



HAL
open science

Journalistic Dataspaces: Data Management for Journalism and Fact-Checking (Keynote Talk)

Ioana Manolescu

► **To cite this version:**

Ioana Manolescu. Journalistic Dataspaces: Data Management for Journalism and Fact-Checking (Keynote Talk). EDBT/ICDT 2019 Joint Conference, Mar 2019, Lisbonne, Portugal. hal-02081430

HAL Id: hal-02081430

<https://inria.hal.science/hal-02081430>

Submitted on 27 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Journalistic Dataspaces: Data Management for Journalism and Fact-Checking

Ioana Manolescu

CEDAR team, Inria Saclay and Ecole polytechnique

<http://pages.saclay.inria.fr/ioana.manolescu>, @ioanamanol

EDBT Conference, Lisbon, 2019



MOTIVATION

Bad memories: Romania, 1989



Bad memories: Romania, 1989



Ceaușescu re-elected
at the 14th congress!

Bad memories: Romania, 1989



Ceaușescu re-elected at the 14th congress!

He had held power since 1965.

Bad memories: Romania, 1989



Bad memories: Romania, 1985



1990: Things get better



... kind of



Democratic societies crucially need the press

❑ To debate and express dissent



❑ To analyze, confirm or refute public statements

Fact-checking

(Data) journalism

❑ To expose and explain society functioning



Democratic societies crucially need the press

❑ To debate and express dissent



❑ To analyze, confirm or refute public statements

Fact-checking

(Data) journalism

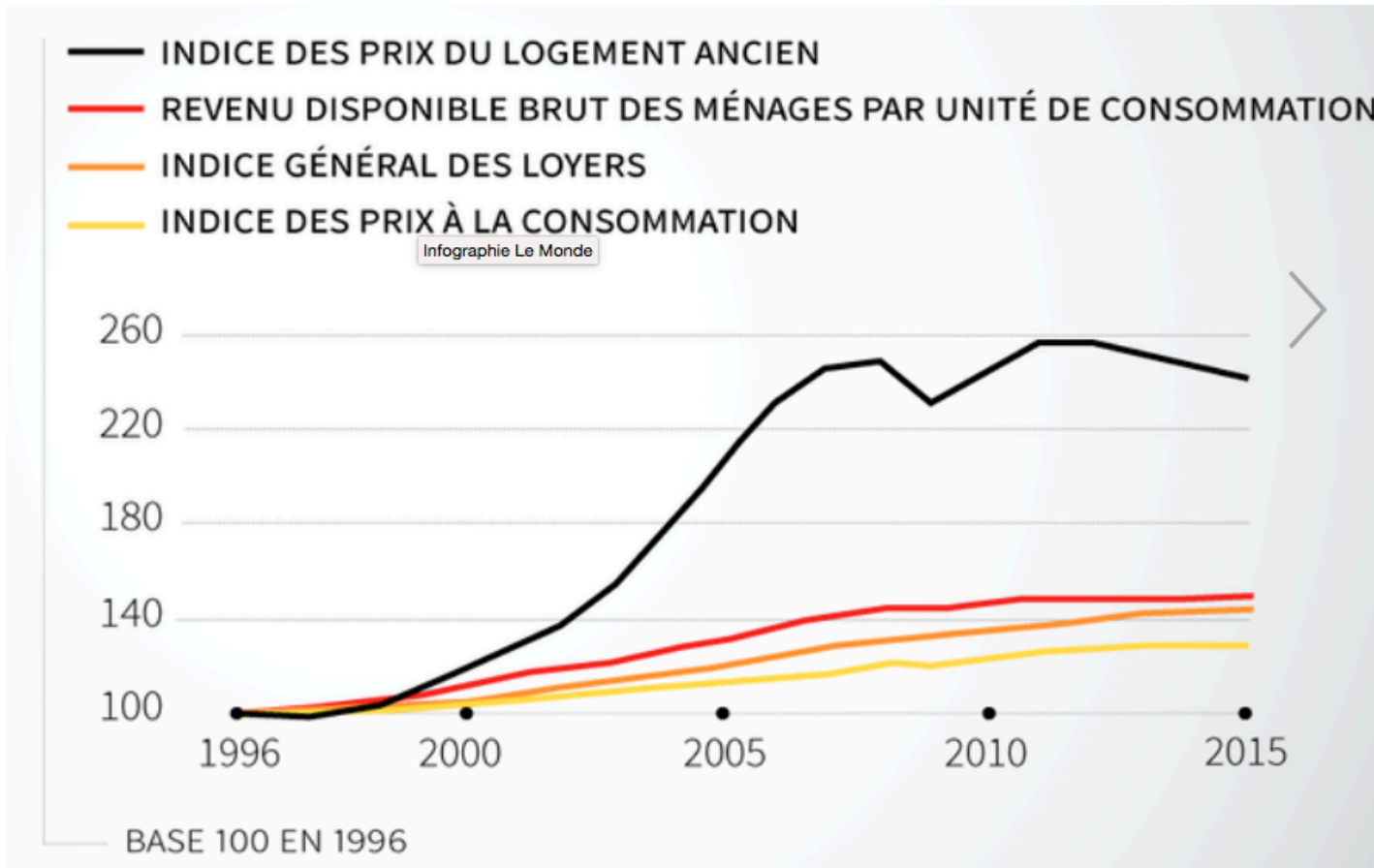
❑ To expose and explain society functioning



**DATA JOURNALISM,
JOURNALISTIC FACT-CHECKING,
FAKE NEWS DETECTION**

Data journalism

Investigative journalism based on **complex and/or large data**



http://abonnes.lemonde.fr/les-decodeurs/portfolio/2017/04/18/les-fractures-francaises-1-5-le-logement-les-raisons-de-la-crise_5112859_4355770.html

Ioana Manolescu, EDBT Conference, 27/03/2019

Data journalism

Panama Papers (International Consortium of Investigative Journalism, ICIJ)

The screenshot shows a web browser displaying the ICIJ website. The URL is https://panamapapers.icij.org/the_power_players/. The page features a profile for Jérôme Cahuzac, a French politician, and a corporate ownership diagram.

