# Global Future Cities Programme

## Smart GIS Training: Session 3
## Advanced Analytics

7 December 2021

# Introductions

Jonathan McCallum
Senior Consultant

George Doughty
Data Scientist

Matthew Fredericks
Data Scientist

# Agenda

1. Overview
2. Advanced Analytics
3. Application
4. Practical

## Sentiment Survey

Please complete quick survey for later… "sentiments"

https://forms.office.com/r/WmuPKEYiUd

FUTURE CITIES
Global Future Cities Programme

# Overview

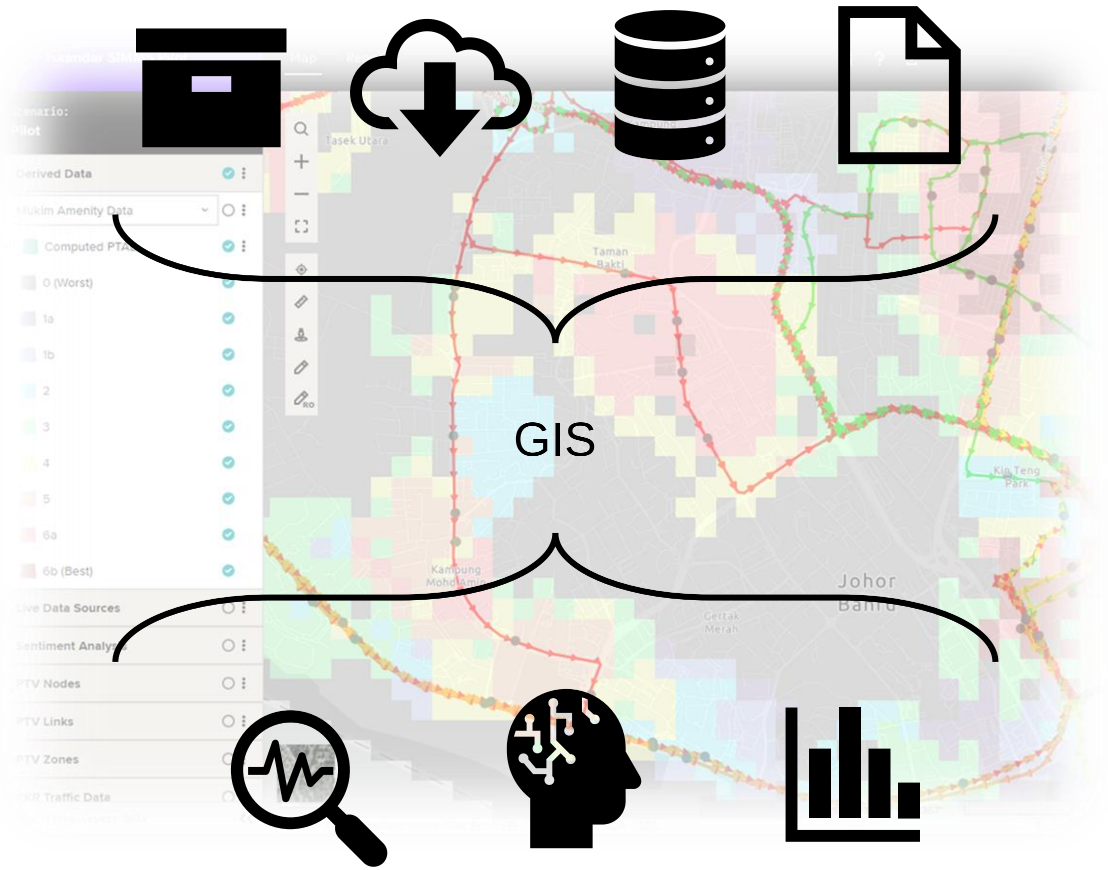
# Overview

Over the next 3 sessions we will look at the work performed to produce the Smart GIS as part of the pilot project on the Iskandar intervention:

- Collect Data
- Process it into GIS formats
- Apply analytics
- Produce visualisations
- Generate additional functionality

We will cover:

- GIS Fundamentals
- Derivation of Data
- Advanced Analytics



# Overview

## Recap from Session 1 and 2

### Theory

- GIS Fundamentals (best practice, naming conventions, data formats)
- Online storage and interaction (direct links, APIs)

### Application (via SIMMS)

- ArcGIS servers (Moata Platform)
- Data collection

### Practical

- Basic GIS operations
- Publishing to ArcGIS Online for data collection

### Theory

- Urban and transport planning metrics
- Concepts of connectivity, entropy and mobility

### Application (via SIMMS)

- How the metrics were derived and use of a master shapefile
- Granularity as a critical aspect of work
- Dashboards and geospatial tools

