

Information extraction from data with statistically enhanced LLMs

Antoine Soetewey & Cédric Heuchenne
Beamm.Conf26, June 30 2026

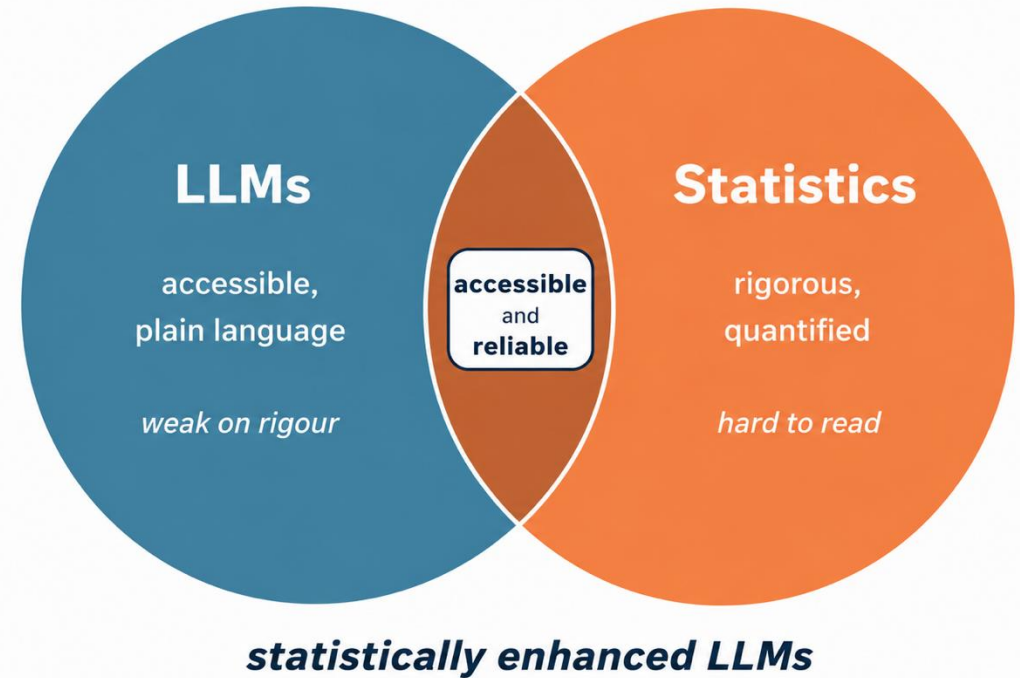
Rich data, but hard to use...

- **Belgium and Europe collect large-scale administrative and survey data** on key dimensions of socio-economic life:
 - Income and taxation (IPCAL, EU-SILC)
 - Labour market and employment (LFS, CBSS)
 - Housing (EU-SILC, Housing data)
 - Consumption (HBS)
 - Mobility (OVG, Beldam, Monitor)
 - Household finances (HFCS)
 - Time use (HETUS)
 - etc.
- **Data is not public for confidentiality reasons**, but Beamm researchers have access under strict data-protection rules
- **Even with access, it is hard to use:**
 - Variables hide behind opaque names (e.g., HY040G, PL031) and descriptions are documented in codebooks of 200 to 600 pages
 - Even with a perfect description for each variable, it would be hard to navigate across all of them, and even more complicated to see associations between them

