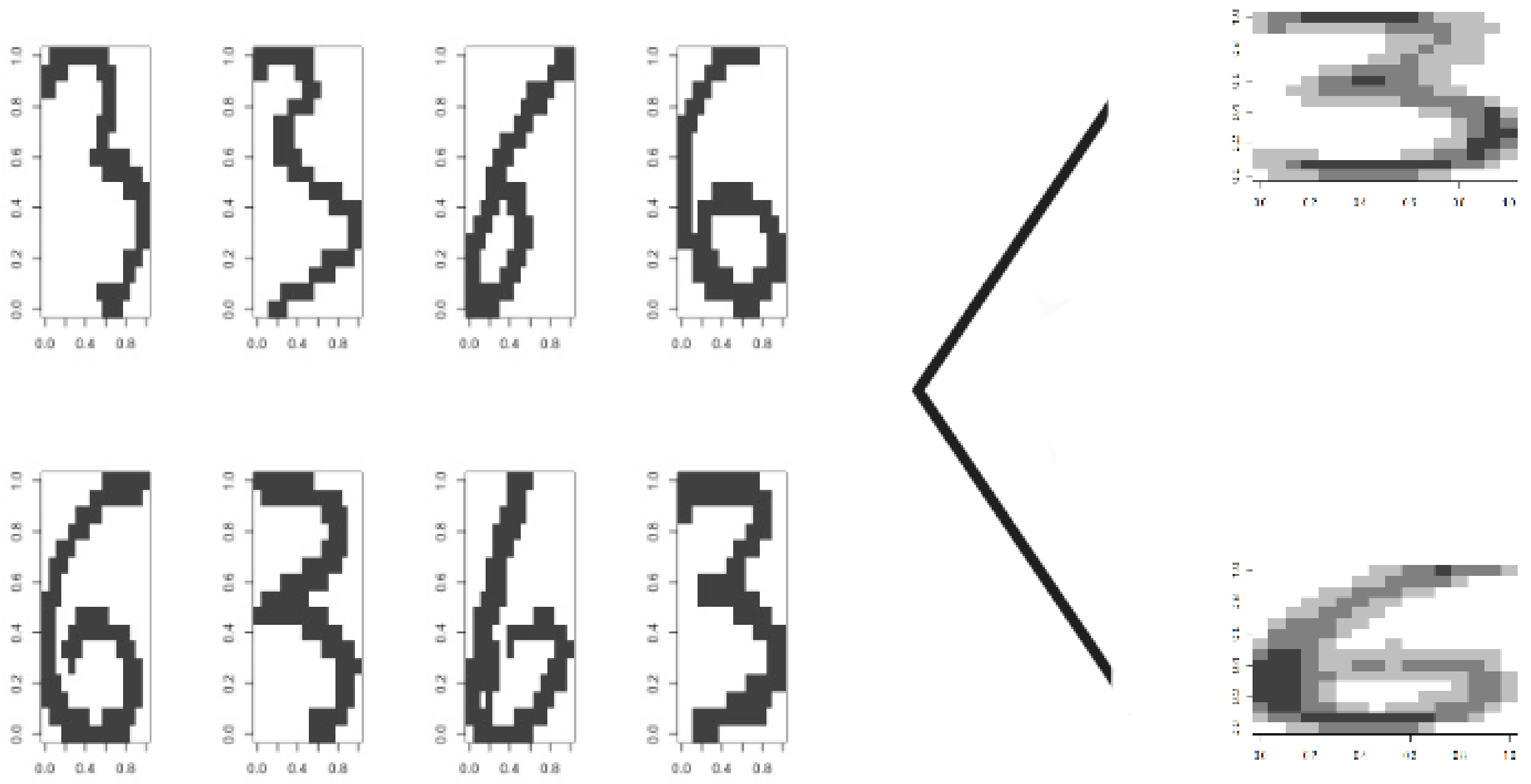
MIXTURE MODELS IN R

Bernoulli Mixture Models

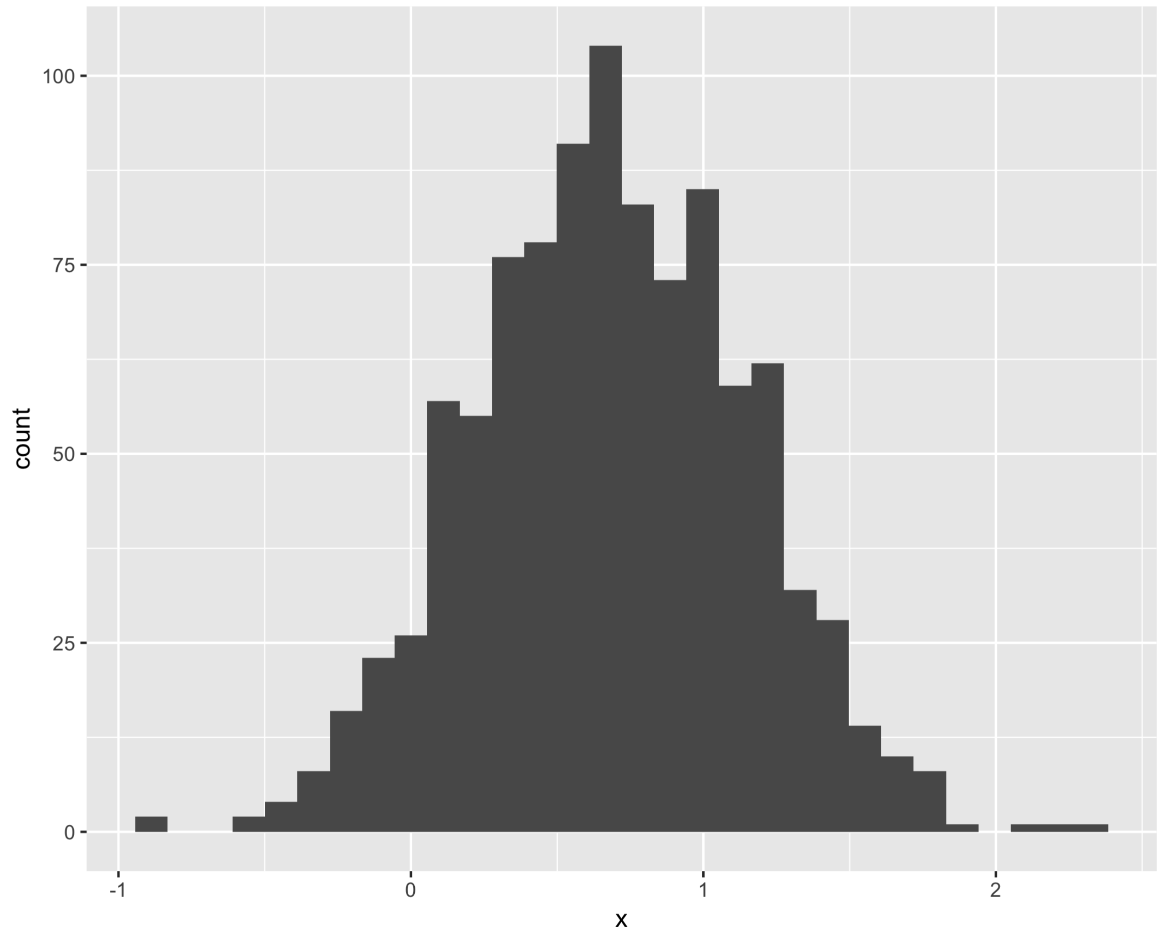
Victor Medina
Researcher at SBIF

The handwritten digits dataset

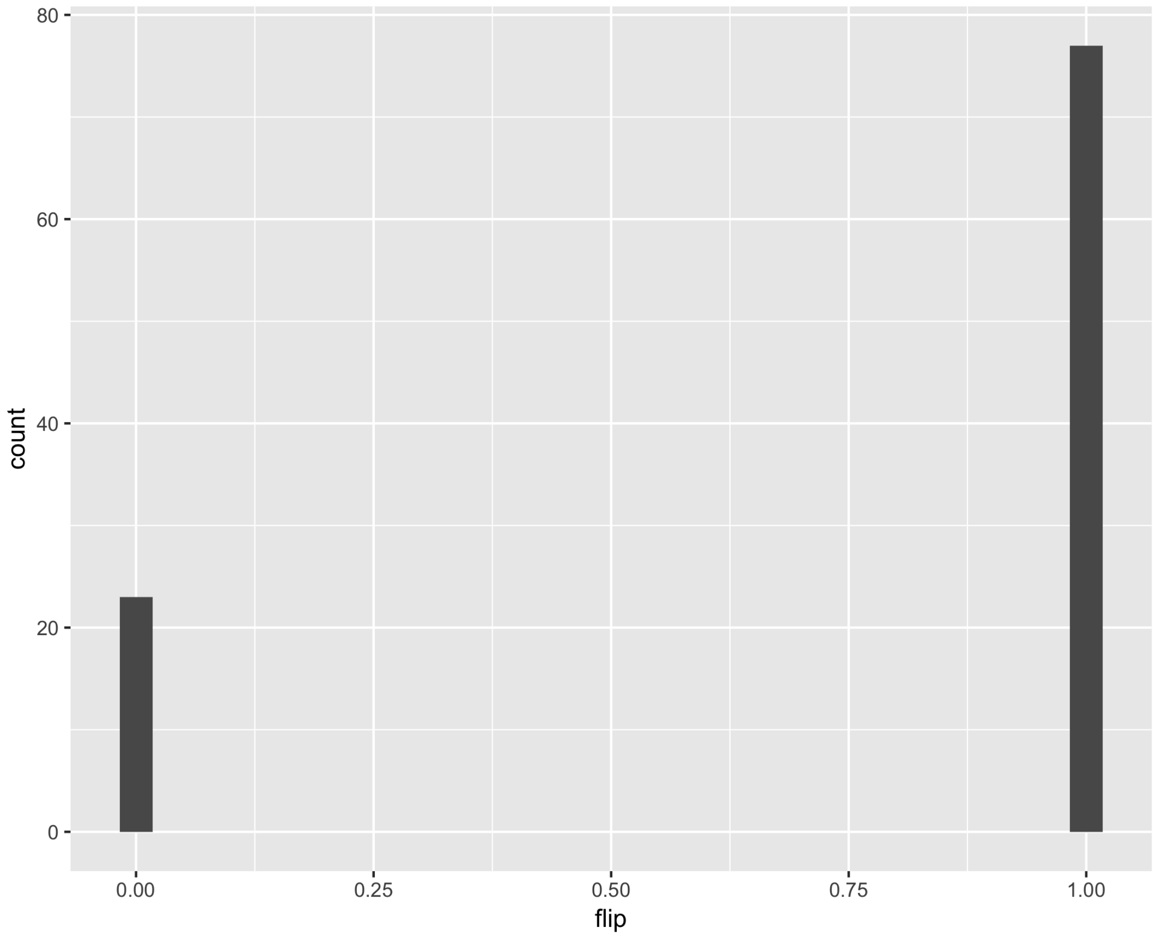


Continuous versus discrete variables

Gaussian distribution



