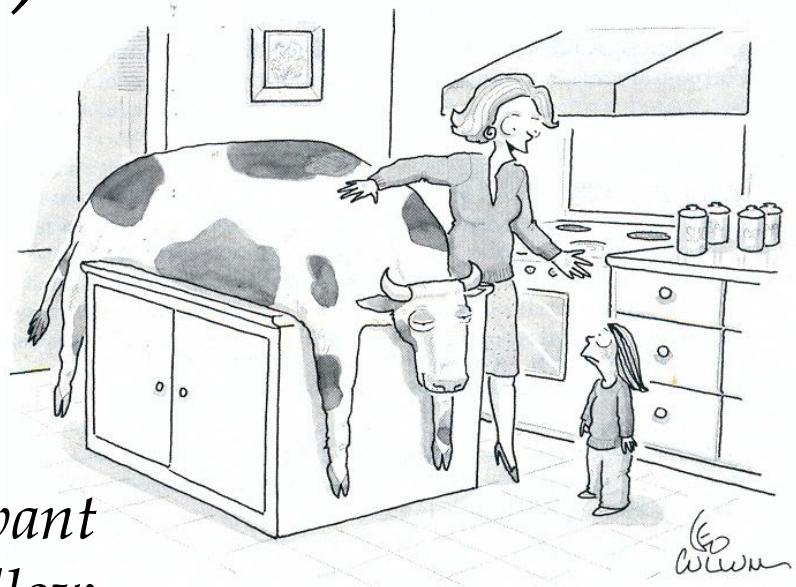
Exploring a Database (and exposing data)

Problem Sets 1 are due before midnight tonight in both sections.

Those in section 002 might want to access Jupyter today to follow along



"Mommy wants you to know where your food comes from."

Welcome Comp521-001!

What I know…

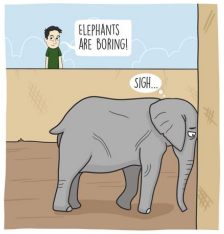
1) Professor Bishop fell ill and called off class last Thursday
2) He has since been hospitalized and expects to be incapcitated for AT LEAST this entire week.
3) Get problem set #1 turned in tonight

What I plan to do…

1) Use the next two lectures to sync up the two sections.
2) For today, assume that neither course has changed. I will be examining both courses syllabi and infrastructure

What if…

1) If Professor Bishop does not return, we will merge the two sections. I will also address discrepancies (grading, # of midterms, etc. …). Changes for both sections.
2) I hope to know more by Thursday



NCCOVID19.db from PS #2

jupyter ProblemSet02Play Last Checkpoint: 2 minutes ago (autosaved)

File Edit View Insert Cell Kernel Help

```
In [1]: import iSQL

query = iSQL.parser("NCCOVID19.db")
```

SQL: .schema

Submit prev

```
CREATE TABLE County (
   fips INTEGER PRIMARY KEY,
   name TEXT,
   region TEXT,
   cog TEXT,
   msa TEXT)

CREATE TABLE Covid19 (
   fips INTEGER,
   date DATETIME,
   cases INTEGER DEFAULT 0,
   deaths INTEGER DEFAULT 0,
   PRIMARY KEY (fips,date),
   FOREIGN KEY(fips) REFERENCES County(fips))

CREATE TABLE Demographics (
   fips INTEGER,
   year INTEGER,
   race TEXT,
   sex TEXT,
   agelo INTEGER,
   agehi INTEGER,
   count INTEGER,
   PRIMARY KEY (fips,year,race,sex,agelo),
   FOREIGN KEY(fips) REFERENCES County(fips))

CREATE TABLE Hospital (
   hid INTEGER PRIMARY KEY,
   name TEXT,
   city TEXT,
   beds INTEGER,
   icu INTEGER,
   discharges INTEGER,
   patientdays INTEGER,
   revenue REAL)

CREATE TABLE HospitalCounty (
   hid INTEGER,
   fips INTEGER,
   incounty INTEGER,
   FOREIGN KEY(hid) REFERENCES Hospital(hid),
   FOREIGN KEY(fips) REFERENCES County(fips))
```

Demographics

fips	year	race	sex	agelo	agehi	count

County

fips	name	region	cog	msa

Covid19

fips	date	cases	death

Hospital

hid	name	city	beds	icu	discharges	patientdays	revenue

HospitalCounty

hid	fips	incounty

You might want to...

- Open your PS #2
- Create a new cell to open the NCCOVID.db database

Simple Questions

- How many NC counties?

```
SELECT COUNT(*)
FROM County
```

- How many hospitals?

```
SELECT COUNT(*)
FROM Hospital
```

- How many races are used in the demographic data?

```
SELECT DISTINCT race
FROM Demographics
```

- How many years worth of demographics data are included.

```
SELECT DISTINCT year
FROM Demographics
```

More useful queries

- What were the 10 NC counties with the most COVID-19 confirmed cases on 2020-08-20?

```
SELECT C.name, V.date, V.cases
FROM County C, Covid19 V
WHERE C.fips=V.fips AND date="2020-08-20"
ORDER BY cases DESC
LIMIT 10
```

- What are the metropolitan areas are in each of these counties?
- On what days did Orange county have its most reported cases?

Another query

- Find the breakdown of North Carolina's 2020 population by race

```
SELECT D.race, SUM(D.count)
FROM Demographics D
WHERE D.year=2020
GROUP BY D.race
```

- Find the total 2020 populations by county

- List the all NC counties that in 2020 are not majority "white"

Queries Continued

- How many hosptials are in each county?

```
SELECT C.name, COUNT(*)
FROM County C, Hospital H, HospitalCounty HC
WHERE C.fips=HC.fips AND H.hid=HC.hid AND HC.incounty=1
GROUP BY C.fips
ORDER BY COUNT(*) DESC
```

- Modify the query above to include every county? (Hint: think LEFT JOIN)

Where are college age folks?

