Information Overload/ Fake News: Are we overwhelmed?

Paul Wälti / 13 May 2019

Overview

1. Rapid increase of information - key figures

2. What does that mean? Special newest developments

   a) Disinformation due to information overload

   b) Filter bubbles

   c) Fake News

   d) Cybercrime, data protection regulations (e.g., GDPR)

   e) Artificial intelligence on the rise

3. How do we separate the wheat from the chaff?

Distribution of information - historical development

Palmyra 2000 BC stone, ceramics

Sokrates 350 BC only oral - documented
by Platon (papyrus fragments)

Archimedes 250 BC

Monasteries 360 first Christian Monastery (Egypt)

Gutenberg 1450 book printing with movable types

Radio/TV since 1900 / 1950

Computer since 1954 first computer (Konrad Zuse,
John v. Neumann)

Internet since 1989 commercialization by WWW
(Tim Berners-Lee, CERN), email
since 1993 rapid upswing by browsers
since 2003 social media: LinkedIn, XING, Facebook, Twitter

Volume and growth of electronically stored data

Study of IDC (November 2010):

New data is growing 50 to 100% per year, i.e., doubling every 1 to 2 years.

International Data Corporation (**IDC**), Massachusetts,

is a renowned international market research and consulting company in the field of

information technology and telecommunications with branches in more than 110 countries.

Increase of information at a doubling in 1-2 years

time	rel. info. mass	required material for storage
today	1	10^{6} kg = 1'000 tons = 1 tank comp.
in 3 years	4	$4*10^{6}$ kg
in 6 years	16	$16*10^{6}$ kg
in 9 years	64	$64*10^{6}$ kg
...		…
in 30 years	10^{6} = 1 mio.	10^{12} kg
in 60 years	10^{12} = 1 mrd.	10^{18} kg
in 90 years	10^{18} = 1 tri.	10^{24} kg ← weight of the earth = $6*10^{24}$ kg
in 120 years	10^{24} = 1 qua.	10^{30} kg

Limits of miniaturization: e.g. speed of light, Planck's constant

2. Some Practical Consequences

a) Information overload

Press agencies		Media, Internet		
UPI, Reuters, }	→	{ various newspapers, radio,	→	human
DPA, AFP, etc. }	→	{ TV, Info portals, emails,	→	being
		{ social media etc.	→	

→ many almost identical articles on many media: one drowns

→ partly formulated differently or manipulated: what to believe?

"Information overload is - particularly in the working world - a trigger for mental fatigue and stress" (**arte,** 13 August 2018)

Note: The bluff and misdirection by a flood of confusing infomation is not new (Clausewitz)

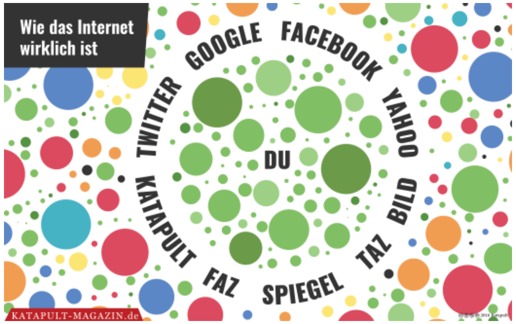
b) Coming up of Filter Bubbles

- Internet media attempt to estimate the preferences and susceptibilities of the users

- Thereafter, they forward primarily those information to the user that matches his topics of interest (=“**personalized infor-mation**“)
  → isolation against other opinions

- In extreme cases: <u>manipulation</u> by playing with the weaknesses of the user

- 2016: Filter Bubbles = word of the year



Source: Wikipedia

c) Fake News

- Un-word of the year 2014: Liar press
- Un-word of the year 2017: Alternative News (for Fake News)

Wikipedia

Kellyanne Conway 2017
Trump Campaign Manager
Inventor of "Alternative News“

Wikipedia

Sarah Sanders, Press speaker
of the White House
Trump's liar-in-chief

Examples of Fake News in the American Election Campaign

1. "Pope Francis recommends the election of Donald Trump."
   (960 000 shares on Facebook)

2. "Hillary Clinton has sold weapons to the Islamic State (IS)."
   (789 000 shares on Facebook)

3. "Hillary Clinton's e-mails to the IS have become public"
   (754 000 Shares).

