24-09-2024

How to study data that you don’t have access to?

Irina Bejan (OpenMined)

OpenMined

OpenMined

We are a non-profit foundation who builds open source technology to help ease the challenges of researching sensitive data.

KEY PROBLEM:

Our ability to answer important questions is limited

because we can’t access existing data

OpenMined

KEY PROBLEM:

Our ability to answer important questions is limited

because we can’t access existing data

Another country

Another org

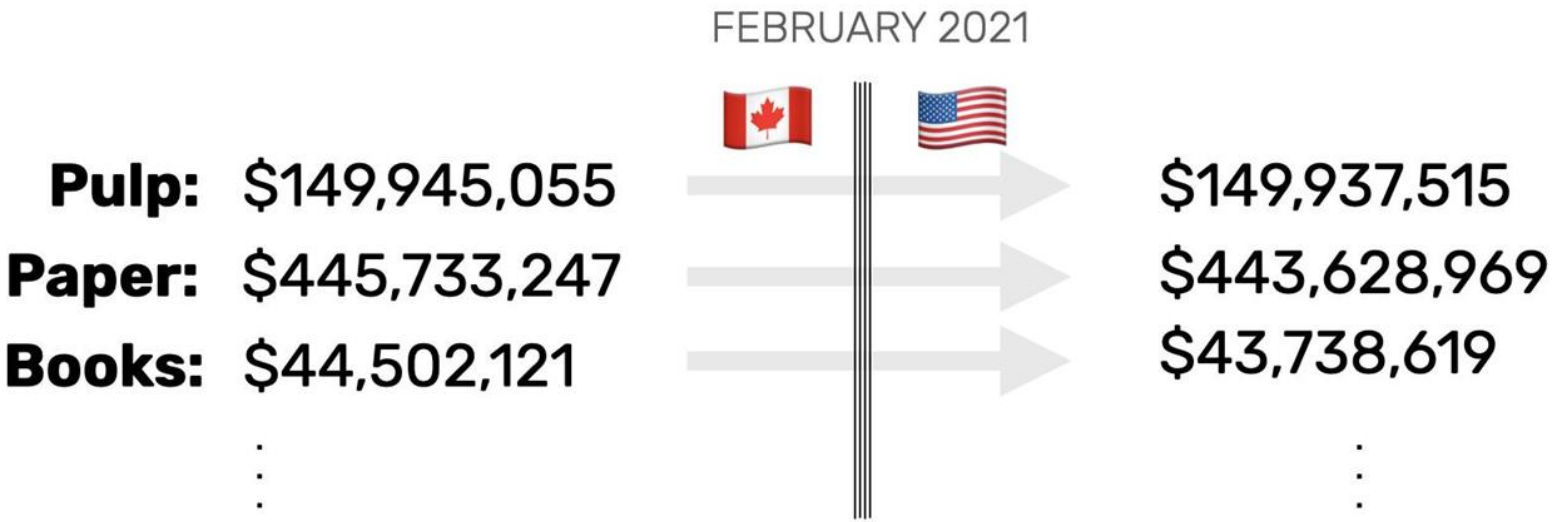
Another department

OpenMined

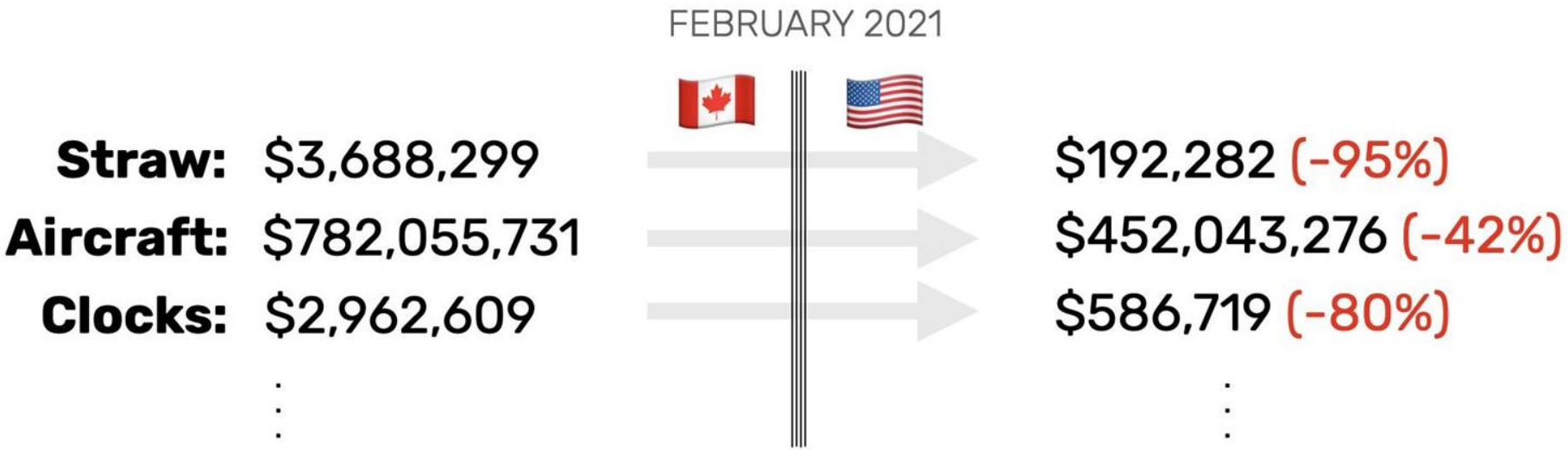
Use Case 1: **Global Trade**

OpenMined

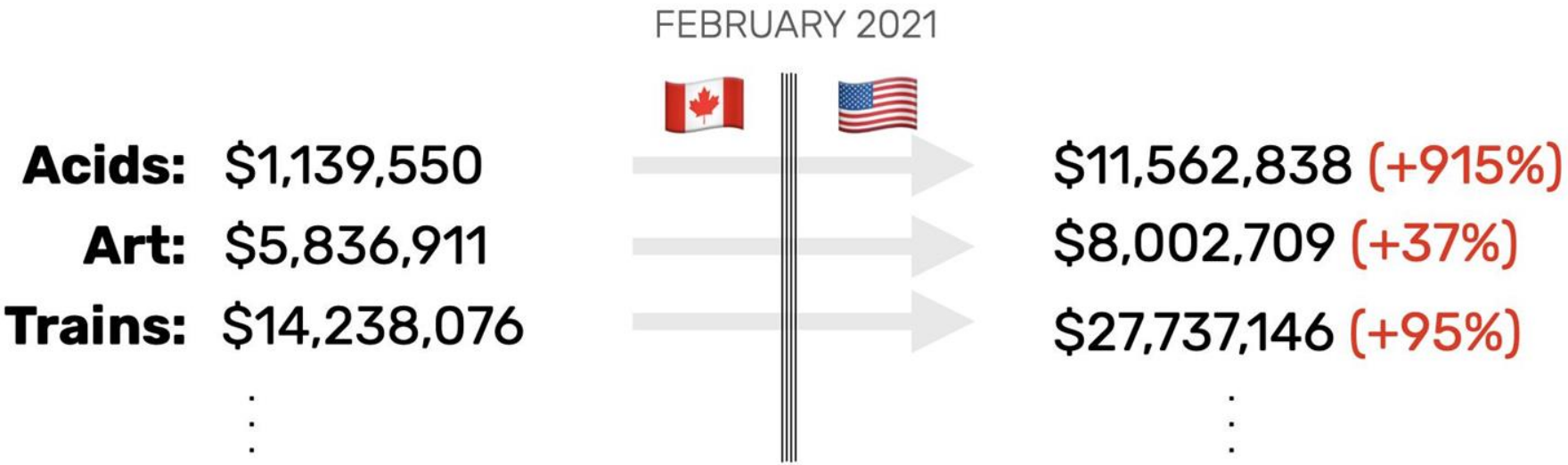
Canada's exports to the USA ***usually*** ~match USA's imports from Canada



… but sometimes goods ***disappear***



… other times goods ***multiply***



KEY PREMISE:

us can't access most of the world's data about its imports...

...because it's stored in other country

OpenMined

KEY PREMISE:

CA can't access most of the world's data about its imports...

...because it's stored in other country

OpenMined

Economic Surveillance

Trade imbalances

Market inefficiencies

Price fluctuations

Due of the lack of access,

important questions are left unanswered

Supply shortages

Foreign Policy misalignment

Tarrifs and taxes

OpenMined

WHY CAN'T WE TRACK GLOBAL TRADE TOGETHER?

The world's national trade statistics orgs ***can't*** put all their data into **one place**

国税庁
NATIONAL TAX AGENCY

Statistics Bureau of Japan

Use Case 2: **Money Laundering**

OpenMined

UNITED NATIONs OFFICE on DRUGS & CRIME:

$1.6 trillion in criminal proceeds are laundered annually

OpenMined

WHY CAN'T WE STOP IT?

can't access most of the world's data about where its money comes from...

...because it's stored in other banks

OpenMined


BMO Bank of Montreal
RBC
TD Bank
Scotiabank
CIBC
BNP PARIBAS
SOCIETE GENERALE
GROUPE BPCE
Crédit Mutuel
WHY CAN'T WE TRACK MONEY LAUNDERING TOGETHER?
The world's banks
can't put all their ledger data
into one place to follow the money
citi
SUNTRUST
usbancorp
J.P.Morgan
PNC
WELLS FARGO
BANK OF AMERICA
MIZUHO
JP BANK ゆうちょ銀行
Japan Net Bank
SMFG
MUFG

Use Case 3: **Emergency Response**

OpenMined

UNITED NATIONs:

In 2020, approximately 270 million people were affected by disasters worldwide.

OpenMined

The IFRC states that

134 million people annually need humanitarian assistance due to natural disasters


TELECOM
verizon wireless
Sprint
Level(3)
COMMUNICATIONS
Time Warner Cable
Sunrise
AT&T
AXtel
WHY DON'T WE KNOW WHERE TO SEND HUMANITARIAN AID?
The companies that hold mobility data cannot put it into one place to help track the movement
vodafone
TELECOM ITALIA
comcast BUSINESS COMMUNICATIONS
orange
TATA COMMUNICATION
T-Mobile

