# Raising better scientists

by teaching reproducibility
as part of
teaching data analysis

Mine Çetinkaya-Rundel
Department of Statistical Science
Duke University
Duke Tech Expo 2015

**#1**:
Convince
researchers to adopt
a reproducible
research workflow

**#2**:
Train new
researchers who
don't have any other
workflow

**reproducibility**
often comes up in the context of published research and the need to accompany such research with the complete data and analyses, including software/code

**statistics** educators who teach data analysis should be instilling best practices in students before they set out to do research
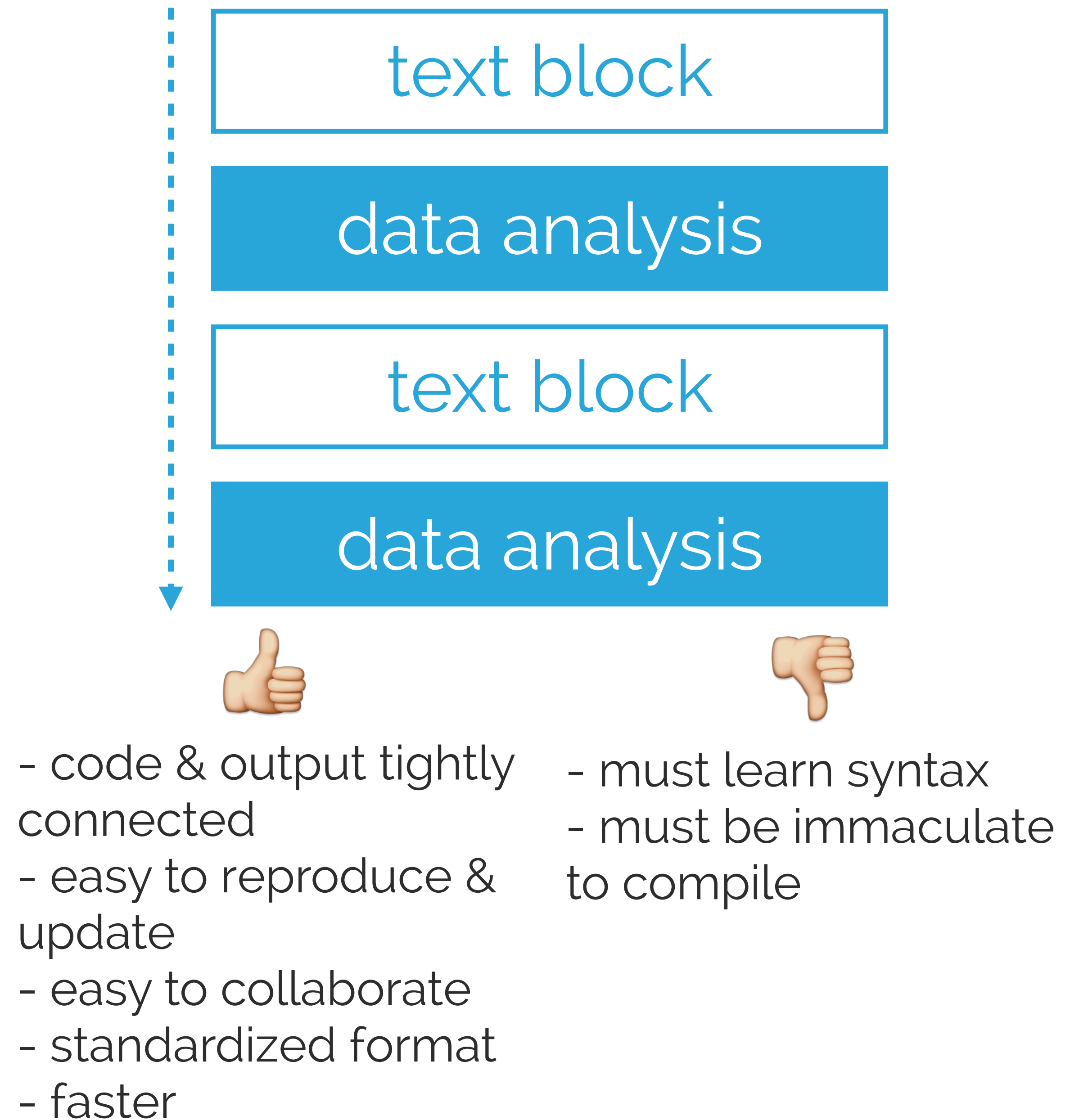
# *traditional*

**data analysis**
- descriptive stats
- plots & tables
- model output

**write-up**
- research question & context
- interpretations
- conclusions

copy-paste

copy-paste

**lab report**

# a better approach

text block

data analysis

text block

data analysis

👍
- code & output tightly connected
- easy to reproduce & update
- easy to collaborate
- standardized format
- faster

