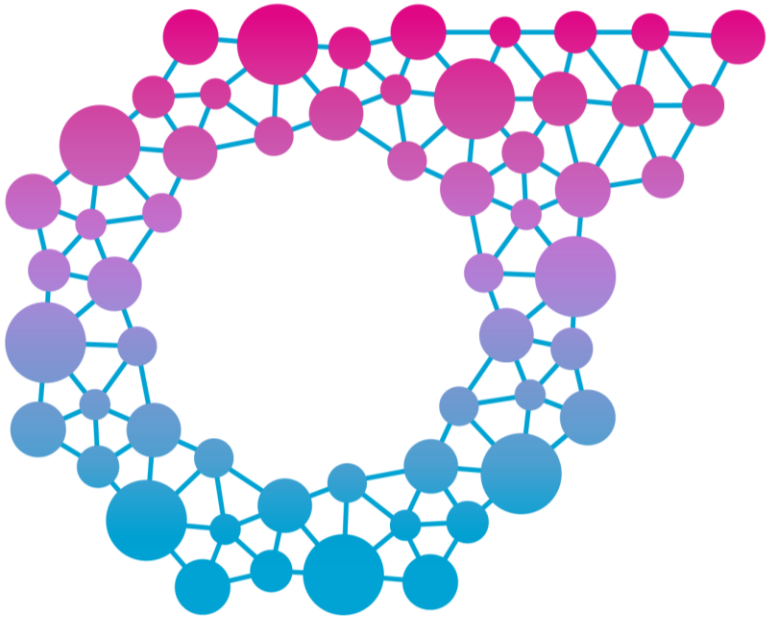
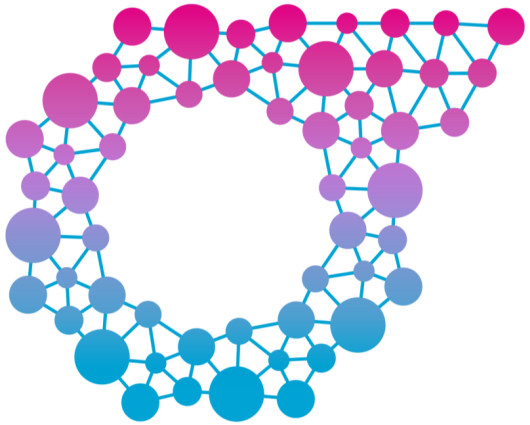


sentometrics

Data Science Meetup Leuven

January 15, 2020



sentometrics

We aggregate textual data into indices

What is an index?

An index is a value that moves over time

We create those indices starting from textual data

VUB-prof onderzoekt de moord op de dino's

BRUSSEL/CHICXULUB – Een internationaal onderzoeksteam heeft eind februari de boringen beëindigd in de krater van de meteoriet die de dinosauriërs om zeep heeft geholpen. VUB-professor Philippe Claeys staat aan het hoofd van een van de zes onderzoeksteams die de gevolgen van de inslag zullen berekenen op basis van de gedane boringen.

De dinosauriërs zijn uitgestorven door de inslag van een asteroïde of een komeet op de aarde, zo'n 65 miljoen jaar geleden. Na een kleine 25 jaar gepalaver zijn de meeste wetenschappers het daar over eens. Sedert twaalf jaar meent men ook te weten welke krater die fatale meteoriet geslagen heeft: de Chicxulubkrater die werd ontdekt nabij de Golf van Mexico. Hij heeft een diameter van 200 kilometer. Dat dit groot gat niet eerder werd opgemerkt, komt omdat de aarde – in tegenstelling tot bijvoorbeeld de maan – een 'levend' hemellichaam is: aan het aardoppervlak doen zich voortdurend grote en minder grote veranderingen voor. Zo hebben zich in de loop der eeuwen heel wat nieuwe aardlagen op de krater vastgezet. Hoewel de bewuste meteoor een diameter van tien kilometer had, is hij relatief spoorloos. Hij genereerde een energie goed voor vijf miljard Hiroshimabommen en verpulverde bijgevolg zichzelf. Daarbij wierp hij zoveel stofdeeltjes op, dat de aarde gedurende weken, zoniet maanden, in duisternis werd gehuld zodat de fotosynthese stilviel, planten stierven en dieren verhongerden. Bovendien stuwde de meteoriet grote hoeveelheden CO2 (het broeikasgas) en SOX (het zure-regengas) de atmosfeer in. Het stof en de gassen zorgden ervoor dat vijftig procent van fauna en flora op aarde afstierf. Tsunamigolven van vijfhonderd meter hoog deden daar nog een schepje bovenop.

Wat de onderzoekers nu gaan doen, is aan de hand van de residuen en reagensstoffen de juiste cijfers van de impact berekenen. Eén VUB-lab gaat aan de hand van radiochemisch onderzoek de precieze ouderdom van de krater trachten te bepalen. Een ander VUB-lab gaat berekenen wat na de inslag de temperatuur van het oceaanwater moet zijn geweest en wat daarvan de gevolgen waren voor de bioproductiviteit in het water. Professor Claeys doctoreerde in de geologie op de Berkeley Universtiteit in Californië en deed onderzoek in heel wat grote internationale universiteiten. Onlangs keerde hij terug naar de vakgroep Geologie van de VUB. Naar eigen zeggen omdat die op een hoog niveau staat. Het Amerikaans weekblad *New Scientist* bracht zopas een groot interview met hem over het onderzoek van de krater.