→ Unlocking this data's full potential requires new tools

Two halves of a solution

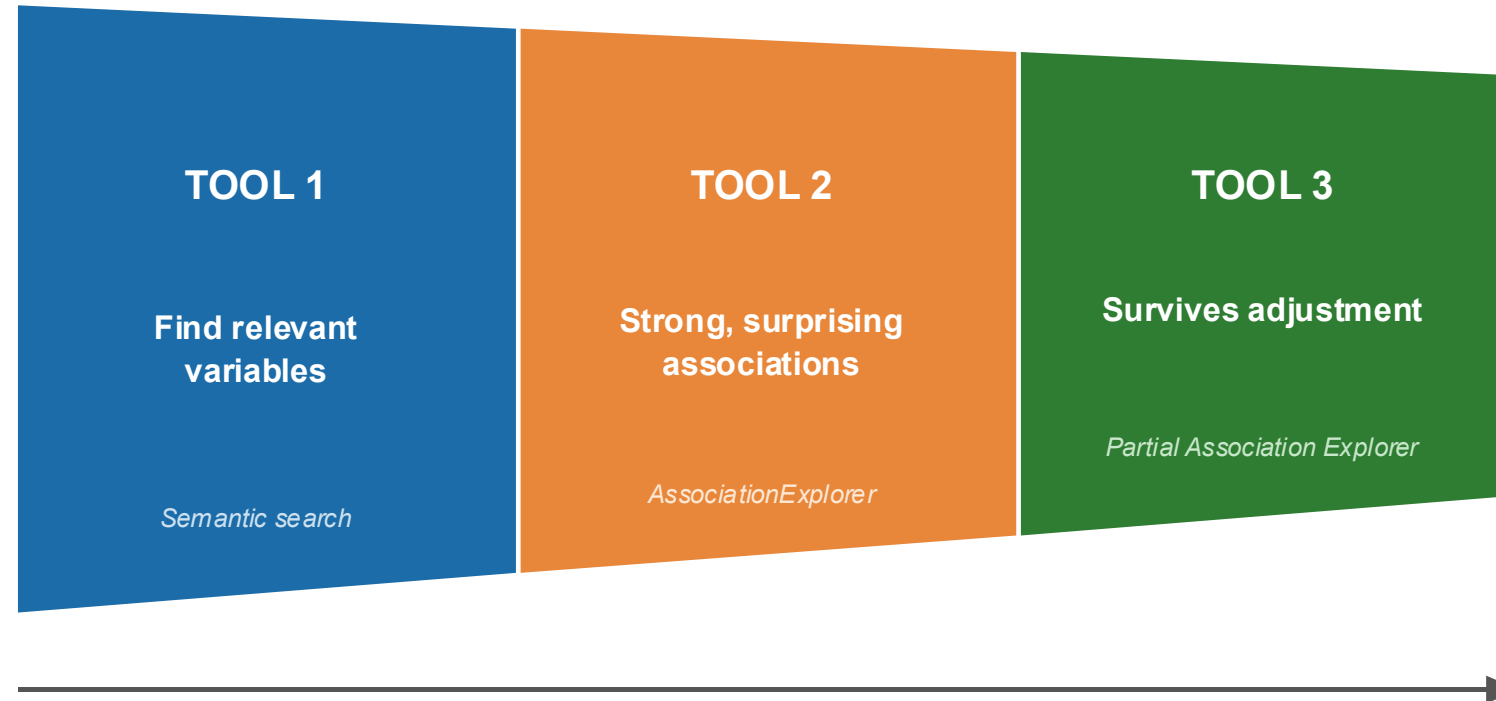
- **LLMs make data accessible** through plain language, but on their own they are inaccurate: they can hallucinate, not reliable at quantifying, and may call a trivial association interesting
- **Statistics are rigorous and quantified**, but opaque to a non-technical audience
- **The project combines the two**: statistically enhanced LLMs



A three-tool workflow

7,000+ opaque variables
across 6 Belgian socio-economic datasets

A reliable,
shareable data story



Tool 1: Semantic search

Or how to navigate among variables?

Collaboration with Carnot Dongmo Mbane and Hugues Annoye

- **6 Belgian datasets:** EU-SILC (income and living conditions), EU-LFS (labour), HBS (household consumption), HFCS (household finances), IPCAL (tax records), DEMOBEL (demographics)
- **7,010 variables in total.** Codes like HY040G, PL031, A0270 mean nothing without the documentation
- **Documentation is spread** across codebooks of 200 to 600 pages, in French, Dutch and English (and with a lot of technical details)
- A researcher studying a specific topic has **no obvious entry point**

7,010

variables across 6 datasets

200 to 600

pages of codebook each

The solution: Simply ask in plain language

- Query the corpus in natural language, instead of codes

- Two directions:
 - Find variables: from a question/topic to the relevant variables (« help you discover »)
 - Describe variables: from a code to a readable description (« help you understand »)

Recherche par Question

Recherche par Code

Posez votre question

Impact du COVID-19 sur les revenus

Nombre de résultats

10

3

20

Validation LLM (Llama 3.2)

Explique la pertinence de chaque variable

Filtrer par enquête (optionnel)

EU-SILC

HFCS

EU-LFS

HBS

IPCAL

DEMOBEL

Rechercher

Exemples

Cliquez pour tester

Quelles variables concernent les revenus locatifs?

Variables sur le patrimoine immobilier

Heures travaillées par semaine

Précarité énergétique en Wallonie

Impact du COVID-19 sur les revenus

Niveau d'éducation

Dépenses alimentaires

Résultats pour: "Impact du COVID-19 sur les revenus"

10 variables trouvées

IPCAL

Score: 0.926

1. B9596

Calcul d'impôt | la partie transférable des heures supplémentaires volontaires | COVID

La variable "Calcul d'impôt | la partie transférable des heures supplémentaires volontaires | COVID" mesure le montant d'impôt qui peut être transféré par un ménage en raison de l'accord sur les heures...

IPCAL

Score: 0.926

2. A9596

Calcul d'impôt | la partie transférable des heures supplémentaires volontaires | COVID

La variable A9596 mesure le calcul d'impôt sur les heures supplémentaires volontaires transférables liées à la pandémie de COVID-19. Cette variable est incluse dans l'enquête IPCAL (Institut national ...

IPCAL

Modifié

Score: 0.925

3. A7986

Calcul d'impôt | Réduction d'impôt | sur les acquisitions d'actions Covid | facturées - 2020

La variable A7986 mesure le calcul d'impôt ou la réduction d'impôt sur les acquisitions d'actions Covid facturées en 2020. Cette variable est issue de l'enquête IPCAL et figure dans le contexte des po...

