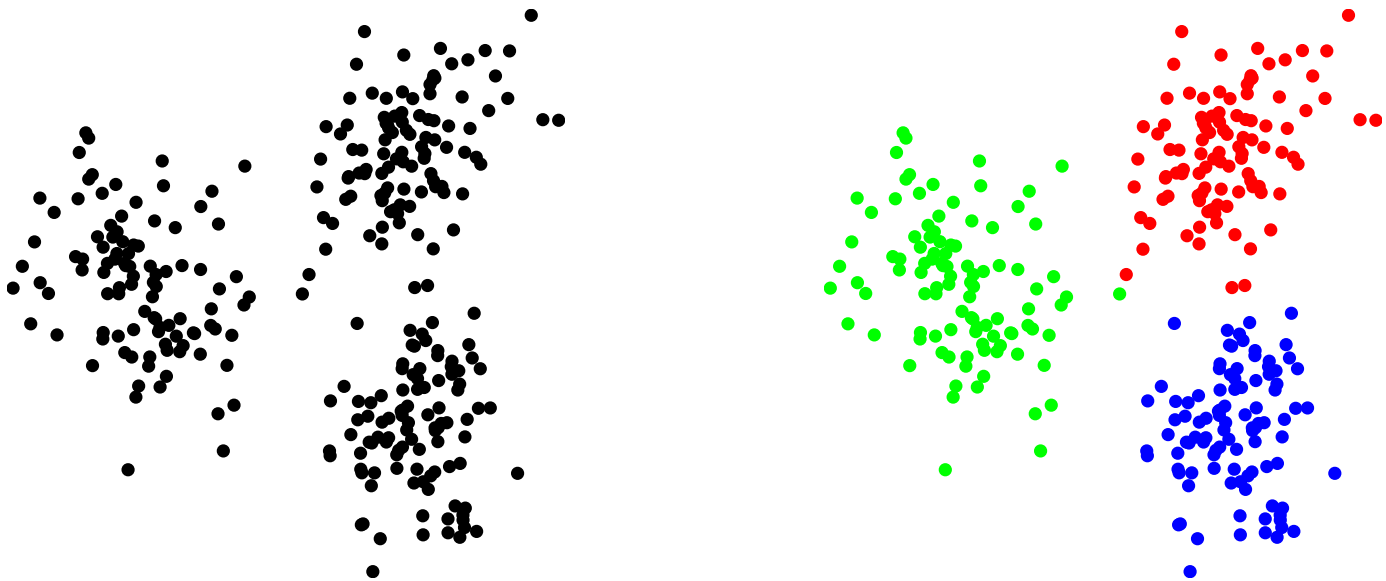
4. Clustering

Outline

Clustering

- given N n-vectors $x_1, \ldots, x_N$
- goal: partition (divide, cluster) into k groups
- want vectors in the same group to be close to one another



Example settings

- topic discovery and document classification
  - x_i is word count histogram for document i
- patient clustering
  - x_i are patient attributes, test results, symptoms
- customer market segmentation
  - x_i is purchase history and other attributes of customer i
- color compression of images
  - x_i are RGB pixel values
- financial sectors
  - x_i are n-vectors of financial attributes of company i

Clustering objective

- $G_j \subset \{1,\ldots,N\}$ is group j, for $j = 1,\ldots,k$
- c_i is group that x_i is in: $i \in G_{c_i}$
- group *representatives*: n-vectors $z_1,\ldots,z_k$
- clustering objective is

$$J^{\text{clust}} = \frac{1}{N}\sum_{i=1}^{N} \|x_i - z_{c_i}\|^2$$

  mean square distance from vectors to associated representative

- J^{clust} small means good clustering
- goal: choose clustering c_i and representatives z_j to minimize J^{clust}

Outline

Partitioning the vectors given the representatives

- suppose representatives $z_1, \ldots, z_k$ are given
- how do we assign the vectors to groups, *i.e.*, choose $c_1, \ldots, c_N$?

- c_i only appears in term $\|x_i - z_{c_i}\|^2$ in J^{clust}
- to minimize over c_i, choose c_i so $\|x_i - z_{c_i}\|^2 = \min_j \|x_i - z_j\|^2$
- *i.e.*, *assign each vector to its nearest representative*

Choosing representatives given the partition

- given the partition $G_1, \ldots, G_k$, how do we choose representatives $z_1, \ldots, z_k$ to minimize J^{clust}?
- J^{clust} splits into a sum of k sums, one for each z_j:

$$J^{\text{clust}} = J_1 + \cdots + J_k, \qquad J_j = (1/N) \sum_{i \in G_j} \|x_i - z_j\|^2$$

- so we choose z_j to minimize mean square distance to the points in its partition
- this is the mean (or average or centroid) of the points in the partition:

$$z_j = (1/|G_j|) \sum_{i \in G_j} x_i$$

k-means algorithm

- alternate between updating the partition, then the representatives
- a famous algorithm called *k-means*
- objective J^{clust} decreases in each step

---

given $x_1, \ldots, x_N \in \mathbf{R}^n$ and $z_1, \ldots, z_k \in \mathbf{R}^n$

repeat

Update partition: assign i to G_j, $j = \operatorname{argmin}_{j'} \|x_i - z_{j'}\|^2$

