

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 human input

For example: Google is a huge relevance feedback engine

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

<p>A person riding a motorcycle on a dirt road.</p> 	<p>Two dogs play in the grass.</p> 	<p>A skateboarder does a trick on a ramp.</p> 	<p>A dog is jumping to catch a frisbee.</p> 
<p>A group of young people playing a game of frisbee.</p> 	<p>Two hockey players are fighting over the puck.</p> 	<p>A little girl in a pink hat is blowing bubbles.</p> 	<p>A refrigerator filled with lots of food and drinks.</p> 
<p>A herd of elephants walking across a dry grass field.</p> 	<p>A close up of a cat laying on a couch.</p> 	<p>A red motorcycle parked on the side of the road.</p> 	<p>A yellow school bus parked in a parking lot.</p> 
Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image

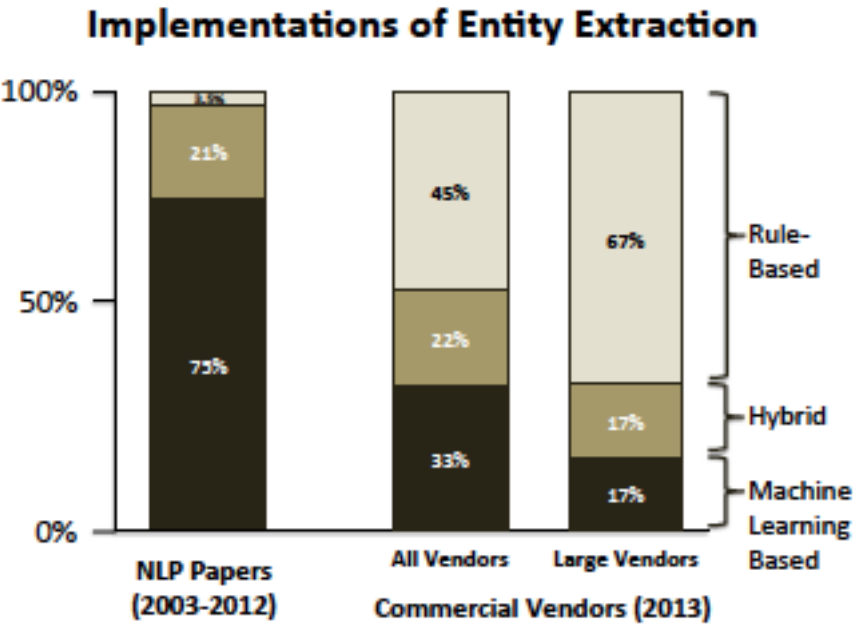
Vinyals, Oriol, et al. "Show and tell: A neural image caption generator."
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

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!

- 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.

Three case studies

1. Web credibility

How human input enables machine learning

2. Data integration

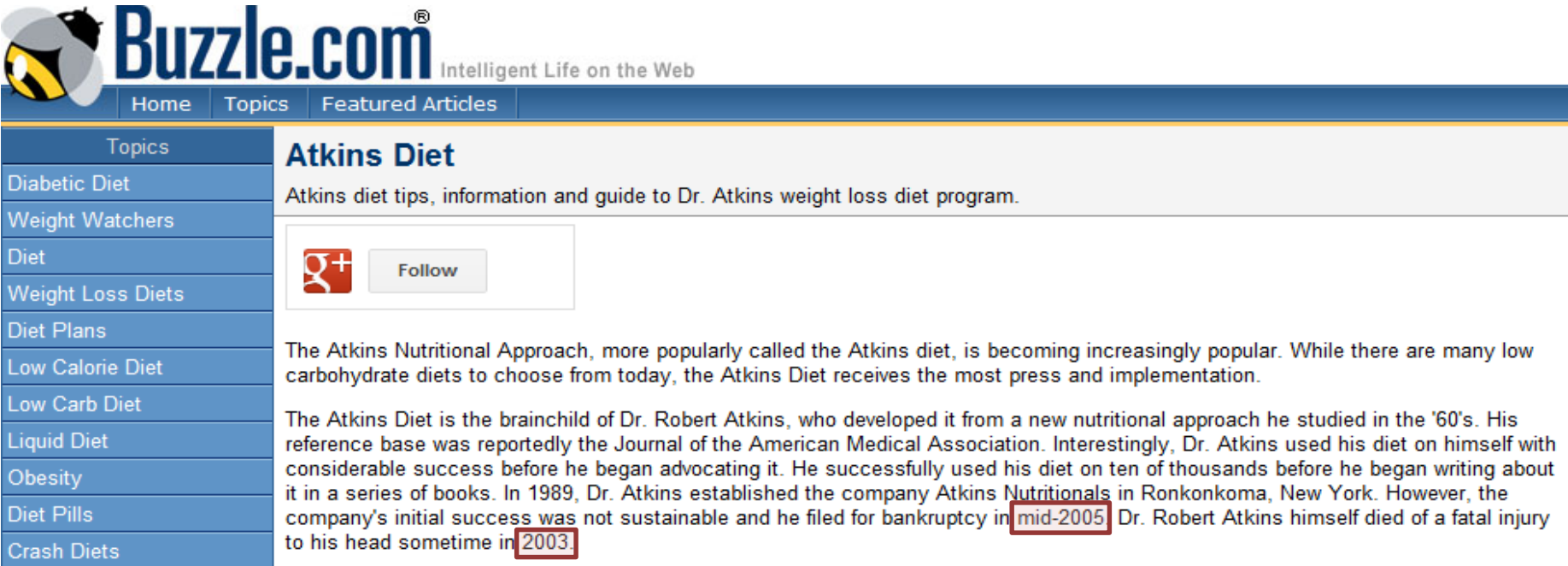
How humans and machines cooperate efficiently in a problem solving task

3. Social Media Analytics

How humans interpret latent structures found by machine learning

CASE STUDY 1: WEB CREDIBILITY

How to Evaluate Web Credibility?



Buzzle.com Intelligent Life on the Web


Home Topics Featured Articles

Topics

- Diabetic Diet
- Weight Watchers
- Diet
- Weight Loss Diets
- Diet Plans
- Low Calorie Diet
- Low Carb Diet
- Liquid Diet
- Obesity
- Diet Pills
- Crash Diets

Atkins Diet

Atkins diet tips, information and guide to Dr. Atkins weight loss diet program.

 Follow

The Atkins Nutritional Approach, more popularly called the Atkins diet, is becoming increasingly popular. While there are many low carbohydrate diets to choose from today, the Atkins Diet receives the most press and implementation.

The Atkins Diet is the brainchild of Dr. Robert Atkins, who developed it from a new nutritional approach he studied in the '60's. His reference base was reportedly the Journal of the American Medical Association. Interestingly, Dr. Atkins used his diet on himself with considerable success before he began advocating it. He successfully used his diet on ten of thousands before he began writing about it in a series of books. In 1989, Dr. Atkins established the company Atkins Nutritionals in Ronkonkoma, New York. However, the company's initial success was not sustainable and he filed for bankruptcy in mid-2005. Dr. Robert Atkins himself died of a fatal injury to his head sometime in 2003.