Jérôme Cahuzac
Budget minister at the Ministry of the Economy, Finance and External Trade (2012-2013); Deputy, National Assembly of France (1997-2002, 2007-2012)

Related countries
France

The lies told by Jérôme Cahuzac in 2013 triggered one of the most spectacular downfalls of a public official in the annals of French scandals. As a government minister waging a campaign against tax evasion, Cahuzac was forced to admit he lied to President François Hollande, former colleagues in Parliament and the French people when he repeatedly denied owning foreign bank accounts. He said he stashed over \$750,000 in a Swiss bank account for 20 years, moving the money to Singapore in 2009. His ex-wife disclosed an account opened in Great Britain in 1997. Cahuzac, who made a fortune as a cosmetic surgeon, resigned his ministry post and awaits trial for tax fraud.

Inside the Mossack Fonseca data » Offshore company held bank account for minister accused of tax fraud [Read more...](#)

Offshore glossary

Corporate Ownership Diagram:

- MONFORT CAPITAL PARTNERS JLT (registered) is connected to CERMAN GROUP LIMITED.
- TALWAY INTERNATIONAL CORP (Shareholder) is connected to CERMAN GROUP LIMITED.
- CERMAN GROUP LIMITED (Beneficial owner) is connected to Jérôme Cahuzac and Mr. Jerome Andre C.
- CERMAN GROUP LIMITED (Beneficiary) is connected to Jérôme Cahuzac and Mr. Jerome Andre C.
- 85 avenue de Breteuil Paris
France (registered address) is connected to Jérôme Cahuzac and Mr. Jerome Andre C.

Fact-checking (since 1930 approx.)

Fact-checking: verification of facts mentioned **in media content**

- ❑ To protect media reputation and avoid legal action

“The day I became a fact-checker at The New Yorker, I received **one set of red pencils** [...] for underlining **passages** on page proofs of articles that might contain **checkable facts**. [...] confirmed **with the help of reference books** from the magazine’s library”



<http://www.nytimes.com/2010/08/22/magazine/22FOB-medium-t.html>

Fact-checking (2012 – ongoing)


Not everyone agrees, however, that Democrats are not flip-flopping on the issue.

Mark Krikorian, executive director of the Center for Immigration Studies, a think tank that advocates for lower immigration, said that because the public doesn't know exactly what border barriers the Trump administration wants to build, Mulvaney's statement is not an "exact" comparison. But, he said, to dismiss it simply on that basis would be "tendentiously literal."

"The fact is that, other than the 'Mexico will pay for it' stuff, Trump is simply channeling the 2006 Secure Fence Act, and Schumer et al. who voted for it out of political calculation are indeed hypocrites for opposing the attempt to finally bring that law to fruition," Krikorian told us via email.

At the surface level, it is true in a broad sense that Democrats including Schumer, Obama and Clinton have in the past supported border fencing. All three voted for the Secure Fence Act of 2006, and all three supported the 2013 Senate immigration overhaul that passed the Senate, and which called for tougher border security including some additional fencing. But to claim that those measures are the same as what Trump is proposing is a stretch.

Share The Facts





Mick Mulvaney
Director, Office of Management and Budget

"We don't understand why the Democrats are so wholeheartedly against [President Trump's border wall]. They voted for it in 2006."

Fox News Sunday – Sunday, April 23, 2017

[SHARE](#) [READ MORE](#)







A Project of The Annenberg Public Policy Center

HOME ARTICLES ASK A QUESTION VIRAL SPIRAL ARCHIVES ABOUT US SEARCH MORE

THE WIRE

Did Democrats Once Support Border Wall?

By Robert Farley Posted on April 26, 2017

Like 835 Tweet Pin It Share 11

White House Office of Management and Budget Director Mick Mulvaney made an apples-to-oranges comparison when he said he couldn't understand why Democrats opposed supplemental funding for a border wall since many of them were for it back in 2006.

Mulvaney is referring to the Secure Fence Act of 2006, which called for construction of 700 miles of fencing and enhanced surveillance technology, such as unmanned drones, ground-based sensors, satellites, radar coverage and cameras. Sen. Chuck Schumer and then-Sens. Barack Obama and Hillary Clinton were among a bipartisan majority that voted in favor of the legislation, and it was signed into law by President George W. Bush.

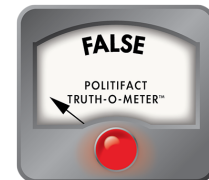
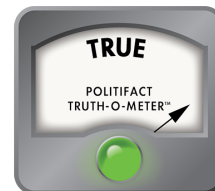
In a very general sense, the Democrats named by Mulvaney supported a bill to build more

ASK FACTCHECK

Like 983 Tweet Pin It Share 98

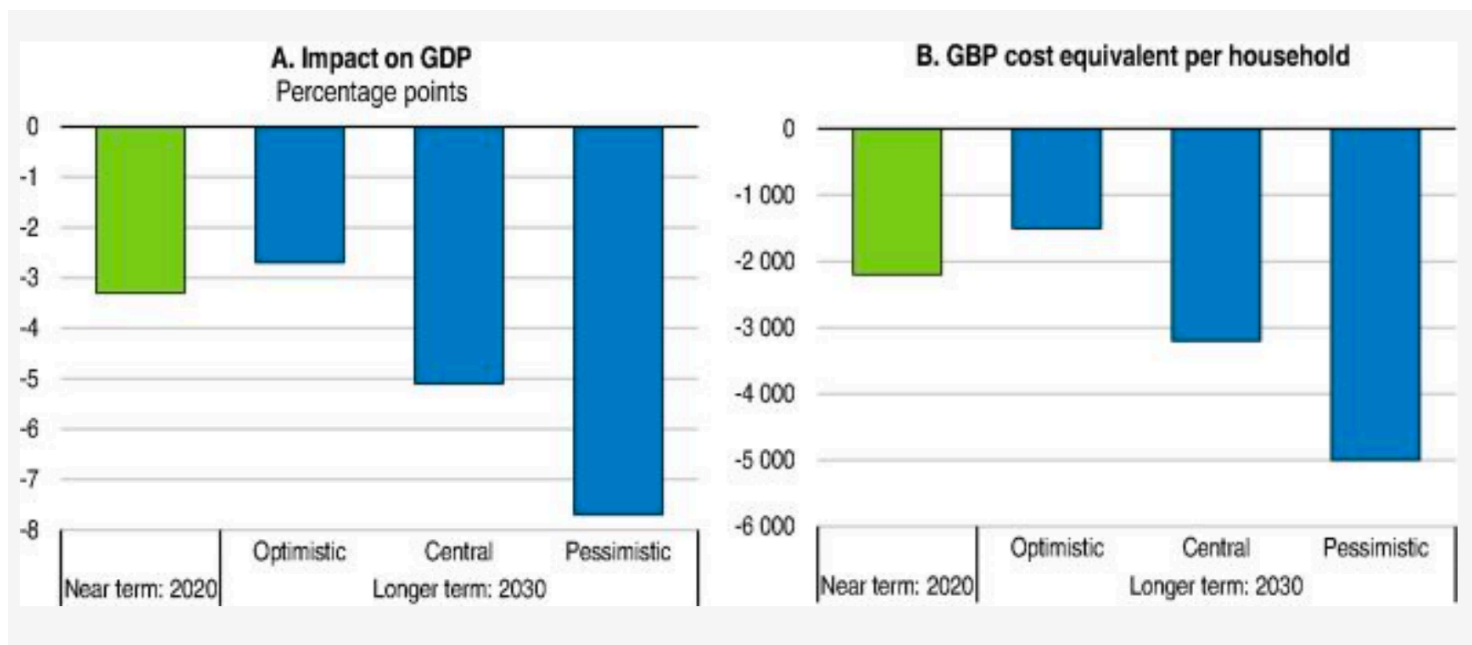
Q: Did the Supreme Court rule that public schools cannot teach students about Islam?

