Chart Advisor

Efficient Algorithm for Recommendation of Data Visualization Tools

Group Members
Cristobal Leiva
Ahmad Amayri
Jorge Ortiz

Supervisor
Fabrizio Orlandi

universitätbonn EIS ENTERPRISE INFORMATION SYSTEMS

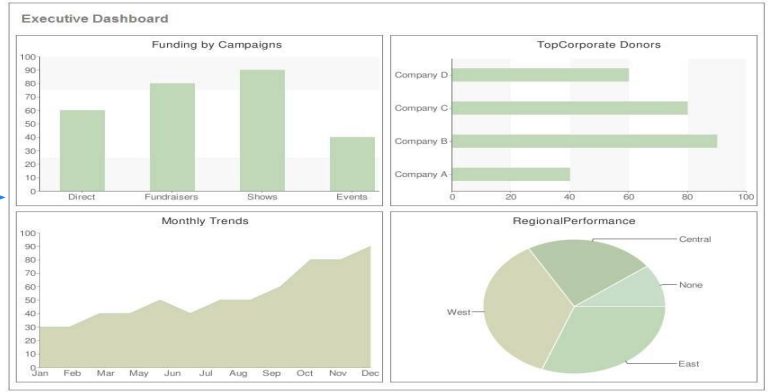
System Overview

Project:

Develop an algorithm to recommend accurate data visualization tools (Charts) based on selected Datasets / Properties

The output of the algorithm is a ranked list of recommended charts such as Bar, Bubble, Line, Geo Charts.

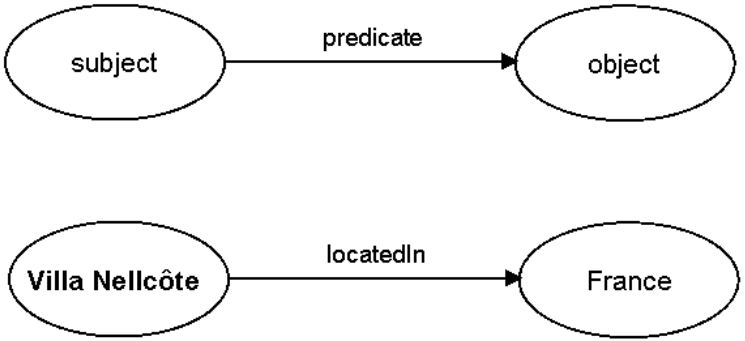
Country	Capital	Latitude	Longitude
Afghanistan	Kabul	34°28'N	69°11'E
Albania	Tirane	41°18'N	19°49'E
Algeria	Algiers	36°42'N	03°08'E
American Samoa	Pago Pago	14°16'S	170°43'W
Andorra	Andorra la Vella	42°31'N	01°32'E
Angola	Luanda	08°50'S	13°15'E
Antigua and Barbuda	W. Indies	17°20'N	61°48'W
Argentina	Buenos Aires	36°30'S	60°00'W



System Overview

Structure of the data:

