welcome to CS165!

prof. Stratos Idreos

HTTP://DASLAB.SEAS.HARVARD.EDU/CLASSES/CS165/
today

big data

data systems

cs165 goals & logistics

when you see this speak up!

+you stop me any time for questions
hmm, my data is too big :(

big data ?
hype vs. business vs. science
(it is all good!)
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 the dawn of humanity to 2003.

[Eric Schmidt, Google]
big data V’s
(it is not about size only)

volume velocity variety veracity
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
data exploration

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. 
Best paper award in Challenges and Visions
tons of opportunities!
research - entrepreneurship - new fields
today
today

tomorrow
more data
more applications
more systems
more h/w
automated data collection
data systems are in the middle of all this
data system?
a data system stores data and provides access to data

(a better definition next time!)
“relational databases are the foundation of western civilization”

Bruce Lindsay, IBM
ACM SIGMOD  Edgar F. Codd Innovations award 2012
how would 1 day without databases look like

how many times did you access a database system today
data systems are everywhere...
declarative interface
ask “what” you want

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

db system

why is this good

5 decades of research
IBM, Microsoft, Oracle, Teradata, etc.
and a gazillion start-ups today
“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

cs165/265 student
you will learn to design and implement db kernels!
you will learn to design and implement db kernels!

res=0;
for(i=0;i<100;i++)
    if (data[i]>v) res++;

what this does is it “good”
you will learn to design and implement db kernels!

res=0;
for(i=0;i<100;i++)
if (data[i]>v) res++;

what this does is it “good”

res=0;
for(i=0;i<100;i++)
res+=(data[i]>v);
In each class we will hold a 30-40 minute interactive session. Breakout groups (3-5 students) to brainstorm on a problem (often open research problem!)
our first in-class quiz: a “simple” example

assume a data set of $N$ integers: find all positions where $x_1 > \text{value} > x_2$

data

result

how to best store and access data?

you may design this in any way you want

report: metric of success, costs, details for storage and access
too many options…
store as fixed width and dense array and scan
store as variable length to save space but scan maybe slower
turn to a tree but is this better than scan
maybe make it a hash table
it depends on what we ask, how often we update, how we update,
do we have enough storage, memory, cpu cycles
do we try to improve latency, throughput or energy utilization
etc. etc. etc.

can we do all (a lot) of the above and then chose the best option dynamically?
Yes, and that is the definition of declarative data systems
data systems architectures

data structures + algorithms

some problems:
how to store data
how to access data

how to best answer a complex query
(e.g., which data to access first and how)

how to answer millions of queries concurrently

how to guarantee correctness and availability

how to spend the least possible energy

...
so what is a good data system?
so what is a good data system?

it depends…

application requirements

performance

hardware

budget

energy profile
conflicting goals
(hardware and requirements change continuously and rapidly)

moving target

application requirements

performance

budget

hardware

energy profile
data systems design (and research) is kind of an art
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
+ a group of awesome TFs

Manos, Postdoc

Brian, PhD, 2nd year

Mike, PhD, 5th year

Niv, Postdoc

Rohit
Undergrad, Senior

Joseph, Undergrad, Senior

office=MD136
EVERYONE CAN BE A DATA-SCIENTIST

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

You will see a lot of research in class & Research Tuesdays
http://daslab.seas.harvard.edu/classes/cs165/

logistics
CS165 topics

~1960s

modern 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, and key value stores

