City Rhythm

Model Background & Validation

Wednesday, May 31st, 2017

A short overview of today's presentation

1. Introduction Delph
2. A new way of looking at data
3. The Mixed Hidden Markov Model
   I. The basics
   II. The output: sequences and states
   III. Going to a Mixed HMM
   IV. Under the trunk
   V. Additional examples
4. Questions

Mark Verhagen

Statistical Modeling

Background in Econometrics, Art History

+31 6 81 38 12 37

mark@delph.io

Ruben Spruit

Statistical Modeling

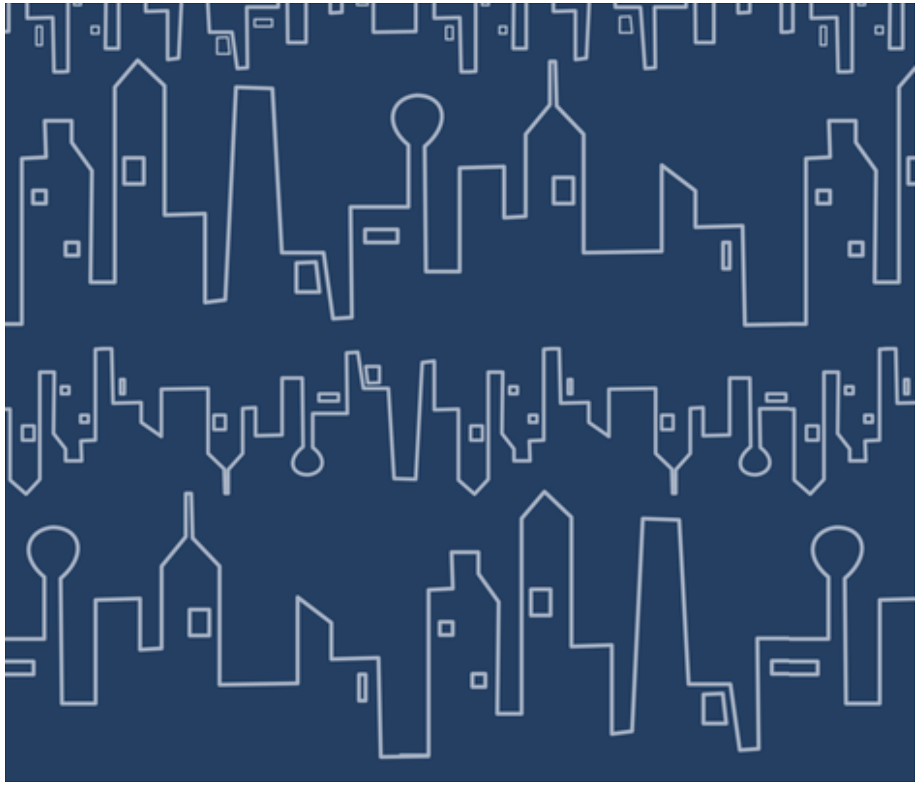
Background in Econometrics, Data Science

+31 6 14 60 50 04

ruben@delph.io

Who are we?

1. **Introduction Delph**
2. A new way of looking at data
3. The Mixed Hidden Markov Model
   I. The basics
   II. The output: sequences and states
   III. Going to a Mixed HMM
   IV. Under the trunk
   V. Additional examples
4. Questions



At Delph we are looking for ways to improve the world **Data**

Optimizing fuel consumption in large ships

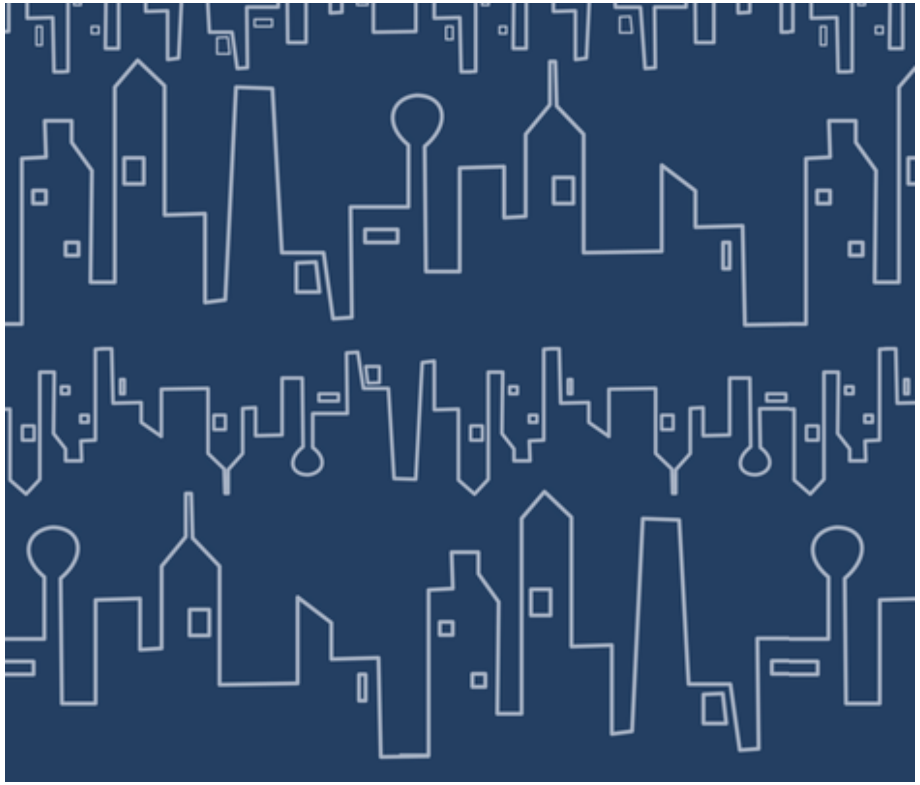
Predicting bridge failures in infrastructure

Helping NGO's reach the right audiences

Help municipalities model their data!

Why is the CR project so interesting?

1. Introduction Delph
2. **A new way of looking at data**
3. The Mixed Hidden Markov Model
   I. The basics
   II. The output: sequences and states
   III. Going to a Mixed HMM
   IV. Under the trunk
   V. Additional examples
4. Questions



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

Movement, food intake, sleep patterns

Interaction between people

Messaging data, GPS movement etc…

Centraal Bureau voor de Statistiek

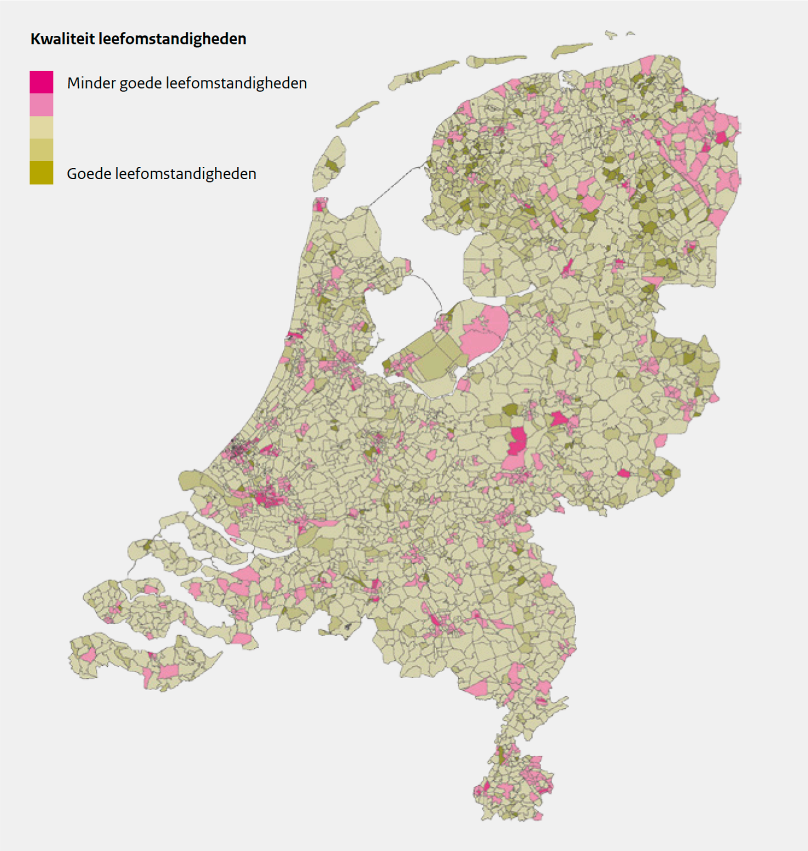
Erosion, nature development

‘Old’ statistics

Traditional data analysis is too static and misses a dynamic and organic perspective.

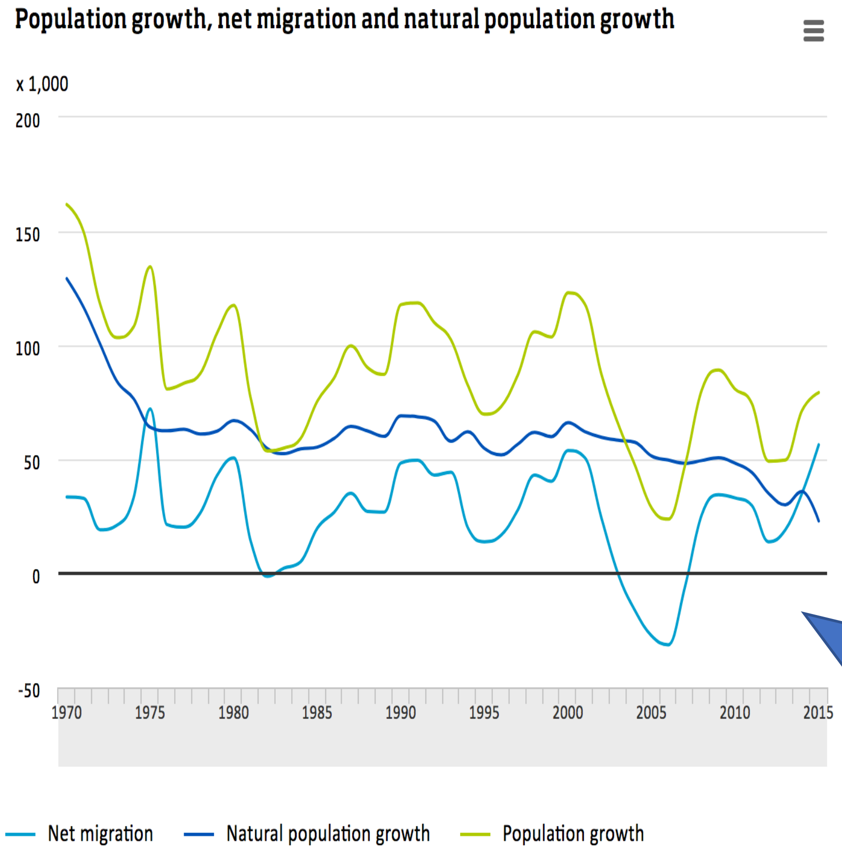
‘Given the available data, this is what your neighborhood was like in 2015…’

- 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?



