Syracuse University

Masters of Science

Applied Data Science

Portfolio Milestone

Peter Mathews

SUID: 455606416

March 19th, 2021

https://github.com/pete_math/portfolio_milestone

Outline of Presentation

I. About Me

II. Introduction

III. Learning Objectives and Example Projects

1. Describe a broad overview of the major practice areas of data science.
2. Collect and organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.
5. Develop a plan of action to implement the business decisions derived from the analyses.
6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice (e.g., privacy).

IV. Conclusion & Thank You!

About Me

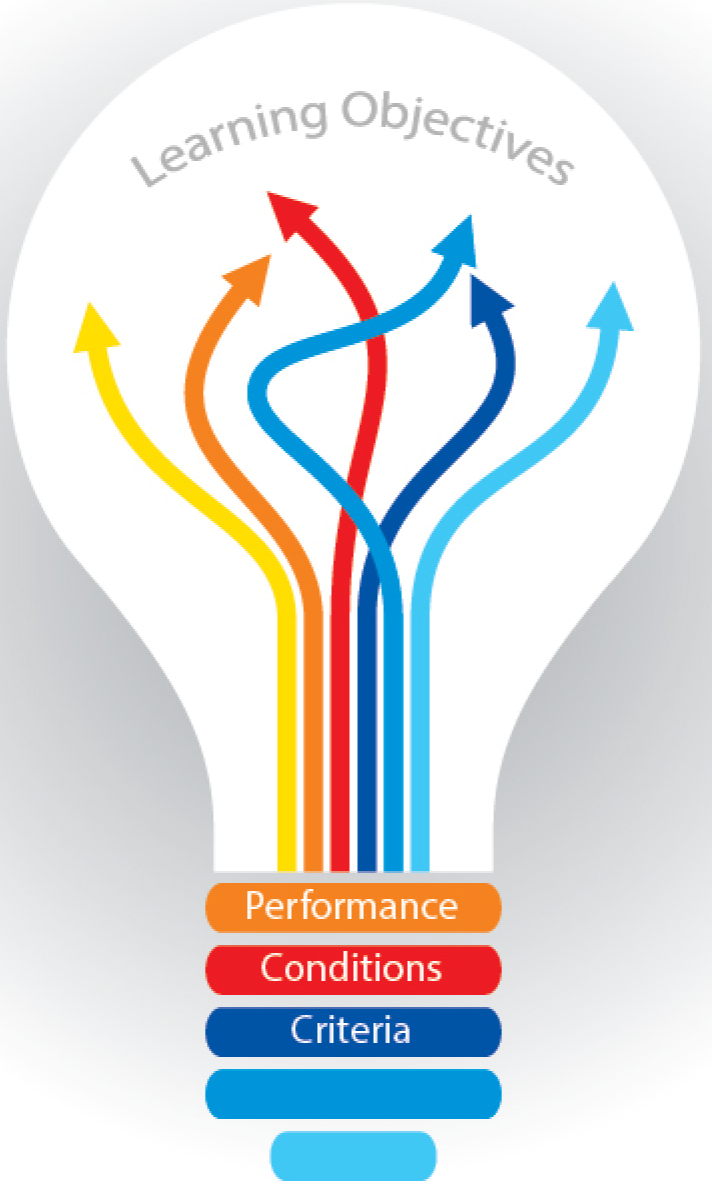
- Born and Raised in Buffalo, NY. Currently reside in NYC.
- Live amateur show junkie: Plays, comedy and jazz are my top 3.
- Enjoy spending time with my wife and daughter.
- Why did I decided to pursue my masters in data science?

Introduction

- The Applied Data Science curriculum at Syracuse University's School of Information Studies provides students the opportunity to collect, manage, analyze, develop, and implement insights using data from a multitude of disciplines using various tools and techniques.
- In general, the curriculum, has provided the relevant and applicable knowledge, training, resources, and academic environment for its students to progress and impart skills that exceed the basic roles of a data scientist
- Using project based learning, students of the School of Information Studies are able to acquire crucial experience in detecting patterns in data, developing and implementing alternative data approaches, demonstrating communication skills that focus on explaining technical outcomes to non-technical recipients, and discussing the ethical scope of data science through courses related to data privacy and policy.

Learning Objectives And Example Projects

The following slides will go through all 7 learning objectives and provide detailed examples of how these objectives were met through a brief synopsis work and projects.

