Using git and GitHub with R

a **statsTeachR** resource

Principles of Reproducible Research – Definition

Reproducible research has been defined in the scientific community as published scientific work that can be recreated using code and data made available by the authors:

- Creating reproducible research requires authors to carefully document approaches used to process, manage, analyze, and visualize data.
- It also requires authors to have a foundational understanding of the uncertainty that underlies the statistical model they use to describe their data.

Principles of Reproducible Research – A Brief History

- Roots of reproducible research can be traced to the concept of literate programming heralded by Donald Knuth

  – Knuth, D. E. (1992). Literate Programming (1st ed.). Center for the Study of Language and Information.

- Concept operationalized in 2002 by Friederic Leisch with introduction of Sweave, a program that allows the user to weave together R code and natural language descriptions

  – Leisch, F. (2002a). Sweave. Dynamic generation of statistical reports using literate data analysis. SFB Adaptive Information Systems and Modelling in Economics and Management Science, WU Vienna University of Economics and Business; Leisch, F. (2002b). Sweave, part I: Mixing R and LaTeX. R News, 2/3, 2831.

- Importance of reproducibility discussed in vast array of fields, from econometrics, epidemiology and biostatistics, bioinformatics, and engineering

  – Koenker, R. (1996). Reproducible econometric research. Retrieved September 17, 2012, from: `http://www.econ.uiuc.edu/~roger/research/repro/repro.html`; Peng, R. D. (2009). Reproducible research and Biostatistics. Biostatistics, 10(3), 405408. doi:10.1093/biostatistics/kxp014; Gentleman, R. (2005). Reproducible research: a bioinformatics case study. Statistical applications in genetics and molecular biology, 4, Article2. doi:10.2202/1544-6115.1034; Vandewalle, P., Barrenetxea, G., Jovanovic, I., Ridolfi, A., & Vetterli, M. (2007). Experiences with Reproducible Research in Various Facets of Signal Processing Research. IEEE International Conference on Acoustics, Speech and Signal Processing. Proceedings, 4, IV1256. doi:10.1109/ICASSP.2007.367304)

Some journals are coming around...

- Publications
- Innovation
- Open Access
- Newsroom
- Community

Data Access for the Open Access Literature: PLOS's Data Policy

Posted on December 12, 2013 by Theo Bloom

Data are any and all of the digital materials that are collected and analyzed in the pursuit of scientific advances. In line with Open Access to research articles themselves, PLOS strongly believes that to best foster scientific progress, the underlying data should be made freely available for researchers to use, wherever this is legal and ethical. Data availability allows replication, reanalysis, new analysis, interpretation, or inclusion into meta-analyses, and facilitates reproducibility of research, all providing a better 'bang for the buck' out of scientific research, much of which is funded from public or nonprofit sources. Ultimately, all of these considerations aside, our viewpoint is quite simple: ensuring access to the underlying data should be an intrinsic part of the scientific publishing process.

Some journals are coming around...

OXFORD JOURNALS CONTACT US MY BASKET MY ACCOUNT

Biostatistics

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE SEARCH

Oxford Journals › Science & Mathematics › Biostatistics › Volume 10, Issue 3 › Pp. 405-408.

Submit your papers now

Reproducible research and *Biostatistics*

1. INTRODUCTION AND MOTIVATION

The replication of scientific findings using independent investigators, methods, data, equipment, and protocols has long been, and will continue to be, the standard by which scientific claims are evaluated. However, in many fields of study there are examples of scientific investigations that cannot be fully replicated because of a lack of time or resources. In such a situation, there is a need for a minimum standard that can fill the void between full replication and nothing. One candidate for this minimum standard is "reproducible research", which requires that data sets and computer code be made available to others for verifying published results and conducting alternative analyses.

The need for publishing reproducible research is increasing for a number of reasons. Investigators are more frequently examining weak associations and complex interactions for which the data contain a low signal–to–noise ratio. New technologies allow scientists in all areas to compile complex high–dimensional databases. The ubiquity of powerful statistical and computing capabilities allows investigators to explore those databases and identify associations of potential interest. However, with the increase in data and computing power comes a greater potential for identifying spurious associations. In addition to these developments, recent reports of

Making your research reproducible

General purpose reproducible research tools

- Version control (e.g. git, subservsion)
- Code in the cloud (e.g. GitHub.com, BitBucket.com)
- Data in the cloud (e.g. GoogleDrive, Harvard Dataverse Network, GenBank, Dryad, FigShare)
- `make`: a convenient command line tool for stitching together large, multi-stage analyses

Reproducible research tools for R

- R/RStudio
- Dynamic documents: knitr, RMarkdown, Sweave
- Package management/version control: `packrat`

Version control systems

Version control systems maintain a database on your computer that allows you to log all changes to text-based files.

