

Semantic Data Enrichment: from Interactive Exploration to Scalable Deployment

Roberto Avogadro *, Flavio De Paoli ^, Dumitru Roman *, Matteo Palmonari ^

Part II: Semantic Data Enrichment, Applications and Requirements



*This work presented in this presentation has received funding from the European Union's Horizon 2020 research and innovation program under grant agreements No 732590 - **EW-Shopp** - and No 732003 – **euBusinessGraph** - and from the European Union's Horizon Europe research and innovation program under grant agreements No 101070284 - **enRichMyData**.*



Outline

- Part II: Semantic Data Enrichment, Applications and Requirements
 - Semantics and KGs for data enrichment
 - The *Link & Extend* enrichment paradigm
 - Interactive exploration and scalability
- Part III: Selected State-of-the-art
 - Data preparation solutions
 - The broader context of data preparation solutions
 - Scalable data pipelines
 - A quick introduction to solutions for scalability
 - Tabular data annotation
 - From heuristic techniques to generative LLMs
- Part IV: Semantic Data Enrichment in Practice with Tools
 - Service-based approach
 - Data model for interoperability
 - Service model for composability
 - Interactive definition of pipelines
 - Exploration with graphical UI
 - Pipeline definition with programmatic UI
 - Pipeline execution at scale
 - Execution with workflow managers (Argo & TAO)
 - Live demos
- Part V: Conclusions and Discussion
 - Wrap-up and take-home messages
 - Discussion



Part II: Semantic Data Enrichment, Applications and Requirements

1) Semantics and KGs for data enrichment

“What is data enrichment and what is the impact of Knowledge Graphs and other semantic technologies?”



Data for AI Applications

- Tabular data

- Proprietary data (business / science)
 - Spreadsheets: ~ 400 million worldwide users (50 to 80% of companies use spreadsheets)
 - RDBS and other NoSQL databases (incl. JSON)
- Web data (encyclopaedic knowledge)
 - Common Crawl*: ~230 million tables in 2016 [Lehmberg et al. 2016]
 - Wikipedia: ~ 3.23 million tables in 2019 [Fetahu et al. 2019]

- Textual data

- Mails
- Social media
- News
- Laws
- Science
- Web content
- User manuals
- ...

- Images

- ...
- Audio
 - ...
- Video
 - ...



Data Enrichment vs Semantics

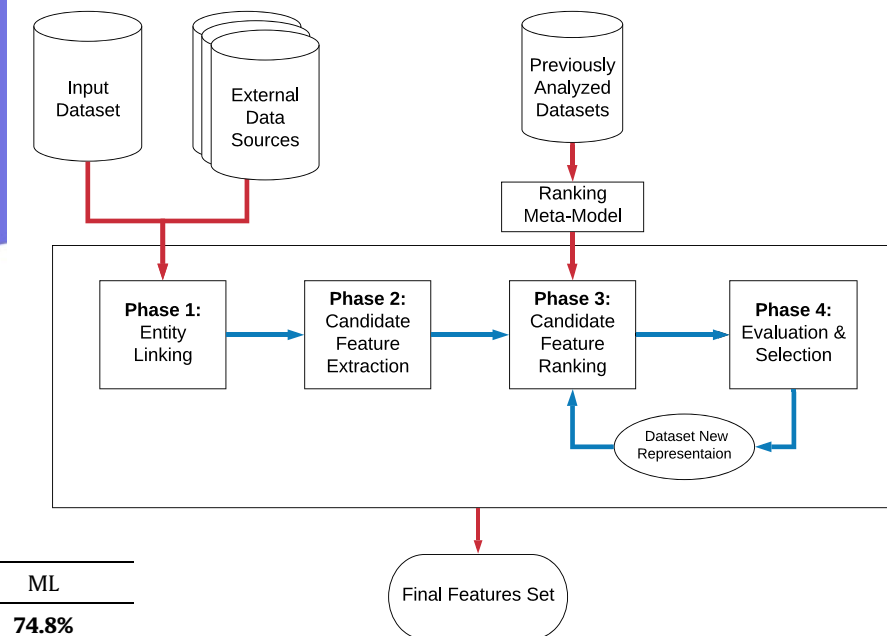
- Part of data preparation
 - Goal: add context to an input data set
 - Related to data cleaning (blurred boundaries)
- Structure and terminology
 - Input: a data set D
 - Output: a data set D' with additional data
 - fetched from external data sources
 - inferred / computed
 - ... also “data augmentation” (tables) and “data annotation” (tables and documents)
- The role of semantics
 - Integration-related tasks and annotations
 - Knowledge graphs as enrichment sources
 - Inference based on semantics, e.g., classification

Categories	Available features
Data discovery	Locate missing values (nulls)
	Locate outliers
	Search by pattern
	Sort data
Data validation	Compare values (selection and join)
	Check data range
	Check permitted characters
	Check column uniqueness
	Find type-mismatched data
	Find data-mismatched datatypes
Data structuring	Change column data type
	Delete column
	Detect & change encoding
	Pivot / unpivot
	Rename column
	Split column
	Transform by example [13]
Data enrichment	Assign semantic data type
	Calculate column using expressions
	Discover & merge external data
	Duplicate column
	Generate primary key column
	Join & union
	Merge columns
	Normalize numeric values
Data filtering	Delete/keep filtered rows
	Delete empty and invalid rows
	Extract value parts
	Filter with regular expressions
Data cleaning	Change date & time format
	Change letter case
	Change number format
	Deduplicate data
	Delete by pattern
	Edit & replace cell data
	Fill empty cells
	Remove extra whitespace
	Remove diacritics
	Standardize strings by pattern
	Standardize values in clusters

Impact of Enriched Features on ML

[Harari & Katz 2022a]

Name	Initial AUC	IG	ML
189 Baseball	0.70	-0.1%	74.8%
9 Autos	0.86	0.6%	0.4%
Aaup	0.77	-0.4%	9.2%
Adult	0.75	-0.1%	0.1%
Anime	0.57	0.0%	-0.1%
Autos	0.50	0.0%	0.0%
Books	0.50	9.4%	39.4%
Conference Attendance	0.50	0.0%	0.0%
WDI	0.69	3.2%	26.3%
Country Codes	0.84	0.5%	12.8%
Movies	0.50	34.9%	69.9%
Netflix Titles	0.50	0.0%	3.0%
Reviewer	0.66	-0.6%	-0.5%
Rmftsa Ctoarrivals	0.89	10.5%	26.1%
S&P 500 Companies	0.66	18.1%	15.9%
Shanghai	0.92	2.6%	4.2%
Waterbody	0.62	0.0%	0.0%
Zoo	0.94	-1.6%	16.7%
Average (Median)		4.3% (0.0%)	16.5% (6.4%)



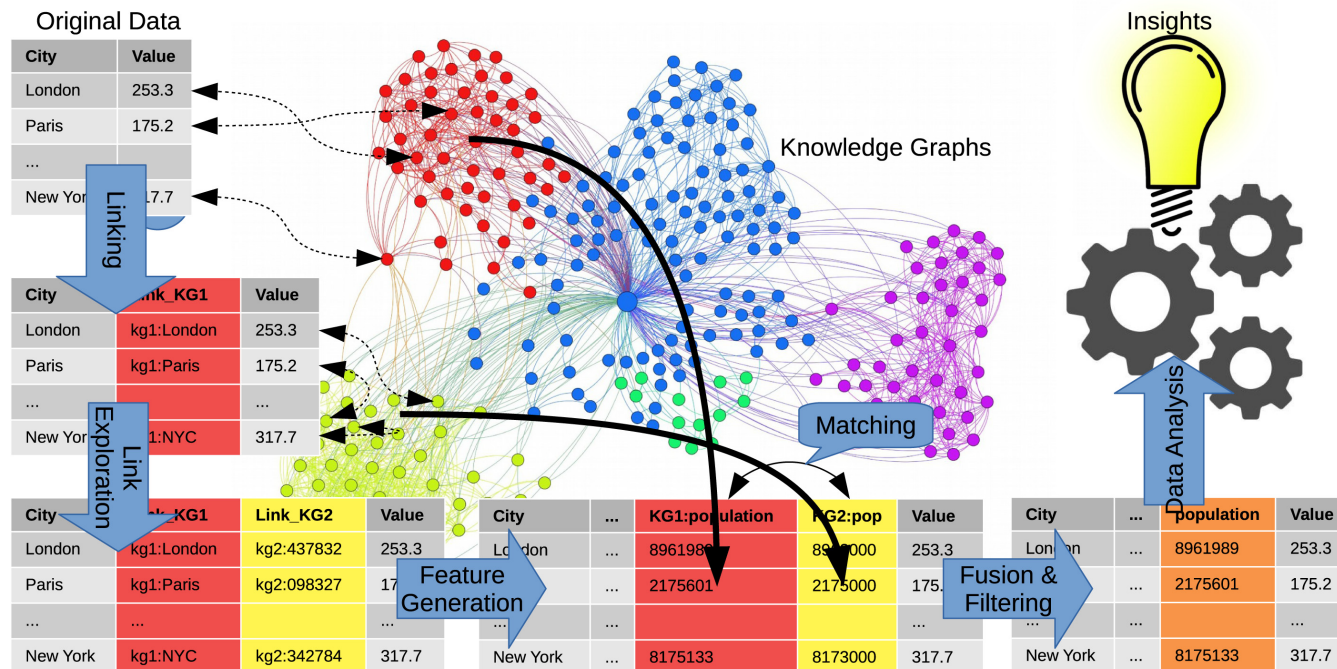
Similar more complex approach
with features from Wikipedia
[Harari & Katz 2022b]

Impact of wrong links (!)



Harari, A., & Katz, G. (2022). Automatic features generation and selection from external sources: a DBpedia use case. Information Sciences, 582, 398-414.

