ISA 401: Business Intelligence & Data Visualization

01: Introduction to BI and Data Viz

Fadel M. Megahed, PhD

Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University

Twitter: FadelMegahed
GitHub: fmegahed
Email: fmegahed@miamioh.edu
Office Hours: Automated Scheduler for Virtual Office Hours

Spring 2022

Learning Objectives for Today's Class

- Describe **course objectives** and **structure**.
- Define **data visualization** and describe its **main goals**.
- Describe the **BI methodology** and its **major concepts**.

Course Design, Expectations, and Overview

The Analytics Journey: Pre-Analytics [1]

- **Pre-Analytics/Data Management:** where one attempts to **extract** the needed *data* for analysis. Data can either be:

- Stale, uninteresting, convenient
- Highly processed and archived
- Example: `iris`, `mtcars`, `titanic`

- Fresh, interesting, challenging
- Impactful
- Examples: Cincinnati Open Data Portal, Ohio Data Portal, US Government's Open Data.

While the highly processed data can be useful in learning basic concepts, **real-world (often messy)** data real are much interesting to work with -- **e.g., we can make useful & meaningful decision from the data.** In this class, we will learn how to scrape, extract and clean messy data in addition to visualizing clean[ed] data.

Source: Slide inspired by Kia Ora's What I mean by "data".

The Analytics Journey: Pre-Analytics [2]

05:00

Non-Graded Class Activity # 1

> *Take 5 minutes to discuss with your partner*

Activity | Your Solution | Fadel's Approach (No Solution Shown)

- Go to https://data.cincinnati-oh.gov/Safety/Traffic-Crash-Reports-CPD-/rvmt-pkmq/data
- Download the data utilizing the export column and answer the following questions:
  - How many **observations/rows** and **columns** do we have in the dataset?
  - How many **crashes** are reported in the dataset?

The Analytics Journey: Pre-Analytics [2]

05:00

Non-Graded Class Activity # 1

> *Take 5 minutes to discuss with your partner*

Activity | **Your Solution** | Fadel's Approach (No Solution Shown)

- Insert your solution here (Use Chrome as your browser to edit this part of the page)

The Analytics Journey: Pre-Analytics [2]

05:00

Non-Graded Class Activity # 1

> *Take 5 minutes to discuss with your partner*

Activity | Your Solution | Fadel's Approach (No Solution Shown)

```
if(require(tidyverse) == FALSE) install.packages("tidyverse")

# Link obtained from site -> Export -> "Right Click on" CSV
crashes = readr::read_csv("https://data.cincinnati-oh.gov/api/views/rvmt-pkmq/rows.csv?acces

# Number of rows and columns
nrow(crashes)
ncol(crashes)
# Or alternatively
dim(crashes)

# Total number of crashes
# Will be discussed in class in greater detail
```

The Analytics Journey: Descriptive [1]

Descriptive Analytics: where one attempts to **understand** the data through **descriptive statistics** and **visualizations**.

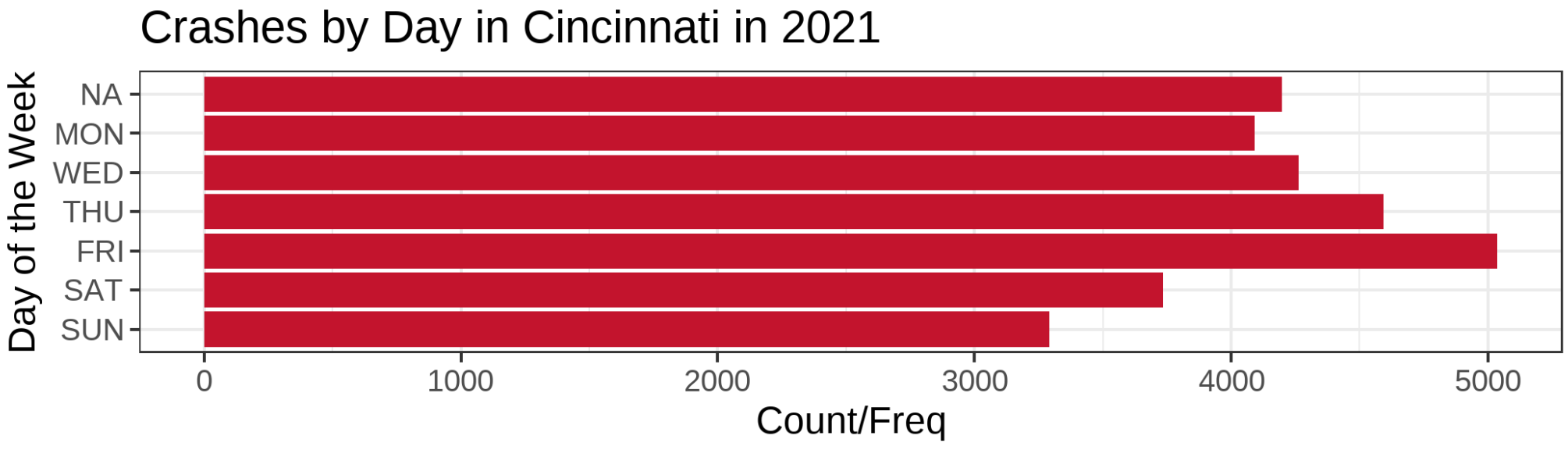
Descriptive Statistics for 2 Categorical Variables

```
## $dayofweek
##
##  FRI  SAT  TUE  MON  WED  SUN  THU
## 5037 3735 4198 4093 4264 3291 4594
##
## $weather
##
##                              1 - CLEAR                                    4 - RAIN
##                                  19904                                        3771
##                             2 - CLOUDY                      99 - OTHER/UNKNOWN
##                                   4550                                         267
##                               6 - SNOW                    3 - FOG, SMOG, SMOKE
##                                    638                                          18
##    8 - BLOWING SAND, SOIL, DIRT, SNOW 9 - FREEZING RAIN OR FREEZING DRIZZLE
##                                      2                                          20
##                         5 - SLEET, HAIL                 7 - SEVERE CROSSWINDS
##                                     41                                           2
```

The Analytics Journey: Descriptive [2]

Descriptive Analytics: where one attempts to **understand** the data through **descriptive statistics** and **visualizations**.

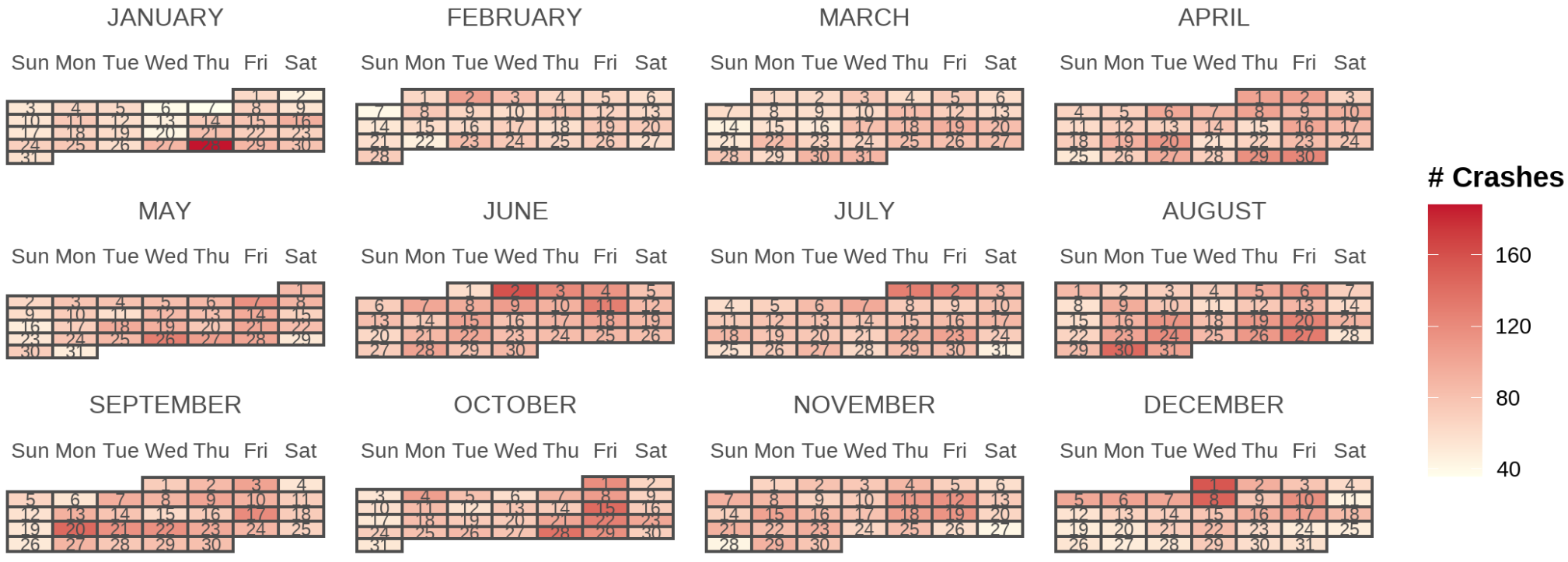
A Simple Visualization - A Bar Chart of Crashes Per Day



Created by: Fadel Megahed | Data source: City of Cincy Open Data Portal (rvmt-pkmq)

The Analytics Journey: Descriptive [3]

