HTTP://DASLAB.SEAS.HARVARD.EDU/CLASSES/CS265/

CS 265 SPRING 2019 SYLLABUS

Prof: Stratos Idreos (stratos@seas.harvard.edu), MD139

After taking CS265 you will be able to understand big data system internals for SQL, NoSQL and Machine Learning and have a handle on what it means to do CS systems research.

WHAT IS THIS CLASS ABOUT?

Big data is everywhere. A fundamental goal across numerous modern businesses and sciences is to be able to utilize as many machines as possible, to consume as much information as possible and as fast as possible. The big challenge is "how to turn data into useful knowledge". This is a moving target as both the underlying hardware and our ability to collect data evolve. In this class, **we will discuss how to design data systems, data structures and algorithms** for key data-driven areas, including **relational systems, distributed systems, graph systems, noSQL, newSQL, machine learning and neural networks**. We will see how they all rely on the same set of very basic concepts and we will learn how to synthesize efficient solutions for any problem across these areas using those basic concepts.

WHAT IS A DATA SYSTEM?

Data systems are literally everywhere. We are using them directly or indirectly every day all day long for numerous basic or not so basic tasks, e.g., when we are buying coffee to when we are booking airplane tickets or training a neural network. They provide the backbone of all modern businesses to manage their data and of course they provide the backbone of online businesses and environments such as social networks and search engines. They are also used increasingly in science as data analytics becomes more and more the fundamental barrier in generating knowledge.

WHAT IS THIS CLASS NOT ABOUT?

This class is not a traditional introduction on how we use a database system and how to write SQL. Instead, this is a systems class about data system design. You will learn how big data systems work at their core and how to design new systems for emerging data-driven applications and hardware. By the way, if you know how systems work, you also become better at using them!

WHY TAKE THIS CLASS?

Data is everywhere. Every year we create even more data. As it stands, every two days we create as much data as much we created from the dawn of humanity up to 2003 [Eric Schmidt, Google]. Sciences, businesses, and everyday life are substantially affected. Data systems are in the middle of all this. Data systems are how we store and access data, i.e., they are the backbone of any data-driven application. It is a $100B industry, growing 10% every year [Economist, “Data, data everywhere”].

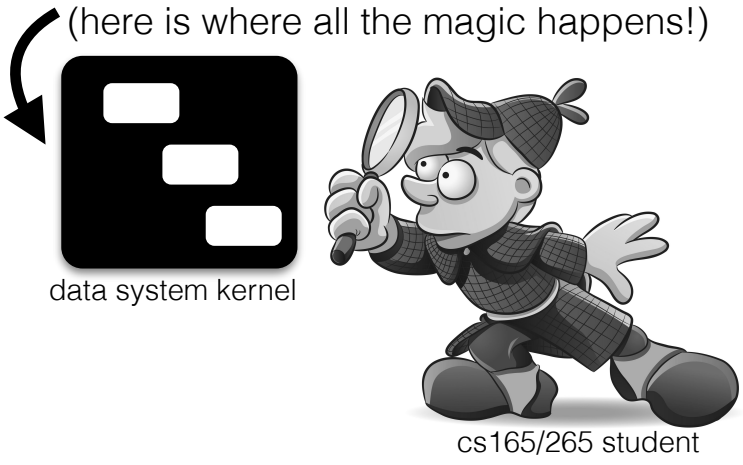
At the same time data systems research and the whole industry are going through a major and continuous transition; given that new data-driven scenarios and applications continuously pop up, there is a continuous need to redefine what is a good data system design in such a dynamic environment.

