From data to journalism

Nelly Barret

4th year PhD student
Supervised by Ioana Manolescu
Inria Saclay and Institut Polytechnique de Paris

January 31, 2024

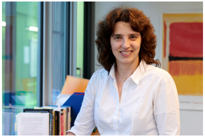
Personal (small) presentation

My background:

- CS Bachelor @ University of Lyon
- CS Master, AI track @ University of Lyon
- CS PhD student @ Inria and Ecole Polytechnique

My thesis is about **user-oriented exploration of semi-structured data**.

It is not only about me:



...and many others!

Context: data is the new gold (1/2)

Context: data is the new gold (2/2)

Our digital world comes with various contexts, needs, actors, ...

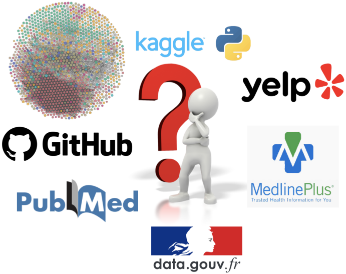
We are **overwhelmed** by (raw) data, we need to bring order

Very **large** and **heterogeneous** data:

- Tables, text, databases, ...

Detection of **entities** of interest:

- People names, places, company names, dates, ...



Data + journalism = data journalism

On one hand, we have:

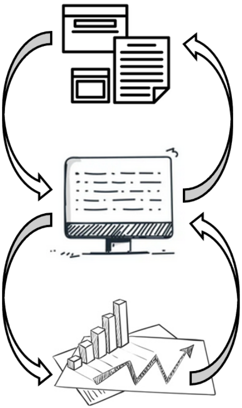
- Facts
- Data

In the middle, we have:

- Computers
- Programs

On the other hand, we have:

- Journalists
- Data investigation
- Fact-checking



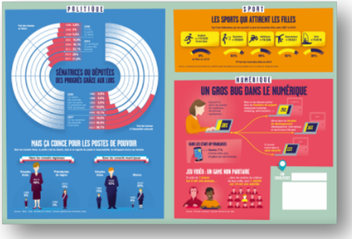
New user needs, especially in data-journalism

With **heterogeneous** data, users need:

1. A **uniform**/integrated view over the data
2. **Efficient** and **intuitive** ways to:
   - Get a global **understanding, description** of the data
   - Get interesting **entity connections**
   - **Query** and **search** for information in the system
3. Produce **insights/tangible results** to share

But:

- They have few or no CS skills
- They do not know what they are exactly looking for
- Their data may be messy/dirty



Vocabulary introduction: dataset, schema, model

Dataset

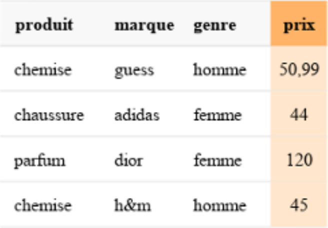
A file reporting data on a precise topic

Data model

How data is represented (table, text, database, ...)

Schema

How data objects are designed and relate



produit	marque	genre	prix
chemise	guess	homme	50,99
chaussure	adidas	femme	44
parfum	dior	femme	120
chemise	h&m	homme	45

Data heterogeneity

At the model and/or schema level

Problem statement

efficient and expressive integration of heterogeneous data

=

Provide a unified access (= put in the same "box")

Problem statement

efficient and expressive integration of heterogeneous data

=

Provide a unified access (= put in the same “box”) to a set of datasets (whatever their provenance, model, schema)

Problem statement

efficient and expressive integration of heterogeneous data

=

Provide a unified access (= put in the same "box") to a set of datasets (whatever their provenance, model, schema), sometimes very large

Problem statement

efficient and expressive integration of heterogeneous data

=

Provide a unified access (= put in the same “box”) to a set of datasets (whatever their provenance, model, schema), sometimes very large, such that this can be understood and used by human users.

Problem statement

efficient and expressive integration of heterogeneous data

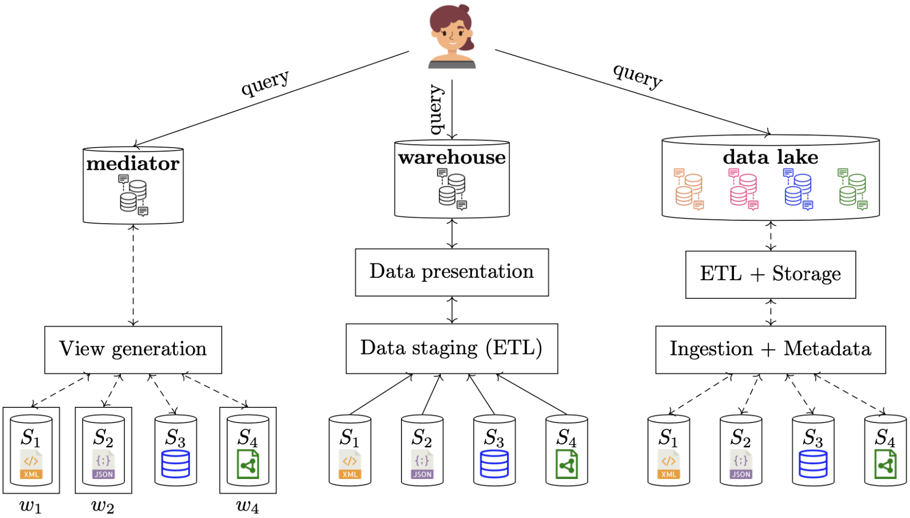
=

Provide a unified access (= put in the same "box") to a set of datasets (whatever their provenance, model, schema), sometimes very large, such that this can be understood and used by human users.

What is data integration?

A system providing a unified interface to access, process and query a set of diverse, and potentially heterogeneous, datasets

Existing architectures for data integration



Yes, but...