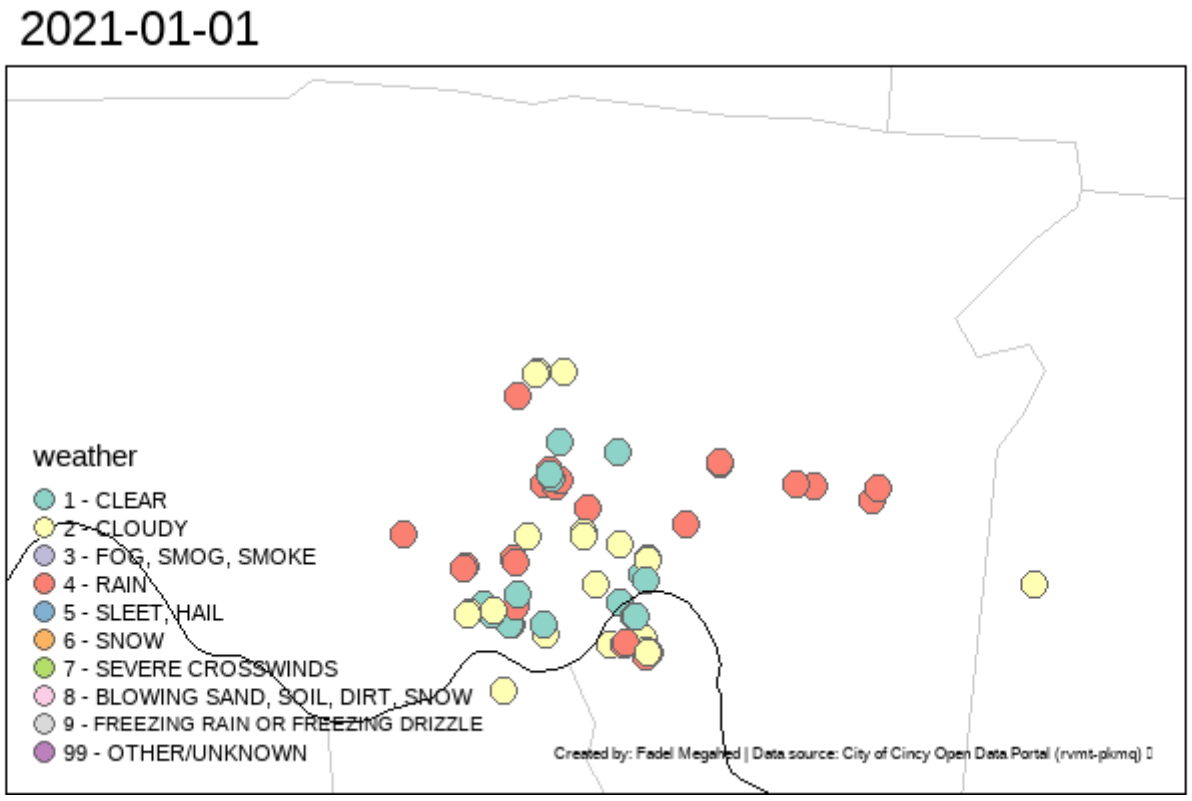
Descriptive Analytics: where one attempts to **understand** the data through **descriptive statistics** and **visualizations**.



Created by: Fadel Megahed | Data source: City of Cincy Open Data Portal (rvmt-pkmq)

The Analytics Journey: Descriptive [4]

Descriptive Analytics: where one attempts to **understand** the data through **descriptive statistics** and **visualizations**.



The Analytics Journey: Descriptive [5]

03:00

Activity | Q1 Solution | Q2 Solution

- How do the previous two graphs complement each other?
- If you were to pick one of the two charts, which one is more informative?
- You will be asked to write down your answers on in the next two panels.

The Analytics Journey: Descriptive [5]

03:00

Activity | Q1 Solution | Q2 Solution

Go to **www.menti.com/spcmam4f23**

How the two graphs complement each other?

Mentimeter

The Analytics Journey: Descriptive [5]

03:00

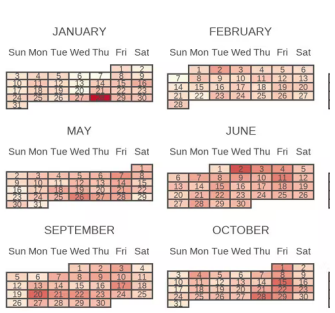
Activity | Q1 Solution | Q2 Solution

Go to **www.menti.com/z15vty41xg**

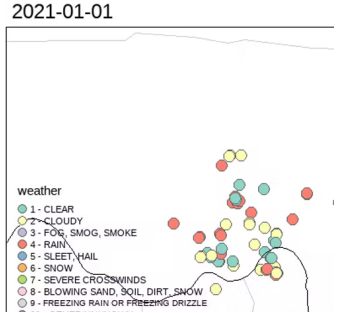
Which one of the charts is more informative?

Mentimeter

0



0



The Analytics Journey: Predictive [1]

Predictive Analytics: where **statistical** and **machine learning** models are used to help us utilize independent variable[s] to predict an outcome variable of choice.

- **Many** consider this component to be the  aspect of the analytics journey.
- IMO, this is not always true, but your success in this stage is **hinged on**:
  - **Correct** ✅ data, i.e.,
    - *Do you actually capture the important predictors?*
    - *Is your data aggregated to the right level?*
  - **Cleaned**  data, i.e.,
    - *Is your data tidy?*
    - *Is your data technically correct?*
    - *Is your data consistent?*

The Analytics Journey: Predictive [2]

Predictive Analytics: where **statistical** and **machine learning** models are used to help us utilize independent variable[s] to predict an outcome variable of choice.

- With the aforementioned constraints/setup, now you can explore how to model the data using statistical and machine learning models?
- **Some recommendations:**
  - Start with the simplest (which is also often the most easy-to-explain) model first.
  - If you are happy with the predictive performance (i.e., no gains would be of practical benefit), you are done 👏.
  - If not, ↩️ and try other models.

The Analytics Journey: Predictive [3]

Predictive Analytics: where **statistical** and **machine learning** models are used to help us utilize indepedent variable[s] to predict an outcome variable.

Transportation Research Part C 126 (2021) 103016

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

The association between crashes and safety-critical events: Synthesized evidence from crash reports and naturalistic driving data among commercial truck drivers

Miao Cai [a], Mohammad Ali Alamdar Yazdi [b], Amir Mehdizadeh [c], Qiong Hu [c], Alexander Vinel [c], Karen Davis [d], Hong Xian [e], Fadel M. Megahed [f], Steven E. Rigdon [e,*]

[a] Department of Epidemiology, School of Public Health, Sun Yat-sen University, Guangzhou, Guangdong 510080, China
[b] Carey Business School, Johns Hopkins University, Baltimore, MD 21218, United States
[c] Department of Industrial and Systems Engineering, Auburn University, Auburn, AL 36849, United States
[d] Department of Computer Science and Software Engineering, Miami University, Oxford, OH 45056, United States
[e] Department of Epidemiology and Biostatistics, Saint Louis University, Saint Louis, MO 63103, United States
[f] Farmer School of Business, Miami University, Oxford, OH 45056, United States

ARTICLE INFO

Keywords:
Truck
Naturalistic driving studies
Safety-critical events
Crashes
Injuries
Fatalities

ABSTRACT

The past decade has witnessed continuous growth of naturalistic driving studies (NDSs). When analysing NDS data, safety-critical events (SCEs) are commonly used as surrogates for safety, since actual crashes are very rare. However, the association between SCEs and crashes is not consistent in previous studies and has not been verified among commercial truck drivers. Based on routinely collected kinematic data from 31,828 truck drivers in a large commercial trucking company, this paper examines the association between four types of SCEs (headway, hard brake, collision mitigation, and rolling stability) and crashes, as well as injuries and fatalities. Compared to existing studies on the subject, that are based on up to about 2 million miles driven, our study involves an estimated 2.3 billion miles driven. Bayesian negative binomial models were applied to examine the association between three outcomes (crashes, injuries, and fatalities) and the four SCEs. It was found that a unit increase in the number of any type of SCEs per 10,000 miles was

Accident Analysis and Prevention 159 (2021) 106285

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Predicting unsafe driving risk among commercial truck drivers using machine learning: Lessons learned from the surveillance of 20 million driving miles

Amir Mehdizadeh [a], Mohammad Ali Alamdar Yazdi [b], Miao Cai [c], Qiong Hu [a], Alexander Vinel [a], Steven E. Rigdon [c], Karen Davis [d], Fadel M. Megahed [e,*]

[a] Department of Industrial and Systems Engineering, Auburn University, Auburn, AL 36849, USA
[b] Carey Business School, Johns Hopkins University, Baltimore, MD 21218, USA
[c] Department of Epidemiology and Biostatistics, Saint Louis University, Saint Louis, MO 63104, USA
[d] Department of Computer Science and Software Engineering, Miami University, Oxford, OH 45056, USA
[e] Farmer School of Business, Miami University, Oxford, OH 45056, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Big data analytics
Crash risk prediction
Naturalistic driving
Near crashes
Safety critical events

ABSTRACT

The emergence of sensor-based *Internet of Things* (IoT) monitoring technologies have paved the way for conducting large-scale naturalistic driving studies, where continuous kinematic driver-based data are generated, capturing crash/near-crash safety critical events (SCEs) and their precursors. However, it is unknown whether the SCEs risk can be predicted to inform driver decisions in the medium term (e.g., hours ahead) since the literature has focused on SCE predictions either for a given road segment or for automated breaking applications, i.e., immediately before the event. In this paper, we examine the SCE data generated from 20+ million miles-driven by 496 commercial truck drivers to address three main questions. First, whether SCEs can be predicted using disparate driving-related data sources. Second, if so, what the relative importance of the different predictors examined is. Third, whether the prediction models can be generalized to new drivers and future time periods. We show that SCEs can be predicted 30 min in advance, using machine learning techniques and dependent variables capturing the driver's characteristics, weather conditions, and day/time categories, where an area under the curve (AUC) up to 76% can be achieved. Moreover, the predictive performance remains relatively stable when tested on new (i.e., not in the training set) drivers and a future two-month time period. Our results can inform dispatching and routing applications, and lead to the development of technological interventions to improve driver safety.

