

Precursors to the Data Explosion: Teaching How to Compute with Data

Nicholas J. Horton

Amherst College, Amherst, MA, USA

October 7, 2013

nhorton@amherst.edu

Acknowledgements

- joint work with Ben Baumer (Smith College), Linda Loi (Smith College), Mine Çetinkaya-Rundel (Duke University) and Andrew Bray (University of Massachusetts/Amherst)
- supported by NSF grant 0920350 (building a community around modeling, statistics, computation and calculus)
- more information at <http://www.mosaic-web.org>

Prelude

Nolan and Temple Lang (TAS, 2010)

The nature of statistics is changing significantly with many opportunities to broaden the discipline and its impact on science and policy. To realize this potential, our curricula and educational culture must change. . . . Computational literacy and programming are as fundamental to statistical practice and research as mathematics. . . . We advocate that our field needs to define statistical computing more broadly to include advancements in modern computing, beyond traditional numerical algorithms.

Nolan and Temple Lang framework

- 1 broaden statistical computing
- 2 deepen computational reasoning and literacy
- 3 compute with data in the practice of statistics

Prelude

- “New Frontier in statistical thinking” (Chamandy, Google)
- teaching precursors to big data/foundations of data science as an intellectually coherent theme (Pruim, Calvin College)
- growing importance of computing and ability to “think with data” (Lambert, Google)
- acknowledgments of the challenges of developing “data habits of mind” (Finzer, TISE 2013)

How to accomplish this?

- start in the first course
- build on this in the second course
- develop more opportunities for students to apply their knowledge in practice (internships, collaborative research, teaching assistants)
- new courses, capstones and co-curricular opportunities focused on “Data Science”

Plan for talk

- background on reproducible analysis
- introduction to R Markdown
- how to make it work in the introductory course
- outcomes and assessment
- what next?
- conclusions

Background on reproducible analysis

The purpose of Sweave is to create dynamic reports, which can be updated automatically if data or analysis change. Instead of inserting a prefabricated graph or table into the report, the master document contains the code necessary to obtain it. When run through a statistics package, all data analysis output (tables, graphs, . . .) is created on the fly and inserted into a final document. The report can be automatically updated if (when!) data or analysis change, which allows for truly reproducible research. (Leisch, R-News 2002)

Reproducible analysis

- integrate code and documentation
- automate documentation and report generation
- helps to structure analysis for new users
- (particularly) useful to students to facilitate appropriate and correct statistical workflow
- helps minimize the pain of iterative analyses
- leaves behind a clear online trail

Background

Rooted in *Literate Programming*

Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to humans what we want the computer to do. (Donald E. Knuth, 1984).

Background

Rooted in *Literate Programming*

Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to humans what we want the computer to do. (Donald E. Knuth, 1984).

Key idea: simplify the cognitive load by providing a structure and foundation for statistical analysis

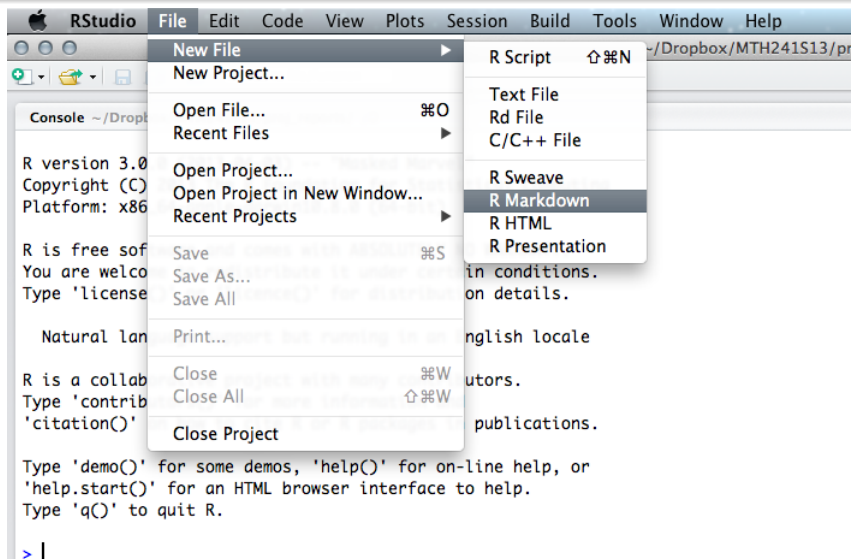
Reproducible statistical analysis (cont.)