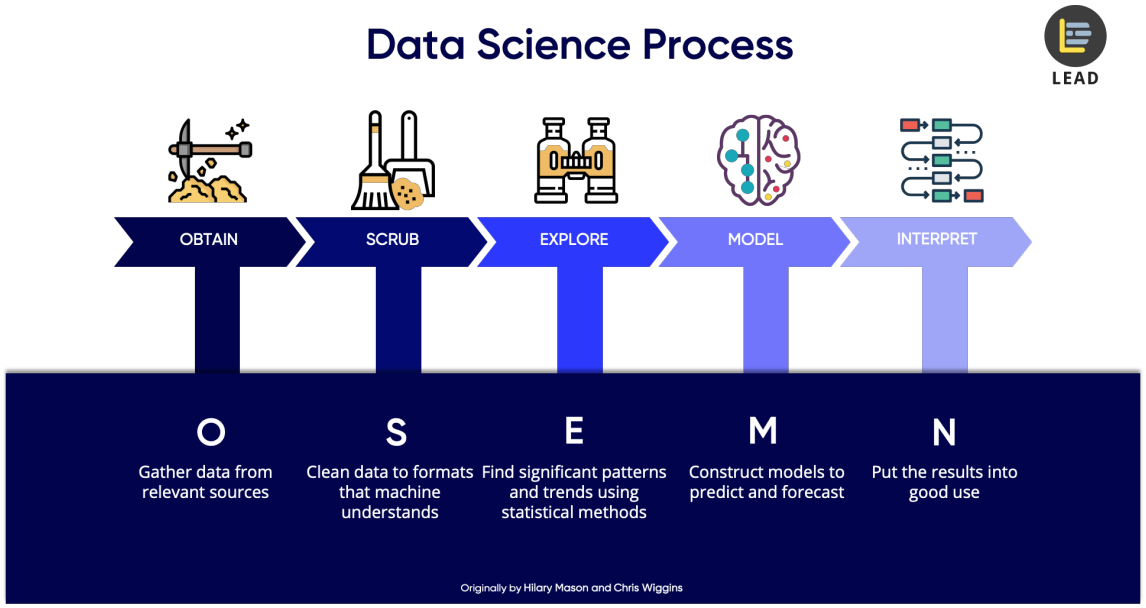


Seven Learning Objectives of the Applied Data Science Program:

1. Describe a broad overview of the major practice areas of data science.
2. Collect and organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.
5. Develop a plan of action to implement the business decisions derived from the analyses.
6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice (e.g., privacy).

Objective 1:

Describe a broad overview of the major practice areas of data science.



Objective 1: Describe a broad overview of the major practice areas of data science.

- In simple terms, a data scientist's job is to analyze data for actionable insights.
- Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.
- Data science practitioners apply machine learning algorithms to text, images, video, audio, and more to build data related products and services.
- In turn, these systems generate actionable insights which ideally will translate business value

Objective 1: Describe a broad overview of the major practice areas of data science.

What does the work really entail?

1. Ask the right questions to begin the discovery process
2. Acquire data
3. Process and clean the data
4. Integrate and store data
5. Initial data investigation and exploratory data analysis
6. Choose one or more potential models and algorithms
7. Apply data science techniques, such as machine learning, statistical modeling, and artificial intelligence
8. Measure and improve results
9. Present final result to stakeholders
10. Make adjustments based on feedback
11. Repeat the process to solve a new problem

Objective 2:

Collect and Organize Data

Objective 2: Collect and Organize Data

Structured and Unstructured data

- Data comes in multiple formats.
- Two primary ways of describing these formats are structured and unstructured data.
- Structured data is highly-organized and formatted in a way so it's easily searchable
- Unstructured data has no pre-defined format or organization, making it much more difficult to collect, process, and analyze.

Objective 2: Collect and Organize Data

Administration Concepts and Database Management (IST-659)

- This project aimed to design a database for a web-based commercial real estate deal (sales and loans) tracking system for use by commercial real estate brokers (both sales and financing).
- I successfully demonstrated skills in data management and organization by taking collected information from a web form to build a SQL database using Microsoft Sql Server.
- Once collected, I was able to query the database using SQL to answer data questions and gain insight.