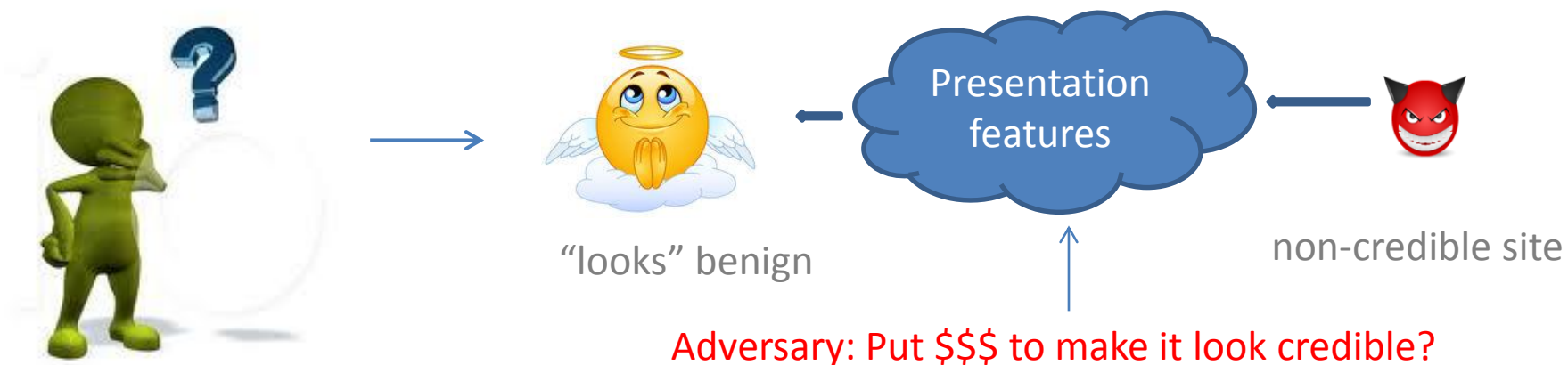
When you browse a webpage, how do you know it's content is valid and accurate?

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

The screenshot shows a web page from TIME Health & Family. The main article is titled "Breakthrough Discoveries of Alzheimer's Genes" by Alice Park, dated Monday, Sept. 07, 2009. The article discusses the discovery of three new genes (clusterin, complement receptor 1 (CR1), and PICALM) linked to Alzheimer's disease. It mentions that these genes were discovered by two separate research groups, one in Wales and one in France. The article also notes that the most common form of the memory disorder, late-onset Alzheimer's, is the type that affects patients in their 60s or later and accounts for about 90% of all Alzheimer's cases. The only other gene connected with the condition, apolipoprotein E (ApoE), was identified in 1993; since then, researchers have tirelessly hunted for other key genes, knowing that 60% to 80% of the progressive, incurable disease is genetically based. The article includes a photo of a person looking at a large map or display. Below the article, there is a "RELATED" section with links to "The Business of Caring for Seniors" and "Home Instead: The Business of Granny Care". To the right of the article, there is a sidebar with a "THE MAGAZINE" section featuring a Pope of the Americas and a "Subscribe" button. Below that, there is a "Replay" section with a video player and a "TORAY Innovation by Chemistry" logo. At the top of the page, there is a navigation bar with links to various sections like NEWSFEED, U.S., POLITICS, WORLD, BUSINESS, TECH, HEALTH, SCIENCE, ENTERTAINMENT, STYLE, SPORTS, OPINION, and PHOTOS. There is also a search bar and social media links for Facebook, Twitter, Google+, and YouTube. A box highlights the social media links.

Social Features

Fetches from third-party websites

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

Content Features

Computed from webpage content or metadata

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

Credibility Features

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



Statistical tests

- Identification of features providing the signals on credibility

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

Use of Grammar

Popularity on Twitter

Popularity on Facebook

Webpage Topic

Readability

Webpage Design

Informativeness

Google Search Ranking

Domain Type (.gov, .edu)

Use of Punctuation

Web Graph Structure

Browsing Patterns

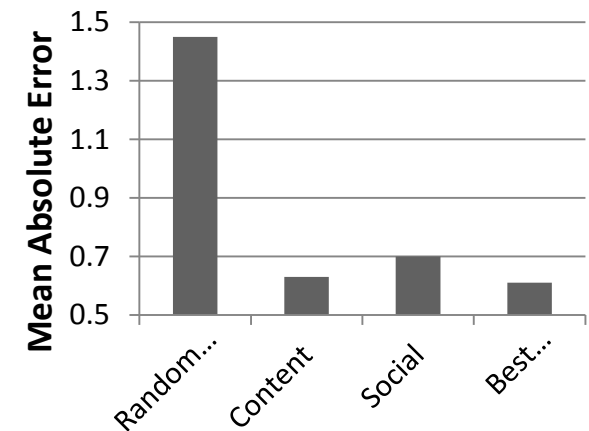
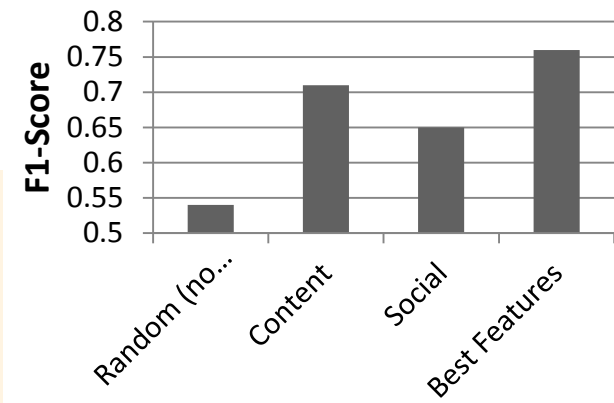
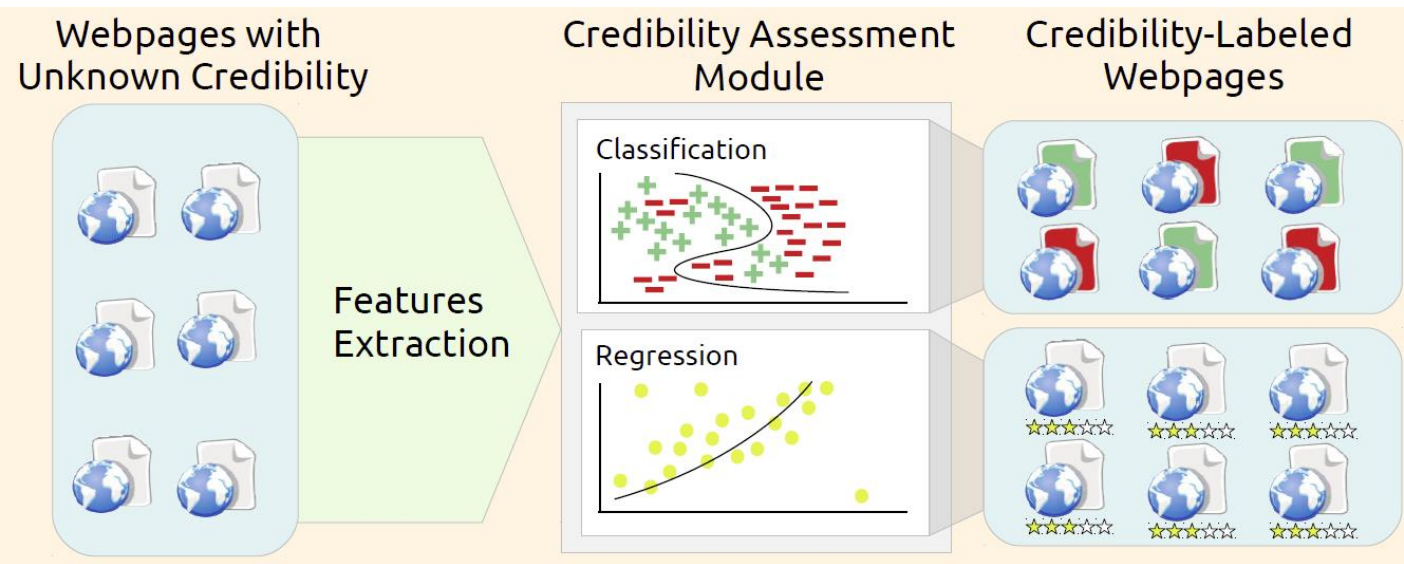
Number of Bookmarks

Ads Prominence

Objectivity

Credibility Evaluation





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




A credible source?