- previous implementations (Sweave/knitr) required knowledge of markup language (such as \LaTeX)
- examples: case studies in R from the *Statistical Sleuth* at <http://www.amherst.edu/~nhorton/sleuth>
- disadvantage: extra degree of difficulty for new students
- much simpler interface using R Markdown

R Markdown

- allows easy authoring of reproducible web reports from R
- features a simple text format with embedded R chunks
- chunks similar to Sweave/knitr
- chunks are executed, saving output and graphics
- files are “woven” into plain markdown files
- these markdown files are compiled into HTML documents
- images are bundled into the HTML, for portability
- simplified interface within RStudio
- disadvantage: less control of output formatting

Creating a Markdown file in RStudio



Example

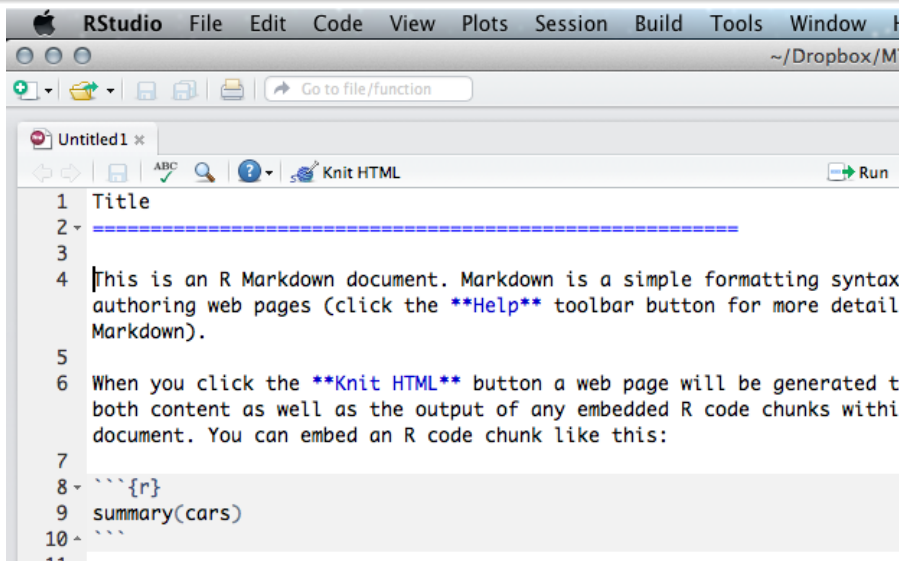
Title

This is an R Markdown document. Markdown is a simple formatting syntax for authoring web pages

When you click the **Knit HTML** button a web page will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
```${r}  
summary(cars)
```
```

Editing a Markdown file



The screenshot shows the RStudio interface with a menu bar (RStudio, File, Edit, Code, View, Plots, Session, Build, Tools, Window) and a toolbar. The main editor window displays a Markdown document titled "Untitled1". The document content is as follows:

```
1 Title
2 =====
3
4 [This is an R Markdown document. Markdown is a simple formatting syntax
5 authoring web pages (click the Help toolbar button for more detail
6 Markdown).
7
8 When you click the Knit HTML button a web page will be generated t
9 both content as well as the output of any embedded R code chunks withi
10 document. You can embed an R code chunk like this:
```


Displaying the formatted results

Title

This is an R Markdown document. Markdown is a simple formatting syntax for authoring web pages (click the **Help** toolbar button for more details on using R Markdown).

When you click the **Knit HTML** button a web page will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed      dist
## Min.   : 4.0    Min.    : 2
## 1st Qu.:12.0    1st Qu.: 26
## Median :15.0    Median : 36
## Mean   :15.4    Mean    : 43
## 3rd Qu.:19.0    3rd Qu.: 56
## Max.   :25.0    Max.    :120
```

You can also embed plots, for example:

```
plot(cars)
```

Additional features

The screenshot displays the RStudio interface with two panes. The left pane, titled 'example.Rmd', shows the source Markdown code with line numbers 1 through 20. The right pane, titled 'RStudio: Preview HTML', shows the rendered HTML output.

