Tirer parti de la structure des données incertaines

Antoine Amarilli

21 avril 2017

Les quatre ‘V’ des données massives

Volume : données très grandes

Variété : données dans des formats différents

Vélocité : données modifiées à grande vitesse

Véracité : données incertaines, imprécises, erronées

Les quatre ‘V’ des données massives

Volume : données très grandes

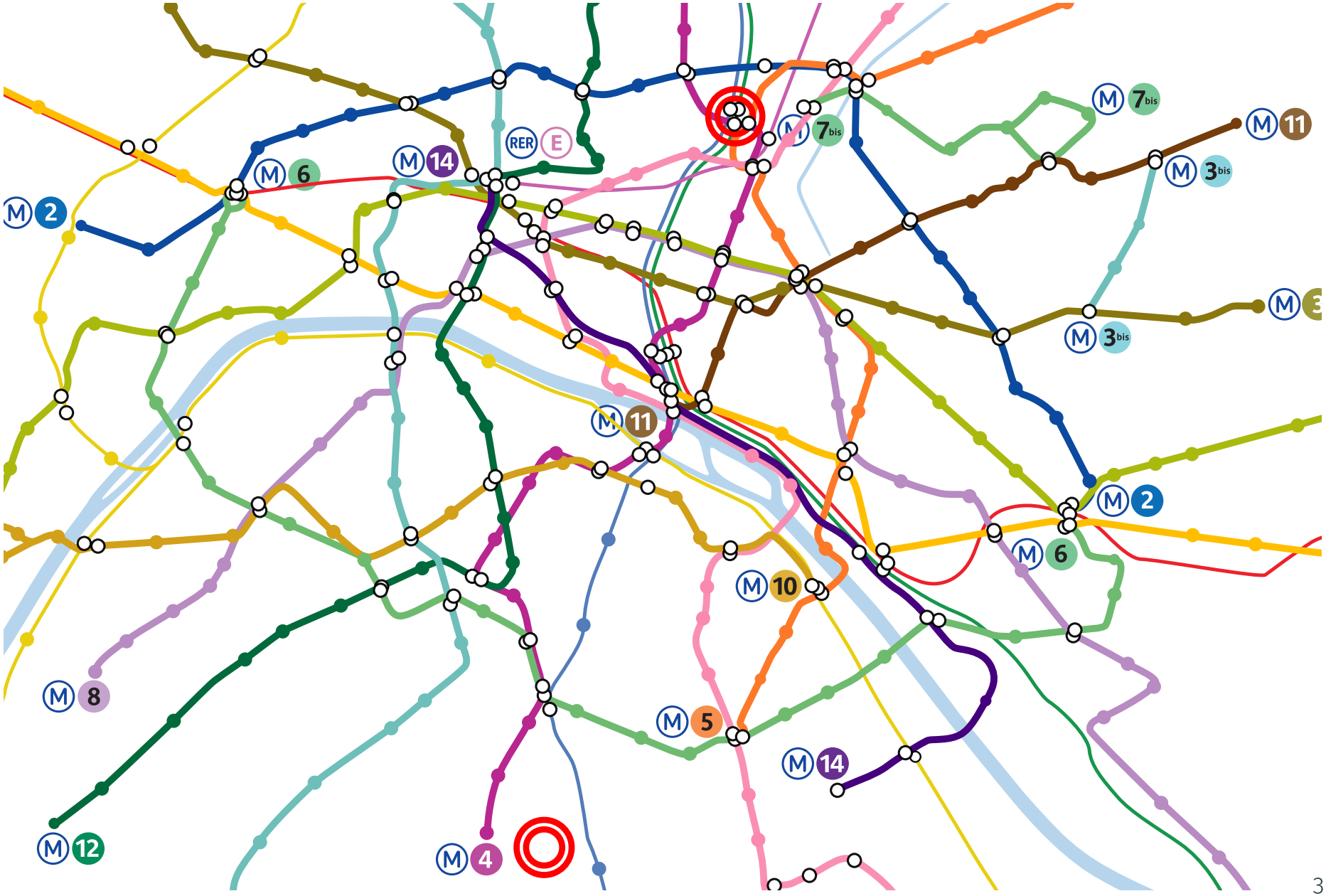
Variété : données dans des formats différents

Vélocité : données modifiées à grande vitesse

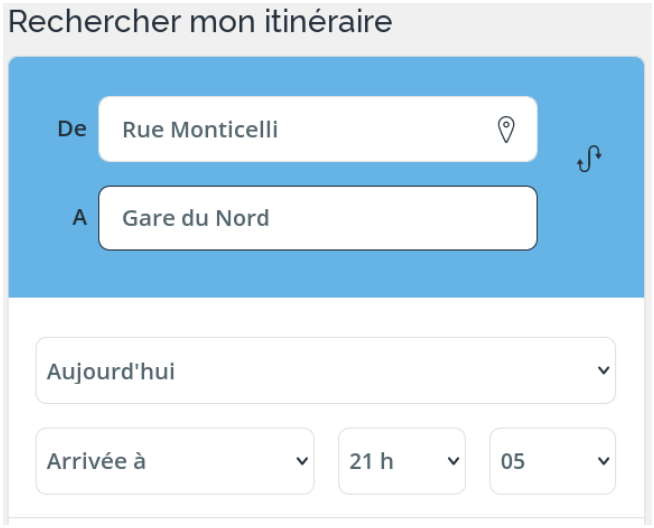
⇒ **Véracité :** données incertaines, imprécises, erronées ⇐

