Visualizing high-resolution exploratory energy maps of energy-performance certificates

Tania CERQUITELLI
Department of Control and Computer engineering, Politecnico di Torino, Italy

Alfonso CAPOZZOLI (DENERG), Elena BARALIS (DAUIN), Marco MELLIA (DET)

Main reasearch objective

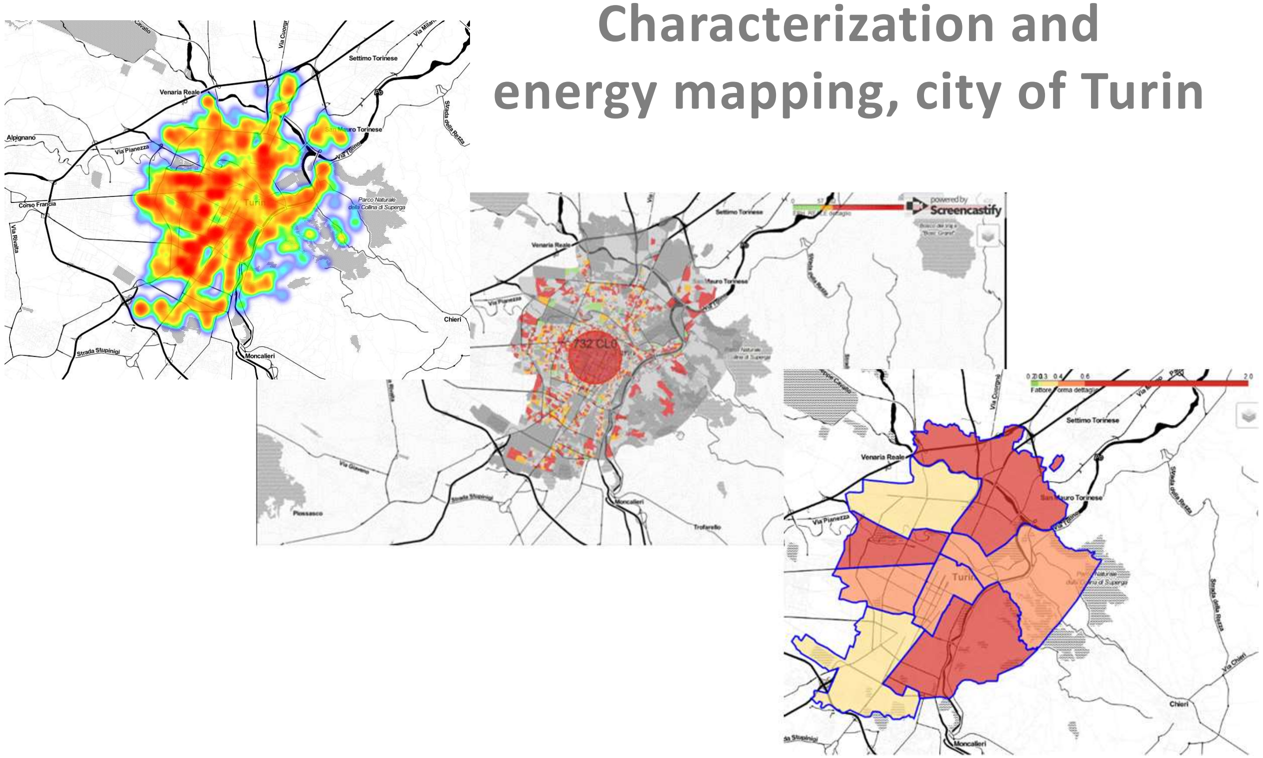
ENERGY DATA

OPEN DATA

Value for different stakeholders

Support and improve decisional processes

Characterization and energy mapping, city of Turin



Main reasearch objective

ENERGY DATA

OPEN DATA

Value for different stakeholders

Support and improve decisional processes

Citizens

Estate agents

Create value from energy open data

Public Administration

Energy companies

Values for the stakeholders

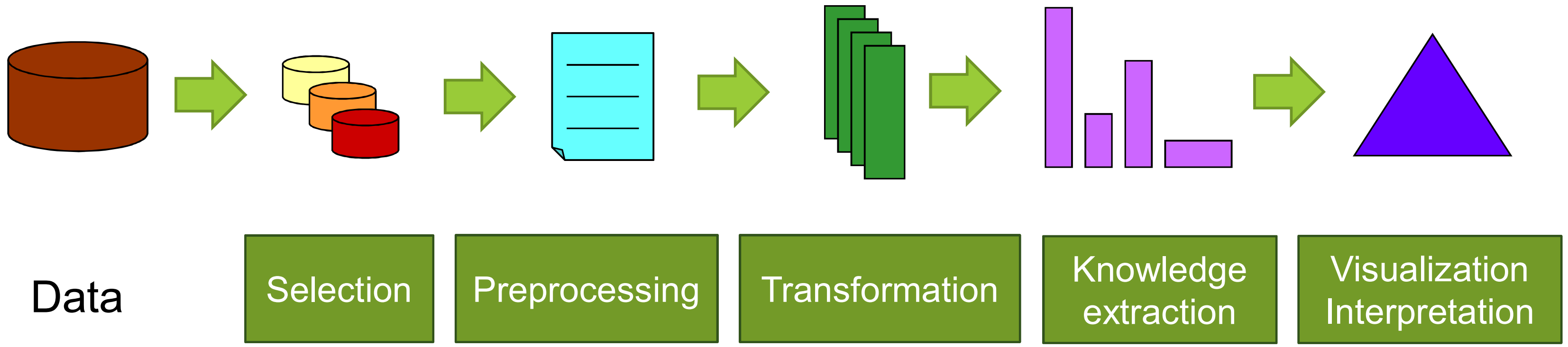
- ✓ **Mapping the energy demand** of buildings at neighborhood or city level
- ✓ **Characterization of metropolitan areas** with respect to energy-efficiency parameters
- ✓ **Targeted incentive policies**
- ✓ Energy planning
- ✓ Development of **more accurate benchmark models**
- ✓ Evaluation of the **effect** obtained through **retrofit measures**
- ✓ Targeted **promotional offers**

POLITECNICO DI TORINO

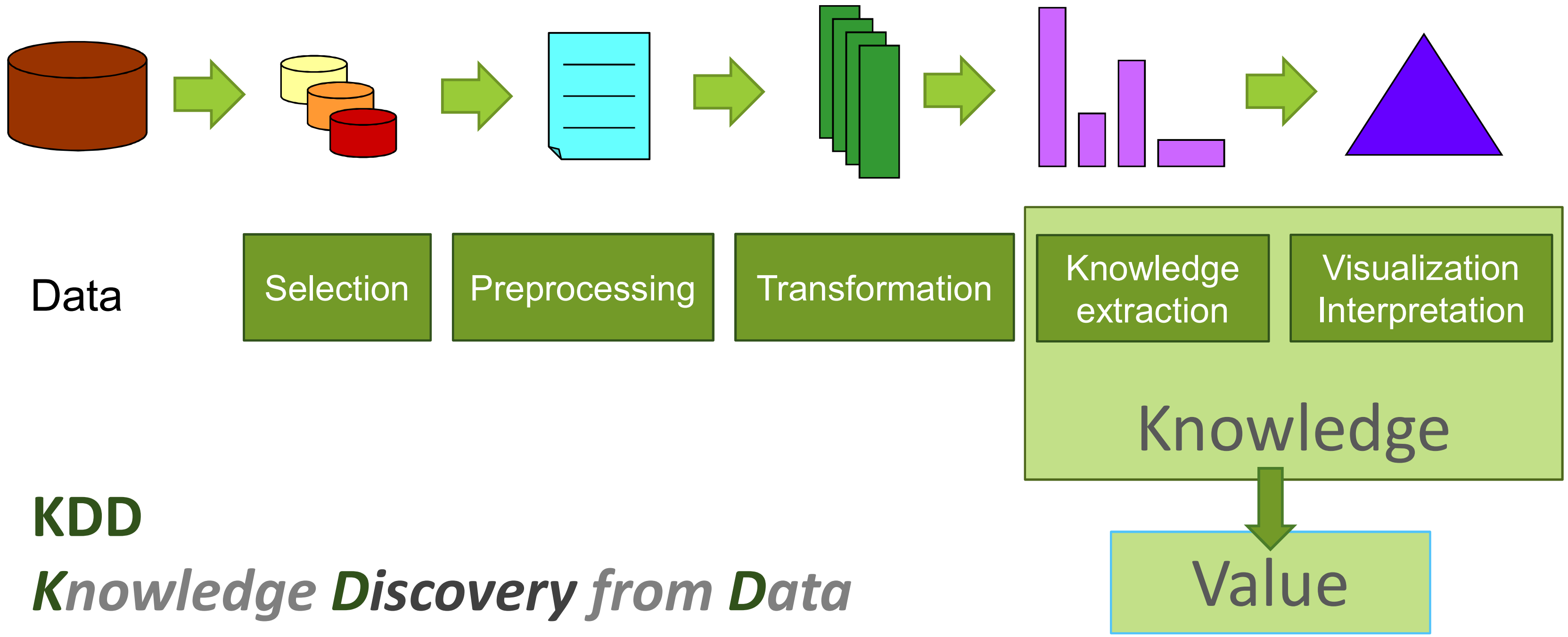
SmartData@PoliTO

SmartData

Knowledge extraction process



Knowledge extraction process



KDD

Knowledge Discovery from Data

KDD from energy data: two key roles

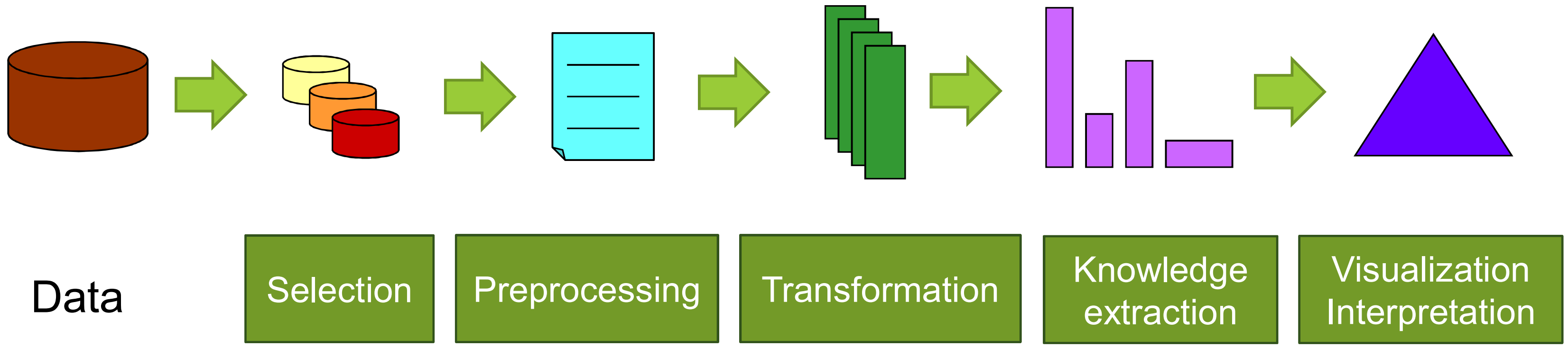
DATA SCIENTIST

- Design **innovative and efficient algorithms**
- Select the **optimal techniques** to address the challenges of the analysis
- Identify the best **trade-off** between knowledge quality and execution time

ENERGY SCIENTIST

- Support **the data pre-processing** phase
- **Assess** extracted **knowledge**
- Strong involvement in the algorithm definition phase which should **respect/include physical laws** and correctly **model physical events**

