

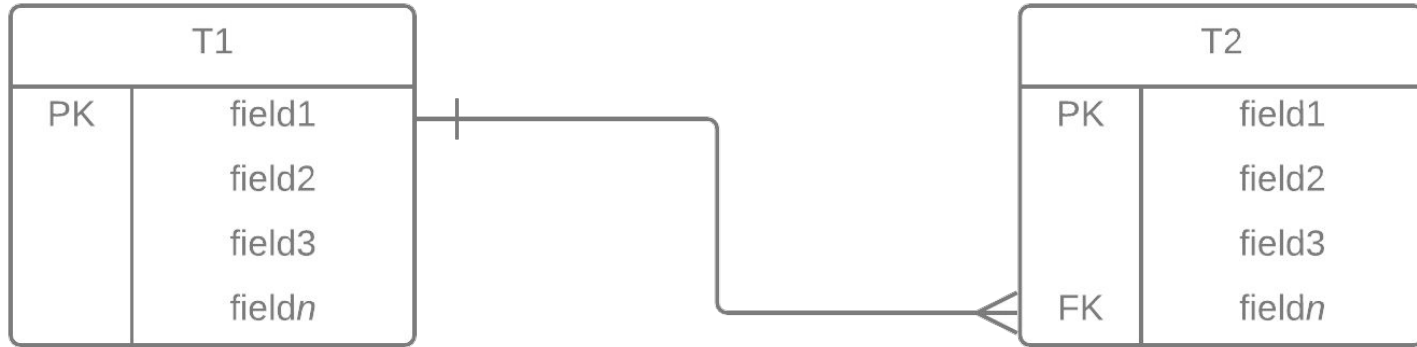
# CS 327E Intro

September 9, 2019

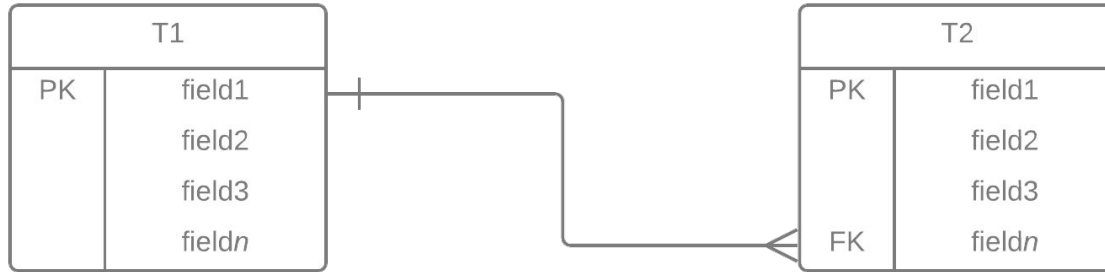
# Basic Terminology

- Relation / Entity Type / Table
- Field / Attribute / Column
- Record / Entity / Tuple / Row
- Cell / Value
- Datatype (e.g. INT, NUMERIC, STRING, BOOL, DATE, TIMESTAMP)
- Constraint (e.g. NOT NULL, Primary Key, Foreign Key)
- Schema
- Database versus Dataset

# Table Relationship: One-to-Many (1:m)



# Table Relationship: One-to-Many (1:m)



Author

<u>id</u>	name	section
1	Mary Tuma	news
2	Michael King	arts
3	Nina Hernandez	news
4	Sunil Kumar	music

Article

<u>id</u>	title	date	authid
1	Turmoil at the Zoo	2019-01-26	1
2	CodeNEXT's New Friend	2019-01-27	1
3	Quote of the Week	2019-01-27	3
4	SXSW News	2019-01-28	2
5	More from Steve Adler	2019-01-28	1

# Table Relationship: One-to-One (1:1)



# Table Relationship: One-to-One (1:1)



Article

<u>id</u>	title	date	authid
1	Turmoil at the Zoo	2019-01-26	1
2	CodeNEXT's New Friend	2019-01-27	1
3	Quote of the Week	2019-01-27	3
4	SXSW News	2019-01-28	2
5	More from Steve Adler	2019-01-28	1

