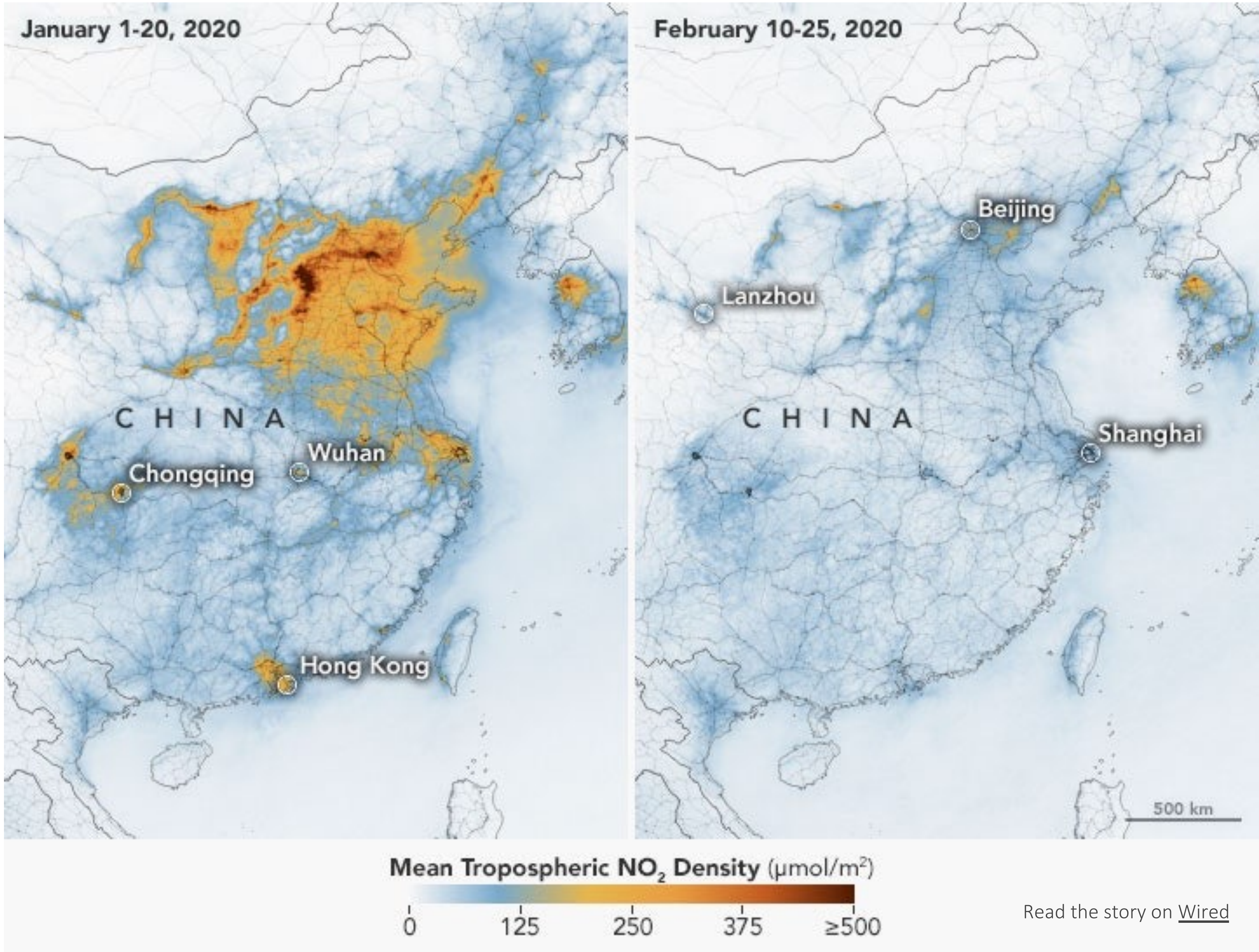
Spatial Data Science

Spatial Density Estimation

(EPA122A)

Lecture 13

Trivik Verma



Read the story on Wired

For Assignment 3

….**include** code from assignment 2

- With modifications
- With improvements
- With whatever is necessary* for us to grade A3

* We will not open A2 to reconcile facts

Final Projects

A : Mobility, Built Environment & Sustainability

B : Identifying the Health Vulnerability in a City

C : Modelling COVID-19 in India

Final Projects - Milestones

1. Group Creation and Project Selection
2. Scope of Work and Preliminary EDA
   - Project statement
   - Preliminary EDA

3. EDA and Revised Project Statement
4. Project Report
   - Template
   - Rubric

Have fun!

Last Time

- Big Data and High Dimensionality
- A Framework For Dimensionality Reduction
- Principal Components Analysis (PCA)

TU Delft

Q: When does high-dimensionality occur?

A. P >> N
B. P << N
C. P = N
D. When data is **N**ormally distributed

Q: What are some of the consequences of high-dimensional data?

A. Issues with regression
B. Variables related to each other
C. Models cannot predict new data
D. All the above
E. None of the above

