Using Semantic Technology to Solve Sparse Training Material Problem in Machine Learning for Classification of Company Websites

Klaus Kater
Deep SEARCH 9 GmbH
Managing Partner

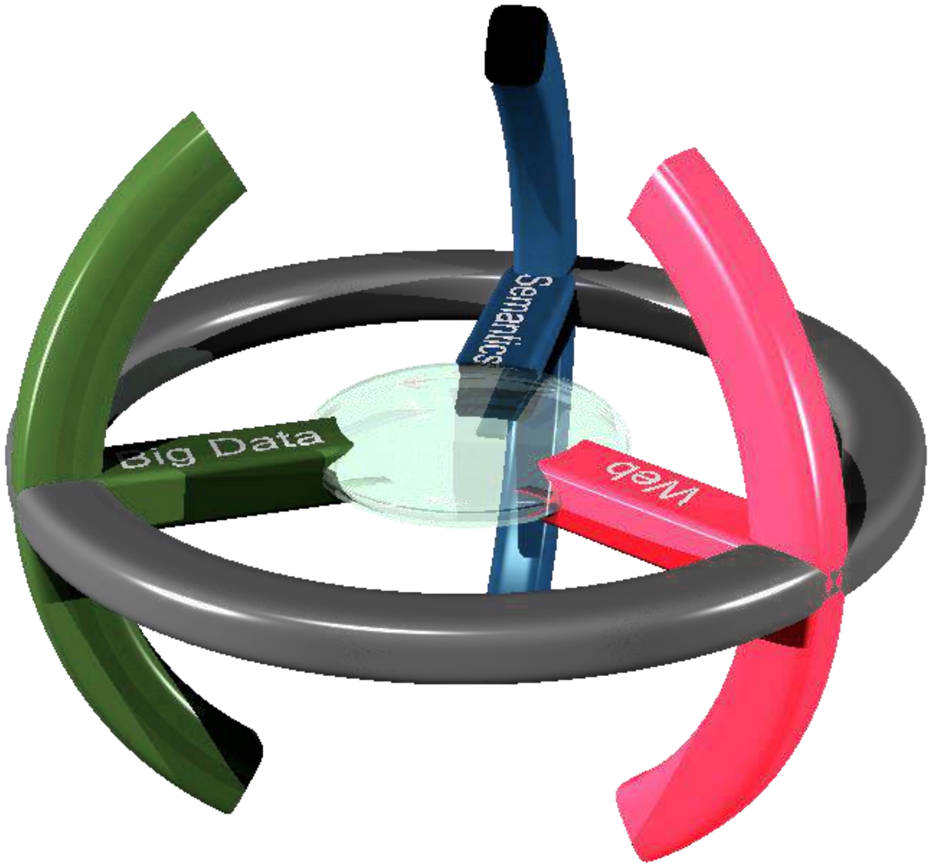
klaus.kater@deepsearchnine.com

SEMANTiCS 2018

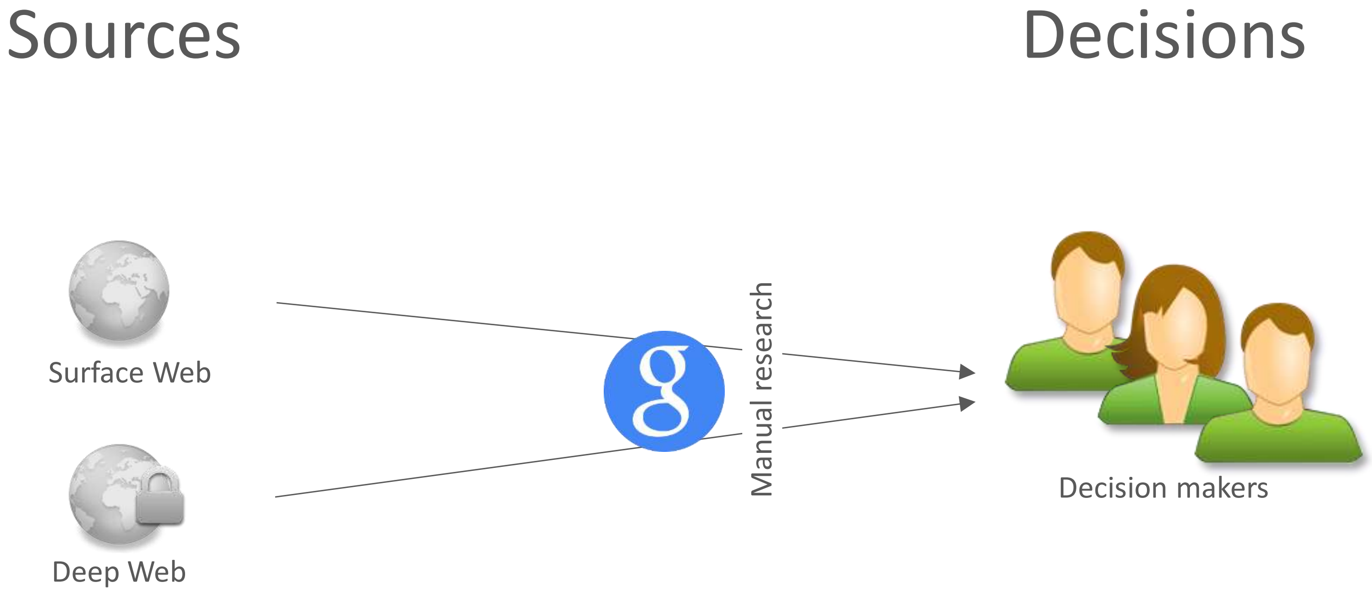
Where Machine Learning Meets Semantics
10th - 13th of September 2018 in Vienna

Deep SEARCH 9

Managed Intelligence.



