Introduction to Data Science and Analytics

Summer School 2015

Srinath Perera

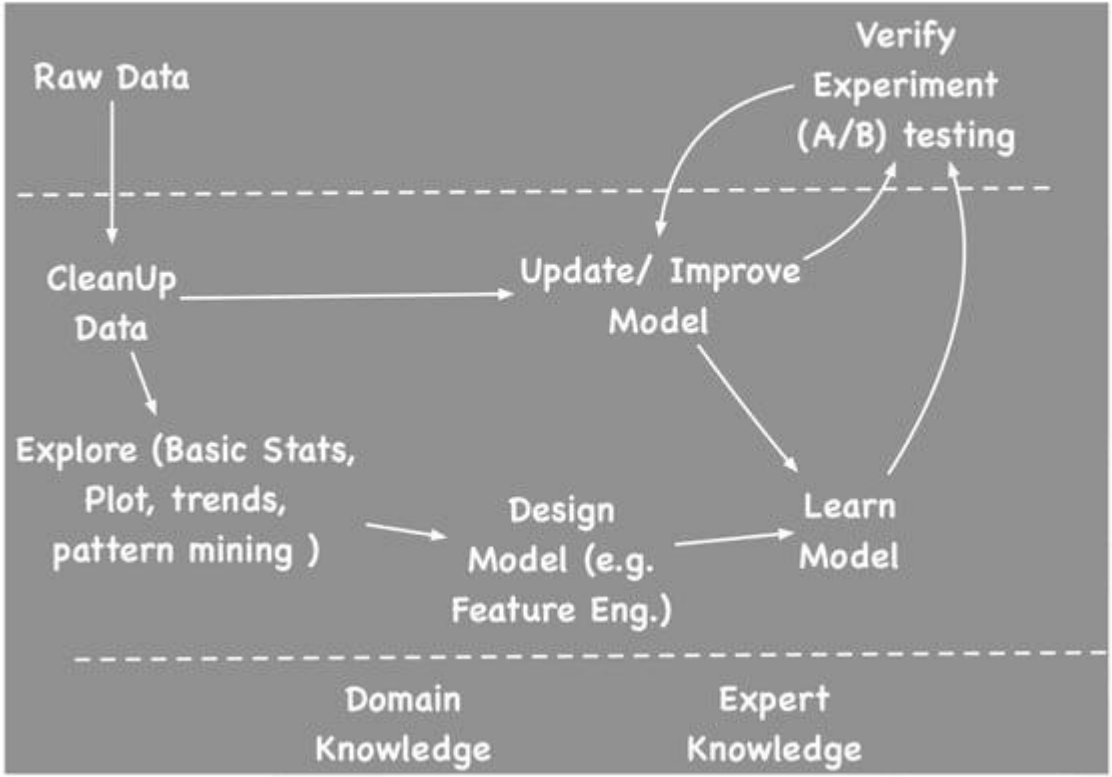
VP Research
WSO2 Inc.

What is Data Science?

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

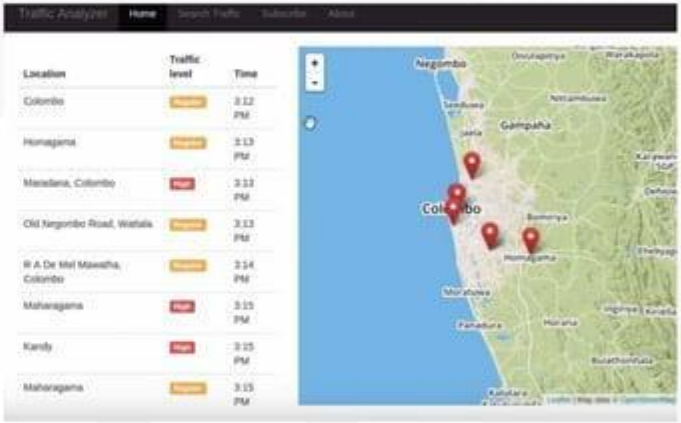
It is a continuation of the fields **data mining** and **predictive analytics**

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.



