Introduction to Data Science Pipeline

Kunal Malhotra
kunal.malhotra1@ibm.com

IBM
CODE

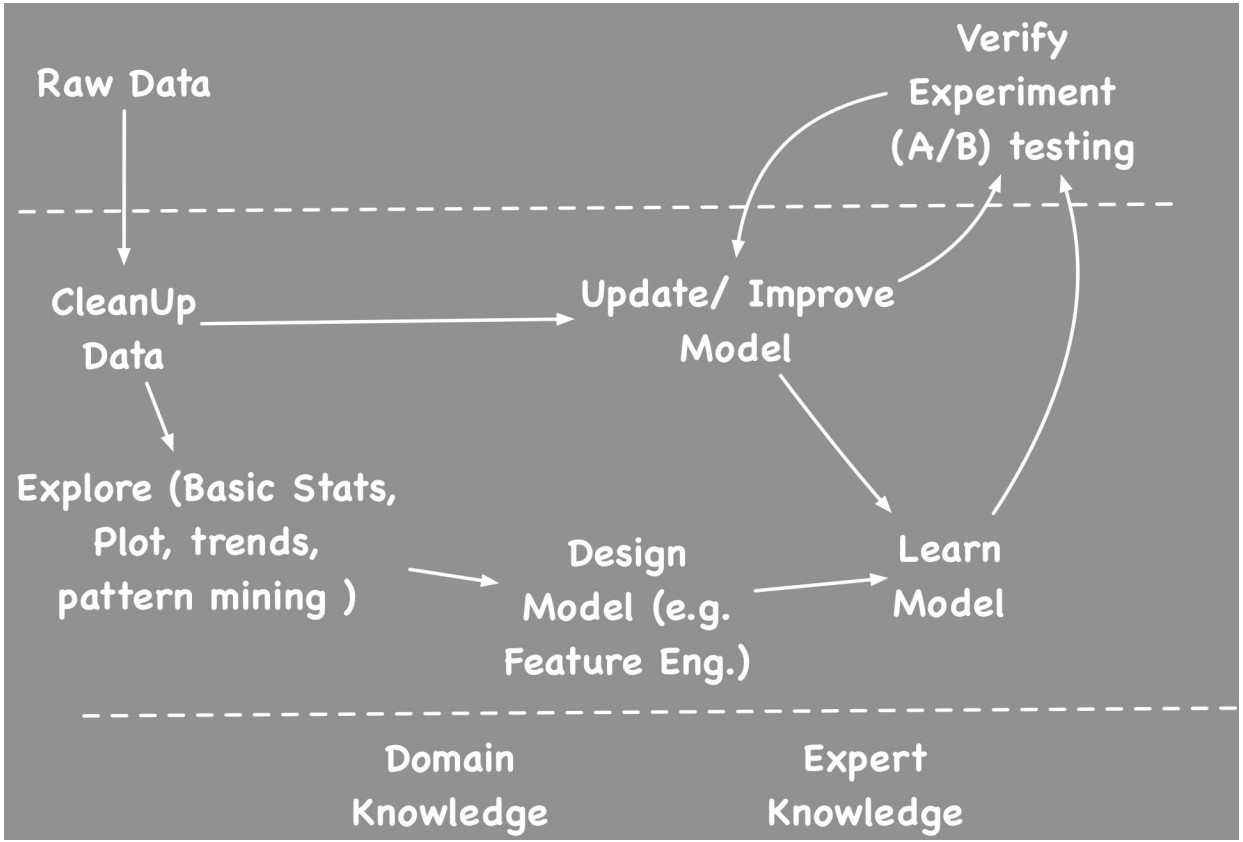
What is Data Science?

Extraction of knowledge from large volumes of data that are structured or unstructured.

