Databricks Inc.
Mountain View, California, CA
+1(510)2309932
yicong.huang@databricks.com
yicong-huang.github.io
linkedin.com/in/yicong-huang
github.com/Yicong-Huang
scholar.google.com/citations?user=9ZmdkkIAAAAJ

Yicong Huang

Curriculum Vitae

Education

2019–2025 **Ph.D., Computer Science**, *University of California, Irvine*, CA, United States.
Lab: Information Systems Group (ISG) Advisor: Dr. Chen Li GPA: 3.96/4.0

Research Interest: Data-Processing Systems, Database Management Systems (DBMS), Data-Intensive Scalable Computing (DISC), Machine Learning Systems.

Dissertation: *UDF-Centric Dataflow Systems for Supporting User-Defined Functions in Collaborative Data Science, AI, and ML.*

2015–2019 **B.S., Computer Science**, *University of California, Irvine*, CA, United States.

Employment

2025 – Pres. **Software Engineer**, *Databricks Inc.*, Mountain View, CA, United States.
Apache Spark, Spark Core, PySpark

2025 – Pres. **Adjunct Assistant Professor**, *CICS at UMass Amherst*, Amherst, MA, United States.

Summer 2024 **Software Engineer Intern**, *Observe Inc.*, San Mateo, CA, United States.
Dataset Transformers, Go, Snowflake

Summer 2022 **Research Intern**, *Visa Research, Visa Inc.*, Palo Alto, CA, United States.
SWAT-RT: Real-time Window Aggregation on Streaming Systems, C++, Redis, Kafka, Flink

Summer 2020 **Research Intern**, *Infrastructure System Lab, ByteDance Inc.*, Mountain View, CA, United States.
HTAP Database System, C++, Kudu, MySQL

Fellowships & Awards

2025 Recipient of ***Most Promising Future Faculty Second Place Award*** from University of California, Irvine.

2025 Recipient of ***Joseph & Dorothy Fischer Memorial Endowed Fellowship*** from University of California, Irvine.

2025 Recipient of ***Beall Family Foundation Graduate Student Entrepreneur Award in Computer Science*** from University of California, Irvine.

2024 Honored ***Best Demo Award Runner-Up Award*** at SIGMOD 2024.

2024 Recipient of ***Graduate Dean's Dissertation Fellowship*** from University of California, Irvine, in recognition of a highly impactful thesis.

2024 Awarded of ***Student Travel Awards*** to attend ***SIGMOD 2024***.

2023 Recipient of ***Public Impact Fellowship*** from University of California, Irvine, for conducting impactful research for the public.

2023 Awarded of ***Student Travel Awards*** to attend ***VLDB 2023***.

2020 Awarded of ***Best Lecturer Award*** for excellence in teaching at CUCS.

Computer Science Publications (Selected)

[1] **Yicong Huang**. "UDF-Centric Dataflow Systems for Supporting User-Defined Functions in Collaborative Data Science, AI, and ML". PhD thesis. 2025, p. 194. ISBN: 9798290963273.
URL: https://www.proquest.com/dissertations-theses/udf-centric-dataflow-systems-supporting-user/docview/3240377738/se-2.

[2] Mengying Wang, Moming Duan, **Yicong Huang**, Chen Li, Bingsheng He, and Yinghui Wu. "ML-Asset Management: Curation, Discovery, and Utilization". In: *Proc. **VLDB** Endow.* 18.12 (2025).
URL: https://www.vldb.org/pvldb/vol18/p5493-wang.pdf.

[3] Xiaozhen Liu, **Yicong Huang**, Xinyuan Lin, Avinash Kumar, Sadeem Alsudais, and Chen Li. "Pasta: A Cost-Based Optimizer for Generating Pipelining Schedules for Dataflow DAGs".
In: *Proc. ACM Manag. Data (**SIGMOD**)* 2.6 (2024), 248:1–248:26. DOI: 10.1145/3698832.
URL: https://doi.org/10.1145/3698832.

