

Semantic Concept Discovery Over Event Data

Oktie Hassanzadeh, Shari Trewin, Alfio Gliozzo
IBM Research AI



Use Case: Question Analysis

- High-Level Goal:

Given a question (set of entities & concepts), find the most relevant entities & concepts needed to generate a high-quality analysis report.

- Example question:
“What are the consequences of Brexit on London’s financial markets?”

Need to discover:

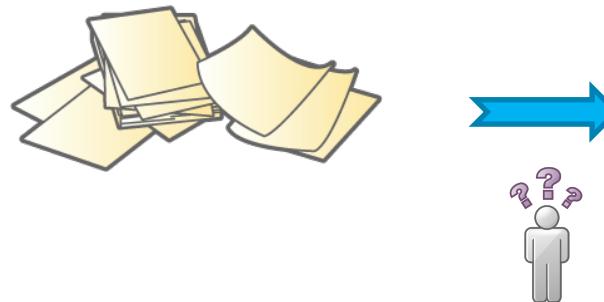
- Key topics (e.g., financial markets, economy, Brexit, Brexit Divorce Bill)
- Key people and organizations involved (e.g., The European Union, decision makers in the EU & UK, people involved in Brexit negotiations)
- All the related events (e.g., Negotiation meetings, Parliamentary elections within the EU, etc.)

Concept Discovery for Deep Analysis

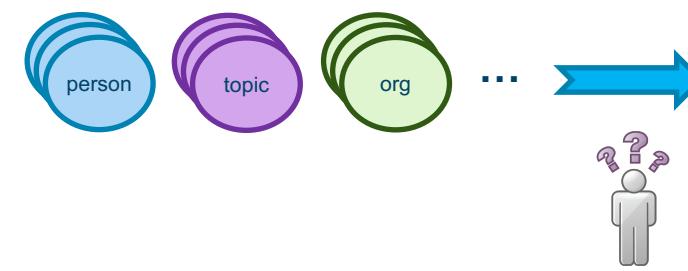
Goal: *Given an analysis question, find the most relevant concepts (topics, entities) needed to make a comprehensive and unbiased analysis*

Classic Solution:

Search over news articles, social media, past reports, other media (text, photo & videos)



Identify relevant concepts mentioned in unstructured sources to establish context



Perform Analysis tasks such as forecasting, hypothesis generation, and scenario analysis, given the context



Problem #1: Mostly manual process, and massive amount of information - so the outcome could be ***biased*** & ***incomplete***

Problem #2: Sources of **search** do not match the available structured data and the corresponding models used for **analytics**

Event Databases

- Structured data representing "events" as reported on the media



GDELT <https://www.gdeltproject.org/>



ICEWS <http://www.lockheedmartin.com/us/products/W-ICEWS.html>



EventRegistry <http://eventregistry.org/>

Event Databases

- Structured data representing "events" as reported on the media



GDELT

ICEWS



Political event databases:
"Event" is an action associated with up to two actors



EventRegistry

Generic event database:
"Event" is a collection of articles on the same topic

Event Databases

- Structured data representing "events" as reported on the media



GDELT 2.0 → {
129+ Million Events
157+ Million Articles Annotated (GKG)
437+ Million Mentions (2+ years)



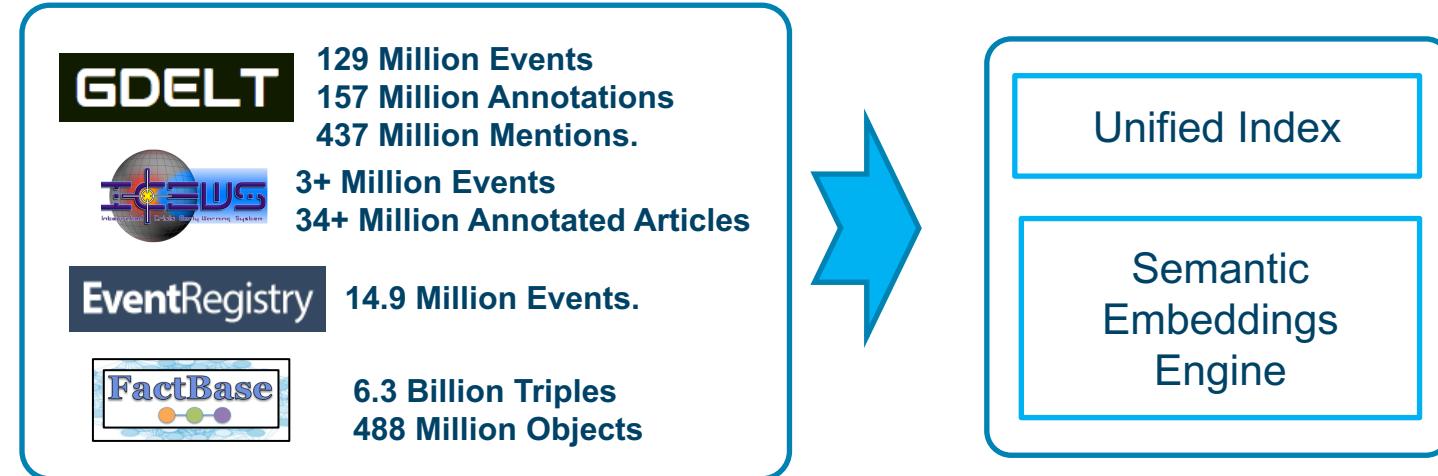
ICEWS → **14.9 Million Events (1995 to 2015)**



EventRegistry → {
5+ Million Events
180+ Million Annotated Articles (2+ years)

Concept Discovery for Deep Analysis: Our Solution

Semantic
Data Curation
Engine:



Analysis Question: *What is the likelihood of violent protest in Caracas, Venezuela?*