Source Code (Left Pane):

```
1 Header 1
2 -----
3 This is an R Markdown document. Markdown is a
4 | simple formatting syntax for authoring web pages.
5 Use an asterisk mark, to provide emphasis such as
6 | italics and bold.
7 Create lists with a dash:
8 - Item 1
9 - Item 2
10 - Item 3
11
12 You can write `in-line` code with a back-tick.
13
14 ```
15 Code blocks display
16 with fixed-width font
17 ```
18
19 > Blockquotes are offset
20
```

Rendered Output (Right Pane):

Preview: ~/example.html

Header 1

This is an R Markdown document. Markdown is a simple formatting syntax for authoring web pages.

Use an asterisk mark, to provide emphasis such as *italics* and **bold**.

Create lists with a dash:

- Item 1
- Item 2
- Item 3

You can write `in-line` code with a back-tick.

```
Code blocks display
with fixed-width font
```

Blockquotes are offset

Motivation

Some may (reasonably) argue:

The introductory statistics course syllabus is already jam-packed. Why include R Markdown in the Intro Course?

Increased importance of computing

Information technologies are increasingly important and should be added to the (statistics) curriculum, as should the ability to reason about computational resources, work with large datasets, and perform computationally intensive tasks (Nolan and Temple Lang, TAS 2010).

ASA Curricular guidelines (circa 2000)

Computational - Working with data requires more than basic computing skills. Programs should require familiarity with a standard statistical software package and encourage study of data management and algorithmic problem-solving.

- Is this sufficient background for students heading out in a world awash with data?
- If not, then when should we start teaching these capacities?

Alternatives?

- separate statistics package and layout package
- cut and paste in between
- huge potential for error
- discourages iteration
- separates computation from analysis

Implementation

- early and often
- start simple
- homework
- projects

Prezi introduction for students

Start early: first homework assignment

- template for homework solution provided to students
- provided commands to load data and create new variables, as needed
- early in the course, included most commands needed to run analyses, as a way of helping students to get up to speed
- main task: annotate and interpret the output
- used for all homeworks in the course

Why projects? (carpentry metaphor due to McGowan)

In week 1 of the carpentry (statistics) course, we learned to use various kinds of planes (summary statistics). In week 2, we learned to use different kinds of saws (graphs). Then, we learned about using hammers (confidence intervals). Later, we learned about the characteristics of different types of wood (tests). By the end of the course, we had covered many aspects of carpentry (statistics). But I wanted to learn how to build a table (collect and analyze data to answer a question) and I never learned how to do that.

Projects in Intro and Intermediate Stat

- projects an effective way to implement many of the GAISE recommendations (Halvorsen, ICOTS8 2010)
 - use real data
 - stress conceptual understanding rather than mere knowledge of procedures
 - use technology for ... analysing data
- generally done in groups of 3-4
- need for audit trail is key given iteration around coding and analyses
- many deliverables throughout the semester
- culminates in a final report

Computing challenges for projects

- data cleaning and variable derivation
- summary statistics and univariate graphical display
- multiple regression model
- assessment of model using variety of diagnostics
- iterative process: many false starts
- students often get muddled as datasets change, coding is modified, and analyses proceed

Grading and feedback

- hardcopy (old school)
- separate feedback (uploaded to Moodle or other CMS)
- integrated comments (requires some knowledge of html)

Grading and feedback

Exercise 8:

The most total number of births in the U.S. was in 1961.

```
which.max((present$boys + present$girls))
```

```
## [1] 22
```

3/3. Think about how to identify the maximum year by referencing the index that

Grading and feedback

```
<pre><code class="r">which.max((present$boys + present$girls))  
</code></pre>
```

```
<pre><code>## [1] 22  
</code></pre>
```

<h3>3/3. Think about how to identify the maximum year by refer

</body>

</html>

Results from Duke University

- n=221 students used R Markdown during the 2012–2013 academic year
- introductory statistics course
- students generally had no computational background
- completed team based project
-

Results from Smith College

| Course | Fall 2012 | Spring 2013 | Fall 2013 |
|---------------------|-----------|-------------|--------------|
| Intro Stats | 42 | 70 | ≈ 40 |
| Regression Analysis | 33 | | |
| Data Science | | | ≈ 25 |
| Total | 75 | 70 | ≈ 65 |

Table : R Markdown exposure at Smith College. 11 of 20 pre-registered students for Data Science have already used R Markdown.