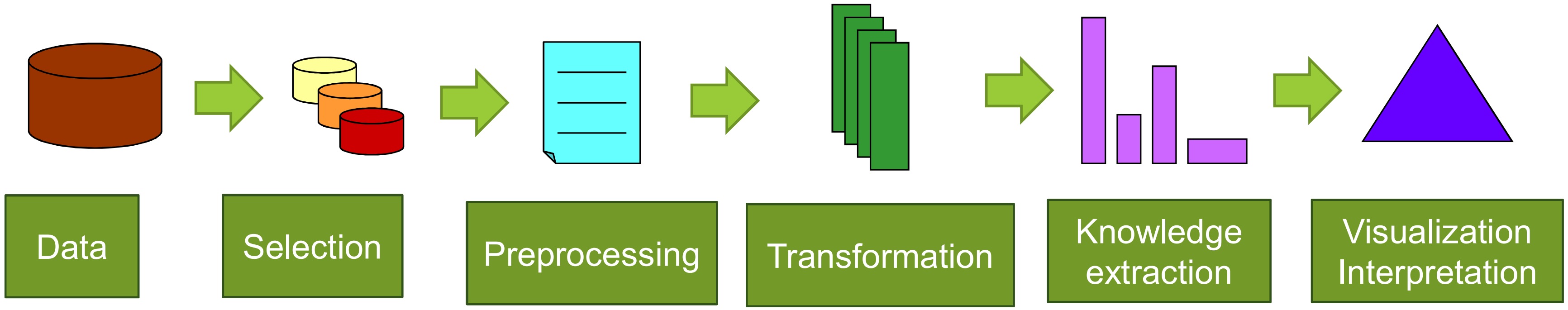
Knowledge extraction process

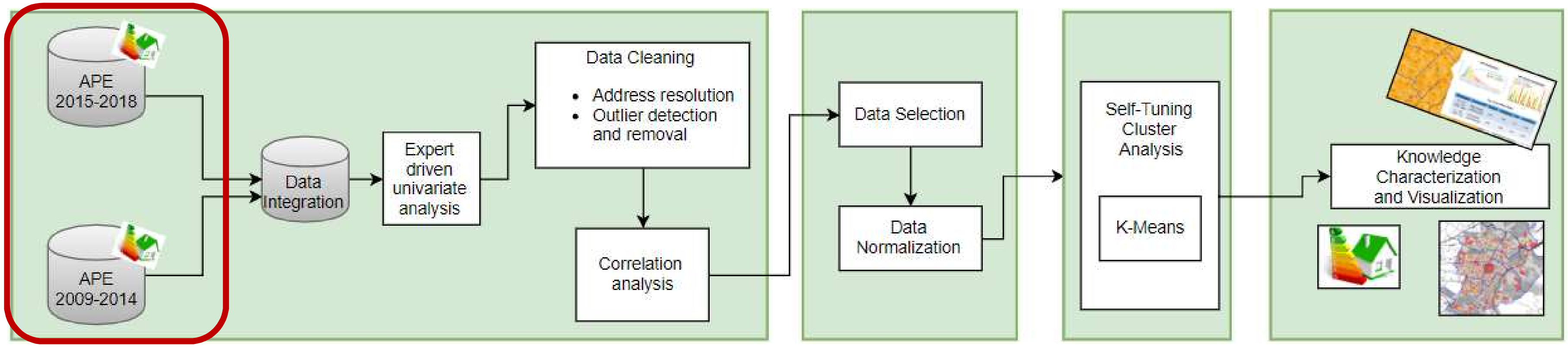


Innovations in the data analytics process

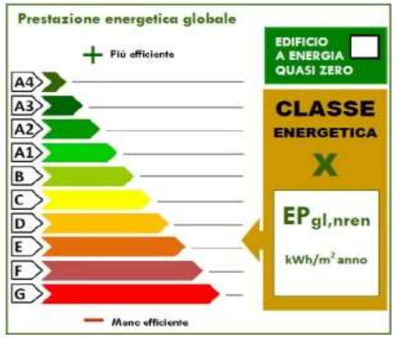
- **Tailor** the **analytic** steps to the different key aspects of **energy data**
- **Automate** the data analytic workflow to **reduce the manual user interventions**
- Translate the domain-expert knowledge into **automated procedures**
- Design **informative dashboards** to support the translation of the extracted knowledge into effective actions

Knowledge extraction process from APE

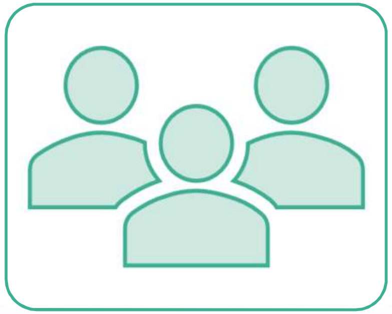




Cerquitelli T., Di Corso E., Proto S,, Capozzoli A., Bellotti F., Cassese M.G., Baralis E., Mellia M., Casagrande S., Tamburini M. ***Exploring Energy Performance Certificates through Visualization*. In Proceedings of the Workshops of the EDBT/ICDT 2019 Joint Conference (EDBT/ICDT 2019) Lisbon, Portugal, March 26, 2019.**



Open data: Energy Certificate of Buildings



Energy analysis of the building

Walling and window characteristics

Geometric features of the building

Hot water production

Environment cooling and heating

Type of plant

Renewable-energy production systems

Energy certificate officer

Qualified technicians granting APE certificates

Use of specific software (this information is not available in open data)

Building purchases

Lease agreements

Interventions to improve the building energy efficiency

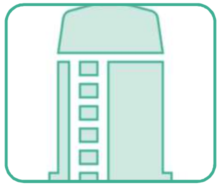
Case study: APE in Piedmont Region

Open data available on the Sistema Piemonte service system *

Each APE is characterized by **175 attributes**, both categorical and numerical

Real building

- **Thermo-physical** characteristics (e.g., Average U-value of the vertical opaque envelope/Average U-value of the windows)
- **Geometric** features (e.g. Heated volume, Heat transfer surface, Aspect ratio)
- **Plant** characteristics (e.g. Efficiencies of the heating plant subsystems)
- **Energy** performance (e.g. Energy demands for different energy services: heating, cooling, ACS e lighting)



Reference building

- Thermo-physical characteristics
- Geometric features
- Plant characteristics
- Energy performance

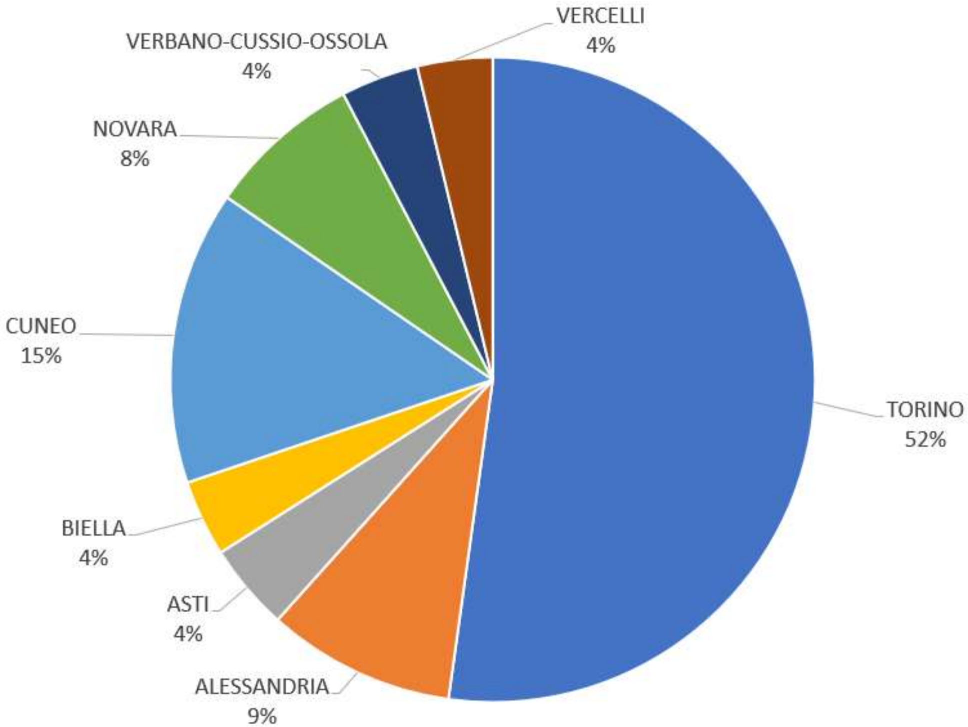
Recommendations

- Possible **actions** to improve energy performance of the building

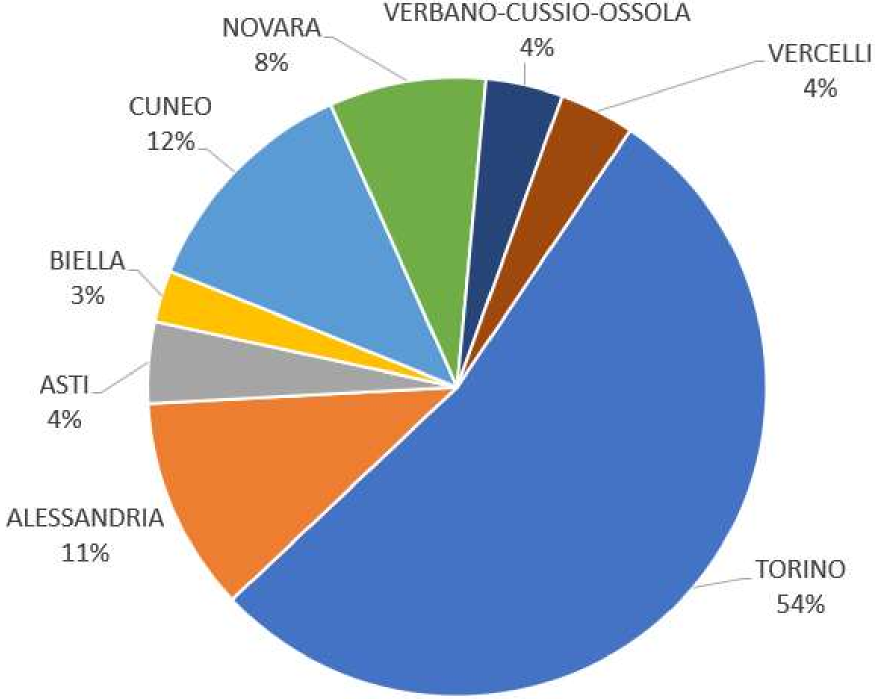
* http://www.sistemapiemonte.it/cms/privati/ambiente-e-energia/servizi/856-sistema-informativo-per-le-prestazioni-energetiche-degli-edifici-sipee

APE in Piedmont Region: 2 data sources

Distribution of the number of APEs by **province**

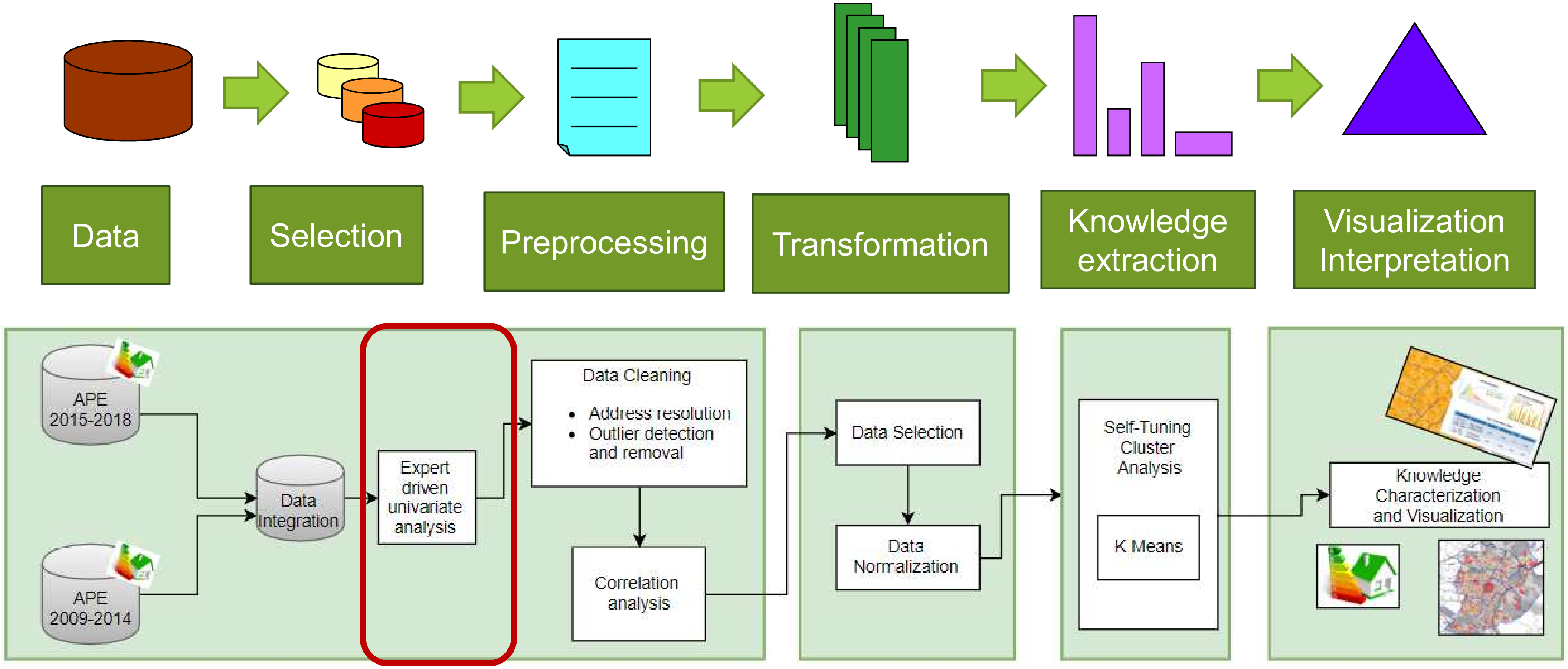


Reference period **2009 – 2014**
APE no. 190,124



Reference period **2015 – 06/2018**
APE no. 78,733

Knowledge extraction process from APE



Expert-driven univariate analysis

E1 (1) buildings used as permanent residence.

Identification of the most important variables

- Average U-value of the vertical opaque envelope
- Average U-value of the windows
- Aspect Ratio
- Efficiency of the plant subsystems
- ...

