



HAL
open science

Computational fact-checking: Problems, state of the art and perspectives

Ioana Manolescu

► **To cite this version:**

Ioana Manolescu. Computational fact-checking: Problems, state of the art and perspectives. 19e
Conférence Francophone sur l'Extraction et Gestion de Connaissances (EGC), Jan 2019, Metz, France.
. hal-01995318

HAL Id: hal-01995318

<https://inria.hal.science/hal-01995318v1>

Submitted on 26 Jan 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.

Computational fact-checking: Problems, state of the art and perspectives

Ioana Manolescu

CEDAR team, Inria Saclay and Ecole polytechnique

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

Extraction et Gestion de Connaissances, Metz, 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



Things get better



... kind of



1000 dead (approx.)
No one convicted.

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

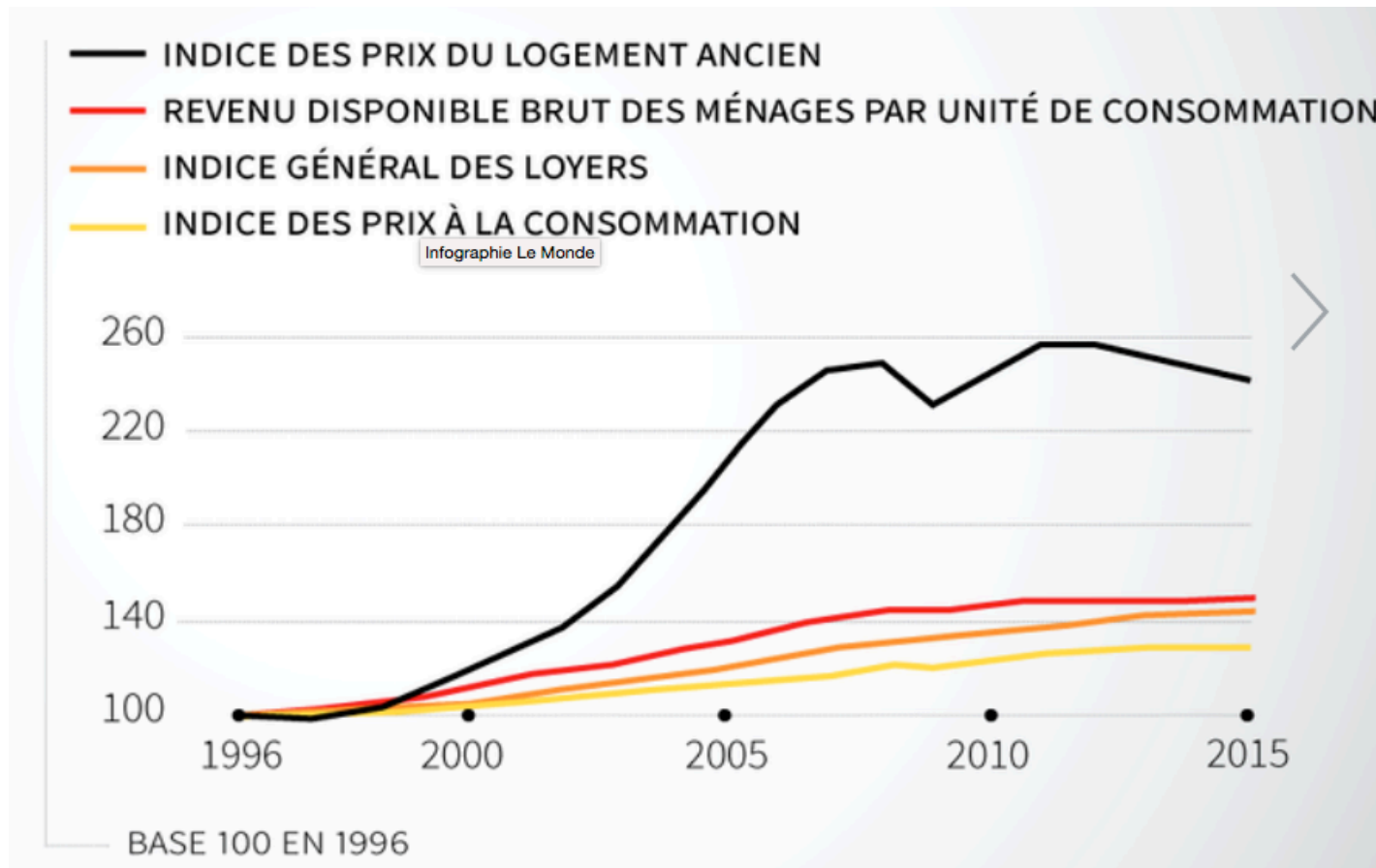
- To expose and explain society functioning



DATA JOURNALISM AND JOURNALISTIC FACT-CHECKING

Data journalism

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



Data journalism

Panama Papers (International Consortium of Investigative Journalism, ICIJ)

The screenshot shows a web browser displaying the ICIJ website. The page title is "The Panama Papers". The main content area features a profile for Jérôme Cahuzac, a French politician. To the right of the profile is a network diagram illustrating his financial relationships. The diagram shows a central node for "CERMAN GROUP LIMITED" with connections to "MONFORT CAPITAL PARTNERS JLT" (registered), "TALWAY INTERNATIONAL CORP" (Shareholder), and "Jérôme Cahuzac" (Beneficial owner). Below the diagram, a "registered address" is listed as "85 avenue de Breteuil, Paris, France".

