Machine Learning in Transportation

Eleni I. Vlahogianni, Ph.D.
National Technical University of Athens
(elenivl@central.ntua.gr)

What we discuss today

Data Science and Machine Learning in Transportation

Sensing and the new opportunities

Outlier detection methods for driving analytics

An Example in python

Data Science in Transportation

The importance of data science from the view of a transportation engineer

- Information from user generated data
- Analysis and forecasting for improving traffic operations and management

"Data science is the civil engineering of data."

Data scientists possess a practical knowledge of tools and methods, coupled with a theoretical understanding of what's possible.

Data Science in Transportation

Data science already a part of transportation science for more 40 years

- We have always made decisions based on quantified information

- Data driven modeling has emerged very early in many transportation research fields (e.g. traffic and travel demand analysis and forecasting)

Data Science in Transportation

Traditionally, turning data into knowledge

Manual analysis and interpretation

⬇

Analyst becomes intimately familiar with data

Highly inefficient & subjective

Good when all is ‘in order’

Data Science in Transportation

…from the modeling side

1. Database size and high dimensionality
2. Overfitting and assessing statistical fit
3. Rapidly changing and imperfect data
4. Complexities and interactions

"I think you should be more explicit here in step two."

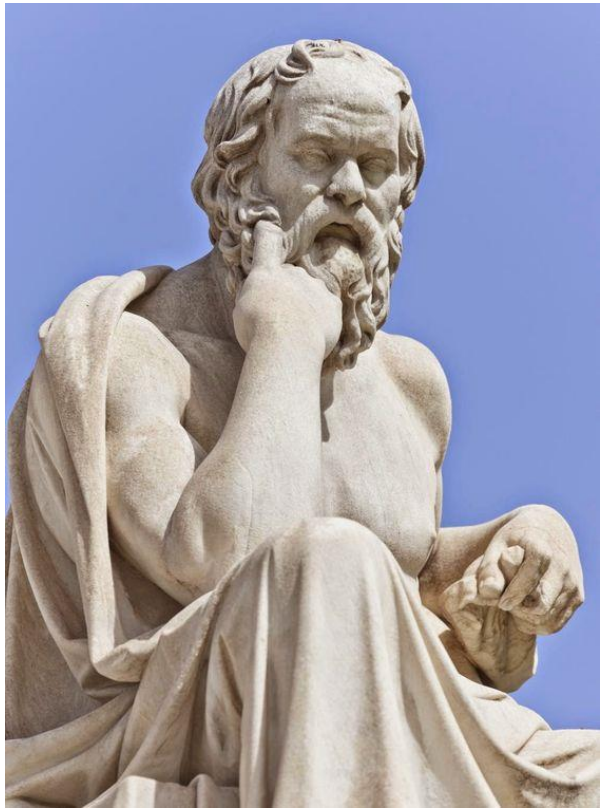
Data Science in Transportation

But…

We cannot wait for data to become ‘well behaved’ before we analyze

We need new modeling paradigms that are

- Robust to data imperfections
- Hypotheses free
- Flexible and powerful



What changed?

Data Accumulation

- Various sources of inferring transportation demand, system changes and their effect to flows (Many Vs…)

The availability of high-end and sometimes low cost technologies to collect data

- Smartphones and crowdsourcing
- The new 3d monitoring environment (the 3rd dimension of drones!)

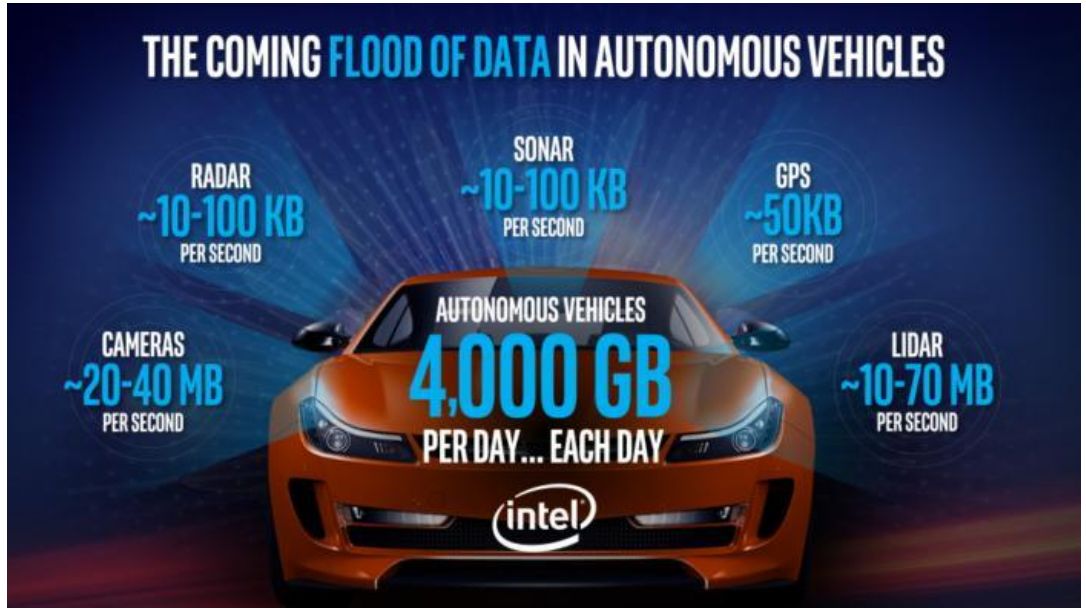
The variety of tools to store, process, visualize and model

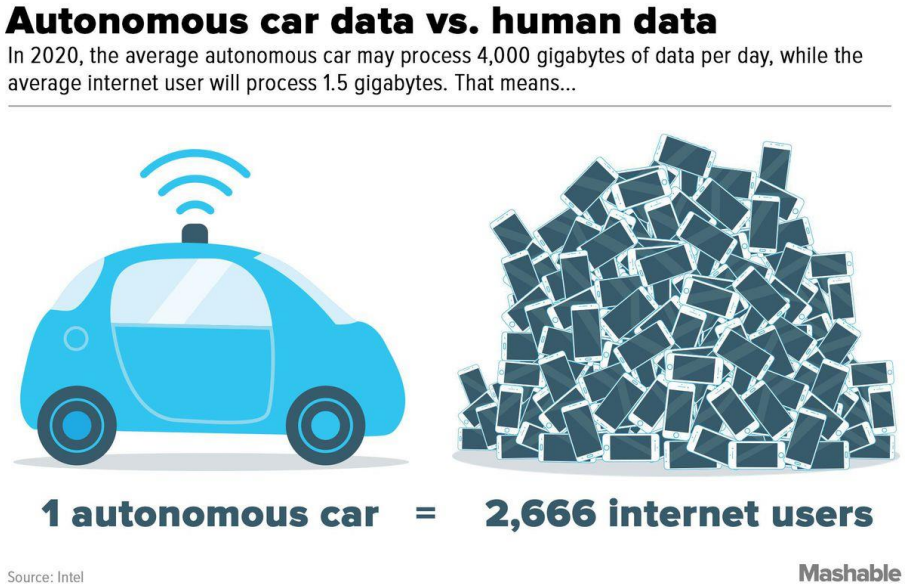
- Support from industry on open source solutions
- Some open source data initiatives
- Many integrated services available

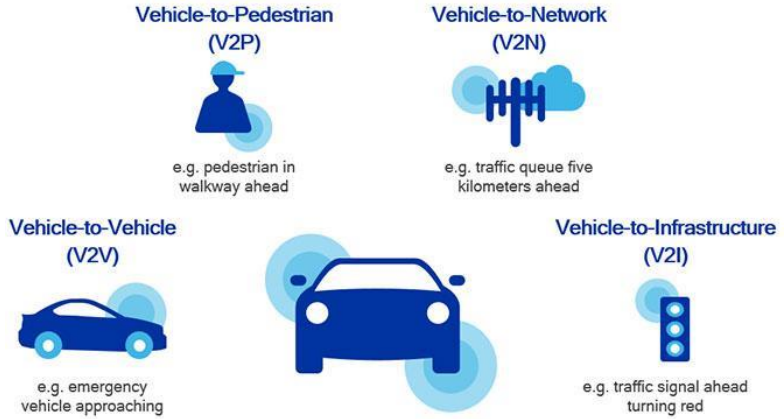
Machine Learning !

The size of some problems

The case of transportation







Uniquely Positioned at the Heart of the Automotive Data Ecosystem

SMART INSURANCE

Usage Based Insurance allow low-risk policy holders to get rewarded for driving safely

ENERGY & UTILITIES

Reliable charging infrastructure to support the Future Urban Mobility

SMART CITIES

Improving mobility, efficiency, & safety with reliable V2X communication

SERVICE & DEALERSHIPS

Customized vehicle maintenance packages for owners based on the actual condition of the car

LOCATION INTELLIGENCE

Location based personalised in-car advertising and offers

SMART PARKING

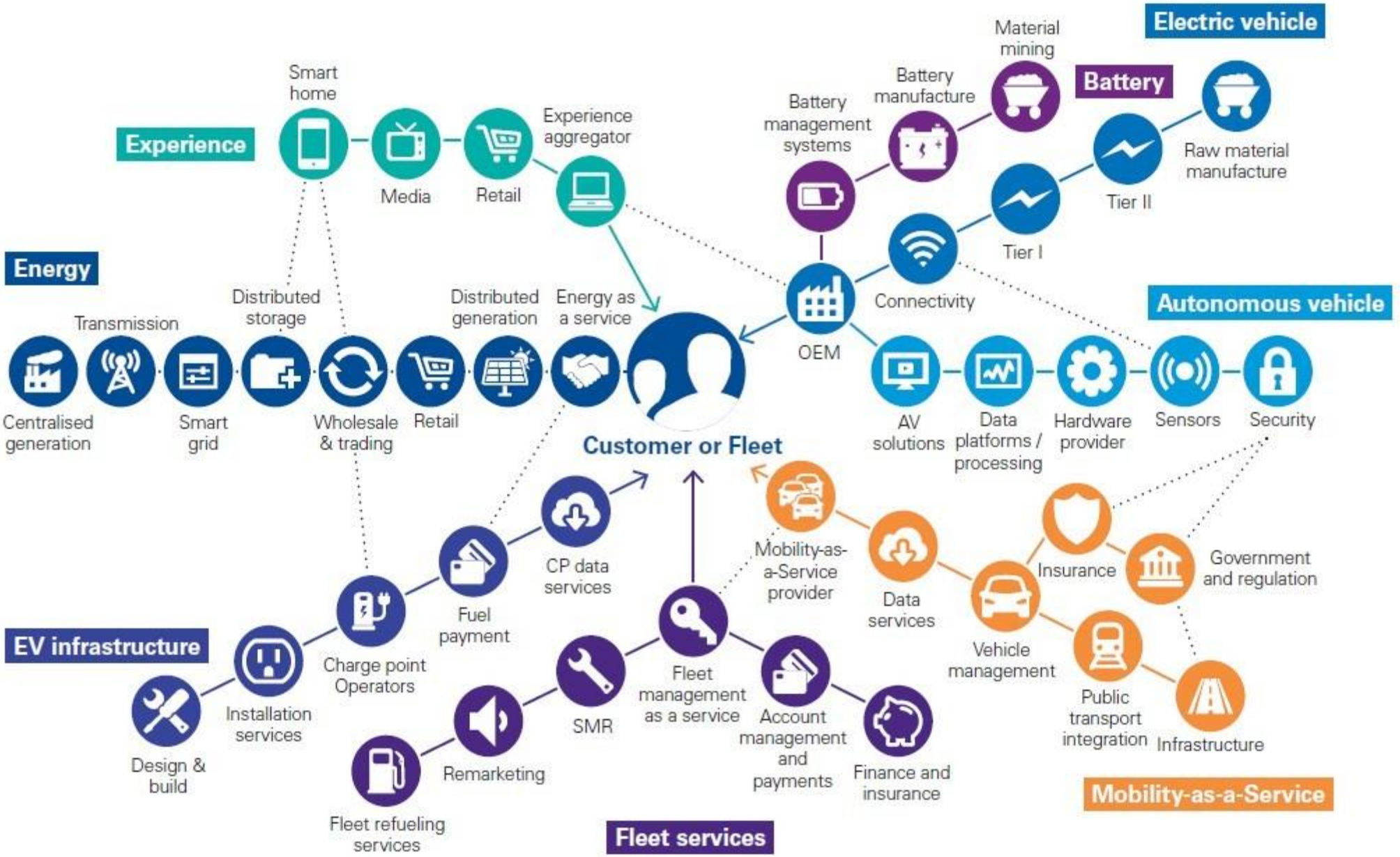
Improving the parking experience with dynamic real-time parking information

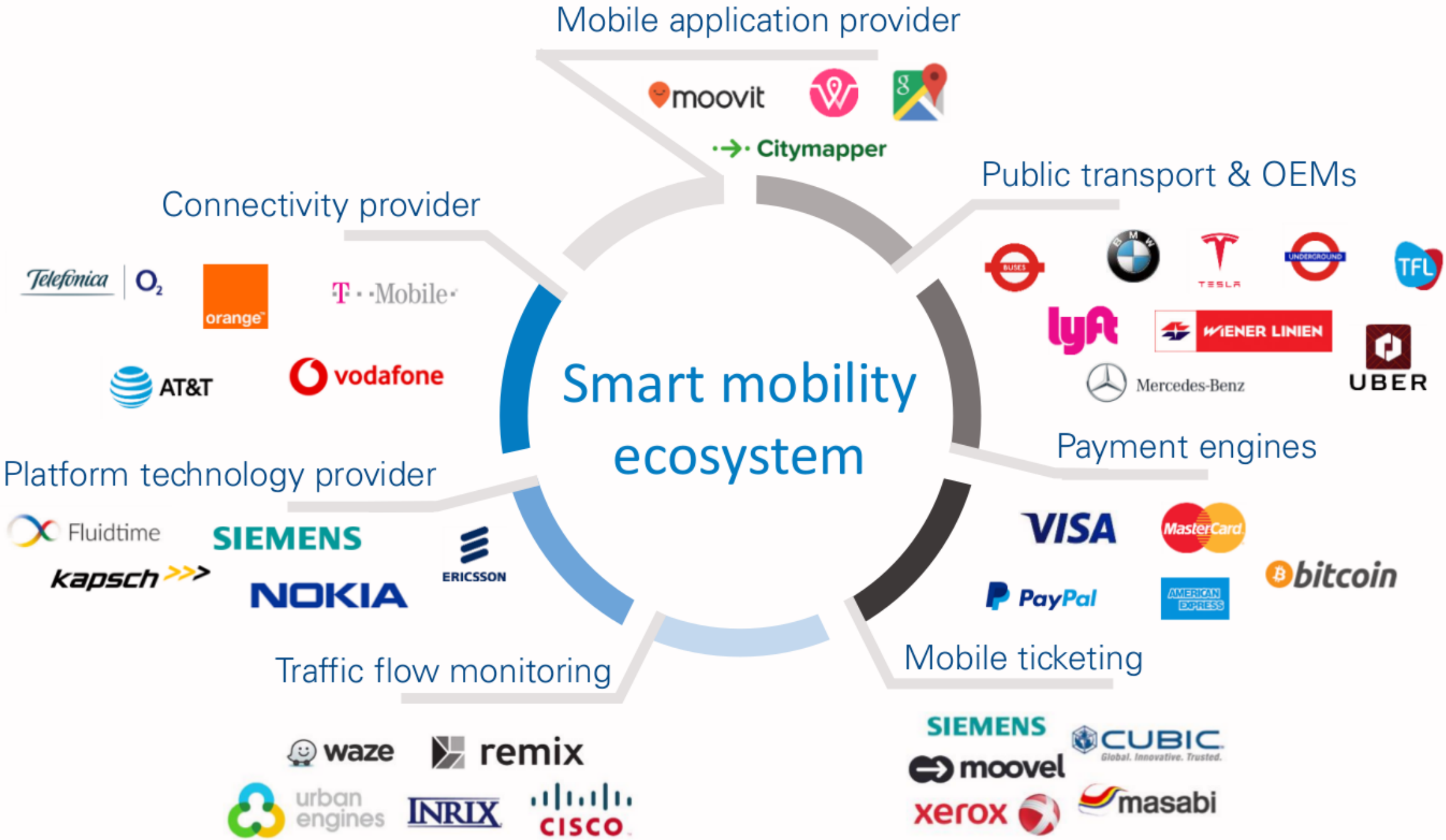
MAPPING & PLANNING

Generating precise mapping and location-centered insights from connected car data