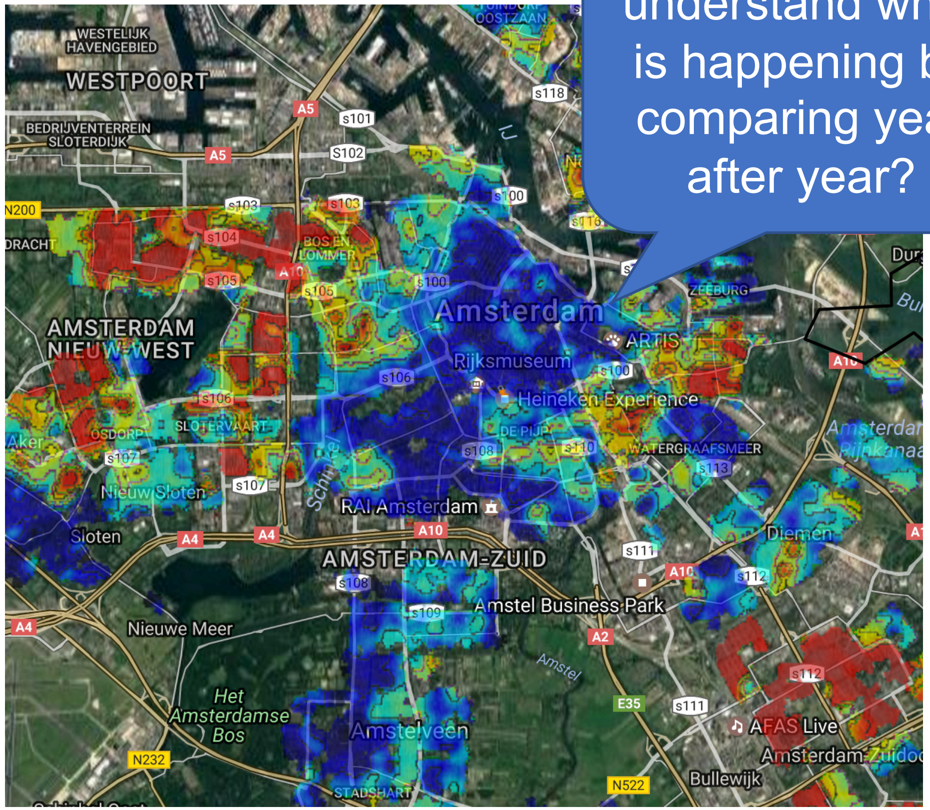
Bron: Tierolf, Gilsing en Steketee, 2017

Ofcourse we can compare one year to another, but we should explicitly model the organic aspects of data!



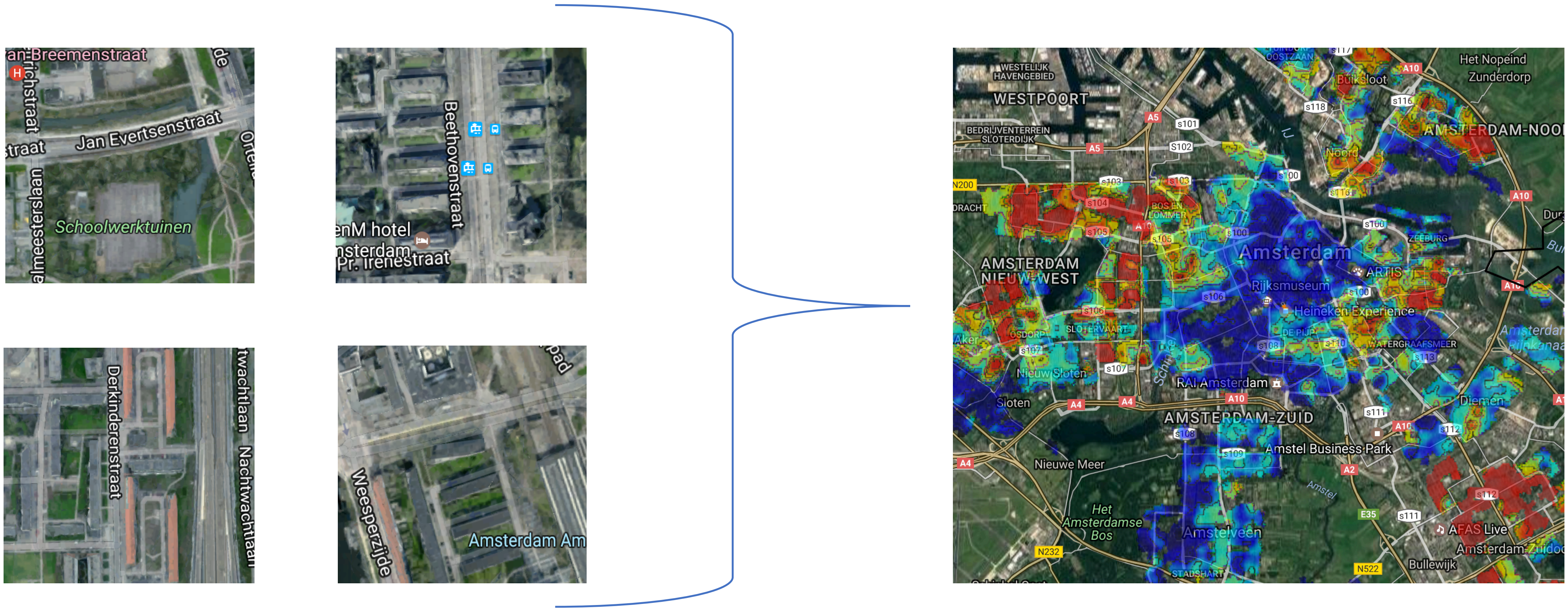
What is **actually** happening?

Are a couple of areas driving these changes or is the entirety of NL changing?

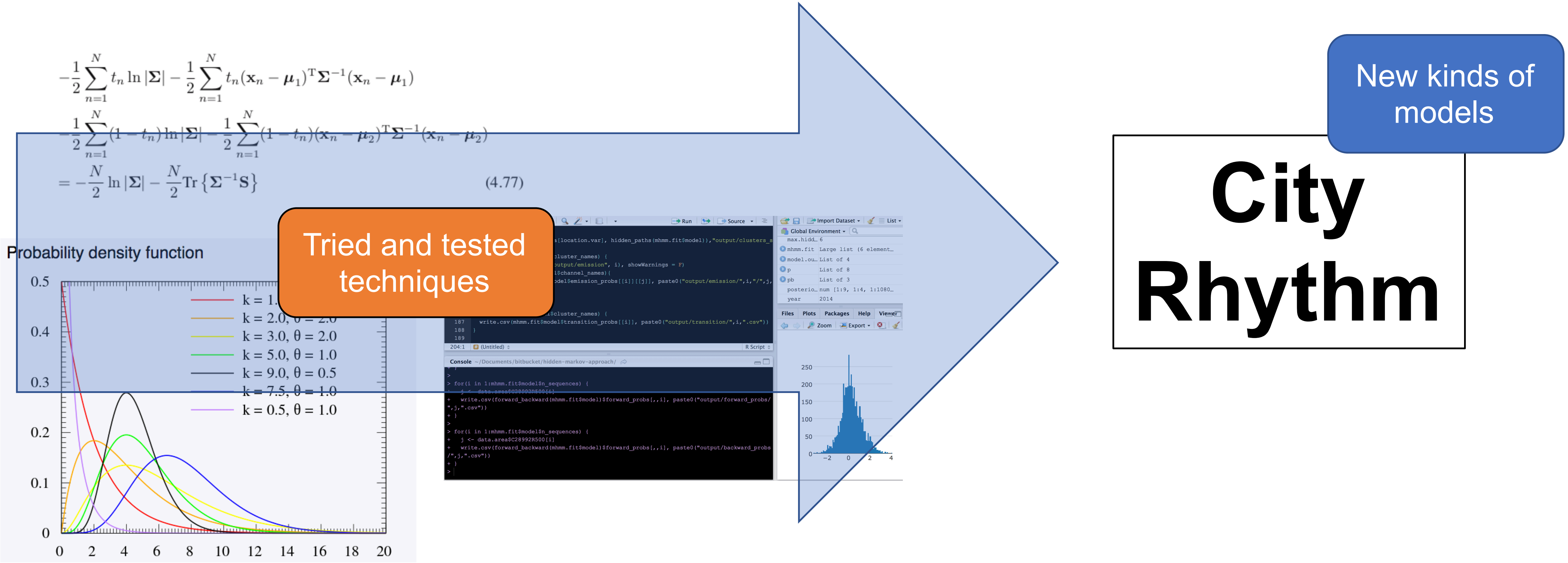


Can we really understand what is happening by comparing year after year?

