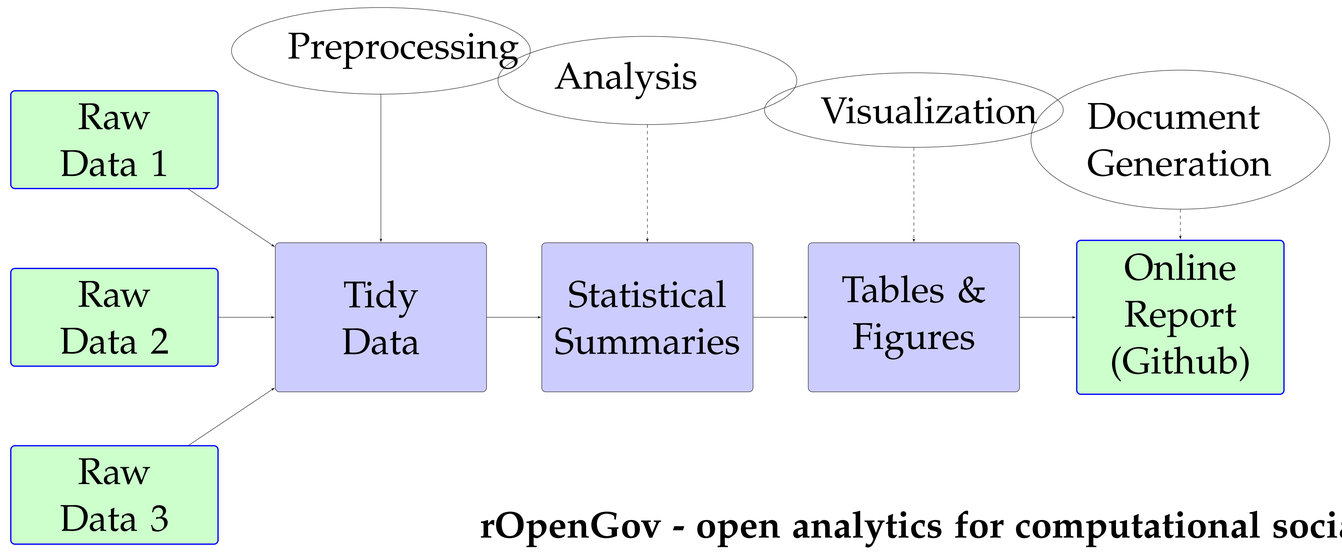
Open Computational Ecosystems and Reproducible Research

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

Contact: http://ropengov.github.io

Open data analysis

The recent explosion in open data availability has created novel opportunities for research, journalism and citizen science.
Efficient data analytical tools are crucial for taking full advantage of digital data streams. 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 customized analysis workflows that are best developed jointly by the user community. Successful open source communities have emerged in research fields, such as bioinformatics and particle physics, and are now shaping up in social sciences and humanities. The open analytics has many benefits (Ioannidis 2014, Morin et al. 2012).

- **Efficiency**: Many standard analysis tasks can automated, leaving more time for individual researchers to focus on their specific problems.
- **Transparency**: Full details of the analysis from raw data to the final results are available.
- **Reproducibility**: Reproducible analyses can be repeated exactly without human intervention. The analysis process is more transparent and can be easily expanded.
- **Standardization**: A community-driven development helps to pool scarce research resources, develop standards and ensure compatibility in data analysis.
- **Open source**: Openly licensed contributions guarantee that the tools are freely available for the international scientific community.

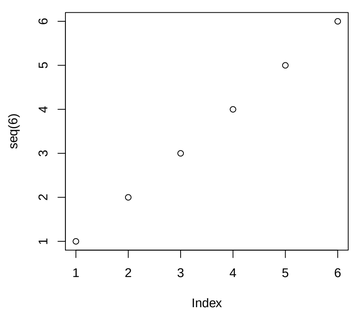
rOpenGov - open analytics for computational social sciences and digital humanities

rOpenGov (rOpeGov core team, 2013) is a statistical ecosystem focused on open source data analysis algorithms relevant to computational social sciences (Lazer et al. 2009) 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 rich statistical modeling and visualization capabilities. Main components of rOpenGov include:

- **Online tutorials** reporting how to access and analyse specific open data sources (see the 'rOpenGov packages' box).
- **Reproducible research blog** at (http://ropengov.github.io) presents case studies on open data analytics.
- **Shared version control** We use GitHub for shared version control and open development.
- **R packages** General-purpose research algorithms are distributed as open source R packages to provide dedicated tools to download, preprocess, integrate, analyse, visualize and report digital data streams in a fully automated and transparent fashion. At present, the collection includes tools for countries (Russia, Finland, Poland, USA), cities (Helsinki), Statistics authorities (Eurostat, Finland, Denmark, Sweden, PX-Web, QOG), Data anonymization, Geographic information (OpenStreetMap, WFS, Finland), Meteorology, Health and demography, Bibliographic analysis (Finland, UK, Europe), Media (Enigma, ProPublica, Sunlight Foundation, New York Times), Political science, Elections and parliament (Austria, Finland, Russia, US Election Polls). For a full listing and links, see http://ropengov.github.io/projects/

Eurostat open data: a reproducible example

```
plot(seq(6))
```



This poster, including the Eurostat analysis example, is fully reproducible. ownload the full source code at http://...

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
   Political Science Review, 107(02), 326–343

We are thankful for a number of developers. For a full list, see http://ropengov.github.io