Common VCS

- git
- subversion (svn)
- mercurial
- ...

Version control and reproducibility

Why version control?

- allows you to roll back to previous versions easily
- allows you to try things out without disrupting working code
- allows you to flag outputs (e.g. analyses, reports) as being generated by certain versions of code
- if in the cloud, everything is backed up!

Version control systems

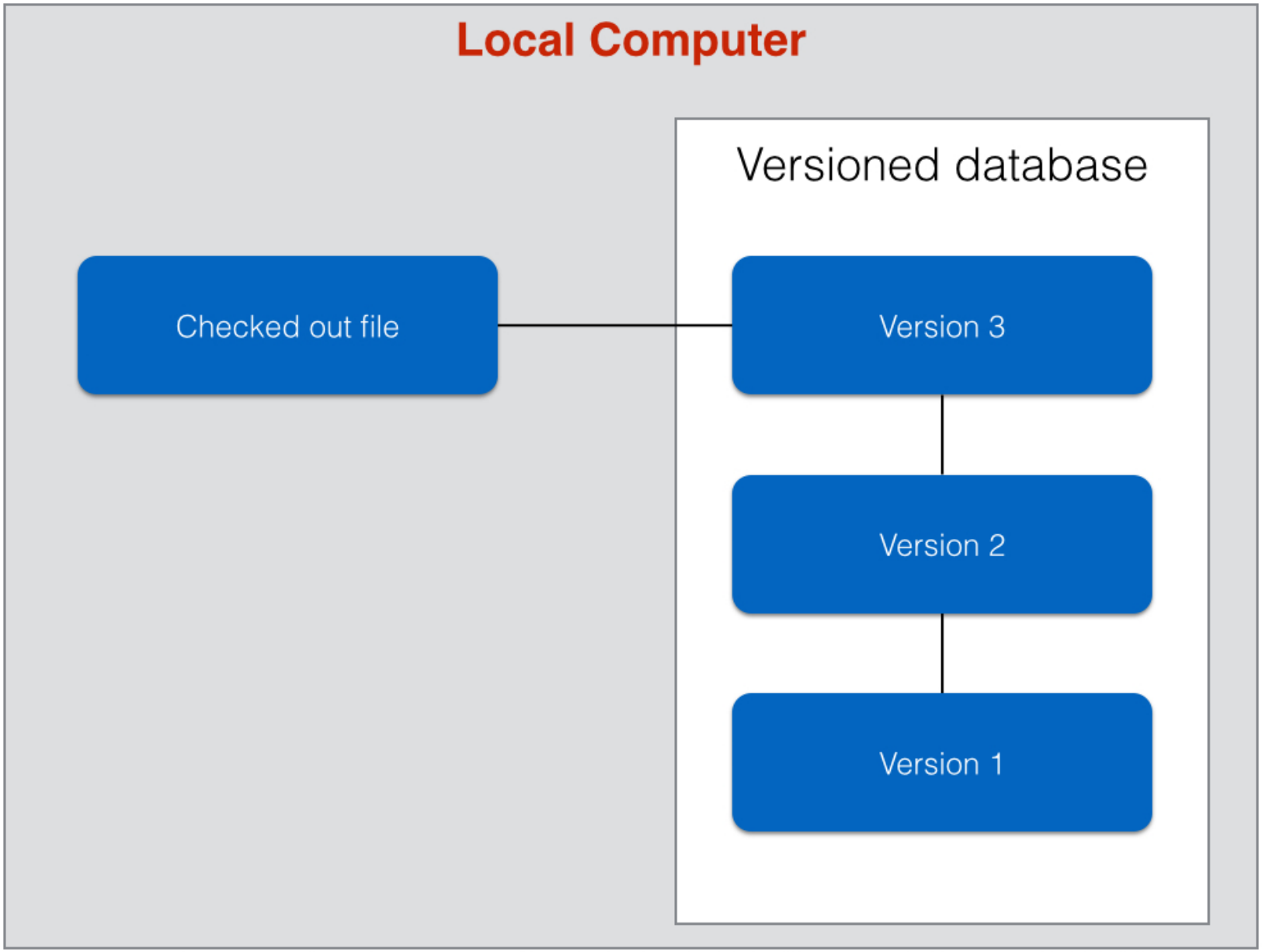

Image adapted from http://git-scm.com/book/en/Getting-Started-About-Version-Control, accessed 6 Feb 2014

