Welcome to DATA 604

Leanne Wu

lewu@ucalgary.ca

Department of Computer Science

UNIVERSITY OF CALGARY

About me

- PhD research in contextual privacy, data privacy, data collection and processing in undergraduate education

- Best way to get ahold of me is email: lewu@ucalgary.ca (I will try to respond within 24 hours or by Monday, if it's a weekend)

- Student hours will be held in ICT 517 (we can use my office upstairs if something requires discretion)

  Mondays: 1 -2 PM

  Wednesdays: 3-4 PM

  (Any other times of week): By appointment (email to set something up)

Your TAs

Abdullah Sarhan

- ICT 506
- asarhan@ucalgary.ca
- Office hours are Wednesdays 2:30 PM to 4:30 PM

Coskun Sahin

- ICT 526
- coskun.sahin1@ucalgary.ca
- Office hours are Tuesdays 1 PM to 3 PM

About this course

Database design, structure, functionality

- SQL (Relational databases)
- NoSQL (Document stores)

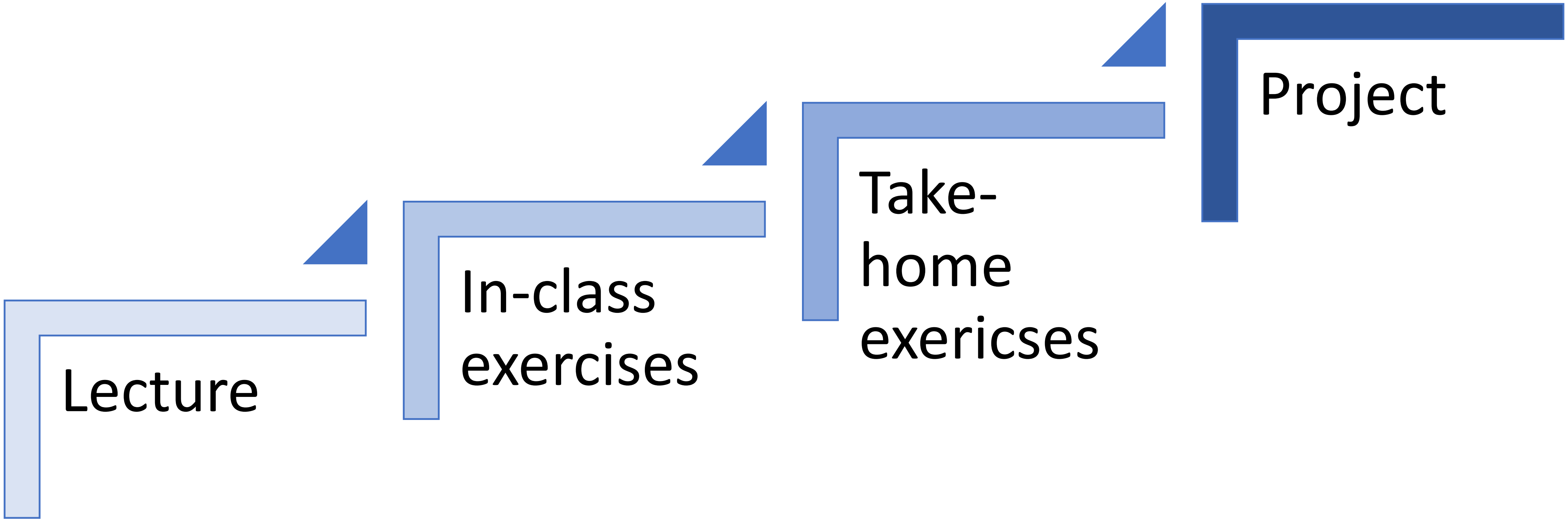
Scaling databases up (and out)

- Distributed databases, MapReduce, moving to the cloud

Specialized data stores

- Graph-based
- Stream-based
- Column-oriented
- Temporal
- Spatial

What to expect



Lecture

In-class exercises

Take-home exericses

Project

Assessment

- Take-home exercises (40% of final grade): Released Wednesday, due on the Tuesday after
  - Week 1 (5%)
  - Week 2 (10%)
  - Week 3 (12%)
  - Week 4 (13%)
- Project (60% of final grade)
  - Proposal (10%)
  - Bibliography (5%)
  - Materials (15%)
  - Presentation (30%)

Let’s start: An introduction to data

- What kinds of data do you deal with regularly?
- How (and what, and where) is it stored?
- What do you like working with? What don't you like working with?
- What technologies would you like to learn to manage this data?

Data

Data? Information? Knowledge?

