welcome to CS265!

BIG DATA SYSTEMS

prof. Stratos Idreos

HARVARD
School of Engineering and Applied Sciences
plan for today

big data + systems

class structure + logistics

who is who

project ideas
data vs knowledge
big data
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
"All models are wrong, but some are useful."

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
today

what is a db?

WANTED
Data Scientists

 tomorrow
soon everyone will need to be a “data scientist”

I should have used a column-store

SELECT max(toys)
FROM store
WHERE mam=won’t yell
big data systems

data systems are in the middle of all this
“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

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

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
essential steps in using a database system

clean → schema → load → tune

query

experts/system admins

user/apps
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**
database kernel

applications

data

data

data

algorithms/operators

data

data

data

sql

cpu

memory

disk
scale up vs scale out
~1960s

late 1990s-early 2000: new designs start appearing

~2010-now: industry adoption and evolution

~2014
data systems design (and research) is kind of an art
cs265 goals

- understanding system design tradeoffs
- be able to design and prototype a data system

-get experience in research:
project: work with instructor in a research problem which will lead to a basis for a publication
Lectures twice a week:
a student presents a research paper and leads a discussion
and brainstorming session

Office hours every week + more on demand
Stratos: Wed 2:30pm-3:30pm, TF office hours: TBA

1 brainstorming session every 3-4 weeks per project

Class participation 10% (presentation, feedback)

No midterms, quizzes, etc.

Research project 90% - groups of 1-4
(research results, ideas, presentation, demo, report/paper)
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
who is who
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, Microsoft Research, EPFL
**Visiting Professor**: National University of Singapore, EPFL Switzerland

**Fav Awards:**
ACM SIGMOD Jim Gray Dissertation Award
ERCIM Cor Baayen Award
a db system
allows you to answer queries fast

a data exploration db system
allows you to find fast which queries to ask
tell me something interesting (fast)

insert data
design db kernels for touch-based exploration
the theory of data systems evolution
**next class:**
Stratos will talk about basic column-store design
(with help from cs165 students)

**as of next week:**
students lead discussion
first batch of papers will be online by Friday
welcome to cs265!
BIG DATA SYSTEMS
prof. Stratos Idreos