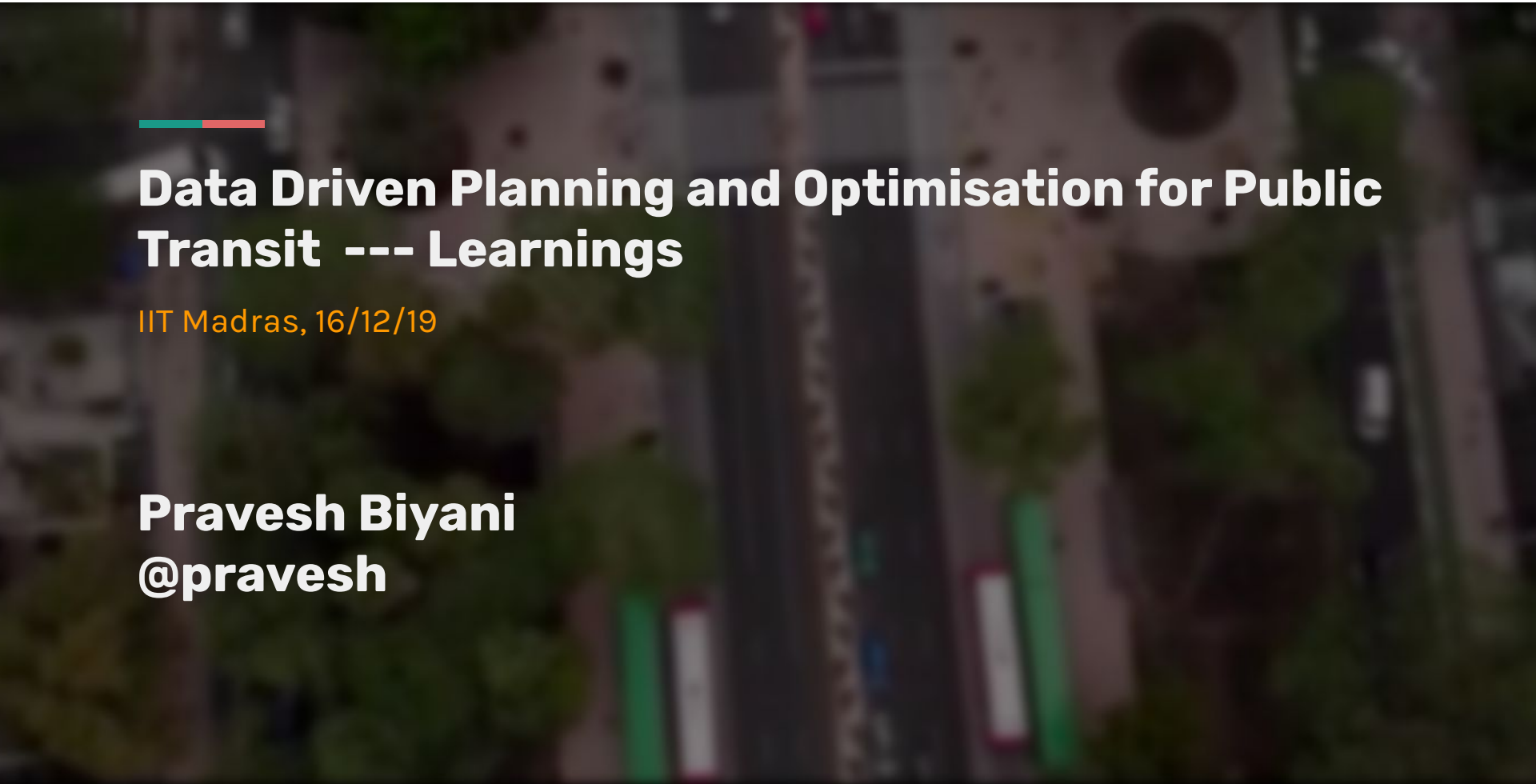


Data Driven Planning and Optimisation for Public Transit --- Learnings

IIT Madras, 16/12/19

Pravesh Biyani
@pravesh

My experience

Mainly with Govt owned public transit agencies

- DTC (Delhi)
- DIMTS (Delhi)
- BMTC (Bengaluru)
- Gurugaman (Gurugram)
- Indore, Bhopal (BCLL), Raipur

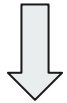
Observations:

1. Severe lack of capacity, specially at IT, ITS and data layers.
2. Data is either not collected or the agency not in control of data
3. The capacity to use data is zero.
4. Fewer avenues to fund projects done by academia
5. Tendering process is completely broken

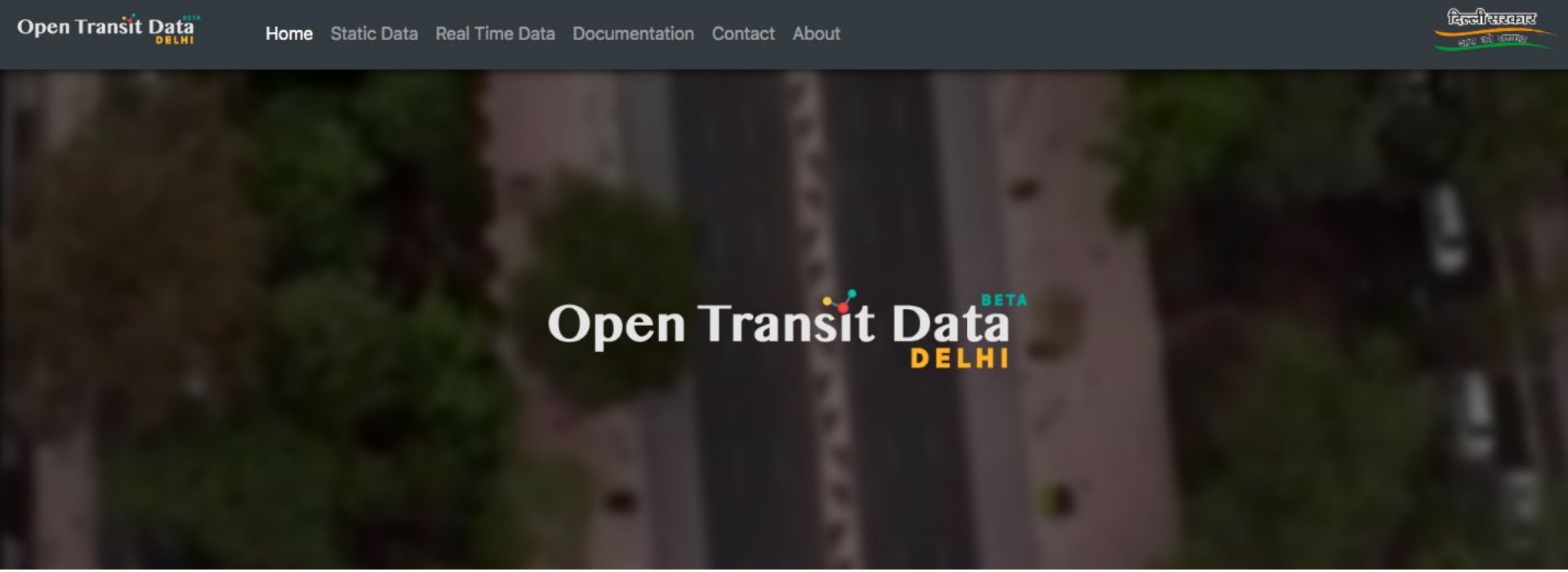
So what can we do? -> Open the transit Data !!

