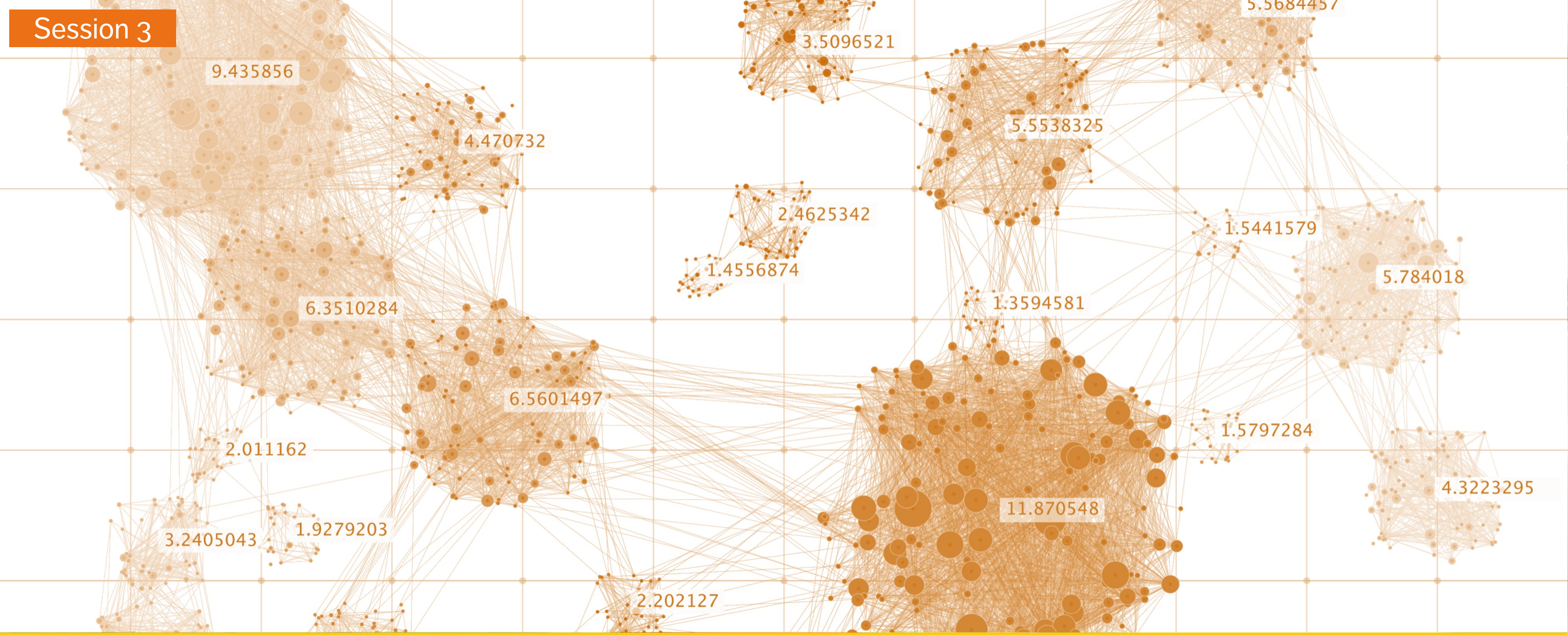


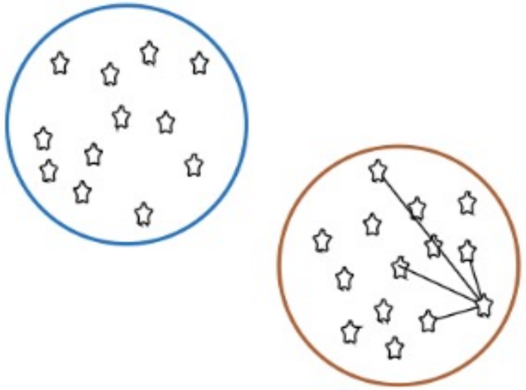
7. Clustering Overview

Overview

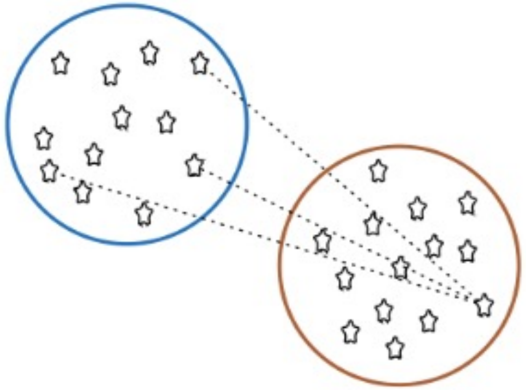
In **clustering**, the data is divided into **naturally occurring groups**. Within each group, the data points are **similar**; from group to group, they are **dissimilar**.

