POLITECNICO
DI TORINO

SmartData@PoliTO

Exploring open data to spread out knowledge: a real-world use case in the energy domain

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

Alfonso CAPOZZOLI
Department of Energy, Politecnico di Torino, Italy

Main reasearch objective

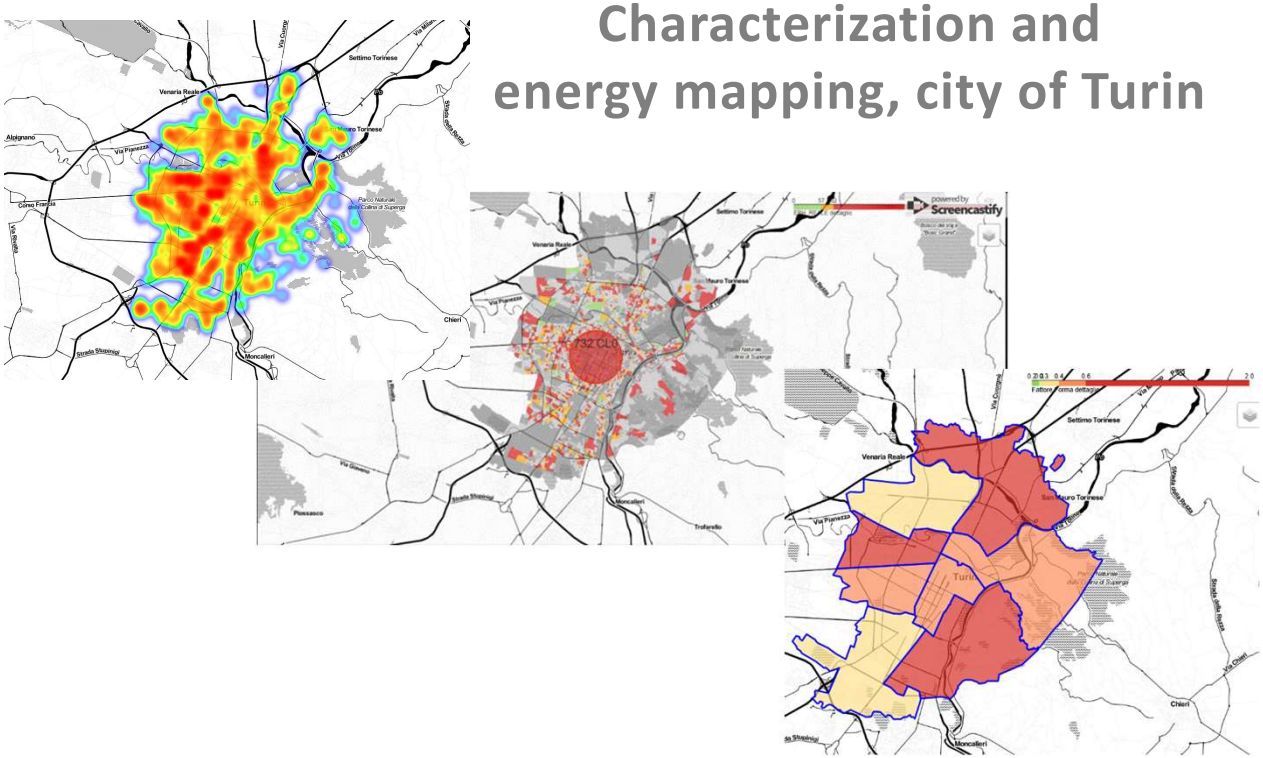
ENERGY DATA

OPEN DATA

Value for different stakeholders

Support and improve decisional processes

Characterization and energy mapping, city of Turin



eDISON EDF GROUP

POLITECNICO DI TORINO

SmartData@PoliTO

SmartData

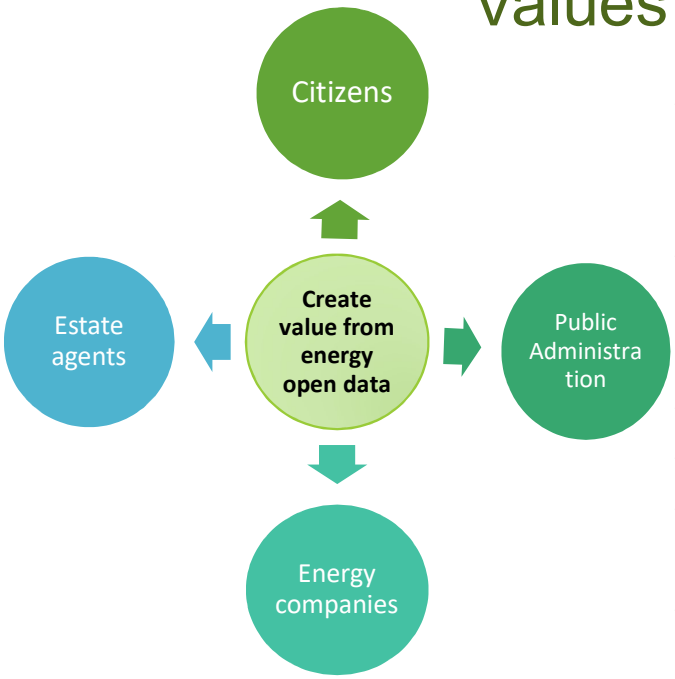
Main reasearch objective

ENERGY DATA

OPEN DATA

Value for different stakeholders

Support and improve decisional processes



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**

EDISON EDF GROUP

POLITECNICO DI TORINO

SmartData@PoliTO

SmartData

Structured and heterogenous energy-related data

Climatic data

- Dry bulb temperature
- Dew point temperature
- Pressure
- Total rainfall
- Humidity
- Total solar radiation
- Wind speed
- ...

Physical parameters

- Floor area
- Heat gross volume
- U-value
- Aspect ratio
- Window-to-wall ratio
- Orientation
- ...

Operational data

- Operational data of HVAC system (supply air temperature and fresh air flow rates)
- Indoor temperature
- Energy consumption
- Energy price
- Renewable energy production
- Indoor environmental quality parameters

User related data

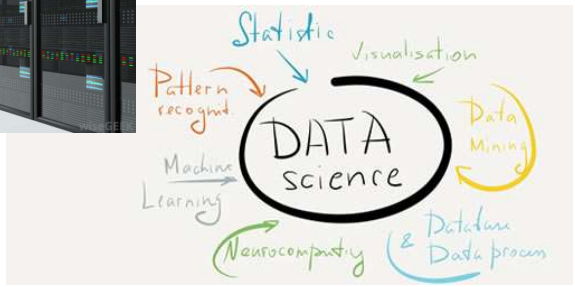
- Occupancy
- Number of occupants
- Occupant activities
- On/off appliances
- Opening/closing windows
- Social and economic factors
- ...

Time variables

- Season
- Month
- Date
- Day of the week
- Hour of day
- ...

Adapted from: Capozzoli A, Cerquitelli T, Piscitelli M S. *Enhancing energy efficiency in buildings through innovative data analytics technologies* (book chapter). Pervasive Computing: Next Generation Platforms for Intelligent Data Collection. Elsevier 2016. Editors: Ciprian Dobre, Fatos Xhafa.

