BitQuery

GitHub API driven and D3 based search engine for open source repositories

---

Lukas Borke (github.com/lborke)
Svetlana Bykovskaya (github.com/polarstern)

Data Science, Statistics & Visualization Conference, July 12-14, 2017

Humboldt-Universität zu Berlin, Lomonosov Moscow State University

Table of contents

Motivation

CRAN Task Views

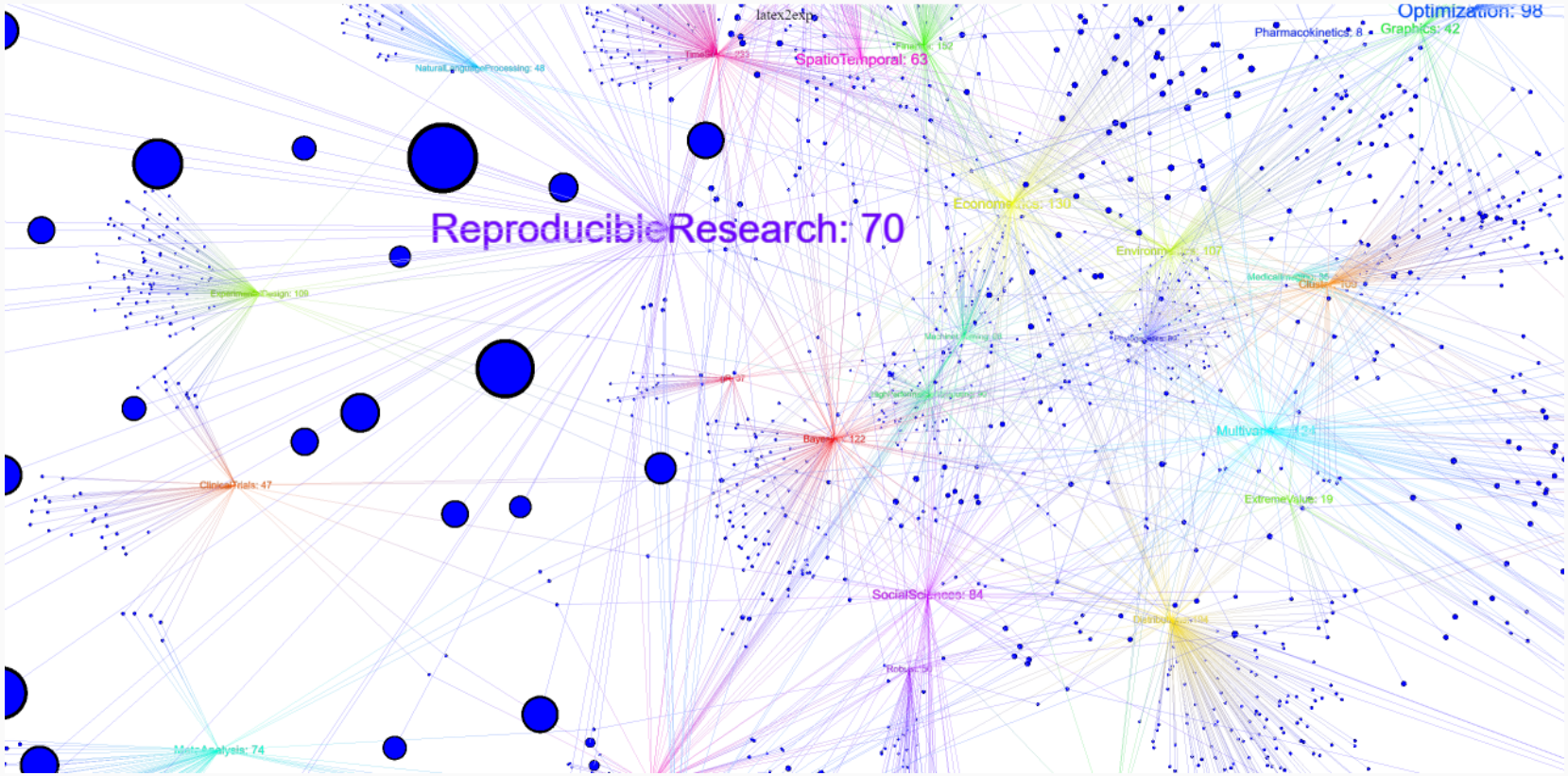

Figure 1: Collaborative reproducible research (CRR): Dynamic 3D visualization of all CRAN Task Views, created by the **taskviewsVA** package

CRAN Statistics

	total number of packages
Entire GitHub	32375
CRAN mirror on GitHub	12298
Official CRAN mirror (Austria)	10811
Bioconductor-mirror on GitHub	1426
Wickham (as search query)	200
Hornik (as search query)	174
Hastie (as search query)	43
Leisch (as search query)	34

Table 1: CRAN@GitHub statistics via the *cran.stats* function from the **rgithubS** package

What is GitHub?