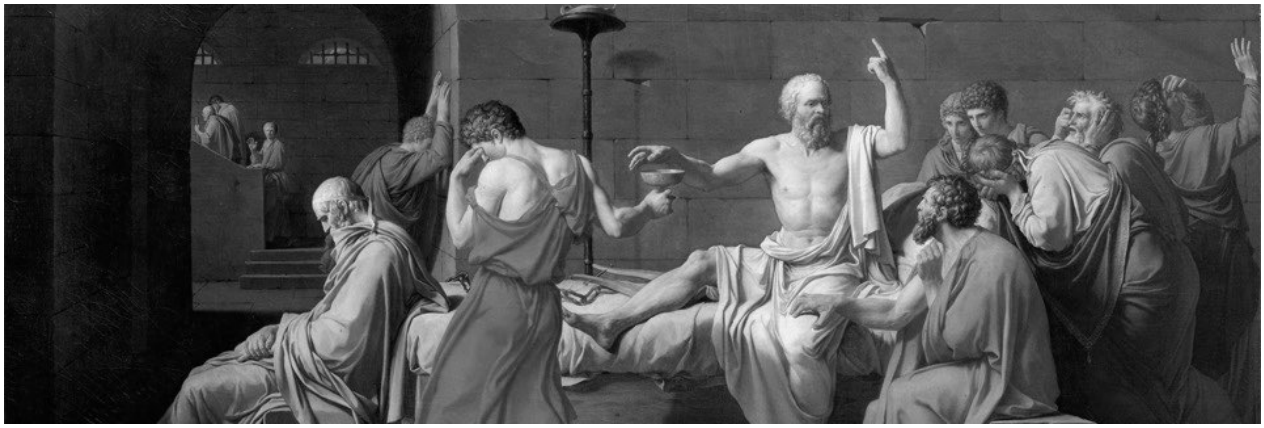
CS265 exposes students to the core internals of data systems making it possible to understand core trends in system design and to be one of the few who know how to design and evaluate systems. In addition, due to the way the course is taught (focus on interactive problem solving, open topics and the latest research results) this is also a great class for those who want to understand what CS research is all about and how to engage in doing research.

WHAT IS THE EXPECTED LEARNING OUTCOME?

1. Learn state-of-the-art research and industry trends in big data systems.
2. Understand the tradeoffs in designing and implementing modern big data systems.
3. Be able to make design decisions in big data driven scenarios.
4. Develop basic research skills: reading, writing and understanding research papers.
5. Deepen C programming, debugging, and performance profiling skills.

Efficient data analytics and system design is all about how we store and access the data. In this class, you are going to see how the same concepts appear again and again in numerous data-driven scenarios from NoSQL to neural networks.





CLASS PHILOSOPHY

CS265 has unlimited office hours, unlimited late days for project deliverables, relies on the latest research papers instead of a standard text book, lectures are based on interaction and discussion instead of just “lecturing”, many of the quizzes and problem sets are actually open research problems and most of all it is fun! The instructor and the TFs are here to help you every day and at all times throughout the semester. You may request as many meetings as you like and as much help as you want.

The course is also geared towards engaging creative thinking and problem solving to give students a feeling of how computer science research takes place. Many of our students in the past have successfully engaged in research projects with DASlab and published research papers.

HOW DOES 265 COMPARE TO 165?

If you took and liked CS165, you will like CS265 as well. From a material point of view CS265 moves on to consider additional topics as a continuation of CS165 such as distributed processing, transaction processing, graph processing, machine learning and more. In terms of the way the class is taught, it is even more interactive, and even more research oriented. Semester projects are actually on open research problems with the potential to lead to a publication and every class is focused on a single research paper, and understanding it in detail.

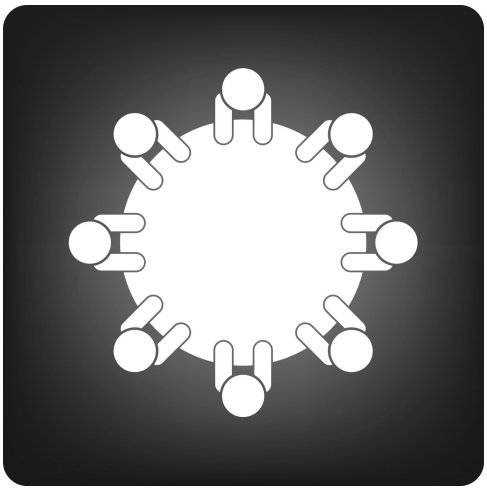
HOW MUCH WORK IS IT?

