



D'où viennent les données ?

La science de la provenance

Pierre Senellart

La Nuit des données, ENS, 25 septembre 2026

Valda (Département Informatique de l'École normale supérieure, CNRS, ENS-PSL, Inria)

Un chiffre dans le journal

Le 31 janvier 2023 à Paris, une manifestation contre la réforme des retraites. Trois chiffres, publiés le soir même :

55 000	cabinet Occurrence, pour un collectif de médias
87 000	préfecture de police
500 000	CGT, organisatrice de la manifestation

Un chiffre dans le journal

Le 31 janvier 2023 à Paris, une manifestation contre la réforme des retraites. Trois chiffres, publiés le soir même :

55 000	cabinet Occurrence, pour un collectif de médias
87 000	préfecture de police
500 000	CGT, organisatrice de la manifestation

Lequel est le bon ?

Un chiffre dans le journal

Le 31 janvier 2023 à Paris, une manifestation contre la réforme des retraites. Trois chiffres, publiés le soir même :

- 55 000 cabinet Occurrence, pour un collectif de médias
- 87 000 préfecture de police
- 500 000 CGT, organisatrice de la manifestation

Lequel est le bon ? Aucun des trois n'est un mensonge : ce sont trois **méthodes**. Des policiers postés en hauteur comptent les rangs qui passent, puis recomptent sur vidéo ; des militants sur le trottoir comptent les rangs et les multiplient par le nombre de personnes qu'ils estiment par rang ; un capteur vidéo, en un point unique du parcours, compte les passages.

Un chiffre dans le journal

Le 31 janvier 2023 à Paris, une manifestation contre la réforme des retraites. Trois chiffres, publiés le soir même :

- 55 000 cabinet Occurrence, pour un collectif de médias
- 87 000 préfecture de police
- 500 000 CGT, organisatrice de la manifestation

Lequel est le bon ? Aucun des trois n'est un mensonge : ce sont trois **méthodes**. Des policiers postés en hauteur comptent les rangs qui passent, puis recomptent sur vidéo ; des militants sur le trottoir comptent les rangs et les multiplient par le nombre de personnes qu'ils estiment par rang ; un capteur vidéo, en un point unique du parcours, compte les passages.

Le chiffre seul ne veut rien dire ; le chiffre avec sa méthode veut dire quelque chose.

Un chiffre dans le journal

Le 31 janvier 2023 à Paris, une manifestation contre la réforme des retraites. Trois chiffres, publiés le soir même :

55 000	cabinet Occurrence, pour un collectif de médias
87 000	préfecture de police
500 000	CGT, organisatrice de la manifestation

Lequel est le bon ? Aucun des trois n'est un mensonge : ce sont trois **méthodes**. Des policiers postés en hauteur comptent les rangs qui passent, puis recomptent sur vidéo ; des militants sur le trottoir comptent les rangs et les multiplient par le nombre de personnes qu'ils estiment par rang ; un capteur vidéo, en un point unique du parcours, compte les passages.

Le chiffre seul ne veut rien dire ; le chiffre avec sa méthode veut dire quelque chose.

Et de tous les chiffres qu'on lit dans le journal, celui-ci est à peu près le seul dont on nous explique comment il a été obtenu.

Le modèle :
la provenance d'une œuvre d'art

Quand un tableau de maître réapparaît



Jules Breton, *Femme à l'ombrelle, baie de Douarnenez*, vers 1870. Petit Palais, Paris Musées, cco.

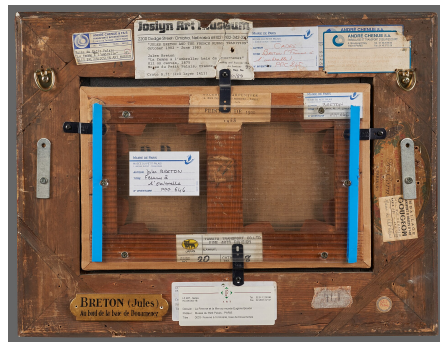
La première question des experts n'est pas « est-il beau ? »

Quand un tableau de maître réapparaît

La première question des experts n'est pas « est-il beau ? », mais :

quelle est sa provenance ?

La **chaîne documentée** de l'atelier du peintre à aujourd'hui : qui l'a possédé, par quelles mains il est passé, quand, et contre quoi.



Dos du même tableau.
Petit Palais, Paris Musées, cco.

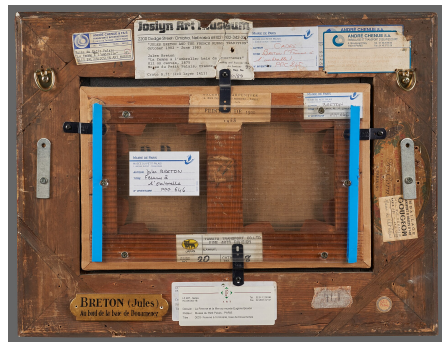
Quand un tableau de maître réapparaît

La première question des experts n'est pas « est-il beau ? », mais :

quelle est sa provenance ?

La **chaîne documentée** de l'atelier du peintre à aujourd'hui : qui l'a possédé, par quelles mains il est passé, quand, et contre quoi.

- Factures, inventaires de succession, registres de ventes
- Et le **dos** du tableau : étiquettes de galeries, cachets de douane, numéros d'inventaire



Dos du même tableau.
Petit Palais, Paris Musées, cco.

Quand un tableau de maître réapparaît

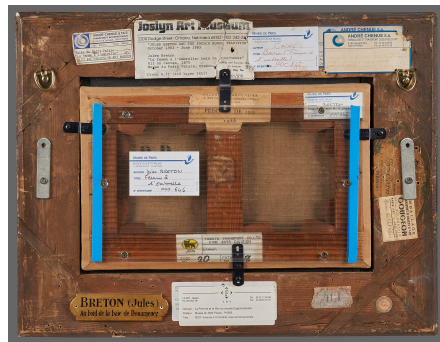
La première question des experts n'est pas « est-il beau ? », mais :

quelle est sa provenance ?

La **chaîne documentée** de l'atelier du peintre à aujourd'hui : qui l'a possédé, par quelles mains il est passé, quand, et contre quoi.

- Factures, inventaires de succession, registres de ventes
- Et le **dos** du tableau : étiquettes de galeries, cachets de douane, numéros d'inventaire

Ce qui en dépend : l'**authenticité**, la **légitimité** de la possession (spoliations de 1933–1945), la **valeur**.



Dos du même tableau.
Petit Palais, Paris Musées, cco.

Une chaîne, et ses trous : le *Salvator Mundi*



Après la restauration de 2011.
Wikimedia Commons, domaine public.

Récit : Wikipédia.

Panneau de noyer de 66 centimètres : le tableau le plus cher jamais vendu, et le plus disputé.

Une chaîne, et ses trous : le *Salvator Mundi*



Après la restauration de 2011.
Wikimedia Commons, domaine public.

Récit : Wikipédia.

Panneau de noyer de 66 centimètres : le tableau le plus cher jamais vendu, et le plus disputé.

Entre 45 livres et 450 millions de dollars, seul le dossier a changé.

Depuis 2017, on ignore où il est, et le consensus sur l'attribution s'est défait.



Et le dossier lui-même peut être fabriqué

Affaire Beltracchi (Allemagne, condamné en 2011). Il peignait, dans la manière d'artistes du début du XX^e siècle, des tableaux inédits que personne n'avait jamais vus. Son vrai talent était le dossier :



Wolfgang Beltracchi, 2024.

Photo Nojin, CC BY 4.0.