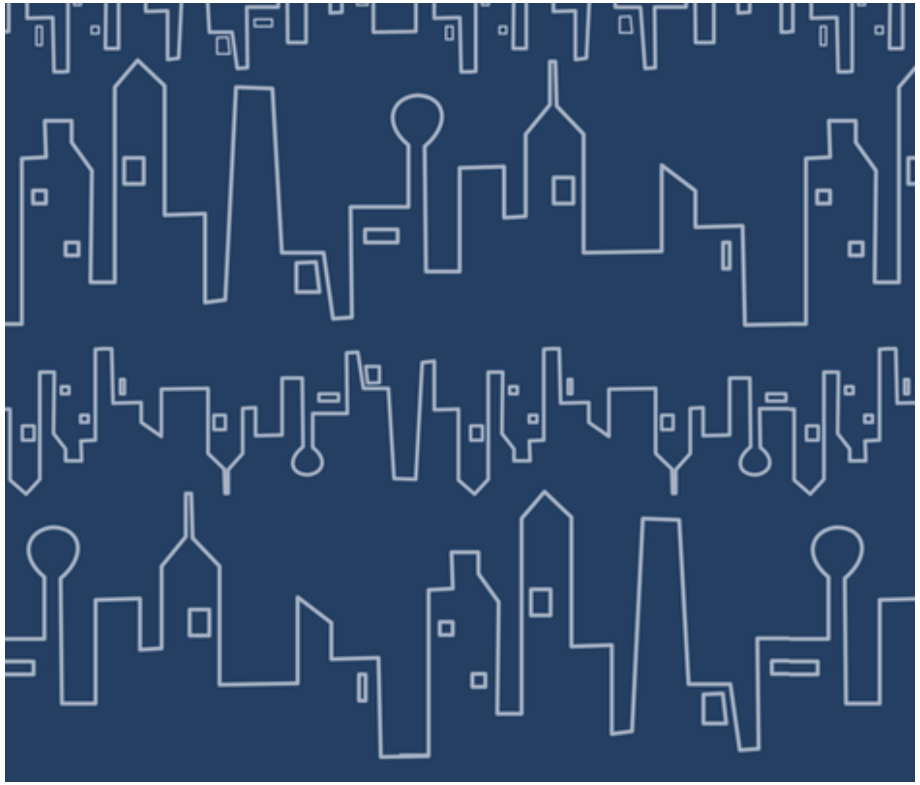
In this project we will instead look at the organic development of low-level data.



This does not mean we have to re-invent the wheel. We can combine classical methods and techniques together with these new ideas!

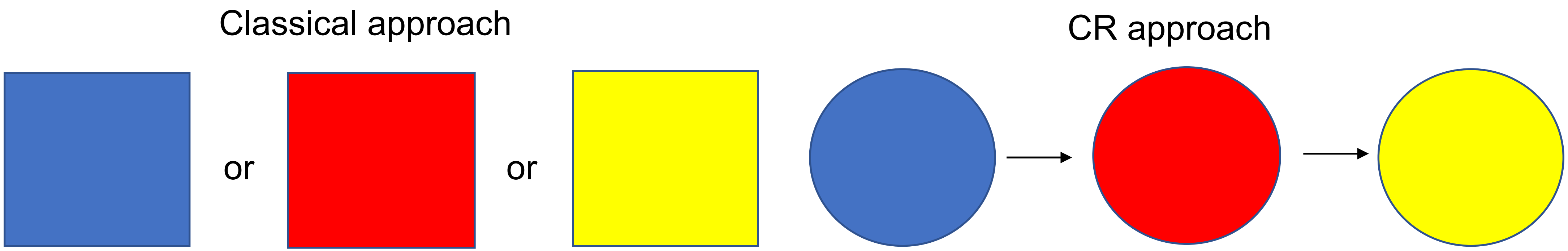


So, what kind of model have we come up with?

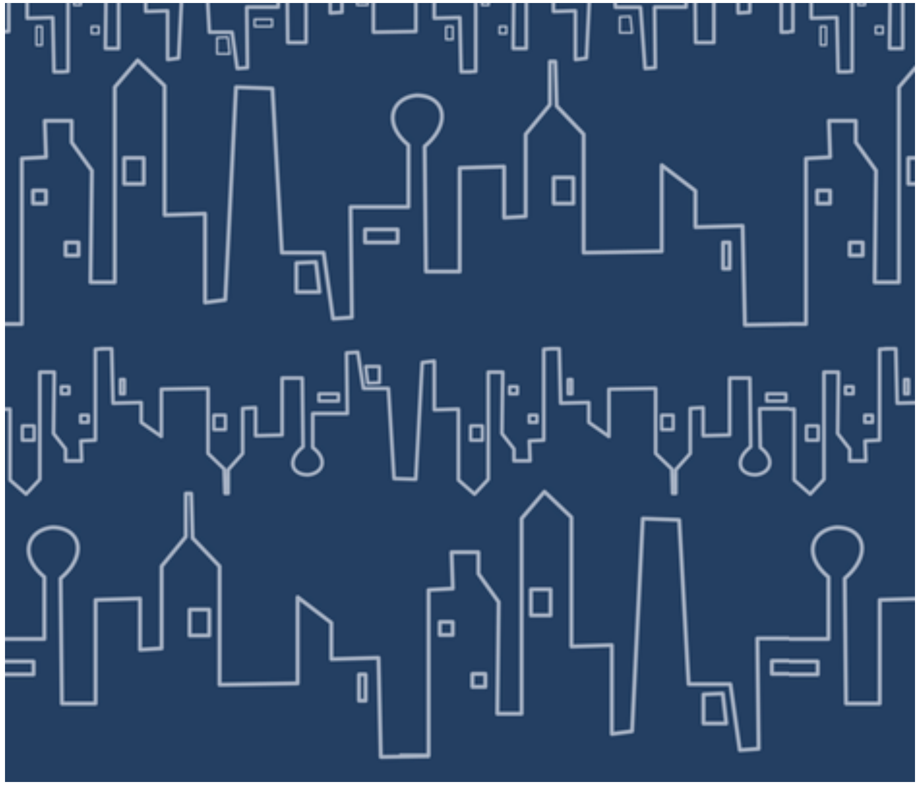


The (Mixed) Hidden Markov Model is designed to capture the **organic** and **dynamic** structure of data.

- 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!



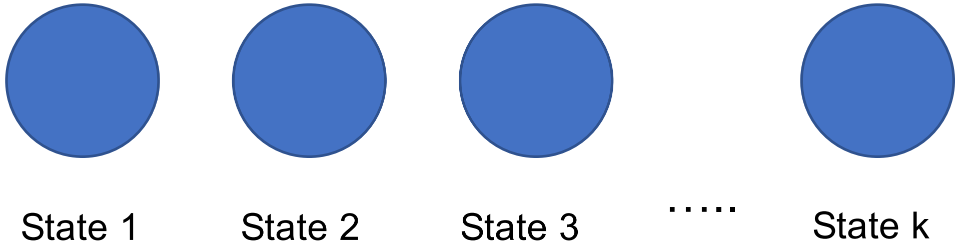
How does the MHMM actually work?



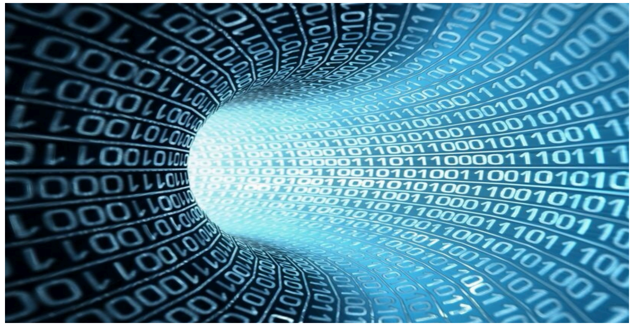
All we need is a number of states: *k* and a dataset.

- 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:

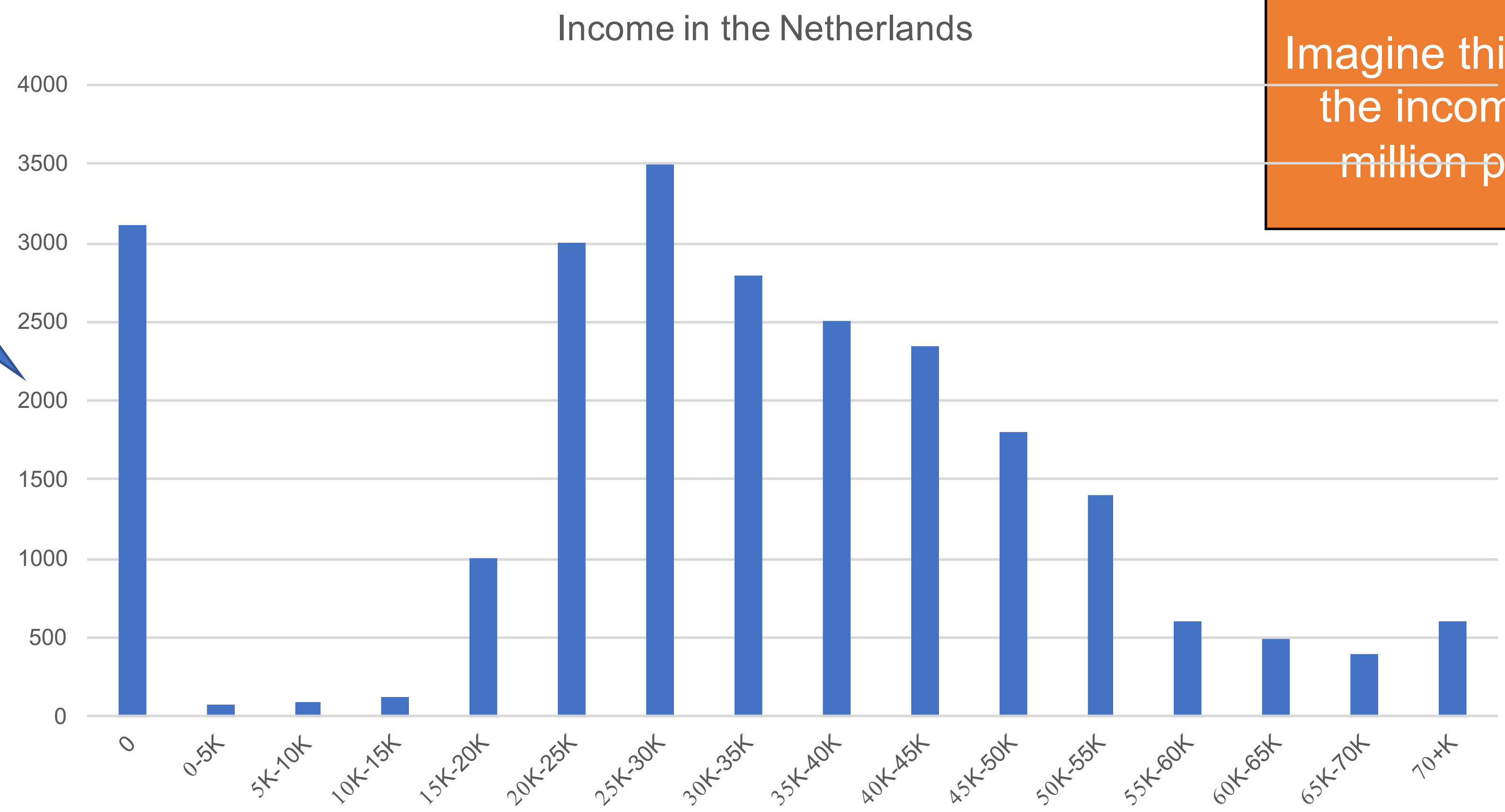


+



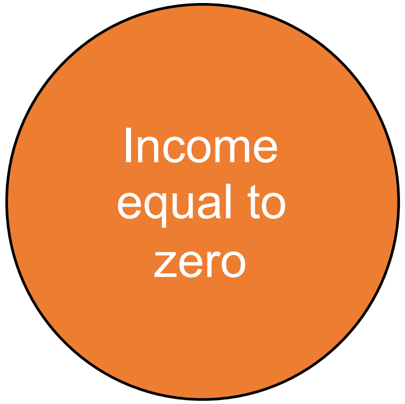
Imagine we have data of the income of everyone in the Netherlands.