Article\_Stats

<u>id</u>	clicks	likes	dislikes	comments
1	120	45	9	13
2	0	0	0	0
3	8	0	0	2
4	30	4	0	1
5	9	1	3	3

# Table Relationship: Many-to-Many ( $m:n$ )



# Table Relationship: Many-to-Many (m:n)



Tag

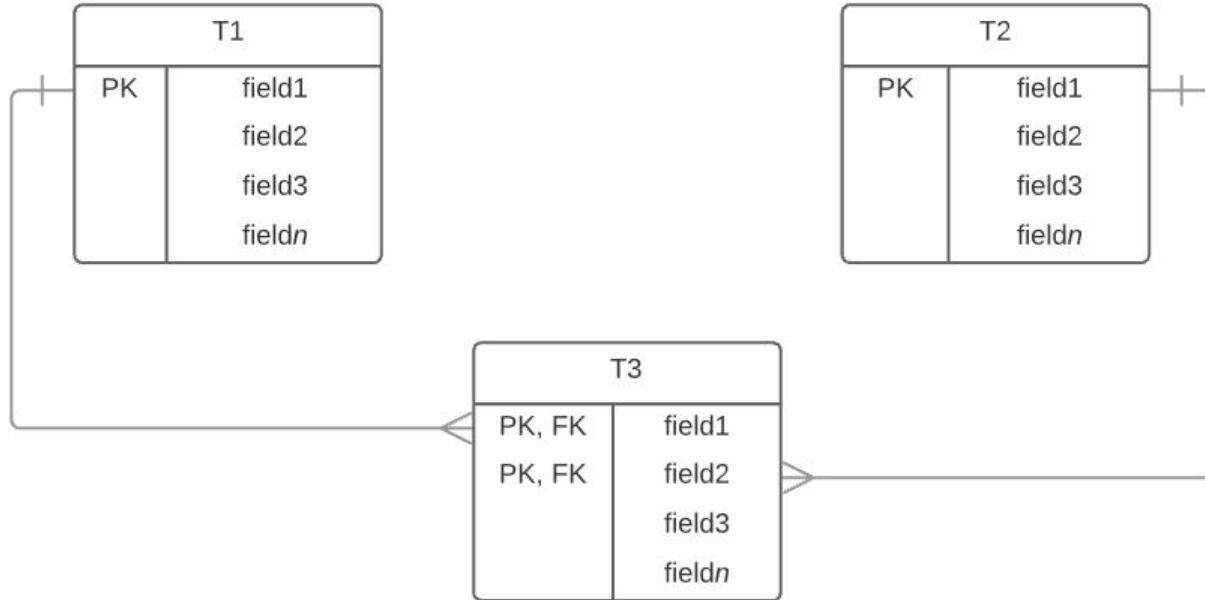
<u>id</u>	tag	<i>aids</i>
1	Politics	1, 2, 5
2	Austin	1, 2, 3, 4, 5
3	Mayor	3, 5
4	Business	1, 2, 5
6	Land Development	2
37	Animals	1

Article

<u>id</u>	title	date	<i>authid</i>	<i>tids</i>
1	Turmoil at the Zoo	2019-01-26	1	4, 37, 2
2	CodeNEXT's New Friend	2019-01-27	1	2, 6
3	Quote of the Week	2019-01-27	3	2, 3
4	SXSW News	2019-01-28	2	2, 40, 7
5	More from Steve Adler	2019-01-28	1	2, 3, 4



# Representation of Many-to-Many ( $m:n$ )



# Table Relationship: Many-to-Many ( $m:n$ )

Tag

<u>id</u>	tag
1	Politics
2	Austin
3	Mayor
4	Business
6	Land Development
37	Animals

Article

<u>id</u>	title	date	<i>authid</i>
1	Turmoil at the Zoo	2019-01-26	1
2	CodeNEXT's New Friend	2019-01-27	1
3	Quote of the Week	2019-01-27	3
4	SXSW News	2019-01-28	2
5	More from Steve Adler	2019-01-28	1

Tagged\_Article

<u>tid</u>	<u>aid</u>
4	1
37	1
2	1
2	2
6	2

# iClicker Setup Test

# iClicker Question

Have you ever spent many hours on a piece of software, made lots of mistakes, and eventually fixed things with some insight?