Récit : Wikipédia.

Et le dossier lui-même peut être fabriqué

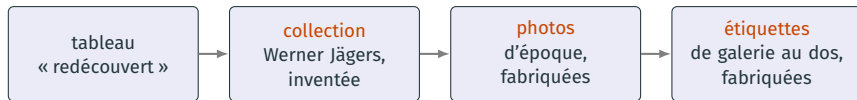


Wolfgang Beltracchi, 2024.

Photo Nojin, CC BY 4.0.

Récit : Wikipédia.

Affaire Beltracchi (Allemagne, condamné en 2011). Il peignait, dans la manière d'artistes du début du XX^e siècle, des tableaux **inédits** que personne n'avait jamais vus. Son vrai talent était le dossier :



Et le dossier lui-même peut être fabriqué

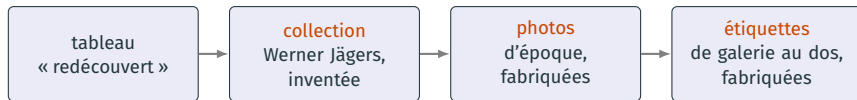


Wolfgang Beltracchi, 2024.

Photo Nojin, CC BY 4.0.

Récit : Wikipédia.

Affaire Beltracchi (Allemagne, condamné en 2011). Il peignait, dans la manière d'artistes du début du XX^e siècle, des tableaux **inédits** que personne n'avait jamais vus. Son vrai talent était le dossier :



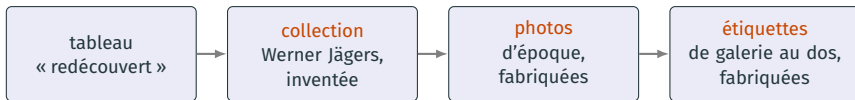
Sa femme, habillée à la mode des années 1930, posant devant les tableaux, sur du papier photographique d'époque : tout se raconte bien.

Et le dossier lui-même peut être fabriqué



Wolfgang Beltracchi, 2024.
Photo Nojin, CC BY 4.0.
Récit : Wikipédia.

Affaire Beltracchi (Allemagne, condamné en 2011). Il peignait, dans la manière d'artistes du début du XX^e siècle, des tableaux **inédits** que personne n'avait jamais vus. Son vrai talent était le dossier :



Sa femme, habillée à la mode des années 1930, posant devant les tableaux, sur du papier photographique d'époque : tout se raconte bien. Ce qui l'a trahi : un **pigment**. Du blanc de titane dans un tableau daté de 1914, alors qu'il n'était pas encore commercialisé.

Et le dossier lui-même peut être fabriqué

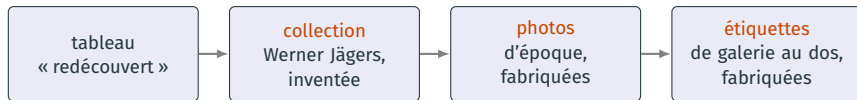


Wolfgang Beltracchi, 2024.

Photo Nojin, CC BY 4.0.

Récit : Wikipédia.

Affaire Beltracchi (Allemagne, condamné en 2011). Il peignait, dans la manière d'artistes du début du XX^e siècle, des tableaux **inédits** que personne n'avait jamais vus. Son vrai talent était le dossier :

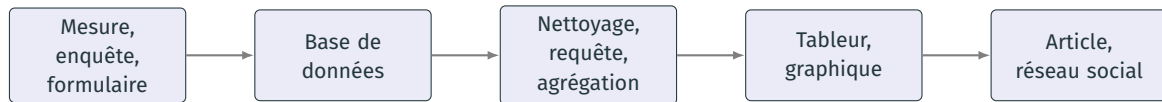


Sa femme, habillée à la mode des années 1930, posant devant les tableaux, sur du papier photographique d'époque : tout se raconte bien. Ce qui l'a trahi : un **pigment**. Du blanc de titane dans un tableau daté de 1914, alors qu'il n'était pas encore commercialisé.

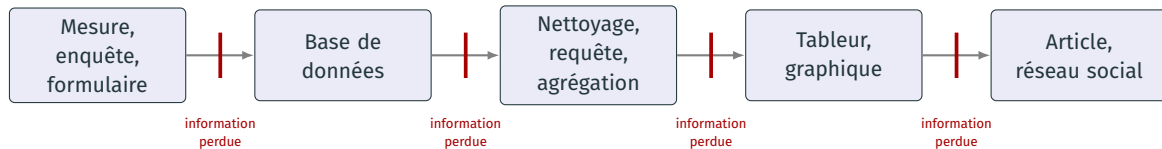
Une provenance vaut exactement ce qu'on peut en vérifier.

Et pour les données ?

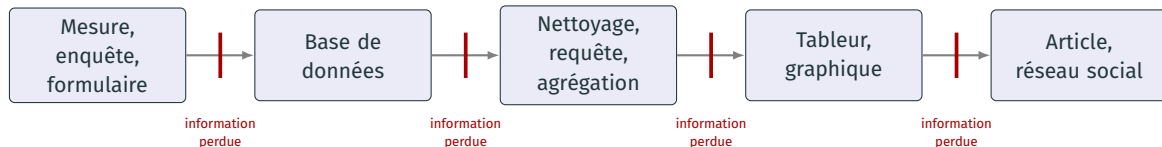
Un chiffre n'arrive jamais seul



Un chiffre n'arrive jamais seul

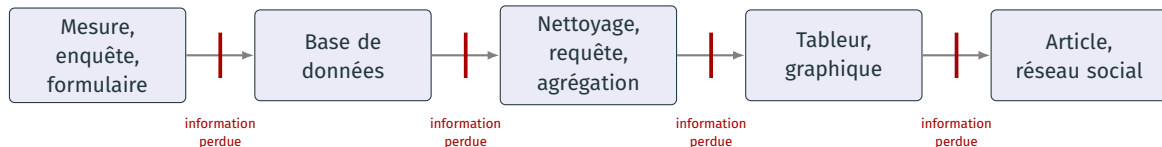


Un chiffre n'arrive jamais seul



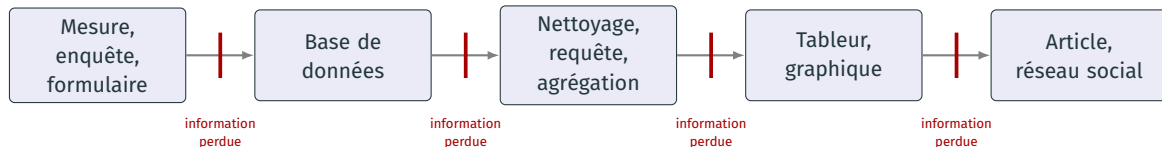
- À chaque flèche, un système passe à un autre un **résultat** – et abandonne tout ce qui a servi à le produire

Un chiffre n'arrive jamais seul

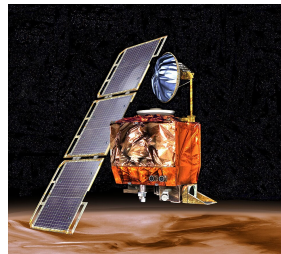


- À chaque flèche, un système passe à un autre un **résultat** – et abandonne tout ce qui a servi à le produire
- Au bout de la chaîne, le chiffre est peut-être exact, mais on ne peut plus le **vérifier** : ni le contester, ni le corriger, ni même l'interpréter

Un chiffre n'arrive jamais seul



- À chaque flèche, un système passe à un autre un **résultat** – et abandonne tout ce qui a servi à le produire
- Au bout de la chaîne, le chiffre est peut-être exact, mais on ne peut plus le **vérifier** : ni le contester, ni le corriger, ni même l'interpréter
- En 1999, la sonde Mars Climate Orbiter est détruite parce qu'un logiciel produit des livres-force et que le suivant lit des newtons. Le nombre circulait sans son unité, c'est-à-dire sans un fragment de sa provenance



Mars Climate Orbiter, vue d'artiste. NASA/JPL, domaine public.