The Python kgextension Package



Data extraction from the KG







Direct semantic data enrichment

Fig. 1. Data analysis pipeline using background knowledge from knowledge graphs

Bucher, T. C., Jiang, X., Meyer, O., Waitz, S., Hertling, S., & Paulheim, H. (2021). scikit-learn Pipelines Meet Knowledge Graphs: The Python kgextension Package. ESWC Satellite Events: Revised Selected Papers 18



Examples from the Industry

Domain	Value	Enrichment Data Sources	Data
eCommerce	Predict impact of events on customer searches 	<u>Events</u> , weather	Tabular
Retail	Workforce/budget optimization 	<u>Events</u> , weather	Tabular
CRM	Workforce optimization 	<u>Events</u> , <u>weather</u>	Tabular
IOT	Customer flow analysis 	Events, weather	Tabular
Digital Marketing	Ad impression prediction for campaign optimization 	Weather	Tabular
Digital Marketing	Ad impression prediction for campaign optimization 	Events	Tabular
Manufacturing	AI-based analytics on welding robot data (tables and user manuals)	Proprietary ~KG	Tabular, Texts
Manufacturing	Troubleshooting and repair based on service manuals, records, log data	Proprietary ~KG	Tabular, Texts
Open data	Construction and maintenance of a European dataset of organizations in procurement from tenders	Proprietary ~KG, Wikidata, Crunch Base	Tabular, Texts
Observatory on AI	Construction and maintenance of a KG to track AI-related innovations from different data sources	Crunch Base, WikiData	Tabular, Texts
Business analysis	Cost-effective enrichment of client datasets' with proprietary company KG	Proprietary KG	Tabular



Tabular data

“Traditional” ML
&
Data Analytics

- Weather-based optimization in digital marketing [[ISWC19](#), [Tech. and Appl. for BDV22](#)]

A Semantic Data Enrichment Example

Digital Marketing at JOT

Google search results for "free tour in Prague". The search bar shows the query and navigation icons. Below the search bar, there are tabs for "Todo", "Maps", "Imágenes", "Shopping", "Noticias", "Más", and "Herramientas". The results show approximately 19,000,000 results in 0.62 seconds. The first result is an advertisement for "Free Tour In Prague" from "http://www.free-tour-prague.eu/". The ad text includes "Every Day at 10AM & 3PM - See all Sights in 2,5 Hours" and "Join the YouTube team Real Prague Guides for a free tour you will never forget!". A red box highlights the ad text, and a red arrow points to it. Below the ad, there are more search results from "Civatis" and "Free Tour In Prague". At the bottom, there is a section "Buscar resultados en" with links to "Civatis", "Tripadvisor", and "FREE WALKING TOUR PRAGUE EU".

Performance Data

clicks	Location	Conversion rate
Keyword	impressions	Date
	Category	Ad Text



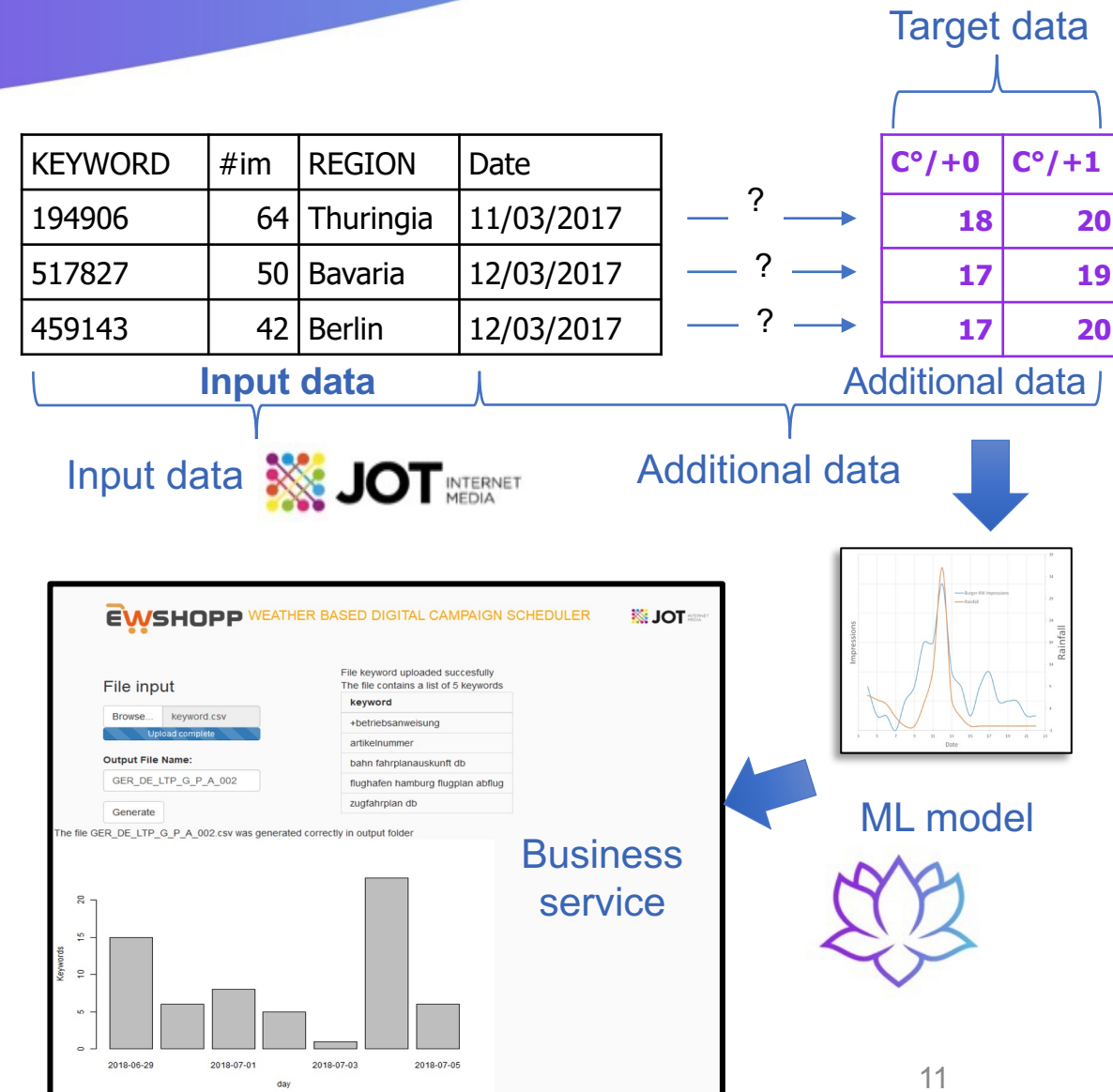
Weather-based Campaign Scheduler

New services for campaign optimization

- Main service: **weather-based campaign scheduler**
 - Predict the best dates to launch the campaign with weather-sensitive keywords
 - in the upcoming week
 - for each region
- + additional services
- Why do we focus on data enrichment?
 - 80% time in data analysis project is spent for cleaning and enriching the data*

Cutrona, V., De Paoli, F., Košmerlj, A., Nikolov, N., Palmonari, M., Perales, F., & Roman, D. (2019). Semantically-enabled optimization of digital marketing campaigns. ISWC

*[Worldwide Semiannual Big Data and Analytics Spending Guide](#) from International Data Corporation (IDC)