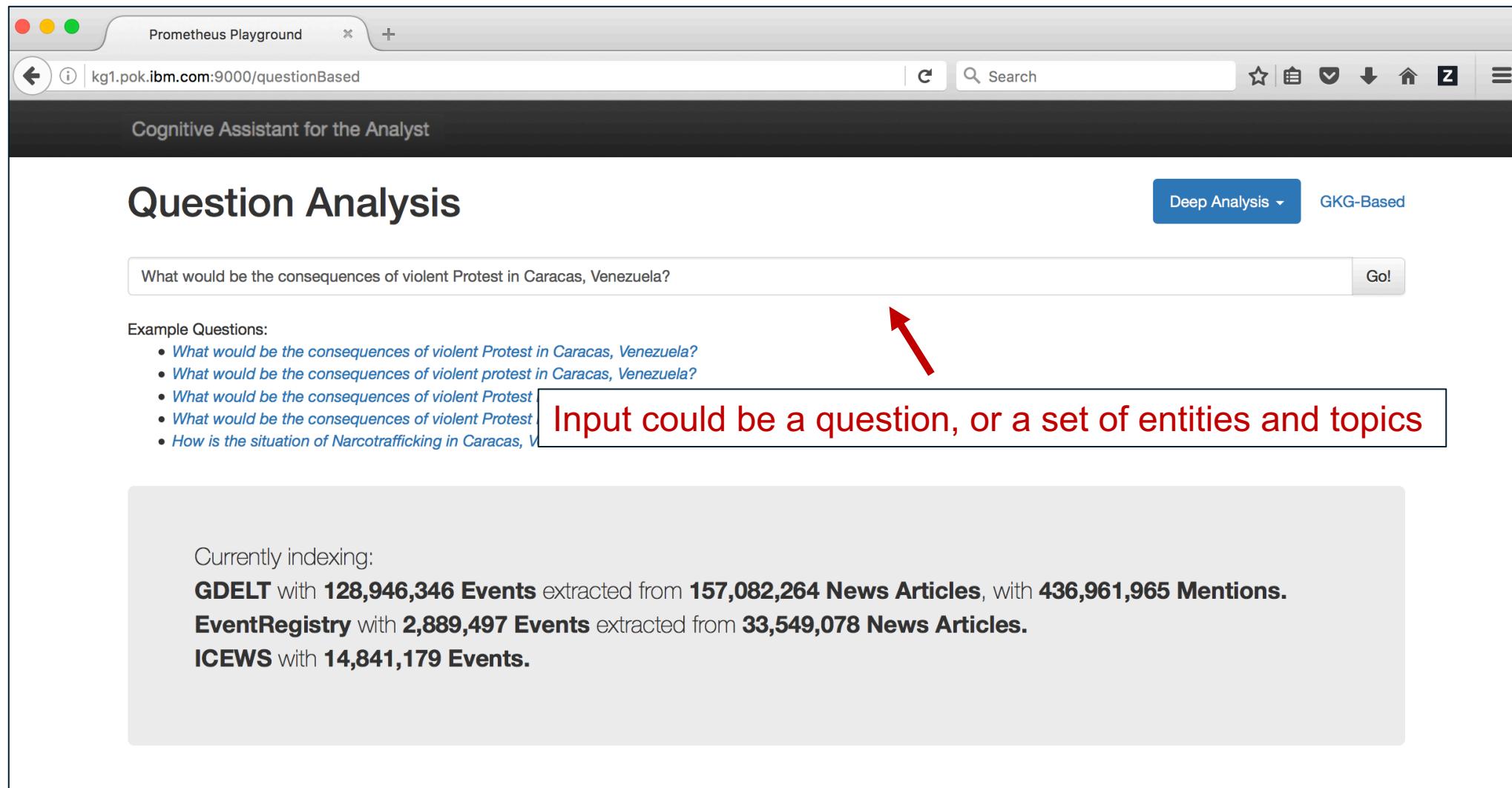
Natural Language Question Understanding: Topic: *Violent Protest* Location: *Caracas, Venezuela*

Semantic Concept Discovery: Query DeepSim API with **Topic#Violent_Protest**
Location#Caracas_Venezuela

DeepSim Analysis
Powered by IBM KRT Socrates & GDELT
DeepSim Context: [Location: Venezuela](#) [Location: Caracas](#) [Theme: PROTEST](#)

Key People	Organizations	Themes
Jesus Torrealba Henrique Capriles Nicolas Maduro Eyanir Chinea David Smilde	Venezuela Supreme Court United Socialist Party Of Venezuela National Electoral Council Venezuelan National Assembly	Tax Ethnicity Venezuelan Tax Ethnicity Venezuelans Tax Political Party Unity Alliance Tax Econ Freetradeagreements Mercosur

Question Analysis: Current Prototype



The screenshot shows a web browser window titled "Prometheus Playground" with the URL "kg1.pok.ibm.com:9000/questionBased". The page header reads "Cognitive Assistant for the Analyst". Below it, a main section is titled "Question Analysis". A text input field contains the question "What would be the consequences of violent Protest in Caracas, Venezuela?". To the right of the input field is a "Go!" button. Above the input field, a dropdown menu shows "Deep Analysis ▾" and "GKG-Based". Below the input field, a list of "Example Questions" is displayed:

- *What would be the consequences of violent Protest in Caracas, Venezuela?*
- *What would be the consequences of violent protest in Caracas, Venezuela?*
- *What would be the consequences of violent Protest*
- *What would be the consequences of violent Protest*
- *How is the situation of Narcotrafficking in Caracas, V*

A red arrow points from the text "Input could be a question, or a set of entities and topics" to the input field. This text is enclosed in a red-bordered box.

Currently indexing:

GDELT with **128,946,346 Events** extracted from **157,082,264 News Articles**, with **436,961,965 Mentions**.

EventRegistry with **2,889,497 Events** extracted from **33,549,078 News Articles**.

ICEWS with **14,841,179 Events**.

