Master of Science Applied Data Science

Syracuse University

Cherngywh Lee
334280765

Charlie Lee

Education

Syracuse University M.S. in Applied Data Science

YuanZe University B.S. in Mechanical Engineering

Professional Background

TPM III at Hyve Solutions

Principal Engineer of iphone team a Foxconn

Mechanical Architects at Supermicro

Mechanical Engineer at INVENTEC


Let's Go Orange!
Syracuse University
SYRACUSE

Portfolio Milestone

M.S. in Applied Data Science

The Applied Data Science program at Syracuse university's School of Information Studies is a practitioner's degree – while the curriculum is founded upon firm theoretical underpinnings, the program is designed to be a professional program with a strong emphasis on the applications of data science to enterprise operations and processes, particularly in the areas of data capture, management, analysis and communication for decision.

Course list

IST 659 Data Admin Concepts & Database Management

IST 769 Advanced Database Management

IST 722 Data Warehouse

IST 687 Applied Data Science

IST 707 Applied Machine Learning

IST 664 Natural Language Processing

IST 772 Quantitative Reasoning in Data Science

IST 718 Big Data Analytics

IST 623 Introduction to Information Security

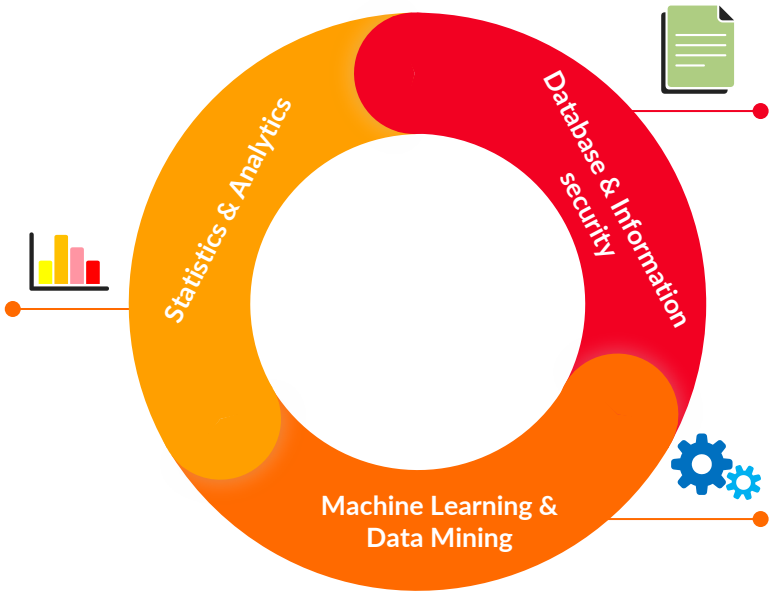
MBC 638 Data Analysis and Decision Making

SCM 651 Business Analytics

The Elements of Applied Data Science

Business Analytics

Analyze data with statistics theory like normal distribution, hypothesis test...etc. as well as tool like fishbone diagram to explore and make a data driven decision



Data Management

Well storage data with solid foundation of database building to various big data applications. Extract data to be data warehouse as the foundation of Business Analytics

Model & Prediction

Find out useful variables, then applied various Machine Learning and Data Mining Model to interpret the data to predict the future

Database & Information Security

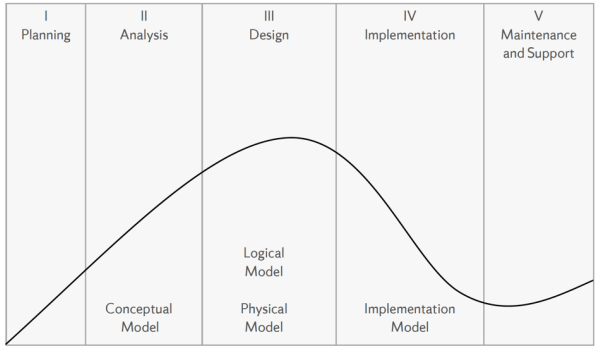
Database is the foundation of Data Science, like a Grocery store to a Chef or Ammunition to the Army

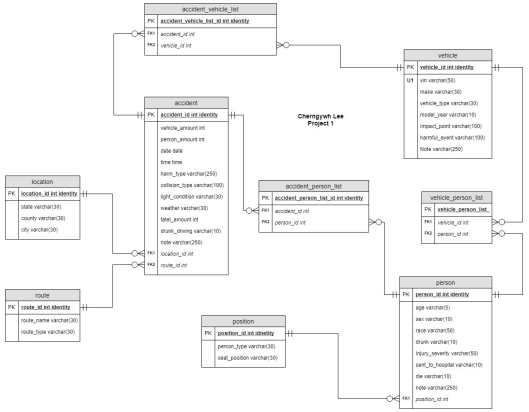
Understand different types of data to build a secured database to clearly store structured data and conduct transaction

Deploy these data on a scalable and robust distributed systems like Hadoop, Data Lake to handle all types of big data

Eventually with Data Warehouse concept to perform ETL, create data sets that precisely fulfill the needs for Business Analytics

Build, Manage & Protect the Database





Examine data structures, file organizations, concepts, and principles of database management systems (DBMS) as well as data analysis, database design, data modeling, database management, and database implementation.