Data integration systems strengths are also their weaknesses:

- **Mediators** convert many sources to a single model, but...
  → Not feasible for dozens of sources
- **Warehouses** lead to a consolidated database, but...
  → Not very flexible with new data
- **Data lakes** allow many data sources to co-exist, but...
  → Rapidly become data swamps

No data integration systems fits all needs!
Data integration takes time, money and requires CS skills

Our proposals: ConnectionStudio and StatCheck

1 ConnectionStudio

- A data lake for novice users
- To load, clean, visualize and query heterogeneous data

→ With "LeMonde" data journalists

2 StatCheck

- A warehouse for centralizing statistical data
- To search for statistics and to analyse political discourses

→ With the "FranceInfo" fact-checking team "Le vrai du faux"

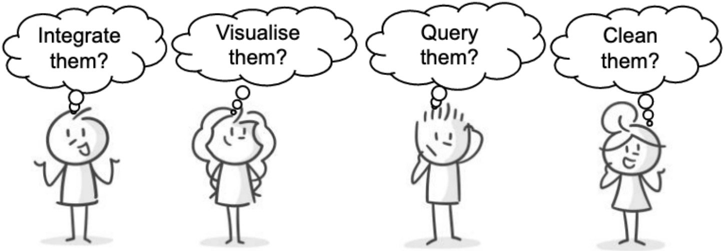
ConnectionStudio

ConnectionStudio problem statement

Question: How to work with heterogeneous (journalistic) data?

ConnectionStudio problem statement

Question: How to work with heterogeneous (journalistic) data?

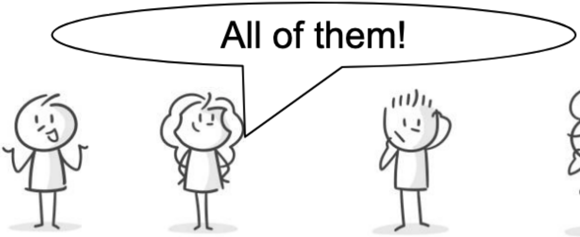


ConnectionStudio problem statement

Question: How to work with heterogeneous (journalistic) data?

ConnectionStudio problem statement

Question: How to work with heterogeneous (journalistic) data?



ConnectionStudio solution in a nutshell

Our answer: A **user-friendly** **data lake** for:

ConnectionStudio solution in a nutshell

Our answer: A **user-friendly** **data lake** for:

1. **Loading heterogeneous datasets**
   → uniform/integrated view over the data

ConnectionStudio solution in a nutshell

Our answer: A **user-friendly data lake** for:

1. **Loading heterogeneous datasets**

   → uniform/integrated view over the data

2. **Statistics and data summaries**

   → global description of the data

ConnectionStudio solution in a nutshell

Our answer: A **user-friendly data lake** for:

1. **Loading heterogeneous datasets**
   → uniform/integrated view over the data
2. **Statistics and data summaries**
   → global description of the data
3. **Keyword search**
   → query and search for information

ConnectionStudio solution in a nutshell

Our answer: A **user-friendly** **data lake** for:

1. **Loading heterogeneous datasets**
   → uniform/integrated view over the data
2. **Statistics and data summaries**
   → global description of the data
3. **Keyword search**
   → query and search for information
4. **Data paths**
   → interesting entity connections

ConnectionStudio solution in a nutshell

Our answer: A **user-friendly** **data lake** for:

1. **Loading heterogeneous datasets**
   → uniform/integrated view over the data
2. **Statistics and data summaries**
   → global description of the data
3. **Keyword search**
   → query and search for information
4. **Data paths**
   → interesting entity connections
5. **Querying the data lake**
   → query and search for information

ConnectionStudio solution in a nutshell

Our answer: A **user-friendly data lake** for:

1 **Loading heterogeneous datasets**

→ uniform/integrated view over the data

2 **Statistics and data summaries**

→ global description of the data

3 **Keyword search**

→ query and search for information

4 **Data paths**

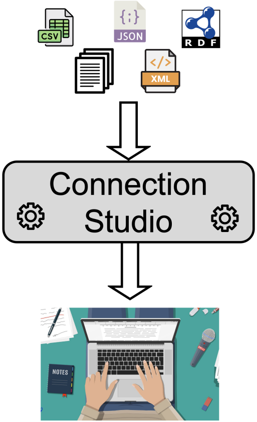
→ interesting entity connections

5 **Querying the data lake**

→ query and search for information

6 **Tabular-looking results**

→ produce insights/tangible results to share



Connection Studio

① Unified data view: a graph

Question: How to commonly represent heterogeneous datasets?

① Unified data view: a graph

Question: How to commonly represent heterogeneous datasets?

Our answer: A graph

① Unified data view: a graph

Question: How to commonly represent heterogeneous datasets?

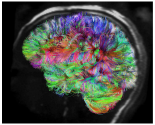
Our answer: A graph

Wait... What is a graph?

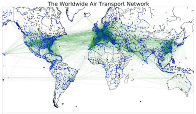
The **graph** paradigm describes:

- Objects (nodes)
- Connected by links (edges)

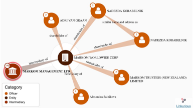
High flexibility → largely used



Brain neurons



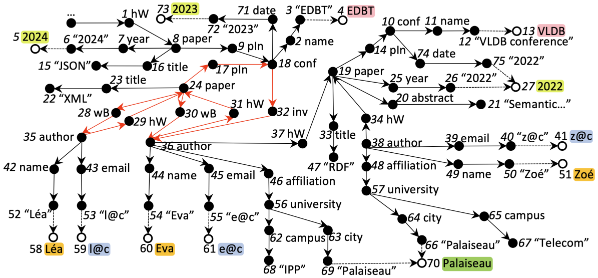
International flights



Panama papers

① Unified data view: a graph

- Ingest any dataset into a **directed graph** (•, →)
- Extract **named entities**, NEs, from the graph values (○, - - →):
  - Temporal: date, time reference
  - Web: URI, email address, hashtag, Twitter citation
  - Complex entities: People, Place, Organization
  - Used pre-trained language models; more recently ChatGPT



① Unified data view: a graph

Question: What if the data graph is huge?

① Unified data view: a graph

Question: What if the data graph is huge?

