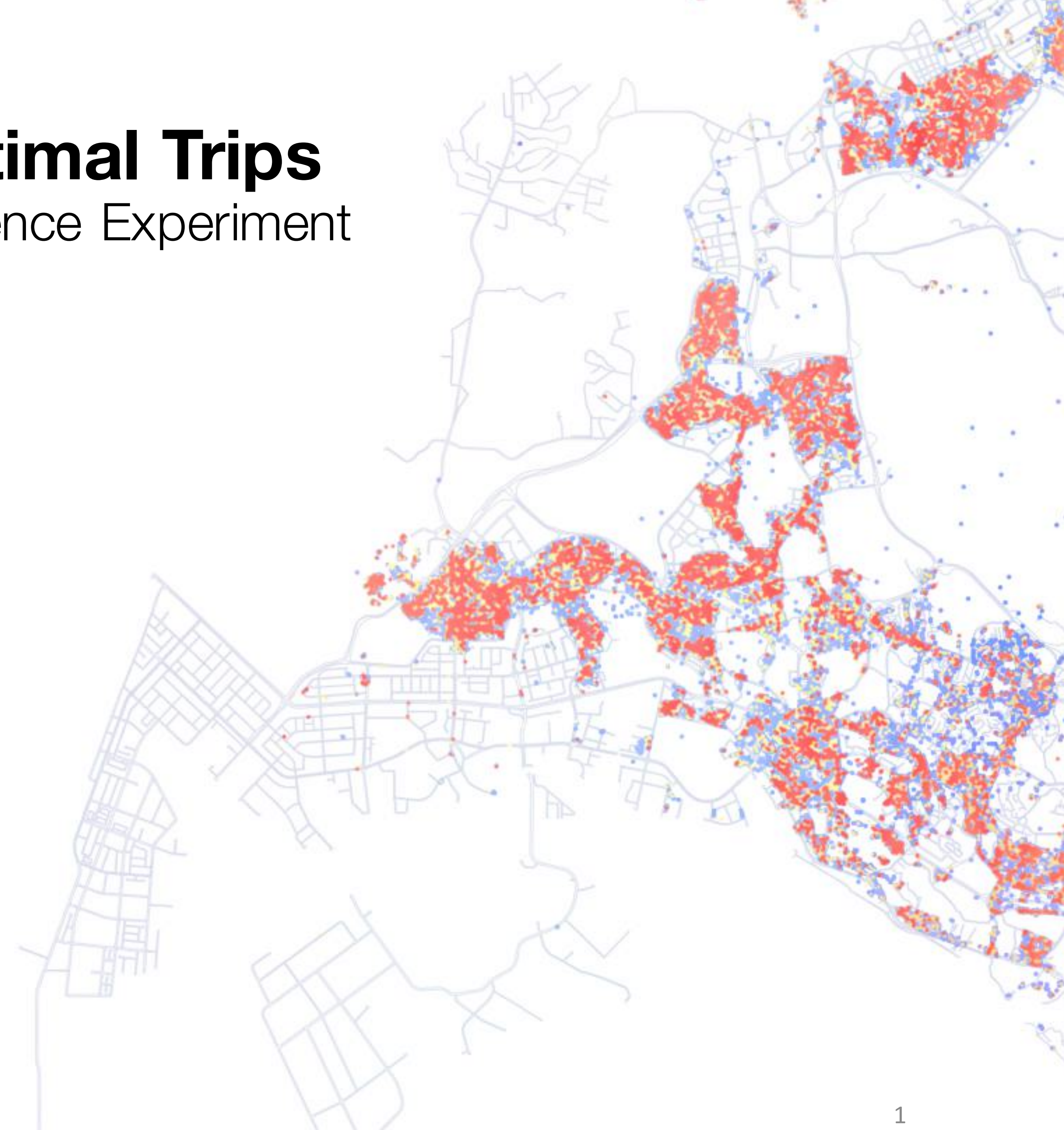
Inferring Activities and Optimal Trips

Lessons from Singapore’s National Science Experiment



Presenter Barnabé Monnot

Authors

B. Monnot, E. Wilhelm, G. Piliouras, Y. Zhou

Singapore University of Technology and Design

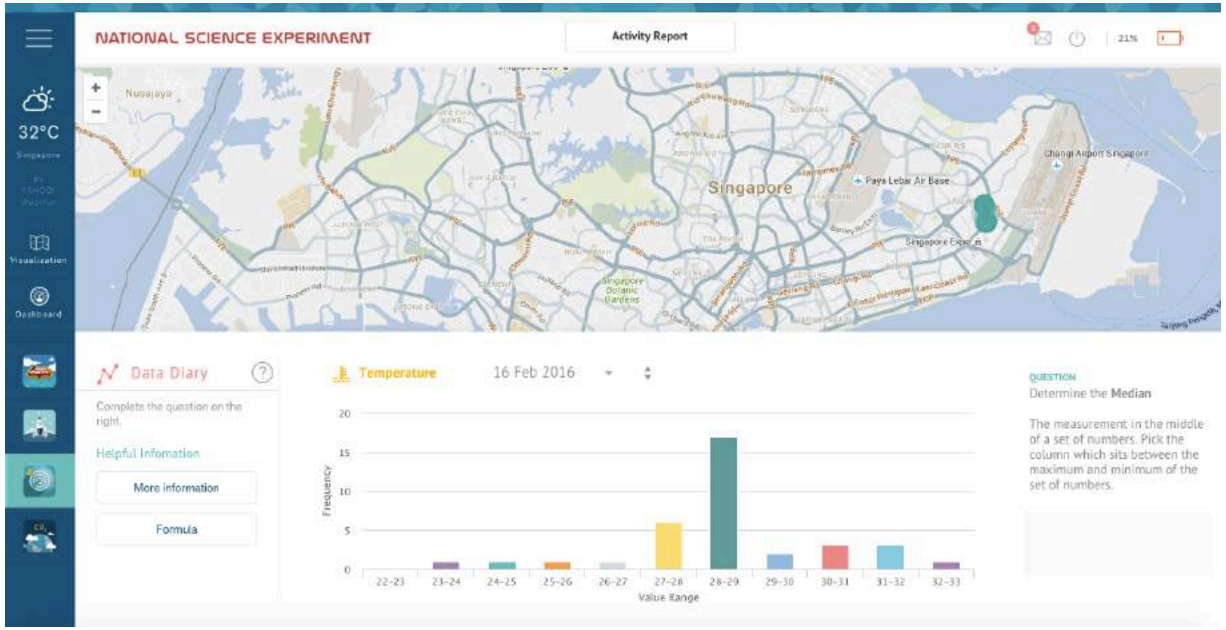
D. Dahlmeier, H.Y. Lu, J. Wang

SAP Singapore

Overview of the experiment

Students were asked to carry the sensor for a week

- Stated goal: introduce students to Internet of Things technology and big data analysis, and allow them to explore their data on an SUTD designed web portal



Overview of the experiment

The sensor is equipped to collect and send multi-dimensional data

- Sensors are custom-designed by SUTD and Delta Electronics Industrial Automation Business Group
- Backed by server infrastructure designed by SUTD with A*Star IHPC group
- Capable of recording temperature, relative humidity, atmospheric pressure, sound pressure, light level and 9-degrees of freedom motion data

Overview of the experiment

Location is determined by looking at Wi-Fi addresses collected by the sensor

- The sensors are equipped with Wi-Fi radio
- Collect MAC addresses from surrounding Wi-Fi points, used to determine location
- Send data points to the server remotely

