Leveraging 3rd – Party Data for Census Modernization

Statistics Modernization Branch

U.S. Census Bureau

Emily Wiley— Branch Chief, Statistics Modernization Branch

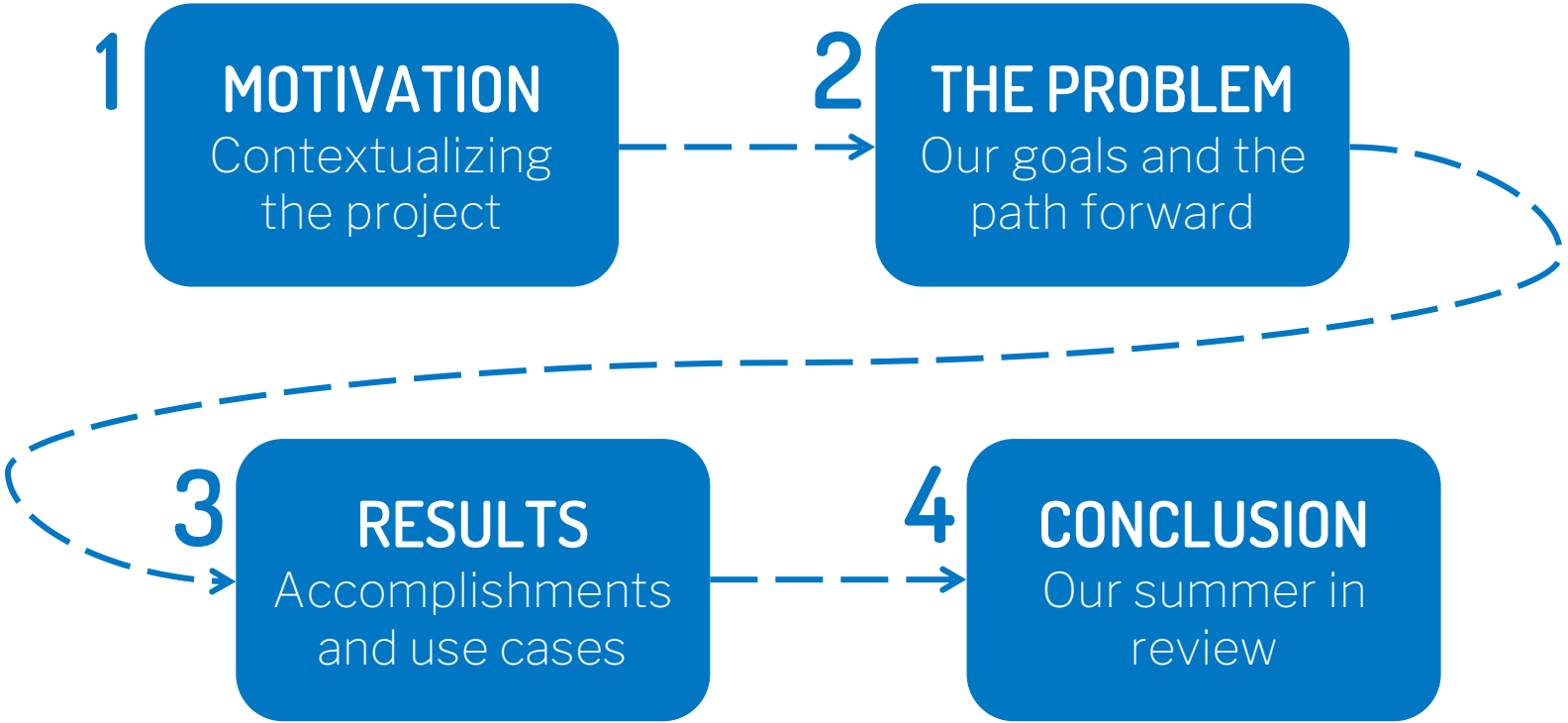
coding it forward >

United States™ Census Bureau

Ayush Kumar
University of Georgia
BS Data Science, BA Economics

Sarah Dzwil
University of Rhode Island
BS Data Science, BS Math

TABLE OF CONTENTS



cif>

MOTIVATION

WHY 3RD – PARTY?

"In order to reduce respondent burden and increase our ability to produce high-quality data products with nimbleness and agility, we will combine traditional methods with alternate data sources and the latest data science principles in a multipronged approach aimed towards meeting our customers' current and future demands."

