



# PORTFOLIO MILESTONE

By: Matthew Laken

Spring 2022

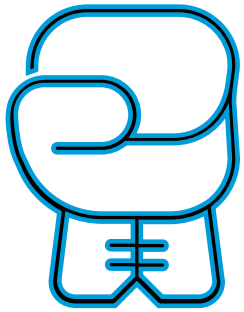
# Who Am I?



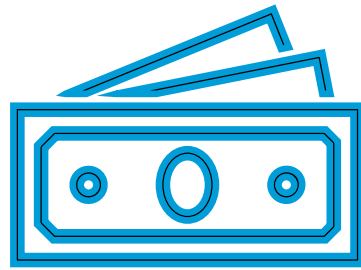
- Undergraduate
  - Graduate of University of Maryland, College Park School of Information Studies, Bachelor of Science in Information Science, May 2020.
- Masters
  - Enrolled in Syracuse University Master of Applied Data Science Program, Fall 2020.
  - Completion of Master of Applied Data Science Program, Spring 2022.
- Objective
  - Expand the application for information and data science by digging deeper.



# My Course Highlights:



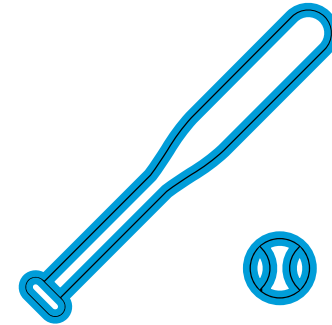
IST 659: Database  
Administration



IST 687: Introduction to  
Data Science



IST 707: Data Analytics



IST:652 Scripting for  
Data Analysis

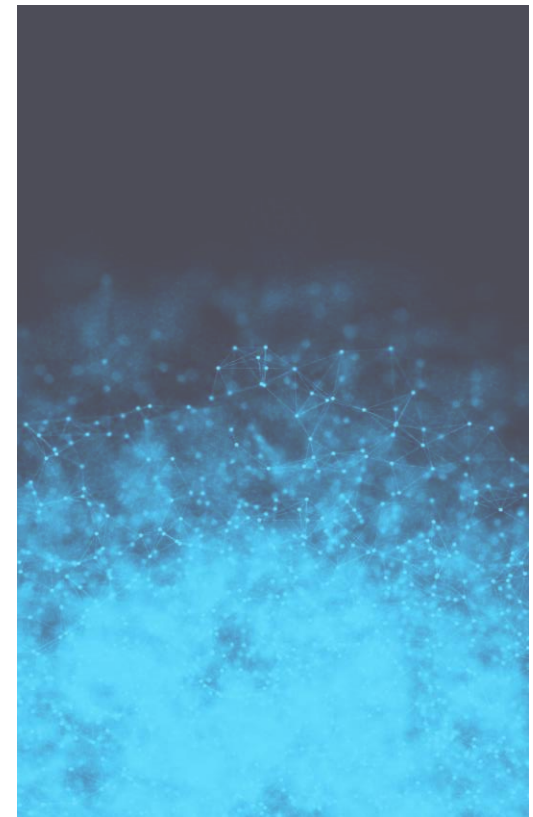
# IST 659: Database Administration

**Goal:** Fundamentals of databases; compiling datasets, building tables, writing queries and storing models.

**Professor:** Chad Harper

**Date:** Fall 2020

**Software:** Microsoft SQL, Access, Excel, R (R Studio)

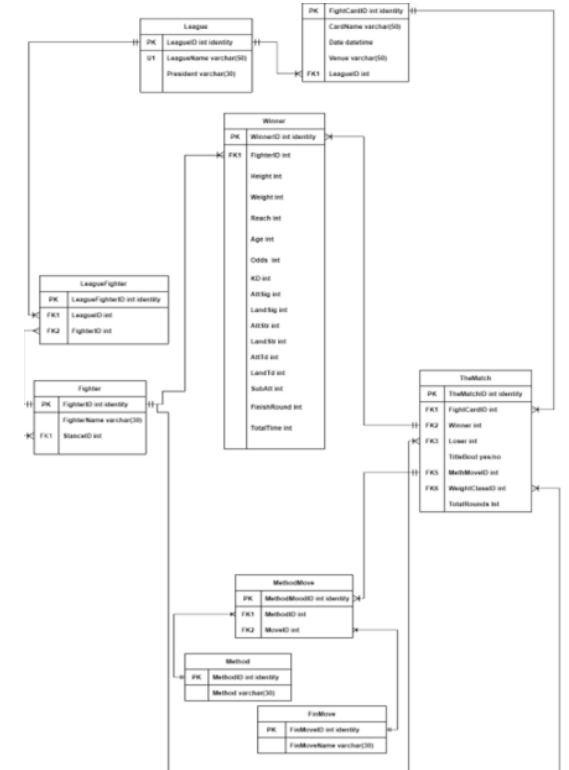


# UFC Relational Database

**Purpose:** Develop a logical relational database to store information on UFC fighters and results of their matches to gain insight on the sport.