By the way: dat er ooit – morgen of binnen dertig miljoen jaar – opnieuw zo een asteroïde inslaat en ons, kleine dingootjes, de grond inboort, is helemaal niet uitgesloten. De professor heeft het zelf gezegd.

MB

Index creation is easy

If you think about it, that cannot be so hard?

Index creation is easy

You assemble the texts

You turn the individual texts into a number

You aggregate the numbers into an index

Index creation is easy

Assume you have 100,000 time-stamped articles

Read it into your favorite programming environment

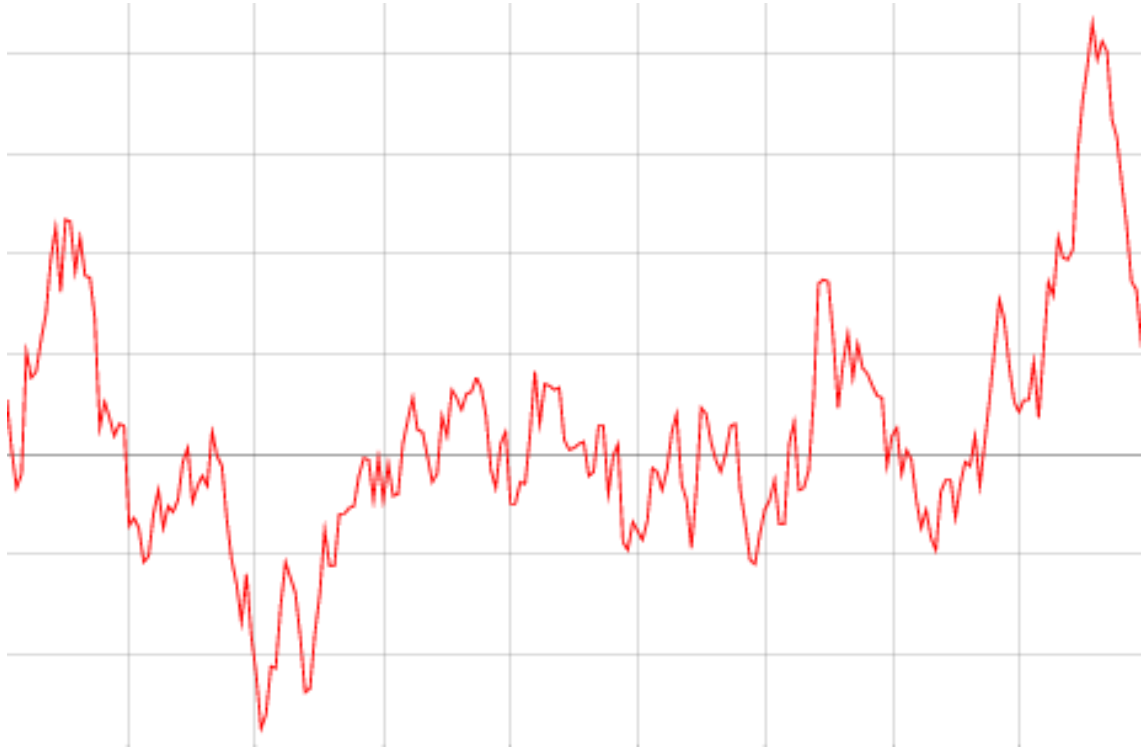
Associate to every text a number

- Just 1, to count the articles
- Maybe even the sentiment – so many APIs out there

You take the average per day and …

Index creation is easy

… an index appears



100, maybe 200 lines of code? Easy!!

Index creation is easy… or not?

Et alors? What does the index mean? Is it useful?

Your index is nothing more than a wild one…

Index creation is NOT easy

If you **really** think about it, index construction from alternative data is quite hard…

Index creation is NOT easy

Data – format, meta, language, source, selection, …

Quantification – sentiment, topics, entities, …

Aggregation – averaging, weighting, combination, …

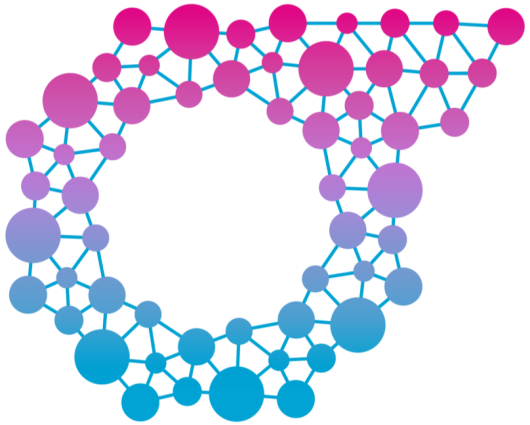
Validation – statistical, human, attribution, integration, …

Designing an index is one thing, maintaining it another

We only talked about the **design** of an index

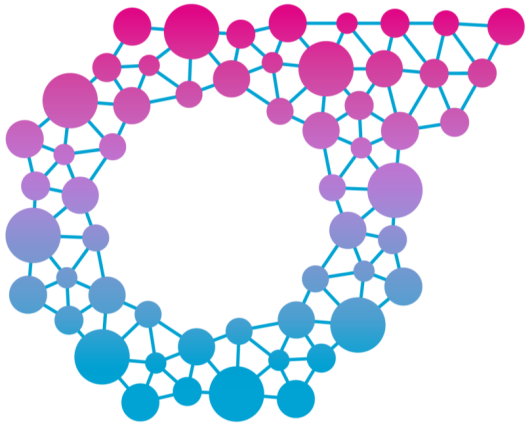
But what about its **maintenance**?