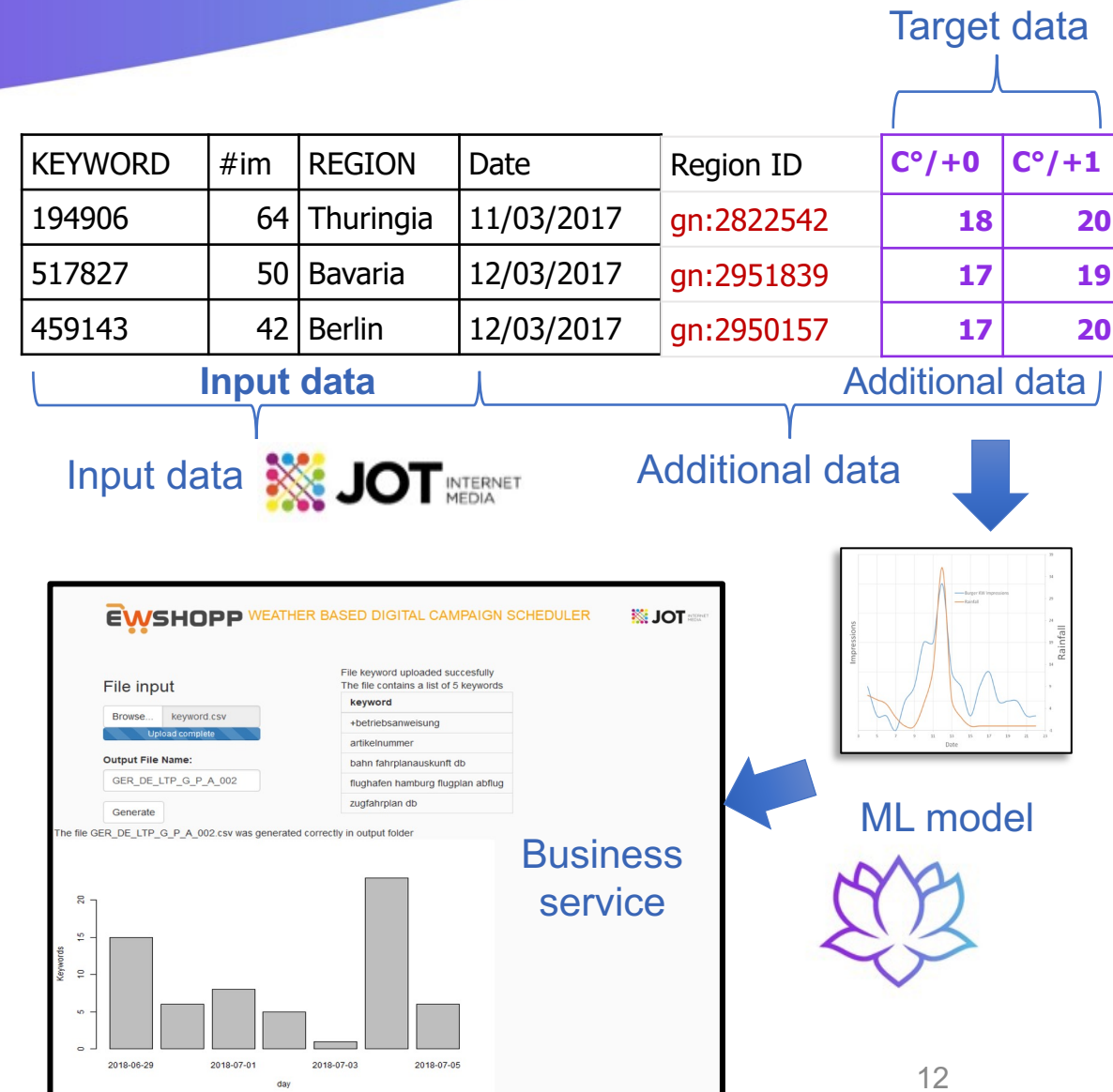
Weather-based Campaign Scheduler

New services for campaign optimization

- Main service: **weather-based campaign scheduler**
 - Predict the best dates to launch the campaign with weather-sensitive keywords
 - in the upcoming week
 - for each region
- + additional services
- Why do we focus on data enrichment?
 - 80% time in data analysis project is spent for cleaning and enriching the data*

Cutrona, V., De Paoli, F., Košmerlj, A., Nikolov, N., Palmonari, M., Perales, F., & Roman, D. (2019). Semantically-enabled optimization of digital marketing campaigns. ISWC

*[Worldwide Semiannual Big Data and Analytics Spending Guide](#) from International Data Corporation (IDC)



Digital Marketing Data Enrichment

Weather Data Source



city: 2950157
- date: 2017-03- 12
2t: 17
- date: 2017-03-13
2t: 20

regionID (GeoNames)

date (ISO 8601)

mismatches



KEYWORD	#im	REGION	Date
194906	64	Thuringia	11/03/2017
517827	50	Bavaria	12/03/2017
459143	42	Berlin	12/03/2017



Digital Marketing Data Enrichment



Weather Data Source



city: 2950157
- date: 2017-03- 12
2t: 17
- date: 2017-03-13
2t: 20

regionID (GeoNames)

date (ISO 8601)

Value
manipulation

Ok



KEYWORD	#im	REGION	Date	Date ISO
194906	64	Thuringia	11/03/2017	2017-03-11
517827	50	Bavaria	12/03/2017	2017-03-12
459143	42	Berlin	12/03/2017	2017-03-12



Digital Marketing Data Enrichment



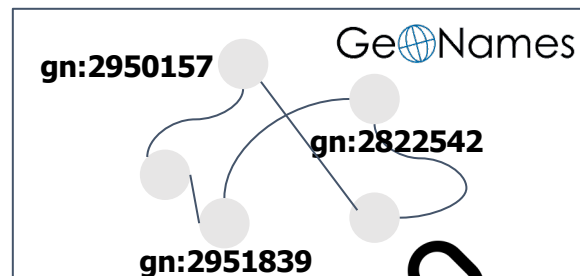
Weather Data Source



city: 2950157
- date: 2017-03- 12
2t: 17
- date: 2017-03-13
2t: 20

regionID (GeoNames)

date (ISO 8601)



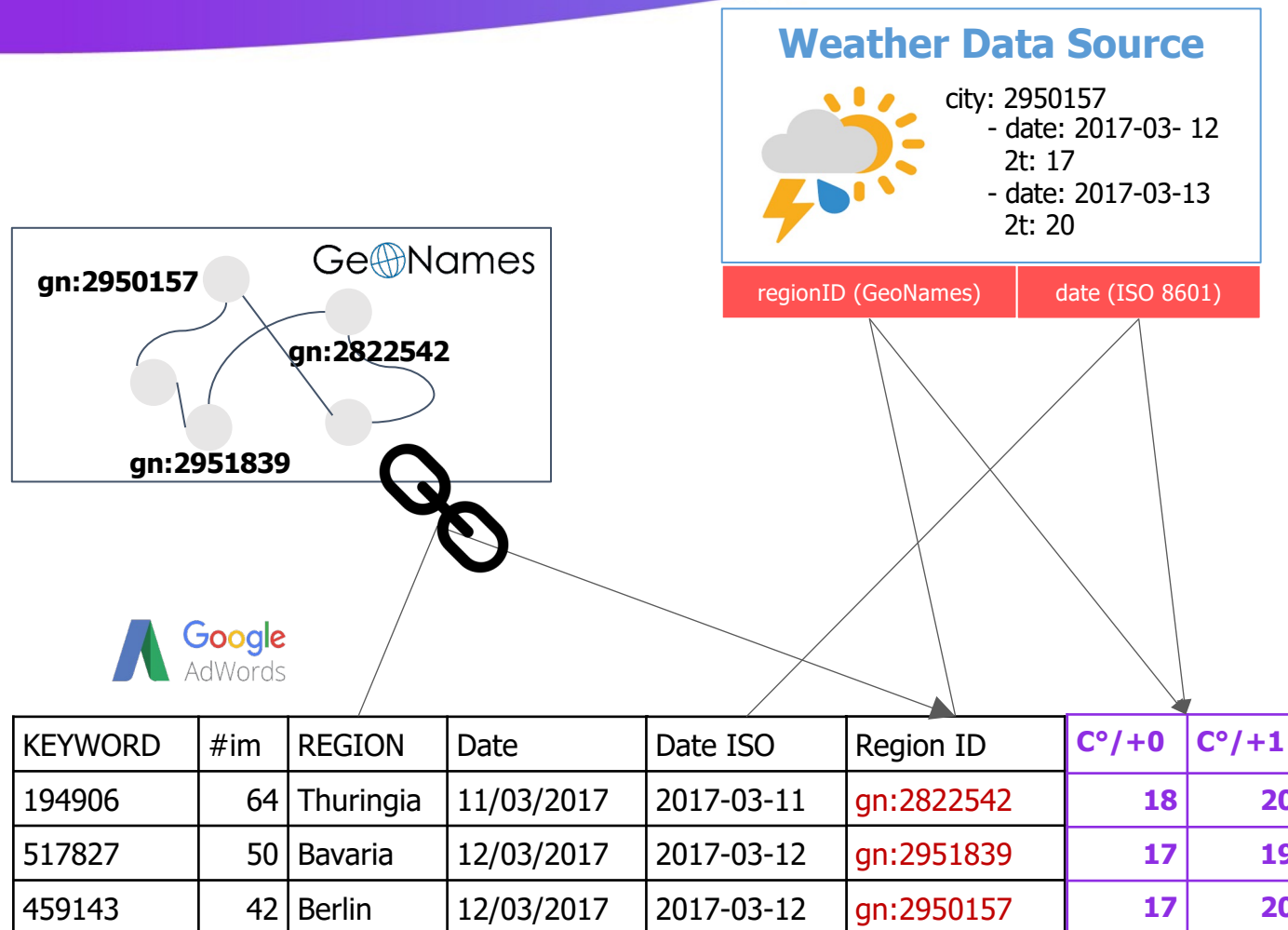
KEYWORD	#im	REGION	Date	Date ISO	Region ID
194906	64	Thuringia	11/03/2017	2017-03-11	gn:2822542
517827	50	Bavaria	12/03/2017	2017-03-12	gn:2951839
459143	42	Berlin	12/03/2017	2017-03-12	gn:2950157

Value
manipulation

Linking



Digital Marketing Data Enrichment



Value manipulation

Linking

Extension





Tabular data

“Traditional” ML
&
Data Analytics

- Weather-based optimization in digital marketing [[ISWC19](#), [Tech. and Appl. for BDV22](#)]

Semantic Data Enrichment: in a Bigger Picture

Semantic Data Integration and KG Construction

BROWSE BY COUNTRY FRANCE ①

ENTITY	JURISDICTION	LINKED TO	DATA FROM
OBOLT GLOBAL LTD	British Virgin Islands	France	Panama Papers
FAIRWINDS SAILING S.A.	British Virgin Islands	France	Panama Papers
SIENNA HOLDINGS INC.	Panama	France	Panama Papers
PROCESOS E INGENIERIA S.A.	Panama	France	Panama Papers
WESCO INTERNATIONAL S.A.	Bahamas	France	Panama Papers
INTERPETROLEUM LTD.	Bahamas	France	Panama Papers
UNICONSTRUCT INC.	Panama	France	Panama Papers
MADINA AL MNWORA AGENCY CORPORATION	Panama	France	Panama Papers
CARLTON PROPERTIES LIMITED	Bahamas	France	Panama Papers
BELROS HOLDINGS LTD.	Bahamas	France	Panama Papers

Entities in OffshoreLeaks linked to France

<https://offshoreleaks.icij.org/search?c=FRA&cat=0>

Business

FRANCE - PANAMA

France investigating hundreds for tax fraud due to Panama Papers

French magistrates are investigating 26 individuals or institutions suspected of hiding huge sums of money in tax havens and tax authorities are looking into 416 suspected tax dodgers in an inquiry sparked by last year's Panama Papers revelations, according to reports.

Issued on: 05/04/2017 - 15:49

PANAMA PAPERS

« Panama papers » : le business offshore du Crédit agricole et de la BNP

Les « Panama papers » mettent en évidence les pratiques opaques des deux grandes banques françaises dans les paradis fiscaux.

