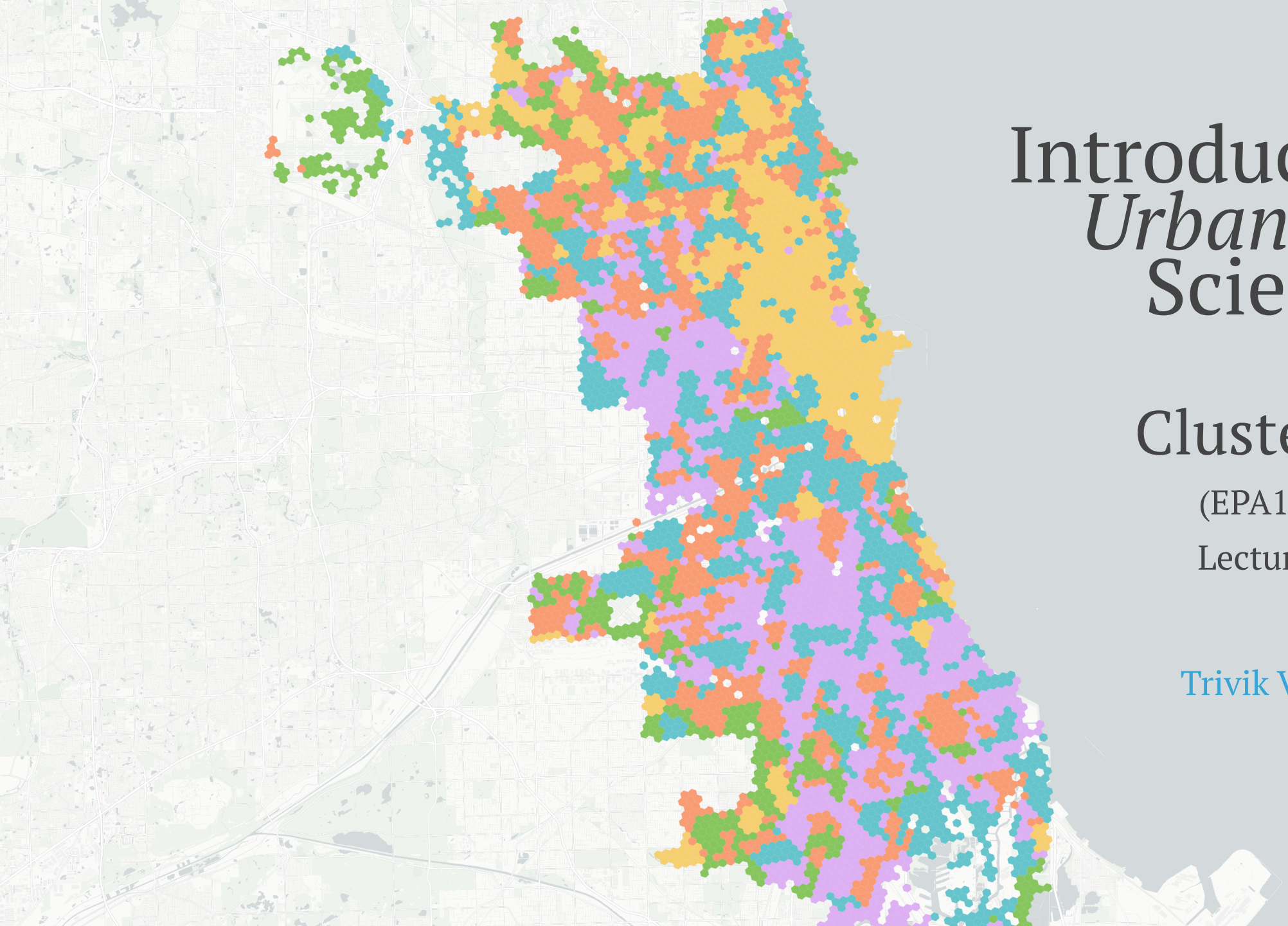


Introduction to *Urban* Data Science

Clustering

(EPA1316)

Lecture 11

Trivik Verma

Last Time

- Linear models
- Estimate of the regression coefficients
- Model evaluation
- Interpretation

Q: I have a pile of socks to sort but I forgot how many colours I own. What kind of learning task am I going to perform for the sorting?

A. Clustering
B. Classification
C. Regression
D. Normalisation

Q: What kind of task is spam-detection?

A. Unsupervised Learning
B. Supervised Learning

Today

- The need to group data
- Geodemographic analysis
- Non-spatial clustering
- Regionalization

The need to group data

The need to group data

- The world is **complex** and **multidimensional**
- **Univariate** analysis focuses on **only one** dimension
- Sometimes, world issues are best understood as **multivariate**. E.g.
  - Percentage of foreign-born Vs. *What is a neighbourhood?*
  - Years of schooling Vs. *Human development*
  - Monthly income Vs. *Deprivation*

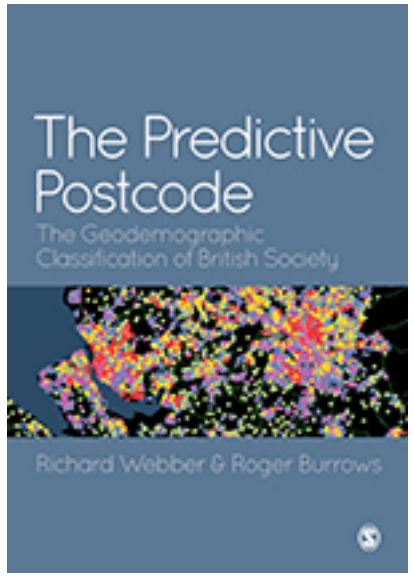
Grouping as simplifying

- Define a given number of categories based on **many characteristics** (multi-dimensional)
- Find the **category** where each observation *fits best*
- **Reduce complexity**, keep all the **relevant information**
- Produce easier-to-understand outputs

*Geo*demographic analysis

Geodemographic analysis

- 1970's, Richard Webber
- **Identify similar neighbourhoods**
  → Target urban deprivation funding
- **Public** Sector (policy) →
  Private sector (marketing and business intelligence)



TUDelft

Consumer Data Research Centre

An ESRC Data Investment

CDRC | Data | Maps | Indicators & Stories | NRDF

CDRC Maps

DATA CHOOSER

Geodem | Indicators | Metrics

Select a map: 2011 Area Classif/n of OAs

MAP OPTIONS

Layers: Land | Labels

Overlays: Pin | Clear

Tip: Try dropping KML or GeoJSON files onto map.

Postcode: Go

CENTRES & CATCHMENTS

JUMP TO CITY

Aberdeen | Birmingham | Brighton | Bristol | Cardiff | Edinburgh | Glasgow | Leeds | Liverpool | London | Manchester | Newcastle | Plymouth

Tweet

Important note: Classifications are an average across the local area, rather than for individual houses, therefore the colour coding on a building is not necessarily indicative of that building.

2011 AREA CLASSIF/N OF OAS

1000 m

Anfield

Liverpool

Birkenhead

Wavertree

Childwall

Sefton Park

About/Attribution

MAP KEY

2011 OAC

The Area Classification of Output Areas (OAC) 2011.

More info about this map

Download these data

- Rural Residents
- Cosmopolitans
- Ethnicity Central
- Multicultural Metropolitans
- Urbanites
- Suburbanites
- Constrained City Dwellers
- Hard-Pressed Living

[source]

How do you segment/cluster observations over space?