KDD from energy data: two key roles



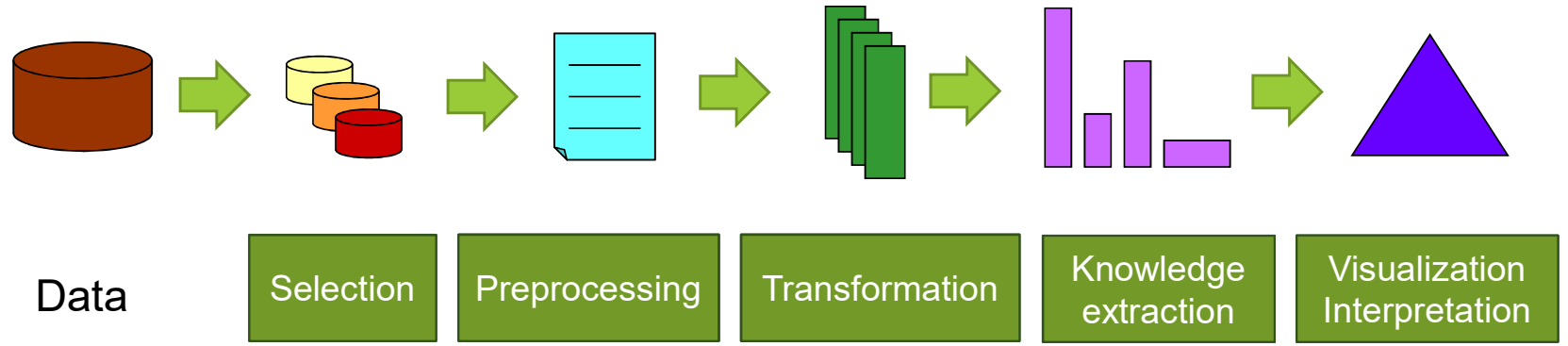
DATA SCIENTIST

- 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

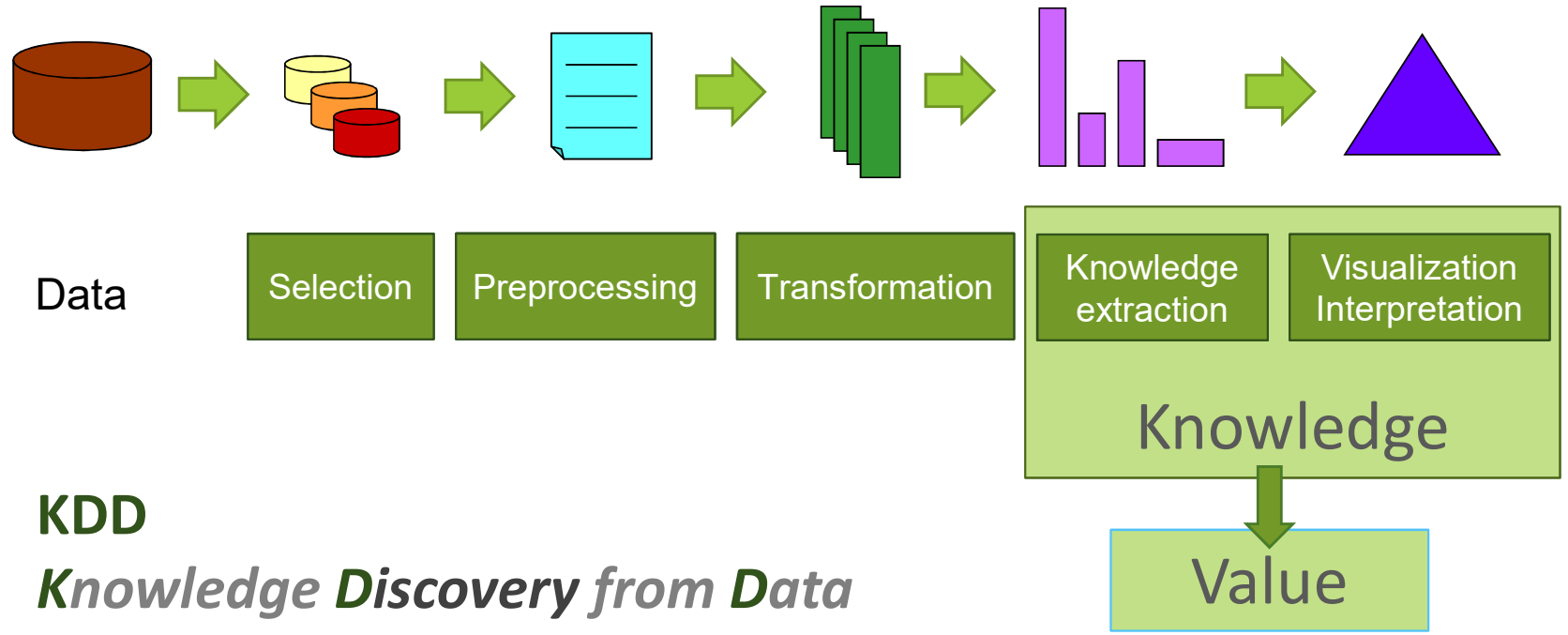
ENERGY SCIENTIST

- 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



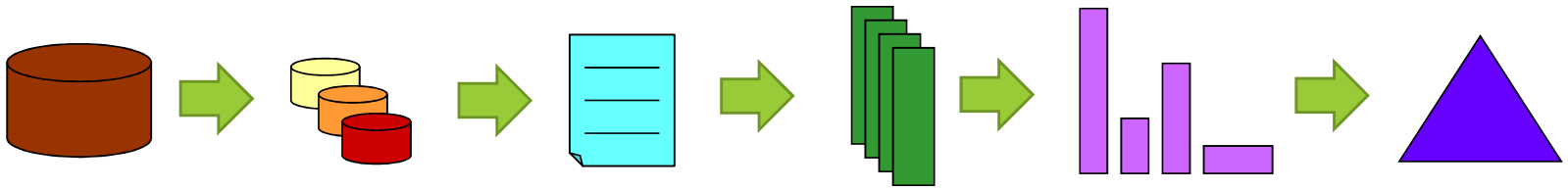
Knowledge extraction process



KDD

Knowledge Discovery from Data

Knowledge extraction process

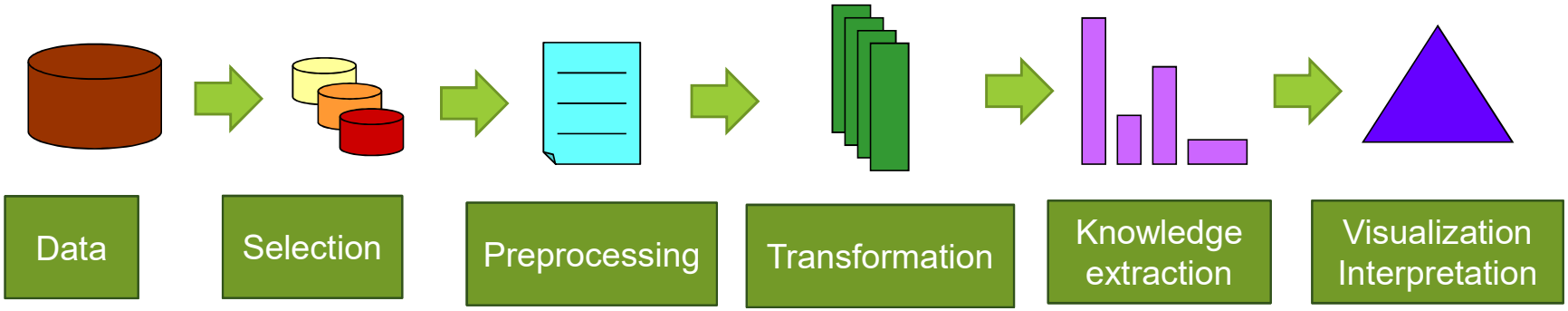


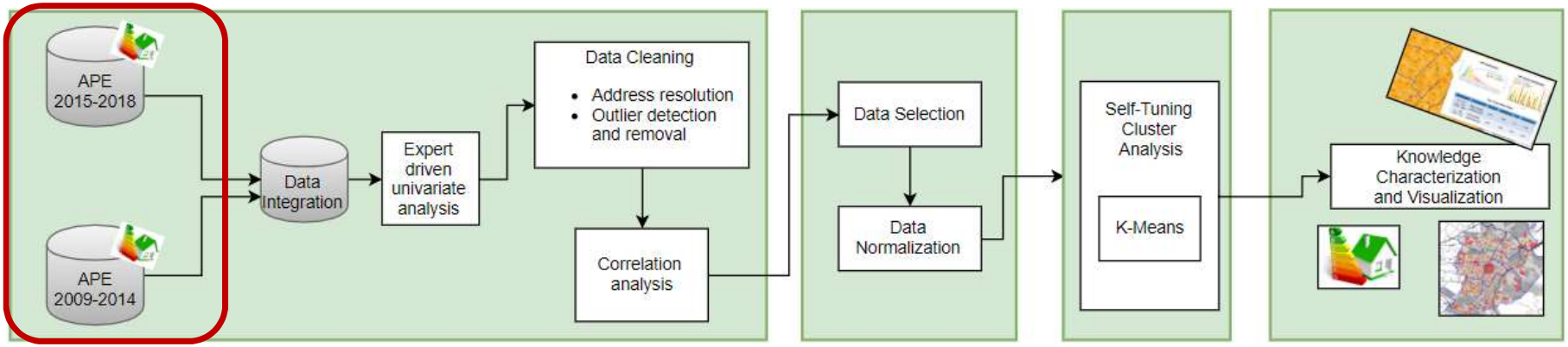
Data → Selection → Preprocessing → Transformation → Knowledge extraction → Visualization Interpretation

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 EPC





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 Performance Certificates

Energy analysis of the building

Wall and window features

Geometric features of the building

Hot water production

Cooling and heating energy needs

Type of plant

Emission impact

Renewable-energy production systems

Energy certificate officer

Qualified technicians granting EPC

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: EPC in Piedmont Region

Open data available on the Sistema Piemonte service system *
Each EPC 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

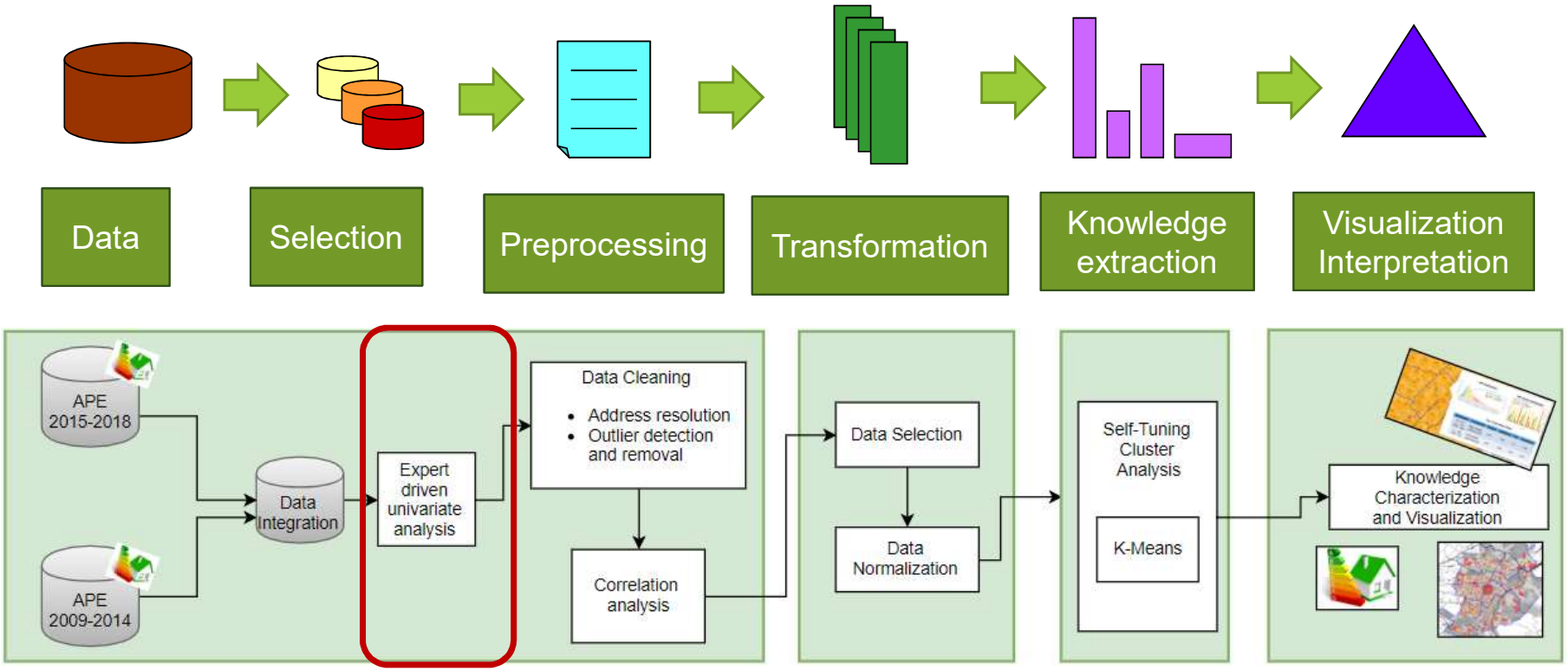
- Thermo-physical characteristics
- Geometric features
- Plant characteristics
- Energy performance