Data

- Observations about events: “A coffee was sold for $2.75”

Information

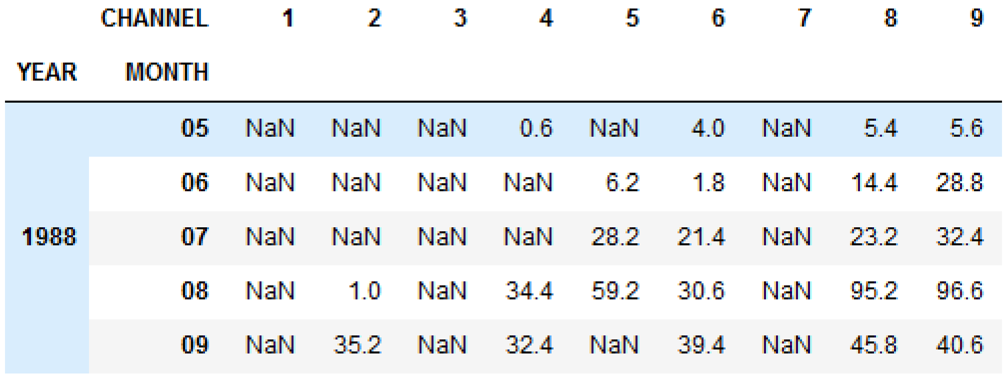
- Data linked with other data:
  - “We sold 58 small coffees, 90 medium coffees, and 75 large coffees”
  - “We sell more coffee in September compared to August”

Knowledge

- Information fit into conceptual models about the world: “Customers prefer coffee when they are in a rush, and espresso-based drinks when they have time to sit”

Structured Data

Tabular data



	CHANNEL	1	2	3	4	5	6	7	8	9
YEAR	MONTH									
1988	05	NaN	NaN	NaN	0.6	NaN	4.0	NaN	5.4	5.6
	06	NaN	NaN	NaN	NaN	6.2	1.8	NaN	14.4	28.8
	07	NaN	NaN	NaN	NaN	28.2	21.4	NaN	23.2	32.4
	08	NaN	1.0	NaN	34.4	59.2	30.6	NaN	95.2	96.6
	09	NaN	35.2	NaN	32.4	NaN	39.4	NaN	45.8	40.6

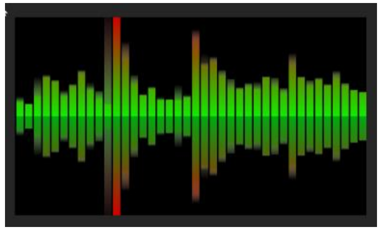
- What was to easy to do with DataFrames?
- What was difficult to do with data frames?
- Other pros/cons?

Unstructured Data

- Text (lorem ipsum etc. etc.)
- Graphics

Source: Open Clip Art Library : openclipart.org

- Audio

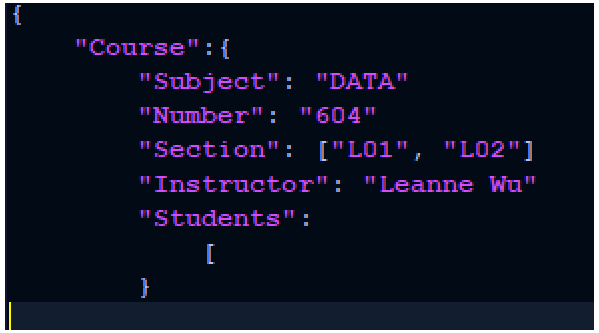


- Video
- Others(?)

- Unstructured data has a structure… but not structured as structured data
- Variety of formats - consider lossy vs. lossless formats

JSON (Javascript Object Notation)

- Common method of transmitting and storing unstructured data
- Most modern programming languages provide some way to use JSON



```
{
    "Course":{
        "Subject": "DATA"
        "Number": "604"
        "Section": ["L01", "L02"]
        "Instructor": "Leanne Wu"
        "Students":
            [
        }
```

JSON Values

- Permitted literals: numbers, strings, Booleans (`true`/`false`)
  - Also allowed to have `null`
- Objects: One or more name/value pairs, separated by commas, enclosed by curly braces
  - `{`*`name1:value1, name2:value12…`*`}`
  - Names must be strings
  - Names do not have to be unique
- Arrays: A sequence of comma-separated values, enclosed by square braces
  - `[`*`value1`*`, `*`value2`*`, …]`
  - Values will stay in the order provided

In-class exercise

Grab the notebook for today's lecture in D2L.

What is Big Data, anyways?

- So much data that traditional data management methods and tools are not useful for processing
- “The amount of data just beyond technology’s capability to store, manage, and process efficiently” (Kaisler *et al.)*
- Commonly summarized as:
    - Volume
    - Variety
    - Velocity

S. Kaisler, F. Armour, J. A. Espinosa and W. Money, "Big Data: Issues and Challenges Moving Forward," *2013 46th Hawaii International Conference on System Sciences*, Wailea, Maui, HI, 2013, pp. 995-1004

How big is big data?

Size	Description	Scale
1 bit	A single 1 or 0	1, 0
1 byte	8 bits, a single character	X
1 kilobyte (kB)	(1000 or 1024) bytes	140-character tweet
1 megabyte (MB)	10^6 bytes	1 minute of audio
1 gigabyte (GB)	10^9 bytes	Half an hour of video
1 terabyte (TB)	10^{12} bytes	Consumer hard drive
1 petabyte	10^{15} bytes	Human brain
1 exabyte	10^{18} bytes	Largest corporate/scientific/government databases
1 yottabyte	10^{21} bytes	Total sensor data from large scientific projects (LHC, SKA)