Ulster University

A Review of Software Tools for Data Analytics

Timothy Patterson B.Sc. Ph.D.

What is data analytics?

“Data analytics is the science of *examining* and *drawing conclusions* from data”

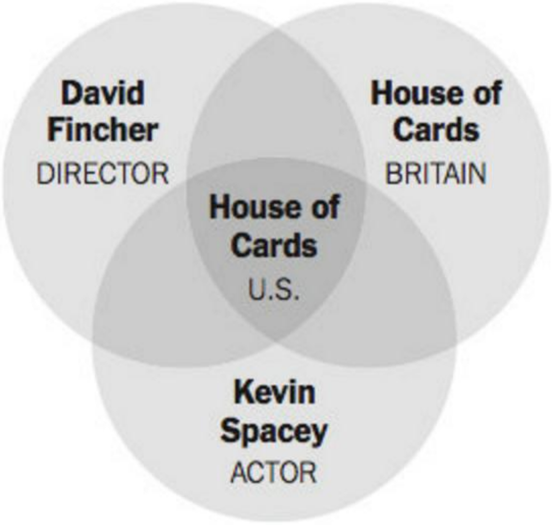
Where is data analytics used?

- Netflix is a data driven company [2]
    - E.g. when you pause, fast forward, rewind; what series do you watch from start to finish
    - Customer retention
    - Content creation

[2] https://blog.kissmetrics.com/how-netflix-uses-analytics/

Circles of Proven Success

Netflix determined that the overlap of these three areas would make "House of Cards" a successful entry into original programming.



THE NEW YORK TIMES

Where is data analytics used?

Walmart

- *"We want to know what every product in the world is. We want to know who every person in the world is. And, we want to have the ability to connect them together in a transaction."* Neil Ashe, Walmart e-commerce CEO (2013)

Ulster University

[3] https://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109

Common tasks in data analytics

- Data exploration
  - Missing values
  - Outlier detection and treatment
  - Visualization
- Feature selection / engineering
- Classification / prediction

Ulster University

Common tasks in data analytics

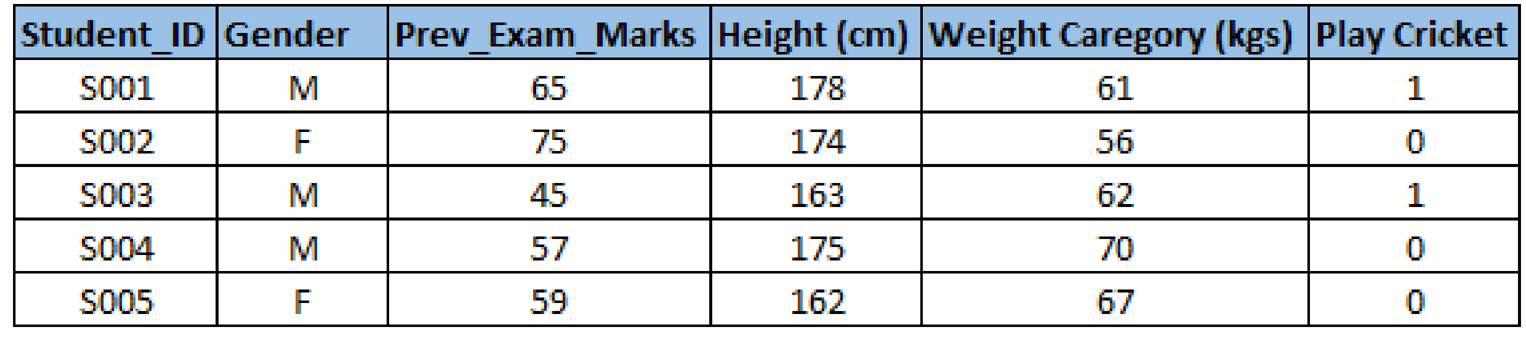
Data exploration

- The quality of your input will determine the quality of your output and may take up to 70% of the project time [4]
- Data exploration tasks:
    - Variable identification
    - Uni/Bi-variate analysis
    - Treatment of missing values
    - Detection of outliers
    - Variable transformation / creation