Update centroids: $z_j = \frac{1}{|G_j|} \sum_{i \in G_j} x_i$

until $z_1, \ldots, z_k$ stop changing

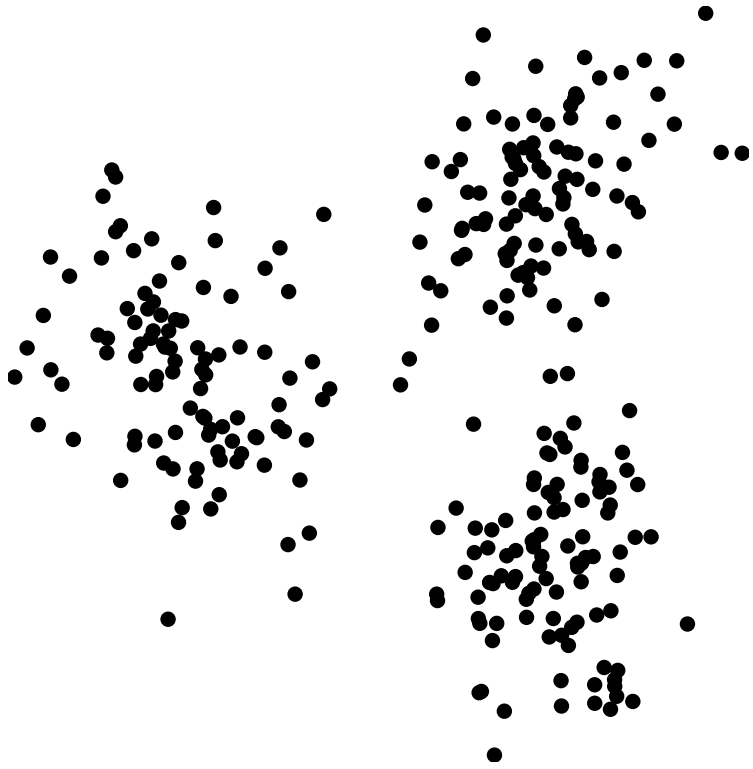
Convergence of k-means algorithm

- J^{clust} goes down in each step, until the z_j's stop changing
- but (in general) the k-means algorithm *does not find the partition that minimizes* J^{clust}

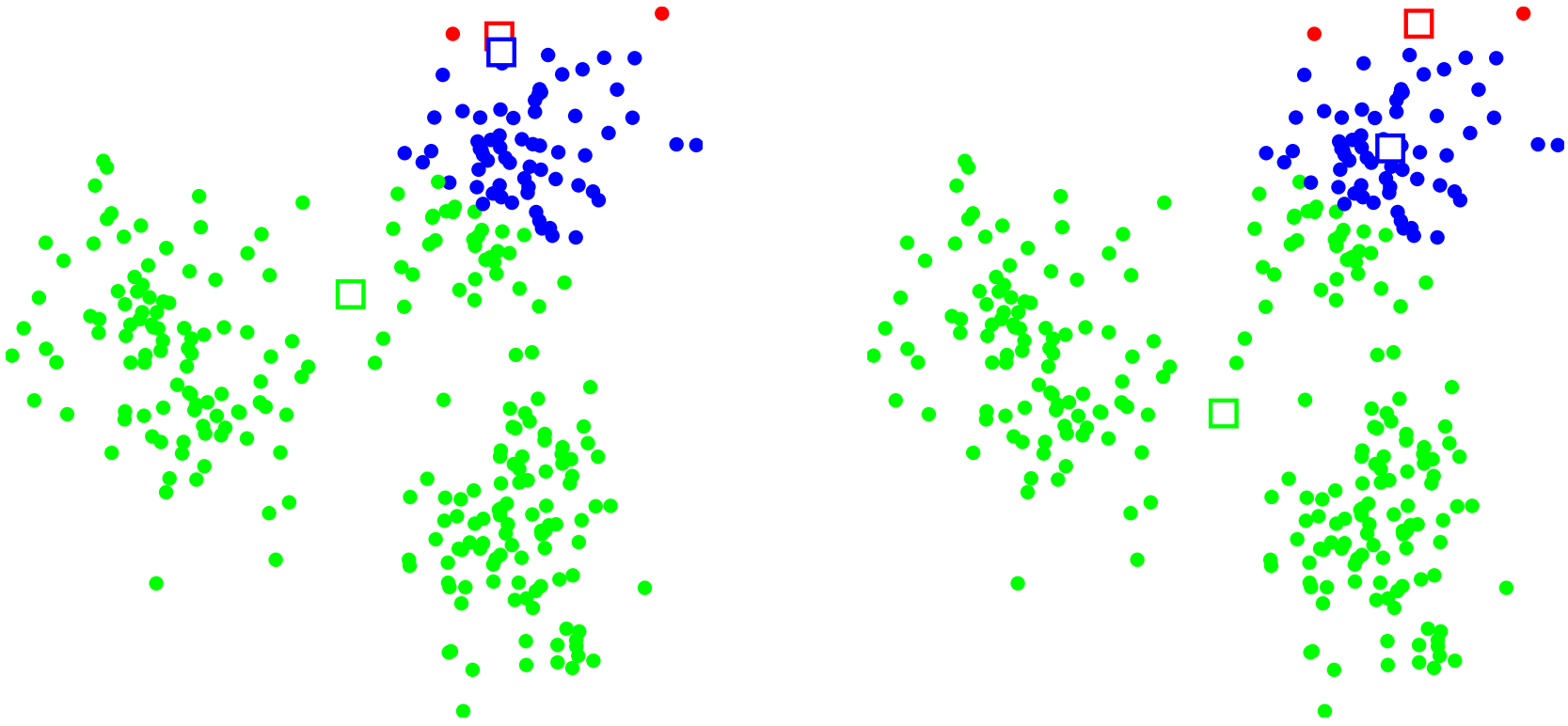
- k-means is a *heuristic*: it is not guaranteed to find the smallest possible value of J^{clust}
- the final partition (and its value of J^{clust}) can depend on the initial representatives
- common approach:
  - run k-means 10 times, with different (often random) initial representatives
  - take as final partition the one with the smallest value of J^{clust}

