Welcome to CS265!

Prof. Stratos Idreos

http://daslab.seas.harvard.edu/classes/cs265/
big data

data systems

cs265 goals & setup
it is all about research
learn to question everything
round table discussion/brainstorming
+ a few lectures
read/understand/review/improve state-of-the-art research
how much work is it?

is this going to be like 165?
big data
data → analysis → knowledge
haven’t we be doing data analysis forever

so what is new

?
data

star(id, name, distance, density, …)

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

star(id, name, distance, density, …)

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

10s/100s

paper - just look at it!

data collection is the key
data

star(id, name, distance, density, …)

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

data size

store - access

10s/100s

paper - just look at it!

data collection is the key

K/M

PC files - shell/excel

learn a bit how computers work
### Data

Data is structured as follows:

```plaintext
star(id, name, distance, density, …)
[1, star1, x1, y1, …]
[2, star2, x2, y2, …]
[3, star3, x3, y3, …]
[4, star4, x4, y4, …]
…
```

<table>
<thead>
<tr>
<th>Data Size</th>
<th>Store - Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>10s/100s</td>
<td>paper - just look at it!</td>
</tr>
<tr>
<td>K/M</td>
<td>PC files - shell/excel</td>
</tr>
</tbody>
</table>

#### Data Collection

- Data collection is the key.
- Learn a bit how computers work.
- Need a bit more tailored analysis.
- Need serious programming skills.

---

CS265, Spring 2016
Stratos Idreos
data

star(id, name, distance, density, …)

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

data size
store - access

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

data collection is the key

K/M
PC files - shell/excel

learn a bit how computers work

need a bit more tailored analysis

B
PC files - custom

need serious programming skills

exploration - many users/updates

T
data sys. - declarative

data-driven analysis
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”

data systems are in the middle of all this
data system?
a data system **stores** data and **provides access** to data
“relational databases are the foundation of western civilization”

Bruce Lindsay, IBM
ACM SIGMOD  Edgar F. Codd Innovations award 2012
dbs are everywhere...
declarative interface
ask “what” you want

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

why is this good
SQL queries

>1 users concurrently

correct + complete answers

security/robustness
“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!
a “simple” example

assume an array of $N$ integers: find all positions where $x_1 > value > x_2$
some questions:
sort, b-tree, hash-table, scan?
what if we have 1000 queries or a million queries?
what if data is compressed?
what if we need to do a max operation after the select?
how to store positions?
multi-core - SIMD?
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

budget

hardware

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

application requirements

moving target

budget

performance

hardware

energy profile
data systems design (and research) is kind of an *art*
~1960s

dbs dbs dbs ...
dbs dbs dbs

late 1990s-early 2000: new designs start appearing

~2010-now: industry adoption and evolution

~2015

"dbs"

history/timeline
As apps become more complex and as apps need to be more scalable, newSQL emerges.
it all starts with how we store data

every bit matters
scale up vs scale out
performance - correctness - data models

using one machine as best as possible

using >1 machines as best as possible
http://daslab.seas.harvard.edu/classes/cs265/

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:**
sharpen systems skills if not there
(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
round table discussion/brainstorming

concepts/algorithms problems/new ideas

2 papers each week one scale up one scale out

students: 1 presentation 5 reviews
options 1: work on log structured merge trees
basic key-value store functionality - work individually
single machine - multi-core design

options 2: work on a research project
self-designing data systems + shape-shifting access methods
research with DASlab researchers - groups of 3
available for cs165 students or otherwise advanced students

3 publications + 2 pending from last course offering
2 students in ACM SIGMOD 2015 undergrad research competition in Melbourne
projects

evaluation
end of semester + midway checkin 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
students may also propose their own project
**background:**
programming
algorithms
data structures
hardware architectures

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

**can I keep up**

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

**can I prepare?**
yes - check next slide
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 (linked list/hash table/tree)
Project: 80%
Discussion: 5%
Presentation: 10%
Reviews: 5%
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)
classes are recorded (links on class website)

extension school notes on website for class participation and OH
web site: http://daslab.seas.harvard.edu/classes/cs265/

piazza: https://piazza.com/harvard/spring2016/cs265/home

office hours: Stratos: Wed/Thur/Fri, 3-4pm, MD139

TF office hours: Mon-Tue, 5-6pm, MD 136

textbook: nope
research papers will be available from the Harvard network
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
DATA SYSTEMS LABORATORY
@ Harvard School of Engineering and Applied Sciences

Designing data systems for the big data era

http://daslab.seas.harvard.edu/
**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

Lukas, PhD, 2nd year

office=MD136
next few classes

class 2
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

class 5+
paper presentations and round table discussions begin
2 classes per week - OH every day
1 presentation/discussion lead - 5 reviews
research (or systems) project
welcome to CS265!

BIG DATA SYSTEMS

prof. Stratos Idreos

office hours by Stratos start tmr:
wed-thur-fri 3-4pm, MD139