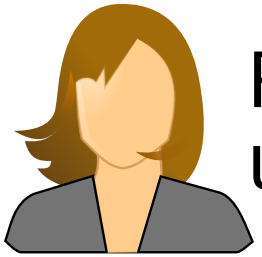
Exemple : calcul de trajets en métro

Exemple : calcul de trajets en métro



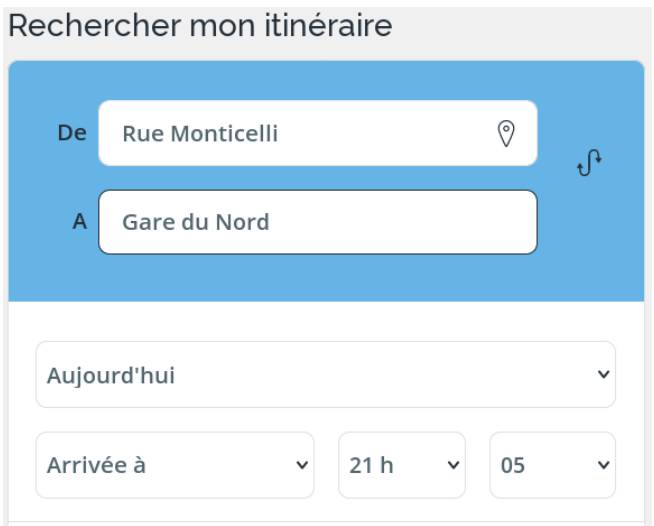
Exemple : calcul de trajets en métro

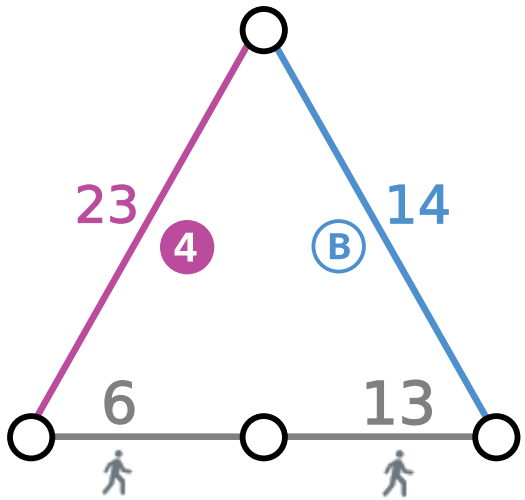


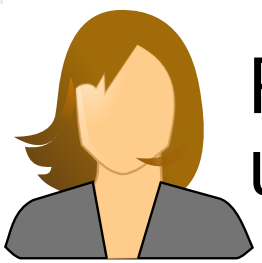


Requête utilisateur

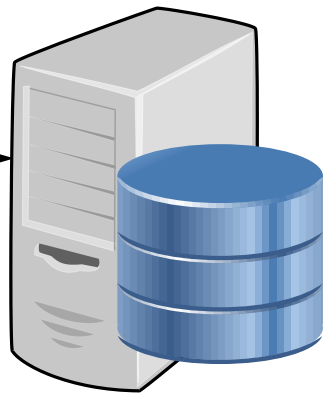
Exemple : calcul de trajets en métro





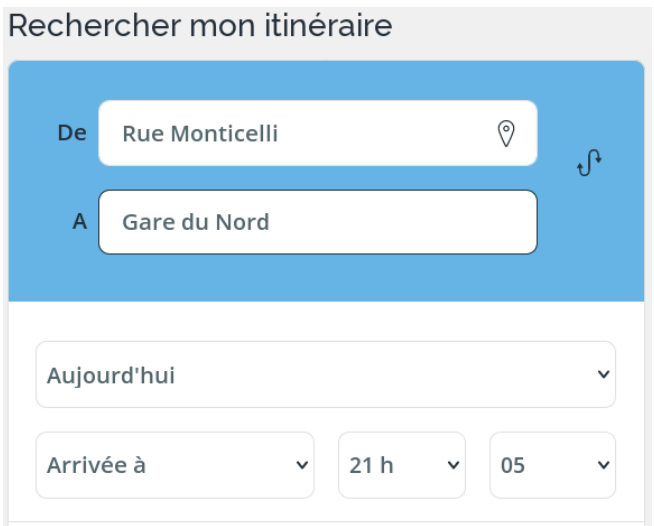


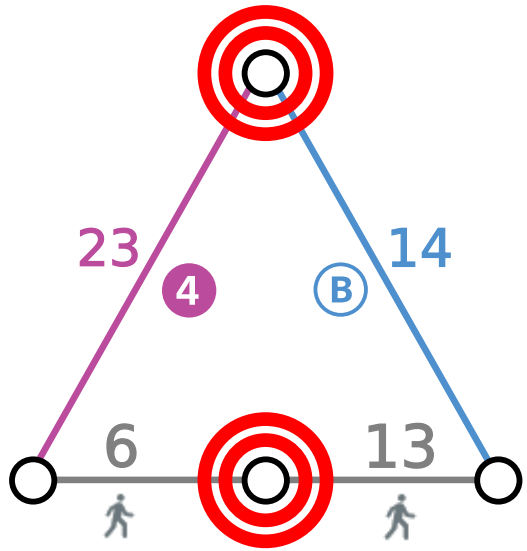
Requête utilisateur

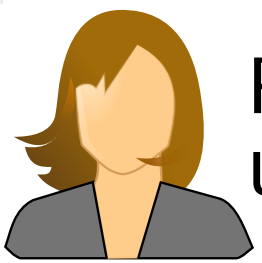


Base de données

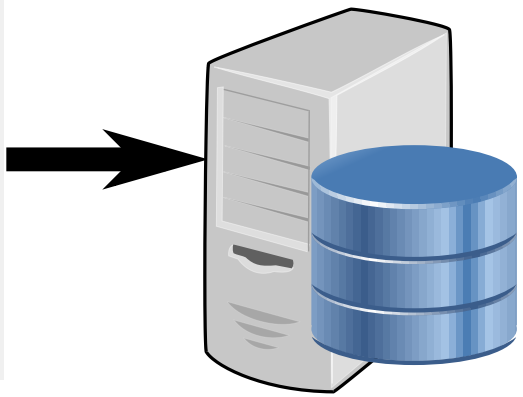
Exemple : calcul de trajets en métro





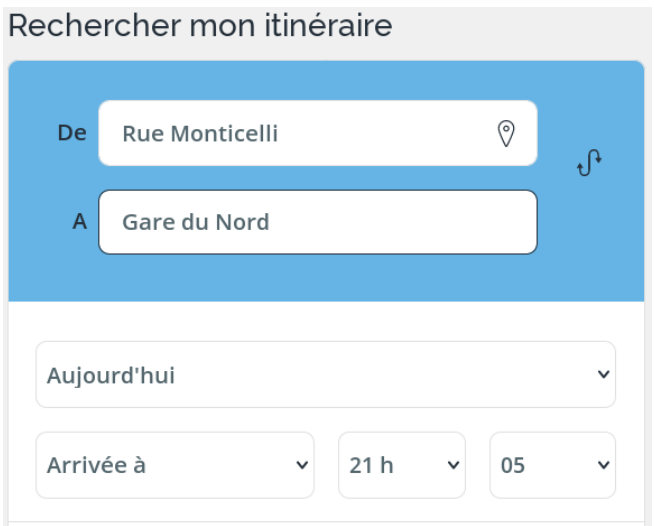


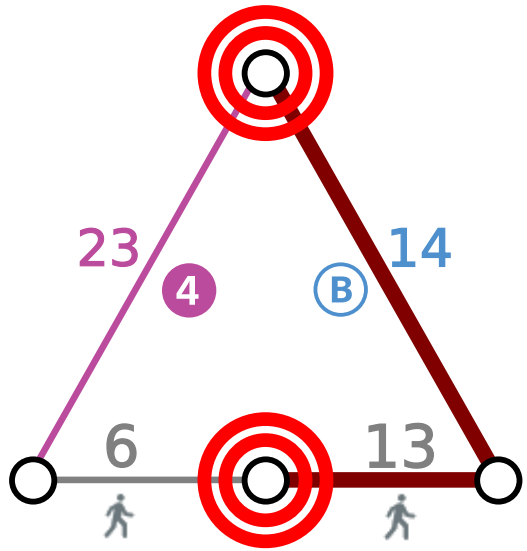
Requête utilisateur

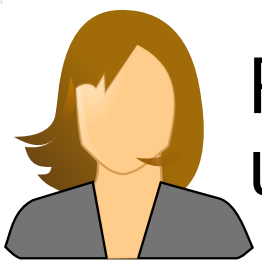


Base de données

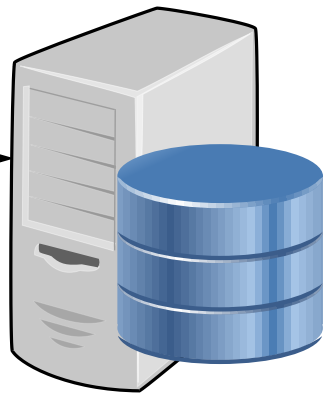
Exemple : calcul de trajets en métro





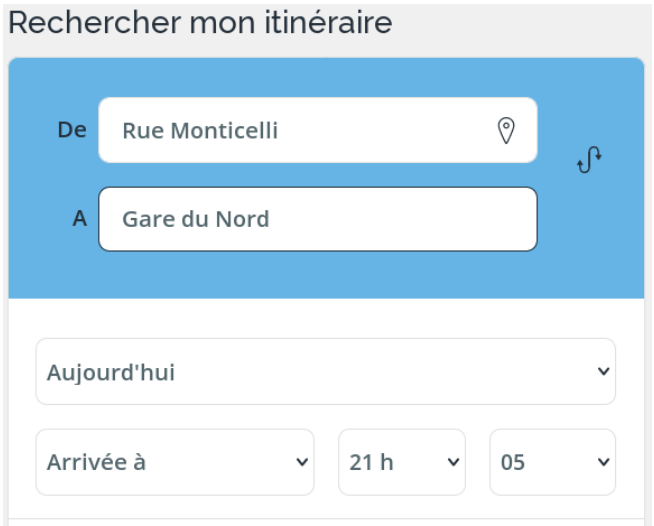


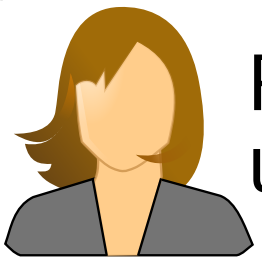
Requête utilisateur



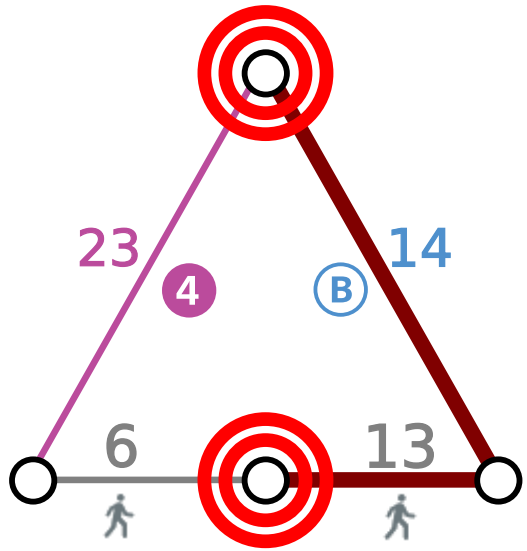
Base de données

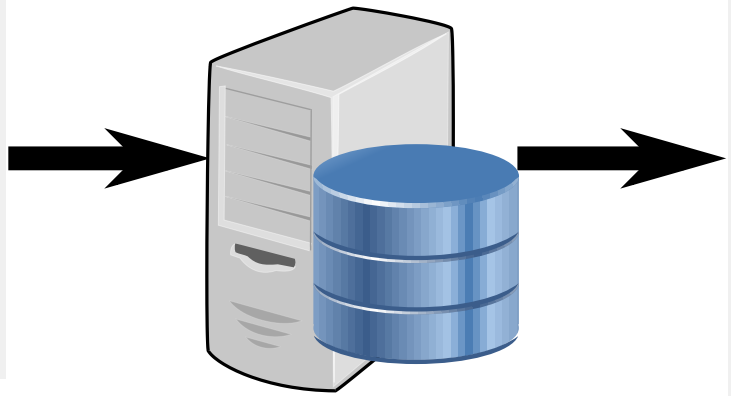
Exemple : calcul de trajets en métro





Requête utilisateur

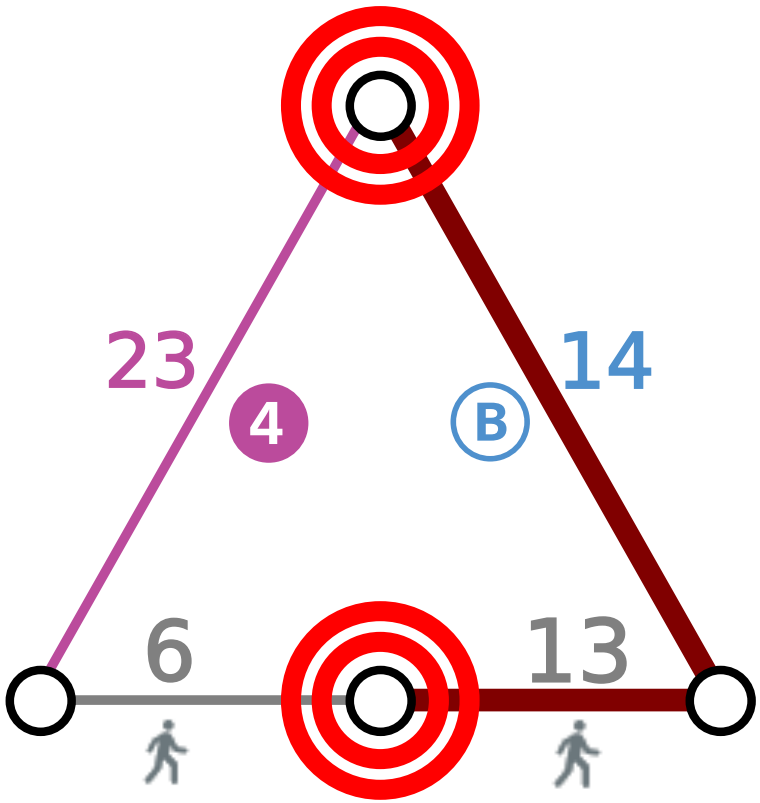




Base de données

Résultat

Exemple : calcul de trajets en métro



Départ

20h17 - 5 rue Monticelli, Paris

1.1 km | 13 min

B

20h30 - CITE UNIVERSITAIRE, Paris

RER B - EPAU
Vers **Aéroport CDG Terminal 2 TGV**

6 arrêts | 14 min

Arrivée

20h44 - GARE DU NORD RER, Paris

Exemple : calcul de trajets en métro

Panne du RER B : trafic interrompu entre Paris et Roissy, des TGV en renfort

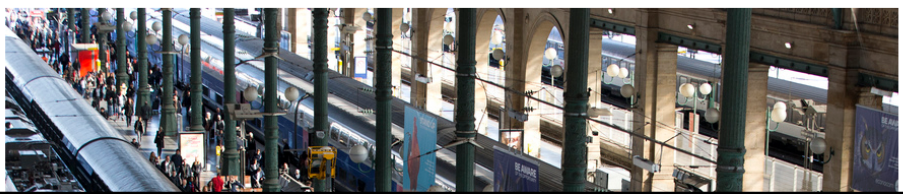
> Transports | 06 décembre 2016, 9h56 | MAJ : 06 décembre 2016, 17h03 |

Exemple : calcul de trajets en métro

Panne du RER B : trafic interrompu entre

Paris : pourquoi il y a autant de perturbations sur le RER B et à Gare du Nord

La circulation de l'ensemble des trains au départ de gare du Nord est totalement interrompue à la suite d'une panne électrique.

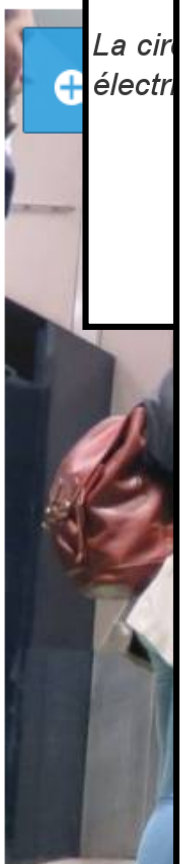


Exemple : calcul de trajets en métro

Panne du RER B : trafic interrompu entre

Paris : pourquoi il y a autant de perturbations sur le

La cir
électr



INCIDENT SUR LE RER B : QUE S'EST-IL PASSÉ CE MATIN ?

Malaise voyageur et application des mesures de sécurité : pour quelles raisons le trafic a-t-il été perturbé ce matin sur la ligne B ?

Pour beaucoup, le voyage a été difficile ce matin. Au fil de vos réactions sur Twitter notamment, je constate que les raisons de ces perturbations ne paraissent pas cohérentes. Je tiens donc à vous apporter des premiers éléments d'explication, que nous pourrons développer

Exemple : calcul de trajets en métro

Panne du RER B : trafic interrompu entre

Paris : pourquoi il y a autant de perturbations sur le

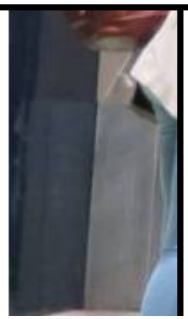
INCIDENT SUR LE RER B · QUE

ACTUALITÉS

Le RER B en panne, les voyageurs n'ont pas eu d'autre choix que de descendre sur les voies

Alors que la circulation alternée a augmenté le nombre de voyageurs dans les transports en commun, le RER B s'est retrouvé à l'arrêt.

