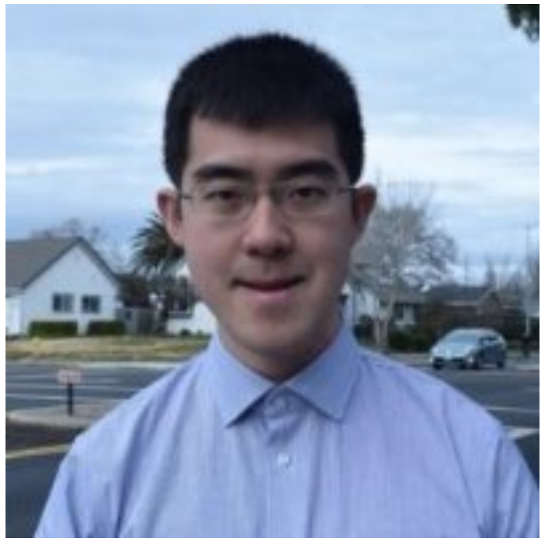
Where does domain expertise stand in transportation AI research?

Yueyue (Yo-Yo) Fan
Professor, Civil and Environmental Engineering
University of California, Davis

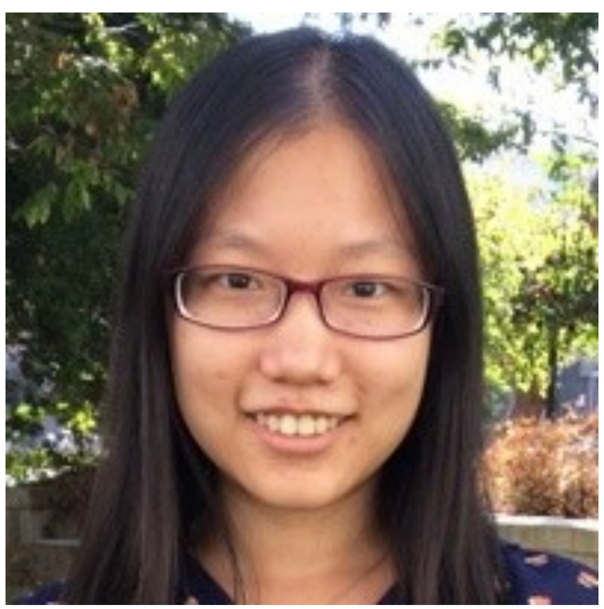
Collaborators



Dr. Yudi Yang
Data Scientist, Amazon

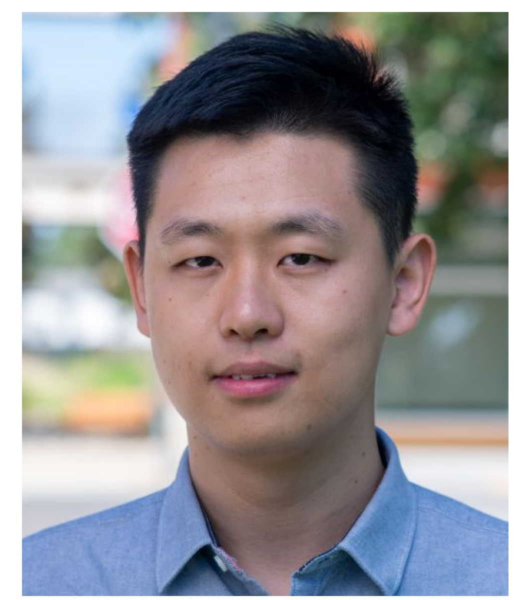
Prof. Roger Wets
Mathematics, UCD

Dr. Ran Sun
Assistant Prof., Utah State U.



Dr. Xiaoyue Li
Software Engineer, Google

Prof. James Sharpnack
Statistics, UCD
AI Research Scientist, Duolingo

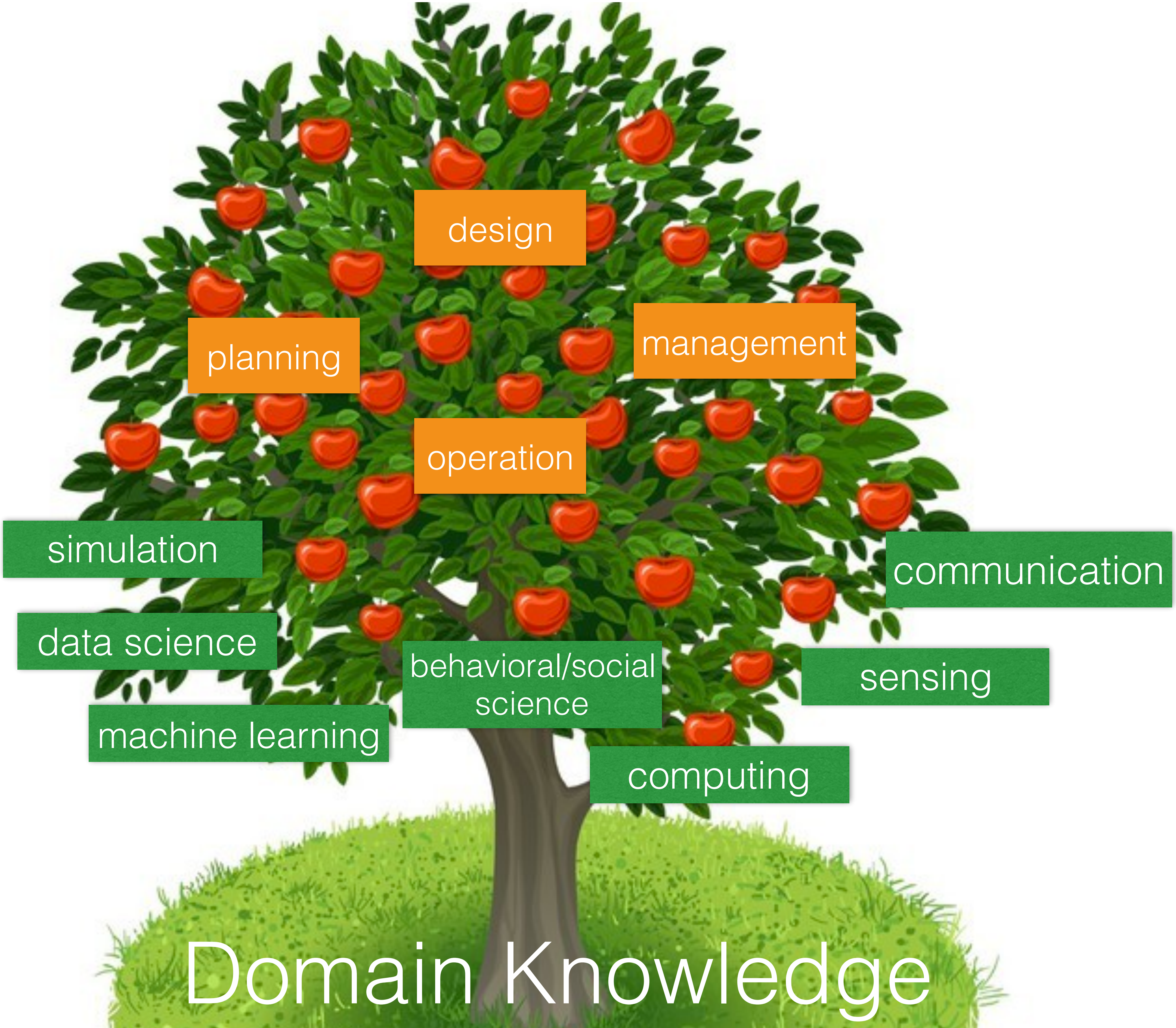


Dr. Xinyue Hu
Data Scientist, Chewy

What motivated this talk?

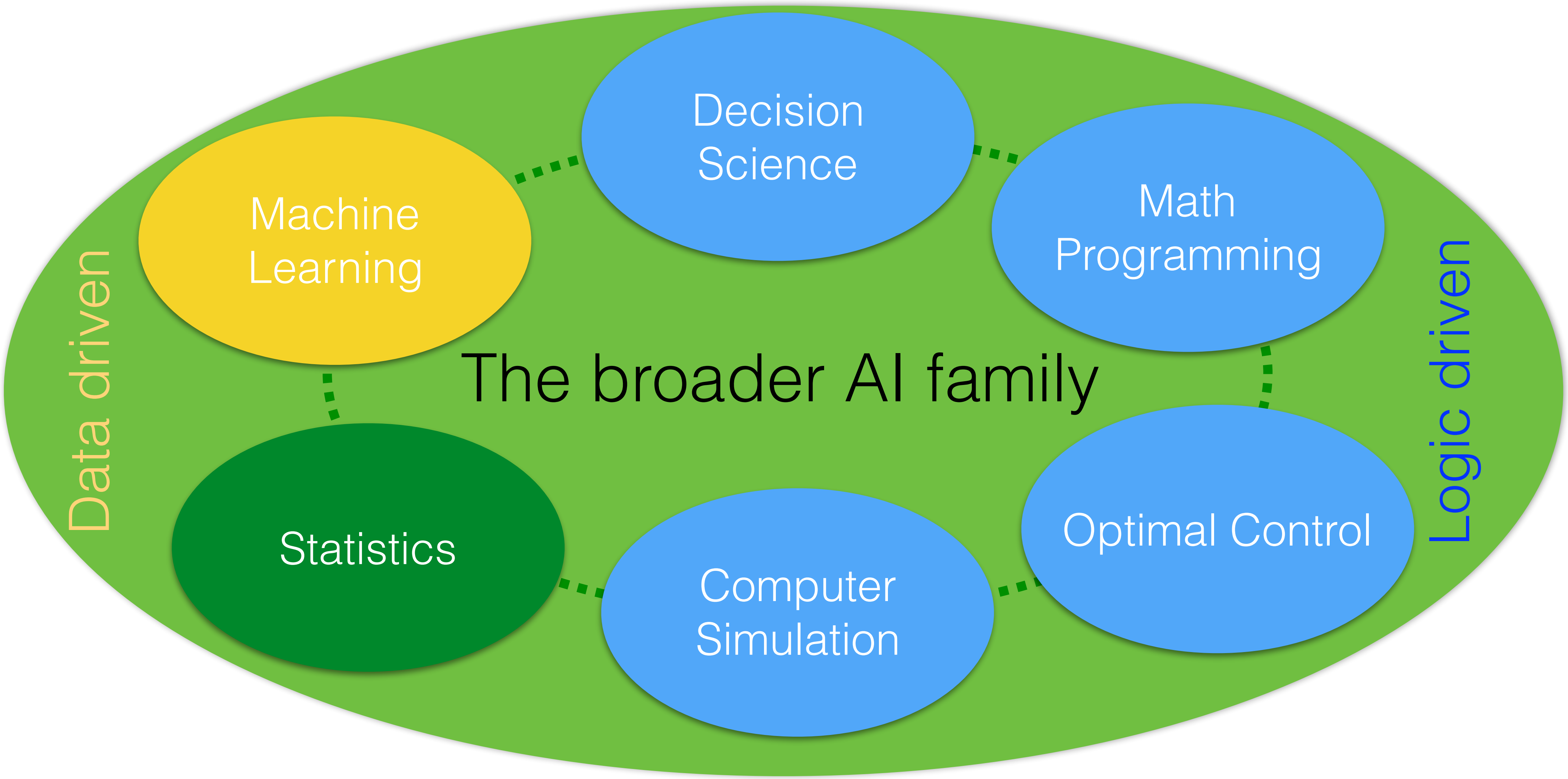
- Where do we (transportation researchers and practitioners) currently stand in the fast growing AI space?
  - competitive advantages - rich domain knowledge, access to data and information, translation to planning/design/construction processes, etc.
  - gaps - limited data/analytics/programming training in traditional CEE curriculum, generic data analytics methods may not fit, open data/tools for testing/validation/bench marking, etc.
- How do we strategically position ourself in preparation for the future?

a healthy transportation research eco system

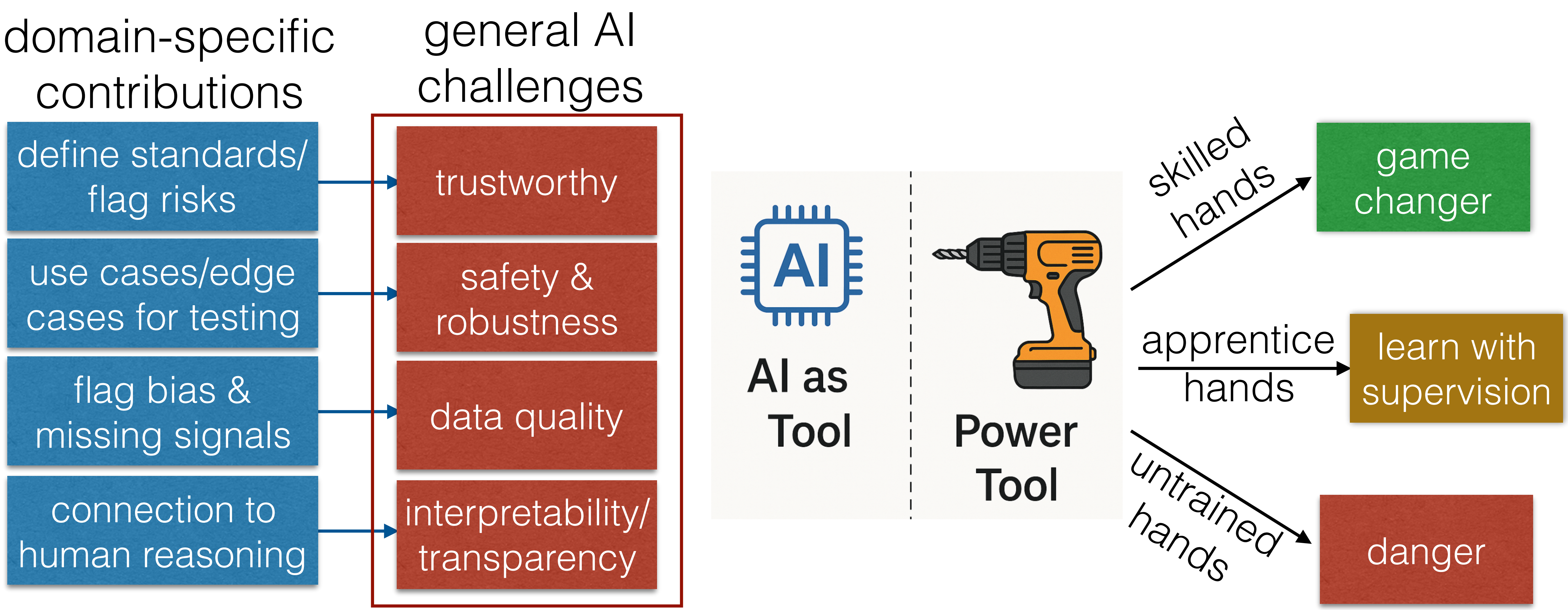


What is AI?

ChatGPT: “AI (Artificial Intelligence) refers to the capability of machines or software to perform tasks that typically require human **intelligence**. These tasks include things like understanding language, recognizing patterns, solving problems, making decisions, and learning from data.”



AI as an amplifier, not a replacer



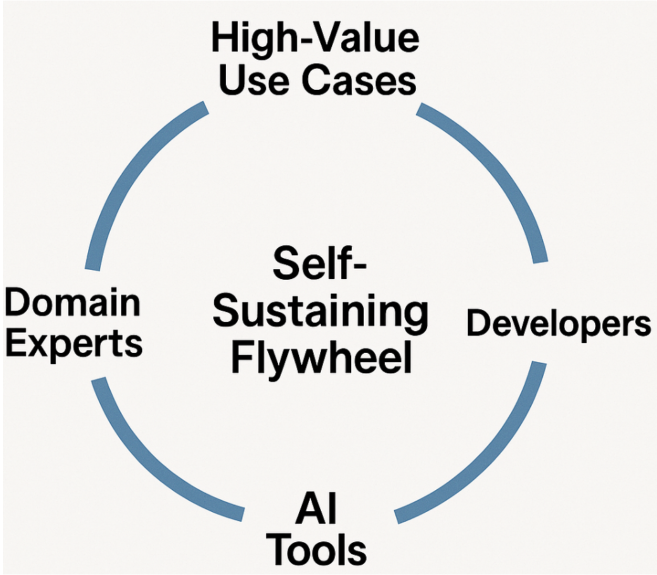
Besides technical issues, let's think about business

- Platform-based economy is widely considered one of the most transformative business innovations of the 21st century: it is an economic architecture shift
    - Well-known examples of platform-based solutions: Amazon Marketplace, Facebook, Airbnb, Youtube

Platform-based economy and network effect

- The AI industry imbeds a platform structure
    - The AI market exhibits bidirectional cross network effects: more AI users ↔ more AI developments
    - AI use cases also generate content (content generating platform)

Network Effects



- Network effects plays a vital role in platform businesses
    - "Cold start" problem
    - A coordinating entity is needed to make strategic decisions to leverage "network externality"
- Who should be the coordinating entity in the transportation AI space?

