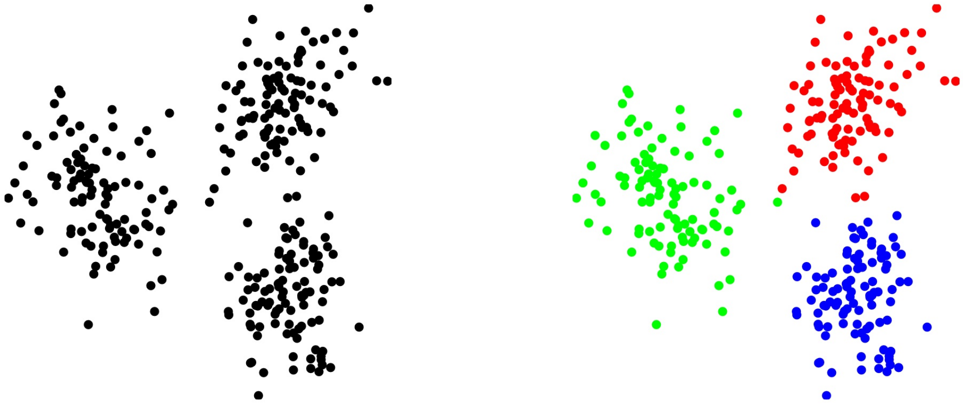
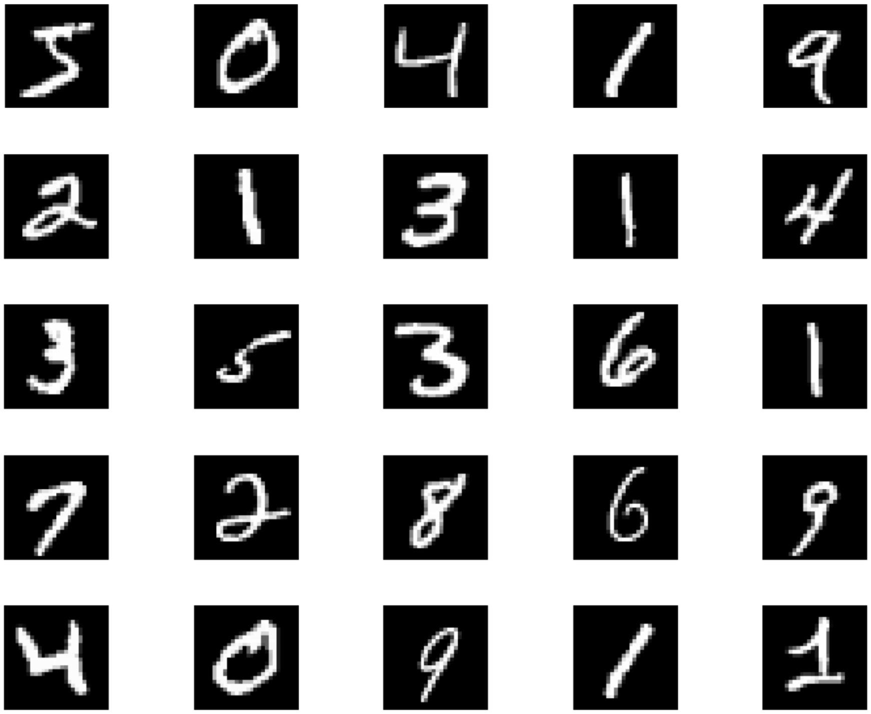
04 Clustering





Unit 1: Vectors, Book ILA Ch. 1-5

- 01 Vectors
- 02 Linear Functions
- 03 Norms and Distances
- ***04 Clustering***
- 05 Linear Independence

