

# *Advanced Human Machine Interaction*

## **(Interaction) Data Analysis**

**Alexandre Pauchet**

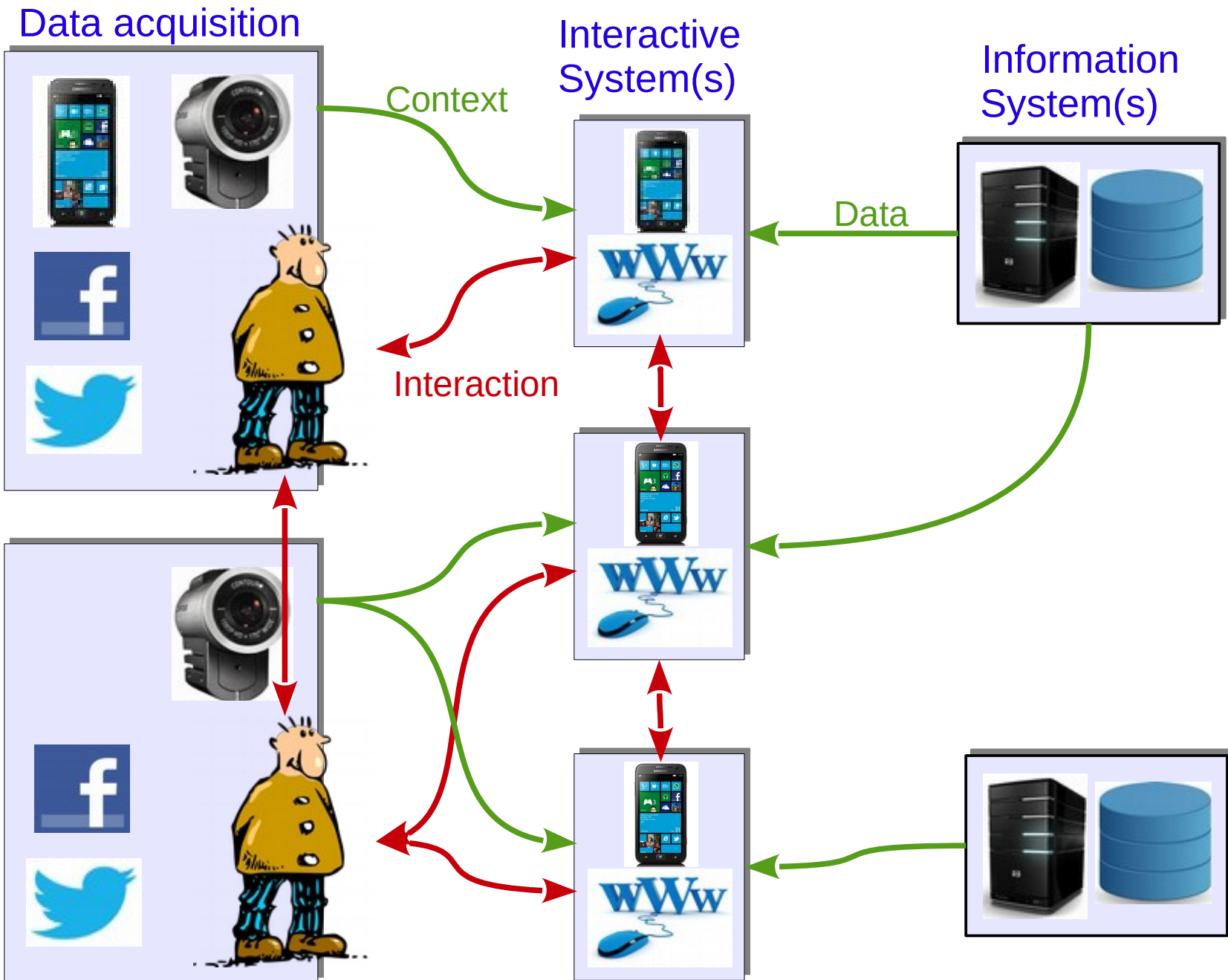
[alexandre.pauchet@insa-rouen.fr](mailto:alexandre.pauchet@insa-rouen.fr) - BO.B.RC18



Normandie Université



# IHME/IDA: objectives





---

# Problem modeling

# Formal modeling of a problem

---

- **Define formally the inputs**

Ex: “*input =  $\{a_1, \dots, a_n\}_{n < 20}$  a sequence of actions with each action  $a_i$  in  $\{up, down, left, right, space\}$ ”*”

- **Define formally the outputs**

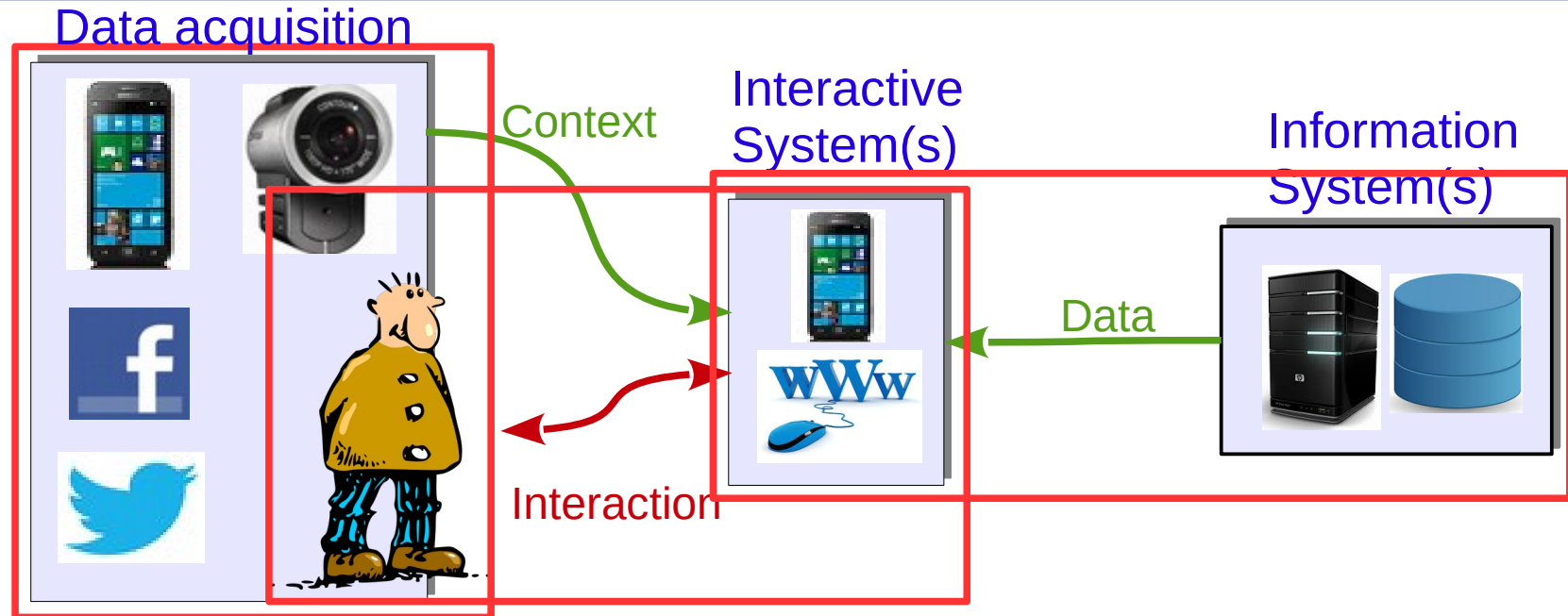
Ex: “*output = a class among  $\{beginner, advanced\}$ ”*”

Ex: “*output = a sequence of words generated as an answer to the user’s query  $\{w_1, \dots, w_n\}_{n > 0}$ ”*”

- **Define a class of problem / algorithm**

- Ex (machine learning): classification problem, clustering problem, regression problem, ...
- Ex (logic): induction / deduction problem, ...
- Ex (algorithm): sorting, graph construction, ...

# Analysis of interaction data



The data to analyze can be collected from:

- External/internal sensors (context of use)
- Interaction data (log of user actions and/or activity)
- External data requested from the interactive system

# Input: discrete data (1/2)

---

- **Discrete sequences**

- Ex: sequence of visited web pages, of user actions
- Representation: ordered set  $\{a_1, a_2, \dots, a_n\}$

- **Discrete sequences of item-sets**

- Ex: actions from left and right hands in a double-handed game or multiple paddles/keyboards
- Representations:

- ordered set of item-sets

$$\{\{a_{1,1}, \dots, a_{m,1}\}, \{a_{1,2}, \dots, a_{m,2}\}, \dots, \{a_{1,n}, \dots, a_{m,n}\}\}$$

- matrix

$$\{a_{i,j}\}_{i=1..m; j=1..n}$$

# Input: discrete data (2/2)

---

- **Independent sequences**

- Ex: actions from two different players
- Representation: set of ordered sets
- $\{\{a_{1,1}, \dots, a_{m1,1}\}, \{a_{1,2}, \dots, a_{m2,2}\}, \dots, \{a_{1,n}, \dots, a_{mn,n}\}\}$

## Remarks:

- Time is not considered and delays may be different (or not) between two actions
- Different sampling are possible in independent sequences



# Input: continuous and mixed data

---

- **Continuous signal**

- Ex: user's voice volume
- Representation: function  $f(t)$

- **Continuous signals**

- Ex: trajectories of two Wiimotes
- Representation: set of functions  $\{f_1(t), \dots, f_n(t)\}$

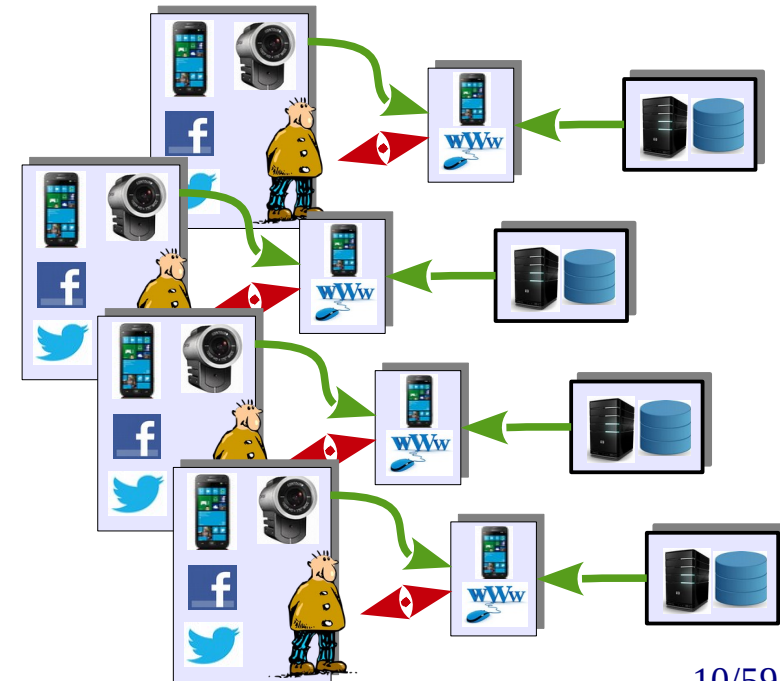
Remarks:

- Any continuous signal can be discretized
- Mix of discrete sequences and continuous signals
- There always is sampling

# Goal / Objective (class of problem)

---

- **Behavioral pattern extraction**
  - Intra- / Inter- user behavioral patterns
  - (Multiple) Time scaling
  - Frequent patterns / similar patterns
- **User classification**
  - Clustering of users' activity
  - Segmentation and classification of parts of users' activity
- **Generation**
  - Simulation of user's behavior
  - Generation of interactive behavior
- **Combination of problems**



# Exercise

---

- Let  $A = \{a_1, \dots, a_n\}$  be the set of  $n$  actions that users can perform. Let  $T$  be a set of  $r$  interaction traces from  $s$  different users  $u_1, \dots, u_s$ :