A public-interest serving platform (governance)?

Public/community Interest Serving Innovations

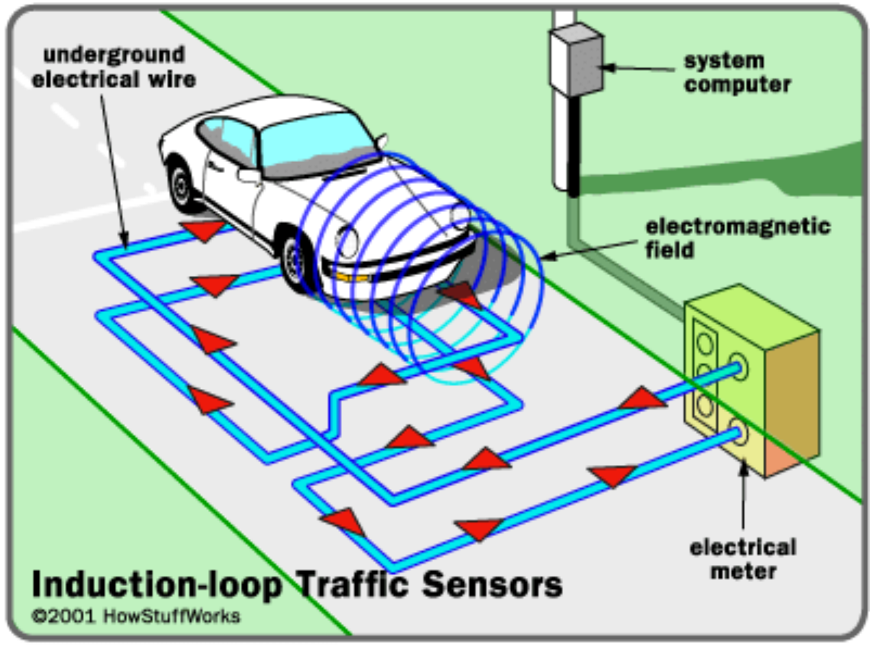
- Challenges in translating AI innovations to impact
  - limited local resources and capacity
  - varying interests and voices
  - need **evidence** for scalable adoption

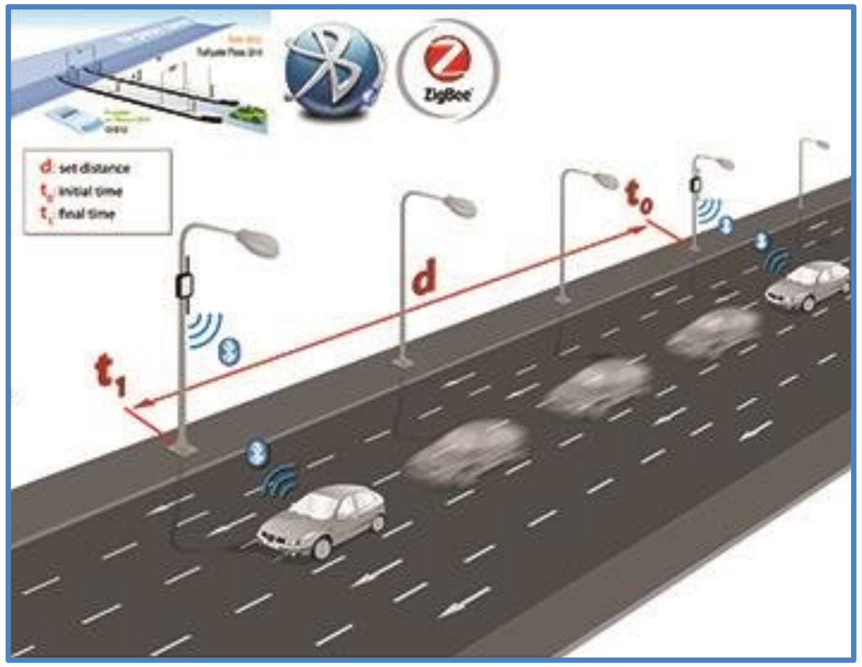
Collective effort: shared testbeds for testing and benchmarking?

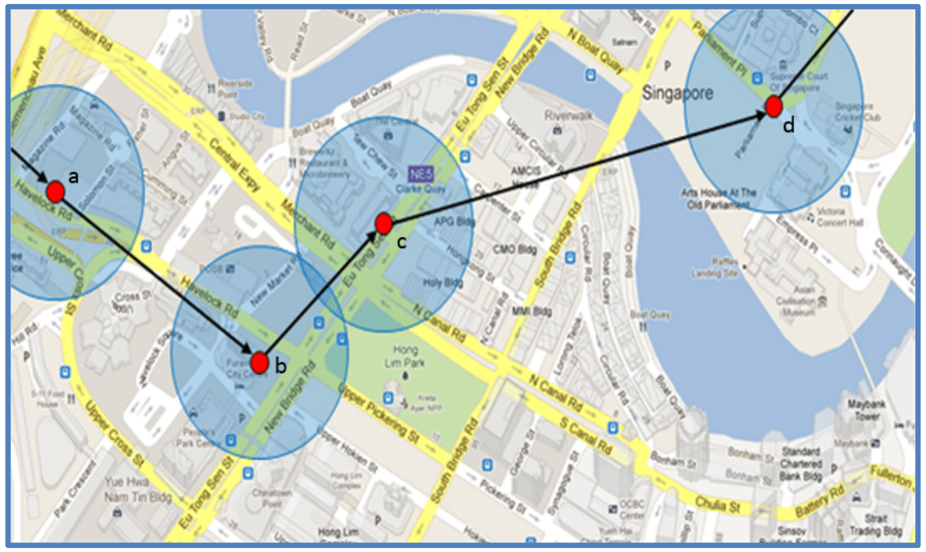
Some examples from my own research to highlight how domain expertise could improve generic data analytics methods

Available data/information

- Sensor data (hard data)
  - Link counts collected by loop detectors
  - Path counts collected via vehicle reidentification techniques
  - GPS trajectories and mobile data







Available data/information

- Domain knowledge (soft information)
  - Relationship based on network topology
  - Relationship based on behavior assumptions
  - Relationship based on the physics of traffic flow

May be expressed by an analytical (linear) model

An example:

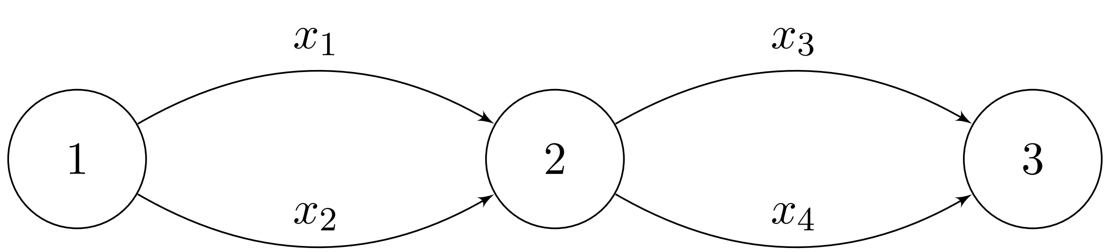


Figure 1: Four link problem

The link space $\mathcal{X}$ is in $\mathbb{Z}_+^4$ or $\mathbb{R}_+^4$.
The demand space $\mathcal{Q}$ is in $\mathbb{Z}_+^3$ or $\mathbb{R}_+^3$.
The path space $\mathcal{F}$ is in $\mathbb{Z}_+^8$ or $\mathbb{R}_+^8$.

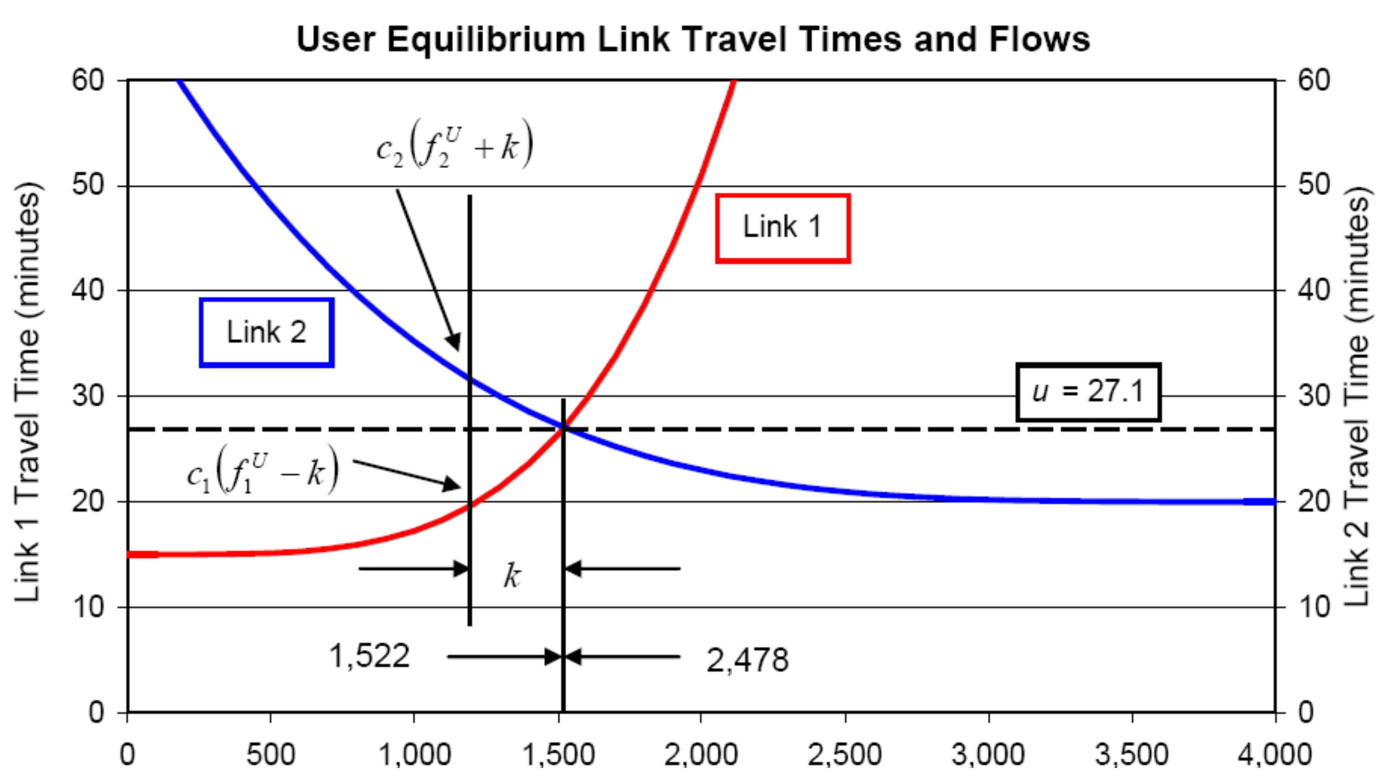
$$\boxed{\begin{aligned} x &= \mathcal{F} f \\ q &= \mathcal{B} f \end{aligned}}$$

$$\mathcal{F} = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Available data/information

- Domain knowledge
  - Relationship based on network topology
  - Relationship based on behavior assumptions
  - Relationship based on the physics of traffic flow

May be expressed by an analytical (nonlinear) model

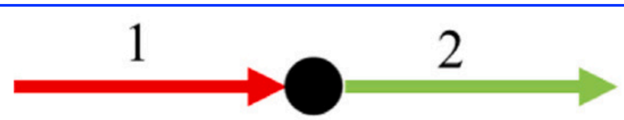


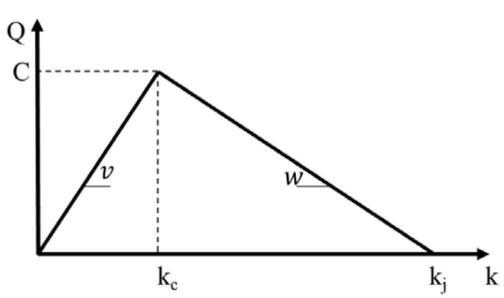
$$
\begin{aligned}
\min_{x} \quad & \sum_i \int_0^{x_i} c_i(s)\,ds \\
\text{subject to} \quad & x = \mathcal{F} f \\
& q = \mathcal{B} f \\
& f \geq 0
\end{aligned}
$$

Available data/information

- Domain knowledge
  - Relationship based on network topology
  - Relationship based on behavior assumptions
  - Relationship based on the physics of traffic flow dynamics

May be expressed as a system of ODEs, PDEs or a simulation model

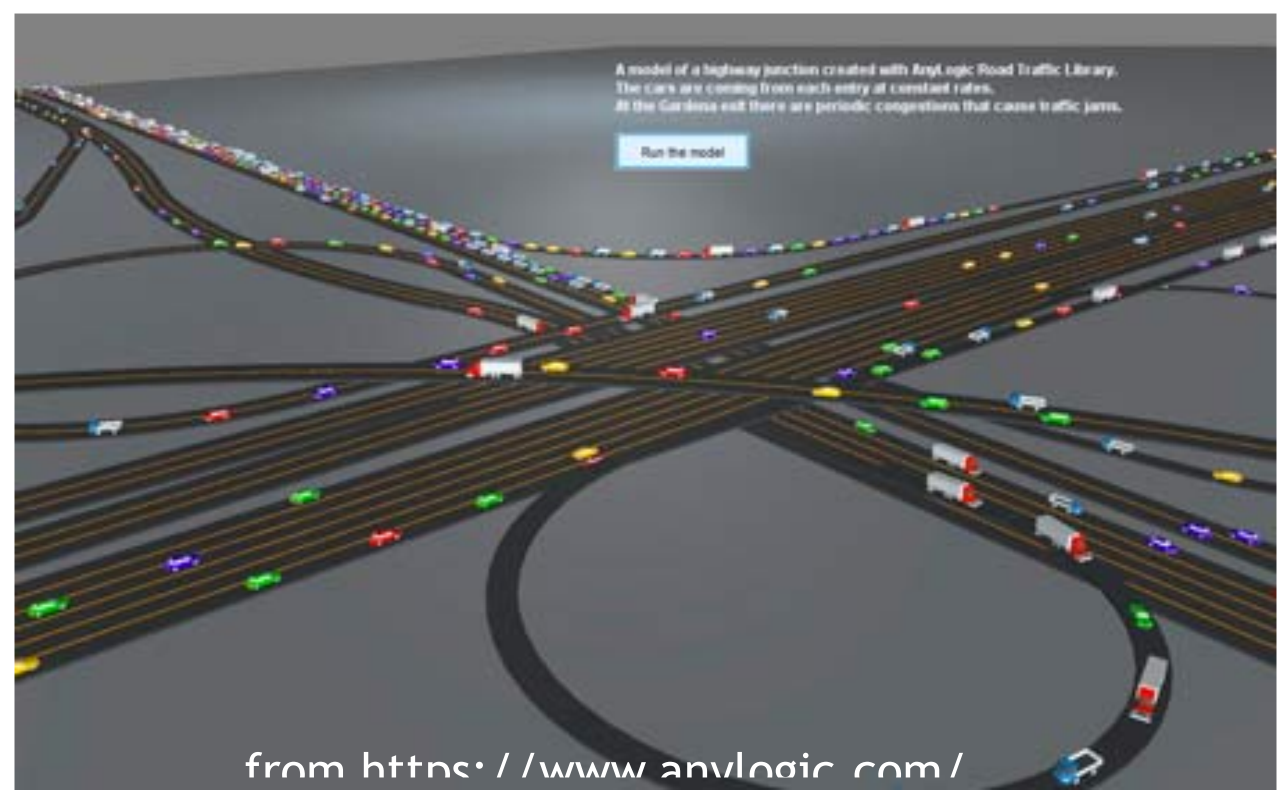




(a) Triangular fundamental diagram

Under one possible congestion mode, the densities can be expressed as:

$$\begin{bmatrix} \dot{k}_1 \\ \dot{k}_2 \end{bmatrix} = \begin{bmatrix} -v_1 & 0 \\ v_1 & 0 \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} + \begin{bmatrix} f_1 \\ -g_2 \end{bmatrix}$$



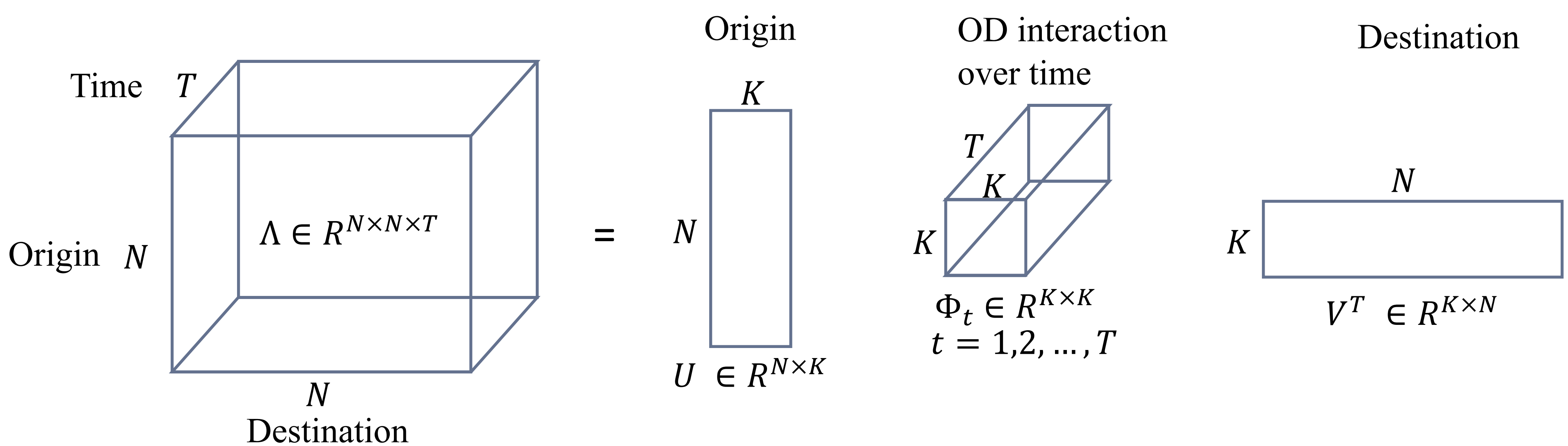
Example 1:
Usage-Aware Data Representation Learning
(an example of reflecting domain knowledge in the loss function design)

More details are in:

1. R. Sun and Y. Fan (2024), Usage-aware representation learning for critical information identification in transportation networks, Transportation Research Part C: Emerging Technologies, Volume 160.
2. X. Li, R. Sun, J. Sharpnack, Y. Fan (2023), Understanding Origin-Destination Ride Demand with Interpretable and Scalable Non-negative Tensor Decomposition, Transportation Science, 57 (6), 1473-1495.

