WEEK 4: DATA COLLECTION AND FORMATS

Dr. Kai Li
School of Information Sciences
University of Tennessee, Knoxville
Spring 2025

Review of Week 3

- Different data types:
  - *categorical vs. numerical vs. anything in between (especially ordinal)*
  - *Boundaries between these categories are sometime vague and they can be transformed into different categories.*
- Research questions:
  - *It is an art and may not have the definitely best solution.*
- Descriptive statistics:
  - *The more basic aspect of statistics: offering summary of the data in one variable.*
  - *It is still helpful as the ground facts and basis of a research. (But we generally need more complex analyses in most cases.)*

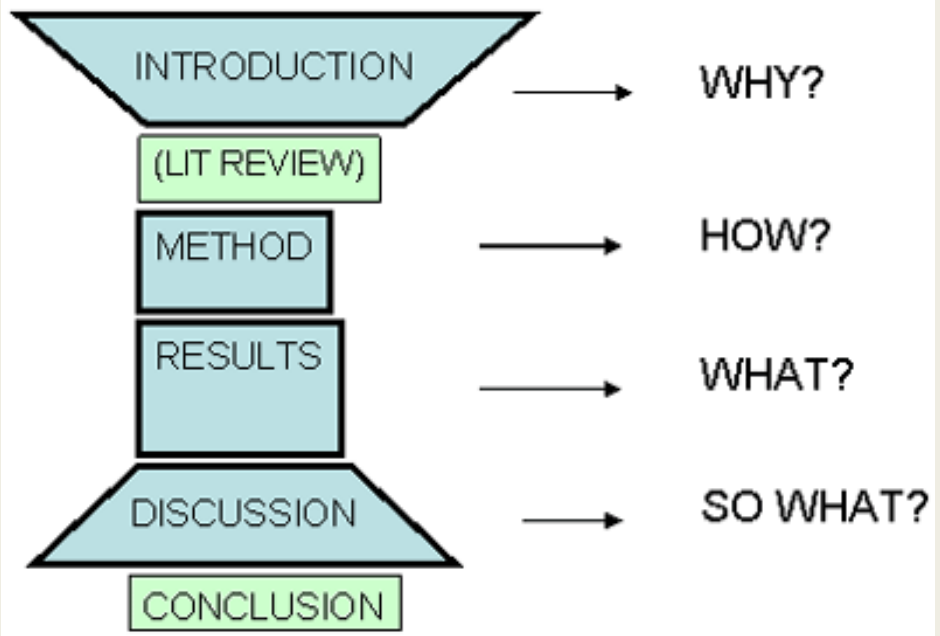
Review of Week 3

- Activity
- Other questions or thoughts?

- I will try not to talk about all the details in the demonstration but give you more time to play with the code and ask questions.

RQ

- A few comments on your answers:
  - *We probably don't want to introduce variables or variable names in the RQs.*
  - *The questions should be as straightforward as possible.*
  - *Can we answer the question?*
  - *One additional issue: double-barreled questions.*



Overview of this week

- Data sources: where to find datasets
  - *Data repositories*
  - *API*
  - *Web scrapping*
- Data formats: various data formats (including database)

Data collection

- Data collection is the first step of data analysis.
- It’s been much easier during the past few years thanks to the open data movement.
- We can find a lot of new data repositories to get access to many different kinds of data.

Data sources

- Some examples include:
  - *Kaggle*
  - *GitHub / Zenodo*
  - *Google Dataset Search:* *https://datasetsearch.research.google.com/*
  - *ICPSR:* *https://www.icpsr.umich.edu/web/pages/*
  - *Dryad:* *https://datadryad.org/*
  - *DataCite:* *https://commons.datacite.org/*
  - *…*

Evaluate your data!

