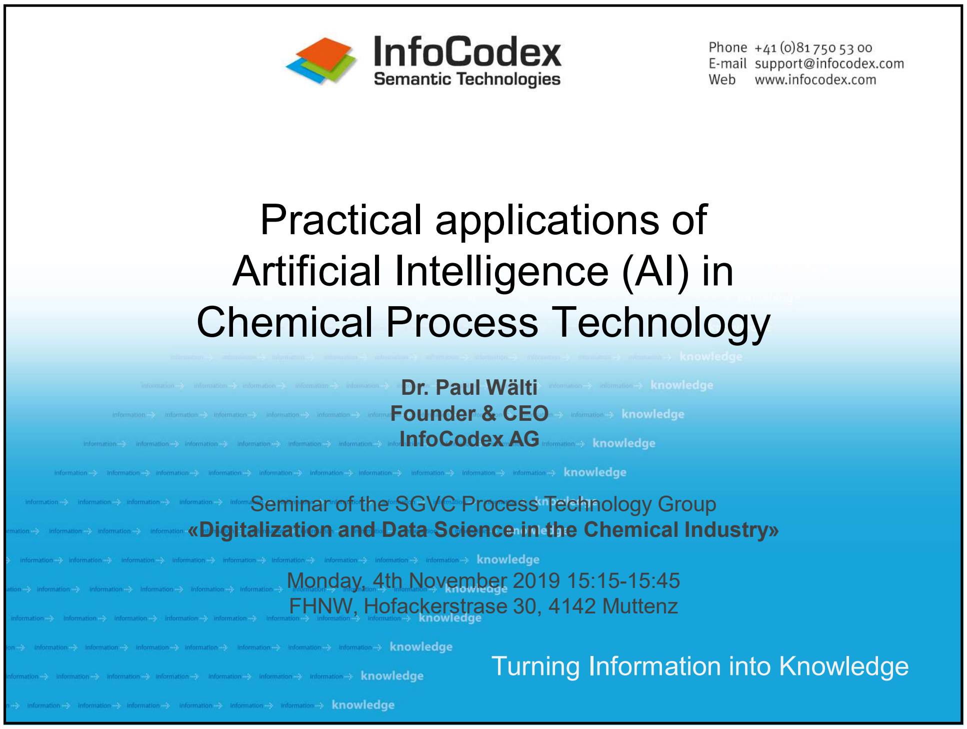


InfoCodex
Semantic Technologies

Phone +41 (0)81 750 53 00
E-mail support@infocodex.com
Web www.infocodex.com

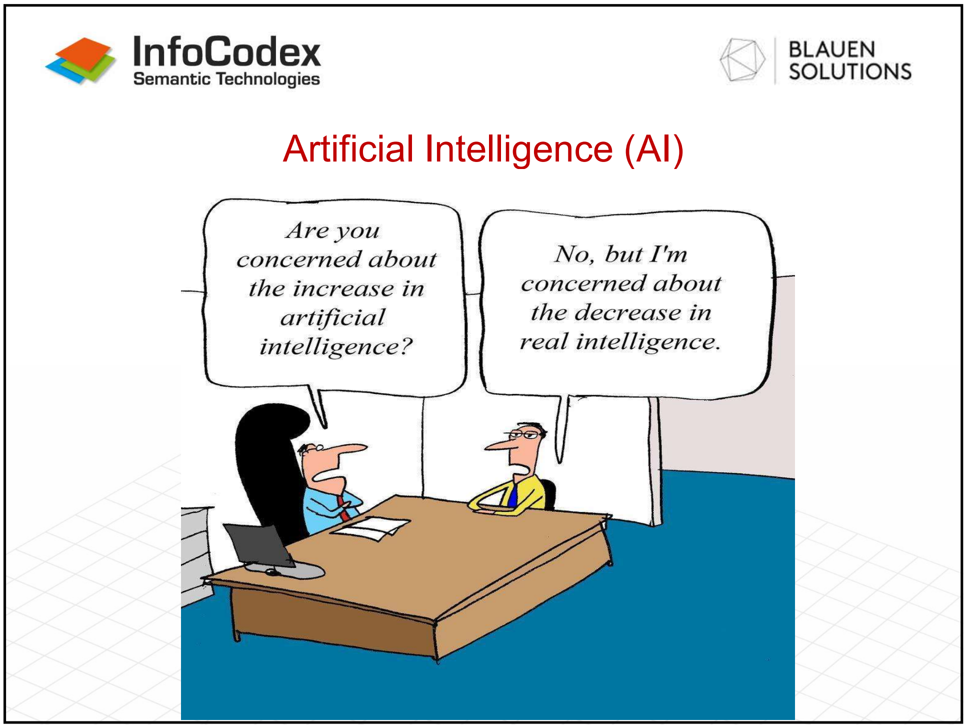
Practical applications of Artificial Intelligence (AI) in Chemical Process Technology

