Stop Purpose Classification from GPS Data of Commercial Vehicle Fleets

source : www.usfleettracking.com/service

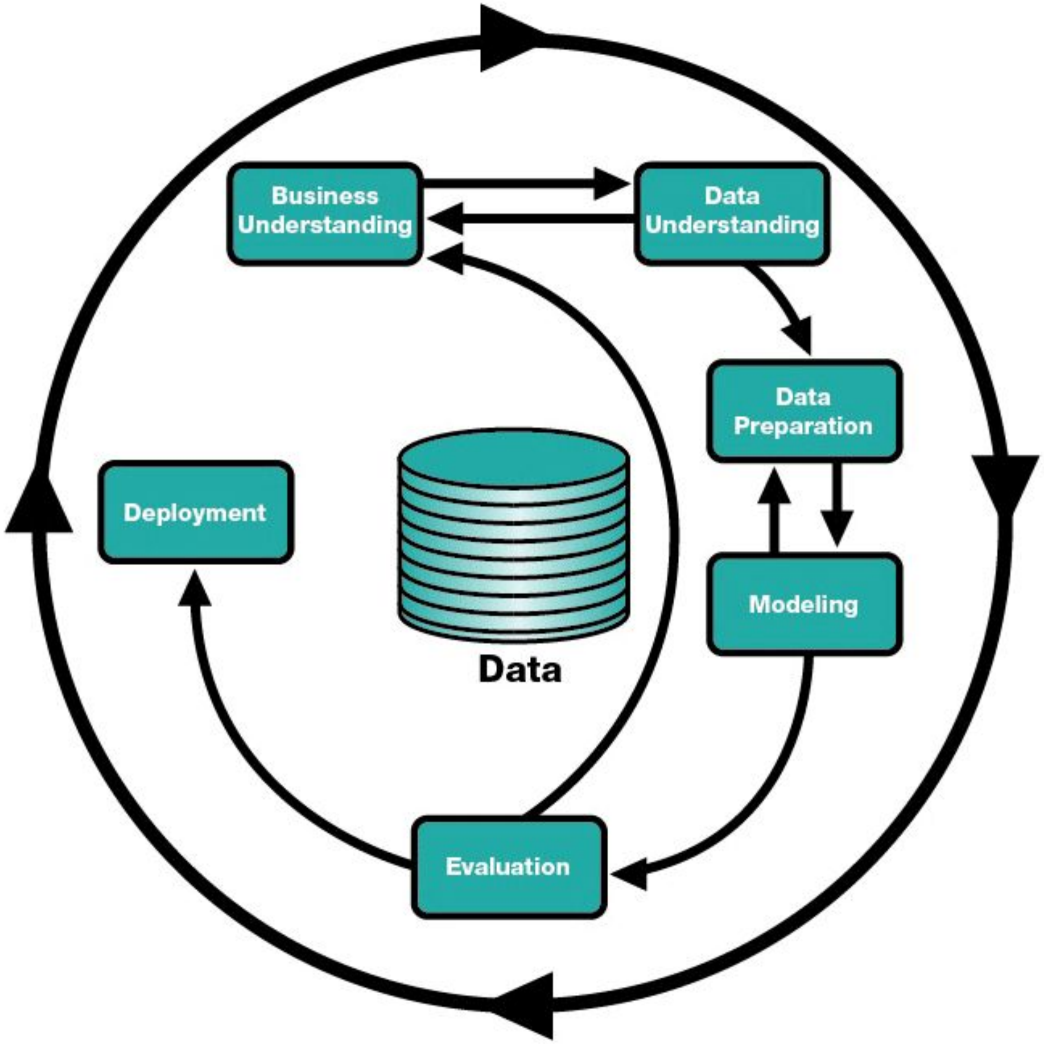
2017 IEEE ICDW Workshops
by L. Sarti et al.

Presented by

61605041 Lada Phonrungwong

CRISP-DM

Cross Industry Standard Process for Data Mining

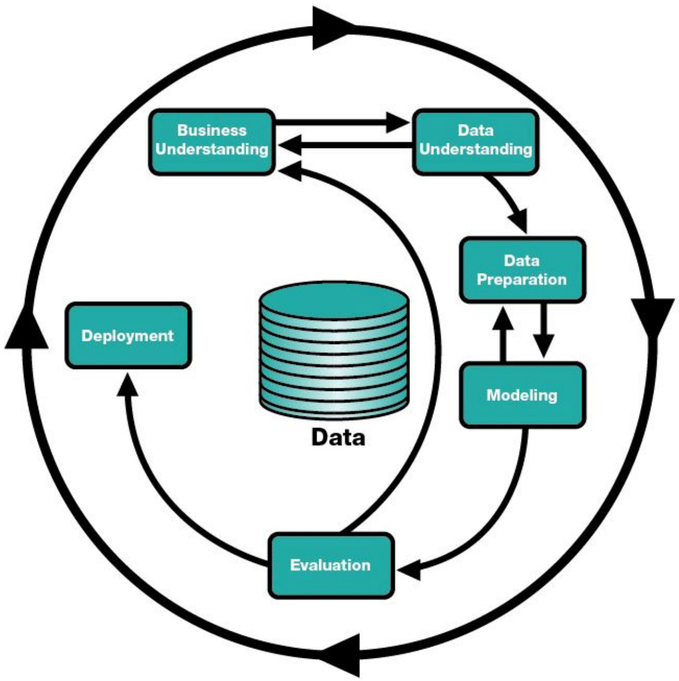


source : crisp-dm.eu/reference-model

CRISP-DM

Cross Industry Standard Process for Data Mining

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment



source : crisp-dm.eu/reference-model

Business understanding

Business understanding

Why Stop Purpose Classification ?

Who uses it ?