IPCAL

Modifié

Score: 0.925

4. A8099

Calcul d'impôt | Réduction d'impôt | sur les acquisitions d'actions Covid | facturées - 2020

La variable A8099 mesure le calcul d'impôt ou la réduction d'impôt sur les acquisitions d'actions Covid-2020 facturées par les ménages ou les individus. Ce type de variable est essentiel pour comprend...

EU-SILC

DUAL_N

HY040N

INCOME FROM RENTAL OF A

Explication simple

La variable "INCOME FROM RENTAL OF A" (HY040N) mesure le revenu généré par l'exploitation locative d'un bien immobilier. Dans le contexte de l'enquête EU-SILC, cette variable est recueillie auprès des ménages pour évaluer leur situation économique et leur capacité à payer les impôts. L'enquête EU-SILC vise à fournir des informations sur la situation socio-économique des populations de l'Union européenne, notamment en ce qui concerne le revenu, les dépenses, les dettes et les conditions de vie. La variable "INCOME FROM RENTAL OF A" représente une valeur numérique qui correspond au montant du revenu généré par l'exploitation locative d'un bien immobilier. Cette valeur est exprimée en euros et est mesurée sur un an. L'unité de mesure est donc l'euro. Un exemple concret de cette variable pourrait être : "Je loue une maison dans le quartier pour 1 200 euros par mois, après avoir payé les frais de location, les impôts et les réparations". Dans ce cas, la valeur de la variable serait de 14 400 euros par an. Cette variable peut être utilisée dans une analyse pour évaluer l'impact du revenu généré par l'exploitation locative sur la situation économique des ménages. Par exemple, on pourrait analyser comment le revenu généré par l'exploitation locative affecte les dépenses de consommation ou les dettes des ménages. Cependant, il est important de prendre en compte les précautions d'interprétation lors de l'utilisation de cette variable. En effet, la valeur de "INCOME FROM RENTAL OF A" peut être influencée par divers facteurs tels que le coût du logement, les impôts, les frais de maintenance et les taux d'intérêt. Il est donc important de prendre en compte ces facteurs lors de l'analyse. La variable "INCOME FROM RENTAL OF A" peut également être liée à d'autres variables telles que le revenu total du ménage, la taille du logement, les coûts de vie et les taux de chômage. En analysant ces relations, on peut mieux comprendre comment le revenu généré par l'exploitation locative affecte la situation économique des ménages. Enfin, cette variable est importante pour la recherche car elle permet de mieux comprendre l'impact du revenu généré par l'exploitation locative sur la situation économique des ménages. Cela peut aider les chercheurs à développer des politiques publiques plus efficaces pour soutenir les ménages qui dépendent de l'exploitation locative pour leur revenu.

DESCRIPTION TECHNIQUE

Income from rental of a property or land (HY040G): Income from rental of a property or land refers to the income received, during the income reference period, from renting a property (for example renting a dwelling not included in the profit/loss of unincorporated enterprises, receipts from boarders or lodgers, or rent from land) after deducting costs such as mortgage interest repayments, minor repairs, maintenance, insurance and other charges. Please note that income from rental of a property or land (HY040G) is a type of property income. Property income refers to all income received, less expenses, occurring during the income reference period by the owner of a financial asset or a tangible non-produced asset (land) in return for providing funds to, or putting the tangible non-produced asset at the disposal of, another institutional unit. In EU-SILC, it is broken down into: - Income from rental of a property or land (HY040G) - Interest dividends, profits from capital investment in an

Recherche par Question

Recherche par Code

Mode 3 — Lookup direct par code

Entrez un code de variable pour obtenir sa **fiche complète** : explication vulgarisée + description technique.

(Majuscules/minuscules acceptées : DI1410I = di1410i)

Code de la variable

HY040N

Explication vulgarisée (Llama 3.2)

