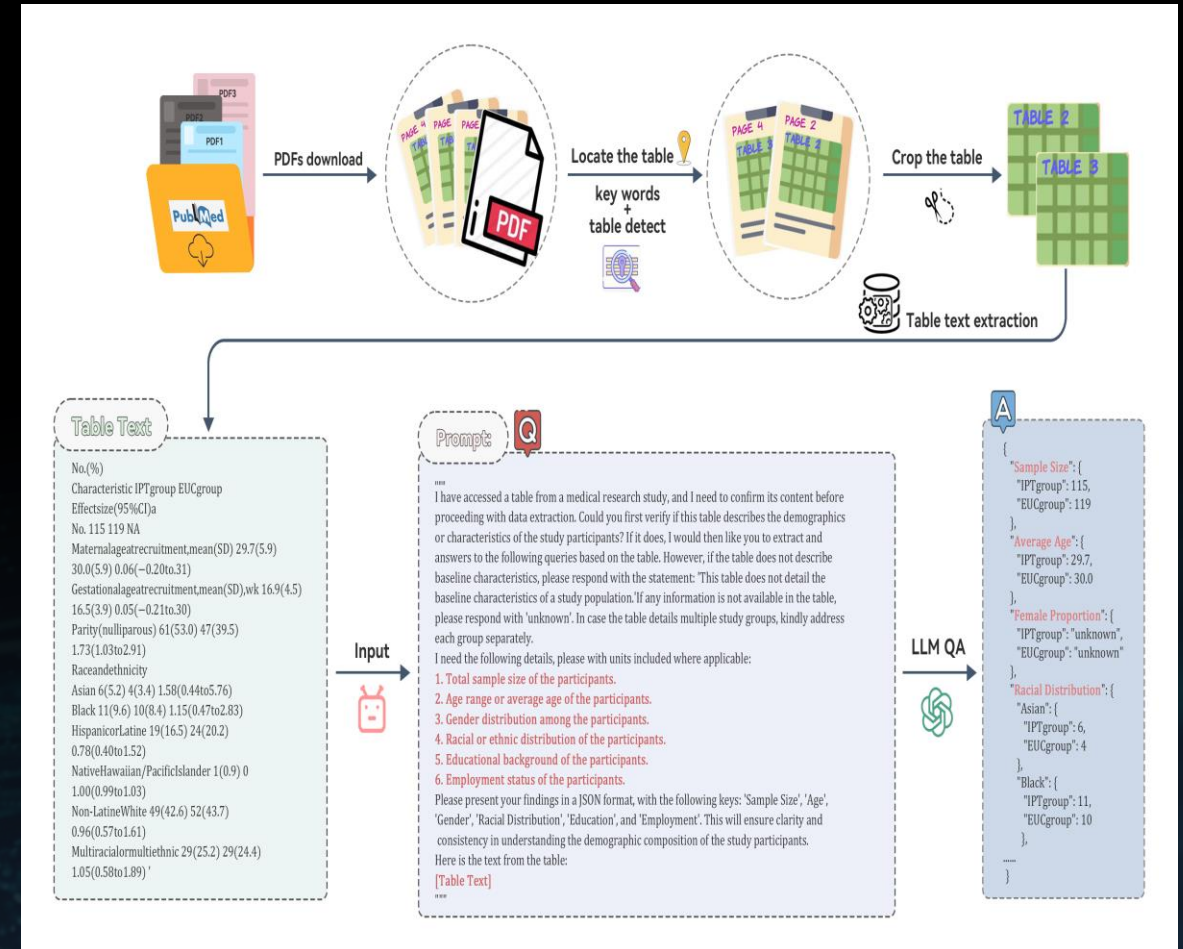
