Session: Reproducible Research

Fabian Schroeder

September 12, 2017

Overview

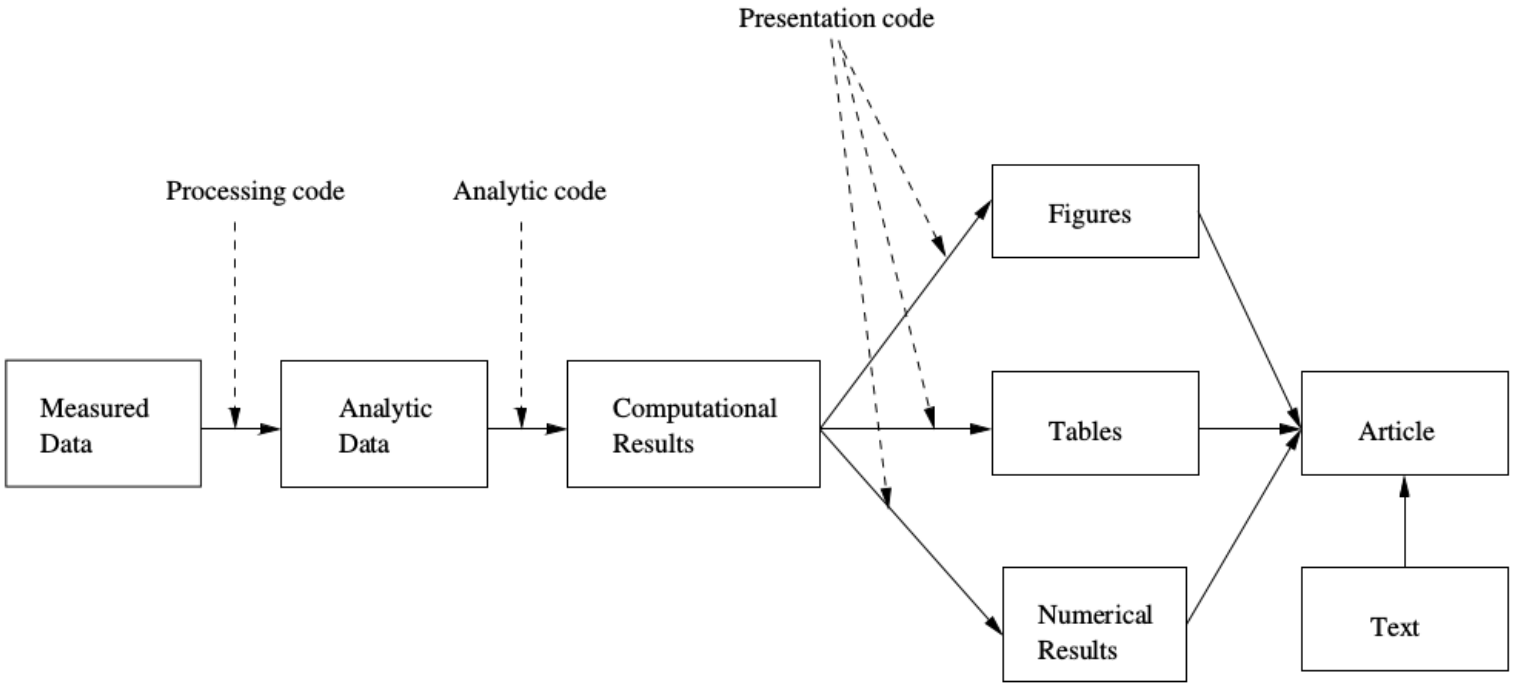
- Some general comments on Reproducible Research
- RMarkdown
- Git/Github

What Makes Research Reproducible?

According to Prof. Roger Peng, Department of Biostatistics and John Hopkins Bloomberg School of Public Health (Peng 2009):

- Analytic data are available
- Analytic code are available
- Documentation of code and data
- Standard means of distribution

Conceptual Look at Reproducible Research



How can RMarkdown aid in Reproducible Research?

- The purpose is "to create dynamic reports, which can be updated automatically if data or analysis change." (Friedrich Leisch)

How can RMarkdown aid in Reproducible Research?

- The purpose is "to create dynamic reports, which can be updated automatically if data or analysis change." (Friedrich Leisch)

- It can map the entire process of statistical research in one document!

What is RMarkdown?

It is an authoring framework for data science combining

- Knitr (Xie 2017) (a further development of Sweave (Leisch 2002)) and

What is RMarkdown?

It is an authoring framework for data science combining

- Knitr (Xie 2017) (a further development of Sweave (Leisch 2002)) and
- Pandoc Markdown (a slightly revised version of the markup language Markdown (by John Gruber) which can handle multiple output formats and has added new functionalities)

What is RMarkdown?

