CHESAPEAKE LARGE-SCALE ANALYTICS CONFERENCE

Encrypted Dataset Collaboration

Using Cryptography for Privacy in Smart Cities

Isaac Potoczny-Jones: ijones@tozny.com
Erin Kenneally: erin.kenneally@hq.dhs.gov
John Ruffing: jruffing@esri.com

TOZNY
https://tozny.com

Opinions expressed are those of the authors and not their respective institutions.

Smart Cities are Collecting Data

- There is huge opportunity here to improve people's daily lives
- Equitable access, transportation, parking, traffic, air quality, safety, …
- We're fans and proponents of smart city data collection!
- But there have been some challenges around privacy

Security and Privacy go Hand-in-Hand

- *Secure* data is only accessible by authorized parties
  - If data is *private*, the user has meaningful say about who is authorized

- You can’t have privacy without good security
  - Data leaks violate the privacy of hundreds of millions every year

- A secure system can have bad privacy
  - Re-identification, 3rd party access, lack of transparency, and no accountability

TOZNY

Does Privacy Matter?

- Do you **do** anything that someone would disapprove of?
- Do you **believe** anything that someone would disagree with?
- Do you **have** anything that someone would want?
- Do you **say** anything that someone would fight against?
- **Are** you anything that someone would hate?

Yes. Privacy matters.

Doing the same thing over & over again…

- When email was first invented, it had no security
  - Everyone knew everyone else and there was no value in hacking it
  - This persisted until SPAM made email almost unusable, 25 years later
  - We've been trying to bolt security on ever since

- We make the same security mistakes for each new technology
  - **Technical**: Bad encryption, bad login security, out of date software
  - **Policy**: Too much trust between systems, bolting-on security
  - **Privacy**: No visibility, no consent, collecting more than we should

Smart Cities stands out: Innovation moving faster than privacy

Deep questions for Smart Cities

- **Ownership**: Who owns the data?
  - A legal question that can be answered with policy
- **Storage**: Who houses the data and where?
  - A practical question about the legal rules for access and security
- **Access**: Who can access the data?
  - A combination of security, access control, and legal policy
- **Subject**: Who is the data about?
  - More often than not, they don’t own it, store it, or even access it.

But the most important question:

Who *Controls* the Data?

Control is the overlap of ownership, storage, access, and subject

- Lots of modern business runs on the premise that you are the product, not the customer.
- In other words, give up your data privacy for free services
- This should not be the model for smart cities.

How can we put the right people in control?

Cities already manage data…
So what's changing?

Data Volume

- Sensors, mobile apps, and other data sources collect *a lot* of data
- At large scale, it's nearly impossible to anonymize human data
- Bad guys always want to get our private data
- And cities are bound by public records laws

TOZNY

Data Anonymization

A Primer

TOZNY

It Used To Be Simpler

Redact personally identifying information ("PII")

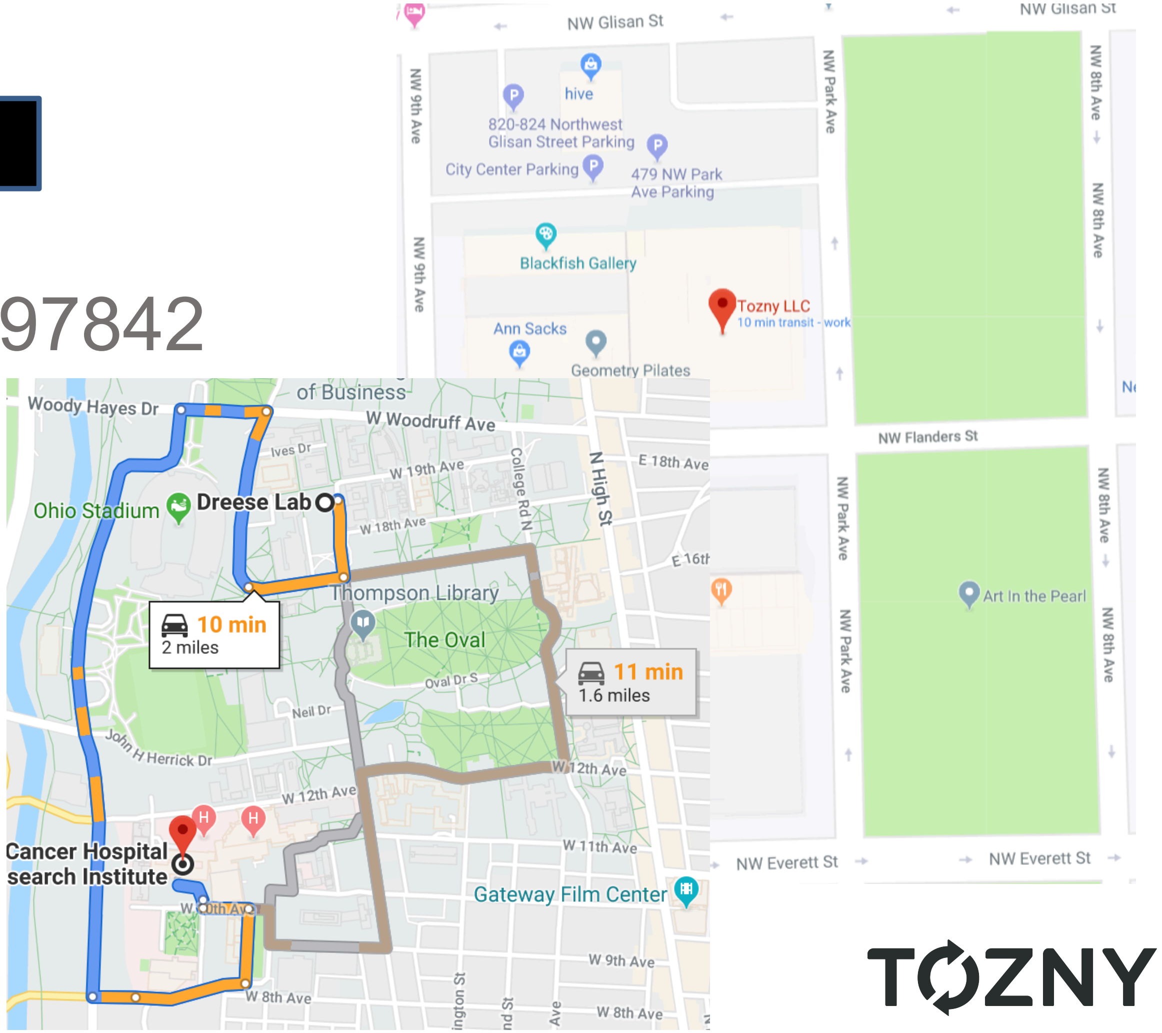
Name:	[redacted]
Purchase:	All Zone Ticket
Location:	Stop ID 3145
Date:	April 23, 2019
Time:	3:32PM

TOZNY

But What Exactly Constitutes ”PII”?