Bernoulli distribution (flipping a coin)



Bernoulli distribution

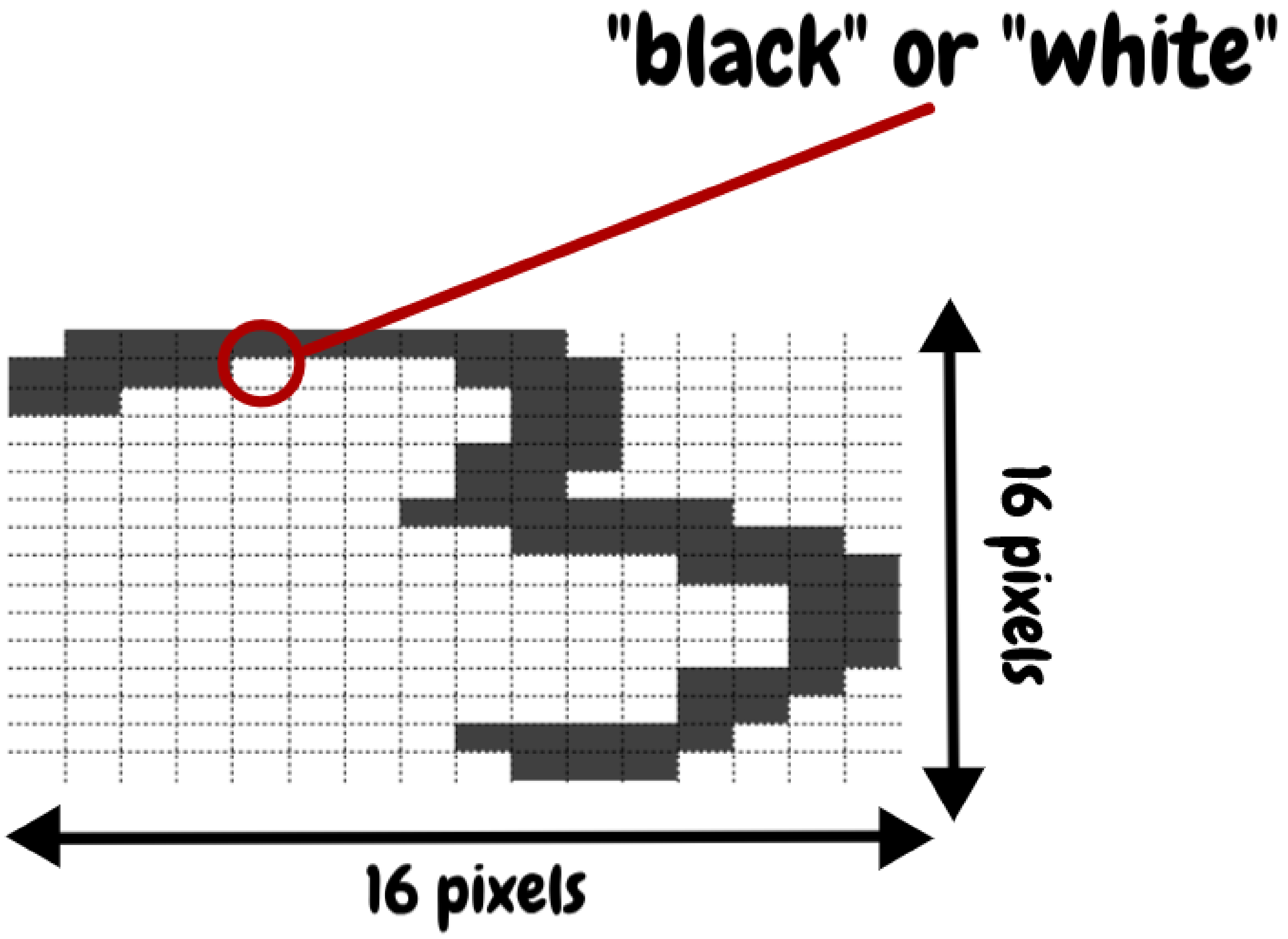
- Two possible outcomes
  - "tails" or "heads"
  - "black" or "white"
- Represented by a probability of "success" $\rightarrow p$
  - $(1 - p)$ = probability for the other option

