

Introduction to Spatial statistics

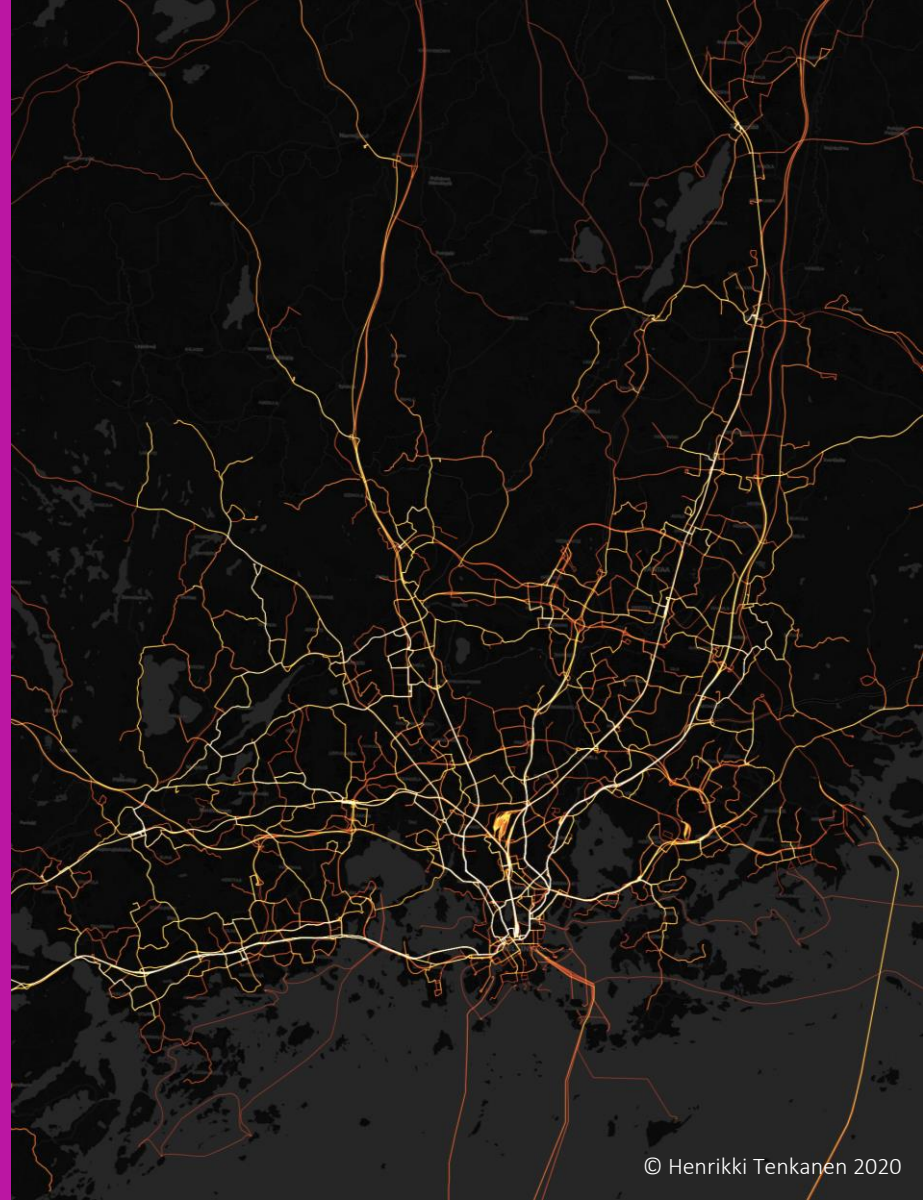
Introduction to Geoinformatics

Henrikki Tenkanen

10.2.2023



Aalto University
School of Engineering



Contents and learning goals

After this lesson you should understand:

- What is the first law of geography? Why geography matters?
- What is spatial statistics and why is it used/useful?
- What is meant by spatial effects and how can we measure/identify them?
 - Spatial dependence vs spatial heterogeneity
 - Autocorrelation / Clustering / hotspot-cold spot
 - Distance decay
- How digitalization and new data sources have changed the way we can understand the world and model different phenomena?

**So what's the deal with
geography? Should I care..?**





Without geography
you're nowhere.

Jimmy Buffett

quote fancy

The tune of the week: [Jimmy Buffet - Margaritaville](#) (video)



Aalto University
School of Engineering

Not convinced?


Think about following two questions based on your own life:

- Think about the person with whom you are most often in contact with?
- How far from each other are you? (in physical distance)

Go to vote! → <https://presemo.aalto.fi/sds/>

Not convinced?

It is very likely, that:

- The persons that you are most often in contact with are in relatively close proximity with you
 - The likelihood of being in contact / interacting with each other decreases as the distance increases (called as “distance decay”):
 - Living in the same apartment / building
 - Living in the same neighborhood / city
 - Living in the same county
 - Living in the same country
- 
- Increasing distance
- You might think that the web has diminished the importance of proximity. It's not. Face-to-face social interaction is integral part of being human. Human activities and interactions are the thriving force of our societies. → Proximity matters.

The First law of geography

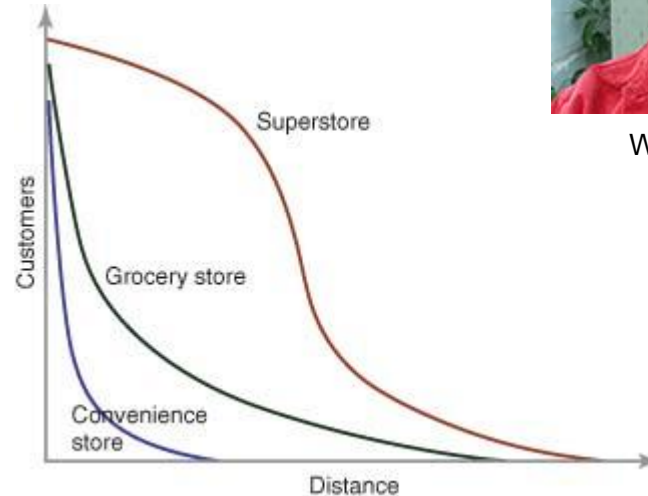
The notion that “proximity matters” is not a new idea:

- Waldo Tobler is the father of the *First Law of Geography* (1970):

The first law of geography

“everything is related to everything else,
but near things are more related than
distant things”

– Waldo Tobler (1970) [3]



Waldo and his hat.