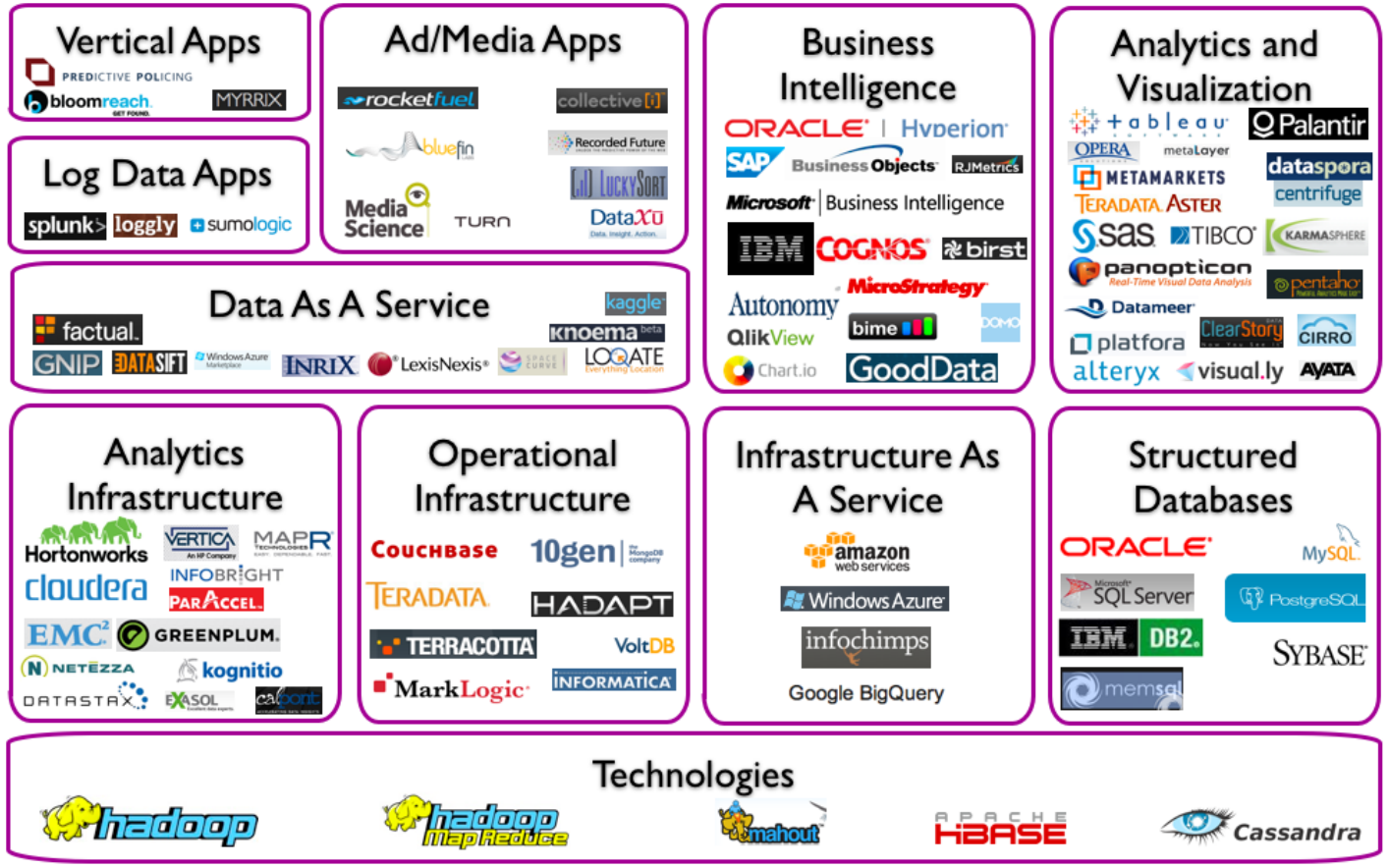
FLEET MANAGEMENT

Fleet data to fine-tune operations and generate new revenue sources


Electric vehicle
Material mining
Battery manufacture
Battery management systems
Battery
Raw material manufacture
Tier II
Tier I
Experience
Smart home
Media
Retail
Experience aggregator
Energy
Distributed storage
Distributed generation
Energy as a service
Transmission
Connectivity
Autonomous vehicle
OEM
Centralised generation
Smart grid
Wholesale & trading
Retail
AV solutions
Data platforms / processing
Hardware provider
Sensors
Security
Customer or Fleet
Mobility-as-a-Service provider
Data services
Insurance
Government and regulation
CP data services
Fuel payment
EV infrastructure
Charge point Operators
Installation services
Design & build
Vehicle management
Public transport integration
Infrastructure
Fleet management as a service
SMR
Remarketing
Fleet refueling services
Account management and payments
Finance and insurance
Fleet services
Mobility-as-a-Service


Mobile application provider
moovit
Citymapper
Public transport & OEMs
TESLA
UNDERGROUND
TFL
WIENER LINIEN
Mercedes-Benz
UBER
Payment engines
VISA
MasterCard
bitcoin
PayPal
AMERICAN EXPRESS
Mobile ticketing
SIEMENS
CUBIC
Global. Innovative. Trusted.
moovel
xerox
masabi
Traffic flow monitoring
waze
remix
urban engines
INRIX
cisco
Platform technology provider
Fluidtime
SIEMENS
kapsch
NOKIA
ERICSSON
Connectivity provider
Telefónica
O2
orange
T-Mobile
AT&T
vodafone
Smart mobility ecosystem

The big data landscape



dave@vcdave.com

blogs.forbes.com/davefeinleib

Machine Learning

Machine learning is

- The ability of a system to gain knowledge from interacting with the environment
- The ability to improve through repeating the execution of an action

Systems with learning abilities:

- Are constantly improving
- Can generalize (they can disregard characteristics and patterns that are not representative of the concept/activity they have learnt)

Machine Learning

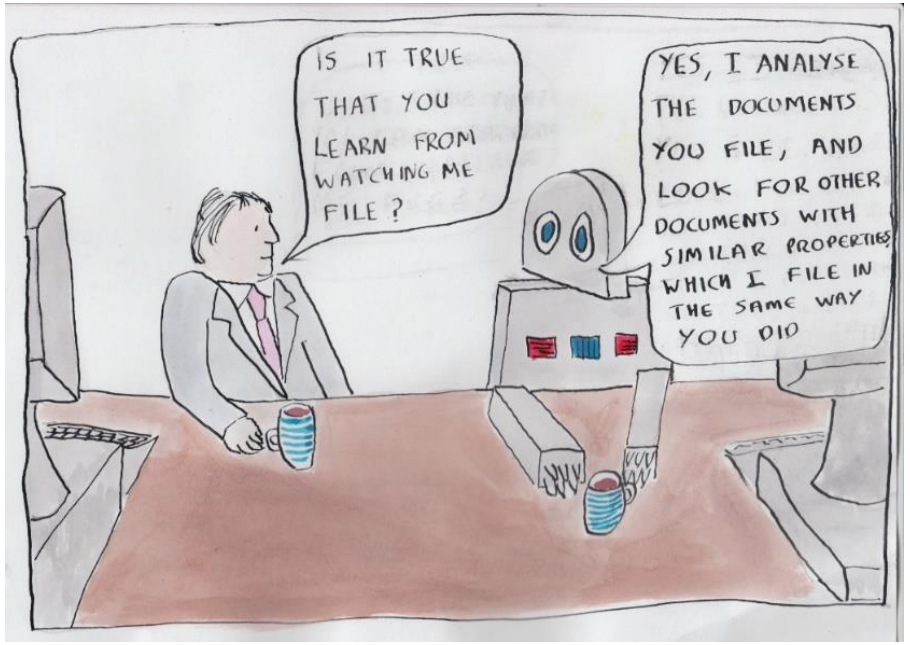
As a search problem

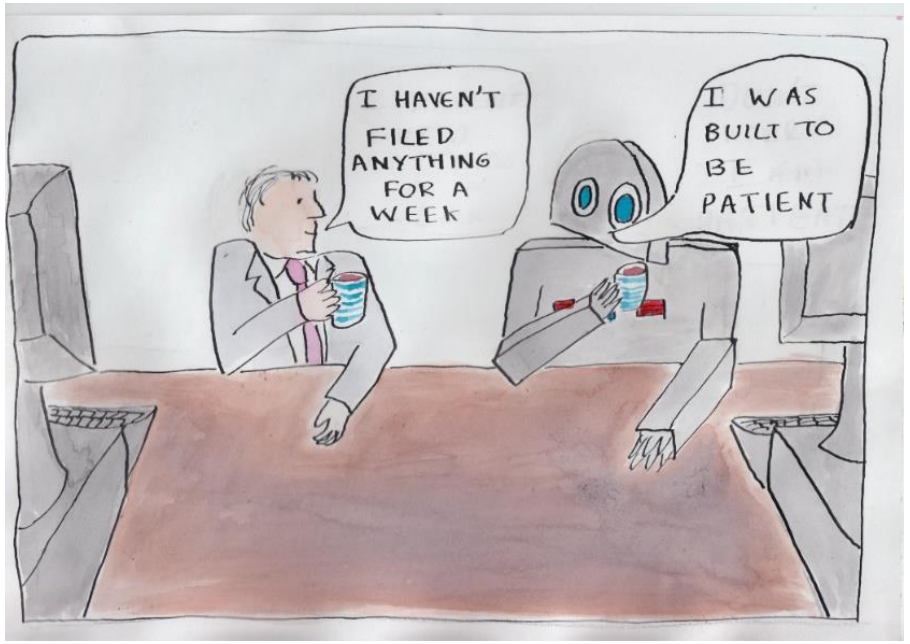
"the search in a space of possible hypotheses, of that hypothesis that best suits the data under consideration and the possibly pre-existing knowledge".

Basic Principle

The resulting knowledge is supported by the examples, but this does not mean that it is necessarily true in the real world.

Beware of "garbage-in garbage-out"





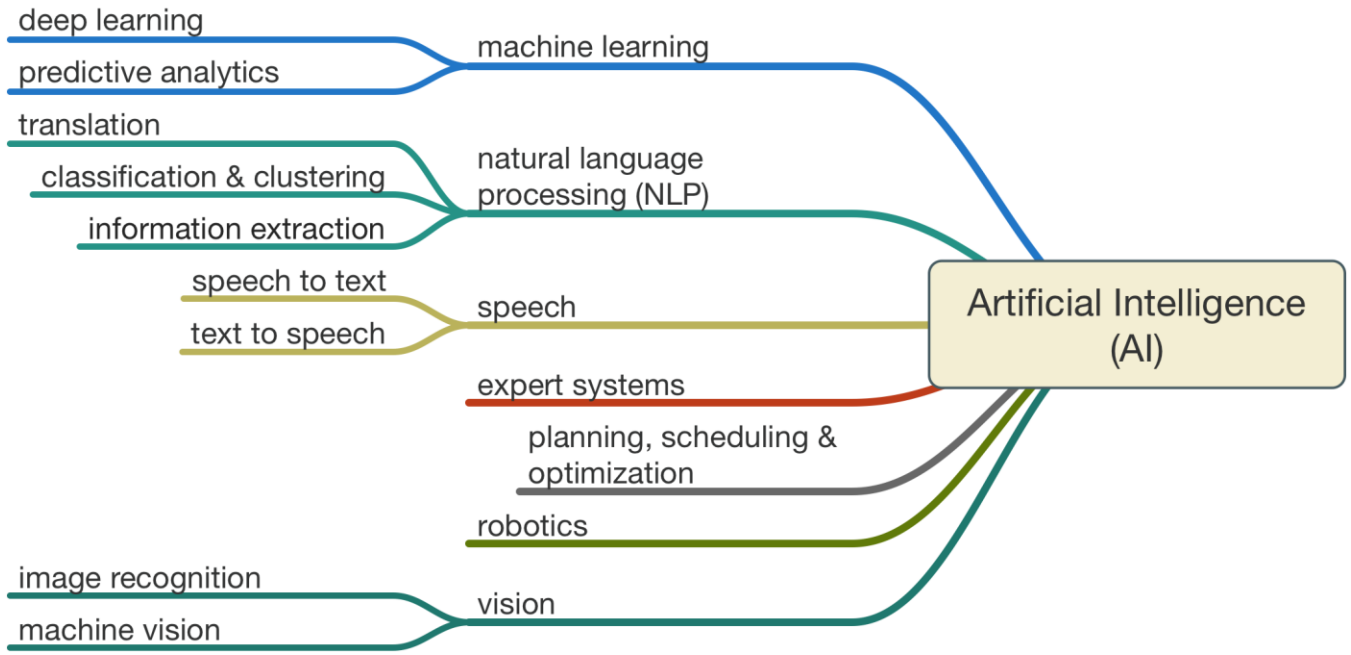
Machine Learning

vs artificial intelligence

- Intelligence: the ability to collect, process and apply knowledge
- Experience: results from exposure to real world conditions (training).
- Knowledge: the information we receive with the experience

Artificial intelligence is the “simulation" of a physical entity, which has the ability to collect, process and apply knowledge through experience

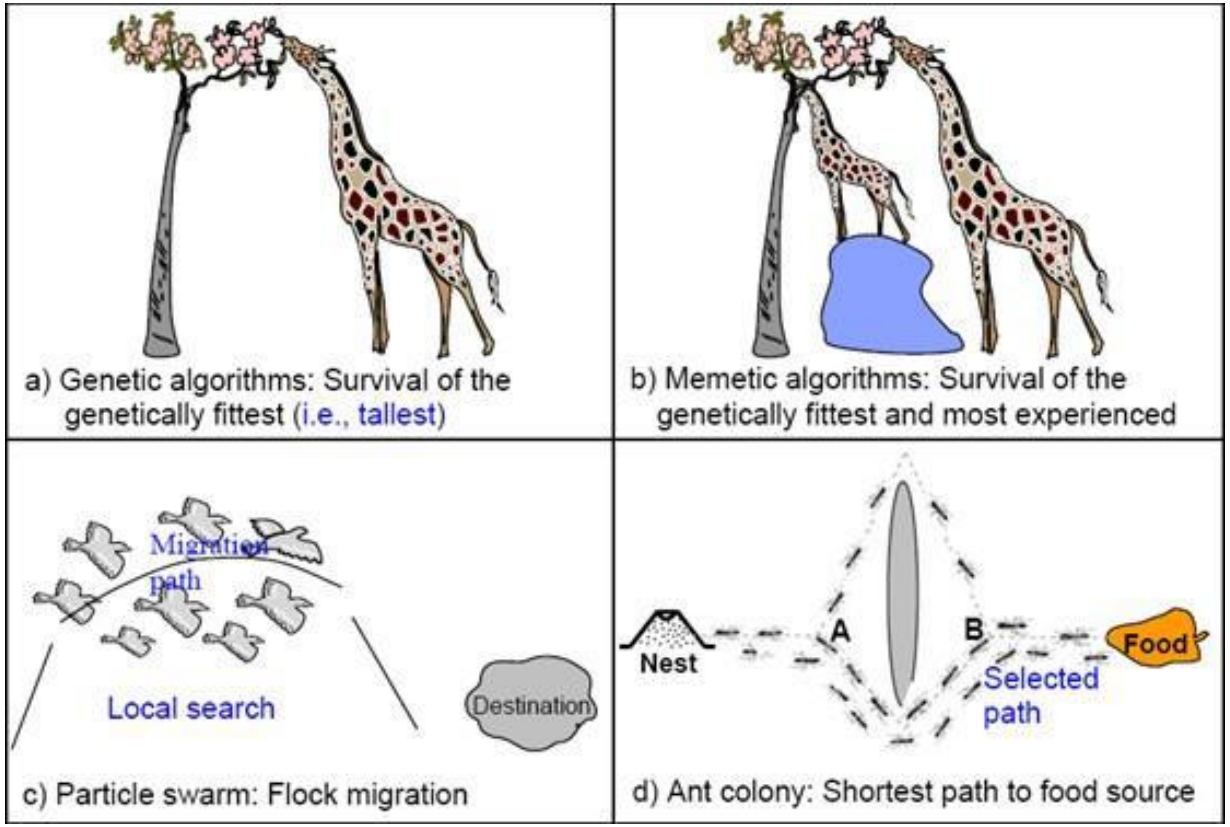
- Reasoning
- Learning
- Problem Solving
- Perception
- Linguistic Intelligence



Machine Learning

vs Computational Intelligence

- The ability to learn a task from real data with algorithms and patterns that mimic nature (Nature-inspired algorithms)



Machine Learning Algorithms

Supervised learning

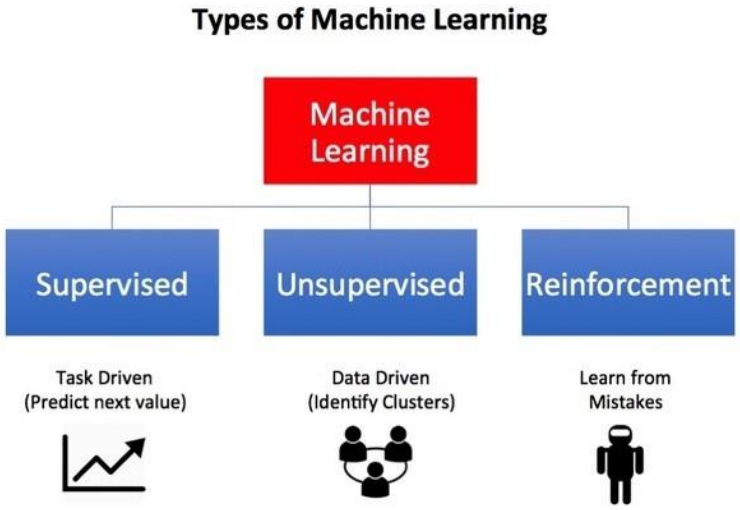
- The system is asked to "learn" a concept or function from a dataset, which is a description of a model.