Large and complex data makes this hard!

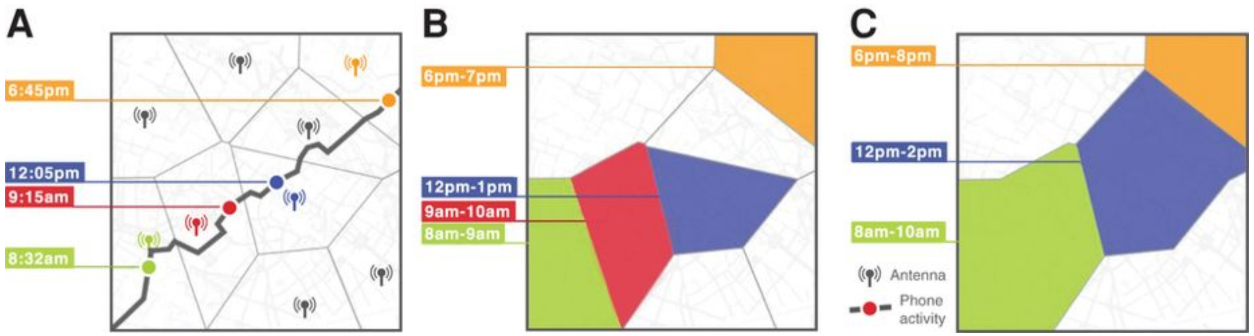
- Name: [redacted]
- Purchase: Scooter Ride
- Location: 45.5262239, -122.6797842
- Date: April 23, 2019
- Time: 3:32PM

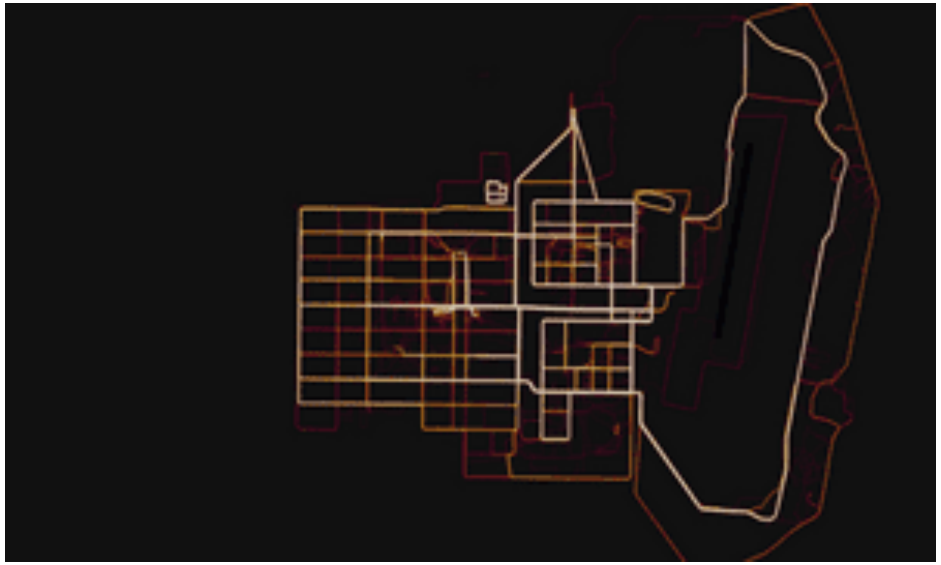


TOZNY

Anonymous Human Mobility Data?

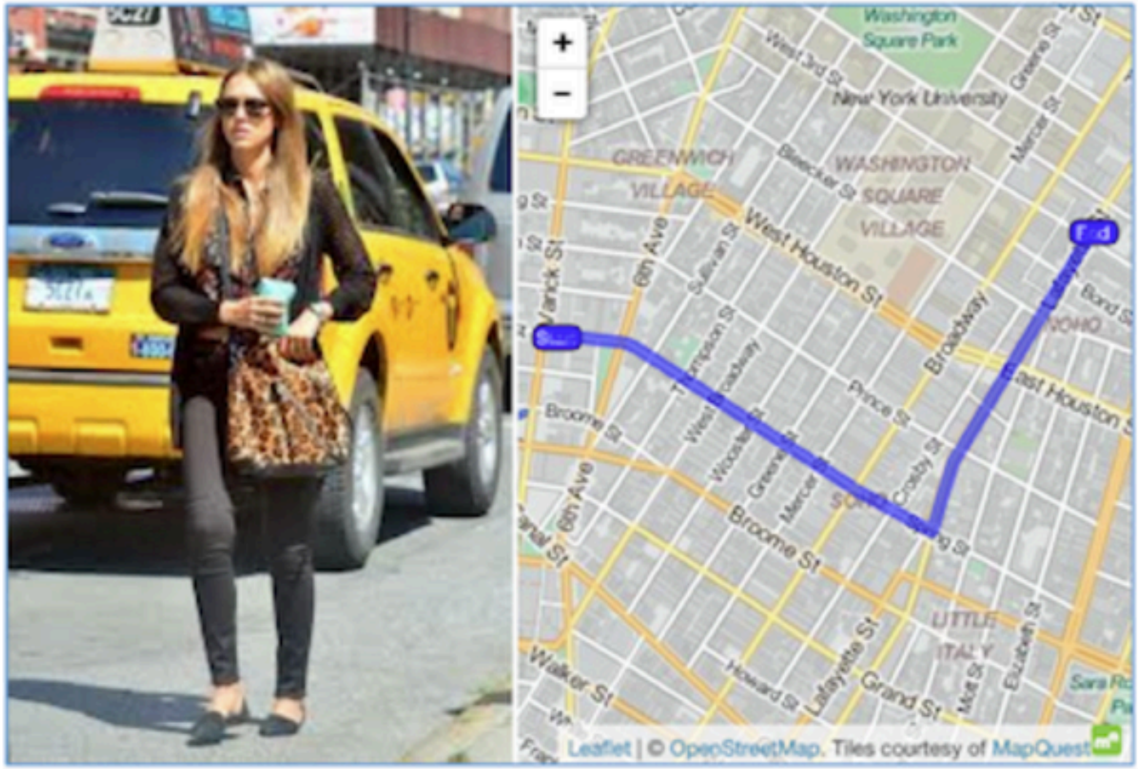
It’s not clear that it’s even possible to anonymize at scale