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

Network Diagram:

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

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

MISLEADING

"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](#)

The screenshot shows the FactCheck.org website interface. At the top, there's a navigation bar with 'HOME', 'ARTICLES', 'ASK A QUESTION', 'VIRAL SPIRAL', 'ARCHIVES', 'ABOUT US', 'SEARCH', and 'MORE'. The main article title is 'Did Democrats Once Support Border Wall?' by Robert Farley, posted on April 26, 2017. The article content discusses Mick Mulvaney's statement and the 2006 Secure Fence Act. To the right, there's an 'ASK FACTCHECK' section with a question and answer about the Supreme Court ruling on public schools. At the bottom, there are six 'POLITIFACT TRUTH-O-METER' gauges: TRUE (green), MOSTLY TRUE (light green), HALF TRUE (yellow), MOSTLY FALSE (orange), FALSE (red), and PANTS ON FIRE! (flaming red).

It's not just **checking**

- Most aspects of modern reality are complex
- **Explaining** can be as important and useful as checking
 - Helps also analyze the future



The screenshot shows the 'LES DÉCODEURS' section of the Le Monde.fr website. The page features a dark header with the site's logo and navigation links. Below the header, there is a main navigation bar with categories like INTERNATIONAL, POLITIQUE, SOCIÉTÉ, ÉCO, CULTURE, IDÉES, PLANÈTE, SPORT, SCIENCES, PIXELS, and CAMPUS. The main content area is titled 'LES DÉCODEURS' in large, colorful letters, with the subtitle 'VENONS-EN AUX FAITS'. Below this, there is a sub-navigation bar with options like 'LES DÉCODEURS', 'Datavisualisation', 'Vérification', 'Nanographix', 'Contexte', 'Evasion fiscale', and 'Le blog du Décodex'. The main article is titled 'Une sortie de l'euro ferait-elle exploser la dette française ?' and includes a sub-headline 'La mesure phare du Front national, à l'échéance repoussée par Marine Le Pen, est entourée de nombreuses zones d'incertitudes.' The article is dated 'LE MONDE | 05.05.2017 à 11h51' and is written by 'Par Maxime Vaudano'. On the right side, there is a sidebar titled 'Les décodeurs, mode d'emploi' which explains the mission of the section: 'Les décodeurs du Monde.fr vérifient déclarations, assertions et rumeurs en tous genres ; ils mettent l'information en forme et la remettent dans son contexte; ils répondent à vos questions.' Below this sidebar, there is a link to 'LIX CHARTRE' with the text 'Lire la charte >'.

Libération *Désintox*: cost of saving Cyprus

The image shows a screenshot of the Désintox blog page. The article title is "Bobards en stock sur les plans de sauvetage européens". The date is "mardi 23 avril 2013". There are 0 comments, 11 tweets, and 34 likes. The article text includes quotes from Nicolas Dupont-Aignan, Florian Philippot, and Marine Le Pen. On the left side, there are three blue arrows pointing to the article text, with labels "\$2-3 bn", "\$10 bn", and "\$15 bn". On the right side, there are sections for "À PROPOS DE CE BLOG", "ALERTE EMAIL", and "RECHERCHER".

\$2-3 bn ←

\$10 bn ←

\$15 bn ←

À PROPOS DE CE BLOG

Créé en 2008, Désintox est un observatoire des mensonges et des mots du discours politique.

Qui sommes nous?

Retrouvez-nous également du lundi au jeudi à 20h05 sur Arte dans l'émission **28 minutes**.

ALERTE EMAIL

Recevez des alertes Désintox par email

votre email

Je m'inscris gratuitement

RECHERCHER

Q- OK

Saving Cyprus: how much does it cost?

The article reads:

- ❑ The money is not **given** but **lent**
- ❑ The European Mechanism of Stability will lend **\$9bn**
- ❑ Out of which France contributes 20% (approximately **\$2bn**)

However, things are complicated because:

- ❑ The initial contribution of France to the EMS (\$16 bn) **counts** toward French public debt
- ❑ The remainder contribution (\$124) **is not**



INTOX En période de rigueur budgétaire, les plan de sauvegarde successifs offrent depuis trois ans un boulevard aux détracteurs de l'euro et de l'Europe. Nicolas Dupont-Aignan et Florian Philippot ont ainsi, de concert, dénoncé récemment les milliards déversés à Chypre au moment où les Français se serrent la ceintures. «Les Français vont devoir donner 2 à 3 milliards d'euros pour des banques à Chypre. D'un côté on supprime les infirmières, on surtaxe les PME en France [...] mais quand il s'agit de donner de

l'argent de l'UE , c'est-à-dire des Français, à des banques pourries à Chypre, on le donne», dénonçait Nicolas Dupont-Aignan. En version Florian Philippot, cela donne : «C'est nous qui allons encore verser de l'argent puisque le MES s'est engagé à hauteur de 10 milliards d'euros, dont 2 milliards au titre de la France et des contribuables français». Un propos qui fait écho -notamment- à celui entendu quelque deux ans avant, dans la bouche de Marine Le Pen, au sujet cette fois du plan d'aide à la Grèce : «Comment peut-on imposer aux Français, aux classes populaires et moyennes, aux petites entreprises, de nouvelles taxes, des taxes sur les sodas, une hausse de la CSG de 550 millions d'euros, et d'un autre côté alimenter de 15 milliards d'euros supplémentaires en Grèce l'incendie de la zone euro ?»