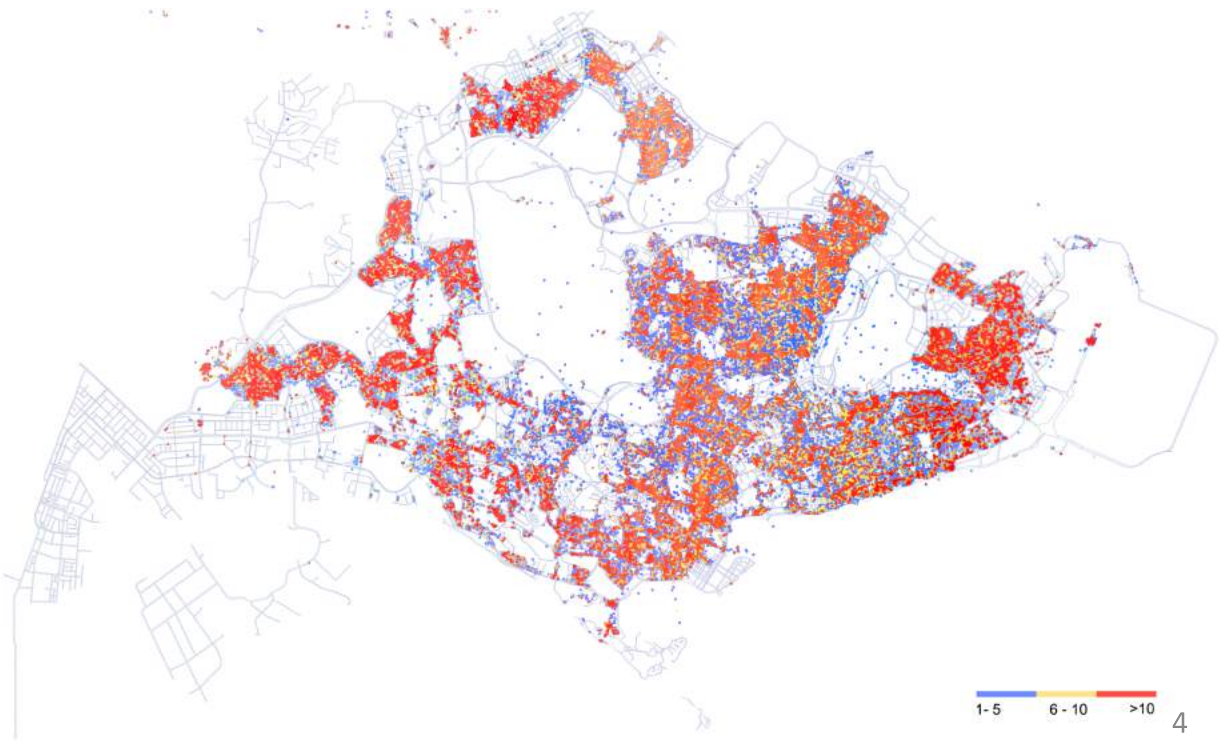
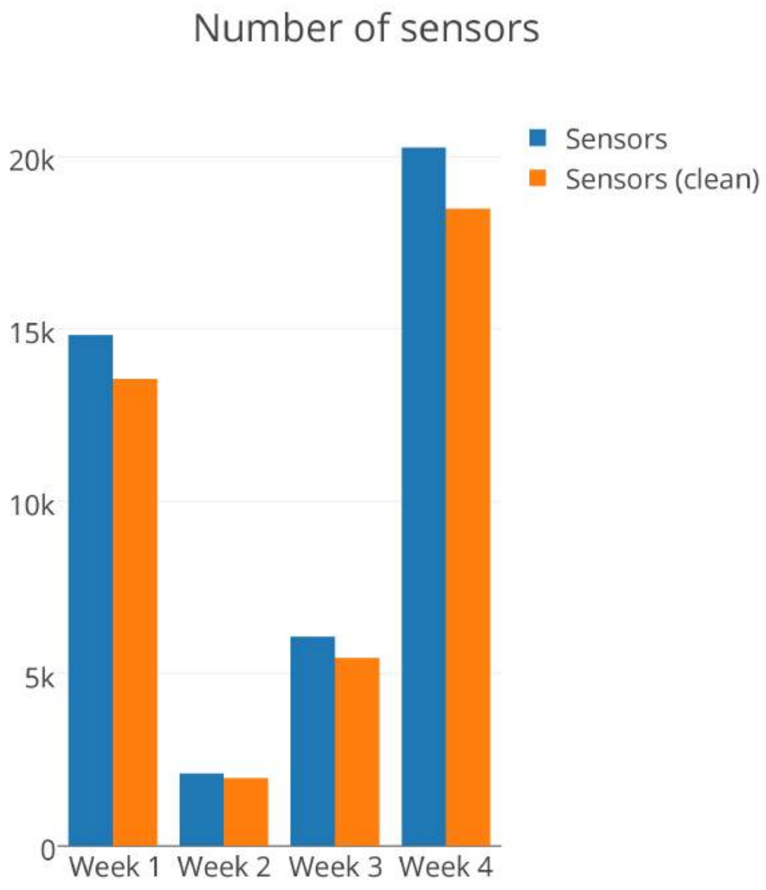
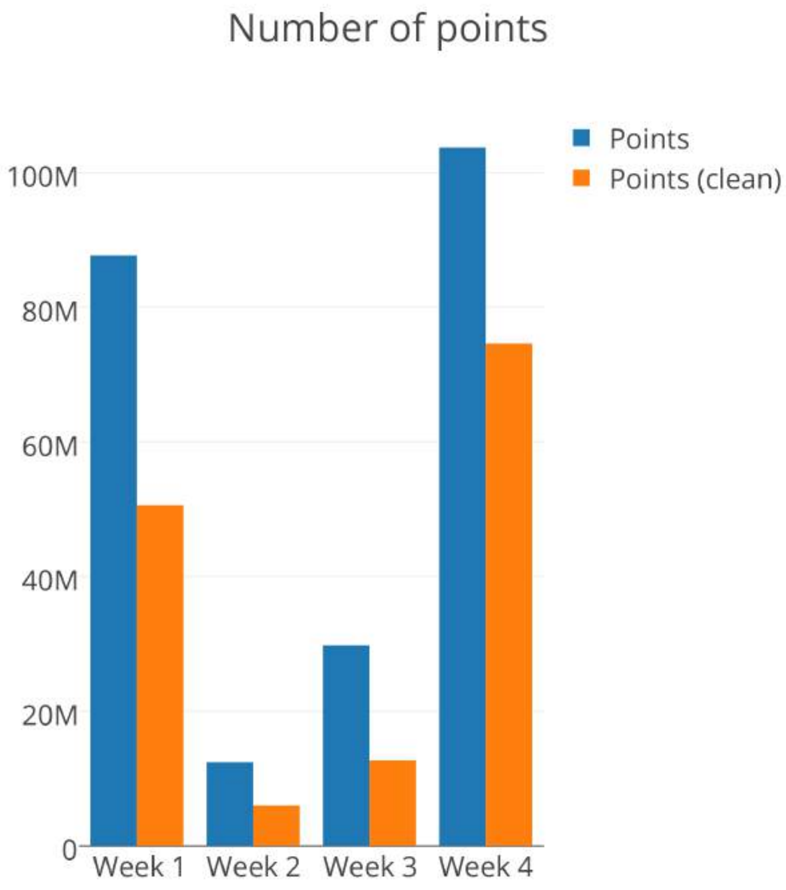


Figure Wi-Fi density in Singapore

Overview of the experiment

Over 40.000 students carried a sensor, totaling over 130 million clean data points





Overview of the experiment

The experiment took shape with the work of many different entities

Sensor design

Server

Analytics

Stories of the smart city

Understanding movement in cities is the first step in assessing human motivations

- Smart cities 'think ahead': they ease interactions with the urban environment by predicting what is next
- To make this prediction, we need to make large data collections available to city planners and researchers
- Individual trips can be aggregated into models of social behavior: who is around me? How much do I interact with them? How? In which conditions?

State of the art

Our experiment is part of a vibrant research community

18.01.2011 - UNESCOPRESS

Global experiment aims to break record for International Year of Chemistry (2011)


© Joe Sullivan/Wikimedia Commons

Primary and secondary school pupils are invited to take part in an attempt to break the record for the biggest scientific experiment ever conducted, as part of the celebration of the International Year of Chemistry (IYC) 2011. The Year will be officially launched by UNESCO and its partner, the International Union for Pure and Applied Chemistry (IUPAC), at a conference on 27 and 28 January at UNESCO's Paris headquarters.