Is there any such thing as de-identification?

- NYC Taxi cab open dataset was combined with other info
- Dataset included: pick up and drop off locations and times
- Researchers correlated this with photos of celebs getting into taxis
  - Figured out their drop-off locations
  - Their fare amounts
  - And whether they tipped

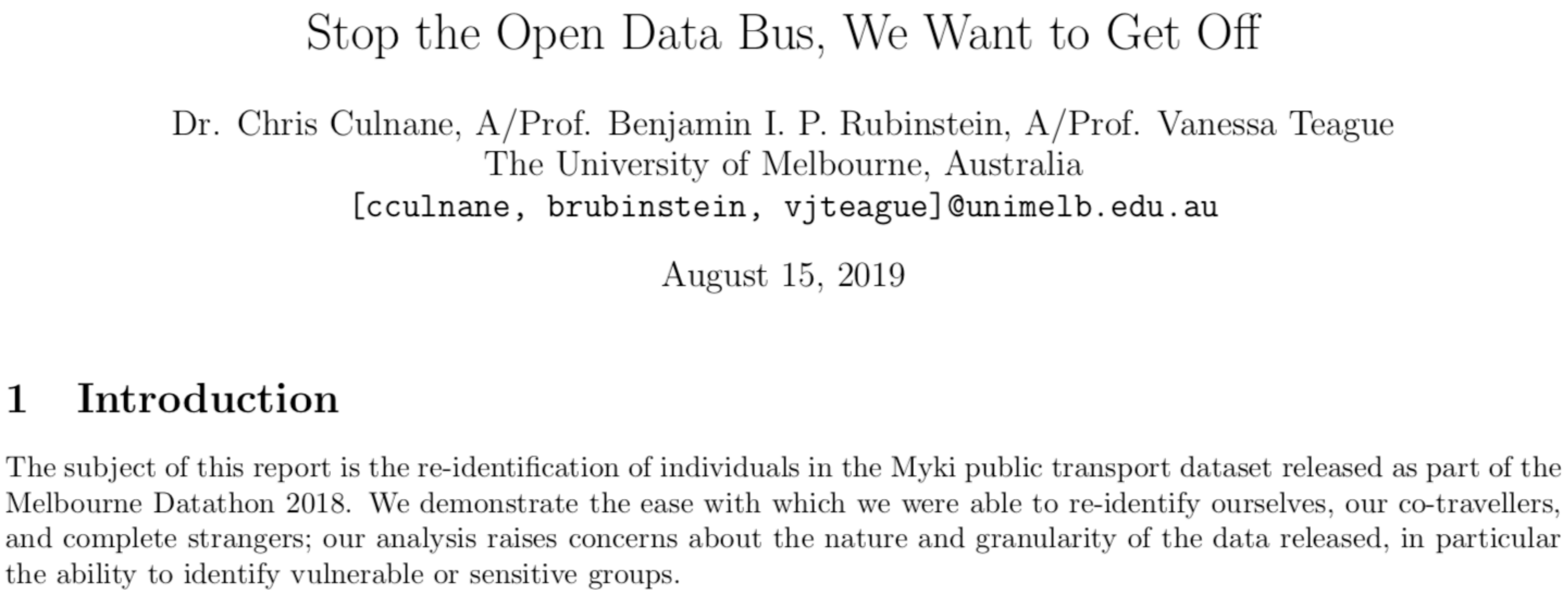


Lessons: Datasets can be correlated with other info to re-identify users

Very hard to predict what's identifying

And we’re still making the same mistake

- Very recent example of “anonymized” public transit data
- Provided by the city to to “hackathon” style event
- Included 3 years of data, 15M people, on an open S3 bucket
- Can identify strangers, co-riders, and MPs based on Twitter



Stop the Open Data Bus, We Want to Get Off

Dr. Chris Culnane, A/Prof. Benjamin I. P. Rubinstein, A/Prof. Vanessa Teague
The University of Melbourne, Australia
[cculnane, brubinstein, vjteague]@unimelb.edu.au

August 15, 2019

1 Introduction

The subject of this report is the re-identification of individuals in the Myki public transport dataset released as part of the Melbourne Datathon 2018. We demonstrate the ease with which we were able to re-identify ourselves, our co-travellers, and complete strangers; our analysis raises concerns about the nature and granularity of the data released, in particular the ability to identify vulnerable or sensitive groups.

You Probably Can’t Anonymize

That Large Data Set*

*Even if you think you can

TOZNY

Is Re-Identification a Problem for People?

Yes. Definitely. Sometimes.

- No one can predict when re-identification will be a problem
  - It’s very personal: Traveling for health treatment? Abusive partner or stranger? Skipping work? Going to a bar?
- No one can predict when other datasets will provide correlation
  - Datasets don’t live in isolation
- Advanced statistics can help, but require advanced expertise
  - Differential privacy would change the way we manage and analyze data

Open Records Laws

Require Release of Data!

