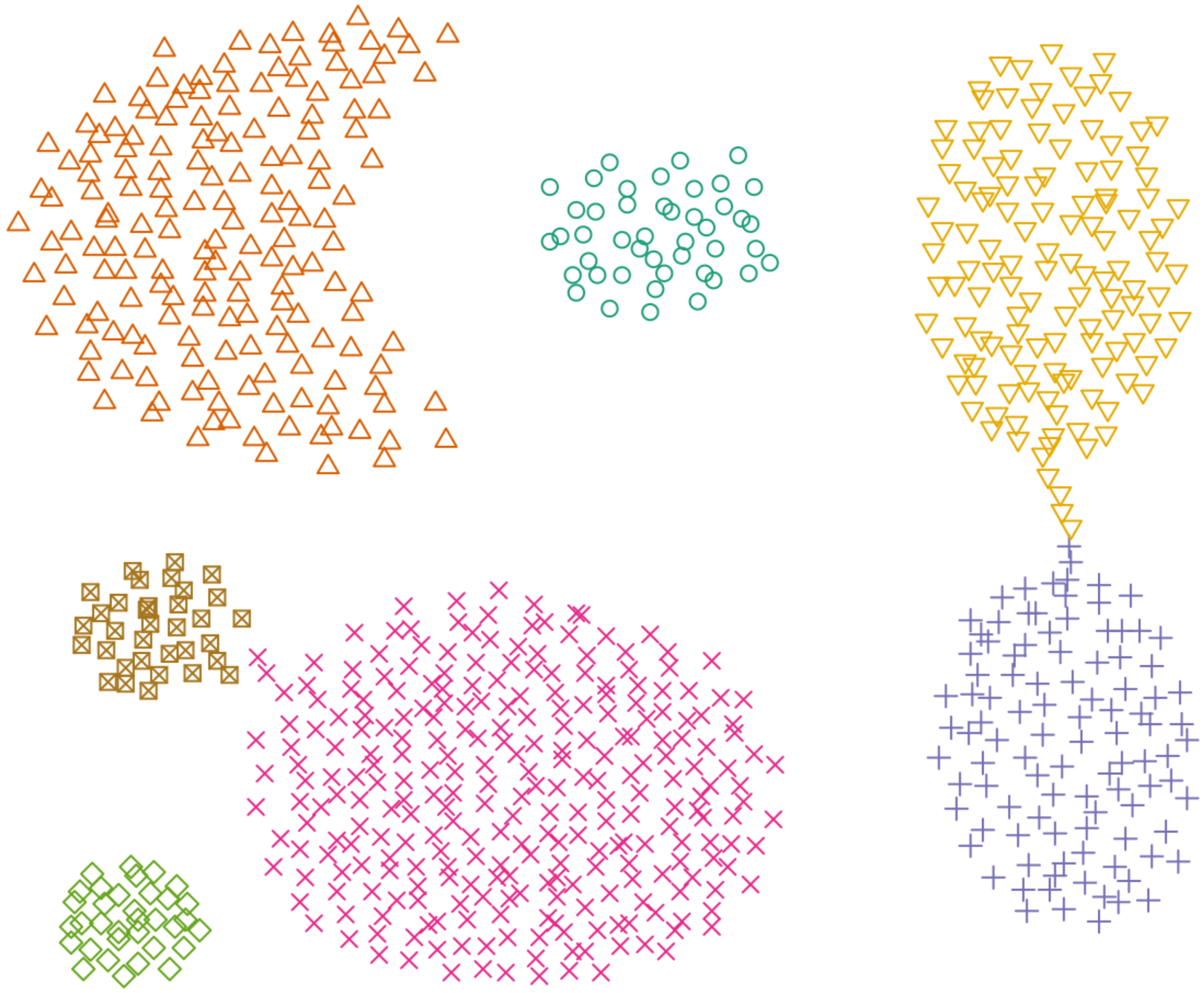
Unsupervised Machine-Learning

Alina Arseniev-Koehler



Outline:

1. ~20 min big picture & concepts
2. ~60 min guided coding, focusing on k-means clustering
3. break
4. ~20 min concepts & applications
5. ~15 min brainstorming applications

Questions/discussion throughout!

Concepts (Part 1)

Unsupervised vs supervised ML

- **Unsupervised:**
  - You want to discover latent structure in your data
  - E.G., you see a diversity of depression symptoms in individuals and want to see if there are “types” of depression
  - Unlabeled

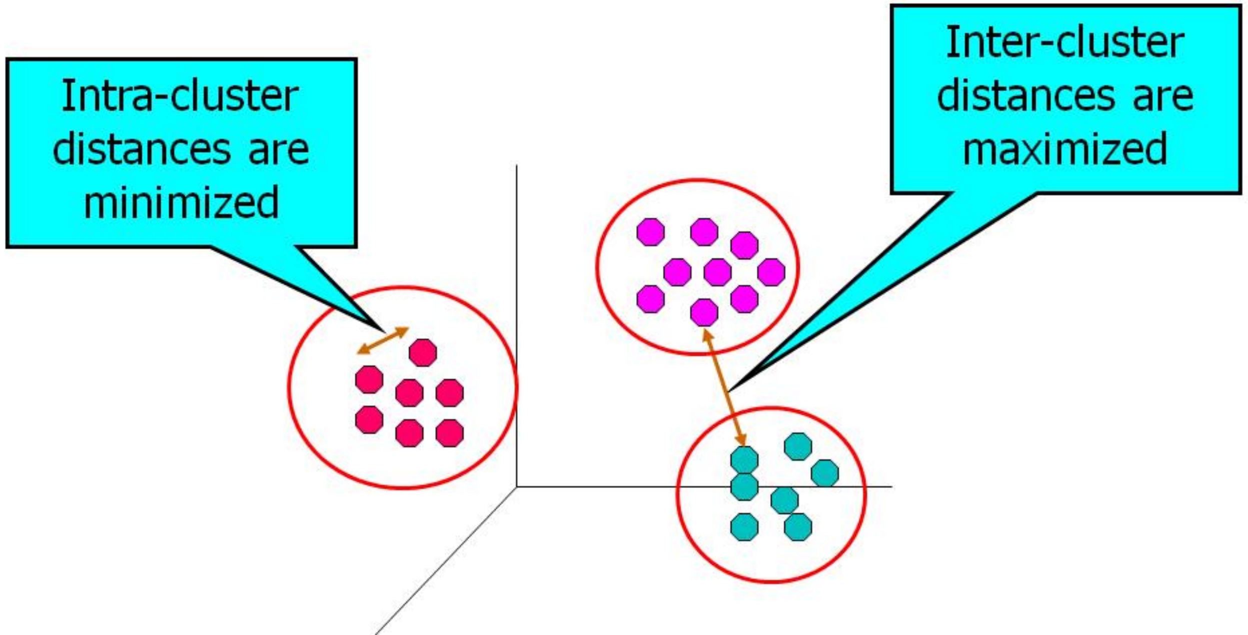
Unsupervised vs supervised ML

- **Unsupervised:**
  - You want to discover latent structure in your data
  - E.G., you see a diversity of depression symptoms in individuals and want to see if there are "types" of depression
  - Unlabeled

- **Supervised:**
  - You want to develop a model to predict Z from W, X, and Y
  - E.G., predict depression level from age, city, race, gender, #friends
  - Labeled (depression score, train)

Clustering Concepts

- Trying to find typologies, or “clusters” of data.
  - But what is a cluster?
- **A cluster of data points are more similar to each other than to points in other clusters**



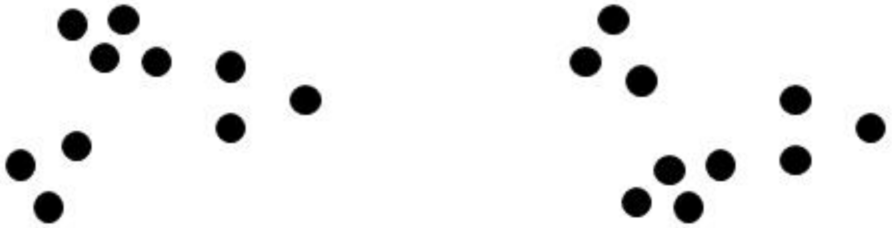
Intra-cluster distances are minimized

Inter-cluster distances are maximized

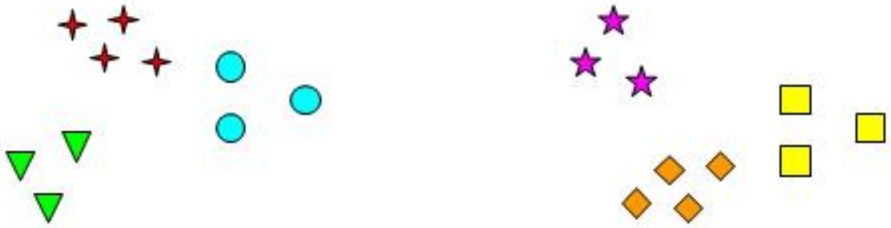
Supervised vs Unsupervised ML

- **Unsupervised:**
  - You want to discover latent structure in your data
  - E.G., you see a diversity of mental health symptoms in individuals and want to see if there are "types" of illnesses
  - Unlabeled