- A distributed version control system (Git)
- A collaboration platform (Hub)
- Offers a variety of open source repositories (OSR)
- The largest host of source code in the world with more than one million organizations: Google, Facebook, Twitter, Yahoo, D3, RStudio, CRAN, Bioconductor ...
- Provides an extensive REST API, which enables scientists to retrieve valuable information about the software and research development life cycles

GitHub Structure & OSR

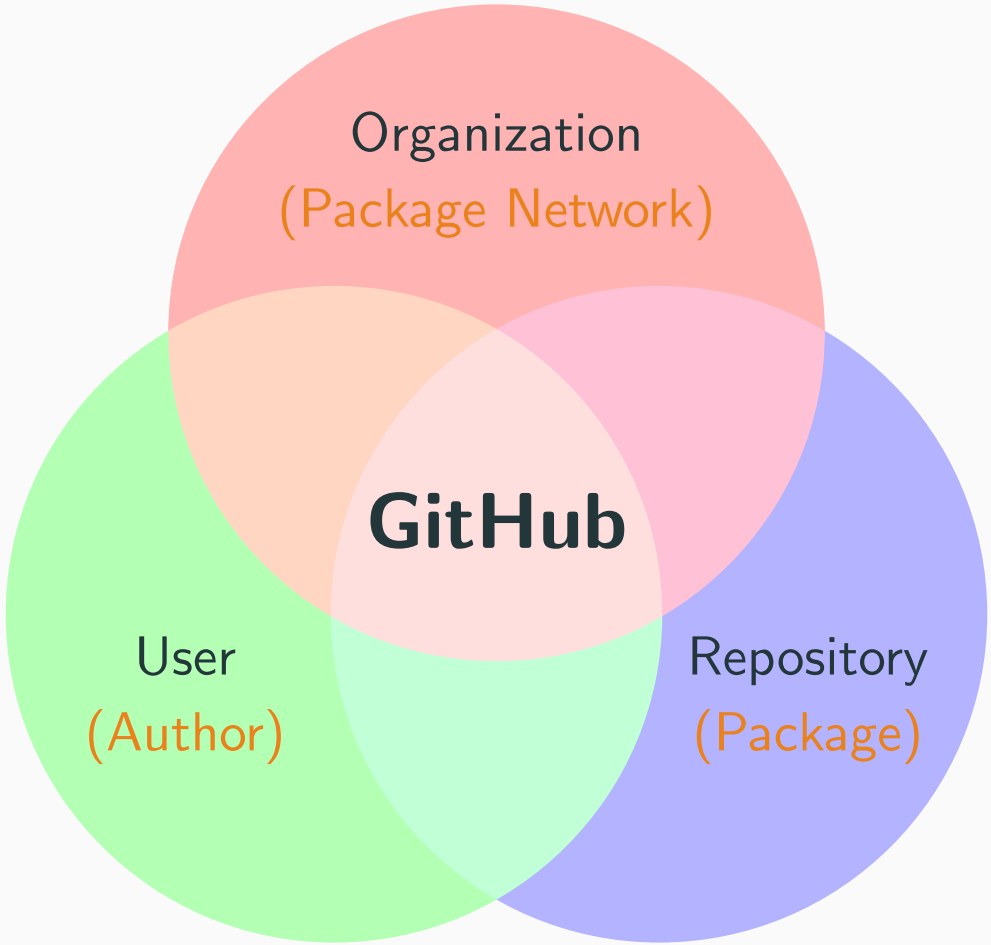

Figure 2: GitHub Structure vs. OSR Structure

Main Concepts and Objectives

Main Concepts and Objectives

BitQuery – a GitHub API driven and D3[1] based search engine for open source repositories.

BitQuery pursues two main objectives:

(I) Provide an automatic OSR categorization system for data science teams and software developers promoting discoverability, technology transfer and coexistence

(II) Establish visual data exploration and topic driven navigation of GitHub organizations for CRR and web deployment

[1] D3.js is a JavaScript library for producing dynamic, interactive data visualizations in web browsers

BitQuery as Visual Analytics app

The BitQuery architecture consists of three abstraction layers (following the visual analytics approach, [4]):

- GitHub API based parser layer (Data Management)
- Smart Data layer (Analysis)
- D3 Visu layer (Visualization)

Visual analytics mantra: Analyze first - show the important - zoom, filter and analyze further - details on demand, [5]

