Extraction de motifs & fouille de données au sein des trajectoires sémantiques

Sur l'enrichissement sémantique des trajectoires spatio-temporelles

*
* *

par Clément Moreau,
Thomas Devogele,
Laurent Etienne

{clement.moreau,
thomas.devogele, @univ-tours.fr
laurent.etienne}

Laboratoire d'informatique
fondamentale et appliquée de Tours

Université de Tours

Introduction

Malgré la complexité et la diversité des trajectoires humaines, (González *et al*., 2008 ; Song *et al*., 2010) montre que les déplacements humains suivent des modèles reproductibles simples.

Oui, mais n'explique pas pourquoi le déplacé est effectué.

Comprendre Pourquoi ?

Pour comprendre l'essence d'un déplacement, il est nécessaire de mener une *inspection contextuelle* de la trajectoire (Parent *et al*., 2013 ; Renso et Trasarti 2013).

Smart Loire

1. Analyse des déplacements touristiques (SMART LOIRE)

   - Projet de recherche d'intérêt régional.
   - Intègre différents pans de l'informatique

     $\Rightarrow$ Services web (composition, personnalisation, …)

     $\Rightarrow$ Optimisation d'itinéraires [touristiques]

     - Contraintes utilisateurs (temporelle, géographiques, préférences)

     $\Rightarrow$ **Fouille de trajectoires**

     - Regrouper des utilisateurs similaires.
     - Perspective de recommandation.

2. Étude des comportements et faits sociologiques (MOBI'KIDS)

$\Rightarrow$ Projet ANR (Grenoble, Rennes, Tours) multi-disciplinaire.

- Géographes
- Urbanistes
- Sociologues
- Informaticiens

$\Rightarrow$ Corroborer des hypothèses sur les comportements des modes de vie urbains.

- $\sim$ 100 familles suivies sur 2×1 semaine.

$\Rightarrow$ Récupération de trajectoires enrichies sémantiquement via des enquêtes sociologiques individuelles.

Introduction

Néanmoins, on veille à la perspective de ***généricité***.

On souhaite s'inspirer de la représentation de la **Time Geography** (Hägerstrand, 1970).

Modélisation centrée sur la notion d'**activité**.

Triptyque dimensionnel :

1. Le Temps
2. L'Espace
3. La Sémantique

Introduction

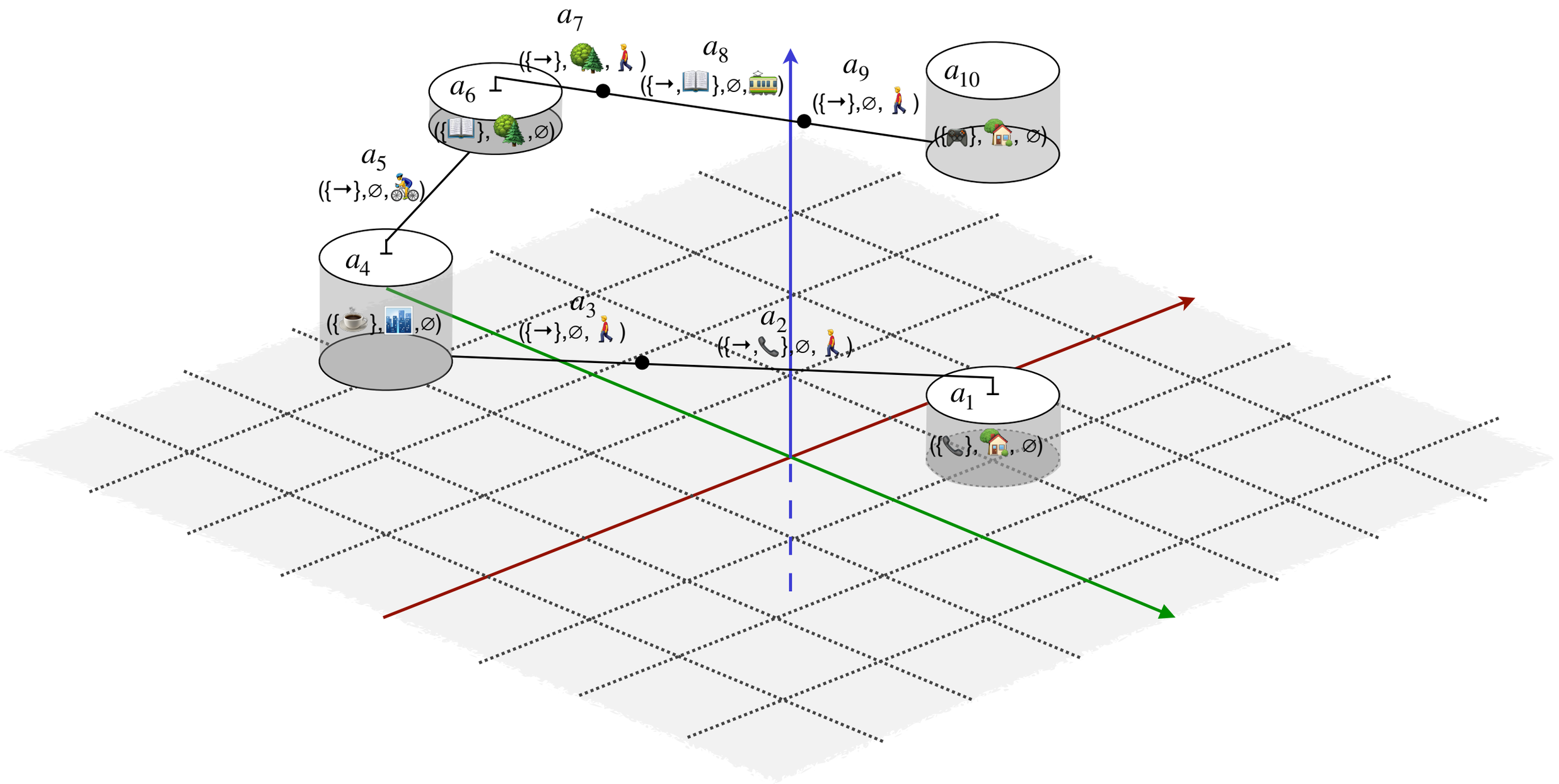


Figure *1* : Représentation de la trajectoire sémantique inspiré de la time geography de Hägerstrand

Représentation sémantique des trajectoires

La *trajectoire sémantique* se basant sur la notion de **séquence d'activités.**

Considération d'un ensemble $\mathcal{O}$ d'ontologies[1] tel que :

$$\{O_a, O_d, O_l\} \subseteq \mathcal{O}$$

- O_a ensemble des concepts d'activités.
- O_d ensemble des concepts de déplacements.
- O_l ensemble des concepts de lieux.

Les dimensions spatiale et temporelle sont portées par la *Trace*.

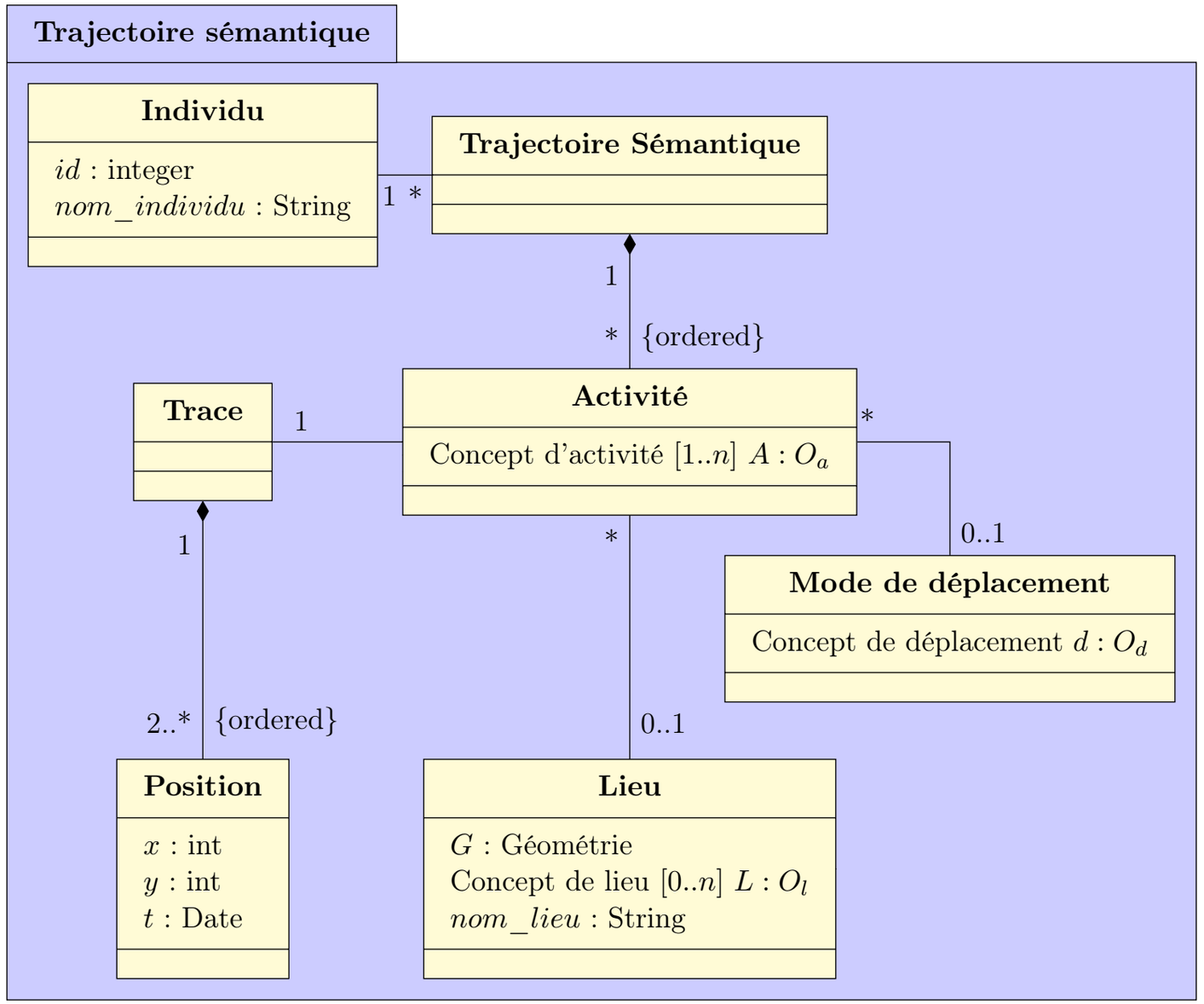


Figure 2 : Diagramme UML de la trajectoire sémantique

[1] **http://www.datatourisme.fr**

À propos de la sémantique

Dans un premier temps, on s'attarde sur la **dimension sémantique**.

