Outside Insight

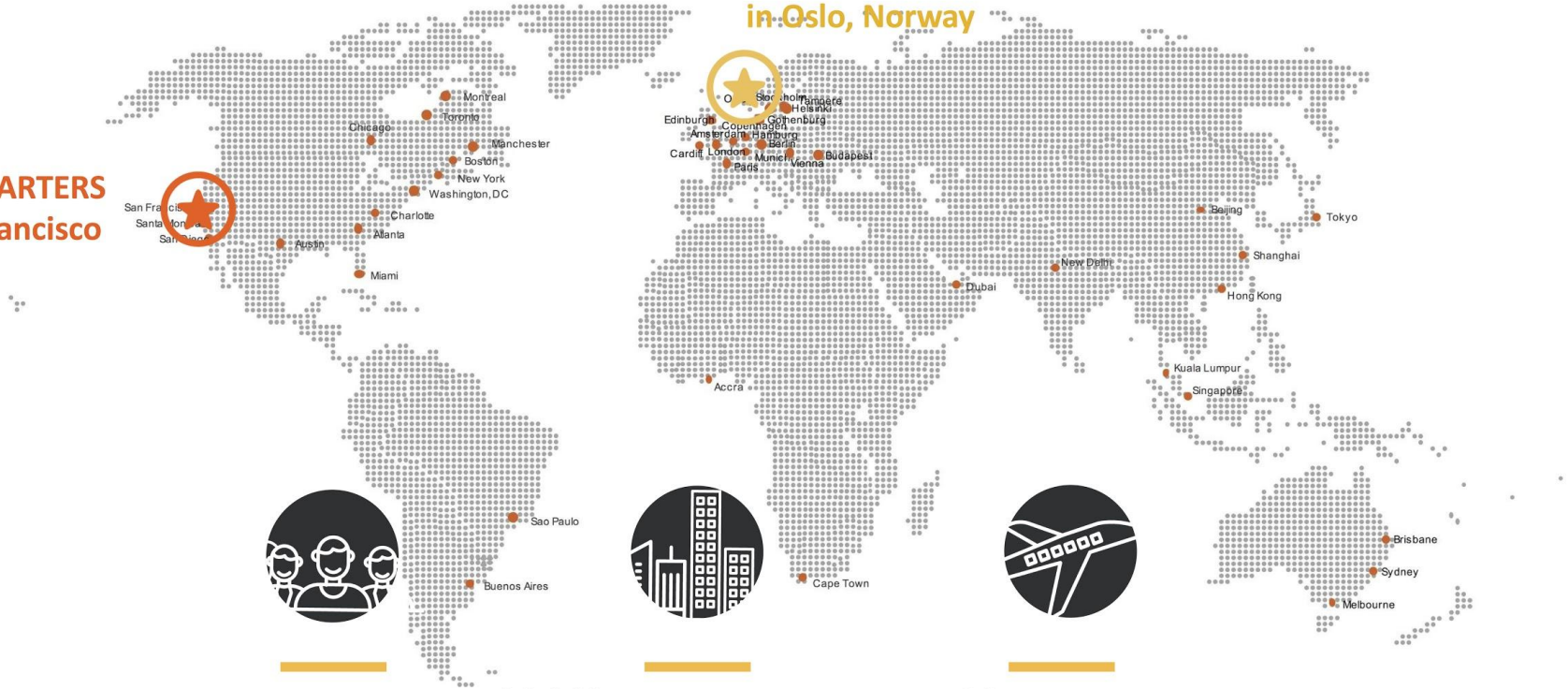
The Meltwater Data Science Platform

Meltwater

Meltwater: Who are we?