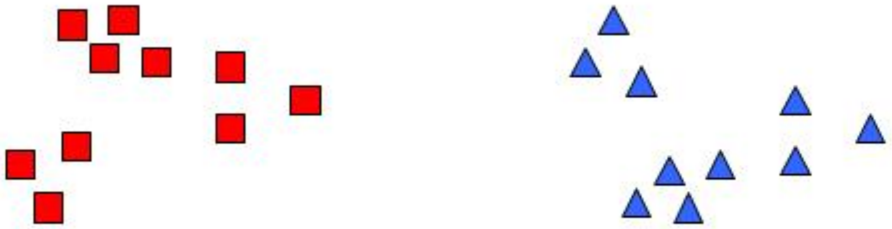
- **Supervised:**
  - You want to develop a model to predict Z from W, X, and Y
  - E.G., in a new population, classify symptoms into learned types of illnesses
  - Labeled (labels=illness)



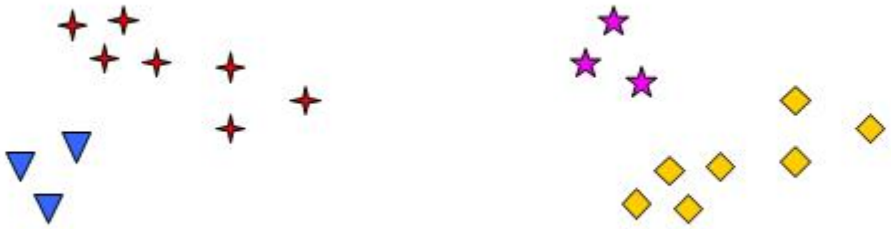
How many clusters?



Six Clusters



Two Clusters



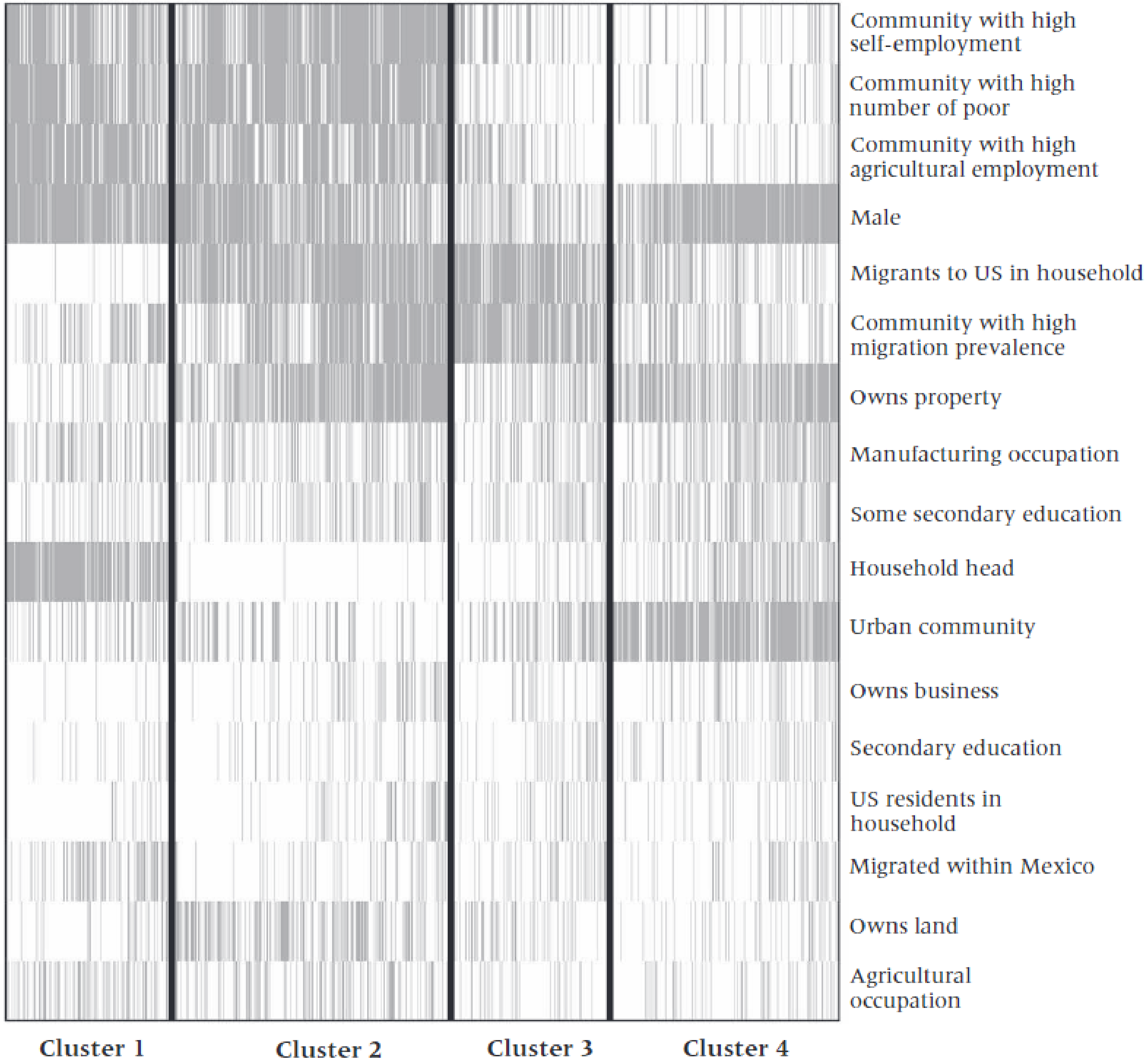
Four Clusters

Sample Research Application

- Garip, Filiz. "Discovering diverse mechanisms of migration: The Mexico–US Stream 1970–2000."
    - “types” of migrants

"…a representative migrant in the second cluster is the son of the house-hold head and has only primary education. He lives in a poor community, where the assets of his household, a property and either a piece of land or a business, place him in the middle or upper wealth category. Given his substantial economic endowments, we posit that this person migrates to diversify the risks to those endowments in the context of Mexico's volatile economic climate. While the migrant, labeled a "risk diversifier" in line with the new economics theory, secures earnings in the United States, the other members of his household, typically the head, manage the roles of subsistence in Mexico. A risk diversifier is expected to migrate temporarily at times of high economic uncertainty. This expected pattern, which we demonstrate in subsequent analysis, is probably facilitated by the migrant's ties to earlier migrants to the US in the family or community." (Garip)