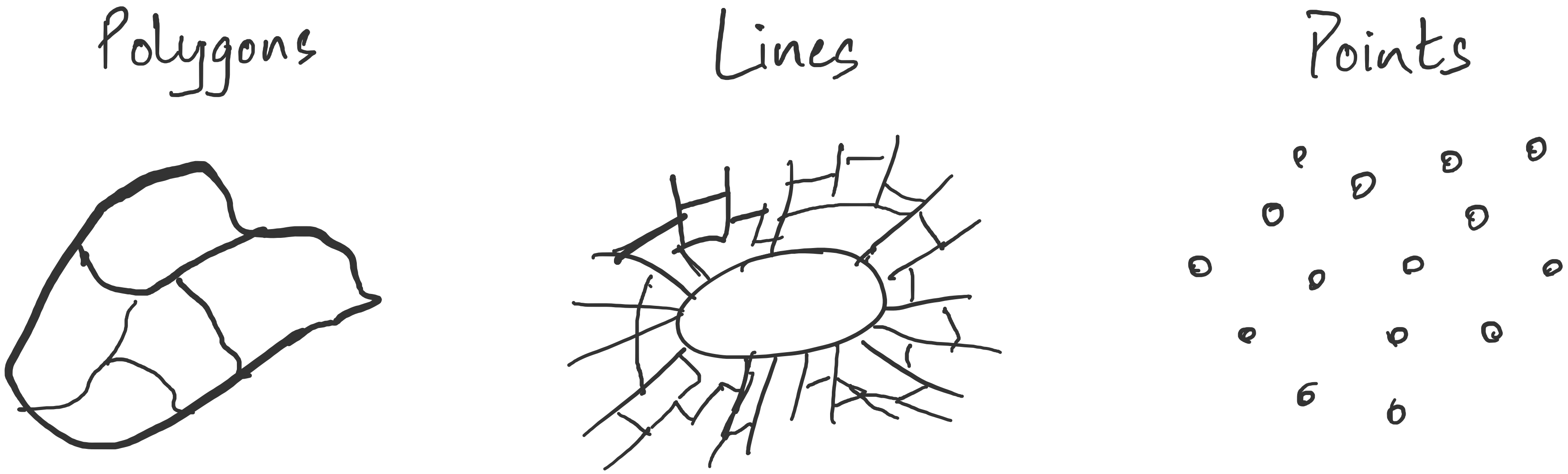
Today

- The *point* of points
- Point patterns
- Visualization of point patterns
- Identifying clusters of points

The *point* of points

Points like polygons

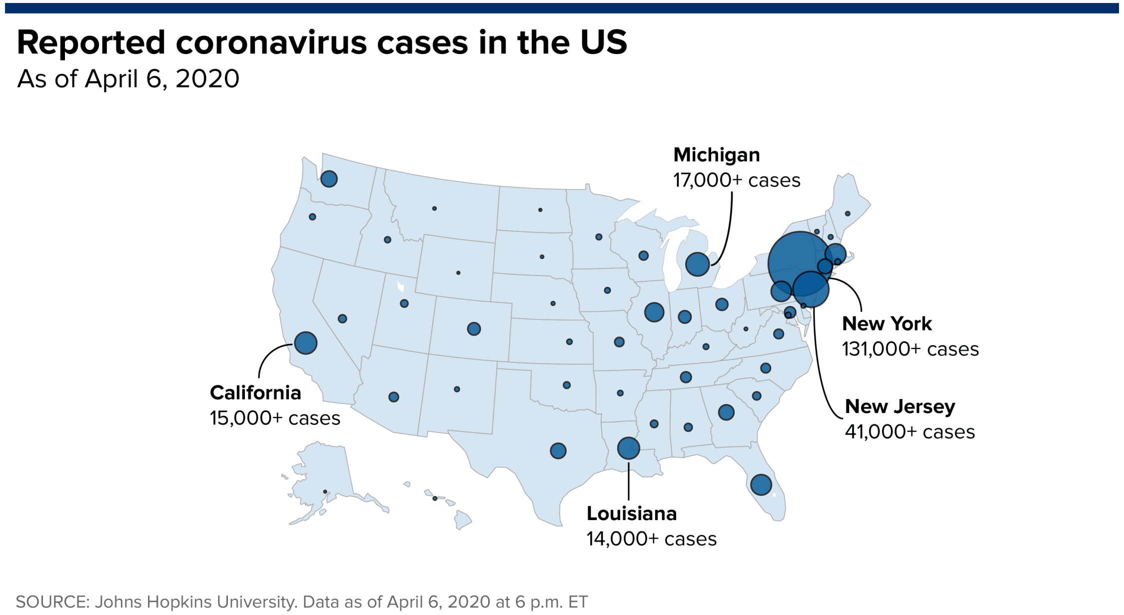
- Points can represent **“fixed” entities**
- In this case, points are **qualitatively** like **polygons/lines**
- The **goal** here is, taking location fixed, to model other aspects of the data

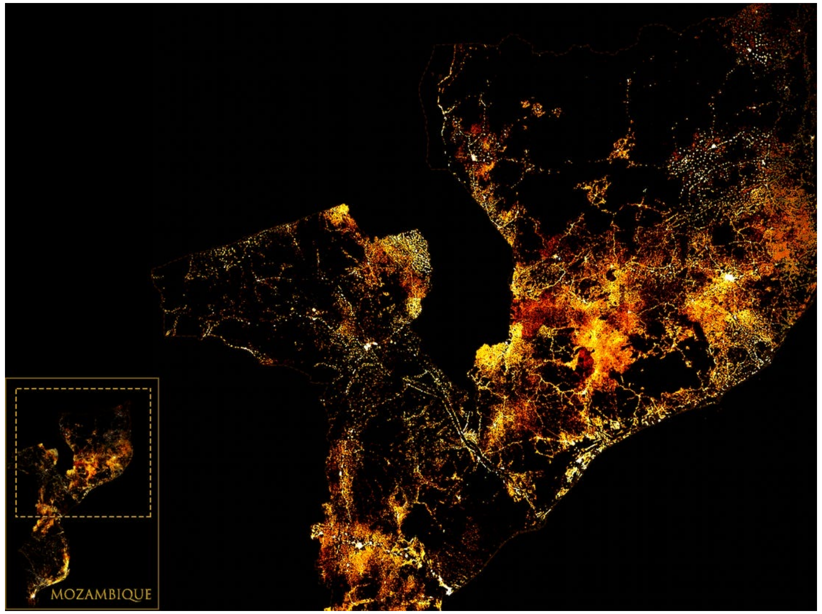


Points like polygons

Examples

- Cities (in most cases)
- Buildings or People (processes to estimate – social media)
- Polygons represented as their centroid …





