COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Lecture 2: R workflow, Git, and R markdown

Data Science for Business Analytics

Thibault Vatter <thibault.vatter@unil.ch>

Department of Statistics, Columbia University and HEC Lausanne, UNIL

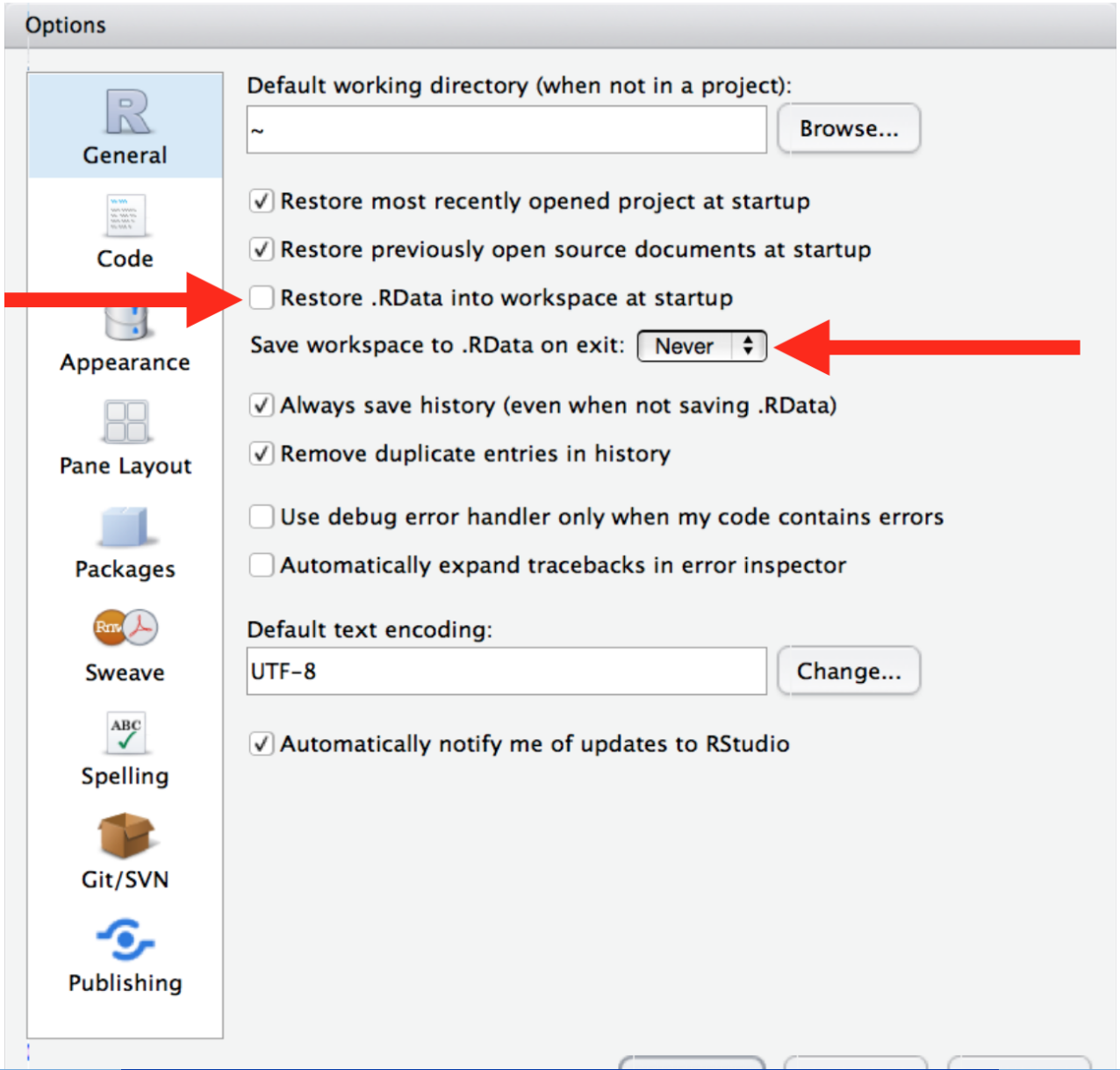
26.02.2018

Outline

Two questions

- What about your analysis is "real"?
- Where does your analysis "live"?

What about your analysis is "real"?



Where does your analysis live?