How it uses?

source : www.usfleettracking.com/service

CRISP-DM	1	2	3	4	5	6

Business understanding

Commercial Vehicle Fleets

source : www.tgmatrix.com/2016/02/01/machine-to-machine-technology-in-fleet-management

CRISP-DM	1	2	3	4	5	6

Business understanding

Commercial Vehicle Fleets

source : www.usfleettracking.com/service

Fleet Manager

Reeeeco , source : www.kisspng.com/users/@reeeeco.html

CRISP-DM	1	2	3	4	5	6

Business understanding

Fleet Manager

Reeeeco , source : www.kisspng.com/users/@reeeeco.html

- vehicle maintenance
- vehicle tracking
- driver management
- speed management
- fuel management

CRISP-DM	1	2	3	4	5	6

Business understanding

driver management

↓

driver behavior

↓

efficient use of their time

↓

"Work Stop" or "Non-work Stop"

↓

automatic classification

↓

Stop Purpose Classification

Data understanding

CRISP-DM	1	2	3	4	5	6

Data understanding

Raw Data ?

Where is it ?

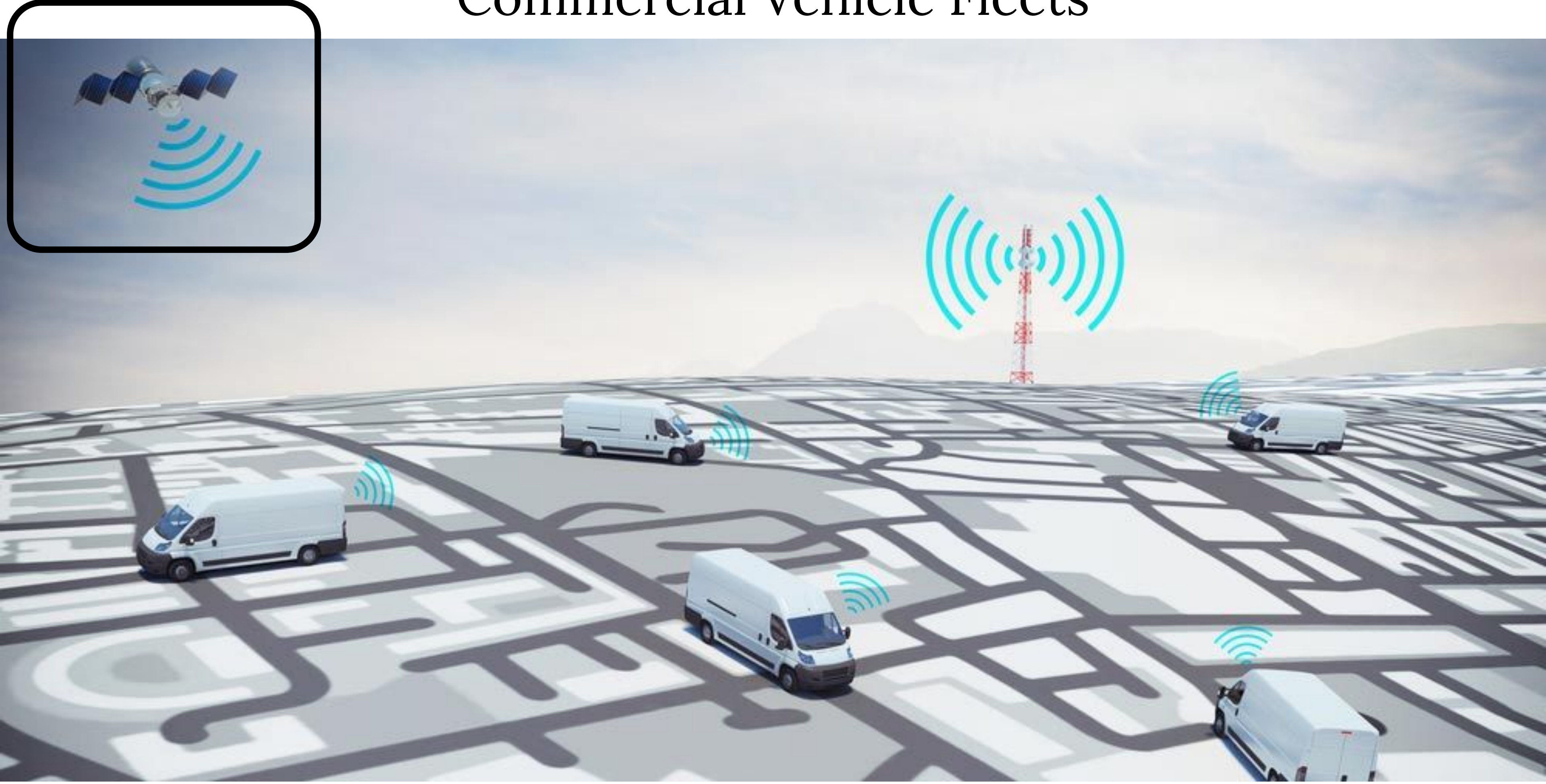
What does it look like?

source : www.usfleettracking.com/service

CRISP-DM	1	2	3	4	5	6

Data understanding

Commercial Vehicle Fleets



source : www.geospatialworld.net/blogs/3-ways-how-gps-can-help-businesses-in-fleet-management

