Taming the “Outside”
in Outside Insight Applications

Meltwater

FAIRHAIR

About me...

Academia

BSc, MSc, and PhD
Computer Science and Engineering
Milan Polytechnic University

Senior Researcher (Oxford Martin Fellow)
Department of Computer Science
Oxford University

Assistant Professor
School of Computer Science
University of Birmingham

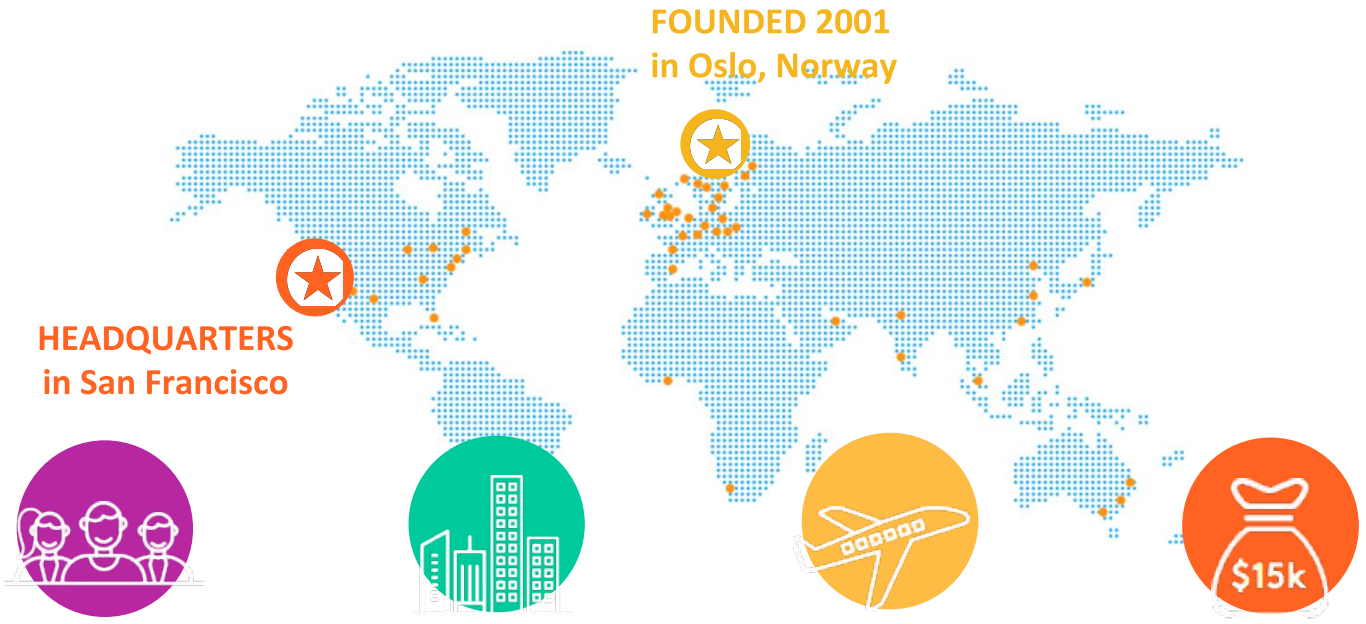
Industry

Research Assistant
Intern

Co-founder
Head of Data Engineering
Wrapidity

Senior Research Scientist
NLP Lead

About Meltwater



1500 employees worldwide

28,000 corporate clients

50 offices across 6 continents

Bootstrapped
No venture funding

Meltwater

Initiatives

SHACK15
The Data Science Hub

6 Data Science Hubs (co-working spaces)

- ✔ London
- ✔ San Francisco
- ✔ Singapore
- ✔ Sydney
- ✔ Berlin
- ✔ New York

mest
meltwater entrepreneurial school of technology

Meltwater Entrepreneurial School of Technology

- HQ in Accra, Ghana
- Training program for African entrepreneurs
- Incubator (25+ startups)
- Networking hub

University collaborations

Stanford University

EURECOM
Sophia Antipolis

UNIVERSITÀ DELLA CALABRIA

Carnegie Mellon University

Meltwater

Meltwater: Media Intelligence

Sources: Editorial, Social, Broadcasts

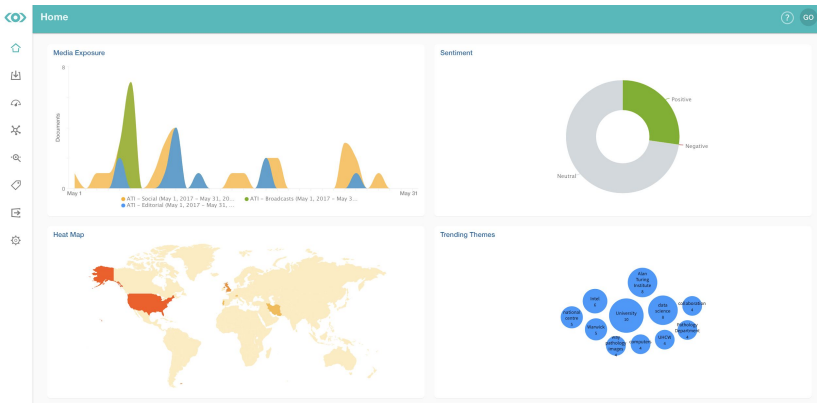
media exposure

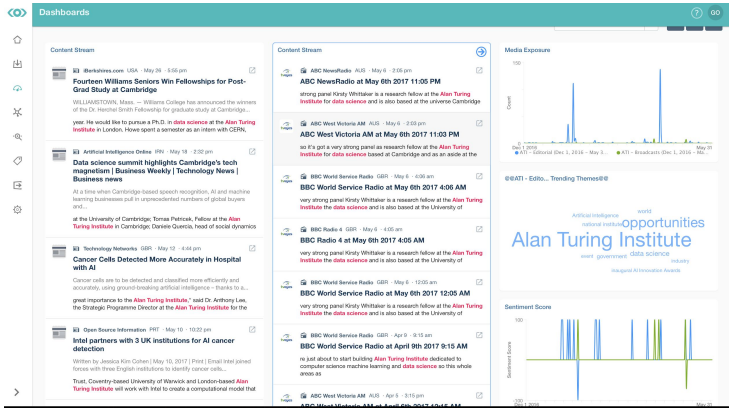
trends

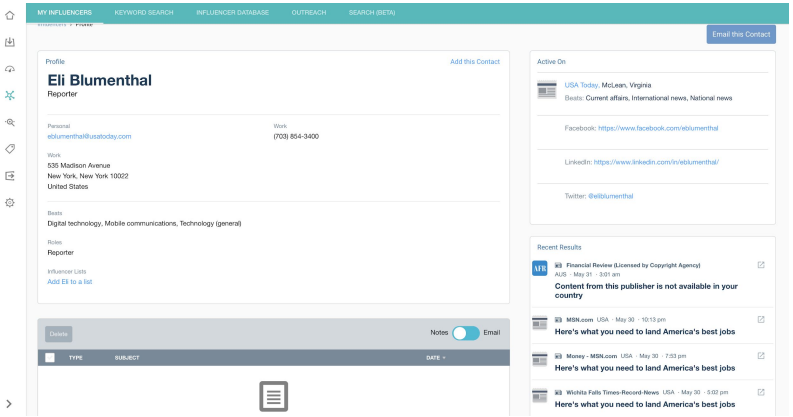
influencers

sentiment analysis

Ks millions of user queries, 300k distinct types







Meltwater

Big data company: We process 100 million documents and hundreds of millions of searches every day

Meltwater

A turning point

Build a world class AI platform for a new software category

Outside Insight

FAIRHAIR AI

Meltwater Outside Insight

https://outsideinsight.com/

Business Intelligence vs Outside Insight

BI is very (today) “ERP” centric. Dashboards and KPIs are built using internal data

- sales volumes
- operational costs
- pricing strategies

The assumption is that the past can predict the future, but there’s a big problem.

Companies are open systems, meaning they are influenced by external agents and factors

If you want to leverage the past you need to know whole of it including what happens (and happened) outside your organization.

From lagging to leading performance indicators

Background: Data, Information, Knowledge

What is Data?

- A set of values of qualitative and quantitative variables
  - x=75
  - y=1,850
  - z=[1,0,2,-3,2]

What is Information?

- Data with a way of interpreting it, e.g., via a schema
  - x=75, y=1,850 + [x → house number, y → birthyear]
  - x=75, y=1,850 + [x → age, y → height(mm)]

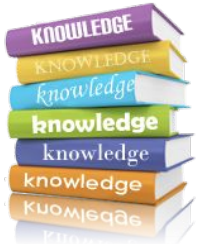
What is Knowledge?

- Available information for rational action
  - x=75, y=1,850 + [x → house number, y → birthyear] -> delete from electoral roll
  - x=75, y=1,850 + [x → age, y → height(mm)] -> provide free medicines

Background: Cognitive Systems

Adaptive and interactive systems that...

enhance human cognition through



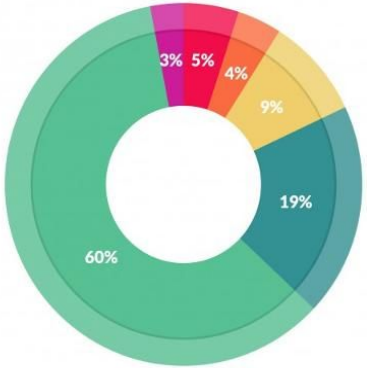
knowledge

reasoning

learning

... give to the right people, the right information, at the right time.

Roadblocks



What data scientists spend the most time doing

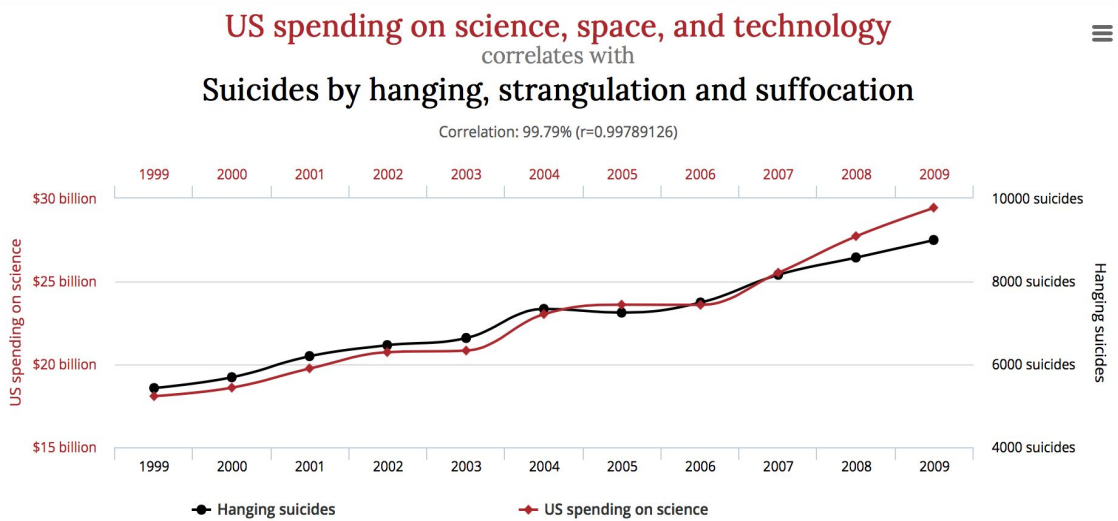
- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Data is deceitful. Need a systematic way to mine, propose, and explain possible insights

- we need factual knowledge
- combine (machine) learning and reasoning

Access to structured data

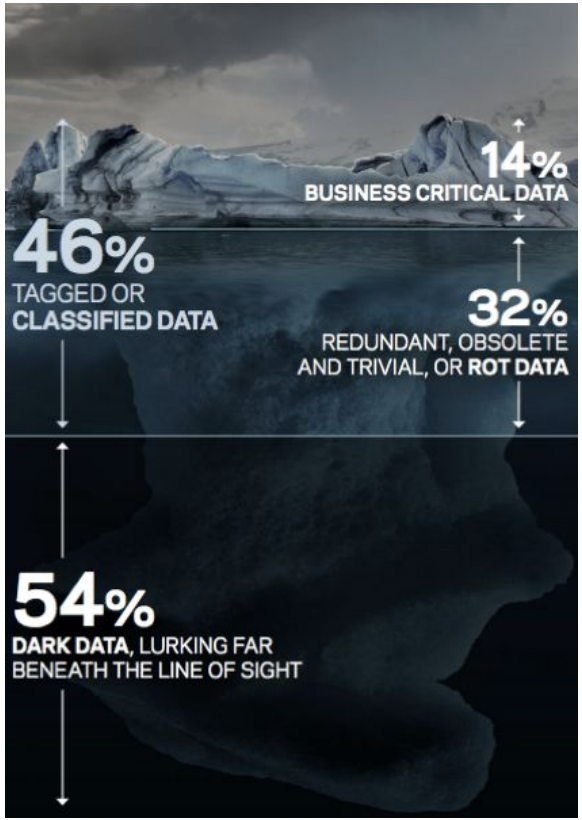
- Make sure the data is clean, complete & normalized
- Make it relevant by connecting the dots
- Bring the methods close to the data



Source: www.tylervigen.com (Spurious Correlations)

Meltwater

Why is that so hard?