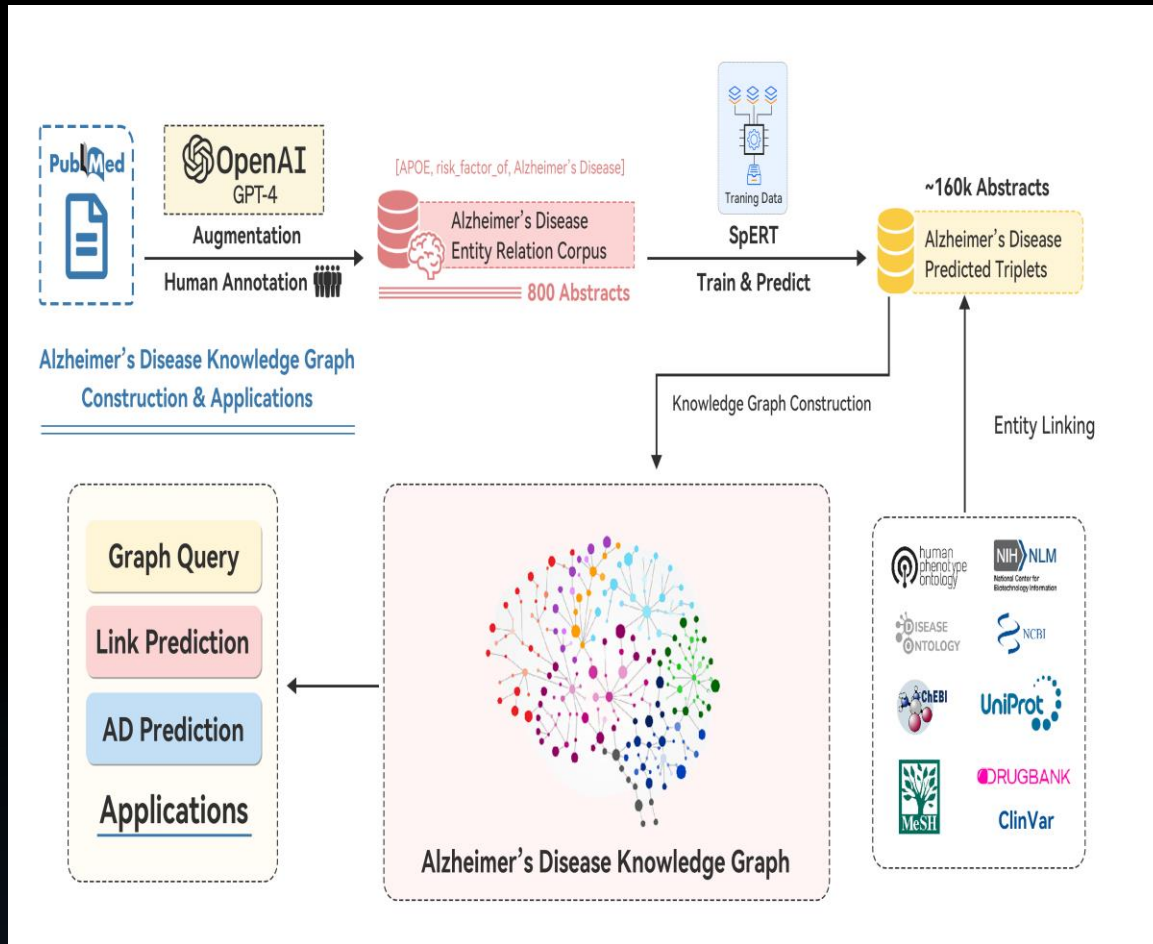