Question Analysis: Current Prototype

The screenshot shows a web browser window titled "Prometheus Playground" with the URL "kg1.pok.ibm.com:9000". The page is titled "Cognitive Assistant for the Analyst" and has a "Question Analysis" section. It includes a search bar, a dropdown menu for "Deep Analysis", and another for "GKG-Based". Below the search bar is a text input field with placeholder text "Enter comma-separated list of context keywords" and a "Go!" button. A red arrow points from a callout box containing the text "Input could be a question, or a set of entities and topics" to the input field. To the left of the input field, there is a list of "Example KIQ Contexts" with items like "Caracas, Venezuela", "Protest, Caracas, Venezuela", etc. At the bottom, a box displays indexing statistics: "Currently indexing: GDELT with 128,946,346 Events extracted from 157,082,264 News Articles, with 436,961,965 Mentions.", "EventRegistry with 2,889,497 Events extracted from 33,549,078 News Articles.", and "ICEWS with 14,841,179 Events."

Enter comma-separated list of context keywords

Deep Analysis ▾ GKG-Based

Example KIQ Contexts:

- Caracas, Venezuela
- Protest, Caracas, Venezuela
- Beijing, China, War
- Beijing, China, THEME:ArmedConflict
- Protest, Beijing, China
- Narcotrafficking, Caracas, Venezuela
- Brazil, Dilma Rousseff, Impeachment
- Switzerland, Money_Laundering

Input could be a question, or a set of entities and topics

Currently indexing:
GDELT with 128,946,346 Events extracted from 157,082,264 News Articles, with 436,961,965 Mentions.
EventRegistry with 2,889,497 Events extracted from 33,549,078 News Articles.
ICEWS with 14,841,179 Events.

Question Analysis: Current Prototype

The screenshot shows a web browser window titled "Prometheus Playground" with the URL "kg1.pok.ibm.com:9000". The page is titled "Cognitive Assistant for the Analyst" and has a "Question Analysis" section. It includes a text input field for "context keywords", a "Go!" button, and two analysis modes: "Deep Analysis" (selected) and "GKG-Based". Below the input field, there's a list of example KIQ contexts. A red callout box highlights a statement about ingested event databases. A red arrow points from this statement down to a summary of currently indexed data.

Example KIQ Contexts:

- Caracas, Venezuela
- Protest, Caracas, Venezuela
- Beijing, China, War
- Beijing, China, THEME:ArmedConflict
- Protest, Beijing, China
- Narcotrafficking, Caracas, Venezuela
- Brazil, Dilma Rousseff, Impeachment
- Switzerland, Money_Laundering

We have ingested 3 major event databases
All are structured (tabular) or semi-structured (JSON) data

Currently indexing:

GDELT with **128,946,346 Events** extracted from **157,082,264 News Articles**, with **436,961,965 Mentions**.

EventRegistry with **2,889,497 Events** extracted from **33,549,078 News Articles**.

ICEWS with **14,841,179 Events**.

Question Understanding

Prometheus Playground

kg1.pok.ibm.com:9000/?q=Brazil%2C+Dilma+Rousseff%2C+Impeachment

Cognitive Assistant for the Analyst

Question Analysis

Deep Analysis ▾ GKG-Based

Brazil, Dilma Rousseff, Impeachment Go!

Global Context

Dilma Rousseff, Impeachment, Brazil

DeepSim® Analysis

Powered by IBM Socrates™ & GDELT™

Context is identified through lookups in our FactBase API

Key People	Organizations	Themes
Dilma Rousseff Eduardo Cunha Renan Calheiros Luiz Inacio Lula Waldir Maranhao Aecio Neves Sergio Moro	Brazil Senate Brazilian Democratic Movement Party Rousseff Worker Party Brazil Supreme Court Rousseff Workers Party Petrobras Brazil Congress	Tax Worldlanguages Temer Impeachment Econ Worldcurrencies Brazilian Real Wb 1073 Fiscal Stimulus And Fiscal Rules Tax Econ Freetradeagreements Mercosur Tax Ethnicity Moro Tax Worldlanguages Latin

Question Understanding: Context Details

The screenshot shows a browser window titled "Prometheus Playground" with the URL "kg1.pok.ibm.com:9000/?q=Brazil%2C+Dilma+Rousseff%2C+Impeachment". A modal window is open, titled "Dilma Rousseff", displaying "Facts from IBM FactBase". The modal contains a table of facts:

Links	http://www.wikidata.org/entity/Q40722 https://en.wikipedia.org/wiki/Dilma_Rousseff
Occupation	Politician, Economist
Position Held	President Of Brazil
Educated At	Federal University Of Rio Grande Do Sul
Award Received	Bertha Lutz Prize, Order Of Isabella The Catholic
Father	Pedro Rousseff
Follows	Luiz Inácio Lula Da Silva
Country Of Citizenship	Brazil
Member Of Political Party	Workers' Party
Place Of Birth	Belo Horizonte
Description	President Of Brazil
Label	Dilma Rousseff
Uri	http://www.wikidata.org/entity/Q40722

The background of the browser shows the "Cognitive Assistant for the Analyst" interface, specifically the "Question Analysis" section. A red arrow points from the "Dilma Rousseff" button in the "Global Context" section of the main interface towards the modal window.