Need for reliable and consistent updating of the index, based on its design rules



sentometrics

We aggregate textual data into indices



sentometrics

We aggregate textual data into practical and explainable index solutions

sentometrics as a research field

Broad:
polarity, bias,
emotion, topic, …

Various outputs in finance, economics and software

Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values

David Ardia [a,b], Keven Bluteau [a,c,*], Kris Boudt [c,d,e]

Managers set the tone: Equity incentives and the tone of earnings press releases ☆

Özgür Arslan-Ayaydin [a], Kris Boudt [b,c], James Thewissen [d,*]

Media and the stock market:
Their relationship and abnormal dynamics around earnings announcements☆

David Ardia[a], Keven Bluteau[b,c,*], Kris Boudt[c,d,e]

Jockeying for Position in CEO Letters: Impression Management and Sentiment Analytics

Kris Boudt and James Thewissen*

Econometrics meets sentiment:
An overview of methodology and applications⋆

Andres Algaba[a,b], David Ardia[c], Keven Bluteau[d,a], Samuel Borms[d,a,*], Kris Boudt[b,a,e]

The R Package sentometrics to Compute, Aggregate and Predict with Textual Sentiment

David Ardia
University of Neuchâtel
HEC Montréal

Keven Bluteau
University of Neuchâtel
Vrije Universiteit Brussel

Samuel Borms
University of Neuchâtel
Vrije Universiteit Brussel

Kris Boudt
Ghent University
Vrije Universiteit Brussel
Vrije Universiteit Amsterdam

Media-augmented consumer confidence index

Andres Algaba[a,b,*], Samuel Borms[a,c], Kris Boudt[a,b,d], Brecht Verbeken[a]

[a] *Solvay Business School, Vrije Universiteit Brussel, Belgium*
[b] *Department of Economics, Ghent University, Belgium*
[c] *Institute of Financial Analysis, University of Neuchâtel, Neuchâtel, Switzerland*
[d] *School of Business and Economics, Vrije Universiteit Amsterdam, The Netherlands*

Next step

From a research group to a viable spin-off

innoviris .brussels
we fund your future

launch
BRUSSELS SPIN-OFF

Vision

"We want to become the reference textual indexing partner for all organizations seeking to valorize textual data. Doing so, we will distribute and maintain reliable indices that grasp the writing pulse of the world."

Team

Founders

PhD business economics, teaching

Product-driven

Andres Algaba

PhD researcher, banking consulting

Organization-driven

Samuel Borms

IT and banking consulting

IT-driven

Jeroen Van Pelt

Prof. econometrics and data science

Innovation-driven

Kris Boudt

David Ardia Prof. data science (HEC Montréal)

Keven Bluteau sentometrics postdoc. researcher

Advisors

Former senior executive at NIBC Bank, KBC Bank, Citibank
Board experience at Rainforest Alliance, Easdaq, Brussels Airport, Guberna

Daniel Couvreur

Chairman of the Board Collibra
Former CEO and president VRT
Former Country Manager IBM
Former General Manager Belgacom

Tony Mary

Chief Information Officer Belga

Tom Wuytack

Market opportunity

Opportunity #1: Textual data science is hard for organizations

Information hidden in internal & external textual data useful for decision-making, but lack of capabilities for analysis and integration

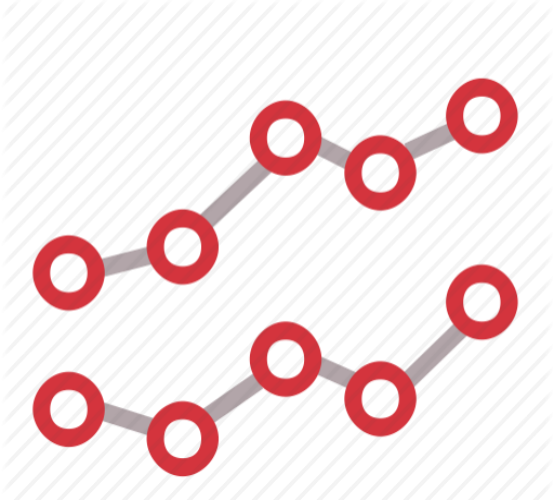
Around 80% of the usable business information originates from unstructured (mostly textual) form

Opportunity #2: Textual indexing agencies are missing

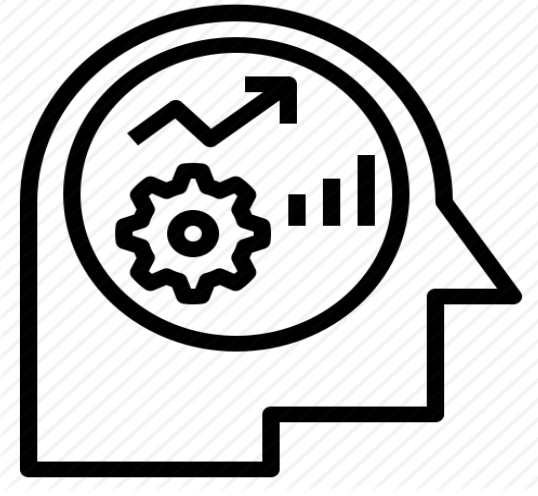
A textual indexing agency publishes reliable textual indices on behalf of its clients in all independence