It is a continuation of the fields **data mining** and **predictive analytics**

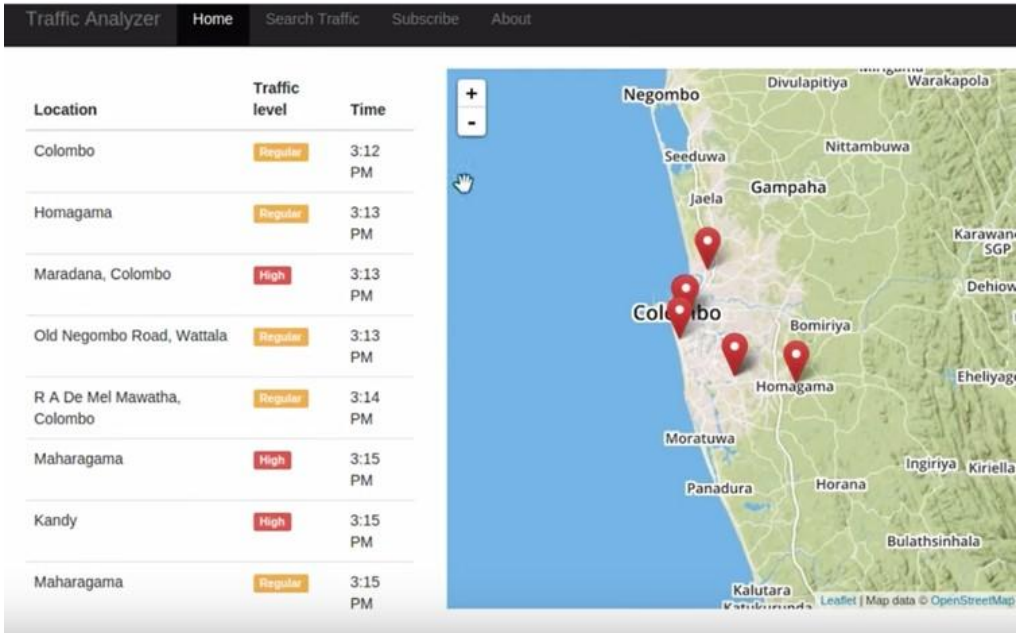
IBM CODE

Data Science Pipeline



Example (Road.lk) traffic Feed

1. Data as tweets
2. Extract time, location, and traffic level using NLP
3. Explore data
4. Model based on time, and it is a holiday
5. Predict traffic given a time and location.



IBM CODE

Data Cleanup

Real data is messay, often needs to cleaned up before useful.

- Bad formats - ignore or treat like missing data
- Missing Data - extrapolate or remove data line
- Useless variables - remove
- Wrong data - e.g. aaa, bbb, joe, some might be deliberate lie, or 99 may be a code for N/A

Data Cleanup (Contd.)

- Transform variables (date formats, String to int)
- Create derived variables
  - Derive country from IP
  - age from ID card number
- Normalize strings
  - e.g. stemm or use phonetic sounds
  - different spellings and nicknames (William->Bill)
- Feature value re scaling (e.g. most ML algorithms needs value to rescaled to 0-1 range).
- Enrich (e.g. lookup and add age from profile)

