

Teaching and Service Statement

Eugene Wu, Columbia University

Teaching and Mentorship

One of the best parts of being a professor is the long-term opportunity to teach, and to see students from different backgrounds succeed towards their personal goals. I believe that students are capable of learning the material, and that our job is to provide motivation, remove stumbling blocks, and support students toward their goals. This defines my approach to teaching in the classroom and mentoring students in my lab. Every semester, I improve my courses based on past feedback and challenges that I have seen students face.

Motivation: Computer science topics can often feel dry and divorced from practice. To motivate students, I try to help them *feel* the value of each topic—both by experiencing the difficulties of solving problems without the appropriate computational tools, and by celebrating the amazing applications and potential that the lecture topics enable. For instance, when I present entity-relationship models in Introduction to Databases, I first lead the class in collectively designing the data model for an application that the students pick. They often point out the difficulty of not having a common modeling language, which directly segues into the lecture topic. Similarly, when discussing database joins, I emphasize that real life concepts such as high-fives, markets, and baking can be viewed as instances of joins. My goal is for students to see that each topic has a purpose and is not a mere academic exercise.

Diverse Backgrounds: Computer science courses draw students with a wide range of technical backgrounds, from expert programmers to curious novices. It is easy for weaker students to fall behind and advanced students to become disinterested. Thus, I position the lectures towards the weaker students, and regularly reference recent research or advanced concepts. Educational research suggests that deliberate and repeated practice is key to knowledge acquisition, so I created websites that auto-generate practice problems so students can practice on their own, and at their own pace. For advanced or more interested students, I wrote a full-featured Python database called DataBass to show how different parts of a database work together, and created extra-credit assignments that can be completed in lieu of regular homeworks. The extra-credit aspect helps relieve them of the feeling that they need to complete them to “get ahead”.

Similarly, in the Data Science Institute (DSI) masters course Big Data Systems (W4121), I encouraged students that want to go beyond the course content to pursue semester-long research projects that incorporate concepts discussed in the course. To do so, I provide extra-credit opportunities for student teams, and personally work with those teams throughout the semester to come up with, and refine, their projects.

Stress: Competition within courses is known to be stressful at Columbia, and there is a perception that grades are zero-sum. I try to provide counteracting incentives. On exams, students are encouraged to “phone-a-friend” for one of the problems, and receive the maximum of their friend and their own grade. This tries to encourage students to meet their colleagues and study together. I also plot the “phone-a-friend” social network for each exam and remark to the students the tangible benefits of studying with each other. I also use extra-credit opportunities so students can show their command of the material beyond assignments and exams. These include optional advanced assignments, opportunities to improve the contents of a shared course wiki, and sharing creative interpretations of course topics.

Undergraduate Mentorship: I have worked with over 20 undergraduate students from Columbia and other universities. I endeavor to help them independently define and lead their own ideas: Hamed Nilforoshan led a project to improve writing quality on social media services, published an ICWSM paper, received a CRA honorable mention, and has started his Ph.D. at Stanford; Kevin Lin initiated a ML debugging project that led to a NeurIPS workshop paper and a SIGMOD paper, and has started a Ph.D. at UC Berkeley; Robert Netzorg

led a project to understand the relationships between behavior and popularity on live-streaming services such as Twitch, submitted an ICWSM paper, and has started a Ph.D. at UC Berkeley; Lauren Arnett co-led the Twitch project, and was admitted to UMass Amherst's Ph.D. program. Sagar Lal, Maneet Khaira, and Rodolfo Raimundo all defined their own areas within the Twitch project. Jacob Fisher currently leads a project to automatically rescale chart axes for interactive visualizations.

Teaching Experience

I regularly teach **Introduction to Databases** (W4111), with 100-130 students in each section. In 2015, I overhauled the content to integrate modern data analysis, and created new lectures, assignments, projects, and exercises; in 2020, I worked with the database faculty to overhaul the syllabus for the first time in a decade. I regularly experiment with optional tracks for the advanced students, extra credit to encourage collaboration and relieve the competitive atmosphere, the use of online programming notebooks to remove confusing configuration and setup issues, and auto-generated exercises to provide practice. The course materials have been adopted by the database course at NorthWestern University and the other 4111 sections at Columbia.

I taught **Systems for Data Science** (W4121) from 2015 to 2018, and modernized the course by developing new lectures and assignments that incorporate modern data processing, visualization, and analysis. I also worked with advanced students on optional research projects. As I describe below (DSI Masters Curriculum), the course is challenging to teach because it is co-taught by 3 instructors. I proposed a curriculum redesign in Fall 2018, and volunteered to teach an extra course to fill gaps in the course offering during its transition.

I created new seminars. **Interactive Data Exploration Systems** (W6995.002) dives into database research and human computer interaction. The content incorporated my visualization research, and the course projects led to InfoVIS and VLDB publications. **Database Topics in Research & Practice** (W6995.005) focuses on landmark database research that transitioned into startups or open source, and I invited 11 speakers to share their first-hand experiences. **Database Research Topics** (W6113) fills gaps in the current DB curricula by covering classic and modern database systems research. For 6113, I extended the W4111 DataBass system to create new assignments that incorporate research ideas such as query compilation and optimization.

