



City Rhythm

Model Background & Validation

Wednesday, May 31st, 2017

A short overview of today's presentation

1. A new way of looking at data
2. The Mixed Hidden Markov Model
3. The MHMM versus descriptive models



Mark Verhagen

Statistical Modeling
*Background in Econometrics, Art
History*

+31 6 81 38 12 37

mark@delph.io



Ruben Spruit

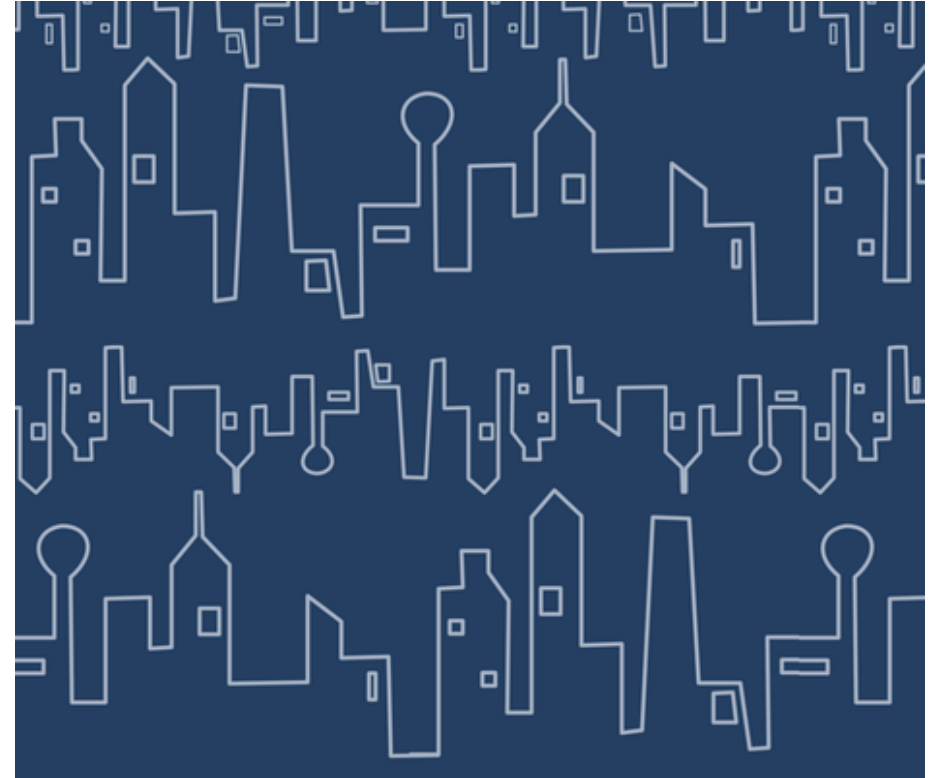
Statistical Modeling
*Background in Econometrics, Data
Science*

+31 6 14 60 50 04

ruben@delph.io

Why is the CR project so interesting?

1. **A new way of looking at data**
2. The Mixed Hidden Markov Model
3. The MHMM versus descriptive models



More and more data is coming in increasingly diverse types.. New approaches are necessary to analyze them!



Messaging data, GPS movement etc...



Movement, food intake, sleep patterns



Interaction between people



Erosion, nature development



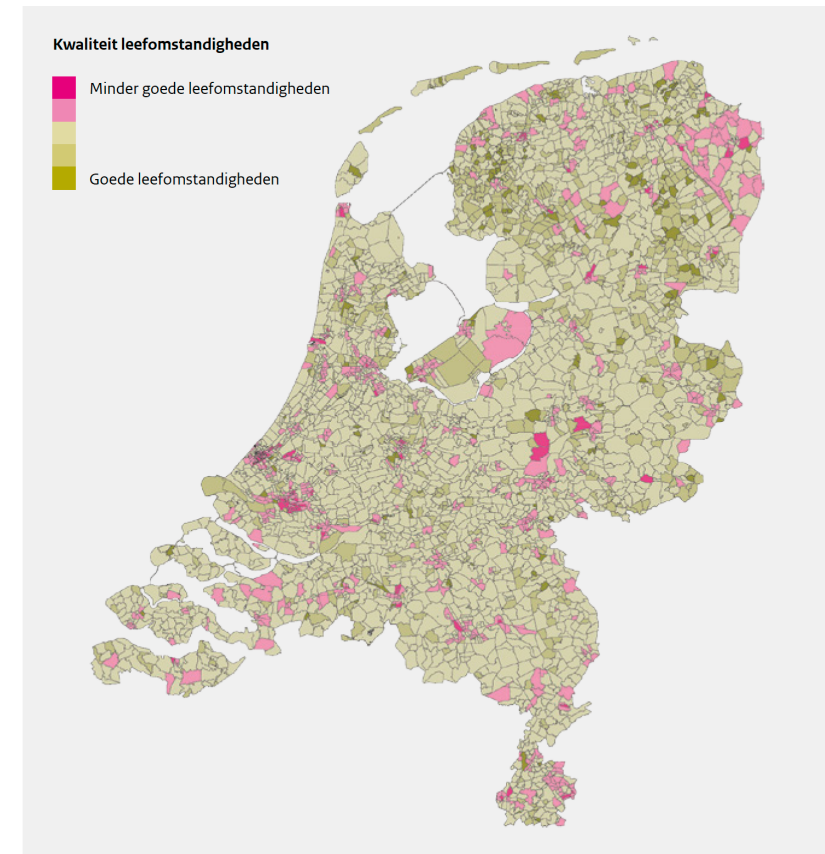
Centraal Bureau voor de Statistiek

'Old' statistics

Past presentation we discussed the problems concerning **aggregation**

‘Given the available data, this is what your neighborhood was like in 2015...’