Wybierz język

Technologia Google Tłumacz



Join | Log in




ABOUTRECOVERYNEWLY DIAGNOSEDRESOURCESBLOGEVENTSSTORE


Who is Generation Rescue?


We're a national non-profit organization providing immediate treatment assistance, information and hope to families affected by autism spectrum disorders.


Donate Today! >




Blake's Journey
Diagnosis to Recovery



Recovery

Prevention

Treatment



enter keywords

Q

- Home
- Recovery
- Newly Diagnosed
- Resources
- Blog
- Events
- Store
- Donate

◀ [Recovery](#) » ◀ [Biomedical Treatment](#) » [Treatments to Explore](#)

- Biomedical Treatment
- Three Steps
- The Big Picture
- [Treatments to Explore](#)

Treatments to Explore

Conventional medicine treats the symptoms of autism.
Biomedical treatment addresses the root cause.

There is a wealth of biomedical therapies that treat the underlying issues of autism inside the body.

The following is a list of biomedical treatments to explore with a physician in order to help heal the body:

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

- Donate
Support the Cause.
- Get Started
Steps to Support and Recovery.
- Biomedical Grant
Apply for a Grant Online.

Events

- Feb. 20, 2014
#ChatAutism with Dr. Bo Wagner
- 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

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

Sponsors

Supporting you



Evaluation of statements

Type in to check credibility of any web content...

Q

Not credible

SomeExpertUser

Logout

Domain » Page » Statement

Statement details

Preview

Your account

Home

Your interests feed

Comments 0

Profile

Achievements

Settings

Logout

Discover content

Quests

Quizzes

Bets

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.

Page: <http://www.generationrescue.org/recovery/biomedical-treatment/treatments-to-explore/>

Linked claim: "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 ..."

Statement details

Follow

Dimensions

Credibility

Intentions

Completeness

Rate

Bar chart showing evaluation rates for 1 to 5 stars. The y-axis ranges from 0.0 to 2.5. The x-axis shows 1, 2, 3, 4, and 5 stars. The bars show approximately 2.0 for 1 star, 1.0 for 2 stars, and very low values for 3, 4, and 5 stars.

Raters: 3 See all

Controversy: No controversy Report controversy

Sources: Supporting: 0 Weakening: 0 Neutral: 0

RECENTLY VIEWED

Treatments to Explore » Generation Re...
www.generationrescue.org

SUGGESTED PAGES

Metastatic Colorectal Cancer Treatmen...
www.xeloda.com

Colloidal Silver - Natural Supplement...
www.regenerativenutrition.com

What Is a Path?
www.users.pjwstk.edu.pl

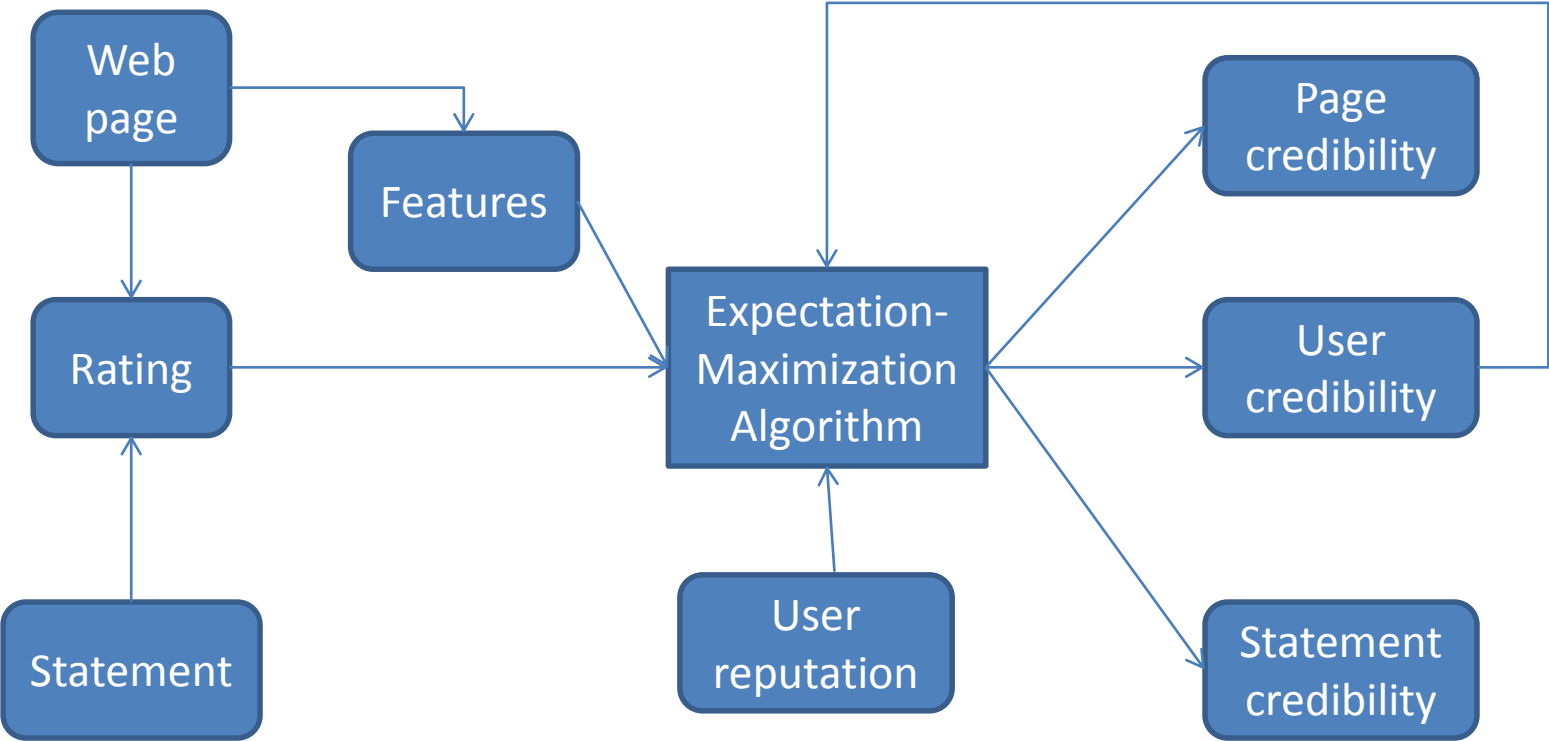
medpedia.com
www.medpedia.com

Migraine Triggers May Not Always Trig...
www.webmd.com

Reconcile project:
<http://reconcile.pl/>

Transfer evaluation to semantically similar statements (claims)

The data fusion algorithm



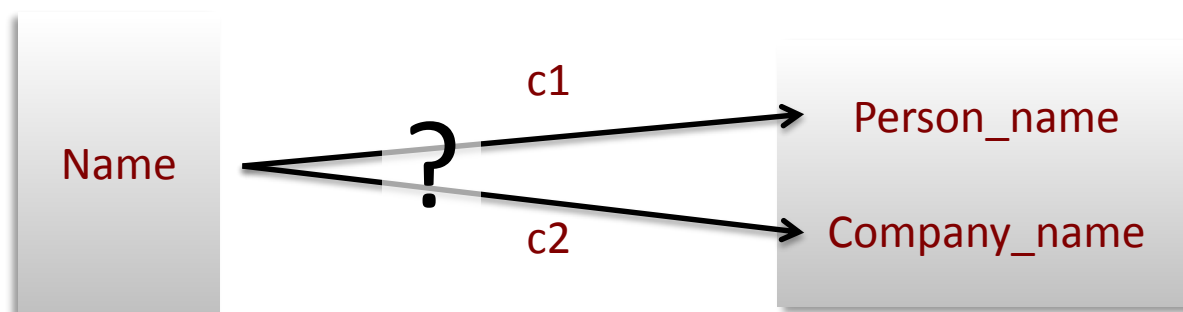
- Human evaluation is at the origin of every automated credibility evaluation task
- Same is true for any semantic or pragmatic task (e.g. translation, image labeling etc.)
- The Big Question: where is the ground truth?
- The answer: ask the crowd or experts