SENSORS FOR THE PEOPLE

How polluted is your home or neighbourhood? Until recently, it was difficult to answer that question because data were available only from networks of expensive sensors in relatively limited locations. The do-it-yourself movement has led to the emergence of low-cost sensors that can be purchased or built from online instructions.

DUSTDUINO
Measures the concentration of pollution particles equal to or smaller than 10 micrometres (PM_{10}) and 2.5 micrometres ($PM_{2.5}$).

AIR QUALITY EGG
Measures nitrogen dioxide and carbon monoxide, humidity and temperature. Streams data online through a separate base station.

SMART CITIZEN KIT
Measures nitrogen dioxide and carbon monoxide, as well as light, humidity, temperature and noise pollution.

- Large-scale experiments involving many different actors are conducted to introduce citizens to scientific possibilities
- Sensors are becoming ubiquitous instruments for personal data collection, but still have low penetration rates

Contents of the talk

- Data analytics
  - Trip recognition
  - Activity matching
  - Efficient routing
- Trip analysis: a visual app
- Future work

Research questions

Making sense of large-scale multi-dimensional data

- Where are people starting and ending their trips?
- Which activities are people traveling to/from?
- Are they taking the most efficient route?

Contents of the talk

- Data analytics
  - Trip recognition
  - Activity matching
  - Efficient routing
- Trip analysis: a visual app
- Future work

From raw data to trips

Automated method for trip recognition to overcome active reporting issues

- Number of studies use active trip reporting: diary kept by participants, retrospective surveys
- Under-reporting issues: noted by Du and Aultman-Hill (2007), up to 50% of loss
- Implement passive method: recognize trips from the location data

Raw data	
Latitude	Longitude
1.3495	103.9305
1.3496	103.9307
1.3492	103.9303

→ **Recognition algorithm**

Trips	
Start point	End point
(1.3495, 103.9305)	(1.3492, 103.9303)

Recognition algorithm

Based on computed sensor velocity

- `Distance(Point A, Point B)` is the distance on Earth between point A and B
- Compare timestamps to get velocity

Raw data		
Latitude	Longitude	Timestamp
1.3495	103.9305	1455270655
1.3496	103.9307	1455270667
1.3492	103.9303	1455270683

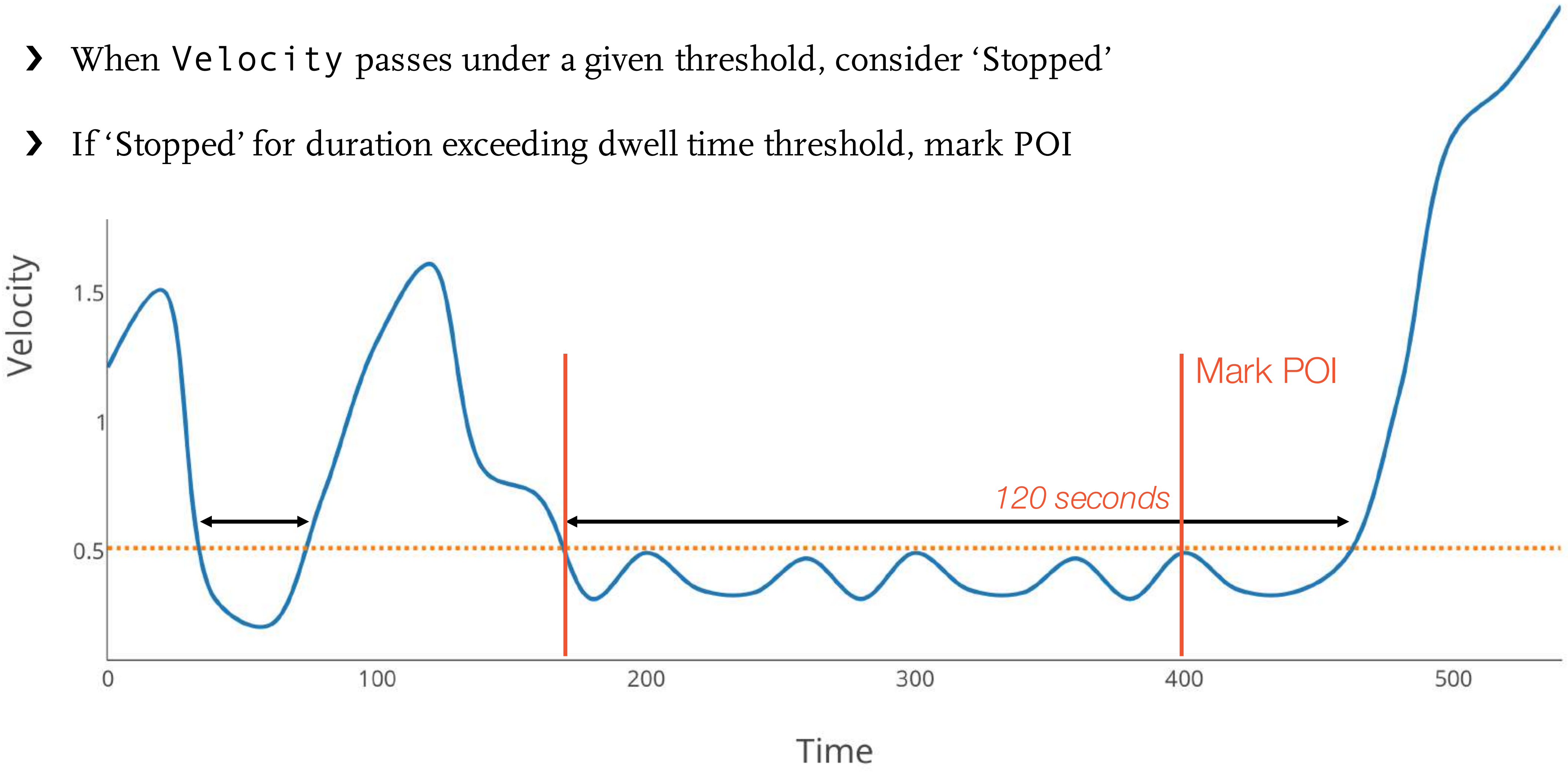
→

$$\text{Velocity} = \frac{\text{Distance(Point A, Point B)}}{\text{Timestamp B} - \text{Timestamp A}}$$

Recognition algorithm

Velocity and time stopped determine Points Of Interest (POI)

- When `Velocity` passes under a given threshold, consider ‘Stopped’
- If ‘Stopped’ for duration exceeding dwell time threshold, mark POI



Recognition algorithm

Our algorithm has detected a large number of trips

6587	Students data analysed
36983	Trips recognized
21465	Points of Interest recognized
~5-6	Trips per student recognized
~3	Points of Interest per student recognized

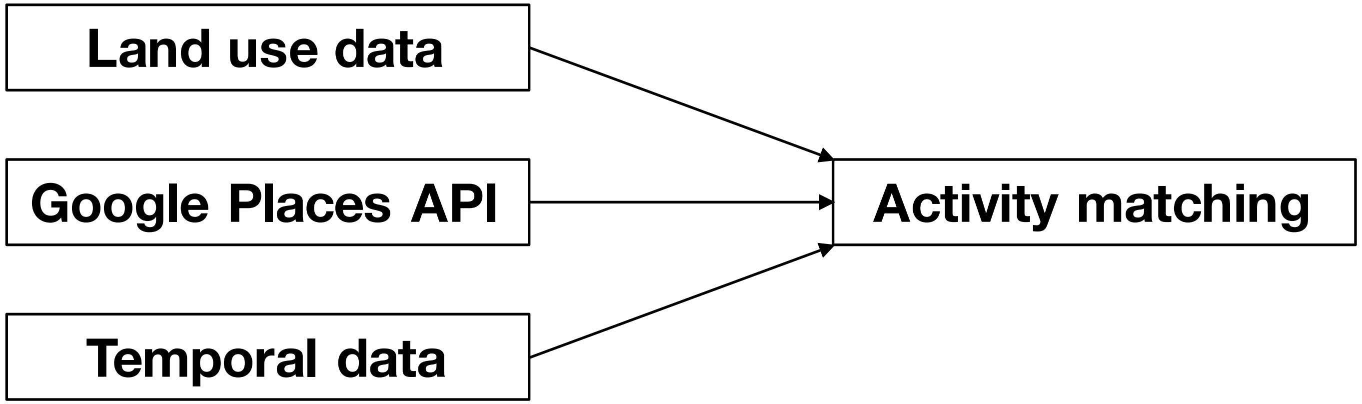
Contents of the talk

- Data analytics
  - Trip recognition
  - Activity matching
  - Efficient routing
- Trip analysis: a visual app
- Future work