Source: Buzzfeed News

Fake News also in the fashion and beauty care

Source: LinkedIn

d) Cybercrime / Data Protection Regulations

- "Hacker attacks have evolved from teenage pranks to a billion dollar growth market."
  (Source: *https://de.malwarebytes.com/hacker/)*

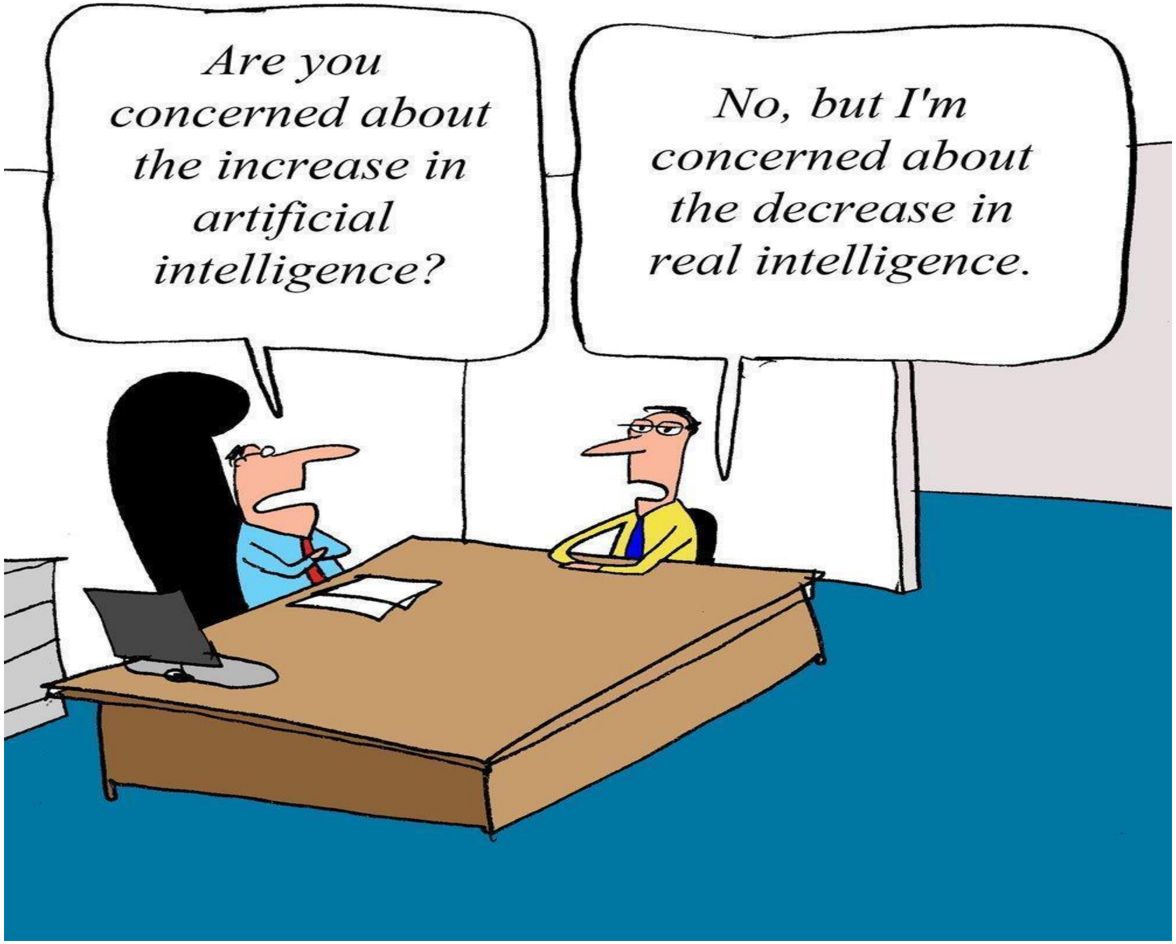
- A regional paralysis of the Internet is no longer excluded
  → new dimension for terrorist attacks.