A: No. That false claim was spread by a network of fake news websites.



Where fact checking meets data journalism

- ❑ Most aspects of modern reality are **complex**
- ❑ **Explaining** can be as important and useful as checking
 - ❑ Also to analyze the future, e.g., Brexit impact:



<http://www.oecd.org/economy/the-economic-consequences-of-brexit-a-taxing-decision.htm>

Do we **need** to understand?

"Populism is telling people that there are simple answers to complex problems"

<https://www.express.co.uk/news/world/1034797/matteo-salvini-italy-budget-crisis-european-union-eu-news-alto-adig>

Salvini's populism on the march as EU-backing north Italy turns on Brussels

<https://www.ft.com/content/53bf2caa-3d6c-11e8-b9f9-de94fa33a81e>

Populism in Europe

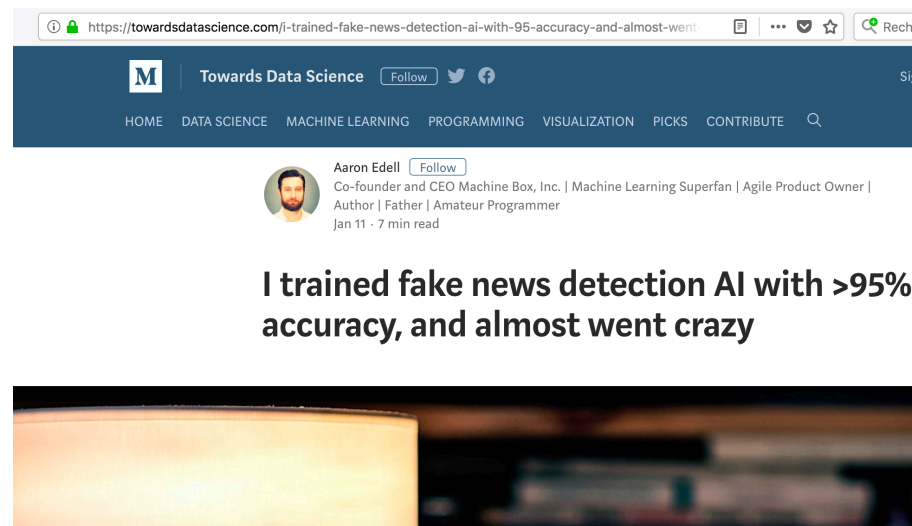
+ Add to myFT

Orban's populism prospers by challenging EU taboos

Hungary's prime minister leads the nationalist charge by challenging liberal taboos

Fact checking vs. fake news detection

- ❑ Fact checking is based on some **background information source**
 - ❑ **Truth commonly agreed upon**
- ❑ Fake news detection may or may not use a source
 - ❑ E.g., text classifier (true, fake) trained on text style



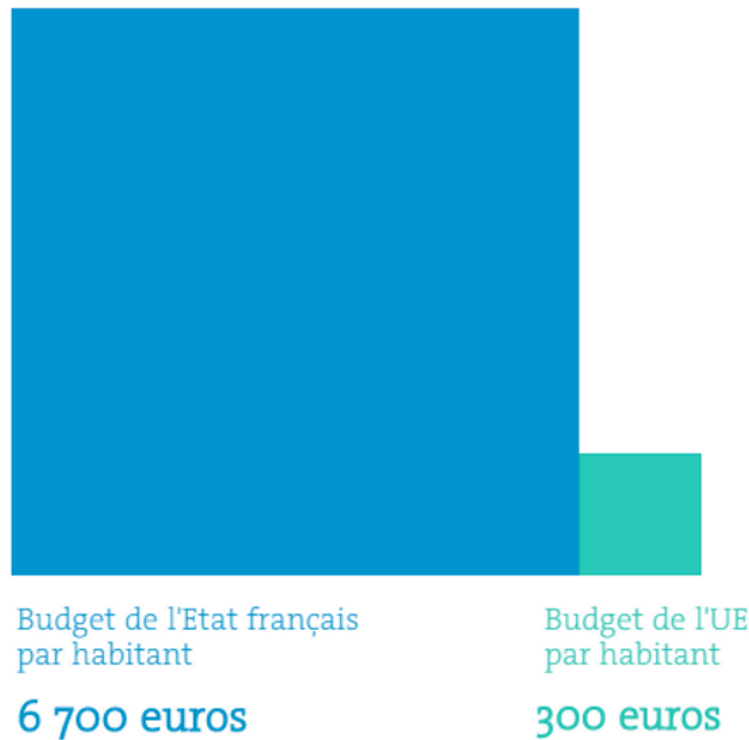
Most common fact-checking scenarios

- "What is the **value** of **metric X** in **space Y** at **time T**"?
 - **X**=youth unemployment, **Y**=Germany, **T**=2018
 - **X**=illegal immigrants, **Y**=Italy, **T**=[2015-2018]
 - **X**=budget for research, **Y**=France, **T**=2019
- Comparison patterns
 - **X1** against **X2**; **Y1** against **Y2**; **T1** against **T2**;
temporal trend etc.