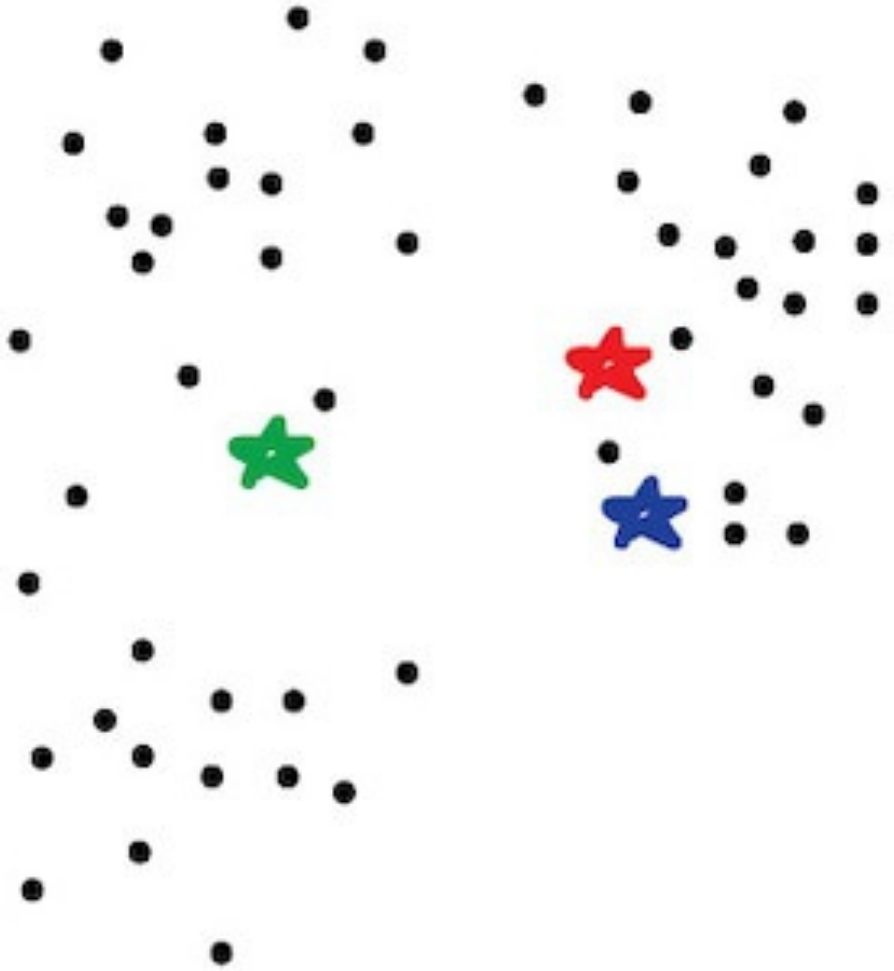
- Statistical clustering
- Explicit spatial clustering (regionalisation)

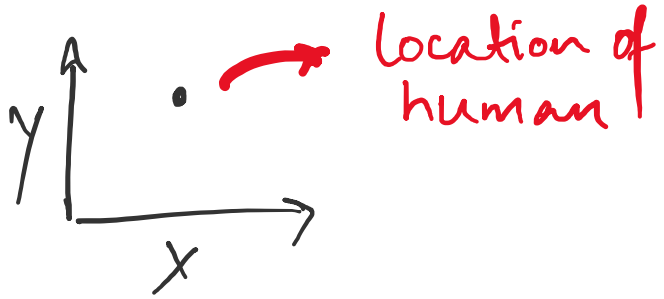
Non-spatial clustering

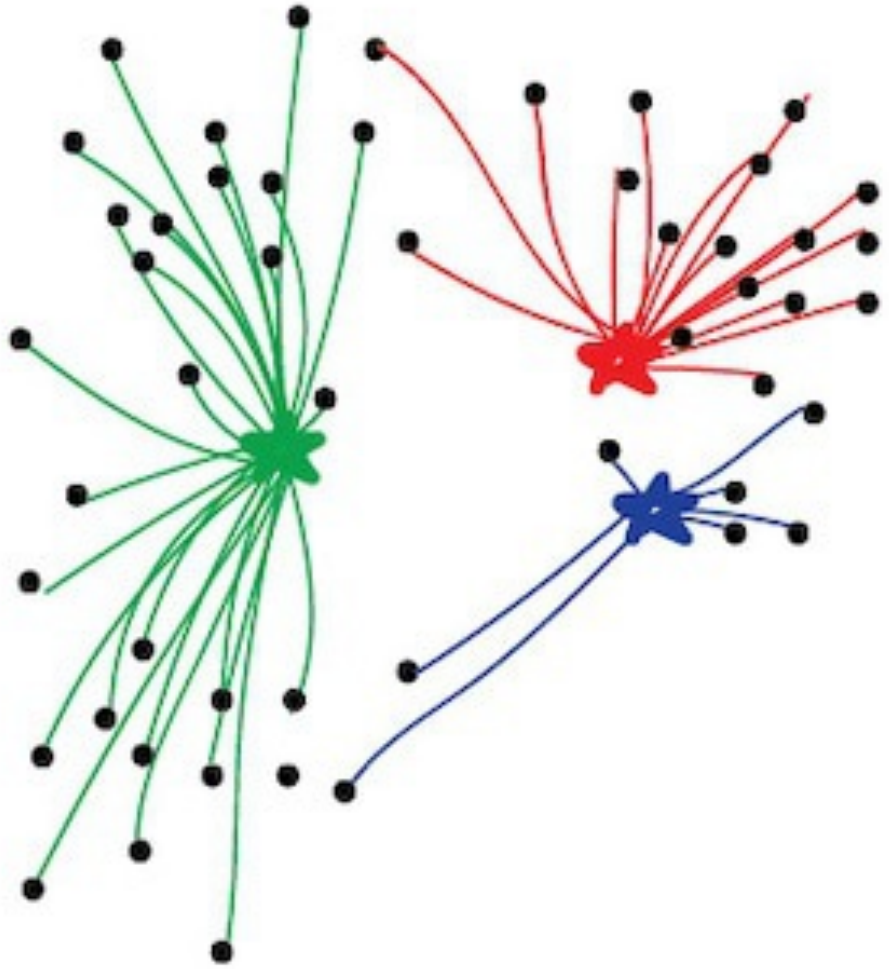
Split a dataset into **groups** of observations that are **similar** within the group and **dissimilar** between groups, based on a series of **attributes.**

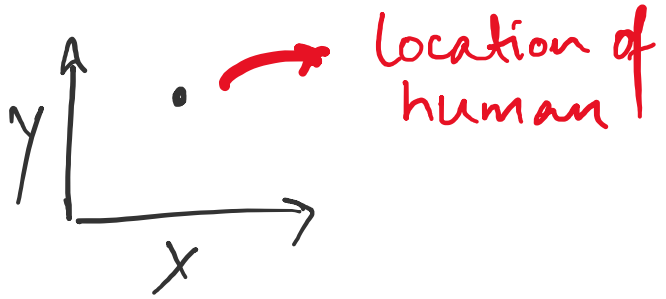
Q: Apple and Google Photos are looking for faces in photos to create albums of your friends. The app doesn't know how many friends you have and how they look, but it's trying to find the common facial features. What task is it?

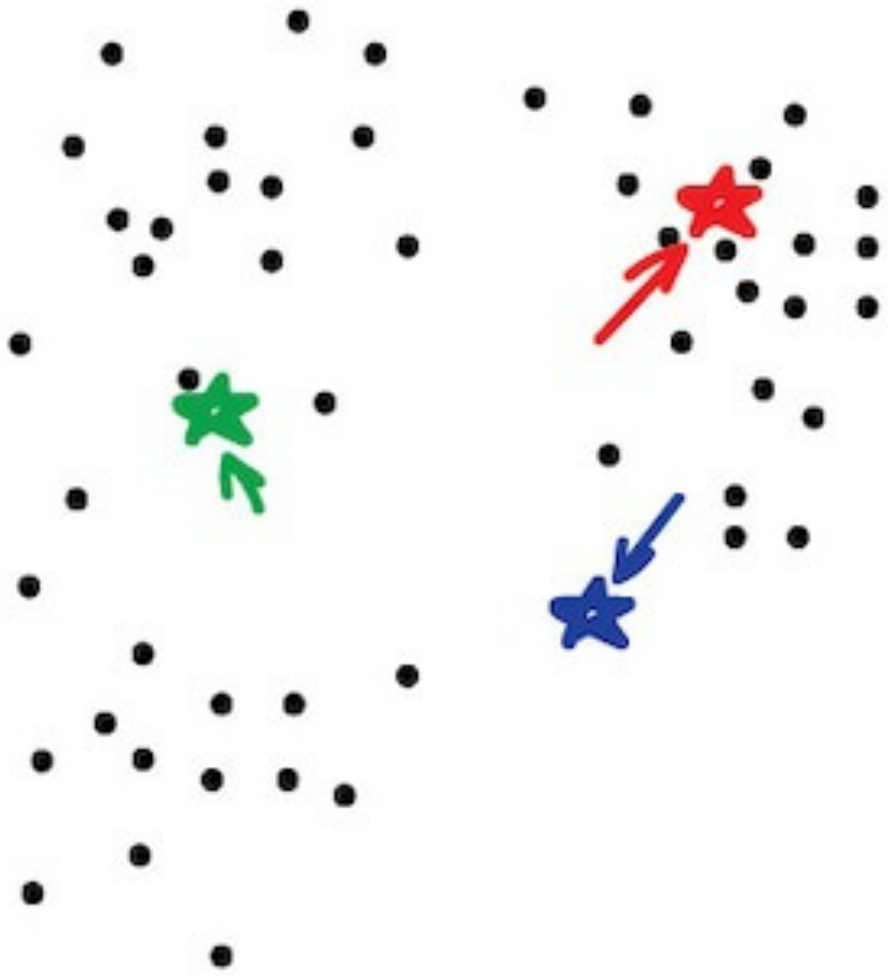
A. Recognition
B. Classification
C. Clustering
D. Multivariate Feature Extraction