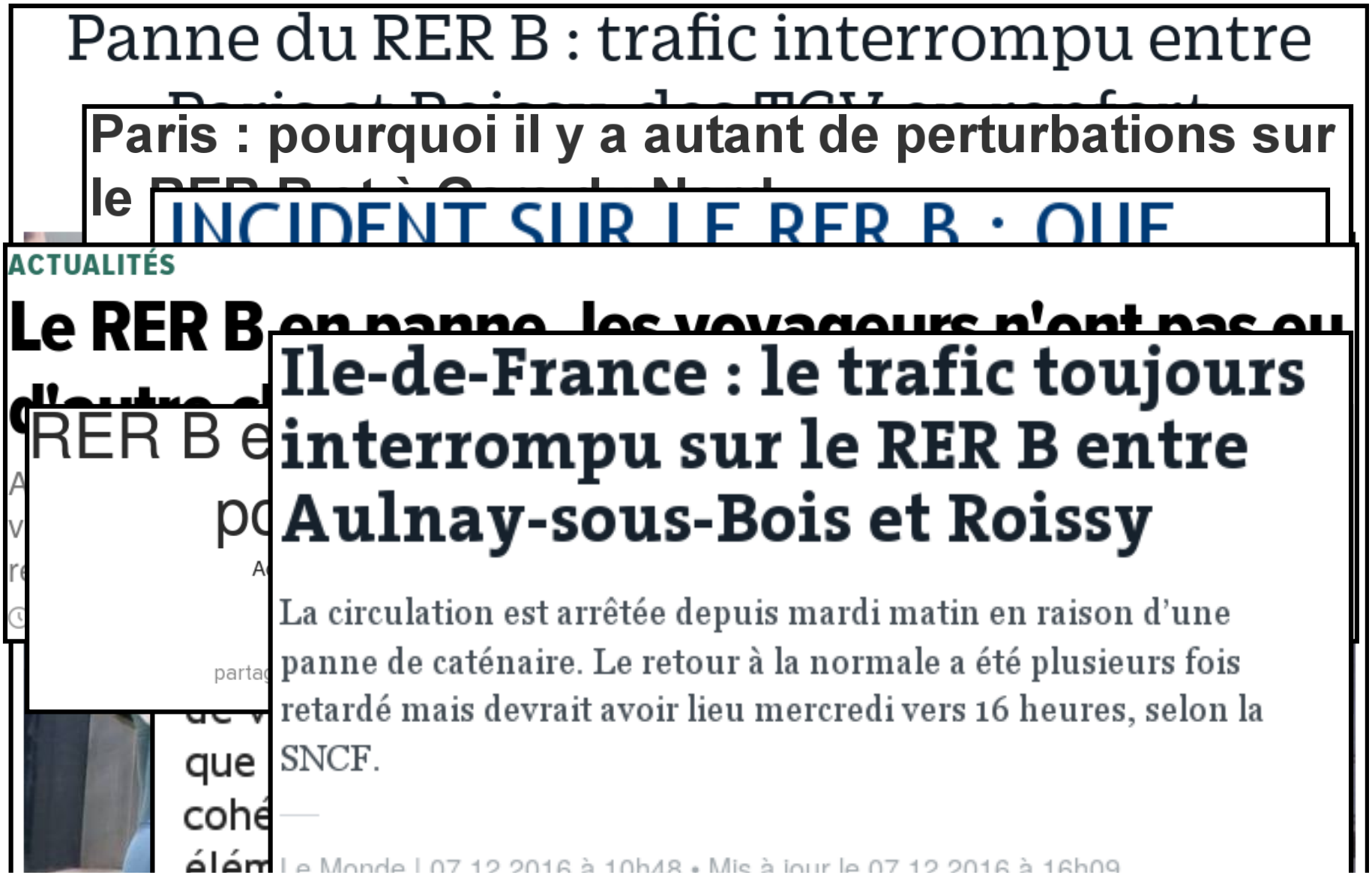
06/12/2016 11:57 CET | **Actualisé** 06/12/2016 20:14 CET



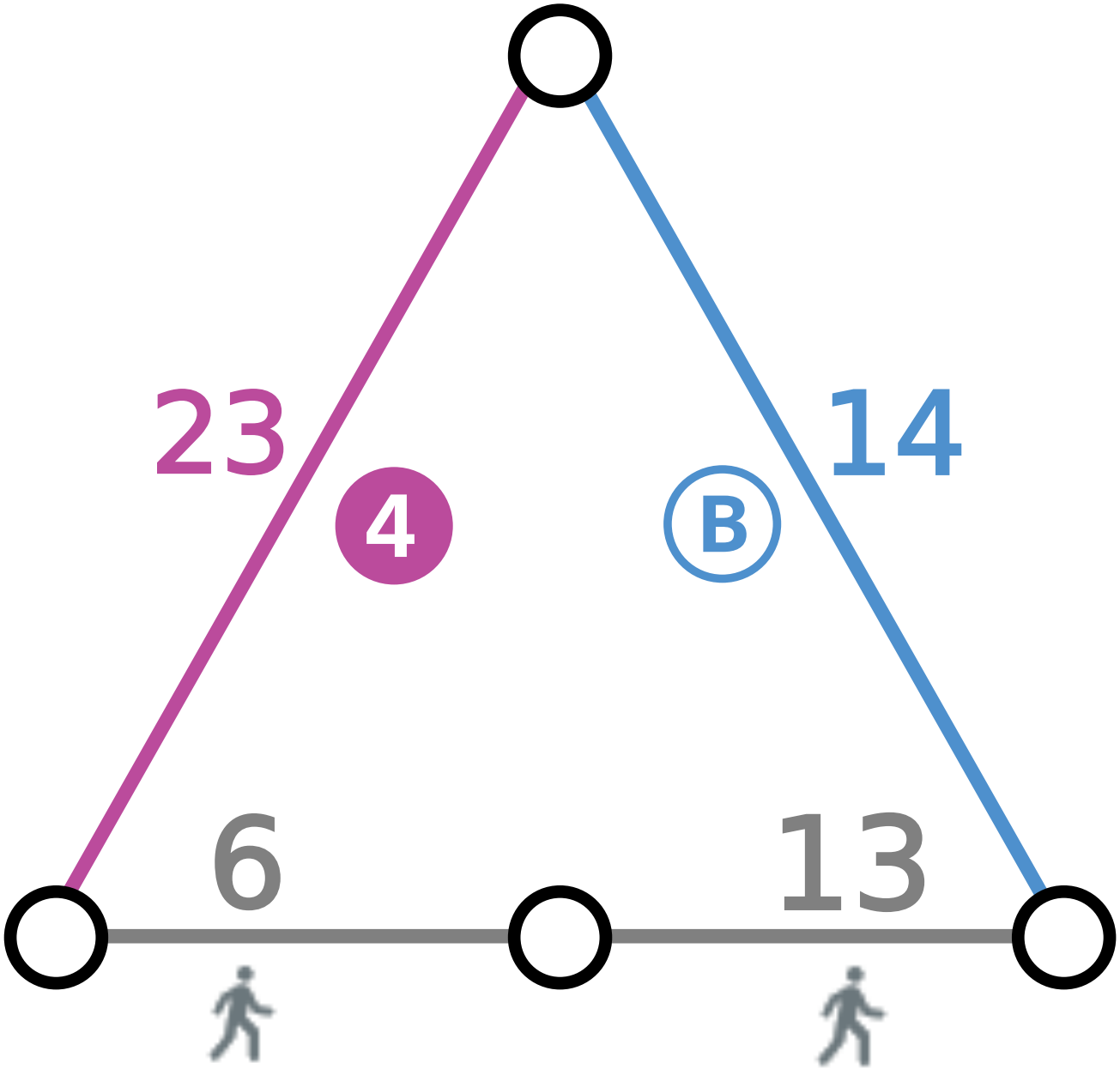
Pour beaucoup, le voyage a été difficile ce matin. Au fil de vos réactions sur Twitter notamment, je constate que les raisons de ces perturbations ne paraissent pas cohérentes. Je tiens donc à vous apporter des premiers éléments d'explication, que nous pourrons développer