The Conundrum of FOIA and Similar Laws

- Governments bound by Freedom of Information Act and similar laws
- Government information is basically in the public domain
- Reporters, concerned citizens, and malicious people can ask for data
- Smart Cities adds terabytes of high-fidelity data to this mix
- Governments are typically required to “redact” private information
- But we just talked about how that’s almost impossible

Cities Address This in Various Ways

- Don’t collect data: But we lose its benefit
- Don’t release the data: But public records laws might require it
- Give it to 3rd parties: They might not respect user privacy
- Differential privacy: Probably too advanced at this point
- Data Trust: A policy and legal framework to govern data…

Policy Approach: Data Trust

- Form a legal entity that stewards the data
- Accountable for its proper access and use
- Address and balance potentially competing concerns
  - Use of data in the public interest
  - Public access to data without violating privacy
  - Access to privately-generated data (e.g. mobility companies)
- A relatively new approach, hasn’t been battle tested yet

You Might Be Required To Release

That Large Data Set*

*Even if you think you shouldn’t

Pilot: Portland Oregon

User Data Wallet

TOZNY

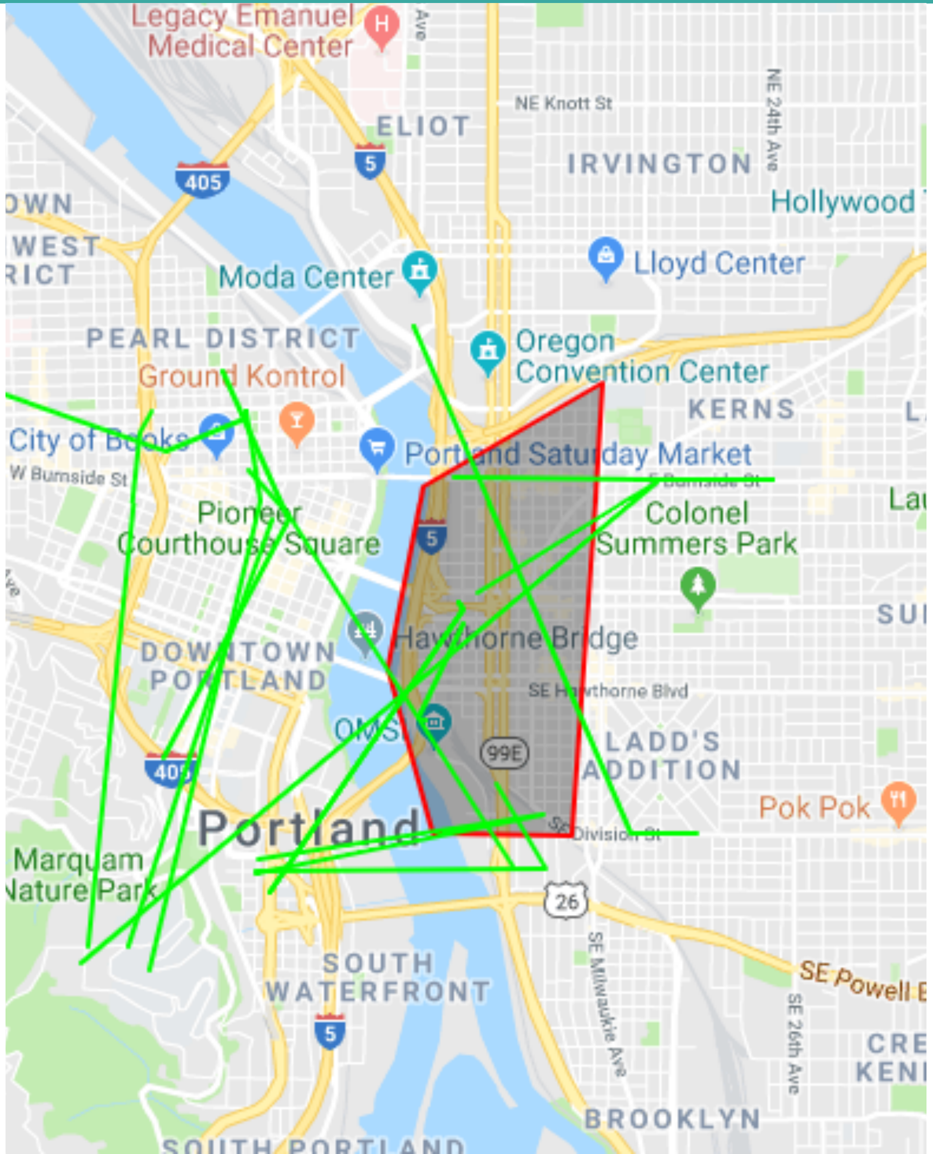
Pilot Partnership Goals

- Collaboration between Tozny and DHS
- To pilot privacy-preserving technical solutions
- Demonstrate a technical capability
- Use this as a model for Smart City privacy in other cities
- Pilot multiple use cases to demonstrate wide applicability

User Data Wallet

- An app, website, and API for users and cities to collaborate
- Controlled, privacy-preserving sharing of user data
- Users can put data in to share with the city
- Cities can put data in to share with the users
- Implemented with end-to-end encryption
- Significantly increases the security and privacy of the data

Use Case: Parking and Transit



- Dense parking district that needs to study how people get around

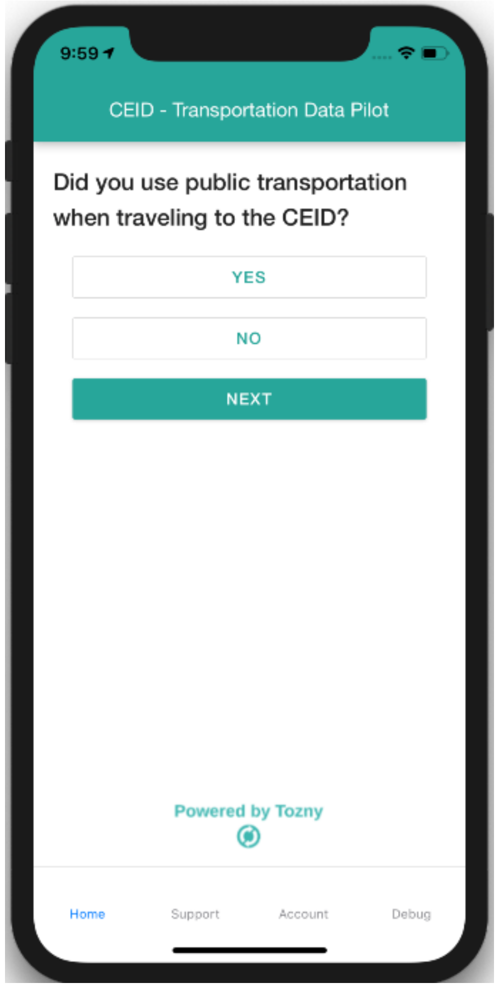
- Want to incentivize efficient transportation and parking