- List the population for the age group from 18 to 24 in every NC county in 2020

```
SELECT C.name, SUM(D.count)
FROM County C, Demographics D
WHERE C.fips=D.fips AND D.agelo >= 18 AND D.agelo <= 24
AND year=2020
GROUP BY D.fips
ORDER BY SUM(D.count) DESC
```

- Do the same query, but also give it as a ratio of the total population

Back to hospitals

- How many hospital and hospital beds are in each county?

```
SELECT C.name, Count(H.hid), SUM(H.beds)
FROM County C, Hospital H, HospitalCounty HC
WHERE C.fips=HC.fips AND H.hid=HC.hid AND HC.incounty=1
GROUP BY C.fips
ORDER BY SUM(H.beds) DESC
```

Can you trust the data?

- How many hospitals have no beds?

```
In [1]: import iSQL

query = iSQL.parser("NCCOVID19.db")
```

SQL:

```
SELECT *
FROM Hospital
WHERE beds=0
```

Submit | prev | next

	hid	name	city	beds	icu	discharges	patientdays	revenue
0	1006	Mission Saint Joseph Campus	Asheville	0	0	0	0	0.0
1	1018	Fayetteville VA Medical Center	Fayetteville	0	0	0	0	0.0
2	1019	Womack Army Medical Center	Fort Bragg	0	0	0	0	0.0
3	1026	Durham VA Medical Center	Durham	0	0	0	0	0.0
4	1030	Novant Health Kernersville Medical Center	Kernersville	0	0	0	0	0.0
5	1031	Brenner Children's Hospital	Winston-Salem	0	0	0	0	0.0
6	1037	Cone Health Wesley Long Hospital	Greensboro	0	0	0	0	0.0
7	1038	Cone Health Women's Hospital	Greensboro	0	0	0	0	0.0
8	1042	Central Harnett Hospital	Lillington	0	0	0	0	0.0
9	1048	FirstHealth Moore Regional Hospital-Hoke Campus	Raeford	0	0	0	0	0.0
10	1053	Johnston Health Clayton	Clayton	0	0	0	0	0.0
11	1060	Atrium Health Mercy	Charlotte	0	0	0	0	0.0
12	1074	Naval Hospital Camp Lejeune	Camp Lejeune	0	0	0	0	0.0
13	1080	Maynard Children's Hospital	Greenville	0	0	0	0	0.0
14	1088	W. G. (Bill) Hefner VA Medical Center - Salisb...	Salisbury	0	0	0	0	0.0
15	1099	Central Prison Hospital	Raleigh	0	0	0	0	0.0
16	1101	North Hospital	Raleigh	0	0	0	0	0.0

Atrium Health Mercy

Atrium Health

Geography	
Location	Charlotte, North Carolina, United States
Coordinates	
Organization	
Care system	Private, Medicaid, Medicare
Type	General and specialized
Services	
Emergency department	Yes
Beds	185
History	
Opened	1906
Links	
Website	https://atriumhealth.org/locations/detail/atrium-health-mercy
Lists	Hospitals in North Carolina

Actually, this isn't true. I grabbed the hospital data from a website where the hospitals voluntarily reported.

Cleaning the data

- Here is how we fix it

```
UPDATE Hospital SET beds=185 WHERE hid=1060
```

- Now, recompute the number of hospital beds in each county?

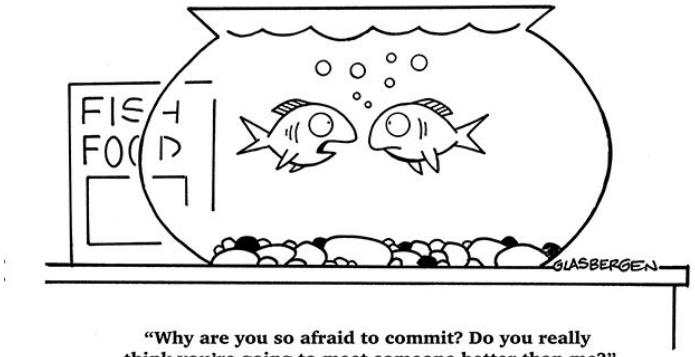
```
SELECT C.name, Count(H.hid), SUM(H.beds)
FROM County C, Hospital H, HospitalCounty HC
WHERE C.fips=HC.fips AND H.hid=HC.hid AND HC.incounty=1
GROUP BY C.fips
ORDER BY SUM(H.beds) DESC
```

Time to COMMIT

- No change is permanently made to a data base until it is commited!

- Now commit

```
COMMIT
```



"Why are you so afraid to commit? Do you really think you're going to meet someone better than me?"

- Now if we restart iSQL the fix will remain.

- Lack of COMMITs is a major source of bugs
  - "locked" database
  - loss of data

Errors, errors, everywhere

- Dirty data is everywhere
  - Some hospitals have closed
  - New ones added
  - The list of counties "served" by hospitals is incomplete
  - It takes a it of sleuthing to figure them out
- However, we need to be able to write correct queries in the presence of imperfect data
- This is the challenge and learning objective of your next problem set
  - I have fixed many errors that appear in the version of NCCOVID19.db that you have
  - I may have also probably introduced dozens more
- Your HW queries need to work correctly on my instance as well as yours.

Next Time

- Embedding SQL queries within a traditional programming language