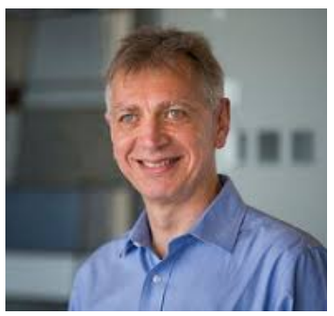
Constraint-based Differential Privacy for Mobility Services

Ferdinando Fioretto

Chansoo Lee

Pascal Van Hentenryck



University of Michigan

AAMAS 2018

Facebook, Cambridge Analytica face lawsuit over privacy loss

Maryland resident Lauren Price has sued over the handling of her personal information. The case could become a class action.

BY **STEPHEN SHANKLAND** / MARCH 22, 2018 12:37 PM PDT

An eight-year Facebook user, Lauren Price, has sued Facebook and Cambridge Analytica over the firm's gathering of private data about more than 50 million people through the social network.

The suit, filed Tuesday in US District Court in San Jose, alleges the companies violated California's unfair competition law. It seeks damages

Facebook CEO Mark Zuckerberg called the Cambridge Analytica data scandal a " a major breach of trust" Wednesday during an interview with CNN.

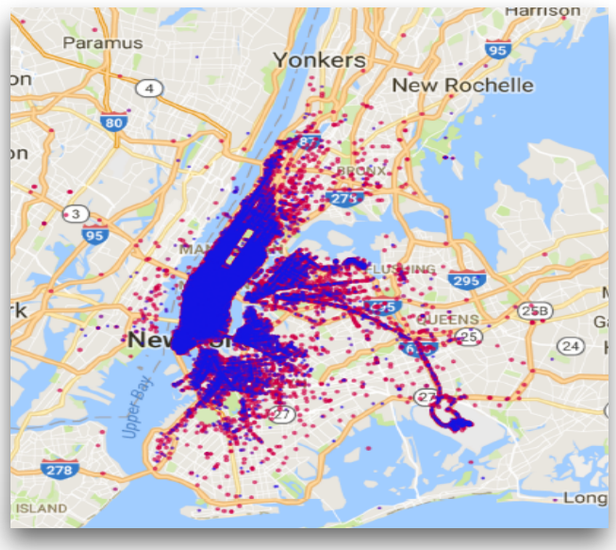
CNN

Aggregated Personal Data is Invaluable

Online advertising

Medical applications

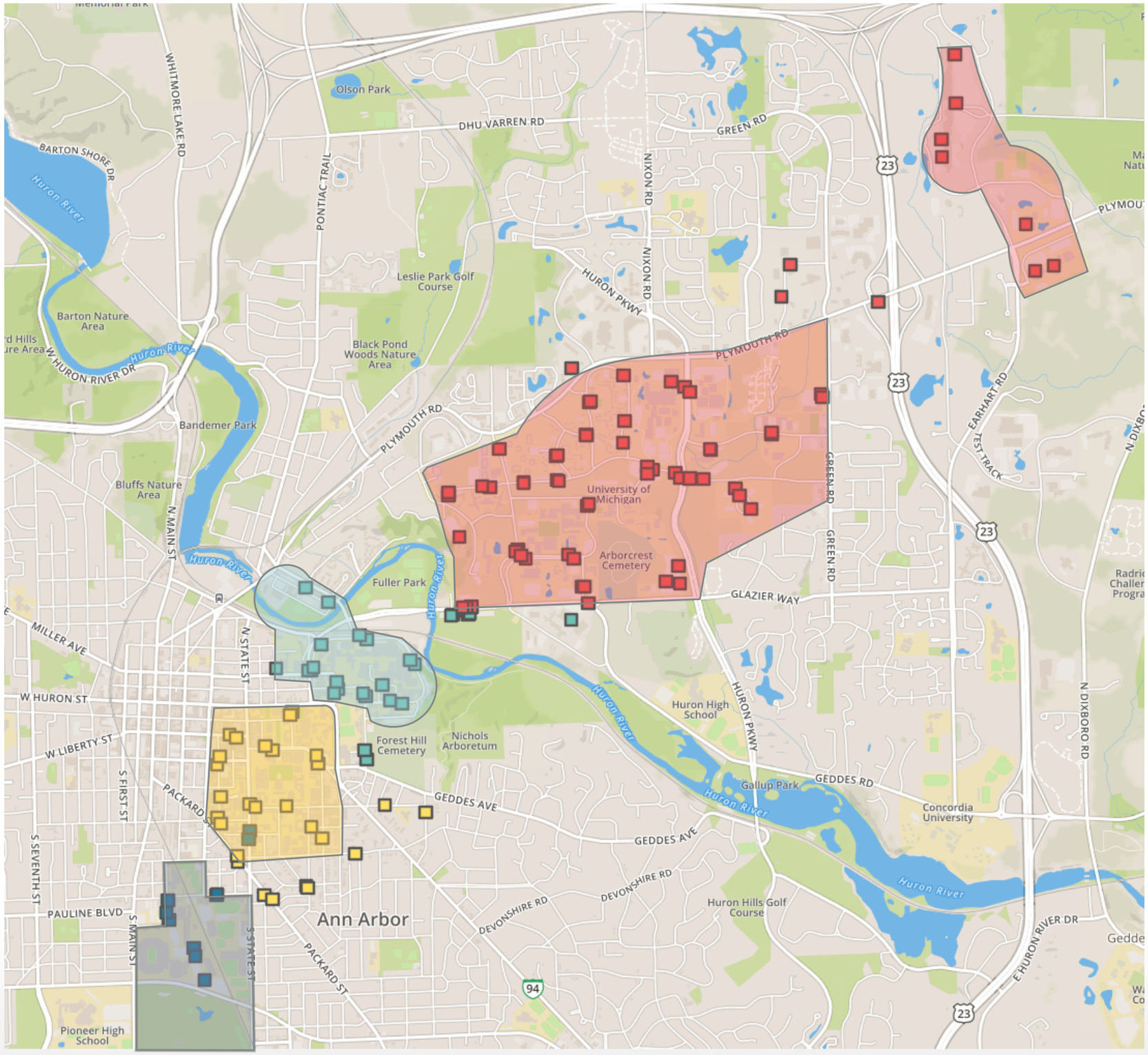
Mobility analysis



Source: https://rcc.uchicago.edu

On-Demand Multi-modal Transportation System

Ann Arbor



On-Demand Multi-modal Transportation System

Ann Arbor

- High ridership: ~40K trips daily
- Low ridership in isolated regions

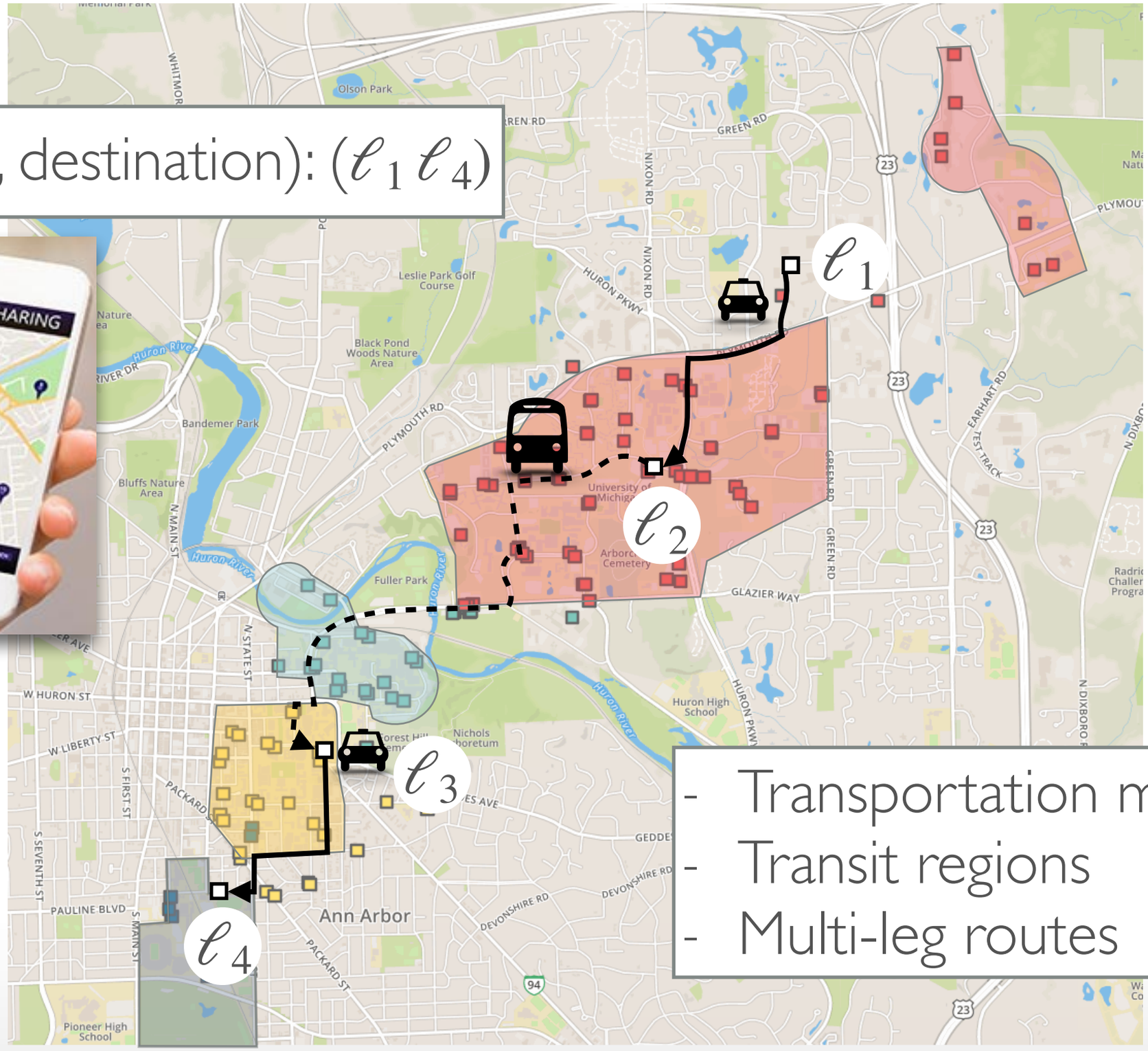
Combine Buses and Shuttles

- shorter, high-frequency bus routes
- shuttles to serve first and last miles
- on-demand requests

