Syracuse University

MS-ADS Portfolio Milestone Presentation

School of Information Studies

Bing-Je Wu

Bwu117@syr.edu

Syracuse University

Agenda

- Introduction
- ADS Program Course Overview
- What is Data Science
- Data Science Life Cycle
  - 1. Business Understanding
  - 2. Data Acquisition & Understanding
  - 3. Modeling
  - 4. Deployment
- Conclusion and Reflection

THE iSCHOOL Syracuse University

Bing-Je Wu

- Born and grew up in Taiwan
- Bachelor degree in Mathematics
- Master degree in Industrial Engineering
- Master of Science in Applied Data Science April 2019 Cohort
- Have more than 3 years working experience in petrochemical industry
- Currently doing as Internal Auditor and Interim Business Analyst

Program Course Overview

Primary Core	Analytics Application Core	Elective Courses
IST 687 - Introduction to Data Science	MAR 653 - Marketing Analytics	**IST 652 - Scripting for Data Analysis**
MBC 638 - Data Analysis and Decision Making		**IST 772 - Quantitative Reasoning for Data Science**
IST 659 - Data Administration Concepts and Database Management (transferred)		**IST 769 - Advanced Database Administration Concepts and Database Management**
SCM 651 - Business Analytics		IST 664 - Natural Language Processing
IST 707 - Data Analytics		**IST 736 - Text Mining**
IST 718 - Big Data Analytics		

What is Data Science?

Data science is a “concept to unify statistics, data analysis, machine learning and their related methods” in order to “understand and analyze actual phenomena” with data.

