ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Distributed Information Systems Laboratory

The Human Factor in Big Data Analysis

Karl Aberer, EPFL
Distributed Information Systems Laboratory
lsir.epfl.ch

Big Data = Volume, Velocity, Variety, Veracity

Variety = Semantics = Meaning
Retrieval, data integration, information extraction, …

Veracity = Pragmatics = Utility
Data quality, credibility, authority, trust, …

Stating the obvious: every semantic and pragmatic information processing task related to human concerns requires <u>human input</u>

For example: Google is a huge relevance feedback engine

Observation

Big Data analysis today

- **Key innovation:** capacity to automatically process and analyse huge volumes of data
- **Key bottleneck:** human input to make the processing meaningful

Example: recent progress in machine translation and image recognition with deep learning

- Rely on huge corpuses with "ground truth"

Example: automatic generation of image captions

Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

An extreme perspective

CHRIS ANDERSON MAGAZINE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

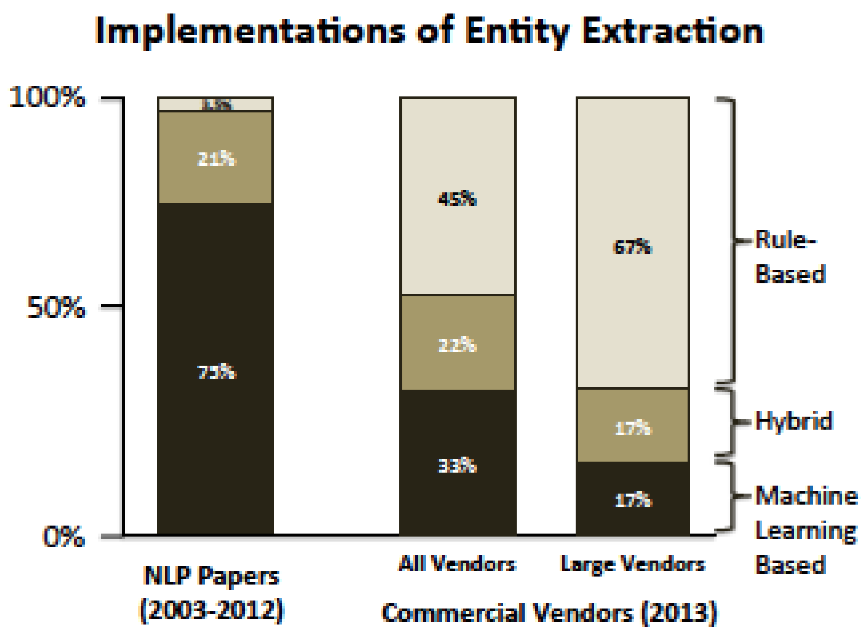
No models!

No causality!

No understanding!

The Reality

- But often no ground truth available, in particular for applications with “not so big data” and involving expert knowledge



Chiticariu, Laura, Yunyao Li, and Frederick R. Reiss. "Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!." *EMNLP*. No. October. 2013.

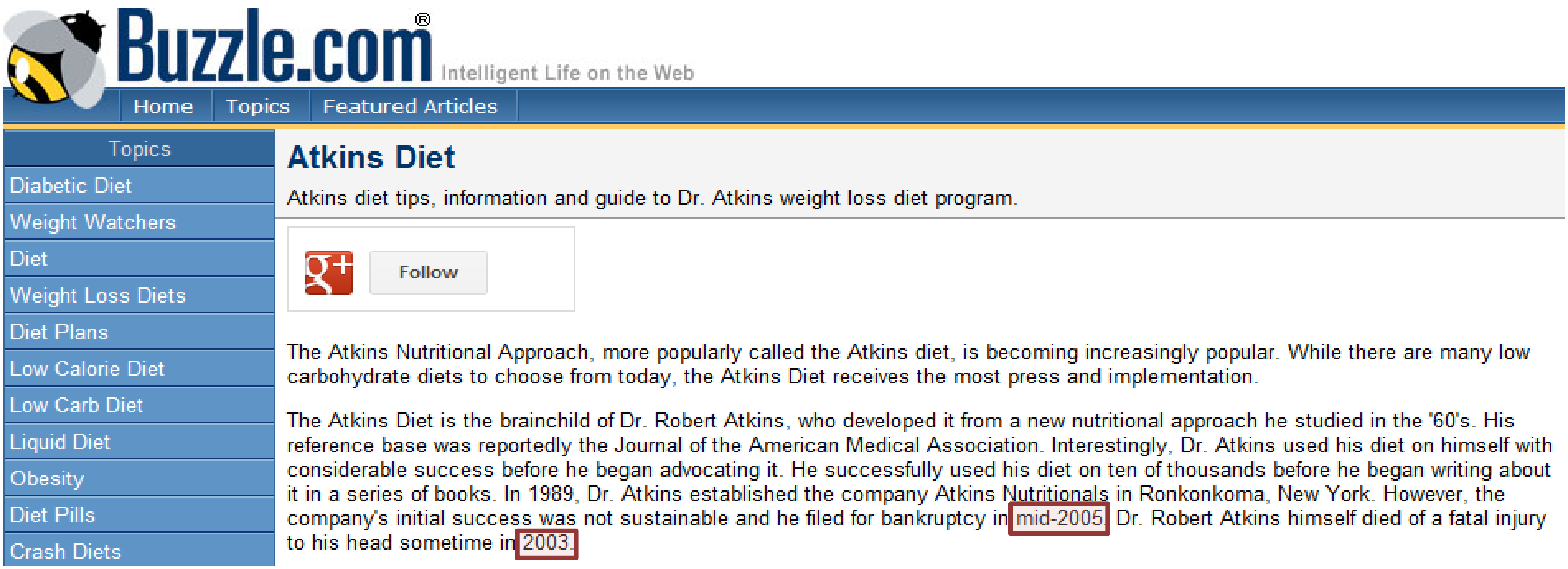
Overview

Three case studies

1. Web credibility
   How human input enables machine learning
2. Data integration
   How humans and machines cooperate efficiently in a problem solving task
3. Social Media Analytics
   How humans interpret latent structures found by machine learning

CASE STUDY 1: WEB CREDIBILITY

How to Evaluate Web Credibility?



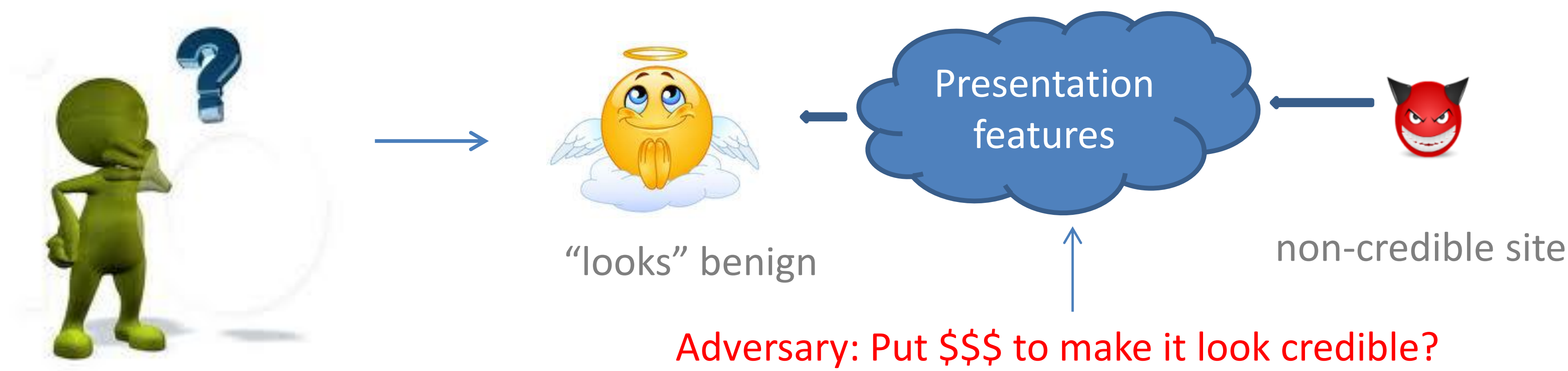
When you browse a webpage, how do you know it's content is valid and accurate?

Web Credibility – the problem

Increasingly difficult to assess credibility of Web content

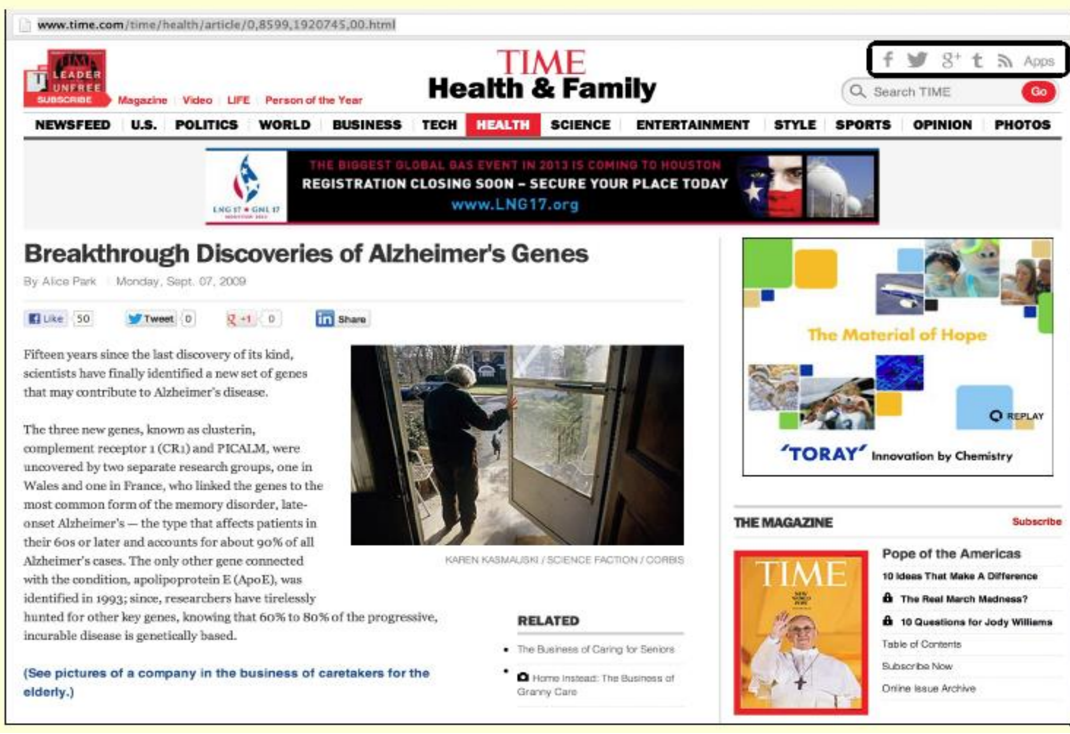
- Economic incentives to manipulate information
  - Marketing, fraud, political motives, etc.
- Enormous volume of web information

User: Believe or not?