- Created a privacy-preserving app that collects location data

- If you start a trip, end a trip, or go through the area, we collect start/end GPS

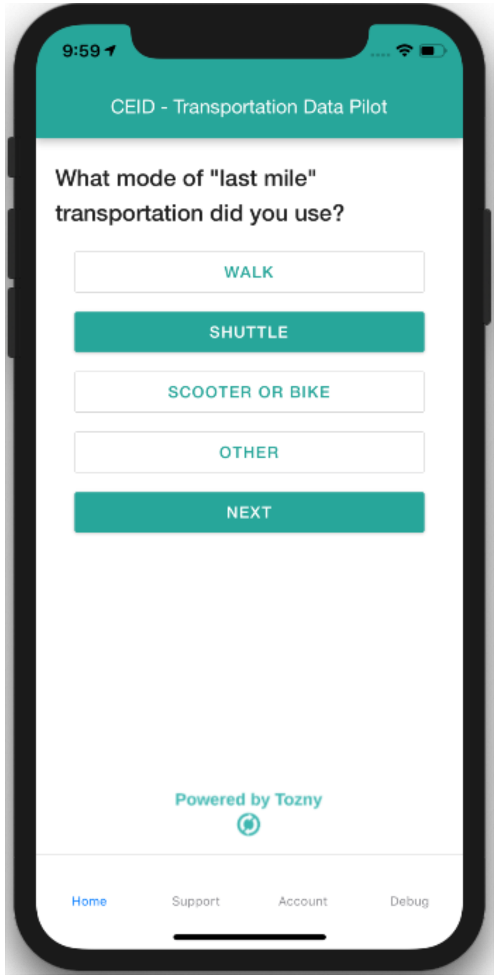
TOZNY

User Surveys



CEID - Transportation Data Pilot

Did you use public transportation when traveling to the CEID?

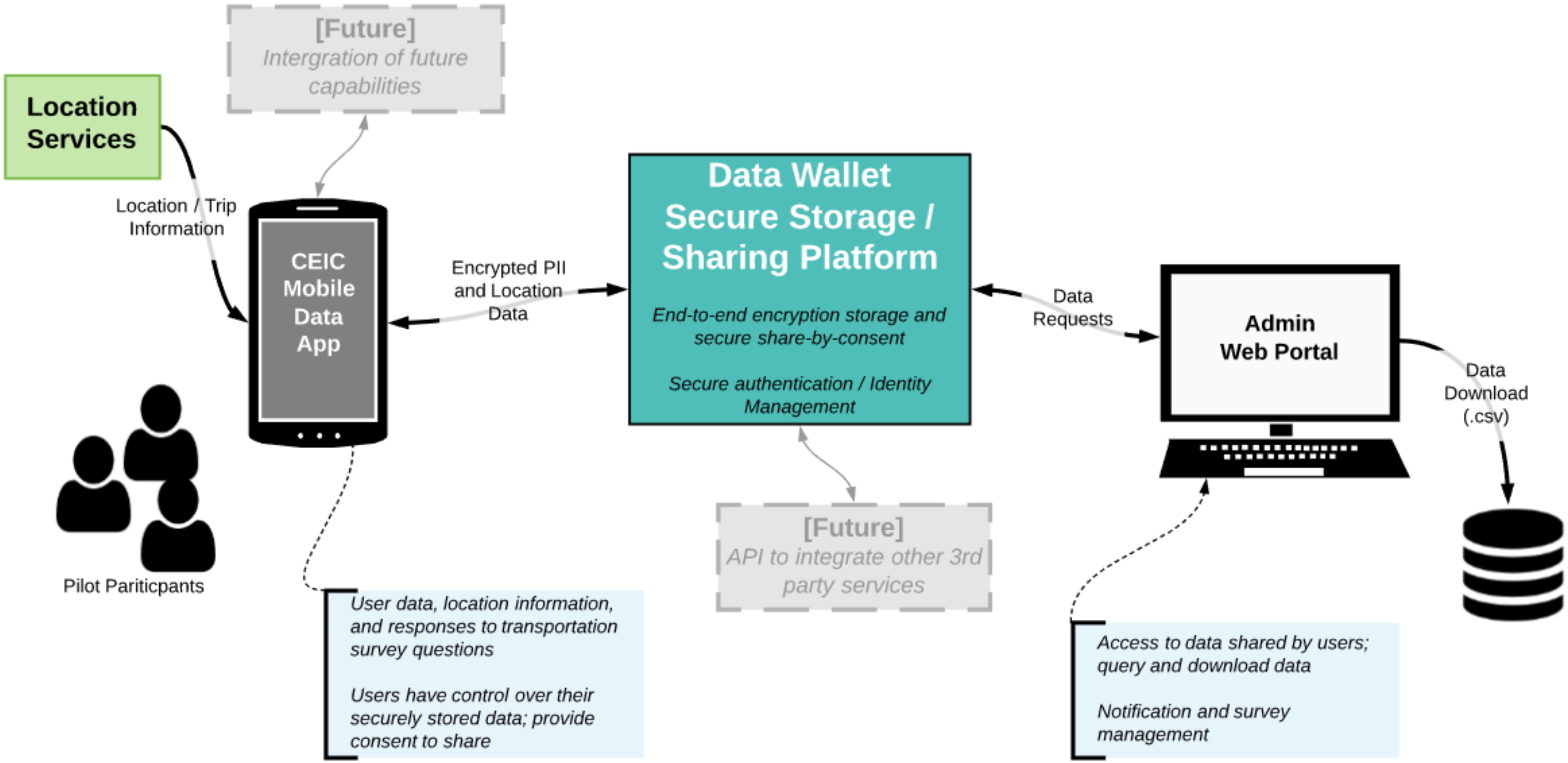


CEID - Transportation Data Pilot

What mode of "last mile" transportation did you use?