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

Knowledge extraction process from EPC



Expert-driven univariate analysis

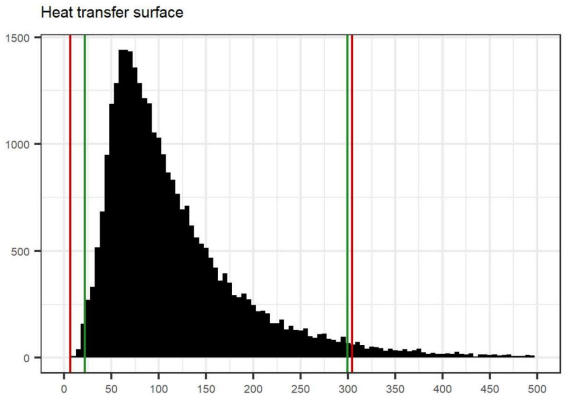
E1 (1) buildings used as permanent flats

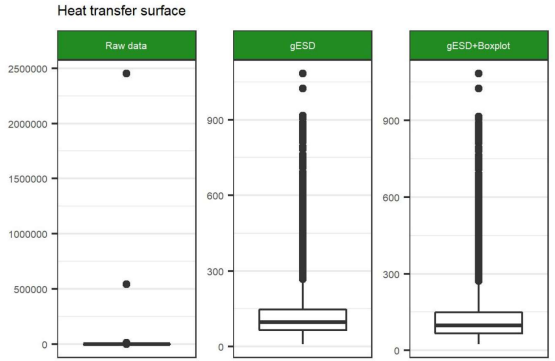
Identification of the most important variables → Identification of the validity ranges for each variable

- 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

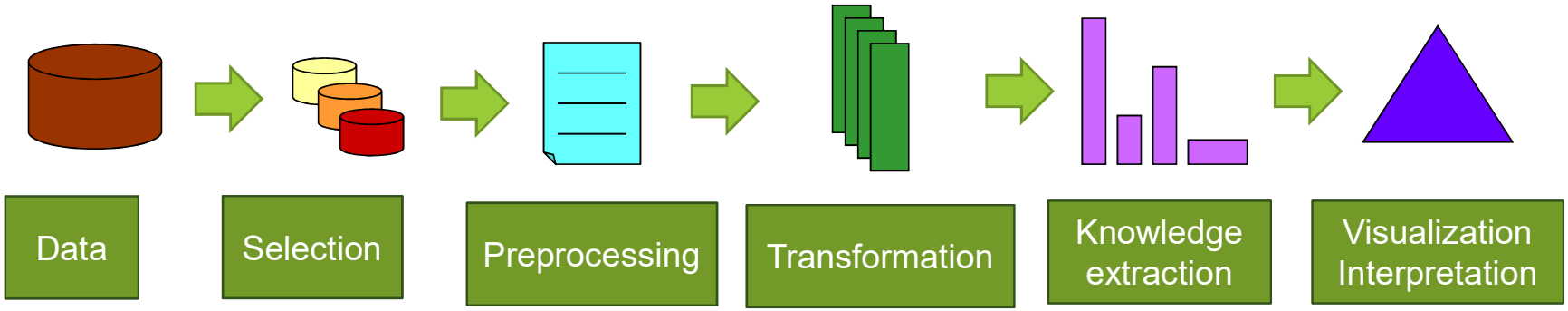
- **Knowledge of domain experts**
- **gESD method** needs as input parameter the upper-bound of potential outliers
- **Boxplot** visually displays a data distribution through its quartiles

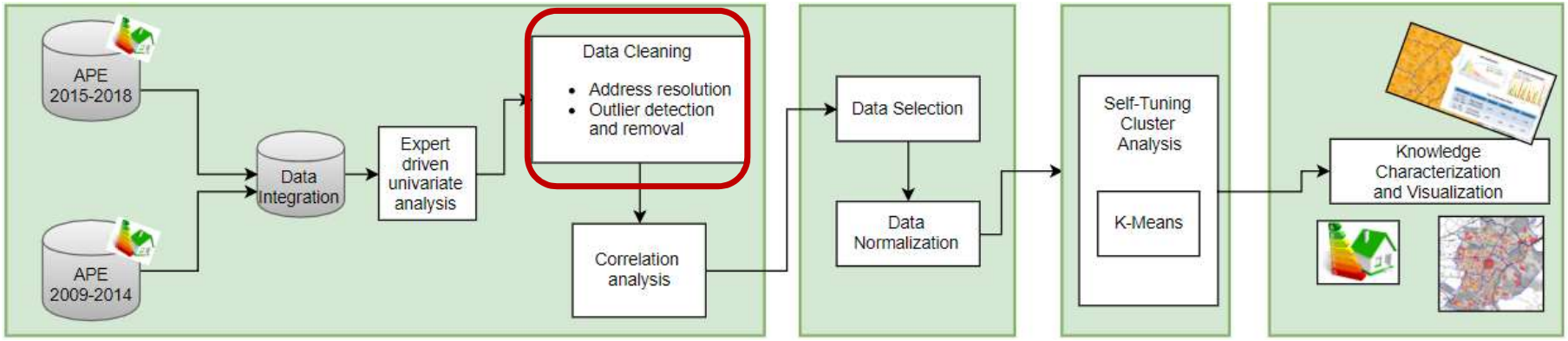




gESD = generalized Extreme Studentized Deviate

Knowledge extraction process from EPC





Data cleaning: address resolution

EPC 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

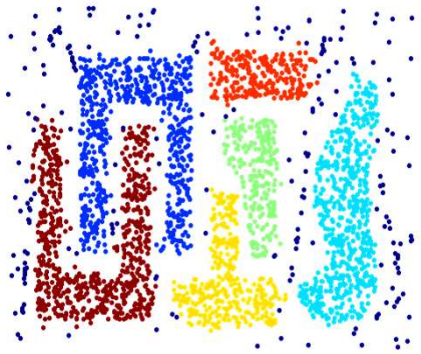
- Addresses in the DB have been **compared** to those stored in the **Turin road list** (from ***Geoportale Comune di Torino***[1])
- **Levenshtein** distance to compute the similarity index between the addresses reported in the APE DB and the reference DB.
  - If the address has been **resolved**, the 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

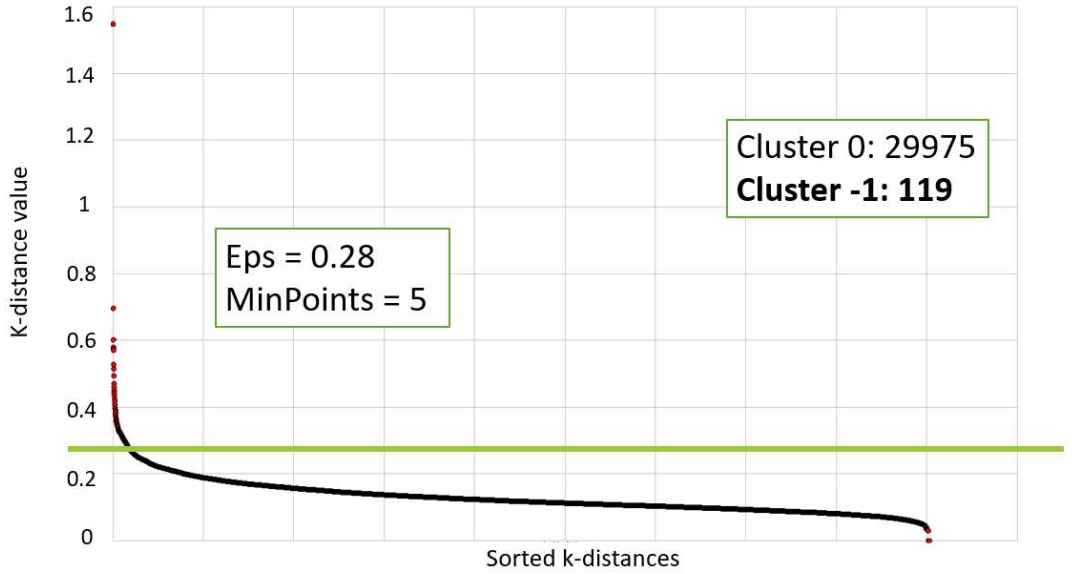
Density-based clustering algorithm: **DBScan**

- 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
- Self-tuning strategy based on k-dist plot
  - sorted distance of every point to its kth nearest neighbor