Multi-Site
Data Adaptation

Prediction Models



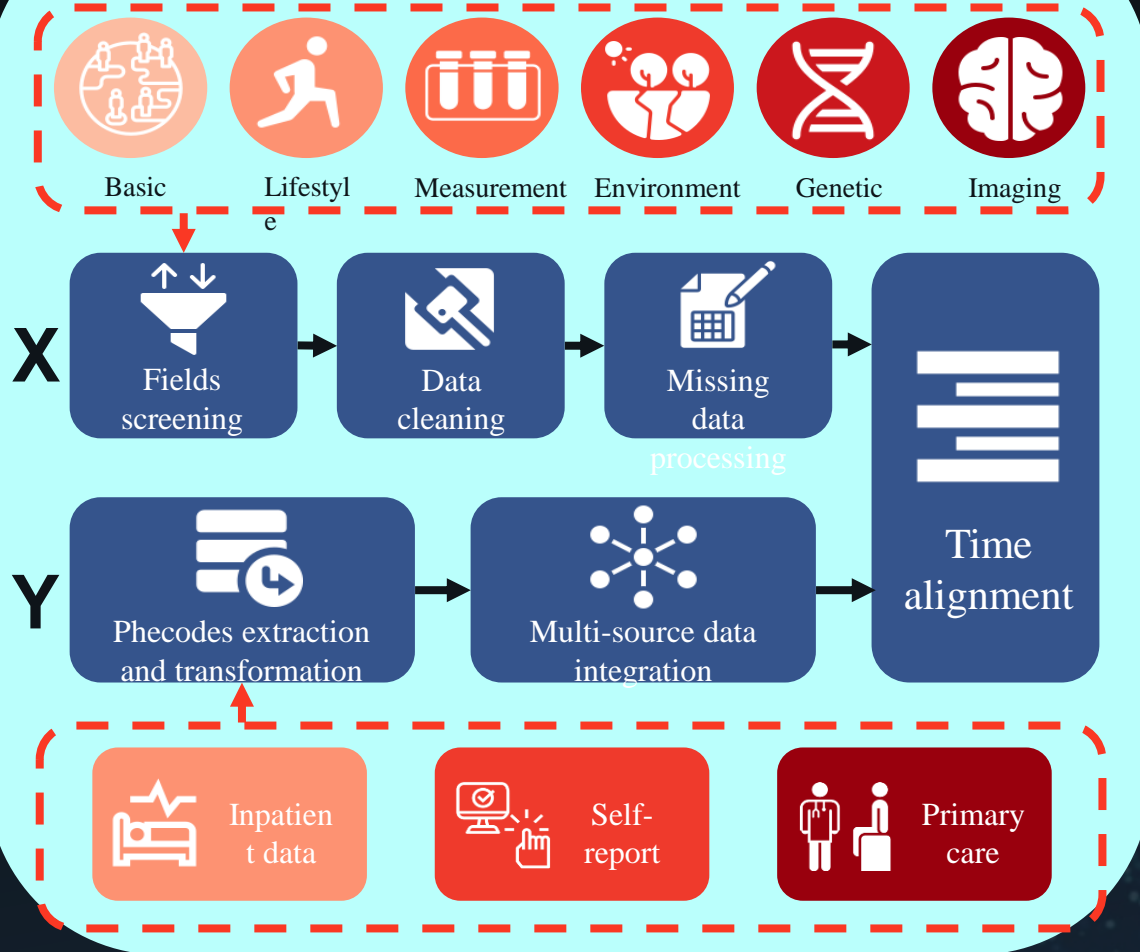
Knowledge Graph Construction



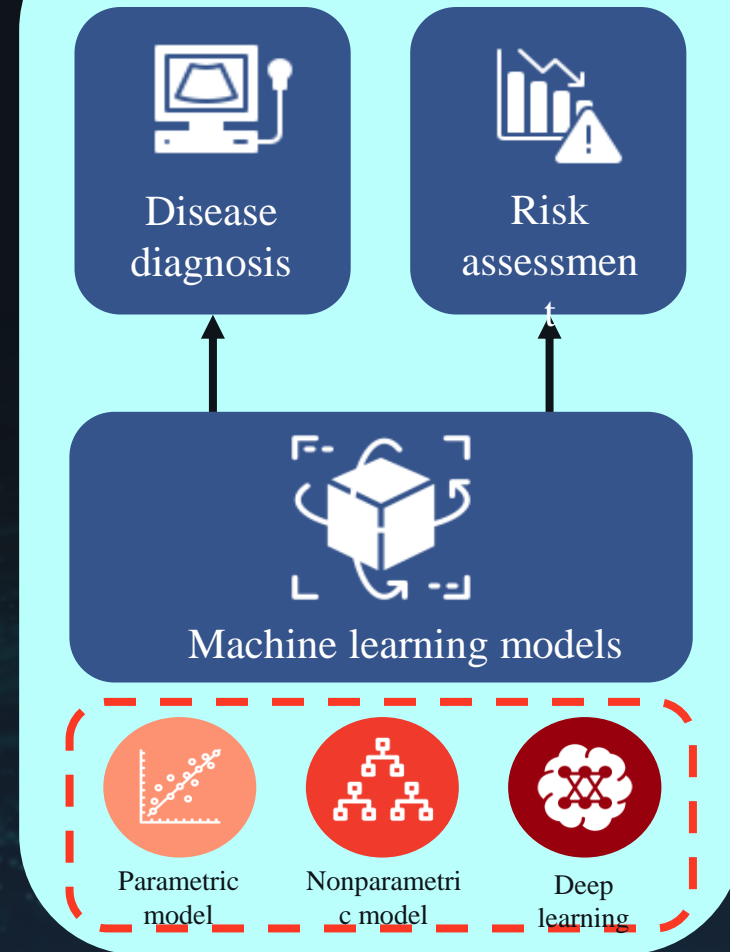
Yang et al., Alzheimer's Disease Knowledge Graph Enhances Knowledge Discovery and Disease Prediction.

