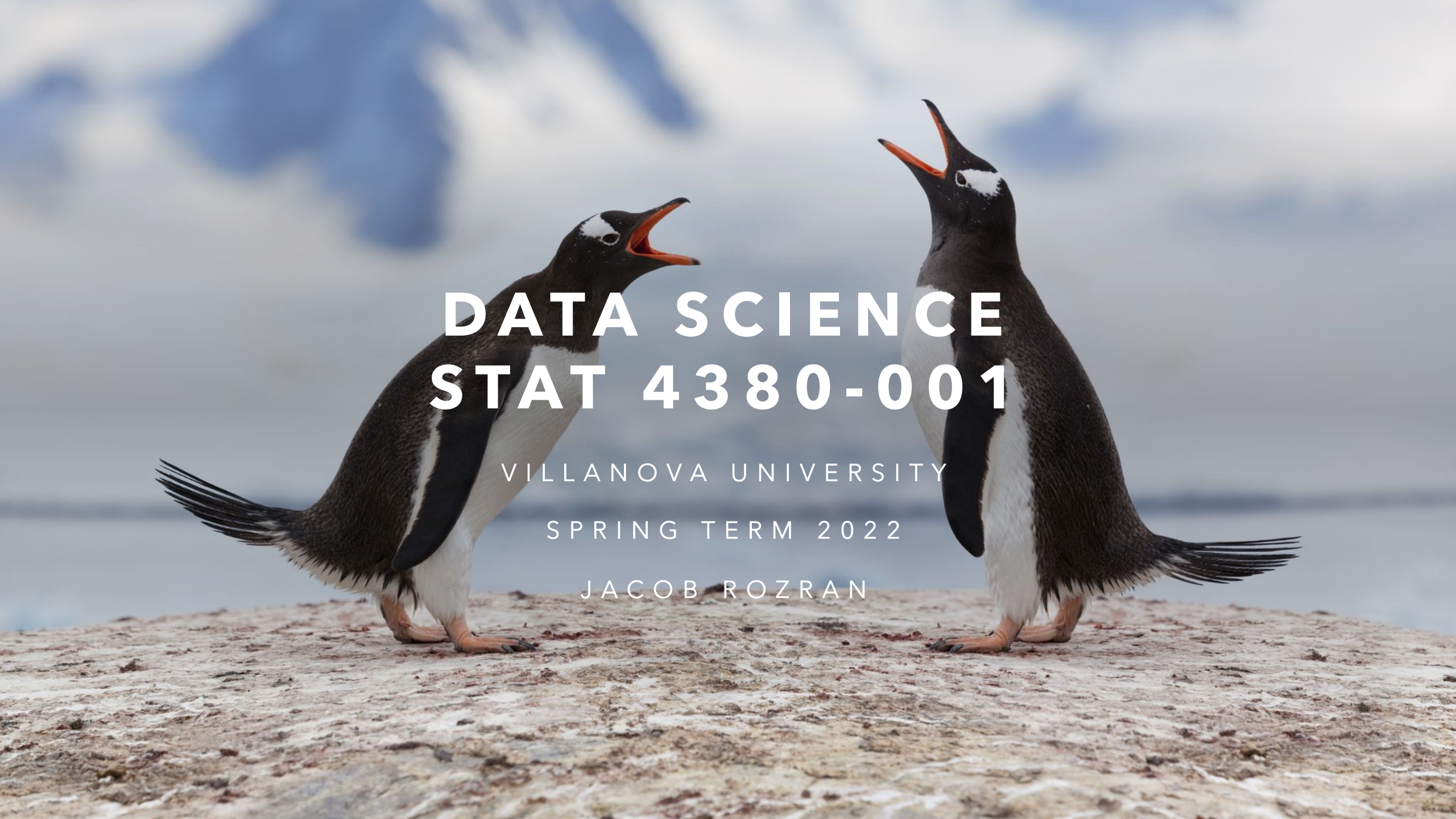


DATA SCIENCE
STAT 4380-001

VILLANOVA UNIVERSITY

SPRING TERM 2022

JACOB ROZRAN

WHO AM I?

- Please call me Jake (Not Dr. Jake, but I'll accept Jake from State Farm)
- Graduated Villanova in 2017 with my Masters in Applied Statistics
- I am an Adjunct Professor – so I work full time at a company called CivicActions as a Data Science Practice Lead
- Dad; Data Nerd; Local

Let’s go around the room and tell each other a little bit about ourselves:

- Where are you from?
- What is your major?
- What is the coolest thing you’ve ever done?
- What do you hope to do when you graduate?

ICE BREAKER!!!!
WHO ARE YOU?!?!

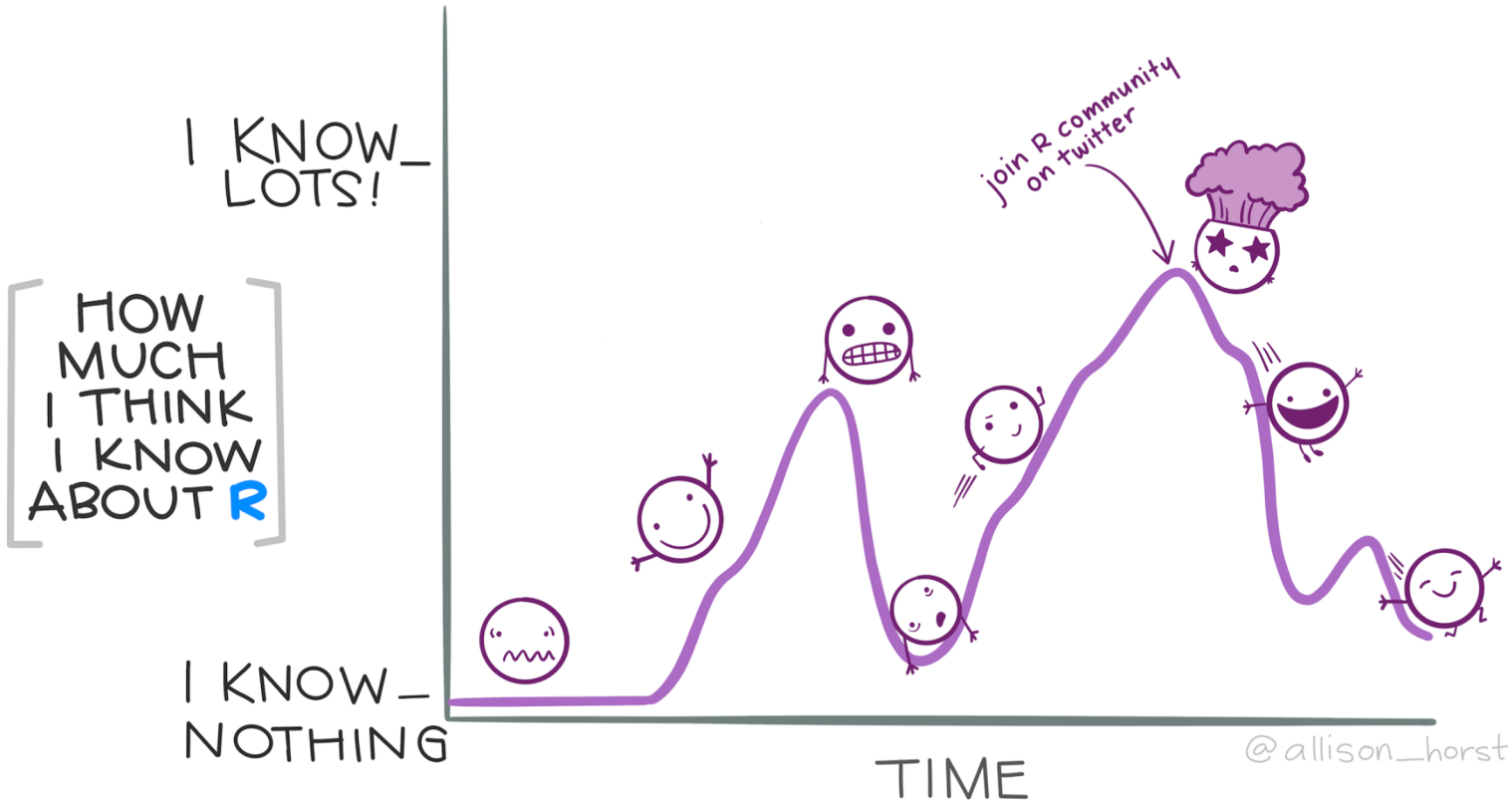
LET'S DO SOMETHING BORING BUT IMPORTANT

LET'S REVIEW THE SYLLABUS!