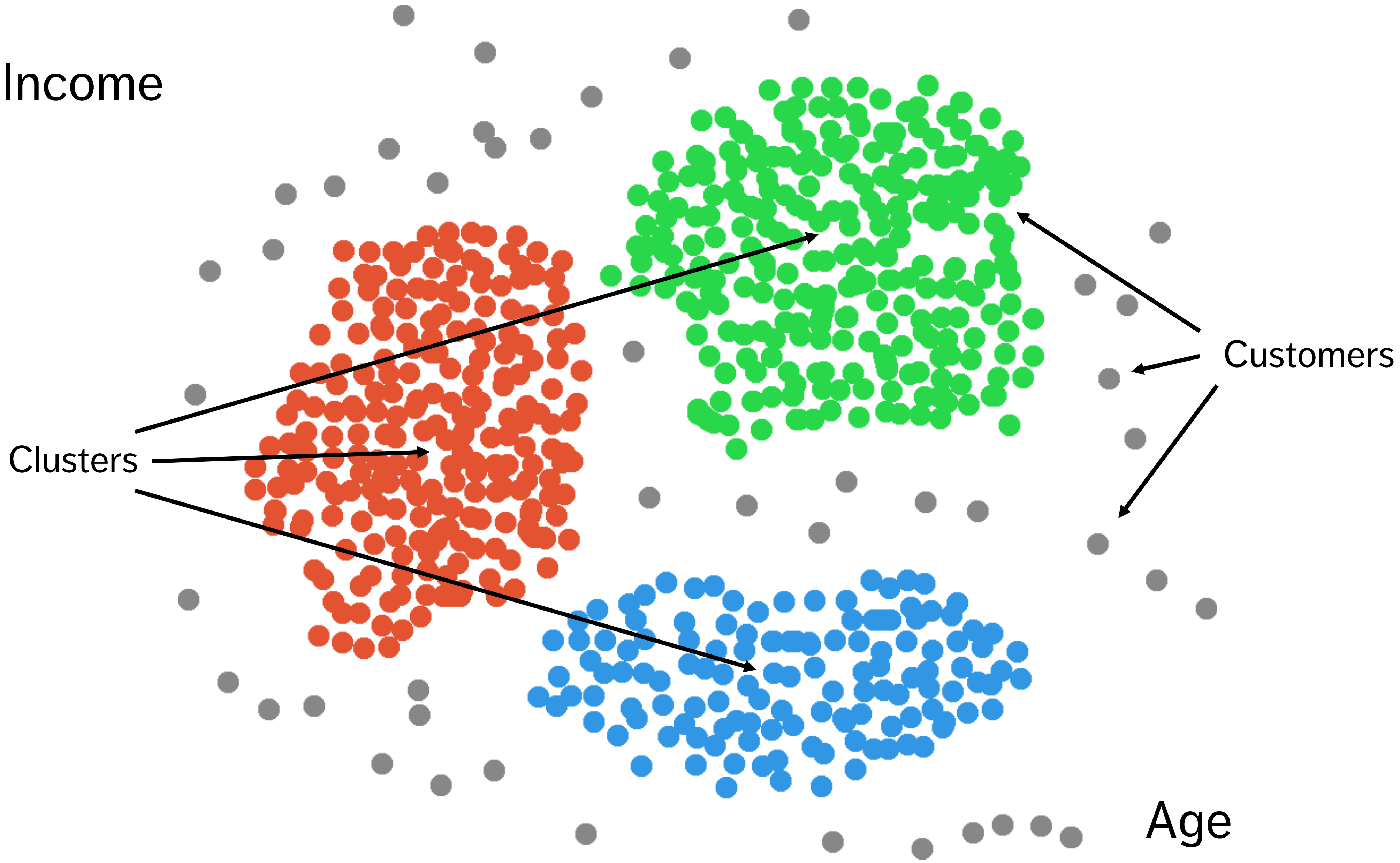
The grouping labels are not determined ahead of time, so clustering is an example of **unsupervised** learning.

average distance to points in own cluster (**low is good**)



average distance to points in neighbouring cluster (**high is good**)




Income
Clusters
Customers
Age

Overview

Clustering algorithms can be **complex** and **non-intuitive**, based on varying notions of similarities between observations.