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 rescaling (e.g. most ML algorithms needs value to rescaled to 0-1 range).
- Enrich (e.g. lookup and add age from profile)

Data Exploration

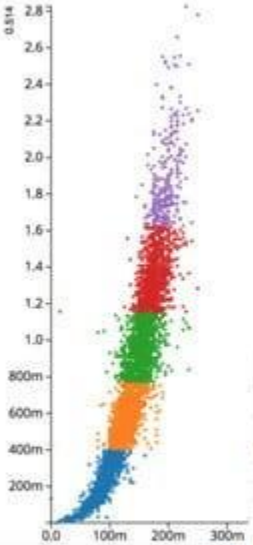
Understand, and get a feel for what is **expected (models => densities, constraints)** and **unexpected/ residuals (errors, outliers)**

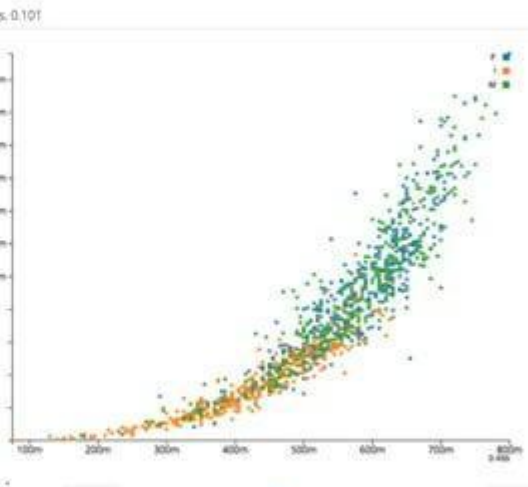
- think what this is data about? domain, background, how it is collected, what each fields mean and range of values.
- head, tail, count, all descriptives (Mean, Max, median, percentiles ..) - Five number Summary. Min. 1st Qu. Median Mean 3rd Qu. Max.
- run a bunch of count/group-by statements to gauge if I think it's corrupt.

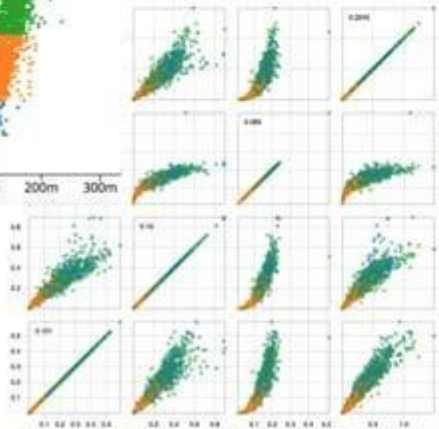
Data Exploration (Contd.)

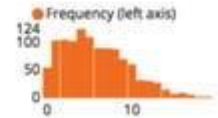
- Plot - take random sample and explore (scatter plot)
  - e.g. Draw scatter plot or Trellis Plot
- Find Dependencies between fields
  - Calculate Correlation
  - Dimensionality reduction
  - Cluster and look visualize clusters
- Look at frequency distribution of each field and try to find a known distribution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

Data Exploration (Contd.)





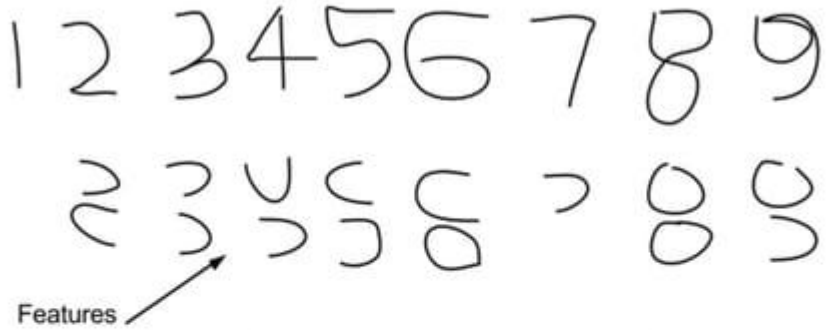




Mean: 0.182 Median: 0.165

Std: 0.111 Skewness: 0.602

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
- ...