WE ARE NOT COMPETING HERE, WE ARE COLLABORATING. WE WILL GET TO THE FINISH LINE TOGETHER

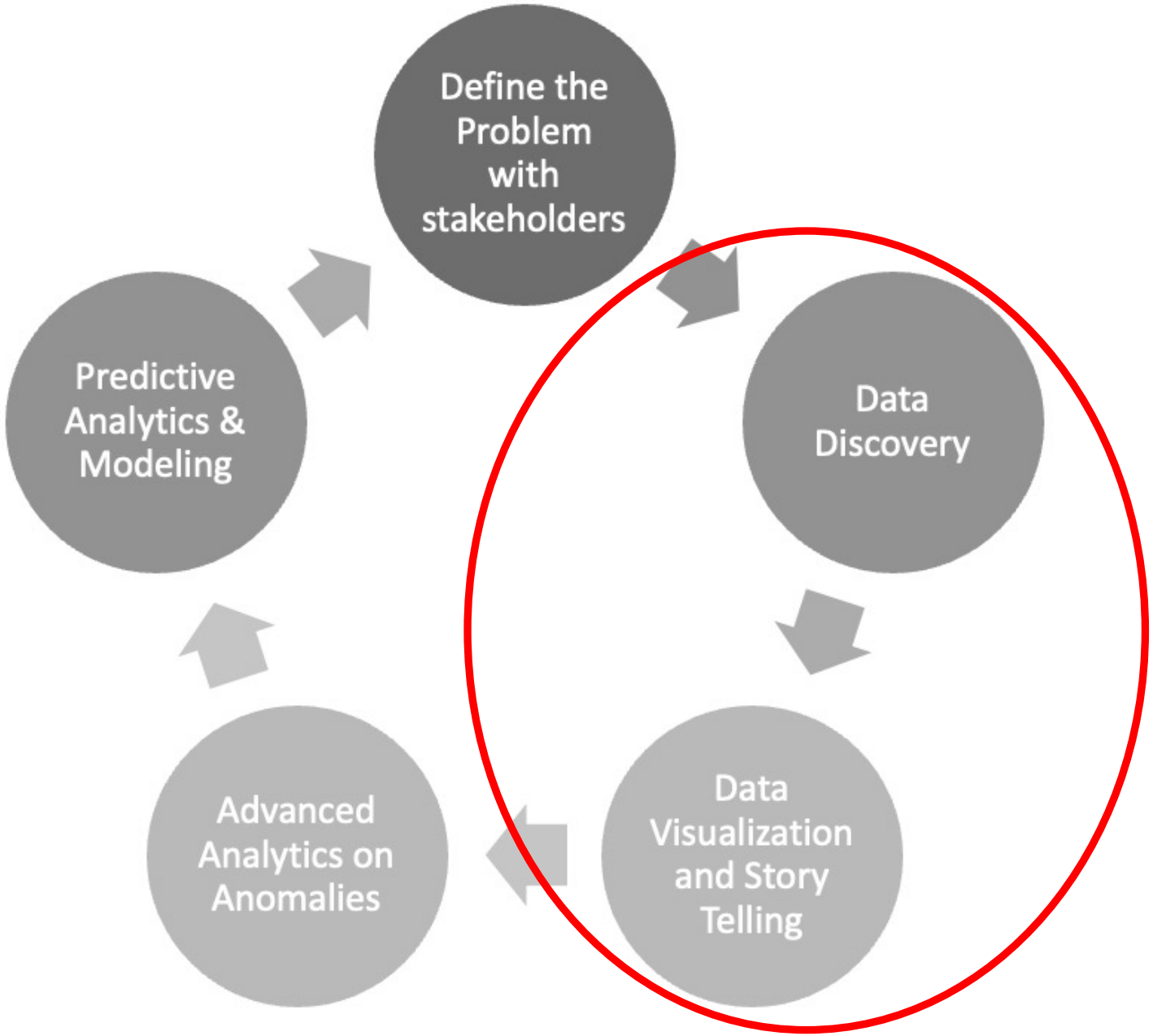
WHAT IS DATA SCIENCE, IRL?

SOME ASSEMBLY REQUIRED

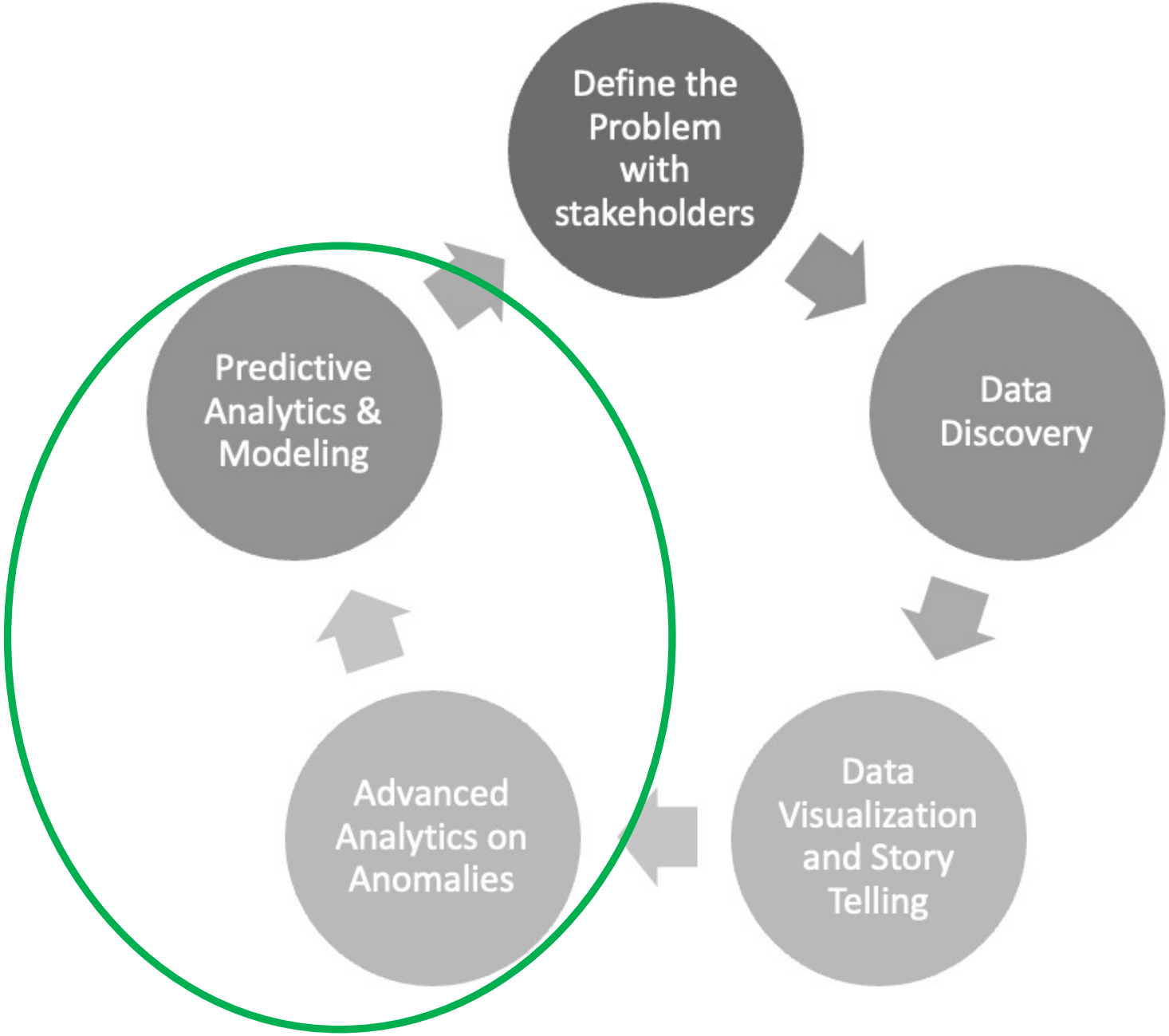


Artwork by @allison_horst

WE ARE GOING TO PRIMARILY FOCUS ON THIS…



AND SOME OF THIS…



NOT IN THE LIFECYCLE, BUT FOREVER PRESENT…

MY THOUGHTS WITH REAL EXAMPLES AT THE SAME TIME!

REAL ANALYSIS!

WHAT IS DATA SCIENCE?