US Census Bureau Strategic Plan (Fiscal Year 2022-2026)

cif>

WHY 3RD – PARTY?

- The Modern Data Environment – Rich data is produced in abundance through administrative records and the internet
- Analyst Time is Precious – Census Bureau analysts spend a lot of time reviewing respondent information and trying to contact respondents; much of this information can be cross-referenced with existing records
- Product Quality – Faster analysis with more information will lead to better quality data products for the public

cif>

THE PROBLEM

LEVERAGING PUBLIC DATA

- A common analyst workflow in the Economic Directorate involves pulling up financial information for a particular entity. This information is often hard to locate and understand
- Every nonprofit above $500,000 in gross receipts must file a 990 form with IRS detailing their revenues, expenses, grants, and investments. Over 550,000 filings annually
- All publicly traded corporations file their financials with the SEC every quarter using form 10K/10Q. Around 6,000 filings every quarter

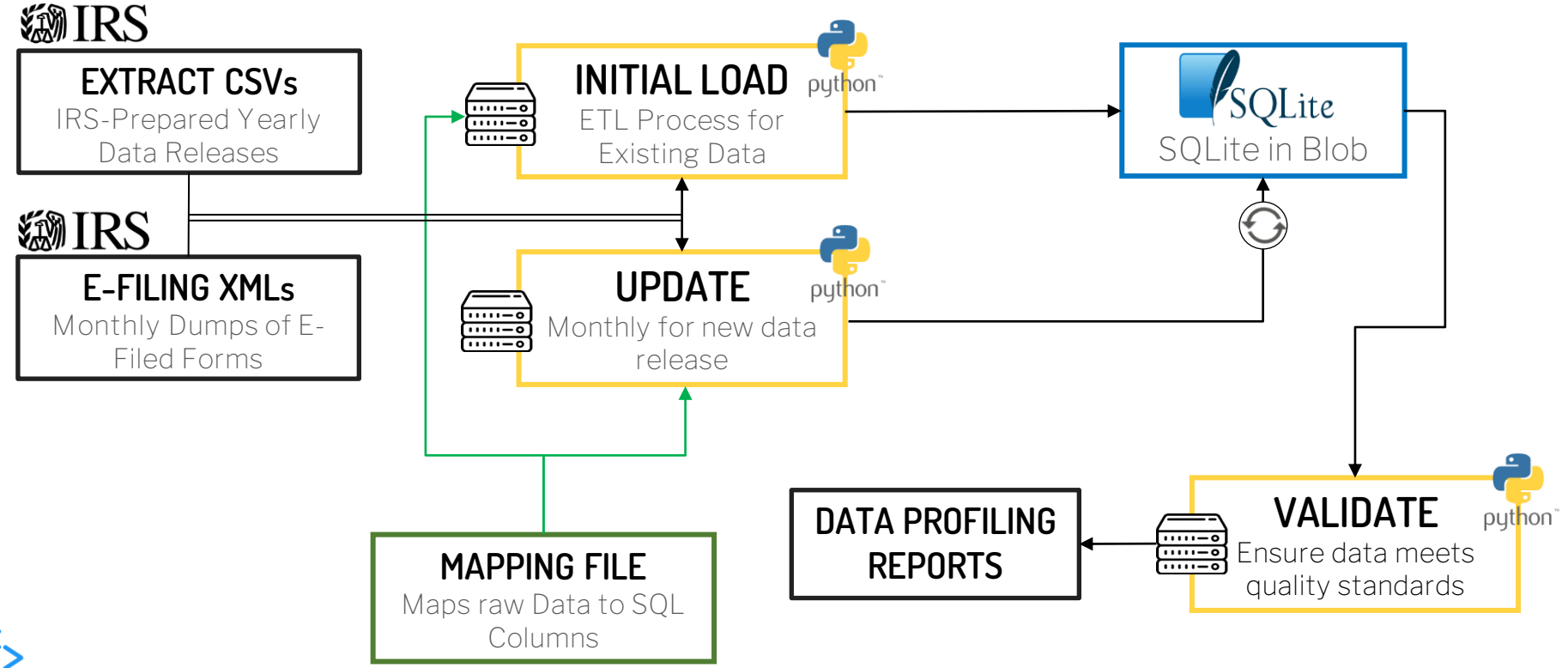
cif>

OUR GOAL

Build extract-transform-load (ETL) pipelines from publicly-available data on nonprofits (990 Series Tax Returns) and publicly-traded corporations (SEC) to accessible databases for analysts.