Supervised Learning

CASE STUDY 2: DATA INTEGRATION

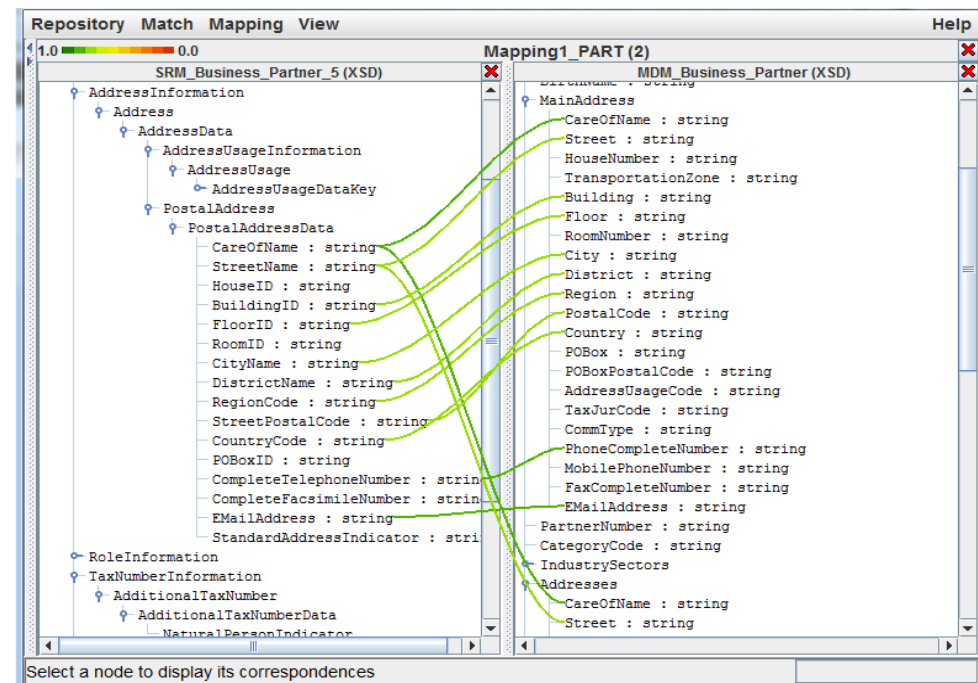
Example: Schema Matching

- 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”?