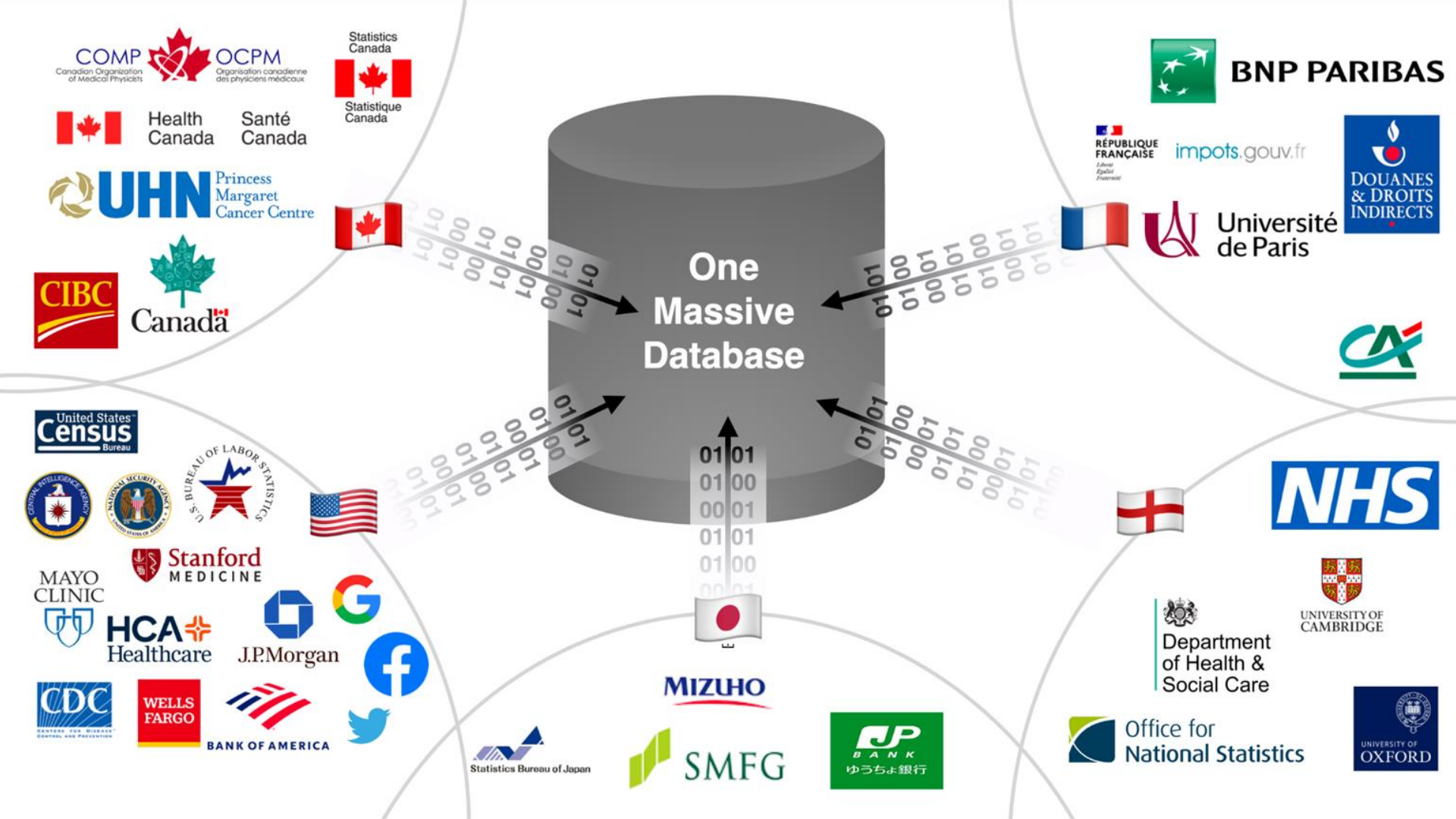
We could continue with more …

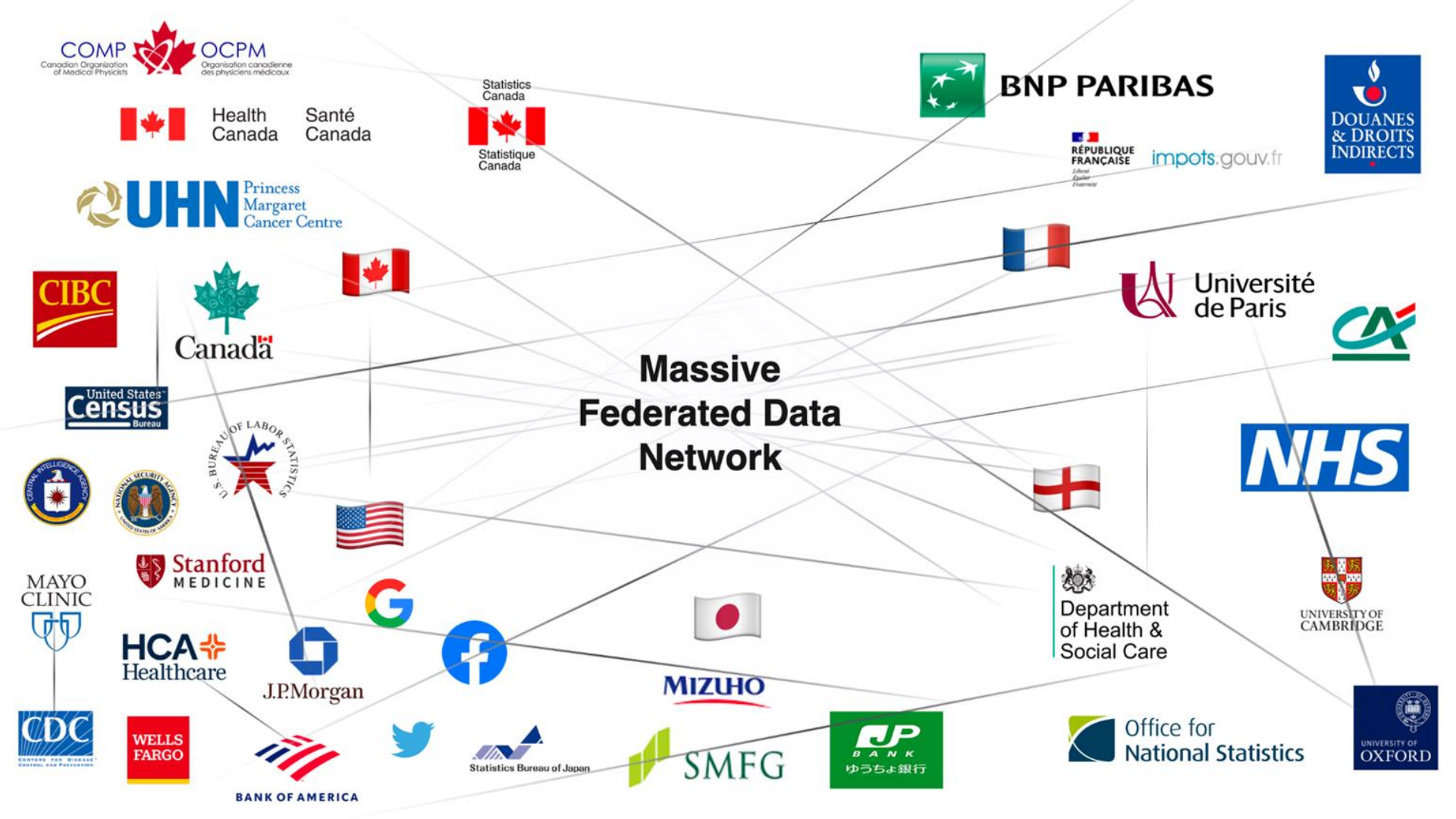
Use Case 1: Global Trade
Use Case 2: Money Laundering
Use Case 3: Emergency Response
Use Case 4: Breast Cancer
Use Case 5: Algorithmic Accountability
Use Case 6: Disinformation


COMP OCPM
Canadian Organization of Medical Physicists
Organisation canadienne des physiciens médicaux
Statistics Canada
Statistique Canada
Health Canada
Santé Canada
UHN Princess Margaret Cancer Centre
CIBC
Canada
BNP PARIBAS
RÉPUBLIQUE FRANÇAISE
impots.gouv.fr
DOUANES & DROITS INDIRECTS
Université de Paris
We could solve <x> important challenge using **existing data** but no-one can access any data but their own
United States Census Bureau
U.S. BUREAU OF LABOR STATISTICS
Stanford MEDICINE
MAYO CLINIC
HCA Healthcare
J.P.Morgan
CDC
WELLS FARGO
BANK OF AMERICA
NHS
UNIVERSITY OF CAMBRIDGE
Department of Health & Social Care
Office for National Statistics
UNIVERSITY OF OXFORD
MIZUHO
Statistics Bureau of Japan
SMFG
JP BANK ゆうちょ銀行

Because..

…data access is *really hard*


COMP OCPM
Canadian Organization of Medical Physicists
Organisation canadienne des physiciens médicaux
Statistics Canada
Statistique Canada
Health Canada
Santé Canada
UHN Princess Margaret Cancer Centre
CIBC
Canada
BNP PARIBAS
RÉPUBLIQUE FRANÇAISE
impots.gouv.fr
DOUANES & DROITS INDIRECTS
Université de Paris
One Massive Database
United States Census Bureau
U.S. BUREAU OF LABOR STATISTICS
Stanford MEDICINE
MAYO CLINIC
HCA Healthcare
J.P.Morgan
CDC
WELLS FARGO
BANK OF AMERICA
NHS
UNIVERSITY OF CAMBRIDGE
Department of Health & Social Care
Office for National Statistics
UNIVERSITY OF OXFORD
MIZUHO
Statistics Bureau of Japan
SMFG
JP BANK ゆうちょ銀行


COMP OCPM
Canadian Organization of Medical Physicists
Organisation canadienne des physiciens médicaux
Health Canada
Santé Canada
Statistics Canada
Statistique Canada
BNP PARIBAS
RÉPUBLIQUE FRANÇAISE
impots.gouv.fr
DOUANES & DROITS INDIRECTS
UHN Princess Margaret Cancer Centre
CIBC
Canada
Université de Paris
United States Census Bureau
U.S. BUREAU OF LABOR STATISTICS
NHS
Massive Federated Data Network
Stanford MEDICINE
MAYO CLINIC
Department of Health & Social Care
UNIVERSITY OF CAMBRIDGE
HCA Healthcare
J.P.Morgan
MIZUHO
CDC
WELLS FARGO
BANK OF AMERICA
Statistics Bureau of Japan
SMFG
JP BANK ゆうちょ銀行
Office for National Statistics
UNIVERSITY OF OXFORD

How are we solving it?

OpenMined

HOMOMORPHIC ENCRYPTION

FUNCTIONAL ENCRYPTION

FEDERATED LEARNING

BLOCKCHAIN?

SECURE ENCLAVES

Privacy Enhancing Technologies (PETs)

ZERO-KNOWLEDGE PROOFS

K-ANONYMIZATION

SYNTHETIC DATA

SECURE MULTI-PARTY COMPUTATION

DIFFERENTIAL PRIVACY

OpenMined

PETS MAKE IT POSSIBLE TO:

answer a question using data owned by someone else

- [x] In Another Country
- [x] In Another Org
- [x] In Another Dept.

OpenMined

PETS MAKE IT POSSIBLE TO:

answer a question using data owned by someone else

- ☑ In Another Country
- ☑ In Another Org
- ☑ In Another Dept.

This is the ability that matters!

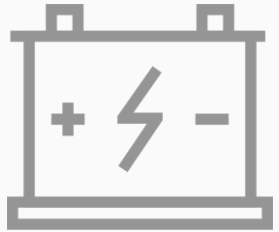
- HOMOMORPHIC ENCRYPTION
- FUNCTIONAL ENCRYPTION
- SYNTHETIC DATA
- FEDERATED LEARNING
- SECURE MULTI-PARTY COMPUTATION
- DIFFERENTIAL PRIVACY
- SECURE ENCLAVES
- K-ANONYMIZATION
- ZERO-KNOWLEDGE PROOFS

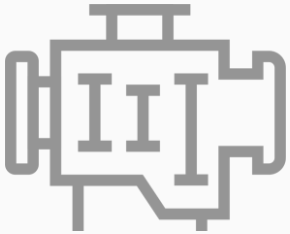
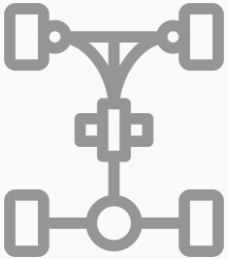
These are just algorithms!

By analogy: everyone is working on car parts…






…but we don’t yet have a car.

KEY QUESTION:

What is the “car” of PETs?

OpenMined

OpenMined / PySyft

Sponsor | Used by 16 | Unwatch 167 | Unstar 3,893 | Fork 869

Code | Issues 162 | Pull requests 17 | Actions | Security | Insights | Settings

A library for encrypted, privacy preserving deep learning

Edit

deep-learning secure-computation pytorch privacy cryptography Manage topics

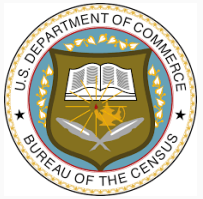
5,337 commits | 42 branches | 5 releases | 1 environment | 180 contributors | Apache-2.0

Branch: dev | New pull request | Create new file | Upload files | Find File | Clone or download

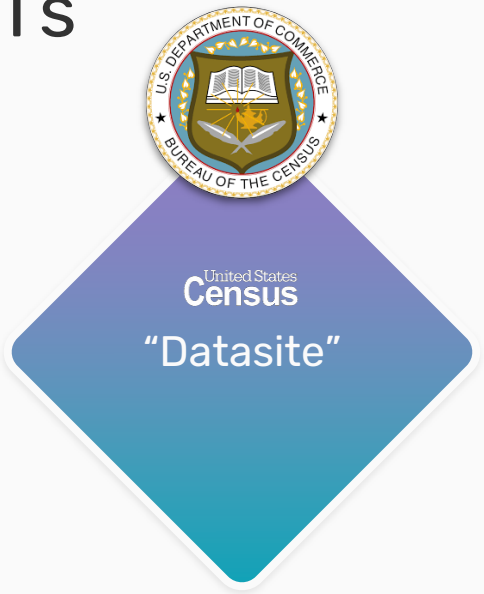
jvmancuso and robert-wagner rm needless integration test (#2629) — Latest commit 8d8baa4 3 days ago

.github/ISSUE_TEMPLATE	Update issue templates	4 months ago
art	Improve the diagram	3 months ago
docker-image	delete unecessary package installation, numpy comes already with pysyft	4 months ago
docs	bumpversion 0.1.27a1 -> 0.1.28a1 (#2619)	13 days ago
examples	Improvements to the Federated Recurrent Neural Network notebook (#2613)	18 days ago
images	Add files via upload	3 months ago
syft	Implementing Protocol (#2605)	7 days ago
test	rm needless integration test (#2629)	3 days ago
.flake8	changed ignore to select	6 months ago

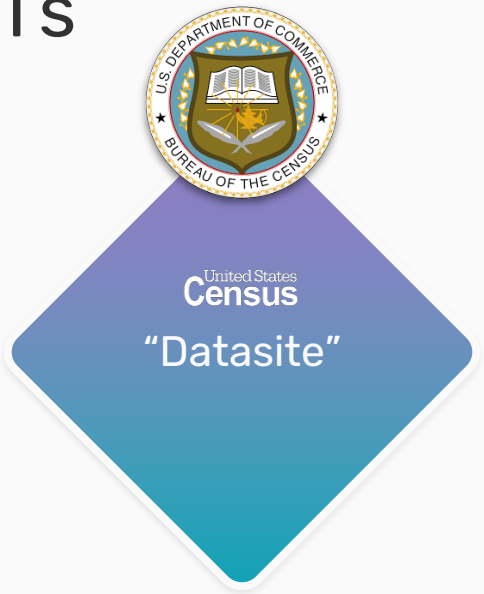
The “car” of PETs



The “car” of PETs



The “car” of PETs



It’s like an Apache Web Server for **private** data

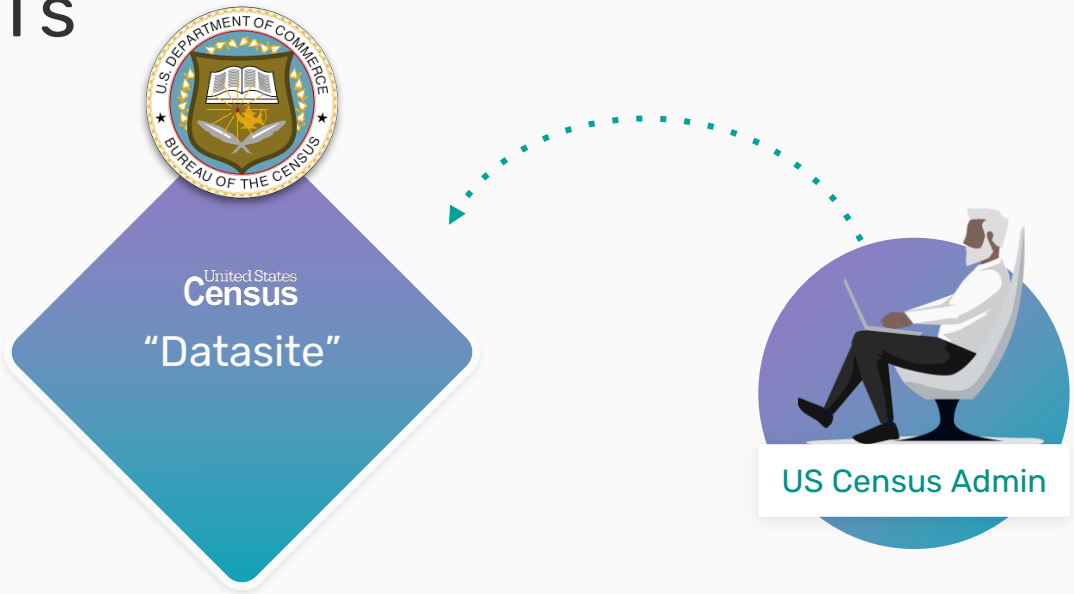
The “car” of PETs

United States Census

“Datasite”

