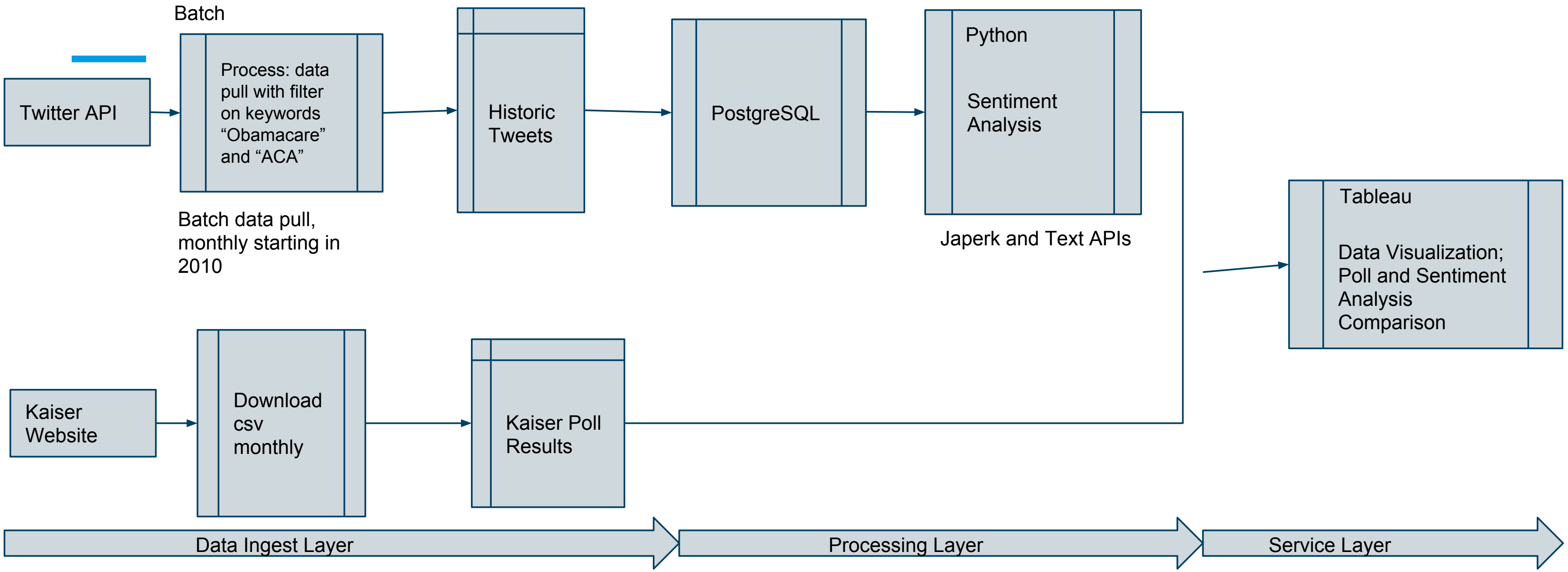
Correlating Poll Results with Twitter Sentiment on Obamacare

W205
Tingwen Bao
Jay Cordes
Alex Jamar
Michelle Liu

Overview of the problem

- Develop a data infrastructure to hold both ACA poll results from Kaiser and Twitter posts
- Analyze the correlation between the result of public polls and sentiment on Twitter
- Twitter claims that the real-time, public orientation of its social network makes it a reliable barometer of the public's constantly changing moods and interests

Overall Architecture



Acquisition of Twitter Data

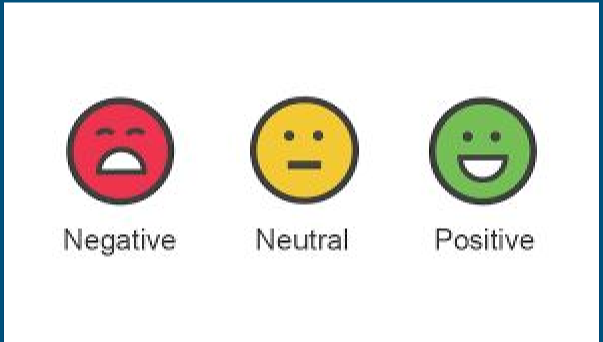
- Jefferson Henrique's Python library, GetOldTweets, was adopted
  - Bypass the time constraint limitations of Twitter API
  - Search for tweets in English about Obamacare or ACA since its introduction in 2010
  - Collects 1000 tweets per day every 7 days.
- Ran on an Amazon Web Services Elastic Cloud Compute virtual machine
- Stored about x million Tweets in x files totaling x GB of data on AWS

