

# Introduction to Detection of Non-Technical Losses using Data Analytics

September 26, 2017

Patrick Glauner, Jorge Augusto Meira, Radu State,  
Interdisciplinary Centre for Security, Reliability and Trust,  
University of Luxembourg

Rui Mano

Choice Technologies Holding, Luxembourg

# Motivation

What is a typical non-technical loss (NTL)?



# Motivation

What is a typical non-technical loss (NTL)?



# Motivation

What is a typical non-technical loss (NTL)?



# Motivation

Worldwide electric utilities  
lose \$96 billion USD (\*)  
annually to fraud/theft

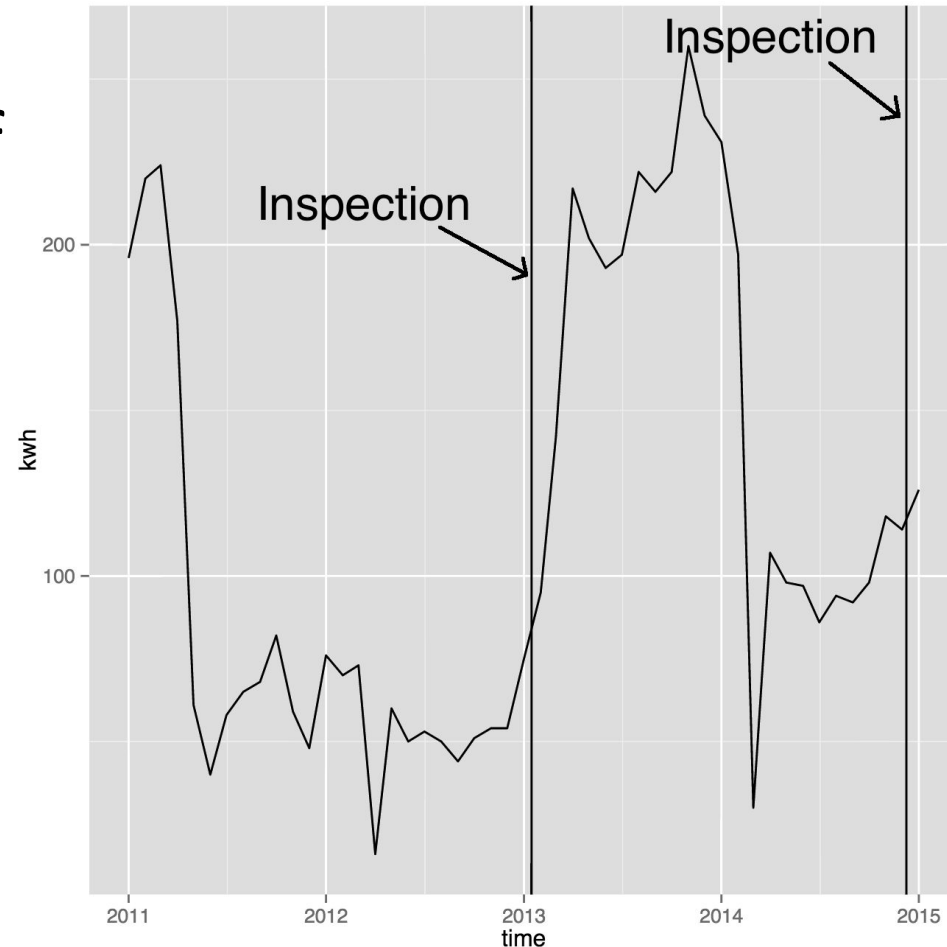
(\*) Electricity Theft and Non-Technical Losses: Global Market, Solutions and Vendors  
May 2017 | Northeast group, llc

[www.northeast-group.com](http://www.northeast-group.com)

(\*) World Bank data

# Motivation

Example of NTL: Two assumed occurrences of NTL due to significant consumption drops followed by inspections (visualized by vertical bars).



# Project Overview

- Joint university-industry research project on detection of NTL
- Started in late 2015
- Goal: applied R&D focused on both publishing papers and deploying features in Choice products



SNT

securityandtrust.lu



UNIVERSITÉ DU  
LUXEMBOURG



# Project Overview

- Published results in:
  - 1 journal paper (survey)
  - 1 paper at ISGT 2016
  - 1 paper at Power and Energy Conference at Illinois (PECI) 2016
  - 1 paper at International Conference on Intelligent System Applications to Power Systems (ISAP) 2017
  - 1 paper at IEEE/ACM International Conference on Big Data Computing Applications and Technologies (BDCAT) 2016



securityandtrust.lu



UNIVERSITÉ DU  
LUXEMBOURG





# Goals of this tutorial

- Introducing the problem of NTL
- Providing a short introduction to machine learning
- Reviewing the state of the art of NTL detection using machine learning methods
- Discussing the challenges of NTL detection
- Presenting a selection of our works
- Providing a forum for discussions

# Interested in NTL?

Join our mailing list:

<https://groups.google.com/d/forum/ntl-community>

-  We plan to organize a NTL detection competition  
By Dr. Ing. Carlos López-Vázquez - 1 post - 4 views
-  Upcoming conferences to discuss NTL detection  
By me - 4 posts - 5 views

# Contents

- Introduction to NTL
- Machine Learning
- State of the art
- Challenges
- Selection of our works
- Conclusions

# Introduction to NTL

Non-technical losses are mainly caused by fraud activities deliberately performed by the consumers.

Besides the financial issues due to non-revenue energy, these frauds lead to a series of additional losses, including damage to grid infrastructure, reduction of grid reliability and may be cause of accidents.

# Conceptualization

Theft is described as connecting directly to energy sources bypassing the metering process.



# Conceptualization

Fraud can be defined as altering the measurement registered by the meter,

i.e. tampering the meter. It is easier with the conventional meters but it is now also performed with electronic, said smart meters.



# Infrastructure

NTL will require reinforced infrastructure to support the additional load from unmetered consumption. Some grid devices not designed for the actual load will deteriorate faster and result in supply interruptions, degrading the quality of service.

# Economics

Tariffs paid by customers are designed to correspond to their consumption and remunerate Utilities for their investments and operational costs. When the metered consumption is adulterated, all the economics of the process are impacted.

(\* More to read on Article “Non-technical Losses in Utility Business – What it is and why it does matter to all of us”.

Rui Mano, 2017 – Metering International magazine, edition 4/2017



# Security

Insecure manipulation by non-authorized personnel may be the cause for damaged appliances and accidents (short circuits, disconnections, electrocution, and fire).

# Regulation

Regulators define penalties for fraud/theft perpetrators, from fines all the way to criminal prosecution.

Stealing energy is, in some countries, defined as a crime.

# Fighting NTL

First step is to discover who is doing a fraud, then Utility needs to inspect and make legal evidence of the fraud. Inspection is a key issue.

# Fighting NTL

**Field audits are the only way to make due evidence of a fraud or theft. Analytical methods are not accepted as evidence.**

# Inspections

Inspections require getting into client premises. This is a) costly and b) not friendly, mostly to honest customers. Accuracy of the detection is key to avoid false positives.

# Analytics

**Analytics** can provide very effective results with short time paybacks. Choice has conducted very successful projects.

# The infrastructure scenario

- **Distribution Automation**
- **Substation Automation**
- **Metering data**
- **Systems**
  - **Corporate, Administrative, Finance**
  - **Technical (Operation, Maintenance, etc.)**
    - **Different Protocols, Standards, Data models**
    - **Data silos**
- **Transmission, Processing and Storage**

**Telecom**

**DATA**



# The NEW infrastructure scenario

- Smart Grid implementation – the new, smarter grid
  - **Distribution Automation**
  - **Substation Automation**
  - **Distributed Energy Resources, Micro Grids**
  - **Smart Metering – AMR / AMI**
  - **Systems (corporate and technical)**
    - Common Standards (ex. 61850)
    - INTEGRATION

**Telecom**

**BIG  
DATA**

**ANALYTICS**



# New Data Scenario

- Technical data (currents, voltages, device status, etc.)
    - Operation, Engineering, Maintenance...
  - Metering Data (Consumption, faults, etc.)
    - Finance, Administrative, Commercial, Customers...
  - Terabytes/month – a data Tsunami!
  - Data Transmission, Processing, Disclosure and Storage
- ANALYTICS**



# New Applications for Utilities

- Better Client Support
- Technical: Planning, Engineering, Operation and Maintenance
- Energy Efficiency, Intelligent Consumption, Hourly Tariffs
- Information and NEW services to the clients
- Techniques and tools for decision support based on knowledge
  - Artificial (Computation) Intelligence – Neural networks, Fuzzy logic, Neuro-Fuzzy, Signal Processing (Wavelets, etc), Clusterization, Machine Learning, Evolutionary Algorithms, etc...

# Data Intelligence

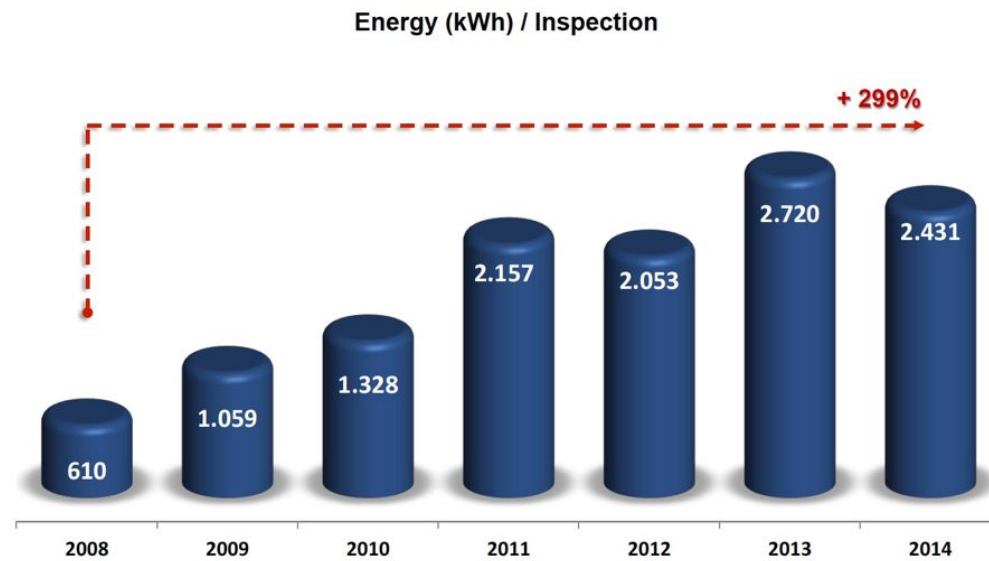
Acquiring data to store and accumulate? Or... Data intelligence? Data Analytics allow to better understand the behavior of consumers, of the load and of the grid, and render better services, with better results for the utility.

# Data Analytics Revenue Protection Projects

Choice project for Light,  
Brazilian 4.5 million  
customers utility resulted in  
recovering 4 times more  
energy per inspection

# Data Analytics Revenue Protection Projects

The chart below summarizes the evolution of the Productivity (= Total Recovered Energy (kWh) / Inspection) at Light.



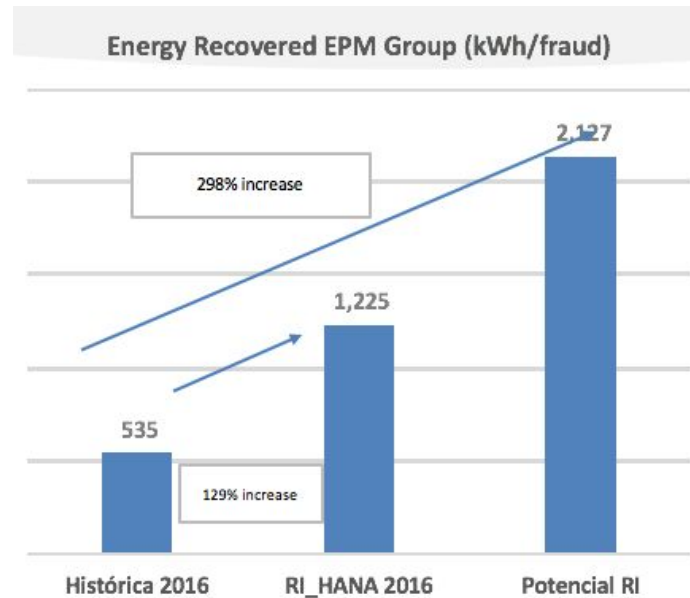
\* Start of Revenue Intelligence in 2008

# Data Analytics Revenue Protection Projects

Choice project for EPM (\*) –  
Colombian multi-country,  
multi-utility covering 7.5 million  
electricity, water, and gas  
customers more than doubled the  
recovery of energy in 6 months  
after go-live.

(\*) <http://choiceholding.com/media/> - MEET JUAN CARLOS DUQUE

# Data Analytics Revenue Protection Projects



Energy  
Recovered in  
2016

(\*) <https://www.youtube.com/channel/UC-azzVfbGnggmqWMvAfOGYA> - Proyecto gestión control pérdidas

# Conclusion

Data is not the problem, anymore. But data is not enough. To make the best from data, Analytics is the very effective technology that opens room to a myriad of new applications, impacting and helping optimize every sector of the utility.



# Conclusion

Energy diversion and fraud impacts customers and taxpayers and helps to damage the environment. This is a very high cost to the utility, to their customers and to the whole society. Utilities have the responsibility to fight non-technical losses with the best tools and methodology (\*).

(\*). More to read on “Electricity Theft and Non-Technical Losses : Global Market, Solutions and Vendors” - May 2017 – Northeast Group. Llc

[www.northeast-group.com](http://www.northeast-group.com)

# Conclusion

***“Big data is not about the data. The value of big data is in the analytics.”***

Harvard Professor Gary King

# Machine Learning

# Machine Learning

“Machine Learning is a subset of Artificial Intelligence techniques which use statistical models to enable machines to improve with experiences”\*

Use cases: *data mining, autonomous cars, recommendation...*

\*<https://rapidminer.com/artificial-intelligence-machine-learning-deep-learning/>

# Definition

**Arthur Samuel (1959):** *“field of study that gives computers the ability to learn without being explicitly programmed”.*

**Tom Mitchell (1998):** *"A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$  if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ ."*

# History

- 1950 — Alan Turing creates the “Turing Test” to determine if a computer has real intelligence.
- 1952 — Arthur Samuel wrote the first computer learning program.
- 1957 — Frank Rosenblatt designed the first neural network for computers (the perceptron).
- 1979 — Students at Stanford University invent the “Stanford Cart” which can navigate obstacles in a room.
- 1981 — Gerald Dejong introduces the concept of Explanation Based Learning (EBL), in which a computer analyses training data and creates a general rule it can follow by discarding unimportant data.
- 1985 — Terry Sejnowski invents NetTalk, which learns to pronounce words the same way a baby does.
- 1990s — Work on machine learning shifts from a knowledge-driven approach to a data-driven approach.
- 1997 — IBM’s Deep Blue beats the world champion at chess.
- 2006 — Geoffrey Hinton coins the term “deep learning” to explain new algorithms that let computers “see” and distinguish objects and text in images and videos.
- 2010 — The Microsoft Kinect can track 20 human features at a rate of 30 times per second, allowing people to interact with the computer via movements and gestures.

...

# Machine Learning

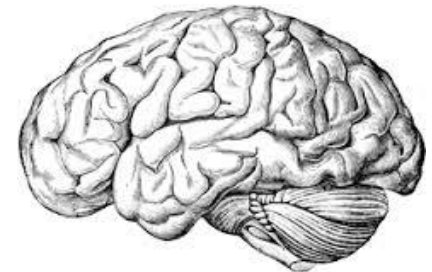
Raw Data



Features



Models



# Algorithms

- Supervised learning
  - find a function that describes labeled training data.
- Unsupervised learning
  - find correlation between "unlabeled" data.
- Reinforced learning
  - reward approach.