### Practical

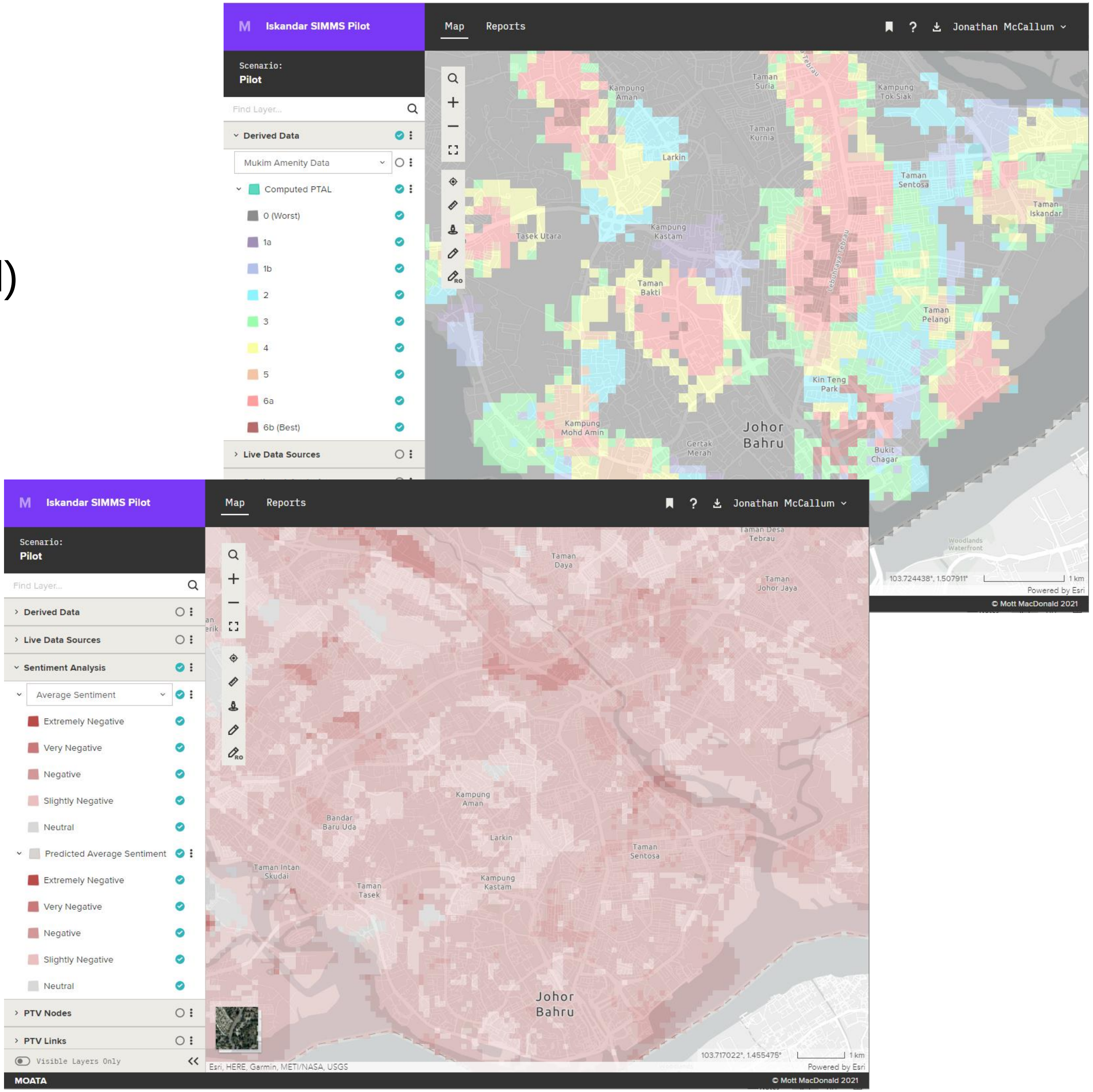
- Performing spatial queries to derive new metrics

Global Future Cities Programme

# Advanced Analytics

# Advanced Analytics

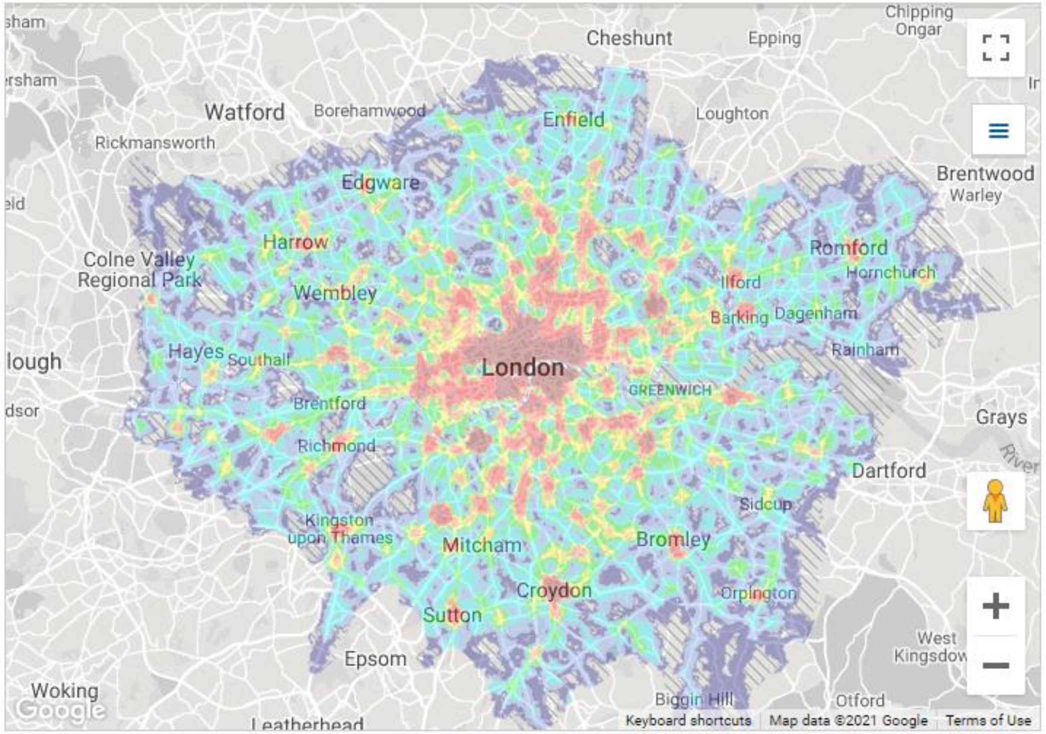
- PTAL (Public Transport Accessibility Level)
- Sentiment analysis of complaints
- Machine learning prediction



# PTAL theory

## A measure of connectivity to Public Transport

- Divide the region into 100m x 100m squares.
- Calculate the distance to the nearest bus stops (within 1km) from the centre of each grid.
- For each bus stop within the radius calculate:
  - the walk time (based on a configurable walking speed): $T$
  - the standard waiting time: $\mathrm{SWT} = \frac{1}{2} \times \frac{60}{Bus\ Frequency}$
  - the average waiting time: $AWT = SWT + 2$
  - Total Access Time: $TAT = T + AWT$
  - equivalent doorstep frequency: $\mathrm{EDF} = \frac{1}{2} \times \frac{60}{TAT},$
  - Access index: $AI = Max(EDF) + 0.5 \times \sum_{Bus\ Stops} EDF$
    (Used for grids with more than one bus stop)
  - Covert to PTAL using bandings of Access Index scores.