The Database

- Created a python script to import tweets tab-delimited file into PostgreSQL
- Used command-line argument with filename for flexibility
- Added surrogate primary key (bigserial data type is auto-incrementing) and indexes
- Exceptions handled gracefully and script keeps importing
- Skipped transactions and kept everything simple and maintainable
- PostgreSQL is very easy to use and facilitated later sentiment scoring / analysis steps

Sentiment Analysis

- Randomly selected 10% of the tweets for analysis
- Used 2 sentiment analysis APIs
  - Sent POST requests containing the Tweet text to the APIs
  - Received JSON objects in return with either “positive,” “negative,” or “neutral” labels
  - Mapped positive → 1, negative → -1, neutral → 0
- Updated Tweets database with the results



Evaluation & Visualization

Why Tableau:

1. Can easily connect both databases and flat files
2. Relative strong visualization with interactive features
3. Various ways to share/access

Drawbacks:

1. Not strong at stats
2. Could become slow after scale up

To evaluate and visualize the result, we create a dashboard with 3 tabs:

- Background: a general introduction of the background and objective of this project

- Comparison: visuals used to compare the 2 different sentiment analysis method and understand its correlation with public poll data

- Explore Tweets: word cloud of top tweets by favorites and its trend by time with interactive filters

https://us-east-1.online.tableau.com/t/tingwen/views/Obamacare/Background?:embed=y&:showShareOptions=true&:display_count=no&:showVizHome=no

Insert Tableau Dashboard Screenshot

Or show interactive dashboard

Results

- Using different sentiment analysis method yields different conclusion of Obamacare related tweets. Japerk’s score indicates that the positive and negative tweets of Obamacare are close to a tie. Textanalysis’s score indicates that it thinks there are about twice as many negative tweets as positive.

- KPP poll shows public has a nearly even spread on their opinion on Obamacare. It is more positive at the start and then more negative, and most recently, almost even.

- There is no significant correlation between sentiment analysis score and poll results

- We only pulled a portion of tweets by time. It is surprised to see that the tweets with most favorites and retweets happened in 2016 rather than when it was first introduced. It was right before election when people mentioned Obamacare to express their political stands.

Future Work

Roadmap for improving the solution with increased usage and increasing data size: improving the ways in acquiring the data, storing and processing it more efficiently, and overall, scaling up our solution.

- Replace all the tabs in the tweets with spaces before writing it to CSV.
- So instead of pulling the same amount of tweets every week, pulling in in proportion with the total number of related tweets might improve the analysis.
- When creating the PostgreSQL table, we should default the sentiment scores to null instead of 0.
- automate and import the polling data into a PostgreSQL table as well. The more data we have in easy-to-use database tables, the better. That would also allow us to do part of the stat calculation and aggregation in PostgreSQL rather than in Tableau.
- Instead of connecting directly to raw tweets data in Postgres, we could do some aggregation and calculation in the serving layer to reduce the processing time of Tableau.

Conclusion

- Text sentiment analysis is nowhere near perfect
- Did not find correlation between poll results and Twitter sentiment
- The 3Vs that define big data are important challenges in our project: long term needs of the solution will have to control the sheer *volume* of the data coming in from Twitter, and to increase the *velocity* of doing sentiment analysis and processing of the data.
- Test any new sentiment analysis approaches that are developed and presumably, their conclusions will become closer and closer to poll results as progress is made. If we could demonstrate a correlation between Twitter sentiment scores and polling data, we would have provided evidence for a powerful new tool, essentially the ability to do real-time polling on any subject.

