PA 446

Will be starting at 6:05pm to give folks time to sign on

PA 446

Coding for Civic Data Applications

Course Logistics

Course Logistics

Prerequisites

1. **PA 434 is strongly recommended**
2. Only ~13 returnings students took PA 434 and enrollment size larger than expected
3. The course assumes that all students have prior experience working in R, ggplot and tidyverse
4. All grading standards will also assume this
5. There will only be minimal review of prior content
6. PA 446 will likely be offered again in the Spring, so please consider taking it then if you do not feel confident in your ability to write code in R

Course Logistics

Instructional Format

- Due to the Delta variant and in order to ensure all students can attend class without concerns about their health, **courses will be fully online until further notice**
  - Blackboard > PA 446 > “Online class links” > [Today’s Date]
- Possible in-person classes later in semester depending on COVID rates and your fellow students’ comfort - will send out poll after class to gauge current comfort
- Class recordings will be available via Blackboard

Course Logistics

Grades

- Homework exercises (5 total at 12% per assignment, 60%)
- Data science “midterm” (20%)
- Final project (20%)
- Class attendance (10%)
    - Since courses will be online, I am waiving the class attendance grade

Course Logistics

Grades

- You have to code in R
  - Failure to do so results in maximum grade of C
- The majority of data manipulation need to be done in tidyverse, not base R
  - Failure to do so results in maximum grade of C
- For homeworks
  - A: all answers are accepted and majority of R commands are correct
  - B: majority of answers are accepted and some incorrect R syntax
  - C: some incorrect answers and significant incorrect R syntax
  - Descriptions purposefully vague as I assess proficiency in R

Course Logistics

Homework

- **Where to get:** Blackboard > PA 446 > “Homework Assignments” module page
- **What you need to know:** homework will be based on the reading for that week. So make sure you do your readings.
- **Due date/time:** Wednesdays, before when the class starts (6pm). See syllabus for details
- **Submission Format**
  - **Code:** please type your code into the original homework R script and re-submit that R script. Please include your final
  - **Answers:** I will provide a Google Form

Course Logistics

Class on 9/15 and 11/24

9/15 alternative date/time:

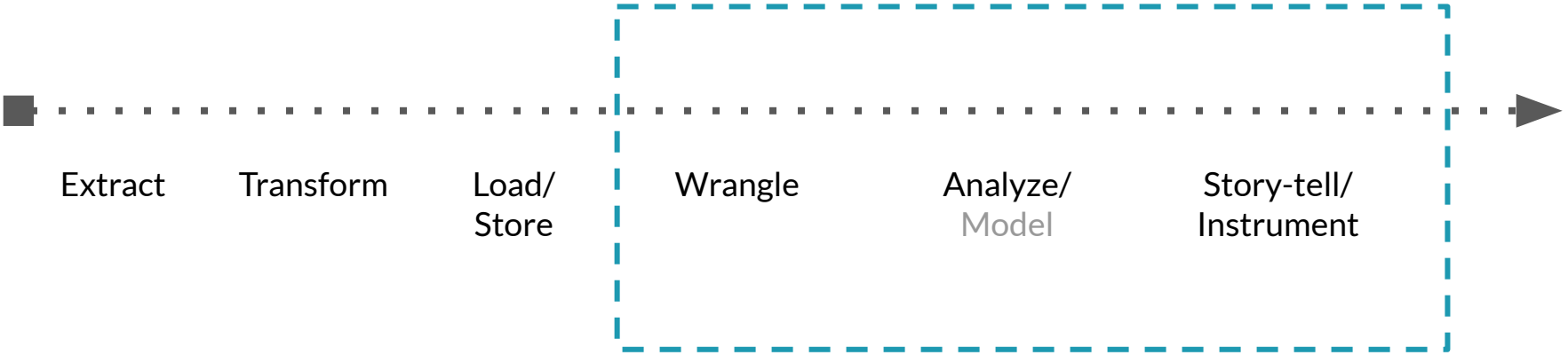
- Friday 9/10, 6-8:50pm
- Tuesday 9/14, 6-8:50pm
- Thursday 9/16, 6-8:50pm
- Friday 9/17, 6-8:50pm

All else fails, I will pre-record the class and you can watch it at your convenience

11/24 (day before Thanksgiving): there will be content that day, please plan to attend

Course Introduction

Data Science “workflow”



Focus of PA 446

Course Content and Timeline

Rough Timeline	Themes	Description
8/25 - 9/15	Wrangle	Data cleaning and transformation
9/22 - 10/6	Analyze/Model	Data analysis and causal inference
10/13 - 10/27	Story-tell	Data presentation
11/3 - 12/1	Instrument	Data dashboarding