- The console
- R scripts
- **RStudio projects:** *make it straightforward to divide your work into multiple contexts, each with their own working directory, workspace, history, and source documents.*

DEMO!

The workflow with RStudio projects

- Create an RStudio project for each data analysis project.
- Keep data files there.
- Keep scripts there.
- Save your outputs (plots and cleaned data) there.
- Only ever use relative paths, not absolute paths.

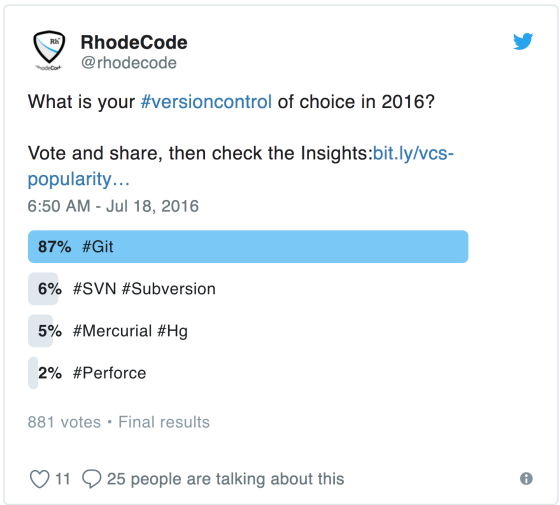
Everything you need is in one place, and cleanly separated from all the other projects that you are working on.

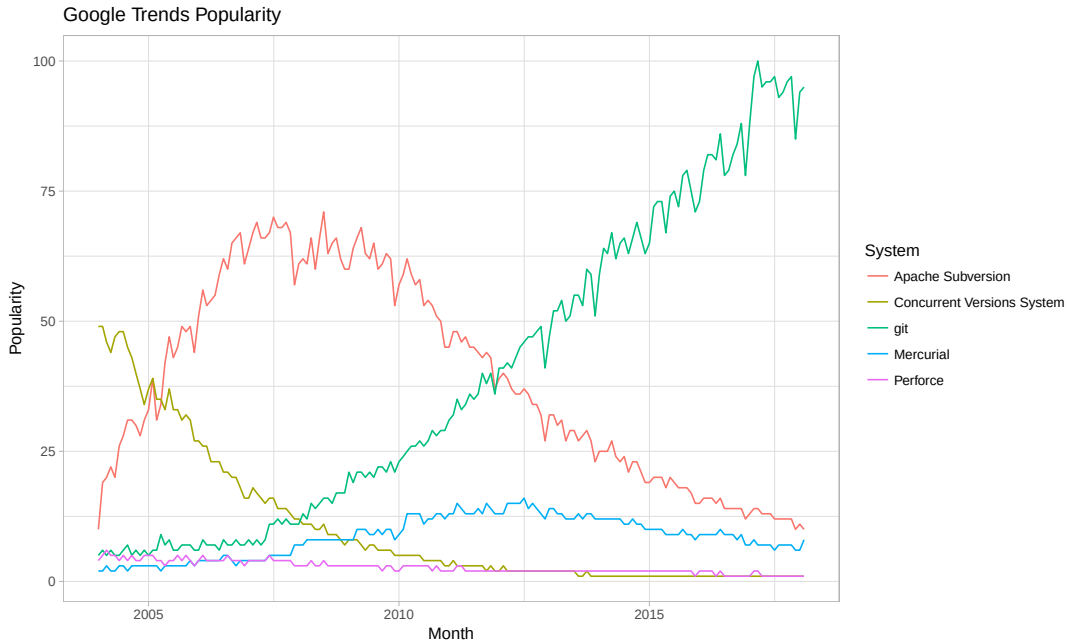
Outline

Version control

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

- **wikipedia:** *management of changes to documents, computer programs, large web sites, and other collections of information*