- A. Yes, I know this feeling!
- B. Maybe, but it's been a while since I've written code
- C. No, my code usually runs without too much work

# Term Project

- Create a data warehouse
- 12 Milestones
- [Milestone 1](#) due this Friday

# Datasets

- A set of related data files produced from the same source
- Dataset types: main dataset and secondary dataset
- Choose data you want to analyze to gain some insights on a particular topic or domain

# Main Dataset

- aka *Dataset1*
- Comprised of  $N$  files ( $3 < N < 30$ )
- Tabular format (e.g. CSV)
- At least 10K rows per file
- At least 5 columns per file
- Multiple parent/child relationships present in the data
- Uncleansed data

# Secondary Dataset

- aka *Dataset2*
- Related to your main dataset (aka *Dataset1*)
- Comprised of  $N$  files ( $3 < N < 30$ )
- At least 10K rows per file
- At least 5 columns per file
- Tabular format (e.g. CSV)
- Uncleansed data



# A Few Examples

Topic	Main Dataset	Secondary Dataset
<b>Transportation</b>	<a href="#">Airline on-time performance</a> (source: Bureau of Transportation Statistics)	<a href="#">Storm events</a> (source: NOAA)
<b>Housing</b>	<a href="#">Short-term rentals in various cities</a> (source: Airbnb)	<a href="#">Long-term rentals nationwide</a> (source: Zillow)
<b>Employment</b>	<a href="#">H1B visa applications</a> (source: US Citizenship and Immigration Services)	Corporate Registrations (source: Secretary of State, accessible via Enigma Public) <a href="#">Occupational Employment Survey</a> (source: Bureau of Labor Statistics)
<b>Movies</b>	<a href="#">Hollywood Movies, Directors, Actors</a> (source: IMDB)	<a href="#">Bollywood Movies, Actors and Songs</a> (source: Kaggle)
<b>Music</b>	<a href="#">Artists and Songs</a> (source: MusicBrainz)	<a href="#">Artists, Labels, Recordings</a> (source: Discog)

## Main Dataset: H1B Visa applications

### Source:

US Dept. of Labor

### Tables:

2015 table: 241 MB, 618,804 rows

2016 table: 233 MB, 647,852 rows

2017 table: 253 MB, 624,650 rows

2018 table: 283 MB, 654,162 rows

### Schemas:

-A few schema variations between the tables (column names, data types).

### Project Work:

-Imported files into staging tables

-Milestone 2

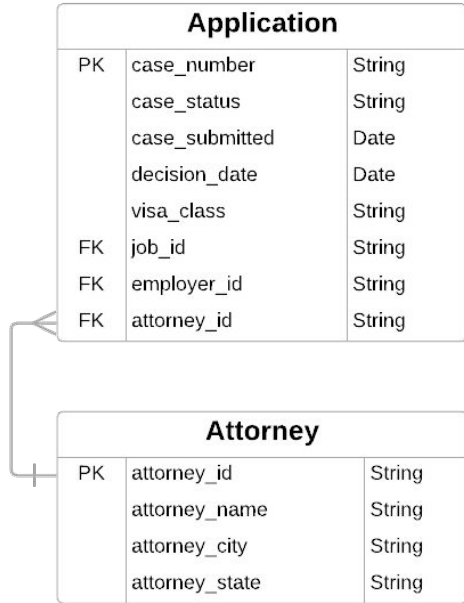
## Table Details: H1B\_Applications\_2017

