enRichMyData

SINTEF *

^ INSID&S Lab
*Department of Informatics,
Systems and Communication
University of Milan - Bicocca*

Semantic Data Enrichment: from Interactive Exploration to Scalable Deployment

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

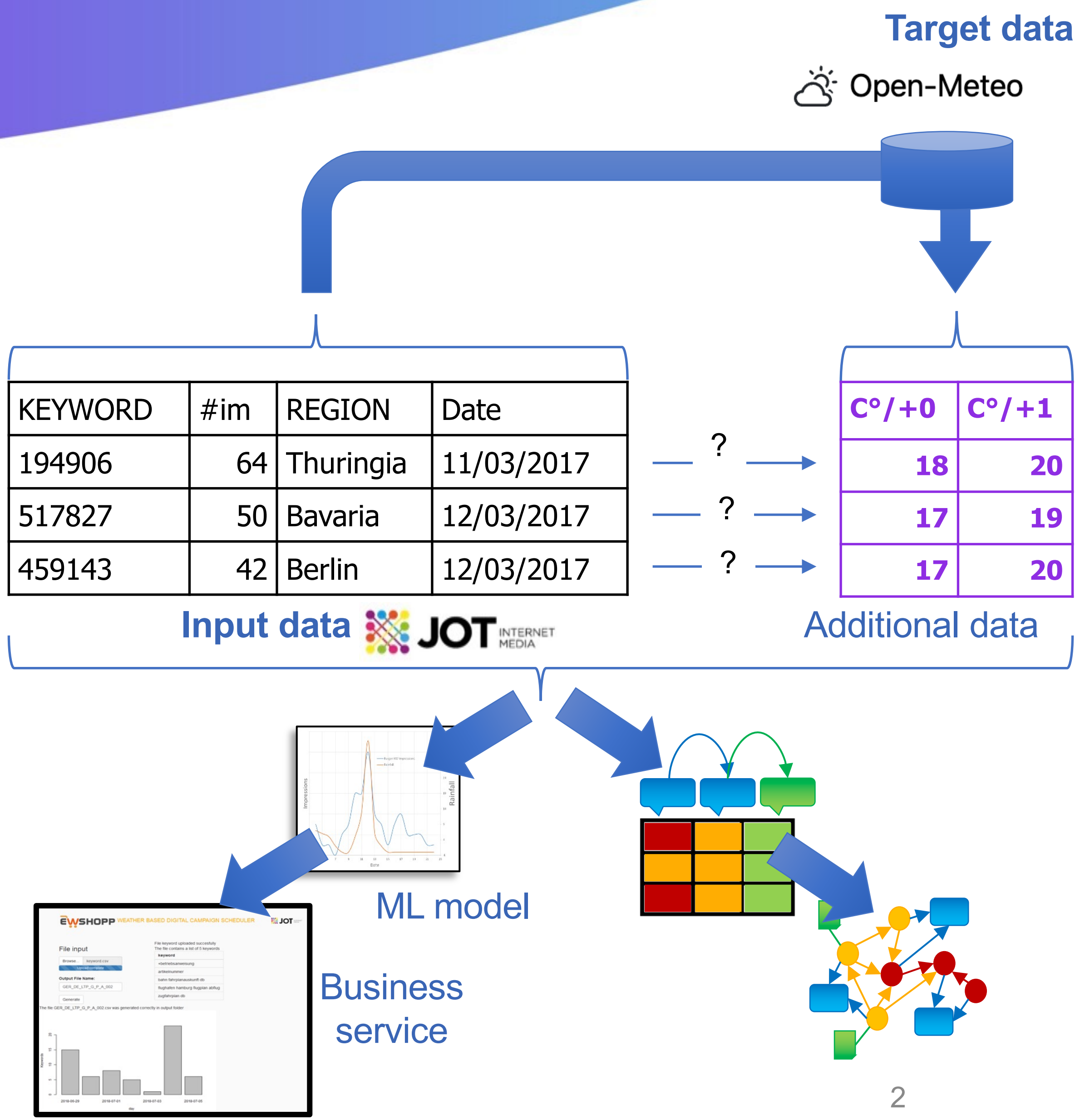
Part 1 – Introduction and Outline

*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**.*

Tutorial @ ESWC2024

Data Enrichment
vs Knowledge Graphs (KGs)

- Data enrichment
  - Add context to the data of an organization, i.e., add more data to an input dataset
    - User *A* wants to enrich her dataset *D* to make a dataset D’
    - *… data D’-D typically fetched from a third-party source S or inferred*
- Knowledge graphs for data enrichment:
  - **Data annotation**: data published with semantic annotations, i.e., shared vocabularies and systems of identifiers
  - **Data augmentation**: access to third-party sources mediated by KGs