Visual Analytics layouts of BitQuery

VA-App - CRAN Edition

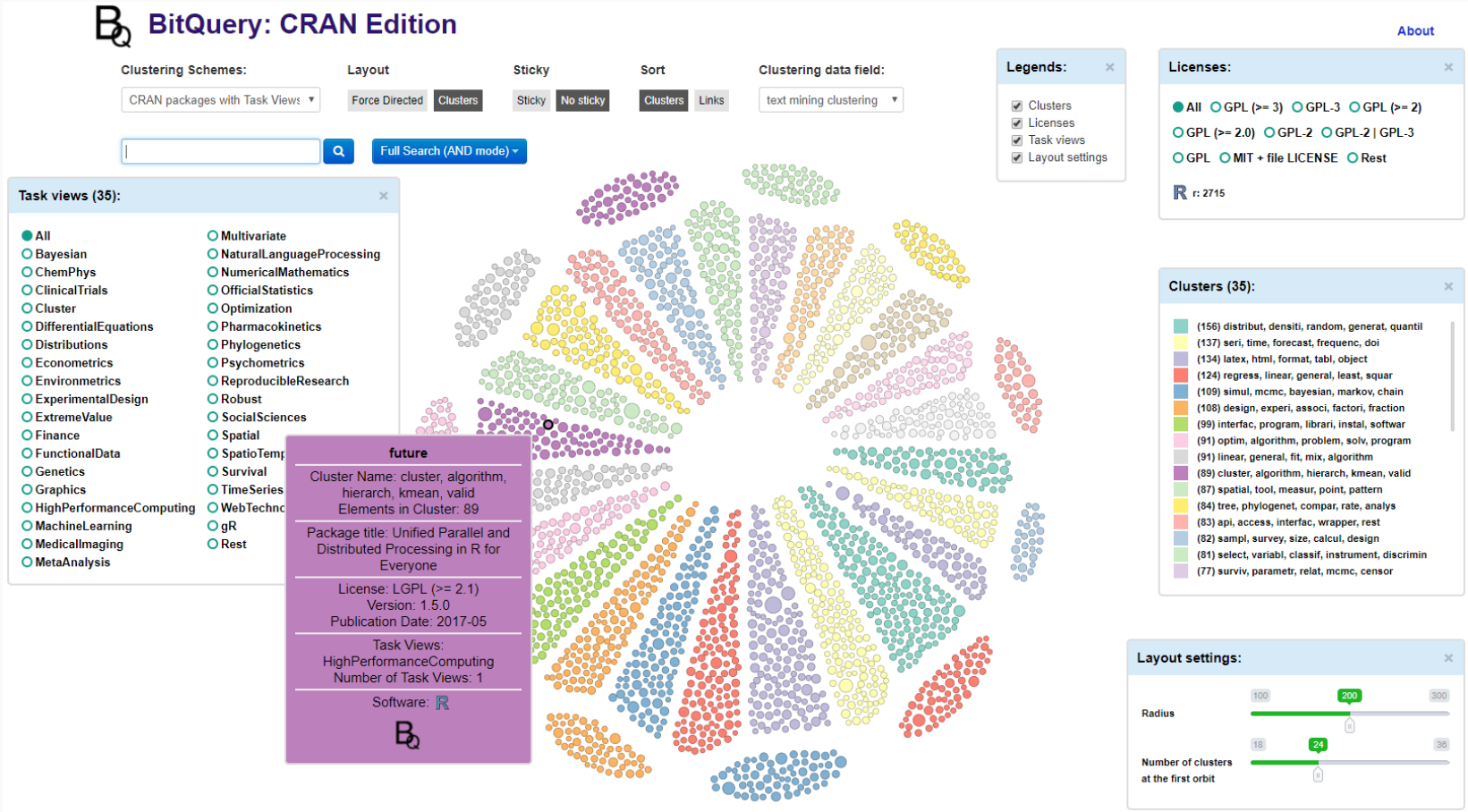

Figure 3: CRAN Edition: visual exploration of the R universe, a massive collection of all R packages on GitHub including CRAN and Bioconductor

