Data Engineering for Data Science

Oscar Romero
Database Technologies and Information Management (DTIM)
Universitat Politècnica de Catalunya-BarcelonaTech (UPC)

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

DTIM www.essi.upc.edu/dtim

DTIM
Database Technologies and Information Management

Members
- 3 senior lecturers (Alberto Abelló, Cristina Gómez, Oscar Romero)
- 3 post-docs (Petar Jovanovic, Besim Bilalli, Sergi Nadal)
- 5 PhD
- 6 MSc / BSc students

- Departments
  - Services and Information Systems Engineering
  - Statistics and Operational Research
- Experience in data and knowledge management
  - Research
  - Development (Transfer of Technology)
- Application fields
  - Big Data
  - Business Intelligence

Vision
Democratize information management and analysis to solve current societal challenges

Mission
Gain and create knowledge on information management and analysis to contribute to global progress and development by (1) leading and contributing to local and international research and innovation projects, (2) disseminating scientific and technological knowledge, and (3) building intersectoral partnerships.

https://www.essi.upc.edu/dtim/

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

DTIM www.essi.upc.edu/dtim

Erasmus Mundus Joint Doctorate (H2020)

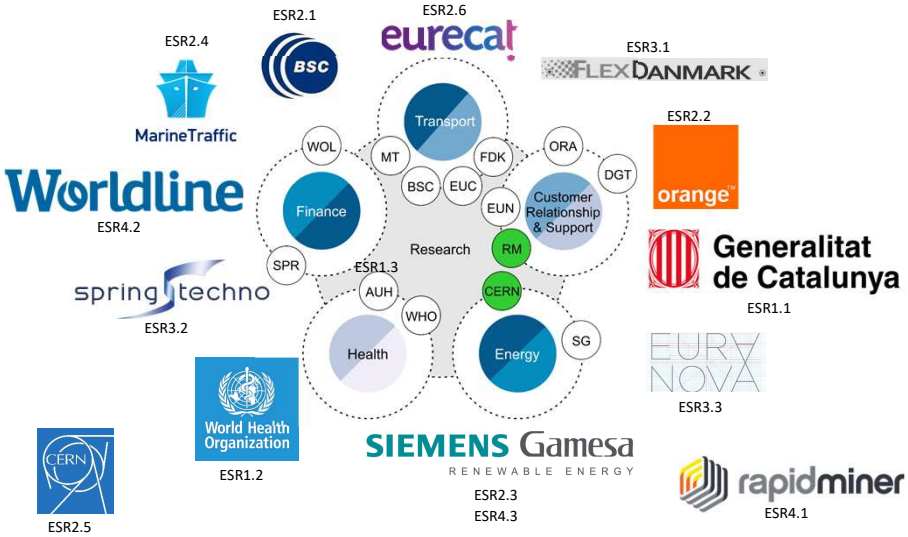
Data Engineering for Data Science

https://deds.ulb.ac.be/

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

DTIM www.essi.upc.edu/dtim

Disciplines and Associate Partners

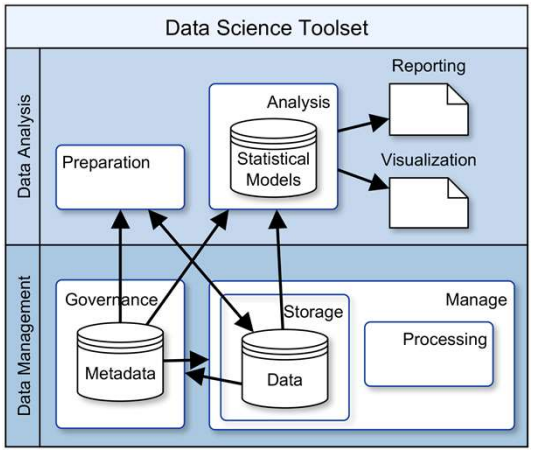


Data Engineering for Data Science

Data Science flows

Big Data Management System

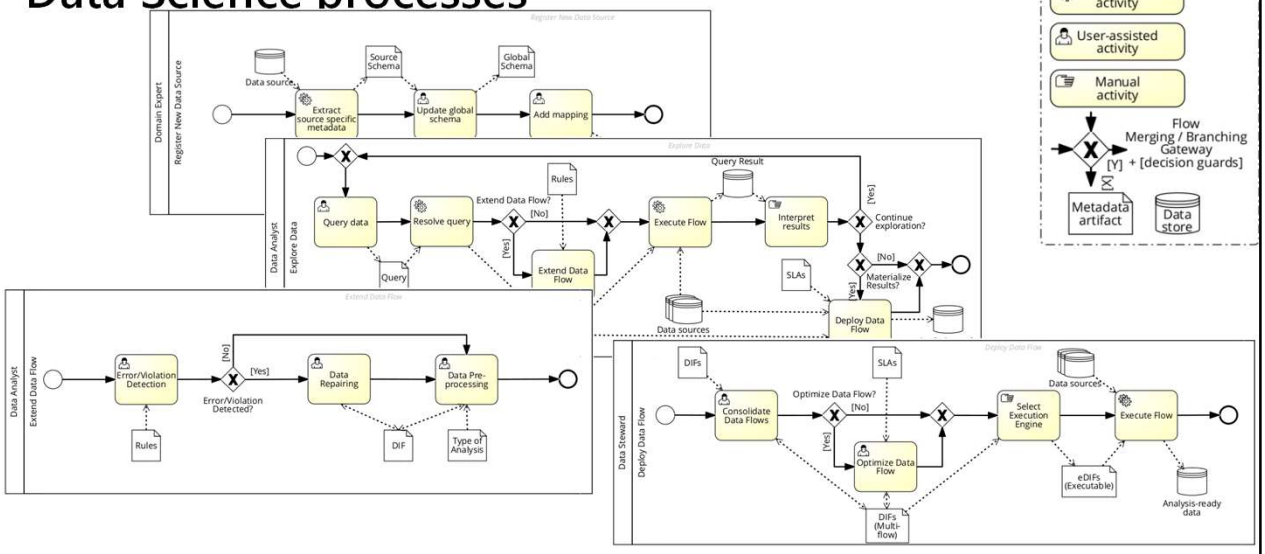
Big Data Management System



Petar Jovanovic, Sergi Nadal, Oscar Romero, Alberto Abelló, Besim Bilalli: Quarry: A User-centered Big Data Integration Platform. Inf. Syst. Frontiers 23(1): 9-33 (2021)

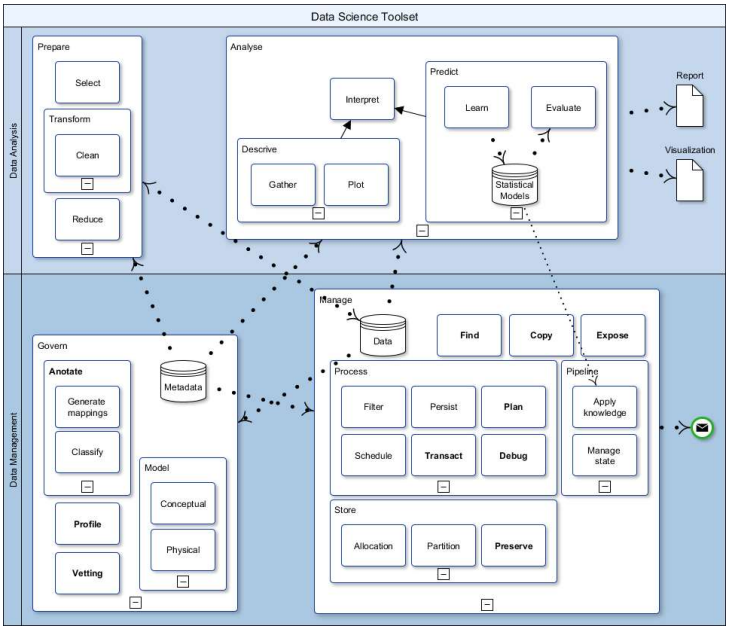
Sergi Nadal, Oscar Romero, Alberto Abelló, Panos Vassiliadis, Stijn Vansummeren: An integration-oriented ontology to govern evolution in Big Data ecosystems. Inf. Syst. 79: 3-19 (2019)