Exemple : calcul de trajets en métro

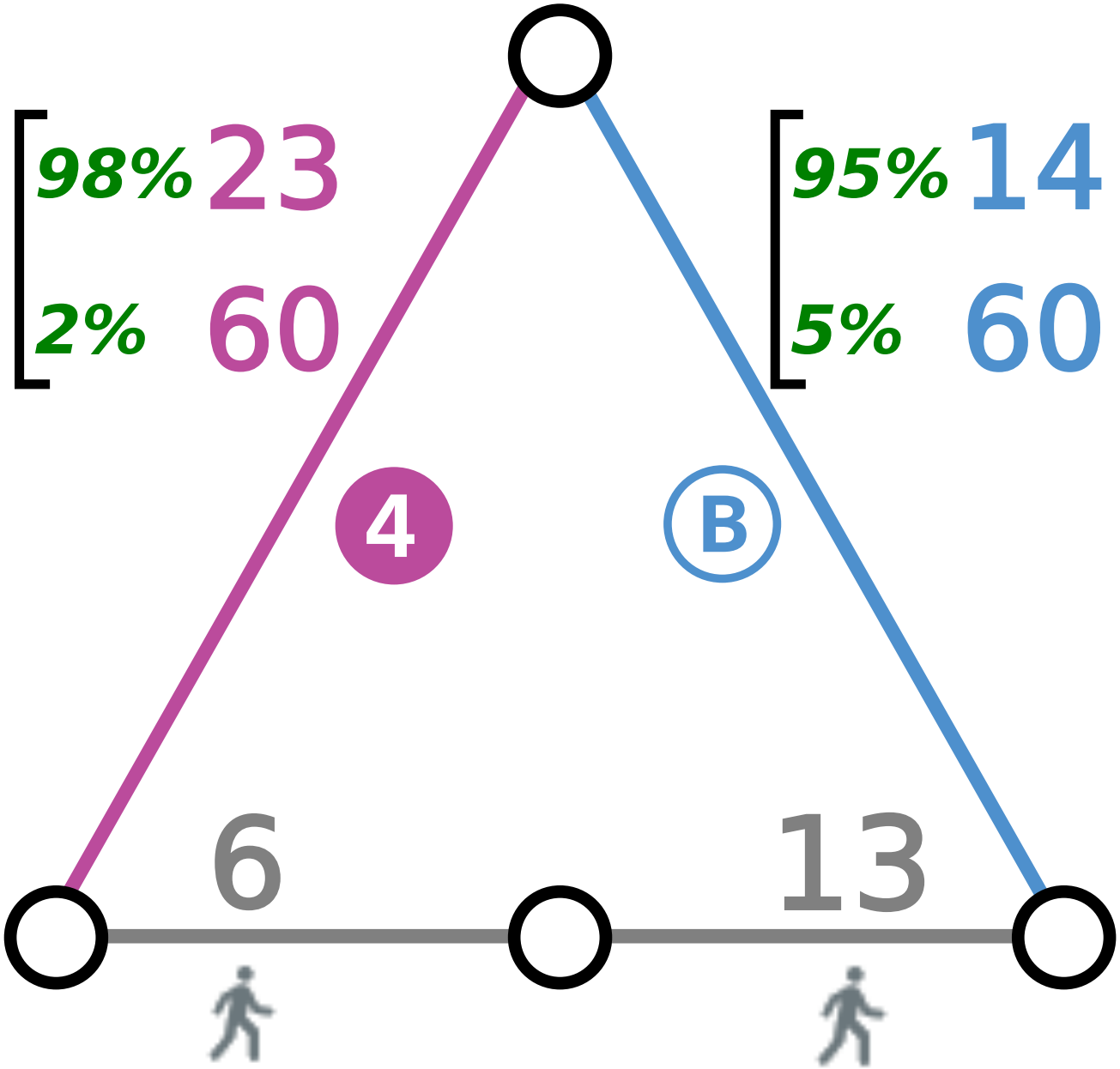
Exemple : calcul de trajets en métro



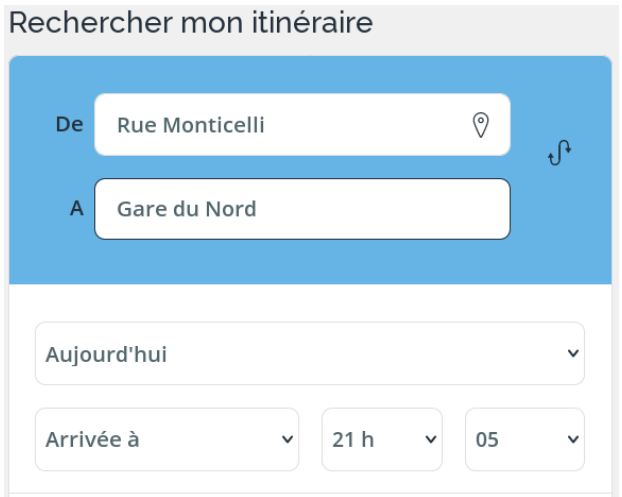
Exemple : calcul de trajets en métro

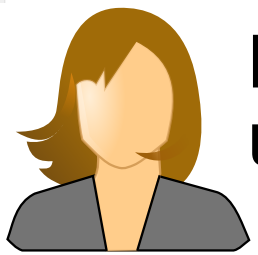


Exemple : calcul de trajets en métro

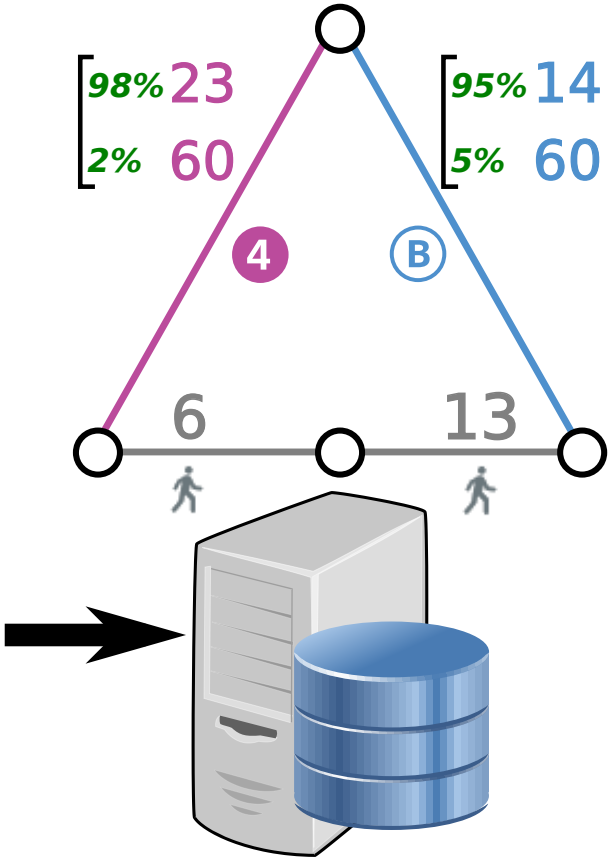


Exemple : calcul de trajets en métro



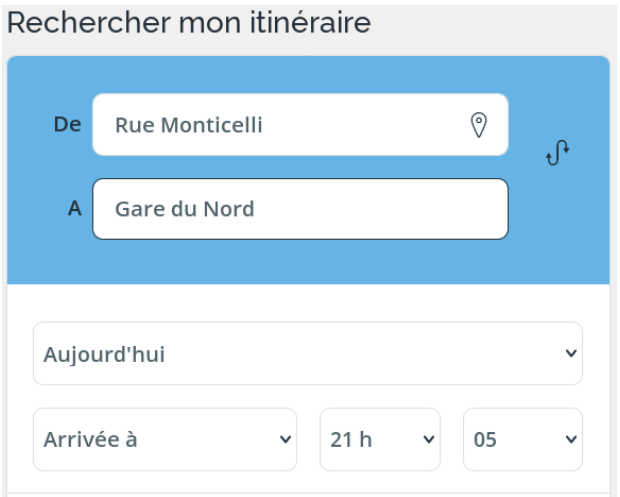


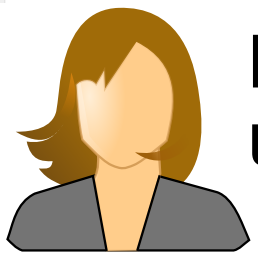
Requête utilisateur



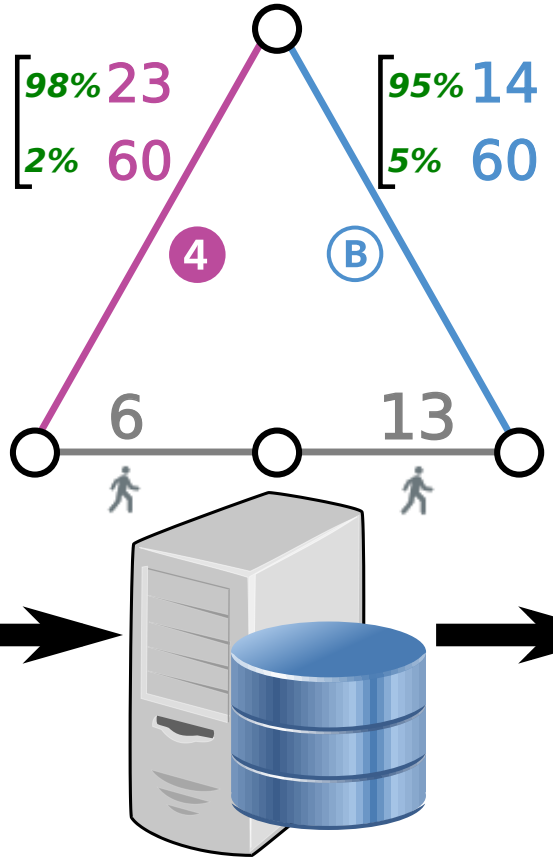
Base de données probabiliste