About Me

Before Graduate School

- Architect
- Helped to start a nonprofit
- St. Louis

During Graduate School

- MBA
- Master of Public Policy
- University of Michigan

After Graduate School

- Senior Applied Data Scientist @ Civis Analytics

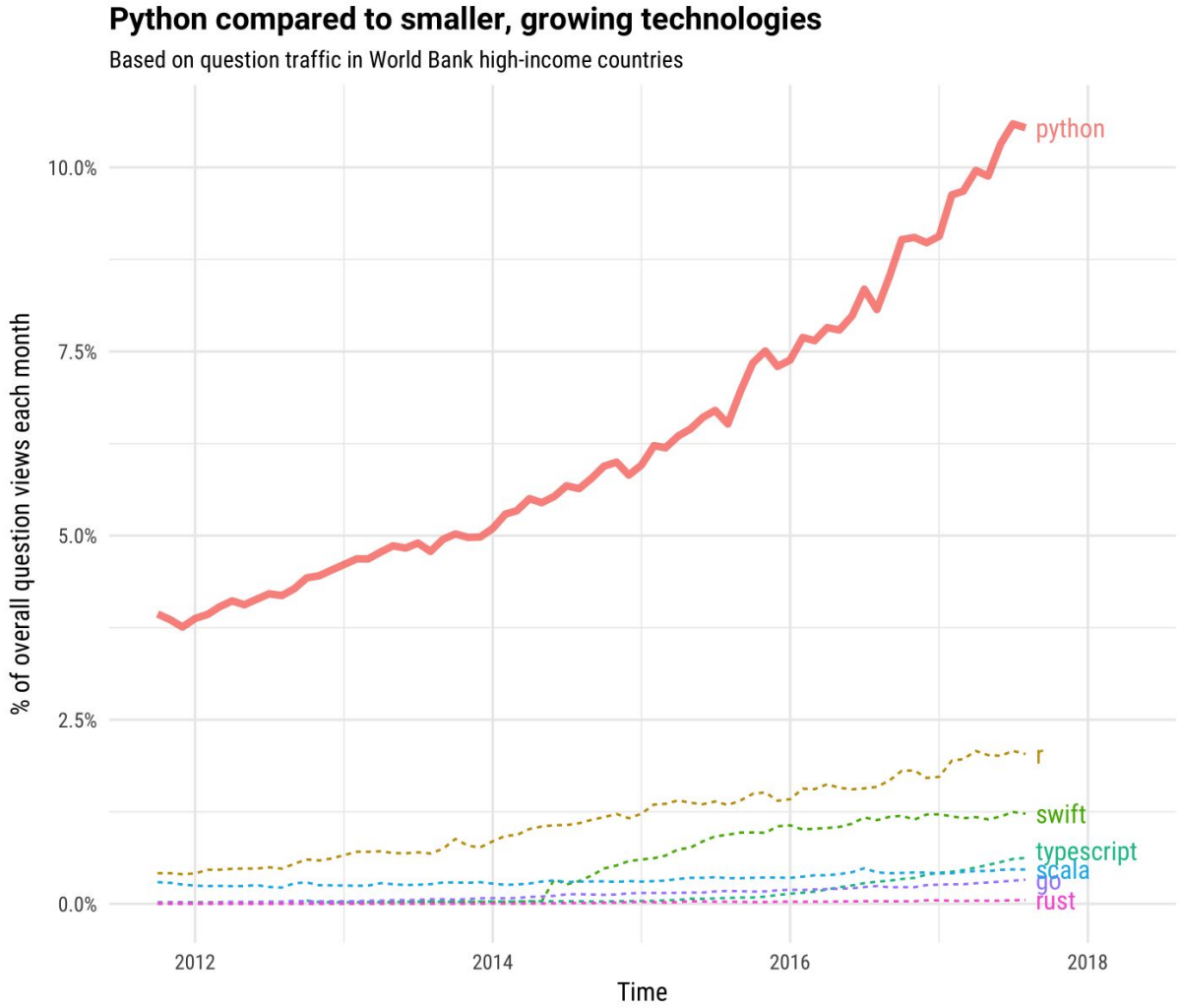
Throughout

- Instructor (Washington University in St. Louis and University of Michigan)

My Ask:

R, Not My Primary Language

1. Your understanding and patience - I'm going to forget syntax
2. Python: don't sleep on it

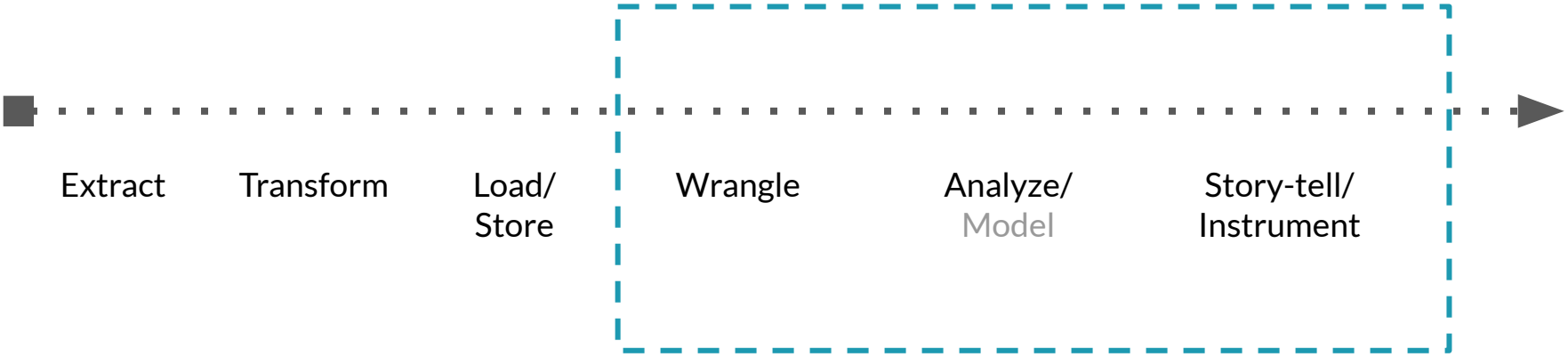


15-min break

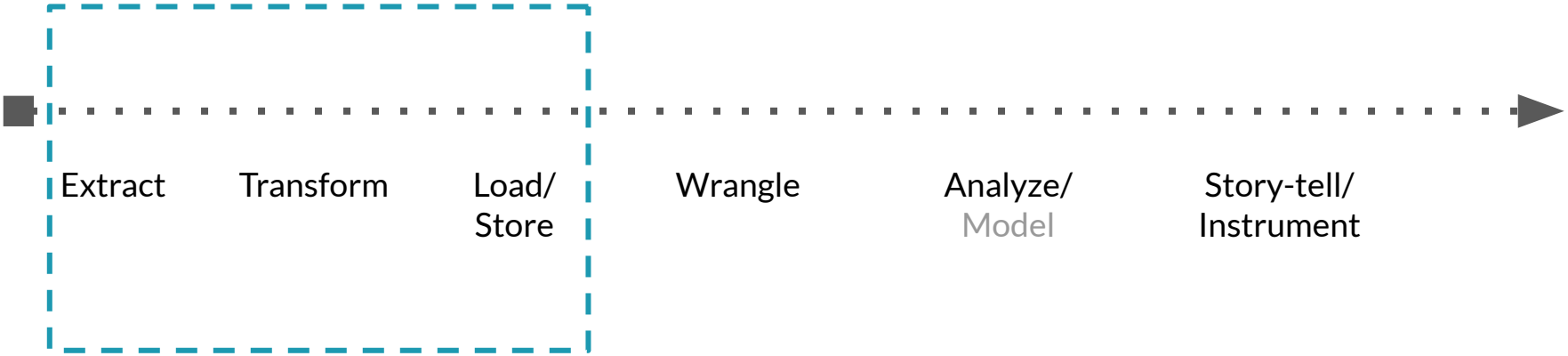
Please be back at 7pm

Data Engineering

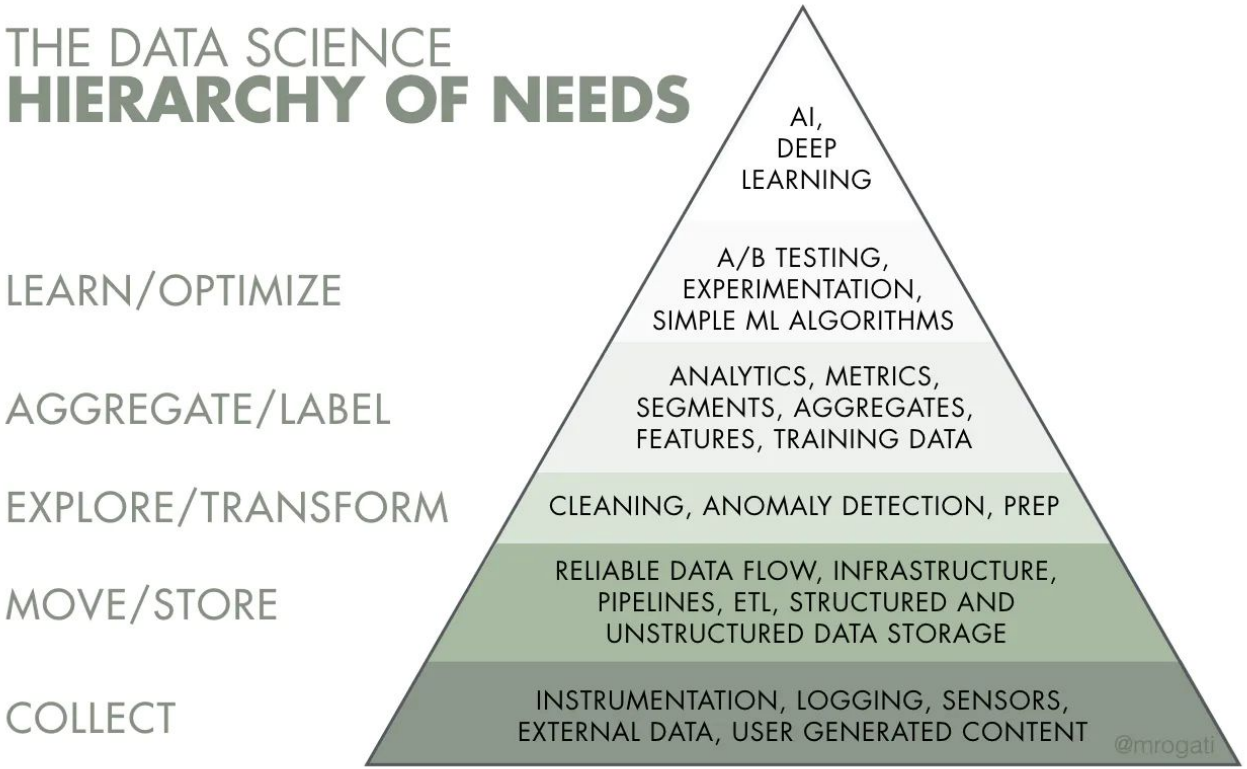
Job Titles in the Public Sector



