Extending RapidMiner with Data Search and Integration Capabilities

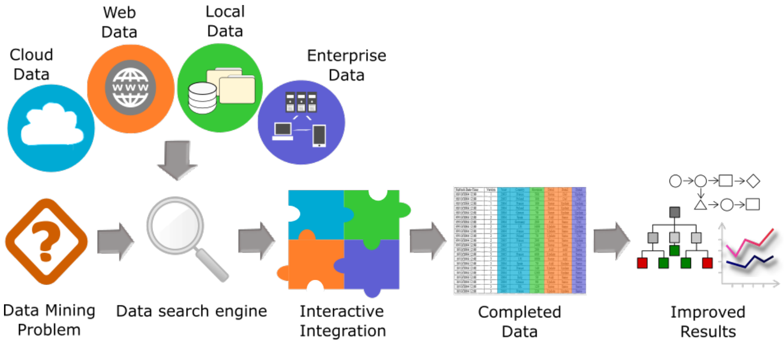
Anna Lisa Gentile[1], Sabrina Kirstein[2], Heiko Paulheim[1], and Christian Bizer[1]

[1] Data and Web Science Group, University of Mannheim

[2] RapidMiner, Germany

44

The Vision



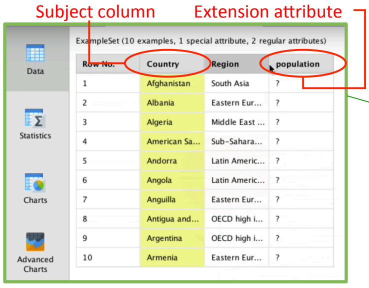
PROBLEM: data mining projects require data which exists somewhere (on the Web or in an organization's intranet) but is difficult to find.

CURRENT SOLUTIONS: state of the art data mining tools offer a wide range of powerful data mining methods but hardly support analysts in searching for suitable data as well as in integrating data from multiple sources.

CONTRIBUTION: The RapidMiner Data Search extension enables analysts to search for relevant datasets and integrate discovered data with data that they already know.

Table Search and Table Extension

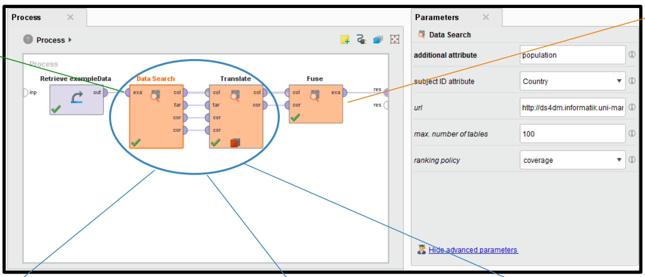
Query table



ExampleSet (10 examples, 1 special attribute, 2 regular attributes)

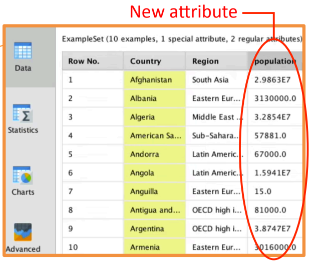
Row No.	Country	Region	population
1	Afghanistan	South Asia	?
2	Albania	Eastern Eur...	?
3	Algeria	Middle East ...	?
4	American Sa...	Sub-Sahara...	?
5	Andorra	Latin Americ...	?
6	Angola	Latin Americ...	?
7	Anguilla	Eastern Eur...	?
8	Antigua and...	OECD high i...	?
9	Argentina	OECD high i...	?
10	Armenia	Eastern Eur...	?