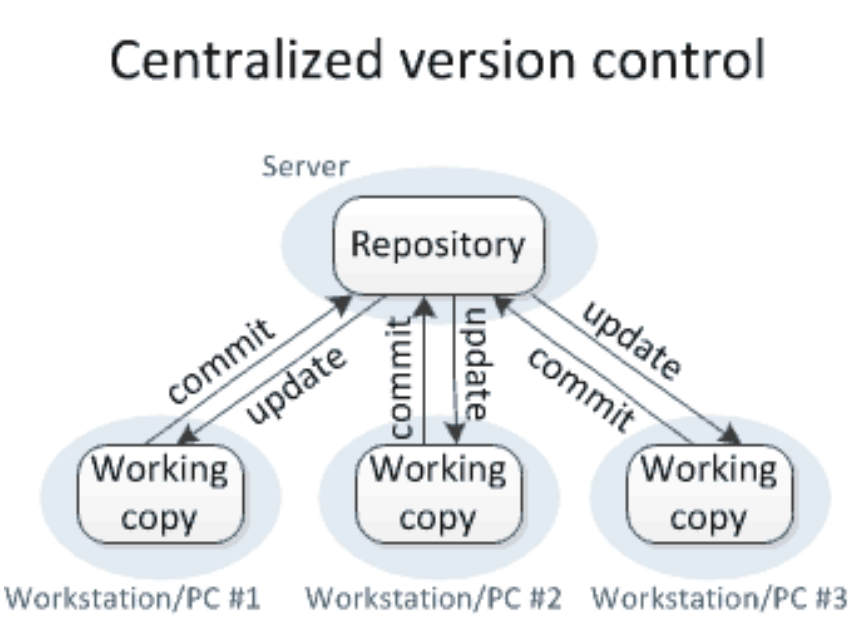


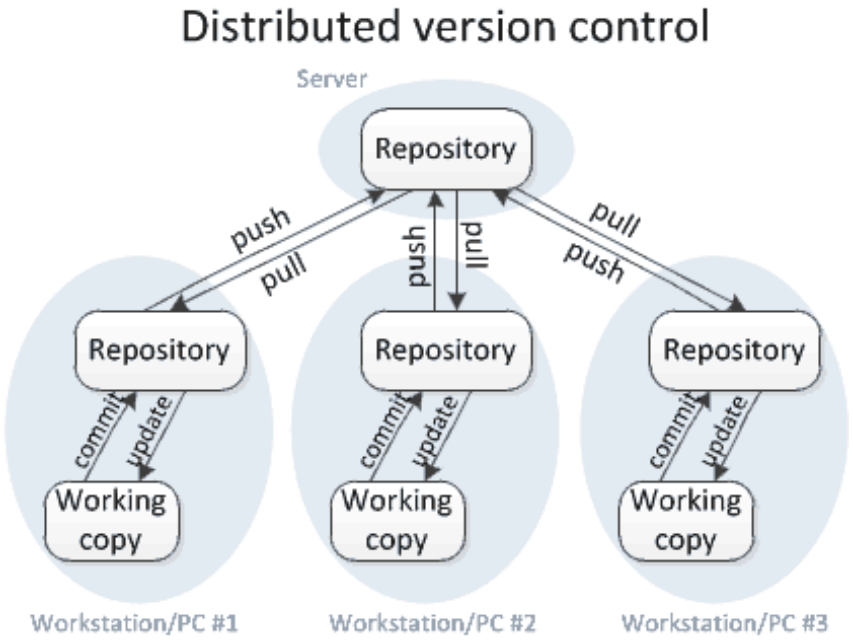


Git: “the stupid content tracker”

- Created by Linus Torvalds in 2005, his criteria:
  - Patching should take < 3 seconds
  - CVS as an ex. of what not to do (in doubt, do the opposite)
  - Distributed workflow
  - Strong safeguards against corruption (accidental or malicious)
- Maintained by Junio Hamano since 2005
- Part of the GNU free software project
- Source code written primarily in C, Shell, Perl, Tcl, Python
- Available for Windows, macOS, and Linux

