Computational Ecosystems for Social Science

Contact: ropengov.github.io

Markus Kainu, Joona Lehtomäki, Juuso Parkkinen, Juha Yrjölä, Måns Magnusson, Mikko Tolonen, Niko Ilomäki, Leo Lahti

Open data creates opportunities

The recent explosion in open data availability has created novel opportunities for research, journalism and citizen-science. High-quality machine readable data streams are increasingly available on political decision making, historical processes, welfare, traffic, and other aspects of society.

Need for analytical tools

Taking advantage of these new data streams in computational social science, digital humanities, and related fields needs novel analytical tools. First software libraries are now emerging and have a huge potential to contribute to transforming these fields. However, these resources are currently highly scattered and come in various formats, hindering wider adoption. Specific web-based tools are available, but more flexible computational tools are urgently needed for fully powered data processing and analysis.

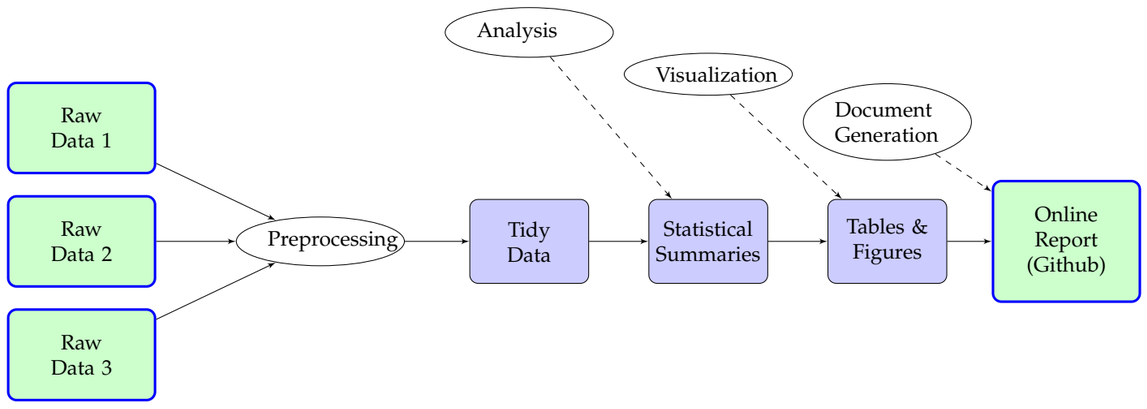

Figure 1: **Reproducible research workflow** Raw data sets are downloaded from original sources, tidied up, and integrated with other information. Statistical summaries, analyses and visualization are then automated with the aid of custom open source software libraries. The results are reported in web-based documents via automated document generation. The complete analysis workflow, including full access to every single detail from raw data to visualization, is shared publicly in distributed version control system (Github). The rOpenGov ecosystem provides dedicated R libraries to support reproducible research in the fields of computational social science and digital humanities. For the full source code to reproduce this poster, see PLACE FINAL VERSION in ROPENGOV

Eurostat open data: a reproducible example

```
p1 <- p + geom_smooth(method="lm")
p2 <- p + geom_smooth(method="glm")
p3 <- p + geom_smooth(method="loess")
p4 <- p + geom_smooth(method="gam")
grid.arrange(p1,p2,p3,p4, ncol = 2)
```

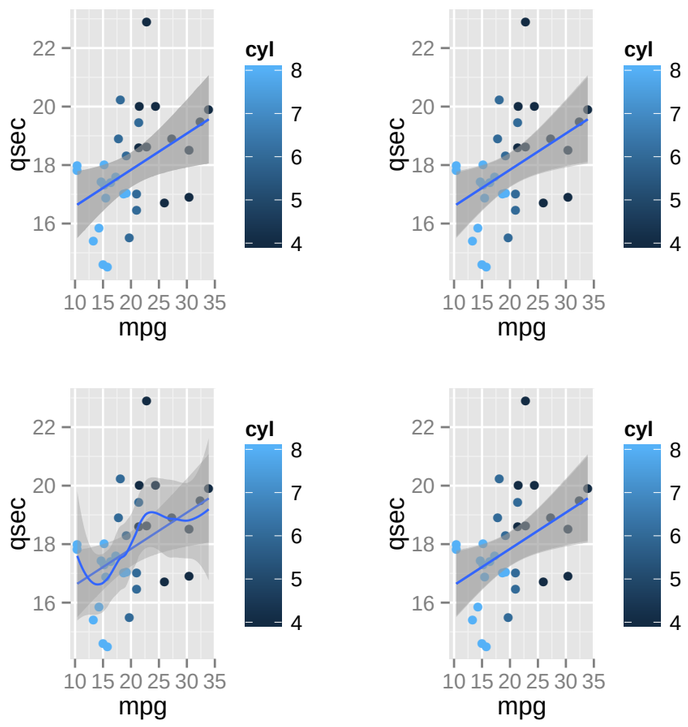

Figure 1: This is figure caption, too!