What do you think this spike is?

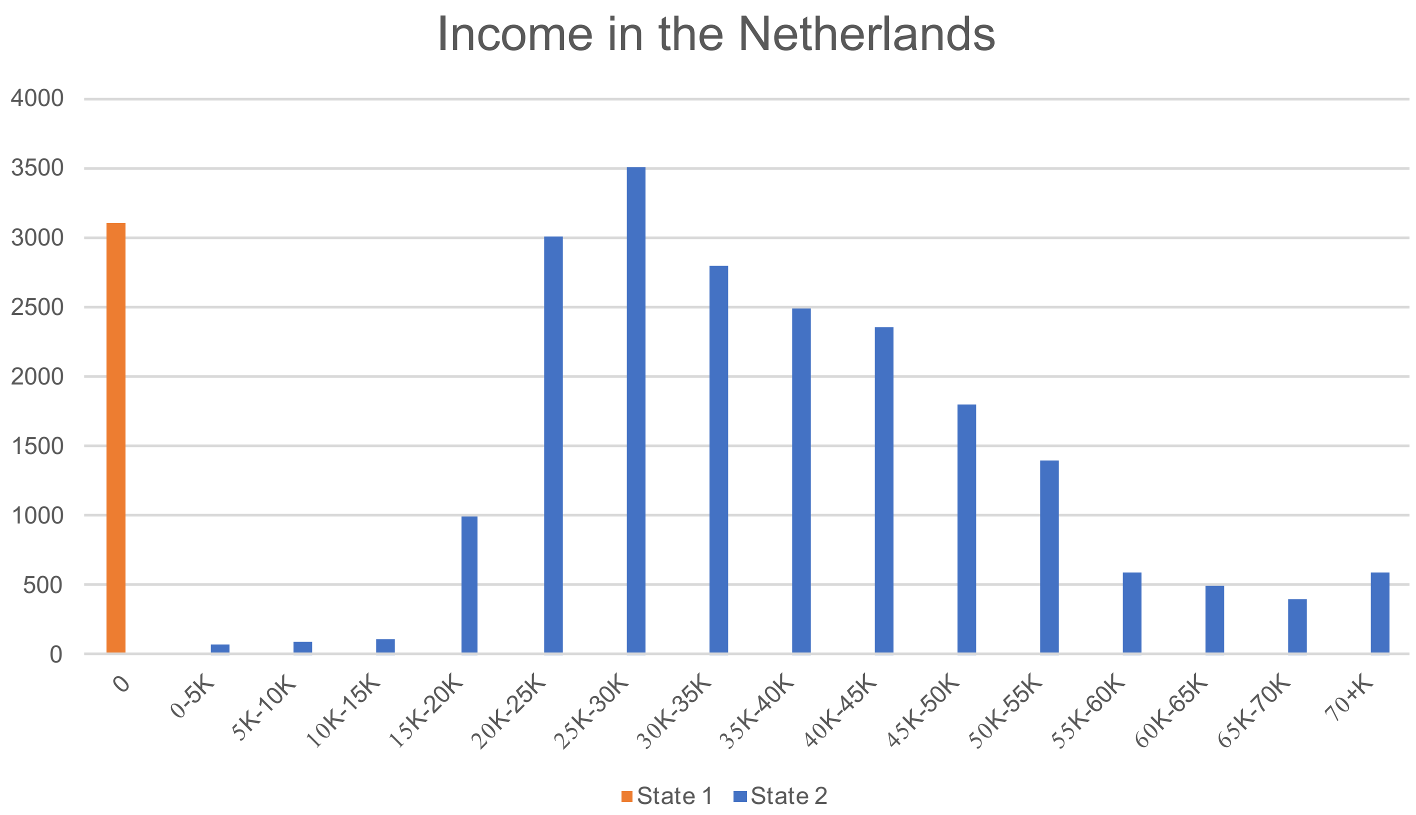


Imagine this is a graph of the income of **all** 16.7 million people in NL

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



State 1

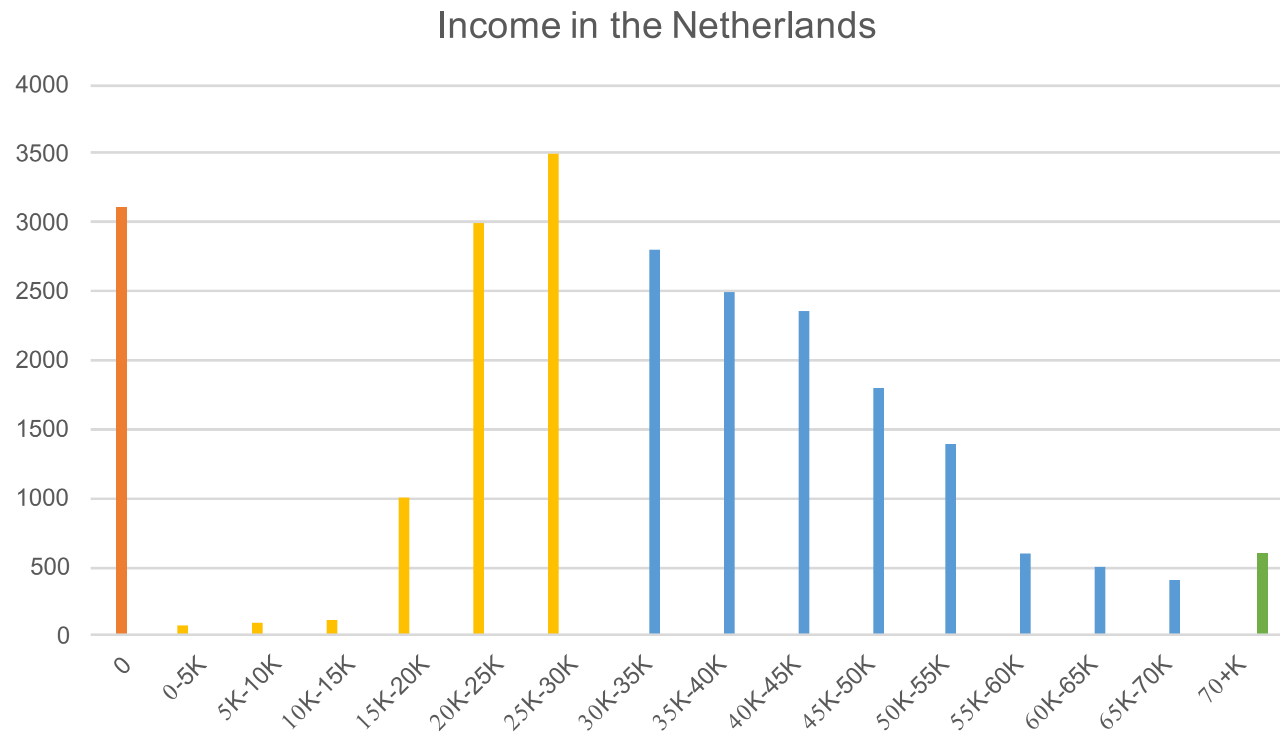


State 2

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

Income is zero

State 1



Income between 30.000 and 70.000

State 3

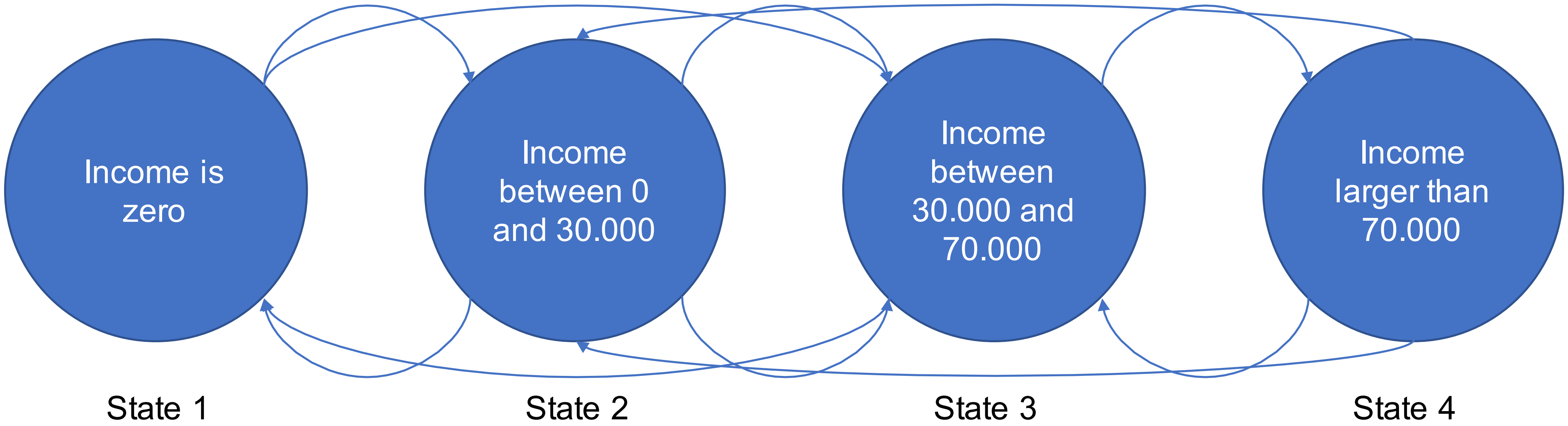
Income until 30.000

State 2

Income larger than 70.000

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.