FIGURE 3 Heat map of migrant attributes by cluster membership



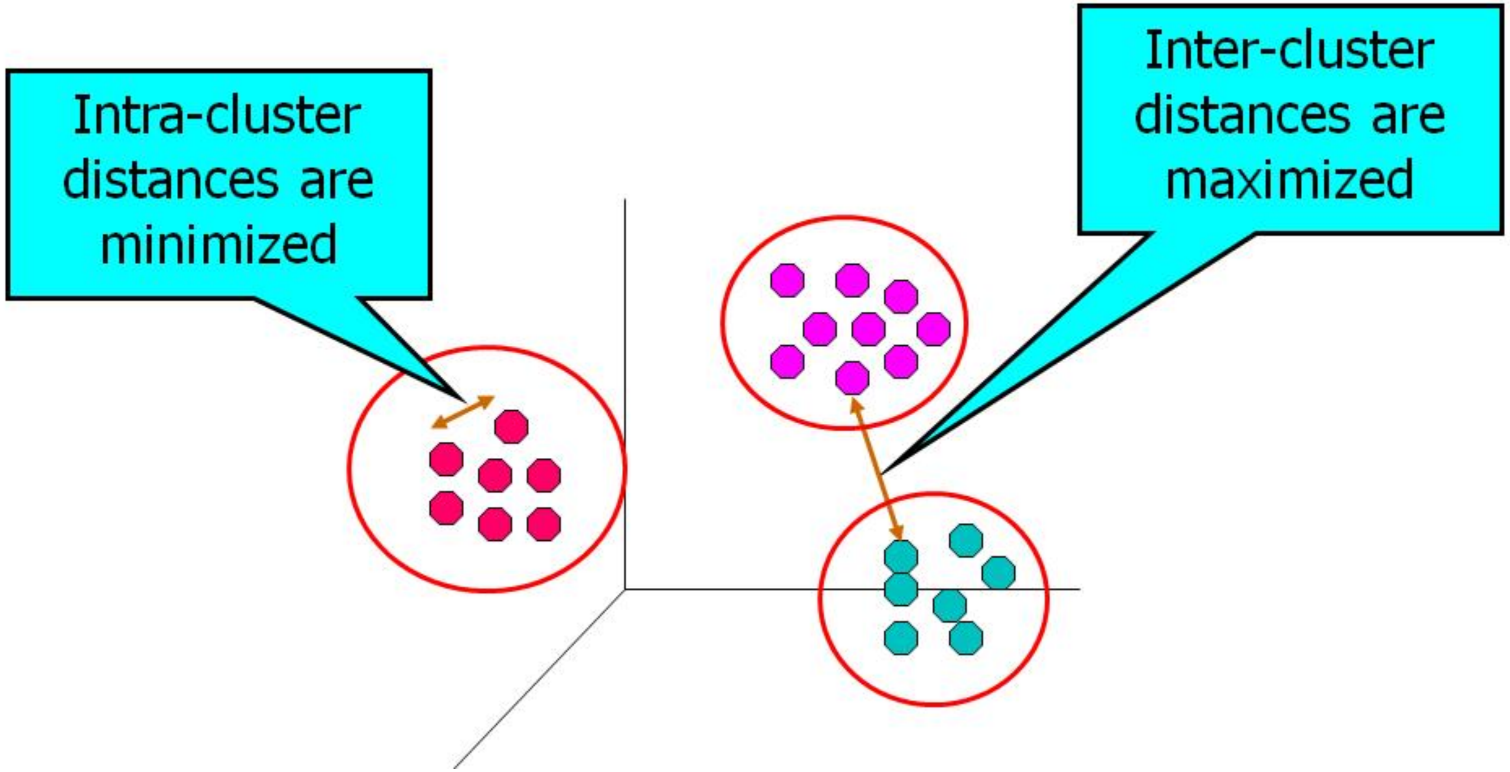
NOTE: The heat map color codes the attributes (rows) of all migrants (columns). Gray indicates the presence of the attribute, and white indicates its absence. The vertical black lines separate the four clusters identified with cluster analysis.

Sample Research Applications

- **Clustering for your research interests?**

Clustering Algorithm: K-means

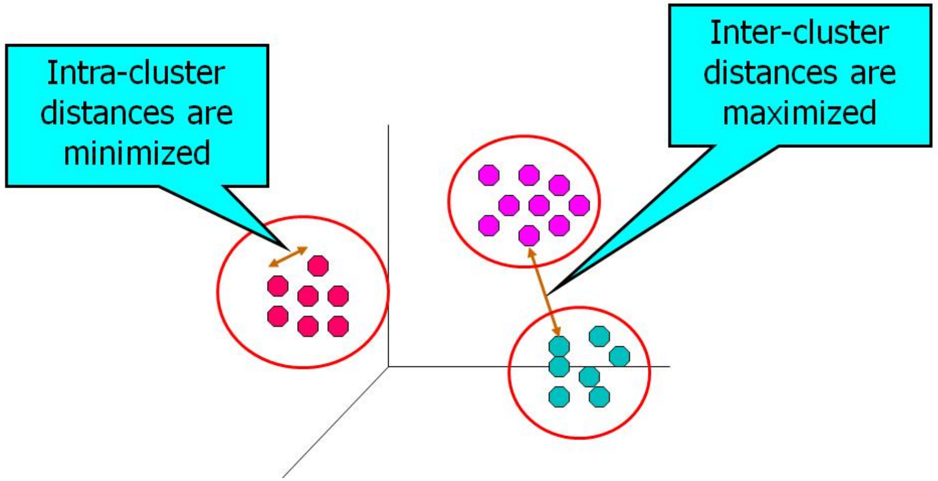
- Clustering goal:



Intra-cluster distances are minimized
Inter-cluster distances are maximized

https://uc-r.github.io/kmeans_clustering

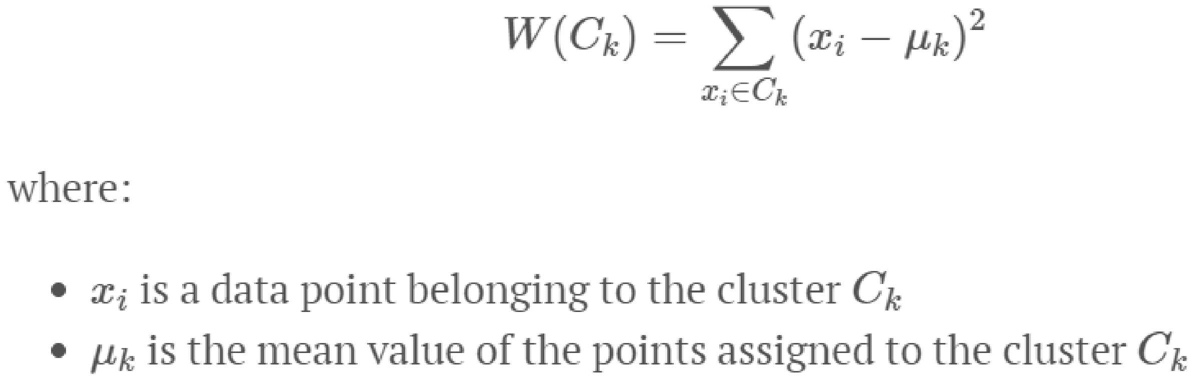
Clustering Algorithm: K-means



Intra-cluster distances are minimized

Inter-cluster distances are maximized

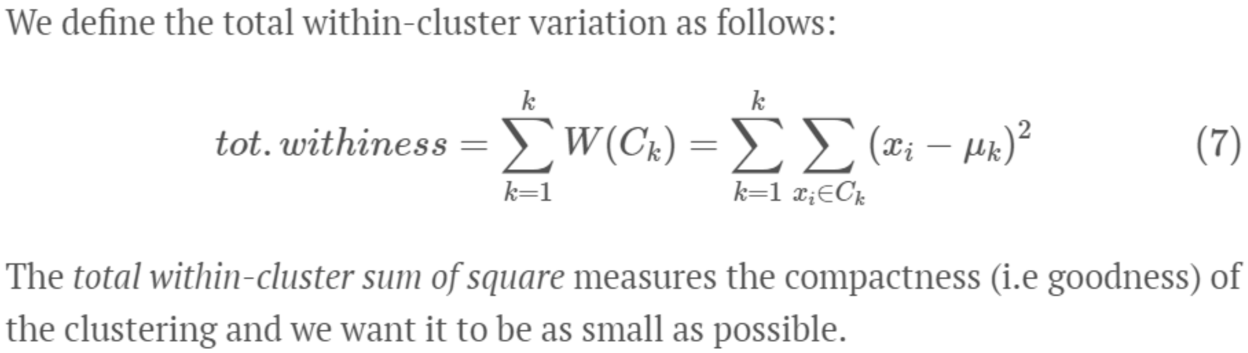
- Minimize errors within clusters



$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

where:

- x_i is a data point belonging to the cluster C_k
- μ_k is the mean value of the points assigned to the cluster C_k



We define the total within-cluster variation as follows:

$$tot.\,withiness = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \qquad (7)$$

The *total within-cluster sum of square* measures the compactness (i.e goodness) of the clustering and we want it to be as small as possible.

https://uc-r.github.io/kmeans_clustering

Clustering Algorithm: K-means

1. Randomly pick k centroids from the sample points as initial cluster centers
2. Assign each sample to the nearest centroid, to yield clusters
3. Recalculate the new centroids, based on the samples that were assigned to each cluster
4. Repeat 2 and 3 until the cluster assignments do not change
   - (or, don't change within a tolerance, or reach max # iterations)

(Sebastian Raschka, Python machine learning)

Watch: https://www.youtube.com/watch?v=4b5d3muPQmA

Clustering Algorithm: K-means

- How do we measure "nearest?"