Dr. Paul Wälti
Founder & CEO
InfoCodex AG

Seminar of the SGVC Process Technology Group
«Digitalization and Data Science in the Chemical Industry»

Monday, 4th November 2019 15:15-15:45
FHNW, Hofackerstrase 30, 4142 Muttenz

Turning Information into Knowledge



InfoCodex
Semantic Technologies

BLAUEN SOLUTIONS

Artificial Intelligence (AI)

Are you concerned about the increase in artificial intelligence?
No, but I'm concerned about the decrease in real intelligence.

AI … the Euphoria

Artificial Intelligence is Part of Every Software in 2020

(Simon Wegmüller, ITRESELLER, 19. July 2017)

Market euphoria and an increasing interest in artificial intelligence are driving software vendors more and more into integrating technology into their product strategies.

Market analysts at Gartner believe that

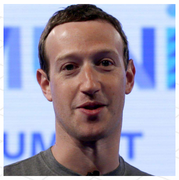
- by 2020 AI technology is part of almost all software products
- AI by then is among the top five investment priorities for more than 30 percent of the CIOs

AI ... the Disillusionment

Machine Intelligence is Massively Overestimated

(Heinz Scheuring, NZZaS, 19.08.2017)

The warning that machines are soon superior to man and enslave him is irresponsible and absurd.



Silicon Valley's progressive ideologists are suddenly terrified: like sorcerers' apprentices, they are afraid of the alleged gains of Artificial Intelligence (AI).

InfoCodex
Semantic Technologies

BLAUEN
SOLUTIONS

What are the Needs of Chemical Process Technology on Digitization and Data Scienes?

- Enterprise Information Management
  - Knowledge centers, overview on masses of free text documents
  - Secure the relevant company knowledge
  - Avoid reinvention of the wheel
- Being informed about the market and new technologies
  - Continuous oberservation of customer needs, competitors and new technologies; get alerts on new facts
  - Early recognition of trends and risks
- Content recognition, summary generation, facts extraction
- Automation and robotics

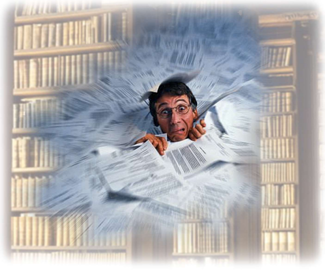
InfoCodex
Semantic Technologies

BLAUEN
SOLUTIONS

Today‘s Information Management Problem

How can we manage the flood of information?

- **Electronically stored data** growth 50 to 100% per year, i.e., **doubling every 1 to 2 years**
  (International Data Corporation IDC, November 2010)
- **85%** of the electronically stored corporate information is **unstructured information**
  (IBM study, August 2005)



InfoCodex
Semantic Technologies

BLAUEN
SOLUTIONS

Current Text Analytics Tools

The market leaders (IBM Watson and HP Idol or former Autonomy) use NLP (Natural Language Processing) based methodologies.

1. Can these categorize documents according to their thematic content without human intervention?

   Yes in known situations, but **No in new, unknown** cases
   (→ need weeks/months of training with large amounts of data to build the knowledge structure)

2. Are they capable to discover novel relationships through analyzing large amounts of literature and realworld data?
   No, the NLP based methods can extract **only known facts** with their sentence-by-sentence analysis