Unit 2: Matrices, Book ILA Ch. 6-11 + Book IMC Ch. 2

Unit 3: Least Squares, Book ILA Ch. 12-14 + Book IMC Ch. 8

Unit 4: Eigen-decomposition, Book IMC Ch. 10, 12, 19

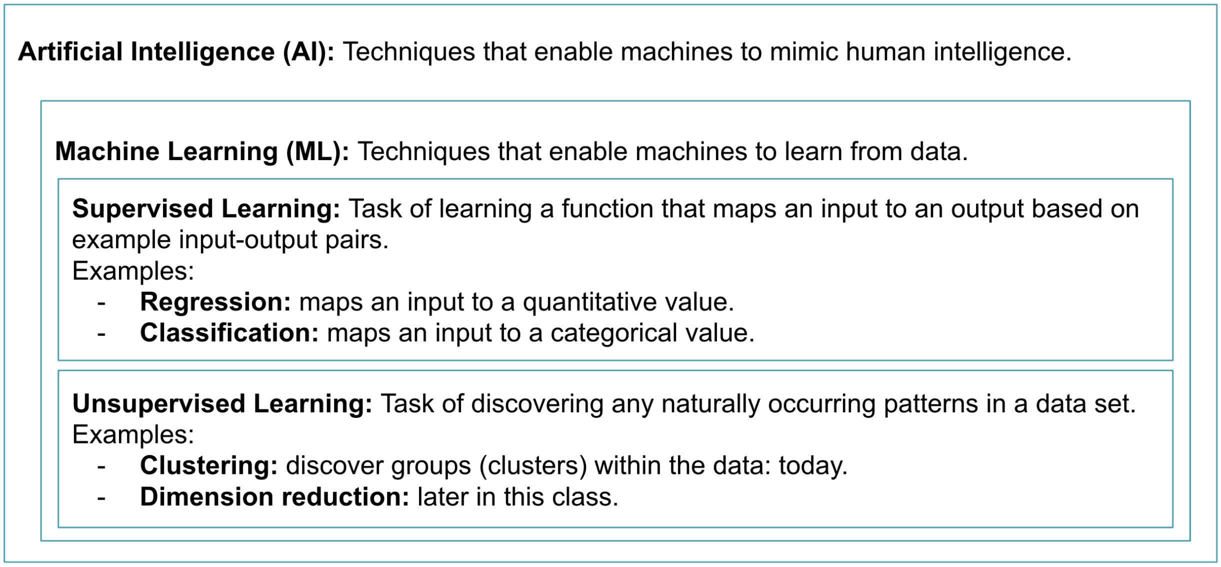
Outline: 04 Clustering

- Clustering
- Algorithm
- Examples
- Applications

Outline: 04 Clustering

- **Clustering**
- Algorithm
- Examples
- Applications

Clustering in Machine Learning



Artificial Intelligence (AI): Techniques that enable machines to mimic human intelligence.

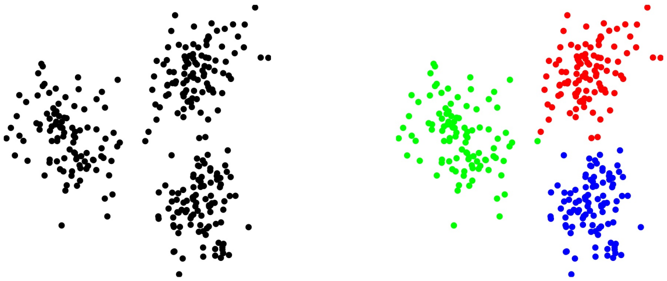
Machine Learning (ML): Techniques that enable machines to learn from data.

Supervised Learning: Task of learning a function that maps an input to an output based on example input-output pairs.
Examples:
- **Regression:** maps an input to a quantitative value.
- **Classification:** maps an input to a categorical value.

Unsupervised Learning: Task of discovering any naturally occurring patterns in a data set.
Examples:
- **Clustering:** discover groups (clusters) within the data: today.
- **Dimension reduction:** later in this class.

Clustering: Goal (Intuition)

- Given: (i) dataset of N n-vectors $x_1, \ldots, x_N$, (ii) integer k,
- Goal (Intuition):
  - Partition/Group/Cluster N vectors into k groups/clusters...
  - ... such that: vectors in the same group are "close".



Exercise: What is k in the figure above? What is n? What is N?

