Reproducibility in Action

11 September 2017

Richard Schwinn, PhD
A Breakthrough at Major University University

Figure 1:
MD Anderson’s Dr. Baggerly Attempts to Verify
Meanwhile Major University Jumped Straight into Clinical Trials

Figure 3:
The First Investigation

Figure 4:
Results Vindicated

Figure 5:
The Clinical Trials Continued

Figure 6:

WALTER JACOBS:
I'm recording this with your permission.
The Clinical Trials Continued

Figure 7:

DR. POTTI: Absolutely. That's a good thing, 'cause you're gonna miss a lot.
Video

60 Minutes Video
Undeterred

Figure 8:
Undeterred

Figure 9:
Finally... After 5 Years of Fighting

Figure 10:
Follow up

Figure 11:
Why were Dr. Baggerly and the whistle blower ignored?
Because the audience for cutting edge research is small.
The Audience for Cutting Edge Research is Small

Figure 13:
What can be done?

- In the past, the only solution was to rely on
  - trust among patients and
  - honor among doctors.

- An economist however would recommend
  - increasing the expected cost of deception
  - \( E[C(\omega)] = \text{expected cost} \)
    - \( \omega = \text{level of deception} \)
  - or the probability of detection.

\[
E[C] = \int_{\Omega} C(\omega)P(d\omega)
\]
Retraction Watch Increases the Cost of Deception

Figure 14:
Retraction Watch

Websites like retractionwatch.com spread the word

1. They increase the severity of the stigma and reputational effects for fraudulent researchers.
2. The other option is to increase the probability of detection.

\[ E[C] = \int_{\Omega} C(\omega) P(d\omega) \]
Combating Deceptive Research

- By increasing the probability of detection, reproducible research
  - reduces the incentive to commit fraud and
  - it makes identifying subtle, unintentional errors easier.

- Reproducible research has a precise definition:
- Research is considered *reproducible* if
  - it is published with both
  - data
  - and code
  - so that it is easy for a non-expert to reproduce the results.

\[
E[C] = \int_{\Omega} C(\omega)P(d\omega)
\]
What are the tools of Reproducibility?

- Reproducibility software
  - Generates all statistical results from the original data in one step.

- Full reproducibility includes
  - all figures
  - tables
  - and language integration so that changes in the data makes meaningful changes to the text.
Software Options

- Literate programming languages
  - such as LaTeX
- Combined with statistical software
  - like SAS
  - and iPython Notebook
- R-Studio integrates a number of programming languages under the extremely easy to use rmarkdown language.
RMarkdown Cheatsheet

5. Embed Code
   Use knitr syntax to embed R code into your report. R will run the code and include the results when you render your report.

   **inline code**
   Surround code with back ticks and `R`. R replaces inline code with its results.

   ```
   Two plus two equals 4.
   ```

   **code chunks**
   Start a chunk with ````R`. End a chunk with ````.

   ```
   dim(iris)
   #> [1] 150 5
   ```

   **display options**
   Use knitr options to style the output of a chunk. Place options in brackets above the chunk.

   ```
   dim(iris)
   ```

6. Render
   Use your .Rmd file as a blueprint to build a finished report.

   Render your report in one of two ways
   1. Run `markdown::render("<file path>")`
   2. Click the `knit HTML` button at the top of the RStudio scripts pane

   When you render, R will
   - execute each embedded code chunk and insert the results into your report
   - build a new version of your report in the output file type
   - open a preview of the output file in the viewer pane
   - save the output file in your working directory

7. Interactive Docs
   Turn your report into an interactive Shiny document in 3 steps

   1. Add runtime: shiny to the YAML header
   2. In the code chunks, add Shiny input functions to embed widgets. Add Shiny render functions to embed reactive output
   3. Render with `rmarkdown::run` or click `Run Document in RStudio`

   *Note: your report will be a Shiny app, which means you must choose an html output format, like `html_document` for an interactive report or `slides_presentation` for an interactive slideshow.*

8. Publish
   Share your report where users can visit it online

   **Rpubs.com**
   Share non-interactive documents on RStudio's free R Markdwon publishing site www.rpubs.com

   **ShinyApps.io**
   Host an interactive document on RStudio's server. Free and paid options
   www.shinyapps.io

9. Learn More
   Documentation and examples - markdown.rstudio.com
   Further Articles - shiny.rstudio.com/articles
   - blog.rstudio.com
   - @rstudio

---

**Figure 15:**
Figure 16:
US Small Business GDP

2.3.12 Other Services, Except Governmental

Small businesses earned $806.8 billion, or 82.8%, of the $980.0 billion other services, except government services. In 2011, making it the 8th largest industry among small businesses (12th overall). The industry paid 14% in employee per small business, earning $8.8 as compared to $10 per small business across all industries.