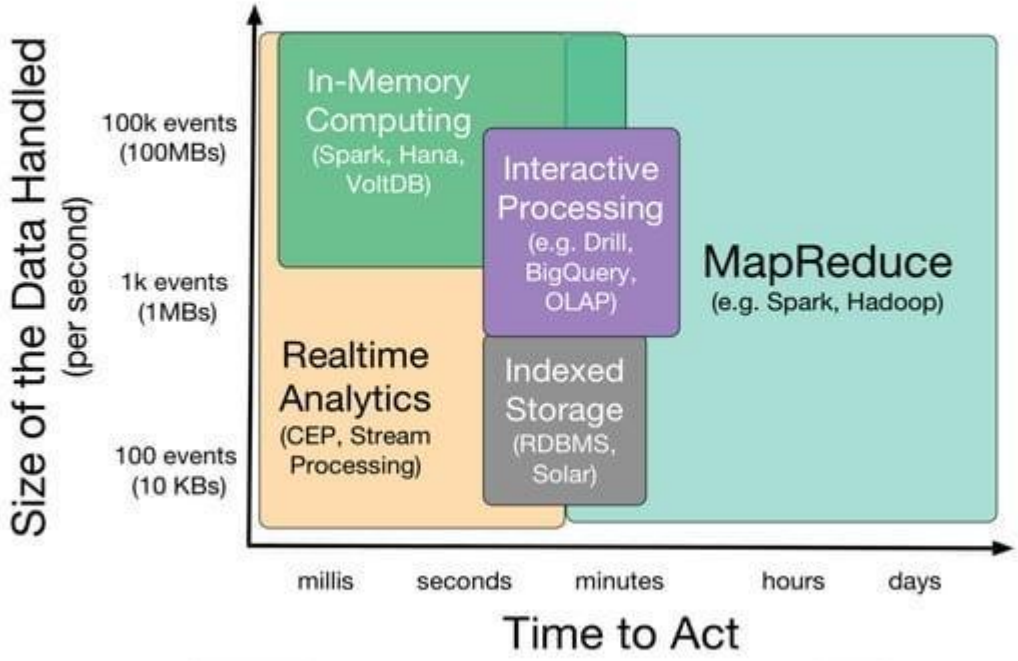
Analysis

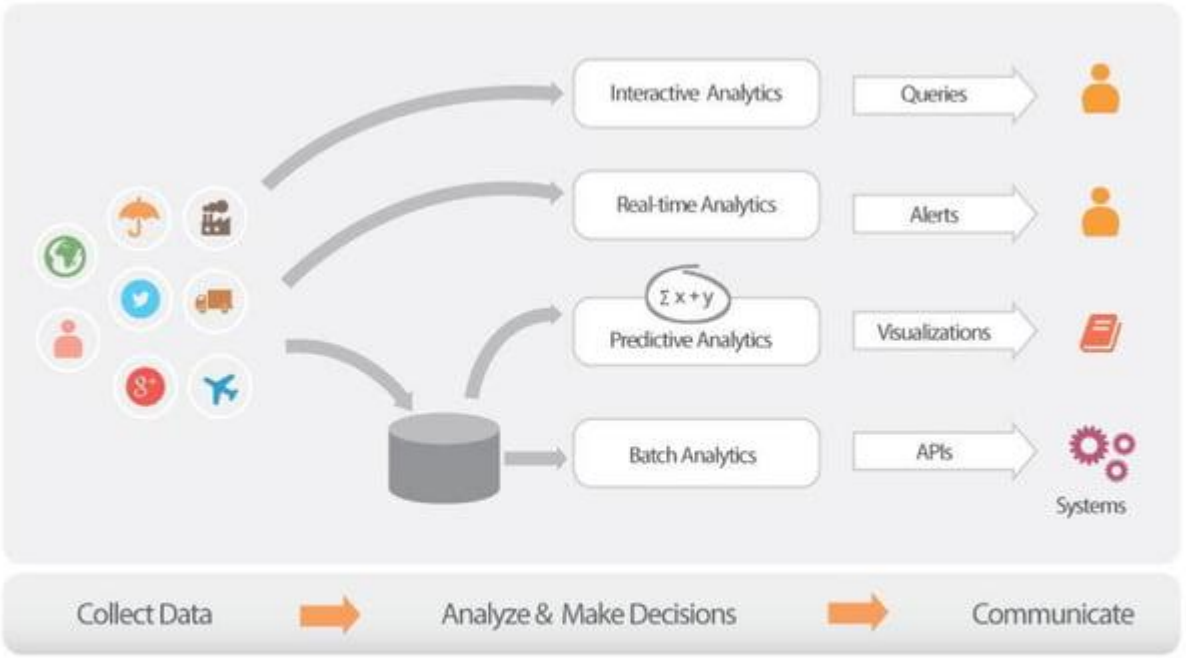
- Goal of analysis is to extract knowledge
- This knowledge usually come in one of the two forms
  - KPI (Key Performance Indicators)
    - Describe key measurement for what is being measured. (e.g. revenue per year, profit margin, revenue for sqft in retail, revenue per employer)
  - Models to describe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