On-Demand Multi-modal Transportation System

Optimizing a Multi-leg Trip: Features

(origin, destination): $(\ell_1\, \ell_4)$



- Transportation modes
- Transit regions
- Multi-leg routes

The problem

A Procurement/Competition Setting

On-Demand Multimodal System Design

Transit Agency

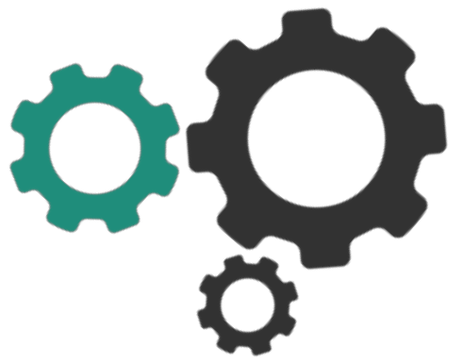
$M(D)$

$\tilde{D}$

training

Solution Provider

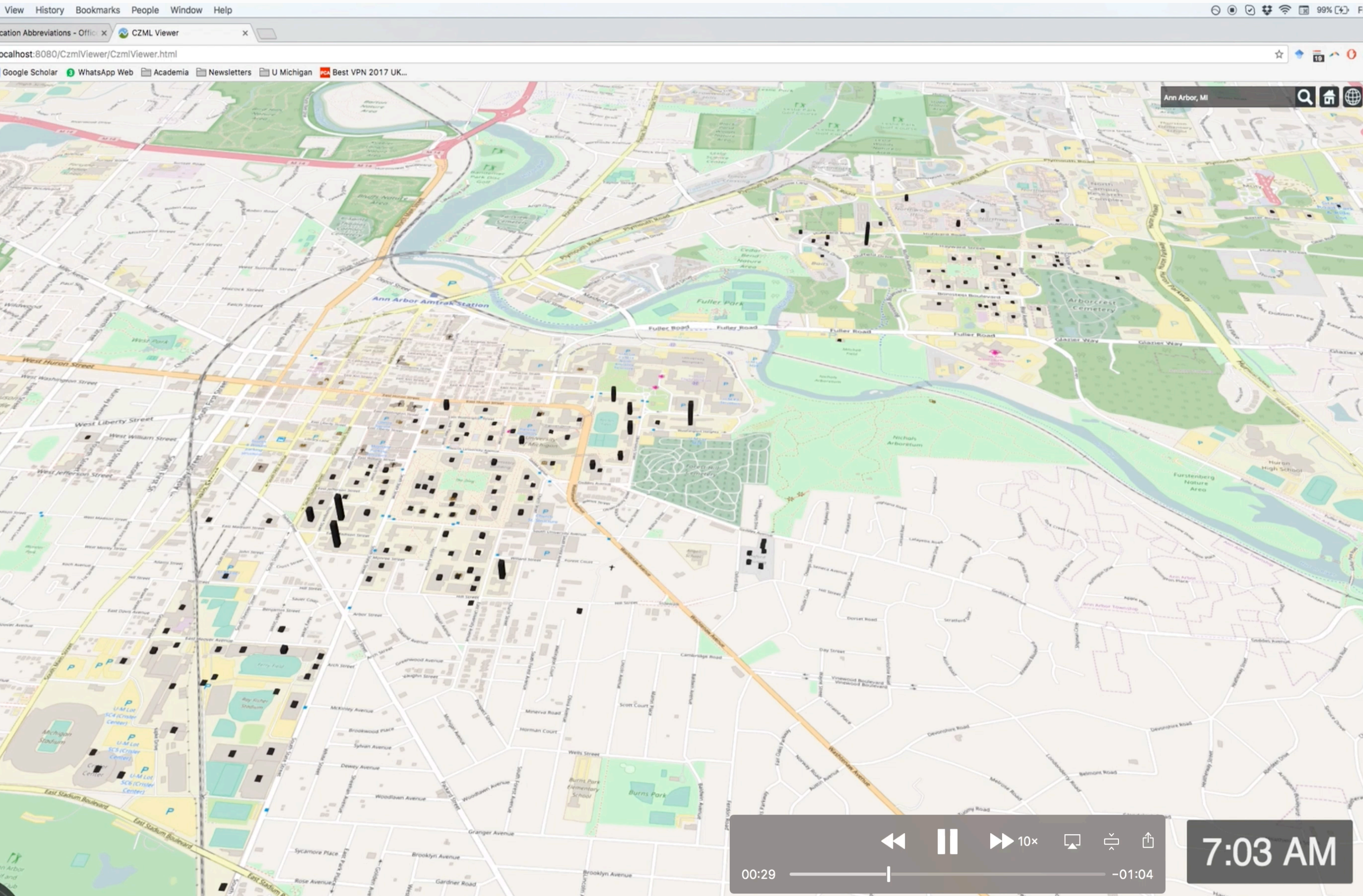
$\pi(\tilde{D})$



model

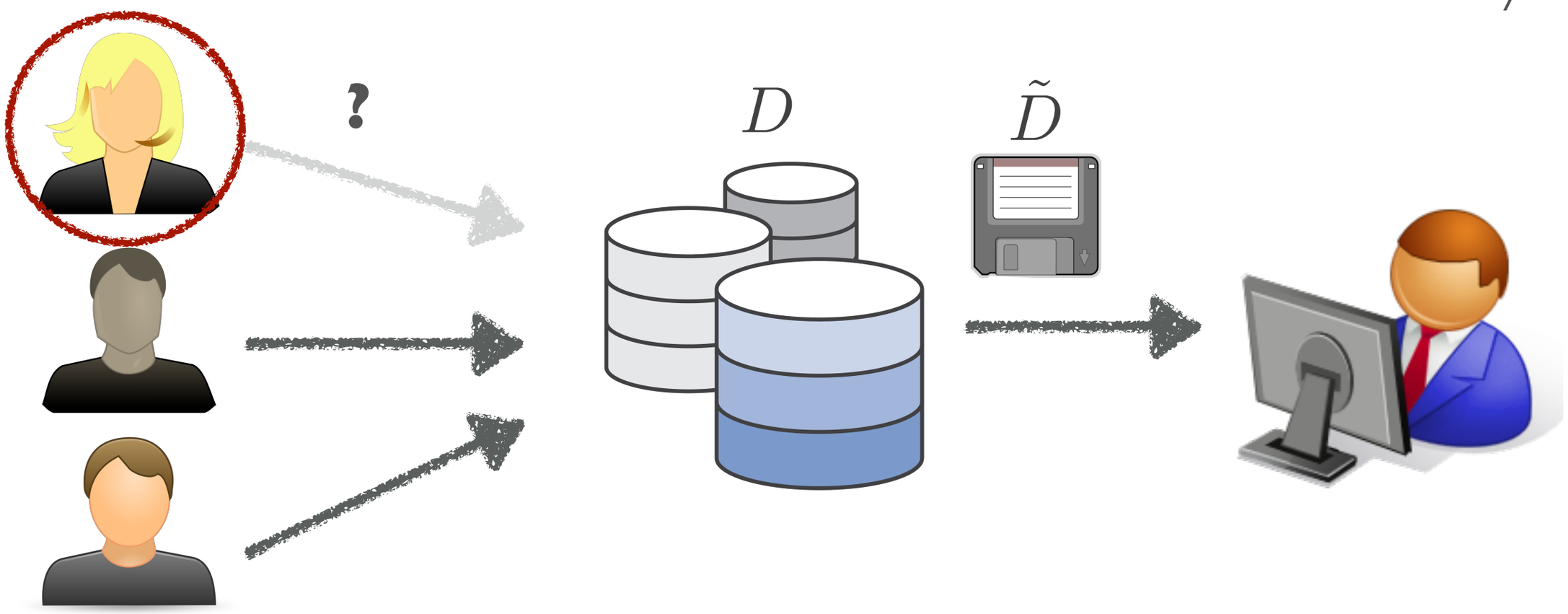
(schedule)

Data Collection and Dataset



Differential Privacy

(Informal)

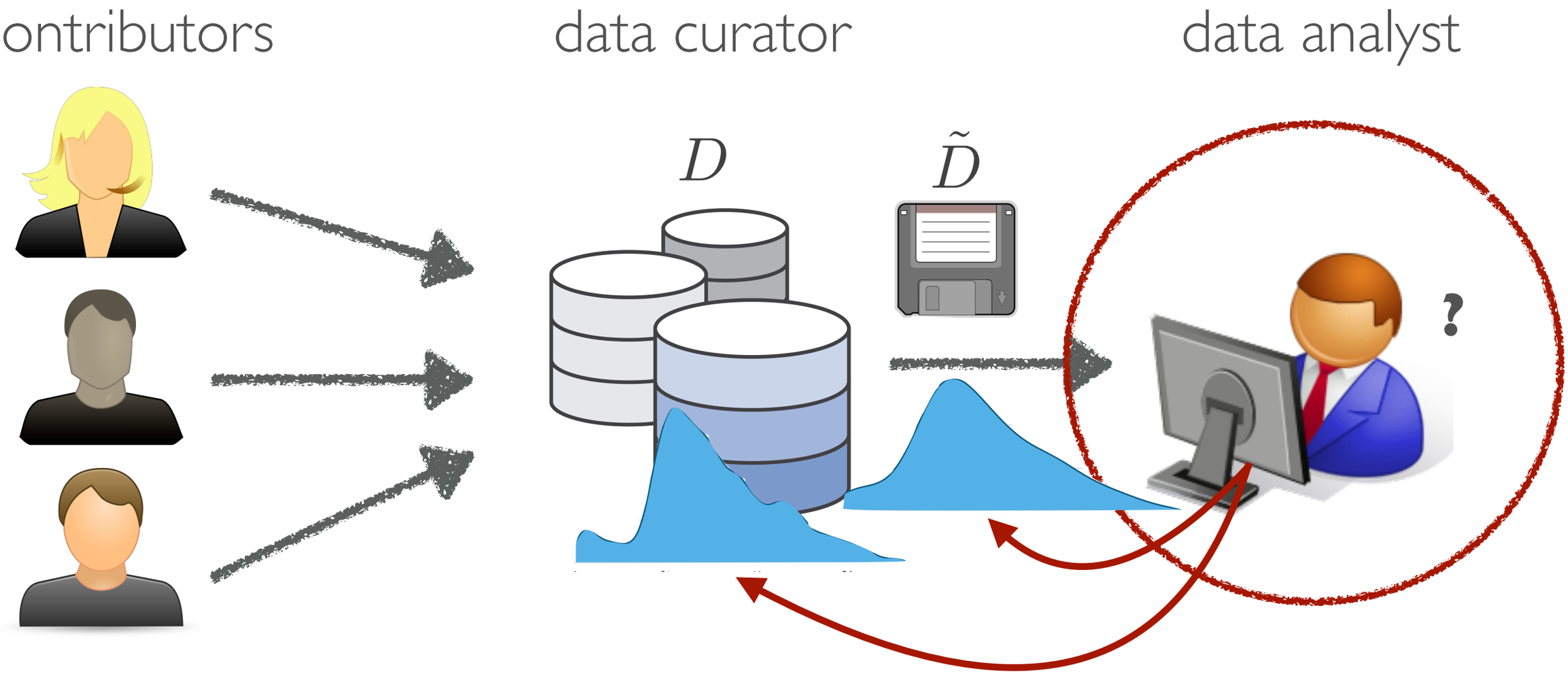


Contributor: Small participation risk (privacy loss)