FOUNDED 2001
in Oslo, Norway

HEADQUARTERS
in San Francisco



1500 employees worldwide

26,000 Business customers in **108** countries

60 offices across **27** Countries

Meltwater: Media Intelligence

Sources: Editorial, Social, Broadcasts

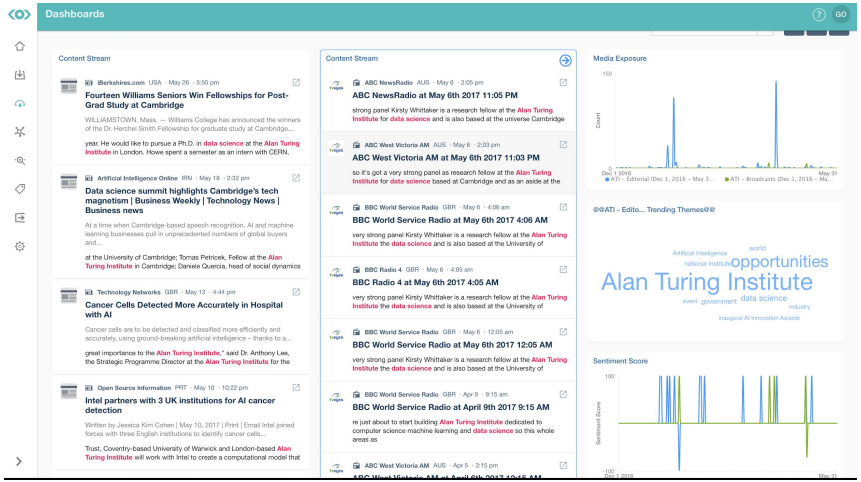
media exposure

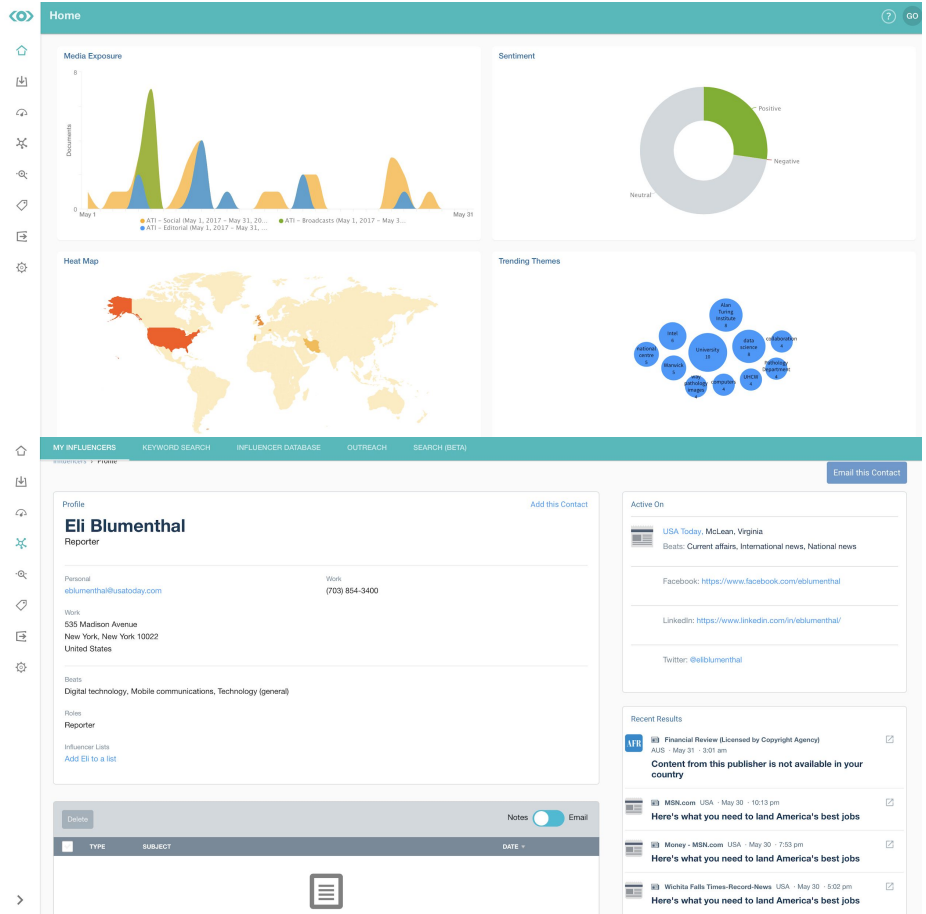
trends

influencers

sentiment analysis

More than 300k different types of user queries





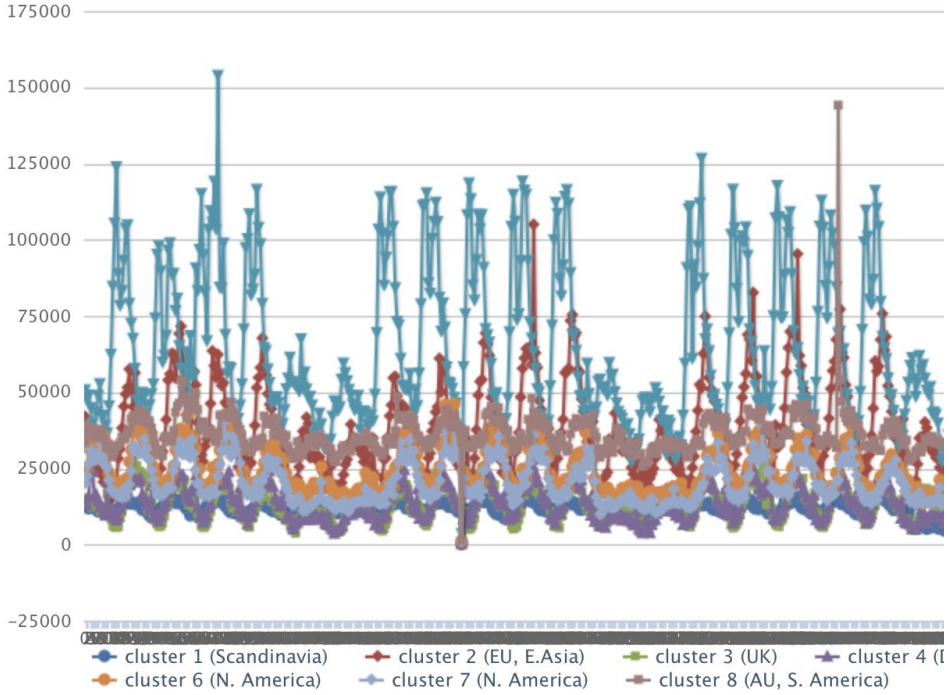
Meltwater: Some numbers

Our ingestion fetches about 3.3M documents / day from 190k editorial sources, re-crawled every 30 minutes.