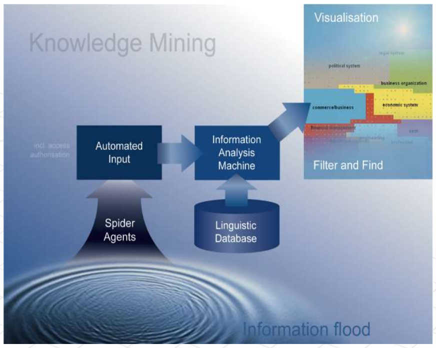
InfoCodex
Semantic Technologies

BLAUEN
SOLUTIONS

Solution by InfoCodex: Semantics + Neural Networks + Statistical Analysis

The essential features

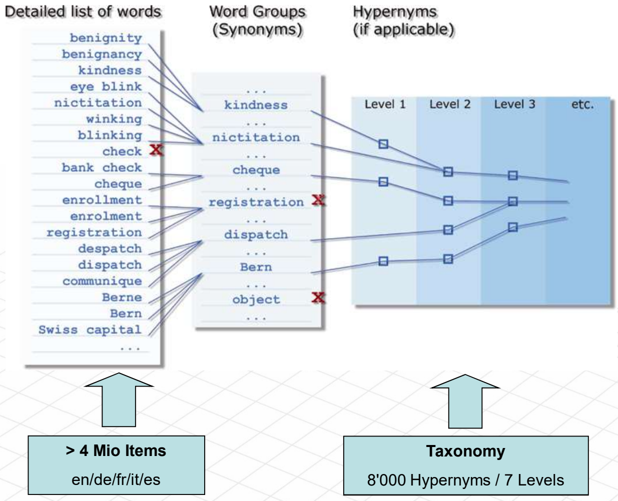
- Unique, large Linguistic Database linked to a universal Taxonomy (4 Mio items, en/de/fr/it/es)
- Combined with linguistic and statistical analyses and self-organizing neural networks
- Patented in the EU and the USA



InfoCodex
Semantic Technologies

BLAUEN
SOLUTIONS

Basis for the Semantic Analysis: Linguistic Database



> 4 Mio Items

en/de/fr/it/es

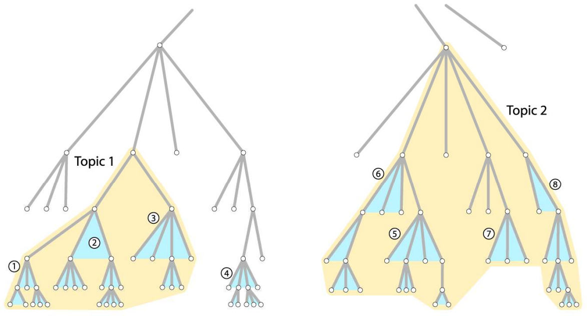
Taxonomy

8'000 Hypernyms / 7 Levels

InfoCodex
Semantic Technologies

BLAUEN
SOLUTIONS

Clustering of Documents



Combining Agglomeration Clusters into Thematic Topics

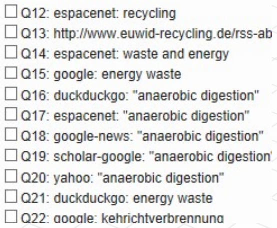
InfoCodex Semantic Technologies

BLAUEN SOLUTIONS

Combination of Dispersed Information Sources

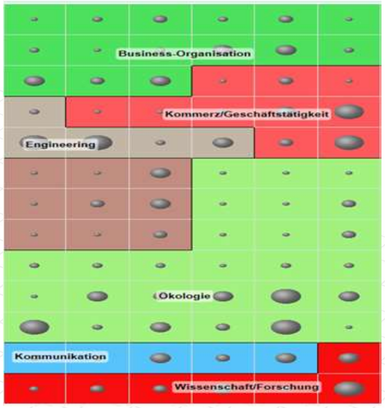
Define interesting and relevant data sources

- Websites
- Internet Databases
- Search Engine Requests
- RSS feeds
- File servers
- E-Mails
- …



□ Q12: espacenet: recycling
□ Q13: http://www.euwid-recycling.de/rss-ab
□ Q14: espacenet: waste and energy
□ Q15: google: energy waste
□ Q16: duckduckgo: "anaerobic digestion"
□ Q17: espacenet: "anaerobic digestion"
□ Q18: google-news: "anaerobic digestion"
□ Q19: scholar-google: "anaerobic digestion'
□ Q20: yahoo: "anaerobic digestion"
□ Q21: duckduckgo: energy waste
□ Q22: google: kehrichtverbrennung