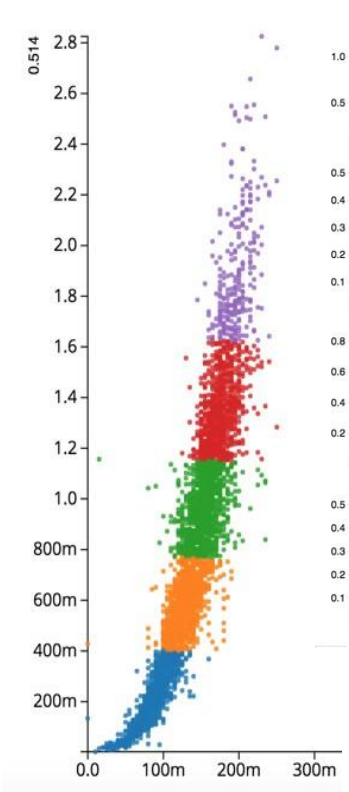
Data Exploration

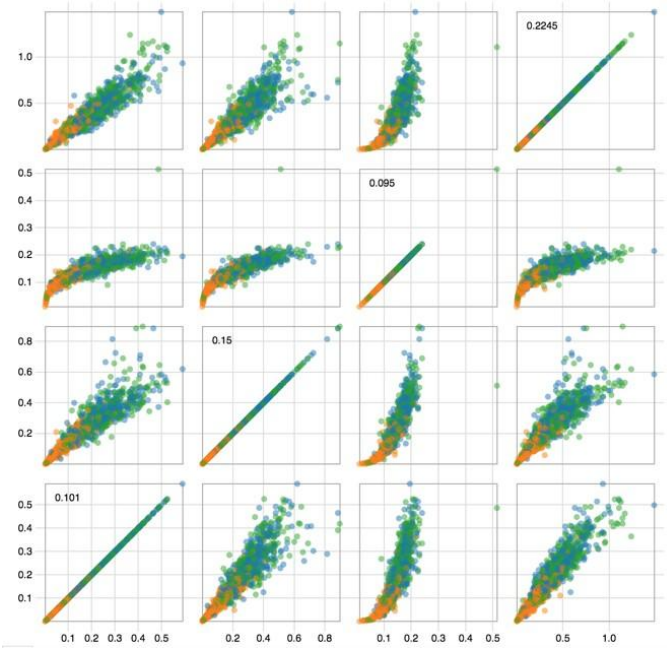
- Understand, and get a feel for what is **expected (models => densities, constraints)** and **unexpected/ residuals (errors, outliers)**
- Think what this is data about? domain, background,
  - how it is collected, what each fields mean and range of values.
- Head, tail, count, all descriptives (Mean, Max, median, percentiles ..) - Five number Summary. Min. 1st Qu. Median Mean 3rd Qu. Max.
- Run a bunch of count/group-by statements to gauge if I think it's corrupt.

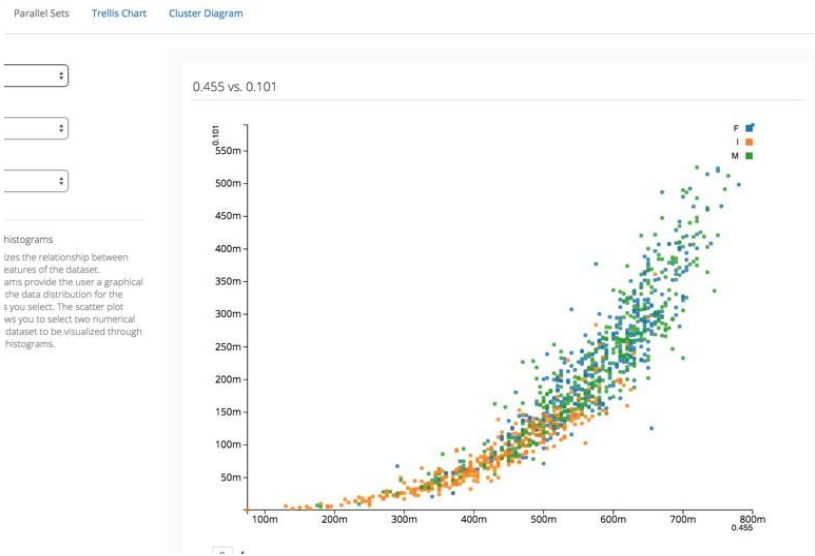
Data Exploration (Contd.)

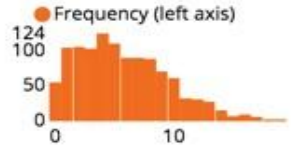
- Plot - take random sample and explore (scatter plot)
  - e.g. Draw scatter plot or Trellis Plot
- Find Dependencies between fields
  - Calculate Correlation
  - Dimensionality reduction
  - Cluster and look visualize clusters
- Look at frequency distribution of each field and try to find a known distribution if possible.

Data Exploration (Contd.)







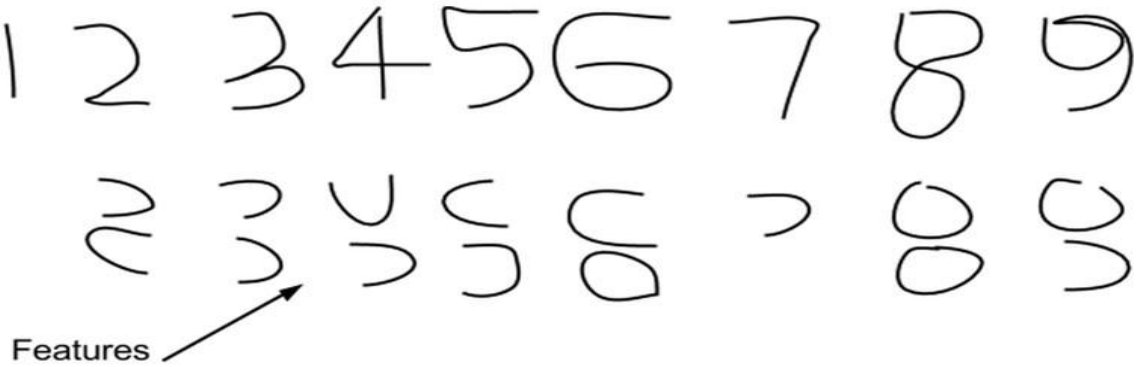


Mean: 0.182 Median: 0.165

Std: 0.111 Skewness: 0.602

IBM CODE

Feature Engineering



- Feature engineering is the art of finding feature that leads simplest decision algorithm. (Good features allow a simple model to beat a complex model.)
- Best features may be a subset, or a combination, or transformed version of the features.