Converging trends in data management:

- Dark Data: semi / unstructured data, buried in web pages, PDFs, plain text, …
- Data Preparation: preparing and maintain data for mining and analytics

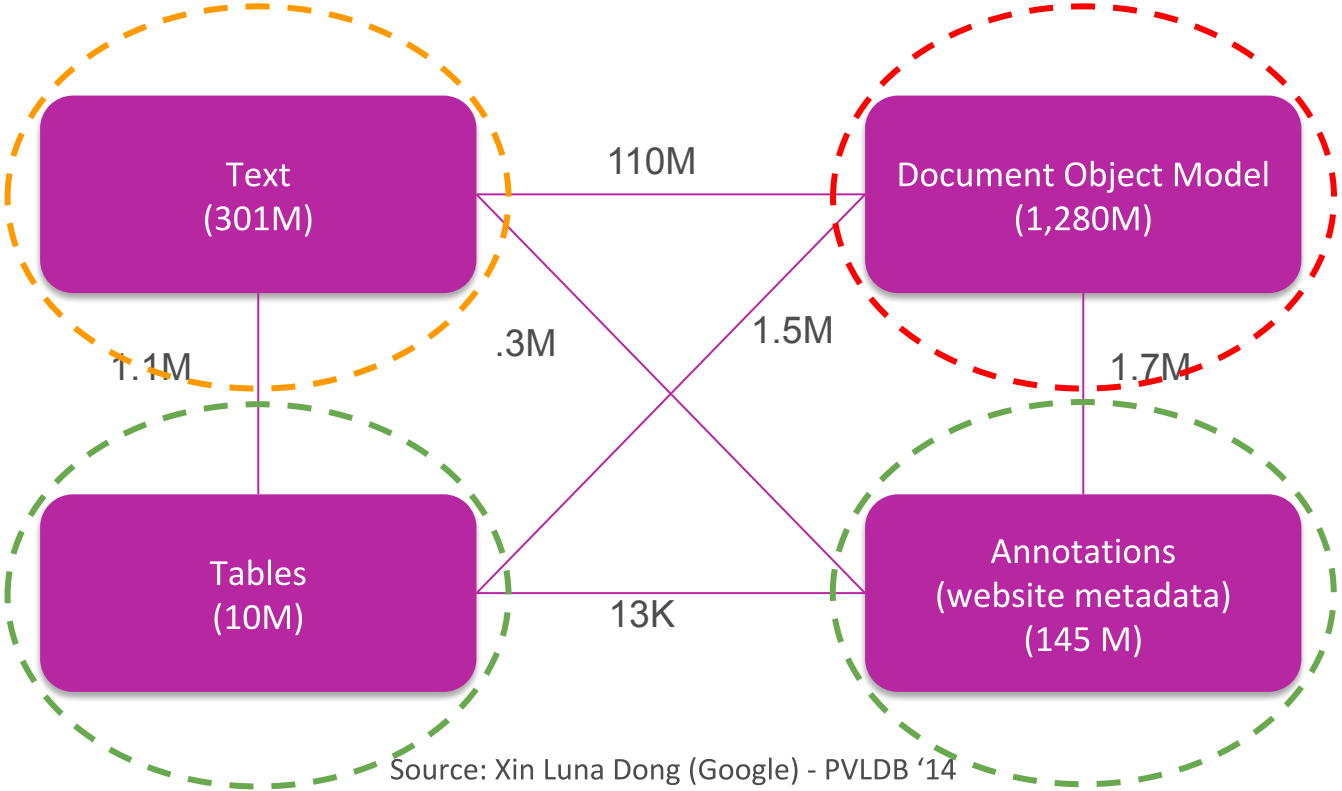
- Outside Insight: combine internal data with external and dark data for comprehensive knowledge

What is outside data

Scenario: build an app that compares your company against its competitors over time

Online News	Social Media	Financial Filings	Corporate Websites
Web traffic	Press Releases	Interest Rates	Blogs
Share Price	Job Postings	Patent Filings	Product Reviews
App Downloads	Online ad Spend	Court Documents	Forums
Oil Prices	Weather Data	Real Estate Rates	Unemployment

Where is information usually found?



Source: Xin Luna Dong (Google) - PVLDB '14

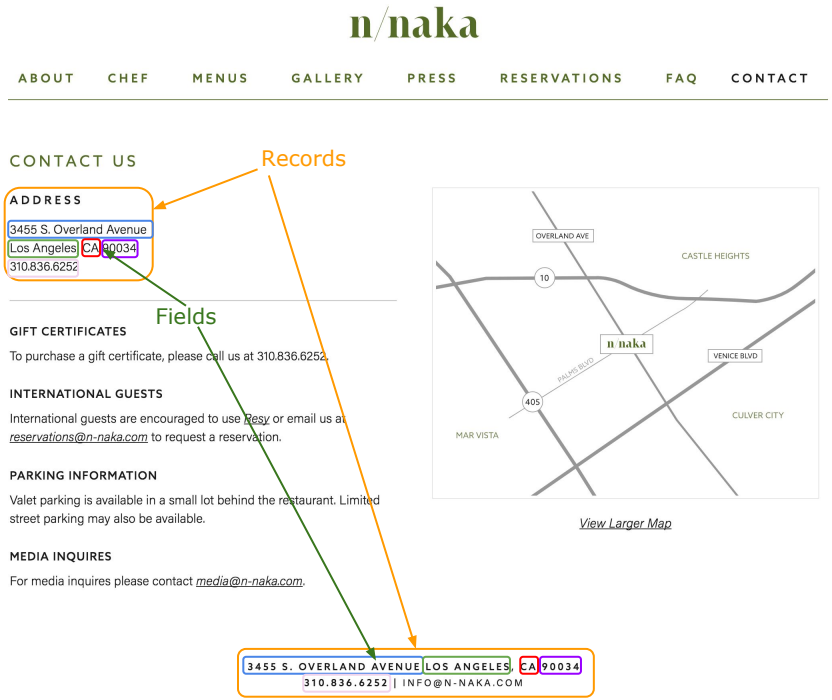
Web Data Extraction

Process or turning semi-structured (templated) web data into structured data

refcode	postcode	bedrooms	bathrooms	available	price
33453	OX2 6AR	3	2	15/10/2013	£1280 pcm
33433	OX4 7DG	2	1	18/04/2013	£995 pcm

Web Data Extraction vs Information Extraction

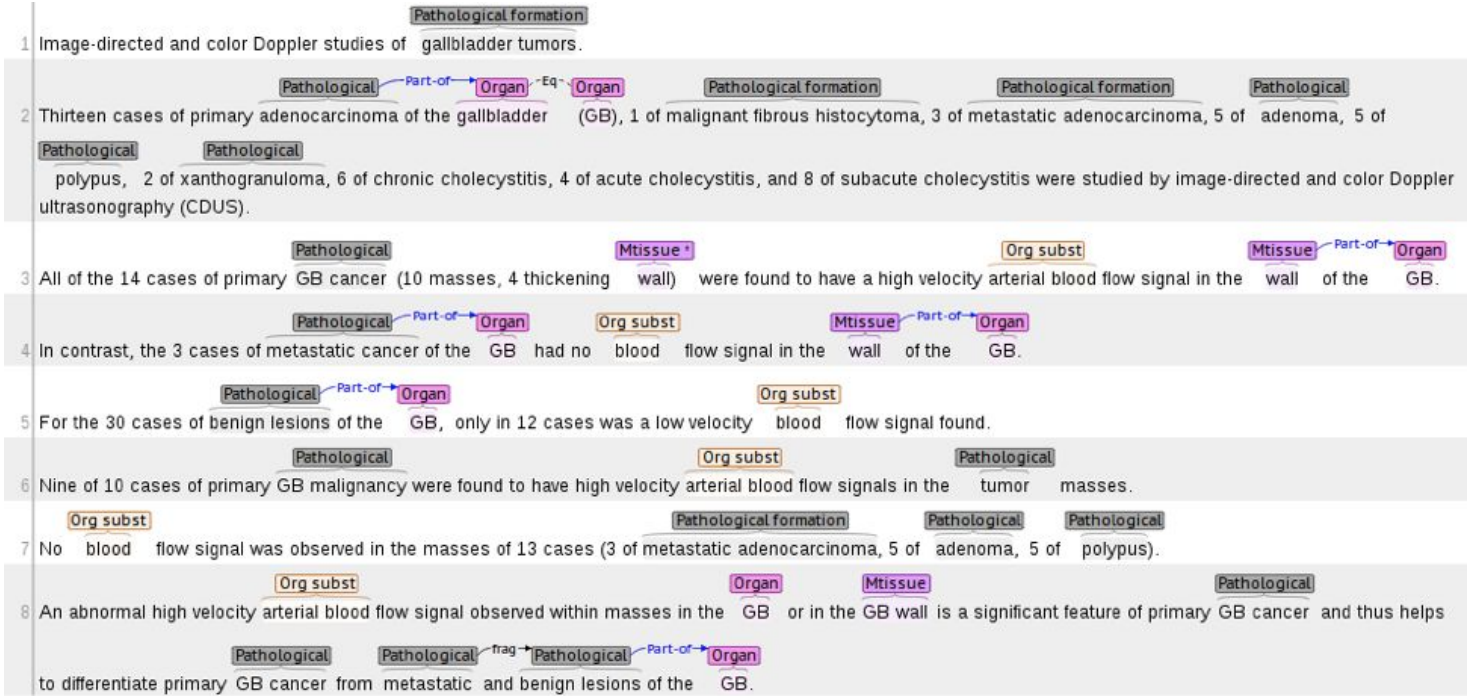
Data is structured according to templates, annotated, or semantically-related blocks



Meltwater

Web Data Extraction vs Information Extraction

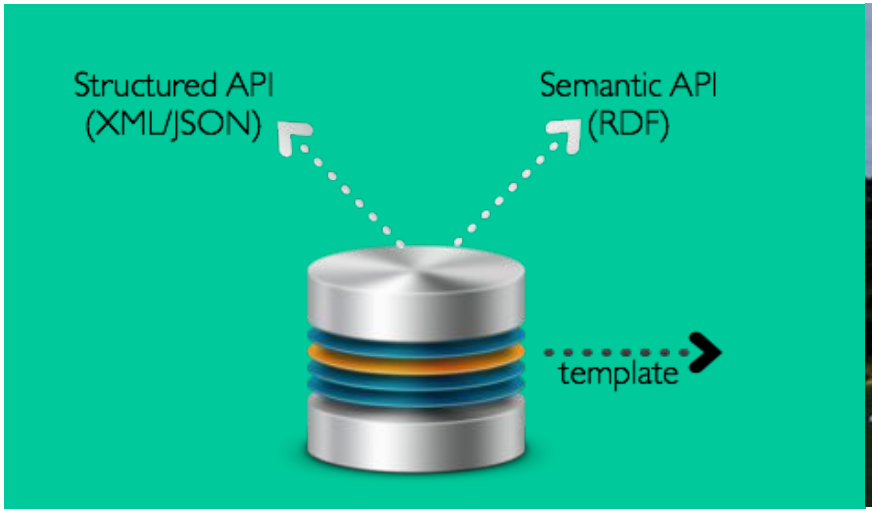
Data is hidden in linguistic structures (entities, relations). Culture and language specific