Clustering in ECE

- topic discovery and document clustering
  - x_i is word count vector for document i
- patient clustering
  - x_i are patient attributes, test results, symptoms for patient i
- customer market "segmentation"
  - x_i is purchase history and other attributes of customer i
- financial sectors
  - x_i are n-vectors of financial attributes of company i

Clustering: Goal (Math)

- Notations:
  - Group G_j for $j = 1, \ldots, k$: Set of indices in $1, \ldots, N$ representing which vectors belong to the group.
  - Assignment c_i for $i = 1, \ldots, k$: Group that x_i is in: $i \in G_{c_i}$
  - Group representative z_j for $j = 1, \ldots, k$: n-vector that represents a typical element of the group G_j.
- Goal (Math): Find c_i and z_j to minimize $J^{clust} = \frac{1}{N} \sum_{i=1}^{N} ||x_i - z_{c_i}||^2$, i.e. the mean square distance from vectors to their representatives.

Outline: 04 Clustering

- Clustering
- **Algorithm**
- Examples
- Applications

K-means algorithm

- Alternate between:
    - (i) update the groups, i.e the group assignments $c_1, \dots, c_N$,
    - (ii) update the representatives $z_1, \dots, z_k$.
- Such that the objective J^{clust} decreases at each step.

(i) Update the groups

- Given: representatives $z_1, \dots, z_k$
- Goal for (i): Assign vectors to groups, i.e. choose $c_1, \dots, c_N$
    - We assign each vector to its nearest representative. Justification:
        - Observe: c_i only appears in term $||x_i - z_{c_i}||^2$ in J^{clust}
        - Conclude: to minimize over c_i, choose c_i so $\|x_i - z_{c_i}\|^2 = min_{j \in \{1,\dots,k\}} \|x_i - z_j\|^2$.

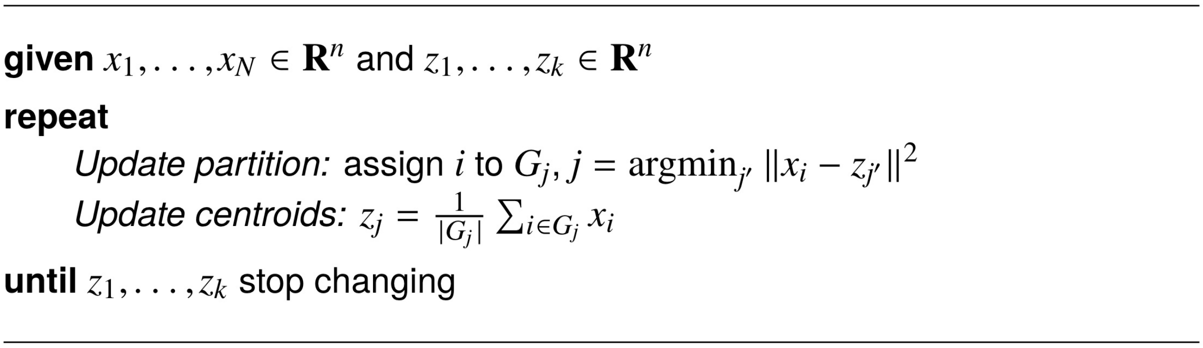
(ii) Update the representatives

- Given the partition $G_1, \dots, G_k$
- Goal for (ii): Choose representatives $z_1, \dots, z_k$
    - Choose z_j = mean of the points in group j. Justification:
        - Observe: J^{clust} splits into a sum of k sums:

$$J^{clust} = J_1 + \cdots + J_k, \quad J_j = \frac{1}{N} \sum_{i \in G_j} \|x_i - z_j\|^2.$$

        - Conclude: Choose z_j to minimize its J_j: $z_j = \frac{1}{|G_j|} \sum_{i \in G_j} x_i$ = mean/center/centroid.