Technology

Compute



Aggregate



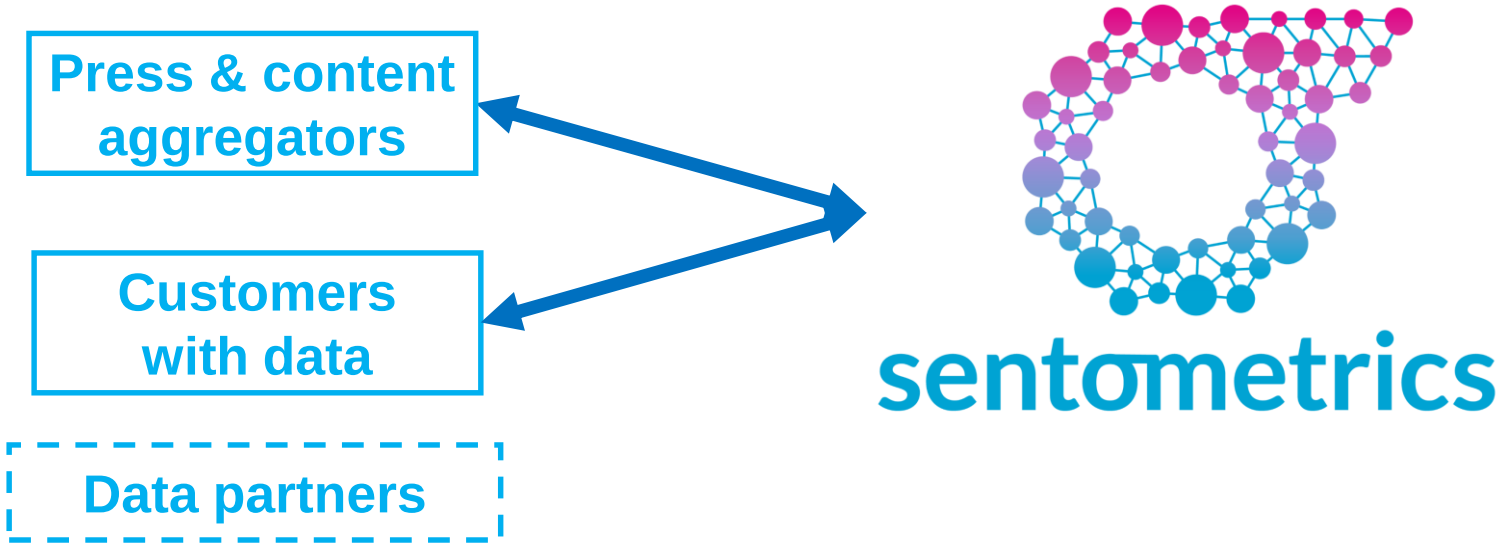
Predict

Text2Feature

Feature2Index

Application-specific (language, domain & purpose)
and **easily interpretable** output

Business model (1/3)



Data partnership with Belga

Functional **data partnership** with Belga in light of three projects

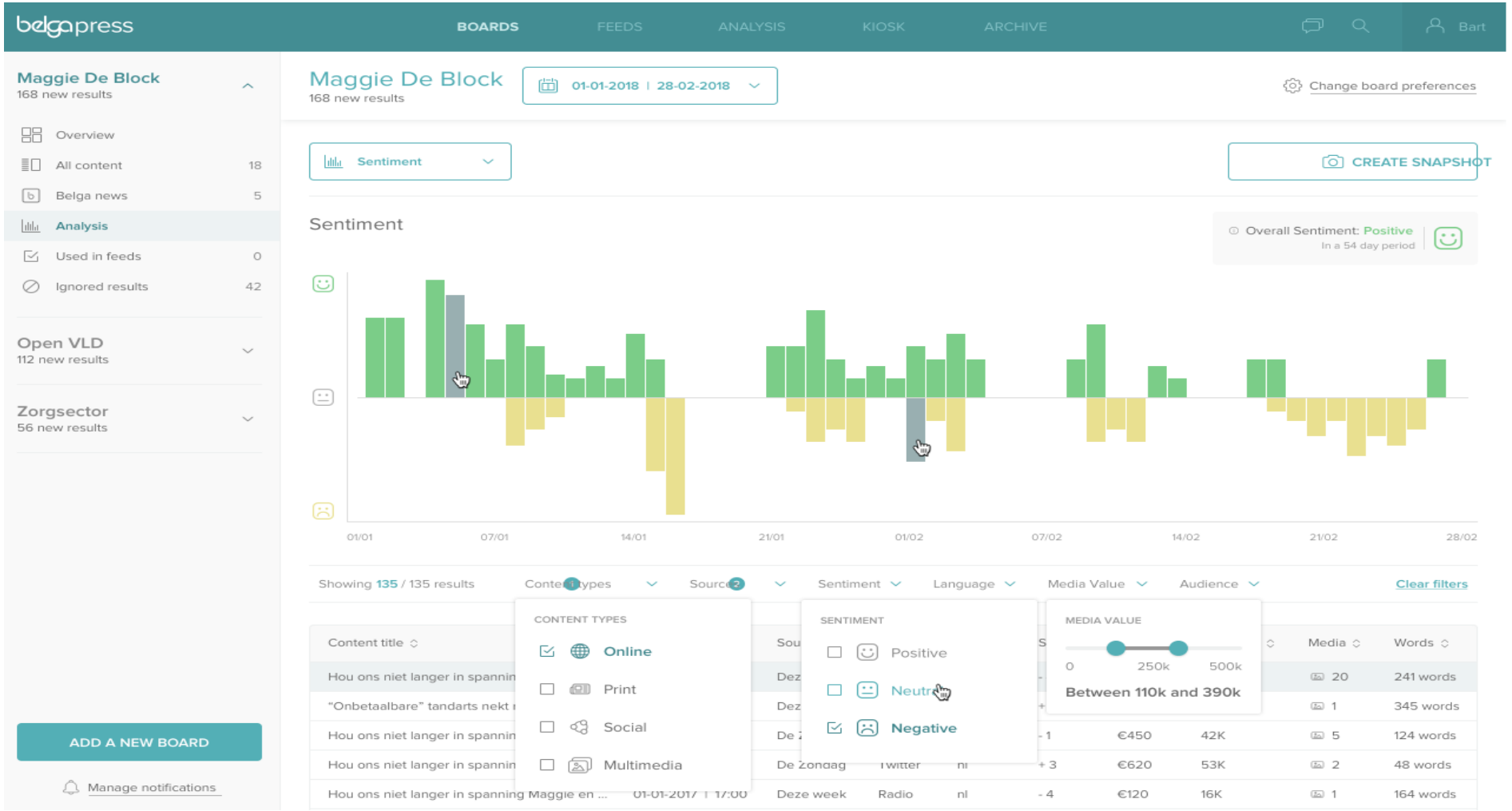
The Gopress/Belga archive combined has over 40 million articles from 1988