Define the Problem

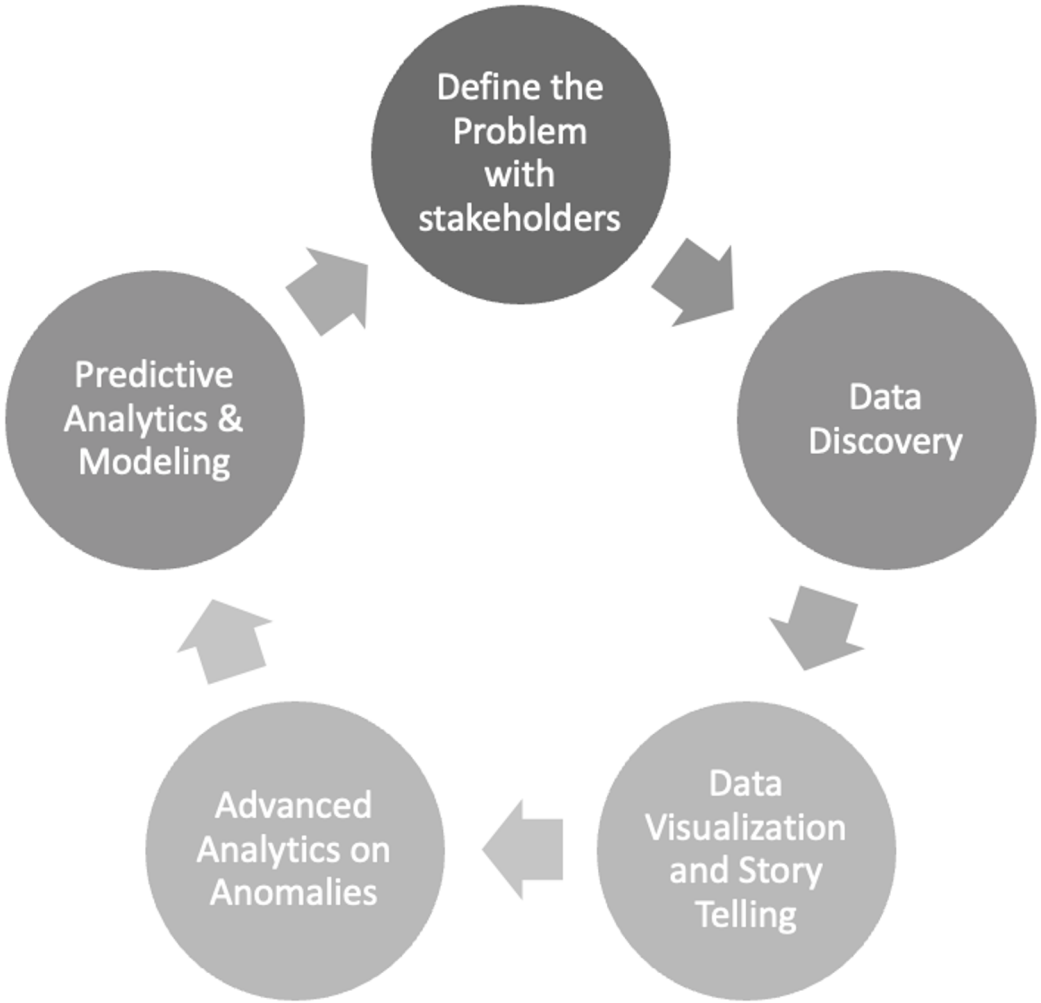
Data Discovery

Story Telling

Advanced Analytics

Predictive Analytics

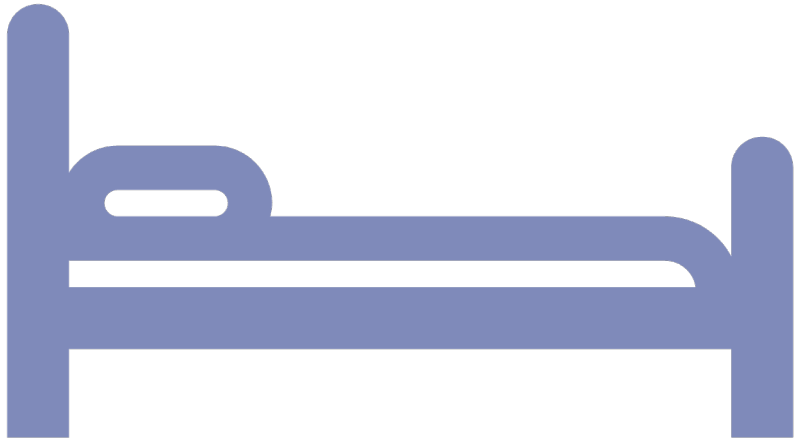
Ethics



STEP 1: DEFINING THE PROBLEM

What is the Question to Answer?

- For this analysis, I've used the Estimated ICU Beds Occupied by State Timeseries dataset from the HHS Open Data Site
- Because I'm doing this analysis as an example, there are no stakeholders for me to interact with
- As this data has ICU Bed Occupation data, I'll investigate what the overall ICU bed occupation rate looks like in the US and split that out by state
- My goal is to inform my class on what data science is and how it can be implemented practically



WHAT IS DATA SCIENCE?

Define the Problem

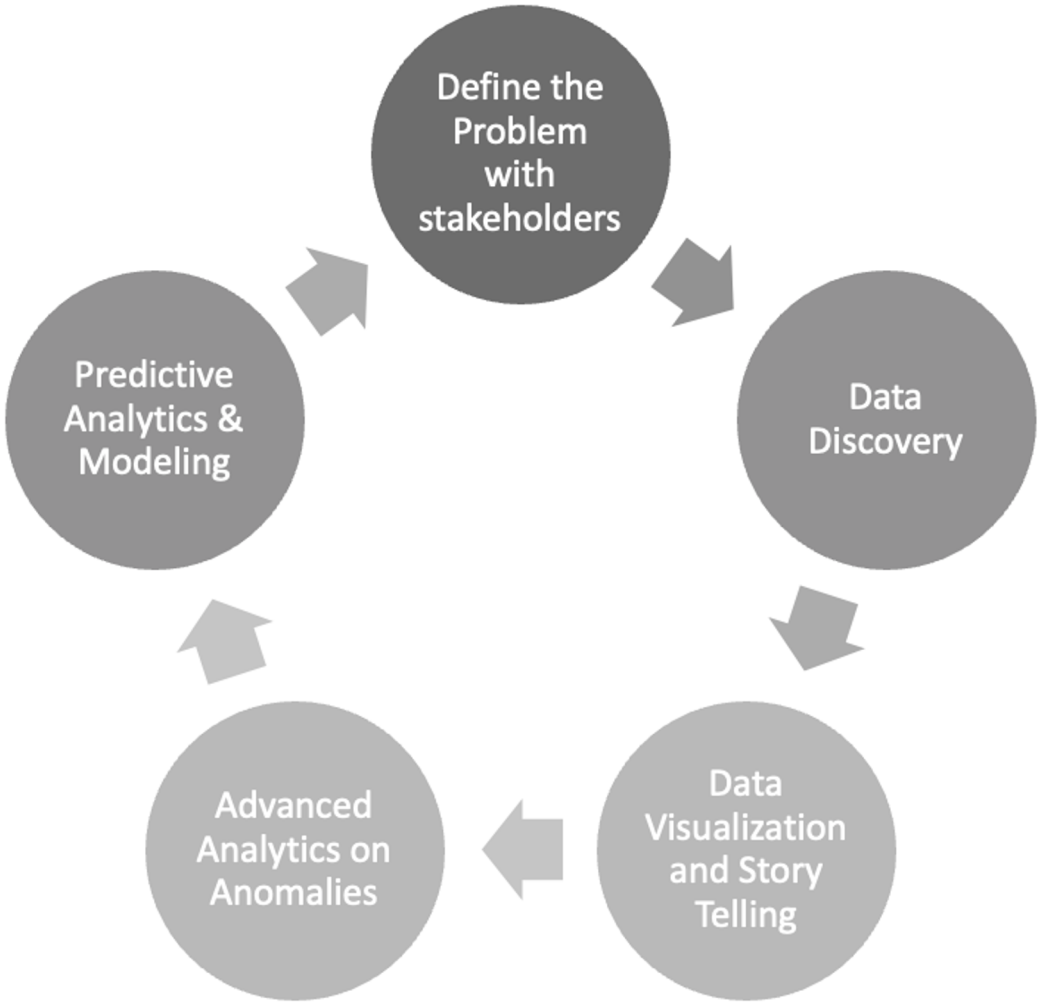
Data Discovery (Data Wrangling)

Story Telling

Advanced Analytics

Predictive Analytics

Ethics



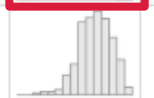
First Look at the Data

State	Collection Date	Occupied ICU Beds (count)	Lower Level (count)	Upper Level (count)	Occupied ICU Beds (%)	Lower Level (%)	Upper Level (%)	Total ICU Beds (count)	Lower Level - Total ICU Beds (count)	Upper Level - Total ICU Beds (count)
CW	2021-01-24	66,538	66,529	66,547	76.16	76.14	76.19	87,297	87,095	87,499
CW	2021-01-25	66,664	66,636	66,692	76.11	76.04	76.19	87,508	87,306	87,710
CW	2021-01-26	68,459	68,454	68,464	77.82	77.80	77.83	87,905	87,704	88,107
CW	2021-01-27	68,556	68,533	68,579	78.04	77.93	78.15	87,753	87,554	87,952
CW	2021-01-28	68,120	68,110	68,129	77.53	77.50	77.57	87,761	87,568	87,954
CW	2021-01-29	67,830	67,811	67,850	77.23	77.17	77.29	87,762	87,575	87,949

* For the presentation, I've updated the column titles

Data Summary

1. It may be that CW is the sum of all the rest
2. Some of the upper limits on the percentages are higher than 100% - technically not illegal for a confidence interval, but still nonsensical

