Classification des utilisateurs de Twitter en fonction de leurs comportements à l'aide d'un algorithme des *K-means*

Vincent Brault, Jean-Marc Francony, Adeline Leclercq-Samson et Matthieu Meynet

Mardi 27 novembre

UNIVERSITÉ Grenoble Alpes

LABORATOIRE JEAN KUNTZMANN
MATHÉMATIQUES APPLIQUÉES - INFORMATIQUE

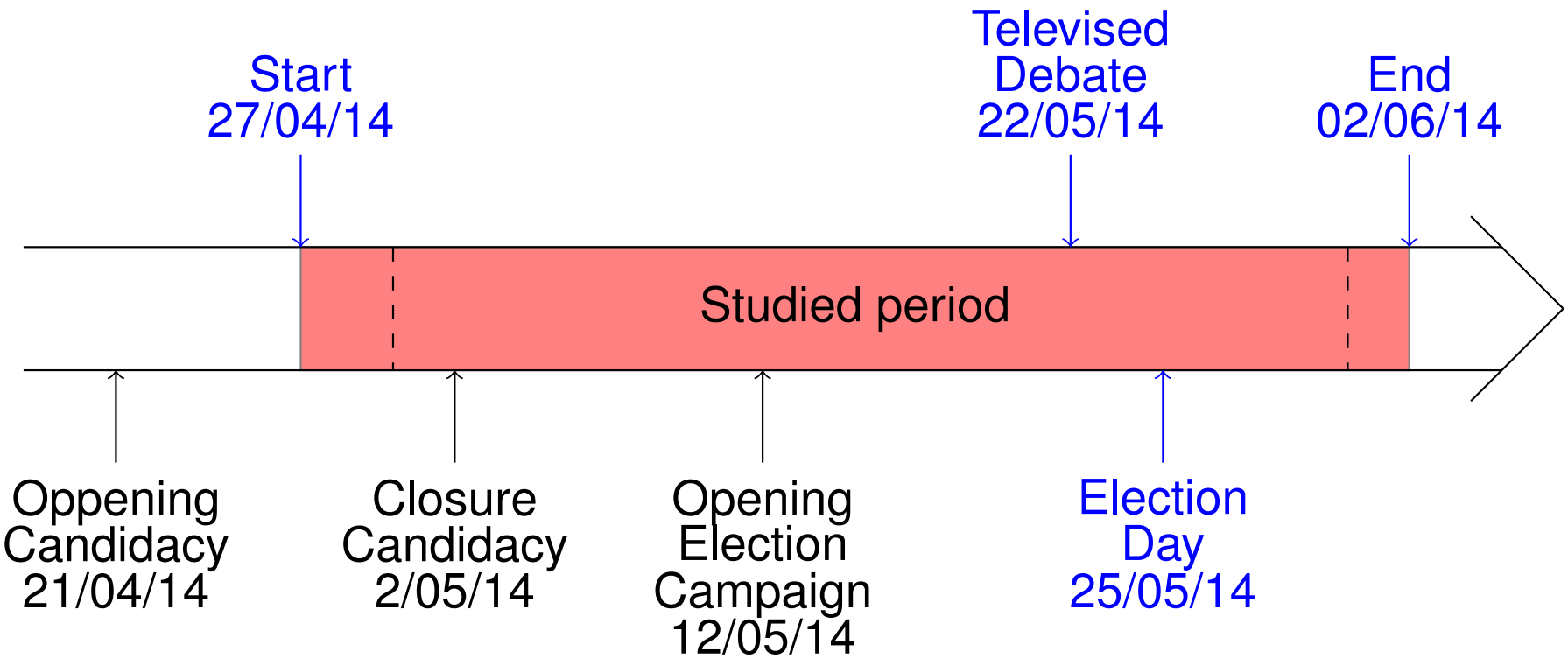
Pacte
UNITÉ MIXTE DE RECHERCHE CNRS

GRESEC
UNIVERSITÉ Grenoble Alpes

Plan

Plan

Context


Start
27/04/14
Televised
Debate
22/05/14
End
02/06/14
Studied period
Oppening
Candidacy
21/04/14
Closure
Candidacy
2/05/14
Opening
Election
Campaign
12/05/14
Election
Day
25/05/14

Julyan Arbel
@JulyanArbel

Suivre

Incoming #RandomGraphTwitter workshop on random graphs with applications to Twitter users behaviour, April 26-27, at @UGrenobleAlpes.

Brault Vincent @Lionning13
L'affiche du workshop #RandomGraphTwitter est disponible avec le programme provisoire.
www-ljk.imag.fr/membres/Vincen...

Traduire le Tweet

16:55 - 22 mars 2018

5 Retweets **5** J'aime

5 5

SFdS

Tweeter votre réponse

- Hashtag : **#EE2014** et **#Europeennes2014**

- Hashtag : **#EE2014** et **#Europeennes2014**
- *Mention* : **@UPR_Asselineau**, **@EELV**, **@DemoryFlorian**...

- Hashtag : **#EE2014** et **#Europeennes2014**
- *Mention* : **@UPR_Asselineau**, **@EELV**, **@DemoryFlorian**...
- *Retweet* (maybe with *mention*)

Question

Are there communication strategies ?

- Bruns, A. (2012). *Journalists and Twitter : How Australian news organisations adapt to a new medium*. Media International Australia, 144(1), 97-107.
- Heinderyckx, F. (2011). *Obama 2008 : l'inflexion numérique*. Hermès, La Revue, (1), 135-136.

Plan

Notations

April 27 at 10am for the individual i

i 3

Notations

April 27-28 at 10am

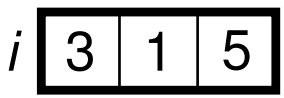
for the individual i

i	3	1

Notations

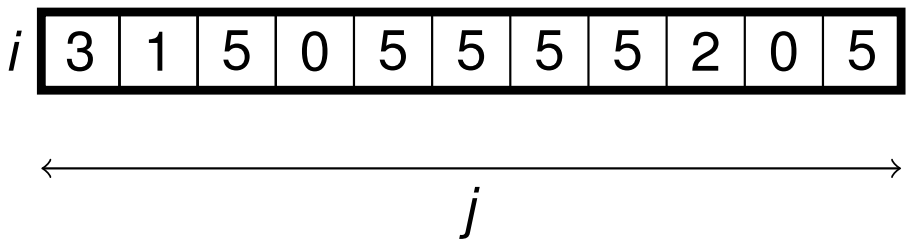
April 27-29 at 10am

for the individual i



Notations

at 10am for the individual i



Notations

at 10am for the individual i and $i+1$

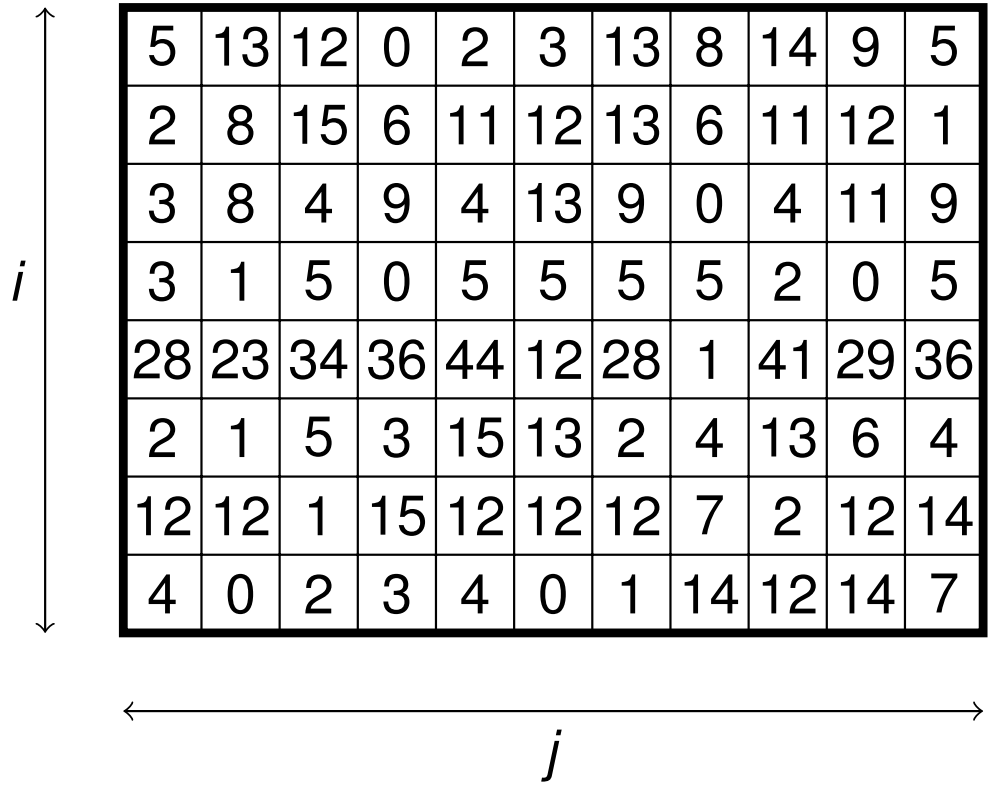
i	3	1	5	0	5	5	5	5	2	0	5
$i+1$	28	23	34	36	44	12	28	1	41	29	36

$\longleftrightarrow$

j

Notations

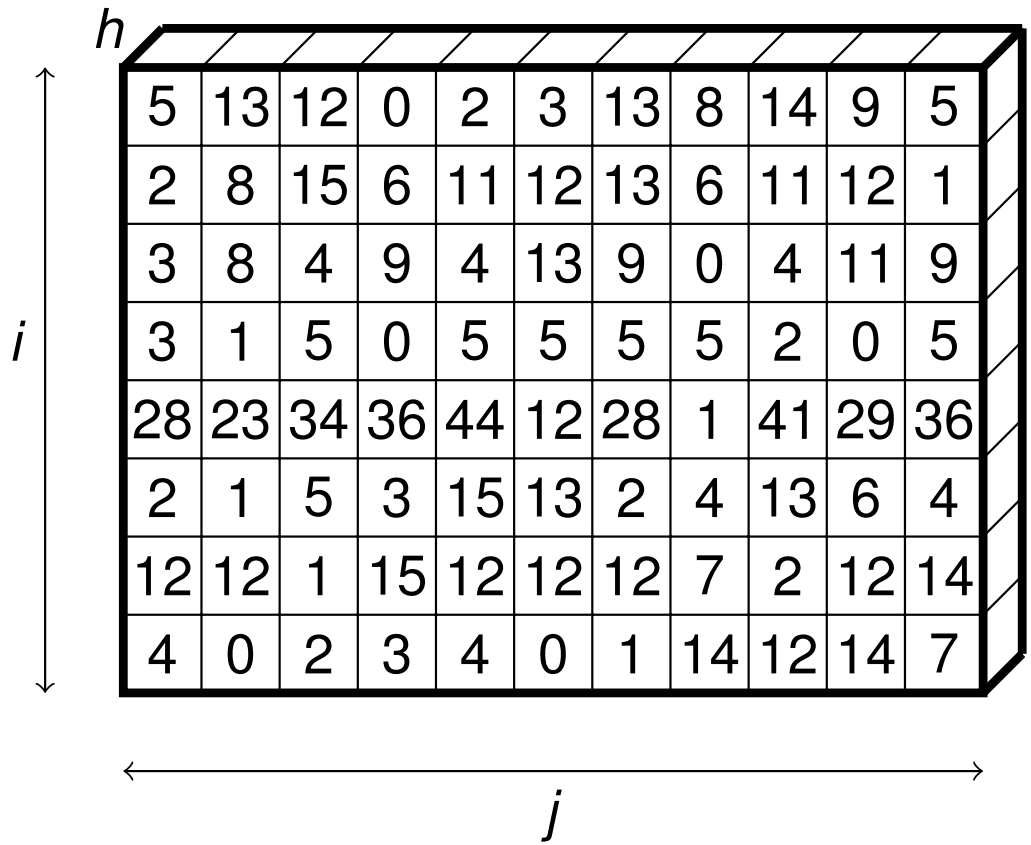
at 10am



5	13	12	0	2	3	13	8	14	9	5
2	8	15	6	11	12	13	6	11	12	1
3	8	4	9	4	13	9	0	4	11	9
3	1	5	0	5	5	5	5	2	0	5
28	23	34	36	44	12	28	1	41	29	36
2	1	5	3	15	13	2	4	13	6	4
12	12	1	15	12	12	12	7	2	12	14
4	0	2	3	4	0	1	14	12	14	7

