人工智能时代统计学的挑战和机遇：
一位统计工作者的思考

University of North Carolina at Chapel Hill

Hongtu Zhu

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

CONTENTS

Part I

统计学面临八大挑战

“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.”

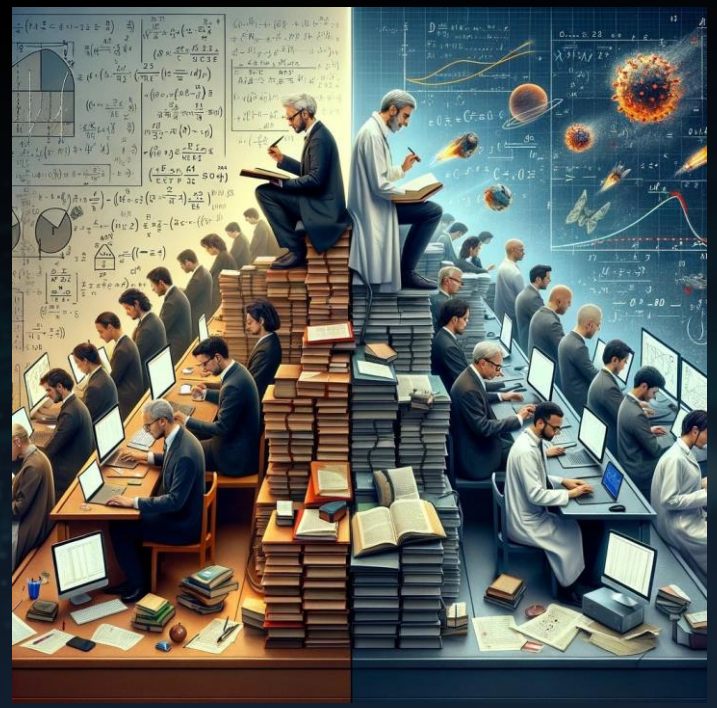
- Leo Breiman -

统计学的挑战一

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."*



统计学的挑战一

Breiman教授的话

统计就是一门收集、分类、处理并且分析事实和数据的科学。

Fisher相信统计的存在是为了预测、解释和处理数据的。

就统计应用的角度而言，我知道工业机构和政府在发生些什么，
但是目前进行的学术研究却似乎离我们无比遥远，
好像只是抽象数学的某一分支一样。

统计学的核心是应用和数据，就是通过分析数据来深刻地探索这个世界，
并通过产品为人类服务。

统计学的挑战一

如何处理高复杂度的问题，现在的数据模型是一筹莫展的

在从数据到结论的过程中，有两种统计建模文化。第一种是数据模型，假设数据是通过给定的随机数据模型生成的。另一种是算法模型，将数据生成机制视为未知。一直以来，统计界几乎完全使用数据模型。这种情况造就了无关紧要的理论、有问题的结论，并使统计学家无法研究广阔、有趣的现实问题。算法建模，都在统计学之外的领域飞速发展。它既可以用于大型复杂的数据集，也可以用于小型数据集。

数据模型是包含与线性模型复杂度类似的所有统计模型， 然而算法模型可能是比线性模型复杂度大许多的所有模型。从模型的角度，它们之间的主要差异是模型的复杂度，然后才是在应用场景、计算力和理论的差别。