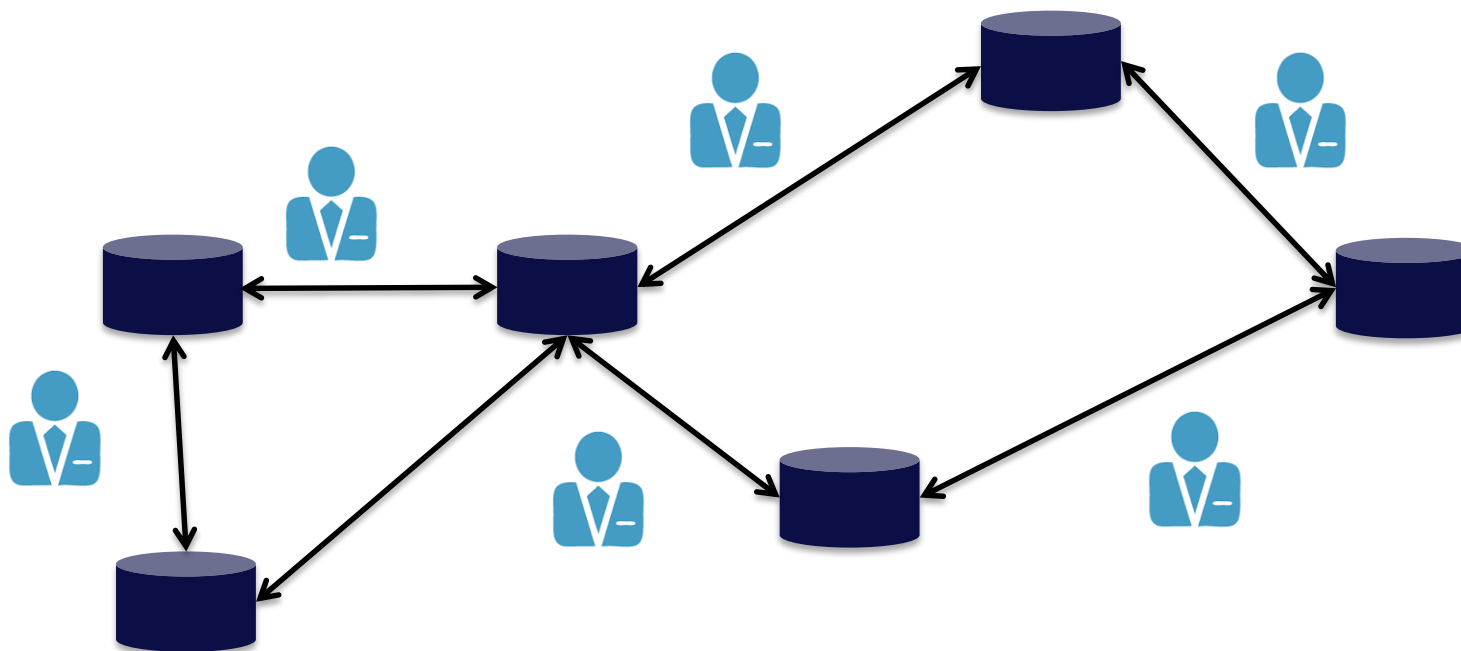


Approaches for Identifying Correspondences

- **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

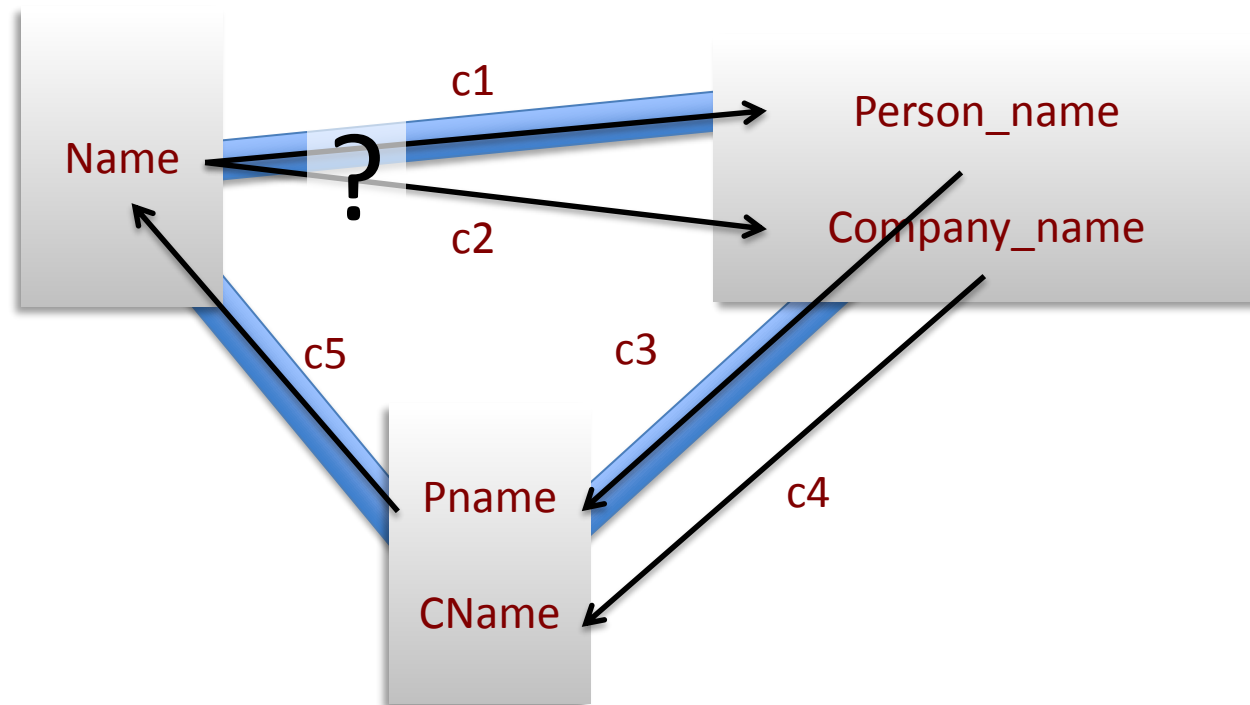


Data integration networks:
different experts may contribute partial matches



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

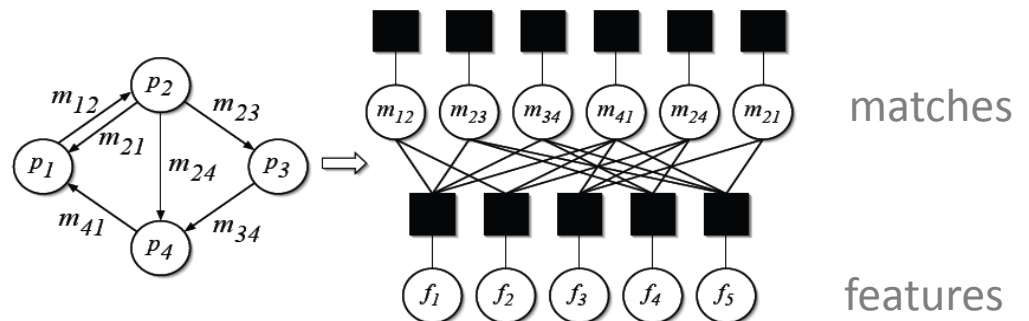
- By combining different matches in a network we can construct evidence for the correctness of those matches
 - For example, a matching contributing to a “bad cycle” less likely to be correct
- Idea: combine all this evidence and use probabilistic reasoning to select the most likely matchings

variable x to local factor f :

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

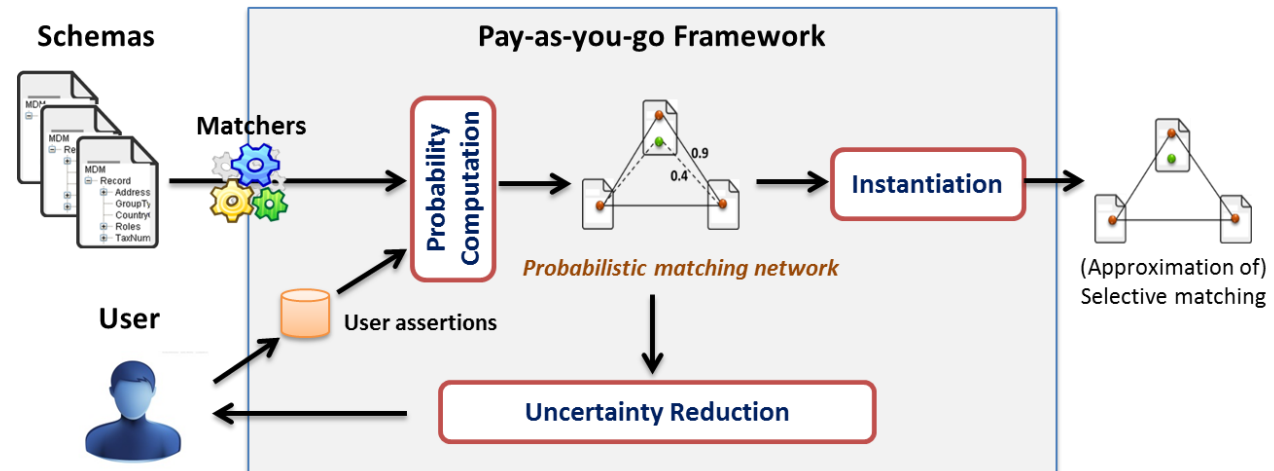
local factor f to variable x

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



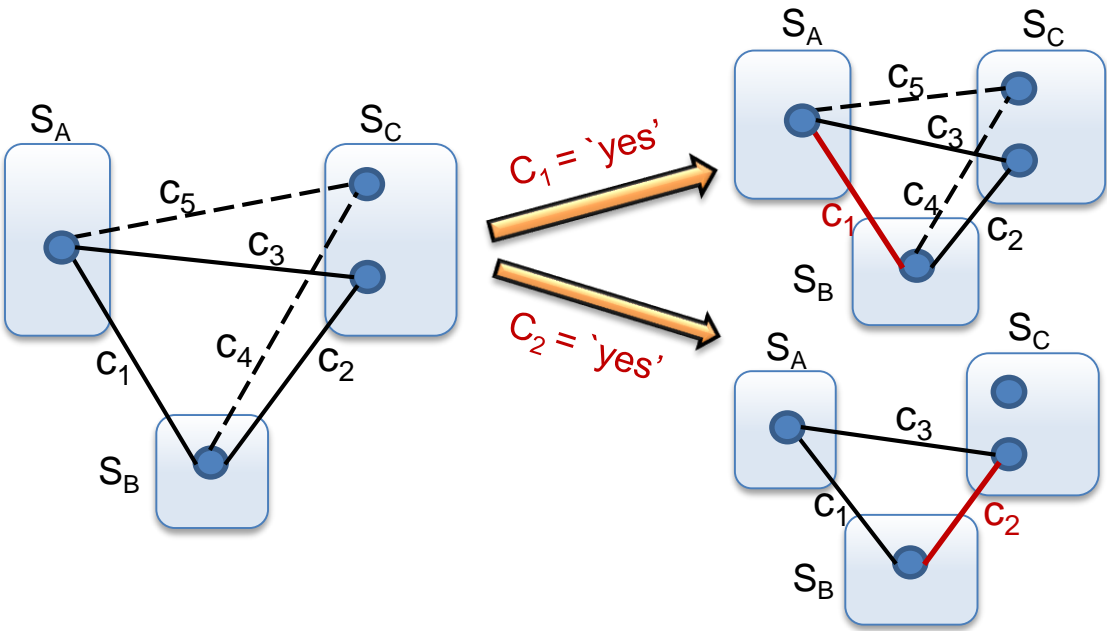
Empowering the User

- 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



Minimal Effort User Feedback?