4 Analysis types by time to decision

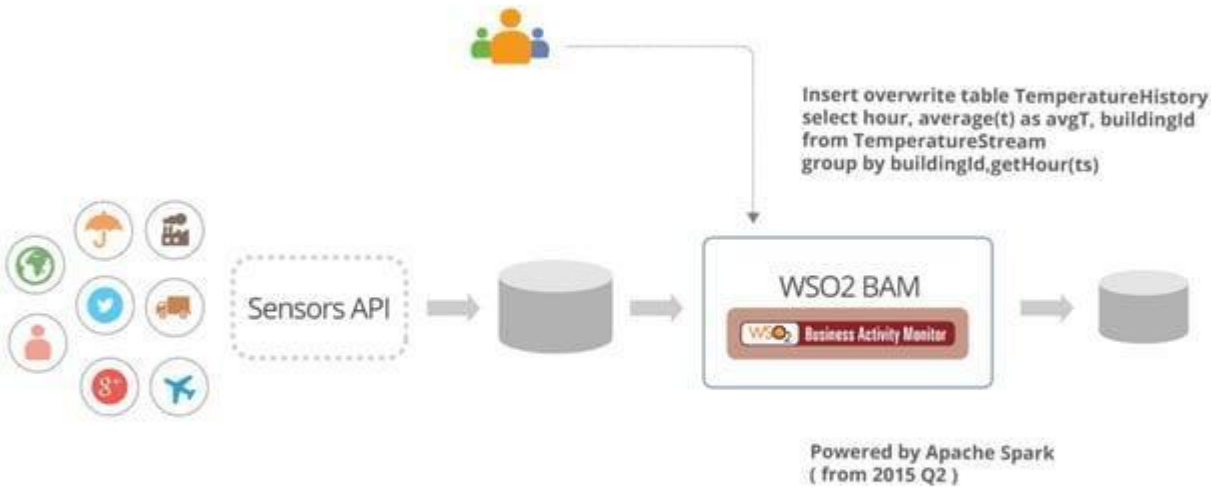
- Hindsight (what happened?)
  - Done using Batch Analytics like MapReduce
- Oversight (what is happening?)
  - Done using Realtime Analytics technologies like CEP
- Insight (why things happening?)
  - Done with Data Mining and Unsupervised learning algorithms like Clustering
- Foresight (what will happen?)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