Par Anne Michel, Jérémie Baruch et Maxime Vaudano

EVASION FISCALE

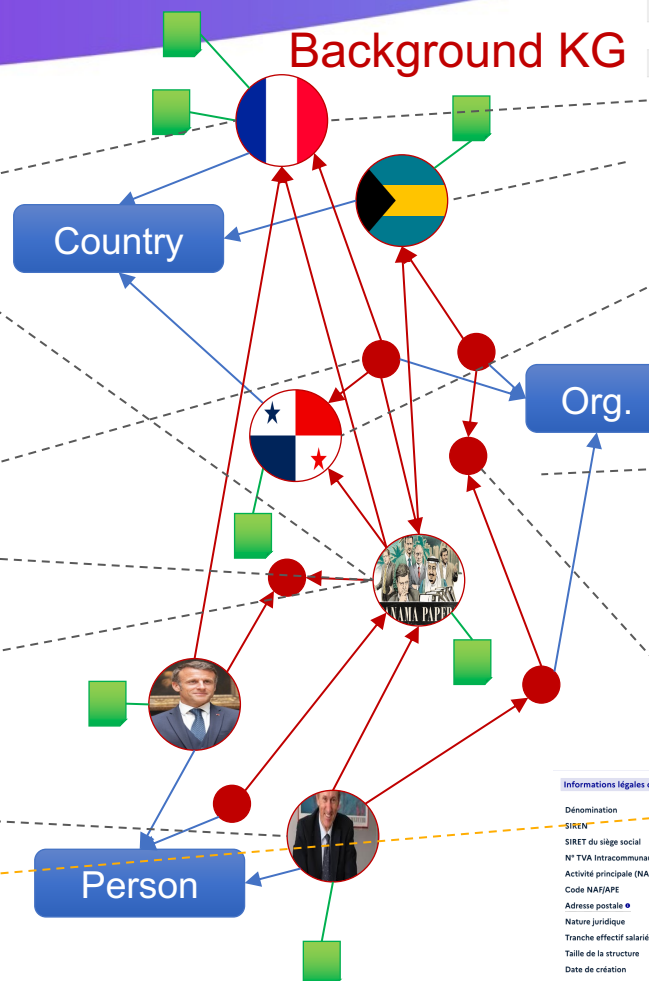
« Panama Papers » : les mauvaises affaires offshore de l'industriel Gérard Autajon

En plus d'un redressement fiscal, d'une amende et d'une peine de prison avec sursis, le PDG a dû faire face à des frais non anticipés sur deux sociétés.

Par Jérémie Baruch

Publié le 21 juin 2018 à 10h54, modifié le 22 juin 2018 à 07h12 - Lecture 3 min.

... Sienna ...



Headquarters	Bank	Number of foundations
Luxembourg	Experta Corporate & Trust Services (100% subsidiary of BIL)	1,659
Luxembourg	Banque J. Safra Sarasin - Luxembourg S.A.	963
United Kingdom	Credit Suisse Channel Islands Limited	918
Guernsey	HSBC Private Bank (Monaco) S.A.	778
Monaco	HSBC Private Bank (Suisse) S.A.	733
Switzerland	UBS AG (subsidiary Rue du Rhône in Geneva)	579
United Kingdom	Coutts & Co Trustees (Jersey) Limited	487
Jersey	Société Générale Bank & Trust Luxembourg	465
Luxembourg	Landsbanki Luxembourg S.A.	404
United Kingdom	Rothschild Trust Guernsey Limited	378
Guernsey	Banco Santander	119
Spain	BBVA	19

Foundations firms 'offshore' customers through banks in Wikipedia

Informations légales de SIENNA REAL ESTATE HOLDING FRANCE

Dénomination	SIENNA REAL ESTATE HOLDING FRANCE
SIREN	492 220 553
SIRET du siège social	492 220 553 00027
N° TVA Intracommunautaire	FR492 220 553
Activité principale (NAF/APE)	Fonds de placement et entités financières similaires
Code NAF/APE	64.30Z
Adresse postale	18 RUE DE COURCELLES, 75008 PARIS 8
Nature juridique	SAS, société par actions simplifiée
Tranche effectif salarié de la structure	Unité non employeuse
Taille de la structure	Petite ou Moyenne Entreprise (PME), en 2020
Date de création	20/09/2006
Dernière modification des données Insee	06/08/2022
Justificatif(s) d'existence	<ul style="list-style-type: none"> Télécharger l'extrait RNE Avis de situation Insee

Company data from

.data.gouv.fr

<https://annuaire-entreprises.data.gouv.fr/entreprise/sienna-real-estate-holding-france-492220553>

Annotations

Named Entity Recognition

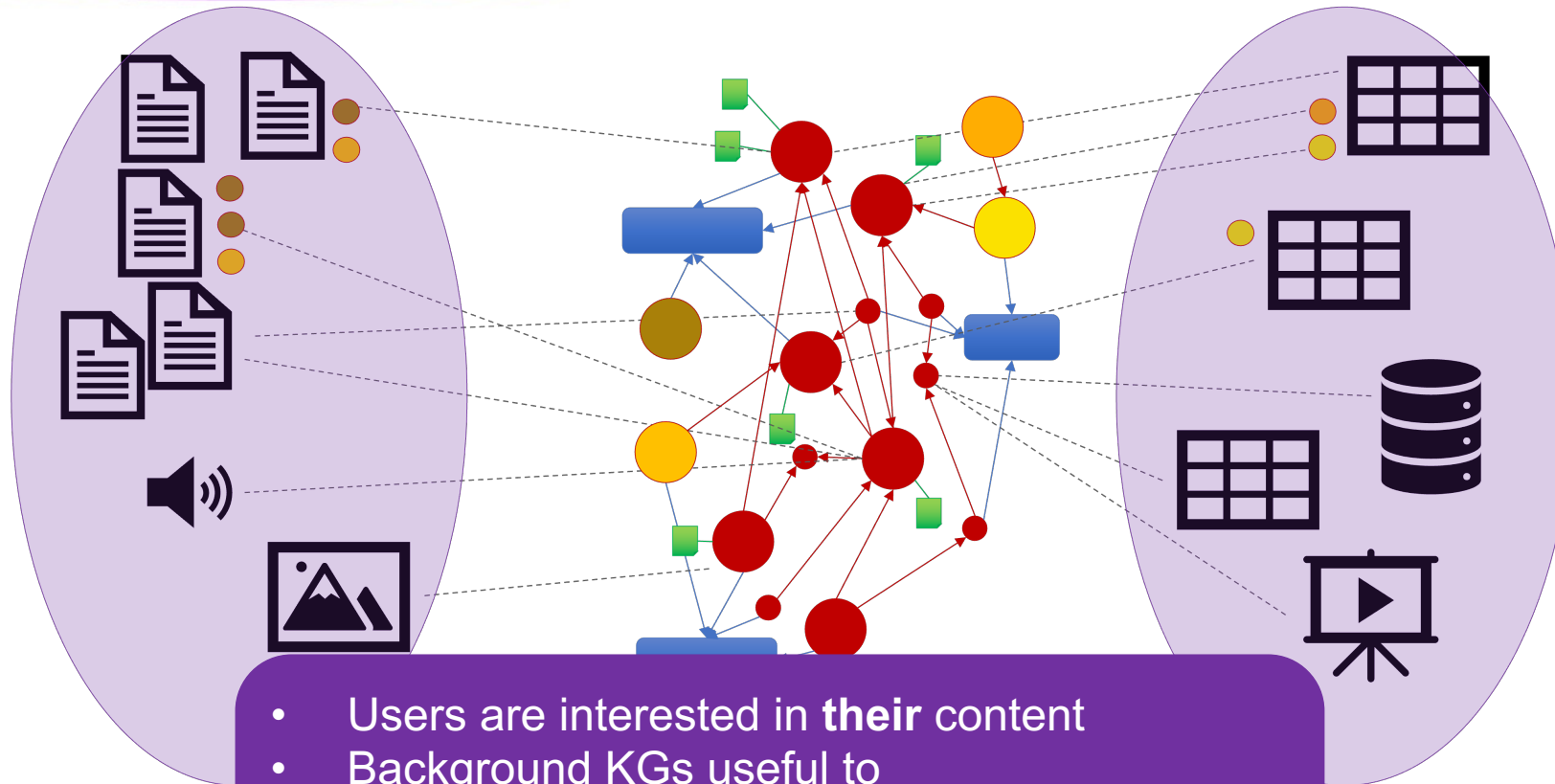
Named Entity Linking

NIL Prediction

Tutorial @ ESWC 2024

18

Semantic Enrichment vs Data Integration and KG Construction



- Users are interested in **their** content
- Background KGs useful to
 - support integration
 - extend their content with additional data
- The construction of a KG **can be** a byproduct



EventRegistry: Event Tracking at Scale with KGs

The screenshot shows the EventRegistry web application interface. The browser address bar displays the URL: `eventregistry.org/intelligence?tab=items&searchMode=simple&type=events&conditions=2--1-1-Giorgia%20Meloni%20&`. The search bar contains the text "Giorgia Meloni". The interface includes a sidebar with navigation options: List of Events, Top Concepts, Tag Cloud, Timeline, Event Locations, News Sources, Article authors, Sentiment, Concept Graph, Concept Trends, and Event Categories. The main content area shows a list of events with 2,416 results found. Two event cards are visible:

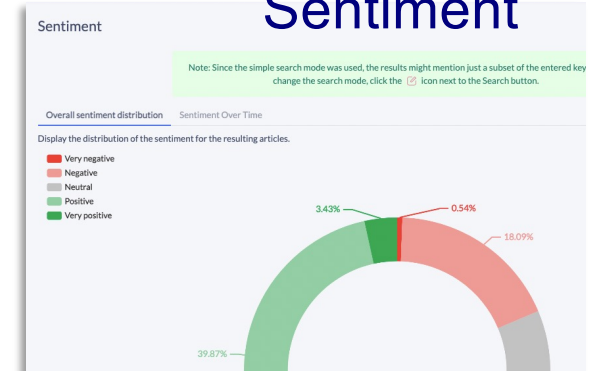
- Event 1:** "Pilgrims to Mussolini's birthplace pray that new PM will resurrect a far-right Italy".
 - When: Fri, October 21, 2022
 - Where: Rome, Italy
 - Articles: 452
 - Virality: 72
 - Sentiment: ●●
 - Tags: Giorgia Meloni, Italy, Brothers of Italy, Coalition government, Silvio Berlusconi, Far-right politics, Sergio Mattarella, Forza Italia, Right-wing politics, Matteo Salvini, Ukraine, Rome, Italy, Russia, Mario Draghi, Prime Minister of the United Kingdom, Cabinet (government).
 - Snippet: "In Predappio, supporters celebrate victory of their first female prime minister Giorgia Meloni, leader of a party with neo-fascist origins Dressed in... (The Guardian)"
- Event 2:** "Macron a rencontré Meloni à Rome, avec qui il promet 'dialogue et ambition'".
 - When: Fri, October 21, 2022
 - Where: Rome, Italy
 - Articles: 214
 - Virality: 73
 - Sentiment: N/A
 - Tags: Giorgia Meloni, Government, Italy, Matteo Salvini, Prime minister, Silvio Berlusconi, Rome, Italy, Sergio Mattarella, Coalition government, Forza Italia, Mario Draghi, Benito Mussolini, President of France, Quirinal Palace, Populism, Antonio Tajani, Euroscepticism, Europe.
 - Snippet: "INTERNATIONAL - À peine entrée en fonction, la nouvelle Première ministre italienne Giorgia Meloni a fait ses débuts sur la scène internationale ce di... (Le Huffington Post)"

Events
clusters of news about a same topic, based on semantic annotations with entities, concepts, time, place

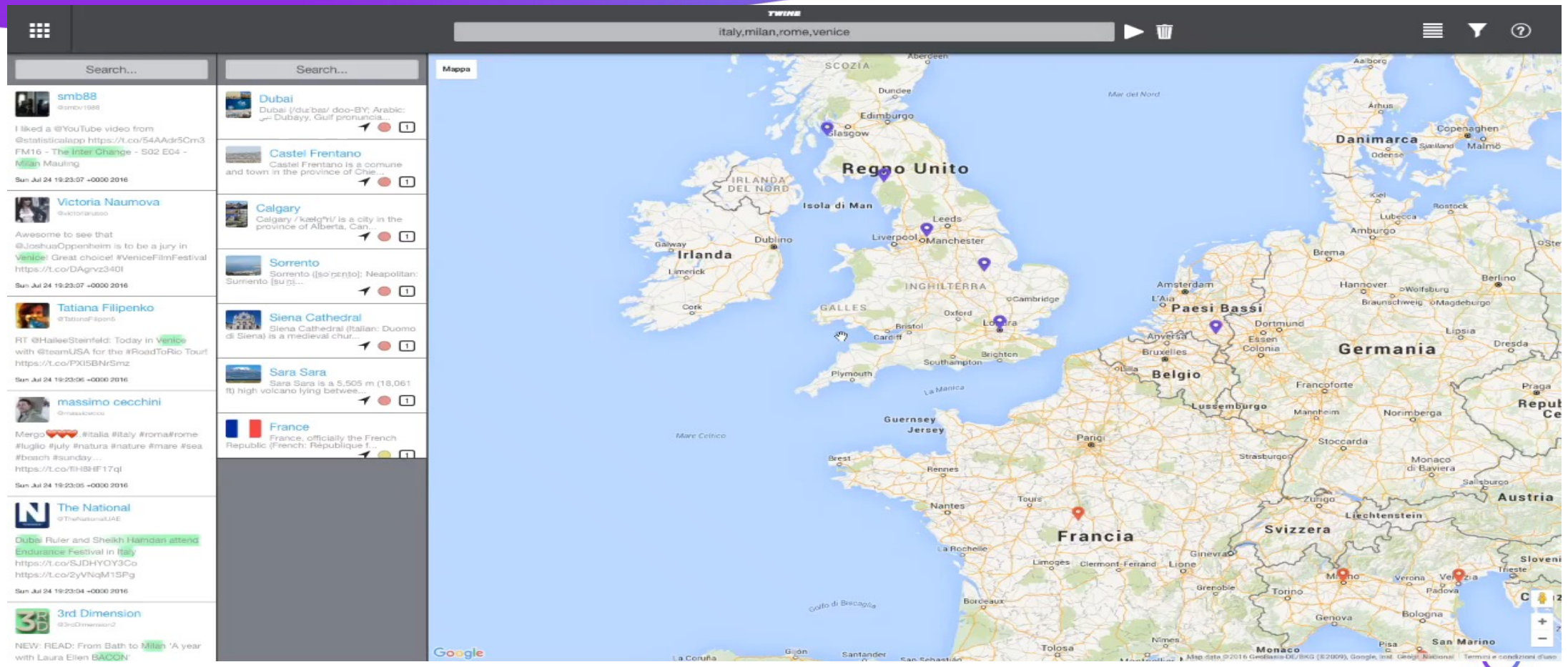
Graph



Sentiment



Textual Data Enrichment for Twitter Analysis in TWINE



[Nozza et al. 2017] D Nozza, F Ristagno, M Palmonari, E Fersini, P Manchanda, E Messina: **TWINE: A real-time system for TWEet analysis via INformation Extraction**. EACL (DEMO) 2017, 25

Data Enrichment Operations

- Linking & Integration
 - Annotations: assign semantic data type
 - Discover & merge external data: matching)
 - Join & union
- Extension
 - Calculate column using expressions
 - Calculate column using inference / classification
 - Discover & merge external data: fetching and merging
 - Fill empty cells
- Manipulations: data structuring
 - Change column data type
 - Delete or duplicate column
 - Detect & change encoding
 - Change format
- Manipulations: values
 - Structure-based values
 - Merge or split columns
 - Normalize numeric values
 - Generate primary key column
 - Filtering
 - Delete/keep filtered rows
 - Delete empty and invalid rows
 - Extract value parts
 - Filter with regular expressions
 - Cleaning
 - Change date & time format
 - Change letter case / number format
 - Delete by pattern
 - Edit & replace cell data
 - Remove extra whitespace
 - Standardize strings by pattern
 - Standardize values in clusters
 - Deduplicate data



Tabular Data Enrichment with Text Classification: the SN Example

- SpendNetwork:
 - built and maintains the largest database of open public tenders in the world: over 180 million lines of tabular data
 - needs better contextualization of the data: disambiguation of organizations and classification of the tenders/organizations
- Enrichment task:
 - Disambiguate key entities: buyers → Wikidata
 - Classify the tenders and the corresponding organizations against canonical classification systems



The SN Example: Input Data

buyer	title	description (SN)	category
derbyshire county council	CST109 Aids to Rehabilitation (for Occupational Health)	Derbyshire County Council is out to tender for aids to rehabilitation following Occupational Health assessments. These aids are aimed at staff members requiring adjustments to their workstation(s) - ie. bespoke chairs, desks, additional support cushions, and other desktop items. The Council requires a fully delivered service with installation and set-up options.	Miscellaneous medical devices and products
derbyshire county council	PLACE430H Supply of UPVC Pipes, Ducting & Access Chambers	Derbyshire County Council is seeking a supplier/s for the supply of UPVC pipes, ducting & access chambers. The contract will be split into four lots:- Lot 1 is for the supply of hdpe twin wall pipes and fittings Lot 2 is for the supply of single wall underground drainage upvc underground drainage pipes to bs en1401-1 Lot 3 is for the supply of upvc ducting Lot 4 is for the supply of Access boxes covers and frames	Pipeline, piping, pipes, casing, tubing and related items
derbyshire county council	CCP100 Secretariat Services for the f40 Group	Derbyshire County Council, on behalf of the f40 Group, is out to tender for Secretariat Services for the group - working to raise issues around education funding and to campaign for change at a national level. The f40 Secretariat supports the f40 Group with strategy and campaign planning, media relations, social media and content creation as well as event management and general administrative tasks.	Business and management consultancy services



The SN Example: Target Output

buyer	title	description (SN)	category	Classifier Taxonomy: Standard context	Classifier Taxonomy: IPTC	buyer (Wikidata ID)	name	description	Classifier Taxonomy: Standard context	Classifier Taxonomy: IPTC
derbyshire county council	CST109 Aids to Rehabilitation (for Occupational Health)	Derbyshire County Council is out to tender for aids to rehabilitation following Occupational Health assessments. These aids are aimed at staff members requiring adjustments to their workstation(s) - ie. bespoke chairs, desks, additional support cushions, and other desktop items. The Council requires a fully delivered service with installation and set-up options.	Miscellaneous medical devices and products	job market	Industrial accident and incident Disaster, accident and emergency incident/Accident and emergency incident/Industrial accident and incident	Q5261561	derbyshire county council	local authority for the english county of derbyshire	institutions 24.7 public administration 23.79	No data for "IPTC Media Topics" taxonomy.
derbyshire county council	PLACE430H Supply of UPVC Pipes, Ducting & Access Chambers	Derbyshire County Council is seeking a supplier/s for the supply of UPVC pipes, ducting & access chambers. The contract will be split into four lots:- Lot 1 is for the supply of hdpe twin wall pipes and fittings Lot 2 is for the supply of single wall underground drainage upvc underground drainage pipes to bs en1401-1 Lot 3 is for the supply of upvc ducting Lot 4 is for the supply of Access boxes covers and frames	Pipeline, piping, pipes, casing, tubing and related items	information technology	Hardware Economy, business and finance/Economic sector/Computing and information technology/Hardware	Q5261561	derbyshire county council	local authority for the english county of derbyshire	institutions 24.7 public administration 23.79	No data for "IPTC Media Topics" taxonomy.
derbyshire county council	CCP100 Secretariat Services for the f40 Group	Derbyshire County Council, on behalf of the f40 Group, is out to tender for Secretariat Services for the group - working to raise issues around education funding and to campaign for change at a national level. The f40 Secretariat supports the f40 Group with strategy and campaign planning, media relations, social media and content creation as well as event management and general administrative tasks.	Business and management consultancy services	business 0.69 finance 0.69	Civil and public service Politics/Government/Civil and public service 29.4% Campaign finance Politics/Election/Political campaigns/Campaign finance 26.46% Social problem Society/Social problem 26.46%	Q5261561	derbyshire county council	local authority for the english county of derbyshire	institutions 24.7 public administration 23.79	No data for "IPTC Media Topics" taxonomy.