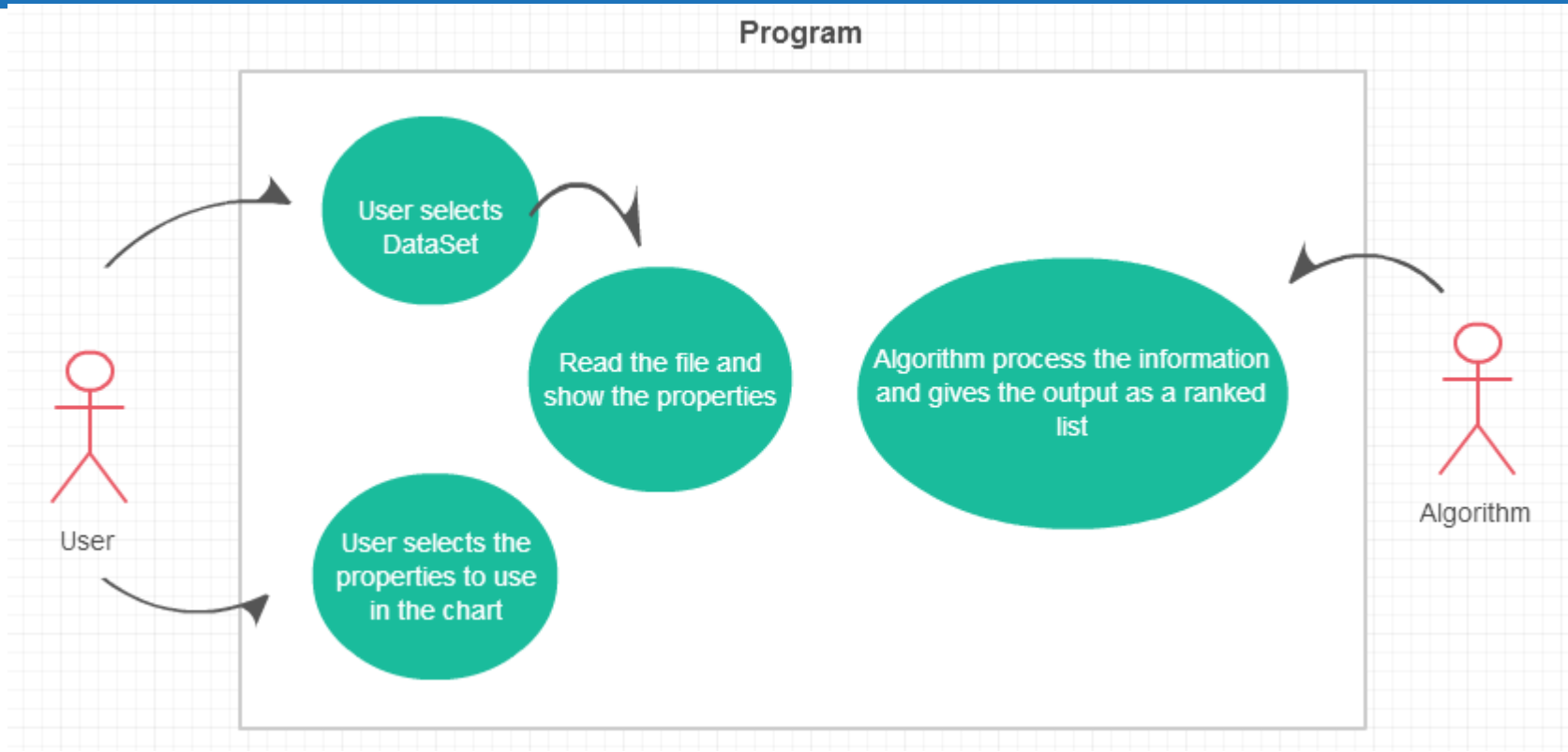
- Data Sets of triples:



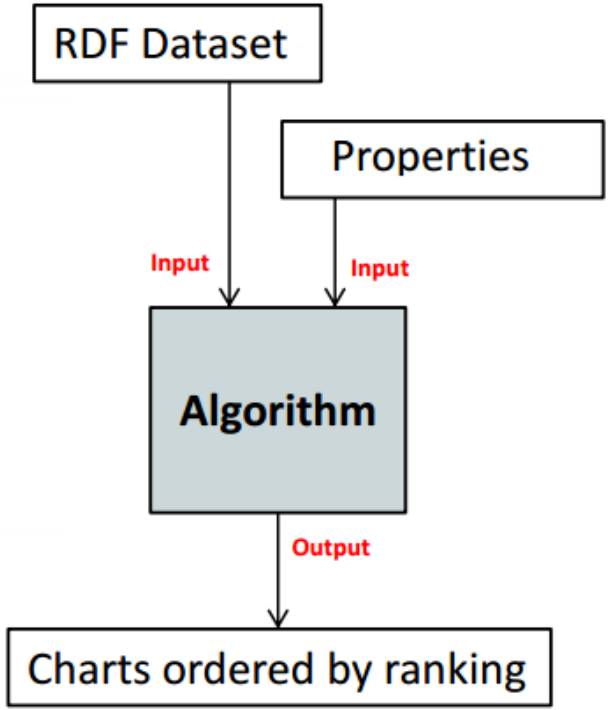
- Tabular data (CSV):

Country	Latitude	Longitude
Afghanistan	34°28'N	69°11'E
Albania	41°18'N	19°49'E

Use Case - General Scenario



Use Case - General Scenario



General steps in the algorithm

1. List and categorize properties.
2. Find out the pattern of the properties selected. (Quantitative, Qualitative or Ordinal).
3. Validation of data selected.
4. Suggest visualizations.