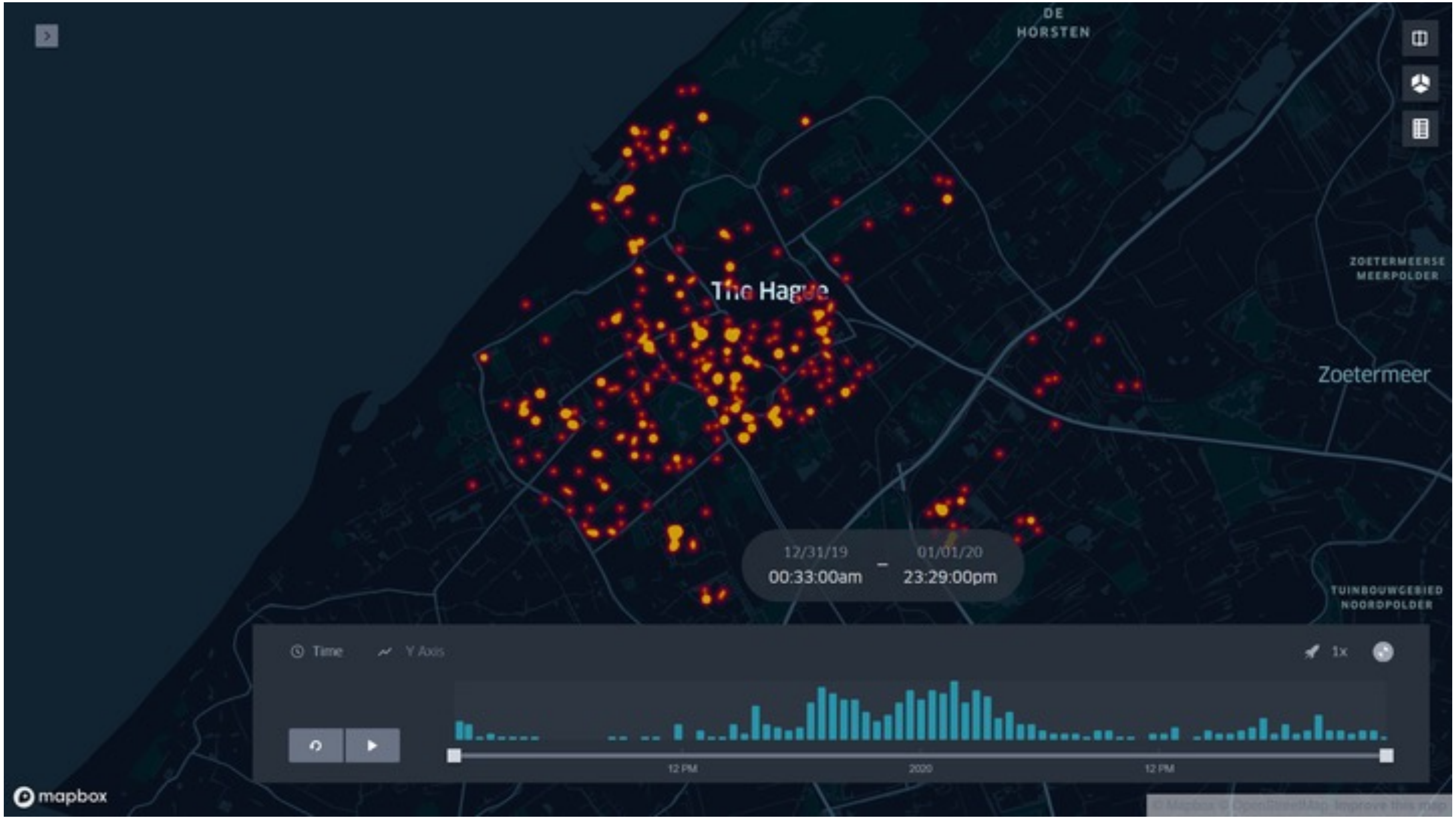
Source: Facebook

When points are not polygons

Point data are not only a different geometry than polygons or lines…

… Points can also represent a fundamentally different way to approach spatial analysis

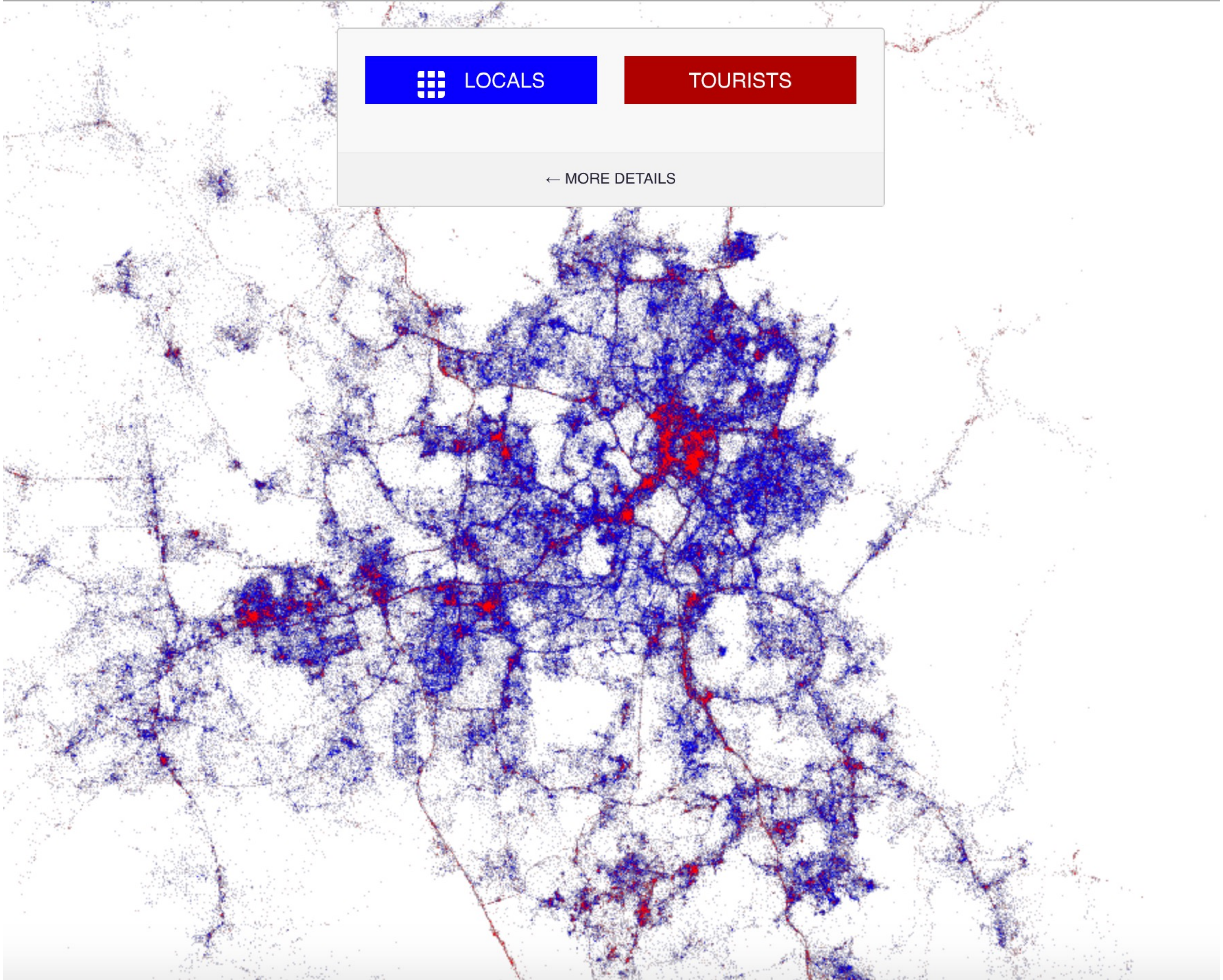
Points unlike polygons



Source: Analysis of Emergency Data in the Netherlands

NYC Street Trees by Species
New York City's urban forest provides numerous environmental and social benefits, and street trees compose roughly one quarter of that canopy. This map shows the distribution and biodiversity of the city's street trees based on the last tree census.
Read more.
Weehawken
Hoboken
New York
Bayonne