算法模型开启了一条通过大应用，到大数据，
再到新算法新的发展模式。

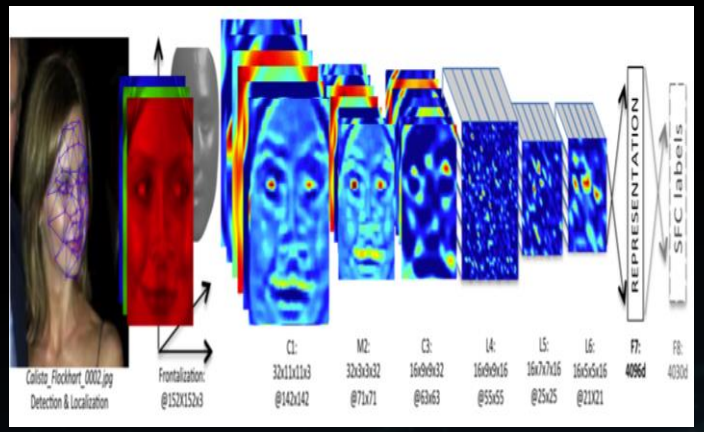
统计学的挑战二

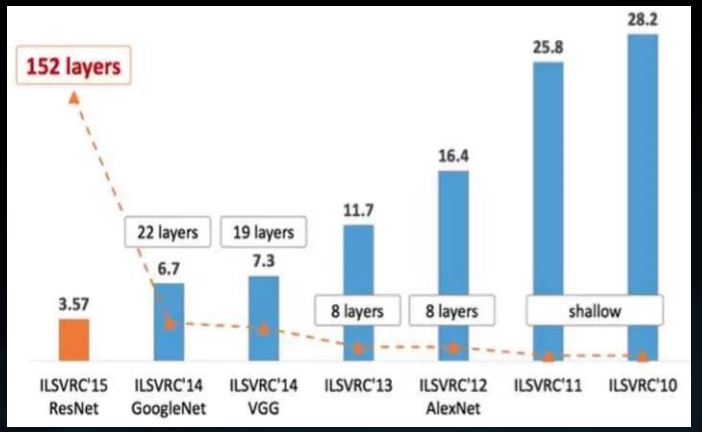
过度抽象化造成对数据和应用缺乏理解，特别是标注数据的重要性

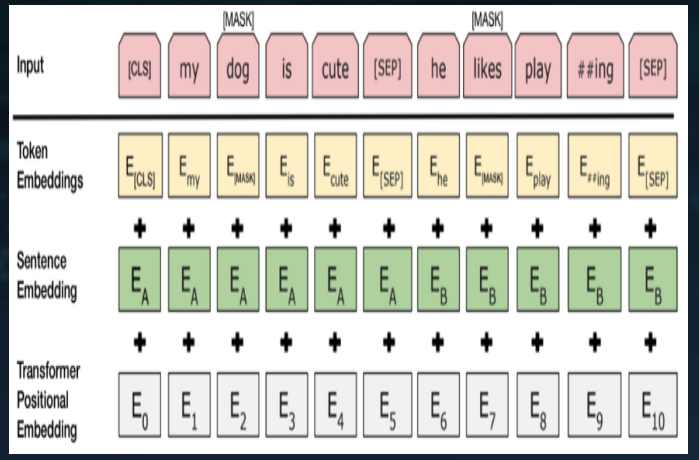
Algorithmic modeling =算法创新

Deep Learning

IMAGENET



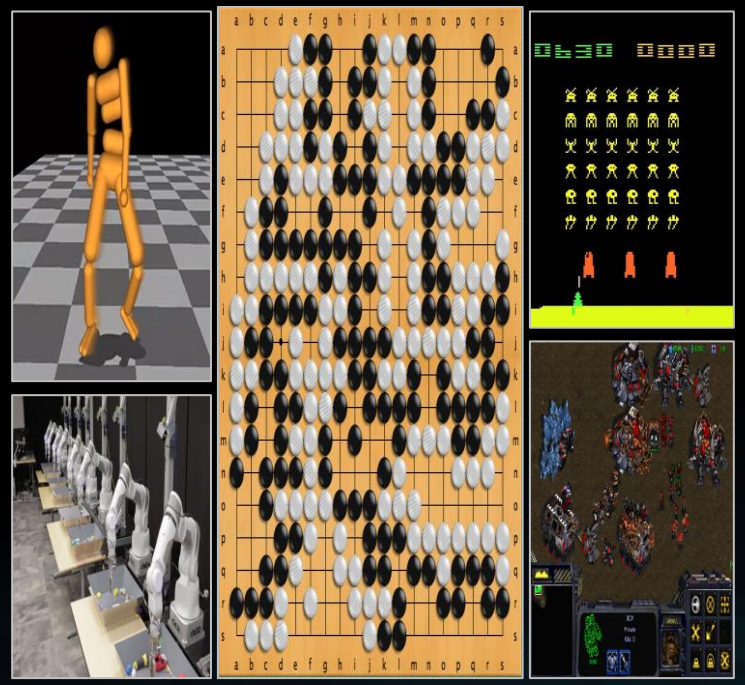




统计学的挑战三

算法模型的进一步创新

Deep Reinforcement Learning

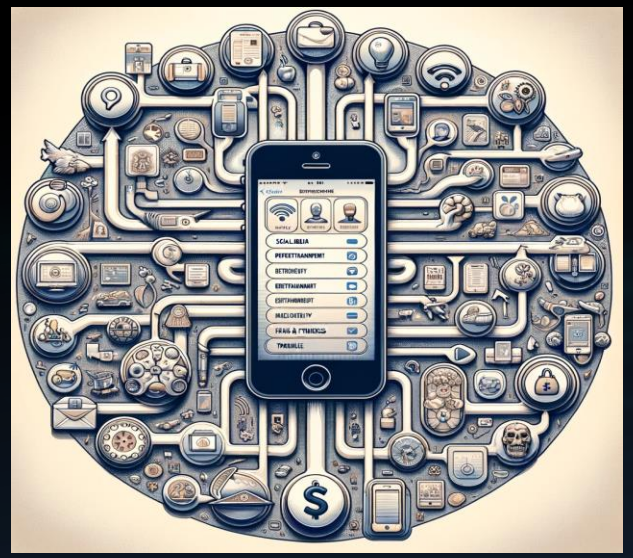


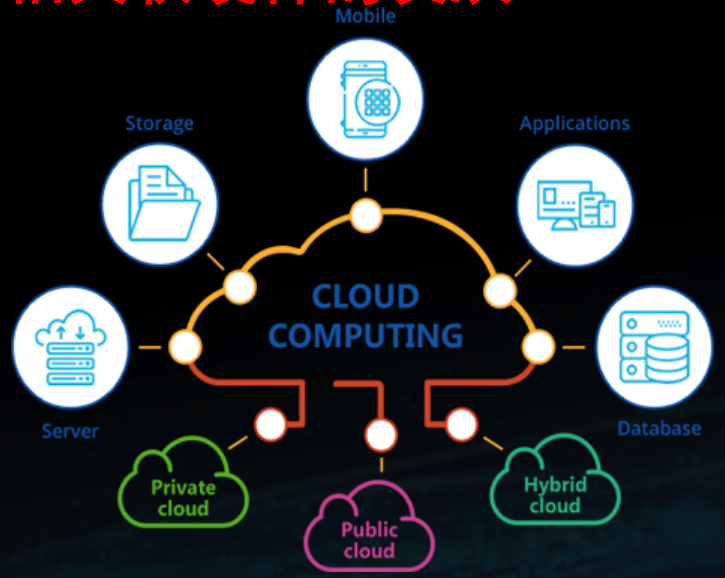
AI Products

Deepmind
OpenAI

统计学的挑战四

大数据的规模效应和
相关软硬件的发展

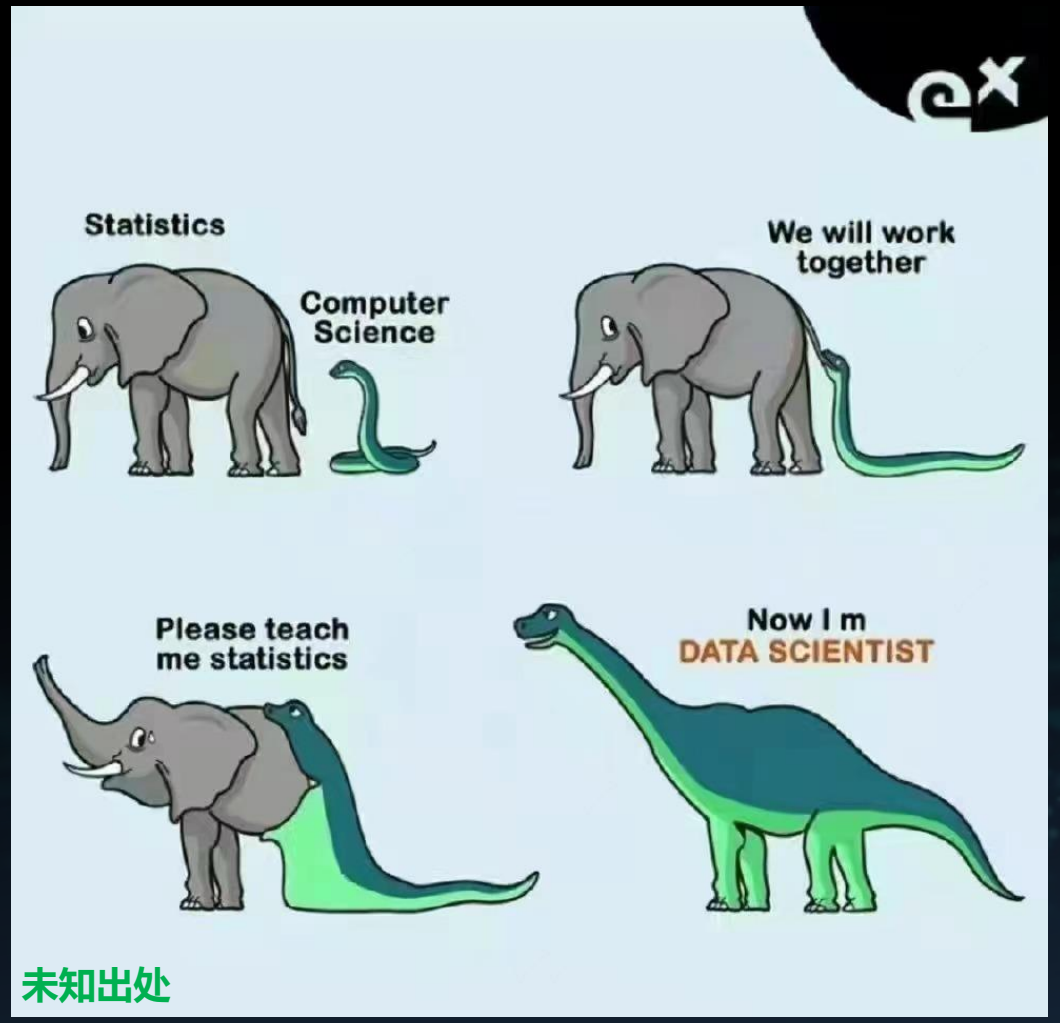




https://chinafilminsider.com/mobile-games-online-videos-movies-national-pastimes-chinese-new-year/

统计学的挑战五

统计学



未知出处

大型数据处理过程中
积累的经验和人才培养
AI=AS?

数据科学

统计学的挑战六

数据产品开发中积累的经验，人才培养，和影响力

智能型数据产品：通过收集和挖掘数据的价值来为受众（用户，企业，和政府）创造价值（比如，某种决策/行为）的一种产品形式。

报表型

工具型

定制服务型

智能型数据产品

遥感影像

红外线图像

医学图像

视频

肿瘤基因检测

司法亲子鉴定

优生优育检测

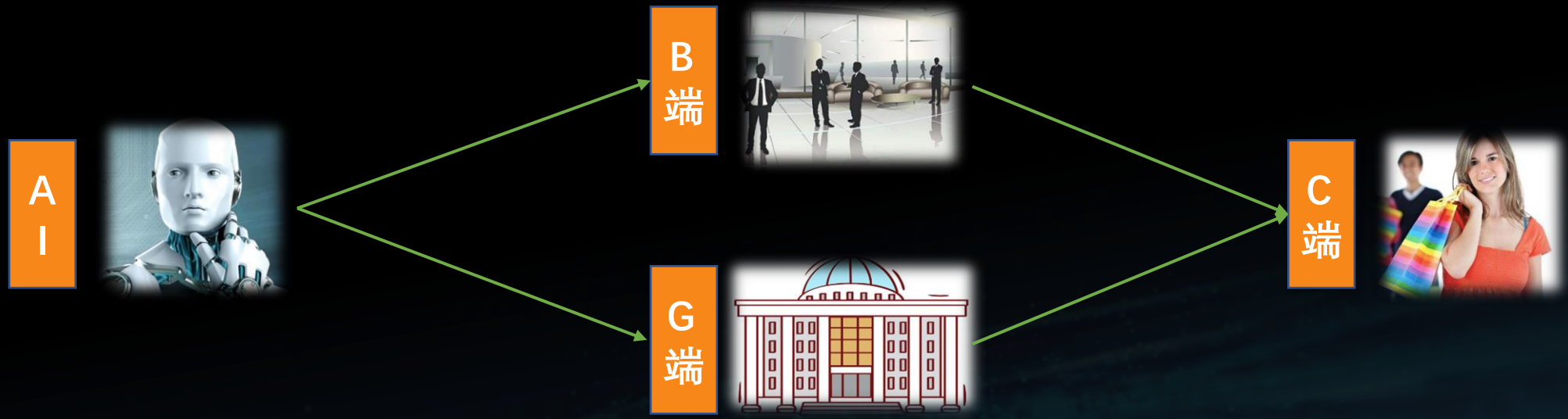
新生儿检测

DNA档案

UNC Biostatistics

统计学的挑战六

人工智能产品的核心是提效降本，更好的服务于人类社会；更好的为B端企业、为G端政府提效降本，更好的服务于C端民众



智能社会

智能经济

智能商业

智能劳动

UNC Biostatistics

统计学的挑战七

学科内部需要在人才培养，知识体系建设，数据产品的开发，和学科发展路径需要深度思考

统计学的挑战八

如何面对政府，学校，公司，和家长的期待需要深度思考

- Government
  NSF/NIH/DoD
- Universities
- Private Sector
- Parents

统计学的挑战八

NATIONAL AI RESEARCH INSTITUTES

The NSF-led National AI Research Institutes Program is the nation's largest AI research ecosystem and is supported by a partnership of federal agencies and industry leaders.

