Berkeley School of Information

Craigslist Price Comparison

Final Project

DATASCI W205 Storing and Retrieving Data
Nate Black, Arthur Mak, Malini Mittal, & Marguerite Oneto
28 April 2015

---

http://blogs.psychcentral.com
http://www.craigslist.org

Problem Statement

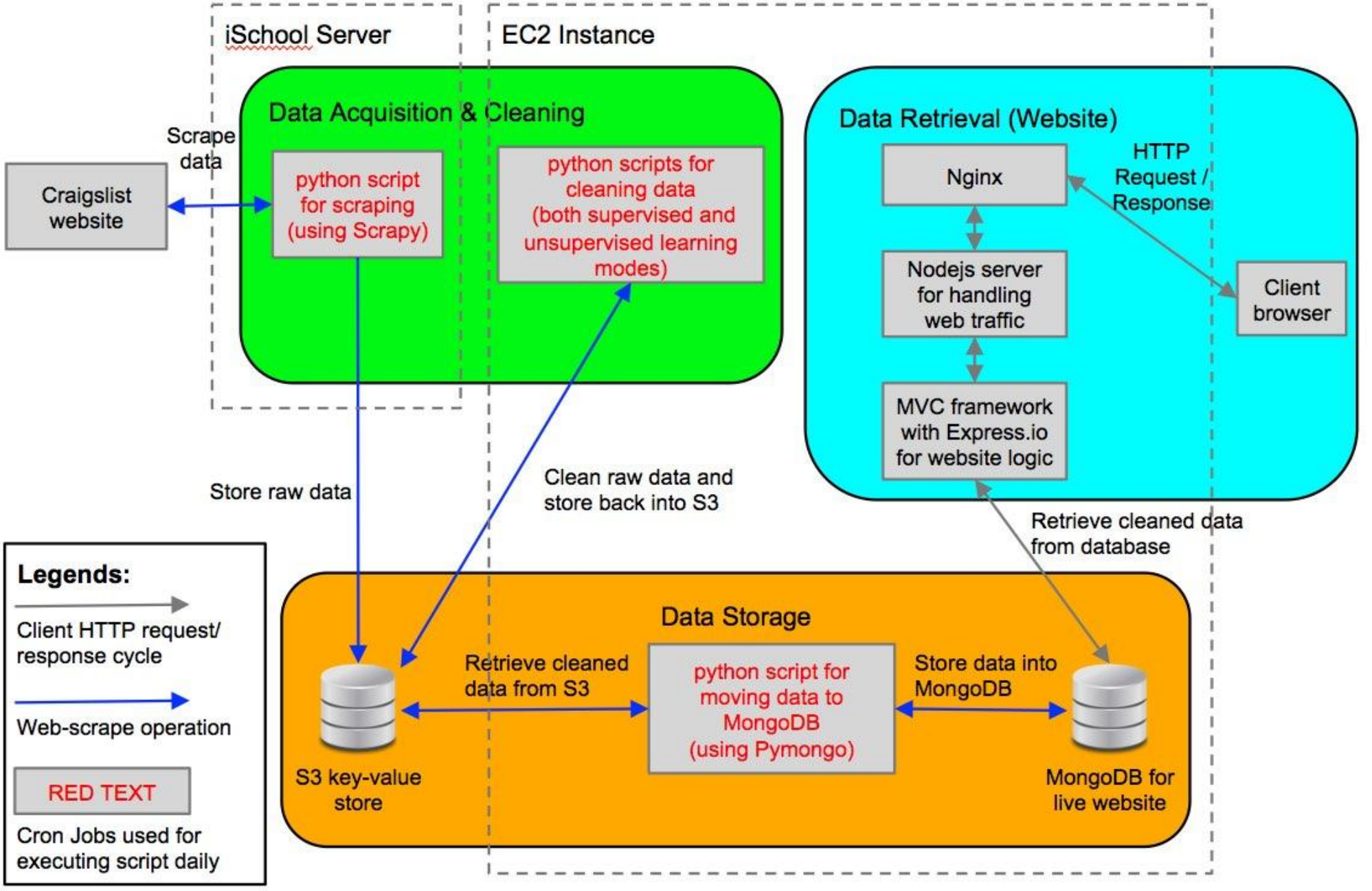
Berkeley School of Information

The goal of this project is to produce a user-friendly application that provides a price comparison interface whenever a user conducts a search on a particular item. The website will be able to answer the following questions:

1. What is the fair market value of the item?
2. What is the price trend for a given city?
3. What is the average price across the country?