Activity matching

Find an automated process to determine the purpose of the trips

- We use the previous algorithm to recognize trips from the raw data
- The next step is to understand the purpose of the trip: which activity is the agent performing?
- We use aggregate data collected from various sources to determine the activity



Home and school recognition

Home and school are recognized using temporal data

- Most of the time is spent by the student either at school or at home
- We use temporal data to assign the location of both: no active reporting

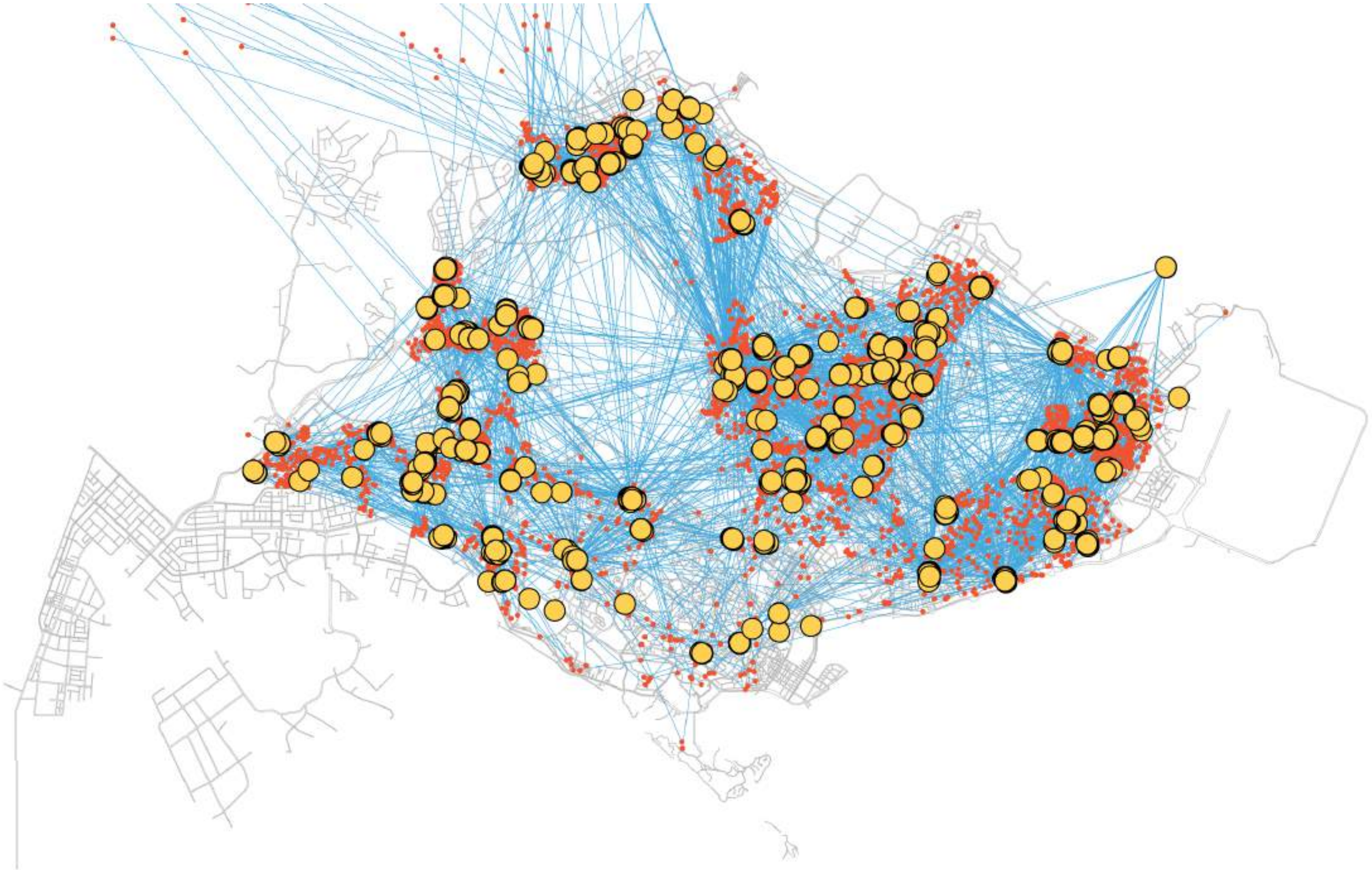


Figure Homes and schools

Weighted average assignment

To decide which activity is performed, synthesize the information collected

- If the student is neither at home or school, decide between ‘food’, ‘outdoor’ or ‘commercial’ for the activity
- Land use and Google Places API data are compiled to give a score to each option
- We then select the activity with the highest score

Score		
Food	Outdoor	Commercial
4.3	2.1	3.7

Land use data

We use the geography of parks and outdoor recreational space

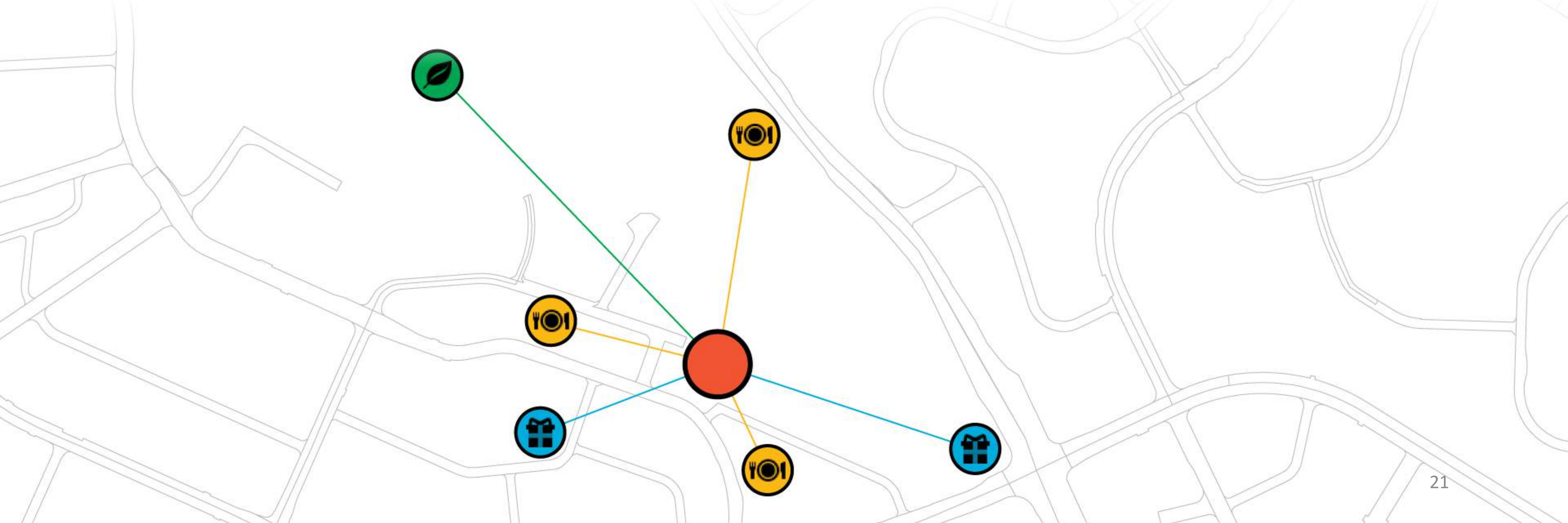
- Land use data is recorded by the URA (Urban Redevelopment Authority)
- We select the outdoor installations of Singapore to classify points inside these bounds as participating in an outdoor activity