Sample of Bernoulli distribution

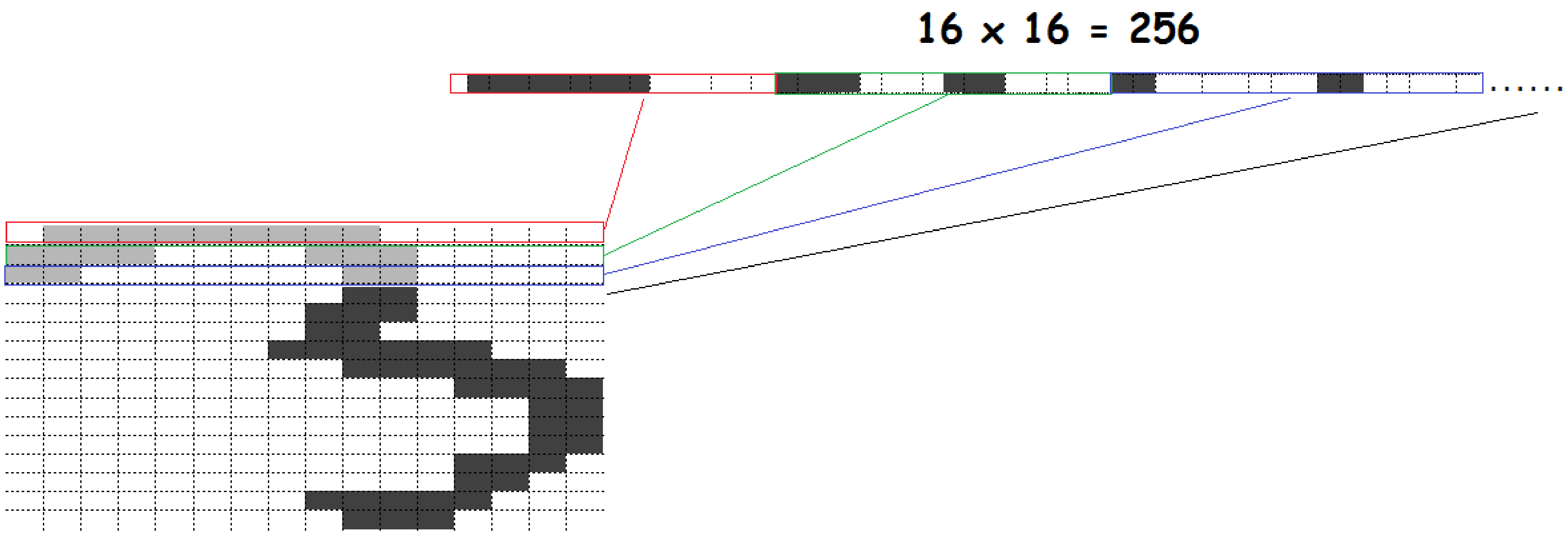
```
> p <- 0.7
> bernoulli <- sample(c(0, 1), 100, replace = TRUE, prob = c(1-p, p))
> head(bernoulli)
```

```
[1] 1 1 1 0 0 1
```

Binary image as Bernoulli distributions



Binary image as Bernoulli vector



Sample of multivariate Bernoulli distribution

```
> p1 <- 0.7; p2 <- 0.5; p3 <- 0.4
>
> bernoulli_1 <- sample(c(0, 1), 100, replace = TRUE, prob = c(1-p1, p1))
> bernoulli_2 <- sample(c(0, 1), 100, replace = TRUE, prob = c(1-p2, p2))
> bernoulli_3 <- sample(c(0, 1), 100, replace = TRUE, prob = c(1-p3, p3))
>
> multi_bernoulli <- cbind(bernoulli_1, bernoulli_2, bernoulli_3)
>
> head(multi_bernoulli)
```

```
     bernoulli_1 bernoulli_2 bernoulli_3
[1,]           1           0           0
[2,]           0           0           0
[3,]           0           0           1
[4,]           1           0           0
[5,]           1           1           1
[6,]           1           0           0
```

```
> p_vector <- c(p1, p2, p3)
```

Bernoulli mixture models

Handwritten digits dataset:

1. Which is the suitable probability distribution?
    - (multivariate) Bernoulli distribution.
2. How many subpopulations should we consider?
    - Let's try with two. That is two binary vectors of size 256.
3. Which are the parameters and their estimations?
    - Each p for each binary vector. Also the two proportions.