I also recruited Agnes Chang, who won the 2019 Pulitzer prize for national reporting, to create and teach the first Introduction to Visualization course in the CS department. She has taught the 40-person course regularly since Fall 2018; it is hugely popular, in its 5th offering, and attracts students across the university.

Student Comments:

For W4111: *"One of the best professors at Columbia. Really helping students understand the materials", "has a great ability to make the knowledge clear to us, he is very creative and can always come up with good examples to make the problem easy to understand."*

For DataBass: *"I learned more things . . . than just from reading textbook", "As a non-cs background student, I learn so much from this assignment. Thx :D!"*

For W4121: *"Professor Wu was extremely prepared and effective during lectures", ". . . genuinely invested in presenting his topics in a way that everyone can understand despite how complicated it can get".*

Service

DSI Center Co-chair: Since Fall of 2016, I have been Co-Chair¹ of the Data Science Institute's Center for Data, Media and Society. I have worked to expand the visibility and collaboration between the engineering-heavy Data Science Institute with the humanities departments. In 2016, we changed the name and focus of the center from "Center for New Media" to "Center for Data, Media & Society", to increase the inclusivity with

¹Along with Laura Kurgan (GSAPP 2016–2017), and Susan McGregor (Journalism, DSI 2017–now).

university departments (especially humanities). I have also organized numerous cross-discipline workshops, poster sessions, and speaker seminars to bring together members across the university. Further, in 2018, we initiated a Scholars program, which has funded several dozen undergraduate and masters researchers to support projects that span data science and the humanities. Each project combines PIs from data science and humanities departments, such as Journalism, English, Social Science, the Earth Institute, History, and more.

DSI Center Committee: I served as committee member of the Center for Computing Systems for Data-Driven Science from its inception in 2017 as a DSI working group to its addition as an official center in the Data Science Institute at the end of 2018. The center aims to serve as the University's intellectual locus for massive-scale systems research and applications. I helped identify collaboration opportunities with Lamont and Brookhaven National Laboratories.

DSI Masters Curriculum: In Fall 2018, I led the redesign of the systems track curriculum for the Data Science Institute's masters program. The existing curriculum (W4121) received consistently poor reviews because, for legacy reasons, it was co-taught and lacked ownership. I worked with the CS Systems faculty to draft a new, more sustainable, curriculum, and then worked with the DSI education team to make the transition in Spring and Fall 2019. I taught an extra section of Intro to Databases (W4111) in Spring 2019 to aid the transition.

CS-Journalism Dual Degree: Since Fall 2016, I have been the CS chair for the CS-Journalism dual degree masters program. This is unique across the world, in that it trains students to be deeply versed in both CS as well as journalism, and ultimately define the future of media. In this capacity, I have worked with my journalism counterpart Susan McGregor to advise the students, formulate outreach and recruiting initiatives, and define the program's CS curriculum.

CS Graduate Admissions: I served on the Computer Science Graduate Admissions Committee from Fall 2015 to 2020. Each spring, it assesses over 1500 Ph.D. candidates to identify the 10-20 strongest and most promising applicants that should be nominated for the prestigious Presidential Fellowship.

Database Workshop Chairing: I have co-created and/or co-chaired numerous workshops in the database research community. In 2018, I co-created the New York Database Day workshop, which brings together academic and industry data management researchers across the wider New York area. The workshop was sponsored by Google and DSI, and will be an annual event. I also co-chaired the Human-in-the-Loop Data Analytics (HILDA) workshop co-located at SIGMOD. It sits squarely in my research area, between HCI, Visualization, Data Mining, and Databases, and had nearly 100 attendees. It was the third largest workshop at SIGMOD, after the decades-old workshop DaMoN, and the new machine learning workshop DEEM.

Database Service: I have served to increase student participation at the major database conference, SIGMOD. This has included co-chairing the SIGMOD Travel Award Committee (2016, 2017), which solicits funding to sponsor students that cannot otherwise afford to attend the conference, and the SIGMOD New Researcher Symposium (2017, 2018), which encourages and connects junior researchers in their Ph.D. or early careers with more senior researchers. I also co-chaired the SIGMOD Student Research Competition (2019, 2020), which highlights and promotes undergraduate and graduate student research, and is part of the broader ACM student research competition. I serve as associate editor of ICDE 2017, VLDB 2021, and VLDB 2022, and am on the program committees for VLDB, SIGMOD, ICDE, HILDA, WWW, HCOMP.

Outreach: I've mentored the Columbia Egleston Scholar's program since 2016, and have mentored two high school students. In 2017, I participated in Columbia Engineering's Inside Engineering program, which brought elementary and middle school students to Columbia. I regularly give talks at companies, such as 2Sigma, BlueCore, Google, AT&T research, and IBM. My research ideas such as query explanation has been used or adopted by several companies, including Tamr, and Honeycomb.