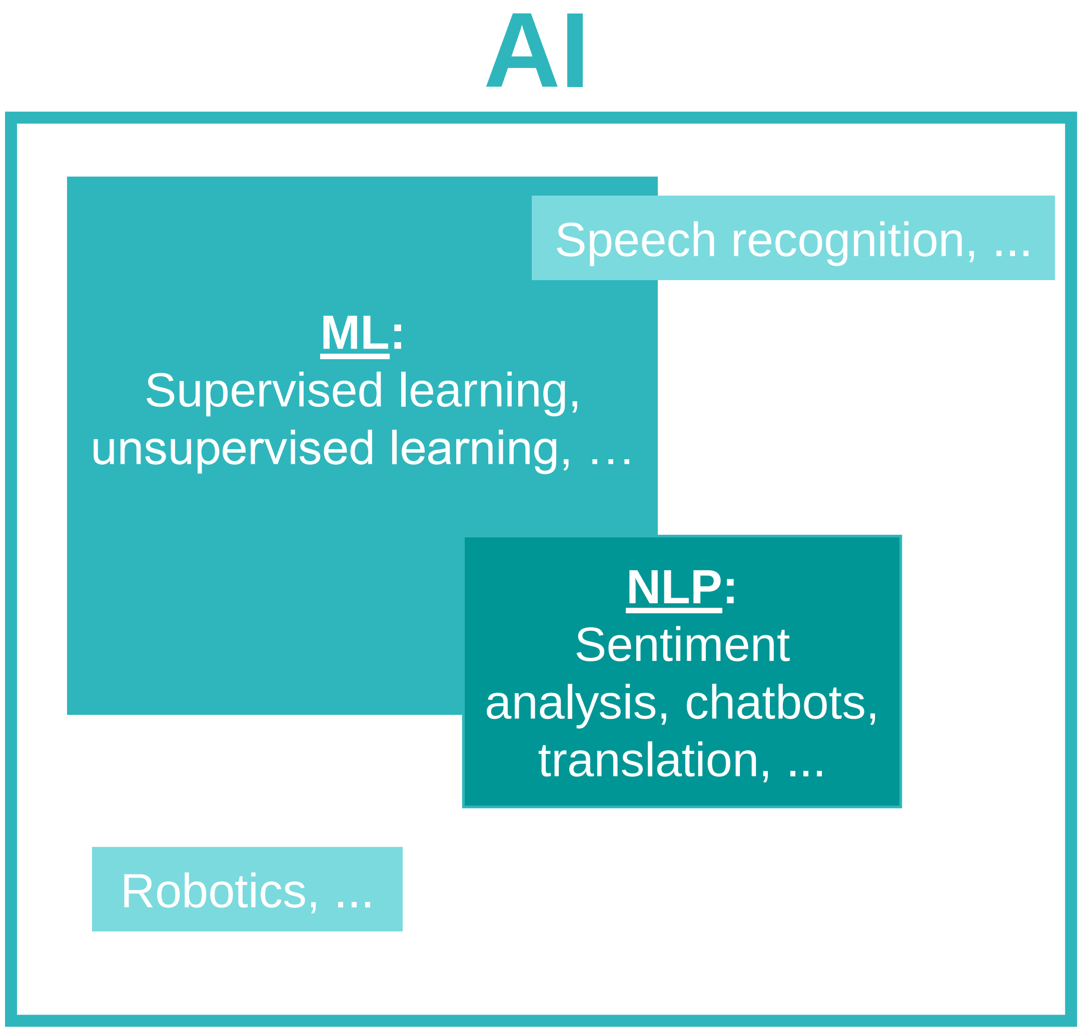

More information can be found at: https://content.tfl.gov.uk/connectivity-assessment-guide.pdf

# Some definitions

- **Artificial intelligence (AI)** - use of computing to mimic human decision-making and functions
- **Machine learning (ML)** - use of algorithms to "learn" patterns in data
  - **Supervised learning** – teaching with examples.

    e.g. Identifying cats vs dogs.
  - **Unsupervised learning** - teaching without examples

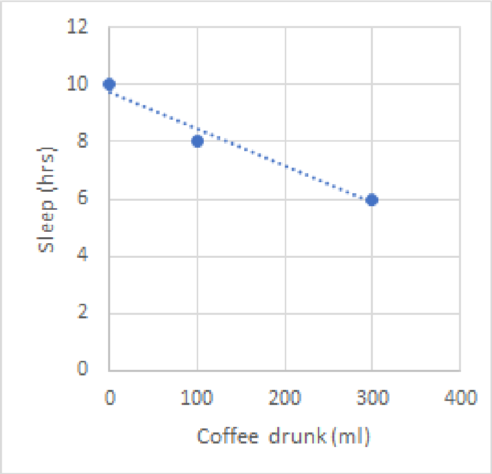
    e.g. grouping similar pets together.
- **Natural language processing (NLP)** - use of algorithms to process human language
  - **Sentiment analysis** - aims to quantify the sentiment (attitude or emotion) behind text