Centralized vs distributed workflows





Version control concepts and best practices by Michael Ernst

Github

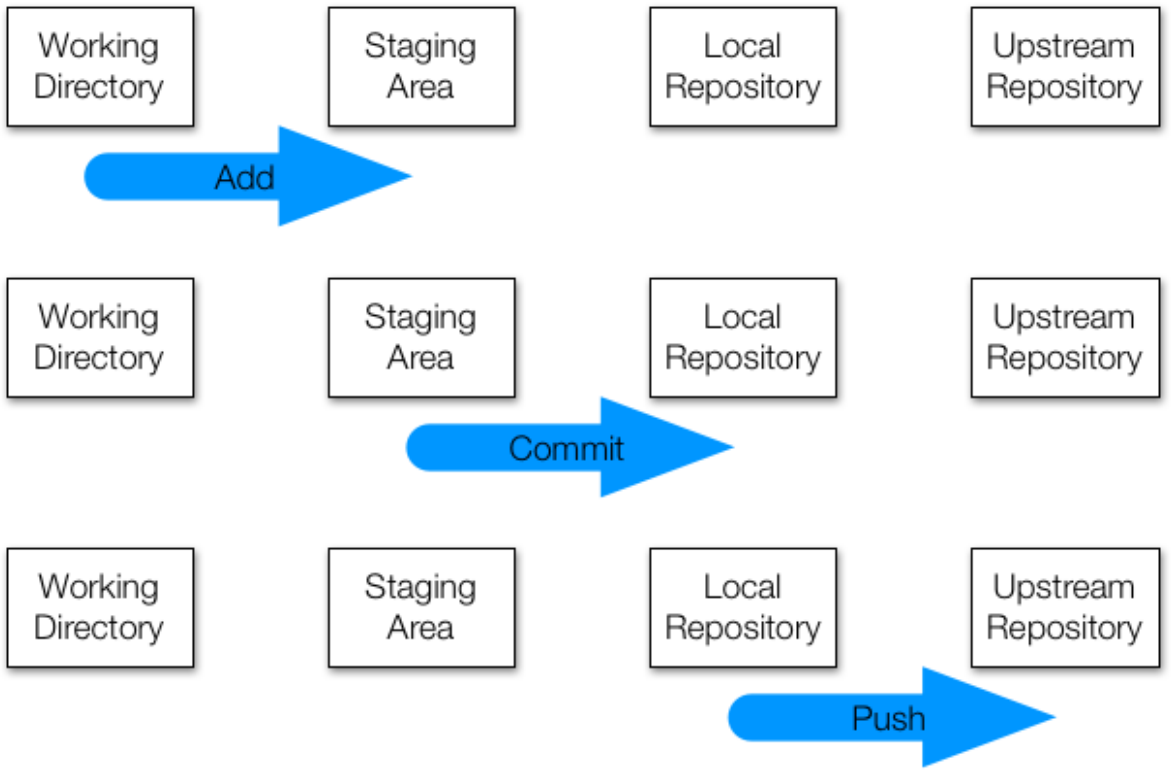
- Web-based version control service using git
- Bug tracking, feature requests, task management, and wikis for every project
- 20 million users and 57 million repositories (April 2017)
- Private and public repos
- GitHub Student Developer Pack
- Create account & email user to Natasha by **March 19**.

RStudio projects with version control

DEMO!

Workflow for your assignments

- Work on your assignment
- Commit changes to your local repository
- Push the changes to the github repo



source: github.com/datasciencelabs

Best practices

- Use a descriptive commit message
- Make each commit a logical unit
- Avoid indiscriminate commits
- Incorporate others' changes frequently
- Share your changes frequently
- Coordinate with your co-workers
- Remember that the tools are line-based
- Don't commit generated files
- Understand your merge tool

See "Version control concepts and best practices" by Michael Ernst

Final comments

- Last commit before midnight of due date as final submission
  - If there are commits after midnight, we will take the last commit up to the due date at 11:59 pm as the final version
- Check that the final commit is showing in your Github repo
  - "I forgot to push" is not an acceptable excuse
- Detailed tutorials (with lots of pictures):
  - Setting-up Github
  - Git and RStudio
  - Github and RStudio (alternative)
  - Github and RStudio (alternative 2)

Outline

R Markdown

The two components:

- Literate programming
- Markdown

Literate programming

- Motivation: helps peers understand and replicate your results, find errors and suggest enhancements
- Introduced by Donald Knuth
- **wikipedia:** *a program is given as an explanation of the program logic in a natural language, such as English, interspersed with snippets of macros and traditional source code, from which a compilable source code can be generated [... It] represents a move away from writing programs in the manner and order imposed by the computer, and instead enables programmers to develop programs in the order demanded by the logic and flow of their thoughts*

What does this R code do?


```r
data(women)
plot(women)
fit <- lm(weight ~ height, data = women)
abline(fit)
```

And this one?


```r
# Analysis of the 'women' dataset in R
data(women) # Load the data
attach(women) # Attach data to path
plot(weight ~ height) # Make a scatter plot
fit <- lm(weight ~ height) # Fit linear model
abline(fit) # Add a line of best fit to the plot
```

Two competing "views"

Real programmers don't comment their code. If it was hard to write, it should be hard to understand. – unknown

OR

If you can't write clearly, you probably don't think nearly as well as you think you do. – Kurt Vonnegut

Can't we do better?