- Squared Euclidean Distance between vectors **X** and **Y**:

$$d(x, y)^{\,2} = \sum_{j=1}^{m} (x_j - y_j)^2 = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

Clustering Algorithm: K-means

- How do we measure “nearest?”

- Squared Euclidean Distance between vectors **X** and **C**:

X	C
x_1	c_1
x_2	c_2
x_3	c_3

$$dist(x,c)^{2} = \sum_{j=1}^{m} (x_j - c)^2 = \|\boldsymbol{x} - \boldsymbol{c}\|_2^2$$

$$= (x_1 - c_1)^2 + (x_2 - c_2)^2 + (x_3 - c_3)^2$$

Clustering Algorithm: K-means

- How do we measure “nearest?”

- Squared Euclidean Distance between vectors **X** and **C**:

X	C
x_1	c_1
x_2	c_2
x_3	c_3

$$dist(x,c)^{\,2} = \sum_{j=1}^{m}(x_j - c)^2 = \|\boldsymbol{x} - \boldsymbol{c}\|_2^2$$

$$SSE = \sum_{i=1}^{K} \sum_{x \in c_i} dist(x, c_i)^2$$

Input/output of clustering

- Input: matrix of attributes
  - for k-means, continuous
  - for categorical, or mixed data, try other algorithms

- Outputs:
  - Each data point is assigned to a cluster
  - The “average” case (centroid) of each cluster

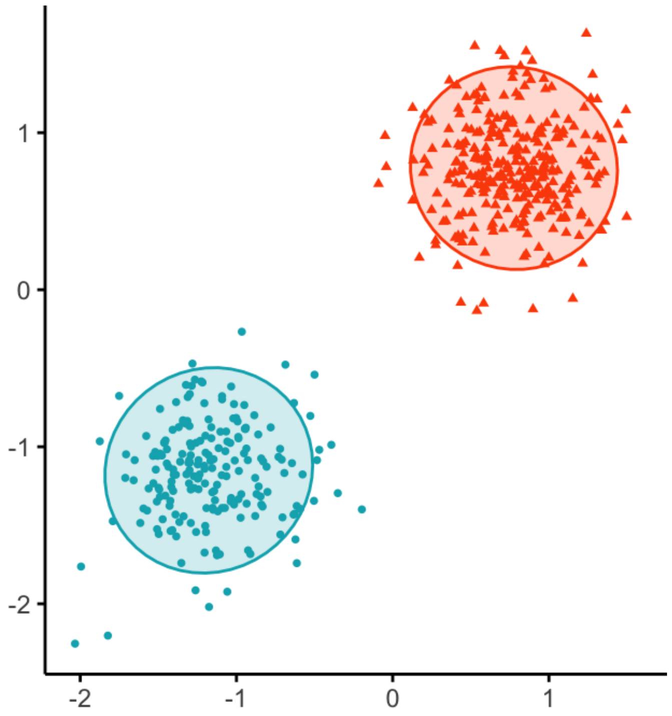
Questions?

Next: Try out k-means in Python

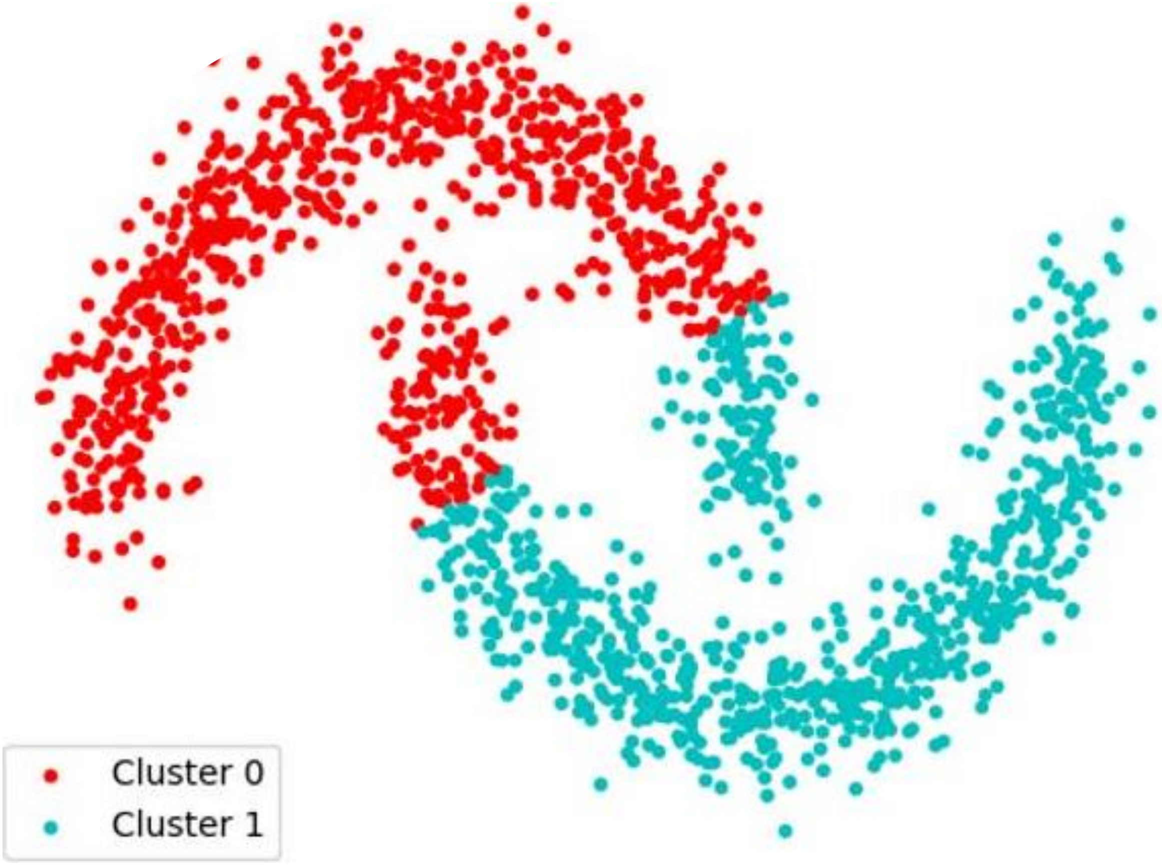
Concepts (Part 2)

Clusters come in all shapes and sizes

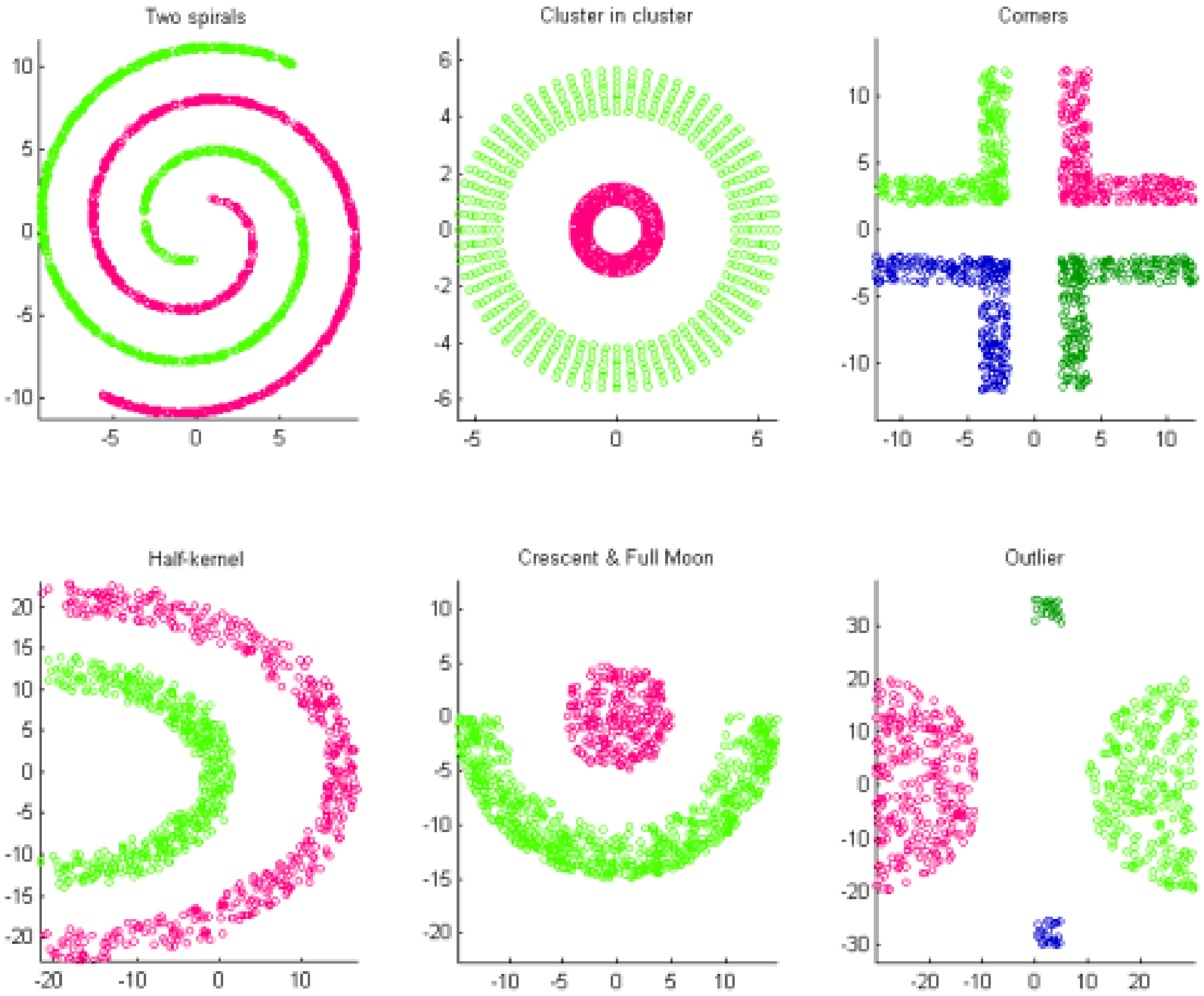
K-means assumes data looks like::