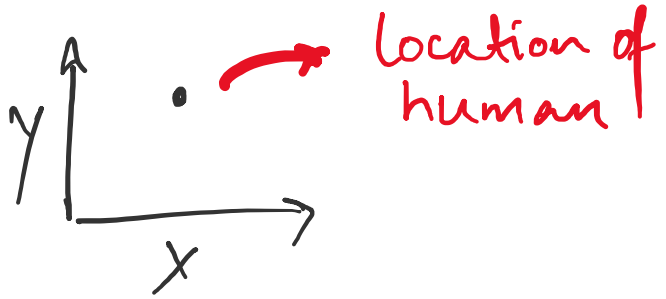


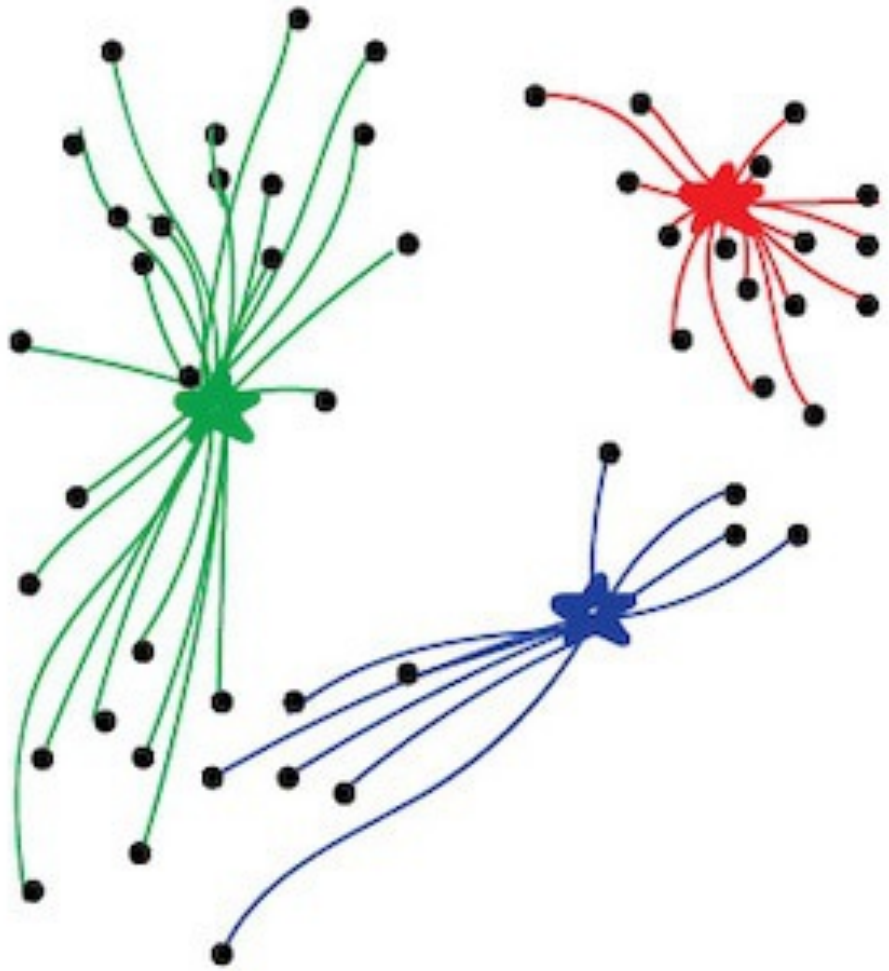

y
x
location of human

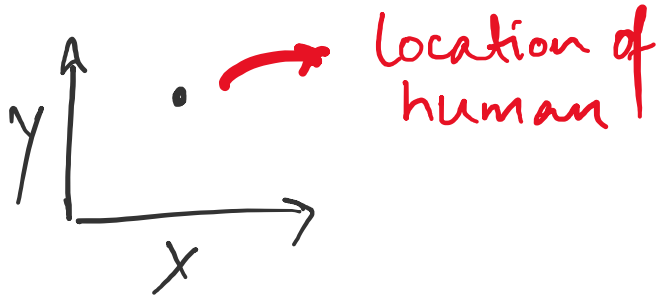


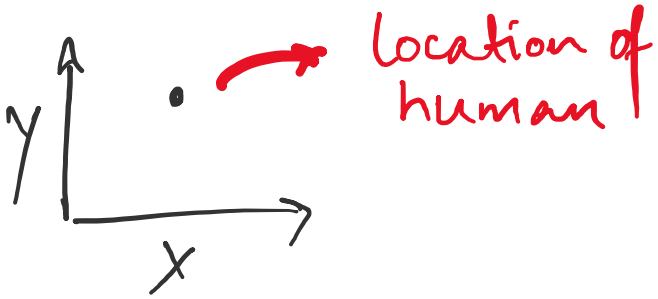

y
x
location of human

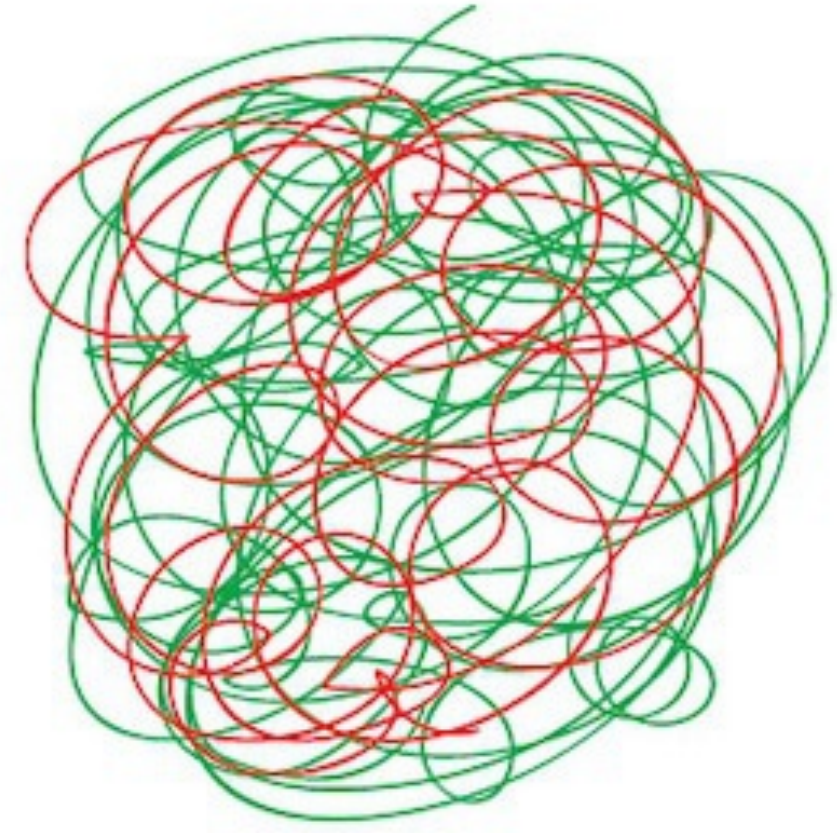


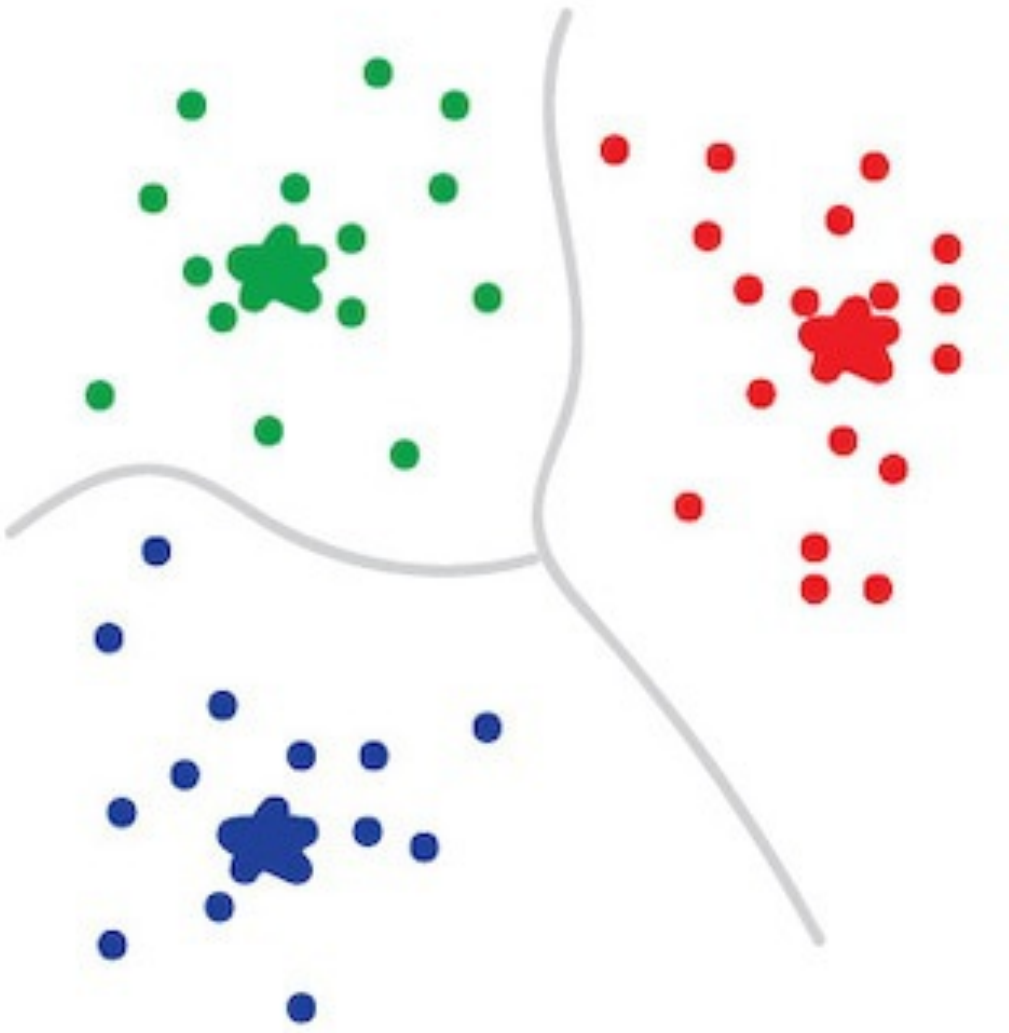

y
x
location of human

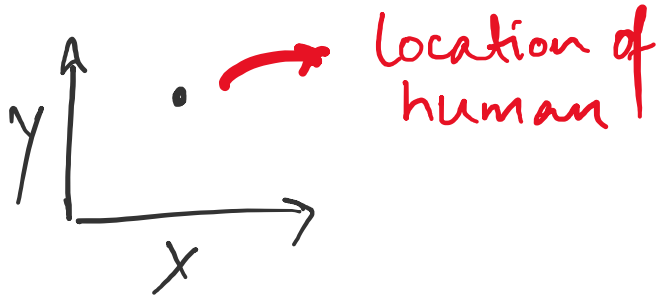



y
x
location of human


y
x
location of human






y
x
location of human

k-means

Randomly initialise k cluster centroids

$\mu_1 \mu_2 \cdots \mu_k \in \mathbb{R}^n$

k-means

Randomly initialise K cluster centroids

$\mu_1 \mu_2 \cdots \mu_K \in \mathbb{R}^n$

Repeat {

for $i = 1$ to m

$c(i) :=$ index (1 to K) of cluster centroid closest to $x^{(i)}$

k-means

Randomly initialise K cluster centroids