Exemple : calcul de trajets en métro





Requête utilisateur

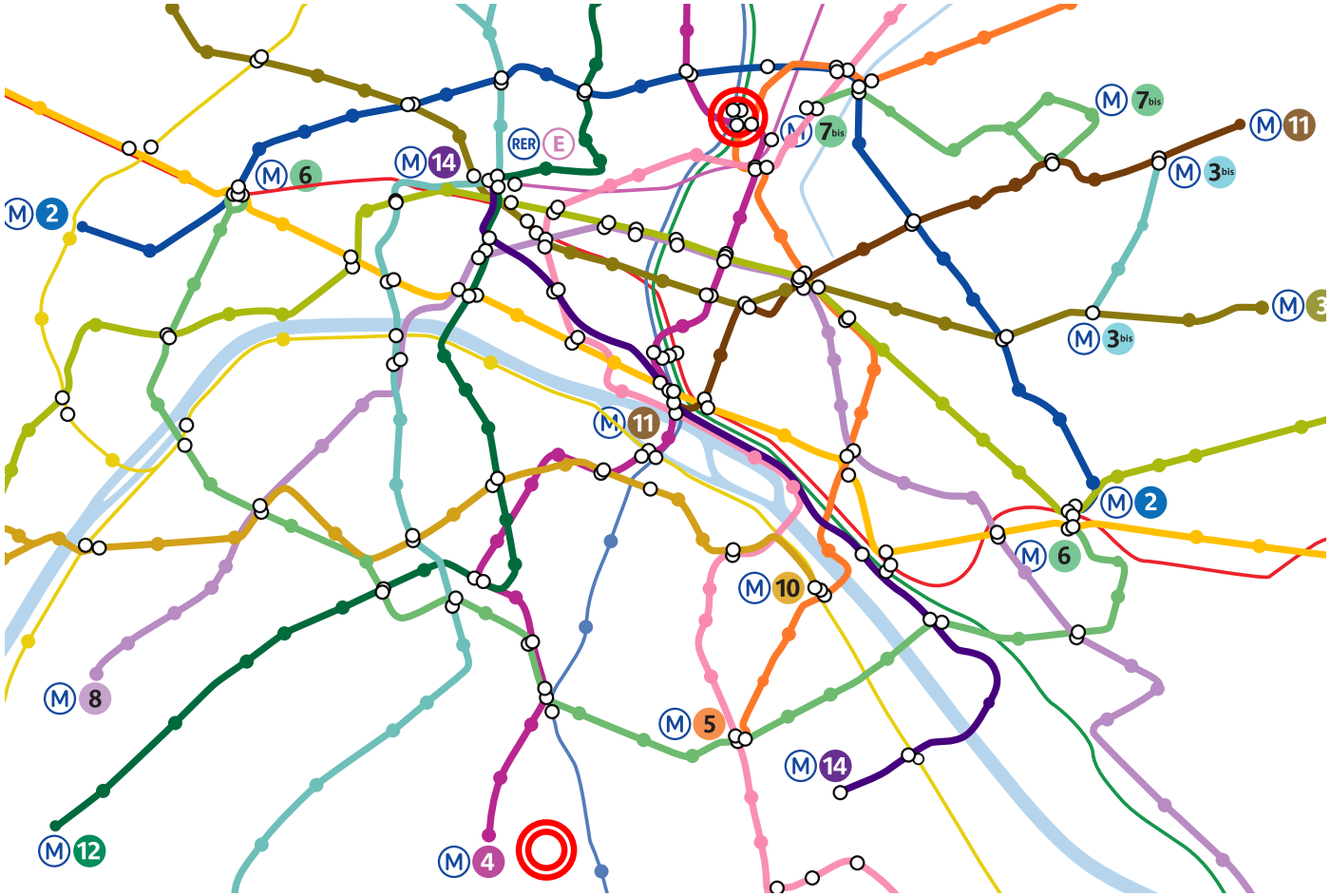


Base de données probabiliste

98% de chances d'être à l'heure

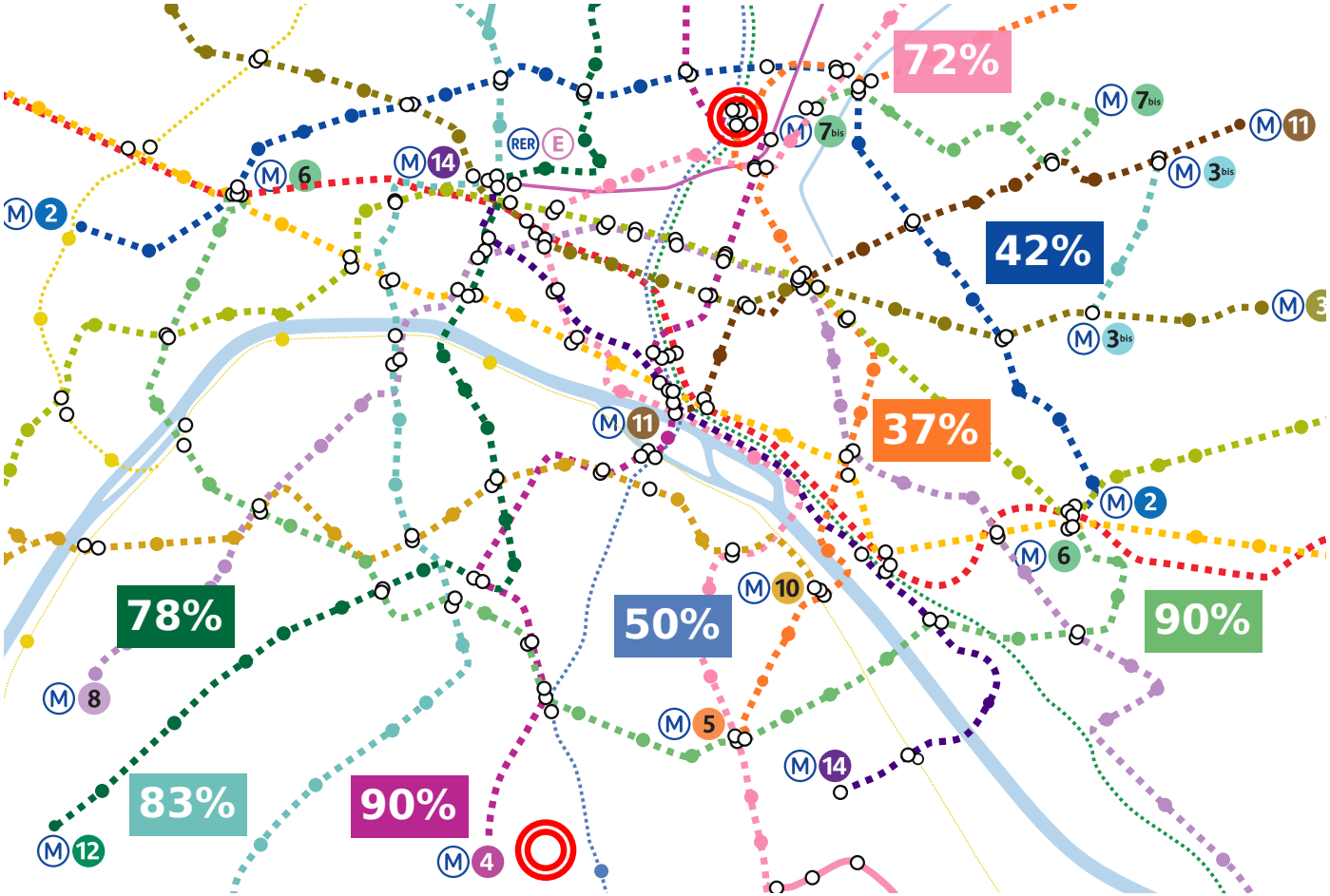
Résultat probabiliste

Problème : difficultés de calcul



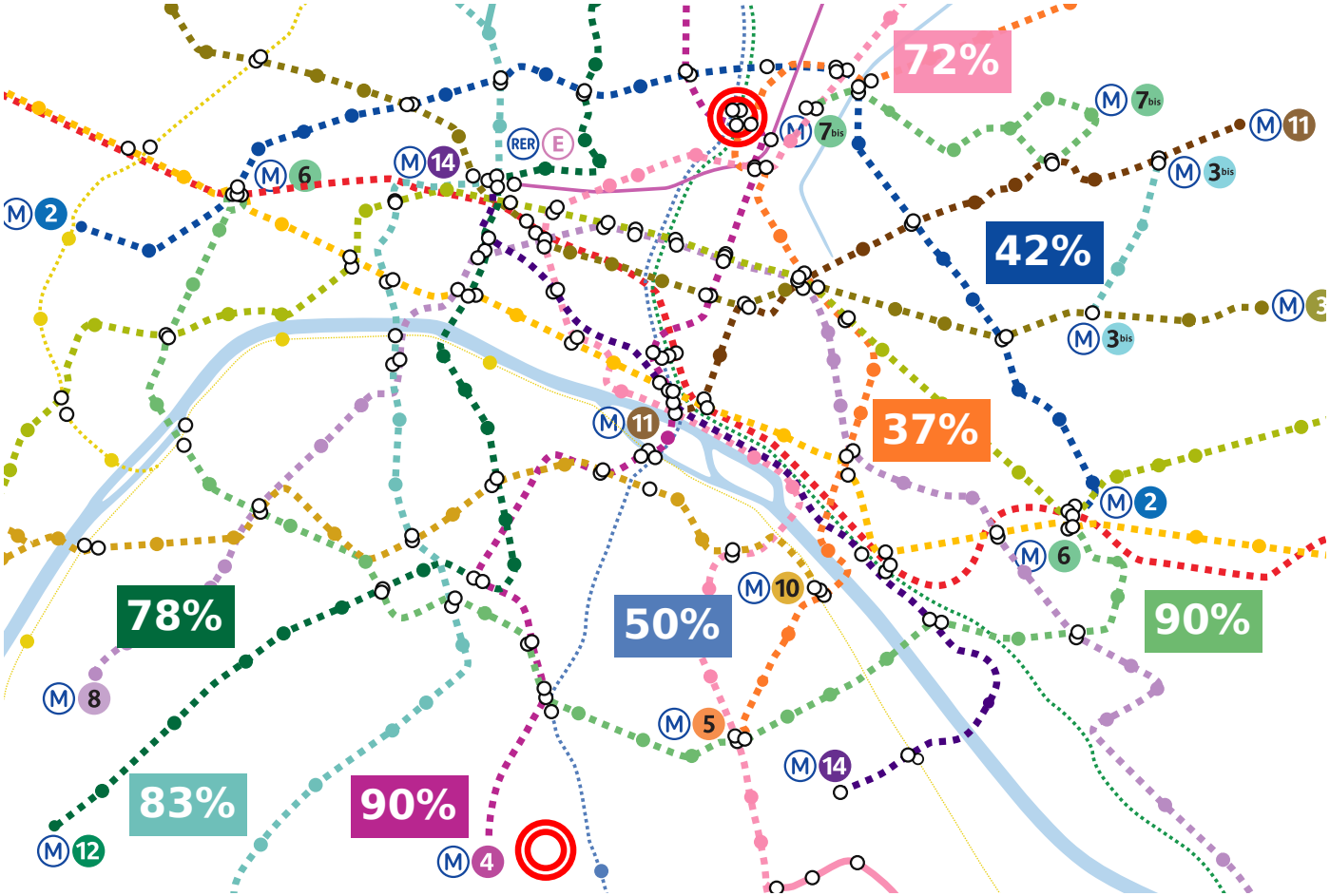
- Plus court chemin sur un grand graphe :
  → Algorithmes existants, problème bien étudié

Problème : difficultés de calcul



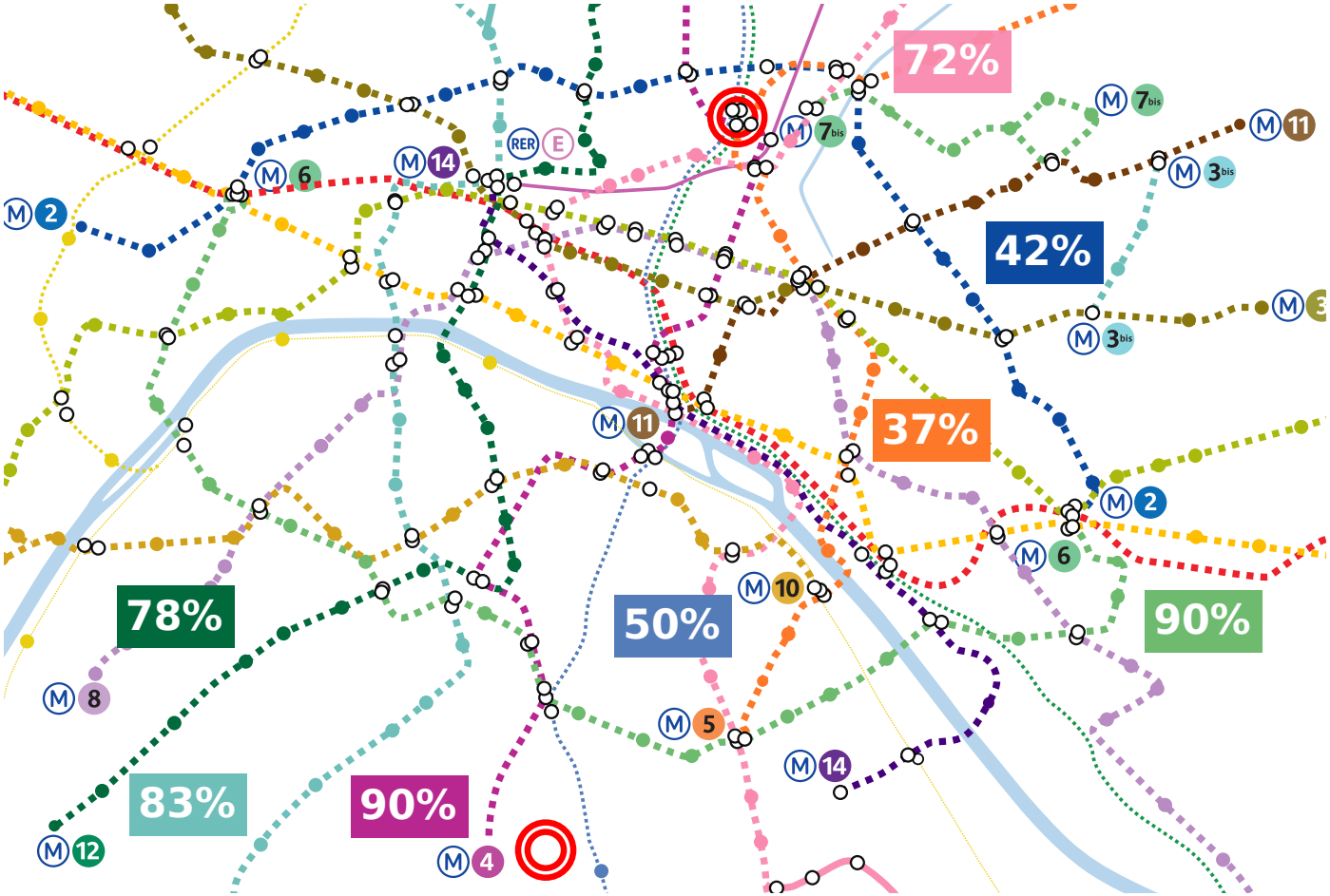
- Plus court chemin sur un grand graphe probabiliste :
  $\rightarrow$???