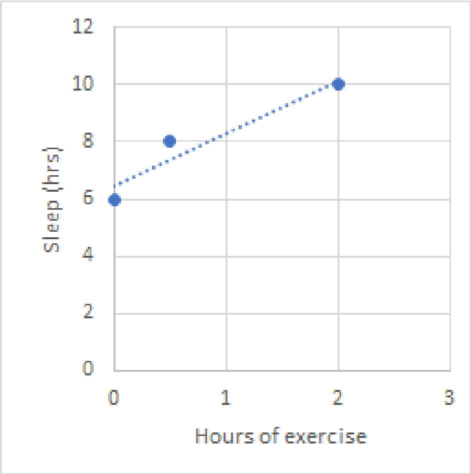


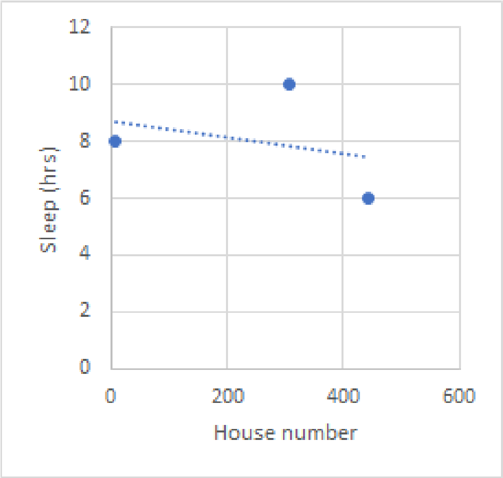
# Machine learning theory

## Linear regression

| Coffee drunk (ml) | Hours of exercise | House number | Hours of sleep |
|---|---|---|---|
| 300 | 0 | 442 | 6 |
| 100 | 0.5 | 7 | 8 |
| 0 | 2 | 308 | 10 |
| 500 | 1 | 25 | **?** |




- **Linear regression** is a ML algorithm that fits a linear relationship between **features** and the **label**
  - Minimise the loss in the objective function:

  Hours of sleep = **A** × coffee drunk + **B** × hours of exercise + **C** × house number + **D**

  - For the example above, after machine learning we might get**:**

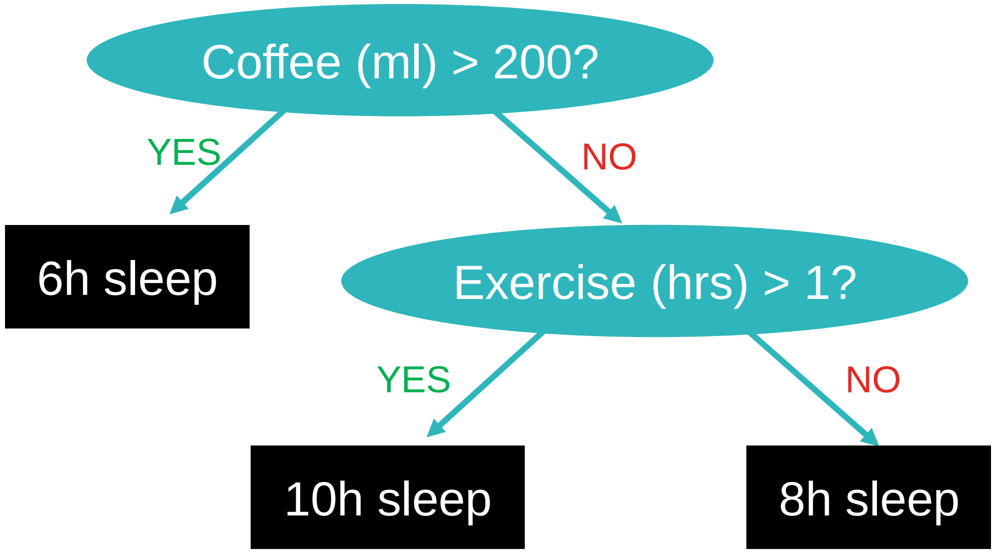
  **A = -0.01, B = 2, C = 0.0001, D = 8**

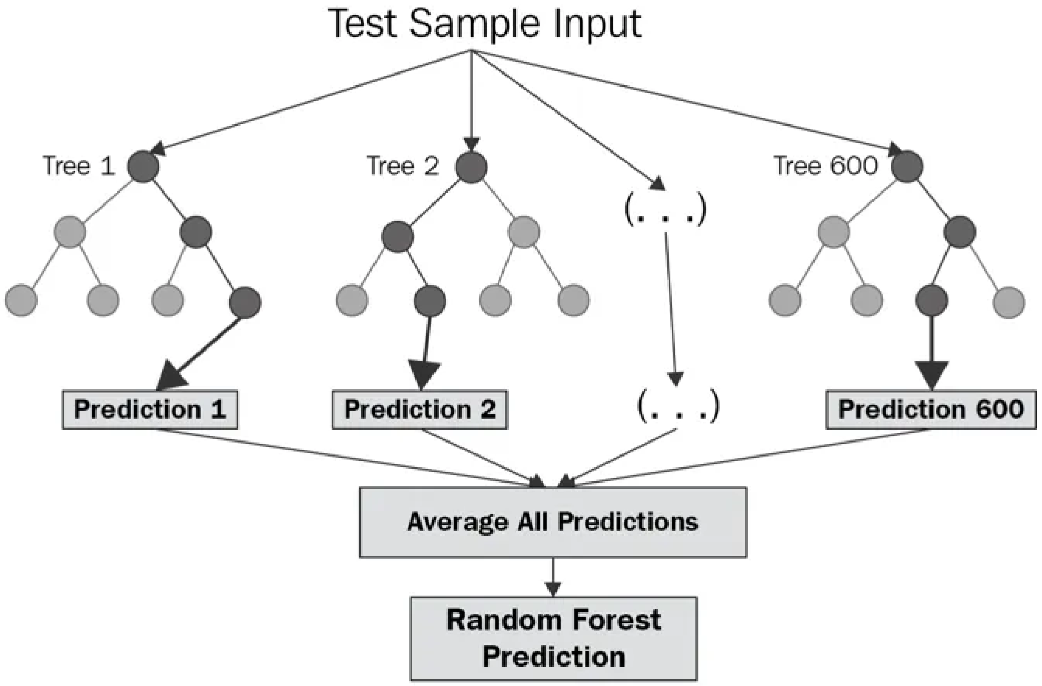
  Meaning for the highlighted row, the model predicts:

  (-0.01)*500 + 2*1 + 0.0001*25 + 8 = **5.0025 hrs of sleep**

# Machine learning theory

## Decision trees and random forests




A **decision tree** is a simplistic ML algorithm that makes a prediction by following a series of splits based on features. For a continuous label, the prediction is just the average of the training labels at that end node.

A **random forest** is a ML algorithm that uses a collection of decision trees which each independently find the best way to separate the training labels.

- A feature's importance is measured by how well it's able to affect the split of training labels.

# Sentiment analysis theory

## Training a word scoring model

**Result**

**Training data**

"I like pizza" : positive

"I love pizza" : positive

"I hate pizza" : negative

"I": neutral

"like": positive

"love": positive

"hate": negative

"pizza": neutral

**Applying the model to new data**

"I like oranges": **?**

Neutral + positive + neutral

= positive

# Sentiment analysis theory

## Grammar and syntactical rules

- However, human languages aren't that simple!

"I don't like oranges": **?**
Neutral + negative + positive + neutral
= neutral

But this should be negative!

"I like oranges": positive
"I love oranges": positive
"I adore oranges": positive

We want to score/rank these!

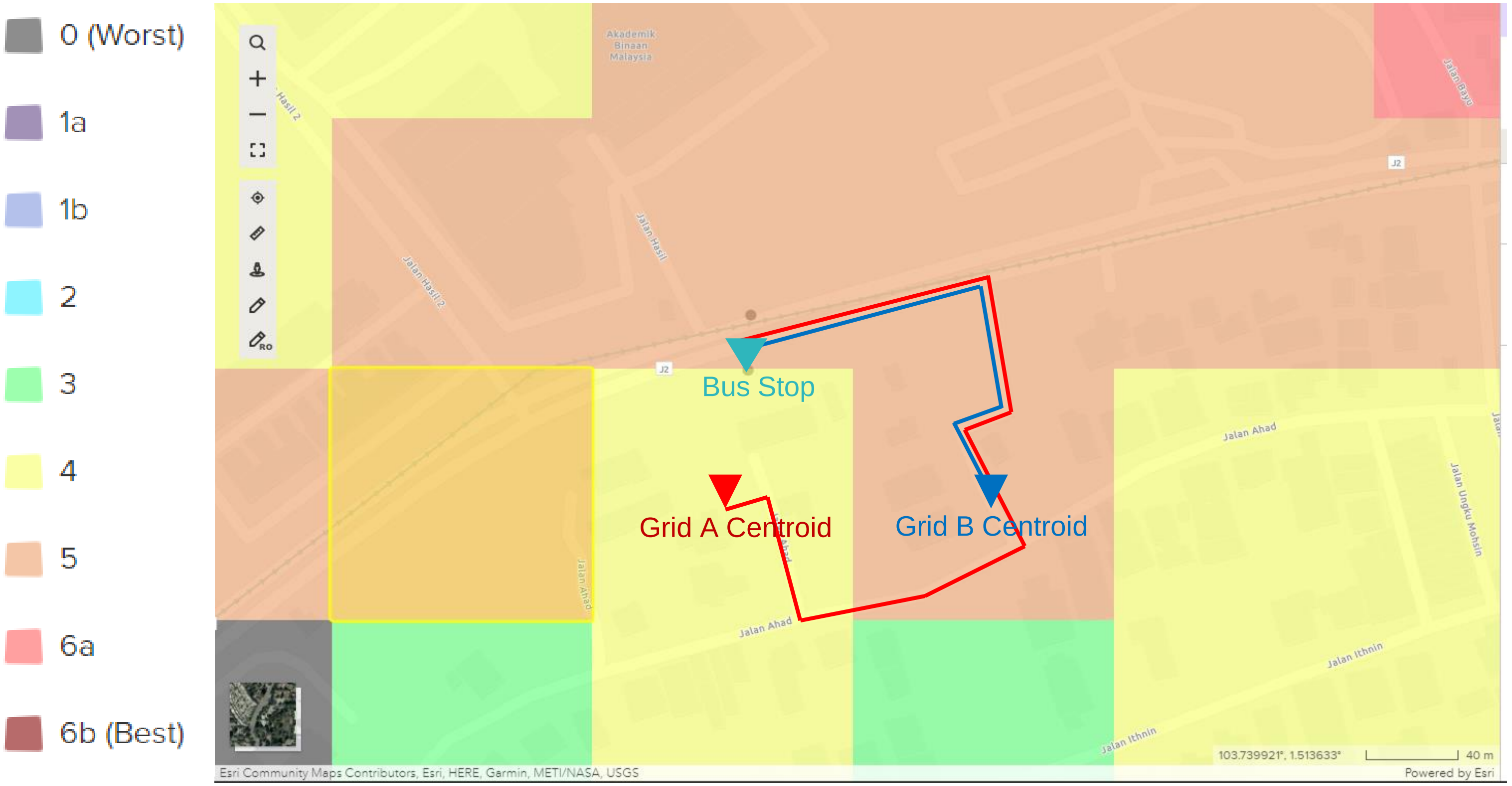
- Solution: word vectors and other NLP techniques to capture word relationships.

