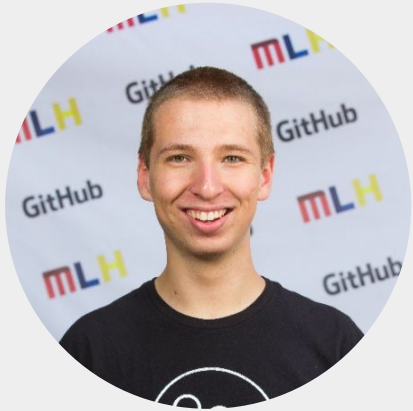
Open Source Analysis

Drew Casner



Carl Cortright

Shubha Swamy

Oliver Collins

Description

- Open source projects
- libraries.io
- 311 million data points

Questions

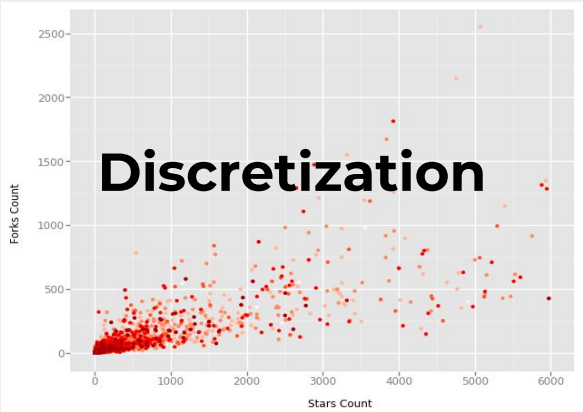
Insights	What insights can we gain to improve the open source community further?
Improvements	How can we identify areas in the open source community that need improvement?
Predictions	Can we predict upcoming popular repositories?

Data Preparation Work

	Unnamed: 0	Fork	Created Timestamp	Updated Timestamp	Last pushed Timestamp	Homepage URL	Size
0	1	2014-09-15 01:21:34 UTC	2016-12-28 16:33:17 UTC	2016-12-18 18:31:32 UTC	http://brianmhunt.github.io/knockout-modal/	512	7
1	2	2010-11-01 09:27:43 UTC	2018-02-11 10:04:55 UTC	2017-06-21 22:54:45 UTC	NaN	924	543
2	3	2014-09-13 03:14:07 UTC	2017-03-18 22:40:02 UTC	2015-01-14 02:01:03 UTC	NaN	472	1
3	4	2014-12-27 21:02:09 UTC	2016-12-28 16:45:20 UTC	2015-01-07 18:04:42 UTC	http://zonuexe.github.io/aozora-ruby-parser.js/	536	3
4	5	2014-12-04 21:13:48 UTC	2017-03-18 22:40:04 UTC	2014-12-11 16:12:08 UTC	http://rawgit.com/immense/knockout-pickatime/m...	192	1
5	6	2014-12-04 17:04:45 UTC	2016-11-03 09:12:51 UTC	2015-10-16 18:13:50 UTC	http://rawgit.com/immense/knockout-pickadate/m...	285	3
6	7	2014-11-24 22:08:11 UTC	2017-08-07 19:18:03 UTC	2016-06-20 20:09:51 UTC	NaN	771	0

Data cleaning

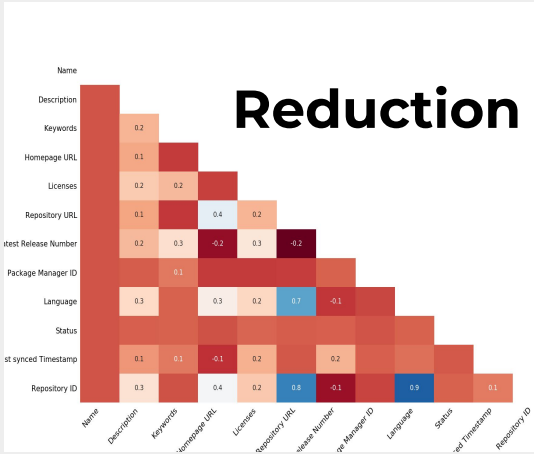
Integration




```
# Get the data in form YYYY-MM-DD
dates = []
for x in range(10000):
    dates.append(a[x][0:10])
```