How to do Feature Engineering?

- Manually pick by domain experts and trial and error.
- Search the possible combinations by training and combining subsets (e.g. Random Forest)
- Use statistical concepts like correlation and information criteria
- Reduce the features to a low dimension space using techniques like PCA.
- Automatic Feature Learning though Deep Learning

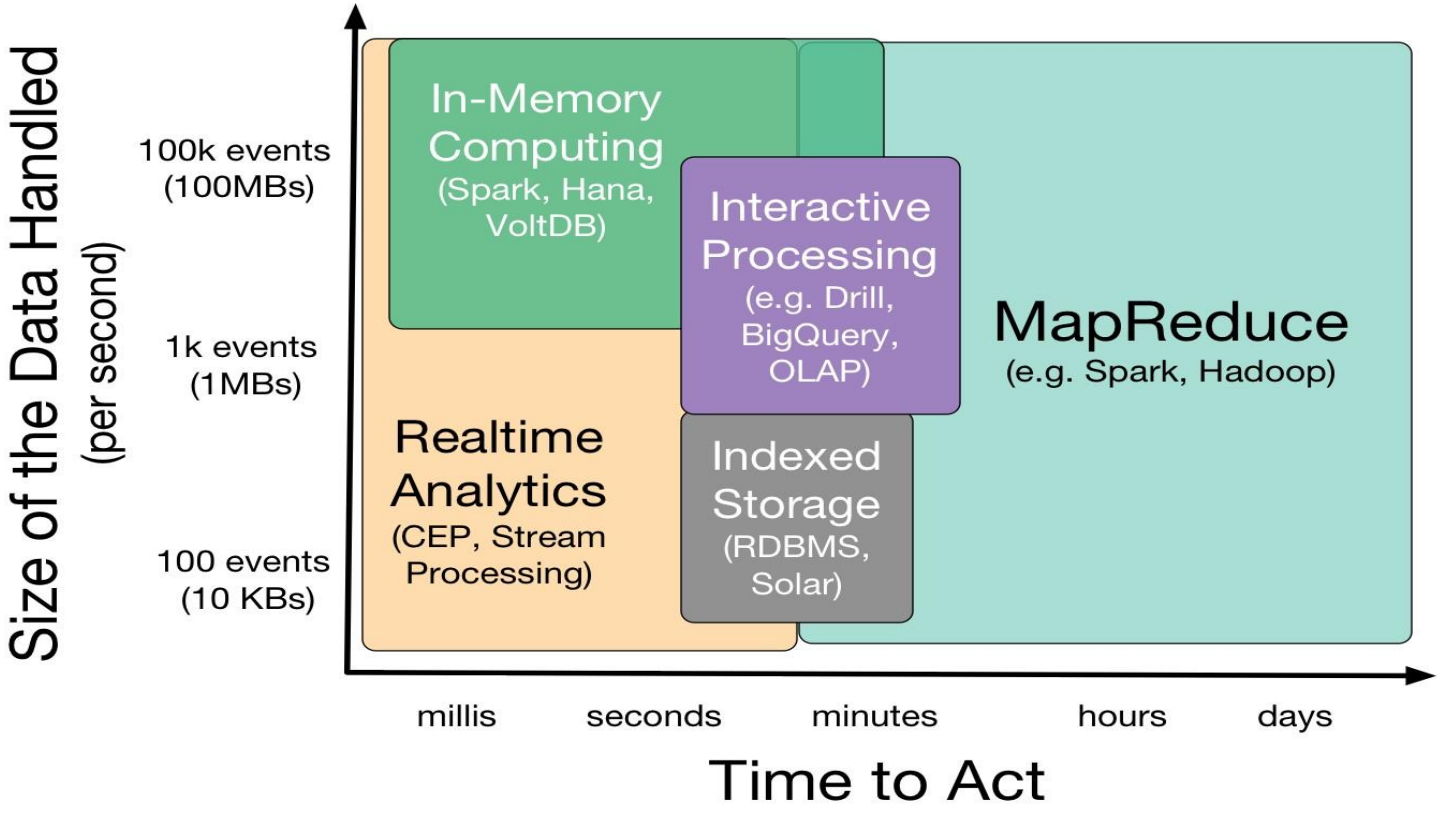
Analysis