US Census Admin

The “car” of PETs

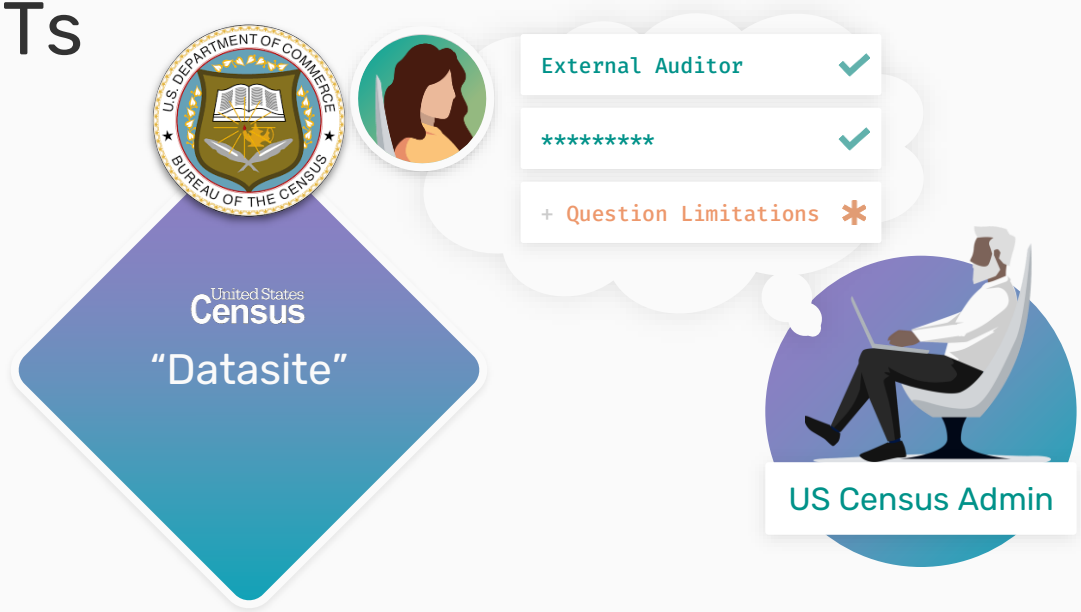


1 Load data into datasite

2 Creates an account for ext. researcher

3 … goes and has a coffee… …(or tea)

The “car” of PETs

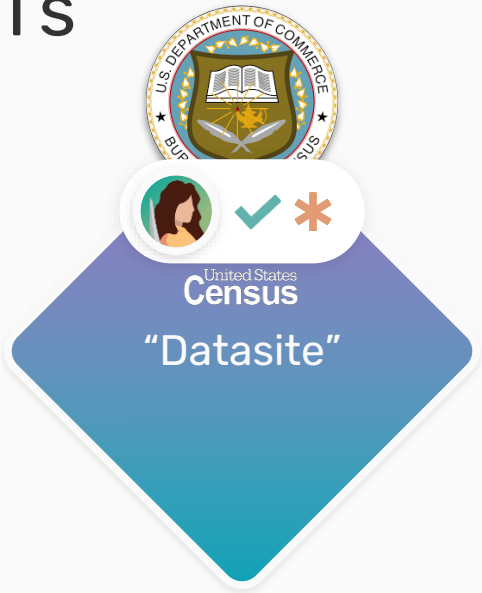


1 Load data into datasite

2 Creates an account for ext. researcher

3 … goes and has a coffee… …(or tea)

The “car” of PETs



US Census Admin

bye!

1 Load AI or user logs into datasite

2 Creates an account for ext. researcher

3 … goes and has a coffee…
…(or tea)

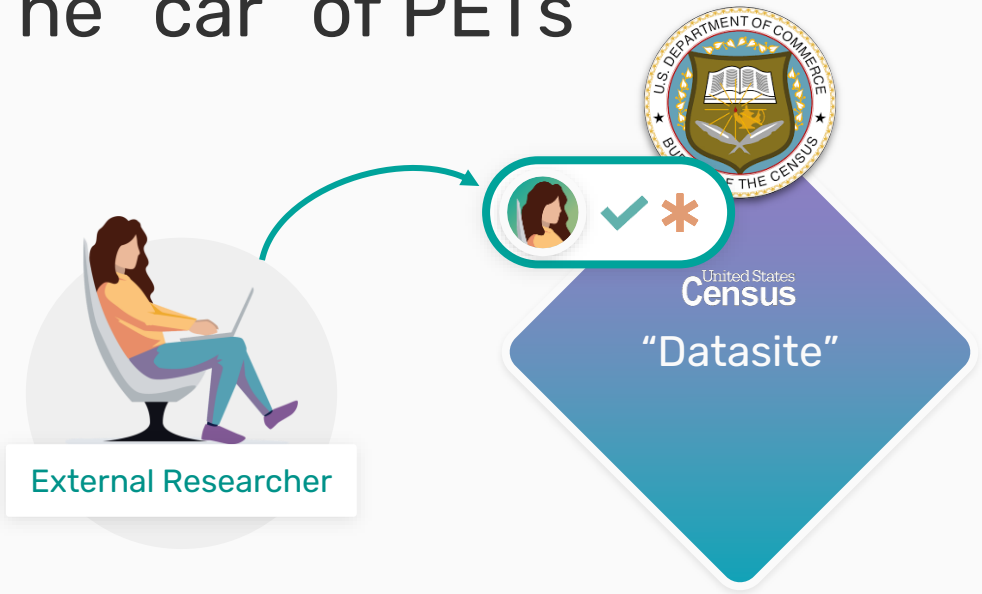
The “car” of PETs

External Researcher

United States Census

“Datasite”

The “car” of PETs

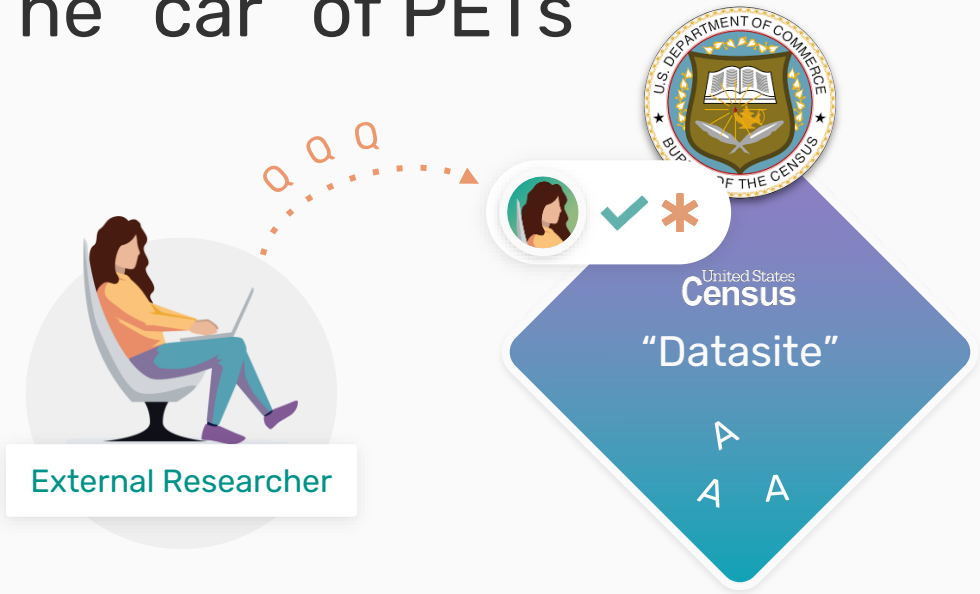


1 Login to Datasite

2 Get answers to allowed questions

3 Download Answers

The “car” of PETs



This is the PETs part.

- SECURE ENCLAVES
- FUNCTIONAL ENCRYPTION
- HOMOMORPHIC ENCRYPTION
- SECURE MULTI-PARTY COMPUTATION
- FEDERATED LEARNING
- ZERO-KNOWLEDGE PROOFS
- K-ANONYMIZATION
- DIFFERENTIAL PRIVACY
- SYNTHETIC DATA