| | |
|---|---|
| "I like oranges" | [1 0.6 1] |
| "I love oranges" | [1 0.8 1] |
| "I adore oranges" | [1 0.9 1] |
| "I don't like oranges" | [1 -0.6 1] |

# Application

# PTAL example

## Worked example of PTAL Calculation




| | Walk Time | Average Wait Time (AWT) | Equivalent Doorstep Frequency (EDF) | PTAL Score |
|---|---|---|---|---|
| Grid A | $7\ Minutes$ | $(0.5\ \times 20) + 2 = 12\ Minutes$ | $\frac{60}{2\ \times 19} \approx 1.579$ | $1.5 \leq EDF < 1.75\ \rightarrow PTAL = 4$ |
| Grid B | $4\ Minutes$ | $(0.5\ \times 20) + 2 = 12\ Minutes$ | $\frac{60}{2\ \times 16} = 1.875$ | $1.75 \leq EDF < 2 \rightarrow PTAL = 5$ |

# Sentiment analysis example

## Iskandar citizen complaints data

lampu jalan rosak

the street lights are broken

-0.47

... di jalan kemuliaan

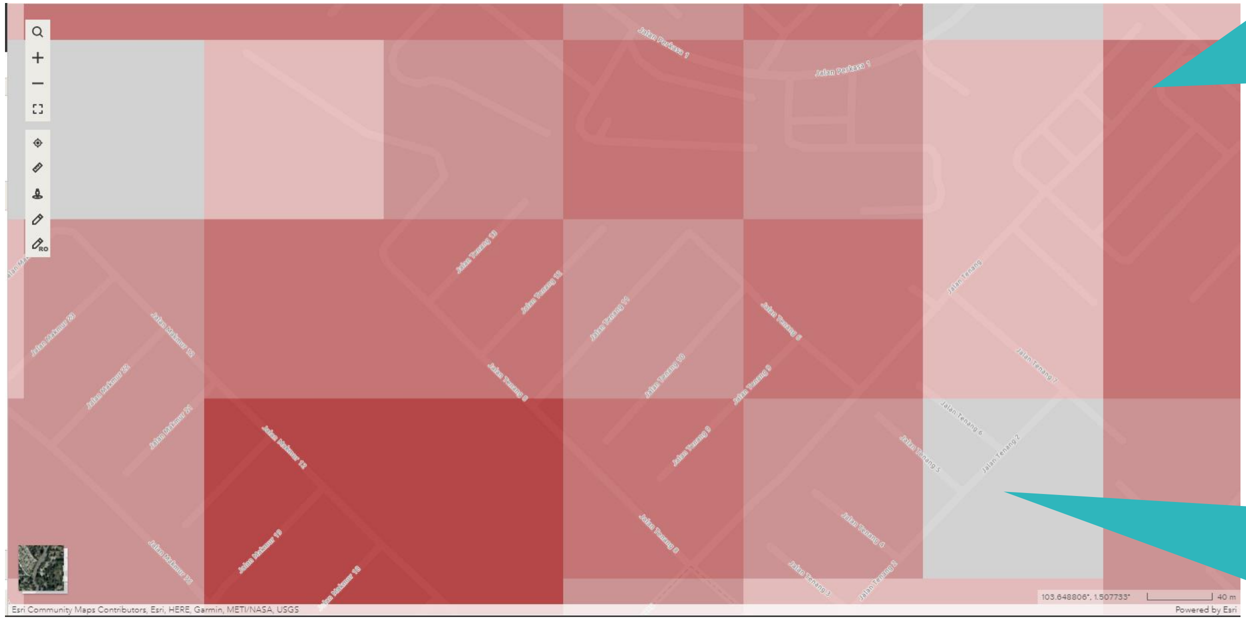
... on glory street

2.15

- Common problems (e.g. translating place names, or unusual words not having scores)
- We expanded the vocabulary of our sentiment "lexicon" using a pre-trained model (on many GB of Google News data) to check unknown words for "similar" known words