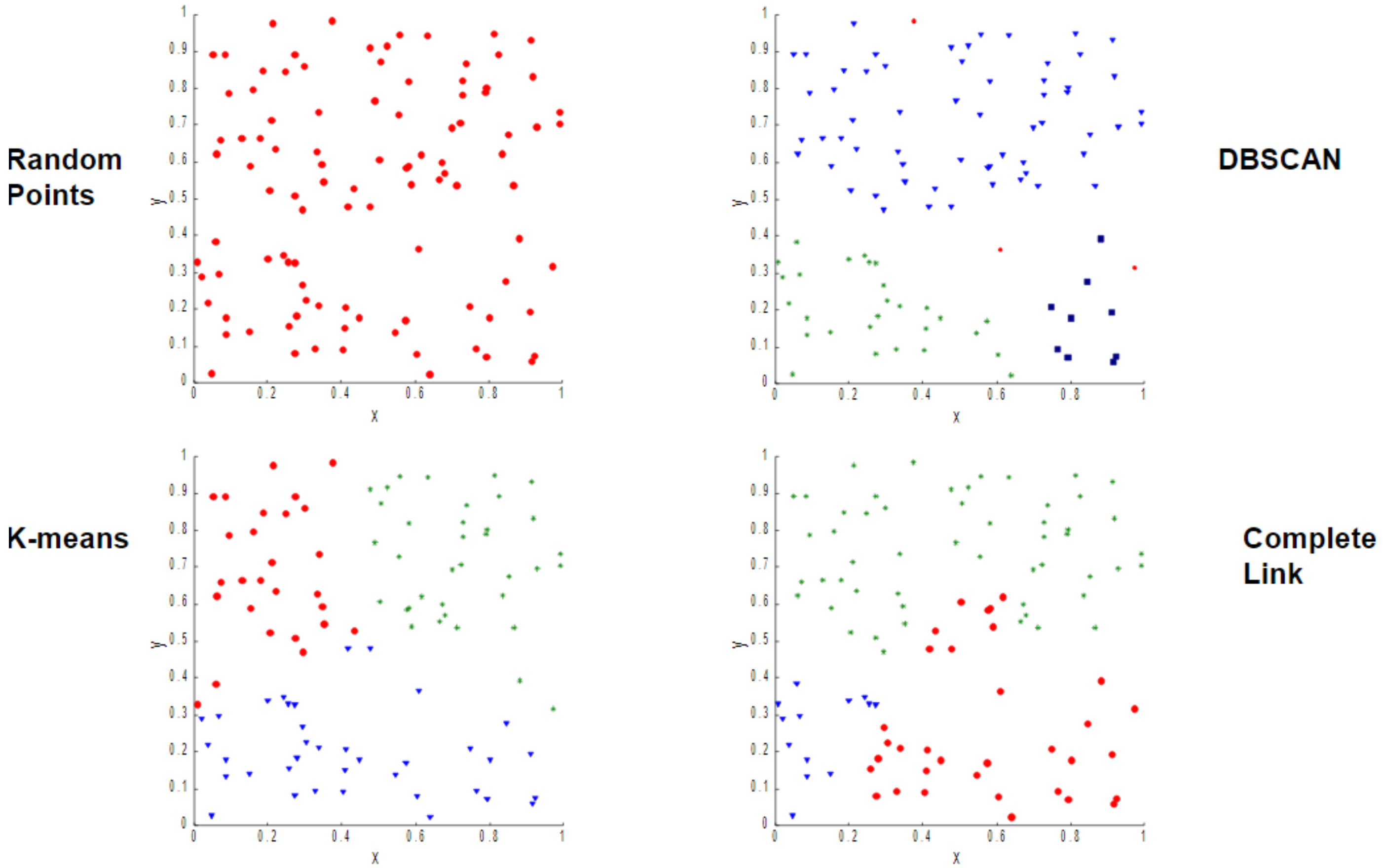
But data might look like:



Clusters come in all shapes and sizes



Challenges: Am I just imagining things?



Random Points

DBSCAN

K-means

Complete Link

Image: http://www.cs.kent.edu/~jin/

Challenges

- Often, different algorithms and parameters produce different solutions (any objective truth?)

- How do we know if the clusters are "real"?
    - We don't have labels on what cluster each data point is supposed to belong to (usually)

- How do we know how many clusters we should expect?

Sample: Polysemy

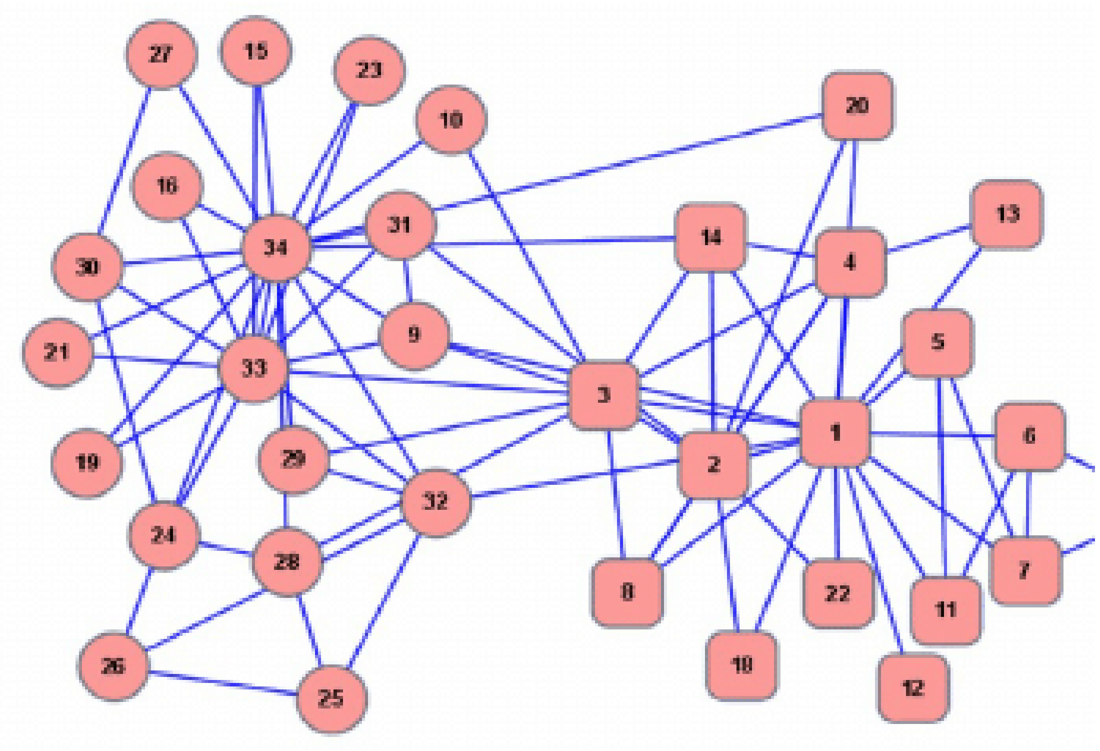
- Alina: What are the meanings of "fat"? Cluster the context words of "fat"
  - Expect 3 clusters:
    - Cluster 1: *"A fat little puppy" "Americans are not becoming more and more fat"*
    - Cluster 2: *"I prefer full-fat milk" "Avocados have a lot of fat" "Fries have the bad fat"*
    - Cluster 3: *"Fat and proud of it" "Fat, fit and healthy"*
  - Also found: "fat" as a texture

- Generalize→ What are the 'senses' of words? How many senses/word?

