Taxi network data

Taxi dataset: Introduction

- 442 taxis running in the city of Porto (Portugal)
- **Motivation**: Predict the passenger demand

How the data was collected

- A new trip is dispatched from **A** to **B**

- **GPS coordinates** are collected every 15 seconds

  They are stored in a list of pairs (x, y) → the attribute **POLYLINE**

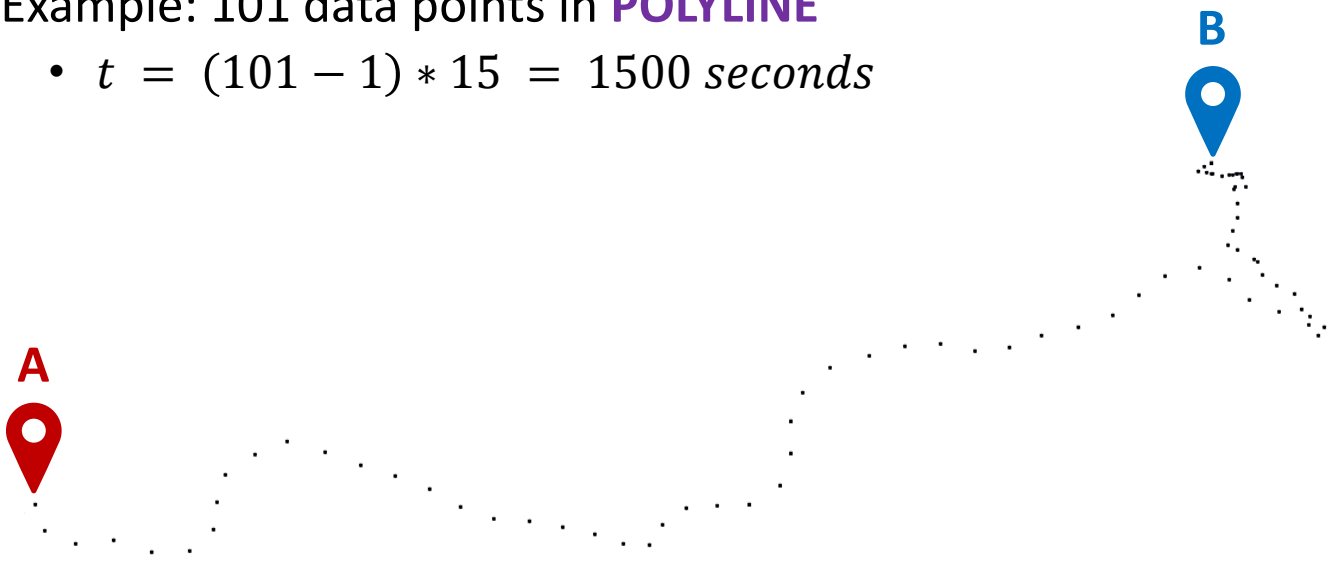
  $[(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)]$