Which features indicate credibility?

- Numerous candidate features could indicate credibility?
- How to determine?
- Let experts annotate a collection of documents



Social Features
Fetched from third-party websites

Facebook Likes		50
Facebook Shares		5
Twitter Tweets		0
Bitly Clicks	bitly	10
Alexa Rank	Alexa	676
Google PageRank		4

Content Features
Computed from webpage content or metadata

Readability	6th Grade
Size of Ads	24%
Number of Ads	3
Number of Spelling Errors	10
Objectivity	Objective
Domain Type	.com
Category	Health

Credibility Features

Topic	Query Terms	Expert URL Filters	# of Users
Health	Atkins diet effectiveness P90x exercise program H1N1 vaccine side effects Alzheimer's genes Autism warning signs	ncbi.nlm.nih.gov/pubmed pubmedcentral.nih.gov	254,175
Finance	Is it a good time to invest in gold? What mutual funds to invest in Reduce personal debt Mortgage refinancing Is it a good time to invest?	bloomberg.com edgar-online.com hoovers.com sec.gov	201,014
Politics	Iran election rigged Cash for clunkers eligibility Obama birthplace Death Panels Tea Party	foreignaffairs.com theatlantic.com foreignpolicy.com hir.harvard.edu economist.com	66,155
Celebrity News	Lady Gaga Adam Lambert Nadya Suleman Floyd Landis Michael Jackson	ew.com usmagazine.com people.com	692,611
Environmental Science	Renewable energy Green jobs Climate change Cap-and-trade Organic Eating	pewclimate.org epa.gov rff.org nrdc.org whitehouse.gov/administration/ceq	83,476
All Users		*(none)*	*50,473,520*

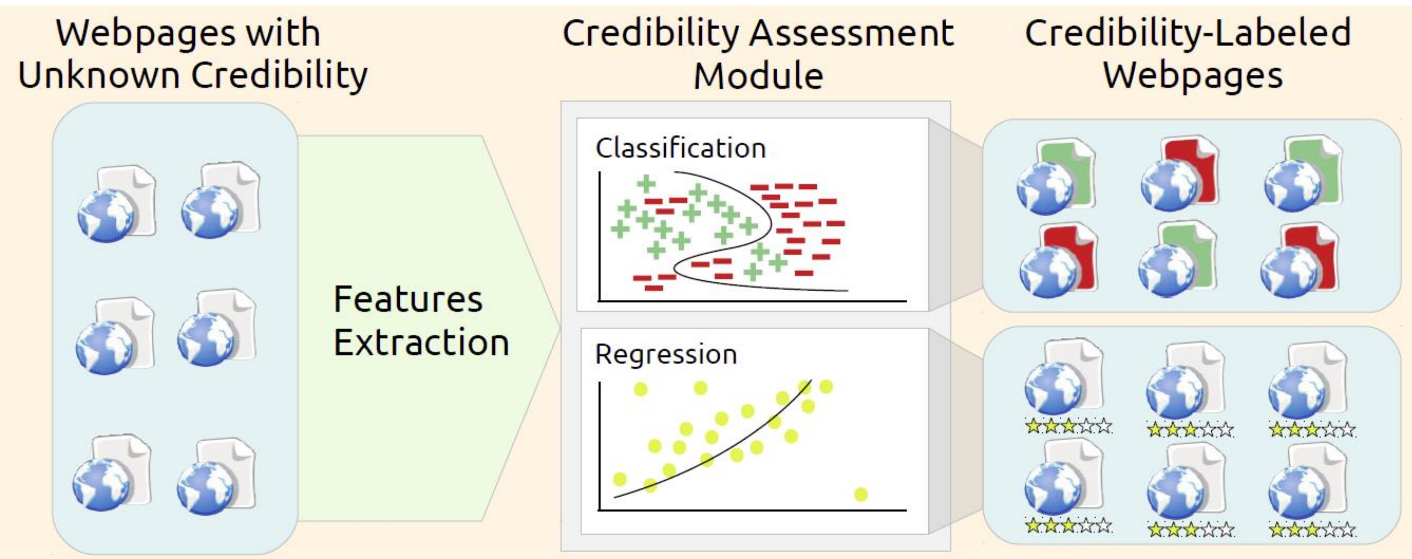
- Corpus of 1000 documents
- Evaluated by domain experts
- (prepared by MS Research)

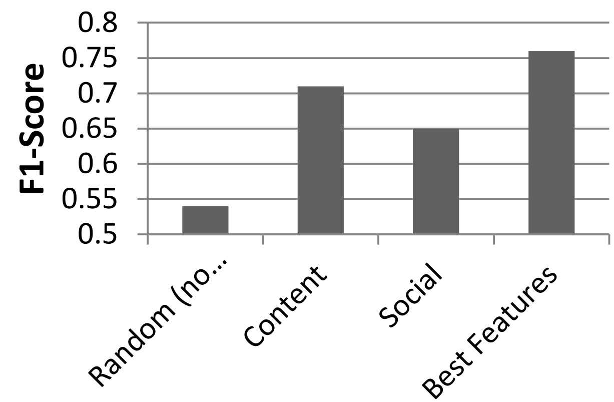
Statistical tests

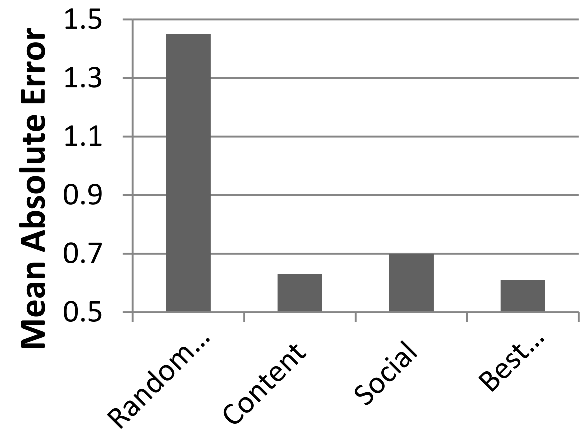
- Identification of features providing the signals on credibility

Credibility Evaluation

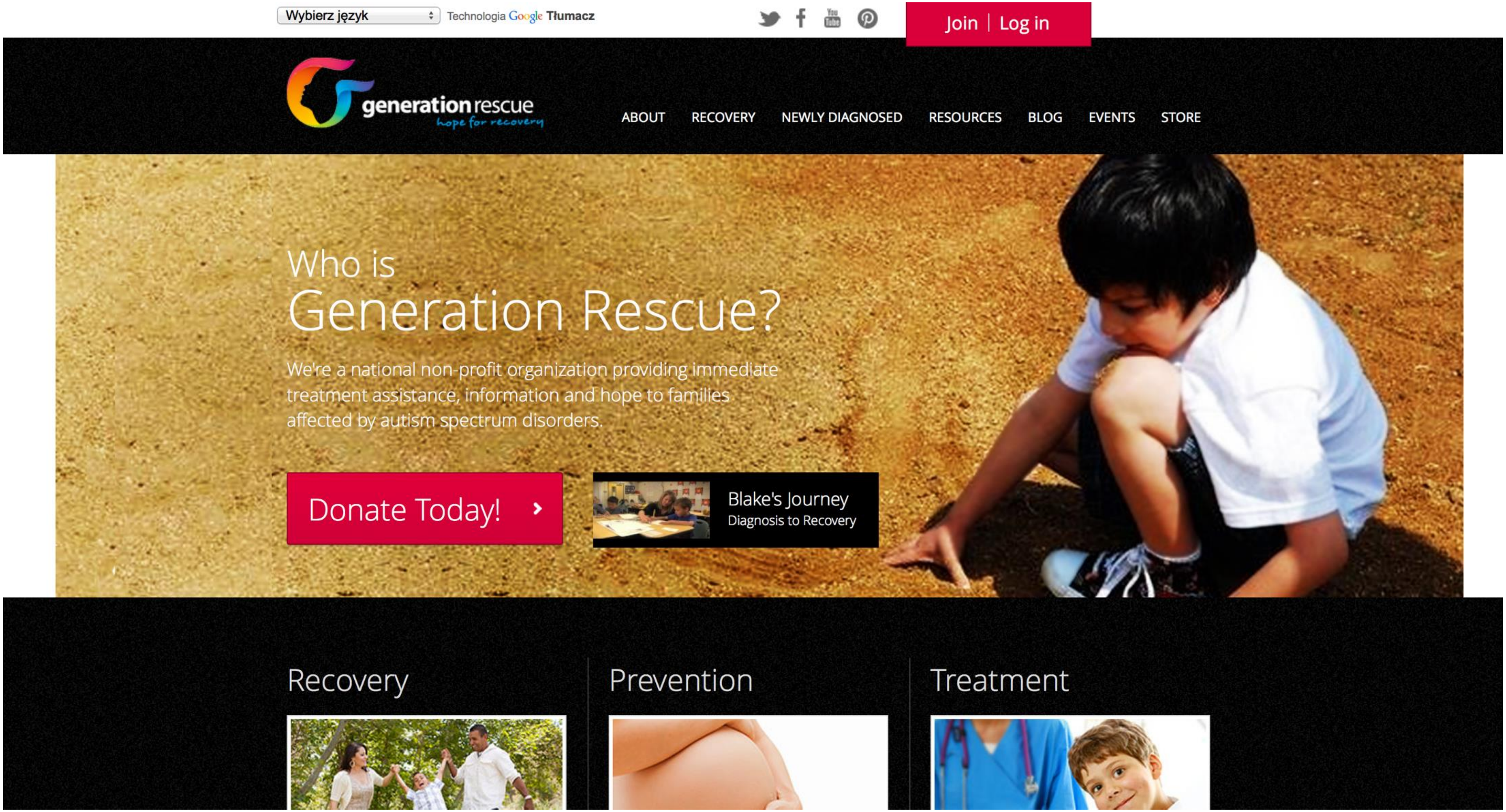
The content of a webpage as well as the social popularity offer signals for credibility