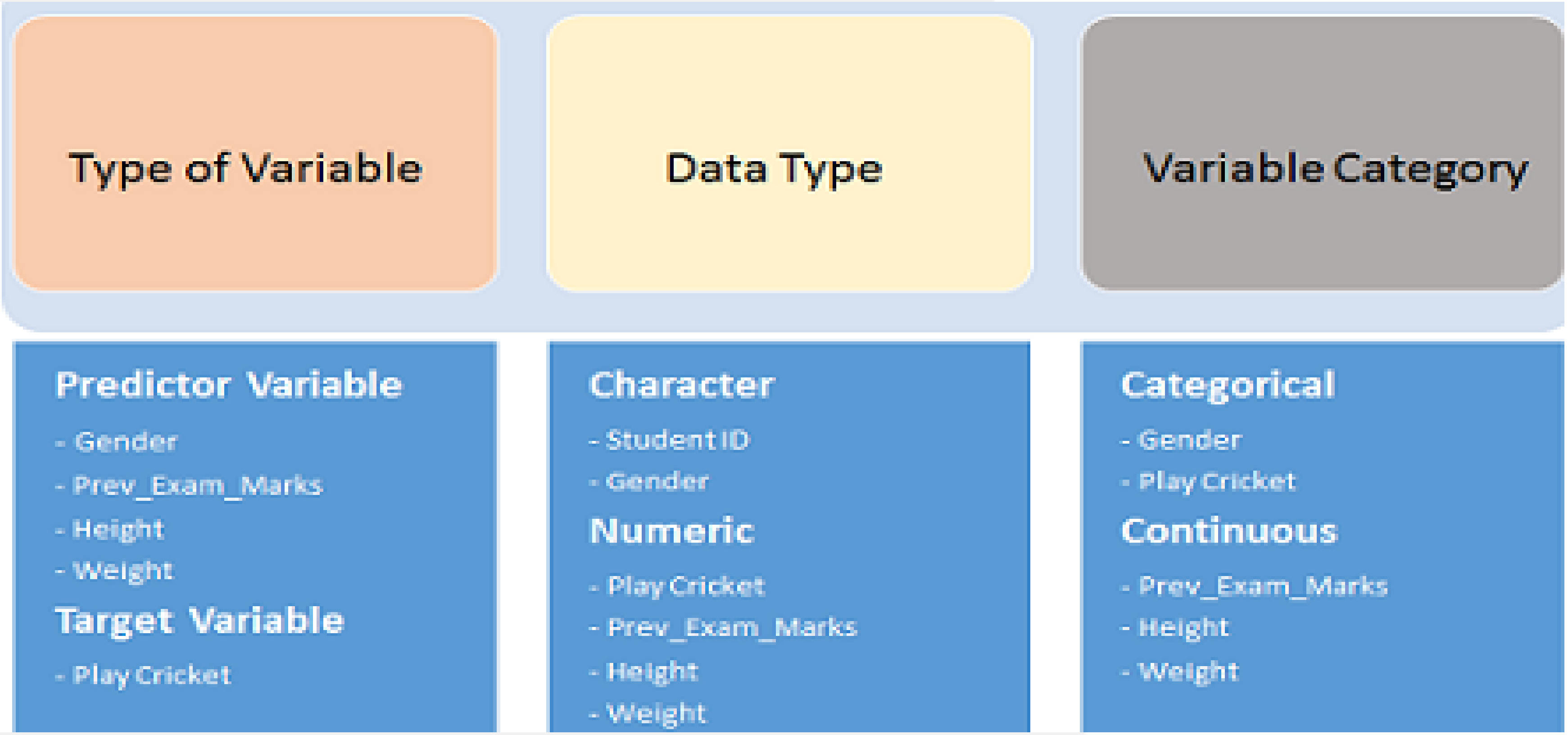
[4] http://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

Data exploration

Variable identification



Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0



Type of Variable

Data Type

Variable Category

Predictor Variable

- Gender
- Prev_Exam_Marks
- Height
- Weight

Target Variable

- Play Cricket

Character

- Student ID
- Gender

Numeric

- Play Cricket
- Prev_Exam_Marks
- Height
- Weight

Categorical

- Gender
- Play Cricket

Continuous

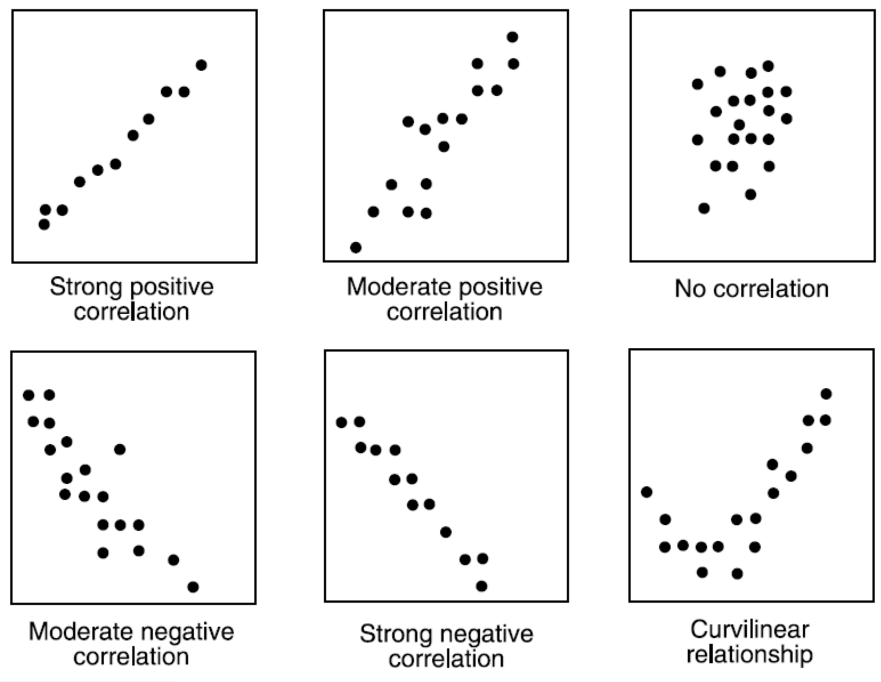
- Prev_Exam_Marks
- Height
- Weight

Data exploration

Variable analysis

- Univariate analysis
  - Continuous variables
  - Categorical variables

- Bivariate analysis
  - Continuous & Continuous
  - Categorical & Categorical
  - Categorical & Continuous