# Supervised

Regression

vs

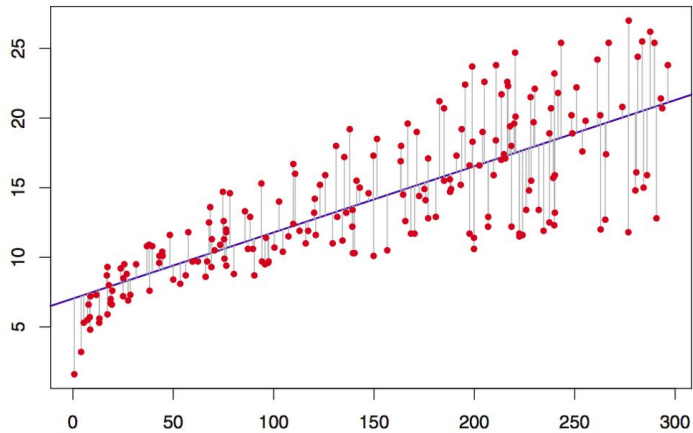
Classification

# Supervised

Regression

vs

Classification

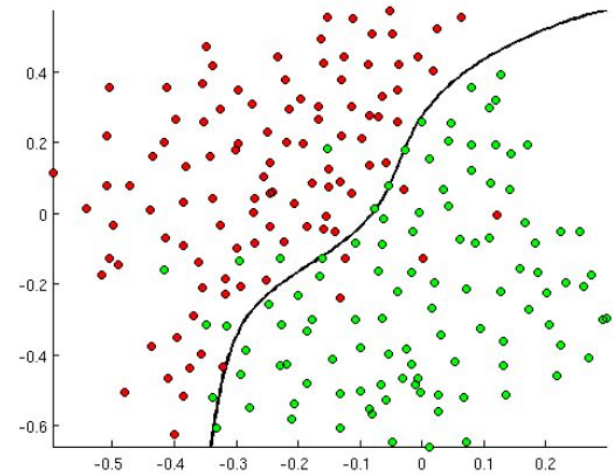
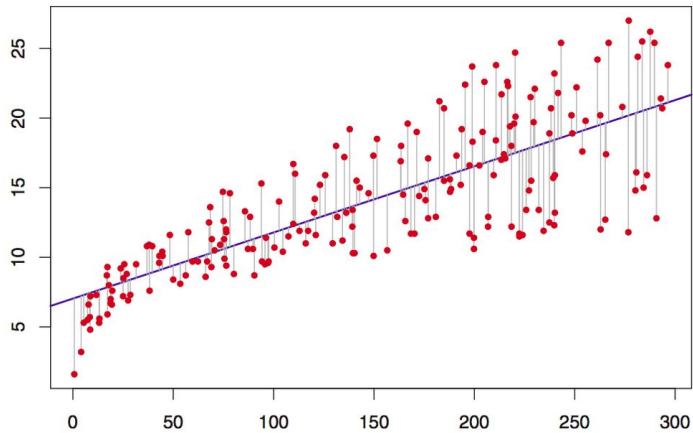


# Supervised

Regression

vs

Classification

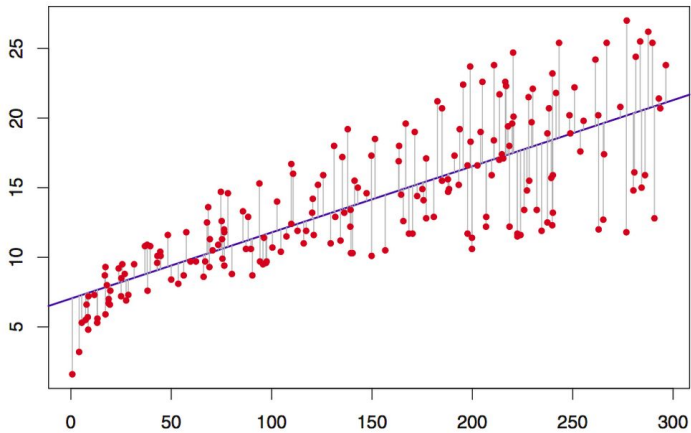


# Supervised

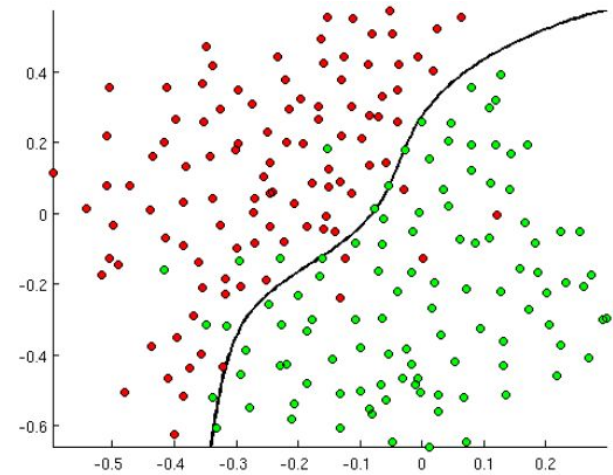
Regression

vs

Classification



*Continuous values*



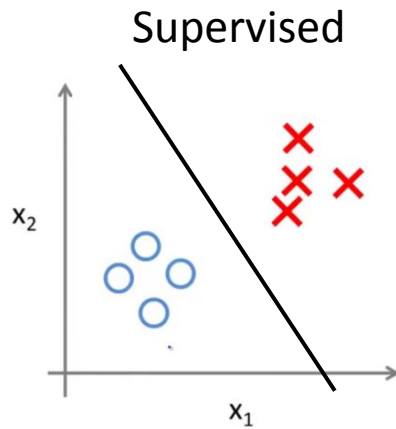
*Discrete values (labels, categories)*

# Supervised

## Use cases:

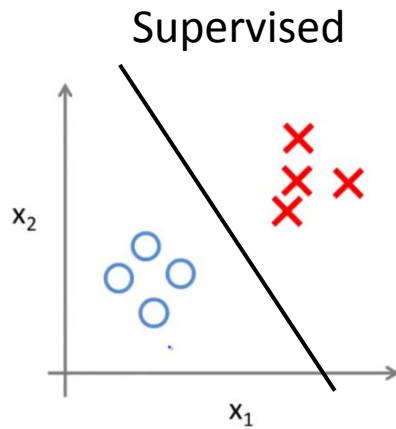
- Diagnosis of disease, anomaly detection, forecasting ...

# Unsupervised

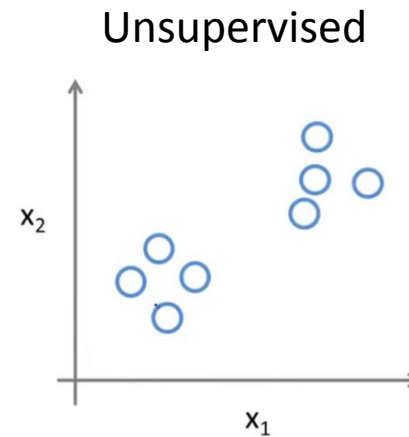


Known labels

# Unsupervised

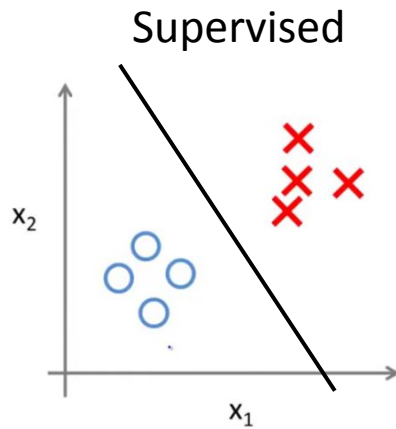


Known labels

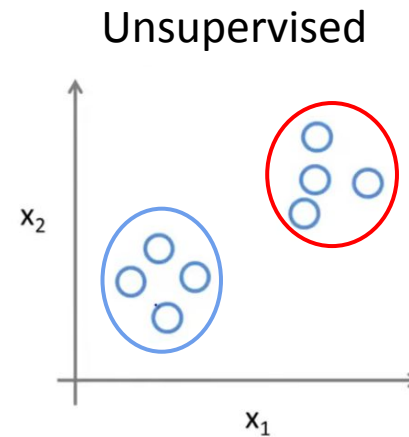


Unknown labels

# Unsupervised



Known labels

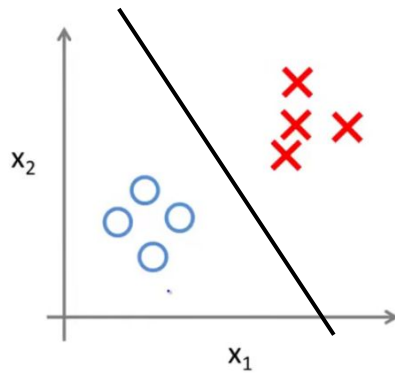


Unknown labels



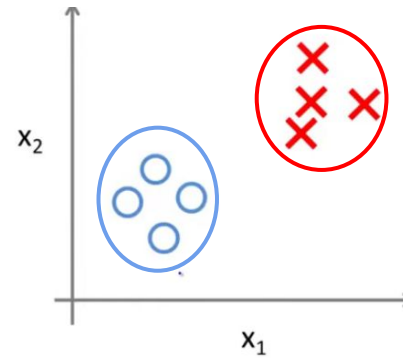
# Unsupervised

Supervised



Known labels

Unsupervised



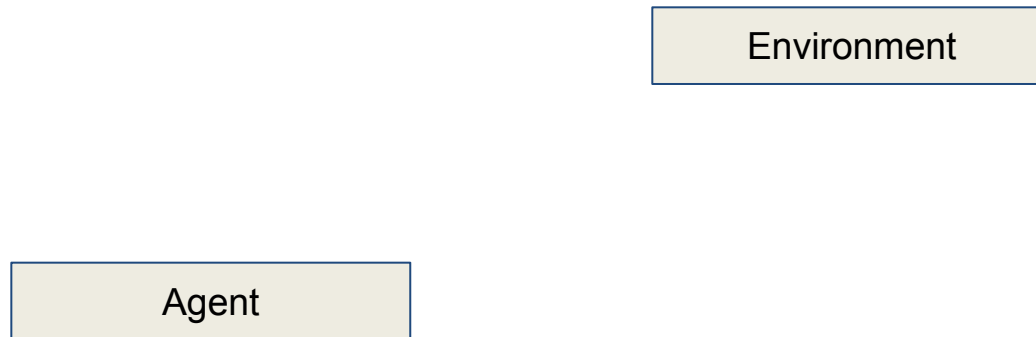
Unknown labels

# Unsupervised

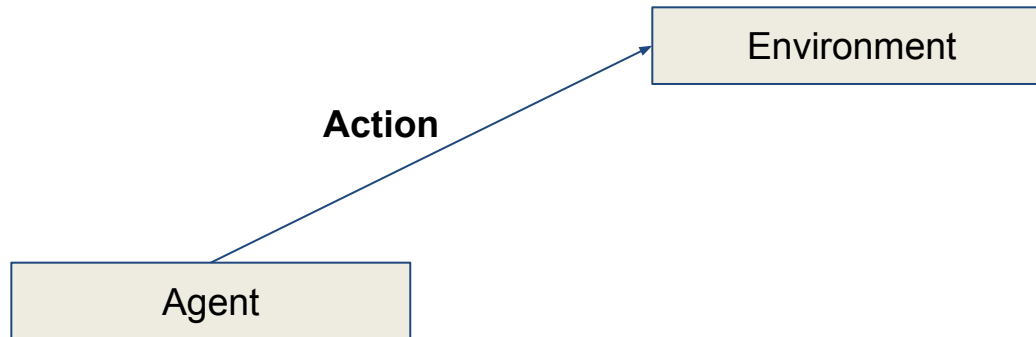
## Use cases:

- Market segmentation, social media analysis (behavior), organize computing clusters ...

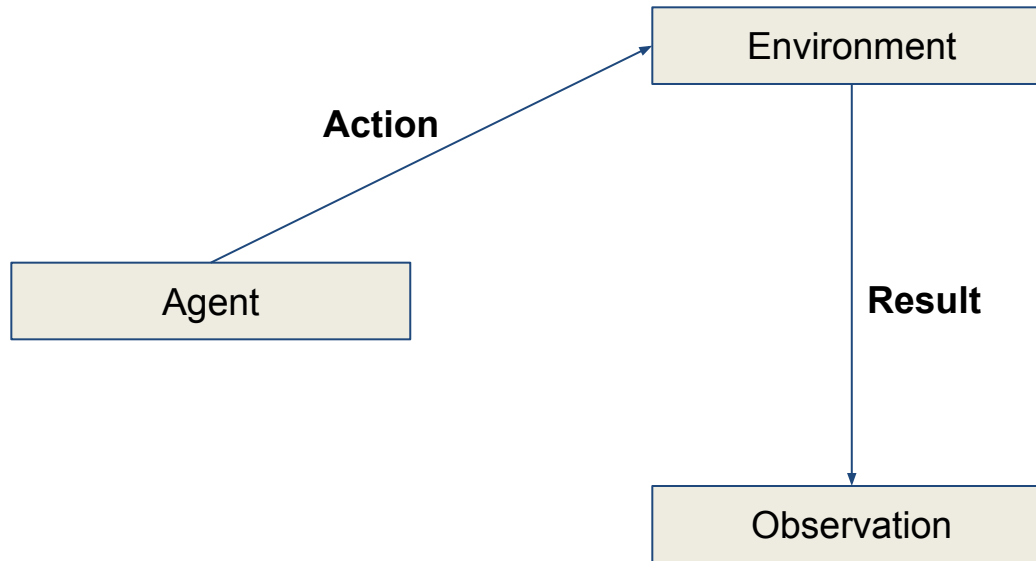
# Reinforced



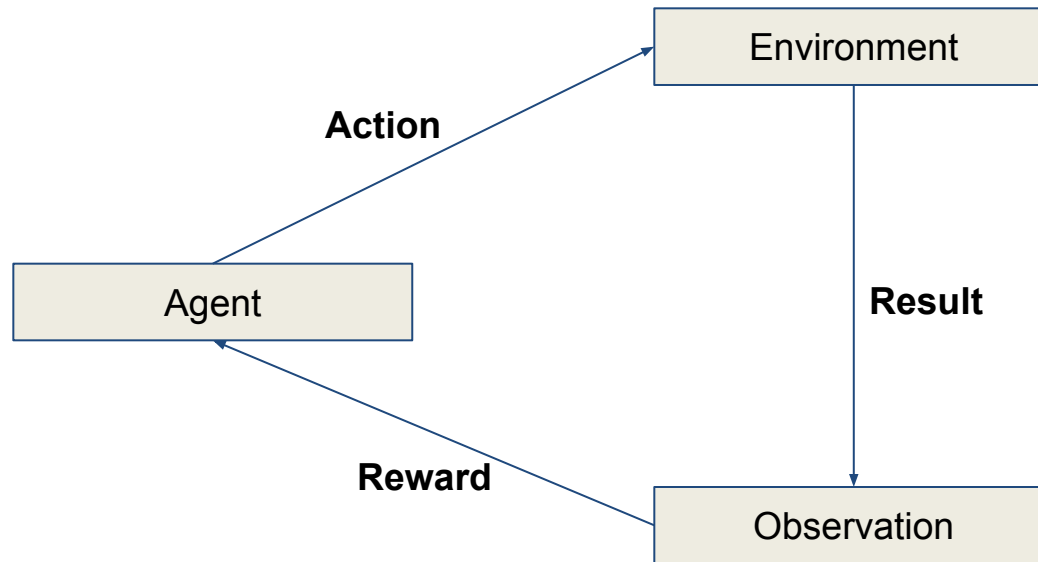
# Reinforced



# Reinforced



# Reinforced

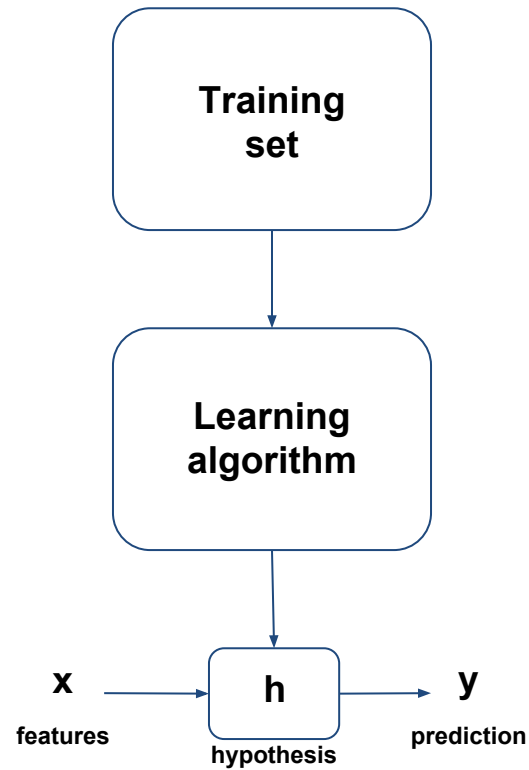


# Reinforced

Use cases:

- Trading strategy, manufacturing, game playing ...