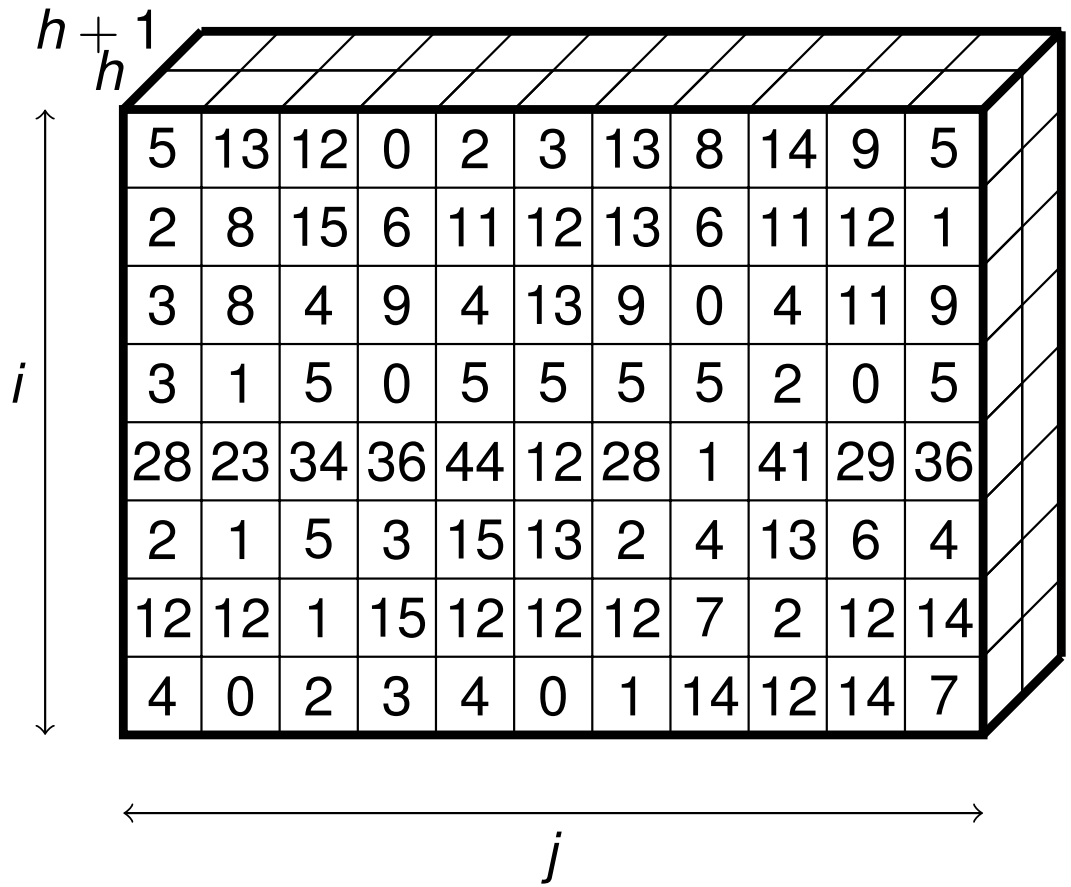
Notations

at 10am

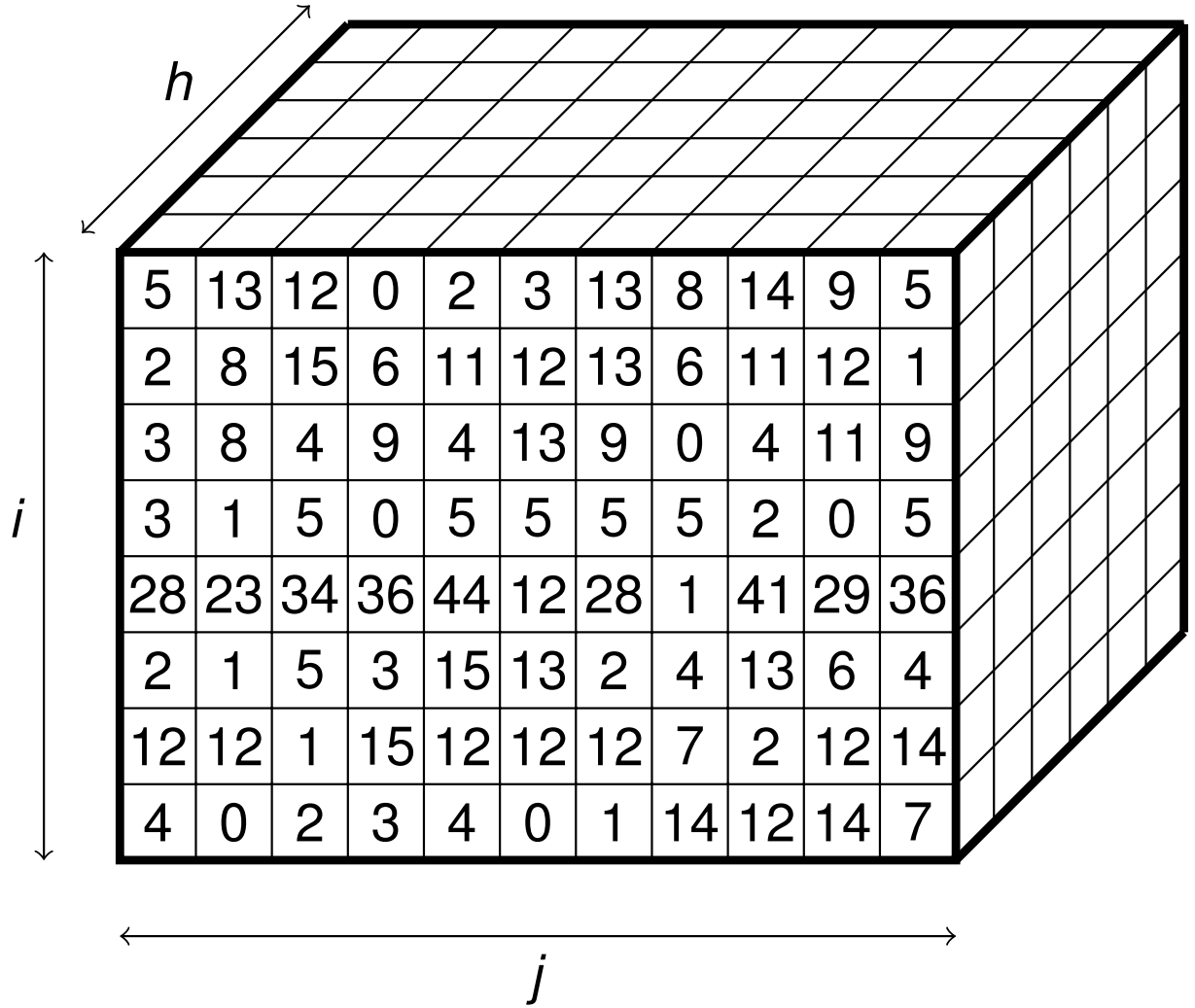


Notations

at 10am and 11am



Notations


h
i
j
5 13 12 0 2 3 13 8 14 9 5
2 8 15 6 11 12 13 6 11 12 1
3 8 4 9 4 13 9 0 4 11 9
3 1 5 0 5 5 5 5 2 0 5
28 23 34 36 44 12 28 1 41 29 36
2 1 5 3 15 13 2 4 13 6 4
12 12 1 15 12 12 12 7 2 12 14
4 0 2 3 4 0 1 14 12 14 7

Notations

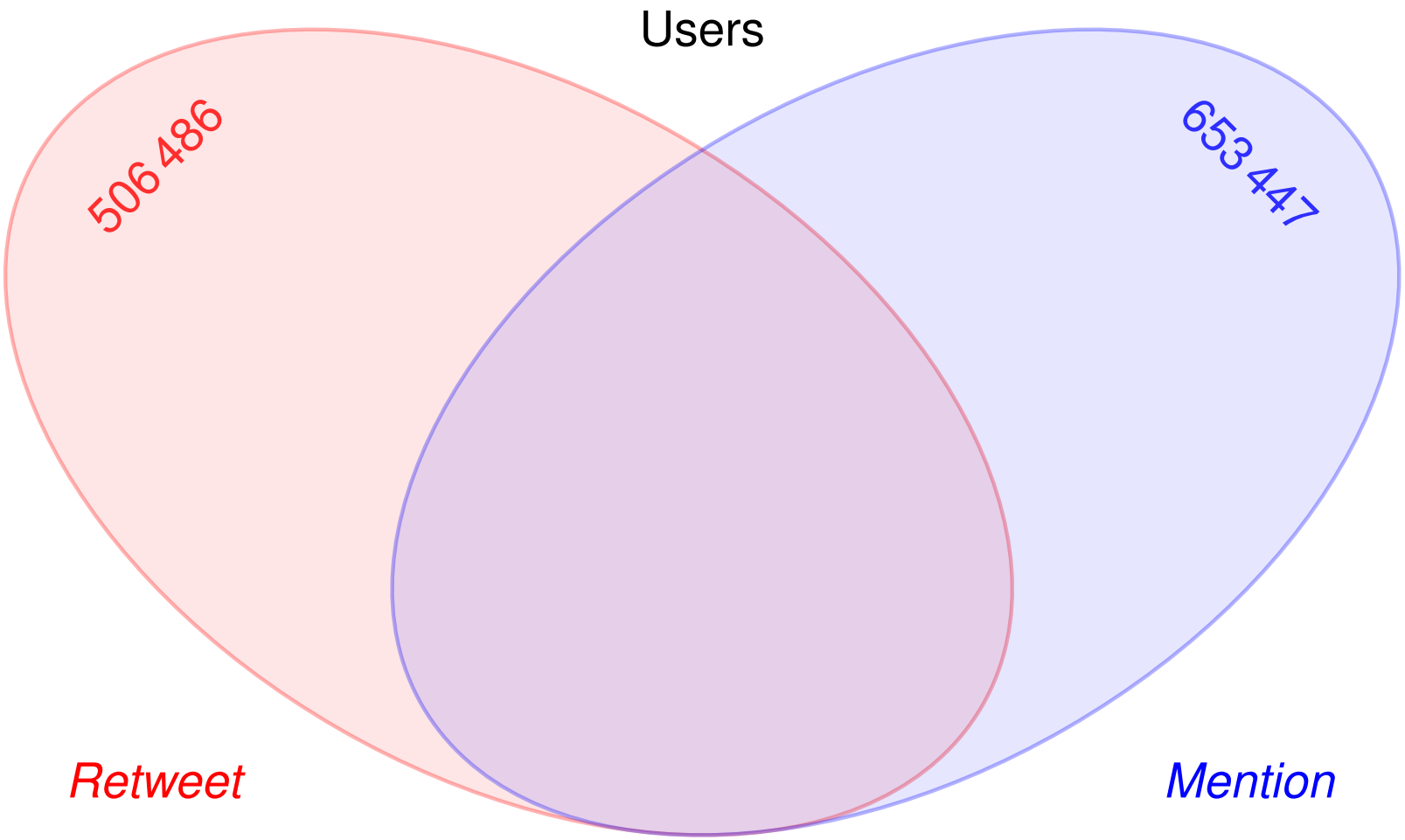
- $i \in \{1, \ldots, I\}$: user.
- $j \in \{1, \ldots, J\}$: day.
- $h \in \{0, \ldots, 23\}$: hour.
- $N^R_{i,j,h}$ (resp. $N^M_{i,j,h}$) : number of *retweets* (resp. *mentions*) by the user i at the time h of the day j.

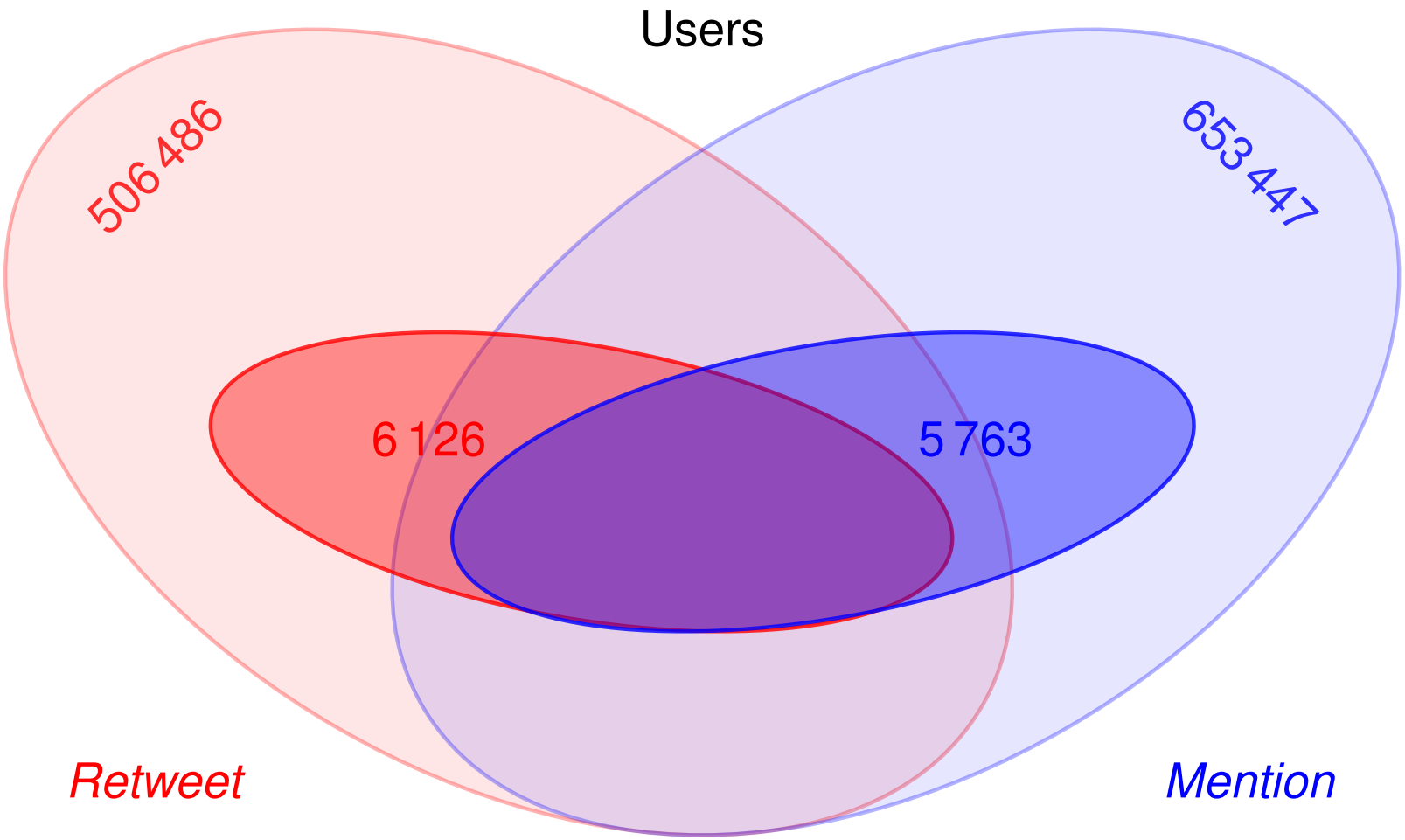
Notations

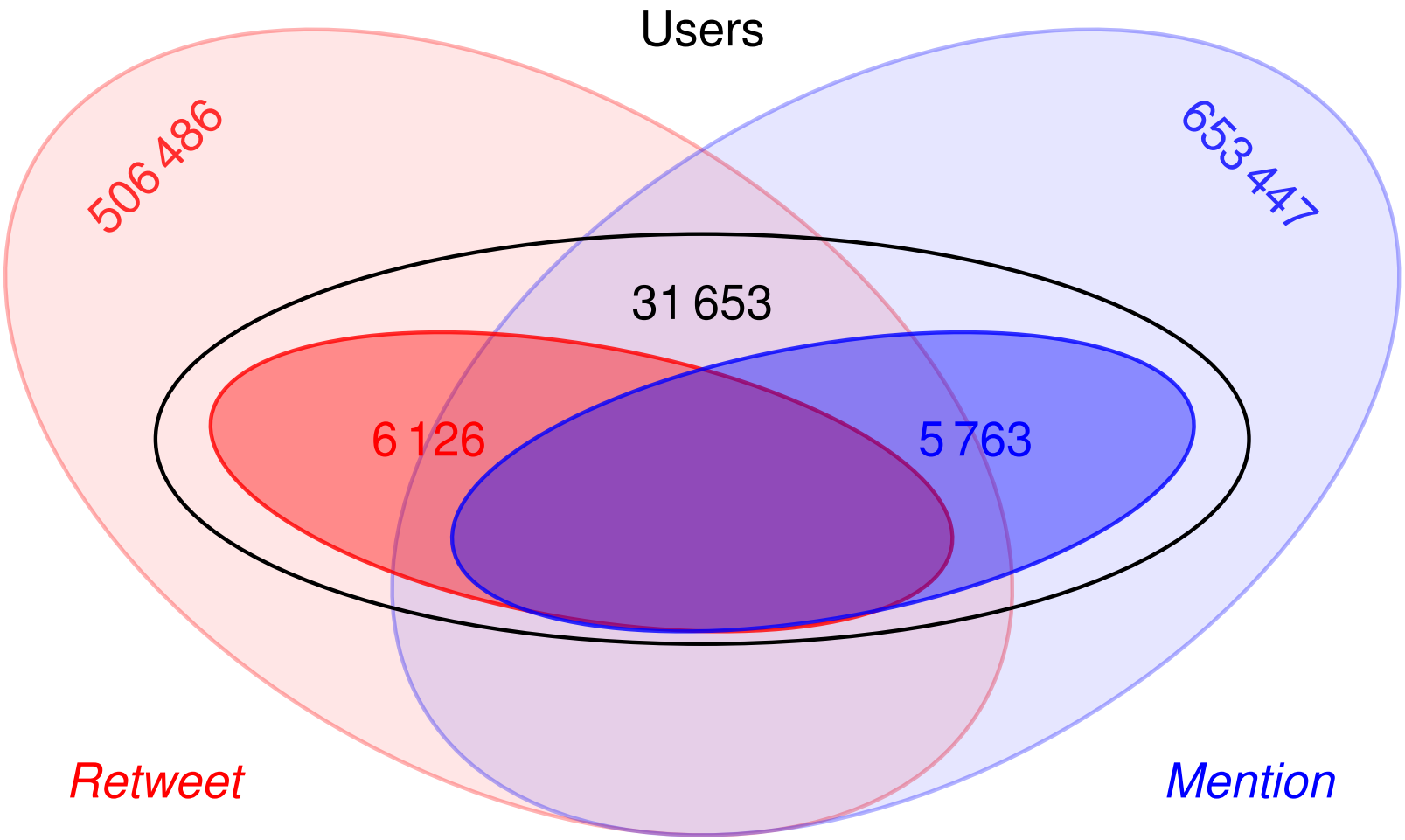
- $i \in \{1, \ldots, I\}$: user.
- $j \in \{1, \ldots, J\}$: day.
- $h \in \{0, \ldots, 23\}$: hour.
- $N^R_{i,j,h}$ (resp. $N^M_{i,j,h}$) : number of *retweets* (resp. *mentions*) by the user i at the time h of the day j.