1 Login to Datasite

2 Submit research questions

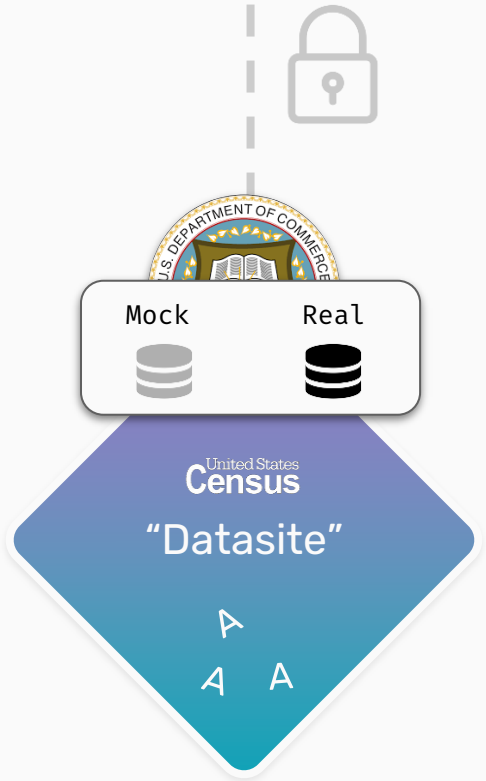
3 Download Answers

Open-access mock data

External Researcher

Securely-hosted private data

US Census Admin

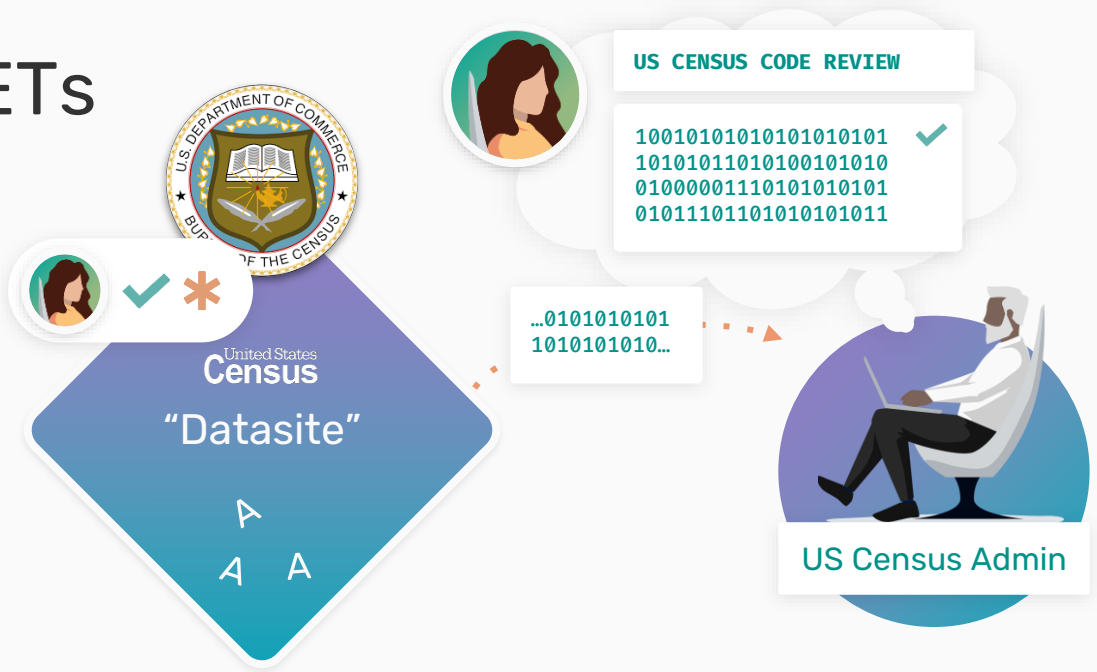


1 Login to Datasite

2 Submit research questions

3 Download Answers

The “car” of PETs



1 Loads data into datasite

2 Creates an account for ext. researcher

3 Code review and approval

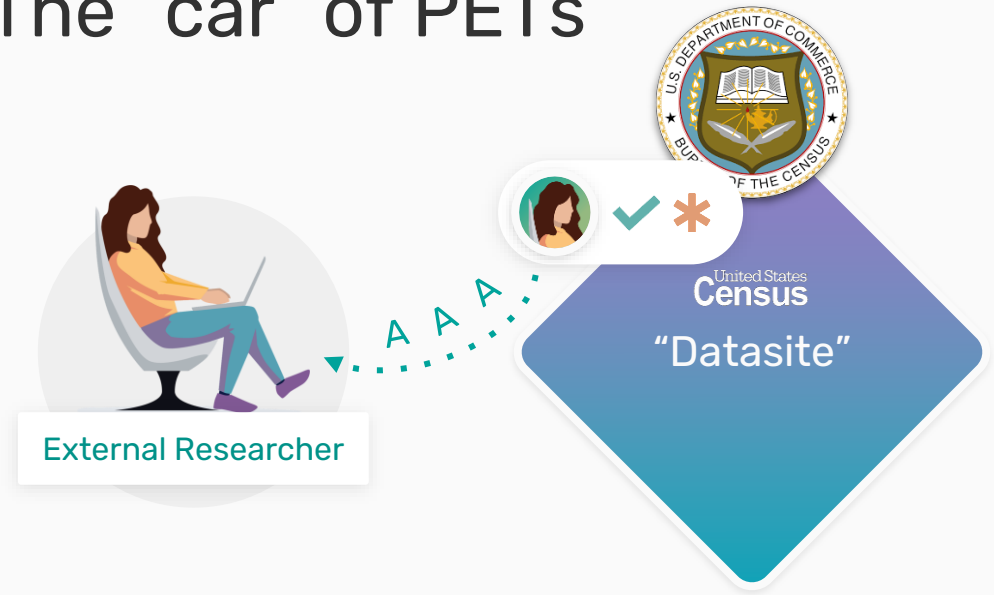
The “car” of PETs

1 Load data into datasite

2 Creates an account for ext. researcher

3 Code review and approval

The “car” of PETs

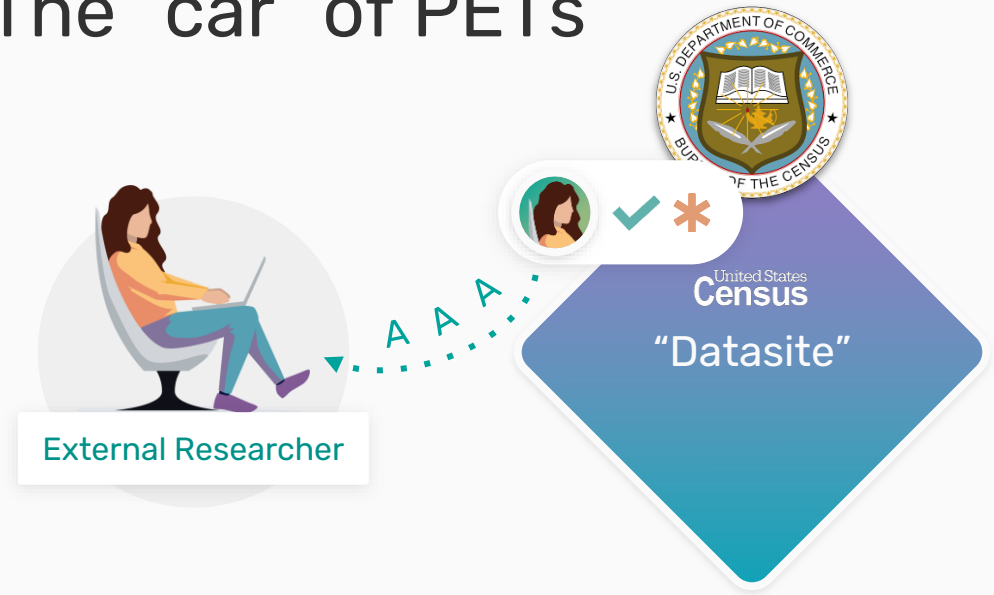


1 Login to datasite

2 Submit research questions

3 Download Answers

The “car” of PETs



Notice what’s missing!

- **Privacy Risk:** researcher only sees results
- **Security Risk:** researcher doesn’t have internal access to secure systems
- **IP Risk:** researcher only sees results (no other IP), as checked by admin.

1 Login to Domain Server

2 Submit research questions

3 Download Answers

The “car” of PETs

External Researcher

United States Census

“Datasite”

UN PET Lab

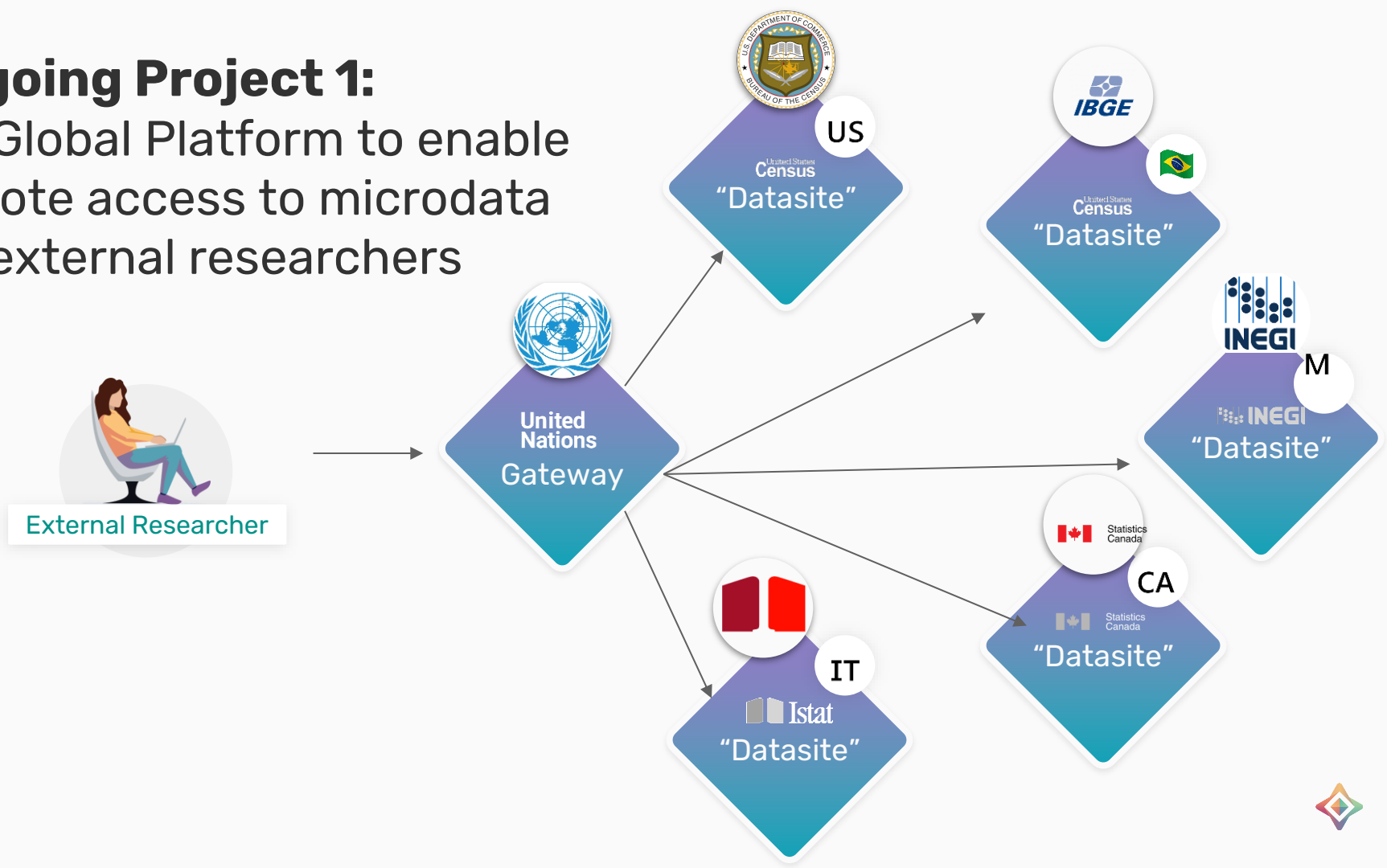
OpenMined

PETs for NSOs

- Mandate to serve the public interest by
  - Protecting citizen data
  - Maximising the public benefits of the data
- To maximise the public benefits, NSOs need third-party data, such as:
  - administrative data from other agencies
  - data held in the private sector

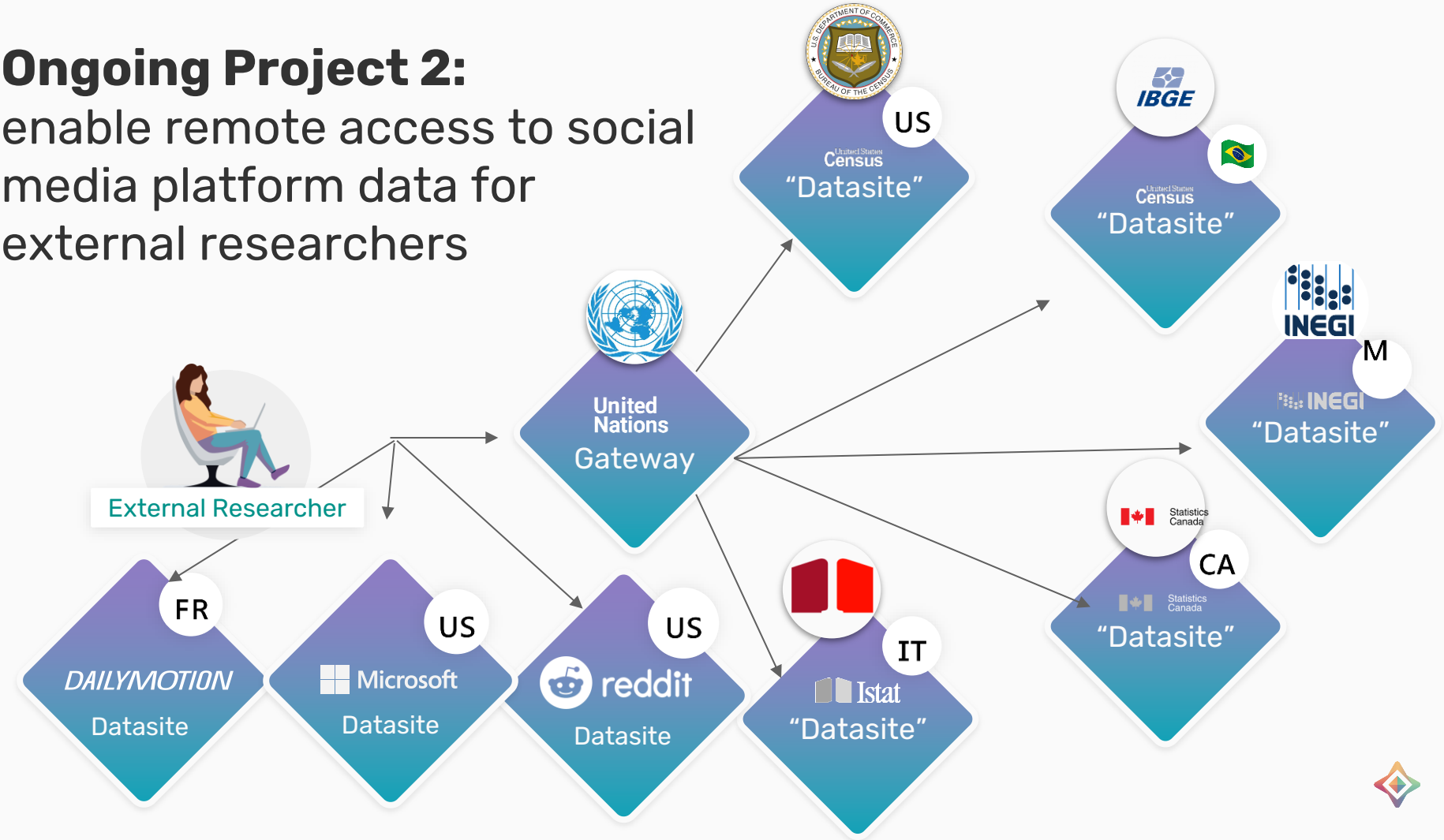
Ongoing Project 1:

UN Global Platform to enable remote access to microdata for external researchers



Ongoing Project 2:

enable remote access to social media platform data for external researchers



PETS MAKE IT POSSIBLE TO:

answer a question using data owned by someone else

- [x] In Another Country
- [x] In Another Org
- [x] In Another Dept.

OpenMined

But there’s another problem…

PROBLEM

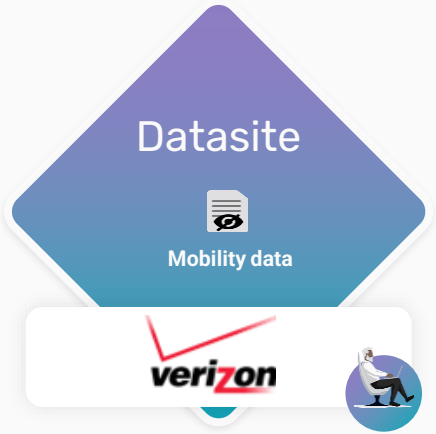
To jointly study private data owned by two or more organisations, the data needs to be put on the same computer.