Génère une explication simple + exemple concret

Afficher la fiche

Exemples

Cliquez pour tester

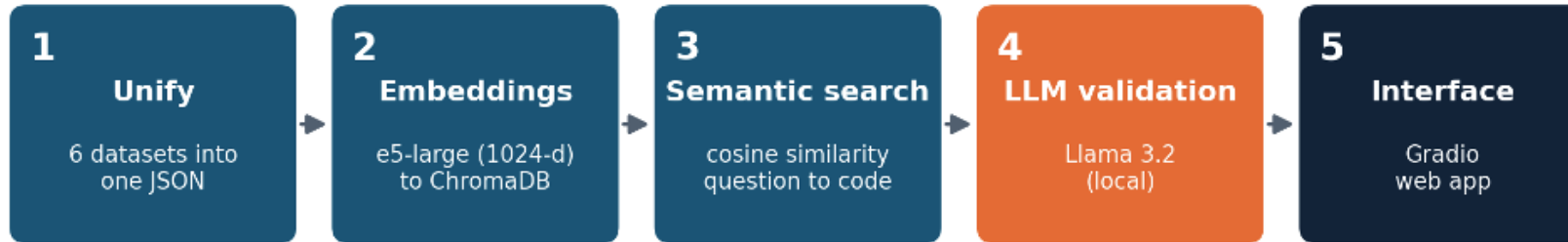
HY040N

PL073

DA1000

HWUSUAL

The process: A five-step pipeline



- 1. Unify**: each dataset has its own format; a per-dataset function maps it into one common JSON dictionary, one record per variable (with code, source, short and long description, category, etc.)
- 2. Embeddings**: every description becomes a 1024-dimensional vector with multilingual-e5-large (XLM-RoBERTa encoder, ~560M parameters), stored in ChromaDB; similar meanings are close together, even across languages
- 3. Semantic search**: the user's question is embedded too, and variables are ranked by cosine similarity (i.e., cosine of the angle between vectors)
- 4. LLM validation**: a local model (Llama 3.2 via Ollama) reads the top candidates and classifies each as truly relevant, somewhat relevant, or off-topic, with a short justification
- 5. Interface**: a Gradio web app ties it together

Two design choices: RAG, and running locally

Fine-tuning or RAG?

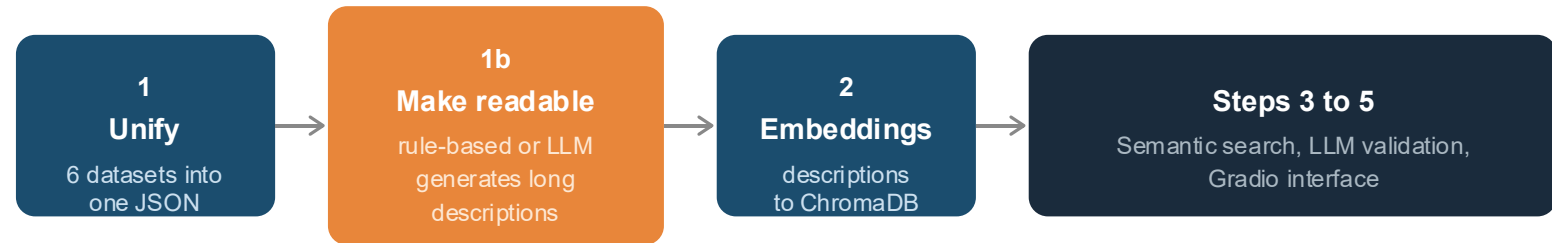
- **Fine-tuning** means retraining the model on our 7,000 variables so it “memorises” them
- **RAG** (Retrieval-Augmented Generation) works differently:
 - the model never sees our variables during training;
 - instead, when a question arrives, we first search for the most relevant variables (step 3, the **retrieval** part),
 - then hand them to the model as a note to read before answering (step 4, the **augmented generation** part)
- We started with the fine-tuning approach (LoRA on Llama-3-8B) but moved to RAG because:
 - adding a new dataset requires no retraining; just embedding and indexing the new variables (faster)
 - we always know exactly which variables were given to the model (more transparent)
 - it can only refer to variables we actually gave, not invent codes from an imperfect memory (fewer hallucinations)

Local or hosted?

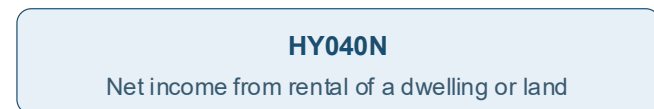
- **Local execution:** both the embedder and Llama 3.2 run locally through Ollama; might be weaker than a hosted model, but we accept it in return for:
 - Price (free)
 - Privacy (no change in the architecture if confidential metadata is added later)

Making variables more readable

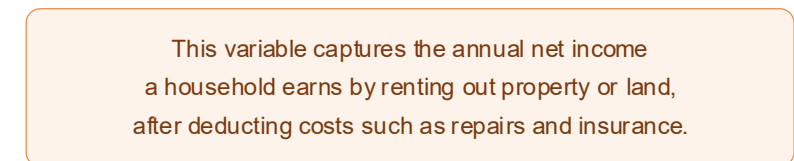
- Descriptions are often too **short and/or technical** to give good search results
- A second loop generates a long, **plain-language description** for each variable
- Two generators:
 1. a LLM-based (rich, with an example)
 2. a rule-based (instant, always available)
- These **enriched descriptions** are used as inputs for step 2, making the semantic search much more effective



Before (codebook label)



After (enriched description)



Tool 2: AssociationExplorer

From variables to associations

- We now have a shortlist of relevant variables
- Which of them are **related**, and **which links deserve a closer look**?
- **AssociationExplorer** (Soetewey et al., 2026) fills this gap:
 - Computes associations of every pair among the **selected variables**
 - But also: every pair between the selected variables **and the rest of the dataset** to highlight unexpected associations
- First, how do we actually compute associations?

