# Tony Ding

***xiayiding@hsph.harvard.edu · tding@mit.edu · 213-245-5570 · LinkedIn · Personal Website***

## EDUCATION

**Harvard University** *Aug 2022 – May 2024 (Expected)*

*Master of Science* in Health Data Science

- GPA: 4.0/4.0; Recipient of: Lyman & Grew Memorial Scholarship AND Harvard Central Grant Scholarship
- Cross-registrations at **MIT**: *ML in Healthcare, Advances in Computer Vision, Advanced NLP*; GPA: 5.0/5.0

**University of Southern California** *Aug 2018 – May 2022*

*Bachelor's* Degrees in Data Science and in Neuroscience

- Data Science major GPA: 4.0/4.0; Cumulative GPA: 3.9/4.0
- Presidential Scholar (half-tuition awarded throughout 4 years); Renaissance Scholar Distinction

## RELEVANT PROJECTS

Imaging-based prediction of transcriptional subtypes in Alzheimer's Disease**, Harvard & MIT CSAIL**

- Applied and fine-tuned Vision Transformers (ViTs) on Alzheimer's MRI imaging dataset using PyTorch. Also used ViT in 3D brain MRI image analysis by treating each 3D patch as a vector to feed into the transformer. The best ViT model achieved a 98.57% test accuracy. Extracted feature representations of brain MRI images and combined them with genetic deep learning models to further improve model performance using a self-supervised contrastive learning pipeline.

Pleural effusion diagnosis: multimodal approaches using deep neural networks and transformer-based architectures,**MIT EECS**

- Combined patients' clinical reports with X-ray images to examine the best fusion strategies for implementing a multimodal approach to diagnose pleural effusion. Used two sets of fusion strategies, namely early fusion and late fusion. Eventually achieved the best AUC, 0.9887, by using VGG16 with DistilBERT in addition to the late fusion multimodality model with an elastic net logistic regression model before classification.

## INDUSTRY EXPERIENCES

*R&D Data Science Intern*, **Johnson & Johnson**, Cambridge, MA *June 2023 – August 2023*

- Completed a referral network analysis project, using claims data, by implementing the Leiden algorithm to detect patient referral communities and by fine-tuning weighted PageRank algorithms to quantitatively rank the Healthcare organizations (HCOs). Identified top influential HCOs for multiple myeloma patient referrals to aid in clinical trial site selection.
- Built five large-scale interactive network visualizations and dashboards using the Dash framework in Python.

*Data Scientist Intern*, **AstraZeneca**, Shanghai, CN *May 2021 – August 2021*

- Built a supervised machine learning pipeline to model and predict users' behavior on AZ_MedInfo (a medical information exchange platform with over 1M users). Increased the AUC score by 26.3% by fine-tuning a Random Forest classifier. Successfully identified 12 out of 161 most significant and meaningful variables that impact users' decisions.
- Designed a weighted association rule mining program, for Prof. Binghe Xu, MD, by extracting keywords like "ctDNA" and "Breast Cancer" from publications' titles and abstract sections to identify and rank his associations with fellow colleagues and determine his individual academic rankings among all researchers in that field.

*Data Science & Visualization Intern*, **Takeda Pharmaceutical Company**, Cambridge, MA *May 2020 – August 2020*

- Extensively pre-processed EHR and clinical data and implemented unsupervised machine learning models on raw baseline data for PANDA, a Takeda's oncology program for Ponatinib. Successfully identified 4 significant patient subpopulations and the key risk drivers of MACE(Major Adverse Cardiac Event) among Chronic Myelogenous Leukemia patients.
- Utilized MS SQL Server Management Studio and Excel to quantitatively analyze Takeda R&D partnerships. Created a new R&D partnerships visualization paradigm for Takeda and designed 8 time-based interactive network visualizations.

## RESEARCH EXPERIENCE

*Statistics Research Assistant*, **USC Health, Emotion, & Addiction Lab** *Dec 2020 – Jul 2022*

- Wrote SPSS syntax and conducted various statistical tests and analysis in R for ADVANCE (*Assessing Developmental Patterns of Vaping, Alcohol, Nicotine, and Cannabis Use and Emotional Well-being*) School Reports project to determine the significance of associations among variables. Calculated RCADS(Revised Children's Anxiety and Depression Scales) scores and applied my syntax to create analytical reports for all 6 high schools in the Greater Los Angeles area.

## ADDITIONAL SKILLS & AWARDS

- Expert in Python, SQL, R, machine learning algorithms and libraries(PyTorch, Tensorflow, sklearn, opencv/cv2, etc.)
- Proficient in database modeling and ETL; Proficient in using Google Cloud Platform, SAS, Tableau, and MongoDB
- Bright Futures Award (2nd Place in US) in 2023 NNLM Data Visualization Challenge - Complex Visualization Category
- Alpha Prize (2nd Place in the World) in AoCMM Mathematical Modeling Contest