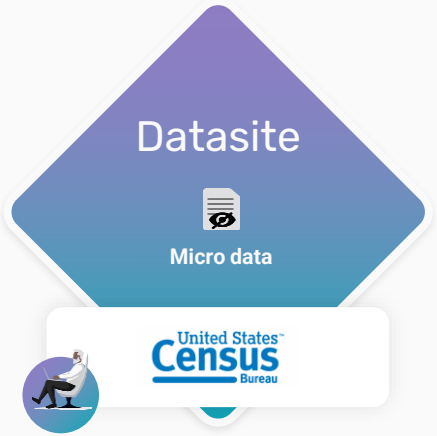
PROBLEM

They need **mutual secrecy**

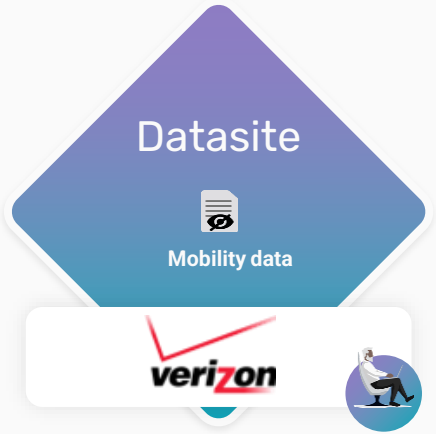
OpenMined

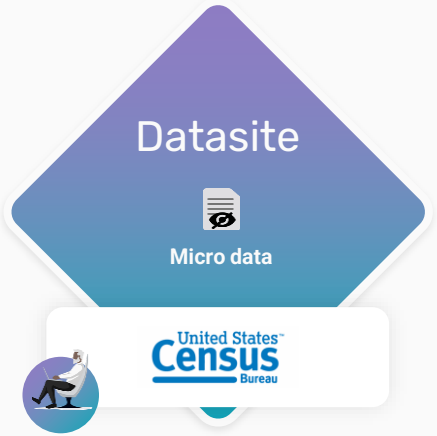
StatCan & US Census launch datasites..



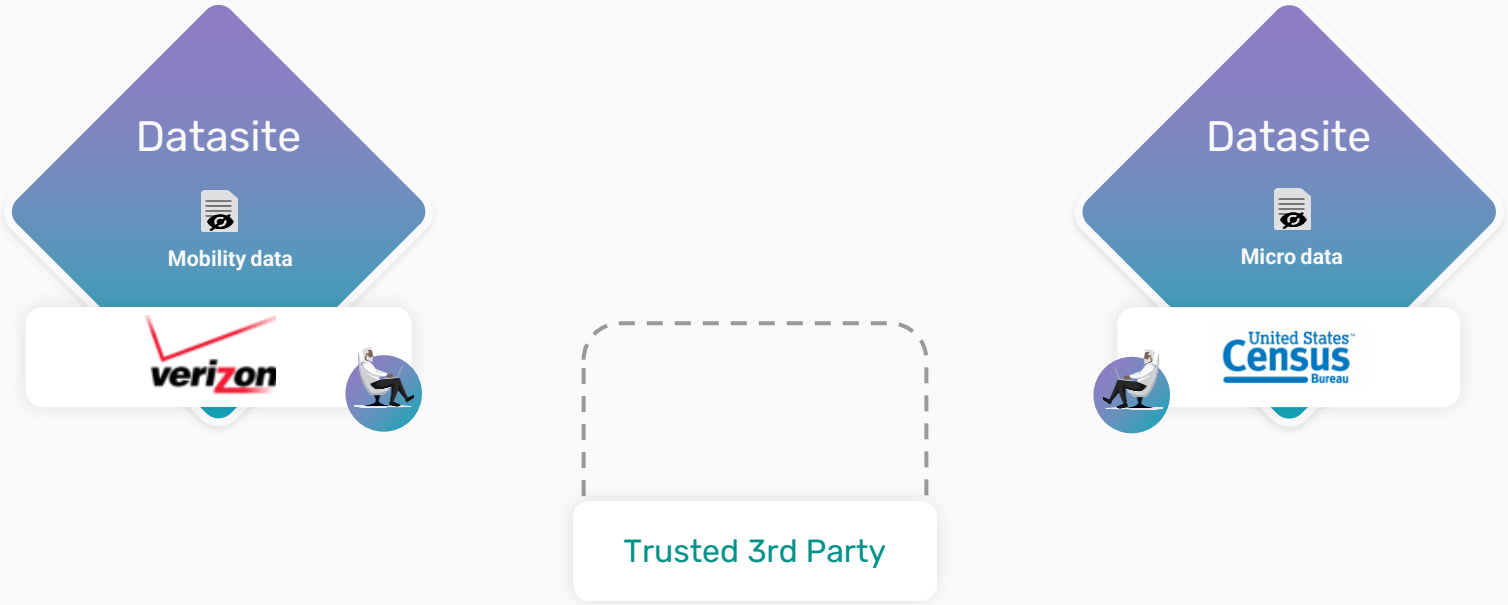


…but how can mutual secrecy be achieved?

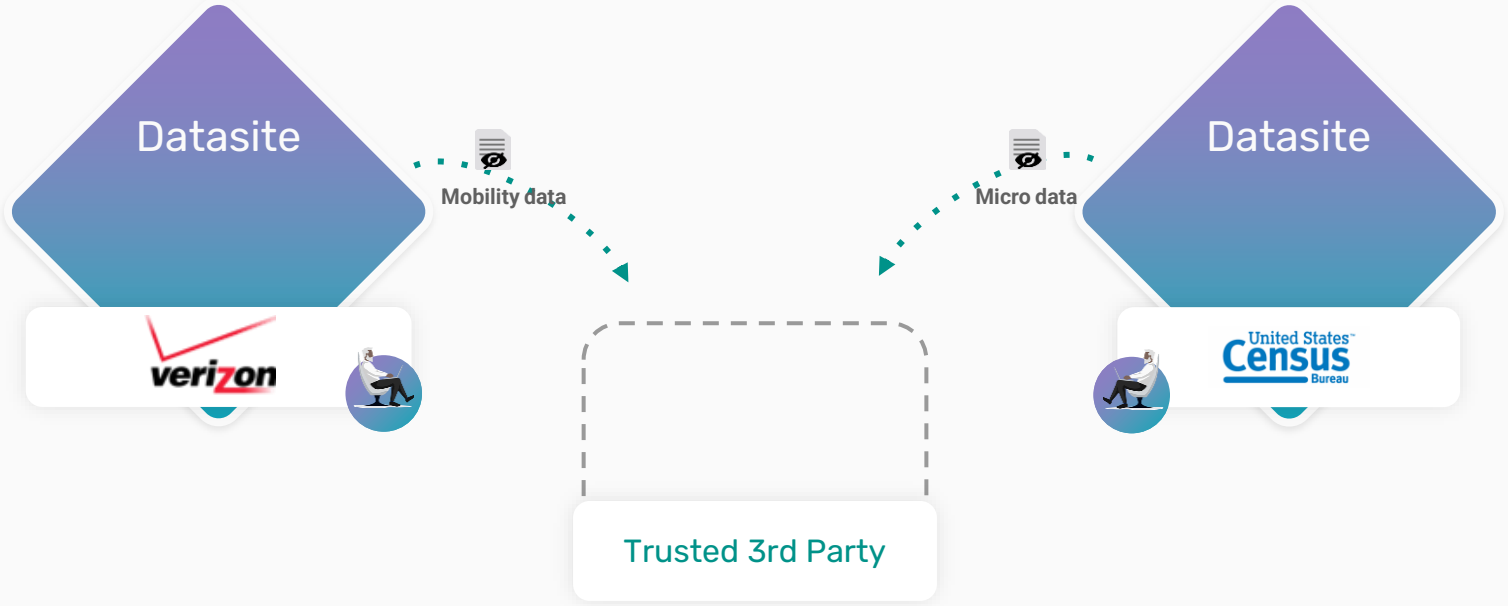




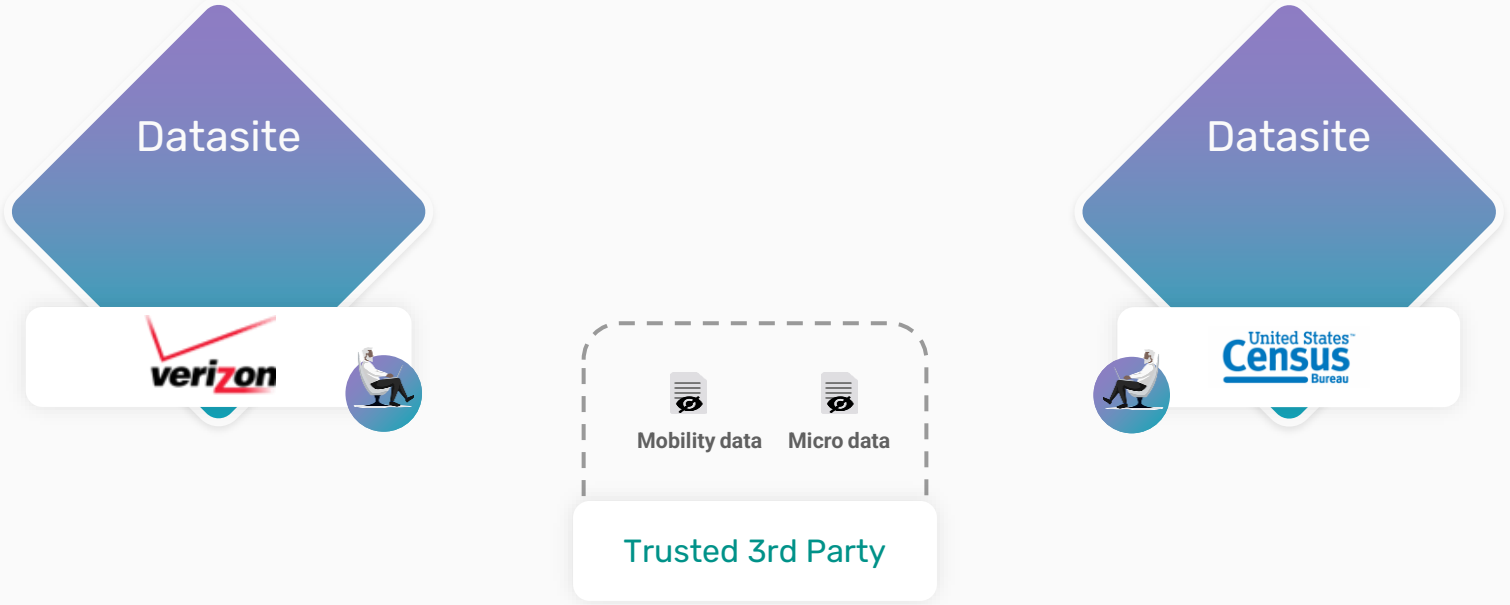
…but how can mutual secrecy be achieved?



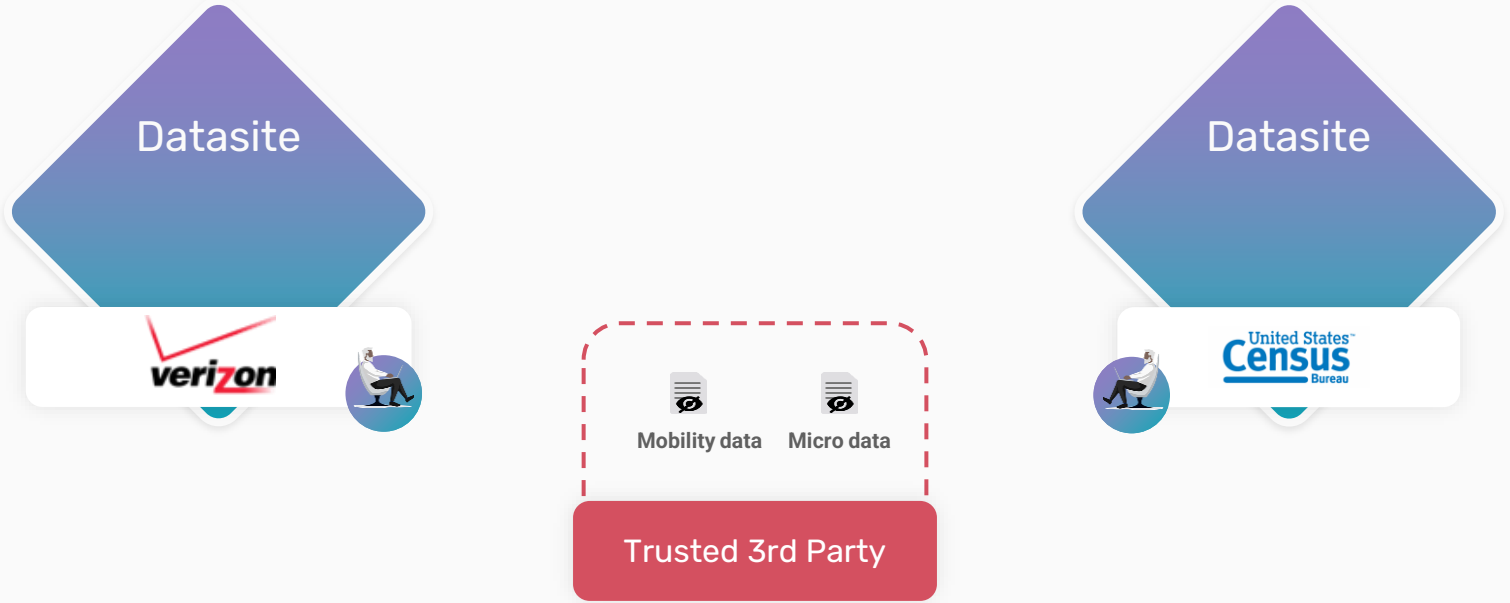
…but how can mutual secrecy be achieved?



…but how can mutual secrecy be achieved?