Most common fact-checking scenarios

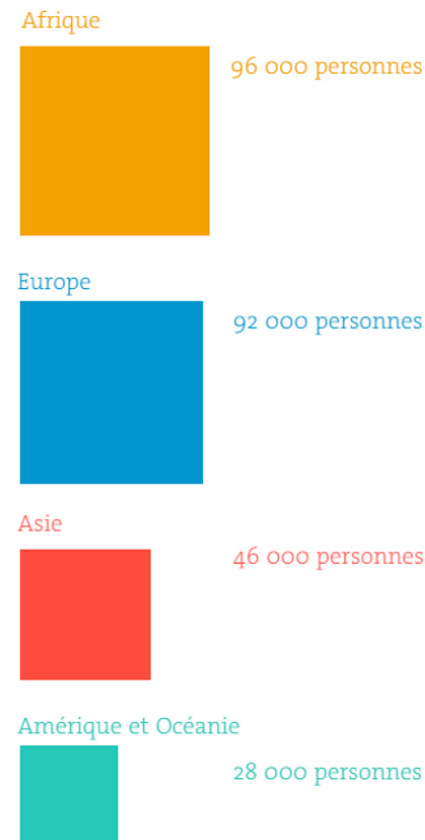
Le budget européen par habitant pèse nettement moins que celui de la France

Budgets pour l'année 2018 de l'UE et de la France rapportés à leurs populations respectives.



Parmi les immigrés en France, presque autant d'Européens que d'Africains

Lieu de naissance des personnes entrées sur le territoire national en 2017



Source : Insee

Most common fact-checking scenarios

- ❑ "What did **X** say about **Y** [at time **T**]"?
- ❑ "Is **X** related [in sense **S**] to **Y**?"



FRANCE — ENQUÊTE

Nicolas Sarkozy a bien servi les intérêts de Kadhafi. Voici les preuves

4 AVR. 2018 | PAR FABRICE ARFI ET KARL LASKE



Nicolas Sarkozy et Mouammar Kadhafi devant la maison du second bombardée par les Américains. © Reuters

Contrairement à ce qu'il a affirmé devant les juges puis dans les médias, Nicolas Sarkozy, actuellement mis en examen pour corruption dans l'affaire des financements libyens, a objectivement servi les intérêts du régime de Kadhafi entre 2005 et 2011. La preuve en cinq actes.

A CONTENT MANAGEMENT PERSPECTIVE

Lines of past and current research

1. **Model fact-checking** through a data and information management perspective
2. Identify **applicable tools and techniques**
 - In a special journalistic context (see next)
3. Devise **new models, tools and techniques** for fact-checking and data journalism problems

Projects and collaborations

- ❑ **Google Award** (2015) with X. Tannier (LIMSI)
- ❑ **ANR ContentCheck** (2016-2019) with X. Tannier (Sorbonne Université), S. Cazalens, P. Lamarre, J.-M. Petit, M. Plantevit (U. Lyon), F. Goasdoué (U. Rennes 1), Les Décodeurs (Le Monde)



LES DÉCODEURS
VENONS-EN AUX FAITS

<http://contentcheck.inria.fr>

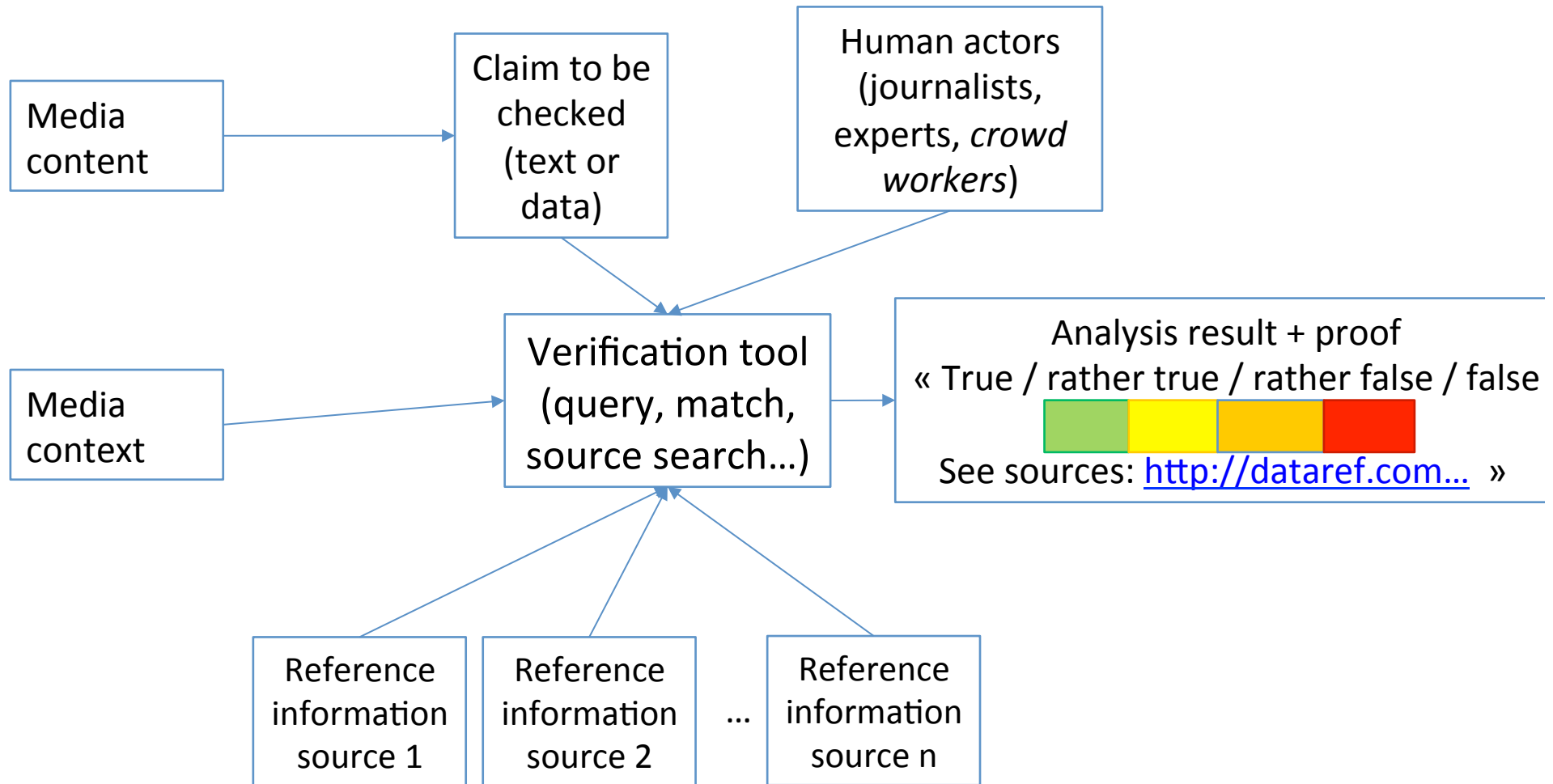
Inria

- ❑ **Inria Associated Team WebClaimExplain** with AIST Japan (Julien Leblay)
- ❑ **Collaborations** with H. Galhardas (University of Lisbon), former PhD S. Zampetakis and others