Web Information Analysis



- 100s of emails…
- 1,000s of websites…
- Once a week, daily, every other hour?
- Keep sitting there, hitting F5 ;-)

Web Information Analysis

Sources

Decisions

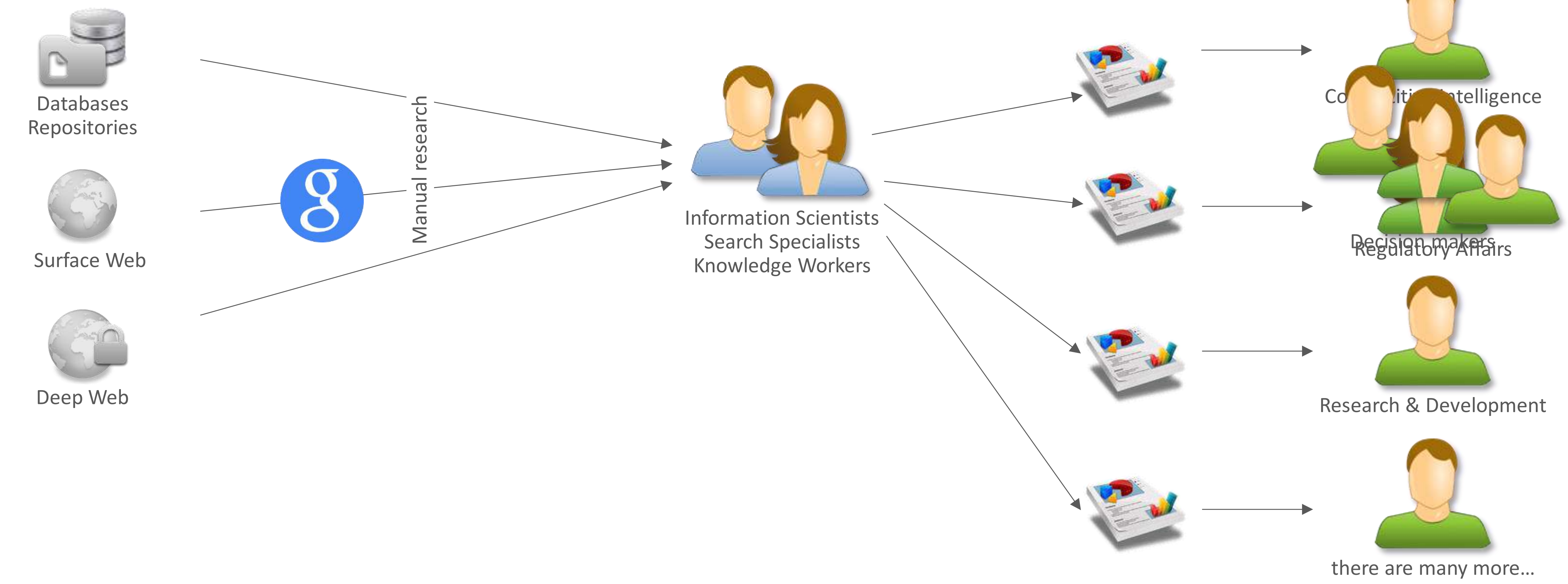
Web Information Analysis

Web Big Data Deep SEARCH 9 Semantics Web

Sources

Expert Search

Decisions



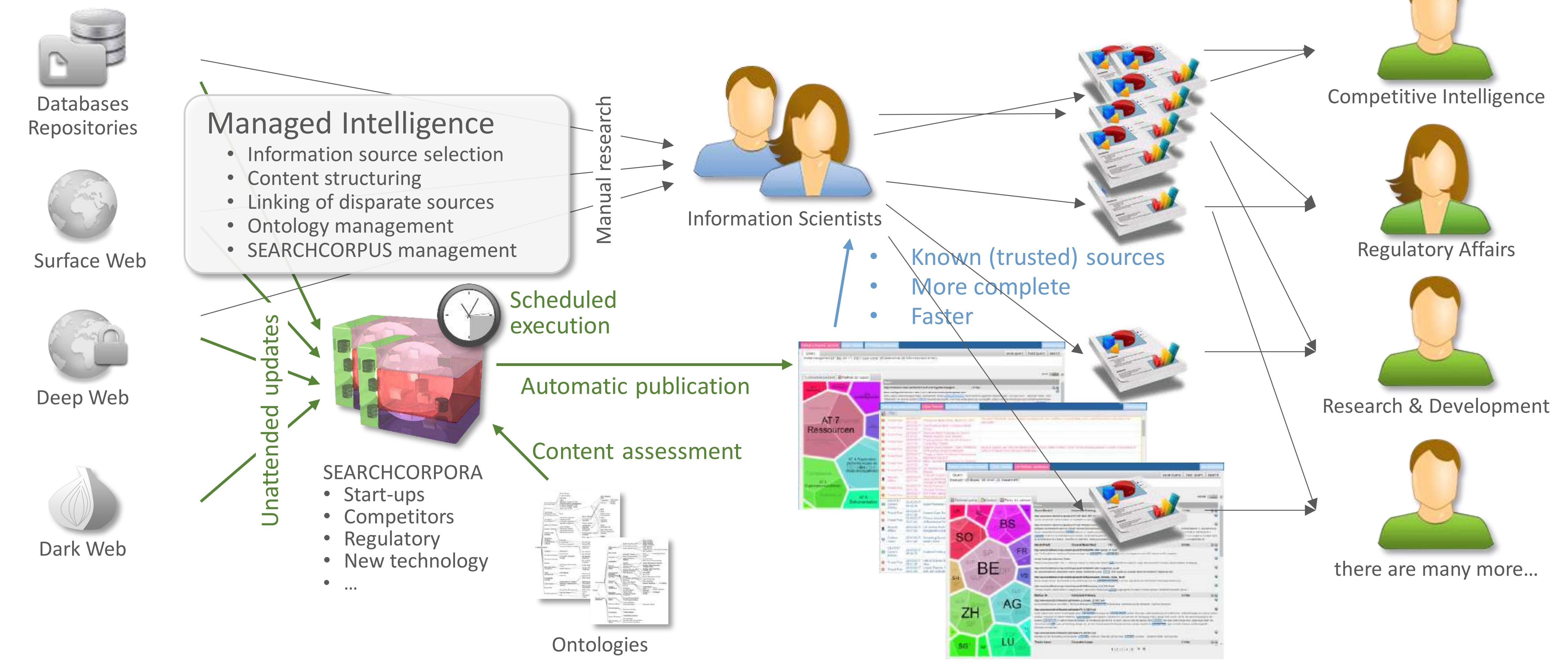
Web Big Data Deep SEARCH 9 Semantics Web