rOpenGov - An emerging ecosystem for CSS & DH

rOpenGov is a statistical ecosystem focused on open source data analysis algorithms relevant to computational social sciences and digital humanities. We build on experiences learned from similar initiatives in other fields, such as Bioconductor and rOpenSci. We use the R statistical programming language, which has a versatile computational ecosystem with rich statistical modeling and state-of-the-art visualization capabilities.

Reproducible research blog

The rOpenGov blog at ropengov.github.io is an emerging collection of example case studies that showcase the opportunities of reproducible open data analytics. The general-purpose research algorithms with a wider applicability are distributed as open source R packages. This provides **well-documented tools to download, preprocess, integrate, analyse, visualize and report digital data streams in a fully automated and transparent fashion**. This complements the existing R ecosystem by **focusing on methodologies relevant to Computational Social Science and Digital Humanities**.[1]

Open Government Data

- Regions: eurostat
- Countries: Russia, Finland, Poland, USA
- Cities: Helsinki
- Statistics Authorities (Finland, Denmark, Sweden, PX-Web)

Data processing and visualization

- GIS tools (Finland, OpenStreetMap, WFS..)
- Data anonymization
- Meteorology
- Health and Demography
- Political Science

Digital humanities and media

- Bibliographic analysis (Finland, UK, Europe)
- Media (Enigma, ProPublica, Sunlight Foundation, New York Times)

Parliamentary monitoring

- Election data (Austria, Finland, Russia, Huffpost Pollster)
- Quality of Government Institute

[1]JOS SAISI KAKSI PALSTAA, JOISTA TOISELLA OLISI MUUTAMA ESIMERKKIKUVA NÄISTÄ PAKETEISTA, OLIS KOVA - katotaan aluksi toi Eurostat-esimerkki, ja sen jälkeen et miten mahtuis + sopii. vois toimia mut pitää samalla varoa ettei tunge liikaa kamaa yhteen posteriin / LL]

Power of open source communities

Efficient data analysis relies on customized, reproducible analysis workflows that are best developed jointly by the user community. Availability of ready-made algorithms for standard data analysis tasks allows an individual researcher to avoid reinventing the wheel, leaving more time to solve the specific research problems. Solutions have emerged in data intensive research fields, such as bioinformatics and particle physics, based on open source statistical programming languages. The resulting communities, such as Bioconductor (http://bioconductor.org/) have proven highly successful, acting as an example for other fields to follow.

Ecosystem benefits

The ecosystem enables rapid development of scalable and interoperable software and provides tools to expand the quantitative methods base. The advantages of the open development model include:

- **Open source**: We use GitHub for shared version control. All contributions are openly licensed. This guarantees that the tools are freely available and the international scientific community remains the owner of the research software.
- **Reproducible documentation**: High-quality documentation is critical for package usability. We provide online tutorials with fully reproducible documentation on how to access and analyse specific data sources, and to report the statistical results.
- **Transparent research**: The programmatic approach makes it possible to publish the data analysis steps from raw data to the final results in full detail. To exemplify this, we publish reproducible case studies based on open data and algorithms in the rOpenGov blog.
- **Standardization**: A community-driven approach helps to pool scarce research resources and develop common standards for data analysis. Joint development ensures that the applicability of the tools extends beyond individual data sets and is compatible with other tools. Whereas different research projects can utilize the same standard algorithms to access and preprocess the data, the source code can be flexibly adapted to different tasks.

References

We are thankful for a number of developers for supporting this community. For a full list, see ropengov.github.io.

1. rOpenGov core team (2013). R ecosystem for open government data and computational social science. NIPS Machine Learning Open Source Software workshop (MLOSS). December 2013, Lake Tahoe, Nevada, US
2. S. Fortunato and C. Castellano (2012). Physics peeks into the ballot box. Physics Today 65:74
3. G. King, J. Pan and M. E. Roberts (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression. American Political Science Review, 107(02), 326–343
4. M. L. Jockers (2013). Macroanalysis: Digital Methods and Literary History. University of Illinois Press.
5. S. Chou, W. Li and R. Sridharan, Democratizing Data Science.
6. D. Lazer, et al. (2009). Computational Social Science 323, 721–723