1. Act as a technology partner to transit agencies
2. Bridge between transit agencies and student/developer community
   a. Water == data
3. Data → leads to funding from other sources → reduces the tendering problems to a certain degree

Opening the transit data for everybody to use is the best fix for all problems



Leverage worldwide academia and network of enthusiastic IT/Data Professionals



We have created Open Transit Data for around 2300 (cluster/orange) buses in Delhi

Kailash Gahlot
@kgahlot

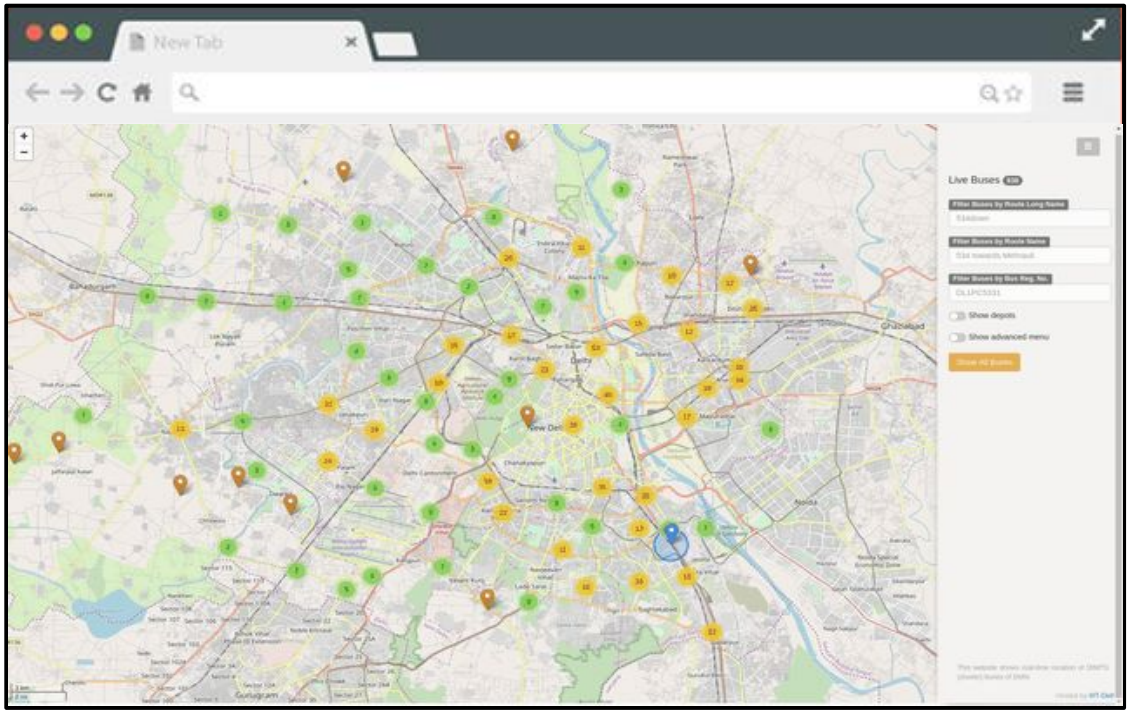
Delhi's Open Transit Data platform provides free of cost static and real time datasets of Delhi's buses for app developers and researchers in machine-readable format. Call out all researchers and developers to join us in transforming public transport 2/2
otd.delhi.gov.in

2:56 PM · Nov 23, 2018 · Twitter for iPhone

MyBus Dashboard (mybus.chartr.in

A real-time dashboard to monitor bus movements all over Delhi.

- Built on **Open Transit Data for cluster buses** in Delhi
- Used for
  - Real-time bus bunching mitigation
  - Bus breakdown identification
  - Prediction
  - Scalable to any GTFS data



Data from Transit Agencies -- Main Features

1. Missing data → requires imputation

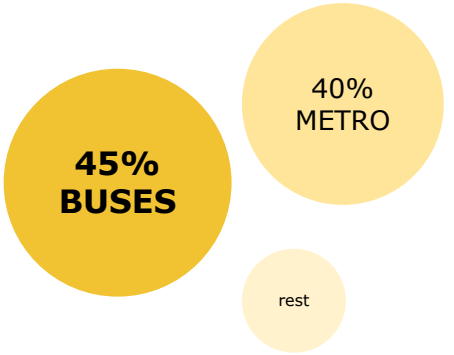
2. Noisy → Require both preprocessing and post processing depending on the task

3. Not Big → Sparse and incomplete

Open Transit Data: Why?

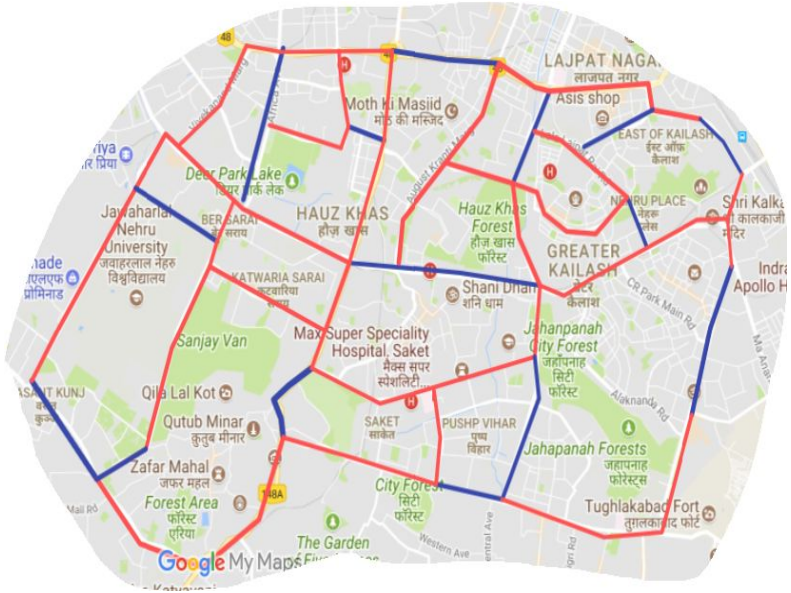
- Similar numbers for other cities like Kolkata, Mumbai, Bangalore..
- Low ridership in buses and poor IT infrastructure
  - Almost no information to passengers
  - Highly suboptimal and inefficient system
- Lack of information to passengers
  - Passengers abandon mass transit and take lower capacity vehicles
  - More information lead to passenger satisfaction

Indians using public transit.