VA-App - Special Edition

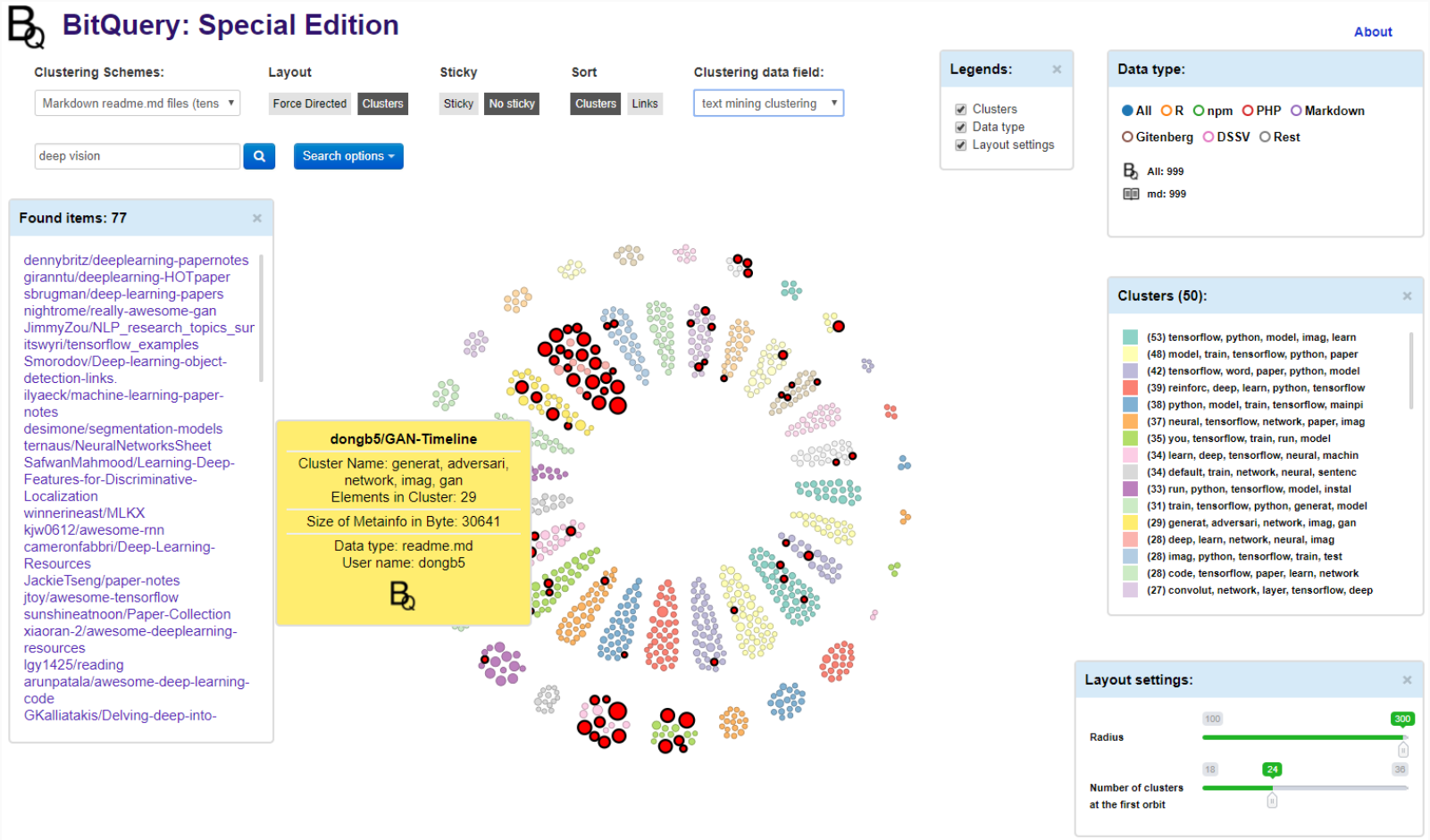

Figure 4: Special Edition: visual data mining of 4 different metadata types: YAML, DCF, JSON, Markdown with applications for the DSSV 2017 overview