Follow the process below to build up a database from scratch:

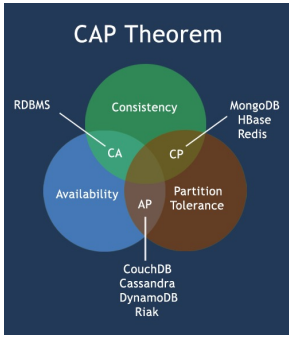
Conceptual → Logical → External → Internal → Physical

- Define the business rules and identification of the entities and relationships in the model. Set up attributes and keys and their types.
- Solve Problems by constructing database queries using SQL
- Recommend and justify strategies for managing data security, privacy, audit/control, fraud detection, backup and recovery with MS Access and SQL Server

Advanced concept and SQL skills with the relational data base model, such as transactions, concurrency control, performance and security. To build up a robust database can safely handle dynamic changes like temporal table.

- T-SQL Transaction Commands like Commit, Rollback to meet ACID (Atomic, Consistent, Isolated, Durable)
- Serializability and Locking Mechanisms to maintain concurrency control
- Define permissions and securable to ensure database security

Big Data Application

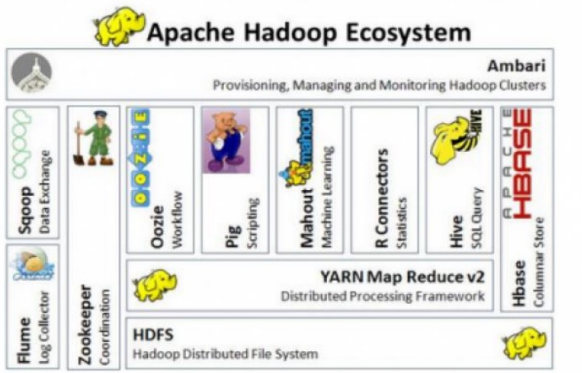


CAP theorem can be used on all kinds of database system

- CA – RDBMSs like Oracle, MySQL
- CP – Single-master systems like MongoDB, HBase and HDFS
- AP – Eventual consistency systems like CouchDB, Cassandra and Dynamo

No database can perfectly fulfill CAP theorem. However, we can freely combine any of them based on our needs per Polyglot Persistence

- MongoDB – Product catalog and blog
- Redis – Shopping cart and web page cache
- Cassandra – Audit and activity logs
- MySQL – Order processing and payments
- Hadoop – Data analytics
- Neo4j – Social graph



Hadoop is the way from data to big data by the scale out design. Following HDFS

Data is stored as it is and chunked into blocks and distributed to data nodes across a cluster and managed by a name node.

- Data stored as they are, schema applied when data are read
- Name Node is the master node to determine and maintain how the blocks of data
- Data split into blocks and are distributed over physical Data Nodes
- Default replication for failover


ANALYSIS OF THE SOURSE FILE
0,0587230-9
0,05
0,04
0,03
0,02
0,01
OBJECT: B/31/99
STATUS:
mode:
32/673
FILE MANAGER
FILE DIRECTORY
026/57
SYSTEM FILE
PROFILE
79933-2B
15%
IDENTIFICATION
PROFILES
SETTINGS
HISTORY
LOG
DOCUMENTS
MATCH
MATCH

Data Warehouse Project – Questions & Actions

Data Warehouse is the foundation of BI (business intelligence).

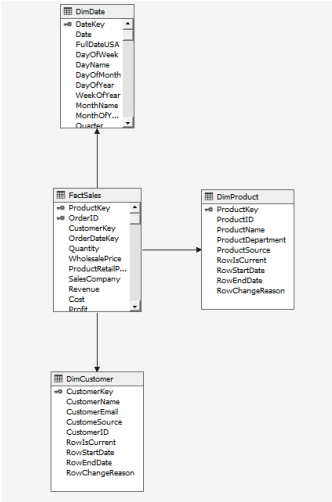
In this project, we performed ETL (Extract, Transform, Load) form OLTP database to BI.