It is an authoring framework for data science combining

- Knitr (Xie 2017) (a further development of Sweave (Leisch 2002)) and
- Pandoc Markdown (a slightly revised version of the markup language Markdown (by John Gruber) which can handle multiple output formats and has added new functionalities)
- which has been well integrated into the RStudio IDE.

Where can I find information regarding RMarkdown?

[RMarkdown] (`http://rmarkdown.rstudio.com`)

What can I use RMarkdown for?

[RMarkdown Gallery]
(`http://rmarkdown.rstudio.com/gallery.html`)

Let us do an example!

What should I use RMarkdown for?

- Homeworks
- Reports
- Not: Dissertation
- Not: Highly Standardized Reports

Version Control I

Version Control II

Version control becomes a LOT MORE DIFFICULT if more people are involved!

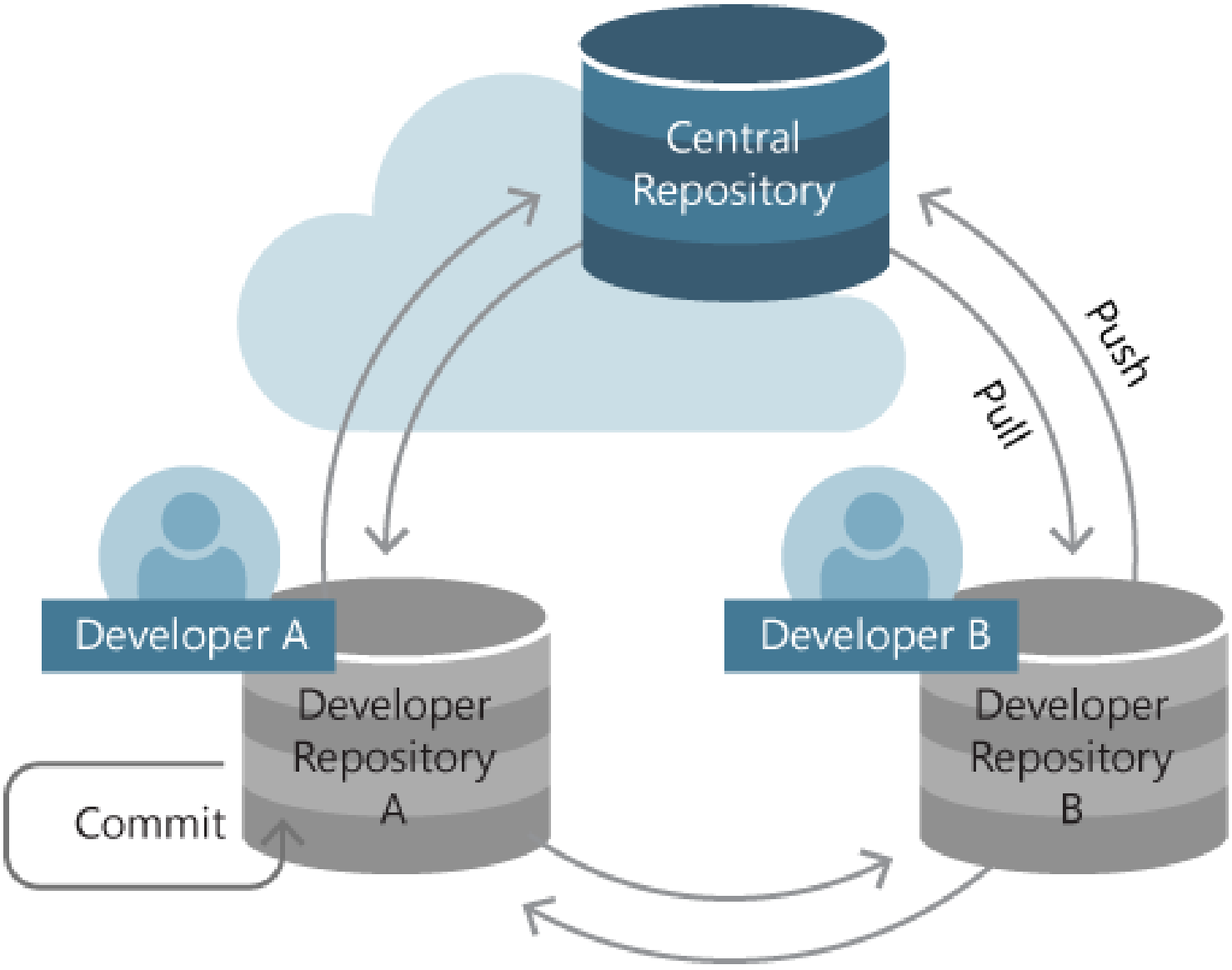
What is Git?

Git is a version control system, originally developed by Linus Torvalds. Its original purpose was to help groups of developers work collaboratively on big software projects. Git manages the evolution of a set of files – called a repository – in a sane, highly structured way.