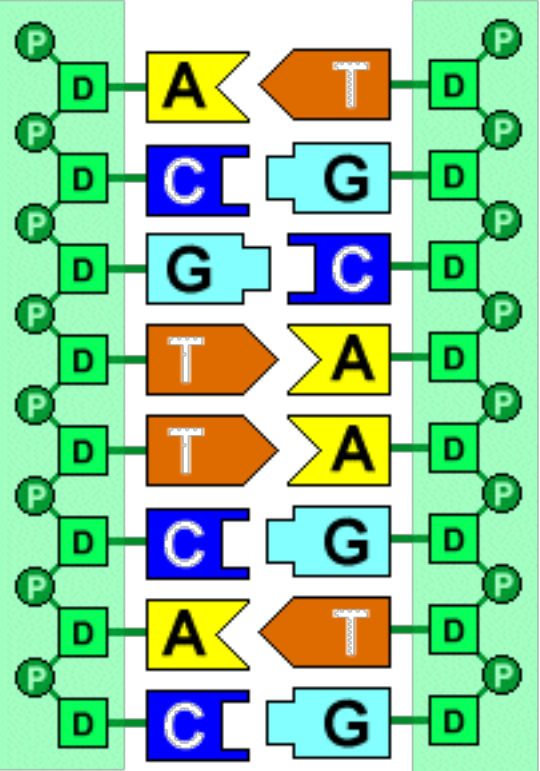
$$T = \{ \{a_{1,x_1}, \dots, a_{m_1,x_1}\}_{u_{x_1}} \dots \{a_{1,m_r}, \dots, a_{m_r,x_r}\}_{u_{x_r}} \}$$

- eg. HMI with 3 possible actions:  $\{ \{a_1 a_2 a_1 a_3\}_{u_1} \{a_2 a_2 a_2 a_1 a_3\}_{u_2} \{a_2 a_1 a_2 a_2 a_1 a_1 a_3\}_{u_1} \{a_1 a_1 a_3\}_{u_2} \{a_1 a_1 a_2 a_1 a_3\}_{u_1} \}$
- Formalize the problem that, from a given sequence of actions, predicts the next action






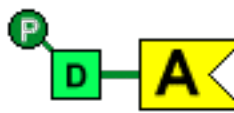
---

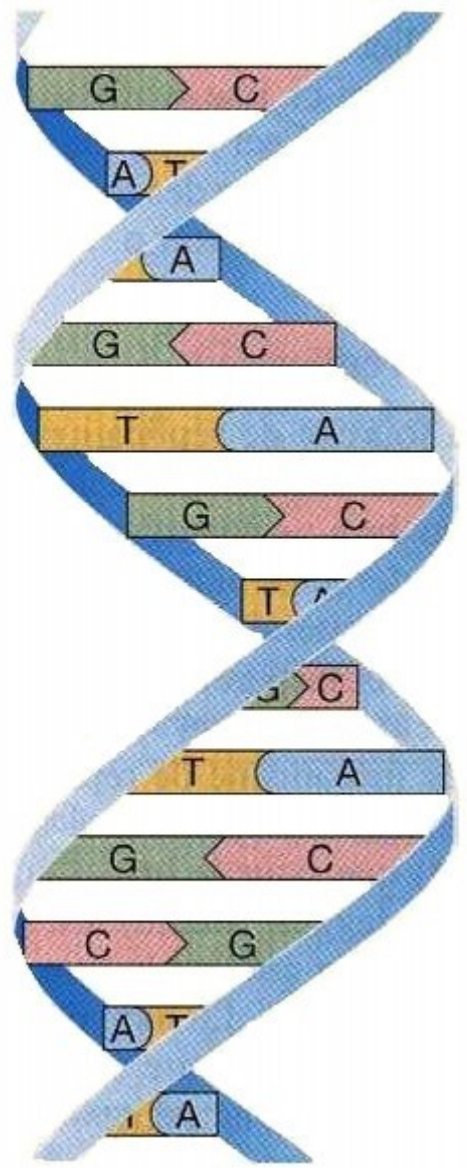
Discrete sequences  
Pattern extraction  
Similarity-based approach

# Sequence alignment



montant      barreau      montant

-  adénine
-  thymine
-  cytosine
-  guanine
-  acide phosphorique  
désoxyribose
-  nucléotide



... ACTGTCATCATCTTACTCATCATCTTACGCT  
 ATAGCTATAGTCATCATCTTACGCTATAGCAC  
 GCTACGAATCTCTGAATAACGCTACGAATCC...

... CTGTCCTGCATCACTGGATGTACCATCTTAC  
 TATGGTACATCTATGTACTACTCAC-TTTTAAAC  
 TCTTACGCTATAGCTATAGTCTACGAATCAC...

# String mining and sequence alignment

---

- Alignment of two sequences of characters
  - Used to compare 2 sequences  $S_1$  ( $|S_1|=m$ ) and  $S_2$  ( $|S_2|=n$ )
  - How  $S_1$  can be transformed into  $S_2$ ?
  - Based on a distance or a similarity measure
  - A sequence alignment can be computed using dynamic programming in  $O(mn)$
- 2 types of alignments:
  - Global
  - Local (Smith & Waterman, 1981)

# Sequence alignment

---

- $\begin{pmatrix} \text{A C G - - A} \\ \text{A T G C T A} \end{pmatrix}$  is an alignment of the two sequences “**ACGA**” and “**ATGCTA**”.
- Algorithmically, it corresponds to an edition script (i.e. a computer program)

Operation	Resulting sequence
Substitution of <b>A</b> by <b>A</b>	<b>A</b>
Substitution of <b>C</b> by <b>T</b>	<b>AT</b>
Substitution of <b>G</b> by <b>G</b>	<b>ATG</b>
Insertion of <b>C</b>	<b>ATGC</b>
Insertion of <b>T</b>	<b>ATGCT</b>
Substitution of <b>A</b> by <b>A</b>	<b>ATGCTA</b>

# Local alignments

---

- 3 editing operations
  - Substitution of a symbol from  $S_1$  at a given position by a symbol from  $S_2$
  - Deletion of a symbol from  $S_1$  at a given position
  - Insertion of a symbol in  $S_2$  at a given position
- Scores
  - $Sub(a,b)$ : score to substitute symbol  $a$  by symbol  $b$
  - $Del(a)$ : score to delete symbol  $a$
  - $Ins(a)$ : score to insert symbol  $a$



# Similarity measure

---

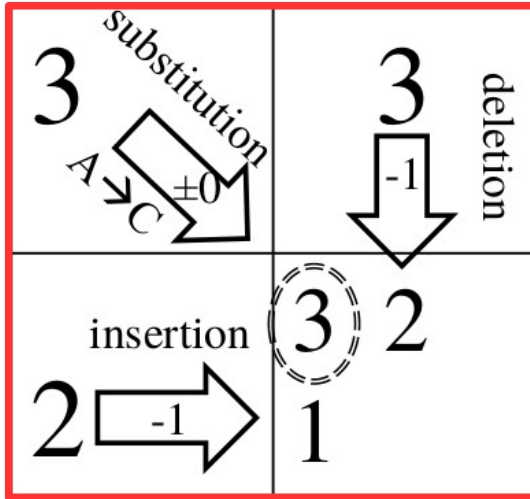
## Similarity measure between 2 sub-sequences

$$s(x,y) = \max \{ \text{score of } e \mid e \text{ in } E_{x,y} \}$$

- $E_{x,y}$ : series of editing operations that transform  $x$  into  $y$
- A score of  $e$  is computed as the sum of all its elementary editing operations

# Dynamic programming

## Smith & Waterman



$$\begin{aligned}
 t[-1, -1] &= 0, \\
 t[i, -1] &= 0, \\
 t[-1, j] &= 0, \\
 t[i, j] &= \max \begin{cases} t[i-1, j-1] + \text{Subs}(x[i], y[j]), \\ t[i-1, j] + \text{Del}(x[i]), \\ t[i, j-1] + \text{Ins}(y[j]), \\ 0 \end{cases}
 \end{aligned}$$

$\text{sub}(x, x) = 1$   
 $\text{sub}(x, z) = 0$   
 $\text{Ins} = \text{del} = -1$

	A	C	G	T	C	G	A	C	G
A	1	0	0	0	0	0	1	0	0
C	0	2	1	0	1	0	0	2	1
T	0	1	2	2	1	1	0	1	2
C	0	1	1	2	3	2	1	1	1
A	1	0	1	1	2	3	3	2	1
C	0	2	1	1	2	2	3	4	3
G	0	1	3	2	1	3	2	3	5

### Remarks:

- Subs/Del/Ins have negative values
- The value at (i,j) position in table t only depends on the 3 adjacent positions
- An optimal alignment (i.e. of maximum score) can be produced by performing a **trace back** in the values of table t from the (maximal) values up to a position of 0.

# Example

	A	C	G	T	C	G	A	C	G
A	1	0	0	0	0	0	1	0	0
C	0	2	0	0	1	0	0	2	0
T	0	0	0	1	0	0	0	0	0
C	0	1	0	0	2	0	0	1	0
A	1	0	0	0	0	0	1	0	0
C	0	2	0	0	1	0	0	2	0
G	0	0	3	1	0	2	0	0	3

	A	C	G	T	C	G	A	C	G
A	1	0	0	0	0	0	1	0	0
C	0	2	1	0	1	0	0	2	1
T	0	1	2	2	1	1	0	1	2
C	0	1	1	2	3	2	1	1	1
A	1	0	1	1	2	3	3	2	1
C	0	2	1	1	2	2	3	4	3
G	0	1	3	2	1	3	2	3	5

Two accumulated similarity tables obtained using the Smith-Waterman algorithm. The left has been calculated using a similarity score of 1 for matches, and dissimilarity penalties of -2 for non-matching substitutions and indels. The right table has this penalty reduced to -1. In each case, the alignments with a similarity score of at least 3 have been highlighted. Note how the higher penalty leads to smaller, more local alignments.

# Exercise

---

- Formalize the problem corresponding to the pattern extraction from discrete sequences using a similarity-based approach

# From patterns to frequent patterns

Sequence alignments

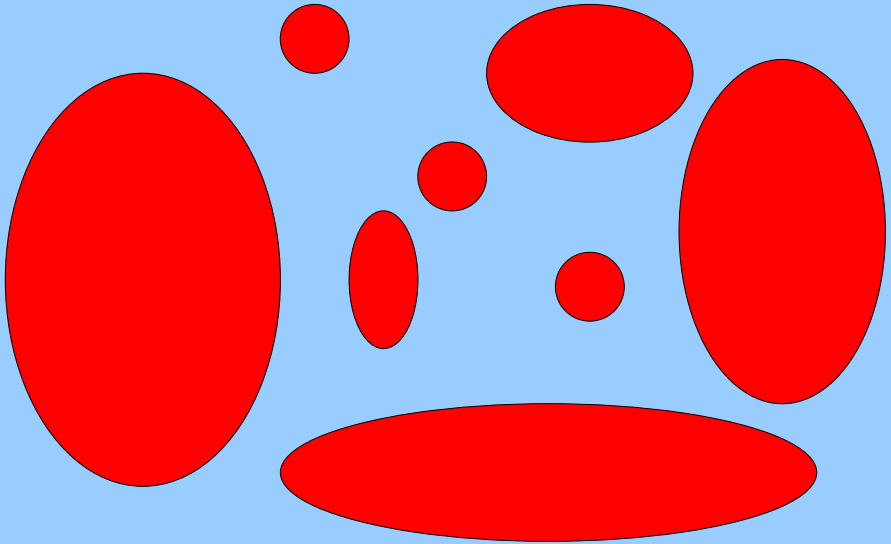


- (pattern 1, sequence 1)
- (pattern 2, sequence 2)
  
- (pattern 3, sequence 1)
- (pattern 4, sequence 5)
  
- ...
  