System Architecture

The software architecture is a model view controller.

Model.

Responsable of the data and models from the files with use of JENA libraries, which is a free and open source Java framework for building semantic web and Linked Data applications

The framework is able to process different formats of files such as .ttl, .nt, .rdf, . jsonld, .csv between others.

System Architecture

Main functions in the model.

- Create models from a file.
- Load model in other files with a different format.
- Get properties of the dataset and their data type.
- Execute sparql queries in the file.

System Architecture

View.

Graphic user interface created with the libraries of SWING JAVA.

Controller.

Core of the algorithm.

HOW IT WORKS - ALGORITHM

Input:

- List of selected properties to visualize (User Selection)

Output:

- Recommended charts names and their accuracy in (RDF, XML, TXT)
  - Accuracy: determined by the number of attributes the suggested chart can visualize.

HOW IT WORKS - ALGORITHM

Input:

ID	Population	Region	Fertility Rate	Life Expectancy
DEU	81902307	Europe	1.36	79.84
CAN	33739900	North America	1.67	80.66
DNK	5523095	Europe	1.84	78.6
...	...	...	...	...

HOW IT WORKS - ALGORITHM

Categorize Properties:

- Determine the type and level of measurement
- Dictionary.rdf

Property	Type	Level of Measurement
ID	String	Categorical
Population	Number	Quantitative
Region	String	Categorical
Fertility Rate	Number	Quantitative
Life Expectancy	Number	Quantitative

HOW IT WORKS - ALGORITHM

Generate and Validate Allocations

- Allocations are the combinations of the input properties.
  - Allocation is of the form x→y
  - Let n be the number of selected attributes, the total number of generated allocations is:

$$\sum_{k=1}^{n-1}\left(\frac{n!}{k!(n-k)!} * \sum_{m=1}^{n-k}\left(\frac{(n-k)!}{m!\,(n-k-m)!}\right)\right)$$

- For 5 selected attributes, 180 allocations are generated.

HOW IT WORKS - ALGORITHM

Generate and Validate Allocations

- Valid Allocation:
  - Left-Total: for any x in X, there is y in Y such that xRy
  - Right-Unique: xRy, xRz → y=z
- ID, Population → Fertility Rate, Life Expectancy (Valid)
- ID, Population, Region, Life Expectancy → Fertility Rate (Valid)
- ID → Fertility Rate (Valid)
- Region → Fertility Rate (Invalid- not right-unique)
- 12 out of 180 allocations are invalid and dismissed.

HOW IT WORKS - ALGORITHM

Map allocations to charts:

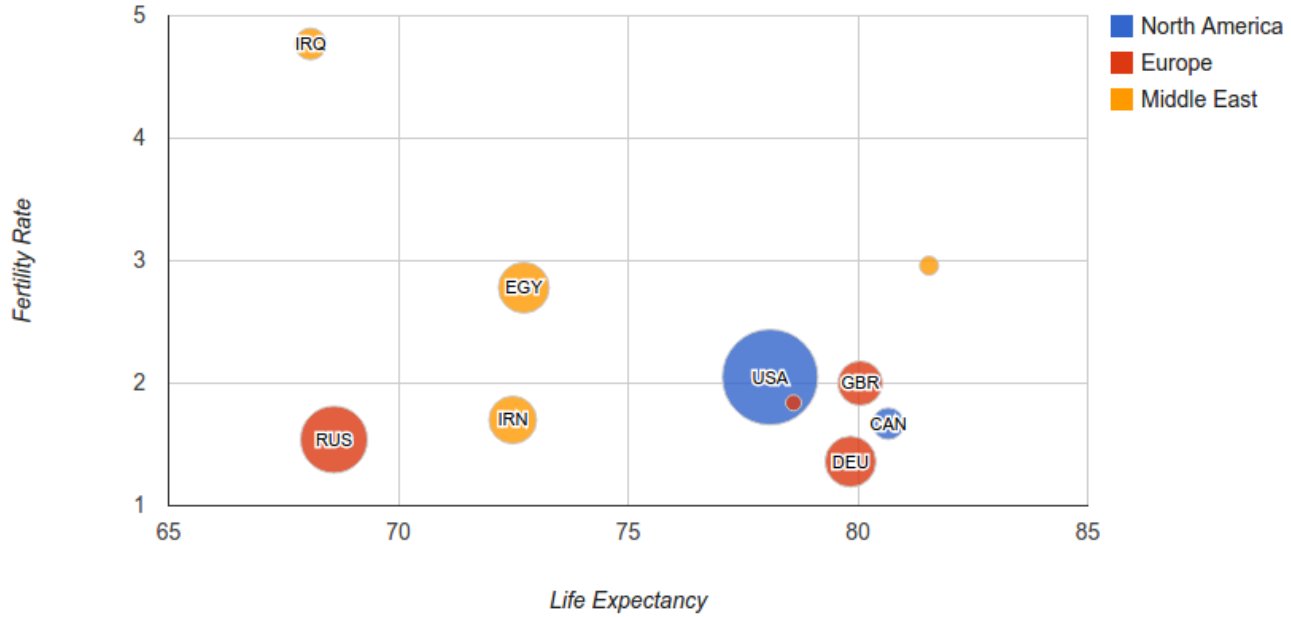
- Every chart capability is stored in chart.rdf
- Map valid allocations to existing charts
- Order by the number of mapped attributes.
  - Bubble Chart, can view all 5 properties, Accuracy 100%
  - Bar, Column Charts, can view up to 4 properties out of 5, Accuracy 80%
  - Geo Chart, can view only 2 properties out of 5, Accuracy 40%

Output:

- Save results to output file (RDF, XML, TXT)