Data Engineering



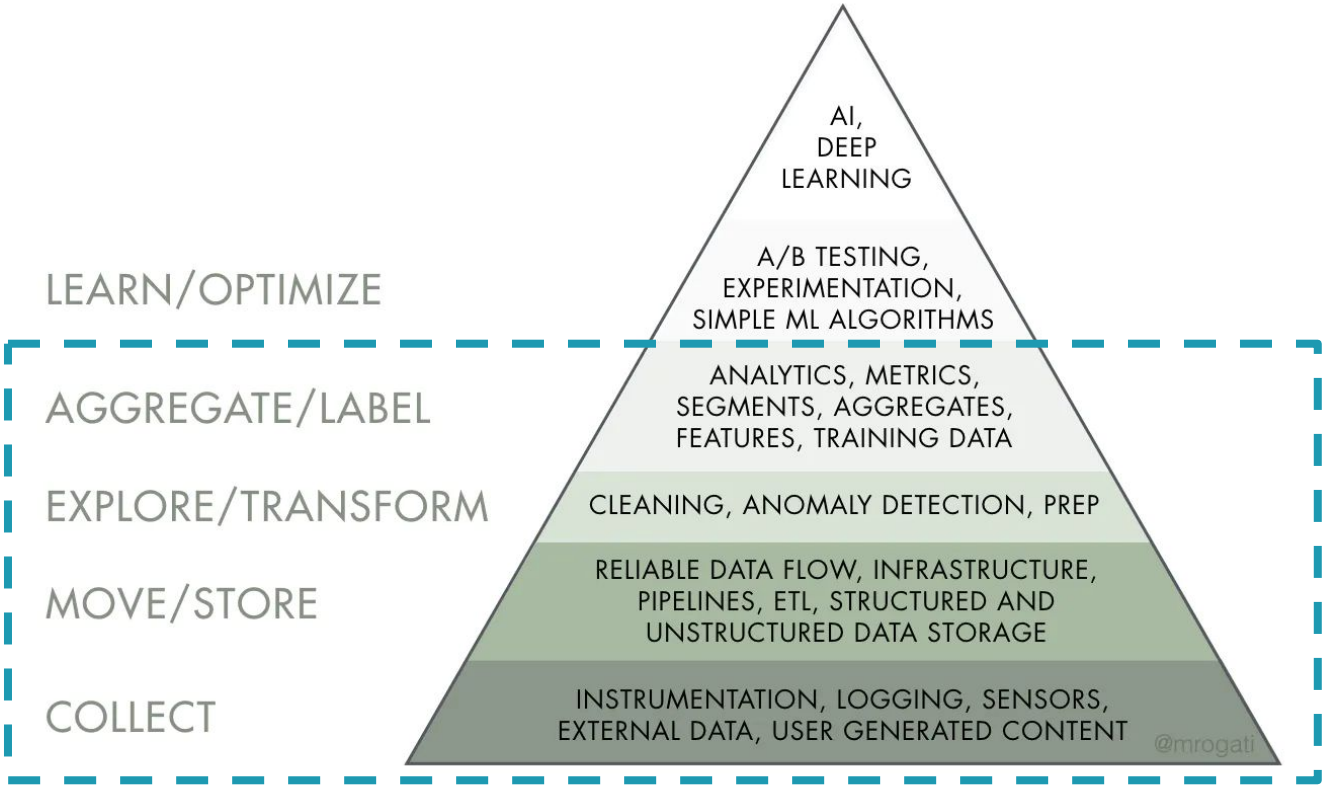
Focus of Today



Take-aways

- Data scientists focus on the top 3 tiers
- Data engineers focus on the bottom 3 tiers
- More work to build the bottom of the pyramid than the top

Public Sector's Biggest Needs

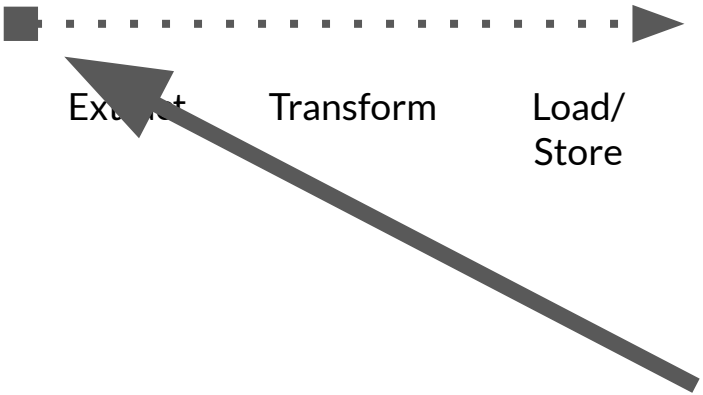


Take-aways