On définit $\Sigma = \{(C_a, c_l, c_d | C_a \in \mathcal{P}(O_a), c_l \in O_l \cup \{\emptyset\}, c_d \in O_d \cup \{\emptyset\}\}$, un alphabet d'activités.

Un symbole $\Leftrightarrow$ Une activité associée à un lieu et/ou un déplacement

Une **séquence sémantique** S est une séquence de symboles telle que $S \in \Sigma^n$.

Soit une séquence sémantique $S = \langle a_1, a_2, ..., a_n \rangle$, la **trajectoire sémantique** TS associée à S est telle que $TS = \langle ts_1, ts_2, ..., ts_n \rangle$ où $ts_i = (a_i, T_i)$. Où T_i est la trace associée à l'activité a_i.

Pour simplifier la compréhension, ici : un symbole $\Leftrightarrow$ un émoticône 😉

Exemple de trajectoires sémantiques

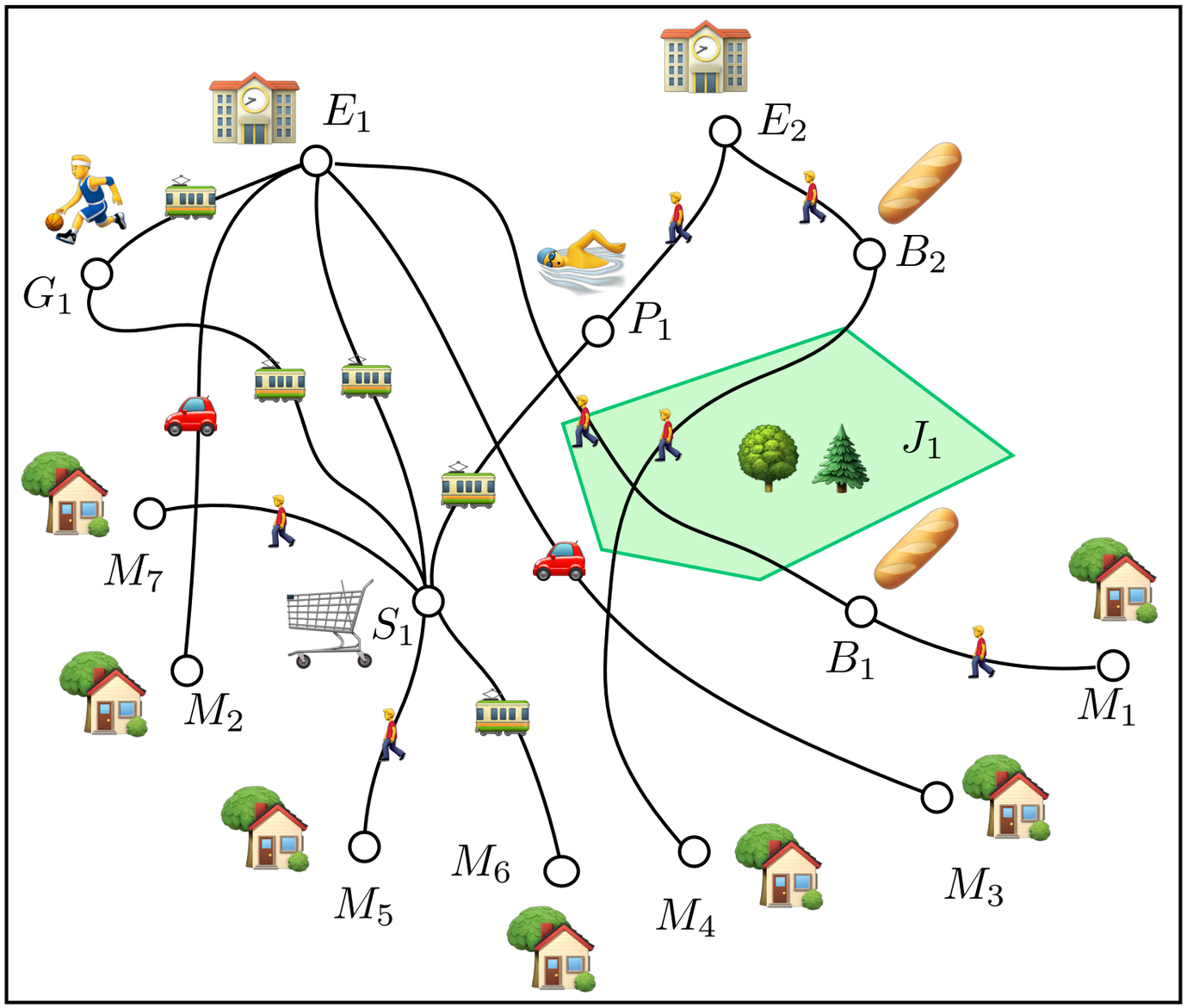


Figure *3* : Exemple de 7 trajectoires sémantiques d'enfants rentrant de leur école à leur maison

La trajectoire de l'enfant i se rendant à la maison M_i est appelée trajectoire sémantique TS_i .

On détaille la séquence sémantique S_6 issue de TS_6 .

$$
\begin{aligned}
S_6 = \langle & (\text{apprendre}, E_2), \\
& (\text{se déplacer, à pieds}), \\
& (\text{nager}, P_1), \\
& (\text{se déplacer, tramway}), \\
& (\text{faire ses courses}, S_1), \\
& (\text{se déplacer, tramway}), \\
& (\texttt{Null}, M_6)\rangle
\end{aligned}
$$

Sur la comparaison des entités

De cette modélisation doit découler une **métrique** permettant de comparer les trajectoires sémantiques entres elles.

1. Comment comparer *deux symboles* de l'alphabet Σ entre eux ?
2. Comment comparer *deux séquences sémantiques* ?
3. Comment comparer *deux trajectoires sémantiques* ?

   Plus globalement, comment intégrer les dimensions temporelle, spatiale et sémantique au sein d'une seule et unique métrique ?

Mesure de similarité sémantique sur le symbole

Grâce aux ***ontologies***, il est possible d'établir des mesures de similarité (ou proximité) entre deux concepts afin d'établir une similarité sémantique.

Dès lors, il est possible de construire une fonction de similarité Sim telle que :

$$\text{Sim} : \Sigma \times \Sigma \rightarrow [0,1]$$

Il existe de nombreuses métriques de similarité conceptuelles répertoriées dans (Aime, 2011 ; Harispe, 2014). Cependant :

1. Les mesures de similarité varient selon l'approche envisagée : intentionnelle (*feature based approach*), extensionnelle (*structural approach*), mixte.
2. Elles sont souvent contextuelles et dépendent du type d'applications.

$\Longrightarrow$ Il n'y a pas de métrique parfait, mais seulement des métriques plus adaptées selon les besoins utilisateurs et les ressources dont il dispose.

Graphe conceptuel

Les ontologies peuvent être résumées en partie par des **graphes conceptuels**.

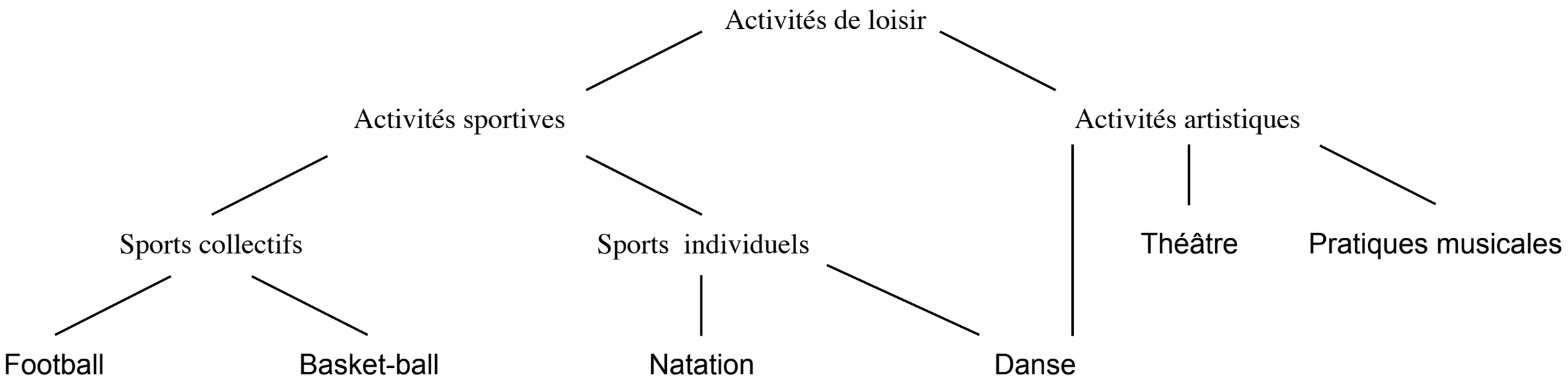


Figure *4* : Exemple de graphe de concepts des activités extra-scolaires

On peut calculer la similarité de deux concepts c_1, c_2 dans un tel graphe à l'aide de la mesure de Leacock par exemple :

$$\text{Sim}_{leacock}(c_1, c_2) = -\log\left(\frac{dist(c_1, c_2)}{max}\right)$$

Distances sur les séquences sémantiques

(Chen *et al*., 2005) propose une variante de la **distance d'édition** de (Wagner & Fischer, 1994), Edit Distance on Real séquence (EDR).