Outline

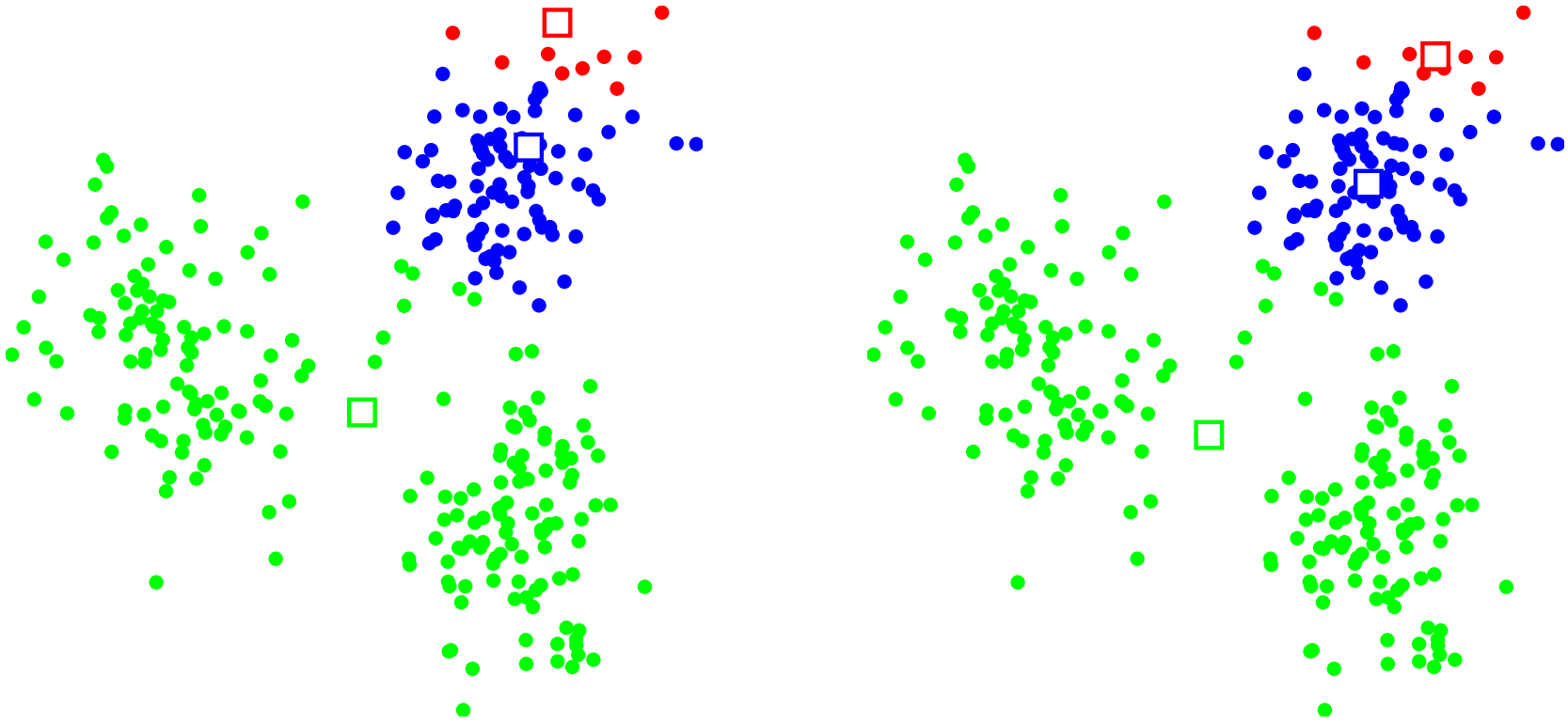
Data



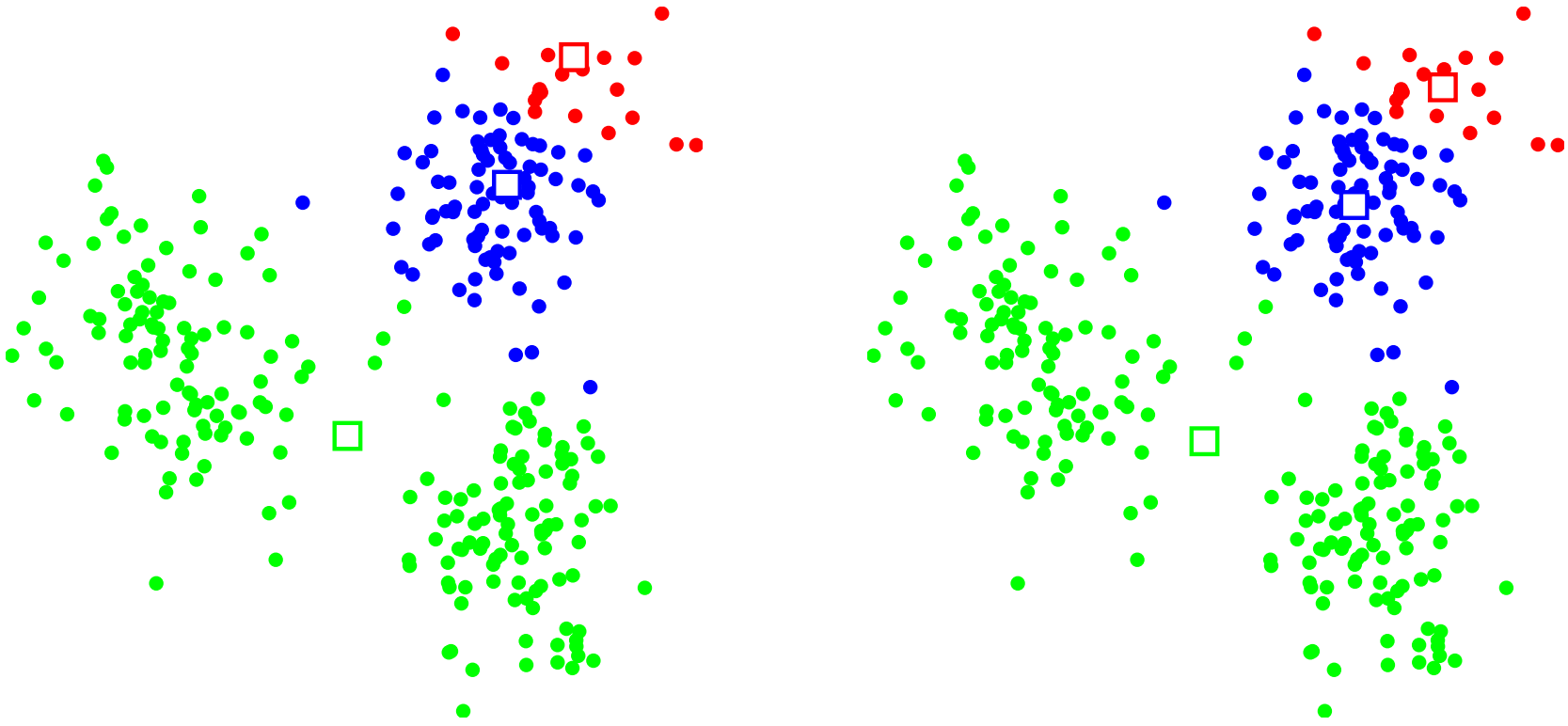
Iteration 1



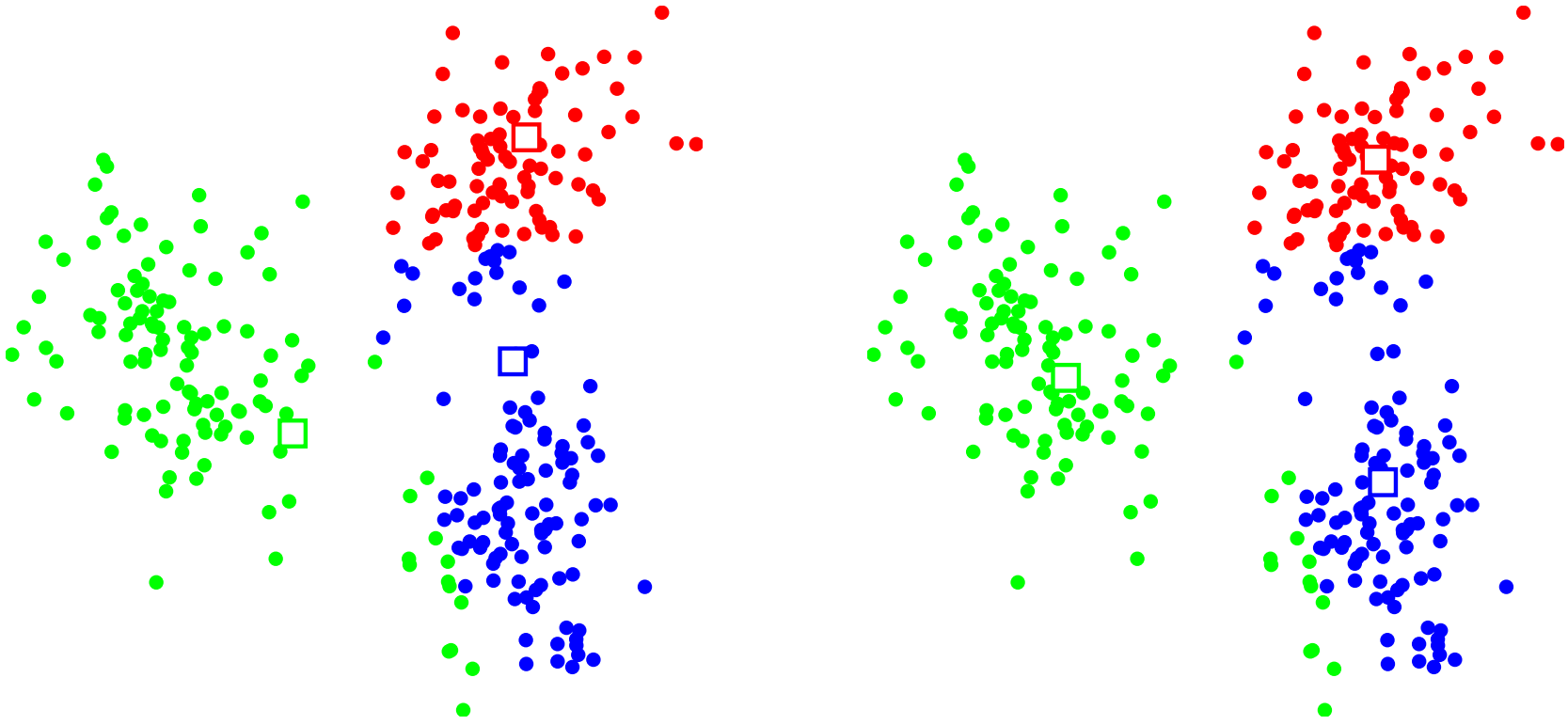