Data Science processes

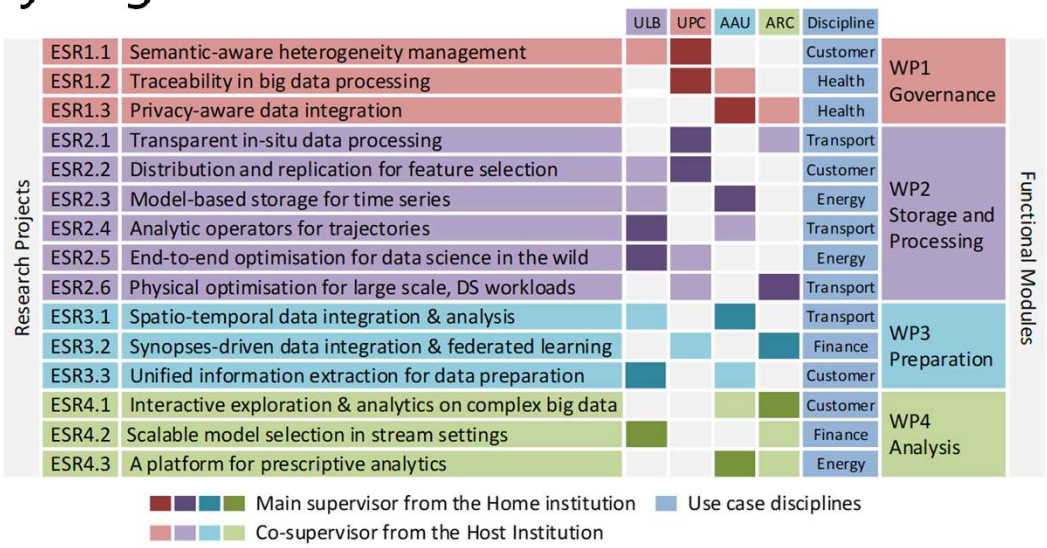


Oscar Romero, Robert Wrembel
Data Engineering for Data Science: Two Sides of the Same Coin. DaWaK 2020: 157-166

Submodules



Early Stage Researchers



UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

9

DTIM www.essi.upc.edu/dtim

Data Discovery

Finding Joinable Datasets in large Data Lakes

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

DTIM www.essi.upc.edu/dtim

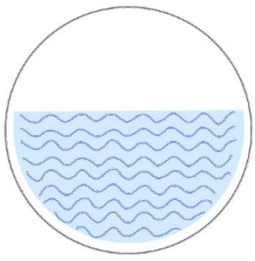
Data Integration: Discovering Joinable Datasets

Reference dataset

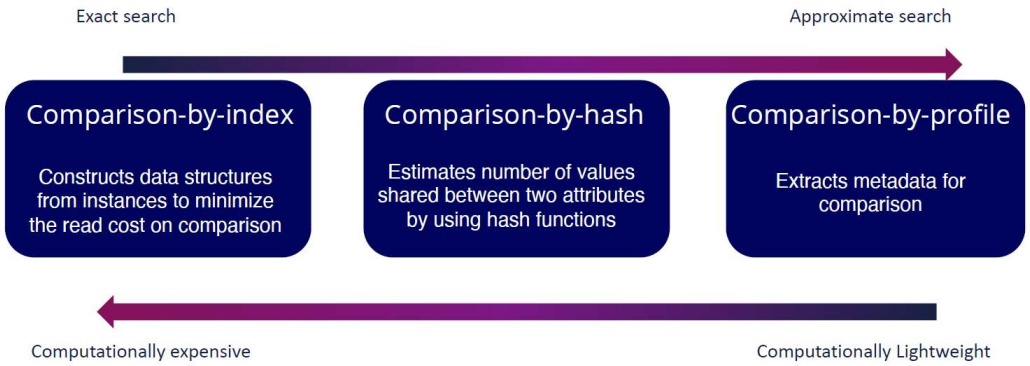
1st Admin. Level	2nd Admin. Level	Store code	Channel
Basque Country	Vitoria-Gasteiz	A00151	Social networks
Castile and Leon	Valladolid	A00248	Transit ads
Catalonia	Barcelona	B00311	Transit ads
Galicia	Santiago de ...	C00094	Social networks
Aragon	Zaragoza	H00202	Social networks
La Rioja	Logroño	L00174	Social networks
Navarre	Pamplona	N00272	Social networks
Asturias	Oviedo	A00078	Transit ads
Cantabria	Santander	C00102	TV and social networks