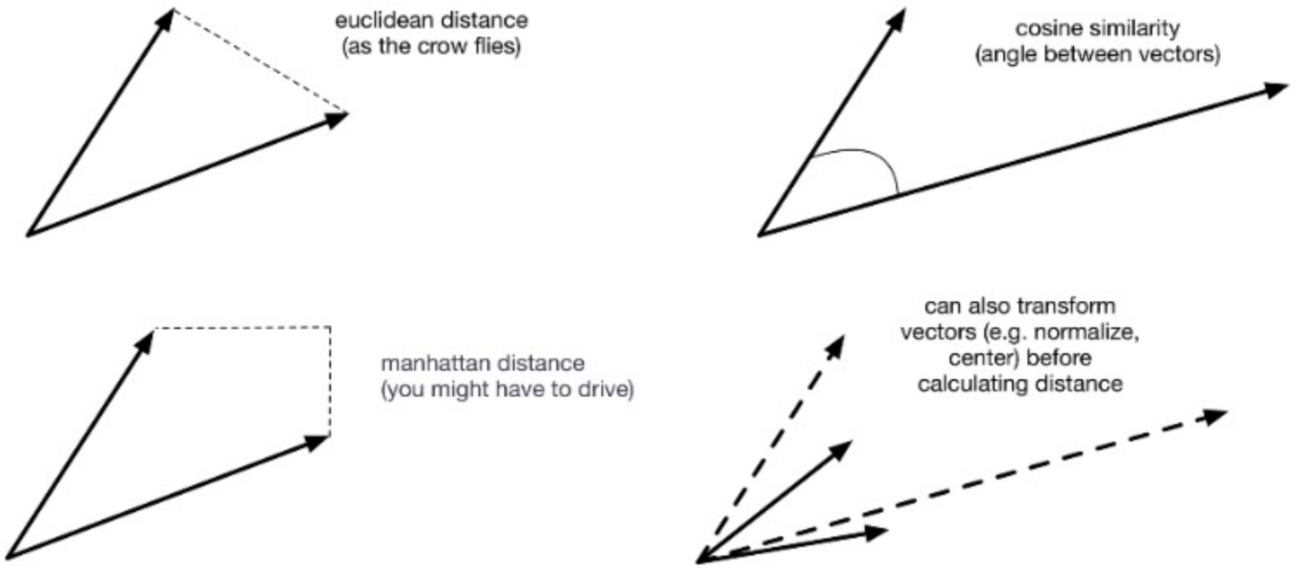
- in spite of that, the temptation to explain clusters *a posteriori* is **strong**

They are also (typically) **non-deterministic:**

- the same algorithm, applied twice (or more) to the same dataset, can discover completely different clusters
- the order in which the data is presented can play a role
- so can starting configurations

Clustering Requirement

A measure of **similarity** w (or a distance d) between observations:



IMPORTANT: data must be scaled before it is fed into clustering algorithms.

Typically, $w \to 1$ as $d \to 0$, and $w \to 0$ as $d \to \infty$.

Distance Measures (Metrics)

Categorical Variables*

- Hamming distance
- Russel/Rao index
- Jaccard
- Dice's coefficient
- etc.