During the previous period from 1998 to 2011, the other services, except government services expanded and contracted across various markets. Small business GDP grew 14.9% to $806.8 billion, the number of small businesses grew 21.0% to 47.3 million, and the number of small business employees grew 21.0% to 47.3 million.

From 2007 to 2011, other services, except government services both expanded and contracted across various markets. Small business GDP grew 14.9% to $806.8 billion, the number of small businesses grew 21.0% to 47.3 million, and the number of small business employees grew 21.0% to 47.3 million.

Across the overall 1998 to 2011 period, small business GDP grew 56.4%, the number of small businesses grew 22.0%, and the number of small business employees grew 56.4%.

82% of other services, except government services are attributable to small businesses

85% of other services, except government services are attributable to small businesses

Figure 17:
2 questions

- By a show of hands, who has never used a car sharing platform?
- Can I also ask, who has wanted to use a car sharing platform but it was unavailable in their area?

Suppose you work for Uber.

- You want to pitch the directors on expanding to new areas in Florida
Car Sharing Demo

Figure 18:
library(knitr) # Generates report
library(dplyr) # Wrangles data
library(choroplethhr) # Creates maps
library(choroplethhrMaps) # County data
library(ggplot2) # Creates graphics
library(gridExtra) # Arranges graphics
library(acs) # Downloads data
library(stringr) # Wraps labels

knitr::opts_chunk$set(...)
Step 2: Data Downloading

demo_df = acs_data_prep(c("B01003", "B19301"))
commute_df = acs_data_prep("B08534", 1:10)
transport_df = acs_data_prep("B08301", c(2,10,16:20))
aggregate_df = acs_data_prep("B08135", 1)
df = rbind(demo_df, commute_df, transport_df, aggregate_df)
Step 3: Create Statewide Maps

maps_list =
c("B01003", # Total Population
    "B19301", # Income
    "B08534", # Number of commuters
    "B08135") # Aggregate Travel Time to Work

plot_maps = function(x) {
  it = filter(df, table_number == maps_list[x], index = 1)
  county_choropleth(it, state_zoom = tolower(state_name)) +
  scale_fill_brewer(palette = x) +
  ggttitle(it$table_title) +
  theme(legend.position = "bottom")
}
Step 4: Create County Level Reports

```r
state_counties = filter(df, state.name == tolower(state_name), table_number == "B08135") %>% arrange(state_rank) # Selects county for the state

make_county_reports = function(x) {
  ...
}

county_reports = lapply((1:nrow(state_counties)), make_county_reports)

county_reports[1:nrow(state_counties)]
```
Knit

Figure 19:
Thank You

- We have seen that
  - Reproducibility tools can combat falsified research
  - That these tools can be used profitably for regional and periodical reporting
  - and that complex and useful reports can be created in a matter minutes
Outline

- What is SUSB?
- History and Generation of SUSB
- Uses of SUSB
- SUSB Data Challenges
- Future Availability via the SUSB Data Explorer
What is SUSB?

The Statistics of U.S. Businesses (SUSB) is an annual dataset that provides data on

- Numbers of businesses
- Employment
- Revenues
- Births and deaths
- Expansions and contractions
- Payroll
  - For firms and for establishments
  - By size, by, industry, and by geography
History and Generation of SUSB

Figure 20: Census logo
Uses of SUSB

Figure 21:
Uses of SUSB

Google

Census SUSB Statistics of Businesses

Scholar

About 19,000 results (0.12 sec)

Articles

[BOOK] An analysis of small business and jobs
B Headd - 2010 - sba.gov
... differing sub-sectors of small business have reacted in previous
Office of Advocacy, from data provided by the US Department of La
BED and US Department of Commerce, Census Bureau, SUSB. ..
Cited by 60  Related articles  All 6 versions  Cite  Save  More

Case law

The growth, decline and survival of small business life cycles
B Headd, B Kirchhoff - Journal of Small Business Management, 20
... Source: US Census Bureau, Statistics of US Businesses, spec
from the SUSB database contained in the research are drawn from
statistics primarily reflect the period of 1992 to 2002 unless otherw
Cited by 110  Related articles  All 5 versions  Cite  Save

My library

Custom range...

Any time

Any time

Since 2017
Since 2016
Since 2013
Custom range...

Sort by relevance
Sort by date

Figure 22:
SUSB Data Challenges
SUSB Data Challenges

- 4162 Geographies
  - 1 national stats
  - 51 states
  - 917 metropolitan statistical areas
  - 3193 counties
- 26 firm sizes
- 2016 industries
- 7 variables (employment, number of firm, etc.)
- 20+ years of data

If data were provided for all permutations, it would represent well over 30 trillion elements. The existing source tables consist of only between 300 and 400 million cells.
SUSB Data Challenges