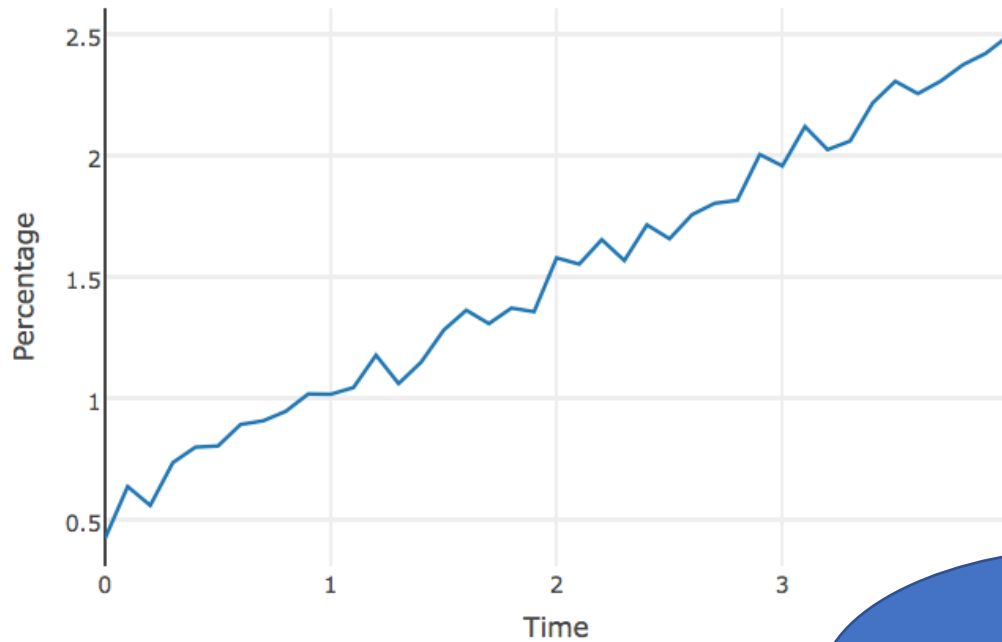
- How about the past?
- How about recurring themes in a neighborhood?
- Is it really static? Black and white?
- What about a more **organic** and **dynamic** flow to data?



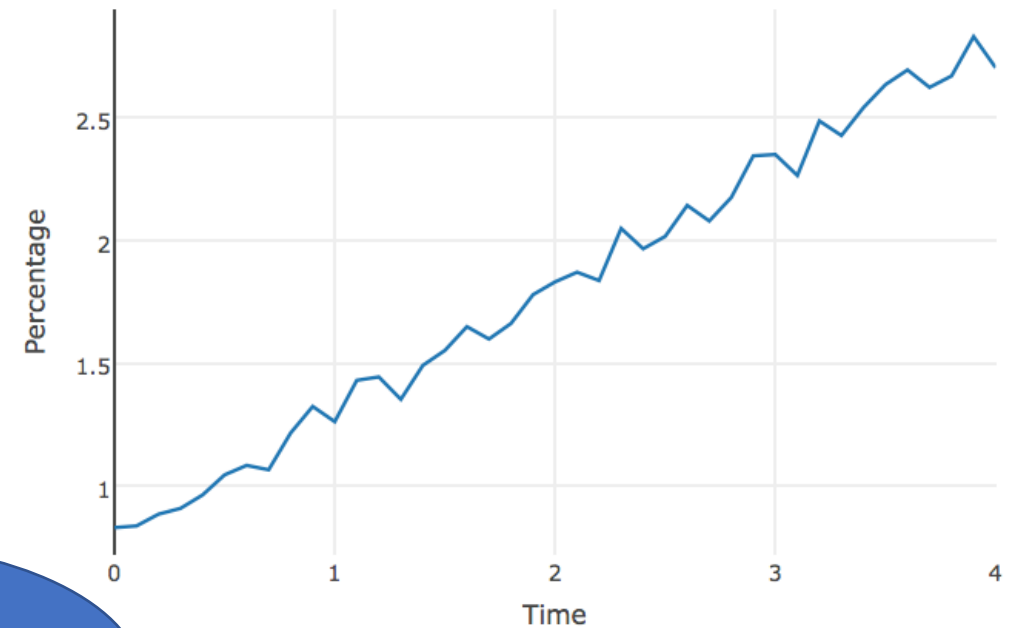
Bron: Tierolf, Gilsing en Steketee, 2017

Usually, analysis is done on **single** variables, but the world is **multivariate**