~2000:
new designs start appearing

~2010-now:
industry adoption and evolution

~2016

“dbs"

cs165 topics

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
CS165 GOALS
understanding system design **tradeoffs**
be able to **design** and **prototype** a data system!
a first idea of what it means to do **systems research**

side-effects:
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
unlimited late days
unlimited office hours
research oriented
open ended questions
discussion oriented
philosophy

unlimited late days
unlimited office hours
research oriented
open ended questions
discussion oriented

some examples:
OH every day + on demand sections: video only for background & help with material and project labs: every day - hands-on help 2+ brainstorming sessions 2+ research Tuesdays guest lectures from industry labs midterms with open books & notes
http://daslab.seas.harvard.edu/classes/cs165/project.html

storage, indexing, fast algorithms, complex algorithms, updates
unlimited late days

we provide:
APIs, tests,
client-server com code

C, individual project
bonus tasks
you >> MySQL
leaderboard & automated testing
+ a benchmark implementation to chase
details about how to get code, register, etc. on class website
testing 3 days a week now, multiples times a day later on
ideal background (CS50, 51, 61):
programming
algorithms
data structures
hardware architectures

check out Project 0 for preparation

how hard is it?

http://daslab.seas.harvard.edu/classes/cs165/self_test.html
how to be successful in CS165?

- ask a lot of questions, ask for a lot of help, come often to OH, labs & extra sessions
- project: pass tests + face to face evaluation
- midterms & quizzes: known answers + open questions
Project 40%
Quizzes and class participation: 20%
Midterms (2): 30%
Midway Check-in: 10%
Bonus points: extra tasks for the project: 10%
Bonus points: best projects: 5%

Extension school
No class participation:
Project 50%
Midterms 40%
Midway Check-in: 10%
MIDTERMS
10/17 & 11/21

open books/notes

we try to make it fun
it is about problem solving
similar to in class quizzes
we value creativity!
no final: face to face evaluation in the end & pass project tests

MIDWAY CHECK-IN: making sure no one is left behind
be able to pass 3 tests of P1 by 10/31, 10% of total grade
how to be involved in research?

work on the open questions + or bring your own
come to OH, research sessions, etc
how to be involved in research?

work on the open questions + or bring your own
come to OH, research sessions, etc

world-wide competition from research labs in data management research
top 5-10 are invited to the conference to present their work
top 3 are given an award

2014: 2 CS165 students in the finals in Melbourne
2015: 2 CS165 students in the finals in San Francisco
First Prize!
slides are not notes!
slides are mainly there to trigger discussion

note keeping is your task:
starting class 3 we will do collaborative note taking:
http://tinyurl.com/CS165-Fall2016-Notes
https://piazza.com/harvard/fall2016/cs165/home

all announcements in piazza
all info in piazza and class website is considered known
classes are recorded
(links on class website)
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)
**web site:**  http://daslab.seas.harvard.edu/classes/cs165/

**project:**  http://daslab.seas.harvard.edu/classes/cs165/project.html

**self-evaluation:**  http://daslab.seas.harvard.edu/classes/cs165/self_test.html

**piazza:**  https://piazza.com/harvard/fall2016/cs165/home

**notes:**  http://tinyurl.com/CS165-Fall2016-Notes

**office hours:**  Stratos: every day, 3-4pm, MD139

**labs:**  Mon/Wed 5:30-6:30 pm, Tue 6-8pm, Thu/Fri 4-5 pm

**textbook:**  Database Management Systems, by R. Ramakrishnan and J. Gehrke
+ several modern data systems surveys and research papers
actions

check out syllabus, website and self-test carefully
register for piazza if you are taking the class
& for note taking
so what does it mean to take cs165?

- algorithms & data structures
- tuned for modern hardware
- **learn to think in a research way**
- do a huge project in C to actually implement these algorithms and data structures
WELCOME TO CS165!

things to remember from today

big data: business term but also backed by cool research

data systems: store and access any data

even “simple” data access problems are far from obvious to solve

data system: collection of algorithms and data structures

declarative processing is the key in data systems
stratos away next week

next class:
history & future
essential db properties
models - SQL
basics on system architecture and design

NEXT CLASS ON SEP 2
2-3:30pm, MDG125

NO CLASS ON SEP 7
TFs will hold a lab

NO OH NEXT WEEK
every day labs +
OH with Stratos on demand
via Skype or Zoom
welcome to CS165!
DATA SYSTEMS
prof. Stratos Idreos

office hours by Stratos start tmr:
every week day 3-4pm, MD139