Feedback

- $n=70$ students in spring Intro Stats completed surveys at the beginning and end of the semester ($n=56$ with data at both times)
- mid-semester assessment (qualitative) and end of semester evaluations
- summary: students got up to speed with R Markdown, but would have welcomed more introduction and structure for first exposure

Sample survey items

R Markdown

Please indicate the response that most closely matches your attitude towards each of the following statements.

1. I find the R Markdown syntax to be simple and understandable.

no opinion strongly disagree disagree indifferent agree strongly agree

2. When my Markdown document does not compile, I know how to go about fixing it.

no opinion strongly disagree disagree indifferent agree strongly agree

3. I am frequently frustrated by R Markdown when doing my homework.

no opinion strongly disagree disagree indifferent agree strongly agree

Survey results

| Question | Initial | | Final | | Change | |
|----------|---------|-------|--------|-------|--------|-------|
| | Mean | SD | Mean | SD | Mean | SD |
| Q1 | -0.300 | 0.931 | 0.241 | 1.053 | 0.527 | 1.124 |
| Q2 | -0.527 | 1.069 | -0.045 | 1.054 | 0.482 | 1.424 |
| Q3 | -0.500 | 1.040 | -0.098 | 1.068 | 0.402 | 1.189 |
| Q4 | -0.682 | 0.899 | -0.205 | 1.167 | 0.455 | 1.263 |
| Q5 | 0.345 | 0.947 | 0.839 | 0.804 | 0.509 | 0.791 |
| Q6 | 0.725 | 0.940 | 0.873 | 1.055 | 0.100 | 0.995 |
| Q7 | 0.349 | 0.988 | 0.545 | 1.020 | 0.240 | 1.027 |
| Q8 | 0.223 | 0.825 | 0.830 | 0.752 | 0.607 | 0.943 |
| Q9 | 0.082 | 0.891 | 0.330 | 0.926 | 0.255 | 1.027 |
| Q10 | -0.045 | 1.001 | 0.300 | 1.003 | 0.345 | 0.990 |
| Q11 | -1.464 | 0.785 | -1.509 | 0.795 | -0.045 | 0.955 |

Survey results (cont.)

- Q 5, 6, 7, 9, and 10 address R Markdown's role in the data analysis workflow: for all five questions, the students responses were favorable at the end of semester, and grew more favorable over the course of the semester
- Q 3, 4, and 8 address the issue of frustration with R and R Markdown.
- While initial frustration with both R and R Markdown was reasonably high, by the end of the semester it had largely dissipated
- Students grew to appreciate R Markdown's ability to streamline their homework workflow
- In particular, students did not prefer to copy-and-paste their work from R into Microsoft Word.

Survey results (cont.)

- There was little to no correlation between a student's attitude towards R Markdown and that student's performance in the course
- Prior exposure to markup languages similar to R Markdown was not an impediment to learning R Markdown

Upper level classes

- build on first exposure in second course, more substantive projects
- leverages existing experiences to move along further in the second course
- instructor can scaffold more complicated manipulations and analysis
- feasible to get AP stats students directly into second courses which demonstrate the excitement of statistics
- leads to capstone, internship or other culminating experience

Other co-curricular opportunities

- utilize R Markdown or `knitr` in collaborative projects (a la St. Olaf Center for Interdisciplinary Reserach) or consulting work with students
- allows storage, reuse and auditing of student code
- program review (by another student)
- technical review: verification of each number in paper (by another student)
- helps develop a cadre of students with confidence and expertise to guide other students

Data Science at Smith College Fall 2013 (Baumer)