Main Map

Select Awards by year

2020 Awards | 2021 Awards | 2023 Awards

Select Awards by Institution

AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES)
University of Oklahoma

AI Institute for Foundations of Machine Learning (IFML)
University of Texas at Austin

AI Institute for Student-AI Teaming (iSAT)
University of Colorado Boulder

AI Institute for Molecular Discovery, Synthetic Strategy, and Manufacturing (Molecule Maker Lab or MMLI)
University of Illinois Urbana-Champaign

AI Institute for Artificial Intelligence and Fundamental Interactions (IAFI)
Massachusetts Institute of Technology

AI Institute for Next Generation Food Systems (AIFS)
University of California, Davis

AI Institute for Future Agricultural Resilience, Management, and Sustainability (AIFARMS)
University of Illinois Urbana-Champaign

AI Institute for Collaborative Assistance and Responsive Interaction for Networked Groups (AI-CARING)
Georgia Tech

AI Institute for Advances in Optimization (AI4OPT)
Georgia Tech

AI Institute for Learning-Enabled Optimization at Scale (TILOS)
University of California San Diego

AI Institute for Intelligent Cyberinfrastructure with Computational Learning in the Environment (ICICLE)
The Ohio State University

AI Institute for Edge Computing Leveraging Next-generation Networks (Athena)
Duke University

AI Institute for Dynamic Systems
University of Washington

AI Institute for Engaged Learning (ENGAGE AI Institute)
North Carolina State University

AI Institute for Adult Learning and Online Education (ALOE)
Georgia Institute of Technology

Institute for Agricultural AI for Transforming Workforce and Decision Support (AgAID)
Washington State University

AI Institute for Resilient Agriculture (AIIRA)
Iowa State University

AI Institute for Agent-based Cyber Threat Intelligence and Operation (ACTION)
University of California, Santa Barbara

AI Institute for Artificial and Natural Intelligence (ARNI)
Columbia University

AI Institute for Societal Decision Making (AI-SDM)
Carnegie Mellon University

AI Institute for Trustworthy AI in Law and Society (TRAILS)
University of Maryland, College Park

AI Institute for Inclusive Intelligent Technologies for Education (INVITE)
University of Illinois Urbana-Champaign

AI Institute for Exceptional Education (AI4ExceptionalEd)
University at Buffalo

Part II

Two Opportunities for Statisticians

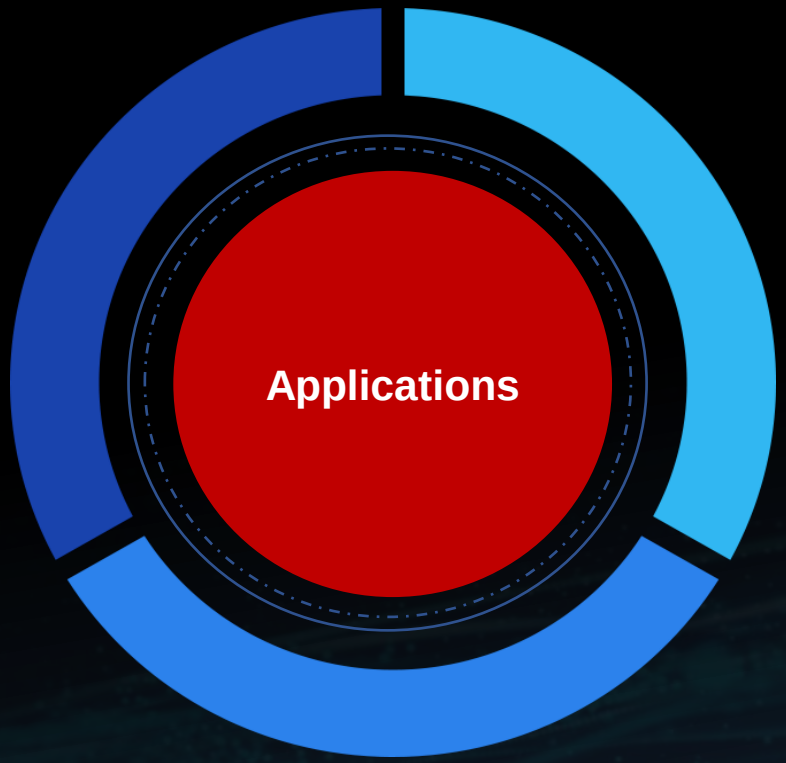
”Oddly, we are in a period where there has never been such a wealth of new statistical problems and sources of data. The danger is that if we define the boundaries of our field in terms of familiar tools and familiar problems, we will fail to grasp the new opportunities.”

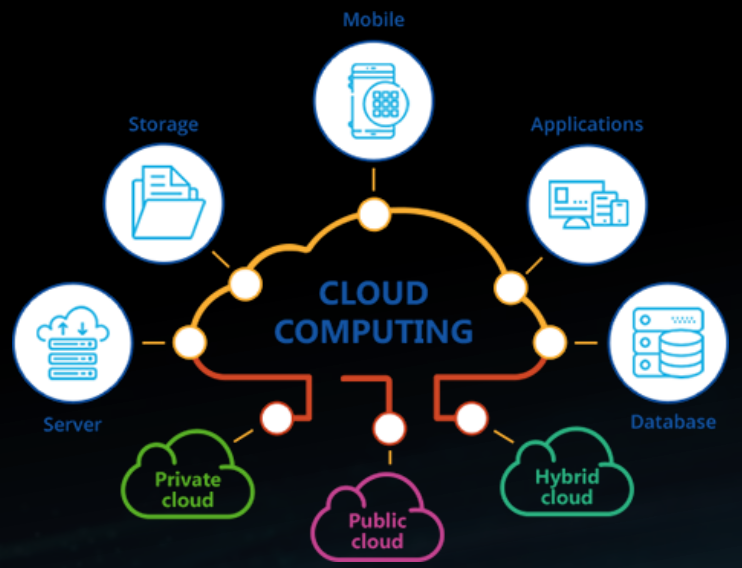
- Leo Breiman -

Deep Applications and Deep Math/Stat

Big Data

http://medium.com





Computing

Analytical Tools

Applied Mathematics

Statistics

Machine Learning

Engineering

Deep Applications

应用层的受众是谁?
数据是已经收集好的吗？能不能用来回答应用中真正有价值的问题
算法的结果对处理数据和应用层提供的信息有多大价值?

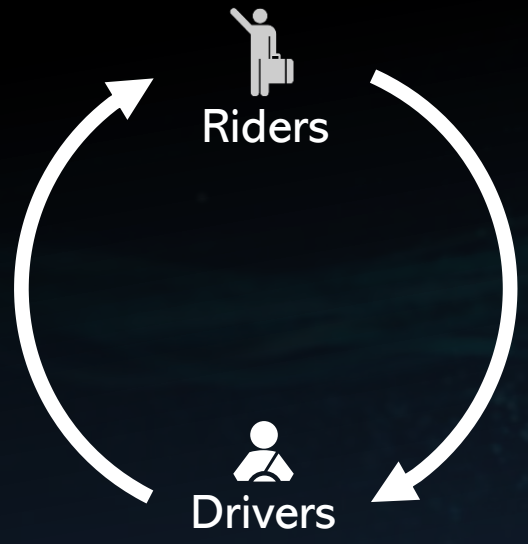
三个核心层相辅相成，相互制约，互相作用，缺一不可。

UNC Biostatistics

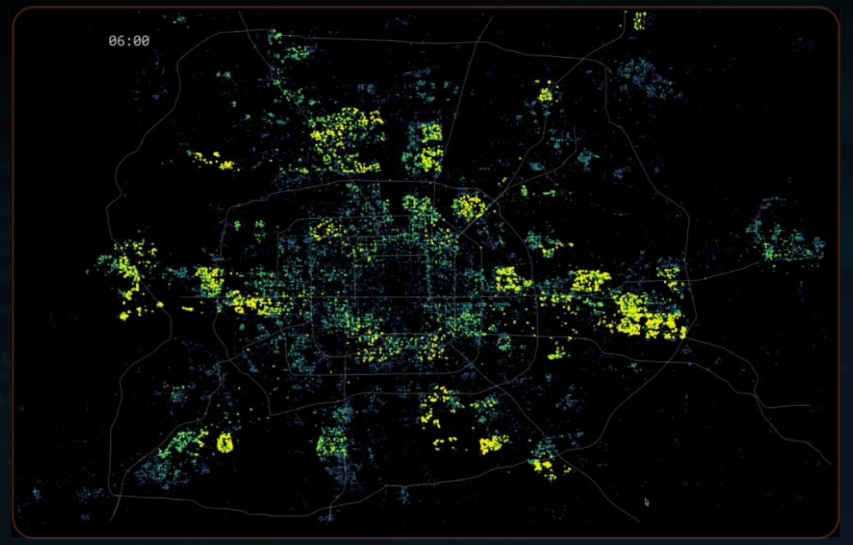
Ride-sharing Platform is a Complex Ecosystem