DÉSINTOX Quel impact ont donc les plans de sauvetage successifs de l'Euro? Et le dernier en date, en direction de Chypre, va-t-il contraindre les Français à allonger 2 milliards d'euros?

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

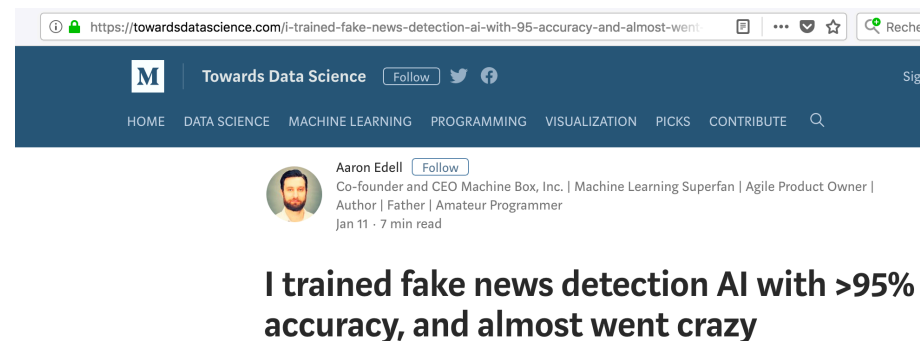
+ Add to myFT

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

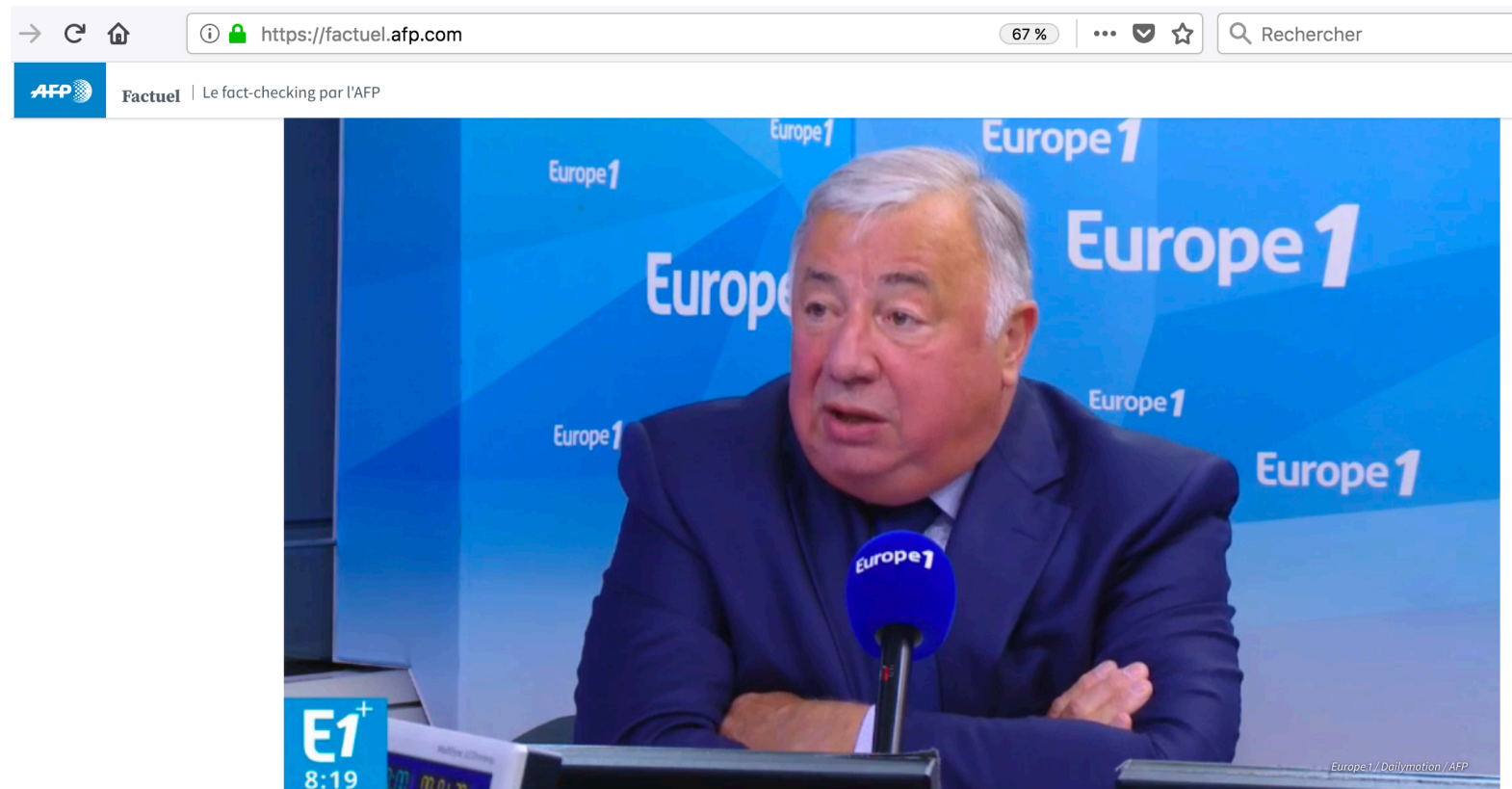
- Fact checking is based on some **background information source**
 - Truth commonly agreed upon
- Fake news detection may or may not use a source
- E.g., text classifier (true, fake) trained with major news agency / fake content (often virulent style)