The SN Example: Services/Steps

buyer	title	description (SN)	category	Classifier Taxonomy: Standard context	Classifier Taxonomy: IPTC	buyer (Wikidata ID)	name	description	Classifier Taxonomy: Standard context	Classifier Taxonomy: IPTC
derbyshire county council	CST109 Aids to Rehabilitation (for Occupational Health)	Derbyshire County Council is out to tender for aids to rehabilitation following Occupational Health assessments. These aids are aimed at staff members requiring adjustments to their workstation(s) - ie. bespoke chairs, desks, additional support cushions, and other desktop items. The Council requires a fully delivered service with installation and set-up options.	Miscellaneous medical devices and products	job market	Industrial accident and incident Disaster, accident and emergency incident/Accident and emergency incident/Industrial accident and incident	Q5261561	derbyshire county council	local authority for the english county of derbyshire	institutions 24.7 public administration 23.79	No data for "IPTC Media Topics" taxonomy.
derbyshire county council	PLACE430H Supply of UPVC Pipes, Ducting & Access Chambers	Derbyshire County Council is seeking a supplier/s for the supply of UPVC pipes, ducting & access chambers. The contract will be split into four lots:- Lot 1 is for the supply of hdpe twin wall pipes and fittings Lot 2 is for the supply of single wall underground drainage upvc underground drainage pipes to bs en1401-1 Lot 3 is for the supply of upvc ducting Lot 4 is for the supply of Access boxes covers and frames	Pipeline, piping, pipes, casing, tubing and related items	information technology	Hardware Economy, business and finance/Economic sector/Computing and information technology/Hardware	Q5261561	derbyshire county council	local authority for the english county of derbyshire	institutions 24.7 public administration 23.79	No data for "IPTC Media Topics" taxonomy.
derbyshire county council	CCP100 Secretariat Services for the f40 Group	Derbyshire County Council, on behalf of the f40 Group, is out to tender for Secretariat Services for the group - working to raise issues around education funding and to campaign for change at a national level. The f40 Secretariat supports the f40 Group with strategy and campaign planning, media relations, social media and content creation as well as event management and general administrative tasks.	Business and management consultancy services	business 0.69 finance 0.69	Civil and public service Politics/Government/Civil and public service 29.4% Campaign finance Politics/Election/Political campaigns/Campaign finance 26.46% Social problem Society/Social problem 26.46%	Q5261561	derbyshire county council	local authority for the english county of derbyshire	institutions 24.7 public administration 23.79	No data for "IPTC Media Topics" taxonomy.

Part II: Semantic Data Enrichment, Applications and Requirements

2) The Link & Extend enrichment paradigm

“Extending the linked data idea for data enrichment”



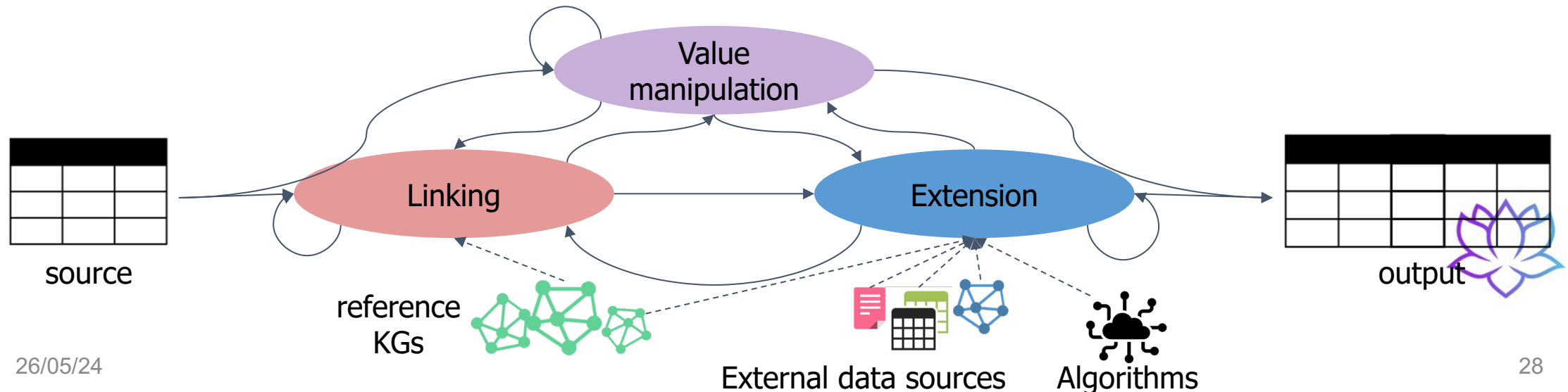
Semantic Data Enrichment: a Sequential View

- Inputs:
 - a **source** dataset D
 - a pool of **reference data sources**

Output:
a dataset D'

Data enrichment: a path on a **data transformations** graph G^T

Semantic data enrichment: at least one node is a **linking operation** or **semantic data sources** for extensionx



Semantic Data Enrichment: a Sequential View

Linking vs extension: not a sharp distinction

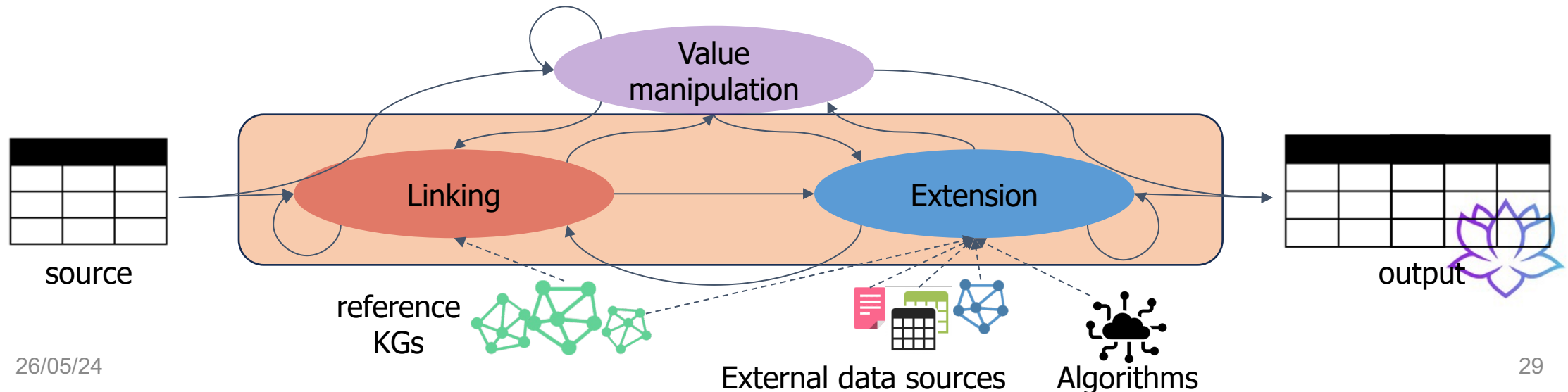
- Extension:
 - for each row, an array or a more complex object (e.g., a table)
- Linking: a function an identifier
 - for each row, at most one identifier (e.g., IRI)

- Inputs:
 - a **source** dataset D
 - a pool of **reference data sources**

Output:
a dataset D'

Data enrichment: a path on a **data transformations** graph G^T

Semantic data enrichment: at least one node is a **linking operation** or **semantic data sources** for extensionx



Semantic Data Enrichment: Combining Steps

Reconciliation against **KGs** on the web

Bridging across different **KGs** by exploiting
links among them

Additional data from **KGs** available/made
available on the web
(large data sources for data enrichment)

Reconciliation by **matching**

Reconciliation by owl:sameAs **links**

Identifiers support **extension**

The diagram illustrates the process of semantic data enrichment through several steps: 1. **Reconciliation by matching**: A red arrow points from the 'Keyword' column to the 'City' column. 2. **Reconciliation by owl:sameAs links**: Red arrows point from 'City' to 'Region' and from 'Region' to 'ID (Geonames)'. 3. **Identifiers support extension**: Blue arrows point from 'ID (Geonames)' to 'Latitude (Geonames)', 'Longitude (Geonames)', and 'ID (Wikidata)'. Another blue arrow points from 'ID (Wikidata)' to 'Population (Wikidata)'. A final blue arrow points from 'Population (Wikidata)' to the 'Temp (ECMWF)' column.

Keyword	#im	City	Region	ID (Geonames)	Latitude (Geonames)	Longitude (Geonames)	ID (Wikidata)	Population (Wikidata)	Temp (ECMWF)	Date
194906	64	Altenburg	Thuringia	2822542	50.98763	12.43684	Q1205		18°	2017-03-11
517827	50	Inglostadt	Bavaria	2951839	48.76508	11.42372	Q980		17°	2017-03-12
459143	42	Berlin	Berlin	2950157	52.52437	13.41053	Q648102		17°	2017-03-12
891139	36	Munich	Bavaria	2951839	48.13743	11.57549	Q980		19°	2017-03-11
459143										2017-03-12

Example of data enrichment by composing different individual linking and extension steps

Link & Extend

- Link

- Reconciliation against IDs that can be used for subsequent queries
 - E.g., Wikidata, spatial queries, etc.
- Critical step in the enrichment process
 - Uncertainty and error propagation
 - Requires control
- Parameter space
 - Configuration
 - Which input?
 - Which output?
 - Which algorithm?
 - Which threshold?
 - ...