In particular :

$$N^R_{i,\cdot,h} = \left(N^R_{i,1,h}, N^R_{i,2,h}, \ldots, N^R_{i,J,h}\right)$$


Users
506 486
653 447
Retweet
Mention

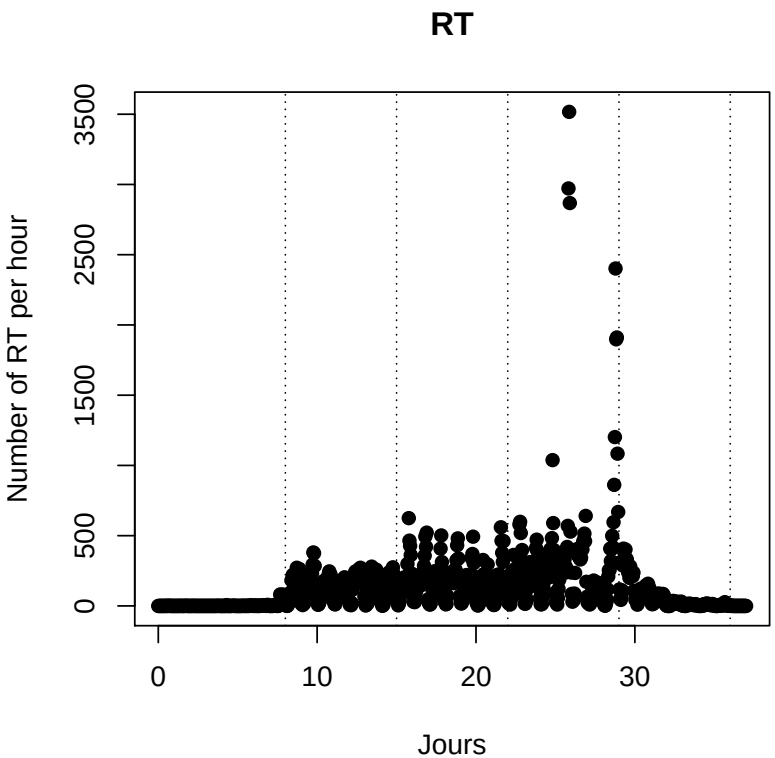

Users
506 486
653 447
6 126
5 763
Retweet
Mention

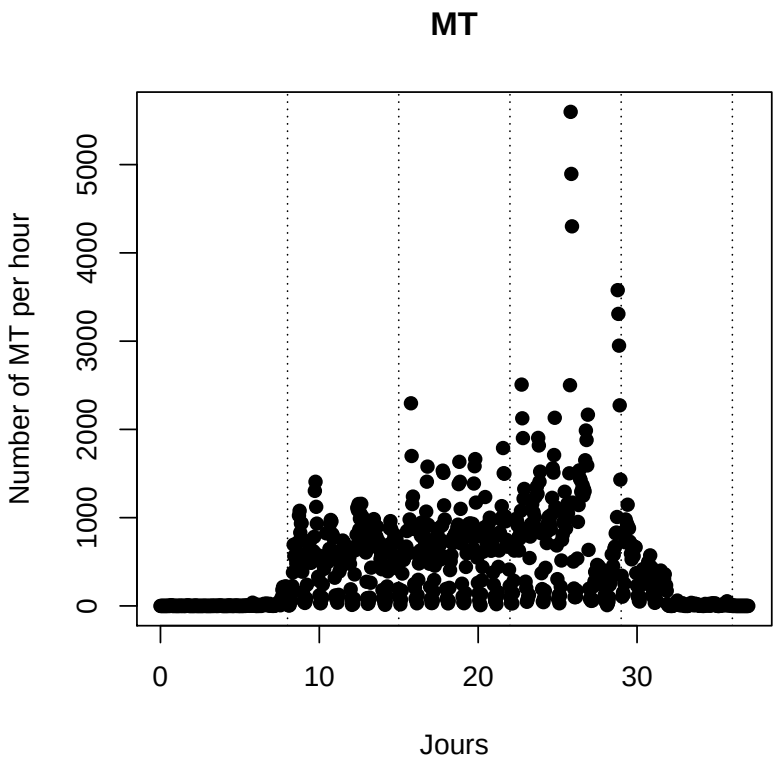

Users
506 486
653 447
31 653
6 126
5 763
Retweet
Mention

Plan

Why to cluster the days ?

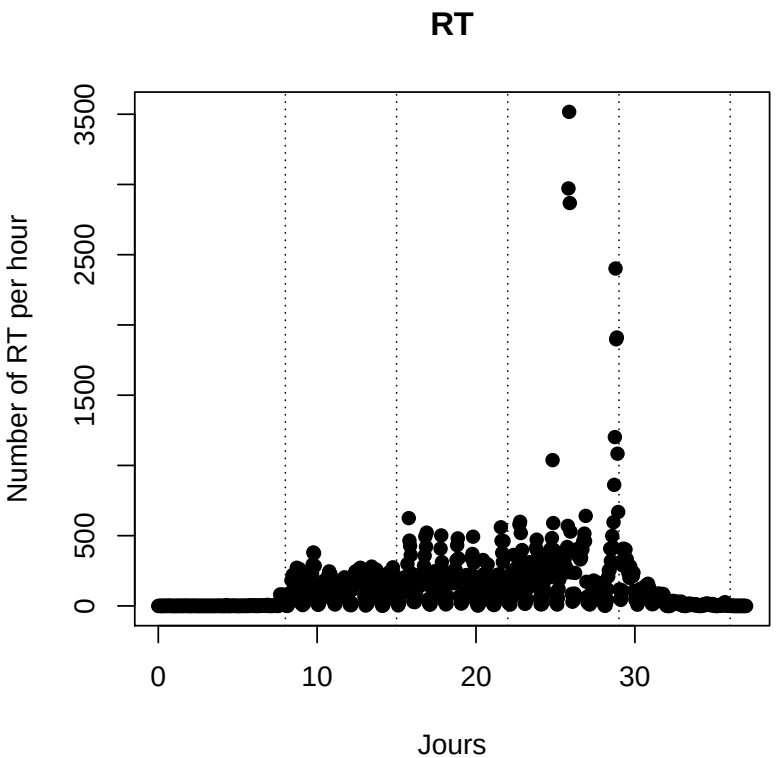
$$N^R_{+,j,h} = \sum_{i=1}^{n} N^R_{i,j,h}$$

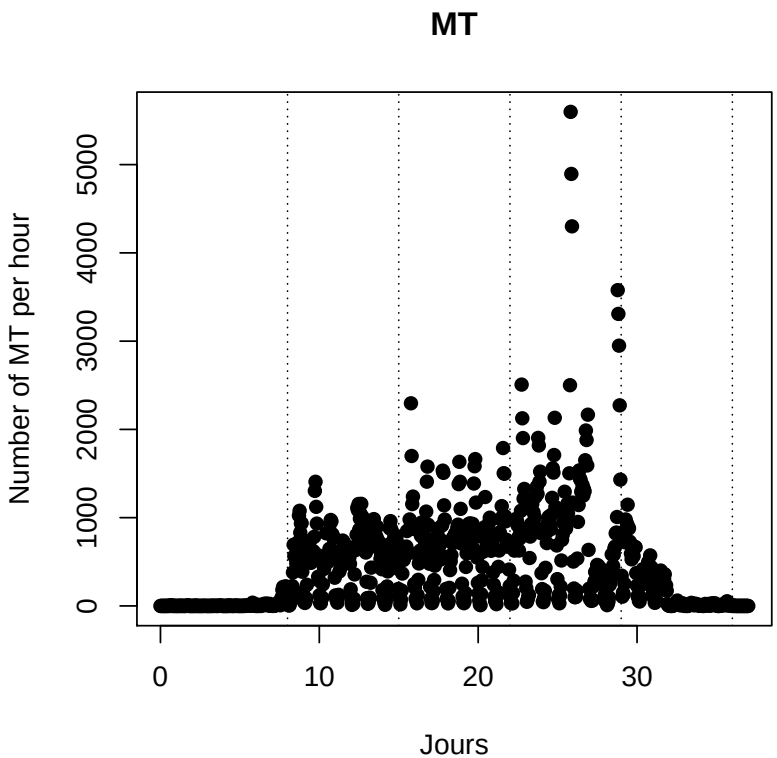




Why to cluster the days ?

$$N^R_{+,j,h} = \sum_{i=1}^{n} N^R_{i,j,h} \quad \Rightarrow \quad N^R_{+,j,.} = \left(\sum_{i=1}^{n} N^R_{i,j,0}, \ldots, \sum_{i=1}^{n} N^R_{i,j,23} \right) \in \mathbb{R}^{24}$$





K-means

$$\mathbf{w} = (w_{jk}) \in \mathcal{M}_{J\times K}(\{0,1\})$$

For example

$$\mathbf{w} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ & & \vdots & \end{pmatrix}$$

The day j is in the cluster k if and only if $w_{jk} = 1$.

K-means

To search $\mathbf{w} = (w_{jk}) \in \mathcal{M}_{J\times K}(\{0,1\})$ such that j and j' are in the same class if

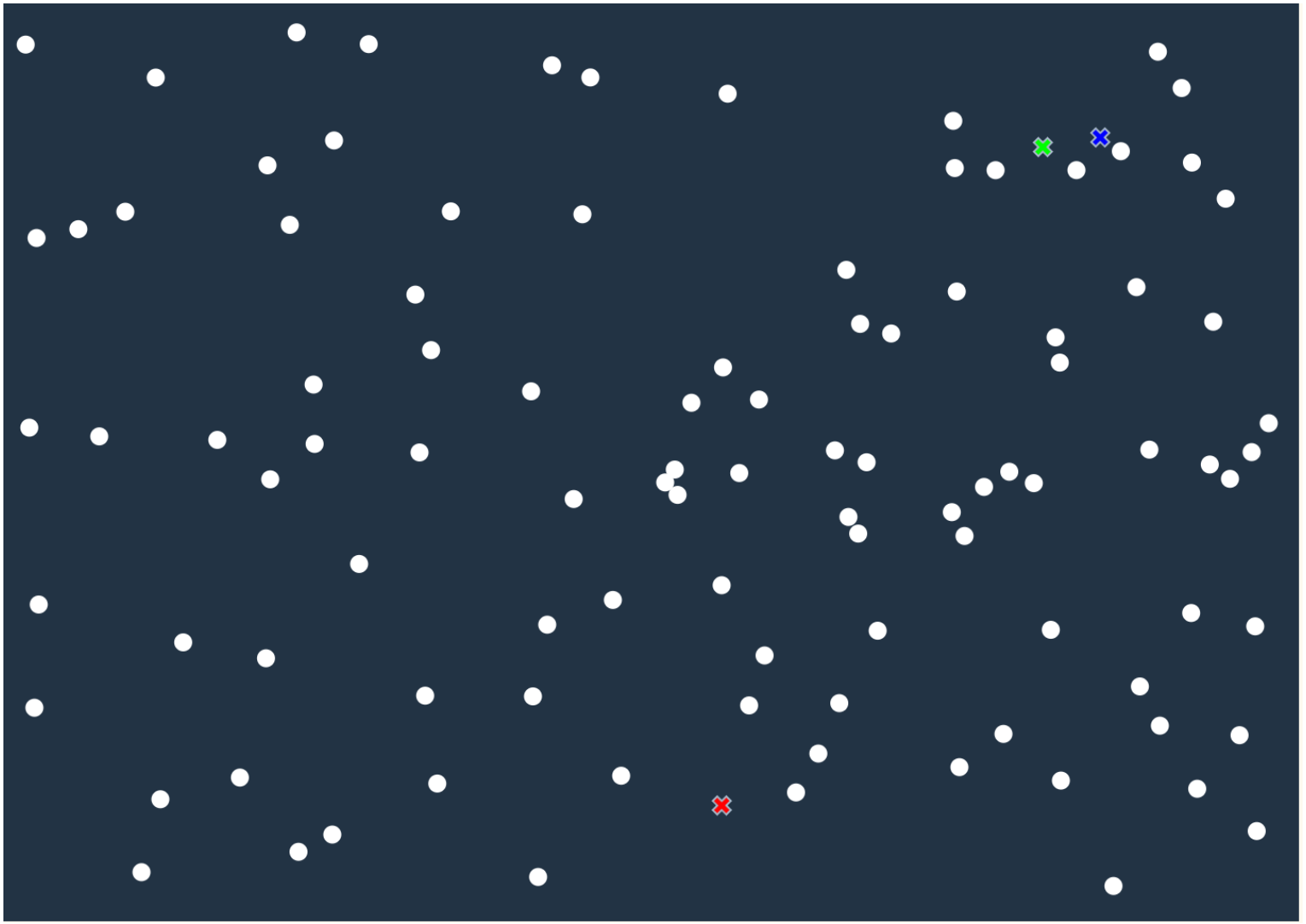
$$\forall h \in \{0,\dots,23\},\ N^R_{+,j,h} \approx N^R_{+,j',h}.$$

For example

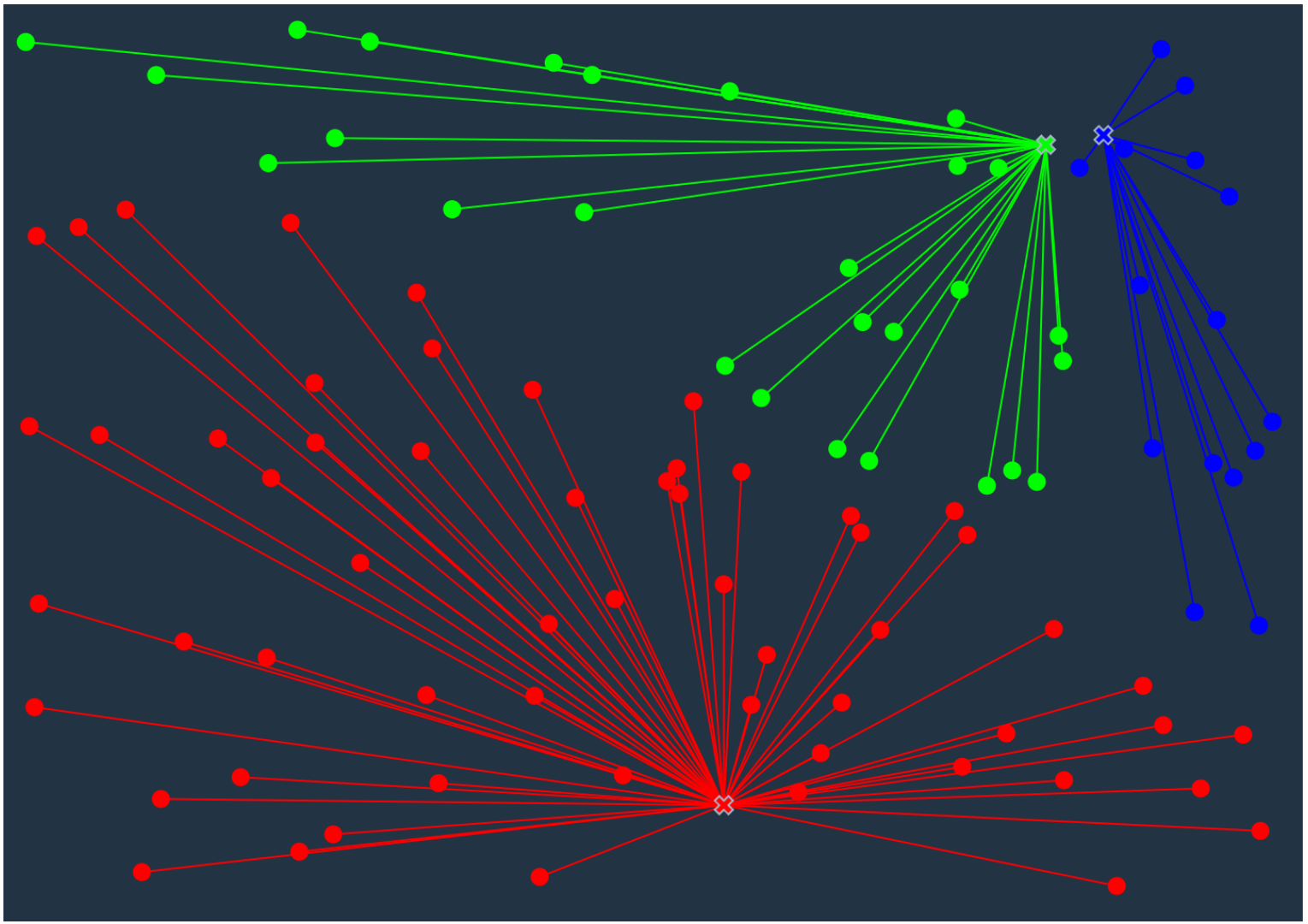
$$\mathbf{w} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ & & \vdots & \end{pmatrix}$$

The day j is in the cluster k if and only if $w_{jk} = 1$.