Meltwater

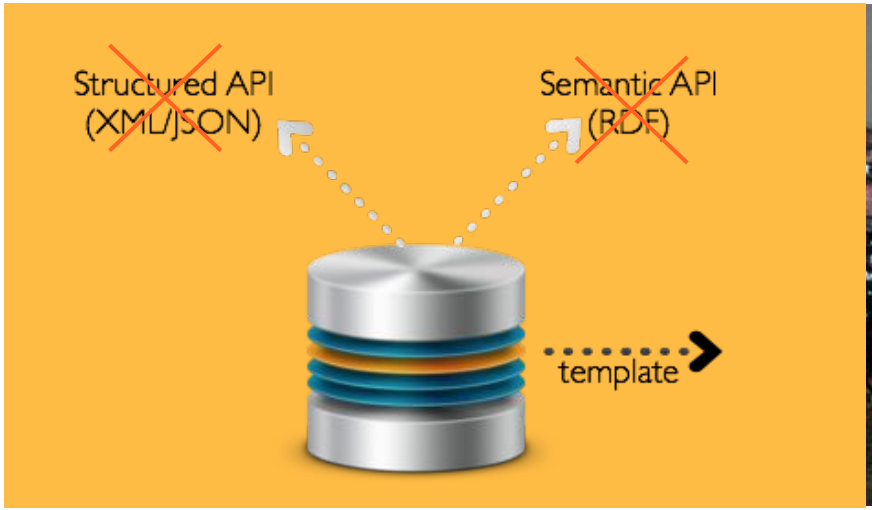
The academic web

- Microdata and the semantic web have solved the problem
- All the data is in web tables
- APIs provide all the structured data you need



The real web

- API's are limited to large websites
- Web tables and microdata are marginal
- The real problem is not one-time extraction, but keeping the data up-to-date over time



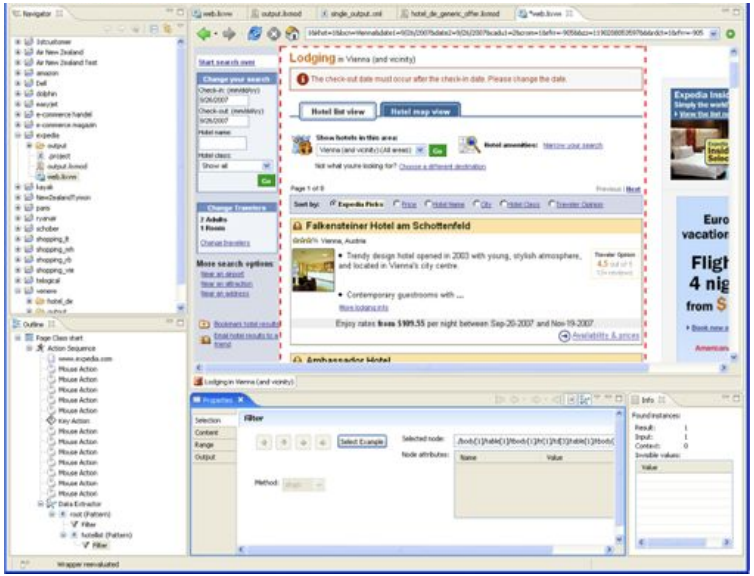
The real web

Manual or supervised wrapper construction is expensive ...

Example 1: A large social network started to harvest all business locations and add them to their social graph...
Result: after 6 months and 10 (expensive) engineers, got 60 restaurant chains in. Wrappers break all the time, company decided to give up and start buying data from third parties, e.g., Factual

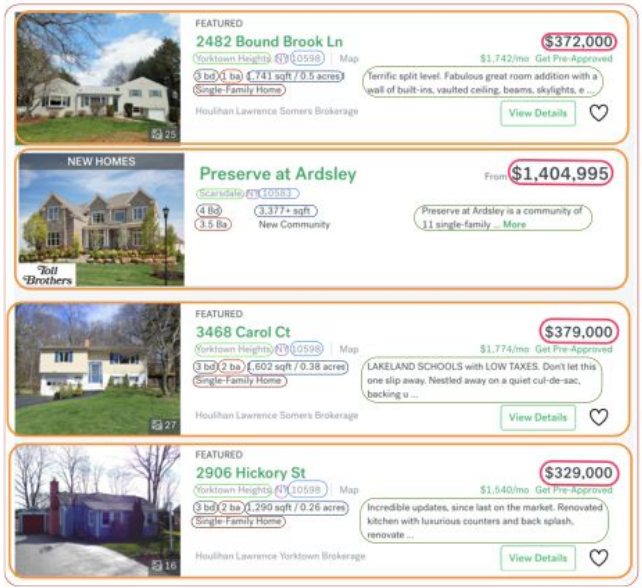
Example 2: A large travel website trying to harvest offers from airlines, hotels, car rentals, and restaurants locations and add them to their travel graph...
Result: after 5 years they have ~900 airlines wrapped, but use ~50 engineers just to maintain the wrappers. They wanted to wrap ~150k Hotel websites... do the math of the engineering resources required.

Web Data Extraction: Techniques



manual / (semi)-supervised

human intensive, not scalable



unsupervised

data intensive, scalable

mozenda

Mozenda Overview

scrapinghub

import.io

diffbot

Meltwater

Web Data Extraction: Techniques

- Wrapper Induction: similar objects are presented in similar structures
  - You need training data (web = high sample complexity + many features) use human to inform system (supervision / crowd)
  - Fully unsupervised methods perform poorly beyond simple structures, e.g., because of regular noise.

Valter Crescenzi, Giansalvatore Mecca:
Automatic information extraction from large websites. *J. ACM 51(5): 731-779 (2004)*

Tim Furche, Jinsong Guo, Sebastian Maneth, Christian Schallhart:
Robust and Noise Resistant Wrapper Induction. *SIGMOD Conference 2016: 773-784*

- Fact redundancy (e.g., Google Knowledge Vault), true facts are repeated many times on different websites
  - works well with highly-redundant common-sense facts (London, Barack Obama)
  - need to take care about copying and source-level trust
  - Ephemeral and infrequent entities get lost or noisy.
    - e.g., price of a product item → ranges instead of single data point

Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Kevin Murphy, Shaohua Sun, Wei Zhang: **From Data Fusion to Knowledge Fusion.** *PVLDB 7(10): 881-892 (2014)*

Fully automated full-site web data extraction

Bringing Web Data Extraction to the real web and at industrial scale

Key Insights:

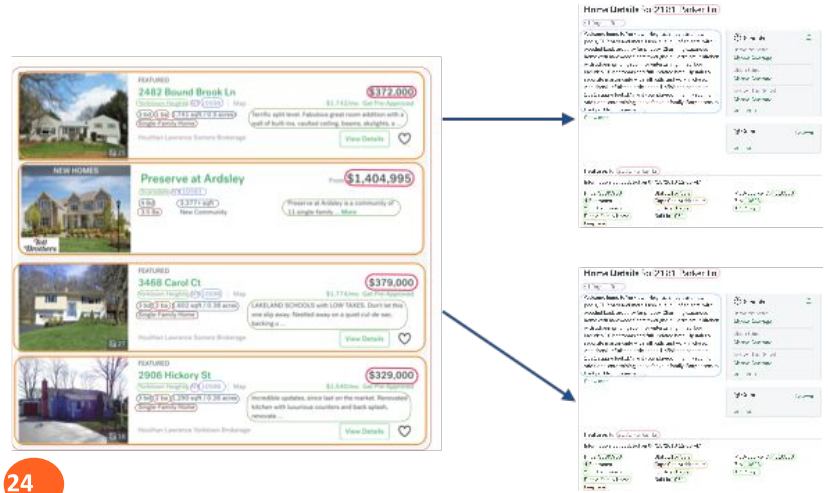
- Replace site supervision with domain knowledge
- Automate the exploration process, e.g., form filling, focused crawling
- Make wrapper induction algorithms knowledge-parametric (both ML and Rules)
- Make wrapper induction algorithms robust to noise

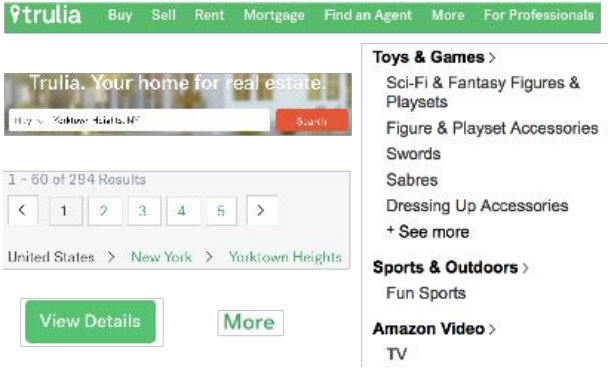
Tim Furche, Georg Gottlob, Giovanni Grasso, Xiaonan Guo, Giorgio Orsi, Christian Schallhart, Cheng Wang:
DIADEM: Thousands of Websites to a Single Database. *PVLDB 7(14): 1845-1856 (2014)*

Full-site web data extraction

Navigation

forms, menus, categories, bread crumbs, pagination, infinite scroll, detail links





Template discovery

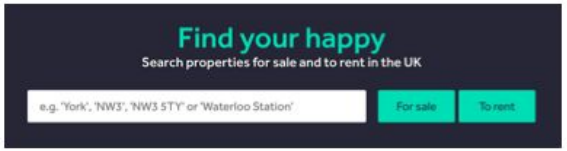
Result pages, detail pages

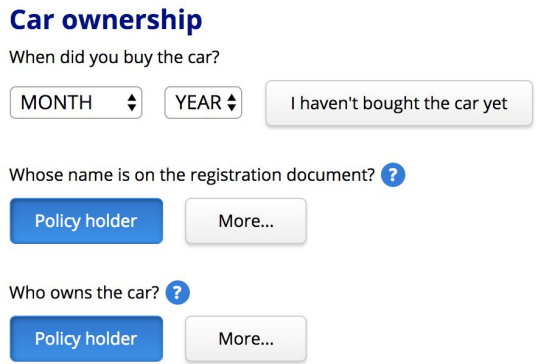
Meltwater

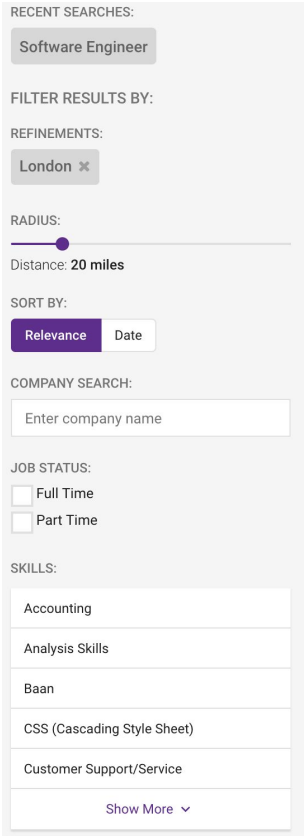
Full-site web data extraction

Forms

The hardest things to deal with in Web Data Extraction







Tasks:

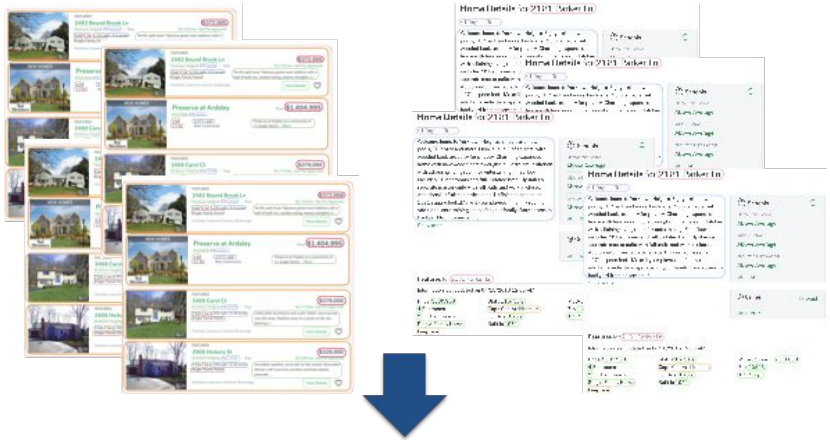
- form Understanding and querying
- form labelling
- field grouping
- form filling / querying

Tim Furche, Georg Gottlob, Giovanni Grasso, Xiaonan Guo, Giorgio Orsi, Christian Schallhart:
The ontological key: automatically understanding and integrating forms to access the deep Web.
VLDB J. 22(5): 615-640 (2013)

Michael Benedikt, Balder ten Cate, Efthymia Tsamoura:
Generating Plans from Proofs. *ACM Trans. Database Syst. 40(4): 22:1-22:45 (2016)*

