Data Science for Smart Cities

CE88

Prof: Alexei Pozdnukhov

Course overview

Cities are complex systems

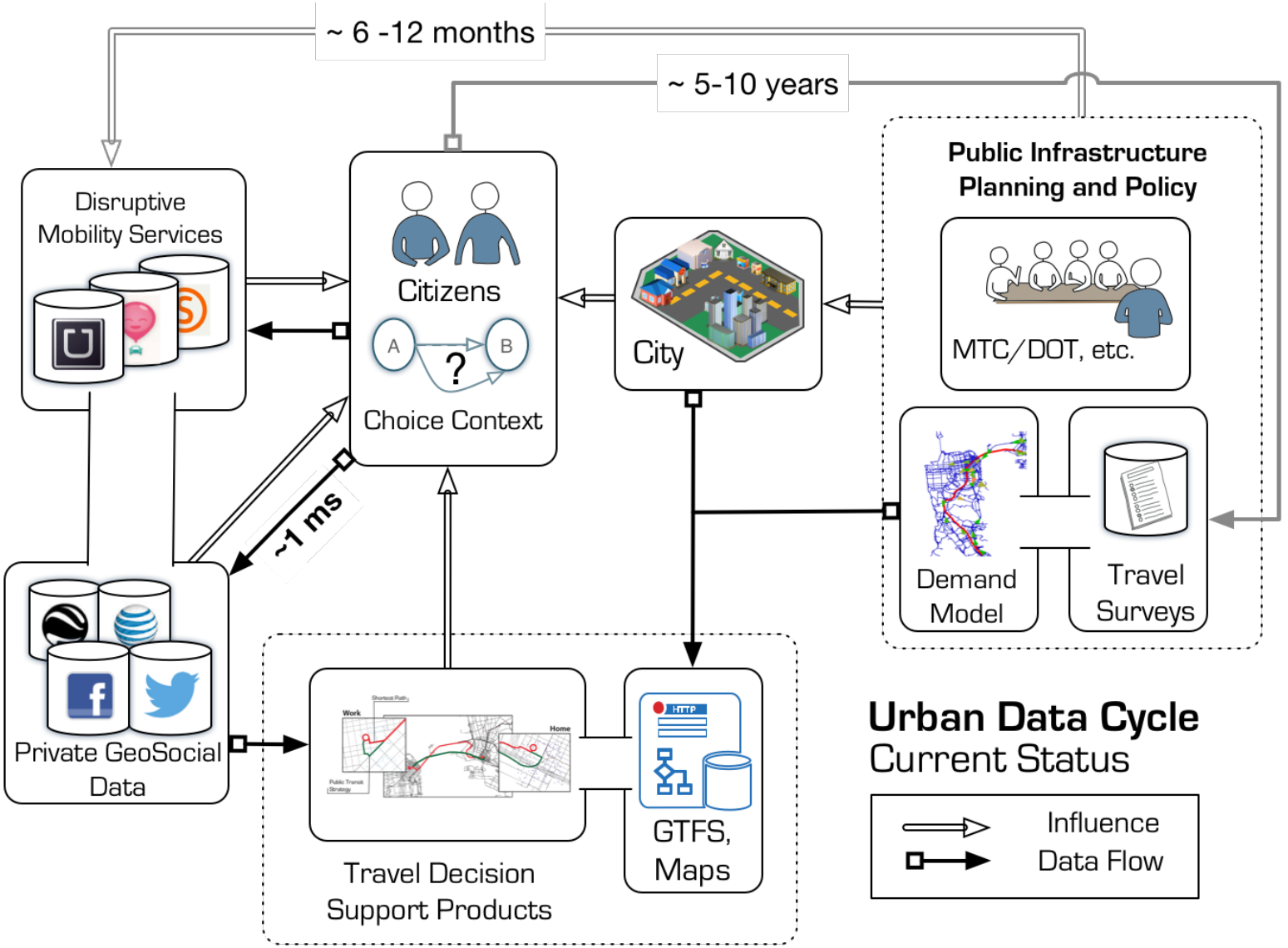
Urban data collection, handling and processing.

Data exploration and analysis. Demand, supply and impact.

Modeling and forecasting. Uncertainty.