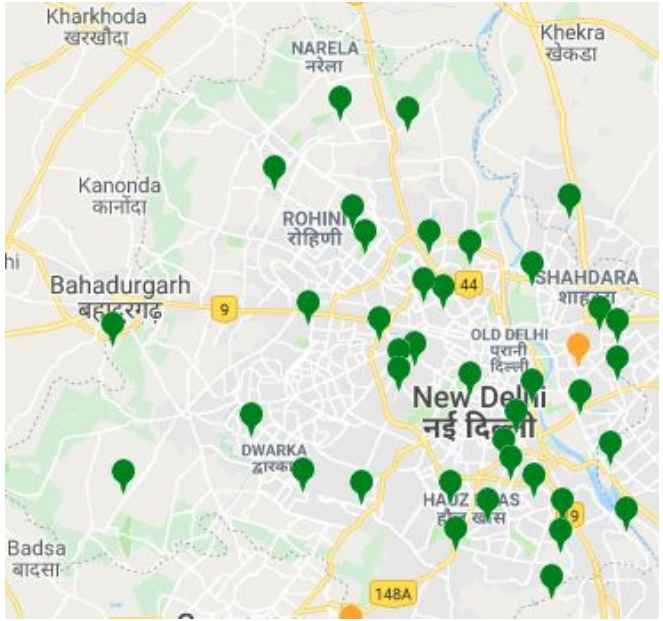


THE FIRST STEP TOWARDS MEASURING IS COLLECTION OF DATA.

Problem 1: Missing Data

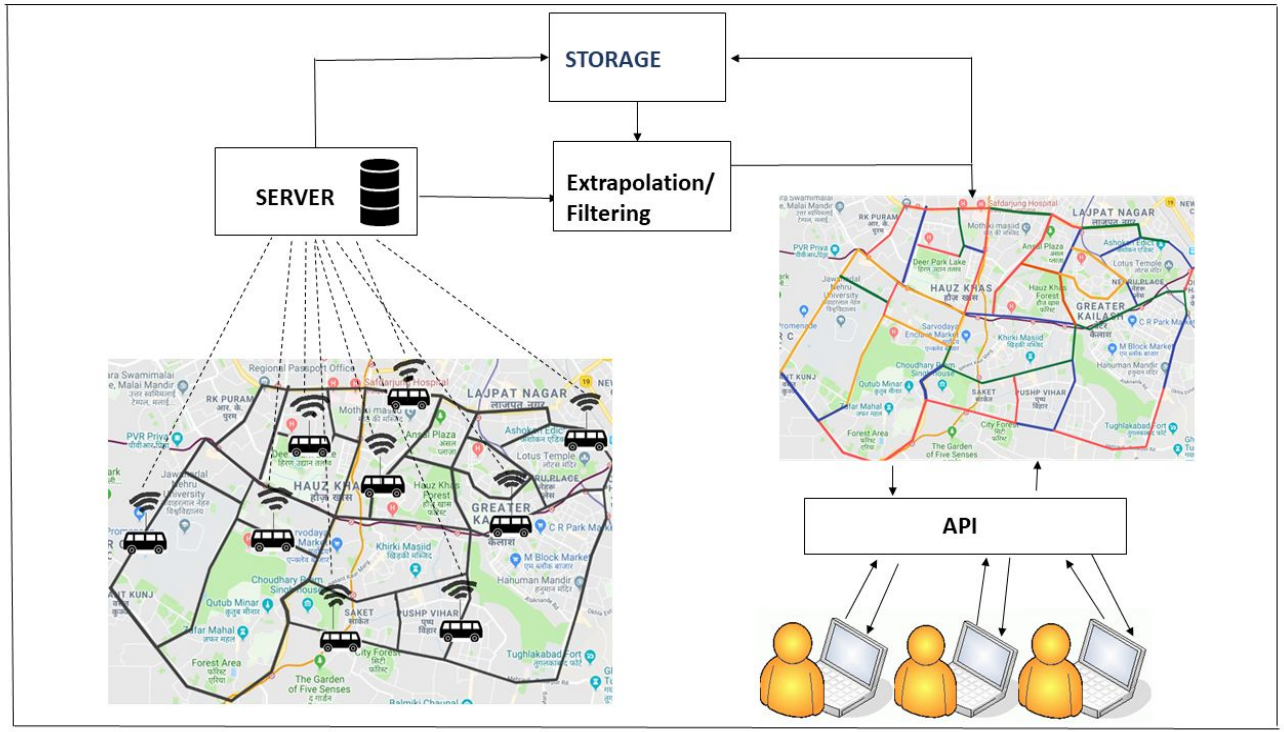


Google API data



Air Quality data (cpcb delhi, china)

Real Time Spatio-Temporal Prediction with Missing Data



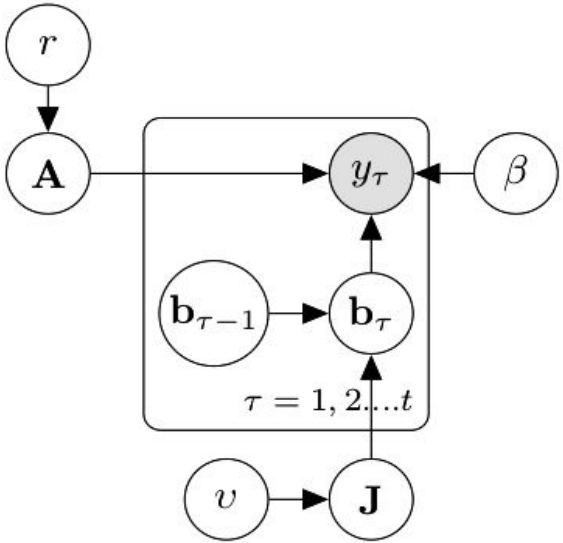
Solution: Variational Bayesian Subspace Filtering