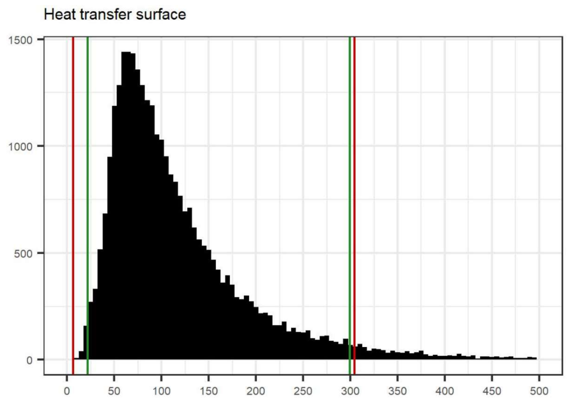
Expert-driven univariate analysis

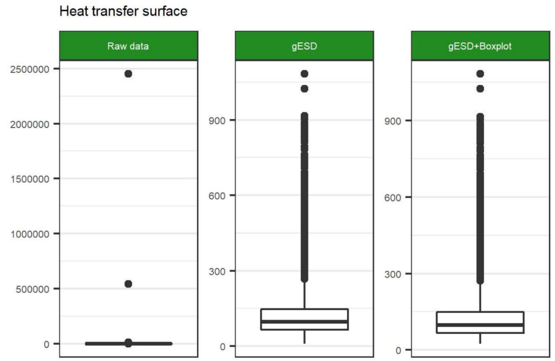
E1 (1) buildings used as permanent residence

Identification of the most important variables → Identification of the validity ranges for each variable

- Average U-value of the vertical opaque envelope
- Average U-value of the windows
- Aspect Ratio
- Efficiency of the plant subsystems
- ...

Outlier detection based on
- **Knowledge of domain experts**
- **gESD method** needs as input parameter the upper-bound of potential outliers
- **Boxplot** visually displays a data distribution through its quartiles





gESD = generalized Extreme Studentized Deviate

Expert-driven univariate analysis

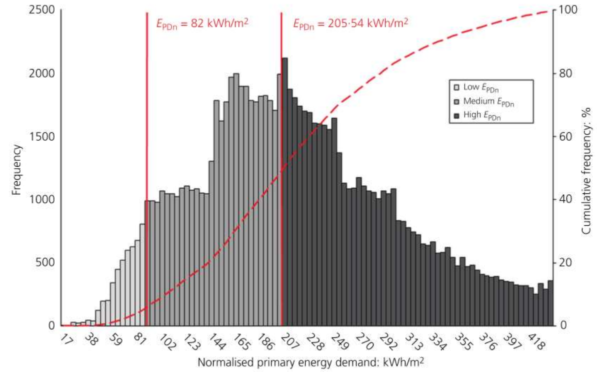
E1 (1) buildings used as permanent residence.

Identification of the most important variables → Identification of the validity ranges for each variable → APE characterized by admissible values (~ 40%) / APE that require further investigation

- Average U-value of the vertical opaque envelope
- Average U-value of the windows
- Aspect Ratio
- Efficiency of the plant subsystems
- ...

Outlier detection based on

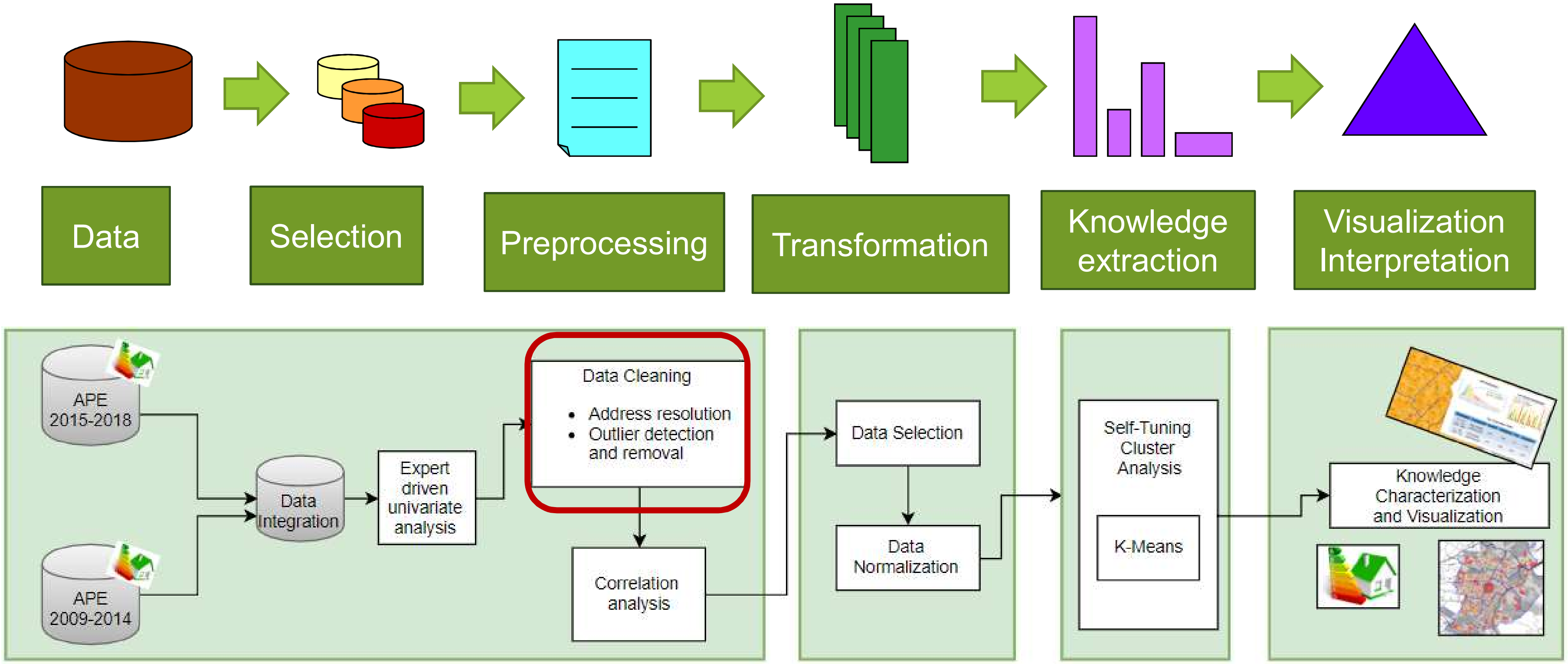
- **Knowledge of domain experts**
- **gESD method** needs as input parameter the upper-bound of potential outliers
- **Boxplot** visually displays a data distribution through its quartiles



Normalized primary energy demand distribution

Source: Capozzoli A, Serale G, Piscitelli MS, Grassi D. *Data mining for energy analysis of a large data set of flats.* (Proc Inst Civ Eng) Engineering Sustain 2017.

Knowledge extraction process from APE



Data cleaning: address resolution

APE with invalid address format

- Typing errors
- Incorrectly-coded characters
- 31.6% of the addresses have a generic 10100 CAP
- Wrong longitude and longitude coordinates

Adopted solution

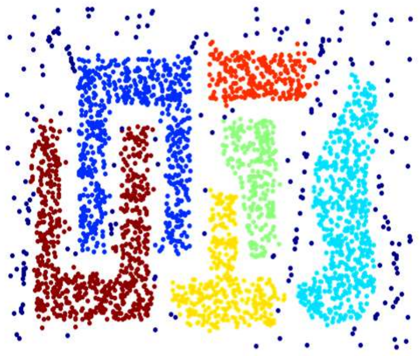
- Addresses in the DB have been **compared** to those stored in the **Turin road list** (from ***Geoportale Comune di Torino***[1])
- **Levenshtein** distance to compute the similarity index between the addresses reported in the APE DB and the reference DB.
  - If the address has been **resolved**, the CAP and the coordinates are saved in our DB eliminating inconsistencies
  - If the address has **not** been **resolved**, the CAP and coordinates are obtained through the Google[2] geocoding API

1 https://developers.google.com/maps/documentation/geocoding/intro
2 http://geoportale.comune.torino.it/web/

Outlier detection: multivariate analysis

Density-based clustering algorithm: **DBScan**

- Splits the database in parts characterized by different densities (dense and sparse)
- **Density** is defined by two parameters (i.e., Eps, MinPoints), that are difficult to set



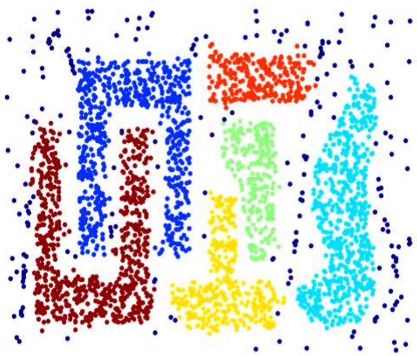
Clustering with DBScan

From: Tan, Steinbach, Kumar, *Introduction to Data Mining*, McGraw Hill 2006

Outlier detection: multivariate analysis

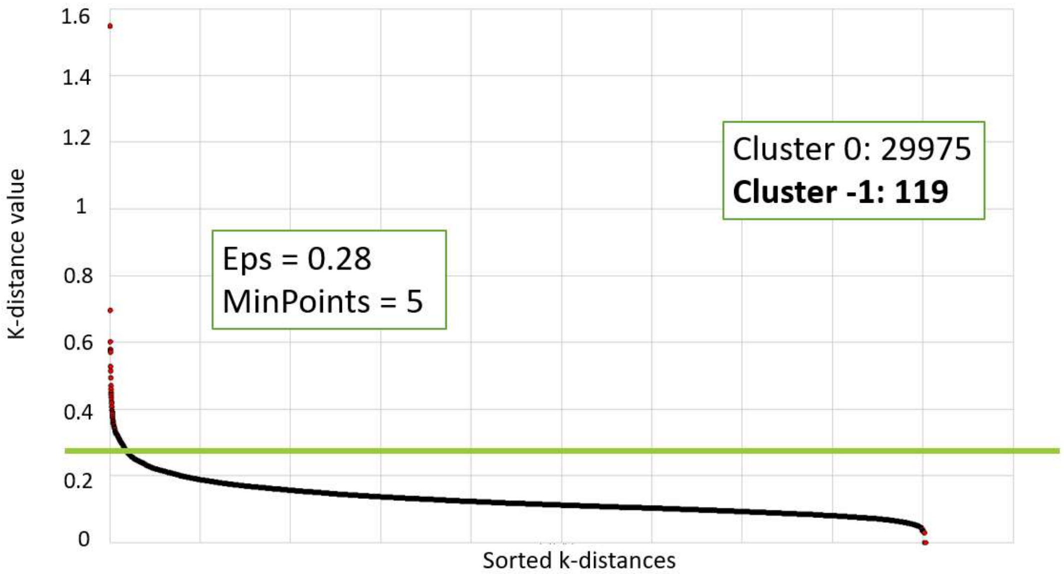
Density-based clustering algorithm: **DBScan**

- Splits the database in parts characterized by different densities (dense and sparse)
- **Density** is defined by two parameters (i.e., Eps, MinPoints), that are difficult to set
- Self-tuning strategy based on k-dist plot
  - sorted distance of every point to its kth nearest neighbor