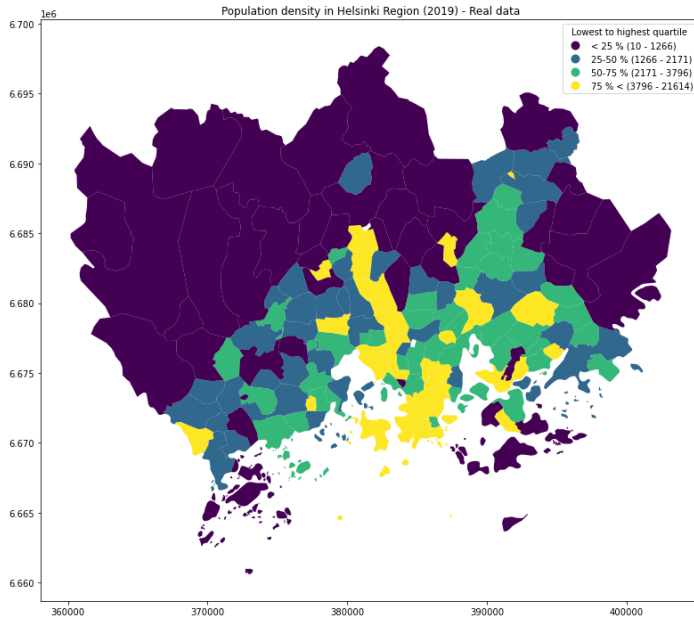
The idea of distance decay. The likelihood of visiting a specific type of service (e.g. food store) decreases as the distance increases. (Image source: [Pun-Cheng, 2016](#))

Statistics and geography?

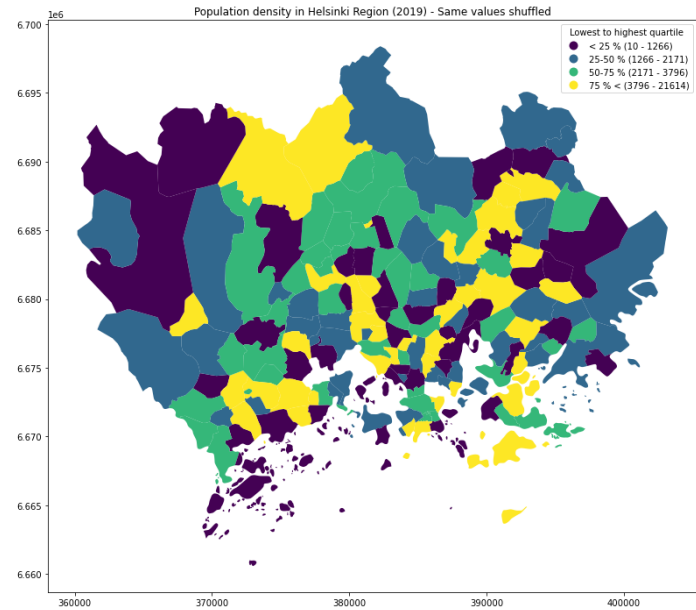
- Statistical methods are widely used to describe, model and predict how different societal and environmental phenomena operates based on observed data.
- Relatively simple statistical models are often underneath even with the most advanced Machine Learning (ML) / Artificial Intelligence (AI) techniques that you might have become familiar with (e.g. automatic image recognition).
- “Traditional” statistical models or more “modern” ML methods cannot escape geography and proximity (it’s in the nature of most data) → GeoAI

Statistics can be misleading ..

Geographical view



Population density (quartiles) - **Real**
Highest population densities clustered

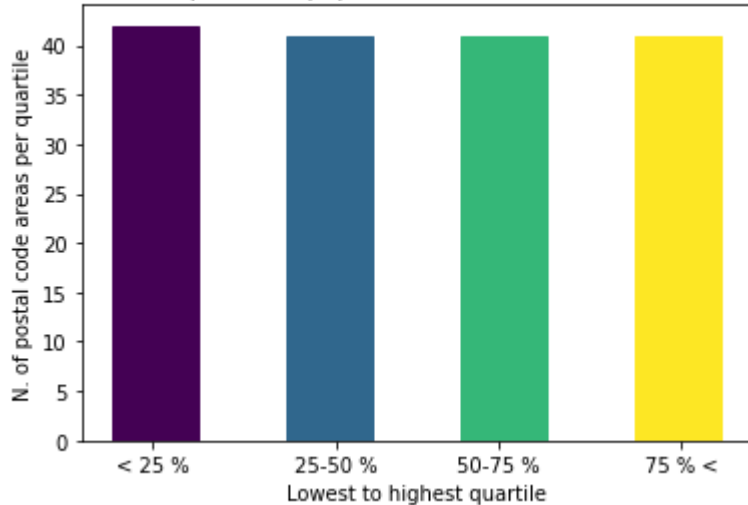


Population density (quartiles) - **Fake**
Highest population densities are randomly distributed in space using the same source data.

Statistics can be misleading ..