MIXTURE MODELS IN R

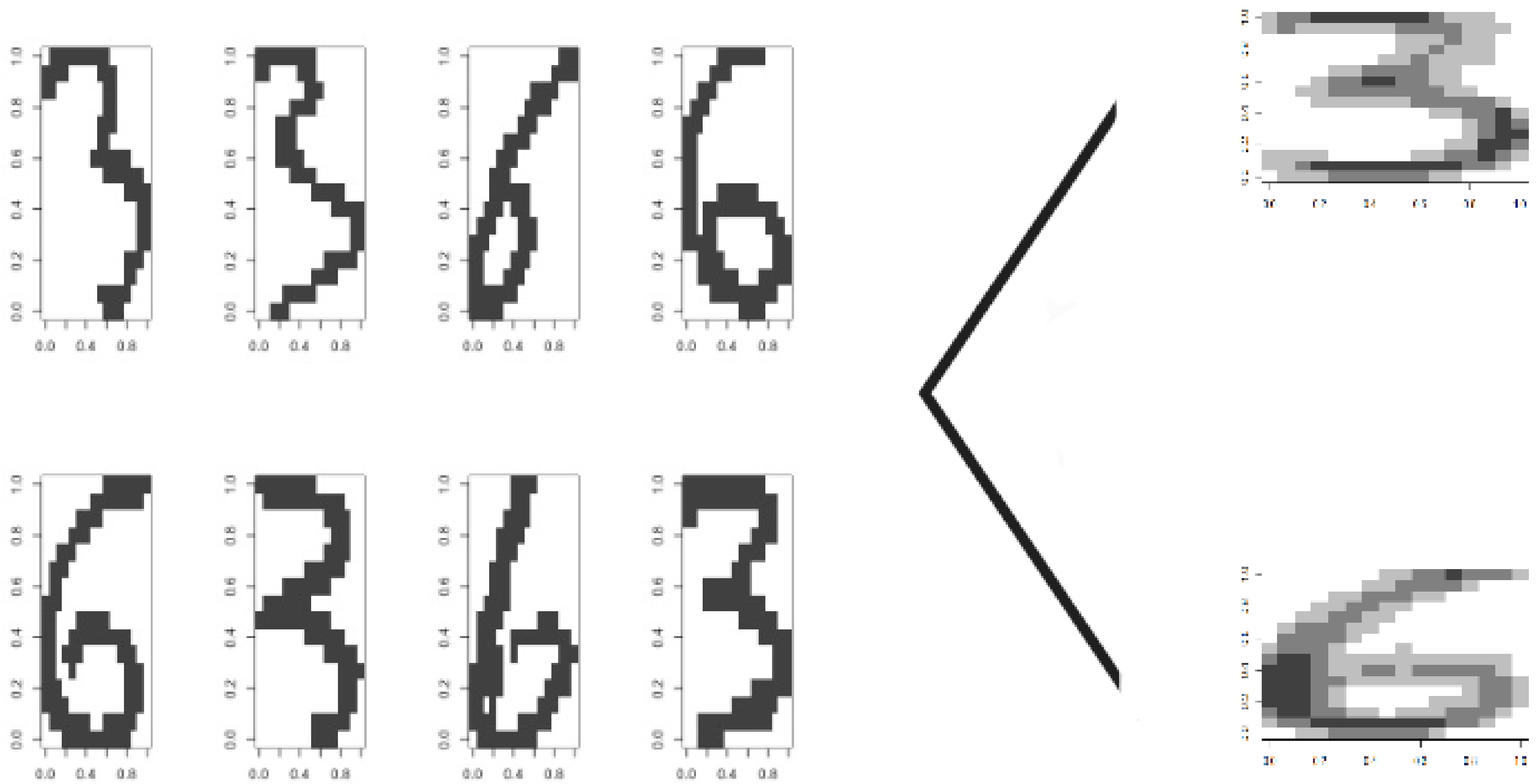
Let's practice

MIXTURE MODELS IN R

Bernoulli Mixture Models with flexmix

Victor Medina
Researcher at SBIF

The problem

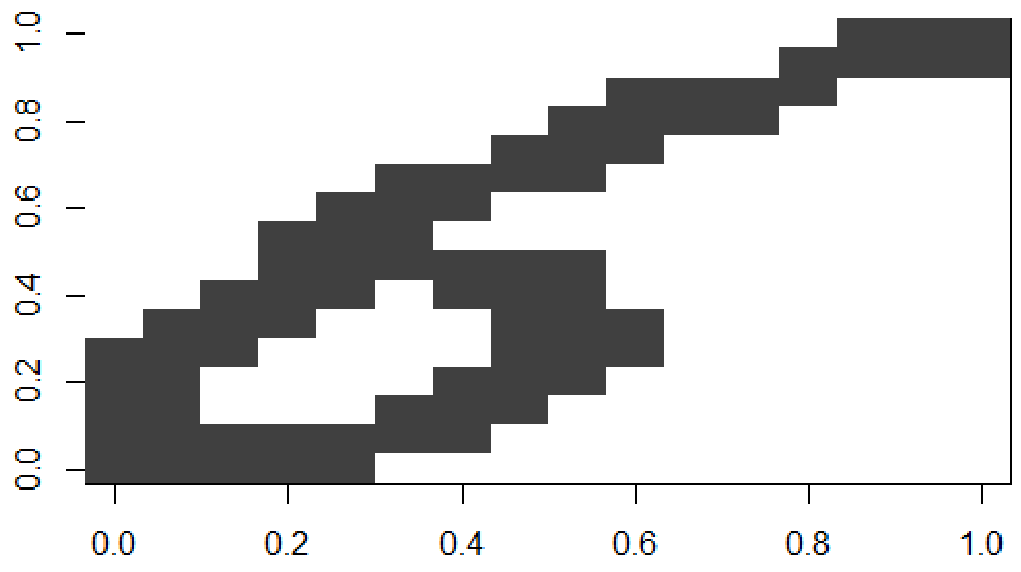


The dataset

```
digits_sample <- as.matrix(digits)
dim(digits_sample)
```

```
[1] 320 256
```

```
show_digit(digits_sample[320,])
```



Fit Bernoulli Mixture Models

```
bernoulli_mix_model <- flexmix(digits_sample~1,
                               k=2,
                               model=FLXMCmvbinary(),
                               control = list(tolerance = 1e-15, iter.max = 1000))
```

- `digits_sample` is a matrix
- `FLXMCmvbinary()` specifies the Bernoulli distribution

The proportions

```
prior(bernoulli_mix_model)
```

```
[1] 0.503125 0.496875
```

parameters function

```
param_comp1 <- parameters(bernoulli_mix_model, component = 1)
param_comp2 <- parameters(bernoulli_mix_model, component = 2)

dim(param_comp1)
```

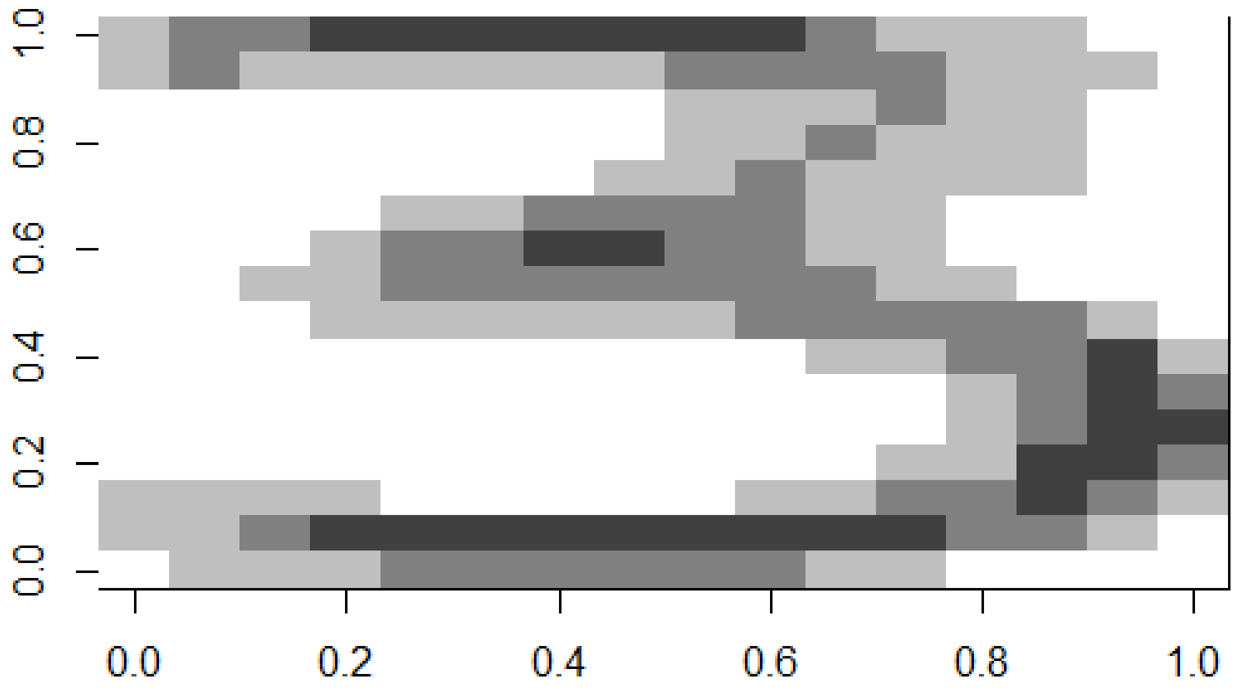
```
[1] 256   1
```

```
head(param_comp1)
```

```
             Comp.1
center.V1 0.3291926
center.V2 0.5093168
center.V3 0.6645963
center.V4 0.7639751
center.V5 0.8136646
center.V6 0.8571428
```