Rationale for Belga?

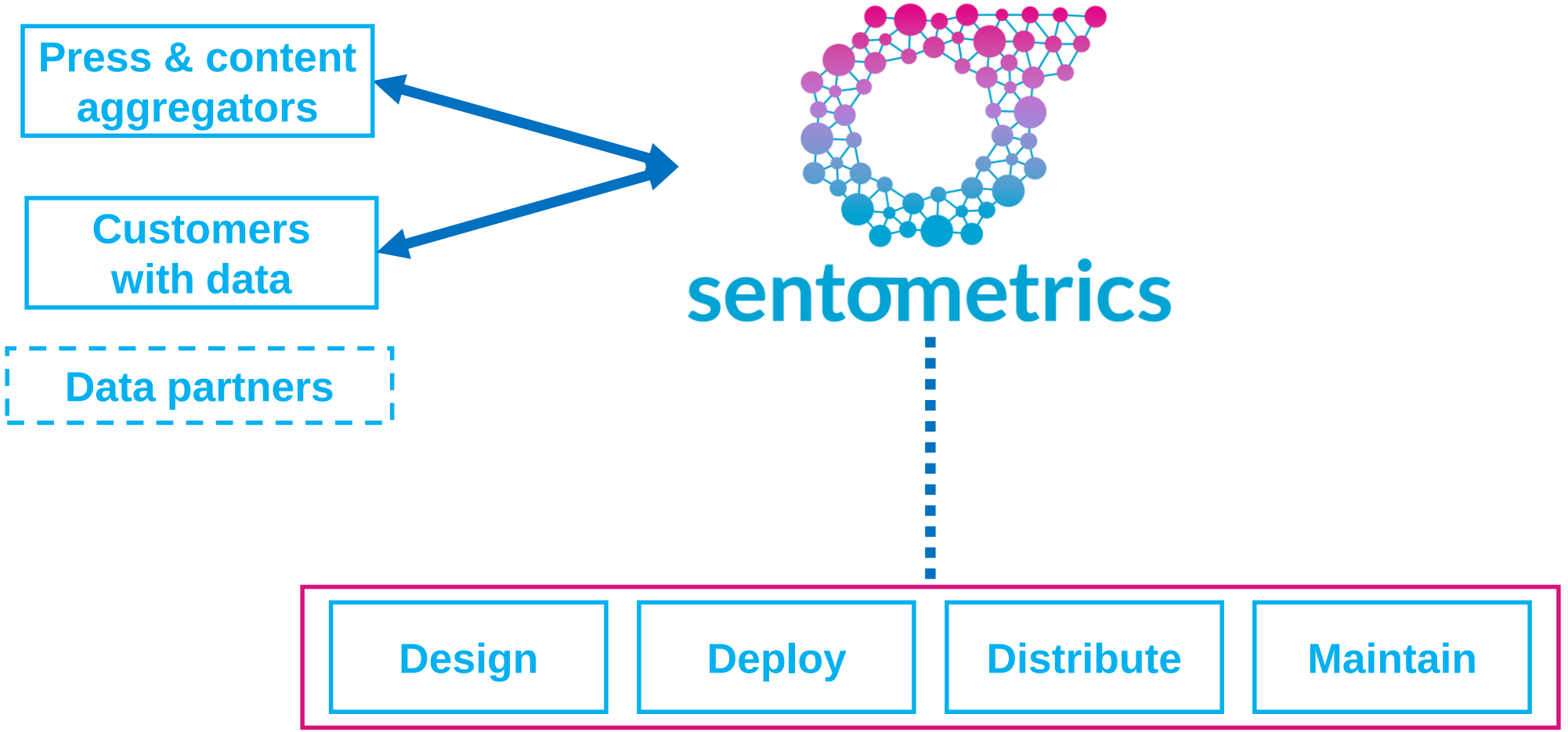
- Local and publisher's content is worth money worldwide
- Low requirements to get started
- Reach out to new market segments