Our answer: Build its compact representation (summary)

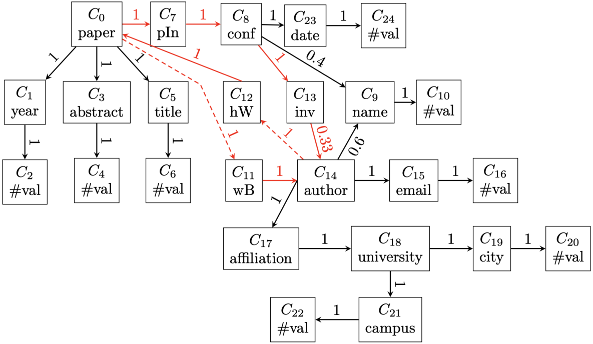
① Unified data view: a graph

Question: What if the data graph is huge?

Our answer: Build its compact representation (summary)

→ We build a **summary graph**, with small information loss

→ **Efficient** algorithms and applications



② Statistics and data summaries

Question: How to get a quick overview of the datasets?

② Statistics and data summaries

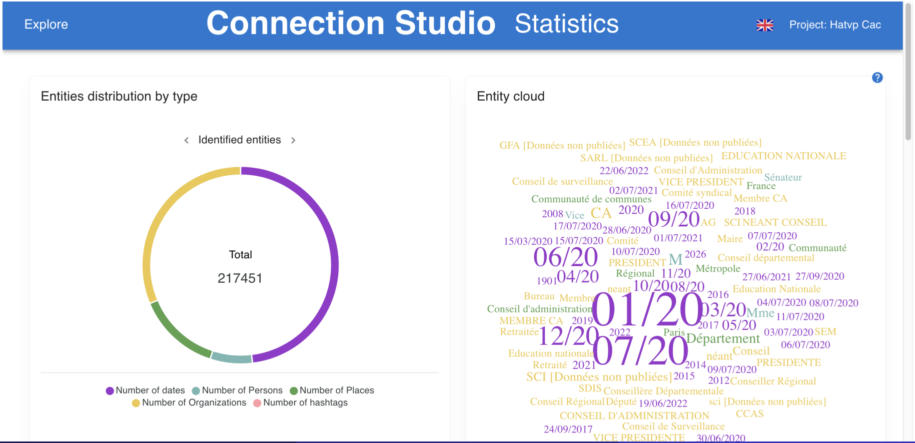
Question: How to get a quick overview of the datasets?

Our answer: Show Named Entities stats in charts, tables and tag clouds

② Statistics and data summaries

Question: How to get a quick overview of the datasets?

Our answer: Show Named Entities stats in charts, tables and tag clouds



② Statistics and data summaries

Question: How to get a quick overview of the datasets?

② Statistics and data summaries

Question: How to get a quick overview of the datasets?

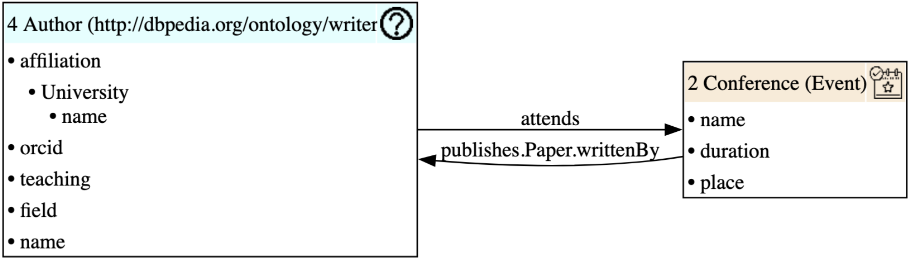
Our answer: Draw an Entity-Relationship diagram representing "important" nodes

② Statistics and data summaries

Question: How to get a quick overview of the datasets?

Our answer: Draw an Entity-Relationship diagram representing "important" nodes

Abstraction of conferences
located at: file:/Users/nelly/Documents/boulot/theseNelly/abstraction-work/../abstraction-data/rdf/testConferences.nt
with 103 normalized nodes, 29 collections, (PR, FLAC), 2/5 main collections, data coverage is 100.0%



4 Author (http://dbpedia.org/ontology/writer

- affiliation
  - University
    - name
- orcid
- teaching
- field
- name

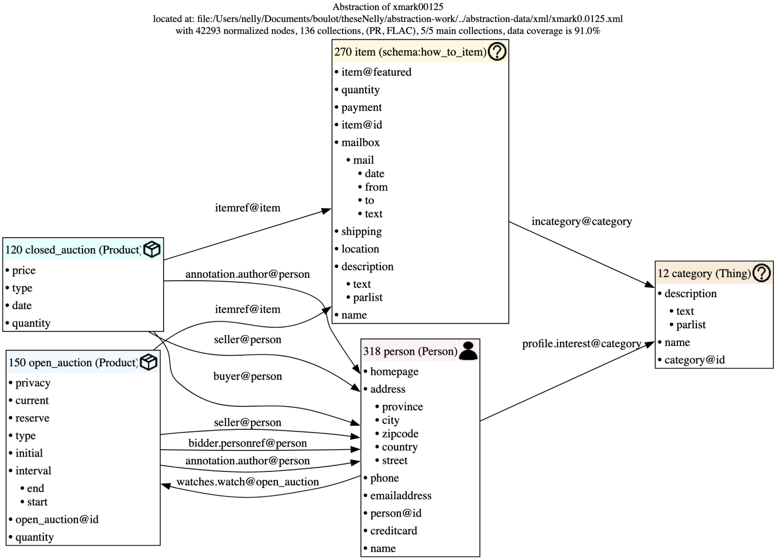
2 Conference (Event)

- name
- duration
- place

attends

publishes.Paper.writtenBy

2 Statistics and data summaries



Abstraction of xmark00125
located at: file:/Users/nelly/Documents/boulot/theseNelly/abstraction-work/../abstraction-data/xml/xmark0.0125.xml
with 42293 normalized nodes, 136 collections, (PR, FLAC), 5/5 main collections, data coverage is 91.0%

270 item (schema:how_to_item)
- item@featured
- quantity
- payment
- item@id
- mailbox
  - mail
    - date
    - from
    - to
    - text
- shipping
- location
- description
  - text
  - parlist
- name

120 closed_auction (Product)
- price
- type
- date
- quantity

150 open_auction (Product)
- privacy
- current
- reserve
- type
- initial
- interval
  - end
  - start
- open_auction@id
- quantity

318 person (Person)
- homepage
- address
  - province
  - city
  - zipcode
  - country
  - street
- phone
- emailaddress
- person@id
- creditcard
- name

12 category (Thing)
- description
  - text
  - parlist
- name
- category@id

5 Querying the data lake

Question: How to "discuss"/ask something to the data lake?