Celle-ci doit être modifiée pour s'adapter aux séquences sémantiques.

La précédente mesure de similarité permet d'établir de **nouveaux opérateurs d'édition** qui se base sur la proximité des activités (i.e. symboles de Σ) entre elles.

Ainsi, on nomme **distance d'édition sémantique** $d_{\mathcal{S}} : \Sigma^n \times \Sigma^p \to \mathbb{R}^+$ la *mesure de proximité entre deux séquences sémantiques* selon la distance d'édition enrichie.

Exemple d'édition d'une séquence sémantique

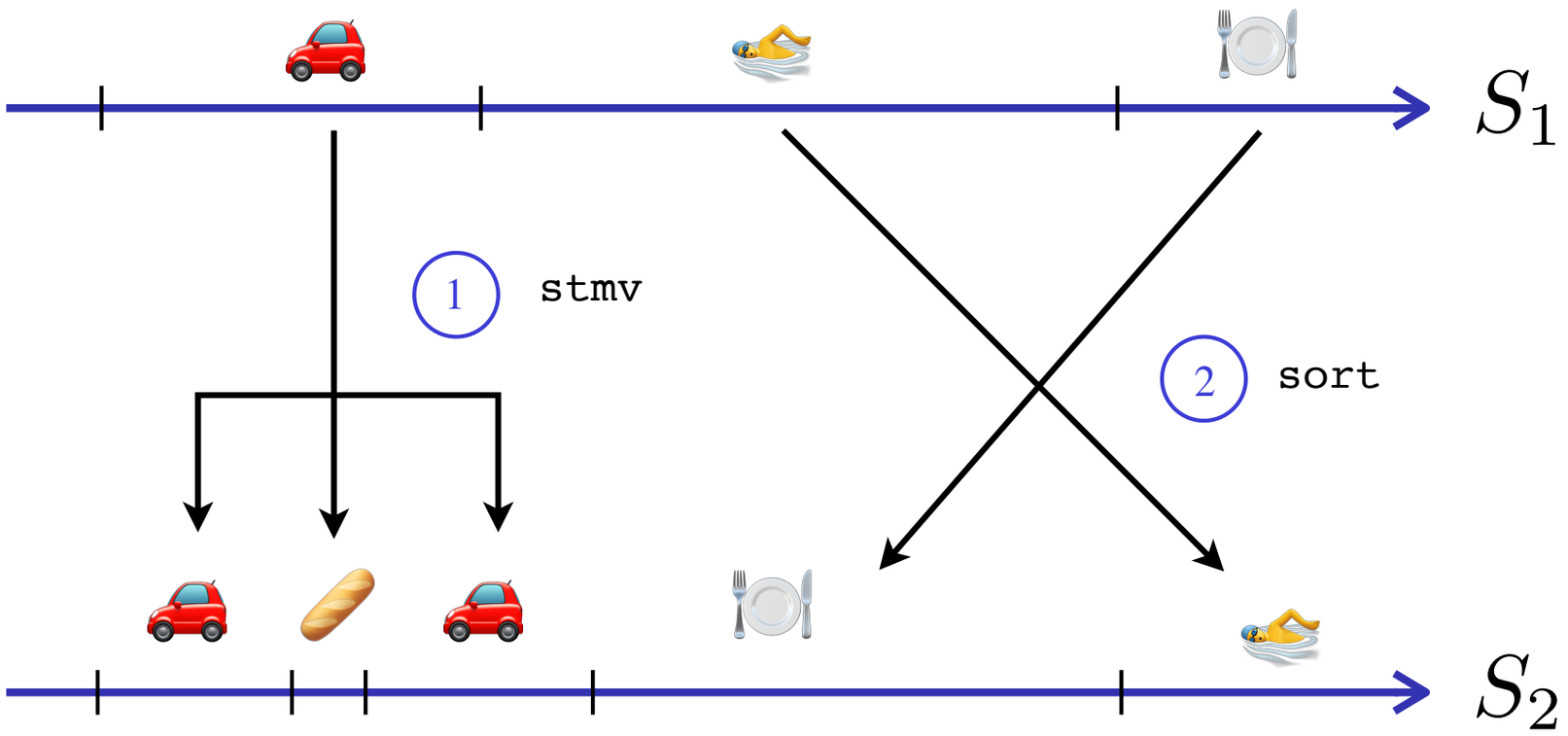


Figure *5* : Edition d'une séquence sémantique S_1 en une S_2

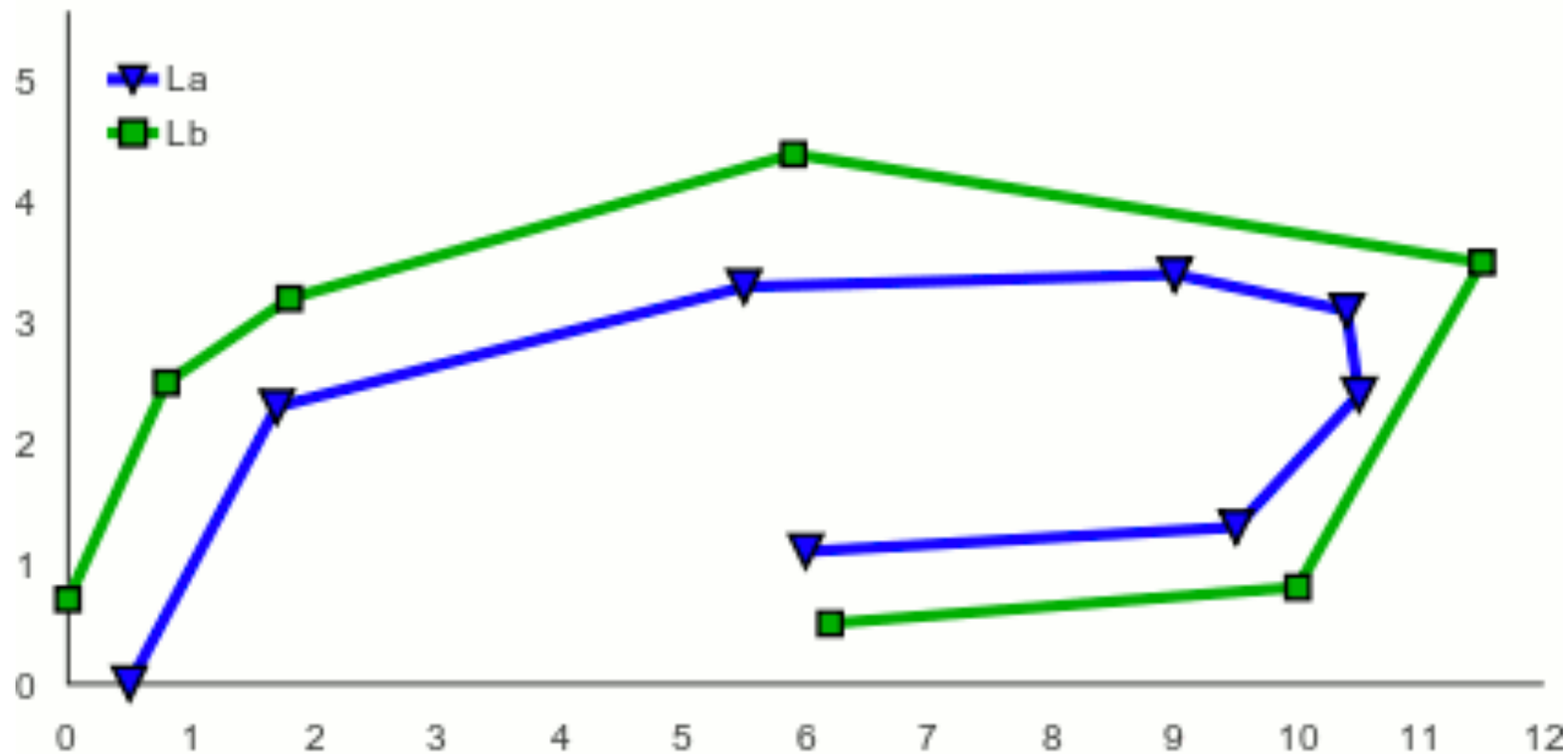
On dispose au *total de 6 opérations d'édition* possibles pour transformer S_1 en S_2.

On applique **2 opérateurs ici contre 4 pour une distance d'édition classique (EDR)**.

Possibilité de faire varier le coût des l'opérateurs selon la vision utilisateur.