Version control systems (git flavored)

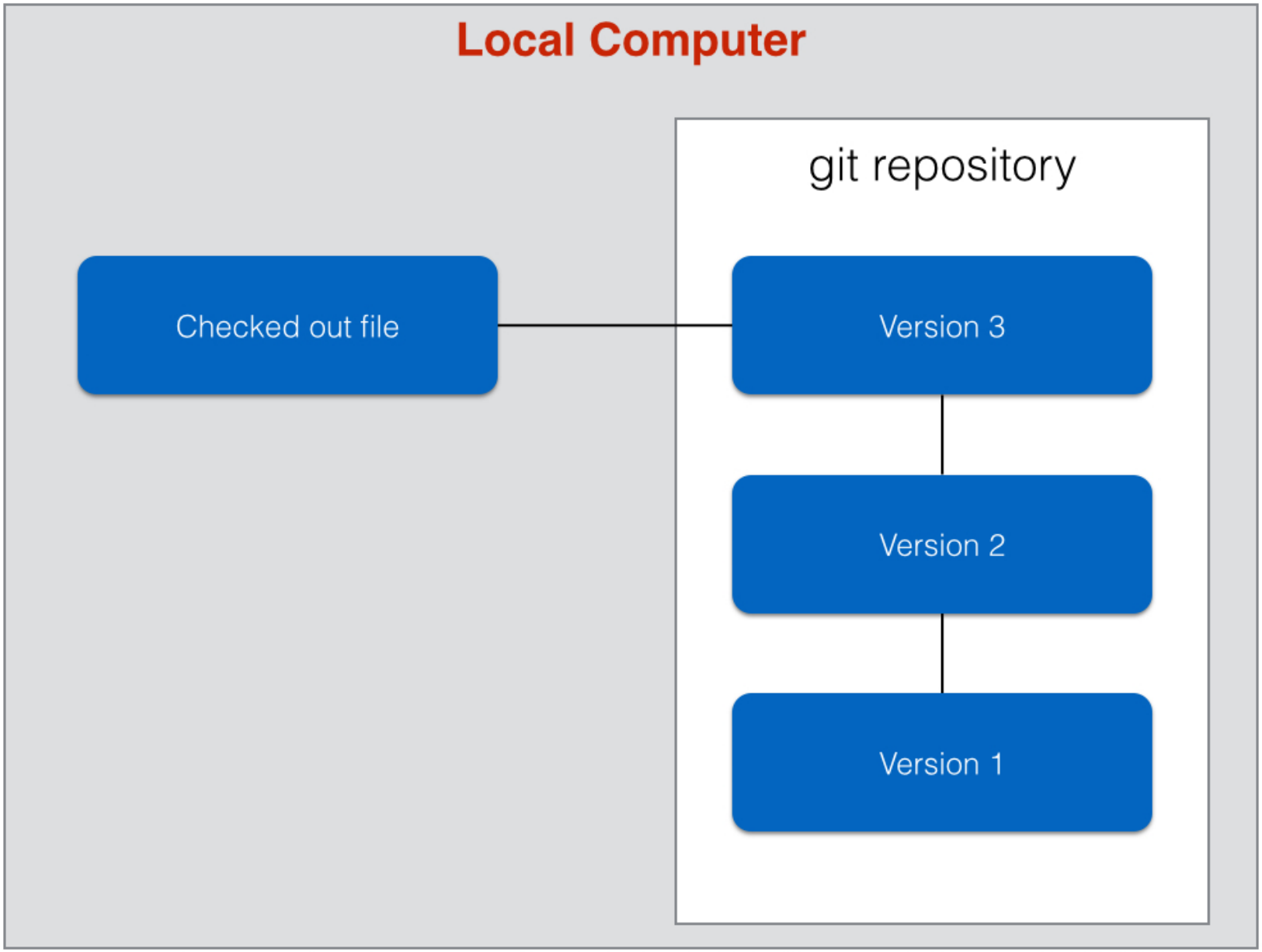

Image adapted from http://git-scm.com/book/en/Getting-Started-About-Version-Control, accessed 6 Feb 2014