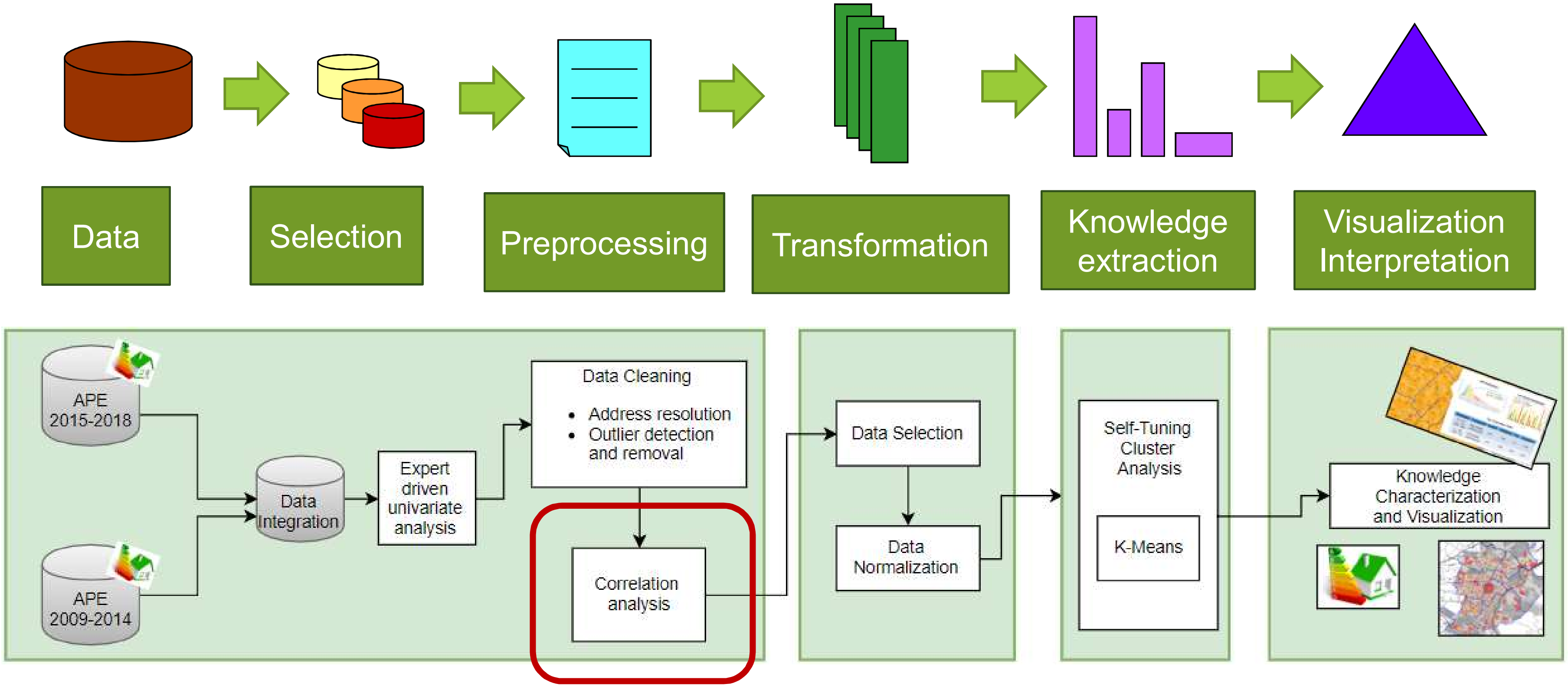


Clustering with DBScan

From: Tan, Steinbach, Kumar, *Introduction to Data Mining*, McGraw Hill 2006



Knowledge extraction process from APE

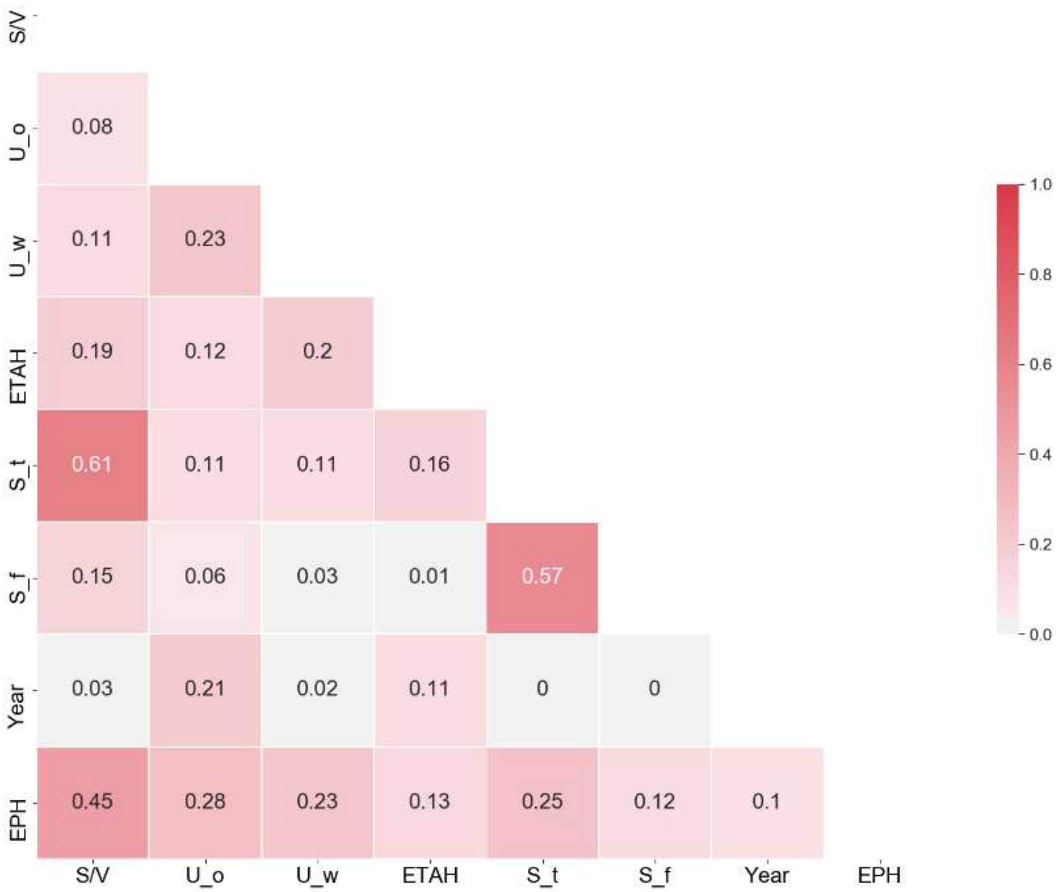


Correlation analyis

Data-driven

- Feature **removal** (correlation-based approach)
  - simplifying the model computation
  - improving the model performance
- Feature **selection**
  - Multicollinearity
    - Variables that can be predicted from the others with a substantial degree of accuracy using a multiple regression model could be discarded from the analysis
  - Correlation Test
    - Features highly-correlated with other attributes (i.e., having dependence or association in any statistical relationship, whether causal or not) could be discarded from the analysis

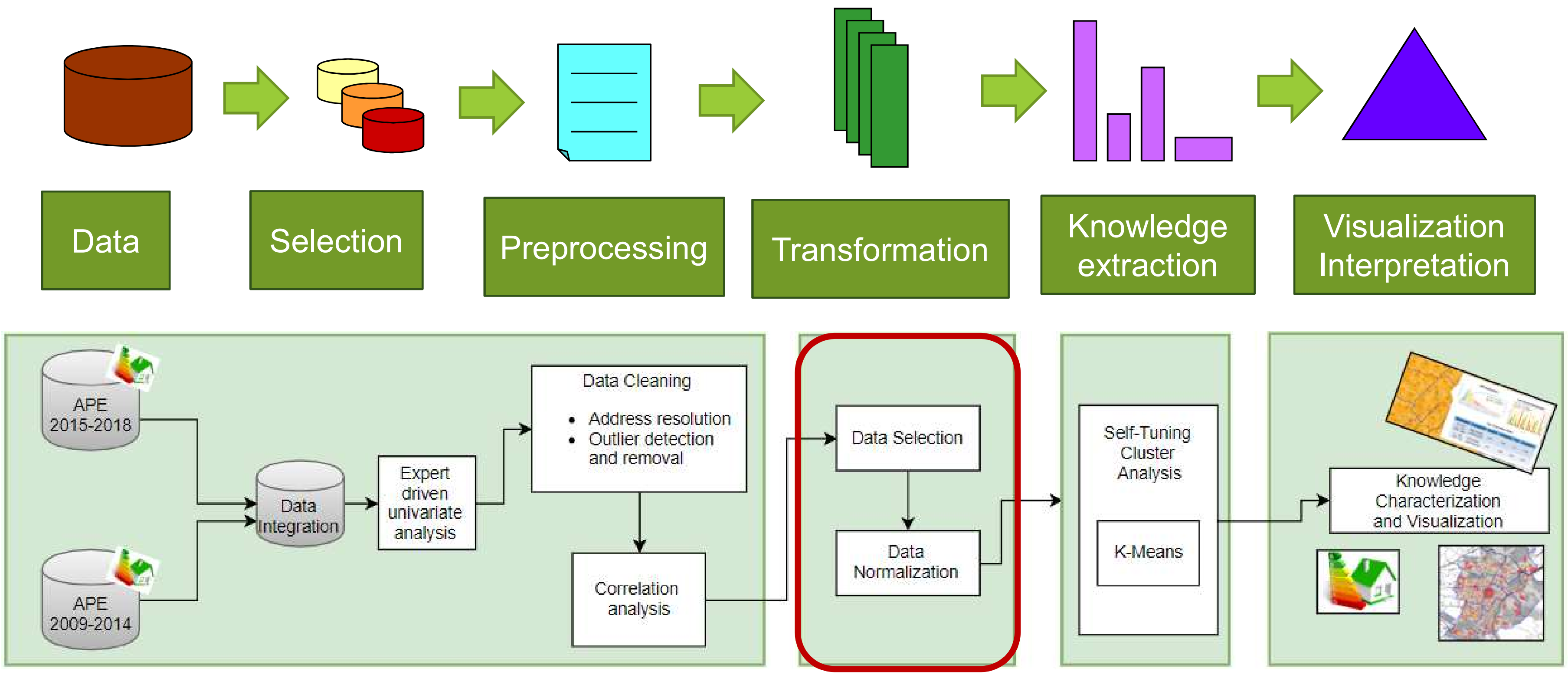
Correlation analysis



- **S/V** Aspect Ratio
- **U_o** Average u-value of opaque envelope
- **U_w** Average u-value of the windows
- **ETAH** Average global efficiency for spacing heating
- **S_t** Heat transfer surface
- **S_f** Floor Area
- **Year** Construction Year

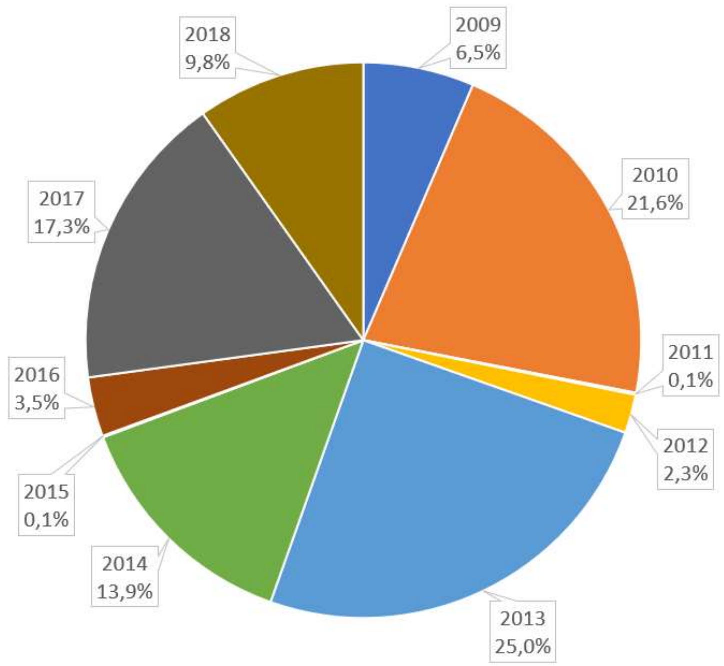
- **EPH** Normalized primary heating energy consumption