You may have heard stories about CS165 and wondering if CS265 is going to be equally hard or you may have taken CS165 and wondering if this is going to be a similar amount of work. CS165 and CS265 are different style of classes. While CS165 is much more focused on implementation leading to a full system prototype, CS265 is more focused on ideas and design. In other words, you may have written 5-10K lines of code (some even more!) for CS165 but in CS265 you are more likely going to write small amounts of code and mostly play with alternative ways to design a specific

functionality, structure or algorithm to highlight the effect of different choices and to find out new ways to solve a specific problem.

LECTURES

The class meets twice a week.



INTERACTION IN EVERY CLASS

While the instructor will do a few lectures through the semester, the class is going to be primarily discussion based. Think of this as an extended brainstorming session, a round table discussion about a specific problem in each class. The goal is to create the maximum possible interaction.

Our discussion will aim at bringing up design trends and tradeoffs, as well as algorithmic issues. Another significant part of our discussions will focus on examining open problems and to highlight opportunities for innovation.

At the very beginning of the semester the instructor will do 4-5 lectures to provide the necessary background. After that, each class will be based on a student presentation about a recent research paper which will work as a trigger for the day's brainstorming. Depending on the needs of the class, the instructor will do additional lectures during class time or during our extra research sessions.

OFFICE HOURS & LABS

Interaction does not stop at lecture time. CS265 is designed to maximize interaction as we truly believe this is the best way to learn; we offer daily office hours and labs.

Starting Week 1, Prof. Idreos will hold office hours during the week and additional OH will be offered periodically during the weekend. Labs are offered by the TFs. Rooms and slots: TBA. The goal of labs is to get hands-on help for the projects (coding). Bring your laptop and your questions about specific project parts you need help with. Labs are the place to go when you have a persistent bug, when you need help with a specific tool for the project (e.g., for debugging or performance testing) or to get feedback about the quality of your coding.

ATTENDANCE

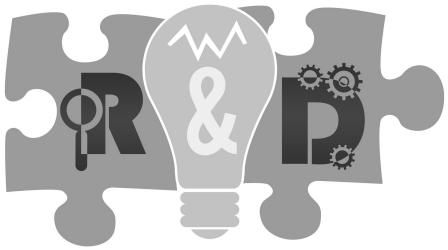
Based on the philosophy of the course, attendance in lectures, labs and office hours is optional. The best way to learn, though, is through discussion and interaction with the instructor and the TFs. Our classes are not about "lecturing" – they are about interaction. We hope to see you there!

CLASS RECORDINGS

All classes and interactive sessions in class will be recorded and will be available online. So even if you miss a class it will be easy to catch up and you can also use these recordings to recite specific material throughout the semester (e.g., to prepare for midterms).

SECTIONS

Another component of the course is sections. Sections are used to deliver material about the class, i.e., to go more deeply into some of the concepts discussed in class, to do additional quizzes, or to deliver background material that is needed to follow next week's class or for the project. There will be no actual section meeting. Instead, all sections will be recored by the teaching staff and videos will be posted online. The material posted will be tailored to present a step by step guide for any of the topics presented to make it easy to follow everything without having to be physically present in an actual section. However, if there are still questions about the material presented in sections, you will be able to ask those questions either during the daily office hours and labs.



RESEARCH SESSIONS

Throughout the semester, on select days the instructor, and DASlab PhDs and postdocs, will discuss about research! First, DASlab researchers will present their recent work on data systems research and connect it with the material you are learning in class. Then, you will get the chance to talk with them about their research, open problems and be exposed to open research opportunities. Snacks and drinks will be provided.