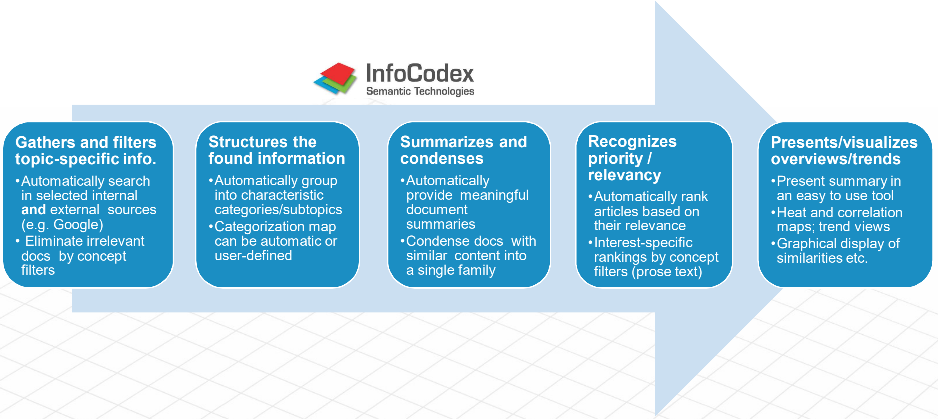
… and InfoCodex generates fully automatic a categorized overview of the content



InfoCodex Semantic Technologies

This is what the Tool Should Do

Automatically read and analyze big amounts of unstructured (incl. structured) information, condense and summarize ("pre-digestion")



InfoCodex Semantic Technologies

Gathers and filters topic-specific info.
- Automatically search in selected internal **and** external sources (e.g. Google)
- Eliminate irrelevant docs by concept filters

Structures the found information
- Automatically group into characteristic categories/subtopics
- Categorization map can be automatic or user-defined

Summarizes and condenses
- Automatically provide meaningful document summaries
- Condense docs with similar content into a single family

Recognizes priority / relevancy
- Automatically rank articles based on their relevance
- Interest-specific rankings by concept filters (prose text)

Presents/visualizes overviews/trends
- Present summary in an easy to use tool
- Heat and correlation maps; trend views
- Graphical display of similarities etc.

InfoCodex Semantic Technologies

BLAUEN SOLUTIONS

The Ants as Role Model

Collective Intelligence. See also Bryan Walsh, TIME, May 27, 2014:
Your Ant Farm is Smarter Than Google

InfoCodex Semantic Technologies

BLAUEN SOLUTIONS

Back the Needs of Chemical Process Technology

- Enterprise Information Management
  https://www.infocodex.com/en/knowledge-management
- Being informed about the market and new technologies
  https://www.infocodex.com/en/competitor-monitoring
- Content recognition, summary generation, facts extraction
  https://www.infocodex.com/en/ediscovery
- Plant Automation & Maintenance / Process Control
  https://www.infocodex.com/en/response-management

InfoCodex Semantic Technologies

BLAUEN SOLUTIONS

Enterprise Information Management

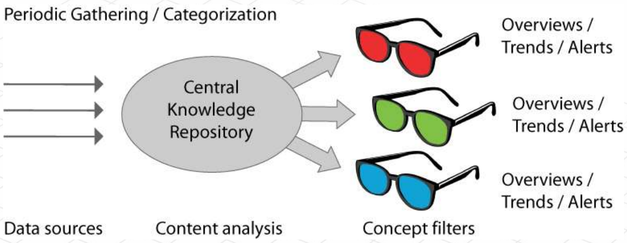
- **Knowledge Centers:** Overview of large quantities of heterogeneous documents and easy access to dispersed information
- **Corporate information management:** e.g. handling of decision-relevant short-term information
- **Secure the relevant company knowledge:** Knowledge transfer, reduce dependence on individual persons