Sentiment scoring for Belga

Belga integrated our Text2*Sentiment* scoring into their client dashboards



Business model (2/3)

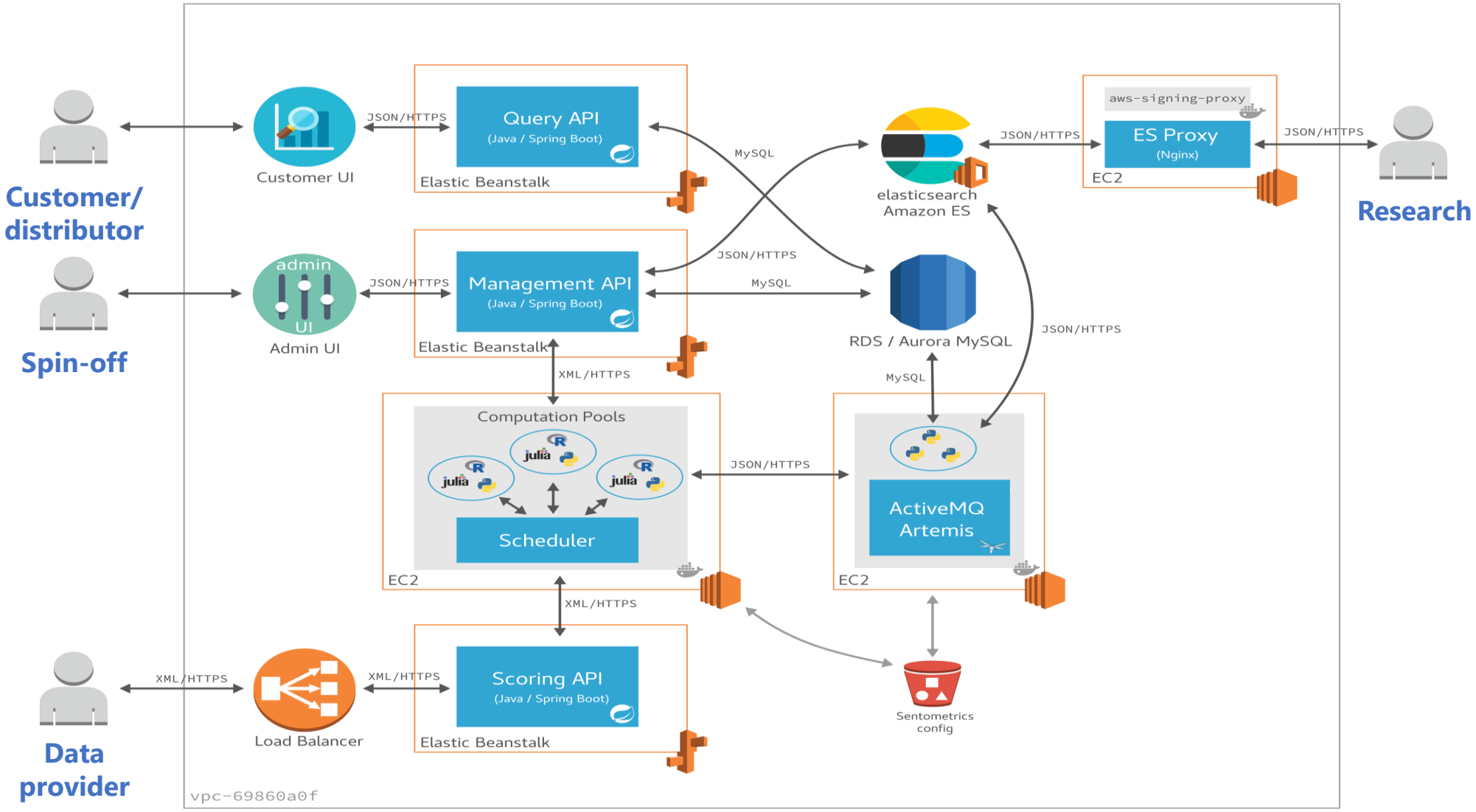


Partnership with Belga spurred the development of our platform

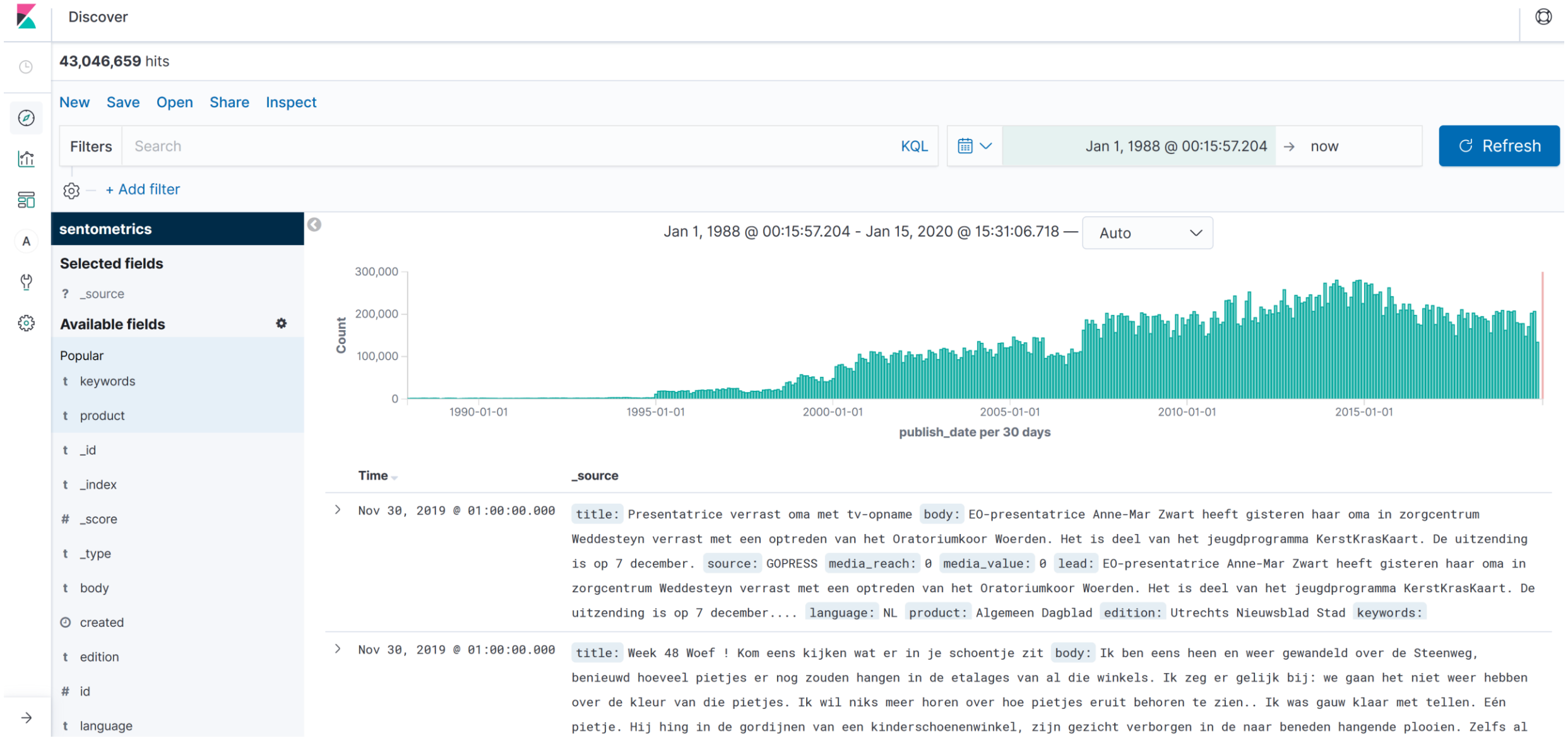