Meltwater

Full-site web data extraction



Wrapper induction
generalisation and weaving

```
doc('http://www.wwagency.com/')//label[@for='sale_type_id']/following-sibling::select/{0 /}
      //form/div[@class='formbtn-ctn'][last()]/button[@class='formbtn']/{click /}
   /.:<data_area>[?.//div[@class='pagenumlinks'][1]//span/text():<number_results=.>]
   ///div[contains(@class,'proplist_wrap')]/following-sibling::div//a[@class='pagenum'][last()]/{nextclick /})*
      //div[contains(@class,'proplist_wrap')]:<record>[? .:<origin_url=current-url()>]
            [? .//span[@class='prop_price']/text():<price=normalize-space(.)> ]
            [? .//span[.='Type:']/following-sibling::strong/text():<property_type=normalize-space(.)> ]
            [? .//div[@class='prop_statuses']//text():<property_status=normalize-space(.)> ]
            [? .//span[.='Bathrooms:']/following-sibling::strong/text():<bathroom_number=normalize-space(.)> ]
            [? .//span[.='Bedrooms:']/following-sibling::strong/text():<bedroom_number=normalize-space(.)> ]
            [? .//strong[@class='orange']/preceding-sibling::text():<location_raw=string(.)> ]
            [? .//strong[@class='orange']/text():<postcode=normalize-space(.)> ]
            [? .//strong/preceding-sibling::strong/text():<street_address=normalize-space(.)> ]
            [? .//@src:<image=normalize-space(.)> ]
            [? .//div[@class='prop_statuses']/following-sibling::a/@href:<url=normalize-space(.)> ]
            [? .//div[@class='prop_maininfo']:<description=normalize-space(.)> ]
```


Tim Furche, Jinsong Guo, Sebastian Maneth, Christian Schallhart:
Robust and Noise Resistant Wrapper Induction. *SIGMOD Conference 2016: 773-784*

Extraction Language: OXPath

https://github.com/oxpath/oxpath.github.io

```
doc('http://www.trulia.com/')//label[@for='sale_type_id']/following-sibling::select/{0 /}
      //form/div[@class='formbtn-ctn'][last()]/button[@class='formbtn']/{click /}
  /.:<data_area>[?.//div[@class='pagenumlinks'][1]//span/text():<number_results=.>]
   /(//div[contains(@class,'proplist_wrap')]/following-sibling::div//a[@class='pagenum'][last()]/{nextclick /})*
     //div[contains(@class,'proplist_wrap')]:<record>[? .:<origin_url=current-url()>]
      [? .//span[@class='prop_price']/text():<price=normalize-space(.)> ]
      [? .//span[.='Type:']/following-sibling::strong/text():<property_type=normalize-space(.)> ]
      [? .//div[@class='prop_statuses']//text():<property_status=normalize-space(.)> ]
      [? .//span[.='Bathrooms:']/following-sibling::strong/text():<bathroom_number=normalize-space(.)> ]
      [? .//span[.='Bedrooms:']/following-sibling::strong/text():<bedroom_number=normalize-space(.)> ]
      [? .//strong[@class='orange']/preceding-sibling::text():<location_raw=string(.)> ]
      [? .//strong[@class='orange']/text():<postcode=normalize-space(.)> ]
      [? .//strong/preceding-sibling::strong/text():<street_address=normalize-space(.)> ]
      [? .//@src:<image=normalize-space(.)> ]
      [? .//div[@class='prop_statuses']/following-sibling::a/@href:<url=normalize-space(.)> ]
      [? .//div[@class='prop_maininfo']:<description=normalize-space(.)> ]
```

Navigation

Record & attributes

Tim Furche, Georg Gottlob, Giovanni Grasso, Christian Schallhart, Andrew Jon Sellers:
OXPath: A language for scalable data extraction, automation, and crawling on the deep web.
VLDB J. 22(1): 47-72 (2013)

Extraction Language: JSON-XPath

A JSON-like wrapper specification

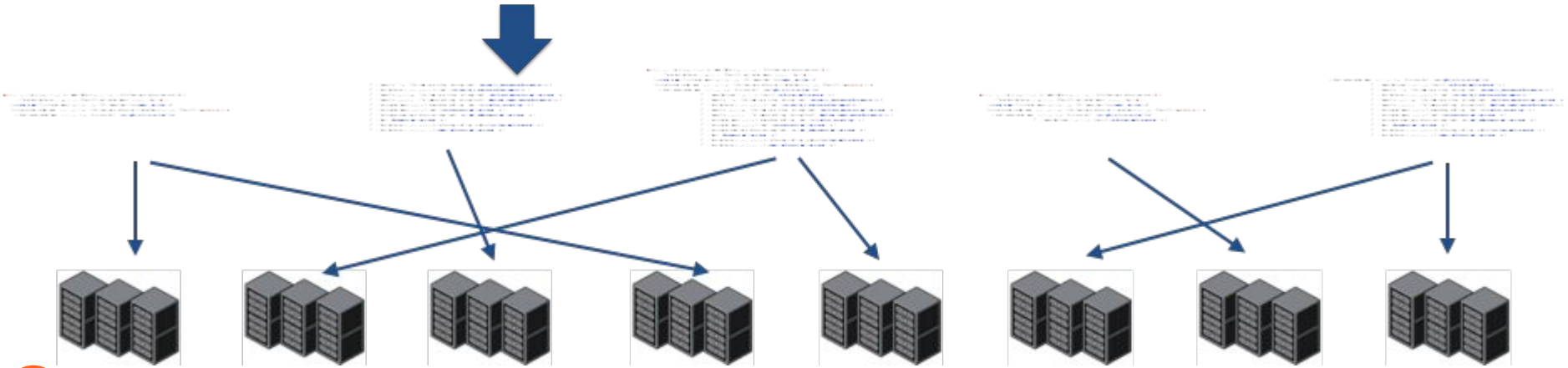
Advantage? You can query the program as well!

```
"articleTpls": [
 {
   "id": "article_template_0",
   "startUrls": [
     "http://www-03.ibm.com/press/us/en/pressrelease/33304.wss",
     "http://www-03.ibm.com/press/us/en/pressrelease/33420.wss",
     "http://www-03.ibm.com/press/us/en/pressrelease/33117.wss",
     "http://www-03.ibm.com/press/us/en/pressrelease/33303.wss"
   ],
   "urlPatterns": [
     "(?<wordset>([a-zA-Z]{1,}[:]{1,}){1,1})//(?<wordnumberset>([\\w]{1,}[\\-.]{1,}){1,3}[\
       \w]{1,})/(?<wordset1>([a-zA-Z]{1,}[/]{1,}){1,3}[a-zA-Z]{1,})/(?<wordnumberset1>([\\w]
       {1,}[\\.]{1,}){1,1}[\\w]{1,})"
   ],
   "titleXpath": "wrty:normalize-space(//h1[@class='ibm-small'])",
   "bylineXpath": "//div[@class='ibm-two-column']//strong",
   "ingressXpath": "wrty:normalize-space(//div[@id='ibm-content-main']/div[@class='ibm-
       container'][1]//p[1])",
   "contentXpath": {
     "includeXpath": "wrty:normalize-space(wrty:string-join(//div[@id='ibm-content-main']//
     div[@class='ibm-container-body']/node()[self::p|self::h2[@class='ibm-inner-subhead']],
     \"\\n\"))"
     },
   "engagementPatterns": [],
   "imagePatterns": [
     {
       "baseXpath": "//img[@width='500']",
       "urlXpath": ".",      }
   ],
   "authorPatterns": [
   {
       "baseXpath": "//div[@class='ibm-two-column']//strong",
       "nameXpath": "wrty:normalize-space(.)",
    }
  ]
}
```

Full-site web data extraction

```
doc('http://www.wwagency.com/')//label[@for='sale_type_id']/following-sibling::select/{0 /}
       //form/div[@class='formbtn-ctn'][last()]/button[@class='formbtn']/{click /}
   /.:<data_area>[?.//div[@class='pagenumlinks'][1]//span/text():<number_results=.>]
   /(//div[contains(@class,'proplist_wrap')]/following-sibling::div//a[@class='pagenum'][last()]/{nextclick /})*
      //div[contains(@class,'proplist_wrap')]:<record>[? .:<origin_url=current-url()>]
            [? .//span[@class='prop_price']/text():<price=normalize-space(.)> ]
            [? .//span[.='Type:']/following-sibling::strong/text():<property_type=normalize-space(.)> ]
            [? .//div[@class='prop_statuses']//text():<property_status=normalize-space(.)> ]
            [? .//span[.='Bathrooms:']/following-sibling::strong/text():<bathroom_number=normalize-space(.)> ]
            [? .//span[.='Bedrooms:']/following-sibling::strong/text():<bedroom_number=normalize-space(.)> ]
            [? .//strong[@class='orange']/preceding-sibling::text():<location_raw=string(.)> ]
            [? .//strong[@class='orange']/text():<postcode=normalize-space(.)> ]
            [? .//strong/preceding-sibling::strong/text():<street_address=normalize-space(.)> ]
            [? .//@src:<image=normalize-space(.)> ]
            [? .//div[@class='prop_statuses']/following-sibling::a/@href:<url=normalize-space(.)> ]
            [? .//div[@class='prop_maininfo']:<description=normalize-space(.)> ]
```

Parallel execution
instantiation, splitting, distribution, monitoring



Full-site web data extraction

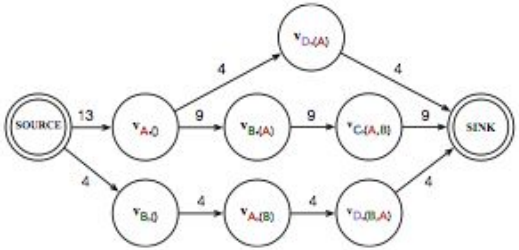
Target signature

title	releaseMonth	releaseDay	releaseYear	rating	genres	producer	runtime	overall score

Wrapper-generated instance

A	B	C	D	E	F	G
Ava's possessions	March 4, 2016	Rated: R	Off Hollywood Pictures	Genre(s): Sci-Fi, Mystery, Thriller, Horror	89 min	51
Camino	March 4, 2016	Rated: Not Rated	Bielberg Entertainment	Genre(s): Action, Adventure, Thriller	103 min	tbd

Data / Wrapper Repair
sequence labeling, sequence alignment, resegmentation



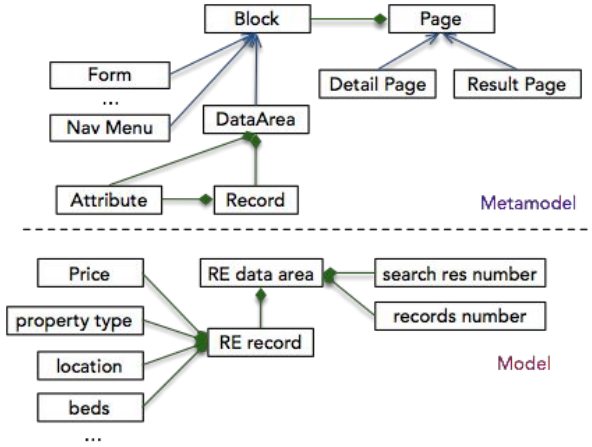
Stefano Ortona, Giorgio Orsi, Tim Furche, Marcello Buoncristiano: **Joint repairs for web wrappers.** *ICDE 2016: 1146-1157.*

Meltwater

Domain Knowledge

Domain knowledge

- Describe target objects via entities, relationships, instances
- Provide a way to identify them on web pages via shallow NLP (dictionaries, regexes)
- Use it to annotate both the visible and invisible parts of the live DOM



Dictionaries

cursymb:instance

£ -> { norm = GBP }
$ -> { norm = USD }
GBP -> { norm = GBP }
USD -> { norm = USD }

curname:instance

pounds -> {norm = GBP}
dollars -> {norm = USD}

price:label

price
amount

Rules

```
(cursymb:instance) number:instance[value>=80k && value<=200M]
|
number:instance[value>80k && value<=200M] (cursymb:instance | curname:instance) -> price:instance
```

Luying Chen, Stefano Ortona, Giorgio Orsi, Michael Benedikt. **ROSeAnn: Reconciling Opinions of Semantic Annotators.** *PVLDB 6(12): 1238-1241 (2013)*

Domain Knowledge

Labels and instances, visible and invisible (HTML structure, Javascript values)


```
<div class="icon first”>
  <img src=”…/beds.jpg” alt="Bedrooms" title="Bedrooms">
  <br>8
</div>
<div class=”icon">
  <img src=”…/bath.jpg” alt="Bathrooms" title="Bathrooms">
  <br>4
</div>
```