Data analyst: Analysis on original and modified data are very similar (data distributions)

Differential Privacy

(Informal)



Contributor: Small participation risk (privacy loss)

Data analyst: Analysis on original and modified data are very similar (data distributions)

Differential Privacy

[Dwork:06]

- *Two datasets D_1, D_2 are said neighbors ($D_1 \sim D_2$) if they differ by at most one tuple $\|D_1 - D_2\|_1 \leq 1$*

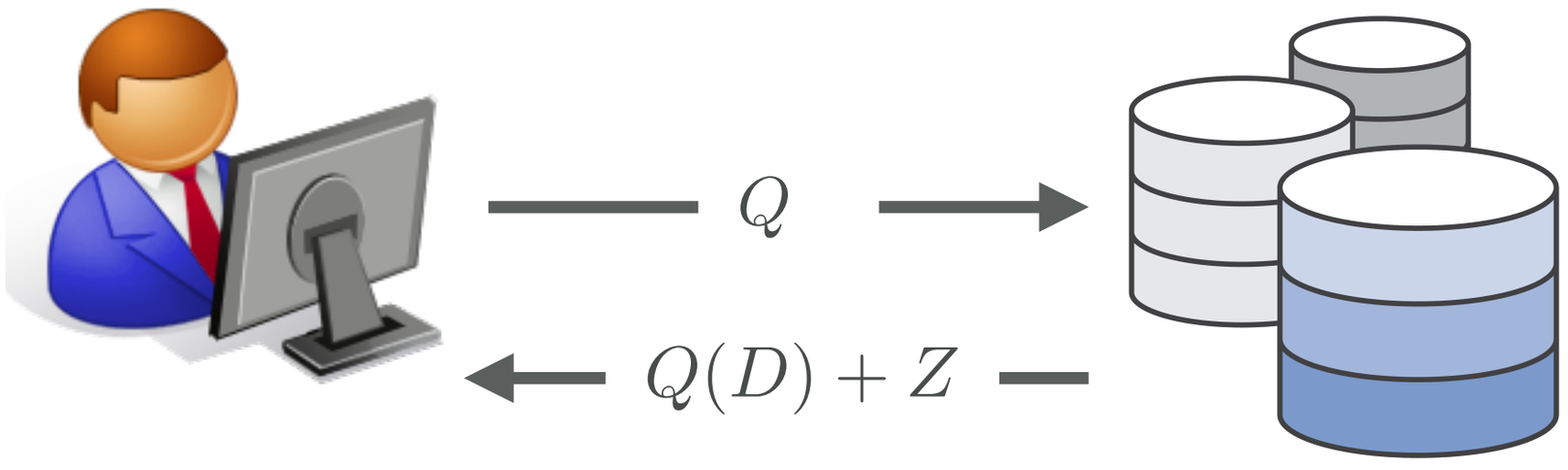
A randomized mechanism $\mathcal{M} : \mathscr{D} \to \mathscr{R}$ is ϵ-differentially private if, for any pair $D_1, D_2 \in \mathscr{D}$ of neighboring datasets and any output $O \in \mathscr{R}$:

$$\frac{\Pr[\mathcal{M}(D_1) = O]}{\Pr[\mathcal{M}(D_2) = O]} \leq \exp(\epsilon), \quad (\epsilon > 0)$$

- *The risk of a user to join the data set is bounded (by* **ε***)*

How Can we Achieve DP? [Dwork:06]

The Laplace Mechanism



$Q(D)$ = true answer

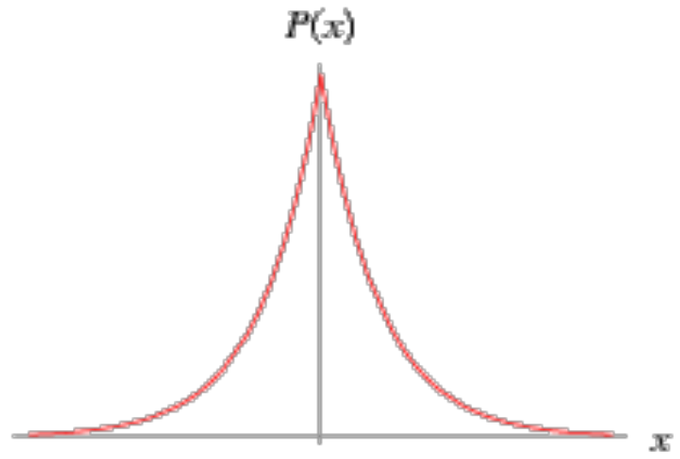
$Z \sim \text{Laplace}(\Delta_Q/\epsilon)$

Theorem (Laplace Mechanism)

Let $Q : \mathscr{D} \to \mathscr{R}$ be a numerical query. The Laplace mechanism $\mathcal{M}(D; Q, \epsilon) = Q(D) + Z$, where $Z \sim Lap(\frac{\Delta_Q}{\epsilon})$ achieves ϵ-differentially privacy.

$$b = (\Delta_Q/\epsilon)$$

How much does the output of Q changes if we add/remove one tuple from D ?



$$f(x \mid \mu = 0, b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$$

PDF

The problem

A Procurement/Competition Setting

On-Demand Multimodal System Design

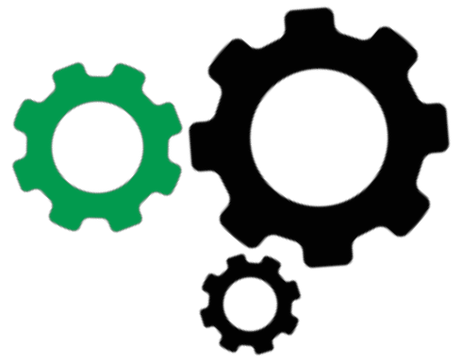
Transit Agency

D

(schedule)