Data Introduction and Analytics Topics:

- 2 OLTP data sets – Fudgemart & Fudgeflix. Fudgemart is and online retailer and Fudgefilx is an online DVD seller by mail/video.
- We have a few questions and improvement targets for these 2 data sets:
  - Who are the most frequent customers for Fudgemart?
  - What are the most profitable products for Fudgemart?
  - How do sales compare between Fudgeflix and Fudgemart? Profits and Revenues?
- Identify Business Process to Model, these represent the data mart in the data warehouse:
  - Business Process and Gain – one row per product order to figure out who is the most frequent customer
  - Dimensions – Sales, Date, Order, Customer
  - Facts – Order amount, product price

Technical Steps:

1. Create a Star Schema, Data mart is implemented as a star schema in a RDBMS. It's also called ROLAP



Data Warehouse Project - Actions

2. Define the attributes in each dimension from high-level to detailed-level

Group 3 *4/20/2020*

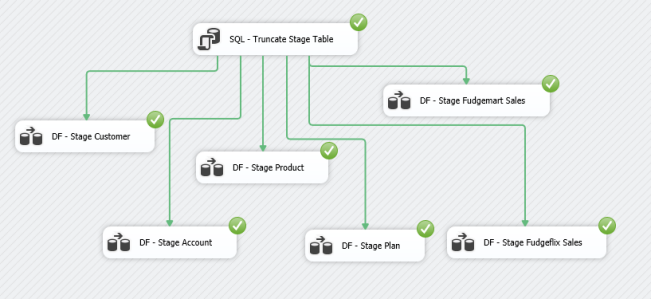
Business Process Name	Fact Table	Fact Grain Type	Granulairty	Facts
most frequent customers	Frequent_customers	accumulating snapshot?	one row per orders	number of orders,
most ordered product per customer	top_product_per_customer	accumulating snapshot?	one row per order detail	number of products
rating of each title	title_rating	transaction	one row per title	Title, Rating, director
accounts per zipcode	accounts_zipcode	Periodic Snapshot	one row per zipcode	Name, Address, Zipcode, Account Plan
fudgeflix vs fudgemart spending	fudgeflix_fudgemart_spending	Accumulating snapshot	row for subscription or order	Customer, Accounts, total sales,
fudgeflix customer subscription tier, spending at fudgemart	fudgeflix_fudgemart_tier_spending	Accumulating snapshot	row for subscription amd order	customer, account, order
Fudgeflix customer plan change	fudgeflix_plan_change	slow changing dimesion	one row per plan	customer, account, order

Table Name	DimProduct
Table Type	Dimension
Display Name	Product
Database Schema	group3
Table Description	Products on an order (with supplier and category info)
Comment	rollup into suppliers and categories
Biz Filter Logic	
Size	
Generate Script?	Y

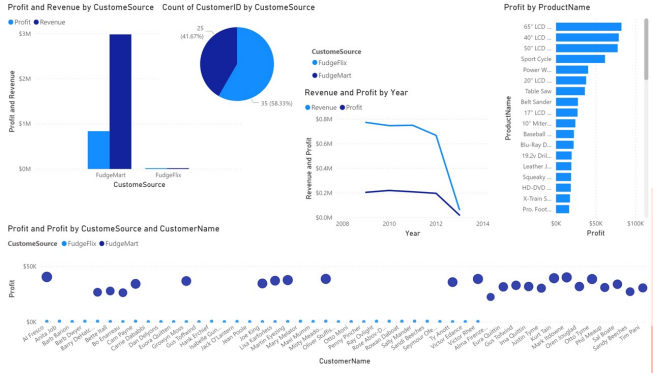
Home Page

Column Name	Display Name	Description	Unknown Member	Example Values	SCD Type
ProductKey	ProductKey	Surrogate primary key	-1	1, 2, 3…	key
ProductID	ProductID	Business key from source system (aka natural key)	-1	1,2,3,…	key
ProductName	ProductName	Name of product	None	Chi Tea	2
ProductDepartment	ProductDepartment	department of product	None	Hardware	2
ProdcutSource	ProdcutSource	From fudgemart or fudgeflix	None	Fudgemart	2
RowIsCurrent	Row Is Current	Is this the current row for this member (Y/N)?	1	TRUE, FALSE	n/a
RowStartDate	Row Start Date	When did this row become valid for this member?	1/1/00	1/24/11	n/a
RowEndDate	Row End Date	When did this row become invalid? (12/31/9999 if current row)	12/31/99	1/14/1998, 12/31/9999	n/a
RowChangeReason	Row Change Reason	Why did the row change last?	N/A		n/a

3. Perform ETL to extract data from OLTA to Data Warehouse based on the ROLAP



4. Conduct Data Analytics by BI tools with data from Data Warehouse



Statistics & Business Analytics

Explore data with statistical methods and technologies for analyzing historical data in order to gain new insight

Clearly figure out the root cause, strategically plan steps and actions to improve the performance to meet the business goal with various soft tools

