Open Data Analytics and Reproducible Research

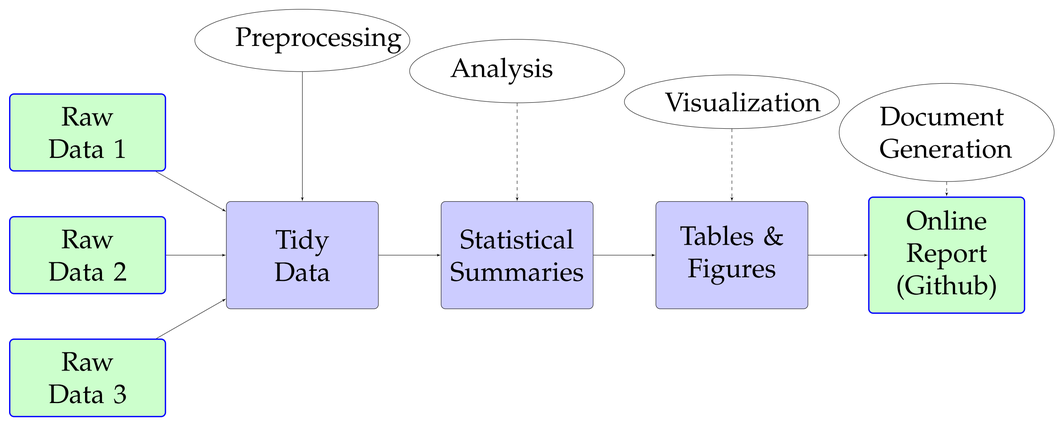
Markus Kainu, Joona Lehtomäki, Juuso Parkkinen, Juha Yrjölä, Måns Magnusson, Mikko Tolonen, Niko Ilomäki, Leo Lahti

Contact: http://ropengov.github.io

Open data analytics The recent explosion in open data availability has created novel opportunities for research. Efficient data analytical tools are crucial for taking full advantage of these new information resources. Custom software libraries are now rapidly emerging and have a huge potential to contribute to transforming computational social sciences, digital humanities, and related fields.

Advantages of the open development model Efficient data analysis relies on custom workflows that are best developed jointly by the user community, as already demonstrated in bioinformatics (Bioconductor), particle physics, and other fields (rOpenSci). Similar communities are now shaping up in social sciences and humanities. The open development model has many benefits (Ioannidis 2014, Morin et al. 2012):