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	State [character]	1. AK 2. AL 3. AR 4. AZ 5. CA 6. CO 7. CT 8. CW 9. DC 10. DE [42 others]	31 (1.9%) 31 (1.9%) 31 (1.9%) 31 (1.9%) 31 (1.9%) 31 (1.9%) 31 (1.9%) 31 (1.9%) 31 (1.9%) 31 (1.9%) 1302 (80.8%)		1612 (100.0%)	0 (0.0%)
2	Collection Date [Date]	min : 2021-01-24 med : 2021-02-08 max : 2021-02-23 range : 30d	31 distinct values		1612 (100.0%)	0 (0.0%)
3	Occupied ICU Beds (count) [numeric]	Mean (sd) : 2474.7 (8798) min < med < max: 32 < 834 < 68556 IQR (CV) : 1198.5 (3.6)	1104 distinct values		1612 (100.0%)	0 (0.0%)
4	Lower Level (count) [numeric]	Mean (sd) : 2471.5 (8792) min < med < max: 31 < 833 < 68533 IQR (CV) : 1200.8 (3.6)	1111 distinct values		1612 (100.0%)	0 (0.0%)
5	Upper Level (count) [numeric]	Mean (sd) : 2477.8 (8803.9) min < med < max: 32 < 837 < 68579 IQR (CV) : 1197 (3.6)	1120 distinct values		1612 (100.0%)	0 (0.0%)
6	Occupied ICU Beds (%) [numeric]	Mean (sd) : 0.7 (0.1) min < med < max: 0.2 < 0.7 < 0.9 IQR (CV) : 0.2 (0.2)	1267 distinct values		1612 (100.0%)	0 (0.0%)
7	Lower Level (%) [numeric]	Mean (sd) : 0.7 (0.1) min < med < max: 0 < 0.7 < 0.9 IQR (CV) : 0.2 (0.2)	1267 distinct values		1612 (100.0%)	0 (0.0%)
8	Upper Level (%) [numeric]	Mean (sd) : 0.7 (0.1) min < med < max: 0.2 < 0.7 < 1.6 IQR (CV) : 0.2 (0.2)	1268 distinct values		1612 (100.0%)	0 (0.0%)
9	Total ICU Beds (count) [numeric]	Mean (sd) : 3324.6 (11781.7) min < med < max: 95 < 1169.5 < 87905 IQR (CV) : 1903.5 (3.5)	900 distinct values		1612 (100.0%)	0 (0.0%)
10	Lower Level - Total ICU Beds (count) [numeric]	Mean (sd) : 3315.6 (11756.3) min < med < max: 95 < 1147 < 87704 IQR (CV) : 1906 (3.5)	924 distinct values		1612 (100.0%)	0 (0.0%)
11	Upper Level - Total ICU Beds (count) [numeric]	Mean (sd) : 3333.7 (11807.1) min < med < max: 95 < 1170.5 < 88107 IQR (CV) : 1900.5 (3.5)	917 distinct values		1612 (100.0%)	0 (0.0%)

Output: Cleaned Data & Improved Data Dictionary

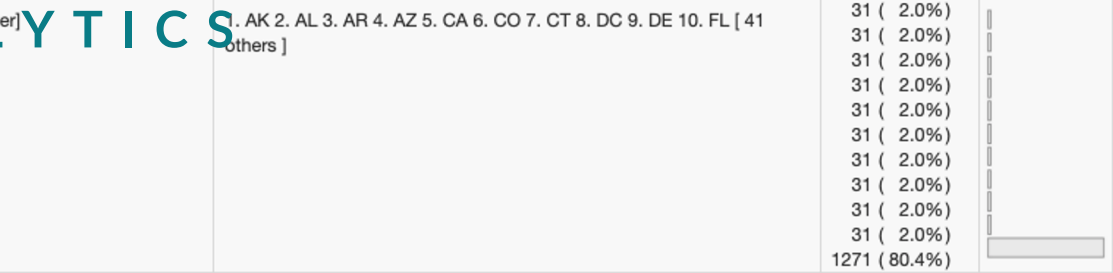
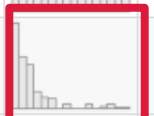
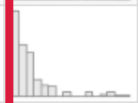
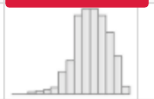
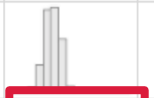
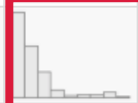
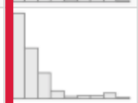
State	Collection Date	Occupied ICU Beds (count)	Lower Level (count)	Upper Level (count)	Occupied ICU Beds (%)	Lower Level (%)	Upper Level (%)	Total ICU Beds (count)	Lower Level - Total ICU Beds (count)	Upper Level - Total ICU Beds (count)
CW	2021-01-24	66538	66529	66547	0.7616	0.7614	0.7619	87297	87095	87499
CW	2021-01-25	66664	66636	66692	0.7611	0.7604	0.7619	87508	87306	87710
CW	2021-01-26	68459	68454	68464	0.7782	0.7780	0.7783	87905	87704	88107
CW	2021-01-27	68556	68533	68579	0.7804	0.7793	0.7815	87753	87554	87952
CW	2021-01-28	68120	68110	68129	0.7753	0.7750	0.7757	87761	87568	87954
CW	2021-01-29	67830	67811	67850	0.7723	0.7717	0.7729	87762	87575	87949

* For the presentation, I've updated the column titles

Diving Deeper on Anomalous Data

1. Is the Outlier from the Data Summary CW?
2. If so, is CW the “Overall” Metric?

Output: Updated Data Dictionary & Complete Understanding of Data

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	State [character]	1. AK 2. AL 3. AR 4. AZ 5. CA 6. CO 7. CT 8. DC 9. DE 10. FL [41 others]	31 (2.0%) 31 (2.0%) 31 (2.0%) 31 (2.0%) 31 (2.0%) 31 (2.0%) 31 (2.0%) 31 (2.0%) 31 (2.0%) 31 (2.0%) 1271 (80.4%)		1581 (100.0%)	0 (0.0%)
2	Collection Date [Date]	min : 2021-01-24 med : 2021-02-08 max : 2021-02-23 range : 30d	31 distinct values		1581 (100.0%)	0 (0.0%)
3	Occupied ICU Beds (count) [numeric]	Mean (sd) : 1261.6 (1497.7) min < med < max: 32 < 810 < 7740 IQR (CV) : 1186 (1.2)	1073 distinct values		1581 (100.0%)	0 (0.0%)
4	Lower Level (count) [numeric]	Mean (sd) : 1259.3 (1495.3) min < med < max: 31 < 802 < 7739 IQR (CV) : 1186 (1.2)	1080 distinct values		1581 (100.0%)	0 (0.0%)
5	Upper Level (count) [numeric]	Mean (sd) : 1263.9 (1500.2) min < med < max: 32 < 813 < 7742 IQR (CV) : 1186 (1.2)	1089 distinct values		1581 (100.0%)	0 (0.0%)
6	Occupied ICU Beds (%) [numeric]	Mean (sd) : 0.7 (0.1) min < med < max: 0.2 < 0.7 < 0.9 IQR (CV) : 0.2 (0.2)	1250 distinct values		1581 (100.0%)	0 (0.0%)
7	Lower Level (%) [numeric]	Mean (sd) : 0.7 (0.1) min < med < max: 0 < 0.7 < 0.9 IQR (CV) : 0.2 (0.2)	1248 distinct values		1581 (100.0%)	0 (0.0%)
8	Upper Level (%) [numeric]	Mean (sd) : 0.7 (0.1) min < med < max: 0.2 < 0.7 < 1.6 IQR (CV) : 0.2 (0.2)	1247 distinct values		1581 (100.0%)	0 (0.0%)
9	Total ICU Beds (count) [numeric]	Mean (sd) : 1694.9 (1822.3) min < med < max: 95 < 1068 < 8592 IQR (CV) : 1852 (1.1)	869 distinct values		1581 (100.0%)	0 (0.0%)
10	Lower Level - Total ICU Beds (count) [numeric]	Mean (sd) : 1689.2 (1813.5) min < med < max: 95 < 1068 < 8592 IQR (CV) : 1852 (1.1)	894 distinct values		1581 (100.0%)	0 (0.0%)
11	Upper Level - Total ICU Beds (count) [numeric]	Mean (sd) : 1700.6 (1831.5) min < med < max: 95 < 1070 < 8592 IQR (CV) : 1867 (1.1)	887 distinct values		1581 (100.0%)	0 (0.0%)

WHAT IS DATA SCIENCE?

