



# Taming the “Outside” in Outside Insight Applications

# 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



# Initiatives



## 6 Data Science Hubs (co-working spaces)

- |                 |            |
|-----------------|------------|
| ✓ London        | ✓ Sydney   |
| ✓ San Francisco | ✓ Berlin   |
| ✓ Singapore     | ✓ New York |



## 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: Media Intelligence

Sources: Editorial, Social, Broadcasts



media exposure



trends



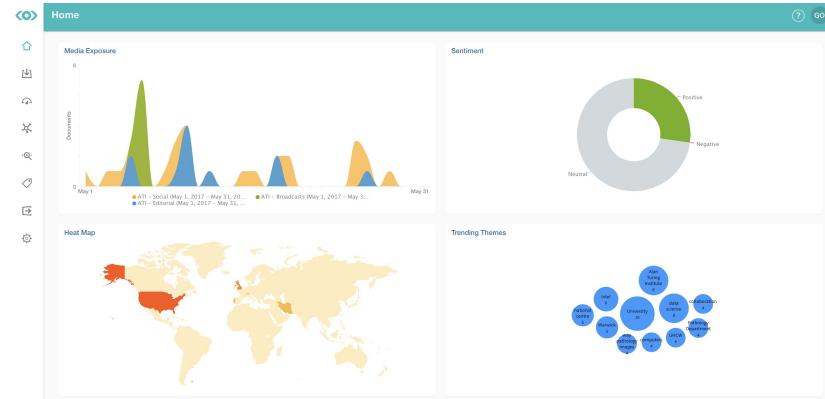
influencers



sentiment analysis

Ks millions of user queries, 300k distinct types

The screenshot shows the Meltwater Content Stream dashboard. It features a 'Content Stream' section with several news items from various sources like BBC News, ABC News, and BBC Radio 4. Below this is a 'Media Exposure' chart showing spikes in document counts over time, with specific data points for 'ATI - Social' and 'ATI - Editorial'. At the bottom is a 'Sentiment Score' chart showing fluctuations between positive, negative, and neutral sentiment levels.



The screenshot shows the Meltwater Influencer profile for 'Eli Blumenthal'. It displays his profile information, including his role as a 'Reporter' at 'USA Today, McLean, Virginia'. Below this is a 'Recent Results' section showing news items from 'Financial Review' and other sources related to ATI.

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



# A turning point

Build a world class **AI** platform for a new software category

## 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

ORACLE®



Palantir



TRIFACTA

tamr

Meltwater

# 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**
  - $x=75, y=1,850 + [x \rightarrow \text{house number}, y \rightarrow \text{birthyear}]$
  - $x=75, y=1,850 + [x \rightarrow \text{age}, y \rightarrow \text{height(mm)}]$

What is **Knowledge**?

- Available information for **rational action**
  - $x=75, y=1,850 + [x \rightarrow \text{house number}, y \rightarrow \text{birthyear}] \rightarrow \text{delete from electoral roll}$
  - $x=75, y=1,850 + [x \rightarrow \text{age}, y \rightarrow \text{height(mm)}] \rightarrow \text{provide free medicines}$

# Background: Cognitive Systems

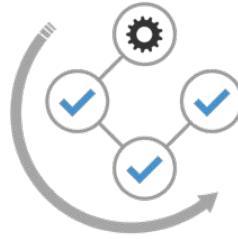
Adaptive and **interactive** systems that...



enhance human **cognition** through



**knowledge**



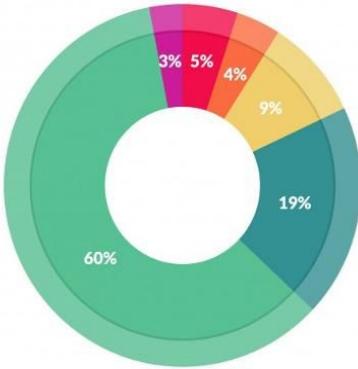
**reasoning**



**learning**

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

# Roadblocks



What data scientists spend the most time doing

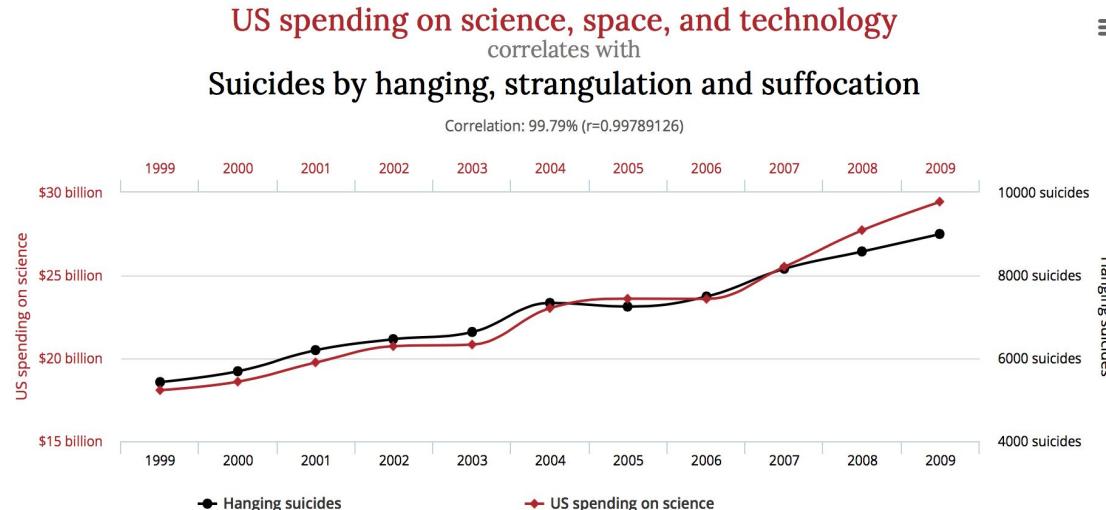
- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

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

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

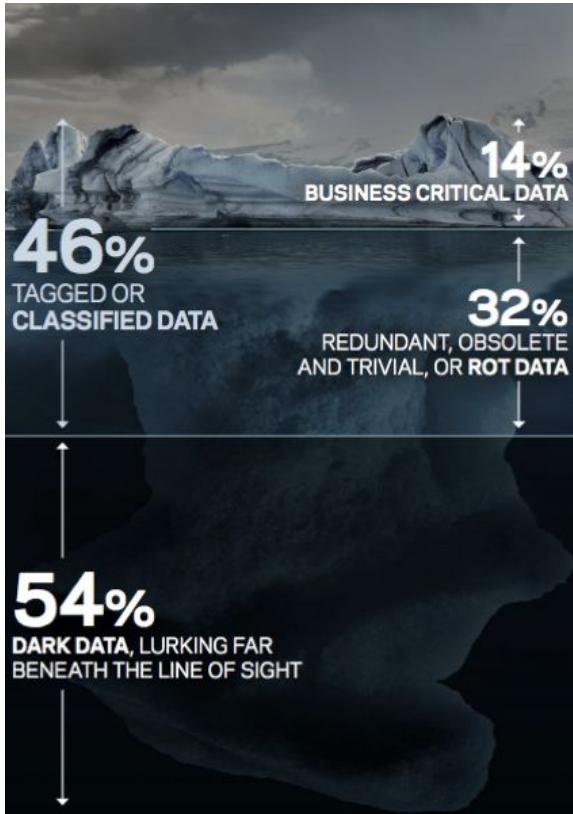
Access to **structured** data

- 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](http://www.tylervigen.com) (Spurious Correlations)