TOZNY

Pilot Architecture



[Future]
Intergration of future capabilities

Location Services

Location / Trip Information

CEIC Mobile Data App

Encrypted PII and Location Data

Data Wallet Secure Storage / Sharing Platform

End-to-end encryption storage and secure share-by-consent

Secure authentication / Identity Management

Data Requests

Admin Web Portal

Data Download (.csv)

Pilot Paritícpants

User data, location information, and responses to transportation survey questions

Users have control over their securely stored data; provide consent to share

[Future]
API to integrate other 3rd party services

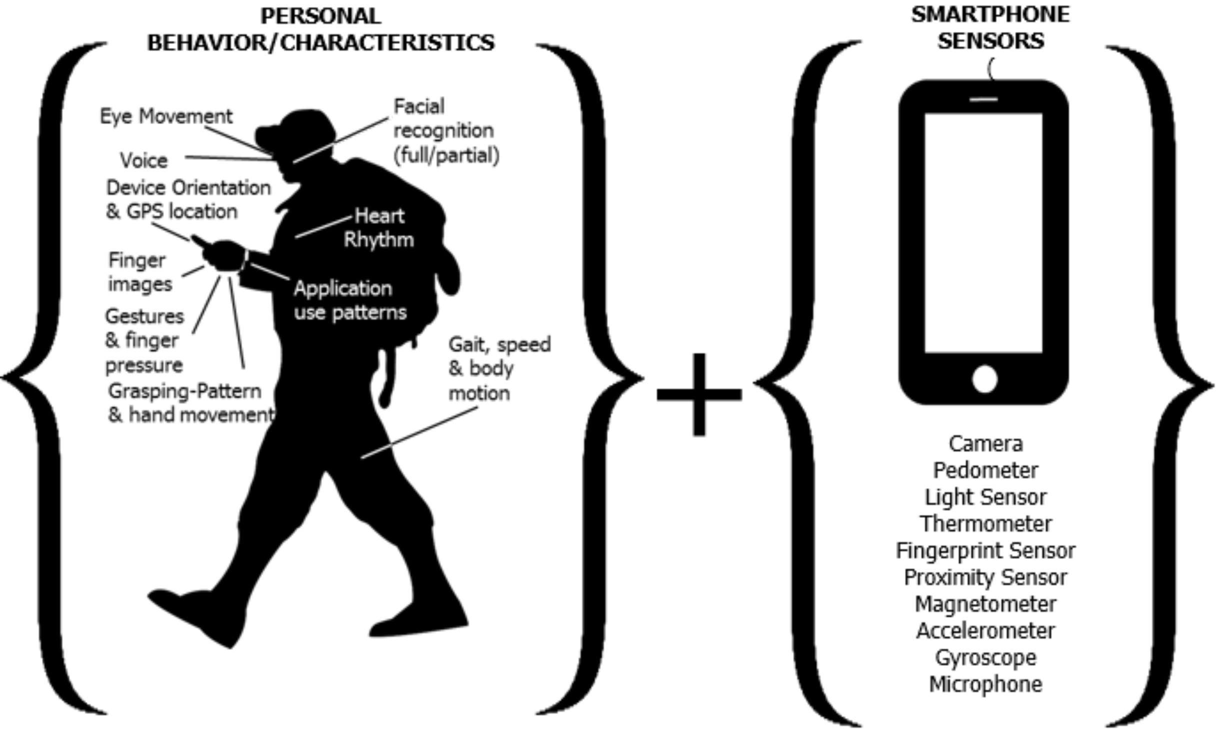
Access to data shared by users; query and download data

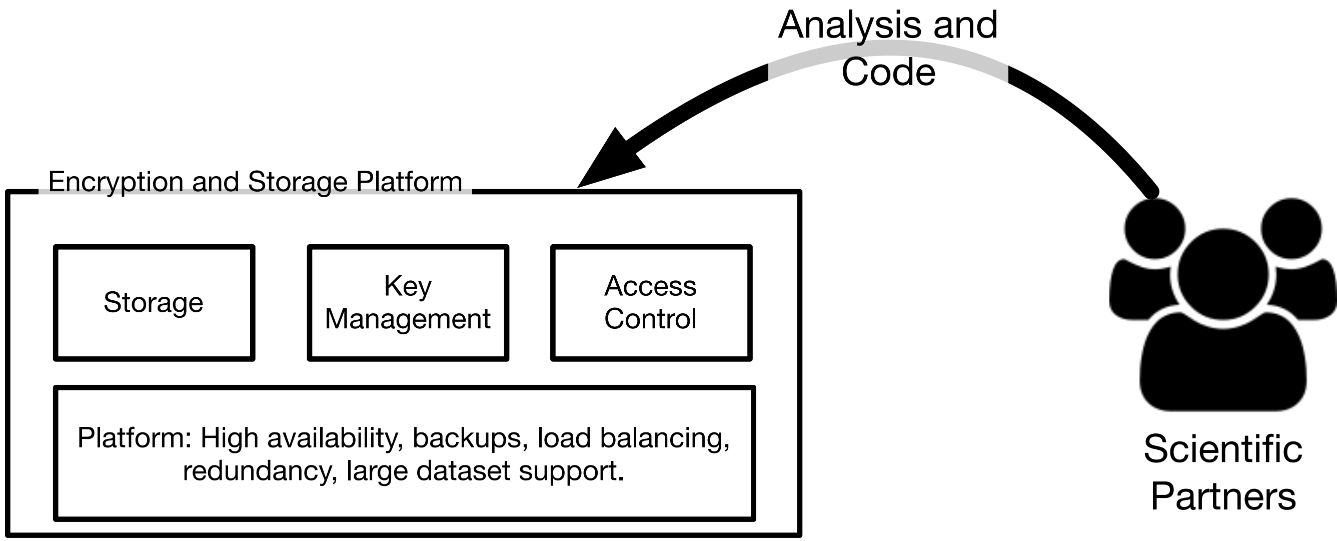
Notification and survey management

TOZNY

Use Case: Human Subject Research

- Current Study IRB / HIPAA / CUI
- We’re the security, privacy, data management team




