# Business Performance Estimate

## How To Start New Businesses Based On Customers' Reviews

CS9233 - Programming for Big Data
Tianyu Wang (tw1707), Yida Zhou (yz4499)

# Outline

- Introduction
- Data Source
- Workflow
- Result
- Conclusion & Future Work

# Outline

- **Introduction**
- Data Source
- Workflow
- Result
- Conclusion & Future Work

# Why *location* is so important?

- Directly affect the source of actual audience
- Hard to estimate w/o actual practice and testing
- Time and money consuming
- Risky for new openings
- Also, difficult to figure out key attributes strongly affecting the rating

# What can *big data* do?

- Suggest a potential opening in designated area
- Estimate performance and rating based on the given location
- Find out the most important factors combined with *machine learning*

# Target Audience?

- Startup or individuals, the ones whom have little or no experience, that willing to take less risks
- Corporate businesses companies, which want to expand their businesses

# Outline

- Introduction
- **Data Source**
- Workflow
- Result
- Conclusion & Future Work

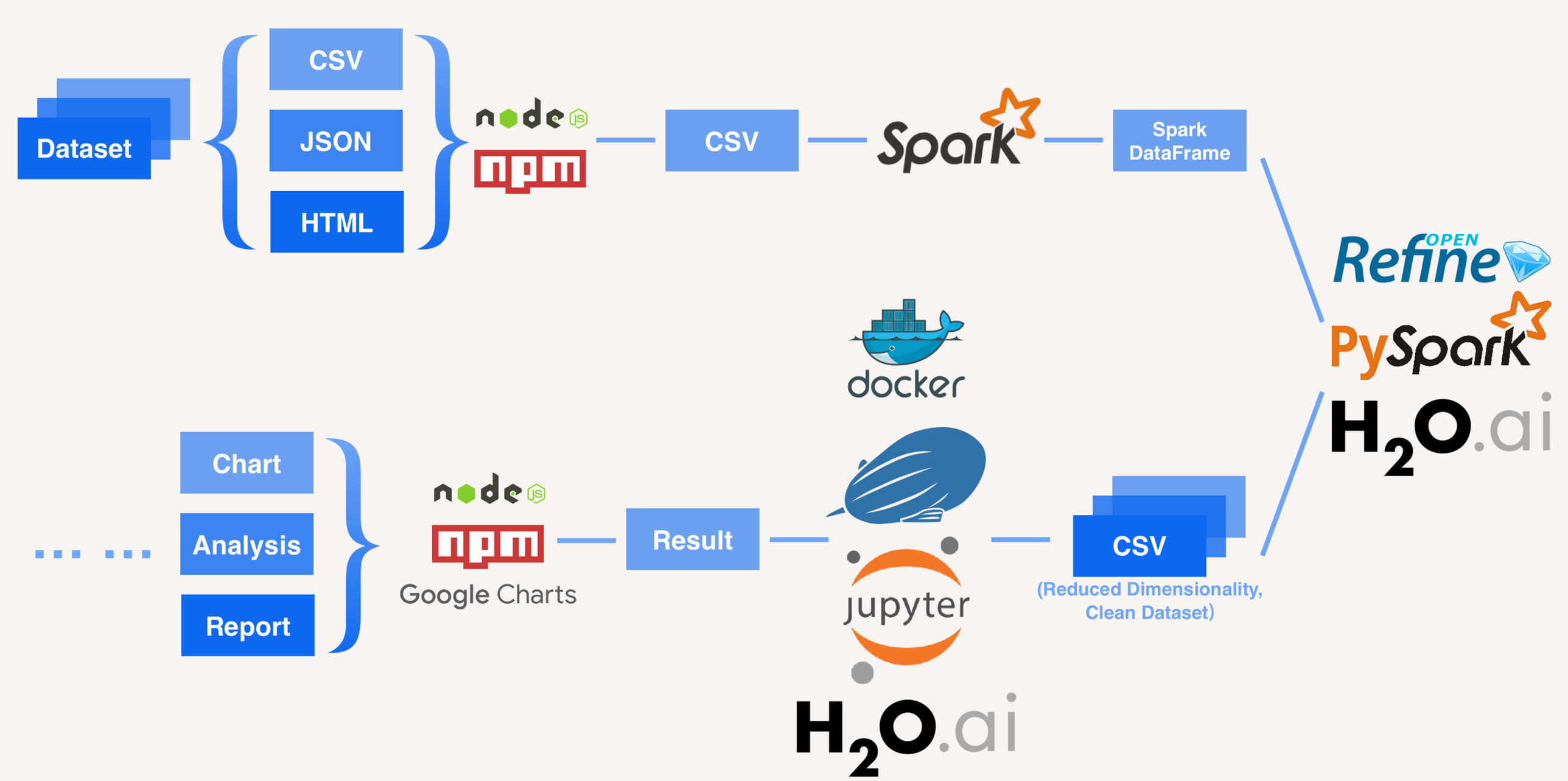
# Data Source

- Yelp / FourSquare
  - 200k businesses
  - 5.2m ratings
- Financial statements from related corporate businesses
- data.gov: statistics of U.S. businesses

