Computational ecosystems for social science*

Markus Kainu †[1], Joona Lehtomäki ‡[2], Juuso Parkkinen §[3], Juha Yrjölä ¶[4], Måns Magnusson ‖[5], and Leo Lahti **[6]

[1]Aleksanteri Institute, University of Helsinki, Finland
[2]Department of Biosciences, University of Helsinki, Finland
[3]Reaktor Innovations Oy, Finland
[4]Kansan Muisti ry, Finland
[5]Linköping University, Sweden
[6]Department of Veterinary Bioscience, University of Helsinki, Finland

May 29, 2015

Keywords: social science; elections; open government data; statistical programming; machine learning

Open availability of scientific and governmental data is increasing rapidly, creating novel opportunities for quantitative social sciences, data journalism, and citizen science. High-quality machine readable data sources relevant to social sciences are now becoming available across the globe related to political decision making, welfare, stratification, traffic, and other fields. In addition to social network analysis that has gained popularity in the field (Lazer et al., 2006), there are ample opportunities to expand the scope of computational social science research based on open governmental data sources. However, these resources are currently highly scattered and come in various formats, hindering their wider adoption. Various web-based tools for specific analysis tasks are available, but more flexible computational tools are urgently needed for more flexible and precise data processing and analyses. In particular, flexible algorithms are needed to clean up raw data and carry out fast expert-driven interactive data exploration and visualization.

There have been few attempts (King et all. 2013, Jockers 2013) within the academia to systematically bring the new data streams and computational methods into use in social science. We emphasize the importance of domain specific computational ecosystems in this task, and propose as one solution the rOpenGov community-driven ecosystem. It is based on the R statistical programming language, which has a strong track record in other fields, a versatile computational ecosystem and an active developer community. Within the R ecosystem there is a well established software release system, support for rich statistical simulation and modeling functions, and state-of-the-art visualization capabilities for addressing the diversity of analysis tasks in social sciences. With rOpenGov, we complement this existing ecosystem by providing a versatile set of R-based tools to access, preprocess and analyze open governmental data streams specifically relevant to social sciences.

The rOpenGov project builds on lessons learned from similar initiatives in other fields such as Bioconductor and rOpenSci. As these examples have shown, a community-driven open source approach is central as data availability is increasing, helping to pool scarce research resources. Using transparent research tools improve research reproducibility by providing standard workflows that can be easily adapted to different analytical tasks. The shared software ecosystem enables rapid development of extensible, scalable, and interoperable software, improvements through bug fixing, and provides tools to explore and expand the quantitative social science research methods.

The rOpenGov project launched in 2013 and is maintained by a core team of PhD-level computational scientists. The core team provides support for package authors and the user community by maintaining the infrastructure, reviewing new packages, and proposing recommended guidelines. An international contributor base has emerged from countries including the USA, Finland, Sweden, Austria, and Poland. The project is maintained in GitHub.

Election data analytics is a specific example of an active community project with implications for research and public outreach. We have created a public website kansanmuisti.fi/ to collect and organize openly available election related data in Finland, including a million social media feeds from election candidates collected since 2012, historical election

*Extended abstract submitted for International Conference on Computational Social Science in June 8-11, 2015 in Helsinki, Finland
†markuskainu@gmail.com
‡joona.lehtomaki@helsinki.fi
§juuso.parkkinen@iki.fi
¶juha.yrjola@iki.fi
‖mans.magnusson@gmail.com
**leo.lahti@iki.fi