Spatio-temporal | Nonlinear | Interactive | Uncertainty | Causal

Two-sided Platform



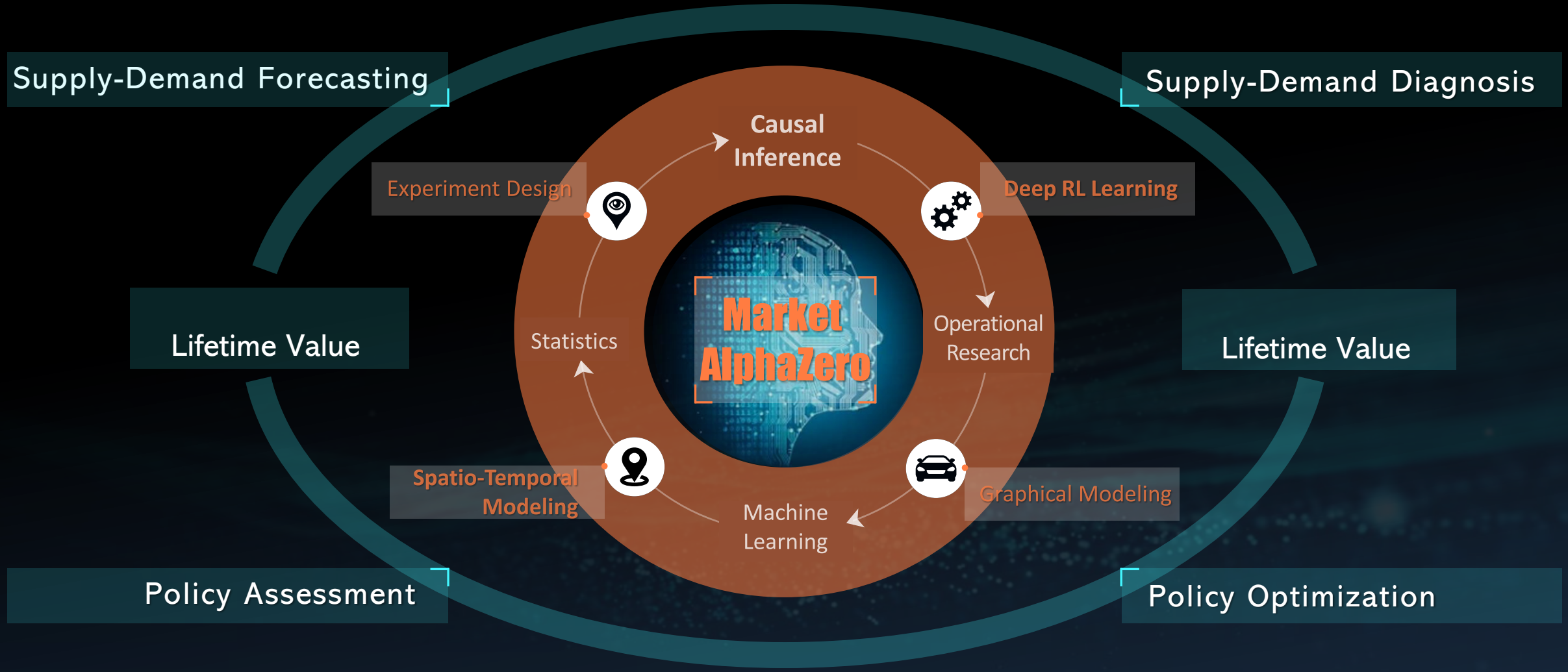
Complex Spatio-temporal System



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

Deep Reinforcement Learning

informs

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.

Synthesis Lectures on Learning, Networks, and Algorithms

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 Biostatics at The University of North Carolina at Chapel Hill.

ISBN 978-3-031-59639-1

springer.com

Qin · Tang · Li · Zhu · Ye

Reinforcement Learning in the Ridesharing Marketplace

Synthesis Lectures on Learning, Networks, and Algorithms

SYNTHESIS COLLECTION OF TECHNOLOGY

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

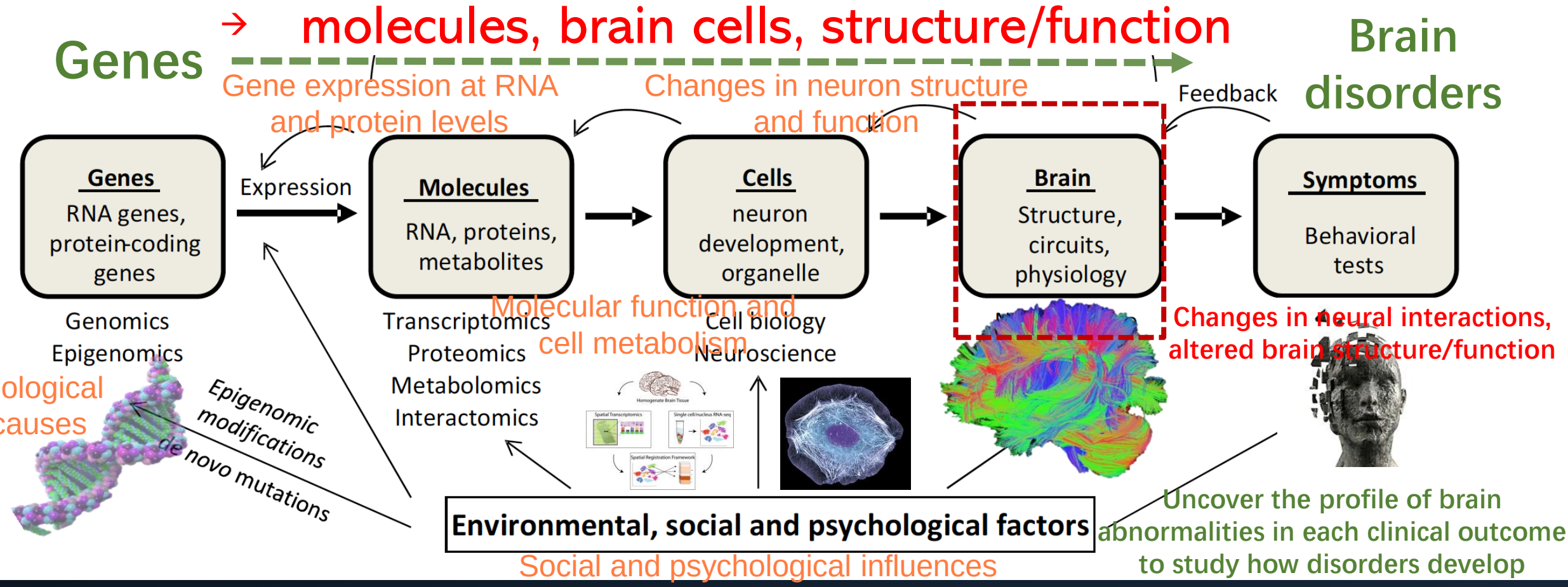
Reinforcement Learning in the Ridesharing Marketplace

Springer

UNC Biostatistics

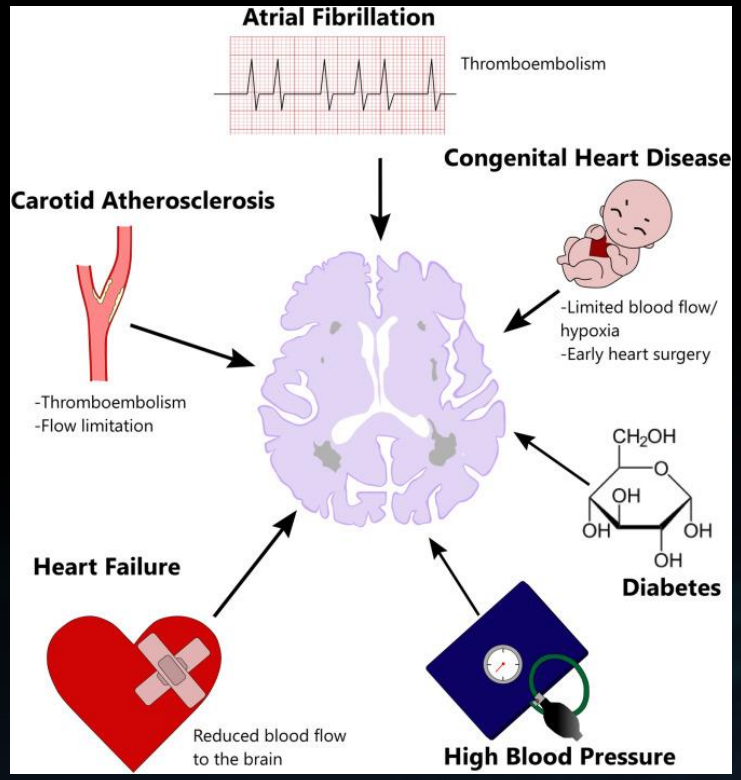
Brain Imaging Genetics Paradigm

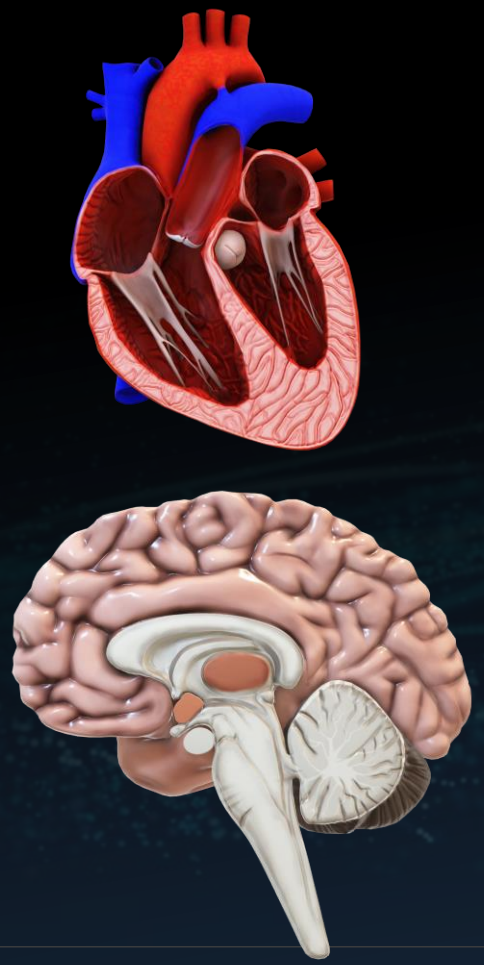
Neuroimaging: an important component to help understand the complex biological pathways of brain disorders

