A New Model for International, Privacy-Preserving Data Science

PEPR Conference, 6/03/2024

xD
https://www.xd.gov

Curtis Mitchell, Emerging Technology Fellow @ xD

United States™ Census Bureau

xD is an emerging technologies group that's advancing the delivery of data-driven services through new and transformative technologies.

xD
https://www.xd.gov

United States™ Census Bureau

Agenda

The US Census Bureau's Mission and Commitment to Privacy

How the Census Bureau Conducts Data Collaborations

Testing New Collaboration Methods with the UN PET Lab

Lessons Learned So Far & Future Goals

xD
https://www.xd.gov

United States™ Census Bureau

1. The Census Bureau's Mission and Commitment to Privacy

The Census Bureau's Mandate and Mission

—

The United States census is mandated under **Article I Section 2** of the United States Constitution

Census Bureau Mission

The Census Bureau's mission is to **serve as the nation's leading provider of quality data** about its people and economy.

We honor **privacy**, protect **confidentiality**, share our expertise globally, and conduct our work openly.

Privacy Laws at the Census Bureau

—

Title 13

Private information is **never published**

Personal information **cannot be used against respondents** by any government agency or court

Census Bureau employees are **sworn to protect confidentiality** and uphold Title 13.

Title 26

Defines how the IRS shares **tax data** with the Census Bureau, and how the Bureau **protects and restricts disclosure** of that data

xD
https://www.xd.gov

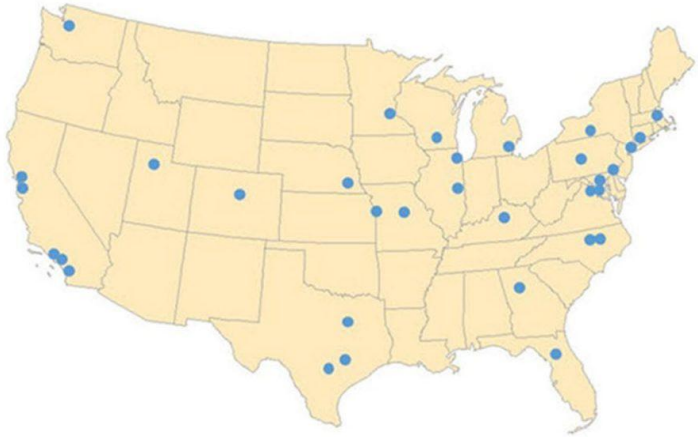
United States™ Census Bureau

2. How the Census Bureau Conducts Data Collaborations

Federal Statistical Research Data Centers (FSRDCs)

—

- FSRDCs provide **secure environments** supporting qualified researchers using **restricted-access data** while protecting respondent confidentiality
- Researchers must obtain **Special Sworn Status** (SSS) to protect confidential information
- 33 facilities around Continental US



Options for International Data Sharing with Census Bureau

—

- International data collaborations are handled on a **case-by-case** basis and usually require establishing a **data-sharing agreement**
- Data is often **aggregated** and already shared with international bodies such as the UN and OECD
- Data that crosses international borders is frequently **subject to export controls** that restrict sharing with other nations or foreign entities
- Monthly trade data can be obtained on the **USA Trade platform** maintained by the Census Bureau
- Different rules and protocols apply to other types of data (health, immigration, etc.)

xD
https://www.xd.gov

3. Testing New Collaboration methods with the UN PET Lab

United States™ Census Bureau

Private Linkage for Enhanced InternAtional Data ExchangeS (PLEIADES)

Objective: Demonstrate international, privacy-preserving data join

Data Phases: **1.** open trade data from UN Comtrade, **2.** synthetic data, **3.** private data

Partners:

UN PET Lab
United Nations
Privacy Enhancing Technologies Lab

Istat | Istituto Nazionale di Statistica

Statistics Canada | Statistique Canada

OpenMined

How OpenMined's PySyft Enables Secure Data Analysis

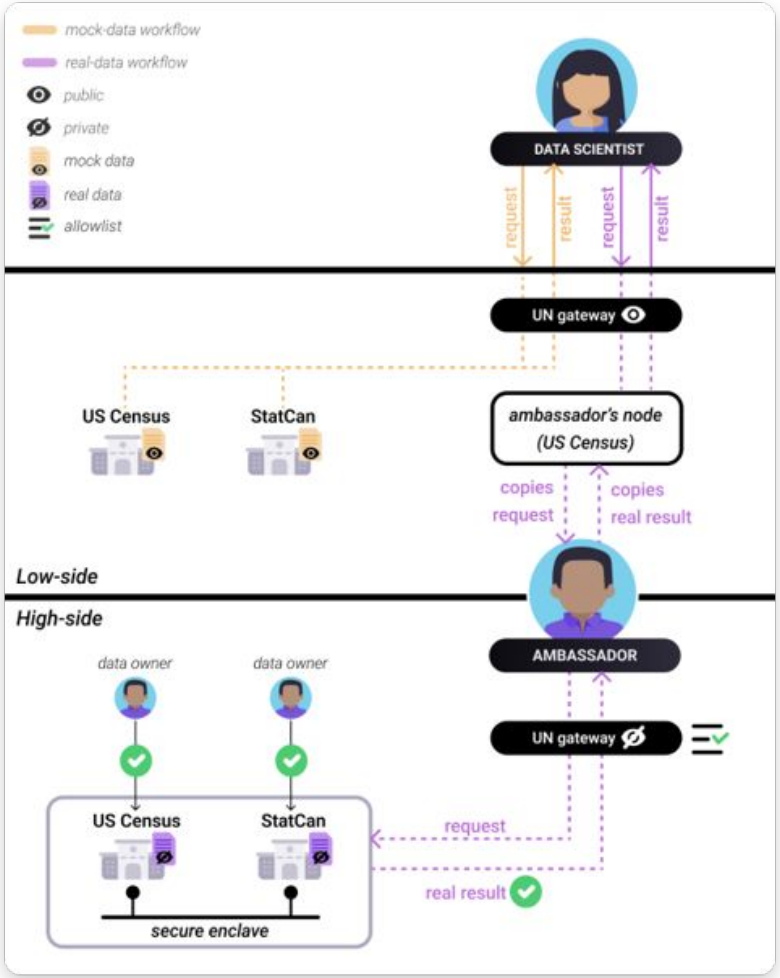
—

"Perform data science on data that **remains in someone else's server**"

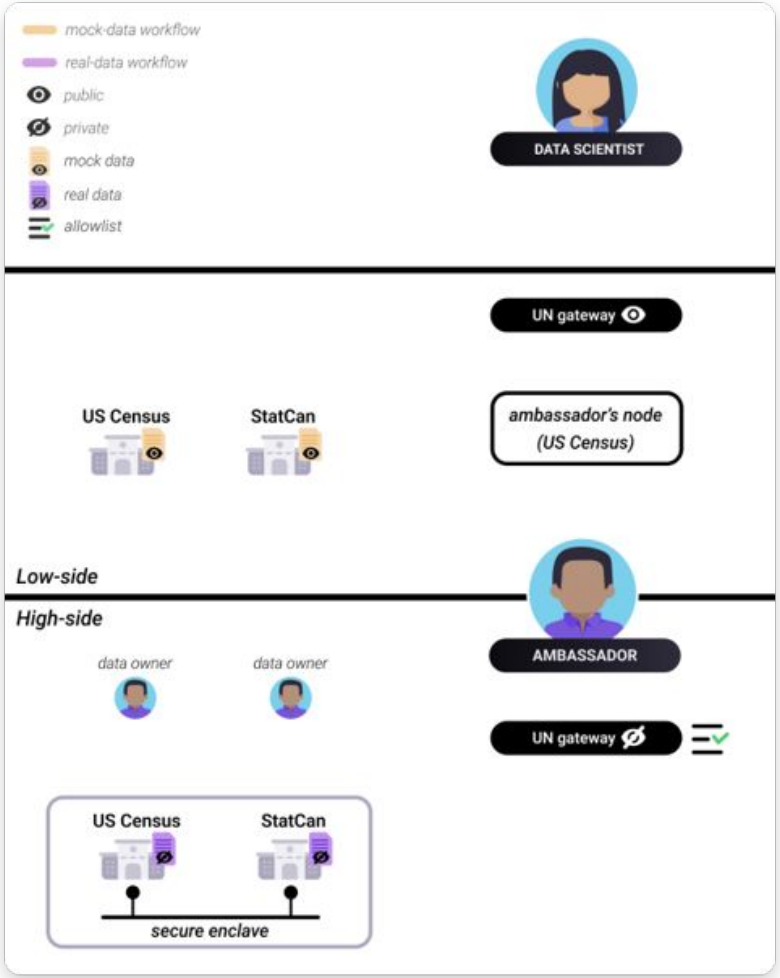
Key Terms:

- **Data Owner** - **Providers of datasets** to be accessed via secure computation
- **Data Scientist** - End users who are **submitting code to get results** from the data owner's datasets
- **Ambassador** – Optional **intermediary** between data scientists and data owners
- **Domain Server** - Enables, with safeguards, the data scientist to submit code and **the data owner to review and approve/deny requests** to run that code on private datasets
    - **Low-Side Domain** - **public** and discoverable
    - **High-Side Domain** - **private** and gated
- **Gateway Server** - **Acts as a bridge between domains, data scientists, and data owners** to enable analysis between multiple participants

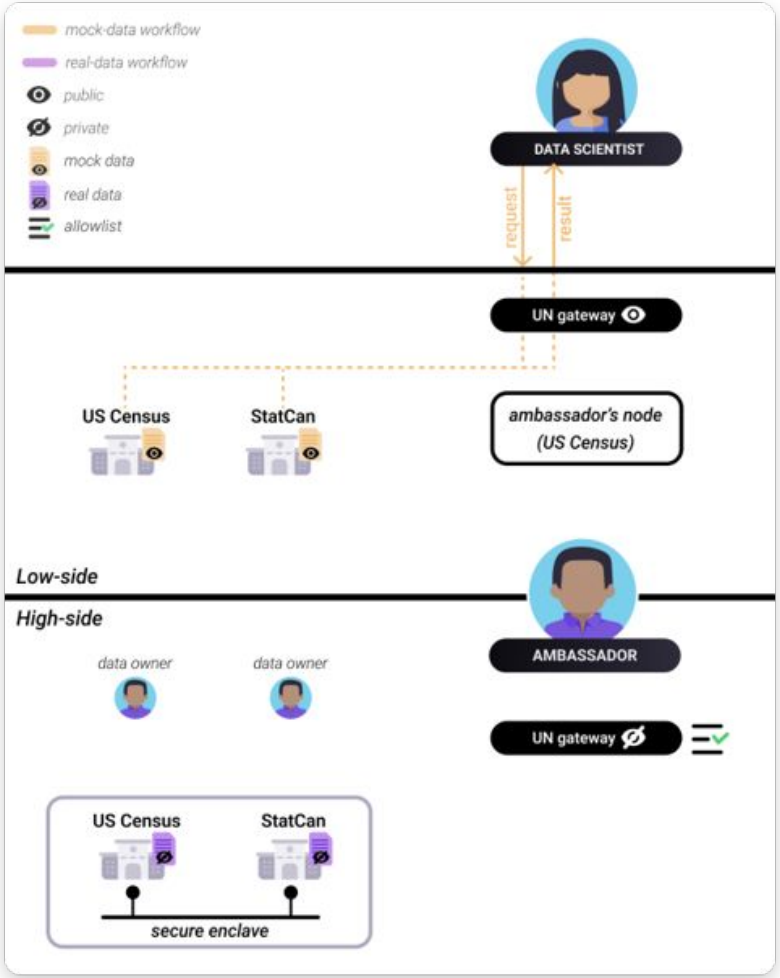
PySyft Data Flow



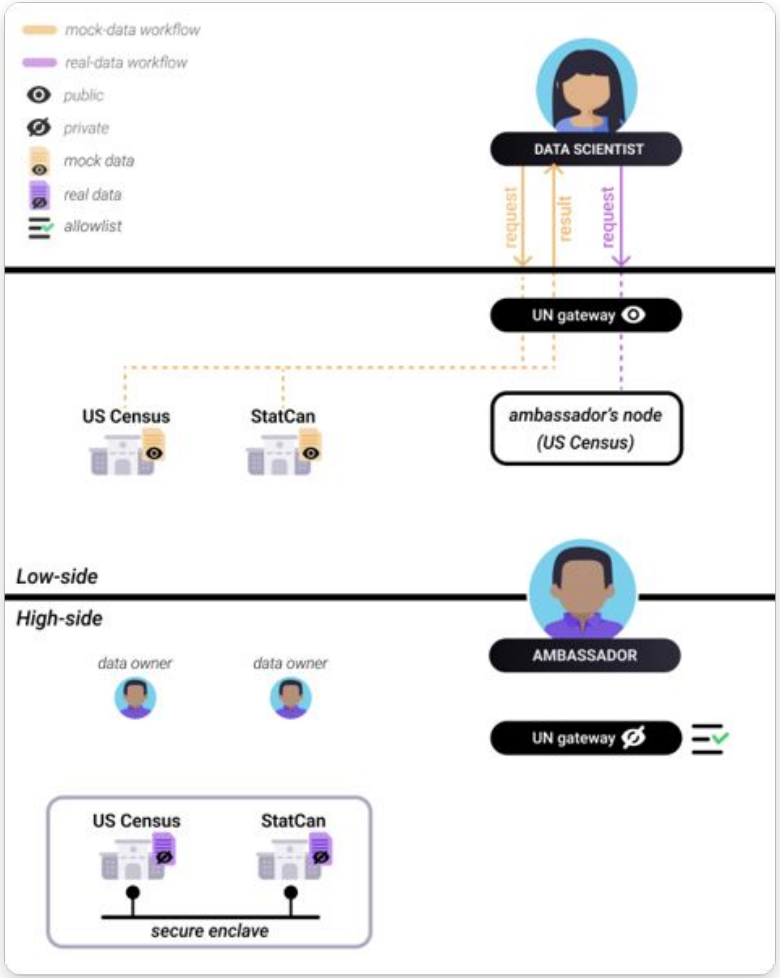
PySyft Data Flow



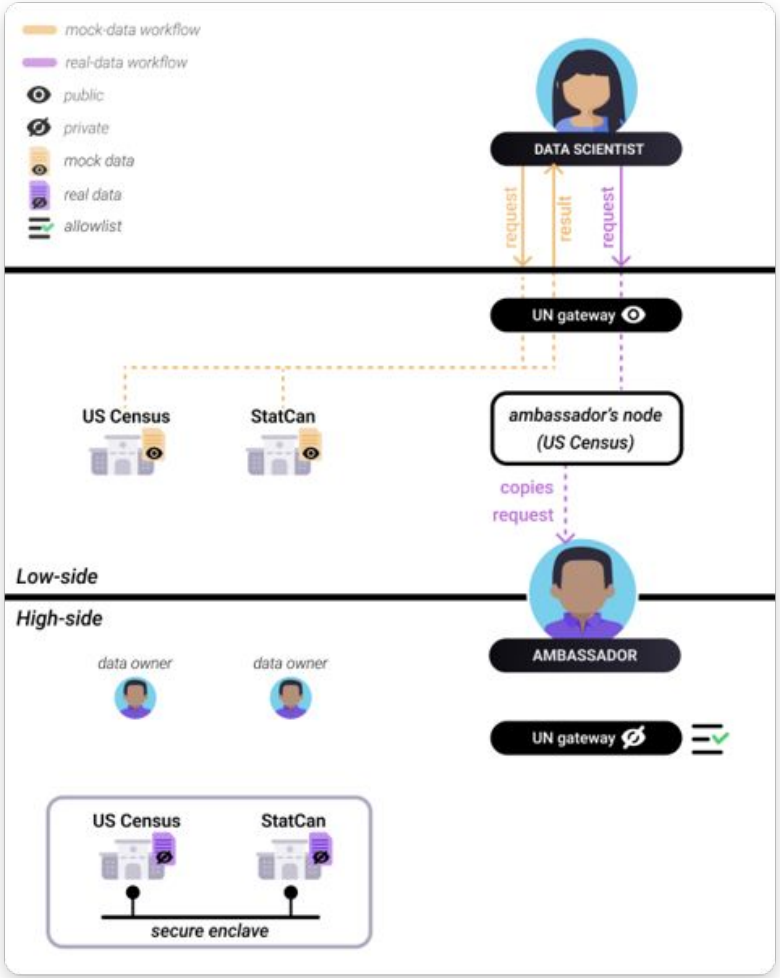
PySyft Data Flow



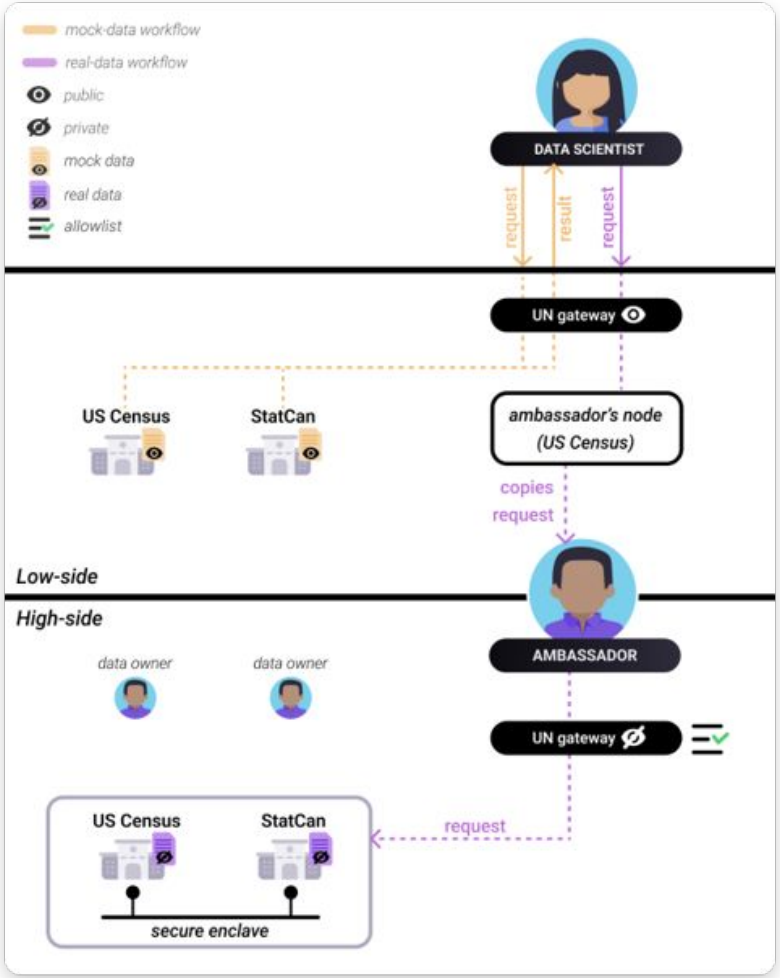
PySyft Data Flow



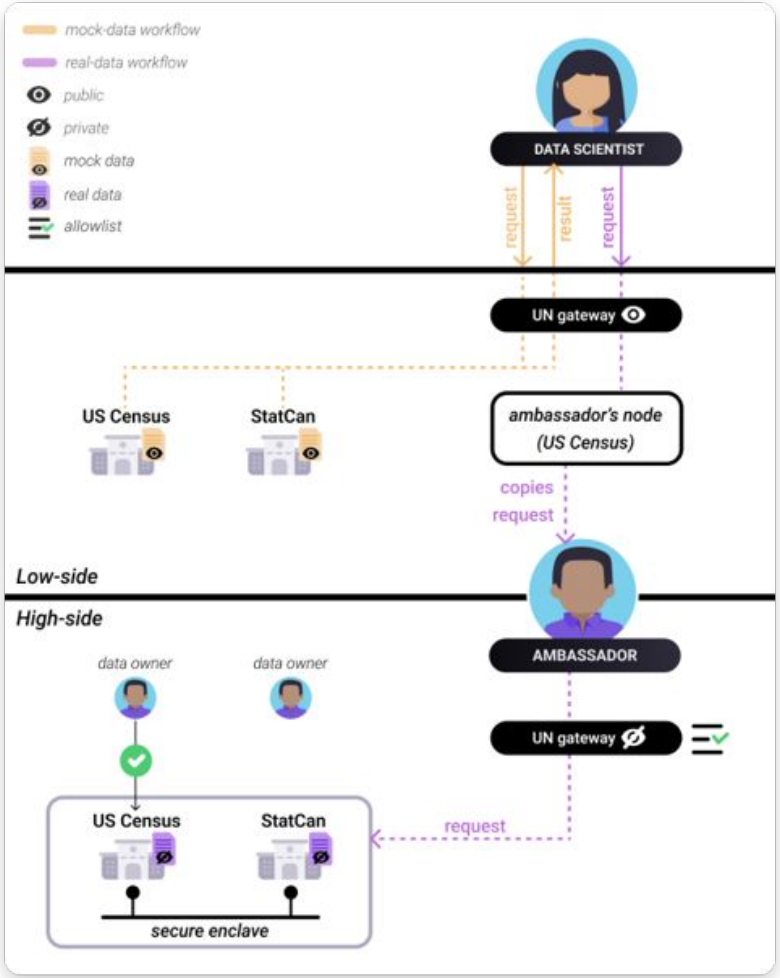
PySyft Data Flow



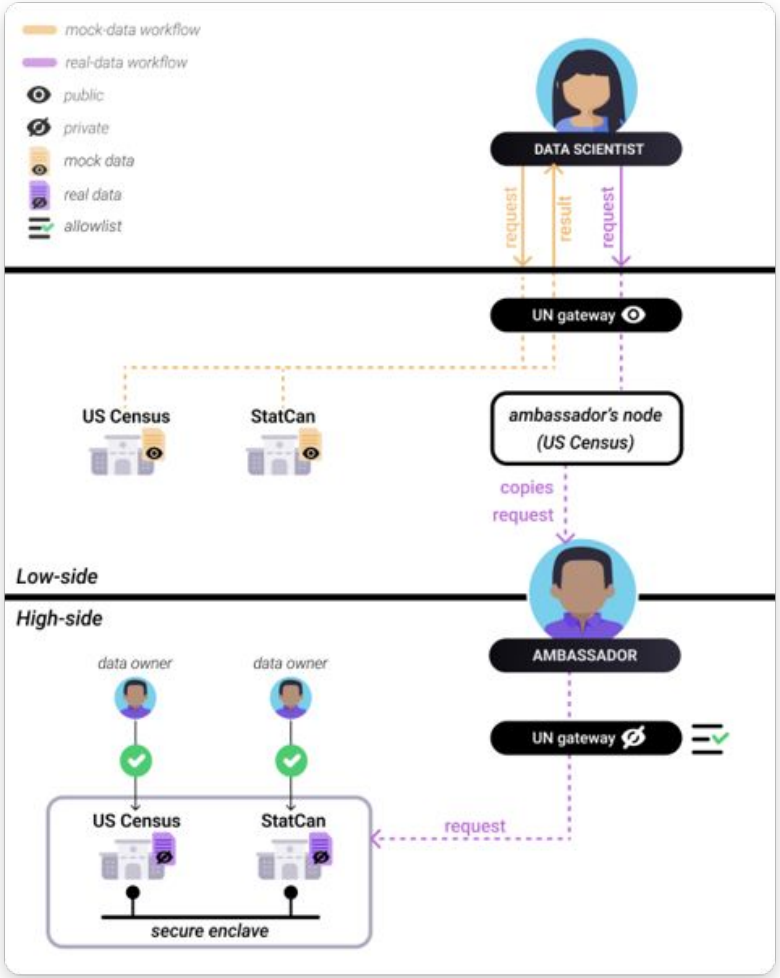
PySyft Data Flow



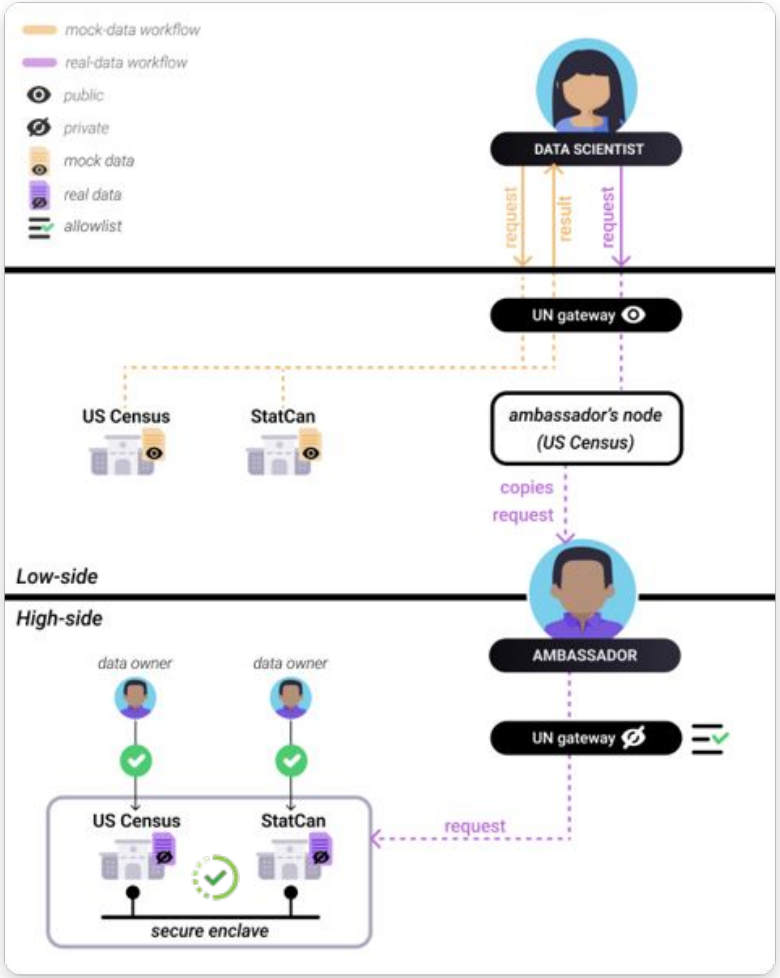
PySyft Data Flow



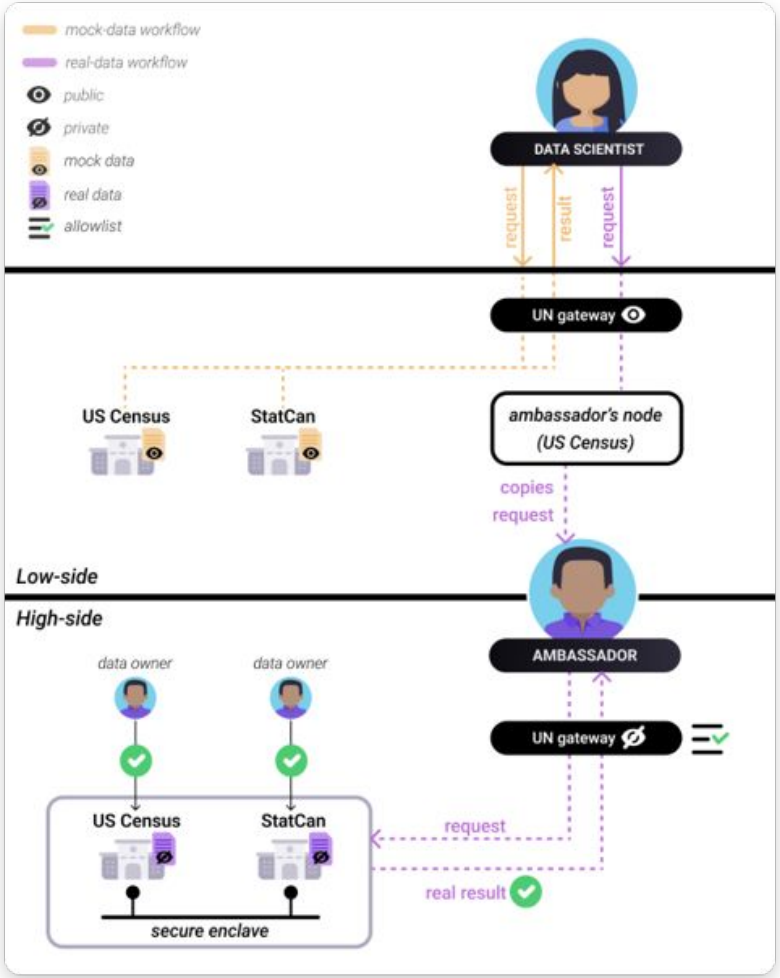
PySyft Data Flow



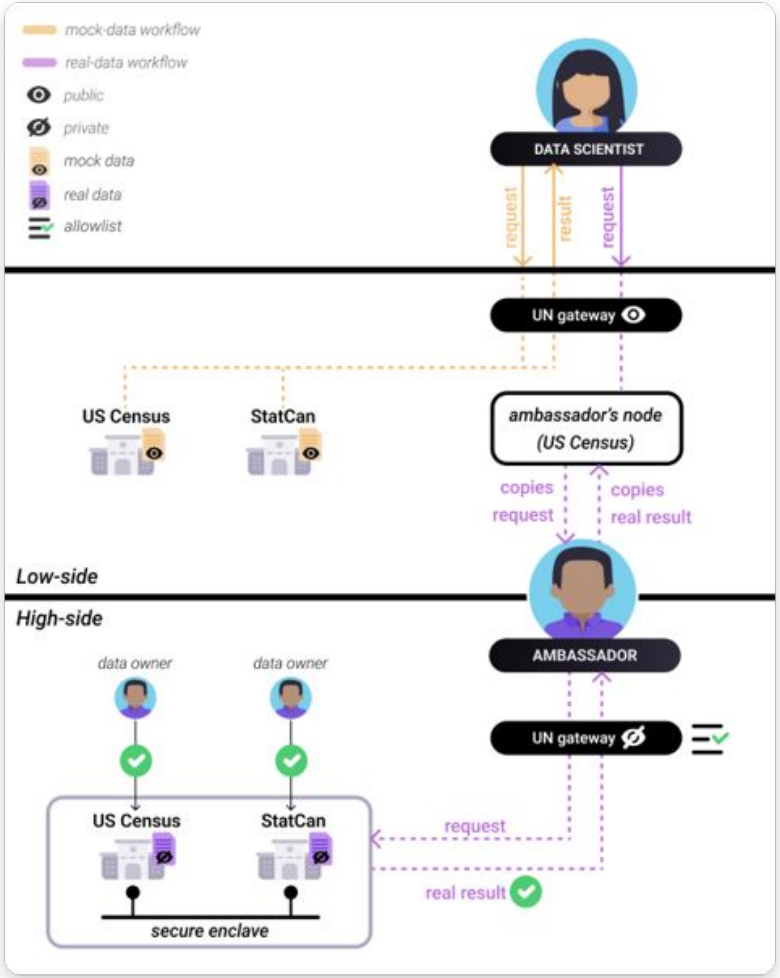
PySyft Data Flow



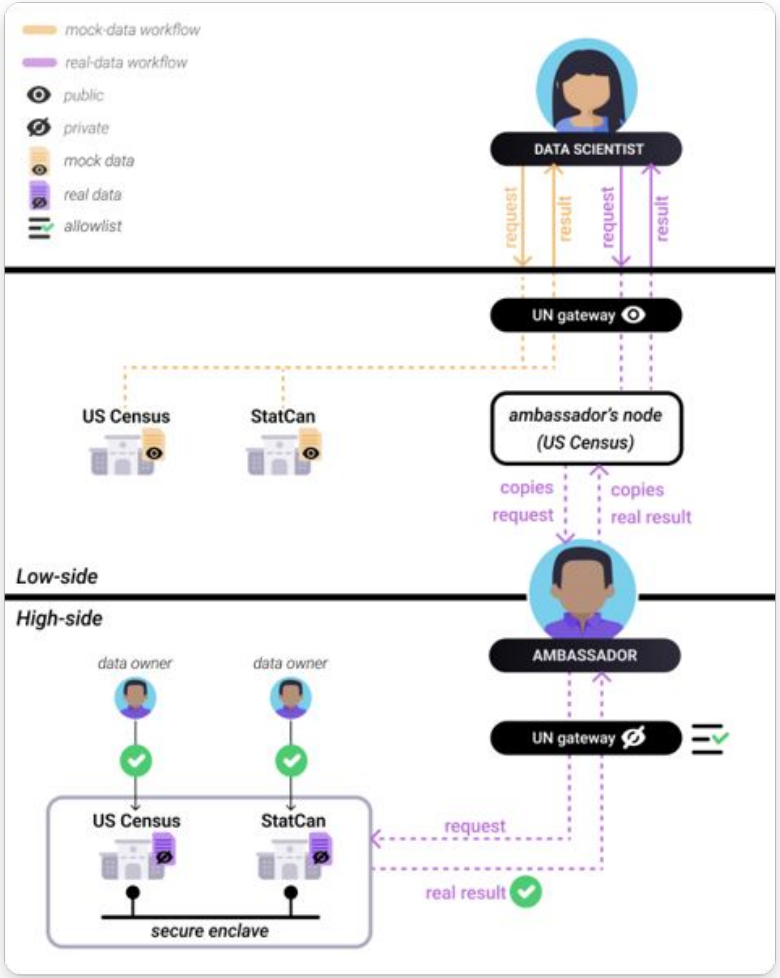
PySyft Data Flow



