

# Reproducibility in Action

26 July 2017

Richard Schwinn, PhD

# A Breakthrough at Major University University



Figure 1:

# A Breakthrough at Major University (Pile of Cash)

- In 2006, a major university saw a billion dollar industry on the horizon.
- One of its primary researchers, we'll call him Dr. P., designed a test to show that
  - DNA signatures could predict whether patients would respond to specific chemotherapy drugs (*P. et al., 2006*).
- Such a test would be hugely valuable.
  - If it said a particular treatment wouldn't work, it could be skipped,
    - instead of uselessly poisoning the patient for 6 weeks.
  - With this test, patients would get the treatments most likely to work for them.

## MD Anderson's Dr. Baggerly Attempts to Verify



Figure 2:

# MD Anderson's Dr. Baggerly Attempts to Verify (Picture of Dr. Baggerly)

- A competing hospital, **MD Anderson Cancer Institute**, was also excited about the possible treatment.
  - But before investing into it, they wanted to be sure that they completely understood the underlying research.
  - So they assigned Dr. Keith Baggerly to replicate the empirical results.
- Everyone was very optimistic especially because
  - Dr. P. had provided data along with his paper.

# Meanwhile Major University Jumped Straight into Clinical Trials

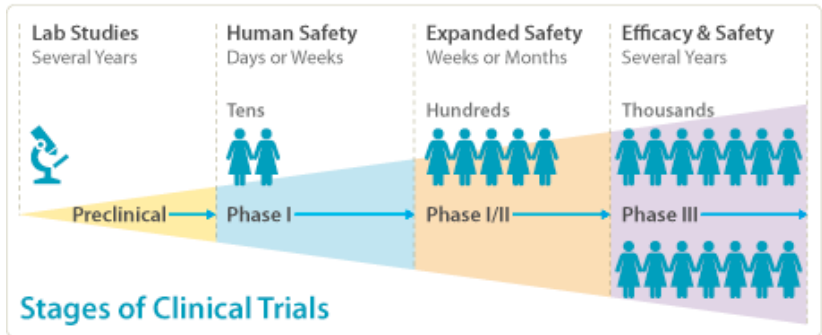


Figure 3:

# Meanwhile Major University Jumped Straight into Clinical Trials

- While Baggerly was trying to make sense of the data, the Major University jumped straight into clinical trials. (PDF and Excel icons)
- After trying to replicate the findings for several months and noticing massive errors Dr. Baggerly contacted Major University's Dr. P.
  - Within 6 months, Dr. P. cut off all communication.

# The First Investigation



Figure 4:



# The First Investigation (magnifying glass)

- In response to Dr. Baggerly's persistence, Major University convened an outside panel to investigate the results
  - And...

## Results Vindicated



Figure 5:

# Results Vindicated

- P.'s findings were vindicated
  - The panel concluded that the research was sound!
  - They said that they could find nothing wrong with it.

# The Clinical Trials Continued



Figure 6:

# The Clinical Trials Continued



Figure 7:

# Video

60 Minutes Video

# The Clinical Trials Continued

- Here are two screen shots from a 60 minutes special this story.
  - The husband of one of Dr. P's patients from the the clinical trial as Dr. P for his
    - permission to record their conversation.
  - Dr. P responds, "Absolutely. That's a good thing, 'cause *you're going to miss a lot.*"
    - This is an important point. Dr. P would the same to anyone asking about his research, even hi colleagues those investigating his work like Baggerly.

# Undeterred



Figure 8:



# Undeterred



Figure 9:

# Undeterred

Baggerly repeatedly showed that Dr. P.'s research was intentionally bad.

- Yet his analyses were either ignored or criticized for being too technical.
- Like the Erin Brockovic of health research, Dr. Baggerly didn't stop raising the red flag.

Finally... After 5 Years of Fighting



Figure 10:

## Finally... After 5 Years of Fighting (Rhodes Scholar Image)

Paul Goldberg, an editor of the Cancer Letter, a small trades publication,

- Who had been previously contacted by Baggerly, noticed that Dr. P. had lied about being a Rhodes Scholar.
  - This was a lie everyone could understand.
  - Finally a story about Dr. P. gets traction!

Follow up



Figure 11:

# Follow up

- In support of his friend, Dr. P.'s co-author, Dr. N.,
  - attempted to reproduce the research from the raw data.
  - And he immediately discovered data manipulations that
  - in his own words, "Couldn't be inadvertent."