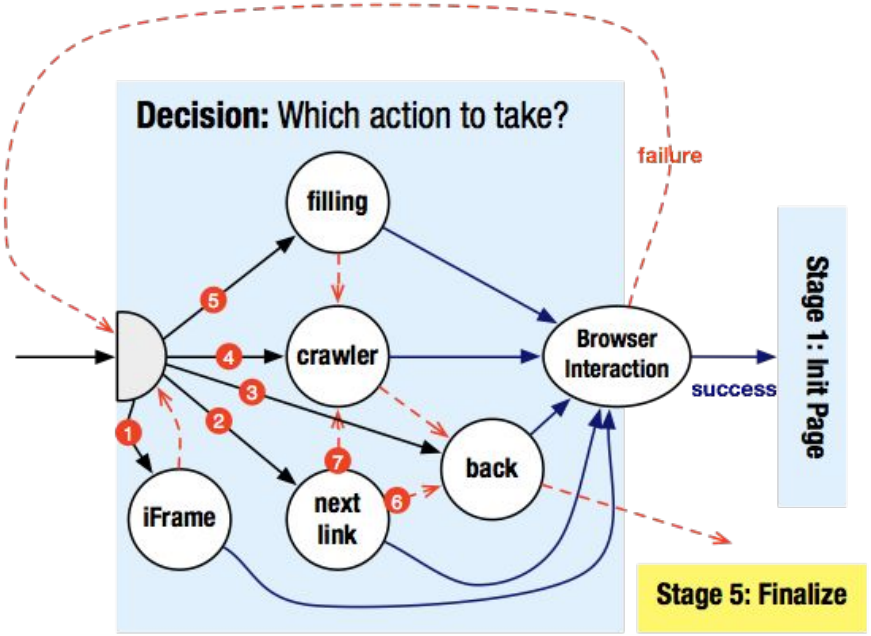
Website Exploration

Block Classification

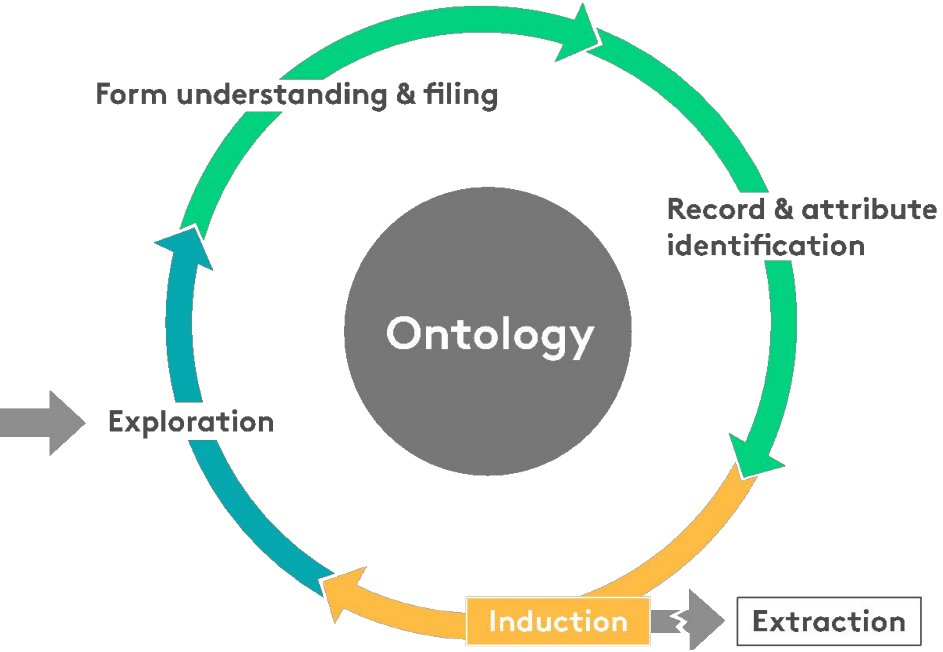
- ML (SVM and Decision trees)
- Features are knowledge-parametric

Exploration strategy

- Knowledge-driven focused crawling
- Relational Transducers to declaratively represent strategies (data driven)
- Everything gets translated into logical facts



A never-ending process



Exploration

- Focused crawling
- Stop conditions
- Relational transducers

Template Discovery

- Data areas detection
- Record segmentation
- Attribute alignment

Domain Modeling

- DOM annotation (dictionaries, regexes)
- Web phenomenology (forms, fields, labels, menus)
- Conceptual models

Form Understanding

- Labelling
- Classification
- Filling

Framework for rule-based feature engineering supporting quick turn around for domain-specific rich features on top of a library of 2.5k pre-built features representing structure, visual rendering, and textual content of a webpage, as well as the link structure and interaction patterns of the entire site.

Effects of full-site web data extraction

80-90% lower
human **effort**

without loss in quality compared with state-of-the-art

3-10x more
attributes

and domains than existing automated solutions and affordable supervised one

10-100x
more sources

e.g., 300k+ news sources, 1M+ of company websites, Job postings, Press releases

And if the data is not structured? Information Extraction

document categorizer	tokenizer	sentence splitter	near dup detection	POS tagger	noun chunking
stemmer	lemmatizer	dictionaries	NER	NED	relation extraction
sentiment analysis	keyphrase extraction	event extraction	clustering	coreference resolution	dependency parsing

Information extraction

Identify, disambiguate, and link mentions of entities of interest in a document.

Tokenizer + Splitter + PoS

Tesla	has	announced	the	full	acquisition	of	SolarCity	which	closed	on	Monday	morning	.
NNP	VBZ	VBN	DT	JJ	NN	IN	NNP	WDT	VBD	IN	NNP	NN	

NER

Tesla has announced the full acquisition of SolarCity which closed on Monday morning .

ORG ORG DATETIME

NED

Tesla has announced the full acquisition of SolarCity which closed on Monday morning .

Tesla Science Center at Wardenclyffe
Tesla (Czechoslovak company)
Tesla, Inc.

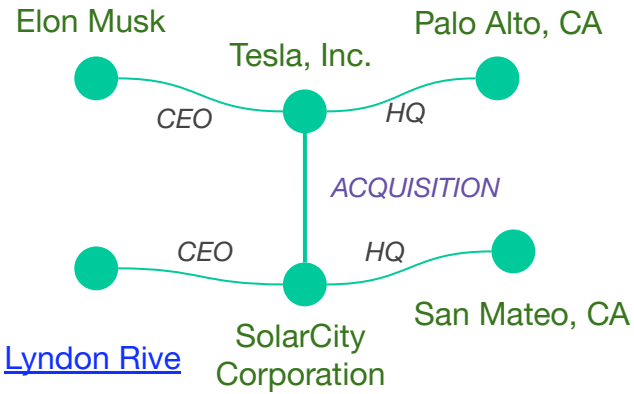
SolarCity Corporation
City Solar AG

Black Monday
Monday

Named Entity Disambiguation and Relation Extraction

- Searching data using keywords can lead to noisy results
- Disambiguated entities can be used to improve search, but they are also **necessary** to produce relations that can link them in database, e.g., a **graph** database. Relation Extraction (RE) is responsible for this task.

Tesla has announced the full acquisition of SolarCity which closed on Monday morning .

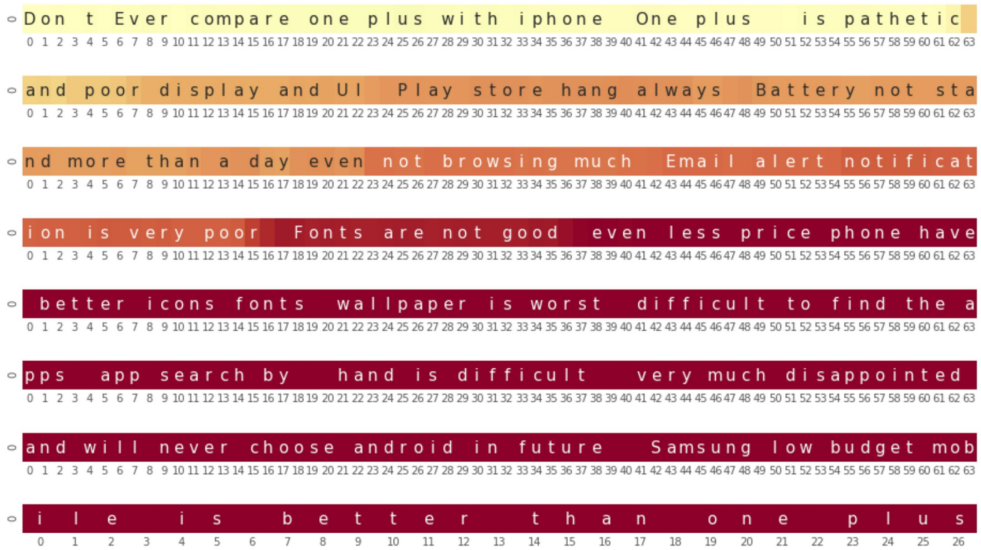


acquire(Tesla_Inc, SolarCity_Corp)

Sentiment Analysis

The task of assigning a polarity (i.e.,, positive, negative, neutral) to a text, entities, keyphrases, or extracted relations

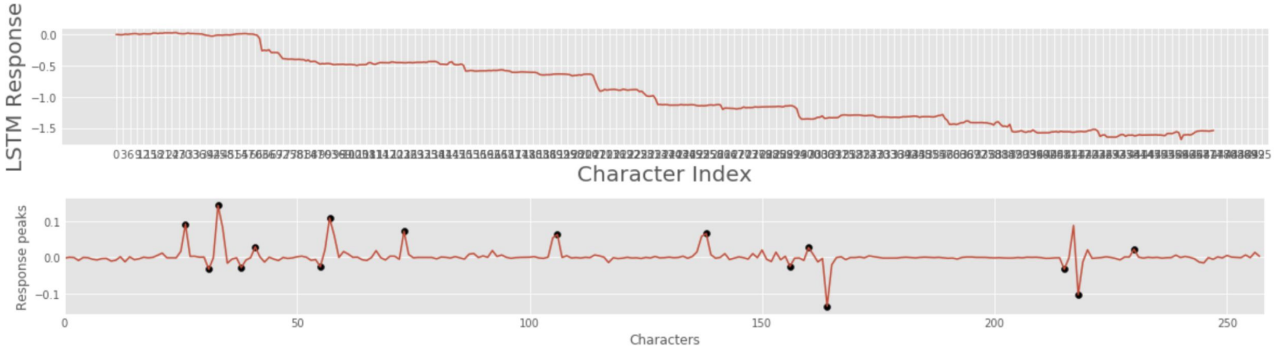
Example



Bing Liu. 2012. **Sentiment Analysis and Opinion Mining**. *Morgan & Claypool.*

Sentiment Analysis

The hot stuff right now is Long-Short Memory Networks (LSTM) with character-based encoding.



Aspect	Sentiment
Display	negative
Email alert notification	negative
Fonts	negative
wallpaper	negative

Markers indicate shifts in the sentiment signal that are used to locate interesting parts of the text carrying sentiment value.

Multiplicative LSTM for sequence modelling. Krause et al., 2017

Learning to Generate Reviews and Discovering Sentiment. *Radford et al., 2017*

Meltwater

Sentiment Analysis: Employee satisfaction

Reviews from Glassdoor, Capterra, Gartner, G2Crowd

Positive Themes

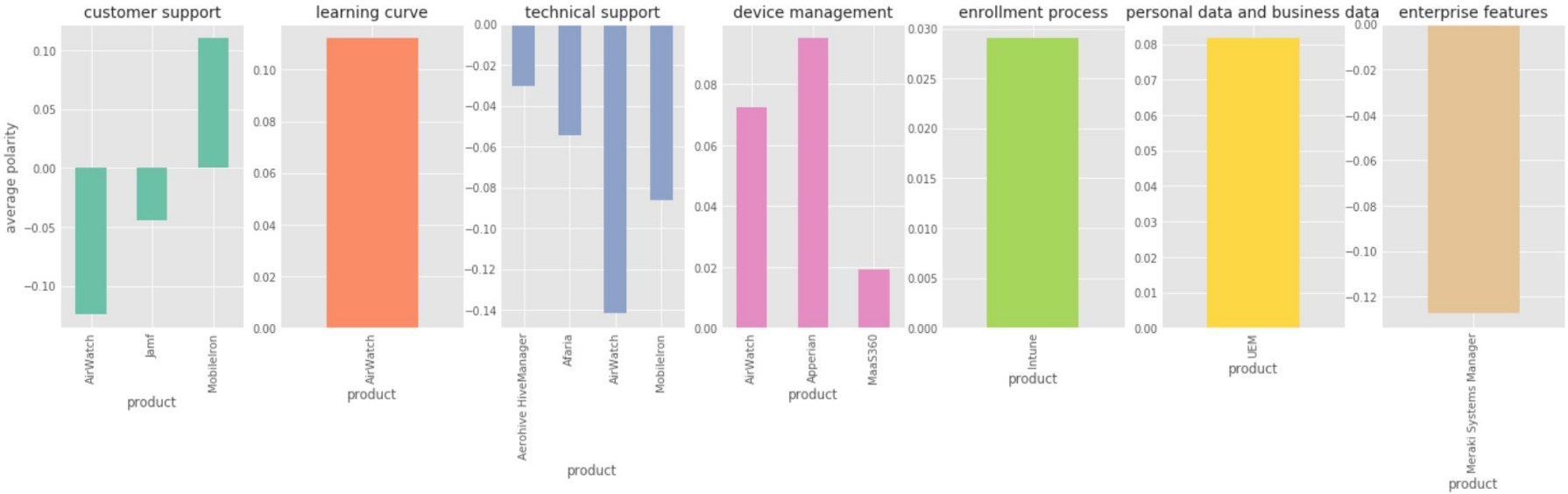
company	keyphrase	
apple	great benefits	60
	great place	27
	great experience	16
	good benefits	11
	great company	11
	great culture	10
	great people	10
	great products	10
	interesting people	10
	retail environment	10
ibm	great benefits	15
	life balance	12
	good benefits	9
	great people	9
	great place	9
	smart people	8
	work life balance	8
	good work environment	7
	good work life balance	7
	flexible hours	6