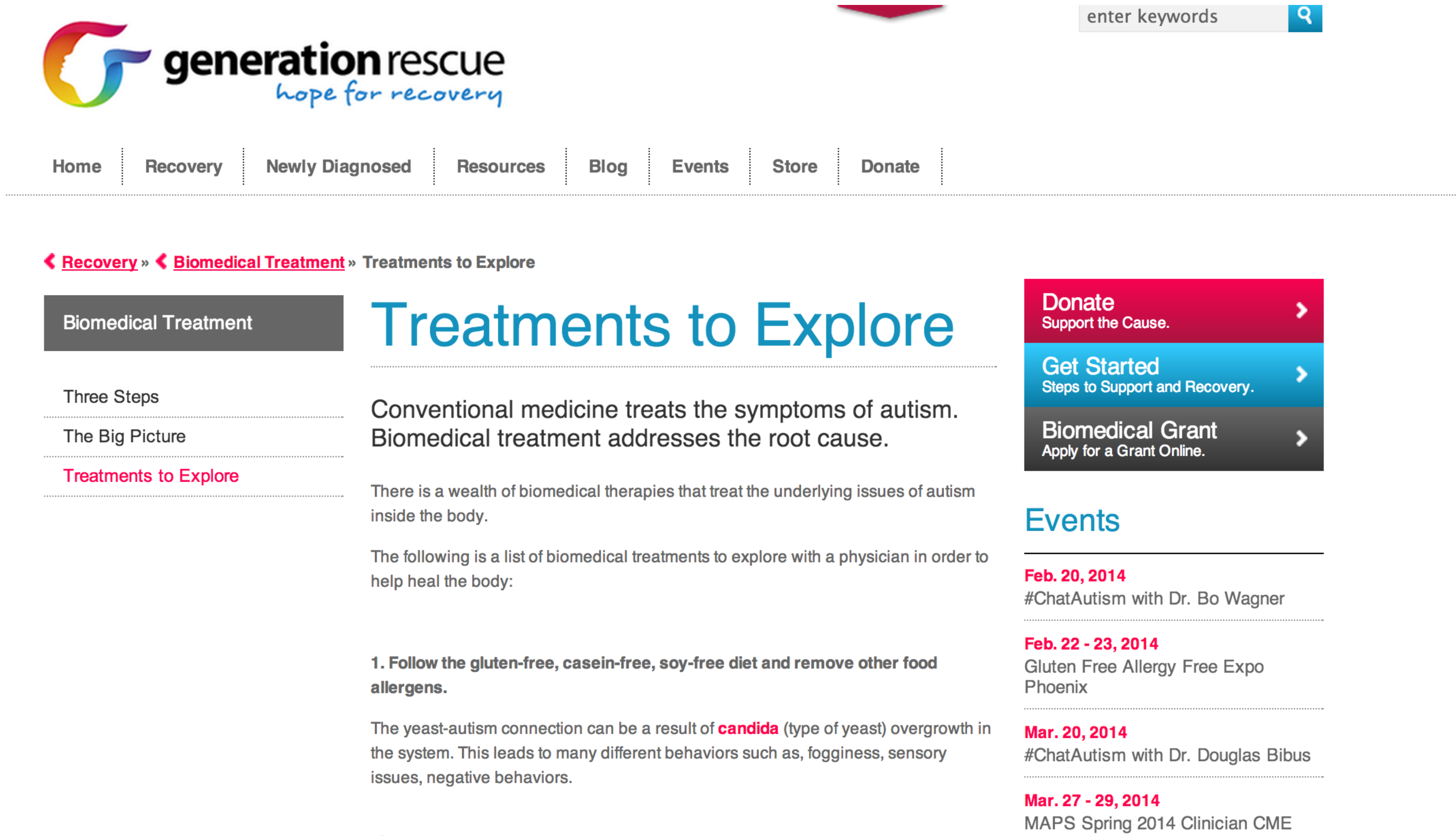


A credible source?



Distributed Information Systems Laboratory

Analysing the content



Not so credible statements

1. Follow the gluten-free, casein-free, soy-free diet and remove other food allergens.

The yeast-autism connection can be a result of **candida** (type of yeast) overgrowth in the system. This leads to many different behaviors such as, fogginess, sensory issues, negative behaviors.

"Gluten and Dairy seem to affect a lot of our children with autism and thus we see a lot of children respond terrifically when these are removed from the diet. The goal behind changing diets is to remove chemicals, toxins and potential neurotransmitters, which are liberated when food are broken down. These substances could be toxic for the brain and cause behavioral trouble in kids who are sensitive. Whether kids test as allergic or not, often they are causing a negative effect on the child and they must be removed. Each child has his or her own set of sensitivities that he or she can't deal with properly. When we change their diets, 80 percent of the kids with autism seem to respond." - Dr. Jerry Kartzinel, from "**Healing and Preventing Autism**" by Jenny McCarthy and Dr. Jerry Kartzinel.

- Effectiveness of the **gluten-free, casein-free diet** for children diagnosed with autism spectrum disorder: based on parental report.
- **Nutrition Guide** and how to implement the GFCF diet;
- Dr. Jerry's blog on why to implement the gluten free, casein free diet - **Parts 1** & **Part 2**

More Resources:

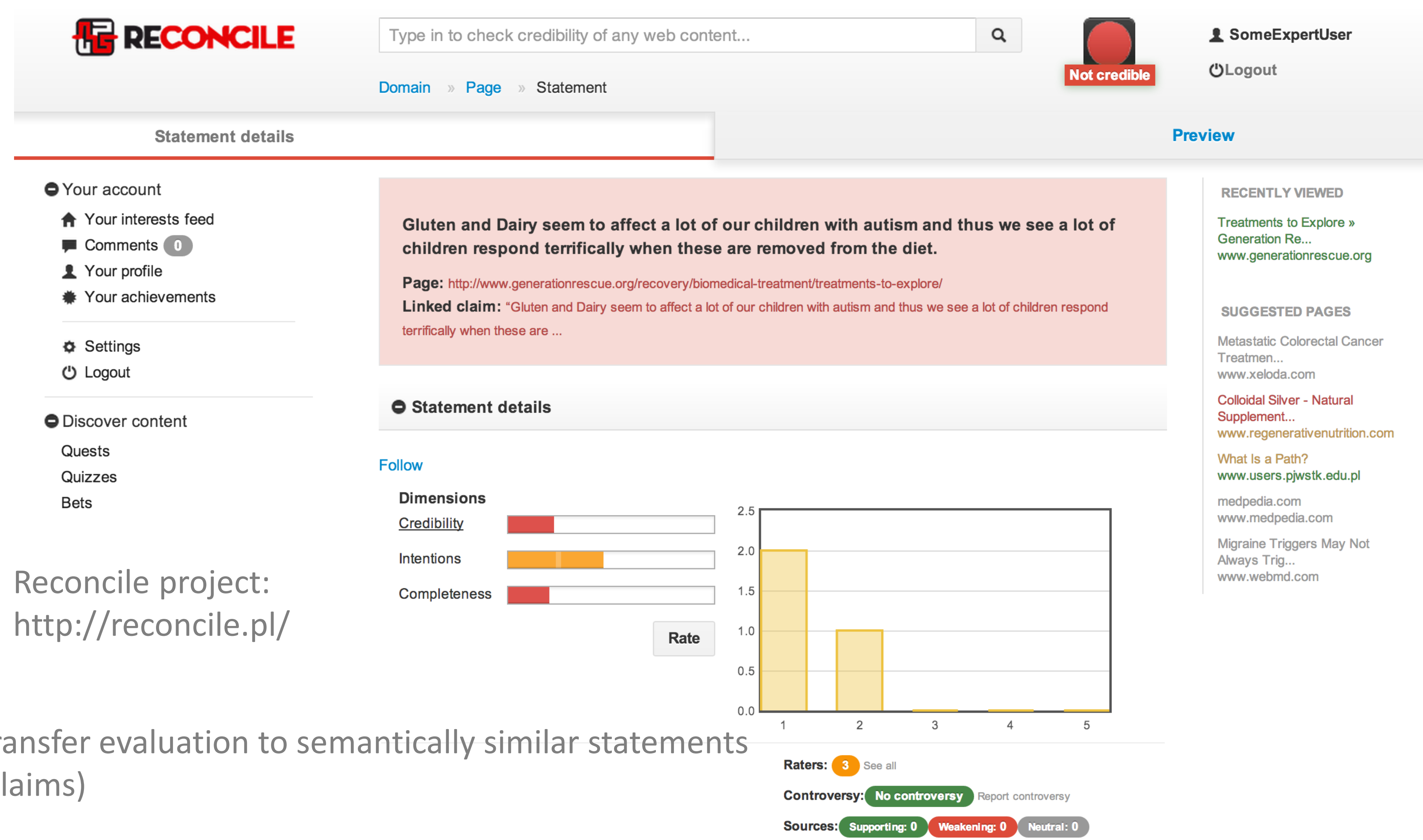
- **GFCFDiet.com**
- **The role of Clostridia and Autism**
- **The Yeast Problem and Bacteria By-products**
- **Improved Diet Helps Children with Autism**