Problème : difficultés de calcul



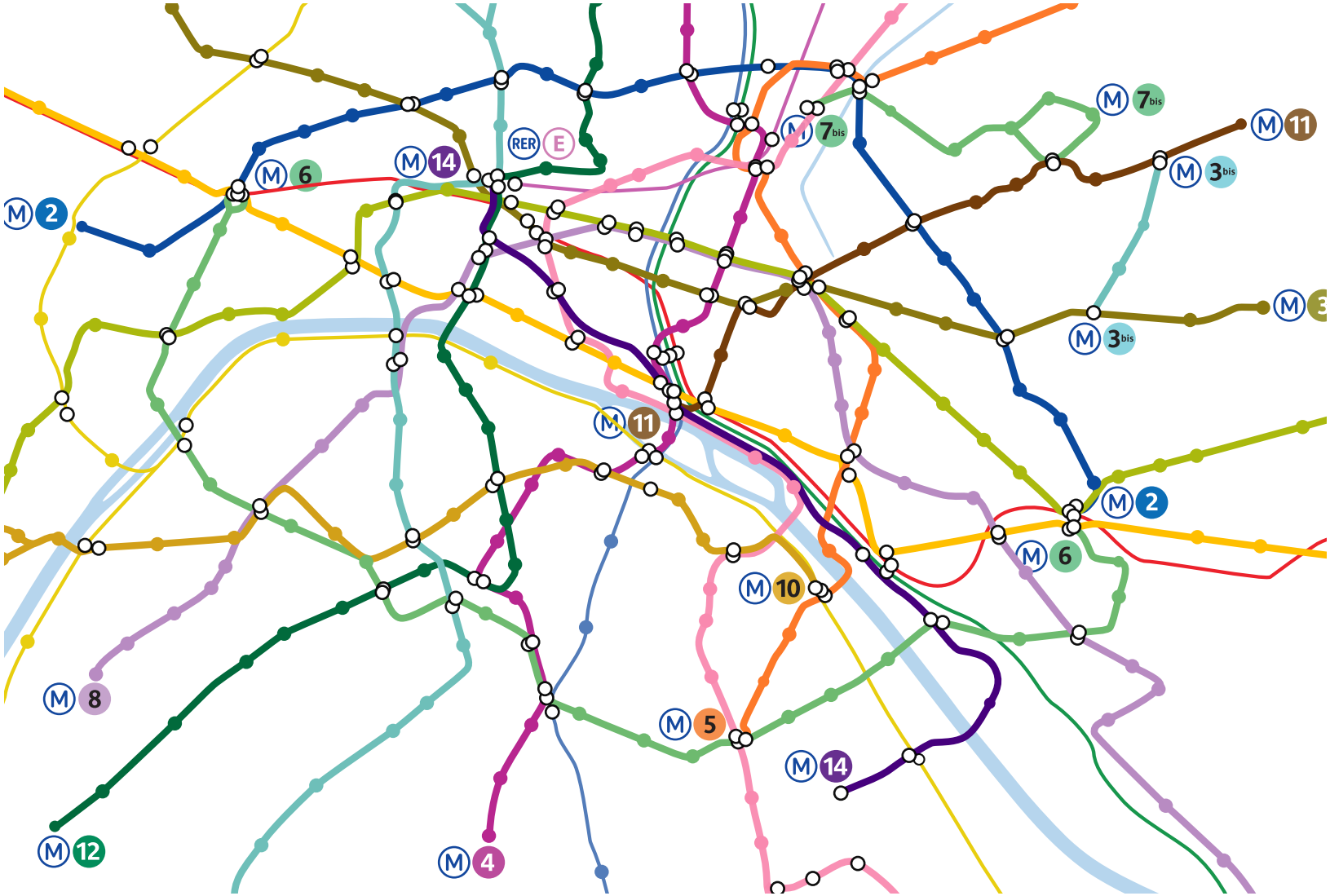
- Plus court chemin sur un grand graphe probabiliste :
  → Énumérer toutes les possibilités : exponentiel

Problème : difficultés de calcul

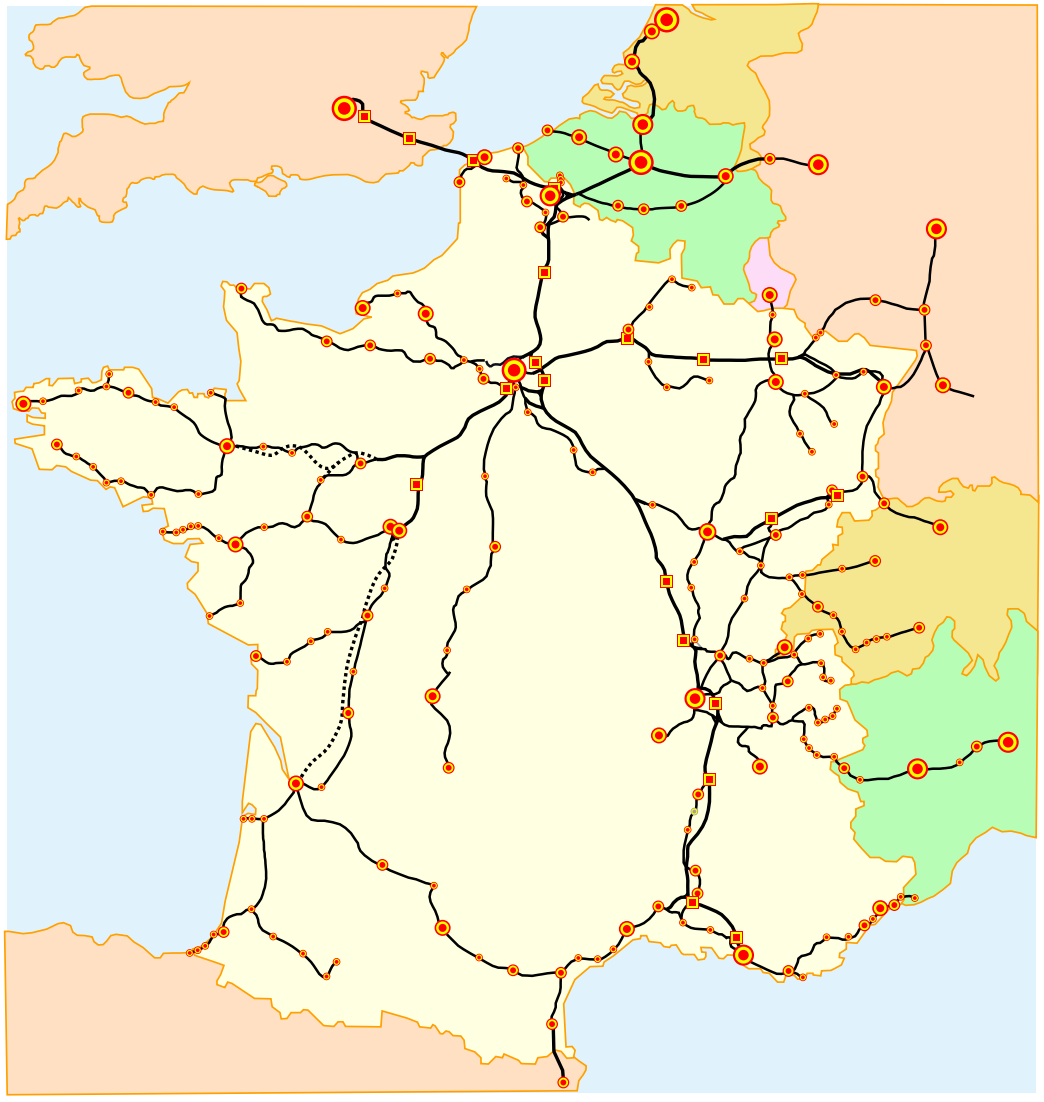


- Plus court chemin sur un grand graphe probabiliste :
  - → Énumérer toutes les possibilités : exponentiel
  - → Théorie : on ne peut pas espérer faire mieux en général !

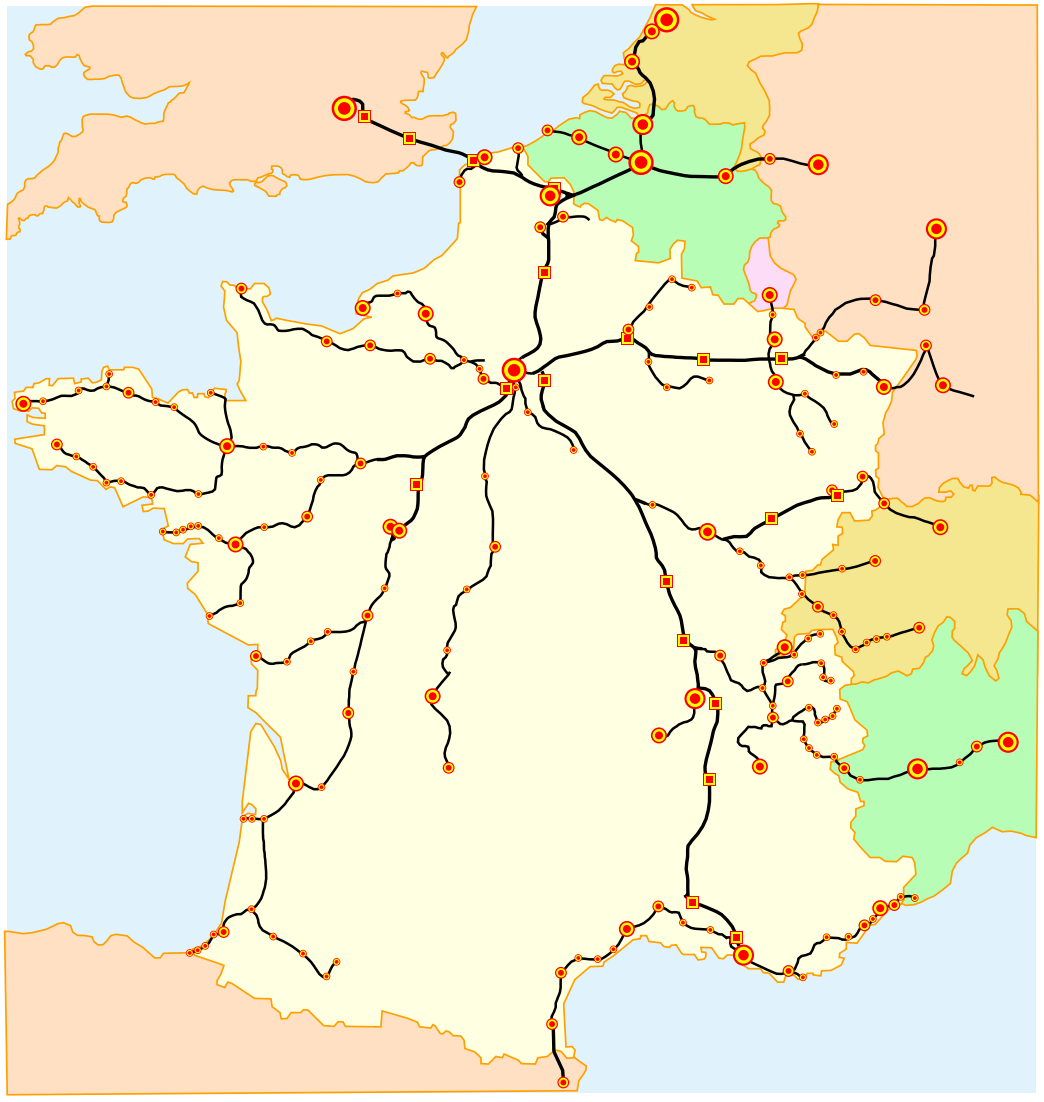
Idée : utiliser la structure des données