Unsupervised learning

- The system must itself discover associations or groups in a dataset, creating patterns, without knowing if there are, how many, and what they are.

Reinforcement Learning

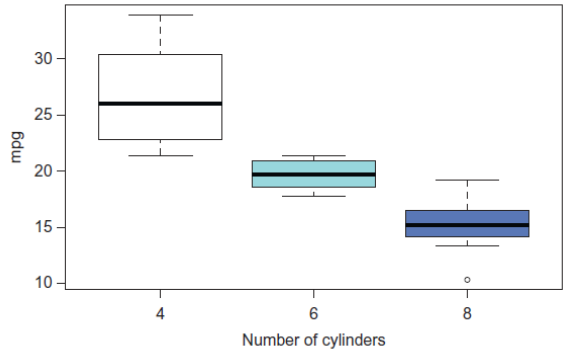
- a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback (in the form of a reward) from its own actions and experiences.



FUNCTIONAL FACETS

Descriptive analytics

- Understand current state of system
- Distributional characteristics, measures of central tendency
- Outlying behavior
- Exploratory data analysis (feature creation)
- Visualization tools



FUNCTIONAL FACETS

Predictive analytics (what might happen?)

- Can we predict future outcomes? Future state of the system?

- Forecast the probability of various events, the evolution of a certain measure etc

  Different methodologies in relation to the physics of the variable being predicted

  Time series approaches

  Pattern recognition approaches

  Function approximation approaches

- Developing relations between variables: (linear) correlation coefficients, mutual information etc

- Feature engineering and selection are critical

FUNCTIONAL FACETS

Prescriptive analytics (what should we do?)

- We need to provide intelligent recommendations

- How to ensure only a chosen or preferred outcome?

  Modeling and evaluating various what-if scenarios through simulation techniques to answer what should be done to maximize the occurrence of good outcomes while preventing the occurrence of potentially bad outcomes.

  Stochastic optimization techniques are used to determine how to achieve better outcomes, among others.

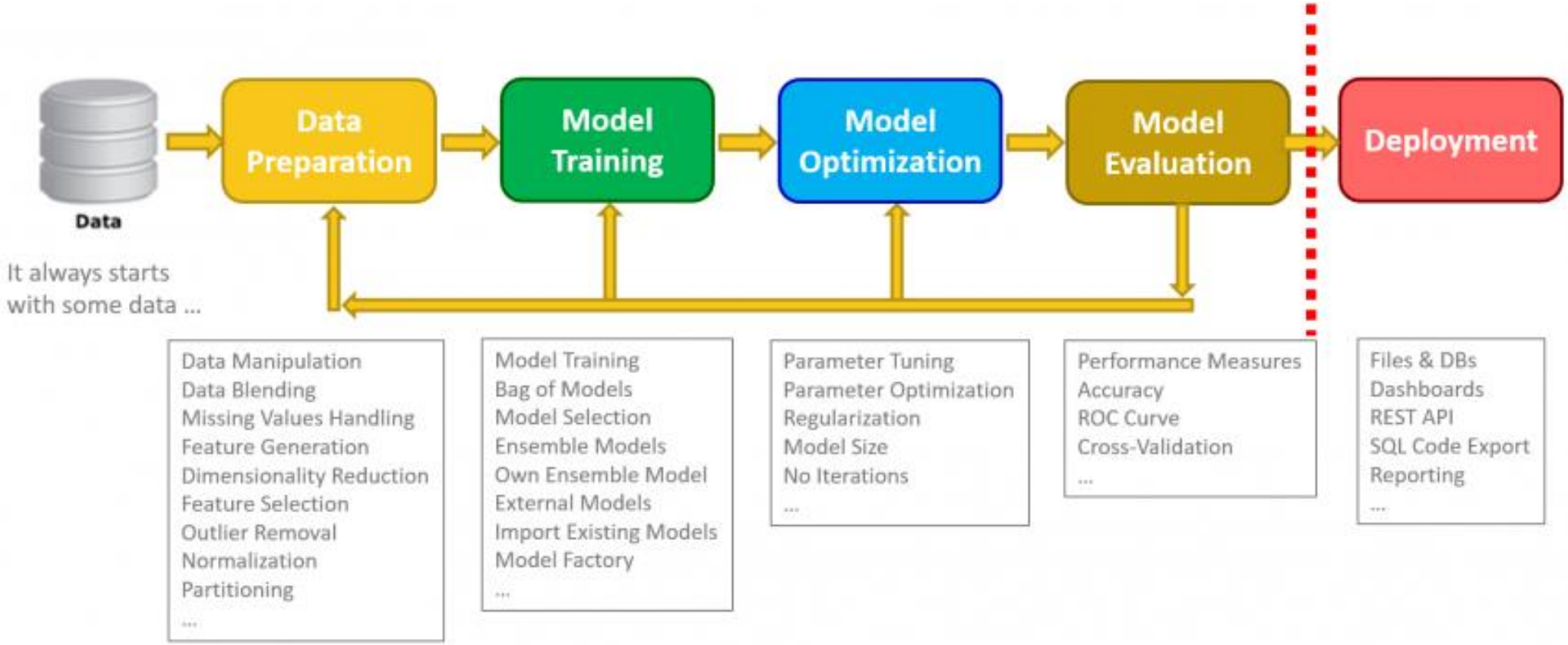
- Prescriptive analytics draws upon descriptive, diagnostic, and predictive analytics.

  Climate Change, Extreme events mitigation and adaptation

  Alleviating traffic congestion in cities

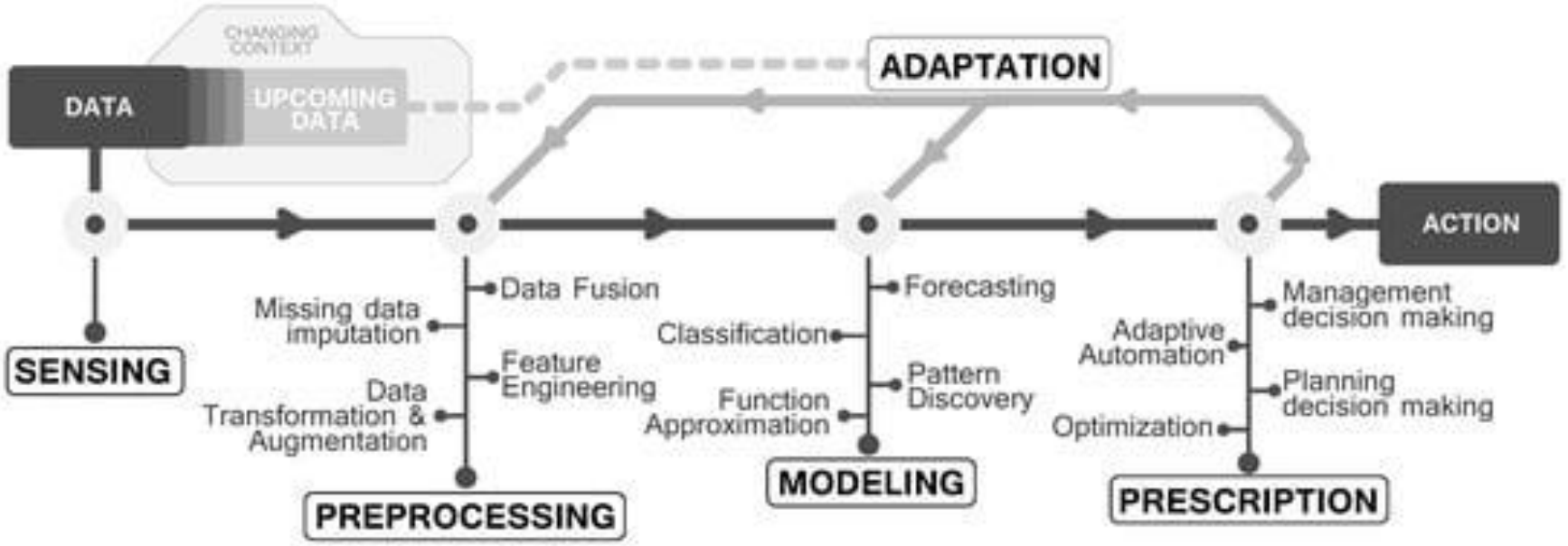
  Optimise a transport service to better serve the demand as well as spatial coverage

What is my action plan when dealing with ML?



From Data to Actions: An Actionable Data-Based Modeling Workflow

Data-based modeling workflow:



Smartphone sensing

The new opportunities

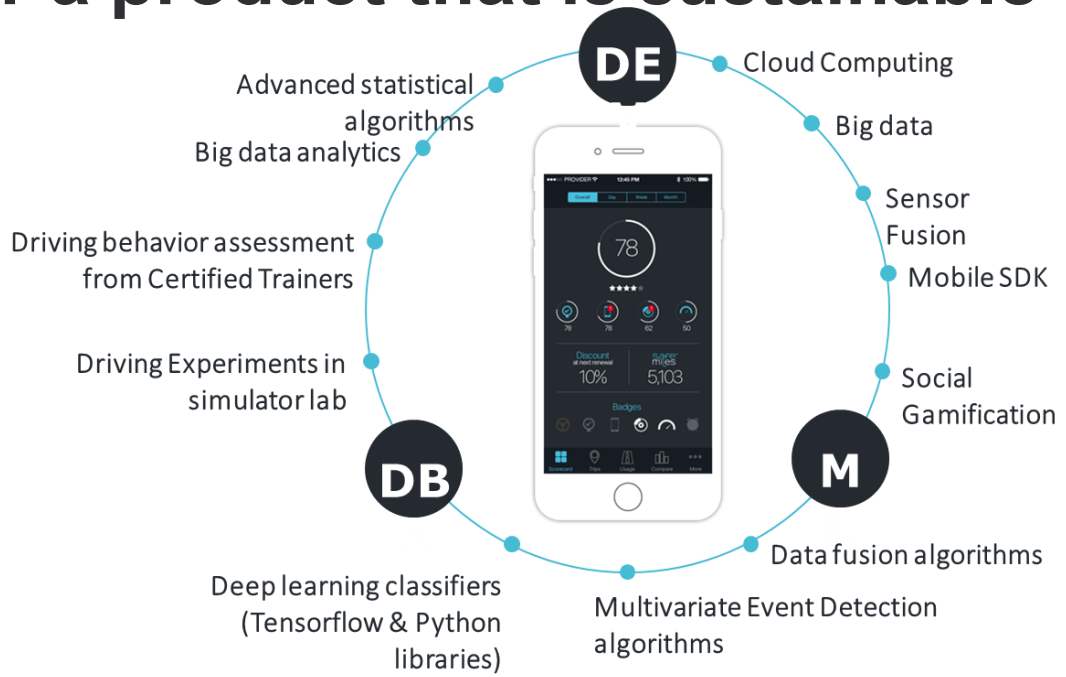
Smartphone based driving and mobility analytics

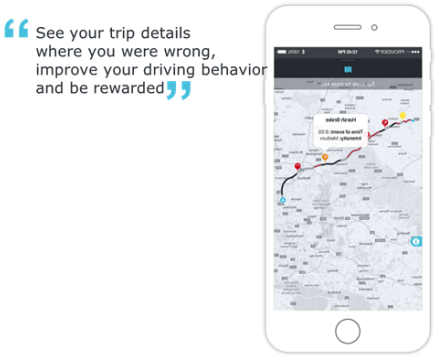
Usefulness to

- Customized pricing policies (Pay-As-You-Drive, Pay-How-You-Drive)
- Raising the awareness of the “efficient” mobility profile
- Identifying the actual human behavior

The constraint: Deliver a product that is sustainable

- Battery consumption
- Accuracy and scalability
- Address industry questions
- Useful and interesting to users





Smartphone based driving and mobility analytics

Detection problems

- Is it a car trip?
- Who drives the car?
- Is it a harsh event?
- Are you distracted?

Harsh events detection

Driver distraction

Clear "Noise" / in car activity

Driver ID recognition

Driver / Passenger / Mass Transit recognition

<u>Modeling and Policy</u>

- How long should I monitor you to know your overall driving behavior?
- How can I develop customized driving policies?
- Do specific user profiles exist?
- Can I use driving analytics for large scale network management?

