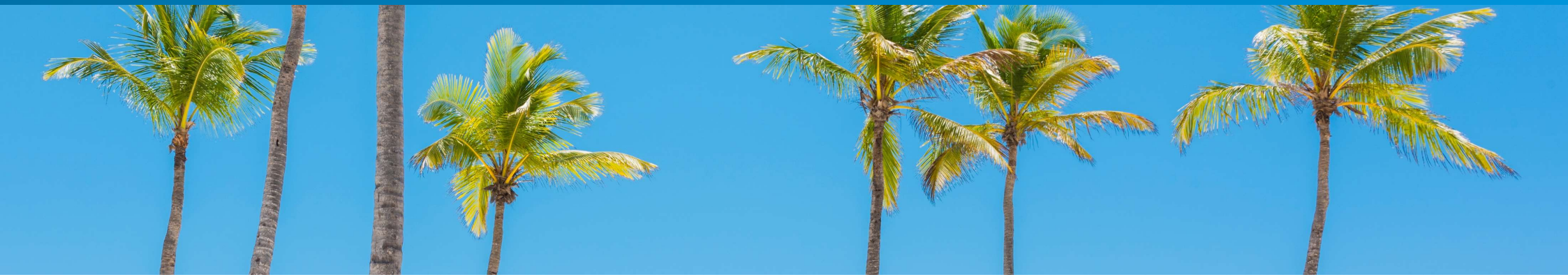
summer scholars

Welcome

summer scholars

Holistic View of Data to Drive Business Decisions

David Staas



David Staas

Current Role:	Chief Architect, Software and Data Platforms
Where I Live:	Camas, WA, USA (near Portland, OR)
Fun Fact:	Avid 3D printer and parts designer (dstaas on Thingiverse)
Strangest Job:	Railroad tanker cleaner for large ketchup company
Last Book Read:	Do long threads on Reddit count?
Favorite Podcast:	Opening Arguments

From Data to Insights

How do big companies like HP turn data into critical business insights?

- How can I start with a set of business questions and systematically answer them with data?
- What are the various steps along the way?

- What are some common concerns and issues?
  - Data quality, consent, design for reuse, understanding the data in context

From Data to Insights

An Example

- For this session, we'll use data to find some specific insights into business questions we have:
  - Who are my top salespeople in the US?
  - Are there any regions in the US where the Covid-19 case rate is rising week to week?
  - Are any of those in my top 100 sales regions?
  - Who are the affected salespeople so I can notify them?

From Data to Insights

An Example

- This will require a mix of data:
  - Sales and salesperson data from our enterprise data systems (we will fake this and generate some)
  - Reference data (US counties/states, populations)
  - Covid-19 case data by county
- We will put this data together to show how we can use it to answer the business questions we have

Example Python Code for Each Step

Jupyter notebook with extensive Python code and SQL queries

- Github repo provided at the end of session

Step 1: Import public reference data for US counties.

We'll need this data to match up FIPS codes (which some of the Covid data uses) to states, which is how our salespeople are assigned. FIPS codes are a 5-digit number that identifies a county within a state, or area within a territory.

I've already grabbed some USDA data.

I already have Postgres installed and running locally, so let's create a table and insert the CSV data from the USDA.

I have saved the data into `data/usda_county_pop_2019.csv`.

The CSV file looks like this:

```
FIPStxt,State,Area_Name,POP_ESTIMATE_2019
01000,AL,Alabama,4903185
...
```

fips

fipstxt	varchar(5)
state	varchar(5)
area_name	varchar(100)
pop_estimate_2019	integer

	fipstxt	state	area_name	pop_estimate_2019
1	01000	AL	Alabama	4903185
2	01001	AL	Autauga County	55869
3	01003	AL	Baldwin County	223234
4	01005	AL	Barbour County	24686
5	01007	AL	Bibb County	22394
6	01009	AL	Blount County	57826
7	01011	AL	Bullock County	10101
8	01013	AL	Butler County	19448
9	01015	AL	Calhoun County	113605
10	01017	AL	Chambers County	33254
11	01019	AL	Cherokee County	26196

We can use the Postgres COPY command to import it directly.