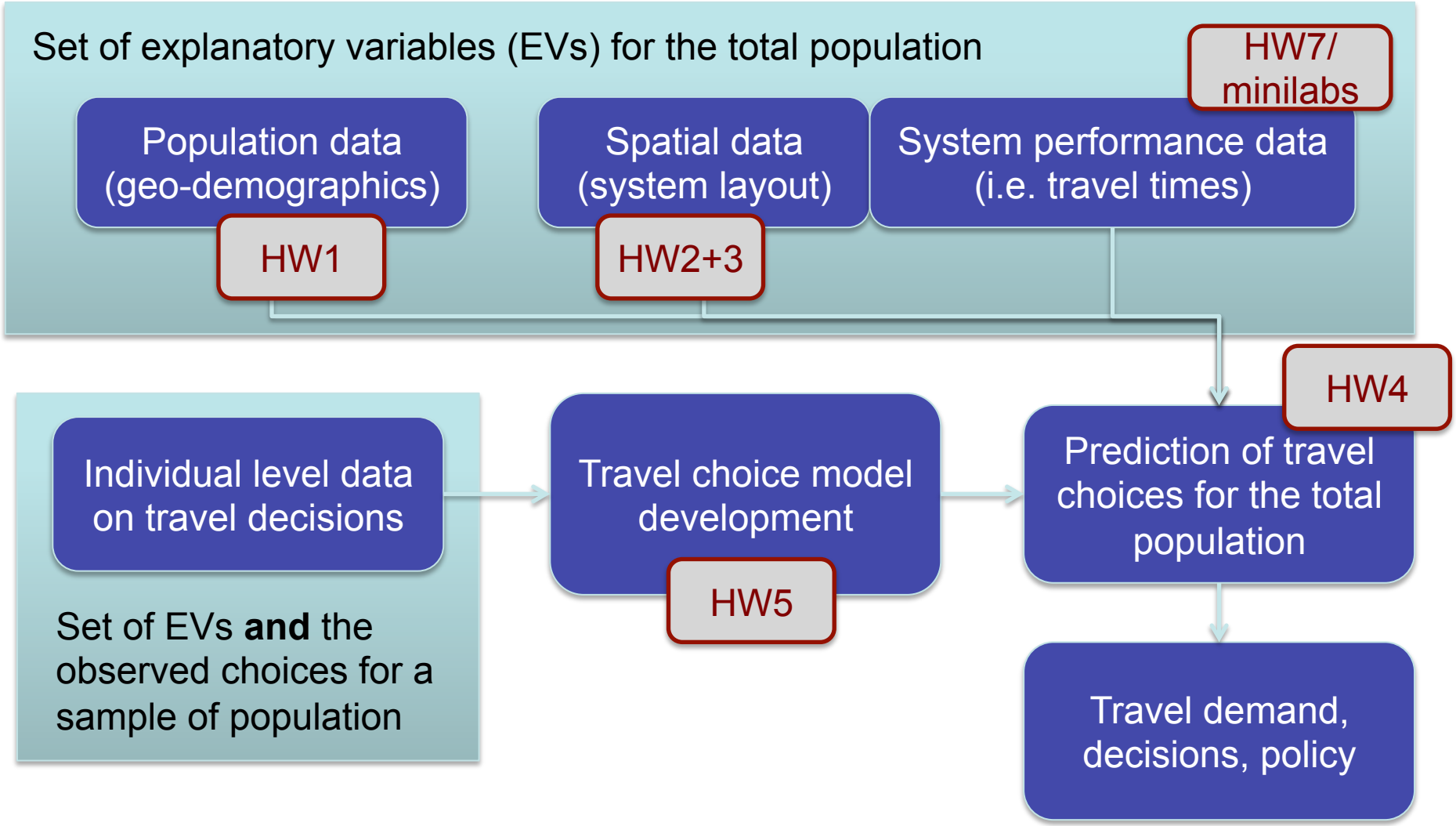
Decision making, planning and governance.

Urban Data Ecosystem

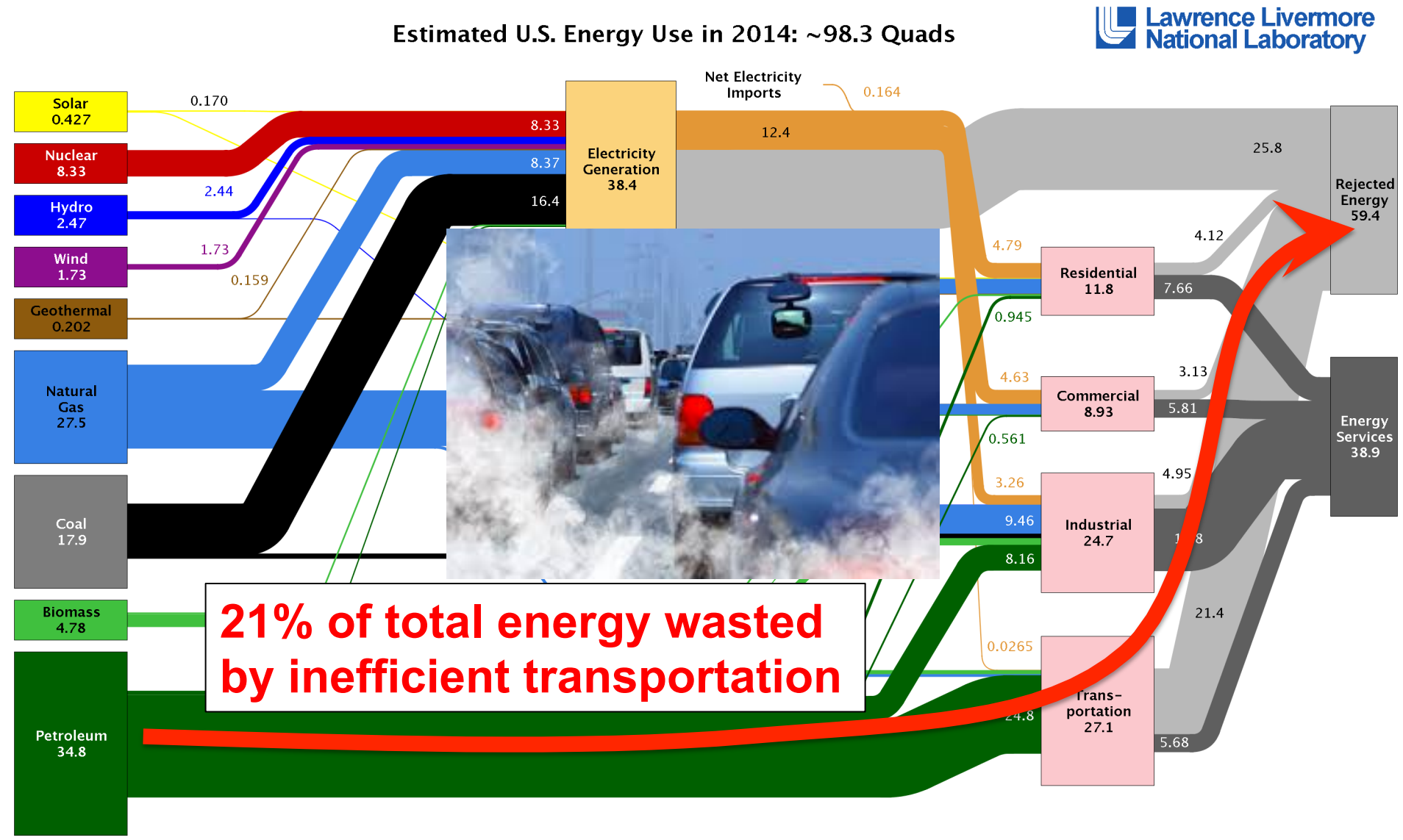


Urban data for decision support

Travel behaviors as an example:

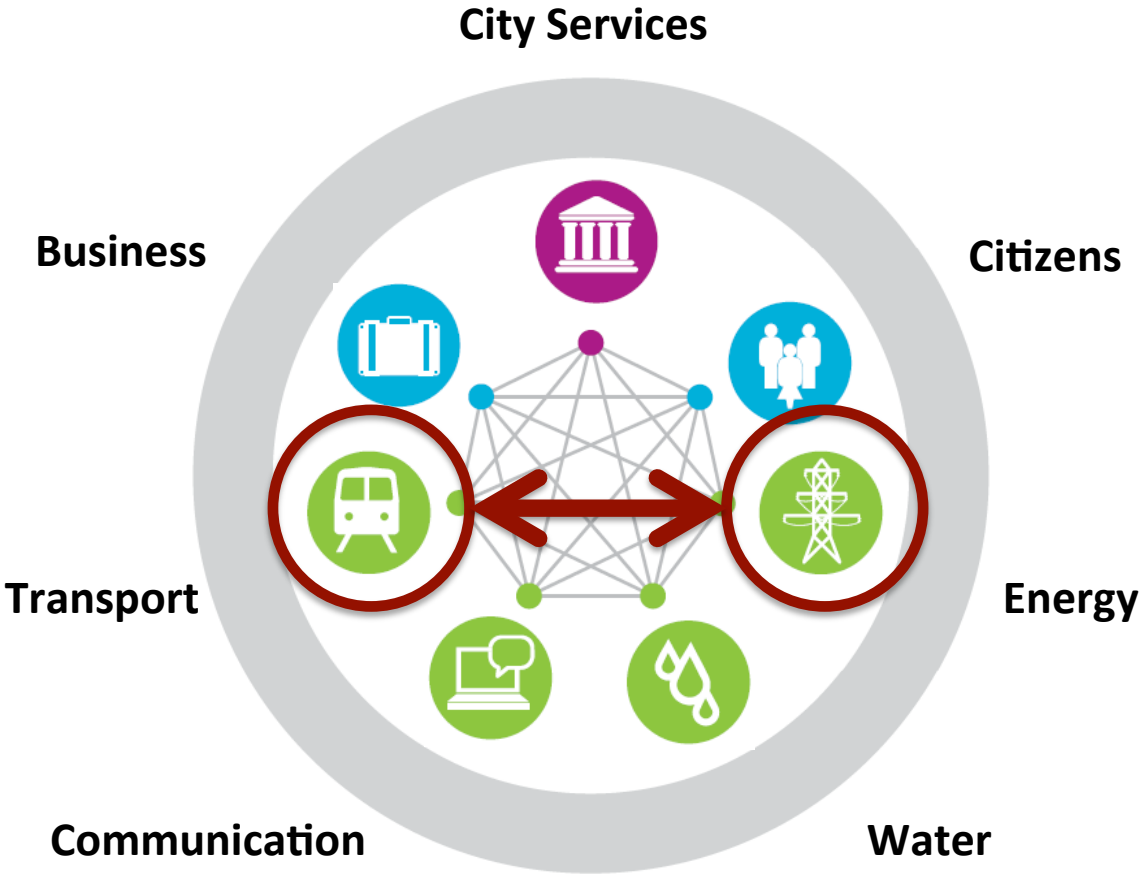


Impact of urbanization: energy use



Source: LLNL 2015. Data is based on DOE/EIA-0035(2015-03), March, 2014. If this information or a reproduction of it is used, credit must be given to the Lawrence Livermore National Laboratory and the Department of Energy, under whose auspices the work was performed. Distributed electricity represents only retail electricity sales and does not include self-generation. EIA reports consumption of renewable resources (i.e., hydro, wind, geothermal and solar) for electricity in BTU-equivalent values by assuming a typical fossil fuel plant "heat rate." The efficiency of electricity production is calculated as the total retail electricity delivered divided by the primary energy input into electricity generation. End use efficiency is estimated as 65% for the residential and commercial sectors 80% for the industrial sector, and 21% for the transportation sector. Totals may not equal sum of components due to independent rounding. LLNL-MI-410527

Electrification of transportation



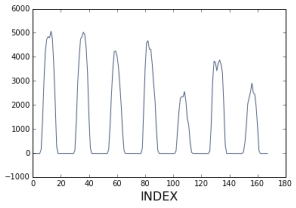
Tesla Model 3 announced
2 years ago, at $35’000, finally available