VA-App - GitHub Edition

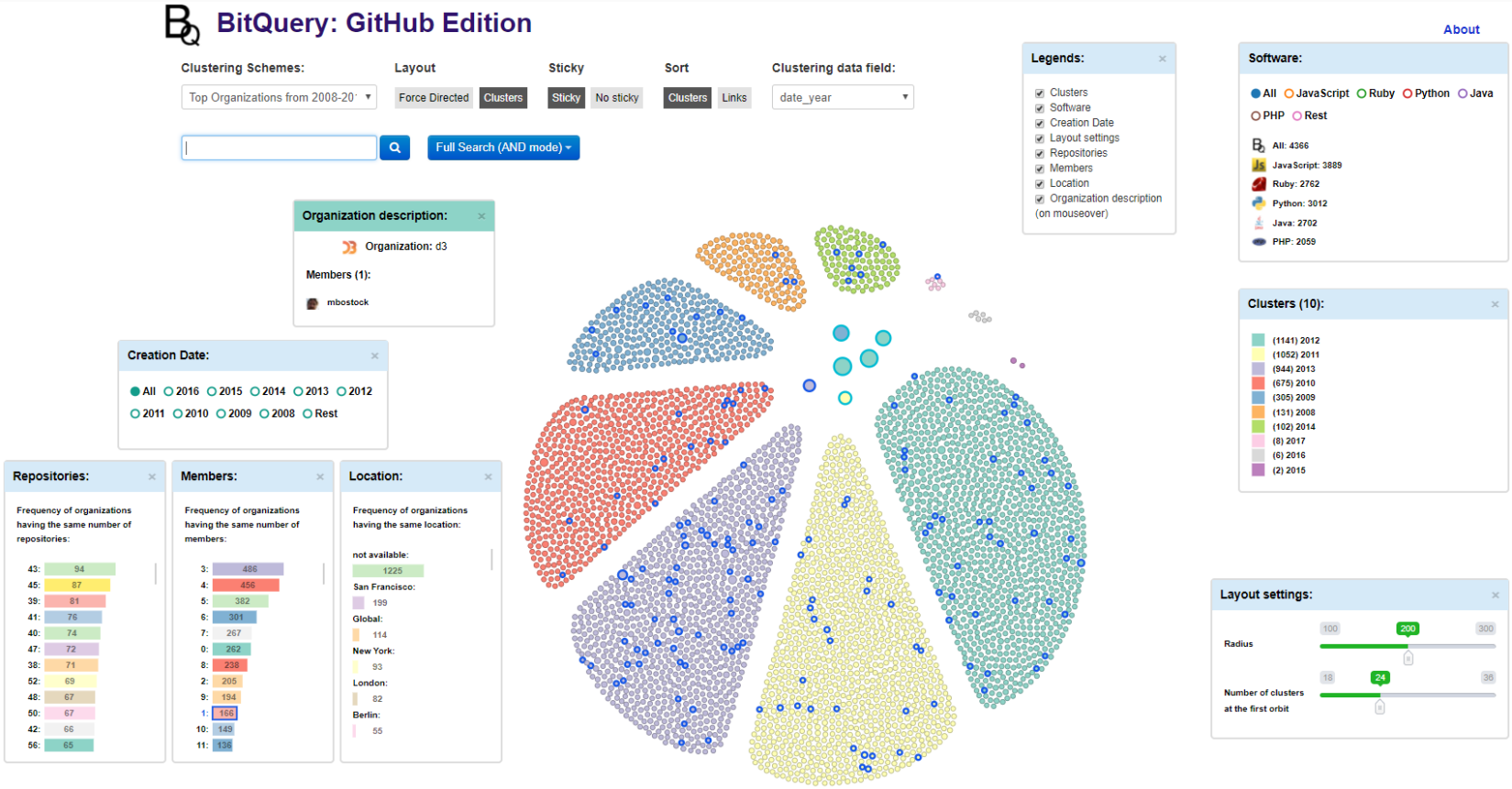

Figure 5: GitHub Edition: interactive visual knowledge discovery of top GitHub organizations (covering the time period 2008-2013)