Architecture

Berkeley School of Information



Data Collection and Storage

Berkeley School of Information

- iPhone postings collected across 23 different cities daily and stored in S3
- Tools: Scrapy, boto
- 23 scrapers, one for each city
- Main roadblock: Craigslist blocking the IP addresses
- Various measures taken
  - autothrottle mode
  - download delay, randomized
  - a set of commonly used user-agent strings, randomized
  - use multiple EC2 servers for the scrapers
    - IP addresses used by these servers blocked en masse by Craigslist.org
  - randomize the order in which the cities are scraped
  - stop scraping the “text” field was removed
    - reduced the total number of requests by a factor of 100 (from 2400+ to ~26)
- Scrapers run daily via a cronjob on iSchool server

Data Cleaning: Supervised Learning

Berkeley School of Information

- Tools: Pandas, Sci-kit Learn, NumPy, re, boto, and cPickle
- Pull raw data files from S3 and perform the general data scrub
  - Exclude outliers and NaNs
  - Format date and price fields
- Postings are classified in a two-step process:
  - Naive Bayes algorithm classifies data as an iPhone or not an iPhone in
    - Algorithm trained on a random sample of 2,000 postings in Sci-kit Learn
    - 98% classification accuracy
  - Regular expressions are used to further classify iPhones by model type (e.g. 4S, 5, 6 etc.)
    - allows for easy calibration and transparency
    - 97.5% overall classification accuracy on test data
- Trained model is serialized in cPickle and applied to new data flowing through the pipeline with clean data being stored back to S3