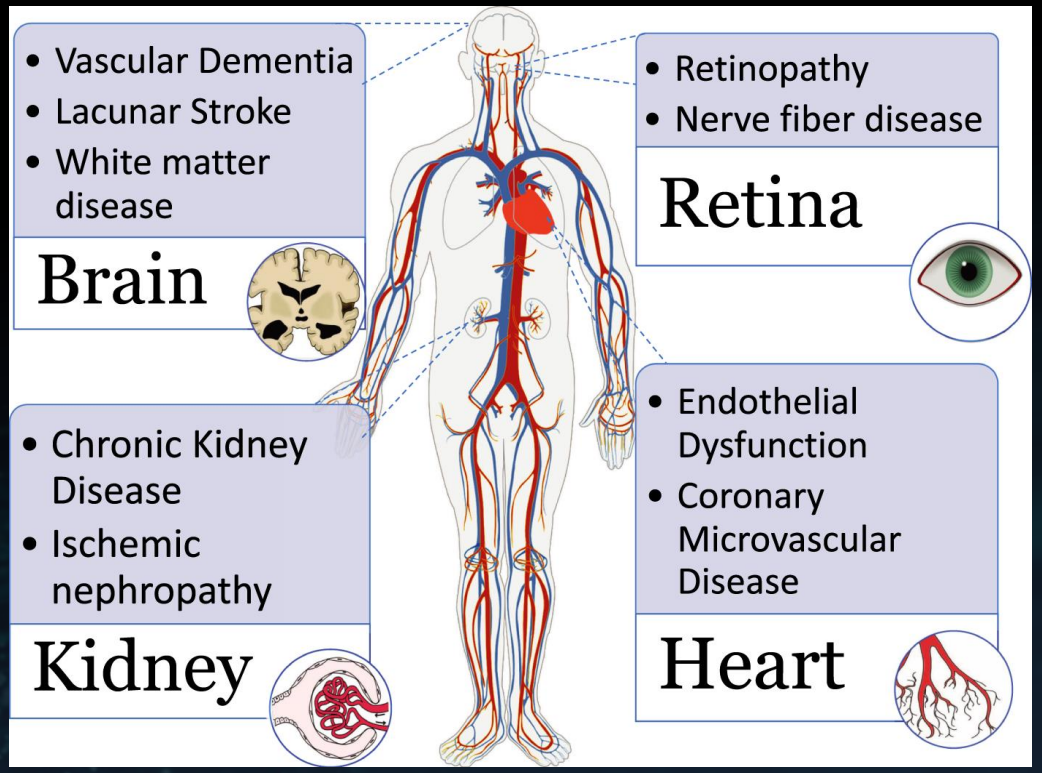


Multi-organ Health

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





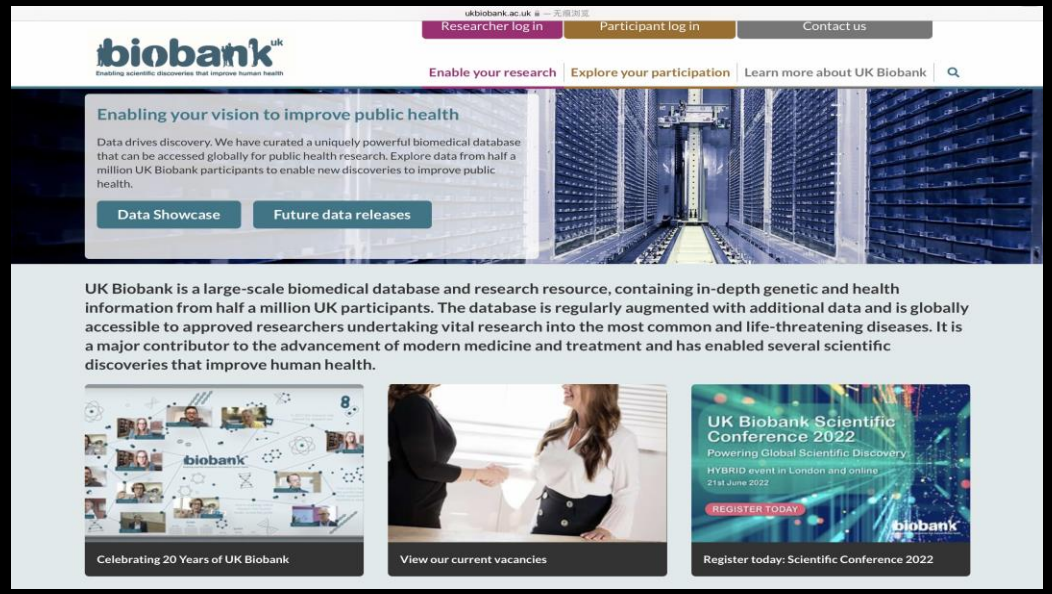


Possible causal factors of brain structure changes, resulting in brain disorders like stroke, dementia and cognitive impairment

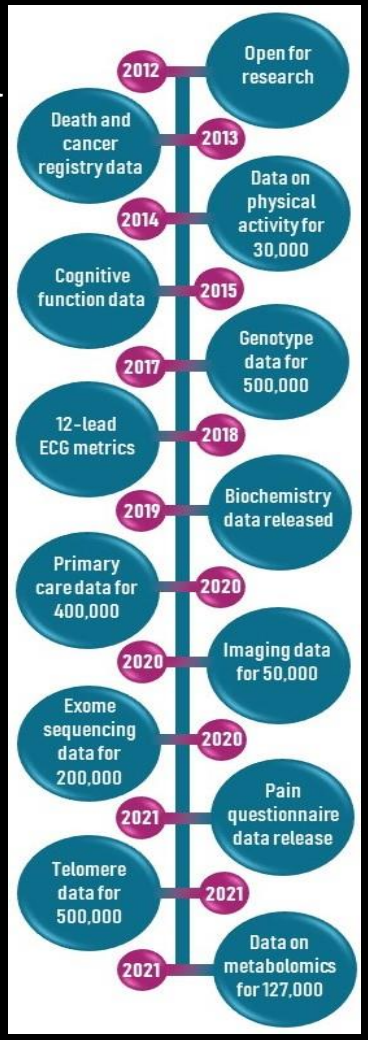
Many diseases (e.g., microvascular disease, high blood pressure) are multisystem disorders

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

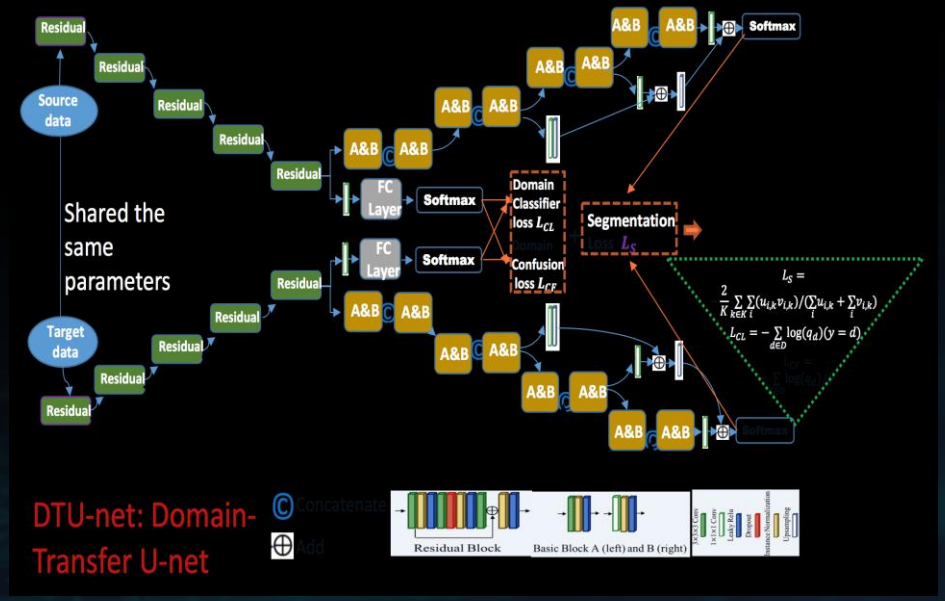


- **Imaging:** Brain, heart and full body MR imaging, plus full body DEXA scan of the bones and joints and an ultrasound of the carotid arteries. The goal is to image 100,000 participants, and to invite participants back for a repeat scan some years later.
- **Genetics:** Genotyping, whole exome sequencing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