Récit : rapport d'enquête de la NASA, 1999.

La cellule qui manquait

2010 – Deux économistes de Harvard, Carmen Reinhart et Kenneth Rogoff : au-delà de 90 % de dette publique, la croissance d'un pays devient **négative**. Le chiffre nourrit trois ans de débats sur l'austérité.

La cellule qui manquait

2010 – Deux économistes de Harvard, Carmen Reinhart et Kenneth Rogoff : au-delà de 90 % de dette publique, la croissance d'un pays devient **négative**. Le chiffre nourrit trois ans de débats sur l'austérité.

2013 – Un doctorant, Thomas Herndon, demande la feuille de calcul pour un devoir de cours. Il l'obtient.

La cellule qui manquait

2010 – Deux économistes de Harvard, Carmen Reinhart et Kenneth Rogoff : au-delà de 90 % de dette publique, la croissance d'un pays devient **négative**. Le chiffre nourrit trois ans de débats sur l'austérité.

2013 – Un doctorant, Thomas Herndon, demande la feuille de calcul pour un devoir de cours. Il l'obtient.

La formule de moyenne s'arrête à la ligne 44 : **cinq des vingt pays n'étaient pas dans la sélection**. Avec eux, la croissance au-delà de 90 % de dette n'est plus négative [Herndon et al., 2014].

La cellule qui manquait

2010 – Deux économistes de Harvard, Carmen Reinhart et Kenneth Rogoff : au-delà de 90 % de dette publique, la croissance d'un pays devient **négative**. Le chiffre nourrit trois ans de débats sur l'austérité.

2013 – Un doctorant, Thomas Herndon, demande la feuille de calcul pour un devoir de cours. Il l'obtient.

La formule de moyenne s'arrête à la ligne 44 : **cinq des vingt pays n'étaient pas dans la sélection**. Avec eux, la croissance au-delà de 90 % de dette n'est plus négative [Herndon et al., 2014].

La provenance de ce chiffre, c'était : quelles cellules ont été moyennées.

	A	L
30	...	...
31	...	...
32	...	...
	(15 pays, lignes 30 à 44)	
43	...	...
44	...	...
45	Australie	3,8
46	Autriche	n.d.
47	Belgique	2,6
48	Canada	3,0
49	Danemark	n.d.

cinq pays hors de la sélection

=MOYENNE(L30:L44)

Colonne L : croissance moyenne au-delà de 90 % de dette (Herndon et al., table 4).

Deux autres histoires vraies

Octobre 2020 – L'agence de santé publique anglaise perd 15 841 tests positifs au COVID. Les résultats des laboratoires étaient rassemblés dans le vieux format .xls, qui s'arrête à 65 536 lignes. Au-delà, rien n'est refusé : les lignes sont simplement absentes.

65534	positif
65535	positif
65536	positif
65537	positif
65538	positif jamais lues
...	...

Deux autres histoires vraies

Octobre 2020 – L'agence de santé publique anglaise perd **15 841 tests positifs** au COVID. Les résultats des laboratoires étaient rassemblés dans le vieux format `.xls`, qui s'arrête à **65 536 lignes**. Au-delà, rien n'est refusé : les lignes sont simplement absentes.

65534	positif
65535	positif
65536	positif
65537	positif
65538	positif jamais lues
...	...

La provenance manquante est celle d'une absence : pourquoi cette ligne *n'est-elle pas* dans le résultat ?

BBC News, 5 octobre 2020, « Excel: Why using Microsoft's tool caused Covid-19 results to be lost »

Deux autres histoires vraies

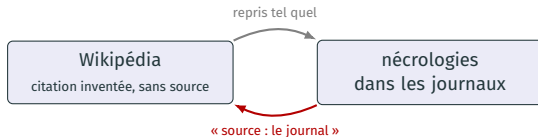
Octobre 2020 – L'agence de santé publique anglaise perd **15 841 tests positifs** au COVID. Les résultats des laboratoires étaient rassemblés dans le vieux format `.xls`, qui s'arrête à **65 536 lignes**. Au-delà, rien n'est refusé : les lignes sont simplement absentes.

65534	positif
65535	positif
65536	positif
65537	positif
65538	positif jamais lues
...	...

La provenance manquante est celle d'une absence :
pourquoi cette ligne *n'est-elle pas* dans le résultat ?

BBC News, 5 octobre 2020, « Excel: Why using Microsoft's tool caused Covid-19 results to be lost »

Mars 2009 – Le compositeur Maurice Jarre meurt. Un étudiant ajoute dans sa notice Wikipédia une citation qu'il vient d'inventer : élégante, plausible, sans source.



Deux autres histoires vraies

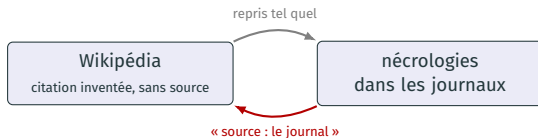
Octobre 2020 – L'agence de santé publique anglaise perd **15 841 tests positifs** au COVID. Les résultats des laboratoires étaient rassemblés dans le vieux format .xls, qui s'arrête à **65 536 lignes**. Au-delà, rien n'est refusé : les lignes sont simplement absentes.

65534	positif
65535	positif
65536	positif
65537	positif
65538	positif jamais lues
...	...

La provenance manquante est celle d'une absence : pourquoi cette ligne *n'est-elle pas* dans le résultat ?

BBC News, 5 octobre 2020, « Excel: Why using Microsoft's tool caused Covid-19 results to be lost »

Mars 2009 – Le compositeur Maurice Jarre meurt. Un étudiant ajoute dans sa notice Wikipédia une citation qu'il vient d'inventer : élégante, plausible, sans source.



Vue de l'extérieur, l'affirmation est sourcée. La chaîne est circulaire, mais rien ne le montre.

The Irish Times, 6 mai 2009, « Student's Wikipedia hoax quote used worldwide in newspaper obituaries »

Ce qu'on voudrait pouvoir demander à une machine

Devant un résultat – un chiffre, une ligne dans un tableau, une réponse à une requête :

- **D'où** vient-il ? De quelle source, de quelle case ?
- **Pourquoi** l'obtient-on ? Quelles données ont été nécessaires ?
- **Le plus rapide**, le moins cher ?
- **Combien** de fois, de combien de manières ?
- **Et si** telle source disparaissait, ou était fausse ?
- **Avec quelle confiance**, si les sources sont inégalement fiables ?
- **Quelle habilitation** faut-il pour le voir ?
- **Quelle source a eu le plus de poids** ?

Ce qu'on voudrait pouvoir demander à une machine

Devant un résultat – un chiffre, une ligne dans un tableau, une réponse à une requête :

- **D'où** vient-il ? De quelle source, de quelle case ?
- **Pourquoi** l'obtient-on ? Quelles données ont été nécessaires ?
- **Le plus rapide**, le moins cher ?
- **Combien** de fois, de combien de manières ?
- **Et si** telle source disparaissait, ou était fausse ?
- **Avec quelle confiance**, si les sources sont inégalement fiables ?
- **Quelle habilitation** faut-il pour le voir ?
- **Quelle source a eu le plus de poids ?**

Un résultat sans sa provenance,
on ne peut que le croire **sur parole**.