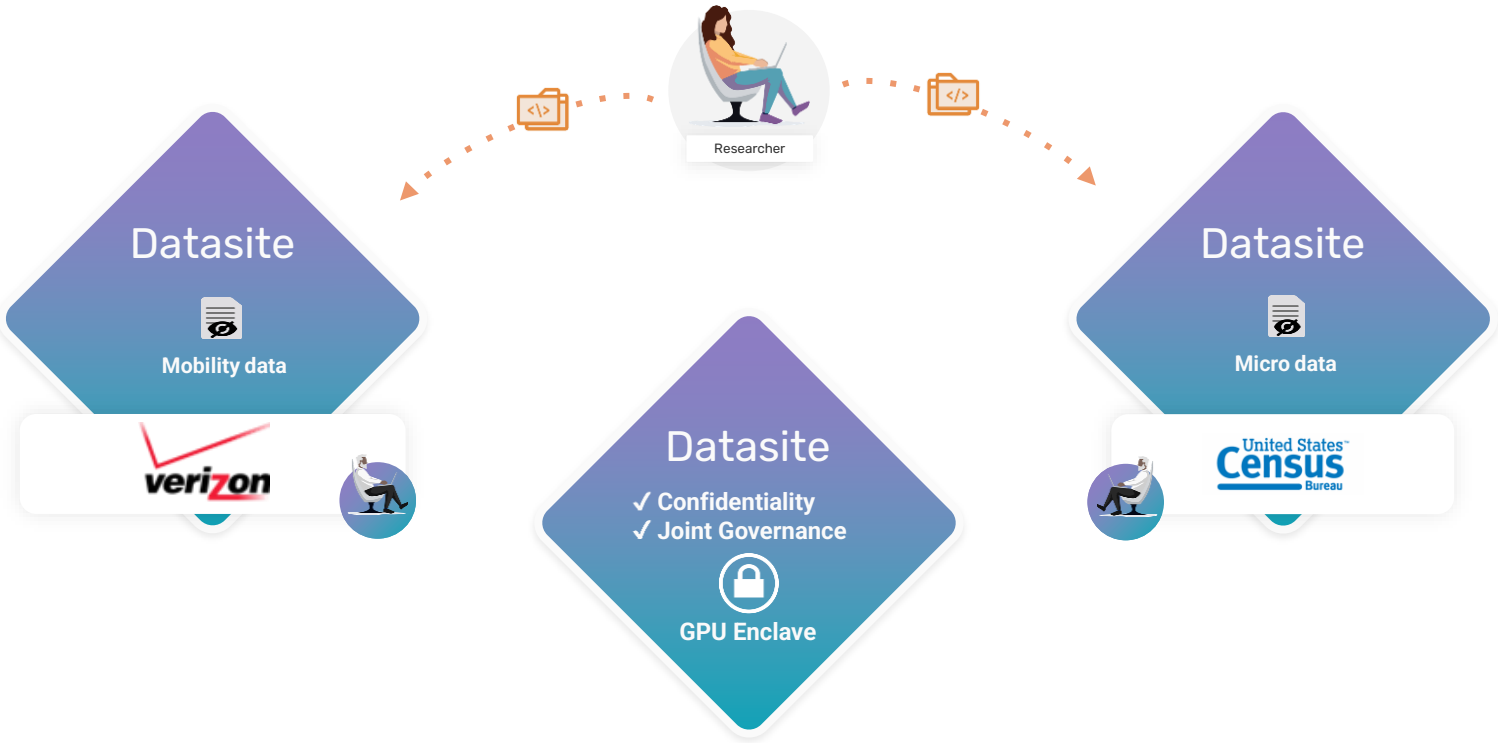
…but how can mutual secrecy be achieved?



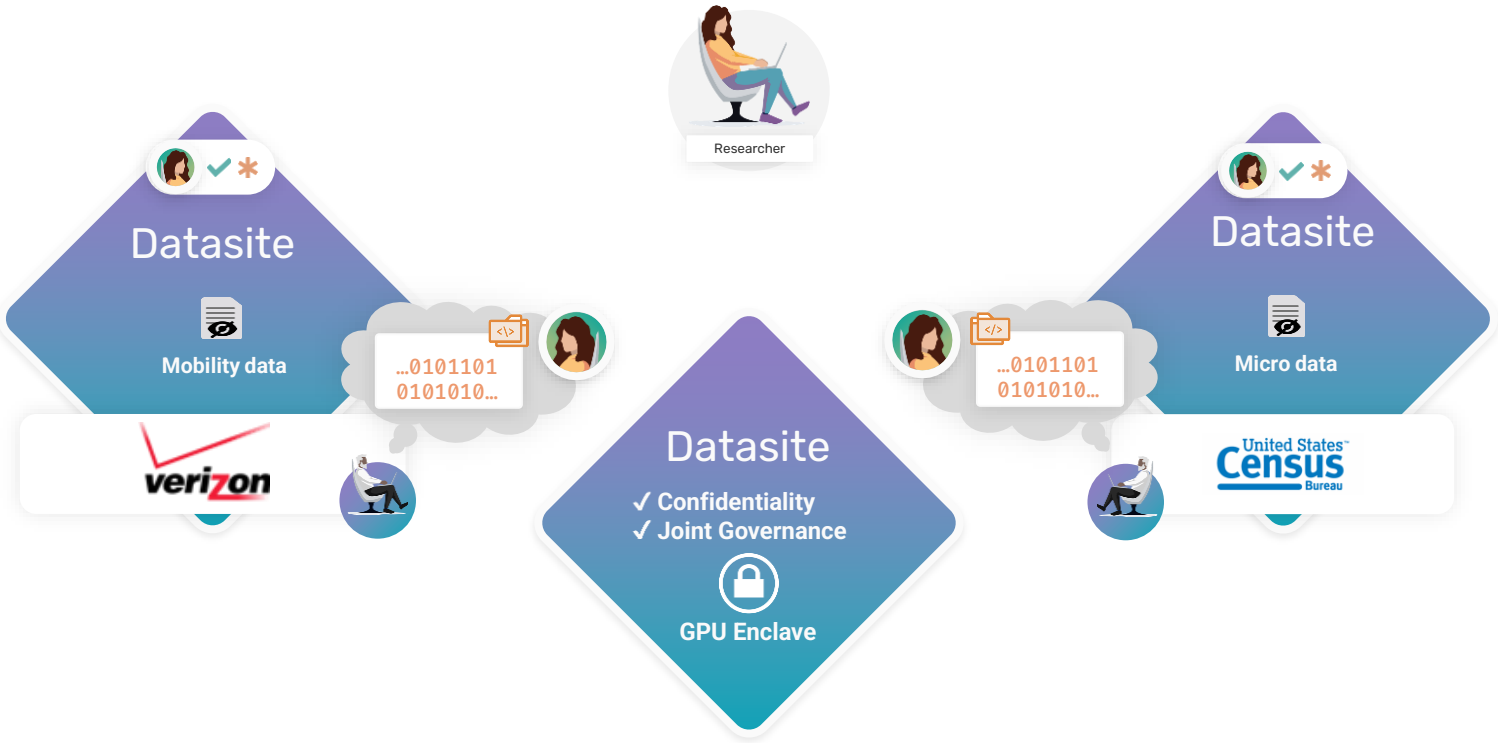
Solution: Launch a secure enclave

How does it work?

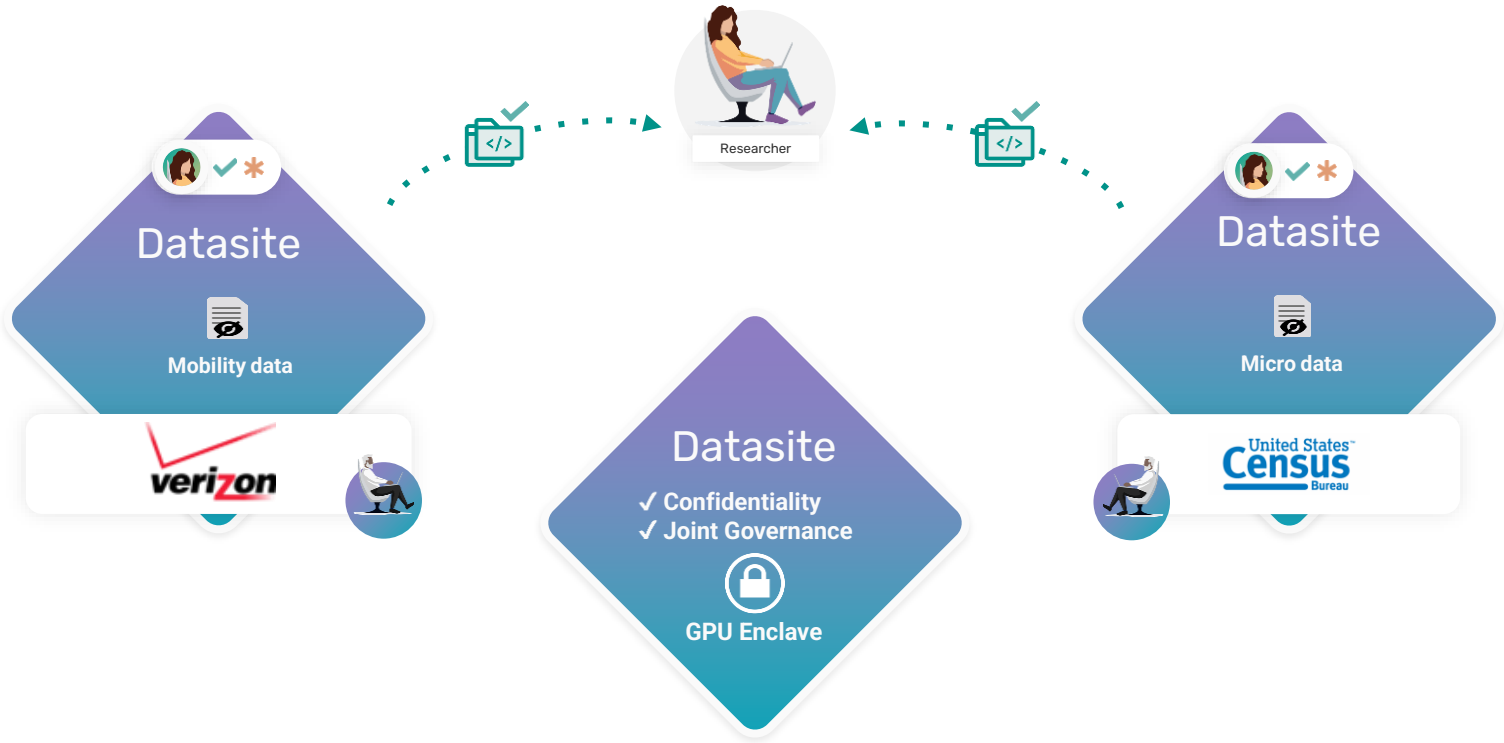
Step 1: External Researcher proposes a project



