

# Realizing Smart Data by automating Tabular Search, Integration and Extraction methods

(Data Search for Data Mining - DS4DM project)

October 12th, 2017

Dr. Edwin Yaqub - Philipp Schlunder - RapidMiner Research

- Pretext
  - Challenge: Too many data sources, document formats and systems
  - Opportunity: More data can lead to smart data if it's accessible
  - Objective: Empower Data Scientists with easy to use tools for:
    - Data Discovery
    - Data Extraction

# Presentation Layout

- Data Discovery
  - How many web tables are data tables?
    - Google study (2008): < 1.1% of 14.1 billion web tables are data tables
  - What software machinery is needed?
    - Back-end: Search Engine technology for table search is in infancy
    - Front-end: Tools to discover data and integrate it with existing table

# Presentation Layout

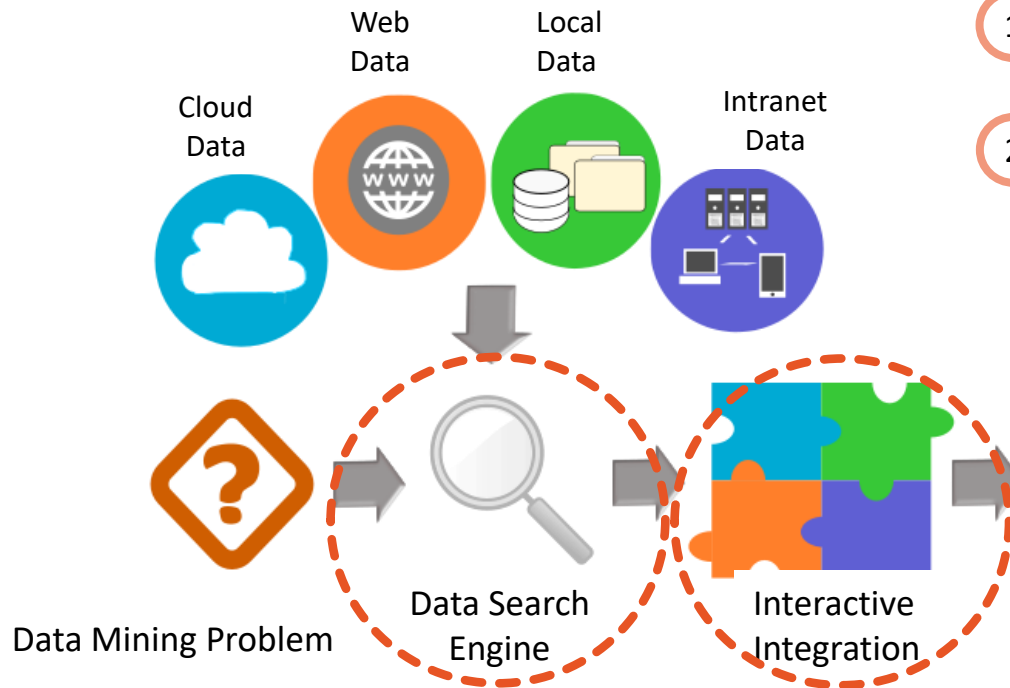
- Data Extraction
  - Target # 1: Document formats
    - Develop generic solutions (easy to use, broadly applicable)
  - Target # 2: Data stores
    - Develop provider-specific solutions (API based)

# Presentation Layout

- Tangible outcomes available as RapidMiner extensions
  - Data Search for Data Mining Extension
  - Web Table Extraction Extension
  - PDF Table Extraction Extension
  - Spreadsheet Table Extraction Extension
  - Sharepoint Connector Extension

# Data Discovery

# Overview of Data Search Extension

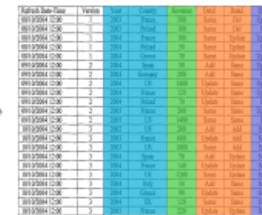


1

Extract and Ingest public/private data into a Data Search Engine

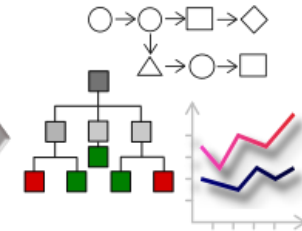
2

Discover new data from Search Engine and Integrate it to existing data interactively



| Dataset Name | Variable    | Value    |
|--------------|-------------|----------|
| Dataset 1    | Variable 1  | Value 1  |
| Dataset 2    | Variable 2  | Value 2  |
| Dataset 3    | Variable 3  | Value 3  |
| Dataset 4    | Variable 4  | Value 4  |
| Dataset 5    | Variable 5  | Value 5  |
| Dataset 6    | Variable 6  | Value 6  |
| Dataset 7    | Variable 7  | Value 7  |
| Dataset 8    | Variable 8  | Value 8  |
| Dataset 9    | Variable 9  | Value 9  |
| Dataset 10   | Variable 10 | Value 10 |
| Dataset 11   | Variable 11 | Value 11 |
| Dataset 12   | Variable 12 | Value 12 |
| Dataset 13   | Variable 13 | Value 13 |
| Dataset 14   | Variable 14 | Value 14 |
| Dataset 15   | Variable 15 | Value 15 |
| Dataset 16   | Variable 16 | Value 16 |
| Dataset 17   | Variable 17 | Value 17 |
| Dataset 18   | Variable 18 | Value 18 |
| Dataset 19   | Variable 19 | Value 19 |
| Dataset 20   | Variable 20 | Value 20 |

Enriched Data Set

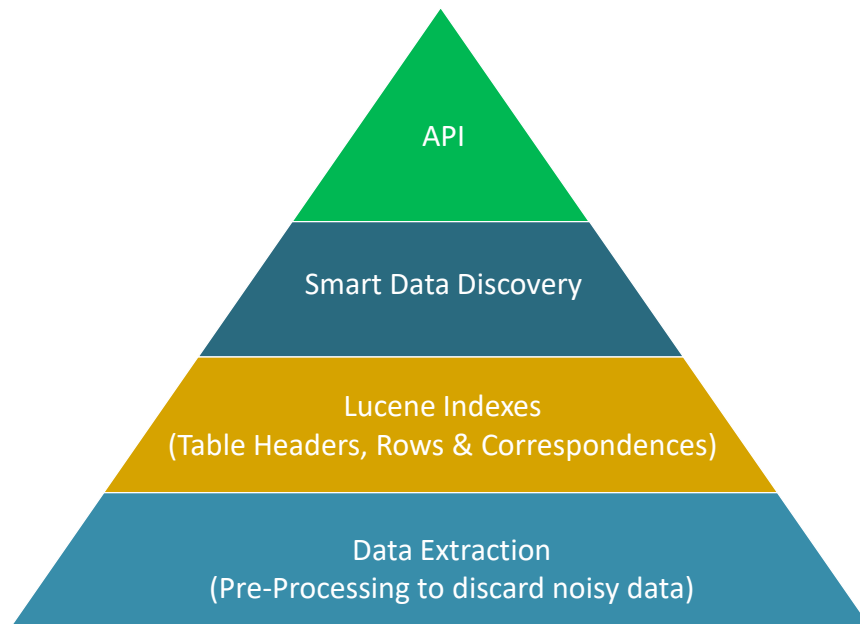


Better Results

GEFÖRDET VOM

# Data Search Engine

- Backend: Layers

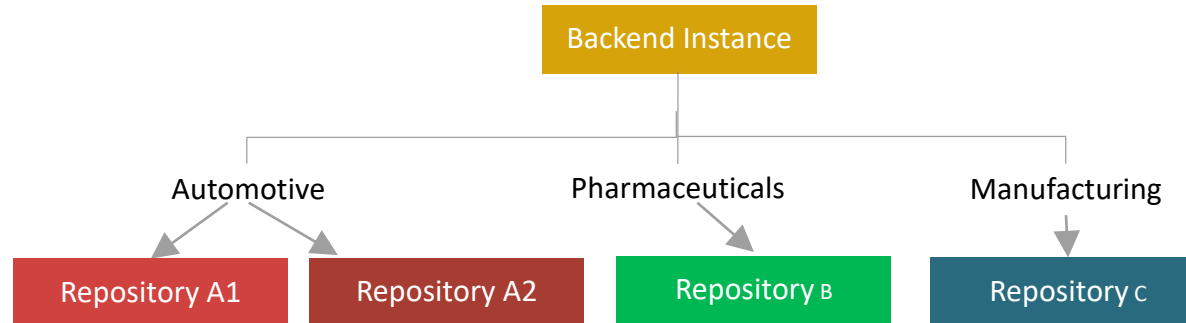


- Webservices manage account and repositories
- Easily invoke data search in models/apps
- Schema Correspondences
- Instance Correspondences
- Table Headers
- Table Rows
- Sources: Wikipedia, Eurostat, etc.
- Formats: CSV, XML, XLSX, PDF, HTML,
- Stores: Corporate (RDBMS, Sharepoint, SAP)



# Data Search Engine

- Backend: User Repository Structure



API

/register

/createRepository

/fetchRepositoryNames

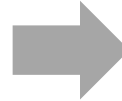
/uploadDataTable

/bulkDataUpload

# Data Search Extension

- *Search-Join* data enrichment method

| Country       | Region        | Language | currency |
|---------------|---------------|----------|----------|
| United States | North America | English  | ?        |
| Germany       | Europe        | German   | ?        |
| Japan         | Asia          | Japanese | ?        |
| China         | Asia          | Mandarin | ?        |



| Country       | Region        | Language | currency  |
|---------------|---------------|----------|-----------|
| United States | North America | English  | US Dollar |
| Germany       | Europe        | German   | Euro      |
| Japan         | Asia          | Japanese | Yen       |
| China         | Asia          | Mandarin | Renminbi  |