DeepSim Analysis – Using Semantic Similarity Anlysis over GDELT GKG

The screenshot shows a browser window titled "Prometheus Playground" with the URL kg1.pok.ibm.com:9000/?q=Brazil%2C+Dilma+Rousseff%2C+Impeachment. The page displays "DeepSim Analysis" powered by IBM Socrates & GDELT. The DeepSim Context is set to Person: Dilma_Rousseff, Location: Brazil, and Theme: IMPEACHMENT. A red callout box points to this context area with the text: "Concepts are mapped from FactBase IDs to GDELT GKG terms". Another red callout box highlights the results section with the text: "Results using embeddings built over 157+ million GDELT GKG records". The results are categorized into three sections: Key People, Organizations, and Themes. The Key People section lists Dilma Rousseff, Eduardo Cunha, Renato, Luiz Inácio Lula da Silva, Waldir, Aécio Neves, Sérgio Moro, Michel Temer, Henrique Meirelles, Jaques Wagner, Antonio Anastasia, Leonardo Picciani, Christopher Garman, Eduardo Cardozo, and Justice Ricardo Lewandowski. The Organizations section lists Petrobras, Brazil Congress, Brazilian Social Democracy Party, Brazil Chamber, Democratic Movement Party, Brazilian Social Democratic Party, Brazilian Democratic Movement, Temer Brazilian Democratic Movement Party, University Of Brasilia, and Brazilian Republican Party. The Themes section lists Tax Worldlanguages Temer, Impeachment, Tax Ethnicity Moro, Tax Worldlanguages Latin, Tax Ethnicity Venezuelans, Tax Political Party Workers Party, Tax Fncact Expresident, Scandal, Tax Fncact Chief Of State, Slfid Dictatorship, Tax Fncact Leftist, and Tax Terror Group African National Congress.

Prometheus Playground

kg1.pok.ibm.com:9000/?q=Brazil%2C+Dilma+Rousseff%2C+Impeachment

Search

Prometheus Playground

DeepSim Analysis

Powered by IBM Socrates¹ & GDELT

DeepSim Context: Person: Dilma_Rousseff Location: Brazil Theme: IMPEACHMENT

Concepts are mapped from FactBase IDs to GDELT GKG terms

Key People

- Dilma Rousseff
- Eduardo Cunha
- Renato
- Luiz Inácio Lula da Silva
- Waldir
- Aécio Neves
- Sérgio Moro
- Michel Temer
- Henrique Meirelles
- Jaques Wagner
- Antonio Anastasia
- Leonardo Picciani
- Christopher Garman
- Eduardo Cardozo
- Justice Ricardo Lewandowski

Organizations

- Petrobras
- Brazil Congress
- Brazilian Social Democracy Party
- Brazil Chamber
- Democratic Movement Party
- Brazilian Social Democratic Party
- Brazilian Democratic Movement
- Temer Brazilian Democratic Movement Party
- University Of Brasilia
- Brazilian Republican Party

Themes

- Tax Worldlanguages Temer
- Impeachment
- Tax Ethnicity Moro
- Tax Worldlanguages Latin
- Tax Ethnicity Venezuelans
- Tax Political Party Workers Party
- Tax Fncact Expresident
- Scandal
- Tax Fncact Chief Of State
- Slfid Dictatorship
- Tax Fncact Leftist
- Tax Terror Group African National Congress

Results using embeddings built over 157+ million GDELT GKG records

Swagger API behind DeepSim Analysis

Index-based Analysis (co-occurrence)

The screenshot shows a web browser window titled "Prometheus Playground" with the URL "kg1.pok.ibm.com:9000/?q=Brazil%2C+Dilma+Rousseff%2C+Impeachment". The page is titled "Cognitive Assistant for the Analyst" and features a section titled "Index-Based Analysis & Events" powered by EventRegistry¹. The "EventRegistry Context" is set to "Dilma_Rousseff Impeachment Brazil".

The page displays three main sections: "Topics", "Key Players", and "Locations".

- Topics:**
 - 100 Impeachment
 - 92 Government
 - 62 Petrobras
 - 64 Political party
 - 56 Democracy
 - 52 Economics
 - 47 Brazilian Social Democracy Party
 - 47 Brazilian Democratic Movement Party
 - 39 Luiz Inácio Lula da Silva
- Key Players:**
 - 100 Dilma Rousseff
 - 34 Fernando Collor de Mello
 - 22 Michel Temer
 - 7 Marina Silva
 - 7 Itamar Franco
 - 6 Escândalo de las mensualidades
 - 5 Tancredo Neves
 - 5 Aécio Neves
 - 5 Aloizio Mercadante
- Locations:**
 - 100 Brazil
 - 62 São Paulo (state)
 - 59 Rio de Janeiro
 - 31 United States
 - 28 Earth
 - 28 Minas Gerais
 - 20 Rio Grande do Sul
 - 20 Belo Horizonte
 - 20 Brazilian military government

A red box highlights the central text: "Results using queries over our event databases (EventRegistry for this view)".

114 Events Found

- Teen libertarian is face of Brazil's young free-market right
- UPDATE 1-Brazil's Petrobras will not sell distribution arm -board member
- In Peru, Acio says (Portuguese)
- TV Cultura virou canal tucano, dizem leitores sobre artigo de ex-presidente (Portuguese)
- Brazil: Ruling Party Treasurer Joao Vaccari Will be Tried over Petrobras Scandal

Swagger API behind Index-Based Analysis



concept-discovery

This is the concept discovery API. It provides a set of functions for expanding a query with related people, organizations and themes.

extract

Show/Hide | List Operations | Expand Operations

POST

/extract/concepts

This api extracts concepts (person, organization, themese, etc.) from a given text string

query

Show/Hide | List Operations | Expand Operations

POST

/query/expand

This api uses GKG data to find key players (people and organizations) for a given country, set of seed people, and set of GDELT themes (topics)

Example input: { "country": "Peru", "players": ["Nicolas Maduro"], "topics": ["IMMIGRATION", "PROTEST"], "candidateRankingMethods": ["similarity", "facetcount"] }