Ce qu'on voudrait pouvoir demander à une machine

Devant un résultat – un chiffre, une ligne dans un tableau, une réponse à une requête :

- **D'où** vient-il ? De quelle source, de quelle case ?
- **Pourquoi** l'obtient-on ? Quelles données ont été nécessaires ?
- **Le plus rapide**, le moins cher ?
- **Combien** de fois, de combien de manières ?
- **Et si** telle source disparaissait, ou était fausse ?
- **Avec quelle confiance**, si les sources sont inégalement fiables ?
- **Quelle habilitation** faut-il pour le voir ?
- **Quelle source a eu le plus de poids ?**

Un résultat sans sa provenance,
on ne peut que le croire **sur parole**.

Et ces huit questions ont une réponse commune : la **provenance**.

**Une même recette,
beaucoup de questions**

1. On donne un **numéro** à chaque donnée de départ : $t_1, t_2, t_3 \dots$ une étiquette, comme un numéro d'inventaire au dos d'un tableau

Une idée toute bête

1. On donne un **numéro** à chaque donnée de départ : $t_1, t_2, t_3 \dots$ une étiquette, comme un numéro d'inventaire au dos d'un tableau
2. On demande au système de **transporter les étiquettes** à travers le calcul

Une idée toute bête

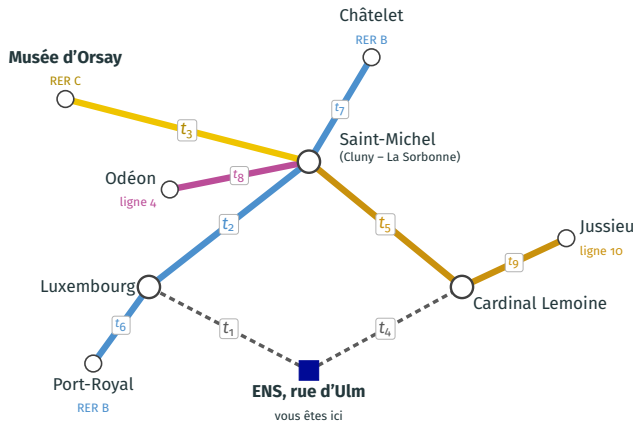
1. On donne un **numéro** à chaque donnée de départ : $t_1, t_2, t_3 \dots$ une étiquette, comme un numéro d'inventaire au dos d'un tableau
2. On demande au système de **transporter les étiquettes** à travers le calcul
3. Chaque résultat ressort avec, accrochée à lui, une petite **expression** qui dit comment les étiquettes ont été combinées

Une idée toute bête

1. On donne un **numéro** à chaque donnée de départ : $t_1, t_2, t_3 \dots$ une étiquette, comme un numéro d'inventaire au dos d'un tableau
2. On demande au système de **transporter les étiquettes** à travers le calcul
3. Chaque résultat ressort avec, accrochée à lui, une petite **expression** qui dit comment les étiquettes ont été combinées

L'idée a un quart de siècle [Buneman et al., 2000, 2001]. Ce qui a changé, c'est qu'on a compris ce que ces expressions *sont*, et ce qu'on peut en faire.

Aller de l'ENS au musée d'Orsay

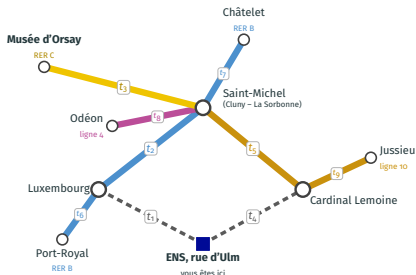


Un morceau du plan du quartier, tel qu'une base de données le stocke : **une ligne par tronçon**, et une étiquette par ligne.

	de	vers	ligne	min
t_1	ENS	Luxembourg	à pied	7
t_2	Luxembourg	Saint-Michel	RER B	3
t_3	Saint-Michel	Musée d'Orsay	RER C	4
t_4	ENS	Cardinal Lemoine	à pied	9
t_5	Cardinal Lemoine	Saint-Michel	ligne 10	5
t_6	Luxembourg	Port-Royal	RER B	2
t_7	Saint-Michel	Châtelet	RER B	2
t_8	Saint-Michel	Odéon	ligne 4	1
t_9	Cardinal Lemoine	Jussieu	ligne 10	1

La question, la réponse, et sa provenance

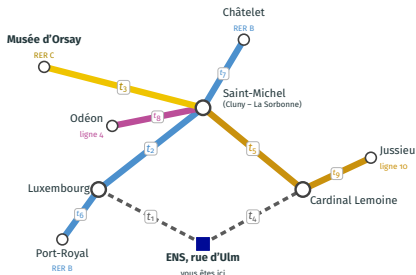
Question : « peut-on aller de l'ENS au musée d'Orsay en trois tronçons ? » – quatre lignes de SQL, le langage informatique des bases de données.



La question, la réponse, et sa provenance

Question : « peut-on aller de l'ENS au musée d'Orsay en trois tronçons ? » – quatre lignes de SQL, le langage informatique des bases de données.

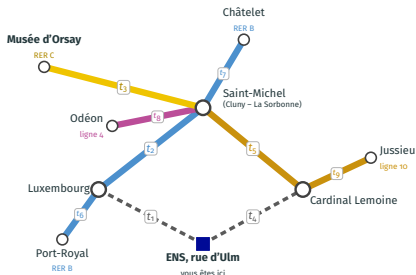
Réponse : oui.



La question, la réponse, et sa provenance

Question : « peut-on aller de l'ENS au musée d'Orsay en trois tronçons ? » – quatre lignes de SQL, le langage informatique des bases de données.

Réponse : oui. Et, avec elle, sa **provenance** :

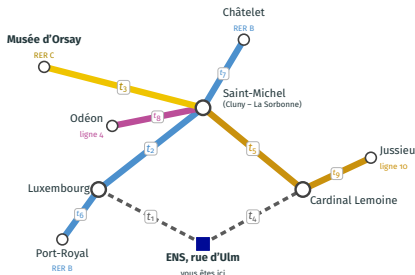


$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

La question, la réponse, et sa provenance

Question : « peut-on aller de l'ENS au musée d'Orsay en trois tronçons ? » – quatre lignes de SQL, le langage informatique des bases de données.

Réponse : oui. Et, avec elle, sa **provenance** :



$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

- Le signe $\otimes$ se lit « **et** » : les trois tronçons d'un itinéraire
- Le signe $\oplus$ se lit « **ou** » : par Luxembourg, ou par Cardinal Lemoine
- Le t_3 , le RER C, est dans les **deux**. Retenez-le.
- Cette formule, personne ne l'a écrite : le système l'a **produite tout seul**, en même temps que la réponse

Première question : et si?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Première question : et si?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Grève sur le RER B. t_2 devient *faux*, le reste *vrai*, $\otimes$ devient *et*, $\oplus$ devient *ou* :

(vrai et faux et vrai) ou (vrai et vrai et vrai) = ✓ *vrai* on passe par Cardinal Lemoine

Première question : et si ?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Grève sur le RER B. t_2 devient *faux*, le reste *vrai*, $\otimes$ devient *et*, $\oplus$ devient *ou* :

(vrai et faux et vrai) ou (vrai et vrai et vrai) = ✓ **vrai** on passe par Cardinal Lemoine

Travaux sur le RER C. t_3 devient *faux* :

(... et faux) ou (... et faux) = ✗ **faux**

Première question : et si ?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Grève sur le RER B. t_2 devient *faux*, le reste *vrai*, $\otimes$ devient *et*, $\oplus$ devient *ou* :

(vrai et faux et vrai) ou (vrai et vrai et vrai) = ✓ **vrai** on passe par Cardinal Lemoine

Travaux sur le RER C. t_3 devient *faux* :

(... et faux) ou (... et faux) = ✗ **faux**

Je n'ai pas relancé la requête : j'ai relu la provenance.