Data Cleaning: Unsupervised Learning

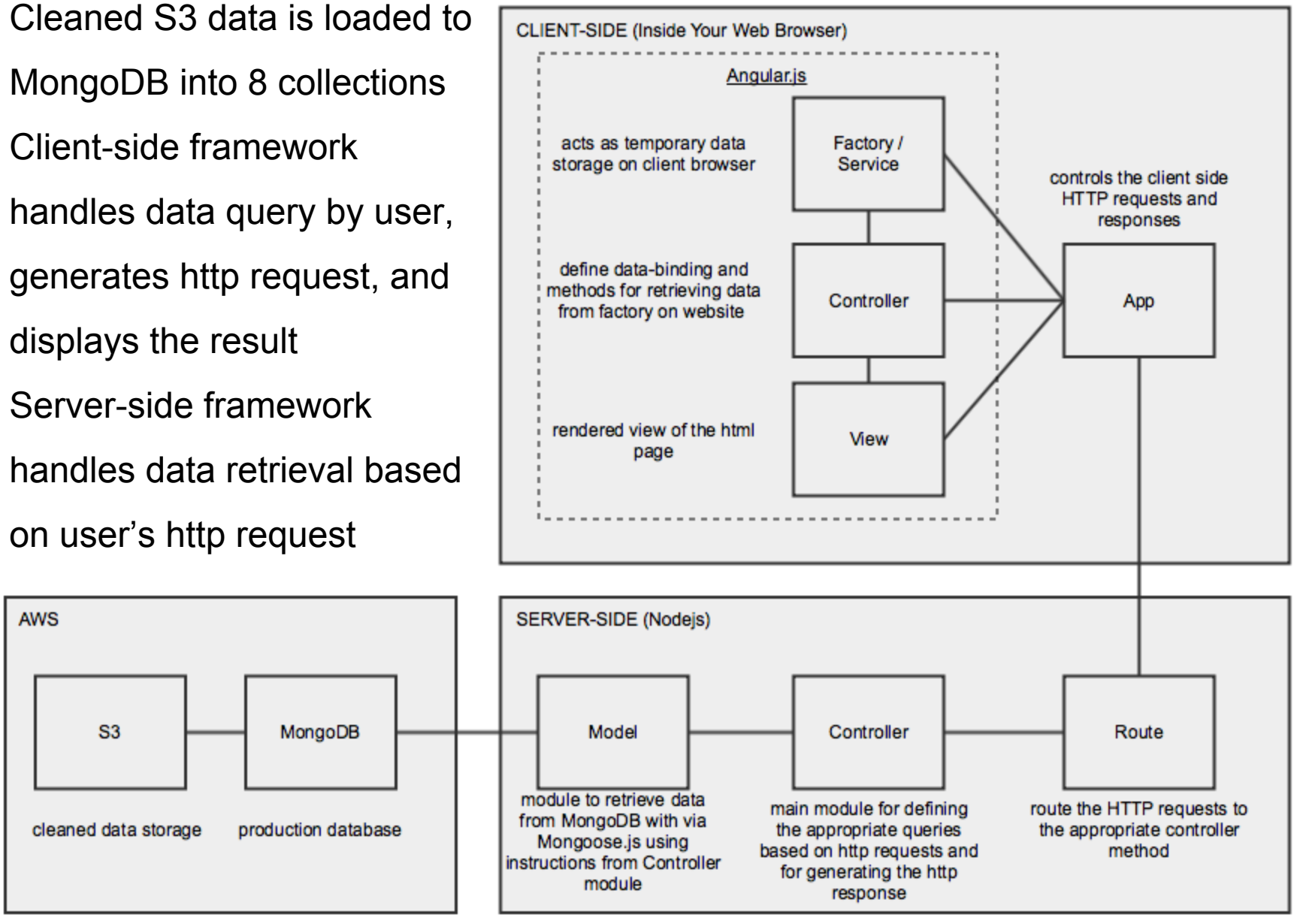
Berkeley School of Information

- Tools: Pandas, NLTK, Gensim, Scipy
- Raw data is pulled from S3 and data scrub performed
- Postings are classified in a two-step process:
  - Latent Semantic Indexing (LSI)
    - Dictionary and Corpus of today's posting titles are created
    - Corpus words turned into weighted vectors using TF-IDF (Term Frequency-Inverse Document Frequency) model
    - TF-IDF vector dimensionality further reduced using LSI model (choose # of topics = dimensions = 100)
  - K-Means Clustering
    - LSI vectors clustered using k-means (choose # of clusters = 10)
    - Each cluster's centroid vector mapped to #1 topic, #1 word in topic (4, 4S, 5, 5C, 5S, 6, 6 Plus)
    - Could only classify 2/3 of postings
    - Model accuracy of 99.4% of what it could classify
  - Cleaned data returned to S3 with predicted iPhone category now attached

Data Retrieval

Berkeley School of Information

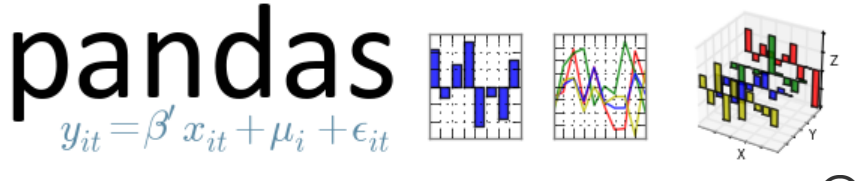
- Cleaned S3 data is loaded to MongoDB into 8 collections
- Client-side framework handles data query by user, generates http request, and displays the result
- Server-side framework handles data retrieval based on user's http request