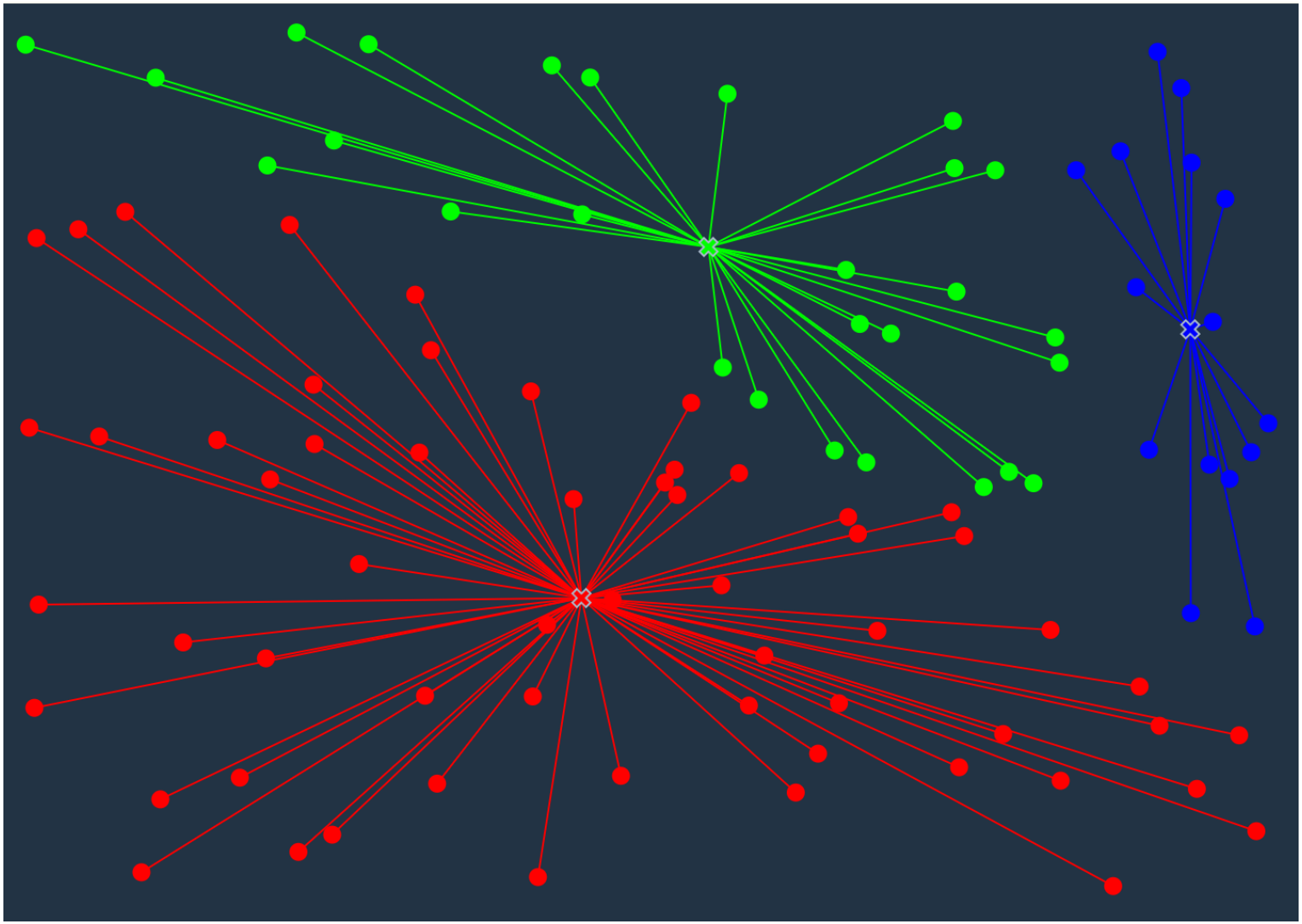
Exemple



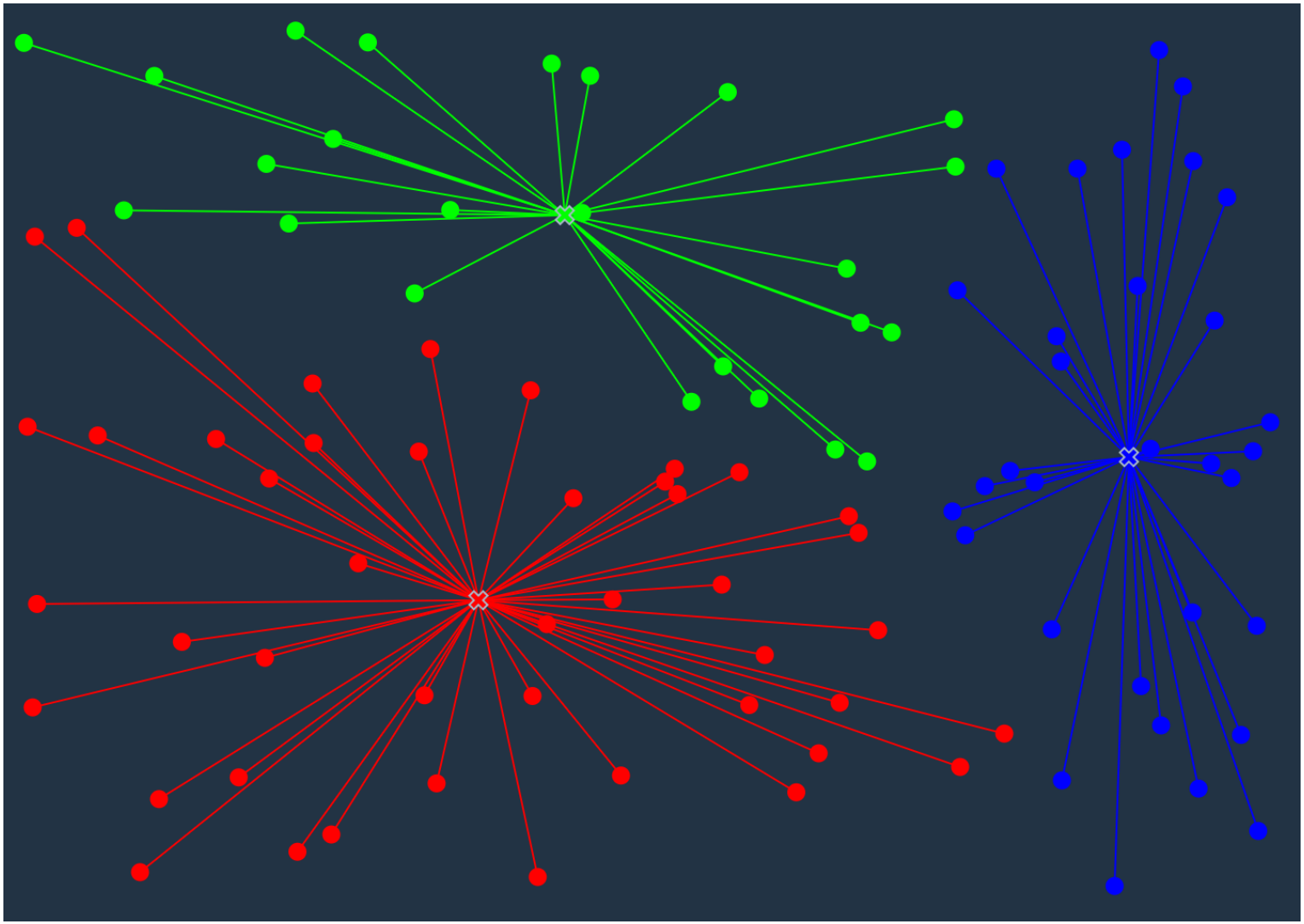
Exemple

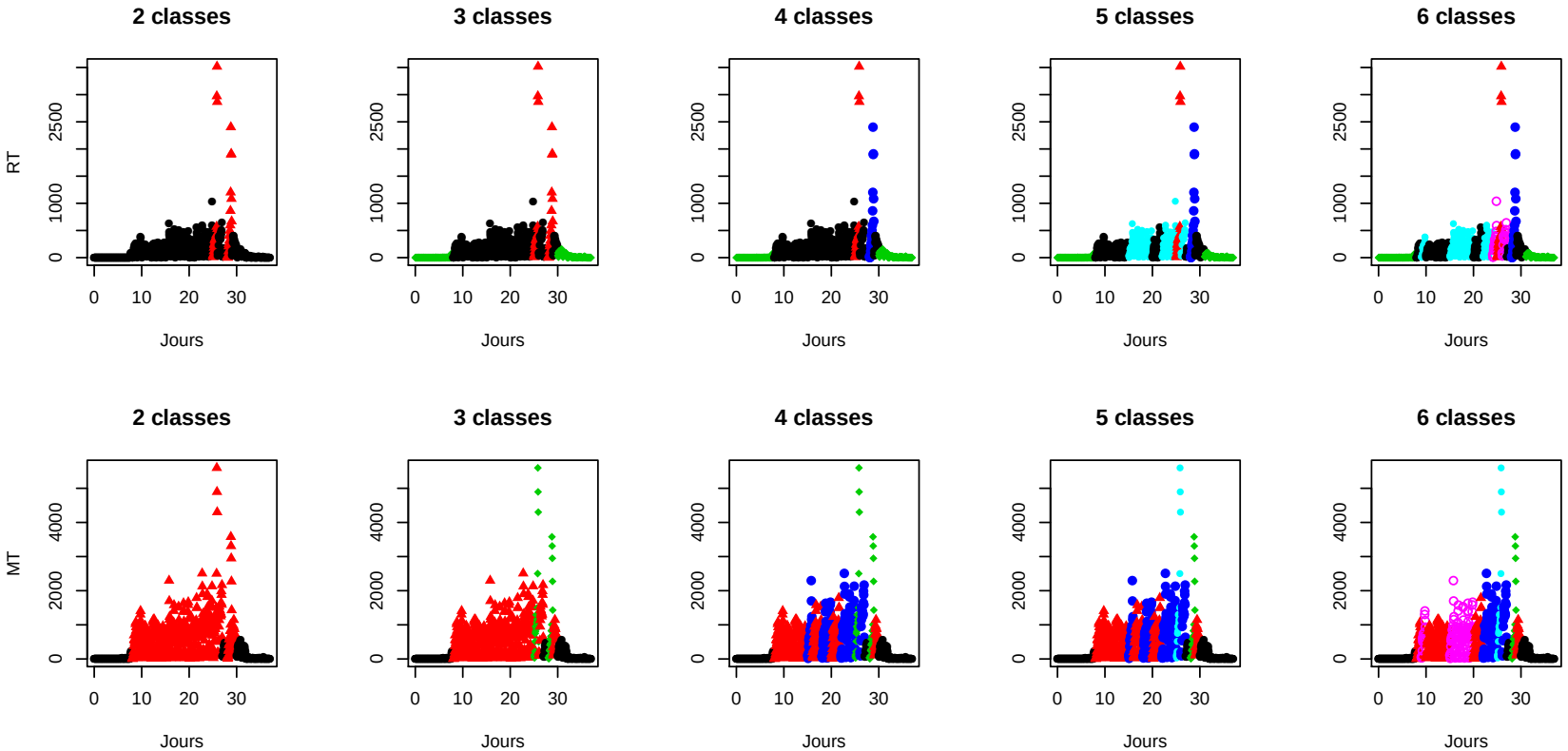


Exemple



Exemple




2 classes
3 classes
4 classes
5 classes
6 classes
RT
MT
Jours
0
10
20
30
1000
2500
2000
4000

Goal

To cluster users such that :

- All users in the same class have the same behaviour.
- Two users in two differents classes have different behaviors.

Goal

To cluster users such that :

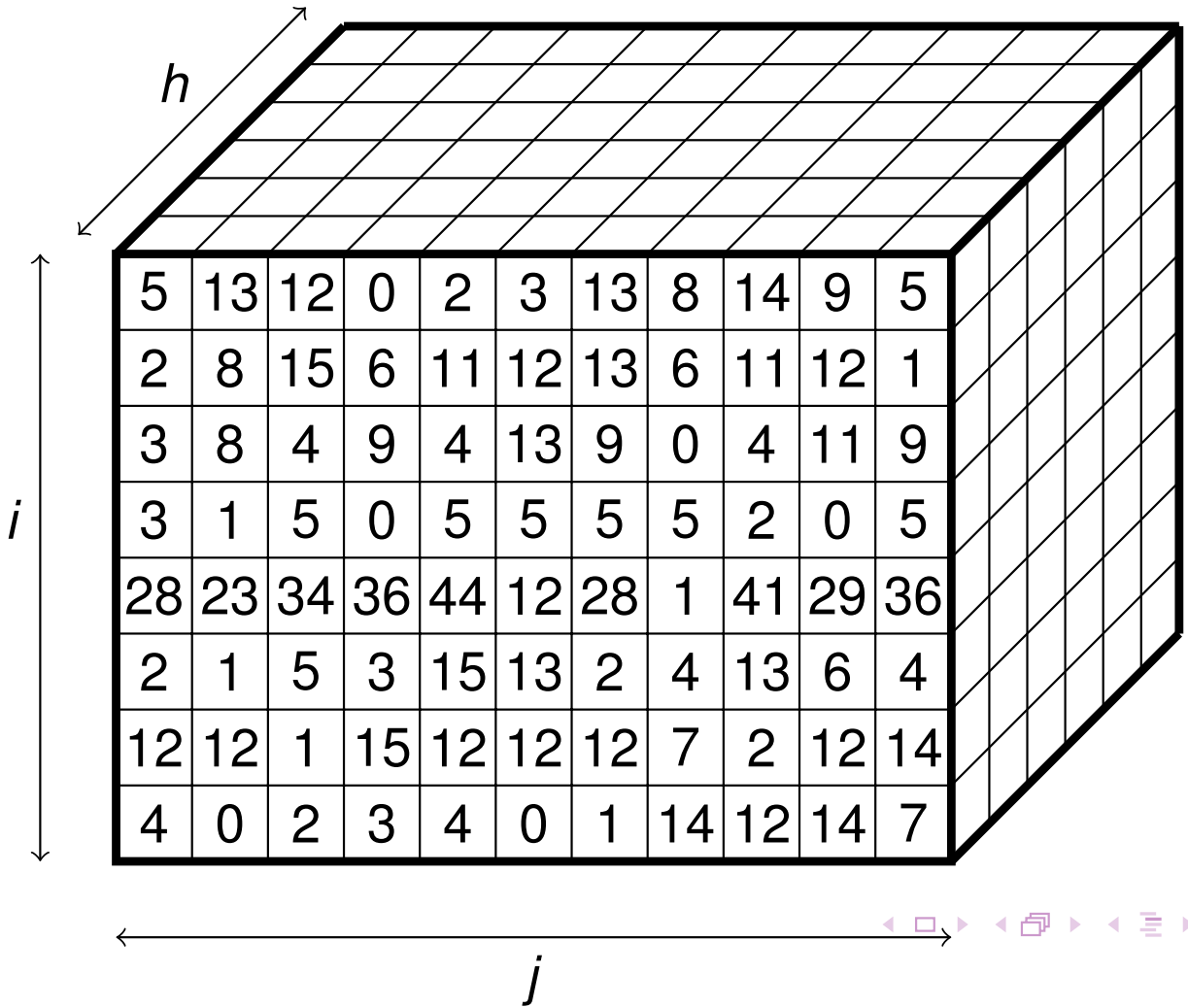
- All users in the same class have the same behaviour.
- Two users in two differents classes have different behaviors.
- Each user has the same behaviour for each day of a day class

Goal

To cluster users such that :

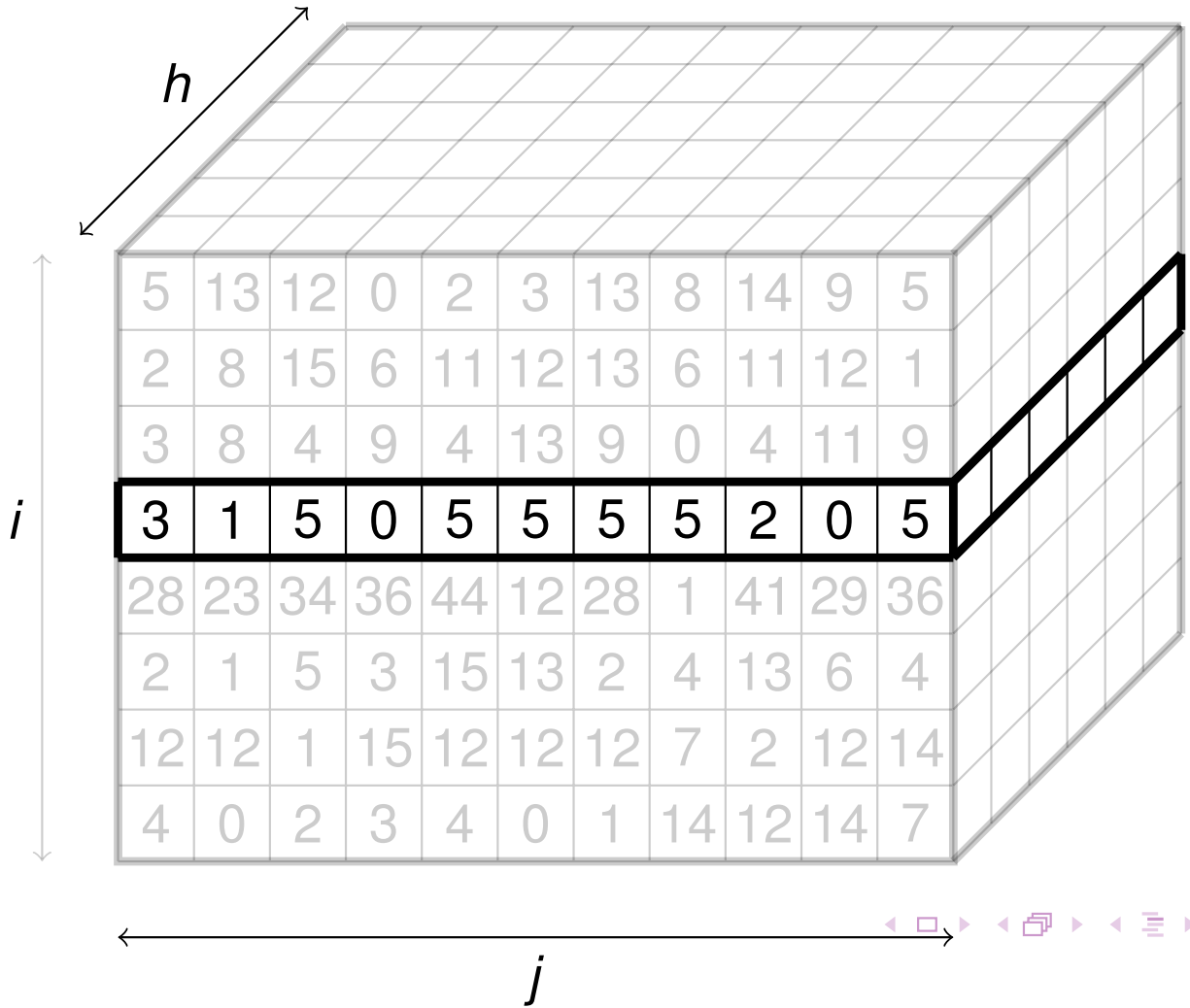
- All users in the same class have the same behaviour.
- Two users in two differents classes have different behaviors.
- Each user has the same behaviour for each day of a day class $\Rightarrow$ using the matrix $\mathbf{w}$.