CRISP-DM	1	2	3	4	5	6

Data understanding

Commercial Vehicle Fleets

source : www.apexsecure.in/products/assets-tracking-solutions

Data understanding

Fleetmatics®
A Verizon Company

GPS

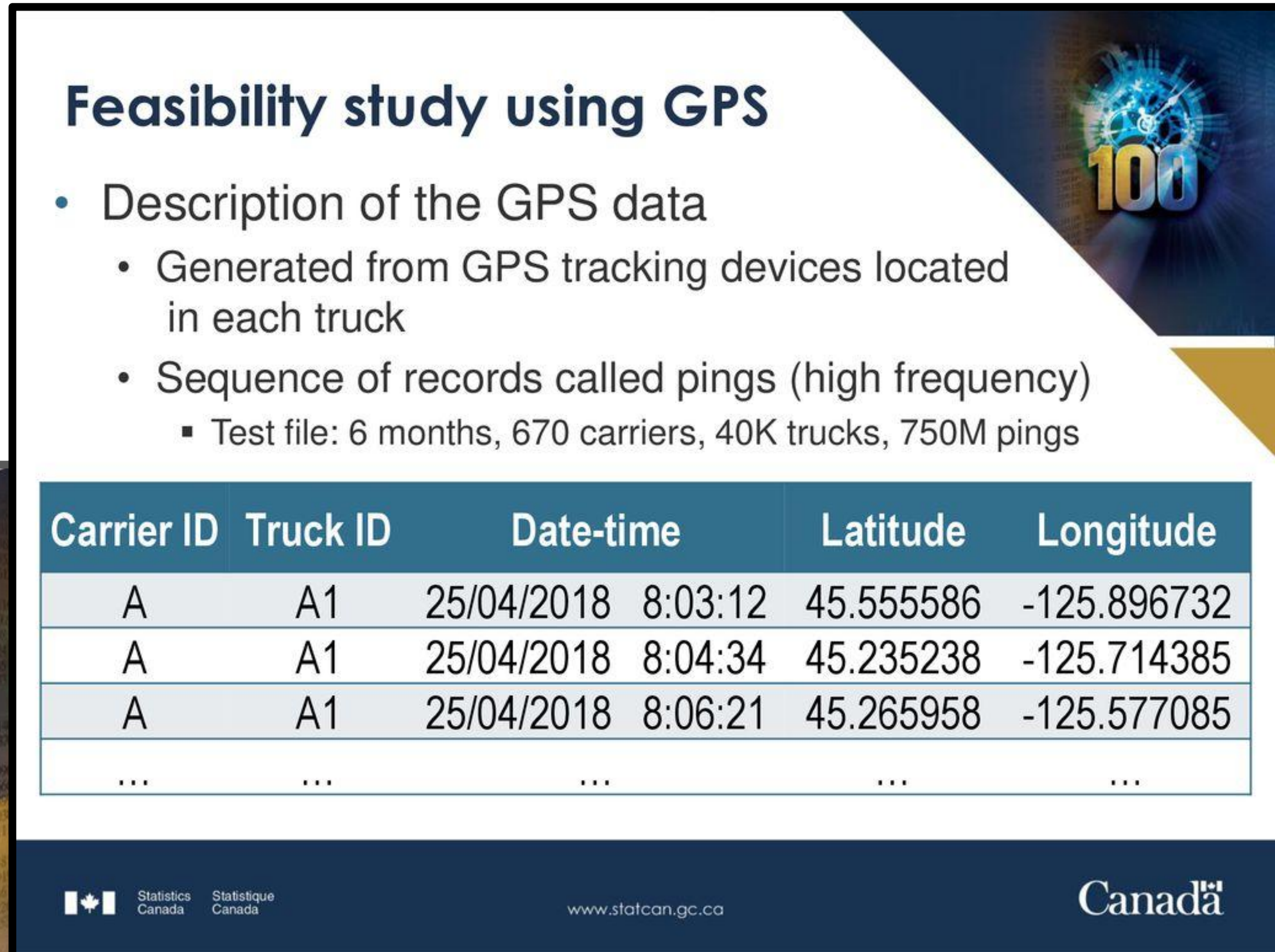
TABLE I
TYPES OF COMPANIES IN THE DATASET.

Type	Count
Heating & air conditioning	45
Plumber & leak detection	18
Protection systems	9
Maintenance & cleaning	7
Electric	6
Irrigation & lighting	5
Delivery	4
Pools installation	2
Locksmith	1
Health assistance	1

CRISP-DM	1	2	3	4	5	6

Data understanding

GPS



Feasibility study using GPS

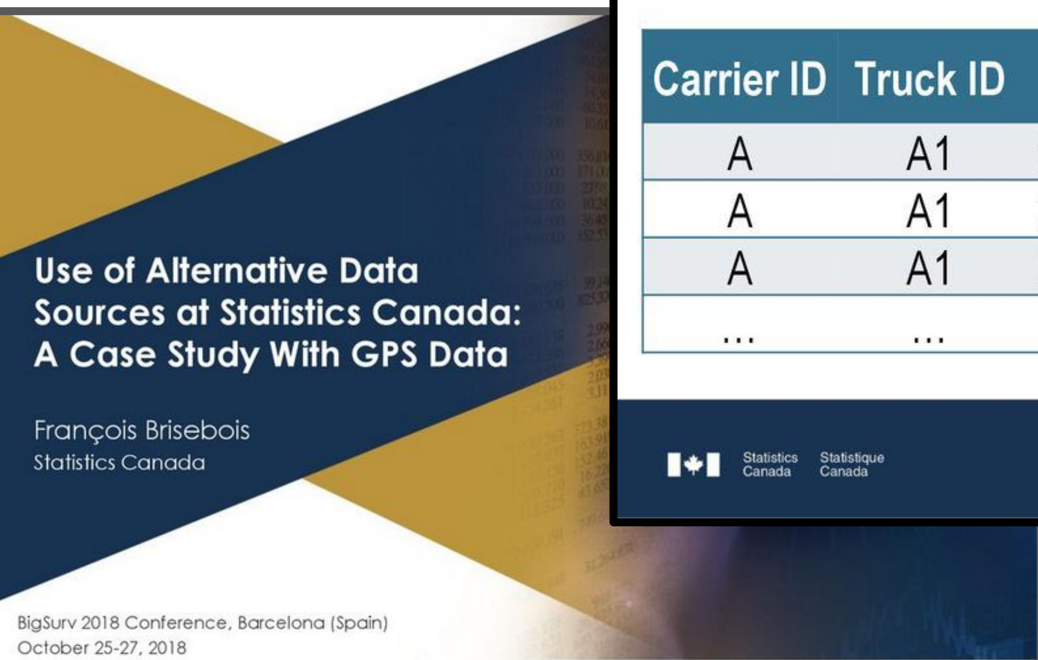
- Description of the GPS data
  - Generated from GPS tracking devices located in each truck
  - Sequence of records called pings (high frequency)
    - Test file: 6 months, 670 carriers, 40K trucks, 750M pings

