

Automatiser le journalisme d'investigation local - Quels enjeux médiatiques?

Arnaud Claes – arnaud.claes@uclouvain.be

Antoine Soetewey – antoine.soetewey@uliege.be

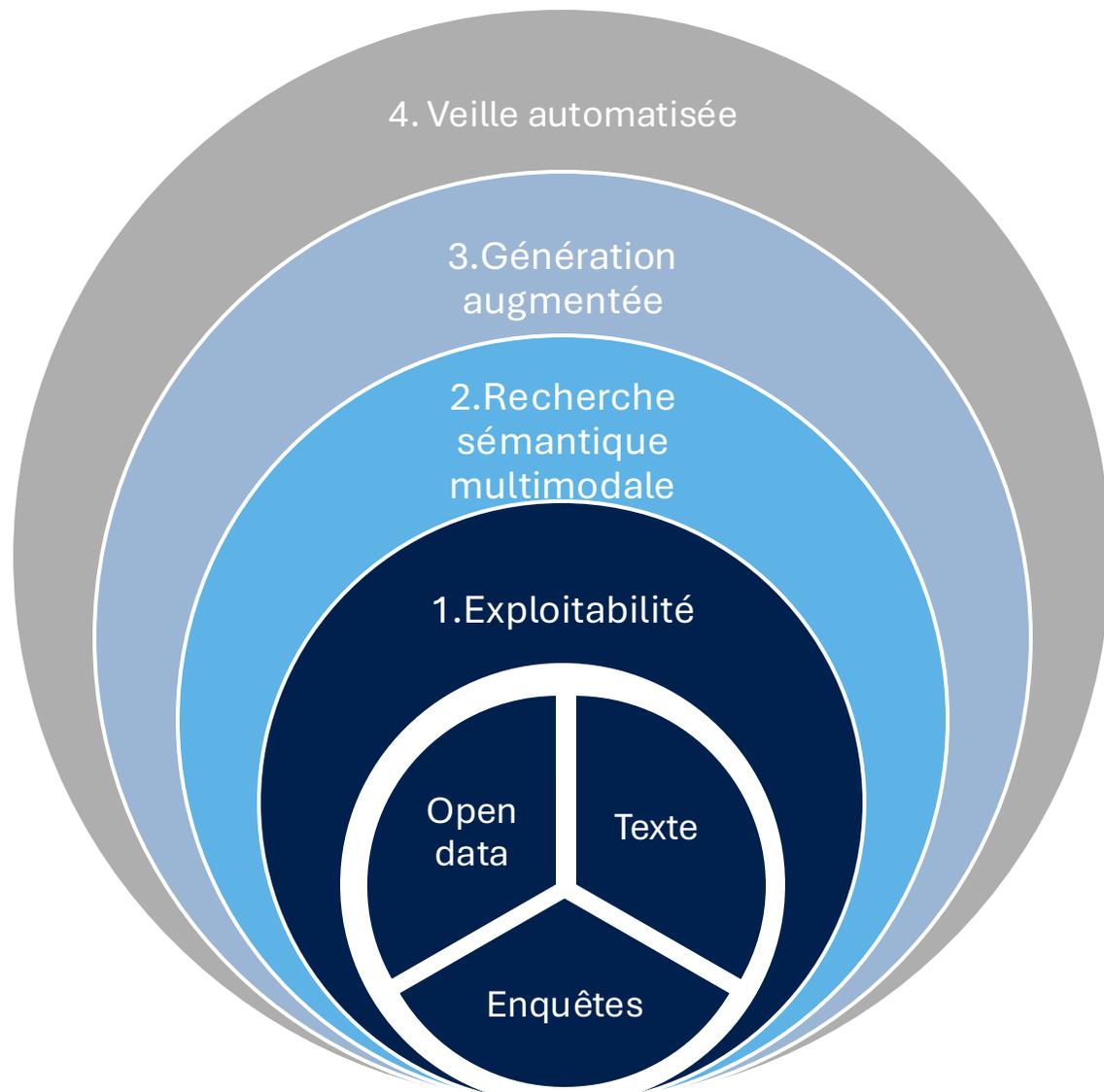
Antonin Descampe – antonin.descampe@uclouvain.be

Cédric Heuchenne – c.heuchenne@uliege.be

6 Mai 2025

Objectif:

Comment automatiser une partie du processus de traitement de l'information d'un journaliste d'investigation dans une rédaction locale?