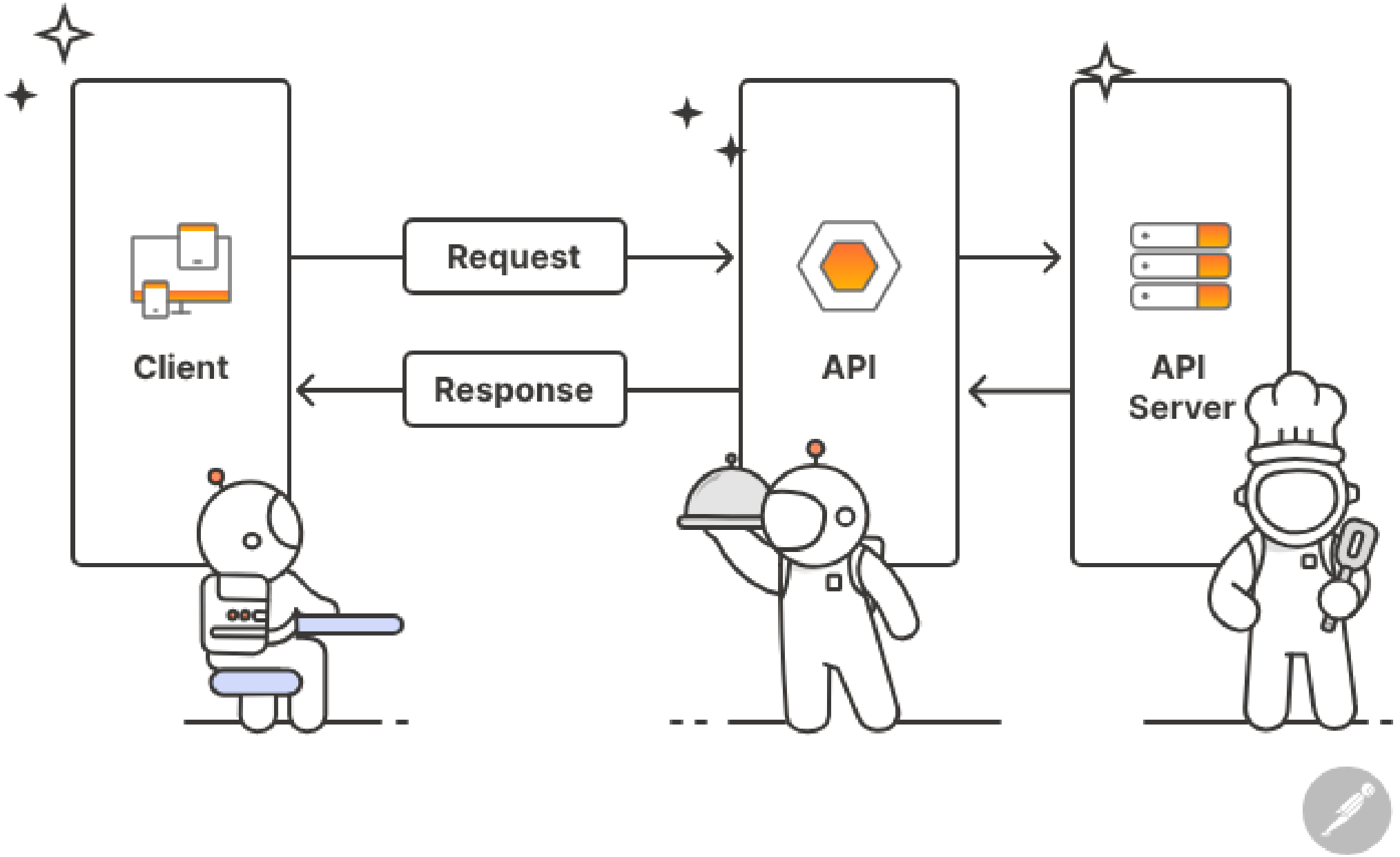
- We need to ask many questions to make sure that the data is usable, relevant and reliable:
  - *Who collected the data?*
  - *How was the data processed before it is published?*
  - *When was the data collected?*
  - *Why was the data collected?*
  - *How was the data collected, on the dataset- and variable-levels?*

Some other ways of collecting data

- We will discuss some other ways of collecting data, such as web scrapping and APIs.
- But using structured databases is still a very easy and useful way to get your data, especially when it is an option.

