USING R ON SMALL TEAMS IN INDUSTRY

Jonathan Nolis, Director of Insights and Analytics

jnolis@lenati.com

@skyetetra

http://jnolis.com

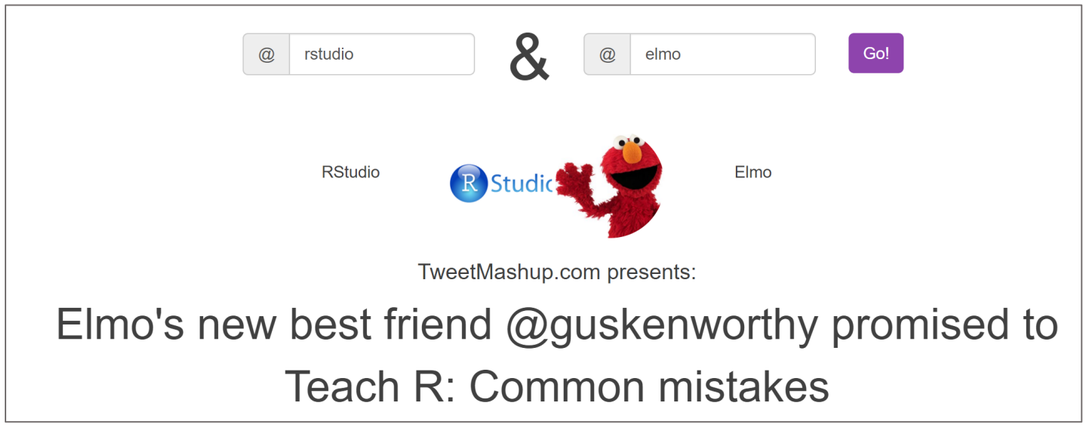
LENATI LLC

2412 7th Avenue West #101

Seattle, WA 98119

Who am I?

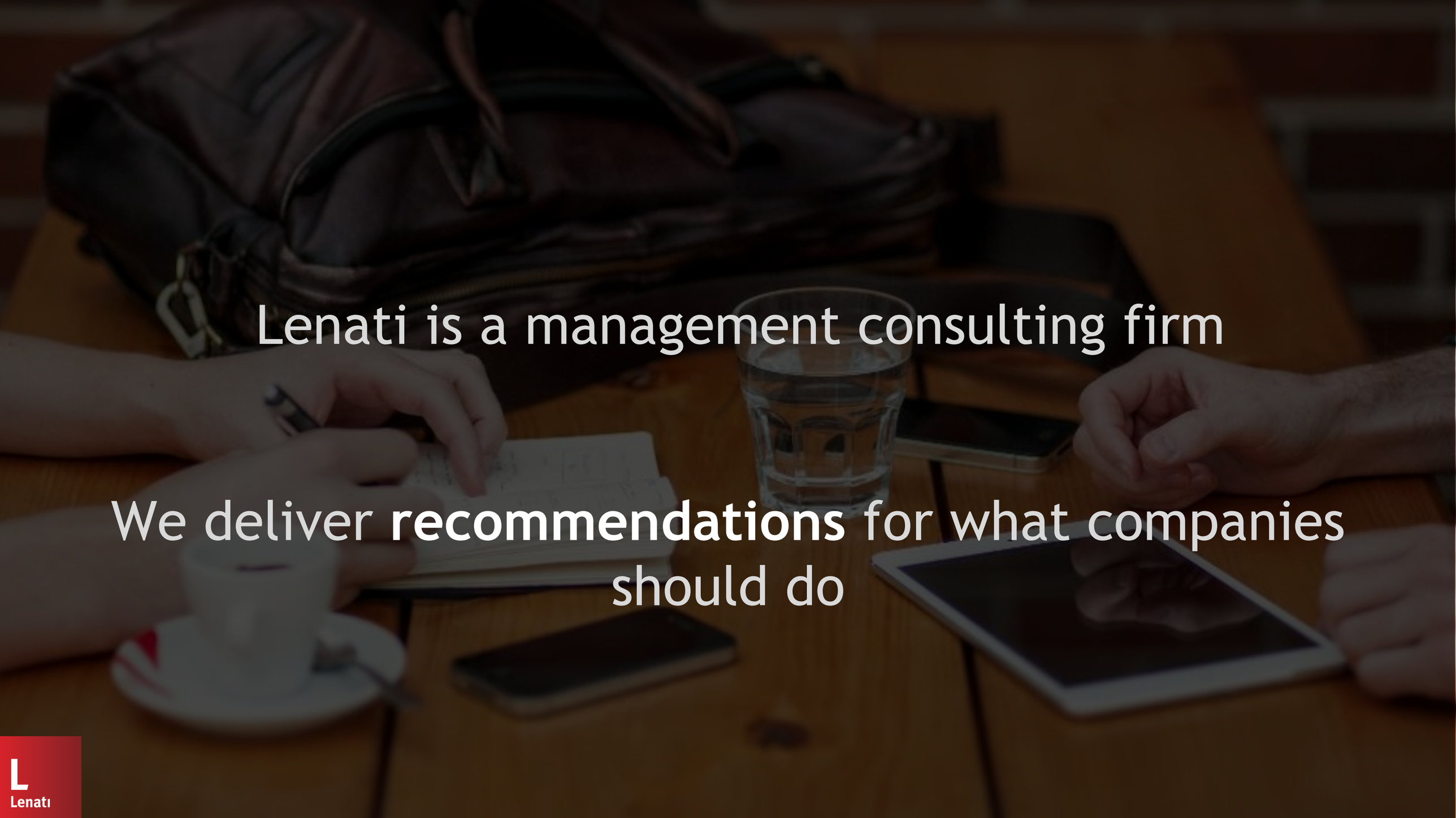
- Director of Insights and Analytics at Lenati
  - Lead a team of five data scientists
- Recovering academic
  - BS & MS in Math from Worcester Polytechnic Institute
  - PhD in Industrial Engineering from Arizona State University
- Weird hobbies
  - Won season 3 of the reality TV show King of the Nerds
  - Created the viral website, tweetmashup.com
- Learned R 5 years ago for a contracting job, never looked back