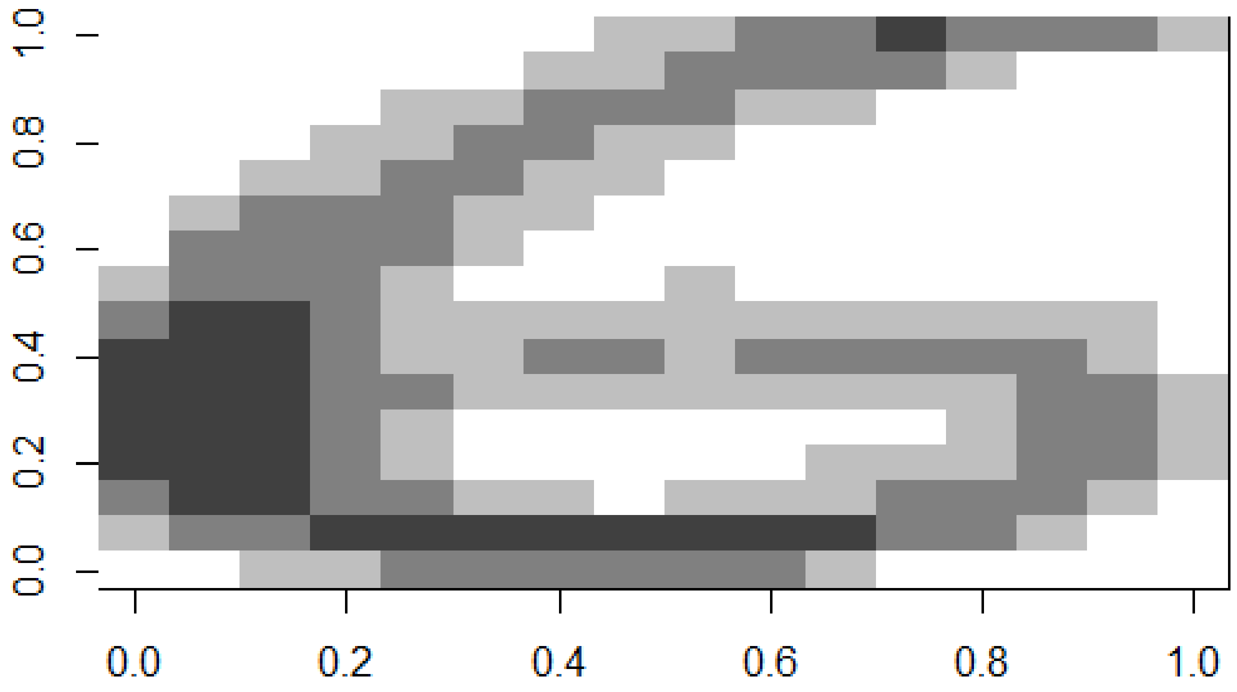
Visualize the component 1

```
show_digit(param_comp1)
```



Visualize the component 2

```
show_digit(param_comp2)
```



MIXTURE MODELS IN R

Let's practice!

MIXTURE MODELS IN R

Poisson Mixture Models

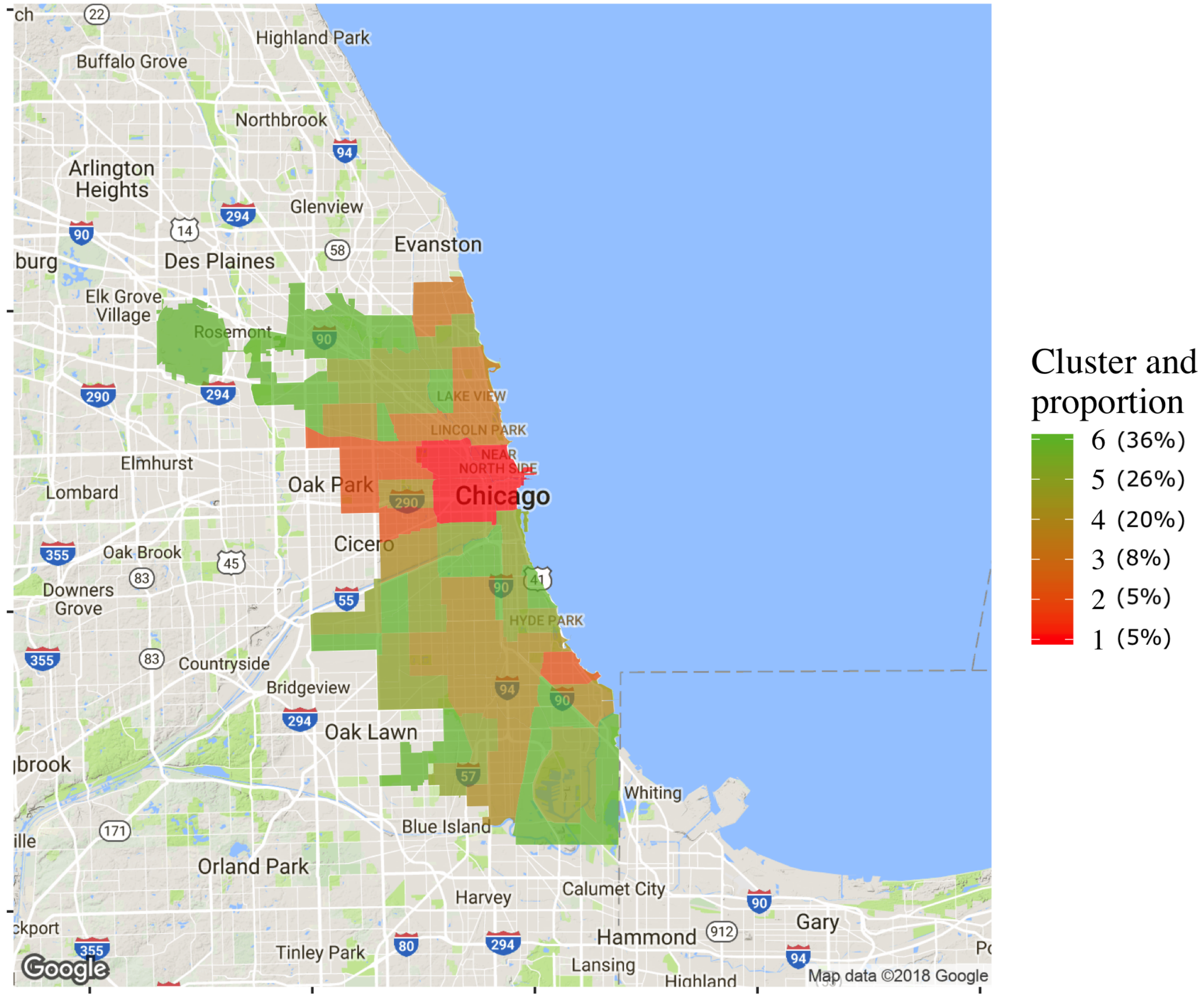
Victor Medina
Researches at SBIF

The crimes dataset

```
# Have a look at the data
glimpse(crimes)
```

```
Observations: 77
Variables: 13
$ COMMUNITY              <chr> "ALBANY PARK", "ARCHER HEIGHTS", "...
$ ASSAULT                <int> 123, 51, 74, 169, 708, 1198, 118, ...
$ BATTERY                <int> 429, 134, 184, 448, 1681, 3347, 28...
$ BURGLARY               <int> 147, 92, 55, 194, 339, 517, 76, 14...
$ `CRIMINAL DAMAGE`      <int> 287, 114, 99, 379, 859, 1666, 150,...
$ `CRIMINAL TRESPASS`    <int> 38, 23, 56, 43, 228, 265, 29, 36, ...
$ `DECEPTIVE PRACTICE`   <int> 137, 67, 59, 178, 310, 767, 73, 20...
$ `MOTOR VEHICLE THEFT`  <int> 176, 50, 37, 189, 281, 732, 58, 12...
$ NARCOTICS              <int> 27, 18, 9, 30, 345, 1456, 15, 22, ...
$ OTHER                  <int> 107, 37, 48, 114, 584, 1261, 76, 8...
$ `OTHER OFFENSE`        <int> 158, 44, 35, 164, 590, 1130, 94, 1...
$ ROBBERY                <int> 144, 30, 98, 111, 349, 829, 65, 10...
$ THEFT                  <int> 690, 180, 263, 461, 1201, 2137, 23...
```