Iteration 2



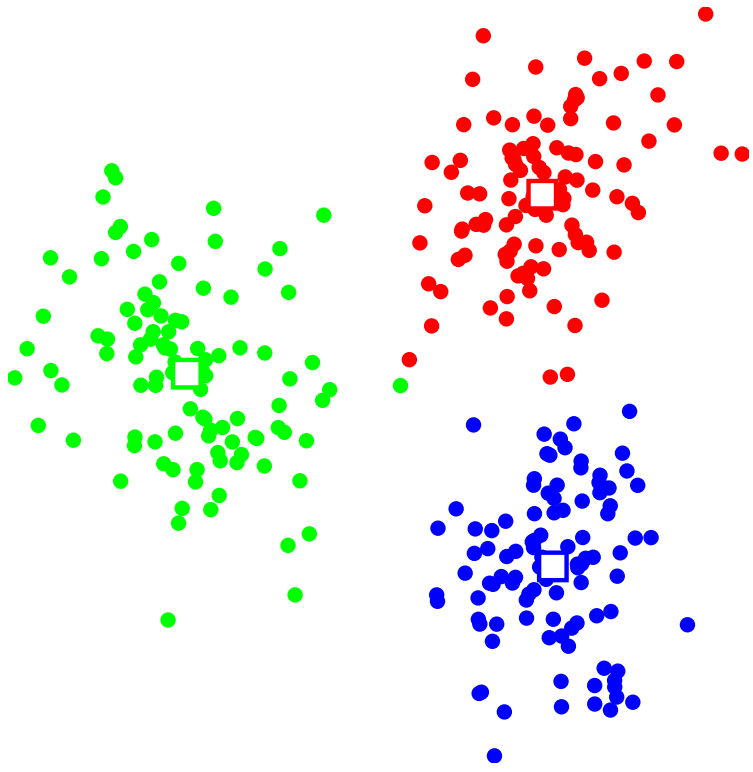
Iteration 3



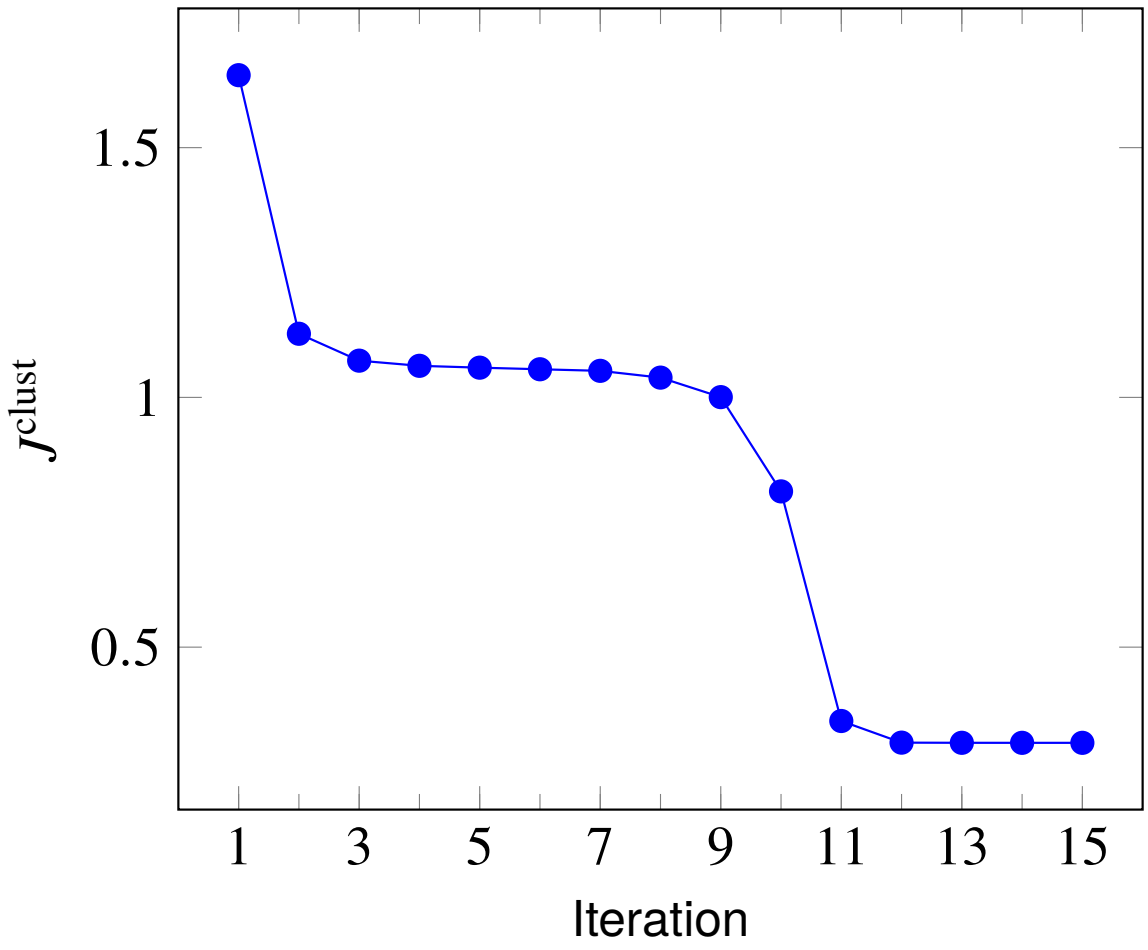
Iteration 10



Final clustering