Data for supporting decisions and evaluate scenarios optimization problem (with constraints)

$$\min_{\beta} \left(\quad \right) =$$

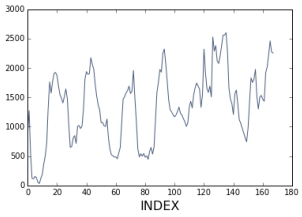
$$\left[\left(\beta_1 \right.\right.$$

Solar



$$+ \beta_2$$

Wind

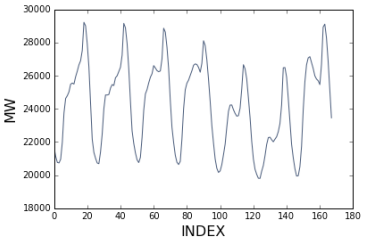


$$\left.\left. \right) \right] -$$

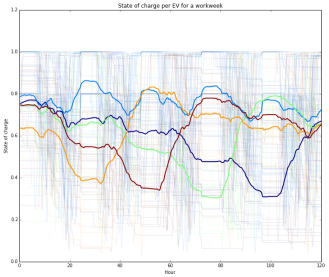
Possible supply from renewables

Demand forecast

$$\left(\alpha_0 \left(\right. \right.$$



$$\left.\left. \text{-15000MWh} \right) + \begin{bmatrix} \alpha_1^1 \\ \alpha_1^2 \end{bmatrix} \right)^2 \Big]$$



Balancing supply and demand ideas

1. With enough **storage**, one could balance over-generation at peaks and satisfy high mid-day demand.

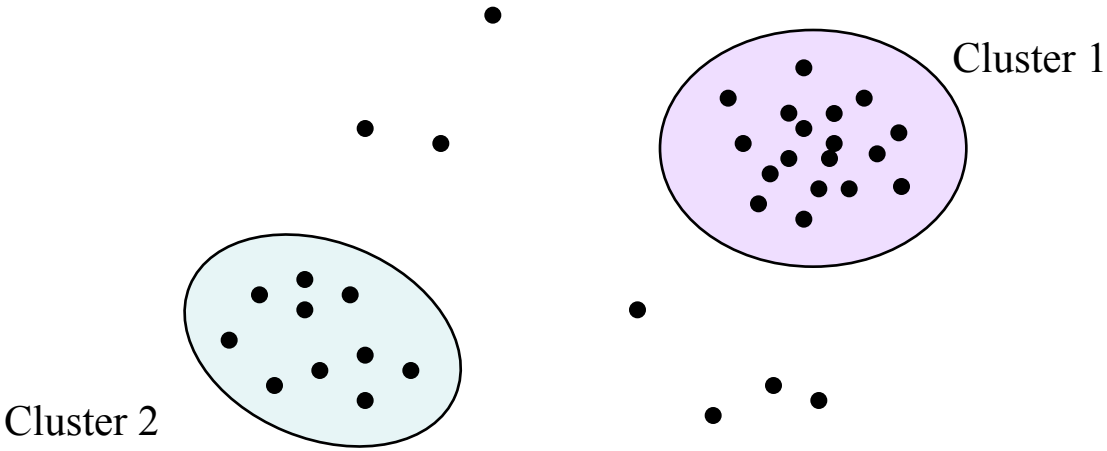
2. One could try and **manage demand** by shifting EV charging patterns in time through economic incentives.

To elaborate on idea 2, we are going to explore charging patterns using **cluster analysis** methods.

What is Cluster Analysis?

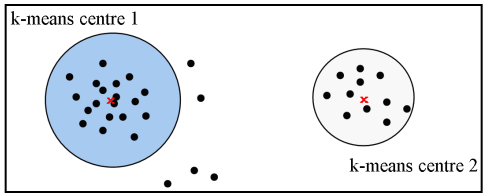
- Cluster: a collection of data objects
  - **Similar** to the objects in the same cluster (intraclass similarity)
  - **Dissimilar** to the objects in other clusters (interclass dissimilarity)
- Cluster analysis
  - Statistical/geometrical method for grouping a set of data objects into clusters
  - A good clustering method produces high quality clusters with high intraclass similarity and low interclass similarity
- Clustering is unsupervised classification
- Can be a stand-alone tool or as a preprocessing step for other algorithms

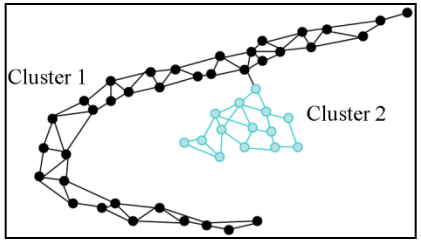
What is Cluster Analysis?