Pseudo-code



```
given $x_1, \dots, x_N \in \mathbf{R}^n$ and $z_1, \dots, z_k \in \mathbf{R}^n$
repeat
    Update partition: assign $i$ to $G_j$, $j = \mathrm{argmin}_{j'} \|x_i - z_{j'}\|^2$
    Update centroids: $z_j = \frac{1}{|G_j|} \sum_{i \in G_j} x_i$
until $z_1, \dots, z_k$ stop changing
```

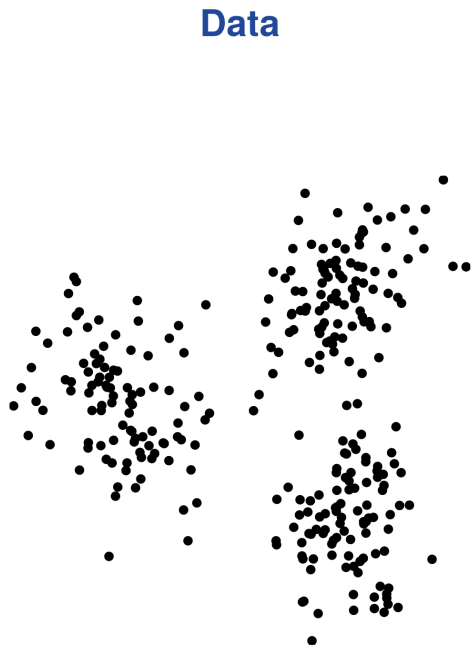
Convergence of K-means

- How many times do we iterate these steps?

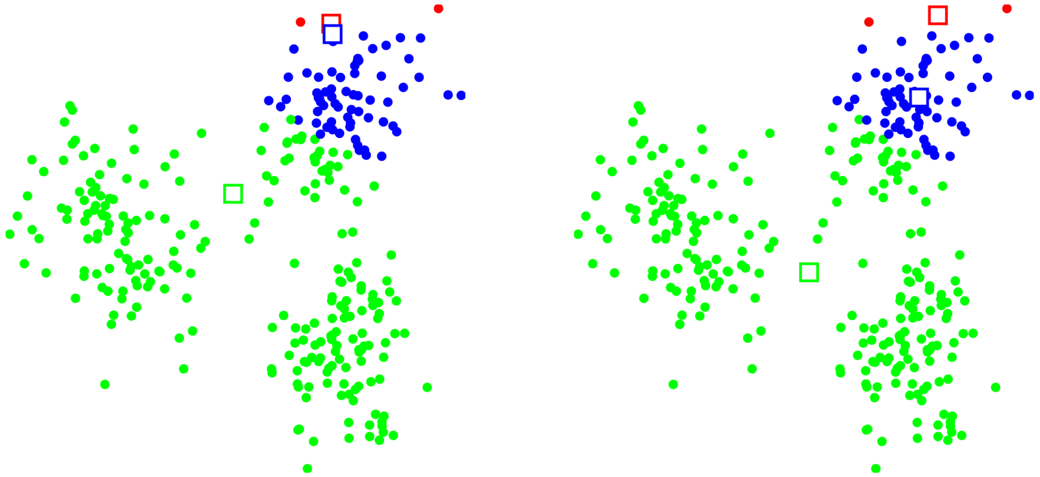
    - Until the z_j's stop changing: "convergence" of the algorithm.
- Remarks:
    - J^{clust} decreases at each step,
    - but in general we don't find partition that minimizes J^{clust},
    - the final partition depends on initial representatives.
- Recommendation:
    - Run k-means 10 times, with different initial representatives
    - Take as final partition the one with smallest J^{clust}