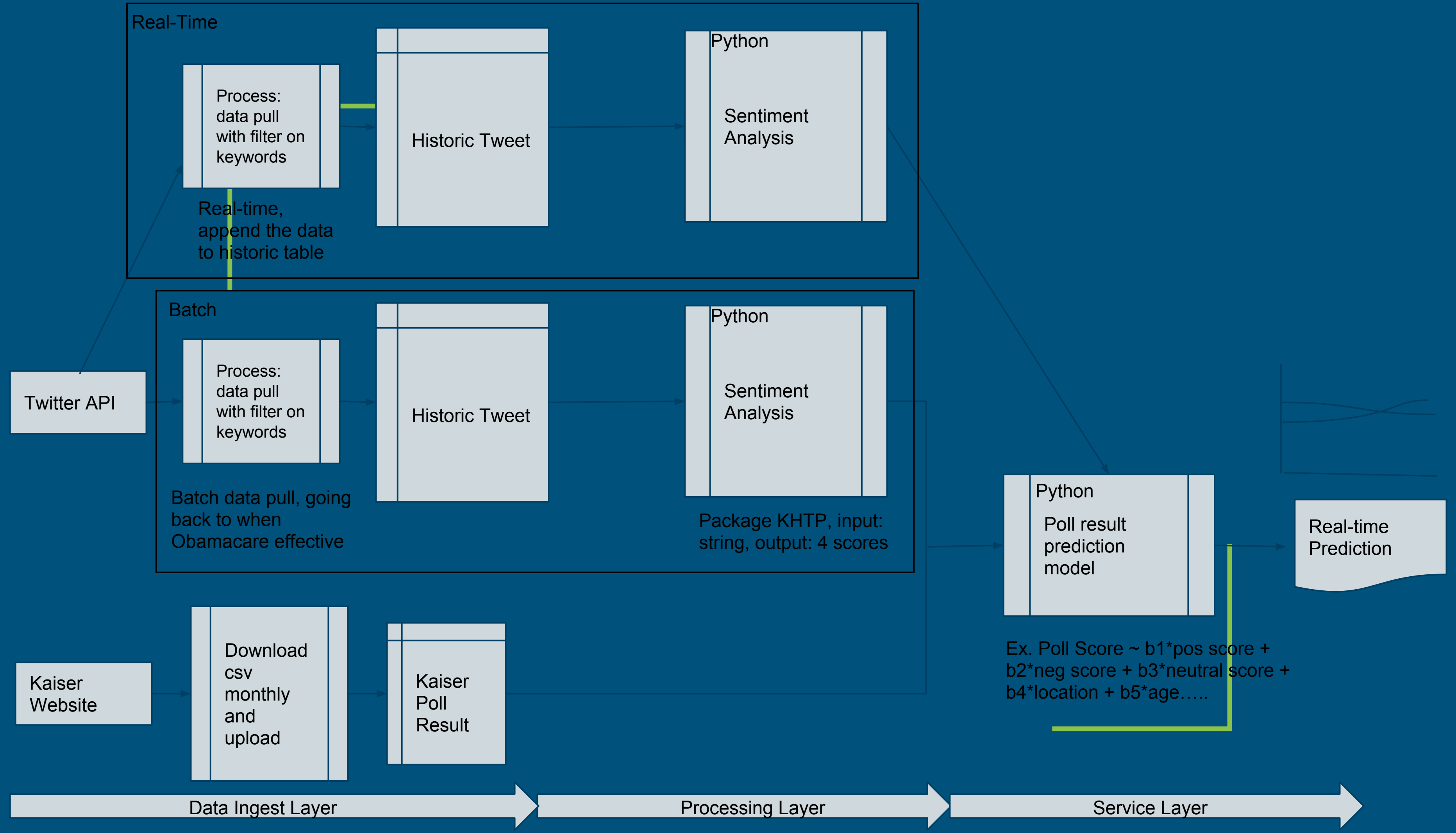
Reference

http://kff.org/interactive/kaiser-health-tracking-poll-the-publics-views-on-the-aca/#?aRange=twoYear

https://github.com/Jefferson-Henrique/GetOldTweets-python

https://market.mashape.com/textanalysis/sentiment-analysis

https://market.mashape.com/japerk/text-processing



Real-Time

Process: data pull with filter on keywords

Real-time, append the data to historic table

Historic Tweet

Python

Sentiment Analysis

Batch

Twitter API

Process: data pull with filter on keywords

Batch data pull, going back to when Obamacare effective

Historic Tweet

Python

Sentiment Analysis

Package KHTP, input: string, output: 4 scores

Python

Poll result prediction model

Ex. Poll Score ~ b1*pos score + b2*neg score + b3*neutral score + b4*location + b5*age…..

Real-time Prediction

Kaiser Website

Download csv monthly and upload

Kaiser Poll Result

Data Ingest Layer

Processing Layer

Service Layer