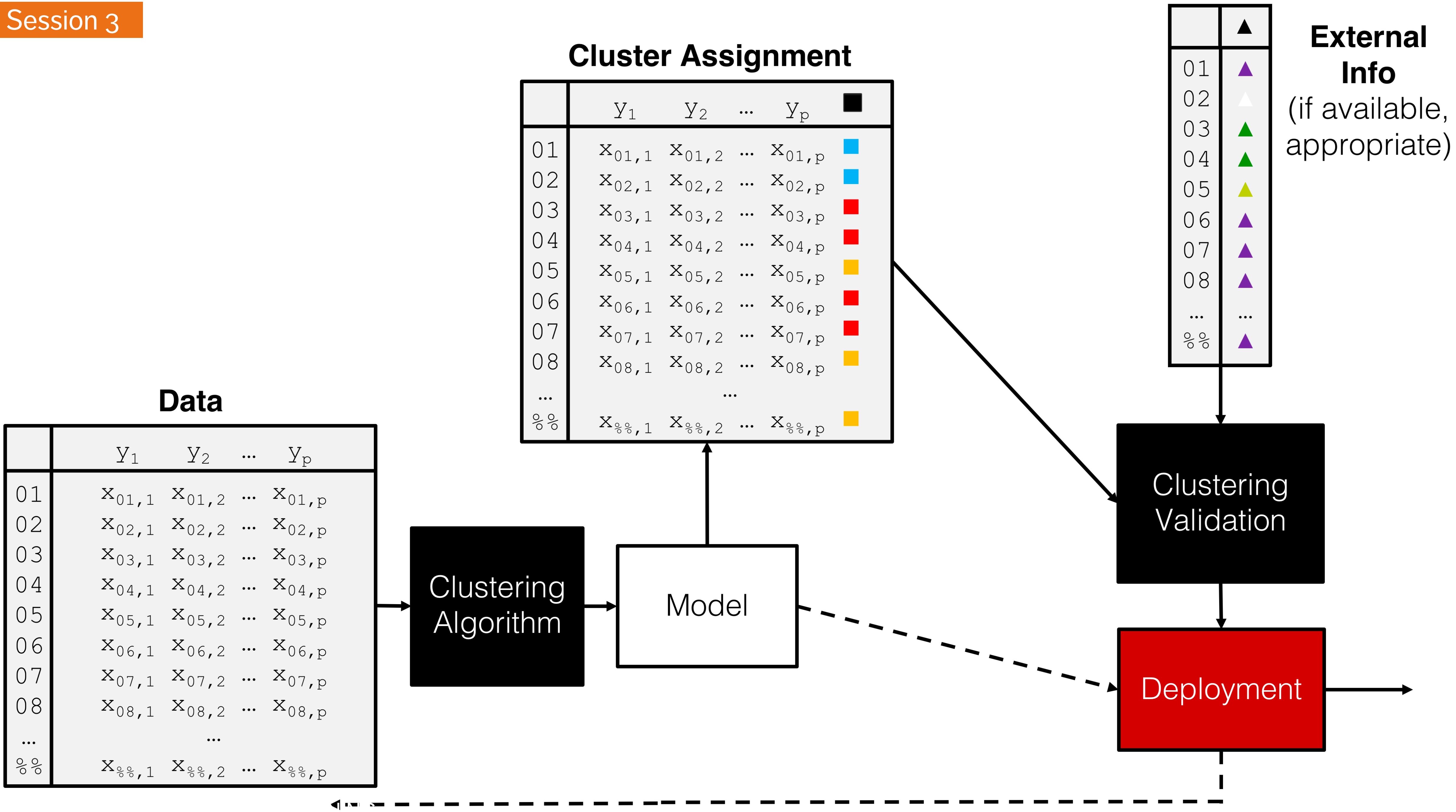
No steadfast rule to determine which distance to use; competing schemes are often produced with diff. metrics.

Numerical Variables

- Euclidean
- Manhattan
- correlation
- cosine
- etc.

We may need to create hybrid metrics for dataset with both categorical and numerical variables.

Session 3


Cluster Assignment
External Info
(if available, appropriate)
Data
Clustering Algorithm
Model
Clustering Validation
Deployment

Applications

Text Documents

- grouping similar documents according to their topics, based on the patterns of common and unusual words

Product Recommendations

- grouping online purchasers based on the products they have viewed, purchased, liked, or disliked
- grouping products based on customer reviews

Marketing and Business

- grouping client profiles based on their demographics and preferences

Applications

Dividing a larger group (or area, or category) into **smaller** groups, with members of the smaller groups guaranteed to have similarities of some kind.

- tasks may then be solved separately for each of the smaller groups
- this may lead to increased accuracy once the separate results are aggregated

Creating taxonomies **on the fly**, as new items are added to a group of items

- this would allow for easier product navigation on a website like Netflix, for instance

Case Study

Livehoods

Cranshaw *et al.*
[The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City]
ICWSM, 2012

Objective

When we think of similarity at the urban level, we typically think in terms of neighbourhoods. Is there some other way to identify similar parts of a city?

The researchers aims to draw the boundaries of **livehoods**, areas of similar character within a city, by using clustering models. Unlike **static** administrative neighborhoods, the livehoods are defined based on the **habits** of their inhabitants.

Case Study

Livehoods

Cranshaw *et al.*
[The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City]
ICWSM, 2012

Methodology

The authors use **spectral clustering** to discover **distinct geographic areas** of the city based on collective **movement patterns**.