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
 - **X**=average monthly wages, **Y**=China, **T**=2018
- Comparison patterns
 - **X1** against **X2**; **Y1** against **Y2**; **T1** against **T2**;
temporal trend etc.

Most common fact-checking scenarios



Deuxième trimestre mis à part, la France n'a pas la "croissance la plus faible de la zone euro"

Most common fact-checking scenarios

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

s.lemonde.fr/es-decodeurs/article/2015/11/09/es-arguments-perimes-de-la-gauche-parisienne-contre-valerie-peccresse_48056

Les arguments périmés de la gauche parisienne contre Valérie Pécresse

Le camp Bartolone a ressorti une déclaration de 2012 de la candidate de droite sur le mariage gay, faisant mine d'oublier qu'elle a, depuis, changé de position sur le sujet.

Le Monde.fr | 09.11.2015 à 15h11 • Mis à jour le 10.11.2015 à 11h53 |

Par Samuel Laurent

Réagir Ajouter Partager (677) Tweeter



L'offensive est-elle concertée ? Spontanée ? Lundi 9 novembre, les socialistes franciliens et leurs soutiens font circuler sur les réseaux sociaux un article visant Valérie Pécresse, tête de liste Les Républicains en Ile-de-France : elle souhaiterait « démarier » les couples homosexuels.

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 **existing tools and techniques** which could be directly applied
 - In a special journalistic context (see next)
3. Devise **new models, tools and techniques** for fact-checking and data journalism problems

Projects and collaborations

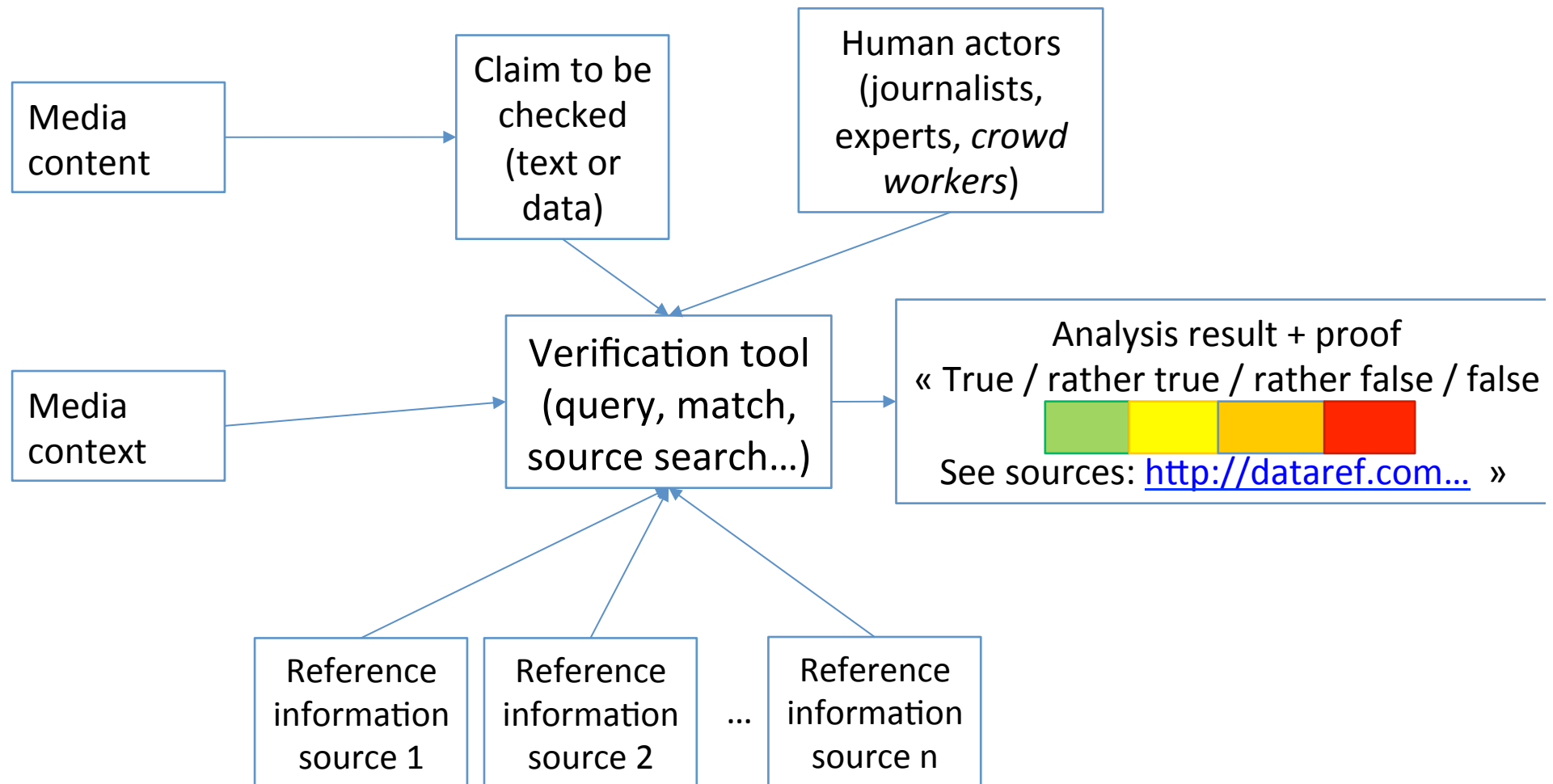
- **Google Award** (2015) with X. Tannier (LIMSI)
- **ANR ContentCheck** (2016-2019) with X. Tannier (Sorbonne Université), S. Cazalens, P. Lamarre, J.-M. Petit, M. Plantevit (U. Lyon), F. Goasdoué (U. Rennes 1), Les Décodeurs (Le Monde)