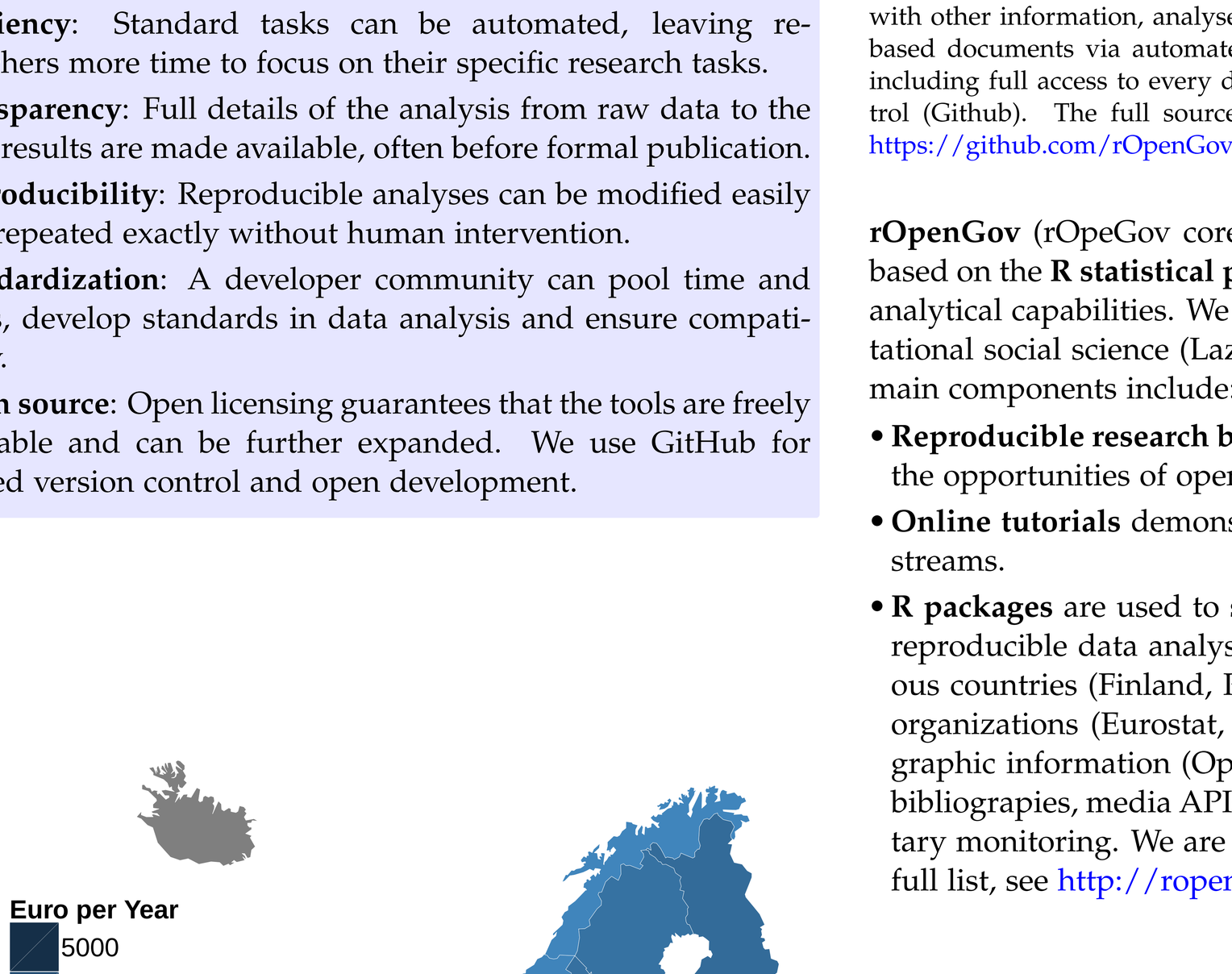
- **Efficiency**: Standard tasks can be automated, leaving researchers more time to focus on their specific research tasks.
- **Transparency**: Full details of the analysis from raw data to the final results are made available, often before formal publication.
- **Reproducibility**: Reproducible analyses can be modified easily and repeated exactly without human intervention.
- **Standardization**: A developer community can pool time and skills, develop standards in data analysis and ensure compatibility.
- **Open source**: Open licensing guarantees that the tools are freely available and can be further expanded. We use GitHub for shared version control and open development.



Reproducible research workflow R is a popular open source statistical programming language. The versatile ecosystem of R packages enables a completely automated analysis from raw data to the final reports. We create dedicated tools to support reproducible research in computational social science and digital humanities, helping to automatize many standard data analysis tasks in these fields. Raw data sets are downloaded from original sources, preprocessed, integrated with other information, analysed and visualized. The results are reported in web-based documents via automated document generation. The complete workflow, including full access to every detail, is shared publicly in distributed version control (Github). The full source code of this poster, for instance, is available at https://github.com/rOpenGov/poster/tree/master/2015-ICCSS.

rOpenGov (rOpeGov core team, 2013) is an open source ecosystem based on the **R statistical programming language** which has rich data analytical capabilities. We develop data analysis methods for computational social science (Lazer et al. 2009) and digital humanities. The main components include:

- **Reproducible research blog** at http://ropengov.github.io highlights the opportunities of open data analytics.
- **Online tutorials** demonstrate how to access and analyse open data streams.
- **R packages** are used to share computational algorithms to support reproducible data analysis. We provide tools for open data in various countries (Finland, Poland, Russia, USA), cities (Helsinki), and organizations (Eurostat, PX-Web, QOG), data anonymization, geographic information (OpenStreetMap, WFS), weather, demography, bibliograpies, media APIs, political science, elections and parliamentary monitoring. We are thankful for a number of developers. For a full list, see http://ropengov.github.io/projects




```
# Load R packages
library("eurostat")

# Read data
map <- readOGR("NUTS_2010_60M_SH/Data")
...

# Draw image
p <- ggplot(data=map, aes(long,lat,group=group))
p <- p + geom_polygon(aes(fill = values), color= alpha("white", 1/2),  size=.2)
...

# Print the map
print(p)
```

The automatically generated map above shows *disposable incomes of private households at NUTS2 level regions in European Union in 2011*. This poster, including the Eurostat analysis example above, is fully reproducible; for the full source code of this poster, see https://github.com/ropengov/poster

References

1. J. Ioannidis (2014). How to Make More Published Research True? PLoS Medicine 11(10): e1001747.
2. D. Lazer et al. (2009). Computational Social Science 323, 721–723
3. A. Morin et al. (2012). Research priorities. Shining light into black boxes. Science 336, 159-160.
4. 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