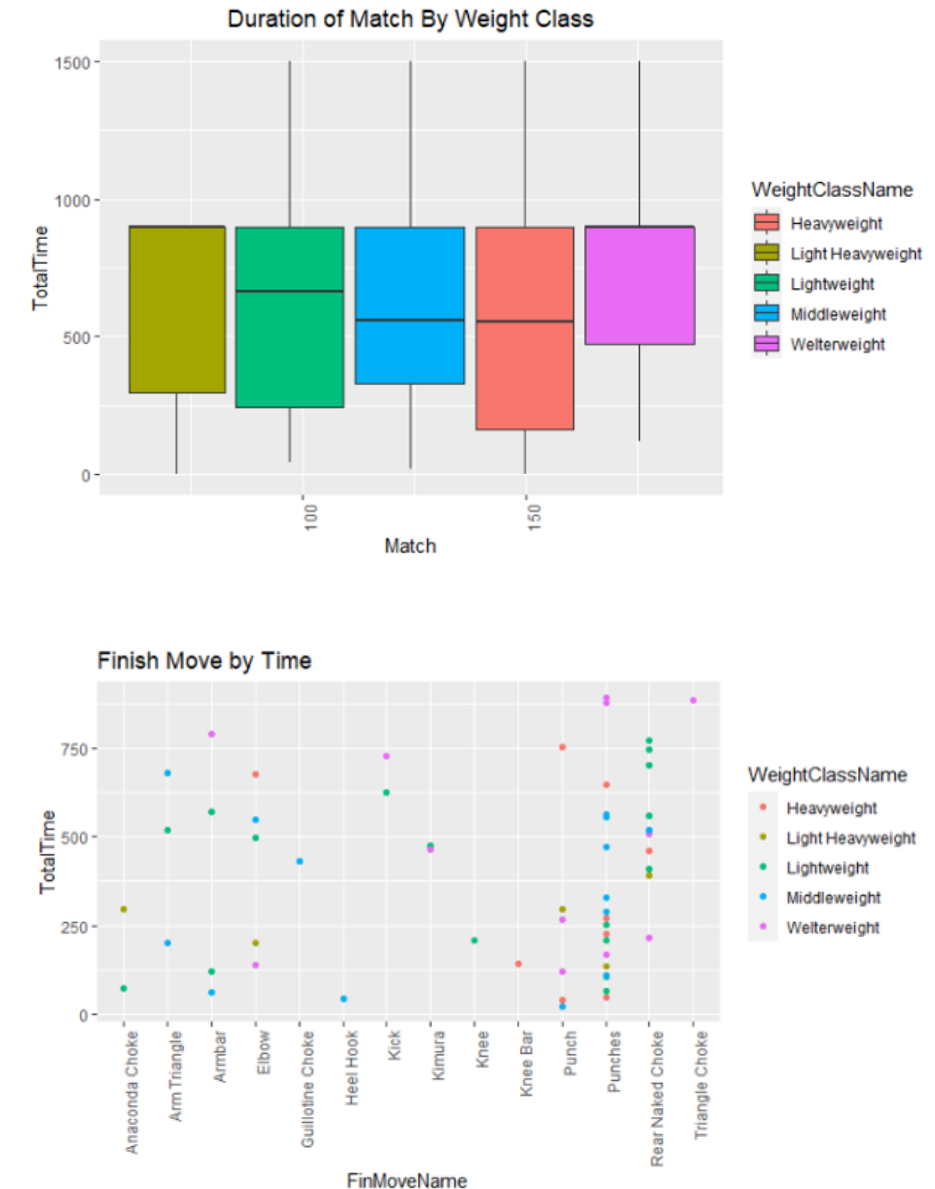
**Data Source:** <http://ufcstats.com/>

**Deliverable:** Comprehensive report including the logical model of the database, questions, views/functions and evaluation.



# Analysis

- Aggregated data by Weight Class:
  - Analyzed aspects of the match to observe the differences between the different divisions.
- Concluded the following:
  - Heavyweight matches end earlier than others, and they throw fewer strikes compared to other divisions.
  - Majority of fights that are stopped early result in the winner by punches.



# Learning Outcomes

- Collect and clean data / develop a database.
- Understand database development lifecycle.
- Construct analyze and solve business problems by using SQL language to create and run queries.
- Enforce integrity of data / database through database design and implementation.

# IST 687: Introduction to Data Science

**Goal:** Exploring the key components of data science by expanding on knowledge of spreadsheets to utilize R to answer business problems.

**Professor:** Jeffery Saltz / Tin Hoang

**Date:** Fall 2020

**Software:** R (R Studio), Microsoft Excel





# Project: Analysis of Cost and Salary Potential of Attending College

**Purpose:** Analyze the cost of going to college and one's salary potential upon graduating to determine the most cost beneficial schools to attend.