Data Integration: Discovering Joinable Datasets

1st Admin. Level	2nd Admin. Level	Store code	Channel
Basque Country	Vitoria-Gasteiz	A00151	Social networks
Castile and Leon	Valladolid	A00248	Transit ads
Catalonia	Barcelona	B00311	Transit ads
Galicia	Santiago de ...	C00094	Social networks
Aragon	Zaragoza	H00202	Social networks
La Rioja	Logroño	L00174	Social networks
Navarre	Pamplona	N00272	Social networks
Asturias	Oviedo	A00078	Transit ads
Cantabria	Santander	C00102	TV and social networks



State of the Art



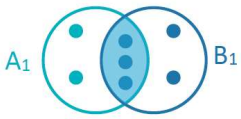
Similarity Metrics

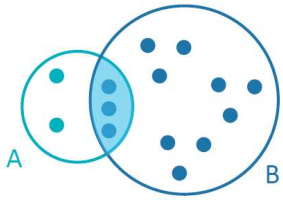
$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Symmetric

$$Containment(A,B) = \frac{|A \cap B|}{|A|}$$

Asymmetric





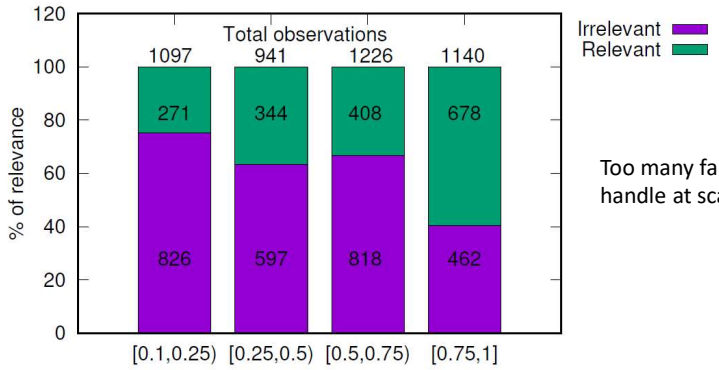
Jaccard(A_1,B_1) = 3/7 = 0.428
Jaccard(B_1,A_1) = 3/7 = 0.428
Containment(A_1,B_1) =3/5=0.6
Containment(B_1,A_1) =3/5=0.6

Jaccard(A_2,B_2) = 3/14 = 0.21
Jaccard(B_2,A_2) = 3/14 = 0.21
Containment(A_2,B_2) =3/5=0.6
Containment(B_2,A_2) =3/12=0.25

Containment as Synonym of Joinability

Experiment: We designed an experiment collecting 138 datasets from open repositories such as Kaggle and OpenML. We collected heterogeneous datasets ranging different topics, which yielded a total of 110, 378 candidate pairs of string attributes, where 4, 404 of those have a containment higher or equal to 0,1.

We then analyzed if containment was enough to identify **semantic** joins.



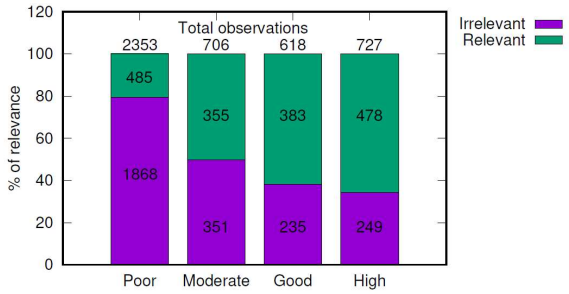
Too many false positives to handle at scale!