- Extend

- Different families of operations
 - Query-based
 - Get info from the target source
 - Inference-based
 - Classification (also associated with uncertainty)
 - Embedding
- Parameter space
 - Configuration (deterministic)
 - Which input?
 - Which output?
 - Which properties to use in the target KG
 - Which query
 - Configuration (uncertain extension)



Link & Extend vs. Services

- Link by **linking services**

- Reconciliation against IDs that can be used for subsequent queries
 - E.g., Wikidata, spatial queries, etc.
- Critical step in the enrichment process
 - Uncertainty and error propagation
 - Requires control
- Parameter space
 - Configuration
 - Which input?
 - Which output?
 - Which algorithm?
 - Which threshold?
 - ...

- Extend by **extension services**

- Different families of operations
 - Query-based
 - Get info from the target source
 - Inference-based
 - Classification (also associated with uncertainty)
 - Embedding
- Parameter space
 - Configuration (deterministic)
 - Which input?
 - Which output?
 - Which properties to use in the target KG
 - Which query
 - Configuration (uncertain extension)

API access



Requirements for Link & Extend

- Setting up reconciliation services

- Interoperability

- W3C Reconciliation Service API v0.2
 - A protocol for data matching on the Web
 - Not a recommendation yet but good step towards interoperability
 - Supported by OpenRefine

- Data providers

- To increase data access, support also linking to your data with by exposing
 - A reconciliation service
 - A lookup service

- Developers

- Can wrap existing reconciliation services for usage
 - E.g., data matching by Atoka (SpazioDati)

- Setting up extension services

- Full-fledged query language, e.g., SPARQL

- What data?

- Ontologies
 - Data profiles, e.g., ABSTAT [Alva Principe & al. 2022, Spahiu & al 2024]

- APIs

- Documentation

- Which parameters?



Link & Extend: Interactive Exploration

- Web-scale search
 - Which data?
 - Dataset search [Chapman & al. 2019]
 - Which services?
 - API recommendation [Nawaz et al. 2022]
- Catalogue-based discovery

KEYWORD	#im	REGION	Date
194906	64	Thuringia	11/03/2017
517827	50	Bavaria	12/03/2017
459143	42	Berlin	12/03/2017

Target data

C°/+0	C°/+1
18	20
17	19
17	20

Diagram showing three arrows with question marks pointing from the rows of the first table to the rows of the second table.

Which source?

How are they accessible?

- Latitude, longitude coordinates
- Strings, e.g., string combination (disambiguation), IDs
- IRI, e.g., Geonames

How to get to have the input they require to be invoked?



...



Explorative analysis first

Link & Extend: Interactive Exploration

- Web-scale search
 - Which data?
 - Dataset search [Chapman & al. 2019]
 - Which services?
 - API recommendation [Nawaz et al. 2022]
- Catalog-based discovery

Which source?

How are they accessible?

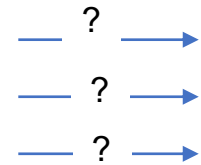
- Latitude, longitude coordinates
- Strings, e.g., string combination (disambiguation), IDs
- IRI, e.g., Geonames

How to get to have the input they require to be invoked?

- String(s) > Geonames > Coordinates IRI via reconciliation
- String(s) > Coordinates IRI via geocoding

How accurate are the links?

KEYWORD	#im	REGION	Date
194906	64	Thuringia	2017-03-11
517827	50	Bavaria	2017-03-12
459143	42	Berlin	2017-03-12



Target data

C°/+0	C°/+1
18	20
17	19
17	20

KEYWORD	#im	REGION	Date	Region ID	C°/+0	C°/+1
194906	64	Thuringia	2017-03-11	gn:2822542	18	20
517827	50	Bavaria	2017-03-12	gn:2951839	17	19

KEYWORD	#im	REGION	Date	Coordinates	C°/+0	C°/+1
194906	64	Thuringia	2017-03-11	50.86111,11.05194	18	20
517827	50	Bavaria	2017-03-12	48.7775 11.43111	17	19
459143	42	Berlin	2017-03-12	52.52437,13.41053	17	20



Explorative analysis first

Link & Extend: Scalability

- Example from digital marketing

- Size of data to enrich
 - 1,000 millions of active keywords
 - 20 TB of historical performance data
 - 1GB of new data every day
- Frequency
 - Event collection: monthly
 - Weather enrichment: weekly
 - Keyword categorization: daily
 - Daily update of new keywords

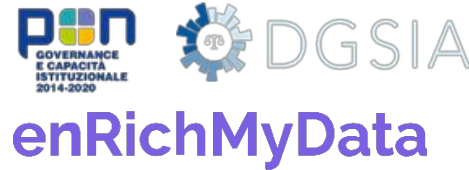
- Execution models

- In browser / notebook with PC
 - Interactive enrichment
- Scripting
 - Batch processing
- Distributed Computation Infrastructures
 - E.g., scheduling, speed-up



Applications of Data Enrichment

Main projects



Main data

Documents

Tabular data

Documents

Applications
and analytical
methods

Query
Answering

Semantic Search
&
Data Exploration

“Traditional” ML
&
Data Analytics

Analyses with
Representation
Learning

...

Contributions:
applications
and novel
analytical
methods

- Criminal investigations [\[SDSM20\]](#)
- Exploring data-contexts to contextualize news articles [\[ISWCdemo15, ESWC17\]](#)
- Enrichment and analysis of social media [\[EACLdemo17\]](#)

- Weather-based optimization in digital marketing [\[ISWC19, Tech. and Appl. for BDV22\]](#)

- Text-based entity embeddings and time-aware entity similarity [\[ISWC18\]](#)
- Entity evolution (+ with [CADE](#) alignment [\[AAAI19\]](#))



This talk

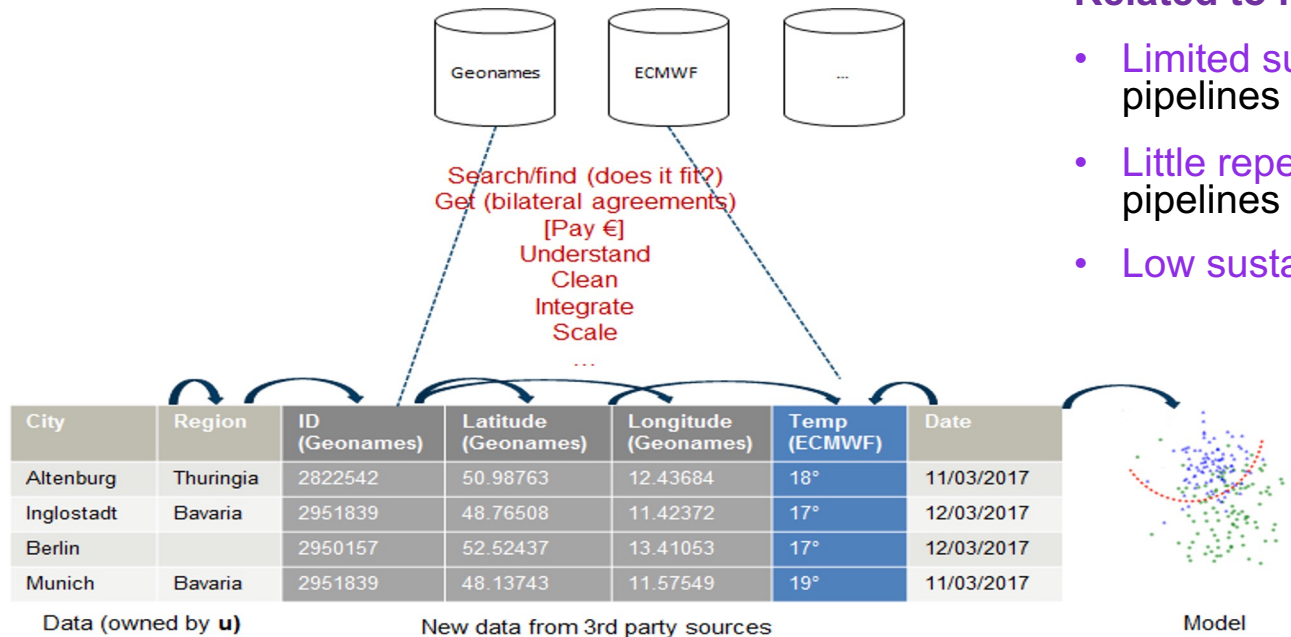
enRichMyData – Addressed Challenges

Related to data enrichment

- Lack of holistic end-to-end support for data enrichment lifecycle
- Steep learning curve for performing data enrichment tasks
- Lack of Humans-In-The-Loop approach

Related to infrastructure for data enrichment

- Limited support for scalable execution of data enrichment pipelines
- Little repeatability and reusability of data enrichment pipelines
- Low sustainability due to inefficient use of resources