## Aftermath - Brad Perez



Figure 12:

# Aftermath

- Some clinical trial patients, who were denied standard treatments based on this research, died.
  - Today, 11 lawsuits have been settled. Others are still pending.
  - Two thirds of Dr. P.'s papers have been retracted
- One of the most shocking aspects of this story is that
  - Brad Perez, a student in Dr. P.'s lab at Major University at the time, had written a thorough whistle blower report discrediting the research back in 2006... and it was ignored!
  - Today Dr. Perez is an oncologist in Tampa, Florida.



# Why?

- 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:

# The Audience for Cutting Edge Research is Small

- Cutting-edge research is
  - written for a small group of experts.
- In many fields, peer reviewers qualified enough to understand cutting-edge research
  - face high opportunity costs and
  - almost never have time to verify research based on raw data.
- These barriers open a door for those willing to commit outright fraud to go unnoticed.

## 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
  - $E[C(\omega)] =$  expected cost
    - $\omega =$  level of deception
  - or the probability of detection.

$$E[C] = \int_{\Omega} C(\omega)P(d\omega)$$

# Retraction Watch



Figure 14:

# Retraction Watch

Websites like [retractionwatch.com](http://retractionwatch.com) spread the word

1. Thus increasing the severity of the stigma and reputational effects for bad researchers.
2. The other option is to increase the *probability of detection*.

# 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.

# 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 **markdown** 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 `r 2 + 2`.

Two plus two equals 4.

## code chunks

Start a chunk with ````{r}`. End a chunk with `````.

Here's some code  
````{r}  
dln(iris)  
````

Here's some code  
`dln(iris)`

`## [1] 150 5`

## display options

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

Here's some code  
````{r eval=FALSE}  
dln(iris)  
````

Here's some code  
`dln(iris)`

Here's some code  
````{r echo=FALSE}  
dln(iris)  
````

Here's some code  
`## [1] 150 5`

option	default	effect
<code>eval</code>	TRUE	Whether to evaluate the code and include its results
<code>echo</code>	TRUE	Whether to display code along with its results
<code>warning</code>	TRUE	Whether to display warnings
<code>error</code>	FALSE	Whether to display errors
<code>message</code>	TRUE	Whether to display messages
<code>tidy</code>	FALSE	Whether to reformat code in a tidy way when displaying it
<code>results</code>	"markup"	"markup", "asis", "hold", or "hide"
<code>cache</code>	FALSE	Whether to cache results for future renders
<code>comment</code>	"##"	Comment character to preface results with
<code>fig.width</code>	7	Width in inches for plots created in chunk
<code>fig.height</code>	7	Height in inches for plots created in chunk

For more details visit [yihui.name/knitr/](http://yihui.name/knitr/)

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

Render your report in one of two ways

1. Run `rmarkdown::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

```
---
title: "Line graph"
output: html_document
runtime: shiny
---
```

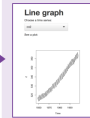
2. In the code chunks, add Shiny input functions to embed widgets. Add Shiny render functions to embed reactive output

```
---
title: "Line graph"
output: html_document
runtime: shiny
---

Choose a time series:
☐ echo = FALSE
selectInput("data", "",
  c("a", "b"))

See a plot:
☐ echo = FALSE
renderPlot({
  d = getInput("data")
  plot(d)
})
```

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 `ioslides_presentation` (for an interactive slideshow).

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

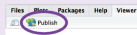
## Rpubs.com

Share non-interactive documents on RStudio's free R Markdown publishing site  
[www.rpubs.com](http://www.rpubs.com)

## ShinyApps.io

Host an interactive document on RStudio's server. Free and paid options  
[www.shinyapps.io](http://www.shinyapps.io)

Click the "Publish" button in the RStudio preview window to publish to [pubs.com](http://pubs.com) with one click.