Negative Themes

company	keyphrase	
apple	work life balance	6
	career opportunities	3
	customer service	3
	life balance	3
	other companies	3
	retail environment	3
	additional perks	2
	bad attendance records	2
	boring real quick	2
	career advancement	2
ibm	work life balance	4
	life balance	3
	positive experience	3
	work-life balance	3
	additional limitations on staff functionality	2
	average benefits	2
	company drives mediocrity	2
	compensation and bonus structure	2
	complete joke	2
	constant fear of layoff	2

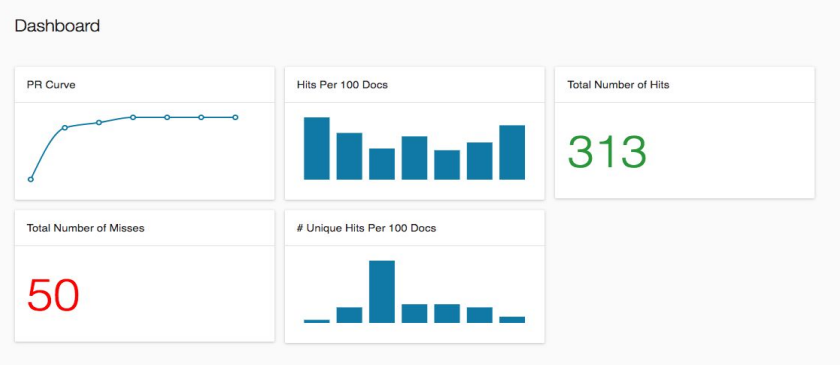
Sentiment Analysis: Product feature comparison

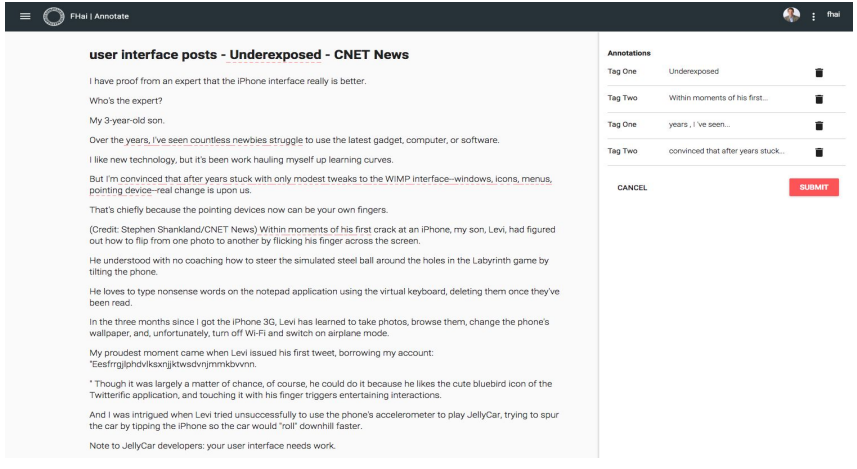
Keyphrases are features, sentiment becomes about those features, enabling comparison.



Meltwater

Human in the loop





- Annotate entities, disambiguations, sentiment, topics in different document formats: Text, HTML, PDF
- Task Assignment, Ranking, Inter Annotator Agreement (IAA)
- Gold set creation for any structured data like NER, NED, Knowledge Fusion, etc.

Very Time Consuming

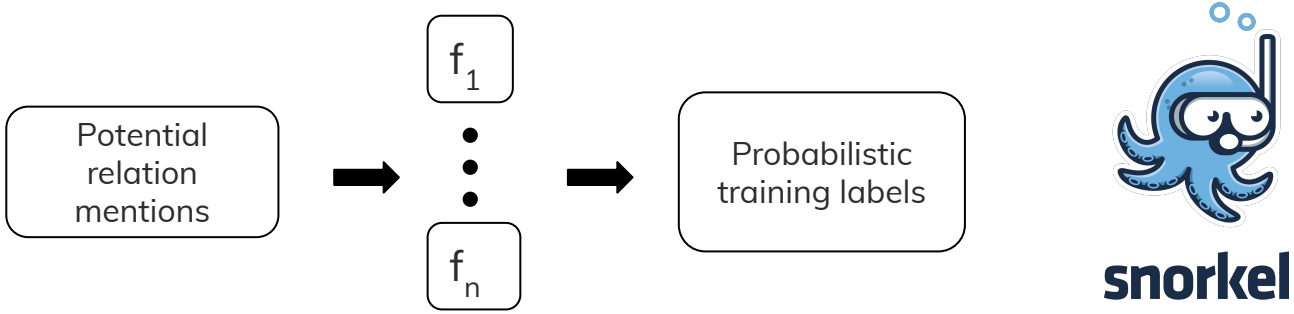
Meltwater

Data Programming

NLP, especially Deep Learning approaches requires large amounts of labelled data, which is an expensive and time-consuming effort.

Facebook *is* *competing* *with* *Google*

Snorkel can help us writing heuristics to programmatically generate training data. However, your training algorithm must be robust to noise



https://hazyresearch.github.io/snorkel/

We have lots of facts, now what?

Infer high-level insights from a set of extracted events/facts. You need to relate them in a Knowledge Graph

- Competitor
- Customer
- Investment
- Lawsuits
- Partnership

- Supplier
- Acquisition
- Out/under performance
- Expanding Operations
- Compliance

- Funding Developments
- Leadership Changes
- New Offerings
- Bankruptcy,
- Restructuring, Cost Cutting

Entities:

- Companies
- Brands
- Products
- Key people
- Influencers

Goal:

- Relate facts
- Data mining
- Cognitive applications
- Contextual Features

Challenges:

- Data Cleaning
- Data deduplication
- Data integration
- Truth Finding

But this is another story… Questions?

How does Meltwater tackle Outside Insight?

Ingestion:

- AI crawling for unstructured web
- Programmatic api's for partnerships
- Over 100M documents everyday

Knowledge Management

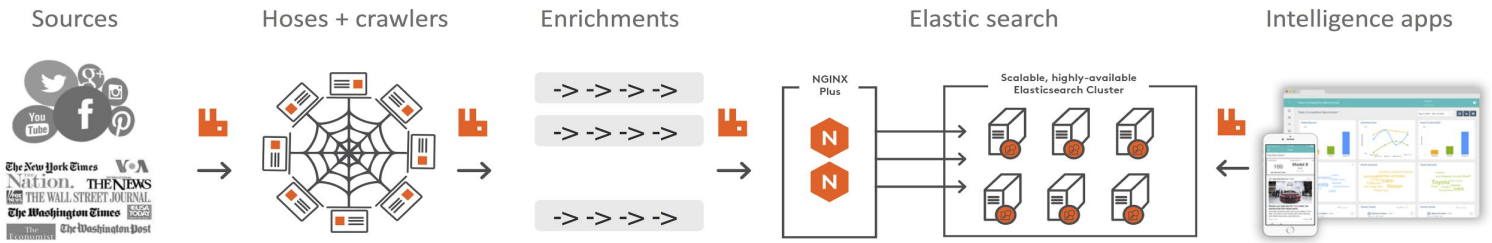
- NER (*person*, *location*, *organization*, ...)
- NED (https://en.wikipedia.org/wiki/Tim_Cook)
- Relation & event extraction
- Truth finding, link prediction, graph mining

Data Augmentation (15 languages):

- Text categorization (*topic*, *language*)
- Keyphrase extraction, summarization
- Sentiment analysis (entity, aspect level)
- Semantic hashing for near duplicate detection

Media Intelligence applications

- 1M complex Boolean queries configured
- Counters, aggregates, drill downs, pivoting, regression
- Vertical Search, news feed, media exposure, alerts based on trends & anomalies, influencers etc



Meltwater

Media Intelligence Apps

Global Monitoring

Analyze & Report

Distribute

Influence & Engage

New Apps

AI-Powered Reporting

Outside Insight

Employee App

Freemium

Enterprise Solutions

Custom solutions

3rd party Apps

PaaS

100M documents ingested daily

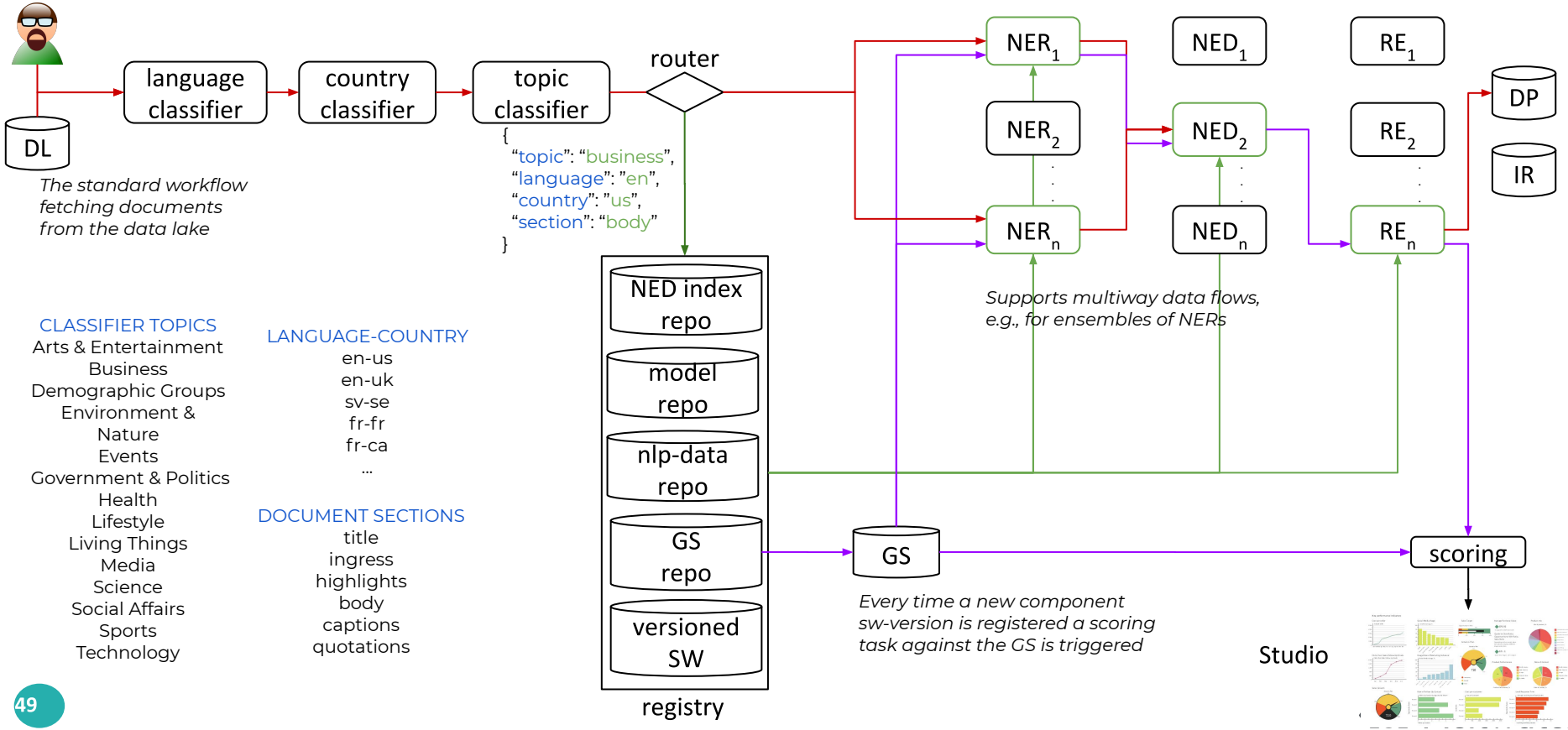
150 NLP/IR pipelines

100's Billions of Searches

Service Layer	**Streaming, Search, Analytics, APIs** Building blocks to leverage the platform
Context Building	**Knowledge Graph** Enable cognitive applications on top of our data by connecting the dots
Enrichment & Analysis	**Data Enrichment Platform** Enrich, analyze & build insights by interoperating with all major players
Outside Data	**AI-Driven Data Acquisition** Bring high quality outside data to our repository with minimal human effort