Lenati

What is this presentation?

- A walkthrough of how we use R as a team at Lenati
- A collection of the things we learned about:
  - Working as an R team
  - Working with business people
- Goal is to inspire you to think about:
  - How to work as a group of R users
  - How to work inside a company of non-data scientists


Lenati is a management consulting firm
We deliver **recommendations** for what companies should do

Lenati

Vans

Helped design Van’s loyalty program

Data questions:

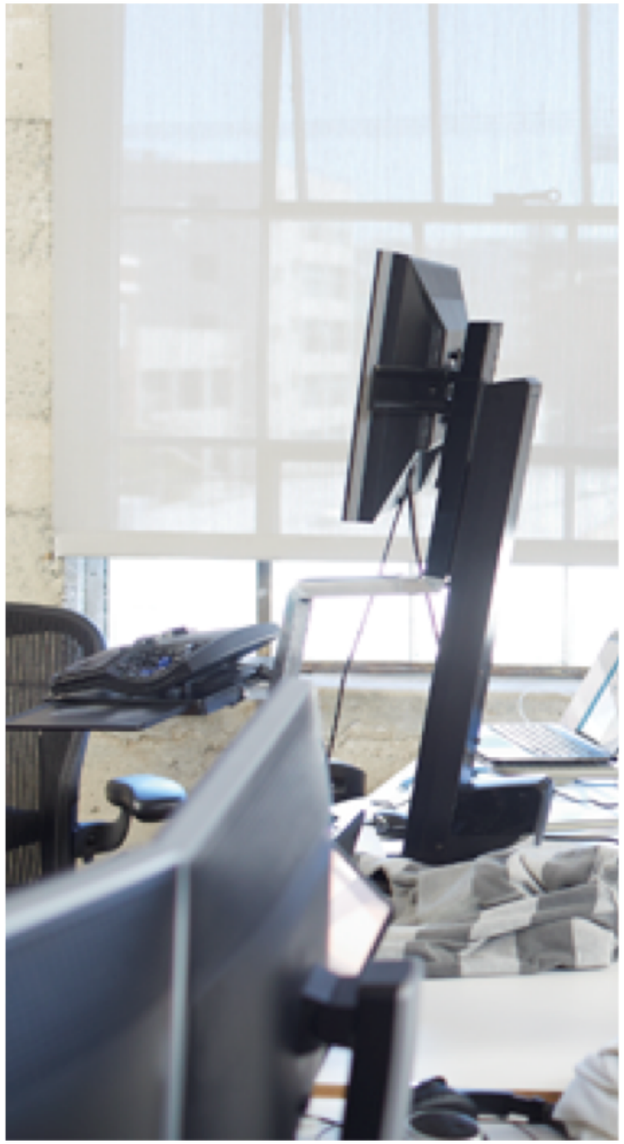
- How do you set tiers at the right level?
- How do loyal customers behave differently?
- How long until points should expire?

Mountain Equipment Co-op

Helped improve customer surveys and analytics

Data questions:

- How do different segments of customers compare in satisfaction?
- How is satisfaction related to purchasing patterns?
- How is satisfaction different for online shoppers?



Microsoft

How can we improve the Microsoft Ignite conference experience

Data questions:

- How do attendees choose what sessions to attend?
- How is satisfaction related to session attendance?
- How can we better target learning paths?