Response Class (Status 200)

successful operation

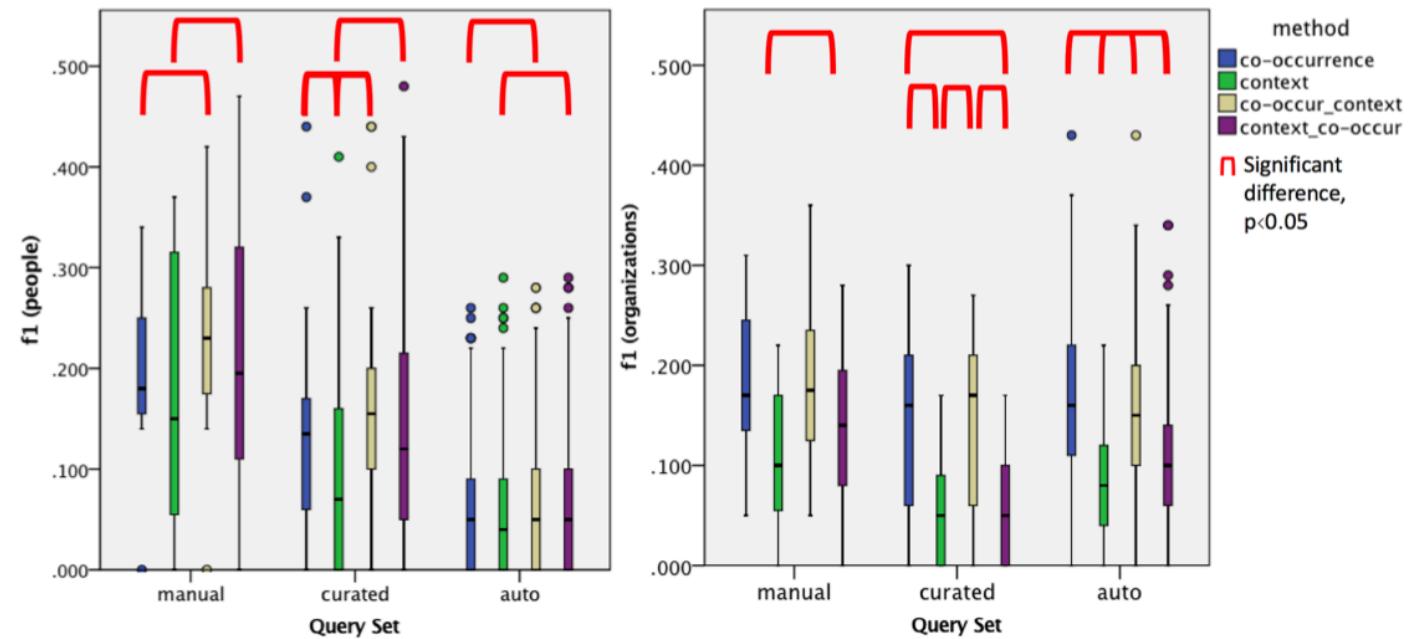
Model | Example Value

```
{  
  "field": "string",  
  "method": "string",  
  "candidates": [  
    "string"  
  ]  
}
```

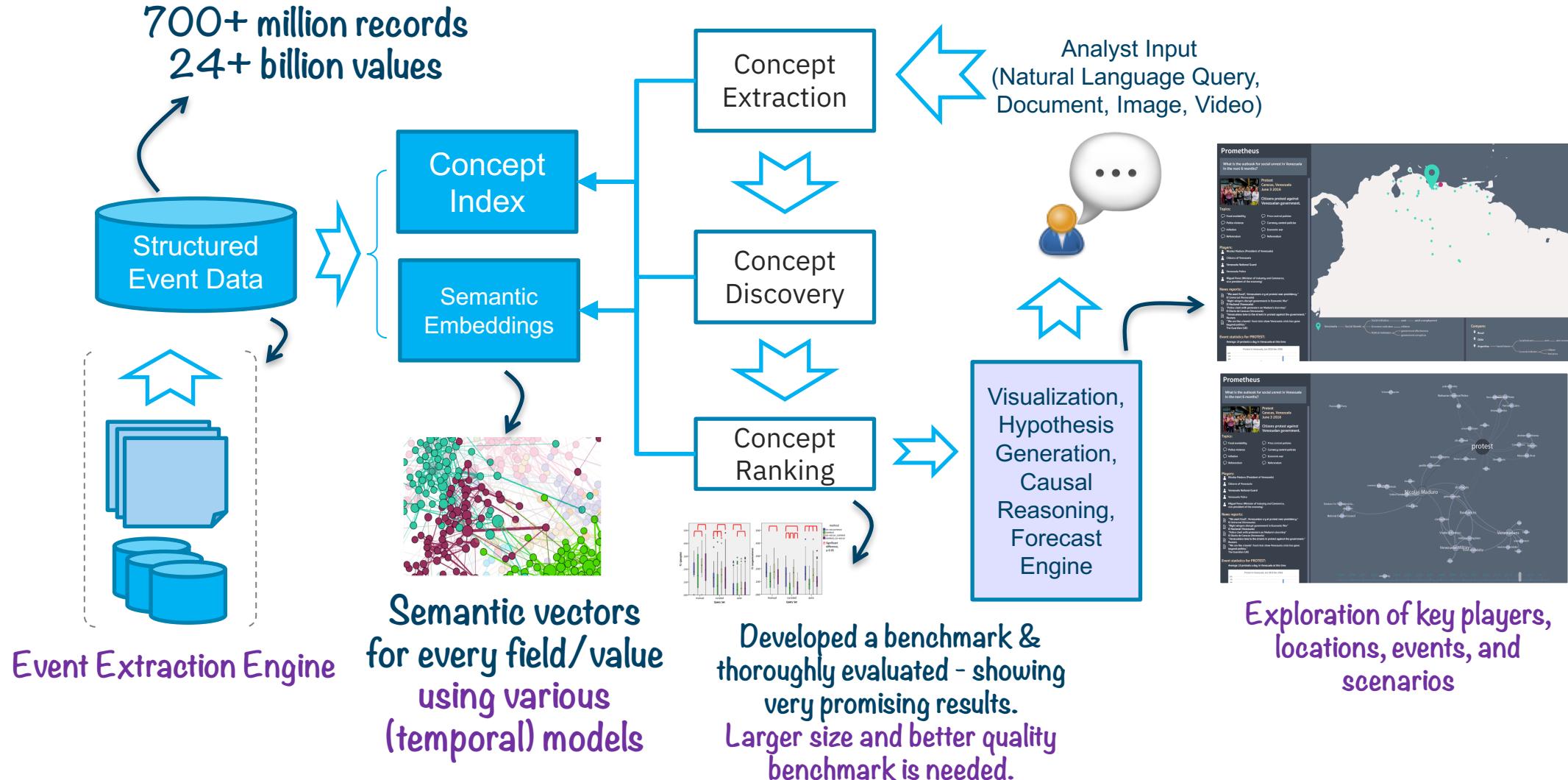
Evaluation

- Benchmark created using reports written by human experts
 - Human Rights Watch organization & Wikipedia articles on events and people
- We measured the ability of different algorithms to find the concepts mentioned in the original reports
- A combination approach works best in most cases

