welcome to CS165!
prof. Stratos Idreos

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

big data

data systems

cs165 goals, logistics & spirit

— — interactive class — —

when you see this speak up!

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

big data?
hype vs. business vs. science
(it is all good!)
data → analysis → knowledge
haven’t we be doing data analysis forever? 

so what’s 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
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
tons of opportunities!

research - entrepreneurship - new fields
today
today

tomorrow
FROM SOCIAL NETWORKS
TO MASSIVELY SCALABLE MONITORING SYSTEMS
data systems are in the middle of all this big data
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 one day without databases look like
how many times did you access a database system today
data systems are everywhere...
5 decades of research
IBM, Microsoft, Oracle, Teradata, etc.
and a gazillion start-ups today

declarative interface
ask “what” you want

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

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);
IF STATEMENTS ARE BAD BAD BAD

```
res=0;
for(i=0;i<100;i++)
    res+=(data[i]>v);
```
learning through interacting and brainstorming

+ OH & Labs: every day
I might not give the right answer

My questions might be “too simple”

I do not want to disturb the class

I just want to sit alone at the back & listen

...
In each class we hold a 30 minute **quiz**. Breakout groups (3-5 students) to brainstorm on a problem *(often open research problem!)*
our first in-class quiz: a “simple” example (break out sessions start in class 2)

assume a data set of \( N \) integers: find all positions where \( x_1 > \text{value} > x_2 \)
our first in-class quiz: a “simple” example
(break out sessions start in class 2)

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

you may design this in any way you want
report: metric of success, costs,
details for storage and access

how to best store and access data?
all modern & past systems: sql, nosql, +
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
is this for one query or many in parallel
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

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

hardware

performance

energy profile

budget
conflicting goals (hardware and requirements change continuously and rapidly)

budget

moving target

application requirements

performance

energy profile

hardware

conflicting goals

moving target

budget

application requirements

performance

energy profile

hardware
conflicting goals
(application requirements change continuously and rapidly)

moving target
(hardware and requirements change continuously and rapidly)

>1 QUINTILLION DESIGN OPTIONS!
data systems design and research is kind of an art
data systems design and research is kind of an art
data systems design and research is kind of an art

CS165/265: STUDYING THE FRONTIERS OF DATA SYSTEMS RESEARCH AND TRADEOFFS AS IT IS APPLIED NOW OR WILL BE IN THE NEAR FUTURE
http://daslab.seas.harvard.edu/classes/cs165/

logistics
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

Kostas, Postdoc

Niv, Postdoc

Brian, PhD

Mike, PhD

Wasay, PhD

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
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

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

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 classes
guest lectures from industry labs
midterms with open books & notes
CS165 WICKED AWESOME SEMESTER PROJECT

Design and build a main-memory optimized column-store

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

storage, indexing, fast algorithms, complex algorithms, updates
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

Seniors talk to Stratos if you did not take CS61
how to be successful in CS165?

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

I got an A in 165!
Project 40%
Class participation: 20%
Midterms (2): 30%
Midway Check-in: 10%
Bonus points: extra tasks for the project: up to 5%
Bonus points: speed: up to 5%

Extension school
No class participation:
Project: 50%
Midterms: 40%
Midway Check-in: 10%
MIDTERMS
10/11 & 11/15

open books/notes
up to 2:30 hours

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 10/10, 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
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 entries in finals in 2017

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
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:
(link on class website)
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!

+ there is enough evidence that laptops and phones slow you down
  (check syllabus for more info)
so what does it mean to take CS165?

- algorithms & data structures
- tuned for modern hardware
- learn to think in a researchy 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 data

a data system is a collection of algorithms and data structures

declarative processing is key

even “simple” data access problems are far from obvious to solve
HTTP://DASLAB.SEAS.HARVARD.EDU/CLASSES/CS165/

project,
self-evaluation
piazza
notes
office hours
labs
readings
**schedule**
read the syllabus, and self-evaluation carefully

talk to Stratos if you have any concerns/questions

if you are taking the class: register for piazza & note taking
welcome to CS165!
DATA SYSTEMS
prof. Stratos Idreos

office hours by Stratos start today:
M/W 5:30-6:30pm, Tue/Thu/F 3-4pm at MD139