- (pattern n-1, sequence x)
- (pattern n, sequence y)

Pattern clustering



**Pattern database**



---

Discrete sequences  
Pattern extraction  
Frequency-based approach

# Approach

---

- Frequent sequential patterns
  - Ex: *users that perform action 'A', often perform action 'B' shortly after*
  - Association rules ('A' => 'B' shortly after)
  - Gap: number of elements (actions) between 'A' and 'B'
  - Confidence: how often a rule has been found to be true
- A simple structure: suffix tree

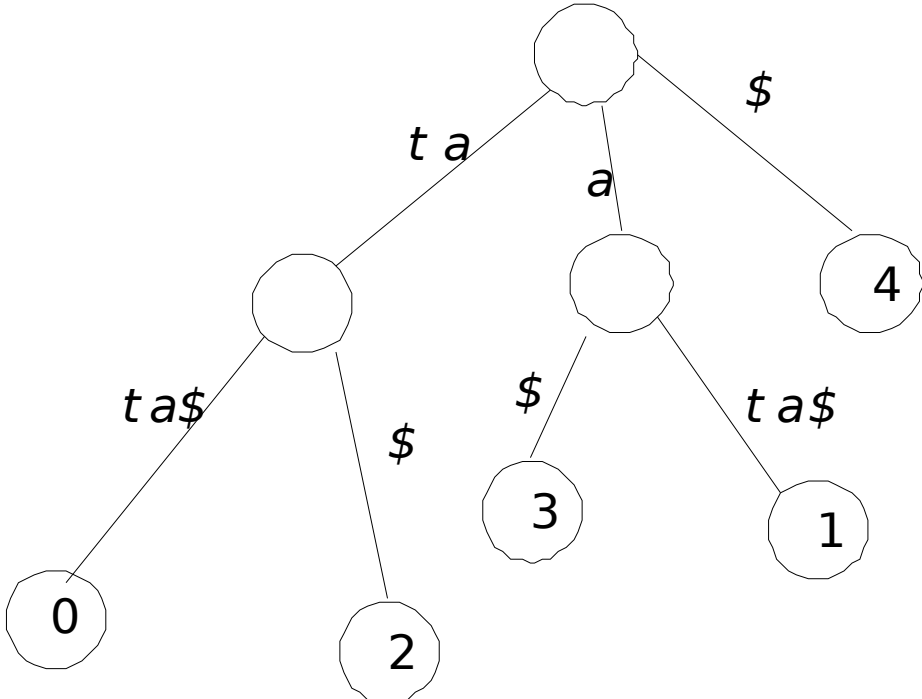
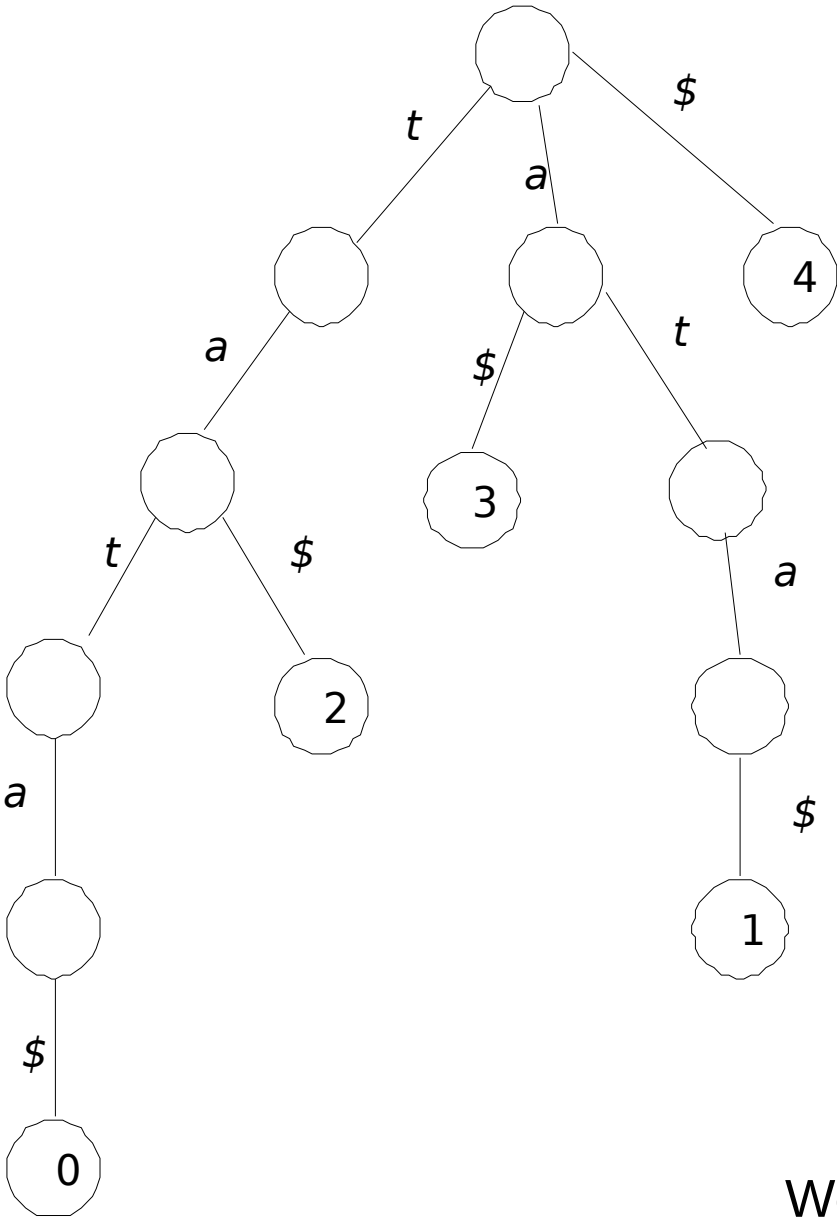
# Suffix trees

---

- The **non-compact suffix tree** of a word  $y$  is the deterministic finite automaton, having a single initial condition called *root* and where the terminal states correspond to the suffix of the word. The language recognized by this automaton is all suffixes of  $y$ .
- In practice a terminator is added at the end of the word (usually denoted \$).
- Leaves are numbered according to the starting position of the suffix they recognize.
- To **compact the tree**, the internal nodes having only a single outgoing branch are removed and the branches are concatenated.



# Example of suffix tree (single word)



Word: "tata"



# Application: CISMeF

- Extraction of recurrent behaviors in the navigations within an online health catalog (CISMeF)

CISMeF [À propos de](#) [Sites et documents médicaux](#) [Terminologies de Santé](#) [Autres outils](#) [Aide](#)

[Se connecter](#)

**DocCISMeF**  
Outil de recherche en santé

Recherche Avancée

asthme



598 entrées trouvées en 0,63 s ★★★ concept(s) identifié(s) : :0804493 - asthme asthme

Vos recherches (1)

Même recherche avec

Voir aussi

Votre sélection

Affiner

Éditeur

- (136) Centre Cochrane Français
- (97) HAS - Haute Autorité de Santé
- (29) Minerva - Revue d'Evidence-Based Medicine
- (21) Revue Médicale Suisse

Type de Ressource

- (142) résumé ou synthèse en français
- (108) article de périodique
- (74) revue de la littérature
- (73) avis de la commission de transparence

Indexation

- (191) asthme
- (238) asthme/traitement médicamenteux
- (195) enfant
- (169) résultat thérapeutique

Niveau d'études

- (15) 2eme cycle / master
- (5) 3eme cycle / doctorat
- (2) 1er cycle / licence

Pays

- (398) France
- (71) Canada
- (54) Belgique
- (50) Suisse

Année

- (23) 2016
- (47) 2015
- (40) 2014
- (20) 2013

1.

## Traitement de l'asthme chronique chez l'enfant : ajout de bêta2-mimétiques à longue durée d'action (LABA) aux corticostéroïdes inhalés (CSI) ?

Minerva - Revue d'Evidence-Based Medicine [Belgique](#) [2016](#)

**\*lecture critique d'article;**

Cette synthèse méthodique avec méta-analyse de la Cochrane Collaboration de bonne qualité méthodologique rassemblant les meilleures preuves actuellement disponibles n'apporte que peu de support factuel à l'ajout d'un LABA aux CSI chez les enfants asthmatiques insuffisamment contrôlés par un CSI seul. Aucune différence quant aux effets indésirables n'a été observée hormis un ralentissement de la croissance dans les groupes utilisant les doses les plus élevées de CSI. Un risque potentiel majoré d'hospitalisation sous LABA mérite une attention particulière à l'avenir.

Voir l'indexation (11)

2.

## INNOVAIR FORMODUAL INNOVAIR NEXTHALER FORMODUAL NEXTHALER 200/6 µg/dose Mise à disposition d'un nouveau dosage (200/6 µg/dose) des en complément du dosage à 100/6 µg/dose. Ces nouvelles spécialités ont une AMM uniquement dans l'asthme

HAS - Haute Autorité de Santé [France](#) [Paris](#) [2016](#)

**\*avis de la commission de transparence;**

"Le service médical rendu par INNOVAIR/FORMODUAL 200/6 µg/dose, solution pour inhalation en flacon pressurisé, et INNOVAIR/FORMODUAL NEXTHALER 200/6 µg/dose, poudre pour inhalation, est important dans les indications de l'AMM. INNOVAIR/FORMODUAL 200/6 µg par dose, solution pour inhalation en flacon pressurisé, INNOVAIR/FORMODUAL NEXTHALER 200/6 µg par dose, poudre pour inhalation, n'apportent pas d'amélioration du service médical rendu (ASMR V) par rapport à INNOVAIR/FORMODUAL 100/6 µg/dose, solution pour inhalation en flacon pressurisé, INNOVAIR/FORMODUAL NEXTHALER 100/6 µg/dose, poudre pour inhalation et aux autres associations fixes corticoïde + bêta-2 agoniste de longue durée d'action dans le traitement continu de l'asthme persistant..."

Voir l'indexation (19)

3.

## Quelle place prend l'immunothérapie par voie sous-cutanée ou sublinguale dans le traitement de l'asthme et de la rhinite allergique? In PHARMACTUEL, Vol. 49, No. 1, 2016

Pharmactuel - la revue internationale francophone de [Canada](#) [2016](#)  
la pratique pharmaceutique en établissement de santé

**\*article de périodique;**

"L'immunothérapie est pertinente lorsque l'arsenal habituel de médicaments ne permet pas d'obtenir une maîtrise acceptable et suffisante des symptômes d'asthme et de rhinite allergique. C'est également une option lorsque l'évitement de l'allergène est impossible ou ne donne pas les résultats attendus. Les allergies affectent actuellement le quart de la population mondiale. Cet article a pour objectif de répondre à la question suivante : « Quelle est la position de l'immunothérapie ciblée par rapport aux standards thérapeutiques pour l'asthme et la rhinite allergique, en monothérapie ou en combinaison? »..."

Voir l'indexation (12)

4.

## Asthme du nourrisson et de l'enfant

SIDES - Référentiel Officiel du Collège des [France](#) [2016](#)  
Enseignants de Radiologie de France et du Collège National des Enseignants de Biophysique et Médecine Nucléaire

**\*cours; épreuves classantes nationales;**

"L'asthme se définit comme la récurrence d'au moins trois épisodes de dyspnée expiratoire sifflante avec sibilants, quel que soit le facteur déclenchant, et l'existence ou non d'un terrain atopique. Il est très fréquent et fait suite le plus souvent à un épisode typique de bronchiolite aiguë virale. L'asthme sévère se traduit par une insuffisance respiratoire aiguë."

Voir l'indexation (5)

5.

## Asthme mal contrôlé sous CSI chez l'adulte : ajout de LAMA ou de LABA ?

Minerva - Revue d'Evidence-Based Medicine [Belgique](#) [2016](#)

**\*lecture critique d'article;**

"Question clinique Chez les patients souffrant d'un asthme mal contrôlé sous CSI seuls, quelles sont l'efficacité et la sécurité de l'ajout d'un LAMA versus l'ajout d'un LABA ?"