Income is zero

Age < 20

State 1
(Still in school?)

Income is zero

Age > 65

State 2
(Retired?)

…

Income between 30.000 and 70.000

Age between 35 and 65

State 15
(Adults?)

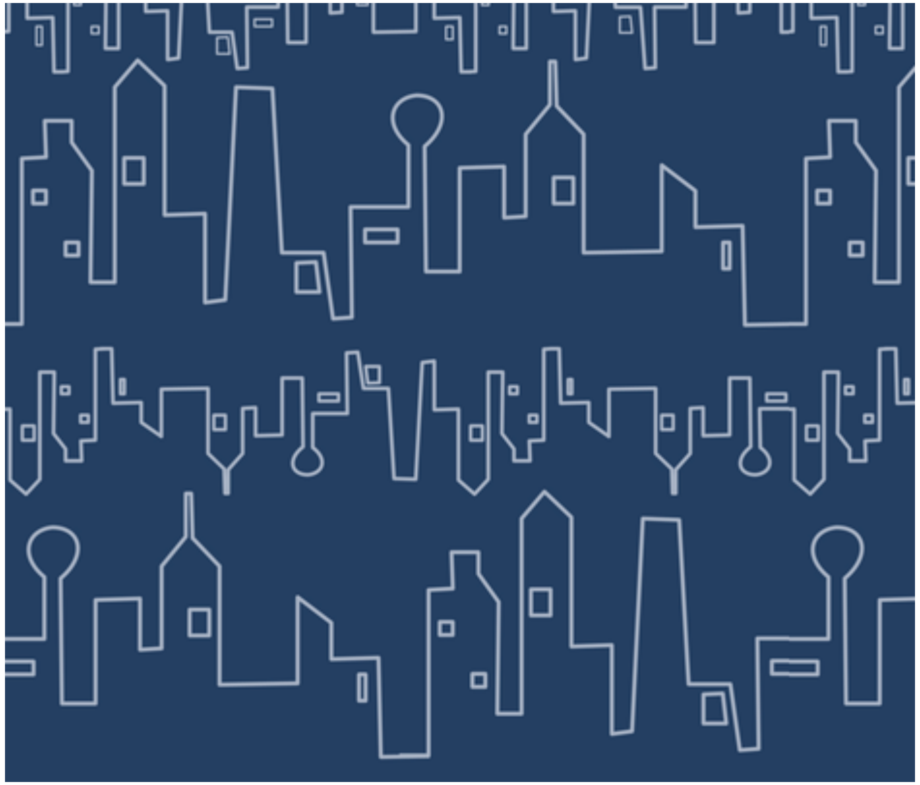
Income between larger than 70.000

Age between 40 and 65

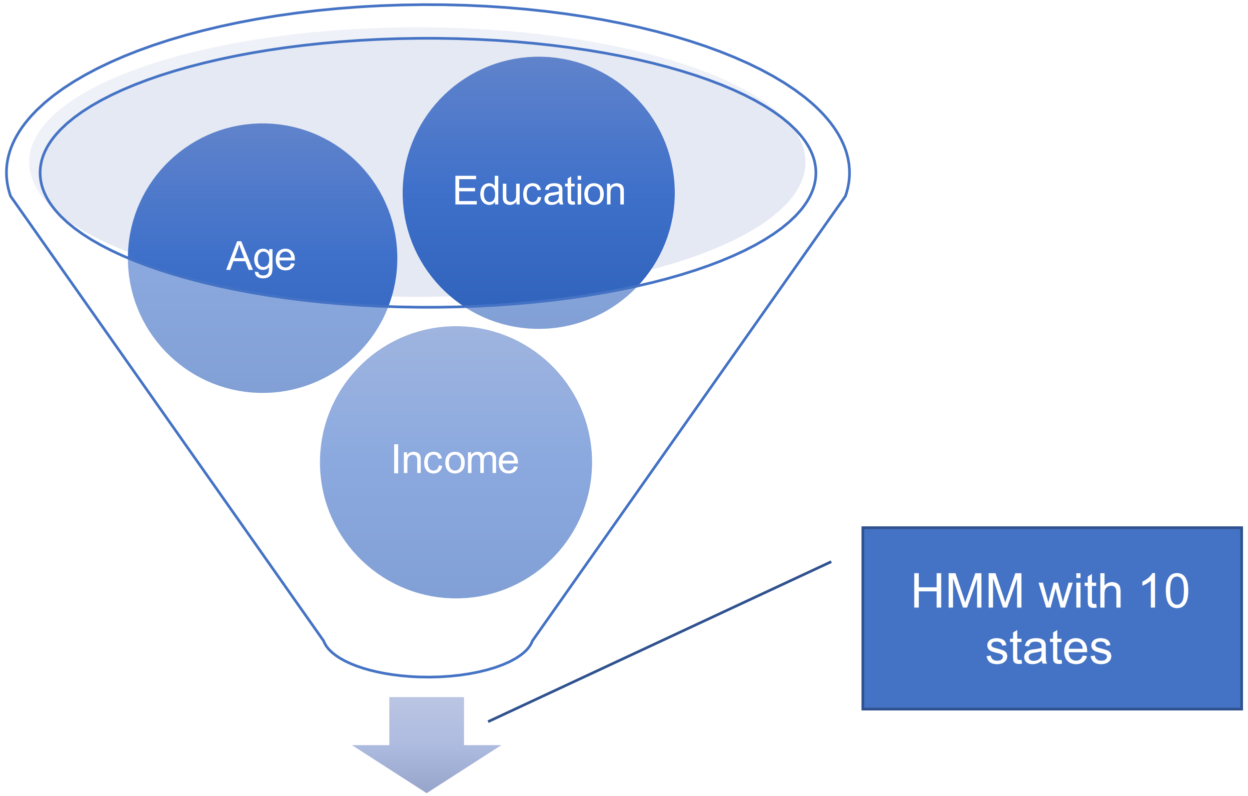
State 16
(Executives?)

What will we actually see?

1. Introduction Delph
2. A new way of looking at data
3. **The Mixed Hidden Markov Model**
   I. The basics
   II. The output: sequences and states
   III. Going to a Mixed HMM
   IV. Under the trunk
   V. Additional examples
4. Questions



Let's assume we have a database over time with individual's age, income and education level.



What will we see?

On the state level, the following could be three of the states the model has created

State 1 generates:
- Low incomes
- Low age
- Few years of education

State 2 generates:
- Low incomes
- High age
- All sorts of education levels

State 3 generates:
- Income between 30- 50K
- Age between 20-30
- Many years of education

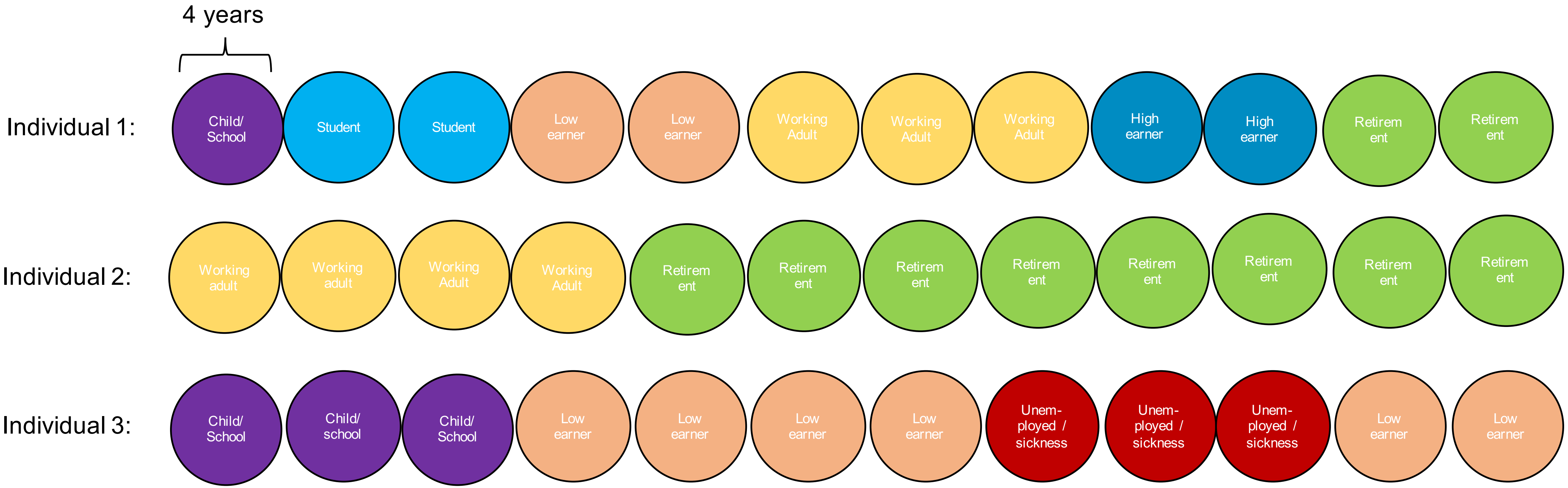
State 1
(Probably still in school?)

State 2
(Probably retired?)