VA-App - Infrastructure

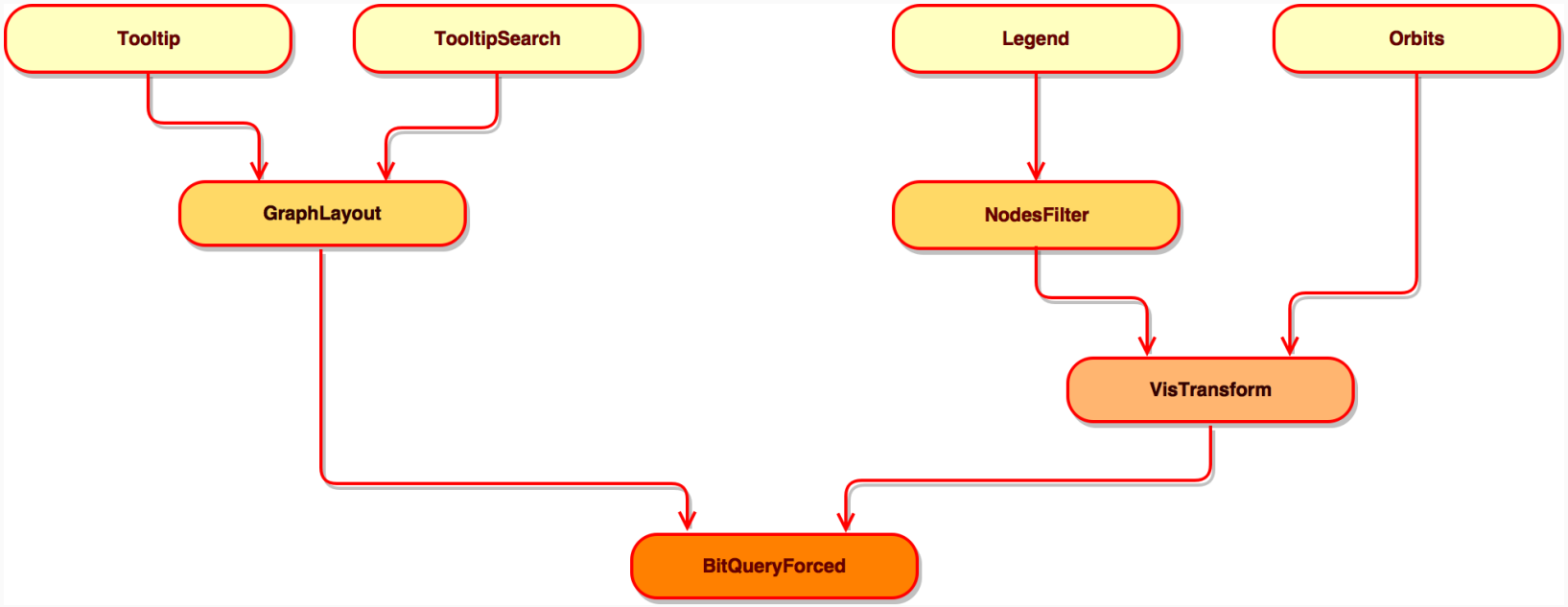

Figure 6: BitQuery VA-App Infrastructure, implemented via CoffeeScript classes. The corresponding libraries may be found on github.com/d3VA

- *overview:* Orbits, GraphLayout, VisTransform, BitQueryForced
- *zoom and filter:* TooltipSearch, Legend, NodesFilter
- *details-on-demand:* Tooltip

Smart Data Layer

Smart Data Layer - Infrastructure

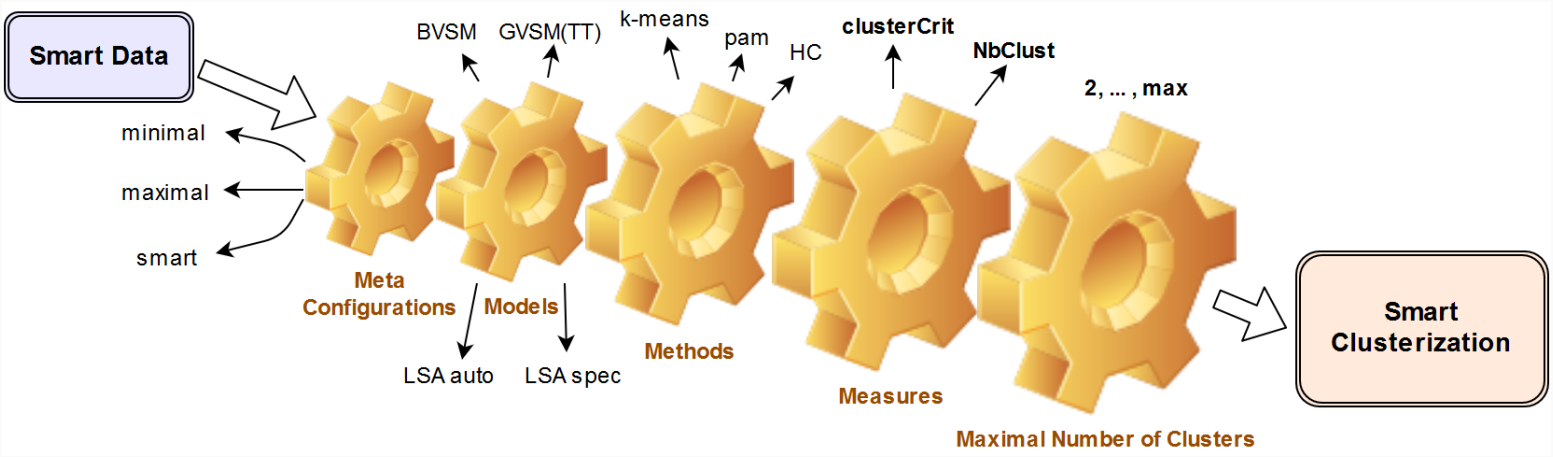

Figure 7: Smart clusterization

Smart clusterization incorporates 5 dimensions:

- **meta information** configurations
- vector space **models** & **text mining**
- **clustering methods**
- **measures** for **clustering quality** validation
- **number of clusters**

GitHub Mining infrastructure in R

1. taskviewsVA: Provides visual analytics tools for CRAN task views and associated packages via various D3.js and Three.js outputs.
2. rgithubS: Provides access to the GitHub v3 API. Special edition: search, statistics, parsers.
3. TManalyzer: Provides IR tools in 3 text mining models: BVSM, GVSM(TT) and LSA. It is complemented by metadata analytics and document clustering functionality.