[source]



[source]

Point Patterns

Point Patterns

Distribution of **points over** a portion of **space**

Assumption is a point can happen anywhere on that space, but only happens in specific locations

- **Unmarked**: locations only
- **Marked**: values attached to each point

Point Pattern **Analysis**

Describe, characterize, and explain point patterns, focusing on their generating process

- *Visual exploration*
- Clustering properties and *clusters*
- *Statistical modelling* of the underlying processes

Visualisation of PPs

Visualisation of PPs

Two routes

1. *Aggregate* $\leftrightarrow$ Histogram
2. *Smooth* $\leftrightarrow$ KDE

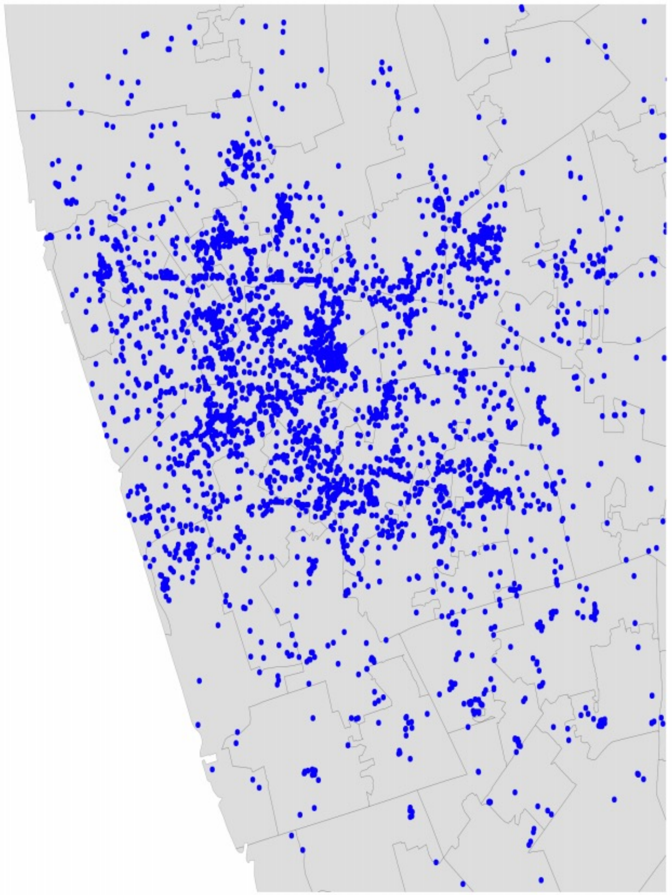
Aggregation

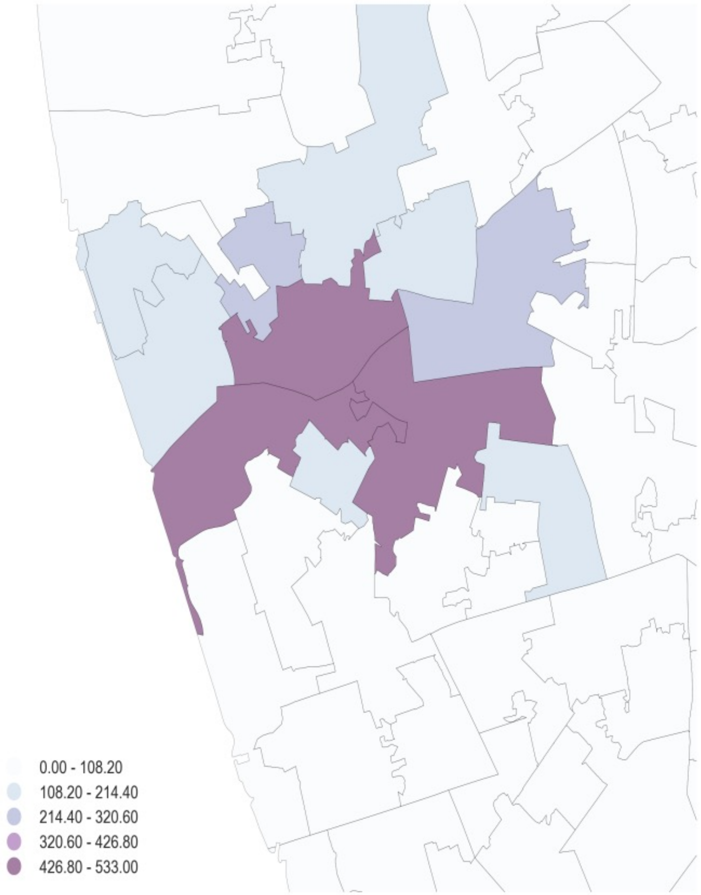
Points meet Polygons

Use polygon boundaries and count points per area

[Insert your skills for choropleth mapping here!!!]

But the polygons need to “make sense” (their delineation needs to relate to the point generating process).





Images taken from: Arribas-Bel, D. (2019). A course on geographic data science. *Journal of Open Source Education*, *2*(16), 42.