Strong positive correlation

Moderate positive correlation

No correlation

Moderate negative correlation

Strong negative correlation

Curvilinear relationship

Data exploration

Treatment of missing values

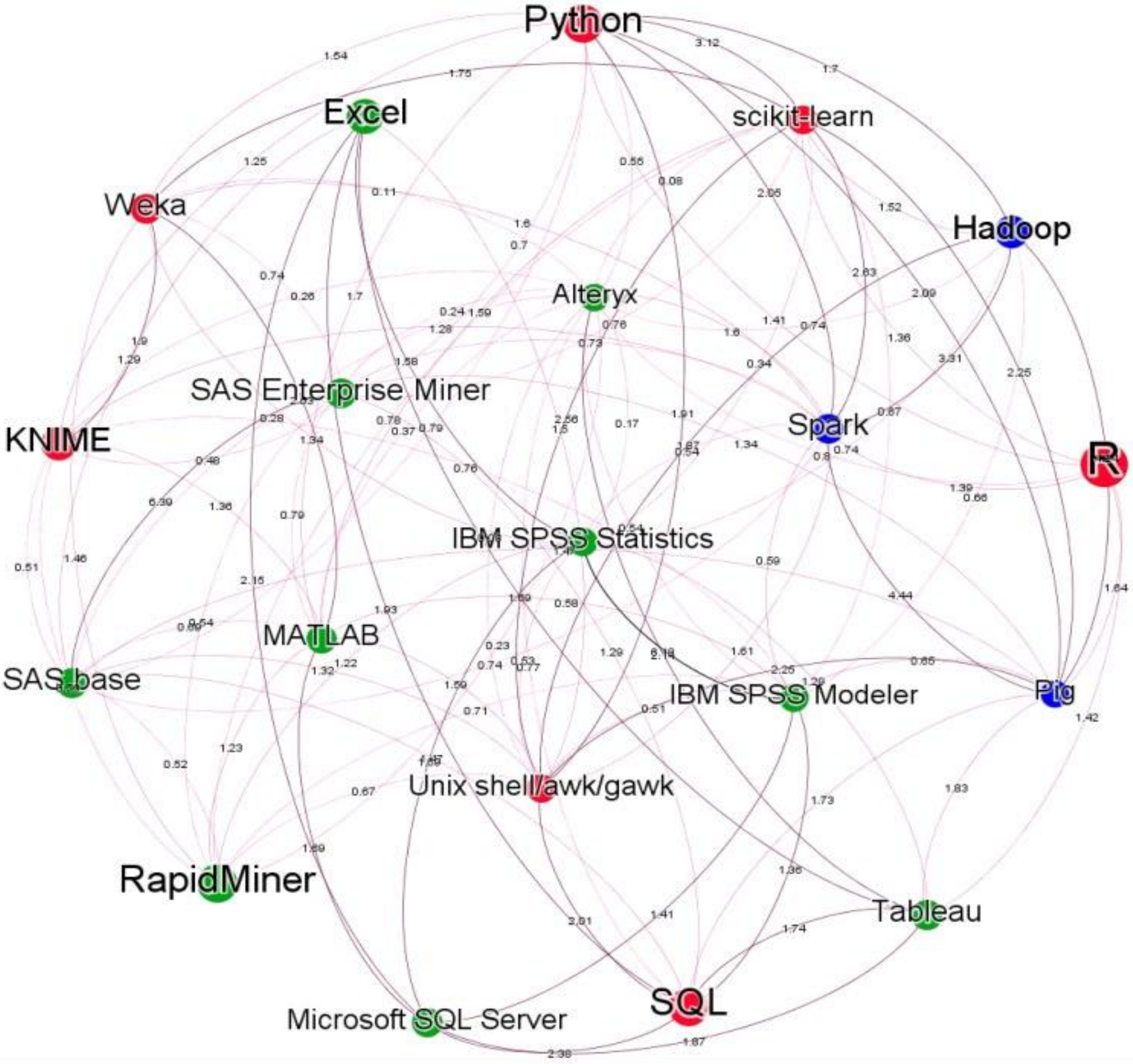
- Reasons for missing data
  - Data extraction
  - Data collection

- Methods for treating missing values
  - Deletion
  - Mean/median/mode imputation
  - Predictive model
  - kNN imputation