An illustration of a basic data representation learning model

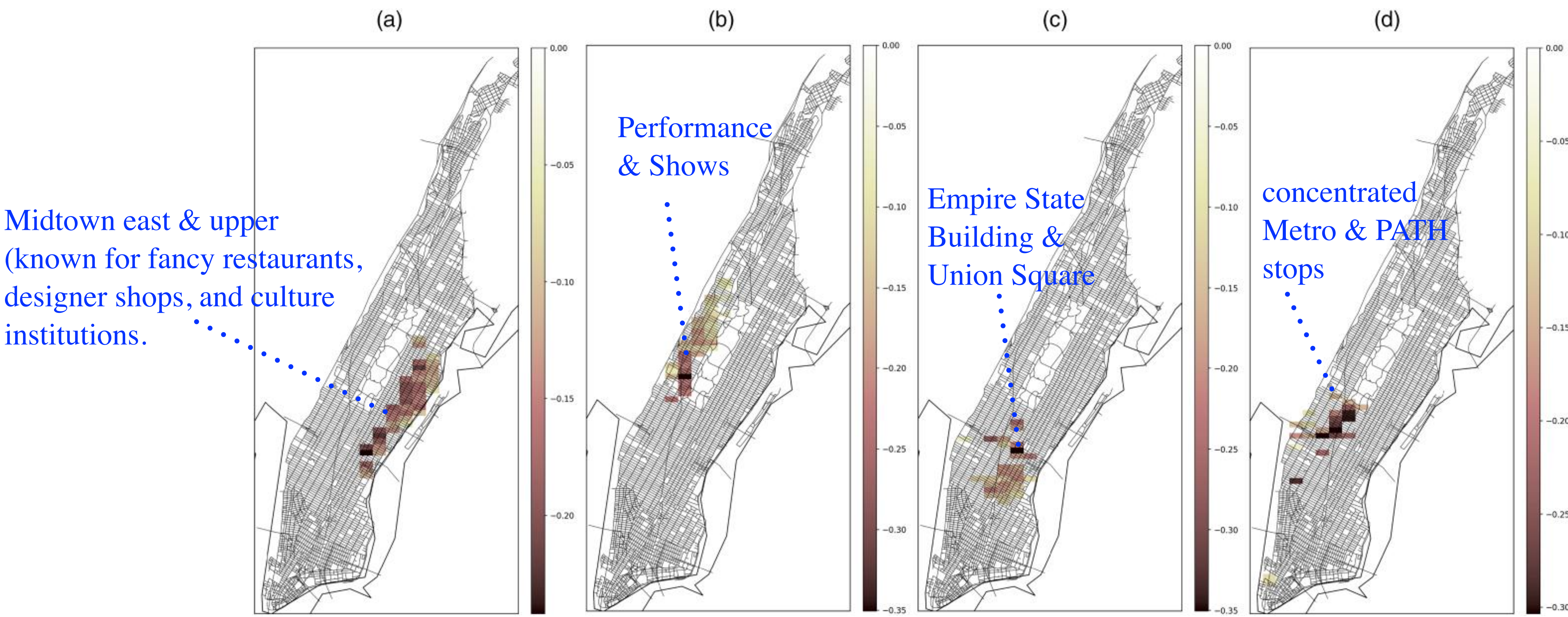
- using travel demand as an example



Motivations for Data Representation Learning

- Discovering important data structure
- Reducing dimension
- Denoising data

Example of Learned Key Destination Factors in NYC Taxi Data (2015)



A usage-aware data representation learning model

loss term based on data itself

loss term reflecting downstream application

$$\boxed{\begin{aligned} \min_{\Phi \in \mathbb{R}^{d\times p}} \quad & \mathcal{Q}(X, \hat{X}) + \gamma \mathcal{R}(X, \hat{X}, \mathcal{A}) \\ \text{s.t.} \quad & \Phi^T \Phi = I_{p\times p} \end{aligned}}$$

for an application aiming to reduce the total network travel time:

UADR
$$\min_{\Phi \in \mathbb{R}^{d\times p}} \quad \|X - X\Phi\Phi^\top\|_F^2 + \gamma \sum_{i=1}^{n} w_i (x_i^T a_i - x_i^T \Phi\Phi^\top a_i)^2$$

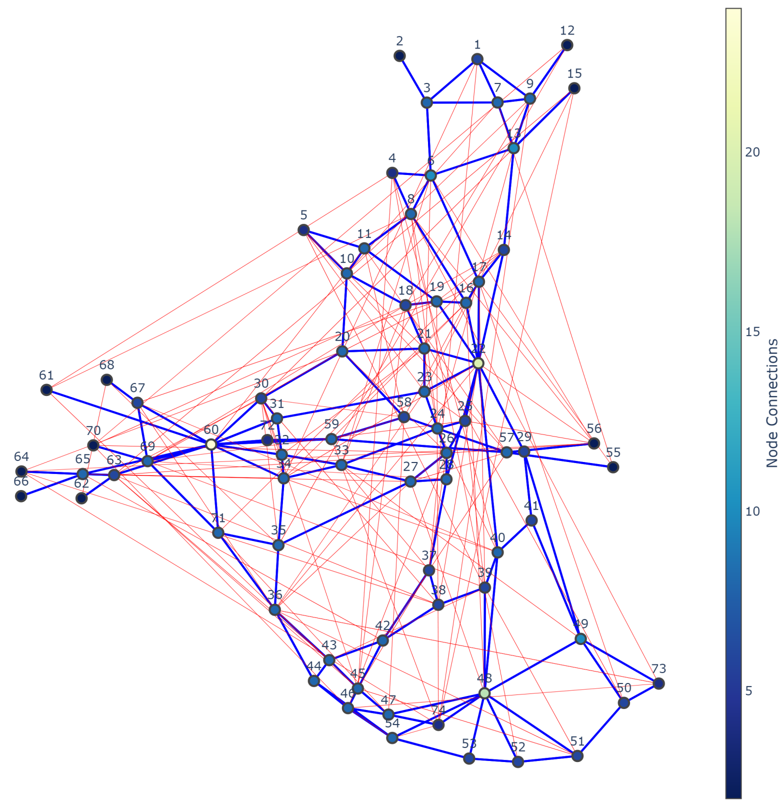
$$\text{s.t.} \quad \Phi^\top \Phi = I_{p\times p}$$

For link e in the network, the link travel time can computed as $c_e = t_e(1 + \alpha(\frac{v_e}{\kappa_e})^\beta)$

$$x_i^\top a_i = x_i^\top H_i^\top c_i$$

where $H_i \in \mathbb{R}^{l\times d}$ is linear proportional mapping of traffic assignment

Application 1: Boston Network
(demand is used for estimate the total network travel time)



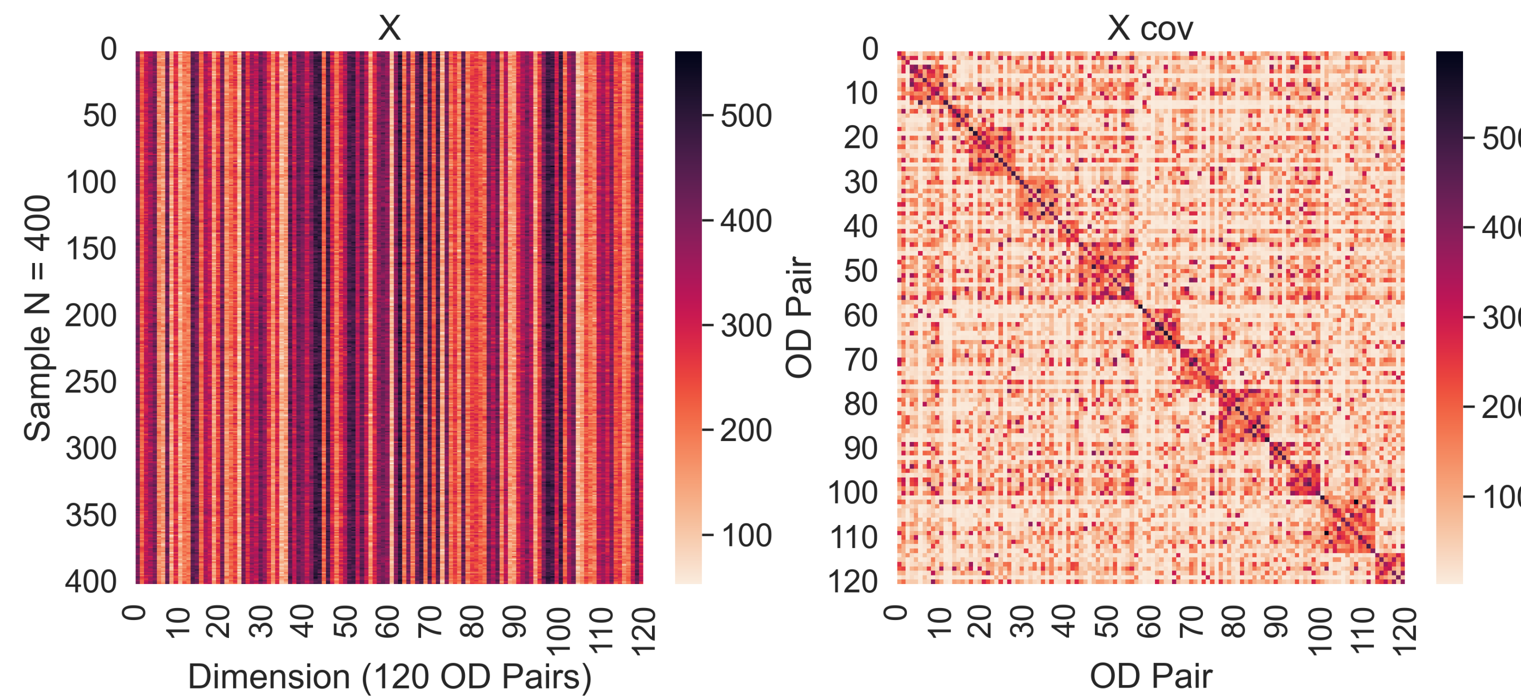
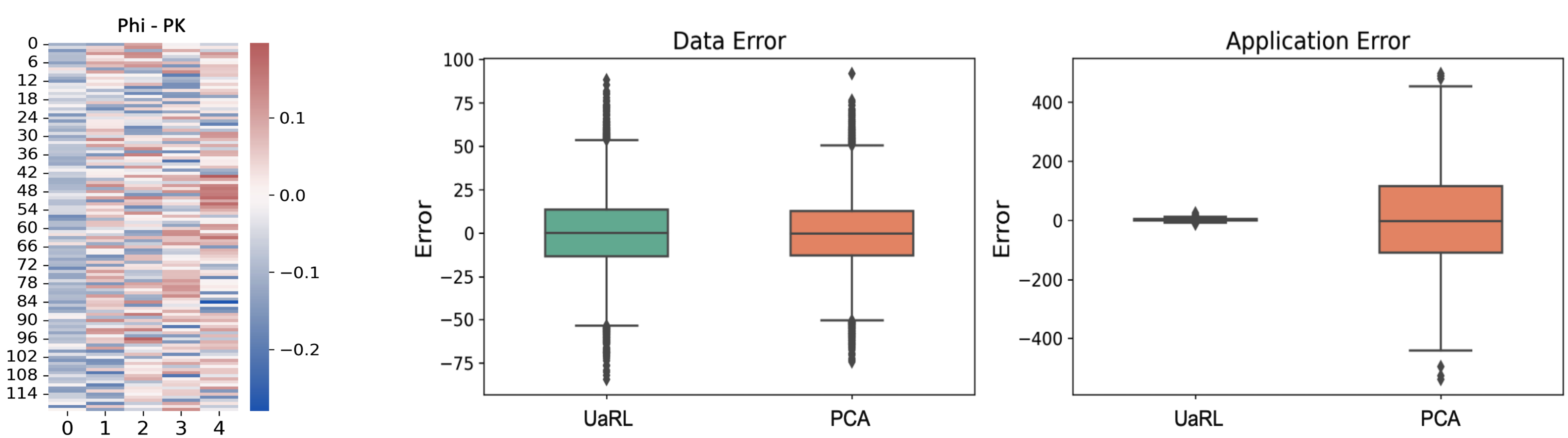