API

- API: application programming interface
  - *The interface to application through a programming language*
  - *It is normally the more ethical, easier, and safer access to an online service.*
- Many information services offer the API access.
  - *APIs are more than just downloading the data: we can also use APIs to upload data, search in the database, or even update information in the system, et al.*


Client
Request
Response
API
API
Server

API services in R packages

- Many API services are already packaged in R or Python libraries.
  - *Such as the ngramr package.*
- rOpenSci project has a pretty good and curated list of API packages (covering many different types of data sources) in R:
  - *https://ropensci.org/packages/data-access/*
- An additional list: https://gist.github.com/zhiiiyang/fc19995f7e350f3c7fb940757f6213cf
- OpenAlex: https://docs.openalex.org/how-to-use-the-api/api-overview

APIs not in R packages

- But in the cases where there is not package, we will need to use the links provided by API to get access to the data.
  - *For example, the API provided by the MET Museum (we will play with it in the demonstration): https://metmuseum.github.io/*

Additional comments on API

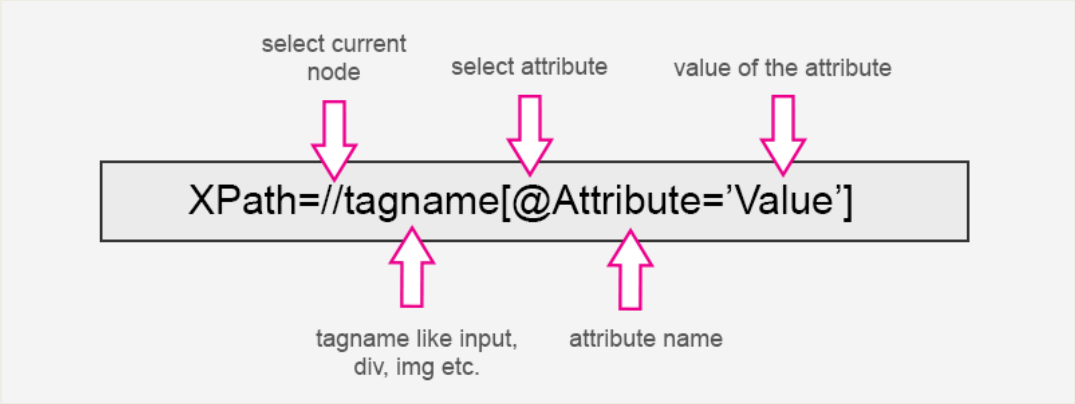
- Most of the APIs should have very detailed documentation.
- Most of the APIs should have limitations (such as the number of queries you can use or the size of data you can download in a period).
- And they may require you to have an account of the service first.
  - *Not every API is for free.*

Web scrapping

- We can directly extract data from the source file of web pages (i.e., HTML file), if:
  - *There is no barrier to the content from the file (such as the website using JavaScript code).*
    - There are more advanced techniques for this scenario though.
  - *The website or service does not offer any API service to more effectively download data.*
    - APIs generally offer easier and more ethical access to the data source if it’s available.

Web scrapping

- The idea is that we can extract certain information from the HTML pages by using patterns expressed by the XML Path Language (XPath).
- We will show an example in R. But Python has more powerful libraries to scrap data from websites (such as *Selenium*).
  - *There is an R version of Selenium, called "selenider".*
  - *https://medium.com/technology-nineleaps/how-to-use-selenium-in-r-b4c92cc3be70*