Define the Problem

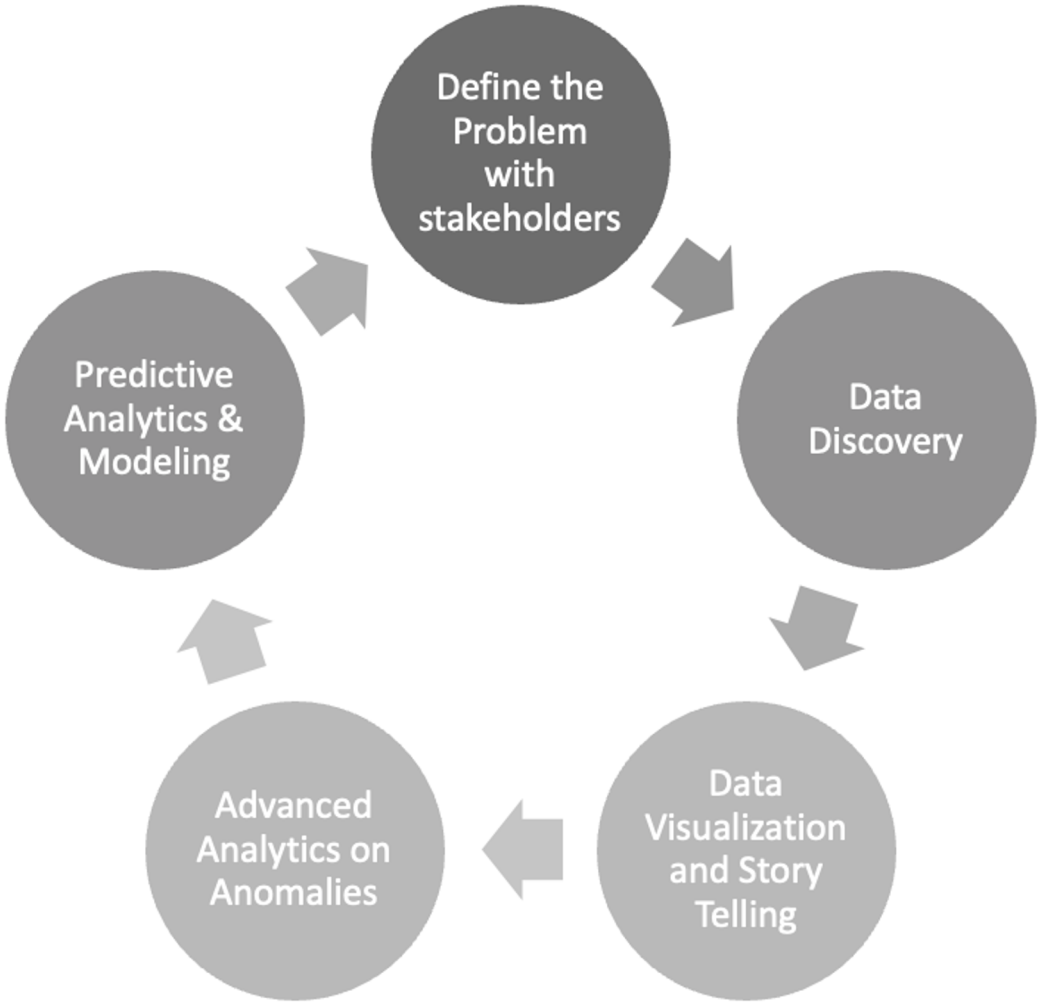
Data Discovery

Story Telling (Data Visualization)

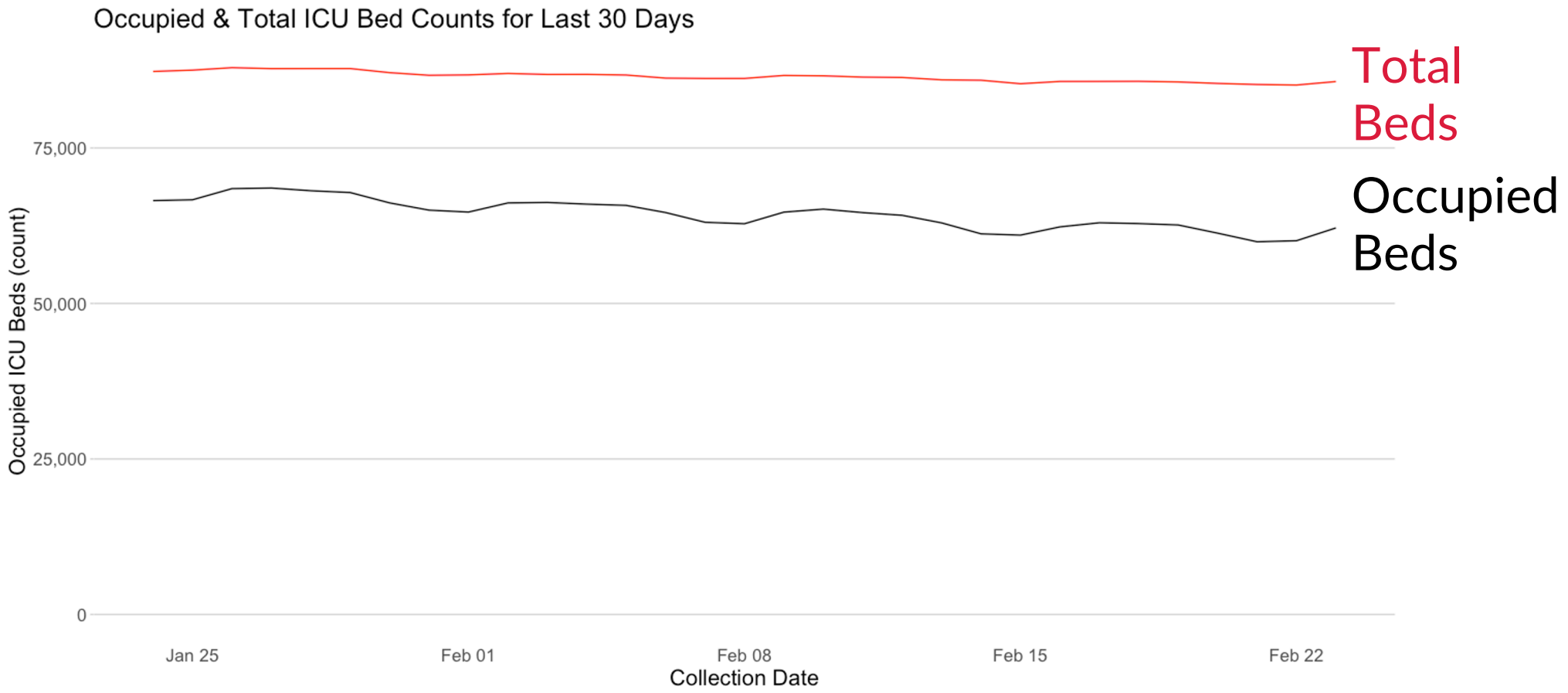
Advanced Analytics

Predictive Analytics

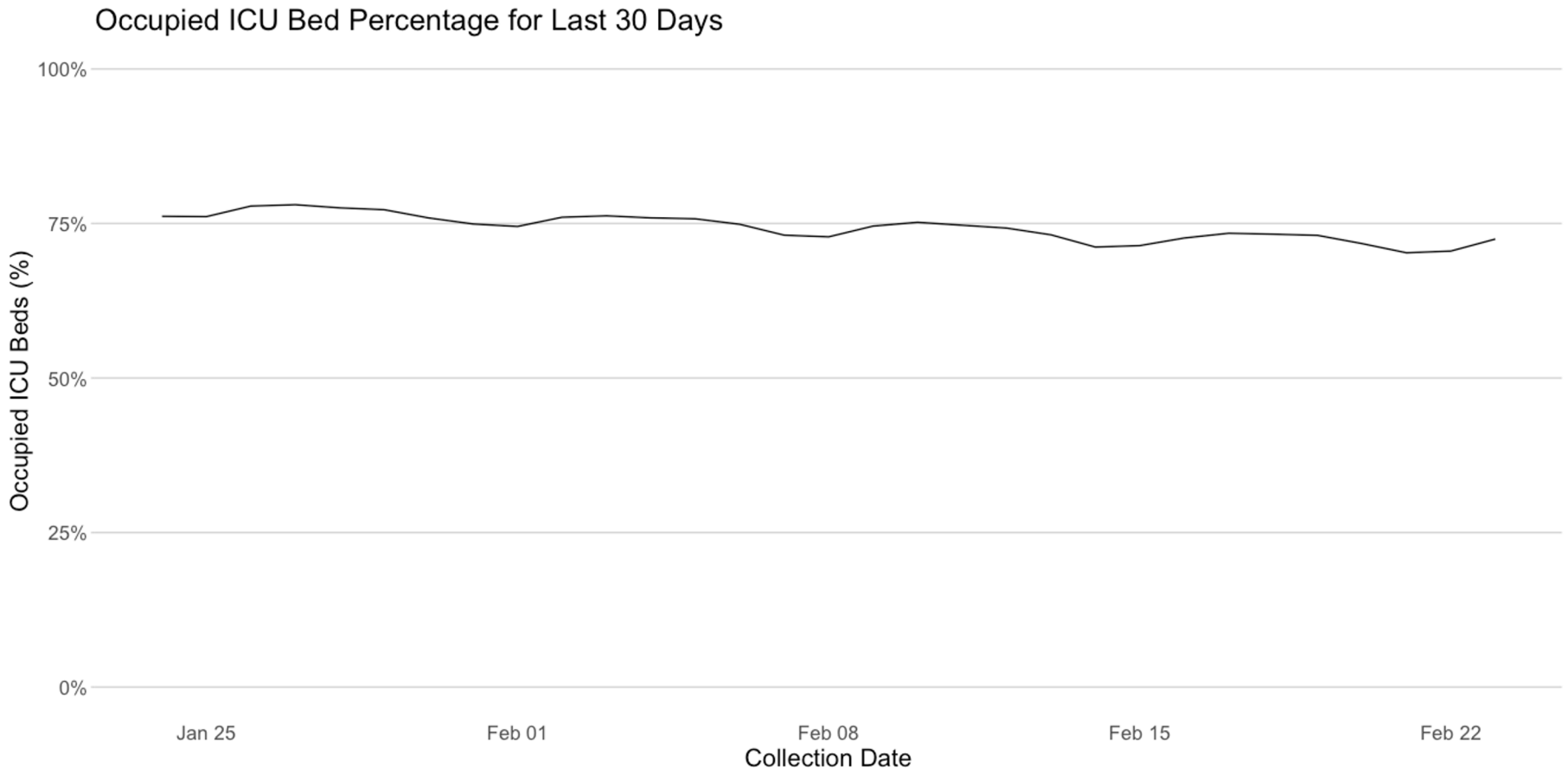
Ethics



How Many ICU Beds are Being Used?