Notations


h
i
j
5 13 12 0 2 3 13 8 14 9 5
2 8 15 6 11 12 13 6 11 12 1
3 8 4 9 4 13 9 0 4 11 9
3 1 5 0 5 5 5 5 2 0 5
28 23 34 36 44 12 28 1 41 29 36
2 1 5 3 15 13 2 4 13 6 4
12 12 1 15 12 12 12 7 2 12 14
4 0 2 3 4 0 1 14 12 14 7

Notations

$$N^R_{i,\cdot,\cdot} = \left(N^R_{i,j,k}\right)_{(j,k)\in\{1,\dots,J\}\times\{1,\dots,K\}}$$



Mean behavior by day cluster

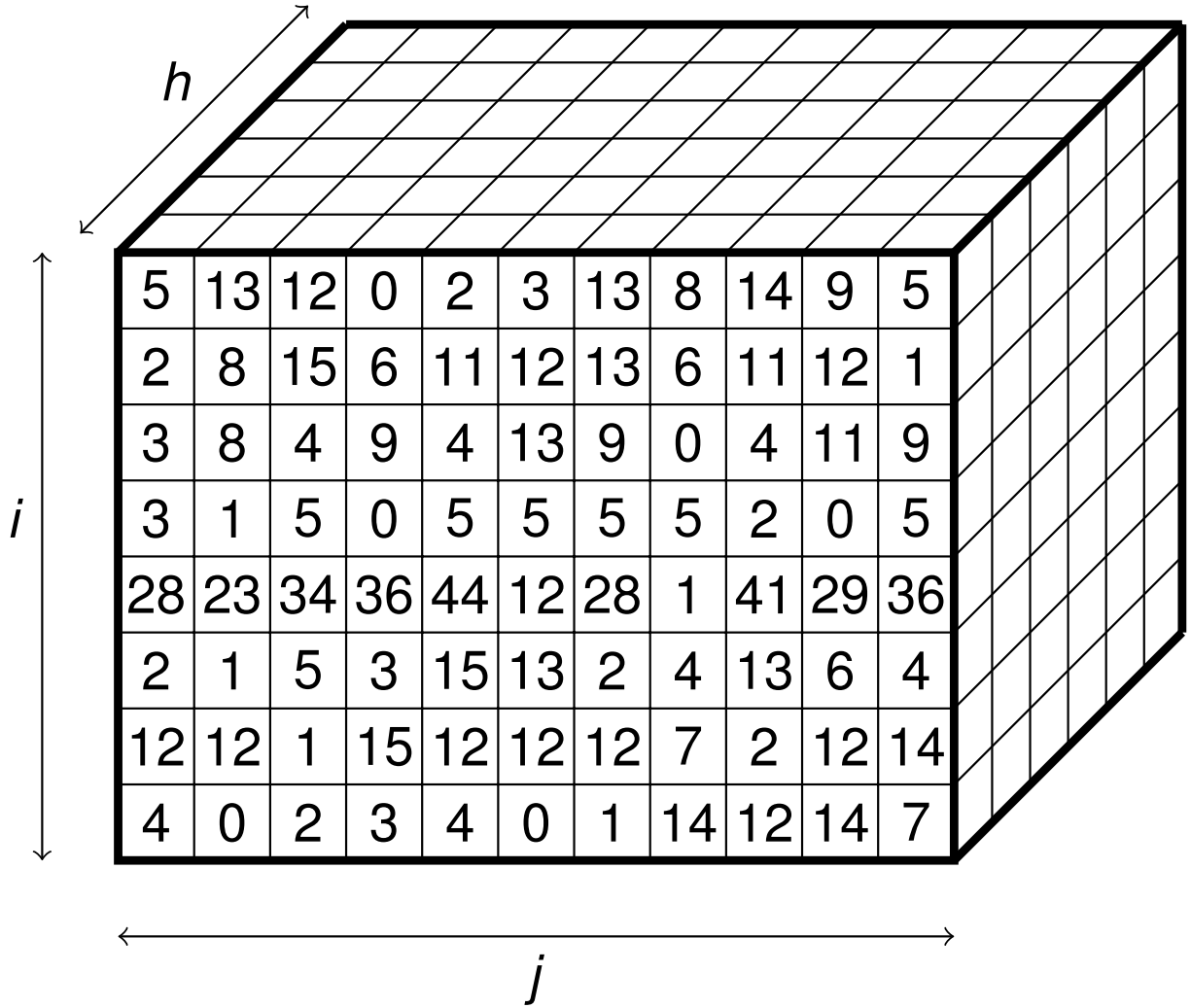
$$\mathbf{c}^i = \left(c^i_{k,k}\right) \in \mathcal{M}_{H\times K}\left(\mathbb{R}\right)$$

Mean behavior by day cluster

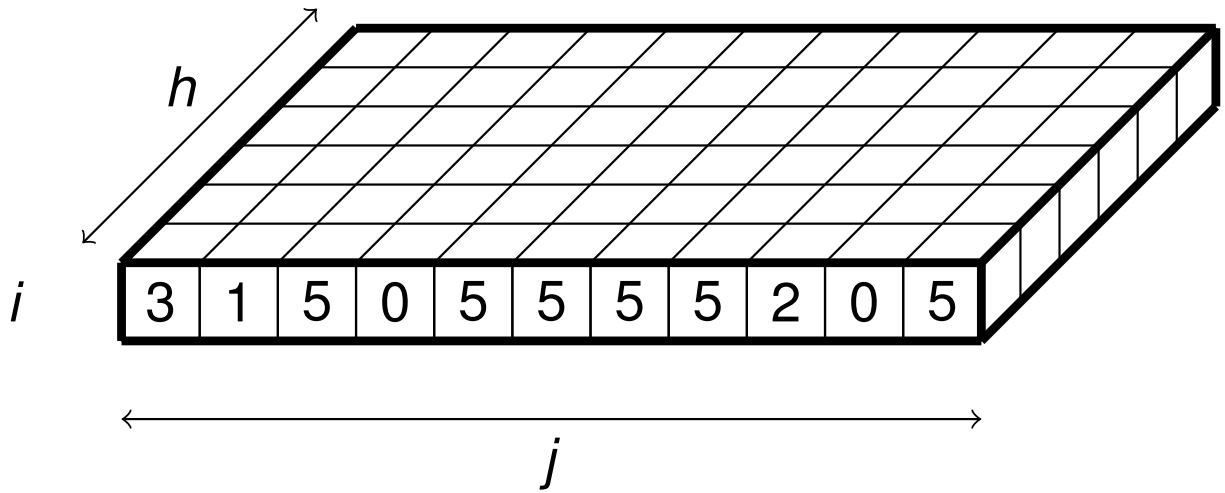
$$\mathbf{C}^i = \left(C^i_{k,k}\right) \in \mathcal{M}_{H\times K}\left(\mathbb{R}\right)$$

$$\mathbf{C}^i = \left(N^R_{i,\cdot,\cdot}\right)^T \overline{\mathbf{w}} = \begin{pmatrix} N^R_{i,1,1} & N^R_{i,2,1} & \cdots & N^R_{i,J,1} \\ N^R_{i,1,2} & N^R_{i,2,2} & & \vdots \\ \vdots & & \ddots & \vdots \\ N^R_{i,1,H} & N^R_{i,2,H} & \cdots & N^R_{i,J,H} \end{pmatrix} \begin{pmatrix} \frac{w_{1,1}}{w_{+,1}} & \frac{w_{1,2}}{w_{+,2}} & \cdots & \frac{w_{1,K}}{w_{+,K}} \\ \frac{w_{2,1}}{w_{+,1}} & \frac{w_{2,2}}{w_{+,2}} & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{w_{J,1}}{w_{+,1}} & \frac{w_{J,2}}{w_{+,2}} & \cdots & \frac{w_{J,K}}{w_{+,K}} \end{pmatrix}.$$