WEEKLY REVIEWS

Each student will provide two paper reviews per week. This prepares you to be ready for the discussion in class. Reviews should be no more than two page long.
Each review should have text for at least the following 9 points:

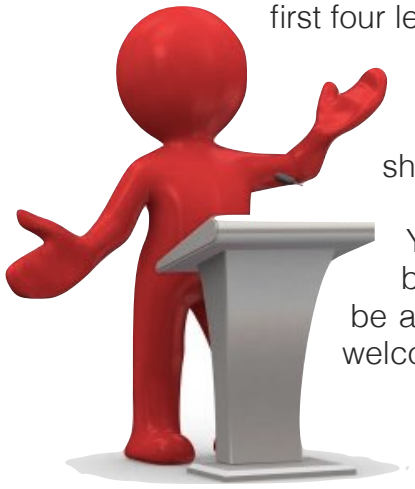
1) what is the problem?
2) why is it important?
3) why is it hard?
4) why existing solutions do not work?
5) what is the core intuition for the solution?
6) does the paper prove its claims?
7) what is the setup of analysis/experiments? is it sufficient?
8) are there any gaps in the logic/proof?
9) describe at least one possible next step.

Reviews should be no more than two pages long. PDF. Single column. 8pt font. 1 inch margins. Submission will be through Canvas. The deadline for each paper review is the starting time of the respective class. This is a hard deadline. The first four reviews will not be graded; we will use them only to provide feedback on the quality of the review and the grade that this review would get. Every second week we will have a special OH meeting to review the student reviews.

PRESENTATIONS

Each student will do at least one paper presentation during the semester. Presentations should follow similar guidelines as the guidelines for reviews. There should be 1-2 slides for each one of the nine core points in the review guidelines. In addition, there should be detailed slides that describe the core idea of the paper.

Your slides should not be a multiple sheets of bullet lists - in fact try to avoid bullet lists altogether - your slides should follow the generic formatting you will see in the first four lectures, that is: make slides as simple as possible - avoid text unless absolutely needed - no full phrases unless you need to give an exact definition of something - use figures and visual examples, one slide one message=each slide should have a single goal that you should be able to describe within a single phrase.



Your slides should be reviewed by the instructor at least 24 hours before the class you are presenting. The final deck of slides should be available 30 minutes before class so we can upload online. You are welcome to join for OH for help >>1 while you prepare your slides!

WHO CAN TAKE THIS CLASS?

IF you have taken CS165, CS161, CS261
    GOTO next syllabus section;
ELSE
    see below;

Background: Naturally, the more background you have the smoother your experience in 265 will be. Prior knowledge of C programming and systems programming, as well as a good understanding of computer architecture and in particular the memory hierarchy (cache memories) is very important for this class. Courses providing systems background (like CS50 and in particular CS61 or equivalent) are essential. Good hacking, algorithm designing, and data structures skills are also required.

If you are graduate student and have taken a mix of systems (database, operating and distributed systems) classes in the past, then you will be OK and we will provide enough background so you can follow. CS265 does satisfy the systems requirement towards a PhD.

If you are a senior in college and this is your last chance to take this class: if you have taken CS61 but no CS161 or CS165 then talk to the instructor to evaluate how fit your are for the class. If you have not taken CS61 but do have significant systems programming experience you may still qualify.

In all other cases, it is a better idea to take CS165 first.

HOW CAN I DO GREAT IN 265?

Just utilize all resources provided. Show up in class to participate in interactive sessions. There are also daily office hours and labs; show up as often as possible so we can help with anything you need! When you find yourself stuck with the project either with a design decision or just a bug, it is normal to struggle for a while — it is part of the learning process — but after some time grab your laptop and come by!

WHAT CAN I DO TO PREPEARE?