Conclusion

Summary

- BitQuery is driven by a powerful GitHub Mining infrastructure in R [2] which allows to incorporate any GitHub organization and repository with its content for Big Data analytics
- 3 VA-App layouts of BitQuery: CRAN Edition, Special Edition, GitHub Edition
- 3 R packages: **taskviewsVA** [3], **rgithubS** [6], **tmAnalyzer** [1]

References i

L. Borke.
TManalyzer: IR tools in 3 text mining models: BVSM, GVSM(TT) and LSA, 2017.
R package version 0.5.0.

L. Borke and S. Bykovskaya.
GitHub Mining Infrastructure in R.
forthcoming, 2017.

L. Borke and S. Bykovskaya.
taskviewsVA: visual analytics tools for CRAN task views, 2017.

R package version 0.4.0.

References ii


D. Keim, G. Andrienko, J.-D. Fekete, C. Gorg, J. Kohlhammer, and G. Melançon.
Visual analytics: Definition, process, and challenges.
Lecture notes in computer science, 4950:154–176, 2008.


D. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler.
Challenges in visual data analysis.
In *Information Visualization, 2006. IV 2006. Tenth International Conference on*, pages 9–16. IEEE, 2006.


C. Scheidegger and L. Borke.
rgithubS: GitHub API bindings for R: search, statistics, parsers, 2017.
R package version 0.9.9.