Carrier ID	Truck ID	Date-time	Latitude	Longitude
A	A1	25/04/2018 8:03:12	45.555586	-125.896732
A	A1	25/04/2018 8:04:34	45.235238	-125.714385
A	A1	25/04/2018 8:06:21	45.265958	-125.577085
...	...	...	...	...

Statistics Canada Statistique Canada

www.statcan.gc.ca

Canada



source : slideplayer.com/slide/15202566

CRISP-DM	1	2	3	4	5	6

Data understanding

Dataset was collected by Fleetmatics

98 small and medium business companies

February 2015 - January 2016

The typical sampling rate ranges from 1 to 2 minutes

More than 55 million GPS pings

source : local-eyes.nl/solution/fleet-management

CRISP-DM	1	2	3	4	5	6

Data understanding

Dataset 700k stops, 160k of which are work stops

The dataset is, to the best of our knowledge,
one of the the largest & most diverse for similar problems

source : local-eyes.nl/solution/fleet-management

Data understanding

RAW DATA

- Vehicle id

- Latitude
- Longitude

- Timestamp

- Event code

Vehicle status information -> engine off Event

- Status code

schedule & progress of the jobs
-> pending, started, completed

Data preparation

CRISP-DM	1	2	3	4	5	6

Data preparation

Why can't feed Raw Data to model ?

What to do with Raw Data ?

Clean data ?
Data transformation ?
Feature engineering ?

source : www.usfleettracking.com/service

CRISP-DM	1	2	3	4	5	6

Data preparation

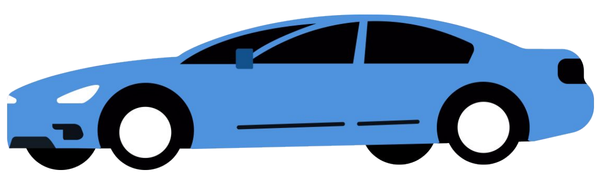
GPS pings

aggregating to activity of vehicle

Engine off

WikimediaImages , source :
pixabay.com/users/wikimediaimages-1185597

Idling



source :
requestreduce.org/categories/uber-car-clipart.html

Journey

source :
clipart-library.com/clipart/rTLo7djkc.htm111

Data preparation

Engine off pings:

These pings are generated the instant the engine is turned off

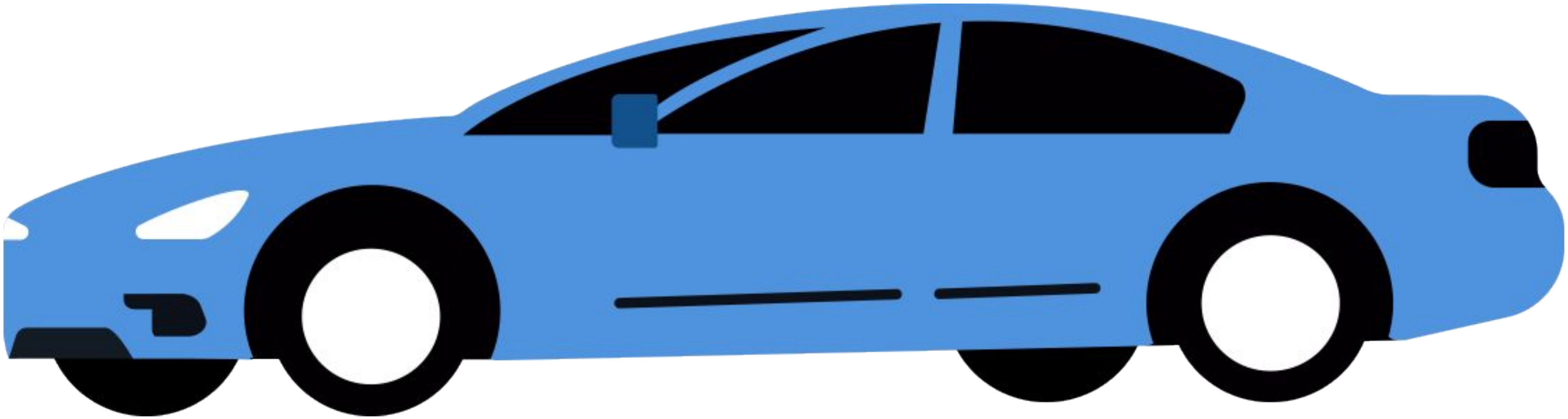
WikimediaImages , source : pixabay.com/users/wikimediaimages-1185597

CRISP-DM	1	2	3	4	5	6

Data preparation

Idling pings :

pings where the engine is on,
but the vehicle is still or moving slowly in a small area.



source : requestreduce.org/categories/uber-car-clipart.html

CRISP-DM	1	2	3	4	5	6

Data preparation

Idling pings :