4. Automatiser la détection de signaux faibles

3. Evaluer les apports de la génération de texte pour le travail d'investigation

2. Permettre une recherche croisée dans les deux modalités

1. Identifier et résoudre les problématiques d'exploitation de

- données textuelles
- données chiffrées publiques
- données chiffrées d'enquêtes

OUVERTURE DE LA SÉANCE

- La séance est ouverte à 14 h

M. le Président. – La séance

ABSENCES MO

M. le Président. – Ont de absence à la présente séance :

- M. Dodrimont, Mmes D raisons de santé ;
- Mme Hanus, en congé de

APPROBATION DE L'O

M. le Président. – Conformè règlement, la Conférence des l'élaboration de l'ordre du jour vous est soumis conformément à

Ce document vous a été adre

J'ai été saisi du dépôt des pro

- la proposition de résolut place une stratégie de construction en bois en MM. Spies, Fontaine, M M. Crampont et Mme Ti (2025) N° 1) ;

COMMUNICATIONS

Données statistiques publiques

1 162 jeux de données

Tri des jeux Les + populaires

Modifié Populaires A-Z

Filtres

Trouver un jeu de données.

Territoire



Région wallonne

Visualiser

Analyse	1 073
Carte	746
Vue personnalisée	294
Image	32
Calendrier	26

Modifié

2025	507
2024	171

Points verts de l'ADEPS

Chaque dimanche e balisées sont la Wallonie et

- sport marc
- activité sp

Parkings

Ce jeu de don ville de Mons en voirie.

- parking sta

Tournai -

Localisation d

Données d'enquêtes

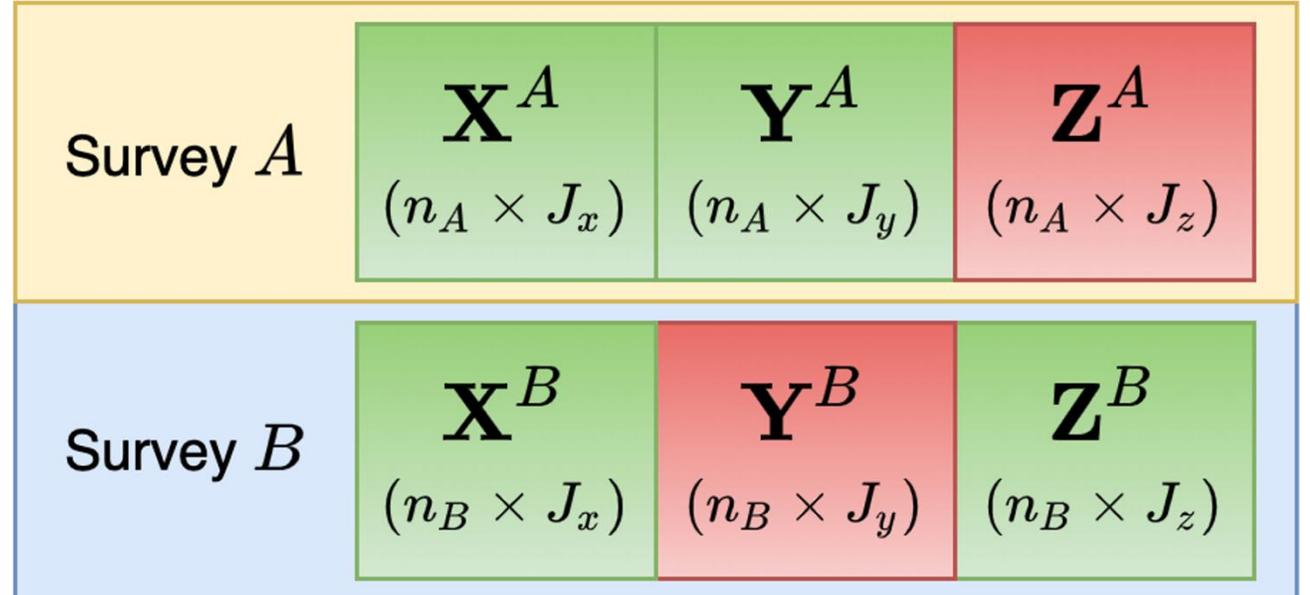


Les bases de données chiffrées d'enquêtes

- Les bases de données chiffrées d'enquêtes :
 1. Proviennent de **plusieurs** enquêtes **indépendantes**, recueillies auprès de la même population
 2. Incluent des données **confidentielles**
- Deux défis se posent :
 1. Comment combiner ces différentes bases de données ? → ***Statistical matching***
 2. Comment créer des données "fictives" pour protéger la vie privée des individus tout en permettant une analyse de qualité de ces données confidentielles ? → **Génération de données synthétiques**