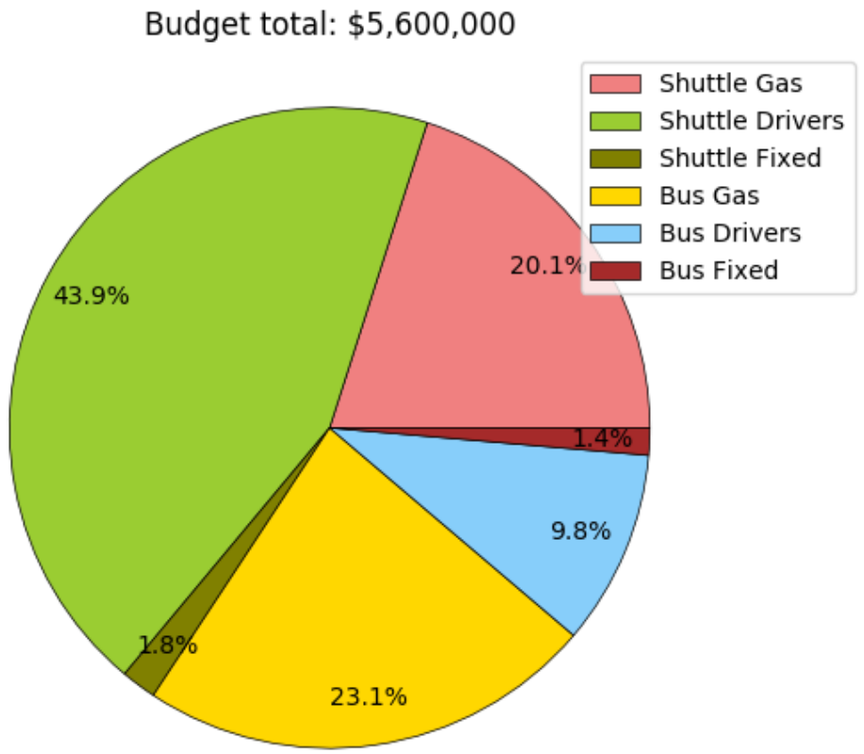
testing

simulation/deployment



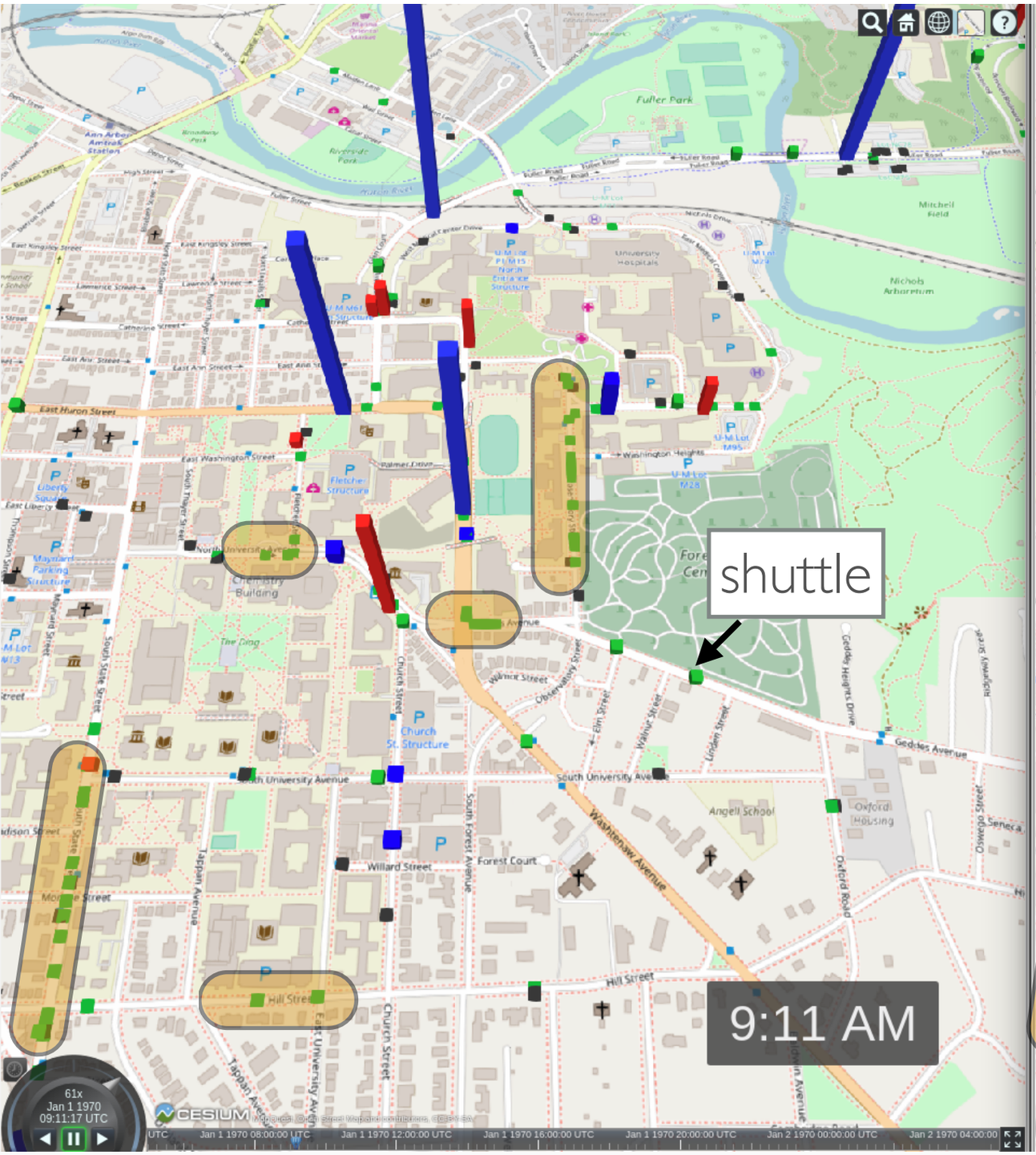
Let's apply the Laplace Mechanism

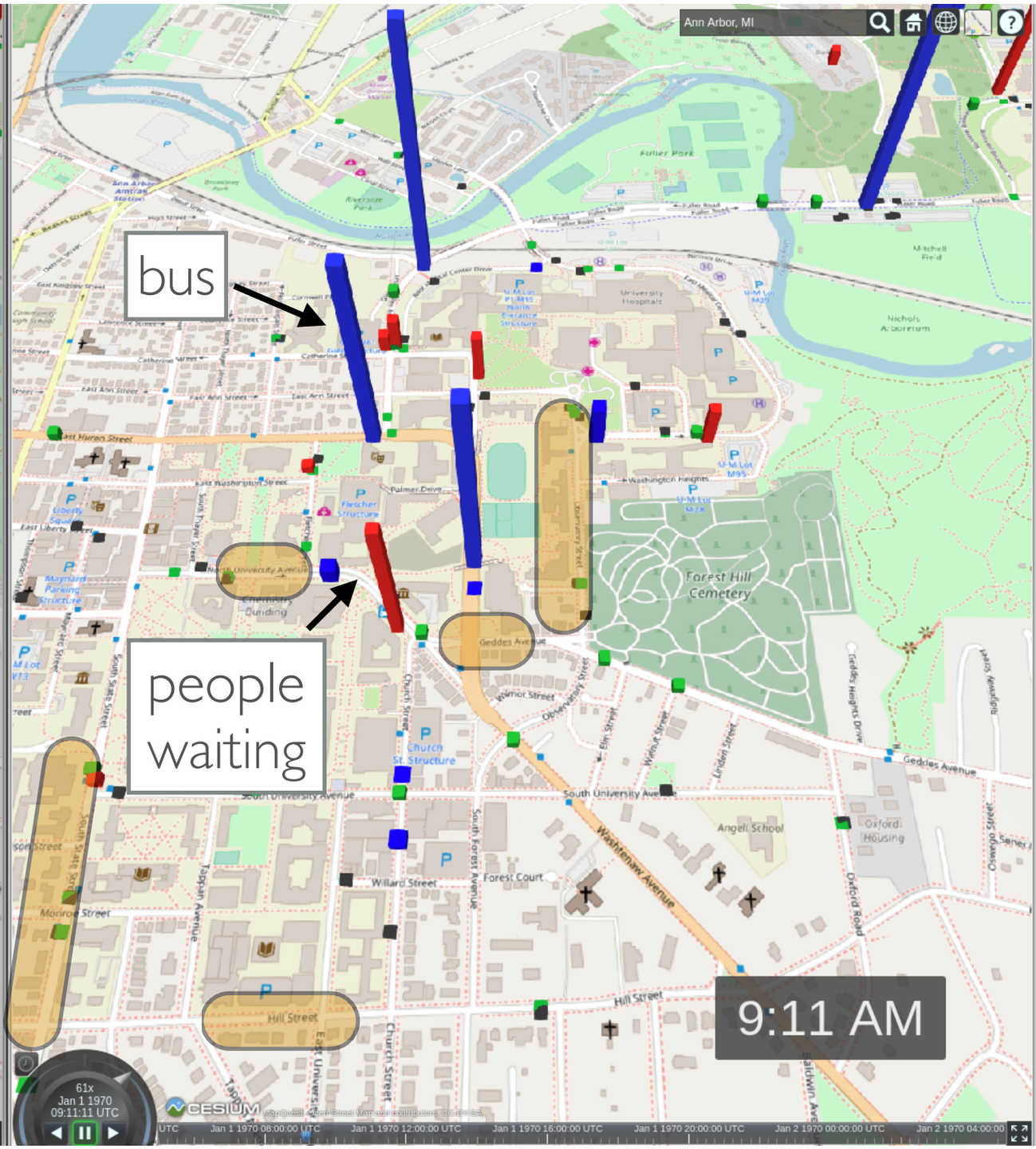
- Original dataset (weekday Oct. 2016)
- 37,714 trips



private data (Laplace)