Voir l'indexation (8)

1 2 3 4 5 6 7 8 9 10 ... 30 >

# Application: CISMeF

---

- **Data preparation**

- Episode extraction: IP + semantic distance between documents + time between requests
- Resource identification: unique ID + delimiter  
example of session: /59451/ /303901/ /170702/

- **Recurrent pattern extraction**

- Generalized suffix tree
- Longest repeated substrings

# Application: CISMeF

Nombre de liens visités	Nombre d'épisodes	Proportion
1	34 005	70,6 %
2	8 254	17,1 %
3	2 940	6,1 %
4	1 284	2,7 %
5	658	1,4 %
6	346	0,7 %
7	216	0,4 %
8	139	0,3 %
9	91	0,2 %
10	60	0,1 %
>10	175	0,4 %

22 days of log analysis:

- 10mn max for an episode
- 48 168 episodes
- 17mn of data processing (2,39GHz/512Mo)

Longueurs des motifs	2	3	4	5
Nombres de motifs	1557	146	20	4

	Episodes contenant un motif			
	longueur 2	longueur 3	longueur 4	longueur 5
Effectifs	4127	326	42	8
% épisodes	8,568	0,677	0,087	0,017
% épisodes (l>1)	29,139	2,302	0,297	0,056

# Discrete sequences

---

- Pattern extraction:
  - Sequence alignments
    - similarity OK, but frequency is difficult to evaluate (should be paired with pattern clustering)
- Prediction:
  - Generalized suffix trees
  - Seq2seq, CRF, HMM, ...
    - frequency OK, but only slight variations (similarity) can be taken into account

# Exercise

---

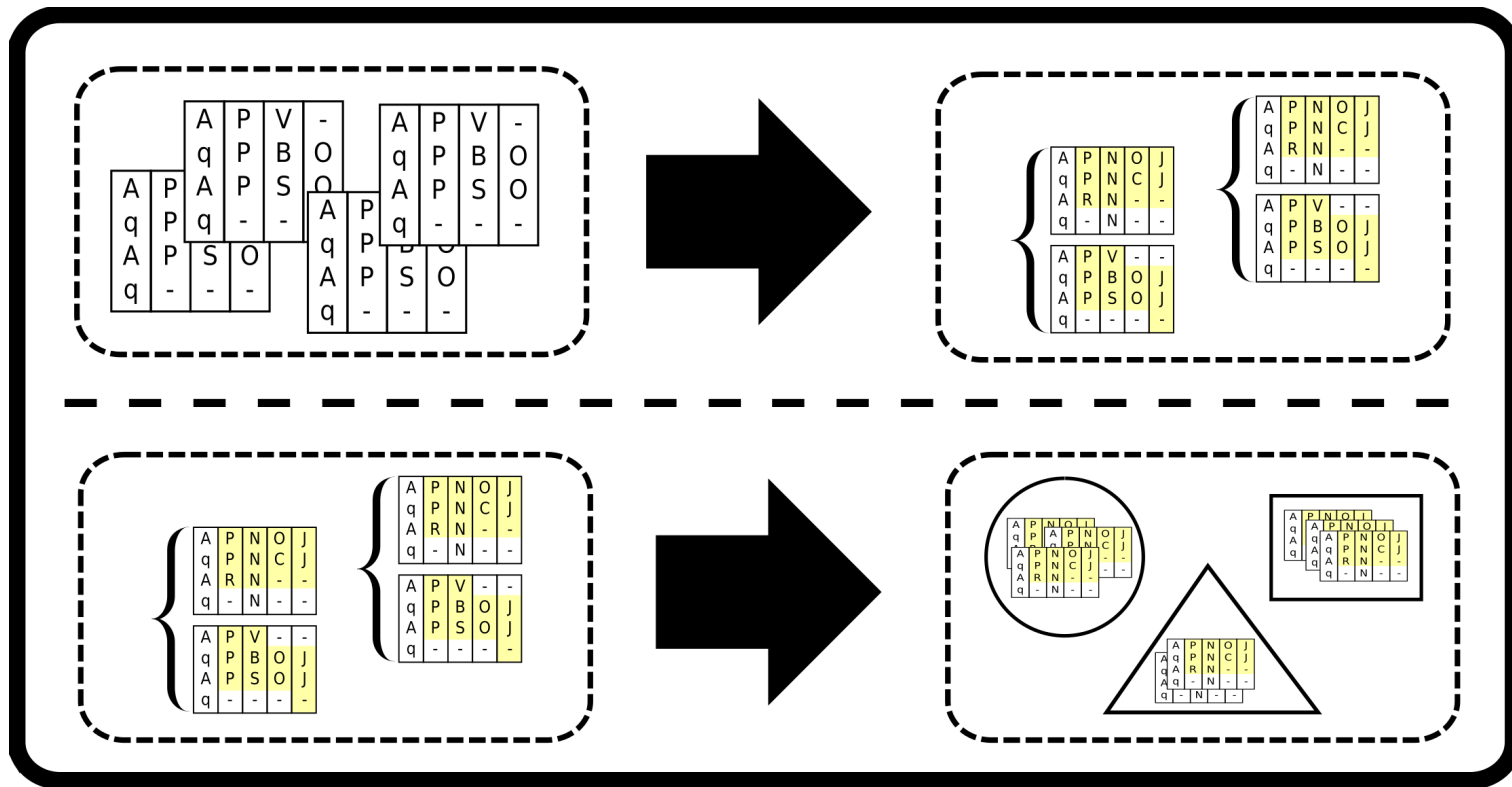
- Formalize the problem that consists in predicting the most probable action of user given the set of previous actions performed?
- What would be an algorithm to construct a (compact) suffix tree?

---

Discrete sequences of item-sets  
Pattern extraction  
Similarity-based approach



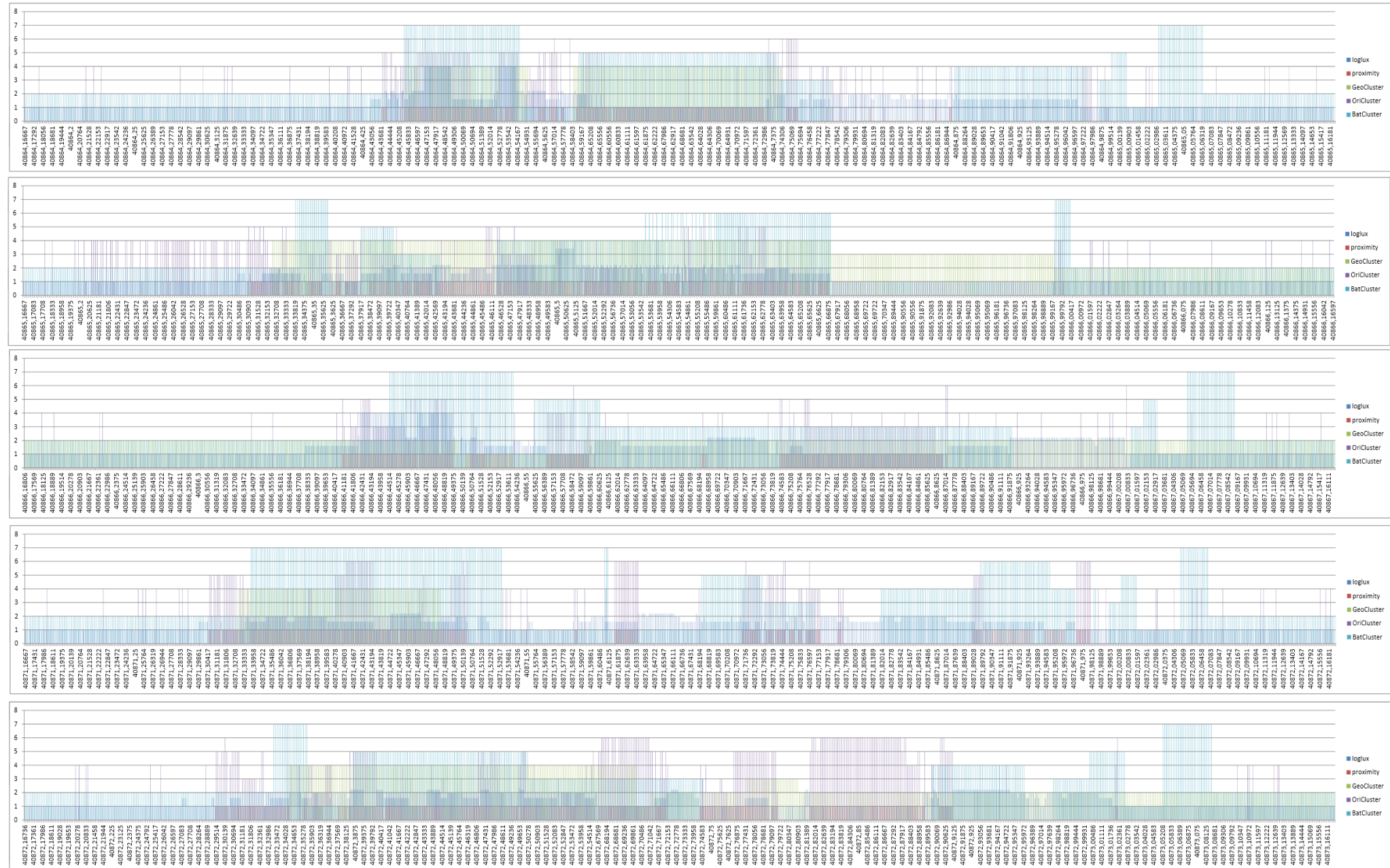
# Approach (similar to sequences)



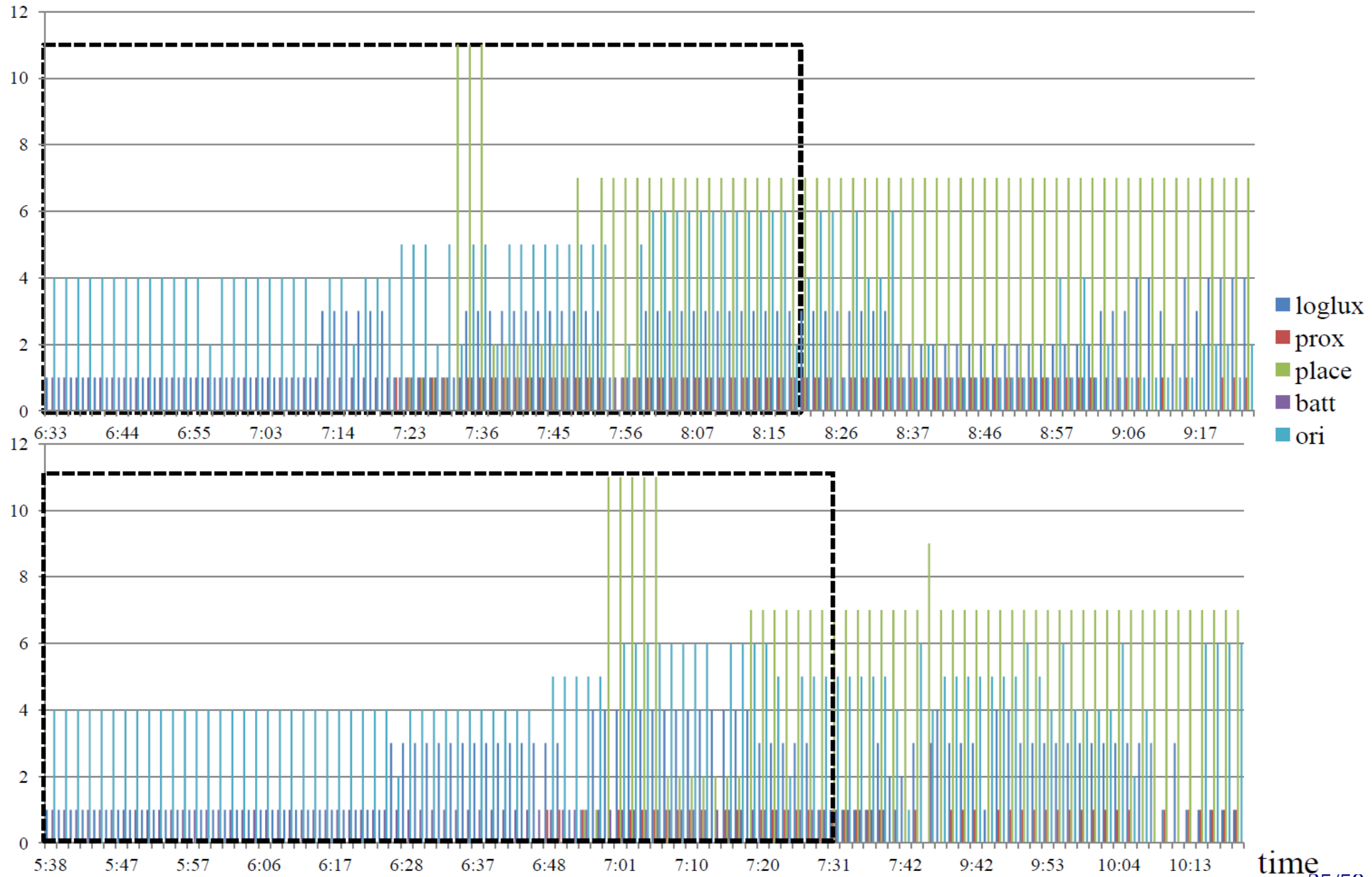
## Decomposition in two separated steps

- Extraction of (pair of) interaction patterns
- Clustering of interaction patterns