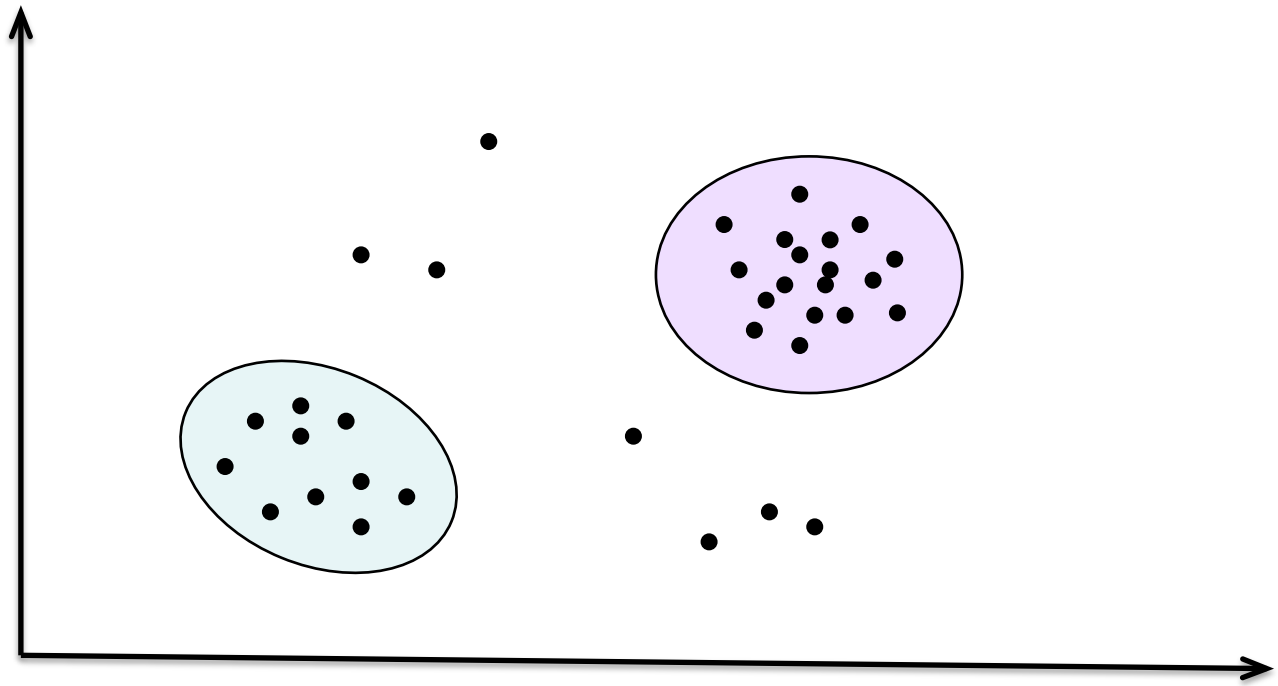
Requirements for Clustering

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal domain knowledge required to determine input parameters
- Ability to deal with noise and outliers
- Insensitivity to order of input records
- Robustness wrt high dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability



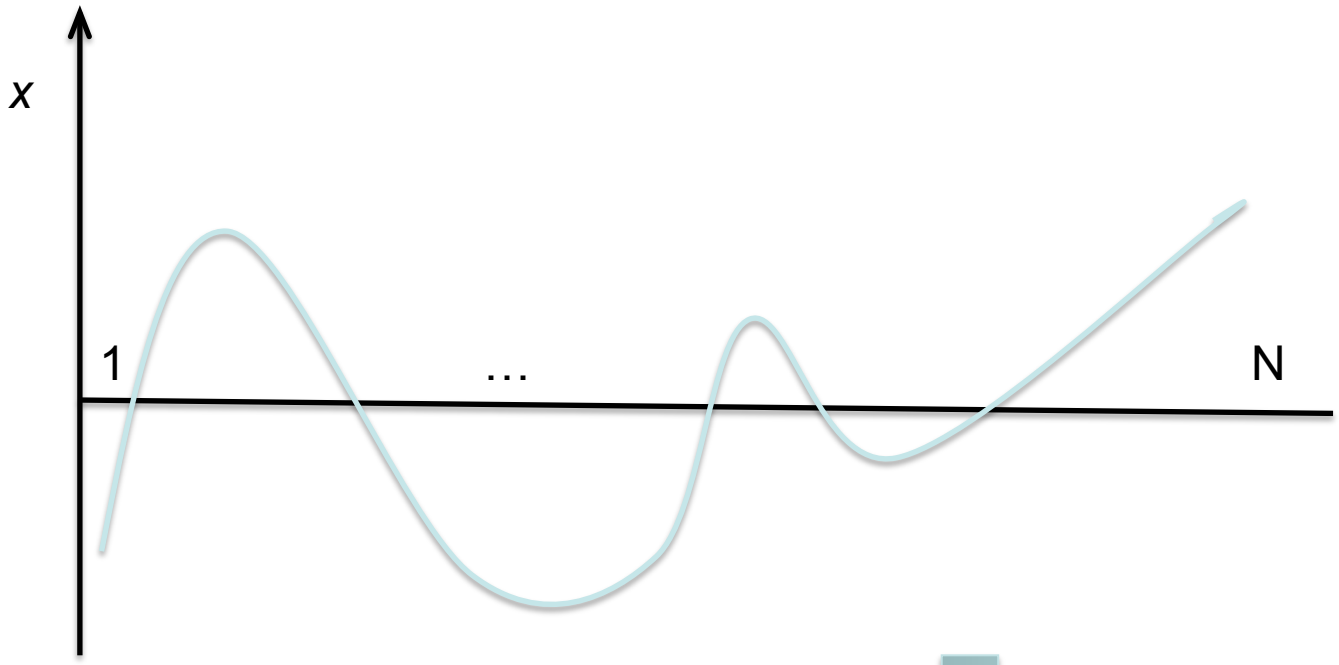


Data representation



$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

Data representation



$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

Data representation

$[\,[x_1\ x_2 \ldots \quad]$

$[x_{32}\ \ x_{33} \ldots \ ]\,]$

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

Similarity measures

- Euclidean Distance

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{n=1}^{N} (x_n - y_n)^2}$$

Similarity measures

- Cosine similarity

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$C_{\text{cosine}}(\vec{x}, \vec{y}) = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i \times y_i}{\|\vec{x}\| \times \|\vec{y}\|}$$