# Workflow

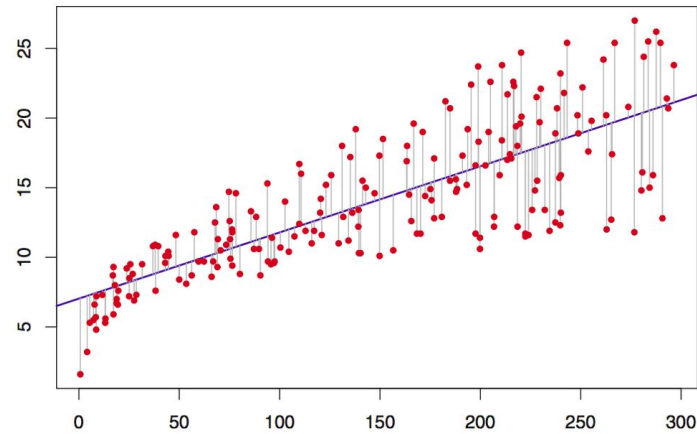




# Regression

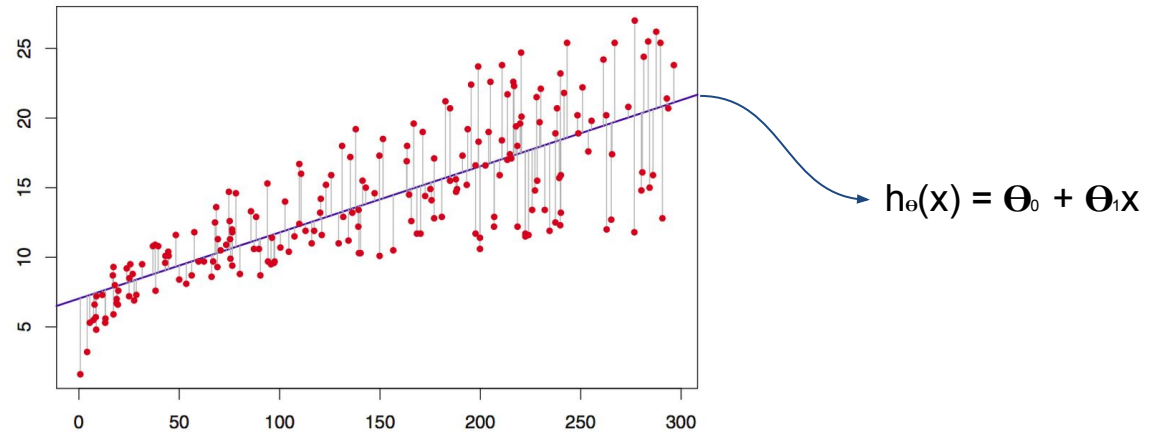
# Regression

How to represent 'h' (hypothesis)



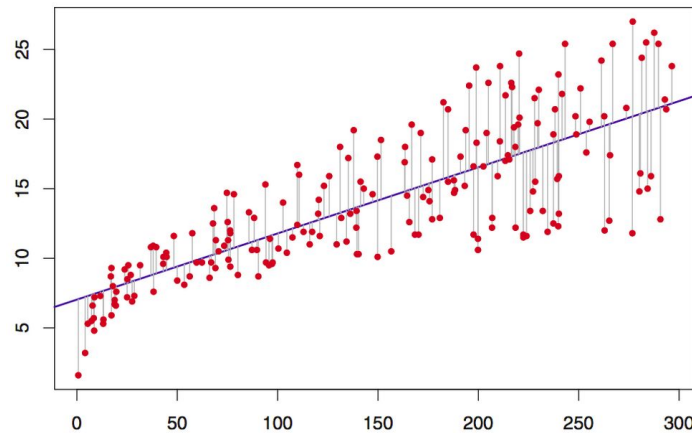
# Regression

How to represent 'h' (hypothesis)



# Regression

How to represent 'h' (hypothesis)



$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

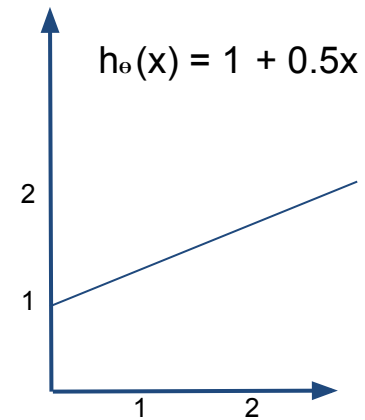
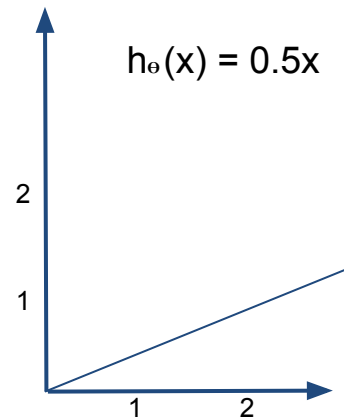
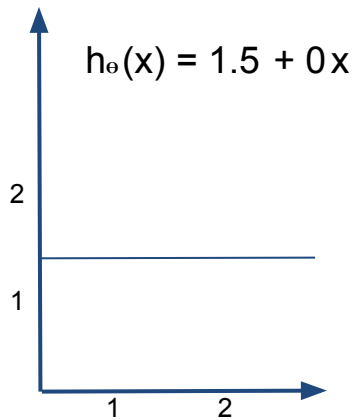
$$y = ax + b$$

For example, housing price.

2014	2015	2016	2017	2018
~180k	~182k	~184k	~186k	???

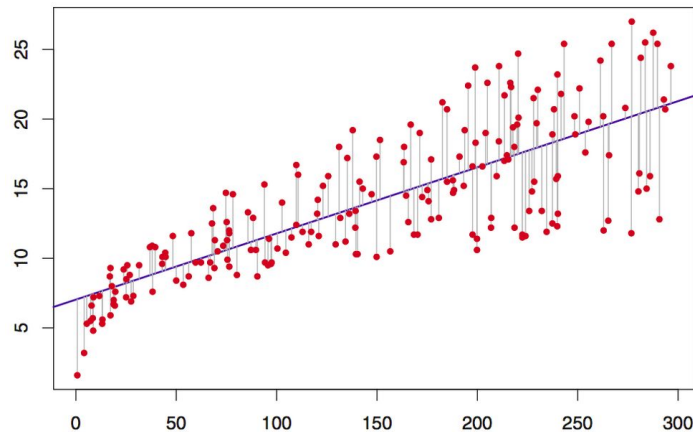
# Regression

How to represent 'h' (hypothesis)



# Regression

## How to represent 'h' (hypothesis)

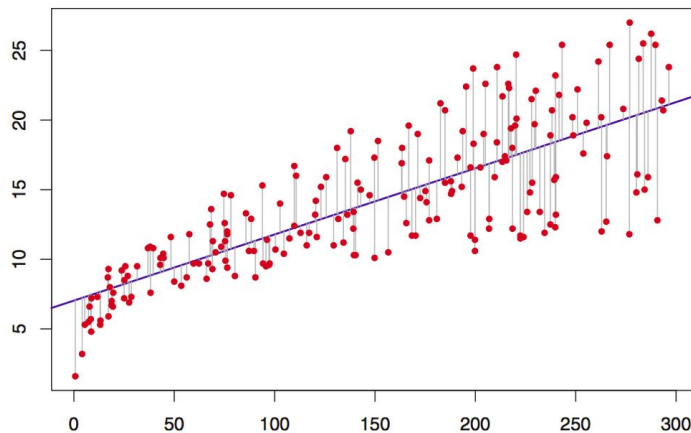


$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

Choose  $\theta_0, \theta_1$  so that  $h_{\theta}(x)$  is close to  $y$  for our training set

# Regression

## How to represent 'h' (hypothesis)



$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

Choose  $\theta_0, \theta_1$  so that  $h_{\theta}(x)$  is close to  $y$  for our training set

The idea is to minimize  $\theta_0, \theta_1$ , so that  $h_{\theta}(x) - y$  tends to decrease.

Thus, we can define the cost function  $J(\theta_0, \theta_1)$  aiming to minimize  $\theta_0, \theta_1$ :

$$J(\theta) = \frac{1}{2} \sum (h_{\theta}(x) - y)^2$$

*square difference*

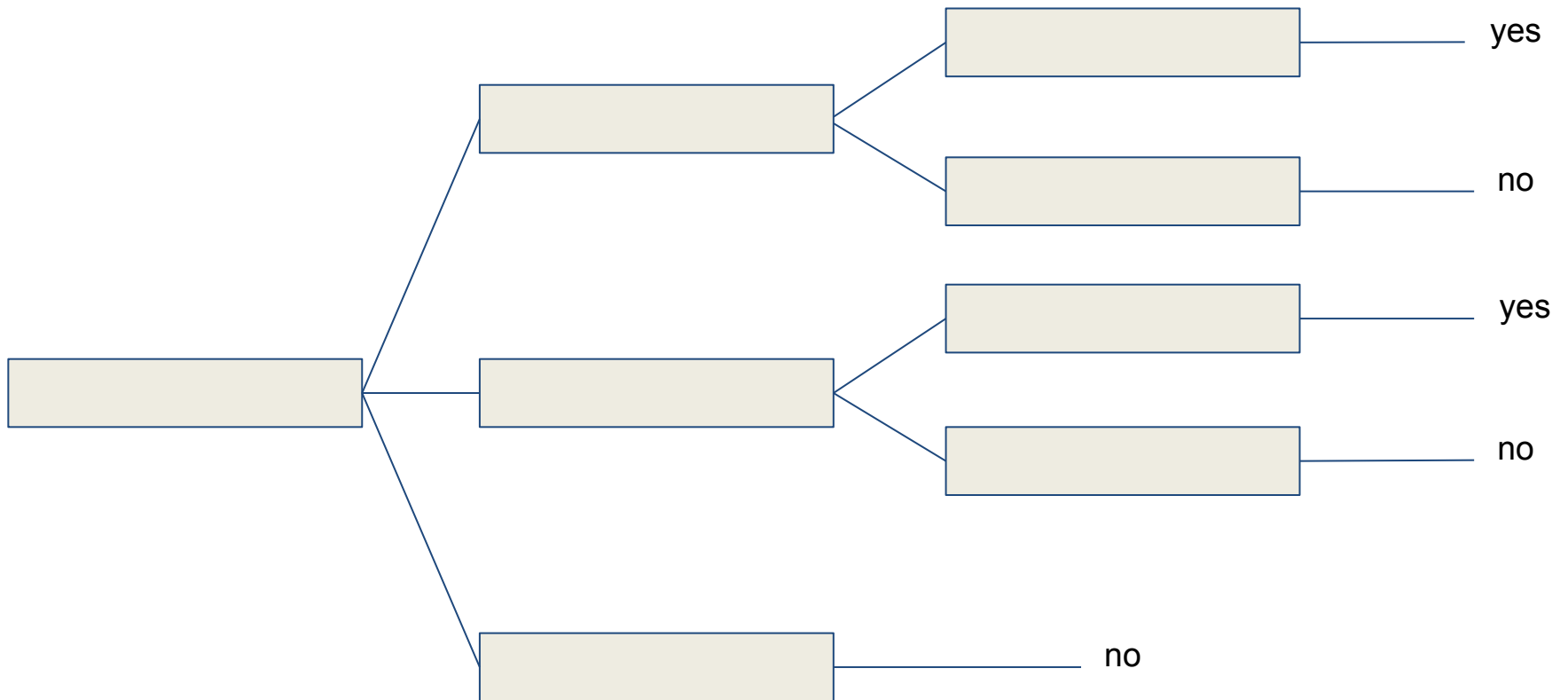
# Classification



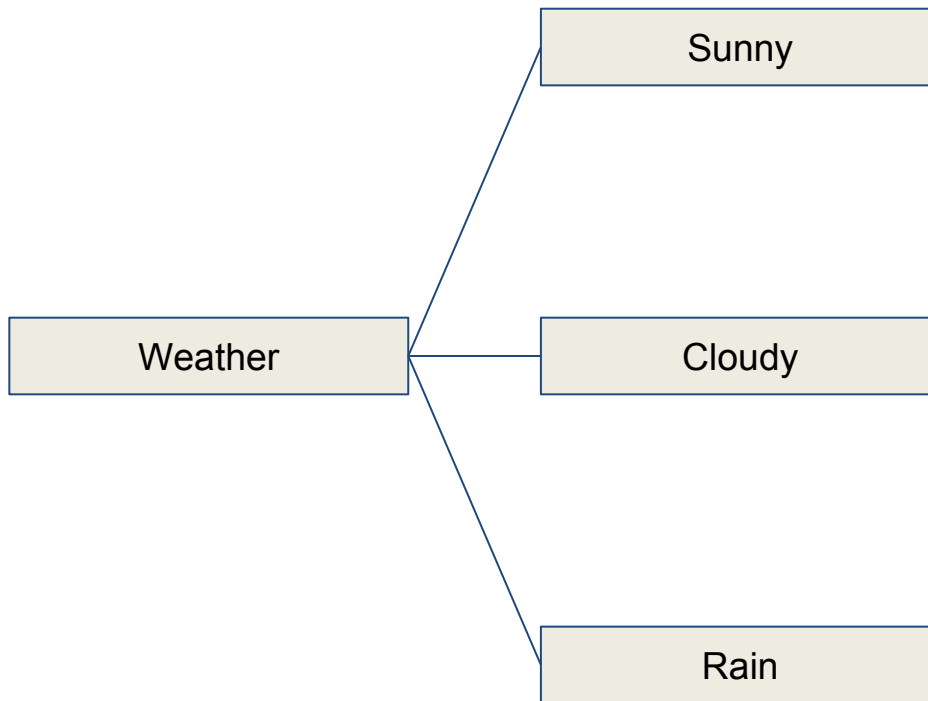
# Classification

- Decision Tree
  - The target is to separate the dataset into classes, for example, 'yes' or 'no'.