DISTANT $\boxed{H(p_i, p_{i-1}) \leq 150\ m}$

SPEED $\boxed{s_i = \frac{H(p_i, p_{i-1})}{t_i - t_{i-1}} \leq 1.4\, m/s}$

roughly 5 km/h

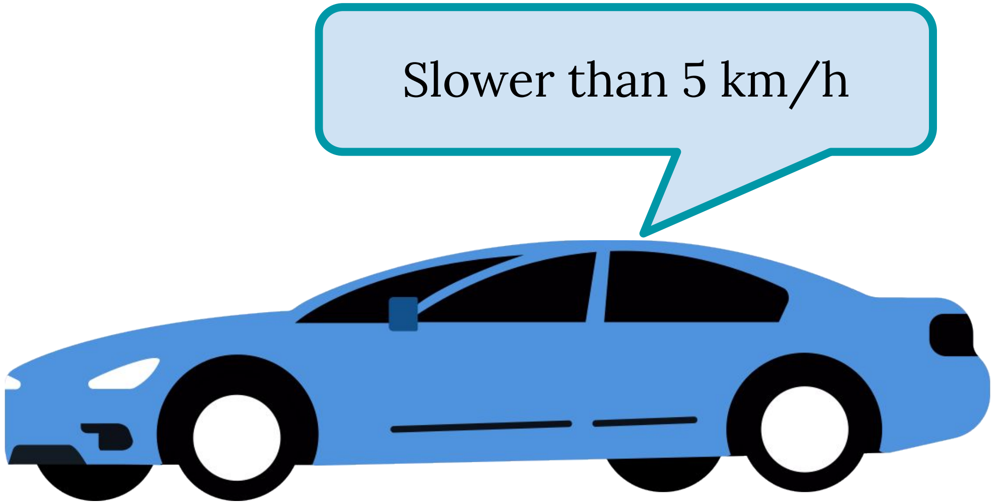
source :
www.dullmensclub.com/
walking-is-good-for-you

CRISP-DM	1	2	3	4	5	6

Data preparation

Idling pings :

=



source : requestreduce.org/categories/uber-car-clipart.html

source : www.dullmensclub.com/walking-is-good-for-you

CRISP-DM	1	2	3	4	5	6

Data preparation

Journey pings :

neither engine off nor idling

source : clipart-library.com/clipart/rTLo7djkc.htm111

CRISP-DM	1	2	**3**	4	5	6

Data preparation

"A stop include an engine off ping,

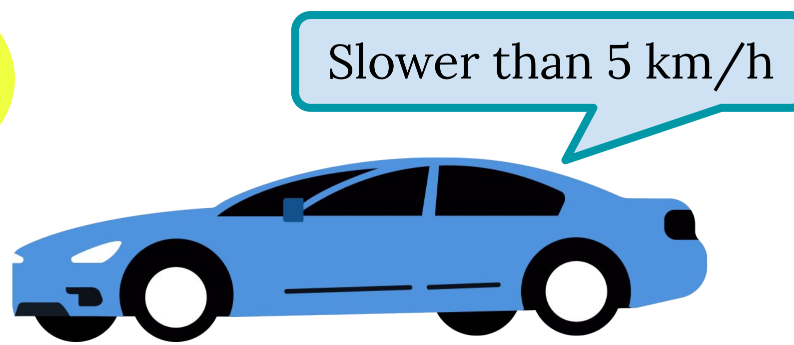
but may contain idling pings only"

Data preparation

delivery companies make their deliveries without turning off

Vehicle is stuck in the traffic or waiting at a traffic light

WikimediaImages , source :
pixabay.com/users/wikimediaimages-1185597



source : requestreduce.org/categories/uber-car-clipart.html

Data preparation

A stop may include
time spent parking also finding parking place

Loaded in multiple load points,
unloading his cargo to a client

source : www.dullmensclub.com/walking-is-good-for-you

CRISP-DM	1	2	3	4	5	6

Data preparation

Extract from them 100 different features then train a Random Forest model

Divide the features into 4 different groups:

1. Stop-wise features (SWF)
2. Points of interest features (POIF)
3. Stop cluster features (CF)
4. Sequential features (SeqF)

source : local-eyes.nl/solution/fleet-management

CRISP-DM	1	2	3	4	5	6

Data preparation

1) Stop-wise features:

"aggregation of multiple idling & engine off GPS pings"

CRISP-DM	1	2	3	4	5	6

Data preparation

1) Stop-wise features: f1:SWF

- Stop duration
  first and last ping belonging to the stop

- Start time features
  hour of day, day of week, day of month, day of year

- Time spent with the engine off
  min, max, mean, variance, sum

Data preparation

1) Stop-wise features: f1:SWF

- Shape
  stop width, stop height, stop area, stop ratio,
  from the bounding box of the included GPS Pings

- Stop type
  engine off if it contains at least an engine off ping,
  idling otherwise

- Odometer distance
  from the first ping to the last ping of the stop

Data preparation

1) Stop-wise features: f1:SWF

- Total count of pings in the stop

- Average speed
  odometer difference between first and last ping in the stop, divided by the duration of the stop

- Number of engine off pings in the stop

CRISP-DM	1	2	3	4	5	6

Data preparation

2) Points of interest features

"points of interest (POI) in the area surrounding each stop"

CRISP-DM	1	2	3	4	5	6

Data preparation

2) Points of interest features f2:POIF

points of interest (POI) in the area surrounding each stop

For each of these POI types,
build a feature that consists of the smallest distance

If such distance >= 200 meters, set ∞

CRISP-DM	1	2	3	4	5	6

Data preparation

2) Points of interest features f2:POIF

Extract the following POI types:

- bank
- university
- hotel
- restaurant
- rest area
- grocery store
- school
- shopping center
- fuel
- open parking area
- vehicle repair facility

Data preparation

3) Stop cluster features

"There are some areas where work orders & non work orders tend to cluster"