- Wikipedia

Data Science Life Cycle

1. Business Understanding
2. Data Acquisition & Understanding
3. Modeling
4. Deployment

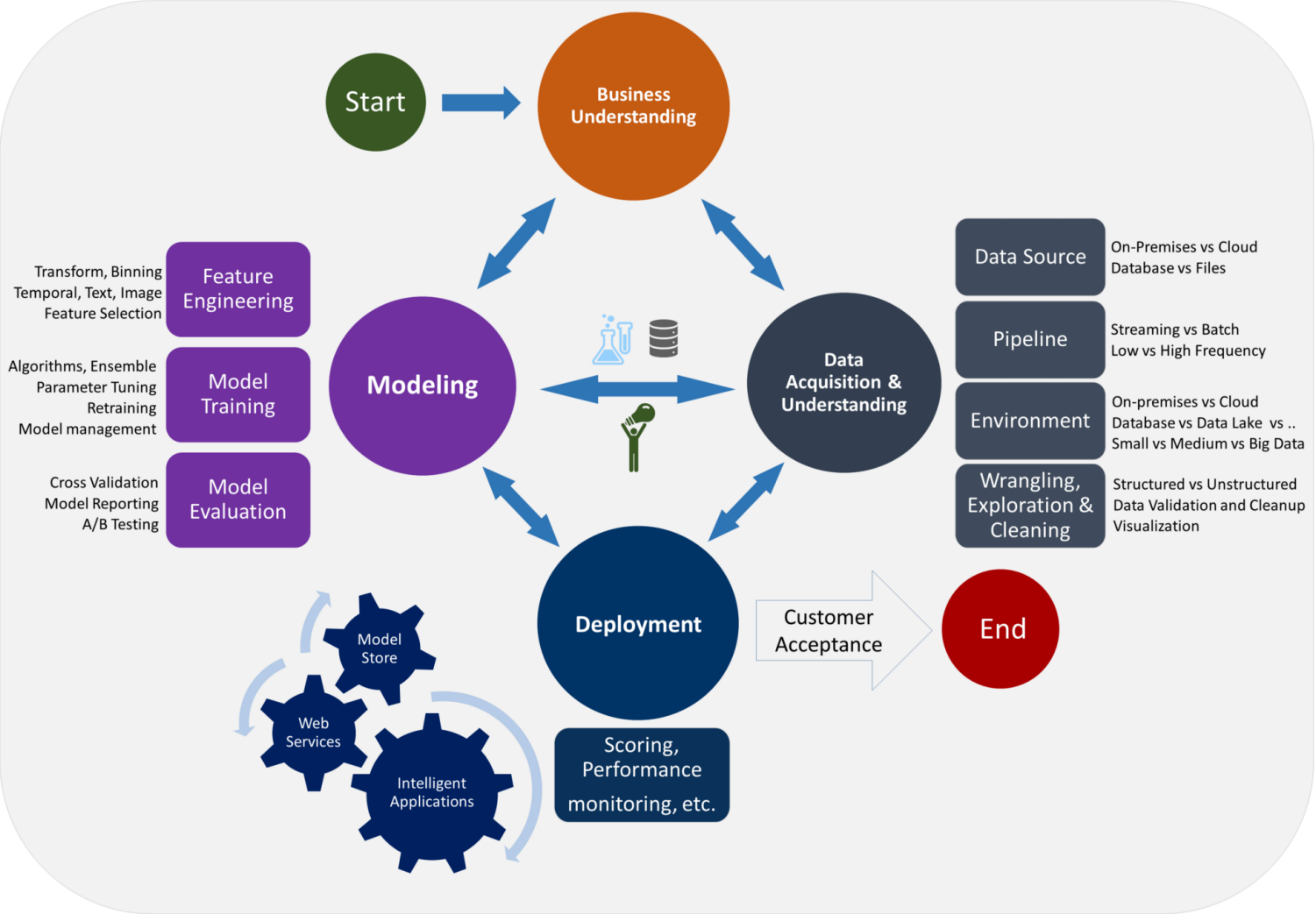
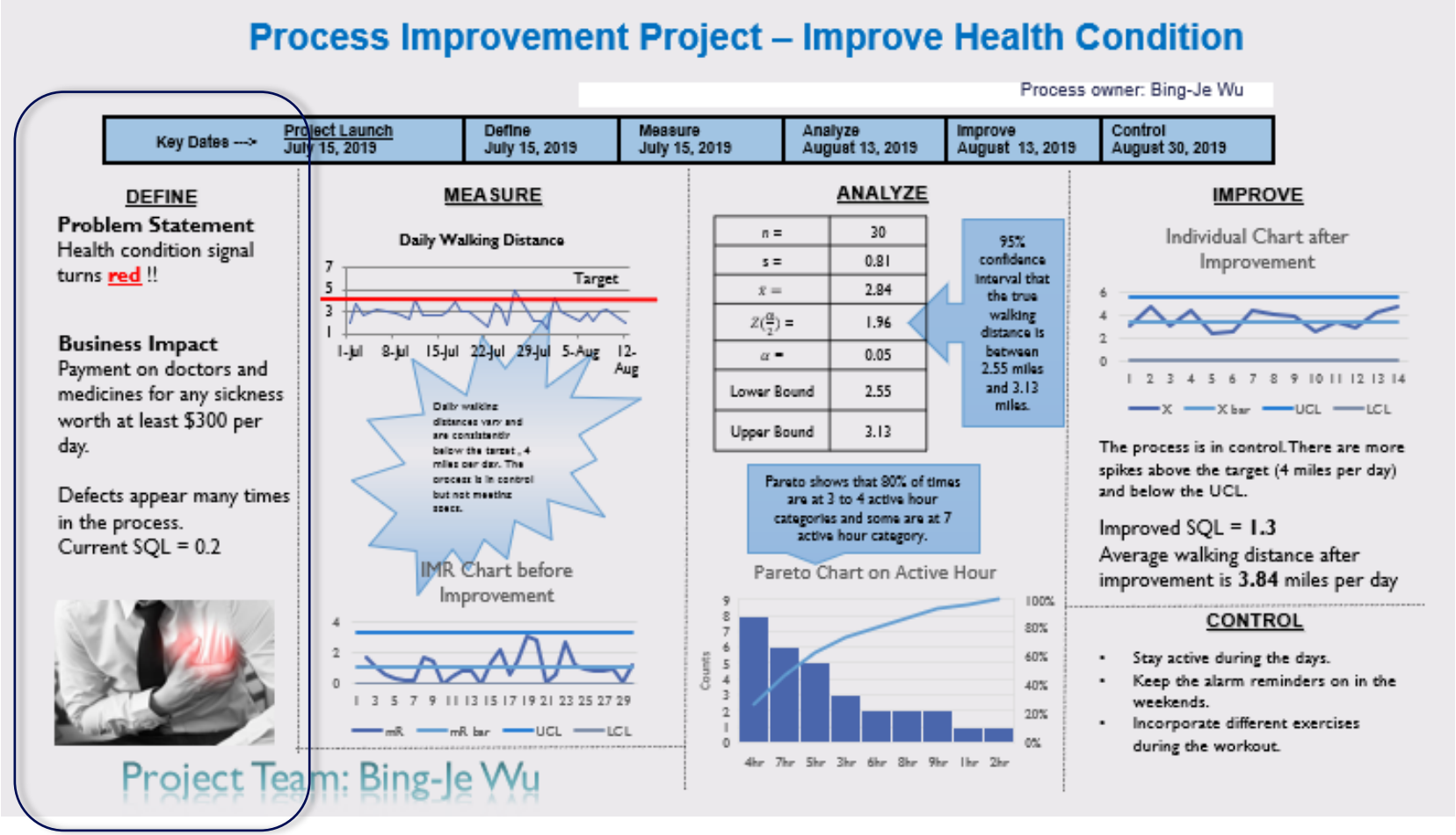