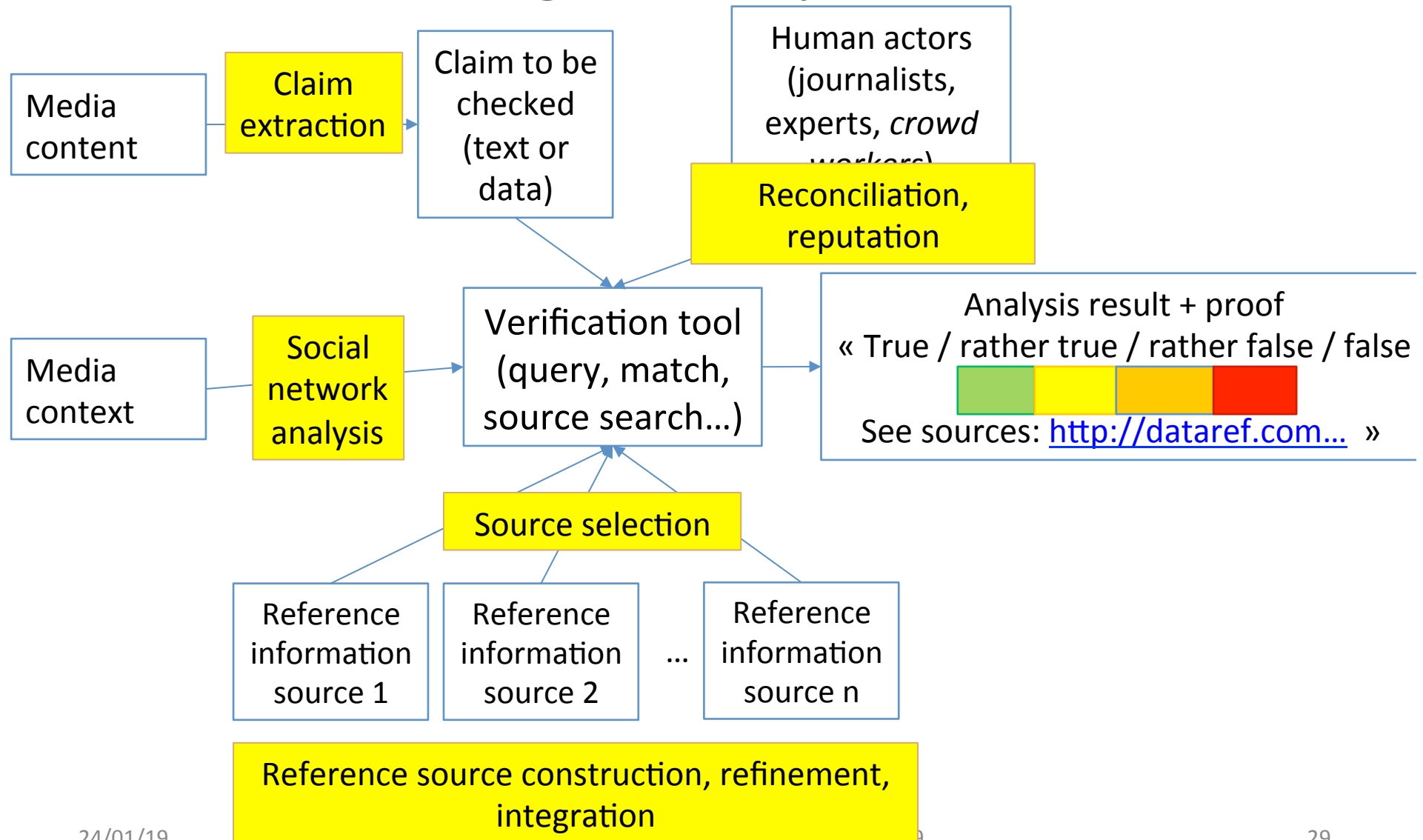
<http://contentcheck.inria.fr>

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

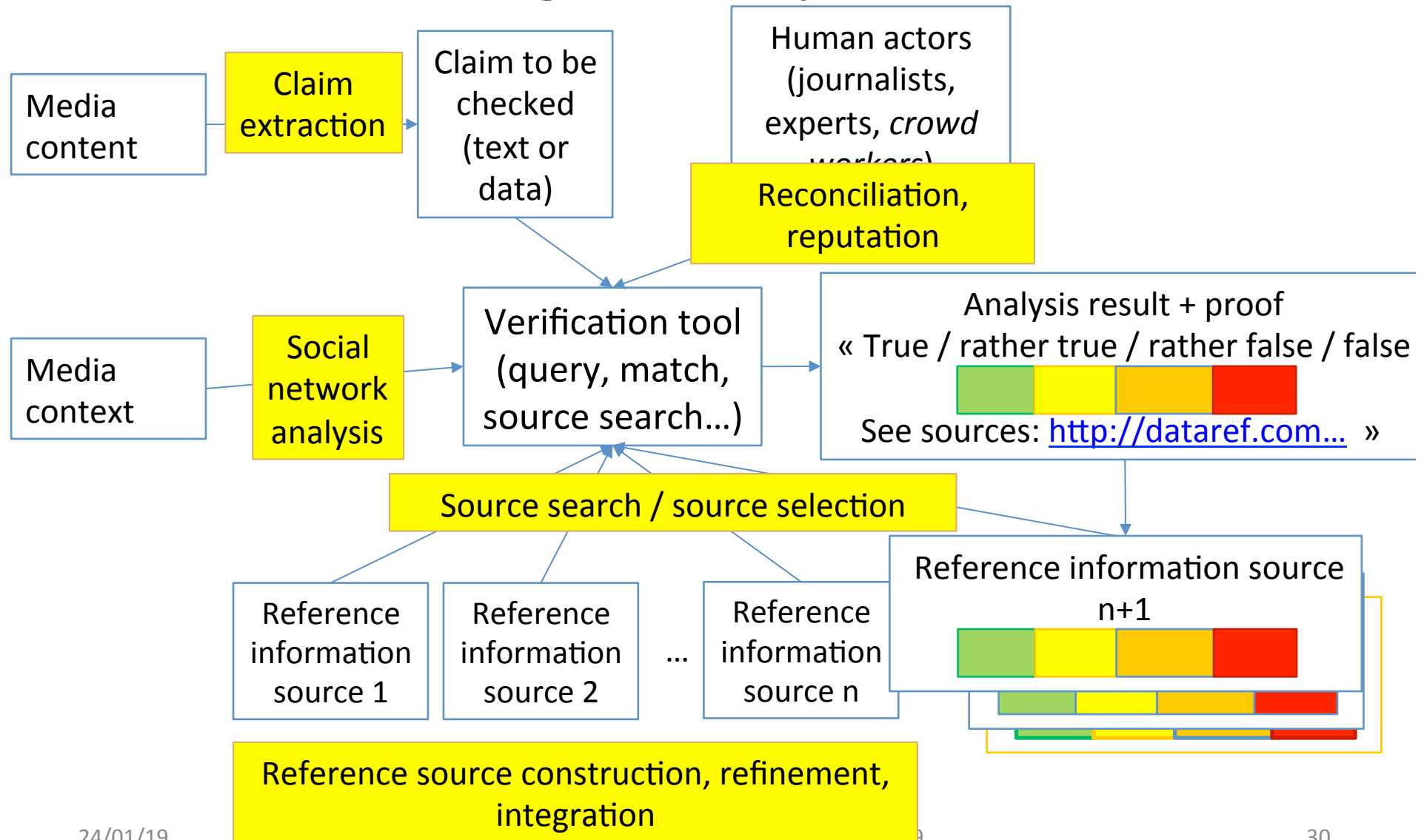
Fact-checking as a content management problem



Fact-checking as a content management problem