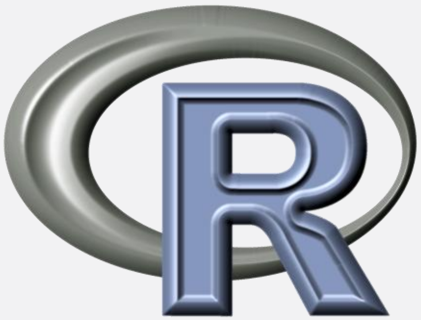
Data exploration

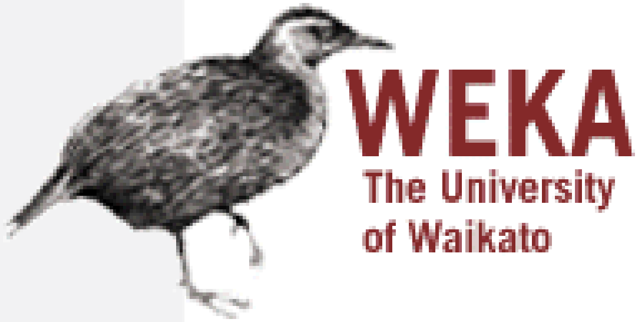
- Feature selection

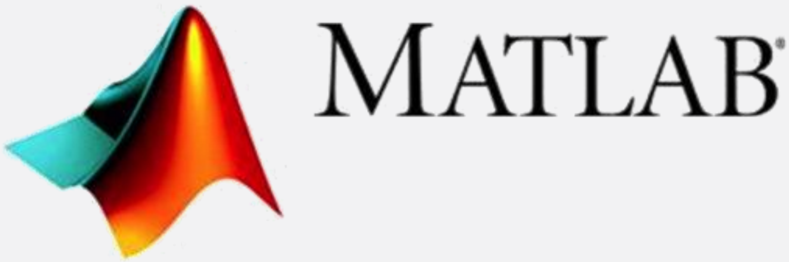
- Classification

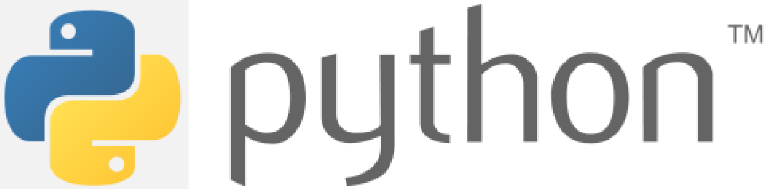

Python
1.54
3.12
1.75
1.7
Excel
scikit-learn
1.25
0.56
0.08
Weka
0.11
2.05
1.52
Hadoop
1.6
0.7
2.63
0.74
Alteryx
0.26
1.7
2.09
0.24
1.59
0.76
1.41
0.74
1.28
1.6
1.36
1.9
0.73
1.29
1.58
0.34
3.31
2.25
SAS Enterprise Miner
2.56
1.91
0.87
Spark
0.28
0.78
1.5
0.17
KNIME
1.34
0.37
0.79
1.87
0.54
1.34
0.74
R
0.48
0.8
0.76
1.39
0.66
6.39
1.36
0.79
IBM SPSS Statistics
0.51
1.46
0.59
2.15
1.64
1.59
0.58
4.44
0.69
0.54
1.93
MATLAB
0.23
1.29
6.19
2.14
1.61
0.65
SAS base
1.22
1.32
0.74
0.53
0.77
2.25
1.29
1.59
IBM SPSS Modeler
Pig
0.71
0.51
1.42
1.23
0.52
1.87
1.88
Unix shell/awk/gawk
0.67
1.83
1.73
1.89
RapidMiner
1.36
Tableau
2.01
1.41
1.74
SQL
Microsoft SQL Server
1.87
2.38

Software tools









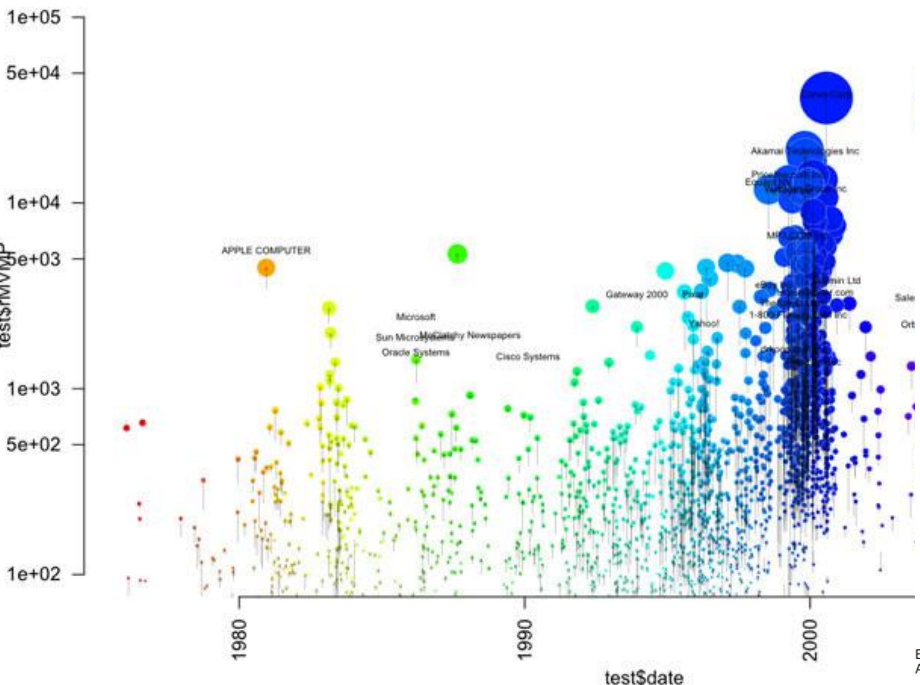
Software Tools

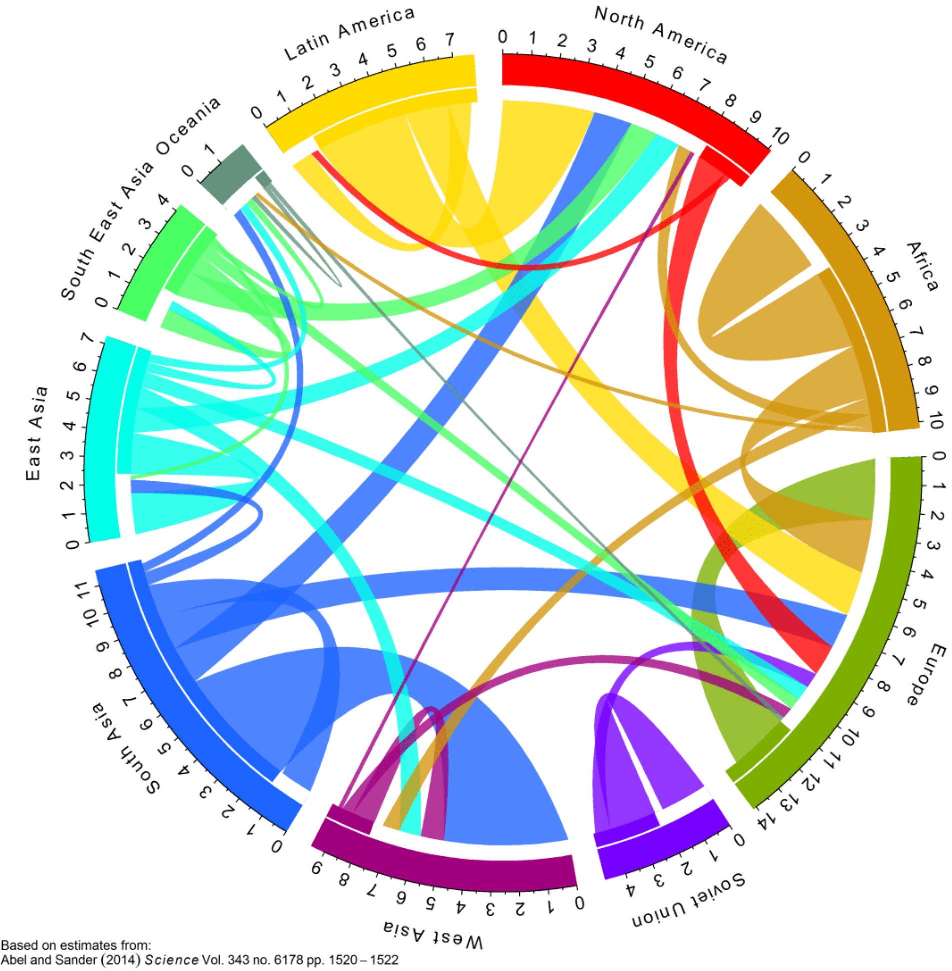
- Overview
- Learning curve
- Trouble shooting / Debugging
- User community
- Cost
- Available libraries
- Data analytics process