How do we know if clusters are "real"?

- Extrinsic validity

- Clustering in a social network (e.g., finding groups of friends)
    - Maybe these clusters predict group fission? E.g., karate club



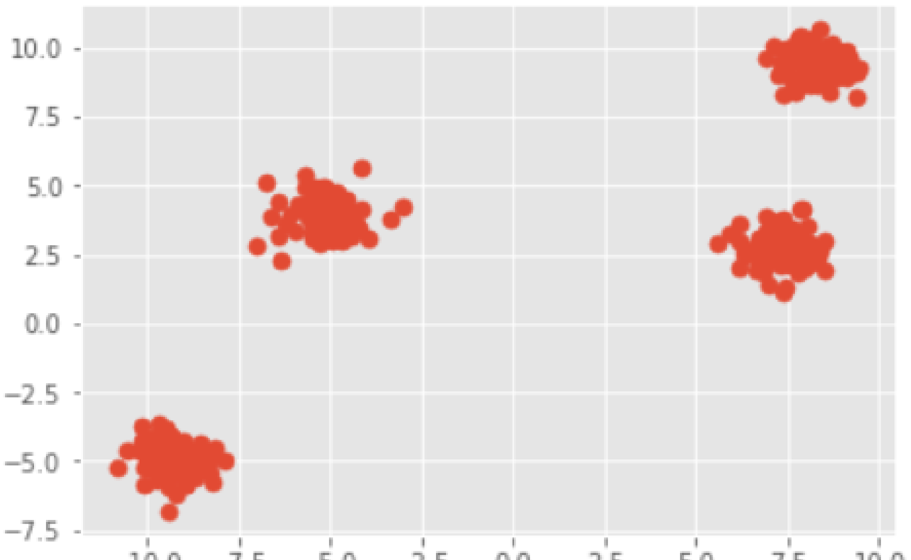
How do we know if clusters are “real”?

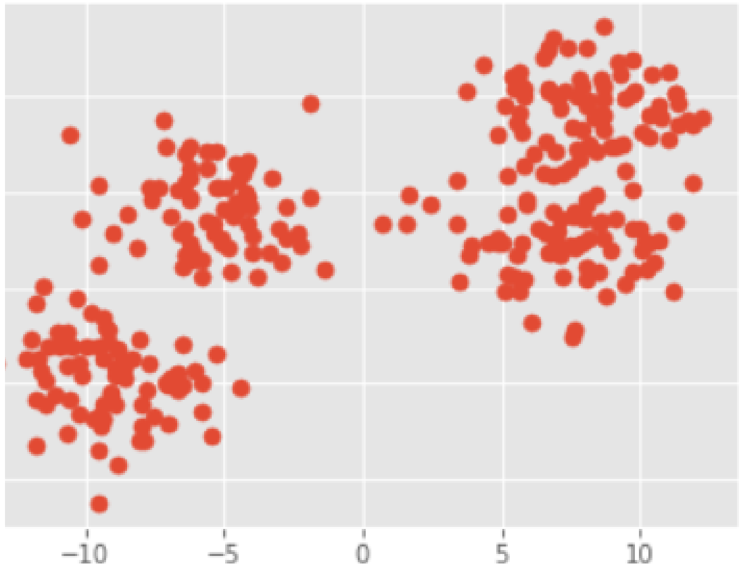
- Intrinsic validity

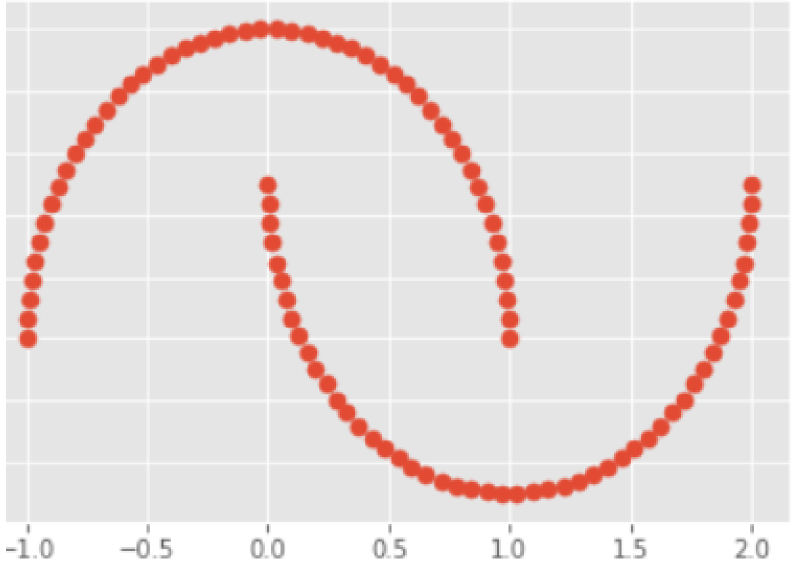
- Differences *within* clusters (SSE) are minimized

How do we know if clusters are “real”?

- To test quality of clustering method- simulate data!