# 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

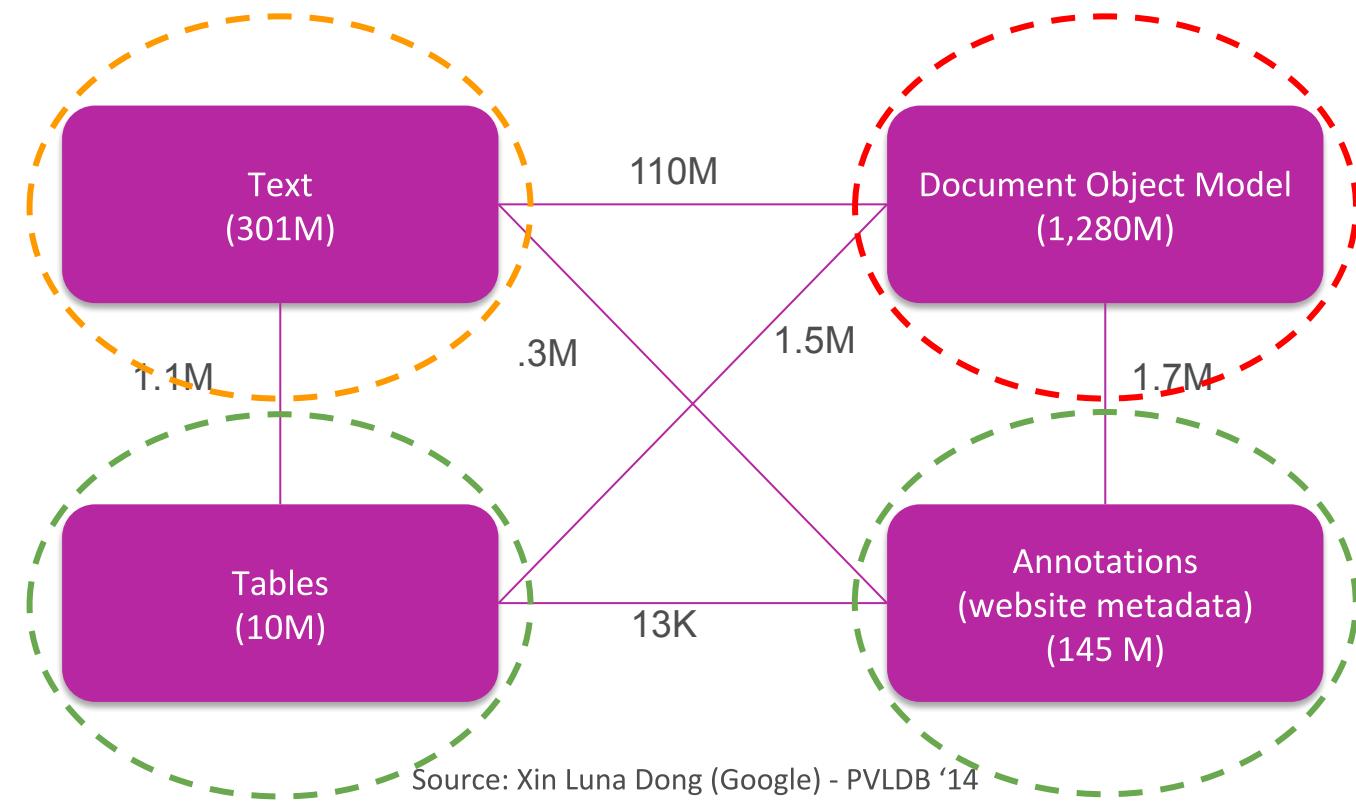


# What is outside data

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?



# 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**

The image shows four real estate listings from the Houlihan Lawrence Somers Brokerage website:

- 2482 Bound Brook Ln** (Yorktown Heights, NY 10598)  
\$372,000  
3 bd | 1 ba | 1,741 sqft / 0.5 acres  
Single-Family Home  
Houlihan Lawrence Somers Brokerage
- Preserve at Ardsley** (Scarsdale, NY 10583)  
From \$1,404,995  
4 Bd | 3.5 Ba | 3,377+ sqft  
New Community  
Preserve at Ardsley is a community of 11 single-family ... More
- 3468 Carol Ct** (Yorktown Heights, NY 10598)  
\$379,000  
3 bd | 2 ba | 1,602 sqft / 0.38 acres  
Single-Family Home  
Houlihan Lawrence Somers Brokerage
- 2906 Hickory St** (Yorktown Heights, NY 10598)  
\$329,000  
3 bd | 2 ba | 1,290 sqft / 0.26 acres  
Single-Family Home  
Houlihan Lawrence Yorktown Brokerage

Annotations point to specific elements:

- Records**: Points to the price boxes (\$372,000, \$1,404,995, \$379,000, \$329,000).
- Data Areas**: Points to the listing cards.
- Fields**: Points to the address fields (e.g., "3455 S. Overland Avenue", "Los Angeles CA 90034").
- Descriptions**: Points to descriptive text within the cards (e.g., "Terrific split level. Fabulous great room addition with a wall of built-ins, vaulted ceiling, beams, skylights, e ...").

The image shows the "CONTACT US" page of the n/aka restaurant website:

**n/aka**

ABOUT CHEF MENUS GALLERY PRESS RESERVATIONS FAQ CONTACT

**CONTACT US**

**ADDRESS**  
3455 S. Overland Avenue  
Los Angeles CA 90034  
310.836.6252

**GIFT CERTIFICATES**  
To purchase a gift certificate, please call us at 310.836.6252.

**INTERNATIONAL GUESTS**  
International guests are encouraged to use Resy or email us at [reservations@n-naka.com](mailto:reservations@n-naka.com) to request a reservation.

**PARKING INFORMATION**  
Valet parking is available in a small lot behind the restaurant. Limited street parking may also be available.

**MEDIA INQUIRIES**  
For media inquiries please contact [media@n-naka.com](mailto:media@n-naka.com).

3455 S. OVERLAND AVENUE LOS ANGELES CA 90034  
310.836.6252 | [INFO@N-NAKA.COM](mailto:INFO@N-NAKA.COM)

Map showing the location of n/aka at 3455 S. Overland Avenue, Los Angeles, CA 90034. The map includes labels for Overland Ave, 10, 405, Palms Blvd, Castle Heights, Culver City, Mar Vista, and Venice Blvd.

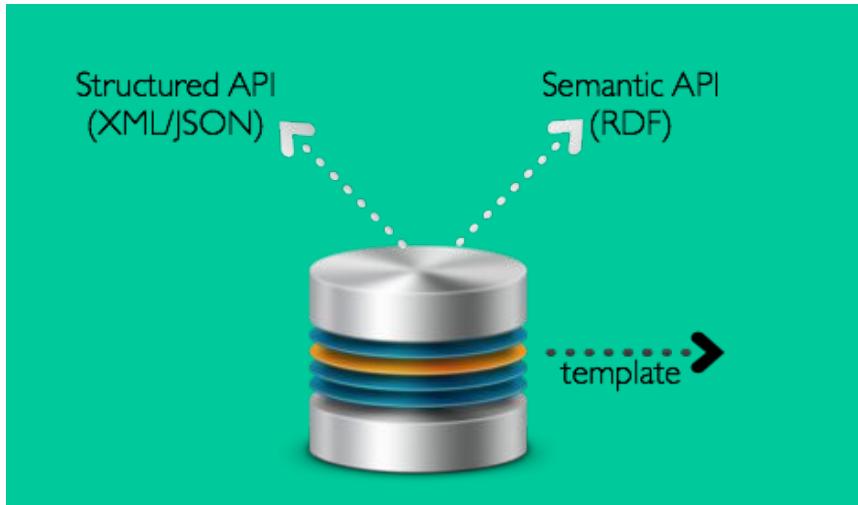
# Web Data Extraction vs Information Extraction

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

- 1 Image-directed and color Doppler studies of gallbladder tumors.
- 2 Thirteen cases of primary adenocarcinoma of the gallbladder (GB), 1 of malignant fibrous histiocytoma, 3 of metastatic adenocarcinoma, 5 of adenoma, 5 of polypus, 2 of xanthogranuloma, 6 of chronic cholecystitis, 4 of acute cholecystitis, and 8 of subacute cholecystitis were studied by image-directed and color Doppler ultrasonography (CDUS).
- 3 All of the 14 cases of primary GB cancer (10 masses, 4 thickening wall) were found to have a high velocity arterial blood flow signal in the wall of the GB.
- 4 In contrast, the 3 cases of metastatic cancer of the GB had no blood flow signal in the wall of the GB.
- 5 For the 30 cases of benign lesions of the GB, only in 12 cases was a low velocity blood flow signal found.
- 6 Nine of 10 cases of primary GB malignancy were found to have high velocity arterial blood flow signals in the tumor masses.
- 7 No blood flow signal was observed in the masses of 13 cases (3 of metastatic adenocarcinoma, 5 of adenoma, 5 of polypus).
- 8 An abnormal high velocity arterial blood flow signal observed within masses in the GB or in the GB wall is a significant feature of primary GB cancer and thus helps to differentiate primary GB cancer from metastatic and benign lesions of the GB.

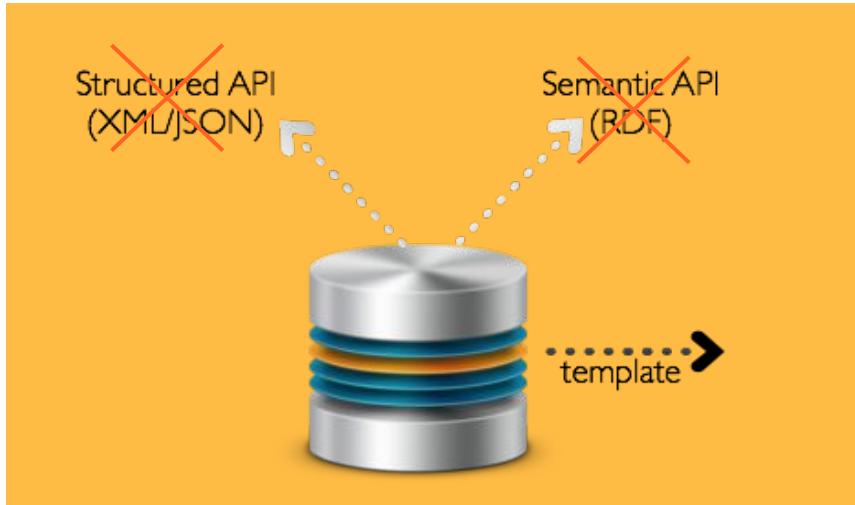
# The academic web

- 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** ...