What Percentage of ICU Beds are Being Used?



WHAT IS DATA SCIENCE?

Define the Problem

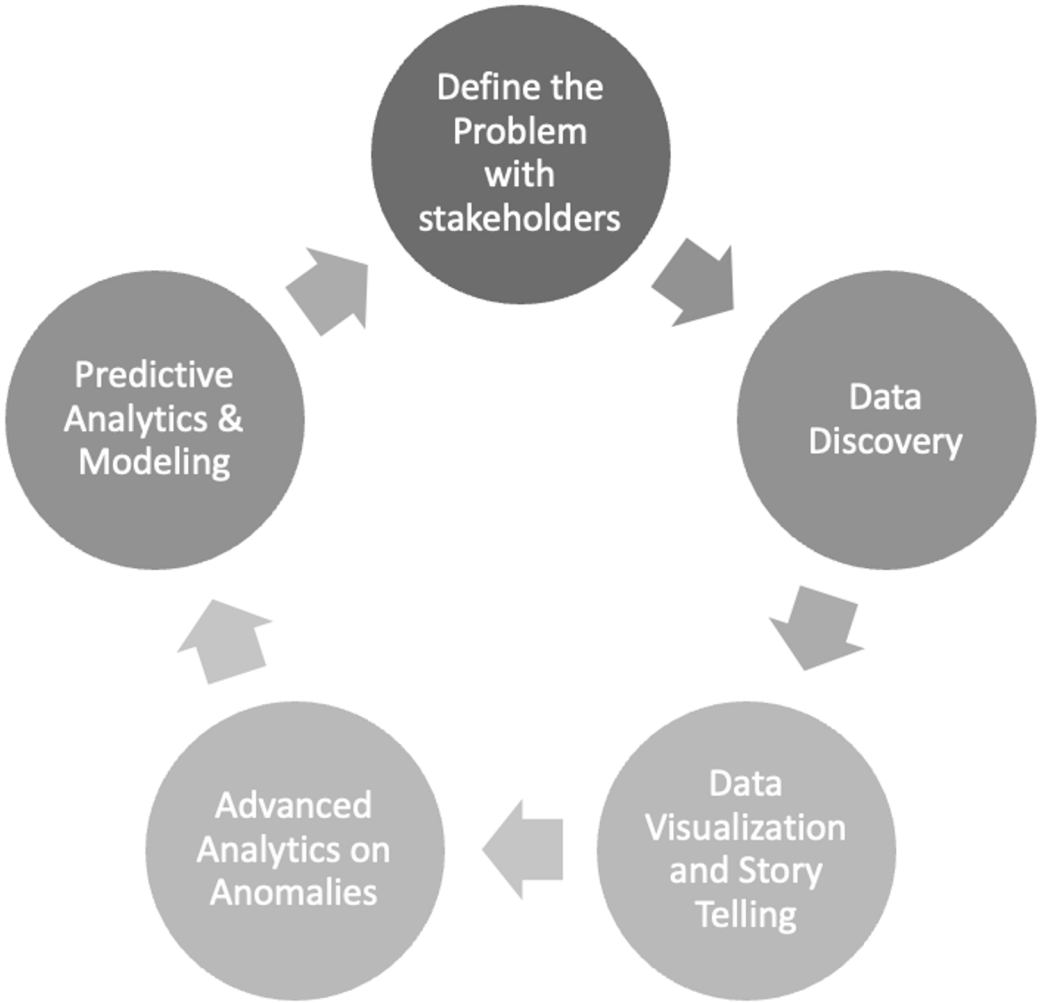
Data Discovery

Story Telling

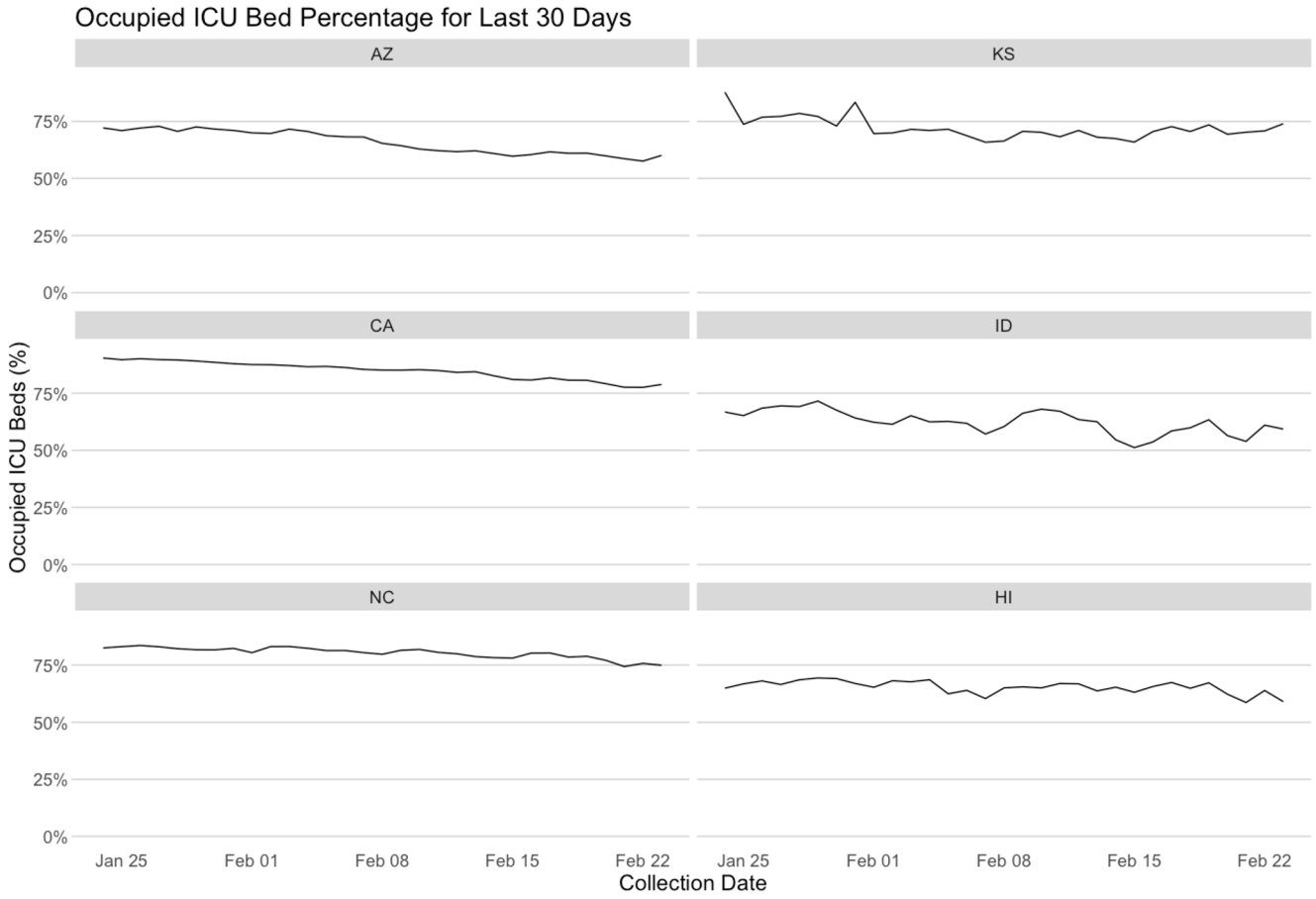
Advanced Analytics

Predictive Analytics

Ethics

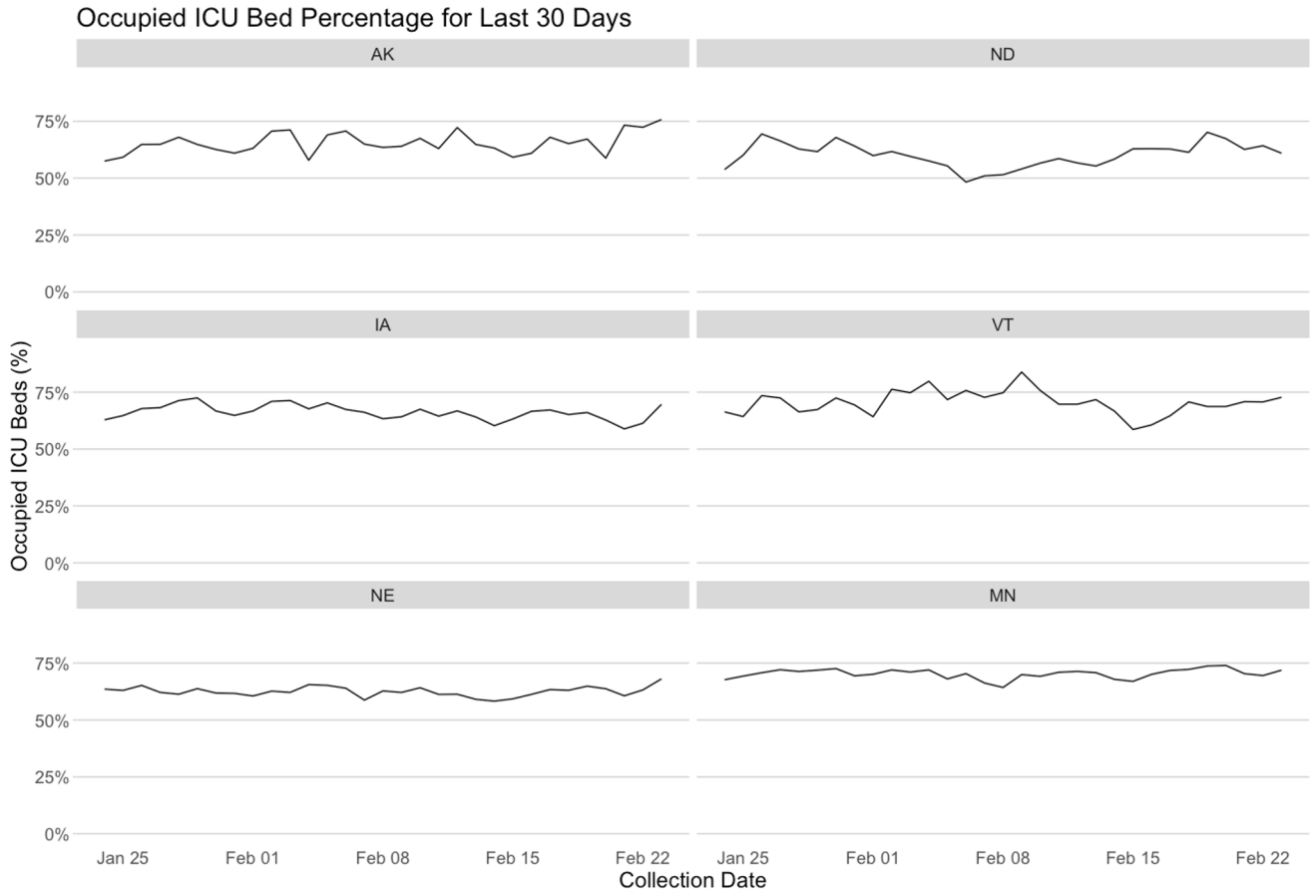


What States are Improving the Most?



State	Change in Percent
AZ	-16.7198%
KS	-15.7283%
CA	-12.7973%
ID	-11.1710%
NC	-9.1967%
HI	-9.0307%

What States are Deteriorating the Most?



State	Change in Percent
AK	31.7565%
ND	13.3346%
IA	10.8035%
VT	9.6487%
NE	7.0799%
MN	6.2057%

WHAT IS DATA SCIENCE?