$\vec{x} = \vec{y}$ +1 ≥ Cosine Correlation ≥ – 1 $\vec{x} = -\vec{y}$

Similarity measures

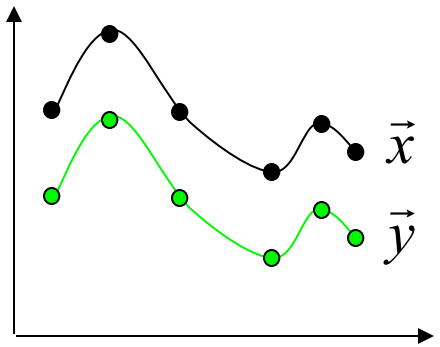
- Pearson Correlation

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

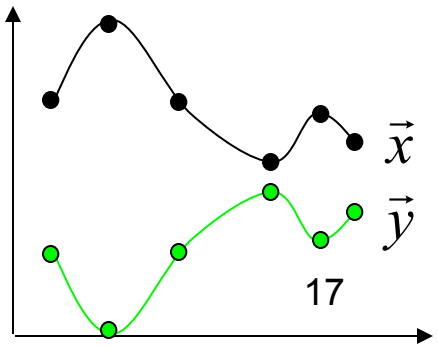
$$C_{pearson}(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{N}(x_i - m_x)(y_i - m_y)}{\sqrt{[\sum_{i=1}^{N}(x_i - m_x)^2][\sum_{i=1}^{N}(y_i - m_y)^2]}}$$

$$m_x = \frac{1}{N}\sum_{n=1}^{N} x_n$$

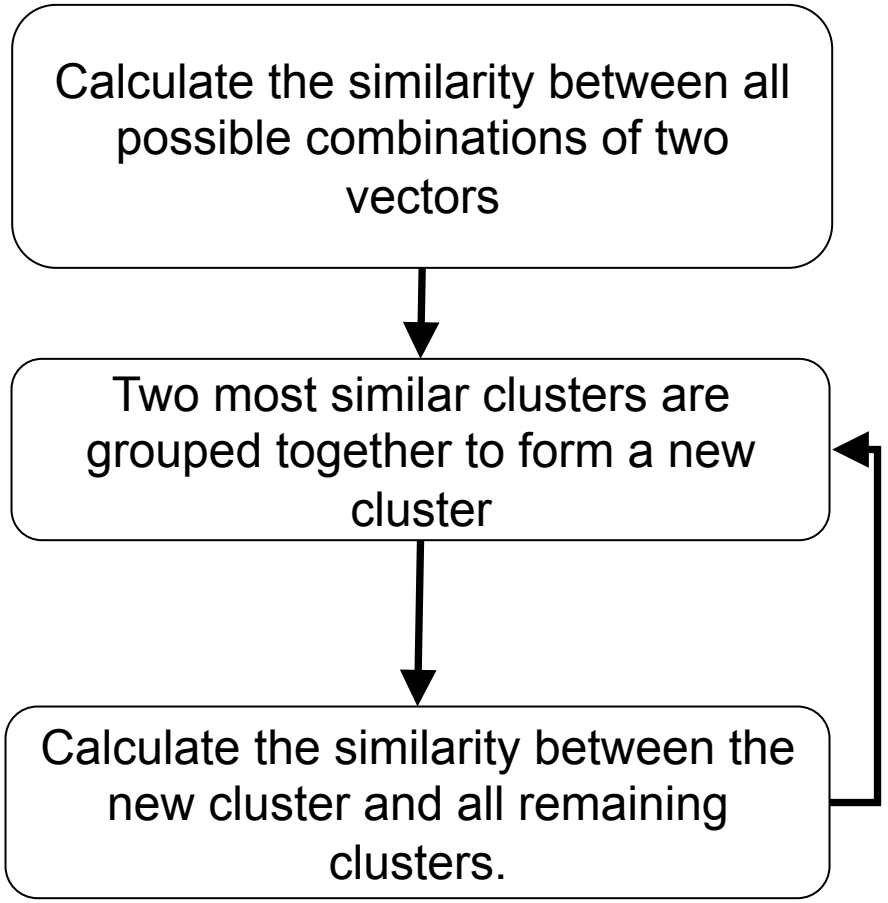
$$m_y = \frac{1}{N}\sum_{n=1}^{N} y_n$$



+1 ≥ Pearson Correlation ≥ – 1



Hierarchical Clustering



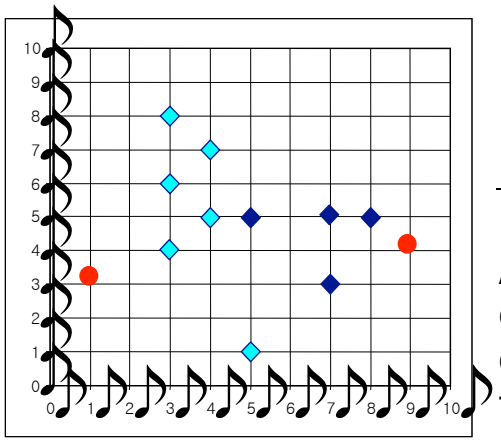
K-Means clustering

- The meaning of ‘K-means’
  - Why it is called ‘K-means’ clustering: K points are used to represent the clustering result; each point corresponds to the centre (geometric mean) of a cluster
- Each point is assigned to the cluster with the closest center point
- The number K must be specified
- Basic algorithm

K-Means clustering

- Given *k*, the *k-means* algorithm is implemented in 4 steps:
  - Partition objects into *k* non-empty subsets
  - Arbitrarily choose *k* points as initial centers (centroids)
  - Assign each object to the cluster with the nearest center
  - Calculate the mean of the cluster and update the center point
  - Go back to Step 3, stop when no more new assignment