Similarité spatiale : Distance de Fréchet

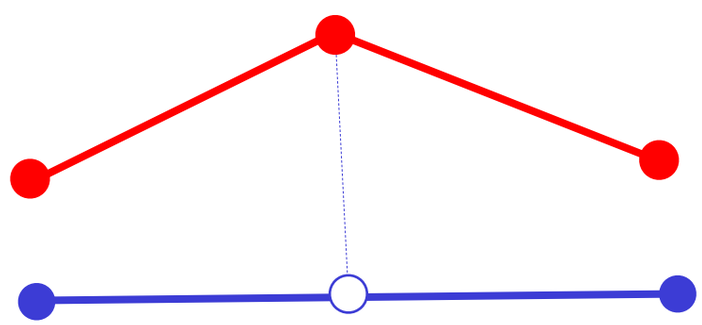
- Distance Maximale
  - Entre deux polylignes
    - Ensemble ordonné de points
- Version discrète
  - Approximation
    - Programmation dynamique
- Uniquement les extrémités sont prises en compte
- Temps de calcul fortement réduit



Optimisation de la distance de Fréchet

(Financement ICVL)

- Filtrage du nombre de points
- Ajout de points saillants
  - **Perpendiculaires aux extrémités** (2017)
    - Passage de + de 8 h à quelques minutes avec une précision équivalent à celle du GPS pour des gros jeux de données réels
  - **Points tournant** (2018)
    - Travaux en cours, des cas limites ne sont pas encore pris en compte
- Travaux réalisés par des stagiaires de Tours et d'Orléans et encadrés par T. Devogele, L. Etienne, E. Melin, S. Robert



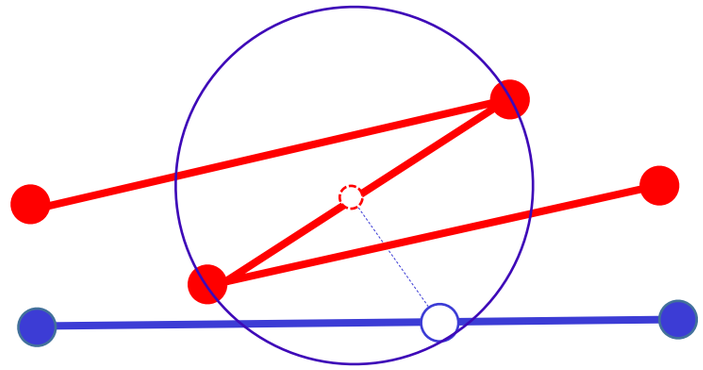
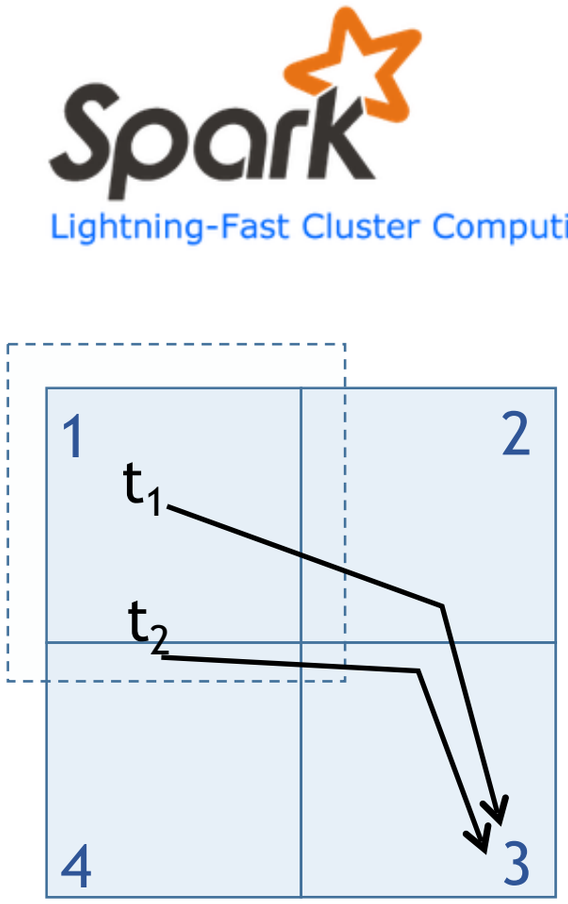


Figure *6* : Exemple d'ajouts de points saillants

Parallélisation (Spark)

- Pour de très larges volumes de trajectoires répartition des mesures de distances à l'aide du framework *Spark*
- Problème d'optimisation
  - Définir la **répartition logique des données** qui permettra de réaliser le minimum de calculs, basée sur :
    - Un **index spatial** (*Fixed Grid avec Overlapping* ou *QuadTree* avec Overlapping)
    - Une **clé double** de répartition (cellule de la position de départ, cellule de la position d'arrivé)
- Premier test (2018)
  - Démontrer l'intérêt de ces indexes
  - Avantage de l'overlapping pour éviter des calculs inutiles
- Travaux en cours

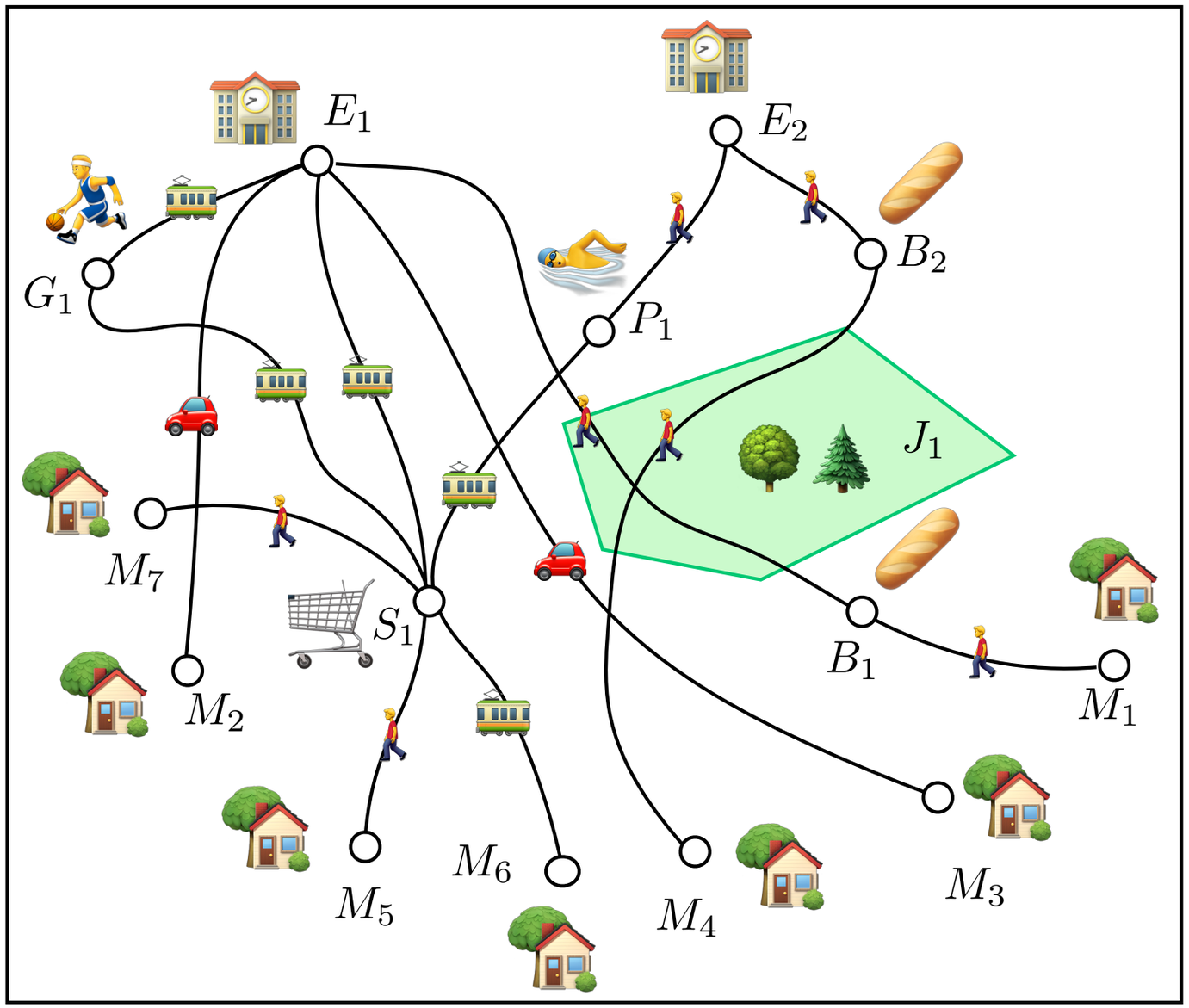


Clé de t_1 : (1,3)

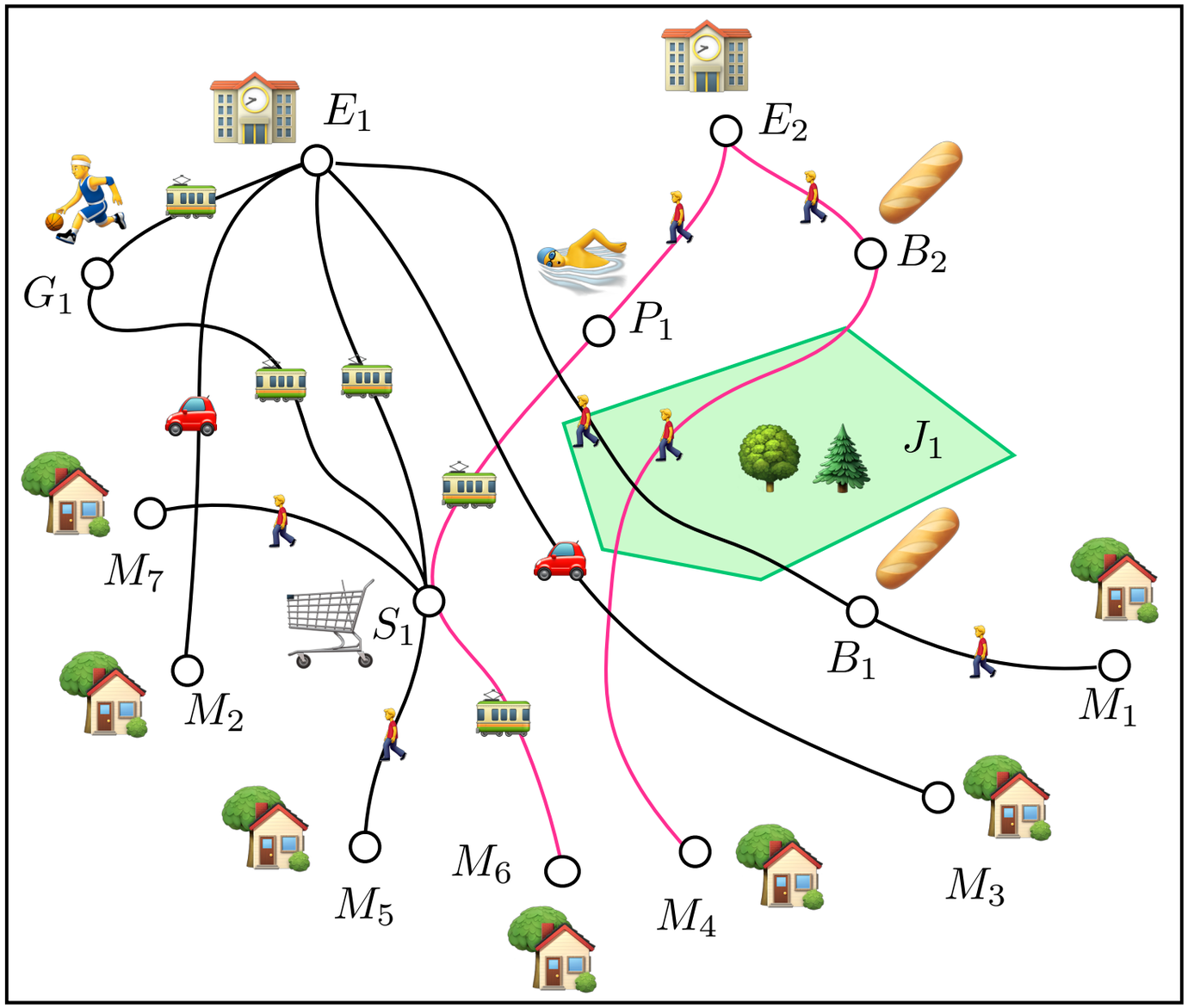
Clé de t_2 : (4,3) et aussi (1,3)