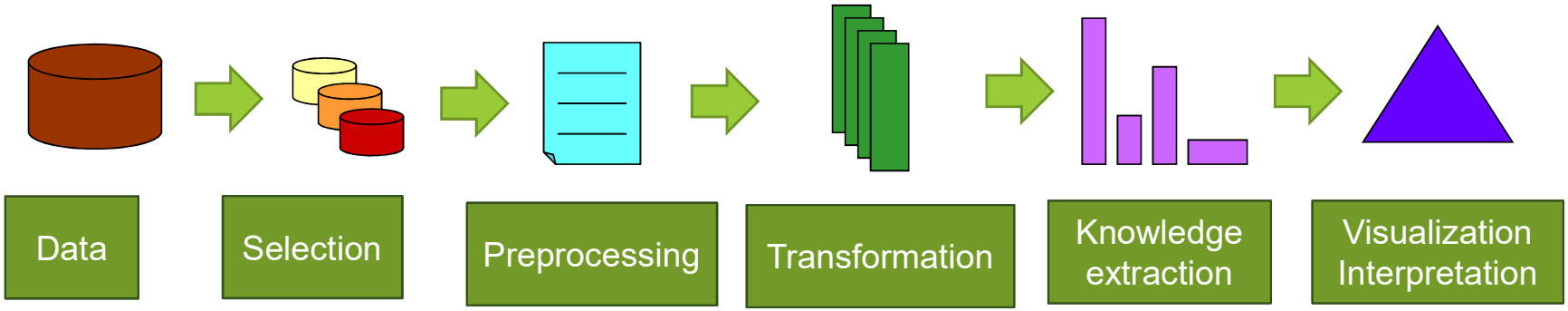


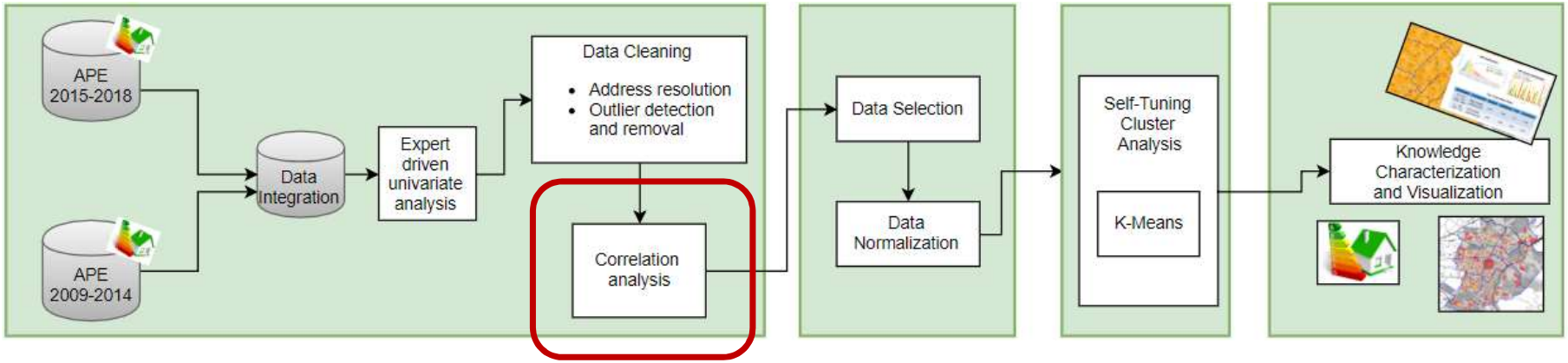
Clustering with DBScan

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



Knowledge extraction process from EPC

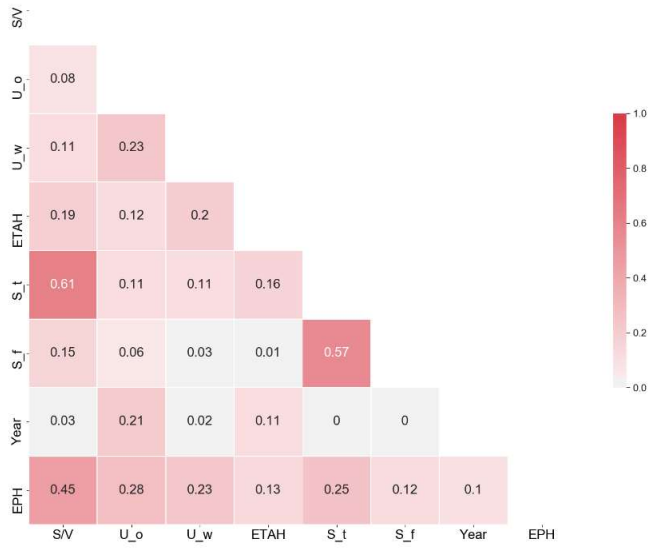




Correlation analysis

Feature **selection** and **removal** (correlation-based approach)

simplifying the model computation
improving the model performance

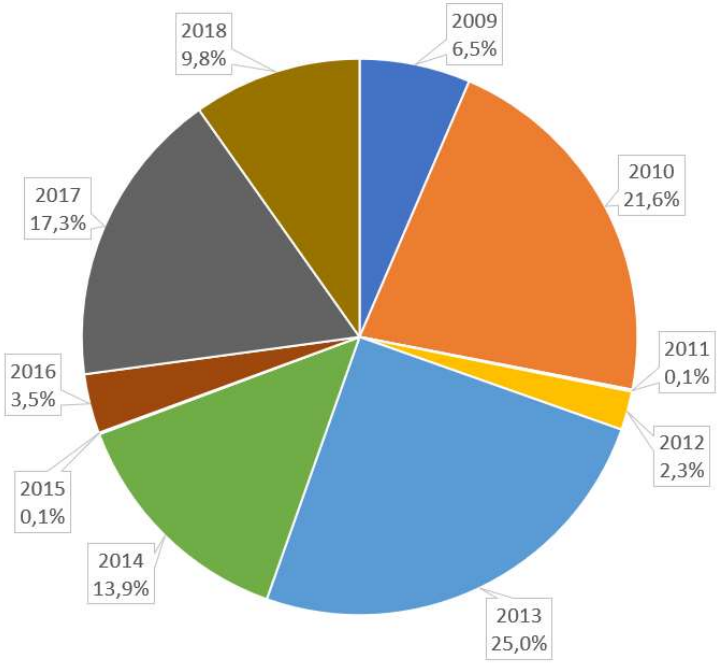


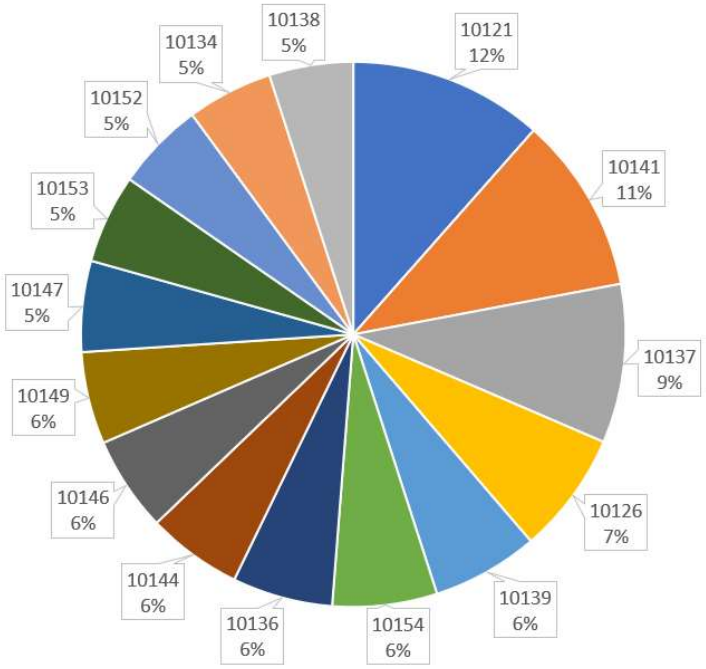
- **S/V** Aspect Ratio
- **U_o** Average u-value of opaque envelope
- **U_w** Average u-value of the windows
- **ETAH** Average global efficiency for spacing heating
- **S_t** Heat transfer surface
- **S_f** Floor Area
- **Year** Construction Year

- **EPH** Normalized primary heating energy consumption

Selected EPC

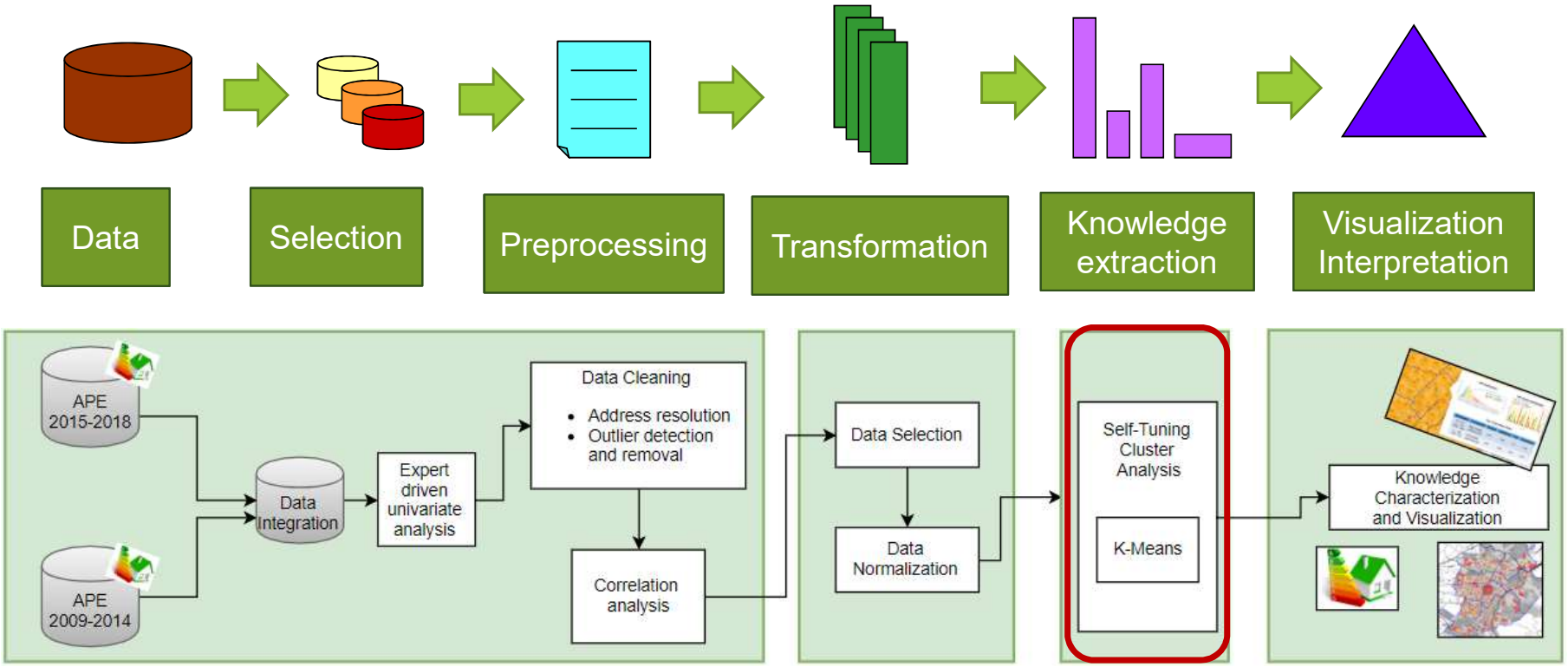
- E1 (1) buildings used as **permanent flats**
- EPC issued in the period: **2009 – 2018**
- EPC for ***particella, foglio*** *e* ***subalterno*** (identifying each single dwelling)
- **Number of selected EPC: ~30.000**