Software Tools

R - Overview

- Open source
- Focus towards statistics
- Strong visualization tools





Based on estimates from:
Abel and Sander (2014) *Science* Vol. 343 no. 6178 pp. 1520 – 1522

Software Tools

R - Overview

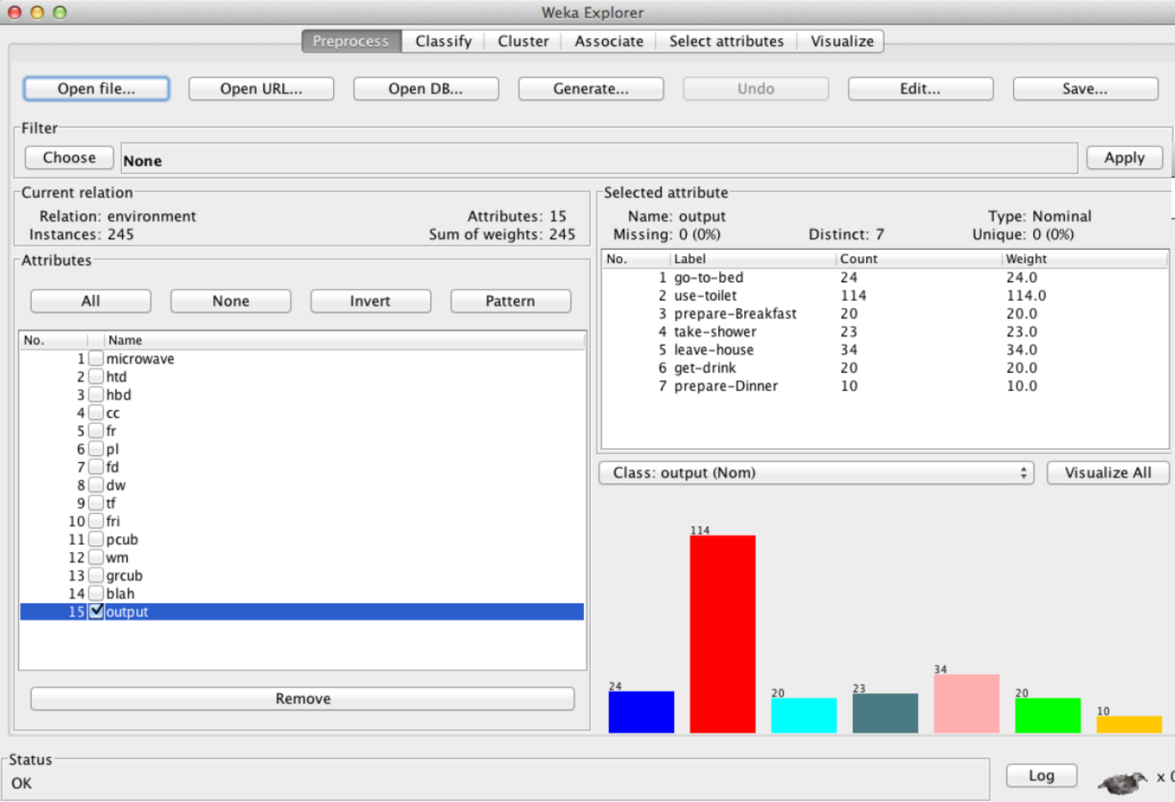
- Rstudio
- Packages
  - zoo – work with time-series data
  - lattice – to visualize data
  - caret – machine learning package

Software Tools

Weka – Overview

http://www.cs.waikato.ac.nz/ml/weka/downloading.html

- Open source
- GUI / Java Library
- Strong visualization tools for data exploration



Filter: None

Current relation
Relation: environment Attributes: 15
Instances: 245 Sum of weights: 245

Attributes

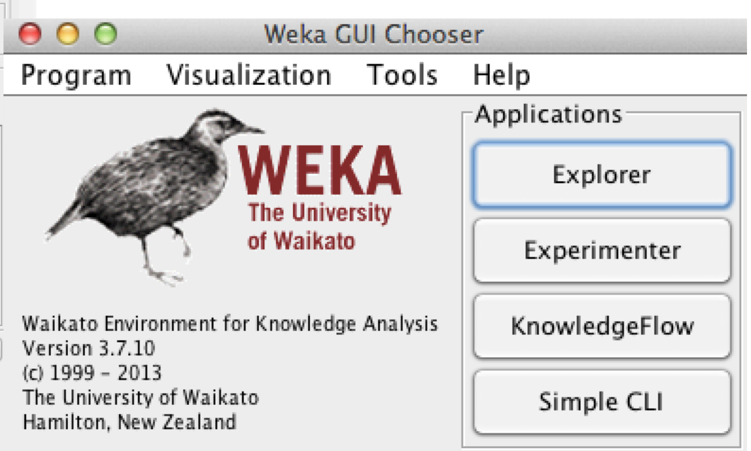
No.		Name
1	□	microwave
2	□	htd
3	□	hbd
4	□	cc
5	□	fr
6	□	pl
7	□	fd
8	□	dw
9	□	tf
10	□	fri
11	□	pcub
12	□	wm
13	□	grcub
14	□	blah
15	☑	output