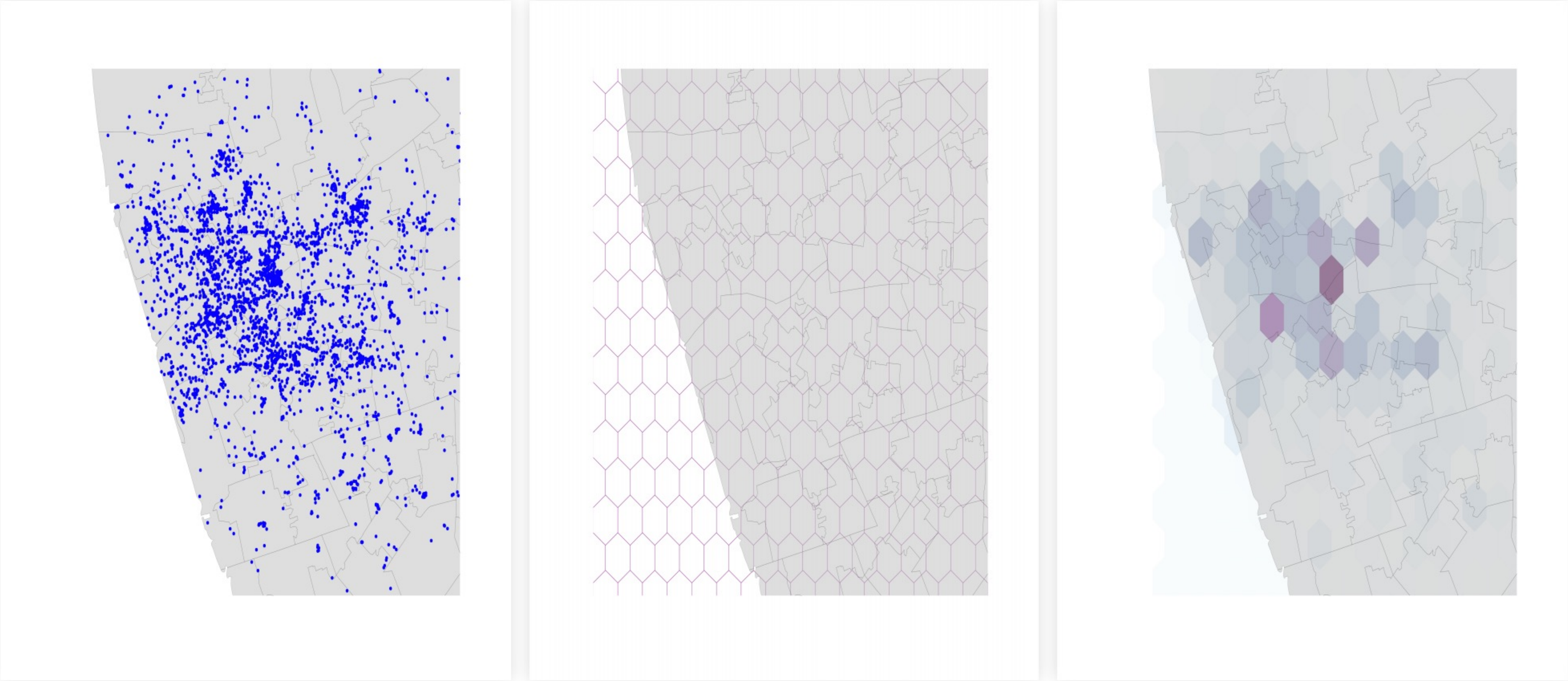
Hex-Binning

If no polygon boundary seems like a good candidate for aggregation…

…draw a **hexagonal** (or **squared**) tessellation!!!

Hexagons…

- Are regular (over census tracts)
- Exhaust the space (unlike circles)
- Have many sides (minimise boundary problems – think queen and rook!)



Images taken from: Arribas-Bel, D. (2019). A course on geographic data science. *Journal of Open Source Education*, *2*(16), 42.

But…

(Arbitrary) aggregation may induce MAUP (see Lecture 6)

+

Points usually represent events that affect only part of the population and hence are best considered as rates (see Lecture 6)

Q: Which processes will you consider hexagon binning for?

… Points showing accessibility to amenities

A. Yes
B. No

Q: Which processes will you consider hexagon binning for?

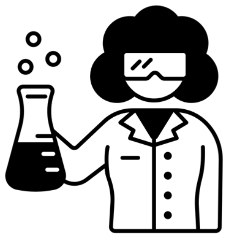
… Points showing cell-phone use in a location