- Blackmailing, forgeries, pornography, etc.

- Examples: Russian interference in elections, GPS interference by Russia in Finland

- Growing regulations, e.g. GDPR of the EU, May 2018, concerning data protection.
  Til today: at least 75 fines have been imposed

e) Artificial Intelligence on the upswing



Artificial Intelligence (AI) ... the euphoria

Market analysts from Gartner Group argued (2017):

- by 2020 AI technology is part of almost all software products
- AI is one of the top 5 investment priorities for 30% of the CIO

-

However: … the disillusionment

- Machine intelligence is substantially overestimated
  Heinz Scheuring, NZZaS 19.08.2017
- Progress is very high in the Robotics (quasi-AI), but
  not in the cognitive areas (analysis of free texts, decision processes, trend recognition, categorization, etc.)
- Computers = blindly obeying machines that are not meaning driven
  Willi Ritschard: computers work only so fast because they do not think
- With innovation/creativity, AI has difficulties (Financial Times, Oct. 2018)

3. How Do We Separate the Wheat from the Chaff?

Various aspects

- “Bring order and knowledge into the flood of data“ (www.chemie.de, 9.4.2019)
- Reduction of the amount of information by bundling similar articles
- Restriction to the topics one is interested in
- Automatic recognition of fake news / spam
    a) in obvious cases
    b) in complex, business relevant cases ← will be treated in more detail
- Make sure that people do not believe blindly in Fake News and spam (important, but this a case for psychologists and politicians)

Recognition of Fake News in obvious cases

Examples

“My name is Mavis Wanczyk, winner of the $ 758.7 million Power Ball Jackpot. I donate € 1.800.000,00 to you. Contact me by e-mail: maviswanczyk0009@gmail.com for info/claim“

“Ich, Herr Shaw .S. peter bietet ein zuverlässiges kreditangebot an, ich meine zuverlässiges garantieroangebot bei einer garantie zu einem erschwinglichen festen zinssatz von nur 1,2 prozent, bieten wir ausleihdarlehen von der mindestmenge von 10.000 euro bis zur maxima von 150.000.000 euro für den zeitraum von 1 bis Nur für 40 Jahre, also für weitere Details, wenn Sie Interesse an uns haben, dann erreichen Sie uns bitte über unsere Email-Adresse Antwort an: edinburghloancompany@yandex.com“

Recognition of Fake News in complex cases

This is about semantic analysis of texts (text mining and content recognition) and about mathematical methods (neural networks etc.).

Possible tools/technologies:

1. Fraunhofer tool:
   Machine learning through training with credible articles or false news, respectively

   (source: Heise.de, Feb. 2019)

2. InfoCodex technology:
   Semantic content recognition + neural network + math. statistics + linguistics

Detection of hidden Fake News with InfoCodex

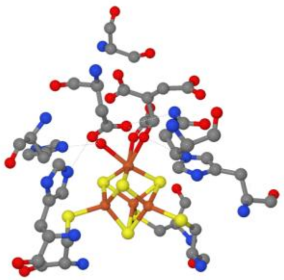
- The potential of InfoCodex in detecting hidden relationships is explained by the example of a comprehensive benchmark conducted by the pharmaceutical company **Merck USA** with InfoCodex.

- Statement of the problem: Recognize previously unknown biomarkers through the analysis of large volumes of medical publications.

- It is not attempted to explain the technology, but only the procedure.

- The same procedure can also be used when detecting hidden false news.