Three cases depending on variable type

A variable can be of 2 types (numeric or categorical), so an association can be of 3 types:

1. **Numeric vs numeric:** Pearson's correlation coefficient (Pearson, 1895): $r = \frac{cov(x,y)}{s_x s_y}$
2. **Categorical vs categorical:** Cramer's V (Cramer, 1946): $V = \sqrt{\frac{\chi^2}{n \cdot \min(r - 1, c - 1)}}$
3. **Numeric vs categorical:** Correlation ratio (Pearson, 1905): $\eta = \sqrt{\frac{SS_{between}}{SS_{total}}}$

From all pairs to the interesting few

3-step workflow:

1. Keep only significant pairs: p -value < 0.05 :

- Numeric vs numeric – Pearson's correlation test: $t_{obs} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2}$
- Categorical vs categorical – Chi-square test of independence: $\chi^2_{obs} = \sum \frac{(O-E)^2}{E} \sim \chi^2_{(r-1, c-1)}$
- Numeric vs categorical – One-way ANOVA F-test: $F_{obs} = \frac{MS_{between}}{MS_{within}} \sim F_{N-k}^{k-1}$

2. Drop negligible ones: minimum strength of at least 1% explained variance (Cohen's (1988) small effect):

- Numeric vs numeric: $r^2 \geq 0.01$
- Categorical vs categorical: $V^2 \geq 0.01$
- Numeric vs categorical: $R^2_{adj} \geq 0.01$

3. Rank by strength, keep the top-N (N = 10)

- Assumptions underlying each test are not tested (non-parametric alternatives could be considered)
- Thresholds are fixed (no user-adjustable slider) to keep the tool accessible for non-specialists

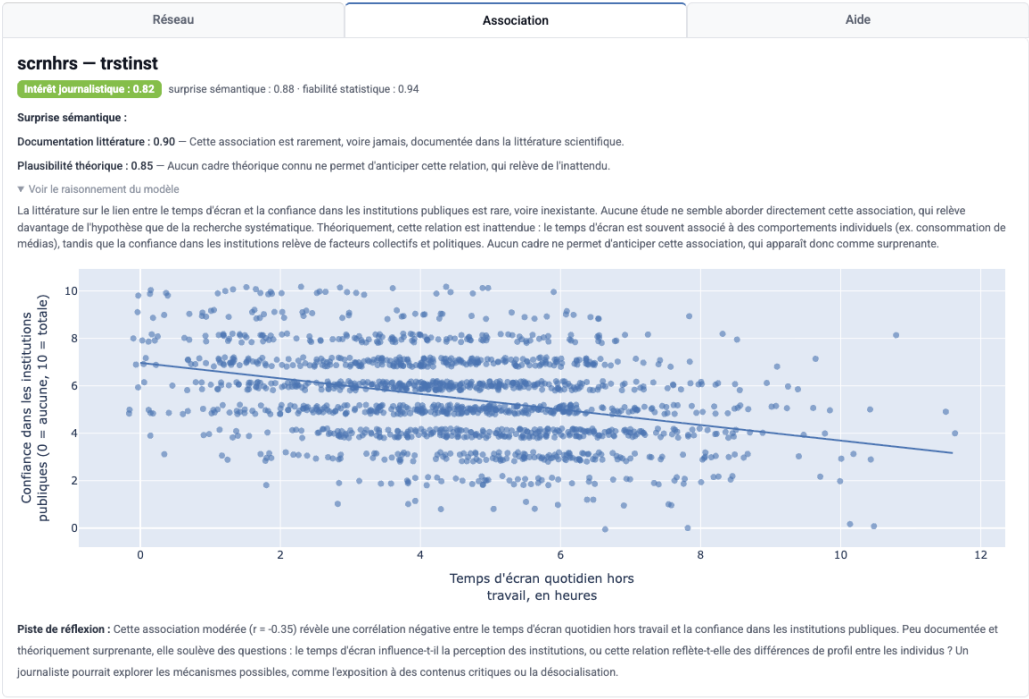
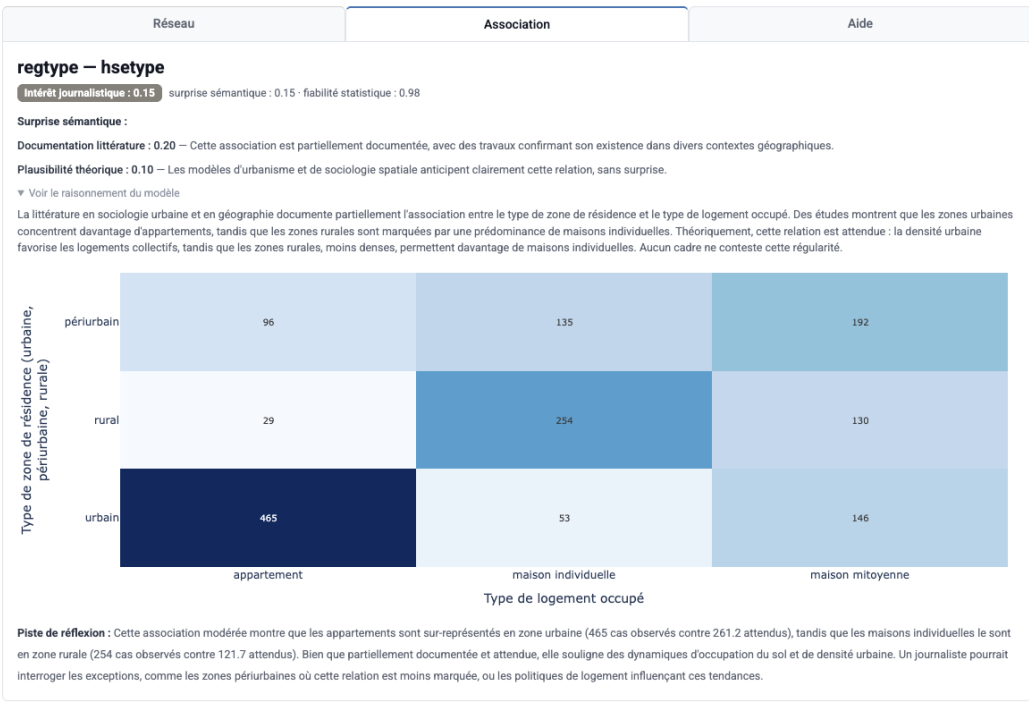
An LLM as a judge