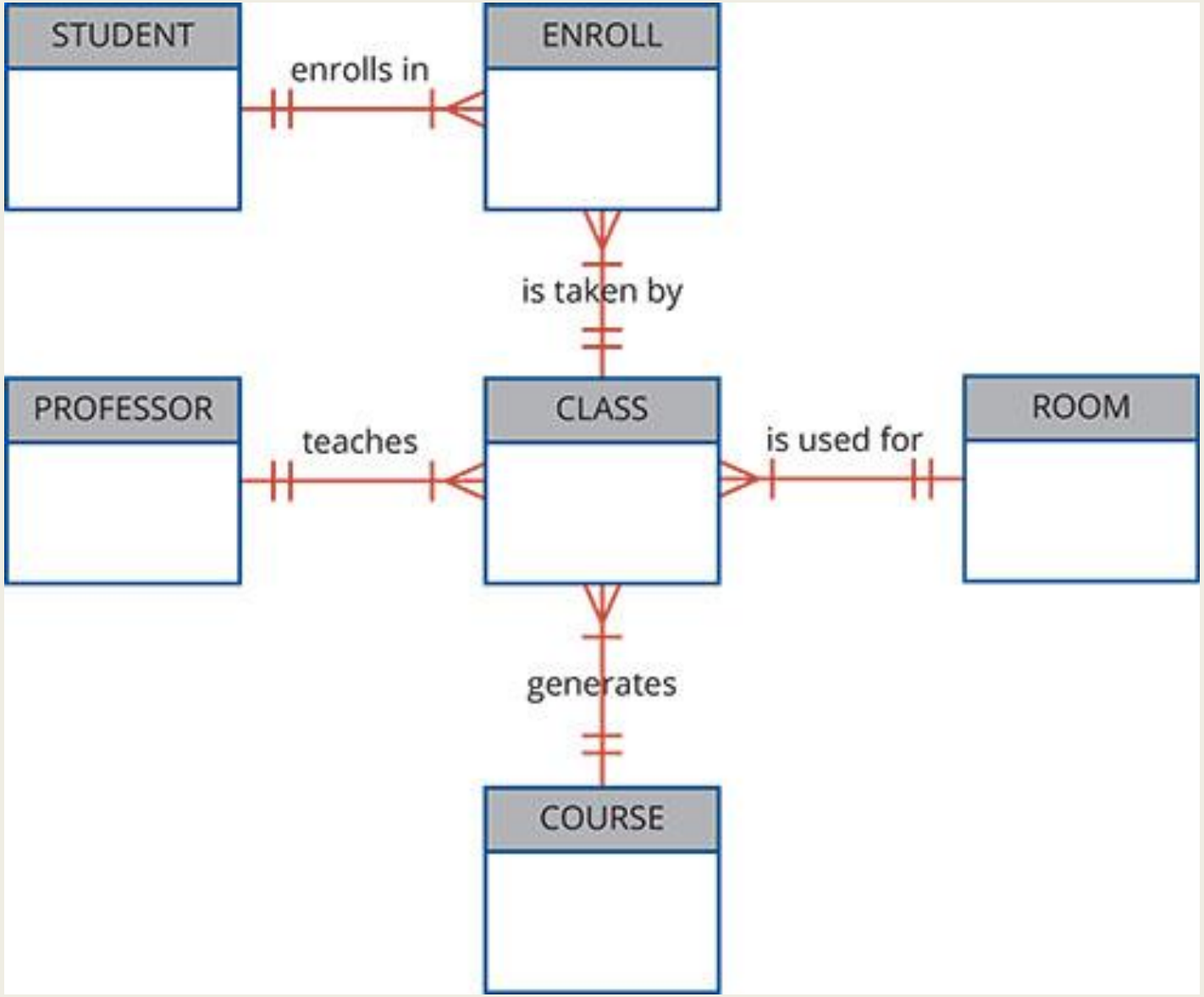


Data formats

- Some popular data formats:
  - *CSV*
  - *JSON*
  - *XLS / XLSX*
  - *TXT*
  - *HTML / XML – more likely textual data*
  - *Databases*

Database

- There are more models of constructing databases today, but relational database (SQL database) is still one of the most popular ones in the industry.
- A **relation** is a two-dimensional structure composed of intersecting rows and columns.
  - *Each row represents an instance.*
  - *Each column represents an attribute.*
- The relational data model is implemented through relational database management systems (RDBMS).