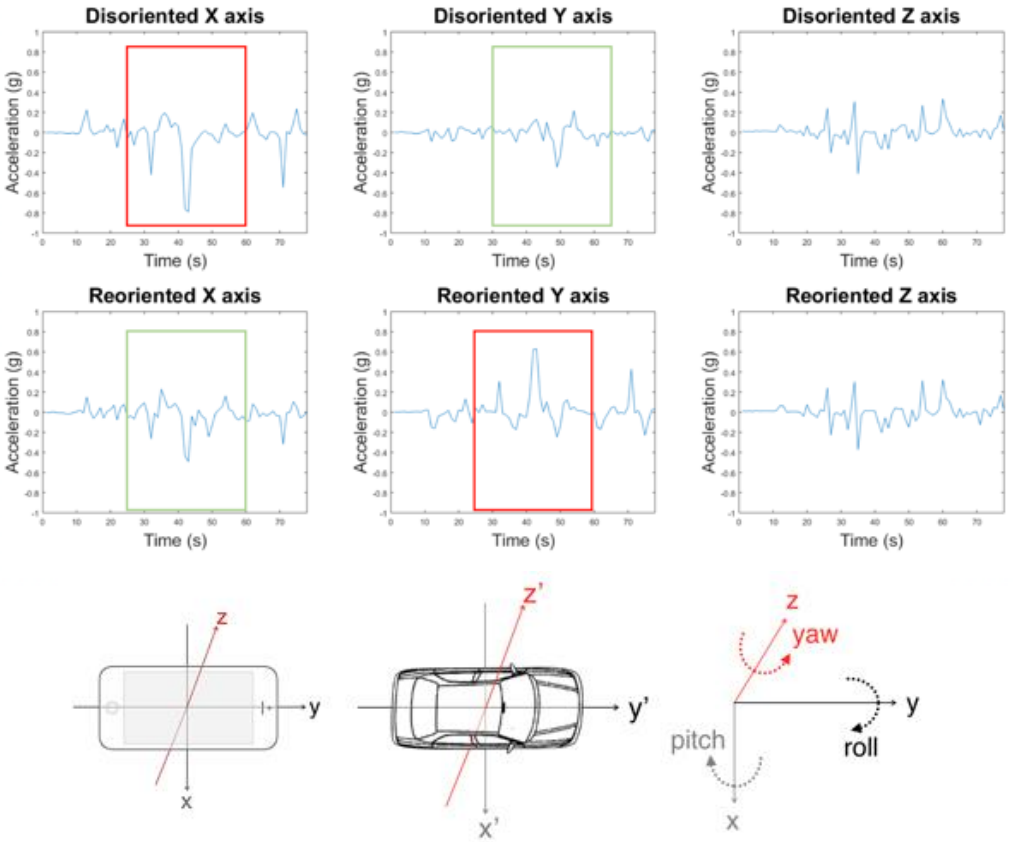
Data Specs and Preparation

Data resolution

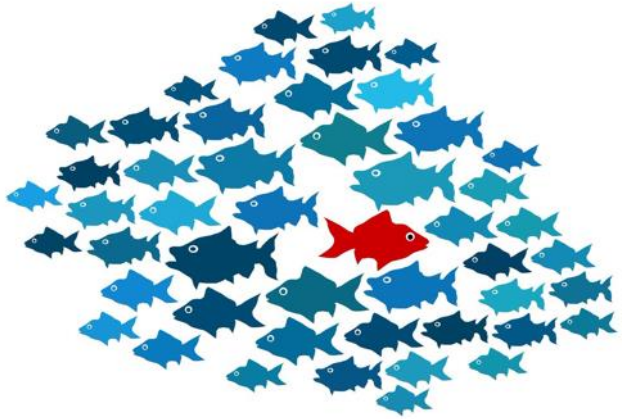
- strong dependence on the type of application

Data preparation

- Device Orientation
- Activity (Walking, Standing etc.)
- Erroneous Values
- etc



Detecting Harsh Driving Events

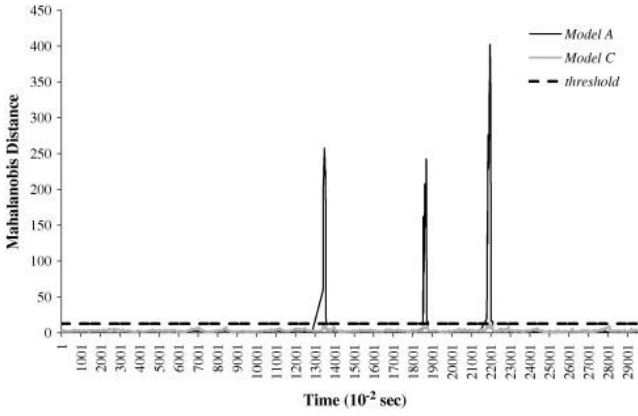


Uncertainties

- The smartphone's position is highly uncertain (on the deck, on the pocket, inside bags etc)
- Different types of smartphones (iOS, android, old, new sensors etc)
- Noisy signals

Harsh Accelerations Harsh Brakes

Harsh Cornerings

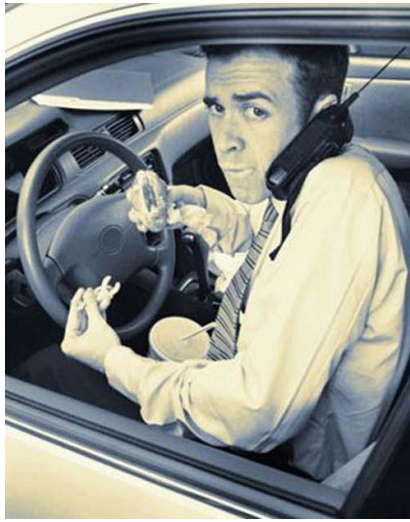


Per Trip Solution → Outlier Detection Strategy

- We define the basic driving pattern and quantify the divergence from it
- We fuse data from other sensors to evaluate the detection (e.g. use GPS data to evaluate acceleration and deceleration patterns)
- The process results to a slightly varying set to thresholds for every trip

Driver's Distraction Detection

Micro movements of the smartphone that are not due to the driving task



Privacy is of great importance.

- no microphone, call logs, or any other application can be used

Assumption: Drivers will generally pick up/leave their phone from/to a stationary position

Solution

- Sensor fusion gyroscope, GPS, accelerometer etc
  If gyroscope is not available, detection is based only on accelerometer (very challenging task)
- We exclude behaviors that look like picking up the phone, but are not (e.g. use GPS data to distinct harsh turns from sudden smartphone movements)

Driver Footprint Detection

Users should not be rated based on their passenger trips, but only on their driving behavior.

What differentiates drivers from passengers?

- Door exited
- Driving footprint (user agnostic approach)
- Users' usual trip information (user specific approach)

Driver Footprint Detection

User agnostic approach

- Supervised classification problem : GBM family of models
  - *Trips annotated by users. Annotation procedure susceptible to errors*
  - *Growing datasets*
- Feature extraction
  - *Meaningful features from sensor data fusion*
  - *Assess the behavior inside and outside the mobile usage periods*

User specific approach

- O-D information
- probability of being a driver

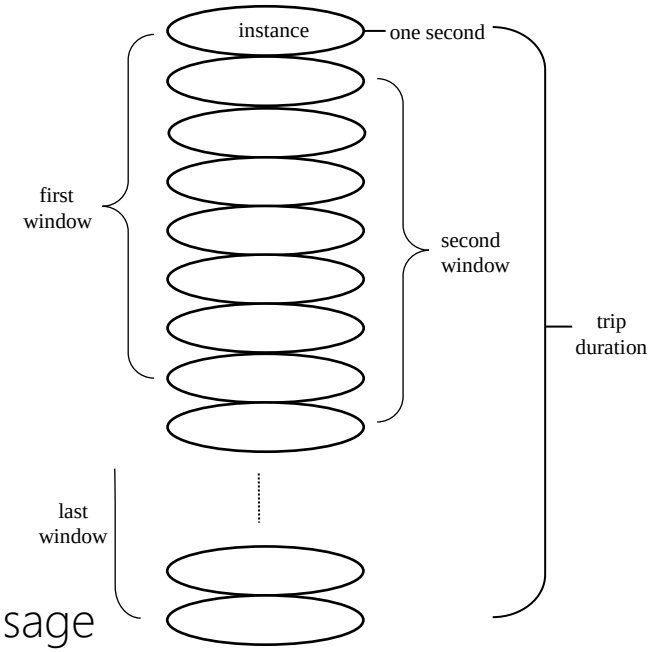
Are you driving?

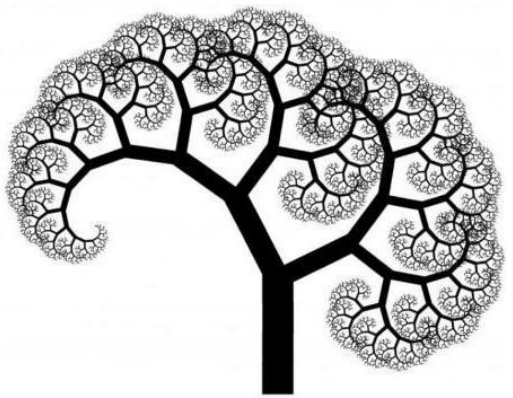
Mode Detection

A classification problem based on time series approach

Variable	Meta-variables	Name
Acceleration	Minimum	accMIN
	Maximum	accMAX
	Standard Deviation	accSD
	Median	accMED
	Mean	accMEAN
Pitch	Minimum	pitMIN
	Maximum	pitMAX
	Standard Deviation	pitSD
	Median	pitMED
	Mean	pitMEAN
Roll	Minimum	rolMIN
	Maximum	rolMAX
	Standard Deviation	rolSD
	Median	rolMED
	Mean	rolMEAN
Angular Velocity	Minimum	gyrMIN
	Maximum	gyrMAX
	Standard Deviation	gyrSD
	Median	gyrMED
	Mean	gyrMEAN
Speed	Smoothened speed based on moving average filter	speed
Driver	Every user called by a unique number	driver

- Entire trip information
- Entire trip excluding driving time with mobile usage
- Entire trip excluding driving time with stops
- Entire trip excluding driving time with both mobile usage and stops
- User specific models
- Models: GBM and some relevant alternatives

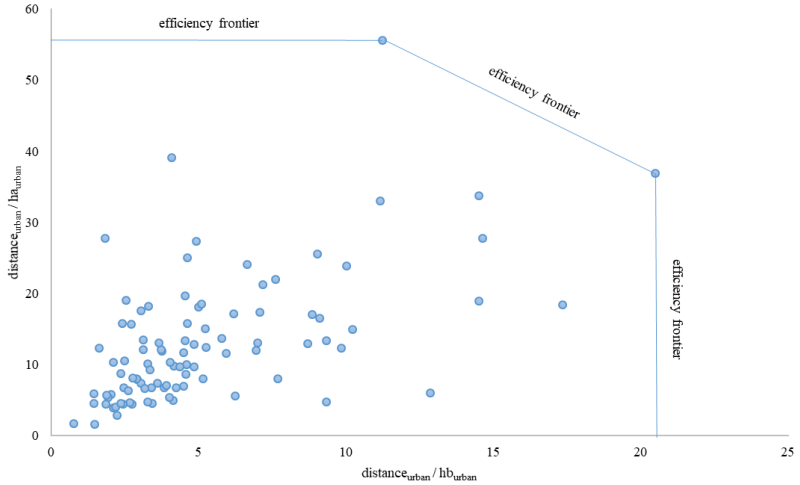


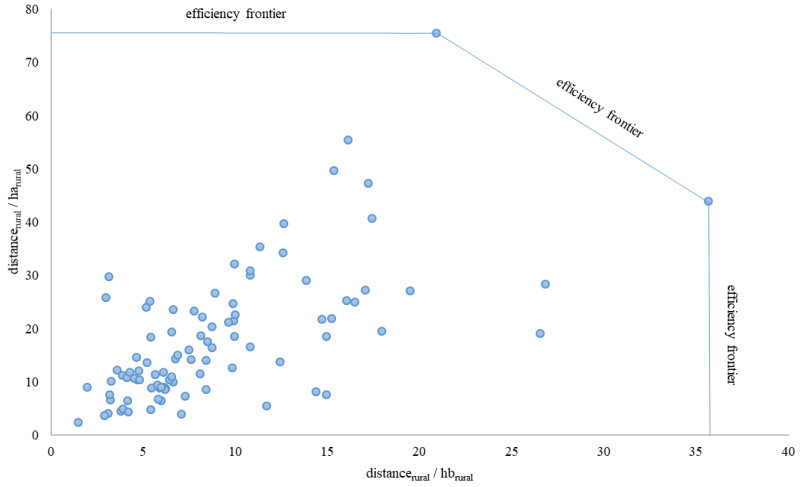


Policy Questions:
How do I rank my sample of drivers based on their performance?

Input-oriented Data Envelopment Analysis based on the driving metrics

- the objective is to minimize the level of driving metrics recorded per driving distance unit
- the driving efficiency problem is considered a constant-returns-to-scale (CRS) problem and the sum of all metrics (inputs) recorded such as the number of harsh acceleration and braking events occurred in each trip changes proportionally to the sum of driving distance (output)



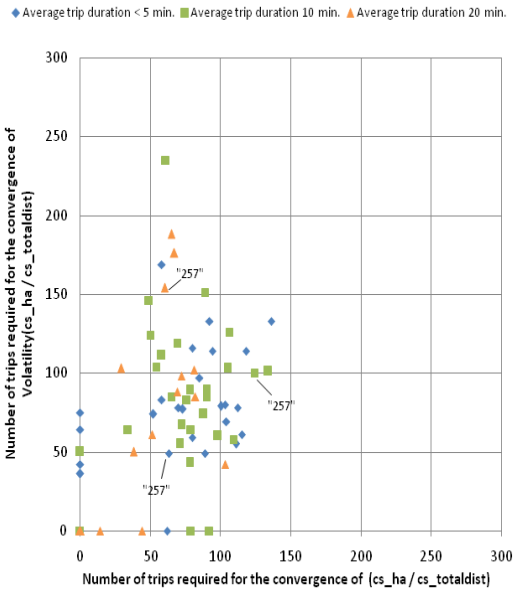


Policy Questions: How much data do we need?

Work with cumulative behavioral metrics (e.g. HA/100km)

Establish driving behavior convergence criteria:

- The moving average of a metric is within the range mean ± 1 * standard deviation.
- For five consecutive trips the percent change (in absolute terms) between successive values of the moving average is less than or equal to 1.5%.
- Examine if a driving variable remains stable over time and within two given upper and lower limits



Answer:

- depends largely on the aggressiveness and volatility of the overall driver's behavior as well as the average duration of the trips being studied.
- Aggressive drivers require less monitoring than cautious drivers do.

Policy Questions:
How can I develop customized policies to specific users?

Reinforcement Learning

- Driving state (poor, good etc)

  An optimum trip is considered one in which all critical driving variables have received zero values.

  The more a driver's trip diverges from optimum driving the more they degrade through states.

- Driving actions

$$(\text{Action}^{k}_{Ha}\,\text{Action}^{k}_{Hb}\,\text{Action}^{k}_{Sp}\,\text{Action}^{k}_{Mu})$$

  Actions are defined based on the variability of each driving metrics

- Reward may be set according to the strategy of the operator.

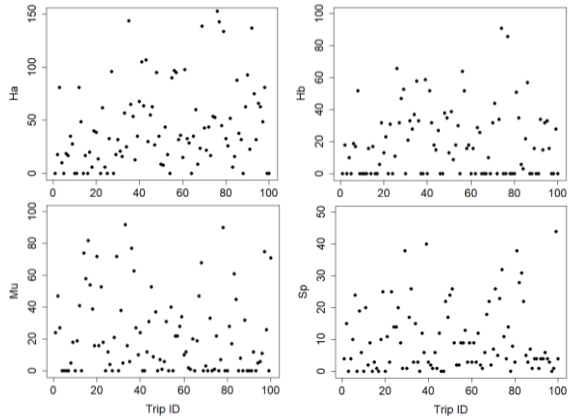


FIGURE 1 ***Representation of the critical driving variables for 100 consecutive trips of one driv from the dataset.***

Driving Profiles and Drivers' Profiles

Research questions:

- Is it possible to produce a meaningful characterization for every trip in relation to the driving behavior?
- Is it possible to extend trip characterization to driver characterization and produce a consistent manner to distinguish different drivers on the road?