# Outline

- Introduction
- Data Source
- **Workflow**
- Result
- Conclusion & Future Work

# Workflow


Dataset
CSV
JSON
HTML
node
npm
CSV
Spark
Spark DataFrame
OPEN Refine
PySpark
H2O.ai
docker
Chart
Analysis
Report
node
npm
Google Charts
Result
jupyter
H2O.ai
CSV
(Reduced Dimensionality, Clean Dataset)

# Workflow (Cont.)

- Formatter: convert `.csv`, `.json` and `.html` file into `.csv`
  - nodeJS & npm: `json2csv` package
  - Manually collect some useful info from the document
- Import `.csv` files into Spark:
  - Flatten attributes
  - Generate Spark DataFrames

# Workflow (Cont.)

- Data Preprocessing:
  - Reduce dimensionality:
    - Feature selection & extraction
    - Principal Component Analysis (PCA)
    - Avoid overfitting
  - Clean data

# Workflow (Cont.)

- Data Preprocessing:
  - Export into multiple .csv files:
    - Different categories
    - Different cities
    - etc.

# Workflow (Cont.)

- Big Data Analysis:
  - Spark SQL module
  - $H_2O$ Machine Learning
- Final Result:
  - Chart
  - Analysis

# Outline

- Introduction
- Data Source
- Workflow
- **Result**
- Conclusion & Future Work

# Top 10 Categories For Each State

Chart

| Category | Rating |
|---|---|
| Shaved Ice | 4.408471787 |
| Gelato | 4.378077256 |
| Coffee Roasteries | 4.374378378 |
| Cupcakes | 4.345326146 |
| Street Vendors | 4.328106969 |
| Poke | 4.2968294 |
| Local Services | 4.22552278 |
| Lebanese | 4.224629437 |
| Internet Cafes | 4.19702228 |
| Polish | 4.19209245 |

# Top 10 Cities For Chinese Restaurant

Chart

| City | Rating |
|---|---|
| Aurora | 3.782264274 |
| Pickering | 3.73193294 |
| Newmarket | 3.729253363 |
| Stuttgart | 3.592544663 |
| Gilbert | 3.560078808 |
| Gastonia | 3.559007997 |
| Edinburgh | 3.558972629 |
| Huntersville | 3.54797137 |
| Fort Mill | 3.47231295 |
| Montreal | 3.465306657 |

# Compare Machine Learning Methods

- Chart (Group By Method)
- Chart (Group By Norm)

| Method | MSE | RMSE | $r^2$ | mean-residual-deviance | mae | rmsle |
|---|---|---|---|---|---|---|
| Random Forest | 0.099864 | 0.316013 | **0.780328** | 0.099864 | 0.230048 | 0.083253 |
| Deep Learning | 0.428654 | 0.654717 | **0.057082** | 0.428654 | 0.503647 | 0.163133 |
| Gradient Boosting | 0.17661 | 0.42025 | **0.611507** | 0.17661 | 0.282799 | 0.108599 |
| XGBoost | 0.408734 | 0.639323 | **0.100901** | 0.408734 | 0.494821 | 0.159414 |

# Find the Best Location For Opening

TO BE ADDED

# Attributes

- `Latitude` comes first:
  - Location is really important

| Attribute | Priority (%) |
|---|---|
| Latitude | 0.0469 |
| Accepts Credit Cards | 0.0433 |
| Price Range | 0.0349 |
| Good For Kids | 0.0132 |
| Good For Groups | 0.012 |
| Reservations | 0.0108 |
| Take Out | 0.0108 |
| Outdoor Seating | 0.0096 |
| Wi-Fi | 0.0096 |
| Noise Level | 0.0084 |
| Alcohol | 0.0084 |

# Estimate Rating Using Machine Learning Model

TO BE ADDED

# Outline

- Introduction
- Data Source
- Workflow
- Result
- **Conclusion & Future Work**

# Conclusion - Estimated Objectives

- For Individuals:
  - New businesses opening suggestions
  - Given category / area
- For Corporate Businesses:
  - Where to expand
  - How to improve ratings

# Conclusion - Estimated Objectives

- For Both:
  - Key attributes affecting on ratings

# Future Work

- If given more data:
  - Time lapse: trends
  - More city data
- Rating vs. Profit
  - Long-term eyesight

# Q&A

# Thank you!