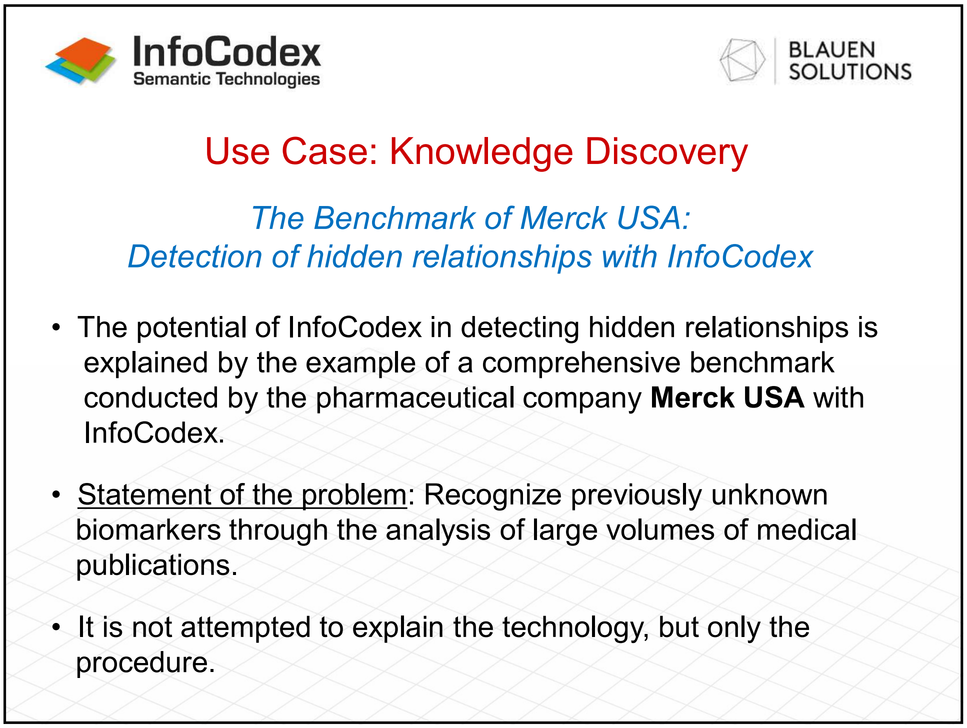
InfoCodex Semantic Technologies

BLAUEN SOLUTIONS

Concept for Market Intelligence

Spidering the Web and gathering data on competitors or on other fields of interest and making the relevant data available in an easy and accessible way.


Periodic Gathering / Categorization
Central Knowledge Repository
Overviews / Trends / Alerts
Overviews / Trends / Alerts
Overviews / Trends / Alerts
Data sources
Content analysis
Concept filters



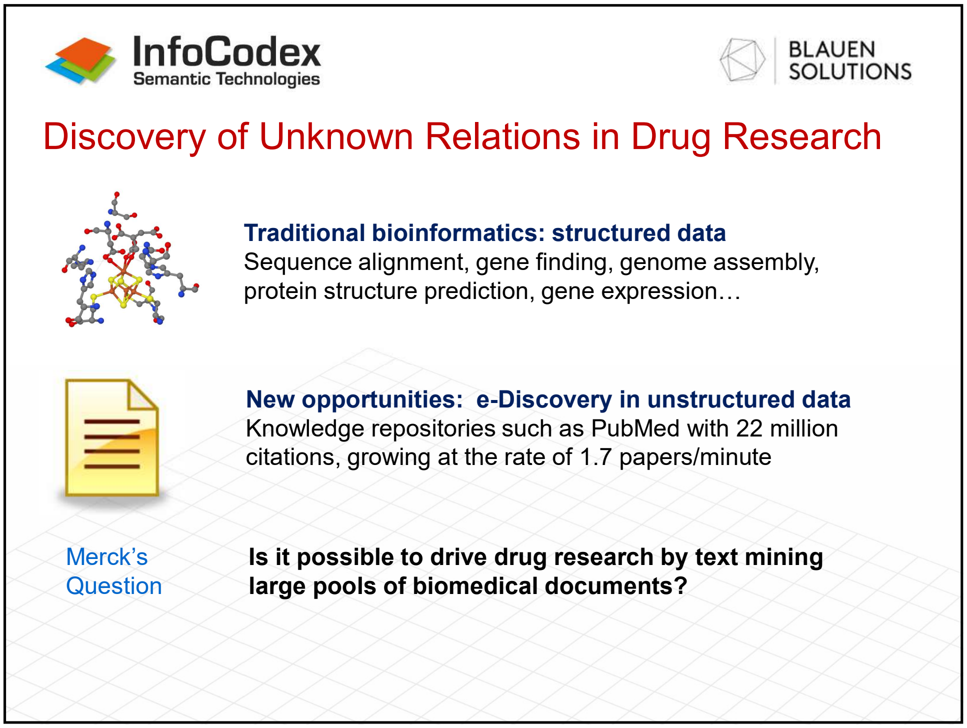
InfoCodex Semantic Technologies

BLAUEN SOLUTIONS

Use Case: Knowledge Discovery

The Benchmark of Merck USA:
Detection of hidden relationships 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.