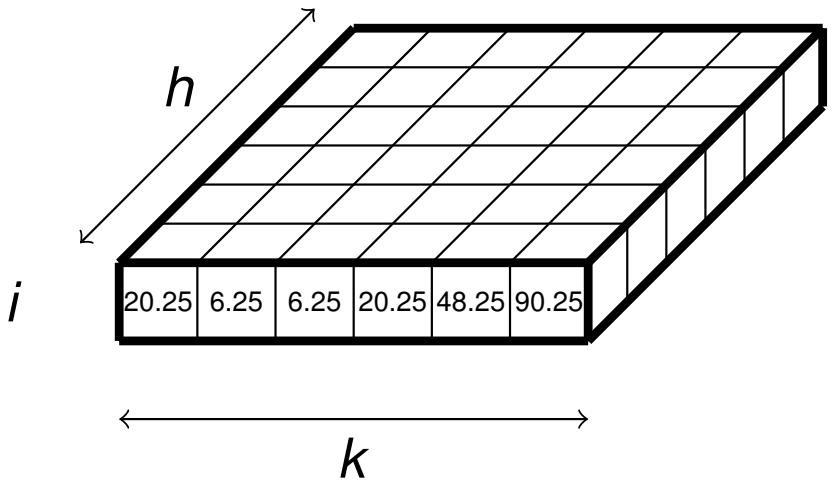
Mean behavior by day cluster



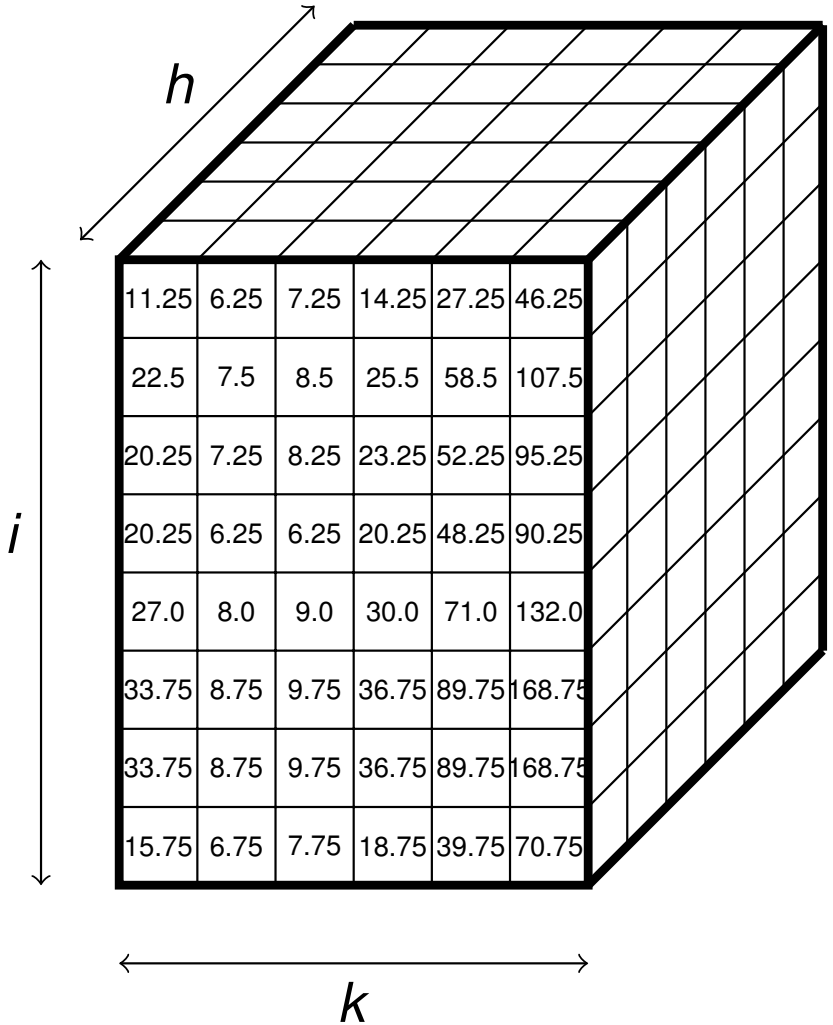
Mean behavior by day cluster



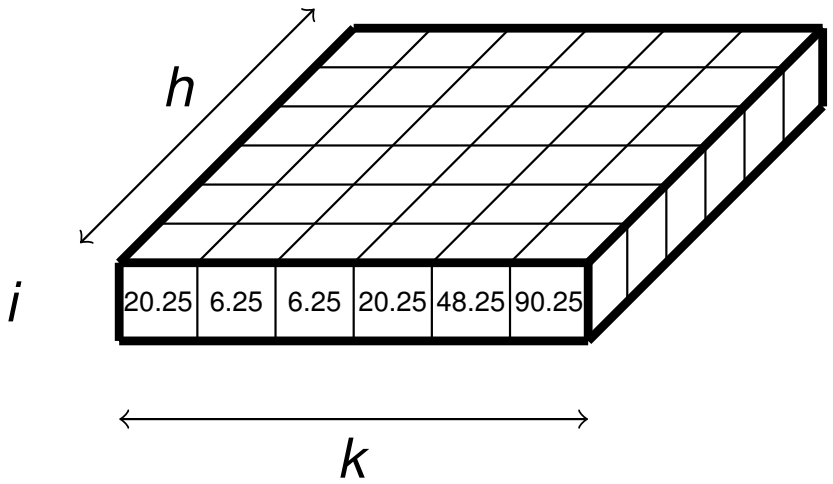
Mean behavior by day cluster



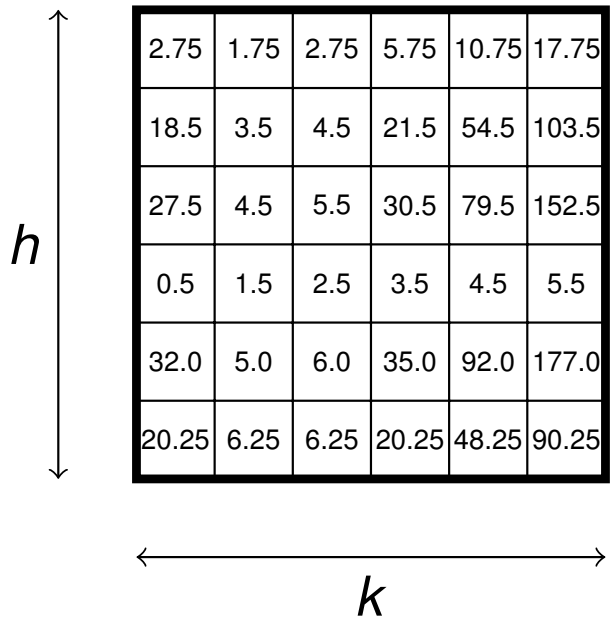
Mean behavior by day cluster



Mean behavior by day cluster

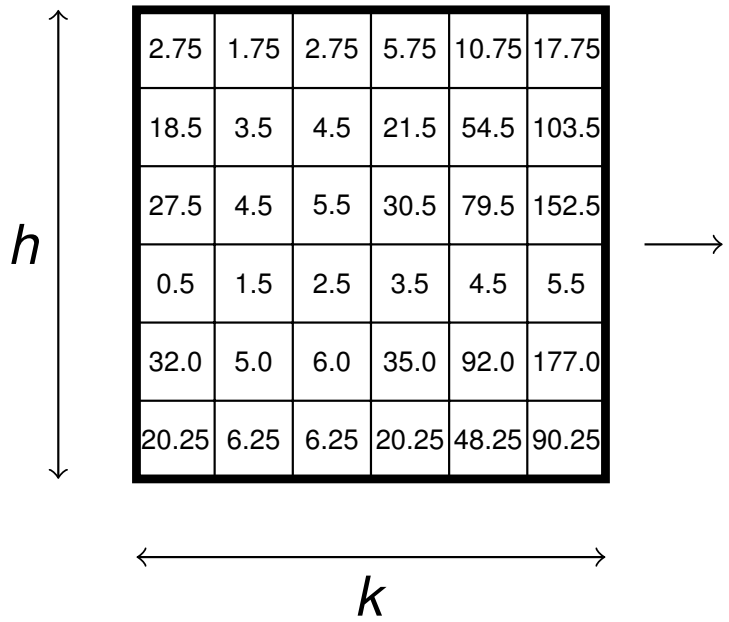


Mean behavior by day cluster



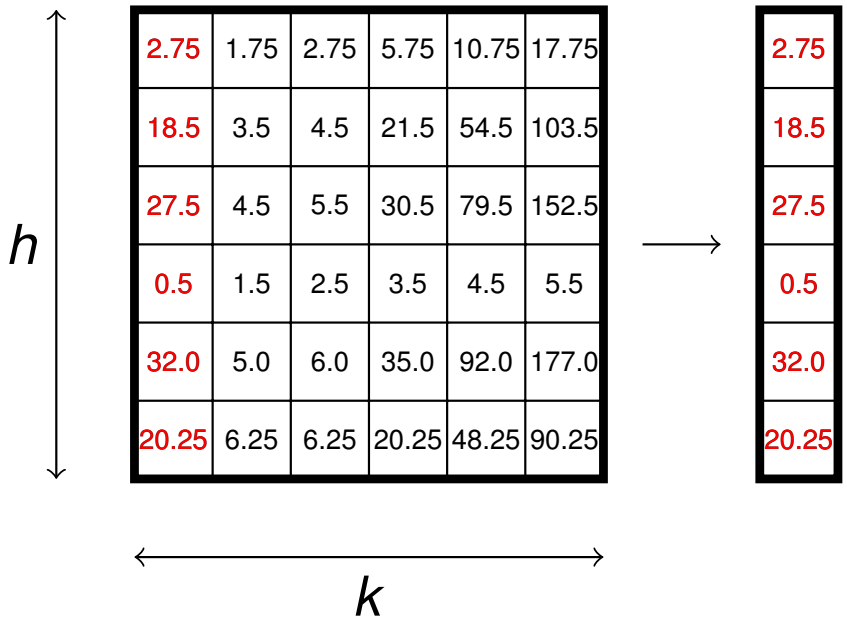
2.75	1.75	2.75	5.75	10.75	17.75
18.5	3.5	4.5	21.5	54.5	103.5
27.5	4.5	5.5	30.5	79.5	152.5
0.5	1.5	2.5	3.5	4.5	5.5
32.0	5.0	6.0	35.0	92.0	177.0
20.25	6.25	6.25	20.25	48.25	90.25

Mean behavior by day cluster

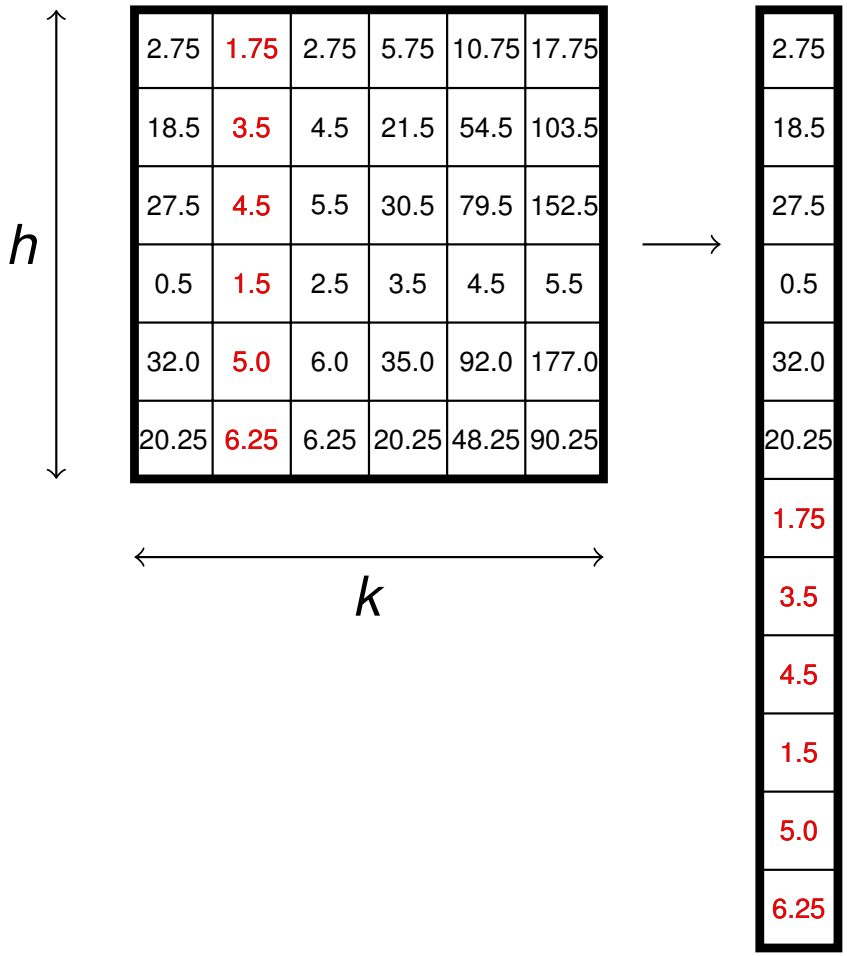


2.75	1.75	2.75	5.75	10.75	17.75
18.5	3.5	4.5	21.5	54.5	103.5
27.5	4.5	5.5	30.5	79.5	152.5
0.5	1.5	2.5	3.5	4.5	5.5
32.0	5.0	6.0	35.0	92.0	177.0
20.25	6.25	6.25	20.25	48.25	90.25

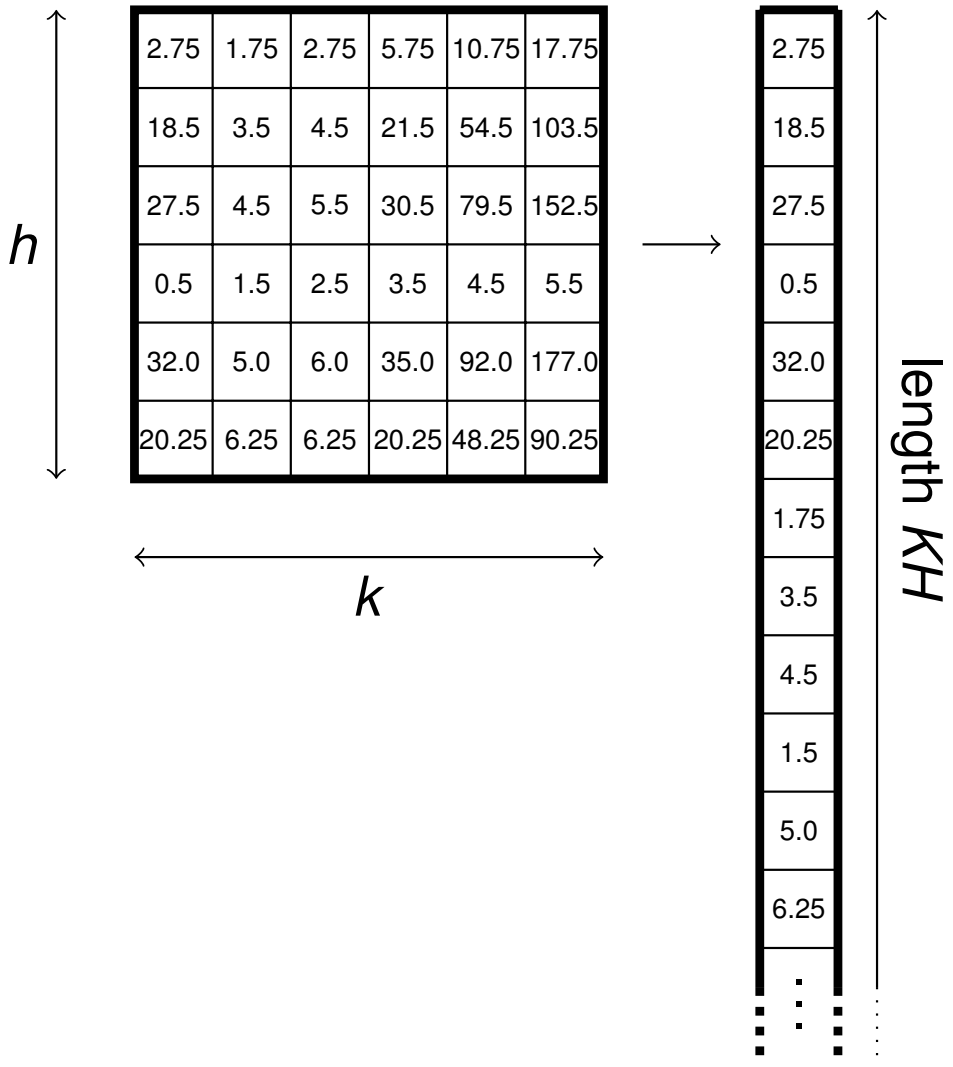
Mean behavior by day cluster



Mean behavior by day cluster



Mean behavior by day cluster



Modelisation

$$\mathbf{C} = \left(\mathrm{Vec}\left(\mathbf{C}^1\right), \ldots, \mathrm{Vec}\left(\mathbf{C}^n\right)\right)$$

but we have the relation

$$\mathrm{Vec}\left(\mathbf{C}^i\right) = \left(Id_H \otimes \overline{\mathbf{w}}^T\right) \mathrm{Vec}\left(N^R_{i,.,.}{}^T\right).$$

Kronecker product

Modelisation

$$\mathbf{C} = \left(\mathrm{Vec}\left(\mathbf{C}^1\right), \ldots, \mathrm{Vec}\left(\mathbf{C}^n\right)\right)$$

but we have the relation

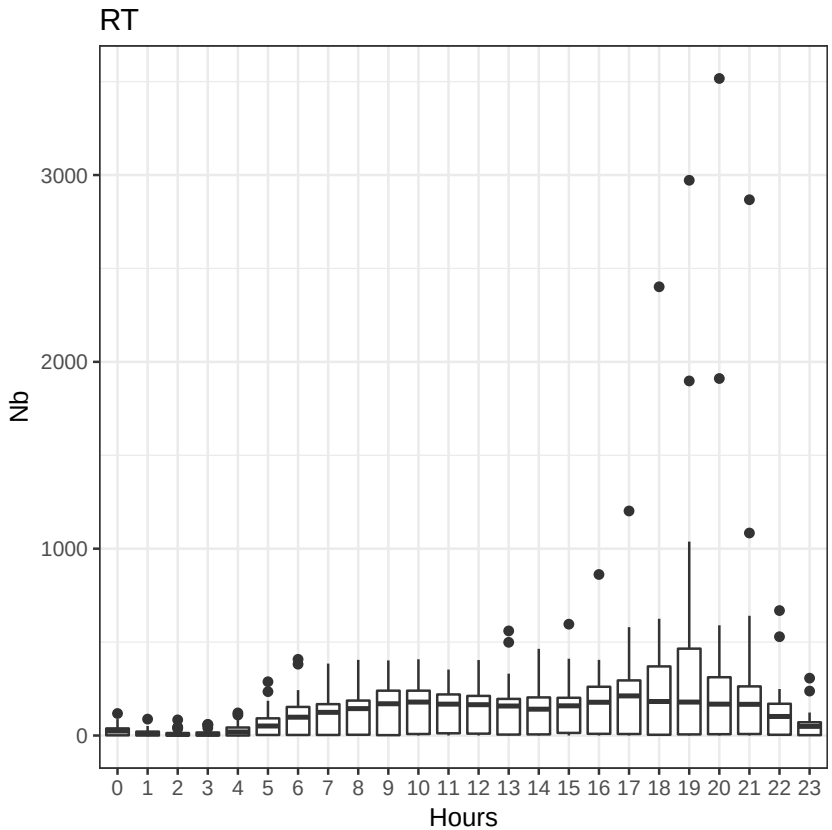
$$\mathrm{Vec}\left(\mathbf{C}^i\right) = \left(Id_H \otimes \overline{\mathbf{w}}^T\right) \mathrm{Vec}\left({N^R_{i,\cdot,\cdot}}^T\right).$$

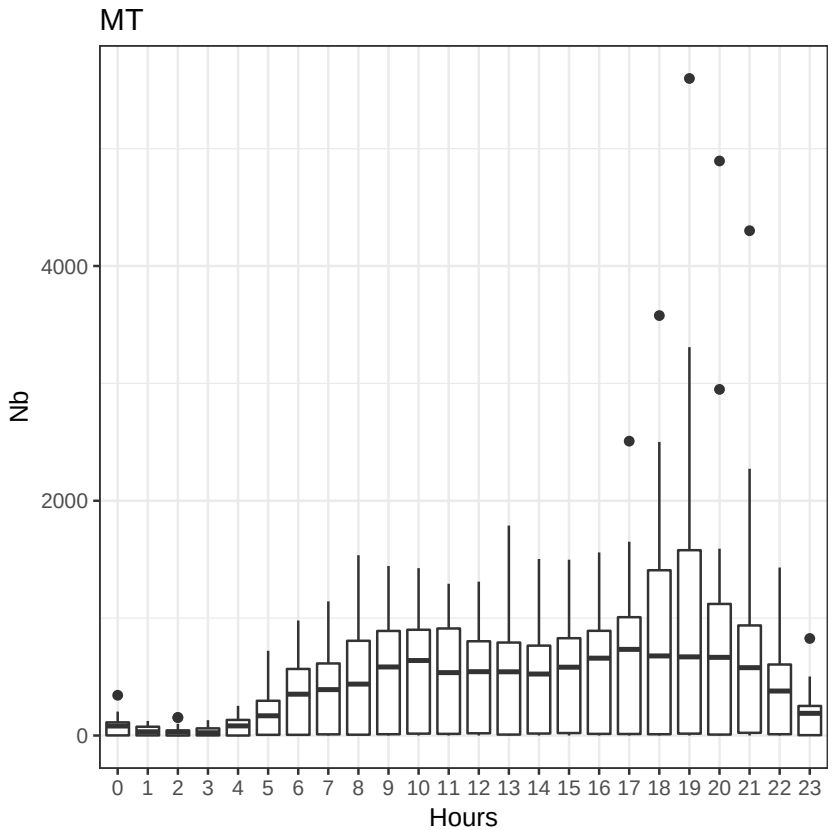
So we have the following relation :

$$\mathbf{C} = \left(Id_H \otimes \overline{\mathbf{w}}^T\right) \left(\mathrm{Vec}\left({N^R_{1,\cdot,\cdot}}^T\right), \ldots, \mathrm{Vec}\left({N^R_{n,\cdot,\cdot}}^T\right)\right).$$

Kronecker product

Heterogeneity of the data





Heterogeneity of the data

- Concatenation of night hours or not : $\mathbf{v} \otimes \overline{\mathbf{w}}^T$ or $Id_H \otimes \overline{\mathbf{w}}^T$.