Figure 7 : Index spatial par Fixed Grid avec Overlapping

Clustering des trajectoires sémantiques



Clustering selon la dimension spatiale



* $C_1 = \{TS_4, TS_6\}$

Figure *8* : Partitionnement en clusters spatiaux

Clustering selon la dimension spatiale

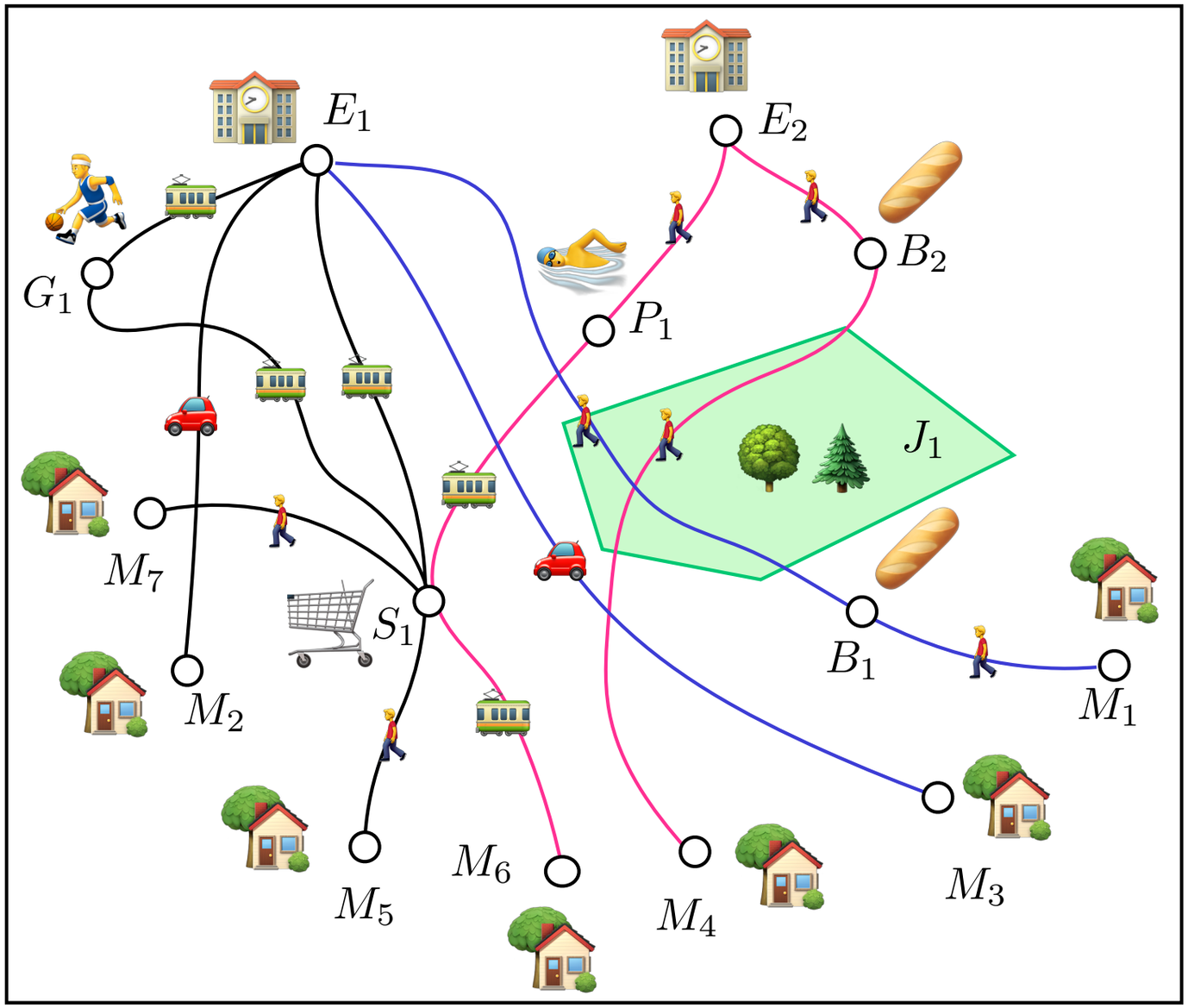
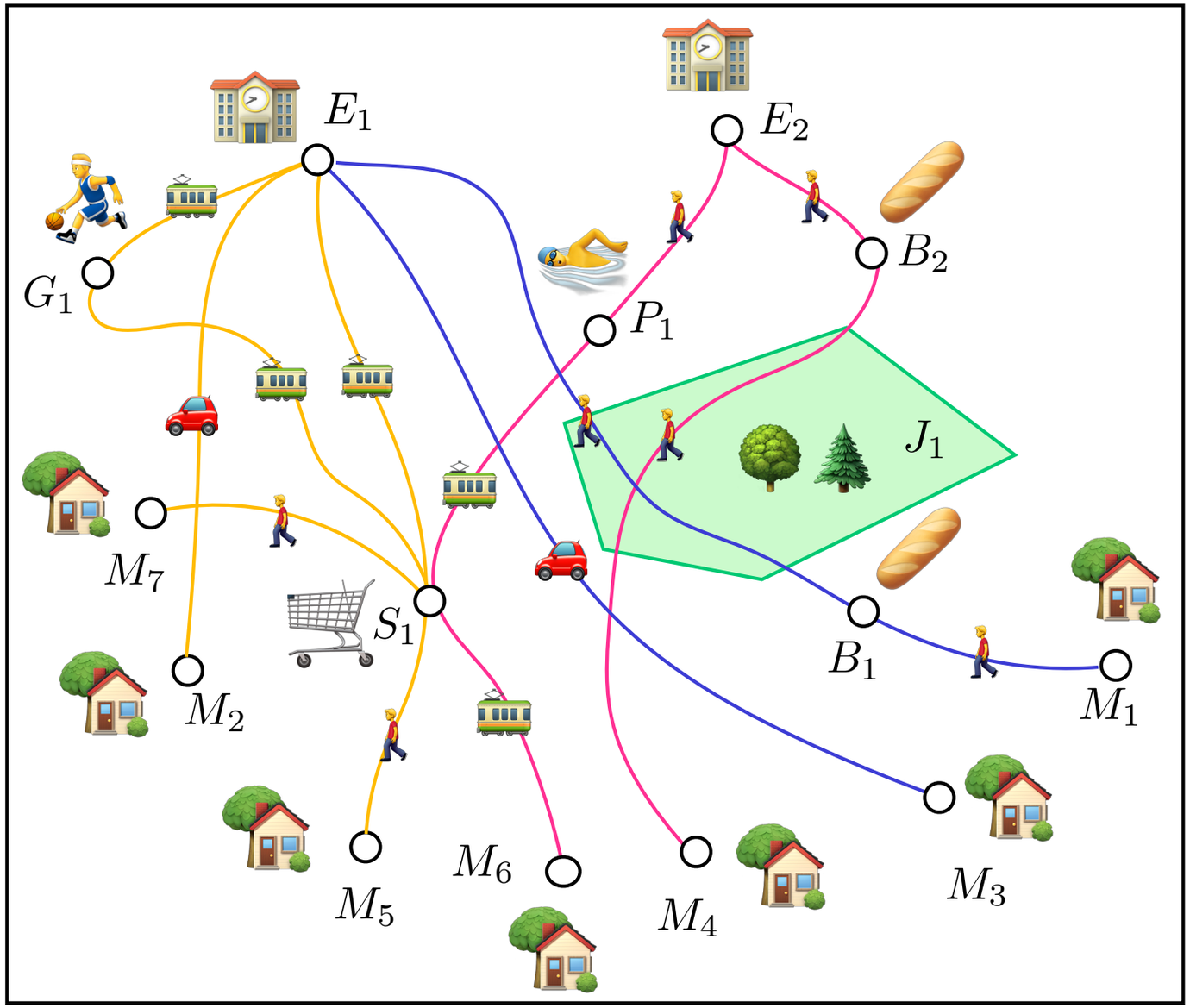