InfoCodex Semantic Technologies

BLAUEN SOLUTIONS

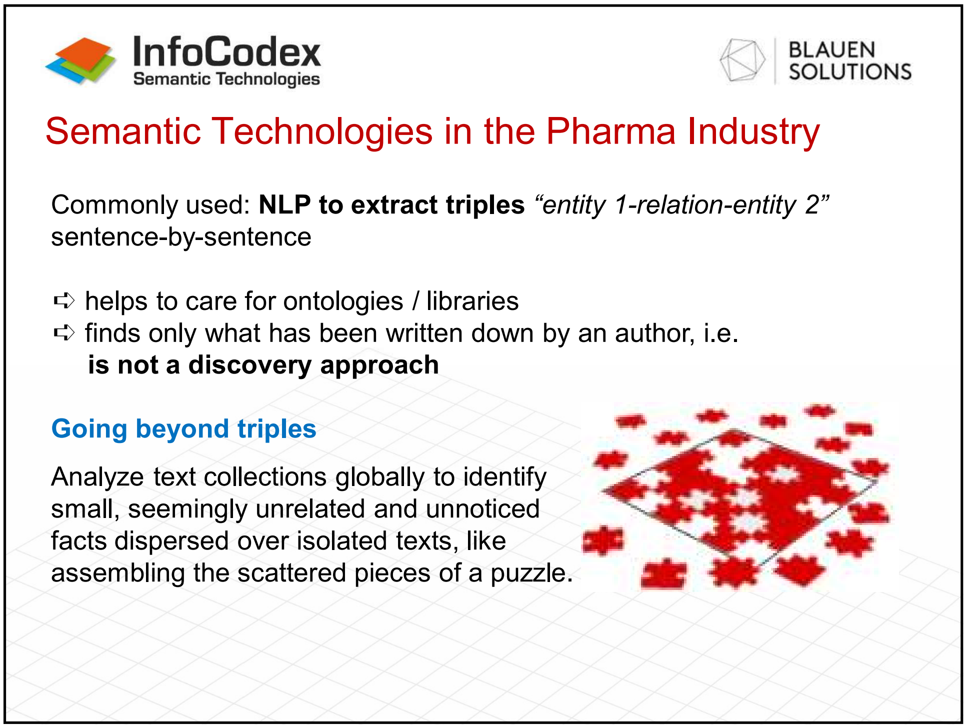
Discovery of Unknown Relations in Drug Research

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?



InfoCodex Semantic Technologies

BLAUEN SOLUTIONS

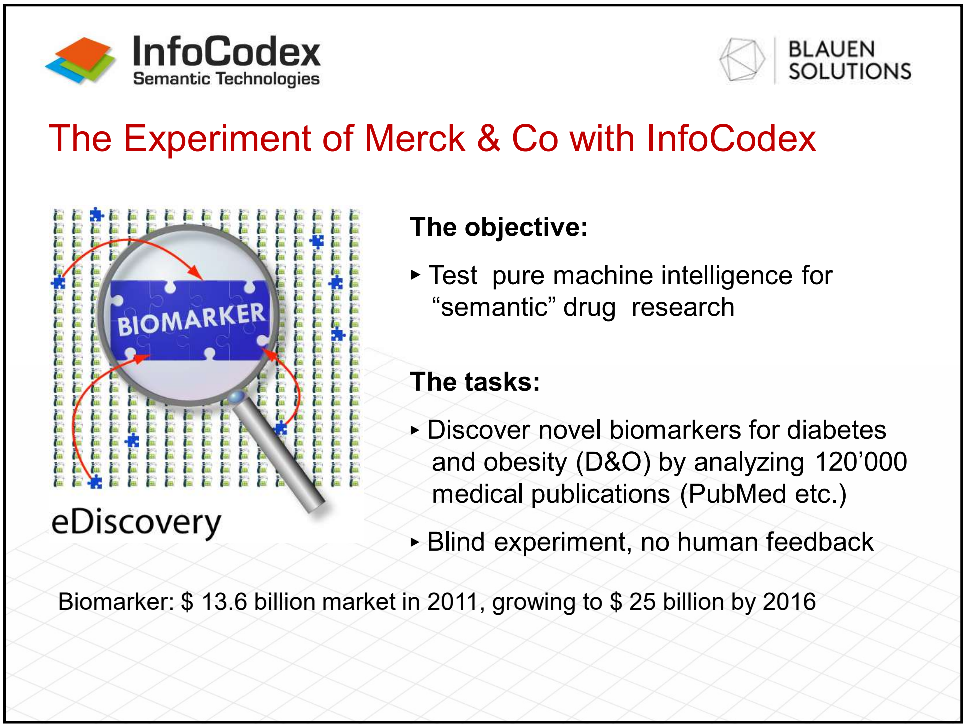
Semantic Technologies in the Pharma Industry

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.