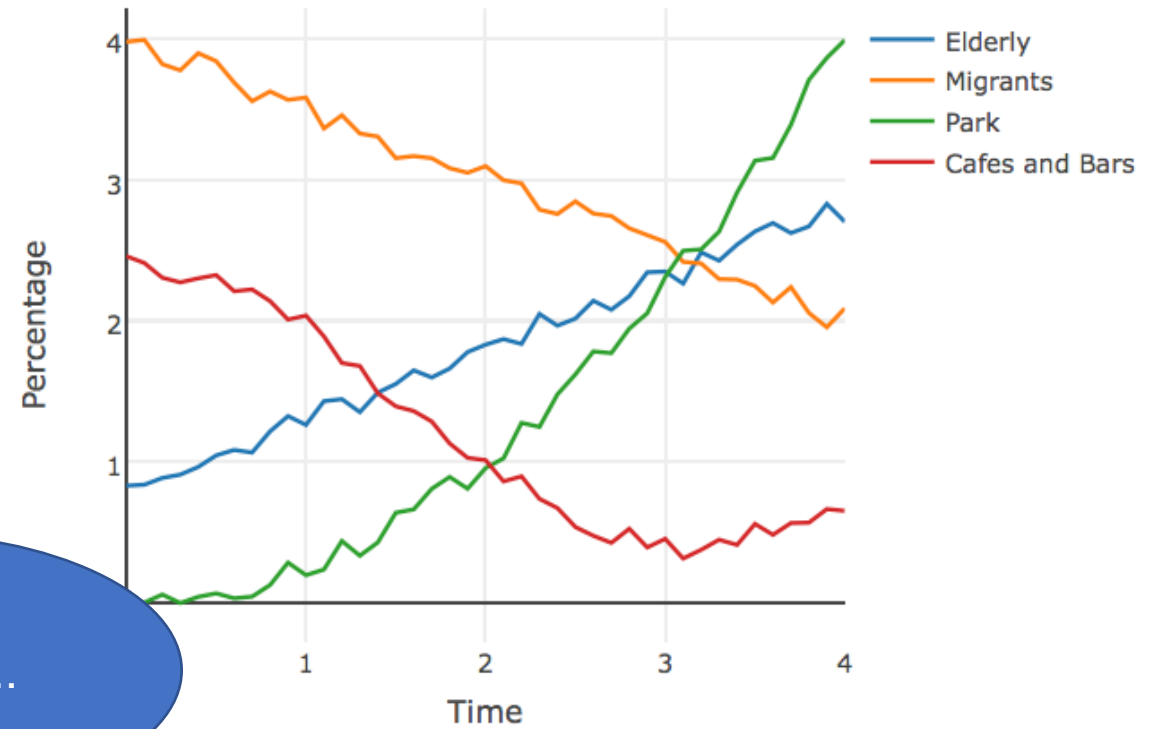
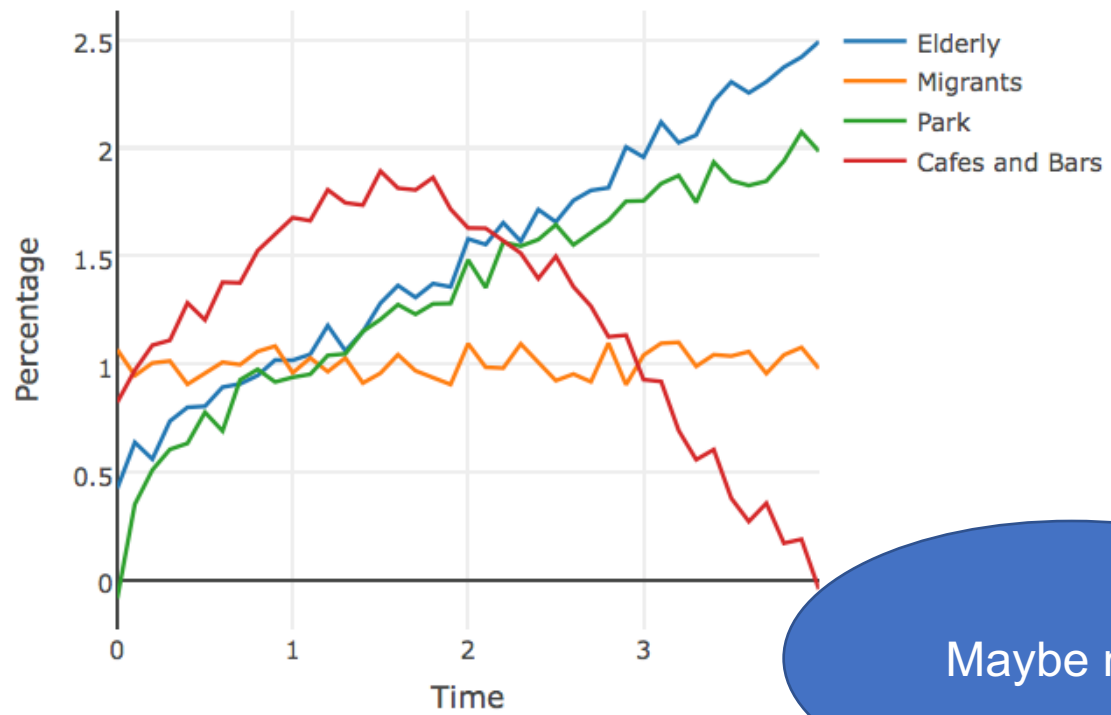
Neighborhood 1