Tools

Berkeley School of Information

UC Berkeley
School of Information

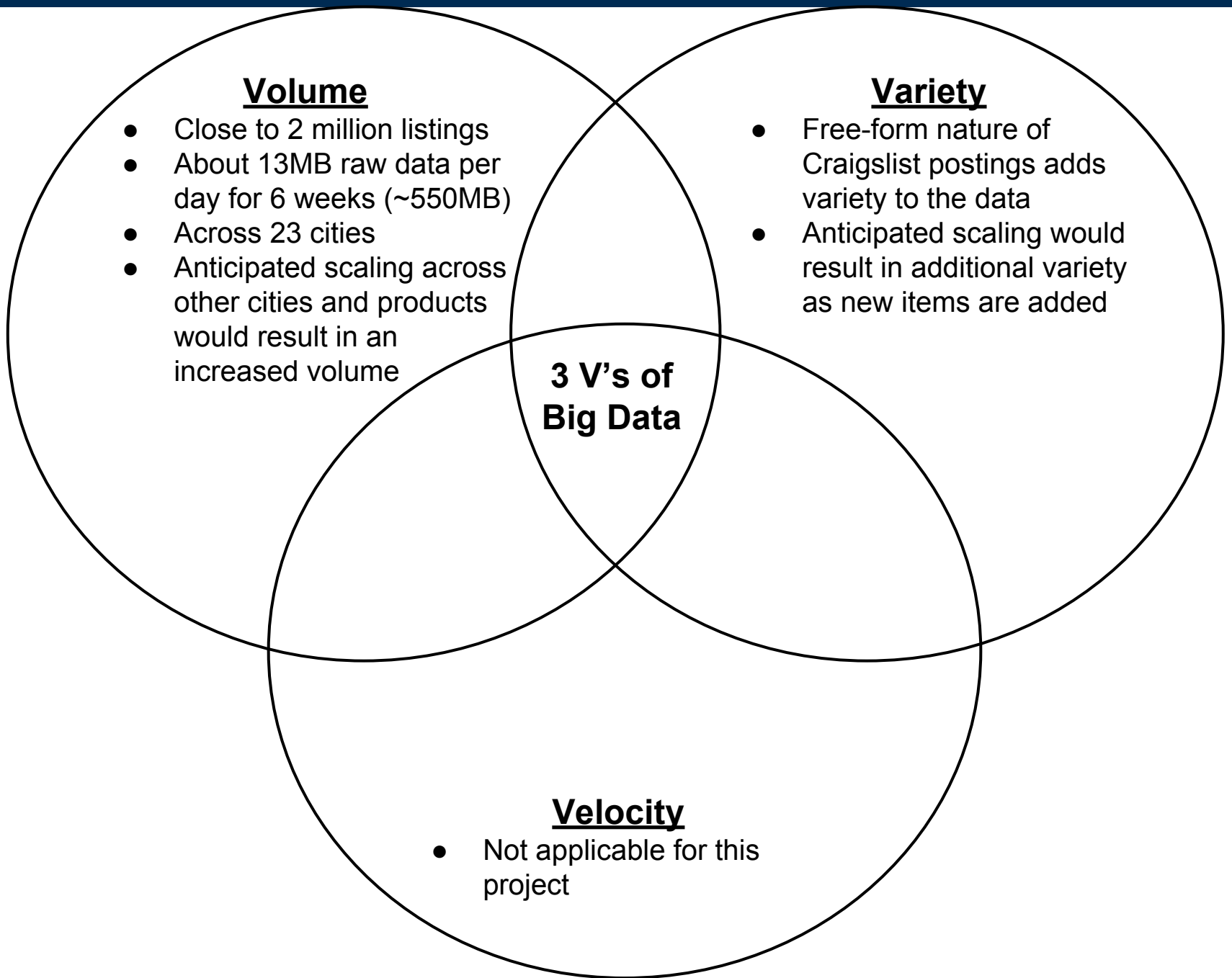


express

mongoDB

3 V's

Berkeley School of Information



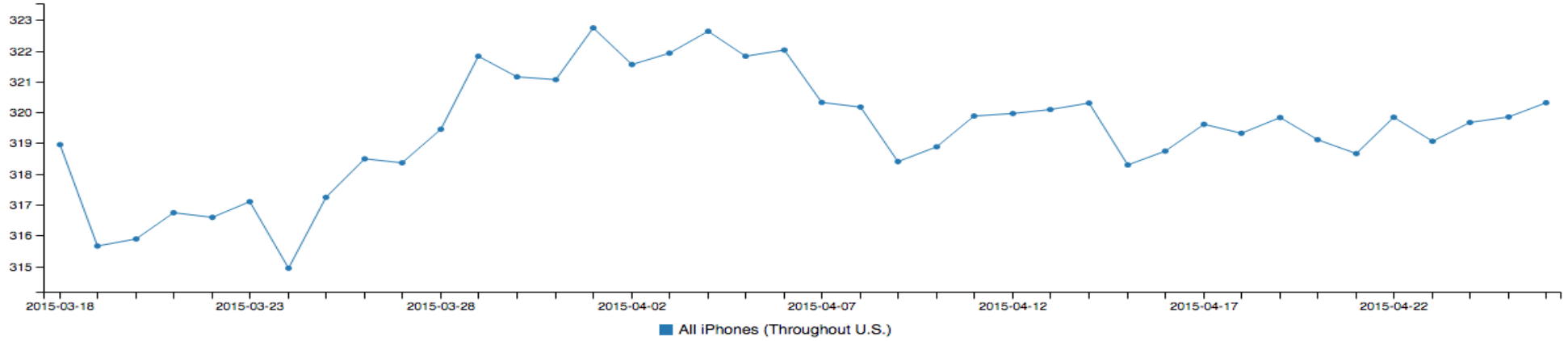
Results

Berkeley School of Information

Your Results Based on 1099000 Listings (supervised mode) ... Data Retrieval Completed