# Data Search Problem

- Search: *relevant* tables to extend a table by new attributes

| Country       | Region        | Language | currency |
|---------------|---------------|----------|----------|
| United States | North America | English  | ?        |
| Germany       | Europe        | German   | ?        |
| Japan         | Asia          | Japanese | ?        |
| China         | Asia          | Mandarin | ?        |

/search



– Schema Correspondences

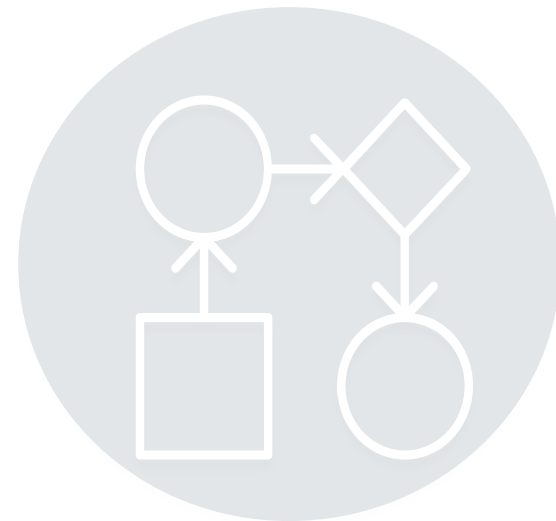


– Instance Correspondences



# Data Search Problem

- Schema Correspondences
  - Matches based on table header
    - Trivial: keyword matching on Lucene index 1
    - Non-Trivial: refine results using distance metrics
    - Other criteria: attribute data type



# Data Search Problem

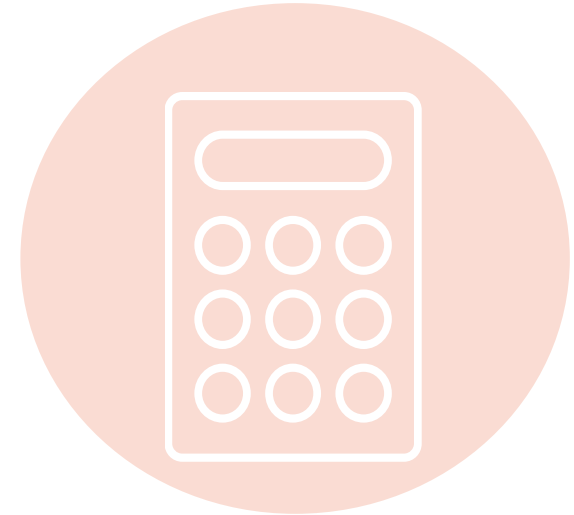
- Schema Correspondences

ExampleSet (10 examples, 0 special attributes, 4 regular attributes)

| Row No. | table                                            | attribute  | targetSchema |
|---------|--------------------------------------------------|------------|--------------|
| 1       | Public_news_channels_337114_672808.csv.gz        | 0_country  | Country      |
| 2       | DVD_107416_215553.csv.gz                         | 9_region   | Region       |
| 3       | DVD_107416_215553.csv.gz                         | 6_language | language     |
| 4       | DVD_107416_215553.csv.gz                         | 0_country  | Country      |
| 5       | Localization_150699_303570.csv.gz                | 2_language | language     |
| 6       | Localization_150699_303570.csv.gz                | 0_country  | Country      |
| 7       | _258630_490773.csv.gz                            | 4_language | language     |
| 8       | _258630_490773.csv.gz                            | 0_country  | Country      |
| 9       | International_broadcasters_601759_1302258.csv.gz | 2_language | language     |
| 10      | International_broadcasters_601759_1302258.csv.gz | 0_country  | Country      |

# Data Search Problem

- Instance Correspondences
  - Matches based on instance (example)
    - Trivial: keyword matching on Lucene index 2
    - Non-Trivial: refine results using distance metrics



# Data Search Problem

- Instance Correspondences

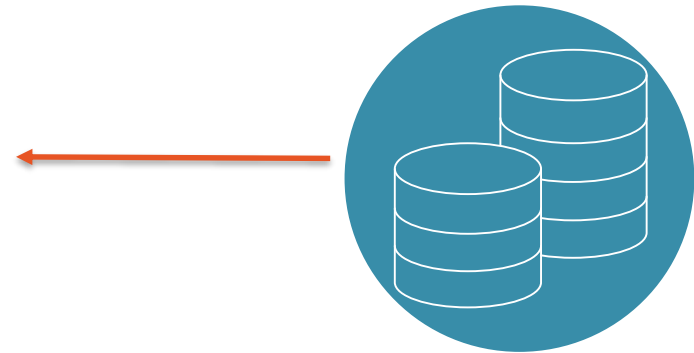
| Row No. | table                                     | instance | instanceFromTargetSchema | content                                      |
|---------|-------------------------------------------|----------|--------------------------|----------------------------------------------|
| 1       | Public_news_channels_337114_672808.csv.gz | 33       | 6                        | 1_network=HispanTV,2_language(s)=Spanish,... |
| 2       | Public_news_channels_337114_672808.csv.gz | 34       | 6                        | 1_network=Al-Alam News Network,2_language... |
| 3       | Public_news_channels_337114_672808.csv.gz | 56       | 2                        | 1_network=VOA TV,2_language(s)=English (V... |
| 4       | Public_news_channels_337114_672808.csv.gz | 24       | 7                        | 1_network=DW-TV,2_language(s)=German E...    |
| 5       | Public_news_channels_337114_672808.csv.gz | 35       | 6                        | 1_network=Sahar TV,2_language(s)=English ... |
| 6       | Public_news_channels_337114_672808.csv.gz | 25       | 7                        | 1_network=EinsExtra,2_language(s)=German,... |
| 7       | Public_news_channels_337114_672808.csv.gz | 15       | 12                       | 1_network=CBC News Network,2_language(s)...  |
| 8       | Public_news_channels_337114_672808.csv.gz | 16       | 11                       | 1_network=CCTV News,2_language(s)=Englis...  |
| 9       | Public_news_channels_337114_672808.csv.gz | 27       | 9                        | 1_network=France 24,2_language(s)=French ... |
| 10      | Public_news_channels_337114_672808.csv.gz | 38       | 10                       | 1_network=PTV News,2_language(s)=Urdu an...  |

# Data Search Problem

- Fetch collection of relevant tables

| Dominion                  | Continents    | Population | Currency   |
|---------------------------|---------------|------------|------------|
| Countries                 | Area          | GDP        | currency   |
| States                    | Regions       | Language   | Money used |
| Countries                 | Continents    | ...        | currency   |
| USA                       | North America |            | US Dollar  |
| Deutschland               | Europe        |            | Euro       |
| Japan                     | Asia          |            | ?          |
| Peoples Republic of China | Asia          |            | Renminbi   |

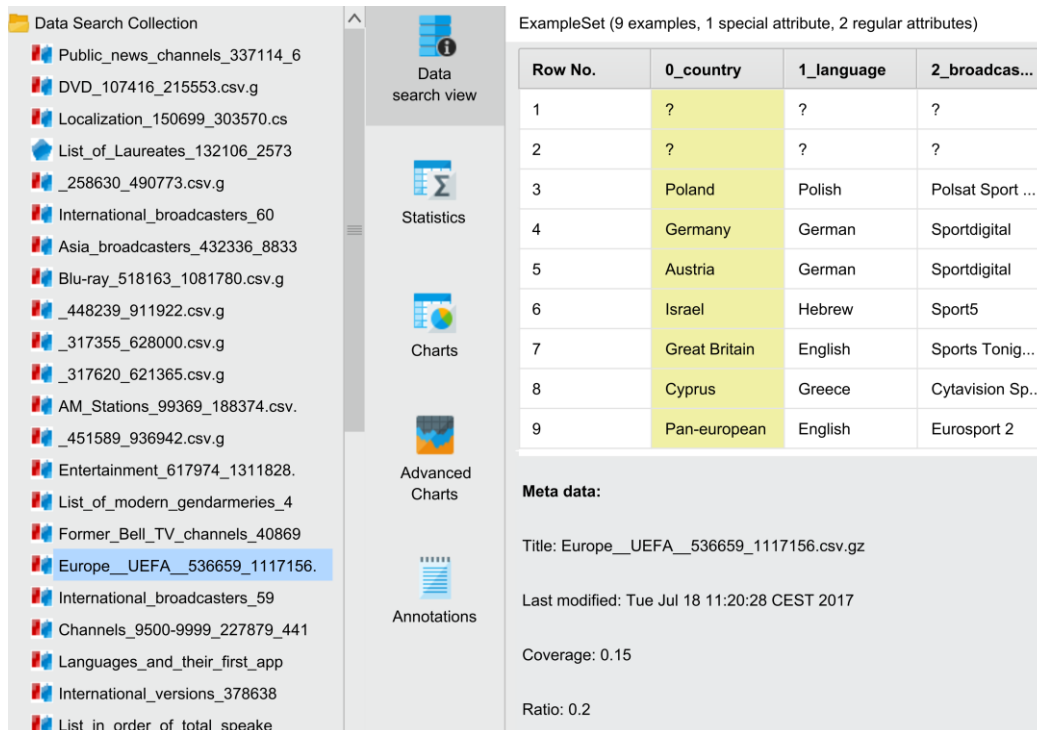
/fetchTable





# Data Search Problem

- Relevant tables discovered by Search are susceptible to noise



Data Search Collection

- Public\_news\_channels\_337114\_6
- DVD\_107416\_215553.csv.g
- Localization\_150699\_303570.cs
- List\_of\_Laureates\_132106\_2573
- \_258630\_490773.csv.g
- International\_broadcasters\_60
- Asia\_broadcasters\_432336\_8833
- Blu-ray\_518163\_1081780.csv.g
- \_448239\_911922.csv.g
- \_317355\_628000.csv.g
- \_317620\_621365.csv.g
- AM\_Stations\_99369\_188374.csv.
- \_451589\_936942.csv.g
- Entertainment\_617974\_1311828.
- List\_of\_modern\_gendarmeries\_4
- Former\_Bell\_TV\_channels\_40869
- Europe\_\_UEFA\_\_536659\_1117156.
- International\_broadcasters\_59
- Channels\_9500-9999\_227879\_441
- Languages\_and\_their\_first\_app
- International\_versions\_378638
- List\_in\_order\_of\_total\_speake