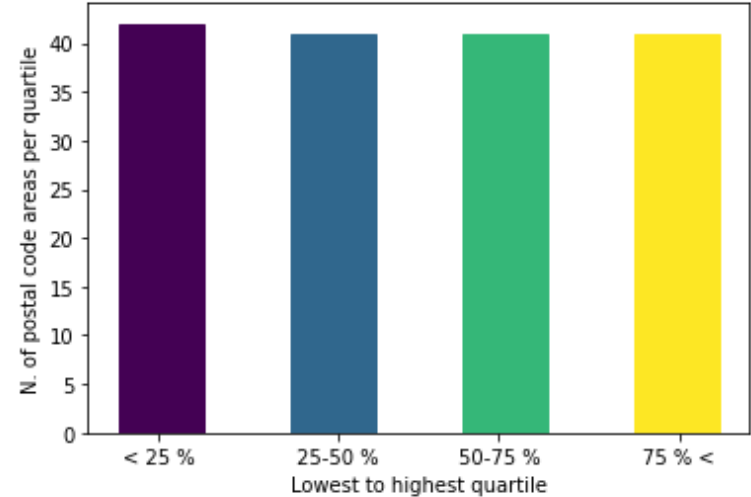
Statistical view

Distribution of postcode population densities in Helsinki - Real data



Population density (quartiles) - Real

Distribution of postcode population densities in Helsinki - Randomly shuffled



Population density (quartiles) - Randomly shuffled

- **The statistical view does not show any difference!** → Always check how the data is distributed in space to understand whether the data makes any sense
- In real world, things tend to cluster due to Tobler's First law of geography

But why is spatial similarity a problem?

Many of the most widely used statistical models (e.g. linear regression) that are used to study the underlying reasons for different phenomena (why things happen) assume that the **data observations are independent** from each other. Due to Tobler's law and the nature of spatial (and temporal) data, this is not the case. We need to take into account the special character of spatial data....with....



Spatial statistics

Spatial statistics

Key functionalities

With spatio-statistical approaches and methods, we can:

- Identify if there are any **spatial effects** involved with our data (which we should consider one way or another):
 - Spatial dependence
 - Spatial heterogeneity
- **Take advantage of proximity** by borrowing information from neighboring values when doing statistical modelling
- Understand not only *where things happen*, but also ***why things happen where they happen***
- **Predict values** to places where we do not have information

Spatial effects #1: Spatial dependence

- Issues related to Tobler's First Law of Geography

Spatial autocorrelation

- Clustering of similar values in geographical space (attribute similarity in space) → Differs from “pure” geographical clustering (only location)
- The presence of spatial autocorrelation can be analytically tested using global or local methods:
 - Moran's I (global measure)
 - Local Moran's I (local measure)
- In statistical modelling, understanding whether your data has spatial autocorrelation is typically one of the first things to do when working with spatial data
- We can take advantage of locational similarity e.g. when doing spatial interpolation (i.e. predicting a value to given location that does not have a value)



Clustering of observations (heatmap) as *fashion design* 😊

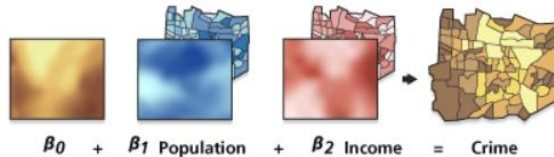
Spatial effects #2: Heterogeneity

- Relates to structural change, structural breaks, varying coefficients, such as different interpretations of space
 - Urban vs rural
 - North vs South
 - Core vs periphery
- **Intrinsic characteristics unevenly distributed over space:**
 - Spatial heterogeneity arises when you cannot safely assume the studied phenomena operates under the same set of rules across the study region.
 - E.g. specific characteristics of a location can influence strongly the variable that we try to model/predict, such as a house or apartment having access/view to the sea makes their price behave differently compared to neighbors without one (Rey et al., 2021).
 - Frequently introduced simultaneously with the concept of spatial dependence (in practice, the two can be difficult to tease apart from each other).
- A dataset in which all subsets have the same statistical properties is considered as **homogeneous** (Haining & Li, 2020). You might e.g. **test for regional homogeneity** to find out if a given phenomena, such as are house prices similar in different neighborhoods? (Anselin, 2017)
- You aim to select / sub-set your data in a way that you have homogeneous study areas, or take into account the heterogeneity in your model.

Spatial Non-stationarity

- Variations in the relationship between an outcome variable and a set of predictor variables across space
- A condition in which a simple “global” model cannot explain the relationships between some set of variables
- The nature of the model must alter over space to reflect the structure within data

→ Geographically Weighted Regression (GWR) can be used to explore these relationships



GWR is a local regression model. Coefficients are allowed to vary.

Image source: [ESRI](#) (2023)

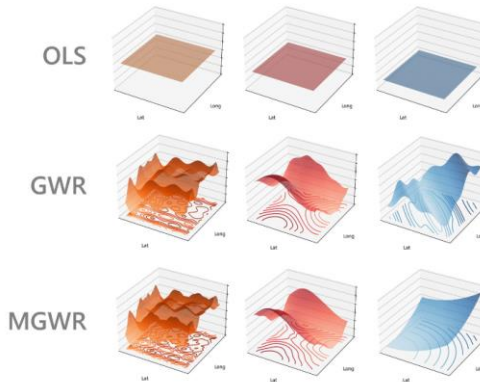


Image source: [ESRI](#) (2023)

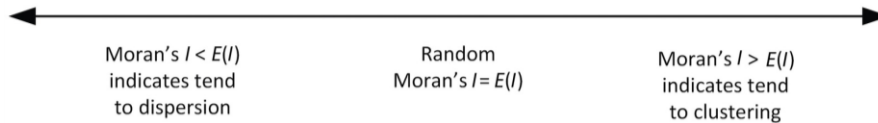
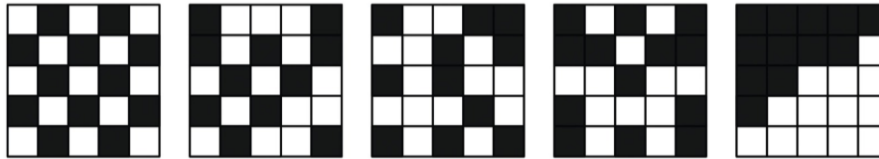
**But how do we really measure /
test for locational similarity, i.e.
spatial autocorrelation?**



Spatial autocorrelation

Moran's I is the most used technique to test whether your data has spatial autocorrelation:

- Can be positive or negative
 - the more similar values cluster, the higher the spatial autocorrelation (Moran's I closer to 1)
 - negative autocorrelation reminds checkerboard pattern (Moran's I closer to -1)
- Measures locational similarity – Tests against **spatial randomness** (the null hypothesis)



Moran's I is defined as

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

where N is the number of spatial units indexed by i and j ; x is the variable of interest; \bar{x} is the mean of x ; w_{ij} is a matrix of spatial weights with zeroes on the diagonal (i.e., $w_{ii} = 0$); and W is the sum of all w_{ij} .