Objective 2: Collect and Organize Data

Administration Concepts and Database Management (IST-659)

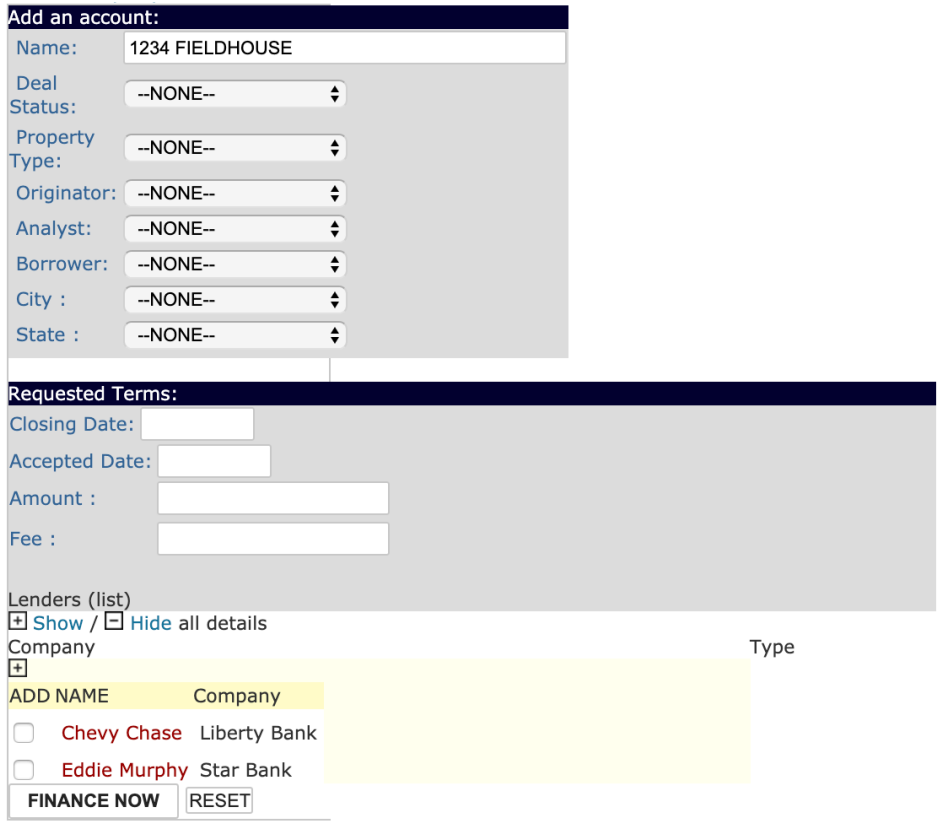


Figure 1: Organization , GUI Front-End of Database

```
> SqlSelectStatement <-
+ "select o.name, c.first_name, c.last_name, avg(d.fee) as fee_average
+  from deal d
+  left join quote q on q.deal_id = d.deal_id
+  left join contact c on c.contact_id = q.contact_id
+  left join organization o on o.organization_id  = c.organization_id
+   where o.name is not null
+ and year(d.date_active)='2019'
+  group by o.name, c.first_name, c.last_name order by avg(d.fee) desc;
+ "
>
> avgResult <- sqlQuery(myconn,SqlSelectStatement)
> avgResult
                  name first_name last_name fee_average
1  Global Markets, Inc.     Steven      Allen     1143200
2         River Capital     George       Bush     1143200
3             Star Bank      Chevy      Chase      422533
4       Financial Group       Dave  Chappelle      403200
5 REIT Investment Trust        Jim     Carrey      330200
6          Liberty Bank      Eddie     Murphy       53200
```