**Data Source:**

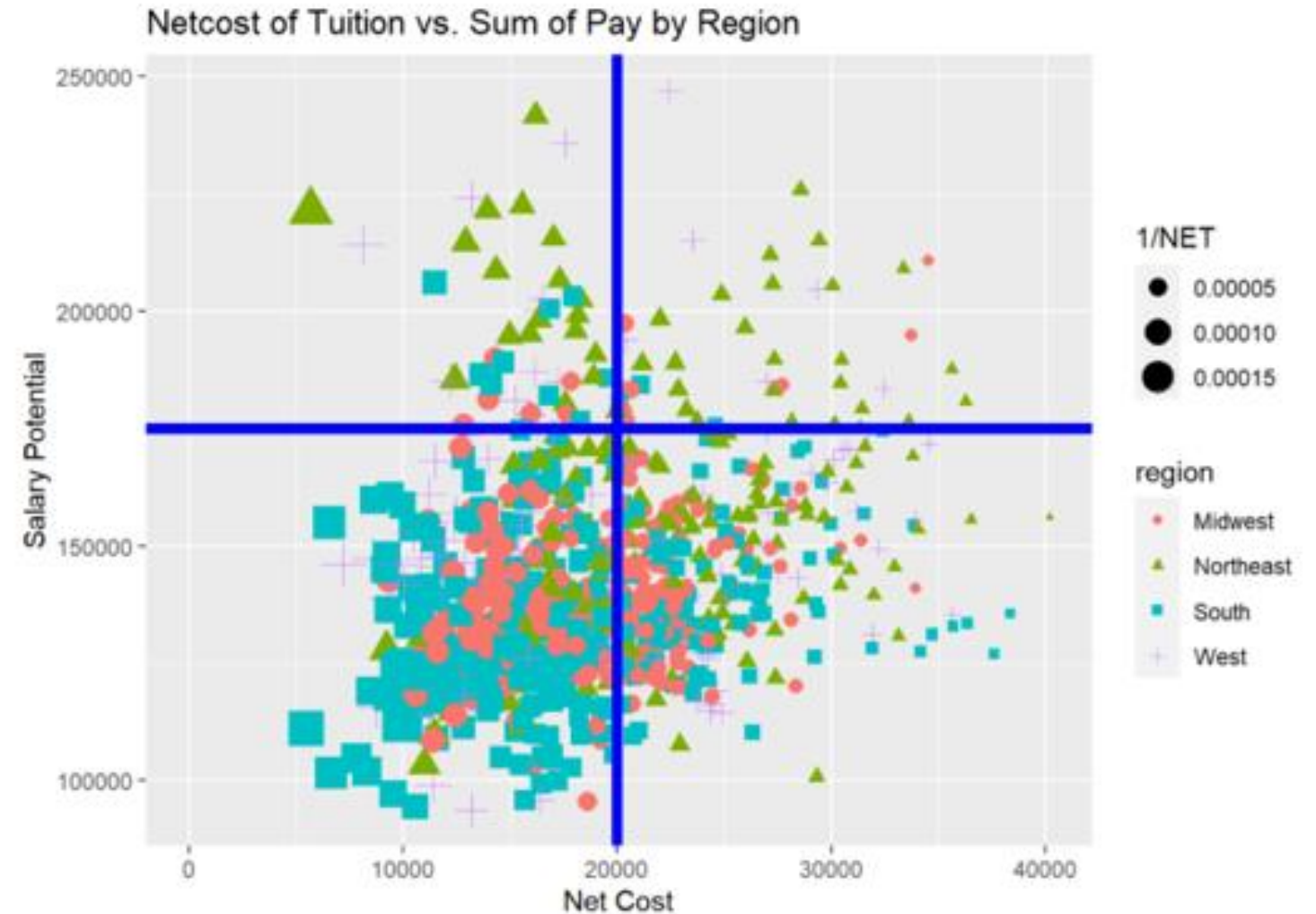
[https://www.kaggle.com/jessemostipak/college-tuition-diversity-and-pay?select=salary\\_potential.csv](https://www.kaggle.com/jessemostipak/college-tuition-diversity-and-pay?select=salary_potential.csv)

**Deliverable:** Comprehensive report including methods of analysis and evaluation.



# Analysis

Ability to determine which schools have the lowest cost and the greatest potential for salary.