With the social fire hoses we go up to 30M documents / day.

Since its foundation, Meltwater has indexed almost 200B documents.

Name	Country	Language	Documents
Notiradar	Mexico	Spanish	40700
福建东南新闻网	China	Chinese (simpl.)	23182
中工网	China	Chinese (simpl.)	20953
매일경제	Korea, Republic Of	Korean	20191
جستجوگر اخبار تی نیوز	Iran (islamic Republic Of)	Persian	18055
Match 生活網	Taiwan	Chinese (trad.)	17512
47NEWS	Japan	Japanese	16966
Nambia Press Agency	Namibia	English	9521
中金在线 - 外汇网	China	Chinese (simpl.)	7957
Onet.pl	Poland	English	7127



Meltwater: How it is done

Ingestion:

- Social media hoses (partnerships)
- Editorial (partnerships + web crawling)
- Broadcasts (views on the above)

Storage and search:

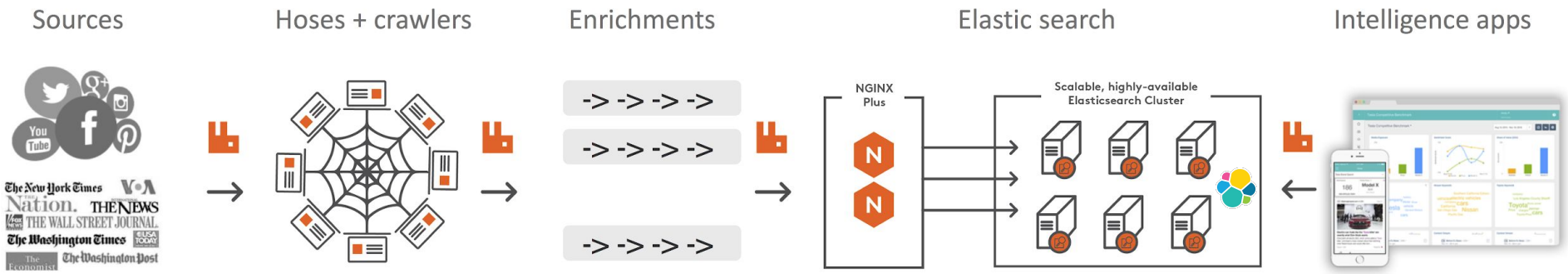
- Elastic search
- Rabbit MQ (distributed queues)
- AWS

Enrichments (15 languages):

- Text categorization (*topic*, *language*)
- NERD (*person*, *location*, *organization*, ...)
- NED (https://en.wikipedia.org/wiki/Tim_Cook)
- Sentiment Analysis

Media Intelligence applications (Custom)

- Boolean queries (*keywords / entities*)
- Counters
- Aggregates
- Drill downs / pivoting



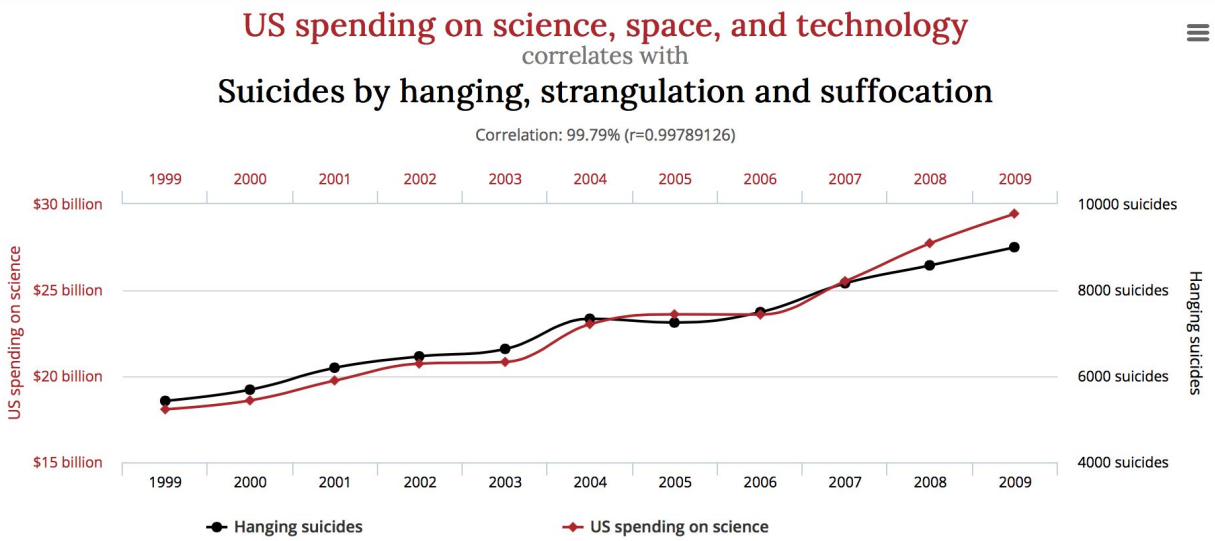
Building Insights

Raw data, statistics, aggregates, summaries, and visualizations are all good but…