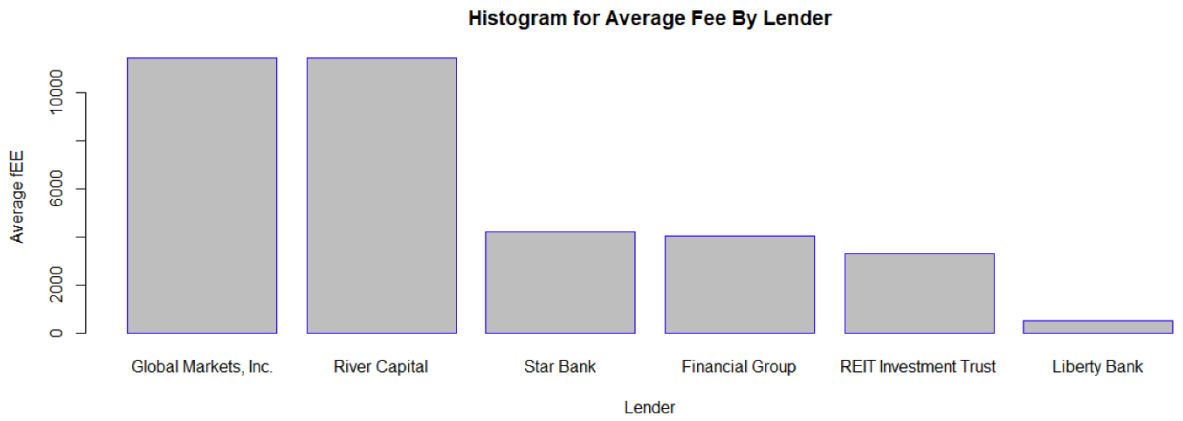
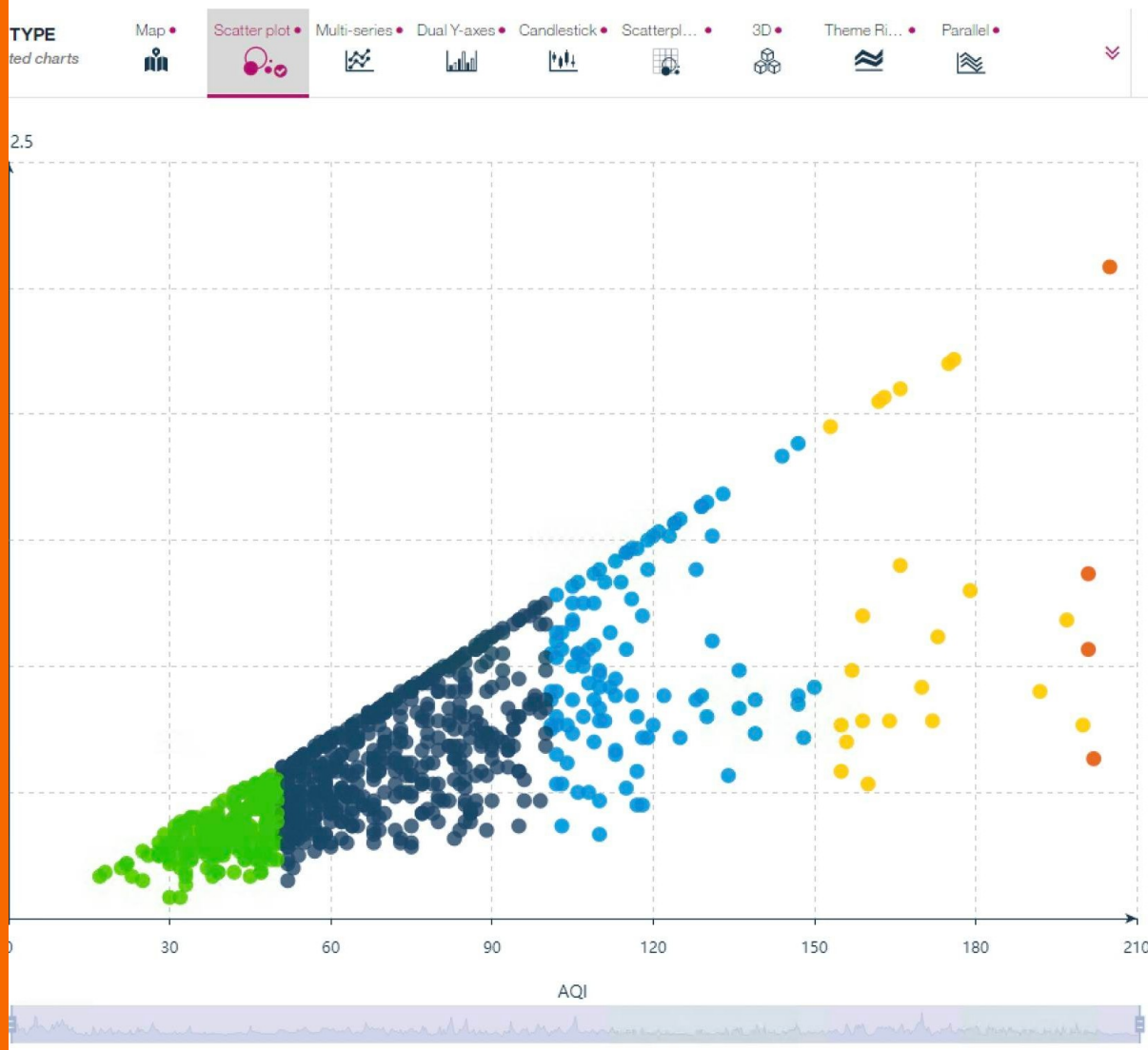