Especially if you have not taken CS165 it is a good idea to spend some time preparing before the semester starts and during the early weeks of the semester even if you consider yourself an expert systems student. The best approach is to browse some fundamental readings in data systems architectures. We propose that you take a look at the following texts from the CS165 readings:

1) Get familiar with the very basics of traditional database architectures:
Architecture of a Database System. By J. Hellerstein, M. Stonebraker and J. Hamilton. Foundations and Trends in Databases, 2007

2) Get familiar with very basics of modern database architectures:
The Design and Implementation of Modern Column-store Database Systems. By D. Abadi, P. Boncz, S. Harizopoulos, S. Idreos, S. Madden. Foundations and Trends in Databases, 2013

3) Get familiar with the very basics of modern large scale systems:
Massively Parallel Databases and MapReduce Systems. By Shivnath Babu and Herodotos Herodotou. Foundations and Trends in Databases, 2013

Test 0: We provide a Test 0 that is designed to 1) help you get an idea about how fit you are for the class and 2) bootstrap your C coding skills. Essentially Test 0 consists of an independent data structure design and implementation in C that will allow you to practice basic system design, coding and debugging skills. In addition, several fundamental section videos are posted on the class website about system coding and profiling to help you with that.

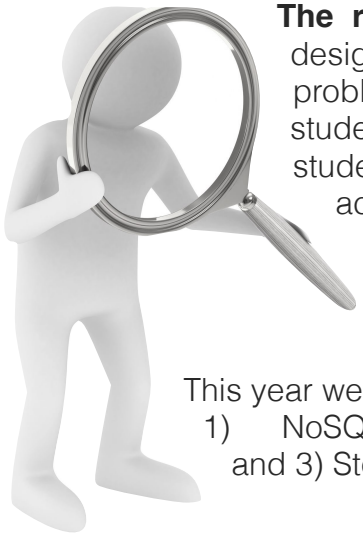
GRADING

- Class discussions: 20%
- Paper reviews: 15%
- Paper presentation: 15%
- Semester project: 40%
- Midway check-in: 10%

SEMESTER PROJECT

Each student completes a semester project. There are two kinds of semester projects:
1) a systems project, and
2) a research project.

Systems projects are tailored to provide background on state-of-the-art systems, data structures and algorithms. They include a design component and an implementation component in C or C++, dealing with low level systems issues such as memory management, hardware conscious processing, parallel processing, managing read/write tradeoffs and scalability. This year's systems project is about designing and implementing a key-value store in the form of a Log Structured Tree that can accommodate fast reads and writes. The key-value store design we will follow will resemble the state-of-the-art design used as a basis in numerous modern key-value stores such as Facebook, LinkedIn, Cassandra, and many more. The first part of the project is about designing the basic structure of an LSM tree for reads and writes, while the second part is about designing and implementing the same functionality in a parallel way so we can support multiple concurrent reads and writes. This is a focused project that while it is not extremely heavy in terms of how much code you have to write it will bring you against basic modern system design issues and tradeoffs. We will upload a detailed description on the class website before the beginning of the semester. Systems projects will be done individually, i.e., each student will have to work on the project on their own.



The research project, on the other hand, is much more tailored on design and proof of concept implementations trying to solve open problems. Research projects are tailored to give a taste of research to students and lead to publications. When working on a research project, students will work closely with the instructor and members of DASlab on active research projects of the lab. Students will work on groups of three. Such projects are mainly about thinking, reading and writing and much less about coding although proof of concept implementations will be our end target in some cases.

This year we will be working on the following research projects:
1) NoSQL Key-value stores, 2) The Periodic Table of Data Structures and 3) Storage for Fast Neural Networks

Research projects will be offered to students who have taken CS165 in the past and students who already have significant systems background. This will be done in consultation with the instructor.

In early February we will hold a special class to introduce both the systems project and the research projects in detail and this will be followed by a series of OH for clarifications. In the meantime students may browse the daslab website and learn more about the projects going on: http://daslab.seas.harvard.edu/,
and the class website for examples of projects from past years.

In special cases where a student wants to work on an alternative research project, i.e., a project which is inspired by existing research that the student is already doing (e.g., as part of a PhD for a grad student or a continuation of the CS165 project for an undergrad) we will work to accommodate such requests on a case by case basis. This will be done in consultation with the instructor and only if students probably would not benefit from doing a systems project as they know this material already. Assuming there is a strong plan and drive for a specific project, such requests will most likely be granted.