Aggressiveness

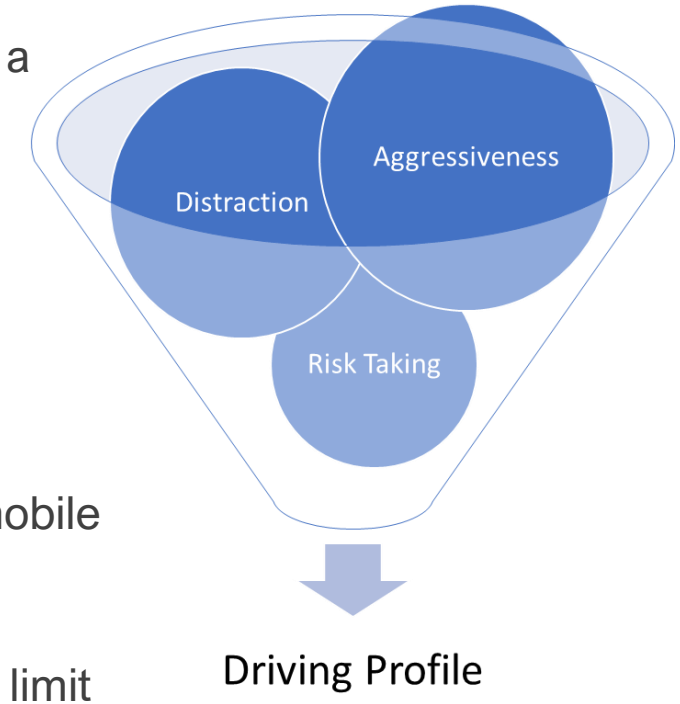
- Harsh accelerations/brakes per km
- Smoothness indicator (Kinetic energy during acceleration)
- Standard deviation of acceleration

Distraction

- Percent of mobile usage (% of trip duration in which the driver interacts with the mobile phone

Risk taking

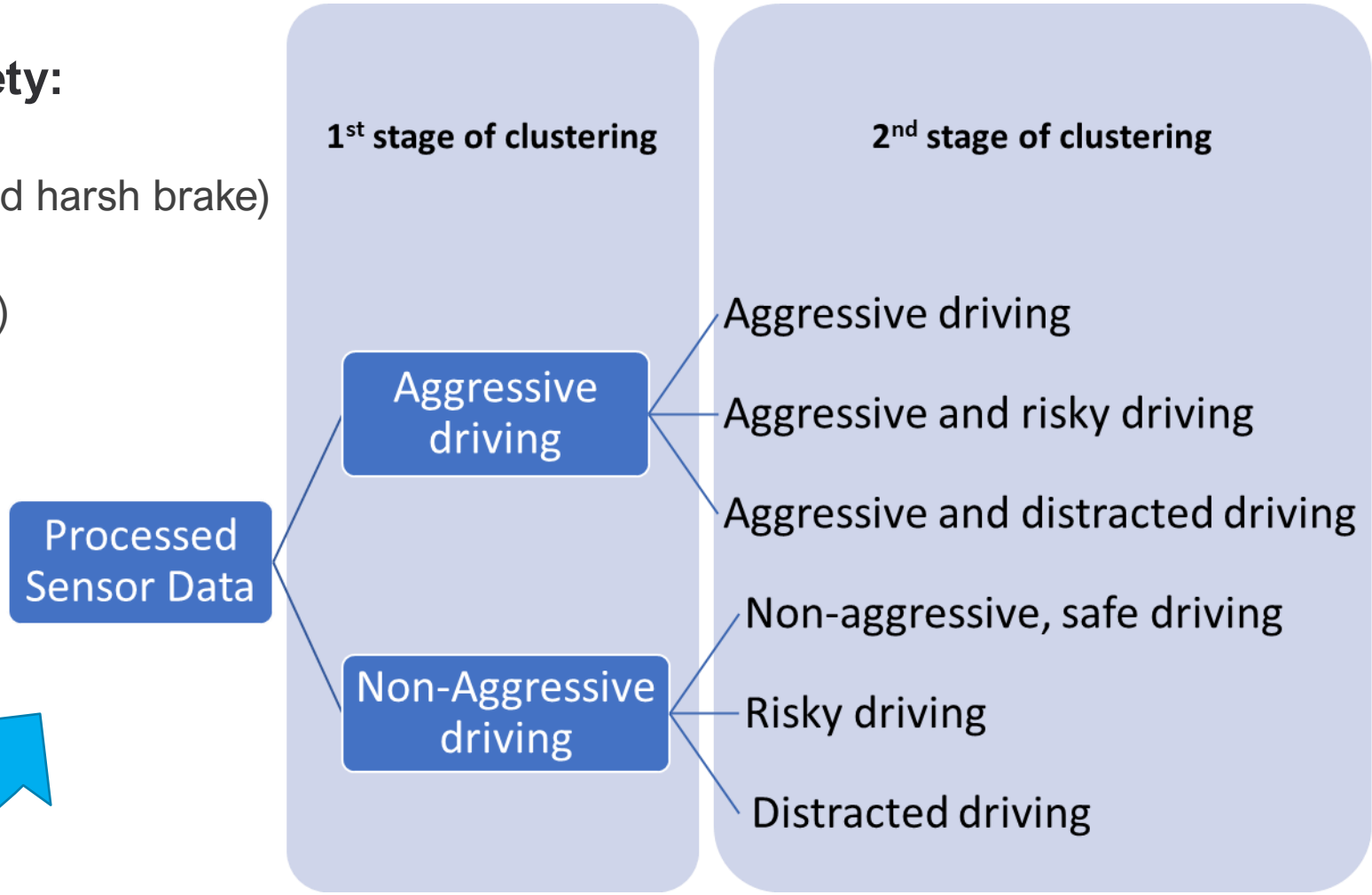
- Percent of speeding (% of trip duration in which the vehicle travels over the speed limit



Driving Profiles and Drivers' Profiles

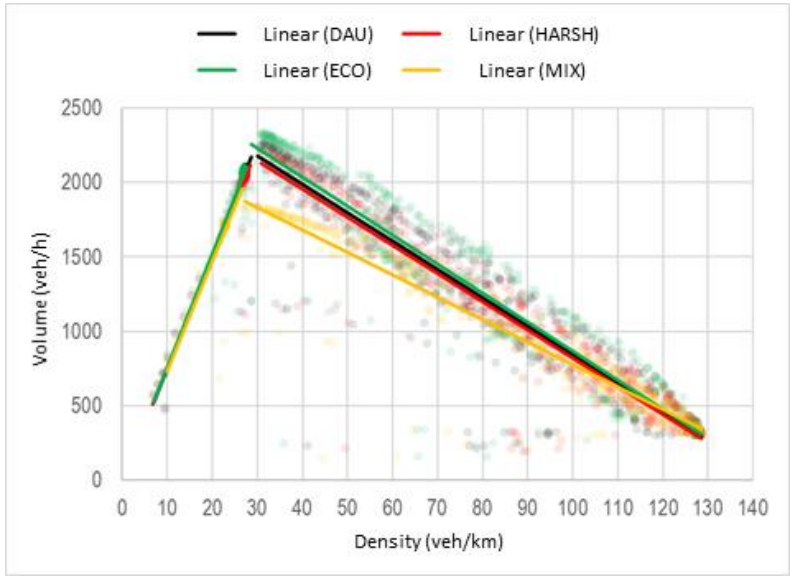
Ranked by importance to driving safety:

- Safe behavior
- Aggressive behavior (harsh accelerate and harsh brake)
- Risky behavior (speed limit violations)
- Distracted Behavior (mobile phone usage)
- Aggressive and Risky behavior
- Aggressive and Distracted behavior



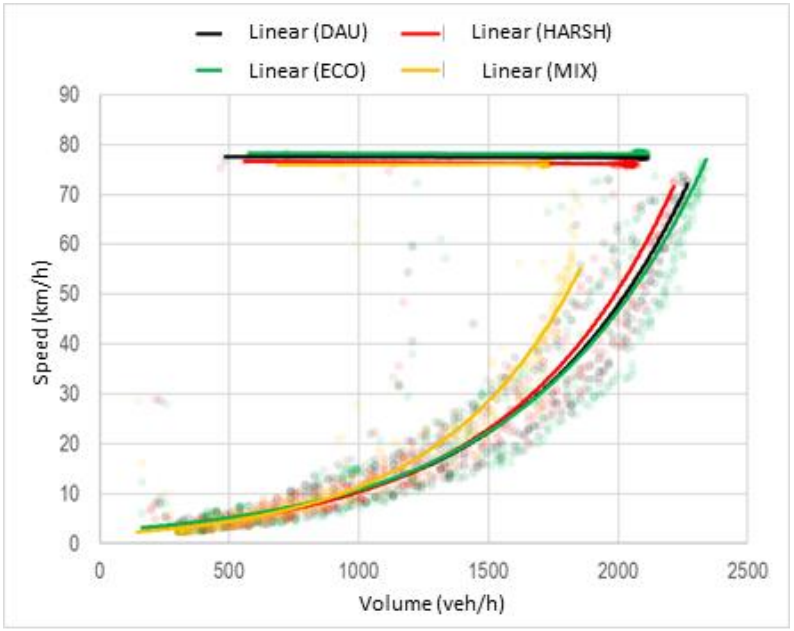
Driving Profiles and Drivers' Profiles

Network Level Impacts



HARSH and ECO model perform better than MIX in terms of mobility

- Homogeneity of traffic
- Reduced micro-variations of speed
- Increase of road capacity and reduce of pollutants



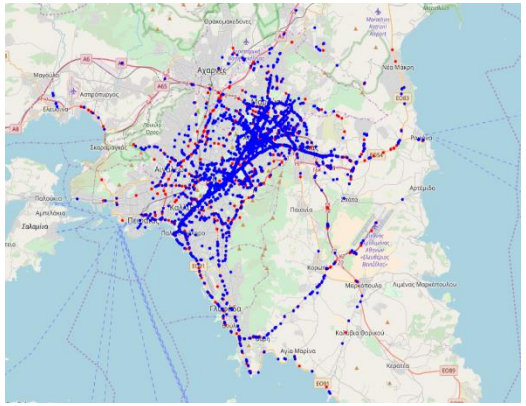
Project Luxembourg SUMO Traffic Scenario 2.0 (LuST)

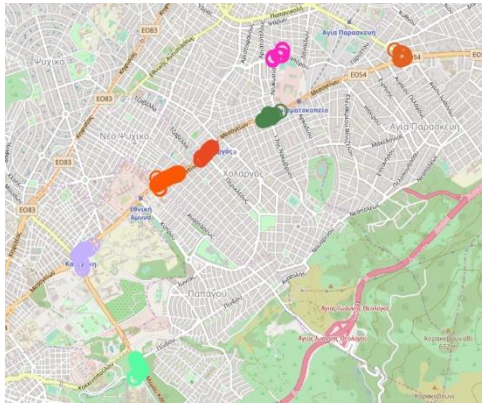
- 24h Simulation
- 5959 edges, 929.5 km
- 4477 junctions
- 203 traffic lights
- 3155 Induction Loops
- Demand Peak: max.5000 vehicles/h

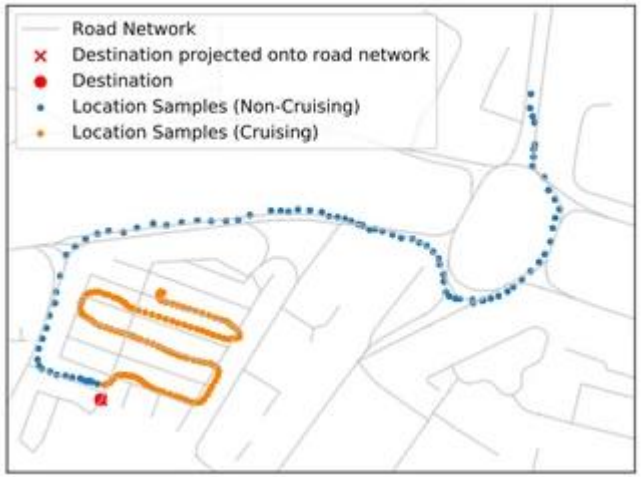
From user patterns to macroscopic information…

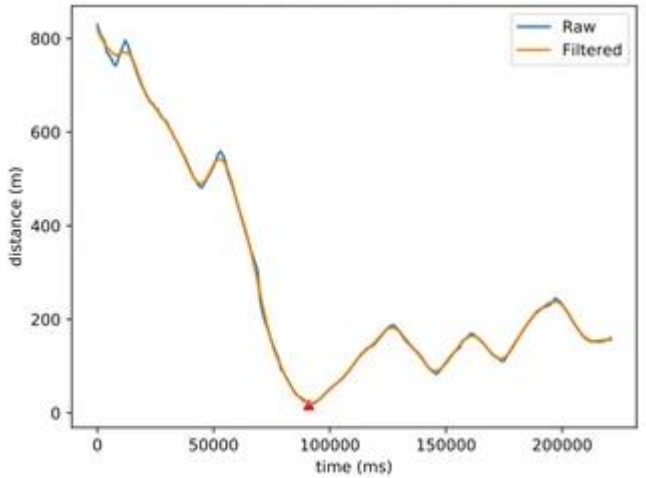
Some applications

- Speeding Information
- OD estimation
- Traffic Safety Maps
- Circulation for Parking









Outlier Detection – Anomaly Detection

Definition of an Outlier

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism" [Hawkins 1980]

Intuition (Statistics-based)

- normal data objects follow a "generating mechanism", e.g. some given statistical process, abnormal objects deviate from this generating mechanism

- Why outliers (and their detection) are important? ***One person's noise could be another person's signal."***

Applications

Fraud detection
- Purchasing behavior of a credit card owner usually changes when the card is stolen card is stolen
- Abnormal buying patterns can characterize credit card abuse

Medicine
- Unusual symptoms or test results may indicate potential health problems of a patient

Security and Surveillance
- abnormal motion detection in a video scene

Detecting measurement errors
- Data derived from sensors (e.g. in a given scientific experiment) may contain measurement errors
- Abnormal values could provide an indication of a measurement error
- Removing such errors can be important in other data mining and data analysis tasks

Traffic Operations and Safety
- Abnormal speed distribution, or time streams of vehicle traffic/crowd flow data may indicate and accident or a non recurrent congestion, or crowd panic conditions

Novelty vs Outlier Detection

"**novelty detection”**: your data set contains only good data, **and** you're trying to determine whether new observations fit within the existing data set

"**outlier detection**", your data set may contain **outliers**, which you want to identify.

Anomaly detection: an example

Users' mobility features:

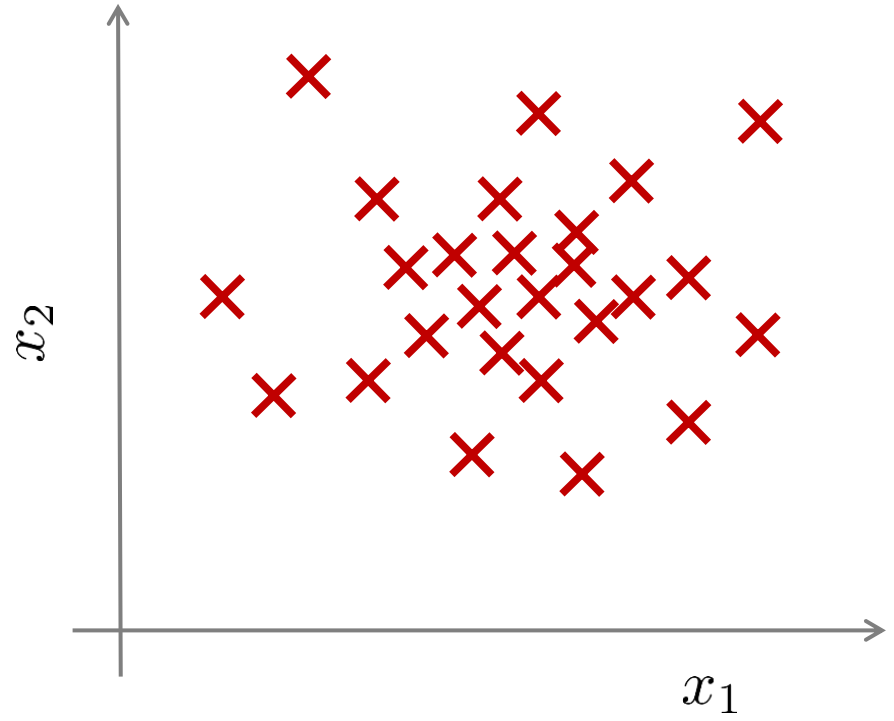
x_1 = trips/day

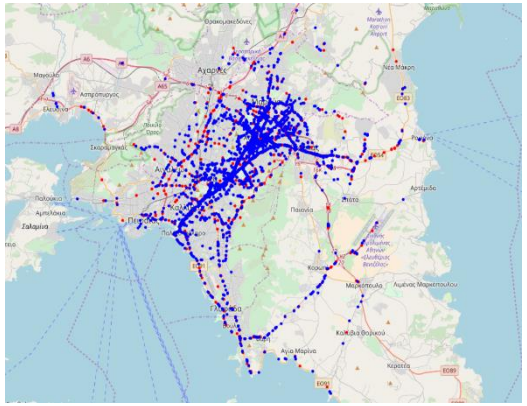
x_2 = duration between trips

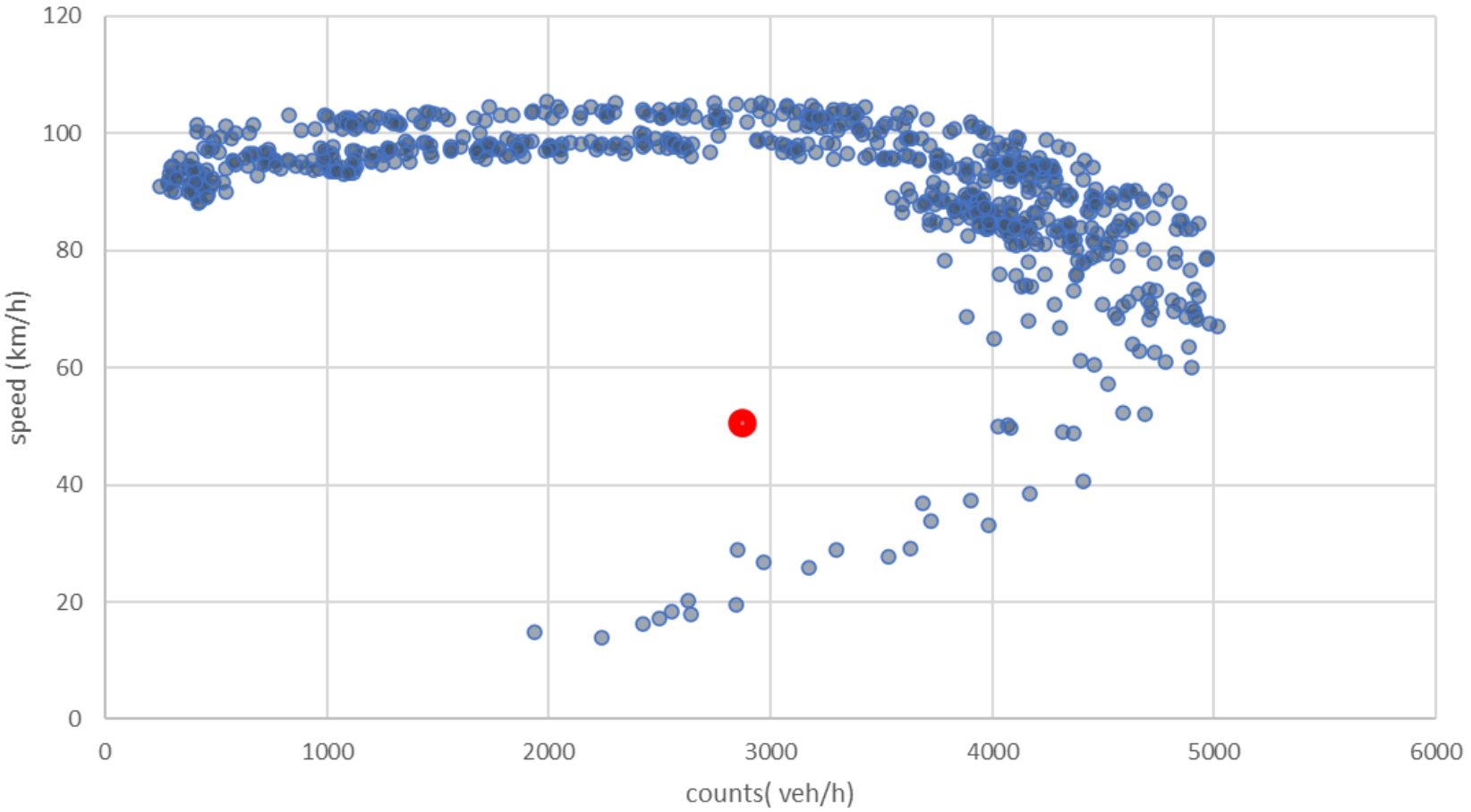
…

Dataset: $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

New user: x_{test}






120
100
80
60
40
20
0
speed (km/h)
0
1000
2000
3000
4000
5000
6000
counts(veh/h)

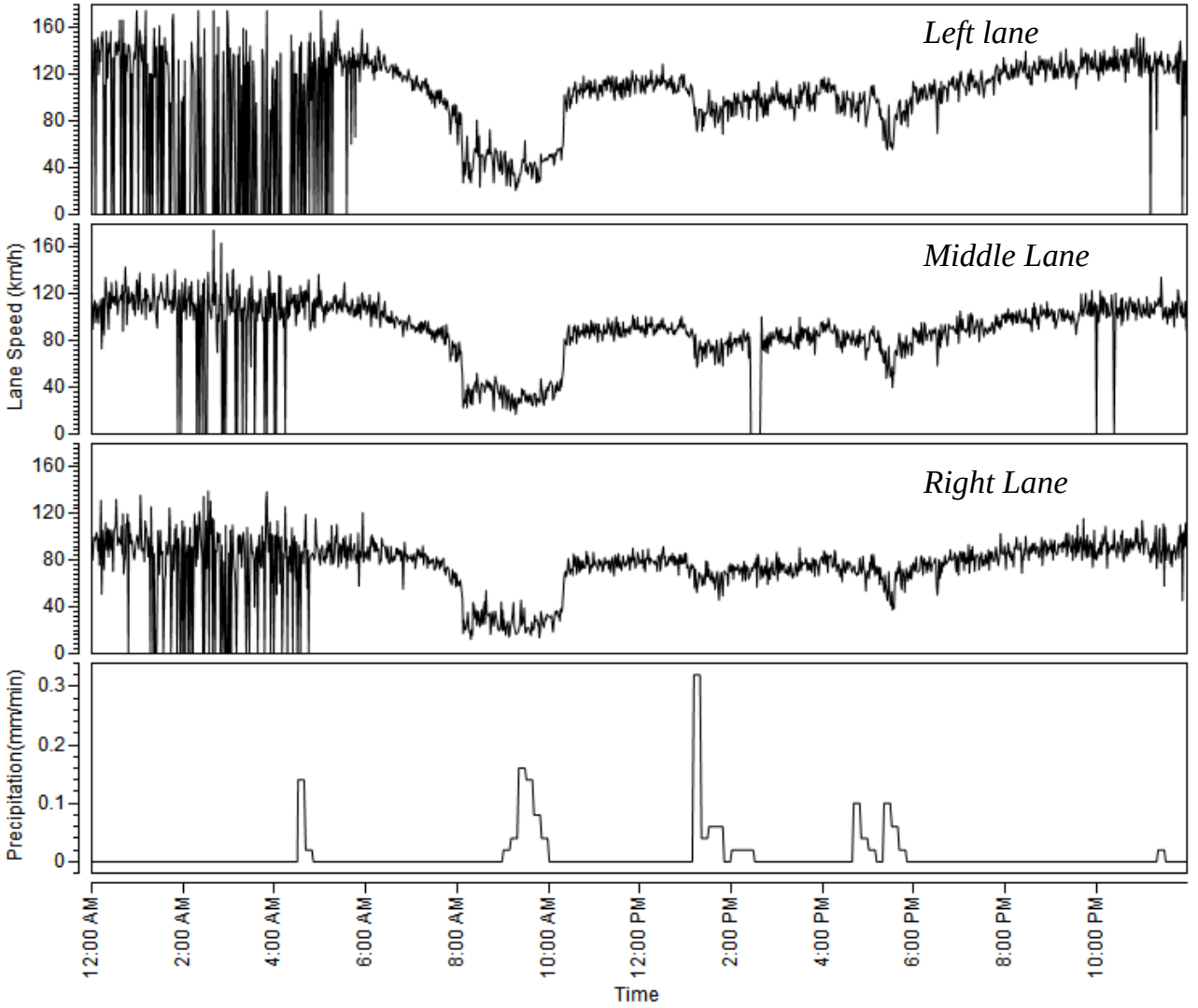
Types of Anomaly

Point anomalies (a point that significantly deviates from others)

Contextual anomalies (anomaly within a context)

Collective anomalies (a collection of points are rare)

- Requires structure (temporal, spatial etc)



Anomaly Detection:

Do we have access to labeled data (including anomalous cases)?

- Enough anomalous cases? supervised learning approach (dataset division into three parts!)
- Very few of no knowledge of anomaly: Unsupervised learning approach (very limited labeled data could be critical!) – How to proceed?

  Training (normal cases e.g. 60% of cases) – CV (normal cases e.g. 20% of cases , 50% of cases with anomaly) – Test (normal cases e.g. 20% of cases , 50% of cases with anomaly)

Some remarks

– Evaluation based on classification matrix…
  Precision/Recall and F_1–score

What features to choose?

Accuracy is not a good measure for imbalanced datasets!!!

Oversampling, Undersampling, Cost Sensitive Learning

Unsupervised Anomaly Detection

No labels assumed

Assumption: anomalies are very rare compared to normal data

General Approach

- Build a profile of “normal” behavior (univariate or multivariate thinking)
- Use the “normal” profile to detect anomalies (observations whose characteristics differ significantly from the normal profile)

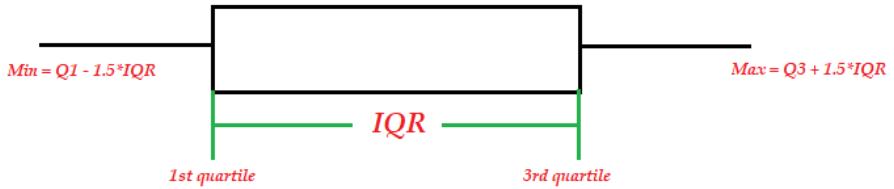
Methods

- Statistical (model based)
- Distance based
- Density based
- Clustering
- other

Statistical based Approaches

Using simple tools

- Boxplot
- *Values that are lower than Q1−1.5IQR or higher than Q3+1.5IQR are considered outliers.*
- *The drawback of this method is that it takes into account only one variable at a time.*



- Z-score
- *Quantifies the distance of a data point from the rest of the dataset. In a more technical term, Z-score equals the number of standard deviations away a given observation is from the mean.*
- *A threshold for the value of Z-score should be defined, in order to detect the outliers.*

$$\text{Z-score} = \frac{x - \text{mean}}{\text{Standard Deviation}}$$

Statistical Based Approaches

Outliers are objects that fit poorly by a statistical model.

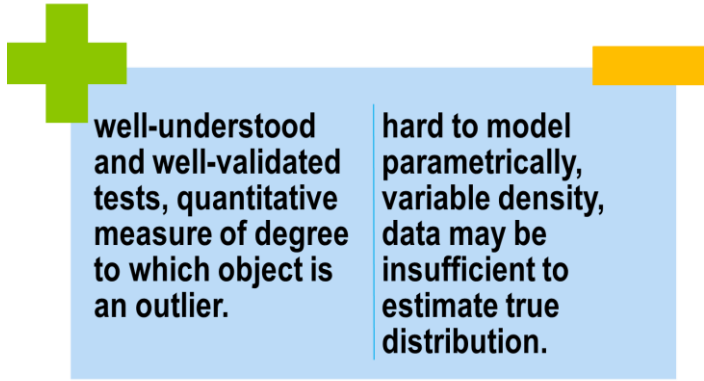
- Estimate a parametric model describing the distribution of the data
- Apply a statistical test that depends on

  Properties of test instance ,
  Parameters of model (e.g., mean, variance)
  Confidence limit (related to number of expected outliers)

Multivariate Gaussian distribution – Outlier defined by Mahalanobis distance > threshold

Grubbs' test (for univariate data, H 0: There is no outlier in data)

- Assume data comes from normal distribution



Statistical Based Approaches

An Example - BACON (Blocked Adaptive Computationally Efficient Outlier Nominators)

Algorithm 3: the BACON algorithm for identifying outliers in multivariate data

Input: An $n \times p$ matrix $\boldsymbol{X}$ of multivariate data.

Output: A set of observations nominated as outliers and the discrepancies for all observations based on (3) relative to the final basic subset.

Step 1: Select an initial basic subset of size m using either V1 or V2 of Algorithm 2.

Step 2: Compute the *discrepancies*

$$d_i(\bar{\boldsymbol{x}}_b, \boldsymbol{S}_b) = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_b)^T \boldsymbol{S}_b^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_b)}, \quad i = 1, \ldots, n, \tag{3}$$

where $\bar{\boldsymbol{x}}_b$ and $\boldsymbol{S}_b$ are the mean and covariance matrix of the observations in the basic subset.

Step 3: Set the new basic subset to all points with discrepancy less than $c_{npr}\chi_{p,\alpha/n}$, where $\chi^2_{p,\alpha}$ is the $1-\alpha$ percentile of the chi square distribution with p degrees of freedom, $c_{npr} = c_{np} + c_{hr}$ is a correction factor, $c_{hr} = \max\{0, (h-r)/(h+r)\}, h = [(n+p+1)/2], r$ is the size of the current basic subset, and

$$c_{np} = 1 + \frac{p+1}{n-p} + \frac{1}{n-h-p} = 1 + \frac{p+1}{n-p} + \frac{2}{n-1-3p}. \tag{4}$$

(When the size of the basic subset r is much smaller than h, the elements of the covariance matrix tend to be smaller than they should be. Thus, one can think of c_{hr} as a variance inflation factor that is used to inflate the variance when r is much smaller than h. Note also that when $r = h$, c_{npr} reduces to c_{np}.)

Step 4: The *stopping rule:* Iterate Steps 2 and 3 until the size of the basic subset no longer changes.

Step 5: Nominate the observations excluded by the final basic subset as outliers.

produces a set of observations nominated as outliers along with the discrepancies based on the Mahalanobis distance relative to a basic subset of the data

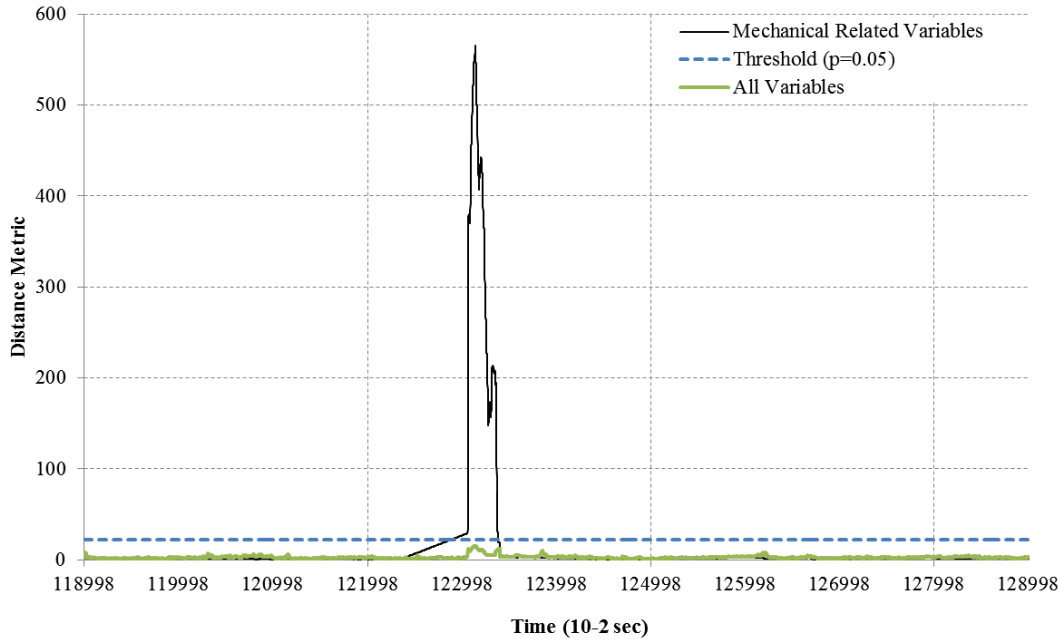
In contrast to the Euclidean distance, the Mahalanobis distance takes into account the correlation structure of the data as well as the individual scales

Results

Detecting Powered Two Wheelers Driving Incidents

- Three models are evaluated using different data:
  - *Model 1: steering, throttle, brake activation and wheel speed.*
  - *Model 2: linear acceleration and speed.*
  - *Model 3: All available variables.*

Distance metric time series of Model 1 and Model 3. Any distance metric value above the 5% threshold value signifies an irregular behavior (outlier).