# Decision Tree

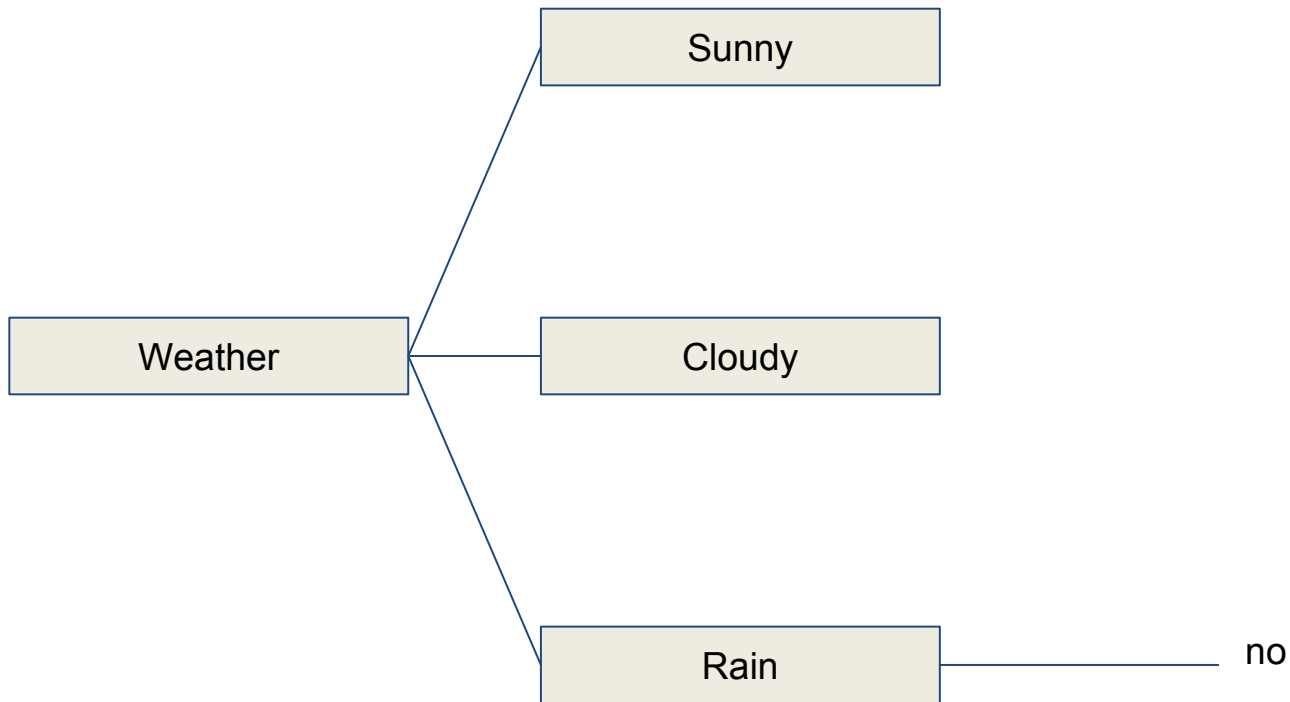


# Decision Tree



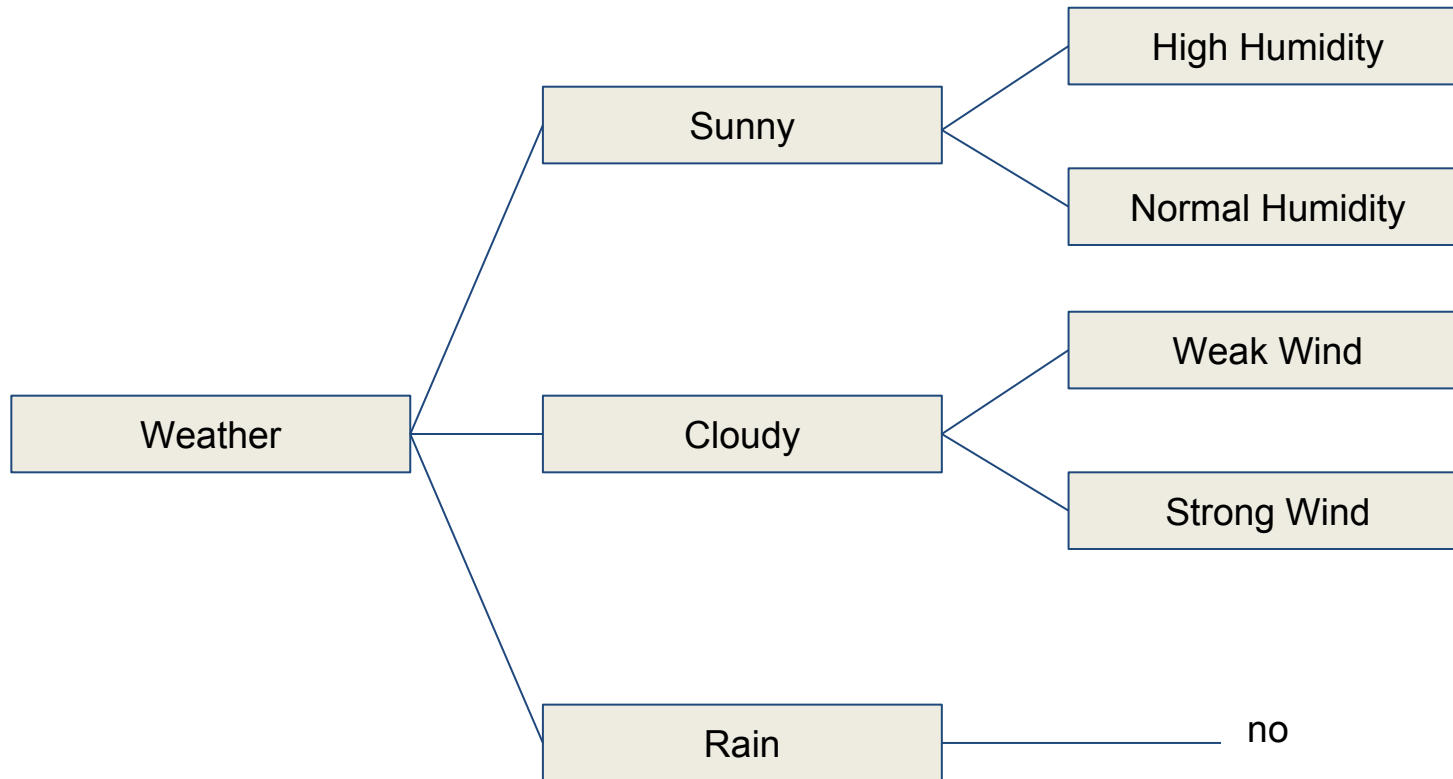
For example, play or not play football?

# Decision Tree



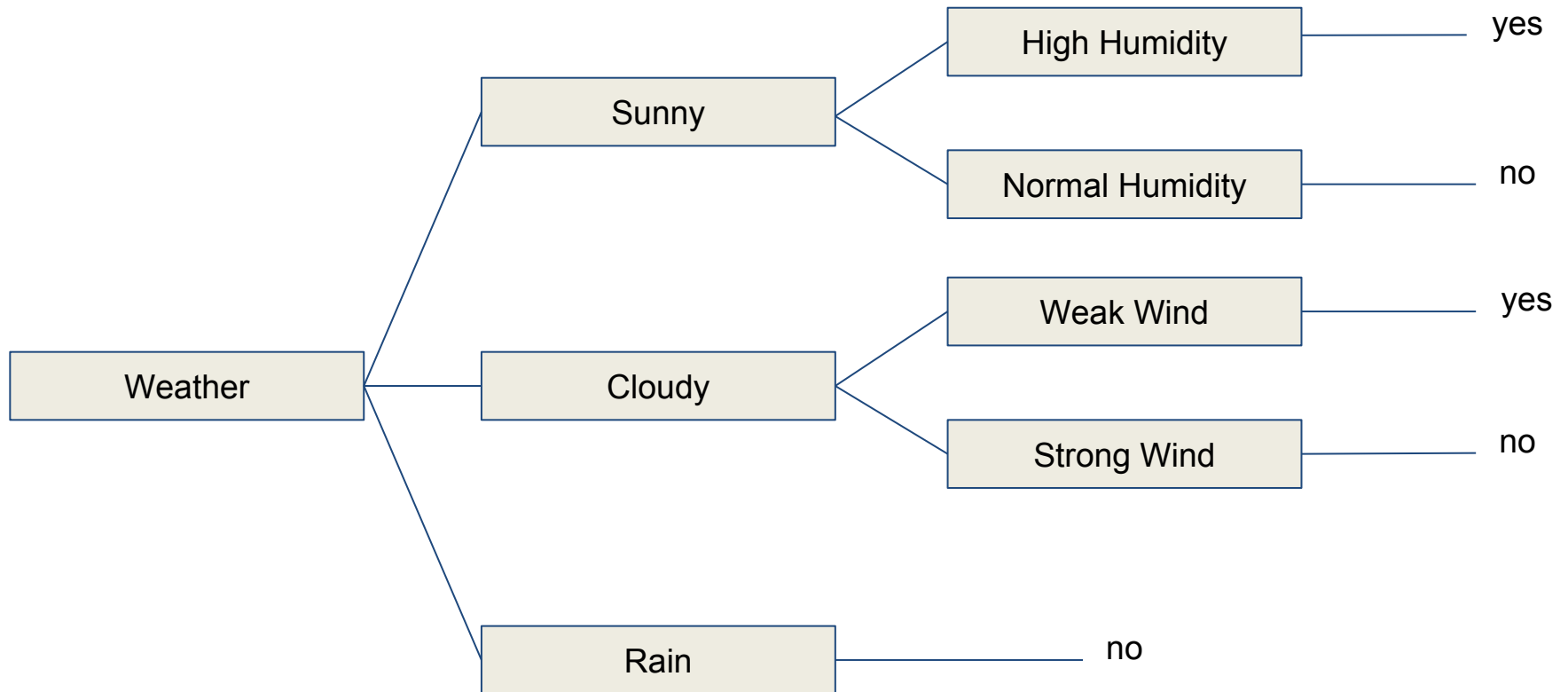
For example, play or not play football?

# Decision Tree



For example, play or not play football?

# Decision Tree

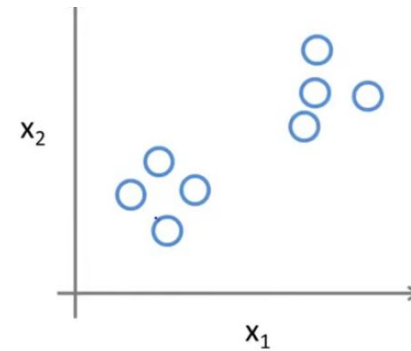


For example, play or not play football?

# Unsupervised (clustering)

## K-means algorithm

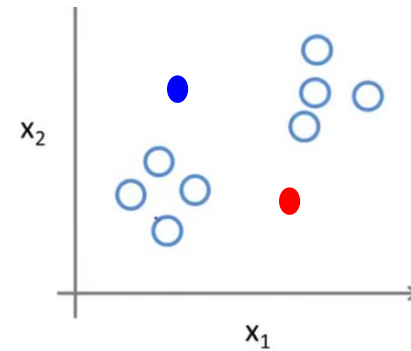
1: Define  $K$  centroids randomly.



# Clustering

## K-means algorithm

- 1: Define **K** centroids randomly.
- 2: Associate every observation according to the nearest centroid.

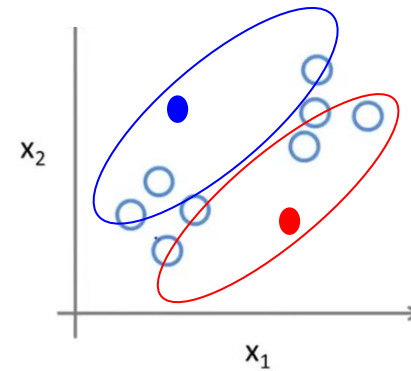




# Clustering

## K-means algorithm

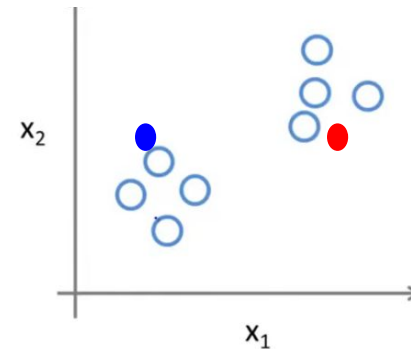
- 1: Define **K** centroids randomly.
- 2: Associate every observation according to the nearest centroid.



# Clustering

## K-means algorithm

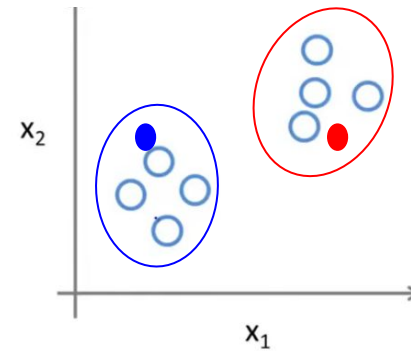
- 1: Define **K** centroids randomly.
- 2: Associate every observation according to the nearest centroid.
- 3: Define new centroids according to the mean of the clusters.



# Clustering

## K-means algorithm

- 1: Define **K** centroids randomly.
- 2: Associate every observation according to the nearest centroid.
- 3: Define new centroids according to the mean of the clusters.
- 4: Repeat step 2 and 3 to converge.