Analysis and Code

Encryption and Storage Platform

Storage

Key Management

Access Control

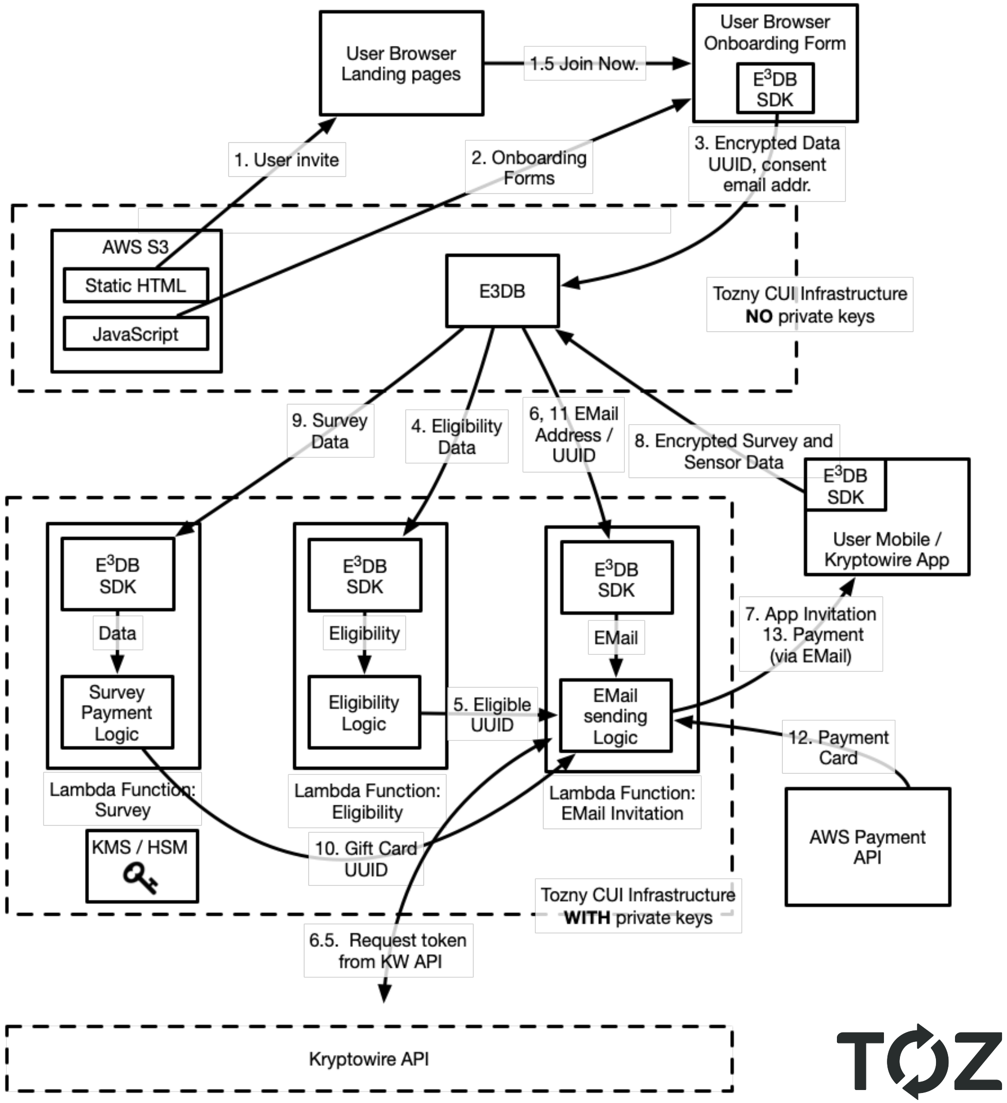
Platform: High availability, backups, load balancing, redundancy, large dataset support.

Scientific Partners

TOZNY

Use Case: Human Subject Research

- New Approach to Human Use Data Collection
  - No human access of any personally identifying information (PII)
  - Only access is to anonymous random unique user identifier (UUID) associated with app
- Completely Anonymized Communication Protocol
  - Anonymous support
  - Anonymous payment

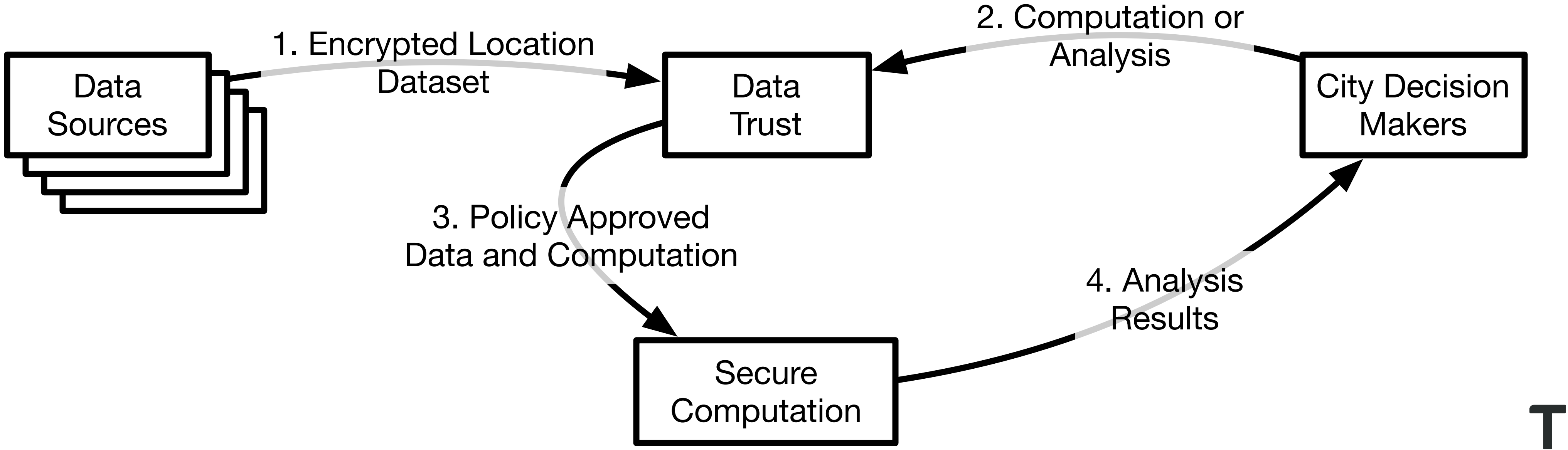


TOZNY

Using Encryption for Privacy, not just Security

Not just about security

- Leverages key management to say who controls this data
- No matter where it’s stored, who owns it, or who it’s about
- Secure Computation to enforce privacy throughout data lifecycle