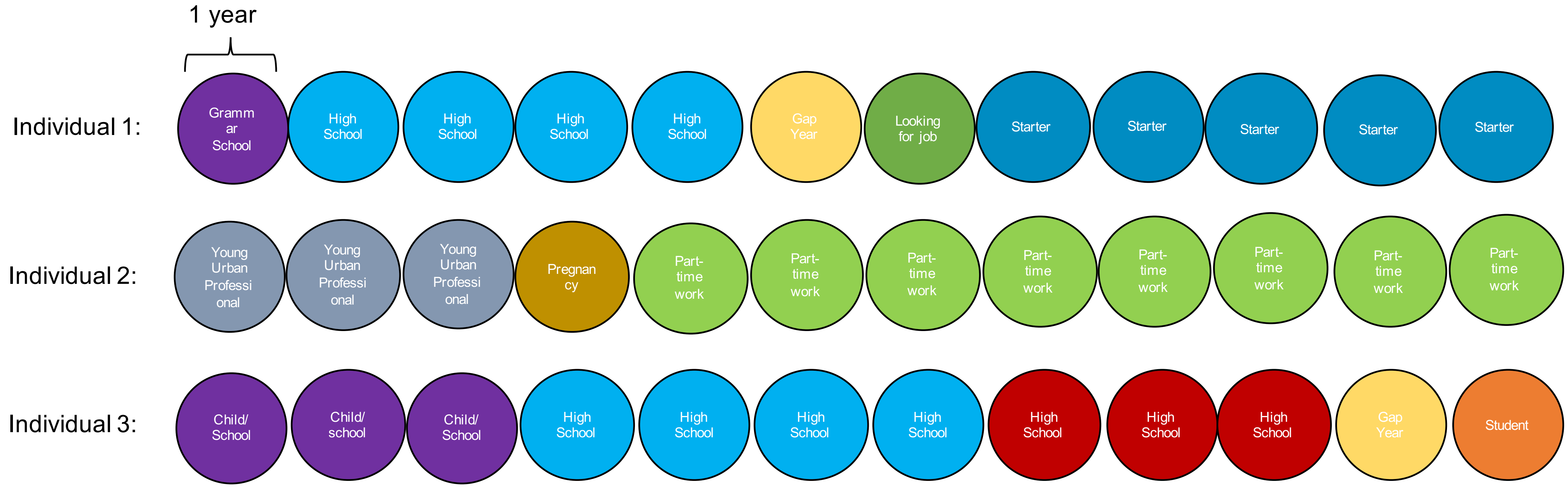
State 3
(Probably highly educated starter?)

For these interpretations, we need expert knowledge from policy-makers and civil servants!

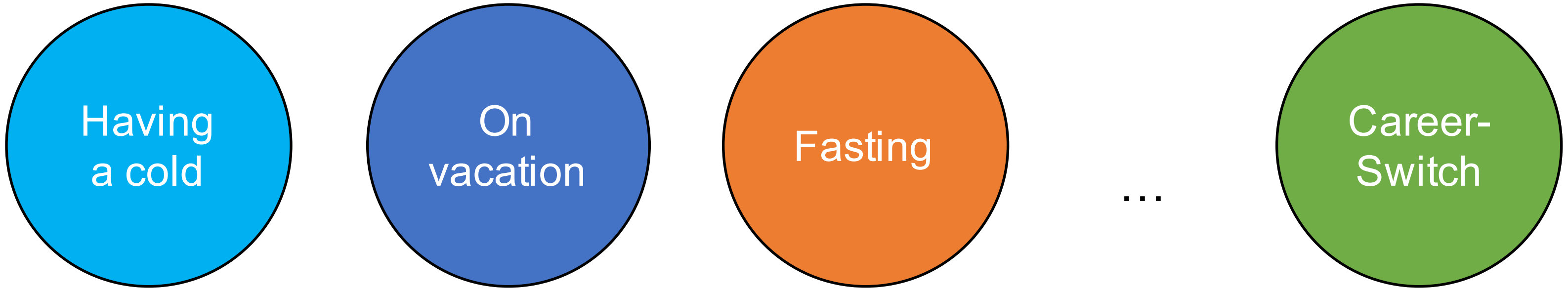
On the observation level, we will see the sequence which best describes the individual observation.



As we add more and more data on a lower level, the insights we extract become more interesting…



As the amount of data increases, the model can track individuals on the monthly, daily or even hourly level!



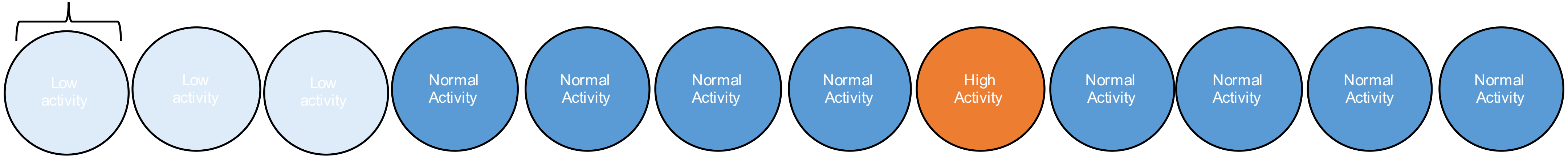
Let's do another example of a HMM with data on movement of individuals in a city.

Babies? The sick? Athletic people? Elderly people?

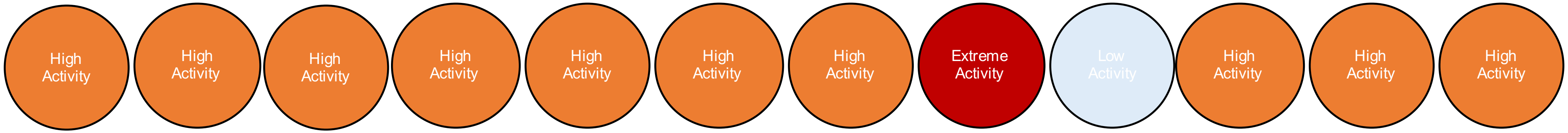
What would happen if we had only activity data, but monthly?

January

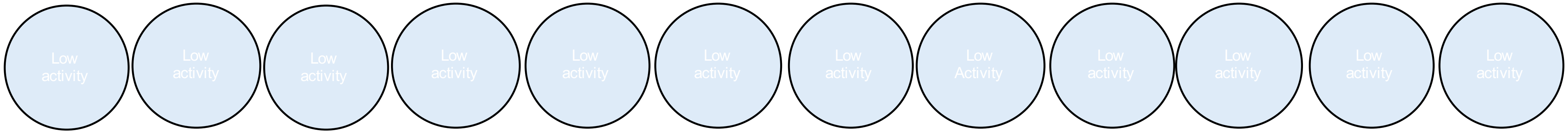
Person that goes on active vacation in the summer?!



Professional athlete training for a marathon?



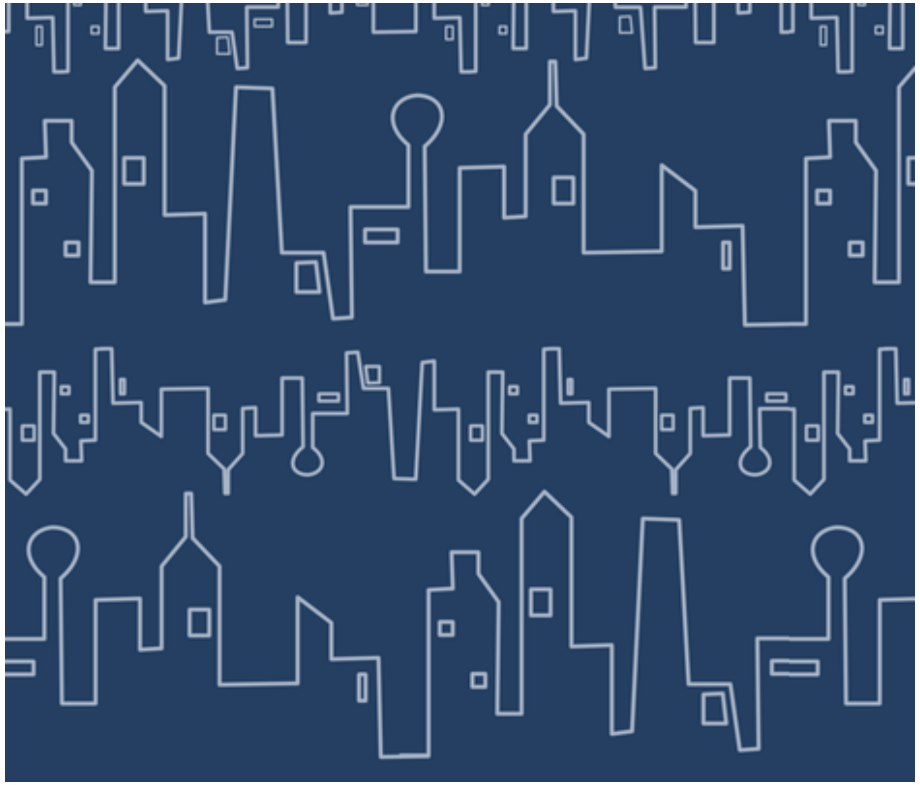
Eldery? Sick?



If we had hourly data, it would enable us to distinguish even more specific types of individuals..

- Young children who only move during the day.
- Workaholics that move steadily but very little throughout the day and parts of the night.
- People that run occasionally.
- The elderly.
- Clubbing youth.
- Professional athletes.
- Gamers.

There are several interesting extensions to the model.



When datasets become very big, the mixed extension will keep the model tractable.

- Imagine we have the same data as before, but over the entirety of Europe! Probably, a farmer in Italy won't be the same as one in the Netherlands, neither will a young professional be the same in Easter-Europe and West-Europe.

Young professional in East-Europa

State 1

Farmer in Italy

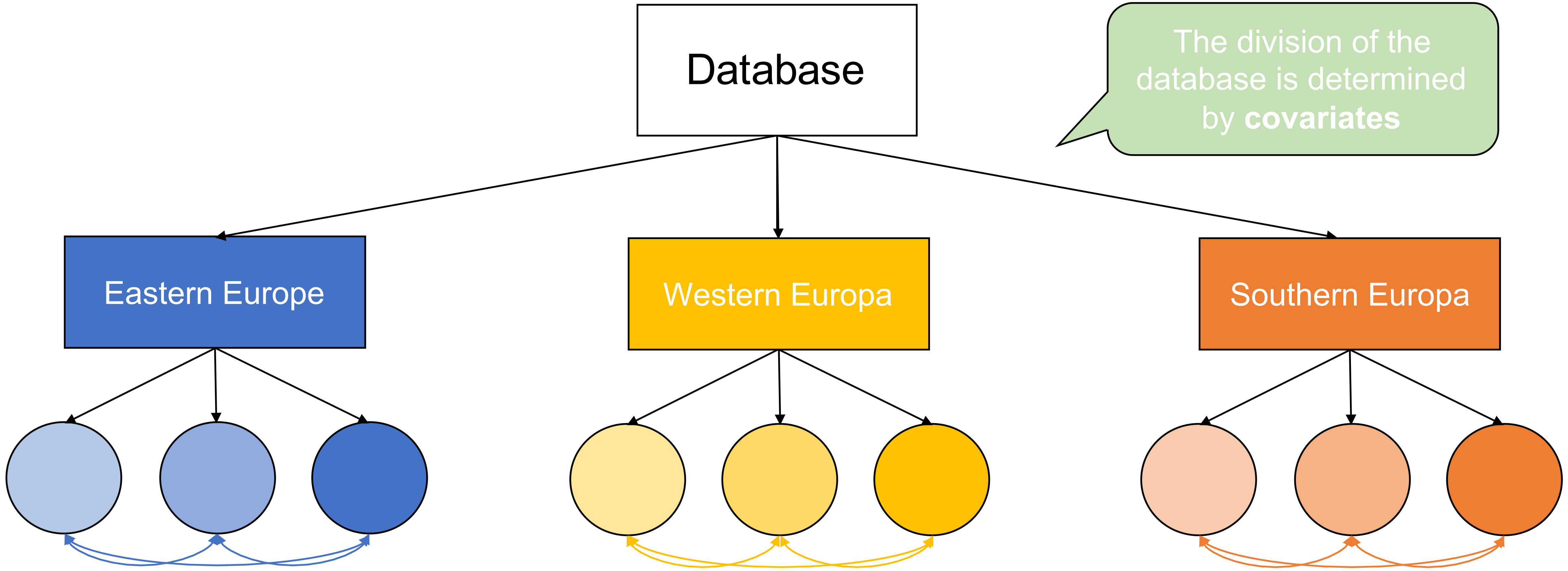
State 245

We will need enormous amounts of states to find interesting results!