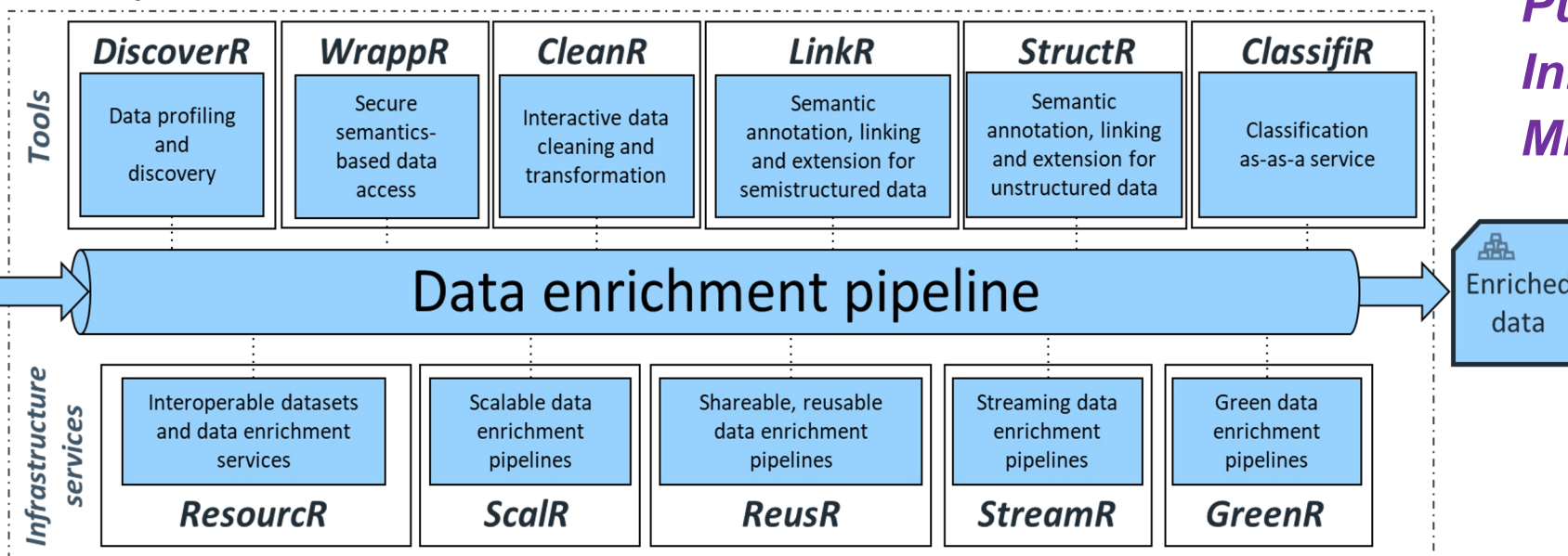


enRichMyData – Contributions

Toolbox for building rich, high-quality, valuable, and FAIR-compliant datasets to feed downstream Big Data and AI applications in the context of Data-sharing Ecosystems

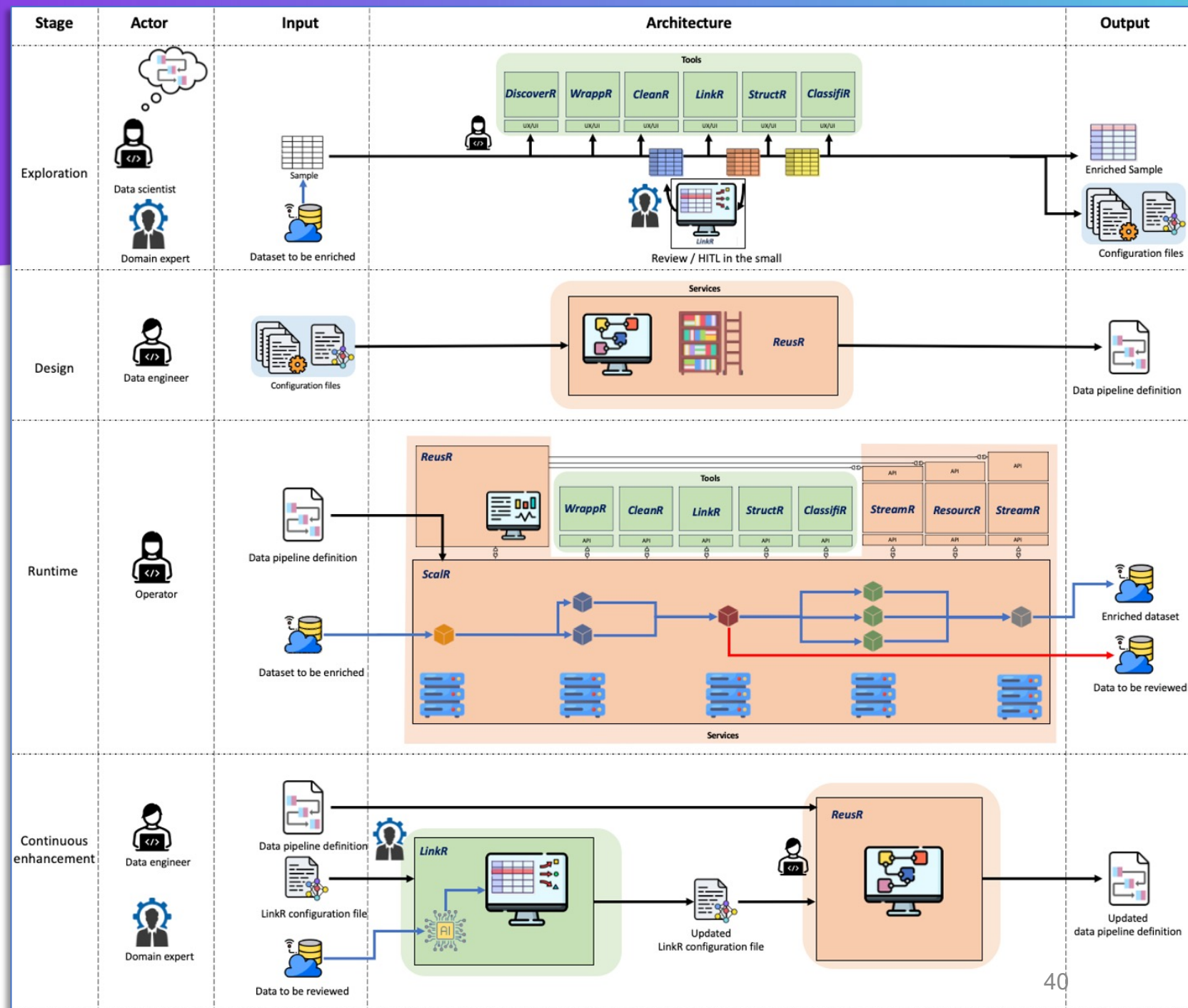
Applications in
Digital Marketing
Manufacturing
Predictive Maintenance
Public Procurement
Innovation Ecosystems
Mineral Processing

enRichMyData toolbox

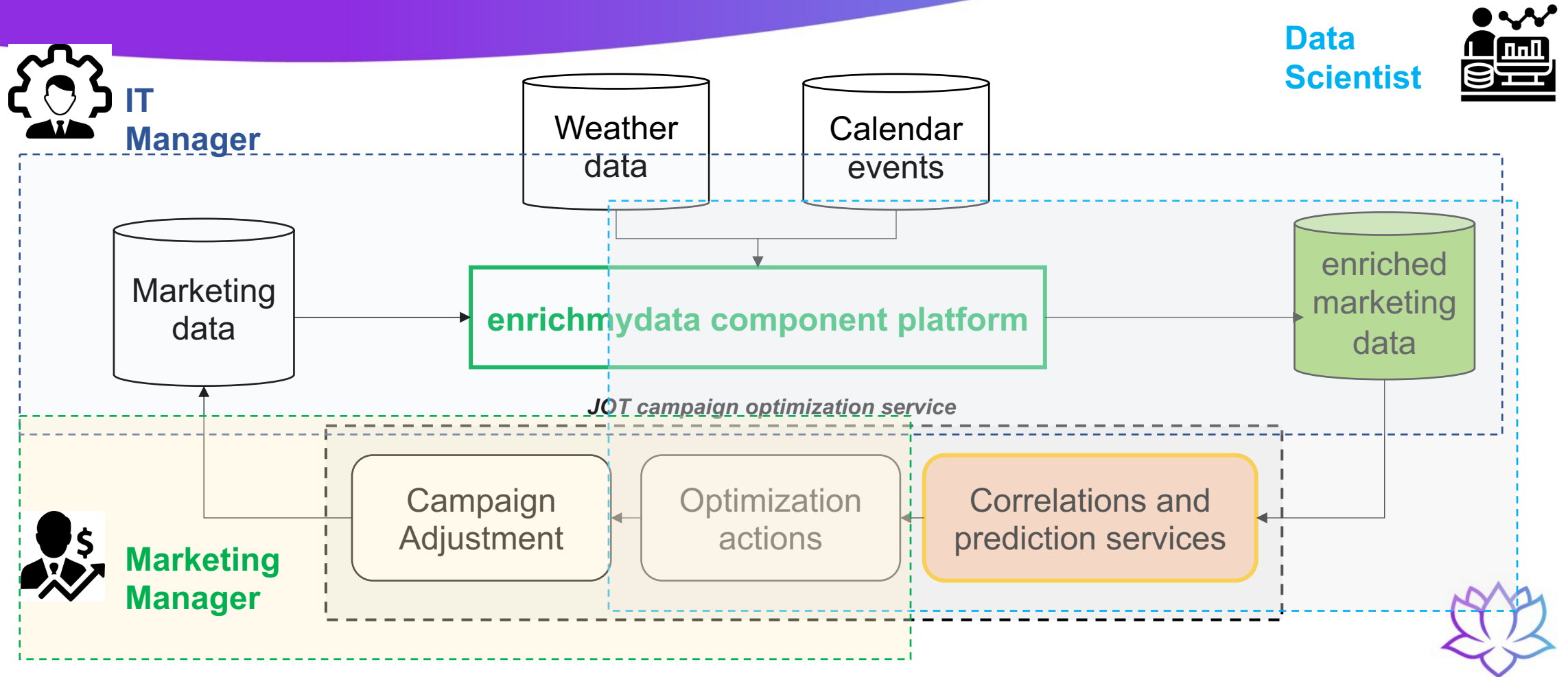


Phases & Architectures

- *Exploration / Discovery*
 - Interactivity, HITL
 - Data and tool exploration
 - Out: enriched sample, task specification, configuration files
- *Design*
 - Workflow definition and management
 - Reusable components, versioning
 - Out: data pipeline definition
- *Runtime*
 - Workflow execution
 - Horizontal scalability
 - Execution monitoring
 - Out: enriched data, data to be reviewed
- *Continuous enhancement (e.g., linking)*
 - Revise uncertain results for some records
 - Out: updated data, updated pipeline



Digital Marketing Data Enrichment



Digital Marketing enRichMyData