Structure of the database

Class Roster

Course: MGT 500 Business Policy Semester: Spring 2010

Section: 2

Name	ID	Major	GPA
Baker, Kenneth D.	324917628	MGT	2.9
Doyle, Joan E.	476193248	MKT	3.4
Finkle, Clive R.	548429344	PRM	2.8
Lewis, John C.	551742186	MGT	3.7
McFerran, Debra R.	409723145	IS	2.9
Sisneros, Michael	392416582	ACCT	3.3

Course	Title	Credit	Level
MGT 500	Business Policy	3	Gr
INSC 504	Research Methods	4	Gr

RegNo	Student	Course
1	324917628	12354
3	476193248	12354
4	548429344	12354
6	392416582	12354
10	551742186	12354
12	409723145	12354
33	409723145	22345
35	324917628	22345

ID	LastName	FirstName	Middle	Major	GPA
324917628	Baker	Kenneth	David	MGT	2.9
392416582	Sisneros	Michael		ACCT	3.3
409723145	McFerran	Debra	Raya	IS	2.9
476193248	Doyle	Joan	Eva	MKT	3.4
548429344	Finkle	Clive	Rikard	PRM	2.8
551742186	Lewis	John	Carl	MGT	3.7

Course	Term	Year	Section
MGT 500	Spring	2010	12354
INSC 504	Fall	2010	22345

Table name: AGENT (first six attributes) Database name: Ch02_InsureCo

AGENT_CODE	AGENT_LNAME	AGENT_FNAME	AGENT_INITIAL	AGENT_AREACODE	AGENT_PHONE
501	Alby	Alex	B	713	228-1249
502	Hahn	Leah	F	615	882-1244
503	Okon	John	T	615	123-5589