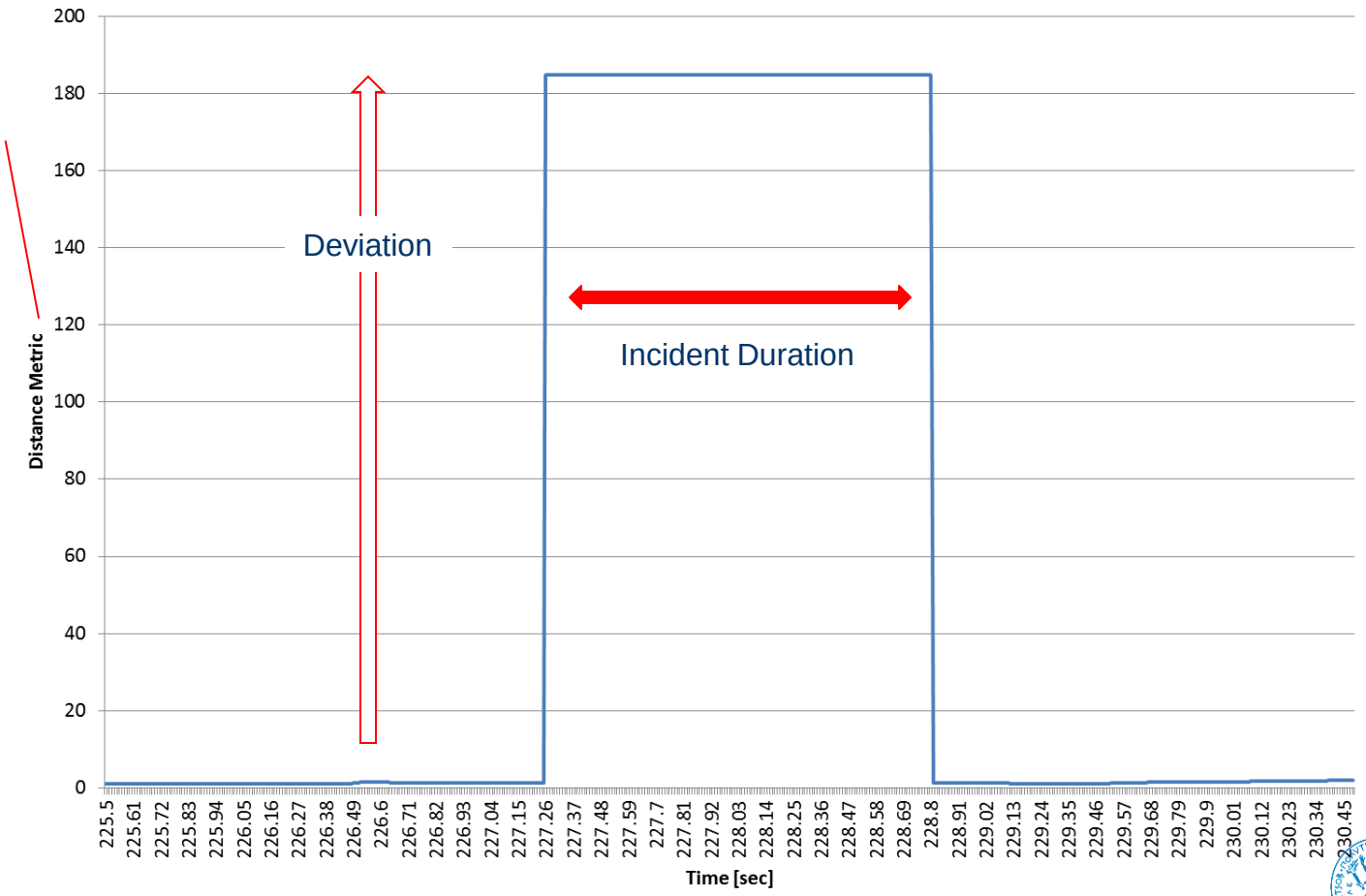


- Model 1 is the optimum
- Traffic related variables have been found less influential in detecting incidents.

Results

An example of detected incident using the proposed methodology (5 sec data, incident duration: ~ 1.5 sec)

Joint consideration of changes in throttle, steering, braking and wheel speed



Distance-based outlier detection

Outliers are objects far away from other objects.

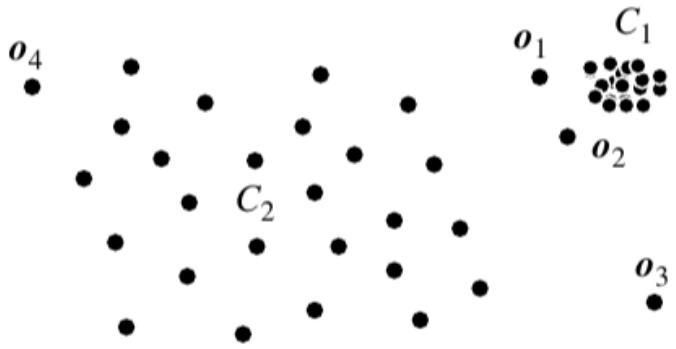
- the greater is the distance of the object to its neighbors, the more likely it is an outlier

Approach:

- Outlier score is distance to kth nearest neighbor.
- Score sensitive to choice of k.

Examples:

- KNN (Outlier Detection based on the distance of an object to its k nearest neighbor)



Easier to define a proximity measure for a dataset than determine its statistical distribution, quantitative measure of degree to which object is an outlier, deals naturally with multivariate data.

$O(n^2)$ complexity, score sensitive to choice of k, does not work well if data has widely variable density.

Distance-based outlier detection

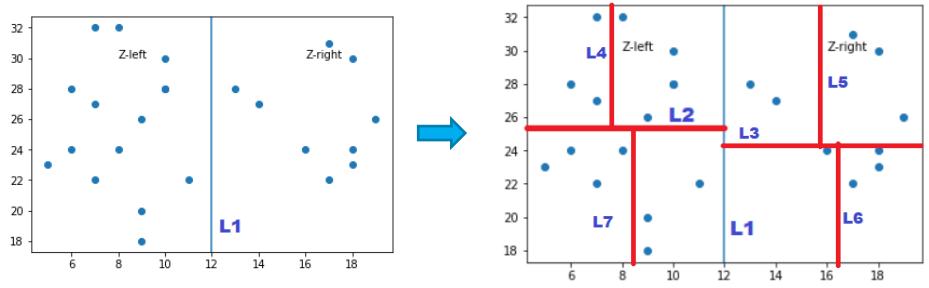
KNN – how it works

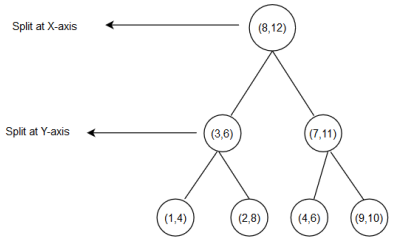
- Compute the distance between every pair of data points (Brute force)
- There are various ways to define outliers:

  Data points for which there are fewer than p neighboring points within a distance D

  The top n data points whose distance to the kth nearest neighbor is greatest

  The top n data points whose average distance to the k nearest neighbors is greatest





Efficient Computation of distances with K-D Tree

Kdtree attributes

1.What dimension to split? Choose the Widest dimension first or the alternating dimensions

2.What value to split? Choose the median value or the Centre of all the values

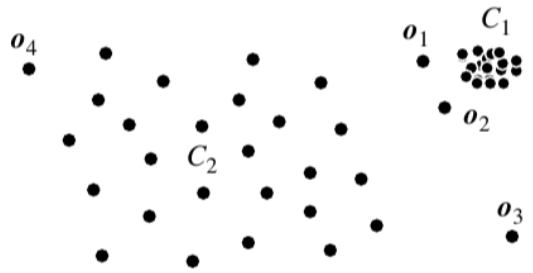
3.When do we stop splitting? When there are fewer data points are left than a specific value say m

Density-based outlier detection

Outliers are objects in regions of low density.

- LOF (Local Outlier Factor) - local density deviation of a given data point with respect to the data points near it
- locality is given by k nearest neighbors
- a normal instance is expected to have a local density similar to that of its neighbors, while abnormal data are expected to have much smaller local density.

LOF(k) ~ 1 means Similar density as neighbors,
LOF(k) < 1 means Higher density than neighbors (Inlier),
LOF(k) > 1 means Lower density than neighbors (Outlier)



Quantitative measure of degree to which object is an outlier, can work well even if data has variable density	$O(n^2)$ complexity, must choose parameters k for nearest neighbor d for distance threshold

LOF explained

1. For each data record q compute the **k- distance(q)** as the distance to the kth-nearest neighbor of q.

2. Compute the reachability distance for each data record q with respect to data record p as follows:

$$\boldsymbol{reach} - \boldsymbol{dist}_k(q,p) = \max\{d(q,p), \mathbf{k} - \mathbf{distance}(p)\}$$

Euclidean distance from q to p

3. Compute the **local reachability density** (lrd) of data record q:

$$lrd_k(q) = 1 / \left(\frac{\sum_{p \in N_k(q)} reach - dist_k(q,p)}{|N_k(q)|} \right)$$

set of k-nearest neighbors of data record q

4. Compute the **LOF** of data record q:

$$LOF_k(q) = \frac{\sum_{p \in N_k(q)} \frac{lrd_k(p)}{lrd_k(q)}}{|N_k(q)|}$$

- Parameter k plays important role. k=20 is usually a proper value
- When the proportion of outliers is high (i.e. greater than 10 %), k should be greater.

Cluster-based outlier detection

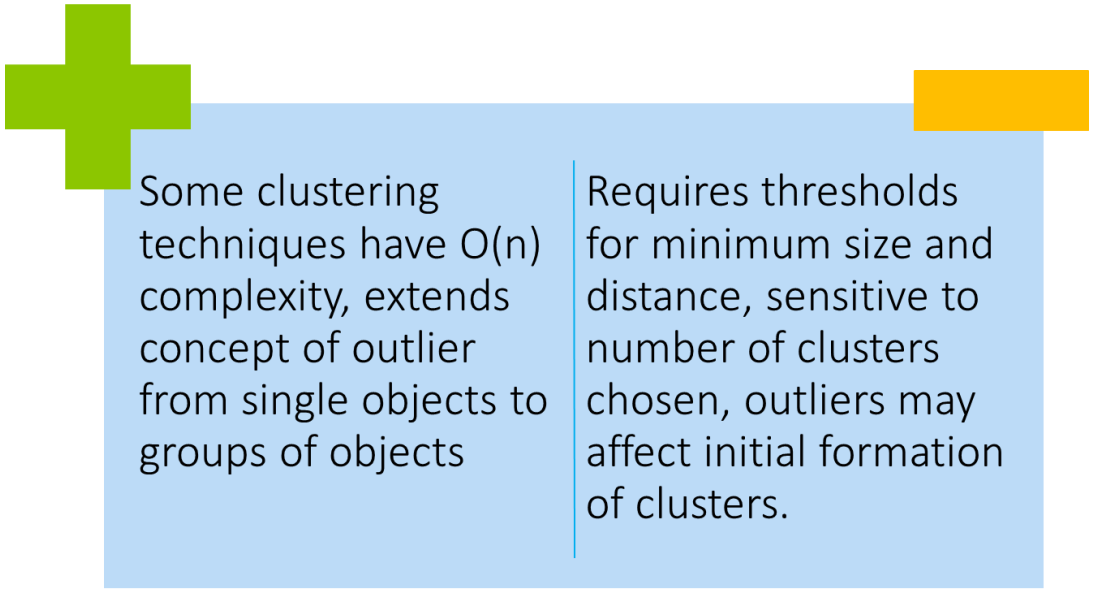
Outliers are objects that do not belong strongly to any cluster.

Approach:

- Assess degree to which object belongs to any cluster
- Eliminate object(s) to improve objective function
- Discard small clusters far from other clusters.

Examples:

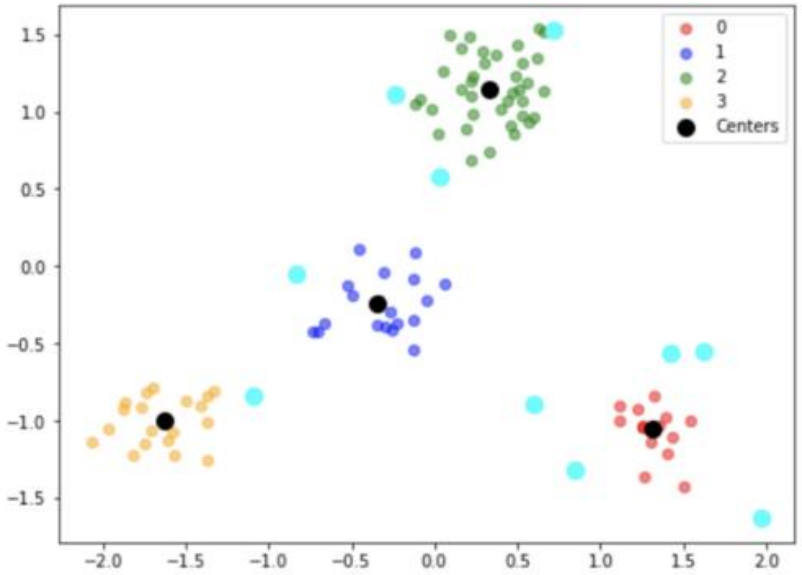
- EMOutlier
- KMeansOutlierDetection



Cluster-based outlier detection

Kmeans (compactness and separation)

- The basic idea is to detect points that are far from any cluster center and/or small clusters (with few points) that are far from the other clusters.
- The dataset is separated into K clusters, and each cluster has its center. All data belong to the nearest cluster.
- For each point, the distance from the center of the cluster it belongs to is calculated.
- The 5% (for example) of the points with the highest distance to the corresponding center are identified as outliers.
- K value is very important! A low number of clusters may result in more outliers (higher distances) than the actual ones, while a high number would have the opposite effect.
- All members of a cluster may be identified as outliers, e.g. if they are very few and far from the other clusters



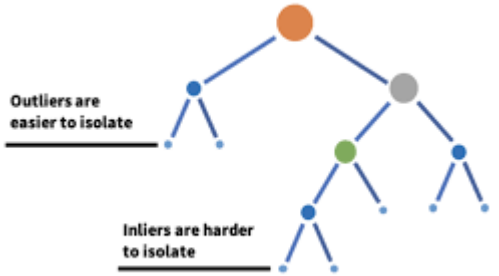
Tree-based outlier detection

Isolation Forest (identifies anomalies by isolating outliers in the data)

- Anomalous data points in a sample have the following properties:

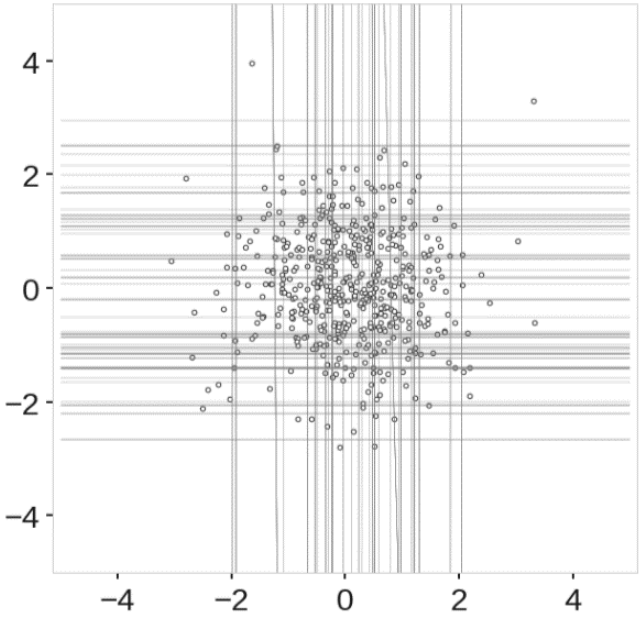
  Few — they are the minority consisting of fewer instances and

  Different — they have attribute-values that are very different from those of normal instances



Fits a tree-based model (e.g. Random Forest) to the data and assigns higher anomaly scores to data points that need few splits to be isolated (leaf nodes).