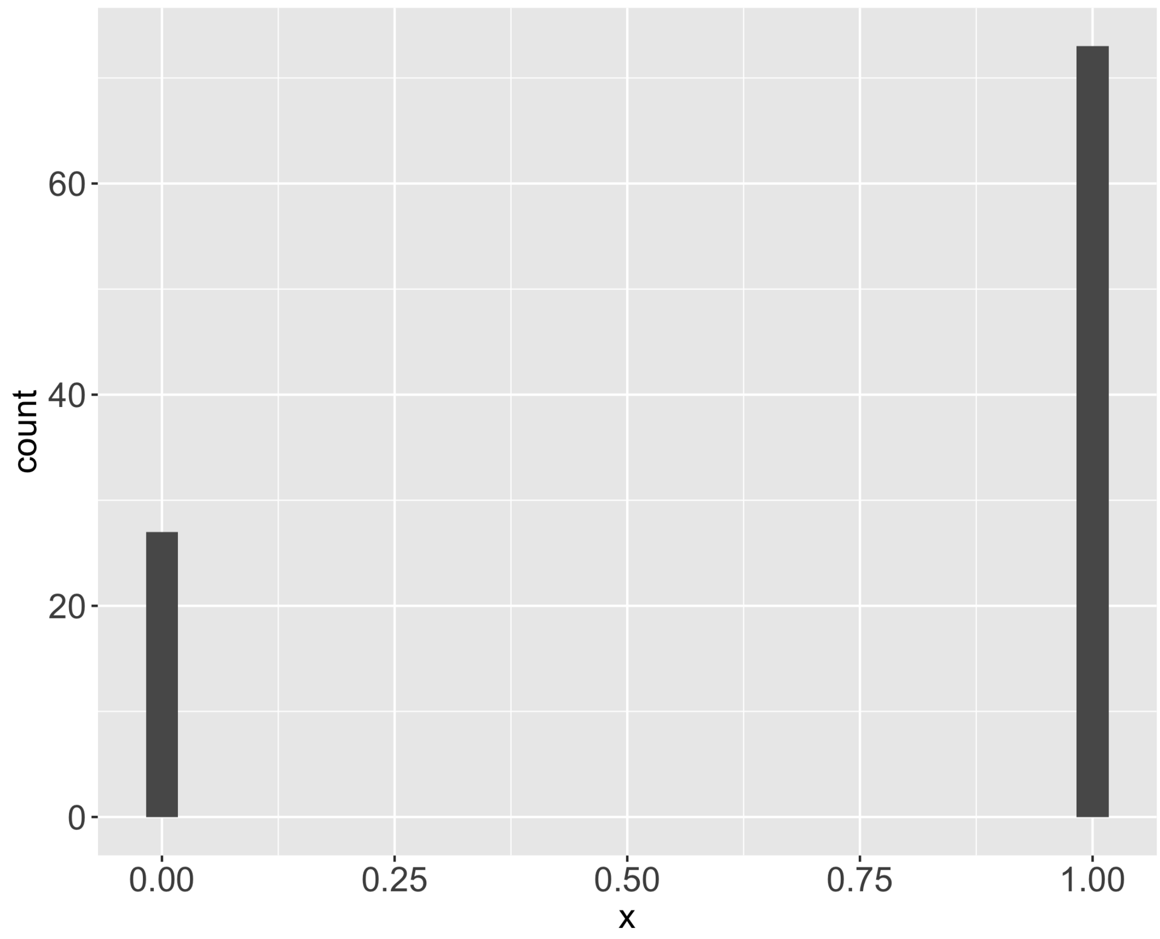
The problem to solve



Comparison of Poisson with Bernoulli

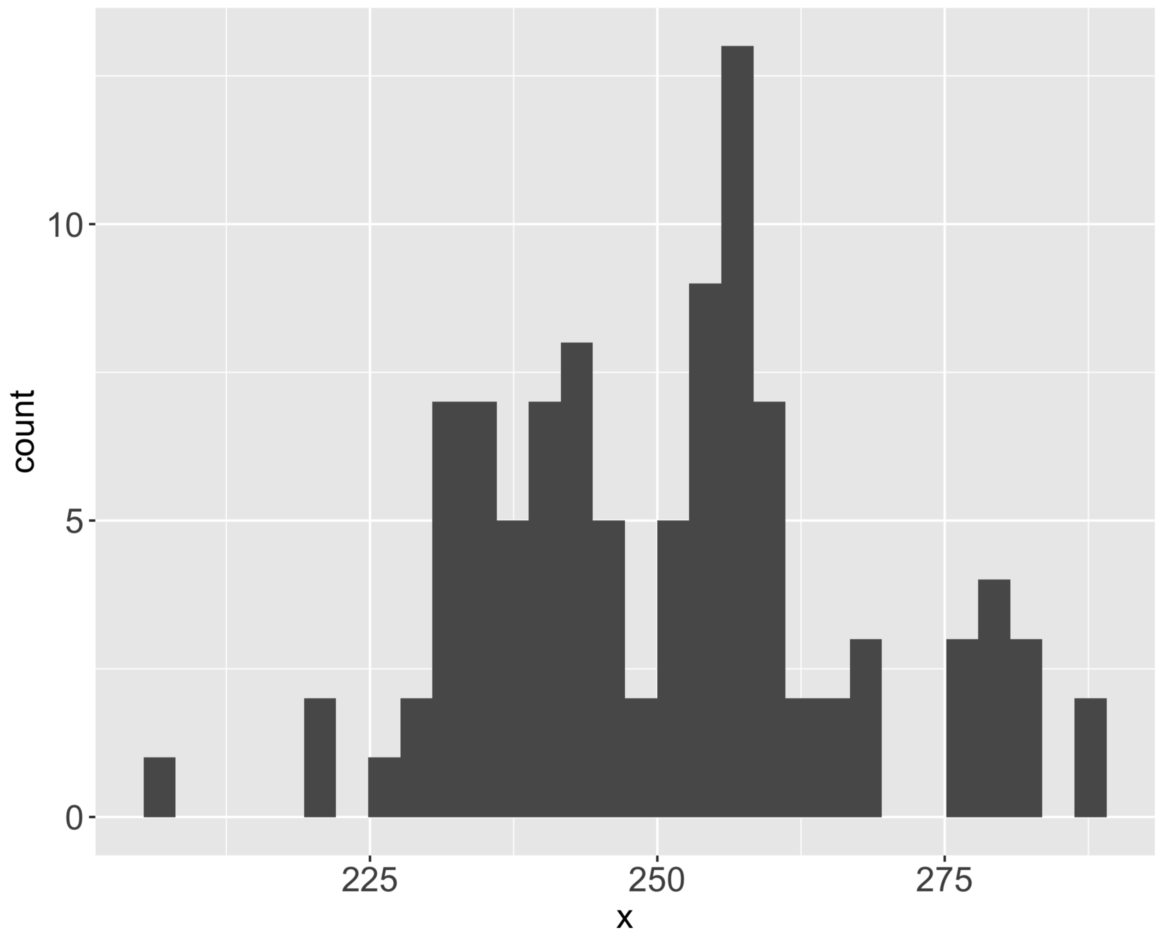
Bernoulli distribution

```
data.frame(x = bernoulli) %>%
  ggplot(aes(x = x)) + geom_histogram(
```



Poisson distribution

```
data.frame(x = rpois(100,  250)) %>%
  ggplot(aes(x = x)) + geom_histogram(
```



Poisson distribution

- Number of times an event occurs in an interval of time
- Examples:
  - Number of car accidents in a year
  - Number of emails received in a day
  - Number of robberies in an area of the city for a period of one year