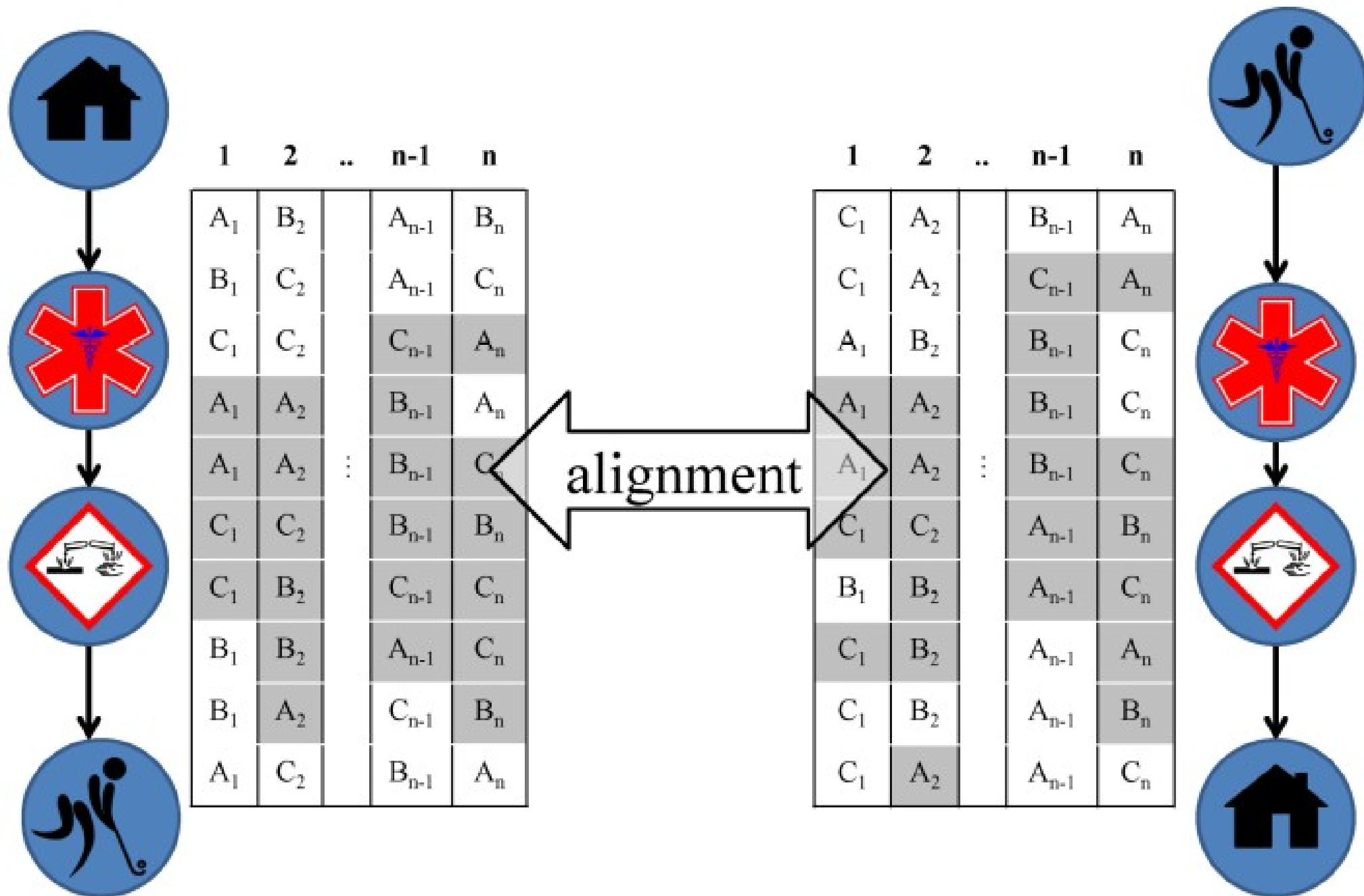
# Routine extraction by matrix alignment



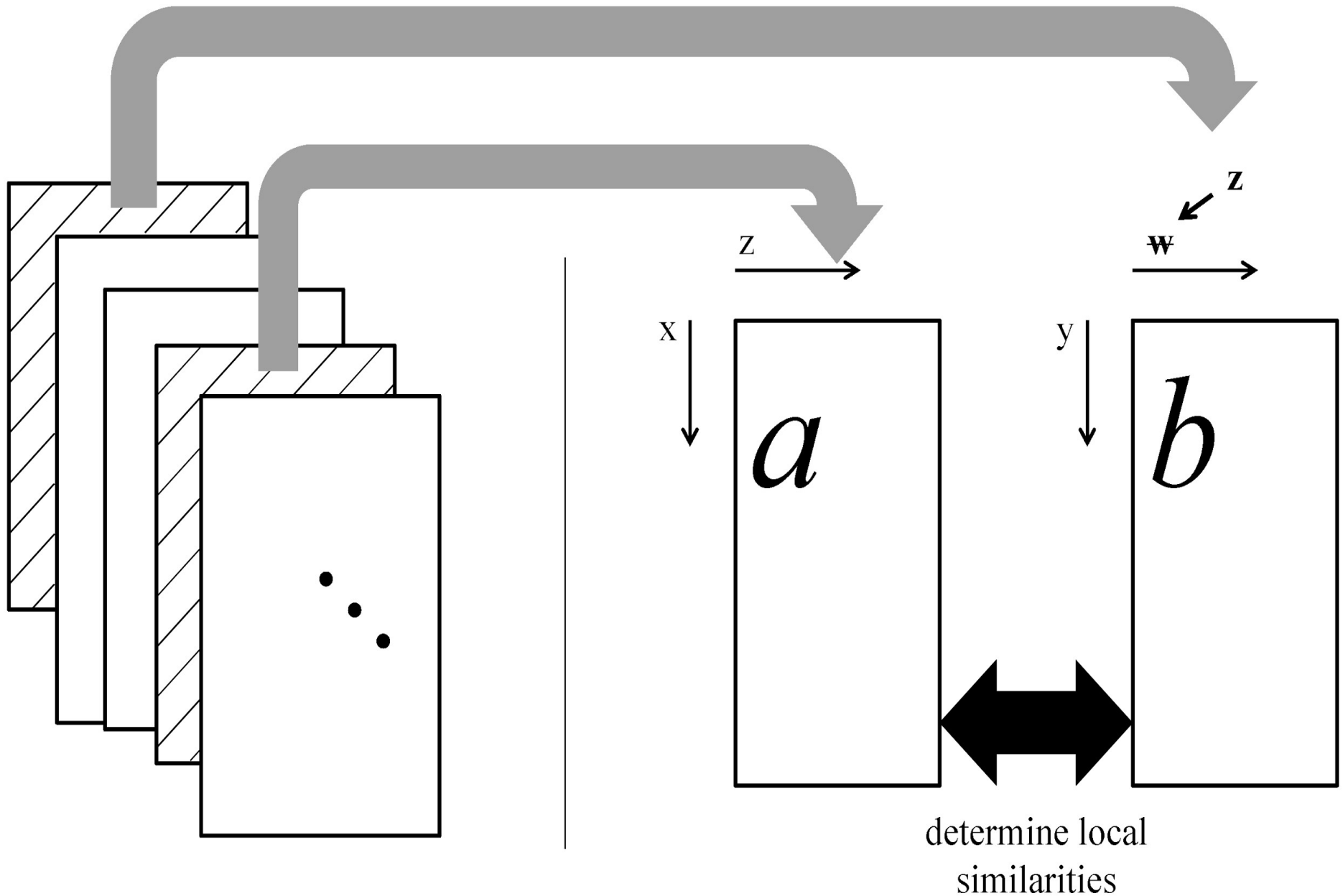
# Routine extraction by matrix alignment



# Routine extraction by matrix alignment

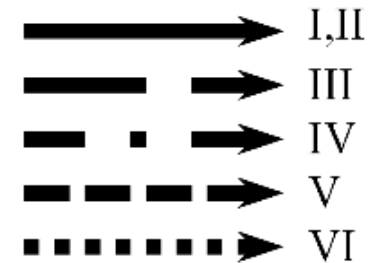
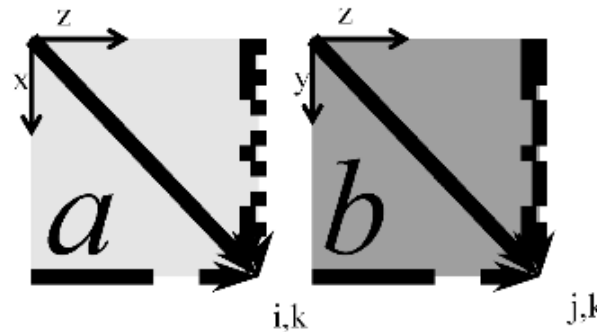
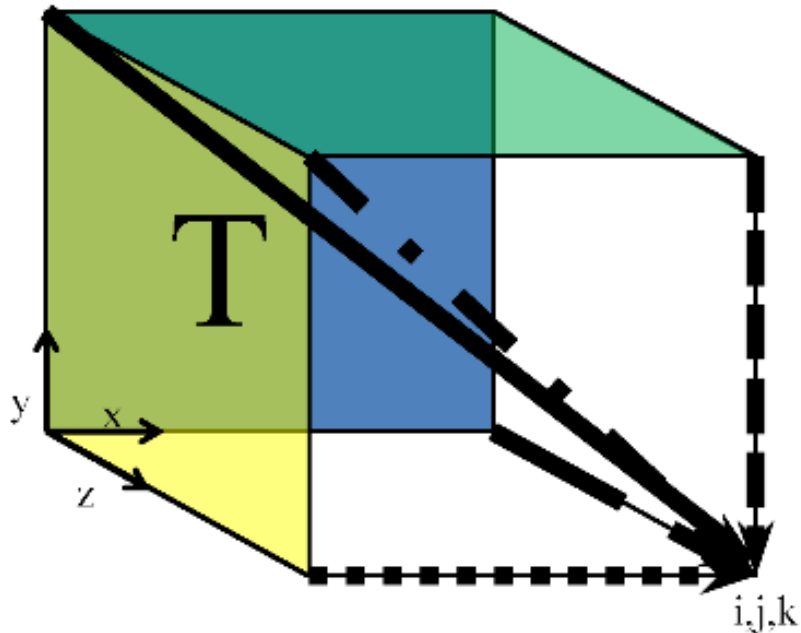