- Insights are still discovered by analysts by munching and interpreting the data
- …and we all know how data can be deceitful

What we need is a systematic way to mine, propose, and explain possible insights

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



Source: www.tylervigen.com (Spurious Correlations)

Vision: Insight Building as a Platform

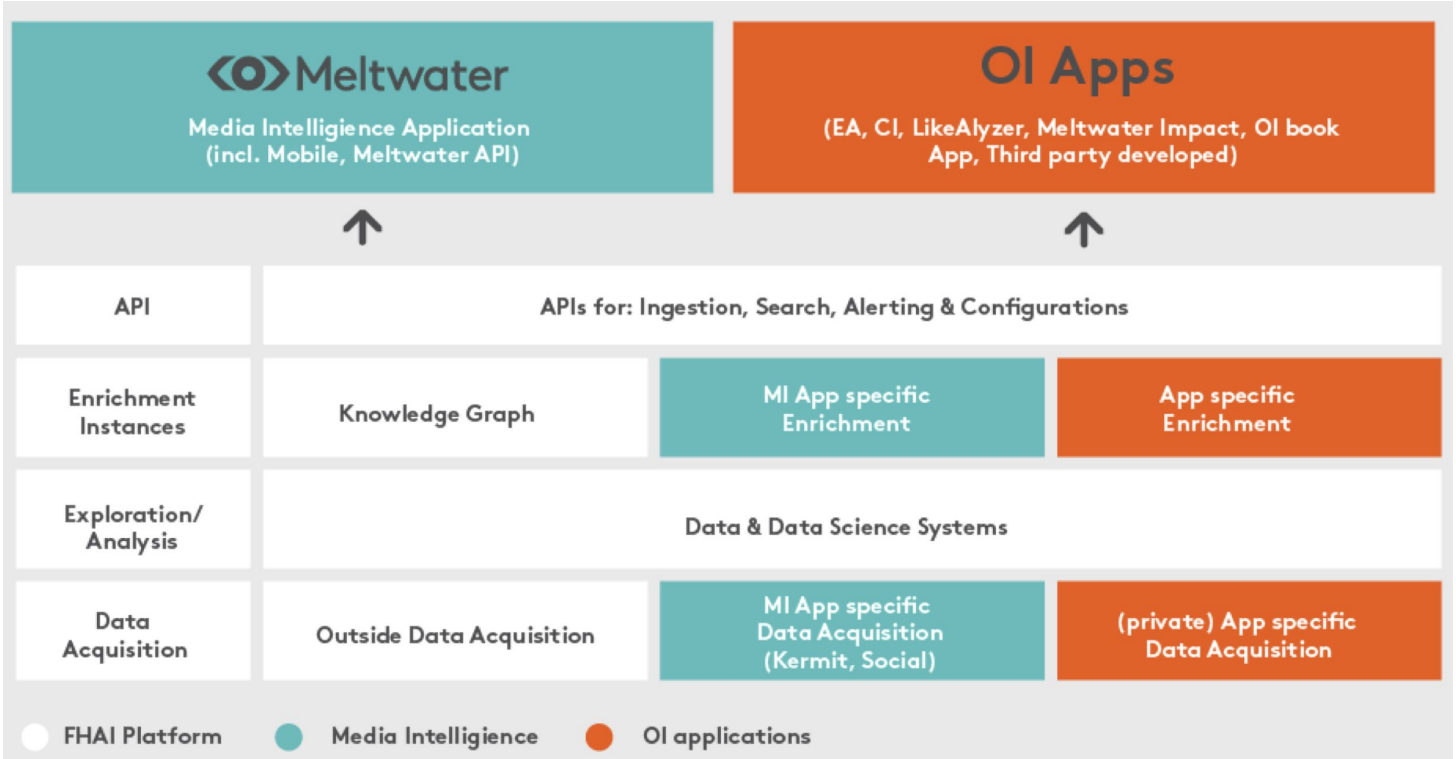
Build a world class AI platform for a new software category