- A statistically significant (via the p -value) and strong (via the strength) association is **not always an interesting association**, if it is expected (e.g., age and retirement, education and salary)
- So we add another perspective: **surprise**
 - An LLM (mistral-medium-latest, via Mistral API) receives the two descriptions and assigns two sub-scores:
 1. How documented is this association in the scientific literature? (0 = well established, 1 = rarely studied)
 2. How theoretically plausible is it? (0 = fully expected, 1 = surprising)
 - The semantic surprise score is the average of the two
- Finally, **journalistic interest = (semantic) surprise × (statistical) reliability**, with $\text{reliability} = \frac{-\log_{10}(p)}{-\log_{10}(p) + 3}$ (to distinguish very small p -values)
 - The final score is high only when an association is **both surprising and statistically solid**; the ideal candidate for a story
- As a bonus: for each pair, the LLM also proposes hypotheses to explore as « **food for thought** »:
 - With the observed data (correlation, mean by group, association strength, etc.)
 - But without concluding or claiming causality

- **Association network:** nodes are variables, edges are associations; closer nodes mean stronger links
- **Edge opacity** proportional to journalistic interest

For the selected association:

- All **scores:** final surprise score, 2 sub-scores, statistical reliability
- Model's **prior reasoning** (written before scoring)
- **Visualization**
- **Food for thought** grounded in the observed data



Tool 3: Partial Association Explorer

Strong and surprising, but is it real?

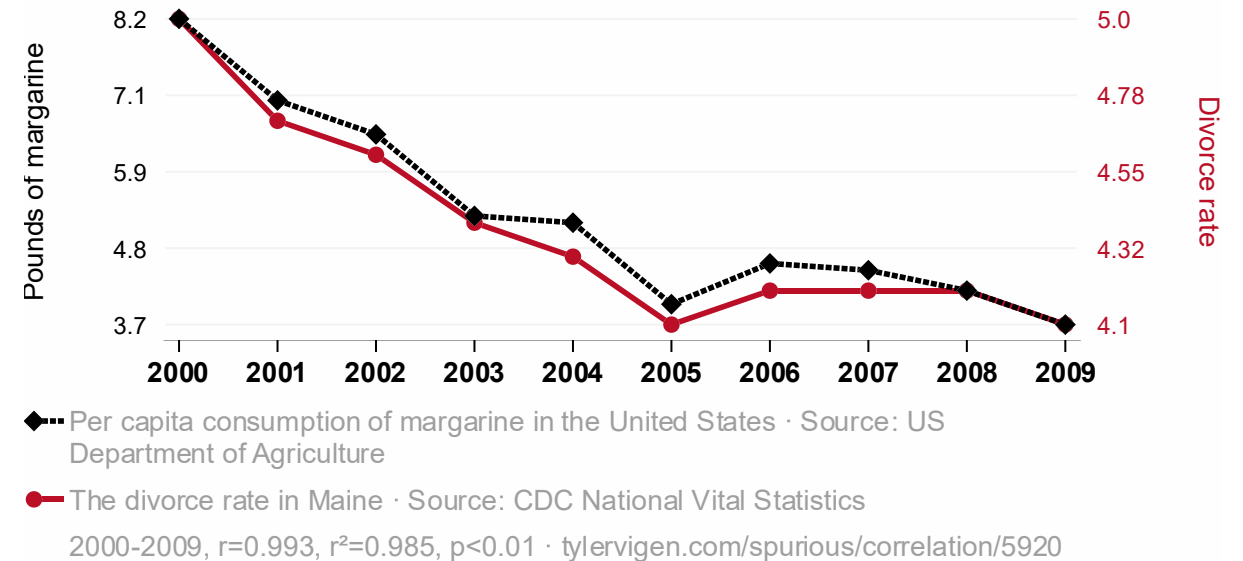
Collaboration with Thaddée D'haegeleer

- A strong, significant and surprising association **can still be an artefact** (spurious correlation)
- **Confounder**: a third variable Z that influences both X and Y, creating an association between two things that are not directly linked
- With many variables, probability of finding that kind of association **just by chance** is high
- To limit it, we add a more careful layer: **control variables**