Figure 1: Data Organization

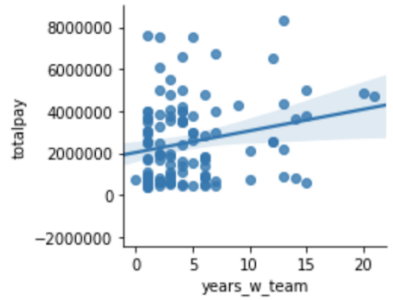
Objective 3:

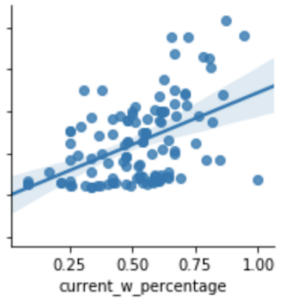
Identify Patterns in Data Via: Visualization, Statistical Analysis, And Data Mining.

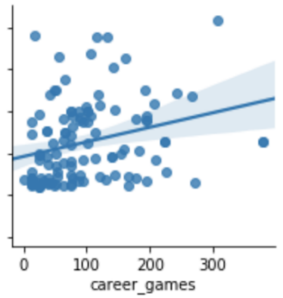


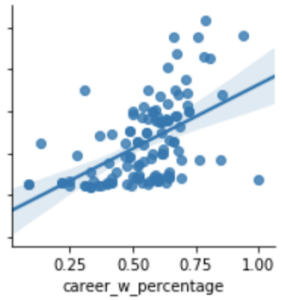
Objective 3: Identify Patterns in Data Via: Visualization, Statistical Analysis, And Data Mining.

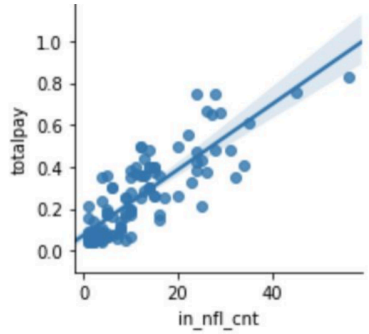
Quantitative Analysis of a NCAA Coach's Salary

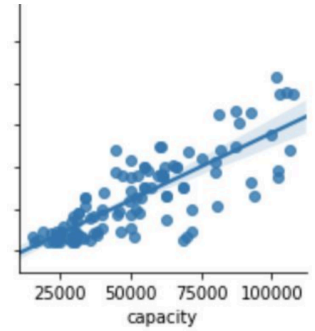




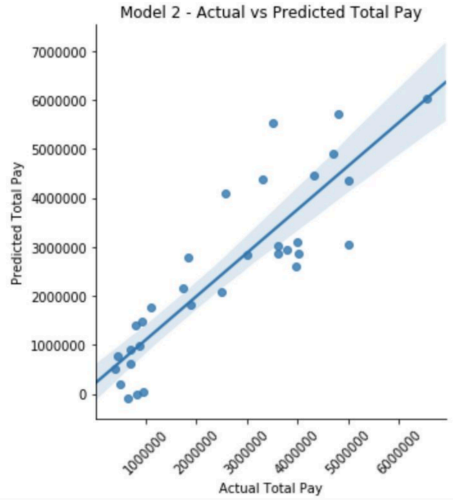









```
Model 2 R-squared:  0.7512790359847417
Model 2 RMSE:  854346.291109091
Model 2 MAE:  690290.407867036
```