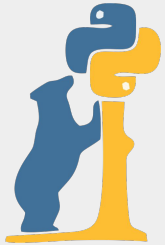
Transformation

Reduction



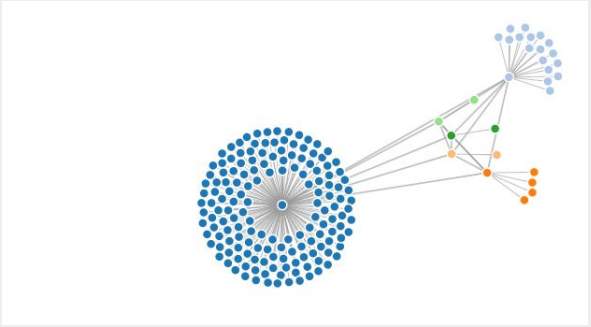
Tools Used

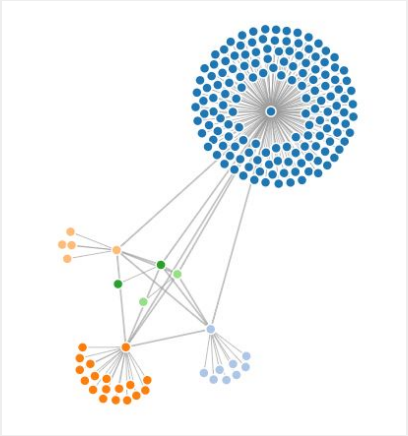
- Pandas
- NumPy
- SciPy
- Jupyter Notebook
- Matplotlib
- ggplot2
- scikit Learn
- D3.js
- Seaborn
- Amazon Cloud9
- Github