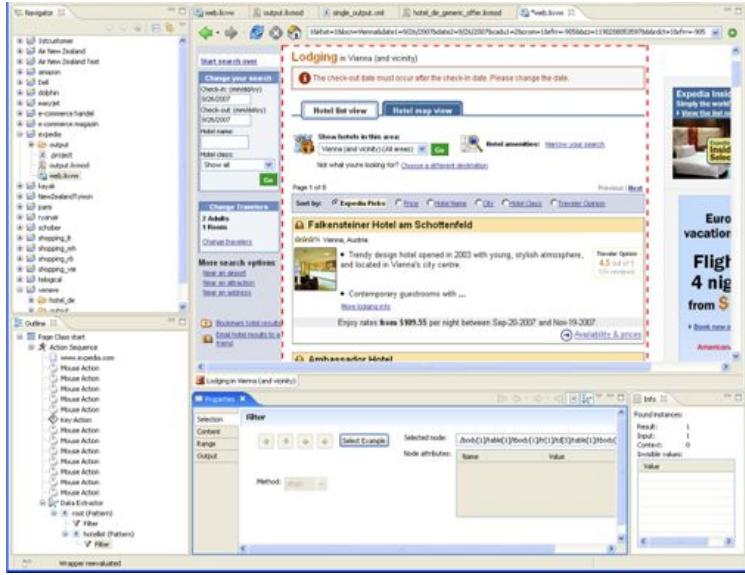
Example 1: A large social network started to harvest all business locations and add them to their social graph...

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

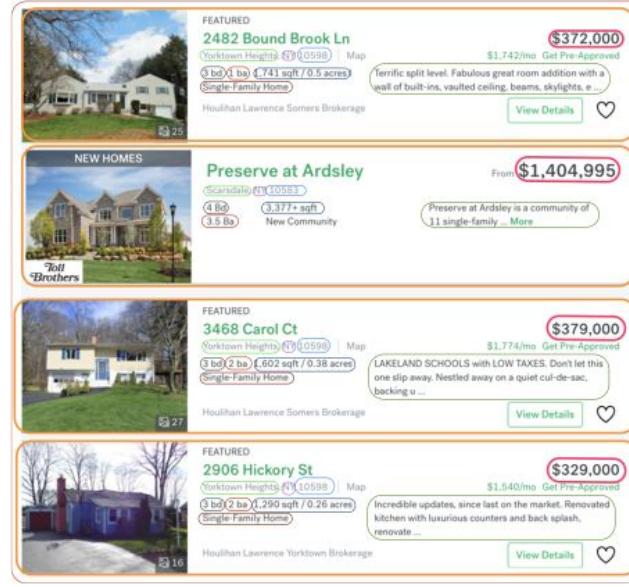
Example 2: A large travel website trying to harvest offers from airlines, hotels, car rentals, and restaurants locations and add them to their travel graph...

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



# 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, Jeremy 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:

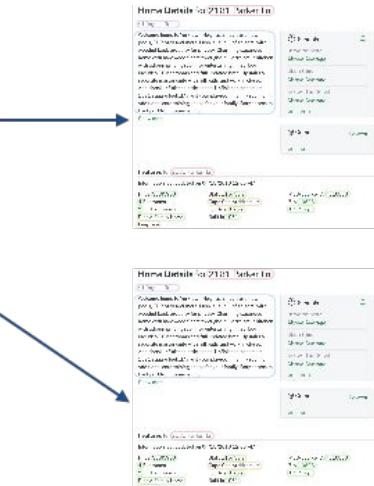
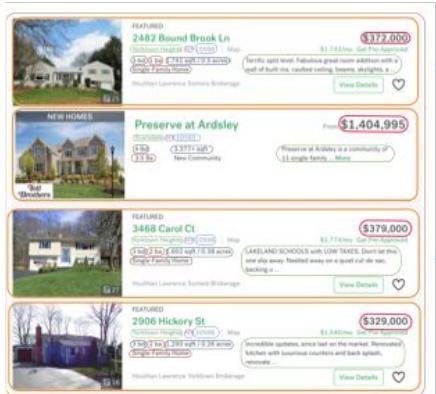
- Replace site supervision with **domain knowledge**
- Automate the exploration process, e.g., **form filling, focused crawling**
- Make wrapper induction algorithms **knowledge-parametric** (both ML and Rules)
- Make wrapper induction algorithms **robust to noise**

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

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

## Navigation



The screenshot shows the Trulia homepage and a search results page for 'Yorktown Heights, NY'. The search results show 1 - 60 of 294 results, with a navigation bar from 1 to 5. The results list includes:

- United States > New York > Yorktown Heights
- View Details** button
- More** button

On the right, there are sidebar categories:

- Toys & Games >** Sci-Fi & Fantasy Figures & Playsets, Figure & Playset Accessories, Swords, Sabres, Dressing Up Accessories, + See more
- Sports & Outdoors >** Fun Sports
- Amazon Video >** TV

## Template discovery

Result pages, detail pages

# Full-site web data extraction

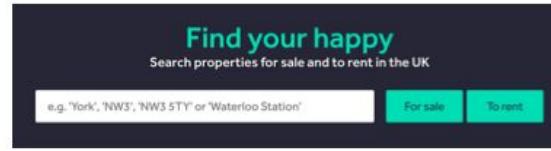
## Forms

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)



### Car ownership

When did you buy the car?

MONTH  YEAR 

I haven't bought the car yet

Whose name is on the registration document? 

Policy holder

More...

Who owns the car? 

Policy holder

More...

RECENT SEARCHES:  
Software Engineer

FILTER RESULTS BY:  
REFINEMENTS:  
London 

RADIUS:  
Distance: 20 miles

SORT BY:  
Relevance Date

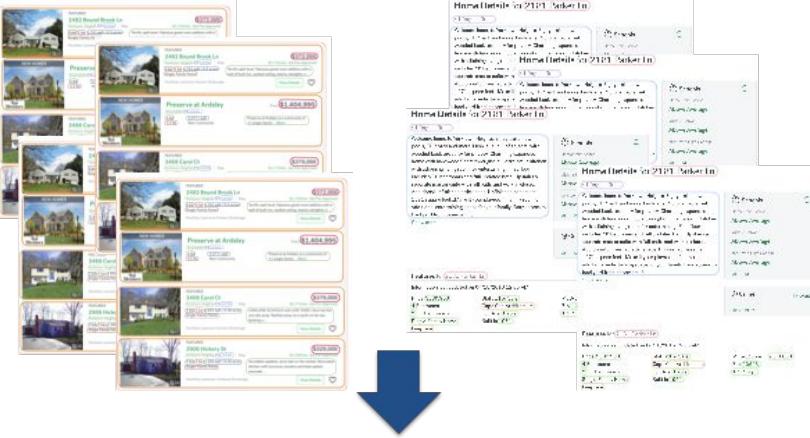
COMPANY SEARCH:  
Enter company name

JOB STATUS:  
 Full Time  
 Part Time

SKILLS:  
Accounting  
Analysis Skills  
Baan  
CSS (Cascading Style Sheet)  
Customer Support/Service