Livehood clusters are built as follows:

1. a **geographic distance** is computed based on pairs of check-in venues' coordinates;
2. a **social similarity** is computed between each pair of **venues** using cosine measurements;
3. spectral clustering produces **candidate livehoods**;
4. interviews are conducted with residents in order to **explore**, **label**, and **validate** the clusters discovered by the algorithm.

Case Study

Livehoods

Cranshaw *et al.*
[The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City]
ICWSM, 2012

Data

The data comes from two sources, combining approximately 11 million check-ins from the dataset of Chen et al. (a recommendation site for venues based on users' experiences) and a new dataset of 7 million Twitter check-ins downloaded between June and December of 2011.

For each check-in, the data consists of the **user ID**, the **time**, the **latitude and longitude**, the **name of the venue**, and its **category**.

In this case study, data from the city of Pittsburgh, Pennsylvania, is examined *via* 42,787 check-ins of 3840 users at 5349 venues.

Case Study

Livehoods

Cranshaw *et al.*
[The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City]
ICWSM, 2012

Strengths and Limitations of the Approach

- The technique used in this study is **agnostic** towards the particular source of the data: it is not dependent on meta-knowledge about the data.
- The algorithm may be prone to "majority" bias, possibly misrepresenting/hiding minority behaviours.
- The dataset is built from a **limited** sample of check-ins shared on Twitter and are therefore biased towards the types of visits/locations that people typically want to share **publicly**.
- Tuning the clusters is non-trivial: experimenter bias may combine with "confirmation bias" of the interviewees in the validation stage – if the researchers are residents of Pittsburgh, will they see clusters when there were none?

Case Study

Livehoods

Cranshaw *et al.*
[The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City]
ICWSM, 2012

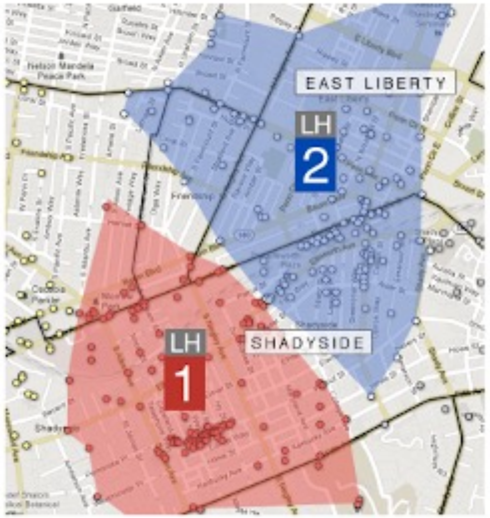
Results, Evaluation, and Validation

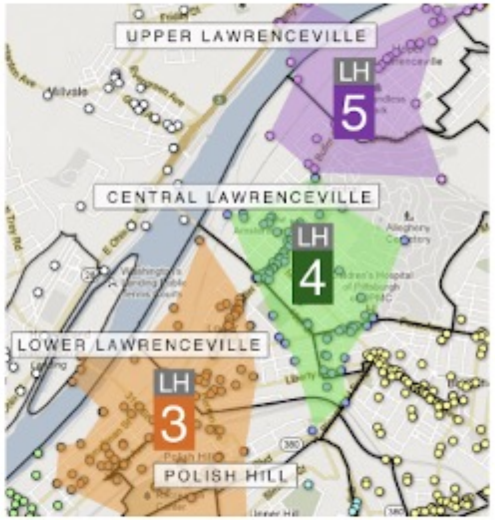
Over 3 areas of the city, 9 livehoods have been identified and validated by 27 Pittsburgh residents

- **Municipal Neighborhoods Borders:** livehoods are dynamic, and evolve as people's behaviours change, unlike fixed neighbourhoods set by the city government.
- **Demographics:** the interviews displayed strong evidence that the demographics of the residents and visitors of an area play a strong role in explaining the livehood divisions.
- **Development and Resources:** economic development can affect the character of an area. Similarly, the resources provided by a region has a strong influence on the people that visit it, and hence its resulting character.