Lenati

Things we deliver:

- Recommendations and strategies (Almost exclusively in PowerPoint)

Things we do not deliver:

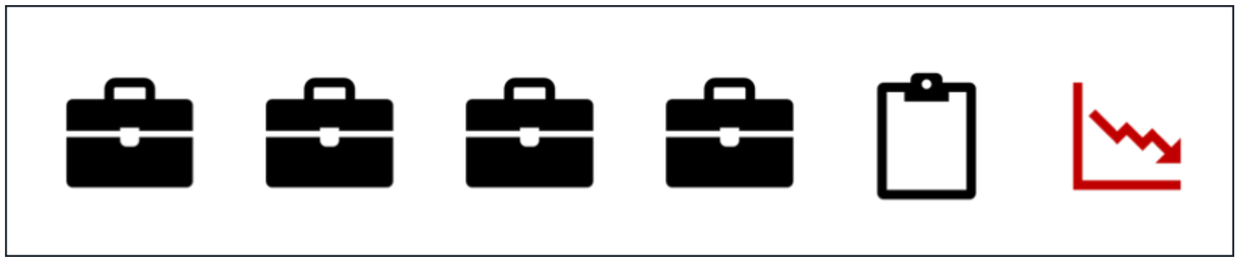
- Continuously running machine learning models
- Code bases for people outside our team to run
- Dashboards and reporting systems (mostly)

We are decision scientists not data scientists

Lenati

A Lenati project

DSW
DESIGNER SHOE WAREHOUSE®



What do we use to do it

R

- Can quickly analyze new datasets
- Code is easy to learn and read (with the tidyverse)
- Output is easy to digest (with ggplot and Rmarkdown)
- Simple to go from EDA to modeling and back
- Can connect to many different locations for data:
  - csv
  - SQL
  - Excel

SQL

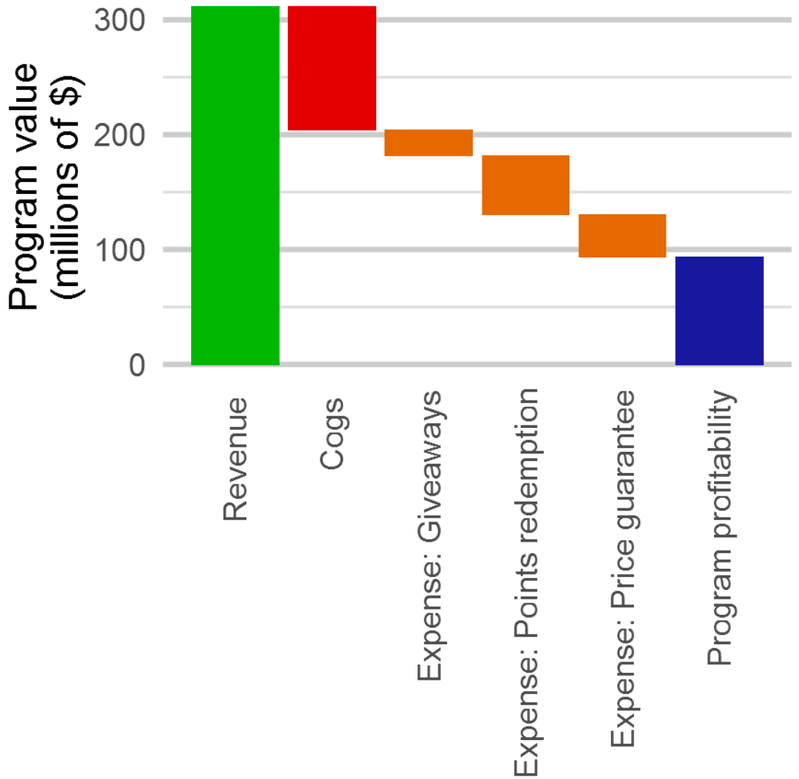
We need a place to store data that won't fit in memory

Our data objective: create a recommendation that…

1. is something we can show to business executives,
2. is quick to make,
3. allows us to trace back our steps,
4. relies on many different data sources,
5. can be re-run by someone else in six months, and
6. has minimal errors

1. Is something we can show to business executives

- Our clients are executives at Fortune 500 companies
- Need to be polished and tell a visual story
- What we use: ggplot2 for everything
- Working as a team:
  - lenati_theme() to style the charts
  - lenati_colors() to create an arbitrary length vector of colors

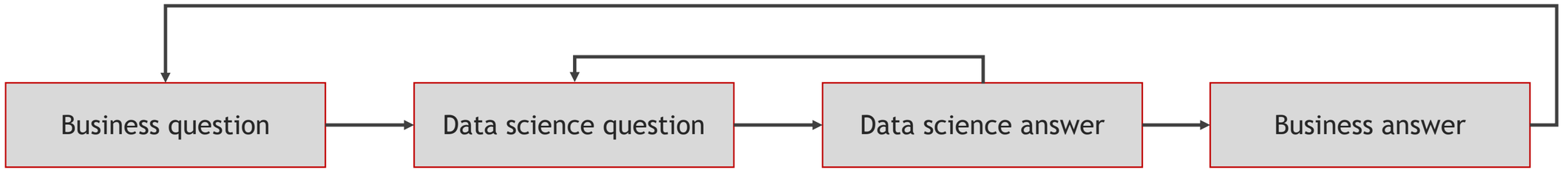


2. Is quick to make

- Questions come to use as “business questions”