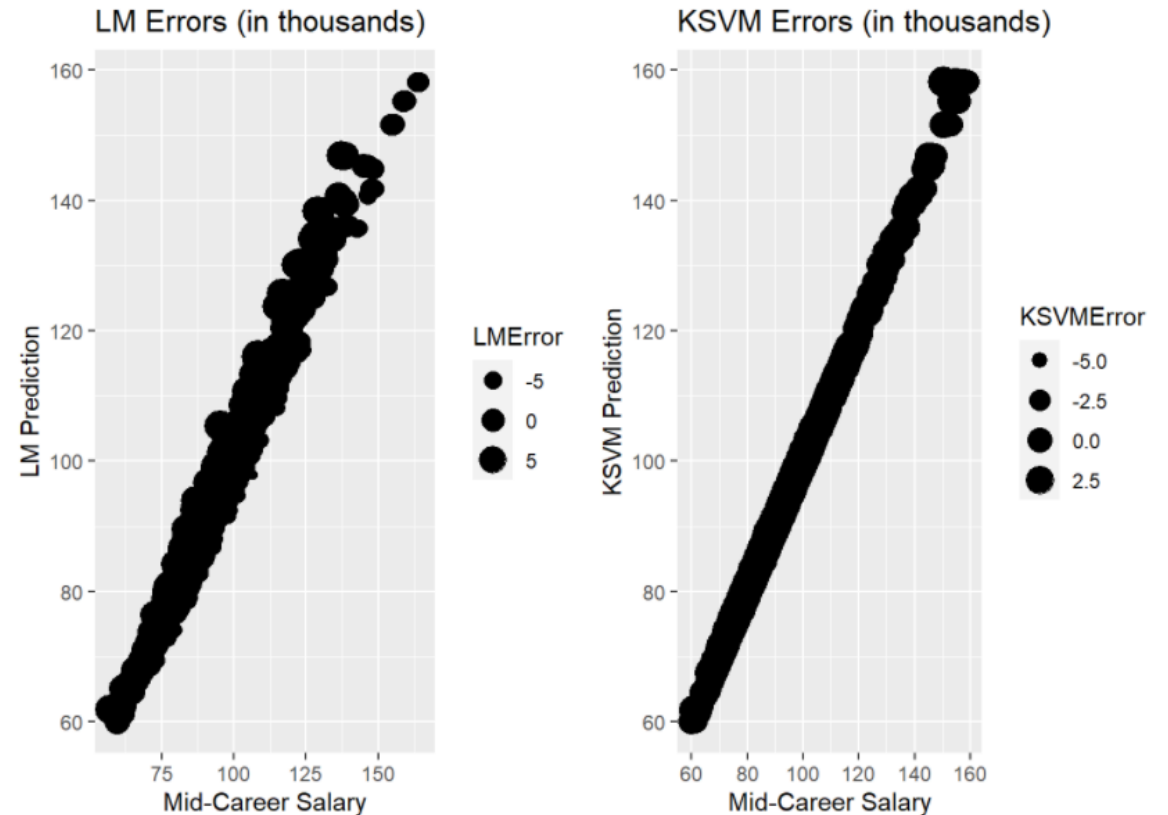


## 1 Rhode Island School of Design	Northeast	RI
## 2 Spelman College	South	GA
## 3 Savannah College of Art and Design	South	GA

# Analysis

Can we predict the mid salary career based on where they go to school?

- KSVM model tracks closer to the true data than the linear model.
- Both models have about the same overall error.
- The narrower range of error makes the KSVM more accurate, but the linear model runs much faster.
- Either model seems sufficient to predict mid-career salary.



# Learning Outcomes

- Fundamentals of data management using R and R-Studio such as data processing, cleaning and merging.
- Identify a problem and the components needed for analyzing and communicating results.
- Collaborate within a team setting to walk through the various stages of a project life-cycle and carry out a finalized product.
- Determine appropriate techniques for analyzing data.



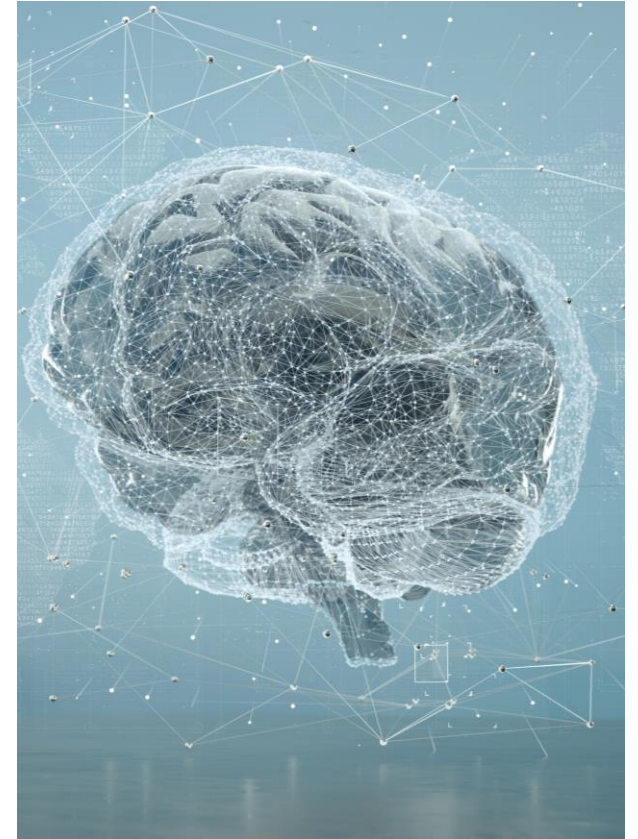
# IST 707: Data Analytics

**Goal:** Utilize the tools and methods of data mining while introducing its real-world application to solving problems in business.