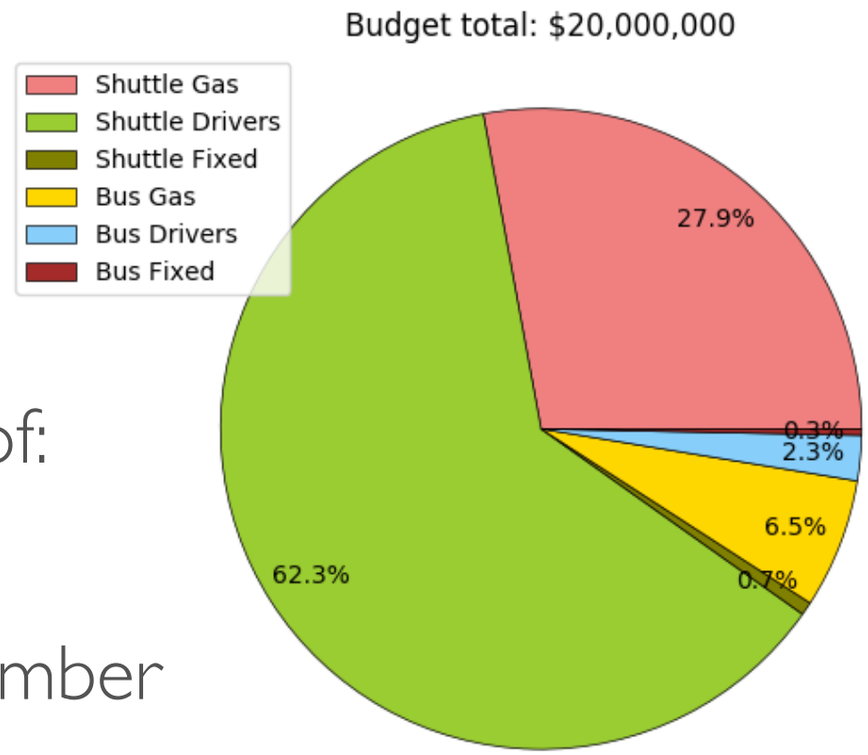
original data

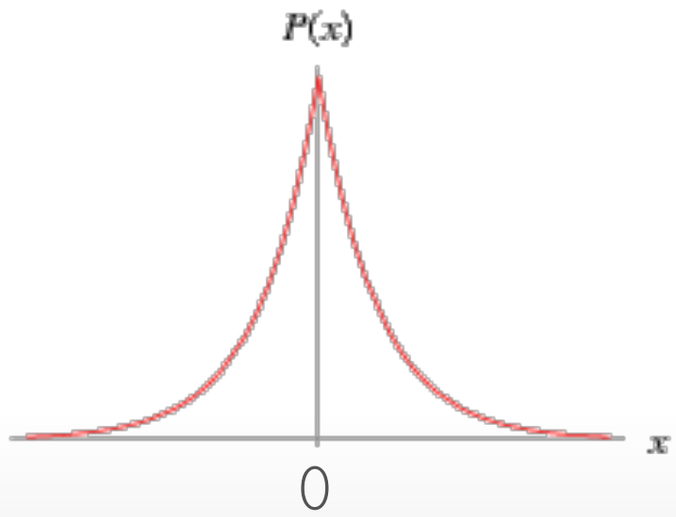




Let's apply the Laplace Mechanism

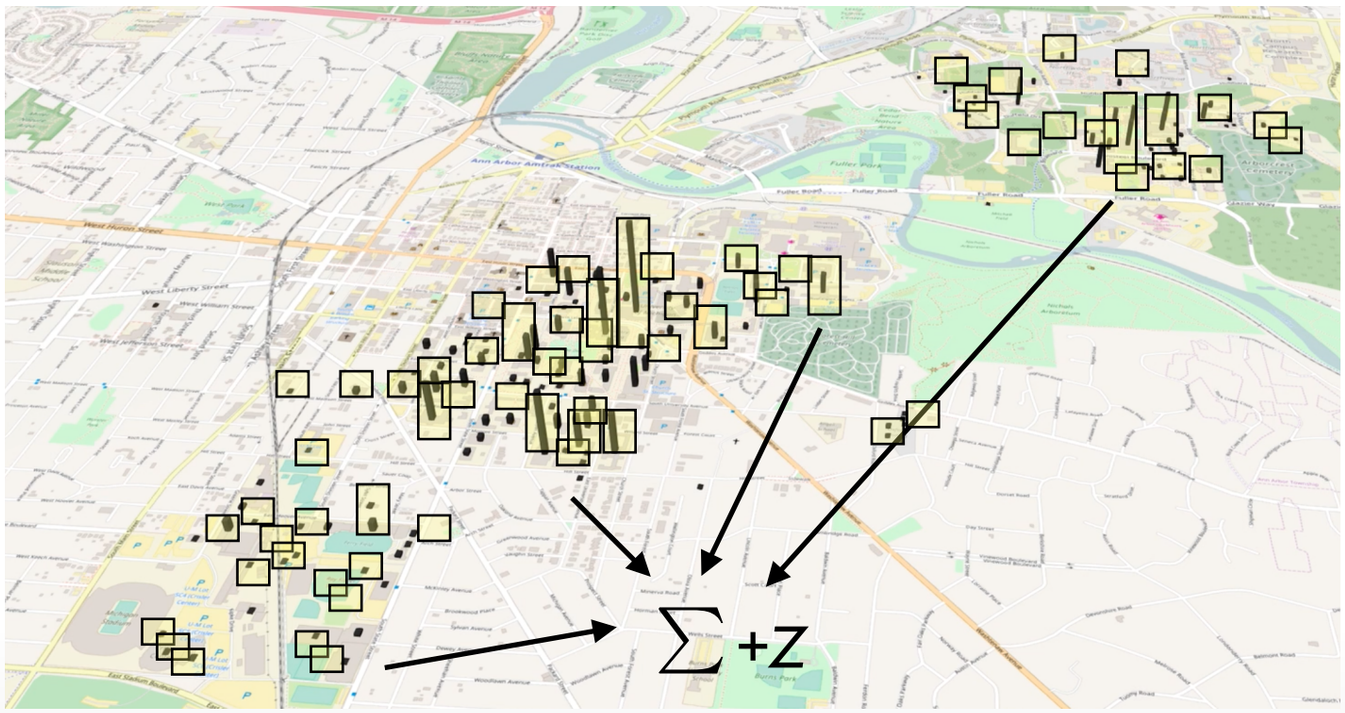
- Original dataset (weekday Oct. 2016)
- 37,714 trips
- Laplace Mechanism generates an average of:
  - $\varepsilon = 1$ — 261,032 trips
- Fleetsizing significantly overestimate the number of shuttles required
- Dataset is highly sparse. Counts can only be non-negative
- Laplace noise with rounding on many 0 counts highly biases the final distribution toward positive values





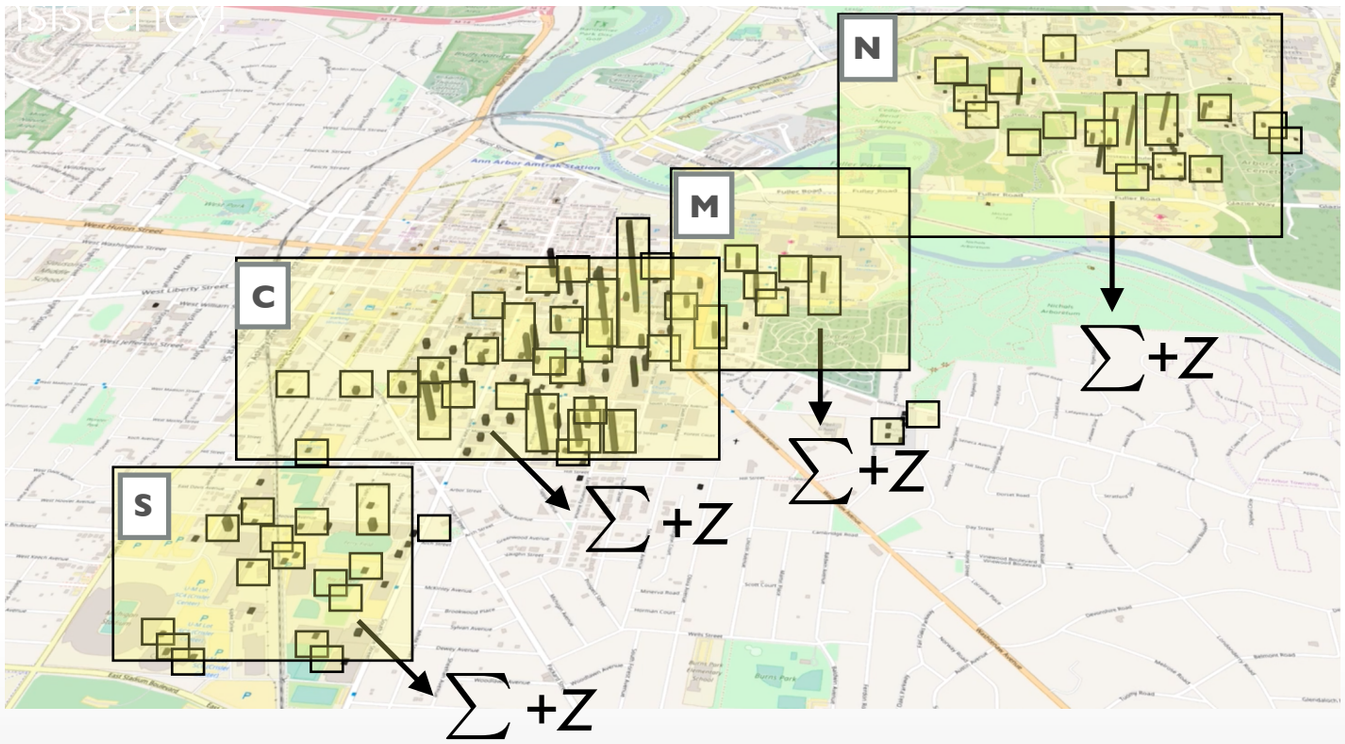
The CBDP Mechanism

- The main idea: Let's exploit the problem structure
- Ask additional queries on aggregate counts (problem *features*):
  - **The total number trips**
  - The number of trips in each zone
  - The number of trips made with each combination of transportation modes



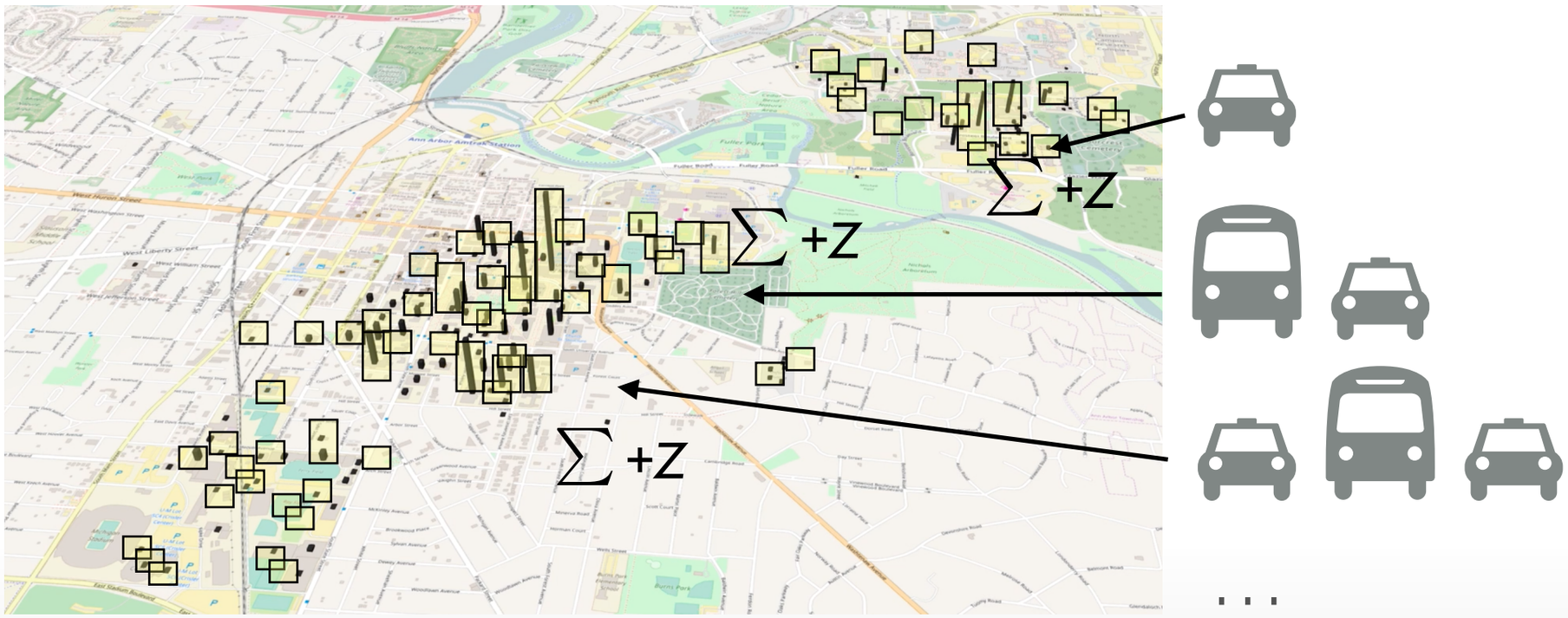
The CBDP Mechanism

- The main idea: Let's exploit the problem structure
- Ask additional queries on aggregate counts (problem *features*):
  - The total number trips
  - **The number of trips in each zone**
  - The number of trips made with each combination of transportation modes



The CBDP Mechanism

- The main idea: Let's exploit the problem structure
- Ask additional queries on aggregate counts (problem *features*):
  - The total number trips
  - The number of trips in each zone
  - **The number of trips made with each combination of transportation modes**



The CBDP Mechanism

- The main idea: Let's exploit the problem structure
- Ask additional queries on aggregate counts (problem *features*):
  - The total number trips
  - The number of trips in each zone
  - The number of trips made with each combination of transportation modes
- Use this (noisy) information to redistribute the noise introduced on the individual trips counts
- Also, enforce consistency!