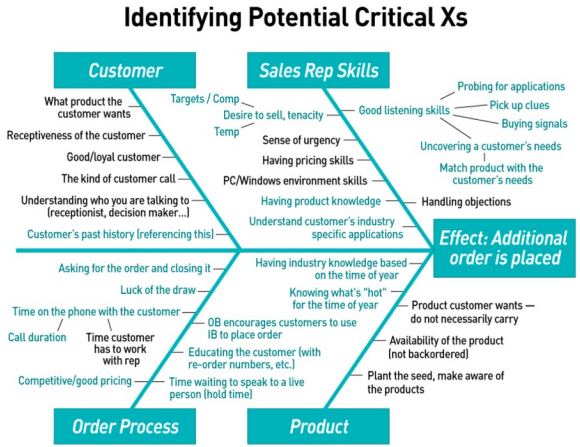
Profoundly understand statistics and probability theory to analyze the nature of the datasets and the tendency underlying the data to make the best business decision

Quantitative Reasoning with Statistics

Soft Tools

Clearly evaluate the factors in an event with critical thinking framework:

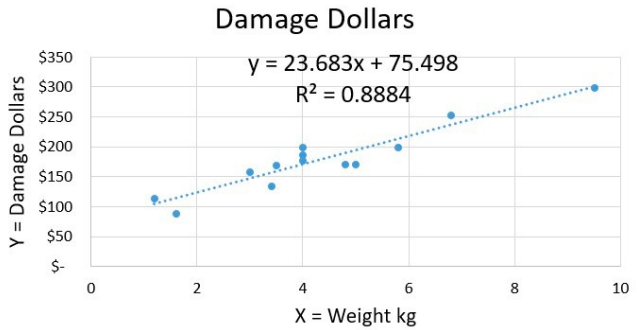
- DMAIC
- Process Map
- Affinity Diagram
- Fishbone Diagram



Hard Tools

Analyze data with statistics mindset and knowledge:

- Various Probability Distribution
- Simple and Multiple Linear Regression
- Logistic Regression
- Performance Evaluation like R-Squared or Sensitivity Analysis
- Bayesian and Traditional Hypothesis Testing
- Tim Series Analysis
- ML & DL application



Define Business Questions and Improve with DMAIC methodology

Define 7/3

- Monthly Saving was $1231.52 in average.
- The goal was to save $518.4 more, Improvement Rare = 42%.
- In order to extract more detail insight, the data set was prepared by week. Mean of Weekly saving is $307.88.

Weekly Expenses	
Mean	307.8791667
Standard Error	53.40803329
Median	358.515
Mode	#N/A
Standard Deviation	320.4481997

- Determined the defect by the Mean. SQL of current saving was 1.85.

Original Process (Nov ~ Jul)	
Defect opportunities per unit (D) =	4
How many units in the data set (U)	9
Total actual defets (D X U)	36
Total actual defects (A)	13
Defect-per-opportunity rate (A ÷ DU = DPO)	36%
Defects per million opportunities (DPMO)	361111
SQL value =	1.85

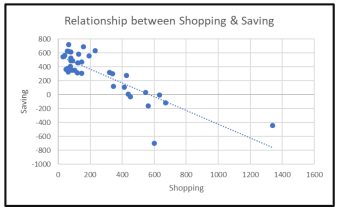
Measure 7/10

- The purpose was to **find the major expense.** Figured out the factor with highest correlation coefficient to leverage the affect of actions we were going to take.

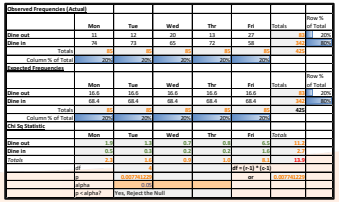
- According to the plot, I decided to do some further Analyses on **Grocery, Restaurant and Shopping.**

Analyze 7/17

- **Shopping** is the major expense category which has a strong relationship with **Saving** because **R = 0.82.**



- **p = 0.0077 < α = 0.05,** Chi-Square Test showed a strong evidence that I dined out more on **Wednesday and Friday.**



Improve 8/15

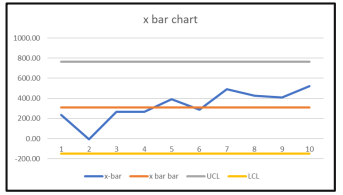
- Stop browsing Amazon and Facebook coupon sharing group. Starting bringing lunch box on Wednesday and Friday.
- Mean of Weekly Saving on August was $520.44.
- Hypothesis Test showed that I made a significant improvement and SQL of saving is 6 now.