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

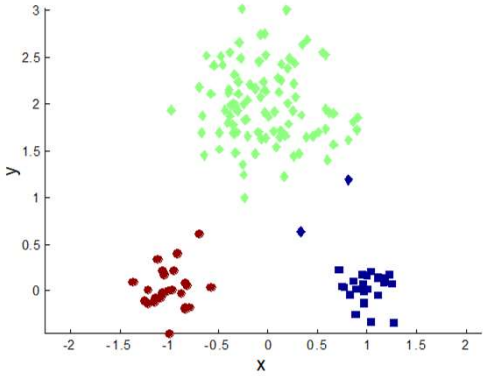
Knowledge extraction process from EPC



Self-tuning cluster analysis

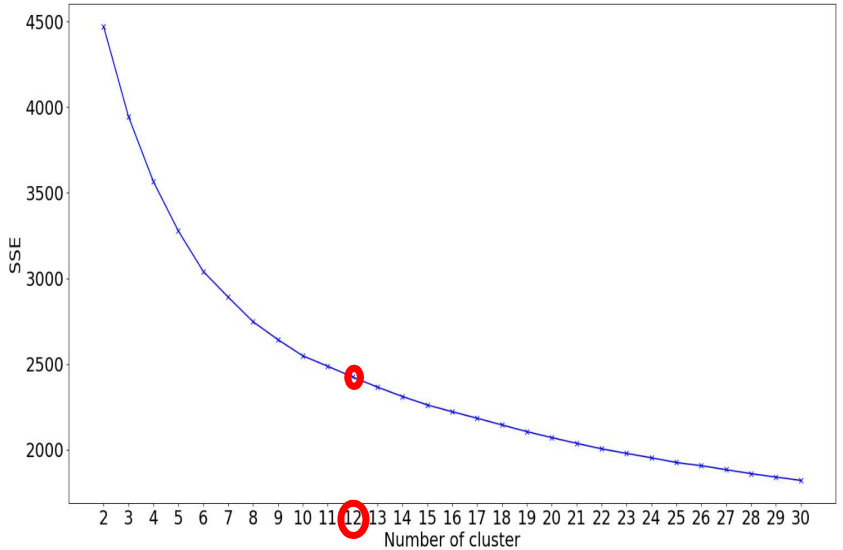
Clustering algorithms enriched by **self-tuning strategies** (i.e., parameter **autoconfiguration**)

- 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

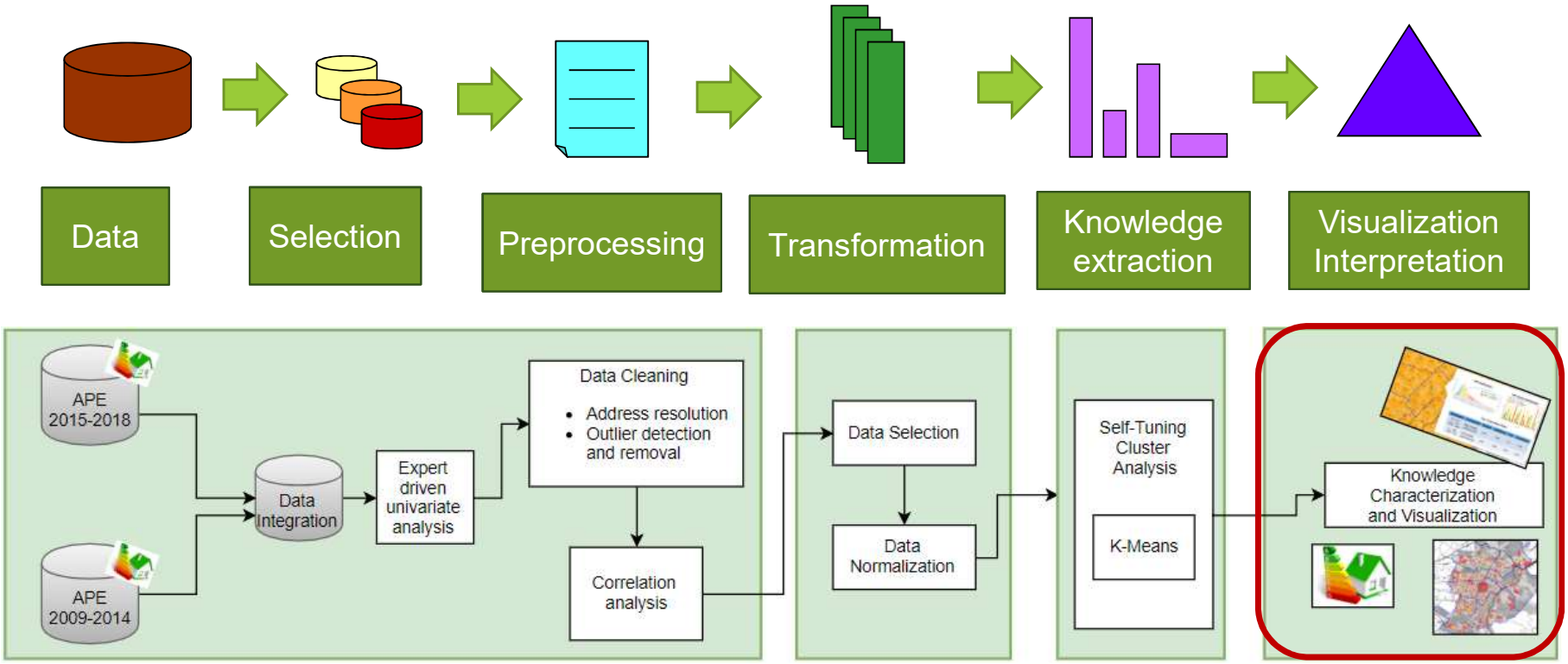


Optimal Clustering with K-Means

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



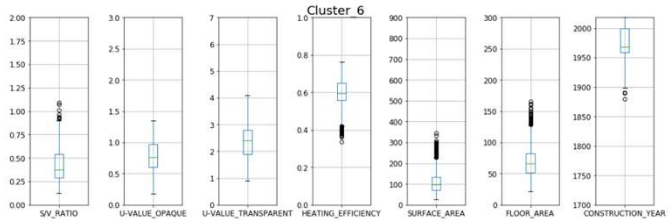
Knowledge extraction process from EPC

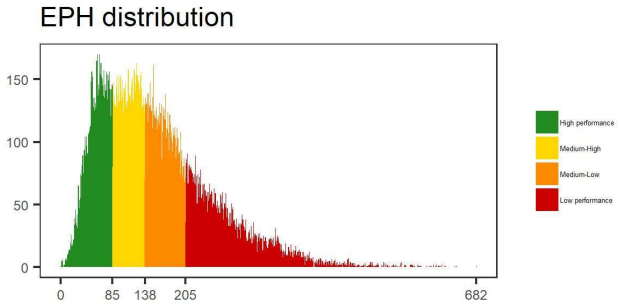


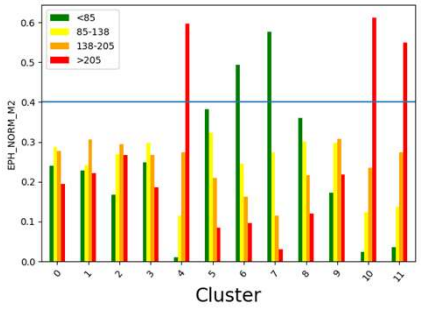
Knowledge characterization

Each discovered cluster of EPCs 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