  A $= (x_1, y_1)$

  B $= (x_n, y_n)$

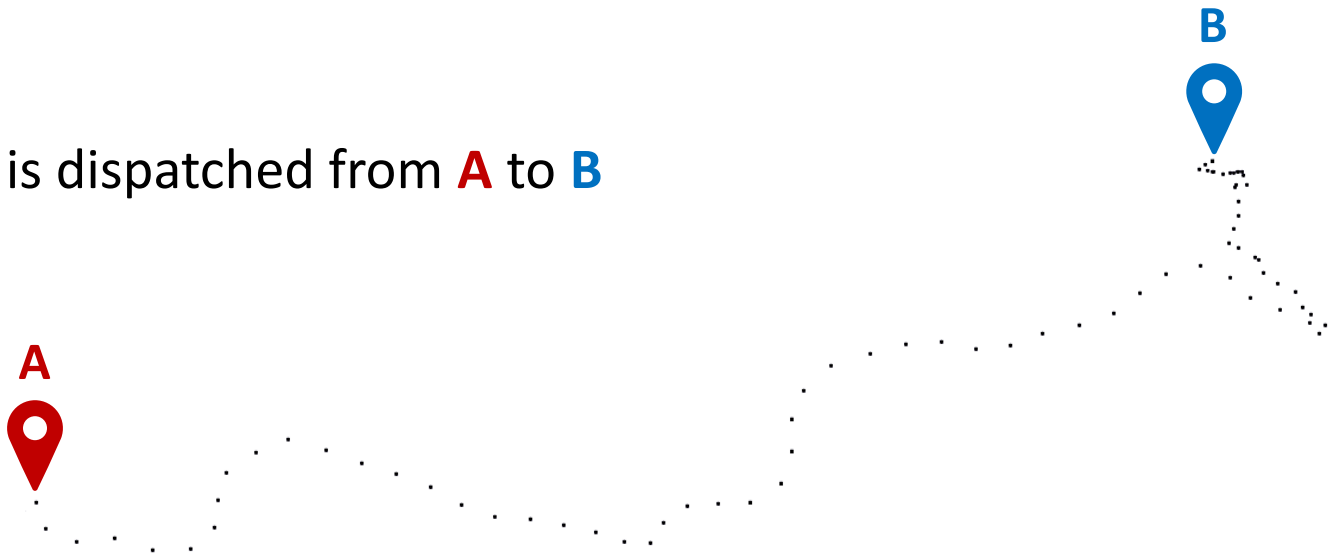
How the data was collected

- Time of the trip
- **Total travel time** t of a trip
  - n is the number of points
  - $t = (n-1) * 15\ seconds$
  - Example: 101 data points in **POLYLINE**
    - $t = (101-1) * 15 = 1500\ seconds$



How the data was collected

- A new trip is dispatched from **A** to **B**



- How was the trip requested?
  - Examples:
    - Dispatched from the central
    - Directly to a taxi driver on a specific stand
    - On a random street
  - Ride categories → the attribute **CALL_TYPE**

The dataset

- 442 taxis running in the city of Porto (Portugal)
- 1 710 670 trips collected!
  - From 01/07/2013 to 30/06/2014
- Ride categories:
  - ‘A’: taxi central based
  - ‘B’: stand-based
  - ‘C’: non-taxi central based

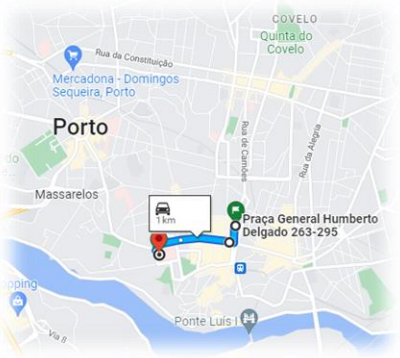
Attributes (original dataset)

- **TRIP_ID**: trip identifier
- **CALL_TYPE**: how the trip was dispatched/demanded
- **ORIGIN_CALL**: identifier for each phone number which was used to demand, at least, one service
- **ORIGIN_STAND**: taxi stand identifier
- **TAXI_ID**: taxi identifier
- **TIMESTAMP**: when the trip started
- **DAYTYPE**: workday, weekend, holiday, …
- **MISSING_DATA**: is there missing points in 'POLYLINE'?
- **POLYLINE**: list of GPS coordinates

Attributes (original dataset)

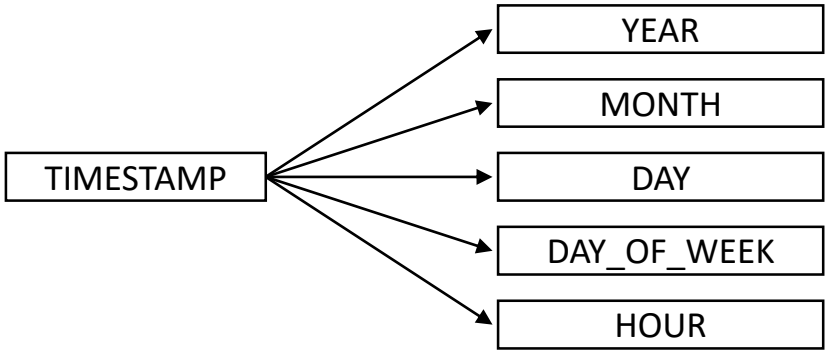
TRIP_ID	CALL_TYPE	ORIGIN_CALL	ORIGIN_STAND	TAXI_ID	TIMESTAMP	DAY_TYPE	MISSING_DATA	POLYLINE
1372636858620000589	C	NaN	NaN	20000589	1372636858	A	False	[[-8.618643,41.141412],[-8.618499,41.141376],[...
1372637303620000596	B	NaN	7.0	20000596	1372637303	A	False	[[-8.639847,41.159826],[-8.640351,41.159871],[...
1372636951620000320	C	NaN	NaN	20000320	1372636951	A	False	[[-8.612964,41.140359],[-8.613378,41.14035],[-...
1372636854620000520	C	NaN	NaN	20000520	1372636854	A	False	[[-8.574678,41.151951],[-8.574705,41.151942],[...
1372637091620000337	C	NaN	NaN	20000337	1372637091	A	False	[[-8.645994,41.18049],[-8.645949,41.180517],[-...
1372639092620000233	C	NaN	NaN	20000233	1372639092	A	False	[[-8.632737,41.168295],[-8.6328,41.16825],[-8....