Image Source: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle

Business Understanding

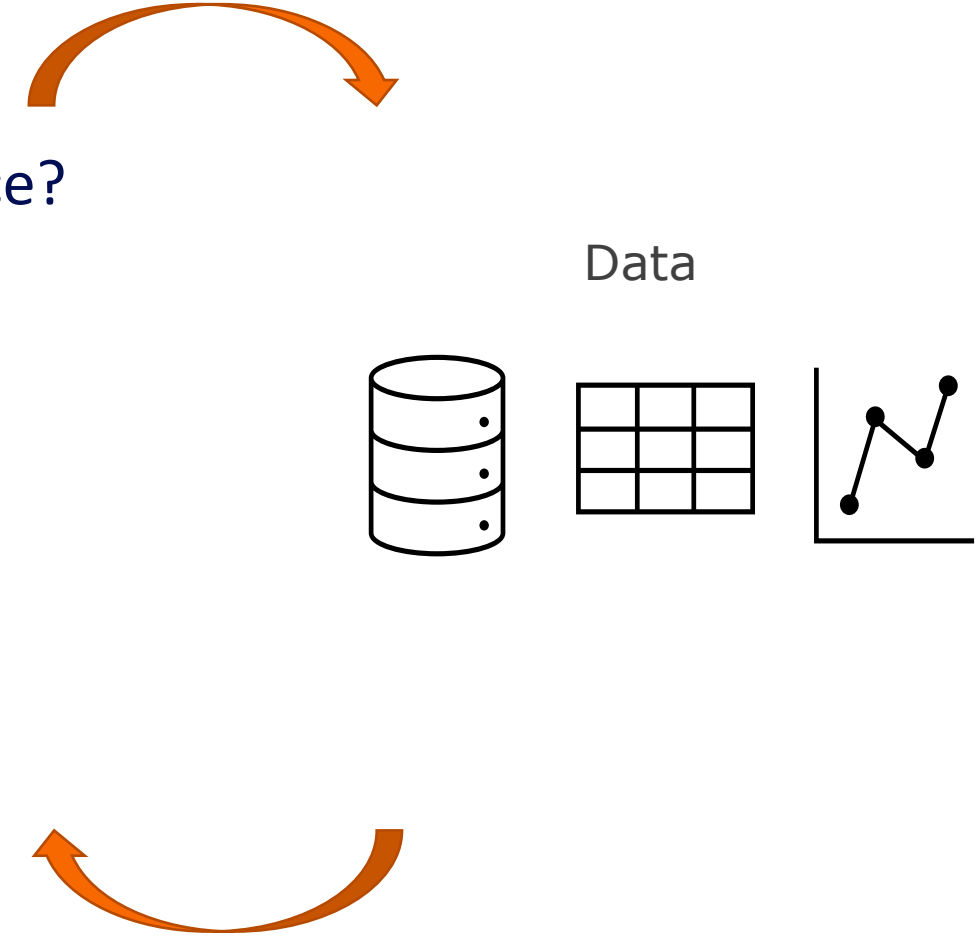
- MBC 638
- Six Sigma Project
- Define Phase
- Measure Phase
- Analyze Phase
- Improve Phase
- Control Phase