# Routine extraction by matrix alignment



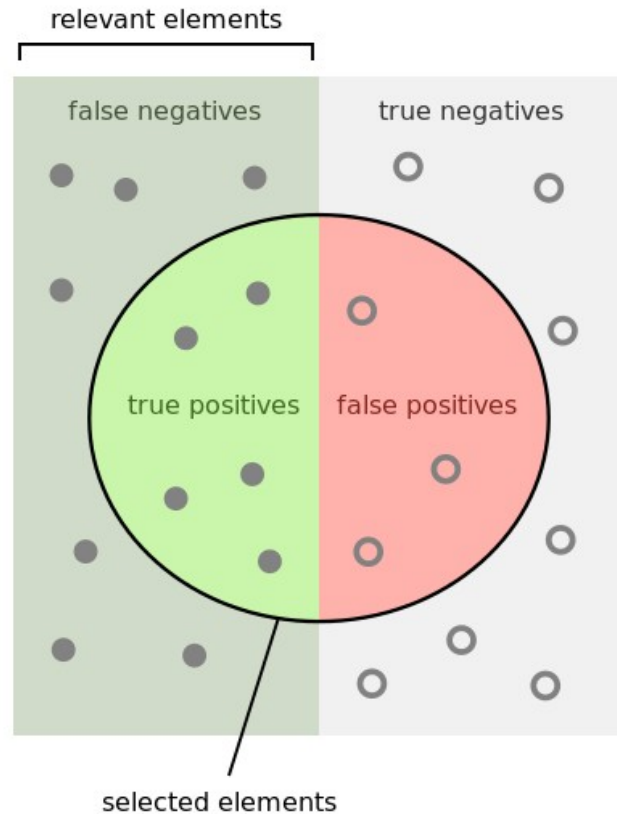
# Routine extraction by matrix alignment

$$\text{sim}(a, b) = \max \left( \begin{array}{l} 0 \\ \text{sim}(a_{1,1}, b_{1,1}) + \text{sim}(a [0, 0] \uparrow, b [0, 0] \uparrow) + \text{sim}(a [0, 1] \leftarrow, b [0, 1] \leftarrow) \quad \text{(I)} \\ \text{sim}(a_{1,1}, b_{1,1}) + \text{sim}(a [1, 0] \uparrow, b [1, 0] \uparrow) + \text{sim}(a [0, 0] \leftarrow, b [0, 0] \leftarrow) \quad \text{(II)} \\ \text{sim}(a_{0,1}, b_{0,1}) + \text{sim}(a [0, 0] \uparrow, b [0, 0] \uparrow) \quad \text{(III)} \\ \text{sim}(a_{1,0}, b_{1,0}) + \text{sim}(a [0, 0] \leftarrow, b [0, 0] \leftarrow) \quad \text{(IV)} \\ \text{sim}(a_{0,0}, b_{1,0}) + \text{indel}(b[0, 0]) + \text{sim}(a[0, 0] \leftarrow, b[1, 0] \leftarrow) \quad \text{(V)} \\ \text{sim}(a_{1,0}, b_{0,0}) + \text{indel}(b[0, 0]) + \text{sim}(a[1, 0] \leftarrow, b[0, 0] \leftarrow) \quad \text{(VI)} \end{array} \right)$$



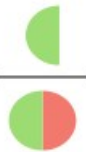
$i$  lies on the x-axis  
 $j$  lies on the y-axis  
 $k$  lies on the z-axis

# Evaluation: usual measures



How many selected items are relevant?

Precision =



How many relevant items are selected?

Recall =



- **Model evaluation:**

- Precision

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

- Recall

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

- F-measure  $F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

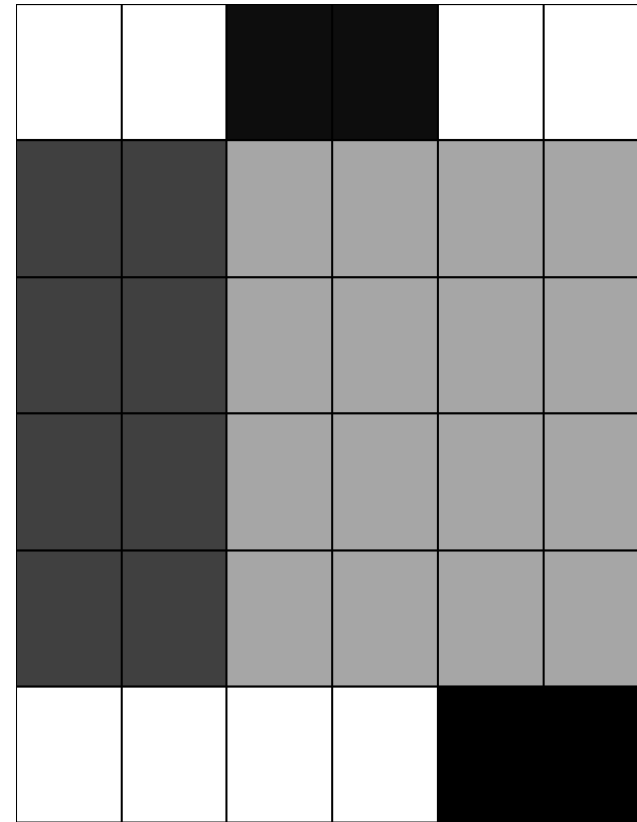
- **Remarks:**

- Unbalanced classes!
- Evaluation by class

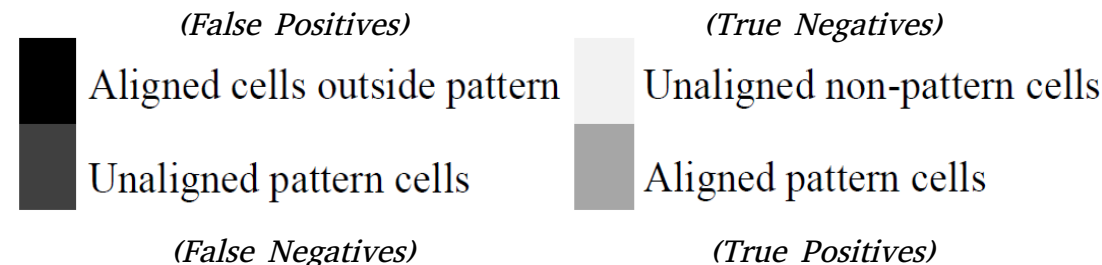
# Routine extraction by matrix alignment

## Evaluation

- Data and ground truth
- Computation time
- Precision and recall for each pattern alignment
- Alignment size
- Number of alignments



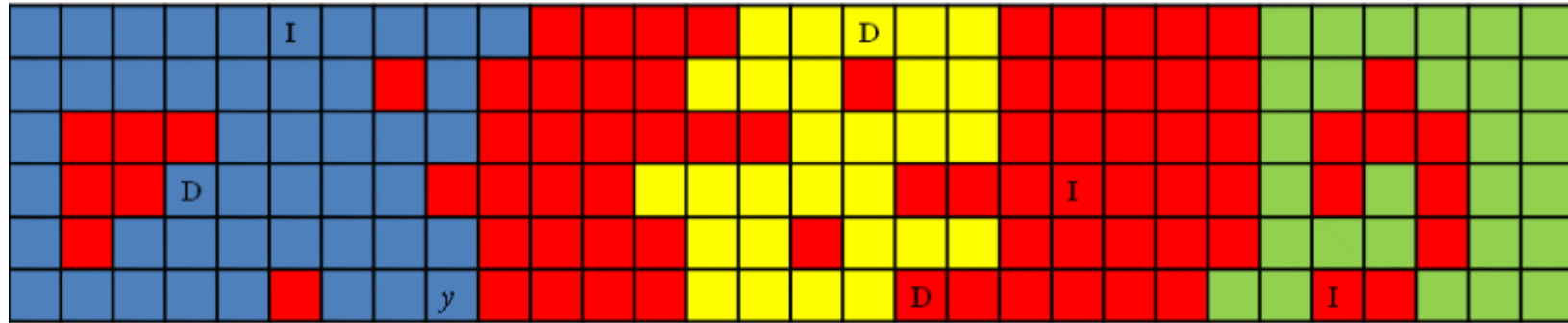
Total pattern cells: 24  
Total aligned cells: 20  
Aligned pattern cells = 16  
Precision =  $16/20 = 4/5$   
Recall =  $16/24 = 2/3$   
Size ratio =  $20/24 = 5/6$





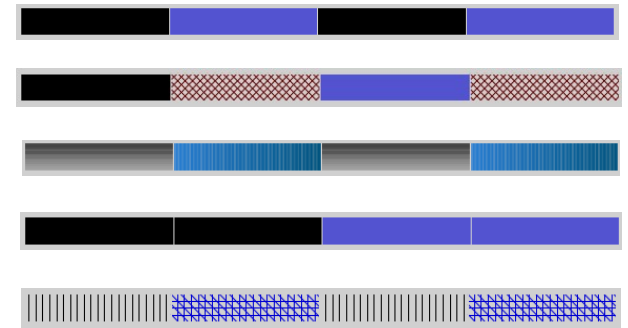
# Routine extraction by matrix alignment

- Synthetic data generation



- Scenarios:

- 1) Regular activity, no noise
- 2) Regular activity, noise between patterns
- 3) Noisy patterns
- 4) Irregular pattern apparition
- 5) Faulty sensors ( $\frac{3}{4}$  random data)



Measure	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
$\frac{\#alignments}{\#pairs\ of\ pat.}$	$0.31 \pm 0.26$	$0.69 \pm 0.29$	$0.041 \pm 0.054$	$0.31 \pm 0.26$	$0.22 \pm 0.32$
precision	$0.54 \pm 0.22$	$0.77 \pm 0.20$	$0.13 \pm 0.06$	$0.52 \pm 0.24$	$0.54 \pm 0.18$
recall	$1.00 \pm 0.00$	$1.00 \pm 0.00$	$0.56 \pm 0.19$	$1.00 \pm 0.01$	$0.20 \pm 0.06$
$\frac{alignment\ size}{pattern\ size}$	$5.66 \pm 4.48$	$2.45 \pm 2.51$	$6.72 \pm 3.96$	$6.65 \pm 6.08$	$1.60 \pm 0.24$

# Exercise

---

- Formalize the problem of behavior pattern extraction from discrete sequences of set of actions, based on a similarity approach (looking for longest and more similar behaviors)

---

Discrete sequences of item-sets  
Pattern extraction  
Frequency-based approach

# Data-mining sequences of item-sets

---

- Frequent item-sets
  - Ex: *users usually select A & B keys*
  - Association rules ('A' => 'B' in the same item-set)
  - Support: how frequently an item-set appears in the dataset
  - Confidence: how often a rule has been found to be true
- Frequent sequential patterns
  - Ex: *users that perform action 'A', often perform action 'B' shortly after*
  - Association rules ('A' => 'B' in two following item-sets)
  - Gap: item-sets appearing inside the association rules

# Definitions

---

- **Item:** minimal element
- **Item-set:** ordered (preference, ...) set of items
- **Sequence of item-sets:** ordered set of item-sets
- **Transaction:** tuple in the database; a set of items or a sequence of item-sets.
  - training set = set of transactions
  - $\{\{a_{1,1}, \dots, a_{1,m_1}\}, \{a_{2,1}, \dots, a_{2,m_2}\}, \dots, \{a_{n,1}, \dots, a_{n,m_n}\}\}$
  - Or a sparse matrix...
- **Association rule:** application  $X \rightarrow Y$  where  $X$  and  $Y$  are disjoint set of items or set of item-sets

# Evaluation of association rules ( $X \rightarrow Y$ )

---

- **Support:** absolute probability  $P(X \cup Y)$

$||X \cup Y|| / ||BD|| = \% \text{ of transactions satisfying the rule}$

- **Confidence:** conditional probability  $P(Y/X)$

$||X \cup Y|| / ||X|| = \% \text{ of transactions}$

verifying the implication

$= \text{support}(XY) / \text{support}(X)$

- An interesting association rule is a rule whose Confidence  $>$  Minconf and Support  $>$  Minsup

# Apriori algorithm (Agrawal & Srikant, 1994)

---

- **Idea:** if an item-set is not frequent, then all its supersets are not frequent:
  - If  $\{A\}$  is not frequent then  $\{AB\}$  cannot be frequent
  - if  $\{AB\}$  is frequent then  $\{A\}$  and  $\{B\}$  are frequent
- **Process:**
  - 1) Generate iteratively candidate item-sets:
    - First pass: search for frequent 1-sets
    - Generate a candidate of size  $k$  from two candidates of size  $k-1$  differentiated by the last element
    - Filter the sets of items with minimum support (keeping frequent item-sets)
  - 2) Use frequent item-sets to generate association rules

# Apriori

---

## Apriori

*Input:  $L_1 = \{\text{frequent 1-itemsets}\};$*

*Output:  $L_k = \{\text{frequent } k\text{-itemsets}\};$*

```
for (k=2;  $L_{k-1} \neq \emptyset$ ; k++) do {  
     $C_k = \text{apriori-gen}(L_{k-1});$   
    // Generate new candidates  
}  
 $L_k = \{ c \in C_k \mid \text{numberOf}(c, \text{DB}) \geq \text{minsup} \};$   
    // Filter candidates  
}  
return  $L_k$ ;
```