Deuxième question : combien ?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Maintenant je décide que $\otimes$ est la **multiplication** des entiers et $\oplus$ l'**addition**, et que chaque tronçon vaut 1 :

$$1 \times 1 \times 1 + 1 \times 1 \times 1 = 2$$

Deuxième question : combien ?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Maintenant je décide que $\otimes$ est la **multiplication** des entiers et $\oplus$ l'**addition**, et que chaque tronçon vaut 1 :

$$1 \times 1 \times 1 + 1 \times 1 \times 1 = 2$$

Deux itinéraires. Et personne n'a écrit de programme pour compter les itinéraires.

Deuxième question : combien ?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Maintenant je décide que $\otimes$ est la **multiplication** des entiers et $\oplus$ l'**addition**, et que chaque tronçon vaut 1 :

$$1 \times 1 \times 1 + 1 \times 1 \times 1 = 2$$

Deux itinéraires. Et personne n'a écrit de programme pour compter les itinéraires.

Si le RER C propose **trois** horaires qui conviennent, je mets 3 à la place de t_3 :

$1 \times 1 \times 3 + 1 \times 1 \times 3 = 6$ voyages possibles.

Deuxième question : combien ?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Maintenant je décide que $\otimes$ est la **multiplication** des entiers et $\oplus$ l'**addition**, et que chaque tronçon vaut 1 :

$$1 \times 1 \times 1 + 1 \times 1 \times 1 = 2$$

Deux itinéraires. Et personne n'a écrit de programme pour compter les itinéraires.

Si le RER C propose **trois** horaires qui conviennent, je mets 3 à la place de t_3 :

$$1 \times 1 \times 3 + 1 \times 1 \times 3 = 6 \text{ voyages possibles.}$$

La même recette. Un autre calcul. Une autre question.

Troisième question : le plus rapide ?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Cette fois, chaque étiquette vaut une **durée** (la colonne *min* de la table), et je décide que

- $\otimes$ devient $+$ (les durées d'un itinéraire s'**ajoutent**)
- $\oplus$ devient $\min$ (entre deux itinéraires, je prends le **meilleur**)

Troisième question : le plus rapide ?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Cette fois, chaque étiquette vaut une **durée** (la colonne *min* de la table), et je décide que

- $\otimes$ devient $+$ (les durées d'un itinéraire s'**ajoutent**)
- $\oplus$ devient $\min$ (entre deux itinéraires, je prends le **meilleur**)

Avec les durées de la table, $7 + 3 + 4$ par Luxembourg et $9 + 5 + 4$ par Cardinal Lemoine :

$$\min(7 + 3 + 4, 9 + 5 + 4) = \min(14, 18) = \mathbf{14} \text{ minutes}$$

Troisième question : le plus rapide ?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Cette fois, chaque étiquette vaut une **durée** (la colonne *min* de la table), et je décide que

- $\otimes$ devient $+$ (les durées d'un itinéraire s'**ajoutent**)
- $\oplus$ devient $\min$ (entre deux itinéraires, je prends le **meilleur**)

Avec les durées de la table, $7 + 3 + 4$ par Luxembourg et $9 + 5 + 4$ par Cardinal Lemoine :

$$\min(7 + 3 + 4, 9 + 5 + 4) = \min(14, 18) = \mathbf{14} \text{ minutes}$$

C'est le plus court chemin : exactement le calcul de Dijkstra, celui qui tourne dans votre application d'itinéraires.

Troisième question : le plus rapide ?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Cette fois, chaque étiquette vaut une **durée** (la colonne *min* de la table), et je décide que

- $\otimes$ devient $+$ (les durées d'un itinéraire s'**ajoutent**)
- $\oplus$ devient $\min$ (entre deux itinéraires, je prends le **meilleur**)

Avec les durées de la table, $7 + 3 + 4$ par Luxembourg et $9 + 5 + 4$ par Cardinal Lemoine :

$$\min(7 + 3 + 4, 9 + 5 + 4) = \min(14, 18) = \mathbf{14} \text{ minutes}$$

C'est le plus court chemin : exactement le calcul de Dijkstra, celui qui tourne dans votre application d'itinéraires.

Des **prix** à la place des durées, et c'est le trajet le moins cher; des **émissions de CO₂**, le moins émetteur.

Quatrième question : avec quelle confiance ?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Chaque tronçon a une **probabilité** d'être assuré : à pied, 1; ligne 10, 0,95; RER C, 0,9; RER B, 0,8.
Avec quelle probabilité arrive-t-on au musée d'Orsay? Nouvelles règles de lecture :

- $\otimes$ devient $\times$ (il faut les deux tronçons : les probabilités se **multiplient**)
- $\oplus$ devient « au moins l'un des deux » $(1 - (1 - p)(1 - q))$: il suffit que **l'un** soit assuré)

mais ces règles ne marchent que si chaque tronçon n'apparaît qu'une seule fois.

Quatrième question : avec quelle confiance ?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Chaque tronçon a une **probabilité** d'être assuré : à pied, 1; ligne 10, 0,95; RER C, 0,9; RER B, 0,8.
Avec quelle probabilité arrive-t-on au musée d'Orsay? Nouvelles règles de lecture :

- $\otimes$ devient $\times$ (il faut les deux tronçons : les probabilités se **multiplient**)
- $\oplus$ devient « au moins l'un des deux » $(1 - (1 - p)(1 - q))$: il suffit que **l'un** soit assuré)

mais ces règles ne marchent que si chaque tronçon n'apparaît qu'une seule fois. Le RER C est dans les deux itinéraires : on le met d'abord **en facteur**, $(t_1 \otimes t_2 \oplus t_4 \otimes t_5) \otimes t_3$, puis on lit :

par Luxembourg, $1 \times 0,8 = 0,8$; par Cardinal Lemoine, $1 \times 0,95 = 0,95$;
au moins l'un des deux, $1 - (1 - 0,8)(1 - 0,95) = 0,99$; puis le RER C, $\times 0,9$, soit **0,89**.

Quatrième question : avec quelle confiance ?

$$t_1 \otimes t_2 \otimes t_3 \oplus t_4 \otimes t_5 \otimes t_3$$

Chaque tronçon a une **probabilité** d'être assuré : à pied, 1; ligne 10, 0,95; RER C, 0,9; RER B, 0,8.
Avec quelle probabilité arrive-t-on au musée d'Orsay? Nouvelles règles de lecture :

- $\otimes$ devient $\times$ (il faut les deux tronçons : les probabilités se **multiplient**)
- $\oplus$ devient « au moins l'un des deux » $(1 - (1 - p)(1 - q))$: il suffit que **l'un** soit assuré)

mais ces règles ne marchent que si chaque tronçon n'apparaît qu'une seule fois. Le RER C est dans les deux itinéraires : on le met d'abord **en facteur**, $(t_1 \otimes t_2 \oplus t_4 \otimes t_5) \otimes t_3$, puis on lit :

par Luxembourg, $1 \times 0,8 = 0,8$; par Cardinal Lemoine, $1 \times 0,95 = 0,95$;
au moins l'un des deux, $1 - (1 - 0,8)(1 - 0,95) = 0,99$; puis le RER C, $\times 0,9$, soit **0,89**.