AI for Image Segmentation

Segmentation Annotation

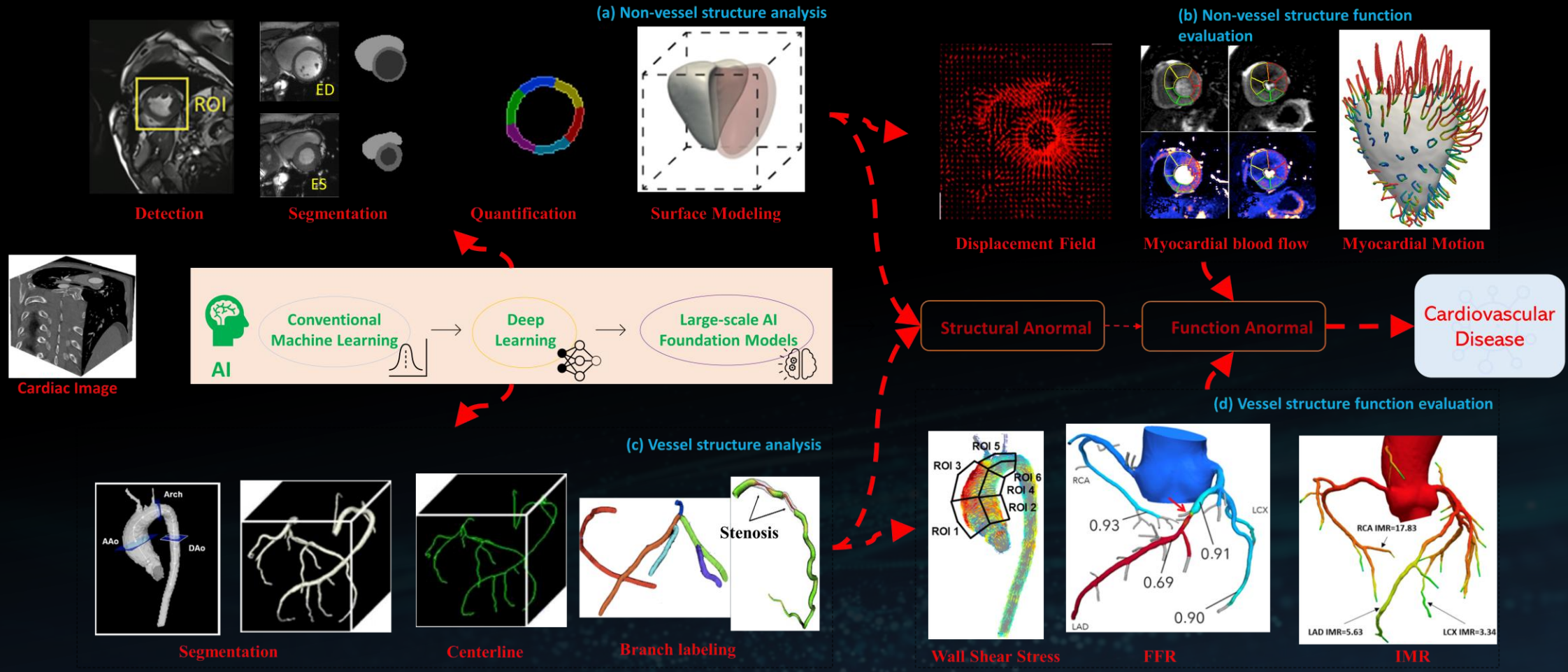
U-Nets



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

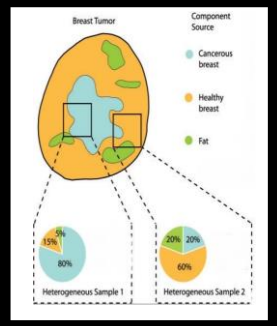
R. Azad *et al.*, “Medical Image Segmentation Review: The success of U-Net.” arXiv, Nov. 27, 2022.
Minaee, Shervin, et al. "Image segmentation using deep learning: A survey." *IEEE PAMI* 44.7 (2021): 3523-3542.

Image Analysis Pipeline



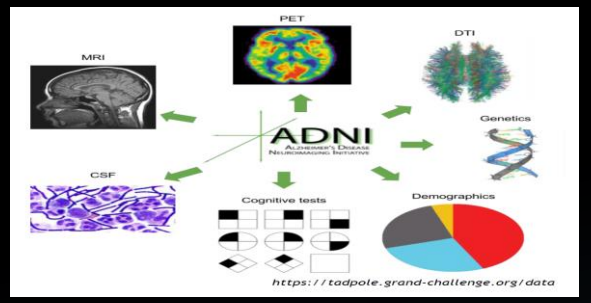
Wang, X. and Zhu, H (2024). Artificial Intelligence in Image-based Cardiovascular Disease Analysis: A Comprehensive Survey and Future Outlook

Ecological Layout for Large-scale Analysis

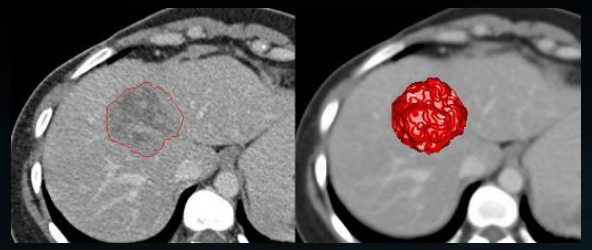


Deconvolution

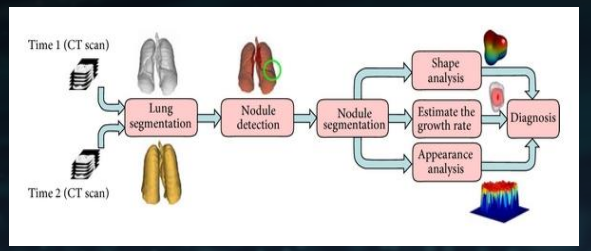
Biobank Database



Integration



Structural Learning



Prediction

Foundation Models for GMAI and Pan-Biobank

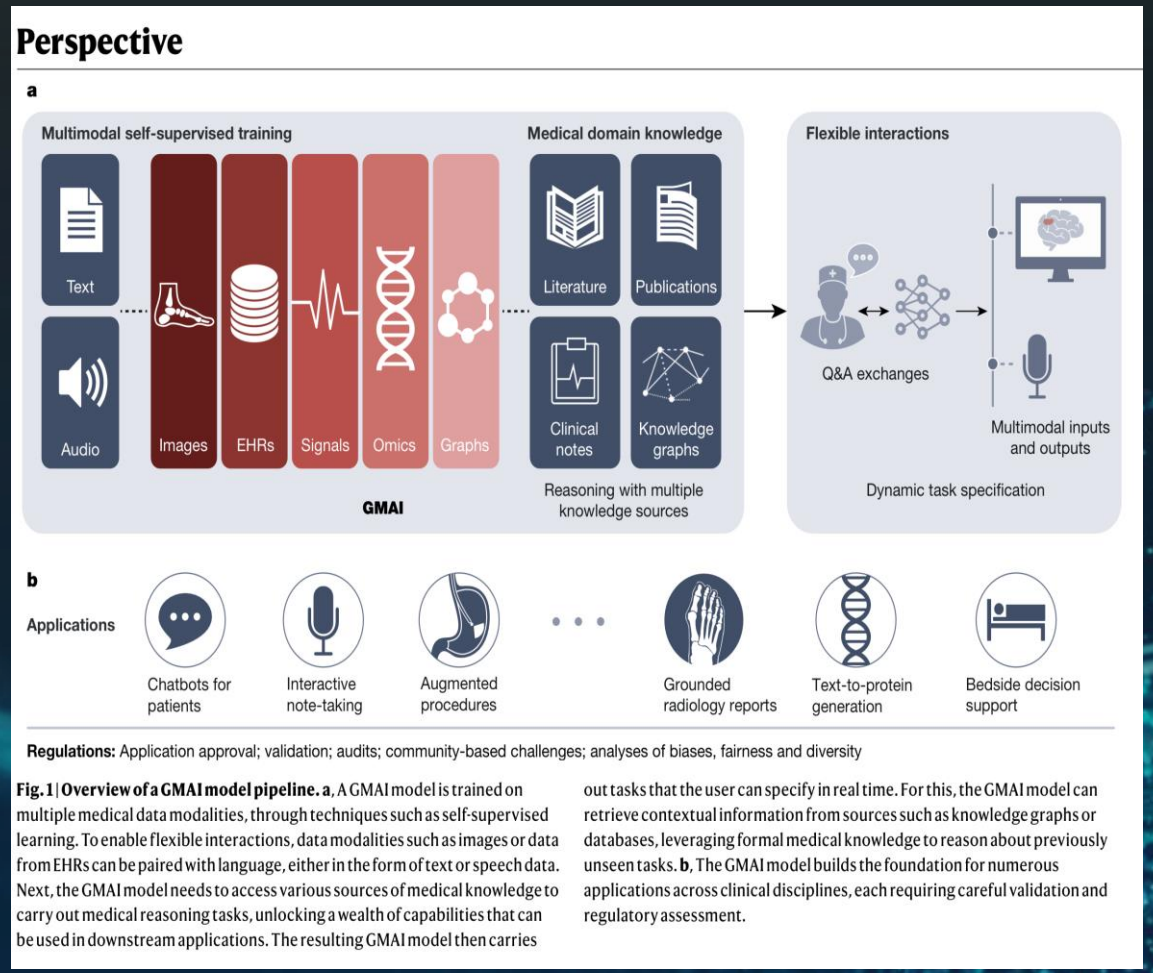
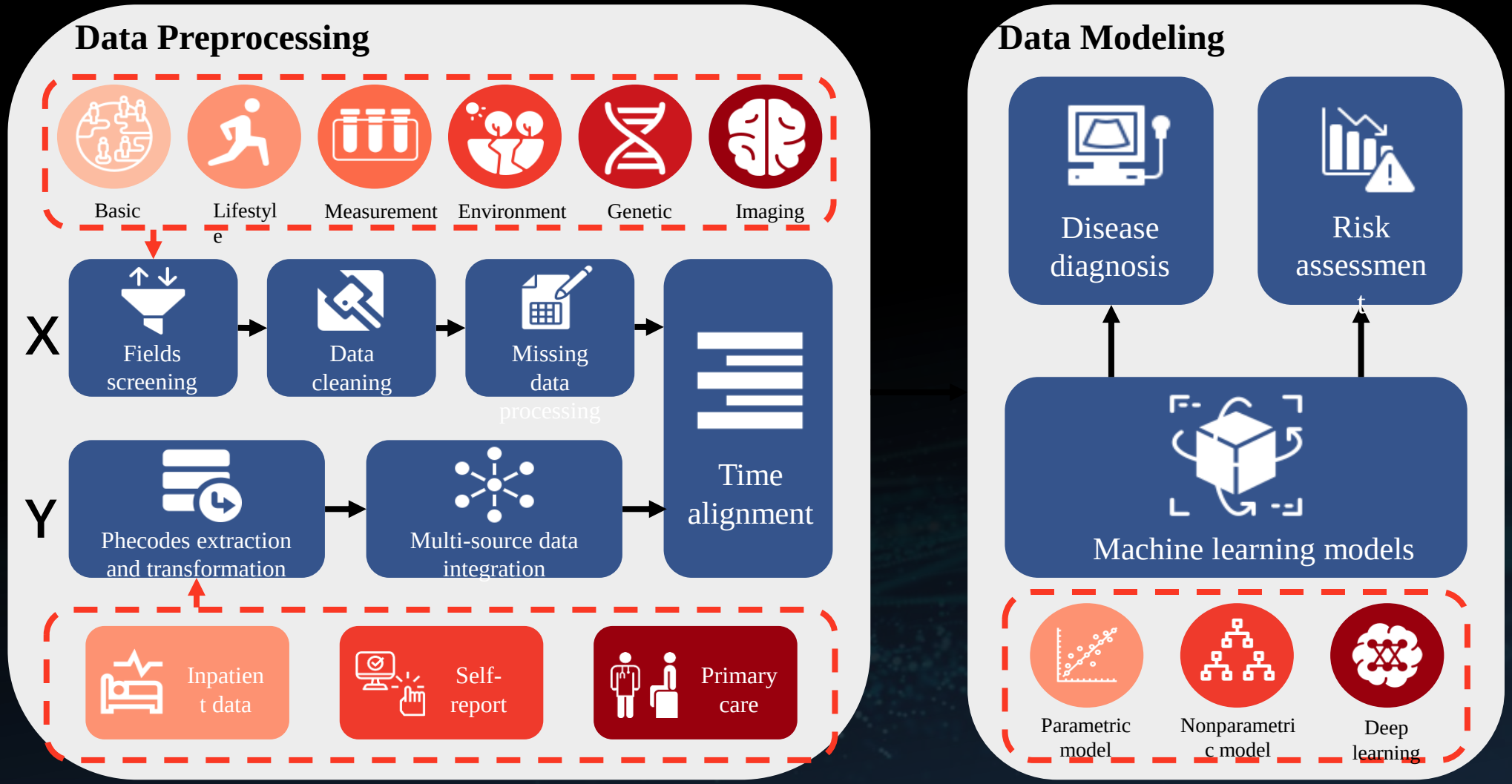