Neighborhood 2



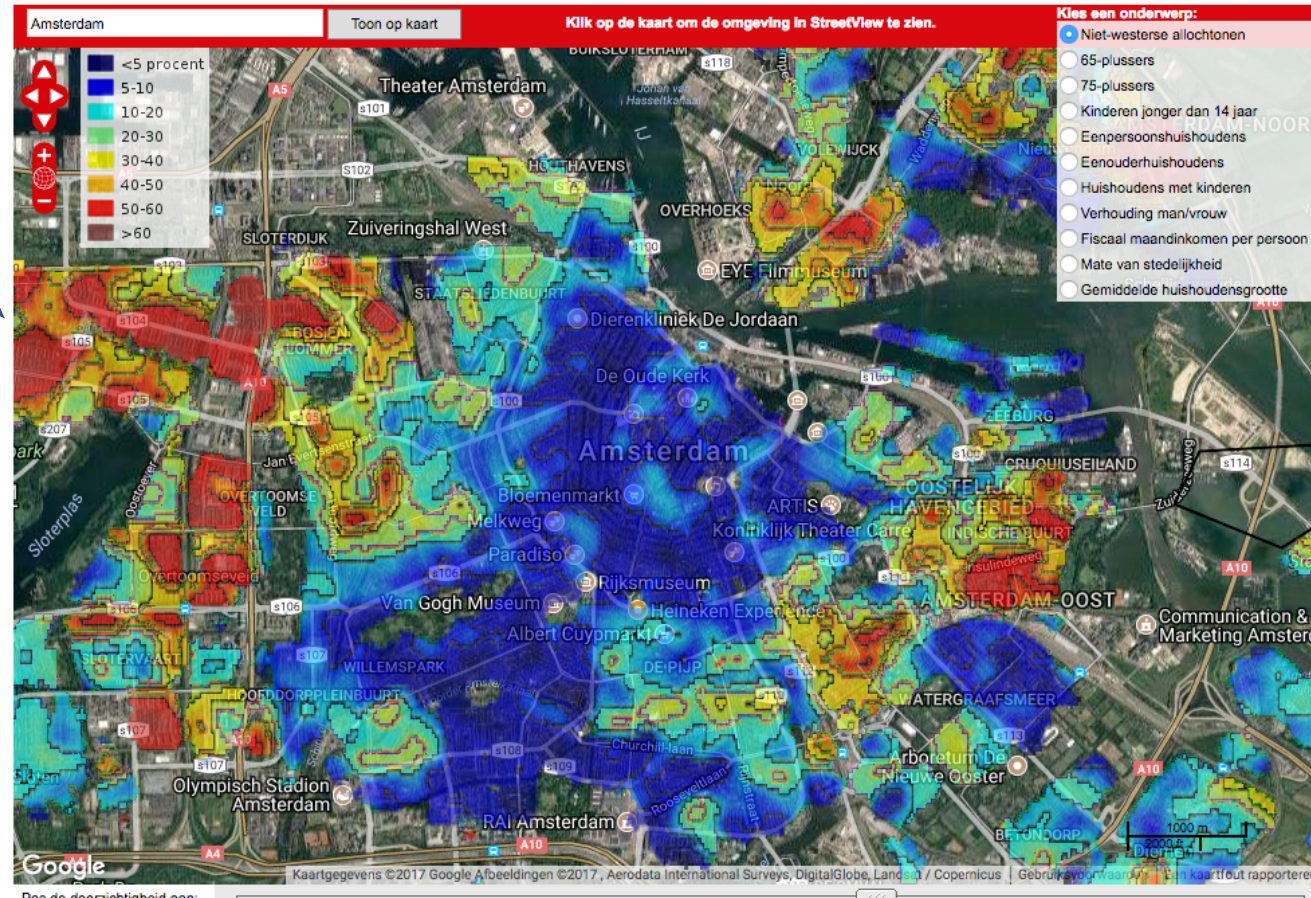
Similar
Neighborhoods?



Maybe not..

Once data comes in so many different ways, can we still understand what is happening through classical techniques?

This map alone can be generated for 11 different variables at different time points.



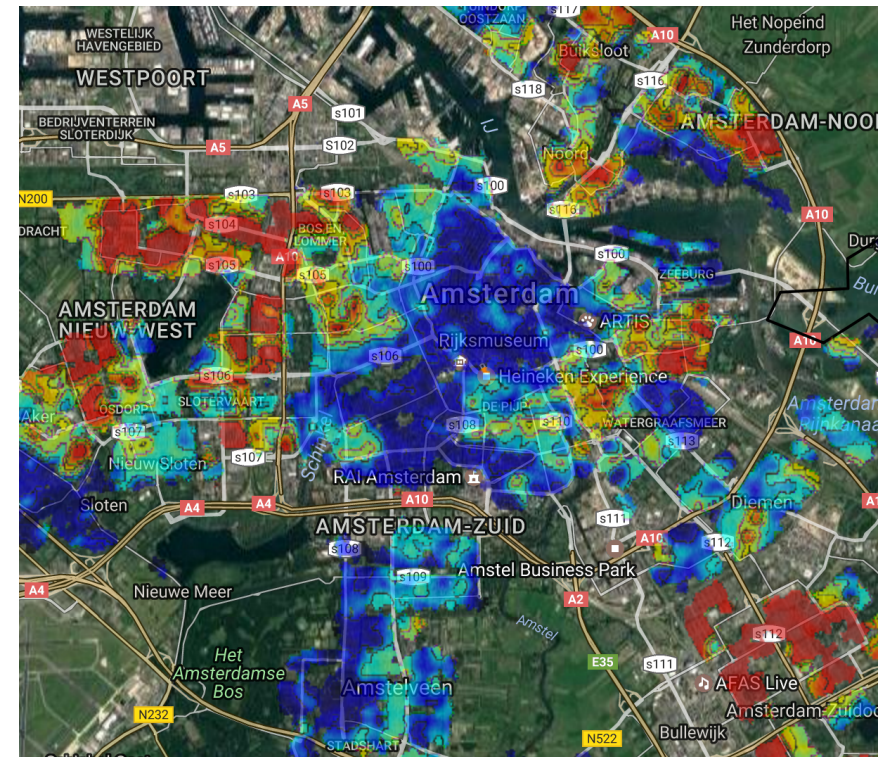
And we haven't talked about the 250+ variables that CBS tracks for every grid in the Netherlands!

Let alone all those other datapoints available..