# Reinforced

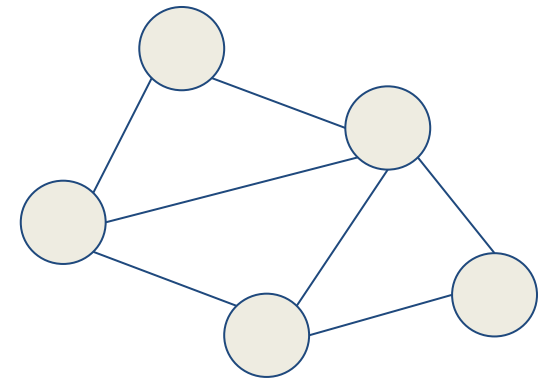
- Markov decision process
  - In short, the target is to maximize the rewards:

$E(r | \pi, s)$ , where:

- Set of states,  $S$
- Set of actions,  $A$
- Reward function,  $R$
- Policy,  $\pi$
- Value,  $V$

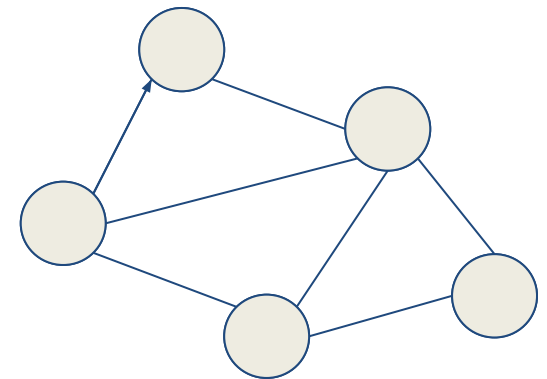
# Reinforced

1. The agent observes an input state



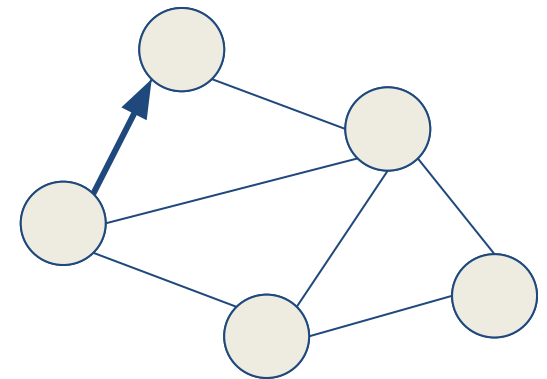
# Reinforced

1. The agent observes an input state
2. An action is determined by a decision making function (policy)



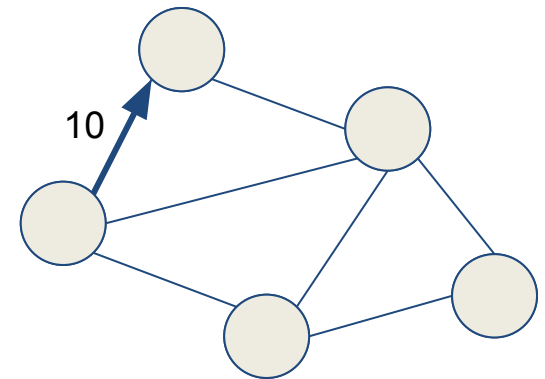
# Reinforced

1. The agent observes an input state
2. An action is determined by a decision making function (policy)
3. The action is performed



# Reinforced

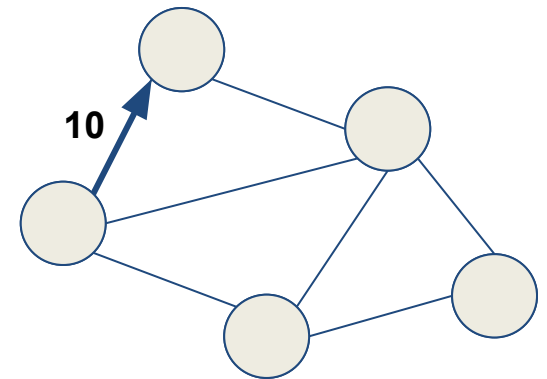
1. The agent observes an input state
2. An action is determined by a decision making function (policy)
3. The action is performed
4. The agent receives a scalar reward or reinforcement from the environment



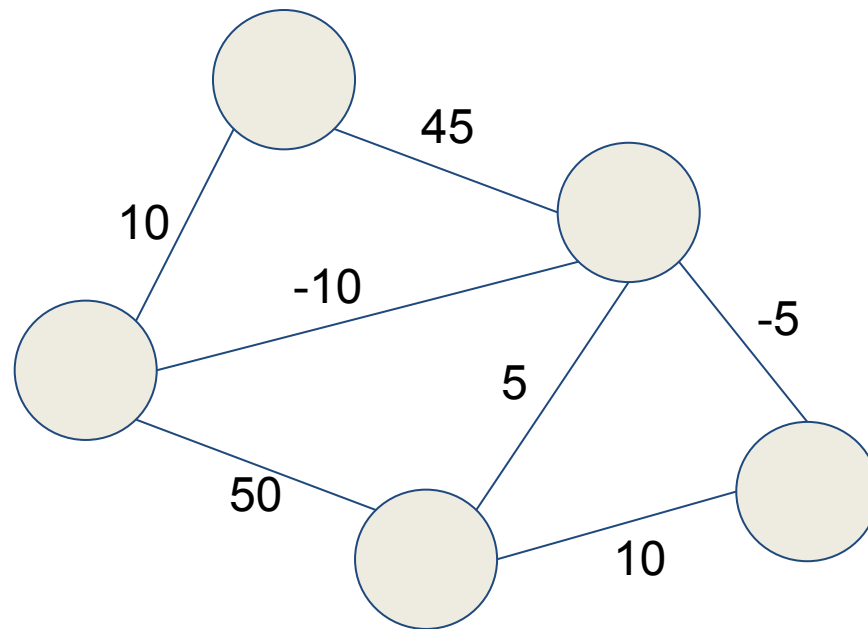


# Reinforced

1. The agent observes an input state
2. An action is determined by a decision making function (policy)
3. The action is performed
4. The agent receives a scalar reward or reinforcement from the environment
5. Information about the reward given for that action pair is recorded



# Reinforced



# State of the art

## **The Challenge of Non-Technical Loss Detection Using Artificial Intelligence: A Survey**

**Patrick Glauner<sup>1</sup>, Jorge Augusto Meira<sup>1</sup>, Petko Valtchev<sup>12</sup>, Radu State<sup>1</sup>, Franck Bettinger<sup>3</sup>**

*Published in:*

International Journal of Computational Intelligence Systems (IJCIS), vol. 10, issue 1, pp. 760-775, 2017.

# State of the art

- Features
- Models
- Comparison

# State of the art: features

- Monthly consumption:

- Daily averages:

$$x_d^{(m)} = \frac{L_d^{(m)}}{R_d^{(m)} - R_{d-1}^{(m)}},$$

- Monthly consumption before the inspection
- Consumption in the same month in the year before
- Consumption in the past three months

# State of the art: features

- Monthly consumption:
  - The customer's consumption over the past 24 months
  - Average consumption
  - Maximum consumption
  - Standard deviation
  - Number of inspections
  - Average consumption of the residential area

# State of the art: features

- Smart meter consumption:
  - Consumption features from intervals of 15 or 30 minutes
  - The maximum consumption in any 15-minute window
  - Load factor is computed by dividing the demand contracted by the maximum consumption
  - Shape factors are derived from the consumption time series including the impact of lunch times, nights and weekends

# State of the art: features

- Smart meter consumption:
  - $4 \times 24 = 96$  measurements are encoded to a 32-dimensional space:
    - Each measurement is 0 or positive
    - Next, it is then mapped to 0 or 1, respectively
    - Last, the 32 features are computed
    - A feature is the weighted sum of three subsequent values, in which the first value is multiplied by 4, the second by 2 and the third by 1



# State of the art: features

- Master data:
  - Location (city and neighborhood)
  - Business class (e.g. residential or industrial)
  - Activity type (e.g. residence or drugstore)
  - Voltage
  - Number of phases (1, 2 or 3)
  - Meter type

# State of the art: features

- Master data:
  - Demand contracted, i.e. the number of kW of continuous availability requested from the energy company and the total demand in kW of installed equipment of the customer
  - Information about the power transformer to which the customer is connected to
  - Town or customer in which the customer is located
  - Type of voltage (low, median or high)

# State of the art: features

- Master data:
  - Electricity tariff
  - Contracted power
  - Number of phases
  - Type of customer
  - Location
  - Voltage level
  - Type of climate (rainy or hot)
  - Weather conditions

# State of the art: features

- Credit worthiness ranking (CWR):
  - Computed from the electricity provider's billing system
  - Reflects if a customer delays or avoids payments of bills
  - CWR ranges from 0 to 5 where 5 represents the maximum score
  - It reflects different information about a customer such as payment performance, income and prosperity of the neighborhood in a single feature

# State of the art: models

- Expert systems and fuzzy systems
- Neural networks
- Support vector machines
- Genetic algorithms
- Rough sets
- Various other methods: optimum path forest, linear regression, etc.

# State of the art: comparison

- Accuracy: 
$$\frac{tp + tn}{tp + tn + fp + fn}$$
- Precision: 
$$\frac{tp}{tp + fp}$$
- Recall: 
$$\frac{tp}{tp + fn}$$
- F1: 
$$2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

# State of the art: comparison

Ref.	Model	#Customers	Accuracy	Precision	Recall	AUC	NTL/theft proportion
1	SVM (Gauss)	< 400	0.86	-	0.77	-	-
7	SVM + fuzzy	100K	-	-	0.72	-	-
16	Bool rules	700K	-	-	-	0.47	5%
16	Fuzzy rules	700K	-	-	-	0.55	5%
16	SVM (linear)	700K	-	-	-	0.55	5%
16	Bool rules	700K	-	-	-	0.48	20%
16	Fuzzy rules	700K	-	-	-	0.55	20%
16	SVM (linear)	700K	-	-	-	0.55	20%
17	SVM	< 400	-	-	0.53	-	-
18	Genetic SVM	1,171	-	-	0.62	-	-
19	Neuro-fuzzy	20K	0.68	0.51	-	-	-
22	NN	22K	0.87	0.65	0.29	-	-
23	Rough sets	N/A	0.93	-	-	-	-
24	SOM	2K	0.93	0.85	0.98	-	-

# State of the art: comparison

25	SVM (Gauss)	1,350	0.98	-	-	-	-
27	Regression	30	-	-	0.22	-	1%
27	Regression	30	-	-	0.78	-	2%
27	Regression	30	-	-	0.98	-	3%
27	Regression	30	-	-	1	-	4-10%
29	SVM	5K	0.96	-	-	-	-
29	KNN	5K	0.96	-	-	-	-
29	NN	5K	0.94	-	-	-	-
30	OPF	736	0.90	-	-	-	-
30	SVM (Gauss)	736	0.89	-	-	-	-
30	SVM (linear)	736	0.45	-	-	-	-
30	NN	736	0.53	-	-	-	-
33	Decision tree	N/A	0.99	-	-	-	-



# Challenges

- Class imbalance and evaluation metric
- Feature description
- Data quality
- Covariate shift
- Scalability
- Comparison of different methods

# Class imbalance, evaluation metric

- Imbalanced classes appear frequently in machine learning, which also affects the choice of evaluation metrics.
- Most NTL detection research do not address this property.

# Class imbalance, evaluation metric

- In many papers, high accuracies or high recalls are reported:

Ref.	Model	#Customers	Accuracy	Precision	Recall
1	SVM (Gauss)	< 400	0.86	-	0.77
7	SVM + fuzzy	100K	-	-	0.72

- The following examples demonstrate why those performance measures are not suitable for NTL detection in imbalanced data sets.

# Class imbalance, evaluation metric

- For a test set containing 1K customers of which 999 have regular use,
  - A classifier always predicting non-NTL has an accuracy of 99.9%
  - While this classifier has a very high accuracy and intuitively seems to perform very well, it will never predict any NTL.

# Class imbalance, evaluation metric

- For a test set containing 1K customers of which 999 have regular use,
  - A classifier always predicting NTL has a recall of 100%.
  - While this classifier will find all NTL, it triggers many costly and unnecessary physical inspections by inspecting all customers.

# Class imbalance, evaluation metric

- This topic is addressed rarely in NTL literature.
- For NTL detection, the goal is to reduce the false positive rate (FPR) to decrease the number of costly inspections, while increasing the true positive rate (TPR) to find as many NTL occurrences as possible.

# Class imbalance, evaluation metric

- We propose to use a receiver operating characteristic (ROC) curve, which plots the TPR against the FPR.
- The area under the curve (AUC) is a performance measure between 0 and 1, where any binary classifier with an  $AUC > 0.5$  performs better than chance

# Class imbalance, evaluation metric

## Large-Scale Detection of Non-Technical Losses in Imbalanced Data Sets

Patrick Glauner\*, Andre Boechat\*, Lautaro Dolberg\*, Radu State\*, Franck Bettinger†, Yves Rangoni†  
and Diogo Duarte†

*Published in:*

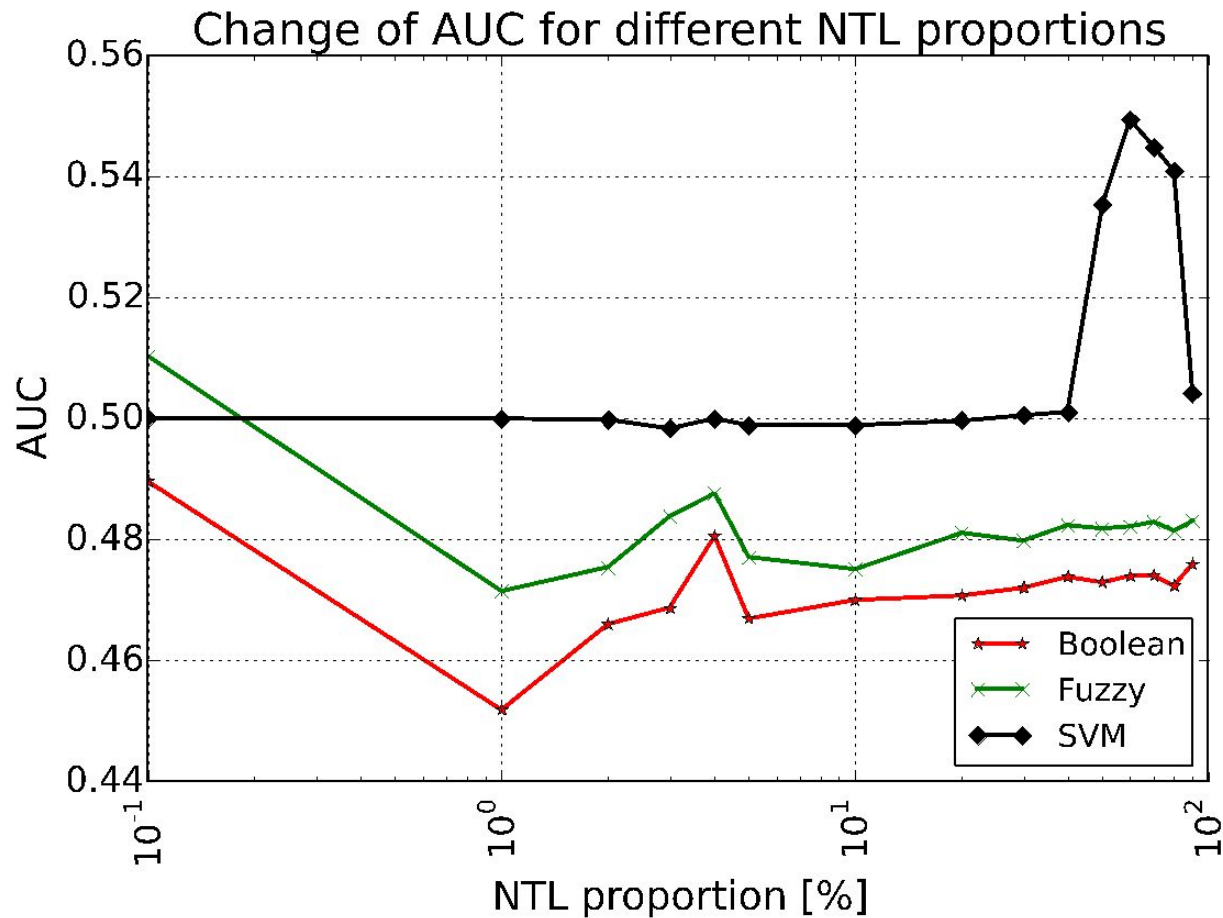
Proceedings of the Seventh IEEE Conference on Innovative Smart Grid Technologies (ISGT 2016), Minneapolis, USA, 2016.