**9. Learn More**

Documentation and examples - [rmarkdown.rstudio.com](http://rmarkdown.rstudio.com)

Further Articles - [shiny.rstudio.com/articles](http://shiny.rstudio.com/articles)

Blog - [blog.rstudio.com](http://blog.rstudio.com)

@rstudio



RStudio® and Shiny® are trademarks of RStudio, Inc.  
CC-BY RStudio is licensed under the CC-BY license  
044-448-1117-rstudio.com

Figure 15:

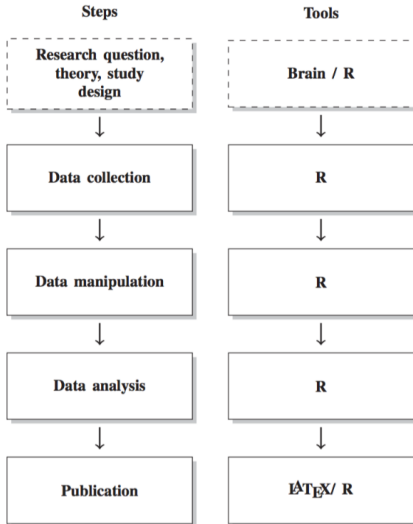
# RMarkdown Cheatsheet

- RMarkdown can be learned by a person with no programming experience in one afternoon.
- RStudio provides a cheat sheet covering all of its important commands that fits on one sheet of paper. (double sided)

# R Work Flow

1. Generating reports in R and RMarkdown decreases the cost of investigating the accuracy of academic research.
2. Data is cheaper than ever.
  - A little study of basic Application Programming Interfaces (APIs) documentation enables
  - researchers to do things that were previously impossible to all but a small group of computer scientists.
3. Reproducibility offers tremendous time-savings opportunities for regional and periodical reporting through parameterized reports.
  - **RMarkdown** enables you to rerun your data-processing and, with a little work, even the textual content of your document can be updated in one click of the **Knit** button.

# R Work Flow



# R Output



# RStudio Output

- RStudio unifies several programming languages
  - through RMarkdown to easily output content to multiple formats.

# US SBA

SBP\_2014.pdf



49



50



51



## Small Business Profile

Advocacy: the voice of small business in government

### Florida

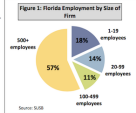
2,180,536 Small Businesses  
404,951 Small Businesses with Employees  
1,775,605 Small Businesses without Employees (Nonemployers)  
5,603,167 Workers Employed by Small Businesses

#### State Economy Overall

- Florida's economy grew at the same rate as the national economy in 2013. Florida's real gross state product and U.S. gross domestic product both increased by 2.8 percent. (Source: BEA)
- The employment picture in Florida has improved. The unemployment rate in Florida declined from 6.7 percent in October 2013 to 6.0 percent in October 2014. This is above the national average of 5.8 percent for the same time period. (Source: BLS)

#### Employment

- Florida's small businesses employed over two-fifths or 3 million of the state's private workforce in 2012. (Source: SUSEB)
- Almost all firms with employees are small. They make up 98.9 percent of all employers in the state. (Source: SUSEB)
- Firms with fewer than 100 employees have the largest share of small business employment. Figure 1 offers further detail.
- In Florida, small businesses created 228,568 net new jobs in 2012. The biggest gain was in the smallest firm size category of 100-499 employees. (Source: BDC)
- The number of people who were primarily self-employed in 2013 decreased by 2.6 percent relative to the previous year.
- The state's private-sector employment growth increased by 3.1 percent over the 12-month period ending in October 2014. (This was above the national average growth rate of 2.9 percent. (Source: BLS))



The Small Business State and Territory Profiles report on the economic status of small business from 2007 to 2014. They include information on the number of firms, employment, demographics and other topics using the most recently available government data. They are a reference tool for researchers, policymakers, and small entities who are interested in how small firms are performing regionally or nationally. Note that this report defines small businesses as firms with fewer than 500 employees.

#### Income and Finance

- The number of banks reported in the Call Reports between June 2013 and June 2014 declined. (Source: FDIC)
- In 2012, 346,245 loans under \$100,000 (and valued at \$4.1 billion) were issued by Community Reinvestment Act lending institutions in Florida. (Source: FFIEC)

SBP\_2017.pdf



41



42



43



## SMALL BUSINESS PROFILE

U.S. SMALL BUSINESS ADMINISTRATION  
OFFICE OF ADVOCACY

ISSUES • RESOURCES • SERVICES

### FLORIDA

2.4 million  
99.9% Small Businesses of Florida Businesses

3.2 million  
42.8% Small Business Employees of Florida Employees

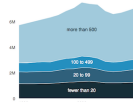


#### Overall Florida Economy

- In the second quarter of 2016, Florida grew at an annual rate of 2.3%, which was faster than the overall US growth rate of 1.2%. Florida's 2015 growth rate of 4.0% was up from the 2014 rate of 2.9%. (Source: BEA)
- In November 2016, the unemployment rate was 4.0%, down from 5.1% at the close of 2015. This was above the November 2014 national unemployment rate of 4.0%. (Source: CTR)