👎
- must learn syntax
- must be immaculate to compile

# Duke UNIVERSITY | VM Manage
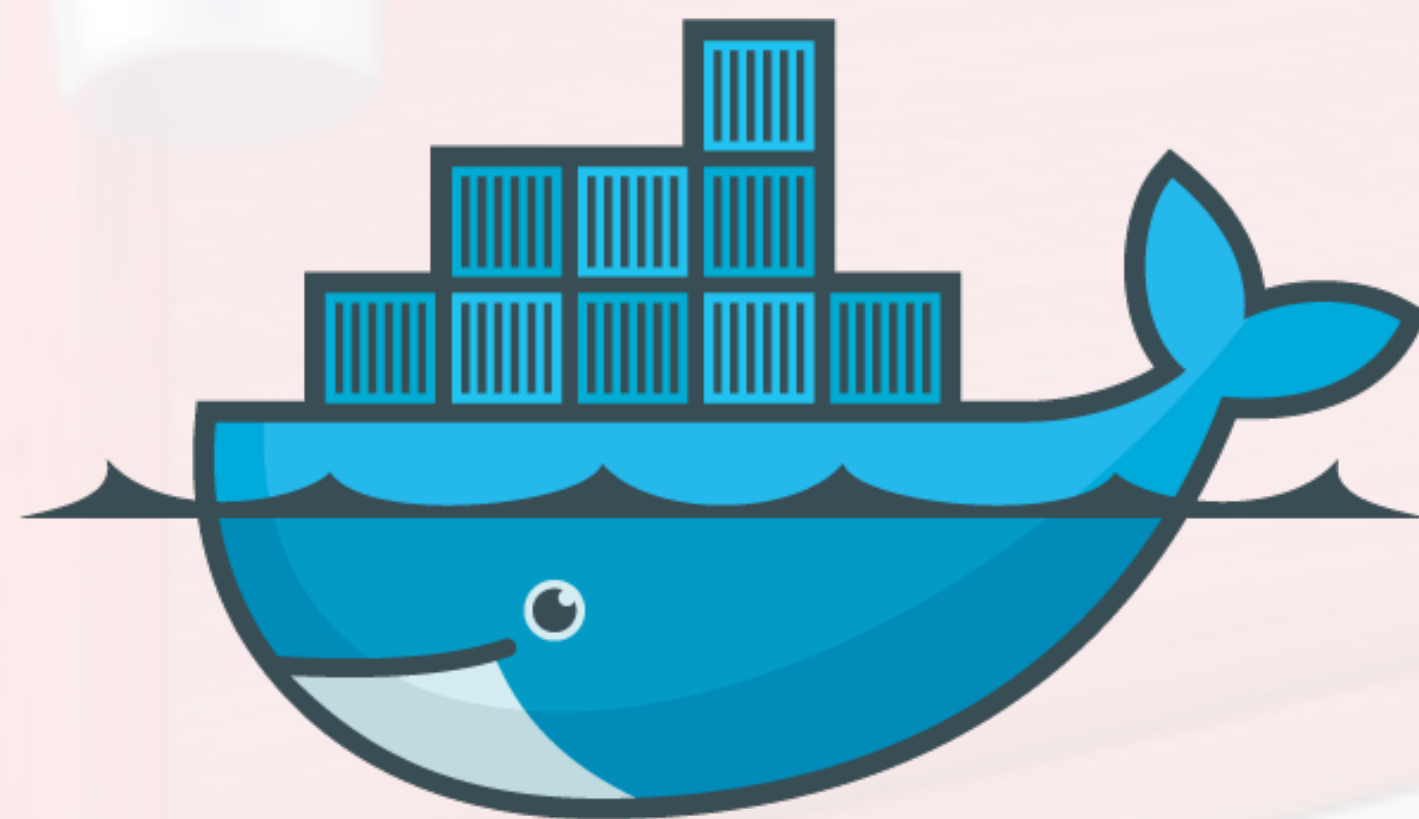
Home  New  **Docker**  Manage  Register Alias  Logout

## Docker Containers

Welcome back Mine Cetinkaya-Rundel!