We developed a **cloud-based platform**:

- To upload textual data
- To compute text-based numbers
- To do the research & index creation
- To store created indices
- To facilitate distribution & maintenance

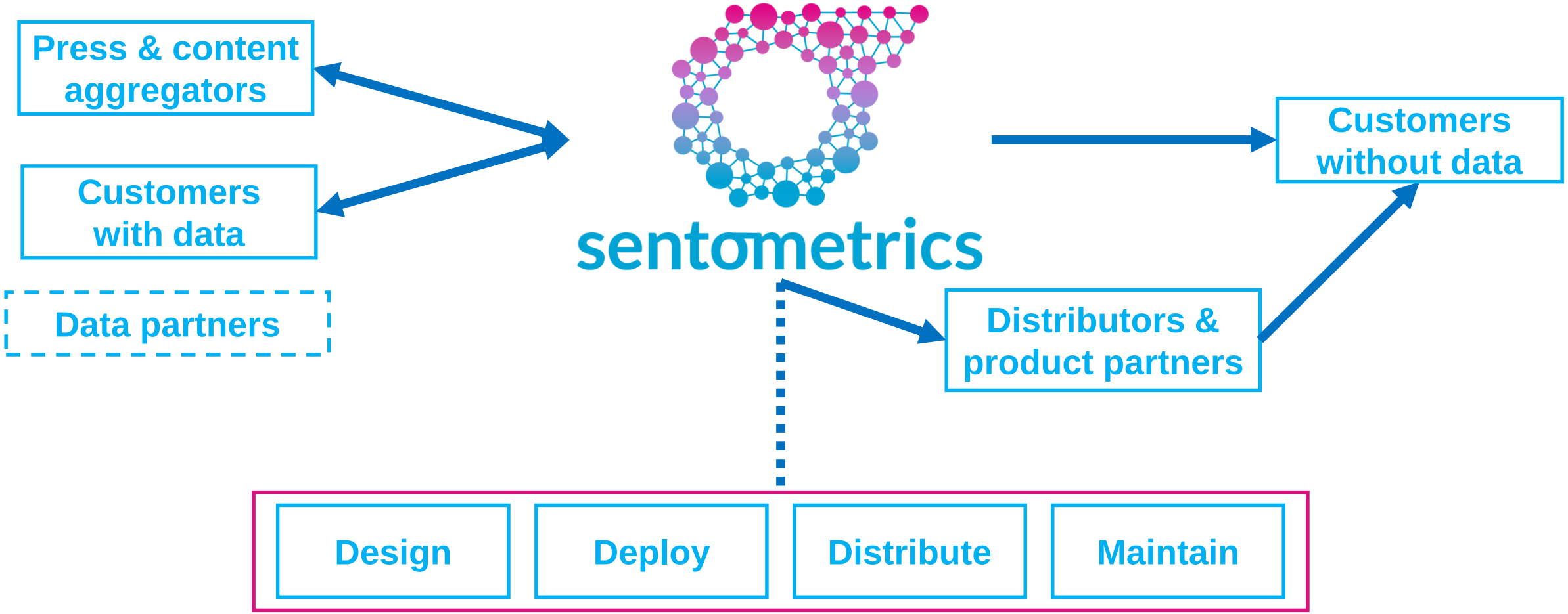
Our platform provides storage, computing power and access