FAIRHAIR

NLP Pipelines



CLASSIFIER TOPICS
Arts & Entertainment
Business
Demographic Groups
Environment & Nature
Events
Government & Politics
Health
Lifestyle
Living Things
Media
Science
Social Affairs
Sports
Technology

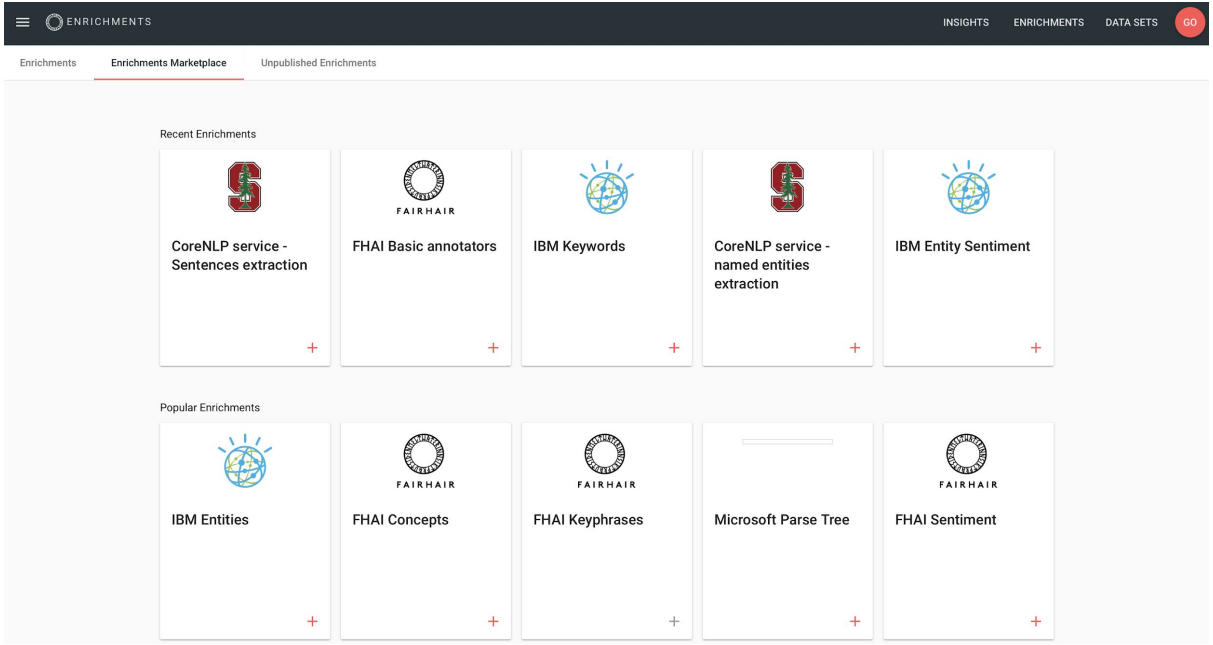
LANGUAGE-COUNTRY
en-us
en-uk
sv-se
fr-fr
fr-ca
...

DOCUMENT SECTIONS
title
ingress
highlights
body
captions
quotations

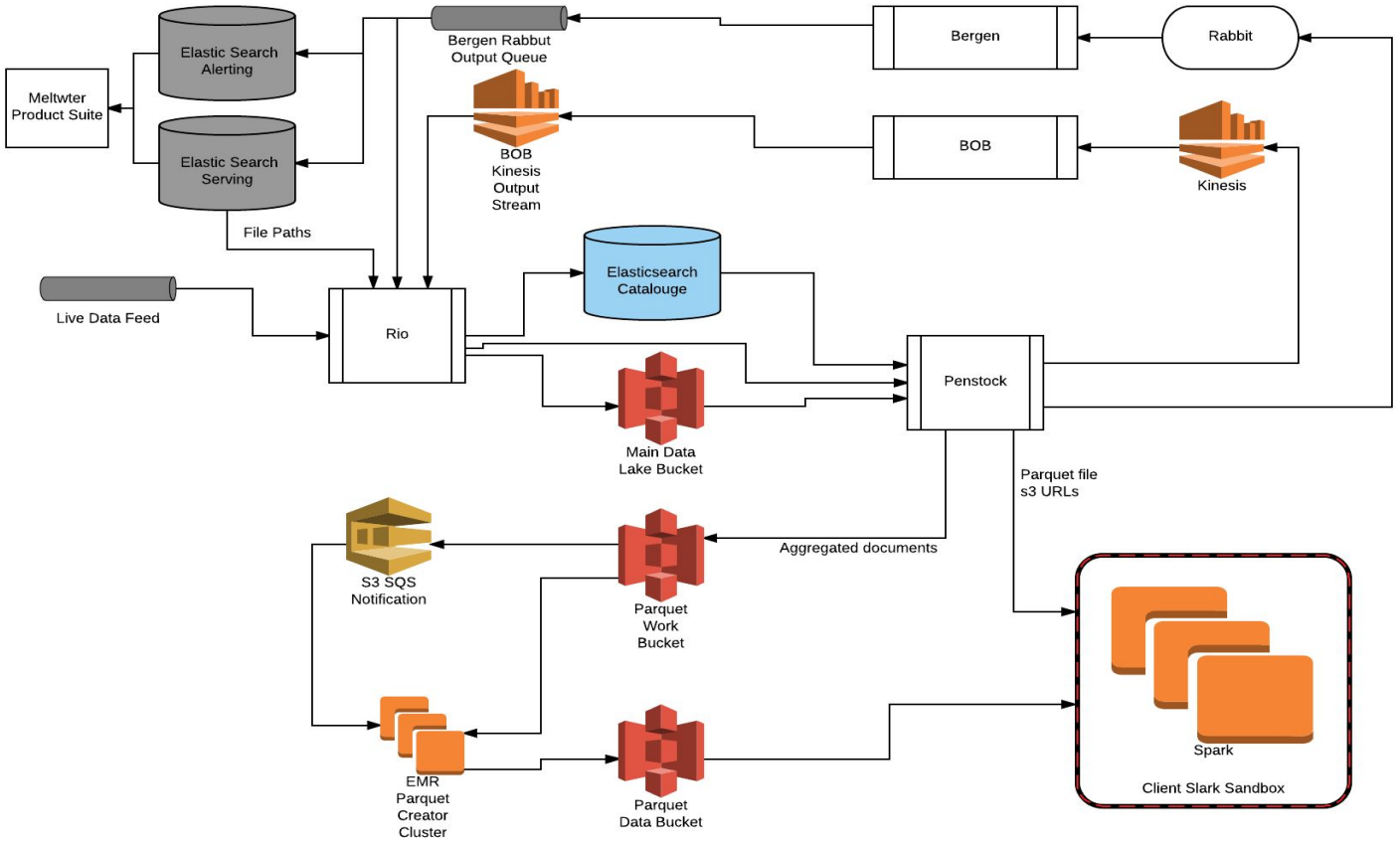
NLP Pipelines

We cannot foresee all uses of our data: Developer APIs to Integrate and orchestrate third party tools.

Personalization is key in Data Science: A flexible data wrangling infrastructure is required.



Data Platform



Meltwater

Data platform

Online News

Social

Patents & Trademarks

Company Websites

Forums

Blogs

Job Postings

●●●

Knowledge Graph

Data Science Platform

Serving Layer

elasticsearch

2.8k vCPU, 21TB RAM, 630TB SSD

Analytics Layer

APACHE Spark™

hadoop

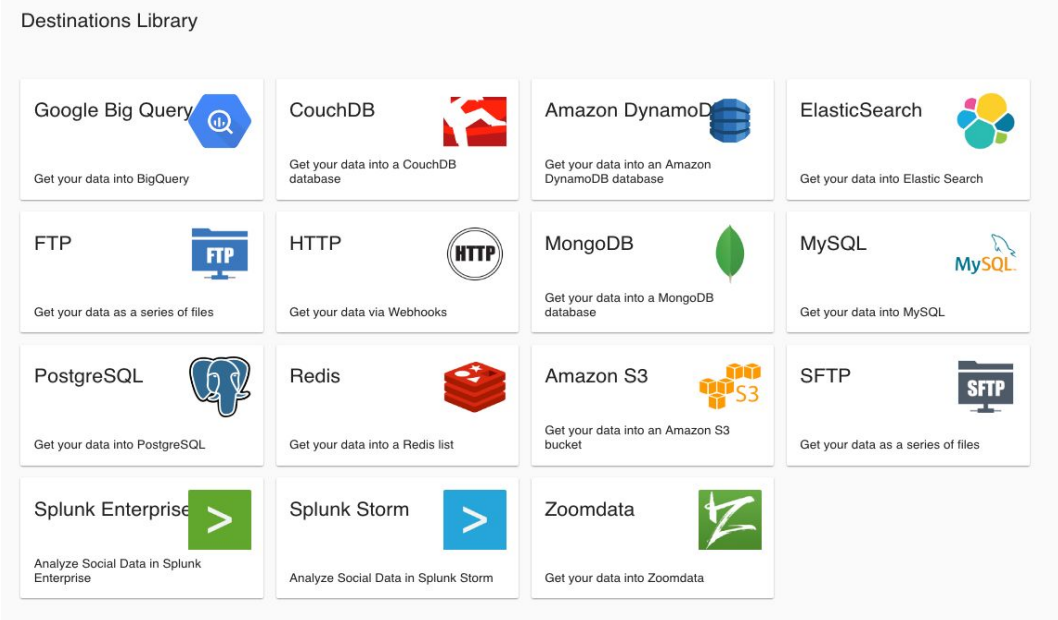
200B Documents

30M Sources

Meltwater

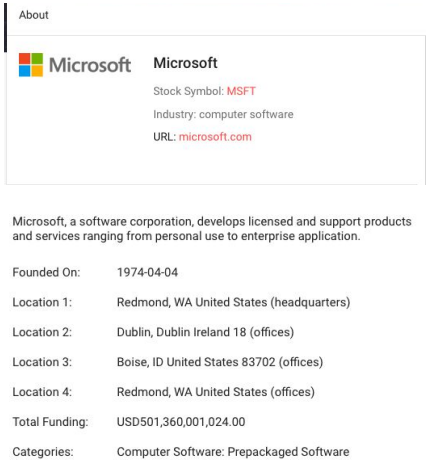
Connectors to serving systems

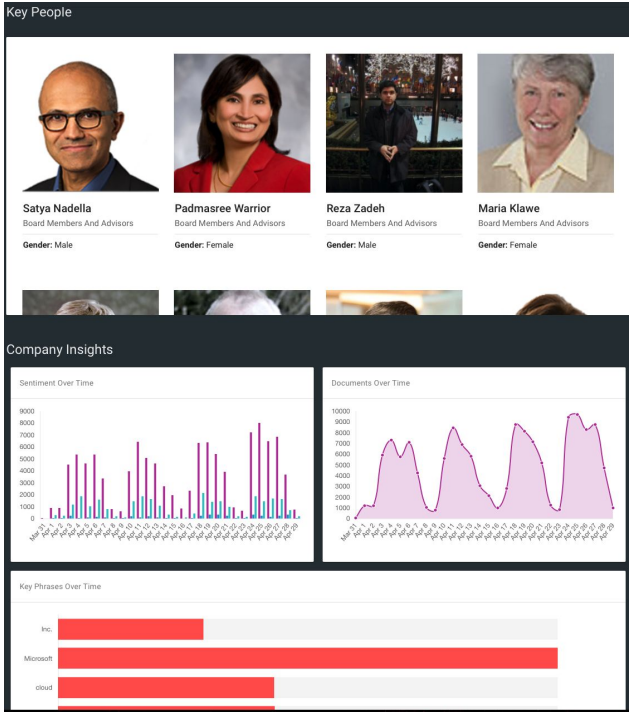
Data Ingestion & Insights Delivery
via schema mapping