$$\mathbf{v} = \begin{pmatrix} 1 & 0 & \mathbf{0}_{1\times 19} \\ \mathbf{0}_{4\times 1} & \mathbf{1}_{4\times 1} & \mathbf{0}_{4\times 19} \\ \mathbf{0}_{19\times 1} & \mathbf{0}_{19\times 1} & Id_{19} \end{pmatrix}$$

Heterogeneity of the data

	N° cluster					
	1	2	3	4	5	6
2	35	2				
3	20	2	15			
4	20	1	15	1		
5	12	1	14	1	9	
6	11	1	14	1	8	2

RT

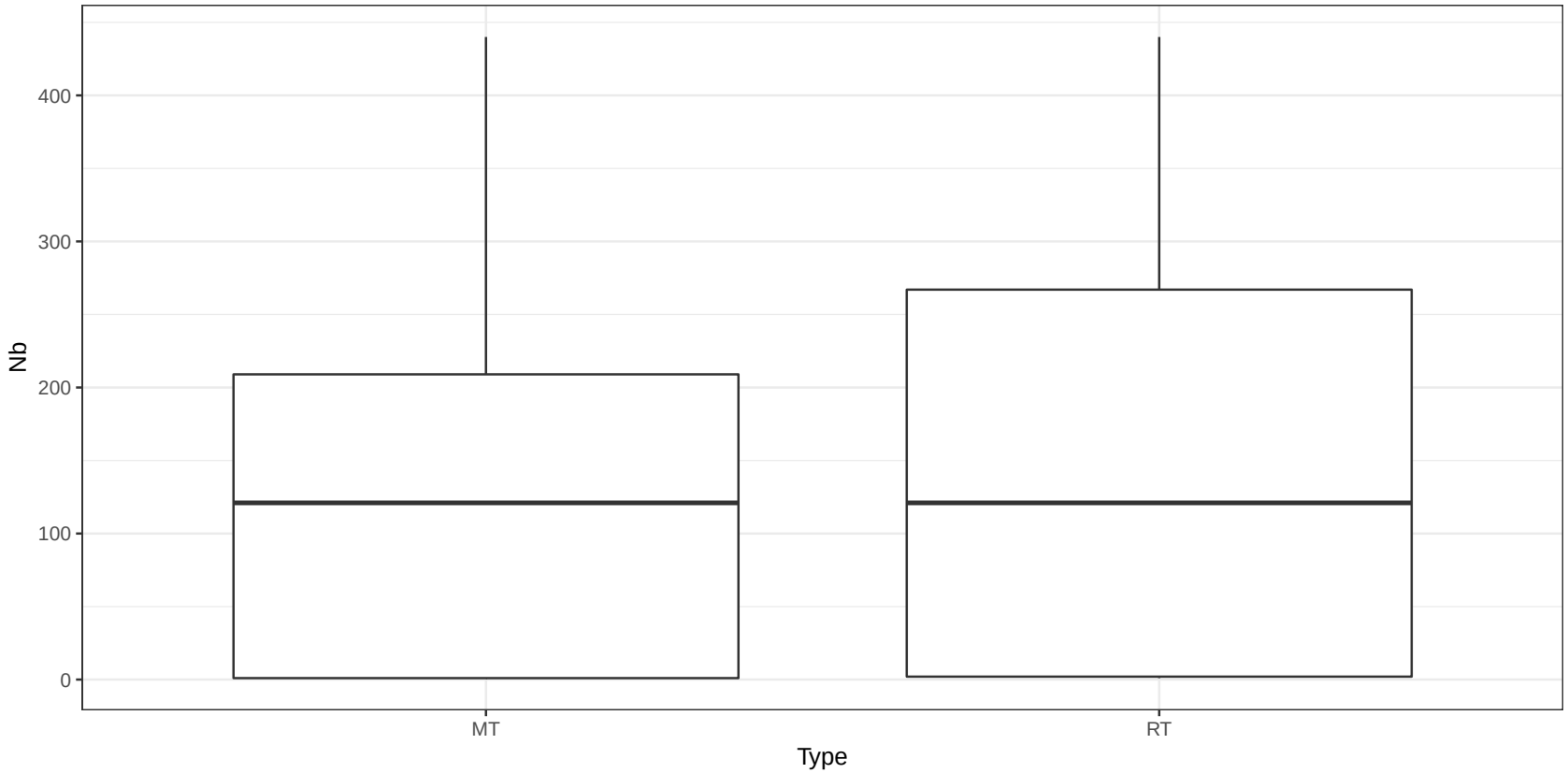
	N° cluster					
	1	2	3	4	5	6
2	16	21				
3	16	19	2			
4	16	12	2	7		
5	16	12	1	7	1	
6	16	9	1	4	1	6

MT

Heterogeneity of the data

- Concatenation of night hours or not : $\mathbf{v} \otimes \overline{\mathbf{w}}^T$ or $Id_H \otimes \overline{\mathbf{w}}^T$.
- Same weight for each cluster or for each day : $Id_H \otimes \overline{\mathbf{w}}^T$ or $Id_H \otimes \mathbf{w}^T$.

Heterogeneity of the data



Heterogeneity of the data

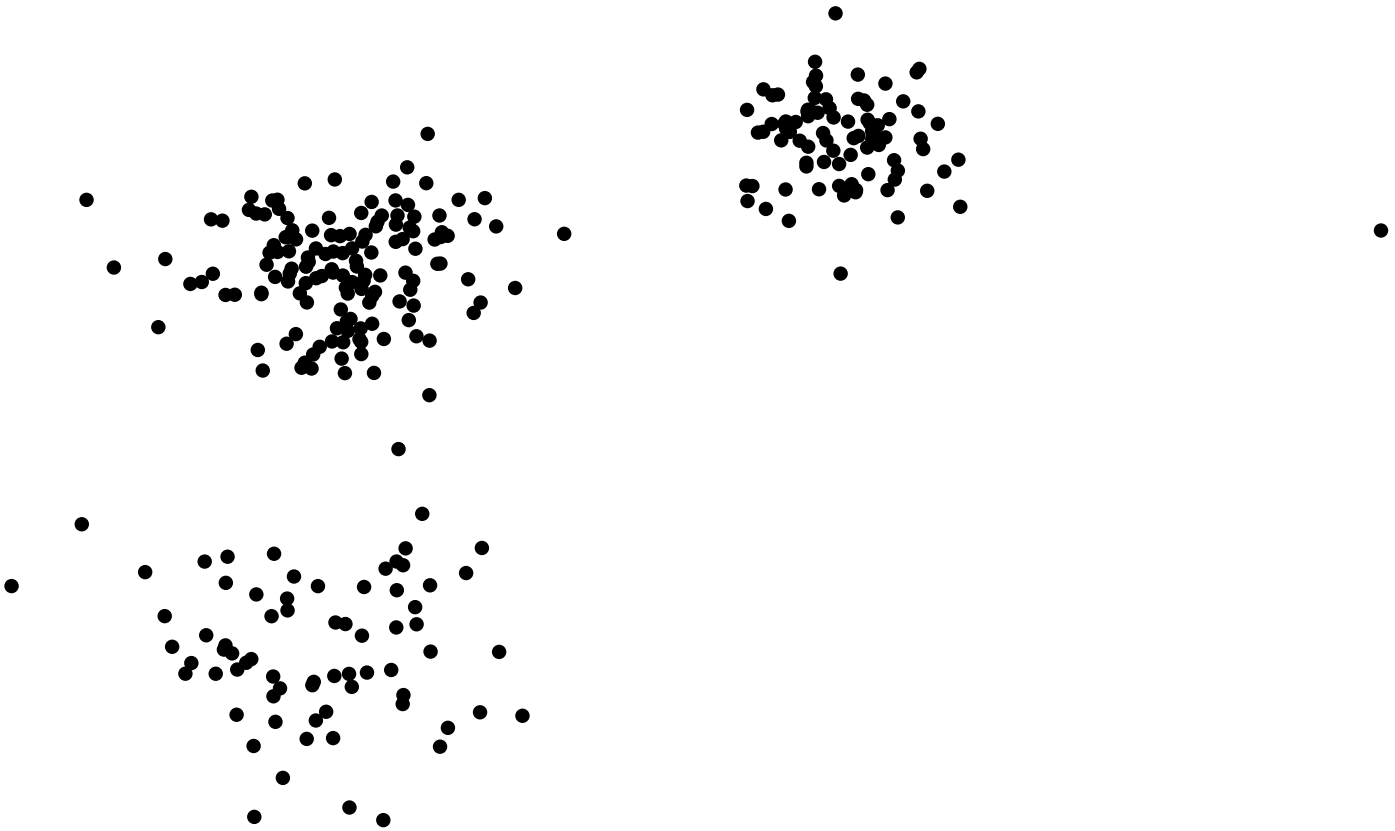
- Concatenation of night hours or not : $\mathbf{v} \otimes \overline{\mathbf{w}}^T$ or $Id_H \otimes \overline{\mathbf{w}}^T$.
- Same weight for each cluster or for each day : $Id_H \otimes \overline{\mathbf{w}}^T$ or $Id_H \otimes \mathbf{w}^T$.
- Raw data or data standardized by the total number per user.

Heterogeneity of the data

- Concatenation of night hours or not : $\mathbf{v} \otimes \overline{\mathbf{w}}^T$ or $Id_H \otimes \overline{\mathbf{w}}^T$.
- Same weight for each cluster or for each day : $Id_H \otimes \overline{\mathbf{w}}^T$ or $Id_H \otimes \mathbf{w}^T$.
- Raw data or data standardized by the total number per user.