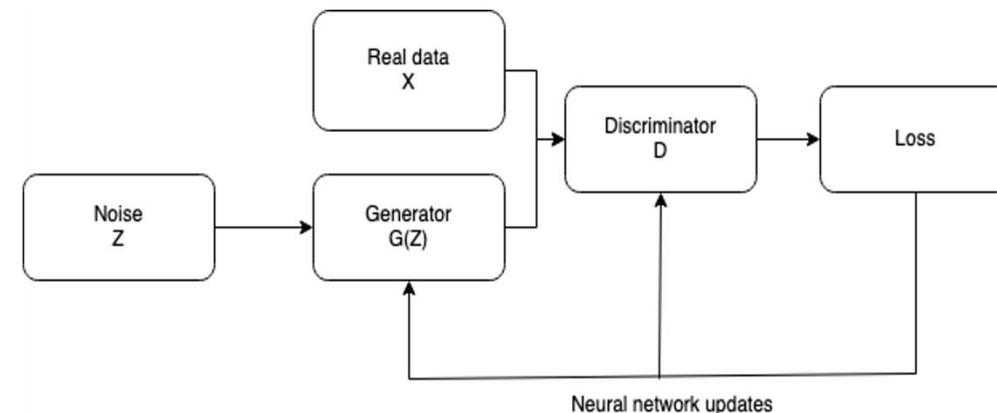
Qu'est-ce que le *statistical matching* ?

- Dans le cas d'enquêtes indépendantes menées sur une grande population :
 - Certaines variables sont communes, d'autres ne sont **pas observées conjointement**
 - La probabilité d'observer le **même individu** dans les différentes bases de données est proche de 0 → l'*appariement exact* est impossible
- En associant des individus de manière statistiquement pertinente, on peut reconstituer une base de données enrichie et plus complète
- Permet donc de **combiner/fusionner des données** provenant de différentes sources (enquêtes, bases de données, etc.), à propos de la même population

