Welcome to CS265!

**Professor:** Stratos Idreos

**url:** http://stratos.seas.harvard.edu  
**office:** MD139  
**email:** stratos@seas.harvard.edu

**TF:** Manos Athanassoulis  
manos.athanassoulis@gmail.com

---

**What is this class about?**

Big data is everywhere. A fundamental goal across numerous modern businesses and sciences is to be able to exploit 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 far from a simple task and 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 and algorithms that can "scale up" and "scale out". Scale up refers to the ability to use a single machine to all its potential, i.e., to exploit properly the memory hierarchy and the multiple CPU and GPU cores. Scale out refers to the ability to use more than 1 machines (typically 100s or 1000s) effectively. We will use examples from several areas, including relational systems and distributed databases, graph processing systems (i.e., for social networks), key value stores, noSQL and newSQL systems as well as mobile computing and interactive analytics (such as dbTouch). In a fast moving industry and research environment such skills are in high demand.

---

**Why take this class?**

Data is everywhere. Every year we create even more data. As it stands, every two years 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 severely affected. Data systems are in the middle of all this. Data systems is how we store and access data, i.e., they are the backbone for 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 in such dynamic environments.

This class will also be ideal for undergrads who ask the question “what is research?”.

**Expected learning outcomes**

- Understanding the basic tradeoffs in designing modern big data systems.

- Being able to design a new big data system given a data-driven scenario.

- Develop basic research skills: reading, writing and understanding research papers.

**Who can take this class?**

The class is accessible to students who have taken CS165 and to students who have a good background of modern data systems and designing algorithms for big data. Contact the instructor if you are unsure about whether you will be able to follow the class or not.

**Lectures**

The class meets twice a week: Wednesdays and Fridays 1:00pm-2:30pm in MD323. Class starts at 1:10pm. We will not have traditional lectures. All classes will be based on a discussion format. Each week we will discuss 2 research papers. Each student will be tasked to lead the discussion for at least 1 paper through the semester.

**Office hours**

Starting Week 1, Prof. Stratos Idreos will hold office hours every Wednesday 2:30pm-3:30pm (time slot can change if it is not convenient for everyone). Students are also welcome to schedule ad-hoc meetings with the instructor as often as needed.

TF office hours will be announced soon.

**Research and Brainstorming Sessions**

We will often schedule "brainstorming sessions". In such sessions, each group may work with the instructor independently on the details of their research project and also each group may get feedback on their progress from the rest of the groups.
Required textbook

We will use recent research papers and surveys which will be posted on the class website and you will have access to them through the Harvard network.

Research project

Each student will work on a research project throughout the semester. Students may work on groups of 2-4. Each group will work on its own project which will be defined within the first month. The projects will be tailored such as they follow both the interests of the students as well as open and challenging questions in the big data systems area. An ideal project will lead to 1) a set of new ideas on how to solve a specific problem, 2) an analysis that demonstrates the effectiveness and benefits of the new approaches as well as 3) a write up that is close to a full research paper (and may lead to a publication).

Assessment and grading

-Class participation: 10%
-Project: 90%

Collaboration Policy

You are welcome to discuss your projects and designs with each other. In fact you are expected to do so during our brainstorming sessions and to give valuable feedback to your classmates. However, all material you deliver (code, design docs and reports) should be produced by you. For ideas that have greatly affected your designs and came outside a project group, you should acknowledge your fellow students in your report. Your final grade will be assessed after several 1:1 meetings with the instructor and the TF where you will be expected to demonstrate not only the end result of your project but also that you "own" the results and the design.

Online discussions

We will use Piazza where you will be able to ask questions and discuss issues related to the course.

sign up url: https://piazza.com/harvard/fall2014/cs265
class url: https://piazza.com/harvard/fall2014/cs265/home