Data search view

Statistics

Charts

Advanced Charts

Annotations

ExampleSet (9 examples, 1 special attribute, 2 regular attributes)

| Row No. | 0_country     | 1_language | 2_broadcas...    |
|---------|---------------|------------|------------------|
| 1       | ?             | ?          | ?                |
| 2       | ?             | ?          | ?                |
| 3       | Poland        | Polish     | Polsat Sport ... |
| 4       | Germany       | German     | Sportdigital     |
| 5       | Austria       | German     | Sportdigital     |
| 6       | Israel        | Hebrew     | Sport5           |
| 7       | Great Britain | English    | Sports Tonig...  |
| 8       | Cyprus        | Greece     | Cytavision Sp... |
| 9       | Pan-european  | English    | Eurosport 2      |

**Meta data:**

Title: Europe\_\_UEFA\_\_536659\_1117156.csv.gz

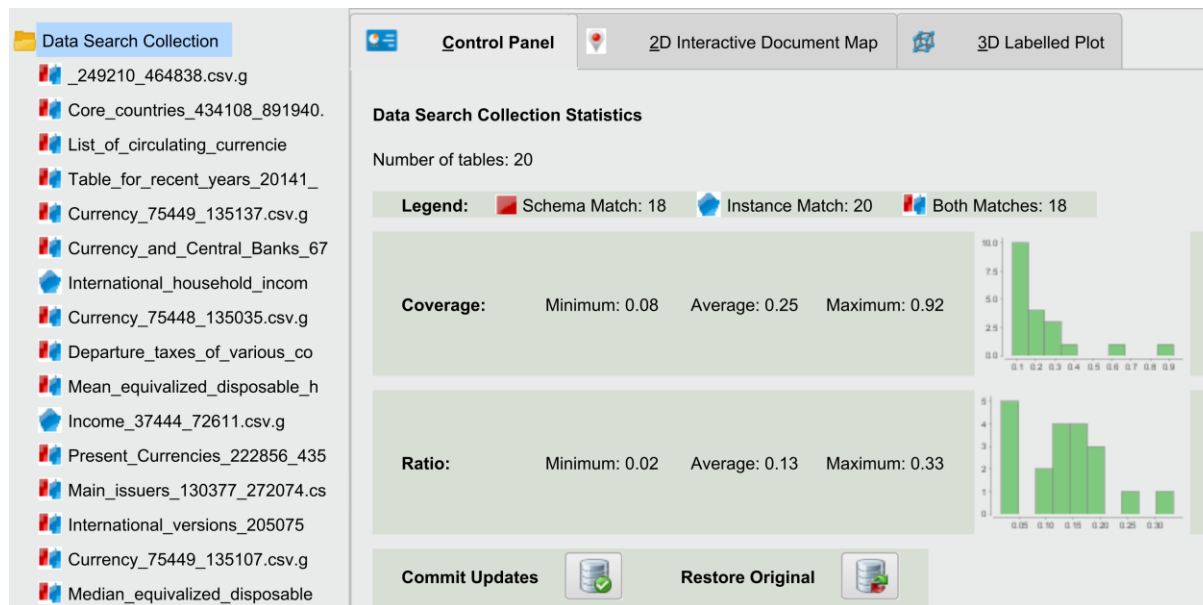
Last modified: Tue Jul 18 11:20:28 CEST 2017