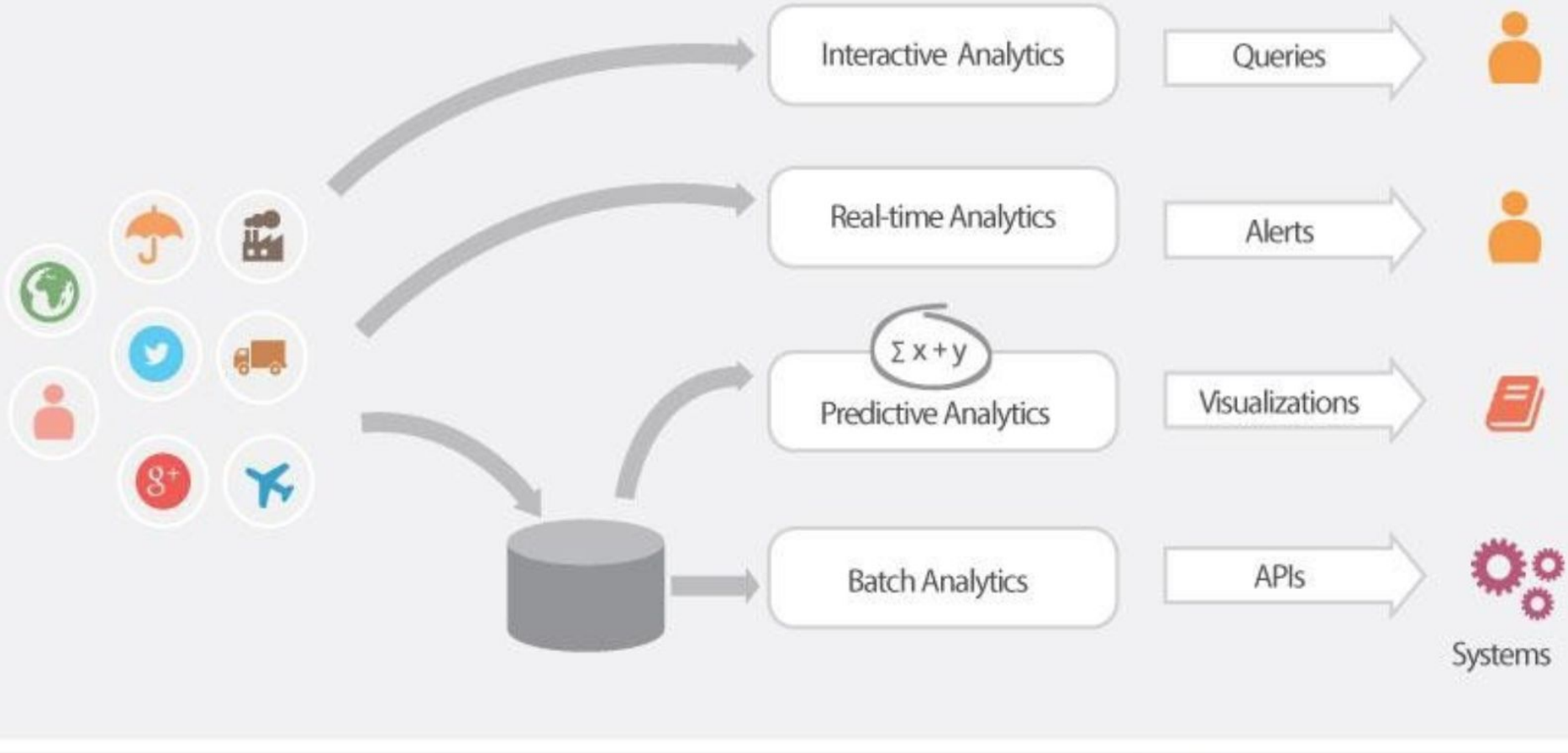
- Goal of analysis is to extract knowledge
- This knowledge usually come in one of the two forms
  - KPI (Key Performance Indicators)
    - Describe key measurement for what is being measured. (e.g. revenue per year, profit margin, revenue for sqft in retail, revenue per employer)
  - Models to describe or predict the data
    - e.g. Machine Learning models or Statistical models