InfoCodex Semantic Technologies

BLAUEN SOLUTIONS

The Experiment of Merck & Co with InfoCodex

BIOMARKER
eDiscovery

The objective:

- Test pure machine intelligence for “semantic” drug research

The tasks:

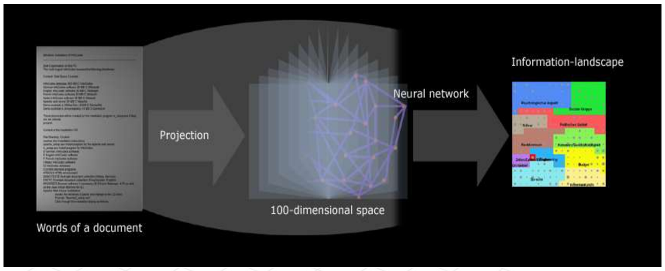
- Discover novel biomarkers for diabetes and obesity (D&O) by analyzing 120’000 medical publications (PubMed etc.)
- Blind experiment, no human feedback

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

InfoCodex Semantic Technologies

BLAUEN SOLUTIONS

Method: e-Discovery in Large Sets of Publications



Keys to success:

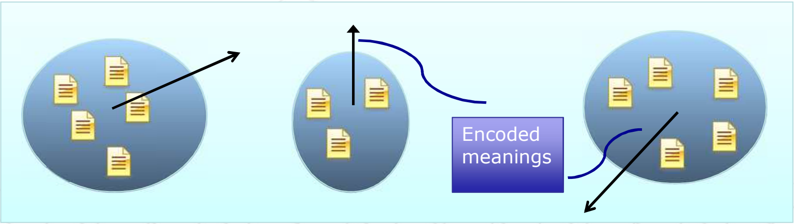
- Ability to categorize unstructured information
  (in a benchmark, InfoCodex reached the very high clustering accuracy of 88%)
- Advanced statistics: combination of unnoticed correlations
  (the sentence-by-sentence analysis of the NLP approaches can detect only those relations that have been written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

InfoCodex Semantic Technologies

BLAUEN SOLUTIONS

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”

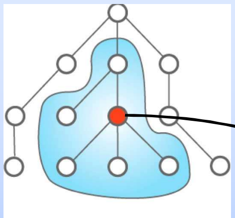


The “Miss Marple” function

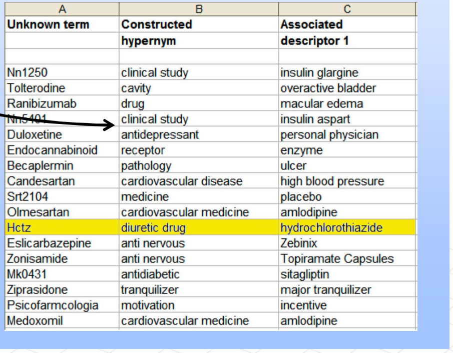
InfoCodex
Semantic Technologies

BLAUEN
SOLUTIONS

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)

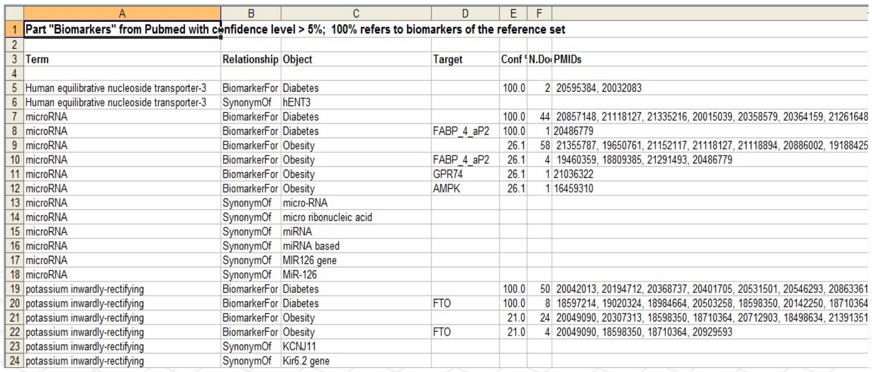
InfoCodex
Semantic Technologies

BLAUEN
SOLUTIONS

Step 3: Construct Potential D&O Biomarkers

(substances close to one of the reference models)