$\mu_1 \mu_2 \cdots \mu_K \in \mathbb{R}^n$

Repeat {

cluster assignment

for $i = 1$ to m

$c(i) :=$ index (1 to K) of cluster centroid closest to $x^{(i)}$

k-means

Randomly initialise K cluster centroids

$\mu_1 \mu_2 \cdots \mu_K \in \mathbb{R}^n$

Repeat {

cluster assignment

for $i = 1$ to m

$c(i)$ = index (1 to K) of cluster centroid closest to $x^{(i)}$

for $k = 1$ to K

μ_k = avg (mean) of points assigned to cluster k

}

k-means

Randomly initialise K cluster centroids

$\mu_1 \mu_2 \cdots \mu_K \in \mathbb{R}^n$

Repeat {

cluster assignment

for $i = 1$ to m

$c(i)$ = index (1 to K) of cluster centroid closest to $x^{(i)}$

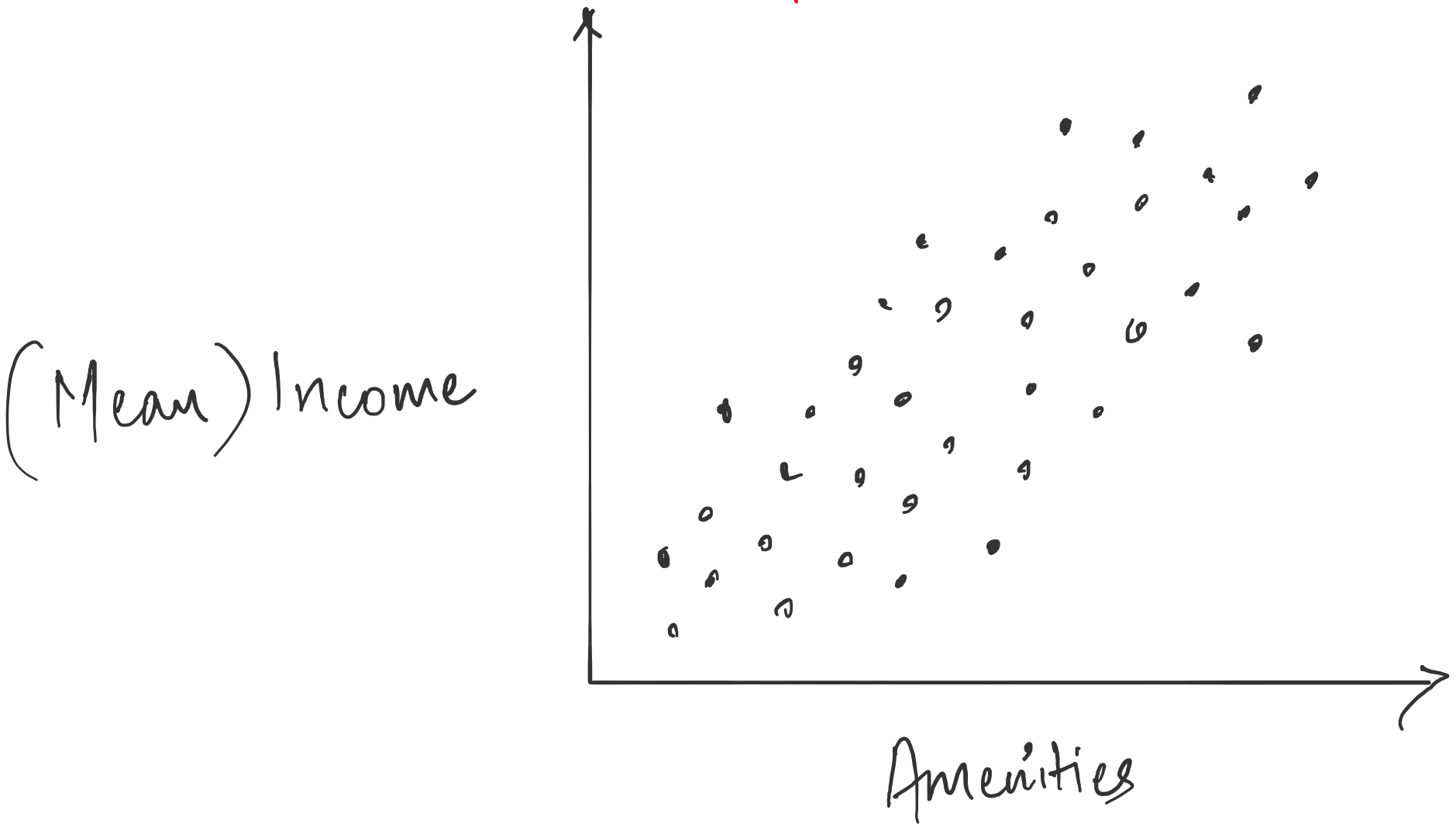
move centroid

for $k = 1$ to K

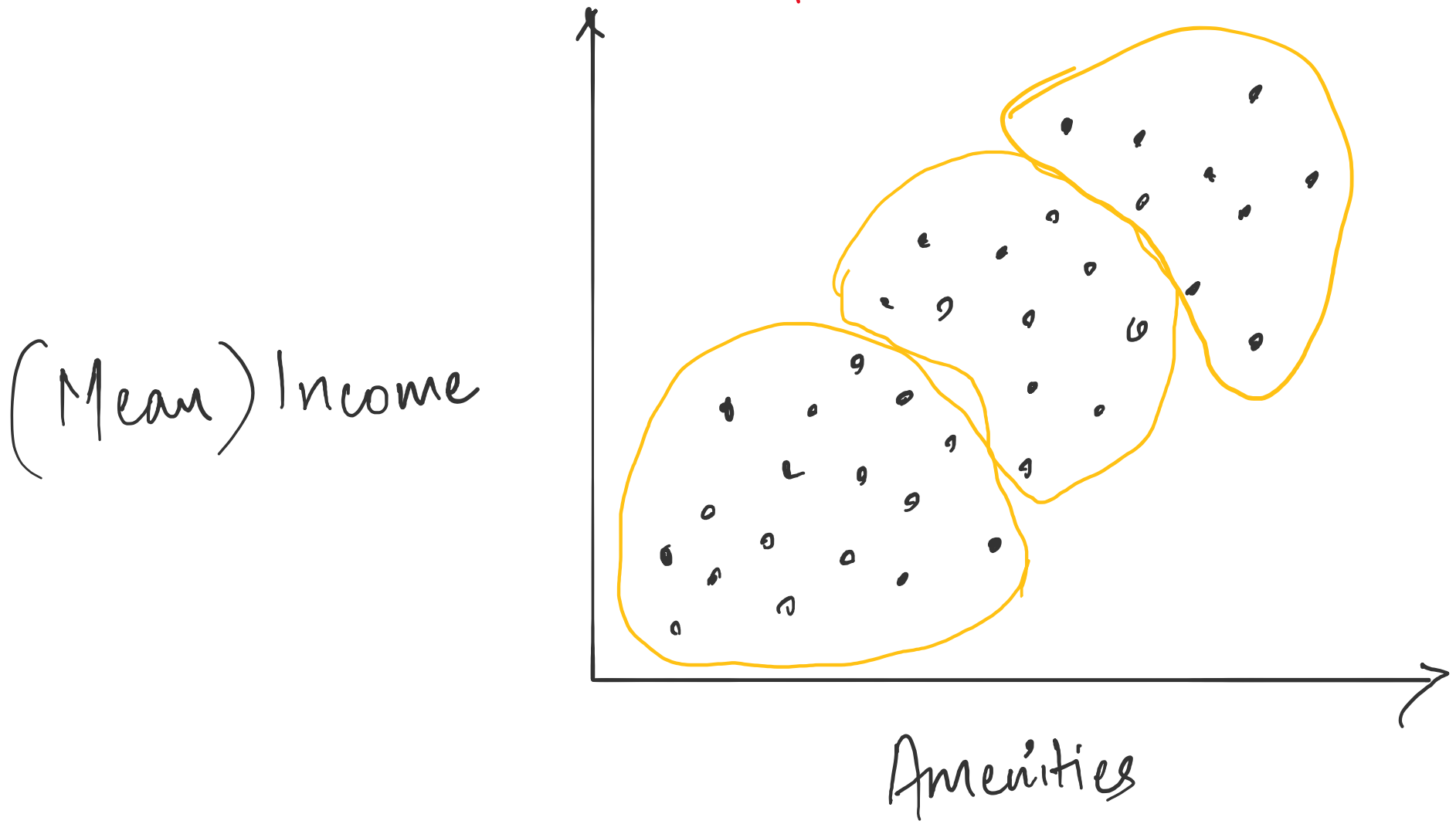
μ_k = avg (mean) of points assigned to cluster k