A. Yes
B. No

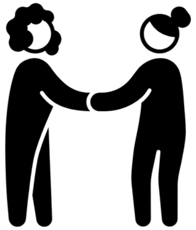
Break

CHILL

WALK



COFFEE OR TEA



MAKE FRIENDS

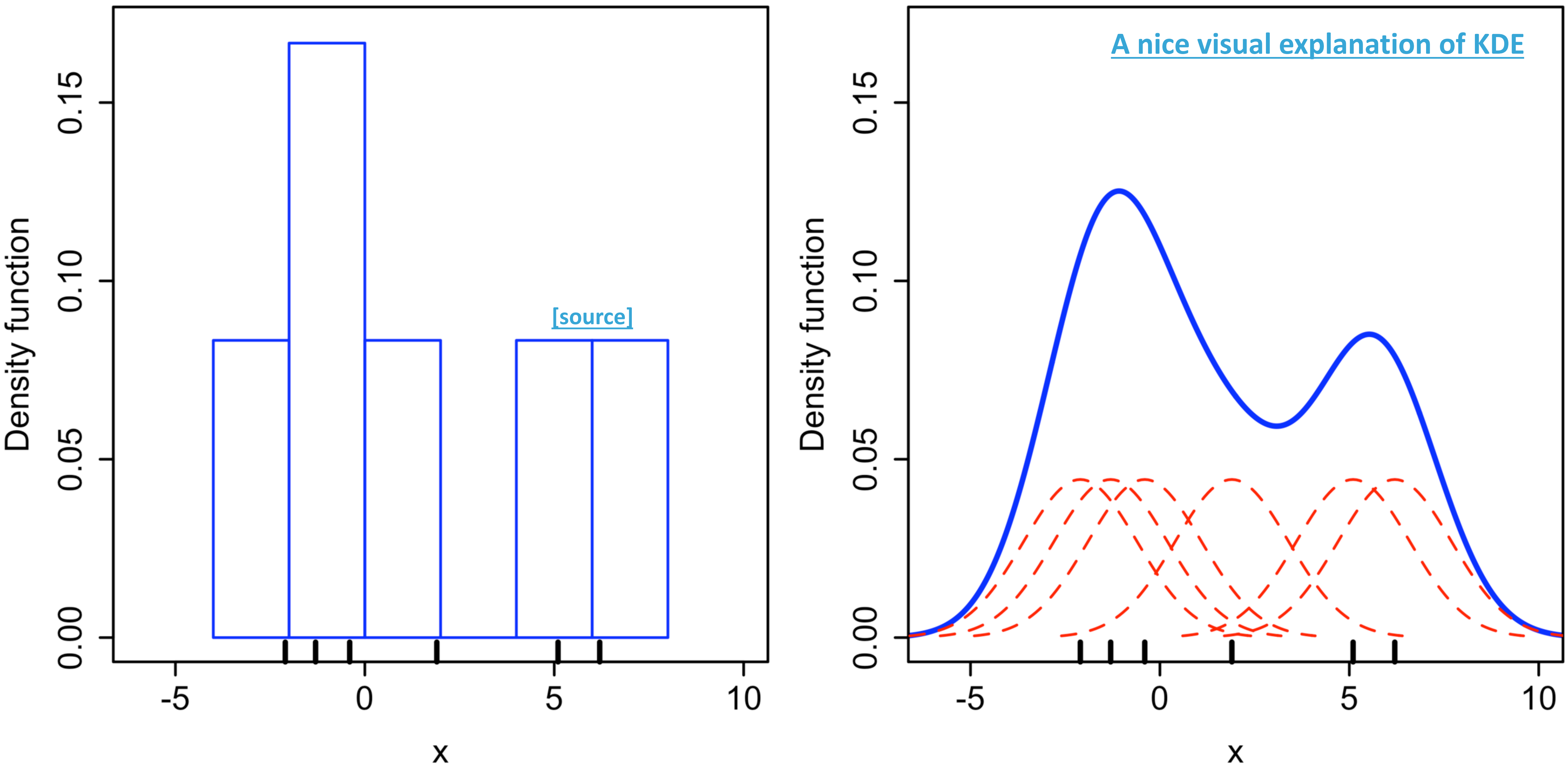
Icons by Iconathon on The Noun Project

Kernel Density Estimation

Kernel Density Estimation

Estimate the (***continuous***) observed ***distribution*** of a variable

- Probability of finding an observation at a given point
- “Continuous histogram ”
- Solves (much of) the MAUP problem, but not the underlying population issue

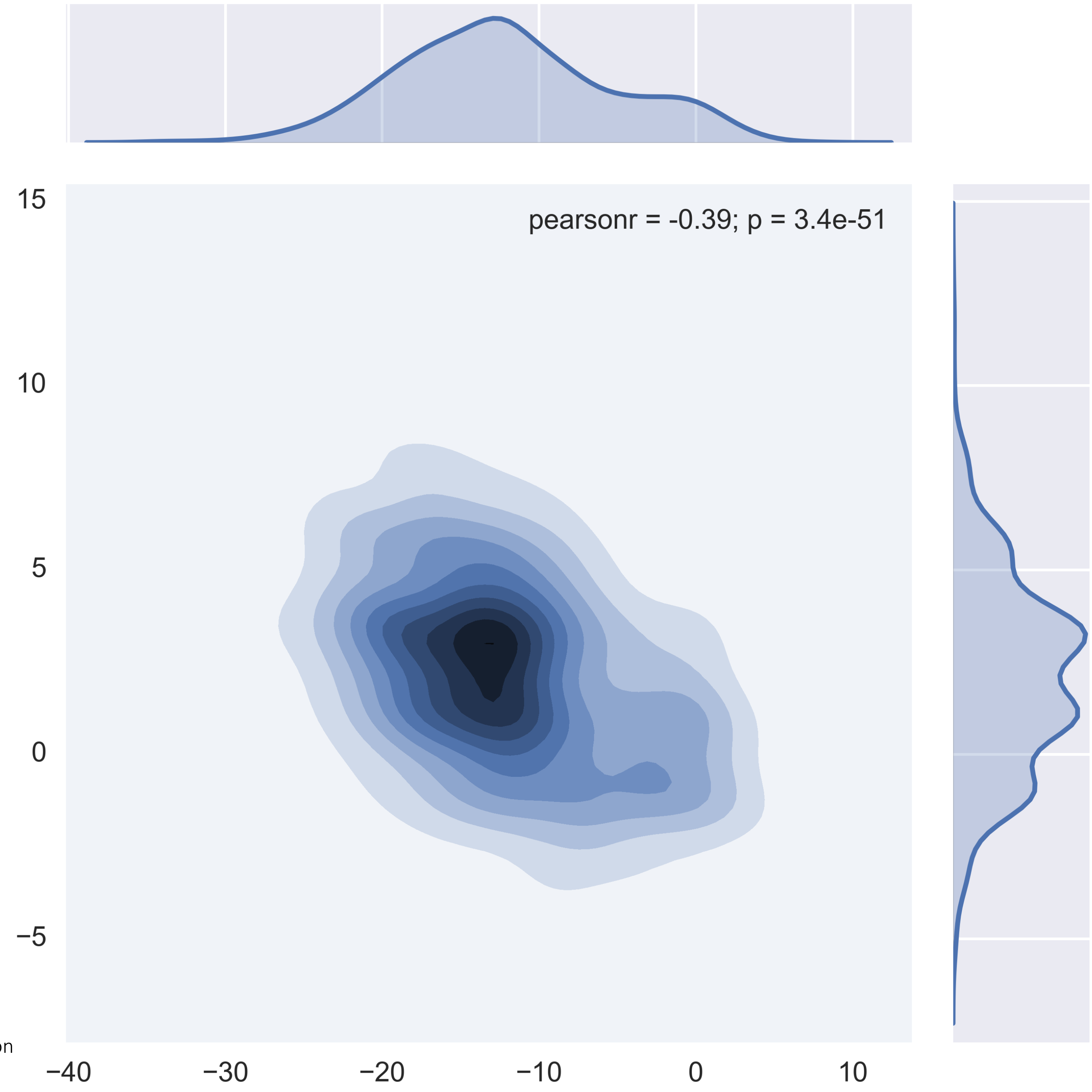


[source]
A nice visual explanation of KDE
Density function
0.00
0.05
0.10
0.15
-5
0
5
10
x