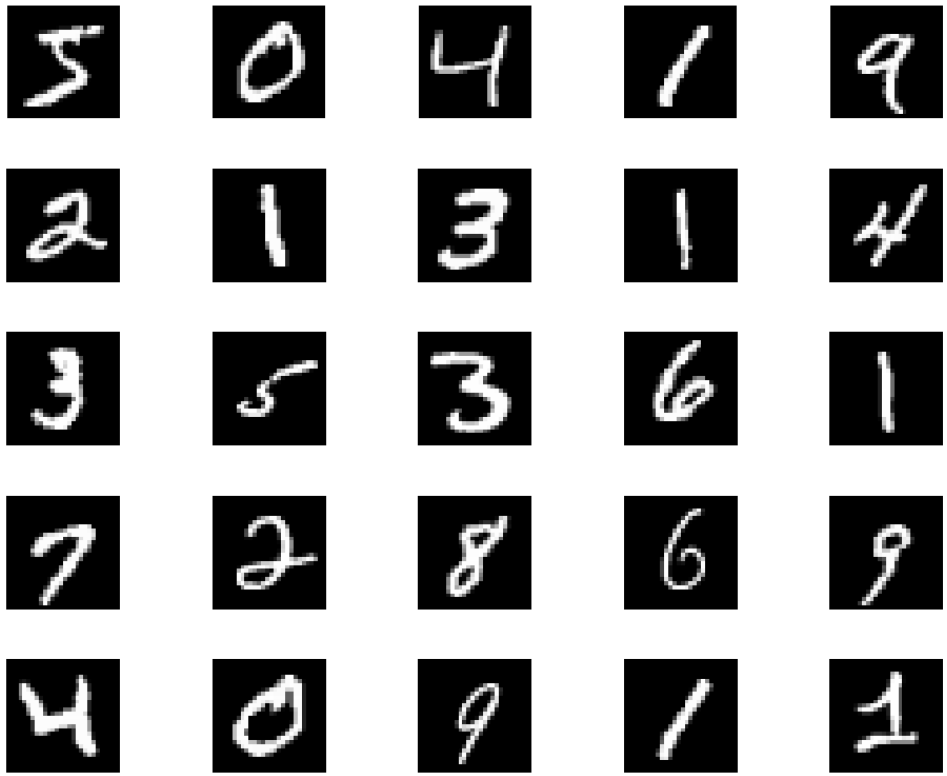
Convergence



Outline

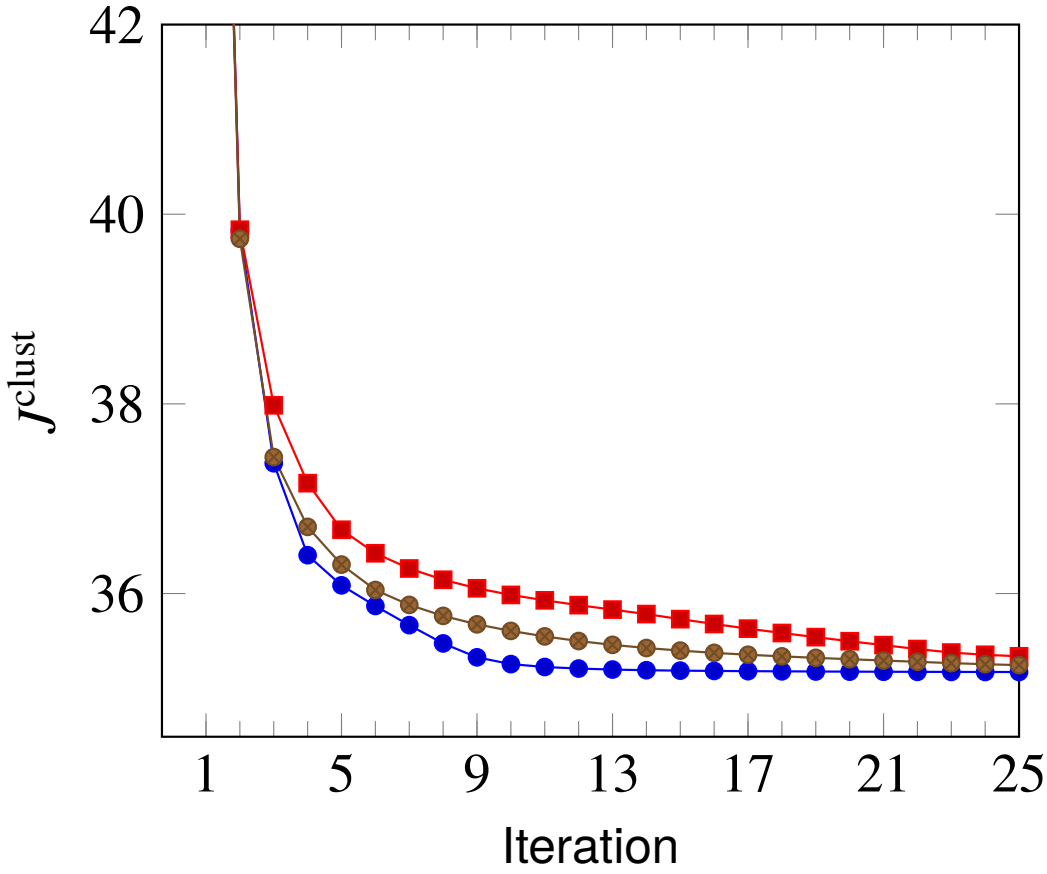
Handwritten digit image set