RapidMiner Data Search Extension



additional attribute	population
subject ID attribute	Country
url	http://ds4dm.informatik.uni-mar
max. number of tables	100
ranking policy	coverage

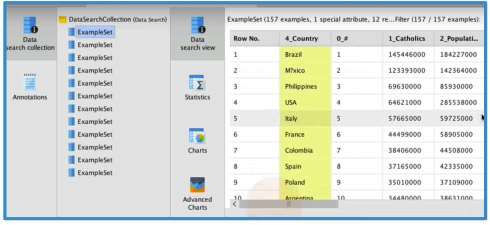
Extended Query Table



ExampleSet (10 examples, 1 special attribute, 2 regular attributes)

Row No.	Country	Region	population
1	Afghanistan	South Asia	2.9863E7
2	Albania	Eastern Eur...	3130000.0
3	Algeria	Middle East ...	3.2854E7
4	American Sa...	Sub-Sahara...	57881.0
5	Andorra	Latin Americ...	67000.0
6	Angola	Latin Americ...	1.5941E7
7	Anguilla	Eastern Eur...	15.0
8	Antigua and...	OECD high i...	81000.0
9	Argentina	OECD high i...	3.8747E7
10	Armenia	Eastern Eur...	3016000.0

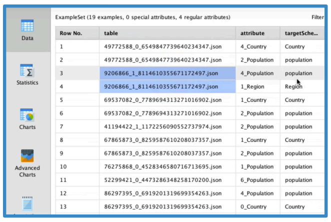
Candidate tables



ExampleSet (157 examples, 1 special attribute, 12 re...Filter (157 / 157 examples):

Row No.	4_Country	0_#	1_Catholics	2_Populati...
1	Brazil	1	145446000	184227000
2	M?xico	2	123393000	142364000
3	Philippines	3	69630000	85930000
4	USA	4	64621000	285538000
5	Italy	5	57665000	59725000
6	France	6	44499000	58905000
7	Colombia	7	38406000	44508000
8	Spain	8	37165000	42335000
9	Poland	9	35010000	37109000
10	Argentina	10	34480000	38631000

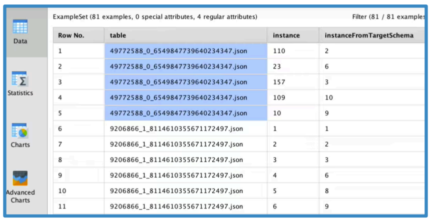
Schema level correspondences



ExampleSet (19 examples, 0 special attributes, 4 regular attributes)

Row No.	table	attribute	targetSche...
1	49772588_0_6549847739640234347.json	4_Country	Country
2	49772588_0_6549847739640234347.json	2_Population	population
3	9206866_1_8114610355671172497.json	4_Population	population
4	9206866_1_8114610355671172497.json	1_Region	Region
5	69537082_0_7789694313271016902.json	1_Country	Country
6	69537082_0_7789694313271016902.json	2_Population	population
7	41194422_1_1172256090552737974.json	2_Population	population
8	67865873_0_8259587610208037357.json	1_Country	Country
9	67865873_0_8259587610208037357.json	2_Population	population
10	76275868_0_4528346580716713695.json	1_Population	population
11	52299421_0_4473286348258170200.json	6_Population	population
12	86297395_0_6919201319699354263.json	4_Population	population
13	86297395_0_6919201319699354263.json	0_Country	Country

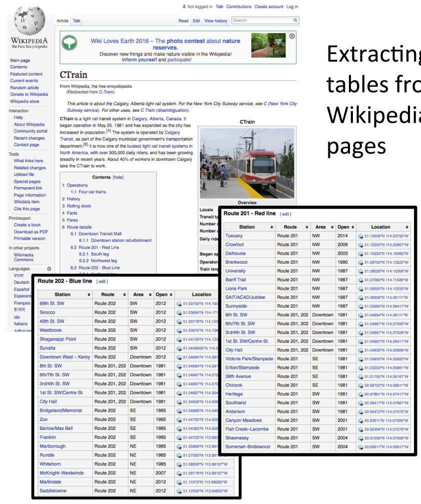
Instance level correspondences



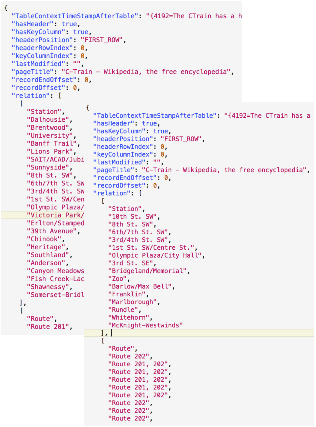
ExampleSet (81 examples, 0 special attributes, 4 regular attributes) Filter (81 / 81 examples