Fact-checking as a content management problem

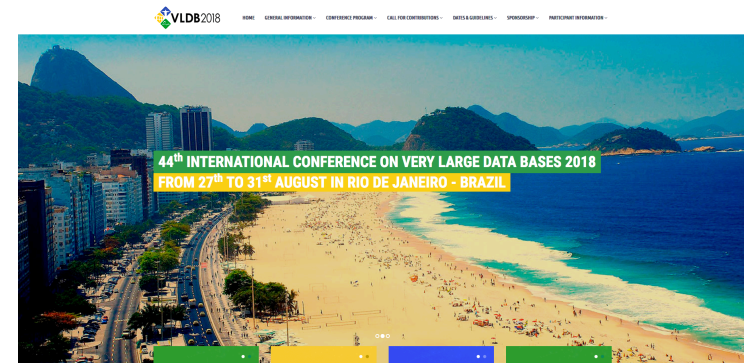
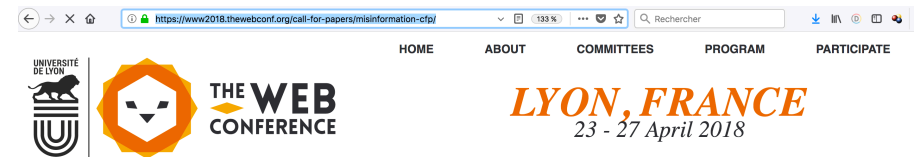


Fact-checking as a content management problem

[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



Which existing technologies can be used?