decision variables

noisy counts (Laplace mechanism)

$$\textbf{minimize: } \|\mathbf{x} - \tilde{\mathbf{c}}\|_{2,w}^2 = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \tilde{c}_{ij})^2 \qquad \text{(O1)}$$

subject to:

$$\forall i', i : \mathbf{F}_{i'} \prec \mathbf{F}_i, j \in [n_i] : x_{ij} = \sum_{l : \mathbf{d}_{i'l} \subseteq \mathbf{d}_{ij}} x_{i'l} \qquad \text{(O2)}$$

$$\forall i, j : x_{ij} \geq 0. \qquad \text{(O3)}$$

The CBDP Mechanism

- The main idea: Let's exploit the problem structure
- Ask additional queries on aggregate counts (problem *features*):
  - The total number trips
  - The number of trips in each zone
  - The number of trips made with each combination of transportation modes
- Use this (noisy) information to redistribute the noise introduced on the individual trips counts
- Also, enforce consistency!

forces different estimates of the same quantities to be consistent

$$\textbf{minimize: } \|\mathbf{x} - \tilde{\mathbf{c}}\|_{2,w}^2 = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \tilde{c}_{ij})^2 \qquad \text{(O1)}$$

subject to:

$$\forall i', i : \mathbf{F}_{i'} \prec \mathbf{F}_i, j \in [n_i] : x_{ij} = \sum_{l : \mathbf{d}_{i'l} \subseteq \mathbf{d}_{ij}} x_{i'l} \qquad \text{(O2)}$$

$$\forall i, j : x_{ij} \geq 0. \qquad \text{(O3)}$$

The CBDP Mechanism

Properties

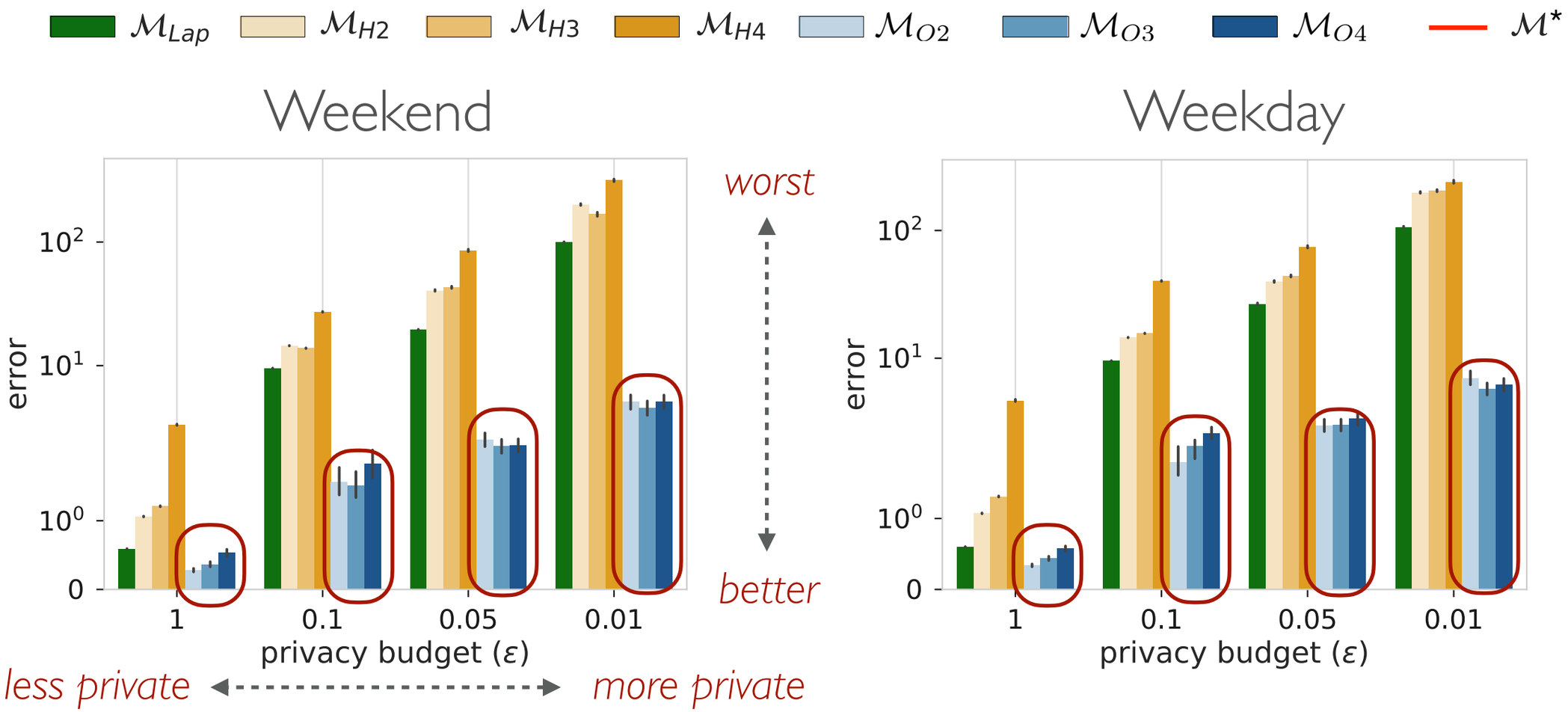
- Privacy: CBDP achieves ε-DP
  - Each set of queries for a problem feature has sensitivity 1
  - If there are k features, each query answer obtained by the Laplace mechanism is k/ε-DP
  - By composition, CBDP is ε-DP
- Efficiency: CBDP runs in polynomial time in the size of the universe and number of features
- Accuracy: The optimal solution to the optimization model of CBDP satisfies:

$$\|\mathbf{x}^* - \mathbf{c}\|_{2,w} \leq 2\|\tilde{\mathbf{c}} - \mathbf{c}\|_{2,w}$$

CBDP is at most a factor 2 away from optimality

Experimental Analysis

Average L1 Error

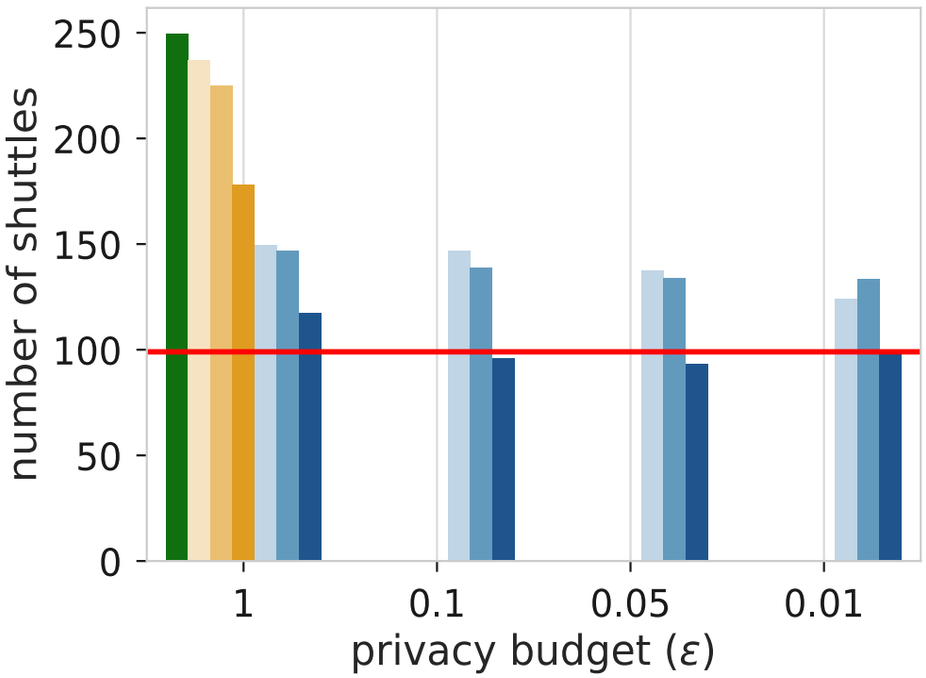


- Features: Total trips (2), Operating zones (3), Transit modes (4)
- Baselines: **Laplace** [Dwork:08], **Hierarchical** [Hay:10, Cormode:13]
- Summary: >= 1 order magnitude improvement