Figure *8* : Partitionnement en clusters spatiaux

* $C_1 = \{TS_4, TS_6\}$
* $C_2 = \{TS_1, TS_3\}$

Clustering selon la dimension spatiale



* $C_1 = \{TS_4, TS_6\}$
* $C_2 = \{TS_1, TS_3\}$
* $C_3 = \{TS_2, TS_5, TS_6\}$

Figure *8* : Partitionnement en clusters spatiaux

Clustering selon la dimension sémantique

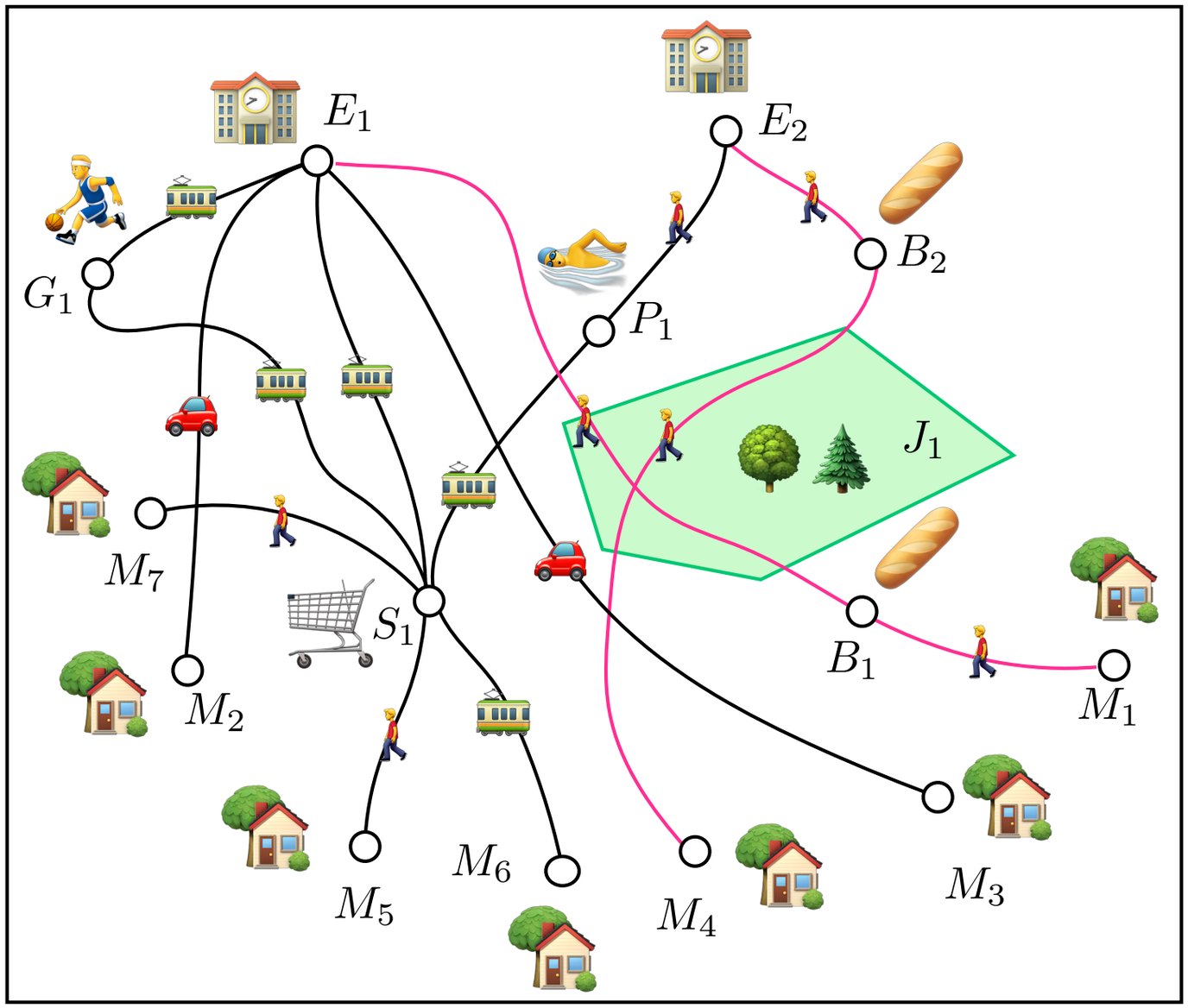
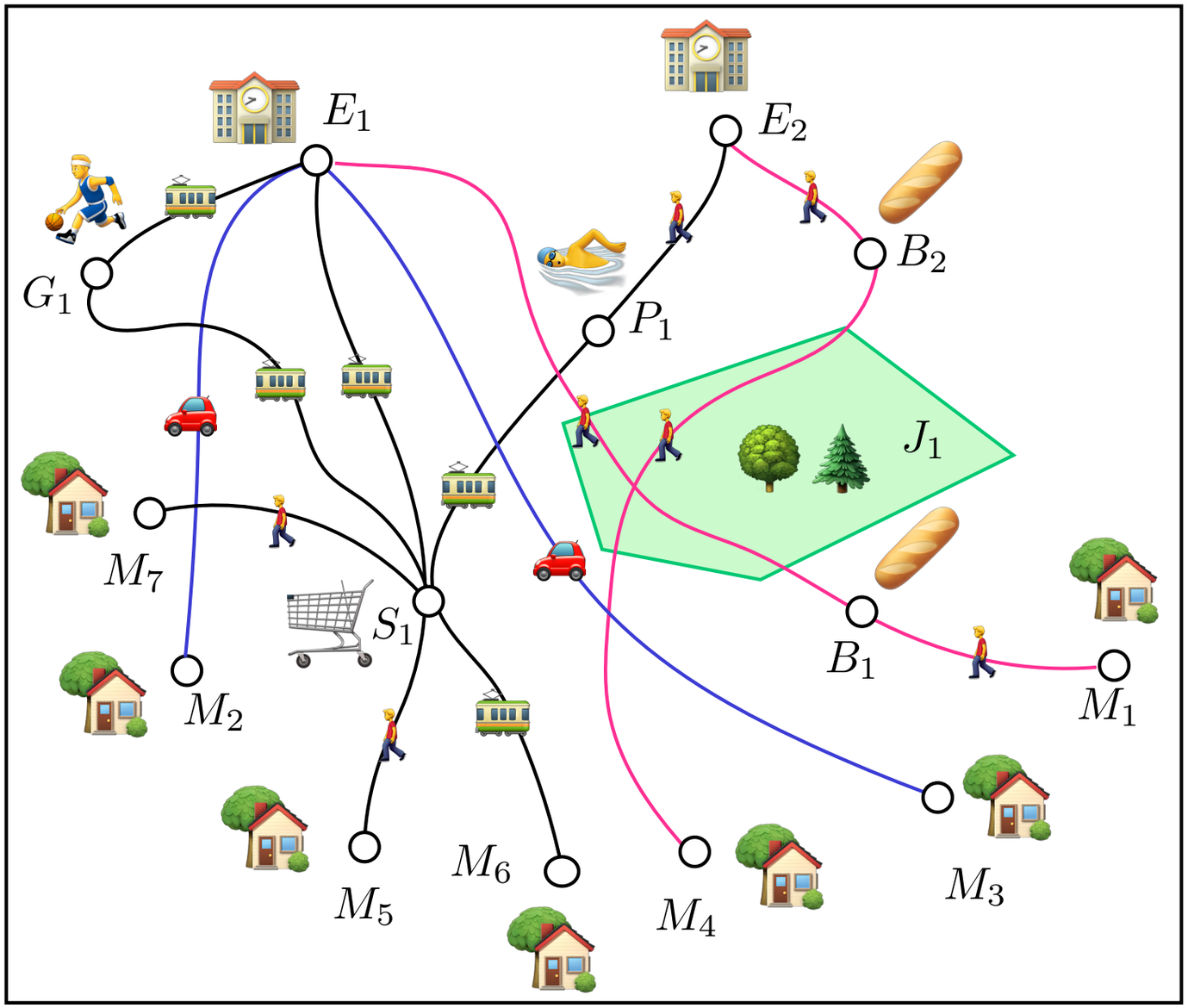


Figure *9* : Partitionnement en clusters sémantiques

* $C_1 = \{TS_1, TS_4\}$

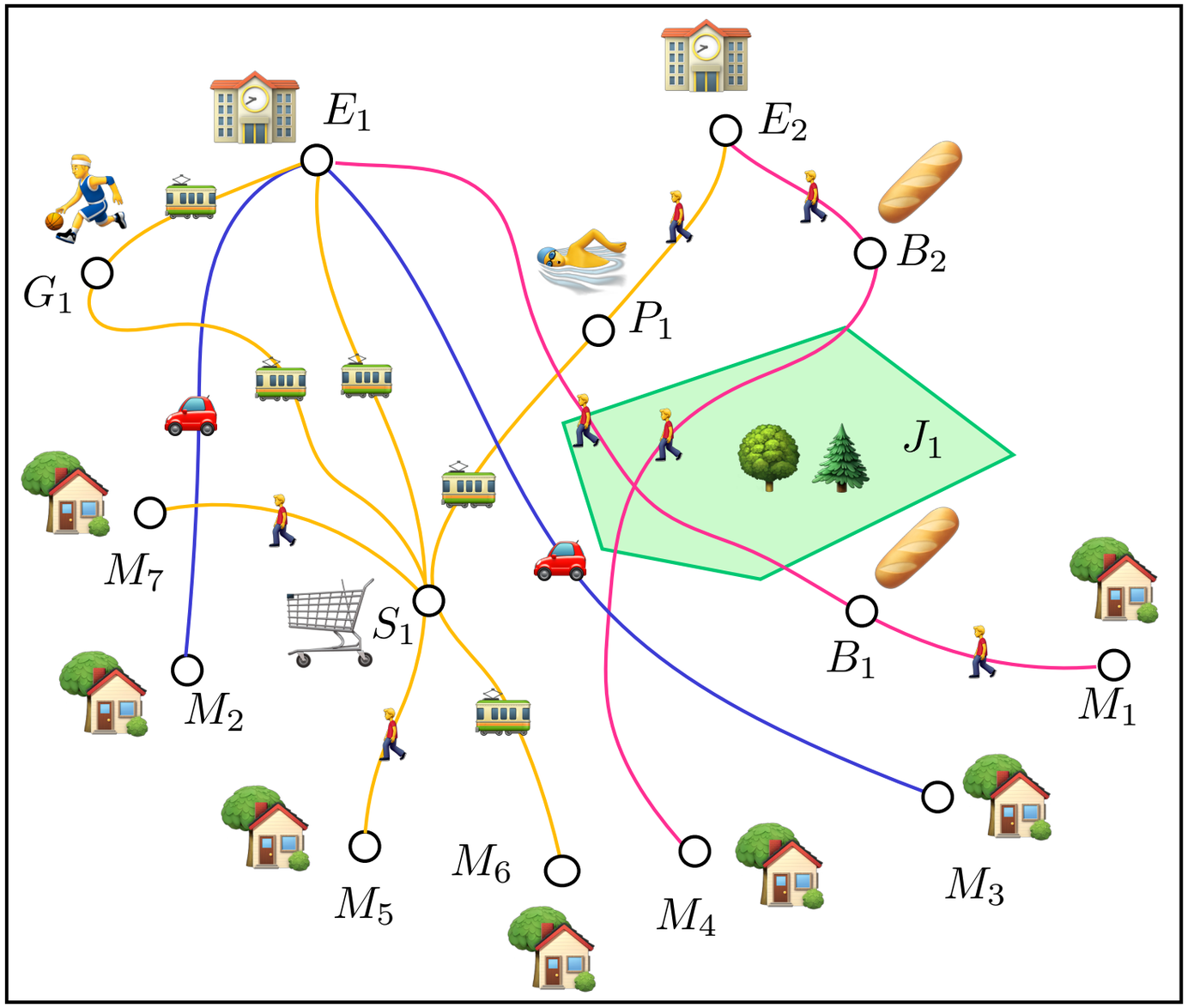
Clustering selon la dimension sémantique