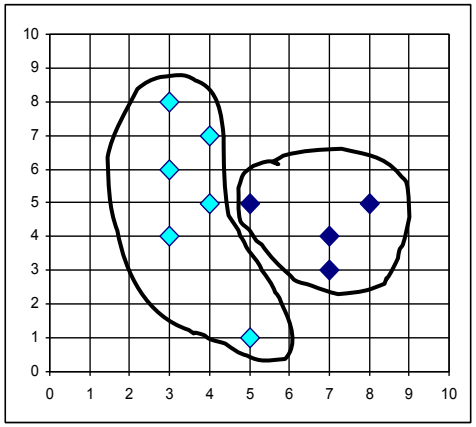
K-Means clustering



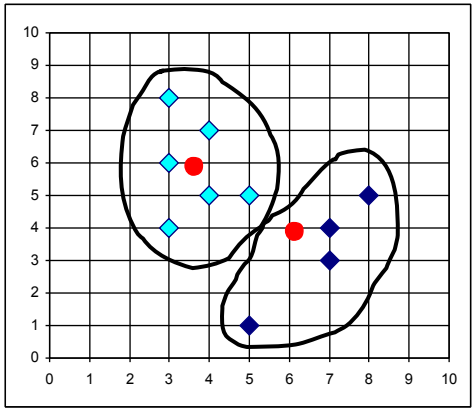
K=2

Arbitrarily choose K object as initial cluster center

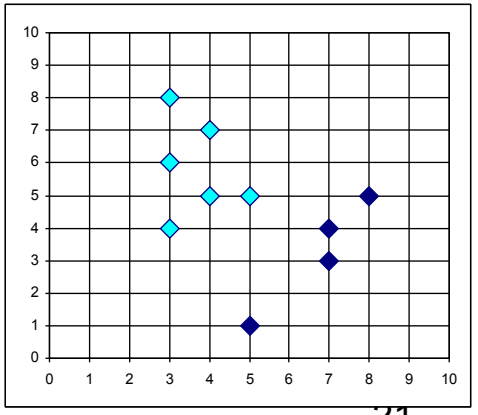
Assign each objects to most similar center



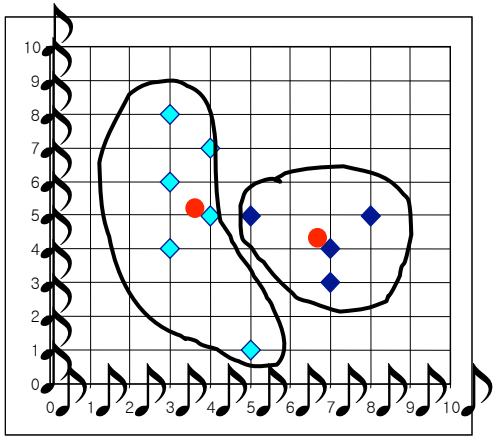
reassign



Update the cluster means



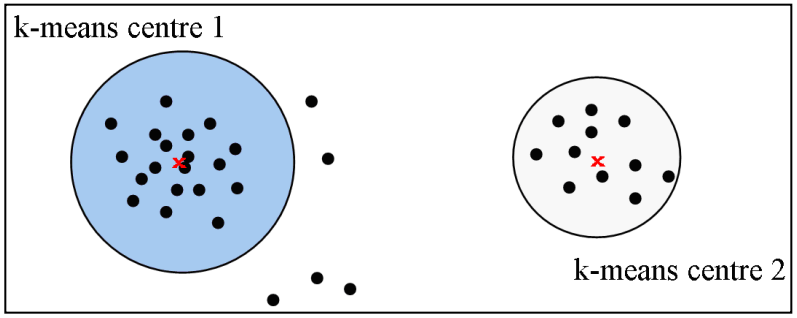
reassign

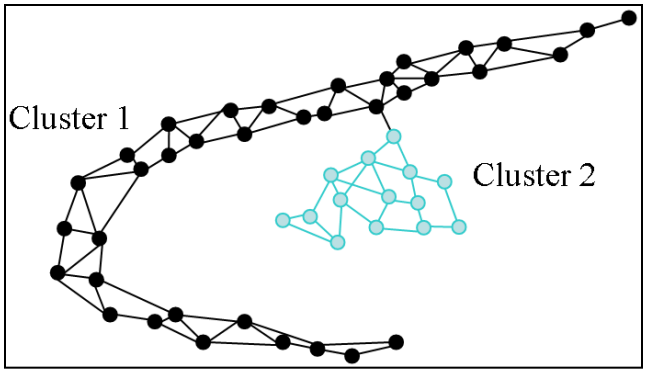


Update the cluster means

Clustering

- Data exploration method
- Can be interpreted as a purely geometrical approach of grouping similar data samples together
- Requires data representation and the definition of similarity
- K-means (and other algorithms)
- Involves parameters choice (number of clusters, etc)