Figure 24:
### SUSB Data Challenges

<table>
<thead>
<tr>
<th>File Name</th>
<th>Size</th>
<th>Date/Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>susb_2012_t7_b.XLSB</td>
<td>165 KB</td>
<td>May 16, 2017, 7:59 PM</td>
</tr>
<tr>
<td>susb_2012_t8.XLSB</td>
<td>4 MB</td>
<td>May 16, 2017, 7:59 PM</td>
</tr>
<tr>
<td>susb_2013_t1_a.XLSB</td>
<td>14 MB</td>
<td>May 16, 2017, 7:59 PM</td>
</tr>
<tr>
<td>susb_2013_t1_b.XLSB</td>
<td>1.7 MB</td>
<td>May 16, 2017, 7:59 PM</td>
</tr>
<tr>
<td>susb_2013_t3_a.XLSB</td>
<td>17.4 MB</td>
<td>May 16, 2017, 8:00 PM</td>
</tr>
<tr>
<td>susb_2013_t3_b.XLSB</td>
<td>13.1 MB</td>
<td>May 16, 2017, 8:00 PM</td>
</tr>
<tr>
<td>susb_2013_t4.XLSB</td>
<td>521 KB</td>
<td>May 16, 2017, 8:00 PM</td>
</tr>
<tr>
<td>susb_2013_t5.XLSB</td>
<td>54 KB</td>
<td>May 16, 2017, 8:00 PM</td>
</tr>
<tr>
<td>susb_2013_t6_a.XLSB</td>
<td>1.3 MB</td>
<td>May 16, 2017, 8:00 PM</td>
</tr>
<tr>
<td>susb_2013_t6_b.XLSB</td>
<td>28 KB</td>
<td>May 16, 2017, 8:00 PM</td>
</tr>
<tr>
<td>susb_2013_t7_a.XLSB</td>
<td>1.6 MB</td>
<td>May 16, 2017, 8:00 PM</td>
</tr>
<tr>
<td>susb_2013_t7_b.XLSB</td>
<td>161 KB</td>
<td>May 16, 2017, 8:00 PM</td>
</tr>
<tr>
<td>susb_2014_t1_a.XLSB</td>
<td>14 MB</td>
<td>May 16, 2017, 8:01 PM</td>
</tr>
<tr>
<td>susb_2014_t1_b.XLSB</td>
<td>1.7 MB</td>
<td>May 16, 2017, 8:01 PM</td>
</tr>
<tr>
<td>susb_2014_t3_a.XLSB</td>
<td>17.5 MB</td>
<td>May 16, 2017, 8:01 PM</td>
</tr>
<tr>
<td>susb_2014_t3_b.XLSB</td>
<td>12.8 MB</td>
<td>May 16, 2017, 8:01 PM</td>
</tr>
<tr>
<td>susb_2014_t4.XLSB</td>
<td>521 KB</td>
<td>May 16, 2017, 8:01 PM</td>
</tr>
<tr>
<td>susb_2014_t5.XLSB</td>
<td>55 KB</td>
<td>May 16, 2017, 8:01 PM</td>
</tr>
</tbody>
</table>

*Figure 25:*
Future Availability via the SUSB Data Explorer

![SUSB Data Explorer](image)

<table>
<thead>
<tr>
<th>naics_code</th>
<th>naics_north_american_industry_classification_system</th>
<th>enterprise_receipt_size</th>
<th>number_of_firms</th>
<th>number_of_establishments</th>
<th>employer</th>
</tr>
</thead>
<tbody>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>&lt;100,000</td>
<td>1015</td>
<td>1016</td>
<td>13</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>100,000-499,999</td>
<td>3431</td>
<td>3433</td>
<td>155</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>500,000-999,999</td>
<td>1933</td>
<td>1943</td>
<td>233</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>1,000,000-2,499,999</td>
<td>2058</td>
<td>2089</td>
<td>553</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>2,500,000-4,999,999</td>
<td>1114</td>
<td>1183</td>
<td>607</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>5,000,000-7,499,999</td>
<td>515</td>
<td>550</td>
<td>411</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>7,500,000-9,999,999</td>
<td>230</td>
<td>272</td>
<td>242</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>10,000,000-14,999,999</td>
<td>167</td>
<td>224</td>
<td>226</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>15,000,000-19,999,999</td>
<td>44</td>
<td>78</td>
<td>69</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>20,000,000-24,999,999</td>
<td>34</td>
<td>63</td>
<td>68</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>25,000,000-29,999,999</td>
<td>24</td>
<td>39</td>
<td>40</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>30,000,000-34,999,999</td>
<td>10</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>35,000,000-39,999,999</td>
<td>6</td>
<td>34</td>
<td>24</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>40,000,000-49,999,999</td>
<td>13</td>
<td>105</td>
<td></td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>50,000,000-74,999,999</td>
<td>15</td>
<td>29</td>
<td>17</td>
</tr>
<tr>
<td>71391</td>
<td>Golf Courses and Country Clubs</td>
<td>75,000,000-99,999,999</td>
<td>12</td>
<td>64</td>
<td>26</td>
</tr>
</tbody>
</table>