Row No.	table	instance	instanceFromTargetSchema
1	49772588_0_6549847739640234347.json	110	2
2	49772588_0_6549847739640234347.json	23	6
3	49772588_0_6549847739640234347.json	157	3
4	49772588_0_6549847739640234347.json	109	10
5	49772588_0_6549847739640234347.json	10	9
6	9206866_1_8114610355671172497.json	1	1
7	9206866_1_8114610355671172497.json	2	2
8	9206866_1_8114610355671172497.json	3	3
9	9206866_1_8114610355671172497.json	4	6
10	9206866_1_8114610355671172497.json	5	8
11	9206866_1_8114610355671172497.json	6	9

Wikipedia Table Processing



Extracting tables from Wikipedia pages



```
{
  "TableContextTimeStampAfterTable": "{4192=The CTrain has a h
  "hasHeader": true,
  "hasKeyColumn": true,
  "headerPosition": "FIRST_ROW",
  "headerRowIndex": 0,
  "keyColumnIndex": 0,
  "lastModified": "",
  "pageTitle": "C-Train - Wikipedia, the free encyclopedia",
  "recordEndOffset": 0,
  "recordOffset": 0,
  "relation": [
    [
      "Station",
      "Dalhousie",
      "Brentwood",
      "University",
      "Banff Trail",
      "Lions Park",
      "SAIT/ACAD/Jubi
      "Sunnyside",
      "8th St. SW",
      "6th/7th St. SW
      "3rd/4th St. SW
      "1st St. SW/Cen
      "Olympic Plaza/
      "Victoria Park/
      "Erlton/Stamped
      "39th Avenue",
      "Chinook",
      "Heritage",
      "Southland",
      "Anderson",
      "Canyon Meadows
      "Fish Creek-Lac
      "Shawnessy",
      "Somerset-Bridl
    ],
    [
      "Route",
      "Route 201",
```

```
{
  "TableContextTimeStampAfterTable": "{4192=The CTrain has a
  "hasHeader": true,
  "hasKeyColumn": true,
  "headerPosition": "FIRST_ROW",
  "headerRowIndex": 0,
  "keyColumnIndex": 0,
  "lastModified": "",
  "pageTitle": "C-Train - Wikipedia, the free encyclopedia",
  "recordEndOffset": 0,
  "recordOffset": 0,
  "relation": [
    [
      "Station",
      "10th St. SW",
      "8th St. SW",
      "6th/7th St. SW",
      "3rd/4th St. SW",
      "1st St. SW/Centre St.",
      "Olympic Plaza/City Hall",
      "3rd St. SE",
      "Bridgeland/Memorial",
      "Zoo",
      "Barlow/Max Bell",
      "Franklin",
      "Marlborough",
      "Rundle",
      "Whitehorn",
      "McKnight-Westwinds"
    ],
    [
      "Route",
      "Route 202",
      "Route 201, 202",
      "Route 201, 202",
      "Route 201, 202",
      "Route 201, 202",
      "Route 201, 202",
      "Route 202",
      "Route 202",
      "Route 202",
```

Standardization of each table (and persistence to JSON format)

Table Indexing

Lucene

Data Indexing

Search and Join facilities via Web service

UNIVERSITÄT MANNHEIM

rapidminer

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

DS4DM
Data Search for Data Mining

http://ds4dm.de