Case Study

Livehoods

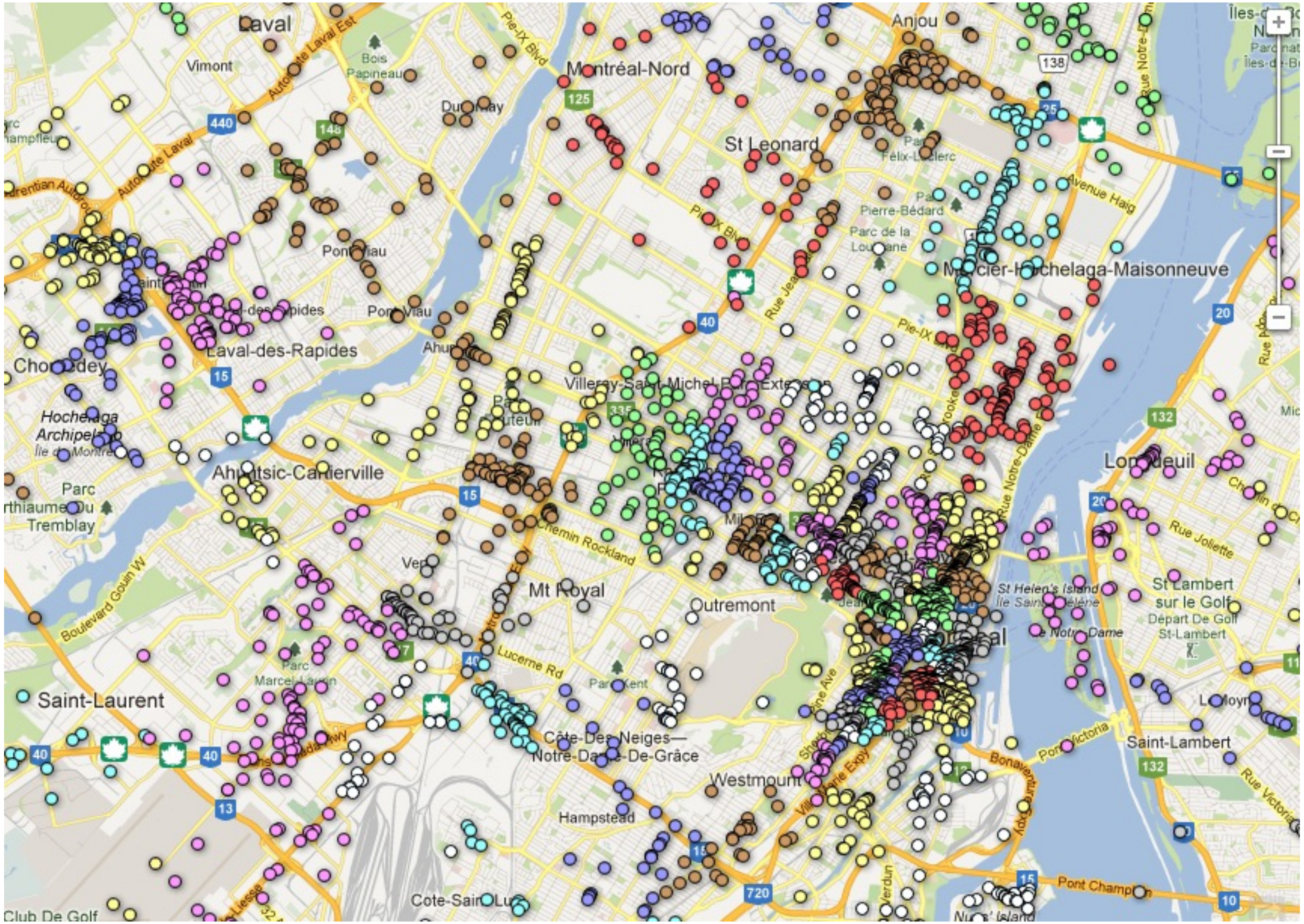




Cranshaw *et al.*
[The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City]
ICWSM, 2012

Case Study

Livehoods



Cranshaw *et al.*
[The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City](#)
ICWSM, 2012

General Remarks

Clustering is a relatively **intuitive** concept for human beings as our brains do it unconsciously:

- facial recognition
- searching for patterns, etc.

In general, people are very good at **messy** data, but computers and algorithms have a harder time.

Part of the difficulty is that there is **no agreed-upon definition of what constitutes a cluster:**

- "I may not be able to define what it is, but I know one when I see one"

Suggested Reading

Clustering Overview

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Clustering

- Overview
- Case Study: Livehoods

Spotlight on Clustering

*Overview (advanced)

- Unsupervised Learning
- Clustering Framework
- A Philosophical Approach to Clustering

Exercises

Clustering Overview

1. What does the (potential) non-replicability of clustering imply for validation? For client and/or stakeholder buy-in?
2. Identify scenarios and questions that could use classification and/or value estimation in your every day work activities.