#### Employment

Figure 1: Florida Employment by Business Size (5 Employees)



- Florida small businesses employed 3.2 million people, or 42.8% of the private workforce, in 2014. (Source: SUSEB)
- Firms with fewer than 100 employees have the largest share of small business employment. See Figure 1 for further details on firms with employees. (Source: SUSEB)
- During the year ending November 2014, private-sector employment increased 0.2%. This was above the previous year's increase of 1.0%. (Source: CPS)
- The number of proprietors increased in 2015 by 3.0% relative to the previous year. (Source: BEA)
- Small business created 152,735 net jobs in 2014. Among the seven BDC size classes, firms employing 20 to 49 employees experienced the largest gains, adding 24,241 net jobs. The smallest net gains were in firms employing 5 to 9 employees, which added 11,095 net jobs. (Source: BDC)

The Small Business Profiles are produced by the US Small Business Administration's Office of Advocacy. Each report incorporates the most up-to-date government data to present a unique snapshot of small businesses. Small businesses are defined as firms employing fewer than 500 employees. Net small business job change, minority and business ownership, and register share statistics are based on the 2014 Business Dynamics Statistics (BDS), 2012 Survey of Business Owners (SBO), and 2014 International Trade Administration (ITA) data, respectively.

SBA Office of Advocacy

41

Florida Small Business Profile, 2017

#### Income and Finance

- The number of banks reported in the Call Reports between June 2015 and June 2016 declined. (Source: FDIC)
- In 2014, 369,076 loans under \$100,000 (valued at \$4.7 billion) were issued by Florida lending institutions reporting under the Community Reinvestment Act. (Source: FFIEC)

Figure 16:



# US SBA

- These tools were created to ensure the quality of academic research
  - but they are incredibly useful outside of academia
- I am an economist at the USA Small Business Administration.
  - One of our stated objectives is to report on the status of small businesses.
  - We issue an annual small business profile to meet this goal.

# US Small Business GDP

- Another one of our statutory goals is report on the contributions of small business.
- Using R Markdown allows us create
  - an adaptive report
  - with new graphical elements
    - such as a modified Tufte layout
    - hierarchical maps
    - spark graphs
    - waffle charts
  - and parameterized industry level profiles
    - which provide dozens of statistics and graphics

## WALL, BENDER, ET AL.

## 2.2 THE SMALL BUSINESS SHARE OF GDP

<sup>a</sup> A portion of the 1998-2000 decline is due to a methodological change discussed in *Section 3*.

## SMALL BUSINESS GROWTH 6



The Educational Services Indicator refers to private sector educational institutions and not public schools.



Industry GDP<sup>a</sup> is attributable to small businesses

85%  
of Other Services, Except Government

employees work for small business

Figure 17:

# Faux Market Research: Car Sharing

- 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

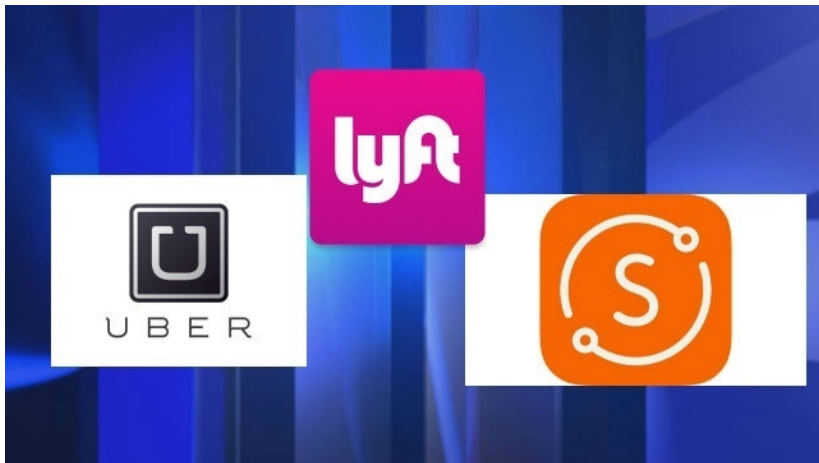


Figure 18:

## Step 1: Load Libraries and knitr options

```
library(knitr)           # Generates report
library(dplyr)           # Wrangles data
library(choroplethr)     # Creates maps
library(choroplethrMaps) # 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 =  
  county_choropleth(it, state_zoom = tolower(state_name)  
  scale_fill_brewer(palette = x) +  
  ggtitle(it$table_title) +  
  theme(legend.position = "bottom")  
}
```



## Step 4: Create County Level Reports

```
state_counties = filter(df, state.name == tolower(state.name))

make_county_reports = function(x) {...}

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

- Definitions
- History
- Uses of SUSB
- Challenges to Accessing SUSB
- Future Availability via the SUSB Data Explorer

# Outline

Here is a quick outline of my introduction to the Statistics of U.S. Businesses

- At first glance, you might think these topics are a little unexciting.
  - But if you give it a moment, and look closely at this list,
  - you will notice that, indeed, it is not exciting.
- Luckily for you, after my discussion on the programmatic manipulation of the underlying data
- You'll hear from Miriam whose detailed mastery of the nuances SUSB makes all of this possible
  - Miriam can tell you about which data exists and why or why not
  - she can explain why reported totals and table sums do not necessarily align, and so much more
  - and today she will tell you about some existing OER tools on the Advocacy webpage

# 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

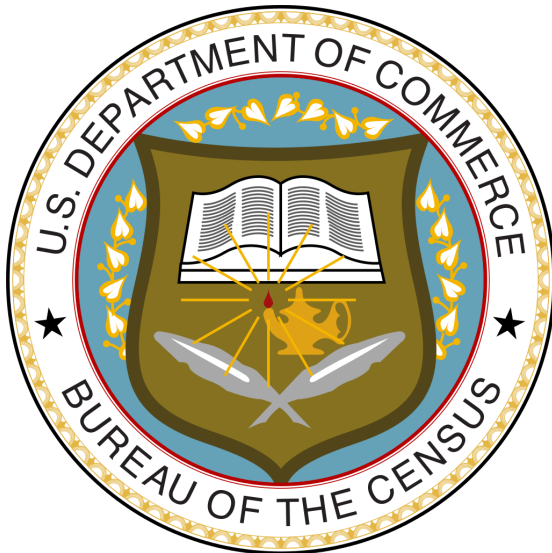


Figure 20: Census logo

## History and Generation of SUSB

- SUSB began in 1989 as an extension of the County Business Patterns (CBP) which itself began in 1964 and also grew out of similar data projects dating back to 1946
- The data underlying SUSB come from the Business Register (BR), which is a database of all known employers maintained by the U.S. Census Bureau
- Additional data come from various Census Bureau programs, such as
  - the Economic Census
  - the Annual Survey of Manufactures
  - the Current Business Surveys
  - the Company Organization Survey
- Noise infusion methodology is applied to protect individual business establishments from disclosure
- Advocacy funds SUSB through an inter-agency-agreement (IAA)



# Uses of SUSB 1



Figure 21:

# Uses of SUSB 1

- SUSB is the premiere source of data on small businesses
- The data are referenced extensively
  - by congress
  - in leading newspapers
  - and, most prominently, in academic research

# Uses of SUSB 2



Census SUSB Statistics of Businesses

Scholar

About 19,000 results (0.12 sec)

Articles

Case law

My library

Any time

Since 2017

Since 2016

Since 2013

Custom range...

Sort by relevance

Sort by date

[\[BOOK\] 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 Labor and US Department of Commerce, **Census** Bureau, **SUSB**. ..  
[Cited by 60](#) [Related articles](#) [All 6 versions](#) [Cite](#) [Save](#) [More](#)

[The growth, decline and survival of small business life cycles](#)  
B Headd, B Kirchoff - [Journal of Small Business Management](#), 2010  
... Source: US **Census** Bureau, **Statistics** of US **Businesses**, **statistics** primarily reflect the period of 1992 to 2002 unless otherwise noted.  
[Cited by 110](#) [Related articles](#) [All 5 versions](#) [Cite](#) [Save](#)

[An analysis of small business size and rate of disc](#)  
T Bates, A Nucci - [Journal of Small Business Management](#), 1989 -

Figure 22:

## Uses of SUSB 2

A quick Google Scholar search yields

