welcome to CS265!

prof. Stratos Idreos

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

data systems

cs265 goals & logistics

+you stop me any time for questions

when you see this speak up!
1 it is all about research
learn to question everything
understanding all possible steps (design space) and the potential impact is the basis of good science

(otherwise our algorithms are a set of mostly ad hoc choices that look like a good option but without any formal reason)
methods a and b should have only a single design difference
round table discussion/brainstorming

4+ lectures then student presentations
you are expected to talk a lot! ask questions, answer my questions

the goal is not instant perfection, our goal is to have interesting discussions

we will only grade class discussions for the second half of the semester
read/understand/review/improve state-of-the-art research
you are expected to write one paper review for each class
(so everyone is properly prepared for the in class discussions)

review and slides should focus on
what is the problem
why is it important
why is it hard
why existing solutions do not work
what is the core intuition for the solution
solution step by step
does the paper prove its claims
exact setup of analysis/experiments
are there any gaps in the logic/proof
possible next steps

* follow a few citations to gain more background

We will grade but do not count the first four reviews
You can skip any two reviews
Every two weeks there will be a special reviewing reviews lab
systems project

individual project
c/c++

research project

groups of three
analysis

only open to cs165 students
(unless proven otherwise)
noSQL key-value stores are everywhere…

project= basic state-of-the-art design
CrimsonDB

A Self-Designing Key-Value Store

www.crimsondb.com
evaluation
end of semester + midway check-in point (for 10%)
face-to-face demo

what is a good project
there will be specific “functionality/questions”
goals for both project options

coming often to OH makes evaluation just a formality
projects will be introduced in detail in class 4
PhD systems students may also propose their own project
ACM Special Interest Group In Data Management (SIGMOD) 
Undergrad Research Competition

cs265/165 students
2 entries in finals in 2015
2 entries in finals in 2016
1 entry in finals in 2017
1 entry in finals in 2018

world-wide competition from research labs in data management research
top 5-10 are invited to the conference to present their work
best work is given an award

+ first prize in 2016 & 2017
DATA SYSTEMS LABORATORY
http://daslab.seas.harvard.edu/

@ Harvard SEAS

Manos Athanassoulis
Postdoctoral Researcher

Niv Dayan
Postdoctoral Researcher

Michael Kester
Ph.D. Researcher

Lukas Maas
Ph.D. Researcher

Abdul Wasay
Ph.D. Researcher

Brian Hentschel
Ph.D. Researcher

Kostas Zoumpatianos
Ph.D. Intern

Ivan Sinyagin
Masters Researcher

Angelo Kastroulis
Graduate Researcher

Dhruv Gupta
Undergraduate Researcher

Demi Guo
Undergraduate Researcher

Xuxiang Tian
Undergraduate Research Intern

Xingding Wei
Undergraduate Research Intern

Heng Chang
Undergraduate Research Intern

Zichen Zhu
Undergraduate Research Intern

Yash Bolisetty
High School Research Intern

Srinidhi Krishnamurthy
High School Research Intern
EVERYONE CAN BE A DATA-SCIENTIST

making data systems
easy to design, tune and use
through adaptation and automation

Accelerating development and research
Accelerating machine learning

You will see a lot of research in class & Research Tuesdays
how much work is it?

is this going to be like 165?

NO (well, it could)

“more lightweight” systems project
even more research
and open ended
up to you how much you push
**background:**
programming
algorithms
data structures
hardware architectures

**can I keep up**

**can I follow the class?**
if not familiar with all the above = No
if some of the above are familiar = maybe

**prerequisites**
undergrad: cs165 || cs161
grad: systems classes
(talk to Stratos otherwise)

**can I prepare?**
yes - check later slides

http://daslab.seas.harvard.edu/classes/cs265/self_test.html
next few classes

class 1, 2
logistics
my take on big data challenges
basics on data system architecture and design
history & future

class 3
more basics on data systems
scale up vs scale out
big data systems

class 4
intro to class projects (key-value stores, automatic design)

class 5+
paper presentations and round table discussions begin

cs165 will see a lot of overlapping material in the first 3 classes
hmm, my data is too big :(  big data ?
haven’t we be doing data analysis forever

so what is new?
Every two days we create as much data as much we did from dawn of humanity to 2003

[Eric Schmidt, Google]
data

star(id, name, distance, density, …)

[1, star1, x1, y1, …]
[2, star2, x2, y2, …]
[3, star3, x3, y3, …]
[4, star4, x4, y4, …]
...

data collection is the key

learn a bit how computers work

need a bit more tailored analysis

need serious programming skills

exploration - many users/updates

data-driven analysis

data size

store - access

10s/100s  paper - just look at it!

K/M  PC files - shell/excel

B  PC files - custom

T  data sys. - declarative
big data V’s
(it is not about size only)

volume  velocity  variety  veracity

actually none of that is really new...

new:
our ability to gather and store machine generated data
broad understanding that we cannot just manually get value out of data
not always sure what we are looking for (until we find it)
“there are good chances we already have the data for the next big breakthroughs in say biology, medicine, etc. but we simply cannot extract the knowledge”

Martin Kersten, Stratos Idreos, Stefan Manegold and Erietta Liarou.  
**The Researcher’s Guide to the Data Deluge:** Querying a Scientific Database in Just a Few Seconds.  
Best paper award in Challenges and Visions
data systems are in the middle of all this
data system?
A data system **stores** data, **provides access** to data, and (ideally) makes data **analysis** easy.
“relational databases are the foundation of western civilization”