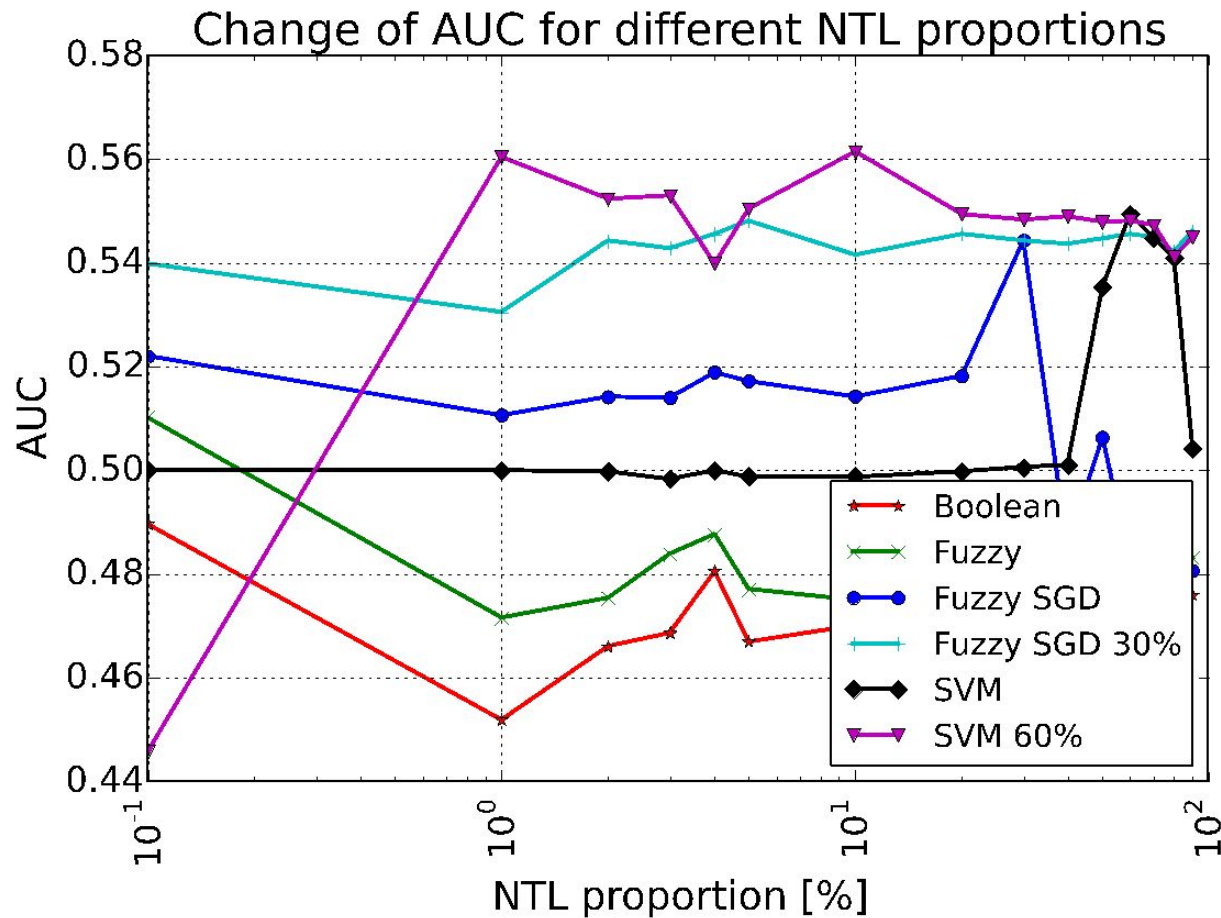
# Class imbalance, evaluation metric

Ref.	Model	#Customers	Accuracy	Precision	Recall	AUC	NTL/theft proportion
1	SVM (Gauss)	< 400	0.86	-	0.77	-	-
7	SVM + fuzzy	100K	-	-	0.72	-	-
16	Bool rules	700K	-	-	-	0.47	5%
16	Fuzzy rules	700K	-	-	-	0.55	5%
16	SVM (linear)	700K	-	-	-	0.55	5%
16	Bool rules	700K	-	-	-	0.48	20%
16	Fuzzy rules	700K	-	-	-	0.55	20%
16	SVM (linear)	700K	-	-	-	0.55	20%

# Class imbalance, evaluation metric



# Class imbalance, evaluation metric



# Feature description

- Generally, hand-crafting features from raw data is a long-standing issue in machine learning having significant impact on the performance of a classifier.
- Different feature description methods have been reviewed in the previous section.

# Feature description

- They fall into two main categories:
  - Features computed from the consumption profile of customers which are from:
    - Monthly meter readings
    - Or smart meter readings
  - And features from the customer master data.

# Feature description

- The features computed from the time series are very different for monthly meter readings and smart meter readings.
- The results of those works are not easily interchangeable. While electricity providers continuously upgrade their infrastructure to smart metering, there will be many remaining traditional meters. In particular, this applies to emerging countries.

# Feature description

- There are only few works on assessing the statistical usefulness of features for NTL detection.
- Almost all works on NTL detection define features and subsequently report improved models that were mostly found experimentally without having a strong theoretical foundation.

# Data quality

- We noticed that the inspection result labels in the training set are not always correct and that some fraudsters may be labelled as non-fraudulent.
- The reasons for this may include bribing, blackmailing or threatening of the technician performing the inspection.
- Also, the fraud may be done too well and is therefore not observable by technicians.



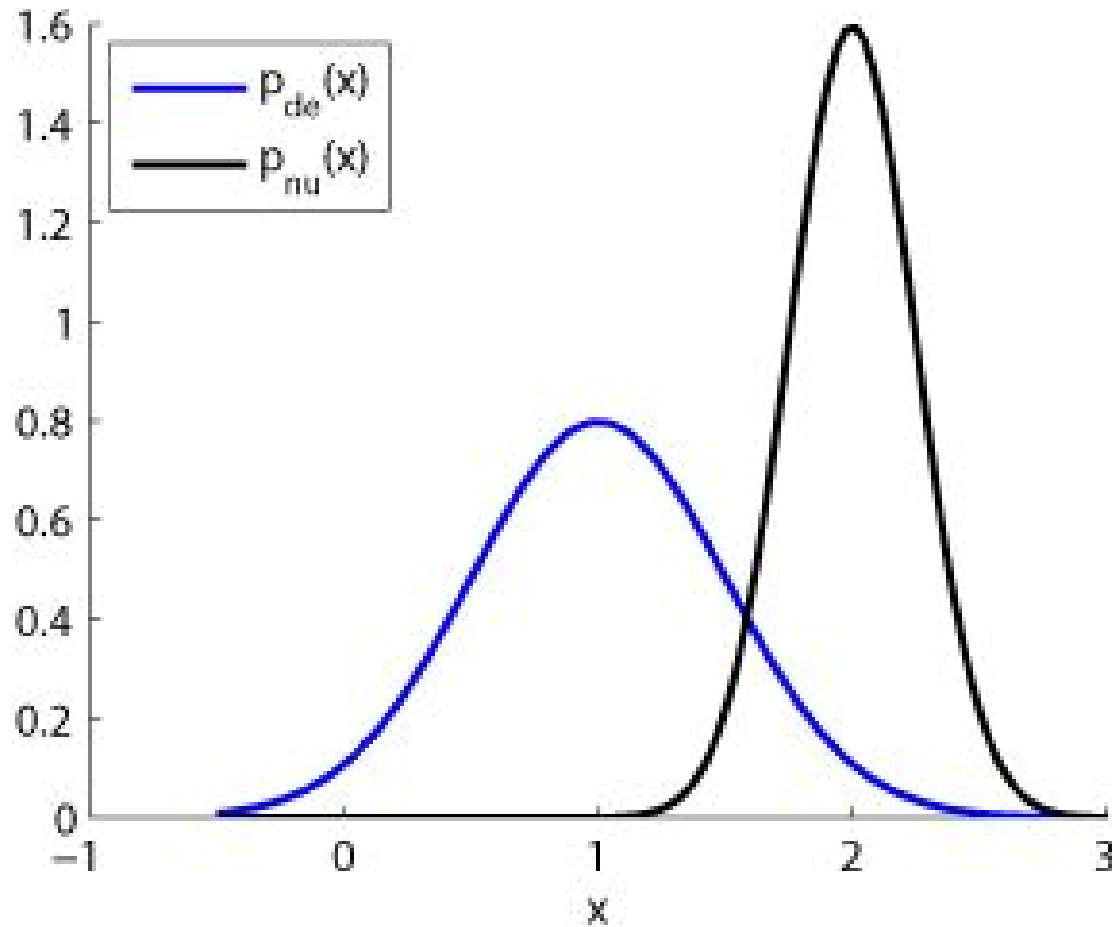
# Data quality

- Another reason may be incorrect processing of the data. It must be noted that the latter reason may, however, also label non-fraudulent behavior as fraudulent.
- Most NTL detection research use supervised methods. This shortcoming of the training data and potential wrong labels in particular are only rarely reported in the literature.

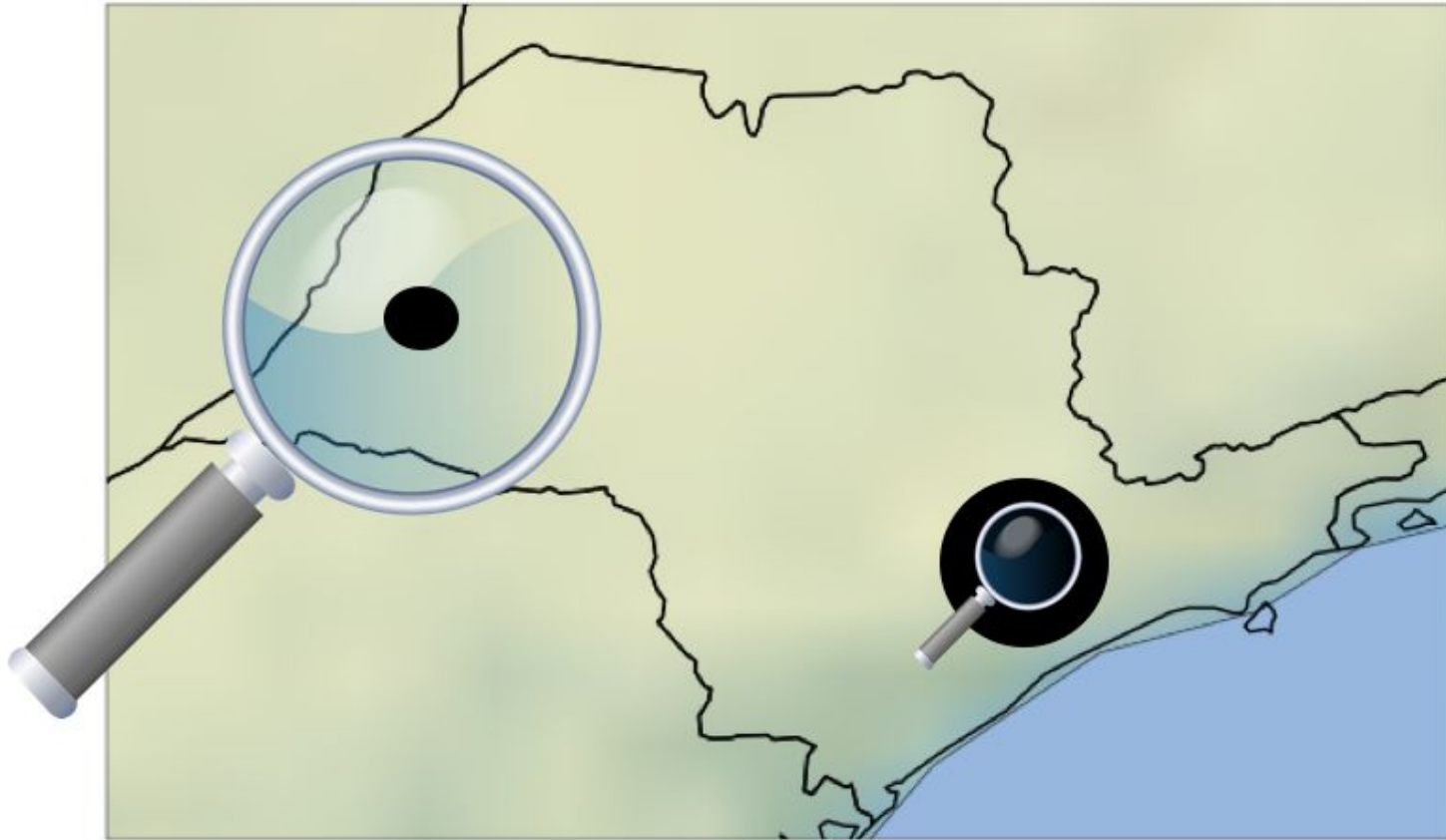
# Covariate shift

- Covariate shift refers to the problem of training data (i.e. the set of inspection results) and production data (i.e. the set of customers to generate inspections for) having different distributions.

# Covariate shift

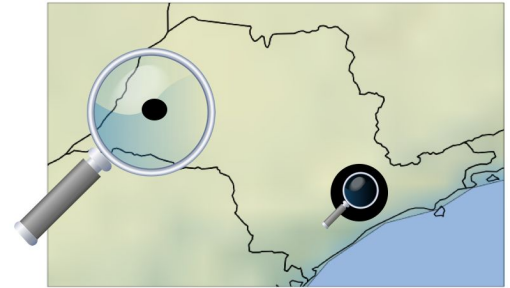


# Covariate shift



# Covariate shift

The large city is close to the sea, whereas the small city is located in the interior of the country. The weather in the small city undergoes stronger changes during the year. The subsequent change of electricity consumption during the year triggers many inspections. As a consequence, most inspections are carried out in the small city. Therefore, the sample of customers inspected does not represent the overall population of customers.



# Covariate shift

- This fact leads to unreliable NTL predictors when learning from this training data.
- Historically, covariate shift has been a long-standing issue in statistics

# Covariate shift

- The Literary Digest sent out 10M questionnaires in order to predict the outcome of the 1936 US Presidential election.
- They received 2.4M returns.
- Predicted Alfred Landon to win.



# Covariate shift

- Nonetheless, the predicted result proved to be wrong.
- The reason for this was that they used car registrations and phone directories to compile a list of recipients.
- In that time, the households that had a phone or a car represented a biased sample of the overall population.



# Covariate shift

- In contrast, George Gallup only interviewed 3K handpicked people, which were an unbiased sample of the population.
- As a consequence, Gallup could predict the outcome of the election very well.

# Covariate shift

## Is Big Data Sufficient for a Reliable Detection of Non-Technical Losses?

Patrick Glauner\*, Angelo Migliosi\*, Jorge Augusto Meira\*, Petko Valtchev\*<sup>†</sup>, Radu State\* and Franck Bettinger<sup>‡</sup>

*Published in:*

Proceedings of the 19th International Conference on Intelligent System Applications to Power Systems (ISAP 2017), San Antonio, USA, 2017.

# Covariate shift

We propose a robust algorithm for measuring covariate shift in data sets.