Managed Intelligence

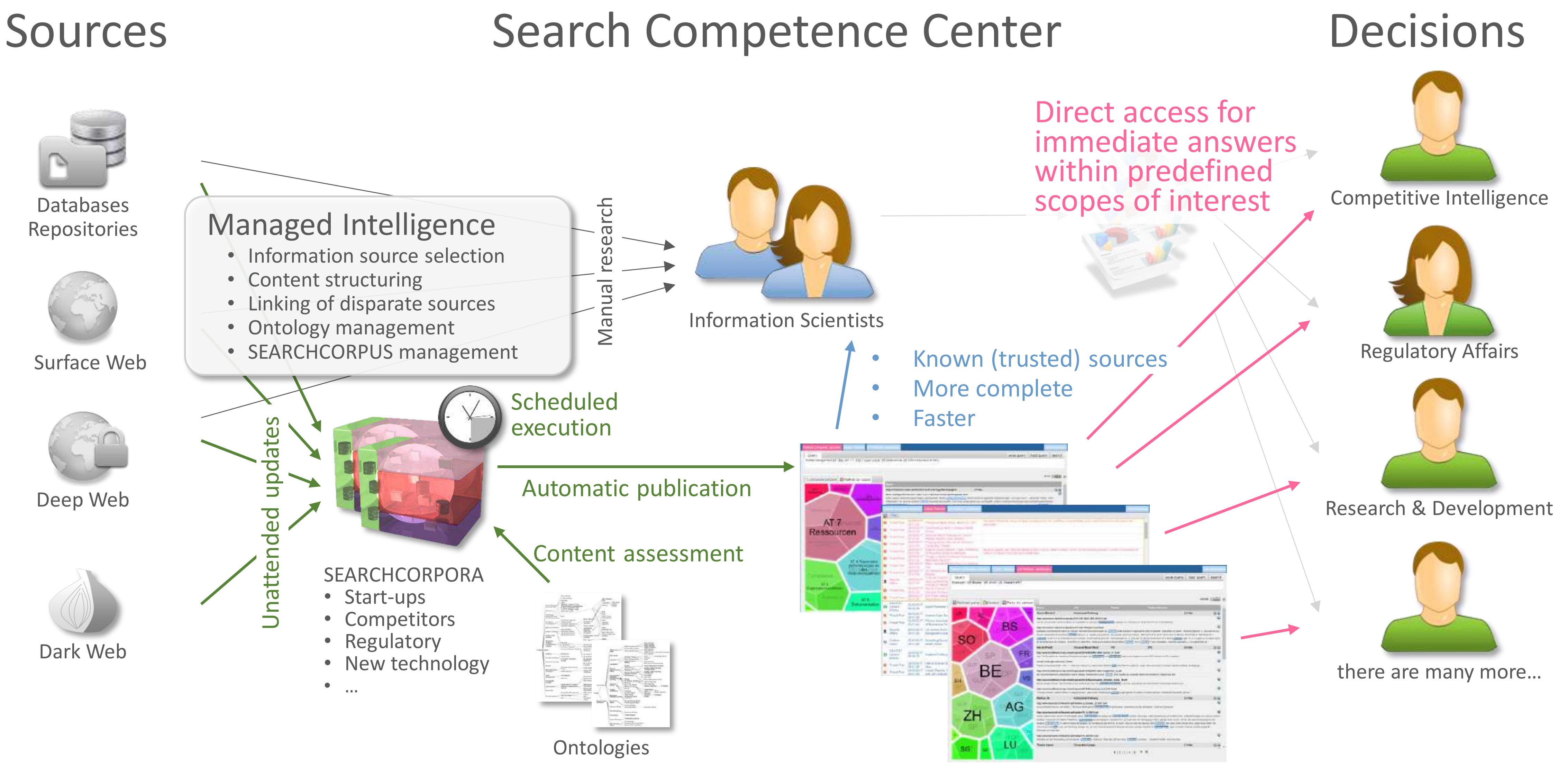
Sources

Search Competence Center

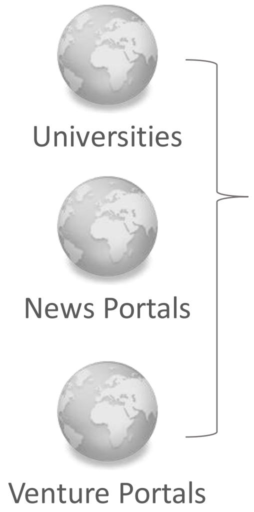
Decisions



Managed Intelligence



Grow the Data Base



Collect company names and URLs of websites from many different sources:

ca. 40.000 company websites

Grow the Data Base

Universities

News Portals

Venture Portals

Collect company names and URLs of websites from many different sources

crunchbase

e.g. 700.000 companies listed on Crunchbase

…

10% of company websites are of interest

Grow the Data Base

Universities

News Portals

Venture Portals

Collect company names and URLs of websites from many different sources

crunchbase

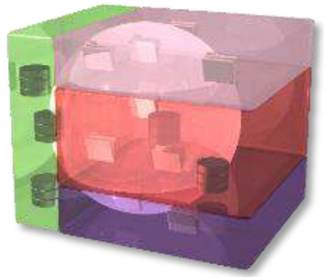
e.g. 700.000 companies listed on Crunchbase

Master SEARCHCORPUS®
- ca. 100.000 websites
- Millions of web pages,
- Documents
- PDFs,
- …

…

10% of company websites are of interest

Focused Crawlers



>5 TB content

Tagging 5 TB?

Web Big Data Deep SEARCH 9 Semantics Web

Universities

News Portals

Venture Portals

Collect company names and URLs of websites from many different sources

crunchbase

e.g. 700.000 companies listed on Crunchbase

ca. 100.000 company websites are of interest

Focused Crawlers

Master SEARCHCORPUS®

- ca. 100.000 websites
- Millions of web pages,
- Documents
- PDFs,
- …

>5 TB content

Tagging with ontologies to build problem specific faceted semantic search engines.