Figure 5: OD demand data for Boston network

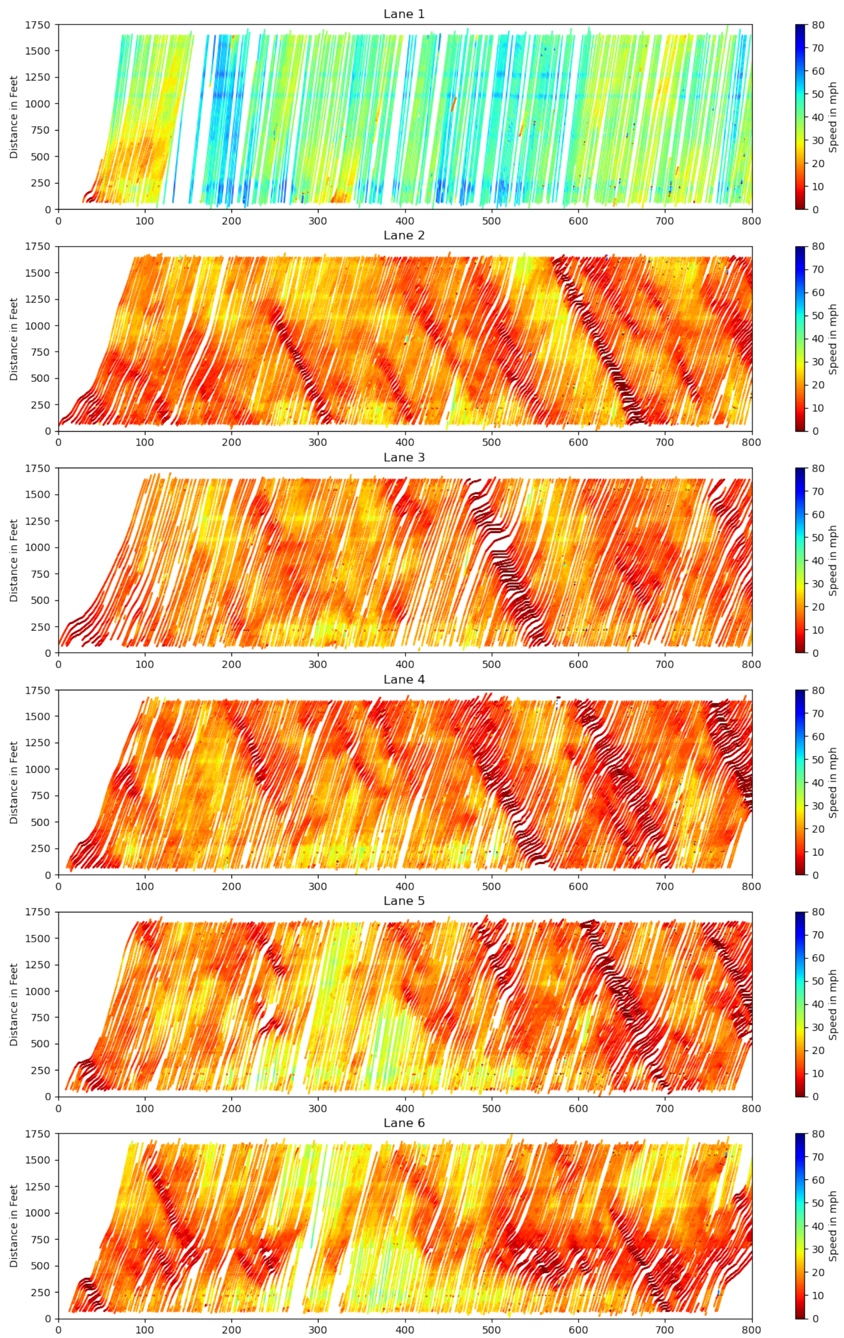


(a) UaRL Phi vs PCA

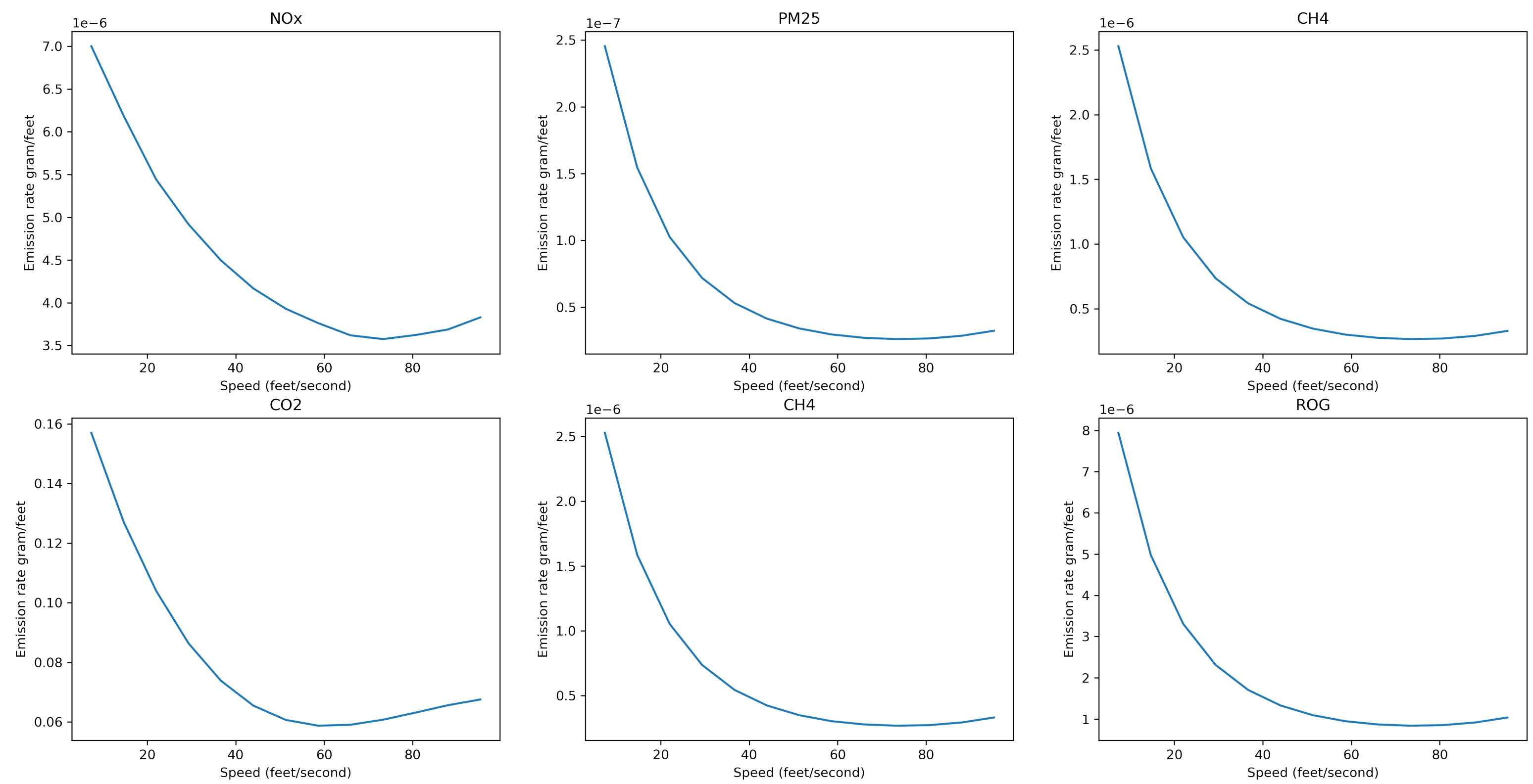
(b) Reconstruction error

Figure 7: Comparison of results from UaRL vs. PCA in Boston network

Application 2: use speed data to estimate emissions



Vehicular trajectories with instantaneous speeds (inspired by Li et al, 2020)



Emission rates

Take PM2.5 as an example, it can be modeled as a polynomial function,

$$y = -1.350 \times 10^{-8}x + 2.000 \times 10^{-10}x^2 - 0.955 \times 10^{-13}x^3 + 3.228 \times 10^{-7}$$

Augmented data matrix to account for the nonlinear relation

$$\tilde{x}_i = [x_i^J, x_i^{J-1}, \ldots, x_i].$$

Application 2: San Jose Network (speed data from 218 sensors, 1st week of October, 2022)

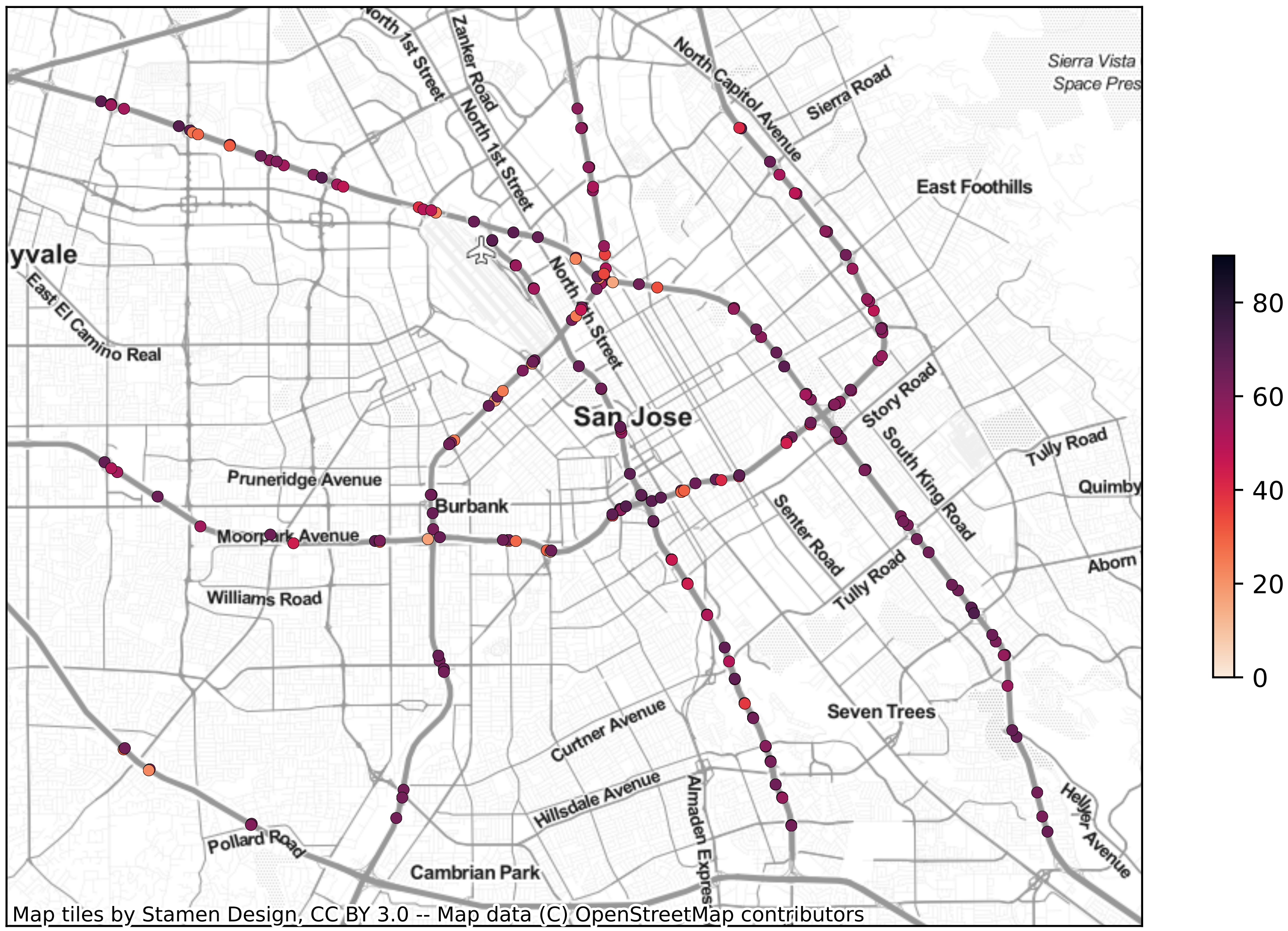


Figure 19: Afternoon snapshot of speed from sensors in downtown San Jose, California

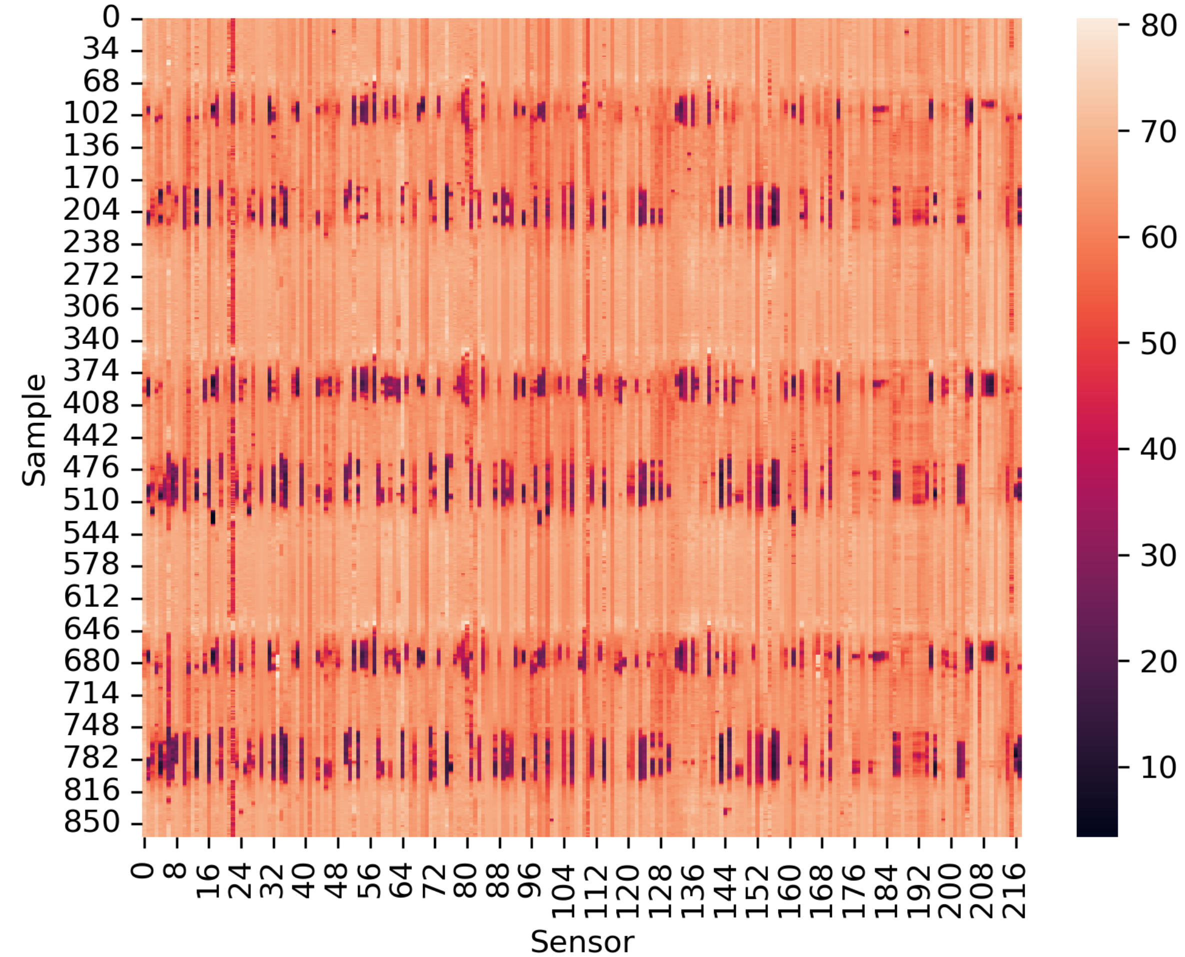


Figure 20: PEMS speed data in downtown San Jose

Performance of UARL in data reconstruction:

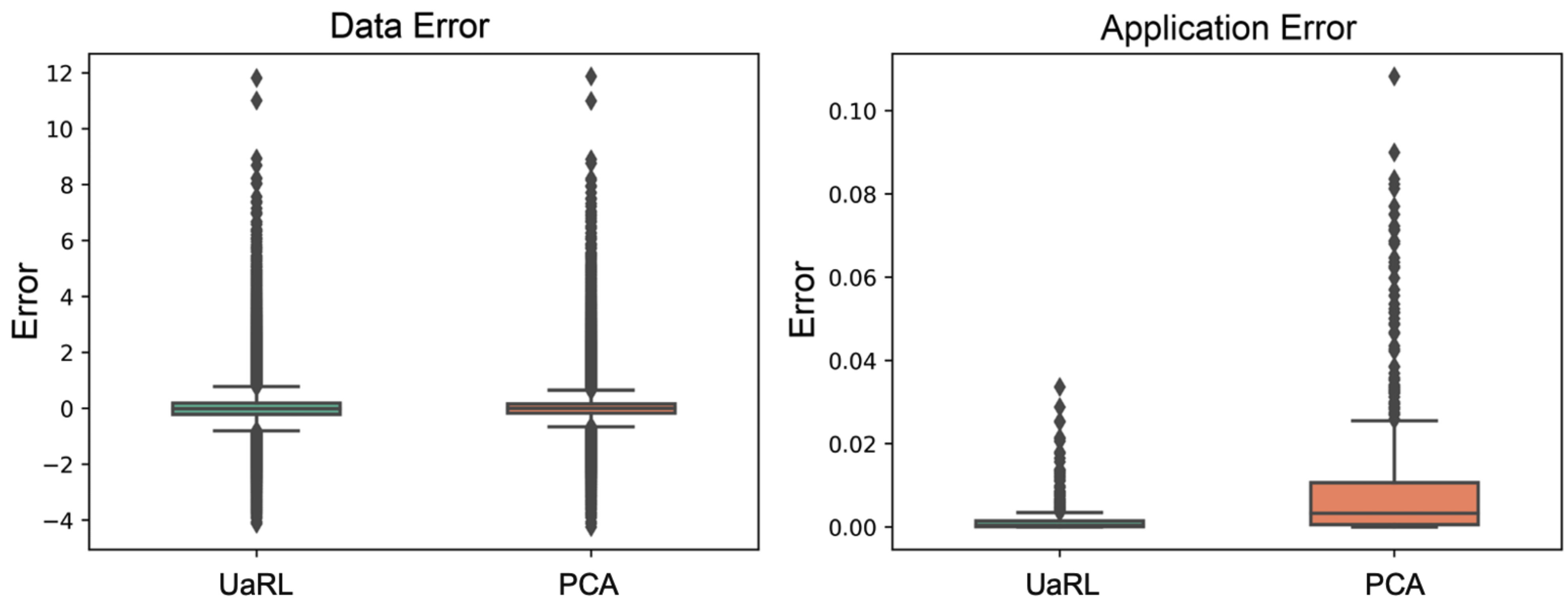


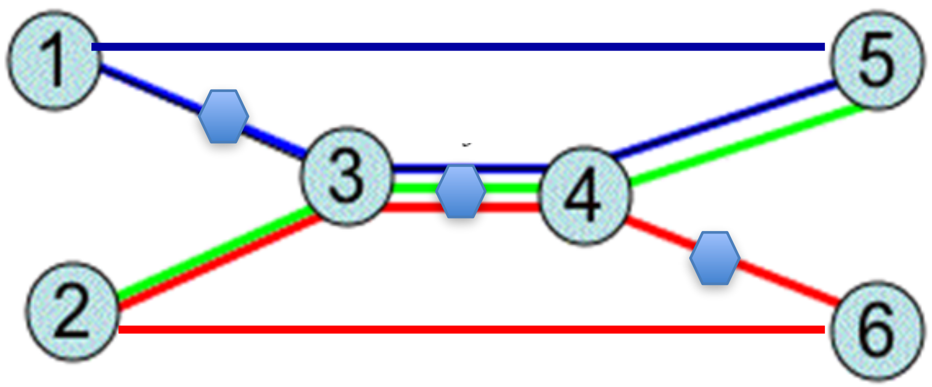
Figure 21: Emission reconstruction error UaRL vs. PCA

Example 2:
Travel Demand Estimation based on Sensor Data
(an example of reflecting domain knowledge in loss function and constraints)

More details are in:

1. R. Sun and Y. Fan (2024), Stochastic OD demand estimation using stochastic programming, Transportation Research Part B: Methodological, Volume 183.
2. Y. Yang, Y. Fan, R. Wets (2018), Stochastic Travel Demand Estimation: Improving Network Identifiability using Multi-day Observation Sets, Transportation Research Part B: Methodological, 107, 192-211.