- Low rankness of the data is captured by the equation

$$y_t = A\, b_t + n$$

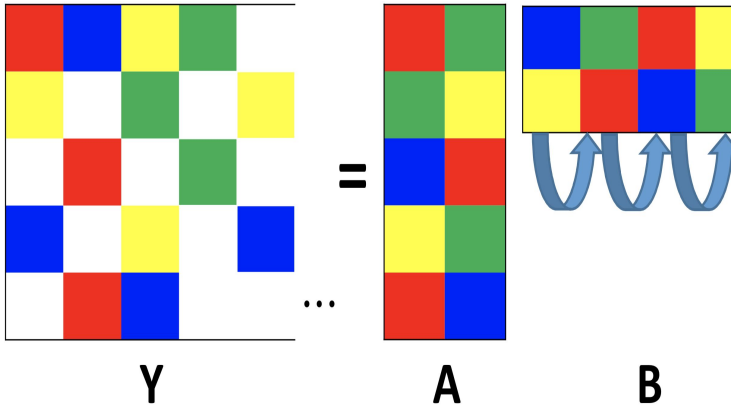
- Further the temporal correlation can be modelled by the equation

$$b_t = J\, b_{t-1} + n$$



Variational Bayesian Subspace Filtering

- Generative model for noisy and incomplete matrix **Y** whose columns arrive sequentially over time.
- VB to learn model parameters.
- Low-rank matrices whose underlying subspace evolves according to a state-space model.
- The key algorithm parameters like rank and various noise power need not be fine-tuned and are learned automatically.



Variational Bayesian framework

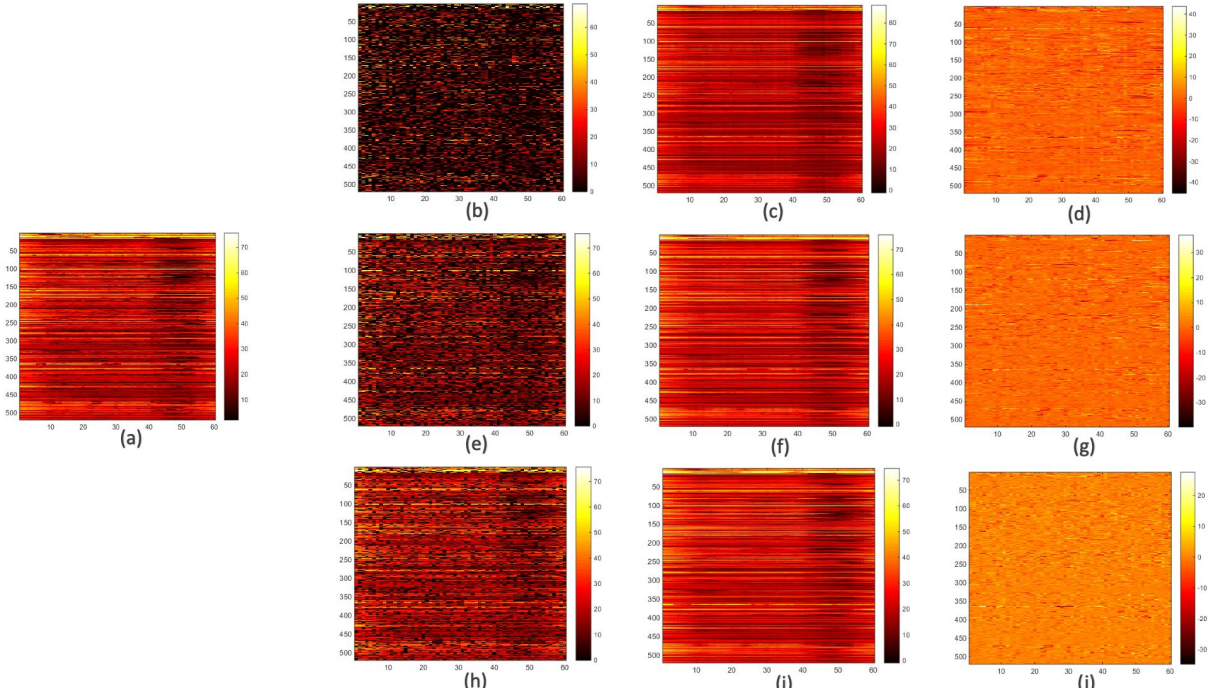
$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$
$$p(y) = \int p(y, \theta)\, d\theta$$

- No closed form, high-dimensional integration
- Solution Variational Bayesian Inference
- Approximate posterior $p(\vartheta|y)$ with q^*

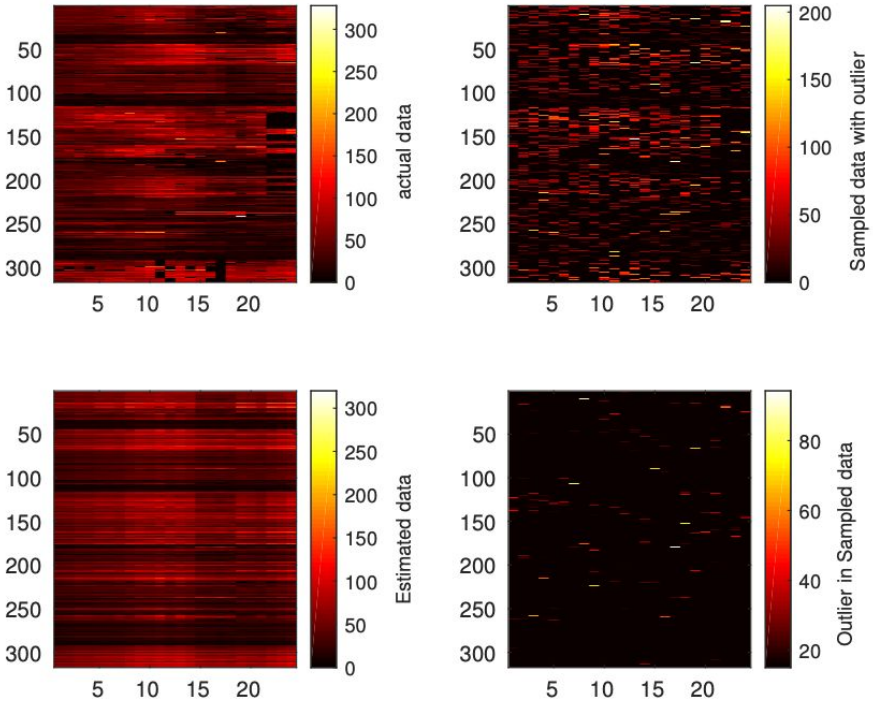
$$q^* = argmin_{q \in Q} f(q(.), p(.|y))$$