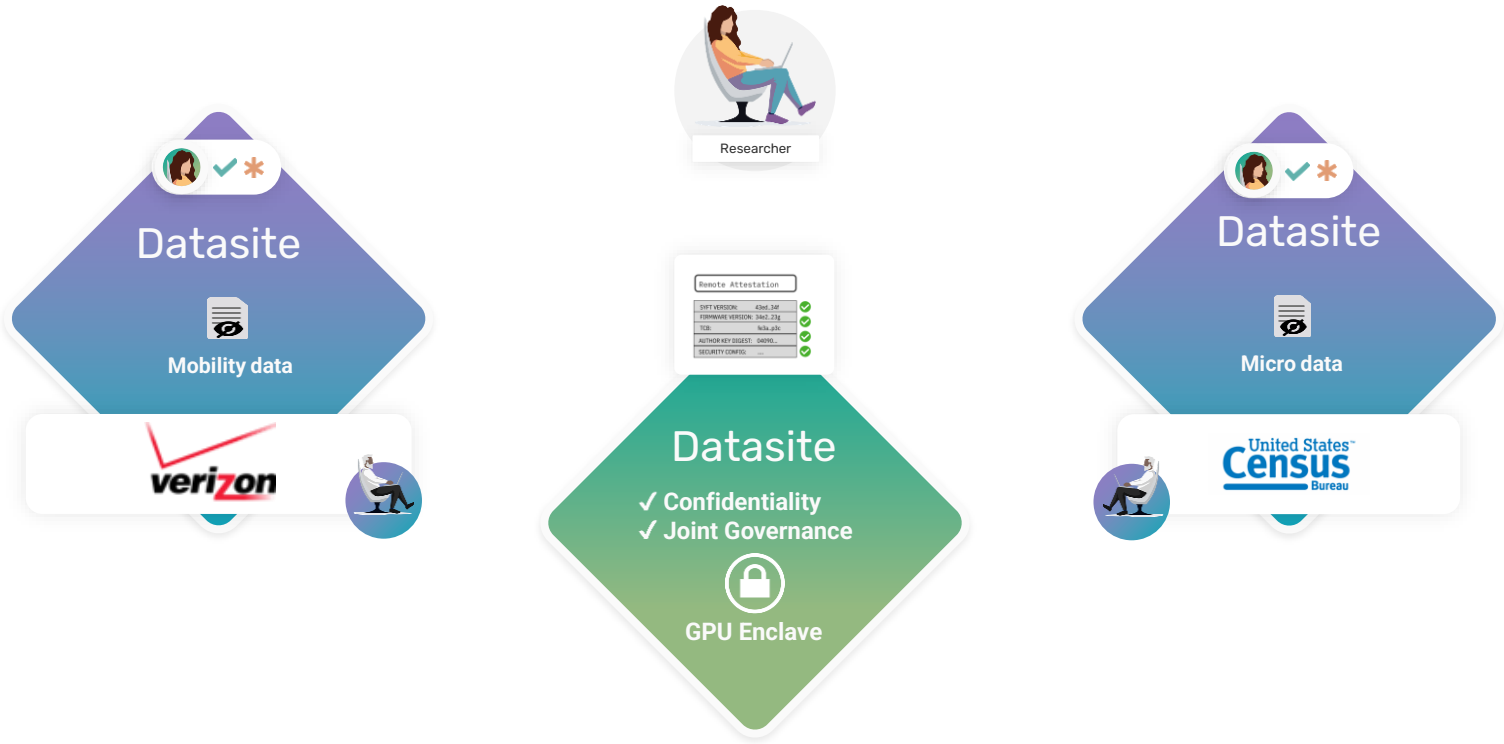
Step 2: Admins review the code



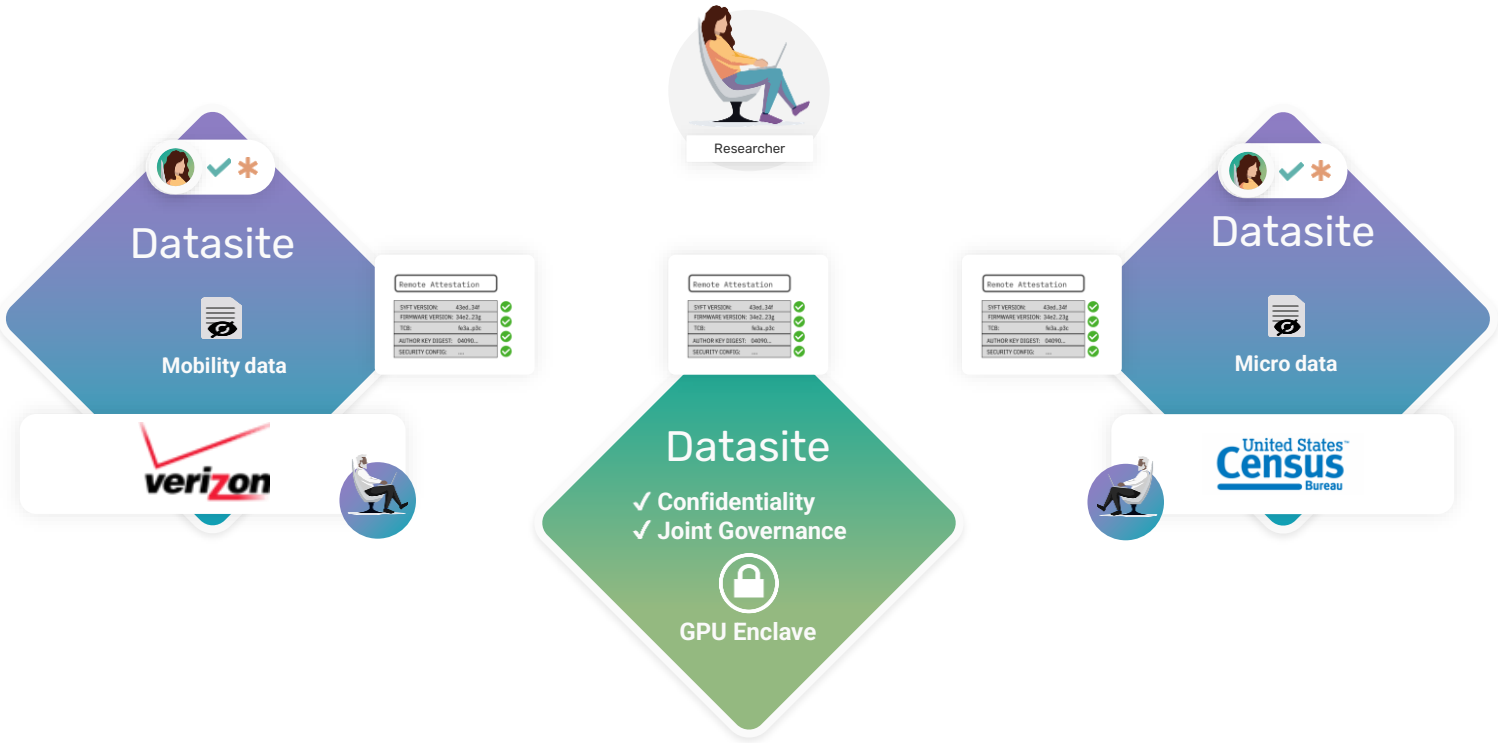
Step 3: Admins approve the code



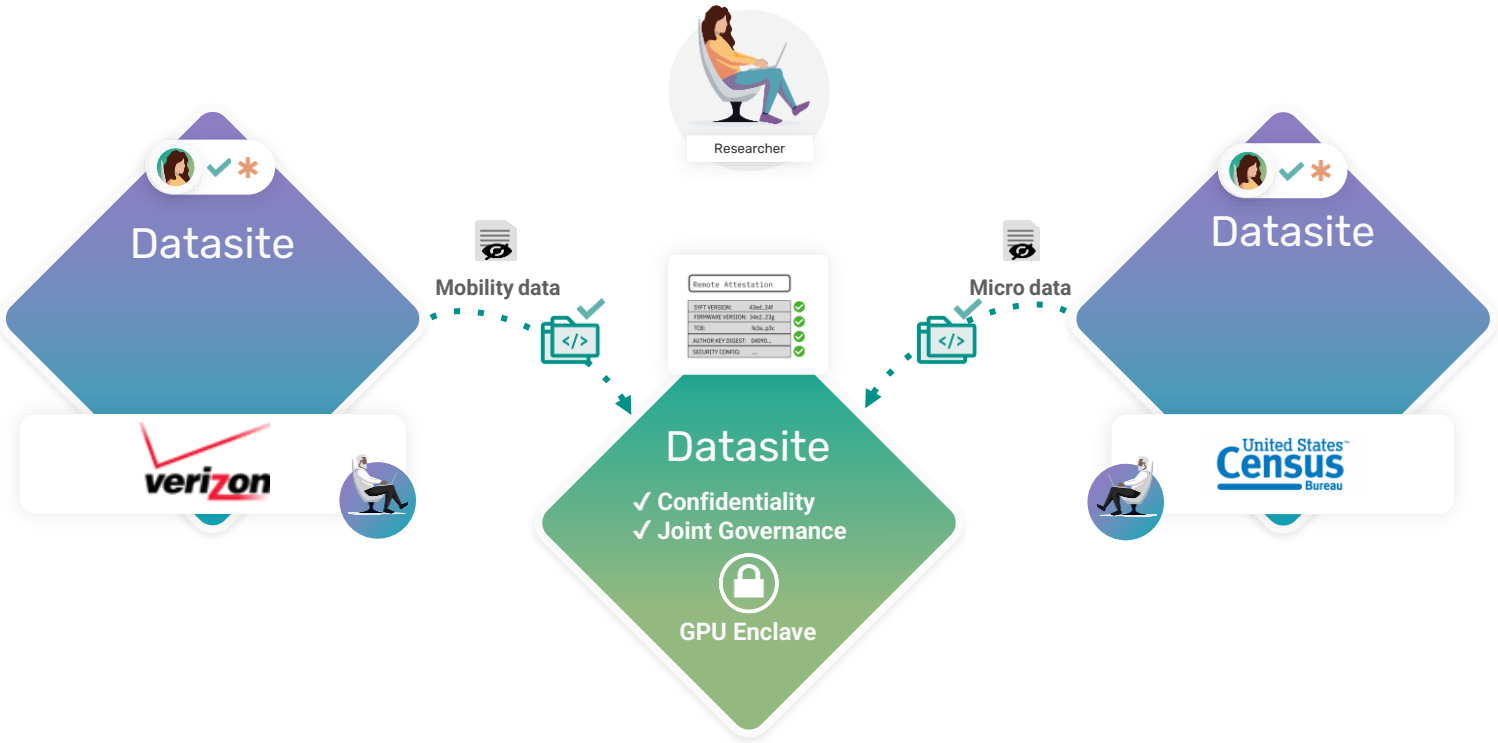
Step 4: Enclave is verified



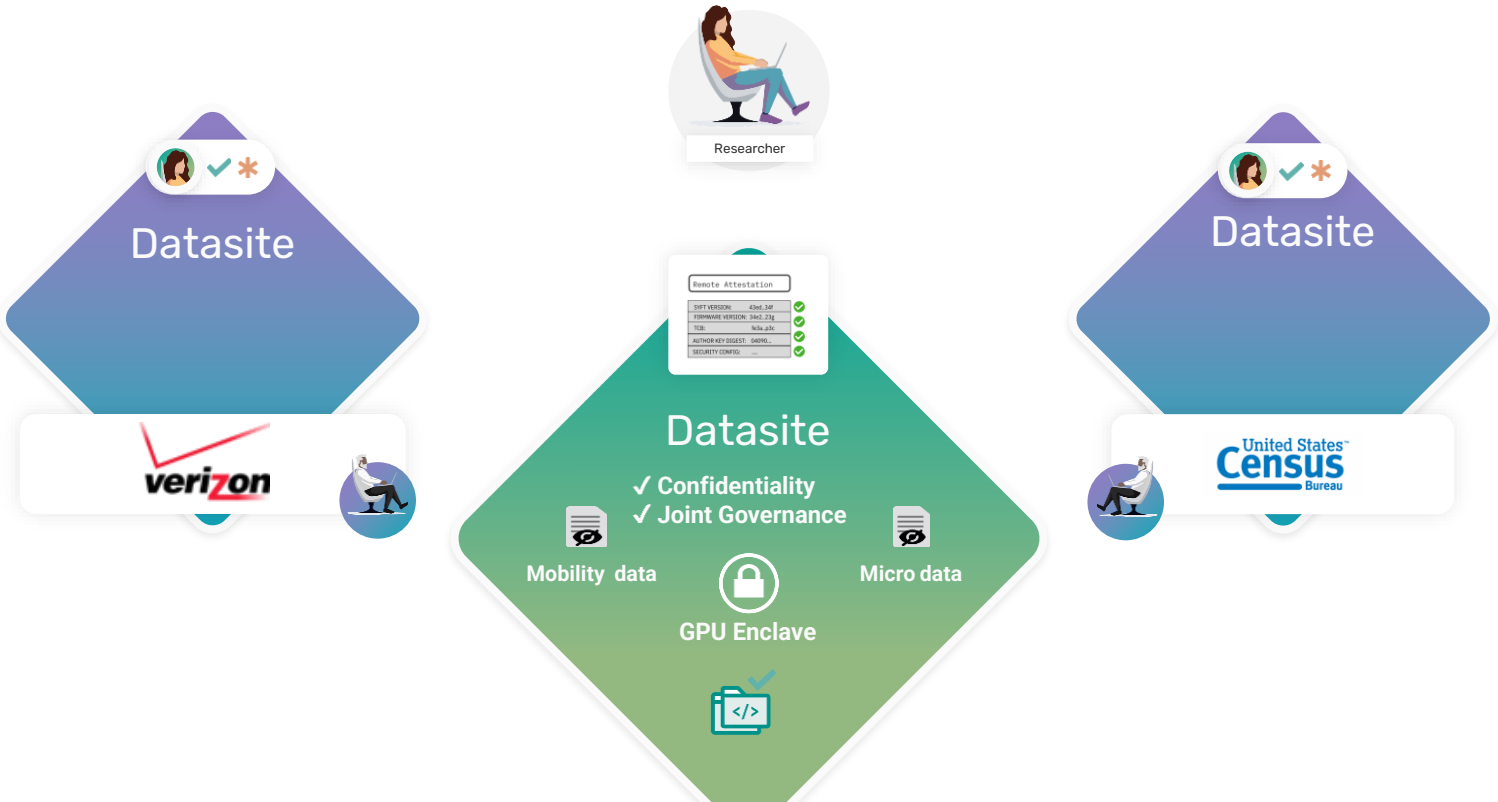
Step 4: Enclave is verified



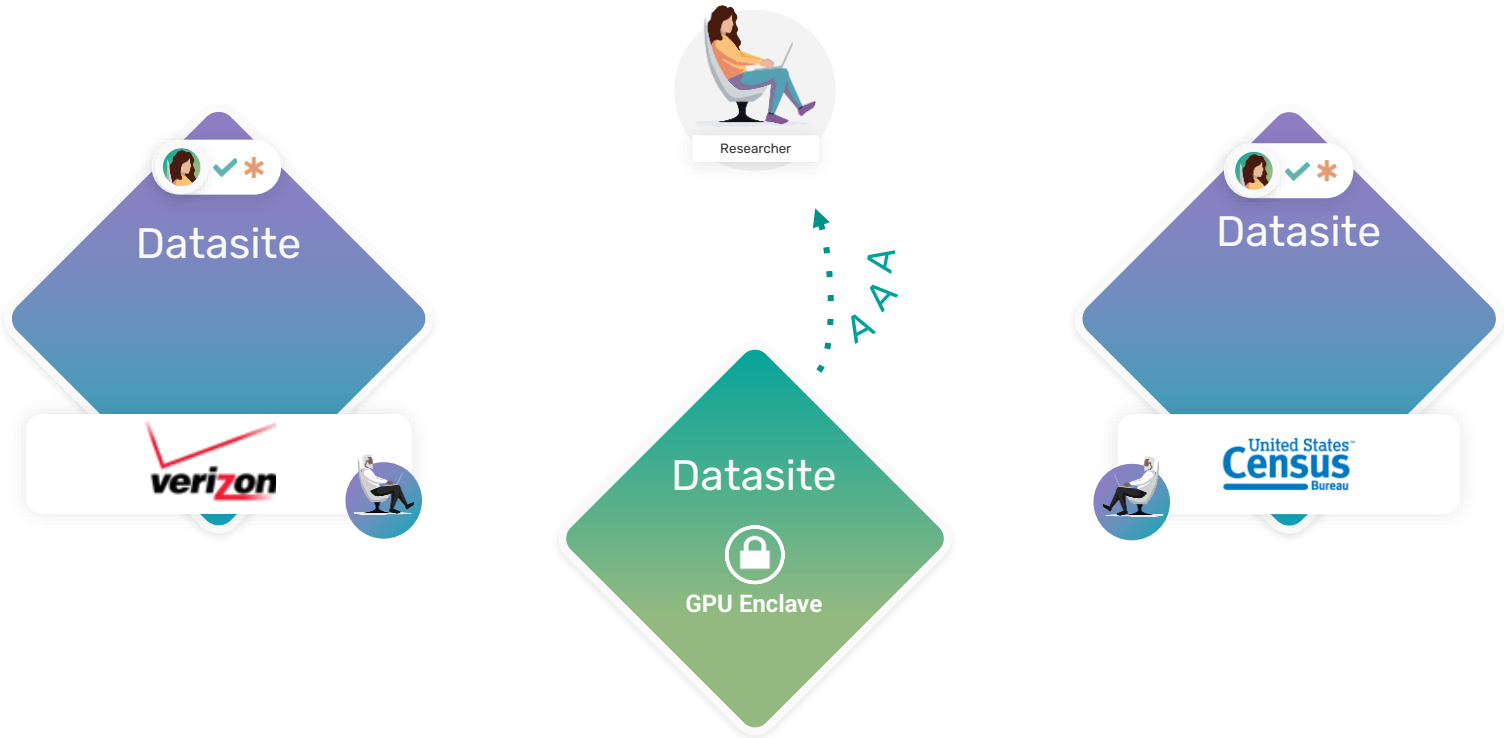
Step 5: Execution happens



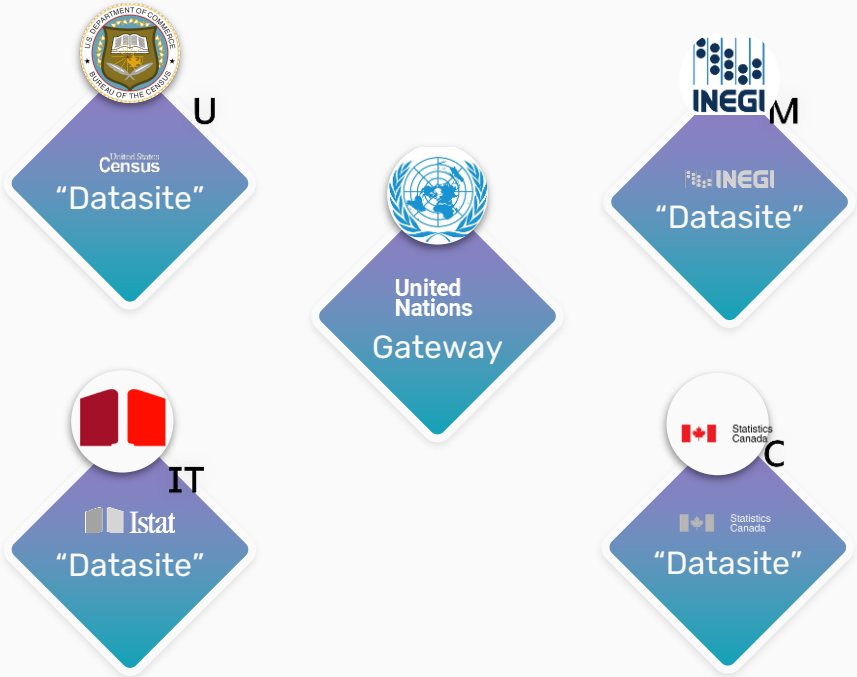
Step 5: Execution happens



Step 6: Researcher gets answers



Ongoing Project 3: UN Global Platform to enable cross-border joins on global trade



Ong o enable
cros

usenix ASSOCIATION

PEPR '24 ATTEND PROGRAM PARTICIPATE SPONSORS ABOUT

A New Model for International, Privacy-Preserving Data Science

Monday, June 03, 2024 - 4:05 pm–4:20 pm
Curtis Mitchell, *xD, US Census Bureau*

Abstract:
Currently when data analysis is performed between National Statistical Organizations (NSOs) such as the US Census Bureau and Statistics Canada, a complex series of arrangements must be agreed to that creates severe yet important restrictions on how and by whom the required data is accessed, thus increasing burden and time.

Here we demonstrate a new approach using remote, privacy-preserving processes via a collaboration between multiple NSOs in conjunction with the United Nations Privacy-Enhancing Technologies Lab (UN PET Lab). The proof-of-concept involves using the open-source data science platform PySyft and establishing the cloud infrastructure necessary such that nodes hosted by the US Census Bureau and other NSOs are facilitated by a network gateway hosted by the UN PET Lab. This architecture enables a private join on synthetic data representing realistic trade data from UN Comtrade, without each NSO needing to directly access the other NSO's data. It also enables investigations into key policy and governance questions as these technologies mature.