5 Querying the data lake

Question: How to “discuss”/ask something to the data lake?

Language to “discuss”/ask something to a database: (mostly) SQL

5 Querying the data lake

Question: How to “discuss”/ask something to the data lake?

Language to “discuss”/ask something to a database: (mostly) SQL

```
SELECT n1.label, n2.label, n3.label
FROM nodes n1, edges e1,
     nodes n2, edges e2, nodes n3
WHERE e1.source=n1.id
      AND e1.target=n2.id
      AND e2.source=n2.id
      AND e2.target=n3.id
LIMIT 3
```

n1.label	n2.label	n3.label
author	name	“Léa”
university	campus	“IPP”
paper	writtenBy	author

⑤ Querying the data lake

Question: How to “discuss”/ask something to the data lake?

Language to “discuss”/ask something to a database: (mostly) SQL

```
SELECT n1.label, n2.label, n3.label
FROM nodes n1, edges e1,
     nodes n2, edges e2, nodes n3
WHERE e1.source=n1.id
      AND e1.target=n2.id
      AND e2.source=n2.id
      AND e2.target=n3.id
LIMIT 3
```

n1.label	n2.label	n3.label
author	name	“Léa”
university	campus	“IPP”
paper	writtenBy	author

Pros and cons of SQL:

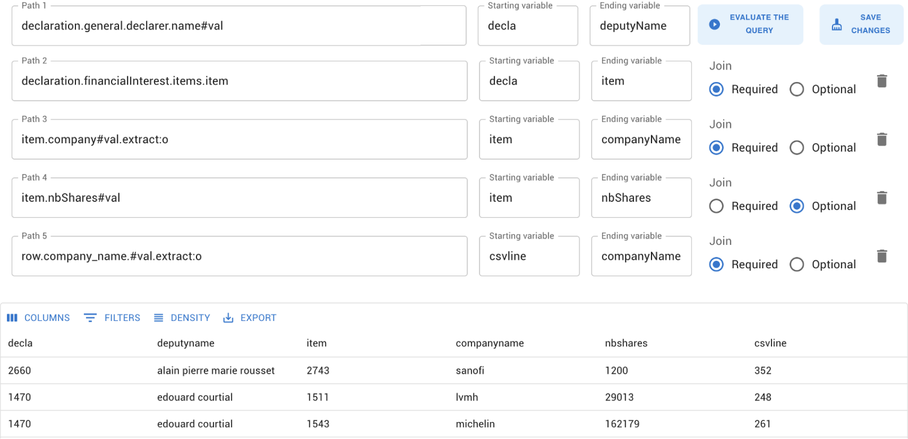
- Needs to be learned (→ time, skills)
- SQL queries are highly performant, scalable, optimizable
- Results shown as tables

⑤ Querying the data lake

Our answer: Get rid of the “SQL writing part”, keep tables as output

5 Querying the data lake

Our answer: Get rid of the “SQL writing part”, keep tables as output



Path 1: declaration.general.declarer.name#val — Starting variable: decla — Ending variable: deputyName

Path 2: declaration.financialInterest.items.item — Starting variable: decla — Ending variable: item — Join: ● Required ○ Optional

Path 3: item.company#val.extract:o — Starting variable: item — Ending variable: companyName — Join: ● Required ○ Optional

Path 4: item.nbShares#val — Starting variable: item — Ending variable: nbShares — Join: ○ Required ● Optional

Path 5: row.company_name.#val.extract:o — Starting variable: csvline — Ending variable: companyName — Join: ● Required ○ Optional

decla	deputyname	item	companyname	nbshares	csvline
2660	alain pierre marie rousset	2743	sanofi	1200	352
1470	edouard courtial	1511	lvmh	29013	248
1470	edouard courtial	1543	michelin	162179	261

5 Querying the data lake

Our answer: Get rid of the “SQL writing part”, keep tables as output

⑤ Querying the data lake

Our answer: Get rid of the “SQL writing part”, keep tables as output

1. Enumerate a set of paths in the summary graph:
   - $p = \{n_0, e_0, n_1, e_1, ..., e_i, n_{i+1}\}$
2. Each selected path p is associated to:
   - A source variable s (the first element in p)
   - A target variable t (the last element in p)
3. Select join predicates (`LEFT JOIN` or `INNER JOIN`)
4. Conversion to a SQL query:
   - Each path leads to a SQL query, reusing s and t
   - Path SQL queries are joined using join predicates

$$p_0 = \{\overbrace{declaration}^{s}, \textvisiblespace, general, \textvisiblespace, declarer, \textvisiblespace, name, \textvisiblespace, \overbrace{\#val}^{t}\}$$

$$\bowtie$$

$$p_1 = \{\underbrace{declaration}_{s}, \textvisiblespace, financialInterest, \textvisiblespace, items, \textvisiblespace, \underbrace{item}_{t}\}$$

6 Concrete results

Question: How to share results found/created in the data lake?

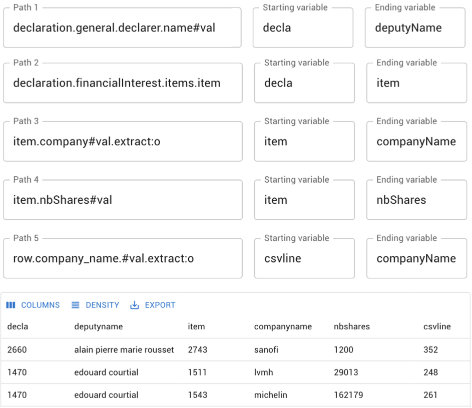
6 Concrete results

Question: How to share results found/created in the data lake?