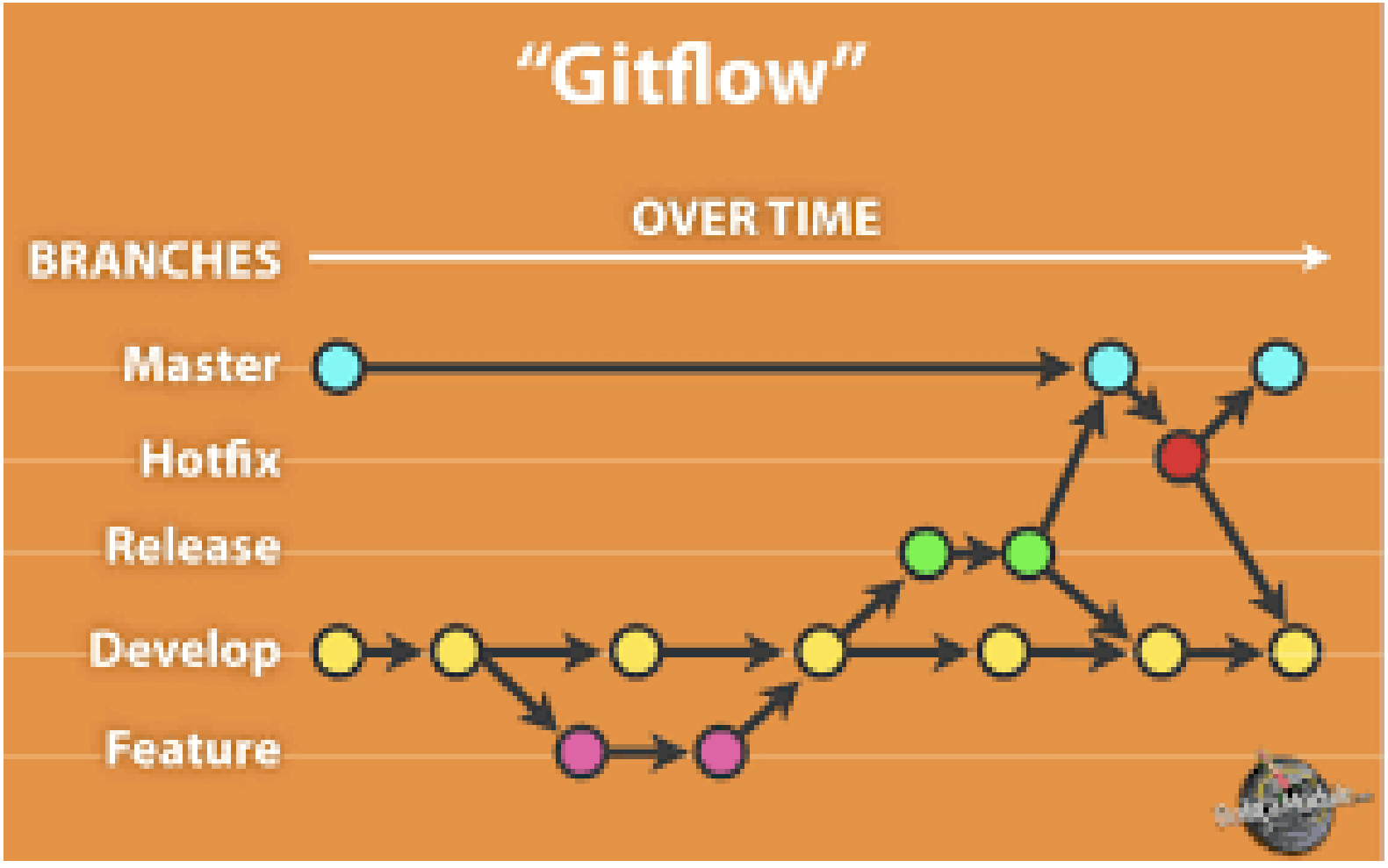
What is GitHub?

GitHub (like Bitbucket or Gitlab) is a hosting service for git-based projects. They add nice web-based interfaces to traditional Git servers.

How does it work?



A typical workflow



Good Resources

`http://happygitwithr.com/` (by Jenny Bryan)

`http://r-pkgs.had.co.nz/git.html` (by Hadley Wickham)

Git for the Data Scientist?

Let us do an example!

References

Leisch, Friedrich. 2002. "Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis." In *Compstat 2002 - Proceedings in Computational Statistics*, edited by Wolfgang Härdle and Bernd Rönz, 575–80. Physica Verlag, Heidelberg. `http://www.stat.uni-muenchen.de/~leisch/Sweave`.

Peng, Roger D. 2009. "Reproducible Research and Biostatistics." *Biostatistics* 10 (3): 405–8. doi:10.1093/biostatistics/kxp014.

Xie, Yihui. 2017. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. `https://yihui.name/knitr/`.