

DS4DM

Data Search for Data Mining

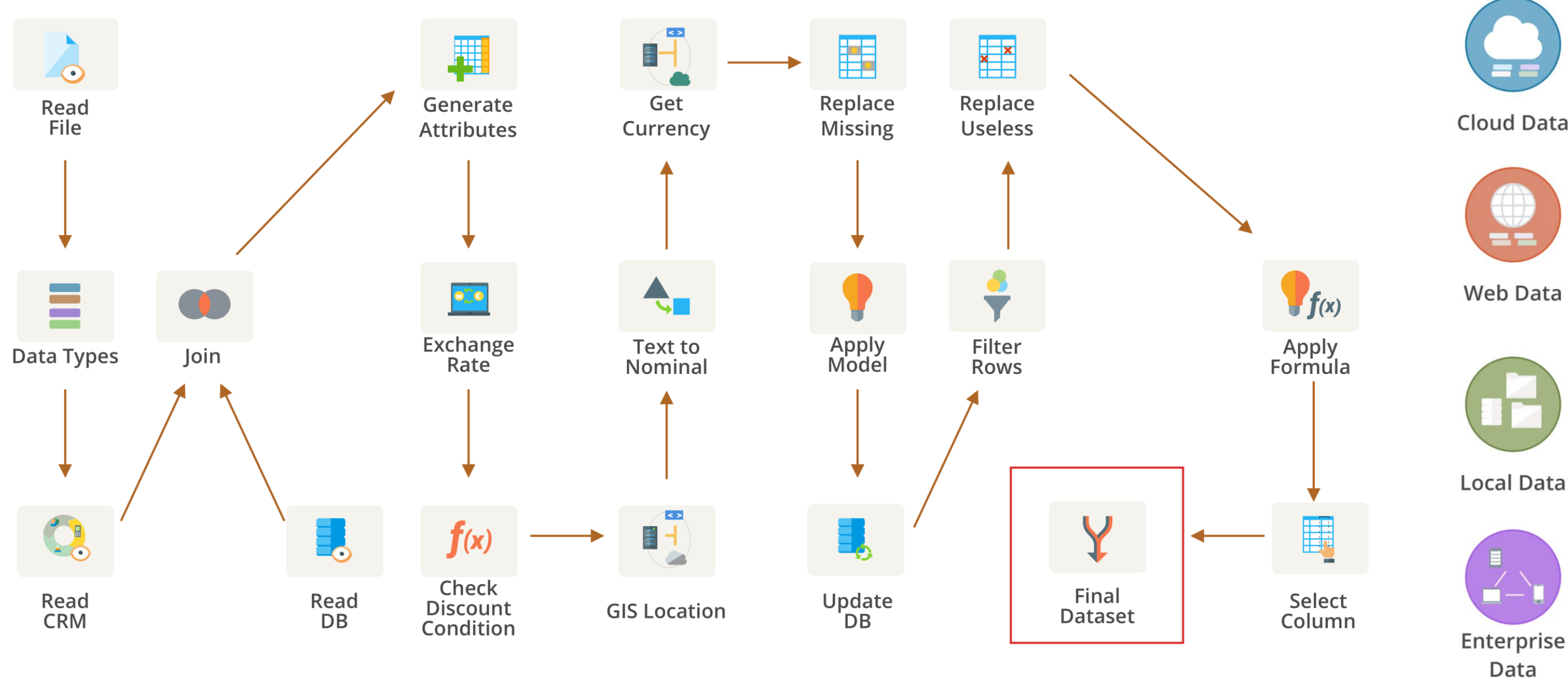
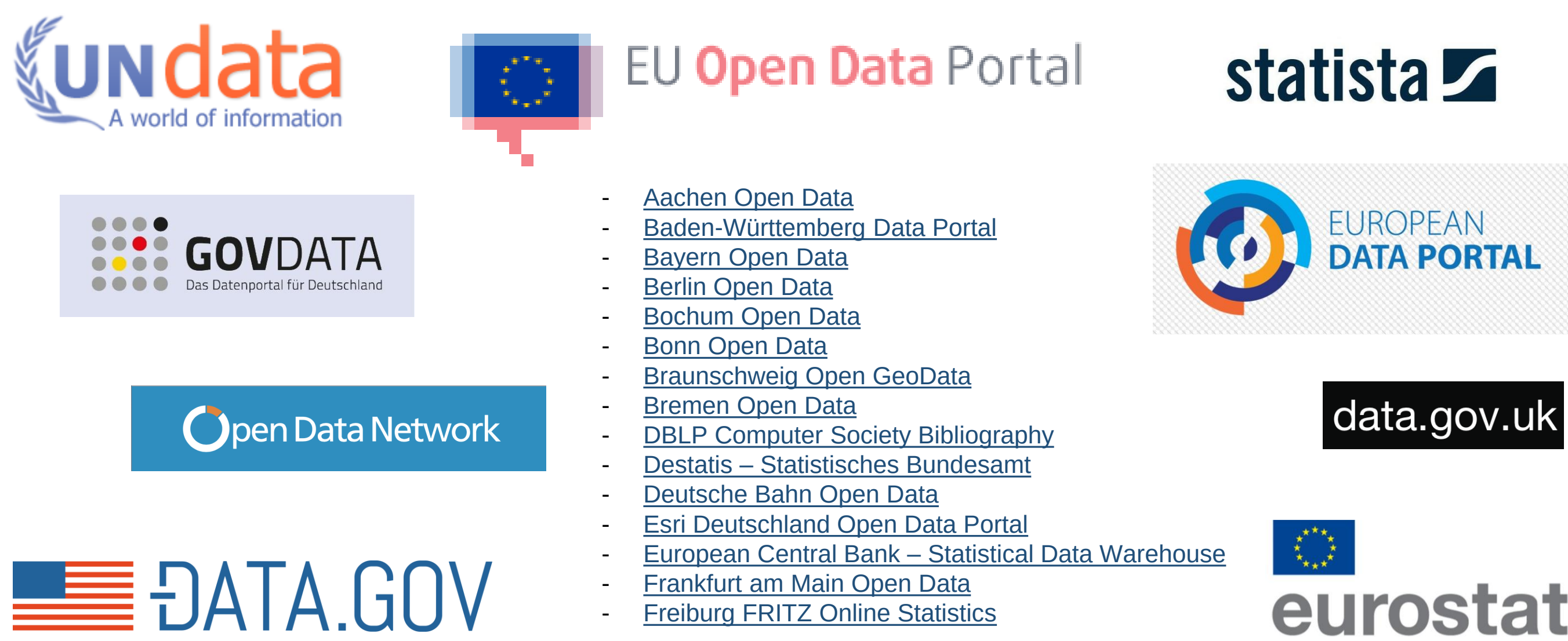
Edwin Yaqub¹, Ralf Klinkenberg¹, Benedikt Kleppmann², Christian Bizer²