Show More 

# 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,'prolist_wrap')]/following-sibling::div//a[@class='pagenum'][last()]/&nextclick /)*
  //div[contains(@class,'prolist_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(.)>
```

# 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

# Extraction Language: JSON-XPath

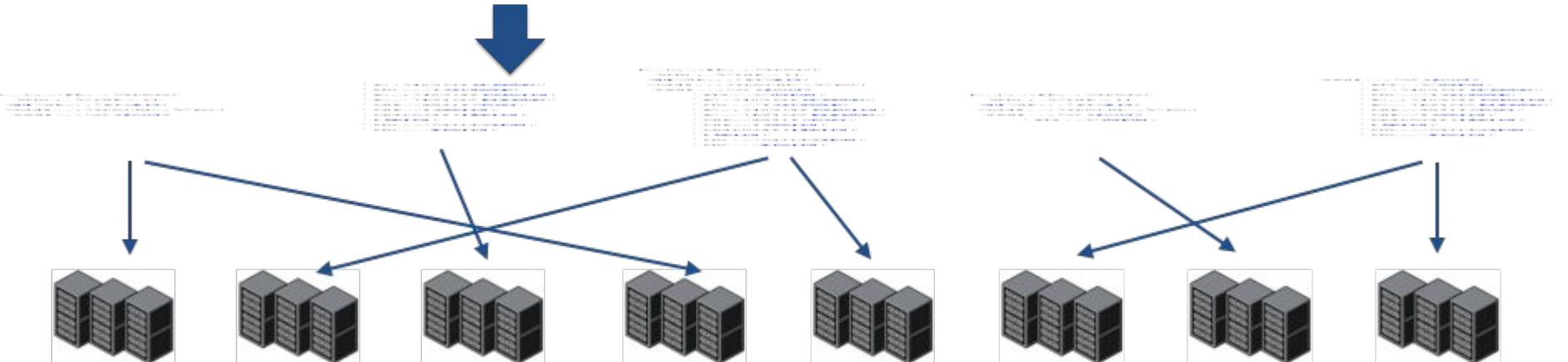
```
"articleTpls": [
  {
    "id": "article_template_0",
    "startUrls": [
      "http://www-03.ibm.com/press/us/en/pressrelease/33304.wss",
      "http://www-03.ibm.com/press/us/en/pressrelease/33420.wss",
      "http://www-03.ibm.com/press/us/en/pressrelease/33117.wss",
      "http://www-03.ibm.com/press/us/en/pressrelease/33303.wss"
    ],
    "urlPatterns": [
      "(?<wordset>([a-zA-Z]{1,}[:]{1,}){1,1})//(?<wordnumberset>([\\w]{1,})[\\.-]{1,}{1,3}[\\w]{1,})/(?<wordset1>([a-zA-Z]{1,})[/]{1,1}){1,3}[a-zA-Z]{1,})/(?<wordnumberset1>([\\w]{1,})[\\.-]{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(.)"
      }
    ]
  }
]
```

A **JSON-like** wrapper specification

Advantage? You can **query the program** as well!

# 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,'prolist_wrap')]/following-sibling::div//a[@class='pagenum'][last()]/nextclick /)*
/div[contains(@class,'prolist_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



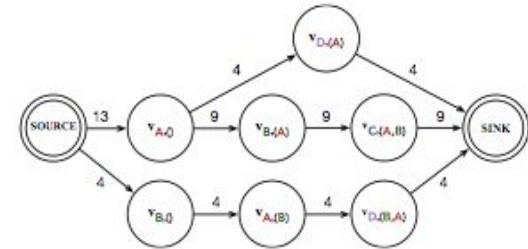
Target signature

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.

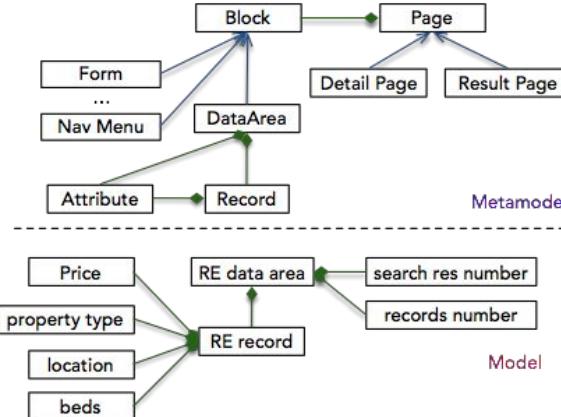
# Domain Knowledge

## 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

## Rules

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



## Dictionaries

cursymb:instance	curname:instance
£ -> { norm = GBP }	pounds -> {norm = GBP}
\$ -> { norm = USD }	dollars -> {norm = USD}
GBP -> { norm = GBP }	
USD -> { norm = USD }	price:label
	price
	amount

# Domain Knowledge

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

labels

Running costs

instances

The screenshot shows a table of running costs with labels and instances. The labels are 'Urban mpg', 'Extra Urban mpg', 'Average mpg', 'CO<sub>2</sub> emissions', and 'Annual Tax'. The instances are '38.7 mpg', '58.9 mpg', '49.6 mpg', '136 g/km', and '£130' respectively.

label	instance
Urban mpg	38.7 mpg
Extra Urban mpg	58.9 mpg
Average mpg	49.6 mpg
CO <sub>2</sub> emissions	136 g/km
Annual Tax	£130

Crick Road, Norham Manor  
Guide Price £6,000,000  
A substantial detached family home built over four floors located in one of the most desirable addre... [Read more](#)

8 4 5 5,683 ft<sup>2</sup>

[Save property](#) [View details](#)

Javascript values

labels

The screenshot shows the Trulia homepage with a search bar labeled 'Search by Neighborhood, City, or Address'. A 'Buy' button is to the left, and a 'Search' button is to the right. Arrows point from the labels 'Javascript values' and 'labels' to the search bar and the 'Buy' button respectively.