Define the Problem

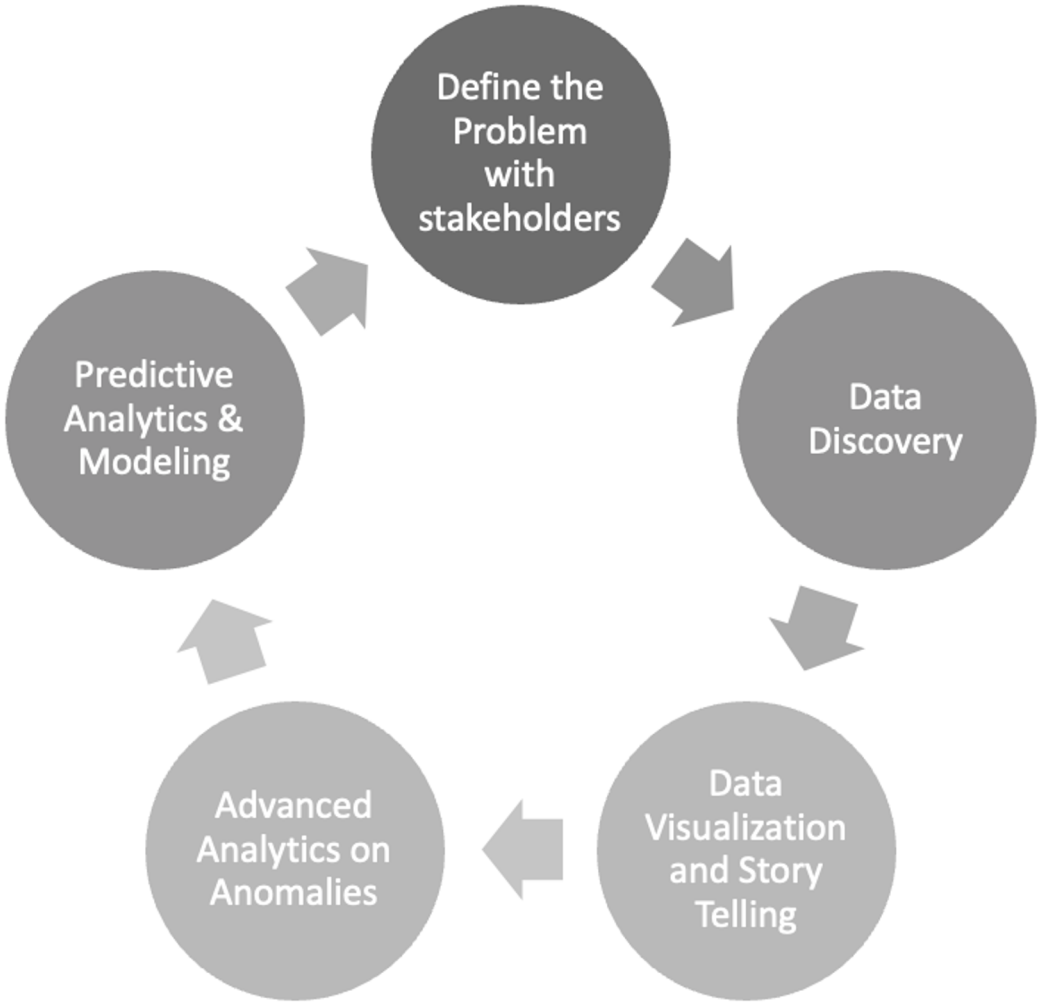
Data Discovery

Story Telling

Advanced Analytics

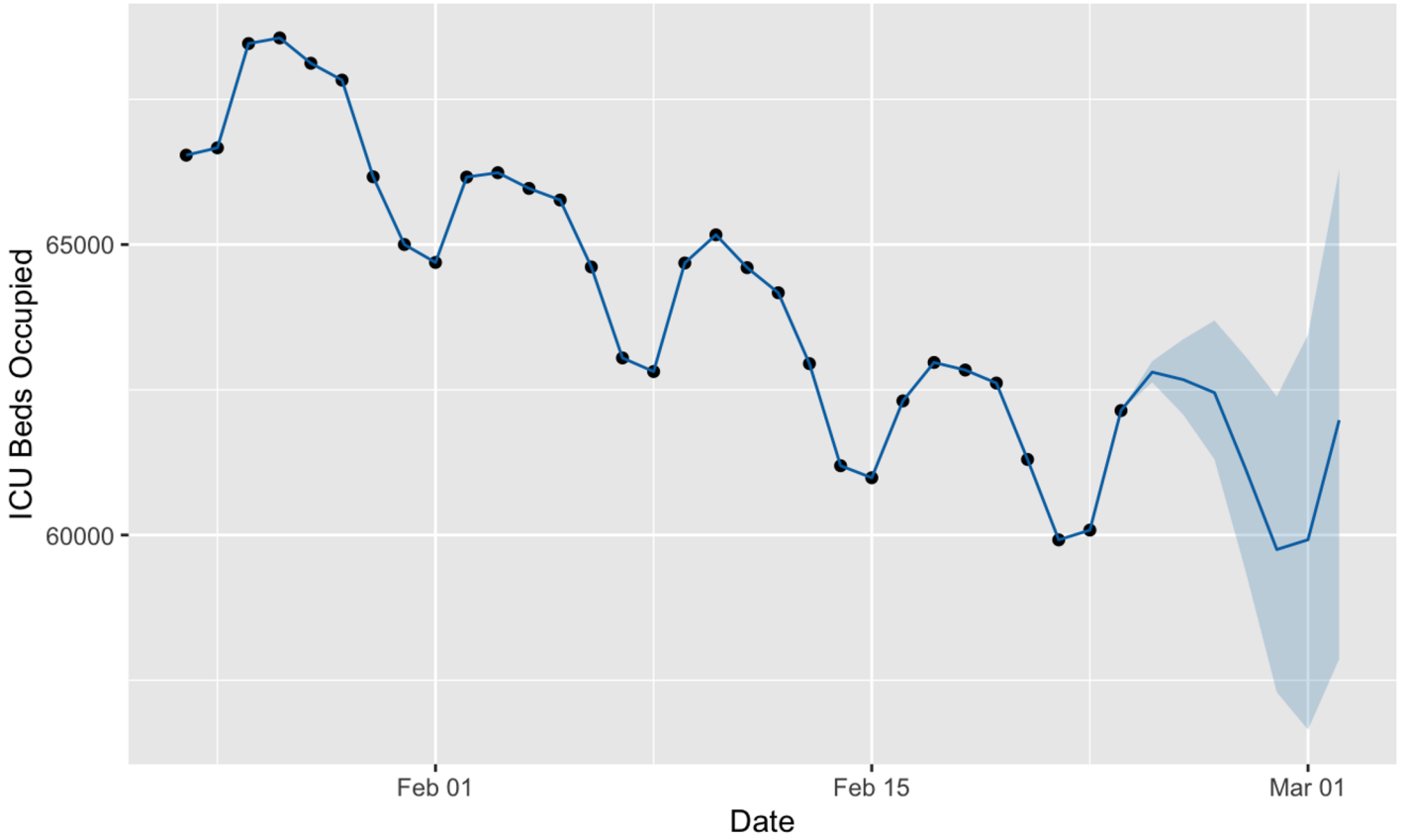
Predictive Analytics

Ethics



Now we know where we've been, where are we going?

Can we get a sense of where this trend will be in a week?



WHAT IS DATA SCIENCE?

Define the Problem

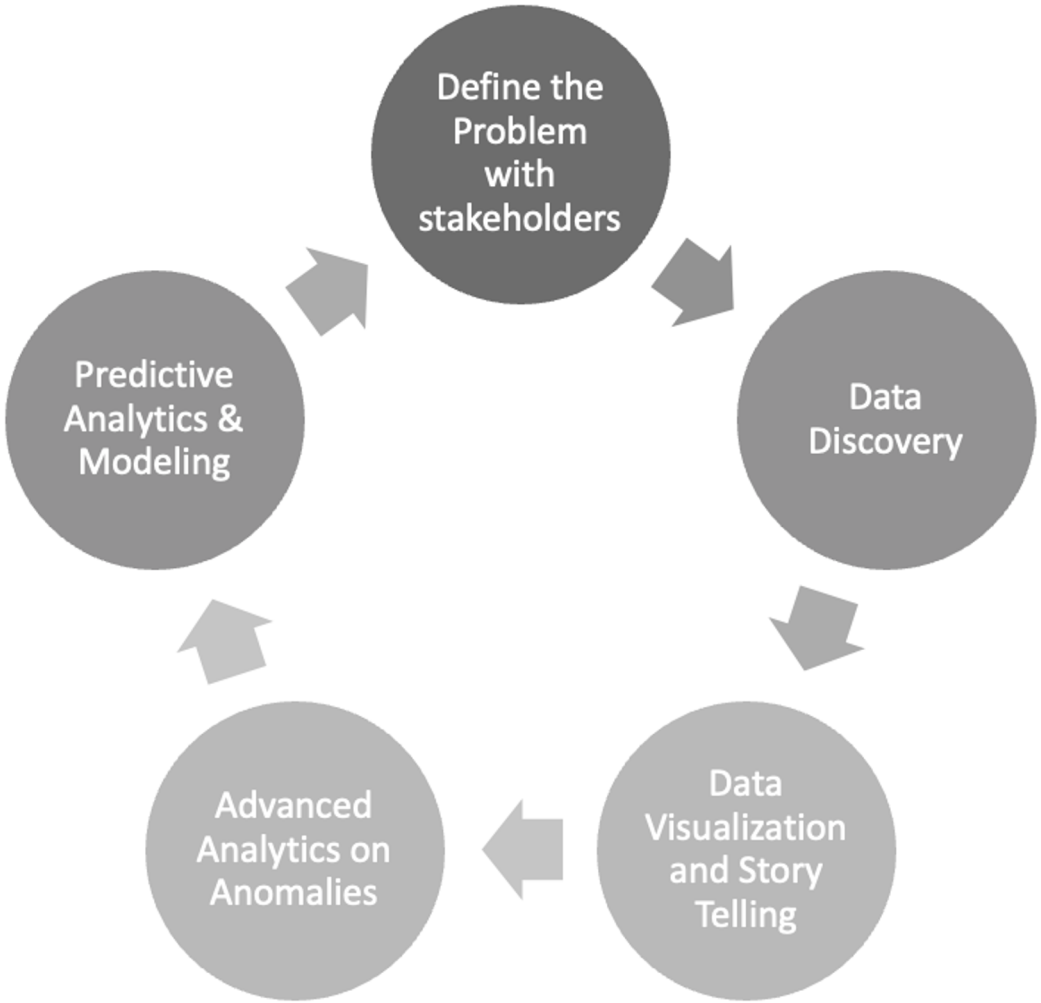
Data Discovery

Story Telling

Advanced Analytics

Predictive Analytics

Ethics