- **Database queries?**
 - Journalists do not, generally, build databases.
 - "Not part of our job"
 - Persisting data is novel to some
 - SQL not user-friendly
 - Not always helped by journal information systems
- Remarkable exception: Ouest France



Which existing technologies can be used?

- **Database queries?**
 - Journalists do not, generally, build databases.
 - "Not part of our job"
- **Curse of the coverage:** they need to cover (almost) any topic
- **Curse of noteworthiness:** write about hot topics of today (or tomorrow)
- They work under **strong time pressure**

Which existing technologies can be used?

- **IR/NLP?**
 - Better options
- The outcome of fact-check has to be **explainable**
 - Not (only) "an ML algorithm said so"
 - But they don't like to train ML systems, either.
- Extremely **picky on data sources**
 - "Many Web sources claim it" is not an option

Automatic source selection: FactMinder

[SIGMOD2013demo] F. Goasdoué, K. Karanasos, Y. Katsis, J. Leblay, I. Manolescu, S. Zampetakis: "Fact-checking and analyzing the Web"

- Plug-in
- Brings relevant data from document and knowledge bases

The screenshot displays the FactMinder interface. On the left, a news article snippet is shown with the headline "Bill Clinton apporte son soutien à Barack Obama". The main content area is partially obscured by a sidebar on the right. The sidebar contains several panels:

- Concepts:** A list of entities extracted from the document, including Bill Clinton (dbpedia:Person), Barack Obama (dbpedia:Person), Des Moines, Iowa (dbpedia:Place), Hillary Clinton (dbpedia:Person), Michelle Obama (dbpedia:Person), Bruce Springsteen (dbpedia:Person), and Pennsylvanie (dbpedia:Place).
- Related stories:** A list of related news items, such as "Ohio, Florida, ... être compliqué" and "Des indicateurs ... d'Obama".
- Facts & figures:** A panel titled "Curriculum" showing biographical information for Bill Clinton, including his birth date (August 19, 1946) and political party (Democratic Party).
- Quotes:** A panel containing quotes from the document, such as "Launching the Africa Regional Media Hub in Johannesburg" and "President Obama speaking LIVE for the last time before the election".
- Sources:** A panel listing the sources of the information, including "Romney... agressifs".

- « Second screen »

Improving access to reference data sources

[BDA2018] T. D. Cao, I. Manolescu, X. Tannier. "Extracting Linked Data from statistic spreadsheets" (WebDB 2018, SBD 2017)

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

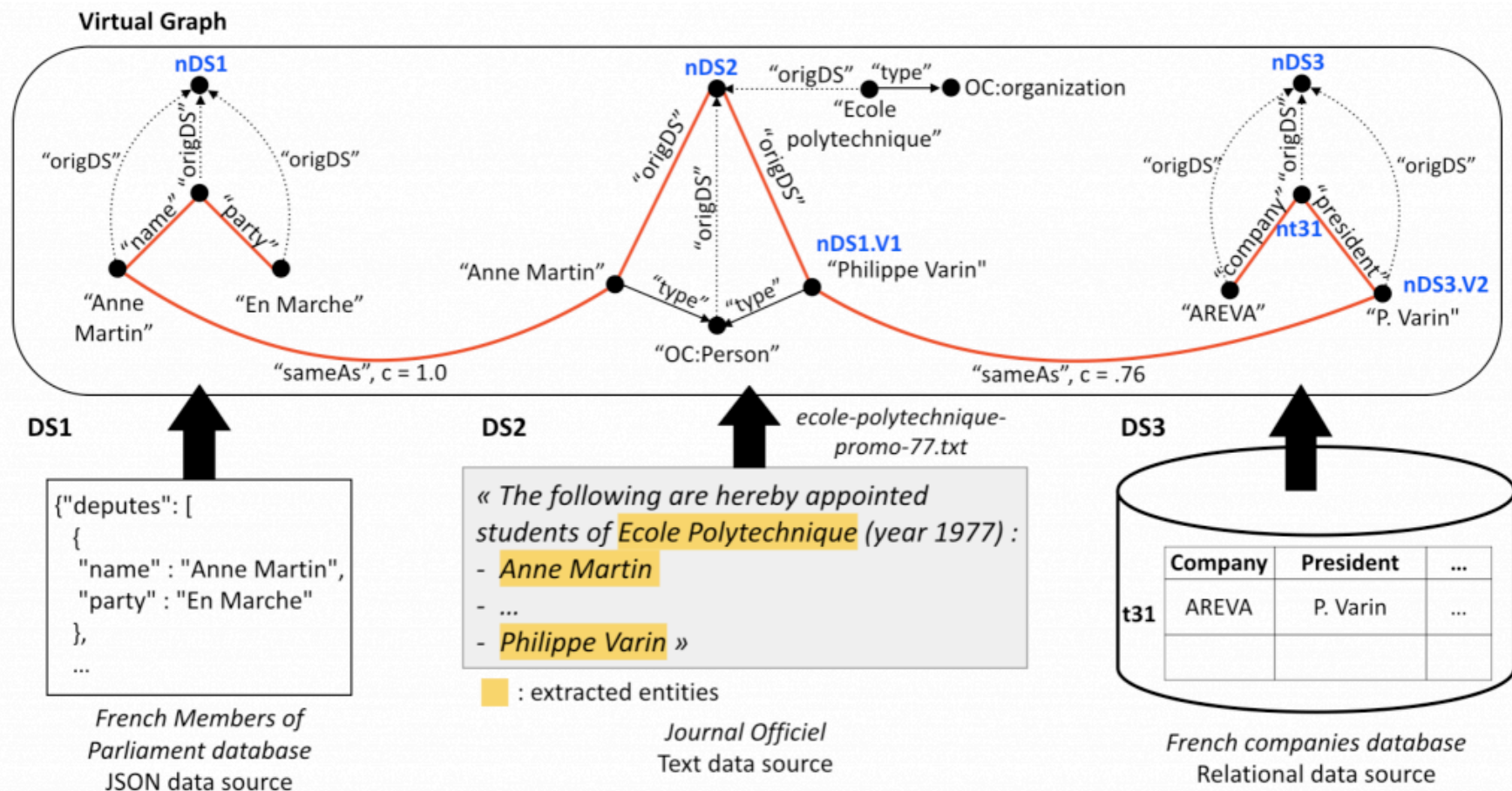
Créations d'entreprises dans quelques pays de l'Union européenne en 2015