Feb. 22 - 23, 2014
Gluten Free Allergy Free Expo Phoenix

Mar. 20, 2014
#ChatAutism with Dr. Douglas Bibus

Mar. 27 - 29, 2014
MAPS Spring 2014 Clinician CME Training Conference

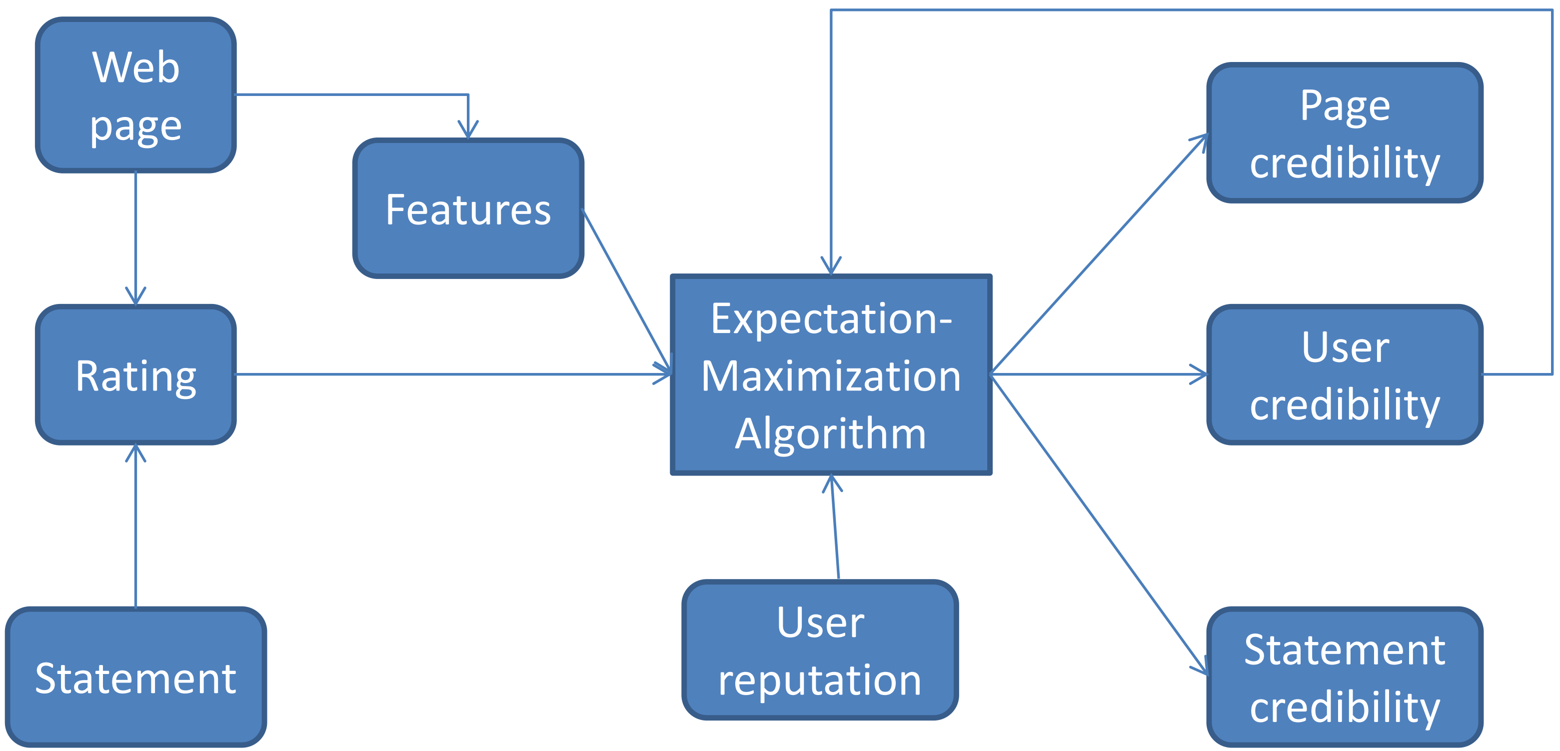
Evaluation of statements



Reconcile project:
http://reconcile.pl/

Transfer evaluation to semantically similar statements (claims)

The data fusion algorithm



Conclusions 1

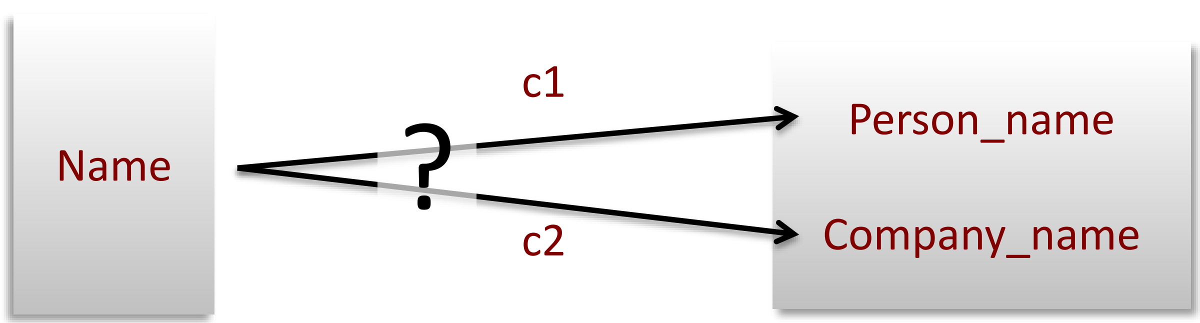
- Human evaluation is at the origin of every automated credibility evaluation task
- Same is true for any semantic or pragmatic task (e.g. translation, image labeling etc.)
- The Big Question: where is the ground truth?
- The answer: ask the crowd or experts

Supervised Learning

CASE STUDY 2: DATA INTEGRATION

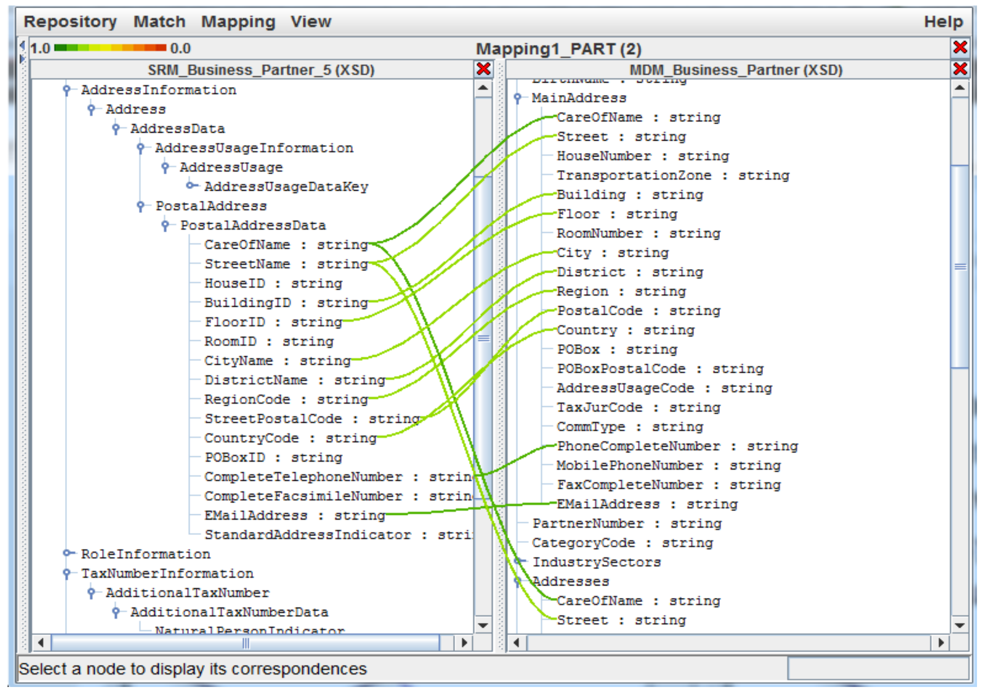
Example: Schema Matching

- Integration of heterogeneous data sources
  - Every project on Big Data analysis first has to integrate data from different, heterogeneous data sources
  - One of the long-standing open problems in data management (both industry and research)
- How to find good "matches"?
- How to choose the "best matches"?



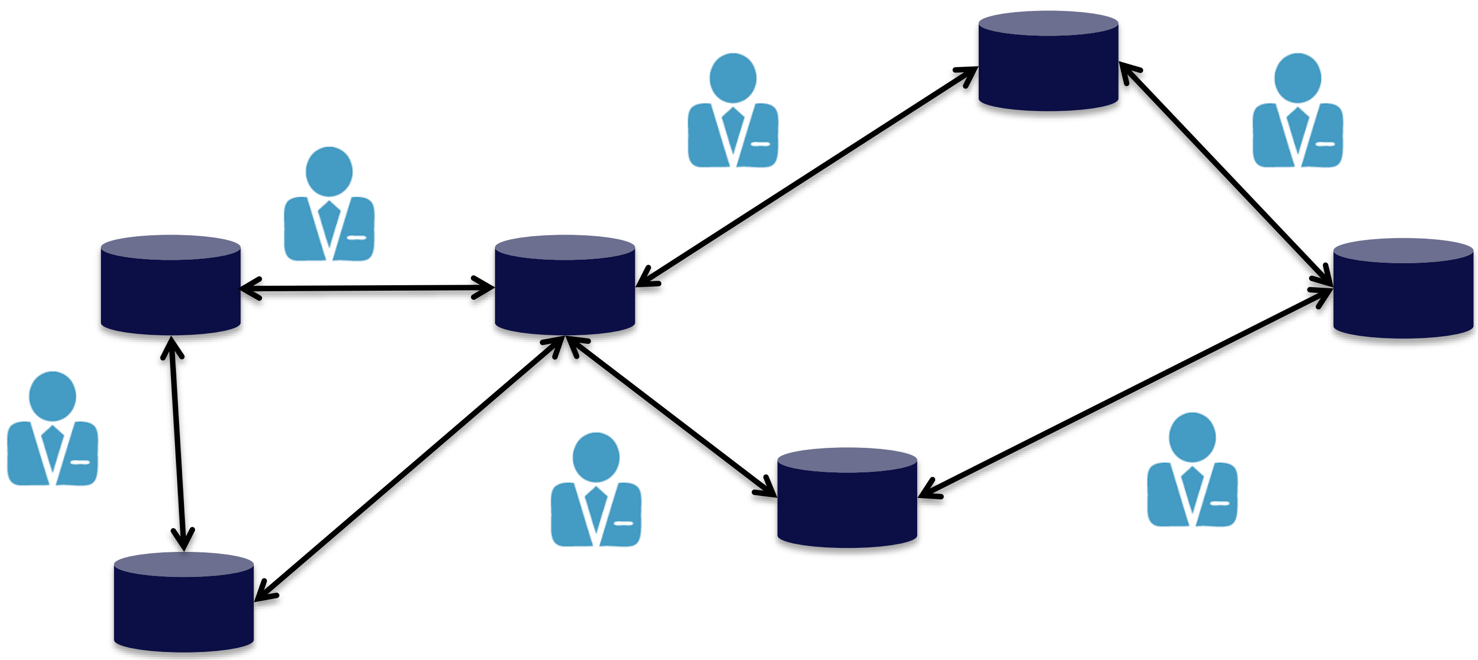
Approaches for Identifying Correspondences