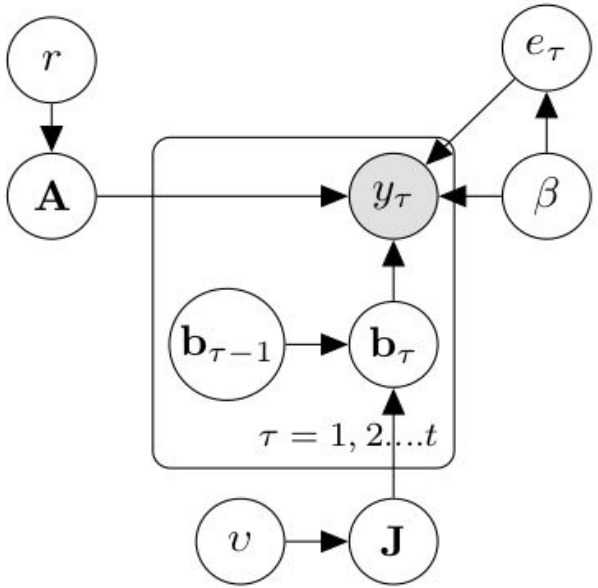
- In Variational Bayes (VB): f is Kullback-Leibler divergence

Results



Robust VBSF for outlier detection (Air Quality Estimation)





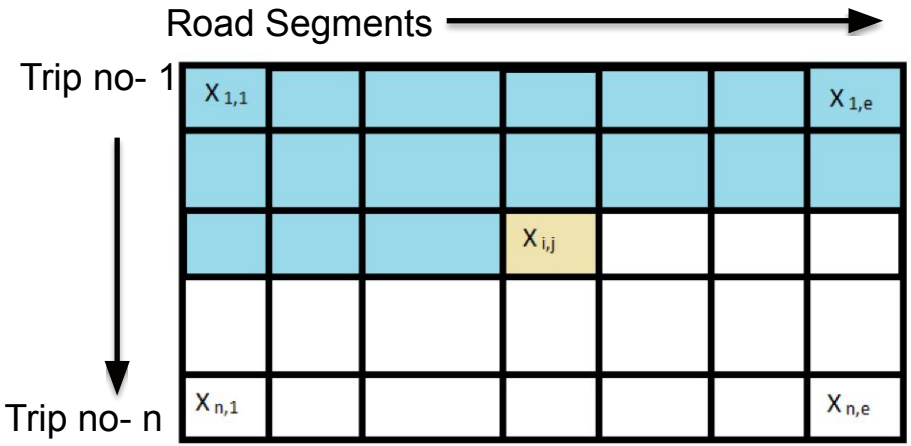
Problem 2: Big Data not available

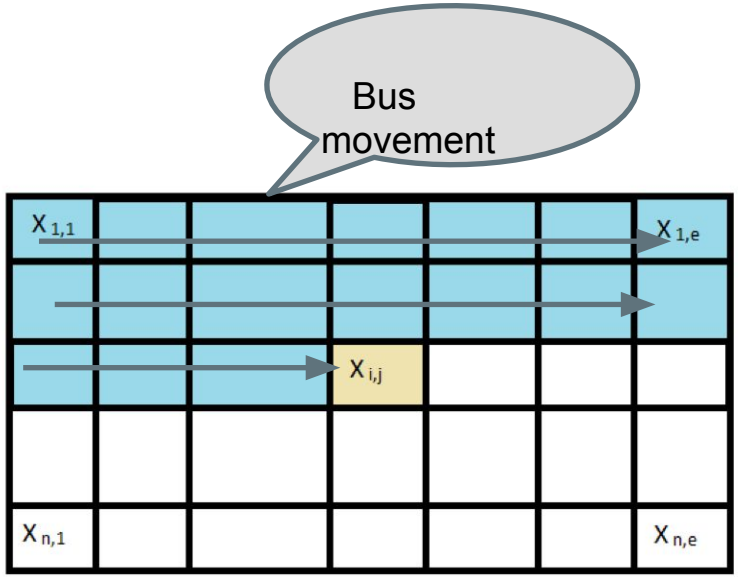
ETA Prediction Problem

Real time bus data (otd.delhi.govt.in)

Given historic ETA data set and the current location of the bus in a trip, the problem is to “predict” in real-time the ETA for all the remaining stops in the trip