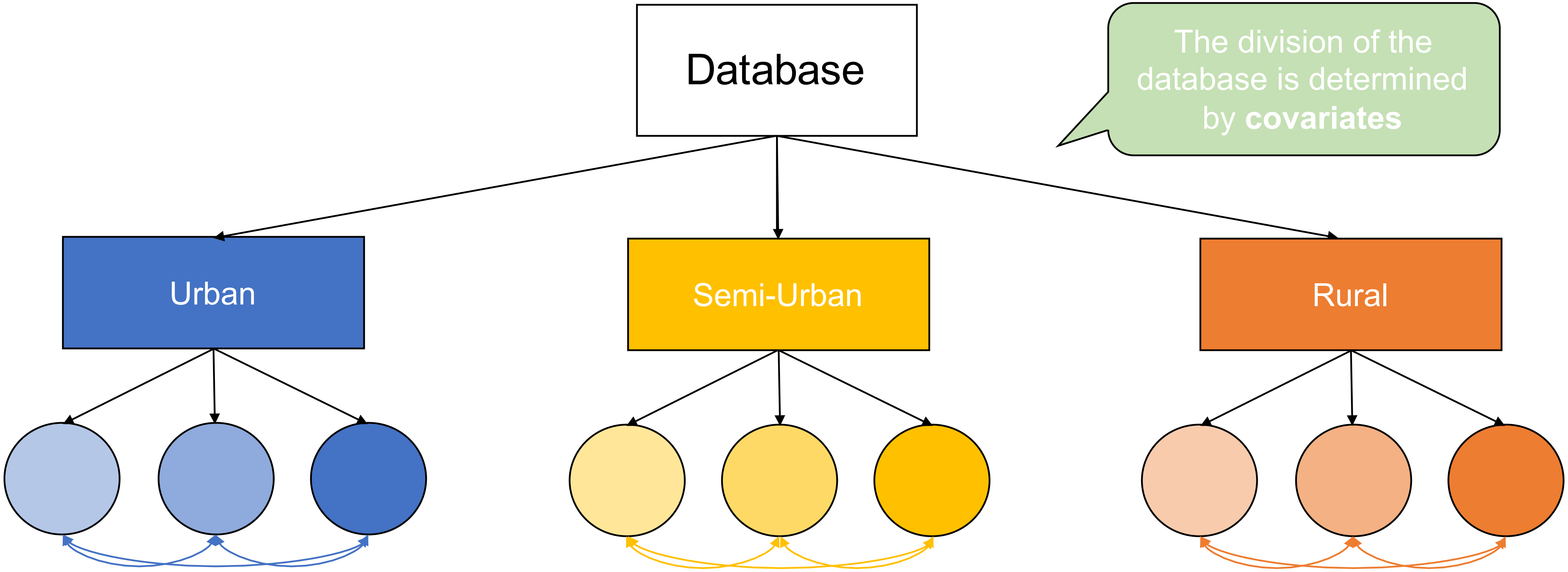
Child in Amsterdam

State 3495

Instead, we can divide the database in sub-parts, like geographic location.



Or by demographic characteristics…



In the end, the (M)HMM goes through the following steps.

Choose the number of HMM's in the model.

↓

Choose the number of states in every HMM.

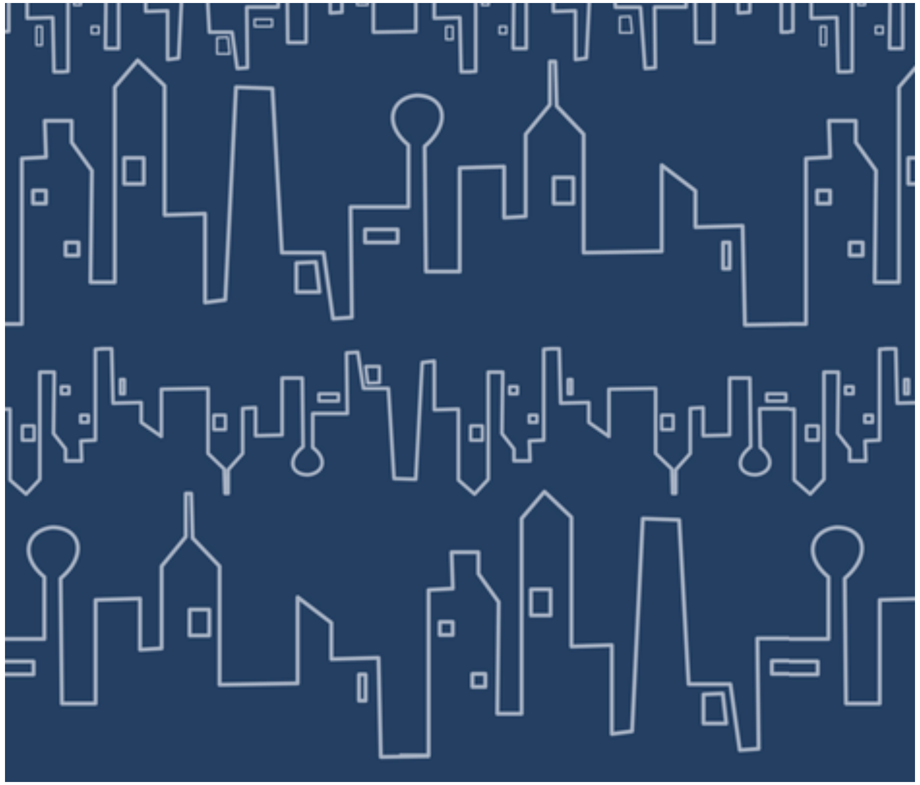
↓

Choose covariates that determine to which HMM an observation belongs

↓

Run the model

So, what happens under the trunk?



What happens under the trunk…?

$$\log L = \sum_{i=1}^{N} \log P(Y_i|\mathcal{M})$$
$$= \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} P(\mathcal{M}^k) \sum_{\text{all } z} P\left(Y_i|z,\mathcal{M}^k\right) P\left(z|\mathcal{M}^k\right) \right]$$
$$= \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} w_k \sum_{\text{all } z} \pi^k_{z_1} b^k_{z_1}(y_{i11}) \cdots b^k_{z_1}(y_{i1C}) \prod_{t=2}^{T} \left[a^k_{z_{t-1}z_t} b^k_{z_t}(y_{it1}) \cdots b^k_{z_t}(y_{itC}) \right] \right].$$

The HMM is expresed as a joint probability that is maximised over all the available parameters

$$P(\mathcal{M}^k|Y_i,\mathbf{x}_i) = \frac{P(Y_i|\mathcal{M}^k,\mathbf{x}_i)P(\mathcal{M}^k|\mathbf{x}_i)}{P(Y_i|\mathbf{x}_i)}$$
$$= \frac{P(Y_i|\mathcal{M}^k,\mathbf{x}_i)P(\mathcal{M}^k|\mathbf{x}_i)}{\sum_{j=1}^{K} P(Y_i|\mathcal{M}^j,\mathbf{x}_i)P(\mathcal{M}^j|\mathbf{x}_i)} = \frac{L^i_k}{L^i},$$

```
mhmm.fit <- fit_model(init.mhmm)

###############
# MODEL OUTPUT #
###############
# interesting model output
summary(mhmm.fit$model)
mhmm.fit$model$transition_probs
hidden_paths(mhmm.fit$model)
posterior_probs <- posterior_probs(mhmm.fit$model)
forward_probs <- forward_backward(mhmm.fit$model)$forward_probs
backward_probs <- forward_backward(mhmm.fit$model)$backward_probs

# Fetch the most probable clusters for a specific grid and year
data.a        ind(data.area, hidden_paths(mhmm.fit$model))
```

The observations are assigned to the various HMM's through fitting a ordered probabilistic model

The entire model is optimised using numerical methods and computational power.

How do we determine the optimal amount of HMM's and states in the model?

$$AIC = -2logL + 2q \quad (1)$$

$$BIC = 2log(L) + qlog(N) \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}\left(dbh_{ij} - \widehat{dbh}_{ij}\right)^2}{\sum_{i=1}^{N} n_i - p}} \quad (3)$$

$$Absolute\ Bias = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}\left|dbh_{ij} - \widehat{dbh}_{ij}\right|}{\sum_{i=1}^{N} n_i} \quad (4)$$

Information criteria give insights in model fit.