Per capita consumption of margarine

correlates with

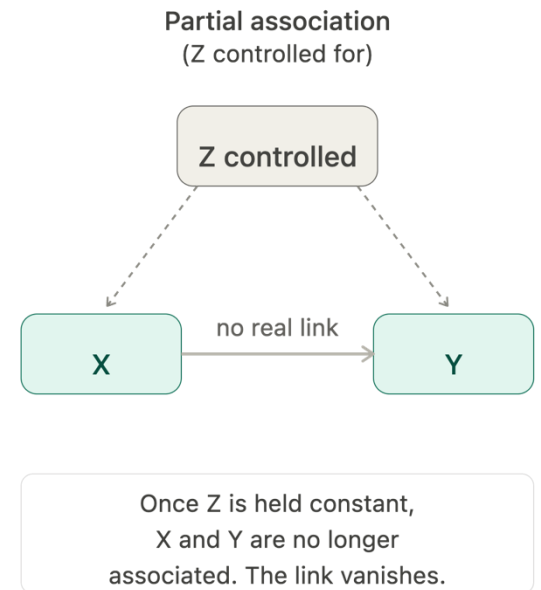
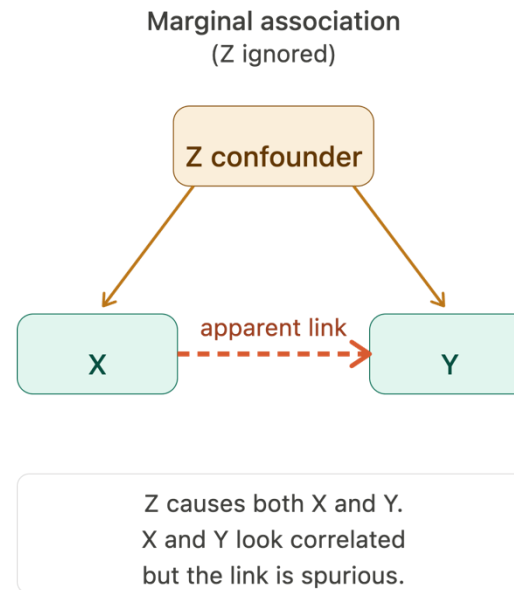
The divorce rate in Maine



Source: tylervigen.com

The idea: adjust for controls

- **Pick control variables Z**, then recompute each association between X and Y after adjusting for Z
- What remains is the **partial association**; the part of the relationship not explained by the controls
- **Intuition:** within groups that are similar on Z, do X and Y move together?
 - If yes, the link remains
 - If no, the link disappears
- How is Z chosen?
 - User and subject knowledge→ Results depend on which controls are picked



Three cases depending on variable type

- **Numeric vs numeric:**

- Marginal: Pearson's correlation r
- Conditional: Partial correlation $r_{XY.Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{1 - r_{XZ}^2}\sqrt{1 - r_{YZ}^2}}$, the correlation of residuals from X on Z and Y on Z; t-test on $n-q-2$ degrees of freedom; strength = r^2

- **Numeric vs categorical:**

- Marginal: η^2 (ANOVA)
- Conditional: partial $\eta^2_{X|Z}$ (ANCOVA), comparing the reduced model Y on Z with the full model Y on X plus Z (extra sum of squares); global F-test

- **Categorical vs categorical:**

- Marginal: a 0 to 1 measure built from a likelihood-ratio statistic $V_L = \sqrt{1 - \exp(-\frac{G^2}{n})}$ with $G^2 = 2(\ell_1 - \ell_0)$
- Conditional: $V_{L|Z}$ fits a multinomial-logit model that includes the controls, advantage over log-linear models: controls may be continuous or categorical

- **Filtering:** on effect size and on the p -value, in both the marginal and the conditional case

Networks before and after adjustment

Marginal network



Dense: many links look strong

Conditional network



Sparser: only links which survive adjustment for cofounders

- Which associations **remain**, which **disappear**?
 - Not “Are X and Y associated?” anymore, but “Are X and Y still associated once we account for age, education, region, etc.?”
- **No causal inference**: a more cautious, transparent exploration tool that shows where deeper modelling is worth investigating

Conclusion

Working today:

- **Tool 1:** a plain-language question returns the relevant variables
- **Tool 2:** strong and non-obvious associations are highlighted and ranked by interest
- **Tool 3:** only the associations that survive adjustment are kept