PySyft Data Flow



PySyft Data Flow



Comparing Traditional Census Collaboration Methods with PySyft

—

Traditional Methods

- Require access to **FSRDC** or creating a formal **data-sharing agreement**
- FSRDC access requires **SSS** (not guaranteed)
- **Export controls** and laws such as **Title 13** limit what data is available outside of the United States
- Some **aggregate data** is available via online platforms such as USA Trade Online

PySyft

- Private **data is hosted on internal domain servers** and is directly accessed only by data owners, **guaranteeing mutual secrecy** for both data owners
- The high-side **domain server is part of a restricted network** accessible only by contributing organizations
- **Code can be executed directly on private data, subject to review and approval** by data owners, lowering barriers for data scientists

4. Lessons Learned So Far & Future Goals

xD
https://www.xd.gov

United States™ Census Bureau

Current Project Status

- Demo joins on open data achieved between **Istat – StatCan** (February) and **Census – StatCan** (May), concluding phase 1

- Census Bureau instance of PySyft hosted on **Cloud.gov**

- **Working on white paper** to be published on Census Bureau research portal

Lessons Learned

- PETs are promising, although current implementations are often still in **pre-production development**
- Deploying PETs systems requires **rethinking how government digital infrastructure is deployed**
- Working with and selling PETs requires **rethinking how we talk about privacy and security** with data, e.g., what does data "access" mean in the context of PETs?

Future Directions and Ideas

- Move towards phase 2: synthetic data based on private data
- Continuing and growing partnerships with **trade data exchanges**
- Exploring pilot projects for other types of data, such as **public health data**
- Perform PySyft join with other PETs such as **differential privacy**
- **Exploring deployment and infrastructure options** for data security, compute flexibility, and cost-savings
- Explore opportunities for **new and updated government policies** related to PETs usage

Links

Census Bureau: https://www.census.gov

xD homepage: https://xd.gov

- xD Emerging Technology Fellowship: https://www.xd.gov/apply

UN PET Lab: https://unstats.un.org/bigdata/task-teams/privacy

UN Comtrade data hub: https://comtradeplus.un.org

StatCan: https://www.statcan.gc.ca

Istat (English site): https://www.istat.it/en

OpenMined: https://openmined.org

- PySyft (Github repository): https://github.com/OpenMined/pysyft

Cloud.Gov: https://cloud.gov

Thank you!