[4] Shengquan Ni, **Yicong Huang**, Zuozhi Wang, and Chen Li. "IcedTea: Efficient and Responsive Time-Travel Debugging in Dataflow Systems".
In: *Proc. **VLDB** Endow.* 18.3 (2024), pp. 902–914.
URL: https://www.vldb.org/pvldb/vol18/p902-ni.pdf.

[5] Alexander K. Taylor*, **Yicong Huang***, Junheng Hao, Xinyuan Lin, Xiusi Chen, Wei Wang, and Chen Li. "Data Science Tasks Implemented with Scripts versus GUI-Based Workflows: The Good, the Bad, and the Ugly". In: *40th International Conference on Data Engineering, **ICDE** 2024 - Dataplat Workshop*.
*The first two authors share equal contributions. IEEE, 2024, pp. 267–277.
DOI: 10.1109/ICDEW61823.2024.00040.

[6] Zuozhi Wang, **Yicong Huang**, Shengquan Ni, Avinash Kumar, Sadeem Alsudais, Xiaozhen Liu, Xinyuan Lin, Yunyan Ding, and Chen Li. "Texera: A System for Collaborative and Interactive Data Analytics Using Workflows".
In: *Proc. **VLDB** Endow.* 17.11 (2024), pp. 3580–3588.
URL: https://www.vldb.org/pvldb/vol17/p3580-wang.pdf.

[7] **Yicong Huang**, Zuozhi Wang, and Chen Li. "Demonstration of Udon: Line-by-line Debugging of User-Defined Functions in Data Workflows".
In: *Companion of the 2024 International Conference on Management of Data, **SIGMOD/PODS** 2024, Santiago AA, Chile, June 9-15, 2024*. **Best Demo Runner-Up Award**. ACM, 2024, pp. 476–479.
DOI: 10.1145/3626246.3654756. URL: https://doi.org/10.1145/3626246.3654756.

[8] **Yicong Huang**, Zuozhi Wang, and Chen Li. "Udon: Efficient Debugging of User-Defined Functions in Big Data Systems with Line-by-Line Control".
In: *Proc. ACM **SIGMOD*** 1.4 (2023), 225:1–225:26. DOI: 10.1145/3626712.
URL: https://doi.org/10.1145/3626712.

[9] Avinash Kumar, Sadeem Alsudais, Shengquan Ni, Zuozhi Wang, **Yicong Huang**, and Chen Li. "Reshape: Adaptive Result-aware Skew Handling for Exploratory Analysis on Big Data".
In: *CoRR* abs/2208.13143 (2022). DOI: 10.48550/arXiv.2208.13143. arXiv: 2208.13143.
URL: https://doi.org/10.48550/arXiv.2208.13143.

[10] Xiaozhen Liu, Zuozhi Wang, Shengquan Ni, Sadeem Alsudais, **Yicong Huang**, Avinash Kumar, and Chen Li. "Demonstration of Collaborative and Interactive Workflow-Based Data Analytics in Texera".
In: *Proc. **VLDB** Endow.* 15.12 (2022), pp. 3738–3741. DOI: 10.14778/3554821.3554888.
URL: https://www.vldb.org/pvldb/vol15/p3738-liu.pdf.

[11] Zhihui Yang, Zuozhi Wang, **Yicong Huang**, Yao Lu, Chen Li, and X. Sean Wang. "Optimizing Machine Learning Inference Queries with Correlative Proxy Models".
In: *Proc. **VLDB** Endow.* 15.10 (2022), pp. 2032–2044. DOI: 10.14778/3547305.3547310.
URL: https://www.vldb.org/pvldb/vol15/p2032-yang.pdf.