A General Statistical Inference Problem over a Network Structure



$$y = G(x) \text{ or } y \in G(x)$$

where y represents link flows that are directly measurable;
x represents the travel demand to be inferred;
$G()$ represents a mapping, a static network assignment model here.

repeated measures on y

?

statistical inference of x

Main schools of thought at the time

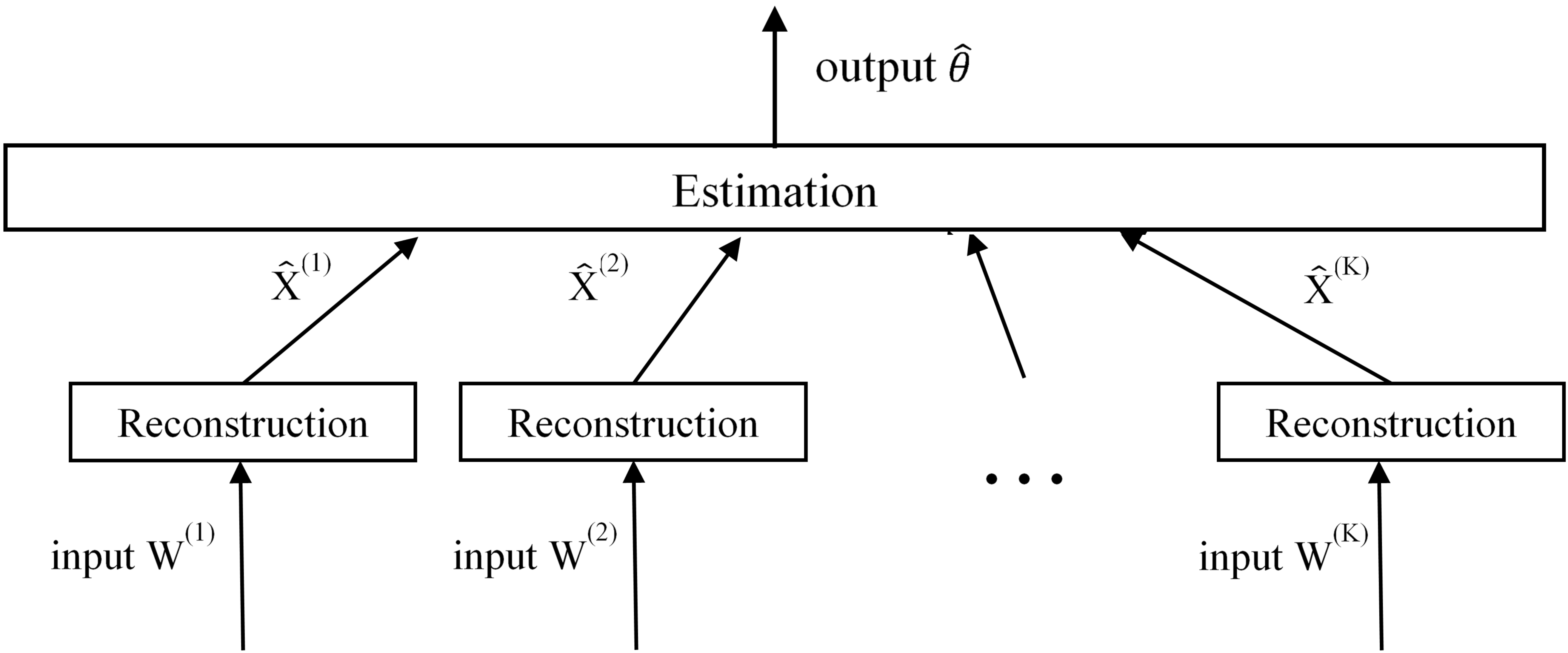
- Approach the problem as a deterministic inverse problem given a set of link flow observation

We know how demand x is distributed to network links $y = G(x)$.
Given an observation of y, what is the corresponding demand x?

- Approach the problem as a statistics problem given multiple observations of link flows

First compress multiple observations of y to its mean and variance. Then ask:
Given the statistics of y, what is the statistics of demand x knowing that $y = G(x)$?

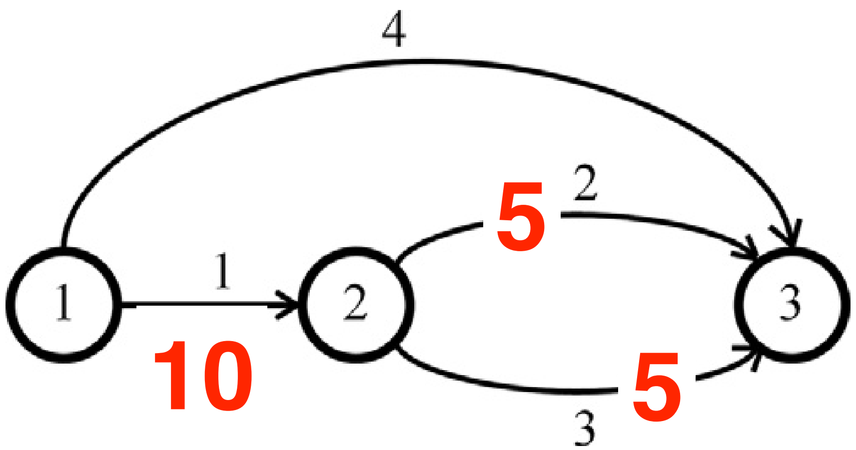
Estimation Framework based on Deterministic Inverse Problem



$$\mathbf{RCP}(w^{(k)}) : \min_{x \geq 0} s(v; w^{(k)}) \ \text{ s.t. } v \in G(x),$$

$$\mathbf{ESP}(\hat{x}^{(1)}, ..., \hat{x}^{(K)}) : \min_{\theta \in \Theta} \sum_{k=1}^{K} r(\theta; \hat{x}^{(k)}).$$

Issue with Deterministic Inverse Approach — Observability



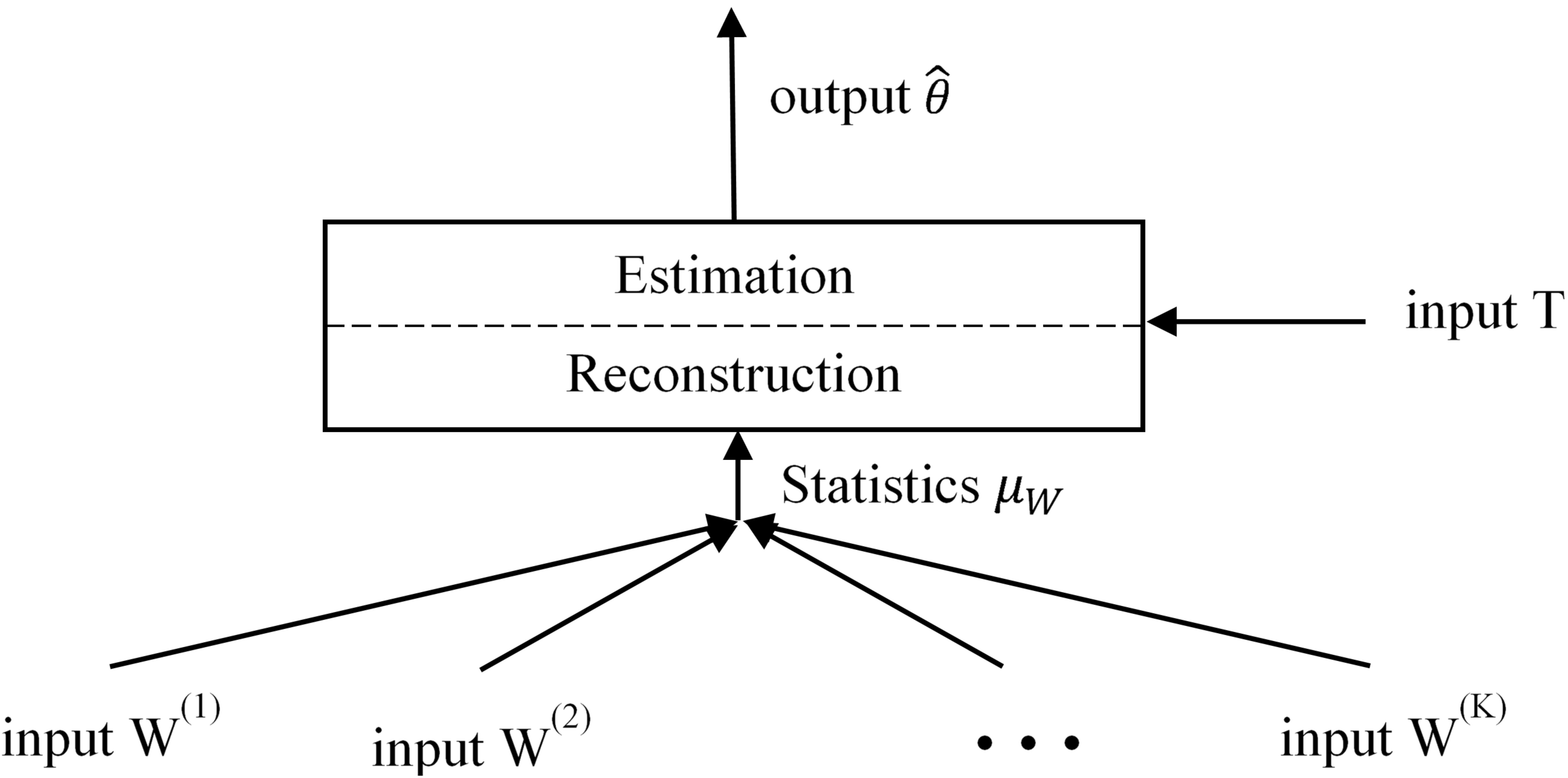
3 origin-destination pairs $x_1(1 \to 2)$, $x_2(1 \to 3)$, $x_3(2 \to 3)$.

We do not know whether:

(1) 10 units of flow goes from 1 to 3, or

(2) 10 units of flow goes from 1 to 2, and 10 units of flow goes from 2 to 3.

Estimation Framework based on Aggregated Statistics

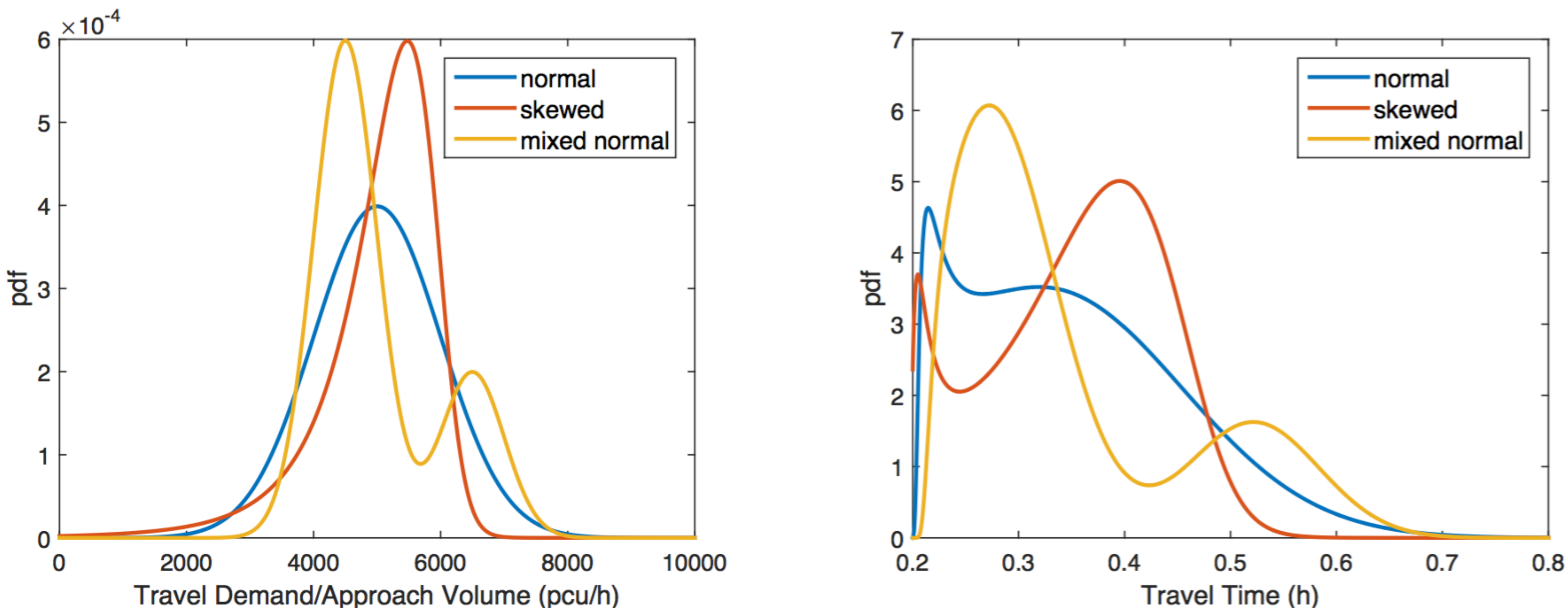


$$\mathbf{STA}(t, \mu_W) : \min_{\theta \in \Theta} h(\theta; t) + Ku(\mu_V; \mu_W) \;\text{ s.t. } \mu_V \in G(\theta),$$

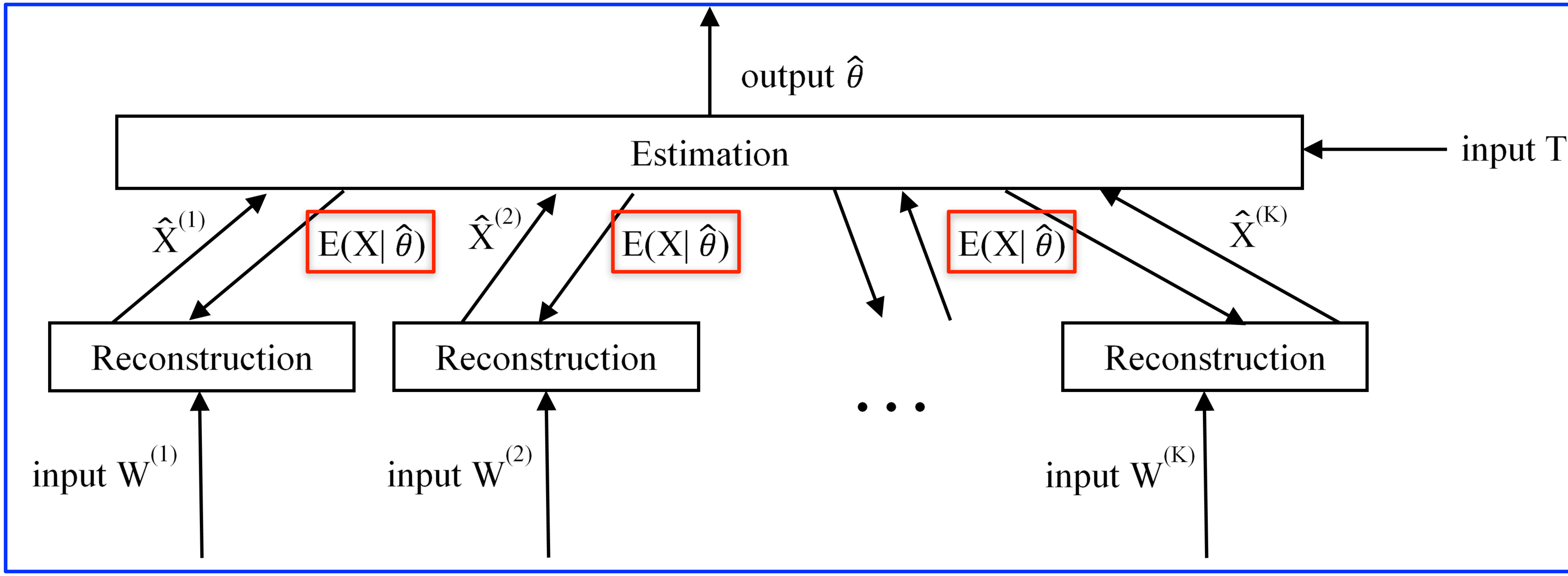
Issue with Aggregated Statistics – Loss of Information

Length = 20km. Designed capacity = 4000 pcu/h. Free flow speed = 100 km/h. Critical speed = 40km/h. $E(X) = 5000$, $\mathrm{Var}(X) = 10^6$.