Selected attribute
Name: output Type: Nominal
Missing: 0 (0%) Distinct: 7 Unique: 0 (0%)

No.	Label	Count	Weight
1	go-to-bed	24	24.0
2	use-toilet	114	114.0
3	prepare-Breakfast	20	20.0
4	take-shower	23	23.0
5	leave-house	34	34.0
6	get-drink	20	20.0
7	prepare-Dinner	10	10.0

Class: output (Nom)

Status
OK



Waikato Environment for Knowledge Analysis
Version 3.7.10
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

Software Tools

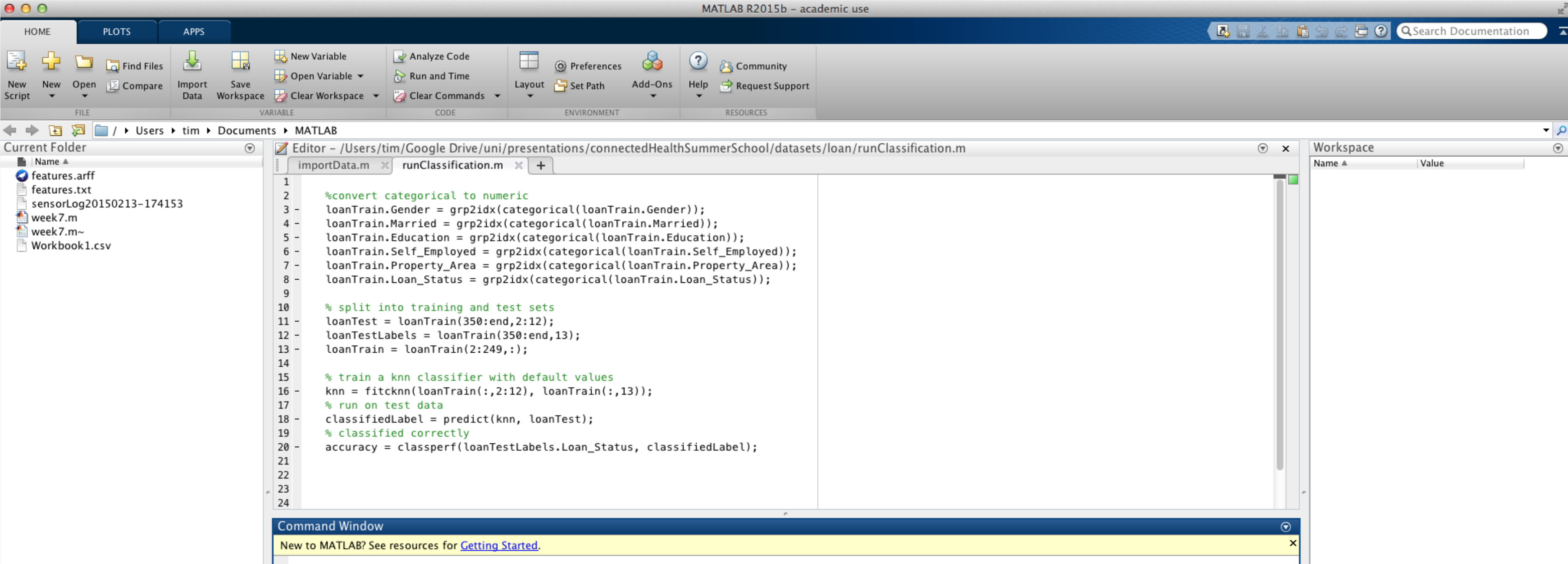
Matlab - Overview

- Owned/maintained/developed by Mathworks
- Costs (base product, add-on-products)
  - Individual £1600, add-on-products e.g. Statistics and Machine Learning £800
  - Home £85, add-on-products £25
  - Student £29, add-on-products £16
- Good for rapid prototyping

Software Tools

Matlab - Overview

- Intuitive GUI
- Debugging capabilities
- User Community
- Learning curve

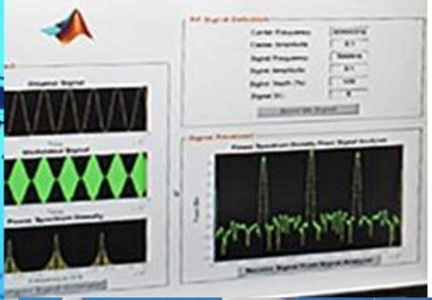


features.arff
features.txt
sensorLog20150213-174153
week7.m
week7.m~
Workbook1.csv

```matlab

%convert categorical to numeric
loanTrain.Gender = grp2idx(categorical(loanTrain.Gender));
loanTrain.Married = grp2idx(categorical(loanTrain.Married));
loanTrain.Education = grp2idx(categorical(loanTrain.Education));
loanTrain.Self_Employed = grp2idx(categorical(loanTrain.Self_Employed));
loanTrain.Property_Area = grp2idx(categorical(loanTrain.Property_Area));
loanTrain.Loan_Status = grp2idx(categorical(loanTrain.Loan_Status));

% split into training and test sets
loanTest = loanTrain(350:end,2:12);
loanTestLabels = loanTrain(350:end,13);
loanTrain = loanTrain(2:249,:);

% train a knn classifier with default values
knn = fitcknn(loanTrain(:,2:12), loanTrain(:,13));
% run on test data
classifiedLabel = predict(knn, loanTest);
% classified correctly
accuracy = classperf(loanTestLabels.Loan_Status, classifiedLabel);
```

New to MATLAB? See resources for Getting Started.