Conference	Sum of TotalPay	Count of Coach	Min of TotalPay	Max of TotalPay	nfl_roster	teams	avg
SEC	$65,008,379.00	14	$2,350,000.00	$8,307,000.00	339	14	$4,643,455.64
Big Ten	$60,256,192.00	14	$1,830,000.00	$7,600,000.00	253	14	$4,304,013.71
ACC	$48,073,154.00	14	$1,831,580.00	$6,543,350.00	215	11	$4,370,286.73
Big 12	$36,163,301.00	10	$1,701,109.00	$5,500,000.00	131	10	$3,616,330.10
Pac-12	$34,681,433.00	12	$1,500,000.00	$4,377,500.00	189	12	$2,890,119.42
AAC	$16,562,677.00	11	$1,000,000.00	$2,600,000.00	100	14	$1,183,048.36
Mt. West	$12,071,254.00	12	$486,504.00	$1,800,000.00	65	12	$1,005,937.83
C-USA	$10,921,091.00	14	$500,000.00	$1,425,000.00	59	14	$780,077.93
MAC	$6,959,534.00	12	$412,500.00	$1,125,000.00	46	12	$579,961.17
Sun Belt	6506500	10	$390,000.00	$850,000.00	21	10	$650,650.00
Ind.	$4,929,080.00	6	$419,640.00	$2,129,638.00	35	6	$821,513.33

Objective 4:

Develop Alternative Strategies Based on The Data

Objective 4: Develop Alternative Strategies Based on The Data

Big Data Analytics (IST-718)

- To better understand the opioid epidemic in New York and to provide a tool to assist health and law enforcement officials in managing the crisis as this is an ongoing, urgent problem throughout our society
- At close to 25 gigabytes, this was a very large dataset. In order to fit the dataset in Google Colab, the original dataset from the DEA ARCOS database was narrowed down to 8.88 million observations by filtering the data to focus only on Hydrocodone and Oxycodone's drugs.
- The results were both accurate and actionable. However, training and testing the model took significant time. If we wanted to provide a real-world tool to law enforcement, we'd need to find a way to decrease training.
- Using Rapids.ai, I was able to retain the same level of accuracy from the model, however, cut the training time down from 4 mins to 3 secs.

Objective 4: Develop Alternative Strategies Based on The Data

Big Data Analytics (IST-718)

BDA with Spark

```
# Run forecasting models for all zip codes with data from 2007-2017
%%time
results = (
  zip_history
    .groupBy('Zip')
    .apply(forecast_zip)
    .repartition(sc.defaultParallelism, ['Zip'])
    ).na.fill(0).cache()

results.show(20)
```

ds	Zip	y	yhat	yhat_upper	yhat_lower	mom_growth
2007-01-31	10301	22086.236	17436.291	29148.967	5674.7827	0.0
2007-02-28	10301	29689.55	19794.701	31451.184	8671.515	0.13525872
2007-03-31	10301	16243.479	21351.828	32487.416	9660.714	0.078663856
2007-04-30	10301	15501.518	19113.617	30961.275	7086.557	-0.104825325
2007-05-31	10301	16640.072	18206.154	30347.621	6949.1294	-0.047477223
2007-06-30	10301	19658.19	20410.861	31969.818	8988.219	0.121096715

```
CPU times: user 30 ms, sys: 13.3 ms, total: 43.3 ms
Wall time: 3min 13s
```

Figure 1: Organization , GUI Front-End of Database

BDA with RAPIDS

IMPORT LIBS

```
[ ] # Install RAPIDS
    !git clone https://github.com/rapidsai/rapidsai-csp-utils.git
    !bash rapidsai-csp-utils/colab/rapids-colab.sh

    import sys, os

    dist_package_index = sys.path.index('/usr/local/lib/python3.6/dist-packages')
    sys.path = sys.path[:dist_package_index] + ['/usr/local/lib/python3.6/site-packages'] + sys.path[dist_package_index:]
```

RAPIDS

```
[ ] %%time
    months_to_predict = 12
    cu_hw = cuES(data, seasonal='multiplicative', seasonal_periods=12, ts_num=877)
    cu_hw.fit()
    cu_preds = cu_hw.forecast(months_to_predict)
    cu_hw_forecast = cu_preds
```

```
CPU times: user 3 s, sys: 584 ms, total: 3.59 s
Wall time: 3.59 s
```

Figure 1: Data Organization

Objective 5:

Develop A Plan of Action to Implement the Business Decisions Derived from The Analyses

Objective 5: Develop A Plan of Action to Implement the Business Decisions Derived from The Analyses.

