Open Software - Restricted Data: A Suicide/Climate Case Study.

Ivan Hanigan[1], David Fisher[2], Steven McEachern[3]

[1]National Centre for Epidemiology and Population Health (ANU)
[2]Information Technology Services (ANU)
[3]Australian Data Archives (ANU)

June 26, 2013

Aim

- Restrictions on data access have increased recently
- Concerns regarding reproducibility of data analyses
- Access to data and analytic software addresses the: **Replicability Crisis**, (Peng 2011, *Science*, 334;6060)
- We built a safe Sever/Client IT environment for this
- We show a Case Study of Suicide and Climate Impacts research

Methods

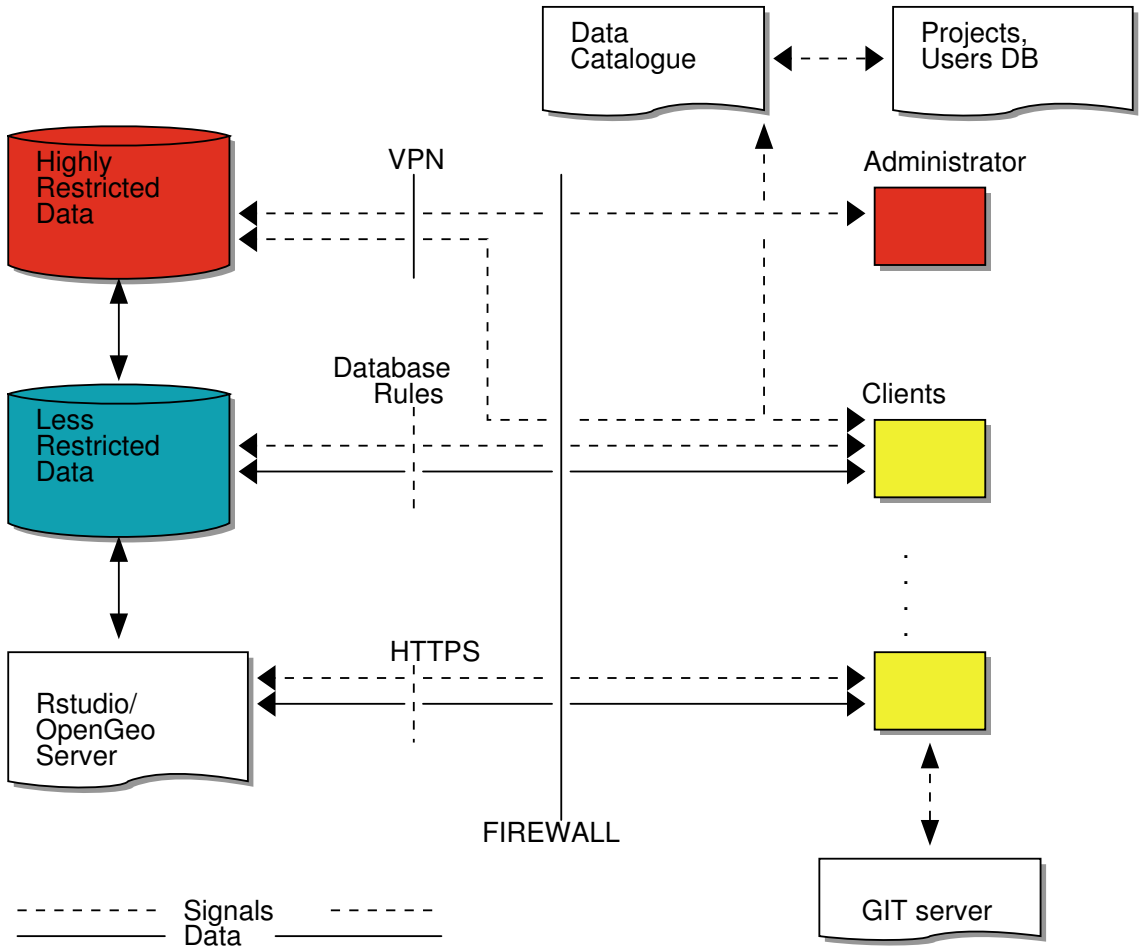

Figure: 1. System Design

Results (Hanigan et al, 2012, *PNAS*, 109;35)

- Restricted Health and Climate data and
- Less Restricted Population data

(Colours refer to data storage and access rules shown in Figure 1).

$$
\begin{aligned}
log(O_{ijk}) \quad = \quad & s(ExposureVariable) + OtherExplanators \\
& + AgeGroup_i + Sex_j \\
& + SpatialZone_k \\
& + sin(Time \times 2 \times \pi) + cos(Time \times 2 \times \pi) \\
& + Trend \\
& + offset(log(Pop_{ijk}))
\end{aligned}
$$

Where:
O_{ijk} = Outcome (counts) by Age_i, Sex_j and $SpatialZone_k$
ExposureVariable = Data with Restrictive Intellectual Property (IP)
OtherExplanators = Other Less Restricted Explanatory variables
s( ) = penalized regression splines
$SpatialZone_k$ = Less Restricted data representing the $SpatialZone_k$
Trend = Longterm smooth trend(s)
Pop_{ijk} = interpolated Census populations, by time in each group

Future (Bambrick et al, 2008, Garnaut Review)

$$Y_{ijk} = \sum_{lm} (e^{(\beta_{ijk} \times X_{lm})} - 1) \times BaselineRate_{jkl} \times Population_{jklm}$$

Where:

β_{ijk} = the ExposureVariable coefficient for $zone_i$, age_j and sex_k

X_{lm} = Projected Future ExposureVariables with Restrictive IP

$BaselineRate_{jkl}$ = avgDeathsPerTime/avgPopPerTime in age_j, sex_k and $zone_l$

$Population_{jklm}$ = projected populations by age_j, sex_k, $zone_l$ and $time_m$ (With Less Restrictions)

Conclusion

This system:

- Enables data analysis in a safe environment
- Allows comparison of multiple climate scenarios and assumptions
- Demonstrated with a Climate/Health Impact Assessment
- And this is Reproducible

Acknowledgements

This project is supported by the Australian National Data Service through the National Collaborative Research Infrastructure Strategy Program and the Education Investment Fund (EIF) Super Science Initiative.
More information from:

- ivan.hanigan@gmail.com
- http://opensoftware-restricteddata.github.io

References

Roger D Peng.

Reproducible research in computational science.

Science (New York, N.Y.), 334(6060):1226–7, December 2011.

I. C. Hanigan, C. D. Butler, P. N. Kokic, and M. F. Hutchinson.

Suicide and drought in New South Wales, Australia, 1970-2007.

Proceedings of the National Academy of Sciences, pages 1112965109–, August 2012.

Hilary J Bambrick, Keith B G Dear, RE Woodruff, Ivan Charles Hanigan, and Anthony J McMichael.

The impacts of climate change on three health outcomes: temperature-related mortality and hospitalisations, salmonellosis and other bacterial gastroenteritis, and population at risk from dengue.

Technical report, Garnaut Climate Change Review, Canberra, 2008.