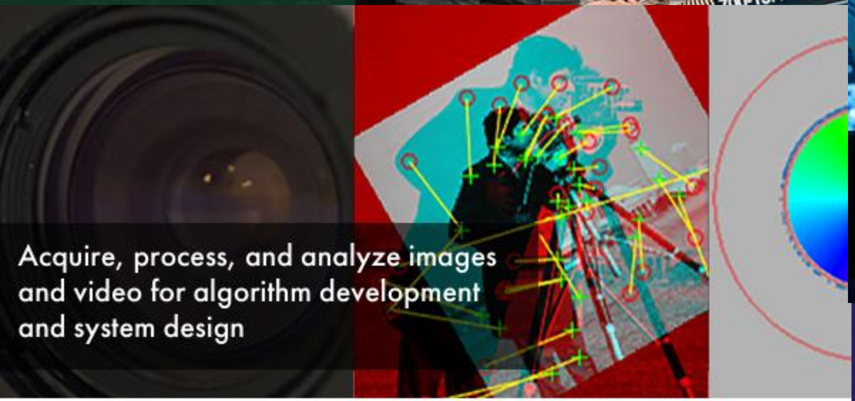
Acquire, analyze, and explore data and automate tests



MATLAB

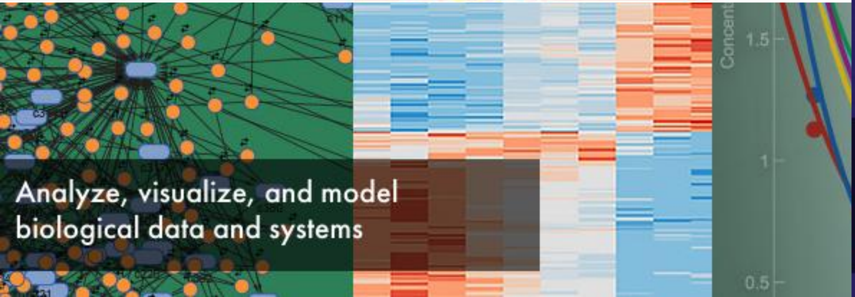
Develop and deploy efficient and robust financial applications

Design, test, and implement control systems

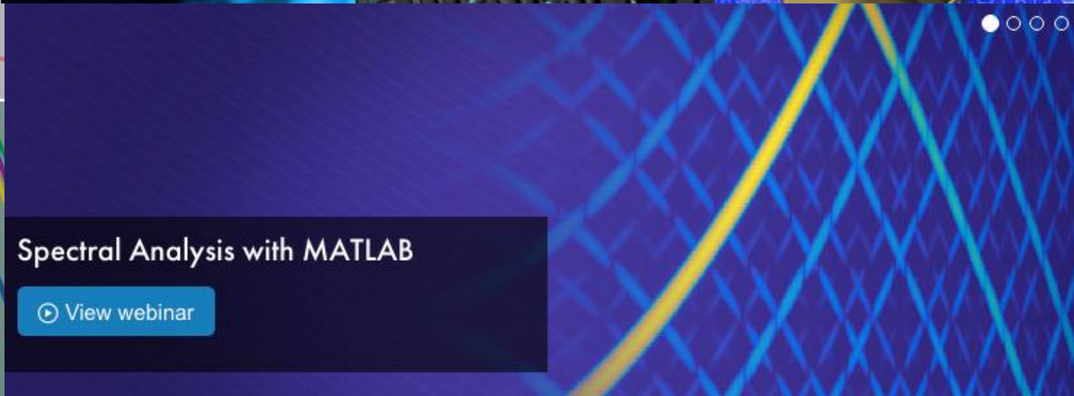


Acquire, process, and analyze images and video for algorithm development and system design

Perform large-scale computations using multicore desktops, GPUs, clusters, grids, and clouds



Analyze, visualize, and model biological data and systems



Spectral Analysis with MATLAB

Software Tools

Data analytics processes

- Importing data
    - Drag and drop, UI for importing and dealing with missing values
- Data exploration
    - *>>histogram(categorical(loanTrain.Property_Area));*
    - *>>scatter(loanTrain.ApplicantIncome,loanTrain.LoanAmount)*
    - Max, min, standard deviation, mean etc.
    - *>>a=categorical(loanTrain.Education)*
    - *>>summary(a)*

Ulster University

Software Tools

Data analytics processes

- Feature selection
    - Cannot be performed in UI [5]
- Classification
    - Can use UI
    - Or code
        - *>>knn=ClassificationKNN.fit(Xtrain, Ytrain)*

[5] http://uk.mathworks.com/help/stats/select-subset-of-features-with-comparative-predictive-power.html

Software Tools

Python

- General purpose programming language
- Open source
- Large user community
- 2 versions, 2.x and 3.x available from https://www.python.org/downloads/
    - Currently more library support for 2.x

[5] http://uk.mathworks.com/help/stats/select-subset-of-features-with-comparative-predictive-power.html

Software Tools

Python

- Popular Libraries
    - NumPy – Numerical Python
    - Matplotlib – Plotting graphs
    - Pandas – Structured data operations
    - Scikit learn – For machine learning
    - OS – Operating system and file operations
    - BeautifulSoup – Scrape webpages