Coverage: 0.15

Ratio: 0.2

# Data Search Problem

- Control Panel view: distribution of key statistics and legends



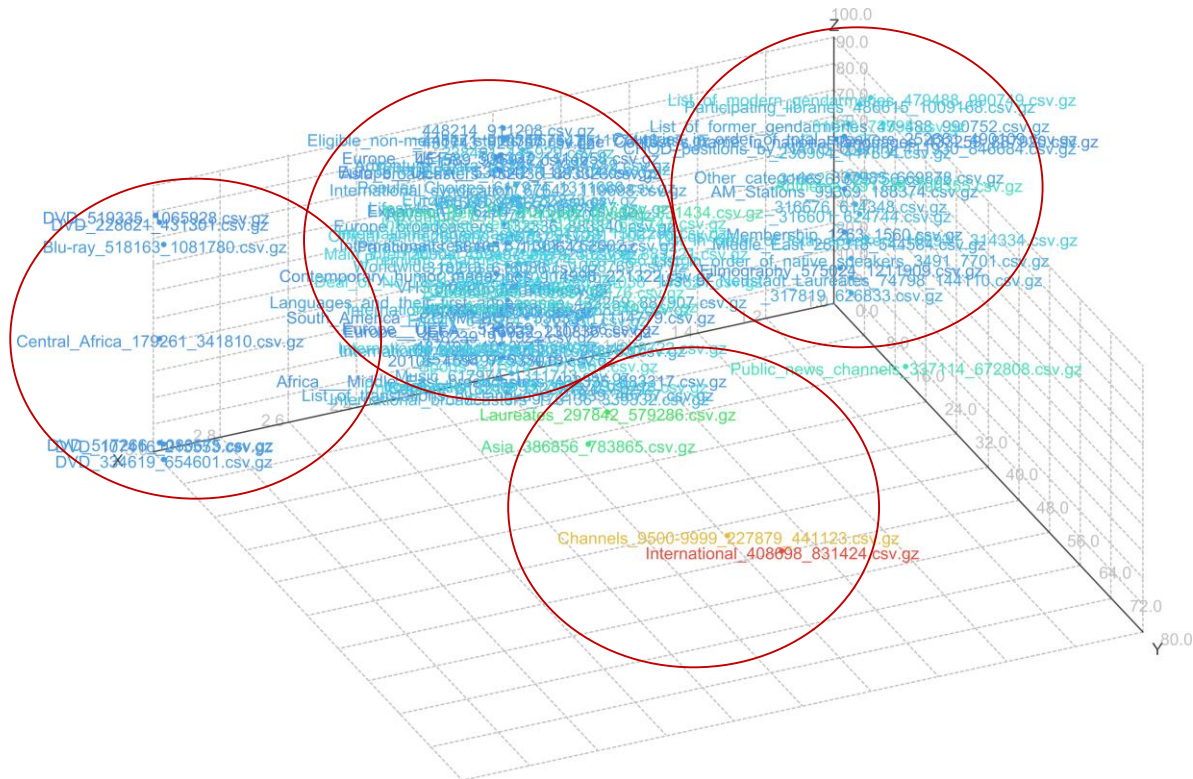
# Data Search Problem

- Map view: Self Organizing Document Map with drill-down



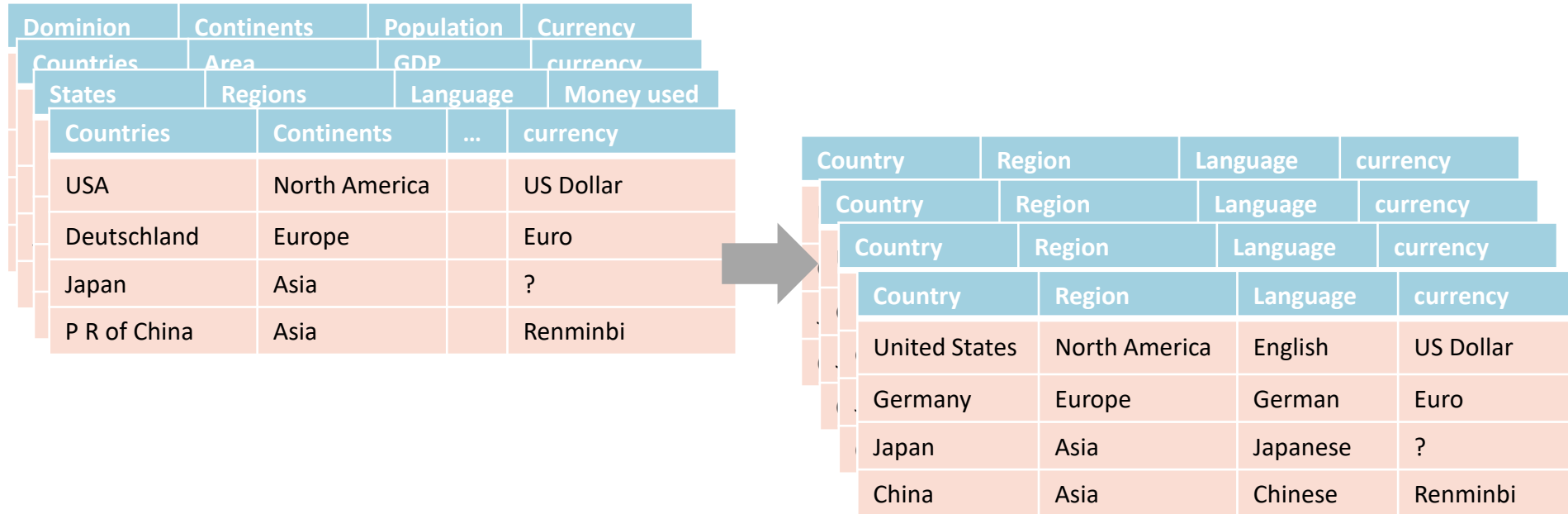
# Data Search Problem

- 3D view: Labelled Scatter Plot



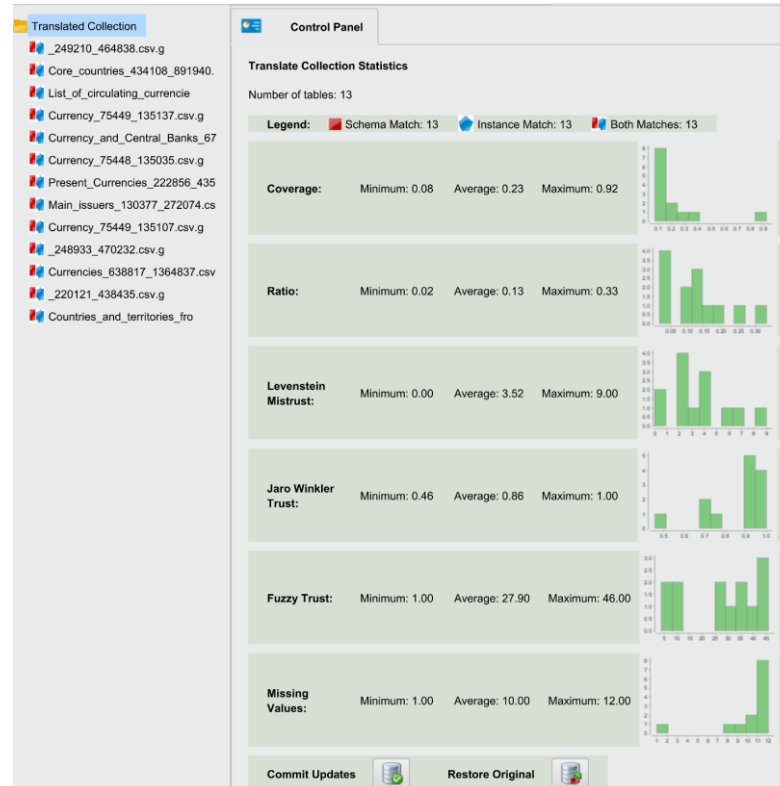
# Data Integration Problem

- Join: Translate schemas of relevant tables to join new attribute with query table



# Data Integration Problem

- Translated tables offer instance-level data for new attribute



# Data Integration Problem

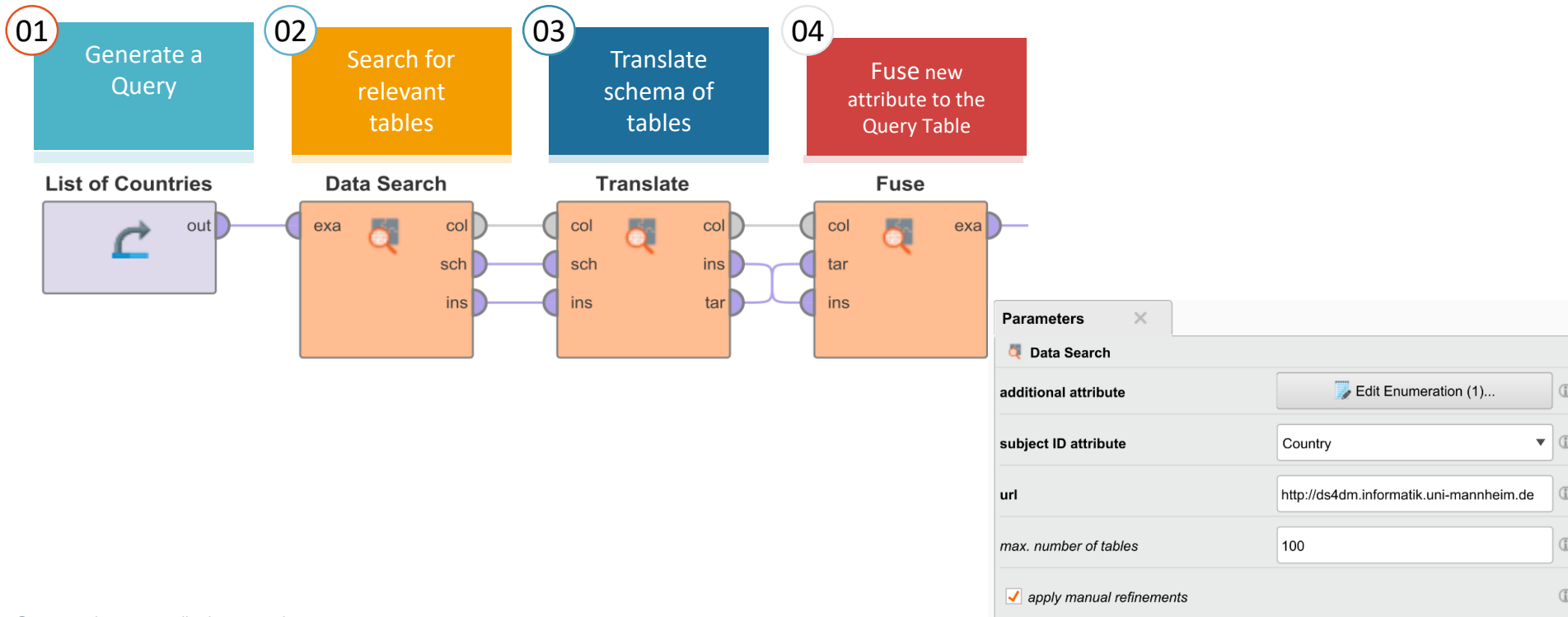
- Join: Fuse instance-level attribute values (conflict resolution)

| Country       | Region        | Language | currency  |
|---------------|---------------|----------|-----------|
| United States | North America | English  | US Dollar |
| Germany       | Europe        | German   | Euro      |
| Japan         | Asia          | Japanese | Yen       |
| China         | Asia          | Mandarin | Renminbi  |

- Fusion policies
  - First fill, Max fill, vote, clustered vote, min, max, average, multi-objective optimization

# Data Search Extension

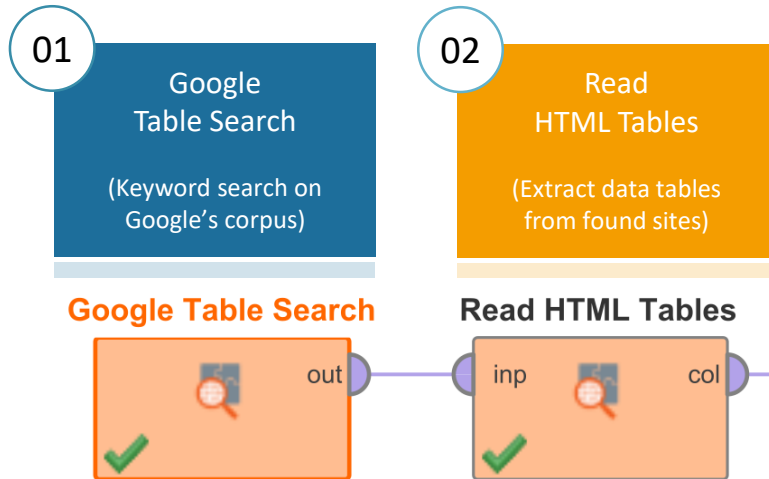
- Search-Join method consolidates all steps in 1 workflow





# Data Search Extension

- Keyword based table search on Google (webtables corpus)



| Parameters                 |                       |
|----------------------------|-----------------------|
| <b>Google Table Search</b> |                       |
| tabular corpus             | Google Web Tables     |
| keywords                   | RapidMiner Datasience |
| pagination start           | 1                     |
| pagination end             | 20                    |