```
<div class="icon first">
  
  <br>8
</div>
<div class="icon">
  
  <br>4
</div>
```

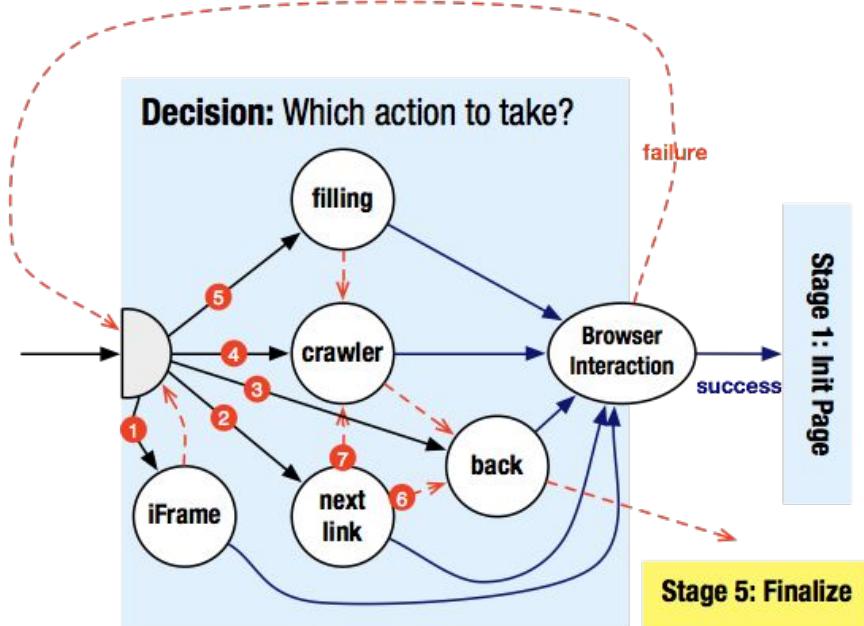
# Website Exploration

## Block Classification

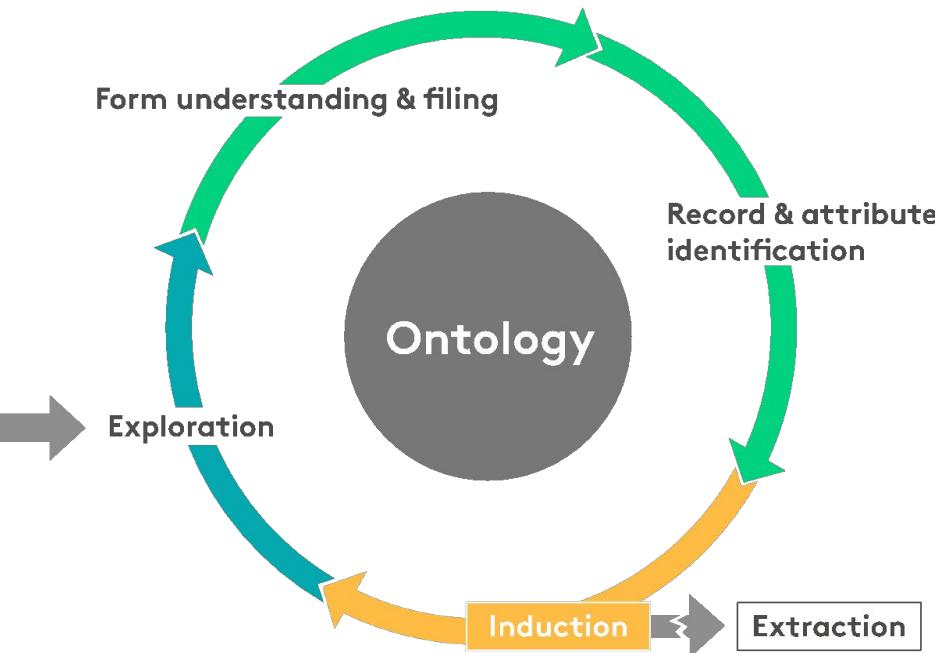
- ML (SVM and Decision trees)
- Features are knowledge-parametric

## 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

SolarCity Corporation

Black Monday

Tesla (Czechoslovak company)

City Solar AG

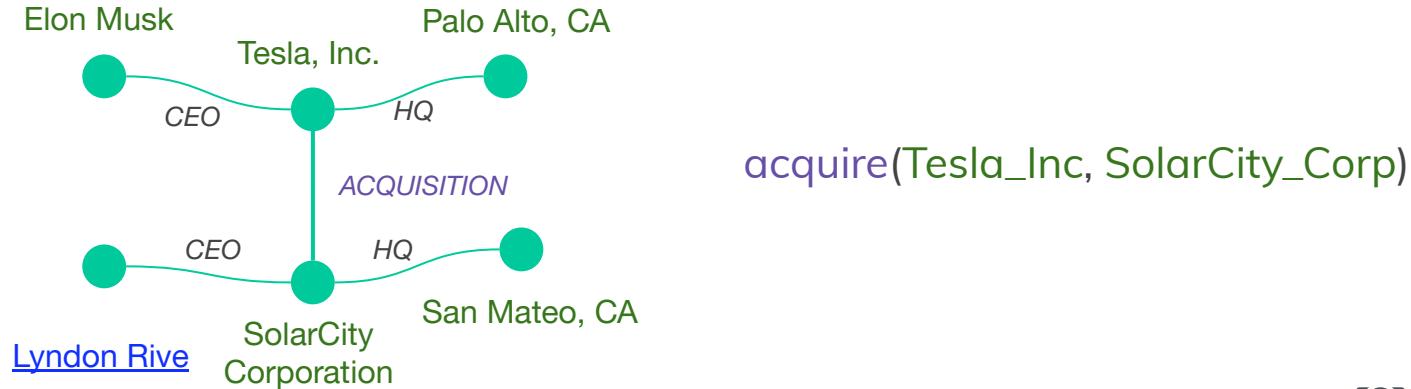
Monday

Tesla, Inc.

# 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 .



# Sentiment Analysis

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

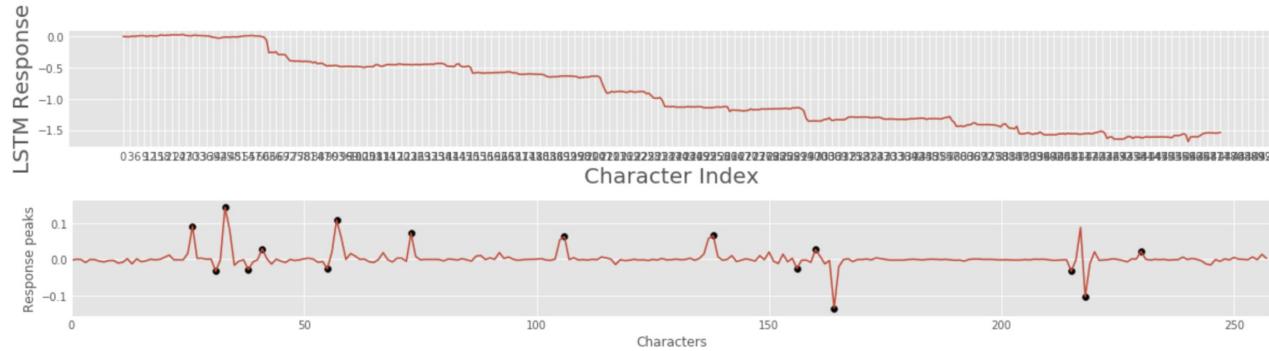
## Example

- Don t Ever compare one plus with iphone One plus is pathetic  
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
- and poor display and UI Play store hang always Battery not sta  
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
- nd more than a day even not browsing much Email alert notificat  
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
- ion is very poor Fonts are not good even less price phone have  
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
- better icons fonts wallpaper is worst difficult to find the a  
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
- pps app search by hand is difficult very much disappointed  
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
- and will never choose android in future Samsung low budget mob  
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
- i lie is better than one plus  
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26

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.

# 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

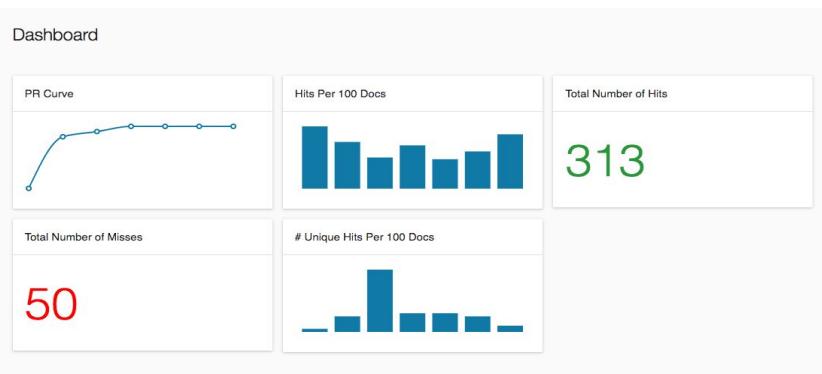
company	keyphrase	
apple	work life balance	6
	career opportunities	3
	customer service	3
	life balance	3
	other companies	3
	retail environment	3
	additional perks	2
	bad attendance records	2
	boring real quick	2
	career advancement	2
ibm	work life balance	4
	life balance	3
	positive experience	3
	work-life 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.



# Human in the loop



Fhai | Annotate

user interface posts - Underexposed - CNET News

I have proof from an expert that the iPhone interface really is better.

Who's the expert?

My 3-year-old son.

Over the years, I've seen countless newbies struggle to use the latest gadget, computer, or software.

I like new technology, but it's been work hauling myself up learning curves.

But I'm convinced that after years stuck with only modest tweaks to the WIMP interface—windows, icons, menus, pointing device—real change is upon us.

That's chiefly because the pointing devices now can be your own fingers.