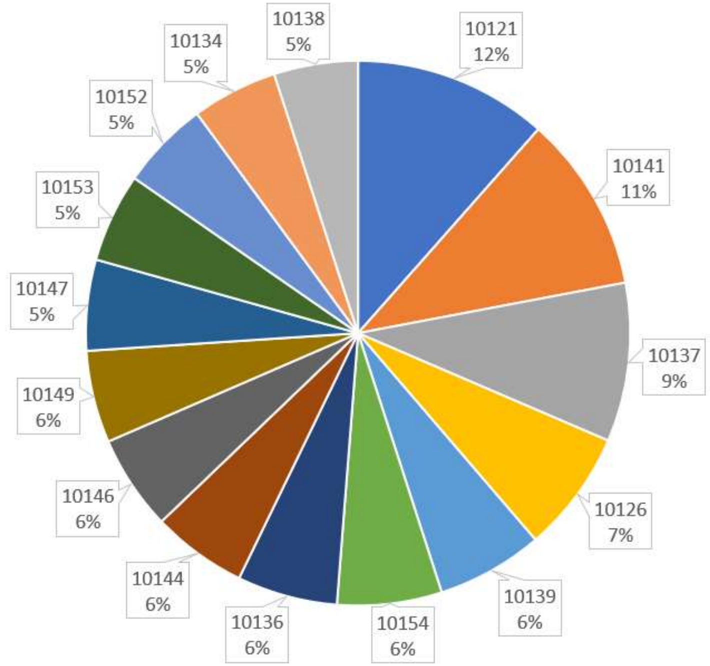
Knowledge extraction process from APE



Selected APE

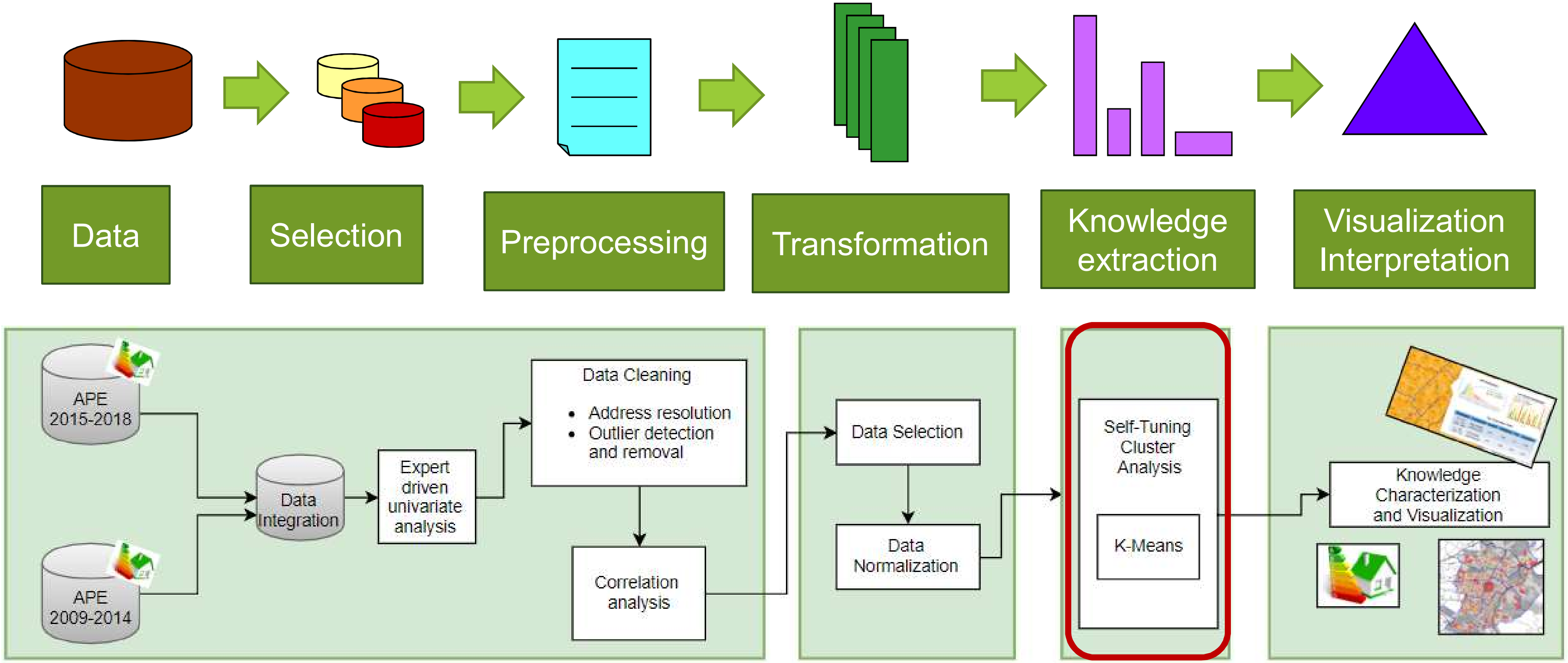
- E1 (1) buildings used as **permanent residence**
- APE issued in the period: **2009 – 2018**
- APE for ***particella, foglio*** *e* ***subalterno*** (identifying each single dwelling)
- **Number of selected APE: ~30.000**





Number of APEs separately by year (left) and by CAP (right)

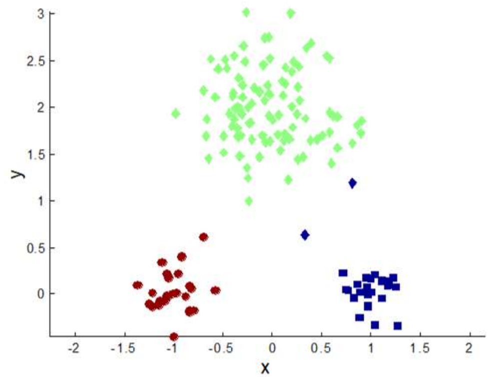
Knowledge extraction process from APE



Self-tuning cluster analysis

Clustering algorithms enriched by **self-tuning strategies** (i.e., parameter **autoconfiguration**)

- Partitional algorithm: **K-Means**
  - Each cluster is represented by a **centroid**
  - The desired **number of clusters** is identified by the user



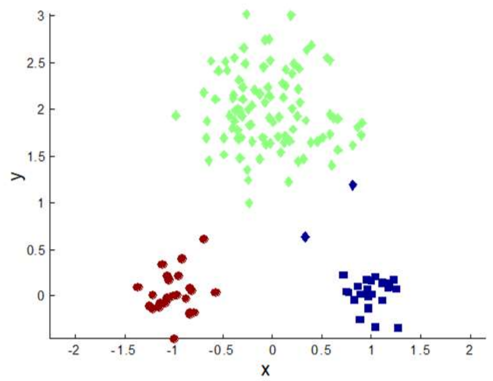
Optimal Clustering with K-Means

From: Tan, Steinbach, Kumar, *Introduction to Data Mining*, McGraw Hill 2006

Self-tuning cluster analysis

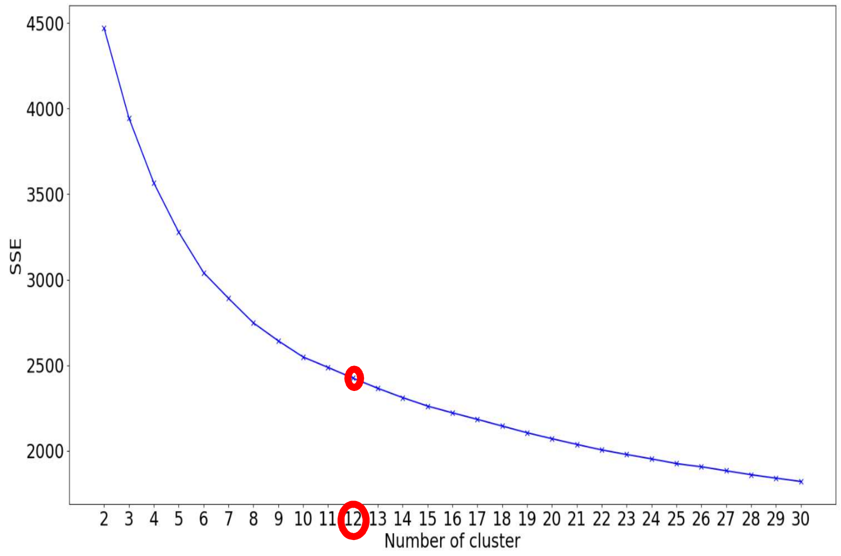
Clustering algorithms enriched by **self-tuning strategies** (i.e., parameter **autoconfiguration**)

- Partitional algorithm: **K-Means**
  - Each cluster is represented by a **centroid**
  - The desired **number of clusters** is identified by the user
- Self-tuning strategy based on the **Elbow plot:** quality-measure trend (e.g., SSE) vs K
  - The gain from adding a centroid is negligible
  - The reduction of the quality measure is not interesting anymore

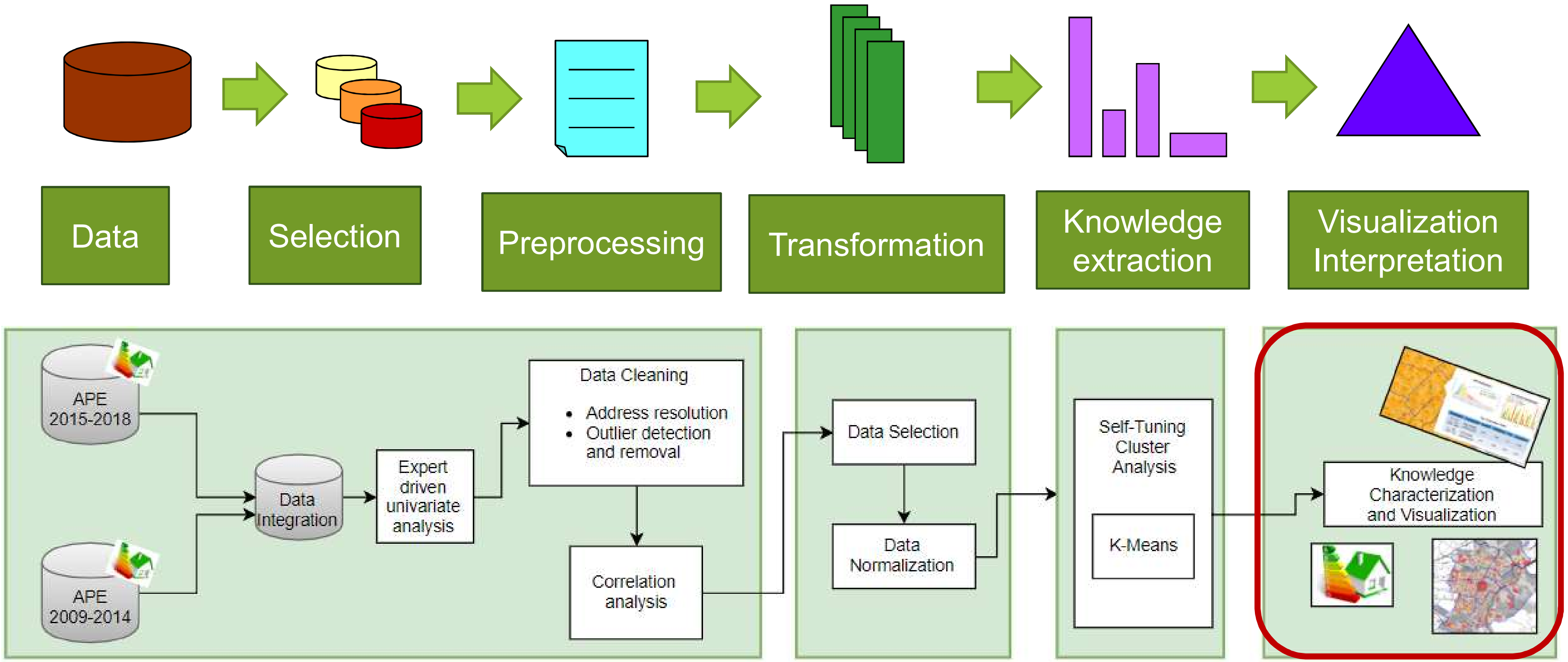


Optimal Clustering with K-Means

From: Tan, Steinbach, Kumar, *Introduction to Data Mining*, McGraw Hill 2006



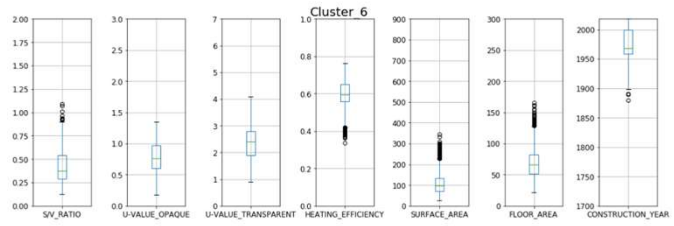
Knowledge extraction process from APE

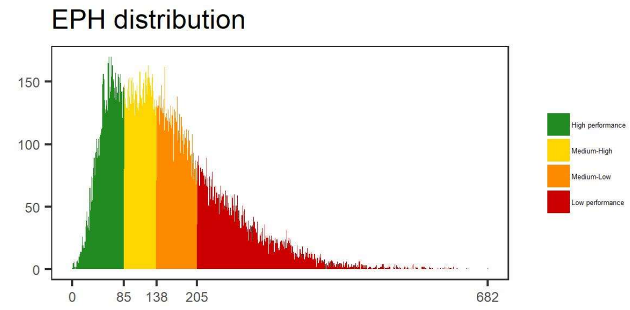


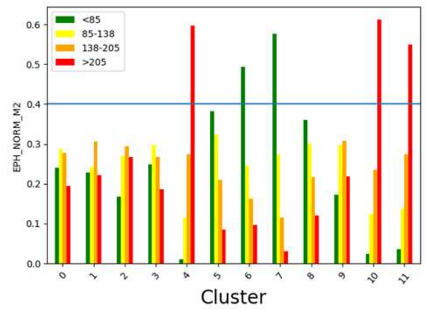
Knowledge characterization

Each discovered cluster of APEs is characterized through

- Centroids represented through radar plots
- Data distribution for each attribute modeled through boxplot
- Cluster label assigned by analyzing the EPH distribution locally