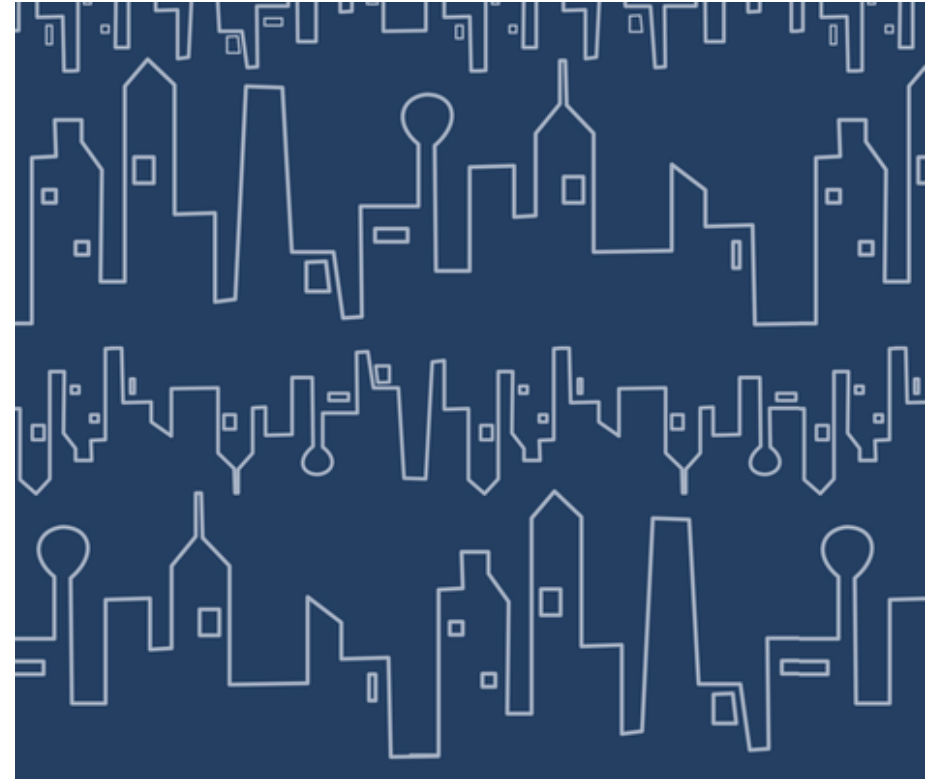
05/07/2017

In this project we will look at the development of grids, but in a **multivariate** and **organic** way.



Short recap: how did the MHMM work again?

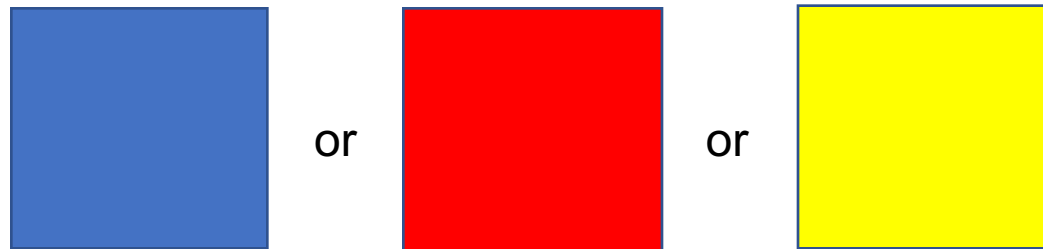
1. A new way of looking at data
- 2. The Mixed Hidden Markov Model**
3. The MHMM versus descriptive models



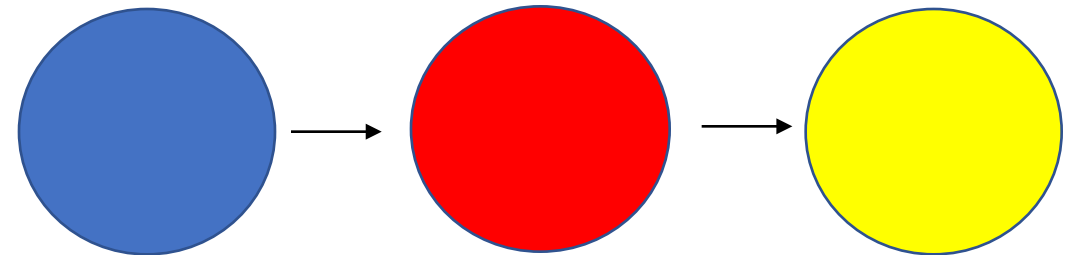
The (Mixed) Hidden Markov Model is designed to capture the **organic** and **dynamic** structure of data.

- Assumes that the things we observe (data) can be related to a handful of **base states**.
- The entities we are tracking (i.e. neighborhoods, reservations) develop through these states.
- We thus observe a sequence of states following each other.
- We let the data speak, instead of the model!

Classical approach



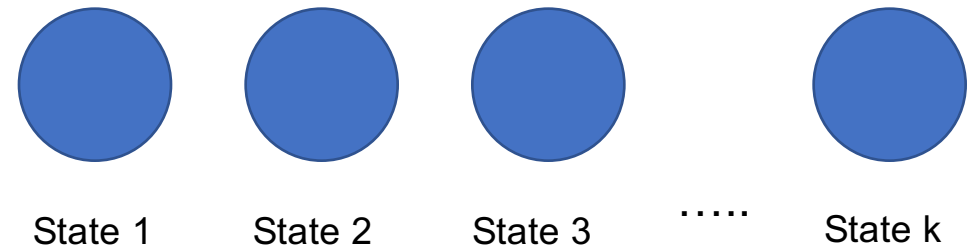
CR approach



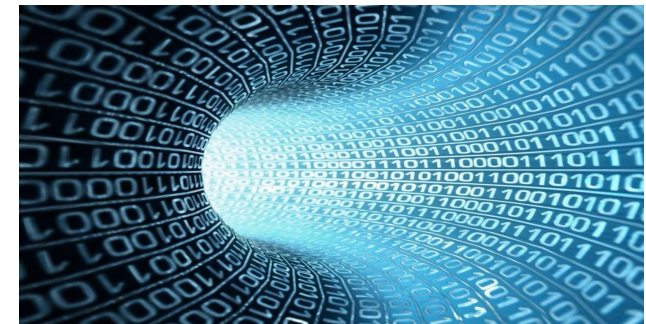
All we need is a number of states: k and a dataset.

- We are tracking observations that change over time (people migrate, things are built etc.)
- We assume that these observations belong to a limited amount of **states**.
- Over time, observations go from one state to another and back.
- We do **not** specify what each state is beforehand, we let the data speak.