What is a successful project? For systems projects we will give out specific functionality and performance metrics you have to achieve as part of the description of the project. For research projects we will give out specific questions you need to answer when we set-up each individual research project.

Evaluation: There is no final or midterms. At the end of the semester each student will have a meeting with the instructor and another meeting with the TFs where students will demonstrate their projects and answer design questions about the project. [Tip: Past experience shows that frequent participation in office hours, brainstorming sessions and sections means that the instructor and the TFs are very well aware of your system and your progress which makes the final evaluation a mere formality for these cases.]

Collaboration policy: The systems project is an individual project: the final deliverable should be personal, you must write from scratch all the code of your system and all documentation and reports. Discussing the design and implementation problems with other students is allowed and encouraged! We will do so in the class and during office hours and brainstorming sessions. Research projects are going to be in groups of three and similar to the systems project we encourage discussions across teams but in the end each team should deliver a project that is clearly theirs.

Late days policy: All projects are due at the end of the semester and this is when they will be graded. The more input you give us, through the semester though, the more we can help you learn. In the systems project description you can find a detailed time-schedule that we propose you follow. Similarly, we will set up specific timelines for each research project. All timelines represent an ideal plan and you have the freedom to adjust according to your schedule.

There are no late days for reviews. This is because reviews are essential for you to follow each class.

Note: Experience says that every year a number of students cannot handle the freedom to self-pace, and end up significantly deviating from the schedule. We will send you frequent reminders but you should know that deviating from the schedule by more than a couple of weeks will most likely mean that you will not be able to finish the whole project by the end of the semester (unless you are an experienced systems student).

Midway Check-in: The goal here is to demonstrate that you are having decent progress and mainly to avoid falling behind. By early March each student working on a systems project should deliver 1) a design document, 2) a 10 minute presentation that describes the intended design for the whole project and, 3) at least one performance experiment that demonstrates an early result (10%). A template of the expected design document will be provided early in the semester.

FEEDBACK

We welcome feedback and ideas about the course at any point during the semester. Just come and chat with us during office hours! Tell us how you are keeping up and how we can make it easier for you.

NO LAPTOP/PHONE POLICY

CS265 is based on interaction. We want students actively participating in class and interactive sessions, asking and answering questions to maximize learning. In each class, we will bring a printed copy of the slides for each one of the students so you can follow along and to keep notes on paper. So you do not need your laptop or phones for notes or looking up the slides online. In fact, recent studies show that even if you only use a laptop for note taking, it can have a negative impact on how well you understand the material in class[1].

[The Pen Is Mightier Than the Keyboard: Advantages of Longhand Over Laptop Note Taking. Pam A. Mueller and Daniel M. Oppenheimer. Psychological Science. 2014, Vol. 25(6) 1159–1168]

[1] There are cases where having a phone or laptop during class is necessary such as when you expect an important call or message or when you need the laptop to better follow the slides due to any issues with your eyes or ears. Just let the instructor know and all such cases will be granted permission to use any tools necessary.

GUEST LECTURES

Every semester we arrange a few guest lectures by leaders in data system design from industry and academia. Past guest lecturers in our classes include: Guy Lohman from IBM Research, Erietta Liarou from EPFL Lausanne, Alkis Simitsis and Georgia Koutrika from HP Labs, Nikita Shamgunov from MemSQL, Laura Haas from IBM Research, Nga Tran from Vertica, Jignesh Patel from University of Wisconsin, Johannes Gherke, from Microsoft, Marcin Zukowski from Snowflake, Richard Hipp from SQLite, Ryan Johnson from Logicblox.