[12] Zhihui Yang, **Yicong Huang**, Zuozhi Wang, Feng Gao, Yao Lu, Chen Li, and X. Sean Wang. "Demonstration of Accelerating Machine Learning Inference Queries with Correlative Proxy Models". In: *Proc. **VLDB** Endow.* 15.12 (2022), pp. 3734–3737. DOI: 10.14778/3554821.3554887. URL: https://www.vldb.org/pvldb/vol15/p3734-yang.pdf.

Interdisciplinary Publications (Selected)

[1] Jiadong Bai, Xiaozhen Liu, Anthony Cuturrufo, Alexander Kundu Taylor, Jeehyun Hwang, Mingyu Derek Ma, Xinyuan Lin, Yanqiao Zhu, **Yicong Huang**, Yunyan Ding, Wei Wang, and Chen Li. "DS4ALL: Teaching High-School Students Data Science and AI/ML Using the Texera Workflow Platform as a Service". In: *Data Science Education K-12: Research to Practice Annual Conference*. Feb. 2025. URL: https://web.cvent.com/event/d641bd9f-6c99-4cbc-951b-33b1ca05d4ed/summary.

[2] Judith Borghouts, **Yicong Huang**, Suellen Hopfer, Chen Li, and Gloria Mark. "Wording Matters: the Effect of Linguistic Characteristics and Political Ideology on Resharing of COVID-19 Vaccine Tweets". In: *Transactions on Computer-Human Interaction (TOCHI)* (2024).

[3] Yunyan Ding, **Yicong Huang**, Pan Gao, Andy Thai, Atchuth Naveen Chilaparasetti, M Gopi, Xiangmin Xu, and Chen Li. "Brain image data processing using collaborative data workflows on Texera". In: *Frontiers in Neural Circuits* 18 (2024), p. 1398884.

[4] Jessie WY Ko, Shengquan Ni, Alexander Taylor, Xiusi Chen, **Yicong Huang**, Avinash Kumar, Sadeem Alsudais, Zuozhi Wang, Xiaozhen Liu, Wei Wang, et al. "How the experience of California wildfires shape Twitter climate change framings". In: *Climatic Change* 177.1 (2024), pp. 1–21.

[5] Judith Borghouts, **Yicong Huang**, Sydney Gibbs, Suellen Hopfer, Chen Li, and Gloria Mark. "Understanding underlying moral values and language use of COVID-19 vaccine attitudes on twitter". In: *PNAS nexus* 2.3 (2023), pgad013.

[6] Joshua Rhee, **Yicong Huang**, Sadeem Alsudais, Shengquan Ni, Avinash Kumar, Chen Li, and David Timberlake. "The marketing and perceptions of non-tobacco blunt wraps on Twitter". In: *Substance Use and Misuse* (2023).

[7] Yawen Guo, Jun Zhu, **Yicong Huang**, Lu He, Changyang He, Chen Li, and Kai Zheng. "Public Opinions toward COVID-19 Vaccine Mandates: A Machine Learning-based Analysis of U.S. Tweets". In: *AMIA 2022, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 5-9, 2022*. AMIA, 2022. URL: https://knowledge.amia.org/76677-amia-1.4637602/f006-1.4642154/f006-1.4642155/516-1.4642396/1066-1.4642393.

[8] Zimu Wang*, **Yicong Huang***, Wanjun Lu, Jiaxin Liu, Xinying Li, Suhua Zhu, Hongbing Liu, and Yong Song. "c-myc-mediated upregulation of NAT10 facilitates tumor development via cell cycle regulation in non-small cell lung cancer". In: *Medical Oncology* 39.10 (2022). *The first two authors share equal contributions, p. 140.

[9] Lu He, Changyang He, Tera L. Reynolds, Qiushi Bai, **Yicong Huang**, Chen Li, Kai Zheng, and Yunan Chen. "Why do people oppose mask wearing? A comprehensive analysis of U.S. tweets during the COVID-19 pandemic". In: *J. Am. Medical Informatics Assoc.* 28.7 (2021), pp. 1564–1573. DOI: 10.1093/jamia/ocab047. URL: https://doi.org/10.1093/jamia/ocab047.