If you have spatial autocorrelation (~50 % of cases):
Then you need a spatial model when doing statistical analysis!

Spatial autocorrelation

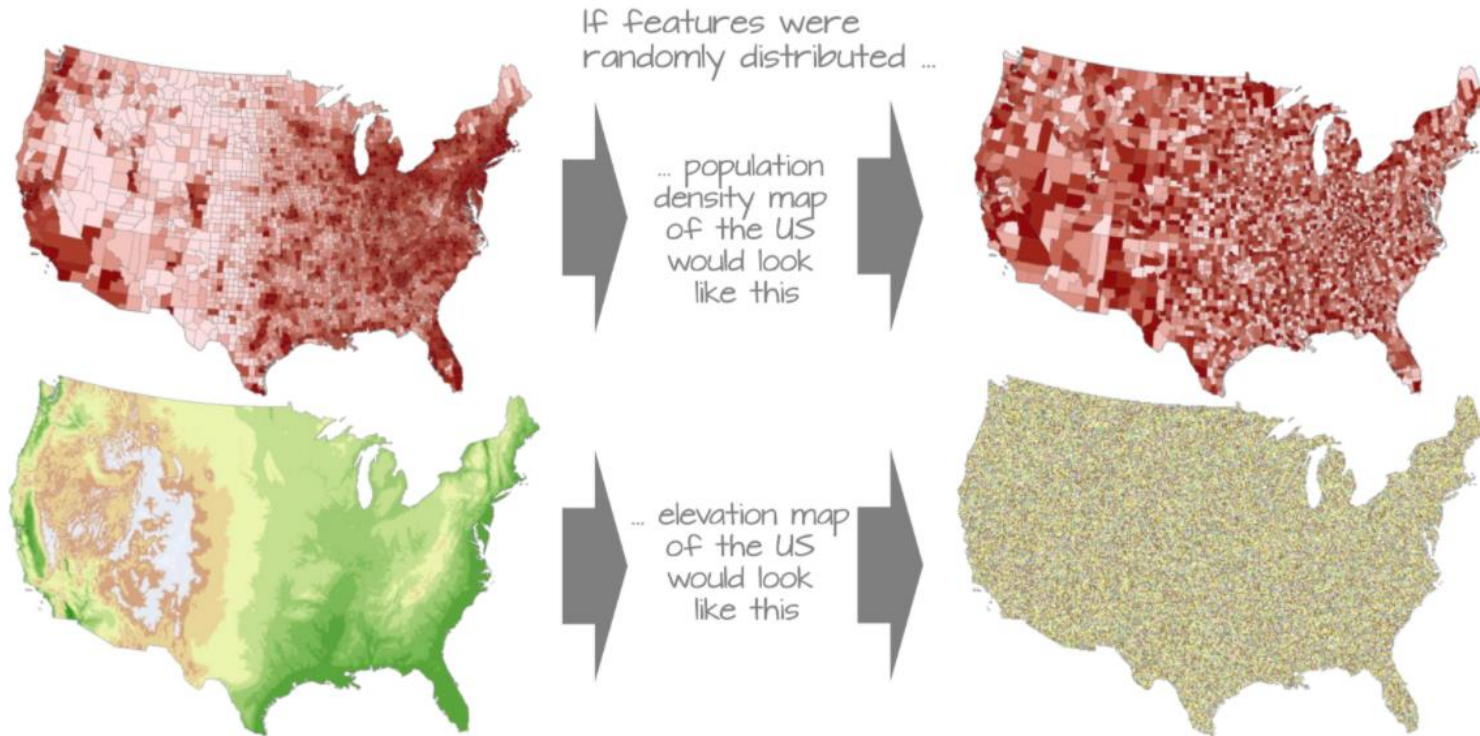


Image source: [Gimond \(2021\)](#)

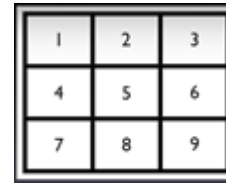
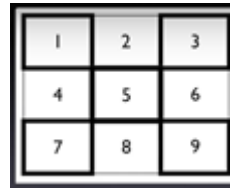
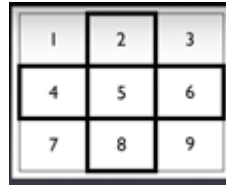
Spatial autocorrelation: **Neighbors**

- In order to understand the locational similarity, we need to have a formal way to define a **neighborhood**.
 - If the neighbors (and neighbors of neighbors) have similar values = positive spatial autocorrelation!
- Formalization of neighbors is done with a concept called **spatial weights** (W_{ij})
- Neighborhood can be defined in different ways, based on:
 - Shared borders → “contiguity weights”
 - Distance:
 - K nearest neighbors (KNN weights)
 - Weighting the influence according the distance decay (Kernel weights)
- With distance based spatial weights, distance is not necessarily measured as Euclidean spatial distance but it can also be e.g. network distance, time, etc.