Regulatory Affairs

Competitive Intelligence

Research & Development

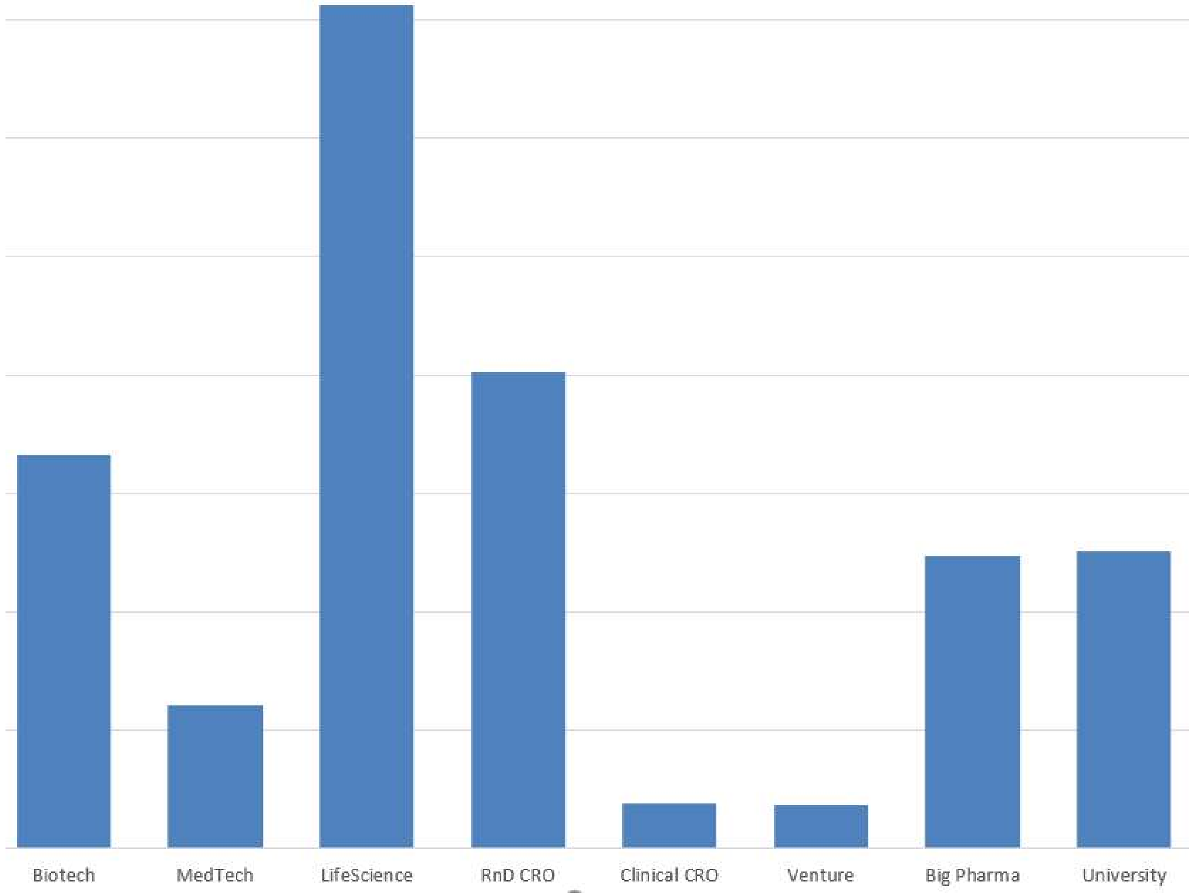
there are many more…

Cut Problem into Pieces

To reduce volume we need to filter early on…

- We use semantic search to filter for interesting research topics like diseases or treatment
- More interestingly:
  We can also filter by business model, development stage,

  i.e. anything that might be of interest



Website Classification

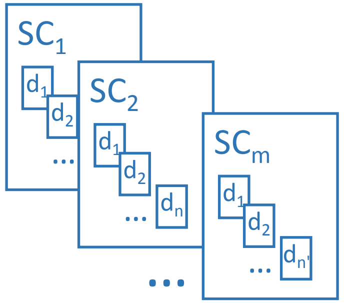
Requirements

- Classes are changing as new scopes of interest come up
- Company websites range from 1 page to 1000s of pages
- Companies may fall into several classes
- Training data could be < 50 samples, depending on class