Data preparation

3) Stop cluster features

f3:CF

characteristics of the stops surrounding the stop

the rationale is that there are some areas where work orders and non work orders tend to cluster

for each stop, we look inside a radius of 250 meters

Data preparation

3) Stop cluster features f3:CF

vehicle entropy, computed as:

$$E_v = -\sum_{v \in V} \frac{n_v}{N} ln\left(\frac{n_v}{N}\right)$$

V = set of vehicles of the given fleet

N = total number stops of all the vehicles , radius of 250 m

nv = total number stops of vehicle v inside the same area

Data preparation

3) Stop cluster features f3:CF

- average, sum, max and min duration of the stops within the cluster

- number of nearby stops centroid is within 250 meters

CRISP-DM	1	2	3	4	5	6

Data preparation

4) Sequential features

"It is unlikely that a driver has two consecutive lunch stops"

CRISP-DM	1	2	3	4	5	6

Data preparation

4) Sequential features f4:SeqF

consider four neighboring stops:

the two previous ones, -2

and the two immediately following it +2

compute a set of features based on their stop-wise and POI features, such as stop duration, etc.

CRISP-DM	1	2	3	4	5	6

Data preparation

4) Sequential features f4:SeqF

In addition,

- time from/to the previous/following stop;
- distance from/to the previous/following stop.

To identify stops that are far from others, meaning that the vehicle had to take a long detour to reach that place

Modeling

CRISP-DM	1	2	3	4	5	6

Modeling

How to build a model ?

Train data vs Test data ?

source : www.usfleettracking.com/service

Modeling

The labeled dataset of 702446 examples
Randomly split into training and test sets

source : local-eyes.nl/solution/fleet-management

Modeling

TABLE I
TYPES OF COMPANIES IN THE DATASET.

Type	Count
Heating & air conditioning	45
Plumber & leak detection	18
Protection systems	9
Maintenance & cleaning	7
Electric	6
Irrigation & lighting	5
Delivery	4
Pools installation	2
Locksmith	1
Health assistance	1

To avoid bias of same fleet,
that are likely to have a similar behavior,

the split is stratified by customer,

The resulting training set is composed of 71 fleets,
While the remaining 27 compose the test set

Modeling

TABLE II
DATASET STRUCTURE.

Dataset split	Work order	Non-work order
Train	128071	396199
Test	33833	144343

CRISP-DM	1	2	3	4	5	6

Modeling

	Work	Non-Work
Train	25%	75%

TABLE II
DATASET STRUCTURE.

Dataset split	Work order	Non-work order
Train	128071	396199
Test	33833	144343

Train	75%
Test	25%

	Work	Non-Work
Test	20%	80%

CRISP-DM	1	2	3	4	5	6

Modeling

Random Forest classifier
consists of an ensemble of decision trees

Freepik , source : www.flaticon.com/free-icon/forest_186474

Modeling

Trained by means of a 10 fold cross-validation

Karl Rosaen , source : karlrosaen.com/ml/learning-log/2016-06-20

CRISP-DM	1	2	3	4	5	6

Modeling

Trained by means of a 10 fold cross-validation

to choose the best

number of trees T and

$$\boxed{T \in \{100, 200\}}$$

their max depth D

$$\boxed{D \in \{12, 15, 20, 22, 25\}}$$

Modeling

In the training phase, each group of features is progressively added.

Each time add a group of features,
train a new classifier with the optimal parameters found with a 10 fold cross-validation.

Evaluation

Evaluation

Baseline model ?

Result ?

source : www.usfleettracking.com/service

CRISP-DM	1	2	3	4	5	6

Evaluation

TABLE III
PERFORMANCE OVER THE TEST SET AND OPTIMAL PARAMETERS OF THE CLASSIFICATION MODELS.

Model	Optimal parameters		ROC AUC
`BLEntr` - Baseline w. Entropy	–		0.678
`BLDur` - Baseline w. Stop Duration	$T = 100$	$D = 12$	0.854
`SWF`	$T = 100$	$D = 12$	0.878
`SWF + POIF`	$T = 100$	$D = 15$	0.890
`SWF + POIF + CF`	$T = 200$	$D = 20$	0.928
`SWF + POIF + CF + SeqF`	$T = 100$	$D = 22$	**0.931**