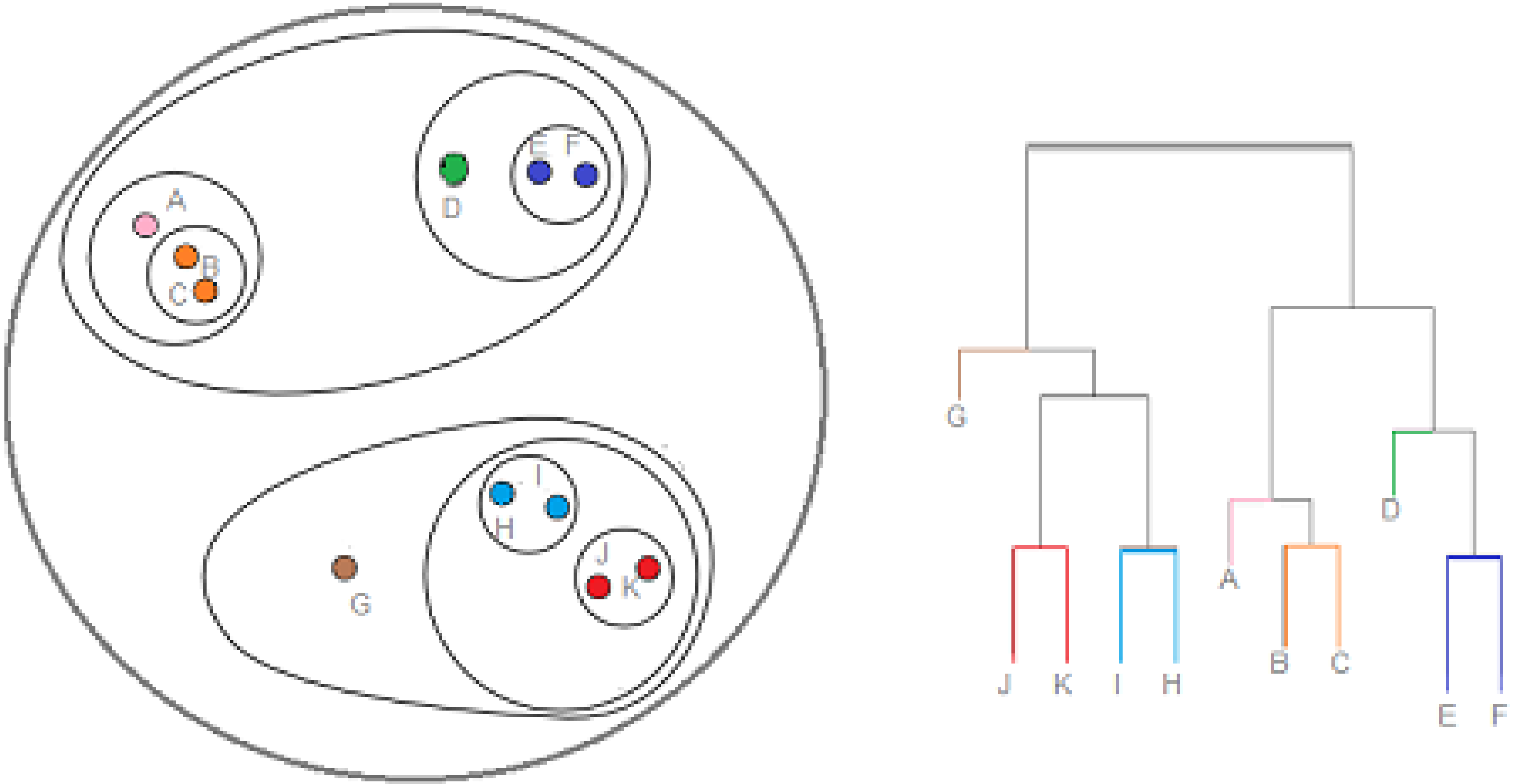
Variants of Clustering Algorithms

- “Hard” vs “soft” clustering

- Hard: each data point belongs to exactly 1 cluster
- Soft: each data point may belong to more than 1 cluster

Variants of Clustering Algorithms

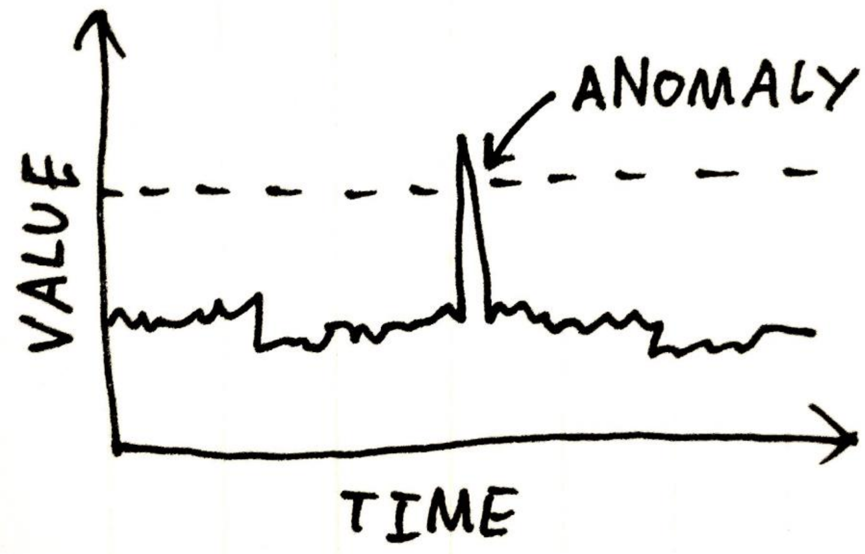
- Hierarchical (clusters are subsets of larger clusters)

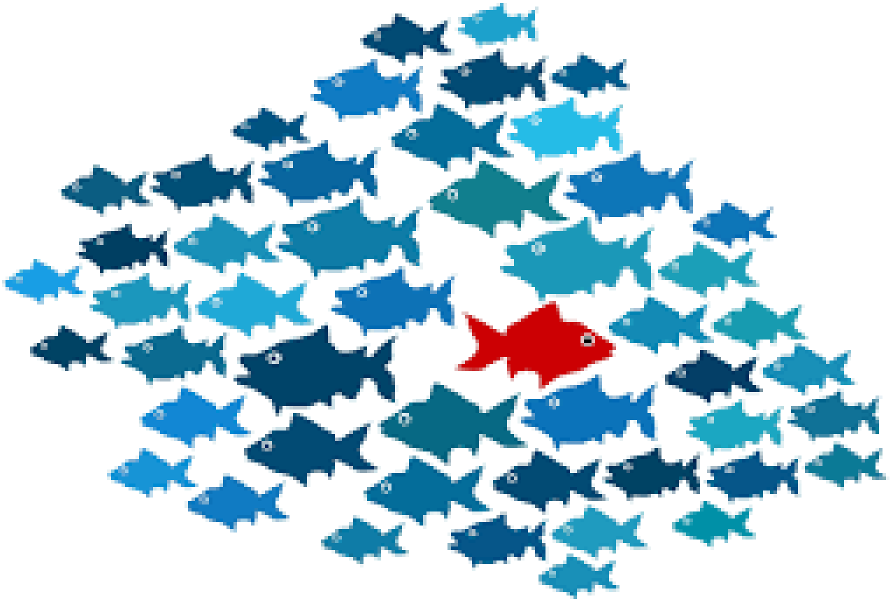


You’ve been clustering your whole life…

- Humans cluster all the time!
- Too many things in the world, clustering = meaning-making
- Examples (hard or soft? Hierarchical or not?):
  - breakfast, lunch, vs dinner foods
  - discovering your research interests in grad school
  - finding “codes” or “themes” in qualitative data analysis
  - psychiatric definitions of mental illnesses (clusters of symptoms)
  - **With your neighbor:** in your 1-3 more examples- are they hard/soft? Hierarchical?

Anomaly Detection





1. Predict value at t+1 from t
2. If your prediction at any time point is *really off*: it might be an anomaly in the data!

- *More generally:* predict a thing from it's context

Dimensionality Reduction

- Sometimes, too much data! (or noise)
- Reduce with:
    - Principal Component Analysis, Singular Value Decomposition, etc.

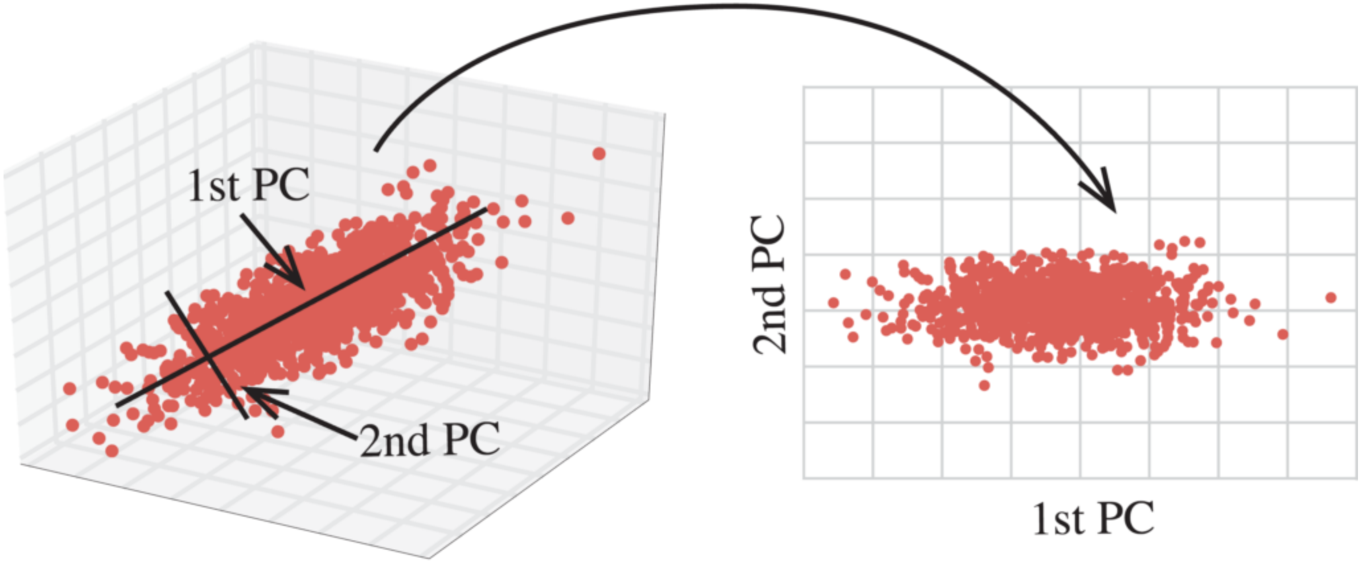
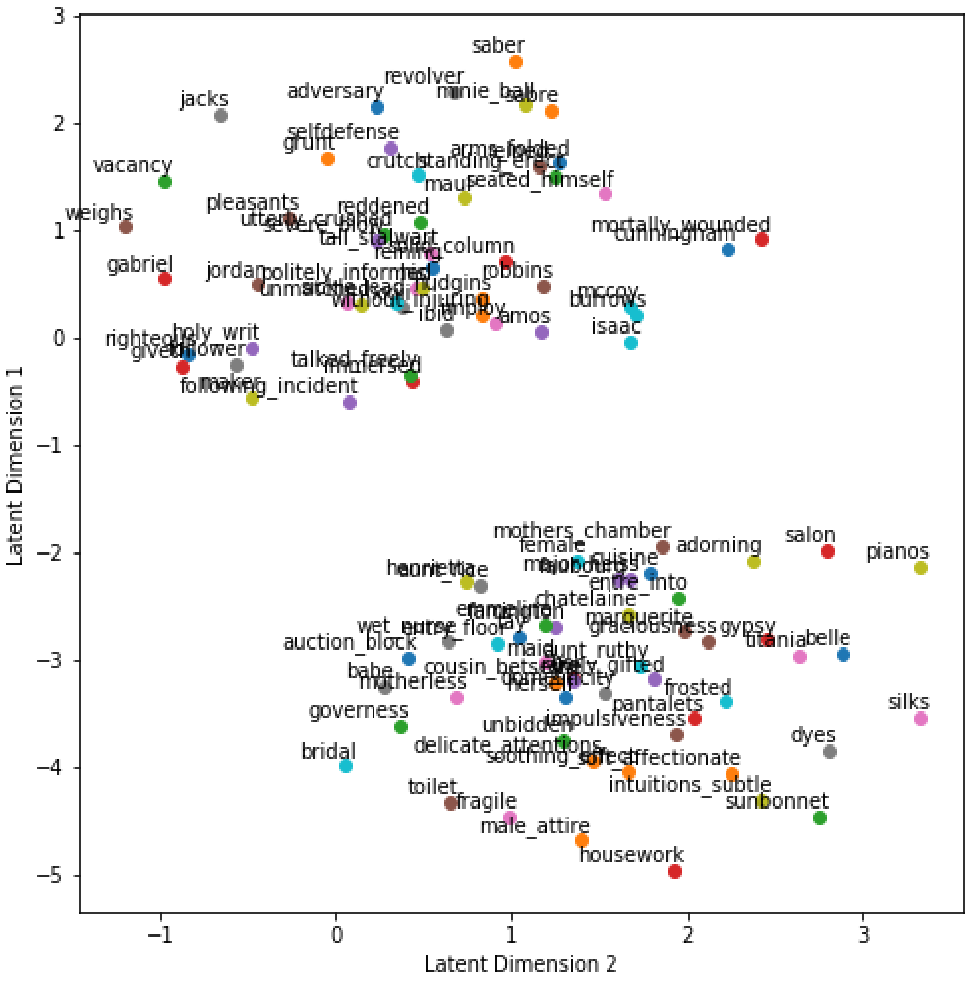