- Over 19,000 references in academic publications
- And there already more than 2,500 citations this year alone
- The data are regularly incorporated into research from leading institutions like the University of Chicago (Hurst), Columbia (Benjamin Pugsley), Harvard, etc.
- You might recognize the author of the top search result for SUSB
- It's a major credit to our office that Advocacy's own, Brian Headd, appears in the top rank
  - Between his book, *An Analysis of Small Business and Jobs*, and his other publications, Brian has hundreds of small business citations for his work based on SUSB

# SUSB Data Challenges 1



# SUSB Data Challenges 1

The Office of Advocacy disseminates the SUSB datasets jointly with the SUSB office at Census

- As a result we can request
  - special tabulations
  - changes in the tables such as breakdowns by various new employment size increments
- Overtime, these little internal changes
  - coupled with external changes like
    - changes to the definitions of industry names and
    - disclosure policies (noise infusion, complementary cell suppression)
  - have resulted in inconsistency in the format of the data over time
  - this makes the data difficult to organize and to use for analyses involving time series

## SUSB Data Challenges 2

- 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 2

- The real problem is that the data are both big and complex
- The raw SUSB data is about 13 GB and exists across over more than 800 different tables
  - I should note that while it doesn't qualify as "BigData" since it fits on a single computer
  - At more than 10 GBs and with single tables too large for excel to open it passes the threshold into the realm of "MediumData"
- We want to make the data accessible to more people, so we have devised a way to sensibly organize the entire database into 12 tables without sacrificing any information
- Each of the 12 tables views the information from a different perspective
  - such as by various geography
  - or by levels of industry detail



# SUSB Data Challenges 3

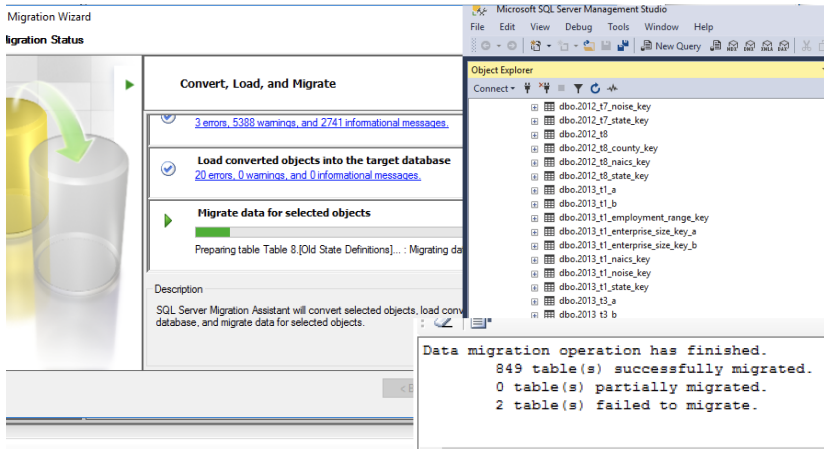


Figure 24:

## SUSB Data Challenges 3

- To achieve this, I used a set of free, open-source tools, originally intended for big data, to wrangle the data into a normalized database
- The great thing about these BigData tools is that they're
  - somehow ready for all the problems which are unknown to me but seem to arise anyway
- After a productive discussion with the SUSB office in April, they sent us a set of the original ACCESS source files
- I then migrated these files into an SQL database
- Next, I normalized naming conventions and merged the key tables with data tables
- So that the original data are now fully represented by ~160 tables complete with descriptive language for key variables
  - So for example, alongside the standard column of state codes numbered 1 to 51 is a column of state names

# SUSB Data Challenges 4




















	susb_2012_t7_b.XLSB	✓	May 16, 2017, 7:59 PM	165 KB	Micros...(.xlsb)
	susb_2012_t8.XLSB	✓	May 16, 2017, 7:59 PM	4 MB	Micros...(.xlsb)
	susb_2013_t1_a.XLSB	✓	May 16, 2017, 7:59 PM	14 MB	Micros...(.xlsb)
	susb_2013_t1_b.XLSB	✓	May 16, 2017, 7:59 PM	1.7 MB	Micros...(.xlsb)
	susb_2013_t3_a.XLSB	✓	May 16, 2017, 8:00 PM	17.4 MB	Micros...(.xlsb)
	susb_2013_t3_b.XLSB	✓	May 16, 2017, 8:00 PM	13.1 MB	Micros...(.xlsb)
	susb_2013_t4.XLSB	✓	May 16, 2017, 8:00 PM	521 KB	Micros...(.xlsb)
	susb_2013_t5.XLSB	✓	May 16, 2017, 8:00 PM	54 KB	Micros...(.xlsb)
	susb_2013_t6_a.XLSB	✓	May 16, 2017, 8:00 PM	1.3 MB	Micros...(.xlsb)