- a normal distribution – 5% probability over critical travel time
- a skewed Pearson family distribution (skewness=-2.75, kurtosis=30) – 1% probability over critical travel time
- and a multimodal mixed normal distribution (mean1=4000, mean2=6500, variance1=variance2=2.5×10^5, mixture weights ratio=3:1) – 10% probability over critical travel time



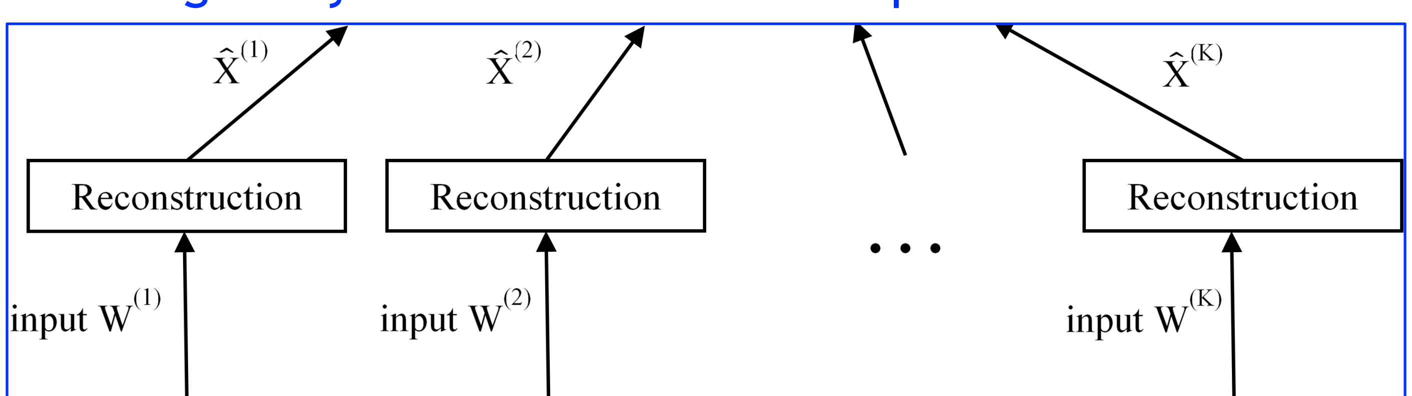
Proposed Estimation Framework



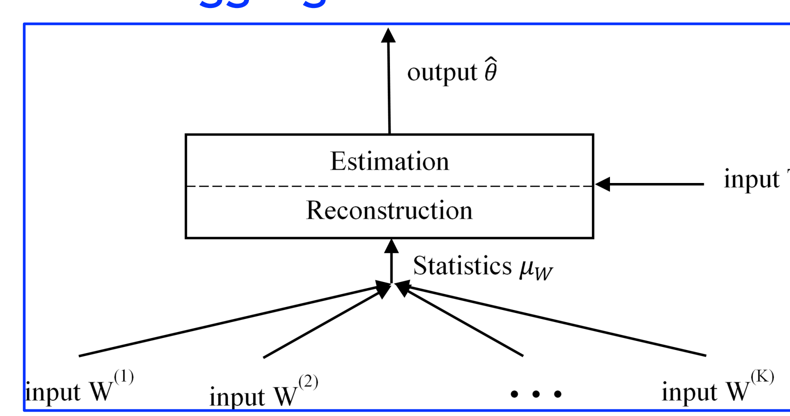
$$\mathbf{SDE}(t, w^{(1)}, \ldots, w^{(K)}) \quad \min_{\substack{\theta \in \Theta \\ x^{(k)} \geq 0,\, k=1,\ldots,K}} \quad \underbrace{h(\theta; t)}_{\text{prior}} + \sum_{k=1}^{K} s(v^{(k)}; \underbrace{w^{(k)}}_{\text{sensor data}}) + \sum_{k=1}^{K} r(\theta, \underbrace{x^{(k)}}_{\text{population statistics}})$$

$$\text{s.t. } v^{(k)} \in G(x^{(k)}), \forall k = 1, \ldots K.$$

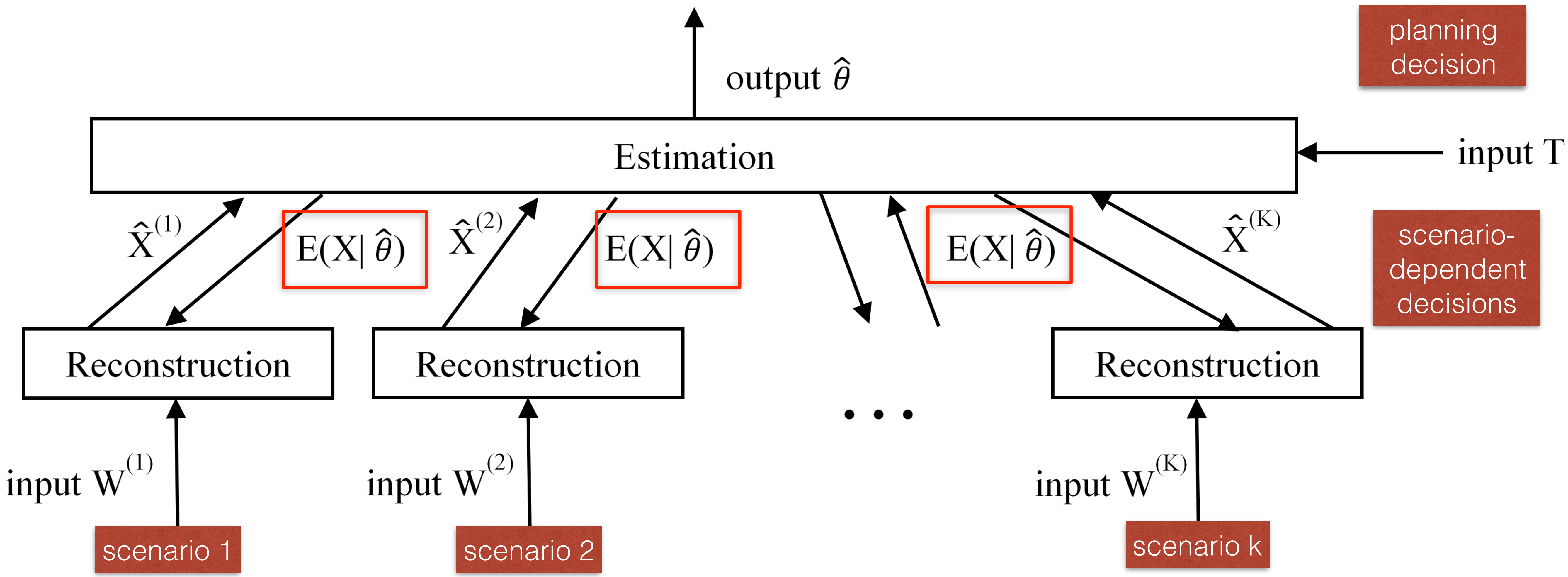
Solving many deterministic inverse problems



Aggregated Statistics



Connection to Stochastic Optimization

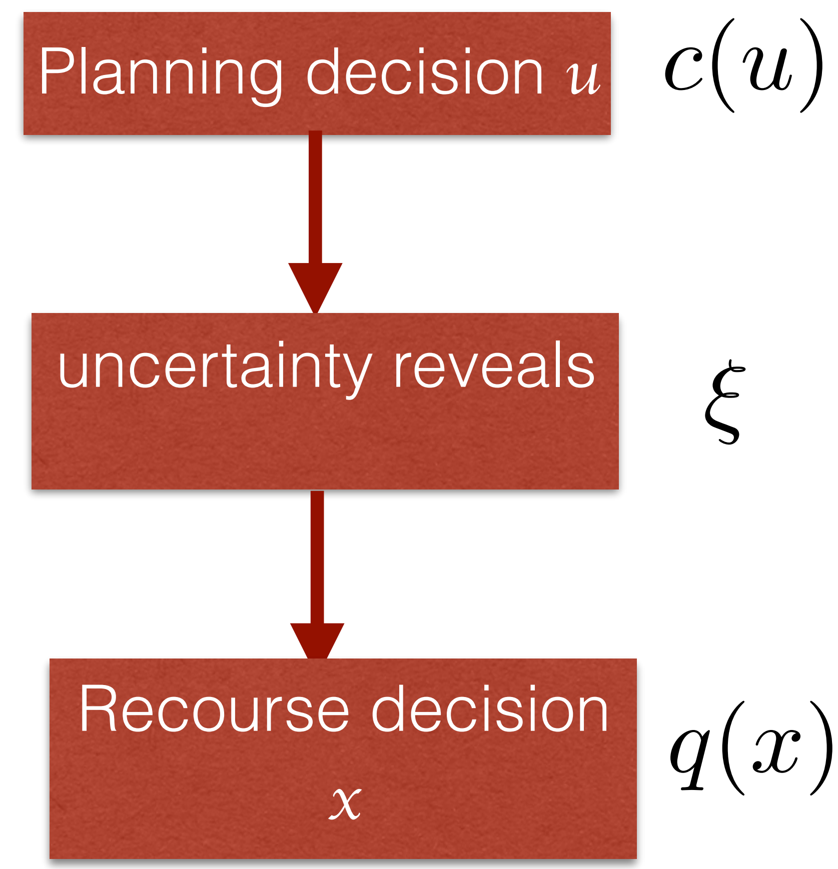


A Two-stage Stochastic Programming Framework

$$\min_{u} \boxed{c(u)} + \boxed{E_{\xi}\{Q(u,\xi)\}}$$

$$\boxed{s.t.\quad u \in U}$$

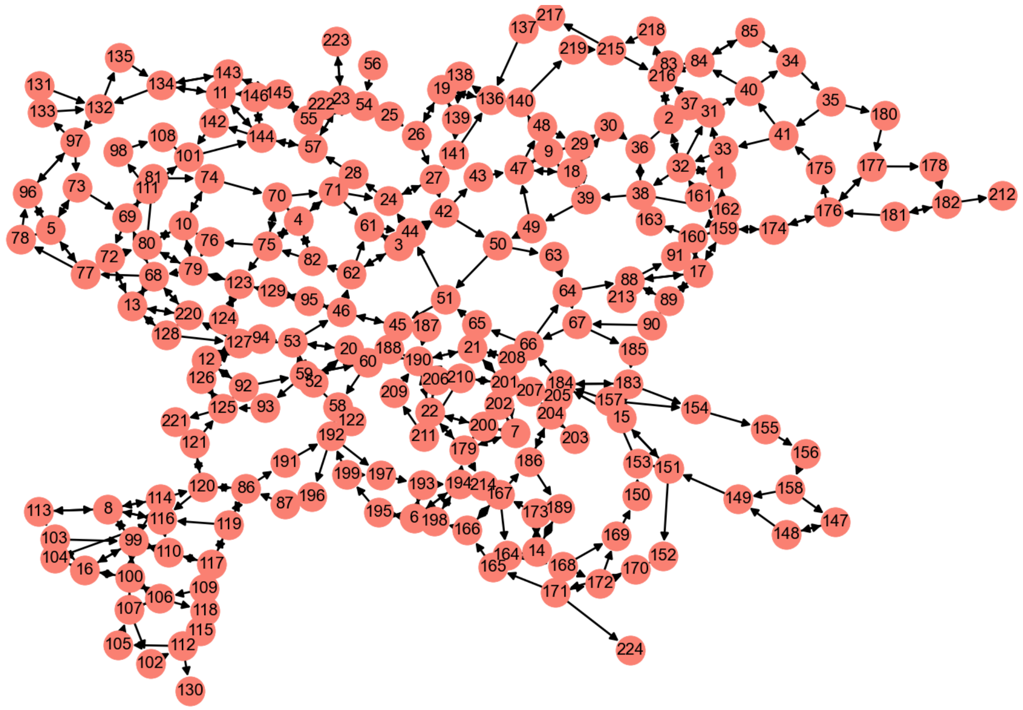
$$\boxed{Q(u,\xi) = \min_{x} q(x) \quad s.t. \quad x \in X(u,\xi)}$$



$$\min_{\substack{\boxed{\theta\in\Theta} \\ x^{(k)}\geq 0,\ k=1,\dots,K}} \boxed{h(\theta;t)} + \boxed{\sum_{k=1}^{K} s(v^{(k)}; w^{(k)}) + \sum_{k=1}^{K} r(\theta, x^{(k)})}$$