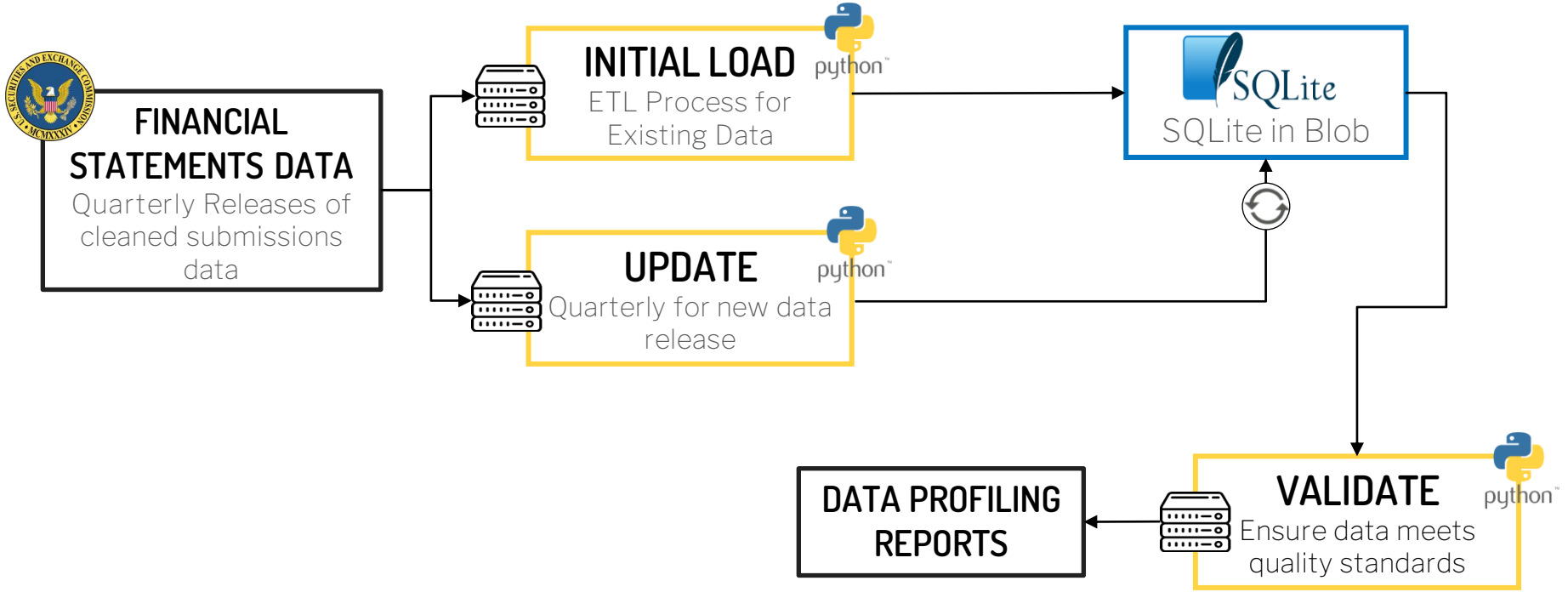
cif>

DATA PIPELINE (990)



cif>

DATA PIPELINE (SEC)



cif>

DATA CHALLENGES

- Complexity – 990 series forms can have over 1,500 fields; SEC submissions can have over 2,000. Firms can file for a single period multiple times and issue corrections.

- Scale – Millions of filings meant that loading data was a time and compute intensive process, bugs that resulted from scale took hours to debug

- Validation – Data accuracy and completeness is a matter of reputation

cif>

RESULTS

RESULTS

Nonprofits 990 Database

- Initial load of over 1,300,000 returns from fiscal years 2021 and 2022
- 9,000,000 financial facts covering 680,000 nonprofits
- Contact information for 585,000 nonprofits (86% Coverage)

SEC 10K/Q Database

- Initial load of 5,800 submissions from 2023Q1
- 273,000 financial facts from 5,600 publicly traded corporations

cif>

USE CASE 1: 990 FIRM-LEVEL ANALYSIS

Field*	Return Period 2020*	Return Period 2021*
Tax Period End	December 2020	December 2021
Tax Return Type	990EZ	990EZ
Total Contributions ($)	300,000	310,000
Total Revenue ($)	430,000	480,000
Grants ($)	125,000	180,000
Business Name	Do-Gooders Inc.	Do-Gooders Inc.
Phone Number	123-456-7890	123-456-7890
Address	100 Main Street	100 Main Street