}

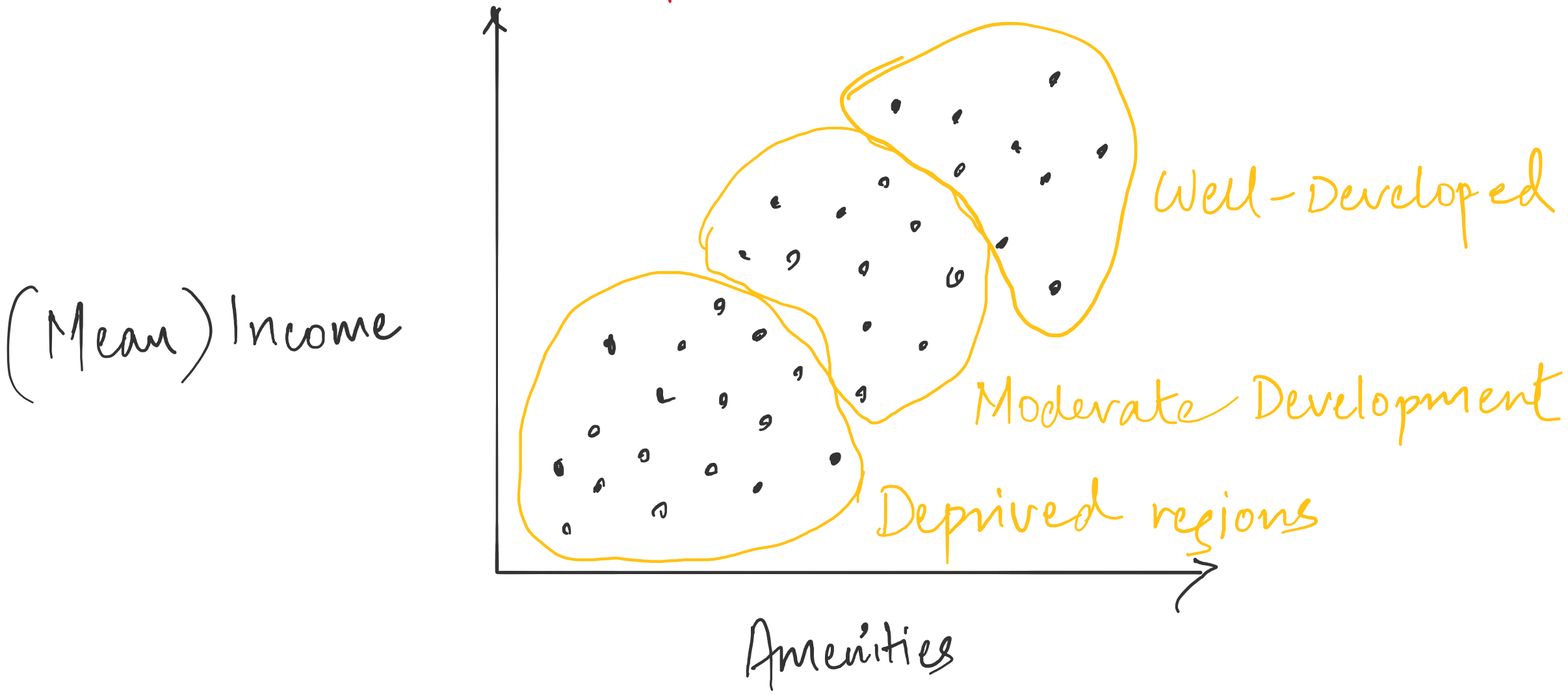
Non-Separated Clusters



Non-Separated Clusters

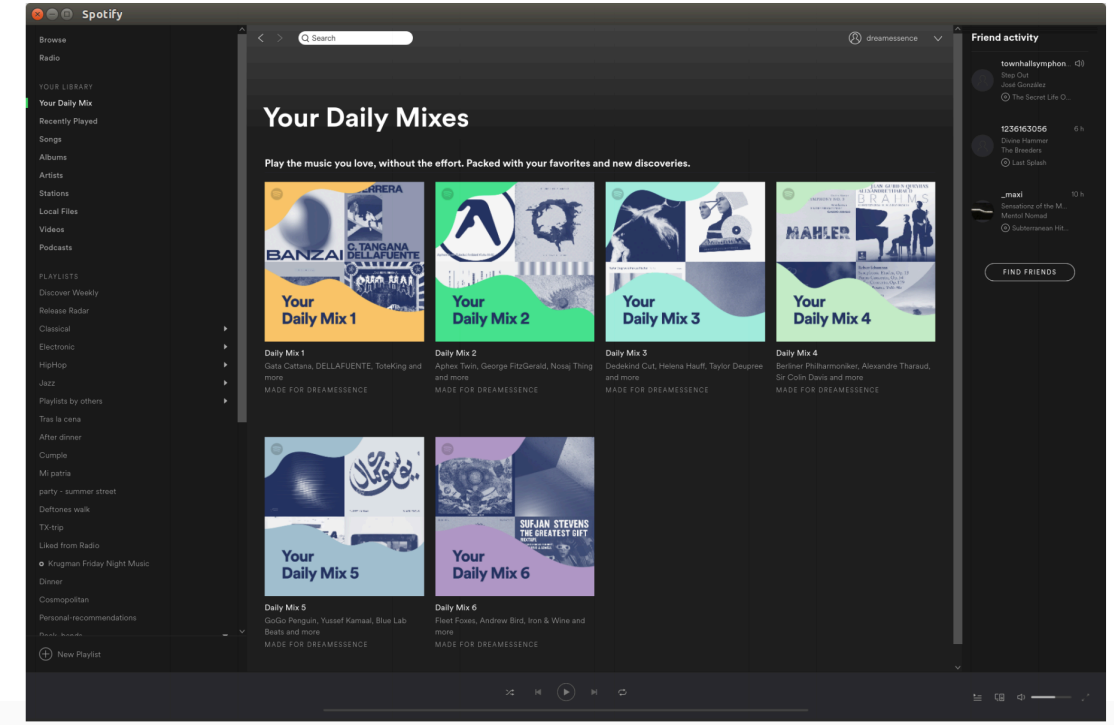


Non-Separated Clusters



Unsupervised Learning

- For market segmentation (types of customers, loyalty) – “Social Dilemma”
- To merge close points on a map
- For image compression
- To analyse and label new data
- To detect abnormal behaviour