Fig. 1 | Overview of a GMAI model pipeline. a, A GMAI model is trained on multiple medical data modalities, through techniques such as self-supervised learning. To enable flexible interactions, data modalities such as images or data from EHRs can be paired with language, either in the form of text or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

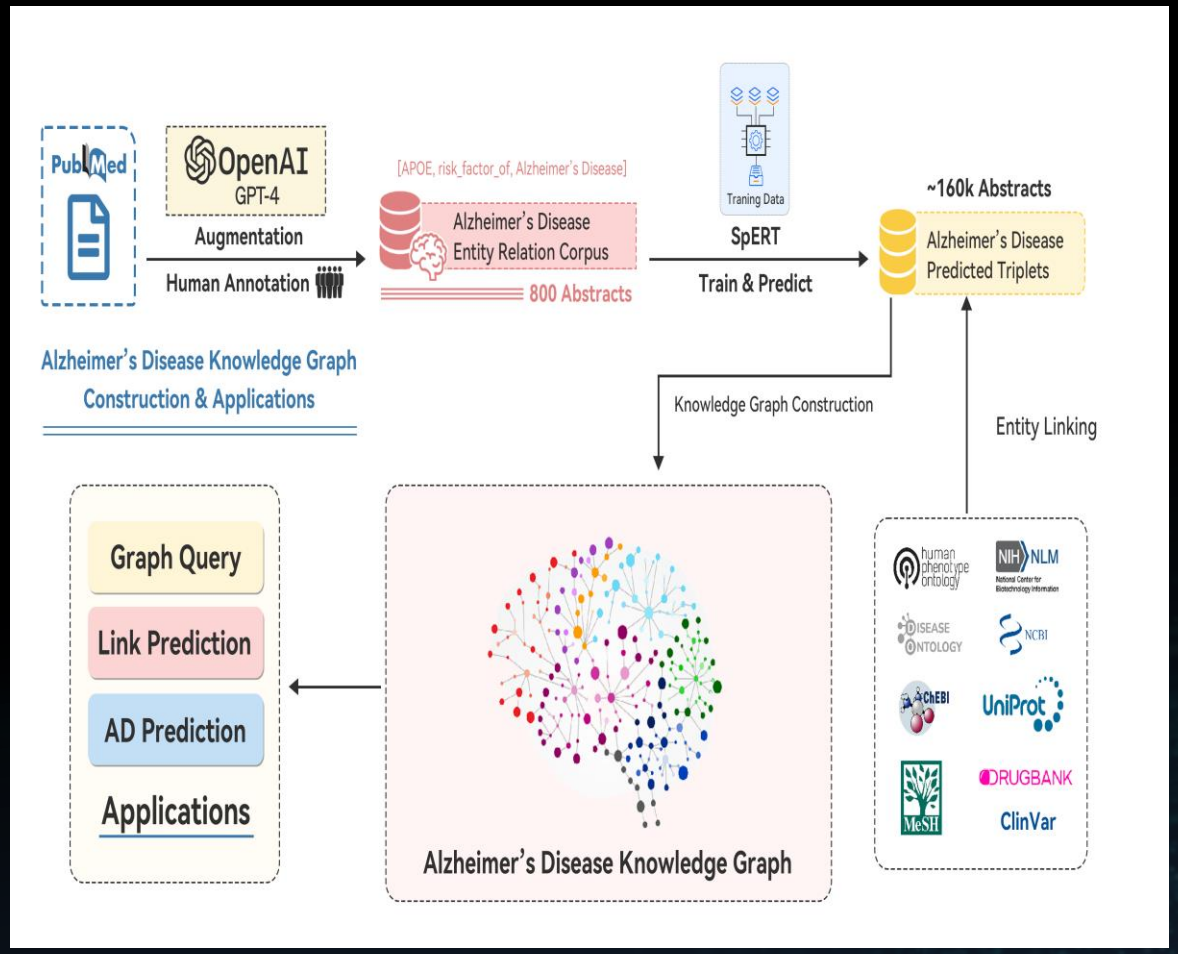
Pan-biobank studies

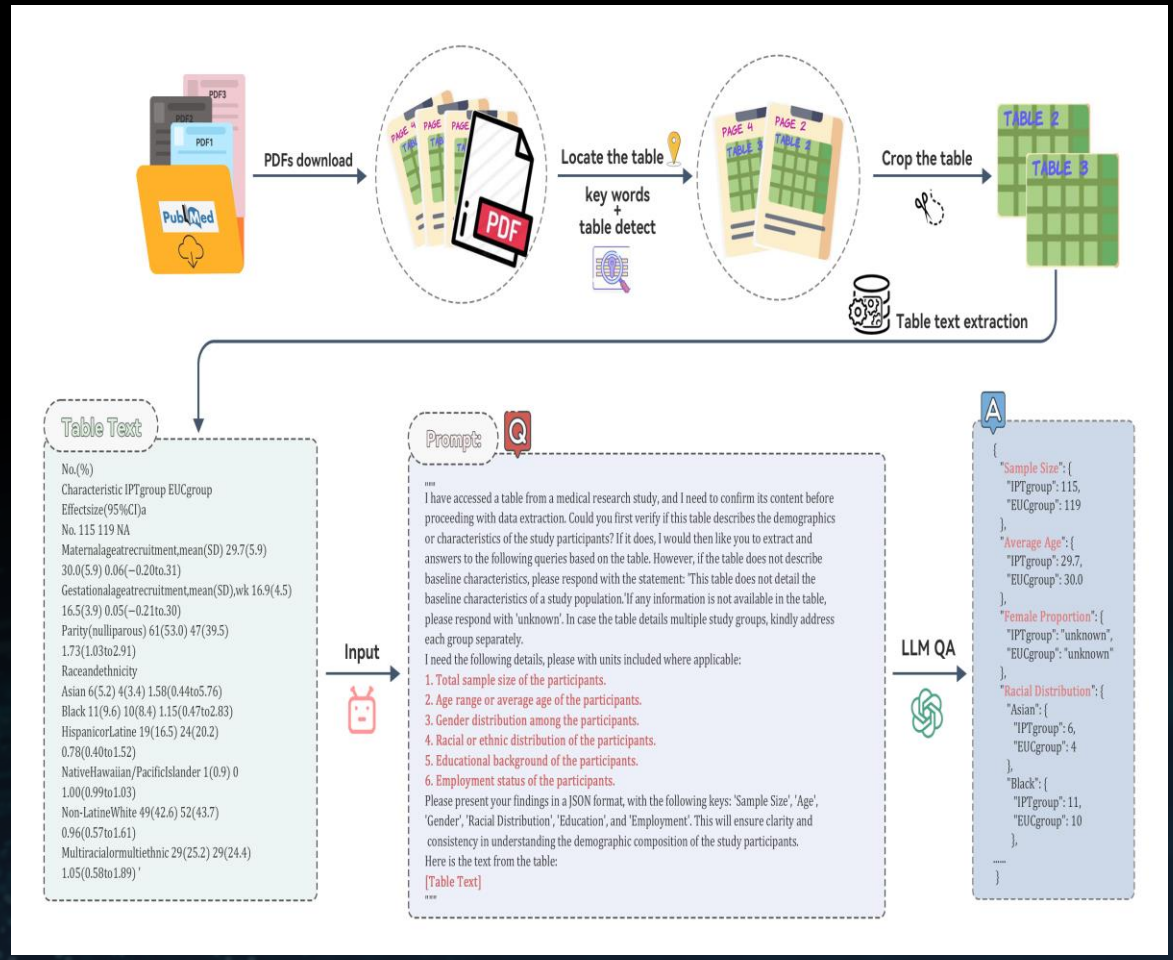
UKBFound



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

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

Deep Mathematics/Statistics

大部分统计/ML方法开始的时候是没有严谨的证明，而是一些直观的想法

- **Descriptive Statistics**
- **PCA**
- **Bootstrap**
- **Linear/logistic/Cox/LASSO regression**
- **Mixed effects**
- **EM/SA algorithms**
- **Causal inference**
- **Bayesian/MCMC**
- **Clustering**
- **Deep learning/DRL**

• **Level 1:** Demonstrates that a method works under restricted conditions.

• **Level 2:** Identifies key components that ensure the method's validity under realistic conditions.

• **Level 3:** Provides theoretical results that guide the further development of the method in more general settings.

Deep Learning Theory

Suh and Cheng (2024)

Data

$$\mathcal{D} := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n} \sim p(x)p(y|x)$$

Model

$$\mathbf{y}_i = f_o(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

Assumption

$$\mathbb{E}(\varepsilon_i | \mathbf{x}_i) = 0$$

Ideal

$$f_\rho := \mathbb{E}(\mathbf{y}|\mathbf{x}) = \operatorname{argmin}_{f \in \mathcal{G}} \quad \mathcal{E}(f) := \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\rho}\left[(\mathbf{y} - f(\mathbf{x}))^2\right]$$

Estimate

$$\widehat{f}_n = \operatorname*{arg\,min}_{f \in \mathcal{F}(L,\mathbf{p},\mathcal{N})} \mathcal{E}_D(f) := \operatorname*{arg\,min}_{f \in \mathcal{F}(L,\mathbf{p},\mathcal{N})} \left\{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - f(\mathbf{x}_i))^2\right\}$$

Risk Error

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f_\rho) \leq \frac{\text{Complexity Measure of } \mathcal{F}}{n} + \frac{\text{Approx. Error}}{\sqrt{n}} + \text{Approx. Error}^2$$

Approx Error

$$\varepsilon_{\text{Apprx}} := \sup_{f_\rho \in \mathcal{G}} \inf_{f \in \mathcal{F}(L,\mathbf{p},\mathcal{N})} \|f - f_\rho\|_{L^p}$$

Complexity

$$\text{VCdim}(\mathcal{F}), \text{Pdim}(\mathcal{F}) \asymp \mathcal{O}(L\mathcal{N}\log(\mathcal{N}))$$

Functional Equivalence

Shen et al. (2024) ICML.

Theorem 3 (Covering number of shallow neural networks)

Consider the class of shallow neural networks $\mathcal{F} := \mathcal{F}(1, d_0, d_1, B)$ *parameterized by* $\theta \in \Theta = [-B, B]^S$. *Suppose the radius of the domain* $\mathcal{X}$ *of* $f \in \mathcal{F}$ *is bounded by some* $B_x > 0$, *and the activation* σ_1 *is continuous. Then for any* $\epsilon > 0$, *the covering number*