The Analytics Journey: Prescriptive [1]

Prescriptive Analytics: where **mathematical models** are used to make recommendations for business actions.

- Our **overarching goal** behind data/business analytics, is to **make informed decisions based on what we have learned from the data**. Hence, this stage is where we build on what we learned during the *descriptive* and *predictive* stages to make more informed decisions.
- Imagine that you are a large trucking company (e.g., Amazon, Fedex, JB Hunt), and you have models that show **both**:
  - Safety critical events that are associated with crashes.
  - The occurrence of safety critical events can be reasonably predicted as a function of: (a) driver characteristics, (b) weather conditions, and (c) Traffic conditions.
- **As a business analyst, what two reasonable questions would you attempt to approach/optimize for?**

The Analytics Journey: Prescriptive [2]

03:00

Prescriptive Analytics: where **mathematical models** are used to make recommendations for business actions.

Non-Graded Class Activity # 3

> *Take 3 minutes to formulate the two best questions with your partner*

Activity | Your Solution | Our Work in this Area

As a business analyst, what two reasonable questions would you attempt to approach/optimize for?

The Analytics Journey: Prescriptive [2]

03:00

Prescriptive Analytics: where **mathematical models** are used to make recommendations for business actions.

Non-Graded Class Activity # 3

> *Take 3 minutes to formulate the two best questions with your partner*

Activity | Your Solution | Our Work in this Area

- Insert your solution here (Use Chrome as your browser to edit this part of the page)

The Analytics Journey: Prescriptive [2]

03:00

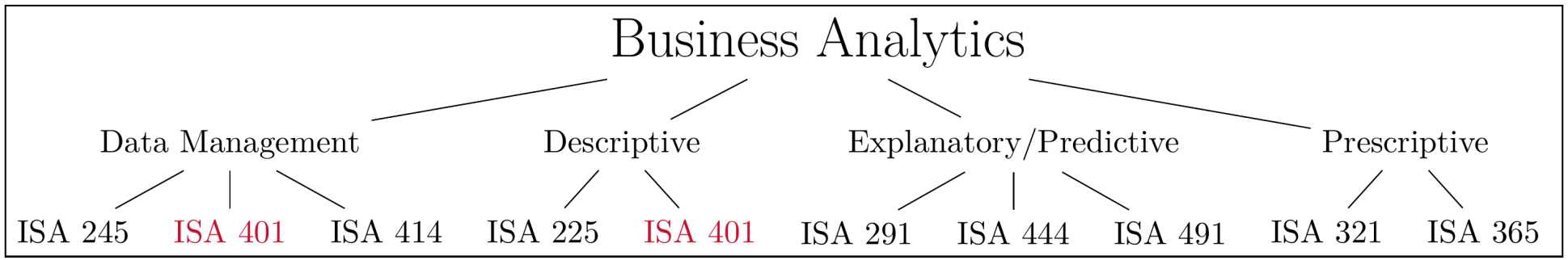
Prescriptive Analytics: where **mathematical models** are used to make recommendations for business actions.

Non-Graded Class Activity # 3

> *Take 3 minutes to formulate the two best questions with your partner*

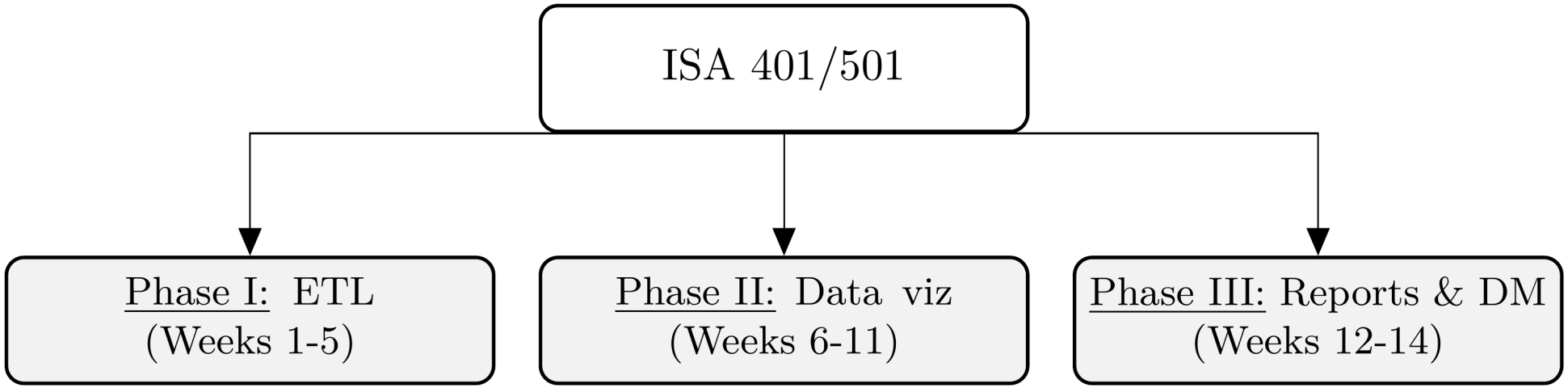
Activity | Your Solution | Our Work in this Area

How does our Curriculum at Miami University Prepare you for this Journey?



My take on the courses within the business analytics major/minor at Miami University

ISA 401/501 Course: An Overview



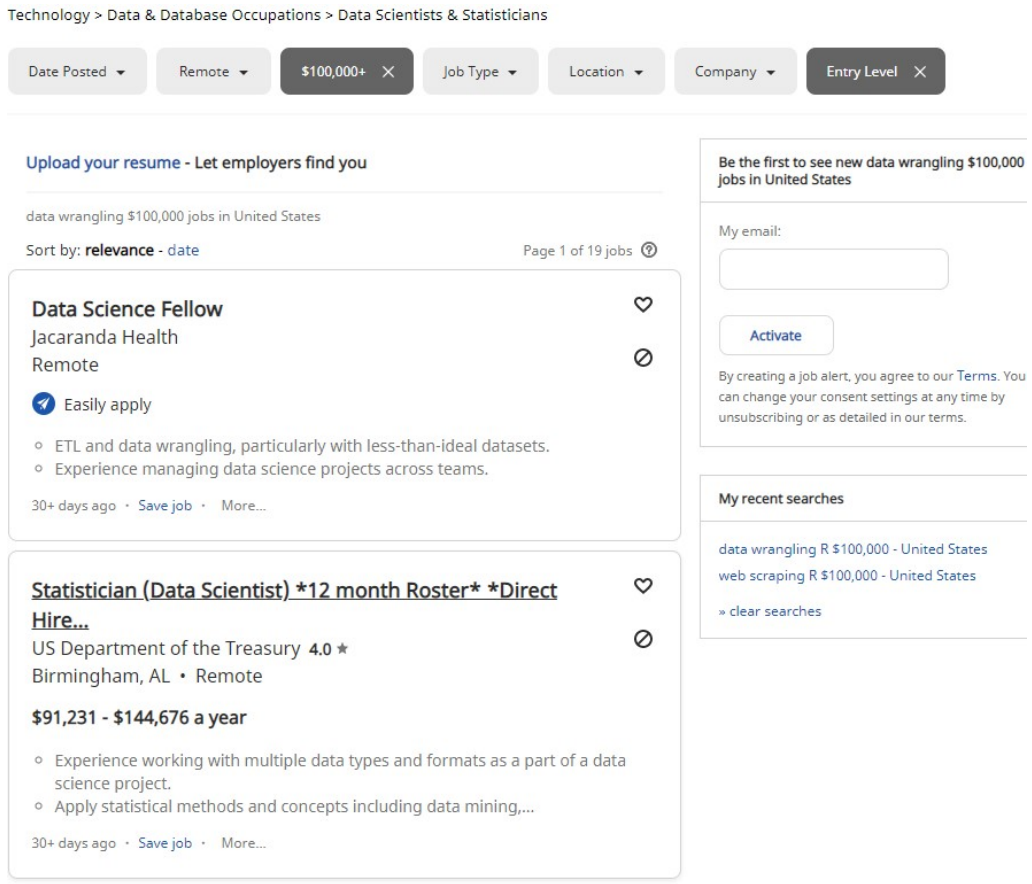
How the ISA 401/501 course is organized.