Data Analytics Tools Landscape

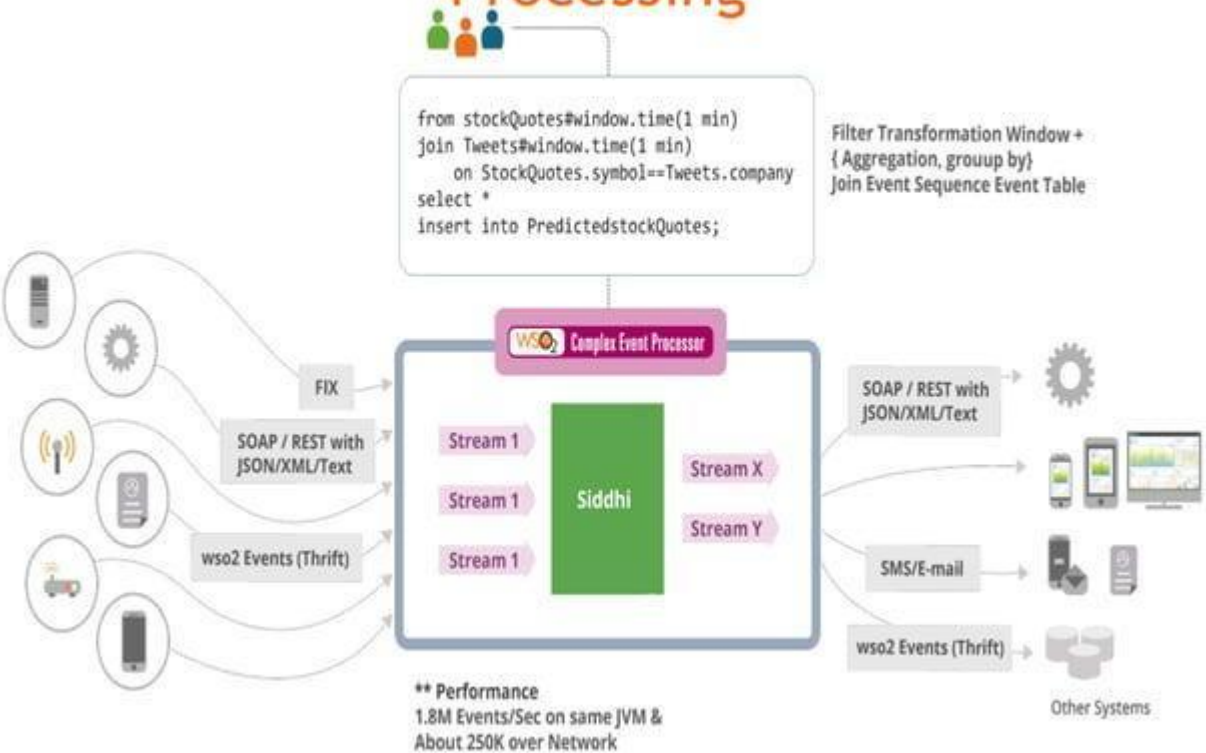



Interactive Analytics
Queries
Real-time Analytics
Alerts
Σ x+y
Predictive Analytics
Visualizations
Batch Analytics
APIs
Systems
Collect Data
Analyze & Make Decisions
Communicate