Figure Outdoor areas in Singapore

Google Places API

Collect information on the surrounding places

- Google Places has a large database of locations with their functions (restaurants, gyms, parks…)
- We scan the surroundings of a POI to detect these locations and help us decide on the activity



Weighted average assignment

The formula gives points to closest clues

Land use data
Point is in outdoor area
Outdoor $+b_O$

Google Places data	
Type	Food
Distance to POI	100m
Food $+b_F*(1-100/r_{max})$	

Score		
Food	Outdoor	Commercial
1.2	10	0

Assign activities to endpoints

Activity matching helps us understand the purpose of the trip

- After assigning activities to all endpoints, we can build the *origin-destination matrix*
- Entry (i, j) is the number of trips from activity i to activity j

Origin-destination matrix

	Home	School	Food	Commercial	Outdoor	Others
Home	0	3	1	2	0	1
School	3	0	1	0	3	1
Food	1	4	0	2	0	2
Commercial	0	0	3	1	2	1
Outdoor	2	1	0	0	1	1
Others	1	0	3	2	1	2

Visualizing the trips

Interactive chord diagrams give a clear presentation of the origin-destination matrix

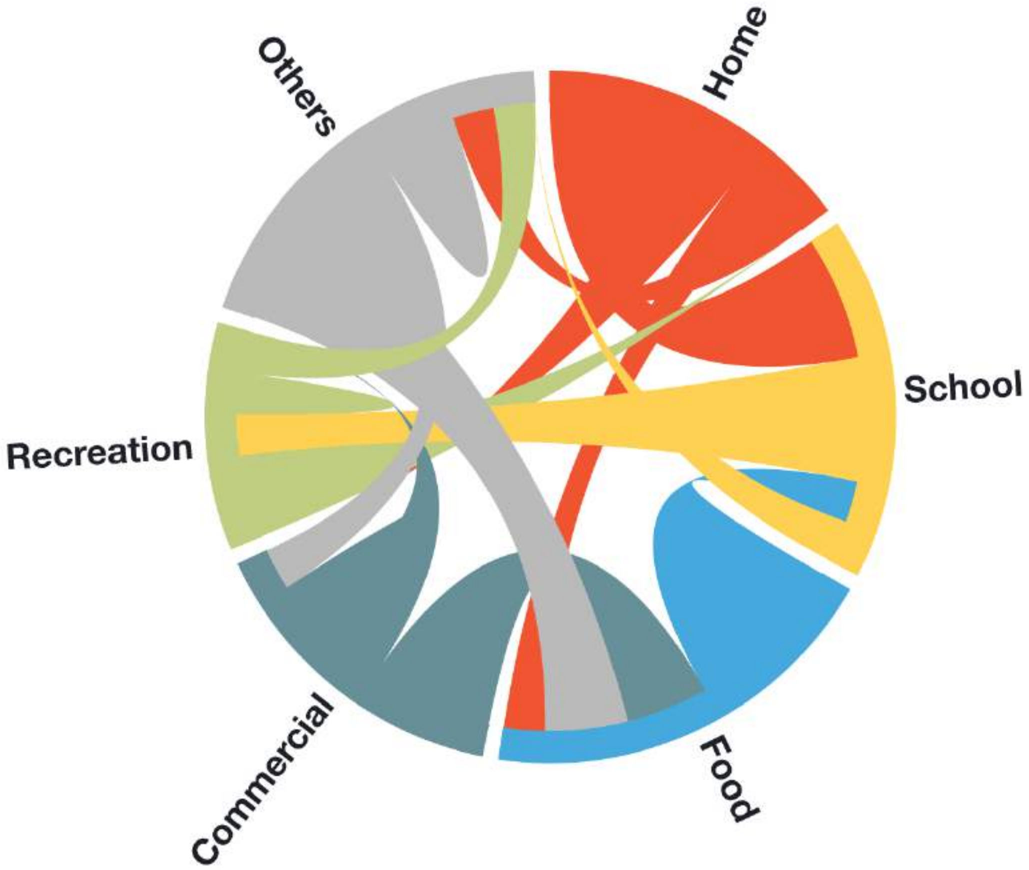
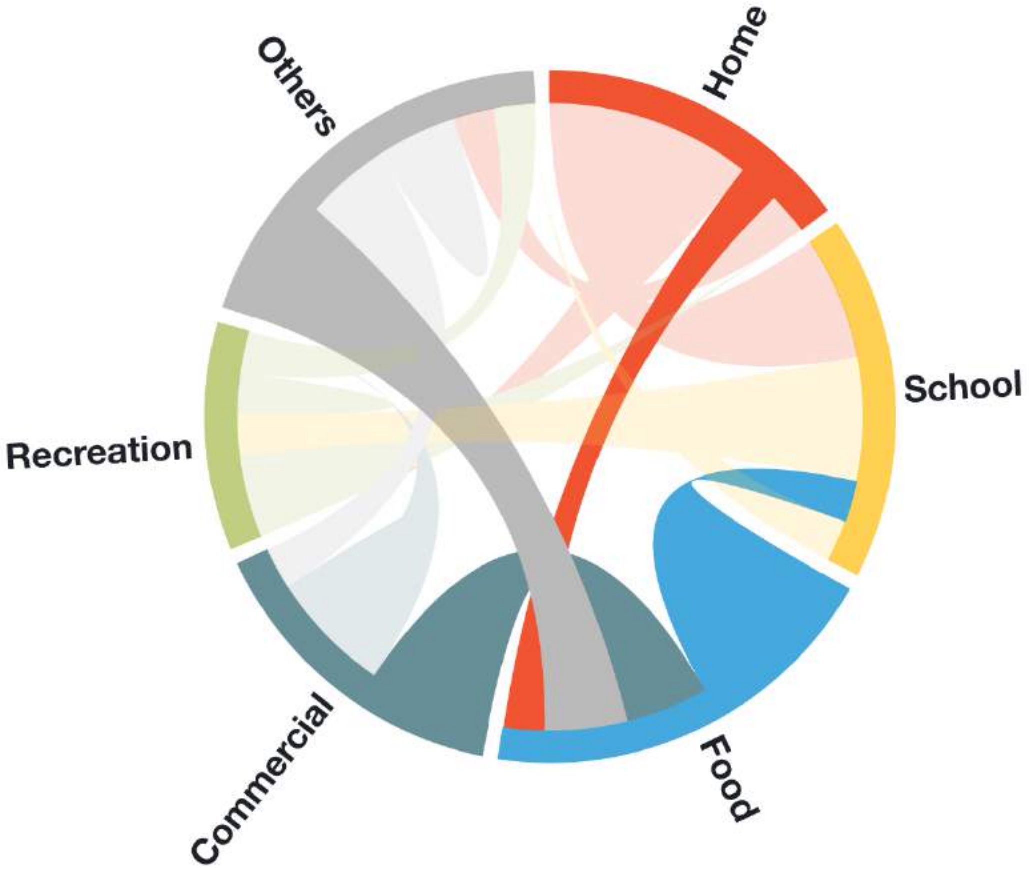