Our answer: They can be exported (statistics, tables, queries, ...)

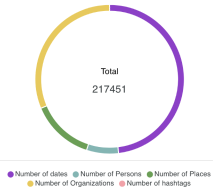
Question: How to share results found/created in the data lake?

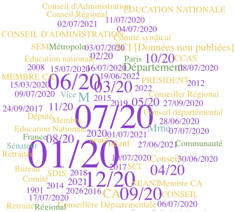
Our answer: They can be exported (statistics, tables, queries, ...)



Path	Starting variable	Ending variable
Path 1: declaration.general.declarer.name#val	decla	deputyName
Path 2: declaration.financialInterest.items.item	decla	item
Path 3: item.company#val.extract:o	item	companyName
Path 4: item.nbShares#val	item	nbShares
Path 5: row.company_name.#val.extract:o	csvline	companyName

decla	deputyname	item	companyname	nbshares	csvline
2660	alain pierre marie rousset	2743	sanofi	1200	352
1470	edouard courtial	1511	lvmh	29013	248
1470	edouard courtial	1543	michelin	162179	261





StatCheck

StatCheck problem statement

Question: How to ease/automate part of the fact-checking process?

StatCheck problem statement

Question: How to ease/automate part of the fact-checking process?

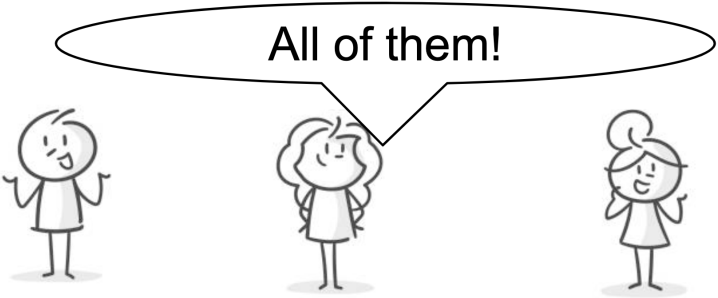
Search in trusted sources?
Analyse discourses in the political sphere?
Cross-check statistics?

ConnectionStudio problem statement

Question: How to ease/automate part of the fact-checking process?

ConnectionStudio problem statement

Question: How to ease/automate part of the fact-checking process?



StatCheck in a nutshell

Our answer: A **user-friendly** **warehouse** for:

StatCheck in a nutshell

Our answer: A **user-friendly** **warehouse** for:

1. Consolidating French statistical data
   → a unique repository for French statistics

StatCheck in a nutshell

Our answer: A **user-friendly** **warehouse** for:

1. Consolidating French statistical data

   → a unique repository for French statistics

2. Statistical search engine

   → search for statistical information

StatCheck in a nutshell

Our answer: A **user-friendly warehouse** for:

1. Consolidating French statistical data
   → a unique repository for French statistics

2. Statistical search engine
   → search for statistical information

3. Automated text analysis 1/2
   → detection of statistical claims

StatCheck in a nutshell

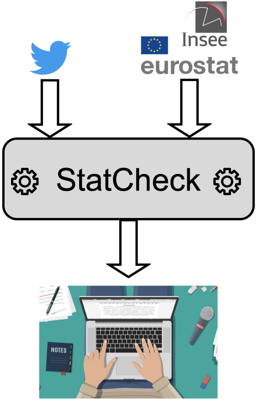
Our answer: A **user-friendly** **warehouse** for:

1. Consolidating French statistical data
   → a unique repository for French statistics

2. Statistical search engine
   → search for statistical information

3. Automated text analysis 1/2
   → detection of statistical claims

4. Automated text analysis 2/2
   → recognise persuasion techniques in political discourses



① Consolidated data for French statistics

Question: how to create consolidated data for French statistics?

1 Consolidated data for French statistics

Question: how to create consolidated data for French statistics?

Our answer: crawling of reference/trusted websites for French statistics (Open Data):

1 Consolidated data for French statistics

Question: how to create consolidated data for French statistics?

Our answer: crawling of reference/trusted websites for French statistics (Open Data):

1. **INSEE**: Institut National de la Statistique et des Études Économiques
2. **EuroStat**: European statistical database

1 Consolidated data for French statistics

Question: how to create consolidated data for French statistics?

Our answer: crawling of reference/trusted websites for French statistics (Open Data):

1. **INSEE**: Institut National de la Statistique et des Études Économiques
2. **EuroStat**: European statistical database

Knowing that:

- Data is in different models (tables, text, ...)
- Data is huge
- ... and many other concerns that we will not cover today

① Consolidated data for French statistics

Initial approach: convert all data into a graph

Yes, but:

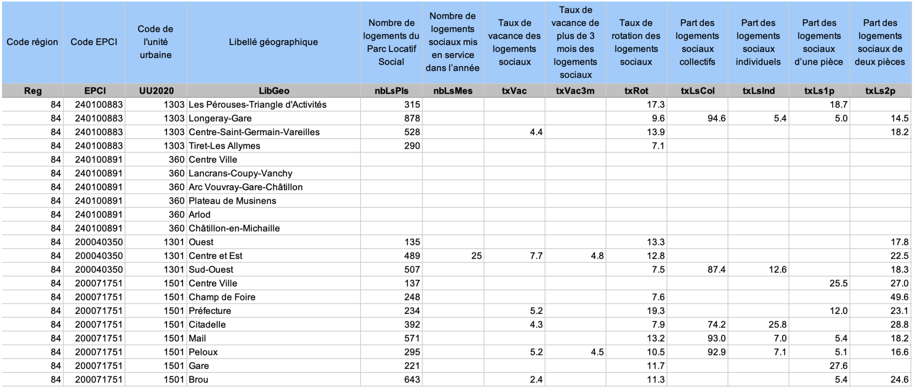
- High cost of storage (graph size: 3Tb)
- Searching the graph was expensive (1/3 of queries were very long)