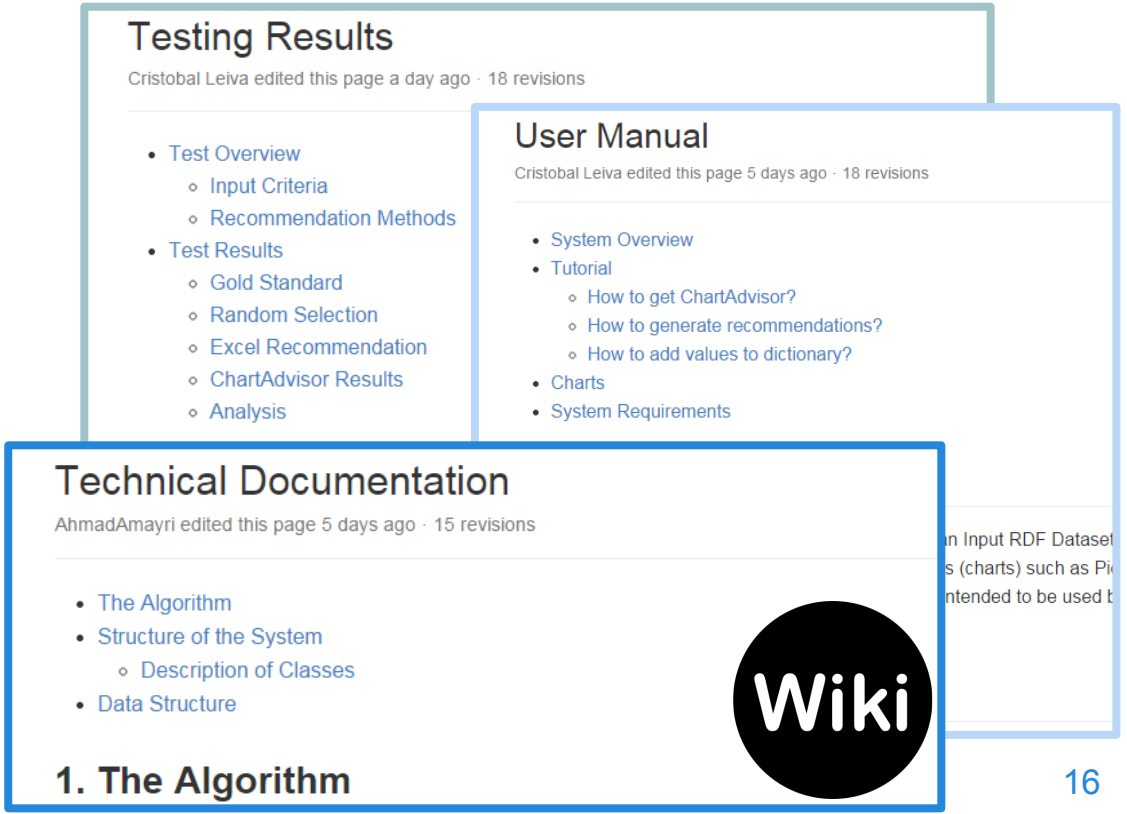
HOW IT WORKS - ALGORITHM

Bubble chart 100%



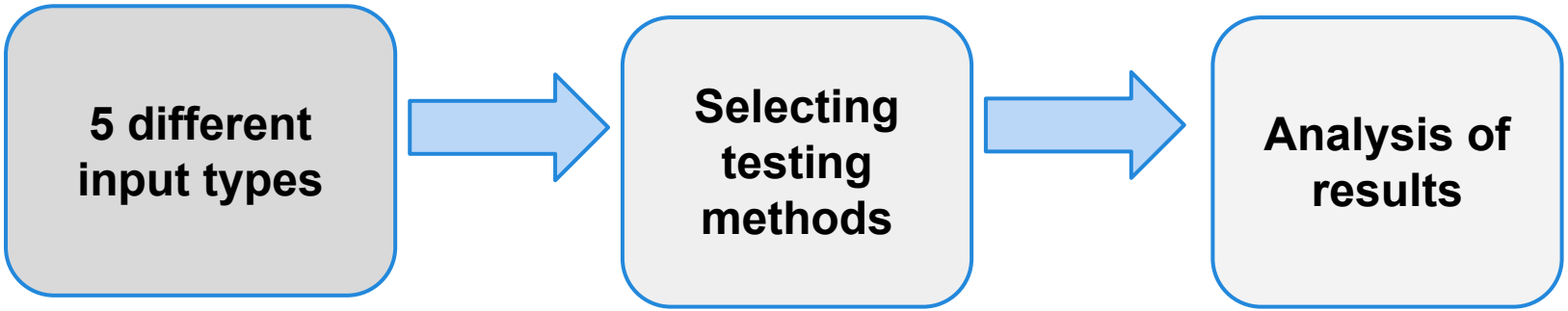
Documentation

- Technical Documentation
  - E.g. UML diagrams, architecture, algorithm explanation, more...
- User Manual
  - E.g. video tutorials, screencasting GIF images, more...
- Test Documentation
  - E.g. results tables, statistics, more...