- Local government's data needs are often more basic: reliable ways to collect, move, store data. Once this is done, simple analysis will often suffice
- An unstated but critical need is to build a culture around decision making with data

Questions?

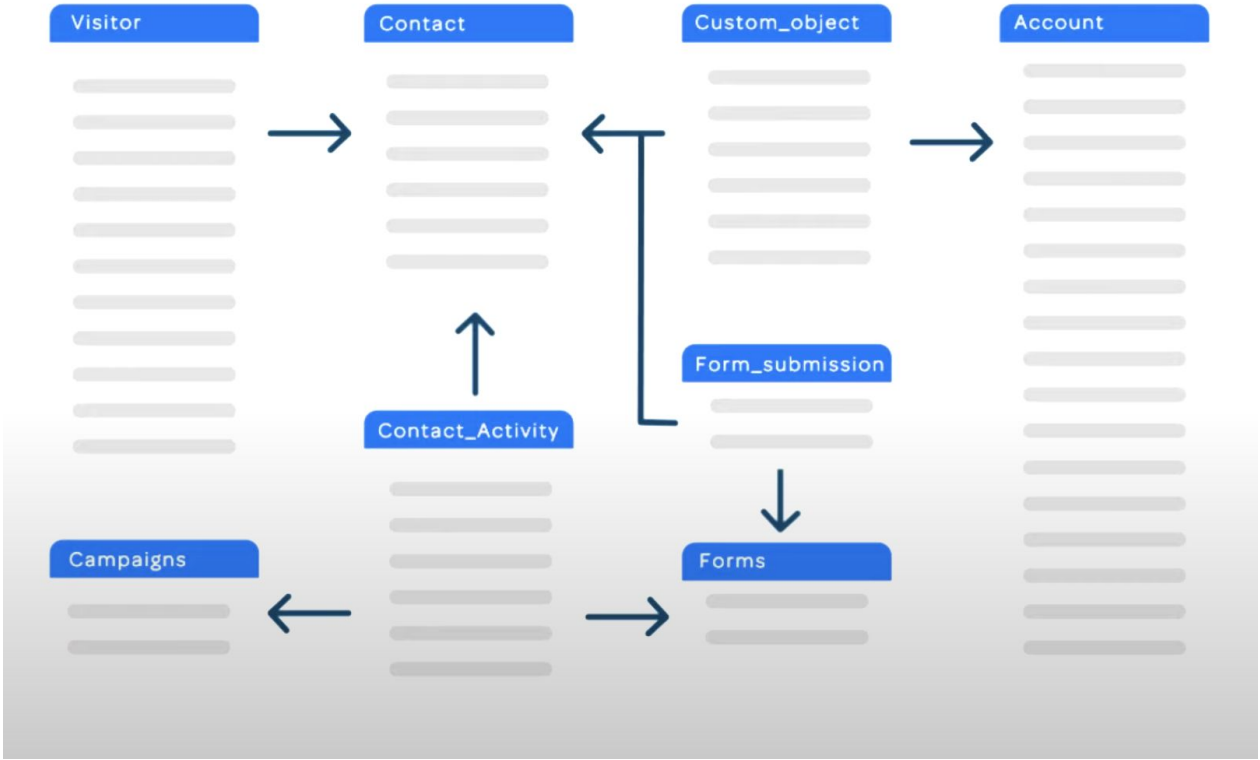
Data Engineering



How does it all begin?