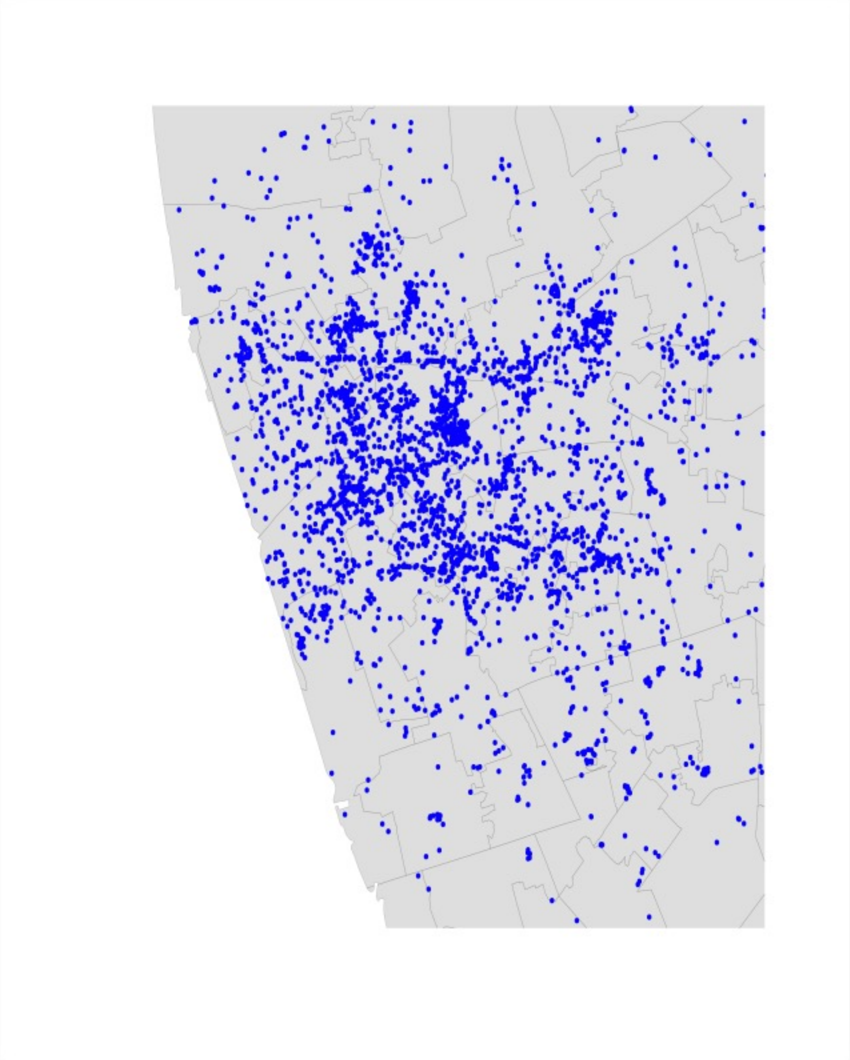
Bivariate (spatial) KDE

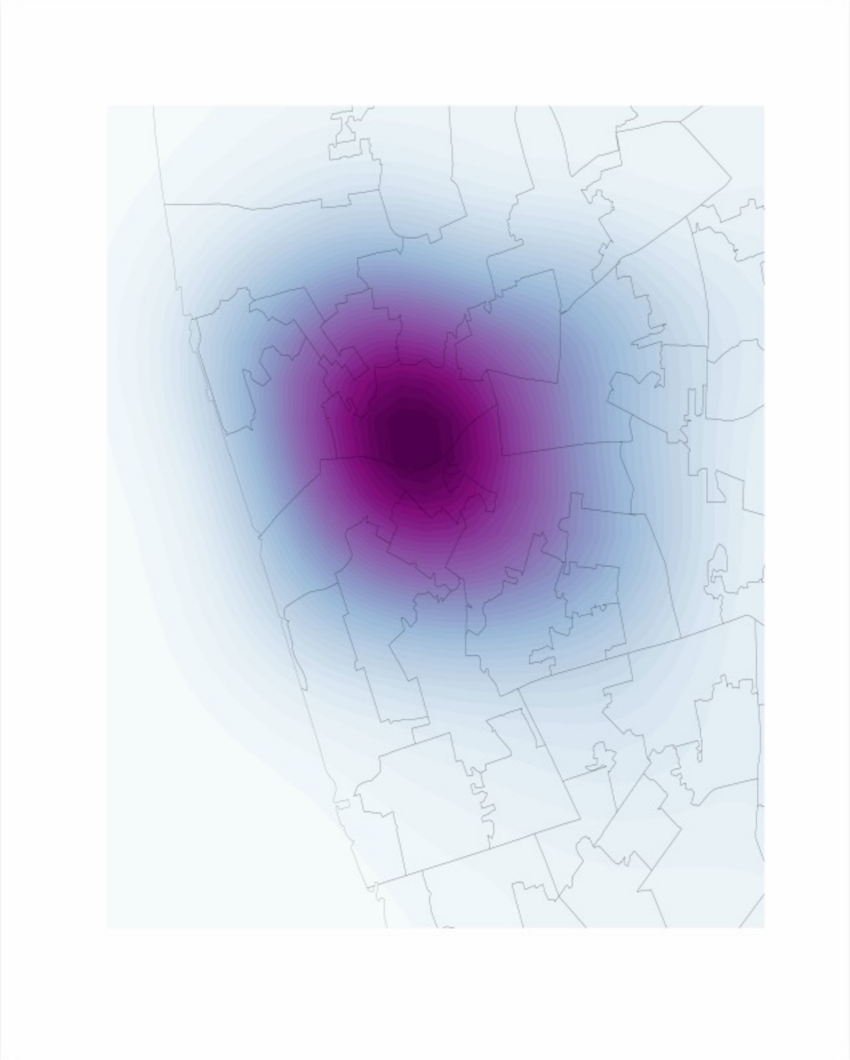
Probability of finding observations at a given point in space

- **Bivariate** version: distribution of **pairs of values**
- **In space**: values are coordinates (XY), locations
- Continuous "version" of a choropleth



Images taken from: Arribas-Bel, D. (2019). A course on geographic data science. *Journal of Open Source Education*, *2*(16), 42.





Images taken from: Arribas-Bel, D. (2019). A course on geographic data science. *Journal of Open Source Education*, *2*(16), 42.

Finding clusters of PPs

Concentrations/agglomerations of points over space, significantly more so than in the rest of the space considered

Huge literature spanning spatial analysis, statistics and computer science.