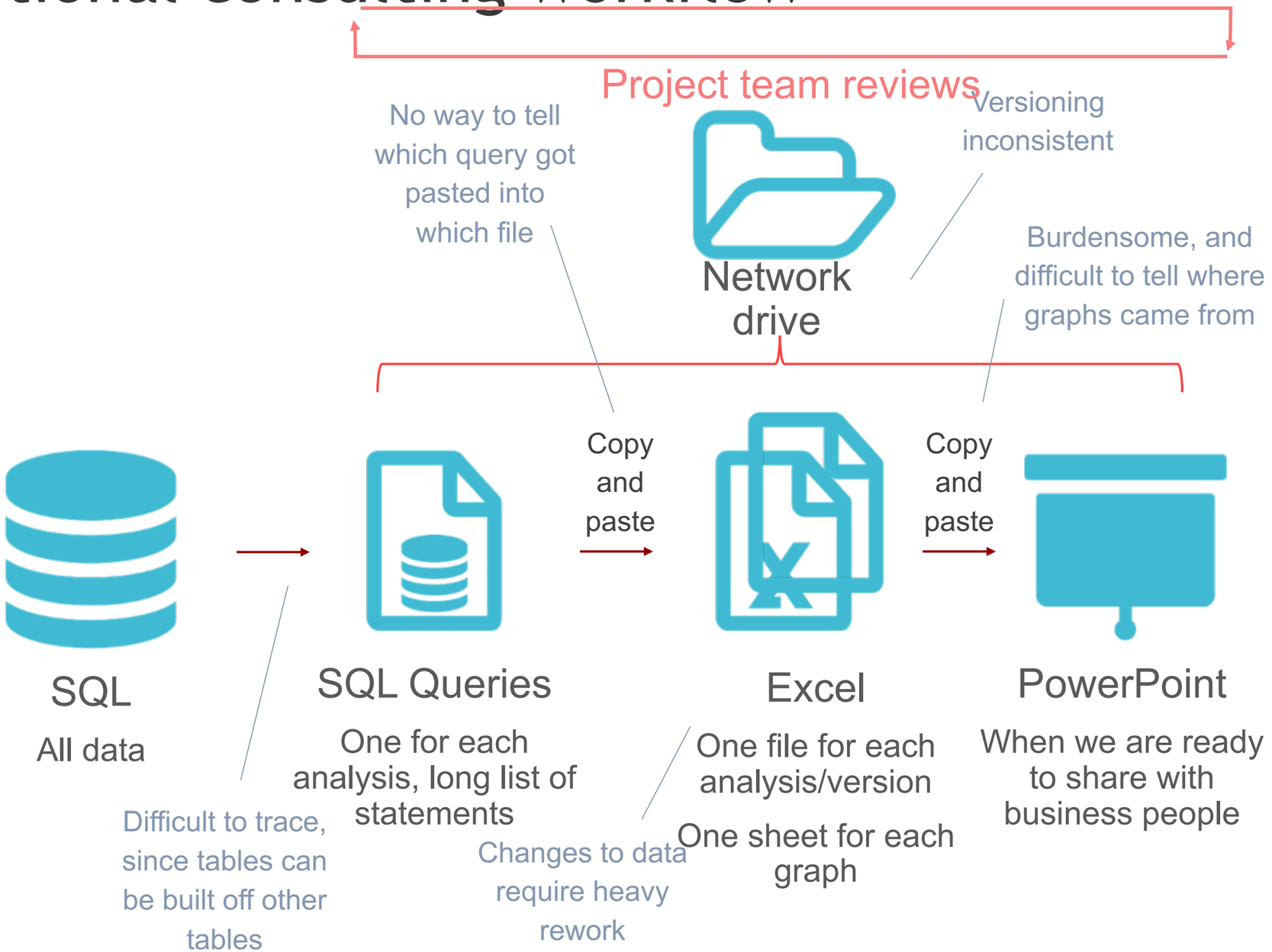
  How do loyal customers differ from regular ones?

- We need to turn that into a data science question, answer it, and turn that into a business recommendation as quickly as possible
- This results in feedback loops