*Examples are mock data

cif>

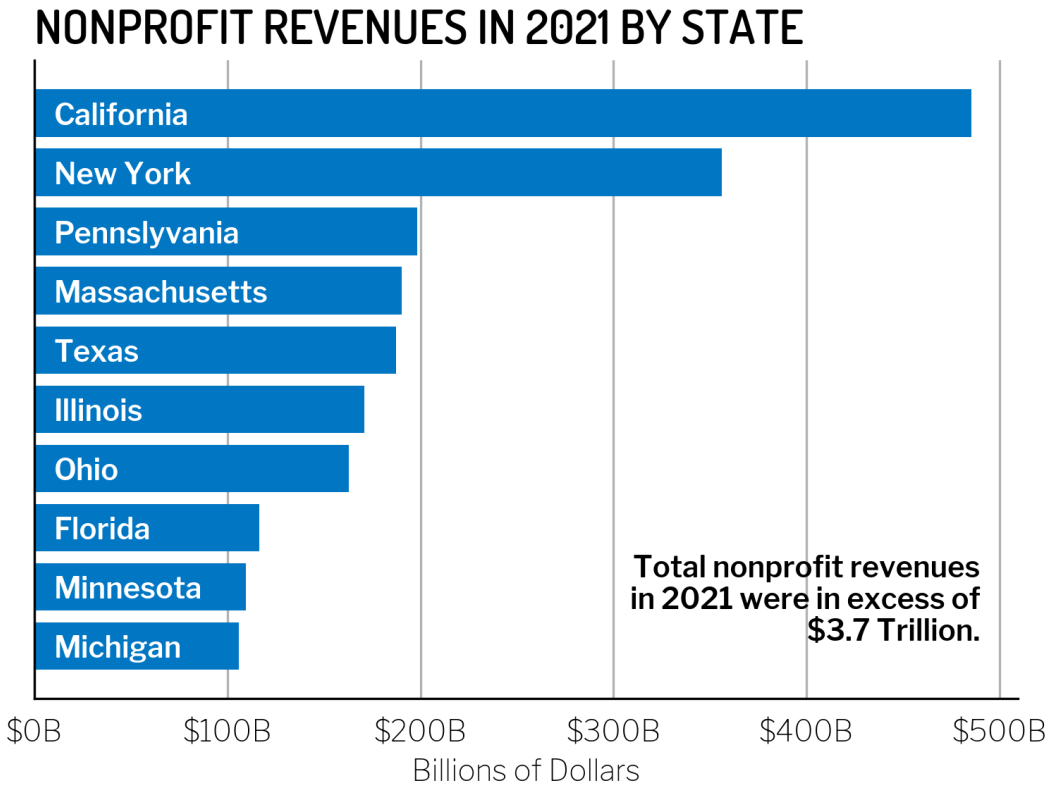
USE CASE 1: SEC FIRM-LEVEL ANALYSIS

Income Statement Element*	Value* ($ Millions)
Revenues	300,000
Research and development	35,000
Sales and marketing	26,000
General and administrative	10,000
Total costs	200,000
Operations Income	70,000
Other income (expense), net	(4,000)
Total	70,000
Provision for income taxes	10,000
Net income	60,000