(Credit: Stephen Shankland/CNET News) Within moments of his first crack at an iPhone, my son, Levi, had figured out how to flip from one photo to another by flicking his finger across the screen.

He understood with no coaching how to steer the simulated steel ball around the holes in the Labyrinth game by tilting the phone.

He loves to type nonsense words on the notepad application using the virtual keyboard, deleting them once they've been read.

In the three months since I got the iPhone 3G, Levi has learned to take photos, browse them, change the phone's wallpaper, and, unfortunately, turn off Wi-Fi and switch on airplane mode.

My proudest moment came when Levi issued his first tweet, borrowing my account: "Eesfrgjljhdvksajkjtwkvdwjnmjkmbwn."

\* Though it was largely a matter of chance, of course, he could do it because he likes the cute bluebird icon of the Twitterific application, and touching it with his finger triggers entertaining interactions.

And I was intrigued when Levi tried unsuccessfully to use the phone's accelerometer to play JellyCar, trying to spur the car by tipping the iPhone so the car would "roll" downhill faster.

Note to JellyCar developers: your user interface needs work.

Annotations

Tag One	Underexposed	<input checked="" type="checkbox"/>
Tag Two	Within moments of his first...	<input checked="" type="checkbox"/>
Tag One	years, I've seen...	<input checked="" type="checkbox"/>
Tag Two	convinced that after years stuck...	<input checked="" type="checkbox"/>

CANCEL      SUBMIT

- 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.

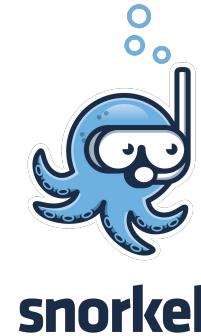
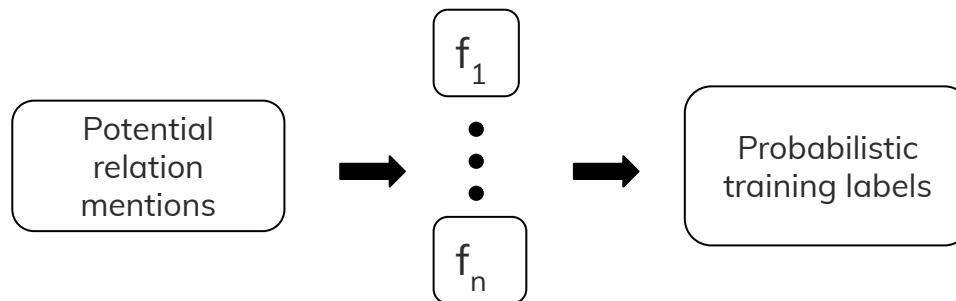
Very Time Consuming

# 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:

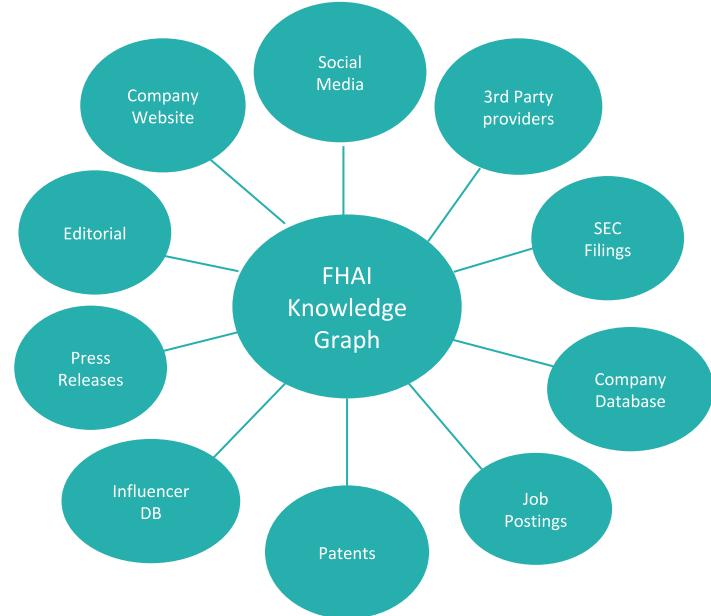
- Companies
- Brands
- Products
- Key people
- Influencers

## 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

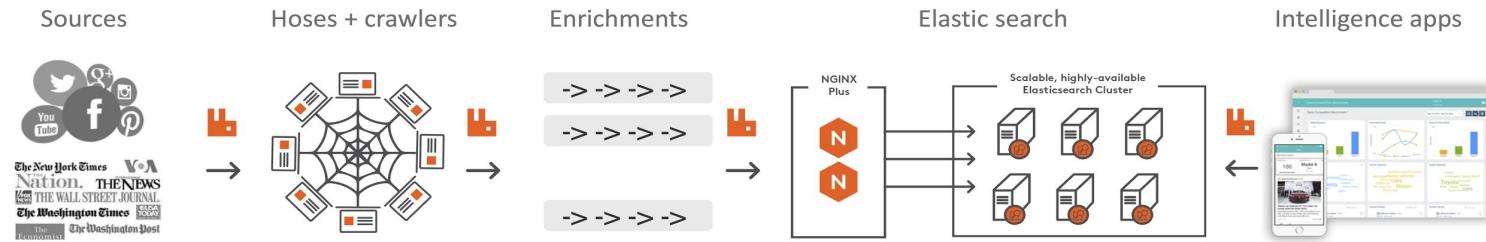
- NER (person, location, organization, ...)
- NED ([https://en.wikipedia.org/wiki/Tim\\_Cook](https://en.wikipedia.org/wiki/Tim_Cook))
- Relation & event extraction
- Truth finding, link prediction, graph mining

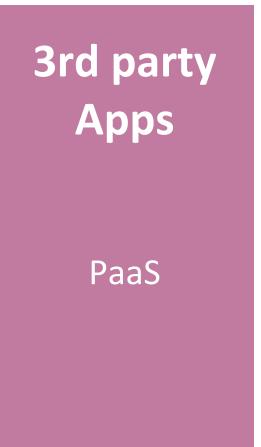
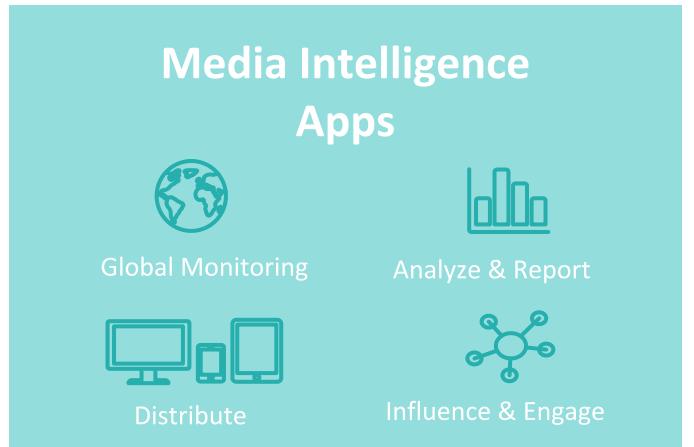
## Data Augmentation (15 languages):

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

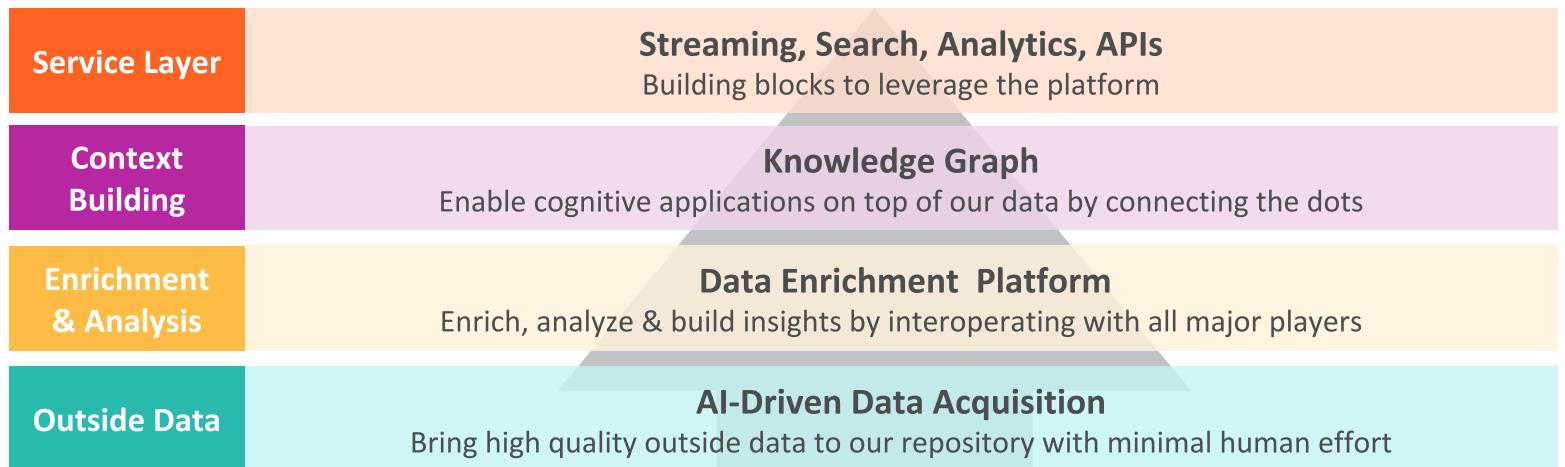
## Media Intelligence applications

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



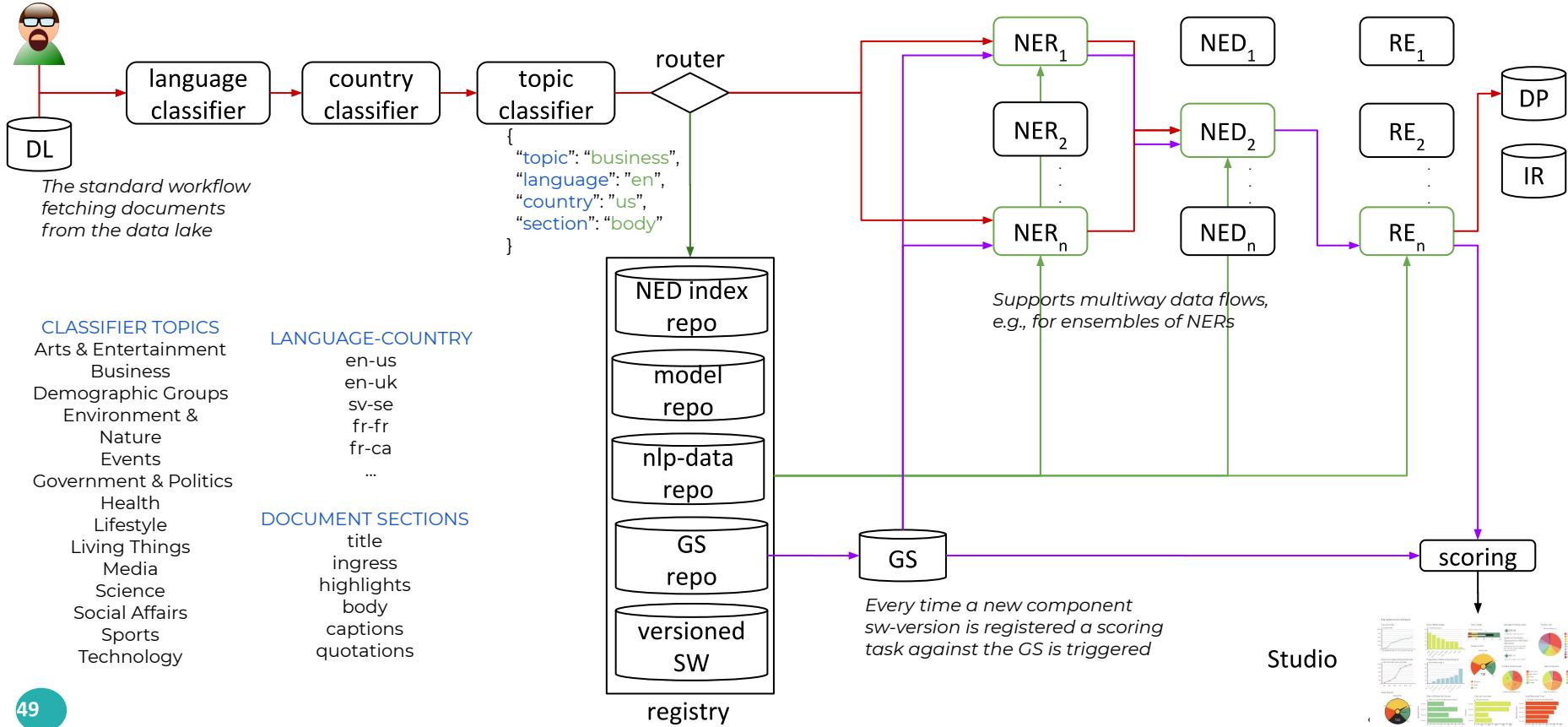


**100M** documents ingested daily  
**150** NLP/IR pipelines  
**100's** Billions of Searches



**FAIRHAIR**

# NLP Pipelines



# NLP Pipelines

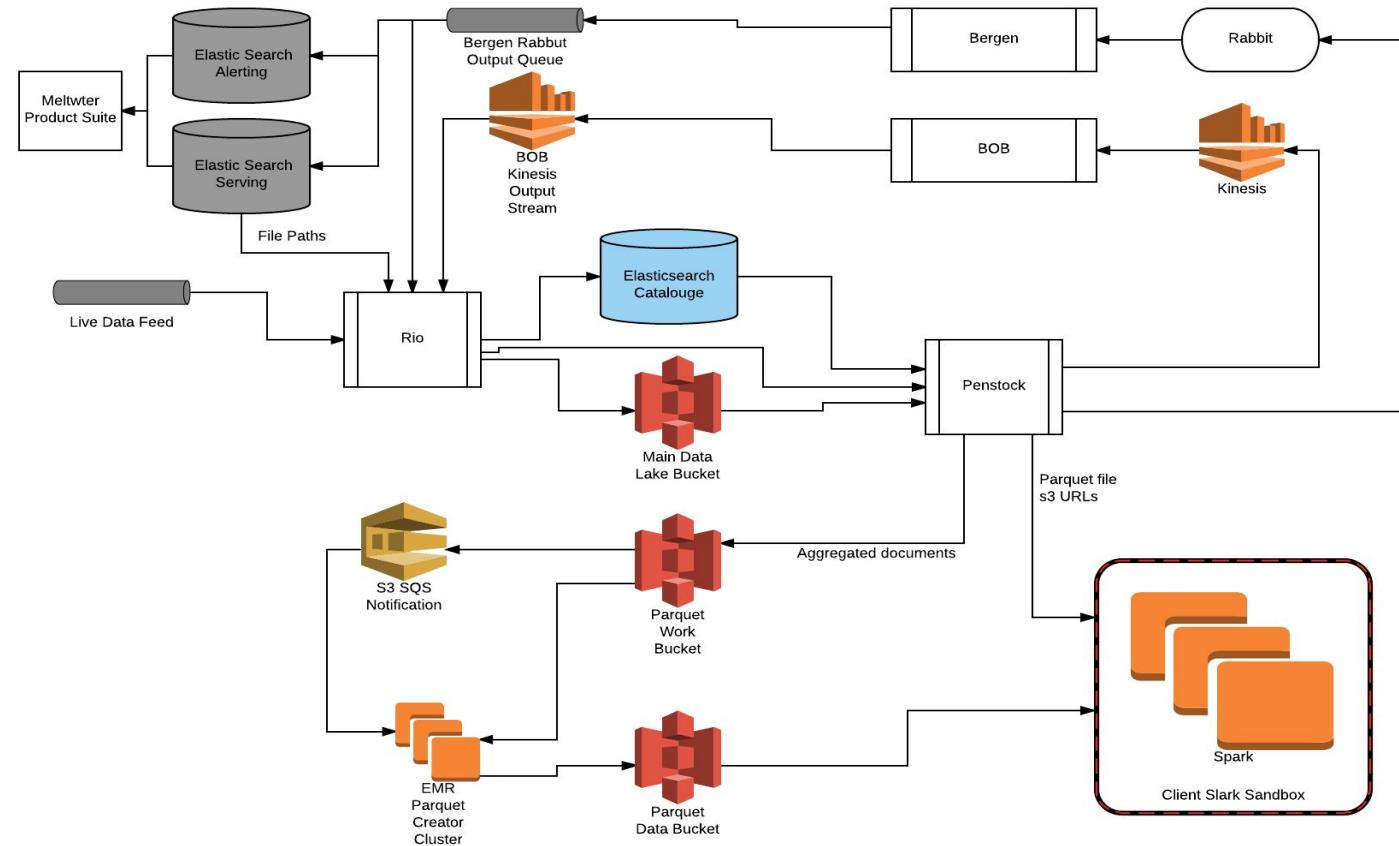
We cannot foresee all uses of our data: **Developer APIs** to **Integrate** and **orchestrate** third party tools.

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