	person				organization			
	co-occur	context	co-occur context	context co-occur	co-occur	context	co-occur context	context co-occur
MAP	0.233	0.199	0.251	0.233	0.179	0.143	0.184	0.189
F1	0.192	0.174	0.228	0.213	0.178	0.107	0.183	0.141
Pr.	0.133	0.121	0.158	0.149	0.117	0.066	0.119	0.089
Re.	0.372	0.328	0.437	0.388	0.436	0.304	0.459	0.374



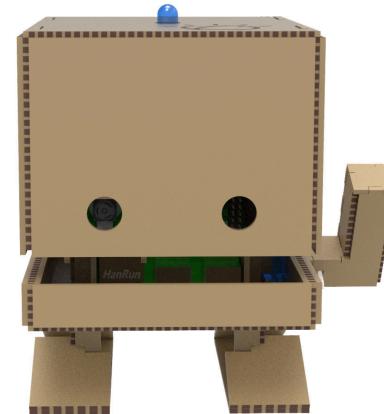
Conclusion & Future Work



We are hiring!

and we very much welcome academic collaborations

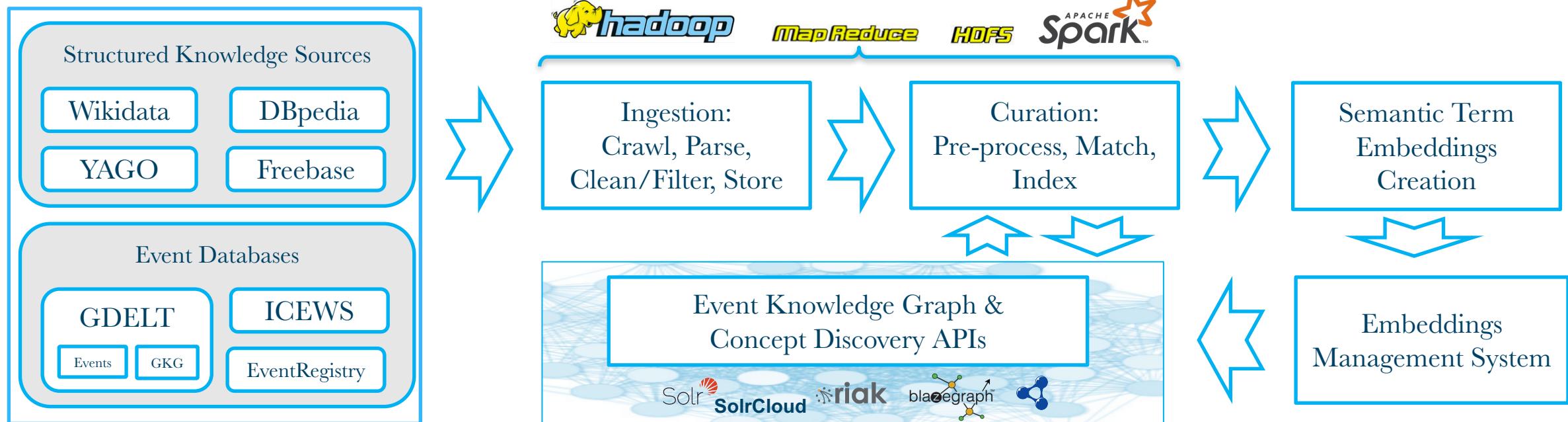
Come to our booth & get in touch.



(You may take one of 🤝 these home!)