**Professor:** Stephen Wallace

**Date:** Spring 2021

**Software:** R (R Studio), Weka, Orange, Alteryx, Excel



# Project: Marvel vs D.C. Analysis

**Purpose:** Experiment with various classification methods by comparison, analyzing data on Superheroes and Villains from Marvel and D.C Comics.

**Data Source:** <https://www.kaggle.com/danniellr/marvel-superheroes>

**Deliverable:** Comprehensive report detailing the methods, proposed questions, and results of the experiment.



# Analysis

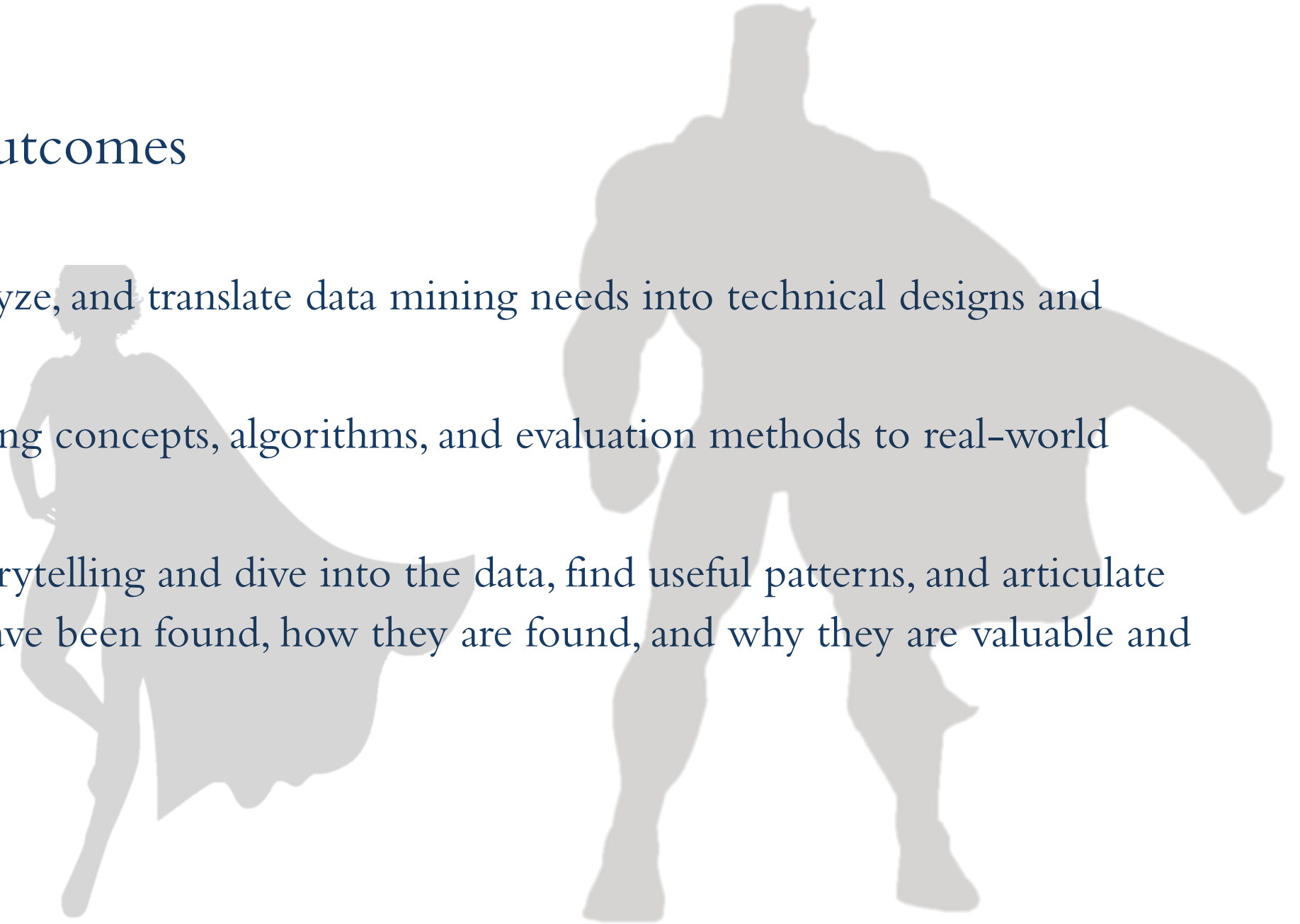
Using SVM, kNN, Random Forest and Naïve Bayes to determine which model is most accurate in classifying the characters.

- 3-fold cross validation to determine how they can predict publisher and alignment for the power dataset, the rating statistics dataset, and the combined dataset.
- SVM algorithm yielded the greatest percent of classification accuracy.