# Machine learning example

## Framing the problem



**Sentiment: -1.55**
Mean vehicle speed: 60km/h
Number of accidents: 15
Number of primary schools: 1
…

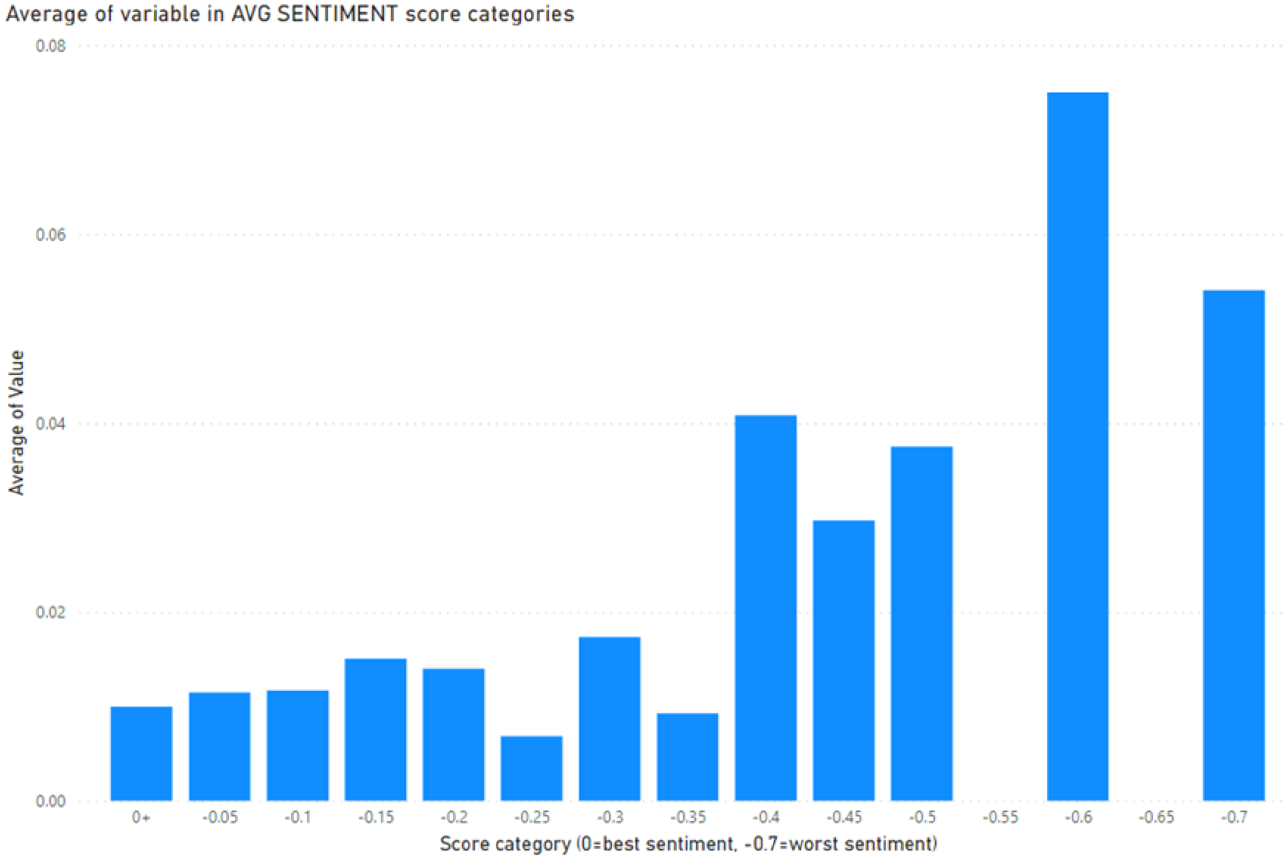
**Sentiment: -0.32**
Mean vehicle speed: 40km/h
Number of accidents: 3
Number of primary schools: 2
…

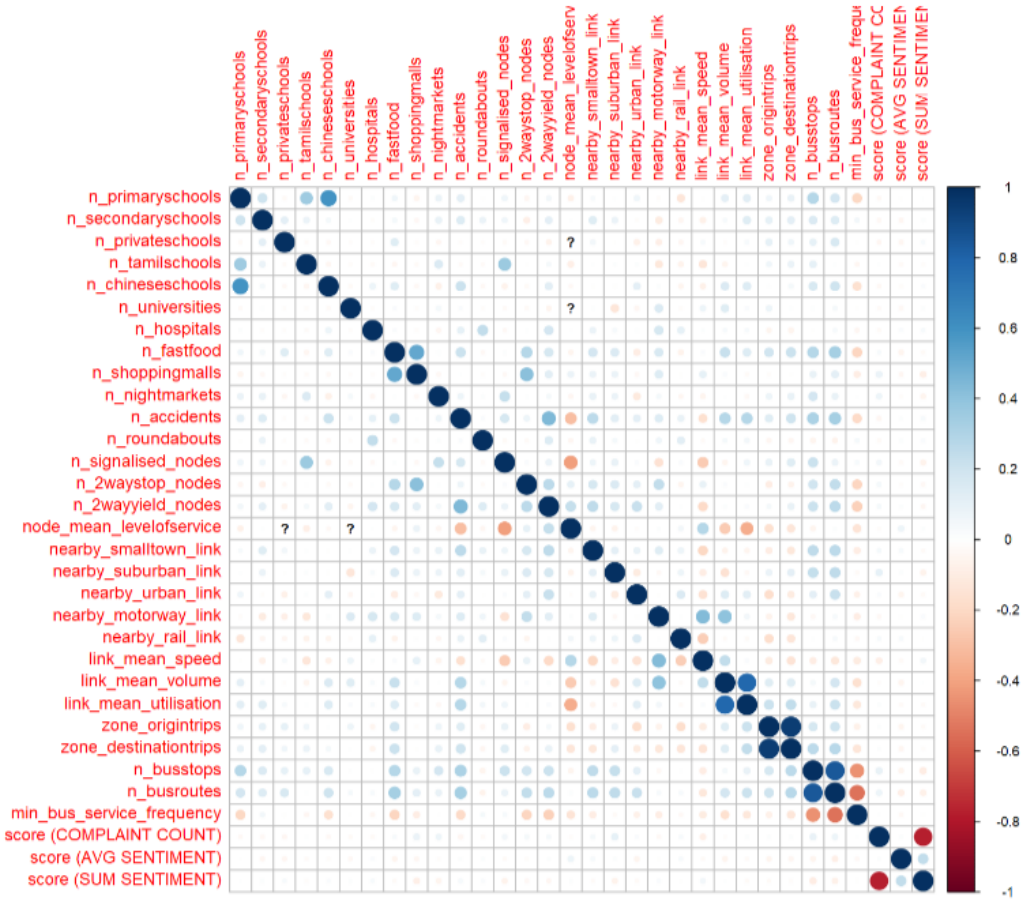
Why machine learning?

- Understand more about causes for complaints.
- Make predictions for areas where we lack complaints data.