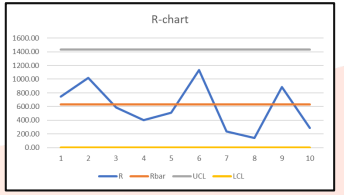
Original	
Mean	307.88
Standard Error	320.45
Count	36
New	
Mean	520.44
Standard Error	126.10
Count	4
α	0.05
Z-Value	-2.57
P-Value	0.005049
Reject Ho	

Original Process (Aug)	
Defect opportunities per unit (D) =	4
How many units in the data set (U)	1
Total actual defets (D X U)	4
Total actual defects (A)	0
Defect-per-opportunity rate (A ÷ DU = DPO)	0%
Defects per million opportunities (DPMO)	0
SQL value =	6

Control 8/29

- The improvement rate was actually 69% and beyond my original goal.
- The trend of saving goes up but still within the limit. It makes me confident with this result. **I'm capable of saving this much money.**

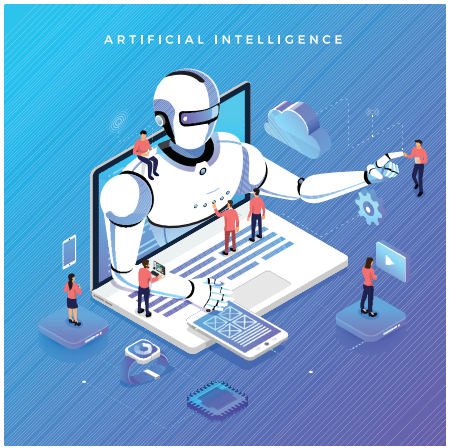




- Will take further actions when the saving become stable.

Machine Learning & Prediction

Predictive modeling solutions are a form of Data Mining by analyzing historical and generating model to predict the future



In order to let computer act like human to complete vast tasks like classify objects, determine true or false, find out unknown pattern and the trend in the future

Train it with all popular ML and DL algorithms with fundamental knowledge to fine tune the hyperparameters to reach a decent performance

ML & DL Algorithms

Supervised Learning

Classification

Learning patterns by labeled data and perform classification based on learned knowledge

- K-Nearest Neighbor
- Logistic Regression
- Decision Tree
- Random Forest
- Naïve Bayes
- SVM
- Artificial Neural Networks

Image Recognition

Computer can recognize the image data and classify

- Convolution Neural Network

Regression

To predict the future based on the data associated with time

- Linear regression
- Tim Series

To determine a mathematical relationship among several random variables

- Multiple Linear regression

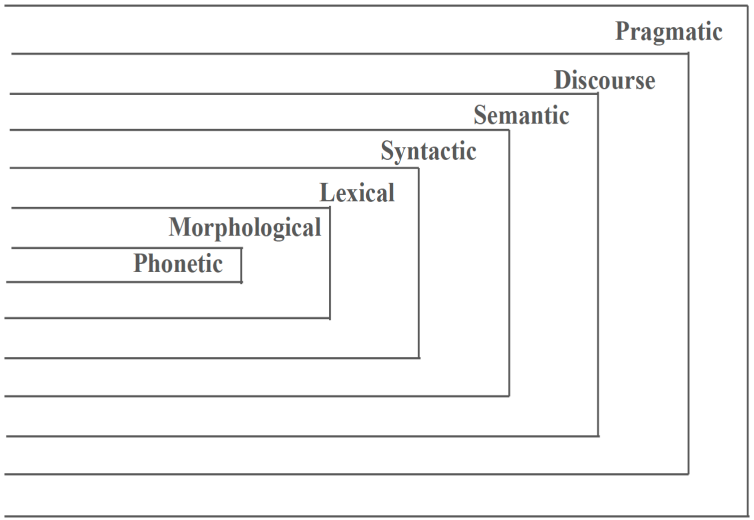
Unsupervised Learning

Directly cluster data without being given an answer. Find out the similarity and define the relationship based on data's characteristics or frequency of occurrence

- Association Rule Mining
- K-Means
- Hierarchical Clustering
- Ensemble Learning

Natural Language Processing - Theory

NLP is a range of computational techniques for analyzing and representing naturally occurring tests at one or more levels of linguistic analysis for the purpose of achieving human-like language processing



- Phonetic – Interpretation of speech sounds within and across word sound waves are analyzed and encoded into a digitized signal
- Morphological – Deals with the componential nature of lexical entities like prefix and suffix
- Lexical – Adding lexical class information to words, Part-of-Speech (POS) tagging tags words with specific noun, verb, adjective, and adverb types
- Syntactic - Analyzing of words in a sentence to uncover the grammatical structure of the sentence. Break down the phase structure rules in a hierarchical way for Noun phrase, Verb phrase to Determiner, Noun…etc.
- Semantic – Determining possible meanings of a sentence, extract the Semantic Relation between words
- Discourse – Determining meaning in texts longer than a sentence and making connections between component sentences. For example, Introduction and conclusions are distinct patterns in a documents
- Pragmatics – The purposeful use of language in situations to explain how extra meaning is read into texts without being encoded in them