Big Data Analytics (IST- 718)

- The goal of this analysis is to identify the 3 top zip codes for Syracuse REIT (SREIT) to evaluate for investment. Given the specific function of SREIT as a builder and seller of single-family homes, this analysis will serve as a potential roadmap for future investments.
- Using FaceBook's Prophet time-series library in conjunction with Apache Spark, I was able to identify the 3 top zip codes for investments and present my findings in a clear manner

Value	Variable	Acceptable Value Ranges:		
Average Passing Touchdowns	0	0 to 5	Projected Fantasy Points:	11.537748
Average Rushing/Receiving Touchdowns	0.5	0 to 4		
Average Rushing Attempts/Targets Per Game	15	0 to 50		
Fantasy Player's Team (Determining Coach Tier)	ARI	0		
Opposing Defense (Determining Defensive Index)	CHI			
Average Market Share of Offense	0.2	0 to 1		

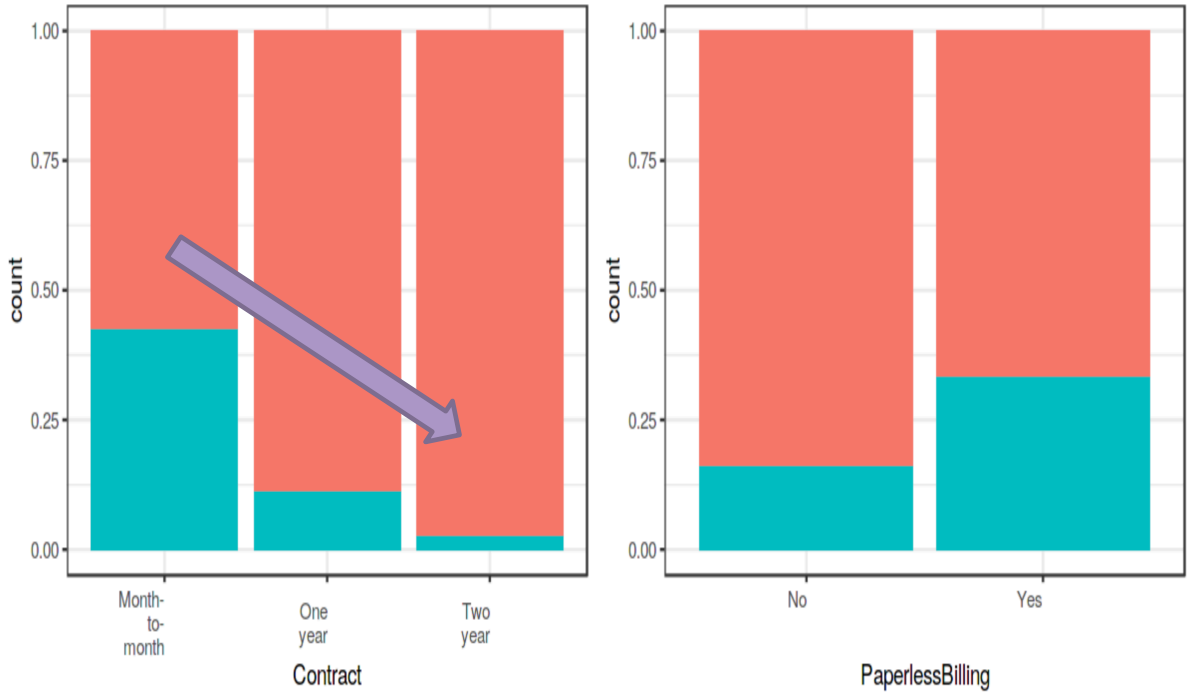
Objective 6:

Demonstrate Communication Skills Regarding Data and Its Analysis for Managers, IT Professionals, Programmers, Statisticians, And Other Relevant Professionals in Their Organization.

Objective 5: Develop A Plan of Action to Implement the Business Decisions Derived from The Analyses.

Marketing Analytics (MAR-653)