Latest Average Price for All iPhones on (2015-04-26): $320.31

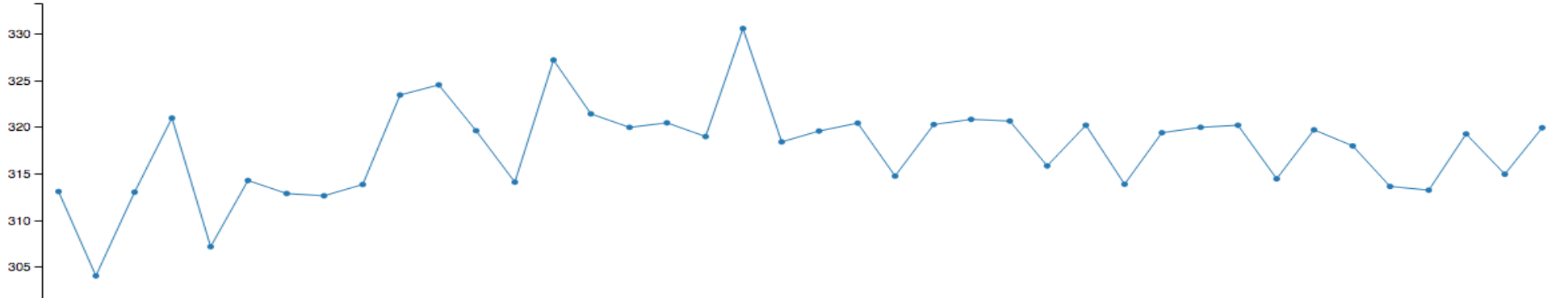
Daily Average Price for All iPhones (Throughout U.S. from 2015-03-17 to 2015-04-27)



Your Results Based on 849731 Listings (unsupervised mode) ... Data Retrieval Completed

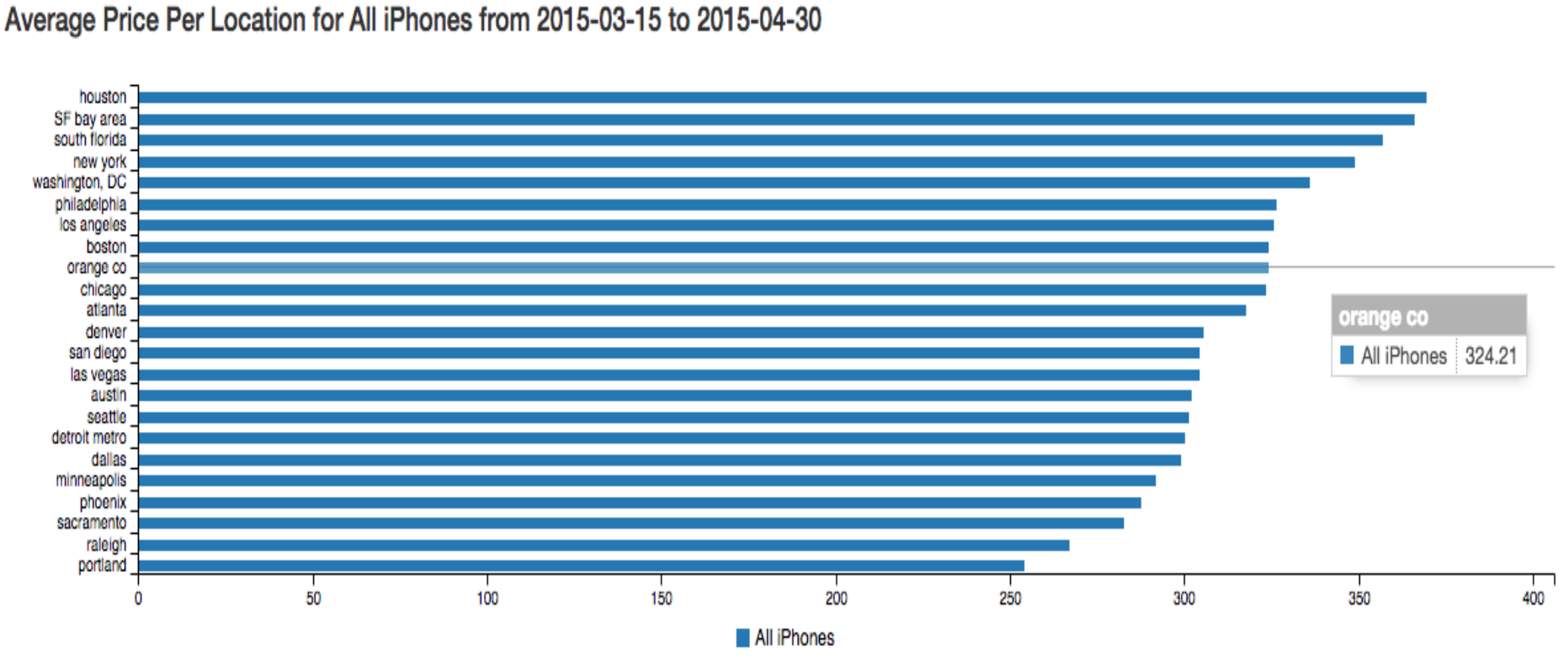
Latest Average Price for All iPhones on (2015-04-26): $319.90