Further research:

- **User feedback:** interviews with journalists will help us design AssociationExplorer (Tool 2)
- The three tools are separate today but designed to be **one pipeline**
- Variable descriptions (Tool 1) could be **enriched with known associations**, giving each variable richer context for both search and surprise scoring
- Semantic surprise is currently scenario-based; adding embedding-based **similarity scores** could make it more formal and objective
- **Comparable strengths:** putting r^2 , V^2 and adjusted R^2 on the same scale is a working approximation; defining a truly comparable measure is worth further investigation
- **Survey design:** accounting for sampling weights would make the statistical significance and strength estimates more representative of the population
- **Multiple testing:** with large samples and many pairs tested, p -values become uninformative; a correction for multiple-testing would make the screening more robust

Thank you!
Questions?

Appendix: RAG vs fine-tuning

Criterion	RAG (used)	Fine-tuning (not implemented)
Principle	Model unchanged, dynamic retrieval	Partially retrain the model weights
Set-up cost	Low, no training	High: GPU, training time, expertise
Add a dataset	Immediate: embed and index	Costly: a new training cycle
Transparency	High: retrieved variables are visible	Low: knowledge diffused in weights
Hallucination	Reduced: answers from real variables	Possible: may invent plausible but fake codes
Status here	Operational on 6 datasets (7,010 vars)	Documented alternative (LoRA, Llama-3-8B)

Appendix: embeddings and retrieval

- **multilingual-e5-large:** XLM-RoBERTa encoder, 24 layers, about 560M parameters; the [CLS] vector summarises each text; contrastive training on about 1 billion multilingual pairs.
- **Prefixes:** indexed texts use passage:, user questions use query:, which aligns short questions with longer descriptions.
- **Store:** ChromaDB with cosine similarity; relevant matches usually score 0.85 to 0.93.
- **Collections:** seven, one per dataset plus a unified one over all 7,010 variables.
- **Size bias:** IPCAL is about 76% of the corpus; an early hybrid weighting was removed in favour of plain cosine.

Appendix: Part 2 measures

Numeric vs numeric (Pearson)

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad \text{strength} = r^2$$

Categorical vs categorical (Cramer's V)

$$V = \sqrt{\frac{\chi^2}{n \min(r-1, c-1)}} \quad \text{strength} = V^2$$

Numeric vs categorical (ANOVA)

$$\eta^2 = \frac{SS_B}{SS_T} \quad R_{\text{adj}}^2 = 1 - (1 - \eta^2) \frac{n-1}{n-k} \quad F = \frac{SS_B/(k-1)}{SS_W/(n-k)}$$

Journalistic interest

$$\text{interest} = \text{surprise} \times \text{reliability} \quad \text{reliability} = \frac{-\log_{10} p}{-\log_{10} p + 3}$$

Appendix: Part 3 conditional measures

Numeric vs numeric: partial correlation

$$r_{XY \cdot Z} = \text{corr}(\hat{\varepsilon}^X, \hat{\varepsilon}^Y) \quad T = r_{XY \cdot Z} \sqrt{\frac{n - q - 2}{1 - r_{XY \cdot Z}^2}} \sim t_{n - q - 2}$$

Numeric vs categorical: partial eta-squared (ANCOVA)

$$\eta_{X|Z}^2 = \frac{RSS_{\text{red}} - RSS_{\text{full}}}{RSS_{\text{red}}} \quad F = \frac{(RSS_{\text{red}} - RSS_{\text{full}})/(k - 1)}{RSS_{\text{full}}/(n - k - q)}$$

Categorical vs categorical: likelihood-ratio measure

$$G^2 = 2(\ell_1 - \ell_0) \quad V_L = \sqrt{1 - e^{-G^2/n}} \quad V_{L|Z} = \sqrt{1 - e^{-G_{|Z}^2/n}}$$

Appendix: a note on evaluation metrics

- **An early evaluation (20 annotated questions) gave** mean similarity about 0.742, MRR about 0.417, recall at 10 about 0.533.
- **These predate two changes:** removing the hybrid weighting, and extending from 4 to 6 datasets.
- **They must be recomputed on the current system**, so treat them as preliminary, not final.

Appendix: widening access with synthetic data

- **The six datasets are confidential research microdata**, not open to the public.
- **Wasserstein GANs generate a synthetic copy** that reproduces the statistical structure of the population while reducing disclosure risk (tested on the Belgian Labour Force Survey).
- **This synthetic layer can widen access** to journalists, students and citizens without exposing real records.
- **References:** Annoye, Heuchenne and co-authors, Applied Intelligence (2025), Knowledge-Based Systems, and Expert Systems with Applications.

Reference

- Soetewey, A., Heuchenne, C., Claes, A., & Descampe, A. (2026). AssociationExplorer: A user-friendly Shiny application for exploring associations and visual patterns. *SoftwareX*, 33, 102483.