ISA 401/501 Course Objectives

Even though software will be extensively used, this is not a software class. **Instead, the focus is on understanding the underlying methods and mindset of how data should be approached.**

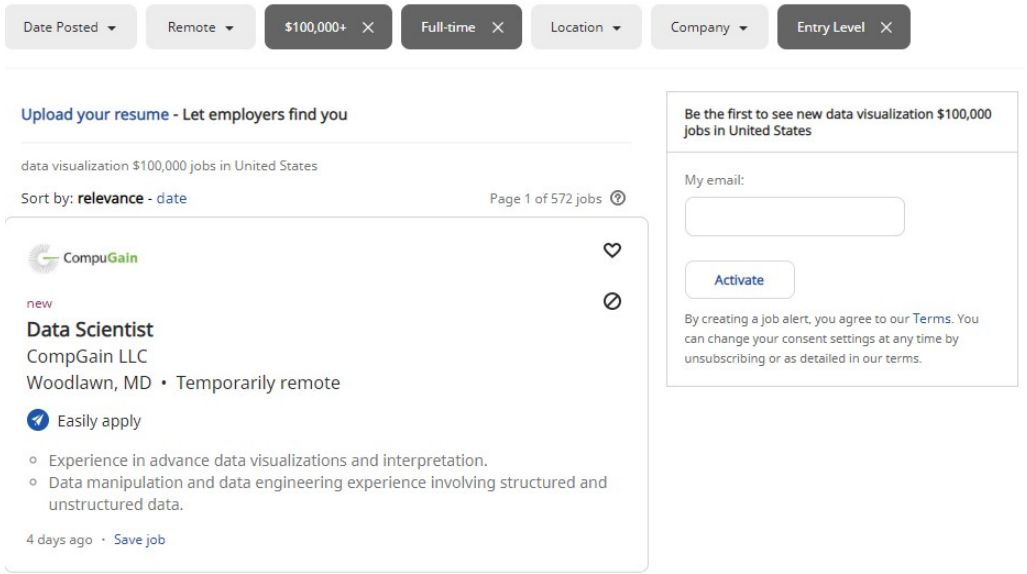
- Be capable of extracting, transforming and loading (ETL) data using multiple platforms (e.g. R, Power BI and/or Tableau).
- Write basic R scripts to preprocess and clean the data.
- Explore the data using visualization approaches that are based on sound human factors.
- Understand how statistical/machine learning can capitalize on the insights generated from the data visualization process.
- Create interactive dashboards that can be used for business decision making, reporting and/or performance management.
- Be able to apply the skills from this class in your future career.

Should you Care? Indeed [1]



Entry-Level Data Wrangling Jobs on Indeed.com

Should you Care? Indeed [2]



Entry-Level Data Visualization Jobs on Indeed.com

Should you Care? Read this Job Ad

04:00

When I have designed this course, I have incorporated a lot of feedback from **industry collaborators, peer/leading academic programs, and state-of-the-art-research advancements.** Thus, this is meant to be a hands-on, practically-relevant course.

Non-Graded Class Activity # 4

Activity | Documentation Space

To demonstrate the practicality of this course, let us consider this job ad.

- Please open the Data Scientist (6257U) - CED Data Scientist position at UC - Berkeley by clicking here.
- Compare the **responsibilities** and the **required qualifications** with the course objectives.
- Read through the required qualifications.
- **Document what you will learn in this course to make you more competitive.**

Should you Care? Read this Job Ad

04:00

When I have designed this course, I have incorporated a lot of feedback from **industry collaborators, peer/leading academic programs, and state-of-the-art-research advancements.** Thus, this is meant to be a hands-on, practically-relevant course.

Non-Graded Class Activity # 4

Activity | Documentation Space

- Insert your solution here (Use Chrome as your browser to edit this part of the page)

Should you Care? Recent Alumni Testimonials

Hi Fadel,

I hope you had a nice summer and are having a great start to the school year! In a bunch of my trainings at Deloitte we have gone over data visualization and wow, I am definitely ahead of the curve. I am also using R every day on my project. So, thank you! Your 401 class prepared me perfectly for the start of my career!

I am reaching out on behalf of Deloitte Consulting's Miami Recruiting team. We are recruiting Miami students this fall for Full-Time and Summer Scholar (internship) positions. The attached page outlines the two full-time roles we are hiring for: Business Analysts (Strategy & Operations competency) and Business Technology Analysts (Technology competency). You should also be receiving additional information via a newsletter from your department chair shortly as well.

If you are interested in learning more about Deloitte Consulting and opportunities for your students, please reach out to our Miami Recruiting Campus Leads, Erica Guidobono or Haley Johnson, they would be happy to connect with you.

Thank you!

Hi Fadel,

I'm not sure if you remember me but I took a couple of your classes (Data Visualization and Optimization) a few years ago.

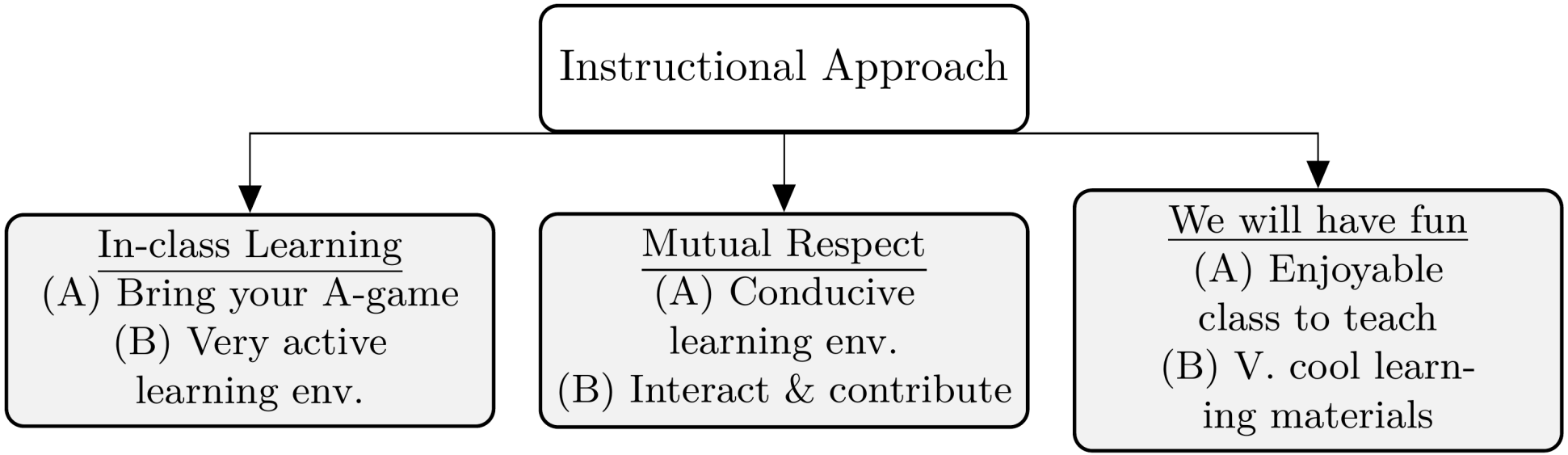
I'm reaching out because I wanted to express how grateful I am that I took your classes and how impactful they have been since I graduated.

When I graduated I got a job as an IT Auditor with EY. I didn't expect to use what I learned in my analytics classes in my job, but consistently throughout my first year here, it has been my analytics background that has set me apart from so many of my peers. The ability to code, to understand data, and to speak confidently about analytics has helped me form close relationships with partners in the firm, helped me to get staffed on projects that nobody in my start class has had the opportunity to get staffed on, and has helped to get my name out there in the firm faster than I could have ever imagined. For example, I was recently staffed on a project using a language I have never touched before to test data I have never seen before and the firm is confident I can handle it because of my experiences at Miami and the results that I've produced so far using that education.

I owe a lot of this to you. You were by far my favorite analytics professor and I find the way that you taught to be the most applicable to the working world.

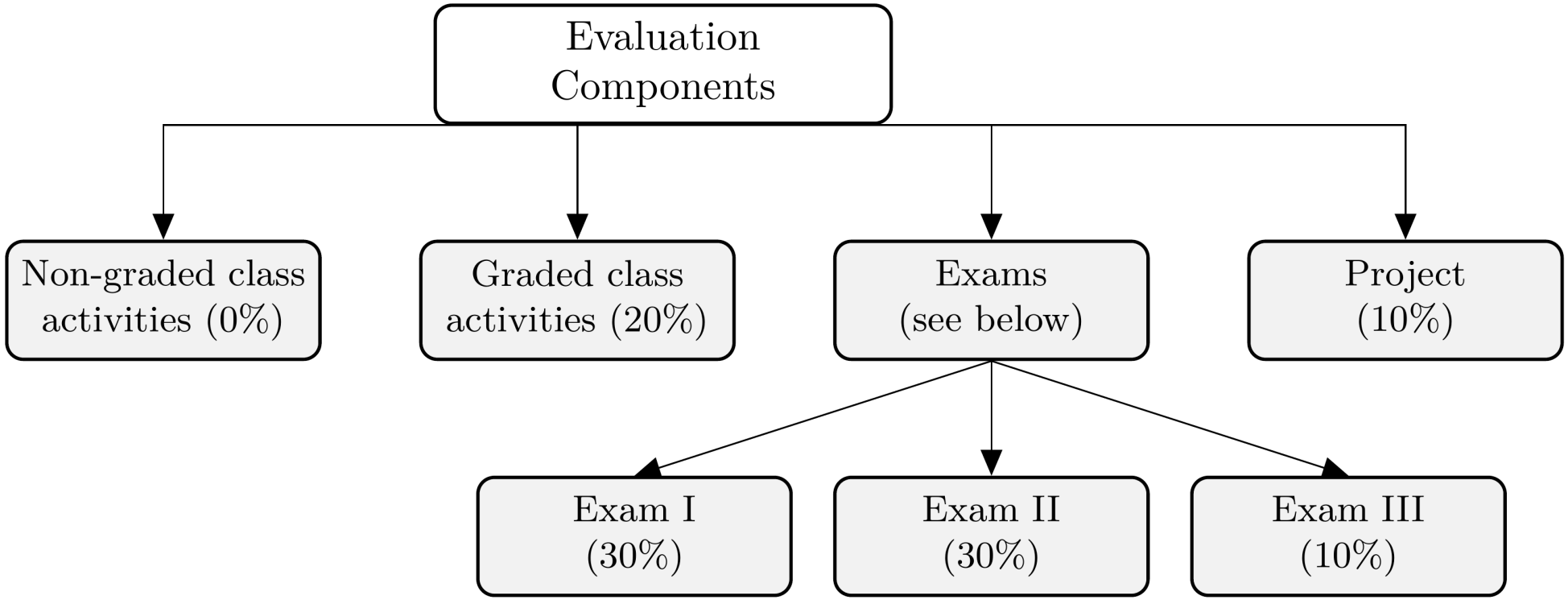
Really I just want to say thank you. Looking back on my time at Miami, you made a huge impact on my time there and now my time at work too. Your classes kept me interested and motivated to learn about analytics and now I see those dividends being paid off. I feel very blessed to have had you as my educator.