The screenshot shows the Meltwater Enrichments Marketplace interface. At the top, there's a navigation bar with 'ENRICHMENTS' (highlighted), 'INSIGHTS', 'ENRICHMENTS', 'DATA SETS', and a red 'GO' button. Below the navigation, tabs for 'Enrichments', 'Enrichments Marketplace' (selected), and 'Unpublished Enrichments' are visible. The main content area is divided into two sections: 'Recent Enrichments' and 'Popular Enrichments'. Each section contains five cards, each representing a different enrichment service with its name and a small icon. In the 'Recent Enrichments' section, the services are: CoreNLP service - Sentences extraction, FAIRHAIR Basic annotators, IBM Keywords, CoreNLP service - named entities extraction, and IBM Entity Sentiment. In the 'Popular Enrichments' section, the services are: IBM Entities, FAIRHAIR Concepts, FAIRHAIR Keyphrases, Microsoft Parse Tree, and FAIRHAIR Sentiment. Each card has a red '+' sign at the bottom right corner.

Recent Enrichments	Popular Enrichments
CoreNLP service - Sentences extraction	IBM Entities
FAIRHAIR Basic annotators	FAIRHAIR Concepts
IBM Keywords	FAIRHAIR Keyphrases
CoreNLP service - named entities extraction	Microsoft Parse Tree
IBM Entity Sentiment	FAIRHAIR Sentiment

# Data Platform



# Data platform



200B Documents

30M Sources



## Serving Layer

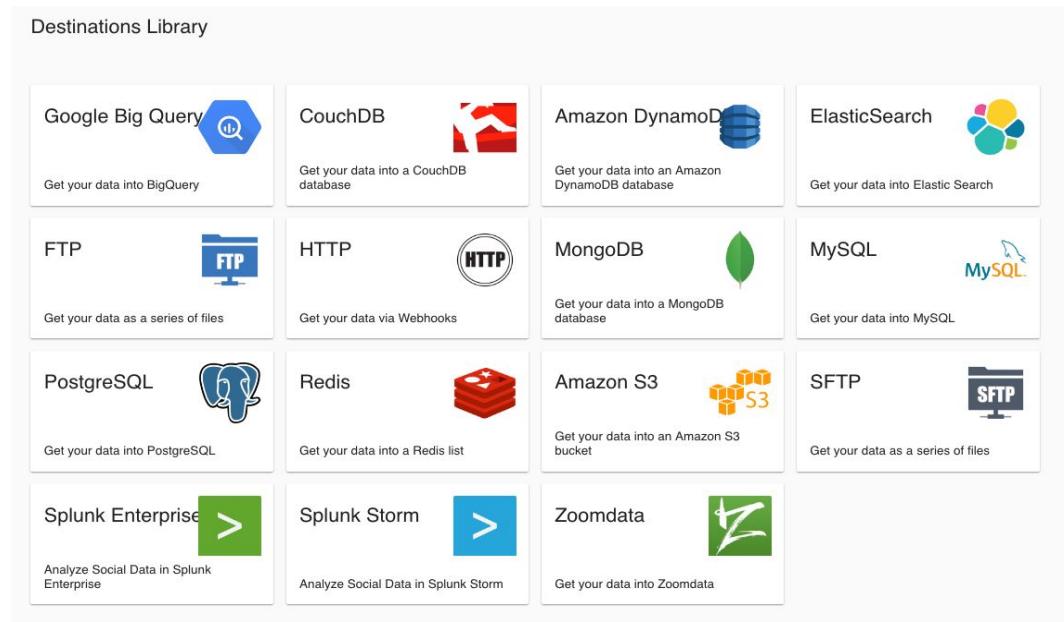


## Analytics Layer



# 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

**About**

 Microsoft Microsoft  
Stock Symbol: MSFT  
Industry: computer software  
URL: [microsoft.com](http://microsoft.com)

Microsoft, a software corporation, develops licensed and support products and services ranging from personal use to enterprise application.

Founded On: 1974-04-04

Location 1: Redmond, WA United States (headquarters)

Location 2: Dublin, Dublin Ireland 18 (offices)

Location 3: Boise, ID United States 83702 (offices)

Location 4: Redmond, WA United States (offices)

Total Funding: USD501,360,001,024.00

Categories: Computer Software: Prepackaged Software

**Key People**



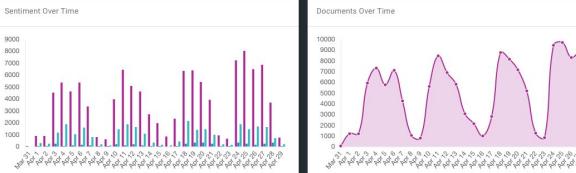
Satya Nadella  
Board Members And Advisors  
Gender: Male

Padmasree Warrior  
Board Members And Advisors  
Gender: Female

Reza Zadeh  
Board Members And Advisors  
Gender: Male

Maria Klawe  
Board Members And Advisors  
Gender: Female

**Company Insights**



Sentiment Over Time

Documents Over Time

Key Phrases Over Time

Phrase	Score
Inc	High
Microsoft	Very High
cloud	Medium

**Business Landscape**

**Competitors**

Company	Score
iLinc	NaN
iNovar Corporation	NaN
BigTwist	NaN
Agily Networks	NaN
PowToon	NaN
PostPath	NaN
Covertix	NaN
LANDesk Software	NaN

**Acquisitions**

Acquisition	Score
Event Zero	NaN
SwiftKey	NaN
Xamarin	NaN
MinecraftEdu	NaN
Secure Islands Technologies	NaN
Talko	NaN
Zikera / Groove	NaN