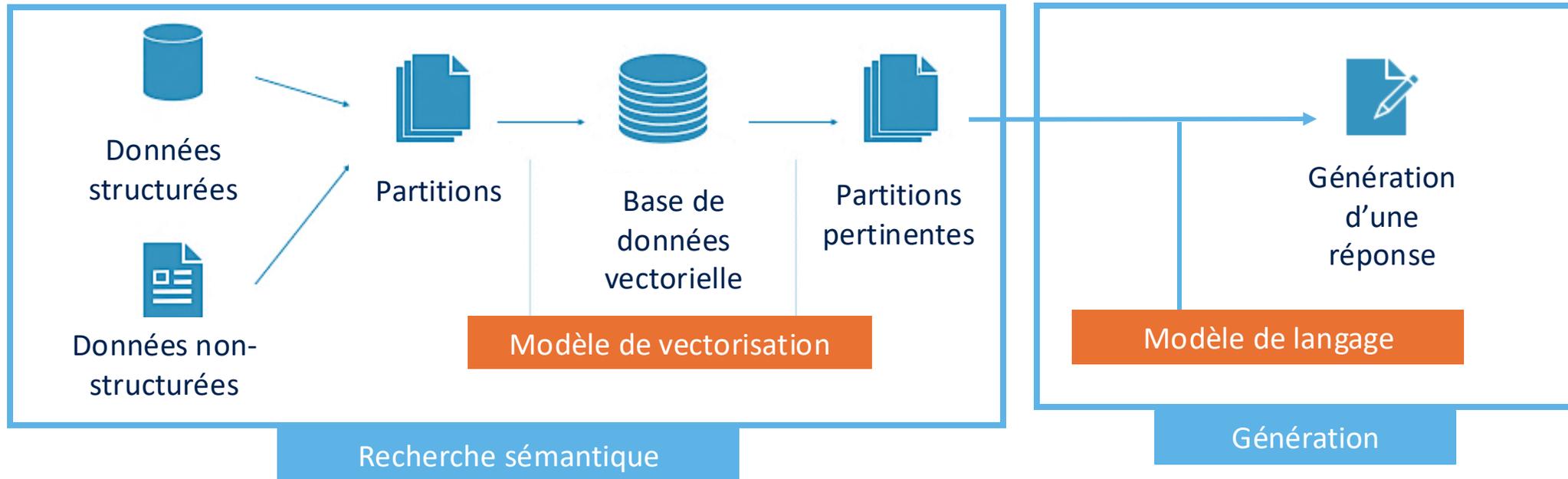


Génération de données synthétiques via un GAN

- Les GANs (*Generative Adversarial Networks*) sont un type de réseau de neurones dont le premier objectif était de **créer de fausses images** avec un haut degré de réalisme
- Au-delà de la création d'images synthétiques, un GAN permet de **générer de nouvelles données** en apprenant à imiter les vraies données
- Ils font partie d'une classe de techniques d'apprentissage automatique composée de :
 1. **Générateur (G)** : génère des nouvelles données à partir d'un bruit aléatoire en essayant de reproduire les caractéristiques des données réelles
 2. **Discriminateur (D)** : évalue l'authenticité des données, en essayant de distinguer les données réelles des données générées
 3. **Fonction de perte** : guide le générateur pour améliorer sa performance



Architecture RAG* pour le journalisme d'investigation



*Retrieval Augmented Generation

Requête: Retrouve-moi toutes les discussions qui concernent des difficultés de mobilité autour de Bruxelles

QUESTION D'ACTUALITÉ DE M. WAHL À M. HENRY, MINISTRE DU CLIMAT, DE L'ÉNERGIE ET DE LA MOBILITÉ, SUR « L'IMPLICATION DE LA POLITIQUE DE MOBILITÉ BRUXELLOISE POUR LES NAVETTEURS WALLONS »

QUESTION URGENTE DE M. WAHL À M. HENRY, MINISTRE DU CLIMAT, DE L'ÉNERGIE, DE LA MOBILITÉ ET DES INFRASTRUCTURES, SUR « L'IMPASSE DU DIALOGUE INTERRÉGIONAL DU PROJET SMARTMOVE »



Intensité du trafic routier de WalStat

Un véhicule.km = le déplacement d'un véhicule sur la distance d'un km sur le réseau routier de l'entité.

L'estimation se fait en concordance avec les chiffres que publiés par le SPFMT('Recensement général de la Circulation 2005' Brochure n° 32 et surtout son Complément Brochure n° 37). qui s'ajoutent aux Recensements quinquennaux publiés depuis 1949 selon les règles du Conseil Economique et So...

mobilité trafic routier statistiques

 Tableau

 Carte

 Analyse

 Export

 API

 Wallonie
 Iweps

Prompt : Rédige un rapport résumant les extraits suivants d'un débat parlementaire.

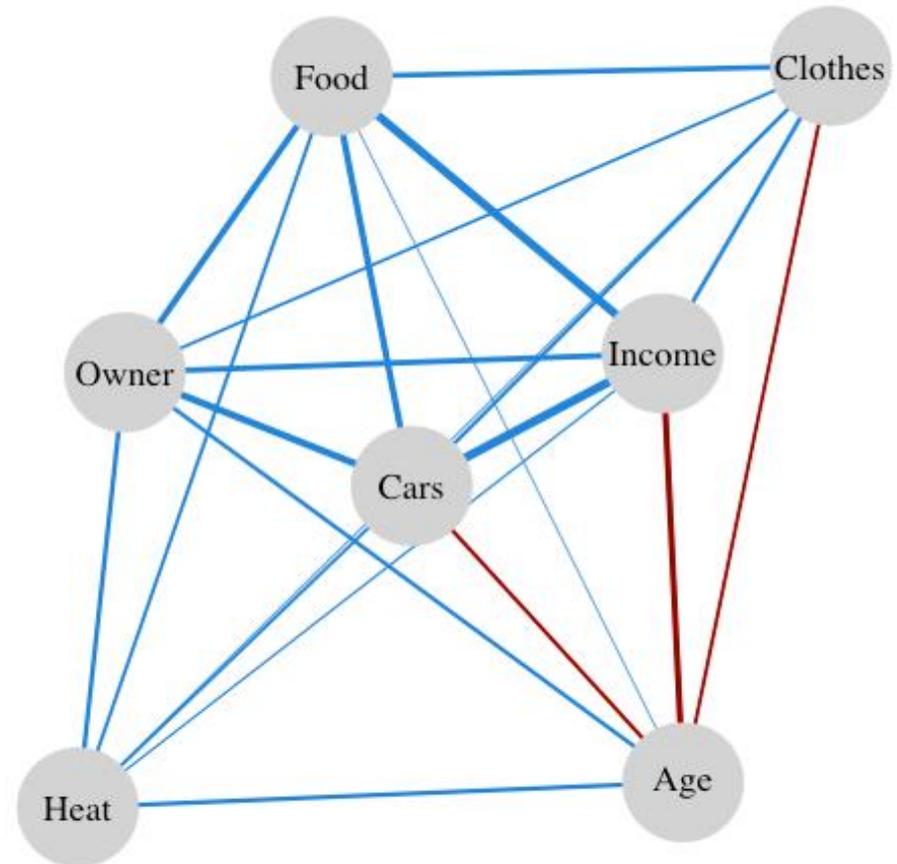
Contexte : MM. Wahl et Hardy interrogent M. Henry sur l'état des discussions interrégionales concernant le projet SmartMove, notamment après une réunion du Comité de concertation.