Today, we ’ll look at DBSCAN

Q: When number of clusters (K) are known, which algorithm do we use for clustering?

A. DBSCAN
B. K-Means
C. K-Nearest Neighbours
D. Support Vector Machines

Density

Based

Spatial

Clustering of

Applications with

Noise

When K is not known

↳ no. of clusters

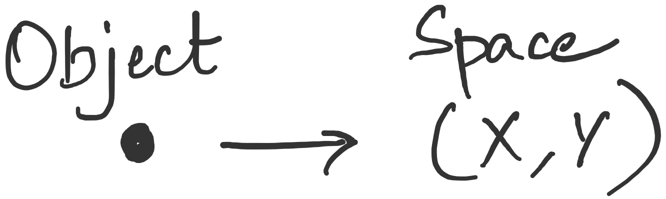
DBSCAN

↳ Set of points in space

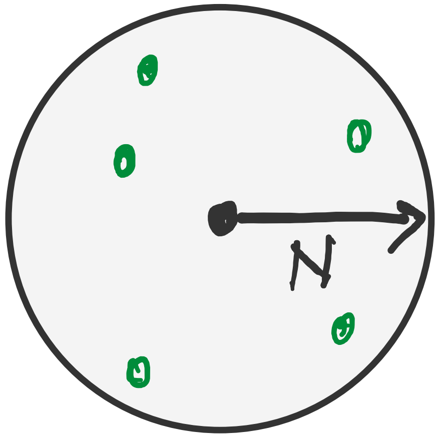
↳ Neighbourhood **N**

↳ Density (minpts)

1. Set of points



2. Neighbourhood



Steps

1. Label each pt. as → Core, Border, Noise

minpts = 4

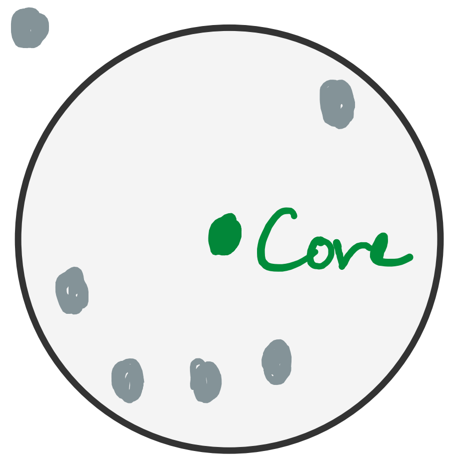
Steps

1. Label each pt. as
   - Core
   - Border
   - Noise

minpts = 4

$|N(x)| = 6$

$4 < 6$



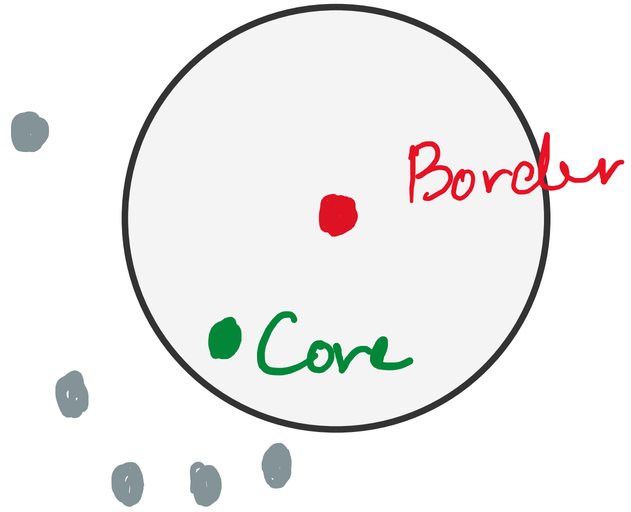
Steps

1. Label each pt. as
   - Core
   - Border
   - Noise

$\underline{\text{minpts} = 4}$

$|N(x)| = 2$

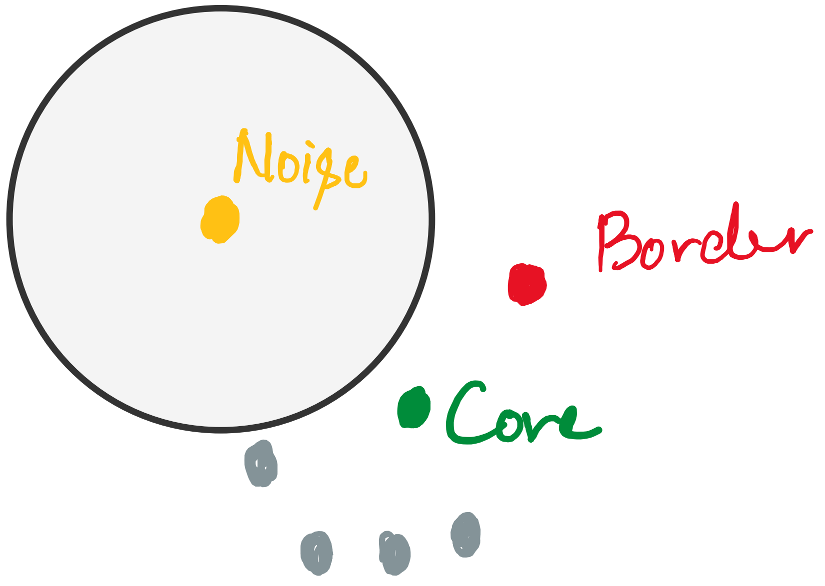
$4 > 2$



Steps

1. Label each pt. as
   - Core
   - Border
   - Noise

minpts = 4



$$|N(x)| = 1$$

$$4 > 1$$

Steps

1. Label each pt. as Core, Border, Noise

   minpts = 4

2. Takes every core and performs a Depth First Search to find its neighbours

DBSCAN

Pros:

- Discover **any** number of clusters
- Clusters of varying **size** and **shape**
- Detect and **ignore outliers** in the data
- Not necessarily spatial
- Very fast to run so → scales relatively well → applicable to large datasets

Cons:

- Sensitive to **Neighbourhood** parameter
    - too small – sparse is noise
    - too large – dense merged together
- Not based on any probabilistic model (no inference)
- Hard to learn about the underlying process

For next class..

Finish Labs to practice programming

Complete Homework for more practice

Check Assignment contents and due date

See “To do before class” for next lecture (~ 1 hour of self-study)