Bruce Lindsay, IBM
ACM SIGMOD  Edgar F. Codd Innovations award 2012
today

data systems are nearly everywhere…

continuous need for new and tailored data systems

tomorrow
more data

more applications

more h/w
5 decades of research
IBM, Microsoft, Oracle, Teradata, etc.
and a gazillion start-ups today

declarative interface
ask “what” you want

data* system
the system decides
“how” to best store
and access data

why is this good
“Three things are important in the database world: performance, performance, and performance”

Bruce Lindsay, IBM
ACM SIGMOD  Edgar F. Codd Innovations award 2012
(here is where all the magic happens!)

data system kernel

265: new research in detail, new systems architectures (not just column-stores), new models (not just relational), scale out, and more
you will learn to design and implement db kernels!
Stratos’ unofficial data systems definition:

A data system is a massive collection of data structures, algorithms, data flow and caching policies. It should all play nice with hardware. Somehow we should always make the right choice about which data structure, algorithm, data flow and caching method we use. Ideally users/apps should be able to just use it.
conflicting goals
so what is a good data system?
(hardware and requirements change continuously and rapidly)
it depends…

application requirements

budget

>1 QUINTILLION DESIGN OPTIONS!
CS165/265: STUDYING THE FRONTIERS OF DATA SYSTEMS RESEARCH AND TRADEOFFS AS IT IS APPLIED NOW OR WILL BE IN THE NEAR FUTURE
build a key-value store
similar to the ones Facebook, Google, etc use

interface supported: put, get, scan, count, get range, load
unique key-value pairs, r>>w but w>>0

how to store and access data
some questions:
what is key and what is value?
are they stored together?
can r/w fluctuate over time?
sort, b-tree, hash-table, scan, skip-list, zone-map?
what if we have 1000 queries or a million concurrent queries?
what if data is compressed?
multi-core, SIMD?
what if data does not fit in memory?
SQL queries

>1 users concurrently

Correct + complete answers

Security/robustness
~1960s

dbs  dbs  dbs  ...  dbs  dbs  dbs

late 1990s-early 2000: new designs start appearing

~2010-now: industry adoption and evolution

~2018

“dbs”

history/timeline
As apps become more complex and as apps need to be more scalable, is there anything really new?

- db: complex, legacy, tuning, expensive, ...
- noSQL: simple, clean, just enough, ...
- newSQL

...
scale up vs scale out
performance - correctness - data models

using one machine as best as possible

using >1 machines as best as possible
more logistics
big data systems:
e.g., column-store and hybrid systems, shared nothing architectures, cache-conscious algorithms, hardware/software co-design, main memory systems, adaptive indexing, stream processing, scientific data management, key value stores, noSQL, newSQL, systems for mobile computing, systems for human computer interaction

past but still relevant topics
e.g., relational model, row-store database systems, optimization, indexing, concurrency control, recovery, SQL

how and why did we get here and where things might go

no textbook - just research papers
cs265 goals
understanding system design tradeoffs
be able to design and prototype a data system!

see how the same concepts appear again and again
(it is all just bits!)

side-effects:
build or sharpen systems skills
(C programming, profiling, debugging and linux tools
algorithms & data structures
modern hardware architectures)

why
data system designer - researcher
any business - any science - any start-up
project due at the end of the semester
unlimited office hours
research oriented
open ended questions
discussion oriented
Project: 40%
Midway Check-in: 10%
Discussion: 20%
Presentation: 15%
Reviews: 15%
how to be successful in CS265?

ask a lot of questions, ask for a lot of help, come often to OH & extra sessions
piazza forum

all announcements & discussions
(link on class website - check out usage guidelines)
classes are recorded
(links on class website)

extension school students have access now
college students will have access as of next week
(but class 1 video is public)
NO LAPTOP/PHONE POLICY

class is based on participation!

we will bring a copy of the slides for every one in each time class so you can follow and keep notes

+ there is enough evidence that laptops and phones slow you down (check syllabus for more info)
+ a group of awesome TFs

Brian, PhD

Kostas, Postdoc

Mike, PhD

Wasay, PhD

office=MD136
prof. Stratos Idreos
other names: Efstratios Ydraios
Ευστράτιος Υδραίος, Στράτος Υδραίος

grew up in Greece - fav non-cs hobby: windsurfing

Diploma and ME Technical University of Crete, Greece
Ph.D. University of Amsterdam, Netherlands
Research Intern: IBM Research California, Microsoft Research Redmond, EPFL Switzerland
Visiting Professor: National University of Singapore, EPFL Switzerland

some awards:
ACM SIGMOD Jim Gray Dissertation Award
ERCIM Cor Baayen Award
IEEE TCDE Early Career Award

http://stratos.seas.harvard.edu/
MD139

OH every W/Th/F 3-4pm
how can I prepare?

1) start browsing some basic texts

Get familiar with the very basics of traditional database architectures:

Get familiar with very basics of modern database architectures:

Get familiar with the very basics of modern large scale systems:

2) play with basic data structures
implementation in C (Project 0)
2 classes per week - OH/Labs every day
1 presentation/discussion lead - 2 reviews each week
research (or systems) project + midway check-in
**Action steps:**
1) Read the syllabus & website carefully,
2) Register to Piazza,
3) Do P0 if you have not taken CS165,
4) Register for paper presentation (week 2),
5) Start submitting your paper reviews (week 3)

**web site:**  http://daslab.seas.harvard.edu/classes/cs265/
**piazza:**  piazza.com/harvard/spring2017/cs265/home
**office hours:**  Stratos: Wed/Thur/Fri, 3-4pm, MD139
**TF office hours:**  check class website
**textbook:**  nope

research papers will be available from the Harvard network
welcome to CS265!

BIG DATA SYSTEMS

prof. Stratos Idreos

next up: more detailed logistics, e.g., midway check-in and start with data system design