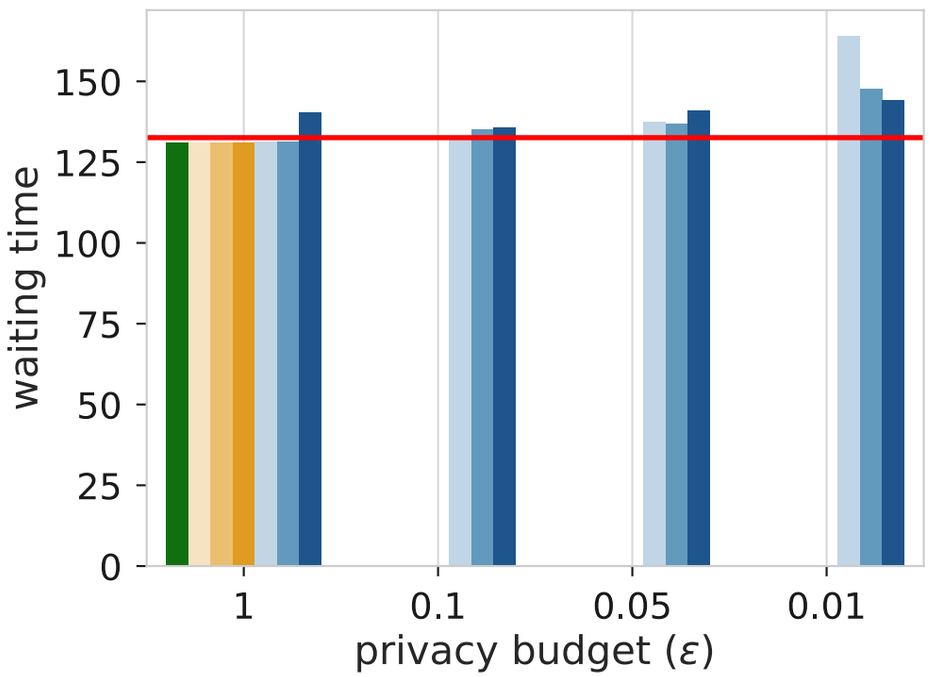
Experimental Analysis

System Cost and Waiting Time

System Cost proxy



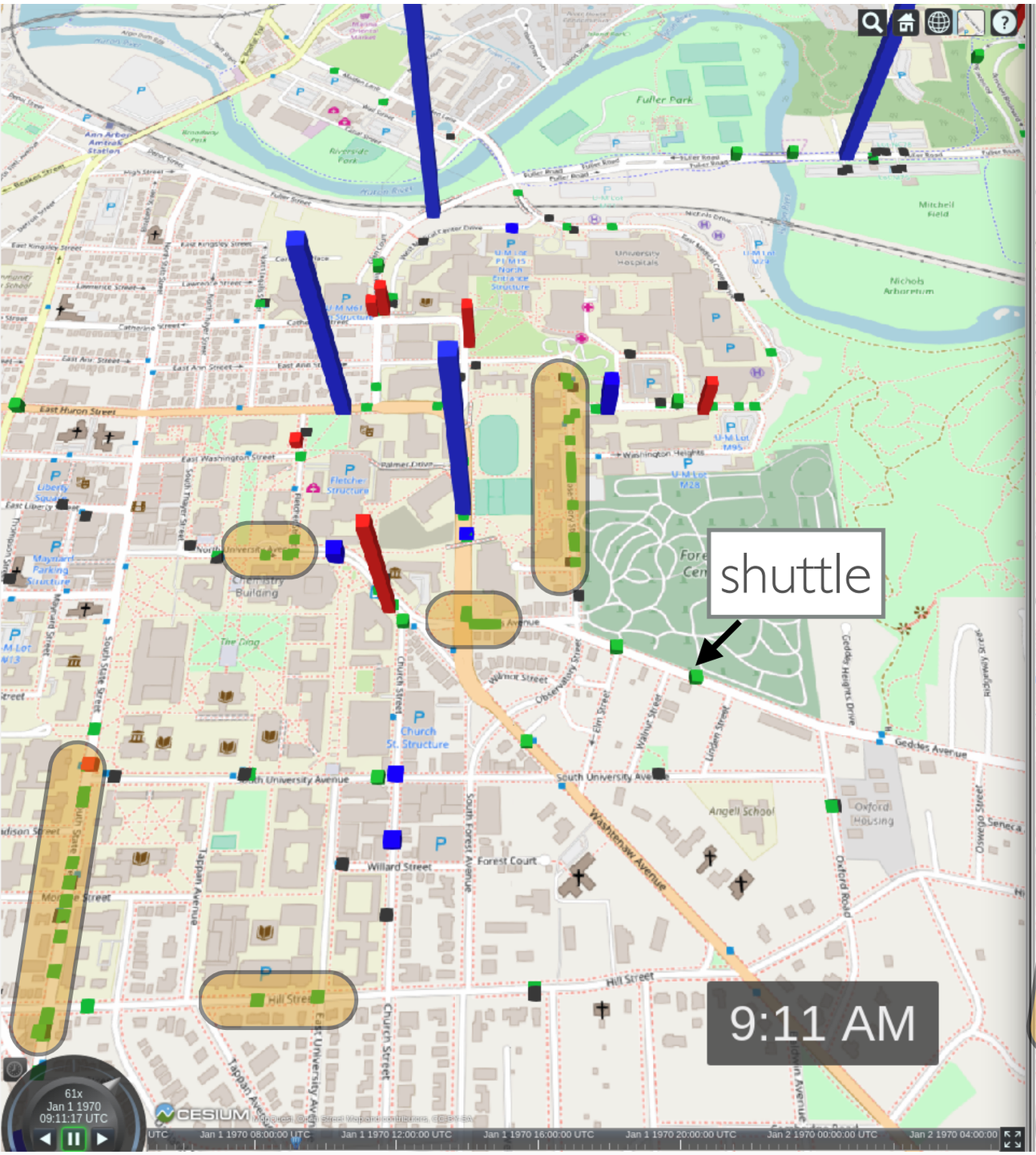
Waiting Time

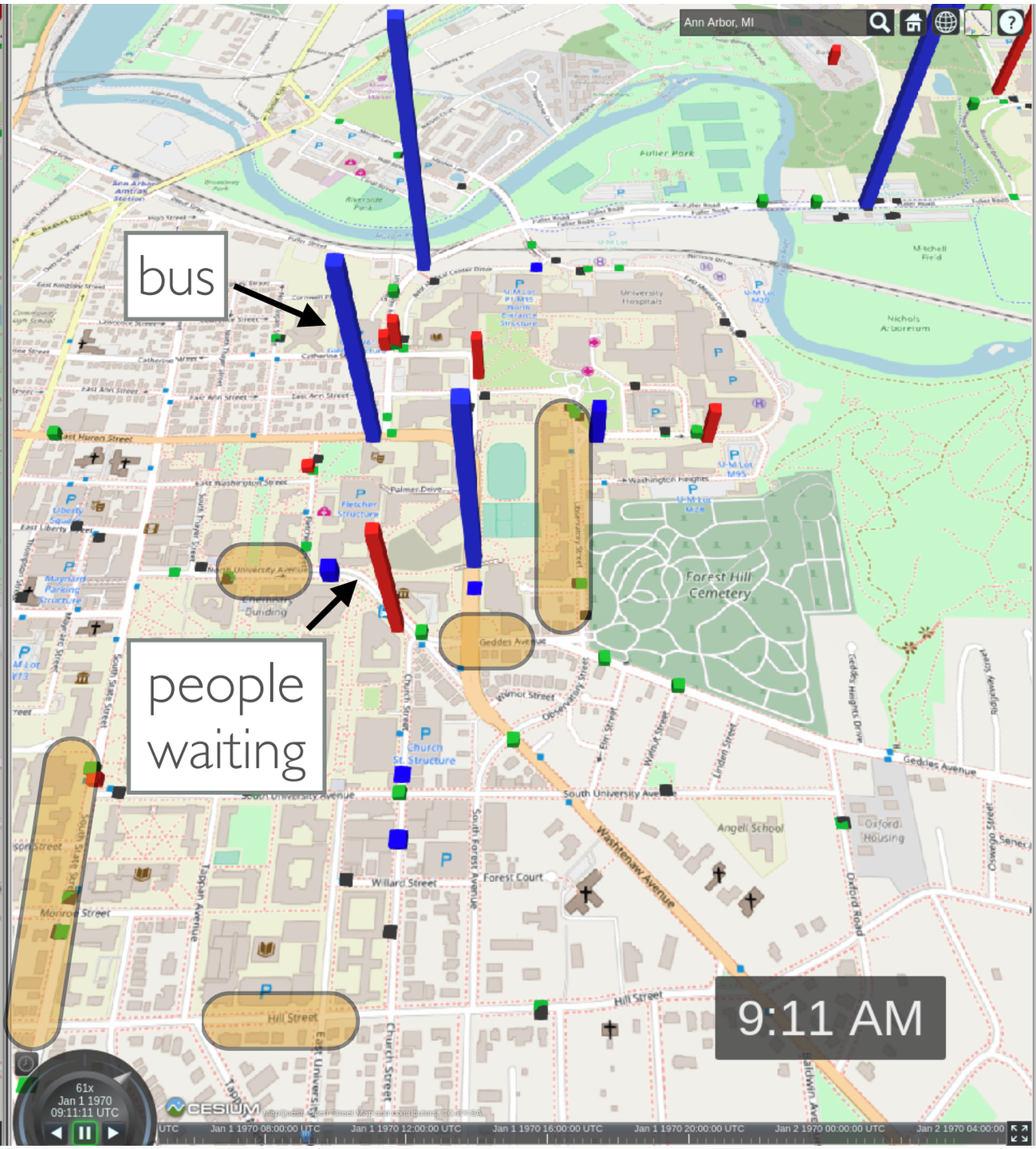


- Summary:
  - CBDP better estimates the system cost preserving the average waiting time.
  - The outputs of both Lap and H's cannot be processed by the simulator due to the large overestimation of total trips.

private data (Laplace)