- MNIST images of handwritten digits (via Yann Lecun)
- $N = 60,000$ 28×28 images, represented as 784-vectors x_i
- 25 examples shown below

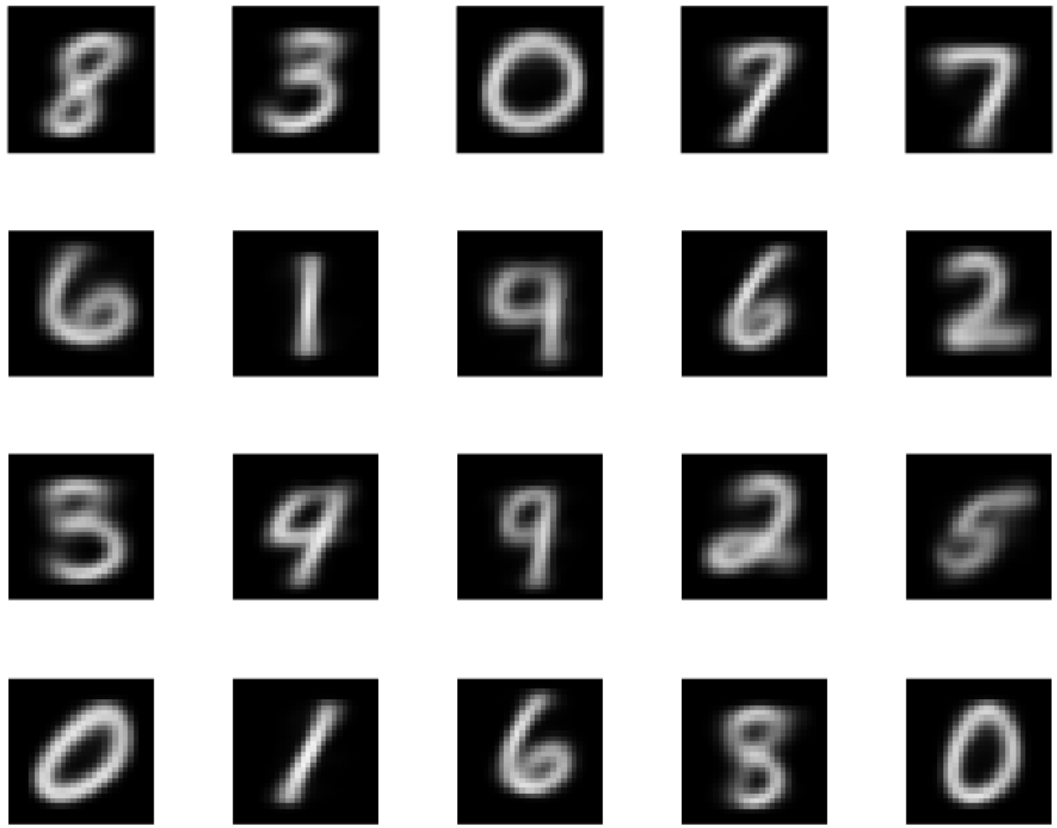


k-means image clustering

- $k = 20$, run 20 times with different initial assignments
- convergence shown below (including best and worst)