C'est le domaine des **bases de données probabilistes** [Suciu et al., 2011].

Une seule provenance, sept manières de calculer

Question	$\otimes$ devient	$\oplus$ devient	Ce qu'on obtient
Est-ce encore vrai si...?	et	ou	robustesse
Combien de façons?	$\times$	$+$	multiplicités, comptage
Le plus rapide, le moins cher?	$+$	min	optimisation, Dijkstra
Avec quelle probabilité?	et / ou, puis un calcul de probabilité sur la formule logique		incertitude
Quelle habilitation faut-il?	la plus élevée	la moins élevée	contrôle d'accès
Quelle fiabilité minimale?	min	max	confiance dans les sources
Quelles sources exactement?	réunion	réunion	ensemble des témoins

Une seule provenance, sept manières de calculer

Question	$\otimes$ devient	$\oplus$ devient	Ce qu'on obtient
Est-ce encore vrai si...?	et	ou	robustesse
Combien de façons?	$\times$	$+$	multiplicités, comptage
Le plus rapide, le moins cher?	$+$	min	optimisation, Dijkstra
Avec quelle probabilité?	et / ou, puis un calcul de probabilité sur la formule logique		incertitude
Quelle habilitation faut-il?	la plus élevée	la moins élevée	contrôle d'accès
Quelle fiabilité minimale?	min	max	confiance dans les sources
Quelles sources exactement?	réunion	réunion	ensemble des témoins

Ces couples d'opérations ont tous les mêmes propriétés : les mathématiciens appellent cela un **semi-anneau**. Et il y a un théorème [Green et al., 2007] : pour une large classe de requêtes, une provenance **universelle**, calculée **une** fois, donne chacune de ces réponses par substitution. La probabilité demande un pas de plus, et c'est là que se cache un piège.

Une seule provenance, sept manières de calculer

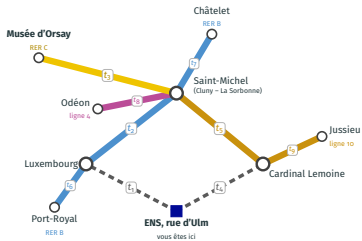
Question	$\otimes$ devient	$\oplus$ devient	Ce qu'on obtient
Est-ce encore vrai si...?	et	ou	robustesse
Combien de façons?	$\times$	$+$	multiplicités, comptage
Le plus rapide, le moins cher?	$+$	min	optimisation, Dijkstra
Avec quelle probabilité?	et / ou, puis un calcul de probabilité sur la formule logique		incertitude
Quelle habilitation faut-il?	la plus élevée	la moins élevée	contrôle d'accès
Quelle fiabilité minimale?	min	max	confiance dans les sources
Quelles sources exactement?	réunion	réunion	ensemble des témoins

Ces couples d'opérations ont tous les mêmes propriétés : les mathématiciens appellent cela un **semi-anneau**. Et il y a un théorème [Green et al., 2007] : pour une large classe de requêtes, une provenance **universelle**, calculée **une** fois, donne chacune de ces réponses par substitution. La probabilité demande un pas de plus, et c'est là que se cache un piège.

Le système n'a pas besoin de savoir quelle question vous allez poser.

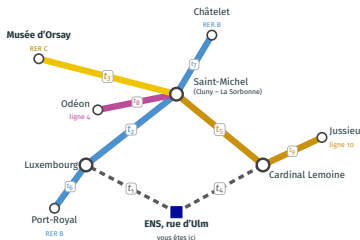
**Pourquoi c'est difficile,
et comment on s'en sort**

Un joli piège

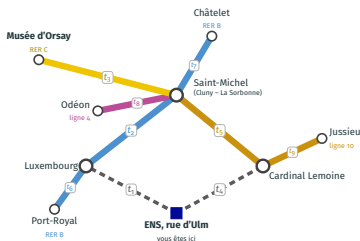


Un joli piège

Sans mettre le RER C en facteur : « par Luxembourg, $0,8 \times 0,9 = 0,72$; par Cardinal Lemoine, $0,95 \times 0,9 = 0,855$; au moins l'un des deux, $1 - (1 - 0,72)(1 - 0,855) = 0,96$ ».

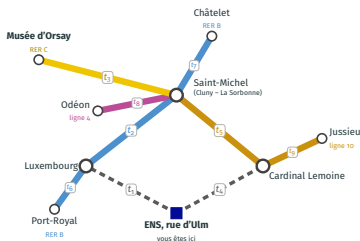


Un joli piège



Sans mettre le RER C en facteur : « par Luxembourg, $0,8 \times 0,9 = 0,72$; par Cardinal Lemoine, $0,95 \times 0,9 = 0,855$; au moins l'un des deux, $1 - (1 - 0,72)(1 - 0,855) = 0,96$ ».

C'est faux – la bonne valeur est 0,89 – parce que les deux itinéraires passent par le **RER C** : s'il tombe, ils tombent **ensemble**. Le t_3 commun aux deux termes de la recette interdit de les traiter comme indépendants.



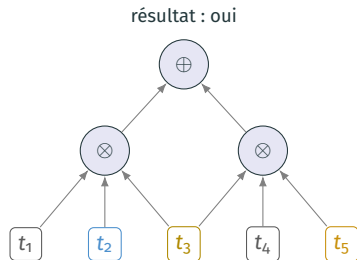
Sans mettre le RER C en facteur : « par Luxembourg, $0,8 \times 0,9 = 0,72$; par Cardinal Lemoine, $0,95 \times 0,9 = 0,855$; au moins l'un des deux, $1 - (1 - 0,72)(1 - 0,855) = 0,96$ ».

C'est faux – la bonne valeur est 0,89 – parce que les deux itinéraires passent par le **RER C** : s'il tombe, ils tombent **ensemble**. Le **t₃** commun aux deux termes de la recette interdit de les traiter comme indépendants.

- « Existe-t-il un itinéraire ? » : facile, même à grande échelle
- « Avec quelle probabilité ? », ou « dans combien de configurations du réseau existe-t-il un itinéraire ? » : **beaucoup** plus difficile, et ce n'est pas la faute des algorithmes : ces problèmes sont #P-complets [Valiant, 1979; Dalvi and Suciu, 2013]

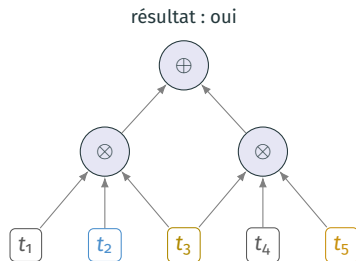
Des circuits, plutôt que des formules

Sur neuf tronçons, l'expression tient sur une diapositive. Sur le réseau entier de la région parisienne, les itinéraires possibles entre deux points se comptent par millions, et une expression écrite à plat les énumérerait un par un. La solution : **partager** les sous-recettes – un graphe, au lieu d'un arbre.



Des circuits, plutôt que des formules

Sur neuf tronçons, l'expression tient sur une diapositive. Sur le réseau entier de la région parisienne, les itinéraires possibles entre deux points se comptent par millions, et une expression écrite à plat les énumérerait un par un. La solution : **partager** les sous-recettes – un graphe, au lieu d'un arbre.