VA-App - D3 HCA Multilevel View

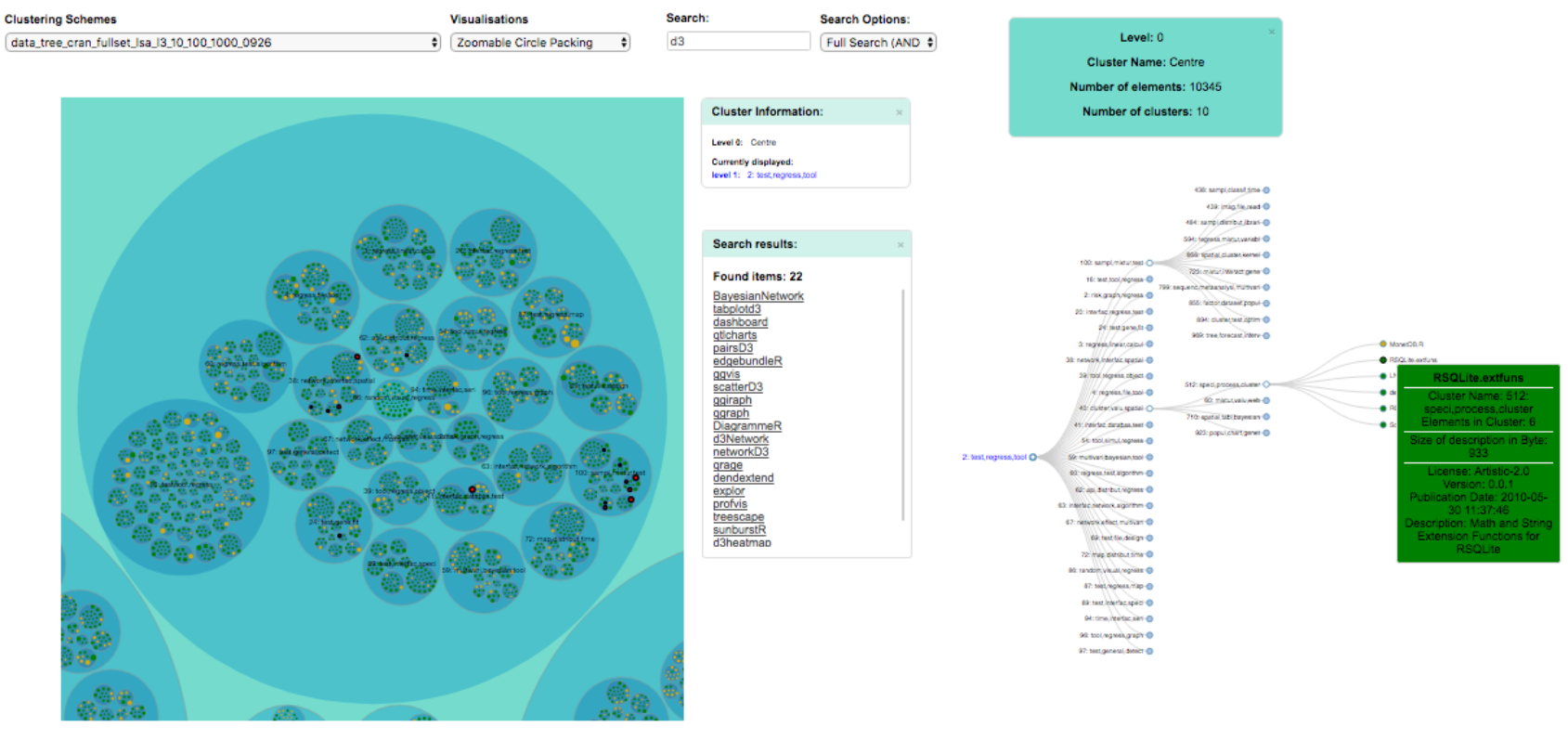

Figure 8: BitQuery VA-App view: D3 HCA Multilevel, powered by D3.js

VA-App - Data Projector View

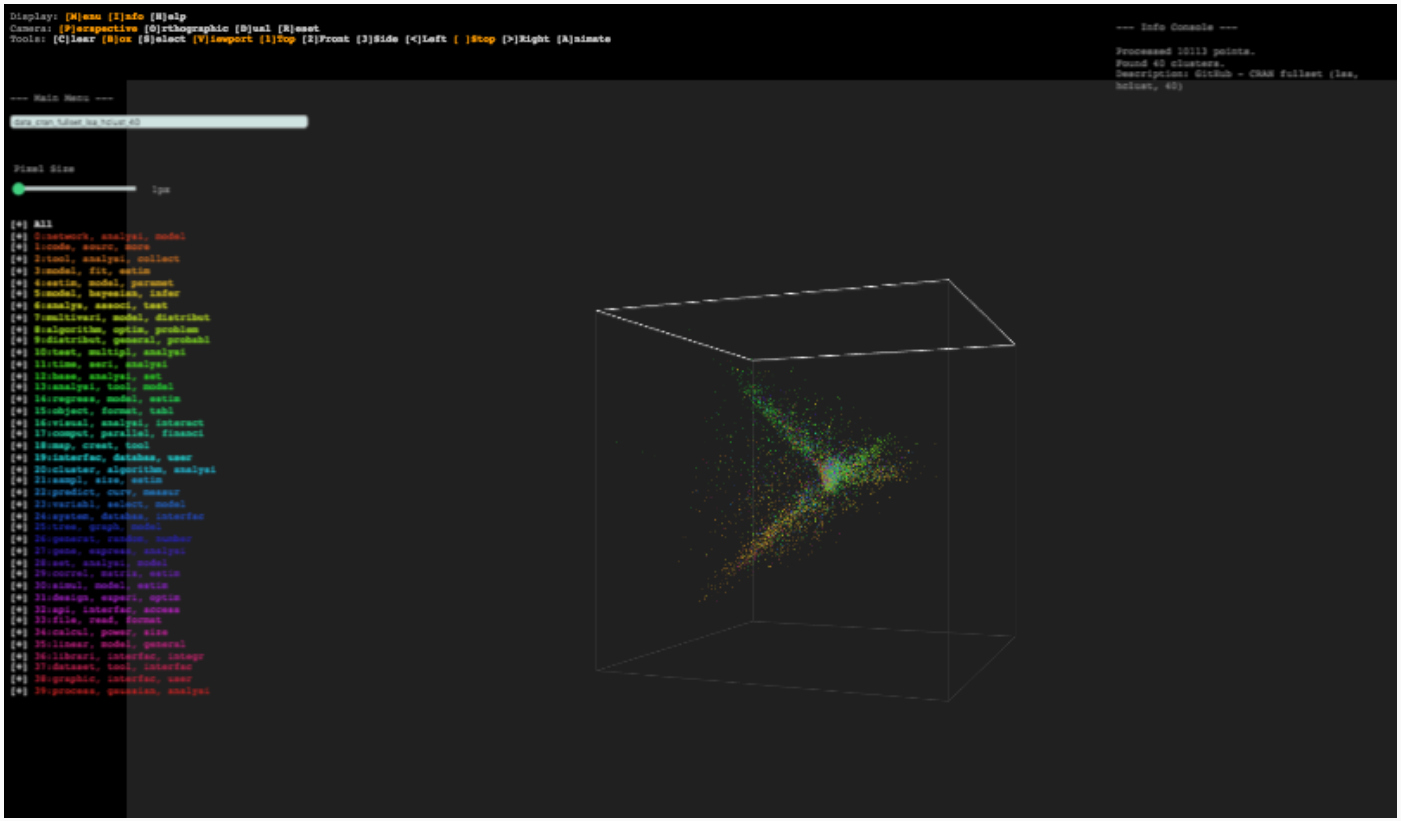

Figure 9: BitQuery VA-App view: Data Projector, powered by Three.js[2]

[2] JavaScript library for displaying animated 3D computer graphics