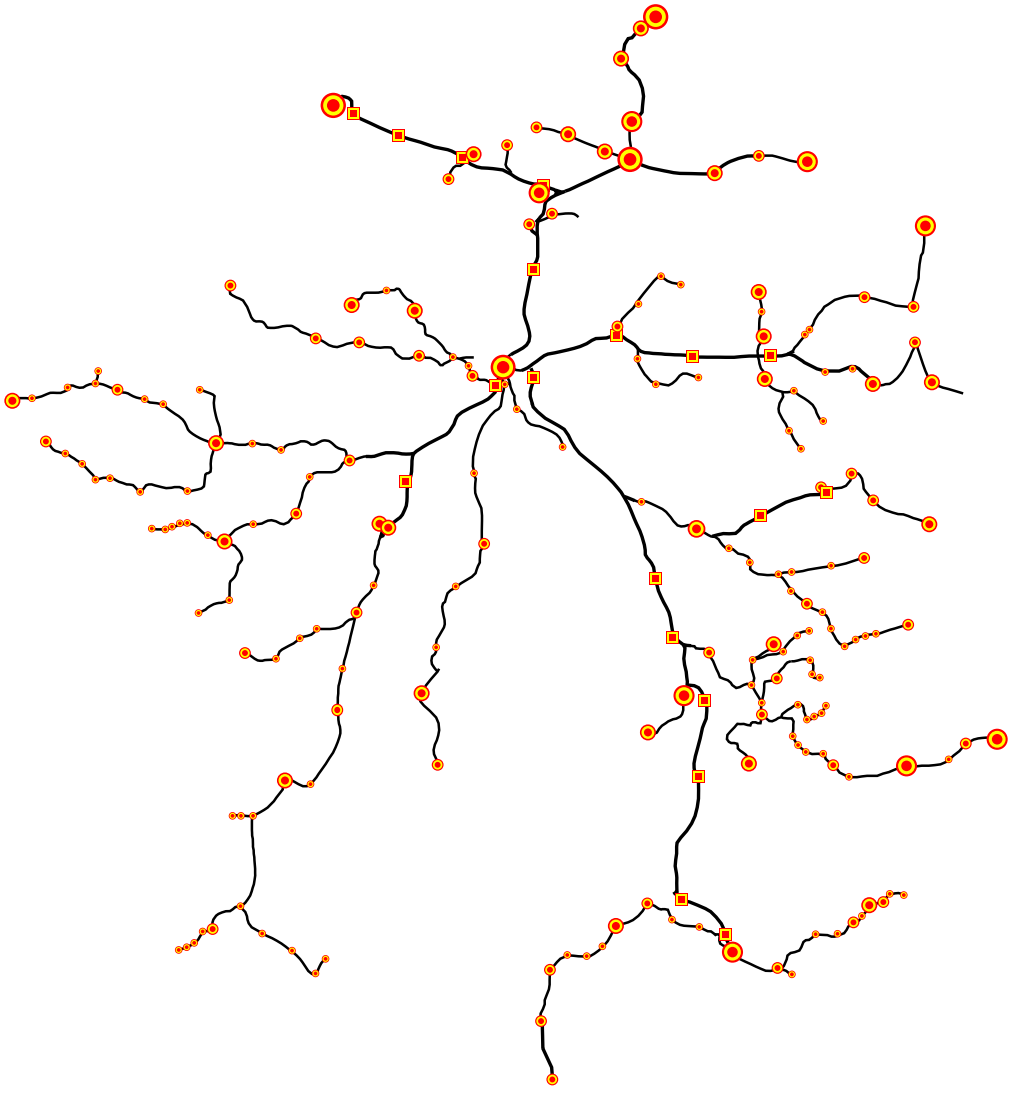
Idée : utiliser la structure des données



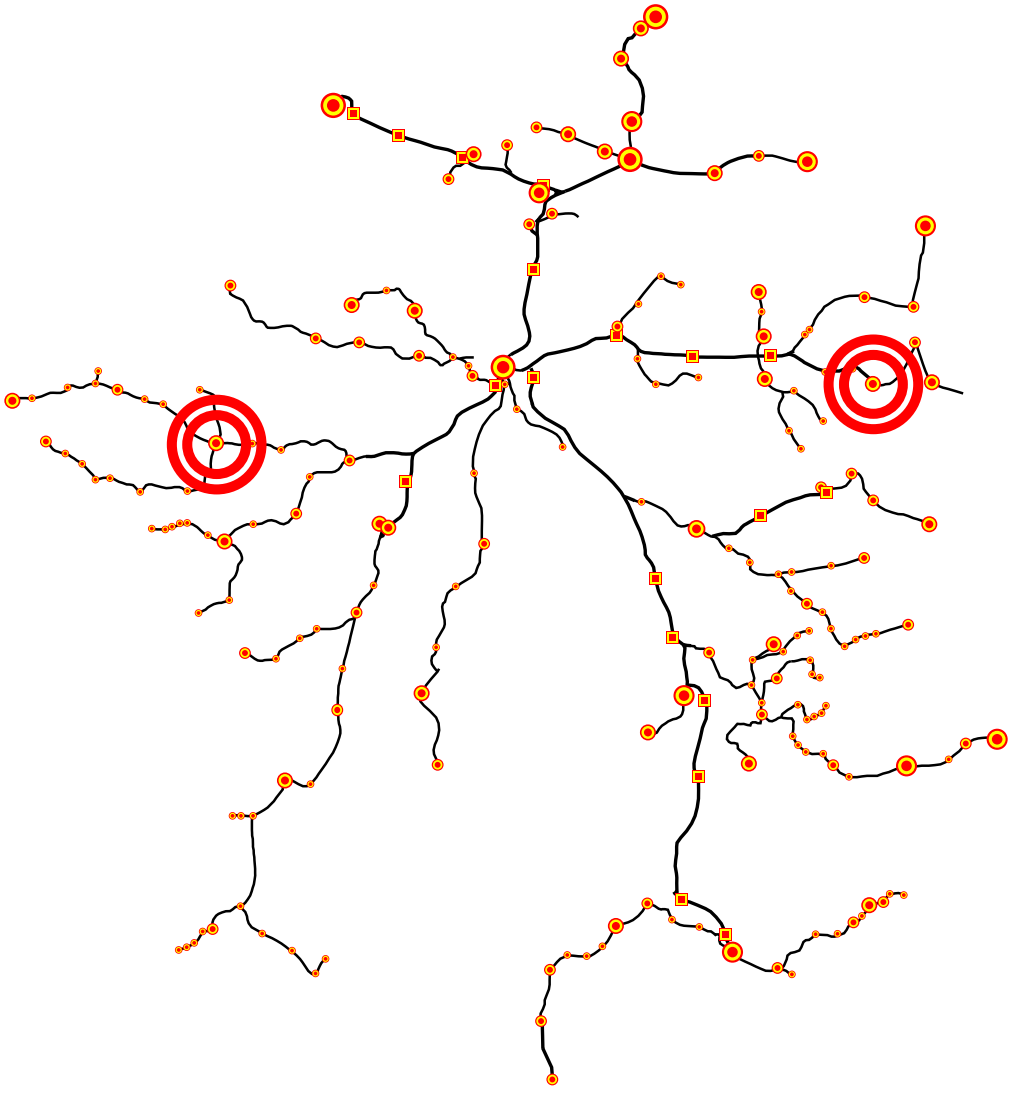
Idée : utiliser la structure des données



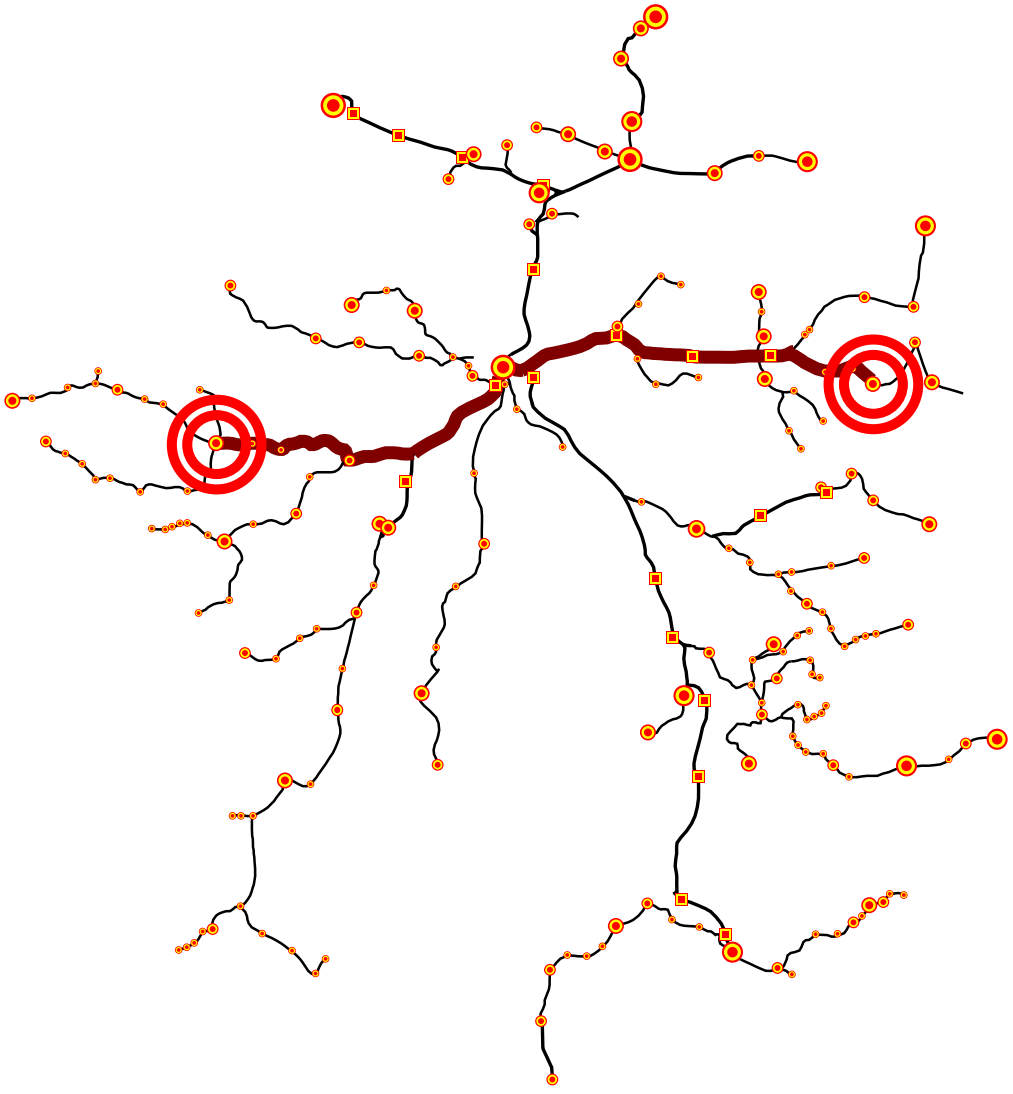
Idée : utiliser la structure des données



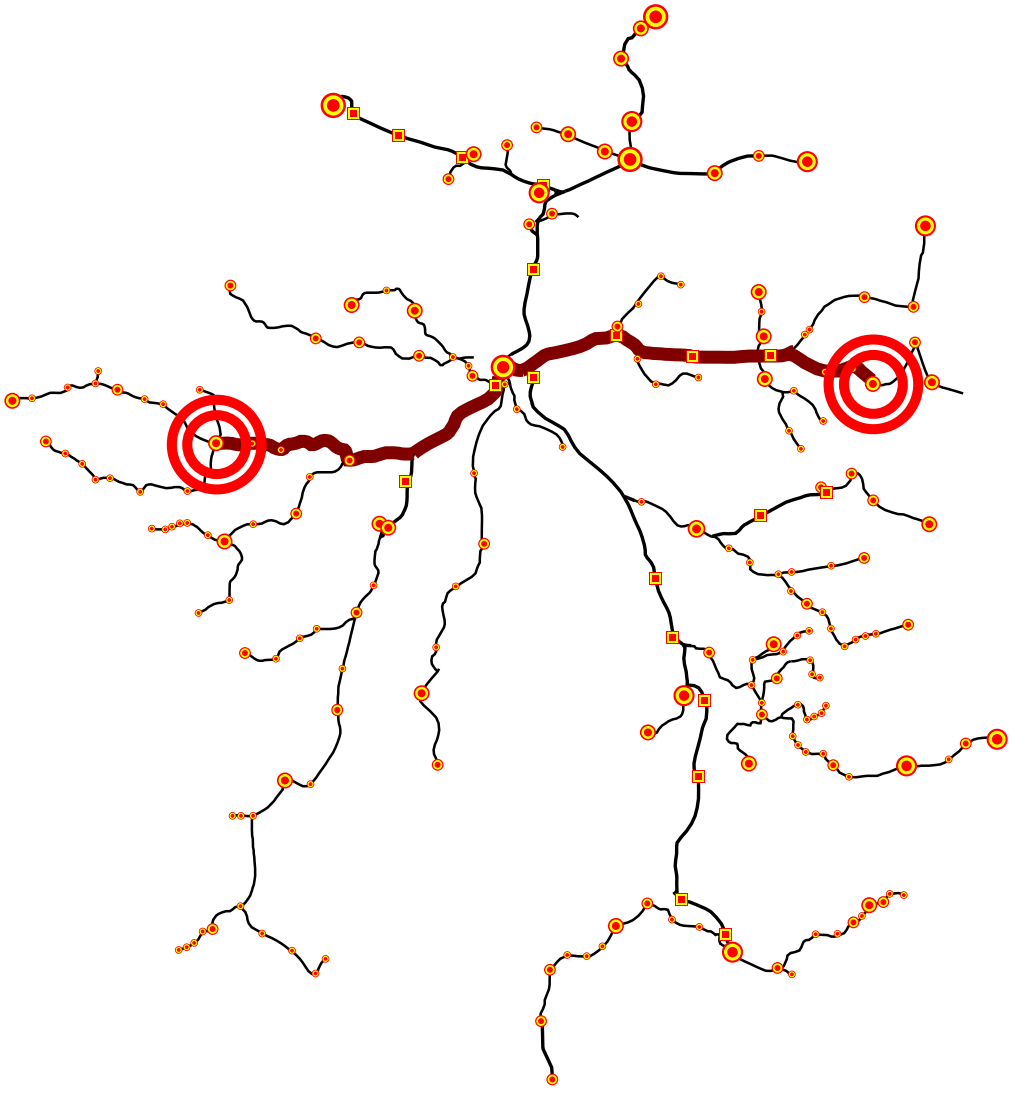
Idée : utiliser la structure des données