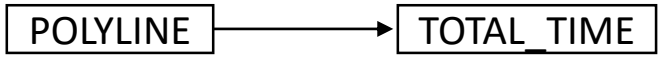
- **Example:** One data instance (one trip)
  - **TRIP_IP** = 1372639092620000233
  - Ride type = ‘C’: non-taxi central based
  - Taxi ID = 20000233
  - Timestamp = 1372639092:
  - Day type = A: This trip was dispatched from the central
  - Missing data = False
  - POLYLINE = [(-8.611065,41.149431),(-8.611209,41.149368),…,(-8.619993,41.146839)]
    - 11 points
    - $t = (11 - 1) * 15 \ = \ 150 \ seconds$



Pre-processing: Modified dataset

- New attributes can be extracted from the current attributes
  - Information from attributes can be merged or divided
- **Why?**
  - Gain/obtain “explicit” information
  - Increase interpretability/explainability!
  - Semantic and format
    - Adjust the data to a suitable format required by an algorithm





Modified dataset: attributes

- **TRIP_ID**: trip identifier
- **CALL_TYPE**: how the trip was dispatched/demanded
- **ORIGIN_CALL**: identifier for each phone number which was used to demand, at least, one service
- **ORIGIN_STAND**: taxi stand identifier
- **TAXI_ID**: taxi identifier
- **TIMESTAMP**: when the trip started
- **DAYTYPE**: workday, weekend, holiday, …
- **MISSING_DATA**: is there missing points in 'POLYLINE'?
- **POLYLINE**: list of GPS coordinates
- **TIME**: total time of the trip (in seconds)
- **YEAR**: 2013 or 2014
- **MONTH**: January, …, December (0 to 11)
- **DAY**: 1 to 31
- **DAY_OF_WEEK**: Monday, …, Sunday
- **HOUR**: 0 to 23

Modified dataset

- **Reduced version**: 1% of the original dataset (**17 106 trips**)

TRIP_ID	CALL_TYPE	ORIGIN_CALL	ORIGIN_STAND	TAXI_ID	TIMESTAMP	DAY_TYPE	MISSING_DATA	POLYLINE	YEAR	MONTH	DAY_OF_WEEK	DAY	HOUR
1380165372620000252	C	NaN	NaN	20000252	2013-09-26 03:16:12	A	False	[[-8.610336,41.153436],[-8.61156,41.153481],[-...	2013	September	Thursday	26	3
1379613883620000295	B	NaN	50.0	20000295	2013-09-19 18:04:43	A	False	[[-8.628741,41.169798],[-8.62857,41.16978],[-8...	2013	September	Thursday	19	18
1376186205620000446	C	NaN	NaN	20000446	2013-08-11 01:56:45	A	False	[[-8.613549,41.147019],[-8.613486,41.146569],[...	2013	August	Sunday	11	1
1399016682620000364	B	NaN	57.0	20000364	2014-05-02 07:44:42	A	False	[[-8.610804,41.145678],[-8.610795,41.145687],[...	2014	May	Friday	2	7

A vous de jouer !

1. **Access** to the **resources**
   - Go to the **page**: http://t.ly/-3kv
   - … or follow the QR Code →
2. **Download** the modified dataset
3. **Launch** the Weka interface
4. **Load** the dataset into Weka