4 Analysis types by time to decision

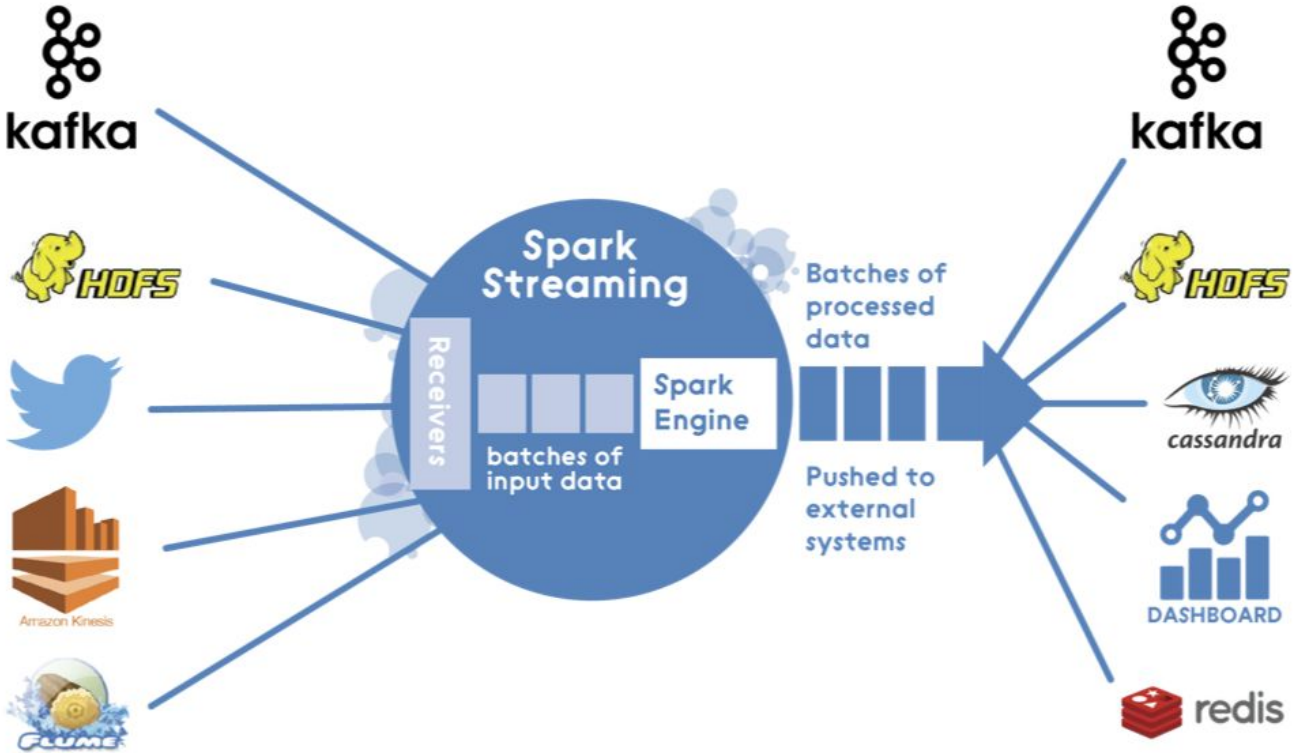
- Hindsight (what happened?)
  - Done using Batch Analytics like MapReduce
- Oversight (what is happening?)
  - Done using Real Time Analytics technologies like CEP
- Insight (why things happening?)
  - Done with Data Mining and Unsupervised learning algorithms like Clustering
- Foresight (what will happen?)
  - Done by building models using Machine learning or one of other techniques

