welcome to CS165!

prof. Stratos Idreos

HTTP://DASLAB.SEAS.HARVARD.EDU/CLASSES/CS165/
big data

data systems

cs165 goals & logistics
big data ?
data → analysis → knowledge
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]
big data V’s
(it is not about size only)

volume  velocity  variety  veracity
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.

**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
future data-driven applications
data systems are in the middle of all this
actually all of CS is in the middle of the big data story!

major opportunity (and responsibility) for interdisciplinary research
data system?
a data system **stores** data and **provides access** to data

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

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

how many times did you access a database system today?
declarative interface
ask “what” you want

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

db system

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

db system

why is this good

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

>1 users concurrently

correct + complete answers

security/robustness

db system
“Three things are important in the database world: performance, performance, and performance”

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

data system kernel

cs165 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] > res)
        res = data[i];
```

what this does is it “good”
a “simple” example

assume an array of $N$ integers:
find all positions where $x_1 > \text{value} > x_2$
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

budget

hardware

performance

energy profile
conflicting goals

application requirements

performance

budget

hardware

energy profile
conflicting goals  moving target

(hardware and requirements change continuously and rapidly)

application requirements

performance

budget

hardware

energy profile
data systems design (and research) is kind of an **art**
essential steps in using a database system

- **clean**
- **schema**
- **load**
- **tune**

**query**

**experts/system admins**

**user/apps**
~1960s

late 1990s-early 2000: new designs start appearing

~2010-now: industry adoption and evolution

~2015

history/timeline
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

Mike, PhD, 4th year

Mike, PhD, 4th year

Alex, Undergrad, Senior

Wasay, PhD, 2nd year

Lukas, PhD, 2nd year

office=MD136
(moving in this week from MD334)
DATA SYSTEMS LABORATORY
@ Harvard School of Engineering and Applied Sciences

Designing data systems for the big data era

http://daslab.seas.harvard.edu/
One size does **not** fit all

Custom solutions are needed for optimal performance

Solutions need to be tuned

Bootstrapping new systems is expensive and time-consuming
data+queries+hardware

self-designing
data systems

data system
dbTouch: Analytics at your fingertips.
Conference on Innovative Data Systems Research (CIDR), 2013

design db kernels for touch-based exploration
I am curious, smart, autonomous, fast and I know what you want.

show me something interesting
http://daslab.seas.harvard.edu/classes/cs165/

logistics
cs165 topics

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

20% textbook - 80% research papers
cs165 goals
understanding system design tradeoffs
be able to design and prototype a data system!

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

**some examples:**
OH every day + on demand
sections 4x per week for background &
help with material and project
2+ brainstorming sessions
2+ research Tuesdays
guest lectures from industry labs
midterms with open books & notes
we provide: APIs, tests, client-server com code leaderboard

C, individual project

bonus tasks

you >> MySQL

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

**background:**
programming
algorithms
data structures
modern hardware architectures

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

http://daslab.seas.harvard.edu/classes/cs165/self_test.html
Project 60%
Quizzes and class participation: 10%
Midterms (2): 30%
Bonus points: extra tasks for the project: 10%
Bonus points: best projects: 5%

grading adapts to your starting level and progress
how to be successful in CS165?

- ask a lot of questions, ask for a lot of help,
- come often to OH and sections & extra sessions
- project: pass tests + face to face evaluation
- midterms & quizzes: known answers + open questions

I got an A in 165!
how to be involved in research?

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

examples from 2014 class:
2 cs165 students in the final round of the ACM SIGMOD undergrad research competition in Melbourne

3 pubs + 2 coming up
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-notes
https://piazza.com/harvard/fall2015/cs165/home

all announcements in piazza
all info in piazza and class website is considered known
web site: http://daslab.seas.harvard.edu/classes/cs165/

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


piazza: https://piazza.com/harvard/fall2015/cs165/home

notes: http://tinyurl.com/cs165-notes

office hours: Stratos: every day, 2:30pm-3:30pm, MD139

sections: Sun-Wed, 6:30-80pm, MD 136

textbook: Database Management Systems, by R. Ramakrishnan and J. Gehrke
+ several modern data systems surveys and research papers
we want you to have fun!
data systems is an exciting field!
tell us how you are keeping up
tell us what you need to better follow the class
tell us your suggestions about how to improve the class
next class:
history & future
essential db properties
models - SQL
basics on system architecture and design

no class on monday
sections as of W3
welcome to CS165!
DATA SYSTEMS
prof. Stratos Idreos

office hours by Stratos start today:
every week day 2:30-3:30pm, MD139