Link through AGENT_CODE

Table name: CUSTOMER

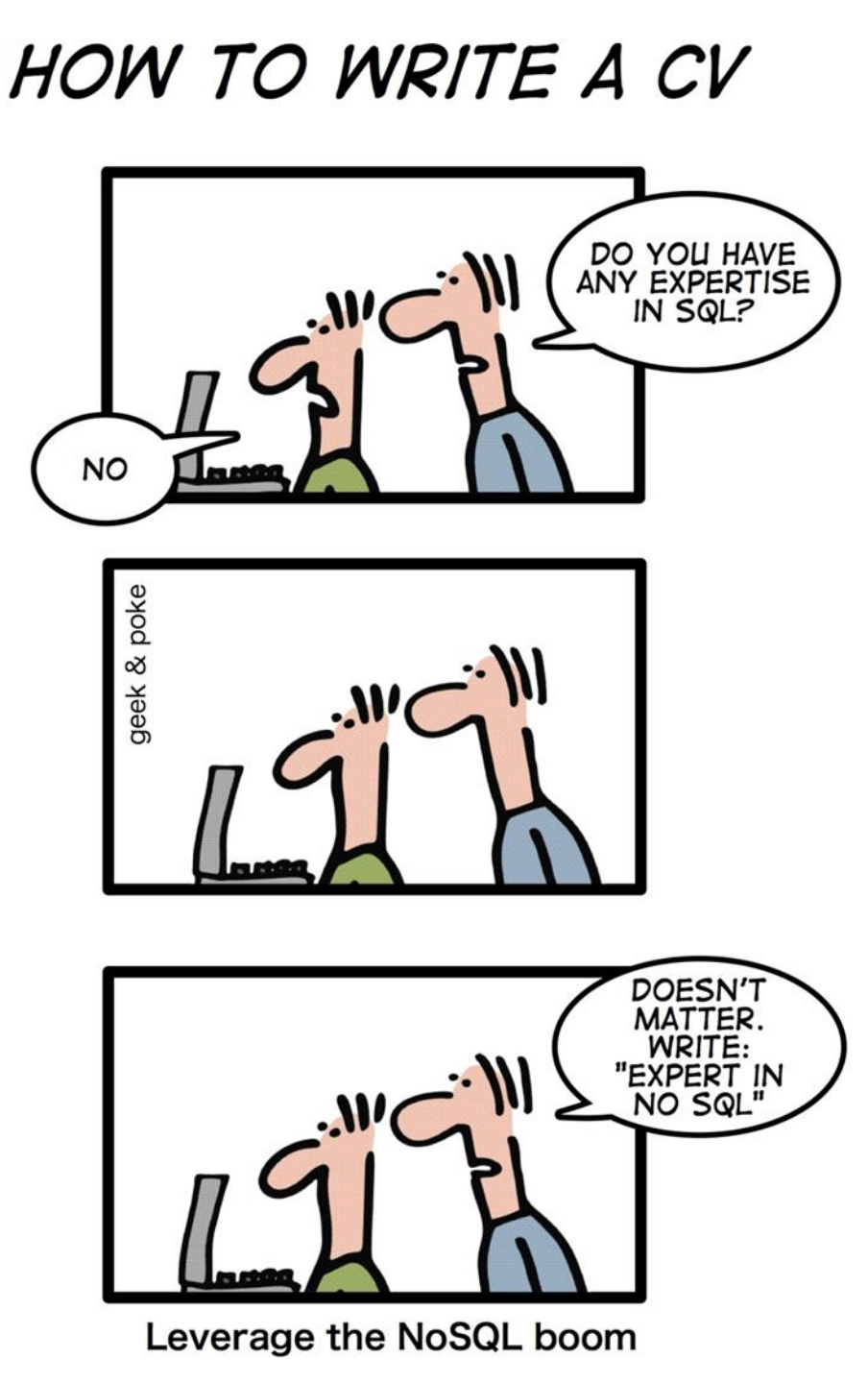
CUS_CODE	CUS_LNAME	CUS_FNAME	CUS_INITIAL	CUS_AREACODE	CUS_PHONE	CUS_INSURE_TYPE	CUS_INSURE_AMT	CUS_RENEW_DATE	AGENT_CODE
10010	Ramas	Alfred	A	615	844-2573	T1	100.00	05-Apr-2018	502
10011	Dunne	Leona	K	713	894-1238	T1	250.00	16-Jun-2018	501
10012	Smith	Kathy	W	615	894-2285	S2	150.00	29-Jan-2019	502
10013	Olowski	Paul	F	615	894-2180	S1	300.00	14-Oct-2018	502
10014	Orlando	Myron		615	222-1672	T1	100.00	28-Dec-2019	501
10015	O'Brian	Amy	B	713	442-3381	T2	850.00	22-Sep-2018	503
10016	Brown	James	G	615	297-1228	S1	120.00	25-Mar-2019	502
10017	Williams	George		615	290-2556	S1	250.00	17-Jul-2018	503
10018	Farriss	Anne	G	713	382-7185	T2	100.00	03-Dec-2018	501
10019	Smith	Olette	K	615	297-3809	S2	500.00	14-Mar-2019	503

Linking between tables

- We may need to connect data tables in the project unless the only table we get is multivariate (i.e., containing all variables that we need)!
- And we will talk about issues of connecting data tables in the next week.

Other types of databases: NoSQL

- Non-relational (NoSQL) databases have been increasingly used in the new data-driven technological landscape.
- This concept covers all kinds of databases that do not follow RDBMS principles.
  - *It is an umbrella terms that covers a broad spectrum of practices in the big data paradigm.*



Demonstration

- 1. Data formats
- 2. API
- 3. Web scrapping
- 4. Activity (next page)

Activity 1: Data for your final assignment

- I want you to start thinking about the questions / data your want to use in your final project!
  - *Ideally, I hope this is a project that (1) can be helpful for your career and/or other projects, (2) you are interested in, and (3) something you can present in your portfolio.*
  - *While it does not have to be a super big/complex project, I hope you can also challenge yourself in the final.*
  - *Possibilities…*
- You don’t really need to do anything this week but to explore some APIs and data sources and start thinking.

Activity 2: Try an API to search and/or download data

- Feel free to find another API that you are interested in and try to download some data (say maybe a few hundred records).
  - *Or, if you are not totally comfortable with moving to a new API, you can certainly keep exploring the MET Museum API!*
  - *Feel free to use the list above or find another one yourself!*
- If you want, feel free to run some basic descriptive analysis on the data that you collected!
  - *If you post this week, I will read your report and leave my comments!*