Outside Insight

FAIRHAIR
AI

Meltwater's AI-First Approach

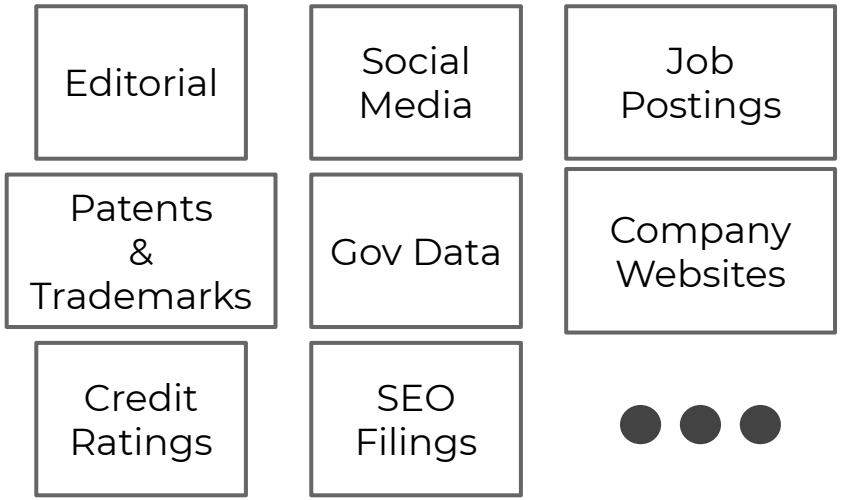
Use Data Science (data, algorithms, infrastructure, tools) to power everything

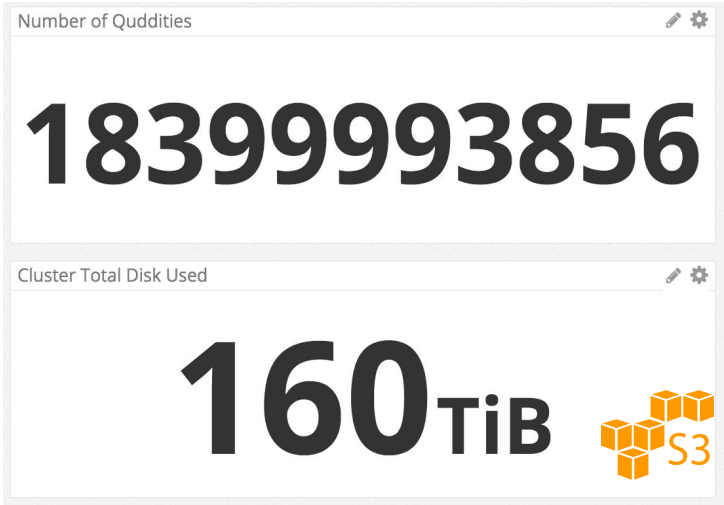


Data and Content Lakes

Factual information: wherefrom?

Need to restrict the domain: focus on the corporate domain, i.e., companies, people, products, ...



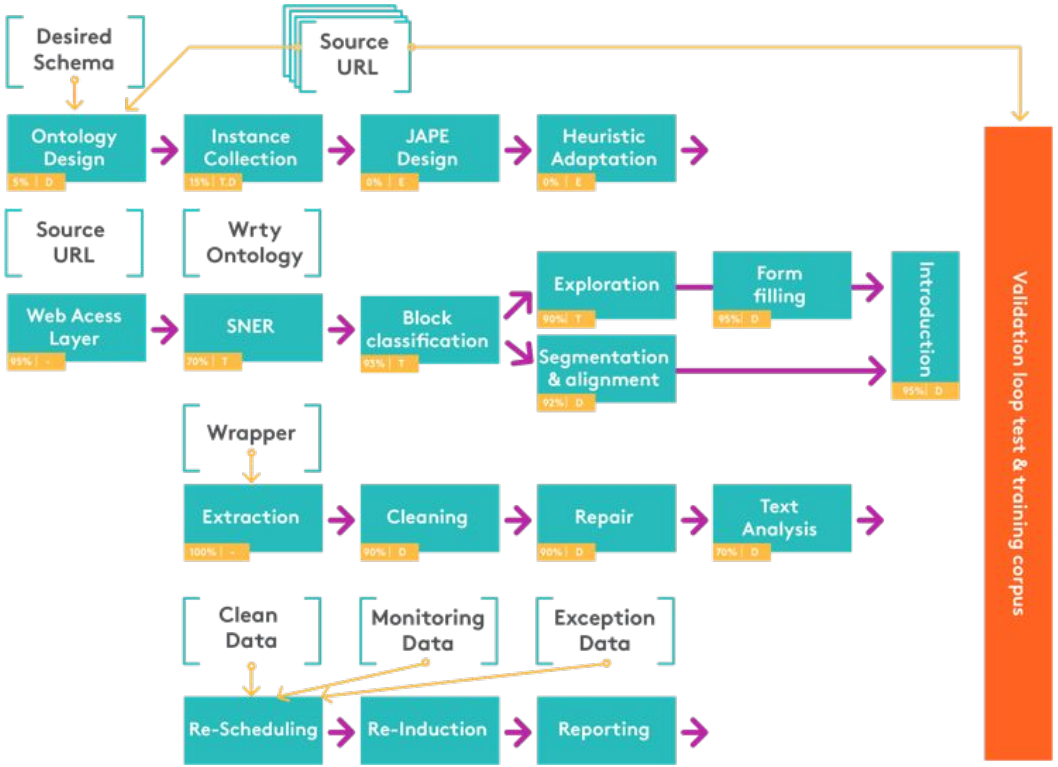


just for last year

Data Lake and Outside Data

Factual information... How? Get as close as possible to the original source of information