Outline: 04 Clustering

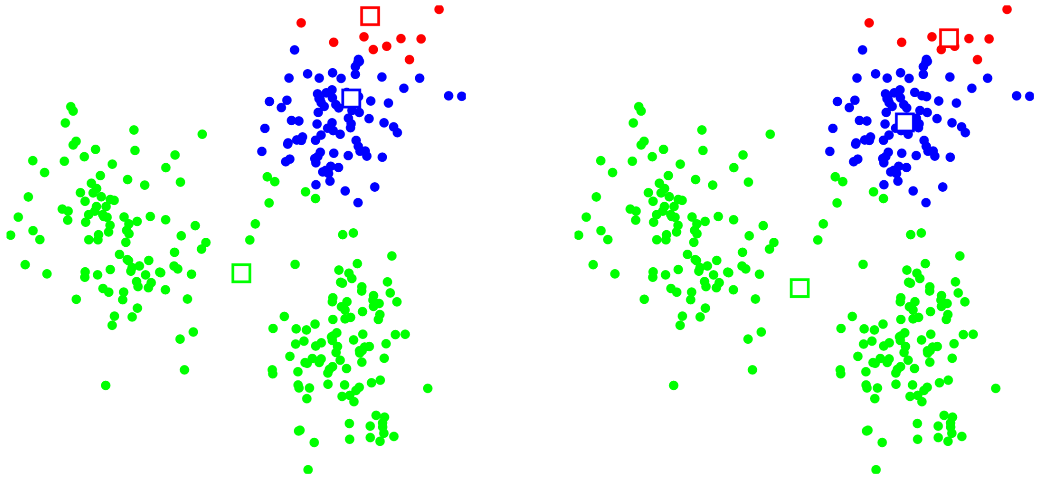
- Clustering
- Algorithm
- **Example**
- Applications