Idée : utiliser la structure des données

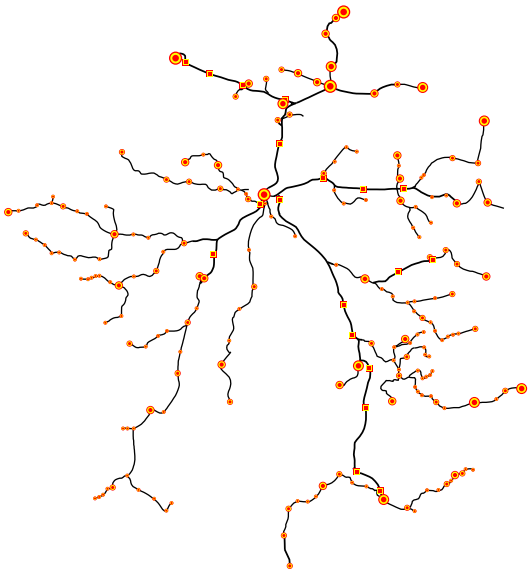


Idée : utiliser la structure des données



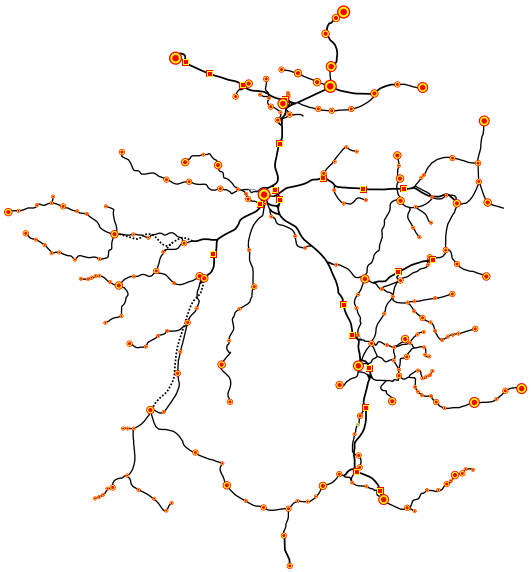
→ Plus court chemin : très facile sur un grand arbre

Résultats de ma thèse



→ Théorème : évaluer des requêtes expressives est facile sur de grands arbres probabilistes

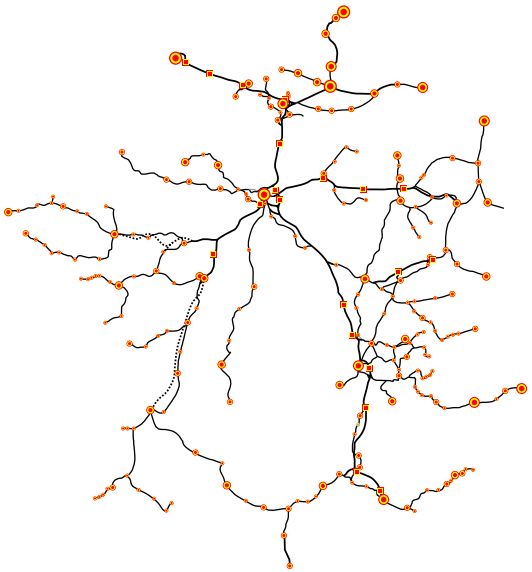
Résultats de ma thèse



→ Théorème : évaluer des requêtes expressives est facile sur de grands arbres probabilistes

→ S'étend aux quasi-arbres (en un sens formel)

Résultats de ma thèse

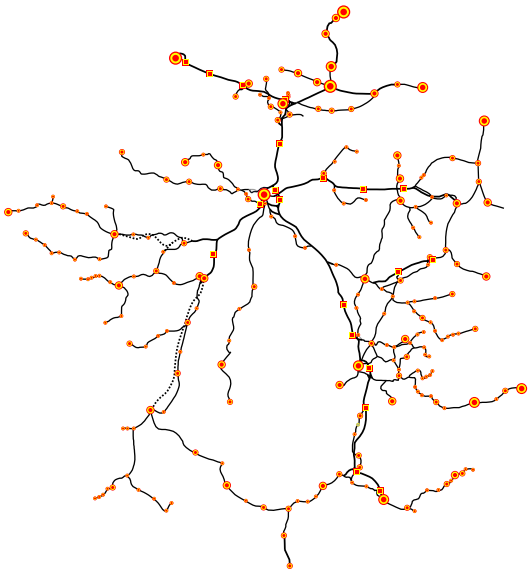


→ **Théorème :** évaluer des **requêtes expressives** est facile sur de grands arbres **probabilistes**

→ S'étend aux **quasi-arbres** (en un sens formel)

→ **Borne inférieure :** on ne peut pas faire mieux (#P-difficile sans borne sur la largeur d'arbre)

Résultats de ma thèse



→ **Théorème :** évaluer des **requêtes expressives** est facile sur de grands arbres **probabilistes**

→ S'étend aux **quasi-arbres** (en un sens formel)

→ **Borne inférieure :** on ne peut pas faire mieux (#P-difficile sans borne sur la largeur d'arbre)

Caractérise les conditions sur la structure des données probabilistes qui permettent de les interroger **efficacement**
(avec P. Bourhis, P. Senellart — ICALP'15, PODS'16)

Perspectives

→ Autres applications :

- Explication de résultats de requêtes par la provenance
- Énumération efficace de résultats de requêtes (ICALP'17)
- Étude plus fine de la complexité (ICDT'17, PODS'17)

Perspectives

→ Autres applications :

- Explication de résultats de requêtes par la provenance
- Énumération efficace de résultats de requêtes (ICALP'17)
- Étude plus fine de la complexité (ICDT'17, PODS'17)

→ Autres aspects de l'incertitude étudiés :

- Sur le raisonnement sur données incomplètes
  (avec M. Benedikt, Oxford — IJCAI'15, LICS'15, IJCAI'16)
- Sur l'incertitude numérique sur les données de la foule
  (avec Y. Amsterdamer, T. Milo, Tel Aviv — ICDT'14, ICDT'17)

Perspectives

→ Autres applications :

- Explication de résultats de requêtes par la provenance
- Énumération efficace de résultats de requêtes (ICALP'17)
- Étude plus fine de la complexité (ICDT'17, PODS'17)

→ Autres aspects de l'incertitude étudiés :

- Sur le raisonnement sur données incomplètes
  (avec M. Benedikt, Oxford — IJCAI'15, LICS'15, IJCAI'16)
- Sur l'incertitude numérique sur les données de la foule
  (avec Y. Amsterdamer, T. Milo, Tel Aviv — ICDT'14, ICDT'17)

Merci pour votre attention !

Références I


Amarilli, A., Amsterdamer, Y., and Milo, T. (2014).
On the Complexity of Mining Itemsets from the Crowd Using Taxonomies.
In *ICDT*.


Amarilli, A., Amsterdamer, Y., Milo, T., and Senellart, P. (2017a).
Top-k Queries on Unknown Values under Order Constraints.
In *ICDT*.


Amarilli, A. and Benedikt, M. (2015a).
Combining Existential Rules and Description Logics.
In *IJCAI*.


Amarilli, A. and Benedikt, M. (2015b).
Finite Open-World Query Answering with Number Restrictions.
In *LICS*.

Références II

- Amarilli, A., Benedikt, M., Bourhis, P., and Boom, M. V. (2016a). **Query Answering with Transitive and Linear-Ordered Data.** In *IJCAI*.
- Amarilli, A., Bourhis, P., Jachiet, L., and Mengel, S. (2017b). **A Circuit-Based Approach to Efficient Enumeration.** In *ICALP*.
- Amarilli, A., Bourhis, P., Monet, M., and Senellart, P. (2017c). **Combined Tractability of Query Evaluation via Tree Automata and Cycluits.** In *ICDT*.
- Amarilli, A., Bourhis, P., and Senellart, P. (2015). **Provenance Circuits for Trees and Treelike Instances.** In *ICALP*.

Références III


Amarilli, A., Bourhis, P., and Senellart, P. (2016b).
Tractable Lineages on Treelike Instances: Limits and Extensions.
In *PODS*.


Amarilli, A., Monet, M., and Senellart, P. (2017d).
Conjunctive Queries on Probabilistic Graphs: Combined Complexity.
In *PODS*.

Crédits photographiques

- Transparent 3 :
  - plan du métro : `https://commons.wikimedia.org/wiki/File:Paris_Metro_map.svg` (édité), utilisateur Umx sur Wikimedia Commons, domaine public
  - captures d'écran de `http://lab.vianavigo.com`, Stif, reproduites en vertu du droit de citation
  - icônes tirées de `https://openclipart.org/download/163711/database-server.svg` et `http://4vector.com/free-vector/female-user-icon-clip-art-117435`
  - articles de journaux reproduits en vertu du droit de citation :
    - `http://www.leparisien.fr/transports/circulation-alternee-a-paris-et-en-banlieue-une-panne-de-rer-et-des-bouchons-06-12-2016-6419610.php`
    - `http://www.rtl.fr/actu/societe-faits-divers/paris-le-trafic-totalement-interrompu-gare-du-nord-7786171150`
    - `https://www.rerb-leblog.fr/incident-rer-b-sest-passe-matin/`
    - `http://www.huffingtonpost.fr/2016/12/06/le-rer-b-en-panne-les-voyageurs-nont-pas-eu-dautres-choix-que/`
    - `http://www.lexpress.fr/actualite/societe/trafic/rer-b-en-panne-retards-du-d-circulation-alternee-deuxieme-journee-de-galere_1857905.html`
    - `http://www.lemonde.fr/entreprises/article/2016/12/07/ile-de-france-le-trafic-toujours-interrompu-sur-le-rer-b-en-direction-de-roissy_5044717_1656994.html`
- Transparent 5 : `https://commons.wikimedia.org/wiki/File:Carte_TGV.svg?uselang=fr` (édité), utilisateurs Jack ma, Muselaar, Benjism89, Pic-Sou, Uwe Dedering, Madcap sur Wikimedia Commons, sous licence CC-BY-SA 3.0