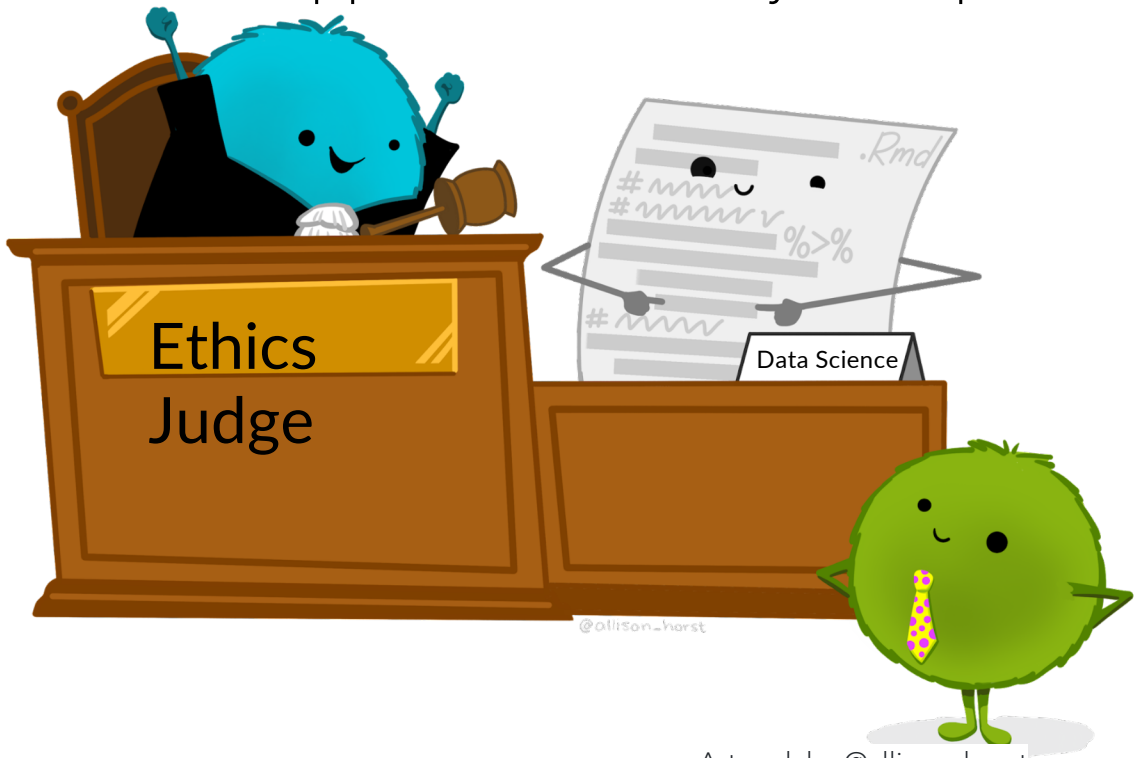


First Do No Harm!

- Presenting that last slide to a client might be reckless
- There isn't really enough data to make the prediction I did, which makes any decisions a client used it for misguided

Opportunities: Ethics is Foundational

Without an ethical approach, none of your capabilities matter



Artwork by @allison_horst

ARE YOU NOT ENTERTAINED?

GET R AND RSTUDIO INSTALLED FOR CLASS ON THURSDAY

FIRST HOMEWORK ASSIGNMENT!


LET THE LEARNING BEGIN!