- Data scientist must be able to create new classes on the fly

Classification using SVM

Support Vector Machine

We started to build feature vectors for SVM training using a classical TF IDF approach



Pick training data

Dictionary Builder

with $tf_{t_i SC_j} > tf_{min}$

	SC_1	SC_2	SC_3	…	SC_m
t_1	$tf_{t_1 SC_1}$	$tf_{t_1 SC_2}$	…		
t_2	$tf_{t_2 SC_1}$				
t_3	…				
…					
t_x					

Calculate TF IDF and train

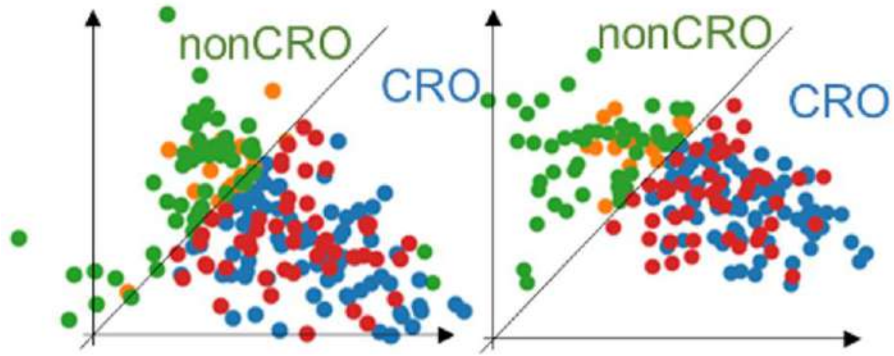
Loop until feature vector has a feasible size (problem specific)

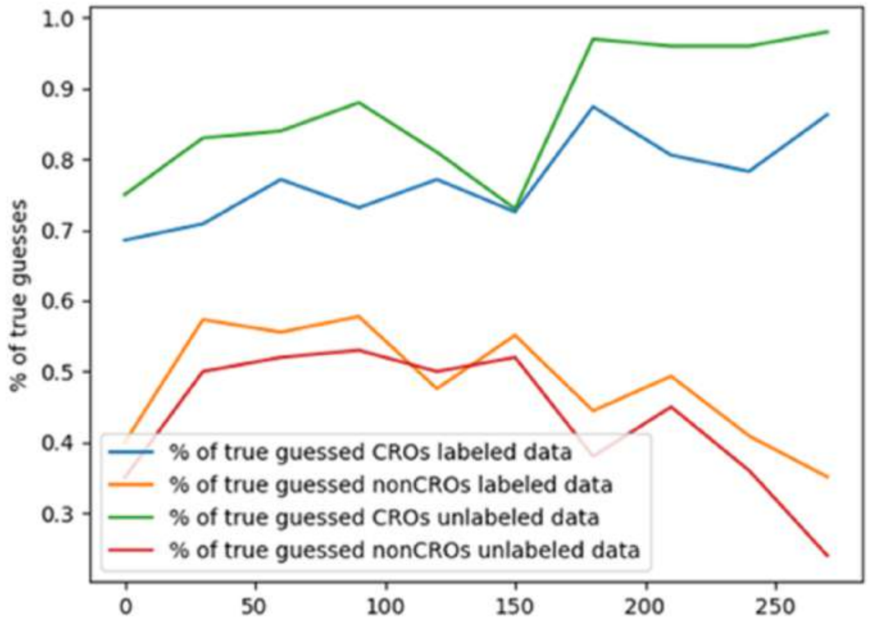
Classification Using SVM

Support Vector Machine

We started to build feature vectors for SVM training using a classical TF IDF approach