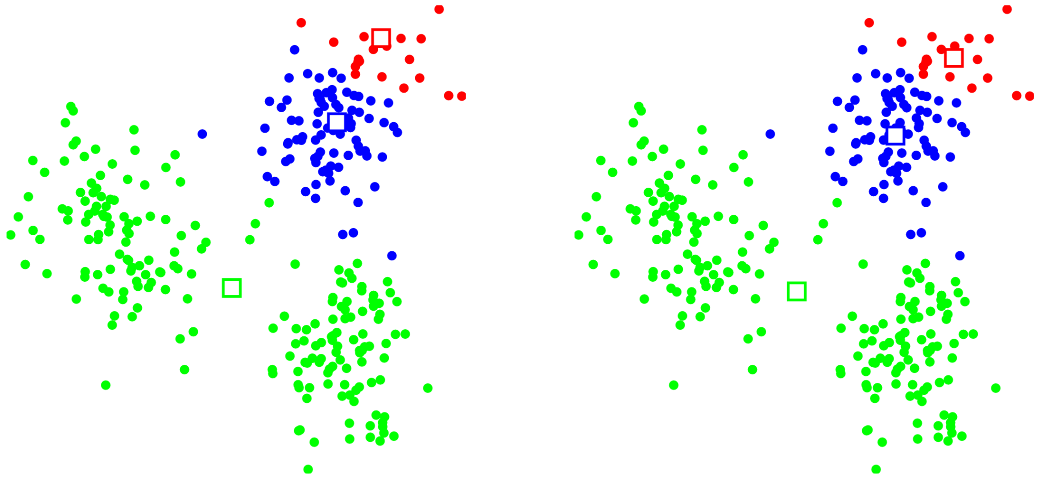
Iteration 1



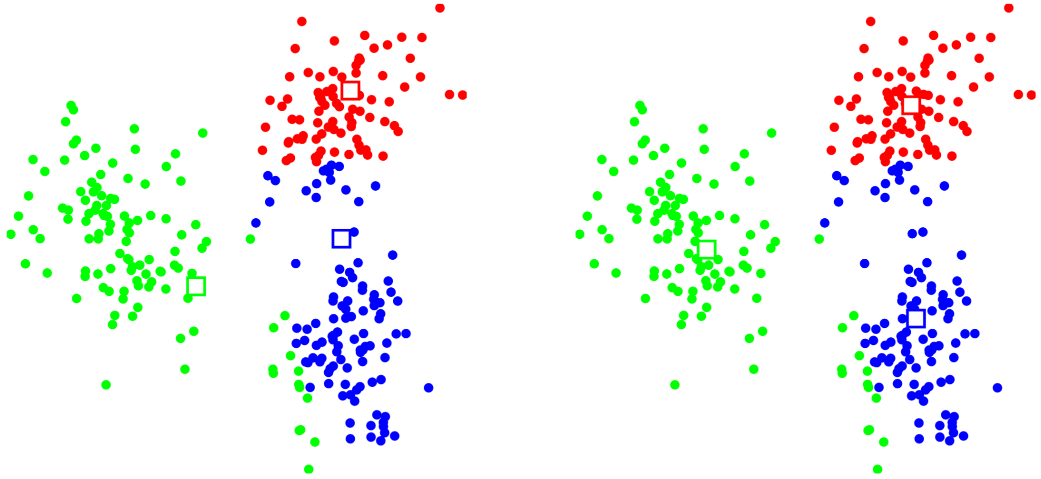
Iteration 2



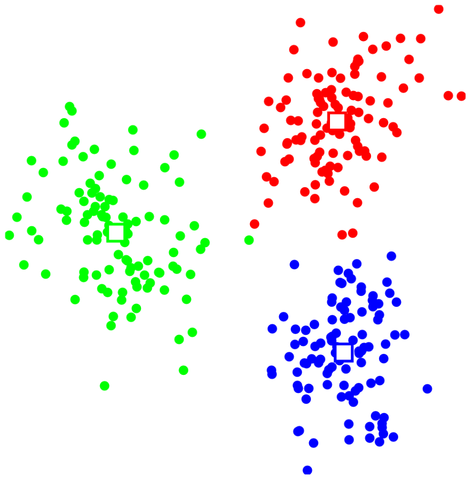
Iteration 3



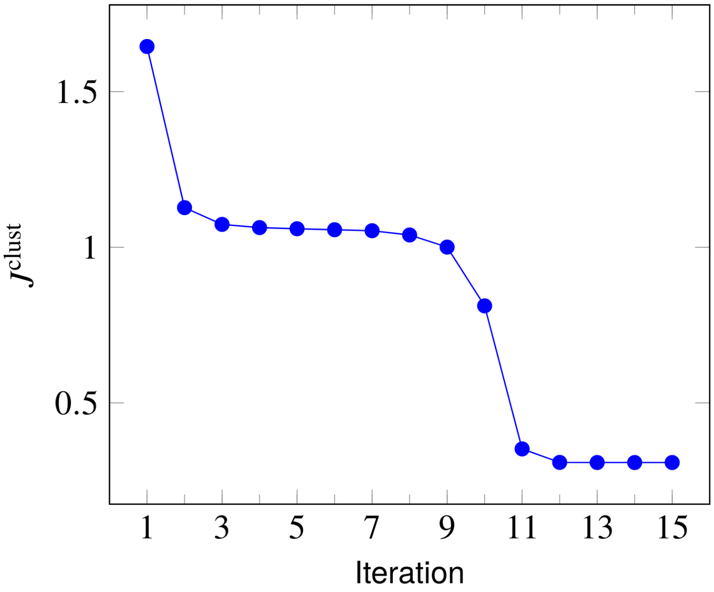
Iteration 10



Final clustering



Convergence



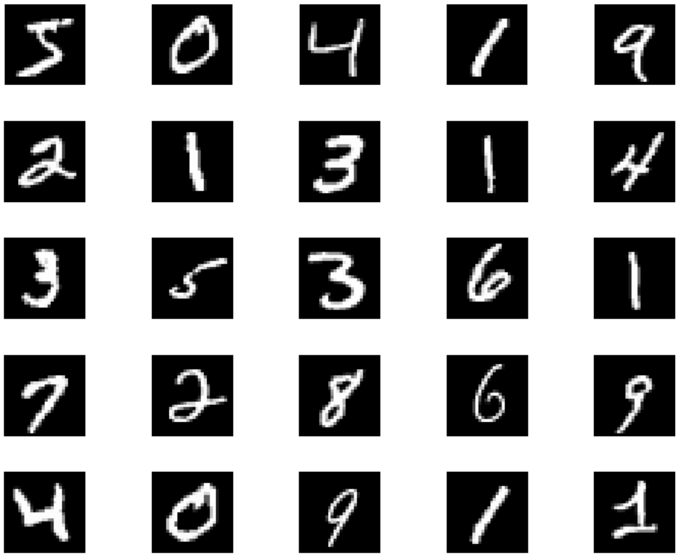
Outline: 04 Clustering

- Clustering

- Algorithm
- Examples
- **Applications**