- provides a practical foundation for students to compute with data
- demonstrates the entire data analysis cycle (forming a statistical question, data acquisition, cleaning, transforming, modeling and interpretation)
- introduce students to tools for data management, storage and manipulation that are common in data science (e.g. R, SQL)
- undertake practical analyses using real, large, messy data sets using those tools
- learn to think statistically in approaching all of these aspects of data analysis
- minimal pre-requisites (intro stats plus intro CS)

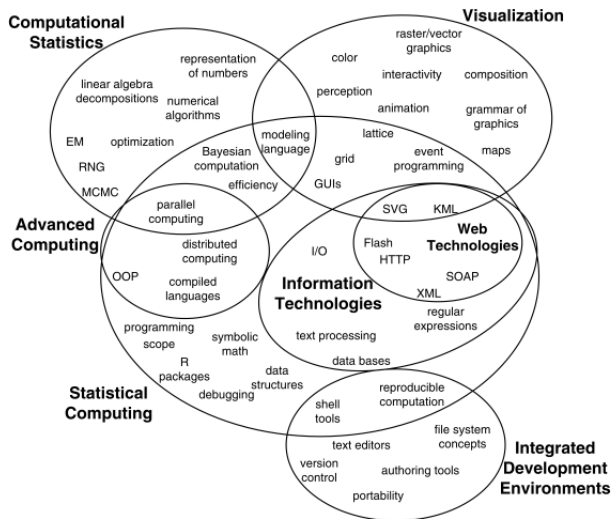
Other possible approaches

- Data Science** Ben Baumer (Smith College), Rachel Schutt's innovative book
- Data Mining** Modern Methods of Data Analysis (STAT 425), Brant Deppa (Winona State)
 - Math Stat** modified Moore method (empirical and analytic solutions)
- Advanced Data Analysis** Cosma Shalizi and Rebecca Nugent (CMU)
 - CIR** and similar models (summer research, year-long projects)

Closing thoughts

- Peck and Chance (2007): "When your students graduate, what is something observable that you think they ought to be able to do?"
- Nolan and Temple Lang (2010): "What ought our students be able to do computationally, and are we preparing them adequately in this regard?"
- Cobb argued (TISE, 2007) "Our courses teach techniques developed by pre-computer-era statisticians as a way to address their lack of computational power"
- we need a major cultural shift to embrace computing and fully integrate it into statistics programs, beginning at the undergraduate level

More from Nolan and Temple Lang



More from Nolan and Temple Lang

In the spirit of Peck and Chance, we ask, What do our students do when they graduate? We have found that Bachelors and Masters students who enter the workforce spend much of their efforts retrieving, filtering, and cleaning data and doing initial exploratory data analysis. These responsibilities increasingly demand working with different data technologies and having general programming skills. The potential for interesting and rich interactions for a statistician is greatly increased if he or she has a good knowledge of information technologies.

More from Nolan and Temple Lang

One former student recently wrote to us about exactly that:

I am currently working at a consulting firm that specializes in statistical and economic research and data analysis for large corporations. . . . Every day I work with data, and whether it is running regressions, cleaning data, finding summary statistics, parsing documents, or working in different database environments, [this statistical computing class] gave me the tools and foundation to succeed in my current position and gave me the confidence to land the job in the first place.

Summary

- reproducible statistical analyses are a (very) good thing
- ability to rerun analyses with different datasets or analytic decisions is a big win
- helps to model appropriate analytic workflow for new users
- reproducibility is emphasized throughout the semester
- requires some startup time to introduce students to this environment
- hurdles can be overcome, and the benefits are well worth it
- critical for students to develop the capacity to “compute with data”
- need for additional faculty development...

Activity

- download the file `markdown.Rmd` from `www.amherst.edu/~nhorton/stolaf`
- upload this to the RStudio server at `rstudio.stolaf.edu`
- run it and work through the questions listed there
- come up with your own questions and tell an interesting story

Precursors to the Data Explosion: Teaching How to Compute with Data

Nicholas J. Horton

Amherst College, Amherst, MA, USA

October 7, 2013

nhorton@amherst.edu