¹ RapidMiner, Germany

² Data and Web Science Group, University of Mannheim

Opportunity: Wealth of Data

Problem: Finding Data and Manual Integration

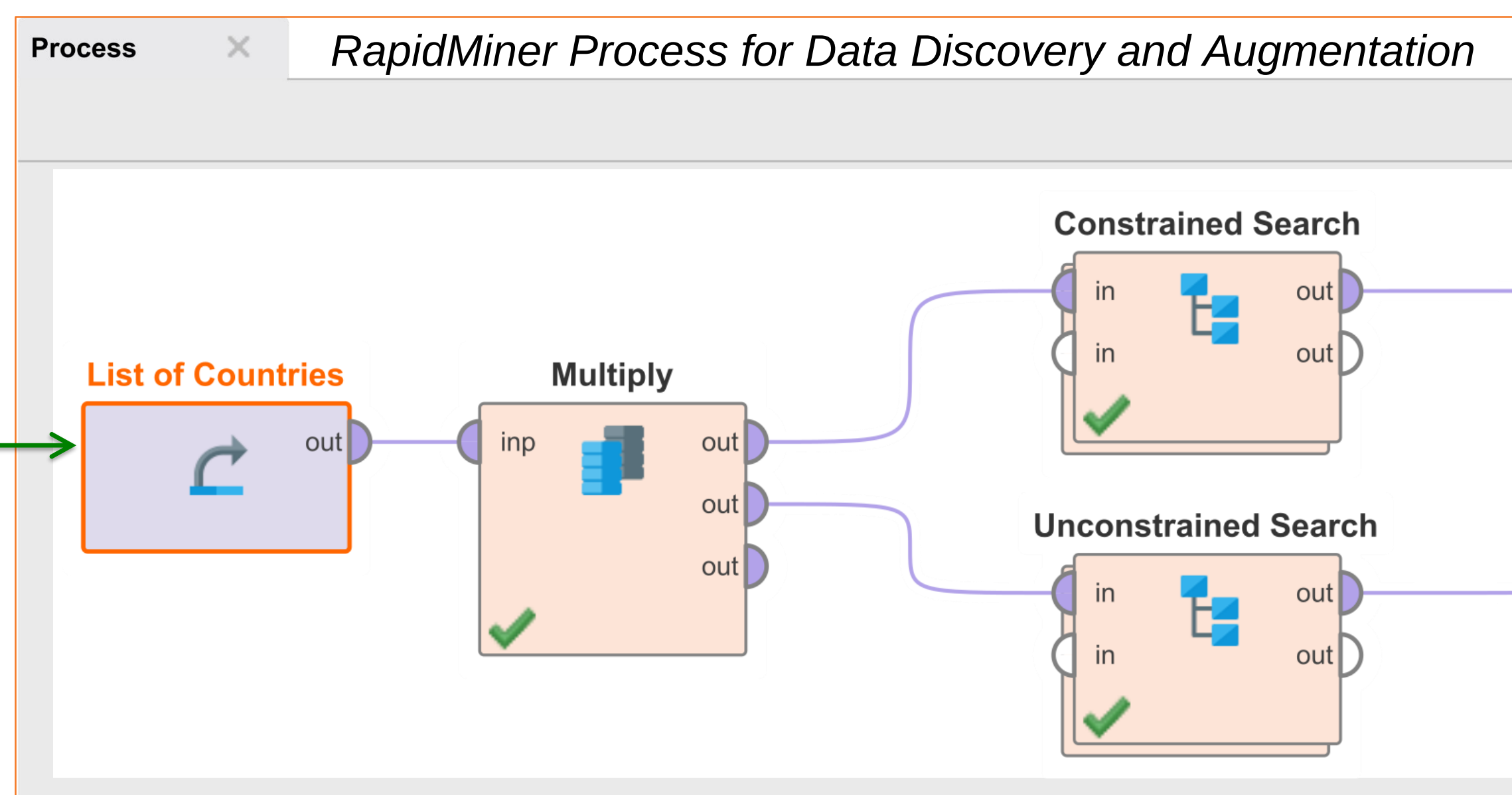


Solution: (Semi-)Automatic Data Discovery and Augmentation

Input Table

Subject Attribute

Row No.	Country	GDP (US\$MM)	Region
1	Albania	15,289	Europe
2	Argentina	625,921	South America
3	Barbados	5,317	Central America/Caribbean
4	Canada	1,798,512	North America
5	France	2,925,096	Europe
6	Germany	4,211,635	Europe
7	Israel	373,751	Middle East
8	Japan	5,167,051	Asia
9	Lebanon	53,620	Middle East
10	Mexico	1,212,831	North America
11	Pakistan	303,993	Asia
12	Romania	245,590	Europe



Extended Tables

Specifically Requested Attributes

Row No.	Country	GDP (US\$MM)	Region	Capital	Currency
1	Albania	15,289	Europe	Tirana	Lek
2	Argentina	625,921	South America	Buenos Aires	Austral (Pesu)
3	Barbados	5,317	Central America/Caribbean	Bridgetown	Dollar
4	Canada	1,798,512	North America	Ottawa	Canadian Dollar
5	France	2,925,096	Europe	Paris	Franc
6	Germany	4,211,635	Europe	Berlin	Deutsche Mark
7	Israel	373,751	Middle East	Tel Aviv	Shekel
8	Japan	5,167,051	Asia	Tokyo	Yen
9	Lebanon	53,620	Middle East	Beirut	Lebanese Pound
10	Mexico	1,212,831	North America	Mexico City	Peso
11	Pakistan	303,993	Asia	Islamabad	Pakistani Rupee
12	Romania	245,590	Europe	Bucharest	Lei

Use Cases

- Extract data from heterogenous sources and formats
- Business Intelligence (fact gathering and analysis)
- Predictive Analytics (AB test of model performance)
- Automatic labeling of data
- Easy utilization of corporate, private or pubic data

Row No.	Country	GDP (US\$MM)	Region	Latitude	Longitude	Area sq.km.	Population 2008...	FIPS	ISO (2)	ISO (3)	ISO (No)	Internet	Unnamed...	Capital
1	Albania	15,289	Europe	41.18°N	19.49°E	28748	3619778	AL	AL	ALB	8	AL	2	Tirana