Links to the relevant PubMed documents



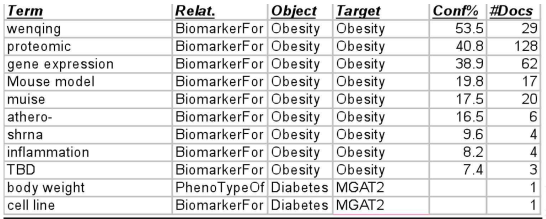
	A	B	C	D	E	F	
1	**Part "Biomarkers" from Pubmed with confidence level > 5%; 100% refers to biomarkers of the reference set**						
2							
3	**Term**	**Relationship**	**Object**	**Target**	**Conf**	**N.Do**	**PMIDs**
4							
5	Human equilibrative nucleoside transporter-3	BiomarkerFor	Diabetes		100.0	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				

InfoCodex Semantic Technologies

BLAUEN SOLUTIONS

Assessment of the Results

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



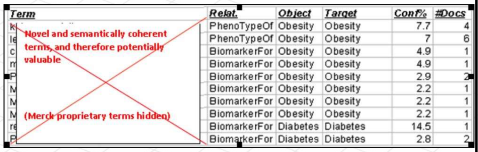
Term	*Relat.*	*Object*	*Target*	*Conf%*	*#Docs*
wenqing	BiomarkerFor	Obesity	Obesity	53.5	29
proteomic	BiomarkerFor	Obesity	Obesity	40.8	128
gene expression	BiomarkerFor	Obesity	Obesity	38.9	62
Mouse model	BiomarkerFor	Obesity	Obesity	19.8	17
muise	BiomarkerFor	Obesity	Obesity	17.5	20
athero-	BiomarkerFor	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*
[redacted]	PhenoTypeOf	Obesity	Obesity	7.7	4
[redacted]	PhenoTypeOf	Obesity	Obesity	7	6
[redacted]	BiomarkerFor	Obesity	Obesity	4.9	1
[redacted]	BiomarkerFor	Obesity	Obesity	4.9	1
[redacted]	BiomarkerFor	Obesity	Obesity	2.9	2
[redacted]	BiomarkerFor	Obesity	Obesity	2.2	1
[redacted]	BiomarkerFor	Obesity	Obesity	2.2	1
[redacted]	BiomarkerFor	Obesity	Obesity	2.2	1
[redacted]	BiomarkerFor	Diabetes	Diabetes	14.5	1
[redacted]	BiomarkerFor	Diabetes	Diabetes	2.8	2

InfoCodex Semantic Technologies

BLAUEN SOLUTIONS

Summary – Take Aways

Value

- **Processing large volumes of data**
  (different formats, all sorts of locations)
- **Analysing real-world data**
  (merging more and less structured data)
- **Knowledge Discovery**
  (discovering unknown relationships)

Different and Distinctive

– **Understanding Context**
  (Linguistic vs NLP-methodology)
– **Cross-language analysis by default**
  (German, English, French, Italian, Spanish)
– **Minimal content training required**
  (days as opposed to weeks/months)

InfoCodex Semantic Technologies

Essential Features

– **Linguistic Database**
  Large Linguistic Database linked to a universal Taxonomy (4 Mio items)
– **Adv. linguistic & statistical analysis**

– **Self-Organizing Neural Networks**

Technology

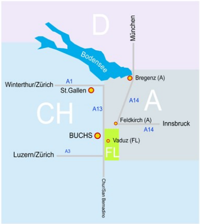
- **Easy Set-up & Highly Scalable**
- **Rapid installation**
  Security: internal / ext. servers, cloud
- **Enterprisewide Integration**
  LDAP & Sharepoint API
- **Resources:**
  Highly efficient database

InfoCodex
Semantic Technologies

BLAUEN
SOLUTIONS

Contact

InfoCodex AG *Semantic Technologies*
Bahnhofstrasse 50
CH-9470 Buchs SG, Switzerland
Phone: +41 81 750 53 00
Dr. Paul Wälti, CEO
wae@infocodex.com



Business Development

Blauen Solutions GmbH
Vogtackerweg 11
CH-4148 Pfeffingen BL / Switzerland
Phone +41 61 701 8181
Beat Meyer, Business Development Manager
beat.meyer@infocodex.com