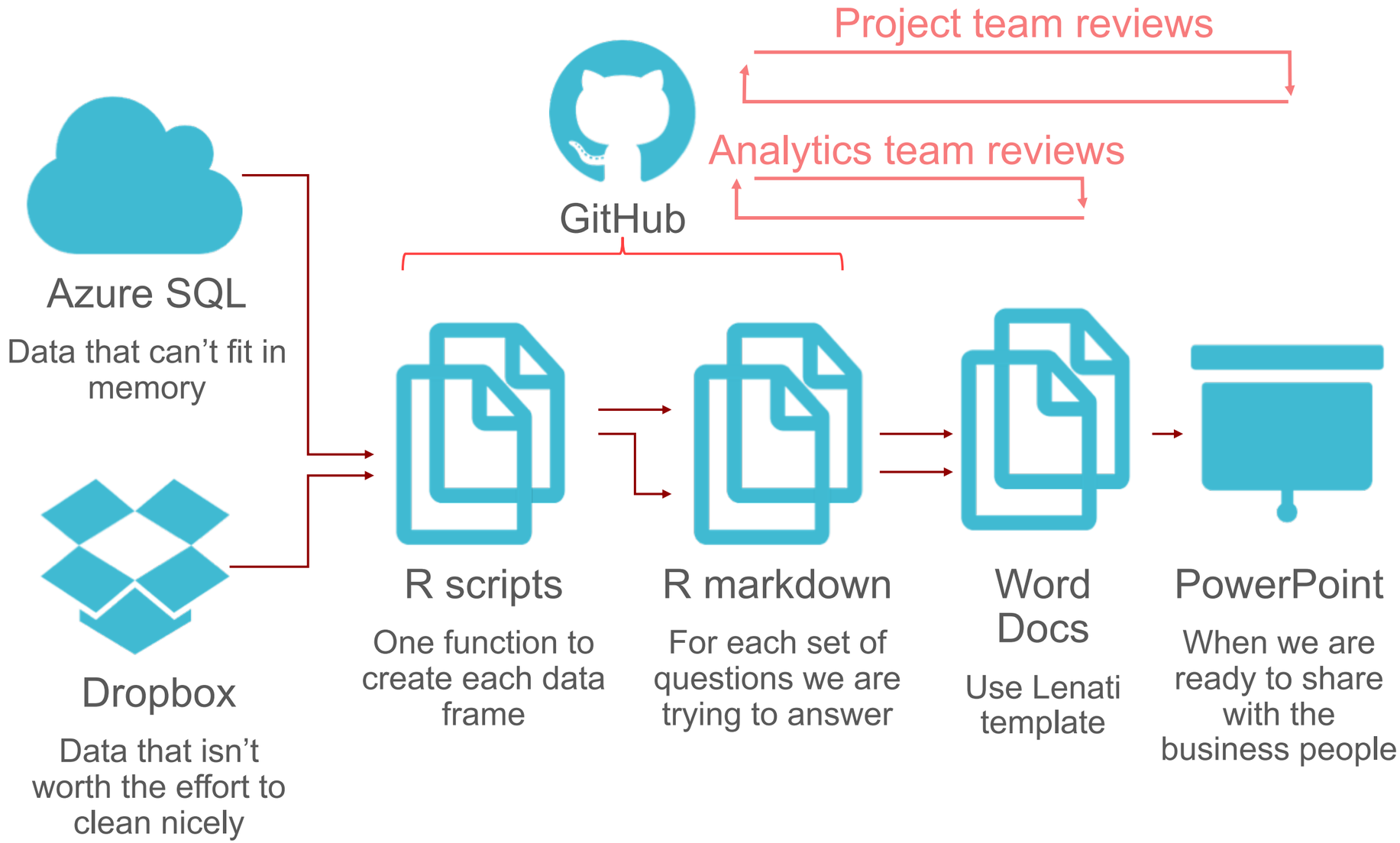
- We need to get through these loops as quickly as possible

Traditional consulting workflow



Lenati

Our approach



Lenati

2. Quick to make (continued)

- 80% of our analysis ends up being a form of “filter -> group -> aggregate”
- dplyr makes this profoundly quick to do (and easy to read)
- Working as a team: any base R (merge) or other methods (Data.Table) are strongly discouraged

```
data %>%
  left_join(spaces,by=c("BoardTitleClean"="SpaceClean")) %>%
  filter(CommunityName == "Excel") %>%
  group_by(ActionKey) %>%
  summarize(Count = n())
```

3. Allows us to trace back our steps,

- Business executives need to understand how we come to our conclusions, not just what they are
- Need to be able to explain precisely what we did at any time
- R makes this dramatically easier:
  - dplyr and tidyverse code is easy to read
  - Can trace a path back from a graph to the original data
- Statistical modeling and machine learning done in the same place we do the data manipulation
- Working as a team: we try to minimize the number of intermediate steps from data to result so others can understand

Analyses and presentations

R markdown

For each set of questions we are trying to answer

Word Docs

Use Lenati template

PowerPoint

When we are ready to share with the business people

- Create one R markdown file per analysis (and an R file to render it)
- Use the Word Docs as intermediate results you are comfortable sharing (and version them)
- Take the best from the Word Docs and put into PowerPoint when ready to share
- Working as a team: have Word Docs ready for when someone walks by your desk and asks how it’s going

4. Relies on many different data sources

- Our clients data can come from:
  - >10 gb csv files
  - Emailed Excel files with many sheets
  - Servers we have to remote desktop into
- R is great since it can connect to so many sources
- We have a strict philosophy:
  - Never edit the original data
  - Do as much as possible within R
  - Split R code and data completely

Data storage

Azure SQL

- If it’s too big for memory then it goes in Azure SQL
- Do not do any processing directly in sql
- Do not make tables that other tables depend on
- R scripts: Load sql data using DBI, odbc, dbplyr

```
con <- DBI::dbConnect(odbc::odbc(),
                      .connection_string = "...")

test <- tbl(con,"Test") %>%
  group_by(Y) %>%
  summarize(TotalX = sum(X,na.rm=T)) %>%
  filter(TotalX > 0)
```