CRISP-DM	1	2	3	4	5	6

Evaluation

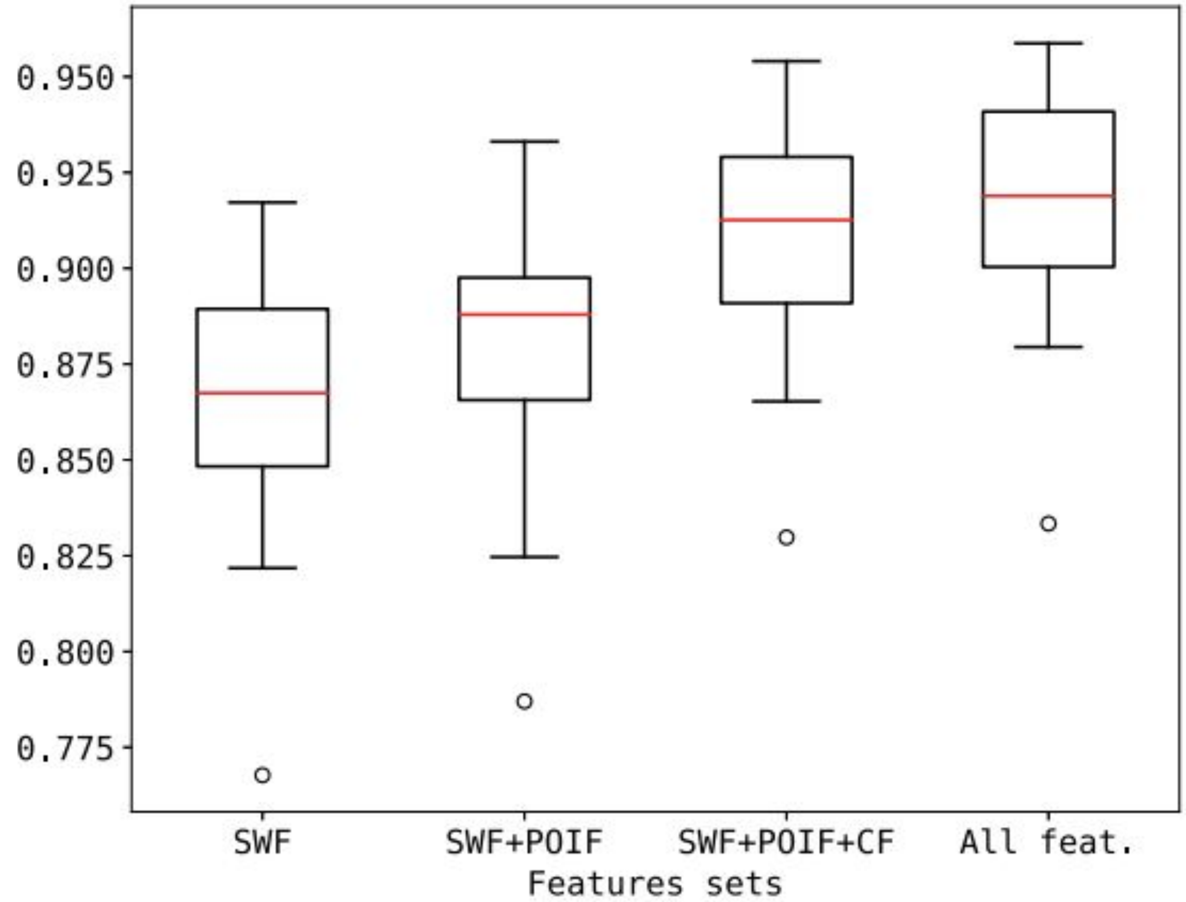


Fig. 2. Boxplots of the ROC AUC over the 10 folds of the cross-validation for each set of features. `All feat.` stands for: `SWF` + `POIF` +`CF` + `SeqF`.

Evaluation

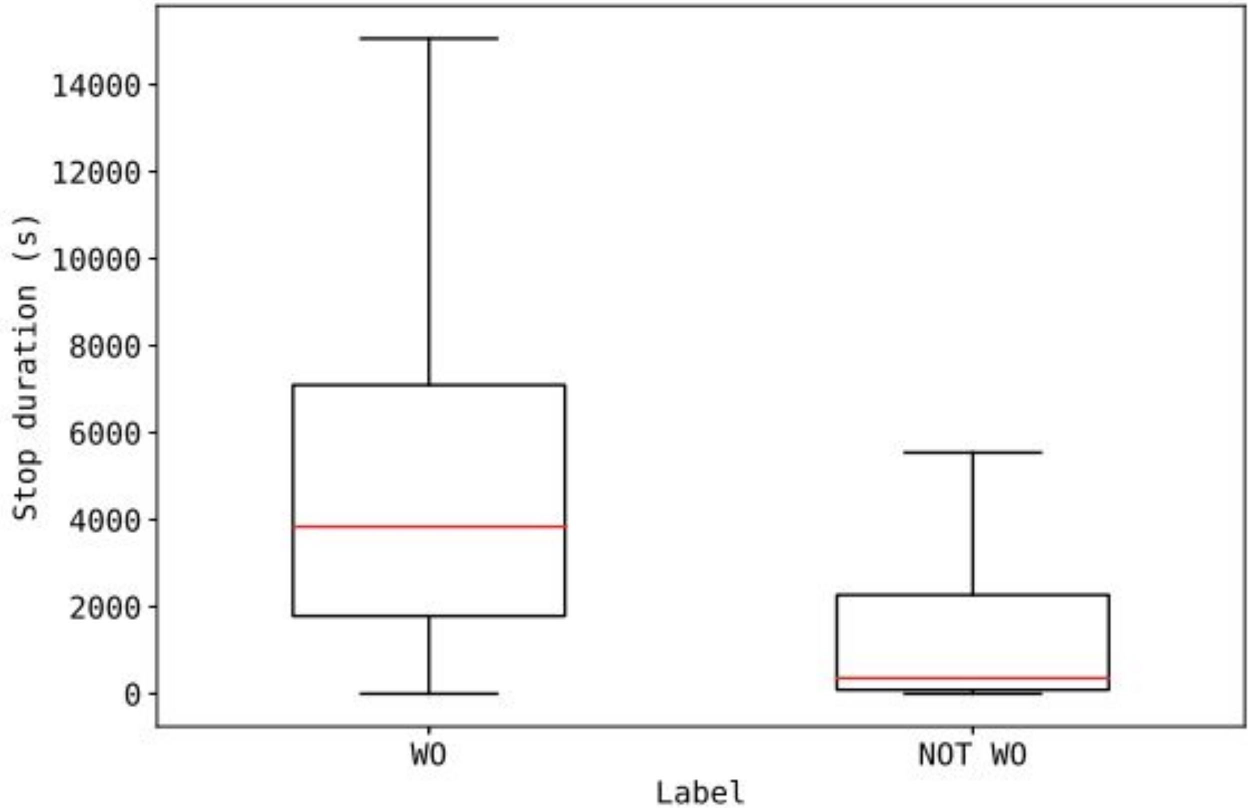


Fig. 1. Boxplots on the training set of the stop duration values for work order (WO) stops and not work order (NOT WO) stops.