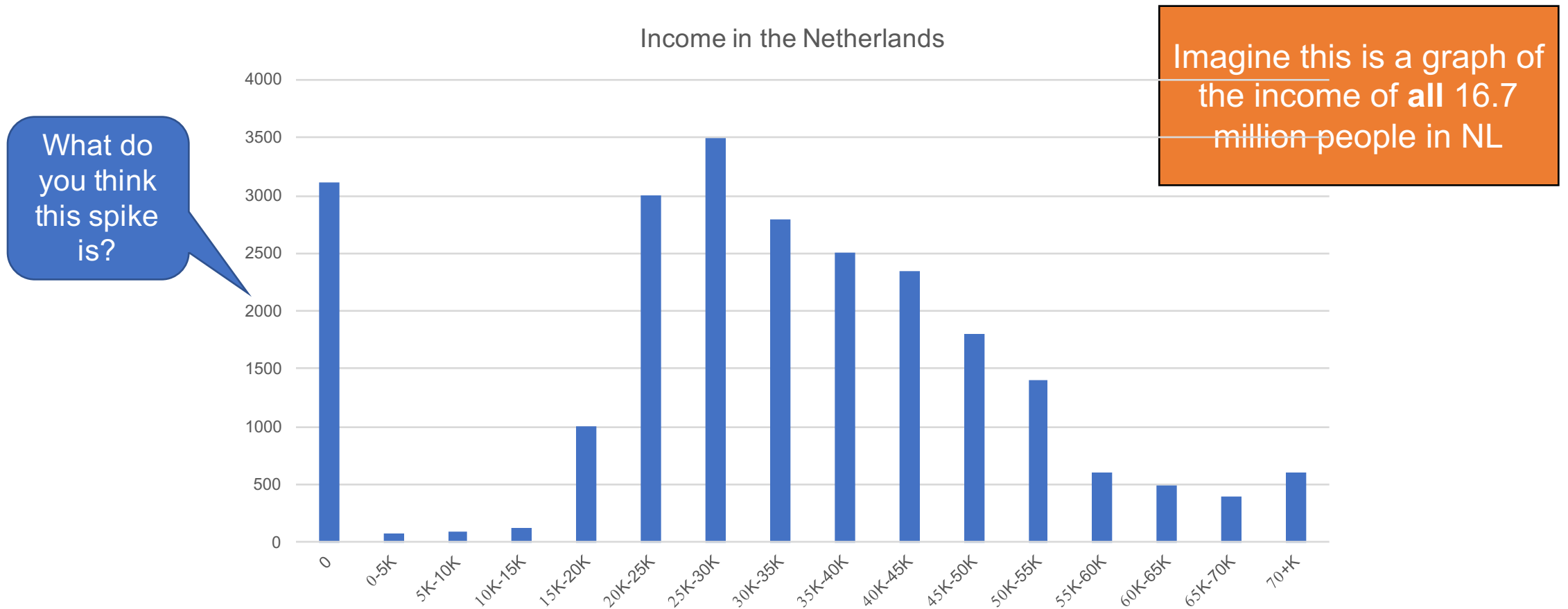
Requirements for the HMM:



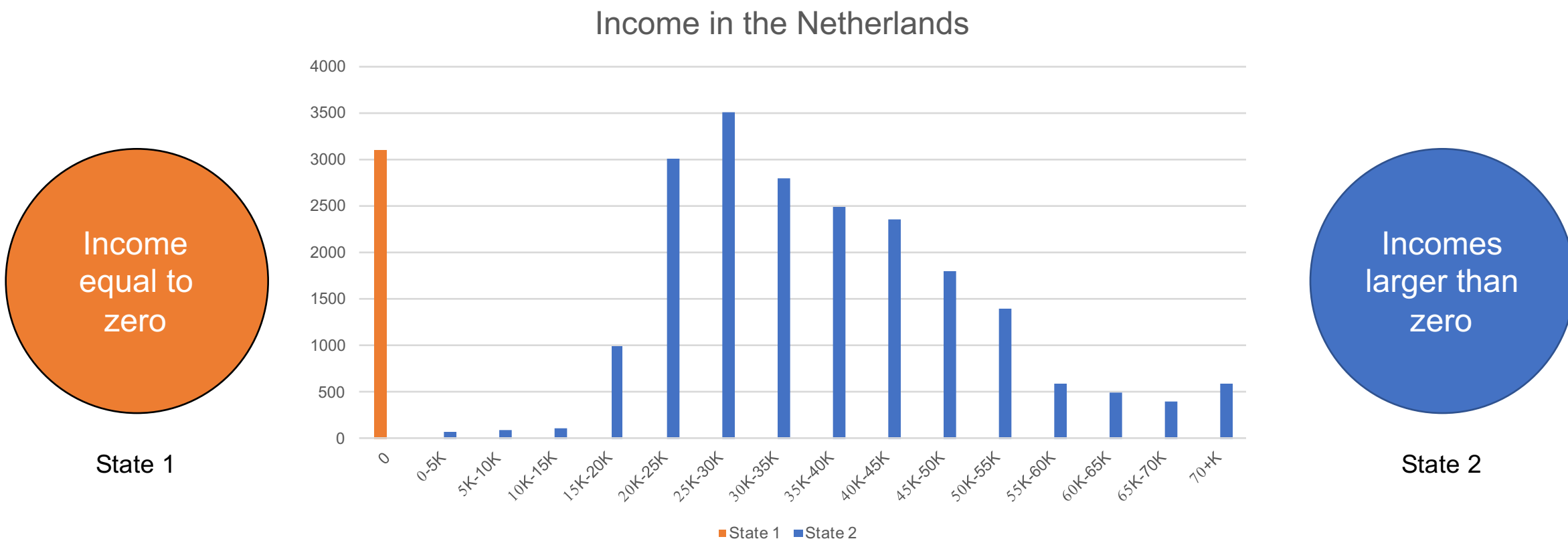
+



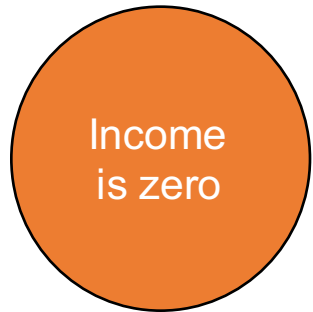
Imagine we have data of the income of everyone in the Netherlands.