Discovery of Unknown Relations in Drug Research



Traditional bioinformatics: structured data
Sequence alignment, gene finding, genome assembly, protein structure prediction, gene expression…

New opportunities: e-Discovery in unstructured data
Knowledge repositories such as PubMed with 22 million citations, growing at the rate of 1.7 papers/minute

Merck's Question

Is it possible to drive drug research by text mining large pools of biomedical documents?

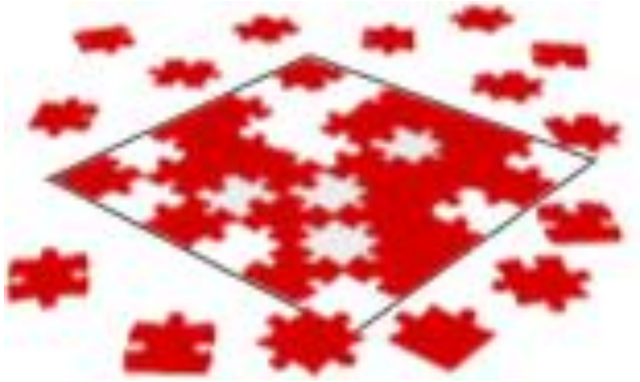
Semantic Technologies in the Pharma Industry

Commonly used: **NLP to extract triples** *“entity 1-relation-entity 2”* sentence-by-sentence

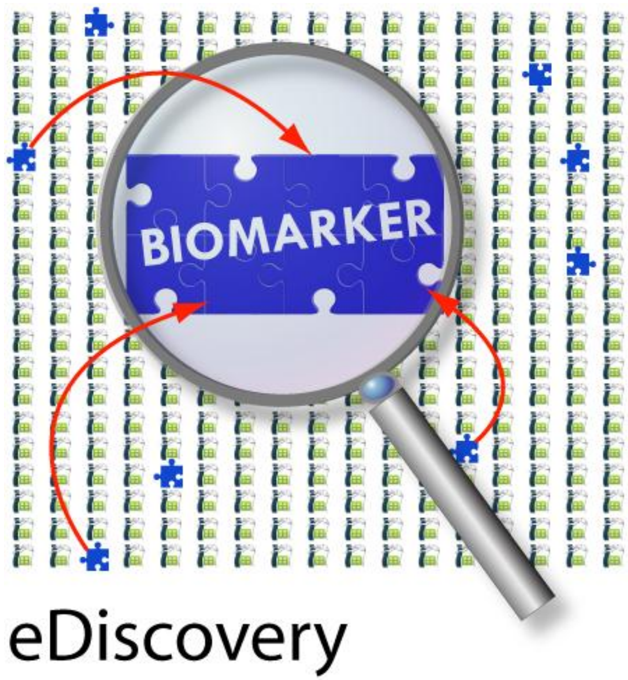
- ➪ helps to care for ontologies / libraries
- ➪ finds only what has been written down by an author, i.e. **is not a discovery approach**

Going beyond triples

Analyze text collections globally to identify small, seemingly unrelated and unnoticed facts dispersed over isolated texts, like assembling the scattered pieces of a puzzle.



The Experiment of Merck & Co with InfoCodex



The objective:

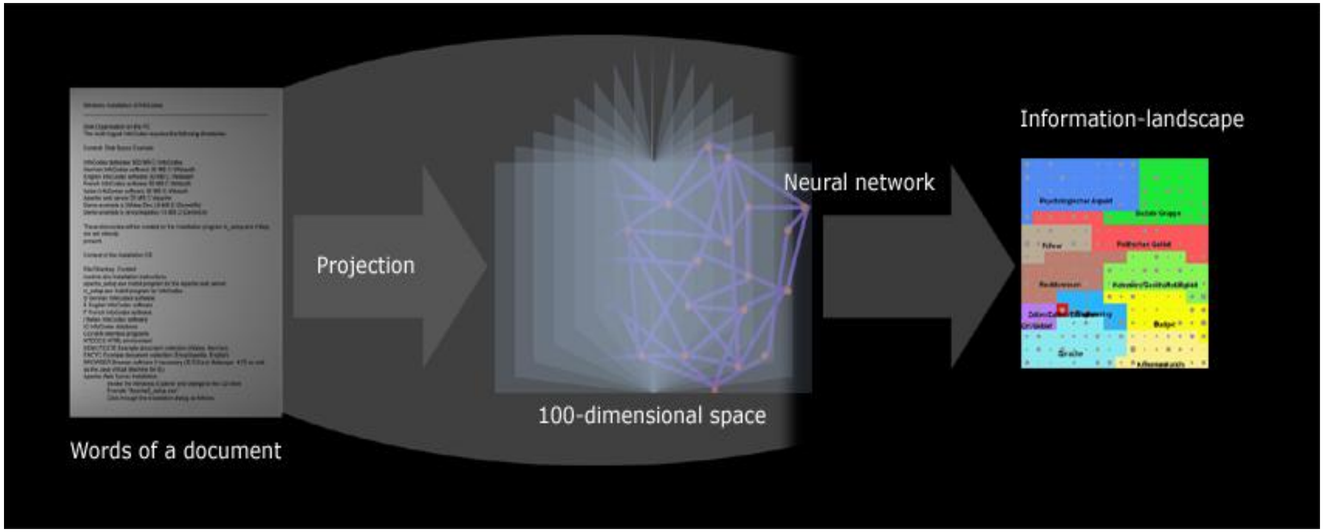
- Test pure machine intelligence for "semantic" drug research

The tasks:

- Discover novel biomarkers for diabetes and obesity (D&O) by analyzing 120'000 medical publications (PubMed etc.)
- Blind experiment, no human feedback

Biomarker: $ 13.6 billion market in 2011, growing to $ 25 billion by 2016

Method: e-Discovery in Large Sets of Publications

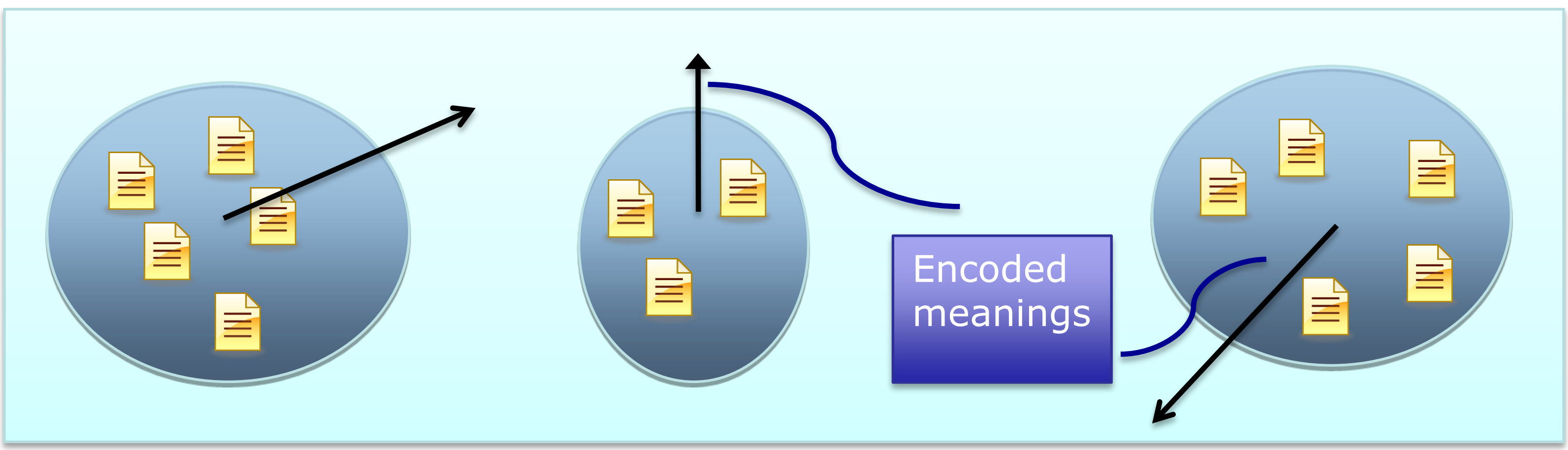


Keys to success:

- Ability to categorize unstructured information
  (in a benchmark, InfoCodex reached the very high clustering accuracy of 88%)
- Advanced statistics: combination of unnoticed correlations
  (the sentence-by-sentence analysis of the NLP approaches can detect only those relations that have been written down by an author, i.e. that are already known)

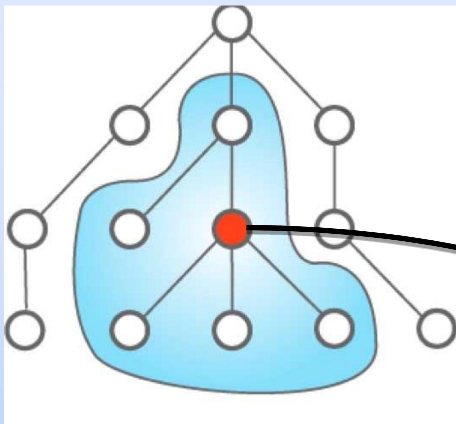
Step 1: Establish Reference Models for Biomarkers

- Collect documents describing known biomarkers for diabetes
- Cluster these documents (build groups of similar documents)
- Each cluster is considered as a reference model for the meanings of “biomarkers for diabetes”



The “Miss Marple” function

Step 2: Determine the Meaning of Unknown Words



Co-occurrences with words in internal knowledge base
→ most probable hypernym → "is a" , "has to do"

A	B	C
Unknown term	**Constructed hypernym**	**Associated descriptor 1**
Nn1250	clinical study	insulin glargine
Tolterodine	cavity	overactive bladder
Ranibizumab	drug	macular edema
Nn5401	clinical study	insulin aspart
Duloxetine	antidepressant	personal physician
Endocannabinoid	receptor	enzyme
Becaplermin	pathology	ulcer
Candesartan	cardiovascular disease	high blood pressure
Srt2104	medicine	placebo
Olmesartan	cardiovascular medicine	amlodipine
Hctz	diuretic drug	hydrochlorothiazide
Eslicarbazepine	anti nervous	Zebinix
Zonisamide	anti nervous	Topiramate Capsules
Mk0431	antidiabetic	sitagliptin
Ziprasidone	tranquilizer	major tranquilizer
Psicofarmcologia	motivation	incentive
Medoxomil	cardiovascular medicine	amlodipine

Example:
"Hctz" is a "diuretic drug" and is
a synonym of "hydrochlorothiazide"

(estimated by machine intelligence
plus the internal knowledge base)

Step 3: Construct Potential D&O Biomarkers

(substances close to one of the reference models)