Many Small Tables

"Normalize Schema"



A lot of small tables

Chicago's 311 Service

311 is Chicago's customer service department. Where anyone can call in with questions about the city

How might the ETL process work in order for Chicago to keep track of all the incoming calls

Chicago's 311 Service

Phone Schema	
Phone Number	**CREATED_DATE**
555-555-5551	08/22/2021 9:14:44 AM
555-555-5552	08/22/2021 9:14:41 AM
555-555-5553	08/22/2021 9:14:01 AM
555-555-5554	08/22/2021 9:13:51 AM

Who is calling?

When are they calling?

Chicago's 311 Service

311 Operator Schema		
ISSUE	**OWNER_DEPARTMENT**	**STREET_ADDRESS**
311 INFORMATION ONLY CALL	311 City Services	2111 W Lexington ST
Clean Vacant Lot Request	Streets and Sanitation	402 S KOLMAR AVE
311 INFORMATION ONLY CALL	311 City Services	2111 W Lexington ST
311 INFORMATION ONLY CALL	311 City Services	2111 W Lexington ST

What are they calling about?

Chicago's 311 Service

What is the outcome of the call?

311 Ticket Management Schema		
LAST_MODIFIED_DATE	**STATUS**	**CLOSED_DATE**
08/22/2021 9:14:44 AM	Completed	08/22/2021 9:14:44 AM
08/22/2021 9:14:43 AM	Open	
08/22/2021 9:14:01 AM	Completed	08/22/2021 9:14:01 AM
08/22/2021 9:12:39 AM	Open	

Chicago's 311 Service

Phone Schema	
Phone Number	**CREATED_DATE**
555-555-5551	08/22/2021 9:14:44 AM
555-555-5552	08/22/2021 9:14:41 AM
555-555-5553	08/22/2021 9:14:01 AM
555-555-5554	08/22/2021 9:13:51 AM

311 Operator Schema		
ISSUE	**OWNER_DEPARTMENT**	**STREET_ADDRESS**
311 INFORMATION ONLY CALL	311 City Services	2111 W Lexington ST
Clean Vacant Lot Request	Streets and Sanitation	402 S KOLMAR AVE
311 INFORMATION ONLY CALL	311 City Services	2111 W Lexington ST
311 INFORMATION ONLY CALL	311 City Services	2111 W Lexington ST

311 Ticket Management Schema		
LAST_MODIFIED_DATE	**STATUS**	**CLOSED_DATE**
08/22/2021 9:14:44 AM	Completed	08/22/2021 9:14:44 AM
08/22/2021 9:14:43 AM	Open	
08/22/2021 9:14:01 AM	Completed	08/22/2021 9:14:01 AM
08/22/2021 9:12:39 AM	Open	

Extract, Transform, Load (ETL)

Extract: Sensor Collecting the Raw Data

“Extraction is the action of extracting data from a source system to be processed at a later stage. This step is focused on obtaining data as efficiently and with as little impact to the source system as possible.”

Source System

- Surveys/polls
- Application logs
- Cameras
- What else?

Extract, Transform, Load (ETL)

Transform: Conversion from Raw to Useful Data

The extracted data is transformed to be more useful to tasks downstream. The transformation includes, but aren't limited to:

- Data cleaning and validation to ensure only quality data is migrated to the new system
- Combining or merging data from multiple source systems and deduplicating that data
- Applying business rules to data

› Examples

Location Schema

WARD	POLICE_DISTRICT	LAT	LONG
28	11	41.875	-87.739
10	4	41.699	-87.545
1	14	41.914	-87.690
48	20	41.980	-87.668

Extract, Transform, Load (ETL)

Load: Saving AND Making Available for Future Use

"The Load step concludes the ETL process with the loading of the extracted and transformed data into the end system"