3.2 Insert salesperson names for all states

In [4]:

```
import pandas
from faker import Faker
import psycopg2.sql as sql

fake = Faker()
connection = my_connect()
cursor = connection.cursor()

# Get the list of states from the zip_city table we created earlier
q = "SELECT DISTINCT state FROM fips ORDER BY state ASC"
df = pandas.io.sql.read_sql_query(q, connection)
states = df['state'].values.tolist()

# Generate a fake salesperson name for each state and insert it if there isn't already one for that state
for state in states:
    name = fake.name()
    q = sql.SQL("INSERT INTO salesperson (name, state) VALUES ({}, {}) ON CONFLICT DO NOTHING;")
    cursor.execute(q.format(sql.Literal(name), sql.Literal(state)))
    connection.commit()

# Print 10 to verify
df = pandas.io.sql.read_sql_query("SELECT * FROM salesperson LIMIT 10", connection)
print(df.head(10))
```

```
  state              name
0    AK  Kimberly Frederick
1    AL       Jaime Mccarty
2    AR        James Thomas
3    AZ        Julie Butler
4    CA       Larry Morales
5    CO       Alexis Palmer
6    CT      Manuel Michael
7    DC         Tammy Braun
8    DE        Jason Nguyen
9    FL      Deborah Walker
```

Data quality concern: flaws in our salesperson table

There's a few things to note about the data set we just created that make it less than ideal for a real-world situation:

- Each state can have one and only one salesperson
- If a salesperson were to handle two states, you'd have to duplicate their name (not normalized)
- Our primary key constraint is on the 'state' column, which would not allow more than one record per state
- We only have the name of the salesperson, no details (e.g. e-mail address)

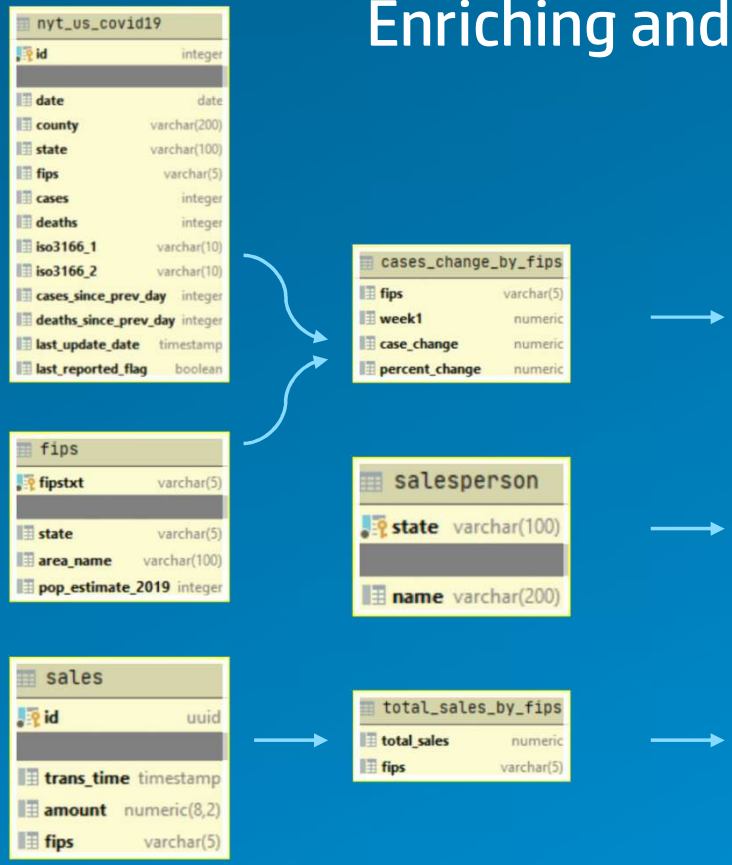
While the dataset is OK for the faked-up example we're doing here, it's important to think about how your data will be used by others. In this case, the dataset is pretty limiting, and it probably means that you'd have to do a bunch of rework later on if something changed (e.g. to allow a backup salesperson per state). Sometimes you can make a small change to the dataset at the beginning that makes it a lot easier. Examples in this case:

- Adding an integer ID field as the primary key instead, allowing multiple state records
- A separate table for salesperson details, including name, e-mail, phone, etc. with an integer ID as the primary key
- Instead of putting the salesperson name, use the salesperson detail ID to normalize this table

Doing a little work up front can make it much easier to share data across your organization. The changes above would make the data set more resilient to reasonable changes without having to do a bunch of schema changes that would break downstream apps, reports, or dashboards.