Business Understanding Cont.

Questions about the process:

- Do I park the car far from building entrance?
- How many steps do I have per day?
- Do I take stairs or elevator?
- What is the weather?
- Do I order a lunch or bring it from home?
- How long do I sit on my chair?
- How much water do I drink per day?
- How often do I workout?



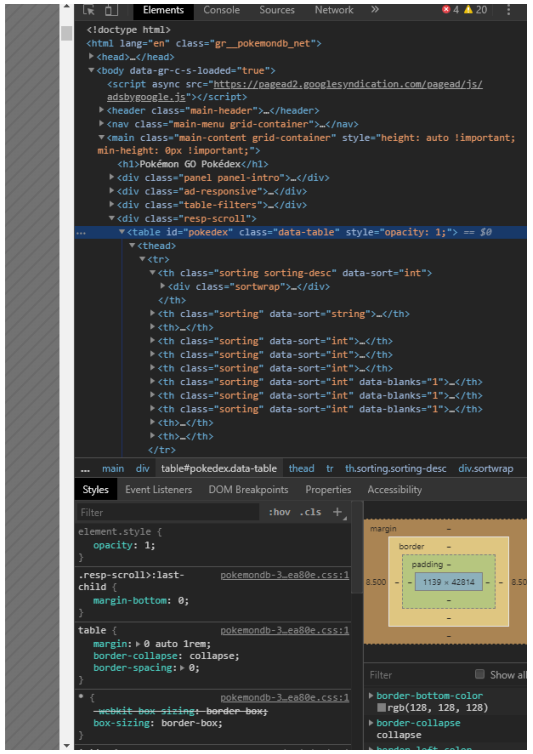
Data Acquisition & Understanding

- IST652
- Pokémon Project
- Unstructured Data
- Programming skills
- Web Scrapping
- API

Unstructured data

Name: Type: - All -

#	Name	Type	Attack	Defense	HP	Catch Rate	Flee Rate	Candy	Fast Moves	Charge Moves
809	Melmetal	STEEL	264	226	190	30%	0%	—	Thunder Shock	Hyper Beam Flash Cannon Rock Slide Thunderbolt
808	Meltan	STEEL	130	118	99	30%	0%	400	Thunder Shock	Flash Cannon Thunderbolt
635	Hydreigon	DARK DRAGON	211	256	188	5%	5%	—	Bite Dragon Breath	Dark Pulse Flash Cannon Dragon Pulse
634	Zweilous	DARK DRAGON	176	159	135	10%	7%	100	Bite Dragon Breath	Dark Pulse Dragon Pulse Body Slam
633	Deino	DARK DRAGON	141	116	93	40%	9%	25	Dragon Breath Tackle	Dragon Pulse Body Slam Crunch
632	Durant	BUG STEEL	151	217	188	20%	7%	—	Bug Bite Metal Claw	Stone Edge Iron Head X-Scissor
631	Heatmor	FIRE	198	204	129	20%	7%	—	Lick Fire Spin	Flamethrower Thunder Punch Power-Up Punch
623	Golurk	GROUND GHOST	205	222	154	15%	5%	—	Mud-Slap Astonish	Shadow Punch Dynamic Punch Earth Power
622	Golett	GROUND GHOST	153	127	92	30%	10%	50	Mud-Slap Astonish	Shadow Punch Brick Break Night Shade
609	Chandelure	GHOST FIRE	155	271	182	5%	5%	—	Hex Fire Spin	Shadow Ball Overheat Energy Ball



Data Acquisition & Understanding Cont.