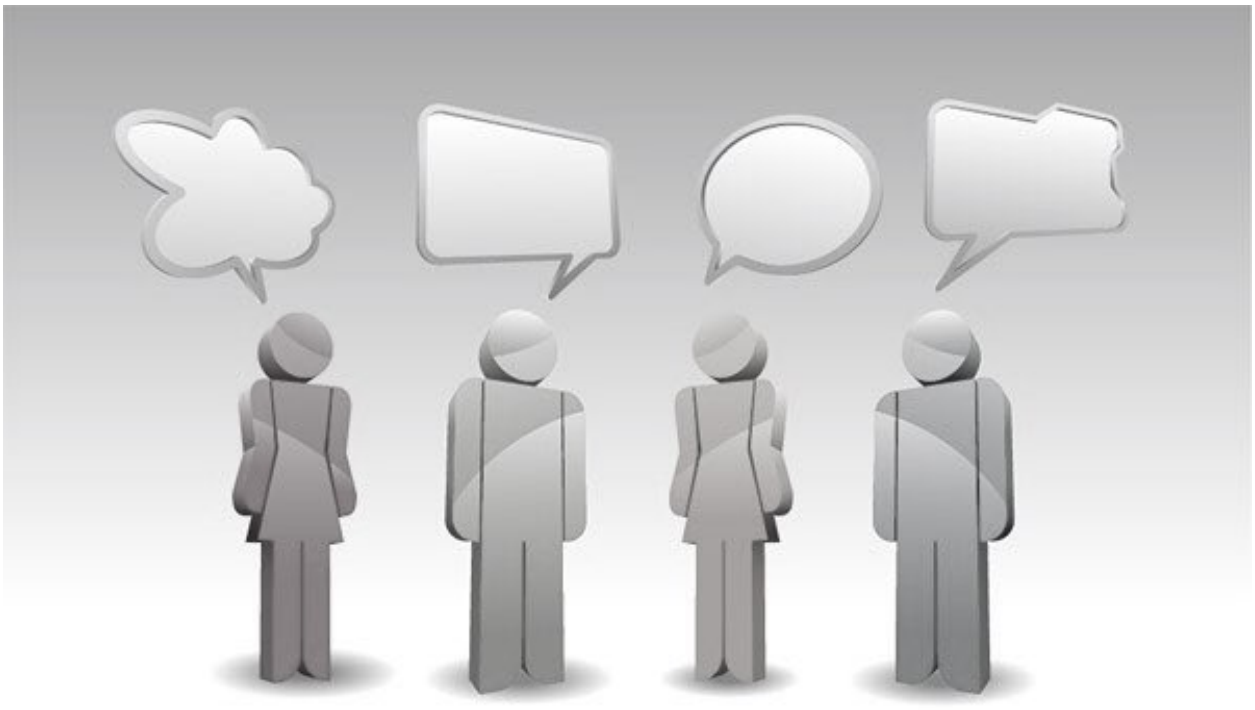
You will get the opportunity to both attend a guest lecture and to actively participate in discussions with our guest speakers.

REQUIRED TEXTBOOK

The class is about state-of-the-art data system design. There is no textbook for that. Thus, we use recent research papers and surveys which will be posted on the course website, which you will have access to through the Harvard network.

FEEDBACK ON PROGRESS

We provide feedback continuously. The main thing that you will need feedback on is your semester project and the paper reviews. The way to get feedback is to show up to our office hours and labs and share your design decisions, code, and test results with the staff or ask us to go with you over your paper reviews. In this way, you will get hands-on help and feedback. Specifically for reviews we will hold a special session every second week to “review the reviews”.



ONLINE DISCUSSIONS

We will use Piazza for online discussions. The links are posted on the class website. You are welcome to post any question that might help you understand the material better or help you with the project. Anonymous posting (to the other students) will be enabled so that students feel more comfortable posting questions.

BASIC RULES FOR PIAZZA: We only have a few basic rules so we can keep the forum functional and useful for the students as well as manageable for the staff.