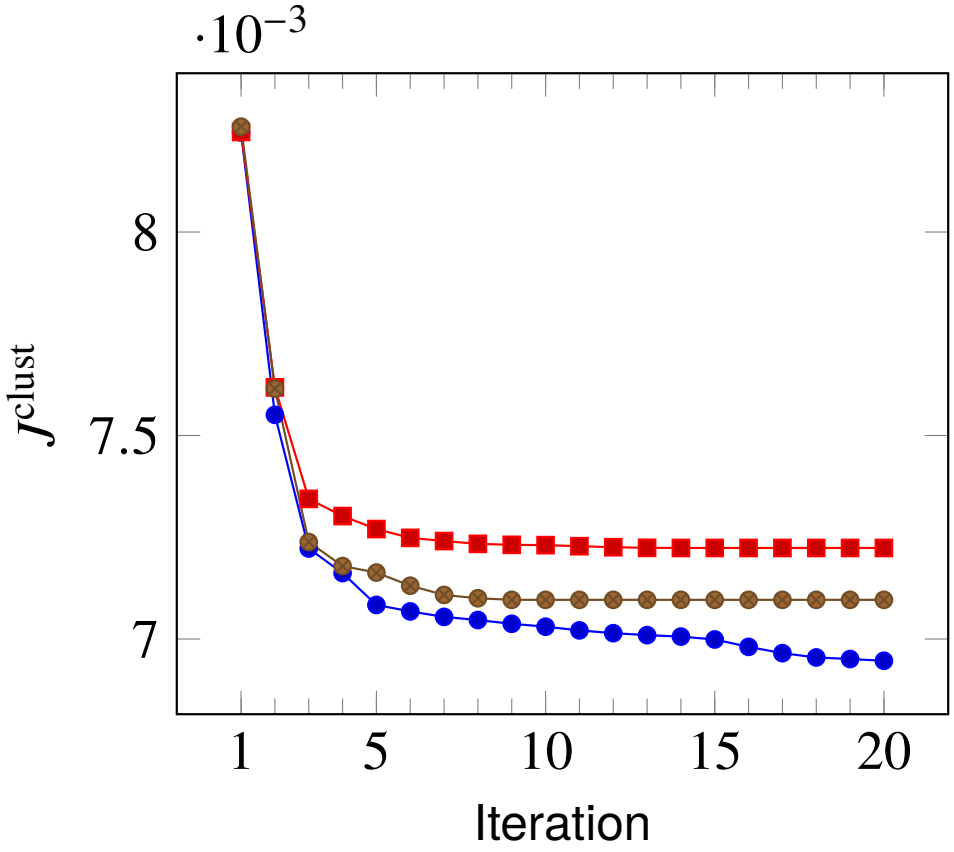


Group representatives, best clustering



Topic discovery

- $N = 500$ Wikipedia articles, word count histograms with $n = 4423$
- $k = 9$, run 20 times with different initial assignments
- convergence shown below (including best and worst)



Topics discovered (clusters 1–3)

- 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”, . . .

Topics discovered (clusters 4–6)

- words with largest representative coefficients

Cluster 4		Cluster 5		Cluster 6	
Word	Coef.	Word	Coef.	Word	Coef.
album	0.031	game	0.023	series	0.029
release	0.016	season	0.020	season	0.027
song	0.015	team	0.018	episode	0.013
music	0.014	win	0.017	character	0.011
single	0.011	player	0.014	film	0.008

- titles of articles closest to cluster representative

4. “David Bowie”, “Kanye West” “Celine Dion”, “Kesha”, “Ariana Grande”, “Adele”, “Gwen Stefani”, “Anti (album)”, “Dolly Parton”, “Sia Furler”, . . .
5. “Kobe Bryant”, “Lamar Odom”, “Johan Cruyff”, “Yogi Berra”, “José Mourinho”, “Halo 5: Guardians”, “Tom Brady”, “Eli Manning”, “Stephen Curry”, “Carolina Panthers”, . . .
6. “The X-Files”, “Game of Thrones”, “House of Cards (U.S. TV series)”, “Daredevil (TV series)”, “Supergirl (U.S. TV series)”, “American Horror Story”, . . .

Topics discovered (clusters 7–9)

- words with largest representative coefficients

Cluster 7		Cluster 8		Cluster 9	
Word	Coef.	Word	Coef.	Word	Coef.
match	0.065	film	0.036	film	0.061
win	0.018	star	0.014	million	0.019
championship	0.016	role	0.014	release	0.013
team	0.015	play	0.010	star	0.010
event	0.015	series	0.009	character	0.006

- titles of articles closest to cluster representative

7. “Wrestlemania 32”, “Payback (2016)”, “Survivor Series (2015)”, “Royal Rumble (2016)”, “Night of Champions (2015)”, “Fastlane (2016)”, “Extreme Rules (2016)”, . . .
8. “Ben Affleck”, “Johnny Depp”, “Maureen O’Hara”, “Kate Beckinsale”, “Leonardo DiCaprio”, “Keanu Reeves”, “Charlie Sheen”, “Kate Winslet”, “Carrie Fisher”, . . .
9. “Star Wars: The Force Awakens”, “Star Wars Episode I: The Phantom Menace”, “The Martian (film)”, “The Revenant (2015 film)”, “The Hateful Eight”, . . .