## Apriori-gen

*Input:  $L_{k-1} = \{\text{frequent } (k-1)\text{-itemsets}\};$*

Items of  $L_{k-1}$  are ordered lexicographically

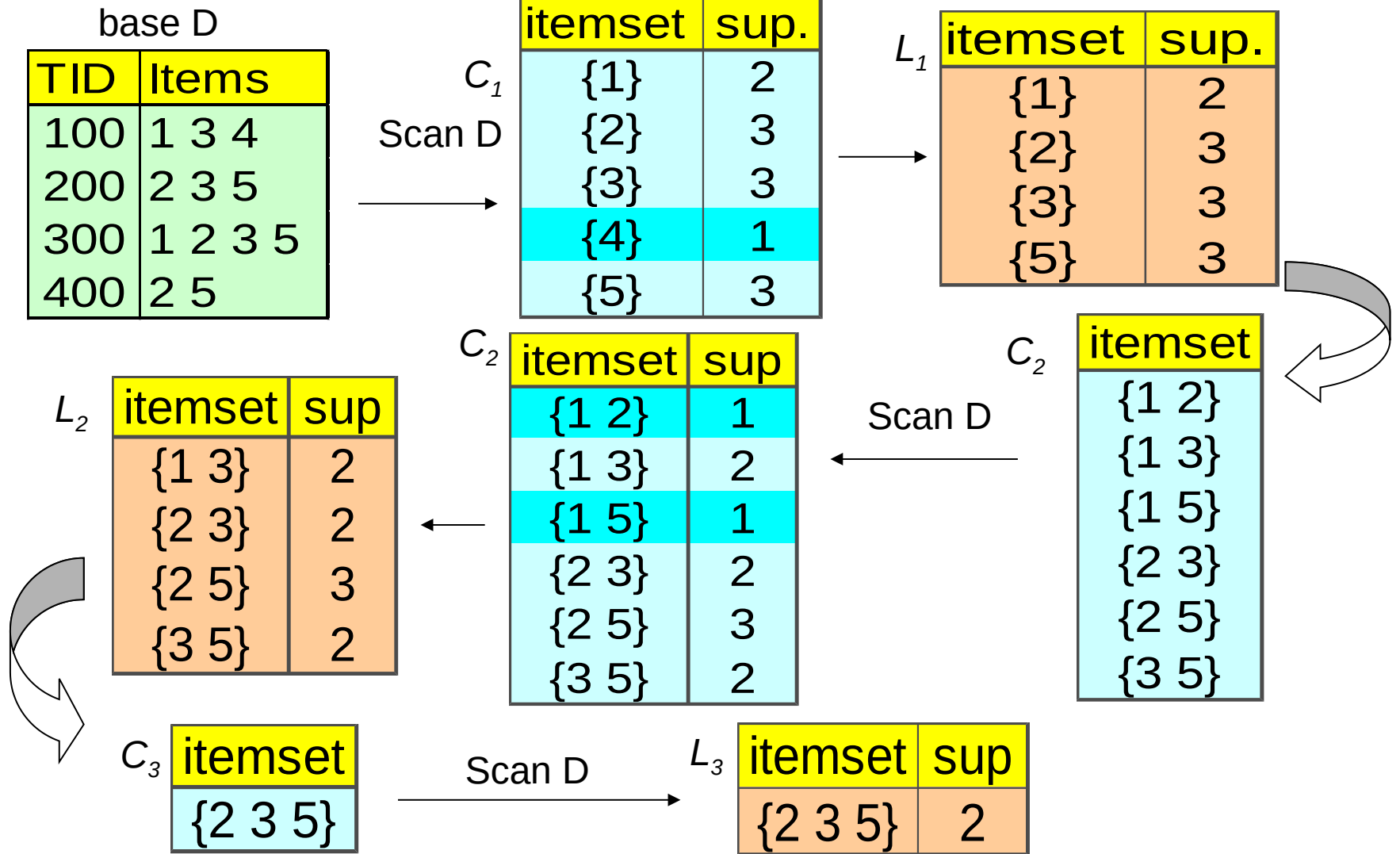
*Output:  $C_k = \{\text{candidates frequent } (k)\text{-itemsets}\};$*

- Step 1: Self-join on  $L_{k-1}$   
For each  $(p_{k-1}, q_{k-1})$  so that  $p_{k-1} < q_{k-1}$  do {  
     $C_k = \{ c_k \mid \text{lexicographically ordered combination of } p_{k-1} \text{ and } q_{k-1} \}$   
}
- Step 2: Pruning  
foreach  $c_k$  in  $C_k$  do {  
    foreach  $(k-1)$ -subsets  $t_{k-1}$  of  $C_k$  do {  
        if ( $t_{k-1}$  is not in  $L_{k-1}$ ) then  
            delete  $t_{k-1}$  from  $C_k$   
    }  
}



# Apriori (example)

min\_support=2



# Apriori (generating association rules)

---

*//Input: MinConf,  $L_k$  (frequent item-sets)*

*//Output: R, set of association rules*

$R = \emptyset$  ;

foreach subsets  $S \neq \emptyset$ ,  $S \neq L_k$  of  $L_k$  do {

    Confidence =  $\text{Sup}(S(L_k - S)) / \text{Sup}(S)$

    If Confidence  $\geq$  MinConf then {

$R = R \cup \{ " S \rightarrow L_k - S " \}$  ;

    }

}

}

return R ;

## Example :

$\{2\ 3\} \rightarrow \{5\}$       confidence=2/2

$\{2\ 5\} \rightarrow \{3\}$       confidence=2/3

...

$\{2\} \rightarrow \{3\ 5\}$       confidence=2/3

...

# Exercise

---

- Formalize the problem of behavior pattern extraction from discrete sequences of set of actions, based on a frequency approach (looking for most frequent behaviors)

---

# Continuous signal(s)

# **(Single) continuous signal**

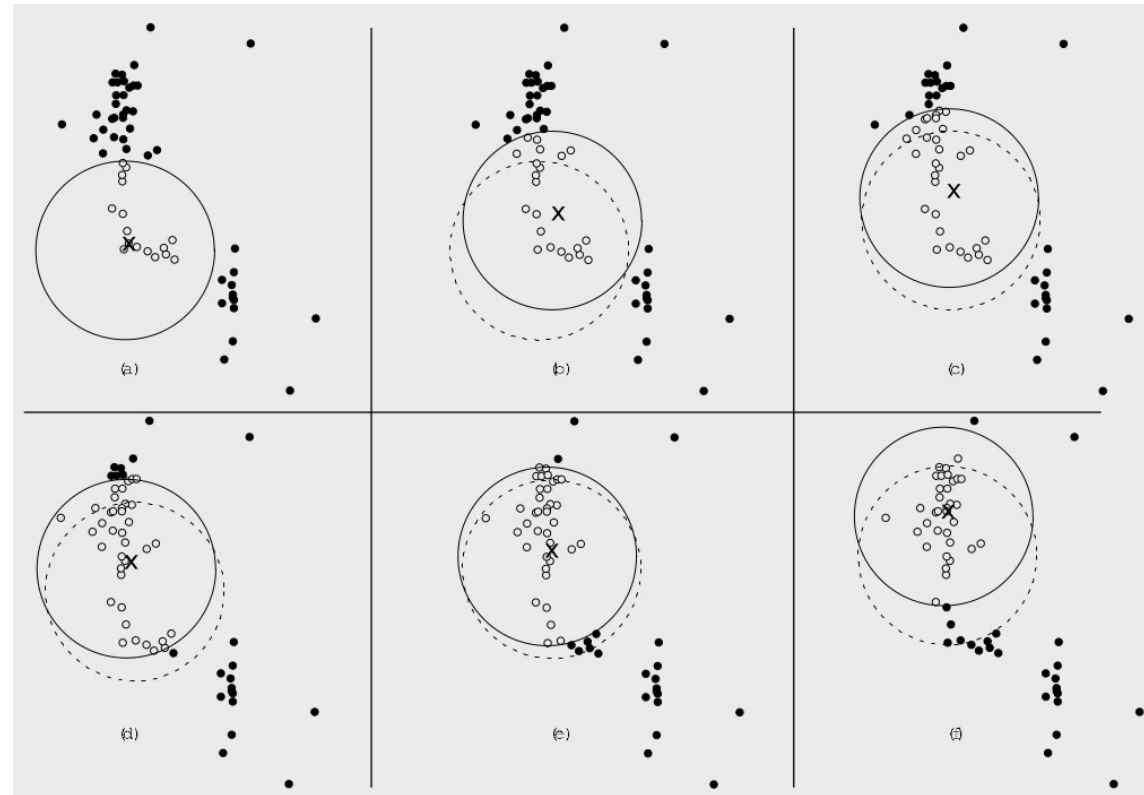
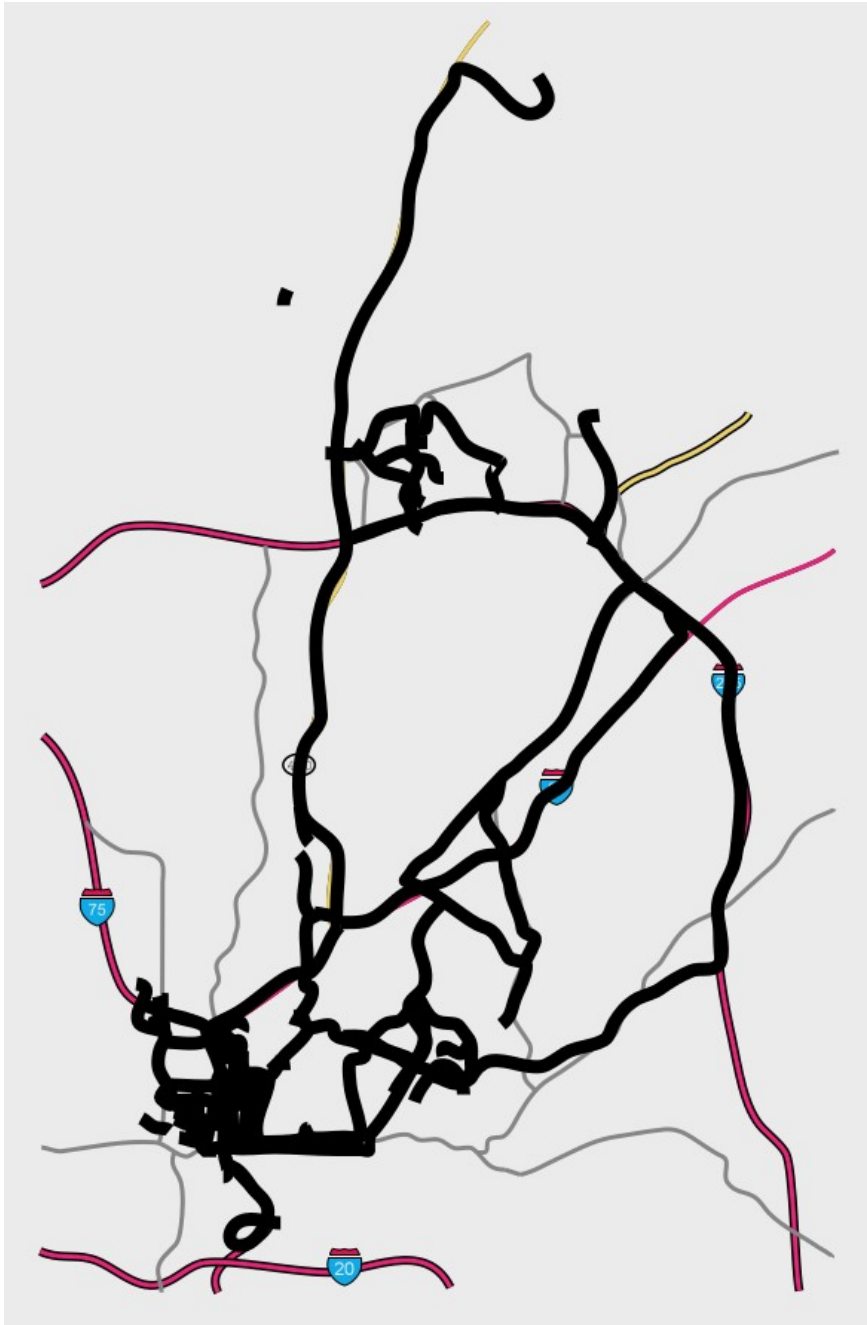
---