Popular algorithms: K-means_clustering, Mean-Shift, DBSCAN

Break

WATER

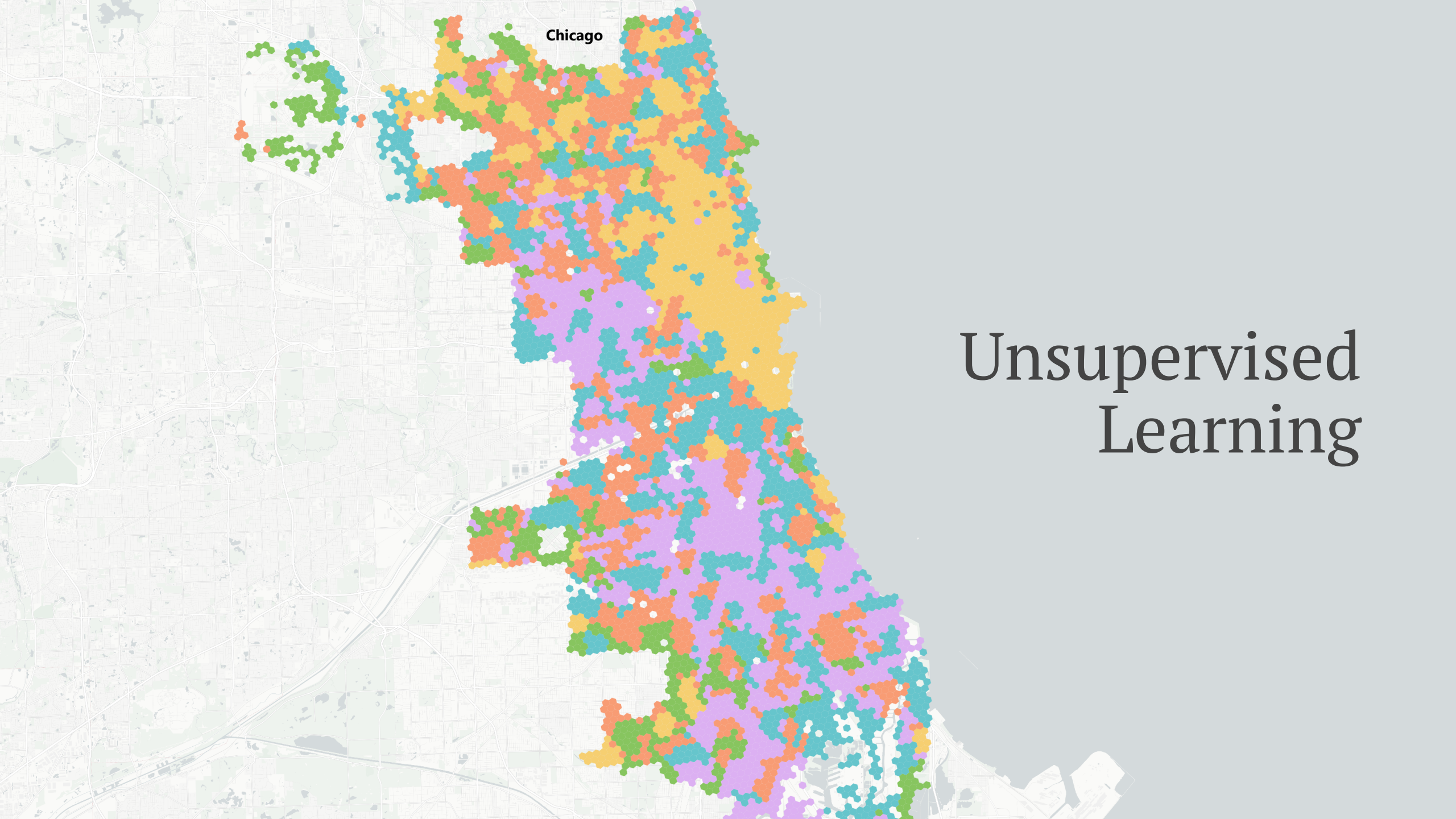
WALK

COFFEE OR TEA

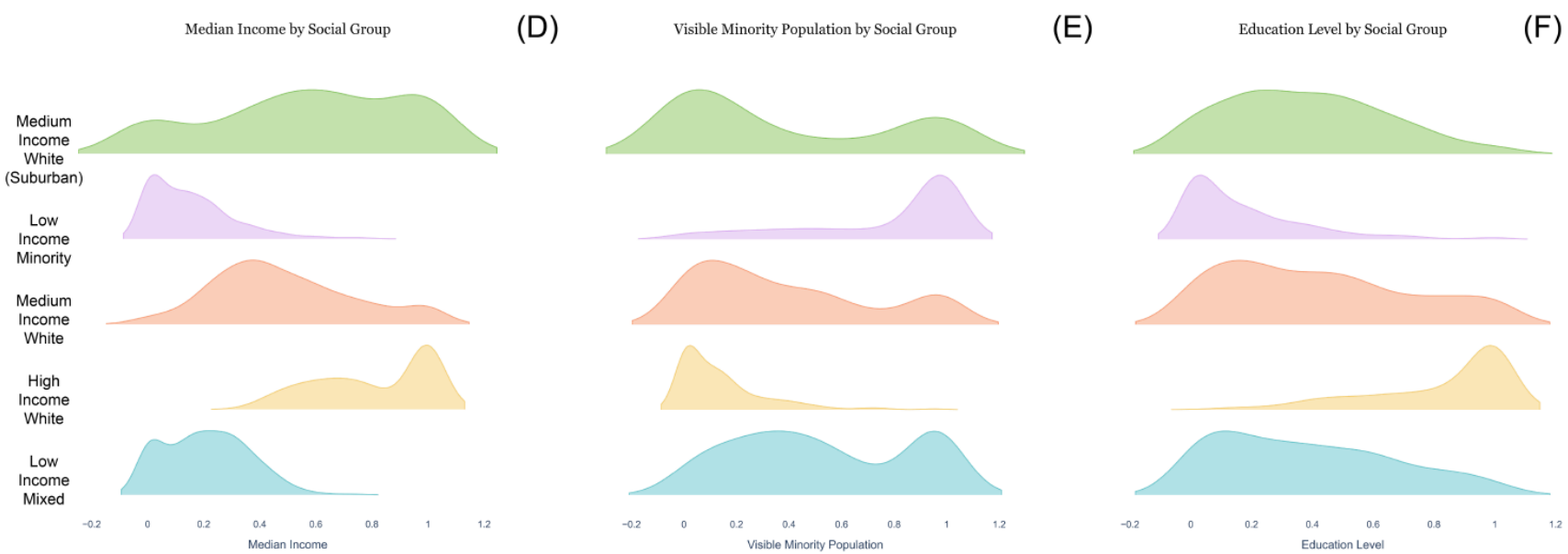
MAKE FRIENDS

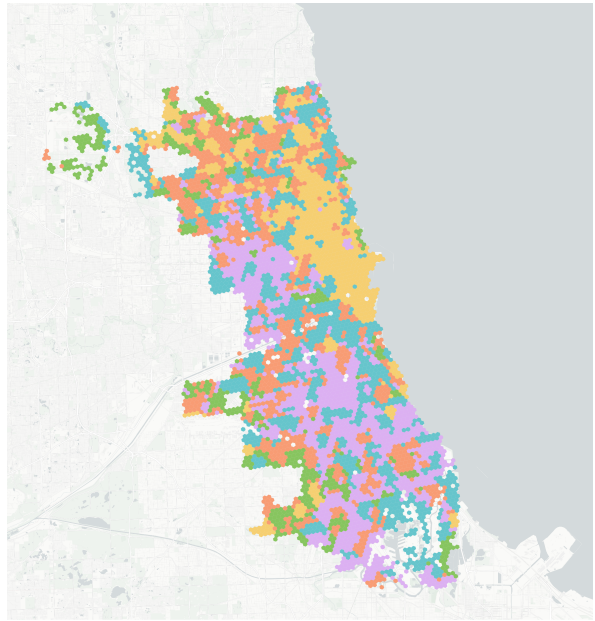
Q: Apple and Google Photos are looking for faces in photos to create albums of your friends. The app doesn't know how many friends you have and how they look, but it's trying to find the common facial features. What kind of algorithm could it be using?

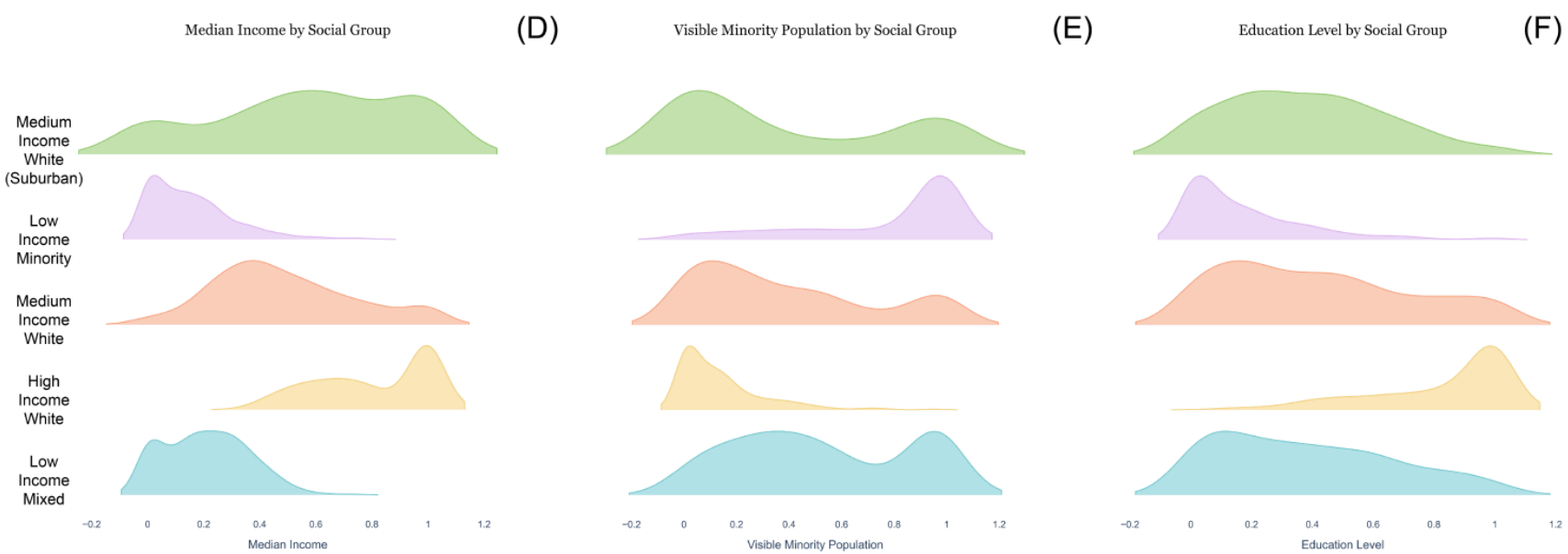
A. K-means
B. DBSCAN
C. Support-Vector Machines
D. Deep-learning

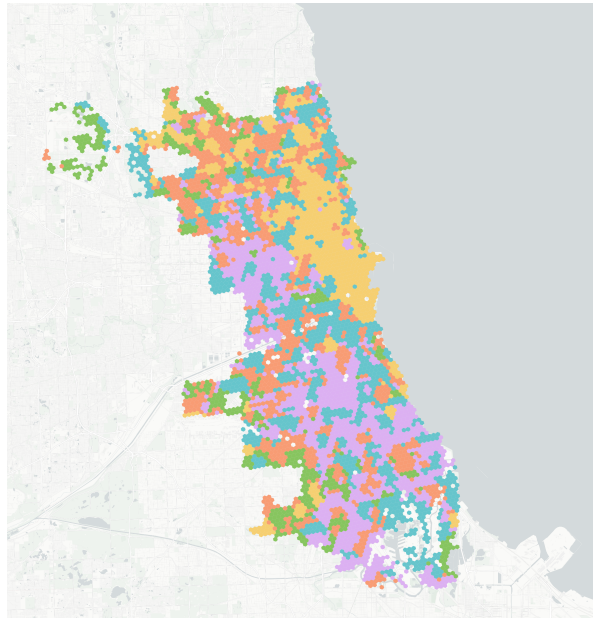


Unsupervised Learning


Median Income by Social Group
(D)
Medium Income White (Suburban)
Low Income Minority
Medium Income White
High Income White
Low Income Mixed
−0.2
0
0.2
0.4
0.6
0.8
1
1.2
Median Income
Visible Minority Population by Social Group
(E)
Visible Minority Population
Education Level by Social Group
(F)
Education Level







Principal Component Analysis – Lecture 12

More Clustering

- Hierarchical clustering
- Agglomerative clustering
- Spectral clustering
- Neural networks (e.g. Self-Organizing Maps)
- DBSCAN
- …

See interesting comparison table

Regionalisation (Duque et al.)