# Making data available

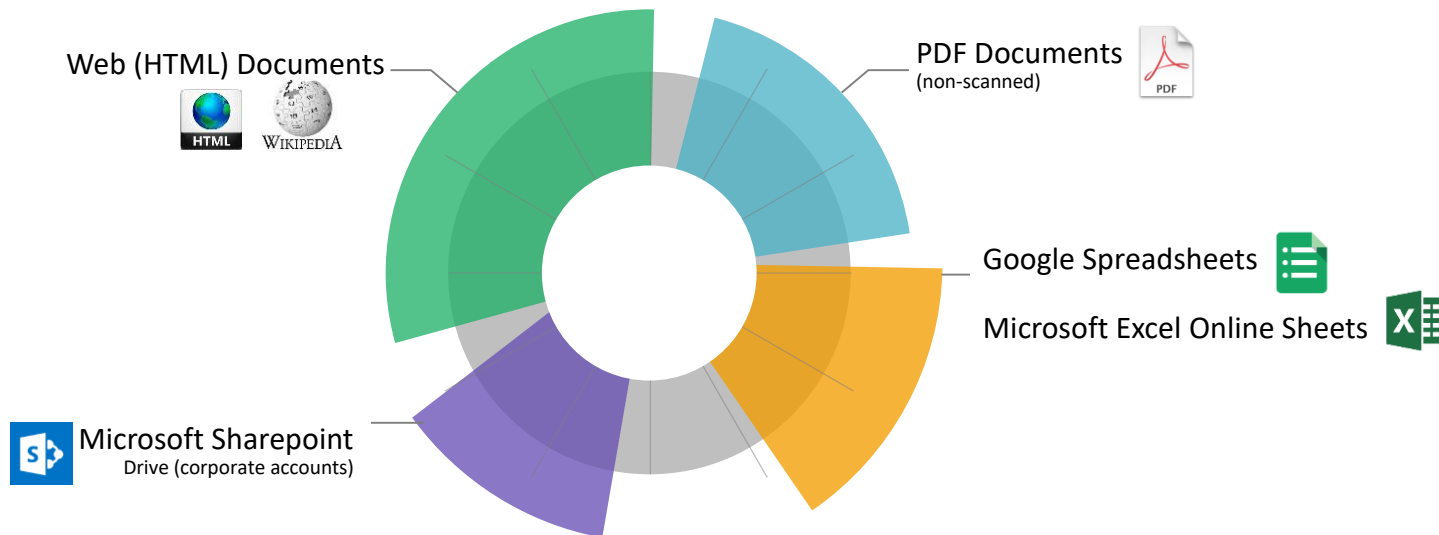
# Making internal & external data available

Making **internal & external** data available  
disregarding the **format**

Discovering **internal & external** data available  
disregarding the **format**

# Data Extraction

# Data Extraction Targets



# Web Table Extraction Extension

- Read HTML Table operator (generic extraction, no scraping)

1

Wiki page with data tables



| Categories               | Brazil | Russia | India | China |
|--------------------------|--------|--------|-------|-------|
| Area                     | 5th    | 1st    | 7th   | 3rd   |
| Population               | 5th    | 9th    | 2nd   | 1st   |
| Population growth rate   | 131st  | 192nd  | 89th  | 152nd |
| Labour force             | 5th    | 7th    | 2nd   | 1st   |
| GDP (nominal)            | 7th    | 10th   | 7th   | 2nd   |
| GDP (PPP)                | 7th    | 6th    | 3rd   | 1st   |
| GDP (nominal) per capita | 57th   | 47th   | 139th | 86th  |
| GDP (PPP) per capita     | 75th   | 55th   | 126th | 94th  |
| GDP (real) growth rate   | 180th  | 177th  | 11th  | 12th  |
| Human Development Index  | 85th   | 55th   | 114th | 101st |
| Exports                  | 6th    | 8th    | 12th  | 1st   |

2

Read HTML page or file. Classify <table> as data table on the fly.

Read HTML Table



Parameters

Read HTML Table

resource type: URL

url: https://en.wikipedia.org/wiki/BRIC

3

Convert data tables to Example Sets

IOObjectCollection ( Read HTML Table )

- ExampleSet
- ExampleSet
- ExampleSet
- ExampleSet
- ExampleSet
- ExampleSet
- ExampleSet
- ExampleSet

Data

Statistics

Charts

Advanced Charts

ExampleSet (27 examples, 0 special attributes, 5 regular attributes)

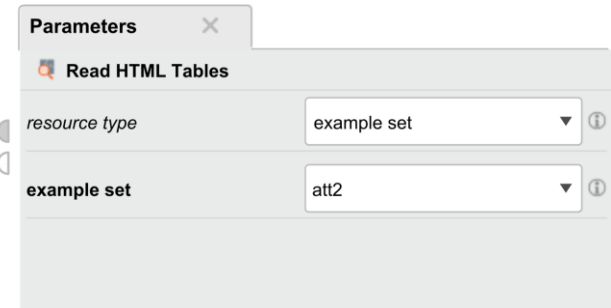
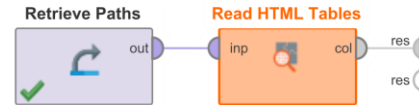
| Row No. | Categories               | Brazil | Russia | India  | China  |
|---------|--------------------------|--------|--------|--------|--------|
| 1       | Area                     | 005th  | 001st  | 007th  | 003rd  |
| 2       | Population               | 005th  | 009th  | 002nd  | 001st  |
| 3       | Population growth rate   | 131st  | 192nd  | 089th  | 152nd  |
| 4       | Labour force             | 005th  | 007th  | 002nd  | 001st  |
| 5       | GDP (nominal)            | 007th  | 010th  | 07th   | 002nd  |
| 6       | GDP (PPP)                | 007th  | 006th  | 003rd  | 001st  |
| 7       | GDP (nominal) per capita | 057th  | 047th  | 139th  | 086th  |
| 8       | GDP (PPP) per capita     | 075th  | 055th  | 126th  | 094th  |
| 9       | GDP (real) growth rate   | 0180th | 0177th | 0011th | 0012th |
| 10      | Human Development Index  | 085th  | 055th  | 114th  | 0101st |
| 11      | Exports                  | 06th   | 08th   | 012th  | 001st  |



# Web Table Extraction Extension

- Read HTML Tables operator
  - Bulk extraction of data tables

1 Input ExampleSet with file paths or URLs



The screenshot shows the 'Parameters' window for the 'Read HTML Tables' operator. It contains two main settings:

- resource type**: A dropdown menu set to 'example set'.
- example set**: A dropdown menu set to 'att2'.

2 Read paths, extract data tables and convert them to Example Sets

ExampleSet (4 examples, 1 special attribute, 2 regular attributes)

| Row No. | att1 | att2                                                                                                                           | att3 |
|---------|------|--------------------------------------------------------------------------------------------------------------------------------|------|
| 1       | ?    | C:\work\R&D-Projects\DS4DM\Extension\Web Table Extraction Test HTML>List of countries by real GDP growth rate - Wikipedia.html | ?    |
| 2       | ?    | C:\work\R&D-Projects\DS4DM\Extension\Web Table Extraction Test HTML\Statewide Detail for Child Day Care Centers and Homes.html | ?    |
| 3       | ?    | C:\work\R&D-Projects\DS4DM\Extension\Web Table Extraction Test HTML\view-source_https___en.wikipedia.org_wiki_BRIC.html        | ?    |
| 4       | ?    | https://en.wikipedia.org/wiki/Iris_flower_data_set                                                                             | ?    |

# PDF Table Extraction Extension

- Read PDF Table operator (generic method, no data scraping)

1

A PDF file with data table (may have surrounding text)

2

Provide file path or URL

Read PDF Table



Parameters

Read PDF Table

resource type

URL

url

https://bitbucket.org/ds4dm/repository-of-pdf/schools.pdf

3

Data tables are extracted from PDF as Example Sets

IOObjectCollection (Read PDF Table)

- ExampleSet
- ExampleSet
- ExampleSet
- ExampleSet
- ExampleSet

Data

Statistics

Charts

Advanced Charts

Annotations

ExampleSet (45 examples, 0 special attributes, 10 regular attributes)

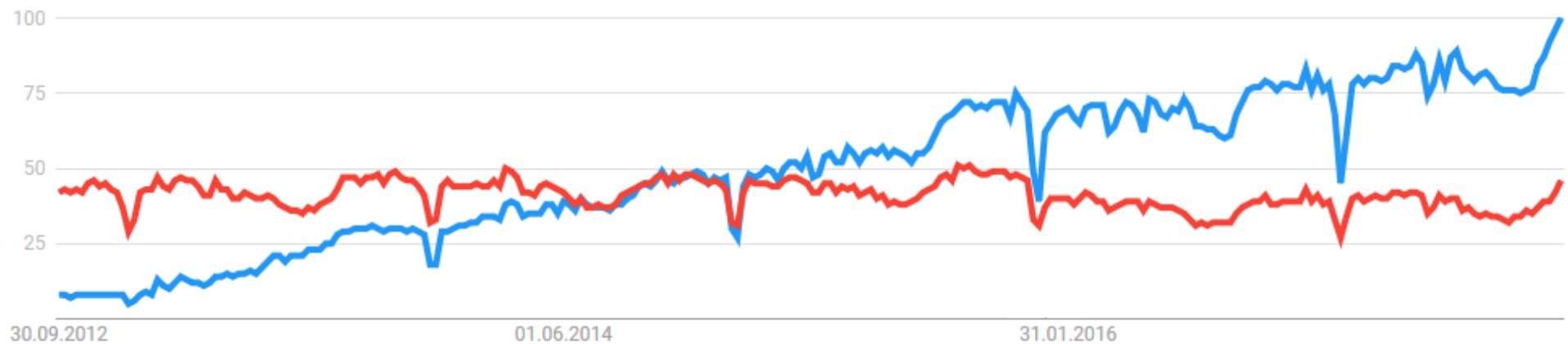
| Row No. | Last Name        | First Name         | Address           | City        | State | Zip   |
|---------|------------------|--------------------|-------------------|-------------|-------|-------|
| 1       | Filice           | Gregory A          | 15 Crocus Pl...   | Saint Paul  | MN    | 55102 |
| 2       | Fisher           | Thomas R.          | 2188 Hendon...    | Saint Paul  | MN    | 55108 |
| 3       | Flynn            | Nancy              | 1726 St Clair ... | St Paul     | MN    | 55105 |
| 4       | Fredrickson      | Jeanette           | 1889 Summit...    | Saint Paul  | MN    | 55105 |
| 5       | Gerhardstein ... | Dennis /Sarah      | 1498 Fairmo...    | Saint Paul  | MN    | 55105 |
| 6       | Guthmann         | Howard & Elizabeth | 683 Wentwort...   | Saint Paul  | MN    | 55118 |
| 7       | Halunen          | Clayton D.         | 7 Heather Pla...  | Saint Paul  | MN    | 55102 |
| 8       | Hengefelt        | Meike C. & Robert  | 756 Ashland ...   | Saint Paul  | MN    | 55104 |
| 9       | Hobbie           | Ann                | 1209 Portlan...   | Saint Paul  | MN    | 55104 |
| 10      | Hunt             | Ruby M.            | 1840 Univers...   | Saint Paul  | MN    | 55104 |
| 11      | Johnson          | Martin             | 1287 SCHEF...     | St. Paul    | Mn    | 55116 |
| 12      | Johnson          | Jana Sue           | 1497 Fulham...    | Saint Paul  | MN    | 55108 |
| 13      | Kanavati         | Louis              |                   |             |       |       |
| 14      | Kayser           | Tom & Marlene      | 800 La Salle ...  | Minneapolis | MN    | 55402 |