The Belga archive at our fingertips



Business model (3/3)



Our aim is to offer a complete service around objective text-based indices

Macroeconomic indices

Capturing underlying drivers of the economy (e.g. consumer confidence)

Sustainability indices

Capturing environmental, social and governance dimensions about companies, sectors and countries

Geopolitical indices

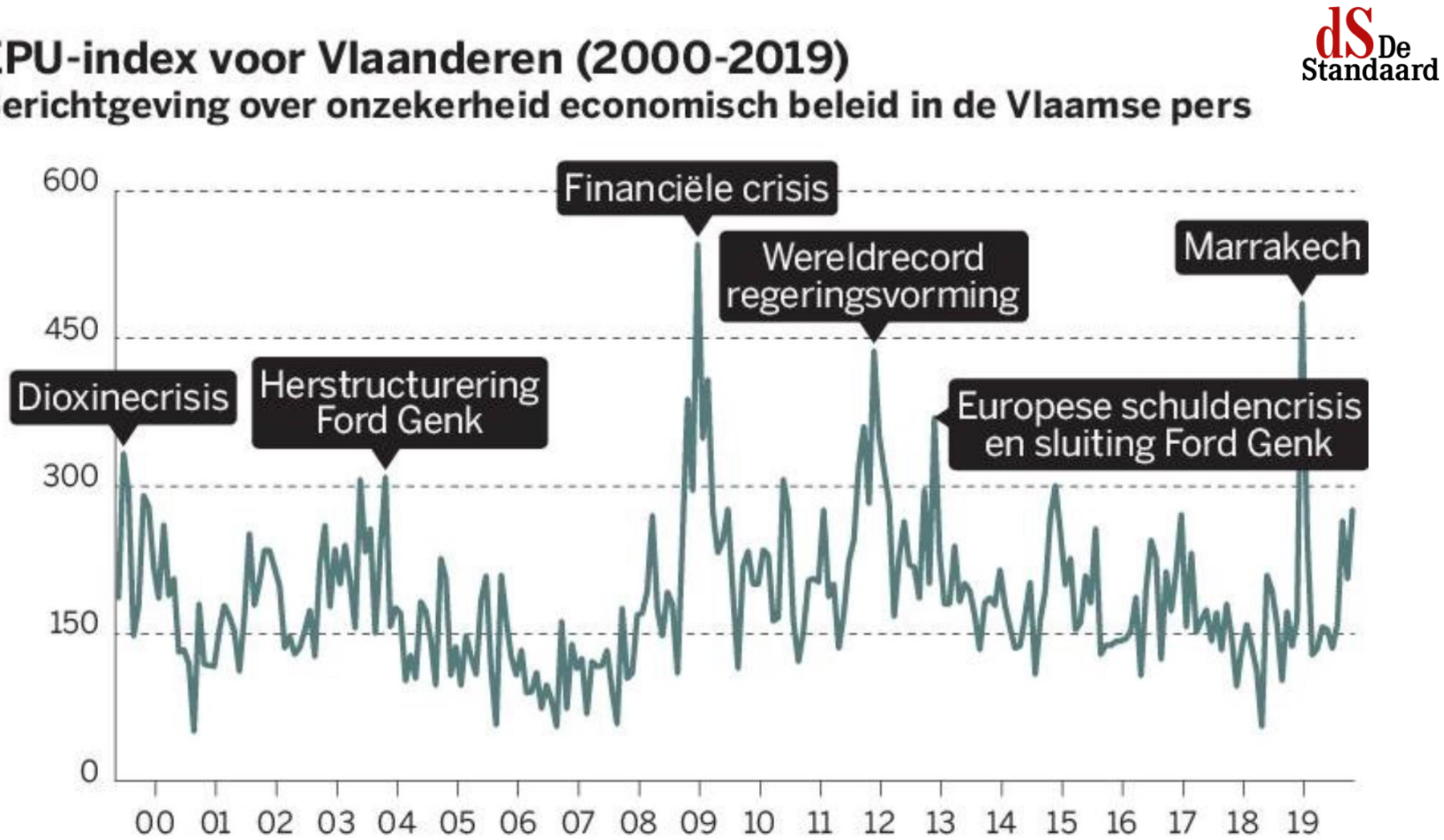
Capturing geopolitical risks (e.g. policy uncertainty)

Financial indices

Capturing asset-specific or market-wide investment dynamics

Showcase: Flemish Economic Policy Uncertainty Index

https://www.policyuncertainty.com/



Stay tuned in 2020

Stay tuned!

Investor

Entrepreneur

Data scientist

Engineer

Researcher

Business

Contact us

Andres Algaba
andres.algaba@vub.be

Samuel Borms
samuel.borms@vub.be

Prof. Kris Boudt
kris.boudt@vub.be

Jeroen Van Pelt
jeroen.van.pelt@vub.be