- No conversion, training sets too small and not representative enough





Normalization of Input Data

Semantic Technologies

Custom Dictionary

- Convert the generated TF based dictionary into an RDF ontology

Thesaurus for Normalization of Input Data

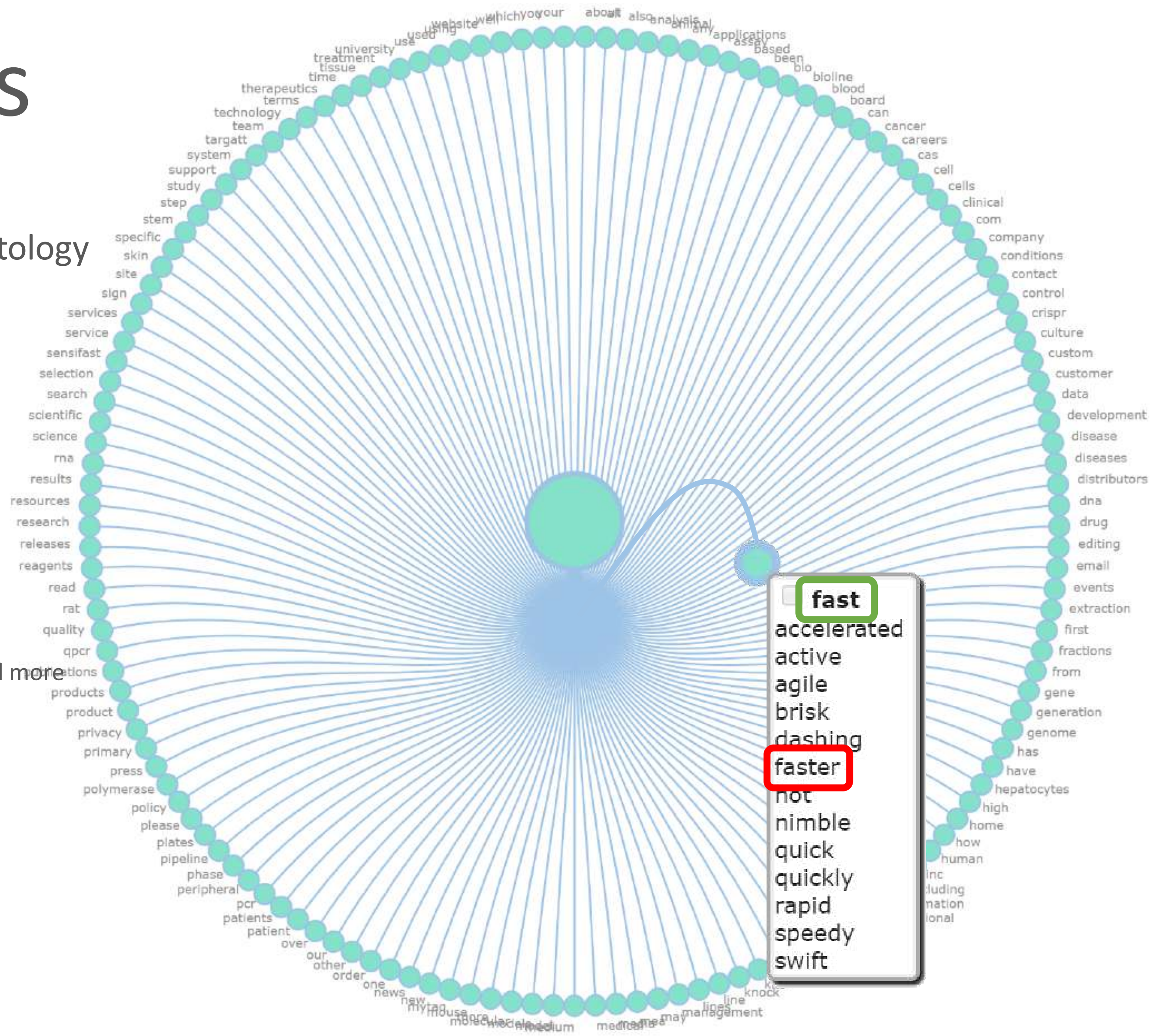
- Automatically fill the ontology with thesaurus data
- Manually optimize thesaurus in editor
- Normalize input data with thesaurus before classification
- Train SVM with normalized dictionary

Sample CRO Website Text

Our unique operation model propels you through the Proof of Concept phase faster and more efficiently, placing your cancer therapy on the road to success.