This data could be represented by a HMM with just two states.



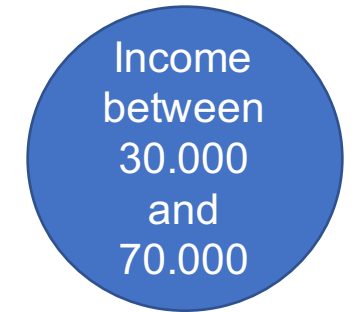
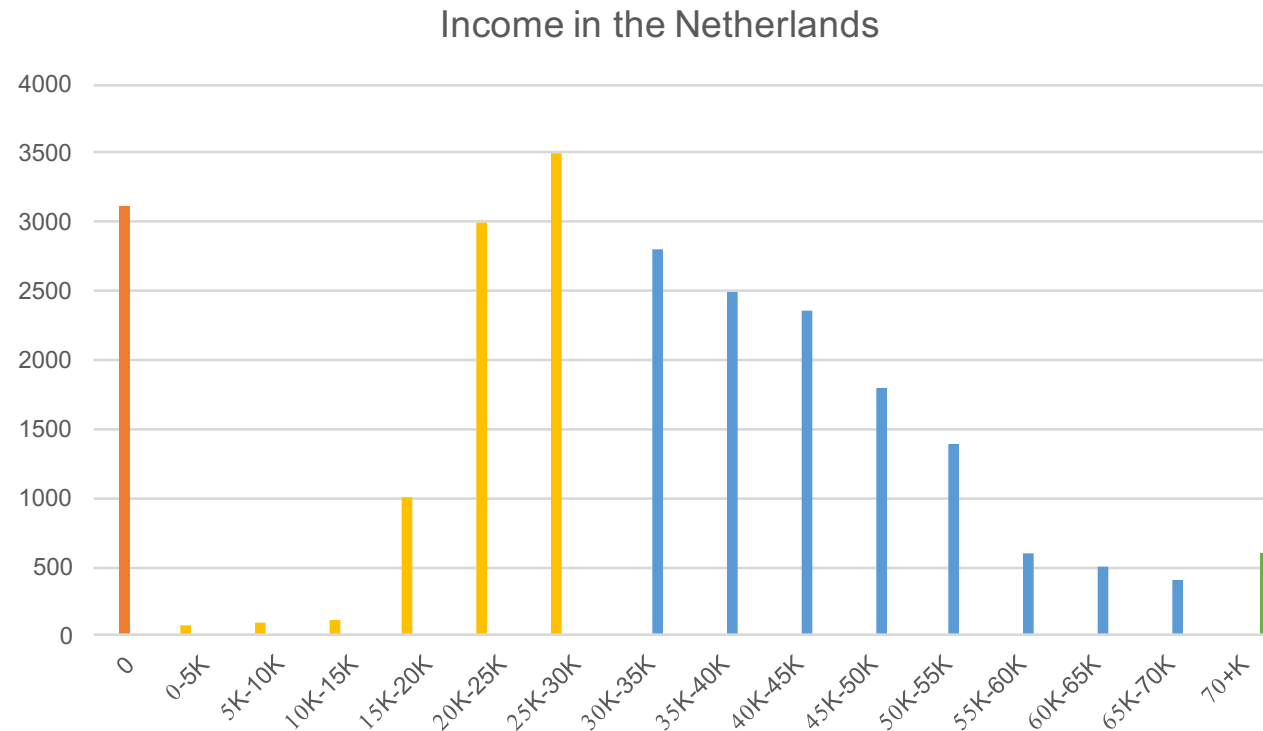
If we add more states, we would get more specific insights on the income distribution.