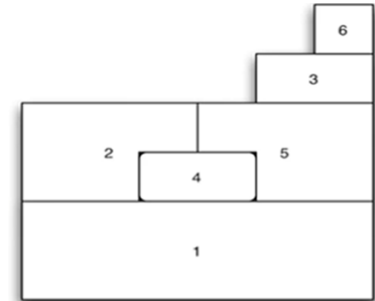
Neighbors: Contiguity weights

Look for neighbors based on common border (Polygon)

- Rook contiguity:
 - You consider horizontal and vertical neighbors
- Bishop contiguity:
 - You consider corners as neighbors
- Queen contiguity:
 - You consider all touching as neighbors



Six regions



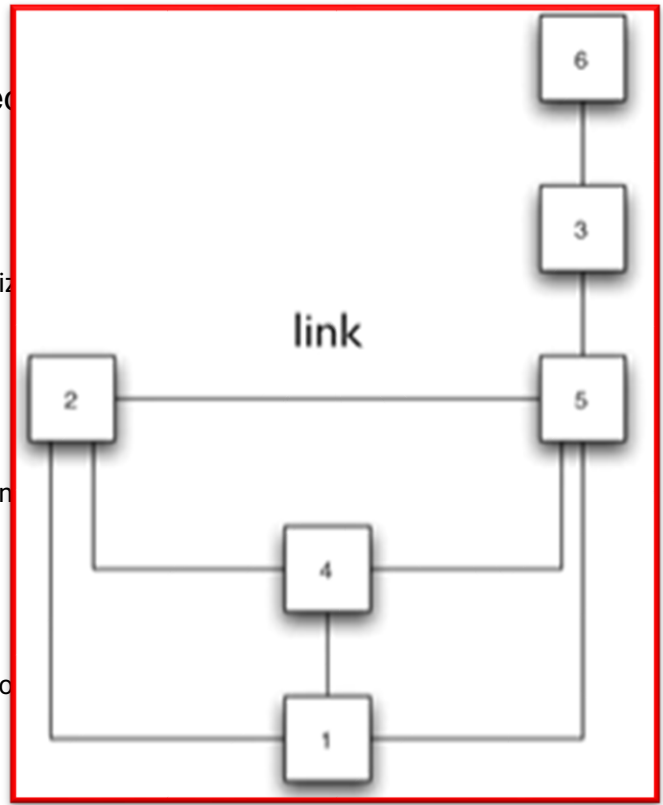
$$W = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Spatial weight matrix is typically used when doing statistical analysis (here: adjacency matrix)

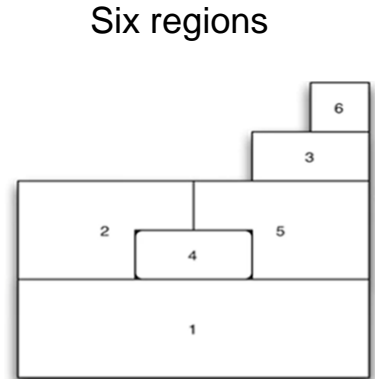
Neighbors: Contiguity weights

Look for neighbors based on

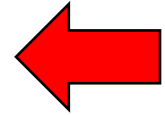
- Rook contiguity:
 - You consider horizontal and vertical neighbors
- Bishop contiguity:
 - You consider corners
- Queen contiguity:
 - You consider all to



Same neighborhood structure as a graph.



Six regions



$$W = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Spatial weight matrix is typically used when doing statistical analysis (here: adjacency matrix)

Neighbors: Contiguity weights

Row standardization

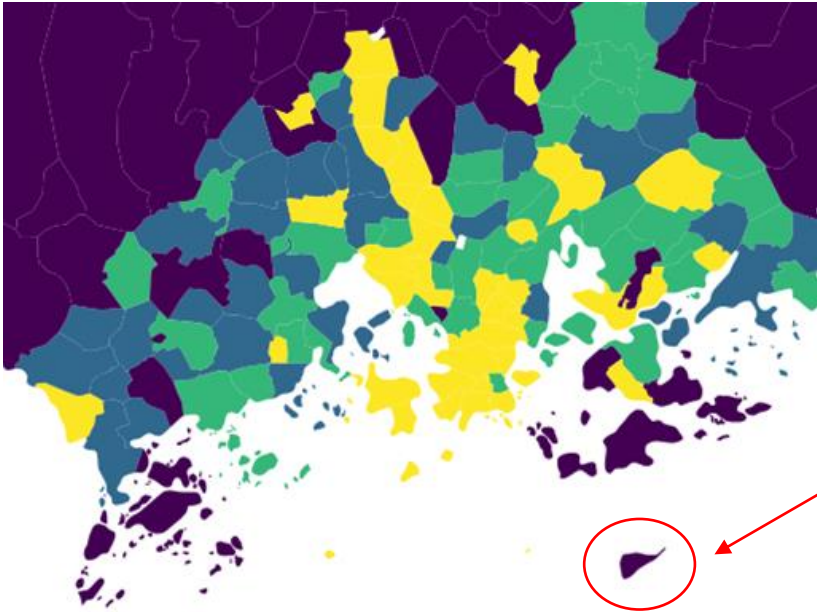
When using spatial weights in statistical analysis, the Boolean values (*is*, or *is not* a neighbor) are typically row standardized (rescaled) so that the sum of all neighbors is 1.

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \quad \rightarrow \quad \mathbf{W}^* = \begin{bmatrix} 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \\ 1/3 & 0 & 0 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

- Row-standardized weights
- Rescale the weights so that the sum equals to 1 $\rightarrow \sum_j W_{ij} = 1$
- Constraints parameter space (values are always between 0-1)
- Make analyses comparable
- Helps in constructing a so called "Spatial lag variable" = average of the neighbors (used in spatio-statistical models)

Neighbors: Contiguity weights

Possible problems



Islands do not have border neighbors! → “isolates”

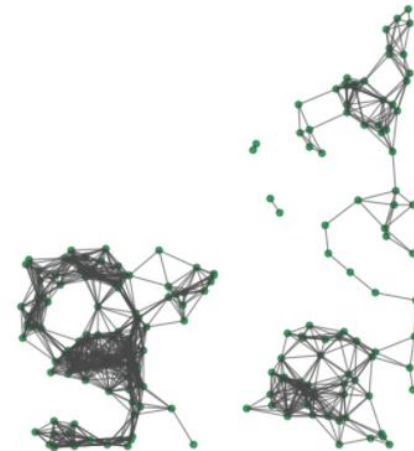
We need to deal with them somehow if doing spatial analysis, e.g. by removing them (lost of data).