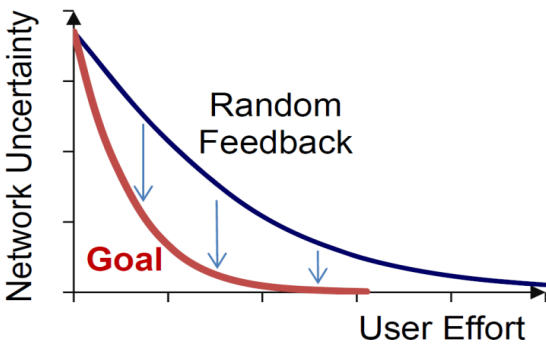
- Asking the right questions is important



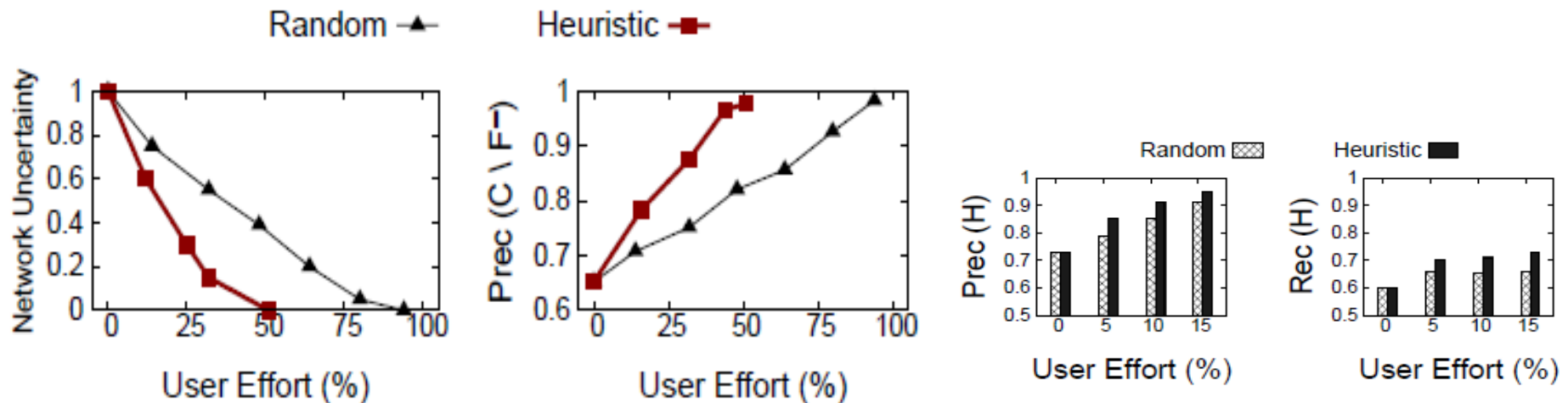
Two possible solutions:
 $\{c_1, c_2, c_3\}$ and $\{c_1, c_4, c_5\}$

- 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



- 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)



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

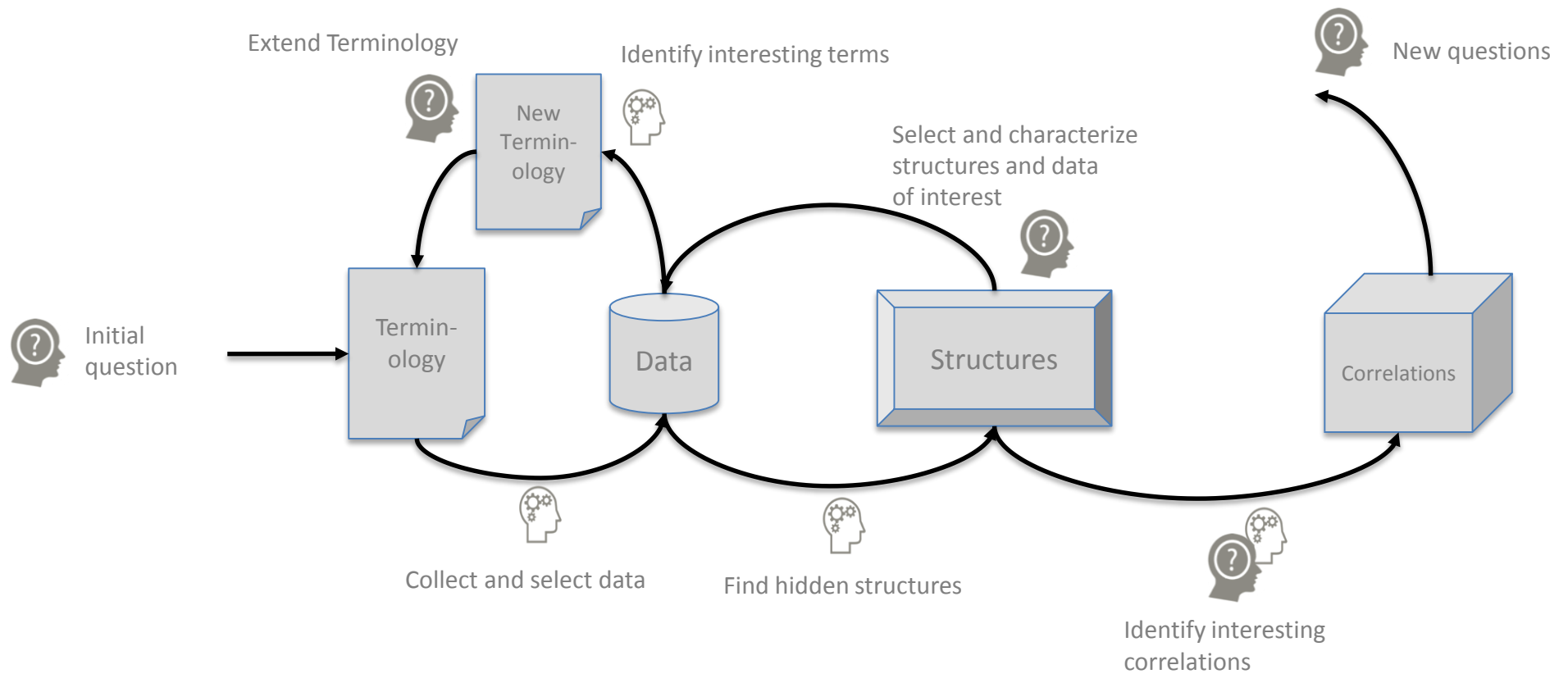
Active Learning

CASE STUDY 3: SOCIAL MEDIA ANALYSIS

- 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)



Creating Terminology

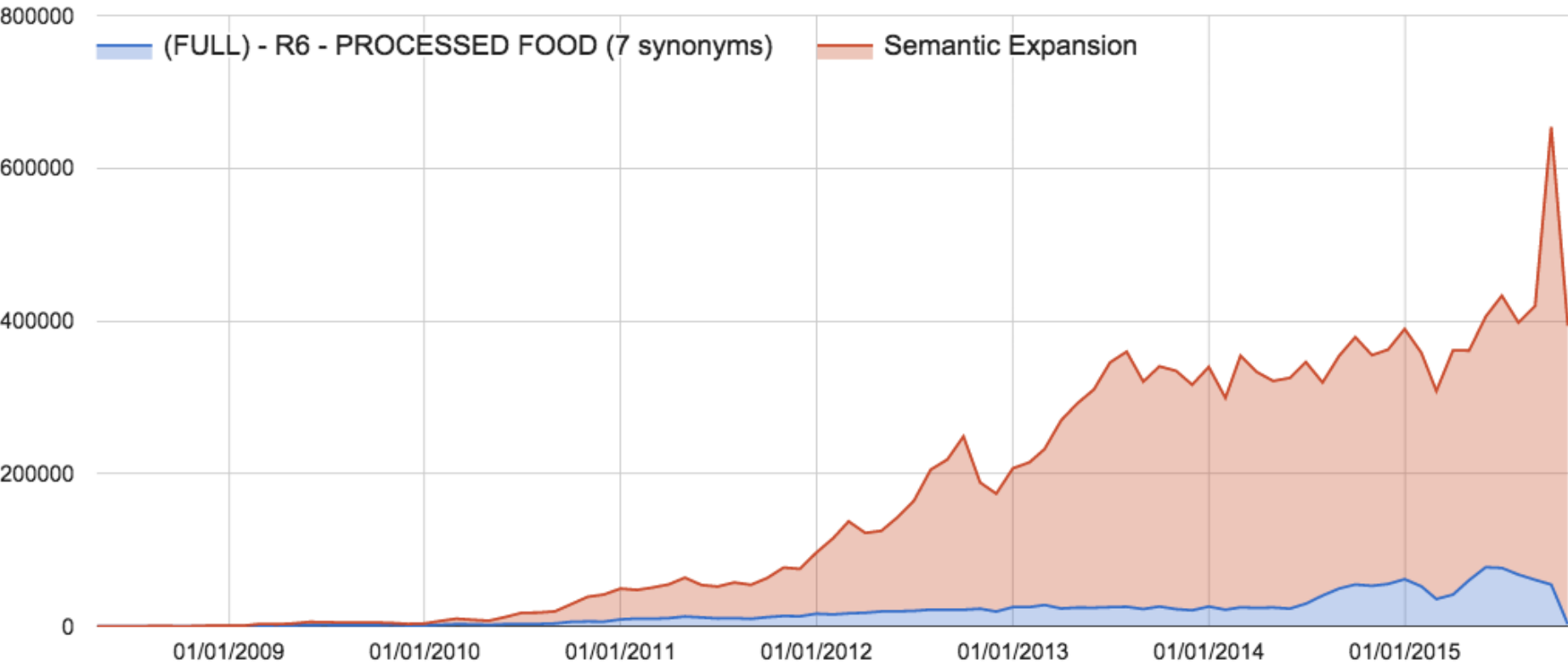
Input:
~50 Mio tweets

Analysis method:
Word embedding
(word2vec)