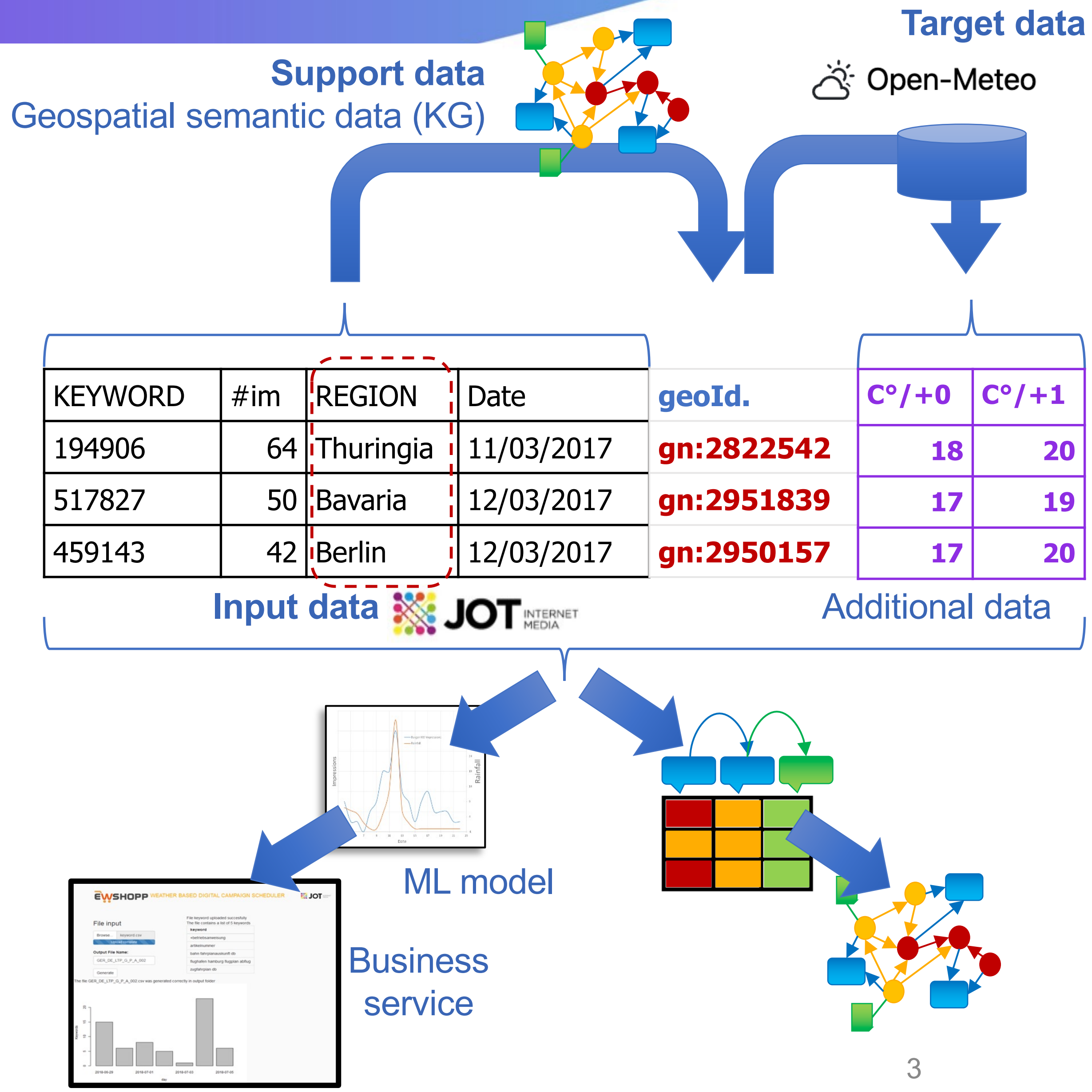


KEYWORD	#im	REGION	Date
194906	64	Thuringia	11/03/2017
517827	50	Bavaria	12/03/2017
459143	42	Berlin	12/03/2017

C°/+0	C°/+1
18	20
17	19
17	20

Data Enrichment vs Knowledge Graphs

- Data enrichment
  - Add context to the data of an organization, i.e., add more data to an input dataset
    - User *A* wants to enrich her dataset *D* to make a dataset D'
    - *… data D'-D typically fetched from a third-party source S or inferred*
- Knowledge graphs for data enrichment:
  - **Data annotation**: data published with semantic annotations, i.e., shared vocabularies and systems of identifiers
  - **Data augmentation**: access to third-party sources mediated by KGs



KEYWORD	#im	REGION	Date	geoId.	C°/+0	C°/+1
194906	64	Thuringia	11/03/2017	gn:2822542	18	20
517827	50	Bavaria	12/03/2017	gn:2951839	17	19
459143	42	Berlin	12/03/2017	gn:2950157	17	20

Semantic Data Enrichment

- A (relatively) novel point of view for exploitation of semantics
  - Extending ideas the semantic web community is familiar with
- Semantics
  - Linking to identifiers as in KGs
  - Fetching information from KG and other sources
  - Service interoperability
  - Representation learning semantics, e.g., LMs and LLMs
- Main take-home messages
  - Highly relevant in the industry
  - The *link & extend* paradigm and its service-based extension
  - Table annotation algorithms for data enrichment
  - Humans-in-the-loop: the role of interactive exploration
  - Volume-aware approaches: the role of scalability

Our focus: enrichment of tabular data

Outline

45’
- Part II: Semantic Data Enrichment, Applications and Requirements
  - Semantics and KGs for data enrichment
  - The *Link & Extend* enrichment paradigm
    - Interactive exploration and scalability

60’
- 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
  - Aaa
  - aaa

60’

- Part VI: Conclusions and Discussion
  - Wrap-up and take-home messages
  - Discussion

15’