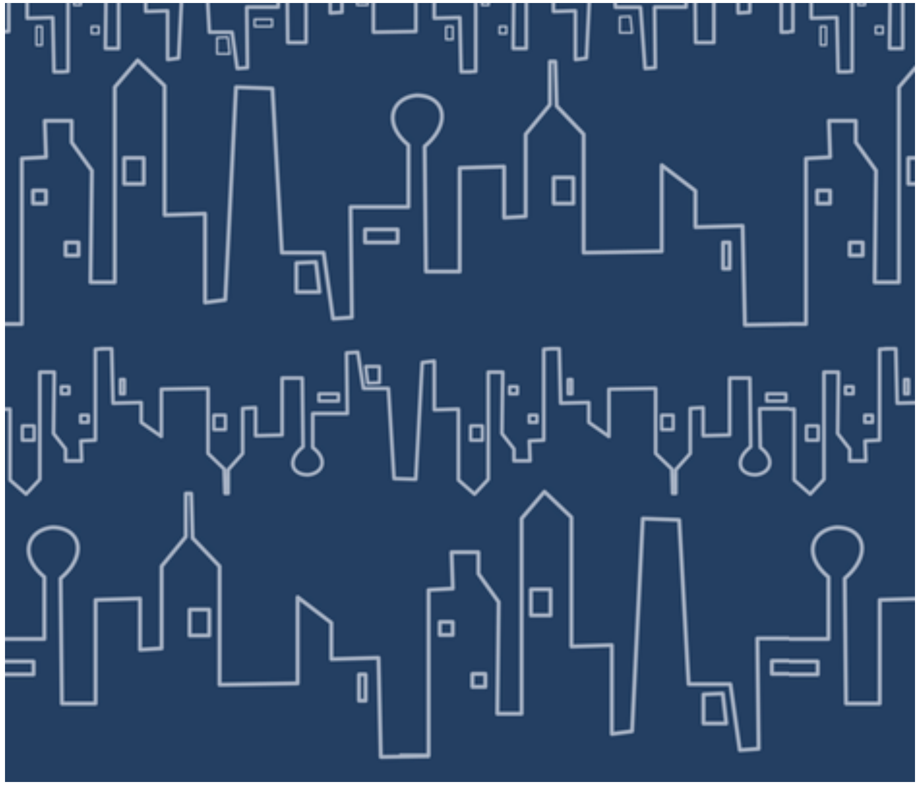
```
close all;
rng(seed);
nmc = 10;                                          % Number of MC replications
nlim = 1000;                                       % Maximum sample size
step = 10;                                         % Stepsize of n increase
res_MC_normal_05 = zeros(nlim / step , 2);            % Results for different n (standard)
res_MC_normal = zeros(nlim / step , 2);            % Results for different n (standard)
res_MC_normal_2 = zeros(nlim / step , 2);            % Results for different n (standard)
res_MC_silver = zeros(nlim / step , 2);            % Results for different n (Silverman)
MC_MSE_normal_05 = zeros(nmc, 2);                     % MC results per n (standard)
MC_MSE_normal = zeros(nmc, 2);                     % MC results per n (standard)
MC_MSE_normal_2 = zeros(nmc, 2);                     % MC results per n (standard)
MC_MSE_silver = zeros(nmc, 2);                     % MC results per n (Silverman)
figure_vec = [10, 50, 100, 250, 500, 750, 1000];% Make figures for n

% calculate density for the total database
[full_points, full_normal, full_silver, bwh] = npdensity_ahmv(Y_log, sub1mid, 0);
```

But policy-makers are the most important source of information!

Brute force computational power can evaluate all possible combinations of states and HMM's

Let's do a couple of examples

1. Introduction Delph
2. A new way of looking at data
3. **The Mixed Hidden Markov Model**
   I. The basics
   II. The output: sequences and states
   III. Going to a Mixed HMM
   IV. Under the trunk
   V. Additional examples
4. Questions



A HMM with data on movement of individuals in a city

Number of steps per year

Babies? The sick?

Almost no steps

'Normal' people?

Many steps

…

Elderly people?

Few steps

Not very insightful yet…

Number of steps per year + age

Babies

Few Steps + young age

Elderly

Few Steps + old age

…

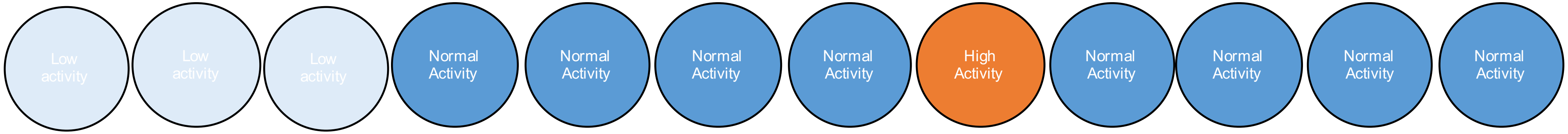
Runners? Athletes?

Extreme amount of steps, middle age

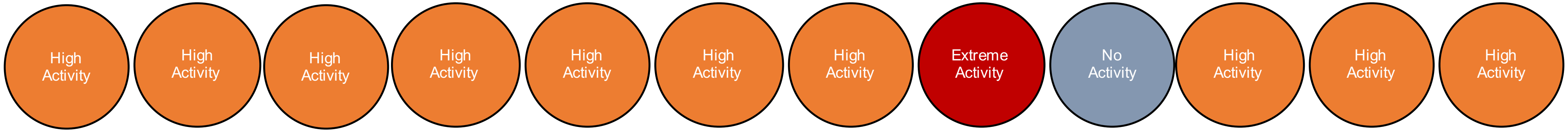
We can try to interpret the states better already…

What would happen if we had only activity data, but monthly?

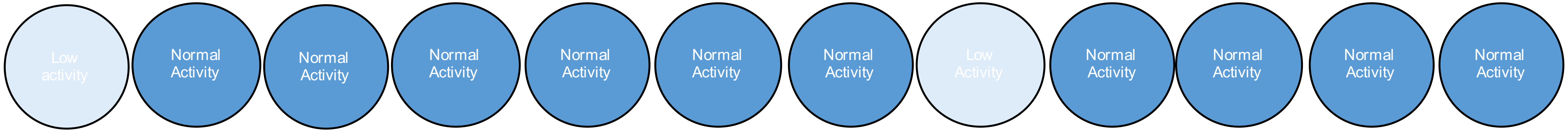
Person that goes on active vacation in the summer?!



Person training for a biking vacation?!



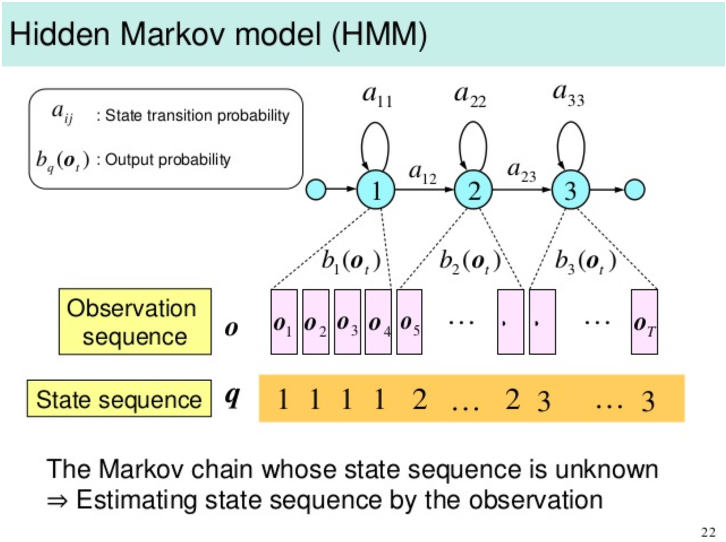
‘Normal’ person that goes on a beach vacation?!



The Mixed Hidden Markov Model has proven its value in several applications

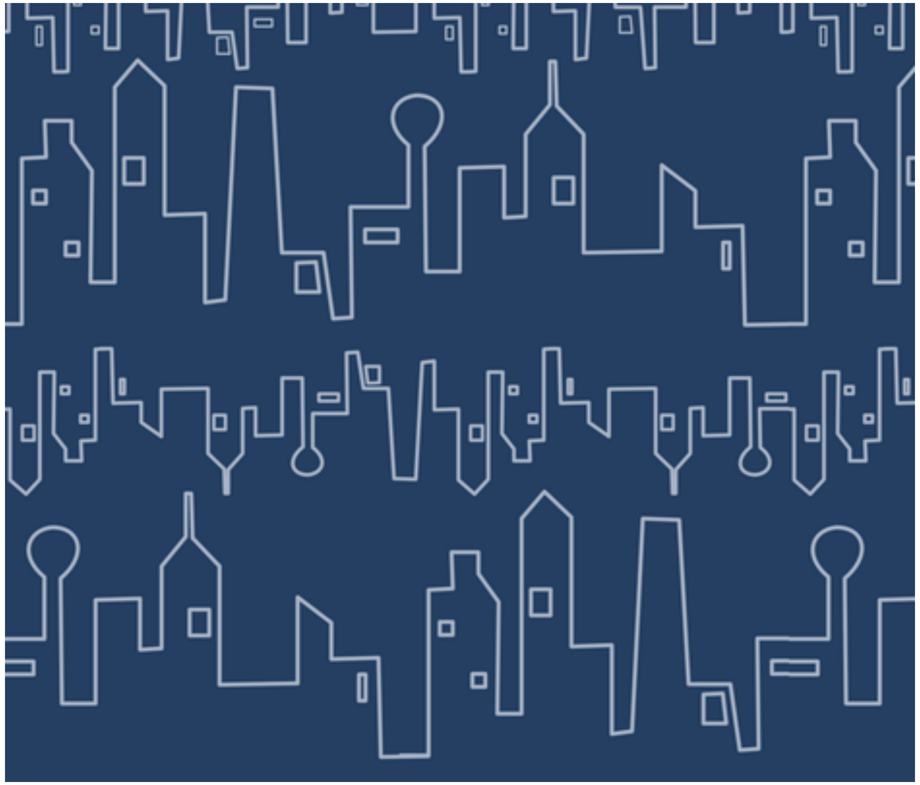
Academic applications of the MHMM:

- Korkmazkiy et al. (1997) speech recognition of diverse speakers
- Ypma and Heskes (2003) joint classification of web pages and user type
- Costa (2009) personalized diagnosis and treatment to medicine
- Darmanjian and Principe (2009) neural signals and kinematics
- Langrock et al. (2012) modeling of animal telemetry data
- Subakan et al. (2014) motion capture data, network traffic data



There are undoubtedly still many questions…

1. Introduction Delph
2. A new way of looking at data
3. The Mixed Hidden Markov Model
   I. The basics
   II. The output: sequences and states
   III. Going to a Mixed HMM
   IV. Under the trunk
   V. Aditional examples
4. **Questions**



Ask them, please!