1) We ask that you first search the forum well before posting a question so that we do not have duplicate entries.
2) Please make sure to stay on top of all staff posts (especially those that are pinned). Anything we post in Piazza we consider “known”.
3) Do not use Piazza to post code or ask help with debugging. While it can work in some cases remote debugging is a pain and takes a lot of time. We have labs every day. Bring your laptop and we will help you on site or join remotely and we will help you via a shared screen mode.
4) Do not use piazza for anything that is not about a technical question or a question abut class logistics. If you want to discuss any concerns about your progress, fit for the class or anything else you should come to OH.

PLAGIARISM

You are responsible for understanding Harvard and Harvard Extension School policies on academic integrity (www.extension.harvard.edu/resources-policies/student-conduct/academic-integrity) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting "the wrong draft", or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity. To support your learning about academic citation rules, please visit the Harvard Extension School Tips to Avoid Plagiarism (www.extension.harvard.edu/resources-policies/resources/tips-avoid-plagiarism), where you'll find links to the Harvard Guide to Using Sources and two, free, online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.

ACCESSIBILITY

Harvard and the Extension School are committed to providing an accessible academic community. The Disability Services Office offers a variety of accommodations and services to students with documented disabilities. Please visit www.extension.harvard.edu/resources-policies/resources/disability-services-accessibility for more information and do not hesitate to contact prof. Idreos directly, by email, with any questions or concerns you might have.

EXTENSION SCHOOL

This section supplements the basic syllabus with additional details that apply to extension school students.

CS265 is a heavily research oriented course that is structured in a very different way than other classes, valuing and promoting critical thinking. For most students this requires a transitions phase. Please check the syllabus and requirements carefully before committing to this course.

In addition, keep in mind that taking this course successfully will in practice require participation in OH and Lab sessions. Even if they are not mandatory, they are critical for students to understand how to think about the material and how to design solutions. Especially if you do not have all the background described in the syllabus (i.e., if you have not taken a research oriented systems course with a heavy systems project), you should budget time for frequent participation in both Labs and OH and many hours of additional work every week to build the foundations needed.

Lecture: Lectures will be broadcasted live. Lectures will also be available for on-demand broadcast within 24 hours after each class. Students will be able to watch the live or recorded broadcast through their browser. The link to the broadcasts for CS265 will be available through the canvas website for this class and will also be posted on the class website before the first lecture.

Participation: Extension school students will be able to participate live in classes, office hours and labs via web-conference tools (we will use Zoom). The course staff will be online with Zoom during each session that is marked as “remote” and you will be able to actively interact with the staff. Other than standard chatting and talking features Zoom also offers screen sharing features which can be used for when you need help with specific issues such as debugging.

Capturing Discussions: Given that a big portion of the class is based on interaction, extension school in cooperation with the class staff is always working to set-up a system with several microphones across the classroom so we can accurately and clearly capture brainstorming discussions and comments during class time. Microphones will “follow” the instructor.

Grading: Even though we encourage extension school students to utilize the opportunity to interact with the staff and participate in class live we know that for practical reasons this will not be possible for all remote students. For this reason for extension school students there will be no “class participation” grade. The rest of the course is exactly the same as what local students do.

For this reason the portion of the class participation grade (20%) will be distributed in project (50%), presentation (20%) and reviews (20%). The 10% for the midway check-in completes the grade distribution for extension school.

Piazza: Given that remote students have usually a different set of needs due to the distance, there is a separate Piazza forum tailored to extension school. Look at the class website for the piazza forum link for extension school students.

Office Hours and Labs: Extension school OH and Labs take place during the weekend. In this way, we can have more flexibility to accommodate students with day jobs that cannot attend during the week. The schedule will be posted at the beginning of the semester on the class website and piazza.

Starting Date: Note that usually extension school shows the class starting date to be one day after the actual starting date. In fact, this is when the first video will be available. However, extension school students will still be able to stream live the first class on the first day of classes and participate live as normal. Check the class website for the exact schedule if you want to participate live.

Graduate Credit: Extension school students who take the course for graduate credit and are on the systems project track should provide a detailed literature review of NoSQL key-value stores. This is due at the end of the semester along with the project and will account for 30% of the project grade.