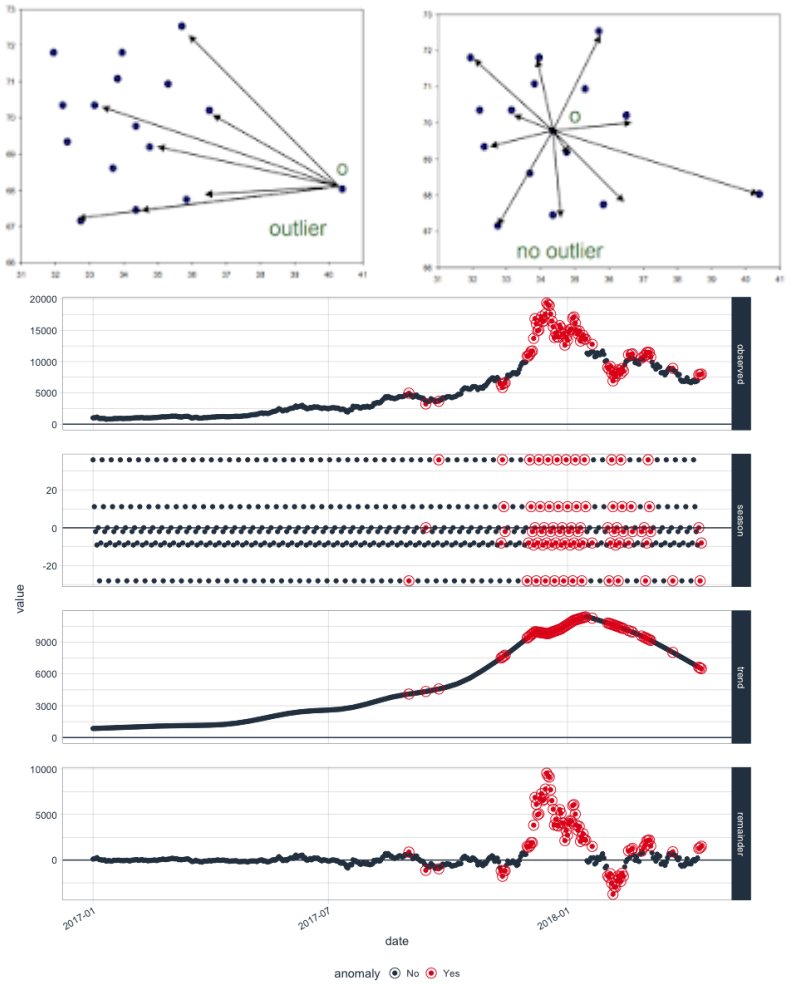
- Simply put, the number of splittings required to isolate a sample instance is equivalent to the path length from the root node to the terminal node.
- This path length, averaged over a forest of such random trees, is a measure of normality and can be used as a decision function.
- Random partitioning produces noticeably shorter paths for anomalies. Hence, when a forest of random trees collectively produces shorter path lengths for particular samples, they are highly likely to be anomalies.



Other approaches

Angle-based outlier detection

- especially useful for high-dimensional data, as angle is a more robust measure than distance in high-dimensional space – abodOutlier package



Time series anomaly detection

- Time series decomposition - anomalize package
- Error reconstruction via time series modeling (including deep learning/recurrent nn)

Graph & Network Outlier Detection

- Outliers (in large graphs) can be portions of the network, which might be nodes, edges, or even subgraphs

https://github.com/yzhao062/anomaly-detection-resources

https://github.com/pridiltal/ctv-AnomalyDetection

Example I

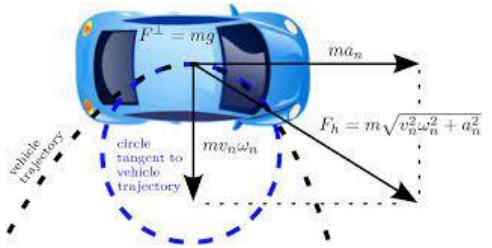
Harsh Driving Detection (Unsupervised Approach)

Example I : harsh event detection (unsupervised approach)

The dataset: driving data from n drivers collected using the accelerometer, gyroscope and GPS of the smartphone

NewAccelX	NewAccelY	NewAccelZ	NewRotRateX	NewRotRateY	NewRotRateZ	locationSpeed
0.001	-0.002	-0.007	0.002	-0.001	0	0
0.007	0	0.011	-0.001	0.001	0	0
0.008	0.004	0.015	0	0	-0.001	0
0.008	0.002	0.004	0.002	-0.001	-0.001	0
0.009	0.001	0.026	-0.002	-0.001	0.002	0
0.012	-0.001	0.003	0	0	-0.001	0
0.007	0	0.004	-0.001	0	0	0
-0.008	-0.006	0.002	0.001	-0.001	0	0
-0.004	0.005	0.022	0	0.001	0.001	0
0.007	0.007	0.016	0	-0.002	0.003	0
0.013	0.011	0.019	-0.001	0	0.001	0
0.013	0.002	0.011	0.001	0.002	-0.002	0
-0.008	-0.003	-0.002	0.002	-0.003	-0.001	0
-0.043	-0.005	0.02	0.004	-0.018	0	0
-0.008	-0.005	-0.002	0.001	-0.002	-0.002	0
0.009	-0.038	0.019	0.002	-0.002	0.001	0
0.028	0.009	0.021	0.004	-0.01	0.014	0.42
-0.013	0.004	-0.07	0.004	0.011	0.006	0
0.017	-0.013	0.009	-0.002	-0.005	0.033	0
0.042	-0.017	-0.018	-0.01	-0.01	0.01	0
-0.039	0.026	-0.008	0.004	-0.004	0.007	0.45

Detect harsh cornering



The signals are re-oriented to match the driving direction (y-axis is the driving direction)

Our sample is annotated (naturalistic driving experiment) – you cannot use the true information to train a model but you can use the true information to evaluate your detection algorithm

Submit:

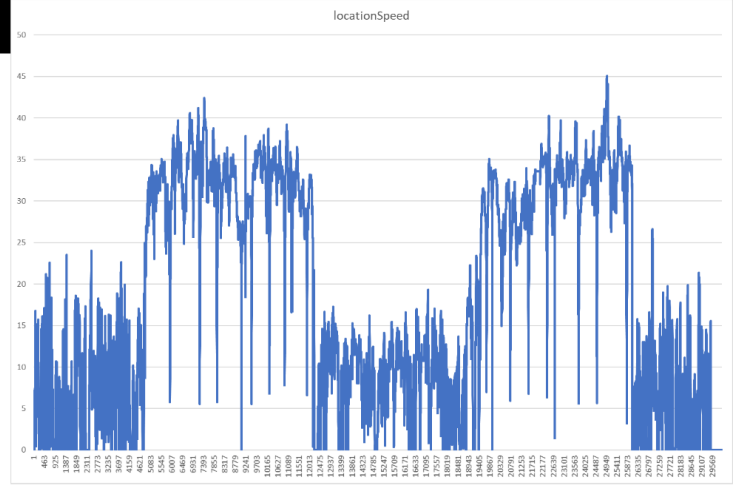
- a script
- a technical report (less than 10 pages)
- csv with the detected corneringsv

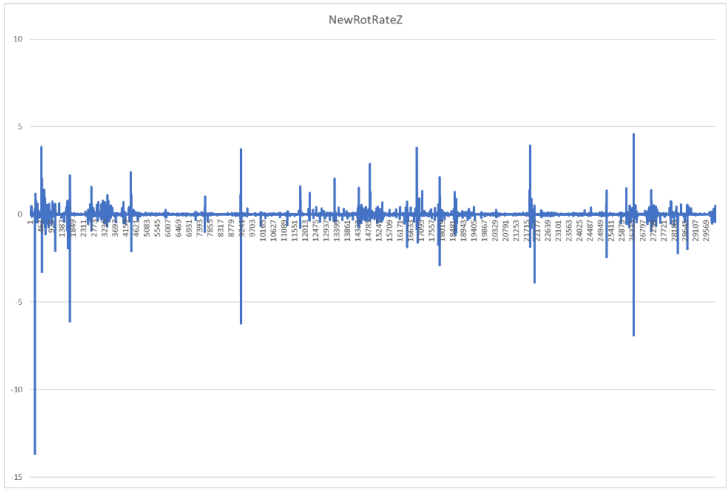
Example I : harsh event detection (unsupervised approach)

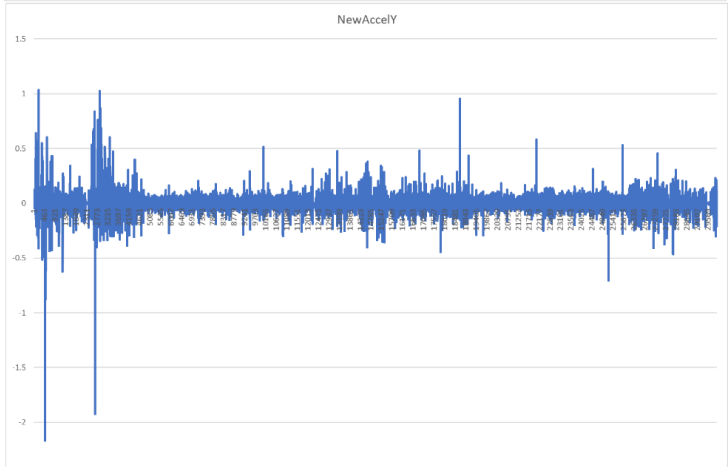
The variables

- Rotated accelerations x/y/z– km/h/sec
- Rotation rates x/y/z - rad/sec
- Location Speed – km/h

- Can you visually detect candidate outliers?





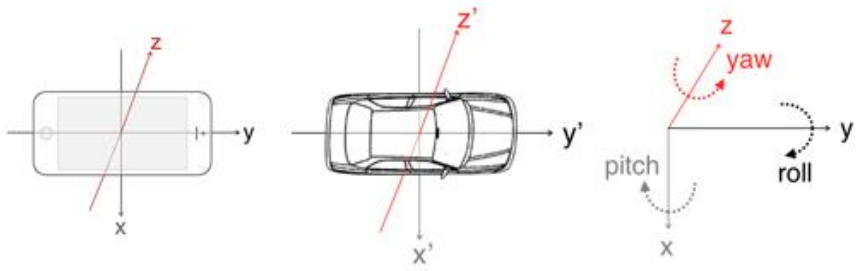


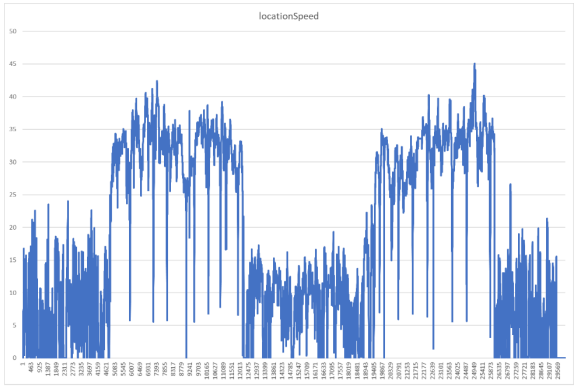
Example I : harsh event detection (unsupervised approach)

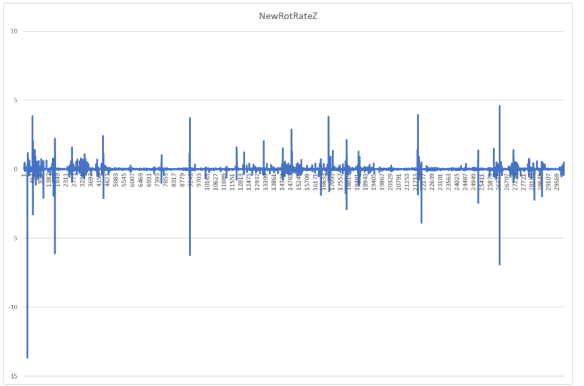
The variables

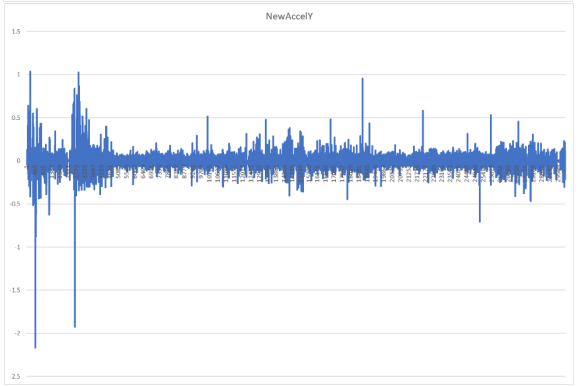
- Rotated acceleration x/y/z – km/h/sec
- Rotation rate x/y/z - rad/sec
- Location Speed – km/h

- What are the most influential variables in harsh cornering detection?









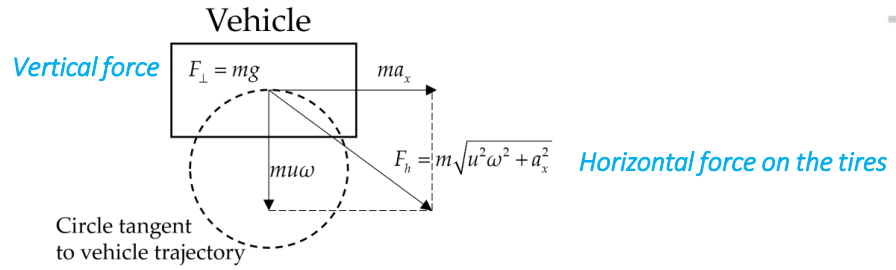
Example I : harsh event detection (unsupervised approach)

The variables

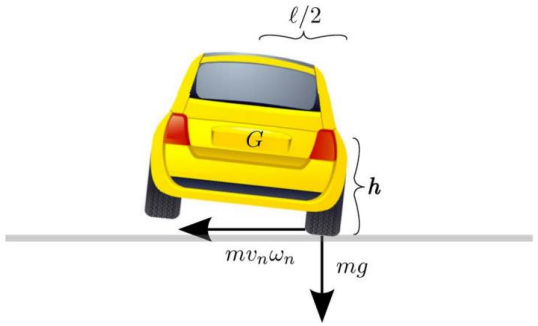
- Rotated acceleration x/y/z – km/h/sec
- Rotation rate x/y/z - rad/sec
- Location Speed – km/h

- Is feature engineering useful?

Developing forces on a turning vehicle based on classical mechanics

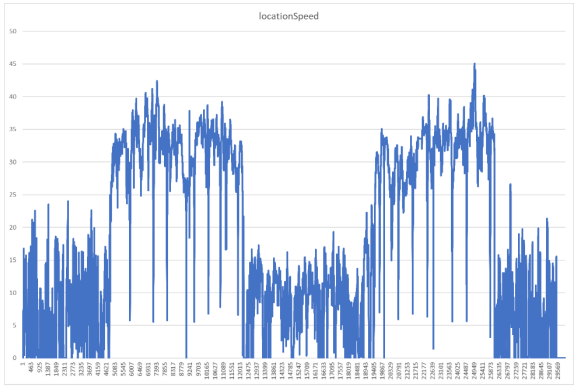


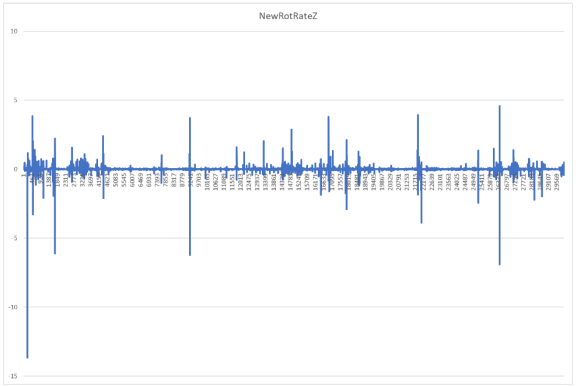
Horizontal force on the tires

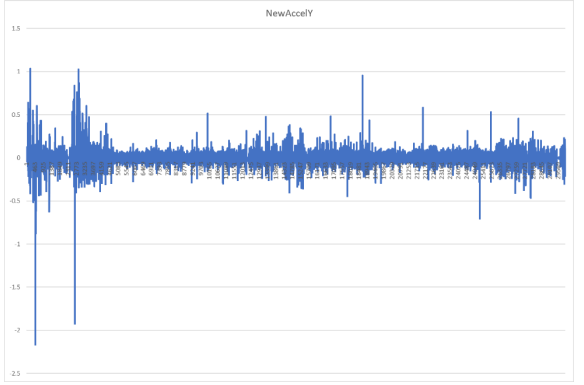


$$T(u, a_x, \omega) = F_h / F_\perp = \frac{1}{g}\sqrt{u^2\omega^2 + a_x^2}$$

Indicator of turning intensity!







Example I : harsh event detection (unsupervised approach)

What you will have to do…

- Descriptive statistics and plotting to observe outliers or abnormal patterns
- Apply outlier detection approaches (statistical based, distance based and density based)

  Select proper parameterization of the algorithms

  Apply on different combinations of input variables

  Compare based on an annotated sample

Next Lecture

Time series Analysis

- Most popular structures
- Deep learning

Traffic (State) Forecasting

- Theory facilitated data-driven models
- Example in python

Machine Learning in Transportation

Eleni I. Vlahogianni, Ph.D.

National Technical University of Athens

(elenivl@central.ntua.gr)