Figure Chord diagram, built with D3

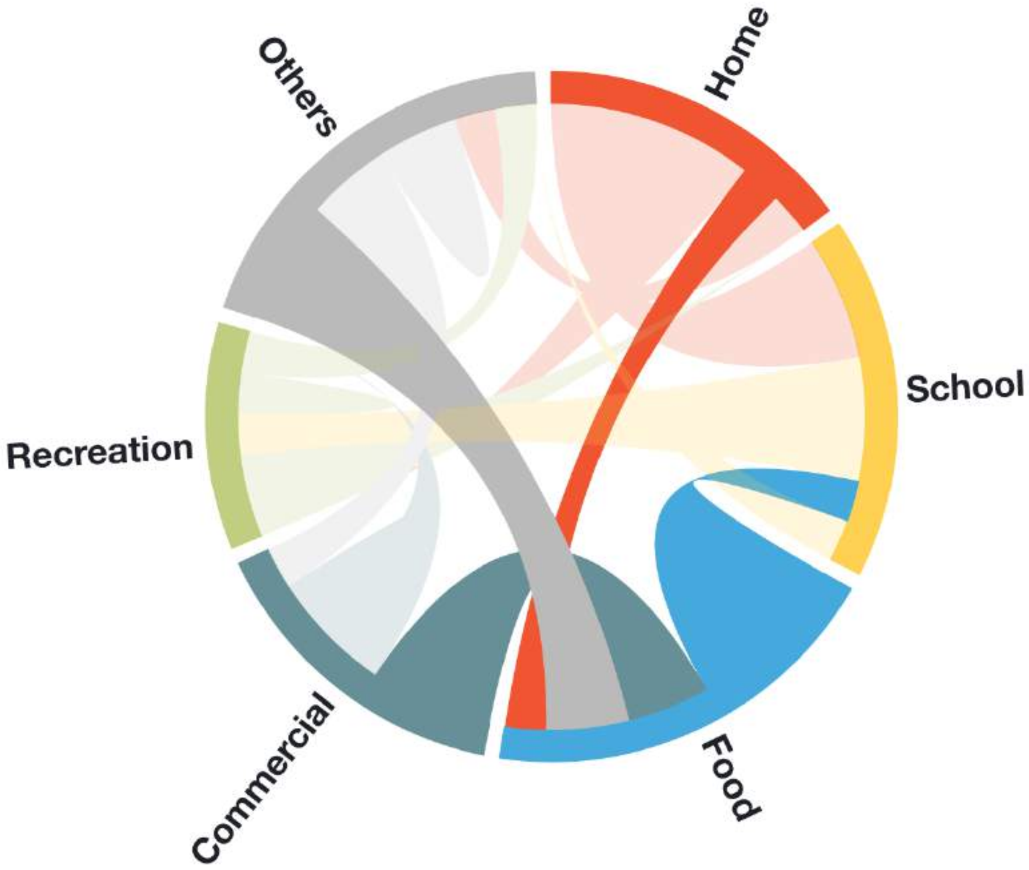
Visualizing the trips

Fading the diagram gives more legibility



Visualizing the trips

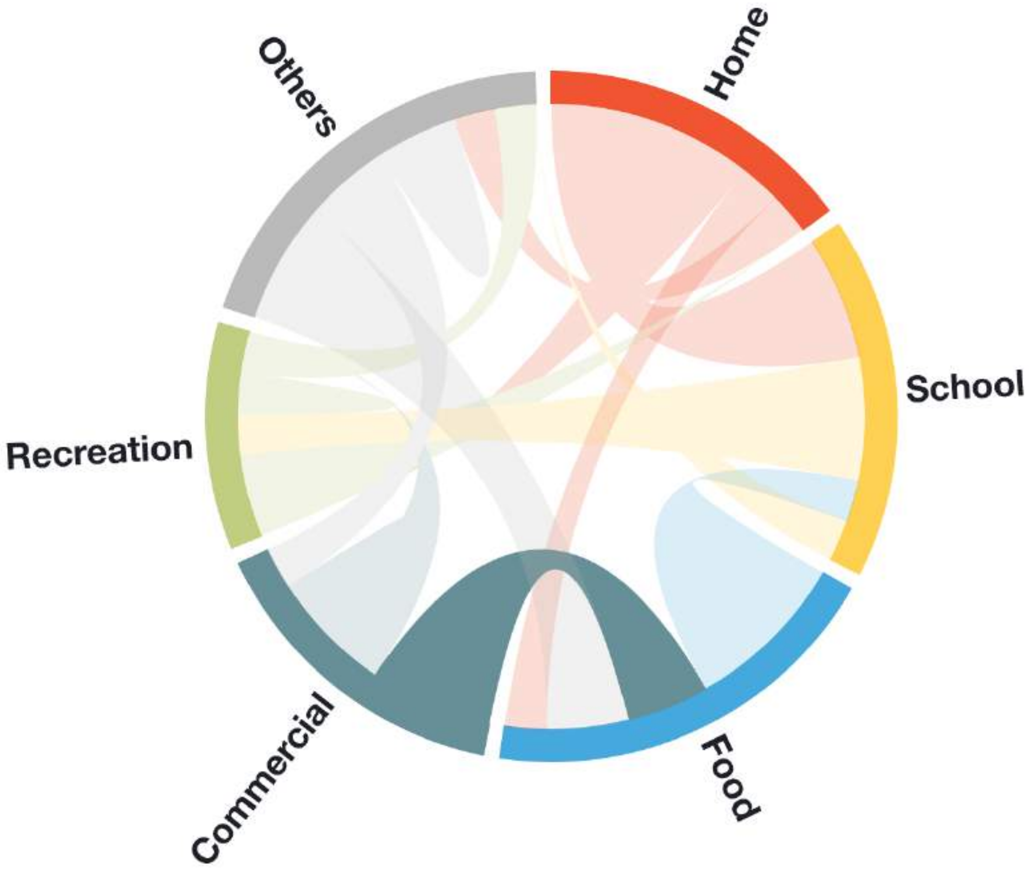
The origin-destination matrix is represented by a chord diagram



	Home	Sch.	Food	Com.	Out	Oth.
Home	0	3	1	2	0	1
Sch.	3	0	1	0	3	1
Food	1	4	0	2	0	2
Com.	0	0	3	1	2	1
Out	2	1	0	0	1	1
Oth.	1	0	3	2	1	2

Visualizing the trips

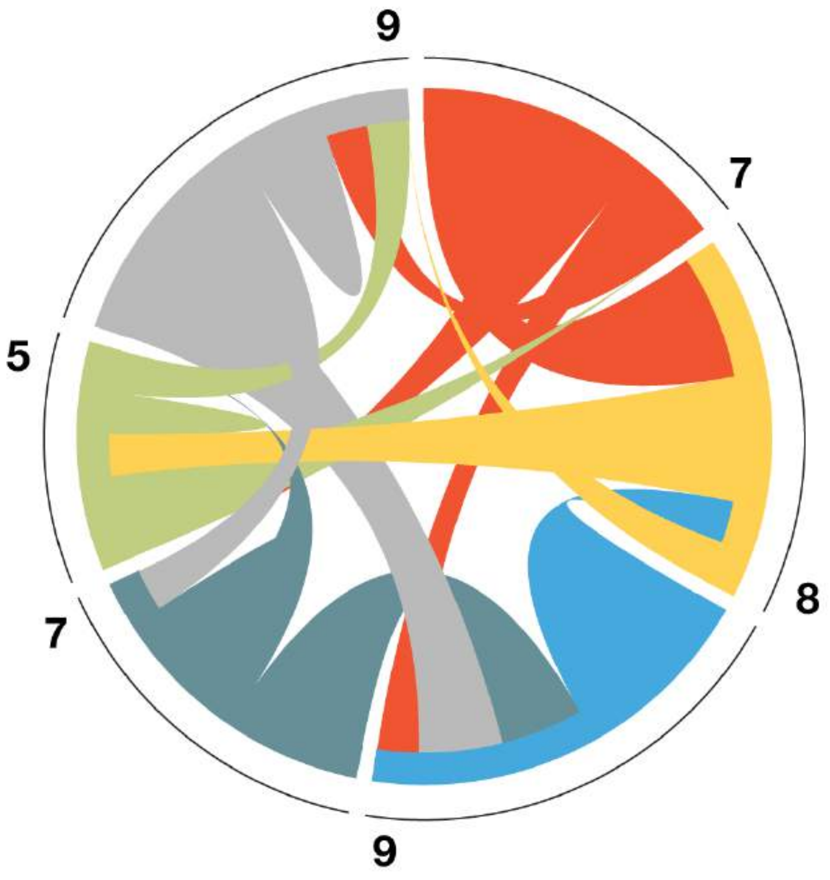
The size of an arc's endpoints are proportional to the number of trips starting from each endpoint