Data Analytics Tools Landscape

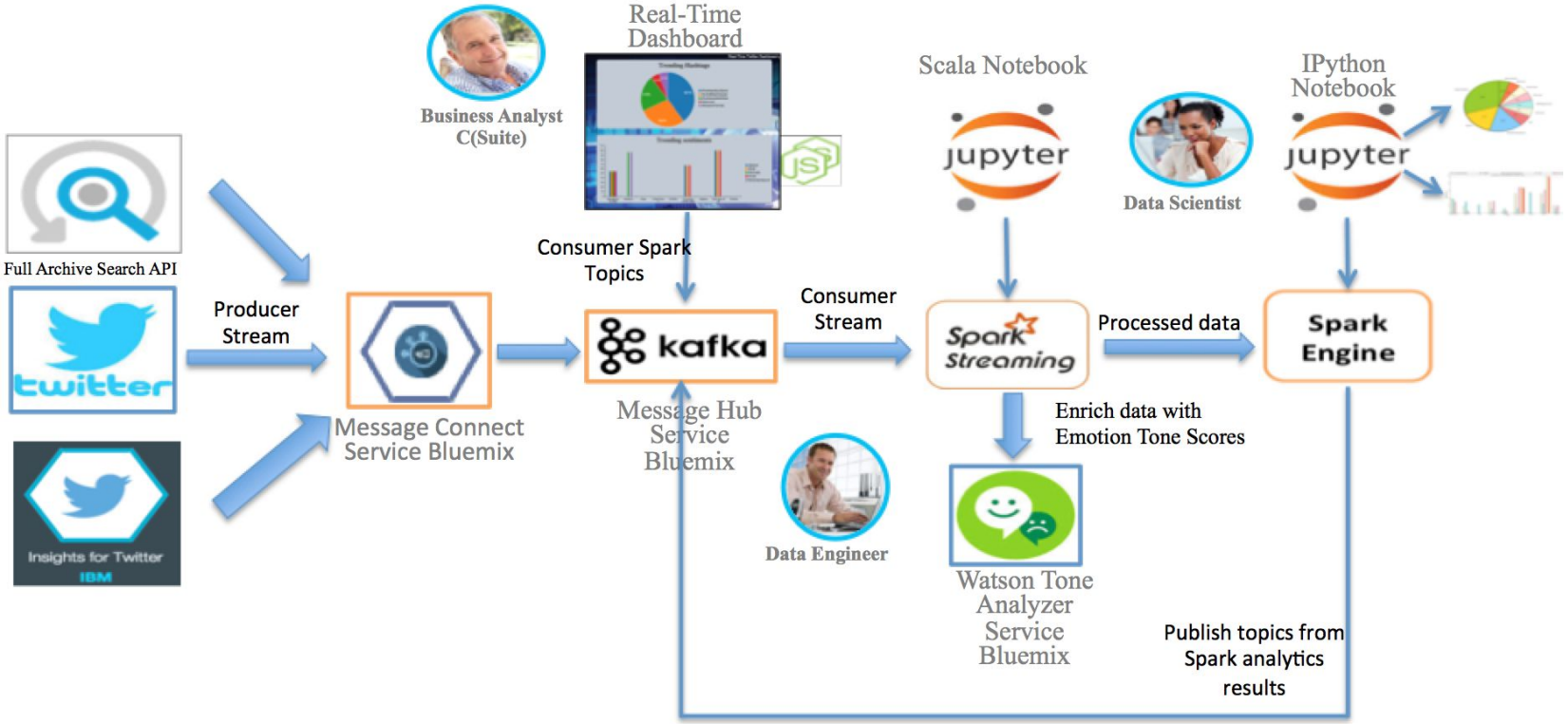



Interactive Analytics
Queries
Real-time Analytics
Alerts
Σ x + y
Predictive Analytics
Visualizations
Batch Analytics
APIs
Systems
Collect Data
Analyze & Make Decisions
Communicate

Batch Analytics



Real Time Analytics: Complex Event Processing



IBM CODE

Building A Decision Model

A model describe how a system behave when input changes. There are many ways to build models.

1. Regression models and ML Models Time series models.
2. Statistical models.
3. Physical Models - based on physical phenomena. For example flight models, space flight models weather models.
4. Mathematical Models

Pitfalls: Experiment vs Observation

1. If you follow scientific method, you would do experiments, and they have control sets (A/B) tests.
2. Big Data does not have a control set, it is rather observations. (we observe the world as it happens)
3. So what we can tell are limited.
4. Correlation does not imply Causality!!
    a. Send a book home example.
    b. All big buyers have free shipping.

Causality: What can we do?

1. Option 1: We can act on correlation if we can verify the guess or if correctness is not critical (Start Investigation, Check for a disease, Marketing).

2. Option 2: We verify correlations using A/B testing or propensity analysis.