$$\mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|_\infty) \leq (16B^2(B_x+1)\sqrt{d_0}d_1/\epsilon)^S \times \rho^{S_h}/d_1!, \tag{3}$$

where ρ *denotes the Lipschitz constant of* σ_1 *on the range of the hidden layer (i.e.,* $[-\sqrt{d_0}B(B_x)+1), \sqrt{d_0}B(B_x+1)]$*), and* $S_h = d_0 d_1 + d_1$ *is the total number of parameters in the linear transformation from input to the hidden layer, and* $S = d_0 \times d_1 + 2d_1 + 1$ *is the total number of parameters.*

- A reduced complexity (by $d_1!$) compared to existing studies [25, 3, 27, 23, 17]. For a shallow ReLU network with $d_1 = 128$, covering number reduced by $\approx 10^{215}$.

Theorem 4 (Covering number of deep neural networks)

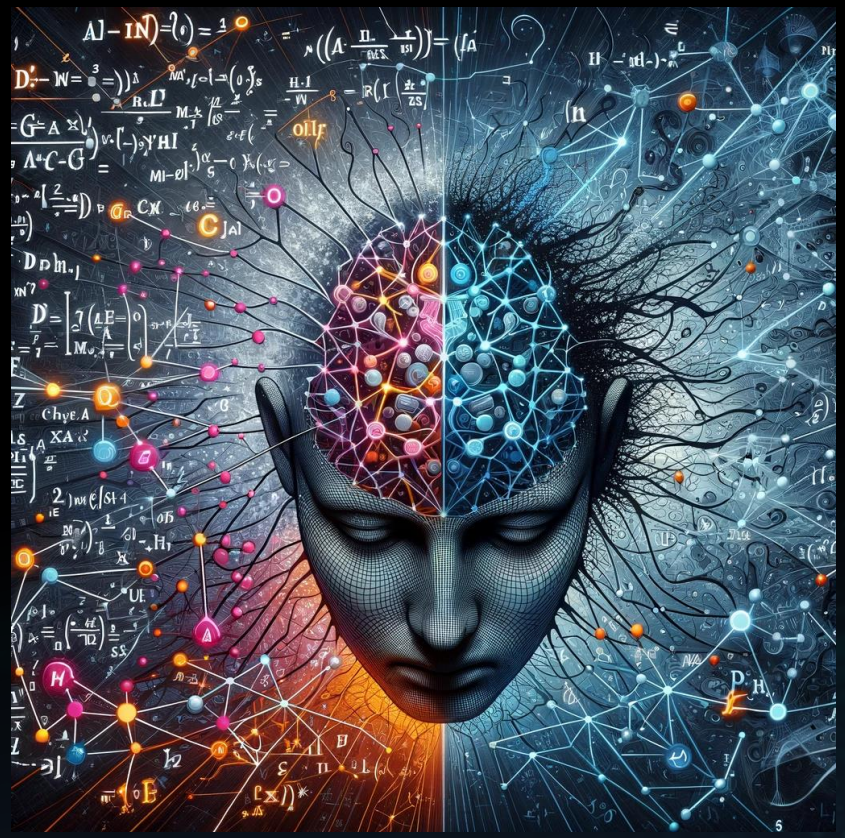
Consider the class of deep neural networks $\mathcal{F} := \mathcal{F}(1, d_0, d_1, \ldots, d_L, B)$ *parameterized by* $\theta \in \Theta = [-B, B]^S$. *Suppose the radius of the domain* $\mathcal{X}$ *of* $f \in \mathcal{F}$ *is bounded by* B_x *for some* $B_x > 0$, *and the activations* $\sigma_1, \ldots, \sigma_L$ *are locally Lipschitz. Then for any* $\epsilon > 0$, *the covering number* $\mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|_\infty)$ *is bounded by*

$$\frac{\left(4(L+1)(B_x+1)(2B)^{L+2}(\Pi_{j=1}^{L}\rho_j)(\Pi_{j=0}^{L}d_j)\cdot\epsilon^{-1}\right)^S}{d_1! \times d_2! \times \cdots \times d_L!},$$

where $S = \sum_{i=0}^{L} d_i d_{i+1} + d_{i+1}$ *and* ρ_i *denotes the Lipschitz constant of* σ_i *on the range of* $(i-1)$*-th hidden layer, especially the range of* $(i-1)$*-th hidden layer is bounded by* $[-B^{(i)}, B^{(i)}]$ *with* $B^{(i)} \leq (2B)^i \Pi_{j=1}^{i-1}\rho_j d_j$ *for* $i = 1, \ldots, L$.

- A reduced complexity (by $(d_1!d_2!\cdots d_L!)$) over existing studies [25, 3, 27, 23, 17].
- Increasing depth L does increase complexity. The increased hidden layer l will have a $(d_l!)$ discount on the complexity.

Deep Mathematics/Statistics



- Existing results cannot explain why deep and/or reinforcement learning methods work in realistic scenarios.
- Existing mathematical and statistical theory is not good enough to validate many algorithmic modelling.
- Many breakthroughs in algorithmic modeling do not have any mathematical reasoning at the beginning.

Deep Mathematics/Statistics

How to theoretically ensure the extraction of signals of interest in real data?

- Start with realistic and challenging scenarios.
- Understand the signal patterns and complexity of a specific problem.
- Formulate the problem and understand how a method really works in simulated and real settings.
- Prove theoretical results from Level 1 to Level 3.

Statistics Up AI Alliance

https://statsupai.org