Data pipeline

- Data cleansing
- Missing Values
- Duplicates
- Inaccurate data (typo)
- Data wrangling
- Transforming data
- Mapping data

Data Cleansing

```
In [10]: # create a dictionary for filling na values
         fill_na = {'Type2':'No value'}
         # assign the the dictionary to replace missing values
         df=df.fillna(value=fill_na)
         # Check to see if missing values have been resolved
         df.isnull().sum()

Out[10]: #             0
         Name          0
         Type1         0
         Type2         0
         Total         0
         HP            0
         Attack        0
         Defense       0
         Sp_Atk        0
         Sp_Def        0
         Speed         0
         Generation    0
         Legendary     0
         dtype: int64
```

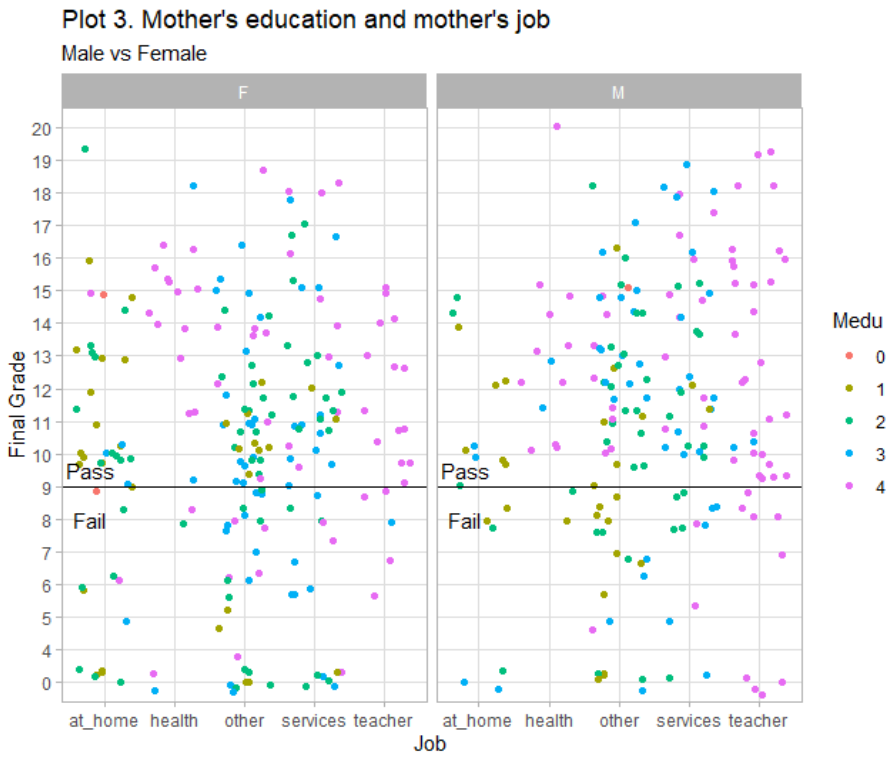
Data Wrangling

```
In [16]: # use apply method from pandas to apply a function to each row
         # Pandas.apply allow the users to pass a function and apply it
         # https://www.geeksforgeeks.org/python-pandas-apply/
         df['Type'] = df['Type'].apply(split_2type)
         df['Type']

Out[16]: 0      [Grass, Poison]
         1      [Grass, Poison]
         2      [Grass, Poison]
         3              [Fire]
         4              [Fire]
                  ...
         565     [Dark, Dragon]
         566     [Dark, Dragon]
         567     [Dark, Dragon]
         568            [Steel]
         569            [Steel]
         Name: Type, Length: 570, dtype: object
```