* $C_1 = \{TS_1, TS_4\}$
* $C_2 = \{TS_2, TS_3\}$

Figure *9* : Partitionnement en clusters sémantiques

Clustering selon la dimension sémantique



* $C_1 = \{TS_1, TS_4\}$
* $C_2 = \{TS_2, TS_3\}$
* $C_3 = \{TS_5, TS_6, TS_7\}$

Figure *9* : Partitionnement en clusters sémantiques

Évaluation des dimensions

On a réussi à évaluer, *a priori*, la composante sémantique de la trajectoire sémantique.

Maintenant, **comment mêler les autres dimensions** (temporelle et spatiale) à cette proximité sémantique établie ?

1. *Étudier chaque dimension indépendamment* l'une de l'autre puis les agréger ensemble.
2. Déterminer une *mesure sémantico-spatio-temporelle* qui traite les trois composantes du même coup.
3. *Poser des contraintes d'analyse* sur chaque dimension afin de se ramener à l'étude principale d'une composante. (Zhang *et al*, 2014)

On choisit [pour le moment] la **première méthode**.

Parti pris également de (Furtado *et al*, 2016) dans son élaboration de métrique multi-dimensionnelle.

Synthèse des clusters

La **synthèse de déplacements** peut-être effectuée d'un *point de géométrique* par des boîtes à moustaches spatio-temporelles (Etienne *et al*, 2014) ; ou bien d'un ***point de vue sémantique***.

Une telle synthèse permet entre autres :

* De *représenter en substance* un ensemble vaste de trajectoires,
* De résumer de *manière anonyme* un ensemble de données,
* Déterminer les *actions caractéristiques* d'un ensemble d'individus.

On exécute une ***inférence grammaticale*** (du Mouza et Rigaux, 2005) sur les partitions afin d'extraire un **automate** correspondant à l'ensemble des séquences sémantiques du cluster.

Automate de synthèse sémantique

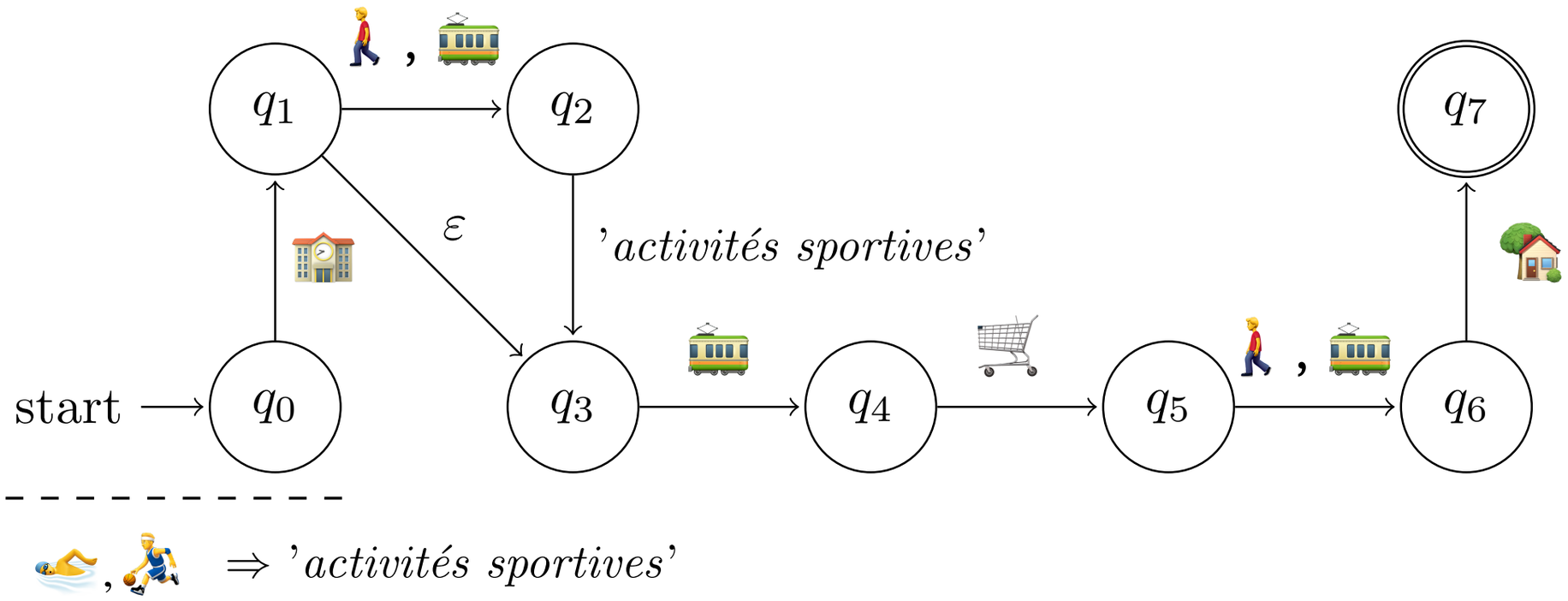


Figure *10* : Automate représentant le motif synthétique de la partition $\{TS_5, TS_6, TS_7\}$ de trajectoires et inférence logique sur les activités sportives

La notion de *grammaire algébrique probabiliste* (Li et Lee, 2017) peut rendre compte de la dimension stochastique des déplacements en considérant la notion de fréquence.

Contributions et conclusion

1. Apport d'un **nouveau modèle** afin de représenter conjointement les dimensions temporelle, spatiale et sémantique.

2. Construction d'une **métrique sémantique** pour évaluer la similarité entre deux instances de ce modèle et de **nouveaux opérateurs d'édition**.

3. Amélioration de la **distance de Fréchet discrète.**

4. **Méthode de synthèse** des clusters de trajectoires selon la dimension sémantique.