Neighbors: Distance weights

- **KNN weights:** Neighborhood can be defined by specifying parameter k which specifies how many nearest points is considered as a neighbor to given location.
- **Distance-based weights:** Neighborhood can be defined by specifying a distance threshold, e.g. all points that are within 1 kilometer are considered as neighbors.
 - In **Kernel weights**, the locations that are closer will have higher weight than the distant ones



KNN connectivity graph with $k=6$



Connectivity graph with distance threshold of ~ 1 km

Spatial autocorrelation using Spatial weights

So how is the spatial autocorrelation calculated again?

- Sum over all observations of an attribute similarity measure with the neighbors

Moran's I is defined as

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Weights matrix (i and j are the locations) \rightarrow how is neighborhood defined?
The variable we are interested in (e.g. number of inhabitants)
Covariance term describes the overall variance (difference from the mean): $\sum(x_i - \bar{x})^2/n$

where N is the number of spatial units indexed by i and j ; x is the variable of interest; \bar{x} is the mean of x ; w_{ij} is a matrix of spatial weights with zeroes on the diagonal (i.e., $w_{ii} = 0$); and W is the sum of all w_{ij} .

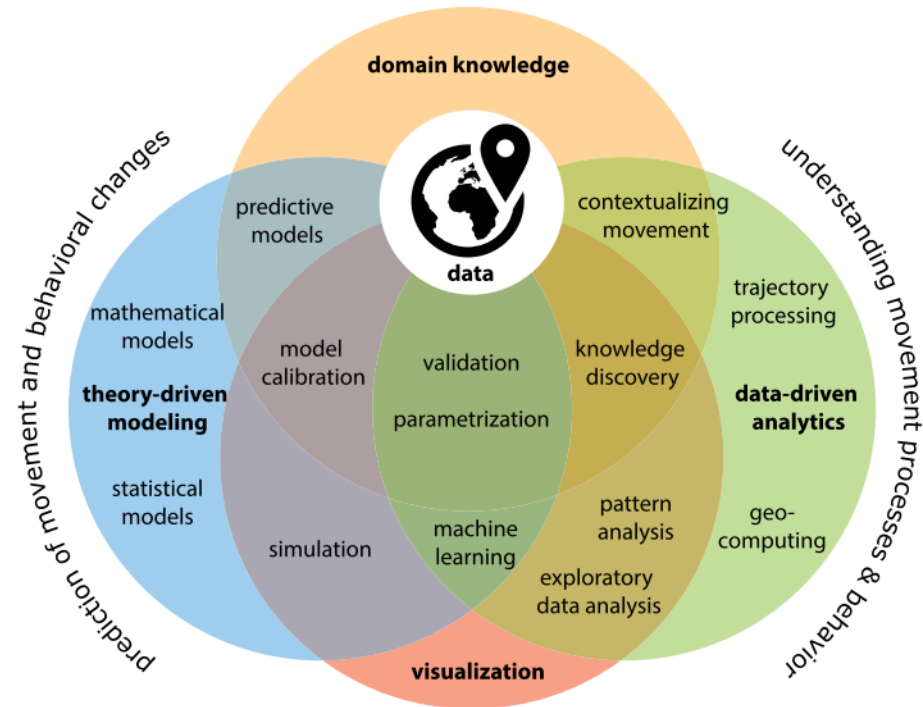
- Notice that the Moran's I can vary significantly depending on how we define/construct our weights matrix

Common tasks in spatial statistics

Common tasks: Why to use spatial statistics?

There are various reasons why you might be interested in applying spatial statistics:

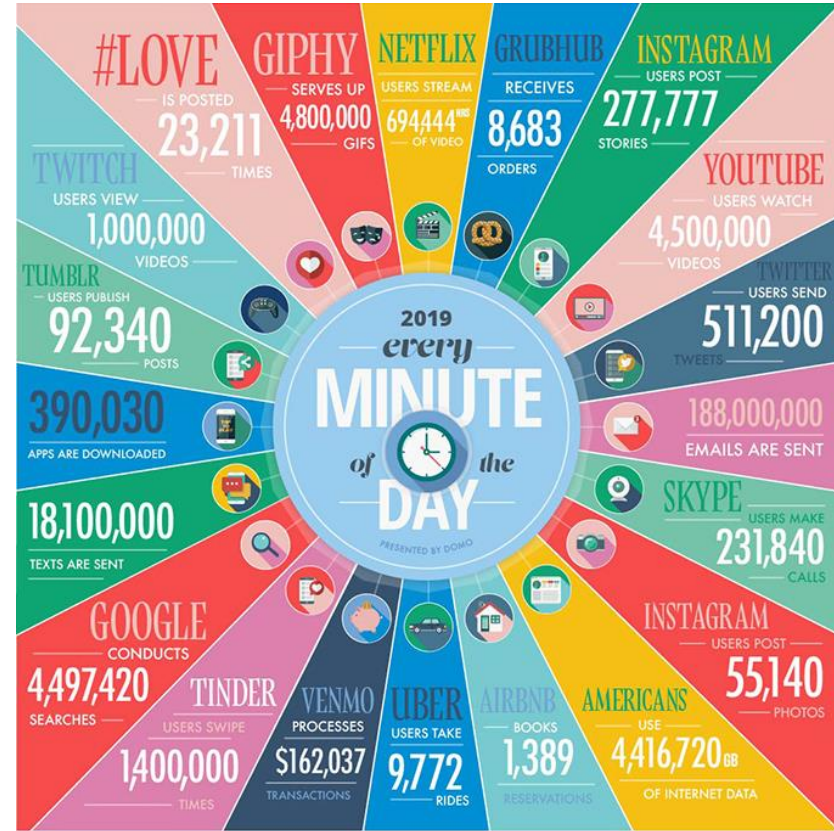
- You might be interested in the **spatio-temporal distribution** of the dataset:
 - E.g. by looking at autocorrelation, clustering, hotspots or detecting outliers
- You might be interested in **relationship and dependency between two or more variables** that are geographically referenced using statistical methods (also between different datasets)
- You might be interested in **understanding / revealing typical behavior** from the data, which enables to build models from the data
 - Can be used to predict future behavior / patterns