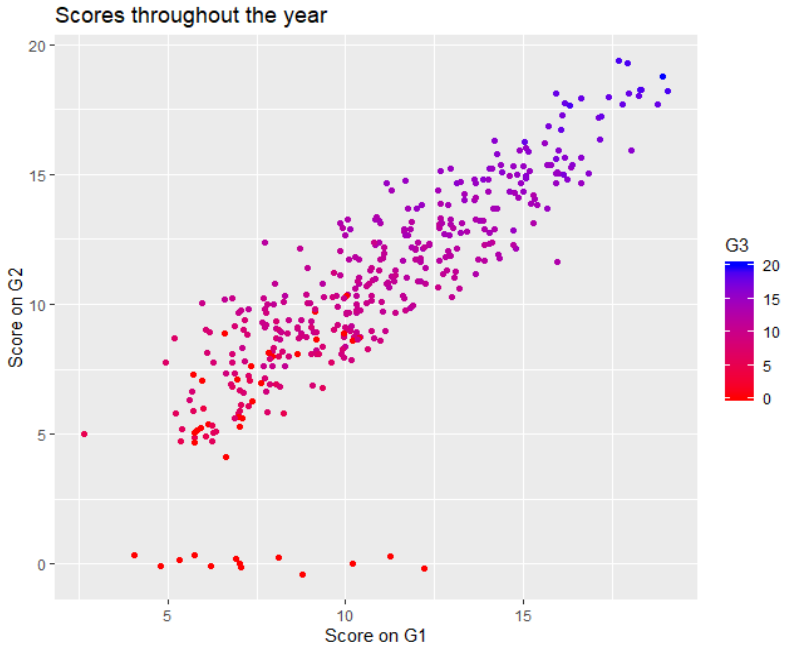
Data Acquisition & Understanding Cont.

- IST707
- Student Performance Project
- Visualization
- Statistical Analysis
- Find patterns

Scatter plot



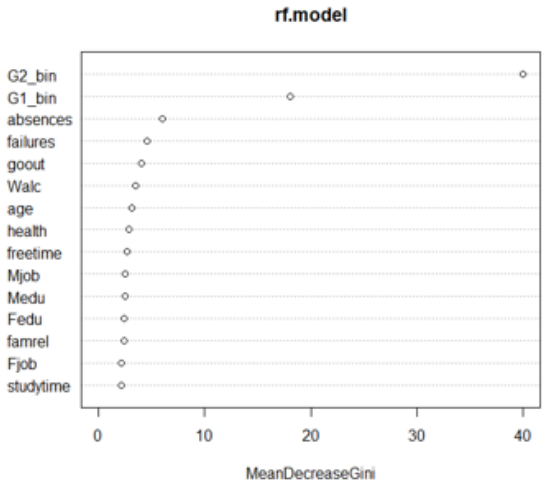
Correlation plot



Modeling

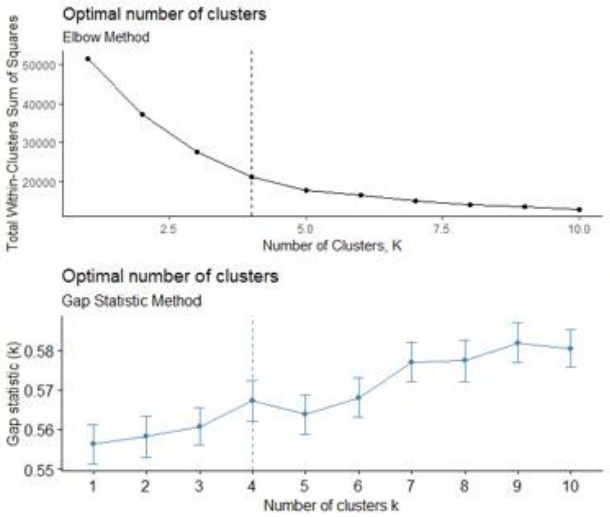
- IST707
- Student Performance Project
- Supervised learning
- Random Forest
- Naïve Bayes
- Support Vector Machine
- Unsupervised learning
- K-means Clustering
- Association Rule Mining