Instructional Approach



An overview of the instructional approach for ISA 401/501.

How will I Evaluate your Learning?



An overview of the evaluation components for ISA 401/501.

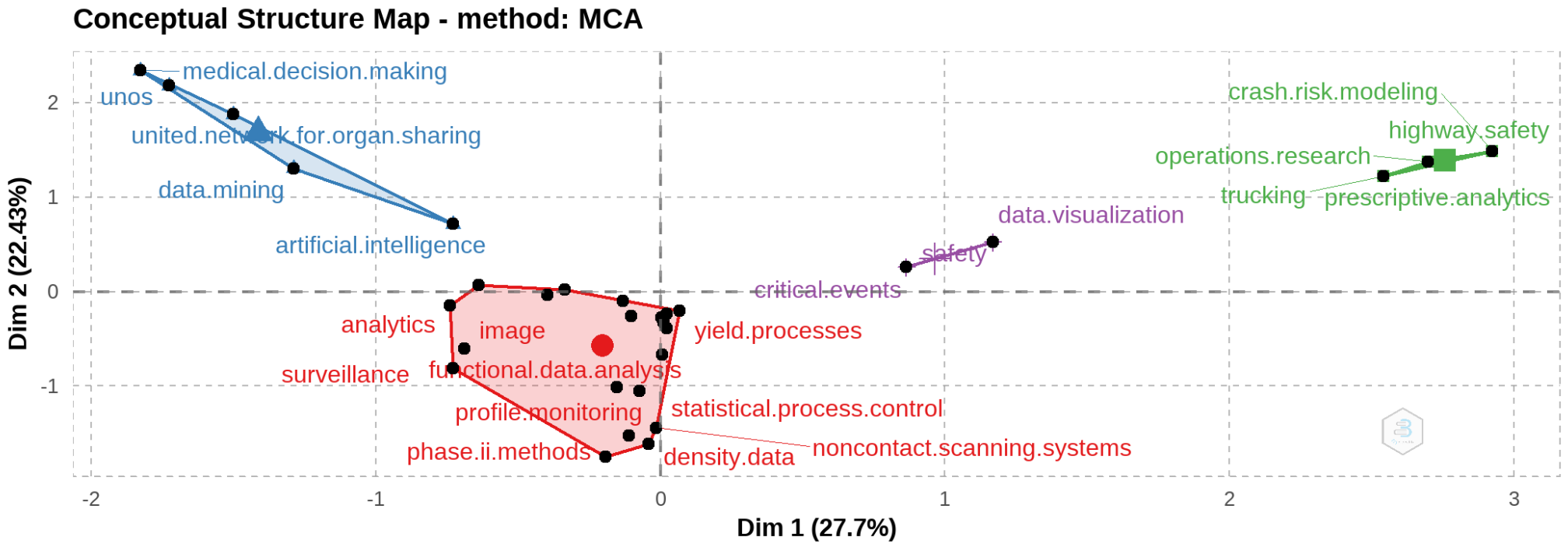
Introductions: Getting to Know Each Other

About Me – My route to Miami University

- Application of data-driven decisions (D3) in 3 continents.
- **Interests:** Applications in logistics, manufacturing, occupational safety & portfolios.
- **Collaborations with:** Aflac, GE Research, Gore, IBM Research, & Tennibot

My journey with data-driven decision making.

An Overview of My Research Portfolio



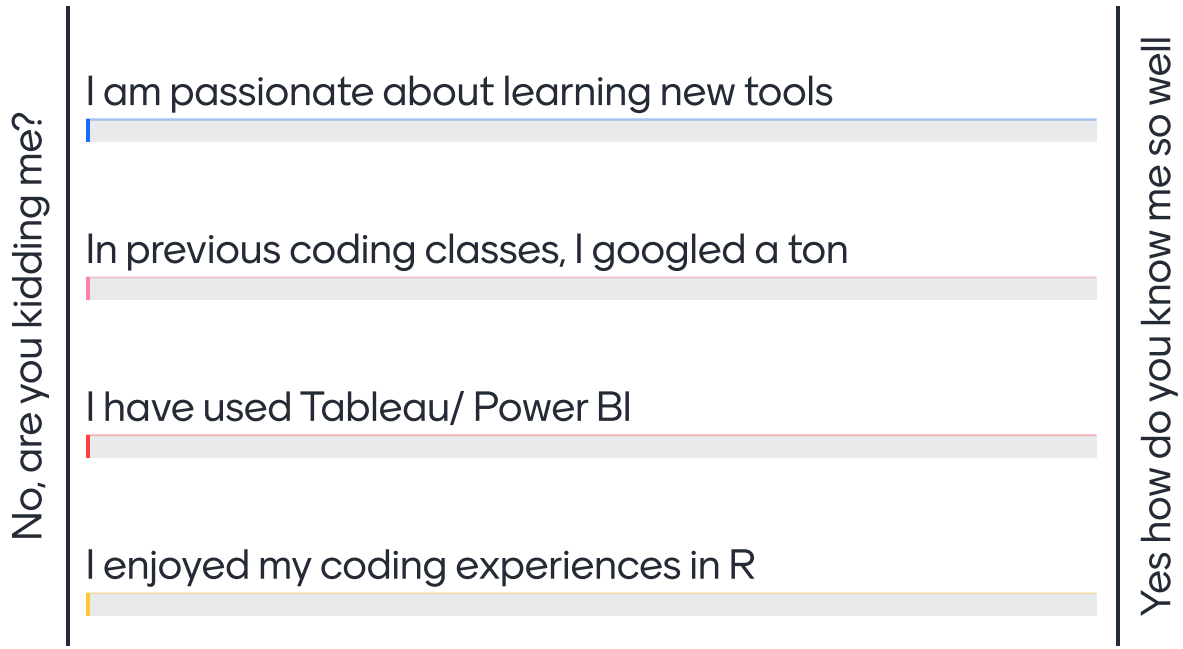
My work can be grouped into four different clusters.

Your Academic/Professional Experience

Go to **www.menti.com/dcc974**

On this scale, please answer the following:

Mentimeter



Getting to Know Your Learning Objectives

Go to **www.menti.com/epeqmvc7rw**

What are you hoping to get out of ISA 401/501?

So What is Data Visualization?

What is Data Visualization?