The figure illustrates the iterative refinement of a food-related ontology through four stages:

- Initial State:** A single category 'food' is shown with a description: "Exploring the topic 'food' the system suggest related terms. We find terms related to food ingredients". A 'Choose action' button is present.
- Refinement 1:** The ontology is expanded to include 'food ingredients' and 'nutrition'. A description for 'food ingredients' is provided: "We create a category 'food ingredients'. The system proposes more related terms.".
- Refinement 2:** A list of food ingredients is displayed, each with a count in parentheses. The list includes: (diet, 41 583), (chemicals, 8725), (cheese, 11 969), (fish, 6832), (fat, 32 555), (gluten free, 4959), (calories, 8865), (candy, 4625), (grain, 2961), (bread, 11 656), (cereal, 3196), (grains, 1861), (cereals, 2730), (cooking, 9299), (feed, 15 741), (diets, 6333), (fruits, 5832), (ground beef, 300), (eggs, 3554), (drink, 24 664), (cspi, 1741), (big organic, 12), (butter, 1), (apples, 1), and (convention, 1).
- Refinement 3:** The list is further refined, showing only the most relevant terms: (sugar, 34 228), (fat, 32 555), (sodium, 2283), (fructose, 3915), (fiber, 2222), (added sugar, 590), (protein, 8042), (salt, 5759), (nutrients, 2517), (saturated fat, 1146), (calcium, 2453), (chemicals, 8725), (hfcs, 1172), (trans fats, 680), (caffeine, 802), (artificial sweeteners, 978), (sweeteners, 2030), (carbohydrates, 233), (vitamins, 2212), (pesticides, 12), (diet soda, 1), (fruit juice, 1), (bacteria, 1), (hormones, 1262), (additives, 1750), and (trans fat, 732).

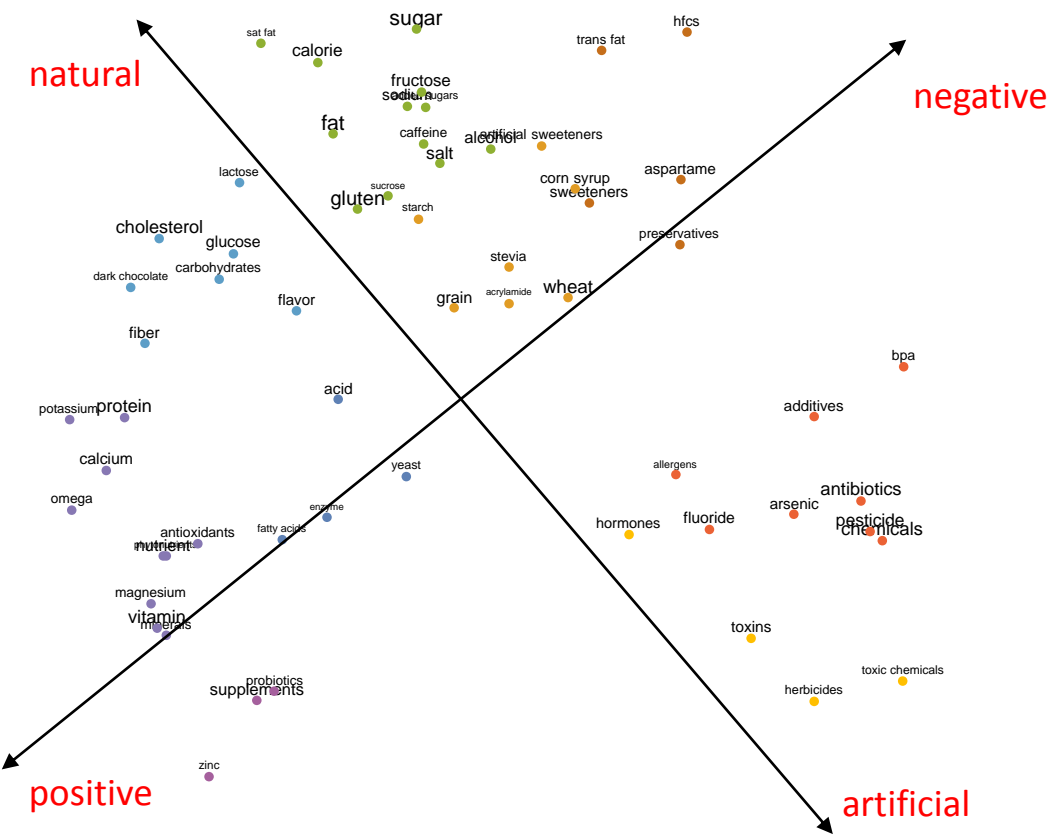
Processed Food - 7 Synonyms vs. Semantic Expansion



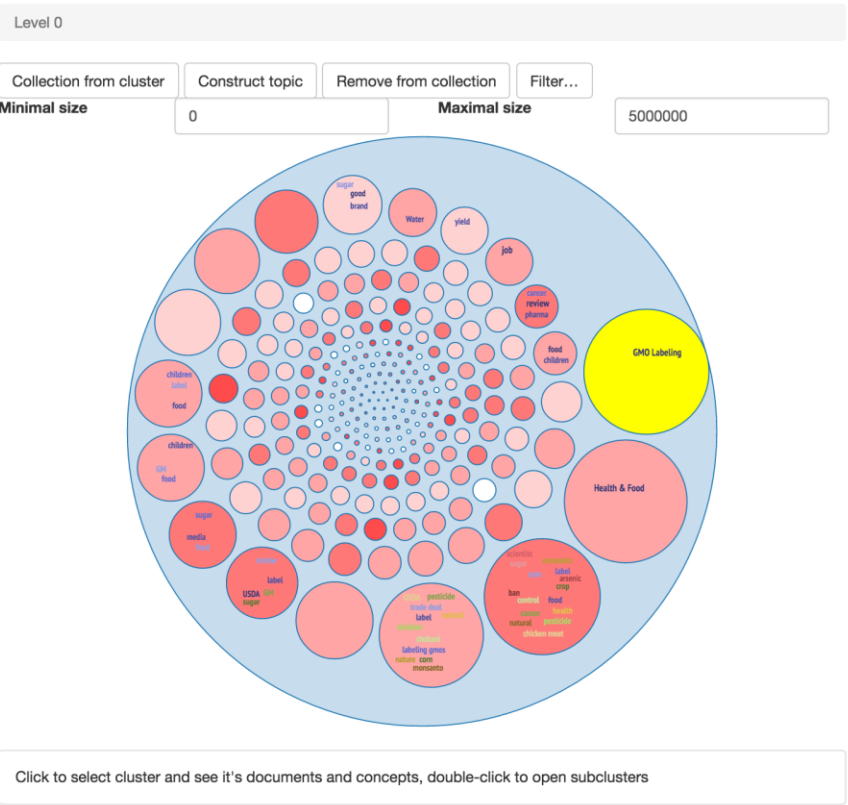
Using a semantically expanded terminology increases coverage significantly!