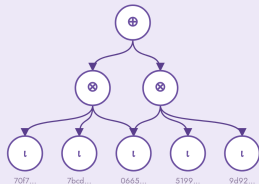
- Ici t_3 est écrit **une** fois et sert **deux** fois
- Ce sont les mêmes objets que les circuits logiques, et la même question : quelle *forme* de circuit rend telle question facile
- Toute une discipline, la **compilation de connaissances**, s'occupe de cela, et c'est elle qui rend les probabilités calculables en pratique

Cela tourne vraiment

Provenance Circuit



8 gates - root d108...



Evaluate → root d108...

Formula

lettre (text)

Run = (t1 * t2 * t3) * (t4 * t5 * t3)

Your Query

search_path: public, provsql

```
SQL SELECT DISTINCT t1.de, t3.vers
FROM troncon t1, troncon t2, troncon t3
WHERE t1.vers = t2.de AND t2.vers = t3.de
AND t1.de = 'ENS' AND t3.vers = 'Musée
d'Orsay';
```

Provenance scheme

Boolean Absorptive **Semiring** Where

☐ update_provenance

History

Send query

Query result

1 TUPLE - EVALUATED IN 20 MS

DE	VERS	PROVSQL	PROV
ENS	Musée d'Orsay	d108...	

Click a UUID / agg_token cell in the result to inspect its circuit.

ProvSQL, notre extension de PostgreSQL, et son interface Studio.

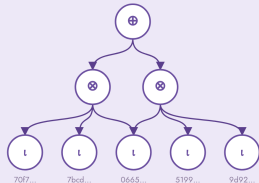
À droite, la question en SQL et sa réponse; à gauche, le circuit, et la formule relue avec les étiquettes.

Cela tourne vraiment

Provenance Circuit



8 gates - root d108...



Evaluate → root d108...

Formula

lettre (text)

Run = (t1 ⊗ t2 ⊗ t3) ⊗ (t4 ⊗ t5 ⊗ t3)

Your Query

search_path: public, provsql

```
SQL SELECT DISTINCT t1.de, t3.vers
FROM troncon t1, troncon t2, troncon t3
WHERE t1.vers = t2.de AND t2.vers = t3.de
AND t1.de = 'ENS' AND t3.vers = 'Musée
d'Orsay';
```

Provenance scheme

Boolean Absorptive Semiring Where

update_provenance

History

Send query

Query result

1 TUPLE - EVALUATED IN 20 MS

DE	VERS	PROVSQL PROV
ENS	Musée d'Orsay	d108...

Click a UUID / agg_token cell in the result to inspect its circuit.

ProvSQL, notre extension de PostgreSQL, et son interface Studio.

À droite, la question en SQL et sa réponse; à gauche, le circuit, et la formule relue avec les étiquettes.

Le nœud qui sert deux fois :

Input gate (i): base tuple

UUID	0665...
DEPTH	2
PROBABILITY	0.9000
troncon	
LETTRE	t3
DE	Saint-Michel
VERS	Musée d'Orsay
LIGNE	RER C
MINUTES	4
PONCTUALITE	0.9

La même recette, quatre calculs, en un clic

Evaluate ? → root d108...

Formula

lettre (text)



Run



$$= (t1 \otimes t2 \otimes t3) \oplus (t4 \otimes t5 \otimes t3)$$

· 12 ms

Evaluate ? → root d108...

Counting

un (integer)

Expects numeric values.



Run



$$= 2$$

· 15 ms

Evaluate ? → root d108...

Tropical (min-plus, nonnegative)

minutes (integer)

Expects nonnegative cost values; absorptive, so it accepts truncated cyclic-recursion tokens (min-cost reachability).



Run



$$= 14$$

· 13 ms

Evaluate ? → root d108...

Marginal probability

exact



Run



$$= 0.8910$$

via sieve · 12 ms

La même recette, quatre calculs, en un clic

Evaluate ? → root d108...

Formula

lettre (text)

Run   = (t1 ⊗ t2 ⊗ t3) ⊕ (t4 ⊗ t5 ⊗ t3)

· 12 ms

Evaluate ? → root d108...

Counting

un (integer)

Expects numeric values.

Run   = 2 · 15 ms

Evaluate ? → root d108...

Tropical (min-plus, nonnegative)

minutes (integer)

Expects nonnegative cost values; absorptive, so it accepts truncated cyclic-recursion tokens (min-cost reachability).

Run   = 14 · 13 ms

Evaluate ? → root d108...

Marginal probability

exact

Run   = 0.8910 via sieve · 12 ms

La recette, le nombre d'itinéraires, la durée minimale, la probabilité :
quatre lectures de la même provenance, calculée une seule fois.

Qui a eu le plus de poids?



Qui a eu le plus de poids?



La **valeur de Shapley**, empruntée à la théorie des jeux [Shapley, 1953], répartit entre les tronçons la responsabilité de la réponse [Livshits et al., 2021].

- Le RER C pèse le plus : il est sur les deux itinéraires
- Cardinal Lemoine pèse plus que Luxembourg : la ligne 10 est plus fiable que le RER B
- Les quatre tronçons qui ne mènent pas au musée d'Orsay pèsent zéro

Qui a eu le plus de poids?



La **valeur de Shapley**, empruntée à la théorie des jeux [Shapley, 1953], répartit entre les tronçons la responsabilité de la réponse [Livshits et al., 2021].

- Le RER C pèse le plus : il est sur les deux itinéraires
- Cardinal Lemoine pèse plus que Luxembourg : la ligne 10 est plus fiable que le RER B
- Les quatre tronçons qui ne mènent pas au musée d'Orsay pèsent zéro

Si vous ne pouvez vérifier qu'une donnée avant de partir, elle vous dit laquelle.

Ce que cela change

Expliquer une décision, effacer une donnée

Dossier refusé, place non obtenue, crédit rejeté. « Pourquoi moi ? » est une question de provenance : quelles données précises, combinées comment, ont produit cette sortie-là.

Expliquer une décision, effacer une donnée

Dossier refusé, place non obtenue, crédit rejeté. « Pourquoi moi ? » est une question de provenance : quelles données précises, combinées comment, ont produit cette sortie-là.

Le droit le demande déjà : en France, depuis 2016, toute personne visée par une décision prise sur le fondement d'un traitement algorithmique peut en obtenir les règles et les principales caractéristiques ; le RGPD encadre les décisions automatisées. **Le droit a formulé l'exigence ; la provenance est la forme technique de la réponse, et elle n'est presque jamais enregistrée.**

Expliquer une décision, effacer une donnée

Dossier refusé, place non obtenue, crédit rejeté. « Pourquoi moi ? » est une question de provenance : quelles données précises, combinées comment, ont produit cette sortie-là.

Le droit le demande déjà : en France, depuis 2016, toute personne visée par une décision prise sur le fondement d'un traitement algorithmique peut en obtenir les règles et les principales caractéristiques ; le RGPD encadre les décisions automatisées. **Le droit a formulé l'exigence ; la provenance est la forme technique de la réponse, et elle n'est presque jamais enregistrée.**

Droit à l'effacement. Supprimer la ligne est facile. La vraie question est : **quels résultats déjà calculés, publiés, mis en cache, copiés dans des tableaux de bord, dépendaient de cette ligne ?**

Expliquer une décision, effacer une donnée

Dossier refusé, place non obtenue, crédit rejeté. « Pourquoi moi ? » est une question de provenance : quelles données précises, combinées comment, ont produit cette sortie-là.