Version control systems

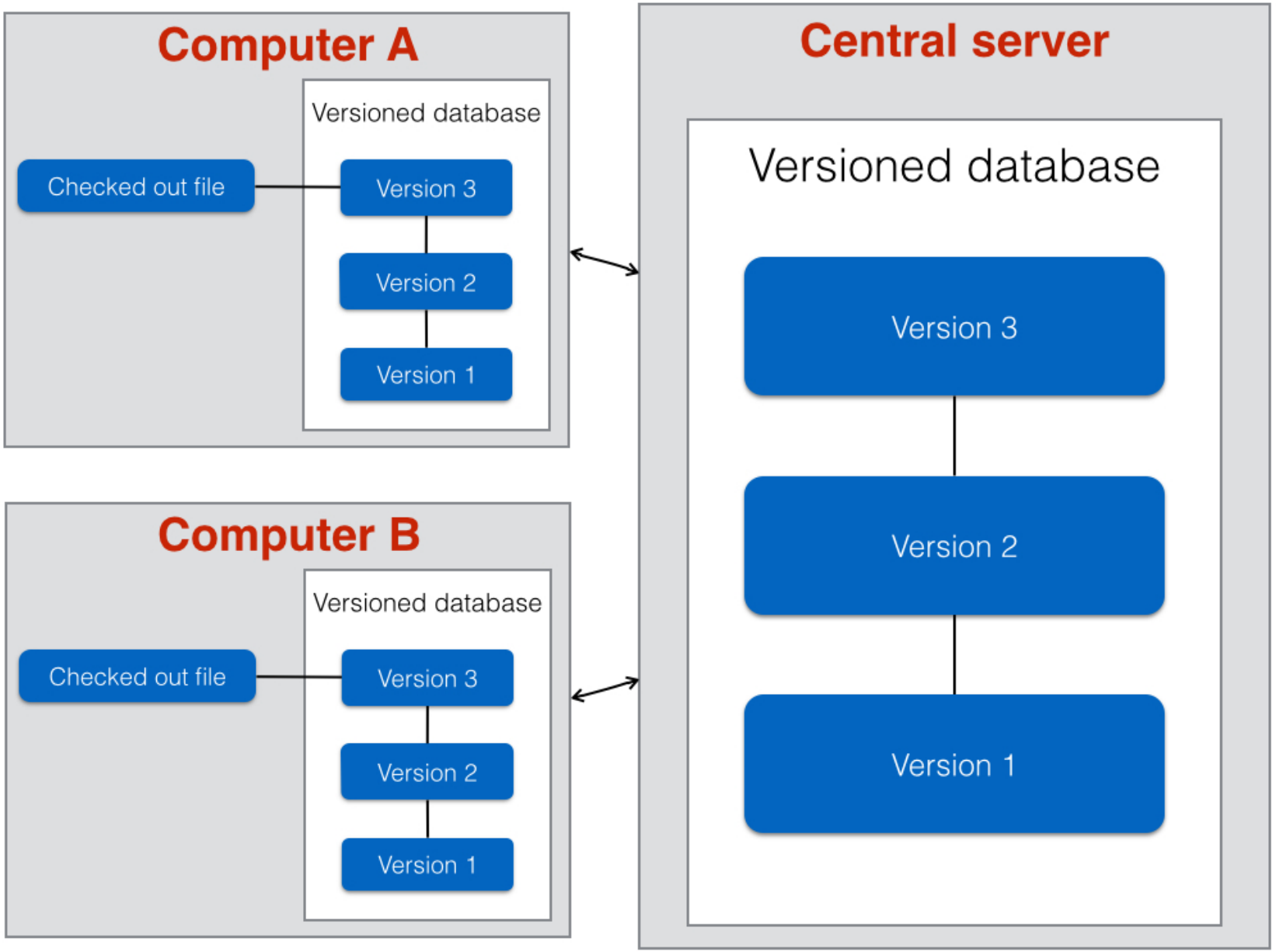

Image adapted from http://git-scm.com/book/en/Getting-Started-About-Version-Control, accessed 6 Feb 2014

Version control systems (git/GitHub flavored)