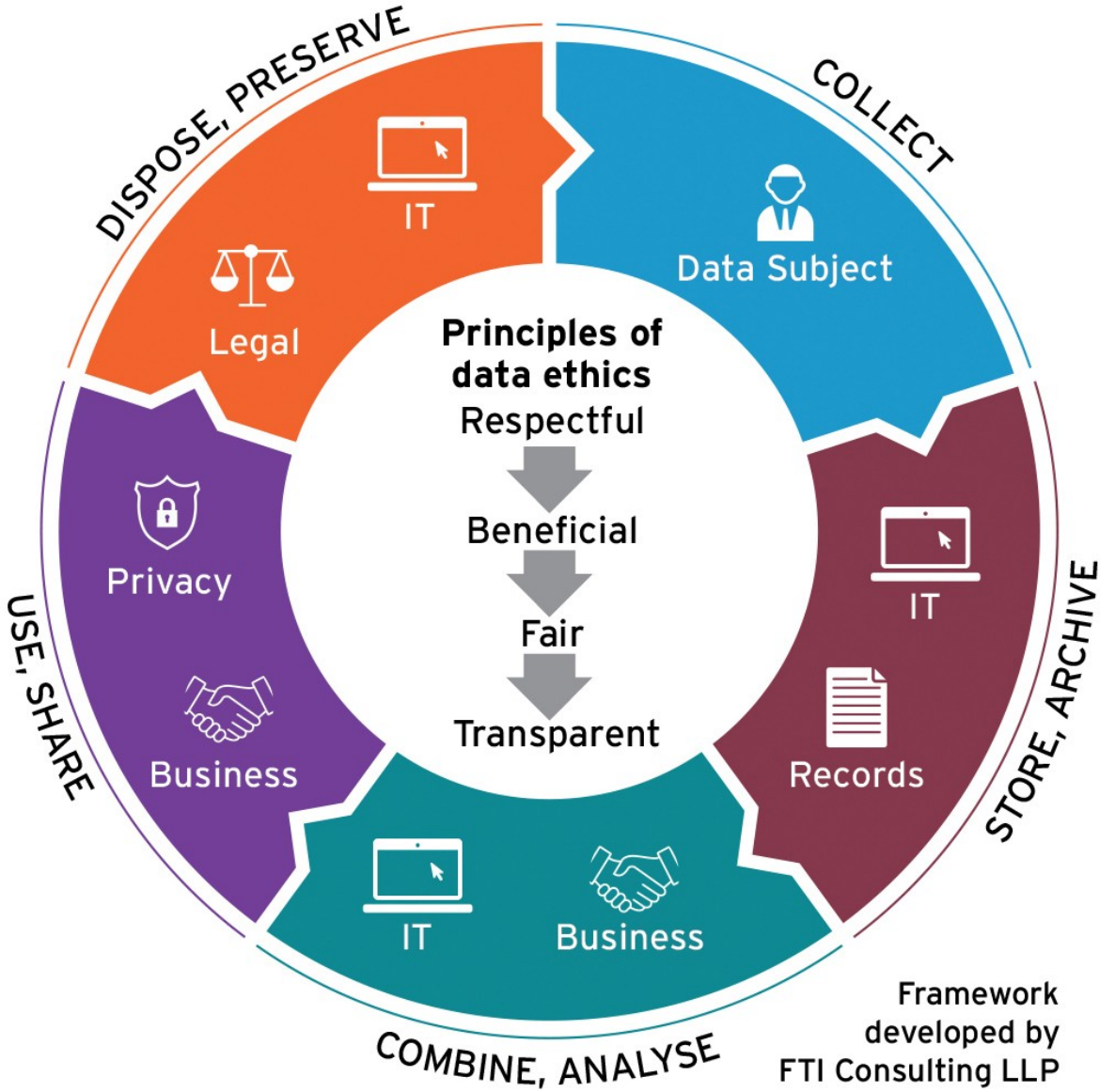
- If a customer has a one-year contract and not using paperless billing, they are unlikely to churn
- If a customer has a month-to-month contract, and is in the tenure group of 0-12 months, and uses paperless billing, they are more likely to churn (78% chance of churn)
- To remain competitive, we recommend locking customers into a contract of at least one year, ensuring terms are agreeable and attractive to customers, to reduce the likelihood of churn.
- We recommend focusing on demographic changes in needs of telecommunications companies, as older generations may be reliable customers for the short terms, while younger generations may churn more quickly in favor of other methods of communication.



As contract length increases, likelihood of churn decreases

Objective 7:

Synthesize the Ethical Dimensions of Data Science Practice



Objective 7: Synthesize the Ethical Dimensions of Data Science Practice.

How can we address privacy, transparency and fairness in our process?

Data Scientist need to conscientiously prepare to address the ethical problems that appear during the use of data

Is our data secure? Have we insured privacy to the best of our ability?

Have we received user consent? Is our process transparent when users initially provide data to us?

Have we studied and understood possible sources of bias in our data? Is our training data fair and representative?

Syracuse University

Thank You!

Overall, the M.S. in Applied Data Science has been a challenging but rewarding experience. Not only did it provide the necessary steps for my future, it allowed me to make lasting friends and connections with fellow students.

This program really represents the springboard to my next stage in my career!