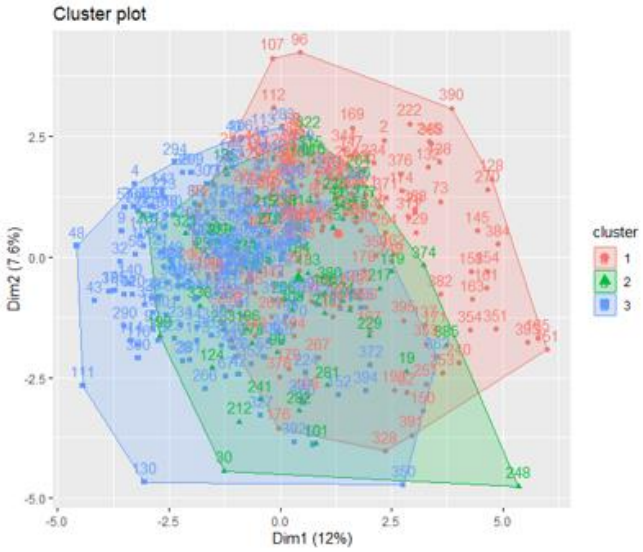
Random Forest




```
Confusion Matrix and Statistics

          Reference
Prediction Fail Pass
      Fail   38    5
      Pass    1   74

               Accuracy : 0.9492
                 95% CI : (0.8926,
0.9811)
    No Information Rate : 0.6695
    P-Value [Acc > NIR] : 1.461e-13

                  Kappa : 0.888

 Mcnemar's Test P-Value : 0.2207

            Precision : 0.8837
               Recall : 0.9744
                   F1 : 0.9268
           Prevalence : 0.3305
       Detection Rate : 0.3220
 Detection Prevalence : 0.3644
    Balanced Accuracy : 0.9555

       'Positive' Class : Fail
```

K-means





Modeling

- IST687
- Google Play Store Apps Project
- Supervised learning
- Support Vector Machine
- Evaluation Metric
- Accuracy Rate
- Precision
- Recall/sensitivity
- F1-Score
- ROC

SVM

```
# Generate a model based on the training data set:
# model 1 --- Radial Basis kernel "Gaussian"
svmOutput <- ksvm(Installs~., data = trainData, kernel = "rbfdot", kpar="automatic",
                  C=5, cross=3, prob.model=TRUE)
svmOutput
```

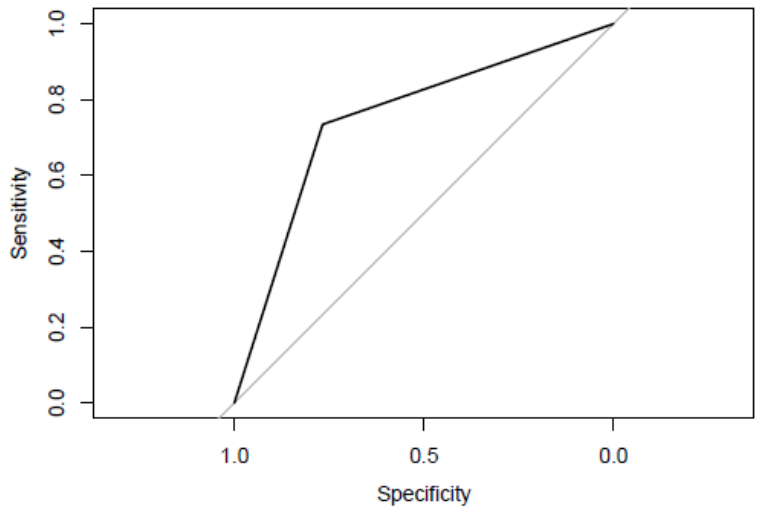
```
## Support Vector Machine object of class "ksvm"
##
## SV type: eps-svr  (regression)
##  parameter : epsilon = 0.1  cost C = 5
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  0.181350269621617
##
## Number of Support Vectors : 3760
##
## Objective Function Value : -11209
## Training error : 0.584834
## Cross validation error : 0.1864
## Laplace distr. width : 0.4722
```

```
predSVM <- round(predict(svmOutput,testData))
cPercent(predSVM, testData$Installs)
```

```
##           Actual
## Prediction    0    1
##          0 1122  341
##          1  338  936
```

```
## [1] "Correct Percentage: 75.19% "
```