Dropbox

- Data is stored in original format
- Don’t rely on temp data
- Working as a team: use a json config file within R project to reference where the Dropbox folder is

```
{
  "Cache": "C:/Users/Jonathan Nolis/Cache/Projects/ProjectX",
  "Data":  "C:/Users/Jonathan Nolis/Dropbox (Lenati)/ProjectX/Data",
  "Output":  "C:/Users/Jonathan Nolis/Dropbox (Lenati)/ProjectX/Output"
}
```

```
location <- function(path,folder="Data"){
  dataFolder <- fromJSON("FolderLocations.json")
  file.path(dataFolder[[folder]],path)
}
```

Data loading

Load data in a single function to:

- Load the data
- Format names
- Modify & add columns

Return a data frame or list of data frames

```
getRegistrants <- function(){
  location("Processed/RegistrantData.csv","Data") %>%
  read_csv() %>%
  namesCamelCase() %>%
  mutate(RegistrationYear = year(RegistrationDate)) %>%
  select(-AltName)
}

data <- getRegistrants()
```

5. Can be rerun in six months

- Projects are distinct – not worth it to reuse most code between projects
  - Anything we do want to reuse we put in a LenatiR package
- Any project may require a refresh at some point in the future
  - Maybe by someone else on the team
- Things we do:
  - Keep consistent folder structure between projects
  - Ensure code is documented before hand-off to different person
- Things we should do:
  - Use packrat to manage what versions of packages we use 😅

Consistent structure

GitHub

Dropbox

Azure SQL

- Scripts/
  - LoadSurveyData.R
  - LoadTransactionData.R
  - MakeSegments.R
- Reports/
  - CustomerSurvival/
    - CustomerSurvivalReport.R
    - CustomerSurvivalReport.Rmd
  - CustomerSpend/
    - CustomerSpendReport.R
    - CustomerSpendReport.Rmd
- Vignettes/
  - WeirdGraphForJane.R
- Project.Rproj
- ProjectConfig.json

- Data/
  - Original/
    - UglyExcelFile2016.xlsx
    - UglyExcelFile2017.xlsx
  - Processed/
    - UglyExcelFile2016.csv
    - UglyExcelFile2017.csv
- Output/
  - CustomerSurvival/
    - CustomerSurvivalReportV1.1.0.docx
    - CustomerSurvivalReportV1.2.0.docx
  - CustomerSpend/
    - CustomerSpendReportV1.2.0.docx
- Docs/
  - WeirdEmailDirections.docx

- org,TransactionData
- org,CustomerData
- dbo.WeirdTempTableForAnalysis

6. Has minimal errors

We do

- Show results to business experts to get feedback
- High level code reviews to make sure we are getting better at coding