Points clés : M. Wahl : souligne le manque d'accord entre les régions et l'importance de la communauté métropolitaine dans la gestion de la mobilité. M. Hardy : Souligne la nécessité de trouver des solutions alternatives pour éviter les impacts négatifs sur les navetteurs wallons. M. Henry : Indique que les discussions sont en cours et que des solutions alternatives de mobilité sont explorées. Il mentionne que la Région bruxelloise est ouverte à des discussions plus larges.

Conclusion : les discussions interrégionales sur le projet SmartMove sont au point mort, mais des efforts sont en cours pour trouver des solutions alternatives. La communauté métropolitaine reste un sujet important mais pas central dans les discussions actuelles.

Hiérarchisation : quelques exemples

- Objectif : **visualiser les corrélations** entre variables pour mettre en évidence les variables les plus fortement liées, et donc potentiellement les liens les plus intéressants pour le journaliste

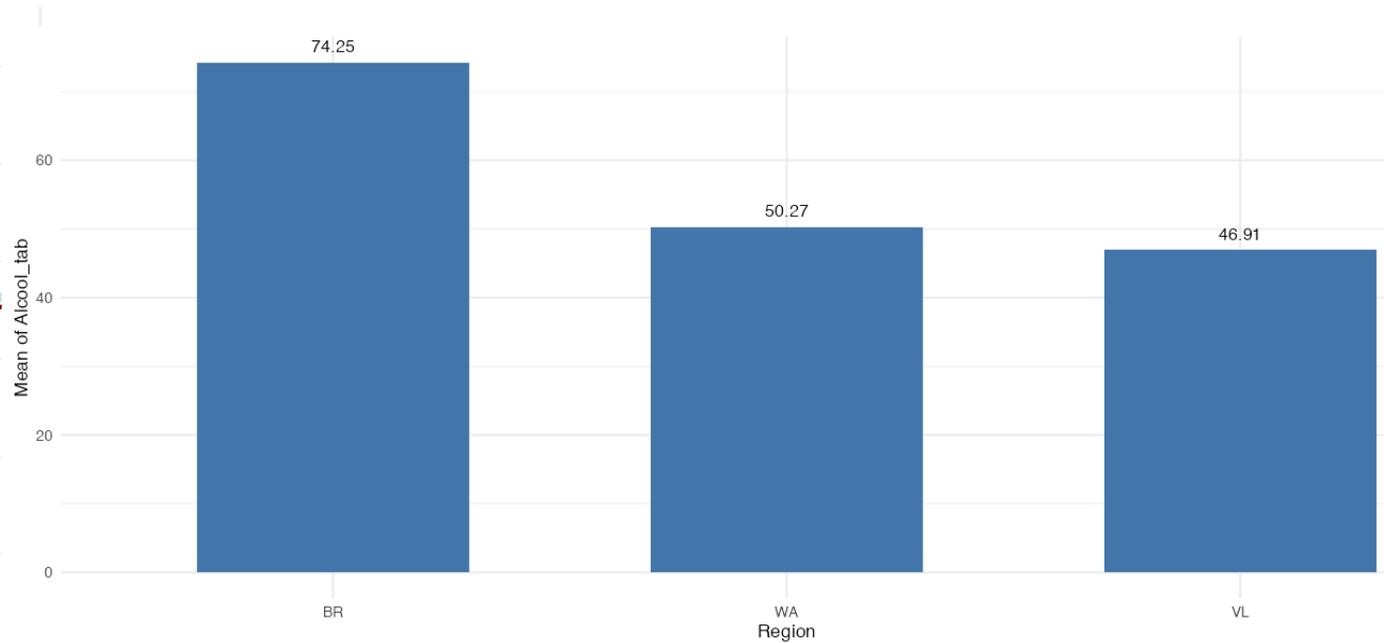
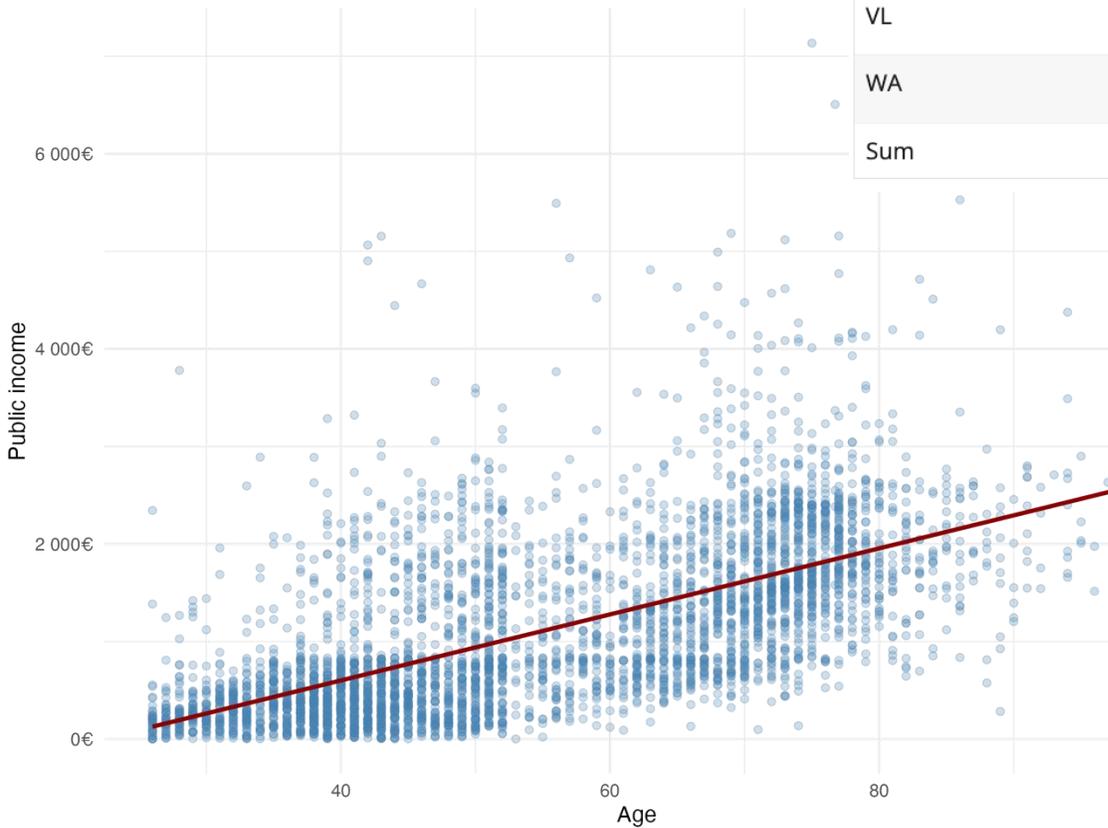


Génération : quelques exemples