- **Continuous numeric acquisition is impossible**
  - Discretize an analogical signal to numerical values along 2 dimensions:
    - Continuous / discrete (alphabet)
    - Time (sampling)
      - NB: a re-sampling can be necessary according to the goal
- **Discretization process**
  - Heterogeneous sensors (unnormalized)
  - Semantic information (analyze & user's feedback)

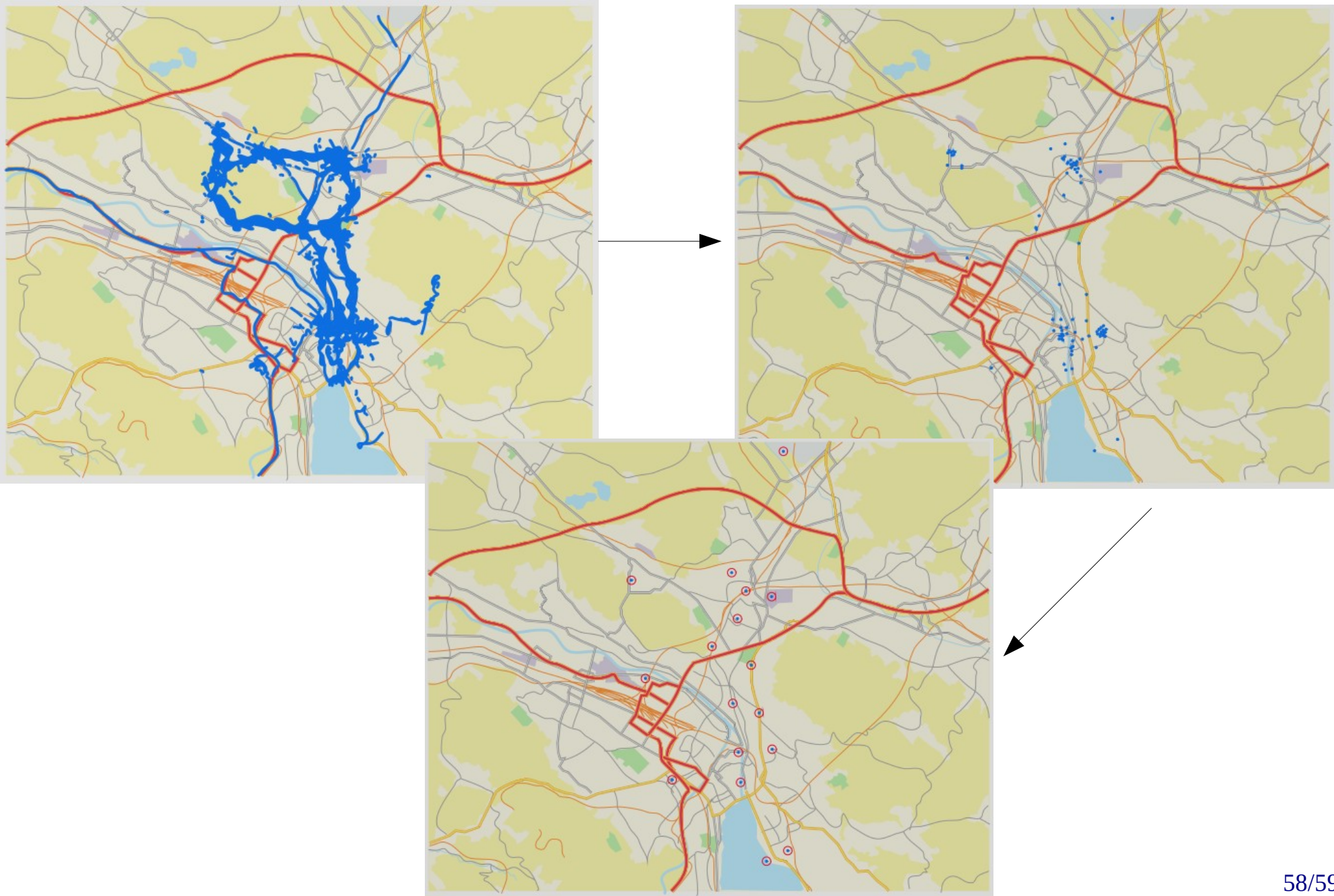
# Example: GPS signal

Daniel Ashbrook & Thad Starner : « Using GPS to learn significant locations and predict movement across multiple users », Personal and Ubiquitous Computing, volume 7, number 5, pp. 275-286, Springer, 2003.

- Finding **significant places / positions** (time dependence: time threshold)
- Clustering places into **locations / keypoints** (spacial dependence: cluster radius)

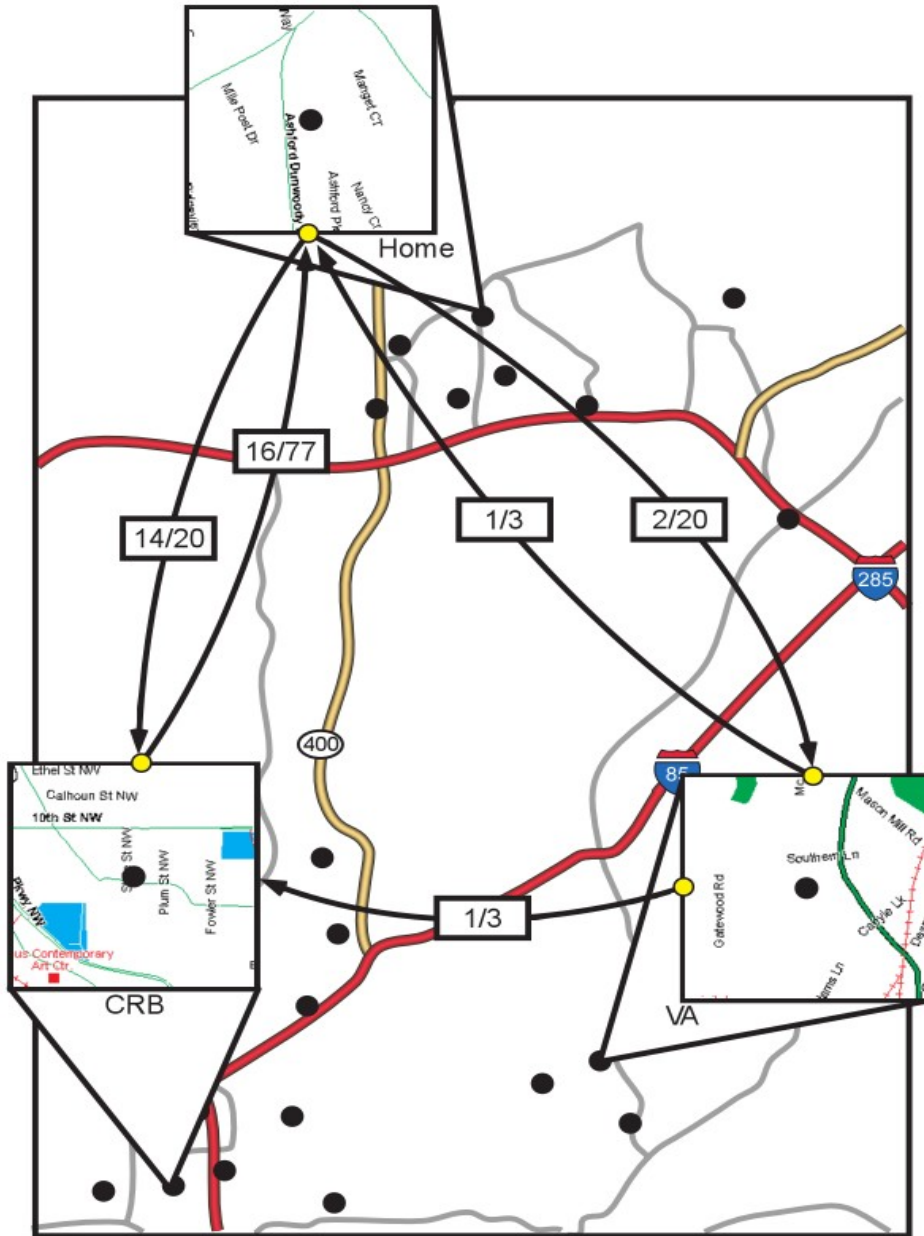


# Example: GPS signal





# Example: GPS signal



Transition	Relative Frequency	Probability
$A \rightarrow B$	14/20	0.7
$A \rightarrow B \rightarrow A$	3/14	0.2142
$A \rightarrow B \rightarrow C$	2/14	0.1428
$A \rightarrow B \rightarrow D$	3/14	0.2142
$A \rightarrow B \rightarrow E$	1/14	0.0714
$A \rightarrow B \rightarrow F$	1/14	0.0714
$A \rightarrow B \rightarrow G$	1/14	0.0714
$A \rightarrow B \rightarrow H$	1/14	0.0714
$A \rightarrow B \rightarrow I$	1/14	0.0714
$B \rightarrow A$	16/77	0.2077
$B \rightarrow A \rightarrow B$	13/16	0.8125
$B \rightarrow A \rightarrow J$	3/16	0.1875
$B \rightarrow C$	10/77	0.1298
$B \rightarrow C \rightarrow A$	6/10	0.6
$B \rightarrow C \rightarrow K$	4/10	0.4
$D \rightarrow B$	5/7	0.7142
$D \rightarrow B \rightarrow A$	2/5	0.4
$D \rightarrow B \rightarrow L$	2/5	0.4
$D \rightarrow B \rightarrow M$	1/5	0.2

Probabilities for transitions in Markov models  
Key: A = "Home"



# Exercise

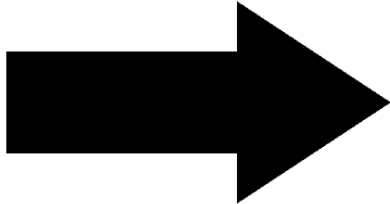
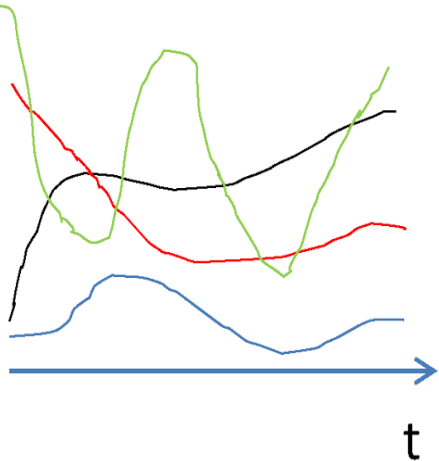
---

- Formalize the problem of finding keypoints from GPS data?
- Formalize the problem of behavior pattern extraction from a sequence of visited places? From GPS data?

# (Multiple) Continuous signals

$$C = C(t)$$

$$\mathbb{R} \rightarrow \mathbb{R}^n$$



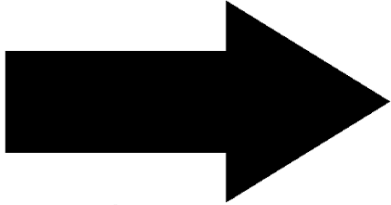
discretisation

$$C = (c_1(t_1), c_2(t_2), \dots, c_n(t_n))$$

$$\mathbb{N}^n \rightarrow \mathbb{N}^n$$

1.1	0.3	0.2	0.8
0.7		0.2	
0.4	0.5	0.3	
0.4	0.6	0.4	
0.6		0.5	0.5
0.8	0.3	0.2	
0.5	0.6	0.2	

$t$



synchronisation  
and classification

$$C = (c_1, c_2, \dots, c_n)(t)$$

$$\mathbb{N} \rightarrow \mathbb{N}^n$$

A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>
B <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>
C <sub>1</sub>	B <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>
C <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>	A <sub>4</sub>
B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>	B <sub>4</sub>
B <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	B <sub>4</sub>
C <sub>1</sub>	B <sub>2</sub>	A <sub>3</sub>	B <sub>4</sub>

$t$

# Multiple sensors: (discrete) approaches