	Home	Sch.	Food	Com.	Out	Oth.
Home	0	3	1	2	0	1
Sch.	3	0	1	0	3	1
Food	1	4	0	2	0	2
Com.	0	0	3	1	2	1
Out	2	1	0	0	1	1
Oth.	1	0	3	2	1	2

Visualizing the trips

The size of the outer circle gives the number of trips starting from the corresponding activity



	Home	Sch.	Food	Com.	Out	Oth.	Total
Home	0	3	1	2	0	1	7
Sch.	3	0	1	0	3	1	8
Food	1	4	0	2	0	2	9
Com.	0	0	3	1	2	1	7
Out	2	1	0	0	1	1	5
Oth.	1	0	3	2	1	2	9

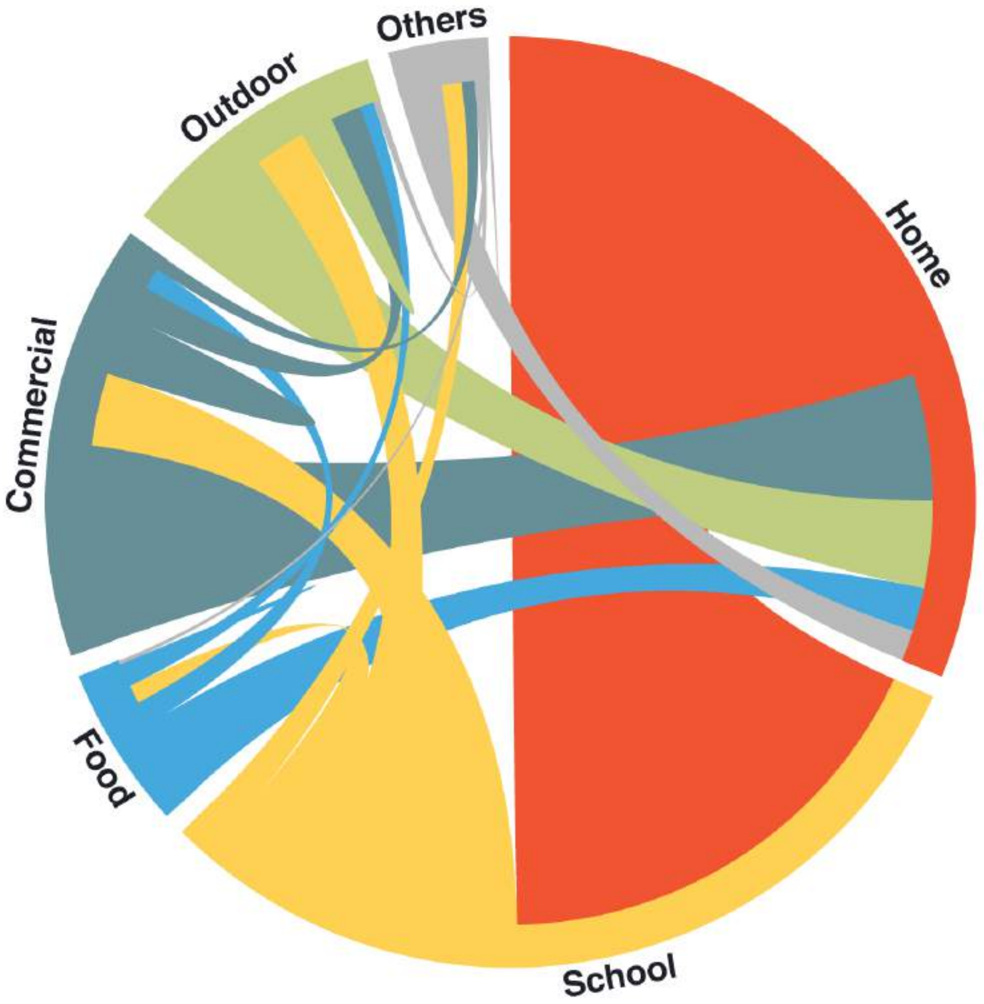
Acitivity matching results

We gain knowledge on the student's environment by looking at its surroundings

6587	Students data analysed
4520	Home/School pairs recognized
1879	Food activities
4642	Commercial activities
2898	Outdoor activities

Total origin-destination matrix

We can build a complete origin-destination matrix based on all students' trips



Contents of the talk

- Data analytics
  - Trip recognition
  - Activity matching
  - Efficient routing
- Trip analysis: a visual app
- Future work

Optimal route selection

How do agents decide on which route to choose?

› To understand the decision-making process of the students, we can check whether they picked the fastest or more direct way to their destination

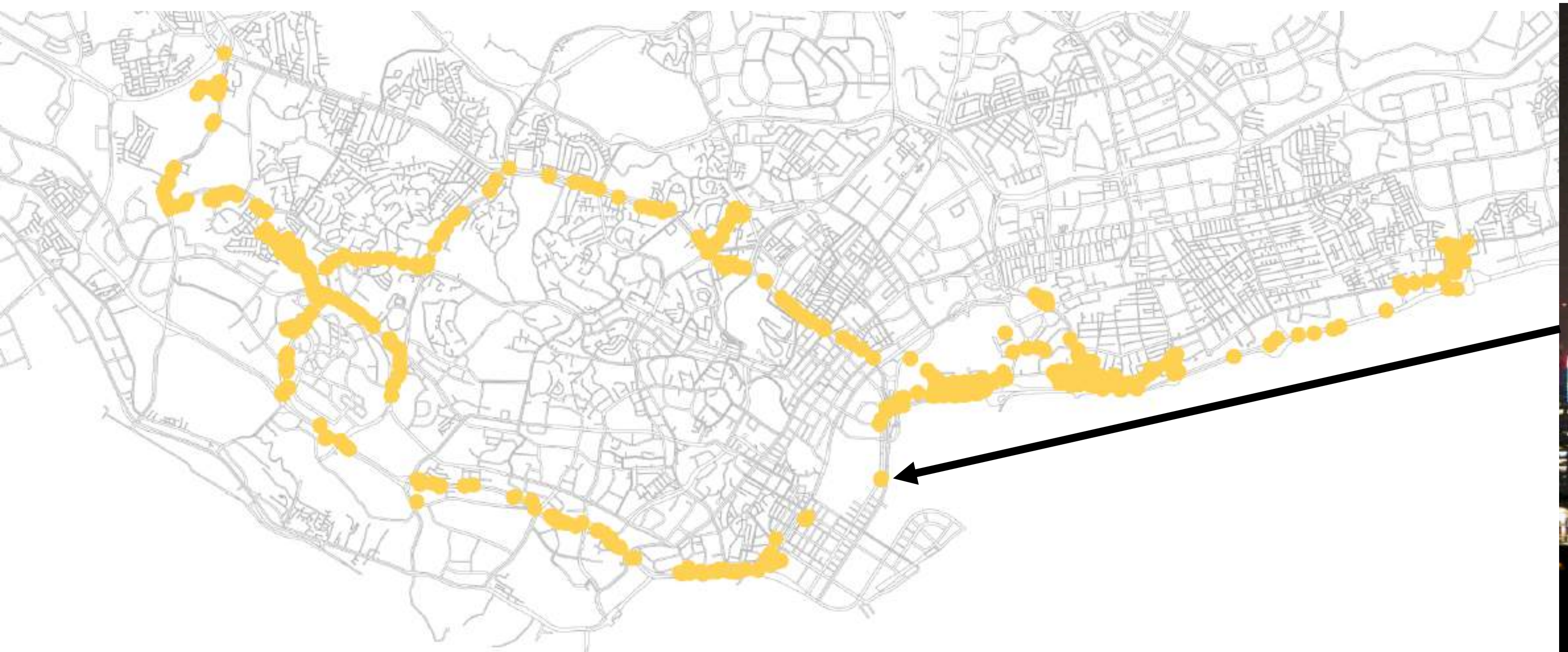
› We use Google Directions API to compare the trip undertaken by the student to the suggested itinerary

Optimal route selection

How do agents decide on which route to choose?