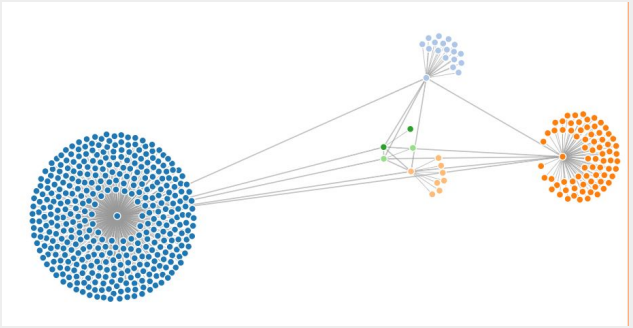


Data Mining Methods

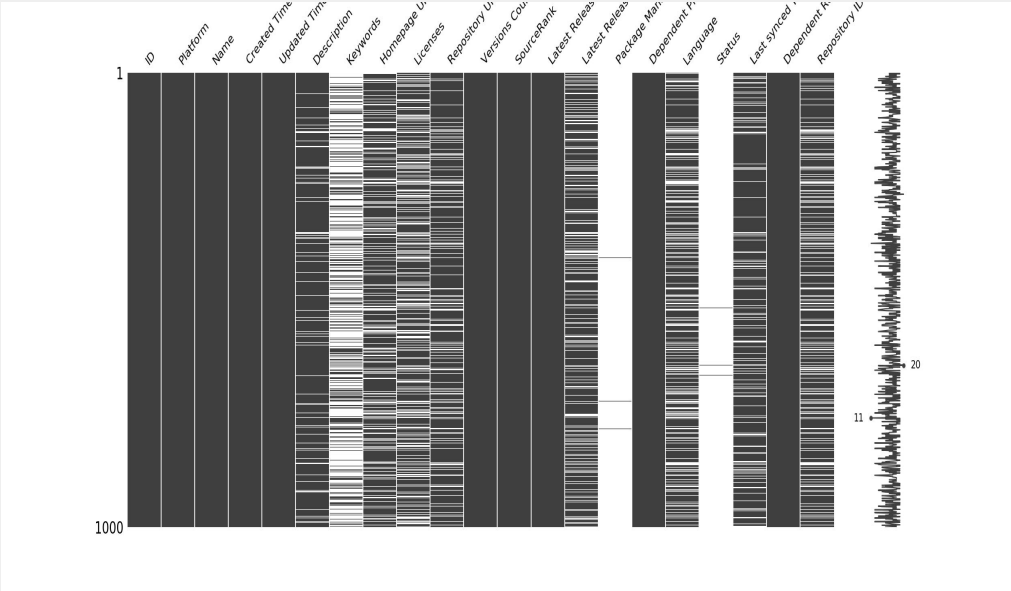
KMeans Clustering

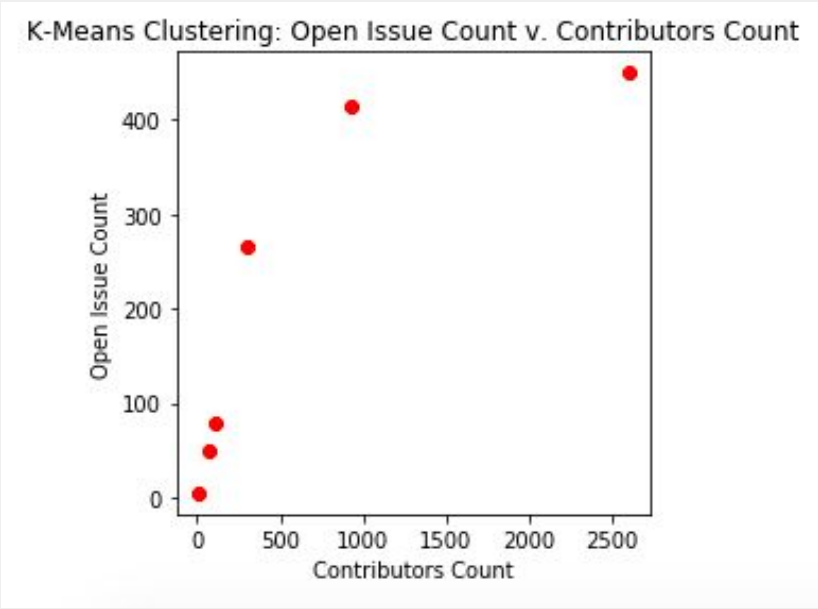






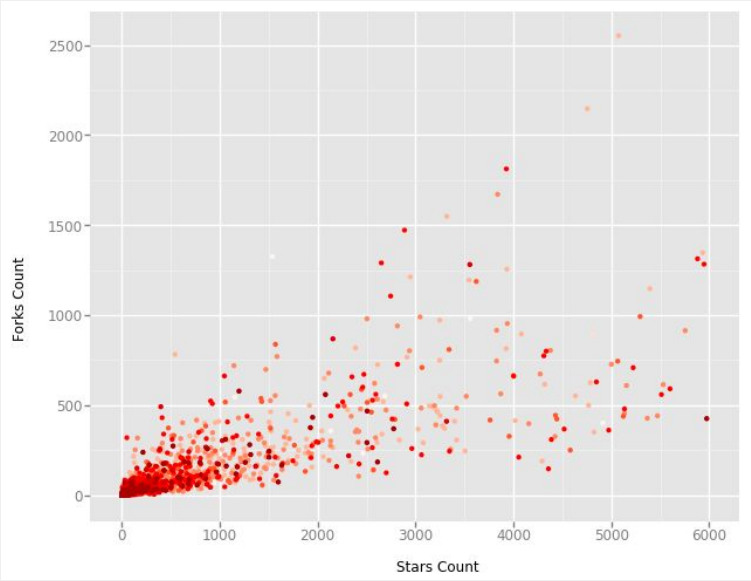
DBSCAN Clustering

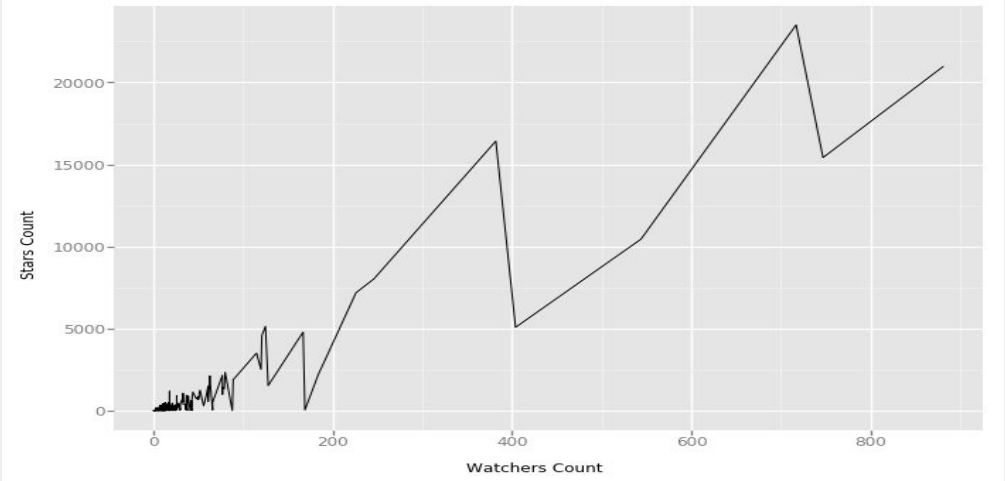

ID
Platform
Name
Description
Keywords
Licenses
SourceRank
Language
Status
1
1000
20
11