	susb_2013_t6_b.XLSB	✓	May 16, 2017, 8:00 PM	28 KB	Micros...(.xlsb)
	susb_2013_t7_a.XLSB	✓	May 16, 2017, 8:00 PM	1.6 MB	Micros...(.xlsb)
	susb_2013_t7_b.XLSB	✓	May 16, 2017, 8:00 PM	161 KB	Micros...(.xlsb)
	susb_2013_t8.XLSB	✓	May 16, 2017, 8:00 PM	4 MB	Micros...(.xlsb)
	susb_2014_t1_a.XLSB	✓	May 16, 2017, 8:01 PM	14 MB	Micros...(.xlsb)
	susb_2014_t1_b.XLSB	✓	May 16, 2017, 8:01 PM	1.7 MB	Micros...(.xlsb)
	susb_2014_t3_a.XLSB	✓	May 16, 2017, 8:01 PM	17.5 MB	Micros...(.xlsb)
	susb_2014_t3_b.XLSB	✓	May 16, 2017, 8:01 PM	12.8 MB	Micros...(.xlsb)
	susb_2014_t4.XLSB	✓	May 16, 2017, 8:01 PM	521 KB	Micros...(.xlsb)
	susb_2014_t5.XLSB	✓	May 16, 2017, 8:01 PM	55 KB	Micros...(.xlsb)

Figure 25:

## SUSB Data Challenges 4

- The data are now organized into roughly 160 tables, down from  $>800$ , with homogenized column names and formatting
- This was achieved
  - using regular expressions (UNIX programming conventions)
  - and open source software
- what makes this approach special is that none of the steps I've taken are point and click
- thus every single command is documented so that if I'd made any error, and thankfully I did not,
  - it would be easy to identify, correct, and rerun the normalization
  - Were we so inclined, we could recreate the normalized data from the Census source files (which represents months of work)

# Future Availability via the SUSB Data Explorer



## SUSB Data Explorer

Download Filtered Data

ving 1 to 25 of 36 entries (filtered from 34,722 total entries) Show 25 entries

Search: golf

	naics_code	naics_north_american_industry_classification_system	enterprise_receipt_size	number_of_firms	number_of_establishments	employer
345	71391	Golf Courses and Country Clubs	<100,000	1015	1016	13
346	71391	Golf Courses and Country Clubs	100,000-499,999	3431	3433	155
347	71391	Golf Courses and Country Clubs	500,000-999,999	1933	1943	233
348	71391	Golf Courses and Country Clubs	1,000,000-2,499,999	2058	2089	553
349	71391	Golf Courses and Country Clubs	2,500,000-4,999,999	1114	1183	607
350	71391	Golf Courses and Country Clubs	5,000,000-7,499,999	515	550	411
351	71391	Golf Courses and Country Clubs	7,500,000-9,999,999	230	272	242
352	71391	Golf Courses and Country Clubs	10,000,000-14,999,999	167	224	226
353	71391	Golf Courses and Country Clubs	15,000,000-19,999,999	44	78	69
354	71391	Golf Courses and Country Clubs	20,000,000-24,999,999	34	63	68
355	71391	Golf Courses and Country Clubs	25,000,000-29,999,999	24	39	40
356	71391	Golf Courses and Country Clubs	30,000,000-34,999,999	10	30	29
357	71391	Golf Courses and Country Clubs	35,000,000-39,999,999	6	34	24
358	71391	Golf Courses and Country Clubs	40,000,000-49,999,999	13	105	
359	71391	Golf Courses and Country Clubs	50,000,000-74,999,999	15	29	17
360	71391	Golf Courses and Country Clubs	75,000,000-99,999,999	12	64	26

## Future Availability via the SUSB Data Explorer

- Inspired by OER's Elizabeth Glass, the OER data team has held several meeting with data specialists to discuss the possibility of creating an online databank
  - Ultimately none were able to provide the solution we wanted
- So I decided to make my our own web interface last week
  - It's working name is the SUSB Data Explorer
  - It allows users to get directly to the data that they want
  - and to download it, and only it, instantly
- Pat and I, as well as potentially Miriam and Sarah,
  - have just enrolled in the Office of Entrepreneurial Development's Human Centered Design course
  - to ensure that the data explorer's interface is as useful as possible to our stakeholders.
- Demo