① Consolidated data for French statistics

Initial approach: convert all data into a graph

Yes, but:

- High cost of storage (graph size: 3Tb)
- Searching the graph was expensive (1/3 of queries were very long)

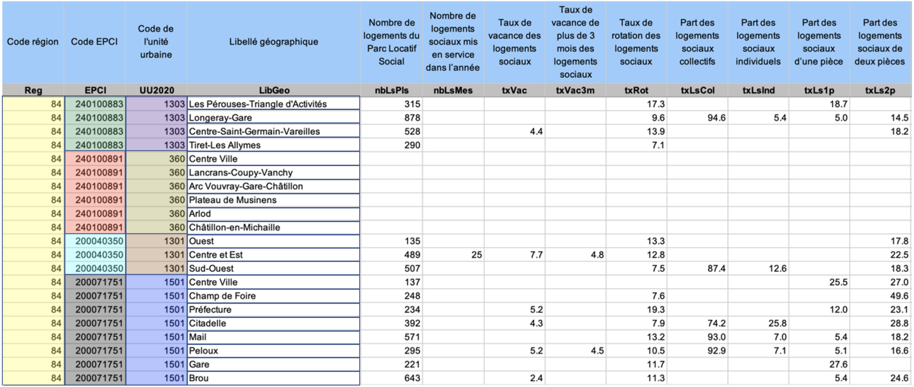


Code région	Code EPCI	Code de l'unité urbaine	Libellé géographique	Nombre de logements du Parc Locatif Social	Nombre de logements sociaux mis en service dans l'année	Taux de vacance des logements sociaux	Taux de vacance de plus de 3 mois des logements sociaux	Taux de rotation des logements sociaux	Part des logements sociaux collectifs	Part des logements sociaux individuels	Part des logements sociaux d'une pièce	Part des logements sociaux de deux pièces
Reg	**EPCI**	**UU2020**	**LibGeo**	**nbLsPls**	**nbLsMes**	**txVac**	**txVac3m**	**txRot**	**txLsCol**	**txLsInd**	**txLs1p**	**txLs2p**
84	240100883	1303	Les Pérouses-Triangle d'Activités	315				17.3			18.7	
84	240100883	1303	Longeray-Gare	878				9.6	94.6	5.4	5.0	14.5
84	240100883	1303	Centre-Saint-Germain-Vareilles	528		4.4		13.9				18.2
84	240100883	1303	Tiret-Les Allymes	290				7.1				
84	240100891	360	Centre Ville									
84	240100891	360	Lancrans-Coupy-Vanchy									
84	240100891	360	Arc Vouvray-Gare-Châtillon									
84	240100891	360	Plateau de Musinens									
84	240100891	360	Arlod									
84	240100891	360	Châtillon-en-Michaille									
84	200040350	1301	Ouest	135				13.3				17.8
84	200040350	1301	Centre et Est	489	25	7.7	4.8	12.8				22.5
84	200040350	1301	Sud-Ouest	507				7.5	87.4	12.6		18.3
84	200071751	1501	Centre Ville	137							25.5	27.0
84	200071751	1501	Champ de Foire	248				7.6				49.6
84	200071751	1501	Préfecture	234		5.2		19.3			12.0	23.1
84	200071751	1501	Citadelle	392		4.3		7.9	74.2	25.8		28.8
84	200071751	1501	Mail	571				13.2	93.0	7.0	5.4	18.2
84	200071751	1501	Peloux	295		5.2	4.5	10.5	92.9	7.1	5.1	16.6
84	200071751	1501	Gare	221				11.7			27.6	
84	200071751	1501	Brou	643		2.4		11.3			5.4	24.6

① Consolidated data for French statistics

Our answer: convert data tables as “areas”

Our answer: convert data tables as “areas”



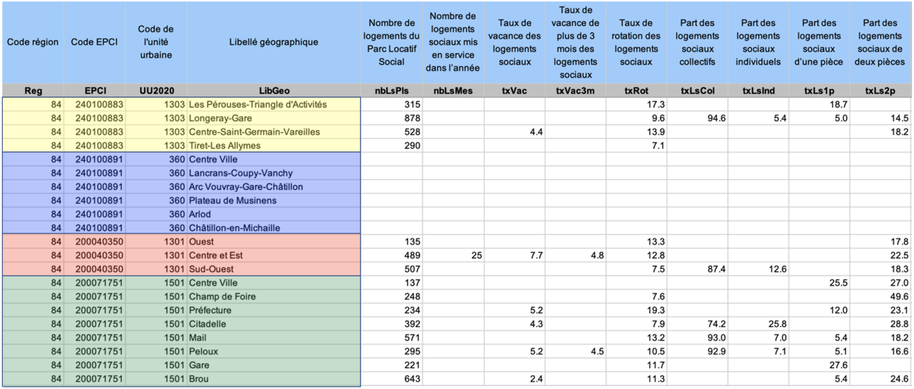
Code région	Code EPCI	Code de l'unité urbaine	Libellé géographique	Nombre de logements du Parc Locatif Social	Nombre de logements sociaux mis en service dans l'année	Taux de vacance des logements sociaux	Taux de vacance de plus de 3 mois des logements sociaux	Taux de rotation des logements sociaux	Part des logements sociaux collectifs	Part des logements sociaux individuels	Part des logements sociaux d'une pièce	Part des logements sociaux de deux pièces
Reg	**EPCI**	**UU2020**	**LibGeo**	**nbLsPls**	**nbLsMes**	**txVac**	**txVac3m**	**txRot**	**txLsCol**	**txLsInd**	**txLs1p**	**txLs2p**
84	240100883	1303	Les Pérouses-Triangle d'Activités	315				17.3			18.7	
84	240100883	1303	Longeray-Gare	878				9.6	94.6	5.4	5.0	14.5
84	240100883	1303	Centre-Saint-Germain-Vareilles	528		4.4		13.9				18.2
84	240100883	1303	Tiret-Les Allymes	290				7.1				
84	240100891	360	Centre Ville									
84	240100891	360	Lancrans-Coupy-Vanchy									
84	240100891	360	Arc Vouvray-Gare-Châtillon									
84	240100891	360	Plateau de Musinens									
84	240100891	360	Arlod									
84	240100891	360	Châtillon-en-Michaille									
84	200040350	1301	Ouest	135				13.3				17.8
84	200040350	1301	Centre et Est	489	25	7.7	4.8	12.8				22.5
84	200040350	1301	Sud-Ouest	507				7.5	87.4	12.6		18.3
84	200071751	1501	Centre Ville	137							25.5	27.0
84	200071751	1501	Champ de Foire	248				7.6				49.6
84	200071751	1501	Préfecture	234		5.2		19.3			12.0	23.1
84	200071751	1501	Citadelle	392		4.3		7.9	74.2	25.8		28.8
84	200071751	1501	Mail	571				13.2	93.0	7.0	5.4	18.2
84	200071751	1501	Peloux	295		5.2	4.5	10.5	92.9	7.1	5.1	16.6
84	200071751	1501	Gare	221				11.7			27.6	
84	200071751	1501	Brou	643		2.4		11.3			5.4	24.6