- **Manual matching**
  - still common practice today
- **Schema matching tools**
  - Based on structural and content features
    - names, domains, structure, values, …
  - Establish correspondences and rank according to quality
    - Errors are frequent and unavoidable
    - Works well for small schemas



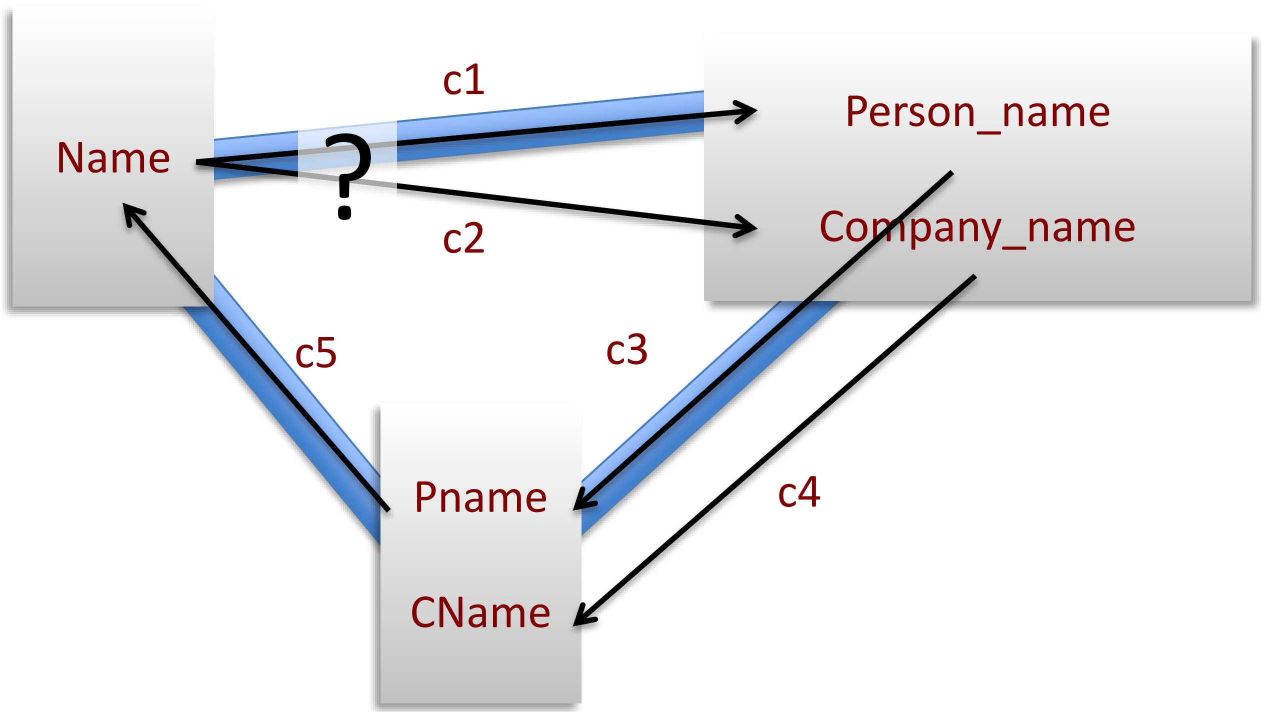
Wisdom of the Network

Data integration networks:
different experts may contribute partial matches



Wisdom of the Network

Which one would you choose?



Instead of considering only one mapping, consider whole networks of mappings: **leverage knowledge from the network!**

Probabilistic Reasoning

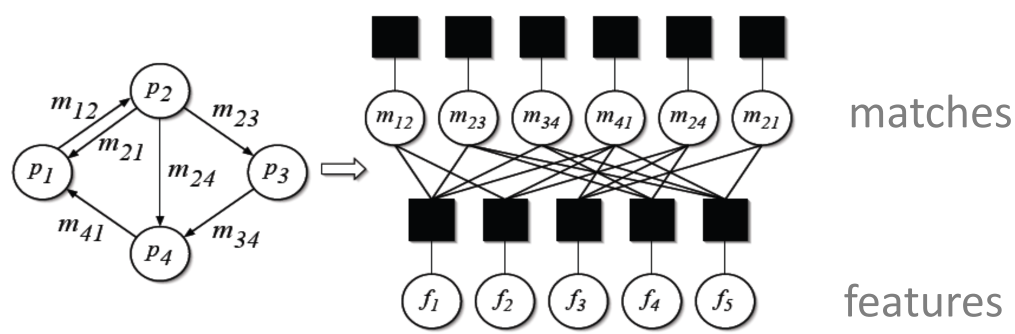
- By combining different matches in a network we can construct evidence for the correctness of those matches
  - For example, a matching contributing to a “bad cycle” less likely to be correct
- Idea: combine all this evidence and use probabilistic reasoning to select the most likely matchings

variable x to local factor f:

$$\mu_{x\to f}(x) = \prod_{h\in n(x)\setminus\{f\}} \mu_{h\to x}(x)$$

local factor f to variable x

$$\mu_{f\to x}(x) = \sum_{\sim\{x\}}\left(f(X)\prod_{y\in n(f)\setminus\{x\}}\mu_{y\to f}(y)\right)$$

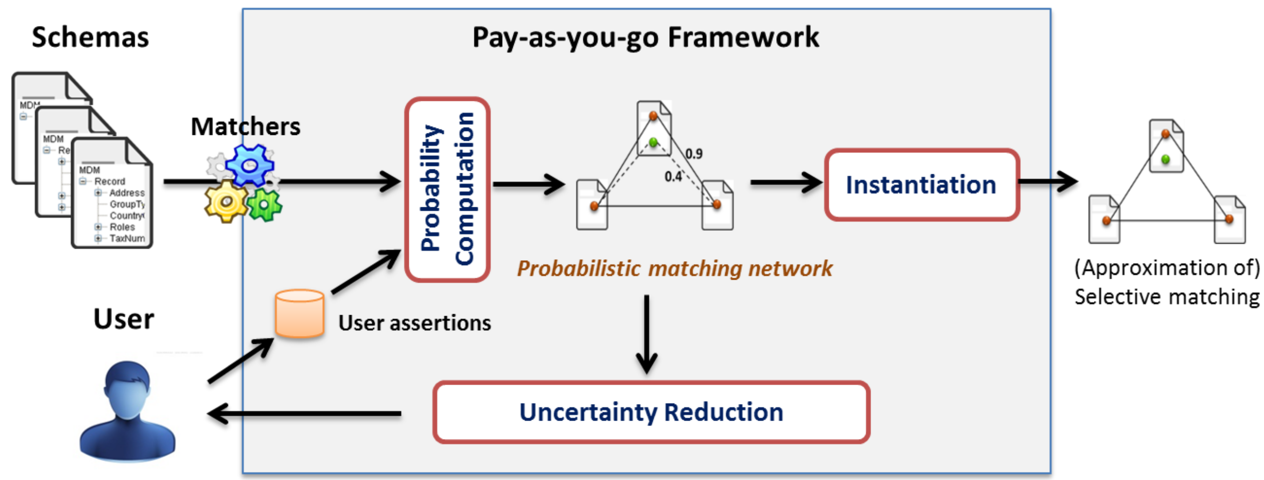


P. Cudré-Mauroux, K. Aberer and A. Feher. Probabilistic Message Passing in Peer Data Management Systems, ICDE 2006.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Distributed Information Systems Laboratory

Empowering the User

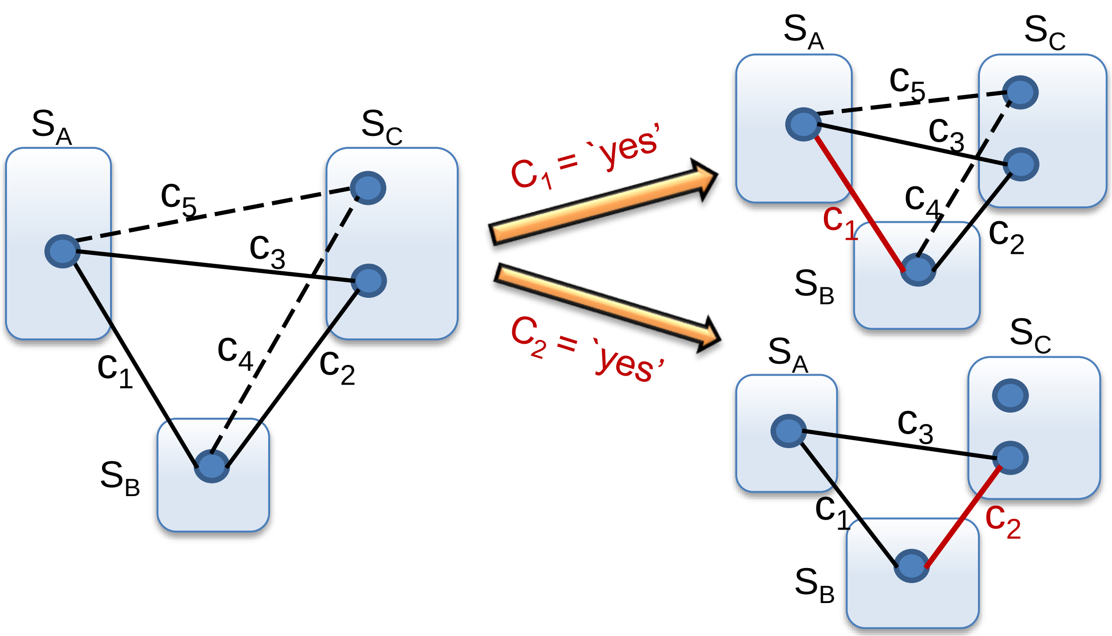
- Probabilistic reasoning results in reasonable improvement of matching quality, but
  - a posteriori analysis can only identify potentially bad choices by experts, but not correct them
- Better approach
  - Let experts make better local decisions by providing them information on global consistency and asking targeted questions



Q. V. H. Nguyen, T. Nguyen Thanh, Z. Miklos, K. Aberer and A. Gal et al. Pay-as-you-go Reconciliation in Schema Matching Networks. ICDE 2014.