Schema Details Preview

case_number	STRING	NULLABLE
visa_class	STRING	NULLABLE
case_status	STRING	NULLABLE
employer_name	STRING	NULLABLE
employer_business_dba	STRING	NULLABLE
employer_address	STRING	NULLABLE
employer_city	STRING	NULLABLE
employer_state	STRING	NULLABLE
employer_postal_code	STRING	NULLABLE
employer_country	STRING	NULLABLE
employer_province	STRING	NULLABLE
employer_phone	STRING	NULLABLE
employer_phone_ext	STRING	NULLABLE
naics_code	STRING	NULLABLE
soc_name	STRING	NULLABLE
soc_code	STRING	NULLABLE
job_title	STRING	NULLABLE
total_workers	INTEGER	NULLABLE
case_submitted	TIMESTAMP	NULLABLE
decision_date	TIMESTAMP	NULLABLE

employment_start_date	TIMESTAMP	NULLABLE
employment_end_date	TIMESTAMP	NULLABLE
full_time_position	BOOLEAN	NULLABLE
prevailing_wage	FLOAT	NULLABLE
pw_unit_of_pay	STRING	NULLABLE
wage_rate_of_pay_from	FLOAT	NULLABLE
wage_rate_of_pay_to	FLOAT	NULLABLE
wage_unit_of_pay	STRING	NULLABLE
worksite_city	STRING	NULLABLE
worksite_county	STRING	NULLABLE
worksite_state	STRING	NULLABLE
worksite_postal_code	STRING	NULLABLE
agent_attorney_name	STRING	NULLABLE
agent_representing_employer	BOOLEAN	NULLABLE
agent_attorney_city	STRING	NULLABLE
agent_attorney_state	STRING	NULLABLE
h1b_dependent	BOOLEAN	NULLABLE
willful_violator	BOOLEAN	NULLABLE
original_cert_date	TIMESTAMP	NULLABLE
new_employment	FLOAT	NULLABLE
continued_employment	FLOAT	NULLABLE
change_previous_employment	FLOAT	NULLABLE
new_concurrent_employment	FLOAT	NULLABLE

# H1B Normalized Database



Job		
PK	job_id	String
FK	employer_id	String
	employment_start_date	Date
	employment_end_date	Date
	job_title	String
	wage_rate_of_pay_from	Float
	wage_rate_of_pay_to	Float
	wage_unit_of_pay	String
	worksite_city	String
	worksite_county	String
	worksite_state	String
	worksite_postal_code	String
	soc_code	String
	soc_name	String
	total_workers	Integer
	full_time_position	Boolean
	prevailing_wage	Float
	pw_unit_of_pay	String
	pw_wage_level	String
	pw_source	String
	pw_source_year	Integer
	pw_source_other	String

Table Sizes (as rows):		
	v1	v2
<b>Employer</b>	348,876	161,759
<b>Job</b>	2,230,779	2,230,625
<b>Application</b>	2,633,426	2,633,156
<b>Attorney</b>	19,861	N/A

## Project Work:

- Merged and split raw tables
- Enforced referential integrity
- Removed duplicate records
- Milestones 4, 5, 6

## Secondary Dataset: Corporate Registrations

### Source:

Secretary of State from 13 states

### Tables:

AZ: 225 MB, 869,943 rows

CA: 1.1 GB, 3,792,457 rows

CO: 38 MB, 160,808 rows

CT: 192 MB, 796,877 rows

GA: 302 MB, 2,076,016 rows;  
116 MB, 2,063,919 rows

MA: 221 MB, 1,066,639 rows

MN: 374 MB, 1,688,714 rows;  
799 MB, 4,072,355 rows

MO: 133 MB, 2,364,476 rows;  
519 MB, 2,115,151 rows

NC: 262 MB, 1,389,877 rows

OH: 497 MB, 2,408,556 rows

NY: 512 MB, 2,587,015 rows

VA: 111 MB, 334,008 rows

WA: 205 MB, 1,152,309 rows

## Table Details: Corporate\_Registrations\_CA

Schema	Details	Preview
--------	---------	---------

so_file_number	STRING
corporation_number	INTEGER
corporation_status	STRING
corporation_classification	STRING
corporation_name	STRING
care_of_name	STRING
mail_address_line_1	STRING
mail_address_line_2	STRING
mail_address_city	STRING
mail_address_state_or_country	STRING
mail_address_zip_code	STRING
corporation_type	STRING
incorporation_date	DATE
so_file_date	DATE
term_expiration_date	DATE
chief_executive_officer_name	STRING