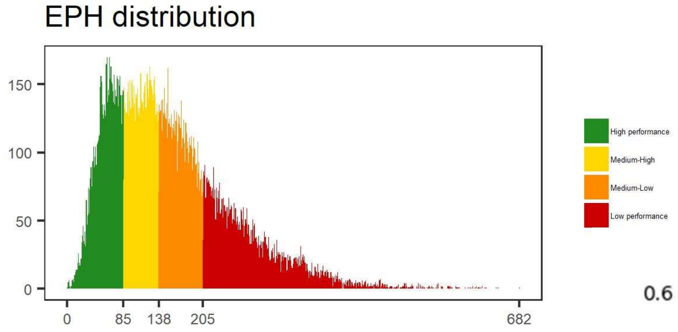


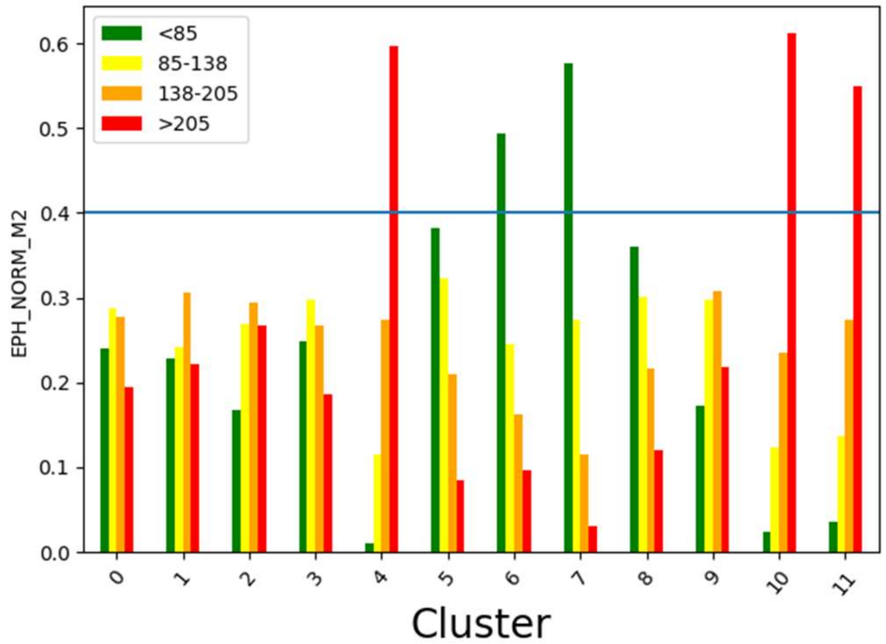




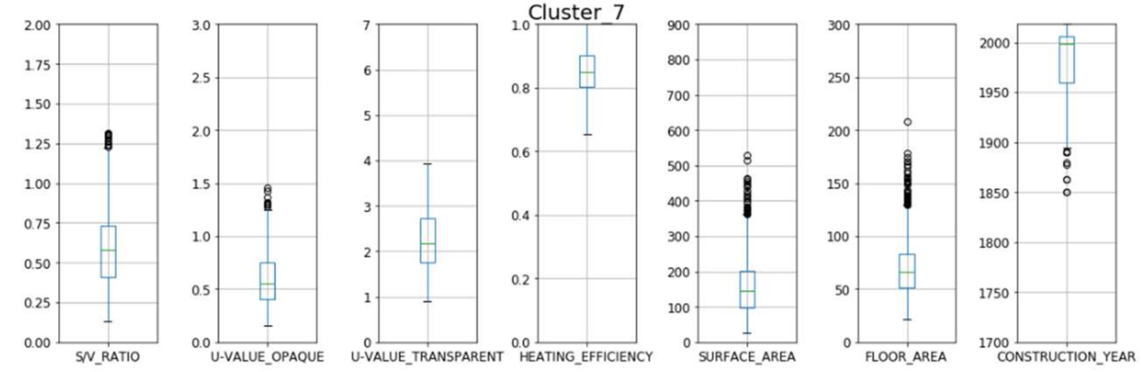
Cluster characterization

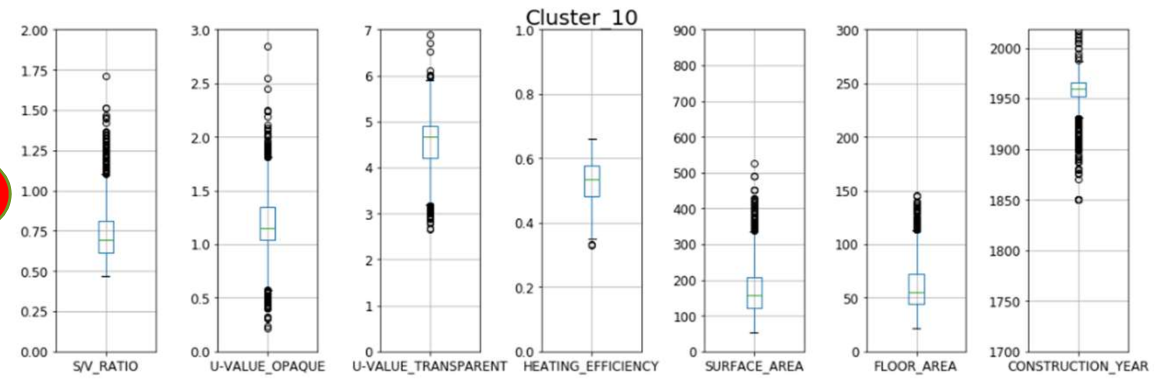
Cluster ID	APE #
Cluster 0	811
Cluster 1	4,321
Cluster 2	1,117
Cluster 3	3,988
Cluster 4	2,080
Cluster 5	2,723
Cluster 6	2,264
Cluster 7	1,723
Cluster 8	3,369
Cluster 9	3,418
Cluster 10	2,042
Cluster 11	2,119

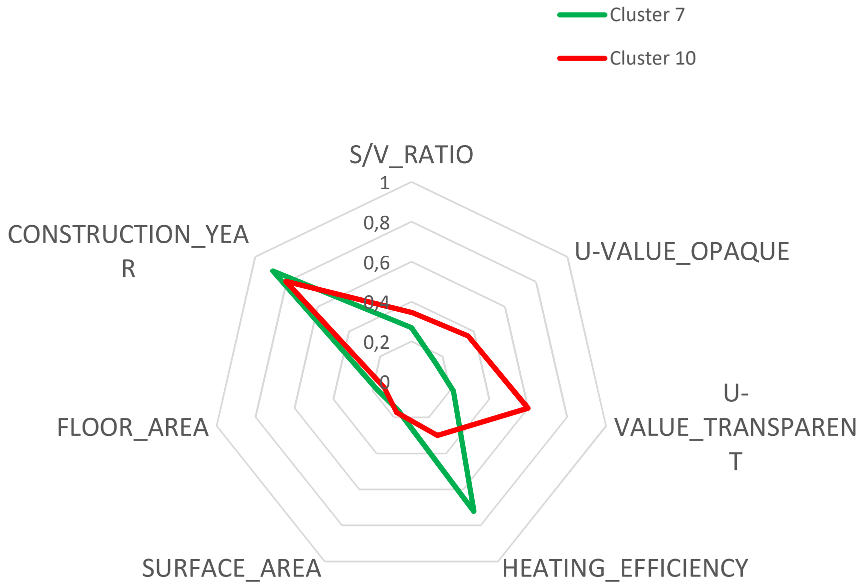




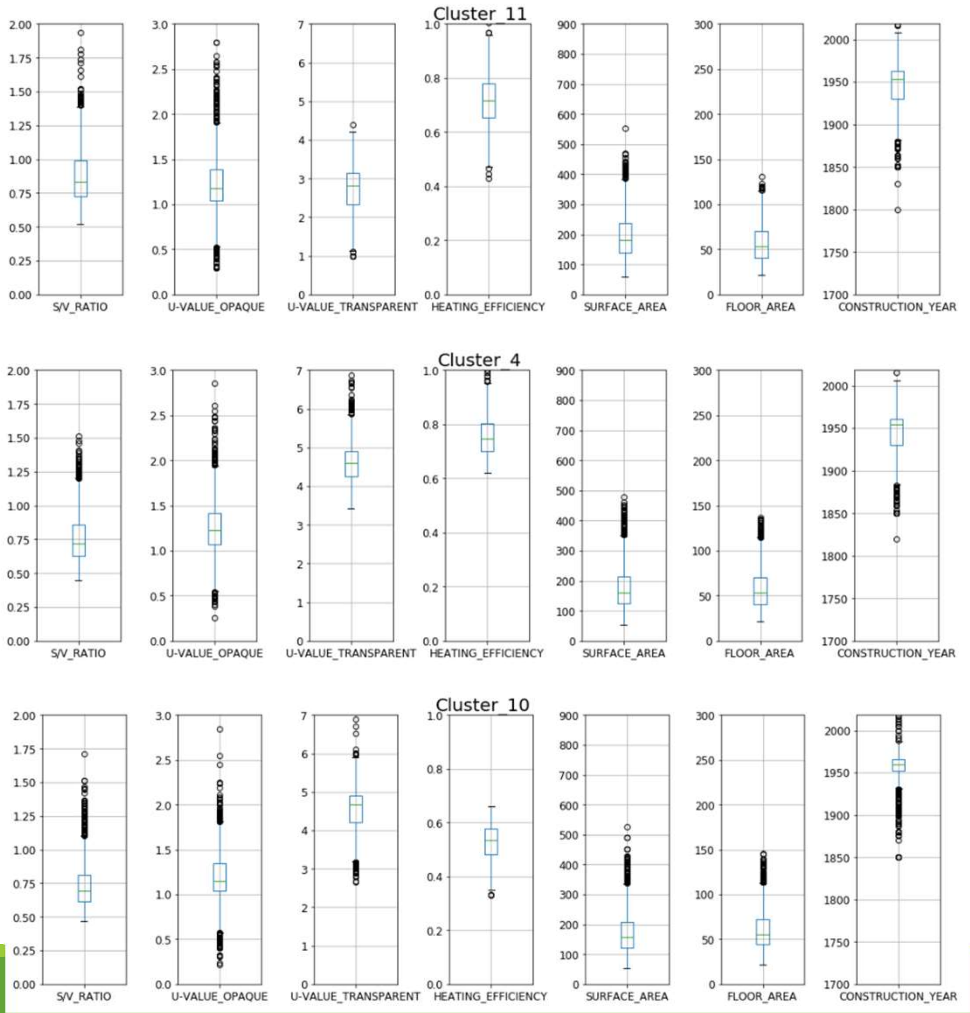
Clusters of APEs: High vs Low energy performance

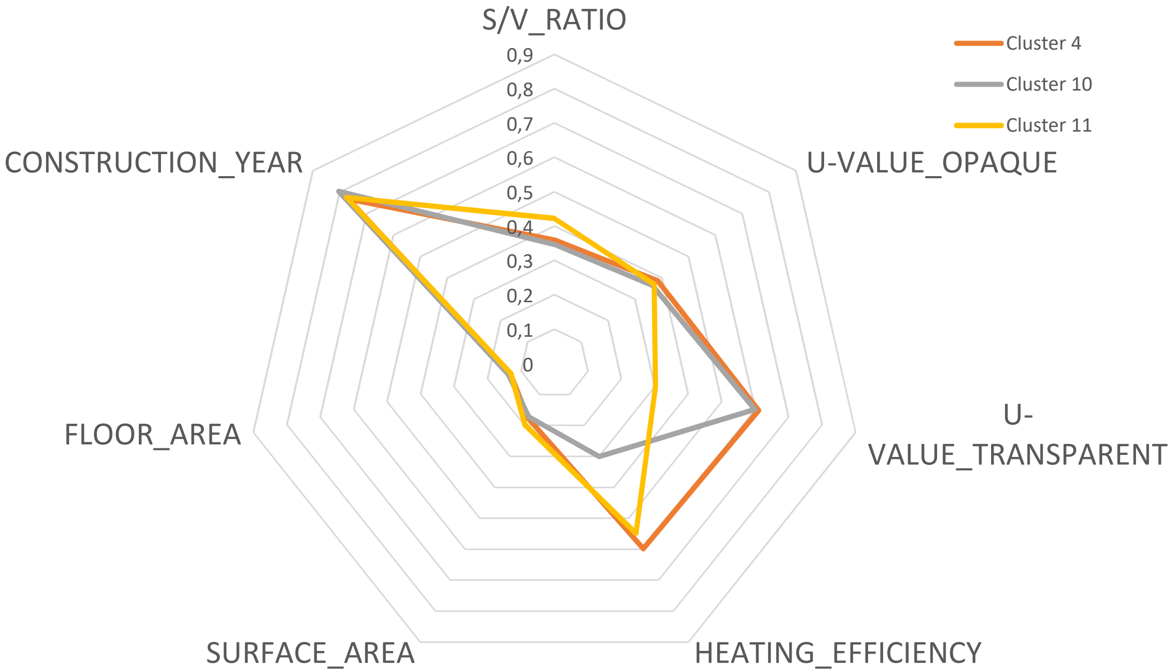




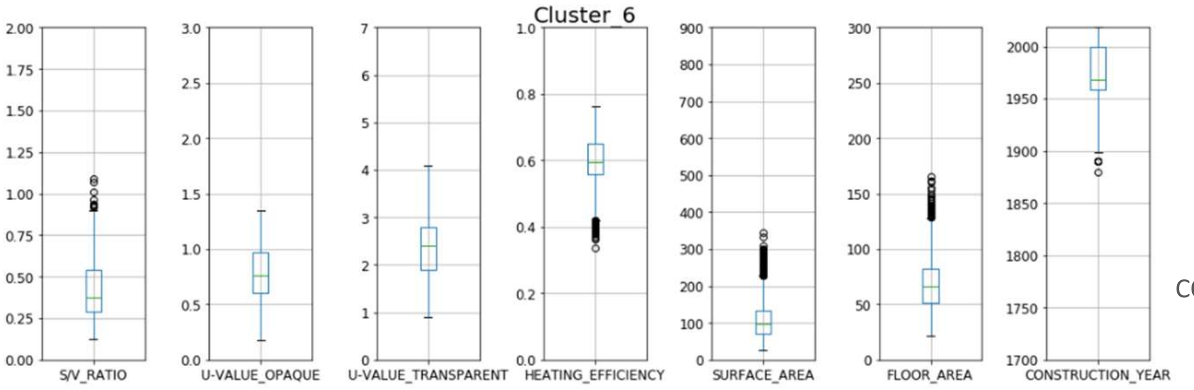


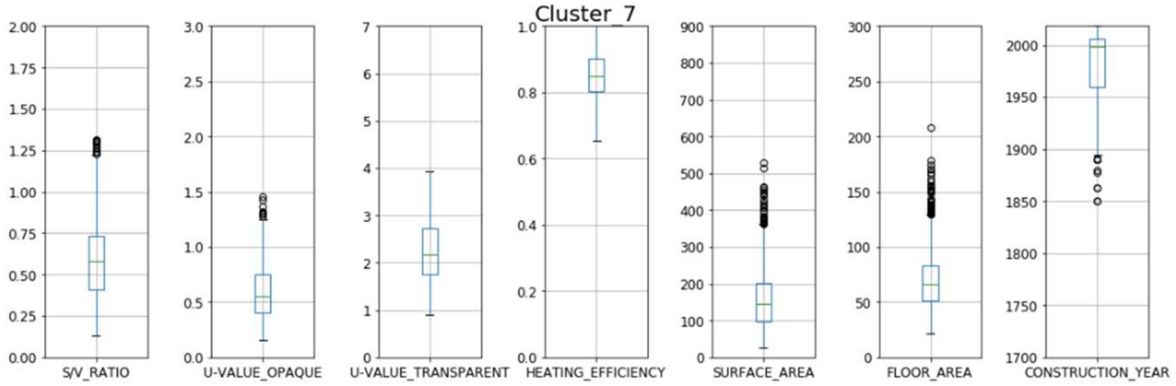
Clusters of APEs: Low energy performance