End System Examples

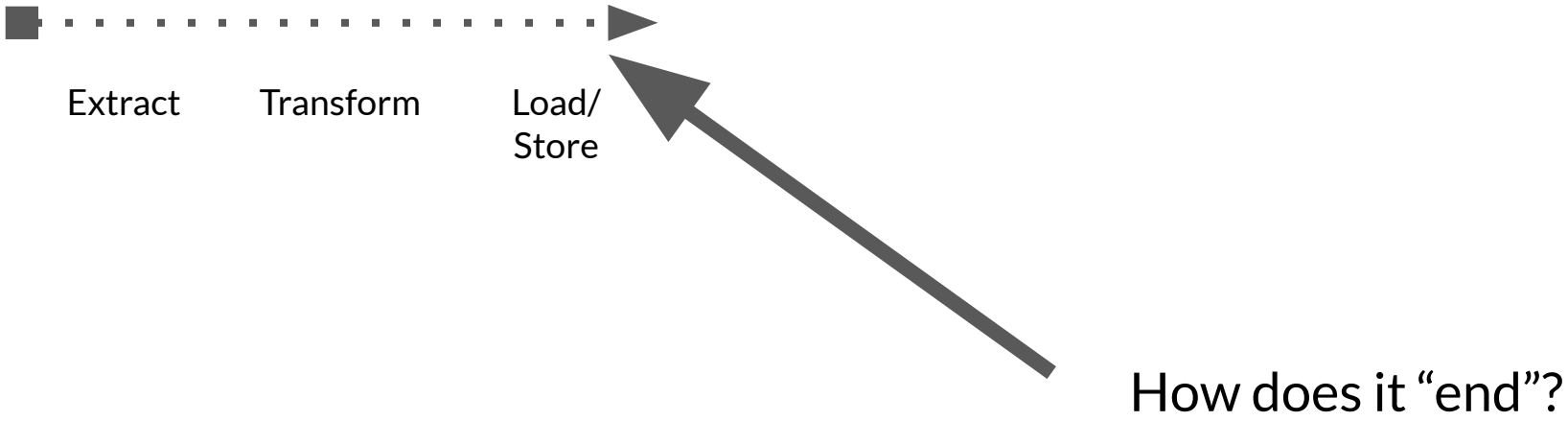
- AWS
- Google Cloud
- On premise servers

When you are trying to share a file with your friends, where do you upload it?

How do you ensure only the right people access the file?

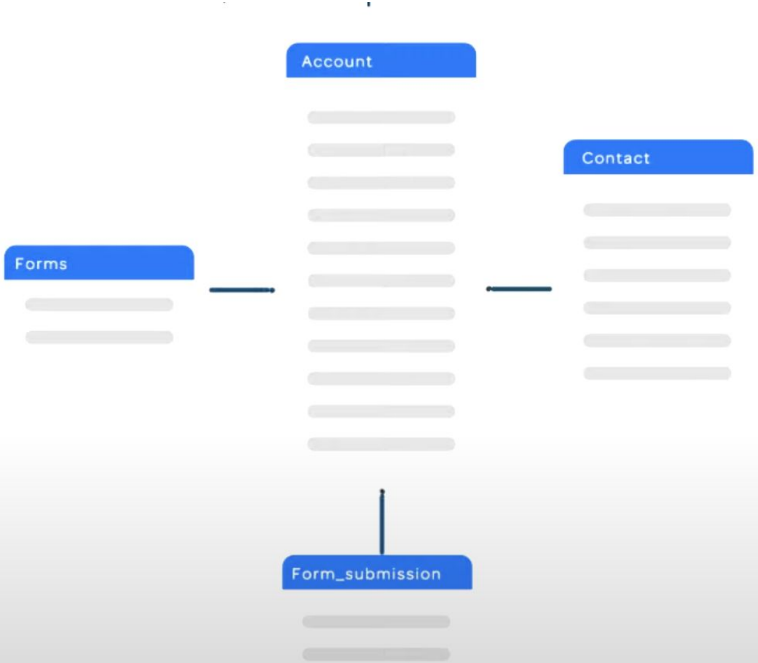
How do you keep your data safe once you load it?