chief_executive_officer_address_line_1	STRING
chief_executive_officer_address_line_2	STRING
chief_executive_officer_address_city	STRING
chief_executive_officer_address_state_or_county	STRING
chief_executive_officer_address_zip_code	STRING
agent_name	STRING
agent_address_line_1	STRING
agent_address_line_2	STRING
agent_address_city	STRING
agent_address_state_or_county	STRING
agent_address_zip_code	STRING
state_or_foreign_country	STRING
ftb_suspension_status	STRING
corporation_tax_base	STRING
transaction_julian_date	DATE
ftb_suspension_string	STRING
filler	STRING

## Secondary Dataset: Occupational Employment Survey

Source: Bureau of Labor Statistics

### Wages Tables:

2015: 29.2 MB, 473,717 rows

2016: 29.9 MB, 484,390 rows

2017: 29.9 MB, 484,390 rows

2018: 29.9 MB, 485,211 rows

### Geography Table Sizes:

2015: 340 KB, 4,765 rows

2016: 357 KB, 4,991 rows

2017: 357 KB, 4,991 rows

2018: 357 KB, 4,991 rows

### Project Work:

-Imported files into staging tables

-Milestone 9

## Table Details: All\_Industries\_Wages\_2018

Schema	Details	Preview
--------	---------	---------

Row	Area	SocCode	GeoLvl	Level1	Level2	Level3	Level4	Average
485200	5100003	27-1022	4	18.57	28.24	37.92	47.59	37.92
485201	5100004	27-1022	4	18.57	28.24	37.92	47.59	37.92
485202	5400001	27-1022	4	18.57	28.24	37.92	47.59	37.92
485203	5400002	27-1022	4	18.57	28.24	37.92	47.59	37.92
485204	6600001	27-1022	4	18.57	28.24	37.92	47.59	37.92
485205	73050	27-1022	4	18.57	28.24	37.92	47.59	37.92
485206	74950	27-1022	4	18.57	28.24	37.92	47.59	37.92