Tuning the Similarity Metric

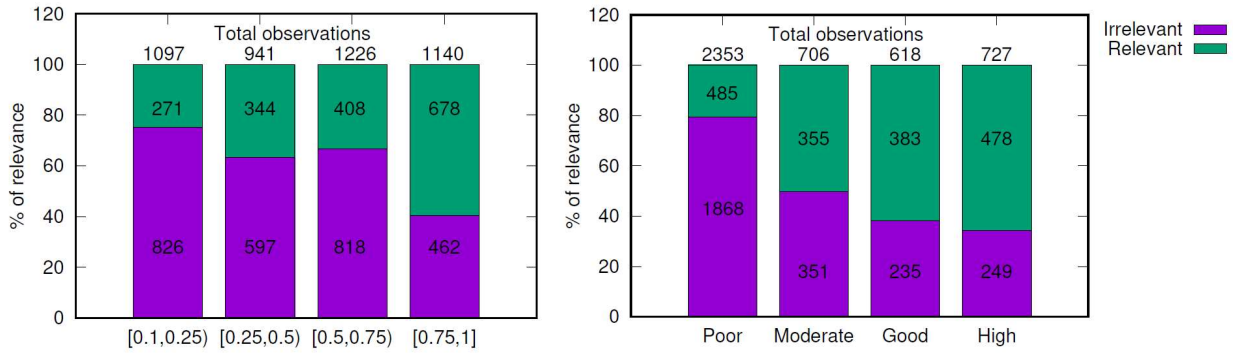
Qualitative metric considering containment and the cardinality proportion between the set of distinct velues of each attribute

$$Quality(A,B) = \begin{cases} (4)\ \texttt{High}, & C(A,B) \geq C_H \wedge \frac{|A|}{|B|} \geq K_H \\ (3)\ \texttt{Good}, & C(A,B) \geq C_G \wedge \frac{|A|}{|B|} \geq K_G \\ (2)\ \texttt{Moderate}, & C(A,B) \geq C_M \wedge \frac{|A|}{|B|} \geq K_M \\ (1)\ \texttt{Poor}, & C(A,B) \geq C_P \\ (0)\ \texttt{None}, & \text{otherwise} \end{cases}$$

Thresholds empirically determined



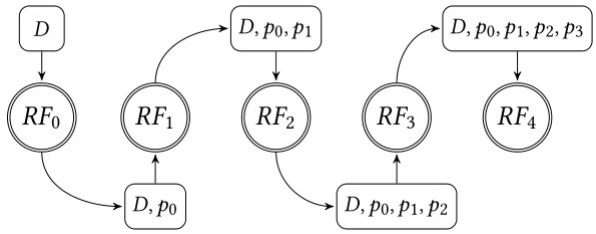
Containment Vs. Our Metric



NextiaJD: A Learning Approach

Category	Meta-feature	Description	Norm.?
Cardinalities	Cardinality	Number of distinct values within an attribute	Yes
	Uniqueness	Measures if the attribute contains unique values	No
	Incompleteness	Measures the number of missing values	No
	Entropy	Measures the variety of an attribute	Yes
Value distribution	Average frequency	Th	
	Min frequency	Th	
	Max frequency	Th	
	SD frequency	Th	
	Octiles	Th	
	Min perc frequency	Th	
	Max perc frequency	Th	
	SD perc frequency	Th	
	Constancy	Fr	
	Frequent words	Th	
	Soundex	Th	
Syntactic	Data type	Th no	
	Specific type	Th	
	Percentage data type	Th	
	Percentage specific type	Th	
	Longest string	The number of characters in the longest string	Yes
	Shortest string	The number of characters in the shortest value in the attribute	Yes
	Average string	Average length of the strings in term of characters	Yes
	Number words	The number of words in the attribute	Yes
	Average words	The average words in the attribute	Yes
	Min words	The minimum words in the attribute	Yes
	Max words	The maximum words in the attribute	Yes
	SD words	The standard deviation in the attribute	Yes
Pair metadata	Best containment	The containment score assuming all distinct values are covered	No
	Flipped containment	Containment assuming all distinct values are covered divided by max cardinality	No
	Name distance	Measures the difference of two attribute names using Levenshtein distance	No