NLP Project – Data Preparation

Use Kaggle competition movie review phrase data (https://www.Kaggle.co/c/sentiment-analysis-on-movie-reviews) and labeled for sentiment to train a ML model for Sentiment Analysis. Socher's group used crowd-sourcing to manually annotate all the subphrases of sentences with a sentiment label ranging over: **Negative, Somewhat Negative, Neutral, Somewhat Positive, Positive.**

Step 1: Pre-processing and filtering

Data cleaning for sentiment analysis. Because sentiment lexicon only contains alphabetic words, so need to remove punctuation and then tokenize words

- Remove punctuation by alpha_filter function

```
def alpha_filter(w):
  # pattern to match word of non-alphabetical characters
  pattern = re.compile('^[^a-z]+$')
  if (pattern.match(w)):
    return True
  else:
    return False
```

- Tokenize the words

```
  for phrase in phraselist:
    tokens = nltk.word_tokenize(phrase[0])
    alphawords = [w for w in tokens if not alpha_filter(w)]
    phrasedocs.append((alphawords, int(phrase[1])))
```

Step 2: Create feature function

- Trigram & POS Tag
- Negation & LIWC

```
def tri_features(document, word_features):
    document_words = set(document)
    features = {}
    for word in word_features:
        features['V_{}'.format(word)] = False
        features['V_NOT{}'.format(word)] = False
    # go through document words in order
    for i in range(0, len(document)):
        word = document[i]
        if ((i + 1) < len(document)) and ((word in negationwords) or
(word.endswith("n't"))):
            i += 1
            features['V_NOT{}'.format(document[i])] = (document[i] in word_features)
        else:
            features['V_{}'.format(word)] = (word in word_features)
    # apply LIWC and count the score for Positive and Negative respectively
    Pos = 0
    Neg = 0
    poslist, neglist = sentiment_read_LIWC_pos_neg_words.read_words()
    for word in document_words:
        if isPresent(word, poslist):
            Pos += 1
        if isPresent(word, neglist):
            Neg += 1
    features['positivecount'] = Pos
    features['negativecount'] = Neg
    return features
```

NLP Project – ML Model & Conclusion

Step 3: Train a Naïve Bay Model

Models predict 5 groups from **Negative** to **Positive**. **Although Neutral is the major category with the most data and the best performance,** we still need to average out all categories to evaluate the whole model.

The experiment with the best Precision score

- Macro Average – Negation & LIWC
  - Precision: 0.431
  - Recall: 0.405
  - F1: 0.401
- Micro Average - Baseline
  - Precision: 0.549
  - Recall: 0.519
  - F1: 0.520

The experiment with the best Recall score

- Macro Average – Alpha Filter
  - Precision: 0.364
  - Recall: 0.406
  - F1: 0.373
- Micro Average – Negation & LIWC
  - Precision: 0.532
  - Recall: 0.545
  - F1: 0.530

Conclusion:

In general, more data and folds has better performance. Randomly select 15000 phrases and run 10 folds on the data we clean in the formal steps.