We believe this project will be an important milestone towards enabling privacy-preserving and remote data science between international government entities and uncovering future aspects of privacy policy and governance.

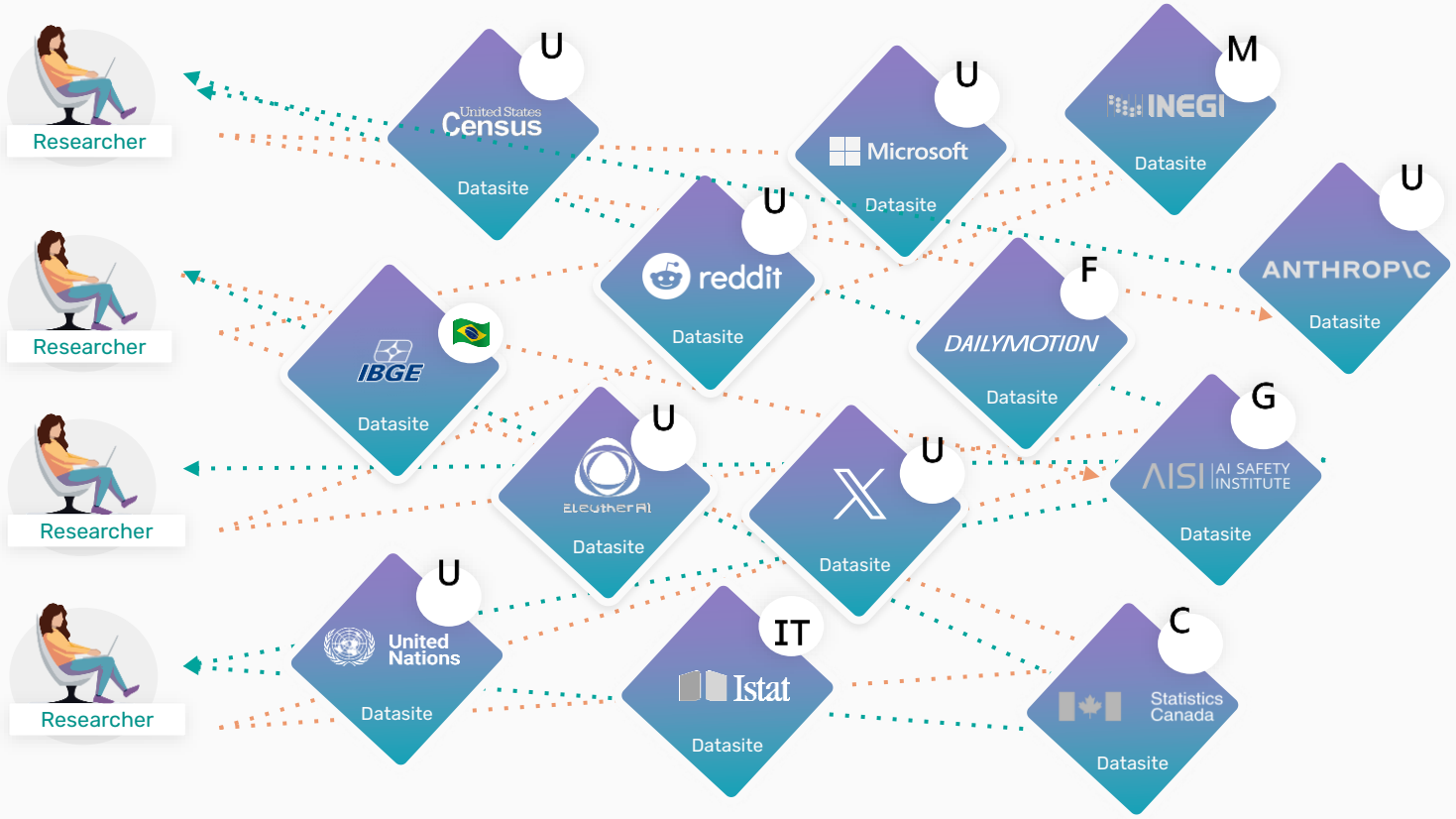
Curtis Mitchell is an Emerging Technology Fellow on the xD team at the US Census Bureau where he is contributing to a variety of projects involving privacy-enhancing technologies, responsible artificial intelligence, and modern web applications. He has over 15 years of experience in software- and data-related roles at small startups, large corporations, and open-source communities. Prior to joining the Census Bureau, he worked at NASA's Ames Research Center.

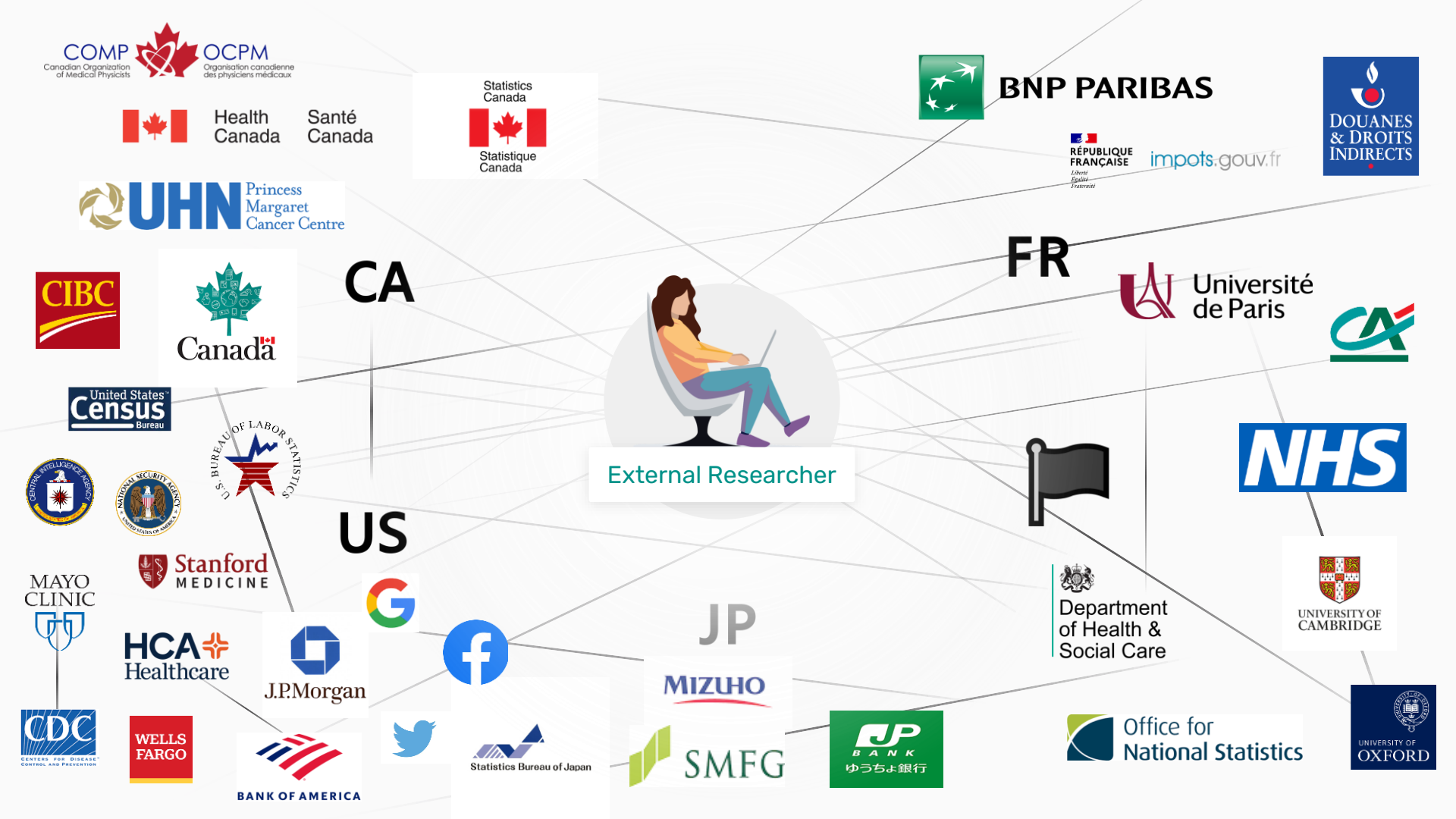
Connect: LinkedIn

BibTeX

Ongoing Project 4: enable joins between public & sector parties to study safety of frontier AI models

Goal: researchers can study any non-public data to answer important questions




COMP OCPM
Canadian Organization of Medical Physicists
Organisation canadienne des physiciens médicaux
Health Canada
Santé Canada
Statistics Canada
Statistique Canada
BNP PARIBAS
RÉPUBLIQUE FRANÇAISE
impots.gouv.fr
DOUANES & DROITS INDIRECTS
UHN Princess Margaret Cancer Centre
CIBC
Canada
CA
FR
Université de Paris
United States Census Bureau
U.S. BUREAU OF LABOR STATISTICS
External Researcher
NHS
Stanford MEDICINE
US
MAYO CLINIC
HCA Healthcare
J.P.Morgan
JP
MIZUHO
Department of Health & Social Care
UNIVERSITY OF CAMBRIDGE
CDC
WELLS FARGO
BANK OF AMERICA
Statistics Bureau of Japan
SMFG
JP BANK ゆうちょ銀行
Office for National Statistics
UNIVERSITY OF OXFORD

Thank you for your attention!

Check out our work at openmined.org

OpenMined