[10] Suellen Hopfer, Emilia J Fields, Yuwen Lu, Ganesh Ramakrishnan, Ted Grover, Quishi Bai, **Yicong Huang**, Chen Li, and Gloria Mark. "The social amplification and attenuation of COVID-19 risk perception shaping mask wearing behavior: a longitudinal twitter analysis". In: *PloS one* 16.9 (2021), e0257428.

Research Experiences

2020 – Pres. ***Texera***, *An open-source system for cloud-based collaborative data science and AI/ML.*
Scala, Typescript, Python, Arrow

- Leading the team effort on designing the system from all layers, including the distributed engine, compiler, scheduler, etc.
- Leading the effort on interactive debugging of Python UDF during the runtime of a workflow.
- Designed and implemented the Python processing engine on top of Akka Actor system, targeting PySpark & PyFlink.
- Intergarted ML and AI to optimize the workflow runtime.
- Contributed in exploration of fault tolerence, version control, resource management and other aspects of the system.
- Maintaining a live service at https://hub.texera.io.
- Research and demo papers accepted by SIGMOD 2024, VLDB 2024, ICDE 2024, SIGMOD 2025, VLDB 2025.

2019 – 2020 ***Cloudberry***, *A middle-ware system for large scale data visualization.*
Scala, Javascript, AWS

- Conducted tweet visualization with an interactive map, aggregating and displaying 4TB data.
- Integrated COVID-data with social media data on the interactive map.
- Built a fully scalable elastic service that can be load balanced on 20+ AWS machines, CoronavirusTwittermap.

2019 – 2020 ***ML-OPT***, *Machine Learning Pipeline Optimization.*
Python

- Explored using Machine Learning models to optimize Machine Learning Pipelines with a confidence guarantee.
- Conducted optimization on video recognition models (e.g., YOLOv3) by 25% with 98% accuracy guarantee, and NLP models (e.g., StanfordNLP) by 45% with 98% accuracy guarantee.
- Full research paper and demo accepted by VLDB 2022.

2019 ***Wildfire***, *Wildfire Detection & Visualization with Social Media.*
Python, Typescript, PostgreSQL

- Led team of 10 masters and undergraduates in detecting wildfires based on tweets and satellite data.
- Built data collection pipelines for real-time tweets, satellite data from NOAA, fire reports from USGS, etc.
- Integrated Machine Learning models such as AllenNLP, StanfordNLP, CNN, RNN, ANN for semantic analysis on text; ResNet50, VGG for images classifications.

2018 ***Blockchain***, *Blockchain in Fin-tech & Blockchain and Smart Contracts.*
Python, C++

- Reviewed over 200 white papers on smart contracts implemented on blockchains such as Ethereum and Bitcoin.
- Developed a prototype smart contract on the Ethereum blockchain.
- Organized the California-Shanghai Innovation Dialogue Conference in 2018.

Contributions to Grants Writing

2024 **dkNET Coordinating Unit: Harnessing the Power of AI and Data Science for Collaborative Discovery and Sharing in the DK Community**.
National Institutes of Health (NIH), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)

- Contact PI: Jeffrey S. Grethe
- Other PIs: Shuibing Chen, Chen Li, Wei Wang
- Period: 2024 – 2029
- Funding: total $10M for 5 years

My Role: Contributed to Aim 3: Computational Core: Develop and deploy unified ML models and workflows.

2022 **Collaborative Research: Collaborative Machine-Learning-Centric Data Analytics at Scale**.
National Science Foundation (NSF)

- Contact PI: Chen Li
- Other PIs: Suellen Hopfer
- Period: 2021 – 2024
- Funding: $913,992

My Role: Contributed to Aim 2: Supporting Debugging of External UDF's during Execution.