### RStudio - statistics application with Rmarkdown and knitr support

**Click here to log in** to your RStudio environment

### MATLAB - mathematics and matrix manipulations

**Click here to log in** to your MATLAB environment

### Mathematica - computational software and symbolic mathematics

**Click here** to create your personal Mathematica environment

### Eclipse - integrated development environment with Canopy python

**Click here** to create your personal Eclipse environment

---

Duke UNIVERSITY

© 2012-2014 Duke University, Durham, NC 27708 |
(919) 684-8111

Accessibility    |    Style Guide    |    About Duke.edu    |    Contact Us

R

File   Edit   Code   View   Plots   Session   Build   Debug   Tools   Help

guest | Sign Out

Go to file/function

Project: (None)

---

sta101_lab1.Rmd ×

? ▾   Knit HTML ▾        Run   Chunks ▾

```
 1  ---
 2  title: "Lab 1"
 3  author: "[Enter team name here]"
 4  date: "[Enter date here]"
 5  output:
 6    html_document:
 7      fig_height: 4
 8      fig_width: 6
 9  ---
10
11  **Author:** [Enter author name here]
12
13  **Discussants:** [Enter discussant names here. Remember, these are other team members
    present in lab.]
14
15  * * *
```

20:1   Chunk 1 ⬍                                        R Markdown ⬍

---

Environment   History

Import Dataset ▾   Clear                              List ▾

Global Environment ▾

Data

arbuthnot                    82 obs. of 3 variables

---

Console ~/

```
R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> source("http://www.openintro.org/stat/data/arbuthnot.R")
```

---

Files   Plots   Packages   Help   Viewer

New Folder   Upload   Delete   Rename   More ▾

Home

| | Name | Size | Modified |
|---|---|---|---|
| | .Rhistory | 15.6 KB | Apr 16, 2015, 10:14 PM |
| | acs.RData | 14.3 KB | Jan 18, 2015, 11:11 PM |
| | cig07.RData | 7.1 KB | Jan 18, 2015, 11:12 PM |
| | gss.Rdata | 2.4 MB | Feb 23, 2015, 10:32 AM |
| | inference.RData | 176.1 KB | Feb 23, 2015, 9:47 PM |
| | R | | |
| | rxntime.csv | 84.2 KB | Mar 4, 2015, 1:37 AM |
| | sta101_lab1.Rmd | 883 B | Jan 12, 2015, 12:48 AM |
| | u5deck1.R | 426 B | Mar 15, 2015, 11:49 PM |
| | u7deck3.R | 839 B | Apr 13, 2015, 12:40 PM |
| | vision.csv | 222 B | Jan 18, 2015, 11:12 PM |

**Load Arbuthnot's data:**

```r
source("http://www.openintro.org/stat/data/arbuthnot.R")
```

## Exercise 1:

```r
# code for Exercise 1 is already entered below as an example
arbuthnot$girls
```

```
##  [1] 4683 4457 4102 4590 4839 4820 4928 4605 4457 4952 4784 5332 5200 4910
## [15] 4617 3997 3919 3395 3536 3181 2746 2722 2840 2908 2959 3179 3349 3382
## [29] 3289 3013 2781 3247 4107 4803 4881 5681 4858 4319 5322 5560 5829 5719
## [43] 6061 6120 5822 5738 5717 5847 6203 6033 6041 6299 6533 6744 7158 7127
## [57] 7246 7119 7214 7101 7167 7302 7392 7316 7483 6647 6713 7229 7767 7626
## [71] 7452 7061 7514 7656 7683 5738 7779 7417 7687 7623 7380 7288
```

```r
plot(arbuthnot$girls)
```



---

**RStudio editor pane (sta101_lab1.Rmd):**

```
1  ---
2  title: "Lab 1"
3  author: "[Enter team name here]"
4  date: "[Enter date here]"
5  output:
6    html_document:
7      fig_height: 4
8      fig_width: 6
9  ---
10
11 **Author:** [Enter author name her
12
13 **Discussants:** [Enter discussan
   present in lab.]
```

1:1    (Top Level)

---

**Console / R Markdown:**

```
R version 3.1.2 (2014-10-31) -- "Pumpkin
Copyright (C) 2014 The R Foundation for
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLU
You are welcome to redistribute it under
Type 'license()' or 'licence()' for dist

R is a collaborative project with many c
Type 'contributors()' for more informati
'citation()' on how to cite R or R packa

Type 'demo()' for some demos, 'help()' f
'help.start()' for an HTML browser inter
Type 'q()' to quit R.
```

| Size | Modified |
|---|---|
| 156.9 KB | Apr 16, 2015, 11:26 PM |
| 14.3 KB | Jan 18, 2015, 11:11 PM |
| 7.1 KB | Jan 18, 2015, 11:12 PM |
| 2.4 MB | Feb 23, 2015, 10:32 AM |
| 176.1 KB | Feb 23, 2015, 9:47 PM |
| 43.2 KB | Apr 1, 2015, 12:26 PM |
| 84.2 KB | Mar 4, 2015, 1:37 AM |
| 883 B | Apr 16, 2015, 11:26 PM |
| 426 B | Mar 15, 2015, 11:49 PM |
| 839 B | Apr 13, 2015, 12:40 PM |
| 222 B | Jan 18, 2015, 11:12 PM |

**reproducible science … and better teaching**

instills early the idea that all analysis must be done in a reproducible framework

eases collaboration on labs and projects

removes ambiguity from grading