*Examples are mock data

cif>

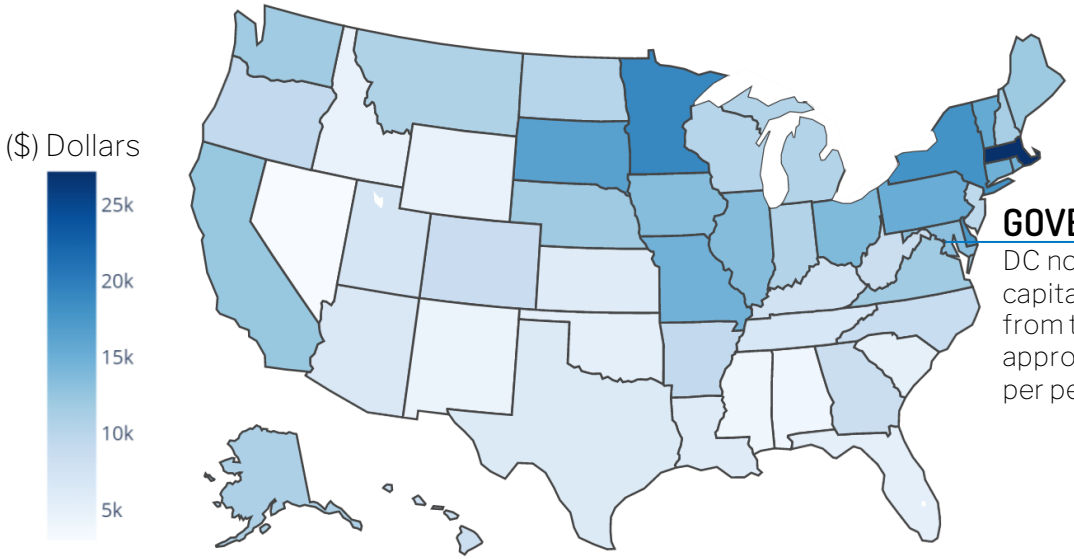
USE CASE 2: ANALYTICS (990)



USE CASE 2: ANALYTICS (990)

THE MOST GIVING-EST STATES

Nonprofit Revenue Per Capita (FY 2021)



GOVERNMENT CARES

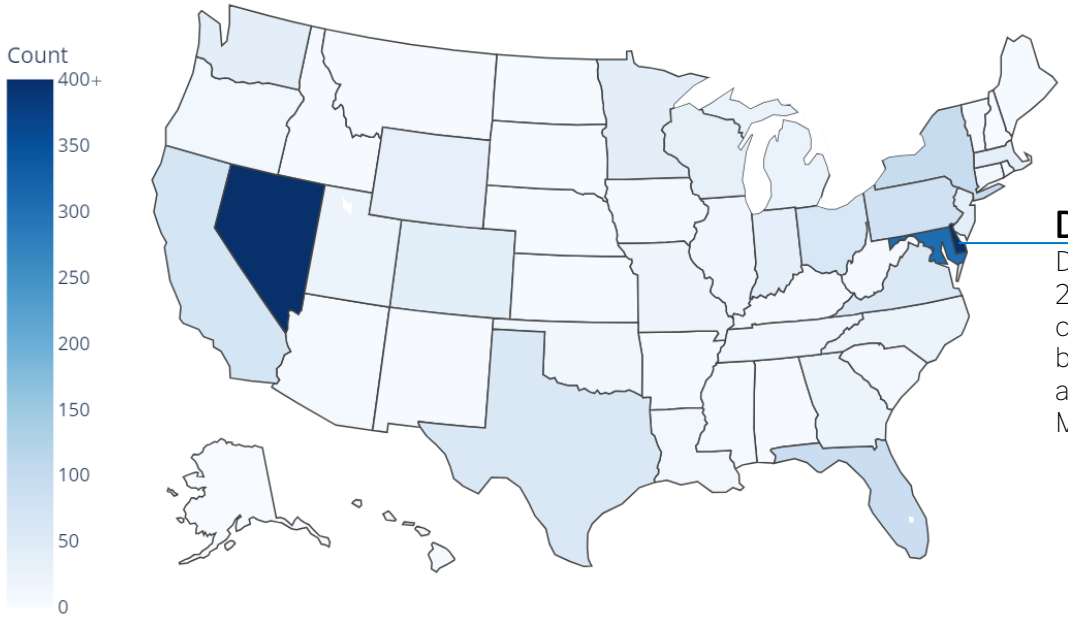
DC nonprofit revenue per capita had to be excluded from the graph because it approached over $100,000 per person.

cif>

USE CASE 2: ANALYTICS (SEC)

THE MOST POPULAR STATES FOR INCORPORATION

Count of State of Incorporation (Q1 2023)



DELAWARE IN THE LEAD

Delaware has a count of over 2,900, making it the most common state for incorporation by far. The next most popular are Nevada with about 450 and Maryland with about 300.

cif>

CONCLUSION

FELLOWSHIP IN REVIEW

New Skills

- Professional software development (git, code review)
- Collaborative setting (database design, code, project scope)

Lessons Learned

- Balancing the government approach to “doing things right” vs. the corporate approach of “getting things done”
- Stakeholder communication is critical to project success

cif>