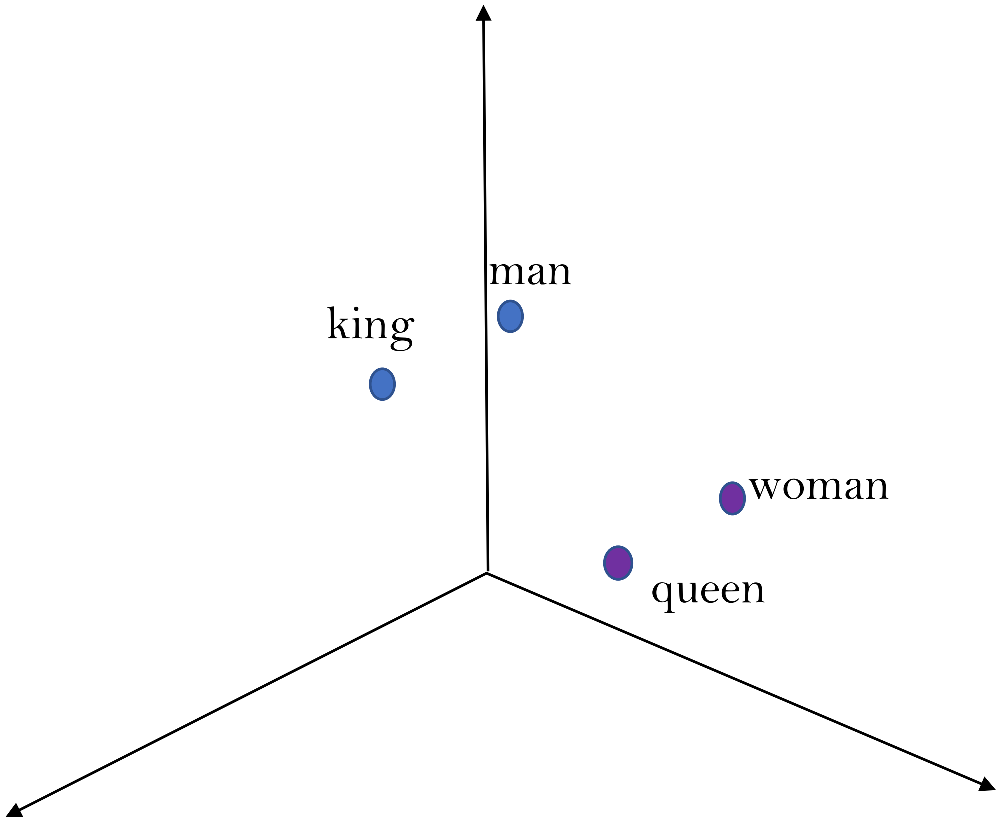
Image: http://blog.kaggle.com/2017/04/10/exploring-the-structure-of-high-dimensional-data-with-hypertools-in-kaggle-kernels/

Supervised or unsupervised?

Word2Vec –
creating labels from context,
learning latent meaning
structures



Word Embeddings



words with more similar meaning are closer in space

closeness= cosine similarity

Meaning of a concept is distributed along three dimensions

Vocabulary Word	Dimension_1	Dimension_2	Dimension_3
King	-.07284	.383918	.0694749
Queen	0.2203	0.03286	-0.032342
Man	0.027485	0.4286	-0.103234
Woman	.28933	.11193	-.11947
Womanly	.9284	-.0535	.10324
Uncle	.4822	.935	-.3531
Friend	-.2842	-.39545	.50225
Girlfriend	.482	.4240	.02841
Neighbor	-.5025	.5018	.9105
…	…	…	…

How does Word2Vec learn word-vectors?

Can you guess the missing word?

“Americans have grown ______ over the last generation, inviting more heart disease, diabetes and premature deaths…”

Answer: fatter

“a word is characterized by the company it keeps” (Firth 1957)

Mikolov, Tomas, et al. 2013

How does Word2Vec learn word-vectors?

Can you guess the missing word?

“ $\overrightarrow{Americans}$ $\overrightarrow{have}$ $\overrightarrow{grown}$ _____
$\overrightarrow{over}$ $\overrightarrow{the}$ $\overrightarrow{last}$ …”

What word is most likely to be the missing word?

→What word has the highest *cosine similarity* to the context words?

How does Word2Vec learn word-vectors?

Can you guess the missing word?

“ $\overrightarrow{Americans}$ $\overrightarrow{have}$ $\overrightarrow{grown}$ ______ $\overrightarrow{over}$ $\overrightarrow{the}$ $\overrightarrow{last}$ …”

What word has the highest *cosine similarity* to the context words?

→We know what the missing word actually is in the NYT (“fatter”)

1. Word2Vec gives correct answer? Then we have good word-vectors
2. Wrong answer? Tweak the word-vectors

Culture, Cognition, and Computational Soc

- Use computational methods to inform (or model) cognitive underpinnings of culture
    - Word2vec example – *learning* meaning (private culture from public culture)

- Clustering
    - Compare clustering in machines & humans, to learn about both
    - Prototype thy vs exemplar thy
    - Polysemy example – we can think of this as meaning-making?

- Anomaly detection
    - Is this what happens when interaction "gets awkward"?

15 min: Brainstorm Research Applications

With your neighbor:

- Do you have data (or an interest area) you could detect "typologies" in?
  - What does a "cluster" represent in your data?
- What are the levels of granularity at which you could cluster? (e.g., mammal vs reptile, or type of mammals)
- What might these clusters predict? (E.g., karate club clusters and group fission)
- How might your own beliefs about the clusters in your data compare to those found by a clustering algorithm?
- What "anomalies" might you find in your data?

Hands-on Clustering in Python

- Jupyter Notebook at: https://github.com/arsena-k/unsupMLclust