# Machine learning example

## Exploratory analysis



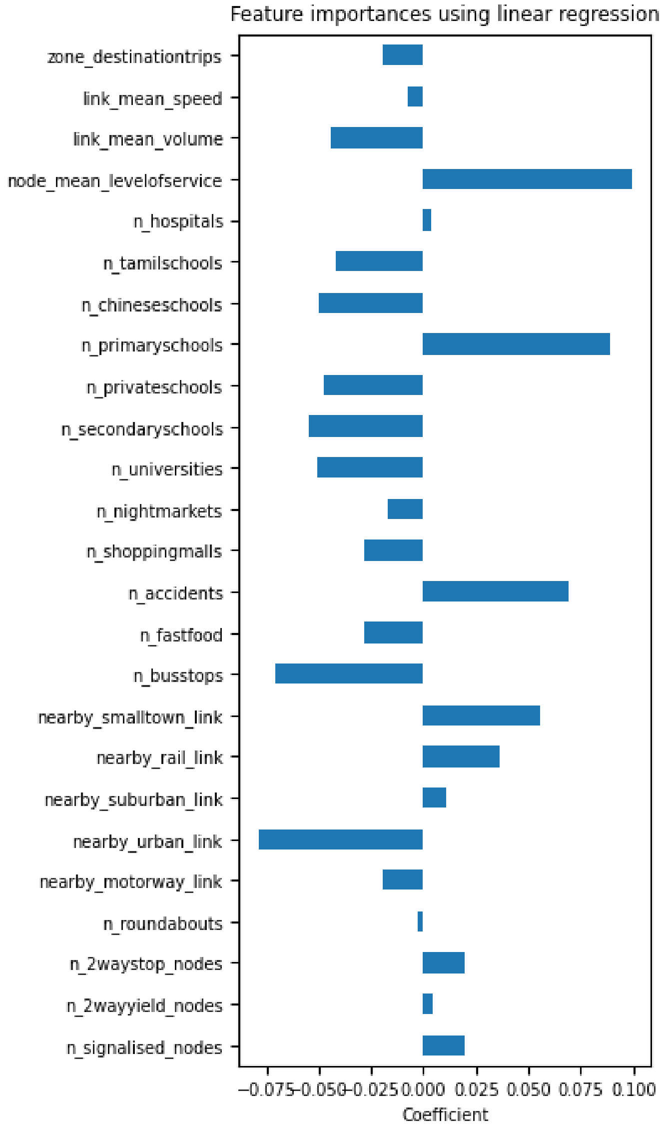
# Machine learning example

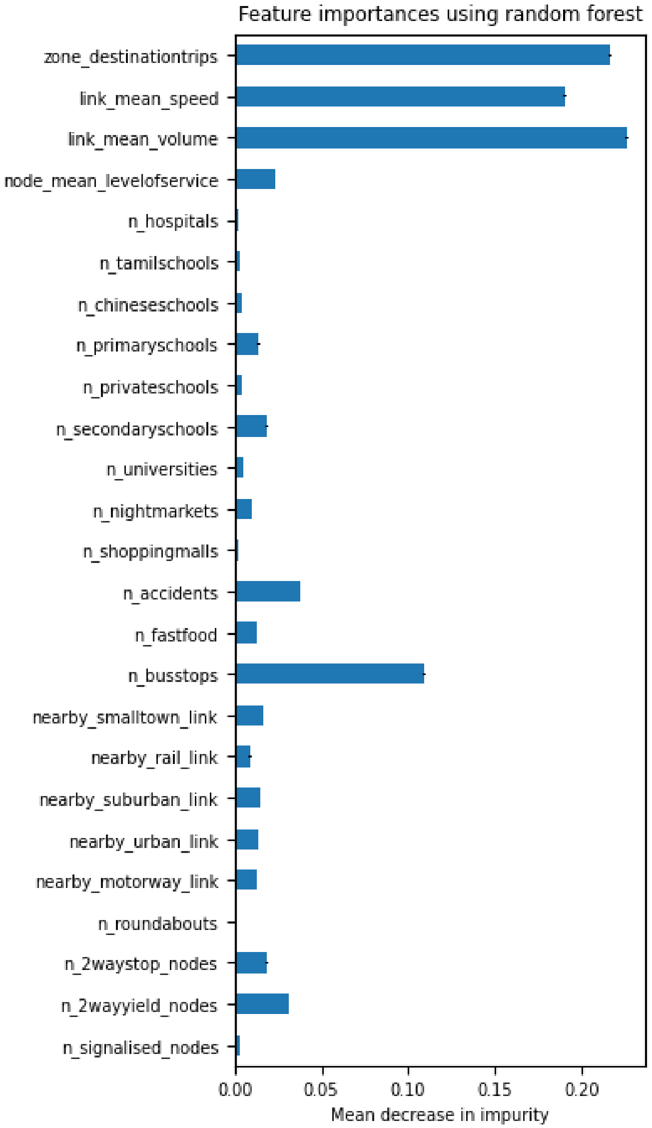
## Prediction of sentiment

- Train the model...

# Machine learning example

## Results




- Areas with busy roads and zones have worse sentiment.
- Sentiment is worse in areas with many bus stops, possibly as they are in busier urban centres with more destinations where complaints could arise.
- Areas near primary schools have better sentiment, but worse for other educational institutes.

# Machine learning example

## Limitations

- Average sentiment is dependent on the likelihood of complaint submission and the accuracy of our sentiment calculation. It is not an observable outcome such as a census result on quality of life.
- Confounding factors such as population would invalidate our insights and mean that we may be inadvertently predicting population hotspots rather than complaint sentiment.
- We are missing suitable feature data on important quantities like population, salary and poverty. The features we do have may not be linked to complaints at all (evidenced by correlogram).

# Practical

# Practical conclusion

## How can we get better results?

(i) useful data (data that is relevant to the question we are trying to answer, for example socio-economic data like qol, poverty levels when trying to predict sentiment);

(ii) high quality data (data that is accurate and well-labeled, with datasets that cover a large shared area).

# Summary

# Summary

Theory

- Machine Learning
- Sentiment Analysis

Application (via SIMMS)

- Deriving PTAL scores
- Citizen complaint data

Practical

- Simple sentiment analysis example

FUTURE CITIES
Global Future Cities Programme

# Thank you