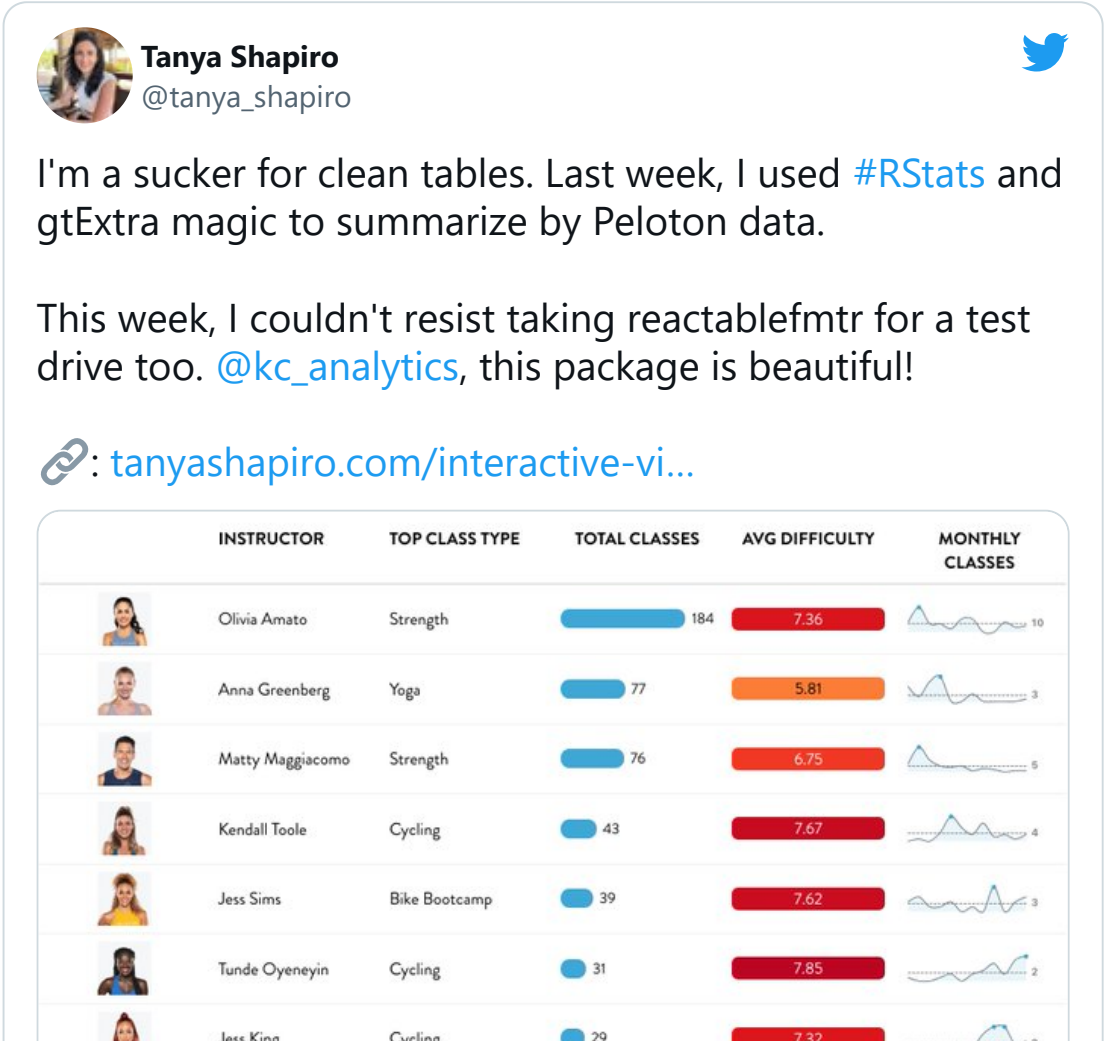
Data visualization involves **presenting data in a graphical format**. It is really a process that starts by getting data, creating initial plot(s) and modifying them to answer questions of interest (and possibly making the plot aesthetically pleasing). For example, see Cedric Scherer's visualization of the UNESCO data on global student to teacher ratios.

The Goals of Data Visualization

- **Record** information
- **Analyze** data to support reasoning
  - Develop and assess hypotheses (EDA)
  - Reveal patterns
  - Discover errors in data
- **Communicate** ideas to others
  - Infographics
  - Statistic charts
  - Interactive charts
  - Dashboards
- **Interact with the data (which supports all the above)**

Record: My Great Grandparents

Record: A More Modern Example



Tanya Shapiro
@tanya_shapiro

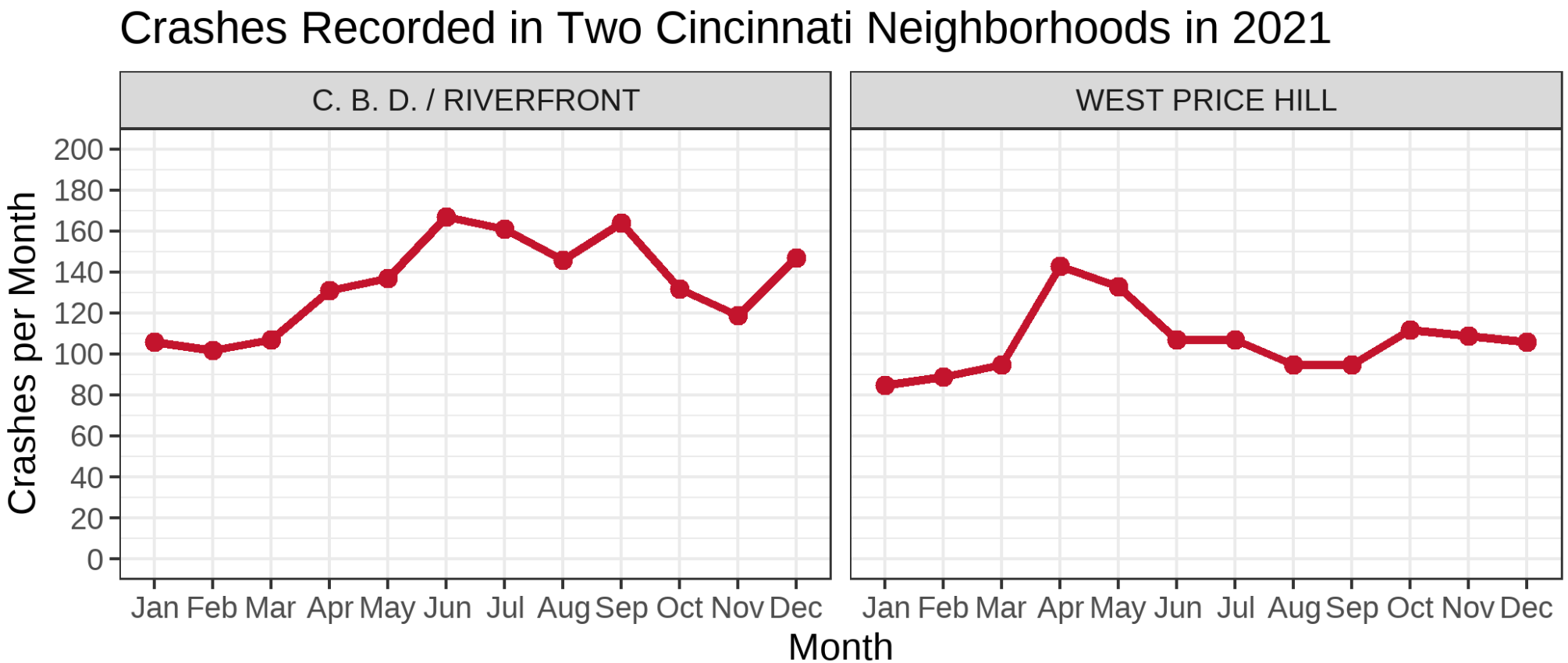
I'm a sucker for clean tables. Last week, I used #RStats and gtExtra magic to summarize by Peloton data.

This week, I couldn't resist taking reactablefmtr for a test drive too. @kc_analytics, this package is beautiful!

🔗: tanyashapiro.com/interactive-vi…

	INSTRUCTOR	TOP CLASS TYPE	TOTAL CLASSES	AVG DIFFICULTY	MONTHLY CLASSES
	Olivia Amato	Strength	184	7.36	10
	Anna Greenberg	Yoga	77	5.81	3
	Matty Maggiacomo	Strength	76	6.75	5
	Kendall Toole	Cycling	43	7.67	4
	Jess Sims	Bike Bootcamp	39	7.62	3
	Tunde Oyeneyin	Cycling	31	7.85	2
	Jess King	Cycling	29	7.32	

Analyze Data



Created by: Fadel Megahed | Data source: City of Cincy Open Data Portal (rvmt-pkmq)