① Consolidated data for French statistics

Our answer: convert data tables as “areas”

① Consolidated data for French statistics

Our answer: convert data tables as “areas”



Code région	Code EPCI	Code de l'unité urbaine	Libellé géographique	Nombre de logements du Parc Locatif Social	Nombre de logements sociaux mis en service dans l'année	Taux de vacance des logements sociaux	Taux de vacance de plus de 3 mois des logements sociaux	Taux de rotation des logements sociaux	Part des logements sociaux collectifs	Part des logements sociaux individuels	Part des logements sociaux d'une pièce	Part des logements sociaux de deux pièces
Reg	EPCI	UU2020	LibGeo	nbLsPls	nbLsMes	txVac	txVac3m	txRot	txLsCol	txLsInd	txLs1p	txLs2p
84	240100883	1303	Les Pérouses-Triangle d'Activités	315				17.3			18.7	
84	240100883	1303	Longeray-Gare	878				9.6	94.6	5.4	5.0	14.5
84	240100883	1303	Centre-Saint-Germain-Vareilles	528		4.4		13.9				18.2
84	240100883	1303	Tiret-Les Allymes	290				7.1				
84	240100891	360	Centre Ville									
84	240100891	360	Lancrans-Coupy-Vanchy									
84	240100891	360	Arc Vouvray-Gare-Châtillon									
84	240100891	360	Plateau de Musinens									
84	240100891	360	Arlod									
84	240100891	360	Châtillon-en-Michaille									
84	200040350	1301	Ouest	135				13.3				17.8
84	200040350	1301	Centre et Est	489	25	7.7	4.8	12.8				22.5
84	200040350	1301	Sud-Ouest	507				7.5	87.4	12.6		18.3
84	200071751	1501	Centre Ville	137							25.5	27.0
84	200071751	1501	Champ de Foire	248				7.6				49.6
84	200071751	1501	Préfecture	234		5.2		19.3			12.0	23.1
84	200071751	1501	Citadelle	392		4.3		7.9	74.2	25.8		28.8
84	200071751	1501	Mail	571				13.2	93.0	7.0	5.4	18.2
84	200071751	1501	Peloux	295		5.2	4.5	10.5	92.9	7.1	5.1	16.6
84	200071751	1501	Gare	221				11.7			27.6	
84	200071751	1501	Brou	643		2.4		11.3			5.4	24.6

① Consolidated data for French statistics

Question: How much more efficient is the novel approach?

① Consolidated data for French statistics

Question: How much more efficient is the novel approach?

Our answer:

1 Consolidated data for French statistics

Question: How much more efficient is the novel approach?

Our answer:

Type	**INSEE**	**EuroStat**	**Total**
Files	96 207	7 094	103 301
Tables	112 966	7 003	119 969
Areas	1 286 603	12 179 533	**13 488 136**
Graph (Mb)	1 864 766	120 425	-
Areas (Mb)	577	8 055	-
Compression	× **3 266**	× **14**	-

① Consolidated data for French statistics

Question: How much more efficient is the novel approach?

Our answer:

Type	**INSEE**	**EuroStat**	**Total**
Files	96 207	7 094	103 301
Tables	112 966	7 003	119 969
Areas	1 286 603	12 179 533	**13 488 136**
Graph (Mb)	1 864 766	120 425	-
Areas (Mb)	577	8 055	-
Compression	× **3 266**	× **14**	-

Lesson learned: the storage should be chosen based on the data/usage

② Searching for statistics in the warehouse

Question: How to retrieve information from textual questions?

② Searching for statistics in the warehouse

Question: How to retrieve information from textual questions?

Our answer: Given a query Q composed of n keywords k_n:

② Searching for statistics in the warehouse

$\boxed{\textbf{Question:}}$ How to retrieve information from textual questions?

$\boxed{\textbf{Our answer:}}$ Given a query Q composed of n keywords k_n:

$$Q = \underbrace{\text{Taux}}_{k_1}\ \text{de}\ \underbrace{\text{chômage}}_{k_2}\ \underbrace{\text{2022}}_{k_3}$$

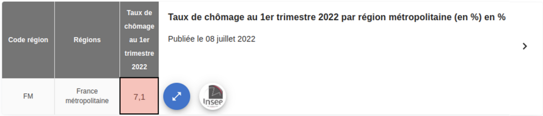
2 Searching for statistics in the warehouse

Question: How to retrieve information from textual questions?

Our answer: Given a query Q composed of n keywords k_n:

$$Q = \underbrace{\text{Taux}}_{k_1}\ \text{de}\ \underbrace{\text{chômage}}_{k_2}\ \underbrace{\text{2022}}_{k_3}$$

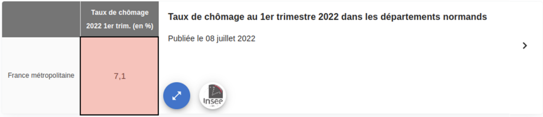
Find the 20 most relevant tables, and possibly the value:



Code région	Régions	Taux de chômage au 1er trimestre 2022
FM	France métropolitaine	7,1

Taux de chômage au 1er trimestre 2022 par région métropolitaine (en %) en %

Publiée le 08 juillet 2022



	Taux de chômage 2022 1er trim. (en %)
France métropolitaine	7,1

Taux de chômage au 1er trimestre 2022 dans les départements normands

Publiée le 08 juillet 2022

3 Automated text analysis

Question: Can we learn things from political discourses?