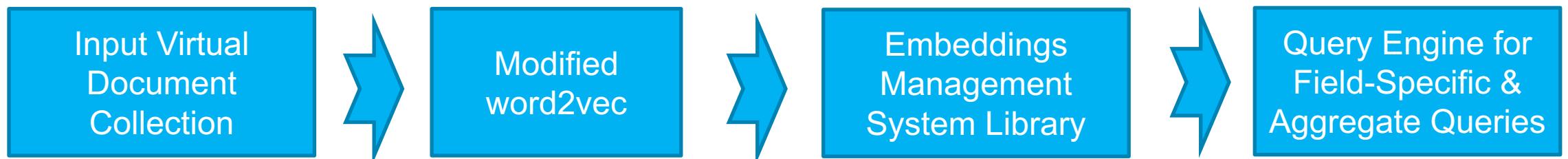
BACKUP

System Architecture



DeepSim (context) Analysis Details – Model Construction & Query Engine

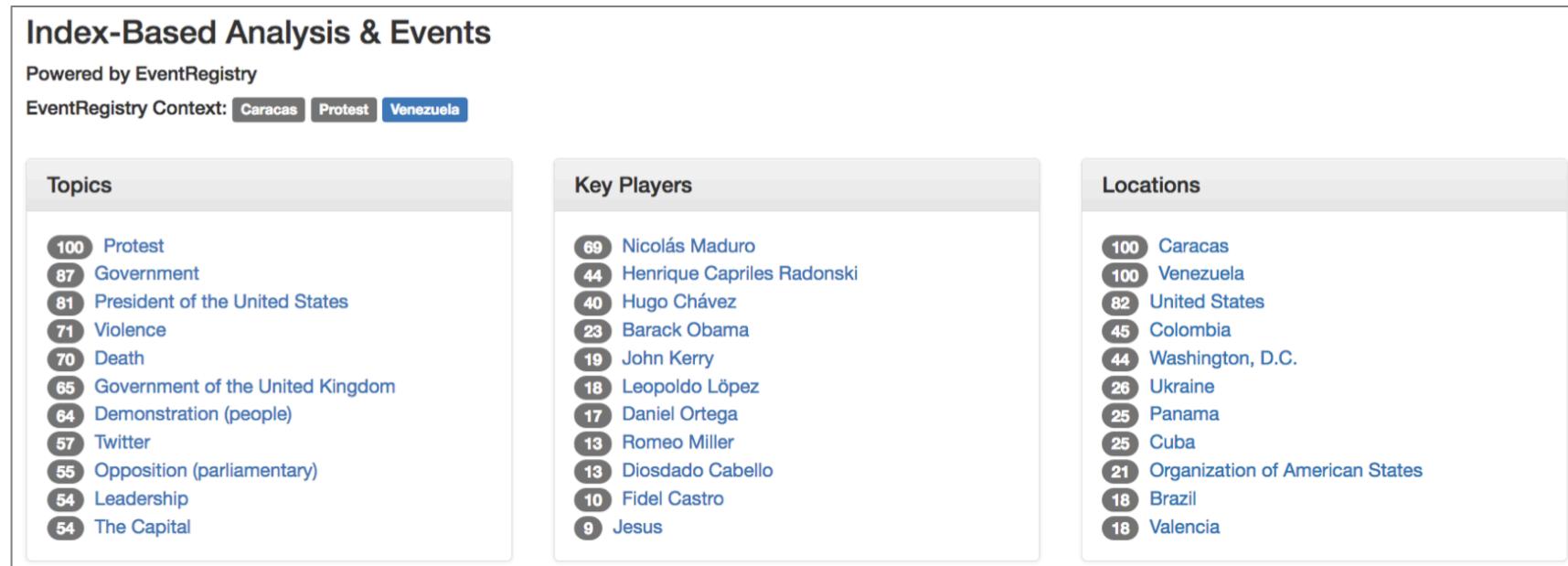
- Step 2: Model Construction & Query Engine



- Word2vec modifications
 - Fixing context window size, rotating window so column order does not affect the outcome
- Embeddings Management System
 - Super fast in-memory approximate nearest neighbor library & API
- Query Engine for DeepSim
 - Prefix-based query for field-specific retrieval (e.g., retrieve "person"s similar to "location")
 - AND query over input terms (e.g., retrieve "person" similar to "location X", "persons Y & Z")

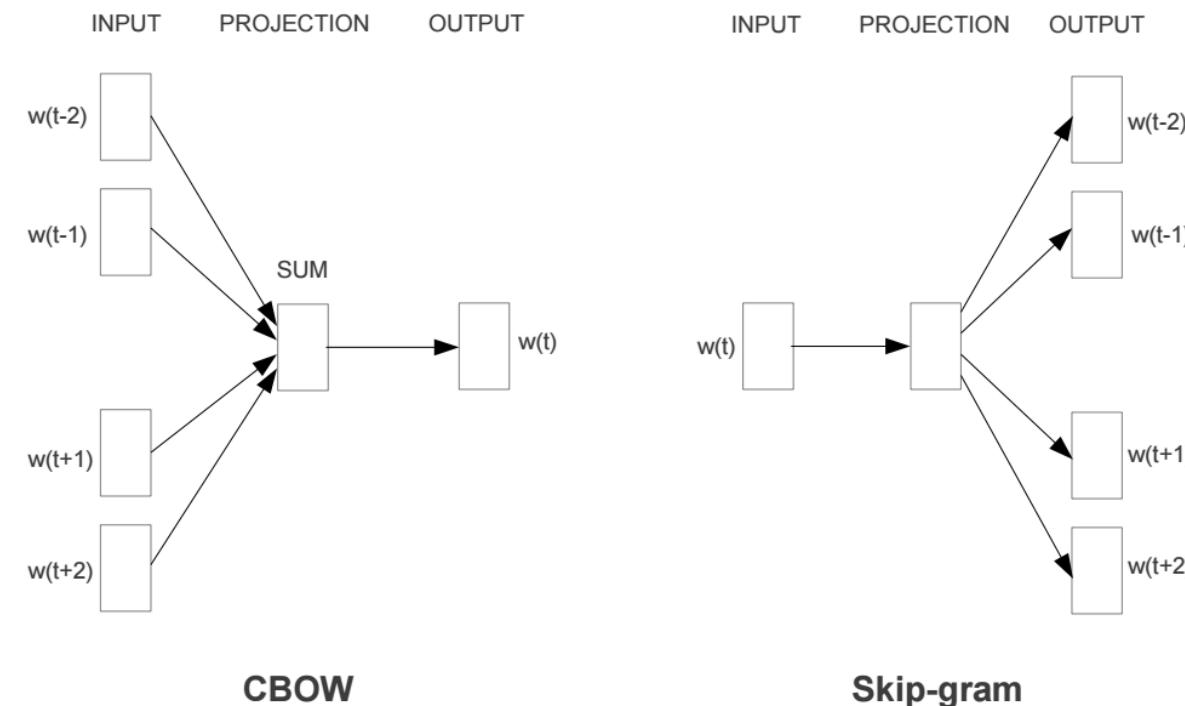
Concept Ranking: Index-Based Method (co-occurrence)

- Index-based Method: Measuring Co-occurrence
 - Formulate a search query using the concepts extracted from the input question
 - Count the concept annotations for every record in the input
 - Return the most frequent annotations of various types (persons, organizations, themes)
 - Use percentage of co-occurrence as the relevance score



Concept Ranking: Deep Similarity Method (context)

- Using word embeddings to capture the semantic similarity of terms (field values)
- Embeddings in NLP: vectors representing the semantic context of each word
 - Similar terms have similar vectors (as per e.g. Cosine similarity between the vectors)
- Method used: modified **Skip-gram word2vec model** [Mikolov et al., 2013]
 - An efficient, shallow Neural Network model
 - CBOW architecture predicts the current word based on the context
 - Skip-gram predicts surrounding words given the current word