$$\boxed{\text{s.t. } v^{(k)} \in G(x^{(k)}), \forall k = 1, \dots K.}$$

Test Case 1: Berlin Friedrichshain Network



Network Size: 224 nodes, 523 links, 529 O-D pairs

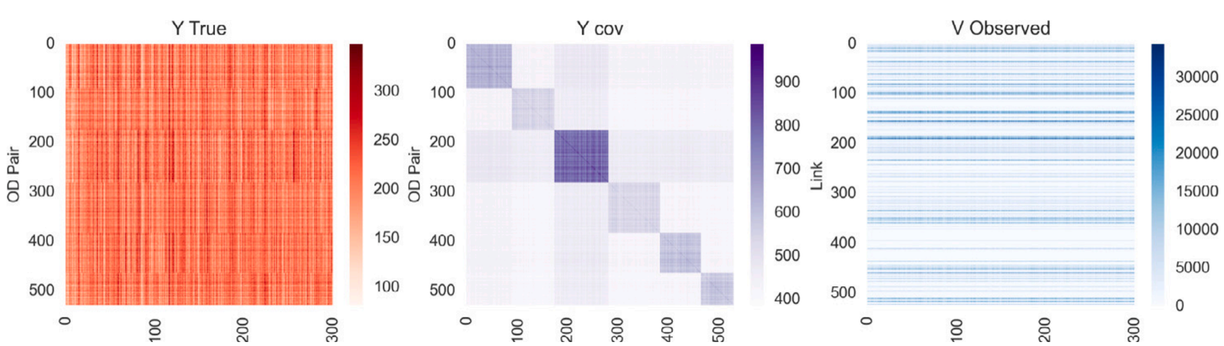
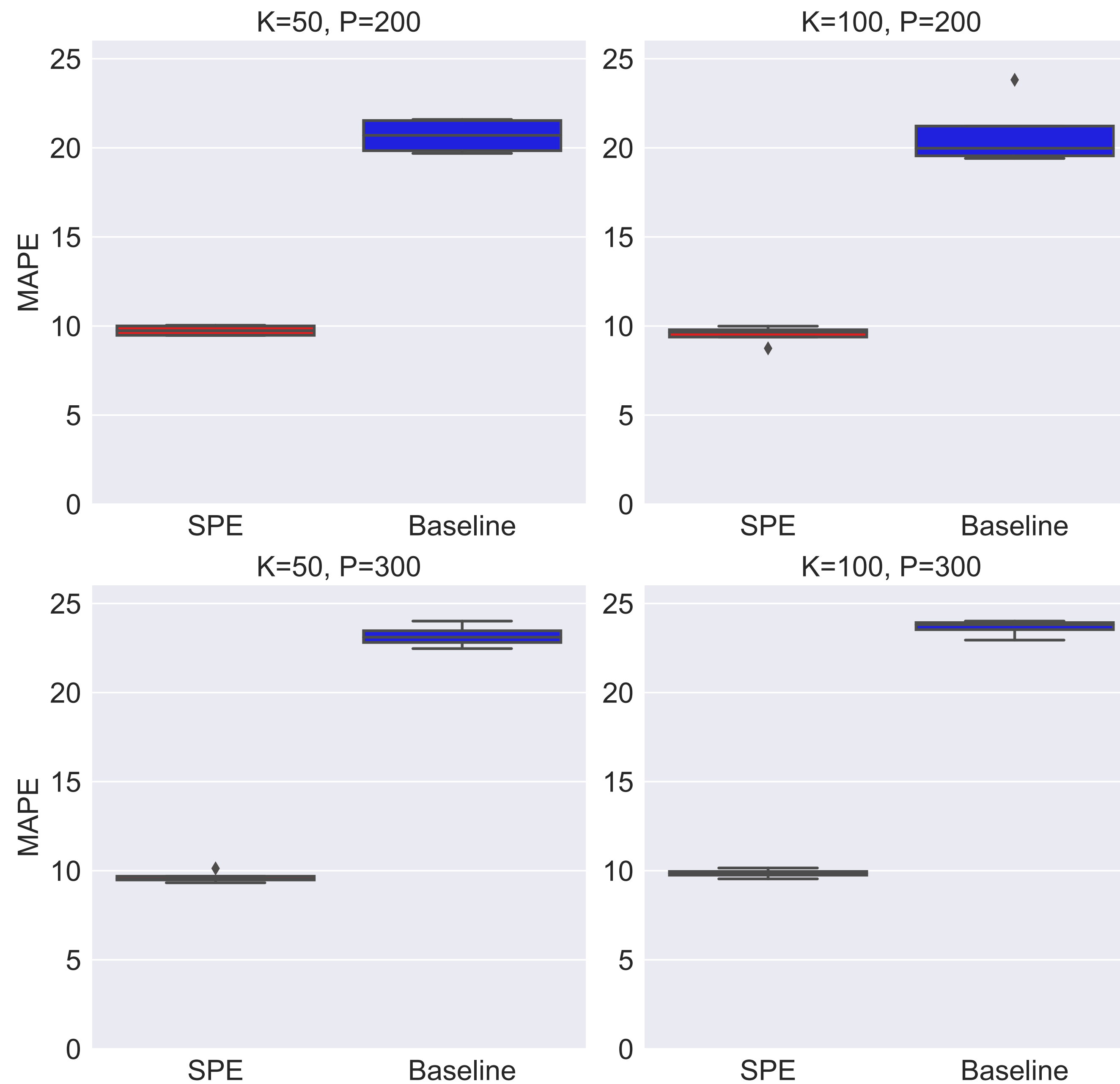
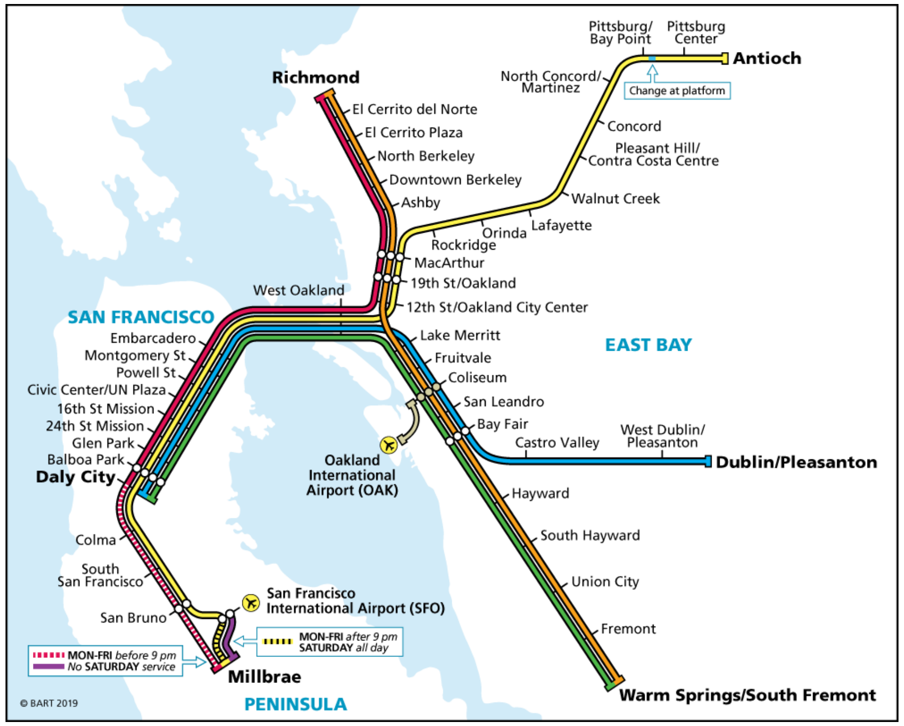


Fig. 2. True OD demand, covariance of OD demand, and observed link flow.



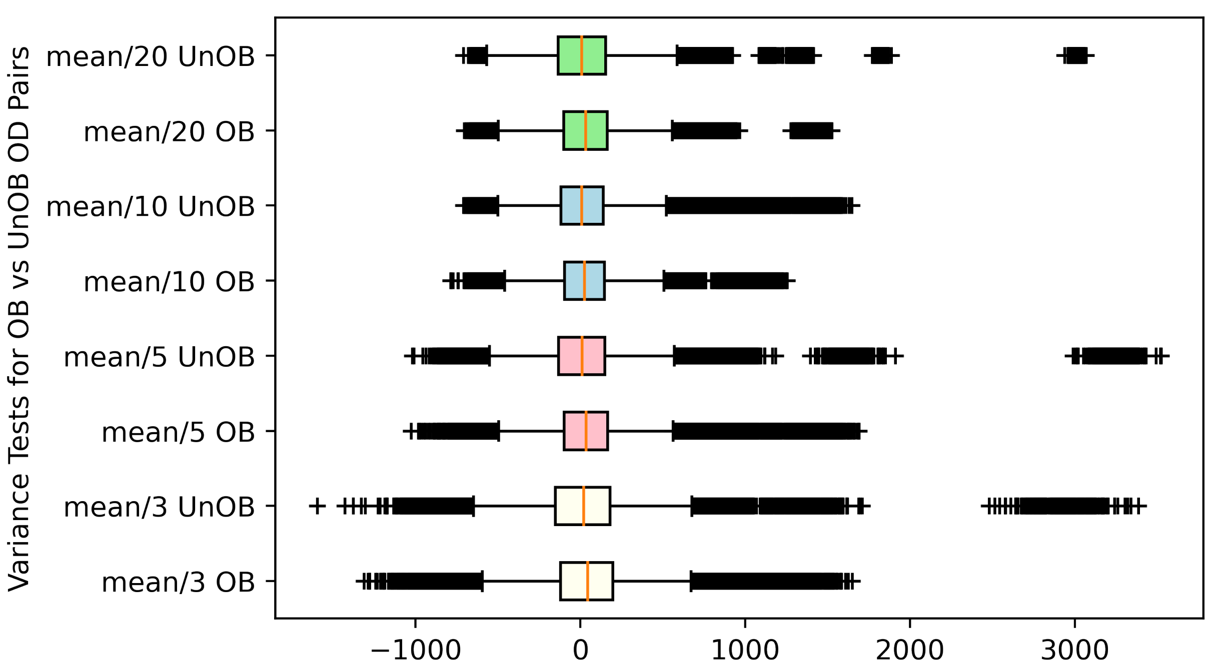
Mean Absolute Percentage Error of the mean estimates
(K: # of data scenarios; P: # of O-D pairs)

Test Case 2:Transit demand estimation based on APC data

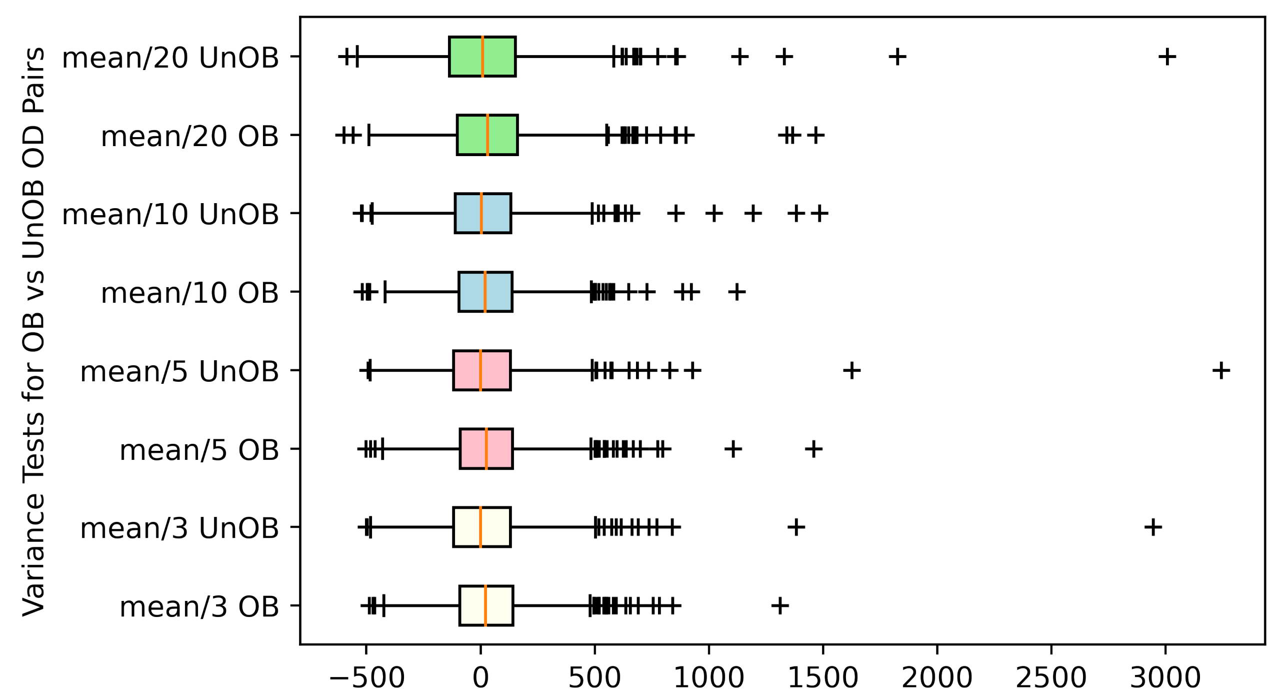


San Francisco BART network
(181 directed links, 1464 route section links, 2256 OD pairs)

	Std.	$\sigma_i = 0$	$\sigma_i = \mu_i/20$	$\sigma_i = \mu_i/10$	$\sigma_i = \mu_i/5$	$\sigma_i = \mu_i/3$
Run time (seconds)	K = 100	299.90	312.67	299.06	345.09	340.41
	K = 250	807.08	816.46	802.16	857.27	918.37
	K = 500	2200.40	2079.79	2137.58	2210.18	2321.05
RMSE	K = 100	216.14	229.08	237.17	251.58	275.44
	K = 250	237.33	242.50	242.42	231.57	260.46
	K = 500	228.52	226.96	220.56	222.18	260.77



Comparison between true and reconstructed demand



Comparison between true mean demand and estimated mean demand

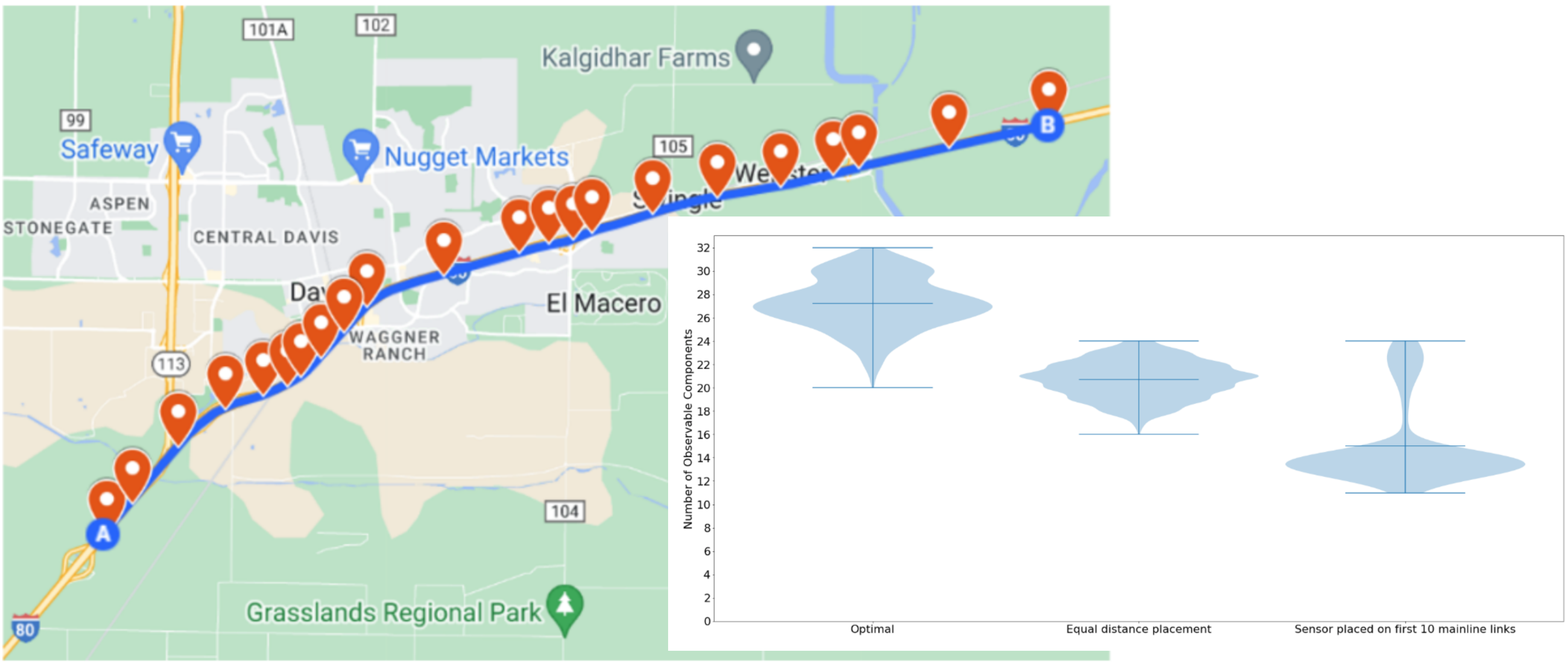
Example 3:
Information Acquisition for Estimating Traffic Dynamics
(an example of reflecting domain knowledge in data preparation)

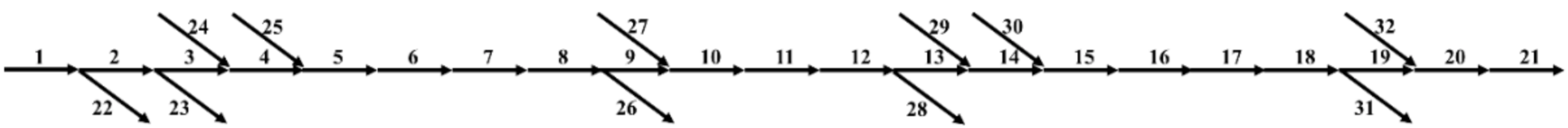
More details are in:

1. X. Hu, Y. Fan, Sensor placement considering the observability of traffic dynamics: On the algebraic and graphical perspectives, Transportation Research Part B: Methodological, Volume 189, 2024.