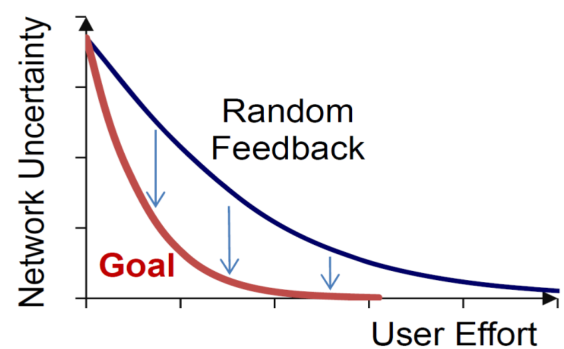
Minimal Effort User Feedback?

- Asking the right questions is important



Two possible solutions:
{c1,c2,c3} and {c1,c4,c5}

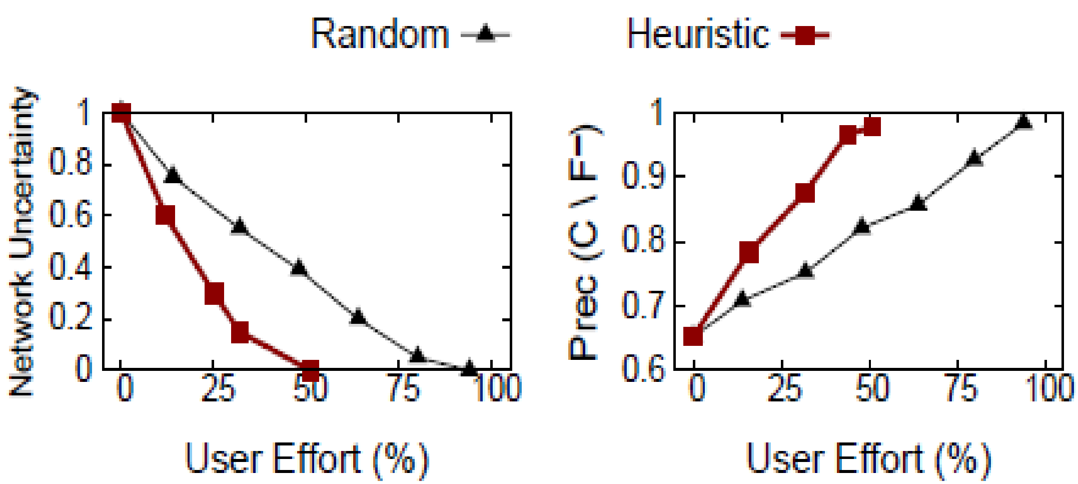
- Ask c_1 first
  - → the network is unchanged
  - → no uncertainty reduction.
- Ask c_2 first
  - → only 1 solution left
  - → the network becomes certain.

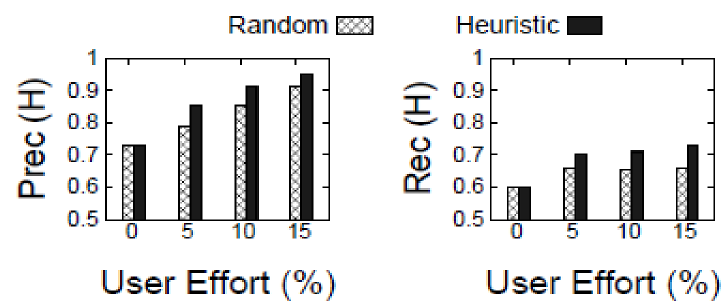


- Idea: optimize information gain with each question

Results

- Information gain ordering strategy achieves savings of up to 48% user effort compared to random ordering
- Outperforms the baseline with an average difference of 15% (precision) and 14% (recall)





Conclusion 2

- Data Integration is a task that combines human and machine intelligence
- The Big Question: How to minimize human effort and maximize information gain?

Active Learning

CASE STUDY 3: SOCIAL MEDIA ANALYSIS

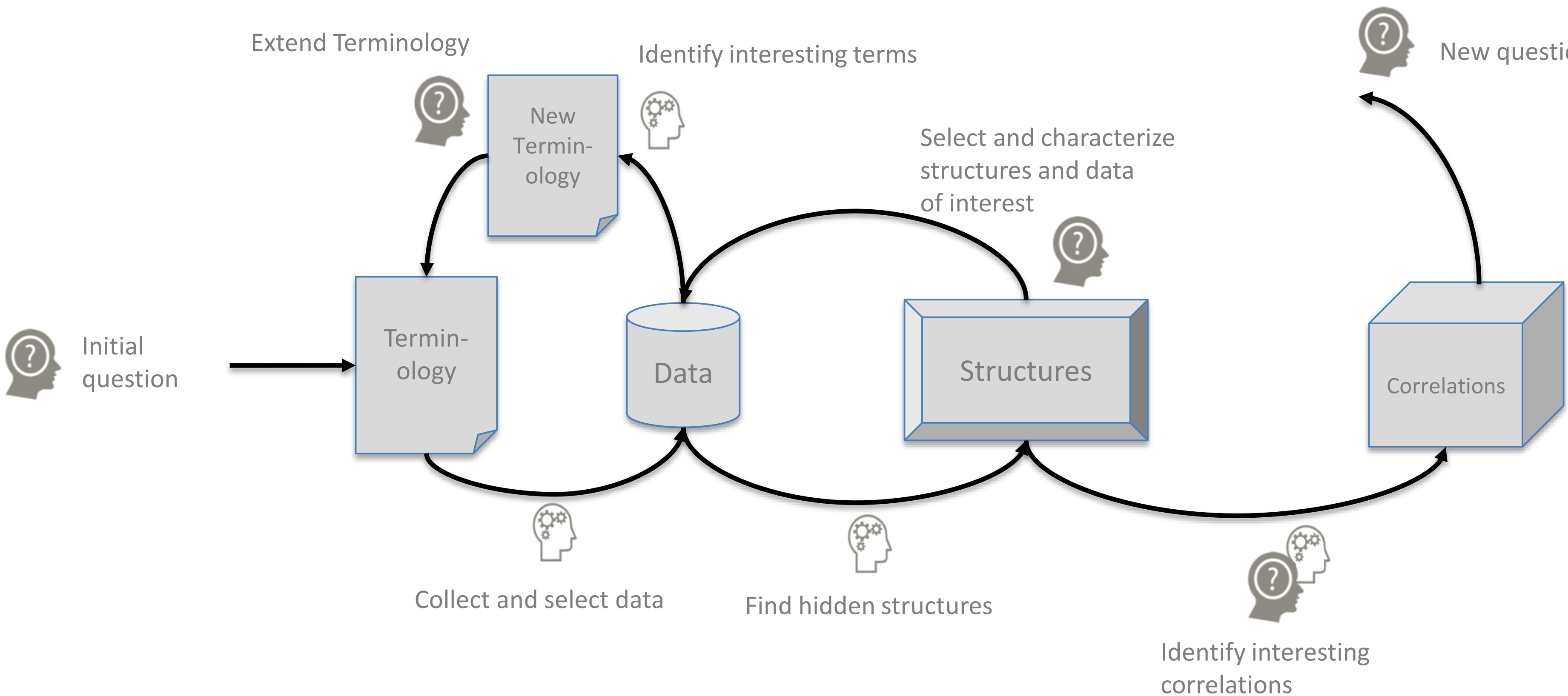
The problem

- Social Media (e.g. Twitter) contains many (hidden) signals on the public perception of issues of general interest
  - nutrition, health, politics, environment etc.
- Goal: identify influencers, their communities, their topics of interest and their stance towards given issues
- Methods
  - Semantic content analysis to capture and classify relevant content
  - Social network analysis to capture and analyze social influence

Typical analysis process

1. Describe the interest (keywords, users, time, geographic)
2. Select (or collect) the data
3. Extract the key Concepts, Entities and Categories
4. Identify Topics and Communities
5. Select relevant Issues, Influencers and Events
6. Produce insights (correlations)