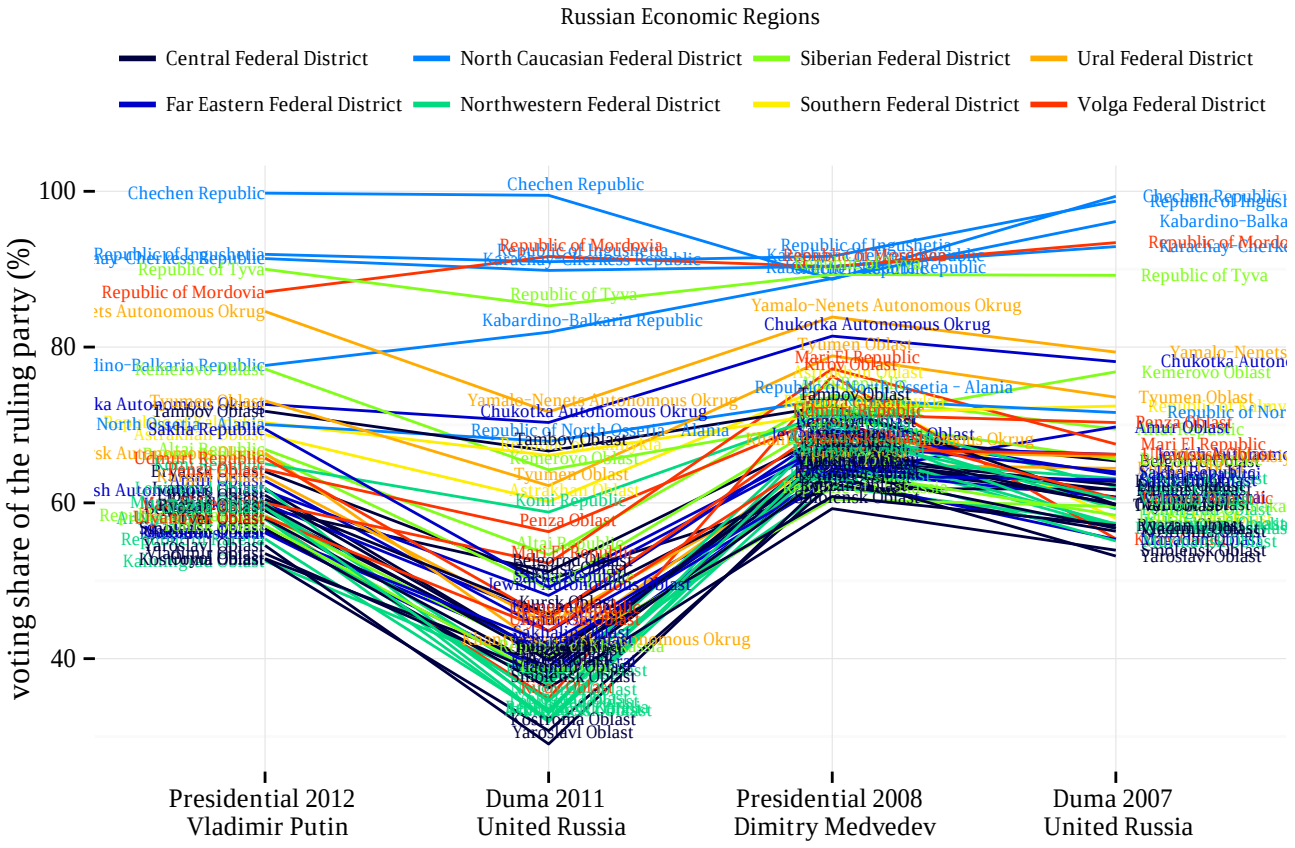

Figure 1: Support for the ruling party *United Russia* in the four latest national elections in regions of Russia (Source: Central Election Commission of the Russian Federation)

statistics, funding information of the candidates, and personal information (age, gender, home town) regarding the election candidates. The data is being collected at a continuous basis and made available in machine readable format through an open API at datavaalit.fi.

In addition to creating a comprehensive collection of election data, we are creating tools for computational analysis of this information. Importantly, we are creating standardized representation formats for election data in R that would allow commensurable data representations across national borders and subsequently accelerating the development of general-purpose computational tools for election data. This borrows the success of analogous representations for gene expression data in bioinformatics during the past decade. As the election data sets are brought within a common representation format, it is possible to create general-purpose analytical tools. We are now building open source algorithms based on previously proposed models to predict election results, detect election fraud, and compare different varieties of election systems such as elections in EU and Russia (see Figure 1).

The launching of the rOpenGov project has been successful, and the current infrastructure provides a strong basis for building up scalable open source ecosystem for flexible data analytics dedicated to computational social sciences.

References

[1] S. Kasberger (2012). Grazwahl: Data Analysis and Visualizations of the communal elections in Graz.*R package*

[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