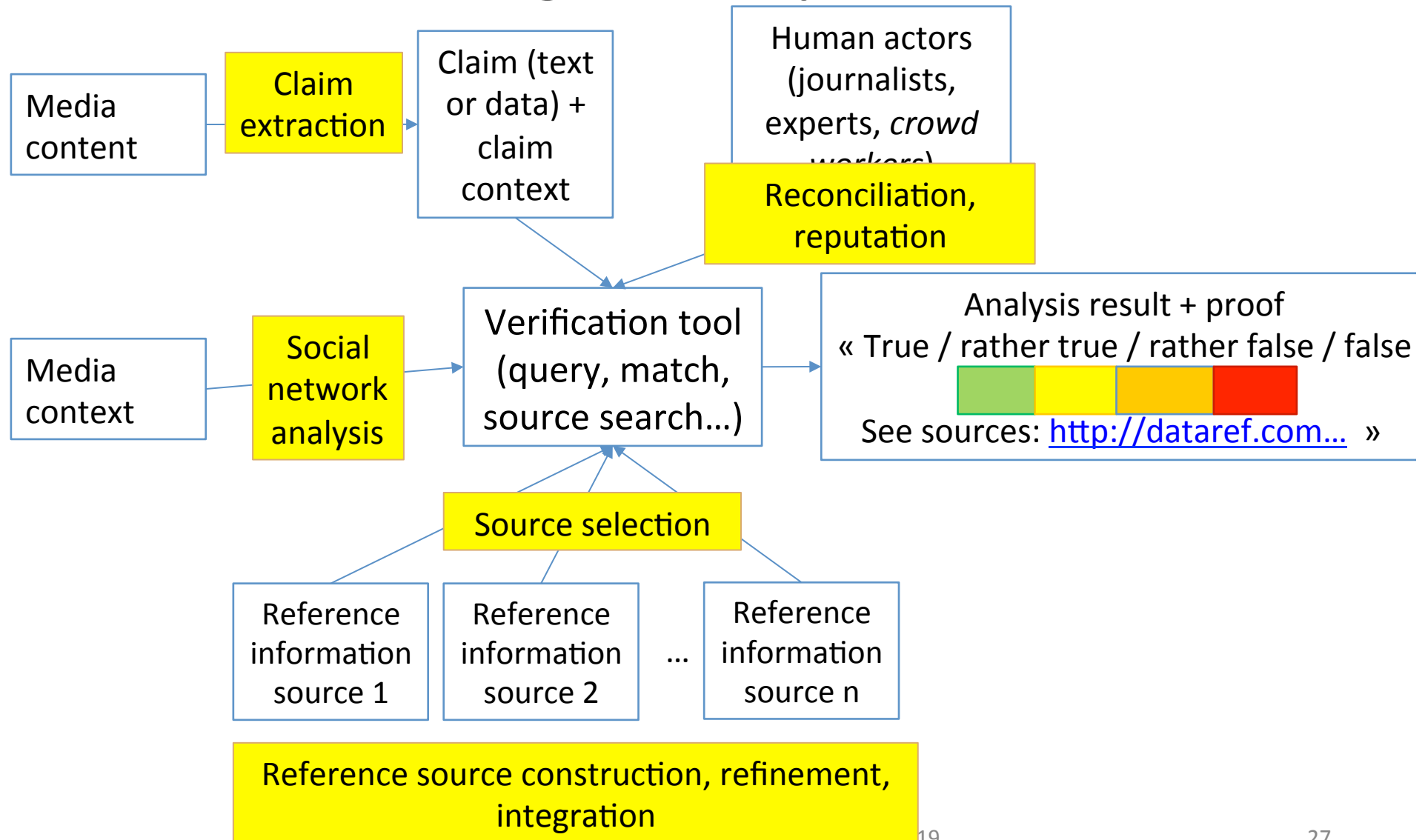
AIST



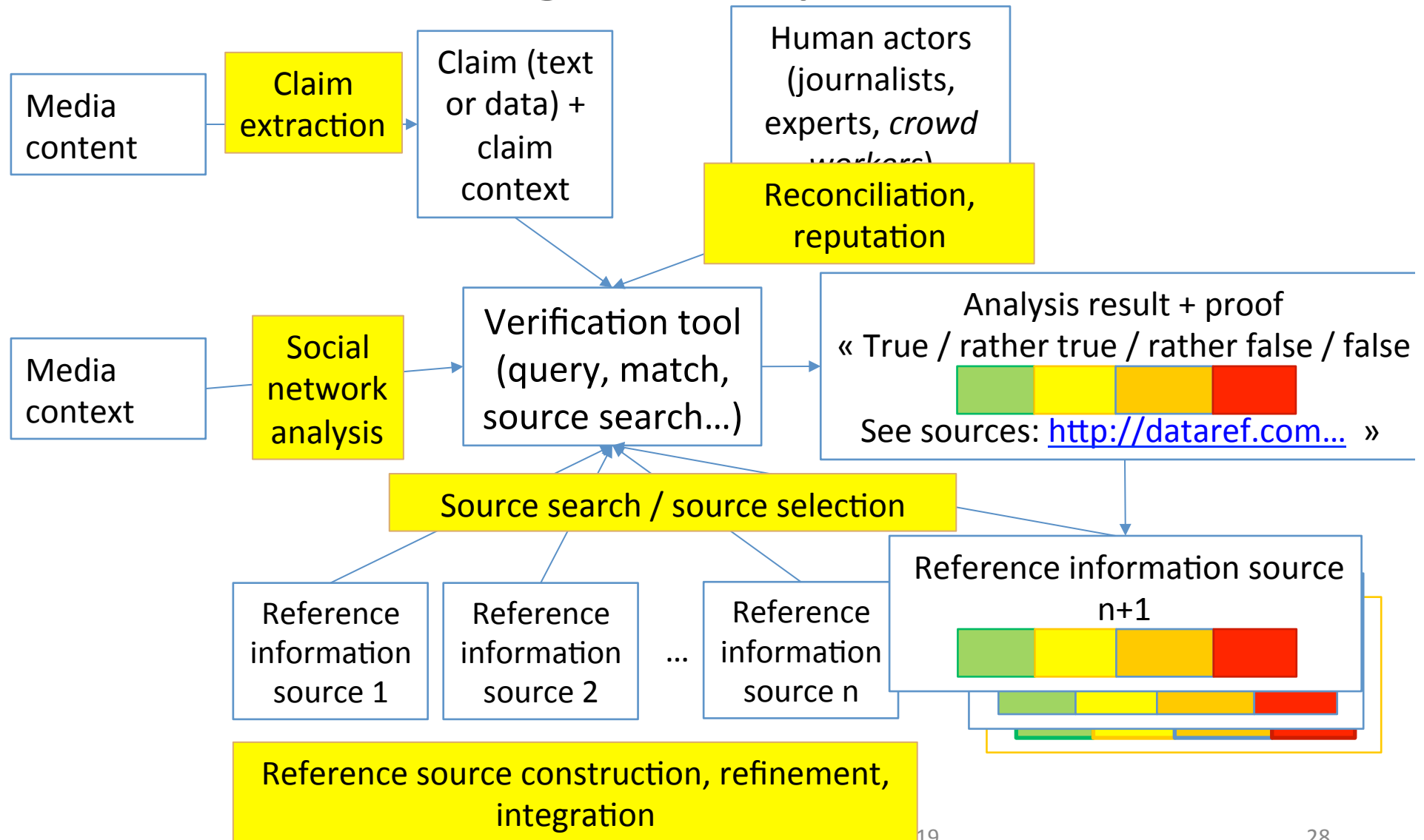
Fact-checking as a content management problem



Fact-checking as a content management problem



Fact-checking as a content management problem



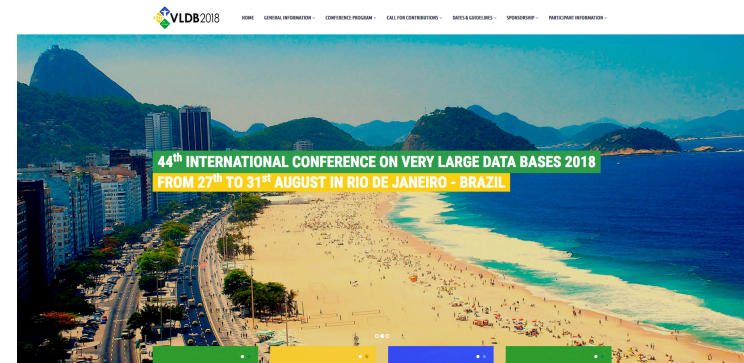
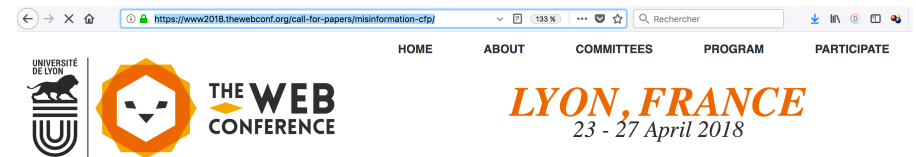
Fact-checking as a content management problem