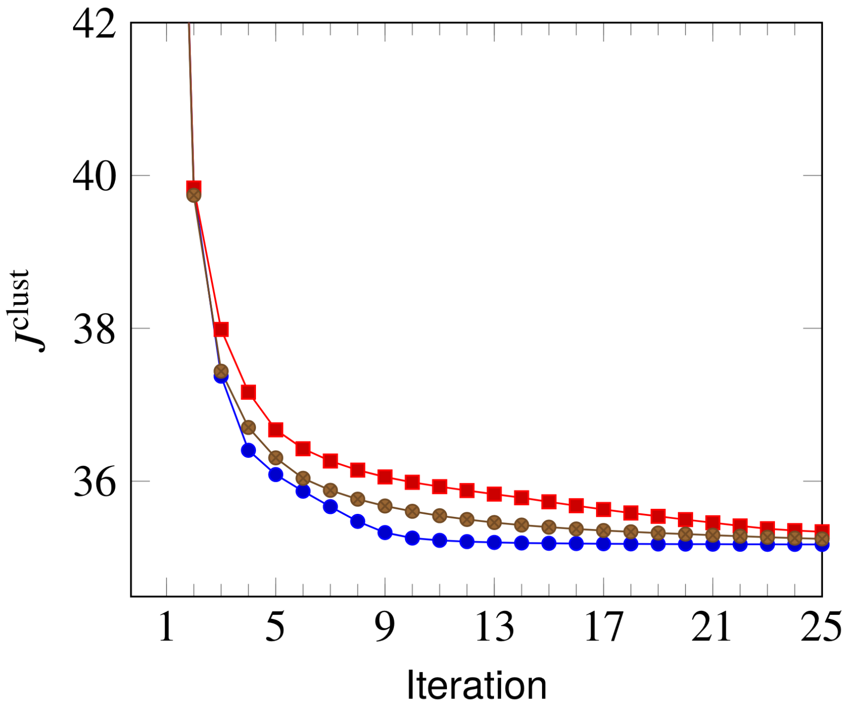
MNIST Dataset: Find Digits

- MNIST images of handwritten digits (via Yann Lecun)
- $60,000$ images of size 28 × 28, represented as 784-vectors x_i

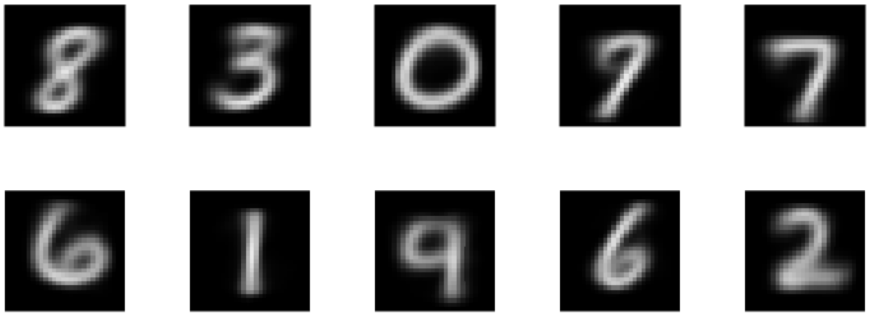


- Goal: Group these images into groups of same digit.
- Exercice: What are k, N, n?

MNIST Results



Convergences: best run (blue), worst run (red), average (brown).



Representatives.