Vote Yes for Strong Schools  
Major Contribution Report  
For the Period Between August 1, 2012 to October 18, 2012

| Last Name                           | First Name            | Address                         | City            | State | Zip        | Occupation            | Employer                       | Date       | Amount     |
|-------------------------------------|-----------------------|---------------------------------|-----------------|-------|------------|-----------------------|--------------------------------|------------|------------|
| D'Aquila                            | James A & Bonnie R    | 6 Crocus Hill                   | Saint Paul      | MN    | 55102      |                       |                                | 9/18/2012  | 500.00     |
| Dady                                | Michael               | 1415 Summit Ave.                | Saint Paul      | MN    | 55105      |                       |                                | 9/14/2012  | 500.00     |
| Dist. 66 DFL Club                   | John                  | 1455 Simpson Street             | Saint Paul      | MN    | 55108      | Filter, Treasurer     |                                | 10/3/2012  | 500.00     |
| Gillett                             | Thomas                | 419 Roosevelt Ave               | Eau Claire      | WI    | 54701      | President             | Indianhead Foodservice Distrib | 9/28/2012  | 500.00     |
| Hang                                | Maykiao Y & Lao Lu    | 4751 Stinson Trail              | Woodbury        | MN    | 55129      | CEO                   | Wilder                         | 10/3/2012  | 500.00     |
| Highland Park Elementary PTA        |                       | 1700 Saunders Ave               | Saint Paul      | MN    | 55116      |                       |                                | 9/17/2012  | 500.00     |
| Hmong Village Inc.                  |                       | 1001 Johnson Pkwy               | Saint Paul      | MN    | 55106      |                       |                                | 10/15/2012 | 500.00     |
| Holman                              | Tom and Kim           | 85 Langford Park                | St. Paul        | MN    | 55108      |                       | Morning Foundation             | 9/14/2012  | 500.00     |
| Kaenemer                            | Martha/Art            | 6 Crocus Hill                   | St Paul         | MN    | 55102      | business owner/pedest | Retired                        | 9/14/2012  | 500.00     |
| Maavo Systems Inc                   |                       | 4330 Centerville Road           | White Bear Lake | MN    | 55127      | Private Organization  |                                | 8/22/2012  | 500.00     |
| McDonough                           | James R. & Carol M    | 815 Montana Ave. E.             | Saint Paul      | MN    | 55106      | commissioner          | Ramsey County                  | 9/28/2012  | 500.00     |
| Murphy                              | Rep. Erin             | 795 Folsom St, 1st floor        | San Francisco   | CA    | 94107      | Private Organization  |                                | 9/18/2012  | 500.00     |
| Pacific Education Group             |                       | 1025 Fairmount Ave              | St. Paul        | MN    | 55105      | CEO                   | Search Institute               | 9/13/2012  | 500.00     |
| Pekel                               |                       | 700 Transfer Road               | Saint Paul      | MN    | 55114      |                       |                                | 10/3/2012  | 500.00     |
| Saint Paul Pipefitters Local 455    |                       | 1455 Simpson Street             | Saint Paul      | MN    | 55108      | Filter, Treasurer     |                                | 10/4/2012  | 500.00     |
| Saint Paul Ward 4 DFL               | John                  | c/o Diane Gerth 909 Bayard Ave  | St Paul         | MN    | 55102      |                       | Committee                      | 9/20/2012  | 500.00     |
| Senate District 64 DFL              |                       | PO Box 4154                     | Saint Paul      | MN    | 55104-0154 |                       |                                | 9/17/2012  | 500.00     |
| Volunteers for John Choi            |                       | 1860 Yorkshire Ave              | Saint Paul      | MN    | 55116      | campaign committee    |                                | 9/22/2012  | 500.00     |
| Walters Recycling & Refuse Inc.     |                       | P. O. Box 67                    | Circle Pines    | MN    | 55014      |                       | Private Organization           | 8/22/2012  | 500.00     |
| Wellington                          | Katherine R           | 2157 Gordon Ave                 | Saint Paul      | MN    | 55108      | professor             |                                | 10/4/2012  | 500.00     |
| ECFE                                | Jill Johnson          | 1624 Hartford Ave               | Saint Paul      | MN    | 55116      |                       |                                | 10/5/2012  | 500.00     |
| Lesch                               | John                  | 1254 Dale St N                  | Saint Paul      | MN    | 55117      | State rep /City atty  | State of MN / City of St Paul  | 10/9/2012  | 500.00     |
| Farnsworth School Organization Inc. |                       | 1290 Arcade St.                 | Saint Paul      | MN    | 55106      |                       |                                | 9/27/2012  | 500.00     |
| Burns                               | Scott                 | 1411 Summit Ave                 | St Paul         | MN    | 55105      | CEO                   | GovDelivery                    | 8/22/2012  | 750.00     |
| SPFT                                |                       | 600 Kelly Ave Suite A           | Saint Paul      | MN    | 55102      |                       |                                | 10/12/2012 | 800.00     |
| Chris Coleman for Saint Paul        |                       | P.O. Box 14447                  | St Paul         | MN    | 55114      |                       | Committee                      | 8/20/2012  | 1,000.00   |
| Sands                               | William A. & Susan    | 1657 Questwood Dr.              | Falcon Heights  | MN    | 55113      | developer             | Nelson, Tietz & Hoye           | 9/28/2012  | 1,000.00   |
| Nelson                              | Russell C. & Nancy    | 4546 Mount Curve Blvd.          | Saint Paul      | MN    | 55116      |                       |                                | 10/10/2012 | 1,500.00   |
| Ken Powell Contributions Account    |                       |                                 |                 |       |            |                       |                                | 8/16/2012  | 2,000.00   |
| McLeely                             | H. G., Jr             | 444 Pine St                     | St Paul         | MN    | 55101      |                       | meritex Enterprise             | 10/3/2012  | 3,000.00   |
| AFSCME MN People Comm Council 5 PAC |                       | 300 Hardman Ave S.              | St Paul         | MN    | 55075      |                       |                                | 10/8/2012  | 3,500.00   |
| Nelson, Tietz & Hoye, Inc.          |                       | 81 South 9th St, Suite 330      | Minneapolis     | MN    | 55402      |                       |                                | 10/2/2012  | 5,000.00   |
| Ecobab                              | Kris Taylor           | 170 Wabasha Street              | Saint Paul      | MN    | 55103      |                       |                                | 10/9/2012  | 5,000.00   |
| CenturyLink (Clay Bailey)           | attn: Controller's Of | P.O. Box 4065                   | Monroe          | LA    | 71211      |                       |                                | 9/20/2012  | 10,000.00  |
| SPFT                                |                       |                                 |                 |       |            |                       |                                | 9/28/2012  | 15,000.00  |
| The Saint Paul Foundation           |                       | 55 Fifth Street East, Suite 600 | Saint Paul      | MN    | 55101      |                       |                                | 9/27/2012  | 25,000.00  |
|                                     |                       |                                 |                 |       |            |                       |                                |            | 111,840.00 |
|                                     |                       |                                 |                 |       |            |                       |                                |            | 17,605.00  |
|                                     |                       |                                 |                 |       |            |                       |                                |            | 120,445.00 |

# Online & Offline Data

## Excel Documents



Source: Google Trends

Microsoft Office - offline - online

# Spreadsheet Table Extraction Extension



- Imports data from
  - Google Spreadsheet sheets
  - Excel Online workbooks
- Define reading range on a cell level
  - No file download necessary
- Spreadsheets can be read from:
  - Google Drive
  - Google Docs
  - Microsoft Corporate OneDrive
  - Microsoft SharePoint Sites