Experiment	SVM CA
Experiment 5: Predict Publisher with Master Dataset	69.1%
Experiment 6: Predict Alignment with Master Dataset	68.2%

# Learning Outcomes

- Document, analyze, and translate data mining needs into technical designs and solutions.
- Apply data mining concepts, algorithms, and evaluation methods to real-world problems.
- Employ data storytelling and dive into the data, find useful patterns, and articulate what patterns have been found, how they are found, and why they are valuable and trustworthy.





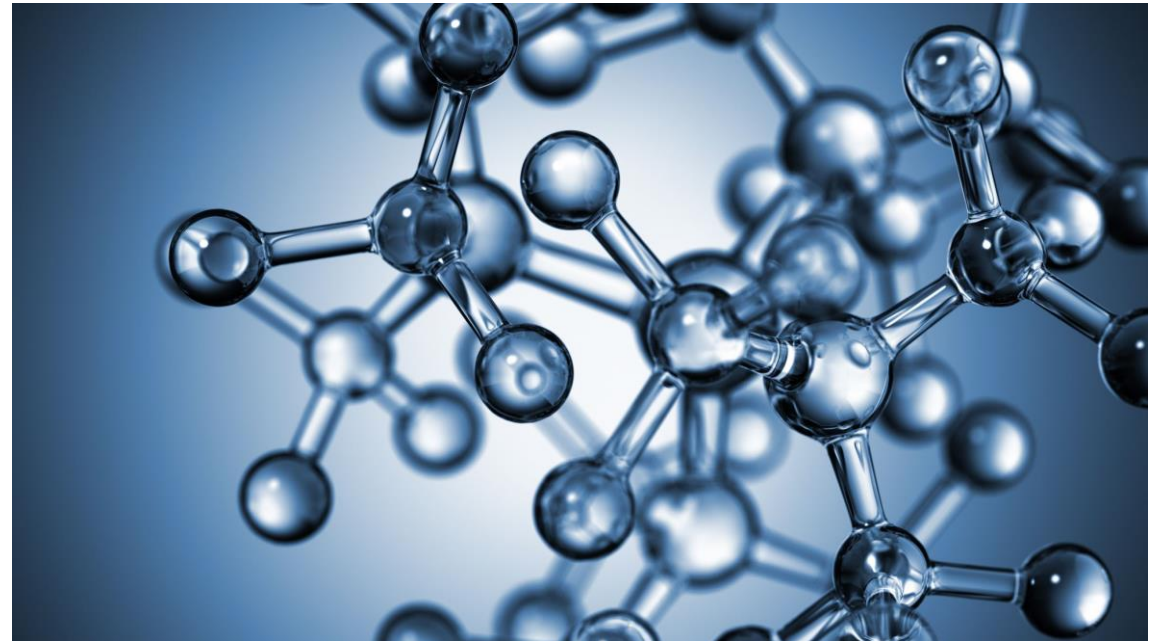
# IST 652: Scripting for Data Analysis

**Goal:** Comprehend the tools and skills of scripting to utilize data science while solving real-world problems by analysis or visualization.