Reveal Patterns: The 1854 Cholera Outbreak

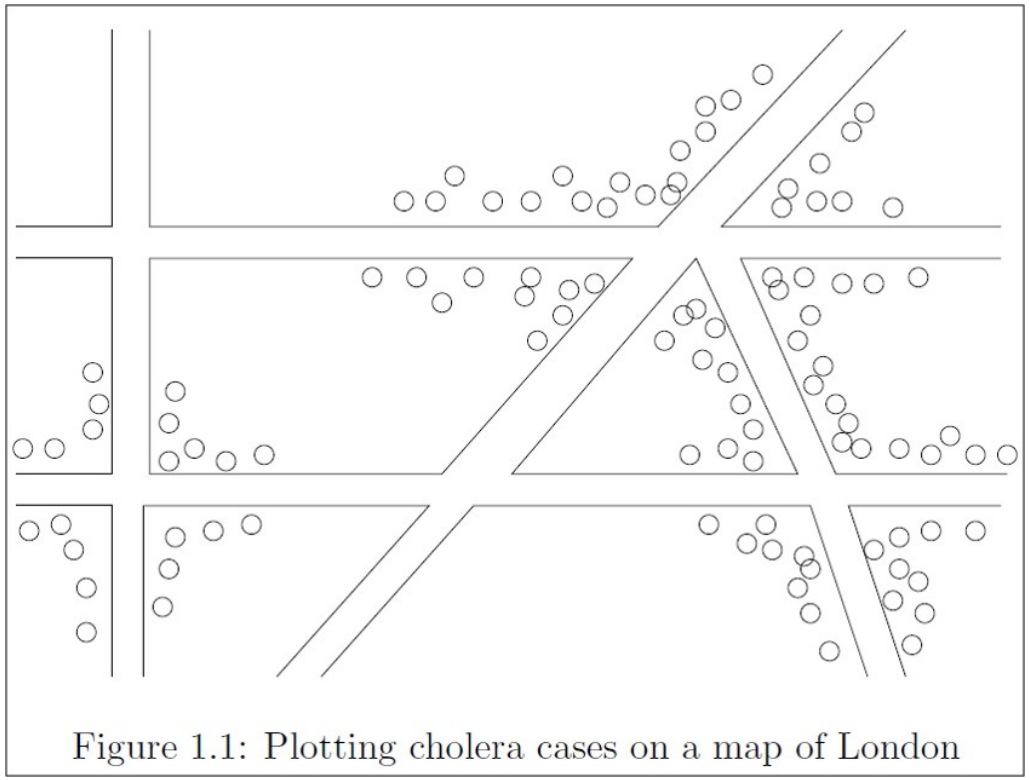
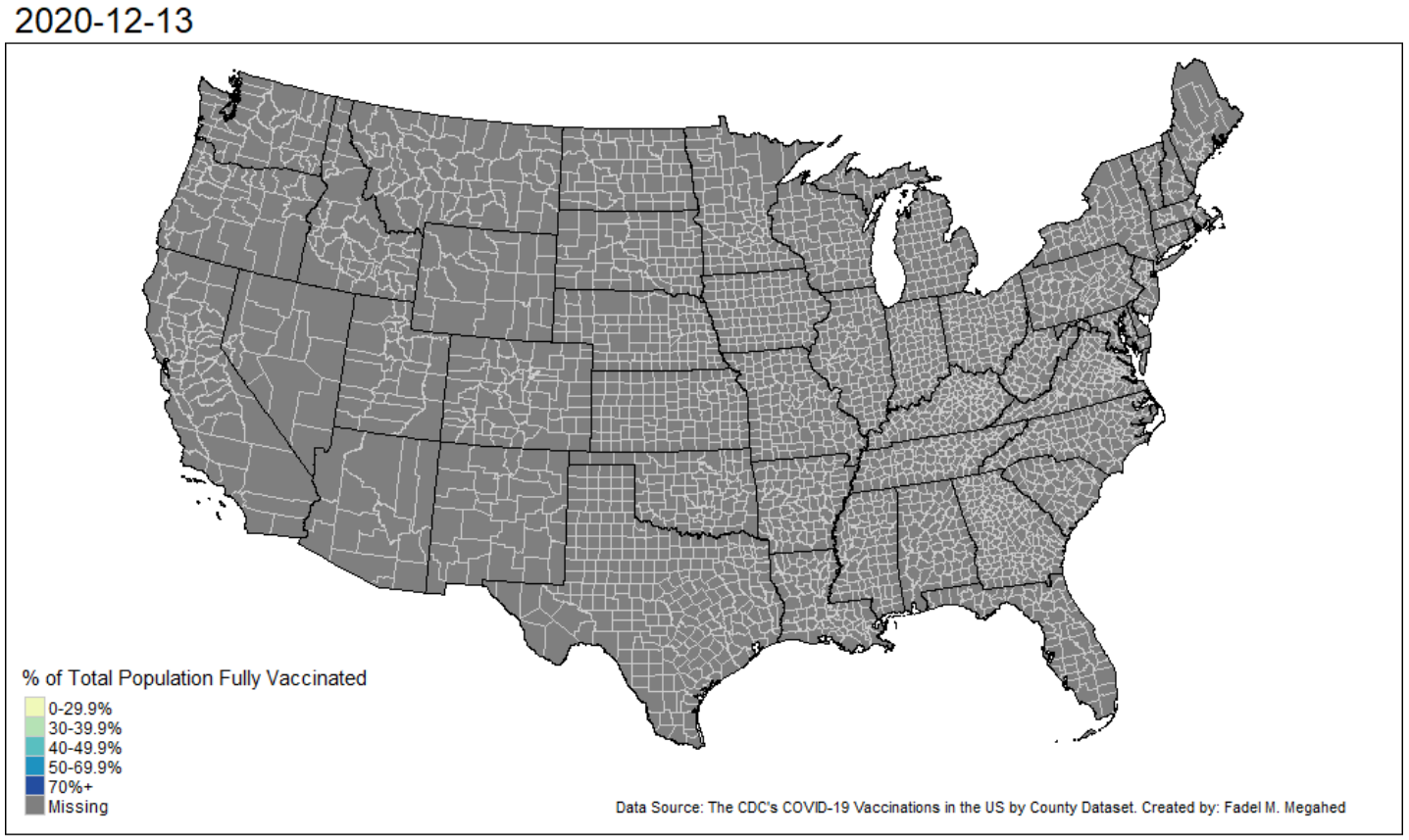


Figure 1.1: Plotting cholera cases on a map of London

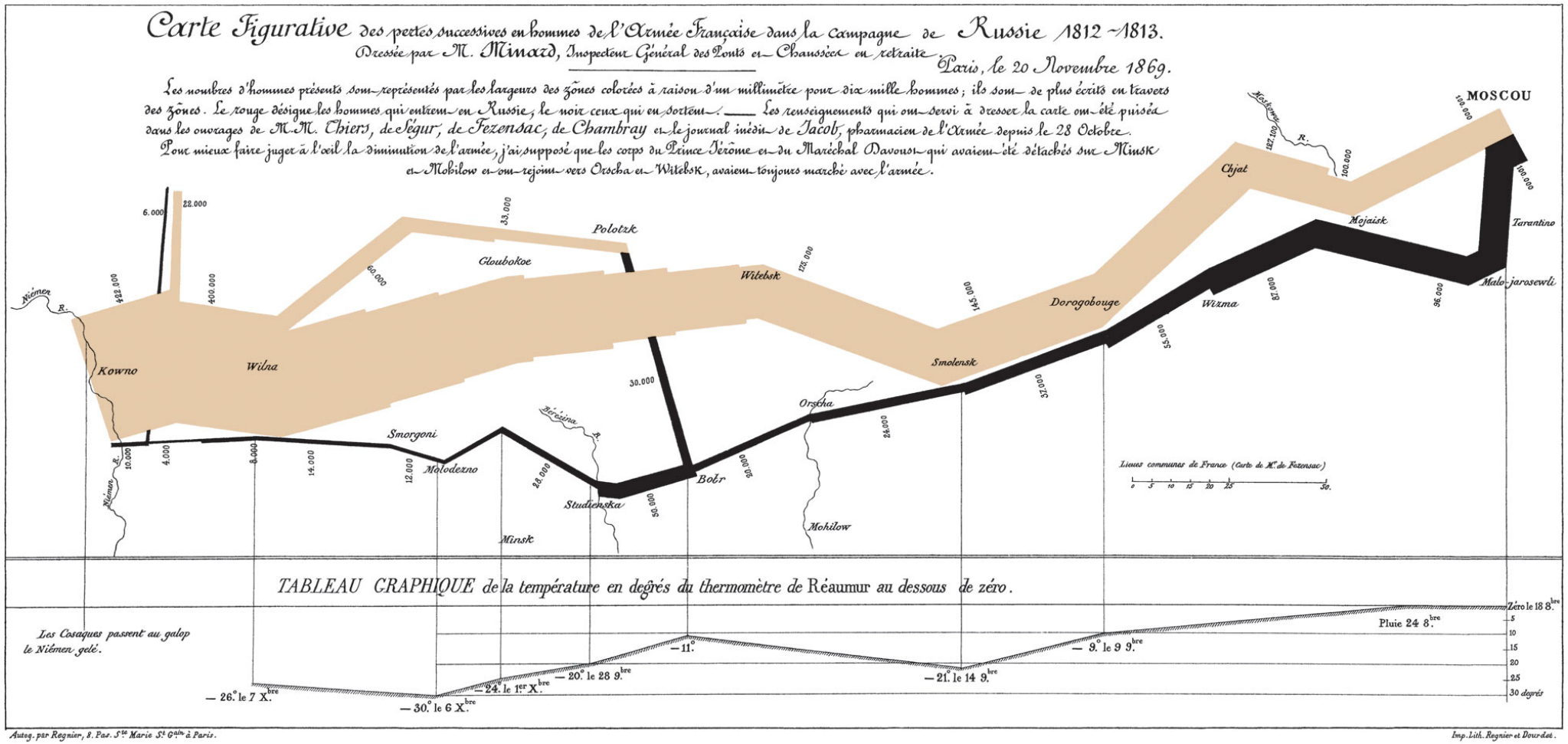
The physician John Snow, dealing with a Cholera outbreak plotted the cases on a map of the city (see schematic above).

Source: Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). Mining of Massive Data Sets (Third Edition). Cambridge University Press. Image is from Chapter 1, which can be accessed here.

Reveal Patterns: COVID-19 Vaccination Rates



Communicate Ideas: C.J Minard 1869



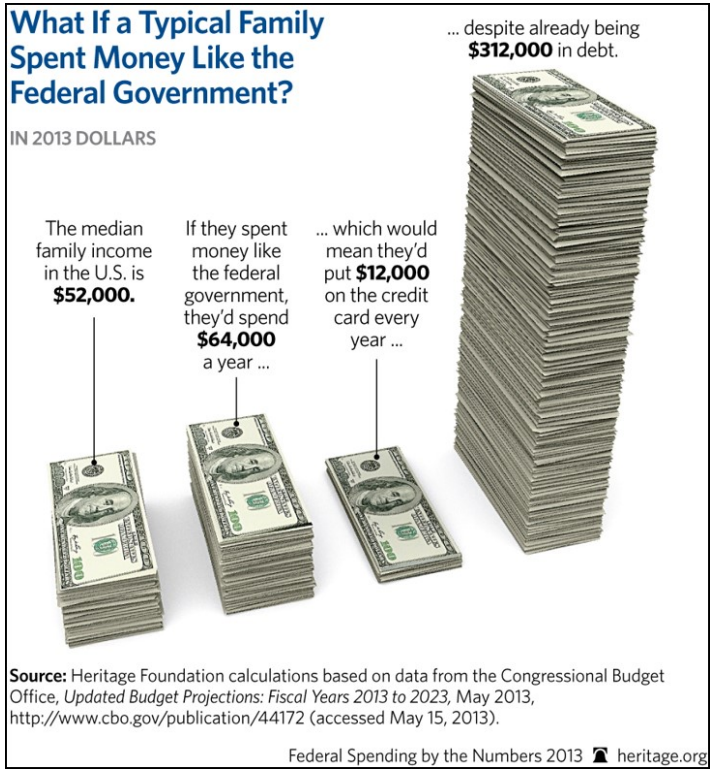
Communicate Ideas

05:00

Activity | Your Solution

Non-Graded Class Activity #5

- Who is the target audience?
- What is the data represented in this visualization? Be Specific.
- How is the data visually encoded?
- Do you like/dislike this visualization? Why?
- Would you do visualization like this for a similar dataset? Why? Why not?