Daily Average Price for All iPhones (Throughout U.S. from 2015-03-17 to 2015-04-27)



Results contd.

Berkeley School of Information



Results contd.

Berkeley School of Information

Listing of cheapest items for All iPhones (Throughout U.S. from 2015-03-15 to 2015-04-30)

Filter listings... **Sort by:** Lowest Price

Product	Description	Price	City	Date Listed
iPhone 4	Iphone 4 8gb Verizon	$63	seattle	2015-04-16
iPhone 4	Iphone 4 8gb Verizon	$63	seattle	2015-04-06
iPhone 6+	Iphone 6 plus like new atnt 64g	$65	SF bay area	2015-04-24
iPhone 4s	iPhone 4S white unlocked	$65	new york	2015-04-24
iPhone 4	Verizon iPhone 4 / Clean Imei	$65	los angeles	2015-04-24
iPhone 4s	Black Sprint iPhone 4s - 16g	$65	south florida	2015-04-24
iPhone 6+	telefono IPhone 4 para Verizon o Page plus	$65	los angeles	2015-04-24
iPhone 4	verizon - IPhone 4 / Clean Imei - ready to activate or port	$65	los angeles	2015-04-24
iPhone 4s	iPhone 4s AT&T. CHEAP!!!!!	$65	portland	2015-04-23
iPhone 5	iPhone 5	$65	philadelphia	2015-04-23
iPhone 3g	iPhone 3GS- Black AT&T	$65	los angeles	2015-04-23
iPhone 3g	IPhone 3GS 16 unlock great condition	$65	dallas	2015-04-23
iPhone 4s	Used IPhone 4s model A1387 8mb	$65	dallas	2015-04-23

Challenges

Berkeley School of Information

- Scraping the data
  - IP block
  - EC2 instance IPs are blocked from Craigslist
  - Creating realistic HTTP requests
  - Scrapy not able to create files >1MB directly on S3
  - Max. items scraped per scraper = 2400
- Setting up mrjob/EMR
  - Correct options in mrjob.conf
  - Debugging a challenge
  - The map/reduce job on EMR took ~15 minutes, as compared to a few seconds on local machine
- Tagging iPhones - NLP required
  - Many non-iPhone postings are scraped from Craigslist
  - Inconsistent Gensim performance
- Writing MongoDB queries
  - MongoDB query function ran out of default memory when aggregating the data
  - Our document output exceeded MongoDB limit of 16MB
  - Python version incompatibilities
  - Need to pre-process data to improve query speed

Possible Future Work

Berkeley School of Information

- Include more items from Craigslist to obtain a price index
- Include more cities
- Include different countries
- Create a Price Index for each country
- Alter the process to look for arbitrage opportunities between cities
- Include more websites like eBay.com, Backpage.com, geebo.com, …
- Cleanup listings when they are no longer active
- Move the process to a platform with an API so it can be expanded upon

Website

Berkeley School of Information

Please visit http://www.priceright.info/ to view the final product and find the used iPhone of your dream!