Motivations for Information Acquisition Design

Google Maps Network
21 mainline links, 11 ramps (32-link network)

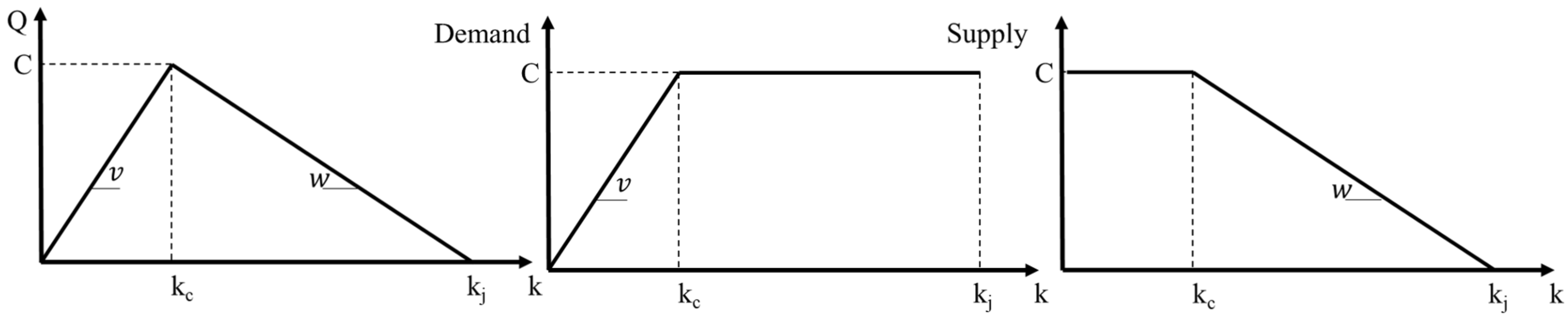




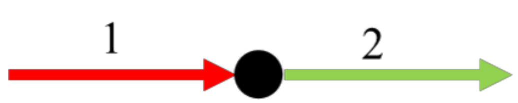
10 sensors are placed. 3 different strategies are compared: 1. Optimal observability 2. Simply placing sensors on first 10 mainline links 3. Equal distance placement on the mainline

Transportation Domain Knowledge

Traffic network density modeled as a dynamical system



Fundamental diagram and demand/supply of a link

	**out-flux** $\boldsymbol{g_1}$ **= in-flux** $\boldsymbol{f_2}$ **= min(**$\boldsymbol{d_1}$, $\boldsymbol{s_2}$**)**
Mode 1 Link 1,2 free flow	$\begin{bmatrix} \dot{k}_1 \\ \dot{k}_2 \end{bmatrix} = \begin{bmatrix} -v_1 & 0 \\ v_1 & 0 \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} + \begin{bmatrix} f_1 \\ -g_2 \end{bmatrix}$
Mode 2 Link 2 congested	$\begin{bmatrix} \dot{k}_1 \\ \dot{k}_2 \end{bmatrix} = \begin{bmatrix} 0 & w_2 \\ 0 & -w_2 \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} + \begin{bmatrix} f_1 - w_2 k_{j,2} \\ -g_2 + w_2 k_{j,2} \end{bmatrix}$

A switching-mode system: $\dot{\boldsymbol{x}}(t) = A(\lambda_t)\boldsymbol{x}(t) + \boldsymbol{b}(\lambda_t) + \boldsymbol{u}(t)$

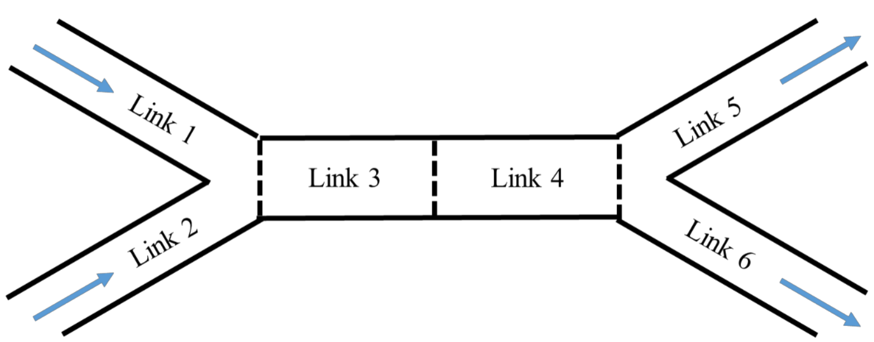
$A(\lambda_t)$ and $\boldsymbol{b}(\lambda_t)$are **mode-specific**

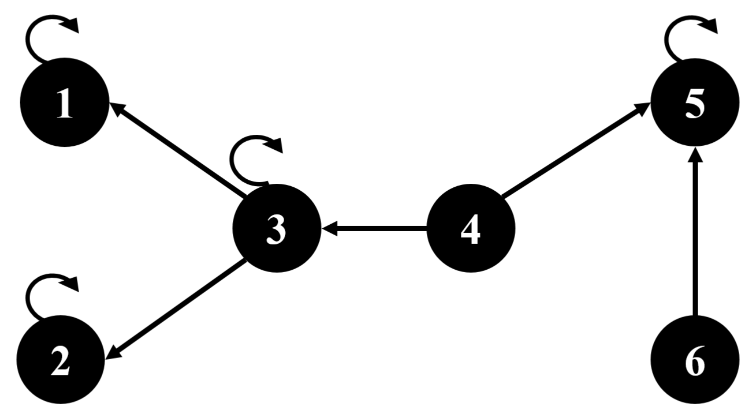
Observation model (state to measurement) **:** $\mathbf{y}(t) = C\boldsymbol{x}(t)$

Graphical Approach for Structural Observability

Inference Diagram

- Each link density is represented as a node in a graph
- Draw a directed edge from node i pointing to node j if $a_{ij} \neq 0$
- Ex. $a_{31} \neq 0$ ③→① $a_{22} \neq 0$ ②→② (***self-edge, loop***)





$$A = \begin{bmatrix} a_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & a_{22} & 0 & 0 & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 & 0 & 0 \\ 0 & 0 & a_{43} & 0 & a_{45} & 0 \\ 0 & 0 & 0 & 0 & a_{55} & 0 \\ 0 & 0 & 0 & 0 & a_{65} & 0 \end{bmatrix}$$

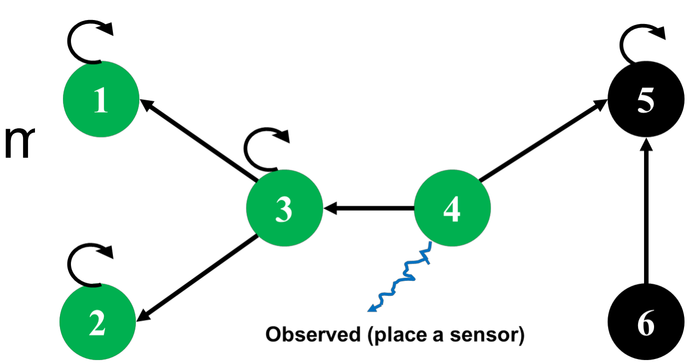
Special about traffic dynamics: existence of **self-edges** due to **flow conservation**

Observable subsystem identification

Any reachable nodes from observed nodes are also (structurally) observable.

The result identifies observable subsystem in a partially observable system

➢ Sensor located on ④ → ③ ② ① observable (indicated in green)



A Stochastic Optimization Model for Maximizing Observability

A mixed-integer program formulation considering various modes

$$\text{let } z_i = \begin{cases} 1, \text{if a sensor is put on link } i \\ 0, \text{otherwise} \end{cases}$$

$$\max_{z,x} \sum_{k=1}^{K} w_k \sum_{i=1}^{n} x_{k,i}$$

$$s.t. \sum_i z_i = p$$

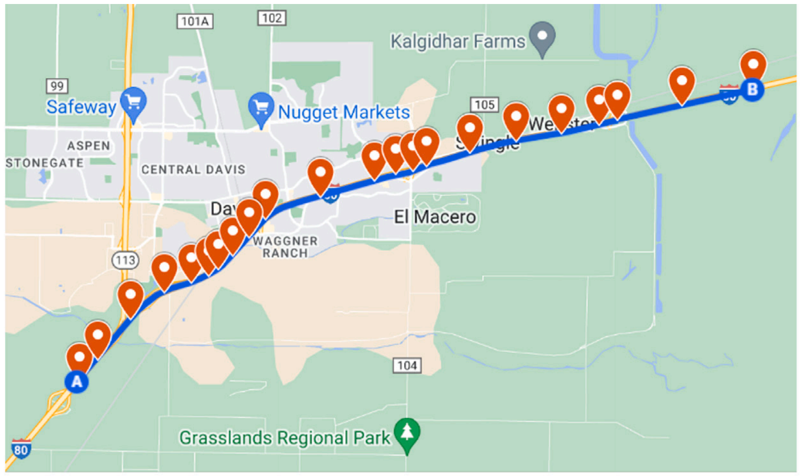
$$x_{k,i} \leq \sum_{(i^-,i)\in\delta_k^-(i)} x_{k,i^-} + z_i, \forall k, i$$

$$z_i \in \{0,1\}, \forall i$$

$$x_{k,i} \leq 1, \forall k, i$$

- Locate p sensors on n links
- $i = 1,2, \dots, n$ and $k = 1,2, \dots, K$
- Total number of modes: K
- $\delta_k^-(i)$: set of incoming edges of node i in the inference diagram for mode k (excluding self-edges)
- w_k: weight factor for mode k
- $x_{k,i}$: will be pushed to be binary; =1 if link i is observable in mode k

Experiment 2: I-80 East (Davis - Sacramento)



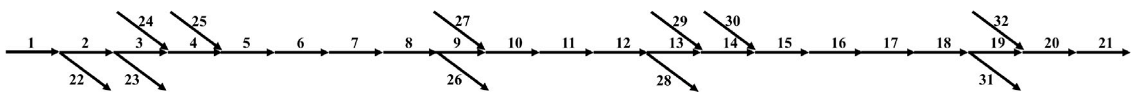


Fig. 12. I80 East Davis–Sacramento.

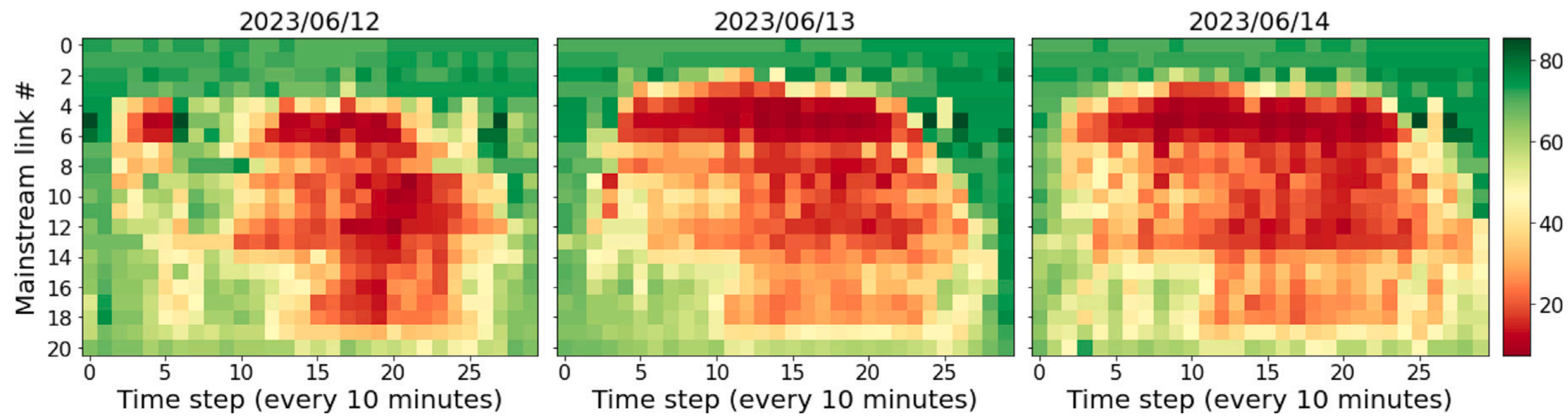


Fig. 13. Speed profile of I80 East evolution in time.

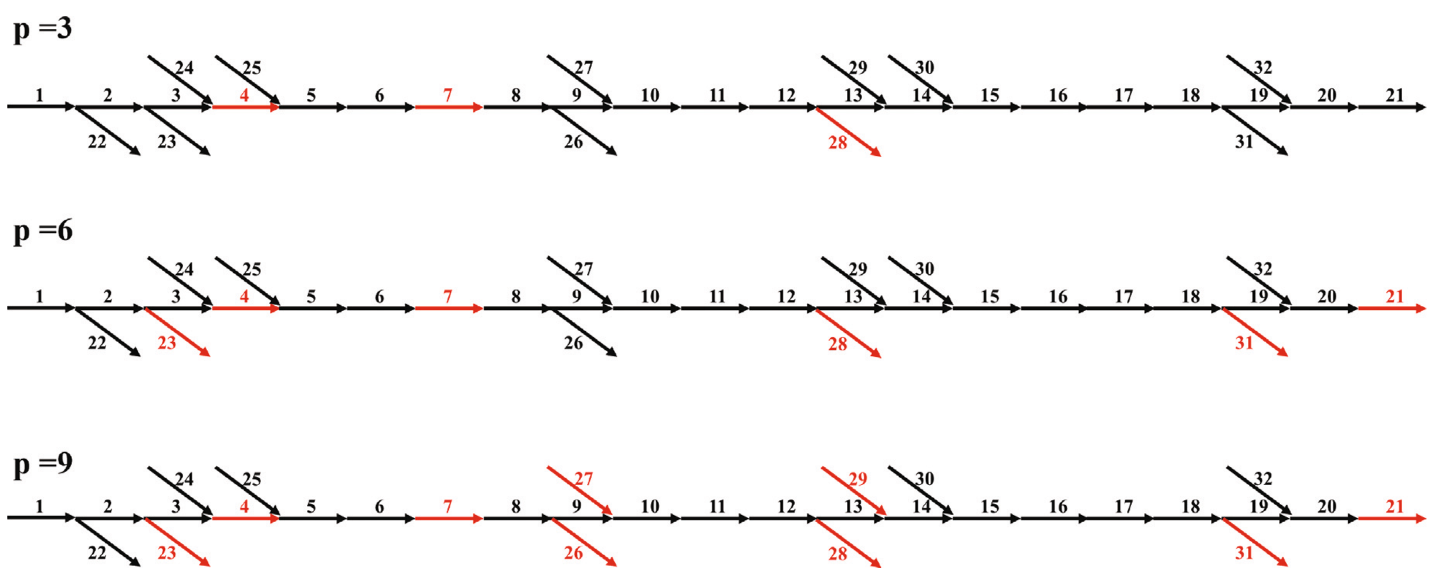


Fig. 14. Optimal sensor locations for varying numbers of sensors to be deployed.

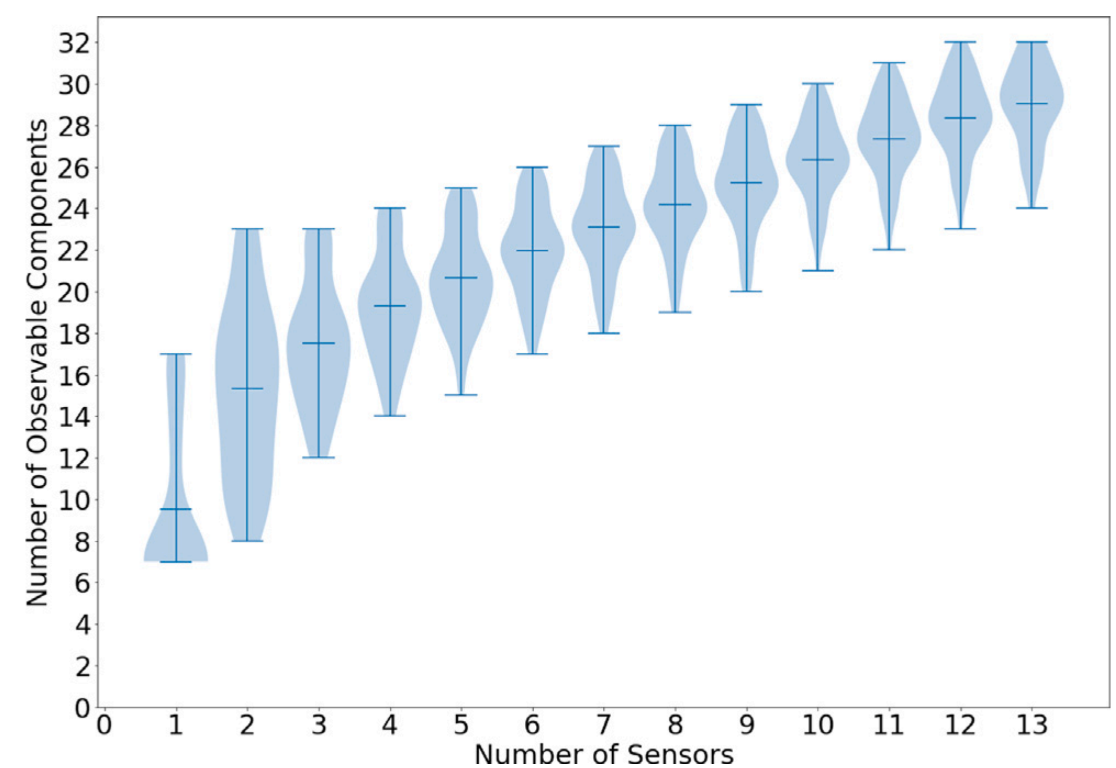


Fig. 15. Violin plot for observability under different numbers of sensor to be deployed.

Discussion

- Position AI in the broad AI family - constrained optimization provides a flexible modeling framework to integrate domain knowledge and hard data/information
- Networked data pose unique data and analytics challenges but also present unique opportunities for domain experts
- AI tools have lowered disciplinary barriers for transportation researchers to benefit from the technologies, but strategic and collective actions are needed for our field to fully exploit and stay at the forefront of the emerging technologies.

data science
machine learning

domain
knowledge

Acknowledgements

- Funding support by National Science Foundation of the United States on grants CMMI 1538263 and CMMI 1825873