# Spreadsheet Table Extraction Extension

## 1 You have an online Google Spreadsheet

Example Spreadsheet

File Edit View Insert Format Data Tools Add-ons Help

View only

Student Name

|    | A            | B      | C            | D          | E          | F       | G                 | H                   | I     | J              | K |
|----|--------------|--------|--------------|------------|------------|---------|-------------------|---------------------|-------|----------------|---|
|    | Student Name | Gender | Class Level  | Home State | Home State | Major   | Extracurricular A | Date                | Index | Decimal Number |   |
| 1  | Alexandra    | Female | 4. Senior    | CA         | CA         | English | Drama Club        | 12/30/1899 11:40:59 | 1     | 1.             |   |
| 2  | Andrew       | Male   | 1. Freshman  | SD         | SD         | Math    | Lacrosse          | 12/30/1899 11:40:59 | 1     | 1.             |   |
| 3  | Anna         | Female | 1. Freshman  | NC         | NC         | English | Basketball        | 12/30/1899 11:40:59 | 1     | 1.             |   |
| 4  | Becky        | Female | 2. Sophomore | SD         | SD         | Art     | Baseball          | 12/30/1899 11:40:59 | 1     | 1.             |   |
| 5  | Benjamin     | Male   | 4. Senior    | WI         | WI         | English | Basketball        | 12/30/1899 11:40:59 | 1     | 1.             |   |
| 6  | Carl         | Male   | 3. Junior    | MD         | MD         | Art     | Debate            | 12/30/1899 11:40:59 | 1     | 1.             |   |
| 7  | Carrie       | Female | 3. Junior    | NE         | NE         | English | Track & Field     | 12/30/1899 11:40:59 | 1     | 1.             |   |
| 8  | Dorothy      | Female | 4. Senior    | MD         | MD         | Math    | Lacrosse          | 12/30/1899 11:40:59 | 1     | 1.             |   |
| 9  | Dylan        | Male   | 1. Freshman  | MA         | MA         | Math    | Baseball          | 12/30/1899 11:40:59 | 1     | 1.             |   |
| 10 | Edward       | Male   | 3. Junior    | FL         | FL         | English | Drama Club        | 12/30/1899 11:40:59 | 1     | 2.             |   |
| 11 | Ellen        | Female | 1. Freshman  | WI         | WI         | Physics | Drama Club        | 12/30/1899 11:40:59 | 1     | 2.             |   |
| 12 | Fiona        | Female | 1. Freshman  | MA         | MA         | Art     | Debate            | 12/30/1899 11:40:59 | 1     | 2.             |   |
| 13 | John         | Male   | 3. Junior    | CA         | CA         | Physics | Basketball        | 12/30/1899 11:40:59 | 1     | 2.             |   |
| 14 | Jonathan     | Male   | 2. Sophomore | SC         | SC         | Math    | Debate            | 12/30/1899 11:40:59 | 1     | 2.             |   |
| 15 | Joseph       | Male   | 1. Freshman  | AK         | AK         | English | Drama Club        | 12/30/1899 11:40:59 | 1     | 2.             |   |
| 16 | Josephine    | Female | 1. Freshman  | NY         | NY         | Math    | Debate            | 12/30/1899 11:40:59 | 1     | 2.             |   |
| 17 | Karen        | Female | 2. Sophomore | NH         | NH         | English | Basketball        | 12/30/1899 11:40:59 | 1     | 2.             |   |
| 18 | Kevin        | Male   | 2. Sophomore | NE         | NE         | Physics | Drama Club        | 12/30/1899 11:40:59 | 1     | 2.             |   |
| 19 | Lisa         | Female | 3. Junior    | SC         | SC         | Art     | Lacrosse          | 12/30/1899 11:40:59 | 1     | 2.             |   |
| 20 | Mary         | Female | 2. Sophomore | AK         | AK         | Physics | Track & Field     | 12/30/1899 11:40:59 | 1     | 3.             |   |
| 21 | Maureen      | Female | 1. Freshman  | CA         | CA         | Physics | Basketball        | 12/30/1899 11:40:59 | 1     | 3.             |   |
| 22 | Nick         | Male   | 4. Senior    | NY         | NY         | Art     | Baseball          | 12/30/1899 11:40:59 | 1     | 3.             |   |
| 23 | Olivia       | Female | 4. Senior    | NC         | NC         | Physics | Track & Field     | 12/30/1899 11:40:59 | 1     | 3.             |   |
| 24 | Pamela       | Female | 3. Junior    | RI         | RI         | Math    | Baseball          | 12/30/1899 11:40:59 | 1     | 3.             |   |
| 25 | Patrick      | Male   | 1. Freshman  | NY         | NY         | Art     | Lacrosse          | 12/30/1899 11:40:59 | 1     | 3.             |   |
| 26 | Robert       | Male   | 1. Freshman  | CA         | CA         | English | Track & Field     | 12/30/1899 11:40:59 | 1     | 3.             |   |
| 27 | Sean         | Male   | 1. Freshman  | NH         | NH         | Physics | Track & Field     | 12/30/1899 11:40:59 | 1     | 3.             |   |
| 28 | Stacy        | Female | 1. Freshman  | NY         | NY         | Math    | Baseball          | 12/30/1899 11:40:59 | 1     | 3.             |   |
| 29 | Thomas       | Male   | 2. Sophomore | RI         | RI         | Art     | Lacrosse          | 12/30/1899 11:40:59 | 1     | 3.             |   |
| 30 | Will         | Male   | 4. Senior    | FL         | FL         | Math    | Debate            | 12/30/1899 11:40:59 | 1     | 4.             |   |

Class Data

## 2 Configure the Read Google Spreadsheet operator

100%

Read Google Spreadsheet

spreadsheet url

sheet name

client secret file

## 3 Sheet is extracted as Example Set

ExampleSet (Read Google Spreadsheet)

ExampleSet (9 examples, 0 special attributes, 11 regular attributes)

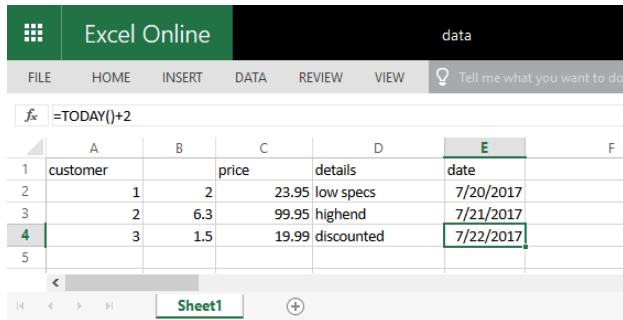
| Row No. | Student Name | Gender | Class Level  | Home State | Attribute 5 | Home State-1 | Major   | Extracurricu... | Date            | Index | Decimal Nu... |
|---------|--------------|--------|--------------|------------|-------------|--------------|---------|-----------------|-----------------|-------|---------------|
| 1       | Alexandra    | Female | 4. Senior    | CA         | CA          | CA           | English | Drama Club      | 12/30/1899 1... | 1     | 1.100         |
| 2       | Andrew       | Male   | 1. Freshman  | SD         | SD          | SD           | Math    | Lacrosse        | 12/30/1899 1... | 1     | 1.200         |
| 3       | Anna         | Female | 1. Freshman  | NC         | NC          | NC           | English | Basketball      | 12/30/1899 1... | 1     | 1.300         |
| 4       | Becky        | Female | 2. Sophomore | SD         | SD          | SD           | Art     | Baseball        | 12/30/1899 1... | 1     | 1.400         |
| 5       | Benjamin     | Male   | 4. Senior    | WI         | WI          | WI           | English | Basketball      | 12/30/1899 1... | 1     | 1.500         |
| 6       | Carl         | Male   | 3. Junior    | MD         | MD          | MD           | Art     | Debate          | 12/30/1899 1... | 1     | 1.600         |
| 7       | Carrie       | Female | 3. Junior    | NE         | NE          | NE           | English | Track & Field   | 12/30/1899 1... | 1     | 1.700         |
| 8       | Dorothy      | Female | 4. Senior    | MD         | MD          | MD           | Math    | Lacrosse        | 12/30/1899 1... | 1     | 1.800         |
| 9       | Dylan        | Male   | 1. Freshman  | MA         | MA          | MA           | Math    | Baseball        | 12/30/1899 1... | 1     | 1.900         |

# Spreadsheet Table Extraction Extension

## Excel Online

1

You have an Excel Online sheet



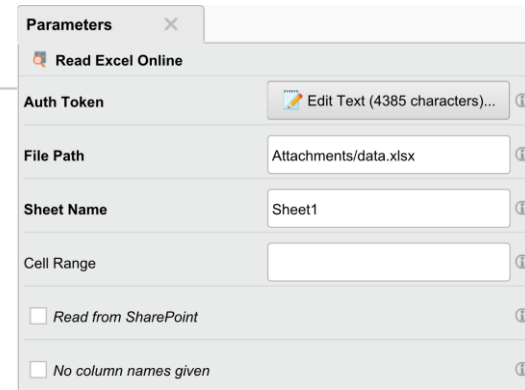
The screenshot shows an Excel Online interface with a spreadsheet titled 'data'. The formula bar shows '=TODAY()+2'. The spreadsheet has columns A through F and rows 1 through 5. The data is as follows:

|   | A        | B   | C     | D          | E         | F |
|---|----------|-----|-------|------------|-----------|---|
| 1 | customer |     | price | details    | date      |   |
| 2 | 1        | 2   | 23.95 | low specs  | 7/20/2017 |   |
| 3 | 2        | 6.3 | 99.95 | highend    | 7/21/2017 |   |
| 4 | 3        | 1.5 | 19.99 | discounted | 7/22/2017 |   |
| 5 |          |     |       |            |           |   |

2

Configure the Read Excel Online operator

Read Excel Online



The Parameters window for the Read Excel Online operator shows the following settings:

- Auth Token: Edit Text (4385 characters)...
- File Path: Attachments/data.xlsx
- Sheet Name: Sheet1
- Cell Range: (empty)
- ☐ Read from SharePoint
- ☐ No column names given