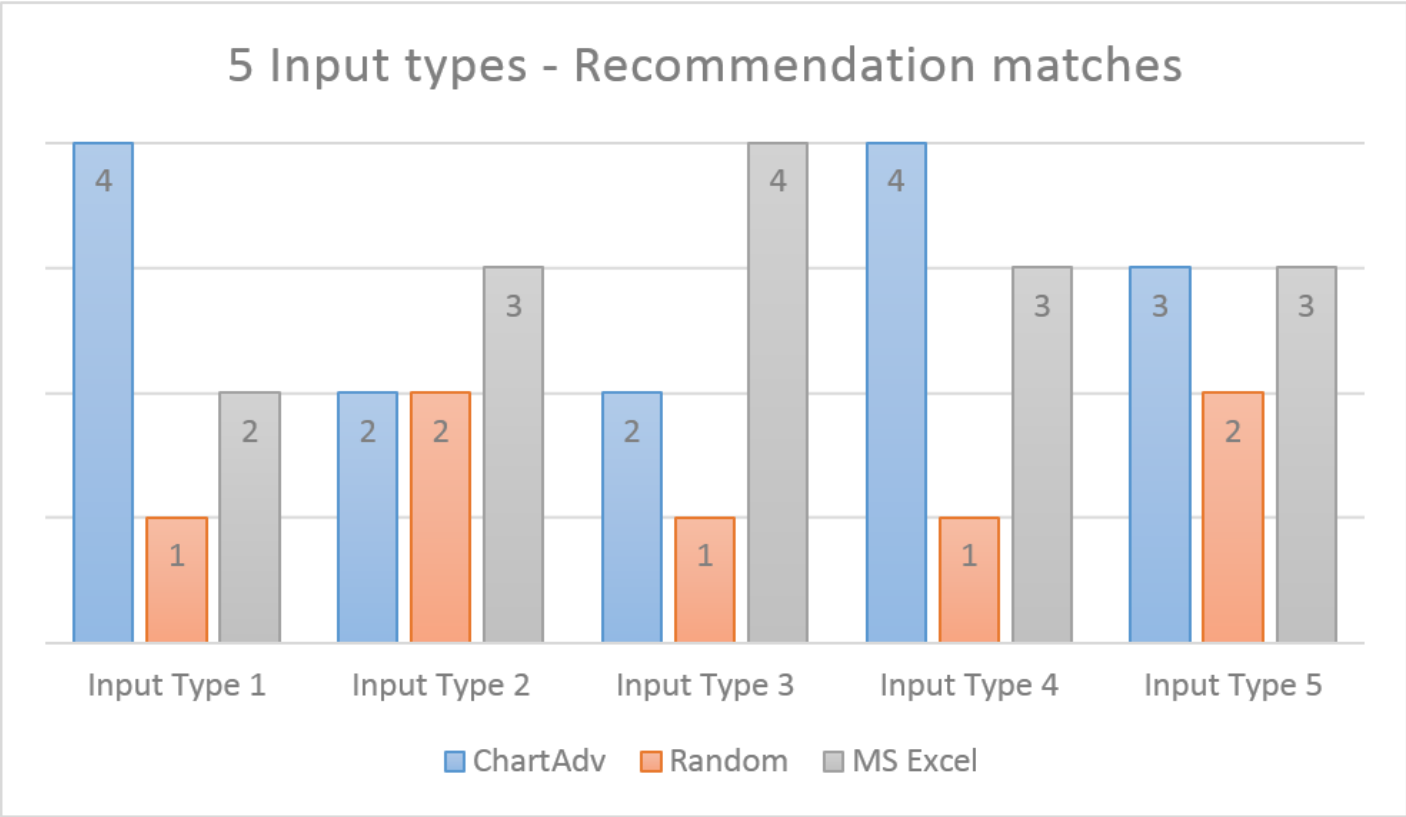


Testing

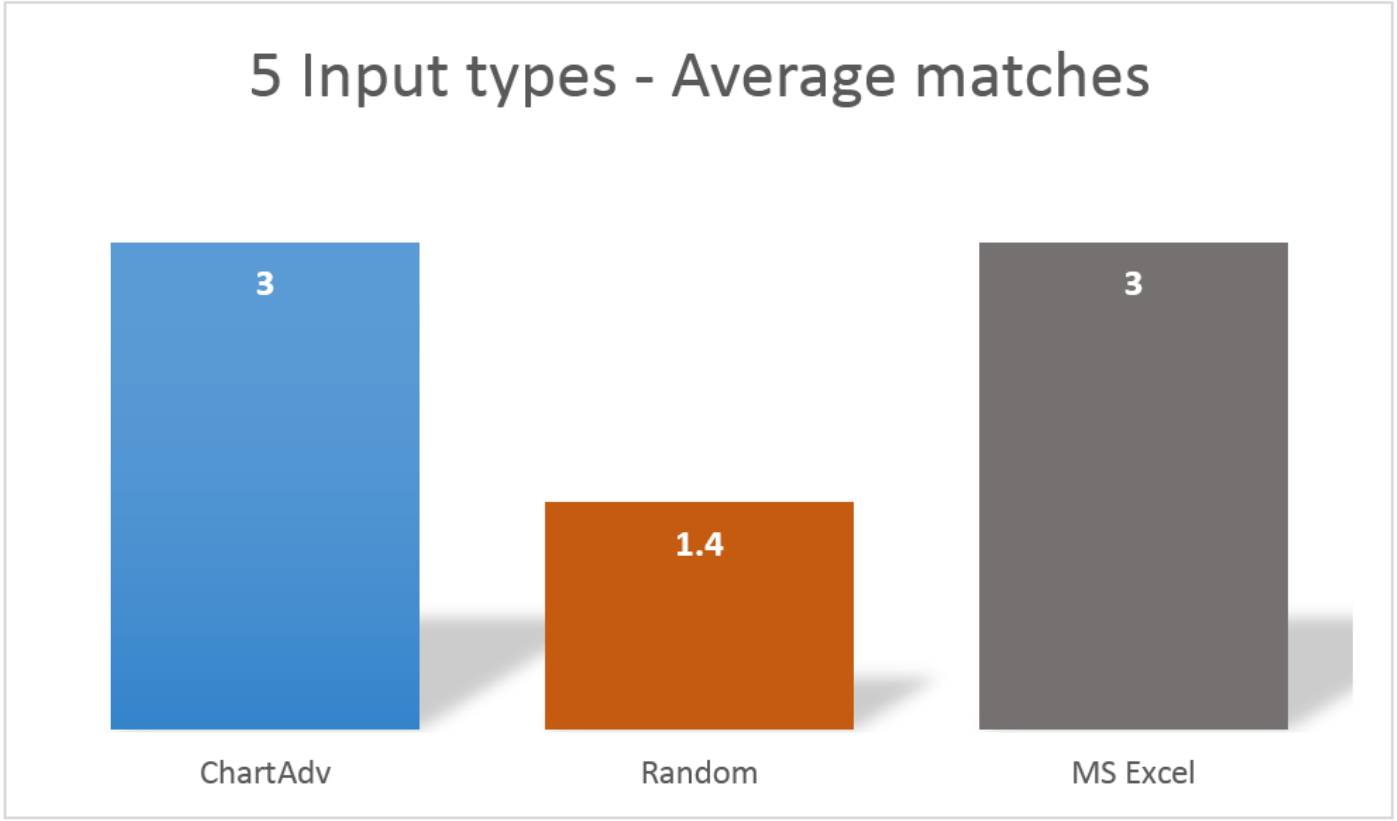
- Selection of input data.
- Creation of gold standard chart recommendations.
  - Charts selected by team members.
- Approaches to test.
  - MS Excel 2013 / Random selection / Chart Advisor
- Analysis of results