Data Preprocessing and Data Modeling

Data Preprocessing



Data Modeling



Foundation Models for GMAI

Perspective

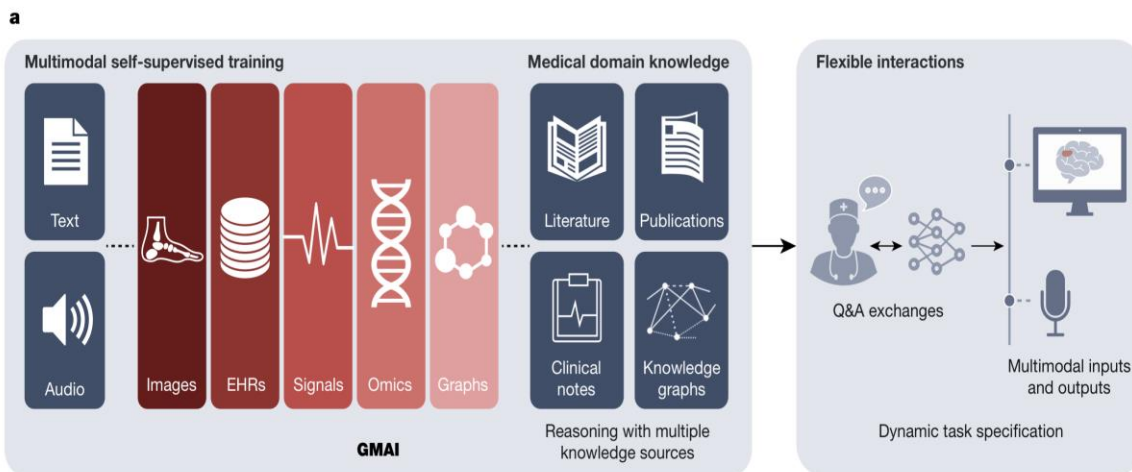


Fig. 1 | Overview of a GMAI model pipeline. **a**, A GMAI model is trained on multiple medical data modalities, through techniques such as self-supervised learning. To enable flexible interactions, data modalities such as images or data from EHRs can be paired with language, either in the form of text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

55 Estimate the risk (in percentages) of developing a cardiovascular disease within 10 years to the person below.
57 year old female, without diabetes, without hypertension, non smoker, total cholesterol 194.6 mg/dL, HDL 58.6 mg/dL, LDL 119.0 mg/dL, triglyceride 63.3 mg/dL, systolic blood pressure 137 mmHg, diastolic blood pressure 86 mmHg, BMI 20.72
Please answer exactly in the format below, without blank lines, and no further information or answer is required.
Risk percentage=(in percentages, round to one decimal place)

Risk percentage=8.2%

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.

medRxiv preprint doi: <https://doi.org/10.1101/2023.05.22.23288842>; this version posted May 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

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>



The screenshot shows the YouTube channel for Stats Up AI. The channel name is "Stats Up AI" with the handle "@StatsUpAI" and 17 subscribers. Below the channel name is a "订阅" (Subscribe) button. The video list includes:

- Part 3 -- Statistical Education in the Age of AI (43:38, 113 views)
- Part 2 -- Statistics, ML, and Data Science Journals in... (35:53, 139 views)
- Part 1 -- Statistical Theory & Methods, Applications and AI (48:45, 378 views)



STATISTICAL SCIENCE IN
ARTIFICIAL INTELLIGENCE
JASA SPECIAL ISSUE

SUBMIT BY
DEC 31, 2024

Information:
www.reallygreatsite.com

Identification of Core AI Problem
Statistical Contributions to AI
Innovative Statistical Theory,
Method and Applications

Acknowledgement



**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.