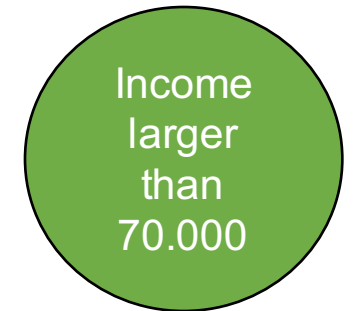
State 1



State 2

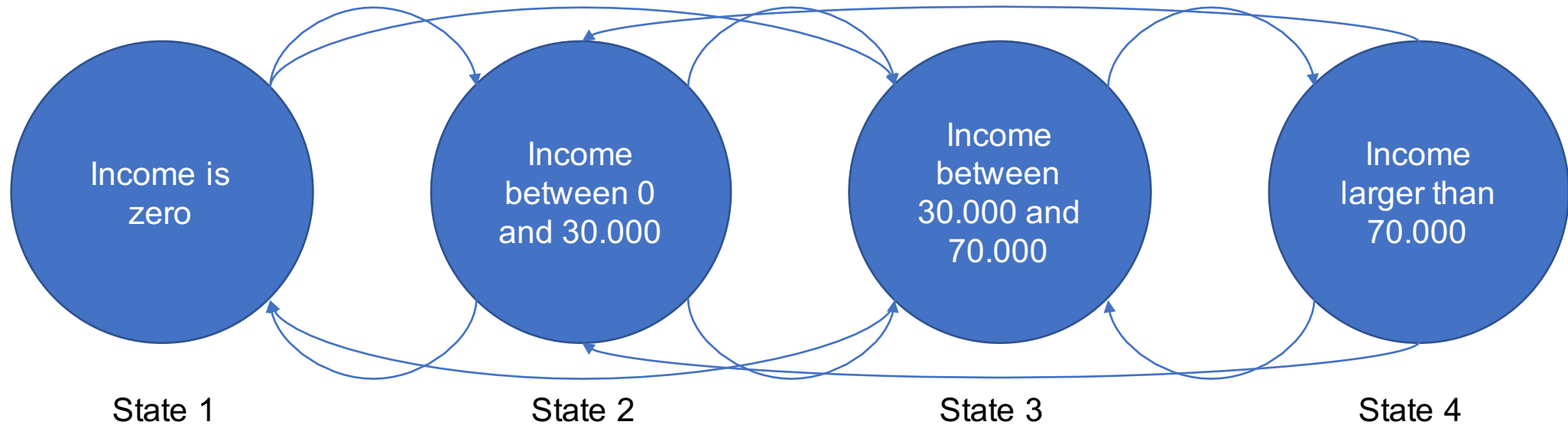


State 3



State 4

It becomes truly interesting when we add time dimensions, since observations can switch states over time.



- We would expect high movement from state 1 to state 2 and little movement from state 1 to state 3 or 4 (unless people are generally very succesful!)
- We would also expect more movement from state 2 towards 3, but similarly we would expect movement back to state 1 at the end of a lifetime.

Finally, adding more data, like age, will enable us to find even more specific states.

An important aspect of CR is that it looks at **all** variables in the model!

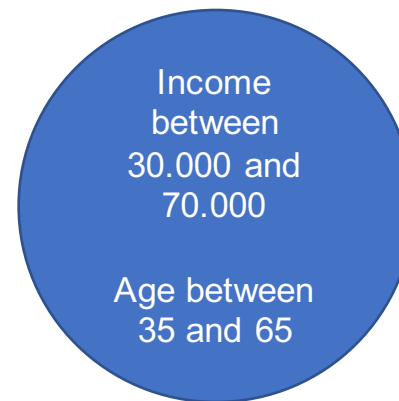


State 1
(Still in school?)

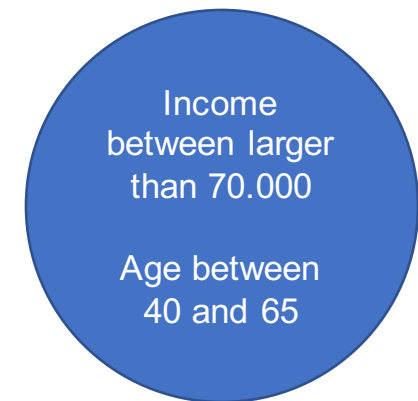


State 2
(Retired?)

...



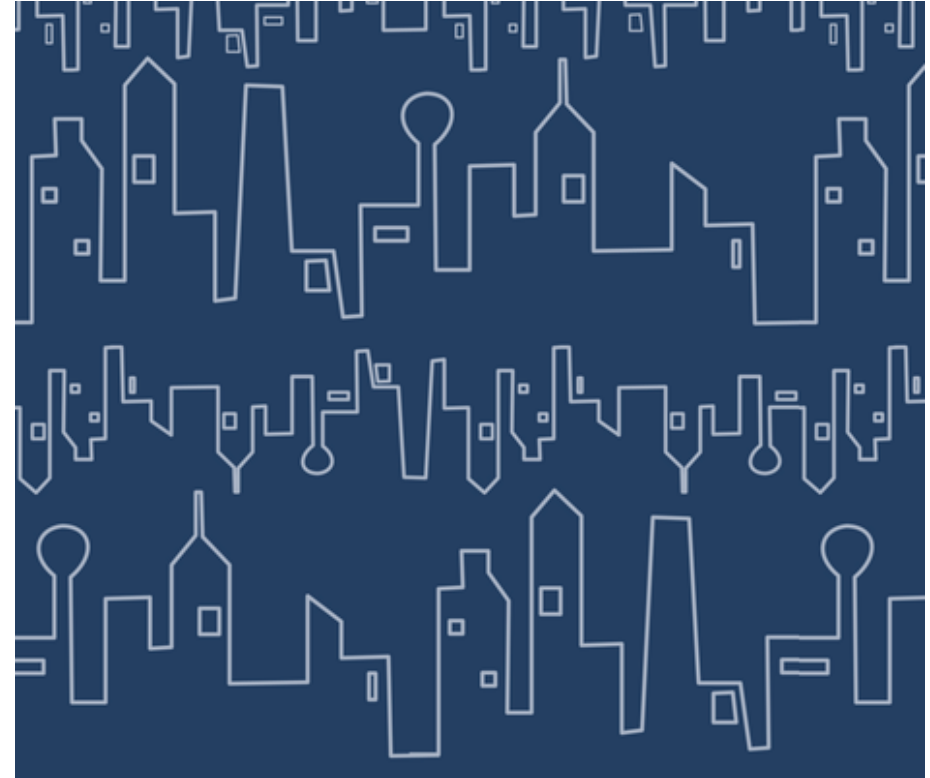
State 15
(Adults?)



State 16
(Executives?)

How does the MHMM compare to descriptive models?

1. A new way of looking at data
2. The Mixed Hidden Markov Model
3. **The MHMM versus descriptive models**



The MHMM enables policymakers to approach **a lot** of data simultaneously without having to rely on indices or aggregation.

Classical Methods

- Aggregates data (losing information)
- Make assumptions about variables (e.g. by making indices)
- Only describes single variables, instead of looking at the entire picture.
- Does not look back or ahead

MHMM

- Uses data as it is generated
- Let's the data **itself** talk and nothing else
- Describes all sorts of variables **jointly**
- Let's us look at the **past**, **present** and **future**