Software Tools

Python

https://try.jupyter.org/

Software Tools

Python

Pandas

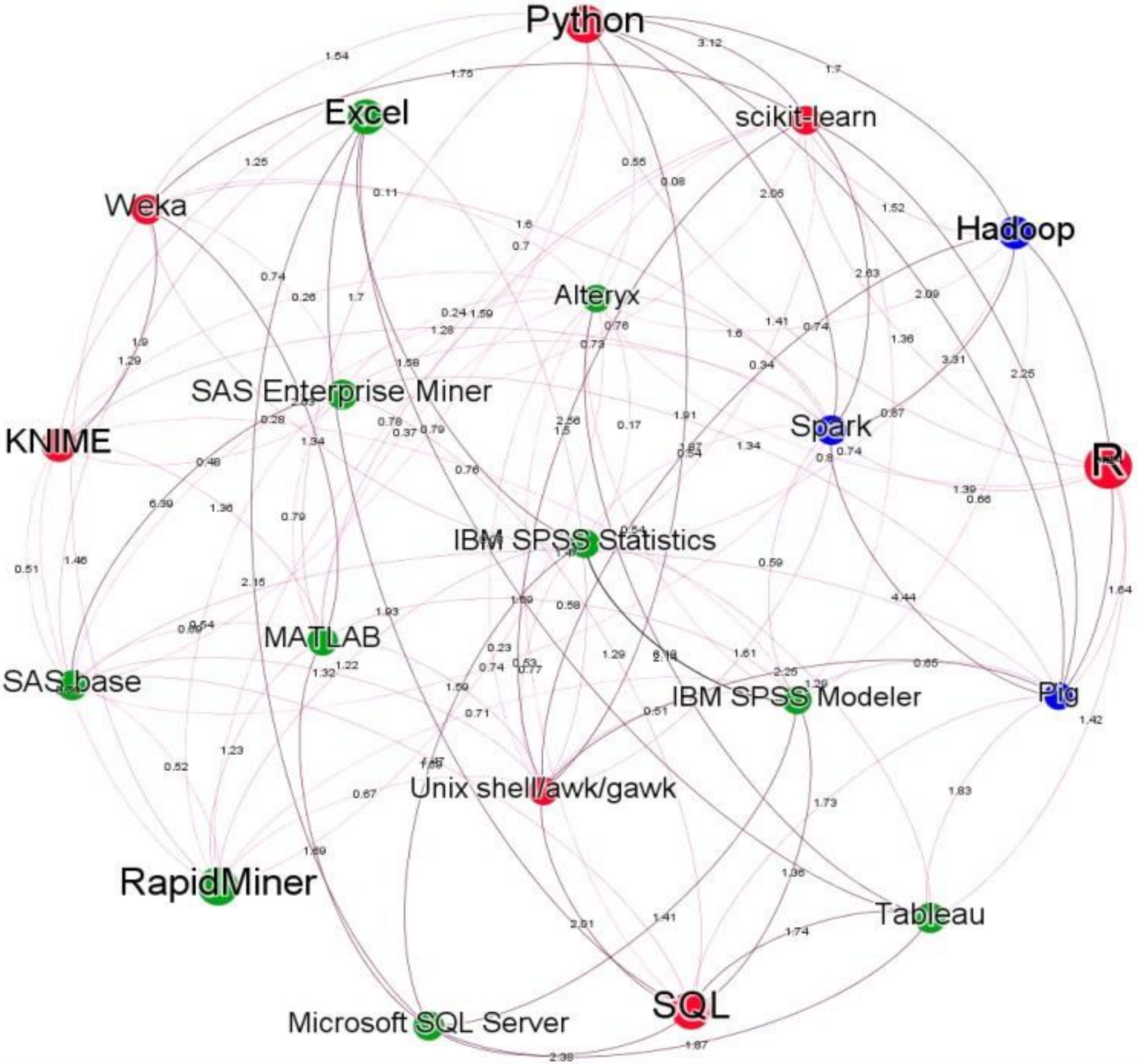
- *df[‘columnName’].hist(bins=n)*
- *df.boxplot(column=’columnName')*
- *df[‘columnName’].plot(kind=‘bar’)*

  df['Property_Area'].value_counts().plot(kind='bar')

Matplotlib

- *plt.pyplot.scatter(df['ApplicantIncome'], df['LoanAmount'])*
- *plt.pyplot.boxplot(x=df['ApplicantIncome'])*

where df is a DataFrame


Python
scikit-learn
Excel
Weka
Hadoop
Alteryx
SAS Enterprise Miner
KNIME
Spark
R
IBM SPSS Statistics
MATLAB
SAS base
IBM SPSS Modeler
Pig
Unix shell/awk/gawk
RapidMiner
Tableau
SQL
Microsoft SQL Server
1.54
1.75
3.12
1.7
1.25
0.11
0.56
0.08
2.05
1.52
1.6
0.7
0.74
0.26
1.7
2.63
2.09
0.24
1.59
1.28
0.76
0.73
1.41
0.74
1.6
1.36
1.9
1.29
1.58
0.34
3.31
2.25
2.03
0.28
0.78
0.37
0.79
2.56
1.5
0.17
1.91
0.87
1.34
1.87
0.54
1.34
0.74
0.8
0.48
0.76
1.39
0.66
6.39
1.36
0.79
0.59
0.51
1.46
2.15
1.69
0.58
4.44
1.64
1.93
0.69
0.54
0.23
1.29
6.18
2.14
1.61
0.65
1.22
1.32
0.74
0.53
0.77
2.25
1.29
1.59
0.71
0.51
1.42
1.23
0.52
0.67
1.87
1.88
1.73
1.83
1.69
1.36
2.01
1.41
1.74
2.38
1.87