

Yiru Chen

<https://www.cs.columbia.edu/~chen1ru/>

☎ +1 3474072116 ✉ yiru.chen@columbia.edu

RESEARCH INTEREST

Data Management, Databases, Visual Data Analytics, Query Optimization, LLM for Data Systems

EDUCATION

Columbia University	New York, USA
Ph.D. in Computer Science	2018 - present
Master of Philosophy	2018 - 2022
M.S. in Computer Science	2018 - 2019
Supported by Google PhD Fellowship	
GPA: 4/4	
Advisor: Prof. Eugene Wu	
Peking University	Beijing, China
B.S. Summa Cum Laude in Computer Science	2014 - 2018
B.A. in Economics	2015 - 2018

WORKING EXPERIENCE

Visiting Scholar	Feb 2024 - present
<i>University of California, Berkeley</i>	<i>California, USA</i>

- **Text Analytics Database based on LLMs**

- *Textdb*: A novel data management system tailored for unstructured textual data based on LLMs.

Graduate Research Assistant	Sep 2018 - present
<i>Columbia University</i>	<i>New York, USA</i>

- **Precision Interface** - <https://github.com/learnedinterfaces/PI2>

A line of works to generate customized interactive data interfaces from intended queries.

- *Data Interface Grammar*: A representation of analysis that is compatible with existing engineering practices, compact to represent any analysis, and simple to translate into an interface design.
- *PI2*: The first system that models the interface generation problem and uses Monte Carlo Tree Search to generate fully functional interactive visualization interfaces from task queries.
- *NL2INTERFACE*: A system prototype that combines *PI2* and large language models to generate fully functional interactive visualization interfaces from natural language questions.

- **Physical Visualization Design**

- *Jade*: A system that automatically generates the system backend for data interfaces, taking into account data structures, data placement, and client-server architecture, to ensure responsiveness.

- **Explainable Artificial Intelligence**

- *Deepbase*: Explanation of neural network behaviors using a unified interface by identifying statistical dependencies between user-provided hypothesis functions and groups of units in the models.

Research Intern	May 2020 - Aug 2020
<i>Data System Group, Microsoft Research, Redmond</i>	<i>Washington, USA</i>

- **Data Explanation**

- *TSExplain*: An engine designed to explain aggregated time series by identifying key evolving contributors across various time segments.

Undergraduate Research Assistant

Peking University

Jan 2016 - July 2018

Beijing, China

- **In-Database Machine Learning**

- *MLog*: A declarative language integrates machine learning into data management systems, allowing the system to manage all data movement, data persistency, and ML-related optimizations.

- **Topic Model**

- *Sys-TM*: A framework featuring a hybrid sampler balances the trade-offs between document length and word frequency, supporting LDA and its variants, including STM, TOT, and CTM.

Research Intern

Social Computing Group, Microsoft Research Asia

Oct 2017 - Apr 2018

Beijing, China

- **Recommender System**

- *Explainable Recommendation*: A model-agnostic reinforcement learning framework designed to explain results of any recommender system, providing flexible control over the explanation quality.

Undergraduate Research Intern

Carnegie Mellon University

June 2017 - Sep 2017

Pennsylvania, USA

- *Social Network*: Designed an algorithm to cluster user groups on Twitter based on their sentiment towards topics and identify the important users within a group.

Open Source Contributor

Google Summer of Code

Apr 2016 - Aug 2016

Beijing, China

- *Biomedical Data Integration*: Developed *Óbidos* - a fast and efficient data integration system allowing lazy materialization, data tracking and data sharing for multi-modal biomedical data.

SELECTED HONORS AND AWARDS

- EECS Rising Star, 2023
- Google PhD Fellowship, 2021
- SIGMOD Travel Award, 2022, 2023
- ICDE Travel Award, 2023
- VLDB Attendance Grant, 2021
- CRA-W Travel Award, 2021
- 2nd prize in ACM SIGMOD Programming Contest, 2017
- China National Scholarship, 2015, 2017
- Founder Scholarship, 2016
- HuaWei Scholarship, 2017
- Outstanding Student of Beijing City, 2018
- Pacemaker to Merit Student, Peking University, 2017
- Outstanding Research Award, Peking University, 2017

- Selected into CS Elite Class (Turing Class), Peking University, 2016
- Outstanding Leader In EECS, Peking University, 2015
- 1st Prize Chinese Mathematical Olympiad in Jiangsu Province (CMOP), 2014, 2013
- Recommended Early Admission into Peking University, 2013

PUBLICATIONS

Yiru Chen, Eugene Wu

Automatic and Efficient Physical Visualization Design

Under Preparation

Yiru Chen, Jeffery Tao, Eugene Wu

DIG: The Data Interface Grammar [\[pdf\]](#)

SIGMOD HILDA workshop 2023

Yiru Chen, Silu Huang

TSEXPLAIN: Explaining Aggregated Time Series by Surfacing Evolving Contributors [\[pdf\]](#)

ICDE 2023

Yiru Chen, Ryan Li, Austin Mac, Tianbao Xie, Tao Yu, Eugene Wu

NL2INTERFACE: Interactive Visualization Interface Generation from Natural Language Queries[\[pdf\]](#)

VIS NLVIZ workshop 2022

Yiru Chen, Eugene Wu

PI2: End-to-end Interactive Visualization Interface Generation from Queries[\[pdf\]](#)

SIGMOD 2022

Jeffery Tao, Yiru Chen, Eugene Wu

Demonstration of PI2: Interactive Visualization Interface Generation for SQL Analysis in Notebook [\[pdf\]](#)