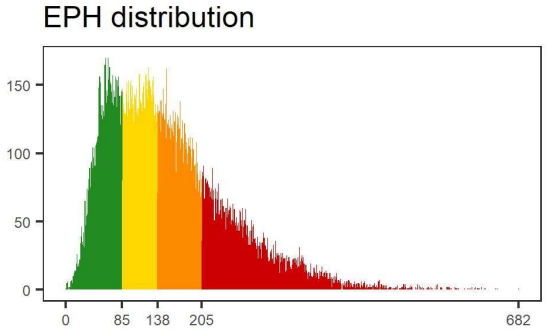


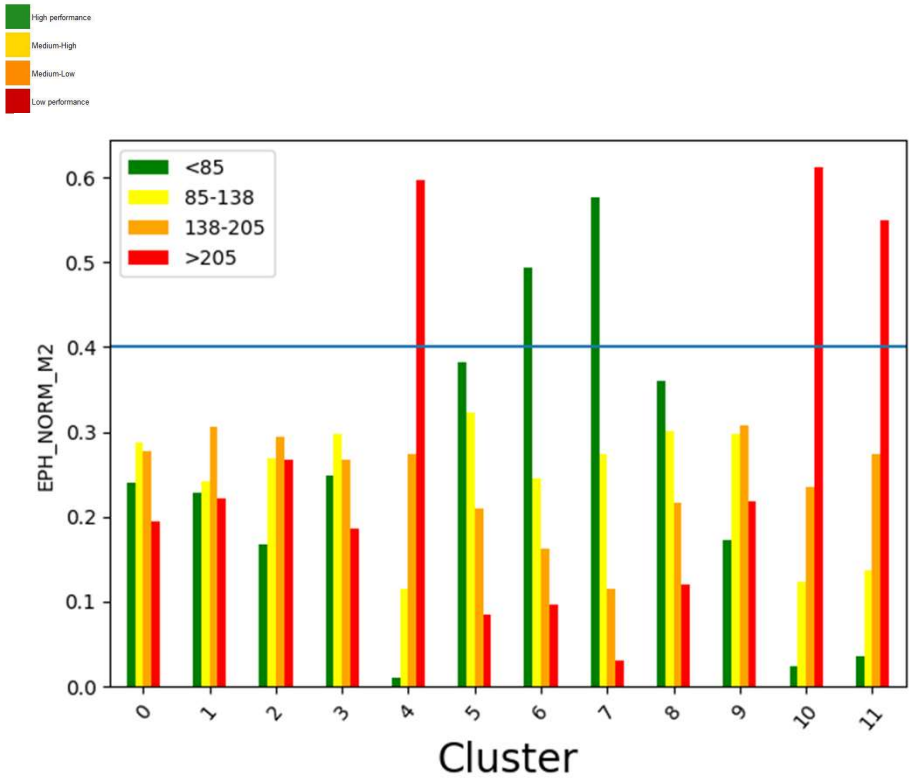




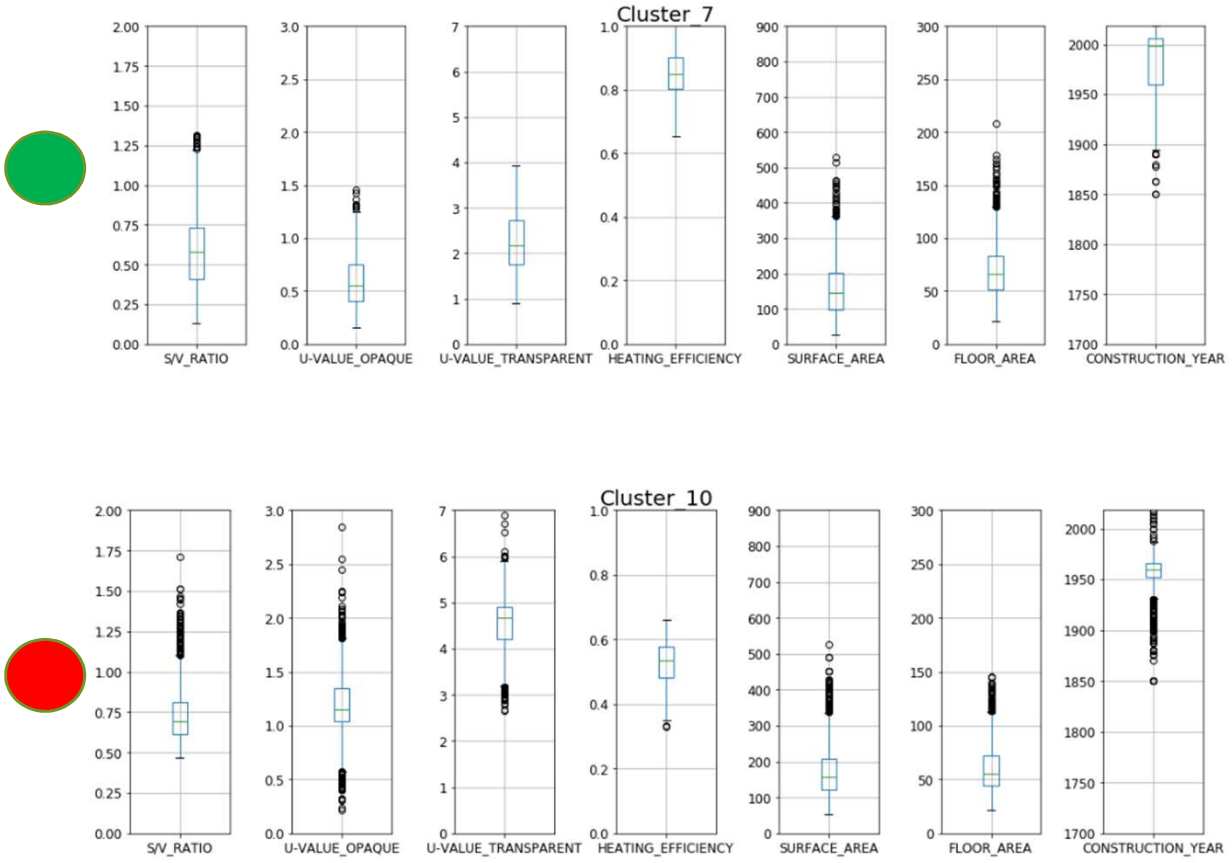
Cluster characterization

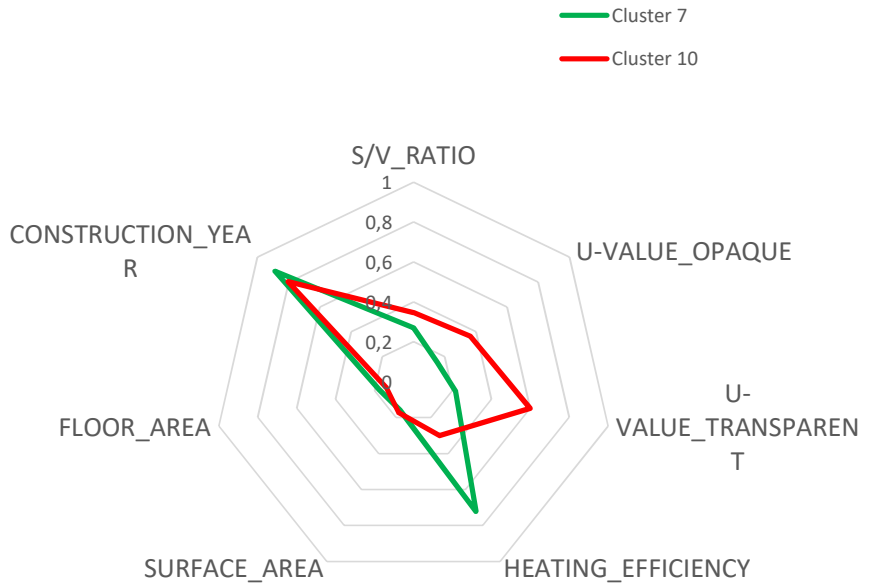
Cluster ID	APE #
Cluster 0	811
Cluster 1	4,321
Cluster 2	1,117
Cluster 3	3,988
Cluster 4	2,080
Cluster 5	2,723
Cluster 6	2,264
Cluster 7	1,723
Cluster 8	3,369
Cluster 9	3,418
Cluster 10	2,042
Cluster 11	2,119





Clusters of APEs: High vs Low energy performance





Knowledge visualization

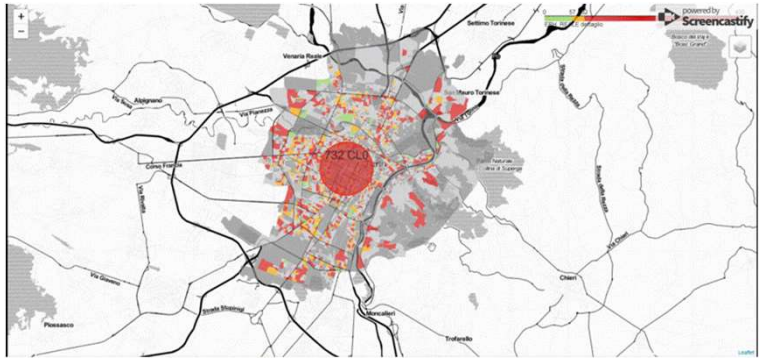
Maps with different spatial granularity

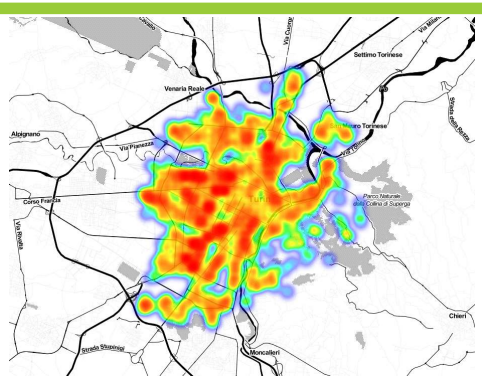
- City
- District
- Neighborhood
- Building