ETA Data as a Matrix

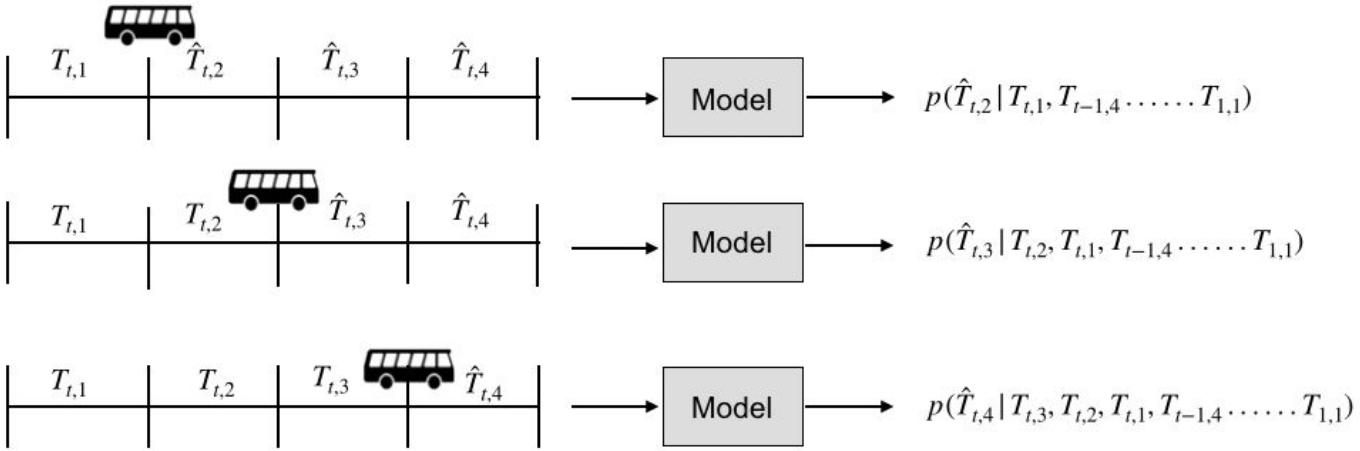




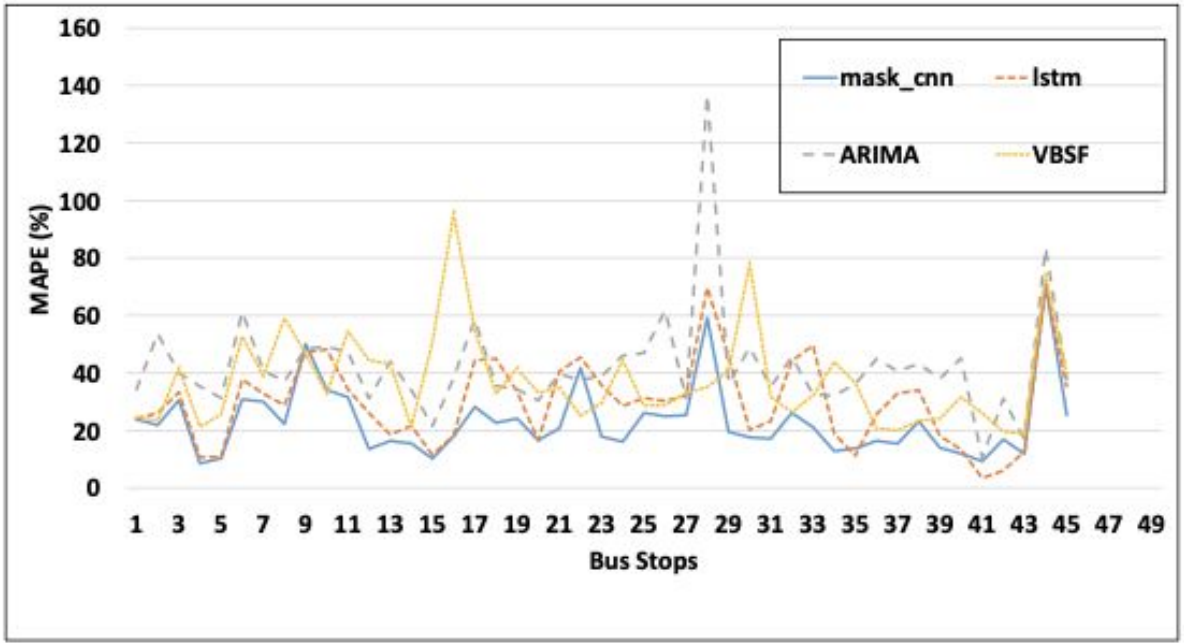
Solution: Mask-CNN

- Propose a system that generates ETA information for a trip and updates it as the trip progresses based on the real-time information
- Use a generative autoregressive model that learns the distribution of the ETA data and conditional on the current trip information updates the ETA information on the go.
- Utilises only the historic GPS data from the same route.
- Parameters can be tuned automatically thereby increasing the ease of implementation in the real-world scenarios
- "Regular" convolution filter may imply that we end up using points for the convolution operation that may not have been generated yet .
- Solution:- Using Masking

Infer the future ETA using Mask-CNN



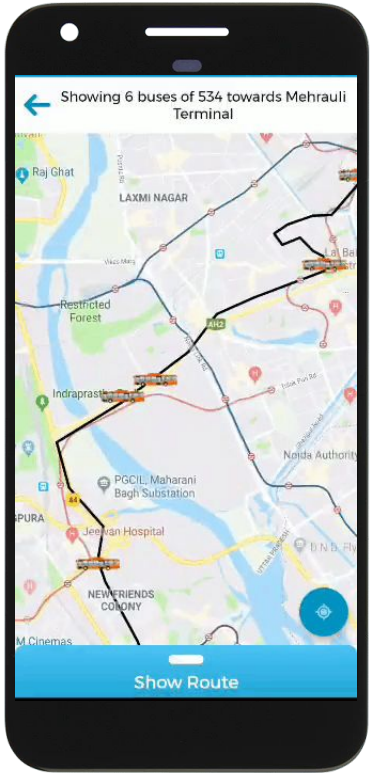
Results


160
140
120
100
80
60
40
20
0
MAPE (%)
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
Bus Stops
mask_cnn
lstm
ARIMA
VBSF

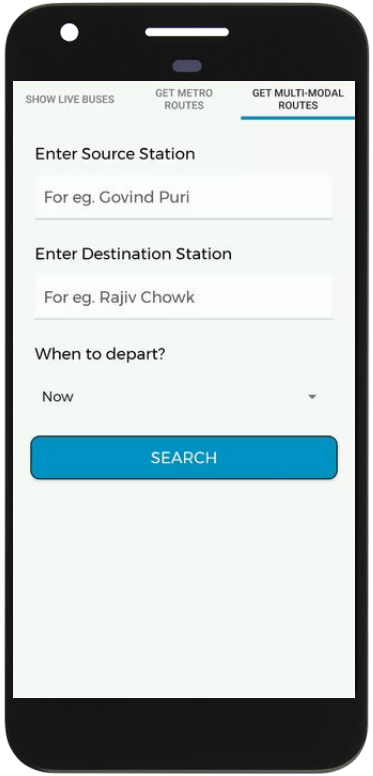
chartr

Android App

- Built on **Open Transit Data for cluster buses** in Delhi
- Shows real-time location and distance of bus from the stop.
- Trip planning with live status of buses
- Multi-modal trip planning. Metro + Bus + Walk using real-time data.



show real time bus location



multi-modal trip planner

Public Information System (PIS)

A Public Information System (PIS) to be installed at different locations for commuters to know about the *incoming buses*.

- Built on **Open Transit Data for cluster buses** in Delhi.
- Easy to install.
- Can be installed at :
  - Bus stops
  - Common shops
  - Educational institutions.

Problem 3: Noisy Data

Non Linear Models

High Dimensional Matrix Completion

If the data defining the matrix belongs to a structure having fewer degree of freedom than the entire dataset, that structure provides redundancy that can be leveraged to complete the matrix. This typical subspace assumption is not always satisfied.

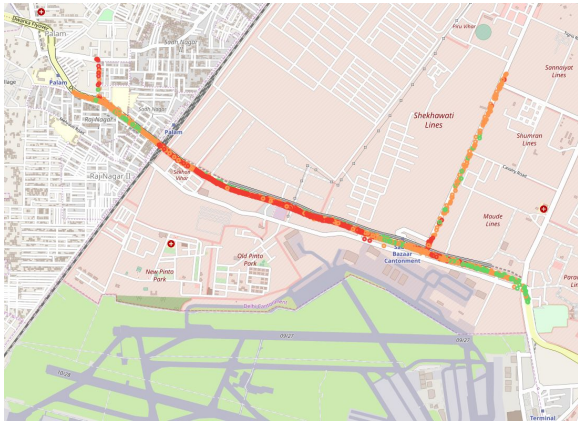
The original matrix is possibly high rank, but it becomes low rank after mapping each column to a higher dimensional space of monomial features. [1]

[1] Ongie, Greg, et al. "Algebraic variety models for high-rank matrix completion." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

[2] Jicong Fan, Madeleine Udell “Online High Rank Matrix Completion.”*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Impact of breakdown on congestion

Breakdown for this vehicle (DL1PD3210) was observed near Dwarka Flyover on 17th October at around 5pm.