Example: From Data to Insights

Enriching and Joining Data for Consumption



Sales Regions: Covid-19 Weekly Case Increases of > 20%

	Total Sales	FIPS	County	State	Increase	Salesperson Name
0	184,061.67	27053	Hennepin County	MN	94%	Daniel Watson
1	210,983.71	12057	Hillsborough County	FL	22%	David Black
2	184,850.89	12103	Pinellas County	FL	36%	David Black
3	193,023.00	17043	DuPage County	IL	30%	Denise Mitchell
4	173,653.79	09003	Hartford County	CT	23%	Jason Gillespie
5	196,527.46	06029	Kern County	CA	52%	Jennifer York
6	175,367.11	06065	Riverside County	CA	53%	Jennifer York
7	202,761.07	06081	San Mateo County	CA	63%	Jennifer York
8	166,539.27	51059	Fairfax County	VA	23%	Jillian Barber
9	189,893.26	13089	DeKalb County	GA	81%	Johnny Summers
10	186,909.55	13121	Fulton County	GA	27%	Johnny Summers
11	207,610.58	13067	Cobb County	GA	58%	Johnny Summers
12	211,224.48	13135	Gwinnett County	GA	35%	Johnny Summers
13	216,434.65	25009	Essex County	MA	35%	Juan Lewis
14	204,717.24	25027	Worcester County	MA	25%	Juan Lewis
15	183,458.34	25017	Middlesex County	MA	42%	Juan Lewis
16	192,031.82	08031	Denver County	CO	31%	Michelle Blankenship
17	204,280.46	48141	El Paso County	TX	36%	Ryan Morgan
18	221,367.39	48439	Tarrant County	TX	63%	Ryan Morgan
19	170,908.30	48453	Travis County	TX	55%	Ryan Morgan
20	227,585.81	39049	Franklin County	OH	51%	Shawn Hudson
21	212,201.84	47157	Shelby County	TN	32%	Timothy Harris

Data Lifecycle

Policy & Ethics

Sourcing

Operating

Analyzing

Acting & Automating

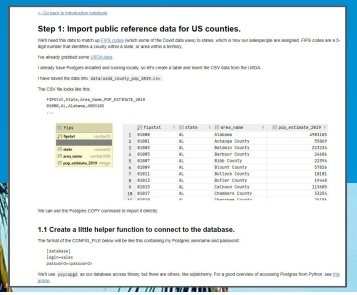
Policy and Ethics

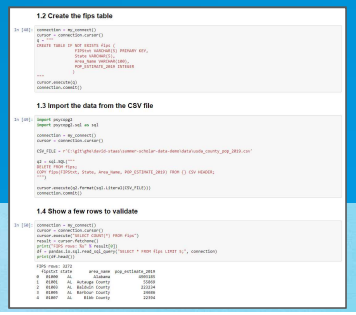
- Personal data is governed by a variety of privacy laws and standards
- Process personal data in accordance with law and with transparency and fairness to the customer
- Certain data such as financial data also has restrictions and legal requirements
- In-depth discussion is beyond the scope of this session, but it's critical to understand your company's requirements for privacy and governance on the data and adhere to them

Sourcing and Quality

- Variety of sources for data: business systems, partners, publicly-available reference data, manual data entry, etc.
- Quality checks and cleansing are critical (missing values, invalid types, data not matching the published schema, values not matching business rules)
- Subject matter experts (SMEs) define and document the data fields in a data catalog
  - Example: sales data. What currency? Conversion rules? Is the sales date in UTC or local time?

Demo: Notebooks #1 and #2





Operations, Technology, and Platform

- Many ways to ingest data: streaming, batch, manual file transfer, partner data interchange, …
- Scheduling, execution, and monitoring should be automated
- Technologies, both on-premise and in the cloud, require specialized expertise to run and manage
  - "DataOps": DB sizing, performance, query optimization, maintenance, automation
  - Many possibilities for technology choices in storage, processing, security, governance, …
- Data platform – warehouse (structured), lake (unstructured), lake house (mix), integration hub (governed sharing)
- Importance of standards for security, governance, technology stack