CRO Website Text after Normalization

Our unique business model help you through the proof phase fast and more effective, placing your cancer therapy on the road to success.

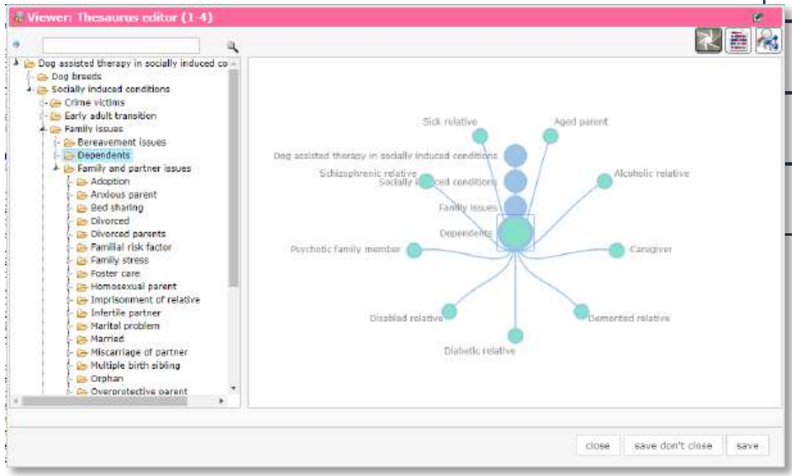


Normalization of Input Data

Support Vector Machine

Group synonyms, remove homonyms, watch out for polysemy, add synonyms from dictionaries, clean

Thesaurus editor



	SC_1	SC_2	SC_3	…	SC_m
~~t_1~~	~~$tf_{t_1 SC_1}$~~	~~$tf_{t_1 S_2}$~~	…		
t_2	$tf_{t_2 SC_1}$				
…					
t_x					

Homonyms removed

Synonyms added

Calculate TF IDF and train

If too many terms are removed from the feature vector, because they are actually synonyms of some other term, we may have to again build another dictionary.

Normalization of Input Data

Support Vector Machine (Trained were 2 classes. Verification against 150/130/250 websites)

20 websites used for training												
#	tf min	Stemmer	Edited thesaurus	% Testing	Nu	Gamma	Epsilon	% correct	% false positive	% false	% not recognized	% correct and false positive
6	50	yes	no	10%	0,1	0,01	0,001	37%	32%	5%	15%	12%
7	50	no	no	10%	0,1	0,01	0,001	39%	31%	7%	13%	10%
8	50	no	curated no synonyms added	10%	0,1	0,01	0,001	39%	29%	6%	16%	11%
9	20	yes	no	10%	0,1	0,01	0,001	64%	1%	0%	35%	0%
10	20	no	no	10%	0,1	0,01	0,001	62%	5%	0%	33%	0%
11	20	no	curated no synonyms	10%	0,1	0,01	0,001	67%	5%	0%	29%	0%
12	20	no	curated with synonyms	10%	0,1	0,01	0,001	72%	9%	1%	16%	1%
Now modifying training and model parameters												
13	20	no	curated with synonyms	25%	0,1	0,01	0,001	70%	11%	1%	16%	2%
14	20	no	curated with synonyms	5%	0,1	0,01	0,001	76%	6%	1%	17%	0%
15	20	no	curated with synonyms	5%	0,15	0,01	0,001	69%	12%	1%	17%	1%
16	20	no	curated with synonyms	5%	0,05	0,01	0,001	71%	9%	1%	18%	1%
17	20	no	curated with synonyms	5%	0,075	0,01	0,001	76%	6%	1%	17%	0%
18	20	no	curated with synonyms	5%	0,0875	0,01	0,001	70%	[illegible]	[illegible]	[illegible]	1%
19	20	no	curated with synonyms	5%	0,075	0,05	0,001	74%	[illegible]	[illegible]	[illegible]	0%
20	20	no	curated with synonyms	5%	0,075	0,1	0,001	71%	0%	0%	29%	0%
21	20	no	curated with synonyms	5%	0,1	0,05	0,001	76%	0%	0%	24%	0%
22	20	no	curated with synonyms	5%	0,2	0,05	0,001	71%	0%	0%	29%	0%
23	20	no	curated with synonyms	5%	0,15	0,05	0,001	72%	0%	0%	28%	0%
24	20	no	curated with synonyms	5%	0,1	0,05	0,01	72%	0%	0%	28%	0%