Traffic speed (4-5pm)	Traffic speed (after 5pm)

Backup Slides

Current Status of Transit Data in most Indian cities

- Almost no collection of transit data in Indian mass transit
- Most processes (like scheduling, time table-ing) are manual

- If some data is collected it is hardly used for:

    - Increasing passenger information

    - Informed decisions

- No technical expertise either

Proposed Project: Open Transit Data for Indian Cities

- **Create an India specific open transit data platform that efficiently and automatically collects all the relevant transit data**
  - Used by map apps like Google, MakeMyTrip, MapmyIndia, Chalo
  - Transit providers -- DTC, BMTC, Bhopal City...
  - Transportation researchers
  - Government agencies

Passenger facing apps → information

can be used to perform several tasks:

- route design,
- bus bunching,
- passenger information systems

Impact

Expect a **minimum 10% rise [2]** in ridership in buses → **more than 4 lakh passengers per day.**

4 L passenger trips per day is the ridership of Mumbai metro today

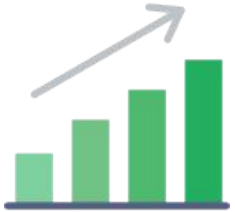
An increase of 10 Cr per month for DTC/BCLL

Less congestion and pollution

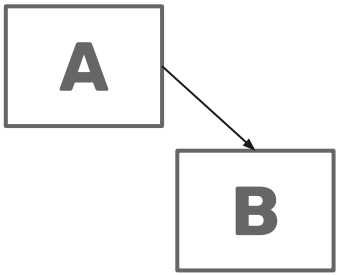
Use Cases for Public

With such live data being publicly available, anyone can build really helpful applications for daily commuters.

Estimated Time of Arrival.
Less Waiting Time!



Increased Ridership
More Revenue for Transit Agencies

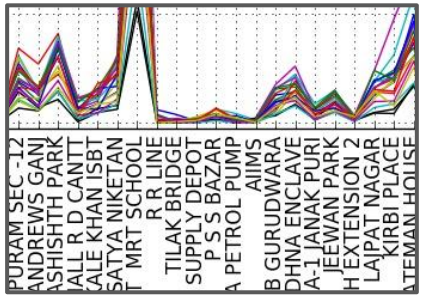


Multimodal Planning.
Better trip planning with real-time data

Use Cases for Transit Agencies

Current traffic system can be made efficient with so much information openly available.

Improve traffic prediction from live bus locations.
Maps for buses!



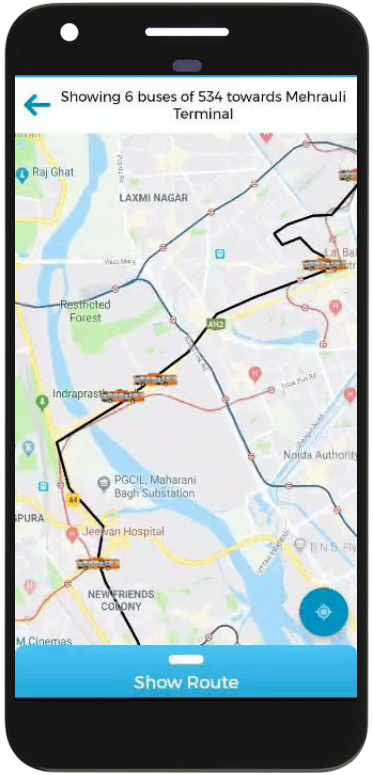
Route analysis, rationalisation.
Understand travel patterns.

Faster bus-breakdown information.
Both for passengers & agencies.

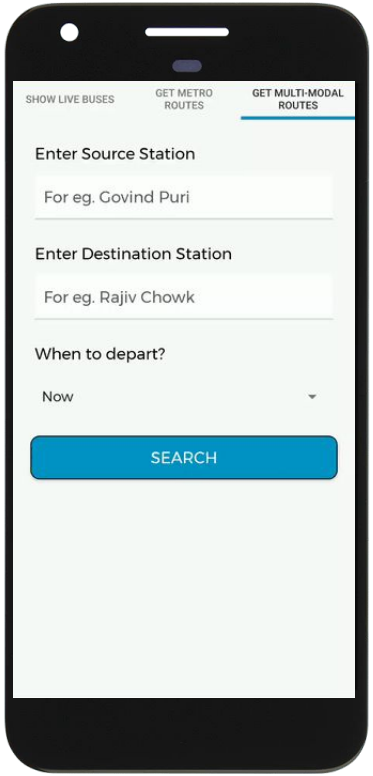
chartr

Android App

- Built on **Open Transit Data for cluster buses** in Delhi
- Shows real-time location and distance of bus from the stop.
- Trip planning with live status of buses
- Multi-modal trip planning. Metro + Bus + Walk using real-time data.



show real time bus location



multi-modal trip planner

Public Information System (PIS)

A Public Information System (PIS) to be installed at different locations for commuters to know about the *incoming buses*.

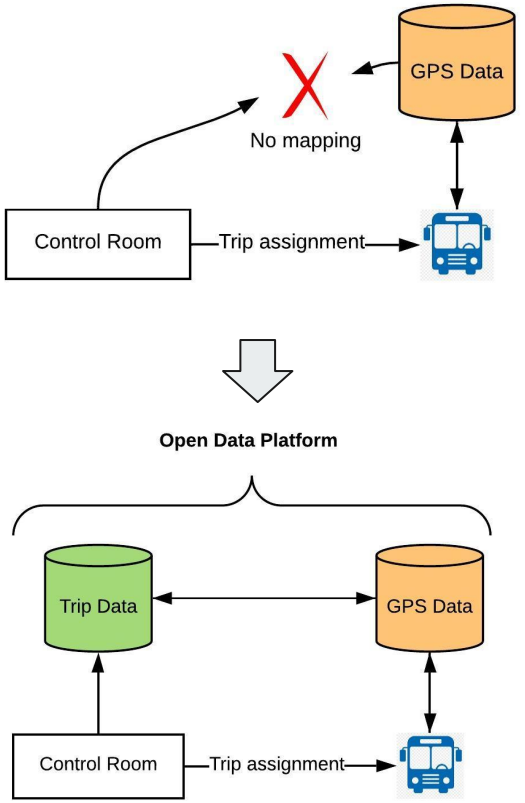
- Built on **Open Transit Data for cluster buses** in Delhi.
- Easy to install.
- Can be installed at :
  - Bus stops
  - Common shops
  - Educational institutions.