Example of a framework where spatial statistics / spatial data science can be used to both understand a phenomena and make predictions based on the common behaviors. (Dodge, 2019)

New sources of data – New challenges for spatial statistics?

Digitalization has pretty much changed everything



Source: domo.com

Physical <> hybrid <> digital space

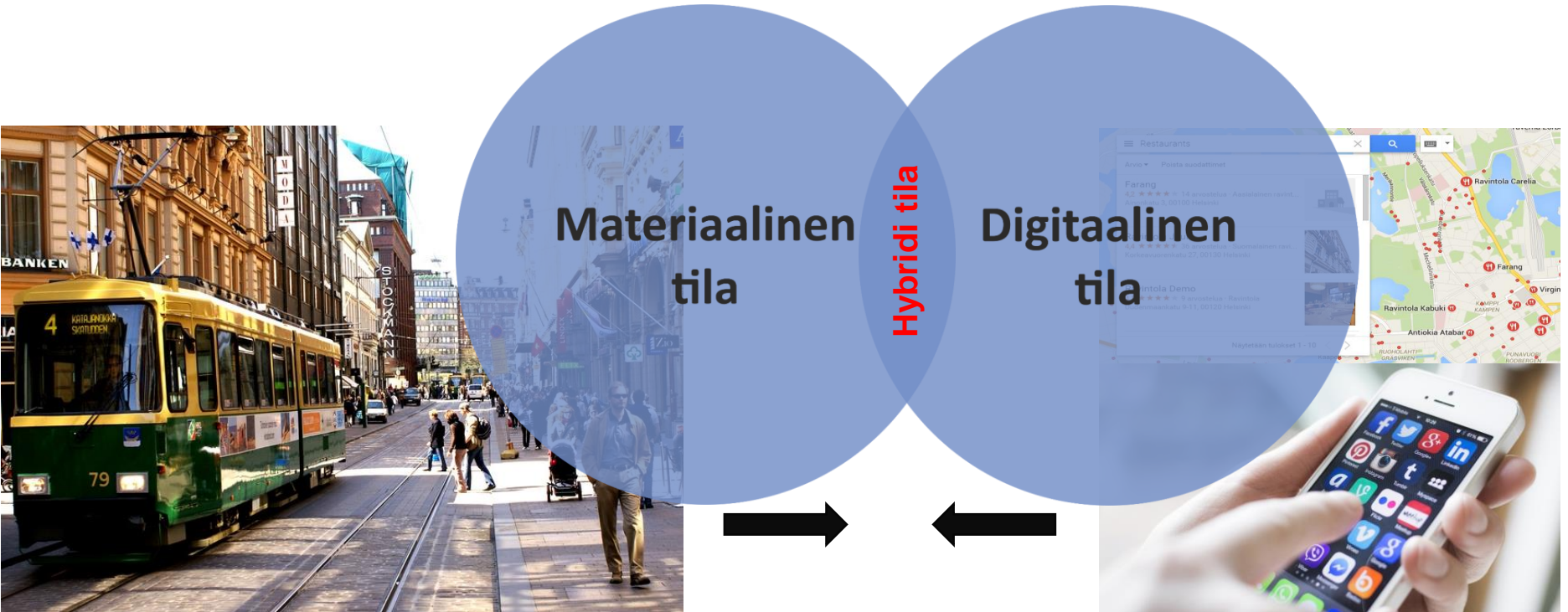


Materiaalinen
tila

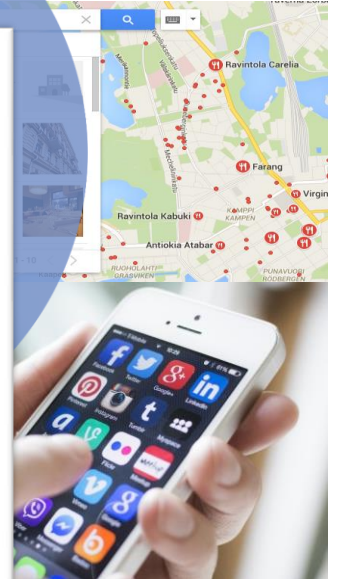
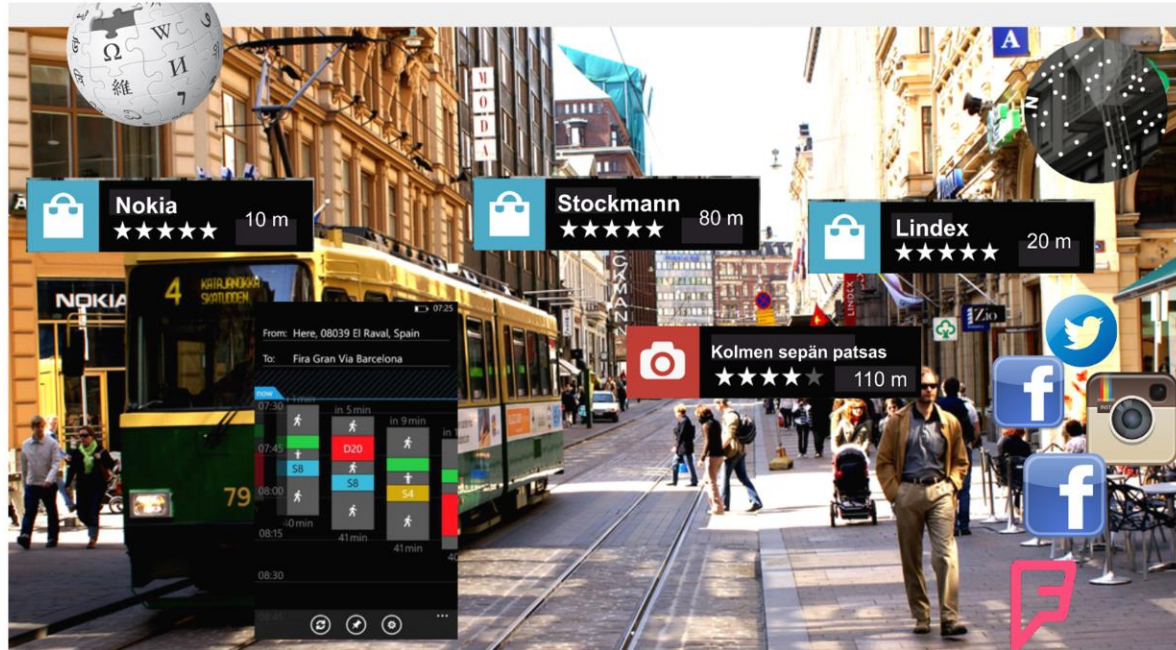