ROC Curve



Deployment

- IST736
- Text Prediction from Review Project
- Web application
- Next word prediction
- Sentiment Analysis

Web Application

Conclusion and Reflection

Data scientists are big data wranglers, gathering and analyzing large sets of structured and unstructured data.

A data scientist's role combines computer science, statistics, and mathematics.

They analyze, process, and model data then interpret the results to create actionable plans for companies and other organizations.

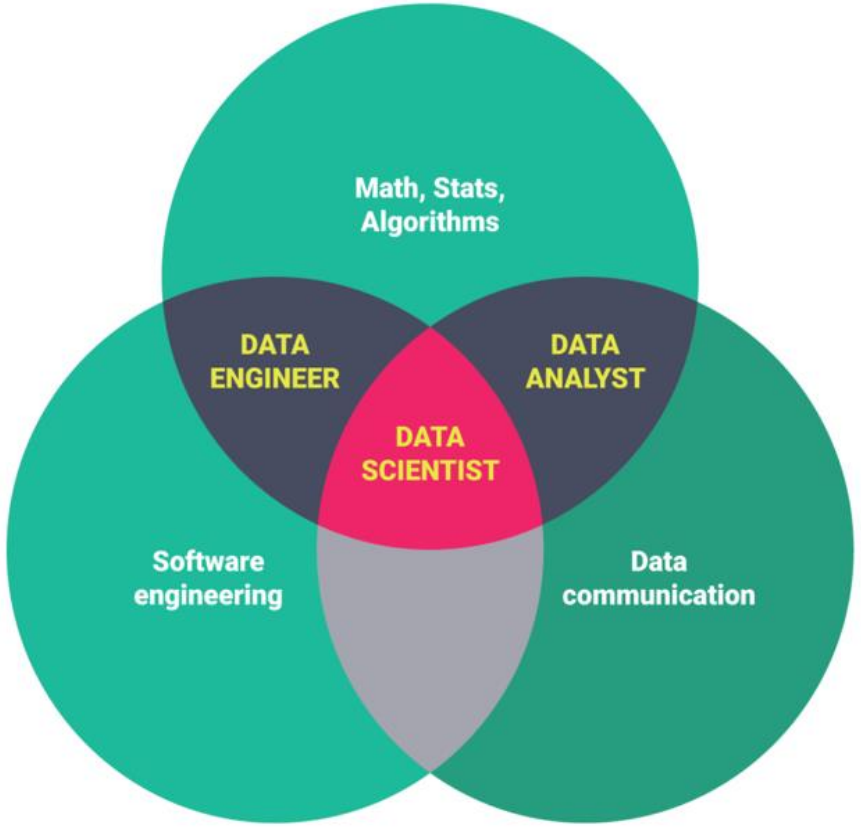


Image Source: https://www.springboard.com/blog/data-science-career-paths-different-roles-industry/

Future planning

- Big data cloud computing
- Deep learning
- Computer vision
- NLP

THE iSCHOOL
Syracuse University

Syracuse University

Thank you iSchool
Thank you SU

Stay connected!