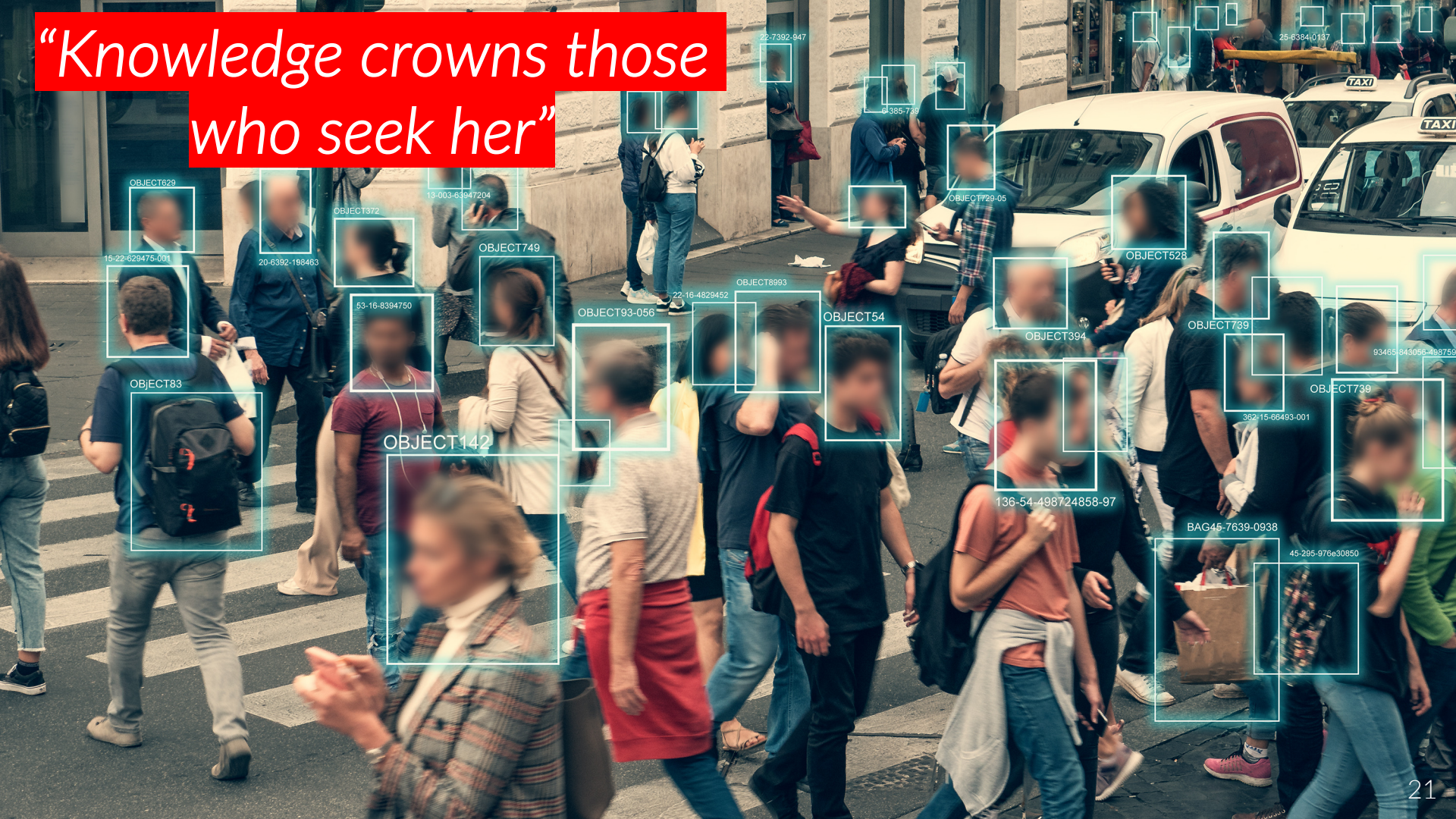
After checking the precision of **Neutral**, we can see a limitation about 0.82. I think this is a fundamental limitation of the methods I used. I need to try more complex and profound methods, **otherwise it may be just kind of a trade-off between Neutral and Non-Neutral (other 4 categories).**

However, because this is about movie review, So **I think we care about the review with strong sentiments more. It's obvious that sentiment lexicons successfully made it.**

Because the sentiment lexicon analyzes the sentence by words. It helps to figure more sentences with stronger sentiment. **But maybe it's too detail so sometimes it may misclassify some natural sentences like "I like the book version more" as positive. Hence it creates more type I error (Neutral but classified as Strong sentiment).**

And because of it, type II error increased on "Neutral". But again, I still think it's a good result because we care about the reviews with strong sentiment more.

In conclusion, the performance of sentiment lexicons like LIWC and AFINN did a better job on finding out more sentences with sentiment. Whereas Alpha Filter and Trigram & POS tags are doing better on the general purpose.


“Knowledge crowns those who seek her”
22-7392-947
25-6384-0137
TAXI
TAXI
OBJECT629
13-003-63947204
OBJECT372
OBJECT729-05
15-22-629475-001
20-6392-198463
OBJECT749
OBJECT528
53-16-8394750
22-16-4829452
OBJECT8993
OBJECT93-056
OBJECT54
OBJECT394
OBJECT739
OBJECT83
OBJECT142
362-15-66493-001
OBJECT739
136-54-498724858-97
BAG45-7639-0938
45-295-976e30850

Big Data Analytics Project

This Data Analytics project is an approach by following OSEMiN to turn data into information

Step 1: Specify

What are the factors leading to higher amounts of confirmed COVID-19 cases? Need to prove our Hypotheses:

- higher vaccination rates = fewer confirmed cases
- more positive vaccine sentiment = higher vaccination rates
- more positive vaccine sentiment = fewer confirmed cases