Unsupervised Spatial Machine Learning

Aggregating basic spatial units (**areas**) into larger units (**regions**)

Split a dataset into **groups** of observations that are **similar** within the group and **dissimilar** between groups, based on a series of **attributes.**

…with the additional constraint that observations need to be **spatial neighbours**

Split a dataset into **groups** of observations that are **similar** within the group and **dissimilar** between groups, based on a series of **attributes.**

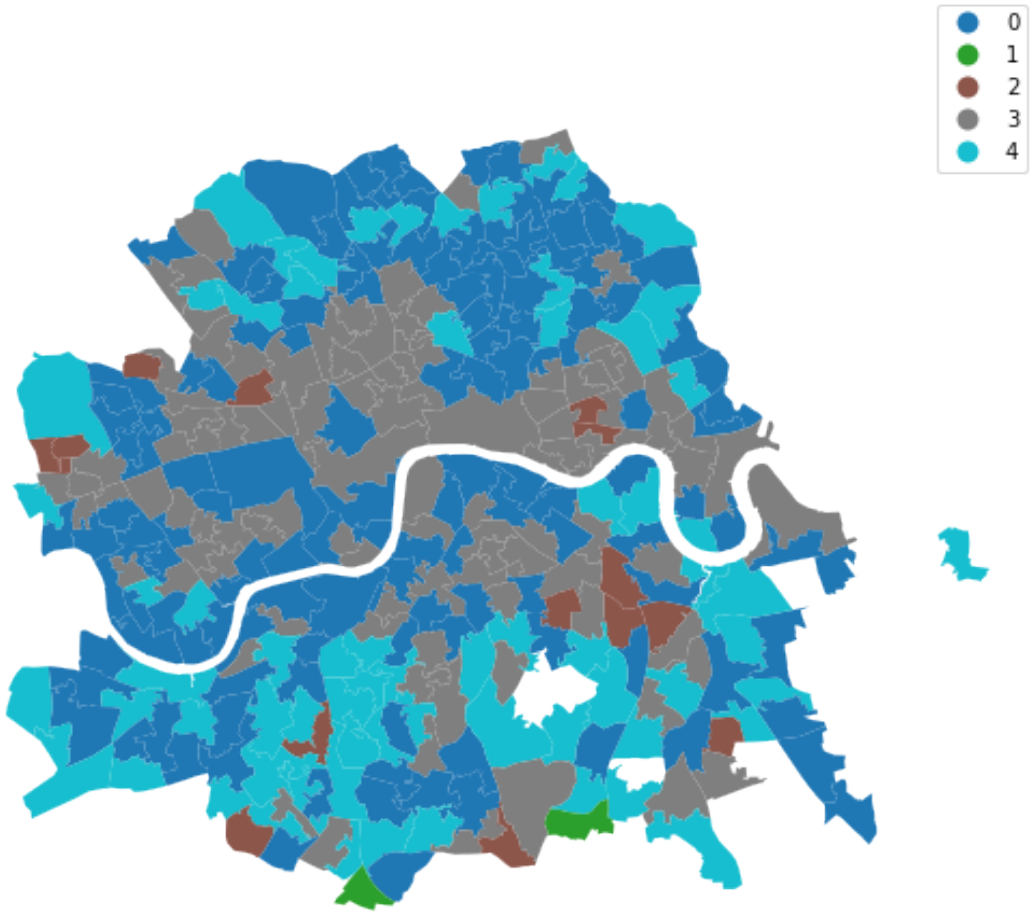
…with the additional constraint that observations need to be **spatial neighbours**

(remember spatial weights?)

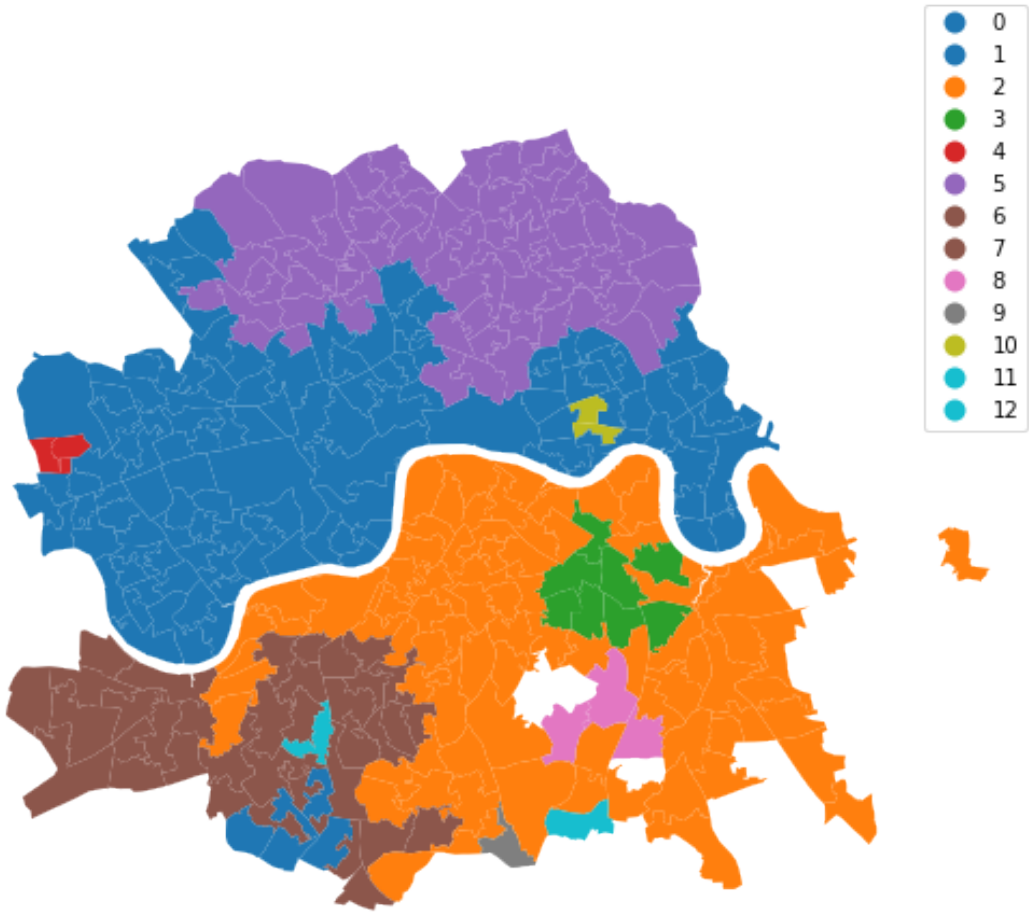
Regionalisation

- All the methods aggregate geographical areas into a predefined number of regions, while optimizing a particular aggregation criterion;
- The areas within a region must be geographically connected (the spatial contiguity constraint);
- The number of regions must be smaller than or equal to the number of areas;
- Each area must be assigned to one and only one region;
- Each region must contain at least one area.

AirBnb Geodemographic classification for Inner London



AirBnb-based boroughs for Inner London



Algorithms (advanced and optional)

- Automated Zoning Procedure (AZP)
- Arisel
- Max-P
- …

See Duque et al. for an excellent, though advanced, overview

Recapitulation

- Some problems are truly **highly dimensional** and univariate representations are not appropriate
- **Clustering** can help reduce complexity by creating **categories** that retain statistical information but are easier to understand
- Two main types of clustering in this context:
  - Geodemographic analysis
  - Regionalisation

Examples in *the wild*

Q: The government wants to know the likelihood of finding regions in a city where communities are deprived of services. Based on census data, which task would you carry out for advising them?

A. Multivariate clustering analysis
B. Regionalisation
C. Polynomial regression
D. None of the above

Q: A popular issue is image compression. When saving the image to PNG you can set the palette, let's say, to 32 colors. It means _____ will find all the "reddish" pixels, calculate the "average red" and set it for all the red pixels. Fewer colors — lower file size — profit! Fill in the blank.

A. I/Me/Human
B. Geodemography
C. One-hot encoding
D. Clustering

Q: When you are looking at an image for compression, you may have problems with colors like Cyan because they don't belong in a palette a machine can read. What kind of task will be useful here?

A. Linear Regression
B. DBSCAN
C. K-Means
D. Neural Nets

For next class..

Finish Lab 07 to practice programming

Submit Homework 07 for peer review on Peer

Submit Assignment 3 – due in **Week 7** on Friday at **2330**

See "To do before class" for every lecture (~ 1 hour of self study)

Read paper for **Discussion** session before every Friday

Post questions on the **Discussion** forum on Brightspace (especially on **Clustering** for this week)