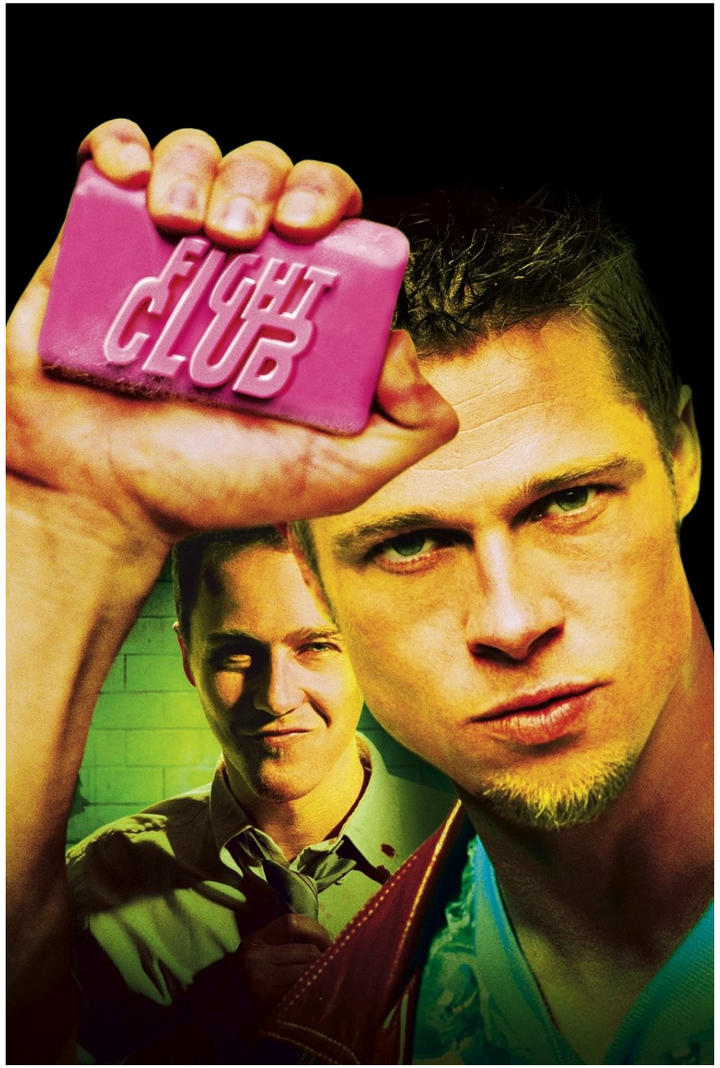
But we cut corners for time

- We don’t do unit testing 😅
  - Lack of repeatability makes this too time consuming
- We don’t do much model validation 😅
  - Usually use simple enough models that don’t need thorough cross-validation etc.
- We don’t consider statistical significance 😅
  - Clients don’t care

Working as a team

R team rules

1. Everyone has to learn R
2. R is the default solution
3. Anyone should be able to understand and run anyone elses' code



Everyone has to learn R

- Don’t need to know R to be hired
- Interview process has a case study based on anonymized project data
  - Any tool can be used in the interview process
- Onboarding new R users done with same case study data
- Given a mentor
- Zero to:
  - handling basic tasks – 2 weeks
  - working independently – 6 weeks
  - at the level of the rest of the team – 12 weeks

The essentials (in our order):

- R vs Rstudio vs Rproj vs .R
- dfs vs lists vs vectors
- dplyr & piping
- ggplot2
- knitr/rmarkdown
- making your own functions
- connecting to SQL

R is the default solution

- Assume that a project will be done in R
  - With Rstudio (no Jupyter or Visual Studio)
- Any deviation requires a good reason
  - Power BI – need a dashboard and Shiny would be overengineering
  - Python – code will be handed off to client team
  - F# - we need something that runs faster and we're weirdos
- "R isn't my language of choice" is not a good reason
- Working as a team: keep languages either "company endorsed" (R, SQL, Power BI) and minimize use of anything else

Anyone should be able to understand and run anyone elses' code

- "understand" – given enough time, can figure out what each chunk is doing
- "run" – can regenerate any report using data in SQL/dropbox and git code
  - Can't rely on temp data existing
  - Needs a clear master script which runs all other scripts
- "anyone else" – our team is cross trained so we have no specialists (although everyone has their own strengths)

Morale building

- We partner with UW Tacoma to give undergrad UW Tacoma students projects to work on in R
- Write articles and allow for non-billable side projects
- Have an offsite every six months where we talk about personal growth and get vulnerable

About me

- Jonathan Nolis
- Director of Insights and Analytics
- Leader in the capabilities group
  - Helps find new projects
  - Provides technical guidance to the practice
  - Provides strategic insights and analytics guidance to the firm

Goals

- 1 year goal: learn how to manage up
- 2-4 year goal:
  - Want to be a nationally recognized expert in running analytics teams within companies
  - Make something cool

Strengths

- Very fast learner
- Compassionate
- Enthusiastic
- Good speaker

Opportunities

- I am easily panicked
- I could be much better at relationship building
- I have trouble giving up control
- I get distracted by unimportant things

Our ongoing struggles

Balancing fun with results

- R is really good at letting you do lots of things:
  - Explore the data from many angles
  - Apply state of the art techniques
- Interesting to analyze != interesting to client
- Working as a team:
  - Create an analysis plan early on in the project
  - Use the most simple approach possible on a project
  - Use the state of the art when not being billed

2017 Ignite Analysis Plan

Overarching Goals:

1. Provide comprehensive analysis of Ignite, identifying key patterns in attendee activity that lead to actionable insight.
2. Create an Ignite guide that includes recommended next steps, factors that drive repeat attendance, and how to improve attendee targeting.

Overview

Provide a high-level overview of Ignite attendance metrics, and introduce most informative insights.

Questions/Analyses:

1. Ignite-by-the-numbers: metrics that include number of attendees (total, as well as across BGs, content categories, etc.), number of sessions, session size and other summary calculations
2. Summary of the most actionable insights from the categories below to help drive next steps (predictive model, most popular clusters/sessions/learning paths)
3. Consolidate key points into a “how we did” document for quick consumption after Ignite

Clustering

Use k-means clustering to determine attendee cluster based on their session attendance. We can then use these clusters to pivot other metrics (satisfaction, companies, learning paths, etc.)