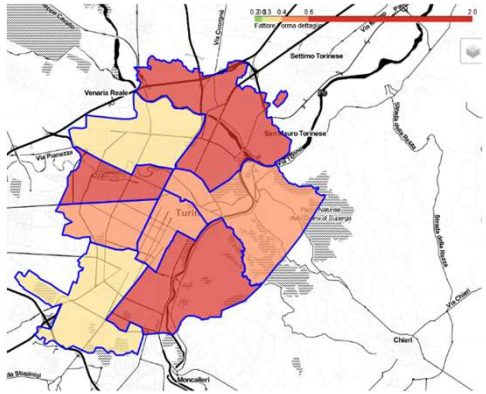
Different types of maps

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

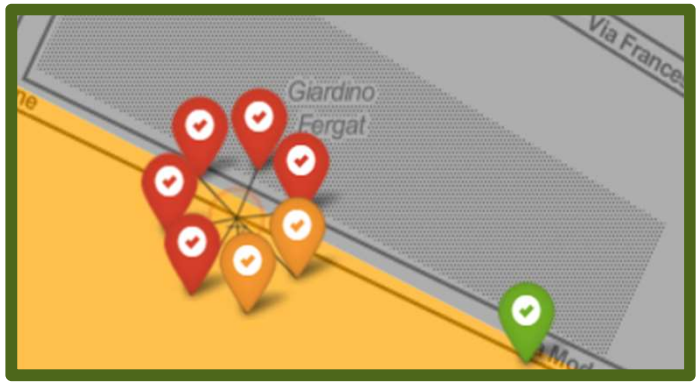
Choropleth maps

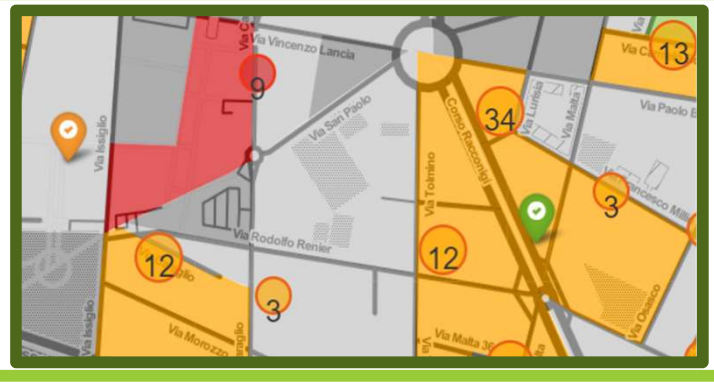
- An aggregation metric is required
  - Majority model
  - Statistical functions to be defined with the domain expert

Scatter maps with individual markers

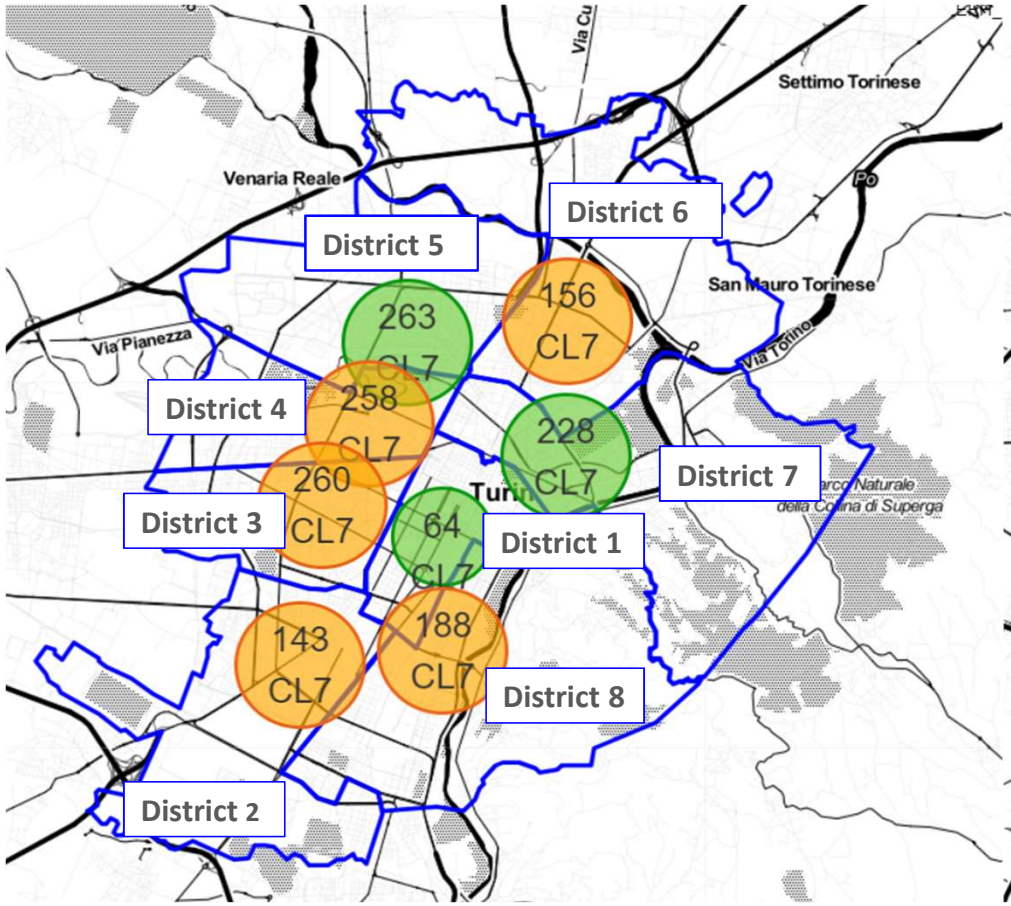
Maps with marker-clusters

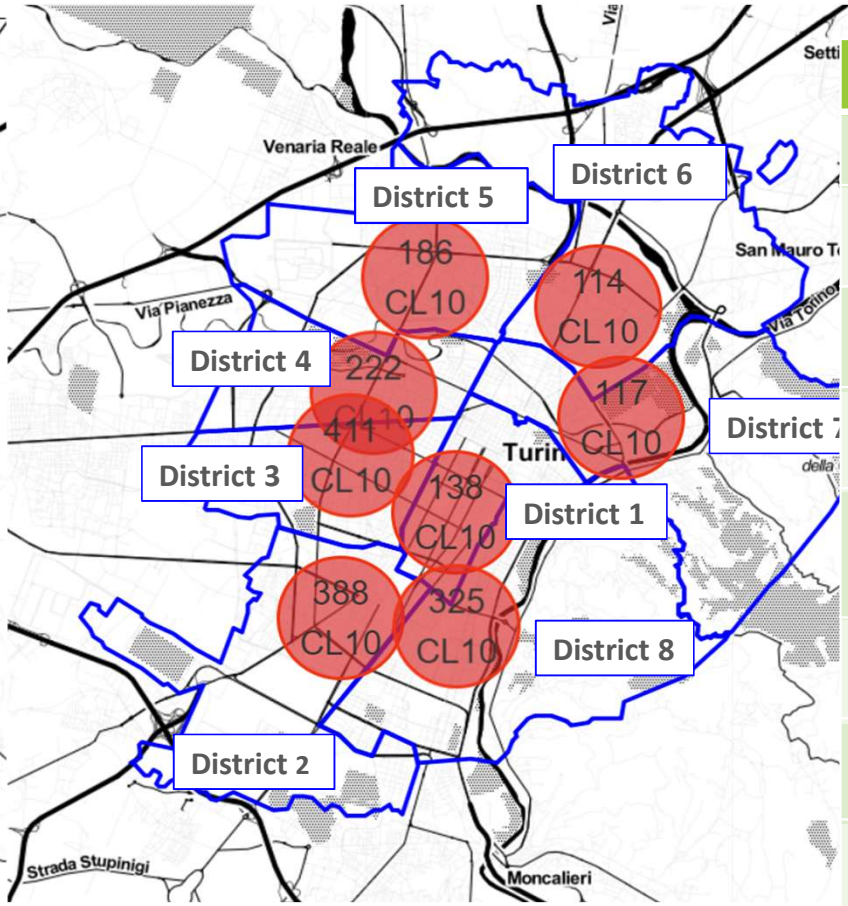
- Dynamic plots to model aggregated EPCs





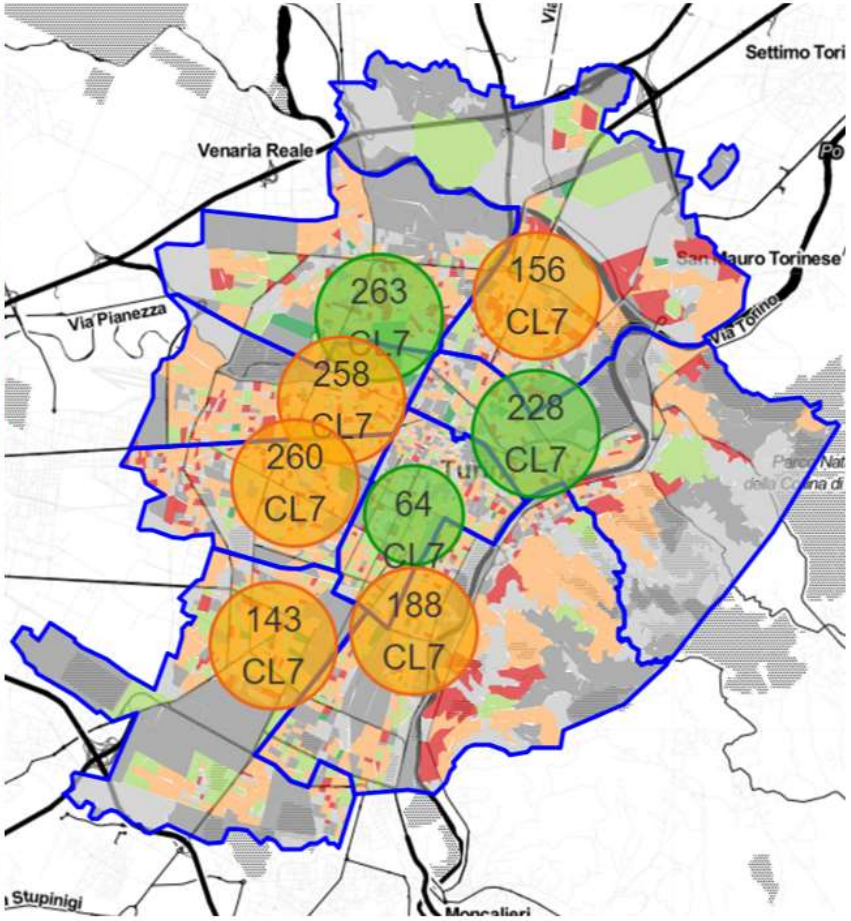
Maps with Marker-Cluster at district level