2	Argentina	625,921	South America	36.30°S	60.00°W	2766890	40482000	AR	AR	ARG	32	AR	9	Buenos Aires
3	Barbados	5,317	Central America/Caribbean	13.05°N	59.30°W	431	281968	BB	BB	BRB	52	BB	19	Bridgetown
4	Canada	1,798,512	North America	45.27°N	75.42°W	9984670	33212696	CA	CA	CAN	124	CA	37	Ottawa
5	France	2,925,096	Europe	48.50°N	02.20°E	643427	64057792	FR	FR	FRA	250	FR	75	Paris
6	Germany	4,211,635	Europe	52.30°N	13.25°E	357021	82369552	GM	DE	DEU	276	DE	82	Berlin
7	Israel	373,751	Middle East	31.47°N	35.12°E	20770	7112359	IS	IL	ISR	376	IL	105	Jerusalem
8	Japan	5,167,051	Asia	?	?	377835	127288416	JA	JP	JPN	392	JP	108	Tokyo

The screenshot shows the Wikipedia page for CTrain. The page title is 'CTrain' and the subtitle is 'The free encyclopedia'. The main text describes the CTrain system as a light rail transit system in Calgary, Alberta, Canada. It mentions that the system was opened on May 25, 1981, and has since expanded to cover a route length of 59.9 kilometers (37.2 mi). The page also includes a table of operations, which lists the routes, stations, and areas served by the system.

Extracting tables from Wikipedia pages

The screenshot shows the Lucene search results for the query 'C-Train - Wikipedia, the free encyclopedia'. The results are displayed in a table format, with columns for the document ID, the document title, and the document content. The first result is for the document 'C-Train - Wikipedia, the free encyclopedia', which is the page shown in the previous screenshot. The second result is for the document 'C-Train - Wikipedia, the free encyclopedia', which is a duplicate of the first result.

The Lucene logo is shown, along with the text 'Data Indexing' and 'Data Set Search facilities via Web service'. Below the text is a globe icon with a gear, representing the search and indexing process.

Standardization of each table
(and persistence to JSON format)

Table Indexing

