CSE 344

INTRODUCTION TO DATA MANAGEMENT

WELCOME!

- **CSE 344**
- **Today's lecture**
  - Course administration
  - What to expect
  - Introduction and motivation

COURSE FORMAT

Lectures

- Location: SIG 134 (moved from MOR)

Sections:

- Content: exercises, tutorials, questions, new materials (occasionally)
- Locations: here
- Please attend
- Bring your laptop!
  - will often walk through software setup

8 homework assignments

7 web quizzes

Midterm and final

GRADING

Homework	40%
Web quizzes	10%
Midterm	20%
Final	30%

(subject to change)

ADMINISTRATION

Web page: http://www.cs.washington.edu/344

- Syllabus (course information)
- Lecture/section notes will be available there
- Homework assignments will be available there

Discussion Board (Piazza or Google Group?)

- Questions and clarification; place to give and get help
- NOT office hours: code can be difficult to debug remotely
- NOT private with staff: no grading questions or other private matters

Gitlab

- Account created this week, for submitting HW assignments

NewGradiance

- Autograded online quizzes, good for practice, unlimited attempts, last score counts

TEXTBOOK

***Database Systems: The Complete Book*,**
Hector Garcia-Molina,
Jeffrey Ullman,
Jennifer Widom

Good reference and alternative explanation

Also, good source for practice problems

EIGHT HOMEWORK ASSIGNMENTS

H1: Sqlite intro (Out tomorrow)

H2: Sqlite basics

H3: Advanced SQL on Azure

H4: Datalog and Relational Algebra

H5: Json and SQL++

H6: Spark on AWS

H7: Schema Design

H8: Transactional Application

Submit via git

ABOUT THE ASSIGNMENTS

You will learn/practice the course material:

- SQL, RA, parallel db, transactions, ...

You will also learn lots of new technology

- Cloud computing: Azure, AWS
- NoSQL: AsterixDB, Souffle
- Git

The time spent learning the new technology is very useful: *write everything on your CV!*

DEADLINES AND LATE DAYS

Assignments are expected to be done on time, but things happen, so…

You have up to 3 late days

- Used in 24-hour chunks

Late days = safety net, not convenience!

- You should not plan on using them
- If you use all 3 you are doing it wrong

Any lateness beyond that = 20% penalty per day

You must notify the staff for assignments 2+ days late

- (otherwise, we might not notice)

SEVEN WEB QUIZZES

- http://newgradiance.com/
- Create account;
  please make sure you use your UW first/last name
- Token to be provided to course email

Short tests, take many times, best score counts

No late days – closes at 11pm deadline

Provides explanations for wrong answers

LECTURES

- **Slides contain vital information for exams**
  - May emphasize tricks or problem types off slides
- **Posted after lecture**
- **Associated readings**
  - Good for alternate explanations

EXAMS

Dates:

- Midterm, Friday, July 27th (tentative)
- Final, Friday, August 17th
- both are *in class*

Final will include some first-half material

SUMMER QUARTER

Changes in summer:

- fewer lectures (and no extra week for finals)
- classes are 10 minutes longer

Implications:

- slightly less time for homework assignments
- schedule may need to change as we go along

ABOUT ME

- **Kevin Zatloukal (kevinz at cs)**
- **UW CSE undergraduate**
- **MIT Ph.D. (quantum algorithms)**
- **15 years in industry: Google, Microsoft, BEA, startup**
- **Part-time Faculty**
  - On campus MWF
  - Otherwise available by email

ABOUT STAFF

- **TAs**

Yao Lu	luyao
Ying Wang	wangy288
Andrew Wei	nowei

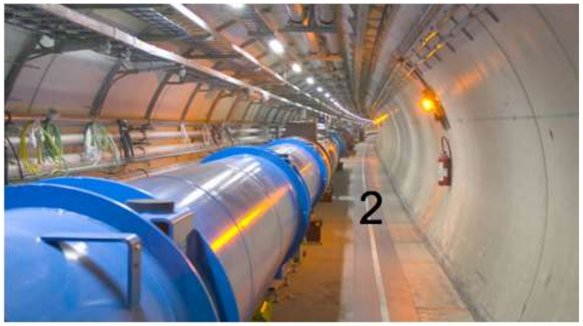
- **First resource for coding / setup problems**
- **Office hours posted soon (none until Friday)**