Le droit le demande déjà : en France, depuis 2016, toute personne visée par une décision prise sur le fondement d'un traitement algorithmique peut en obtenir les règles et les principales caractéristiques ; le RGPD encadre les décisions automatisées. **Le droit a formulé l'exigence ; la provenance est la forme technique de la réponse, et elle n'est presque jamais enregistrée.**

Droit à l'effacement. Supprimer la ligne est facile. La vraie question est : **quels résultats déjà calculés, publiés, mis en cache, copiés dans des tableaux de bord, dépendaient de cette ligne ?**

Sans provenance, il faut tout recalculer, ou espérer. Avec, on sait exactement quoi reprendre – et c'est la même machinerie que la mise à jour des vues et des caches, un problème que les bases de données connaissent depuis longtemps.

- Un modèle de langue produit une phrase. Sa provenance, au sens de cet exposé, **n'existe pas** : les paramètres mêlent tout ce qu'il a lu, sans garder trace de qui a contribué à quoi

- Un modèle de langue produit une phrase. Sa provenance, au sens de cet exposé, **n'existe pas** : les paramètres mêlent tout ce qu'il a lu, sans garder trace de qui a contribué à quoi
- Deux réponses très différentes : **citer**, en affichant des documents à côté de la réponse – à condition que la citation corresponde vraiment à ce qui a été utilisé; et **attribuer**, c'est-à-dire chercher quels exemples d'entraînement expliquent telle sortie. Problème ouvert, et enjeu technique de plusieurs procès en cours sur le droit d'auteur

- Un modèle de langue produit une phrase. Sa provenance, au sens de cet exposé, **n'existe pas** : les paramètres mêlent tout ce qu'il a lu, sans garder trace de qui a contribué à quoi
- Deux réponses très différentes : **citer**, en affichant des documents à côté de la réponse – à condition que la citation corresponde vraiment à ce qui a été utilisé; et **attribuer**, c'est-à-dire chercher quels exemples d'entraînement expliquent telle sortie. Problème ouvert, et enjeu technique de plusieurs procès en cours sur le droit d'auteur
- Pour les images, la réponse retenue par l'industrie est exactement celle des étiquettes au dos d'un tableau : **signer** la prise de vue, puis chaque retouche, et faire voyager la chaîne avec le fichier (format C2PA, 2021)

- Un modèle de langue produit une phrase. Sa provenance, au sens de cet exposé, **n'existe pas** : les paramètres mêlent tout ce qu'il a lu, sans garder trace de qui a contribué à quoi
- Deux réponses très différentes : **citer**, en affichant des documents à côté de la réponse – à condition que la citation corresponde vraiment à ce qui a été utilisé; et **attribuer**, c'est-à-dire chercher quels exemples d'entraînement expliquent telle sortie. Problème ouvert, et enjeu technique de plusieurs procès en cours sur le droit d'auteur
- Pour les images, la réponse retenue par l'industrie est exactement celle des étiquettes au dos d'un tableau : **signer** la prise de vue, puis chaque retouche, et faire voyager la chaîne avec le fichier (format C2PA, 2021)
- Deux limites : une signature prouve **qui affirme quoi**, pas que ce soit vrai; et une capture d'écran efface tout, gratuitement

Trois idées à emporter

1. Un résultat sans sa provenance, on ne peut que le croire sur parole. Impossible de le corriger, de le contester ou de le réutiliser.

Trois idées à emporter

1. **Un résultat sans sa provenance, on ne peut que le croire sur parole.** Impossible de le corriger, de le contester ou de le réutiliser.
2. **La provenance a une structure mathématique.** Une recette, quoi se combine avec quoi, qu'on relit dans des calculs différents. Toute la difficulté est de l'enregistrer au fil du calcul, sans intervention humaine : c'est un problème de recherche.

Trois idées à emporter

1. **Un résultat sans sa provenance, on ne peut que le croire sur parole.** Impossible de le corriger, de le contester ou de le réutiliser.
2. **La provenance a une structure mathématique.** Une recette, quoi se combine avec quoi, qu'on relit dans des calculs différents. Toute la difficulté est de l'enregistrer au fil du calcul, sans intervention humaine : c'est un problème de recherche.
3. **Le réflexe, lui, est à la portée de tous : demander la chaîne, et aller voir un maillon.** Ce chiffre, moyenne ou médiane ? Sur quelle population, quelle période ? Ce prix, relevé quand, chez quel marchand ? Ce lien mène-t-il à une source, ou à un autre article ? Exiger la provenance est un droit, et un acte citoyen.

Trois idées à emporter

1. **Un résultat sans sa provenance, on ne peut que le croire sur parole.** Impossible de le corriger, de le contester ou de le réutiliser.
2. **La provenance a une structure mathématique.** Une recette, quoi se combine avec quoi, qu'on relit dans des calculs différents. Toute la difficulté est de l'enregistrer au fil du calcul, sans intervention humaine : c'est un problème de recherche.
3. **Le réflexe, lui, est à la portée de tous : demander la chaîne, et aller voir un maillon.** Ce chiffre, moyenne ou médiane ? Sur quelle population, quelle période ? Ce prix, relevé quand, chez quel marchand ? Ce lien mène-t-il à une source, ou à un autre article ? Exiger la provenance est un droit, et un acte citoyen.

La première question de l'expert devant un tableau,
posons-la aux chiffres : **d'où vient cette donnée ?**

pierre.senellart@ens.psl.eu - <https://provsql.org/>

Pour aller plus loin

- Une introduction accessible aux idées de cet exposé [[Senellart, 2017](#)]
- Les trois questions « où, pourquoi, comment » [[Cheney et al., 2009](#)]
- Le résultat central sur les semi-anneaux [[Green et al., 2007](#)]
- Les bases de données probabilistes [[Suciu et al., 2011](#)]
- ProvSQL et ProvSQL Studio [[Senellart et al., 2018](#)] : <https://provsql.org/>

Références

- Peter Buneman, Sanjeev Khanna, and Wang Chiew Tan. Data provenance : Some basic issues. In *FSTTCS*, 2000. [lien].
- Peter Buneman, Sanjeev Khanna, and Wang Chiew Tan. Why and where : A characterization of data provenance. In *ICDT*, 2001. [lien].
- James Cheney, Laura Chiticariu, and Wang Chiew Tan. Provenance in databases : Why, how, and where. *Foundations and Trends in Databases*, 1(4), 2009. [lien].
- Nilesh Dalvi and Dan Suciu. The dichotomy of probabilistic inference for unions of conjunctive queries. *J. ACM*, 59(6), 2013. [lien].
- Todd J Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *PODS*, 2007. [lien].
- Thomas Herndon, Michael Ash, and Robert Pollin. Does high public debt consistently stifle economic growth ? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 38(2), 2014. [lien].
- Ester Livshits, Leopoldo Bertossi, Benny Kimelfeld, and Moshe Sebag. The Shapley value of tuples in query answering. *Logical Methods in Computer Science*, 17, 2021. [lien].
- Pierre Senellart. Provenance and probabilities in relational databases: From theory to practice. *SIGMOD Record*, 46(4), December 2017. [lien].
- Pierre Senellart, Louis Jachiet, Silviu Maniu, and Yann Ramusat. ProvSQL : Provenance and probability management in PostgreSQL. *Proc. VLDB Endow.*, 11(12), 2018. Demonstration. [lien].
- Lloyd S. Shapley. A value for n -person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*. Princeton University Press, 1953. [lien].
- Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch. *Probabilistic Databases*. Morgan & Claypool, 2011. [lien].
- Leslie G. Valiant. The complexity of enumeration and reliability problems. *SIAM J. Comput.*, 8(3), 1979. [lien].