Sample of Poisson distribution

```
> lambda_1 <- 100
> poisson_1 <- rpois(n = 100, lambda = lambda_1)
> head(poisson_1)
```

```
[1]  98  98  87  77 102  85
```

Sample of multivariate Poisson distribution

```
> lambda_1 <- 100
> lambda_2 <- 200
> lambda_3 <- 300
>
> poisson_1 <- rpois(n = 100, lambda = lambda_1)
> poisson_2 <- rpois(n = 100, lambda = lambda_2)
> poisson_3 <- rpois(n = 100, lambda = lambda_3)
>
> multi_poisson <- cbind(poisson_1, poisson_2, poisson_3)
>
> head(multi_poisson)
```

```
     poisson_1 poisson_2 poisson_3
[1,]        98       198       296
[2,]        98       213       312
[3,]        87       197       311
[4,]        77       215       299
[5,]       102       189       313
[6,]        85       199       309
```

Count data as (multi) Poisson distribution

```
> head(crimes)
```

```
# A tibble: 6 x 13
  COMMUNITY       ASSAULT BATTERY BURGLARY `CRIMINAL DAMAGE` `CRIMINAL TRESPASS`
  <chr>             <int>   <int>    <int>             <int>               <int>
1 ALBANY PARK         123     429      147               287                  38
2 ARCHER HEIGHTS       51     134       92               114                  23
3 ARMOUR SQUARE        74     184       55                99                  56
4 ASHBURN             169     448      194               379                  43
5 AUBURN GRESHAM      708    1681      339               859                 228
6 AUSTIN             1198    3347      517              1666                 265
# ... with 7 more variables: `DECEPTIVE PRACTICE` <int>, `MOTOR VEHICLE THEFT` <
#   NARCOTICS <int>, OTHER <int>, `OTHER OFFENSE` <int>, ROBBERY <int>, THEFT <i
```

Poisson Mixture Model

1. Which is the suitable probability distribution?
    - (multi) Poisson distribution
2. How many subpopulations should we consider?
    - Let's try from 1 to 15 clusters and pick by BIC.
3. Which are the parameters and their estimations?
    - Each lambda for each of the multi Poisson. Also the proportions.

MIXTURE MODELS IN R

Let's practice!

MIXTURE MODELS IN R

Poisson Mixture Models with flexmix

Victor Medina
Researcher at SBIF

The problem to solve

1. Which is the suitable probability distribution?
   - (multi) Poisson distribution
2. How many subpopulations should we consider?
   - Let's try from 1 to 15 clusters and pick by BIC.
3. Which are the parameters and their estimations?
   - Each lambda for each of the multi Poisson. Also the proportions.

Fit with flexmix

```
crimes_matrix <- as.matrix(crimes[,-1])
```

```
poisson_mix_model <- stepFlexmix(crimes_matrix ~ 1,
                                 k = 1:15,
                                 nrep = 5,
                                 model = FLXMCmvpois(),
                                 control = list(tolerance = 1e-15, iter = 1000))
```

- Use `stepFlexmix` instead of `flexmix` function.
- `k` is now a range of values.
- `nrep` is the number of repetitions the EM algorithm runs for each `k` value.
- The Poisson distribution is `FLXMCmvpois`

Pick the best model

```
best_fit <- getModel(poisson_mix_model, which = "BIC")
```

- Other statistical criteria implemented in `flexmix` are the AIC and ICL.

The proportions

```
prior(best_fit)
```

```
[1] 0.07792208 0.05194805 0.19480519 0.27272727 0.20779224 0.19480517
```

The parameters

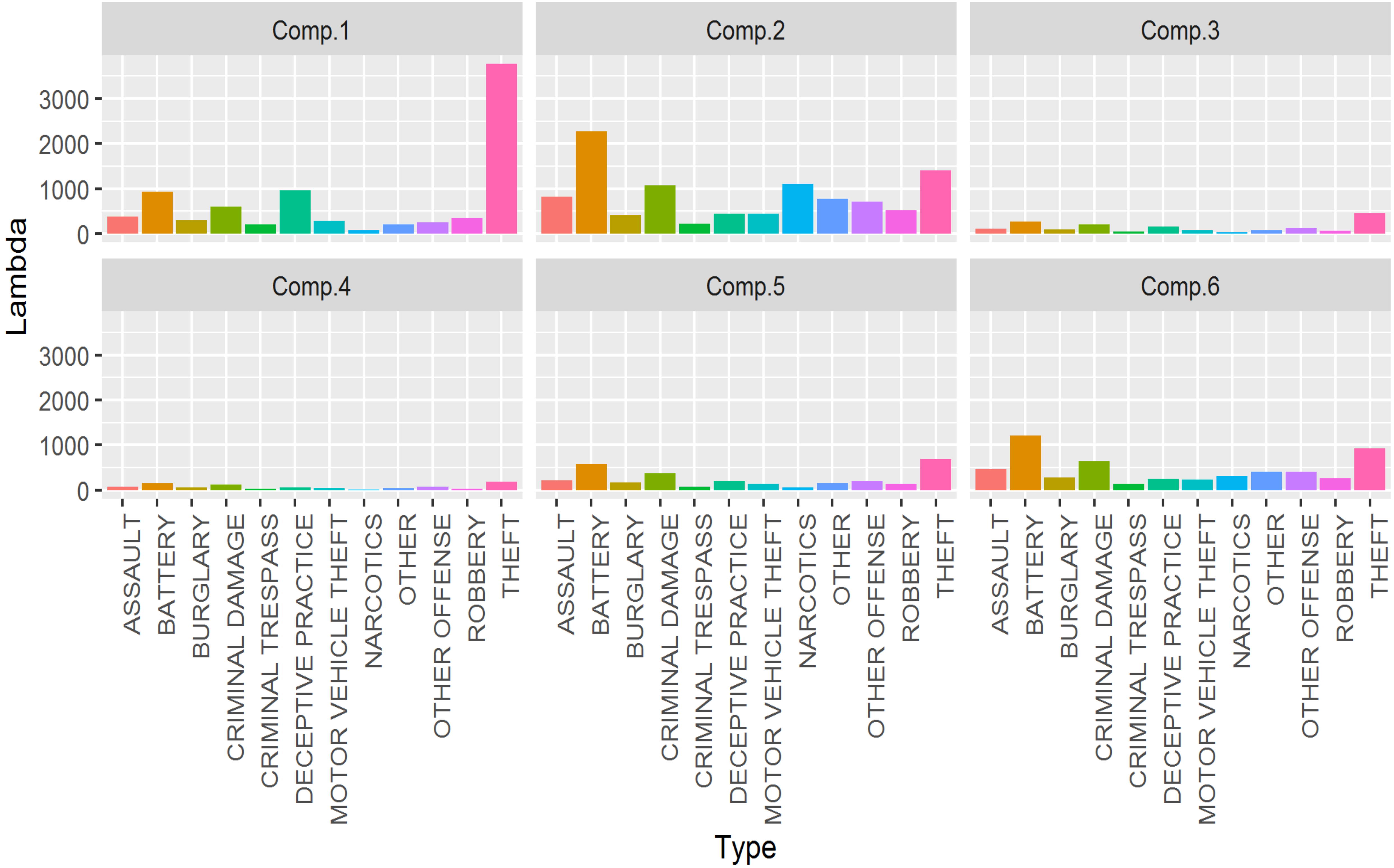
```
param_pmm <- data.frame(parameters(best_fit))
```

```
param_pmm <- param_pmm %>% mutate(Type = colnames(crimes_matrix))
head(param_pmm)
```

```
     Comp.1  Comp.2    Comp.3    Comp.4   Comp.5    Comp.6                 Type
1  380.3333  821.75 112.26667  67.57143 216.9375  475.3334              ASSAULT
2  929.5000 2271.50 268.13333 153.14286 574.7500 1204.8667              BATTERY
3  303.8333  418.00  98.60000  52.04762 174.9375  272.9333             BURGLARY
4  601.3333 1074.50 199.66666 116.90476 370.9375  648.6667      CRIMINAL DAMAGE
5  210.5000  223.75  49.73333  25.00000  81.0625  139.0000   CRIMINAL TRESPASS
6  973.1667  438.00 158.80000  61.95238 196.7500  241.4666 DECEPTIVE PRACTICE
```