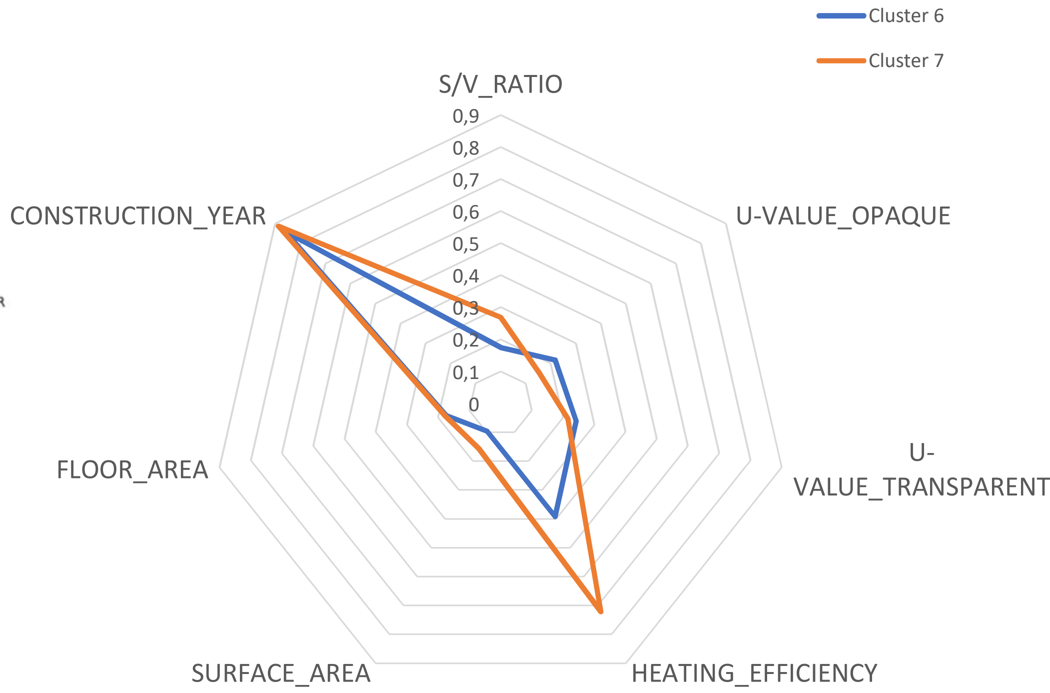




Clusters of APEs: High energy performance







Cluster characterization

Automatically extract knowledge from data, being **directly exploitable** by all stakeholders (including non-experts).

Association rule extraction

- **Exhaustive** analysis of all the possible **correlations**, above a given threshold, among values of the attributes characterizing the APE certificates
- Requires **discretization** for numeric attributes (data- / domain-driven)
- Can be performed at different **granularity** / aggregation levels (hierarchy definition)
- Qualitative indexes to select only the **most relevant** correlations
- **Transparent** self-describing model, directly “readable” by humans

X → Y	**{(Global Mean Efficiency = (0.85, 1.0]), (Average U-value of the vertical opaque envelope = (0.15, 0.45]), (Average U-value of the windows = (1.1, 1.85])} → {High Energy Performance}**

Knowledge visualization

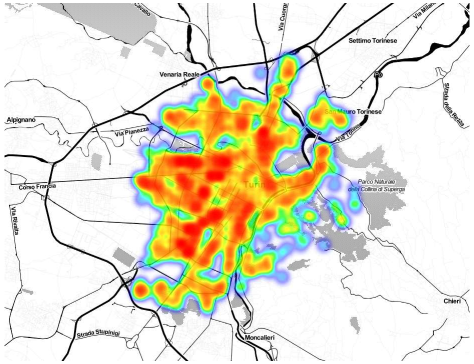
Maps with different spatial granularity

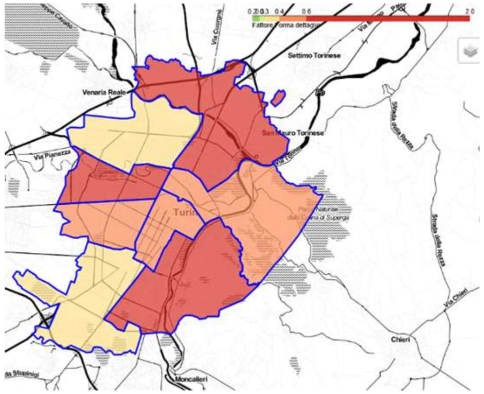
- City
- District
- Neighborhood
- Building

Different types of maps

Choropleth maps

- An aggregation metric is required
  - Majority model
  - Statistical functions to be defined with the domain expert





Knowledge visualization

Maps with different spatial granularity

- City
- District
- Neighborhood
- Building

Different types of maps

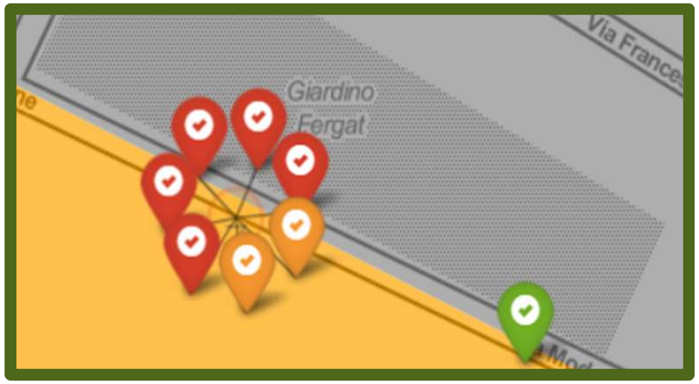
Choropleth maps

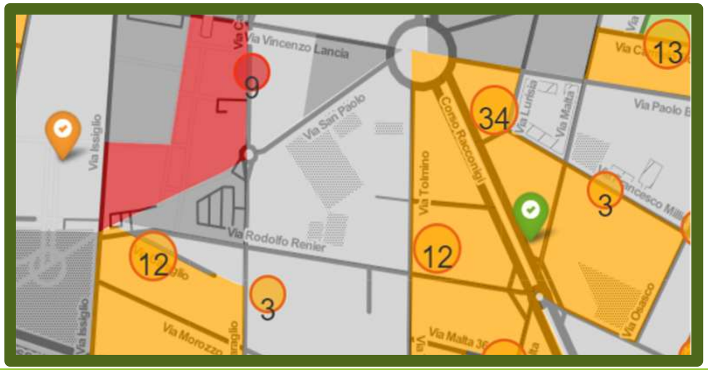
- An aggregation metric is required
  - Majority model
  - Statistical functions to be defined with the domain expert

Scatter maps with individual markers

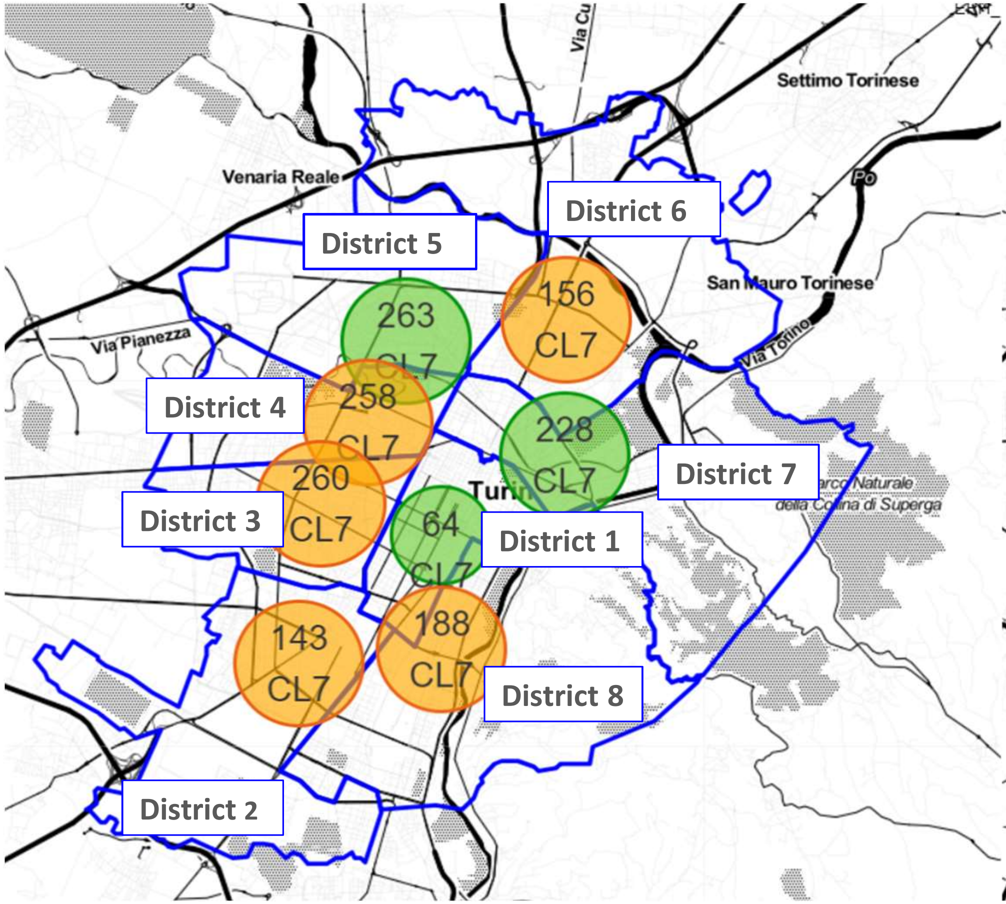
Maps with marker-clusters

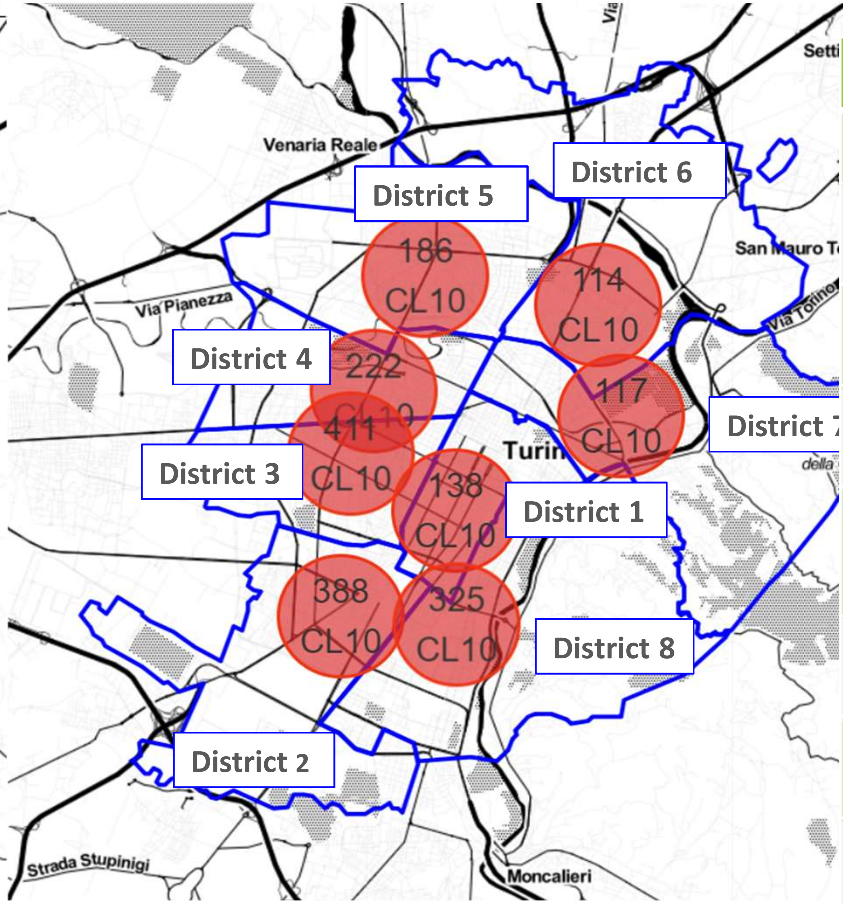
- Dynamic plots to model aggregated APEs