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



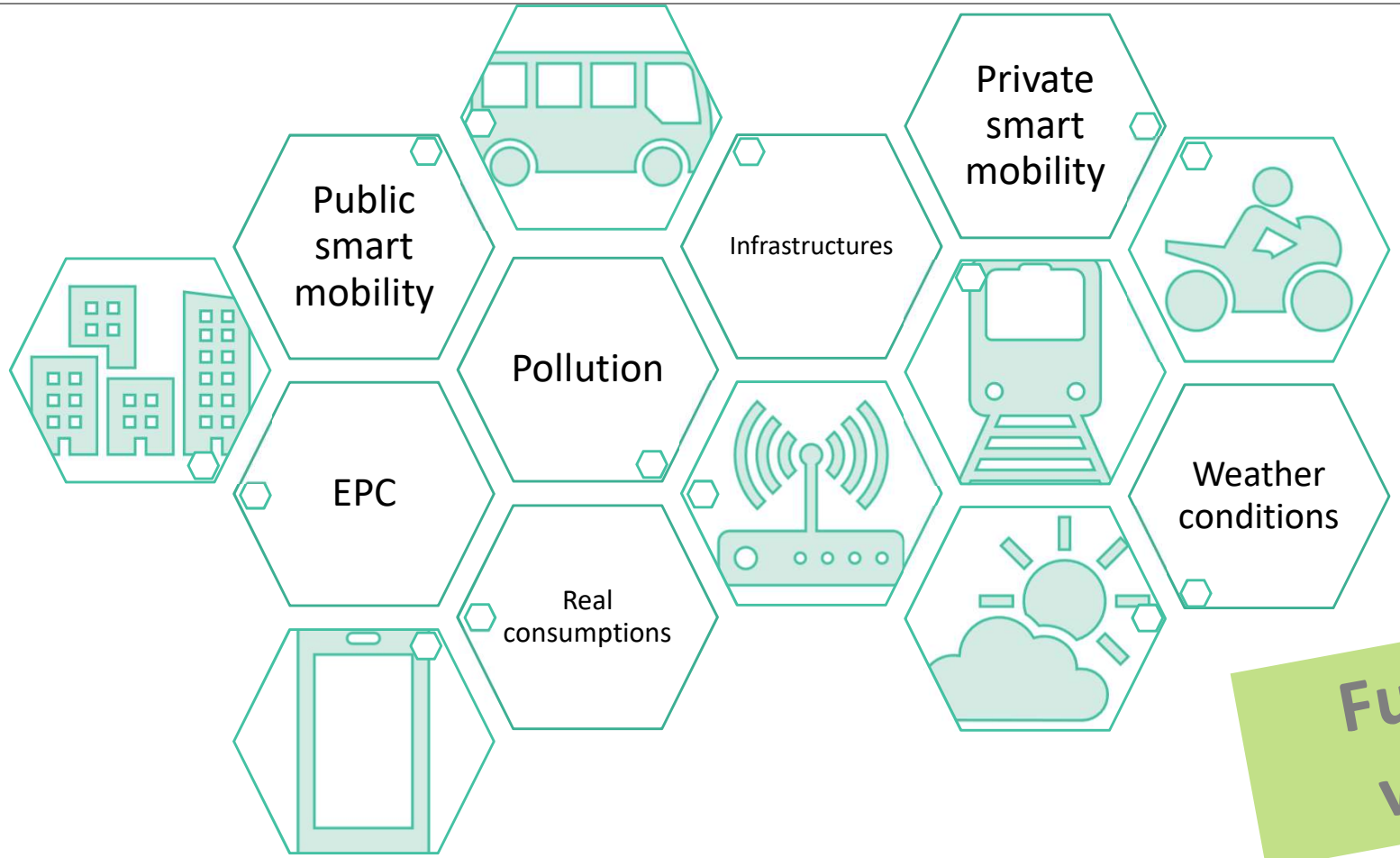
Work-in-progress activities

Tailoring of the informative dashboards to different stakeholders

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

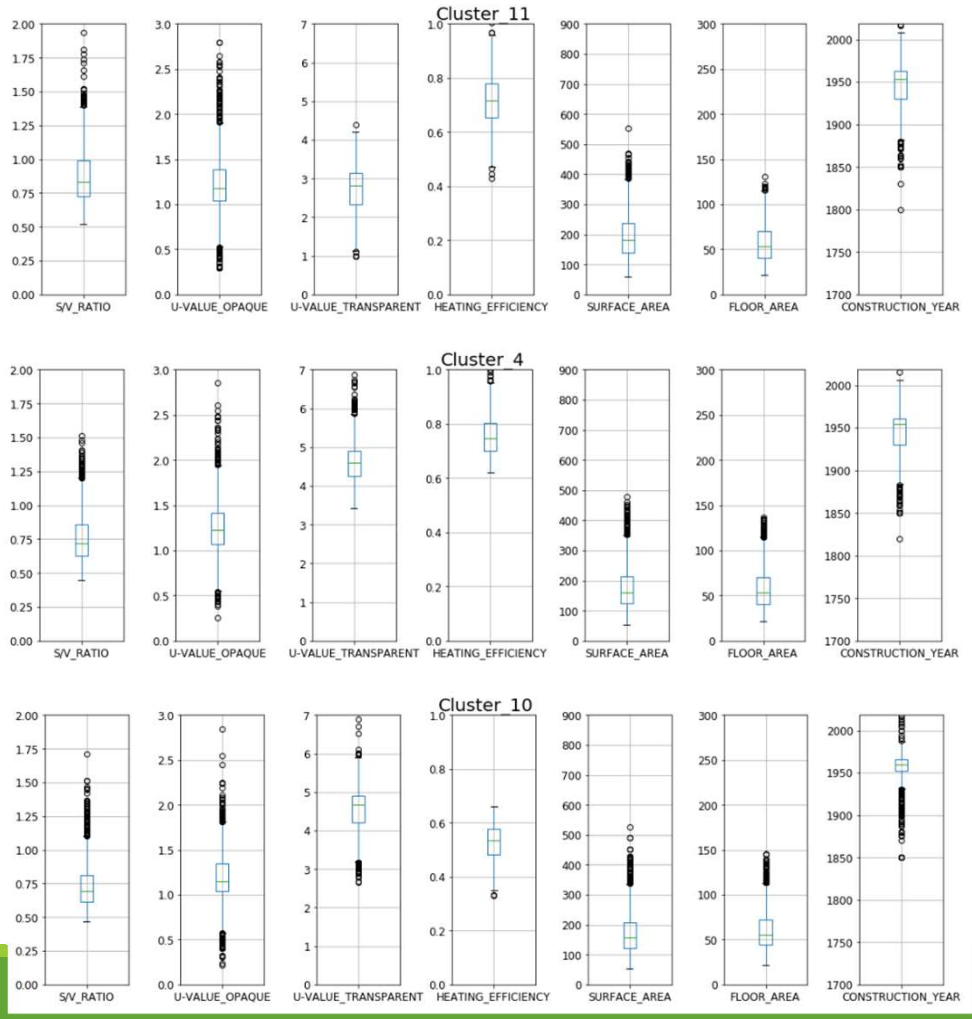


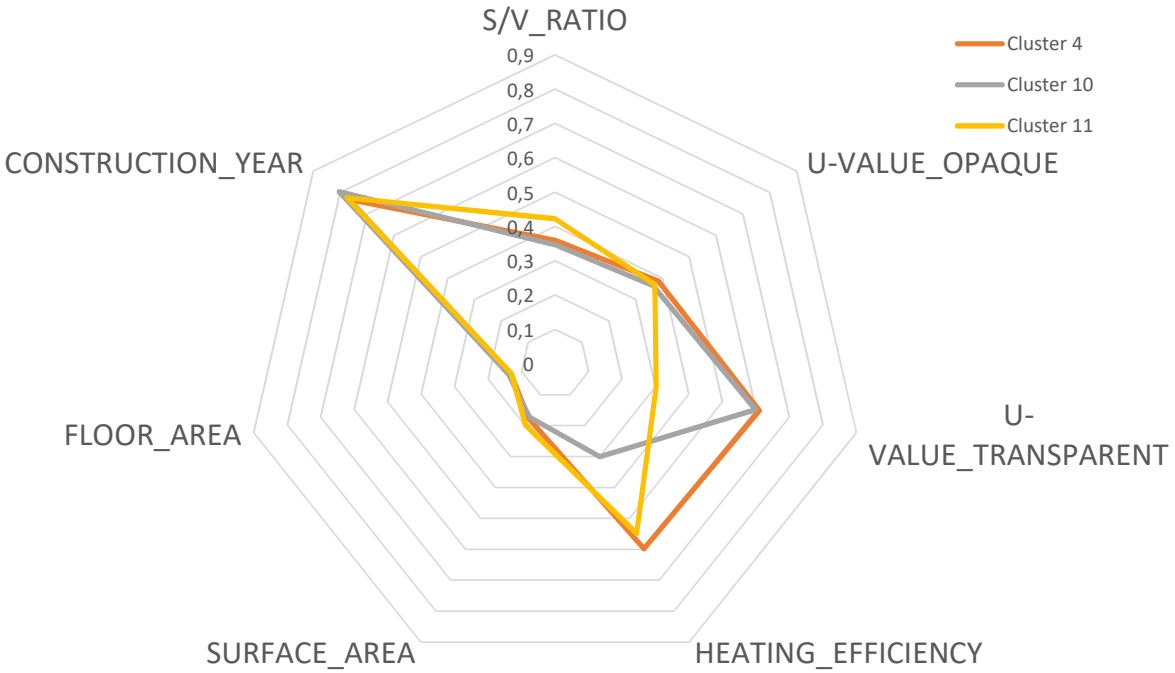
Future work

... questions?

Tania CERQUITELLI and Alfonso CAPOZZOLI

Clusters of APEs: Low energy performance





Clusters of APEs: High energy performance