Evaluation

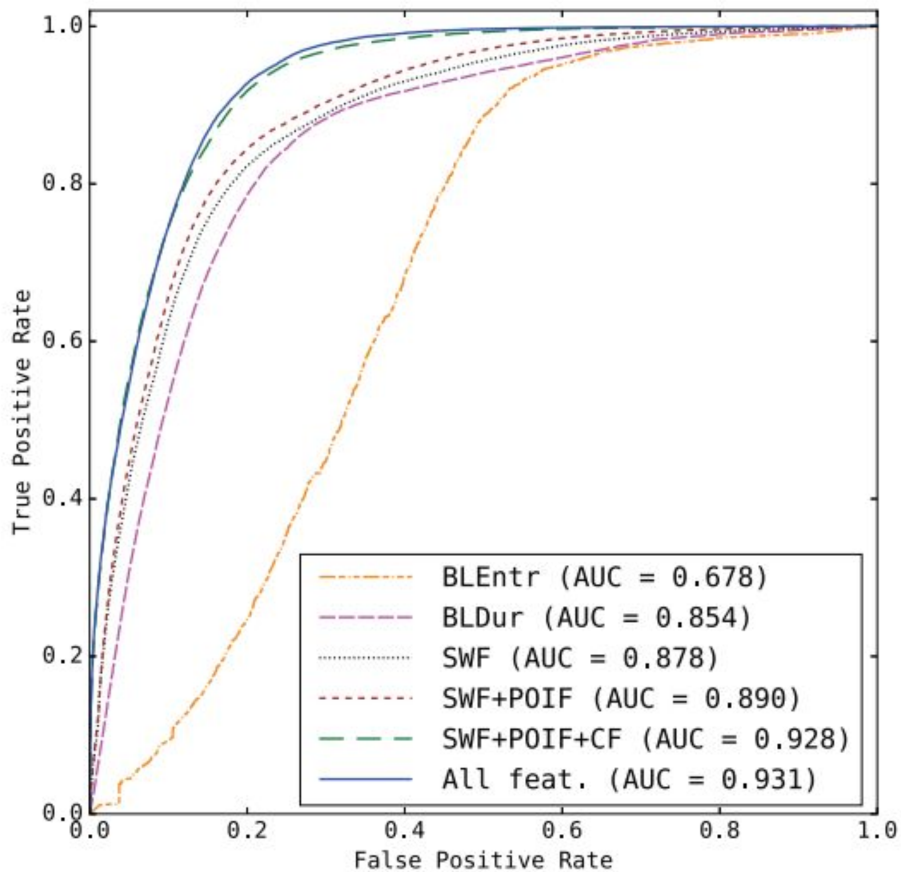


Fig. 3. Plots of the ROC curves of all the classifiers evaluated on the test set. `BLEntr` is the baseline model based on entropy, `BLDur` is the baseline model based on the stop duration. `SWF`, `POIF`, `CF` and `All feat.` refer to the groups of features used in the random forest model.

CRISP-DM	1	2	3	4	5	6

Evaluation

TABLE IV

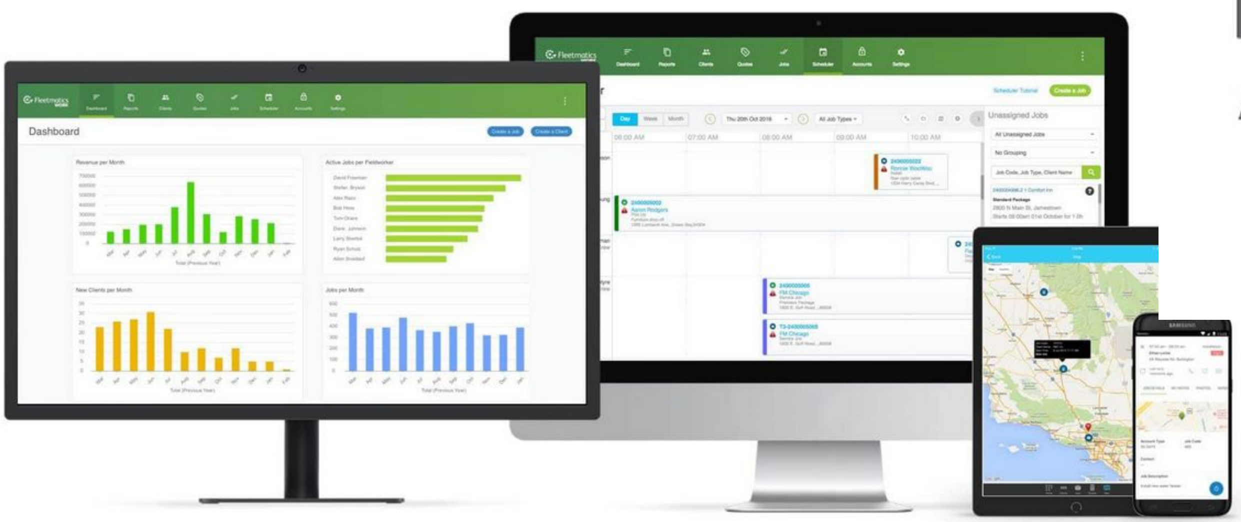
FEATURE RANKING: TOP 3 FEATURES OF EACH GROUP WITH ITS OVERALL RANK AND SCORE.

Overall rank	Name	Feature set	Score
1	Average stop duration in cluster	CF	0.121
5	Max stop duration in cluster	CF	0.048
9	Sum of stop durations in cluster	CF	0.035
2	Stop duration	SWF	0.079
3	Total time with engine off	SWF	0.060
4	Max time with engine off	SWF	0.056
8	Time to next stop	SeqF	0.039
12	Distance from previous stop	SeqF	0.021
13	Distance to next stop	SeqF	0.020
19	Distance to closest fuel station	POIF	0.013
24	Distance to closest restaurant	POIF	0.008
29	Distance to closest vehicle repair facility	POIF	0.006

Deployment

CRISP-DM	1	2	3	4	5	6

Deployment



Fleetmatics®

A Verizon Company

Thank You
&
Discussion