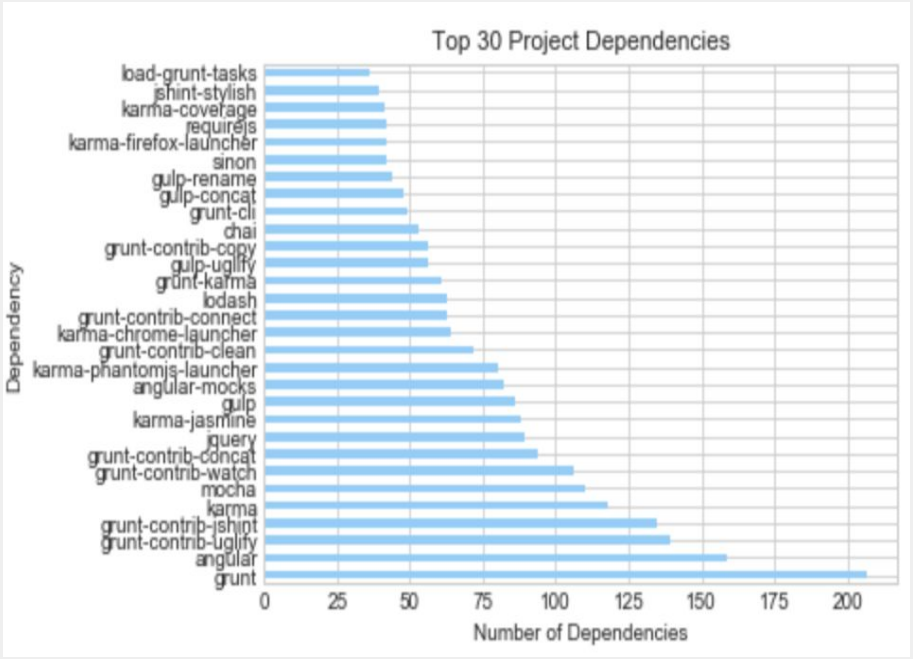

K-Means Clustering: Open Issue Count v. Contributors Count
Open Issue Count
Contributors Count
0
100
200
300
400
0
500
1000
1500
2000
2500

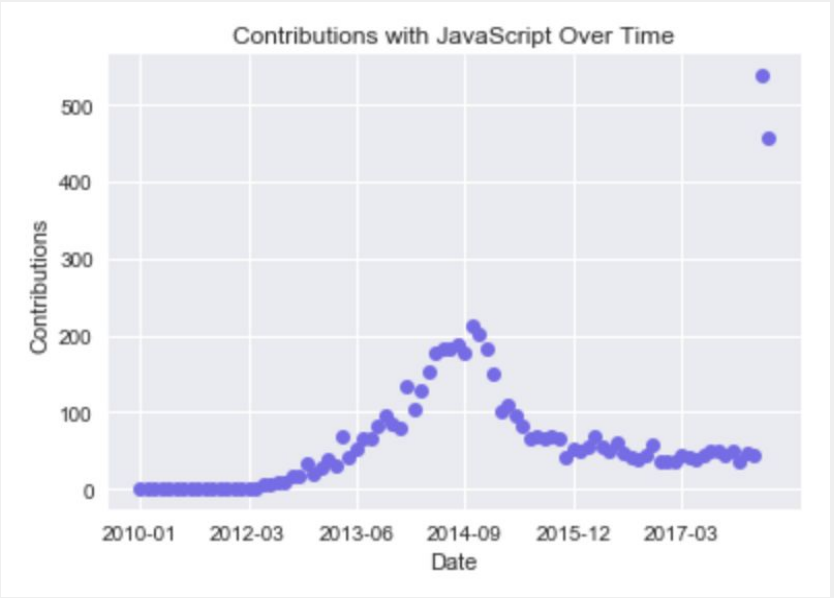
Knowledge Gained

Insights	Trends and patterns in open source development
Improvements	Identified 17 Repositories In “High Need” of contribution
Predictions	Identified 348 “Up and Coming” Repositories we believe will be very prevalent soon


Forks Count
2500
2000
1500
1000
500
0
0
1000
2000
3000
4000
5000
6000
Stars Count


Stars Count
20000
15000
10000
5000
0
0
200
400
600
800
Watchers Count


Top 30 Project Dependencies
load-grunt-tasks
jshint-stylish
karma-coverage
requirejs
karma-firefox-launcher
sinon
gulp-rename
gulp-concat
grunt-cli
chai
grunt-contrib-copy
gulp-uglify
grunt-karma
lodash
grunt-contrib-connect
karma-chrome-launcher
grunt-contrib-clean
karma-phantomjs-launcher
angular-mocks
gulp
karma-jasmine
jquery
grunt-contrib-concat
grunt-contrib-watch
mocha
karma
grunt-contrib-jshint
grunt-contrib-uglify
angular
grunt
Dependency
0
25
50
75
100
125
150
175
200
Number of Dependencies


Contributions with JavaScript Over Time
Contributions
0
100
200
300
400
500
2010-01
2012-03
2013-06
2014-09
2015-12
2017-03
Date

Application of Knowledge

- **Tens of thousands of open source projects run worldwide; millions of users relying on the software**
- **Match Developers with repositories in need of support**
- **Improve knowledge surrounding FOSS**
- **Show future trends, so developers can adapt and prepare for the future**
- **Provide a general direction to work towards when creating new software**
- **programming languages: dying vs. growing**

https://bit.ly/2Fvverq