Data Engineering



Fewer Simpler Tables

“Dimensional Schema”/Data Warehouse

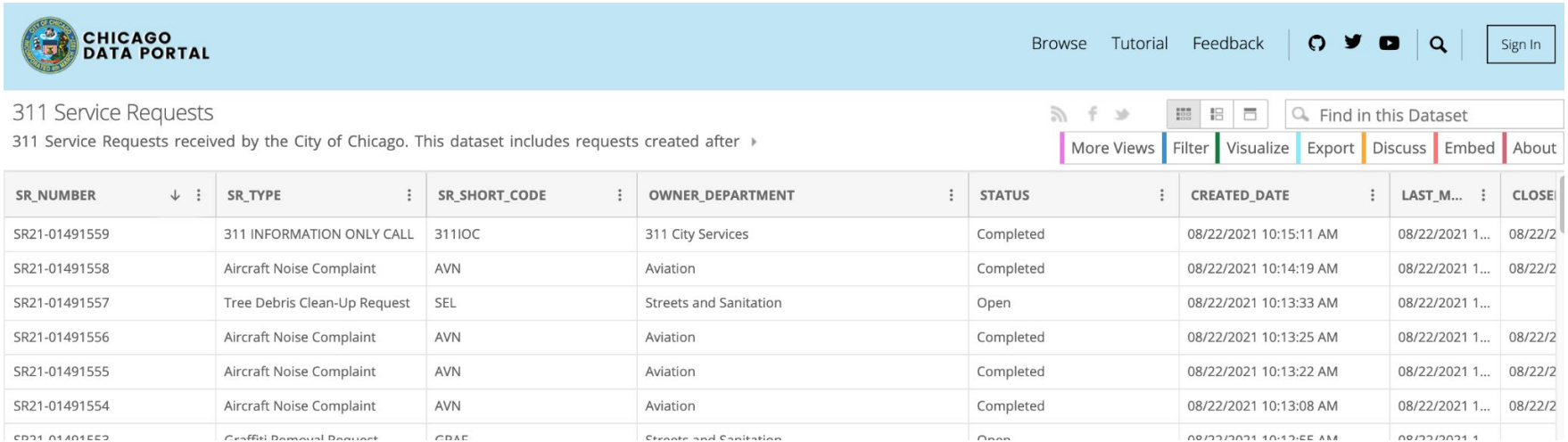


Examples

https://data.cityofchicago.org/Service-Requests/311-Service-Requests/v6vf-nfxy/data

https://data.cityofchicago.org/Service-Requests/311-Service-Requests-Request-Types/dgc7-2pdf/data

Chicago's 311 Service



SR_NUMBER	SR_TYPE	SR_SHORT_CODE	OWNER_DEPARTMENT	STATUS	CREATED_DATE	LAST_M...	CLOSE
SR21-01491559	311 INFORMATION ONLY CALL	311IOC	311 City Services	Completed	08/22/2021 10:15:11 AM	08/22/2021 1...	08/22/2
SR21-01491558	Aircraft Noise Complaint	AVN	Aviation	Completed	08/22/2021 10:14:19 AM	08/22/2021 1...	08/22/2
SR21-01491557	Tree Debris Clean-Up Request	SEL	Streets and Sanitation	Open	08/22/2021 10:13:33 AM	08/22/2021 1...	
SR21-01491556	Aircraft Noise Complaint	AVN	Aviation	Completed	08/22/2021 10:13:25 AM	08/22/2021 1...	08/22/2
SR21-01491555	Aircraft Noise Complaint	AVN	Aviation	Completed	08/22/2021 10:13:22 AM	08/22/2021 1...	08/22/2
SR21-01491554	Aircraft Noise Complaint	AVN	Aviation	Completed	08/22/2021 10:13:08 AM	08/22/2021 1...	08/22/2
SR21-01491553	Graffiti Removal Request	GRAF	Streets and Sanitation	Open	08/22/2021 10:12:55 AM	08/22/2021 1...	

311 Service Requests - Request Types

Based on 311 Service Requests

A list of the unique Service Request types present in the underlying 311 Service Requests dataset.

SR_TYPE	SR_SHORT_CODE	Count of Recor
311 INFORMATION ONLY CALL	311IOC	1,809,202
Abandoned Vehicle Complaint	SKA	86,168
Aircraft Noise Complaint	AVN	778,929
Alley Light Out Complaint	SFA	52,897
Alley Pothole Complaint	PHB	29,009

311 Service Requests - Abandoned Vehicles

All open abandoned vehicle complaints made to 311 and all requests completed since January 1,

Creat...	Status	Completion Da...	Service...	Type of Service Requ...	License ...	Vehicle M...
04/27/2020	Open		20-00034520	Abandoned Vehicle Compl...	PAA2336	Acura
06/29/2020	Open		20-00046108	Abandoned Vehicle Compl...	NO PLATES	Ford
06/12/2020	Open		20-00042834	Abandoned Vehicle Compl...	NO PLATES	Bmw
06/21/2020	Open		20-00044400	Abandoned Vehicle Compl...	AZ 42642	Acura
06/28/2020	Open		20-00045804	Abandoned Vehicle Compl...	INDIANA PLA...	Pontiac

Reliability Is Key

- Automatic - no manual copy and paste
- Scalable
- Auditable
- Secure

The growth of data scientists has been largely fueled by the growth of data. Without data engineers to reliably pre-process that data, data scientists cannot do their jobs

To Do’s for Next Week

- Today’s slides will be posted on to Blackboard after class
- Readings and HW 1 assignment, due next week, available on Blackboard
- Take the poll here.

Final Questions?