Table 11: Meta-features composing a profile



Experiments: Performance

Testbed	*XS*	*S*	*M*	*L*
File size	0 – 1 MB	1 – 100 MB	100 MB –1 GB	> 1 GB
Datasets	28	46	46	19
String attributes	159	590	600	331

Table 13: Characteristics per testbed

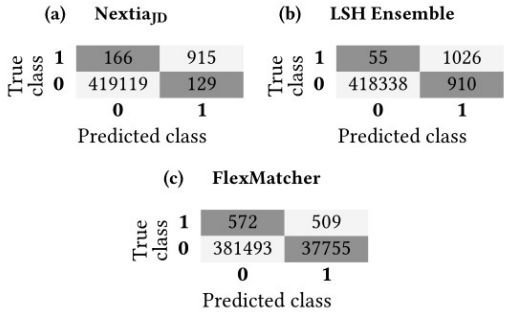


Figure 9: Combined confusion matrices for each system on testbeds XS, S, M

Experiments: Scalability

Testbed	*XS*	*S*	*M*	*L*
File size	0 – 1 MB	1 – 100 MB	100 MB –1 GB	> 1 GB
Datasets	28	46	46	19
String attributes	159	590	600	331

Table 13: Characteristics per testbed

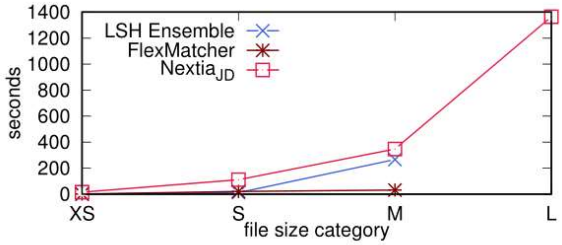


Figure 6: Pre runtime

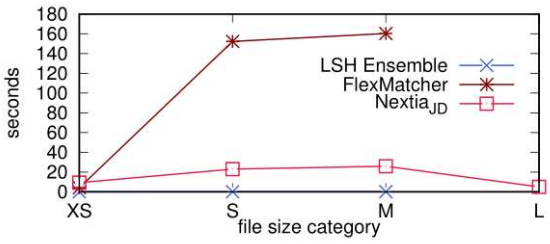


Figure 7: Query runtime (*discovery-by-dataset*)

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
DTIM www.essi.upc.edu/dtim

Experiments: Scalability

Testbed	XS	S	M	L
File size	0 – 1 MB	1 – 100 MB	100 MB –1 GB	> 1 GB
Datasets	28	46	46	19
String attributes	159	590	600	331

Table 13: Characteristics per testbed

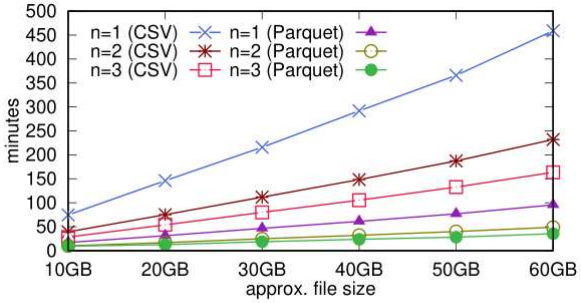


Figure 11: Profiling runtime over an increasing file size

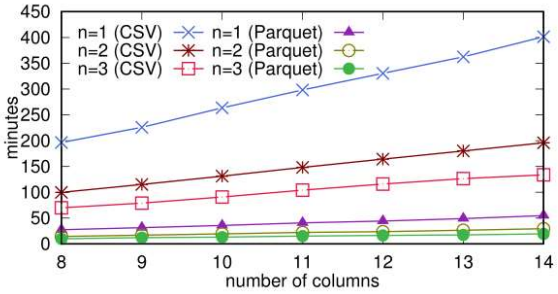


Figure 12: Profiling runtime over an increasing number of columns

NextiaJD: Scalable Data Discovery

Javier Flores, Sergi Nadal, Oscar Romero. Towards Scalable Data Discovery (EDBT 2021 Short Paper)

Javier Flores, Sergi Nadal, Oscar Romero. Effective and Scalable Data Discovery with NextiaJD (EDBT 2021 Demo)

https://www.essi.upc.edu/~snadal/nextiajd.html