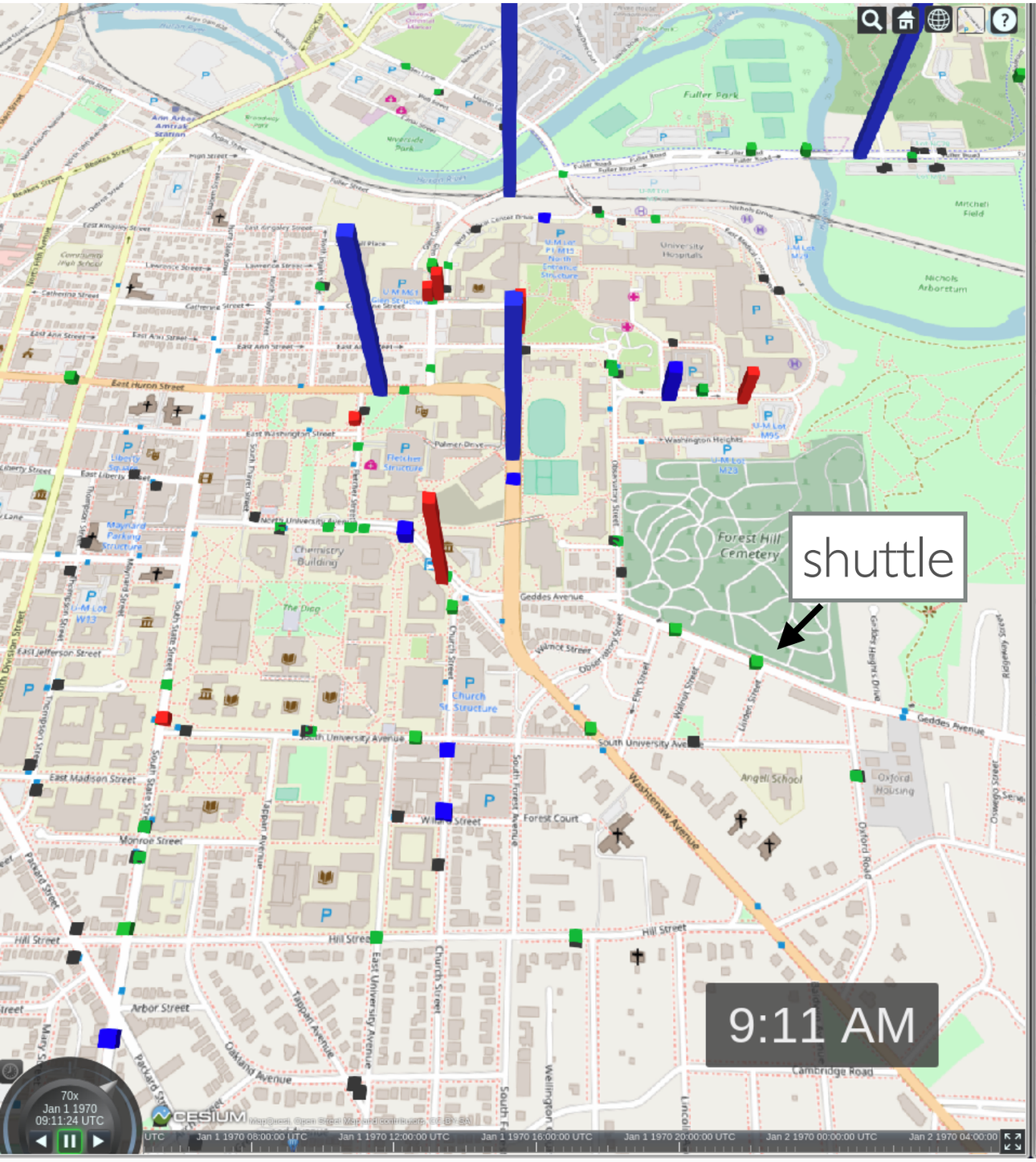
original data

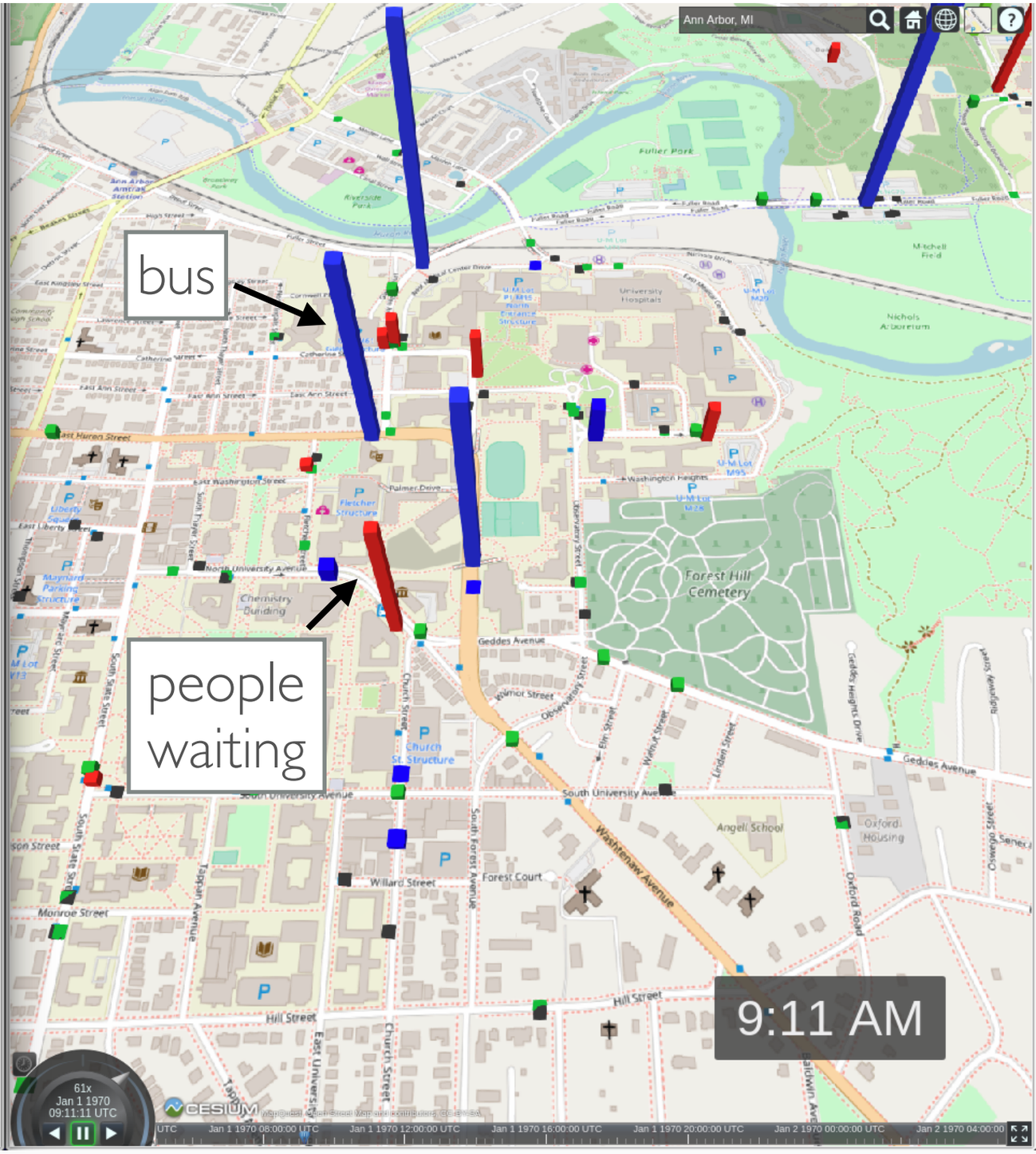




private data (CBDP)

original data





Closing thoughts

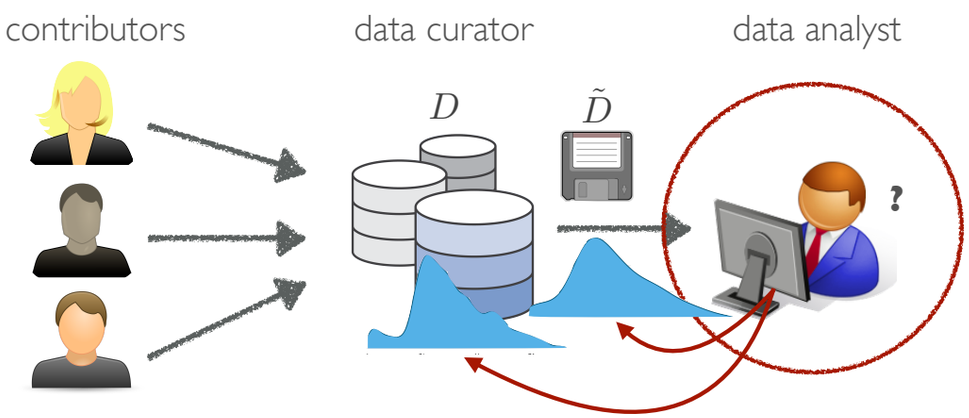
- AI-based services powered by personal data
- Differential Privacy Challenge in ODMTS
- Proposed a CBDP Mechanism which ensures:
  1. Differential private data release
  2. Faithfulness to the features of the problem
  3. Constraint consistency
- Large, complex MAS mobility problems can be solved with good accuracy while protecting the privacy of the individuals

Thank you

Poster #121

Differential Privacy

(Informal)



Contributor: Small participation risk (privacy loss)
Data analyst: Analysis on original and modified data are very similar (data distributions)

The CBDP Mechanism

- The main idea: Let's exploit the problem structure
- Ask additional queries on aggregate counts (problem *features*), e.g.,
  - The total number trips
  - The number of trips in each zone
  - The number of trips made with each combination of transportation modes
- Use this (noisy) information to redistribute the noise introduced on the individual trips counts
- Also, enforce consistency!

$$\textbf{minimize: } \|\mathbf{x} - \tilde{\mathbf{c}}\|^2_{2,w} = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \tilde{c}_{ij})^2 \qquad \text{(O1)}$$

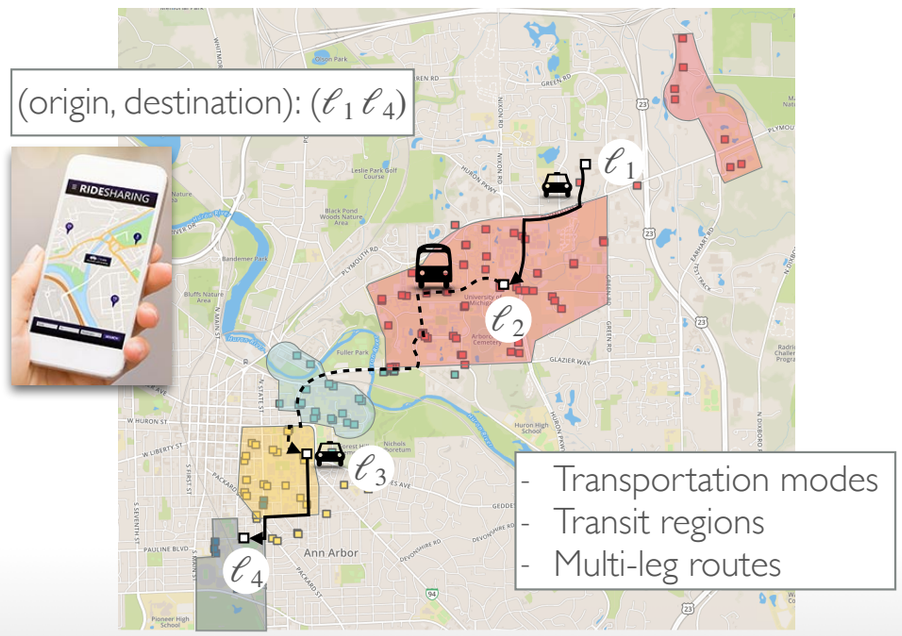
subject to:

$$\forall i', i : \mathbf{F}_{i'} \prec \mathbf{F}_i, j \in [n_i] : x_{ij} = \sum_{l : \mathbf{d}_{i'l} \subseteq \mathbf{d}_{ij}} x_{i'l} \qquad \text{(O2)}$$

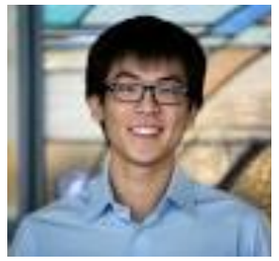
$$\forall i, j : x_{ij} \geq 0. \qquad \text{(O3)}$$

On-Demand Multi-modal Transportation System

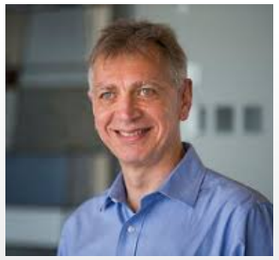
Optimizing a Multi-leg Trip: Features



Ferdinando Fioretto
fioretto@umich.edu



Chansoo Lee
chansool@umich.edu



Pascal Van Hentenryck
pvanhent@umich.edu