Communicate Ideas

05:00

Activity | Your Solution

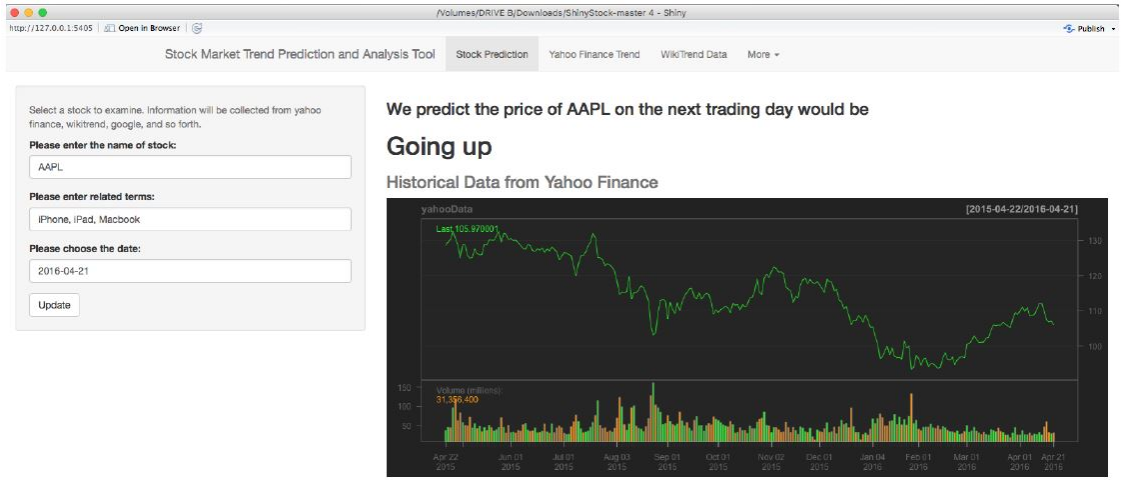
- Insert your solution here (Use Chrome as your browser to edit this part of the page)

Interact: GapMinder/ Hans Rosling Example

Business Intelligence: From Visualizations to Dashboards to Insights

What is Business Intelligence?

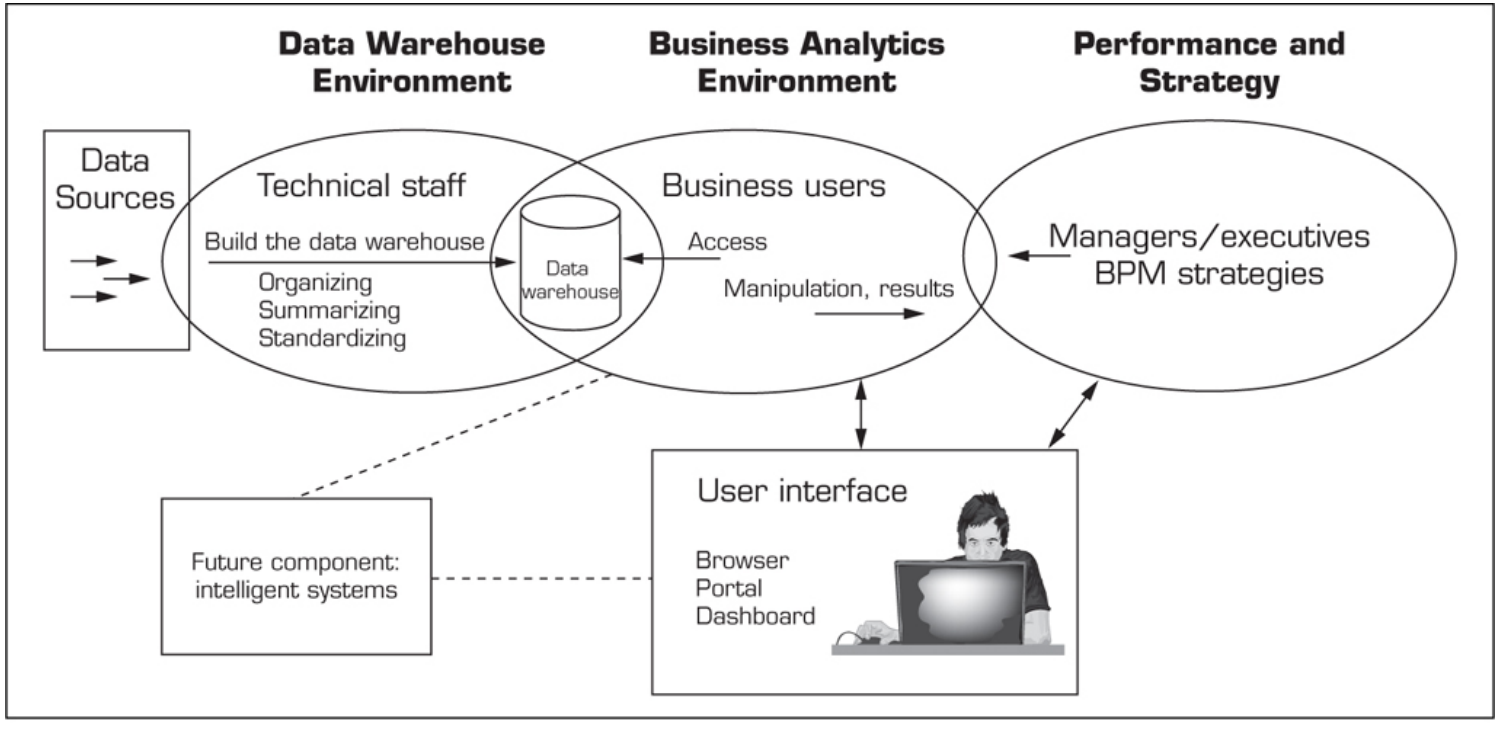
"... to enable **interactive access (sometimes in real time)** to data, to enable manipulation of data, and to give business managers and analysts the ability to conduct appropriate analysis. By analyzing ... data, situations, and performances, decision makers get valuable insights that enable them to **make more informed and better decisions** ... BI is based on the **transformation of data to information, then to decisions, and finally to actions.**"



Quote from Sharda, R., Delen, D., & Turban, E. (2013). Business Intelligence: A managerial perspective on analytics. Prentice Hall Press.

Image Credit: Joint work with Bin Weng.

The BI Process



Source: Based on W. Eckerson, Smart Companies in the 21st Century: The Secrets of Creating Successful Business Intelligent Solutions. The Data Warehousing Institute, Seattle, WA 2003, p. 32, Illustration 5.

• **Image Credit:** Sharda, R., Delen, D., & Turban, E. (2013). Business Intelligence: A managerial perspective on analytics. Prentice Hall Press.

Recap

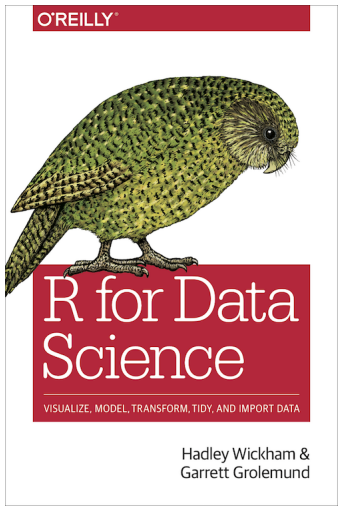
Summary of Main Points

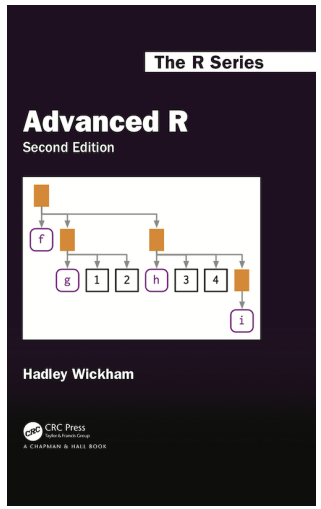
By now, you should be able to do the following:

- Describe **course objectives** and **structure**.
- Define **data visualization** and describe its **main goals**.
- Describe the **BI methodology** and its **major concepts**.

Things to Do to Prepare for Our Next Class

- Go over your notes and complete Assignment 01 on Canvas.
- Rread through the following references in preparation of our next class.





- Workflow: basics
- Workflow: scripts
- Workflow: project

- Names and values
- Vectors
- Subsetting