- Sur base des variables les plus corrélées entre elles, **automatisation** des sorties graphiques, avec pour avantages un gain d'**autonomie** et une meilleure **accessibilité** pour un plus grand nombre de journalistes

Owner

Region	No	Yes	Sum
BR	339	250	589
VL	878	2,726	3,604
WA	580	1,363	1,943
Sum	1,797	4,339	6,136





- La génération impose une interprétation qu'il peut être difficile d'ignorer (une analyse synthétique du texte ou une visualisation de données) tout en réduisant **la « friction » de l'activité de recherche**. L'interface d'opération de ce dispositif est la variable d'ajustement de ce degré de friction.
- Au-delà des craintes d'une « atrophie cognitive » (Bainbridge, 1983), **quels effets sur les pratiques informationnelles?** Des travaux empiriques encore très limités sur l'étude des compétences cognitives mobilisées (e.g. Lee et al., 2025)
- Plutôt que de voir la recherche d'information comme un problème à résoudre, ne faut-il pas le concevoir comme **un processus créatif et itératif?** A cet égard, est-ce qu'un certain niveau de friction n'est pas souhaitable?

Merci !

Questions ?

Arnaud Claes – arnaud.claes@uclouvain.be

Antoine Soetewey – antoine.soetewey@uliege.be

Antonin Descampe – antonin.descampe@uclouvain.be

Cédric Heuchenne – c.heuchenne@uliege.be