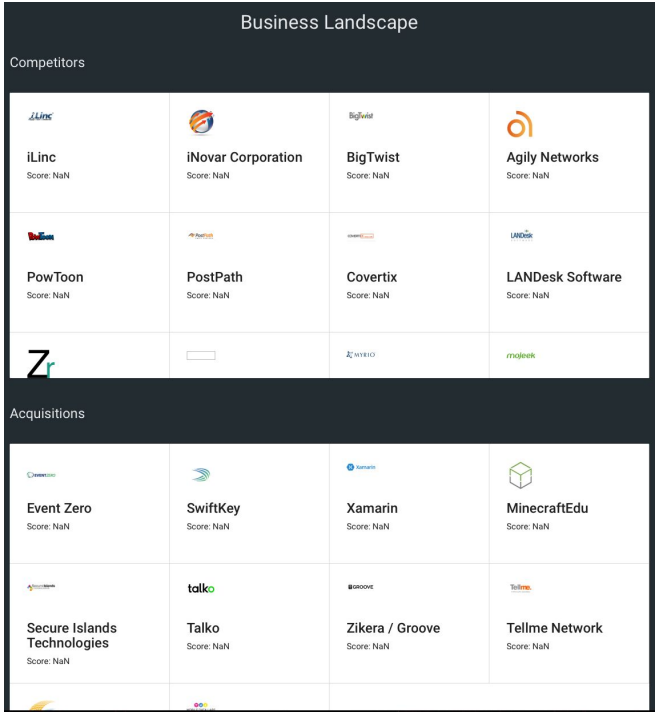


The most obvious use is to give “context” to information

Cognitive Applications complement traditional media intelligence tools

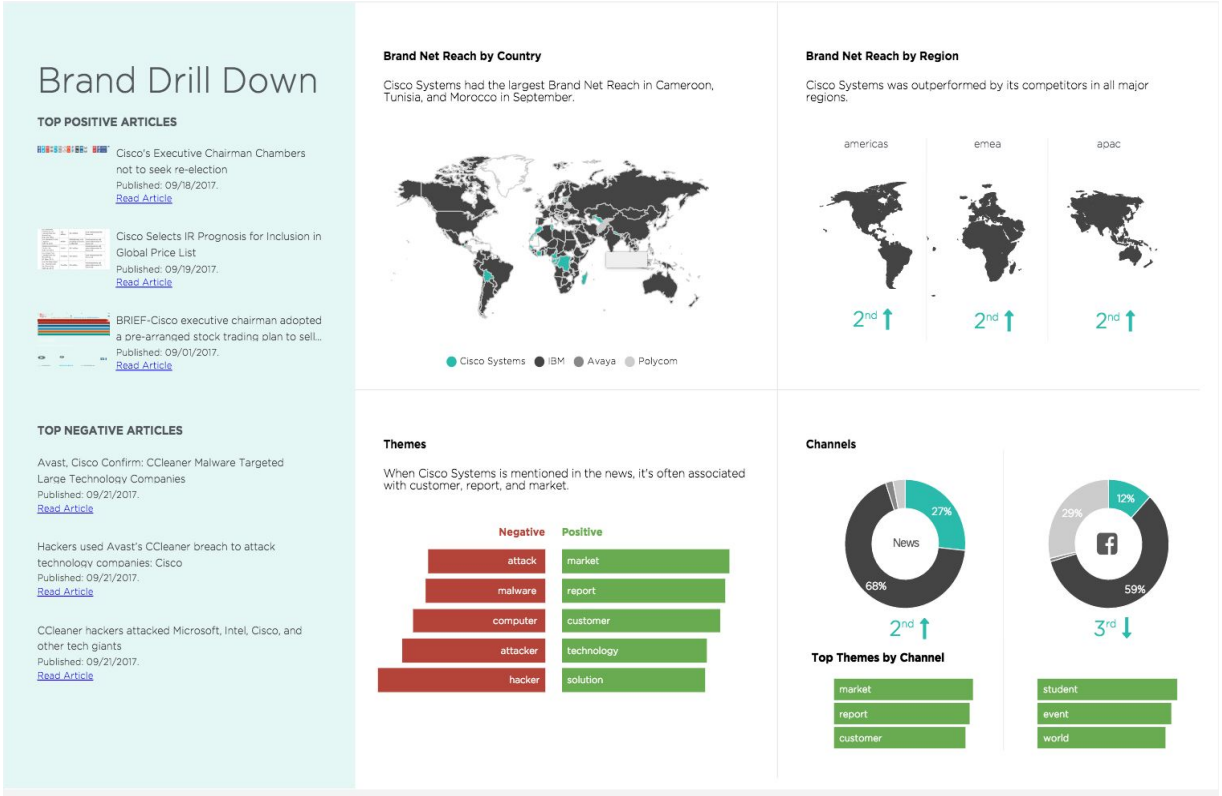






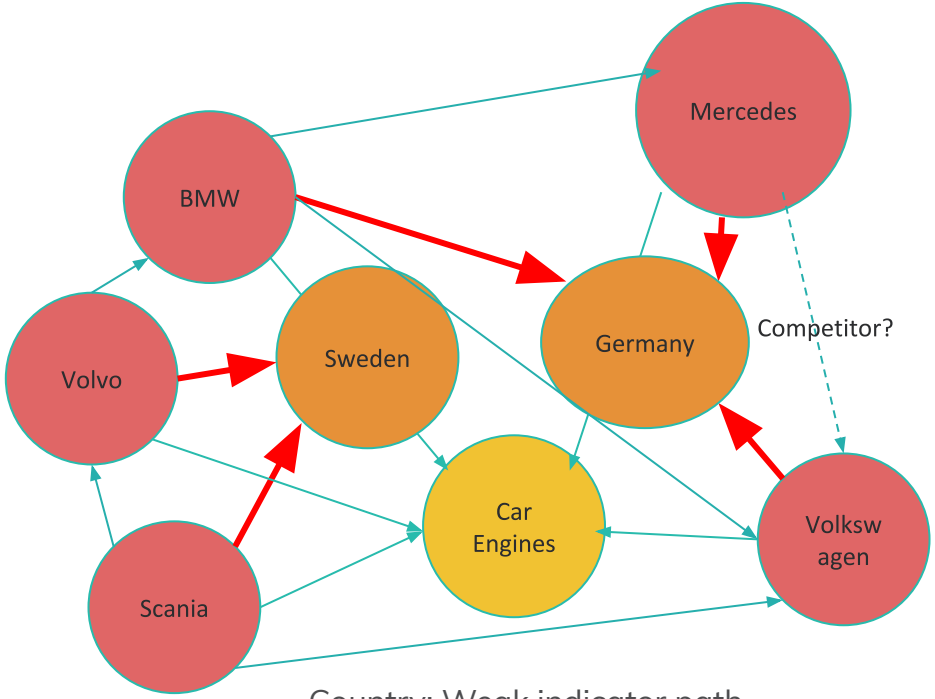
Meltwater

The most obvious use is to give “context” to information

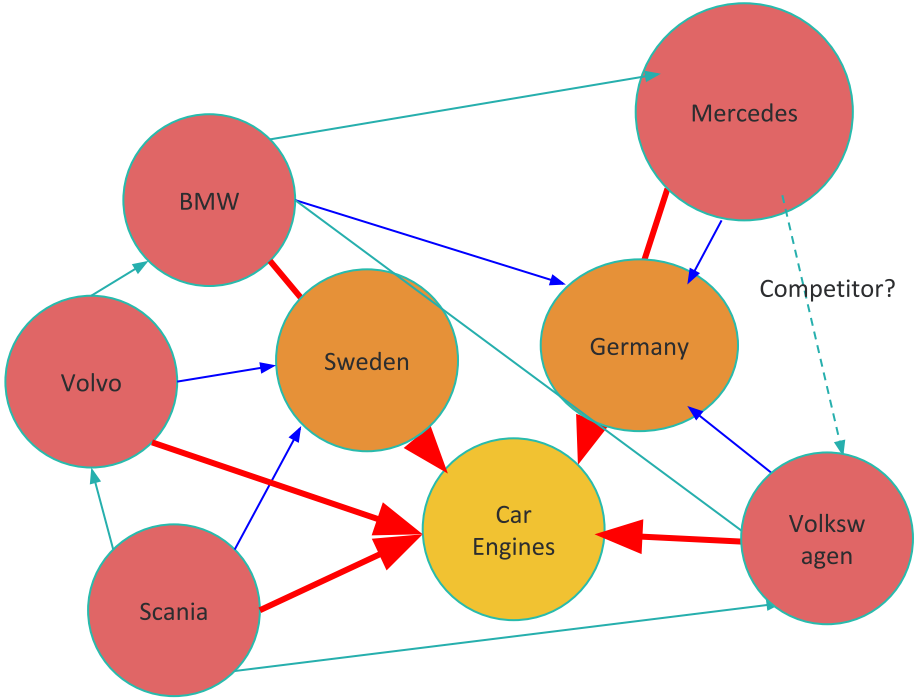


Meltwater

Path ranking



Country: Weak indicator path



Industry: Strong indicator path

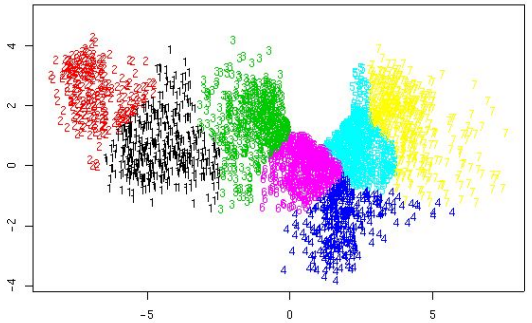
Graph embeddings

- Given a graph (g), entities (e) and relations (r), produce a low rank tensor factorization of the co-occurrence cube of all combinations of (e,r,e)

- Input:
  - Graph
  - Vector size

- Goal:
  - Find vectors for all (e) and (r) that minimizes the scoring function:

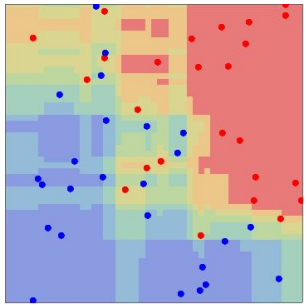
$$s_{LC}(h,\ell,t) = \delta_1^\ell \langle \mathbf{r}_1^\ell | \mathbf{e}_1^h \rangle + \delta_2^\ell \langle \mathbf{r}_2^\ell | \mathbf{e}_1^t \rangle + \delta_3^\ell \langle \mathbf{e}_1^h | \mathbf{D} | \mathbf{e}_1^t \rangle + \delta_4^\ell \langle \mathbf{e}_2^h | \mathbf{R}^\ell | \mathbf{e}_2^t \rangle$$



- Output:
  - Embedding vectors for (e) and (r).

Link prediction using embeddings

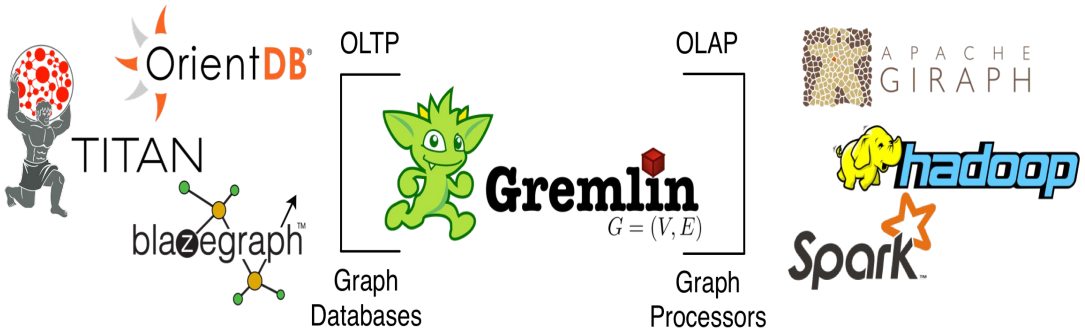
- Given a pair of entities `(e1,e2)`, give a score on how probable that they have a relation
- Input:
  - Embedding vectors of entities and relation `r`
  - Annotated examples of true and false combinations of `<e,r,e>`
- Goal:
  - Find a decision boundary that separates the true and false
    - E.g.: Using a standard classifier (SVM or RandomForest), use Embeddings as features
- Output:
  - A probability score for `<e1,e2,r>`



Meltwater's Fairhair.AI KG

Organization	Person
11,171,077	1,708,796

275M Facts Mined
(Distinct)



Relation	Instances
Competition	33,327,137
Works At	228,070
Investor (Company)	93,980
Founder	67,515
Board Member	43,525
Acquisition	15,532
Investor (Person)	10,420
Sub-organization	4,214