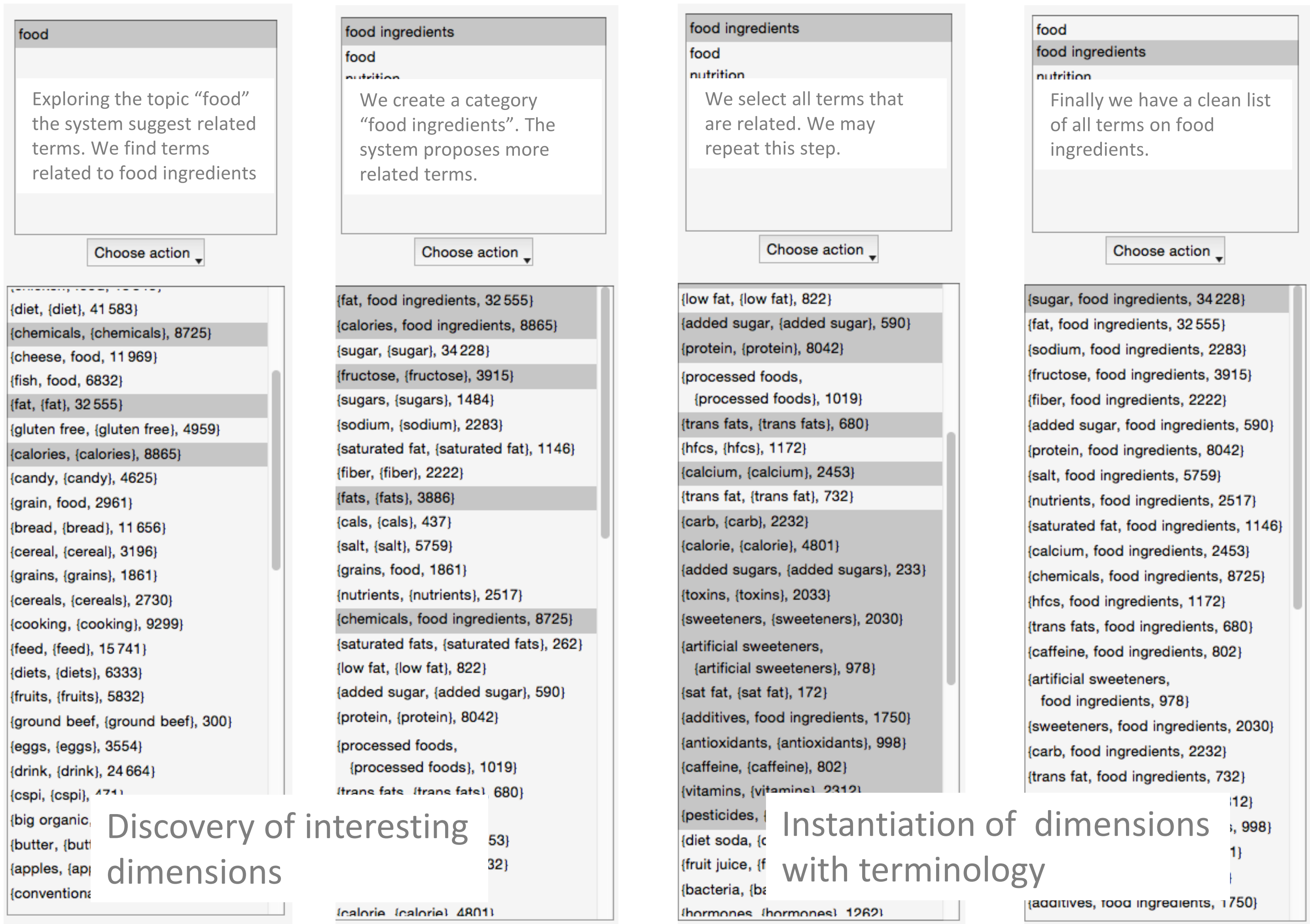
Human-Machine Interaction



Creating Terminology

Input:
~50 Mio tweets

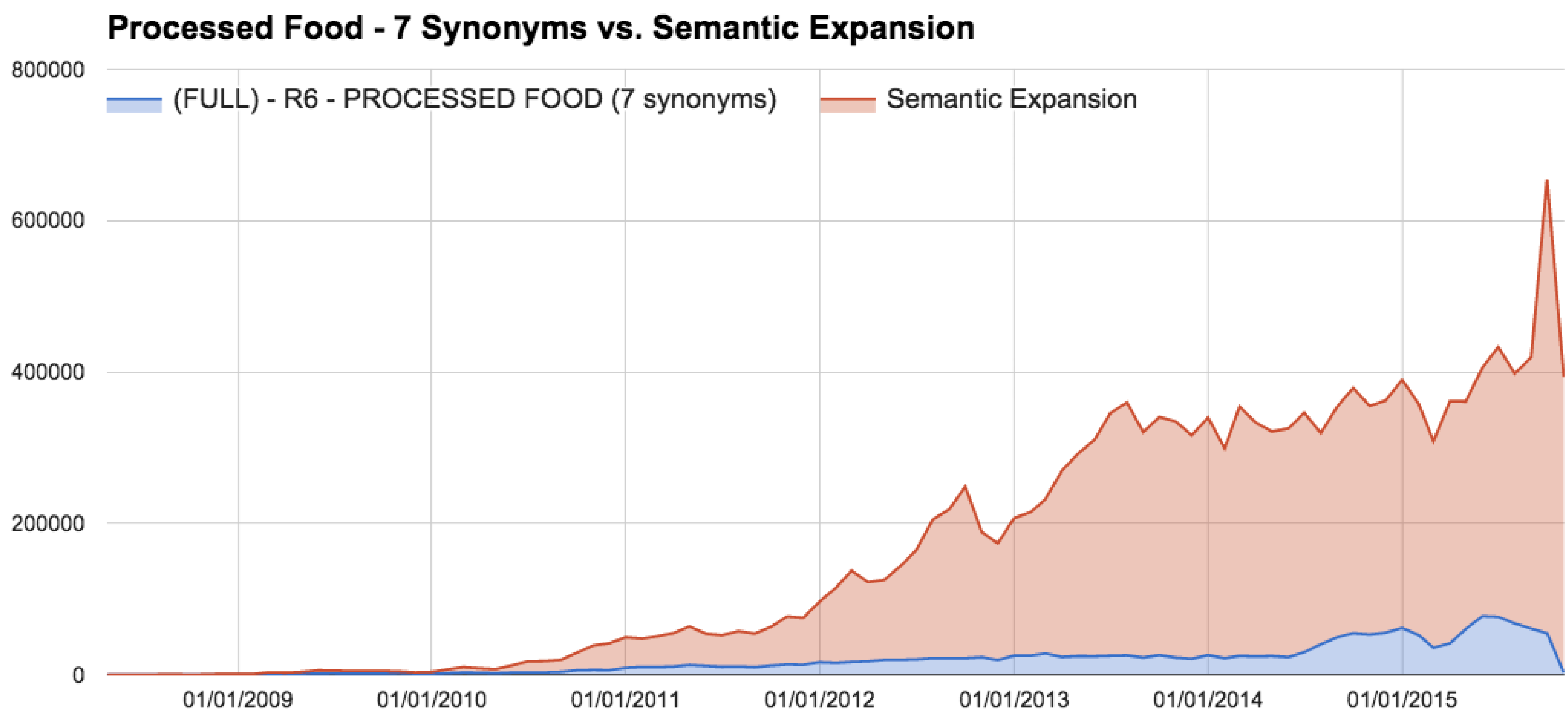
Analysis method:
Word embedding (word2vec)



ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Distributed Information Systems Laboratory

Data Collection



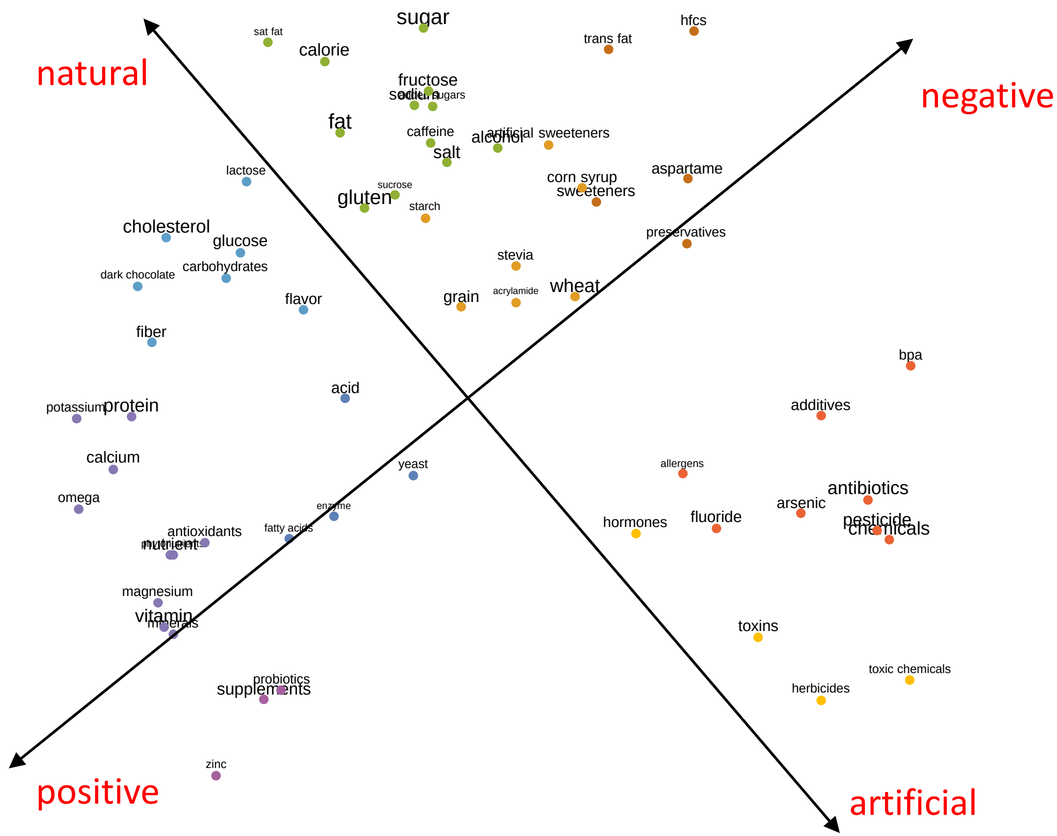
Using a semantically expanded terminology increases coverage significantly!