16 different configurations

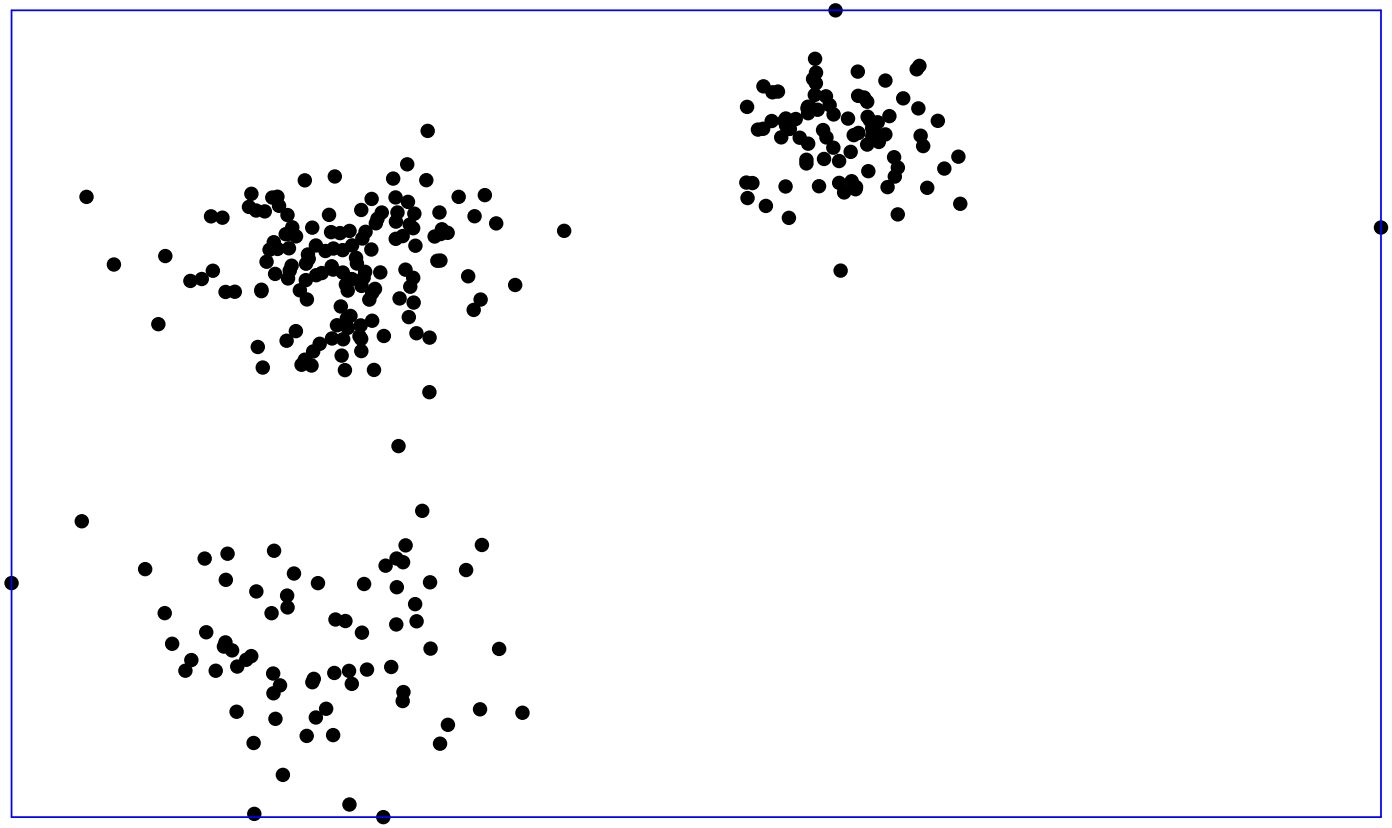
Initialisation for any K



Initialisation for any K

- Basic

Initialisation for any K

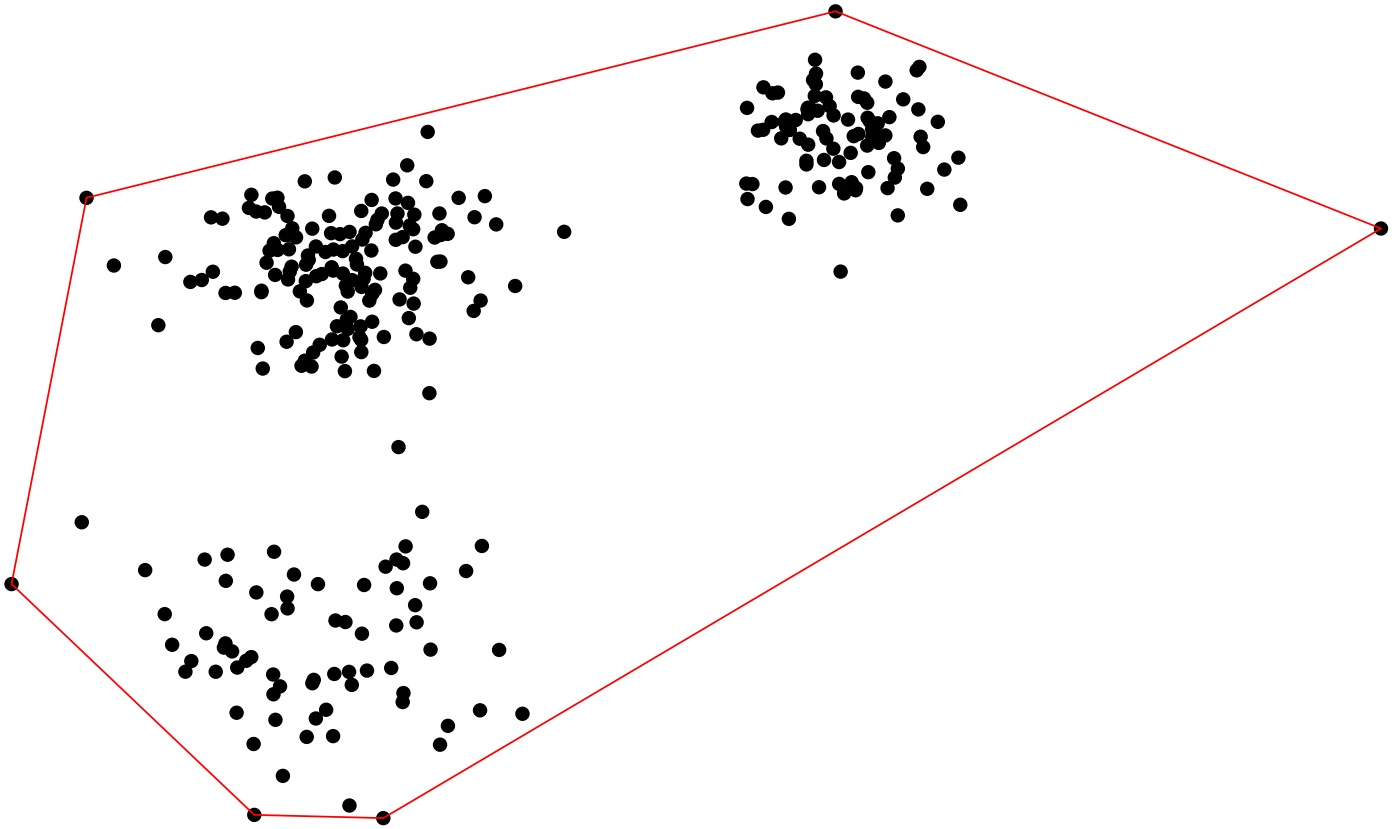


Initialisation for any K

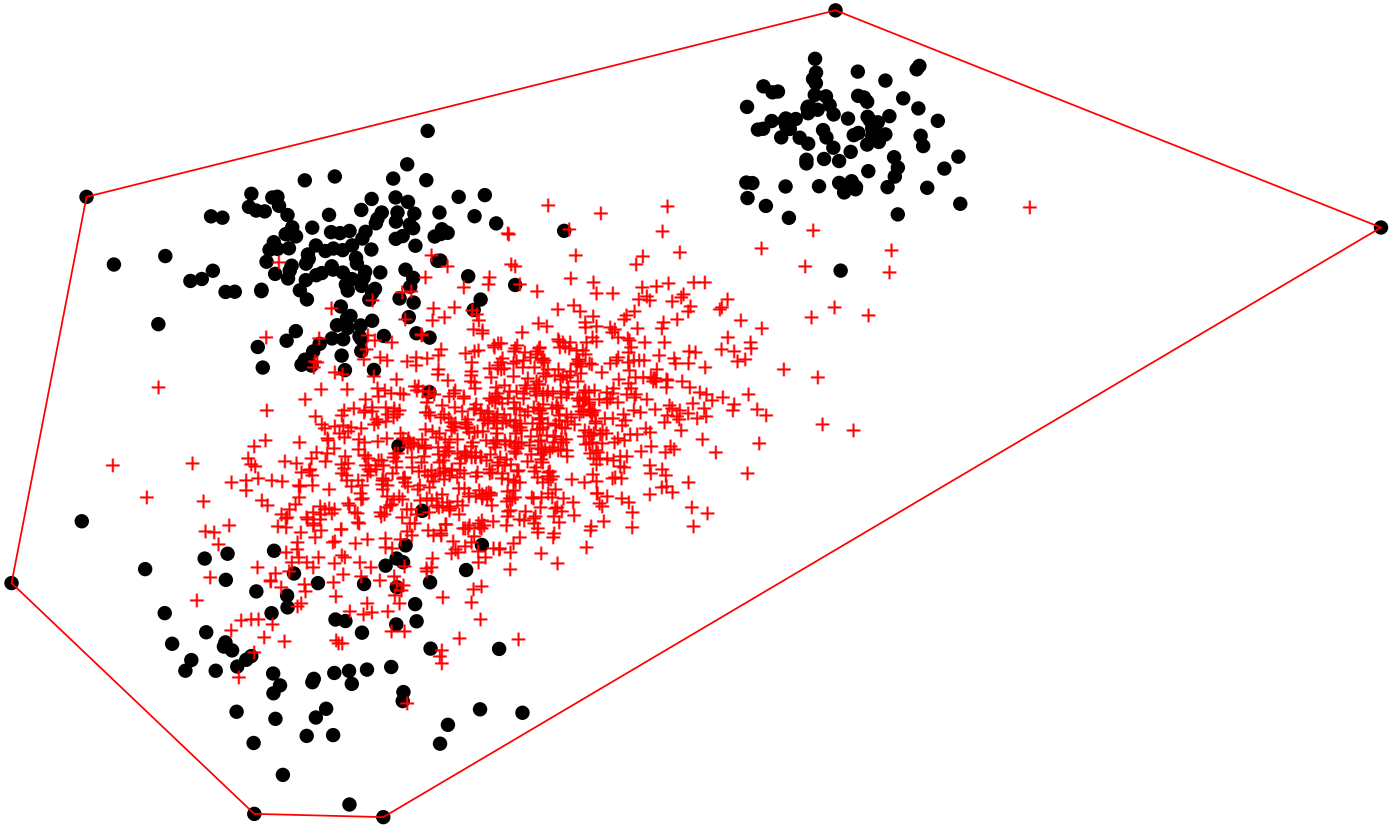
- Basic
- Convex envelope

$$\left\{ \sum_{i=1}^{n} \lambda_i x_i \middle| 0 \leq \lambda_i \leq 1 \text{ and } \sum_{i=1}^{n} \lambda_i = 1 \right\}$$

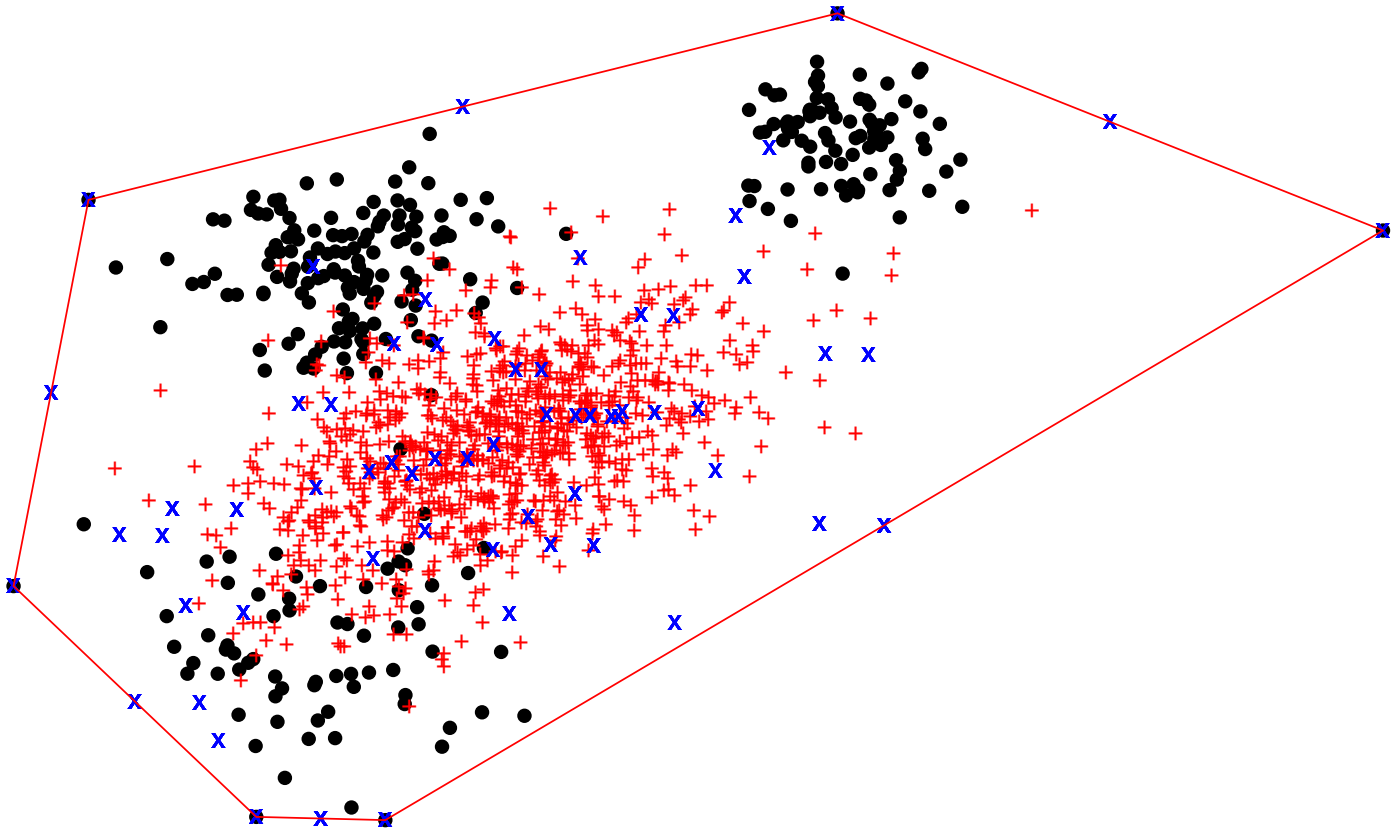
Initialisation for any K



Initialisation for any K



Initialisation for any K

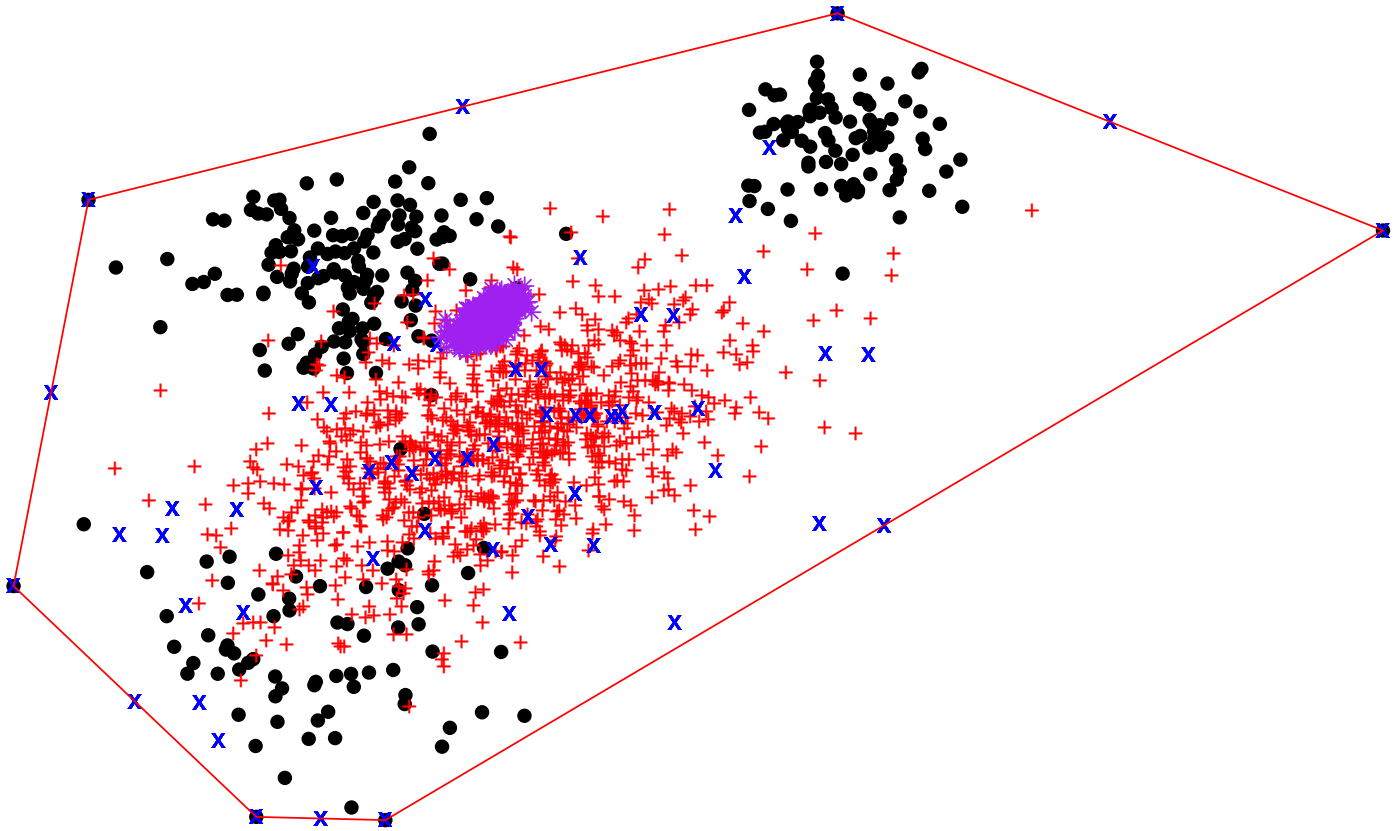


Initialisation for any K

- Basic
- ~~Convex envelope~~ $\Rightarrow$ weight on every observations

$$\left\{ \sum_{i=1}^{n} \lambda_i x_i \middle| 0 \leq \lambda_i \leq 1 \text{ and } \sum_{i=1}^{n} \lambda_i = 1 \right\}$$

Initialisation for any K



Initialisation for any K

- Basic
- ~~Convex envelope~~ $\Rightarrow$ weight on every observations

$$\left\{ \sum_{i=1}^{n} \lambda_i x_i \middle| 0 \leq \lambda_i \leq 1 \text{ and } \sum_{i=1}^{n} \lambda_i = 1 \right\}$$

- Among the observations

Initialisation for any K

- Basic
- ~~Convex envelope~~ $\Rightarrow$ weight on every observations

$$\left\{ \sum_{i=1}^{n} \lambda_i x_i \middle| 0 \leq \lambda_i \leq 1 \text{ and } \sum_{i=1}^{n} \lambda_i = 1 \right\}$$

- Among the observations

If there is a configuration for $K - 1$

- Cutting a class

Initialisation for any K

- Basic
- ~~Convex envelope~~ $\Rightarrow$ weight on every observations

$$\left\{ \sum_{i=1}^{n} \lambda_i x_i \middle| 0 \leq \lambda_i \leq 1 \text{ and } \sum_{i=1}^{n} \lambda_i = 1 \right\}$$

- Among the observations

If there is a configuration for $K - 1$

- Cutting a class
- New class with one observation

Luke

- 16 configurations.
- 9 day clusters.
- K between 2 and 100.
- 5 initialisations.
- Several launches.
- 3 databases.
- K-means complexity : $\mathcal{O}\left(N_{\text{algo}} nK\right)$.

Results

Shiny application

Plan

Prospect

- Mixture model with Poisson or Binomial Negative distributions.
- Mixed effects model for the users.
- Model selection.

Brault Vincent
@Lionning13

#Thank
@you_all

10:11 - 1 juin 2018

1 J'aime

1

Ajouter un autre Tweet

Bibliographie

C. Bouveyron, P. Latouche, and R. Zreik. The stochastic topic block model for the clustering of vertices in networks with textual edges. Statistics and Computing, pages 1–21, 2016.

A. Bruns. Journalists and twitter : How australian news organisations adapt to a new medium. Media International Australia, 144(1) : 97–107, 2012.

J. A. Hartigan. Clustering algorithms. John Wiley & Sons, Inc., New York, NY, USA, 99th edition, 1975. ISBN 047135645X.

F. Heinderyckx. Obama 2008 : l'inflexion numérique. Hermès, La Revue, (1) :135–136, 2011.

Plan

Let $\mathbf{A} \in \mathcal{M}_{n\times m}(\mathbb{R})$ and $\mathbf{B} \in \mathcal{M}_{p\times q}(\mathbb{R})$ two matrices, the kronecker product of $\mathbf{A}$ and $\mathbf{B}$ is a matrix $(np) \times (mq)$ satisfying :

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}\mathbf{B} & a_{n2}\mathbf{B} & \cdots & a_{nm}\mathbf{B} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}b_{11} & a_{11}b_{12} & \cdots & a_{11}b_{1q} & a_{12}b_{11} & \cdots & \cdots & a_{1m}b_{11} & \cdots & a_{1m}b_{1q} \\ a_{11}b_{21} & a_{11}b_{22} & \cdots & a_{11}b_{2q} & a_{12}b_{21} & \cdots & \cdots & a_{1m}b_{21} & \cdots & a_{1m}b_{2q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & & & \vdots & \ddots & \vdots \\ a_{11}b_{p1} & a_{11}b_{p2} & \cdots & a_{11}b_{pq} & a_{12}b_{p1} & \cdots & \cdots & a_{1m}b_{p1} & \cdots & a_{1m}b_{pq} \\ \vdots & \vdots & & \vdots & \vdots & \ddots & & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & \vdots & & \ddots & \vdots & & \vdots \\ a_{n1}b_{11} & a_{n1}b_{12} & \cdots & a_{n1}b_{1q} & a_{n2}b_{11} & \cdots & \cdots & a_{nm}b_{11} & \cdots & a_{nm}b_{1q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & & & \vdots & \ddots & \vdots \\ a_{n1}b_{p1} & a_{n1}b_{p2} & \cdots & a_{n1}b_{pq} & a_{n2}b_{p1} & \cdots & \cdots & a_{nm}b_{p1} & \cdots & a_{nm}b_{pq} \end{pmatrix}.$$

Return vectorisation