[WWW2018] "A Content Management Perspective on Fact-Checking", S. Cazalens, J. Leblay, I. Manolescu, X. Tannier (fact-checking track)

[WWW2018 tutorial] "Computational fact-checking: problems, state of the art, and perspectives", J. Leblay, I. Manolescu, X. Tannier

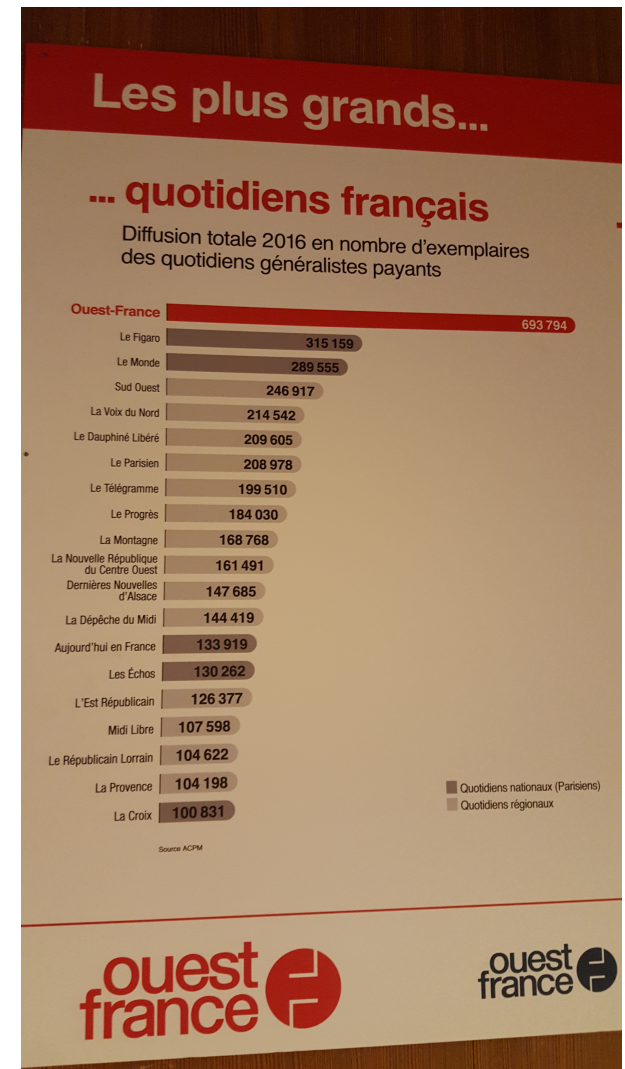
[VLDB2018 tutorial] "Computational fact-checking: a content management perspective", S. Cazalens, J. Leblay, P. Lamarre, I. Manolescu, X. Tannier