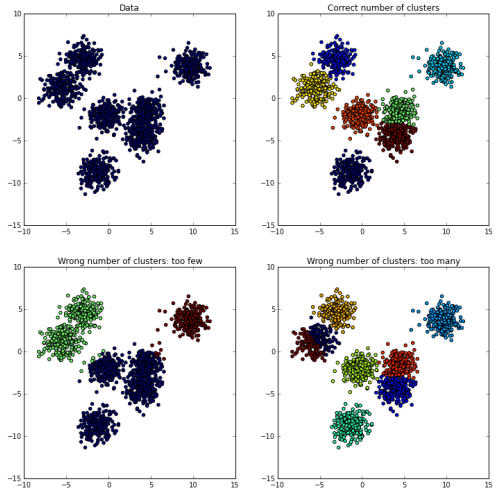
Implementation of clustering methods

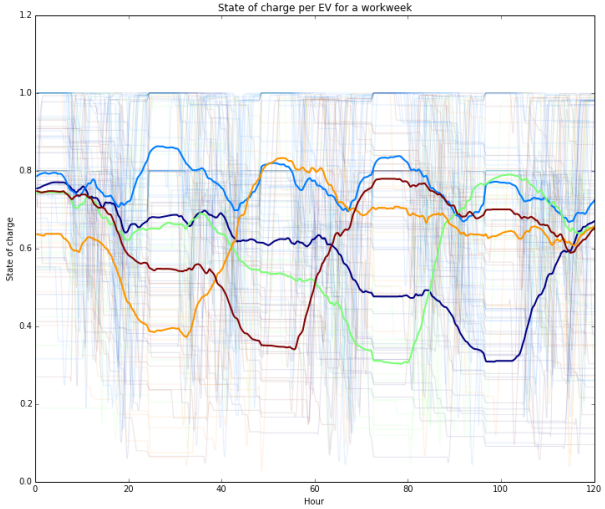
Scikit-learn

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large n_samples, medium n_clusters	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Spectral clustering	number of clusters	Medium n_samples, small n_clusters	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Hierarchical clustering	number of clusters	Large n_samples and n_clusters	Many clusters, possibly connectivity constraints	Distances between points
DBSCAN	neighborhood size	Very large n_samples, medium n_clusters	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers

Minilab

- How to choose parameters: “toy” problem
- Clustering EV owners charging patterns
- Interpretation of clustering results





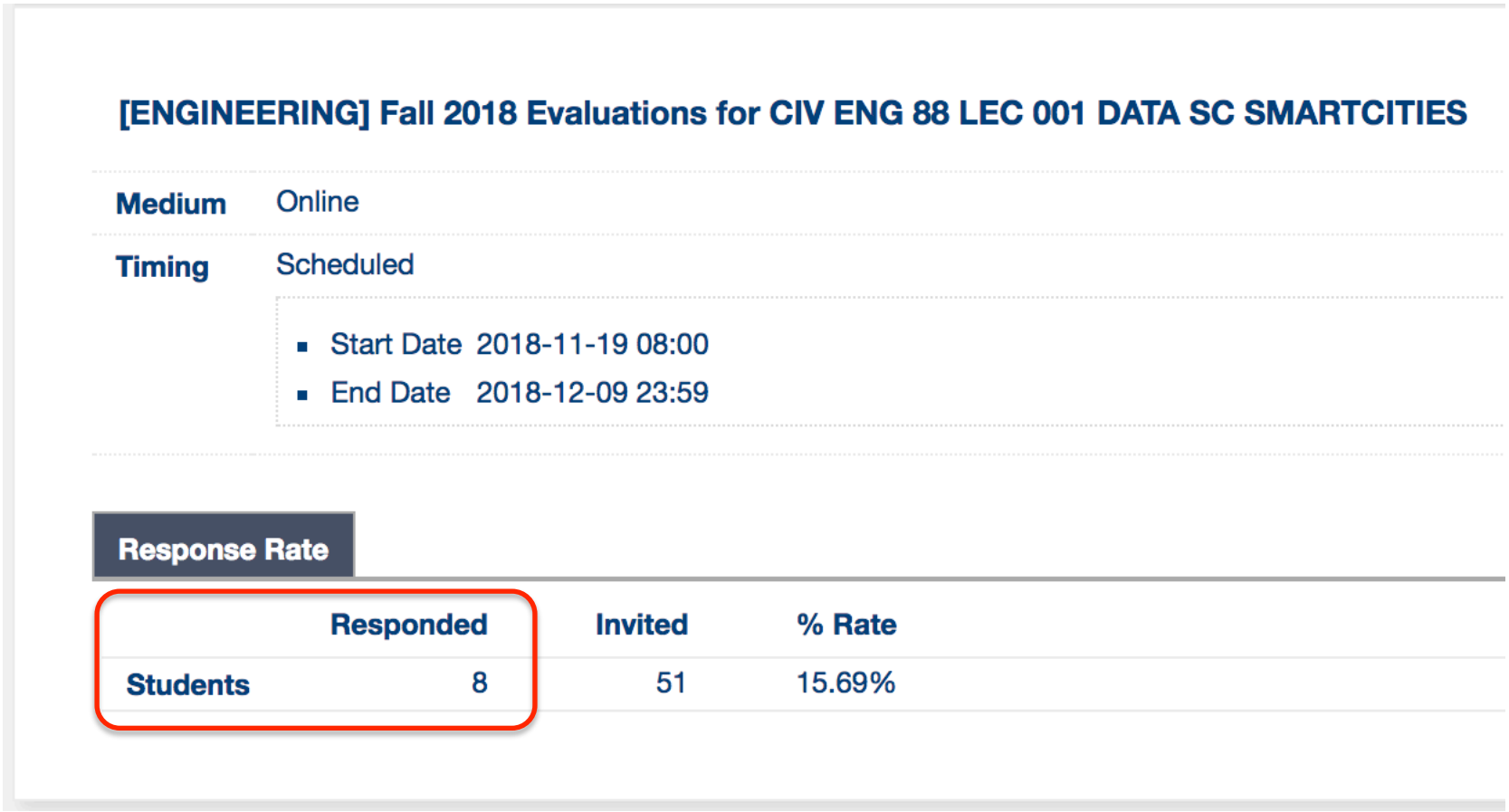
Final notes

Next Monday:

RRR week

Midterm graded + Final grades

Submit your evaluations. It matters!



[ENGINEERING] Fall 2018 Evaluations for CIV ENG 88 LEC 001 DATA SC SMARTCITIES

Medium Online

Timing Scheduled

- Start Date 2018-11-19 08:00
- End Date 2018-12-09 23:59

Response Rate

	Responded	Invited	% Rate
Students	8	51	15.69%

Thank you

Blair Zhang

Bingyi Fan

…and of course to all of you!!!