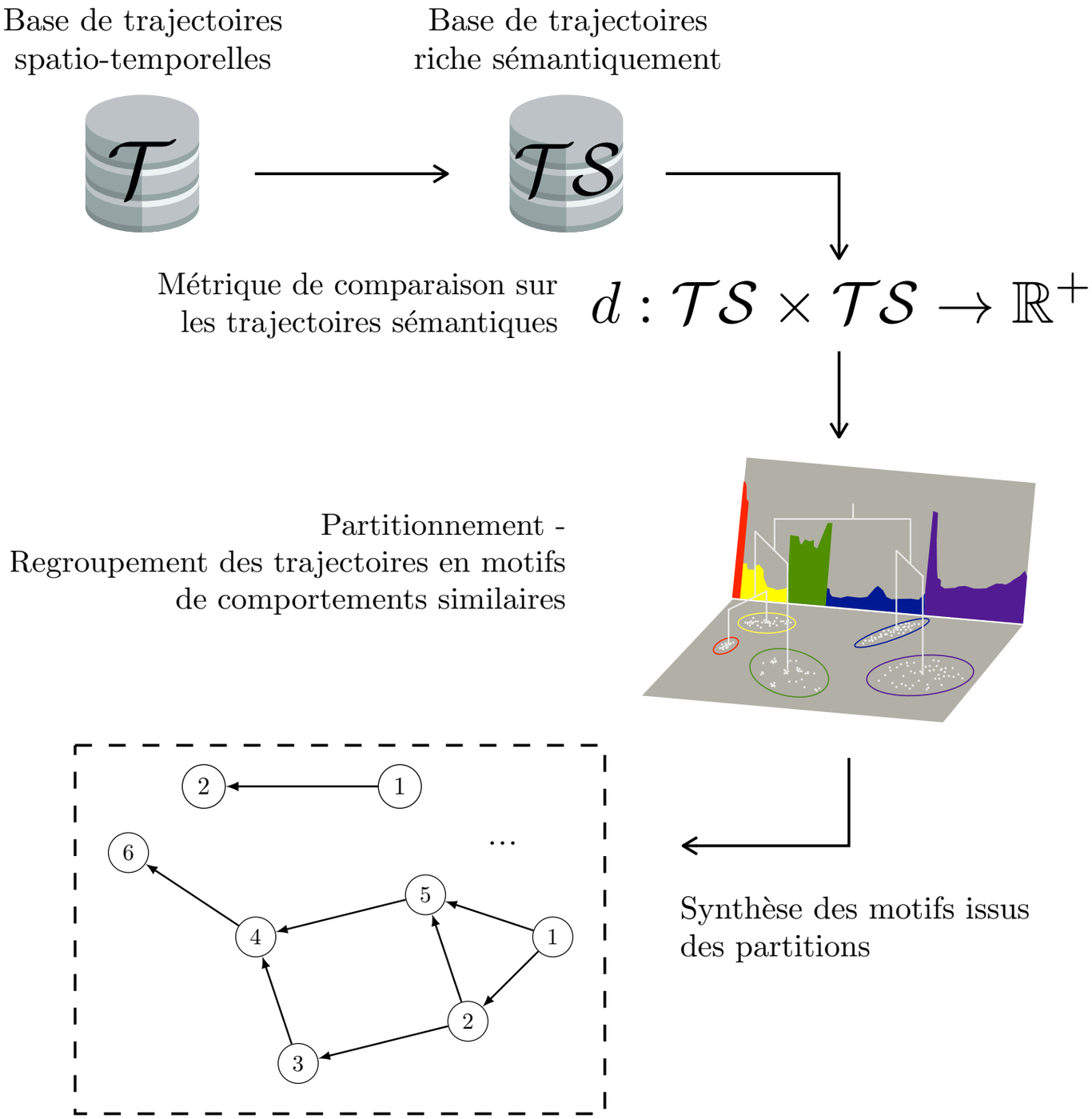
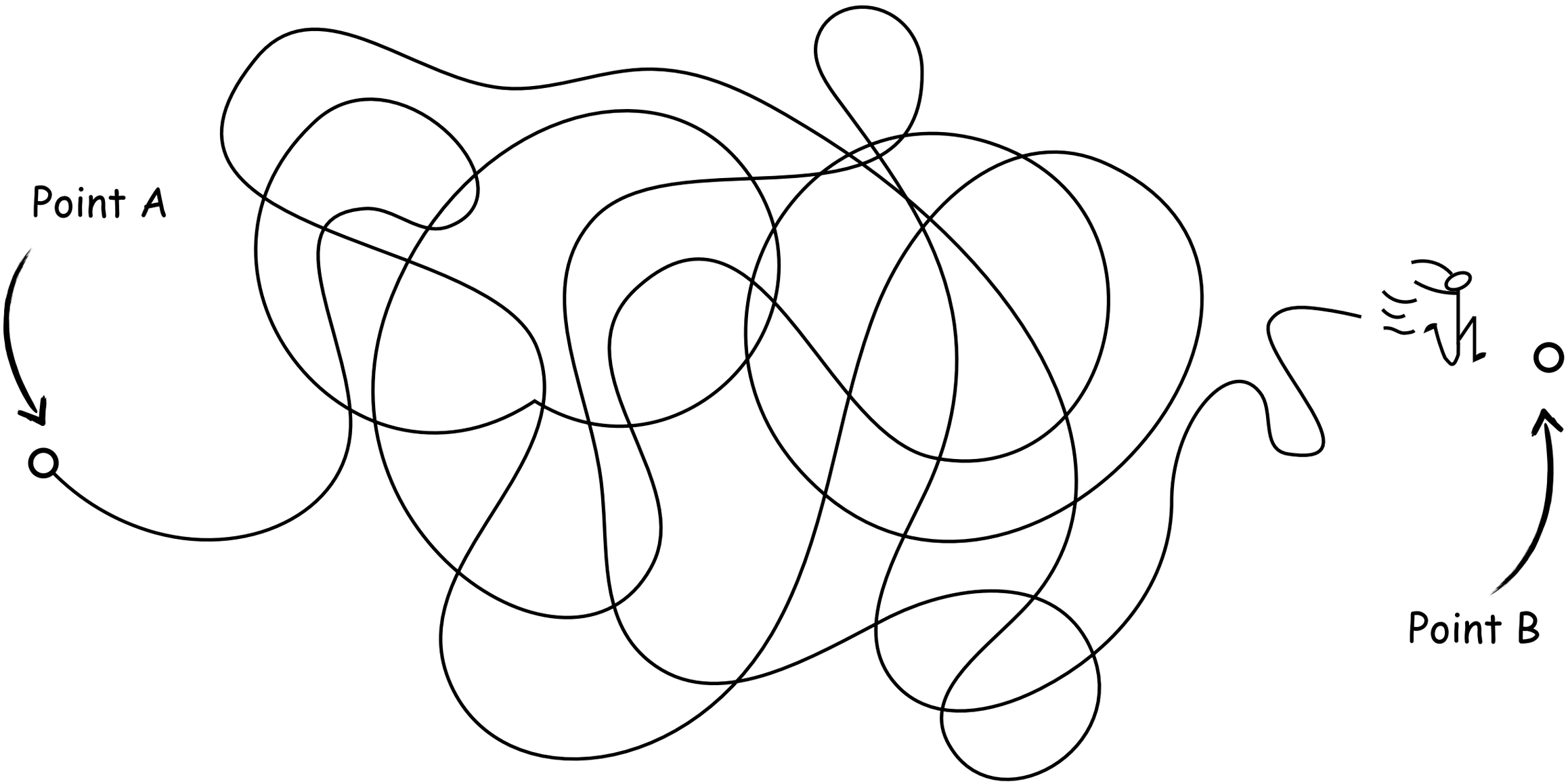


Figure *11* : Chaîne de traitement pour l'extraction de motifs de trajectoires sémantiques similaires

Perspectives

Ces travaux forment un opuscule sur différents challenges et possibilités du point de vue de l'enrichissement des trajectoires sémantiques.

* Approfondir les notions liées à l'orthogonalité des dimensions temporelle, spatiale et sémantique.
* Valider la pertinence des nouveaux opérateurs d'édition définis via une expérience psychologique.
* Explorer plus en détails la notion de grammaire stochastique pour la synthèse de clusters sémantiques.
* *Mise en pratique* des modèles présentés sur les données Mobi'Kids et Smart Loire.
* Parallélisation de la distance de Fréchet (dans le cadre de la collaboration ICVL Tours-Orléans).



Merci de votre attention

Bibliographie

Aime X. (2011). *Gradients de prototypicalité, mesures de similarité et de proximité sémantique : une contribution à l'ingénierie des ontologies*. Thèse de doctorat, Université de Nantes.

Alvares L., Bogorny V., Kuijpers B., Macedo J. de, Moelans B., Vaisman A. (2007). A model for enriching trajectories with semantic geographical information. *Proc. of the 15th annual ACM international symposium on Advances GIS*, nº 22, p. 1–8.

Beber M., Ferrero C., Fileto R., Bogorny V. (2017). Individual and group activity recognition in moving object trajectories. *Journal of Information and Data Management*, vol. 8, nº 1, p. 50–66.

Bogorny V., Renso C., Aquino A. R. de, Lucca Siqueira F. de, Alvares L. (2014). Constant - a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS*, vol. 18, p. 66–88.

Chen L., Özsu M. T., Oria V. (2005). Robust and fast similarity search for moving object trajectories. *Proc. of the 2005 ACM SIGMOD*, p. 491–502.

Choi D., Pei J., Heinis T. (2017). Efficient mining of regional movement patterns in semantic trajectories. *Proc. of the VLDB*, vol. 10, p. 2073–2084.

Devogele T. (2002). A new merging process for data integration based on the discrete Fréchet distance. In Advances in spatial data handling, p. 167–181. Springer.

Etienne L., Devogele T., Buchin M., McArdle G. (2016). Trajectory box plot; a new pattern to summarize movements. *International Journal of GIS*, vol. 30, p. 835–853.

Ferrero C., Alvares L., Bogorny V. (2016). Multiple aspect trajectory data analysis: Research challenges and opportunities. *GeoInfo*, p. 56–67.

Furtado A., Kopanaki D., Alvares L., Bogorny V. (2016). Multidimensional similarity measuring for semantic trajectories. *Transactions in GIS*, vol. 20, p. 280–298.

González M., CA.Hidalgo, Barabási A.-L. (2008). Understanding individual human mobility patterns. *Nature*, vol. 453, p. 779–782.

Gibert K., Valls A., Batet M. (2013). Introducing semantic variables in mixed distance measures: Impact on hierarchical clustering. *Knowledge and Information Systems*, vol. 40, p. 559–593.

Hägerstrand T. (1970). What about people in regional science ?. *Papers of the Regional science association*. 24 (1): 6–21

Harispe S. (2014). *Knowledge-based Semantic Measures: From Theory to Applications*. Thèse de doctorat, Université de Montpellier.

Li, S. et Lee, D. (2017). Learning daily activity patterns with probabilistic grammars. *Transportation*, 44:49–68.

Mouza C. du, Rigaux P. (2005). Mobility patterns. *GeoInfo*, vol. 9, p. 297–319.

Noël D., Villanova-Oliver M., Gensel J., Quéau P. L. (2015). Modeling semantic trajectories including multiple viewpoints and explanatory factors: Application to life trajectories. *Proc. of the 1st International ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, p. 107–113.

Parent C., Spaccapietra S., Renso C., Andrienko G., Bogorny V., Damiani M. *et al*. (2013). Semantic trajectories modeling and analysis. *ACM Computing Surveys*, vol. 45, p. 1–32.

Renso C., Trasarti R. (2013). Mobility data : Modeling, management and understan- ding. In, p. 129–151. Cambridge University Press.

Sakoe H., Chiba S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on ASSP*, vol. 26, p. 43–49.

Song C., Qu Z., Blumm N., Barabási A.-L. (2010). Limits of predictability in human mobility. *Science*, vol. 327, p. 1018–1021.

Wagner R., Fisher M. (1994). The string-to-string correction problem. *Journal of the ACM*, vol. 21, p. 168–173.

Xu Z., Da Q. (2003). An overview of operators for aggregating information. *Inter- national Journal of intelligent systems*, vol. 18, p. 953–969.

Zhang C., Han J., Shou L., Lu J., Porta T. L. (2014). Splitter: Mining fine-grained sequential patterns in semantic trajectories. *VLDB Endowment*, p. 769-780.