EXPECTATIONS ABOUT YOU

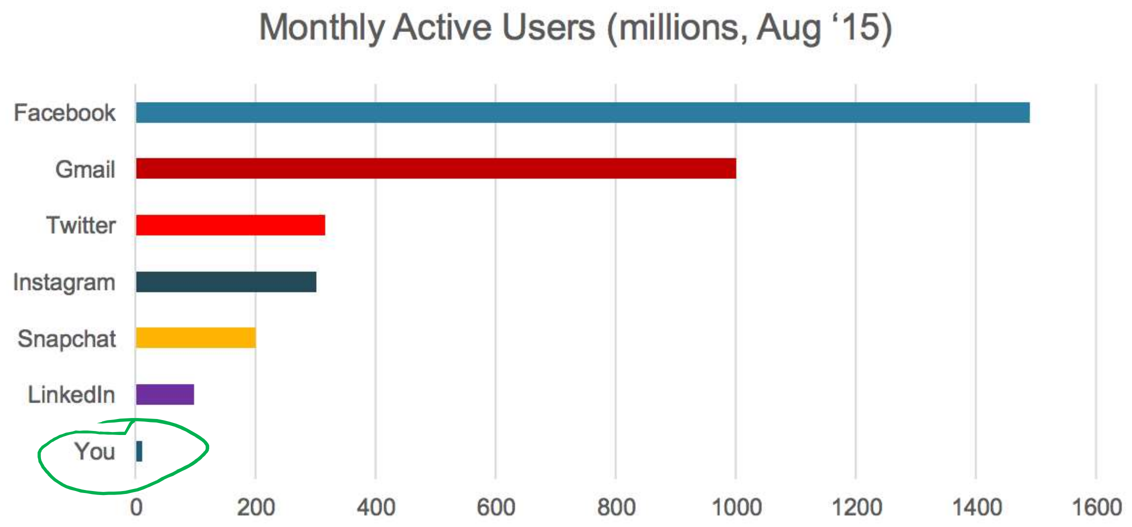
- **CSE majors**
- **(Hopefully) registered**
  - If not, talk with me after
- **Have taken CSE 311**
  - If not, may need to review relations
- **Likely headed to industry after UW CSE**
- **Academic Honesty**

WHY DATA MANAGEMENT?

- **The world is drowning in data!**
  - LSST produces 30 TB of data *per night*
    - Large Synoptic Survey Telescope
    - 9 PB per year
  - LHC produced 25 PB in 2012 finding the Higgs boson
    - Large Hadron Collider
  - Not just large scientific experiments
    - this affects ***almost every*** modern application



YOUR NEW APP



- **Suppose you:**
  - have 10M monthly active / 2M daily active users
  - record 20K per page view
  - have 200 page views per session
- **Analyzing 3 months of data for trends: 1 TB of data**

MORE USERS, MORE PROBLEMS

- **Efficiency problems**
  - takes a long time to read 1 TB of data from disk
- **Hardware problems**
  - disks fail, fiber optic cables fail
  - data centers light on fire
  - need to store data on many, geographically separated disks to avoid losing data
- **Concurrency problems**
  - can’t sell the last seat or last book to two people
  - those people could be on opposite sides of the globe

CLASS GOALS

The world is drowning in data!

Efficiently querying and updating it is hard!

Need computer scientists to help manage this data

- Help domain scientists achieve new discoveries
- Help companies provide better services (e.g., Facebook)
- Help governments (and universities) become more efficient

Welcome to 344: Introduction to Data Management

- Existing tools PLUS data management principles
- This is not just a class on SQL!

DATABASE

What is a database ?

DATABASE

What is a database ?

A collection of files storing *related* data

DATABASE

What is a database ?

A collection of files storing *related* data

Examples of databases:

accounts database
payroll database
UW's students database
Amazon's products database
airline reservation database

DATABASE MANAGEMENT SYSTEM

What is a DBMS ?

DATABASE MANAGEMENT SYSTEM

What is a DBMS ?

A big program written by someone else that allows us to manage efficiently a large database and allows it to persist over long periods of time

Examples of DBMSs

- Oracle, IBM DB2, Microsoft SQL Server, Vertica, Teradata
- Open source: MySQL (Sun/Oracle), PostgreSQL, CouchDB
- Open source library: SQLite

We will focus *mostly* on relational DBMSs quarter

EXAMPLE: YOUR NEW APP

What app should we build?

disease prediction app

What data do we need to store?

list of all diseases (w/ symptoms)

patient ~~data~~ history: list of pulse, etc. age, phone

search history.

session history.

EXAMPLE: YOUR NEW APP

What operations do we need?

What constraints can we put on the data?

EXAMPLE: YOUR NEW APP

- **Suppose we store the data in a regular file...**

- **How do we ensure:**

scale	can we support 100M users? 1B?
efficiency	how do we query it quickly?
fault tolerance	how do we survive failures?
concurrency	how do we support multiple users?
consistency	how do we save users from bugs?
changeability	how do we add new features?

WHAT A DBMS DOES

Describe real-world entities in terms of stored data

Persistently store large datasets

Efficiently query & update

- Must handle complex questions about data
- Must handle sophisticated updates
- Performance matters

Change structure (e.g., add attributes)

Concurrency control: enable simultaneous updates

Crash recovery

Security and integrity

MORALS

Almost any application has lots of important data

Getting the data right is (>) half the battle

- what operations do you want to support?
- what data do you need for that?
- what constraints does the data have?

DBMSs

- make app development easier
- make apps more reliable
- make apps more efficient
- make apps more easily changeable

THE PLAYERS

DB application developer: writes programs that query and modify data (344)

DB designer: establishes schema (344)

DB administrator: loads data, tunes system, keeps whole thing running (344, 444)

Data analyst: data mining, data integration (344, 446)

DBMS implementer: builds the DBMS (444)

WHAT IS THIS CLASS ABOUT?

Unit 1: Intro (today)

Unit 2: Relational Data Models and Query Languages

Unit 3: Non-relational data

Unit 4: RDMBS internals and query optimization

Unit 5: Parallel query processing

Unit 6: DBMS usability, conceptual design

Unit 7: Transactions

Unit 8: Advanced topics (time permitting)

WHAT TO EXPECT SOON

- Course Website
- Syllabus
- Git tutorial / help
- The first HW assignment
- Discussion board
- Canvas page
- Link for online quizzes