---

## Algorithm 1 Quantifying covariate shift.

---

```

1: result  $\leftarrow$  0
2: reliability  $\leftarrow$  0
3: selected  $\leftarrow$  train_data.add_feature(s, 1)
4: not_selected  $\leftarrow$  prod_data.add_feature(s, 0)
5: data  $\leftarrow$  selected  $\cup$  not_selected
6: folds  $\leftarrow$  cv_folds(data, k)
7: for model in get_model_candidates() do
8:   mccs  $\leftarrow$  list()
9:   for fold in folds do
10:     Xtrain, Xtest, ytrain, ytest  $\leftarrow$  fold
11:     classifier  $\leftarrow$  DecisionTree(model)
12:     classifier.train(Xtrain, ytrain)
13:     ypred  $\leftarrow$  classifier.predict(Xtest)
14:     mccs.append(MCC(ytest, ypred))
15:   end for
16:   mcc_mean  $\leftarrow$  mean(mccs)
17:   if mcc_mean > result then
18:     result  $\leftarrow$  mcc_mean
19:     reliability  $\leftarrow$  std(mccs)
20:   end if
21: end for
22: return result, reliability

```

---

# Covariate shift

## ASSESSED FEATURES.

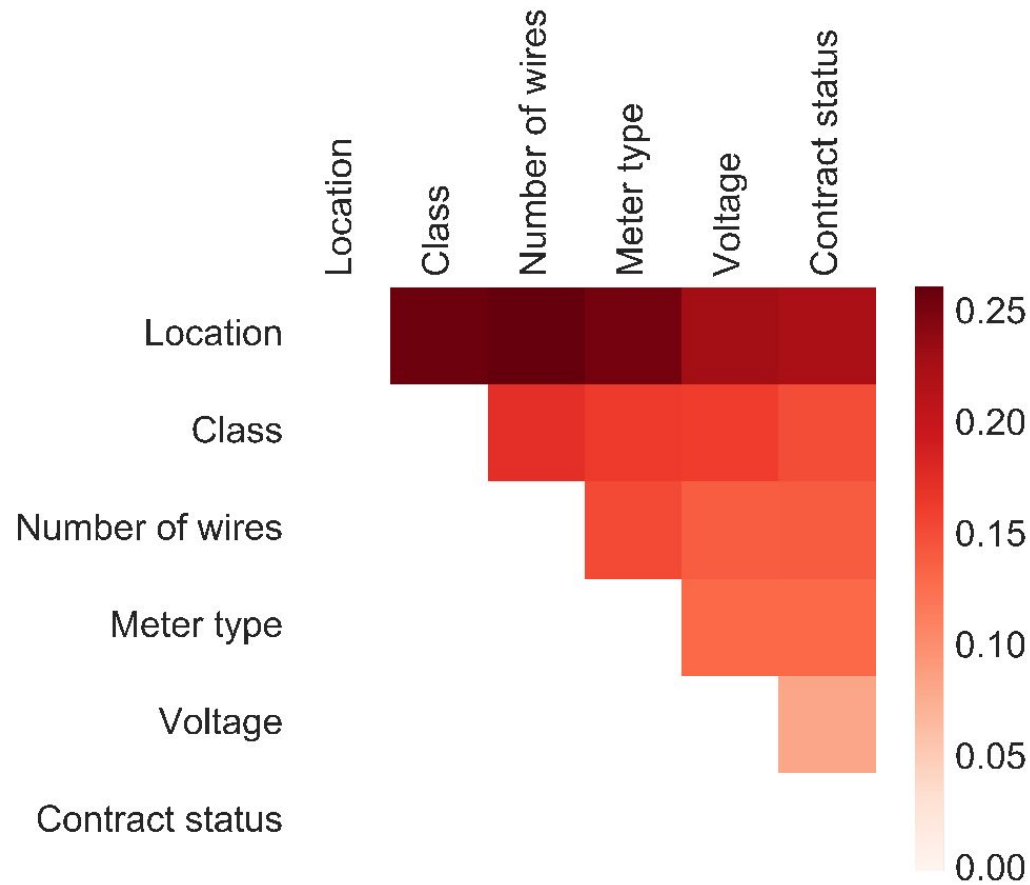
Feature	Possible values
Class	Power generation infrastructure, residential, commercial, industrial, public, public illumination, rural, public service, reseller
Contract status	Active, suspended
Location	Longitude and latitude
Meter type	22 different meter types
Number of wires	1, 2, 3
Voltage	$\leq 2.3\text{kV}$ , $> 2.3\text{kV}$

# Covariate shift

GLOBAL COVARIATE SHIFT OF SINGLE FEATURES.

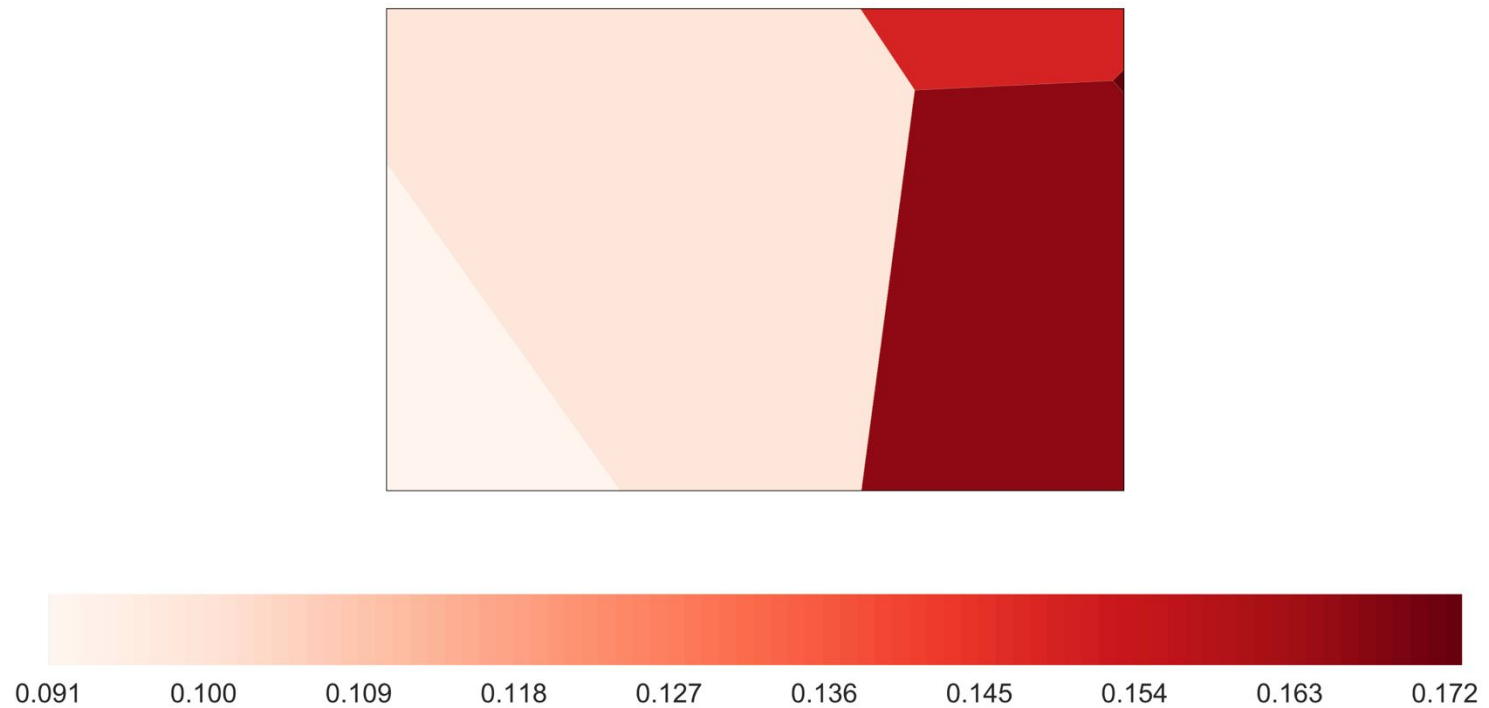
Feature	$\overline{MCC}_{max}$	$\sigma$
Location	<b>0.22367</b>	0.03453
Class	0.16255	0.01371
Number of wires	0.14111	0.00794
Meter type	0.13158	0.00382
Voltage	0.07092	0.02375
Contract status	0.03744	<b>0.09183</b>

# Covariate shift



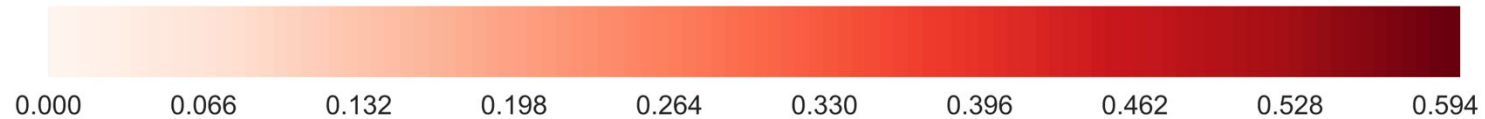
# Covariate shift

Regional level:



# Covariate shift

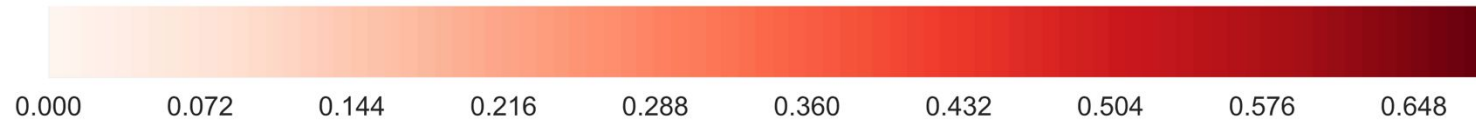
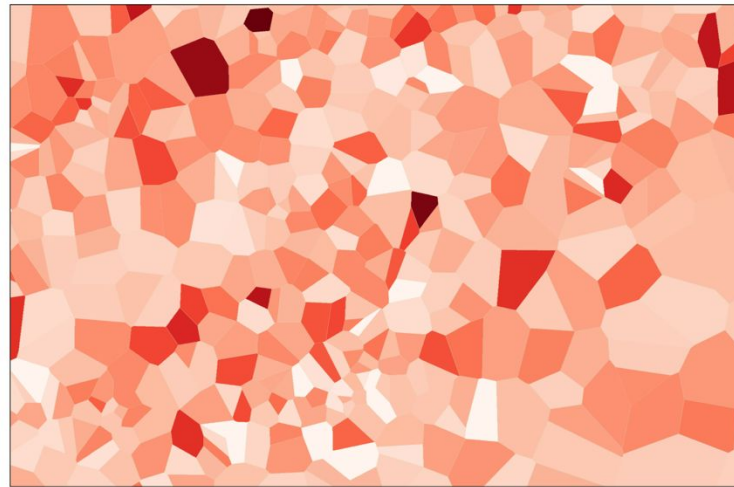
Municipal level:





# Covariate shift

Local level:



# Covariate shift

- We are currently working on reducing covariate shift.
- We aim at detecting NTL more reliably.

# Scalability

- The number of customers used throughout the research reviewed significantly varies.
- Some papers use less than a few hundred customers in the training.
- Some papers use SVMs with a Gaussian kernels. In that setting, training is only feasible in a realistic amount of time for up to a couple of tens of thousands of customers in current implementations.

# Scalability

- Another paper uses the Moore-Penrose pseudoinverse. This model is also only able to scale to up to a couple of tens of thousands of customers.

$$\hat{R} = (H^T H)^{-1} H^T L \quad (4)$$

where

$$H = \begin{bmatrix} \frac{I_1(t_2)^3 - I_1(t_1)^3}{3s_{1,2}} & \frac{I_2(t_2)^3 - I_2(t_1)^3}{3s_{2,2}} & \dots & \frac{I_n(t_2)^3 - I_n(t_1)^3}{3s_{n,2}} & 1 \\ \frac{I_1(t_3)^3 - I_1(t_2)^3}{3s_{1,3}} & \frac{I_2(t_3)^3 - I_2(t_2)^3}{3s_{2,3}} & \dots & \frac{I_n(t_3)^3 - I_n(t_2)^3}{3s_{n,3}} & 1 \\ \dots & \dots & \dots & \dots & 1 \\ \frac{I_1(t_m)^3 - I_1(t_{m-1})^3}{3s_{1,m}} & \frac{I_2(t_m)^3 - I_2(t_{m-1})^3}{3s_{2,m}} & \dots & \frac{I_n(t_m)^3 - I_n(t_{m-1})^3}{3s_{n,m}} & 1 \end{bmatrix}$$

$$L = [L_2 \quad L_3 \quad \dots \quad L_m]^T$$

$$\hat{R} = [\hat{R}_1 \quad \hat{R}_2 \quad \dots \quad \hat{R}_n \quad l_0]^T$$

# Scalability

- A few papers use up to hundreds of thousands or millions of customers.
- An important property of NTL detection methods is that their computational time should scale to large data sets of hundreds of thousands or millions of customers. Most works reported in the literature do not satisfy this requirement.

# Comparison of different methods

- Comparing the different methods reviewed in this paper is challenging because they are tested on different data sets.
- In many cases, the description of the data lacks fundamental properties such as the number of meter readings per customer, NTL proportion, etc.

# Comparison of different methods

- In order to increase the reliability of a comparison, joint efforts of different research groups are necessary.
- These efforts need to address the benchmarking and comparability of NTL detection systems based on a comprehensive freely available data set.

# Comparison of different methods

- Carlos López, Universidad ORT Uruguay, is planning a NTL detection challenge
- Project title: Objective comparison among NTL detection methods in houses

We plan to organize a NTL detection competition

1 post by 1 author 



Dr. Ing. Carlos López-Vázquez

Jul 26



Hello everybody:

My name is Carlos López, and I am based in Universidad ORT Uruguay (Montevideo). As the subject states, we are applying to our NSF (named ANII) in order to fund a competition intended to objectively compare the performance of various NTL detection methods in a given dataset. Below you will find both the title and summary of the project. The project has been presented last year, and despite a good evaluation some weakness were noticed. ANII stated that we have no evidence that external researchers will be willing to participate in the competition. We kindly request your opinion on the initiative, and if you feel it is possible, send a formal letter stating that you are considering participate in the competition once it is launched (not before mid 2019). If you are interested in the details let me know.

Regards

Carlos



# Comparison of different methods

- Goal: Create a simulation environment in which competitors can objectively test and compare their NTL detection methods to the ones of others
- Currently looking for researchers interested in realizing the competition
- Carlos will apply for funding at Agencia Nacional de Investigación e Innovación (ANII)

# Comparison of different methods

- Have a look at our mailing list:  
<https://groups.google.com/d/forum/ntl-community>
- Details:
  - The comparison will be through a Monte Carlo simulation
  - Unlike typical competitions platforms (e.g. kaggle.com) which just require one classification, the algorithms should be run many times

# Comparison of different methods

- Challenges:
  - Derive suitable metrics to assess models
  - Get a sufficiently large dataset that can be put in the public domain

# Locality vs Similarity

Is it possible to provide an accurate detection of non-technical losses by using features only derived from provider-independent data?

# Locality vs Similarity

## Distilling Provider-Independent Data for General Detection of Non-Technical Losses

Jorge Augusto Meira, Patrick Glauner,  
Radu State and Petko Valtchev  
SnT, University of Luxembourg, Luxembourg

Lautaro Dolberg, Franck Bettinger and  
Diogo Duarte  
CHOICE Technologies Holding Sàrl, Luxembourg

*Published in:*

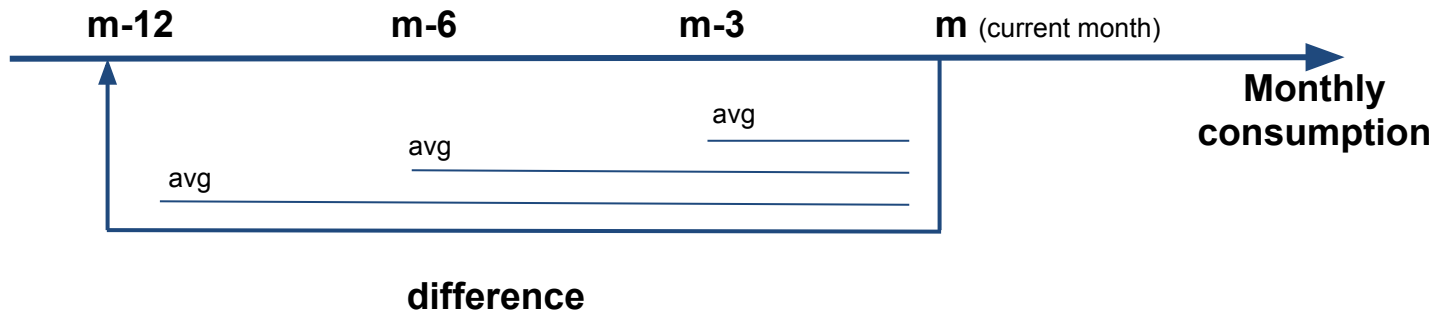
2017 IEEE Power and Energy Conference at Illinois (PECI 2017),  
Urbana, USA, 2017.