We got some pretty good results but could not get any better

Website Classification

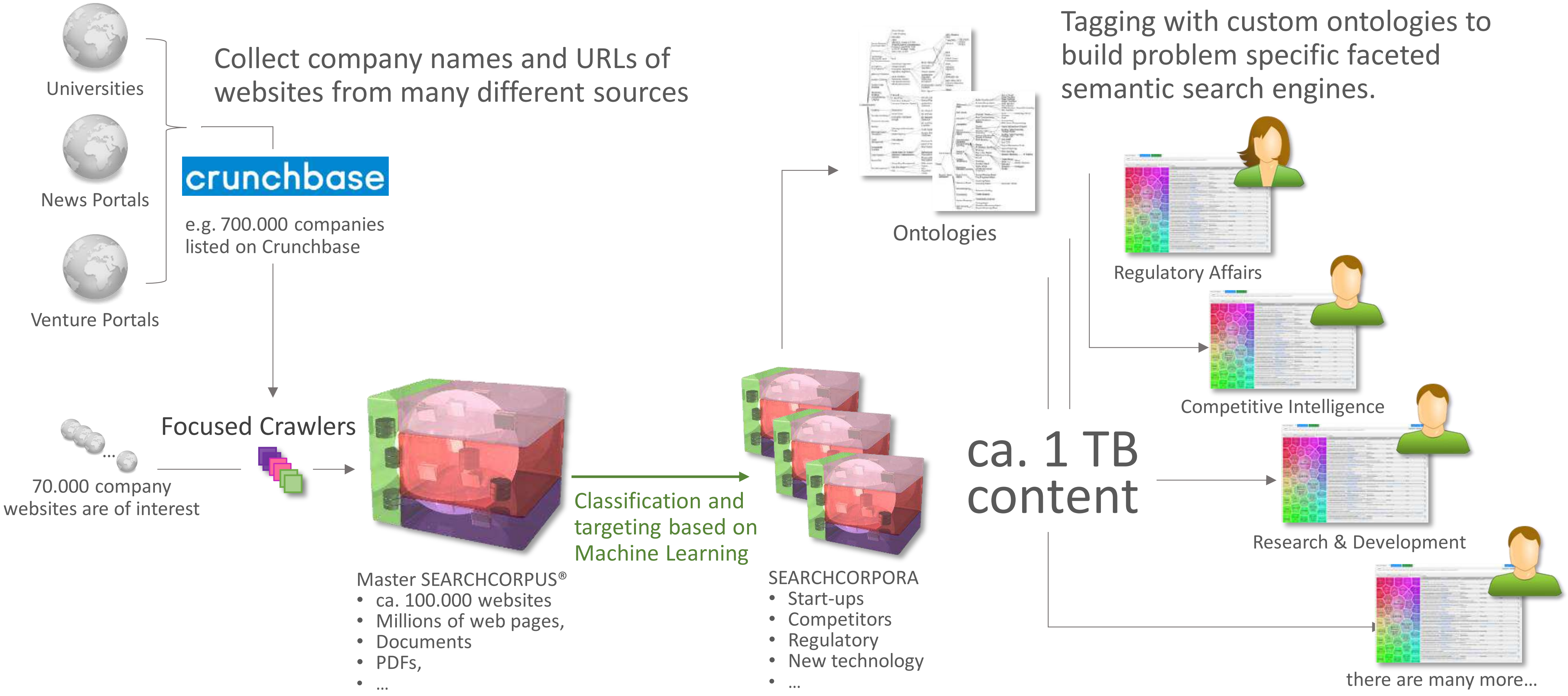
Problem

- We cannot find an exhaustive set of representative negative examples
- Therefore, we need to use 1-class SVM
- But TF IDF is not suitable for 1-class classification because it penalizes terms that appear in many documents
- Instead use Hadamard Product, which reinforces such terms[1)]

$$\begin{pmatrix} tf_{t_i sc_i} \\ tf_{t_{i+1} sc_i} \\ \dots \\ \\ tf_{t_x sc_i} \end{pmatrix} \otimes \begin{pmatrix} tf_{t_i sc} \\ tf_{t_{i+1} sc} \\ \dots \\ \\ tf_{t_x sc} \end{pmatrix} = \begin{pmatrix} tf_{t_i sc_i} tf_{t_i sc} \\ tf_{t_{i+1} sc_{i+1}} tf_{t_{i+1} sc} \\ \dots \\ \\ tf_{t_x sc_x} tf_{t_x sc} \end{pmatrix} \xrightarrow{\text{scale}} [0\,;1]$$

1) See One-class document classification via Neural Networks, Larry Manevitz, Malik Yousef, 2007

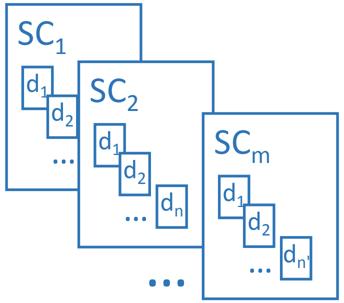
Now we are processing…



Sparse Training Data Classification

Normalized Input SVM

Thesaurus based input data normalization can optimize SVM classification with sparse training data:



→ Normalize with optimized Thesaurus →

	SC_1	SC_2	SC_3	…	SC_m
t_1	tf_{t_1,SC_1}	tf_{t_1,SC_2}	…		
t_2	tf_{t_2,SC_1}				
t_3	…				
…					
t_x					

→ Calculate HP, scale, classify

- Normalize input with manually curated thesaurus,
- Use Hadamard product to generate feature vectors
- Scale
- Then classify

Using Semantic Technology to Solve Sparse Training Material Problem in Machine Learning for Classification of Company Websites

Klaus Kater
Deep SEARCH 9 GmbH
Managing Partner

klaus.kater@deepsearchnine.com

SEMANTiCS 2018

Where Machine Learning Meets Semantics
10th - 13th of September 2018 in Vienna