<http://contentcheck.inria.fr/>



Databases for fact-checking and data journalism

- ❑ Journalists do not, historically, build databases.
 - ❑ "Writers, not techies"
 - ❑ "Not part of our job"
 - ❑ Persisting data is novel to some
 - ❑ Journal information systems not always helpful
- ❑ However, some journals (e.g., OuestFrance) have valuable (reference) databases



Databases for fact-checking and data journalism

- ❑ **Curse of the coverage:** they need to cover (almost) any topic
- ❑ **Curse of newsworthiness:** write about hot topics of today (or tomorrow)
- ❑ They work under **strong time pressure**
- ❑ Good journalists are very **picky** with their data sources

Databases for fact-checking and data journalism

- ❑ Good journalists are very **picky** with their data sources...
- ❑ Which are **heterogeneous**: HTML, JSON, Excel, XML, CSV etc.
- ❑ Journalists won't write queries.
They may not know what they're looking for
- ❑ The fact-checking process and result must be **explainable**
 - ❑ Not (only) "an ML algorithm said so"

Improving access to reference data sources

[WebDB2018, SBD2017] T. D. Cao, I. Manolescu, X. Tannier.

"Extracting Linked Data from statistic spreadsheets"

“Créations d’entreprises en France en 2015” → return:

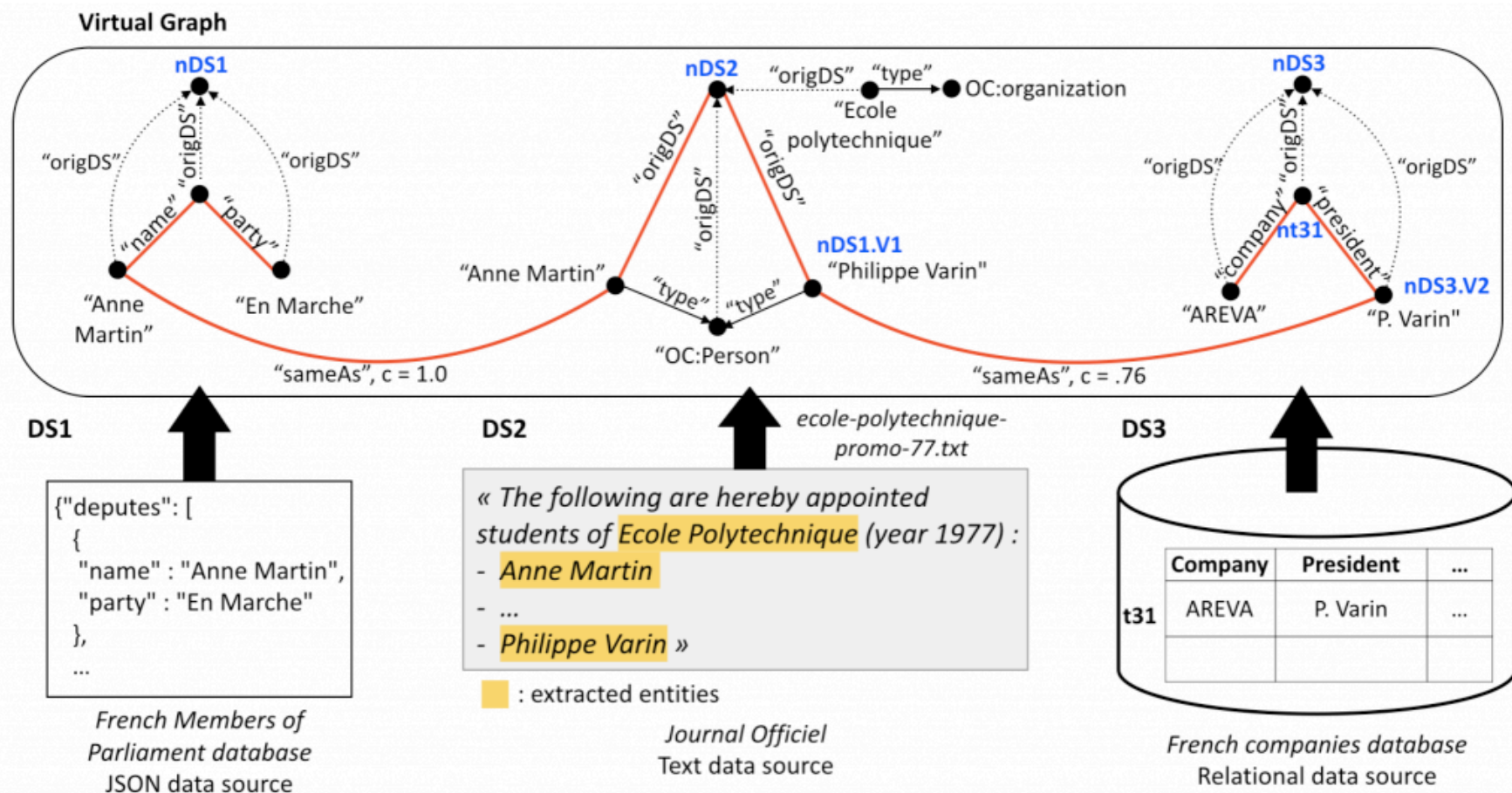
Créations d'entreprises dans quelques pays de l'Union européenne en 2015

en %

Pays	Taux de création
Allemagne	7,1
Belgique	6,2
Espagne	9,5
France (1)	9,5
Italie	7,5
Pays-Bas	10,1
Portugal	15,7
République tchèque	8,2
Royaume-Uni	14,3

Keyword search across heterogeneous sources

[VLDB2018demo] C. Chanial, R. Dziri, H. Galhardas, J. Leblay, M. Le Nguyen and I. Manolescu: "ConnectionLens: Finding Connections Across Heterogeneous Data Sources"



A data model for facts, statement and beliefs

Modeling who said what when

POLITICS January 19, 2018 7:19 pm Updated: January 19, 2018 7:24 pm

10 times that Trump has contradicted himself in his first year in office

By Darlene Superville The Associated Press

Comments 8 Facebook 398 Twitter Email Print ...