Testing - Results

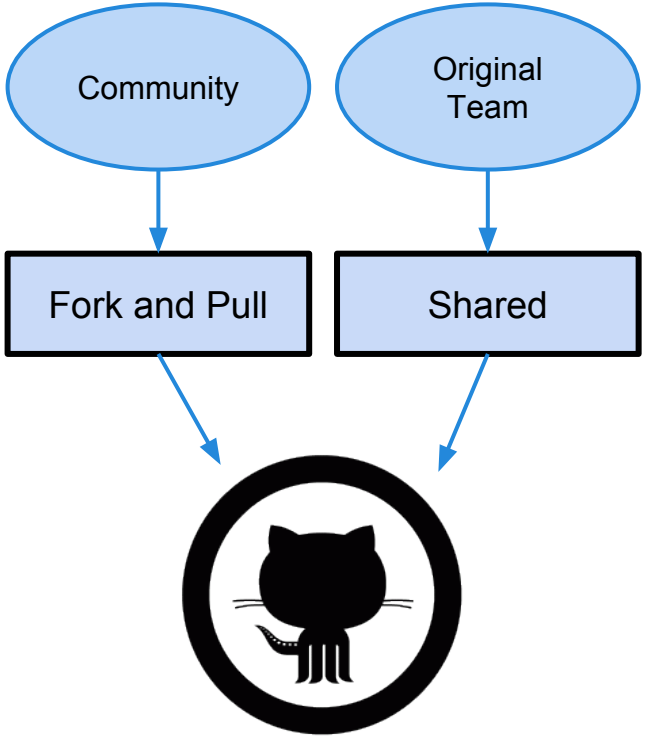


Testing - Results



Deployment on GitHub

- Documentation fully written in wikis.
- Dedicated repository: Summary, source code, documentation.
- README.md with organization of the content on the repo. e.g. documentation index, libraries.
- Project website one-click software download.



Possible Future Enhancements

- Provide a user feed-back for the selection.
  - Allow machine learning capability that learns from user feed-back.
- Expand the number of charts used by the algorithm.
- Feed the dictionary with more properties.
  - Integration of the algorithm to a web platform to centralize dictionary data.

Thank You

For more information or test the software, go to project' s website or check the VM on EIS Lab.

Group Members
Cristobal Leiva
Ahmad Amayri
Jorge Ortiz

Supervisor
Fabrizio Orlandi