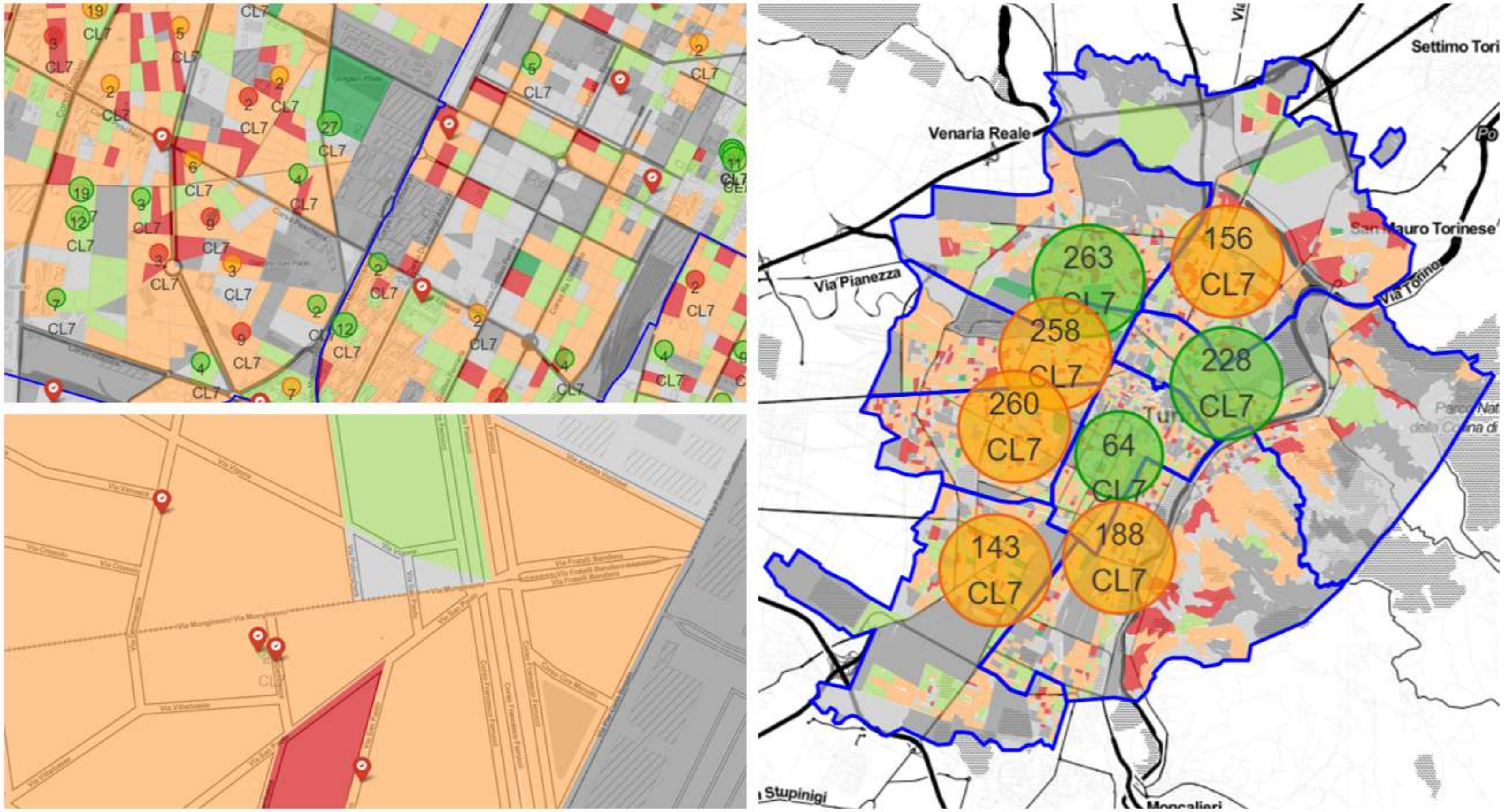
Maps with Marker-Cluster at district level



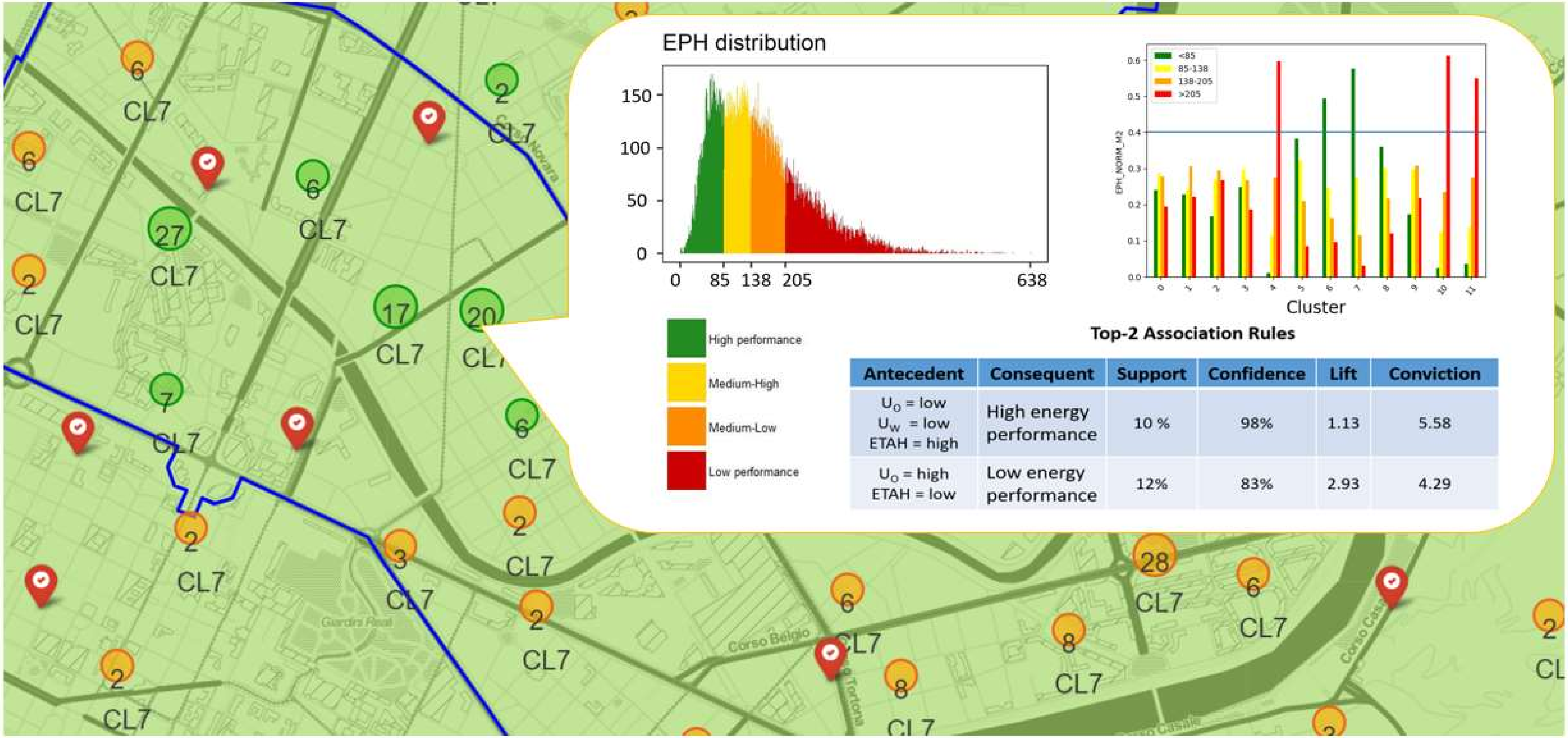


District Number	Name
District 1	• Centro • Crocetta
District 2	• Santa Rita • Mirafiori Nord • Mirafiori Sud
District 3	• Borgo San Paolo • Cenisia • Pozzo Strada
District 4	• San Donato • Campidoglio • Parella
District 5	• Borgo Vittoria • Madonna di Campagna • Barriera di Lanzo
District 6	• Barriera di Milano • Regio Parco • Barca
District 7	• Aurora • Vanchiglia • Borgata
District 8	• San Salvario • Cavoretto • Borgo Po

Maps with Marker-Cluster at different spatial granularity



Dashboard overview



Antecedent	Consequent	Support	Confidence	Lift	Conviction
U_O = low U_W = low ETAH = high	High energy performance	10 %	98%	1.13	5.58
U_O = high ETAH = low	Low energy performance	12%	83%	2.93	4.29

Work-in-progress activities

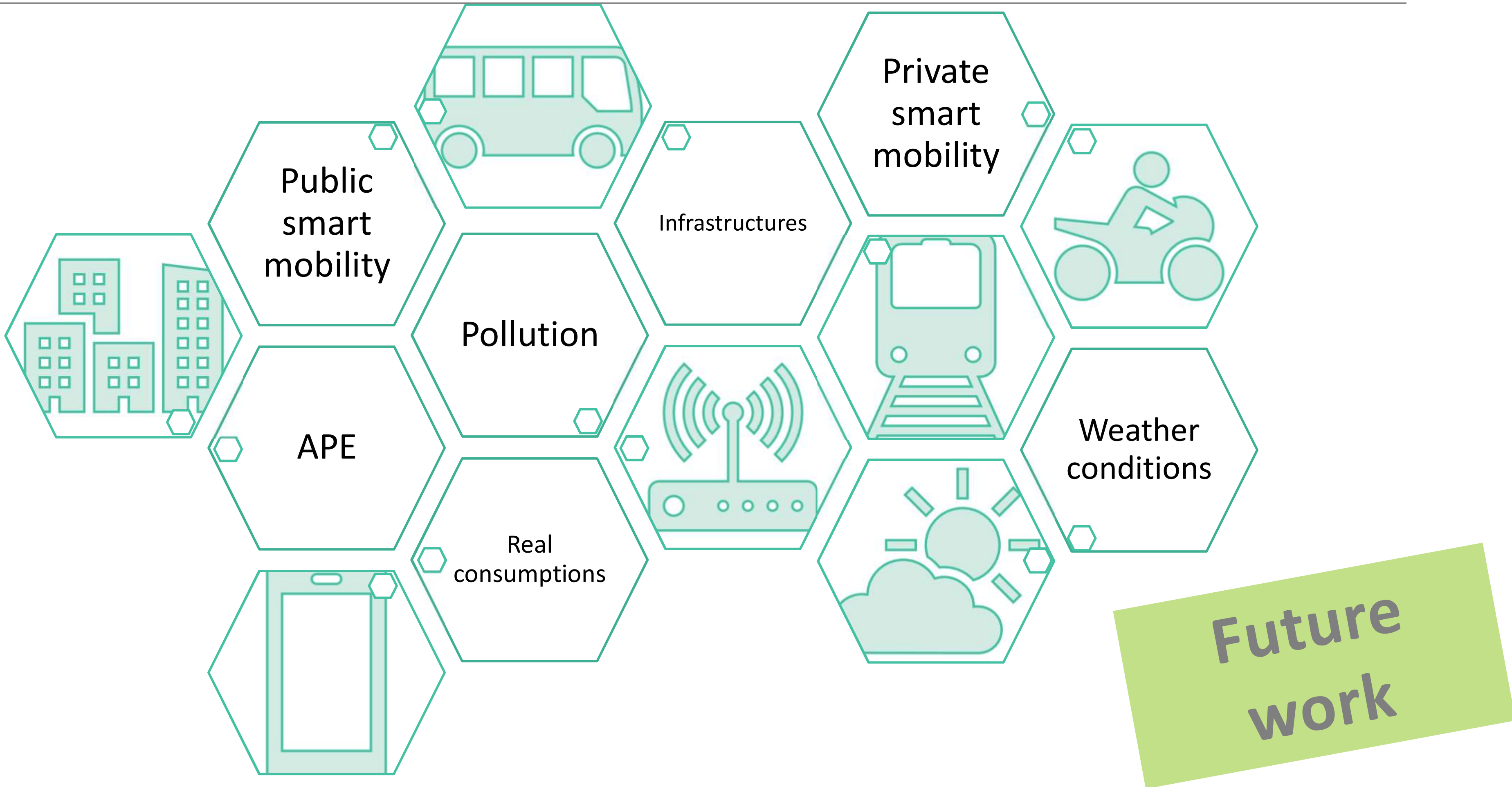
Exploitation of **supervised learning algorithms**

- to enhance the **data cleaning step**
- to include a larger number of APEs in the analysis

Generalization of the extracted knowledge

- through machine learning and statistical methods
- to provide a detailed overview at the city spatial granularity

Transparent and comprehensible cities



… questions?

Tania CERQUITELLI