Wikipedia Dataset: Find Topics

- Wikipedia articles
- 500 articles, where the word count vectors are computed (dictionary of 4423 words)

Game of Thrones

From Wikipedia, the free encyclopedia

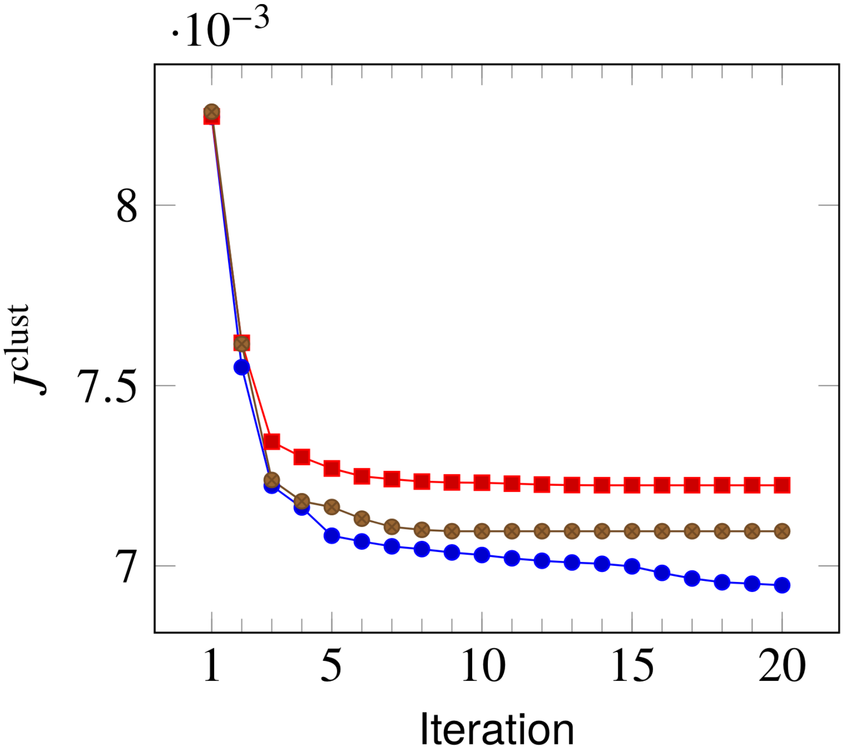
This article is about the television series. For other uses, see Game of Thrones (disambiguatic

Game of Thrones is an American fantasy drama television series created by David Benioff and D Martin, the first of which is *A Game of Thrones*. The show was shot in the United Kingdom, Canad and concluded on May 19, 2019, with 73 episodes broadcast over eight seasons.

Set on the fictional continents of Westeros and Essos, *Game of Thrones* has a large ensemble ca: Throne of the Seven Kingdoms of Westeros through a web of political conflicts among the noble fa

- Goal: Group these articles into groups of same topic.
- Exercice: What are k, N, n?

Wikipedia Results



Convergences: best run (blue), worst run (red), average (brown).

Explore first 3 clusters.

- words with largest representative coefficients

Cluster 1		Cluster 2		Cluster 3	
Word	Coef.	Word	Coef.	Word	Coef.
fight	0.038	holiday	0.012	united	0.004
win	0.022	celebrate	0.009	family	0.003
event	0.019	festival	0.007	party	0.003
champion	0.015	celebration	0.007	president	0.003
fighter	0.015	calendar	0.006	government	0.003

- titles of articles closest to cluster representative

1. “Floyd Mayweather, Jr”, “Kimbo Slice”, “Ronda Rousey”, “José Aldo”, “Joe Frazier”, “Wladimir Klitschko”, “Saul Álvarez”, “Gennady Golovkin”, “Nate Diaz”, . . .
2. “Halloween”, “Guy Fawkes Night” “Diwali”, “Hanukkah”, “Groundhog Day”, “Rosh Hashanah”, “Yom Kippur”, “Seventh-day Adventist Church”, “Remembrance Day”, . . .
3. “Mahatma Gandhi”, “Sigmund Freud”, “Carly Fiorina”, “Frederick Douglass”, “Marco Rubio”, “Christopher Columbus”, “Fidel Castro”, “Jim Webb”, . . .

Questions?

- Clustering
- Algorithm
- Examples

- Applications

Resources: Book ILA, Ch. 4