Depot Trip Management System

crucial

Since a bus can go on different routes everyday, there is a need to build a tool to at least update bus out-shedding information from all the 40 depots in real-time.

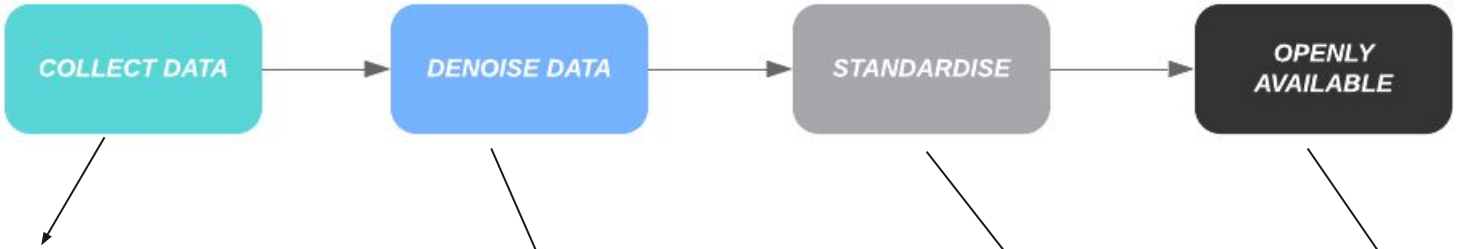
Currently, outshedding information is written in notebooks.



PROPOSED SOLUTION

Since there are more than **5000** buses in Delhi, originating from **40+** bus depot from **5AM** in the morning, sharing **20M** data points every hour, it will be impossible to collect, denoise & share the data with users in real-time using manual techniques.

Automated Open Transit Data Process Flow



Automated platform for each depot. (Across 40 depots in Delhi)

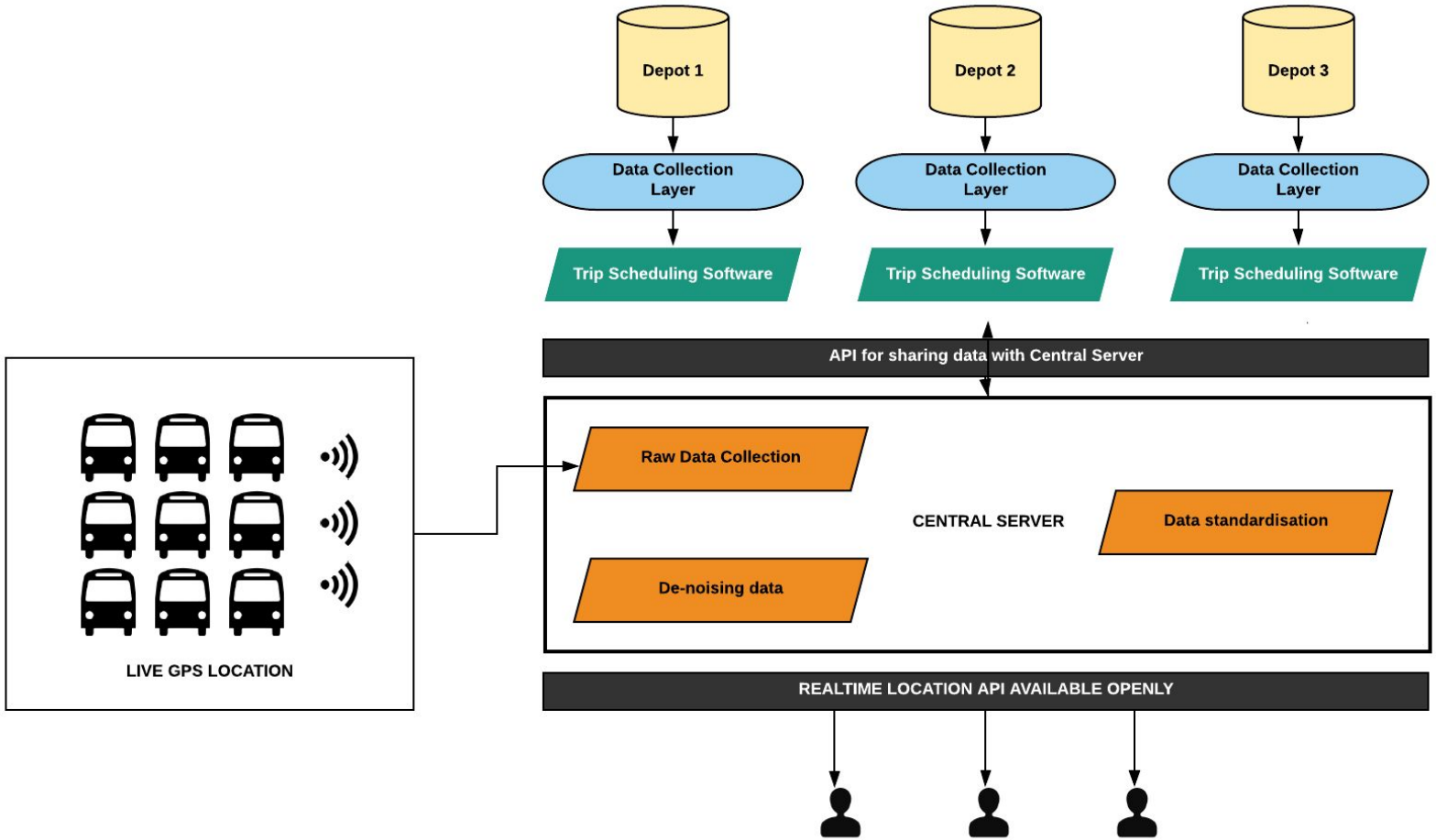
Several research challenges:

1. Incomplete data imputation
2. Denoising
3. Algorithms for Real-time processing

Adopting the globally recognised GTFS standard

APIs → Thousands of users including Google!

Architecture of Proposed Solution



Technical diagram for our end to end infrastructure

System Complexity

6000 buses a day for Delhi itself (10 million?)

20M data-points every hour

85k+ API requests per day for every user, 15+ active users

1M+ requests per day

10GB/day/user served

150GB data served everyday


Thank you.