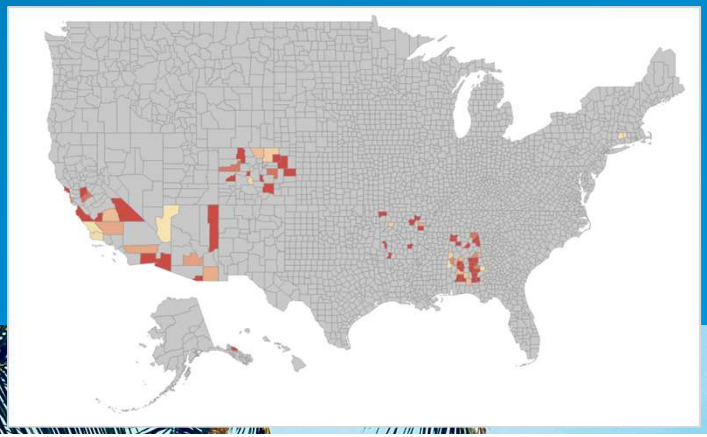
Analytics, Insights, and AI

- Visualization: charts, graphs, and dashboards
- Reporting
- Machine learning: classification, prediction, forecasting (sessions tomorrow)

Demo: Notebooks 3, 4, 5

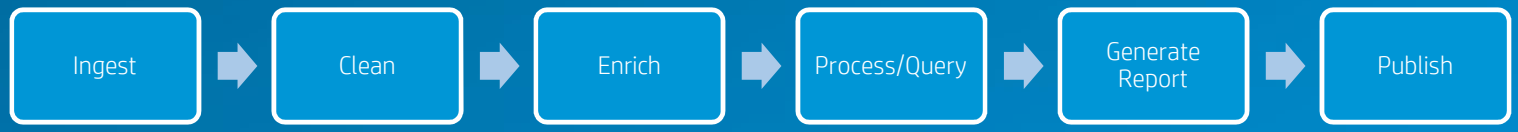
Sales Regions: Covid-19 Weekly Case Increases of > 20%

	Total Sales	FIPS	County	State	Increase	Salesperson Name
0	184,061.67	27053	Hennepin County	MN	94%	Daniel Watson
1	210,983.71	12057	Hillsborough County	FL	22%	David Black
2	184,850.89	12103	Pinellas County	FL	36%	David Black
3	193,023.00	17043	DuPage County	IL	30%	Denise Mitchell
4	173,653.79	09003	Hartford County	CT	23%	Jason Gillespie
5	196,527.46	06029	Kern County	CA	52%	Jennifer York
6	175,367.11	06065	Riverside County	CA	53%	Jennifer York
7	202,761.07	06081	San Mateo County	CA	63%	Jennifer York
8	166,539.27	51059	Fairfax County	VA	23%	Jillian Barber
9	189,893.26	13089	DeKalb County	GA	81%	Johnny Summers
10	186,909.55	13121	Fulton County	GA	27%	Johnny Summers
11	207,610.58	13067	Cobb County	GA	58%	Johnny Summers
12	211,224.48	13135	Gwinnett County	GA	35%	Johnny Summers
13	216,434.65	25009	Essex County	MA	35%	Juan Lewis
14	204,717.24	25027	Worcester County	MA	25%	Juan Lewis
15	183,458.34	25017	Middlesex County	MA	42%	Juan Lewis
16	192,031.82	08031	Denver County	CO	31%	Michelle Blankenship
17	204,280.46	48141	El Paso County	TX	36%	Ryan Morgan
18	221,367.39	48439	Tarrant County	TX	63%	Ryan Morgan
19	170,908.30	48453	Travis County	TX	55%	Ryan Morgan
20	227,585.81	39049	Franklin County	OH	51%	Shawn Hudson
21	212,201.84	47157	Shelby County	TN	32%	Timothy Harris



Acting & Automating

- In our Covid-19 sales regions example, we would likely want to run that report weekly and send an automated e-mail or publish to a dashboard
- It is common to build a processing pipeline to automate these steps



- Many technologies and ways to do this. Examples:
  - Apache Airflow for scheduling and workflow
  - Python, Hadoop, or Apache Spark/Databricks for processing
  - Qlik, PowerBI, Tableau, or Looker for visualization
  - AWS S3, Amazon Redshift, Azure DW, Google BigQuery, MySQL, PostgreSQL for storage

Summary

- The power of data enables businesses to gain insights into customers, sales, markets, and opportunities
- A holistic view of the data that includes policy and ethics, sourcing and quality, operating, analyzing and acting, and automating provides a complete framework
- The Jupyter notebook provides a hands-on example that takes specific business questions and uses data to get the answers and insights we need
  - You can download the code and run it yourself here:

https://github.com/HPInc/hp-summer-scholars-2020

summer scholars

Thank You

Related:

- Statistical Data Analysis & Lab – 1pm PDT today
- Intro to Reinforcement Learning – 11am tomorrow
- Bonus notebook #7: a better Covid-19 query, with mapping instructions