Visualize the clusters

```
param_pmm %>%
  gather(Components, Lambda, -Type) %>%
  ggplot(aes(x = Type, y = Lambda, fill = Type)) +
  geom_bar(stat = "identity") +
  facet_wrap(~ Components) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = "none")
```

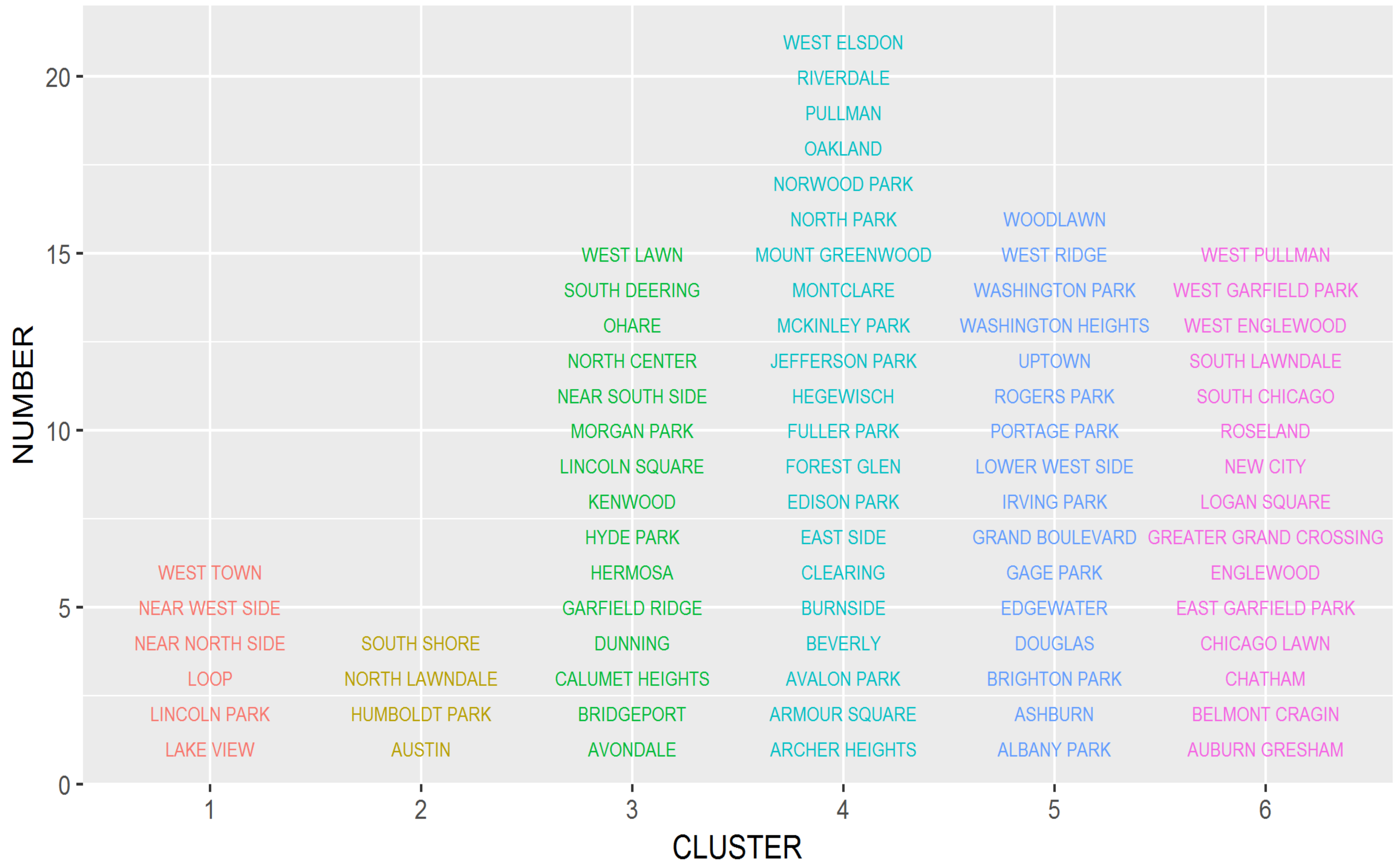

Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6
Lambda
3000
2000
1000
0
ASSAULT
BATTERY
BURGLARY
CRIMINAL DAMAGE
CRIMINAL TRESPASS
DECEPTIVE PRACTICE
MOTOR VEHICLE THEFT
NARCOTICS
OTHER
OTHER OFFENSE
ROBBERY
THEFT
Type

Assign cluster to each community

```
crimes_c <- crimes %>%
  mutate(CLUSTER = factor(clusters(best_fit)))
```

Visualize the clusters with their communities

```
crimes_c %>%
  group_by(CLUSTER) %>%
  mutate(NUMBER = row_number()) %>%
  ggplot(aes(x = CLUSTER, y = NUMBER, col = CLUSTER)) +
  geom_text(aes(label = COMMUNITY), size = 2.3)+
  theme(legend.position="none")
```


WEST ELSDON
RIVERDALE
PULLMAN
OAKLAND
NORWOOD PARK
NORTH PARK
WOODLAWN
WEST LAWN
MOUNT GREENWOOD
WEST RIDGE
WEST PULLMAN
SOUTH DEERING
MONTCLARE
WASHINGTON PARK
WEST GARFIELD PARK
OHARE
MCKINLEY PARK
WASHINGTON HEIGHTS
WEST ENGLEWOOD
NORTH CENTER
JEFFERSON PARK
UPTOWN
SOUTH LAWNDALE
NEAR SOUTH SIDE
HEGEWISCH
ROGERS PARK
SOUTH CHICAGO
MORGAN PARK
FULLER PARK
PORTAGE PARK
ROSELAND
LINCOLN SQUARE
FOREST GLEN
LOWER WEST SIDE
NEW CITY
KENWOOD
EDISON PARK
IRVING PARK
LOGAN SQUARE
HYDE PARK
EAST SIDE
GRAND BOULEVARD
GREATER GRAND CROSSING
WEST TOWN
HERMOSA
CLEARING
GAGE PARK
ENGLEWOOD
NEAR WEST SIDE
GARFIELD RIDGE
BURNSIDE
EDGEWATER
EAST GARFIELD PARK
NEAR NORTH SIDE
SOUTH SHORE
DUNNING
BEVERLY
DOUGLAS
CHICAGO LAWN
LOOP
NORTH LAWNDALE
CALUMET HEIGHTS
AVALON PARK
BRIGHTON PARK
CHATHAM
LINCOLN PARK
HUMBOLDT PARK
BRIDGEPORT
ARMOUR SQUARE
ASHBURN
BELMONT CRAGIN
LAKE VIEW
AUSTIN
AVONDALE
ARCHER HEIGHTS
ALBANY PARK
AUBURN GRESHAM
NUMBER
20
15
10
5
0
1
2
3
4
5
6
CLUSTER

MIXTURE MODELS IN R

Let's practice!