# Features

Based on the following categories:

- Temporal: Seasonal, Monthly, Semiannual, Quarterly, Intra Year;
- Locality: Geographical Neighbourhoods;
- **Similarity: k-means clustering using consumption profile**
- Infrastructure: Transformers;

# Features



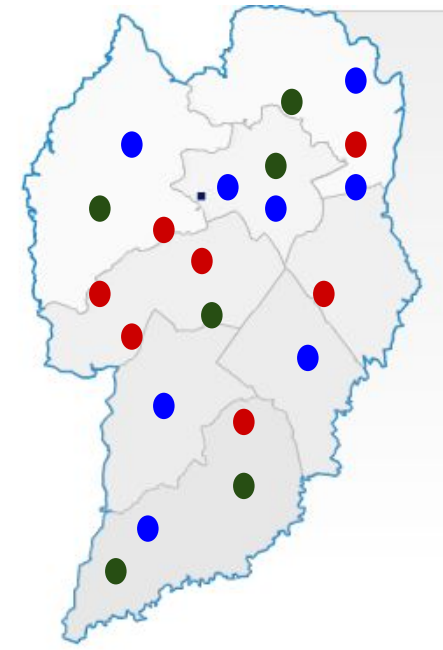
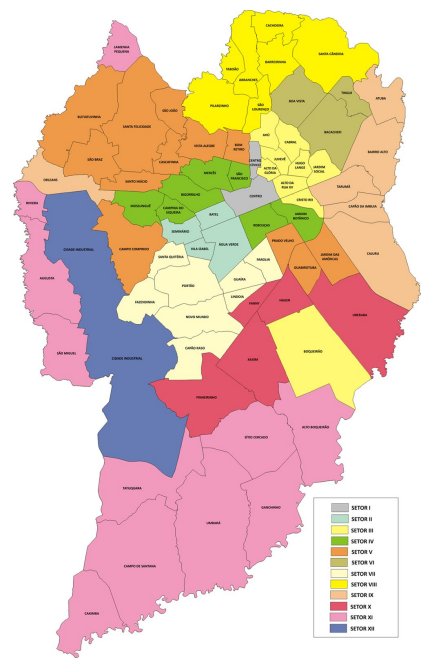
- The temporal features are calculated individually and for each of the three subsequent categories: Locality, Similarity and Infrastructure

# Locality vs Similarity

Neighbourhood (code\_Neig)

vs

Consumption Profile (code\_Neig)



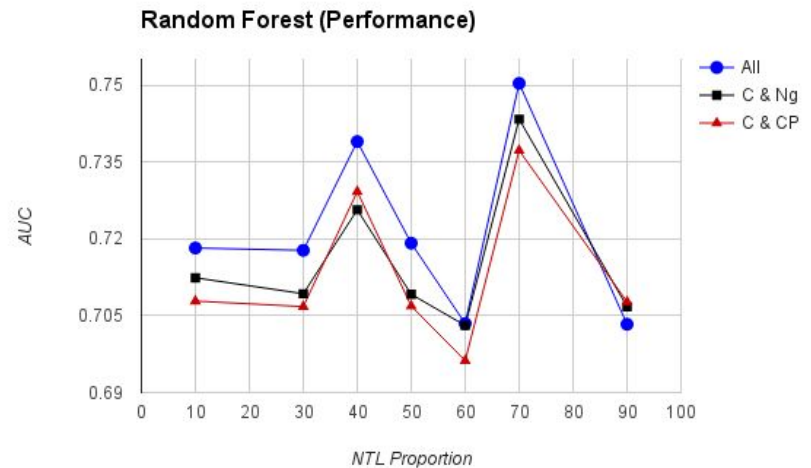
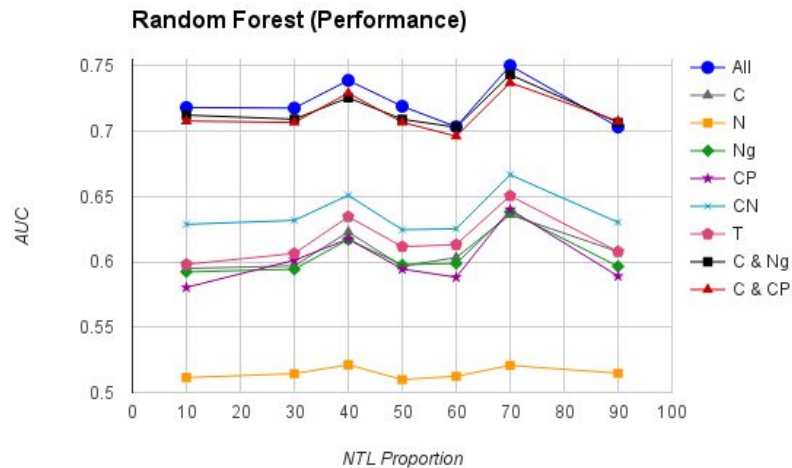
- Cluster 1
- Cluster 2
- Cluster 3



# Locality vs Similarity

Set of features	Description
Notes (N)	Meter reader's notes
Consumption (C)	Fixed Interval + Fixed Lag
Consumption & Notes (CN)	Fixed Interval and Notes
Neighbourhood (Ng)	Intra Group (geographical neighbourhood)
Transformers (T)	Intra Group (Transformers)
Consumption Profile (CP)	Intra Group (k-means clustering)
C & Ng	Consumption and Neighbourhood
C & CP	Consumption and Consumption Profile
All	N+C+Ng+CP+T

# Locality vs Similarity



# Conclusion

- Several sets of features computed using four criteria: temporal, locality, similarity and infrastructure.
- The experimental results show that sets of features supported only by raw consumption data can achieve satisfactory performance when compared with sets composed of "providers' dependent features".

# Challenge

A 100% automatic tool

# Mixed reality

- How to improve targets selection by taking advantage of the domain expert experience?

# Mixed reality

## Identifying Irregular Power Usage by Turning Predictions into Holographic Spatial Visualizations

Patrick Glauner\*, Niklas Dahringer\*, Oleksandr Puhachov\*, Jorge Augusto Meira\*,  
Petko Valtchev<sup>†</sup>, Radu State\* and Diogo Duarte<sup>‡</sup>

*Published in:*

Proceedings of the 17th IEEE International Conference on Data Mining Workshops (ICDMW 2017), New Orleans, USA, 2017.

# Mixed reality

- How to improve targets selection by taking advantage of the domain expert experience?



*HoloLens*

# Mixed reality

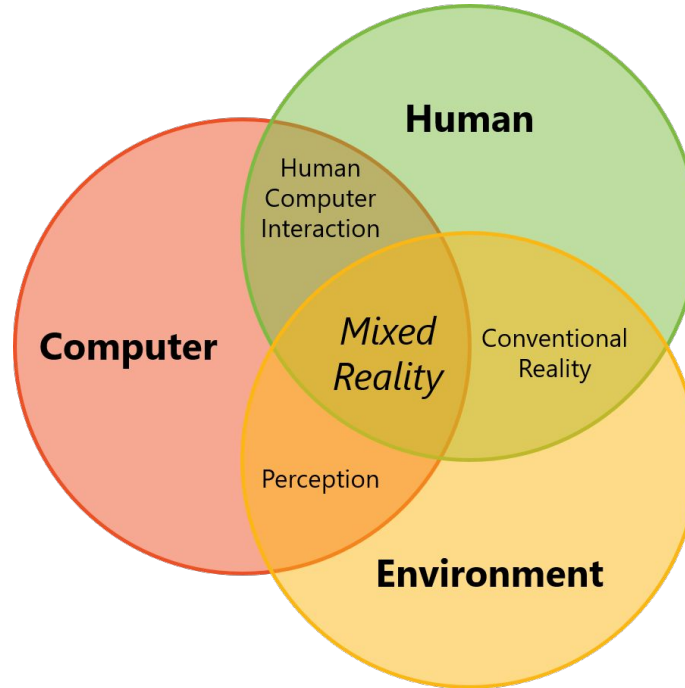
- In 1994, the paper "A Taxonomy of Mixed Reality Visual Displays" introduced the term *Mixed Reality*.

*“... a particular subset of Virtual Reality (VR) related technologies that involve the merging of real and virtual worlds somewhere along the "virtuality continuum" which connects completely real environments to completely virtual ones.”*



# Mixed reality

- In short, MR is the result of blending the physical world with the digital world.



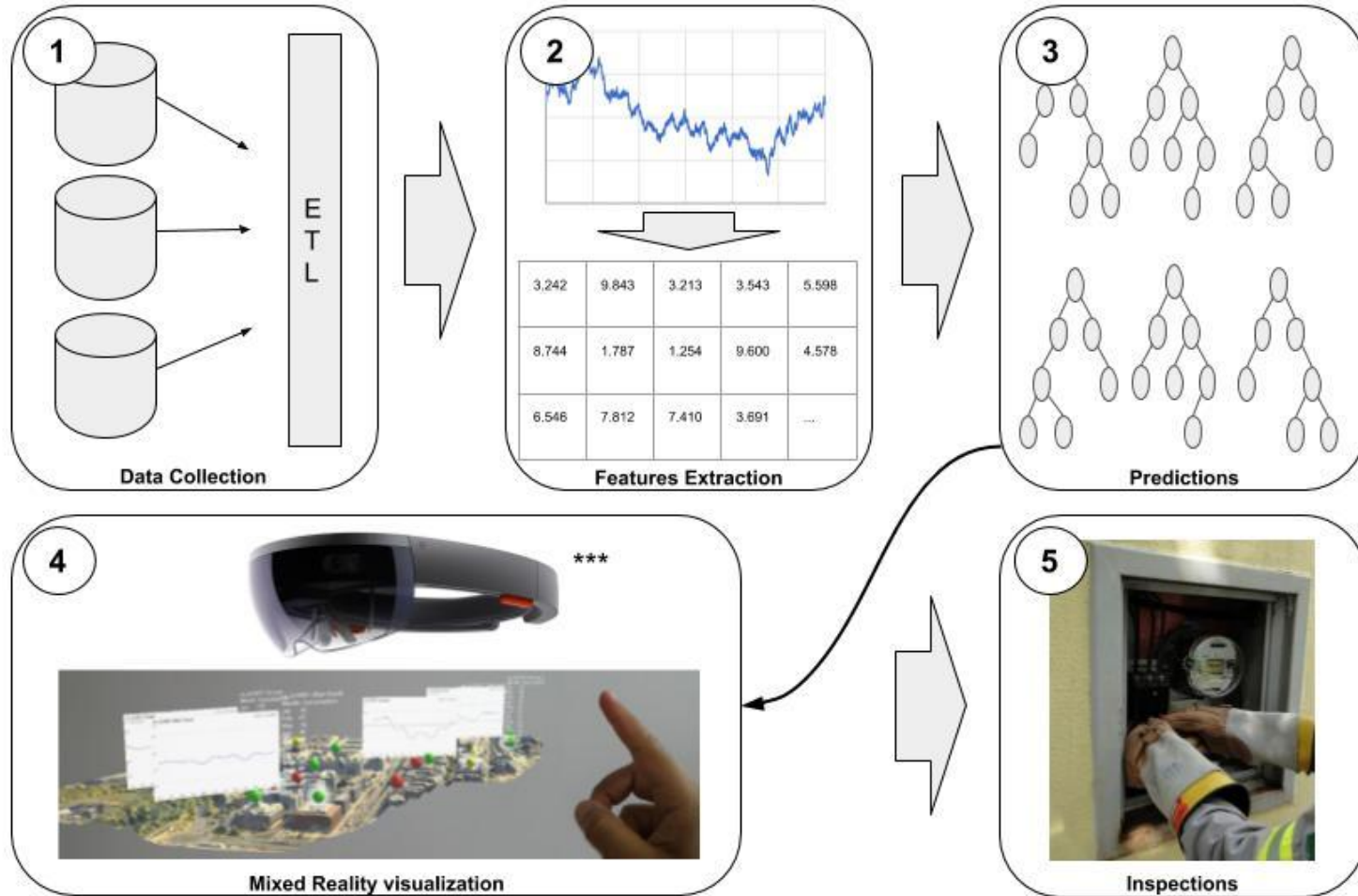
source:

[https://developer.microsoft.com/en-us/windows/mixed-reality/mixed\\_reality](https://developer.microsoft.com/en-us/windows/mixed-reality/mixed_reality)

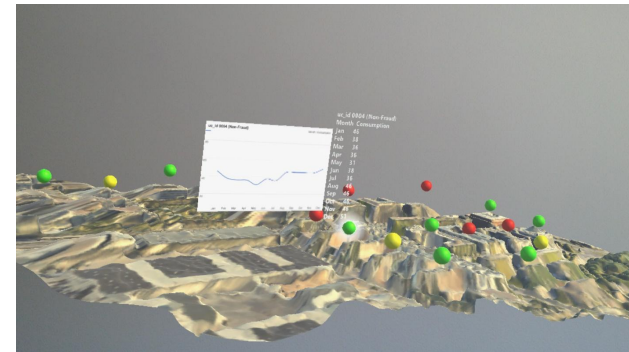
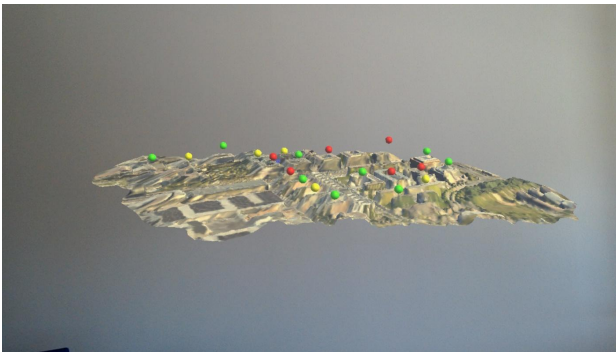
# Mixed reality

- A novel approach to support to visualize the prediction results in a 3D hologram that contains information about customers and their spatial neighborhood.

# Mixed reality



# Mixed reality



# Mixed reality





# Conclusions

- Non-technical losses (NTL) cause major financial losses to electricity suppliers
- Detecting NTL thrives significant economic value
- Different approaches reported in the literature, superior performance of machine learning approaches compared to expert system
- Many open challenges

# Interested in NTL?

Join our mailing list:

<https://groups.google.com/d/forum/ntl-community>

-  We plan to organize a NTL detection competition  
By Dr. Ing. Carlos López-Vázquez - 1 post - 4 views
-  Upcoming conferences to discuss NTL detection  
By me - 4 posts - 5 views

# References

- [1] P. Glauner, et al., "**The Challenge of Non-Technical Loss Detection using Artificial Intelligence: A Survey**", International Journal of Computational Intelligence Systems (IJCIS), vol. 10, issue 1, pp. 760-775, 2017.
- [2] P. Glauner, et al., "**Large-Scale Detection of Non-Technical Losses in Imbalanced Data Sets**", Proceedings of the Seventh IEEE Conference on Innovative Smart Grid Technologies (ISGT 2016), Minneapolis, USA, 2016.
- [3] J. Meira, et al., "**Distilling Provider-Independent Data for General Detection of Non-Technical Losses**", 2017 IEEE Power and Energy Conference at Illinois (PECI 2017), Urbana, USA, 2017.



# References

- [4] P. Glauner, et al., "**Neighborhood Features Help Detecting Non-Technical Losses in Big Data Sets**", Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing Applications and Technologies (BDCAT 2016), Shanghai, China, 2016.
- [5] P. Glauner, et al., "**Is Big Data Sufficient for a Reliable Detection of Non-Technical Losses?**", Proceedings of the 19th International Conference on Intelligent System Applications to Power Systems (ISAP 2017), San Antonio, USA, 2017.

# References

- [6] P. Glauner, et al., "**Identifying Irregular Power Usage by Turning Predictions into Holographic Spatial Visualizations**", Proceedings of the 17th IEEE International Conference on Data Mining Workshops (ICDMW 2017), New Orleans, USA, 2017.