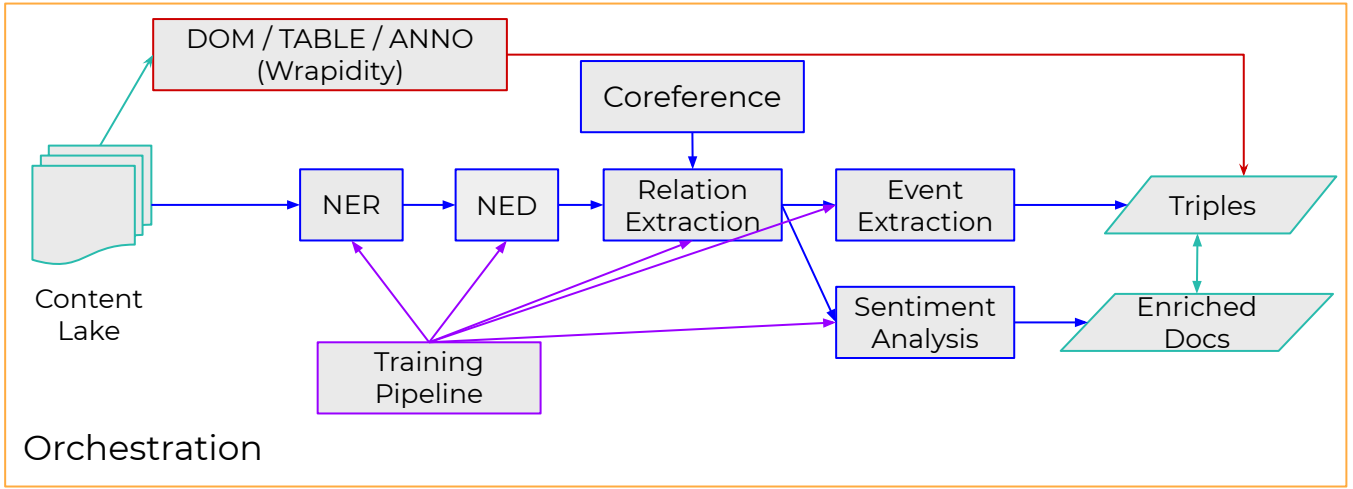
Wrapidity (acq. 2016) automates outside data acquisition via web scraping



Enrichments

Factual information: how?

Linguistic enrichments provide a way to support semantic retrieval and fact extraction.



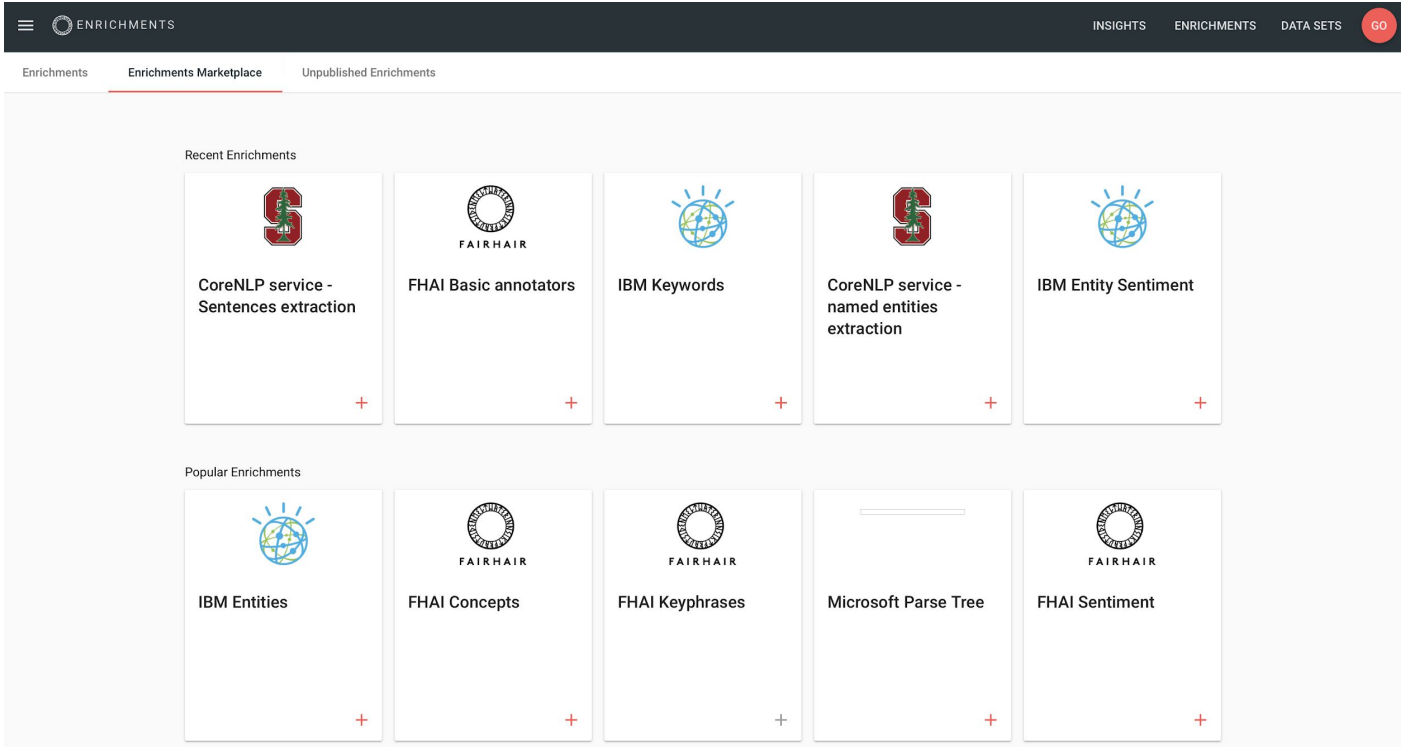
Summary:

- Scalable and distributed enrichment workflows (Cosmify acquisition),
- CRFs for NER, PageRank (variant) for NED, CNNs and LSTMs for Relation/Event extraction, Bayesian classification for sentiment analysis,
- TensorFlow and GPUs for training infrastructure

Enrichments

We can’t 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.



Knowledge Graph

Wait... did you say PageRank, triples? So do you have a (Knowledge) Graph?

Content:

- Companies
- Brands
- Products
- Key people
- Influencers

Goals:

- Relate facts
- Data mining
- Cognitive applications (higher-order reasoning)

Challenges:

- Data Cleaning
- Data deduplication / integration
- Truth Finding

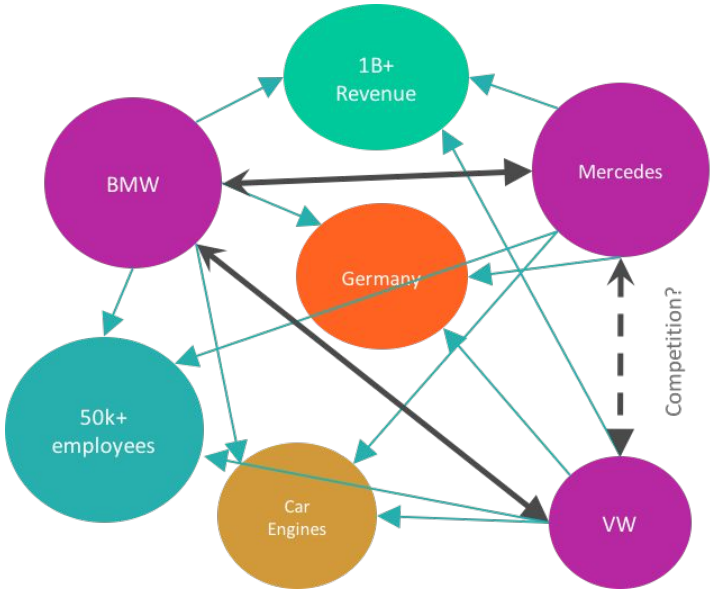
Cognitive Applications

Infer high-level insights from a set of extracted events/facts.

- Competitor
- Customer
- Investment
- Lawsuit/Litigation
- Partnership
- Supplier
- Acquisition
- Out/under performance
- Expanding Operations
- Compliance
- Funding Developments
- Leadership Changes
- New Offerings
- Bankruptcy,
- Restructuring, Cost Cutting

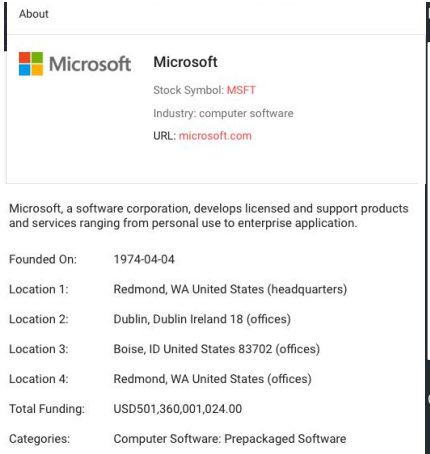
Insight discovery:

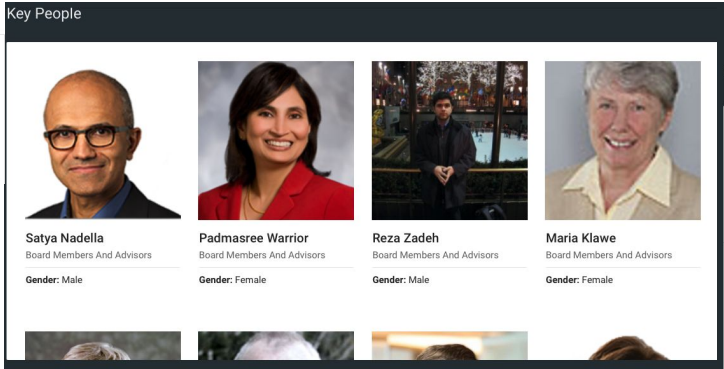
- Rule/Graph mining (data cleaning)
  - GPAR (VLDB '15)
  - RUDIK (internal, paper submitted)
- Link prediction (data enrichment, fact checking)
  - Path Ranking Algorithms (PRA)
  - Probabilistic Soft Logic (PSL)

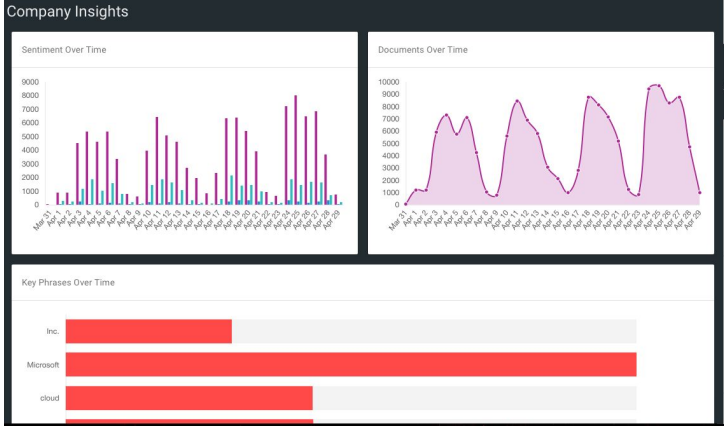


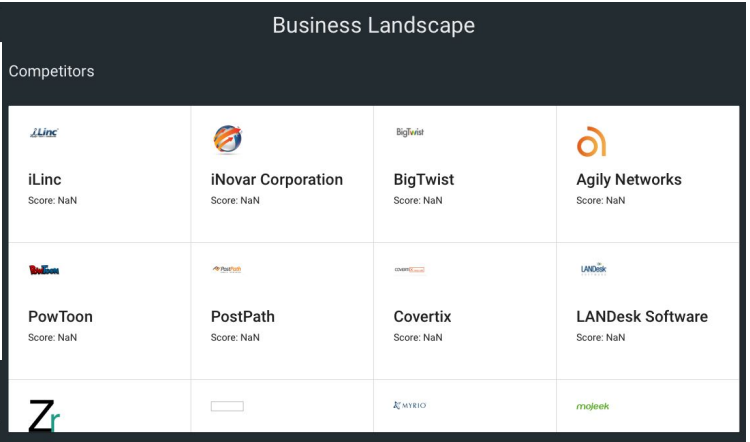
Cognitive Applications

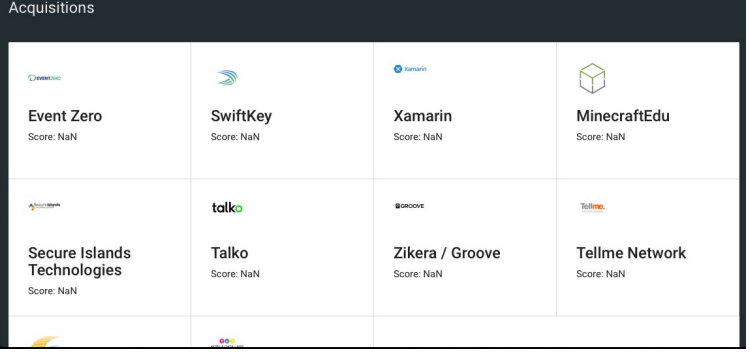
Cognitive Applications complement traditional media intelligence tools











Involve Users, Entrepreneurs, and Researchers

6 Data Science Hubs (co-working spaces)

SHACK15
The Data Science Hub

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

Meltwater Entrepreneurial School of Technology

mest
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

THE ALAN TURING INSTITUTE

Questions?

рахмат
danke
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
mersi
barka
welalin
tack
spas
kia ora
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
kiitos
dankie
dhanyavad
nanni
nandri
maururu
köszönöm
thank you
gracias
хвала
asante
manana
obrigada
tenki
bedankt
bayarlalaa
gràçie
hvala
aciu
dankon
akun
djiere dieuf
tau
mochchakkeram
murakoze
enkosi
dziękuję
дякую
mamnun
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
sobodi
děkuji
taiku
obrigado
sukriya
kop khun krap
arigatô
takk
dakujem
trugarez
mèsi
sagolun
ありがとう
tanemirt
grazie
didi madloba
najis tuke
rahmet
terima kasih
dhanyavadagalu
shukriya
mercè
мерси
kam sah hamnida
rahmat
diolch
ευχαριστώ
merci
감사합니다
xièxie