Digitaalinen
tila

Physical <> hybrid <> digital space



Physical <> hybrid <> digital space



Digital footprint



Tenkanen (2014)

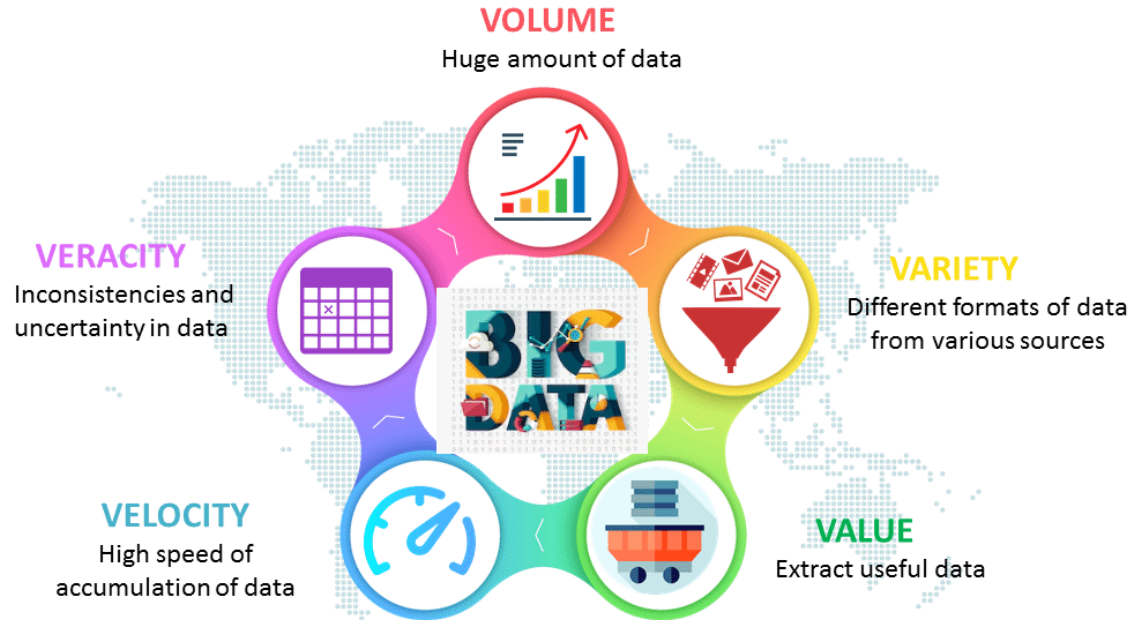
Digital footprint

- Digital footprint is a concept that is used to describe the layers digital content in geographical space that can describe/reveal human functions/actions based on data produced in the space itself.
- The previous example of the "world map" (based on Instagram posts) and the image on the right describing taxi drop-off/pick up locations are examples that describe the functioning of our society that is represented with **massive amount of observations** (*big data*)



Big data

- "Big data" is often used as a term to describe this new avalanche of digital content that surrounds us and is continuously accumulating.



Prospects and Challenges

Prospects

- Allows investigating various phenomena with higher spatio-temporal granularity than ever before
- Allows investigating social interactions as well as human-nature/environment interactions without needing to do expensive surveys/interviews
- Allows study of behavior without self-reporting errors / biases (the fact that you know you're being studied influences)

Challenges

- Many spatio-statistical methods were developed during an era when the data was **sparse**, new approaches are needed to be able to deal with massive datasets (scaling issues)
- **Biases / representativeness** → Who does the data represent? (e.g. if using social media data)
- **Data access issues** → most of the data that could be **used for good** are owned by private companies that are not eager to share the data for research (or do so with high price)
- **Privacy issues** → Cambridge Analytica ... surveillance economy .

References

Useful readings and references:

Anselin, L. & A. Getis (1992). Spatial statistical analysis and geographic information systems. *The Annals of Regional Science*.

Di Minin, E., H. Tenkanen & T. Toivonen (2015). Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*.

Dodge, S. (2019). A Data Science Framework for Movement. *Geographical Analysis*.

Miller, H. & M. Goodchild (2015). Data-driven geography. *GeoJournal*

Singleton, A. & Arribas-Bel, D. (2019). Geographic Data Science

Tobler, W. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*.

Yang, C., K. Clarke, S. Shekhar & C.V. Tao (2019). Big spatio-temporal data analytics: a research and innovation frontier. *International Journal of Geographical Information Science*.