**Professor:** Dr. Debbie Landowski

**Date:** Summer 2021

**Software:** Python, Excel



# Project: MLB Strikeout Investigation



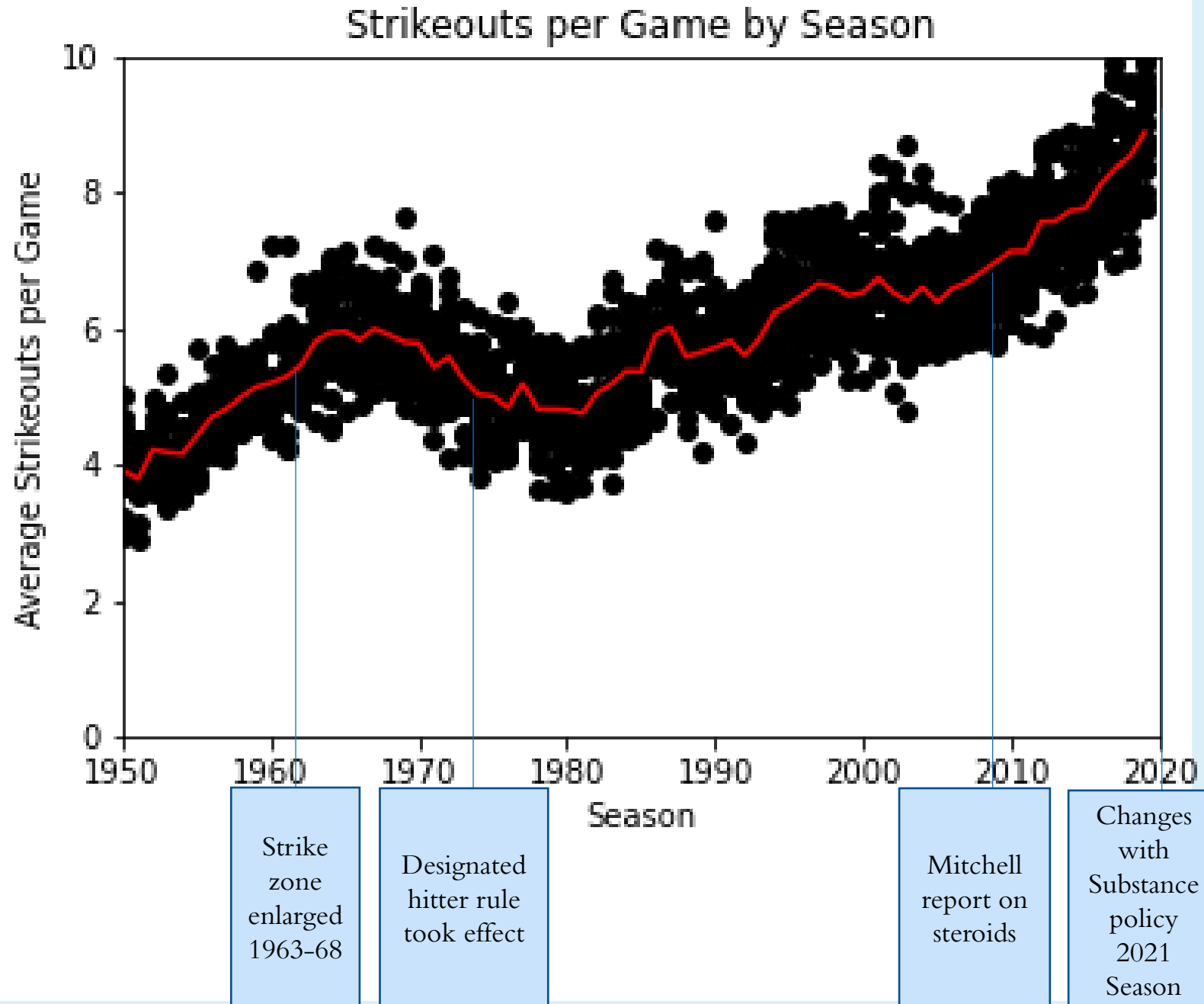
**Purpose:** Collaborate with peer to write Python scripts to investigate data and communicate findings utilizing the techniques for processing data. This is demonstrated using baseball metrics from the pybaseball package analyzing strikeouts throughout history.

**Data Source:** <https://pypi.org/project/pybaseball/>

**Deliverable:** Comprehensive report detailing an initial plan for the group project. Complete with the demonstration of written Python scripts to access, prepare and use data to produce data summaries, lists, and other structures.

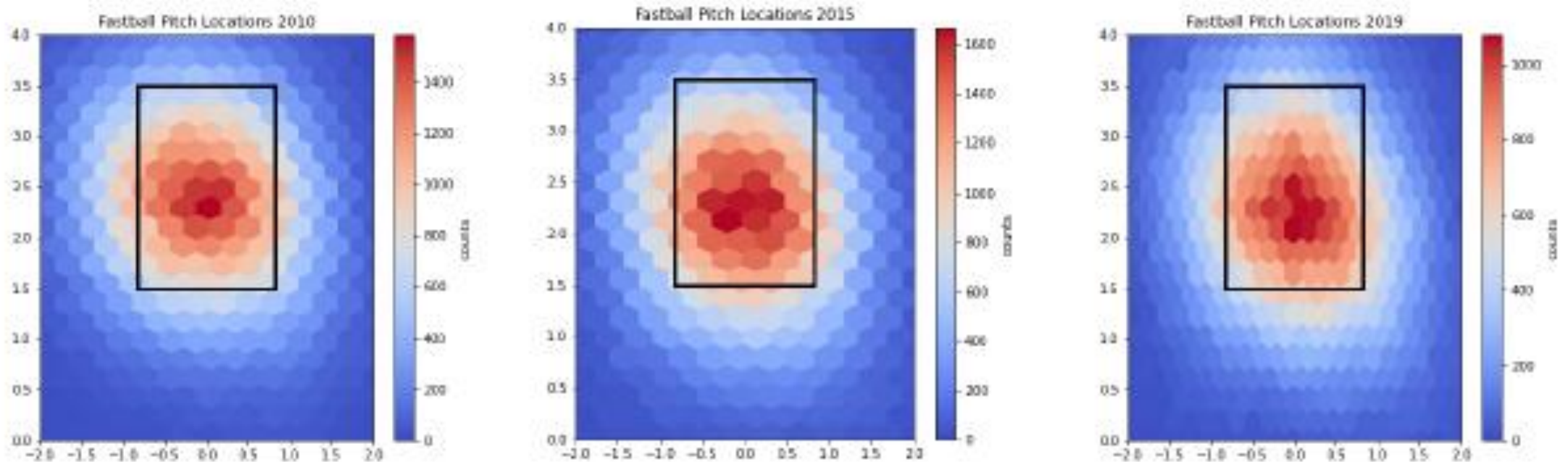
## History

- Over 70+ years of historical data in the MLB season.
- More than 700,000 pitches are thrown in each season.
- Strikeouts have risen significantly since 1980 and continue to increase.