The *World Almanac and Book of Facts* (1975) includes a dataset of heights (in) and weights (lbs) of 15 American women aged 30–39. It is built into R:

```
data(women)
```

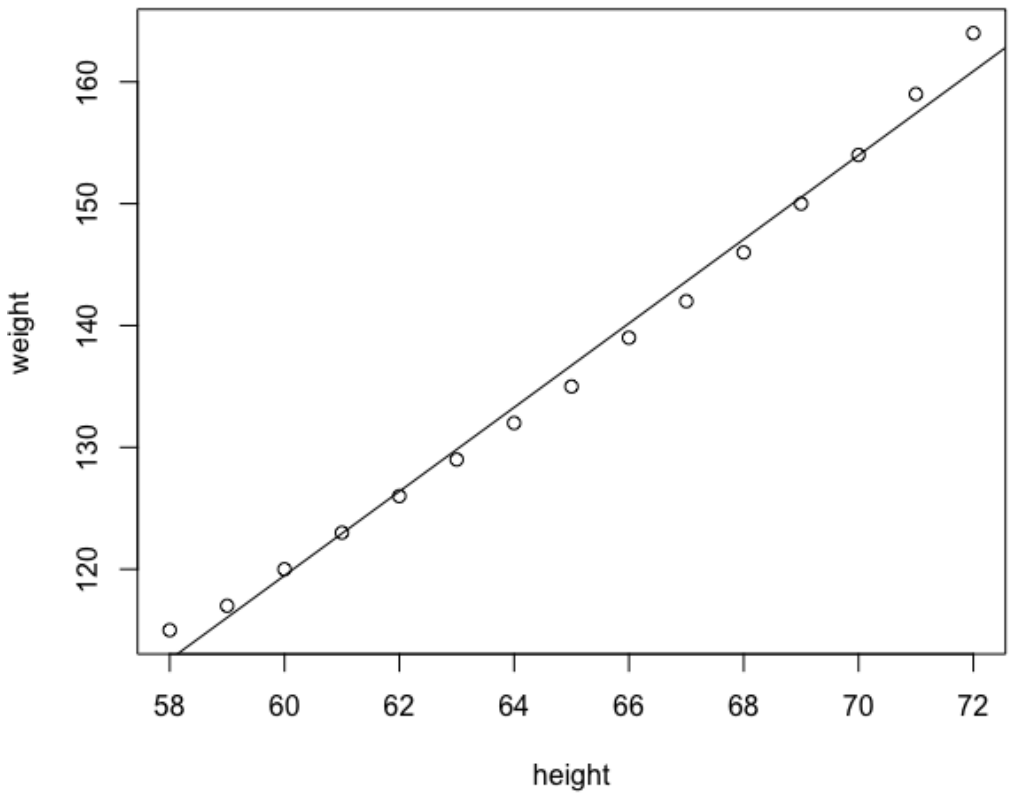
As height increases, weight appears to increases (almost) linearly: every inch in height adds approximately 3.45 lbs. This was determined by fitting a simple linear regression model of weight against height:

```
fit ← lm(weight ~ height, data = women)
```

Can't we do better? cont'd

The resulting least-squares regression line can be drawn on a scatter plot of height against weight, where the models seems appropriate:

```
1 plot(weight ~ height, data = women)
2 abline(fit)
```



A lightweight markup language

- **Markup:**
  - A system for annotating a document in a way that is syntactically distinguishable from the text
  - E.g., LaTeX and HyperText Markup Language (HTML)
- **Lightweight:**
  - A markup language with simple, unobtrusive syntax
  - E.g., Markdown and R markdown

Markup vs lightweight markup

Here is some text:

- in *italics*,
- in **boldface**.

In LaTeX:

```
Here is some text:
\begin{itemize}
\item in \textit{italics},
\item in \textbf{boldface}.
\end{itemize}
```

In Markdown:

```
Here is some text:
* in *italics*,
* in **boldface**.
```

What is R markdown?

A markdown-based literate programming system

DEMO!

Useful resources

- Essential: R Markdown cheat sheet
- RStudio's R markdown website
  - Tutorial (to get you started)
  - Output formats (e.g., HTML, Word documents, PDFs, presentations, etc.)
- stuff written by Yihui
  - knitr and especially its the options page
  - bookdown to write technical reports
  - blogdown to even build your own website