2020 **RAPID: Leveraging Twitter Data for Real-time Public Health Responses to Coronavirus: Identifying Affective Desensitization, Loneliness and Depression, and Trust in Message Sources and Content**.
National Science Foundation (NSF)

- Contact PI: Gloria Mark
- Other PIs: Chen Li, Suellen Hopfer
- Period: 2020
- Funding: $180,000

My Role: Collected data and contributed to the writing of preliminary results.

Teaching Experiences

2024 **Workshop Instructor**.
Cerritos College, Norwalk, CA, United States

Fall 2024 *Workshop of Data Science for Everyone 2024* Workshop Homepage

- Hosted a two-day workshop designed for non-CS students to learn data science and ML without a single line of coding.
- More than 50 community college students and faculty members participated.

2024 **Associate Instructor (Lecturer)**.
University of California, Irvine, CA, United States

Spring 2024 *ICS 80: Data Science and AI/ML Using Workflows* Syllabus

- Designed and taught a new course that helped non-CS students to gain knowledge of data science, AI, and ML in a short period.
- 42 undergraduate students enrolled.

2019 — Pres. **Research Mentor**.
University of California, Irvine, CA, United States

PhD students:
- **2023-2024**: Raj Mohanty (PhD), Jiadong Bai (PhD), Shagoto Rahman Shrestho (PhD);
- **2022-2023**: Xinyuan Lin (PhD), Yunyan Ding (PhD);

Master students:
- **2022-2023**: Aditya Verma (MS), Sreetej Reddy (MS), Dhruv Raipure (MS), Jiaxi Chen (MS);
- **2019-2020**: Yang Cao (MS);

Undergraduate students:
- **2023-2024**: Kevin Wu;
- **2022-2023**: Chengxi Li (MS), Ethan Wong (MS), Tianyun Yuan (MS), Tony Liu (MS);
- **2021-2022**: Zhen Guan – UCSD (MS), Jiashu Zhang – Hong Kong Polytechnic University (PhD), Yinan Zhou – UCI (PhD), Andrew Li (MS), Eric Peng (MS), Jiyang Wu (MS), Zeyu Li (MS);
- **2020-2021**: Chen He – CMU (MS), Bihao Xu – Uchicago (MS), Conghuai Tan (MS), Make Tao (MS), Mingshuo Liu (MS), Qifan Yu (MS);
- **2019-2020**: Dayue Bai – UIUC (MS), Yinan Zhou – UCI (MS), Shiqi Wu – Berkeley (MS), Christine Xinrong Huang – CMU (MS), Tianran Liu – Univ of Washington (MS), Yutong Wang – UCD (PhD), Tingxuan Gu – CMU (MS), Yichi Zhang – NYU (MS), Xinyue Han – UCLA (MS), Qiaonan Huang (Hugo) – Brown (MS), Yuan Fu – CMU (SE-SV), Yuqi Huai – UCI (PhD), Quanzhen Du – UCSD (MS), Shiling (Scarlett) Zhang – Cornell (MS), Zeyad Kelani (MS);

2018 — 2022 **Teaching Assistant**.
University of California, Irvine, CA, United States

- W'22 *CS 222/122C: Principles of Data Management*
- F'21 *CS 122B: Projects in Databases and Web Applications*
- S'21 *CS 122B: Projects in Databases and Web Applications*
- W'21 *ICS 51: Introduction to Computer Organization*
- F'20 *CS 222/122C: Principles of Data Management*
- S'20 *CS 122B: Projects in Databases and Web Applications*
- W'20 *CS 222/122C: Principles of Data Management*
- F'19 *CS 222/122C: Principles of Data Management*
- S'19 *CS 122B: Projects in Databases and Web Applications*
- W'19 *CS 122B: Projects in Databases and Web Applications*
- F'18 *CS 141: Concepts of Programming Languages I*
- S'18 *CS 122B: Projects in Databases and Web Applications*