<https://globalnews.ca/news/3976740/trump-contradicted-first-year/>

[MisInfo@WWW2019] L. Duroyon, F. Goasdoué, I. Manolescu "A Linked Data Model for Facts, Statements and Beliefs"

- ❑ Fact **F** (holds according to the database)
- ❑ Statement of **X** about **F**
 - ❑ It's only according to **X**
- ❑ Statement of **Y** about statement of **X** about **F**
 - ❑ Only according to **Y**

Facts: application-dependent

Statements: writes, says, declares, ...

A data model for facts, statement and beliefs

Modeling who said what when

POLITICS January 19, 2018 7:19 pm Updated: January 19, 2018 7:24 pm

10 times that Trump has contradicted himself in his first year in office

By Darlene Superville The Associated Press

Comments 8 Facebook 398 Twitter Email Print ...



<https://globalnews.ca/news/3976740/trump-contradicted-first-year/>

[MisInfo@WWW2019] L. Duroyon, F. Goasdoué, I. Manolescu "A Linked Data Model for Facts, Statements and Beliefs"

- Fact **F**
- Statement of **X** about **F**
- Statement of **Y** about statement of **X** about **F**

Also:

- Time** (for facts and statements)
- Propagation of information** through communication →
- Who has heard of what when**

Lessons learned

- ❑ Work with **the right data**
 - ❑ The trusted data
- ❑ Work with the data **as it is**
 - ❑ Heterogeneous
 - ❑ Evolving usage and requirements prevent schema design, consolidation
 - ❑ **Extract, structure, connect**
- ❑ **Simplify** use
 - ❑ Keyword queries, canned queries

PERSPECTIVES

CS research for fact-checking

- ❑ DB, KR, IR, NLP
- ❑ In its most general statement, fact-checking supposes perfect NLP → study **sub-problems!**
- ❑ In fact-checking journalism, **human writers** chose topics, angle, style...
 - ❑ "A story wrapped around a query"
- ❑ Vision: build "**perfect data machines**" and give them to talented writers

A vision of journalistic dataspace

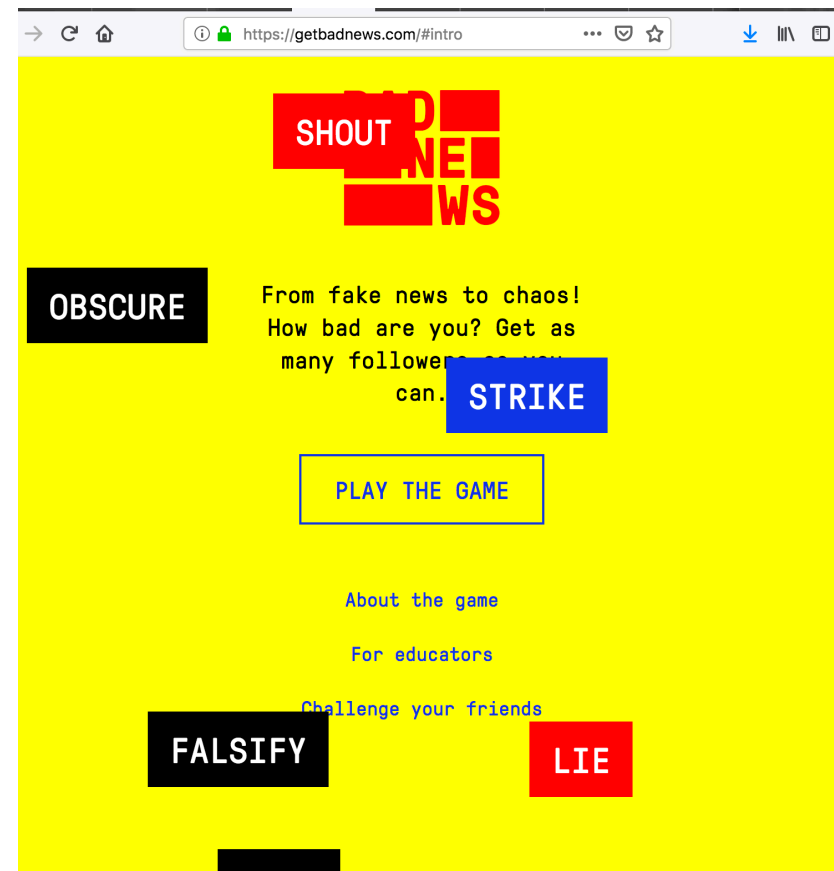
- ❑ "Dataspace": Franklin, Halevy and Maier, SIGMOD Record 2005
- ❑ Ingest **data of any nature**: structured (relational), semistructured (JSON, XML, (social) graphs), unstructured (text), KB...
- ❑ Storage, indexing
- ❑ Search across the data
 - ❑ Dong and Halevy [SIGMOD 2017]: kwd search, result can come from any data source
 - ❑ Bonaque, Cautis, Goasdoué, Manolescu [EDBT2016]: document search with social score component
 - ❑ ConnectionLens [VLDB2018]: find answers in any combination of data sources; "ad-hoc linked data"

Requirements for journalistic dataspace

- ❑ Time
 - ❑ Of data acquisition
 - ❑ Of events described in the data
- ❑ Provenance
 - ❑ Authorship metadata
 - ❑ Annotation by users
 - ❑ Access control based on provenance and annotations
- ❑ Ability to "derive" content (à la views)
- ❑ Semantic annotation and classification
- ❑ Social connections analysis
- ❑ Friendly interfaces
- ❑ Scalability

Roadmap: society

- Educate: the general audience, journalists, other social scientists
 - "Fake news creation" games, e.g.
<http://getbadnews.com>



Roadmap: society

- ❑ Educate: the general audience, journalists, other social scientists
- ❑ Education to media and the internet in schools



*In France, School Lessons Ask:
Which Twitter Post Should You Trust?*



Is this worth it?

“Some people will never be convinced”

- ❑ “Facts have a liberal bias” (Paul Krugman)
<https://www.nytimes.com/2017/12/08/opinion/facts-have-a-well-known-liberal-bias.html>
- ❑ "Scientists and humanity scholars believe in a constructed, logical discourse, and believe **humans yield to reason**. Businesspeople **know this is not true**, in general. Businesspeople have thus an **advantage** in winning political competitions."
George Lakoff, former Berkeley professor
<https://georgelakoff.com/2016/11/22/a-minority-president-why-the-polls-failed-and-what-the-majority-can-do/>
- ❑ Conspiracy theory adepts believe two obviously contradicting theories [Wood et al., 2012]

Thank you / questions?

<http://contentcheck.inria.fr/>