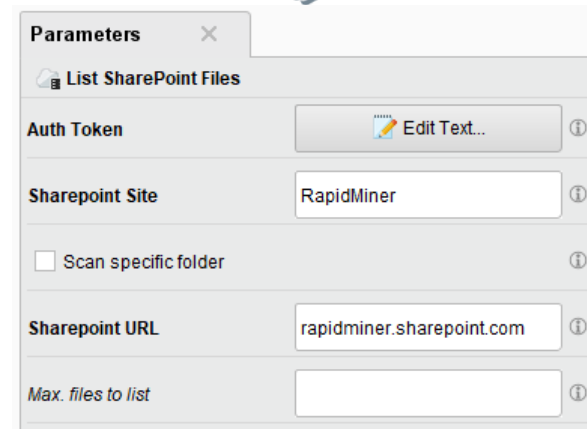
3

Sheet is extracted as Example Set

| Row No. | customer | Attribute 2 | price  | details    | date      |
|---------|----------|-------------|--------|------------|-----------|
| 1       | 1        | 2           | 23.950 | low specs  | 7/20/2017 |
| 2       | 2        | 6.300       | 99.950 | highend    | 7/21/2017 |
| 3       | 3        | 1.500       | 19.990 | discounted | 7/22/2017 |

# SharePoint Connector Extension

- Get list of files & folders from Microsoft SharePoint Site
- Download files & folders to your PC
- Uses Microsoft Graph API
- Intranet Data available for Search



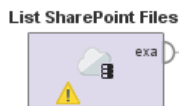
2

Configure the List SharePoint Files operator

Dokumente

Name ▾

- Big Data
- Data Science
- Design
- Funded RnD
- General



| Row No. | filename     | creationDateTime     |
|---------|--------------|----------------------|
| 1       | Big Data     | 2017-03-21T09:51:01Z |
| 2       | Data Science | 2017-03-21T09:50:28Z |
| 3       | Design       | 2017-03-13T13:42:50Z |
| 4       | Funded RnD   | 2017-02-27T16:58:42Z |
| 5       | General      | 2017-02-27T13:57:37Z |

3

Files & Folders are extracted as Example Set

...

| folder |
|--------|
| true   |
| true   |
| true   |
| true   |
| true   |

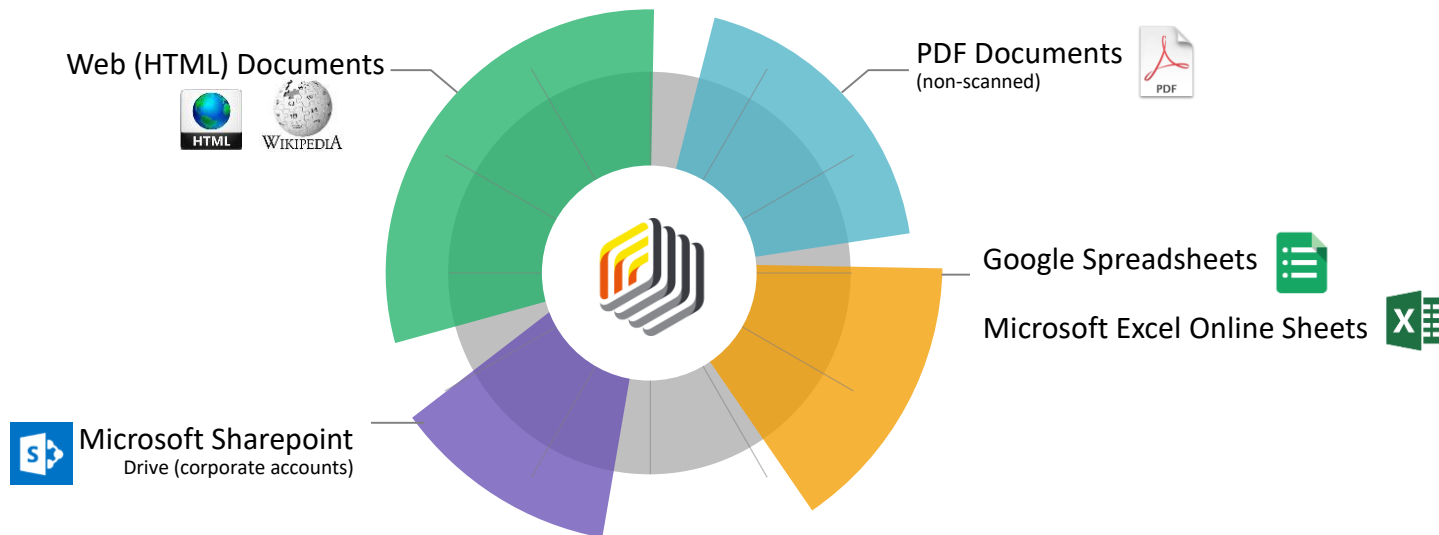


4

Download selected Files & Folders

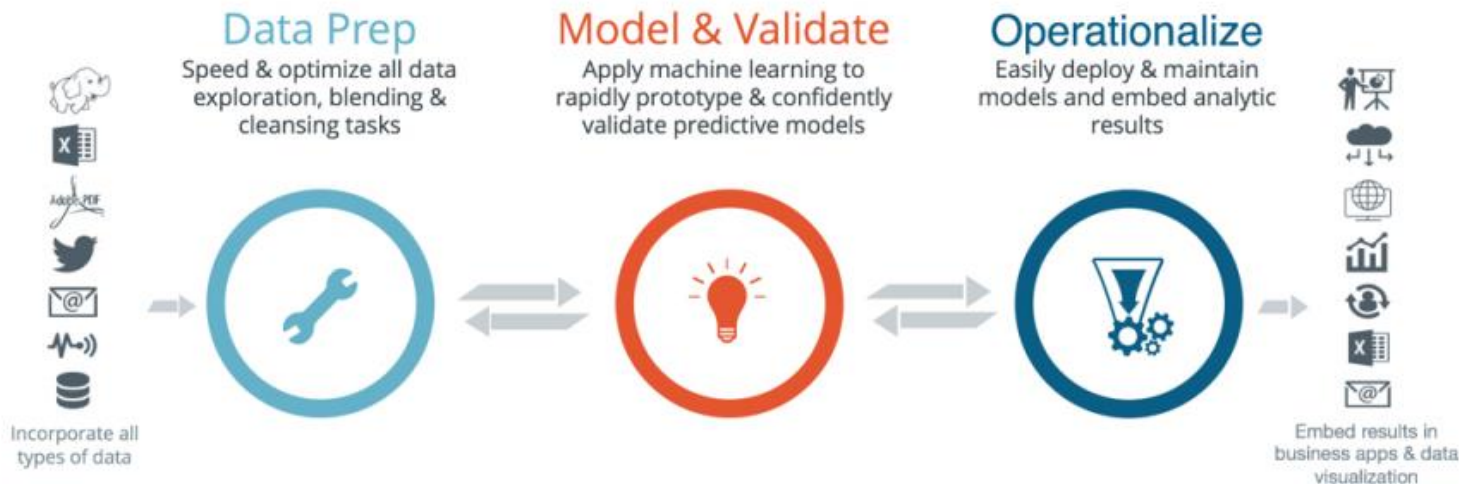
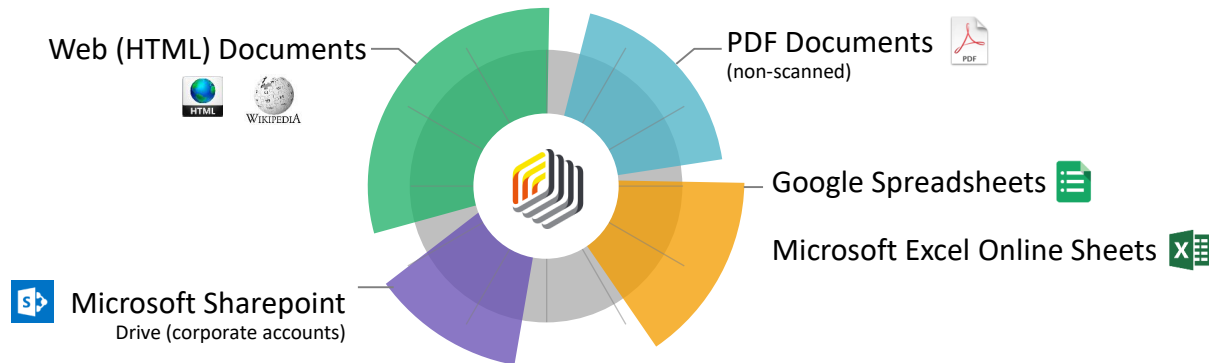
1 You have a SharePoint Site

# Data Extraction Targets





# Data Extraction Targets





[Dr. Edwin Yaqub](#)



[@byond2day](#)



[Philipp Schlunder](#)



[@WortPixel](#)

## Q & A



[inquiries@rapidminer.com](mailto:inquiries@rapidminer.com)



[www.rapidminer.com](http://www.rapidminer.com)



[@RapidMiner](#)

# Extras

- Performance metrics for Search quality
  - Precision: % of found Extension Attribute values that are correct
  - Recall: % of findable Extension Attributes that were actually found
  - F1: harmonic mean of above two values
- Results (for 11 controlled queries)
  - Current version: F1= 0,69 (Precision= 0,91; Recall= 0,55)
  - Upcoming version: F1= 0,77 (Precision= 0,89; Recall= 0,69)

# State of the Art

- Commercial offerings
  - Google Fusion Table service
  - Microsoft Power Query (plugin for Excel)
  - Tamr
- Academic
  - Senbazuru (a prototype spreadsheet database management system)