3 Automated text analysis

Question: Can we learn things from political discourses?

Our answer: Yes, e.g., in tweets

③ Automated text analysis

Question: Can we learn things from political discourses?

Our answer: Yes, e.g., in tweets

- Recognize/extract statistical claims
- Identify well-known persuasion techniques

Number of followed politicians	63
Number of gathered tweets	77 081
Claims detected (since 01/2022)	61 207

③ Statistical claim detection in political discourses

Question: How to identify statistics used in political discourses?

③ Statistical claim detection in political discourses

Question: How to identify statistics used in political discourses?

Our answer: Use a Machine Learning model trained on political debates (from 1950 to 2024)

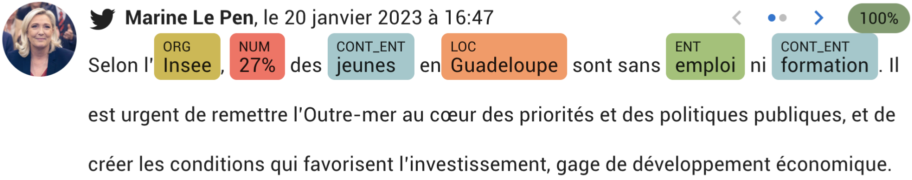
3 Statistical claim detection in political discourses

Question: How to identify statistics used in political discourses?

Our answer: Use a Machine Learning model trained on political debates (from 1950 to 2024)

- statistical claim = a statistical entity + a value + a date

$$\underbrace{\text{Taux de chômage en Guadeloupe}}_{\text{statistical entity}} + \underbrace{27\%}_{\text{value}} + \underbrace{2023}_{\text{date}}$$


Marine Le Pen, le 20 janvier 2023 à 16:47

Selon l' ORG Insee , NUM 27% des CONT_ENT jeunes en LOC Guadeloupe sont sans ENT emploi ni CONT_ENT formation . Il est urgent de remettre l'Outre-mer au cœur des priorités et des politiques publiques, et de créer les conditions qui favorisent l'investissement, gage de développement économique.

③ Persuasion techniques detection in political discourses

Question: Can we detect persuasion techniques in political discourses?

③ Persuasion techniques detection in political discourses

Question: Can we detect persuasion techniques in political discourses?

Our answer: Use ML classifiers to detect those techniques and assign them a category

③ Persuasion techniques detection in political discourses

Question: Can we detect persuasion techniques in political discourses?

Our answer: Use ML classifiers to detect those techniques and assign them a category

Emmanuel Macron on Nov 28, 2023:

$\underbrace{\text{Le cap que je porte a toujours été le même:}}_{\text{affirmation}}$

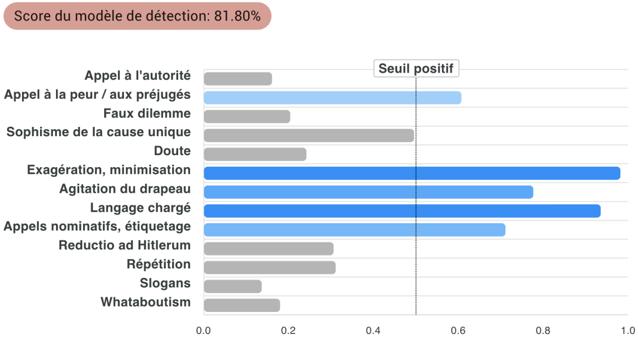
$\underbrace{\text{réindustrialiser la France}}_{\text{flag-waving}}$, $\underbrace{\text{gagner la bataille du plein-emploi}}_{\text{war term}}$,

$\underbrace{\text{être une Nation plus souveraine, industrielle et décarbonée.}}_{\text{flag-waving, hopes}}$

$\underbrace{\text{Soyons des optimistes déterminés.}}_{\text{loaded language}}$

3 Persuasion techniques detection in political discourses

1. Binary classifier:
   - Given a sentence, does it contain persuasive techniques?
2. Multi-class classifier:
   - Given a sentence, which of the 13 persuasive techniques are used?



Score du modèle de détection: 81.80%

Takeaways and future work

Takeaways:

- **ConnectionStudio**: a user-oriented data lake for data exploration
- **StatCheck**: a statistical warehouse for fact-checking

Future work:

- ConnectionStudio:
  - Link data graph Named Entities to Wikipedia/trusted resources
  - Propose new ways to query the data
  - Clean (automatically) data in the data lake
- StatCheck:
  - Gathering other sources than INSEE and EuroStat
  - Cross-check statistical data between sources
  - Investigate recent Machine Learning models

Final words

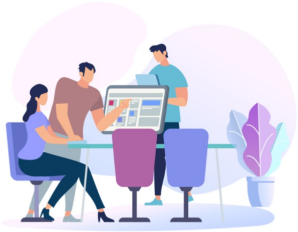
ConnectionStudio

StatCheck

If you are interested in what we are doing in the CEDAR team at Inria

Next: interactive sessions

- 2 groups
- 1 hour each



We will put our journalists hats:

1. Investigate a use-case in ConnectionStudio
2. Browse StatCheck data, tweets and ML outputs

And discuss about your questions, thoughts, ...

Next: interactive sessions

CAC40: CSV dataset

- Describes the top-40 most influential French companies
- Quite small (40 lines, 3 columns)

HATVP: XML dataset

- Describes political members' declarations about their wealth, jobs, financial interesets, ...
- Large ($\sim$ 2M nodes, $\sim$ 2M edges)

They share Named Entities:

- Crédit agricole, Danone, Education Nationale, Bouygues, ...

Let's see what we can do in ConnectionStudio!