Links to the relevant PubMed documents

	A	B	C	D	E	F	
1	**Part "Biomarkers" from Pubmed with confidence level > 5%; 100% refers to biomarkers of the reference set**						
2							
3	**Term**	**Relationship**	**Object**	**Target**	**Conf**	**N.Do**	**PMIDs**
4							
5	Human equilibrative nucleoside transporter-3	BiomarkerFor	Diabetes		100.0	2	20595384, 20032083
6	Human equilibrative nucleoside transporter-3	SynonymOf	hENT3				
7	microRNA	BiomarkerFor	Diabetes		100.0	44	20857148, 21118127, 21335216, 20015039, 20358579, 20364159, 21261648
8	microRNA	BiomarkerFor	Diabetes	FABP_4_aP2	100.0	1	20486779
9	microRNA	BiomarkerFor	Obesity		26.1	58	21355787, 19650761, 21152117, 21118127, 21118894, 20886002, 19188425
10	microRNA	BiomarkerFor	Obesity	FABP_4_aP2	26.1	4	19460359, 18809385, 21291493, 20486779
11	microRNA	BiomarkerFor	Obesity	GPR74	26.1	1	21036322
12	microRNA	BiomarkerFor	Obesity	AMPK	26.1	1	16459310
13	microRNA	SynonymOf	micro-RNA				
14	microRNA	SynonymOf	micro ribonucleic acid				
15	microRNA	SynonymOf	miRNA				
16	microRNA	SynonymOf	miRNA based				
17	microRNA	SynonymOf	MIR126 gene				
18	microRNA	SynonymOf	MiR-126				
19	potassium inwardly-rectifying	BiomarkerFor	Diabetes		100.0	50	20042013, 20194712, 20368737, 20401705, 20531501, 20546293, 20863361
20	potassium inwardly-rectifying	BiomarkerFor	Diabetes	FTO	100.0	8	18597214, 19020324, 18984664, 20503258, 18598350, 20142250, 18710364
21	potassium inwardly-rectifying	BiomarkerFor	Obesity		21.0	24	20049090, 20307313, 18598350, 18710364, 20712903, 18498634, 21391351
22	potassium inwardly-rectifying	BiomarkerFor	Obesity	FTO	21.0	4	20049090, 18598350, 18710364, 20929593
23	potassium inwardly-rectifying	SynonymOf	KCNJ11				
24	potassium inwardly-rectifying	SynonymOf	Kir6.2 gene				

Assessment of the Results

See Trugenberger et al. BMC Bioinformatics 2013, **14**:51

Term	*Relat.*	*Object*	*Target*	*Conf%*	*#Docs*
wenqing	BiomarkerFor	Obesity	Obesity	53.5	29
proteomic	BiomarkerFor	Obesity	Obesity	40.8	128
gene expression	BiomarkerFor	Obesity	Obesity	38.9	62
Mouse model	BiomarkerFor	Obesity	Obesity	19.8	17
muise	BiomarkerFor	Obesity	Obesity	17.5	20
athero-	BiomarkerFor	Obesity	Obesity	16.5	6
shrna	BiomarkerFor	Obesity	Obesity	9.6	4
inflammation	BiomarkerFor	Obesity	Obesity	8.2	4
TBD	BiomarkerFor	Obesity	Obesity	7.4	3
body weight	PhenoTypeOf	Diabetes	MGAT2		1
cell line	BiomarkerFor	Diabetes	MGAT2		1

Weak Points

Many uninteresting candidates
➪ too much noise
(can be easily eliminated)

Strong Points

Lots of *”**needles in the haystack**”*
Tens of extremely interesting and valuable candidates

Novel and semantically coherent terms, and therefore potentially valuable

(Merck proprietary terms hidden)

Term	*Relat.*	*Object*	*Target*	*Conf%*	*#Docs*
K	PhenoTypeOf	Obesity	Obesity	7.7	4
le	PhenoTypeOf	Obesity	Obesity	7	6
c	BiomarkerFor	Obesity	Obesity	4.9	1
m	BiomarkerFor	Obesity	Obesity	4.9	1
P	BiomarkerFor	Obesity	Obesity	2.9	2
M	BiomarkerFor	Obesity	Obesity	2.2	1
M	BiomarkerFor	Obesity	Obesity	2.2	1
M	BiomarkerFor	Obesity	Obesity	2.2	1
re	BiomarkerFor	Diabetes	Diabetes	14.5	1
P	BiomarkerFor	Diabetes	Diabetes	2.8	2

What has the benchmark to do with the discovery of Fake News

- The methodology is not limited to the discovery of biomarkers.

- The benchmark shows that the text-mining of large volumes of unstructured documents, combined with cross-documentary statistical analysis, can uncover unknown relationships.
  (“the Holy Grail of text mining”, what NLP methods cannot offer).

- But it needs a **reference model**: What are credible statements (or examples of fake news) in the considered area.

- Because **the computer cannot think**, it can just compare