- To understand the decision-making process of the students, we can check whether they picked the fastest or more direct way to their destination

- We use Google Directions API to compare the trip undertaken by the student to the suggested itinerary



Optimal route selection

Comparing the trip with the suggested itinerary helps determine the mode of transportation

- The Directions API lets us find the fastest itinerary by car, by public transportation or by walk
- We can supplement the mode identification algorithm with clues from the Directions API to precisely determine the mode of transportation

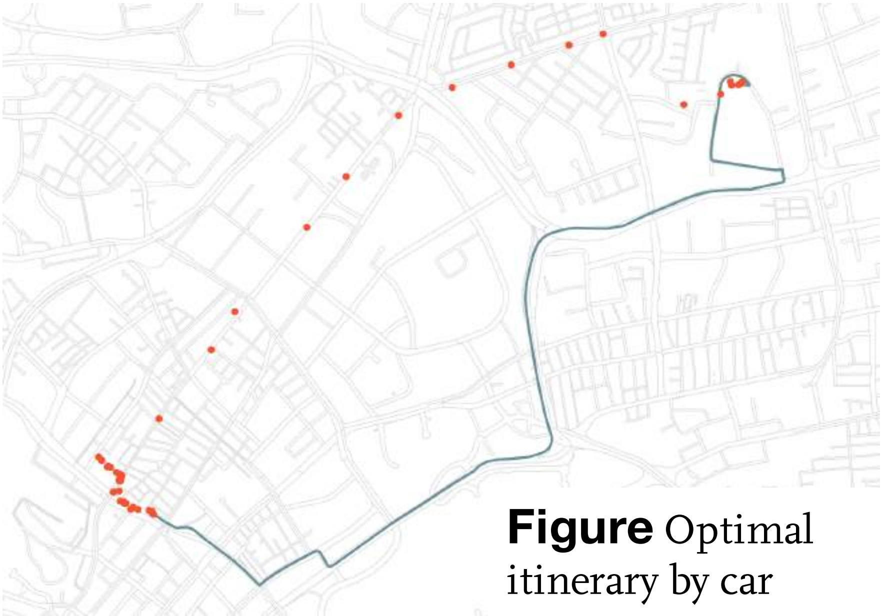


Figure Optimal itinerary by car

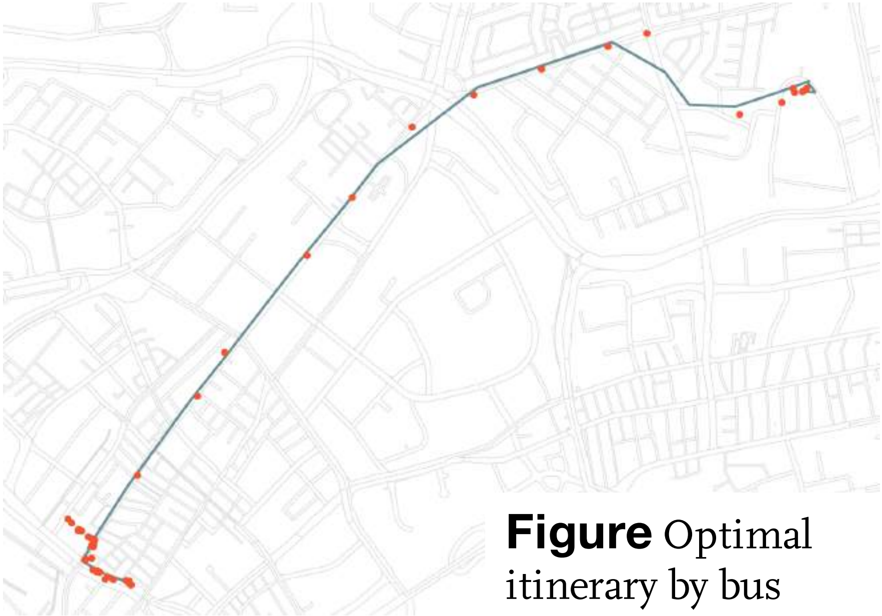


Figure Optimal itinerary by bus

Contents of the talk

- Data analytics
  - Trip recognition
  - Activity matching
  - Efficient routing
- Trip analysis: a visual app
- Future work

Trip analysis app

Building an app to explore trip analysis data

- Visual information makes digging deeper into the data an easier task
- We built an app on Node.js / PostgreSQL to read the collected data and present it visually

localhost

Sensg / Trip Analysis | Trips | Node | Inspect | Random | Others

The Singapore National Science Experiment

Contents of the talk

- Data analytics
    - Trip recognition
    - Activity matching
    - Efficient routing
- Trip analysis: a visual app
- Future work

Algorithmic challenges

Large volumes of data demand different approaches

- Streaming algorithms provide a good alternative to linear-time ones running on all the data
- If we are looking for information on global structures in our data, we want sublinear-time algorithms
- *Property testing* is a powerful tool: testing a smaller input and finding the probability that the data has a particular property

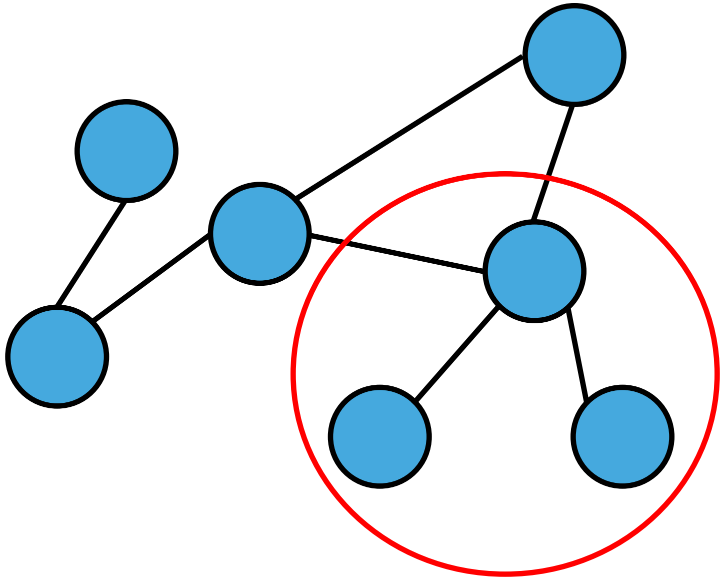
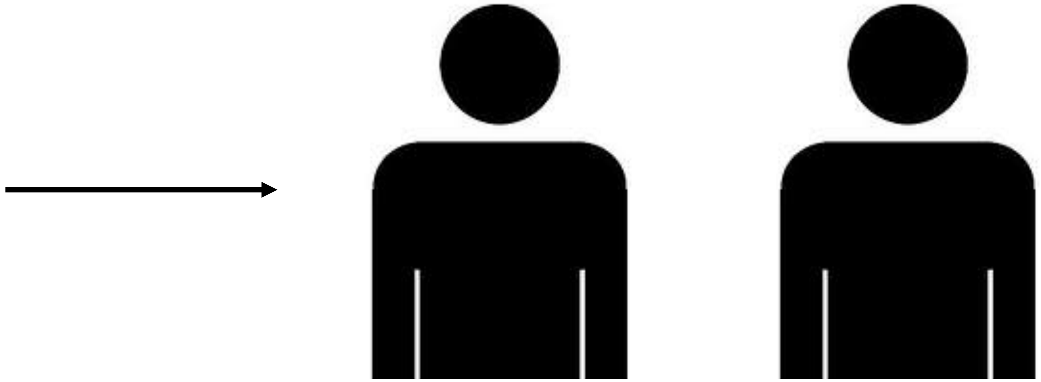


Figure How many cliques of 3 does the graph possesses?

A social challenge: privacy

Data must be anonymous enough to retain privacy while still being useful

- The rise of big data has multiplied privacy issues: linking attacks are more potent
- How is our privacy affected when we participate in such a study? Do aggregate results reveal anything about me?
- These concerns have been addressed by *differential privacy* (C. Dwork, A. Roth, *The Algorithmic Foundations of Differential Privacy*)



Inferring Activities and Optimal Trips

Lessons from Singapore's National Science Experiment

Thank you!

Any questions?