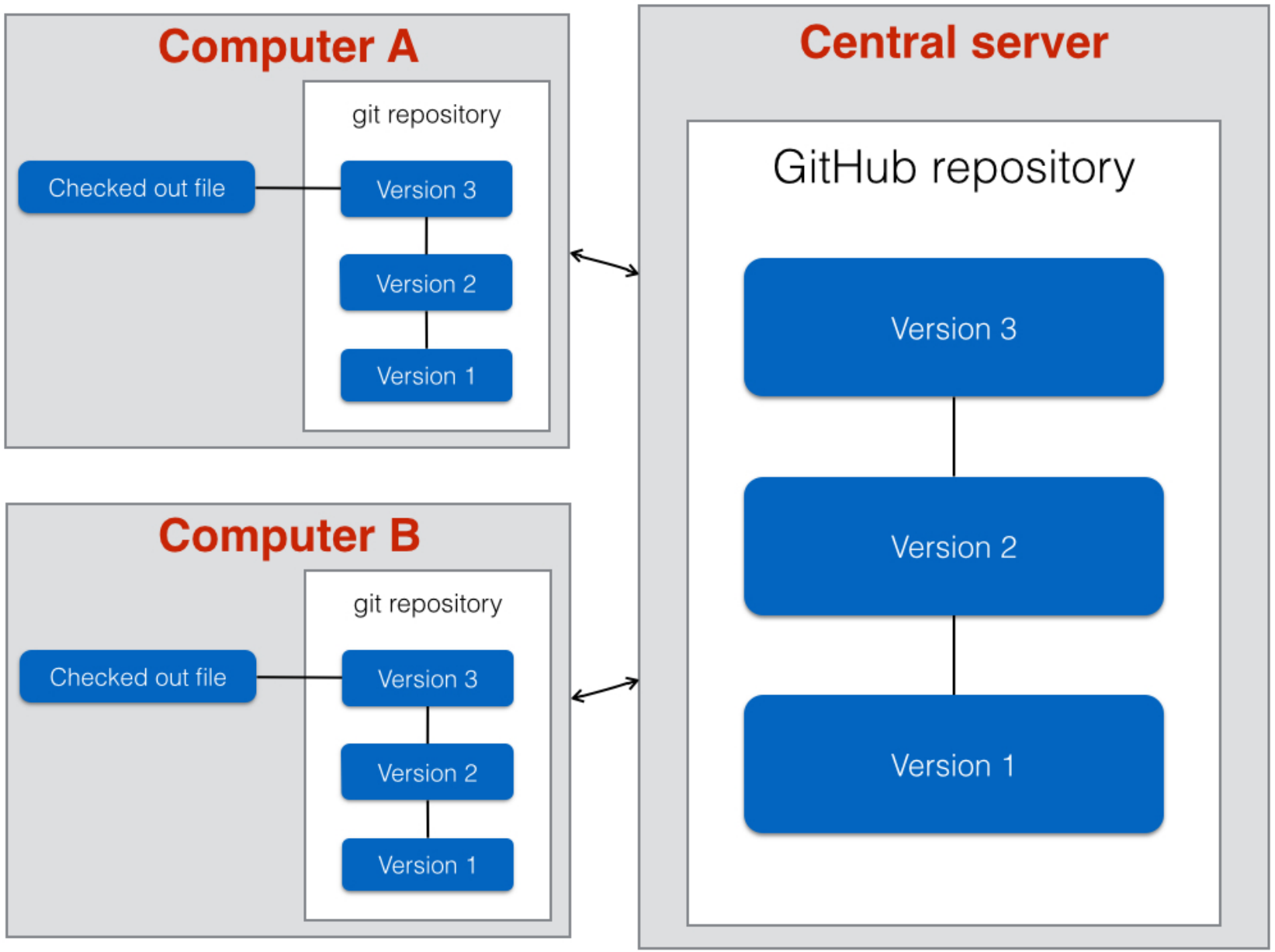

Image adapted from http://git-scm.com/book/en/Getting-Started-About-Version-Control, accessed 6 Feb 2014

Lots of (mostly free) options for cloud-based version controlling

Most services host multiple types of VCS

- sourceforge.net
- github.com
- bitbucket.org
- springloops.io
- ... [what have you used?]

git is a dialect

Key command-line operations

- `git init`: initializes a repository locally
- `git clone`: clones a repository from a remote source (i.e. GitHub.com)
- `git branch`: creates a new "branch" of code
- `git add`, `git rm`: manipulating files
- `git commit`: commits changes you have made

Using git with RStudio

Demo...

- clone the nickreich/statComp2014 repository from GitHub
- simple commit/push/pull examples