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



Analyzing social interactions we can identify clear communities



Info **Users(28101)** Tweets(81720) Documents(7274) Concepts(100) Hashtags(100)

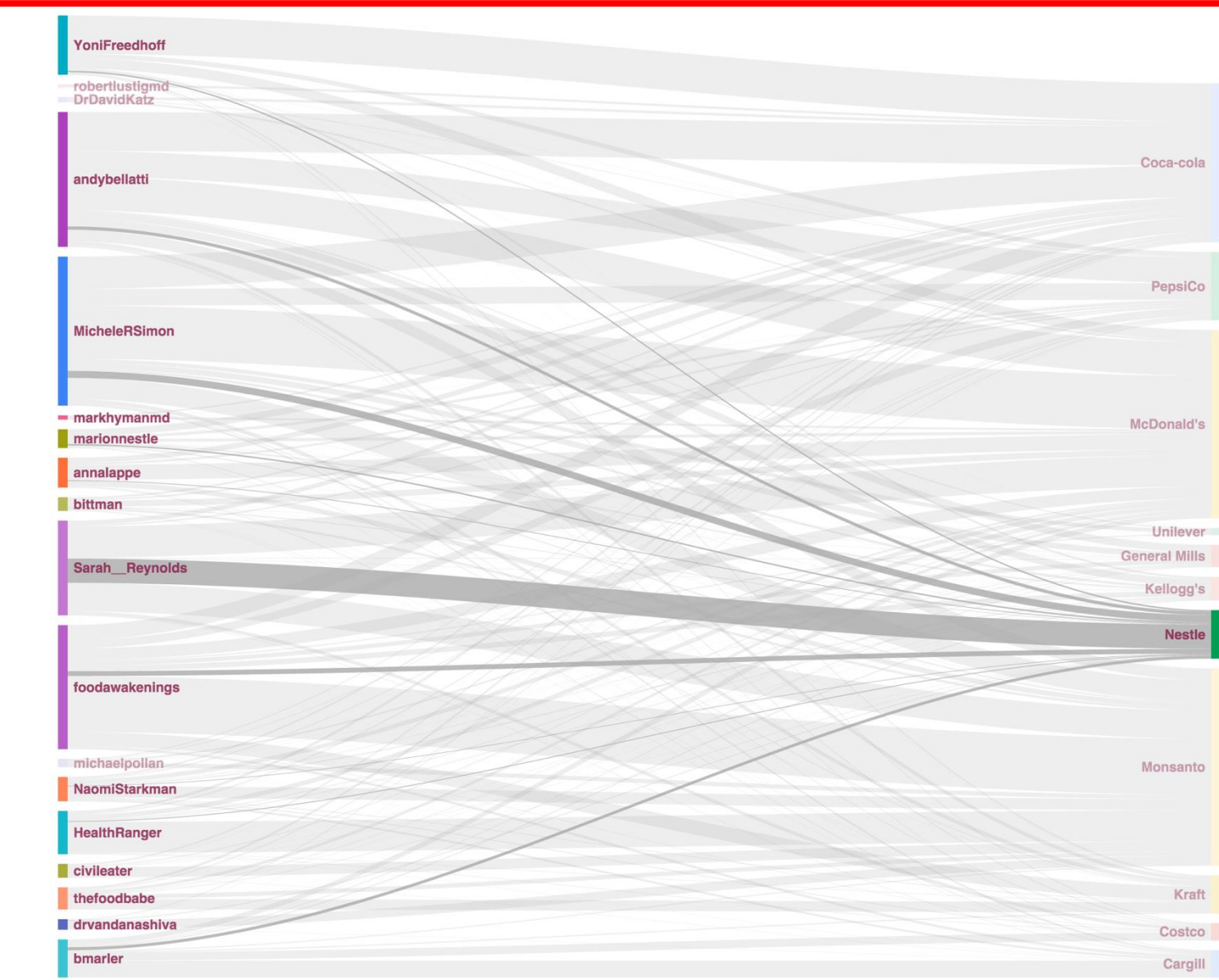
Keywords(205)

#	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/3OkImzLnch Huffington Blogger, Teen Earth Activist	1347	223	1412
5	kevinfolta	Land-grant scientist exploring ways to make better food with less food, less pesticides and less... how to effective	394	11	1363
6	TheGOPJesus	Politicians special place maybe not			
7	GMWatch	Countering			
8	MonsantoCo	Monsanto others to as sustain the			
9	geneticmaize	Mom of Ad https://t.co/3OkImzLnch 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

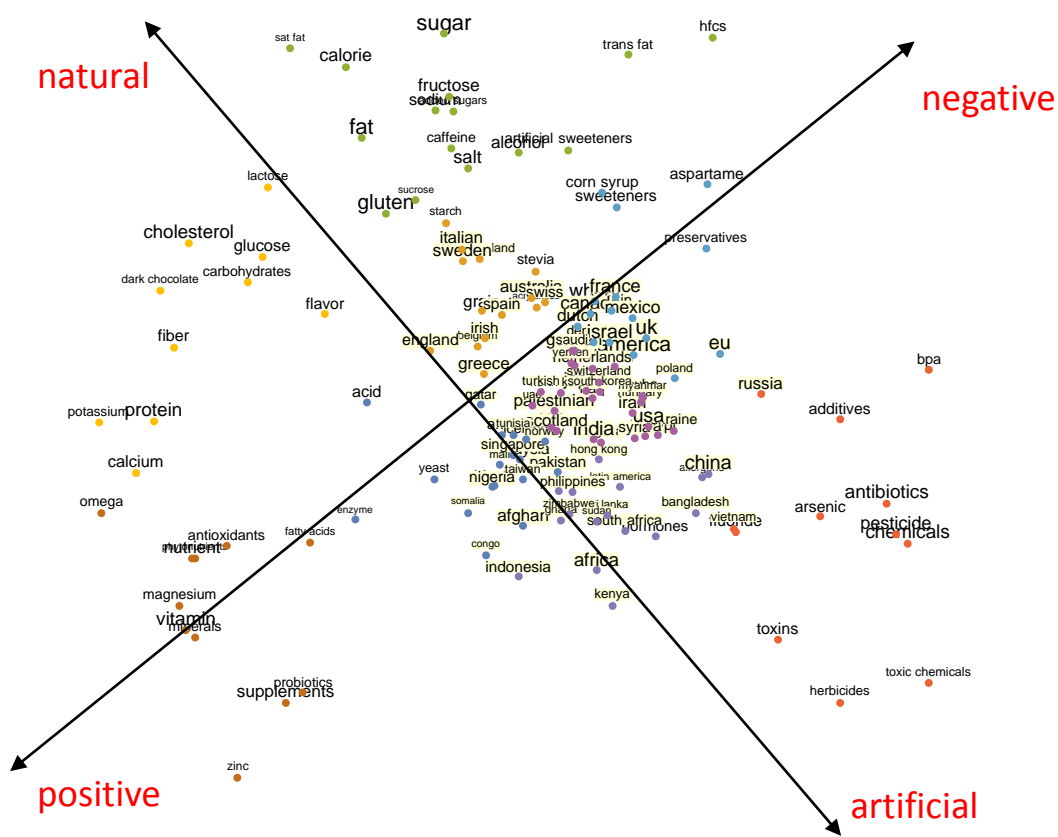
For each community we can identify

- Their influencers
- Their main concepts
- Potentially new interesting terminology

Findings: Company influencers



Findings: role of food ingredients in different countries



- 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

- 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