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Distributed Information Systems Laboratory

Categorization

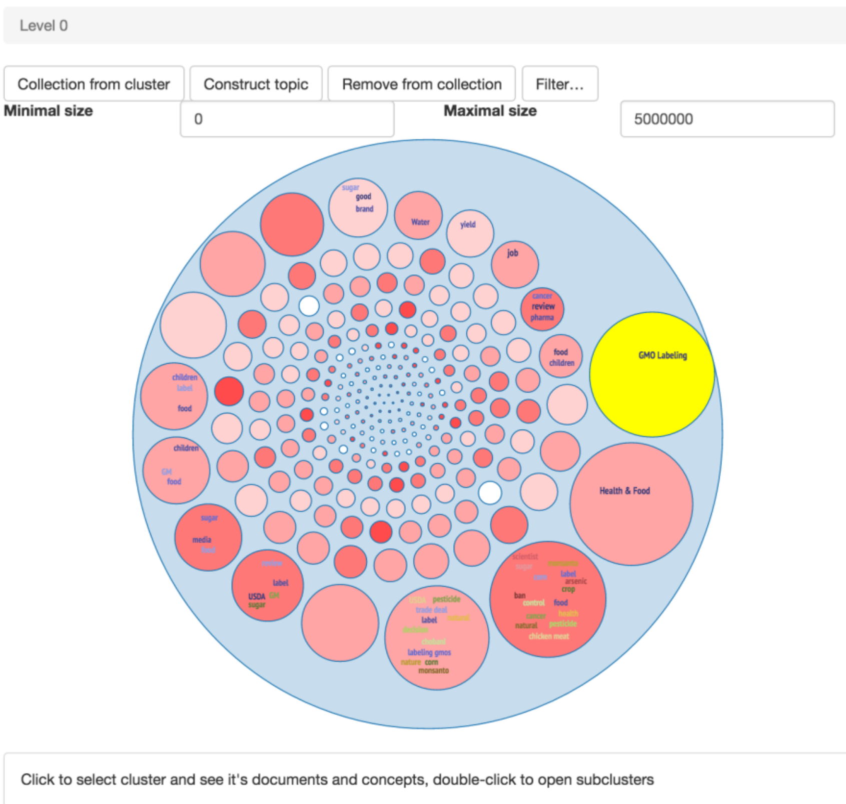
- The system clusters the terms on food ingredients according to similarity
- The expert sees
  - A clear distinction between positive and negative terms
  - Distinction between natural and artificial ingredients
  - Clusters of related terms, e.g. vitamins, additives etc.
- We may use this to create sub-categories of interest



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Distributed Information Systems Laboratory

Social Clustering

Analyzing social interactions we can identify clear communities



Info | Users(28101) | Tweets(81720) | Documents(7274) | Concepts(100) | Hashtags(100) | Keywords(205)

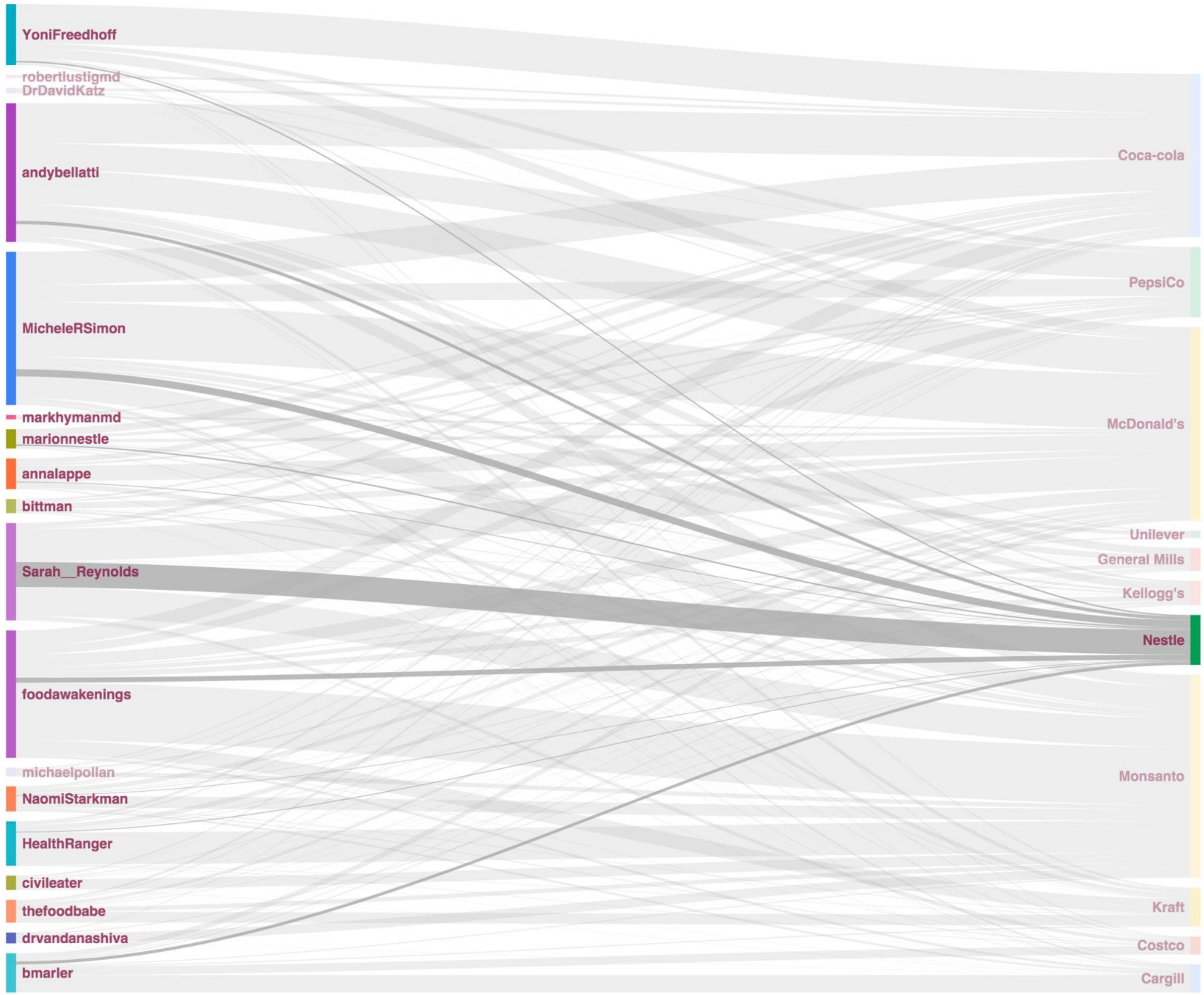
#	Name	Description	Tweets	Retweets	Influence (RT+RE+Mentions)
1	8extremes	Sharing Peace & Understanding with Compassion & Gratitude. Having Love & Respect for all living things. Susan Elaine Los Angeles CA	872	214	2367
2	GMOFreeUSA	We are a national group, educating consumers about the potential hazards of Genetically Engineered foods.	78	25	1679
3	OrganicLiveFood	We provide info regarding #organic/sustainable foods, nutritional facts 4improving #health, #raw foods, #superfoods, #herbs, #GMO harms, #bees & #pesticides	606	28	1439
4	RachelsNews	KIDS RIGHT TO KNOW(Founder)16 DEBATED Shark/Dragon https://t.co/pn1AzZnHKI TEDxTO 2014-SPEAKER https://t.co/3OklmzLnch Huffington Blogger, Teen Earth Activist	1347	223	1412
5	kevinfolta	Land-grant scientist exploring ways to make better food with le... to effective	394	11	1363
6	TheGOPJesus	Politicians ... special pla... maybe not			
7	GMWatch	Countering			
8	MonsantoCo	Monsanto i... others to a... sustain the			
9	geneticmaize	Mom of Ad... https://t.co... sustainable ag. Former US Army public health. Ecomodernist. Words are mine.			
10	SSF_BERF_DEFM	SSF = Support Small Farms -- BERF = By Eating Real Food -- DEFM = Don't Eat Factory Meat	914	0	888

For each community we can identify
- Their influencers
- Their main concepts
- Potentially new interesting terminology

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Distributed Information Systems Laboratory

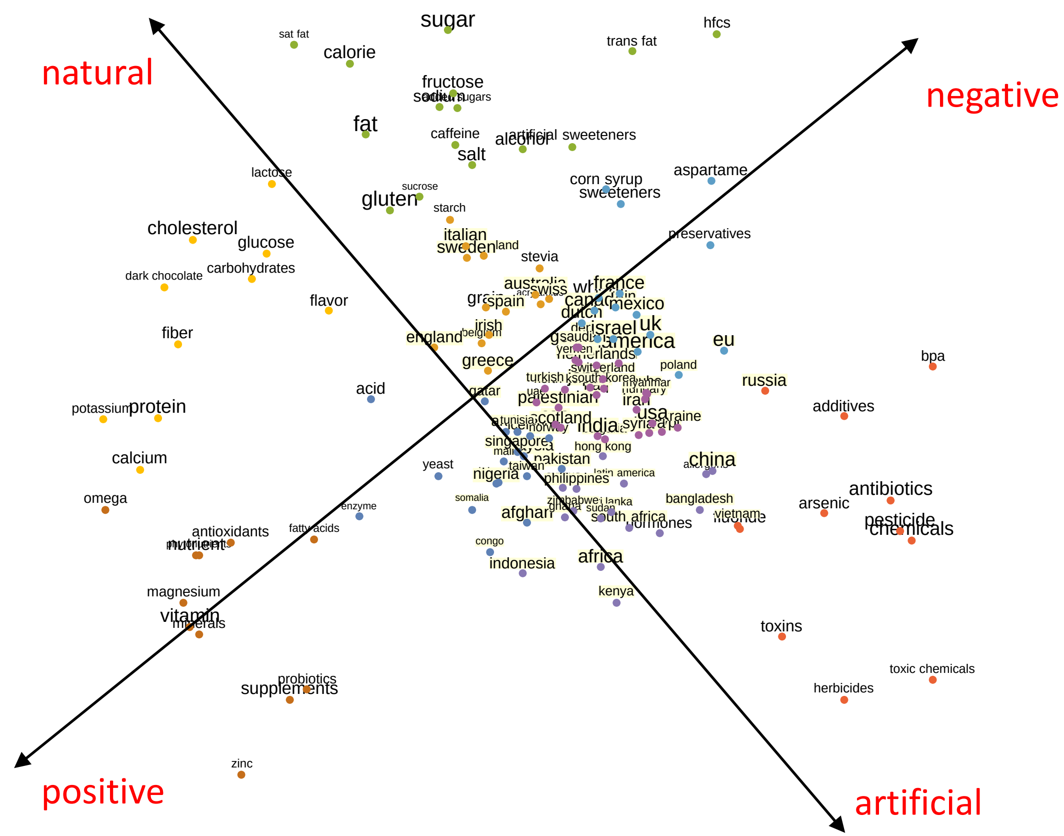
Findings: Company influencers



ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Distributed Information Systems Laboratory

Findings: role of food ingredients in different countries



Conclusions

- Machine learning applied to Big Data can reveal surprising hidden structures with valuable insights
- Big questions:
  - How to guide the machines to the right data and analysis
  - How to make the resulting structures human-interpretable

Unsupervised Learning

Summary

- Big Data has impressive potential to create insights and solve hard problems
- Human intervention in the analysis processes is essential for obtaining meaningful results
- Three main types of intervention
  - A priori: supervised learning
  - Interactive: active learning
  - A posteriori: unsupervised learning
- No one size fits all: their specific implementation depends strongly on the use case