DeepSim (context) Analysis Details – Virtual Document Generation

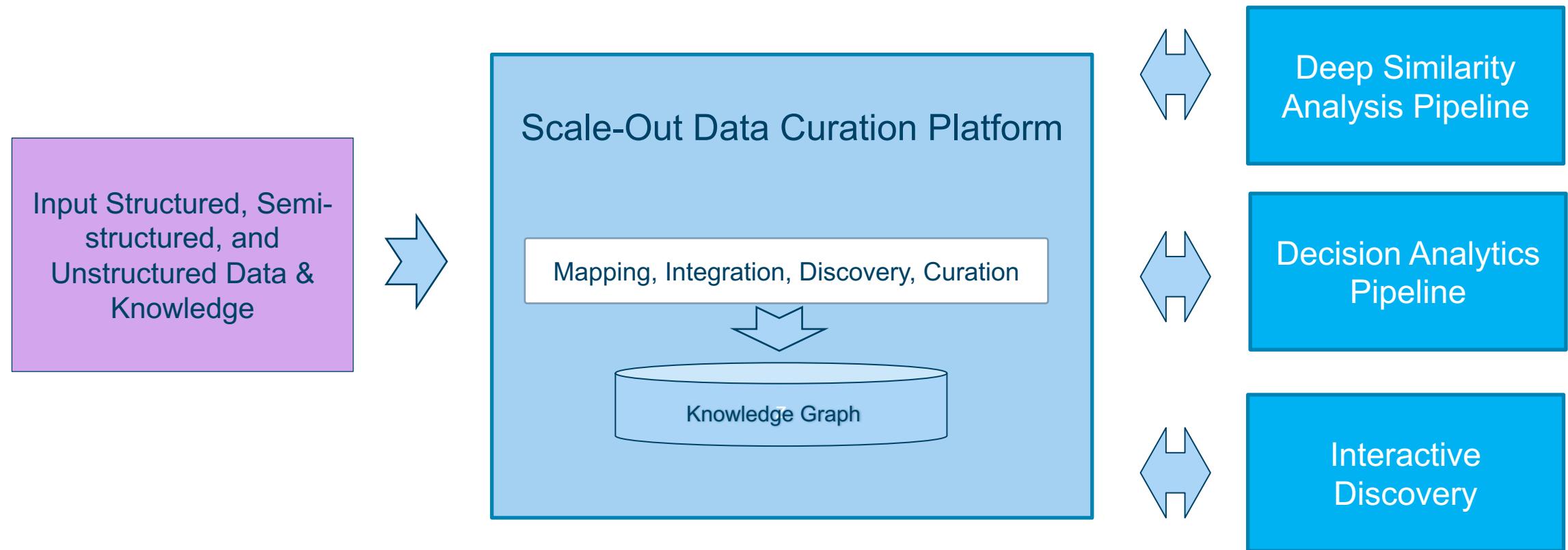
- Step 1: Virtual Document (Context) Generation
 - A term is a concatenation of a field name (column header) and a field value (cell content)
 - Each record in the input data turns into a context
 - We may need transformations such as data binning / bucketing for numerical fields

date	Location_countryCode	Location_fullName	persons	names	themes	...
2017-01-01T08#00#00Z	AS	Goulburn, New South Wales, Australia	Barnaby Joyce, Alastair Starritt, ...	Water Minister Barnaby, Southern Valley, ...	GENERAL_GOVERNMENT, EPU_POLICY_GOVERNMENT, ...	



```
date#2017-01-01T08#00#00Z tone_selfGroupReferenceDensity#0 tone_tone#-2 tone_activity#2
tone_positiveScore#0 tone_negativeScore#0 tone_polarity#0 person#Barnaby_Joyce
person#Alastair_Starritt person#Alastair_Starritt sourceCommonName#stockandland_com_au
organization#Deniboota_Landholders_Association organization#Deniboota_Landholders_Association
name#Deniboota_Landholders_Association name#Alastair_Starritt
name#Deniboota_Landholders_Association name#Murray-Darling_Basin name#Northern_Basin
name#Basin_Plan name#Southern_Basin name#Goulburn_Murray name#Goulburn_Valley name#Basin_Plan
name#Latrobe_Valley name#Water_Minister_Barnaby_Joyce sourceCollectionIdentifier#WEB
theme#GENERAL_GOVERNMENT theme#GENERAL_GOVERNMENT theme#GENERAL_GOVERNMENT
theme#EPU_POLICY_GOVERNMENT ... location_latitudeLongitude#149_721--34_7515 location_countryCode#AS
location_fullName#Goulburn_New_South_Wales_Australia location_type#WORLDCITY
location_featureID#-1576139 location_ADMIN1Code#AS02 location_ADMIN2Code#154641
location_latitudeLongitude#149_721--34_7515 location_countryCode#AS location_fullName#Goulburn
_New_South_Wales_Australia location_type#WORLDCITY location_featureID#-1576139 ...
```

Towards a Generic Pipeline for Semantic Similarity Analysis



The Need for an Event Extraction Engine

- **Shortcomings of existing event data**
 - No associated text articles. Only URLs are available.
 - No associated meta-data (GKG or GKG-like data) over **historic** articles -> we only have GKG data limited to the past two years.
 - Limited definition of "event" in each source. E.g., GDELT & ICEWS are CAMEO coded (and so only cover a specific type of political events) and EventRegistry defines an event as a collection of articles so we do not know the kind of actions and actors in each event.
 - Noise → both random noise and *systematic noise* (a result of rule-based extraction)
 - EventRegistry
 - encoding issues (tofu characters in labels)
 - inaccurate concept annotations
 - GDELT & ICEWS
 - wrong annotations, missing annotations, basically all the problems a rule-based system could have
- **What we need:** a comprehensive, accurate, and up-to-date event database with annotations similar to EventRegistry and GDELT GKG, historic coverage similar to ICEWS, event coding similar to GSR