## Table Details: Geography\_2018

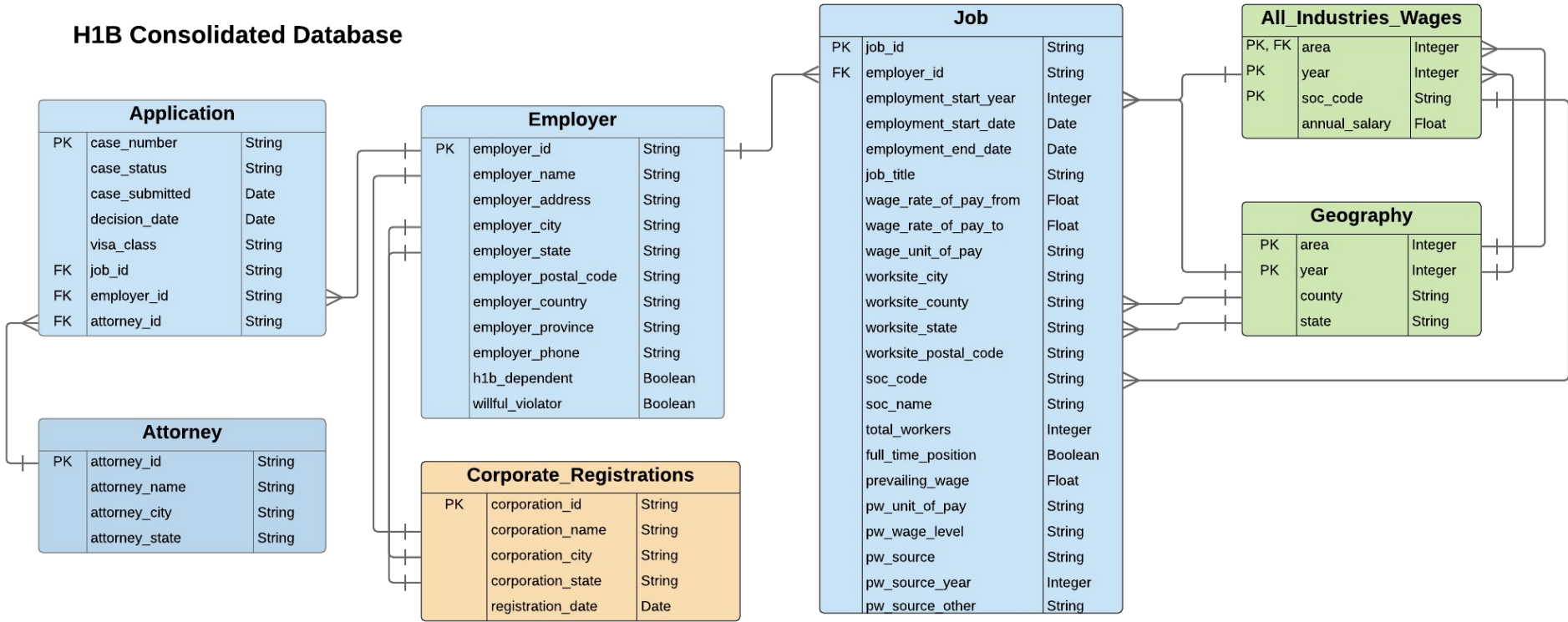
Refresh

Query Table

Schema	Details	Preview
--------	---------	---------

Row	Area	AreaName	StateAb	State	CountyTownName
4416	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (STOUGHTON)
4417	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (FRANKLIN)
4418	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (MEDWAY)
4419	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (NORWOOD)
4420	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (CANTON)
4421	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (DEDHAM)
4422	71654	Boston-Cambridge-Newton, MA NECTA Division	MA	MASSACHUSETTS	NORFOLK (DOVER)

# H1B Consolidated Database



## Project Work:

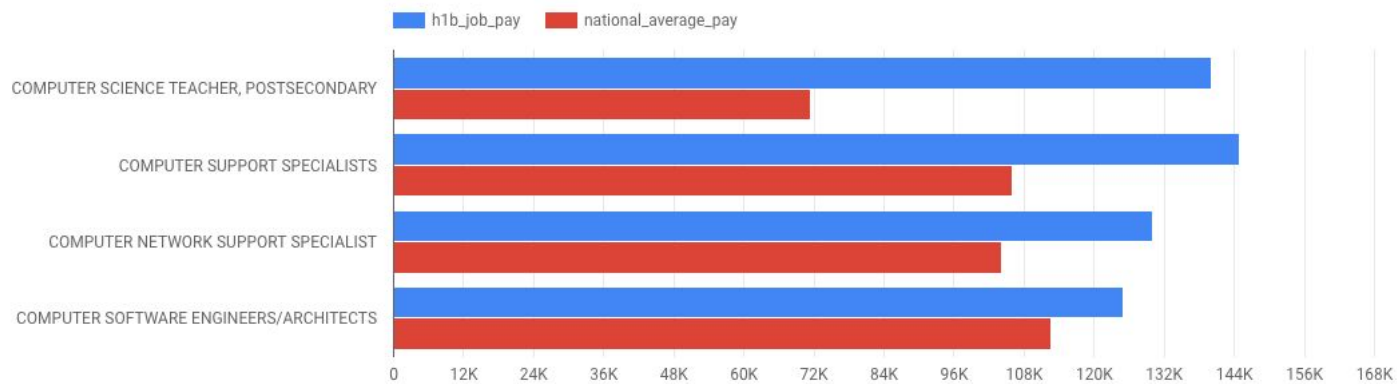
- Merged corp. registration tables
- Merged wages tables
- Merged geography tables
- Normalized corporation name, city, state

Milestones 9 and 10

# Sample Reports

## Pay Gaps by Occupation:

Occupations which pay H1B workers *higher* than domestic workers



Occupations which pay H1B workers *lower* than domestic workers