Tellme Network	NaN

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

## Brand Drill Down

### TOP POSITIVE ARTICLES

 Cisco's Executive Chairman Chambers not to seek re-election  
Published: 09/08/2017.  
[Read Article](#)

 Cisco Selects IR Prognosis for Inclusion in Global Price List  
Published: 09/09/2017.  
[Read Article](#)

 BRIEF-Cisco executive chairman adopted a pre-arranged stock trading plan to sell..  
Published: 09/01/2017.  
[Read Article](#)

### TOP NEGATIVE ARTICLES

Avast, Cisco Confirm: CCleaner Malware Targeted Large Technology Companies  
Published: 09/21/2017.  
[Read Article](#)

Hackers used Avast's CCleaner breach to attack technology companies: Cisco  
Published: 09/21/2017.  
[Read Article](#)

CCleaner hackers attacked Microsoft, Intel, Cisco, and other tech giants  
Published: 09/21/2017.  
[Read Article](#)

### Brand Net Reach by Country

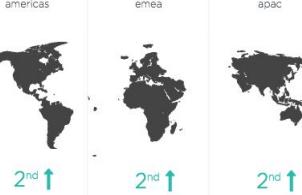
Cisco Systems had the largest Brand Net Reach in Cameroon, Tunisia, and Morocco in September.



Cisco Systems ● IBM ● Avaya ● Polycom

### Brand Net Reach by Region

Cisco Systems was outperformed by its competitors in all major regions.



americas      emea      apac

2nd ↑      2nd ↑      2nd ↑

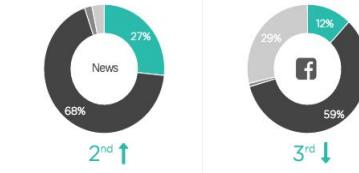
### Themes

When Cisco Systems is mentioned in the news, it's often associated with customer, report, and market.



Negative	Positive
attack	market
malware	report
computer	customer
attacker	technology
hacker	solution

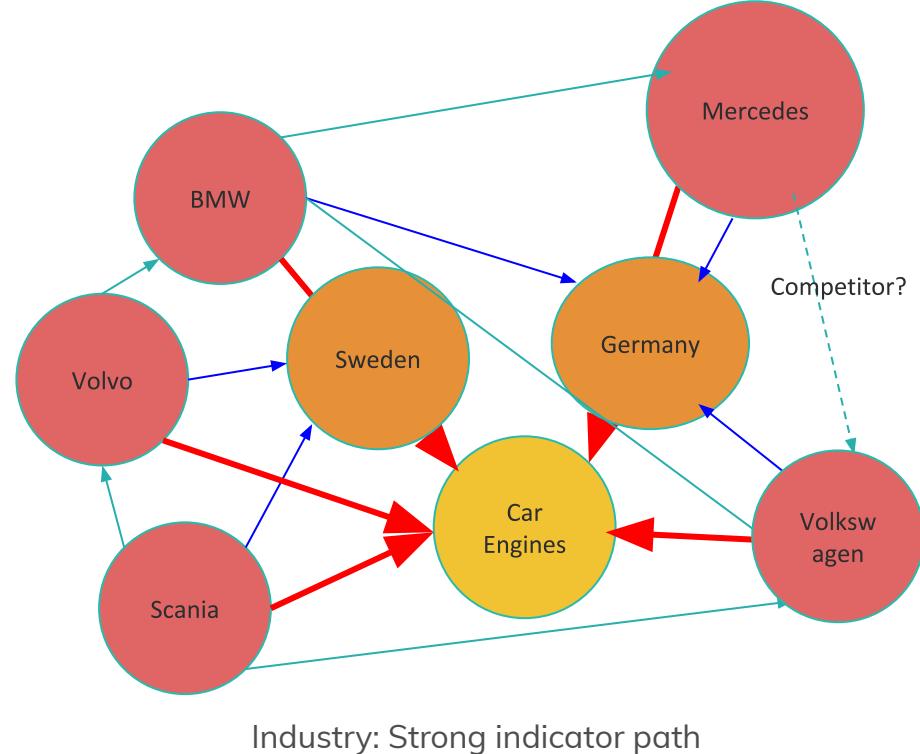
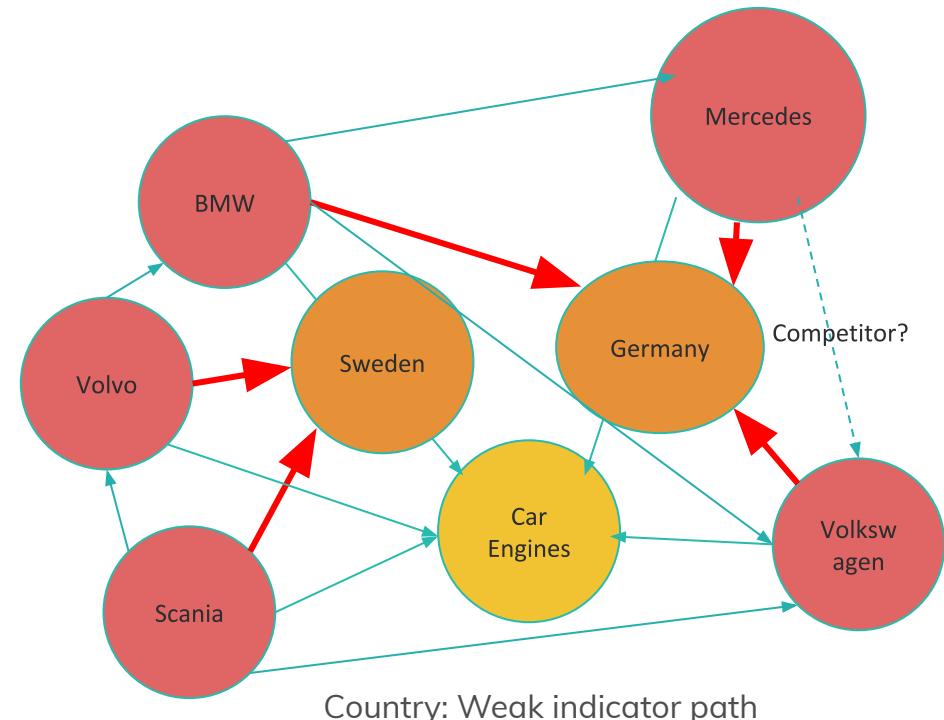
### Channels



Top Themes by Channel

Channel	Theme	Percentage
News	market	27%
News	report	66%
Facebook	customer	59%
Facebook	event	29%
Facebook	world	12%
Facebook	student	3rd ↓

# Path ranking

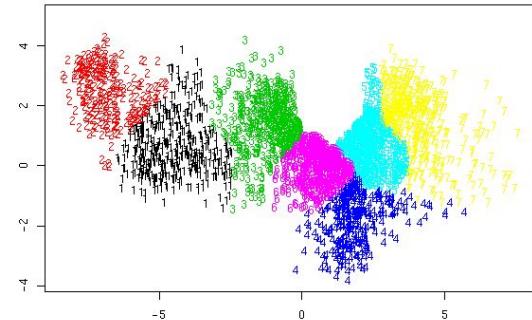


# Graph embeddings

- 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:

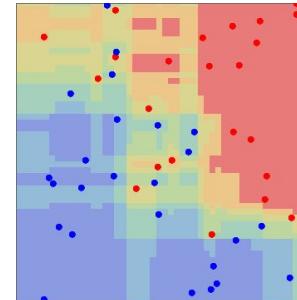
$$s_{LC}(h, \ell, t) = \delta_1^\ell \langle \mathbf{r}_1^\ell | \mathbf{e}_1^h \rangle + \delta_2^\ell \langle \mathbf{r}_2^\ell | \mathbf{e}_1^t \rangle + \delta_3^\ell \langle \mathbf{e}_1^h | \mathbf{D} | \mathbf{e}_1^t \rangle + \delta_4^\ell \langle \mathbf{e}_2^h | \mathbf{R}^\ell | \mathbf{e}_2^t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

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



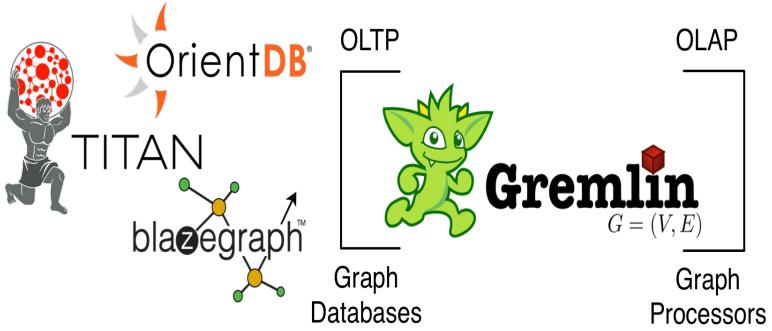
# Link prediction using embeddings

- Given a pair of entities ( $e_1, e_2$ ), 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  $\langle e, r, e \rangle$
- 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  $\langle e_1, e_2, r \rangle$



# Meltwater's Fairhair.AI KG

275M Facts Mined  
(Distinct)



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

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