# Analysis

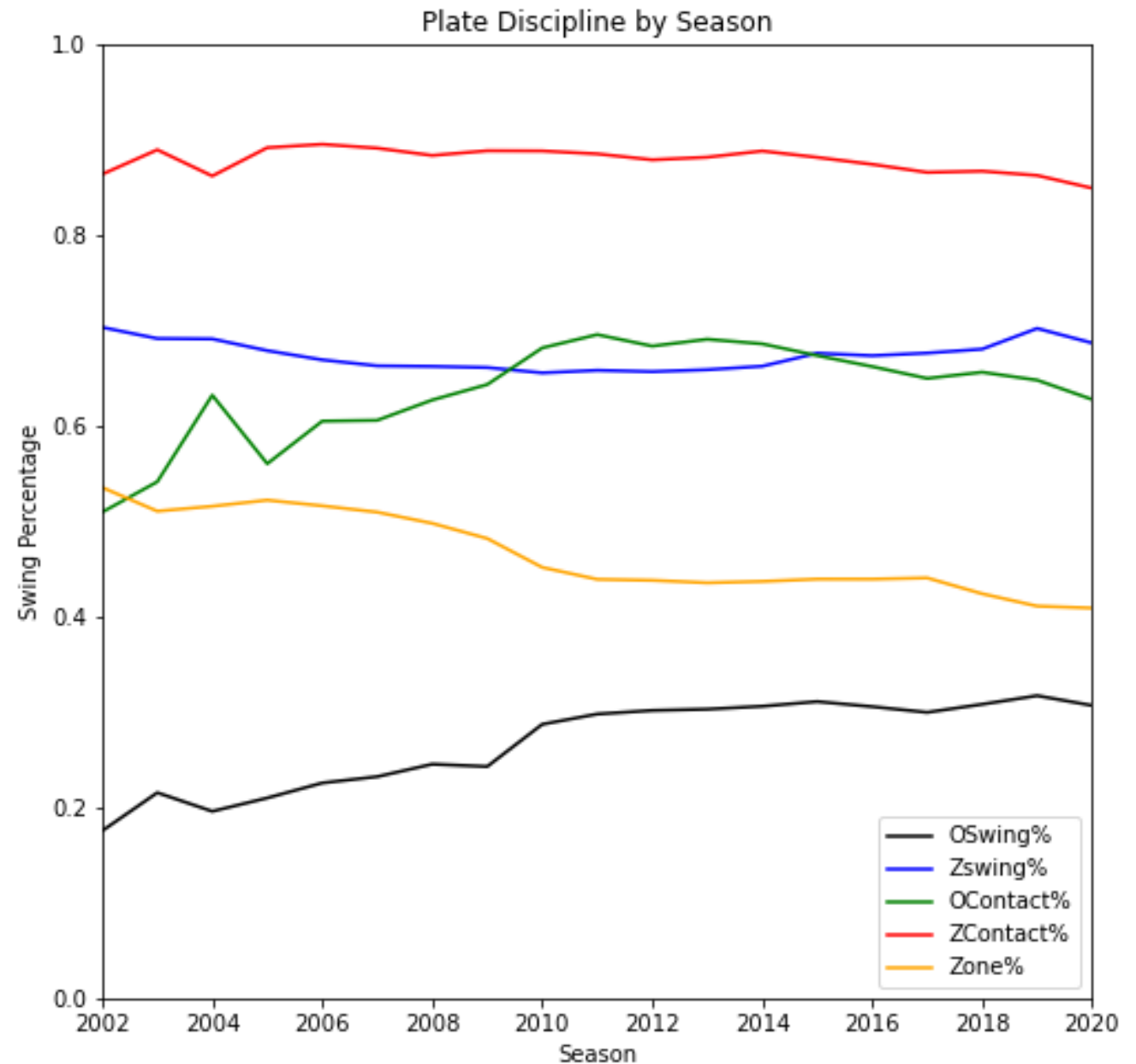
The rise in strikeouts and homeruns can be attested to the way pitchers have gradually raised where they throw fastballs in the strike zone over the past 10 years.





## Analysis

- Aggregated data on the players by year to determine how disciplined hitters are.
- More pitches outside the zone = swinging at more pitches out of the zone.
- Rising strikeouts and decreasing batting averages.
- Pitchers are forcing hitters to swing at more pitches outside of the strike zone thus more strikeouts, lower batting average.



# Learning Outcomes

- Develop python scripts to access information from files of structured data, access files in semi-structured data, and to define and find patterns in unstructured data.
- Prepare and transform data to produce data summaries, lists, and networks utilizing appropriate software packages that can be integrated into the problem solution.
- Analyze and solve data access problems for the three types of data and to find and frame real world data questions and show how they can be answered with data.

# Thank you!

To review these projects in detail as well as additional information about my studies in this Masters program please check out my [GitHub](#) and [LinkedIn](#). I am always open for new connections and conversations!