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

With F. Goasdoué and L. Duroyon (U. Rennes 1)

Modeling who said what when

- "On Jan 22, 2017, Le Canard Enchaîné wrote that François Fillon had stated on Jan 21 that Penelope worked for the NA from 2002 to 2008."
- What has one actor heard of?
- Who knew this at that time?
- Who reversed their position on a topic?

Summary of our work

- Automatically find most relevant reference data for a claim
 - RDF dump of full INSEE statistics
 - Table-aware full text search on the resulting graphs
 - Identifying statistic mentions in text (submitted)
- Identify connections between entities across heterogeneous content: ConnectionLens
- Modelling temporal facts, beliefs and statements

STATUS AND ROADMAP

Status

- Any **question answering framework** which can be plugged on trusted data sources is useful
 - Usually no time nor skills to integrate the data
- No IT infrastructure can be counted upon
 - More "a jumble of tools"
- ML promises a lot but delivery is hard(er)
 - Also: journalists ambivalence
- Strong and increasingly IT-literate community
 - E.g. <https://www.poynter.org/channels/fact-checking>

Roadmap: computer science research

- 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

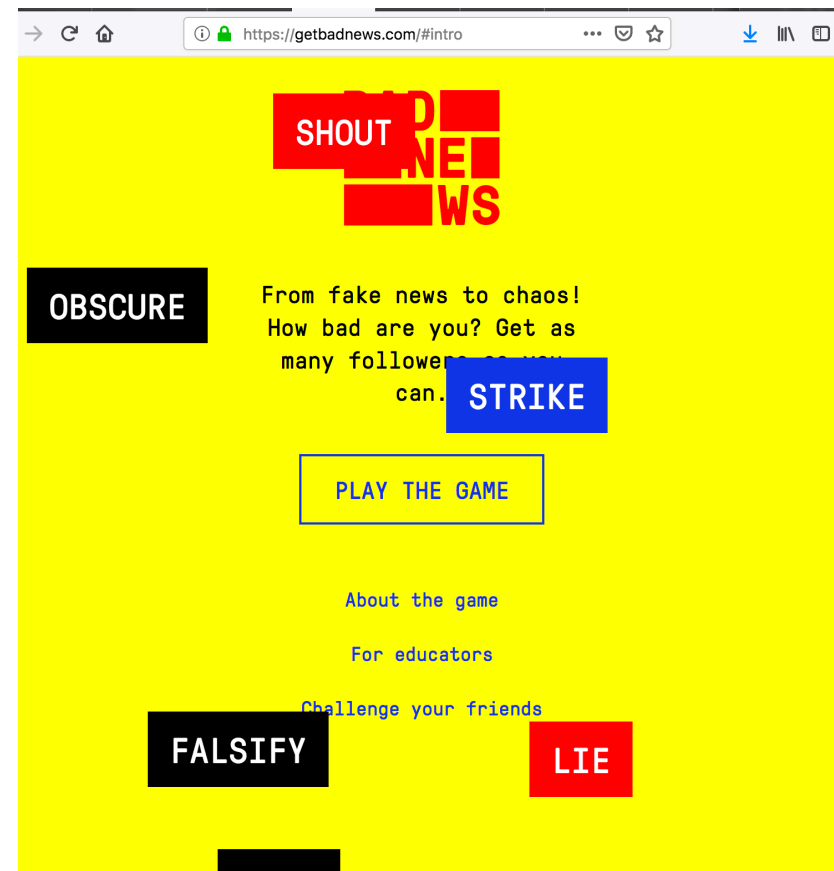
- "Dataspace": Franklin, Halevy and Maier, SIGMOD Record 2005
- Ingest **data of any nature**: structured (relational), semistructured (JSON, XML, [social] graphs), unstructured (text), KB...
- Storage, indexing
- Search across the data
 - Dong and Halevy [SIGMOD 2017]: kwd search, result can come from any data source
 - Bonaque, Cautis, Goasdoué, Manolescu [**EDBT2016**]: document search with social score component
 - ConnectionLens: 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
- Capacity to "derive" content (à la views)
- (Semi-automatic) 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/](http://contentcheck.inria.fr/)