SIGMOD Demonstration 2022

Yiru Chen, Silu Huang

TSExplain: Surfacing Evolving Explanations for Time Series[\[pdf\]](#)

SIGMOD Demonstration 2021

Yiru Chen, Eugene Wu

Monte Carlo Tree Search for Generating Interactive Data Analysis Interfaces[\[pdf\]](#)

AAAI IPA workshop 2020

Yingxia Shao, Xupeng Li, Yiru Chen, Lele Yu, Bin Cui

Sys-TM: A Fast and General Topic Modeling System[\[pdf\]](#)

TKDE 2019

Thibault Sellam, Kevin Lin, Ian Huang, Yiru Chen, Michelle Yang, Carl Vondrick, Eugene Wu

DeepBase: Deep Inspection of Neural Networks[\[pdf\]](#)

Technical Report

Yiru Chen, Yiliang Shi, Boyuan Chen, Thibault Sellam, Carl Vondrick, Eugene Wu

Deep Neural Inspection Using DeepBase[\[pdf\]](#)

NeurIPS Sys for ML workshop 2018

Xiting Wang, Yiru Chen, Jie Yang, Le Wu, Zhengtao Wu, Xing Xie

A Reinforcement Learning Framework for Explainable Recommendation [\[pdf\]](#)

ICDM 2018

Xupeng Li, Bin Cui, **Yiru Chen**, Wentao Wu, and Ce Zhang

MLog: Towards Declarative In-Database Machine Learning [\[pdf\]](#)

VLDB Demonstration 2017

Pradeeban Kathiravelu, **Yiru Chen**, Ashish Sharma, Helena Galhardas, Peter Van Roy, Luís Veiga

On-Demand Service-Based Big Data Integration: Optimized for Research Collaboration [\[pdf\]](#)

VLDB Data Management and Analytics for Medicine and Healthcare workshop 2017

Dominic Stark, Barthélemy Launet, Kevin Schawinski, Ce Zhang, Michael Koss, M Dennis Turp, Lia F Sartori, Hantian Zhang, **Yiru Chen**, Anna K Weigel

PSFGAN: a generative adversarial network system for separating quasar point sources and host galaxy light [\[pdf\]](#)

Monthly Notices of the Royal Astronomical Society

TALKS AND PRESENTATIONS

- **Towards Democratizing Data Interfaces**
Rising Stars in CS Lecture Series at University of Massachusetts Amherst 2023
- **DIG: The Data Interface Grammar**
SIGMOD HILDA workshop 2023
- **NL2INTERFACE: Interactive Visualization Interface Generation from Natural Language Queries**
VIS NLVIZ workshop 2022
- **PI2: End-to-end Interactive Interface Generation from SQL Queries**
North East Database Day 2023
New York Data Science Day 2023
SIGMOD 2022
SIDMOD Demonstration 2022
Jane Street's Symposium 2022
- **TSEXPLAIN: Explaining Aggregated Time Series by Surfacing Evolving Contributors**
ICDE 2023
SIGMOD Demonstration 2021
Microsoft Research 2020
- **Monte Carlo Tree Search for Generating Interactive Data Analysis Interfaces**
Grad Cohort Workshop for Women 2021
AAAI IPA workshop 2020
New York Data Science Day 2020
- **Deep Neural Inspection Using DeepBase**
New York Data Science Day 2019
New York Database Workshop 2019
- **MLog: Towards Declarative In-Database Machine Learning**
VLDB Demonstration 2017
- **SIGMOD Programming Contest Runner Up Poster Presentation 2017**

SERVICE

To the Profession

- Session Chair - VLDB 2022
- Program Committee Member - WSDM Demonstration 2024, VLDB Demonstration 2023, BigVis 2023, SDM 2023, DASFAA 2023, VLDB Demonstration 2022, SOSP Artifact Evaluation 2021

- Reviewer - WSDM Demonstration 2023, TKDE 2022
- Student Volunteer - SIGMOD 2023, ICDE 2023, VIS 2022

To the University

- MS Admissions Committee – Columbia University, 2022
- Graduate Application Mentor - Columbia University, 2020
- CS Department Representative - Engineering Graduate Student Council, Columbia University, 2018

TEACHING ASSISTANT

- COMS W6113 TOPICS IN DATABASE RESEARCH, Spring '23, Columbia University
- COMS W6998 Systems for Human Data Interaction, Spring '20, Columbia University
- COMS W4111 Introduction to Databases, Spring '19, Columbia University
- COMS W4231 Analysis of Algorithms, Fall '19, Columbia University
- ENGI E4900 Summer Masters Research, Summer '20, Columbia University
- COMS W3998 Undergraduate Research, Spring '20, Columbia University

MENTORSHIP

Master Student

- Jeffery Tao 2021.9 - 2023.7
Master Thesis: An Execution Engine for Physical Visualization Design
- Miles Hong 2022.2 - 2023.5
Master Thesis: Tutorial Structure: an Interface Model to Generate Interactive Tutorials
- Teng Jiang 2023.2 - 2023.5
- Xiaoyue Chen 2022.5 - 2022.9
- Alexander Yao 2022.2 - 2022.5

Undergraduate

- Janissa Abreu 2023.5 - 2023.8
- Ryan Li 2022.5 - 2022.9
- Austin Mac 2022.5 - 2022.9
- Brandon Zhang 2020.2 - 2020.5

TECHNICAL STRENGTHS

- Python, C++, SQL, Javascript, TypeScript, HTML/CSS
- Tensorflow, PyTorch