Winter 2018 **Mentor**.
Dreams for Schools APPJAM+, Yorba Linda High School, Yorba Linda, CA, United States

---

2016 — 2018 **Tutor**.
University of California, Irvine, CA, United States

- W'18 *ICS 46: Data Structure Implementation and Analysis*
- F'17 *ICS 45J: Programming in Java as a Second Language*
- S'17 *ICS 33: Intermediate Programming*
- W'17 *ICS 32: Programming with Software Libraries*
- F'16 *ICS 31, Introduction to Programming*

Media & Press (Selected)

Talks

2025 **UDF-Centric Dataflow Systems for Supporting User-Defined Functions in Collaborative Data Science, AI, and ML**.
My final defense at University of California, Irvine.

2025 **Building a Collaborative and Interactive Data System to Broaden Access to Data Science, AI, and ML**.
University at Buffalo.
Lehigh University.
University of Massachusetts Amherst.
Purdue University.
University of Massachusetts Boston.
Amazon Web Services.
Arizona State University.
Santa Clara University.

2024 **Texera and Interactive Debugging in Distributed Data Processing Systems**.
An invited talk at Databricks Inc., Mountain View, CA, United States.

2024 **Texera: A System for Collaborative and Interactive Data Analytics Using Workflows**.
A paper presentation talk at VLDB 2024, Guangzhou, China

2024 **The Journey to Build Texera**.
An invited talk at ISG Reunion, Irvine, CA, United States.

2024 **Udon: Efficient Debugging of User-Defined Functions in Big Data Systems**.
A paper presentation talk at SIGMOD 2024, Santiago, Chile.

2023 **Python User-defined Functions on Workflows**.
A guest lecture at DS4ALL 2023 summer program, University of california, Irvine.

2022 **Data Challenges in Streaming Systems**.
A talk at VISA Research, AI Research team, Palo Alto, CA, United States.

Blogs

2024 **Suggesting Python Code Type Annotations with LLM – A Language Server Enhancement**.
Texera Blog, Minchong Wu and Yicong Huang

2024 **Adding R UDF to Texera: The Journey**.
Texera Blog, Kevin Wu and Yicong Huang

2023 **Enhancing the UDF Editor by Adding Language Server Support**.
Texera Blog, Aditya Verma, Dhruv Raipure, Jiaxi Chen, Sreetej Reddy, and Yicong Huang

2023 **Using Texera to Perform Single-cell RNA Sequencing Analysis with R Language**.
Texera Blog, Yicong Huang

Services

Conference Officer

2027 **Web Chair**, *ACM SIGMOD International Conference on Management of Data*.
Huntington Beach, CA, United States

Reviewer

2024 **External Reviwer**, *IEEE Transactions on Knowledge and Data Engineering (TKDE)*.

2024 **Program Committee**, *ACM SIGMOD International Conference on Management of Data - Artifact Review and Reproducibility (ARI)*.

Student Volunteer

2023 **ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)**.
Long Beach, CA, United States

2023 **IEEE International Conference on Data Engineering (ICDE)**.
Anaheim, CA, United States

2019 **International Conference on Very Large Data Bases (VLDB)**.
Los Angeles, CA, United States

2018 **California-Shanghai Innovation Dialogue Conference**.
Irvine, CA, United States

Computer skills

Proficient	Python, Java/Scala, C/C++, JavaScript/TypeScript, Golang, Lisp, Prolog, MySQL, PostgreSQL
Intermediate	HTML, CSS, SQL, R, Redis
Programming Concepts	MapReduce, OOD, Functional, Logical Programming, Model-View-Control, Multithreading
Frameworks	Arrow, Hadoop, Spark, Flink, Kudu, Protobuf, Spring, Angular 2+, Django, Flask, Express, ReactJS, Elasticsearch
Services	AWS, GCP, GitHub, Docker