Step 2: Scrub

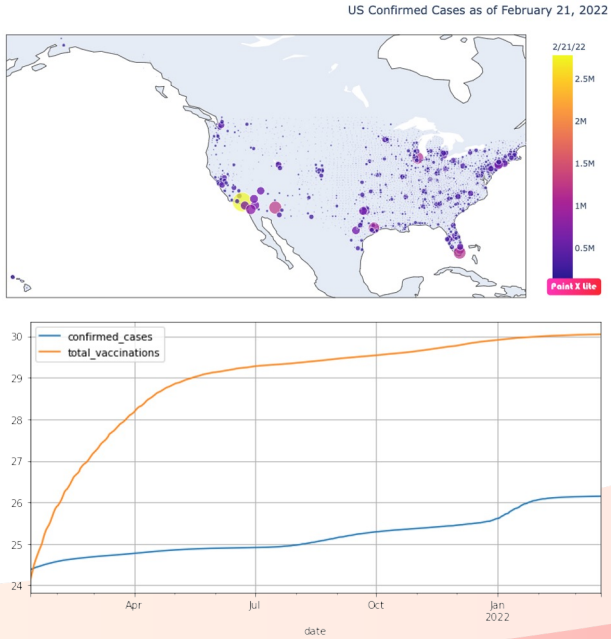
Which is also called Data Wrangling. Looking into the raw data and optimize the data quality by unifying data format, normalizing values and filling in unknown data with a meaningful value like mean or zero…etc.

In this case, we fill in NA with a mean of the previous day and the next day.

```
for i in range(len(us_states_2020_df.index)-1):
    if (i != 15633) & (us_states_2020_df['negative'][i]==0):
        us_states_2020_df['negative'][i] = (us_states_2020_df['negative'][i-1]+us_states_2020_df['negative'][i+1])/2
```

Step 3: Explore

Explore the data by creating visualization or a graph to gain an intuitive insight from the data



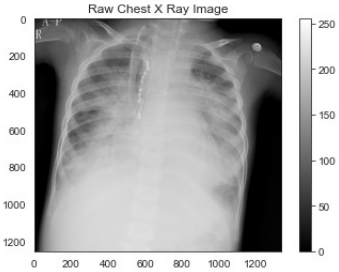
Big Data Analytics Project

Step 4: Model - CNN

Create a Deep Learning model to classify X-Ray image to quickly determine it's Normal or Pneumonia.

Normal

Pneumonia



Train set: PNEUMONIA = 3875, NORMAL = 1341

Test set: PNEUMONIA = 390, NORMAL = 234

Validation set: PNEUMONIA = 8, NORMAL = 8

We directly applied transfer learning by using champion models:

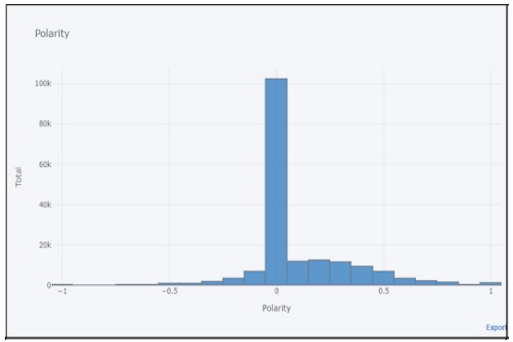
Densenet121, InceptionV3,

- **DenseNet:** utilizes dense connections between layers, through Dense Blocks, where we connect all layers (with matching feature-map sizes) directly with each other.
- **Inception-V3:** makes several improvements including using Label Smoothing, Factorized 7 x 7 convolutions, and the use of an auxiliary classifier to propagate label information lower down the network
- **EfficientNetB4:** a novel model scaling method that uses a simple yet highly effective *compound coefficient* to scale up CNNs in a more structured manner.

Model	Precision	Recall	F1	Accuracy
DenseNet	81.6%	51.5%	41.6%	63.6%
Inception-V3	86.3%	84.6%	85.3%	86.5%
EfficientNetB4	67%	55.2%	39%	44.4%

Big Data Analytics Project

Step 4: Model – Sentiment Analysis



- Polarity from TextBlob Package
- ~45% of tweets as 0
- Majority of other tweets are positive sentiment
- Cluster from 0.05 ~ 0.5

Step 5: Data Observation

- No correlation between cases per capita and state population
- Vaccine sentiment in US leans more positive
- More likely for a negative sentiment Tweet to be re-tweeted than positive sentiment Tweet
- Higher raw numbers of vaccinations and COVID-19 cases correspond with population numbers
- Higher amounts of vaccinations per capita follow political borders

Step 6: Recommendations

- Be wary of Covid-19 info on social media; use reliable sources for information pertaining to Covid-19
- Get vaccinated

Summary of My ADS Journey

The path in the program that I selected is about data engineering. I took all classes from database foundation to real world application like data warehouse as well as more hardware related field like distributed file system (Hadoop). All these classes provide me a comprehensive knowledge from physical computer infrastructure to virtual data information extraction.

Not to mention the core courses which are focus on machine learning and analytics, I gained abundant hands-on experience to explore data, tell a good story with statistical thinking and predict the future with the model that built by AI. I'm confident and ready to be a data scientist to work in the industry.

In this program, I met many wonderful people, made new friends and learned from brilliant, knowledgeable professors from various fields. It's a wonderful journey. Go Orange!!

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash
- Photographs by Sutterstock
- Photographs by https://www.syracuse.edu/about/social-media/images/