1. Encrypted Location Dataset
Data Sources
Data Trust
2. Computation or Analysis
City Decision Makers
3. Policy Approved Data and Computation
4. Analysis Results
Secure Computation

TOZNY

End-to-End Encryption

- You’ve probably heard about end-to-end encryption in the news
- Apple and several others are implementing it as a best practice
- It maintains encrypted control of the data for its entire lifecycle
- It’s more secure than standard approaches to encryption
- But it’s typically more challenging to implement

This platform and pilot works to make encryption easy for cities

TOZNY

Benefits and Residual Risk

- Benefits: Allow use of data with significantly reduced privacy risk
  - Exclusion of PII from data
  - Data cannot be used without consent
  - Mitigates unintentional or accidental data leaks
  - Mitigates compromise of data trust through encryption

- Residual risk is minimal
  - Trojan computations: Mitigate with inspection, differential privacy
  - Compromised secure computation

Status and Next Steps

- The platform has already been developed for DHS, DARPA, and NIST
- It’s robust and deployed in production
- Tozny and DHS are working with the City of Portland and others
- A few use cases have been identified
- A transportation-related pilot is planned for early fall
- We are open to engaging other cities in pilots!

TOZNY

Thank You!

Isaac Potoczny-Jones: ijones@tozny.com
Erin Kenneally: erin.kenneally@hq.dhs.gov
John Ruffing: jruffing@esri.com

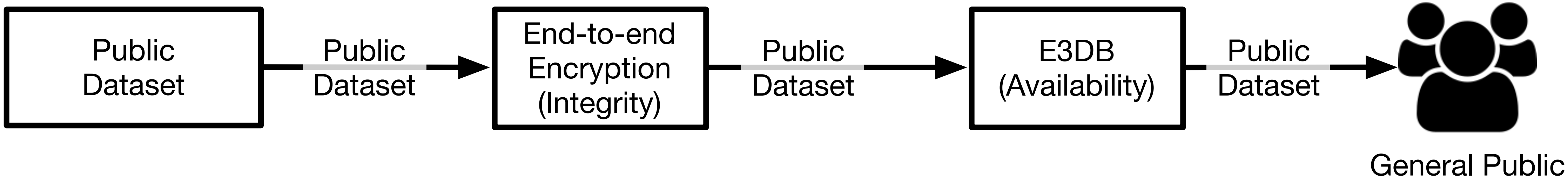
TOZNY

Backup

Slides

TOZNY

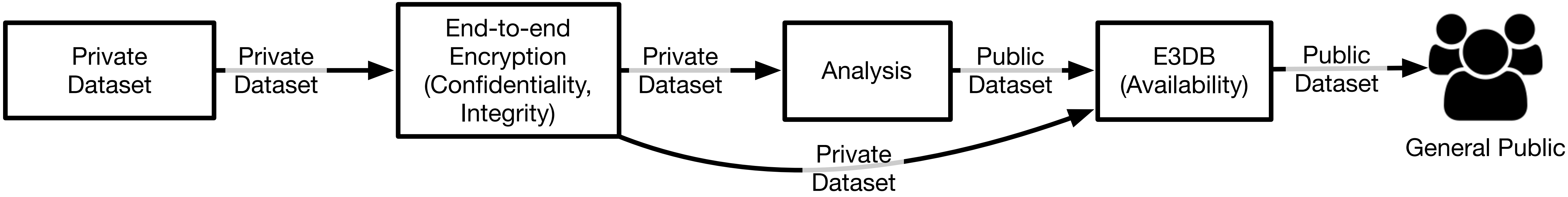
Public Datasets: Control who can change



- Provide integrity and availability
- Easy to access, general purpose API
- But smart city datasets are about more than just public data

TOZNY

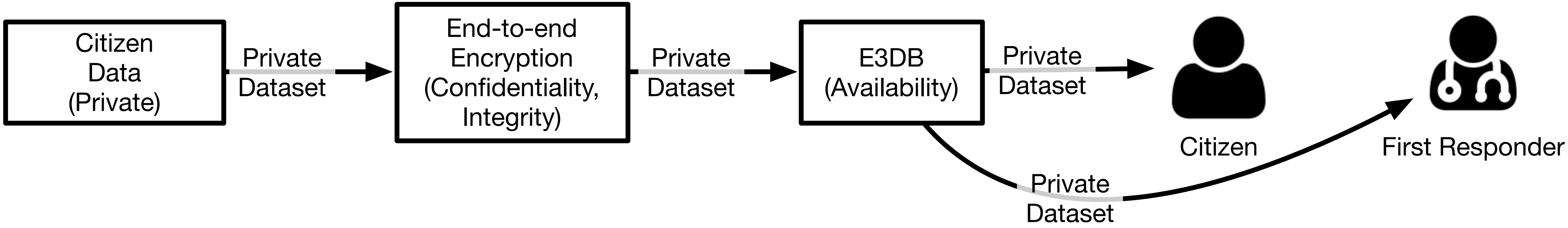
Extracting public data from private data



- Provide security for private data
- Allow privacy-preserving transformations
- Provide integrity and availability to public data

TOZNY

Private Datasets: Control who can access



- Provide confidentiality for private data
- Put citizens in control

TOZNY