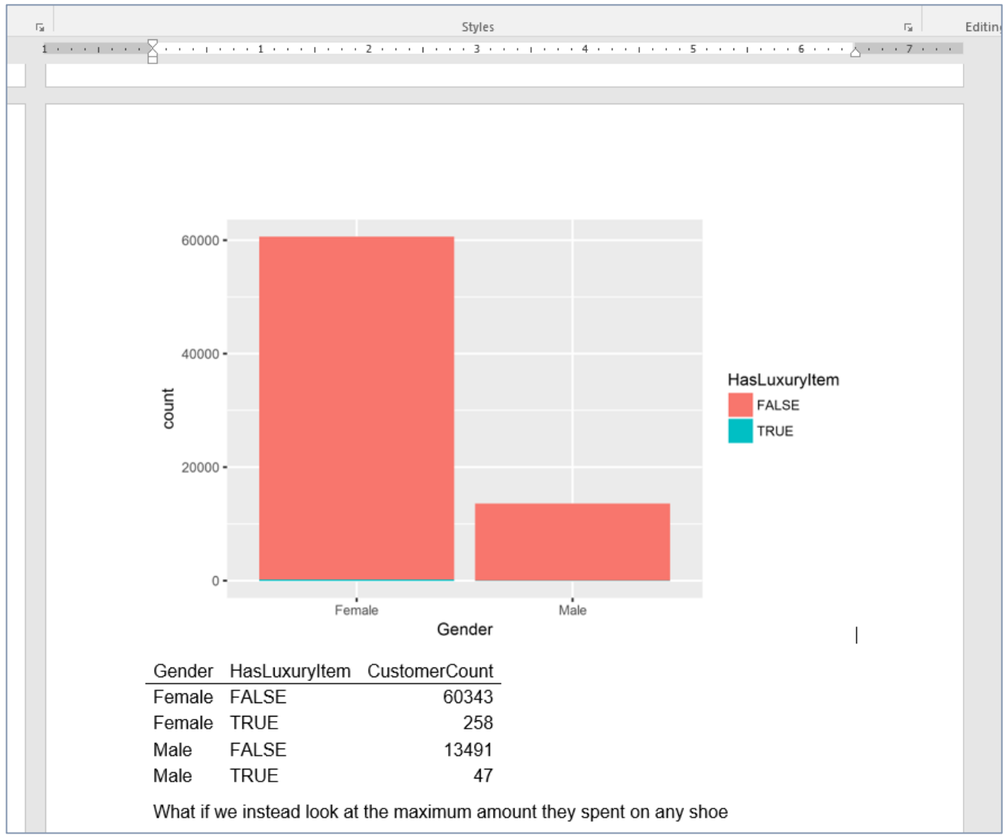
Questions/Analyses (for these analyses, compare to 2016 findings):

1. Conduct cluster analysis on session attendance to assign a distinct cluster to each Ignite attendee
2. Determine cluster distribution across industry, roles, company size and attendee demographics
3. For each cluster, determine what % of session attendance was to sessions in that cluster

Source: Jessica Locke

Allowing edits from non-data scientists

- Current process: data -> R -> ggplot image
- Sometimes people want to edit the image
  - Change labels, colors, formatting
  - Access the exact numbers to use somewhere else
- We don’t have an easy answer
  - Whenever possible, try and use kable() to export the data so we can at least send that



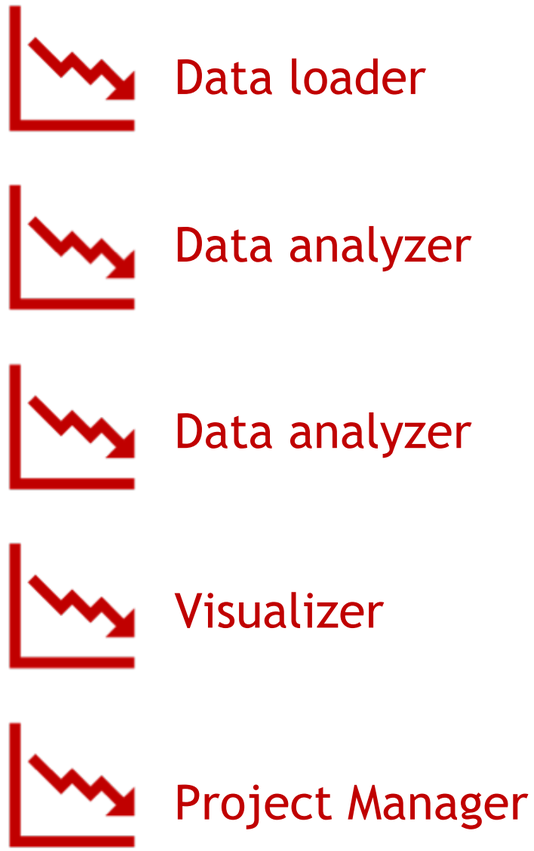
Gender	HasLuxuryItem	CustomerCount
Female	FALSE	60343
Female	TRUE	258
Male	FALSE	13491
Male	TRUE	47

What if we instead look at the maximum amount they spent on any shoe

Team specialization

- Current R team, everyone can do everything
  - Load data from client
  - Analyze data
  - Create visuals
  - Create narrative
  - Answer the business question

  Especially important since each person working independently on a project
- Not everyone likes doing everything
- Open question: should we allow specialization?



Wrapping it up

Conclusions

- R is a great tool for our particular line of work
  - Quick EDA
  - Meaningful modeling
  - Beautiful output
  - Easy to learn
- Our business causes us to write code in different ways
  - Easy for others to use and understand
  - Not necessary to consistently run in production
- Orienting our team around R has been great to manage