Batch Analytics: SparkSQL

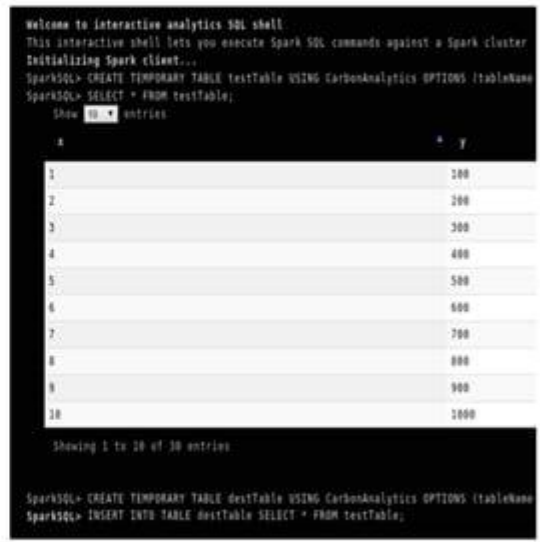


Realtime Analytics: Complex Event Processing



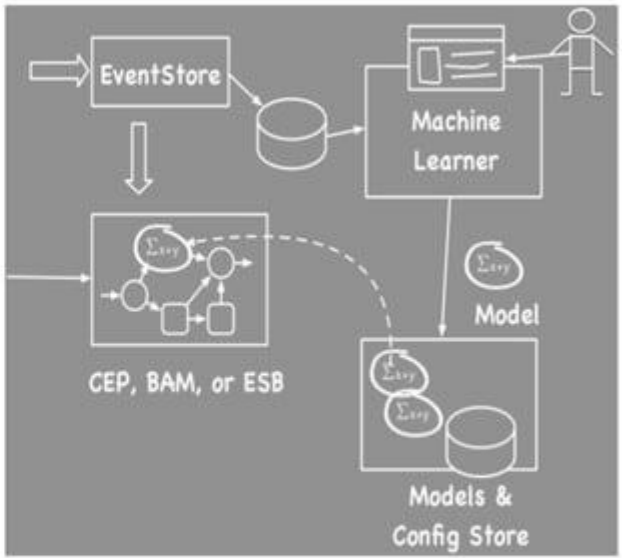
Interactive Analytics

- Define Indexes on Collected data (Streams)
- Issue, dynamic queries and get results right away. (Powered by Apache Lucene)
- Shows multiples events from same activity together using custom defined activity IDs
- Useful for data exploration
- Powered by Apache Lucene, with support for Index Sharding



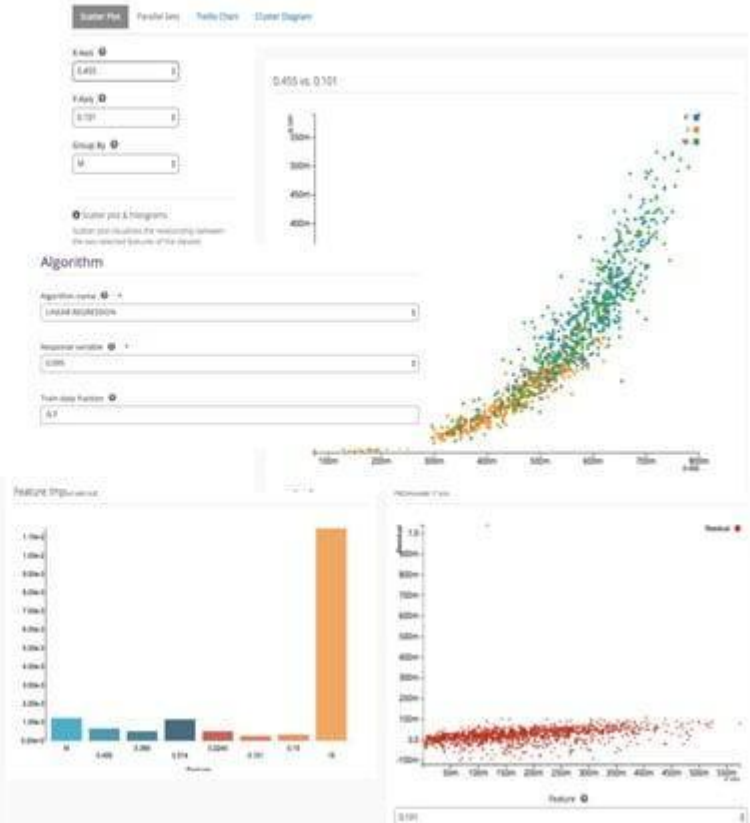
Predictive Analytics

- Build models and use them with WSO2 CEP, BAM and ESB using WSO2 Machine Learner Product (2015 Q3)
- Build model using R, export them as PMML, and use within WSO2 CEP



WSO2 Machine Learner

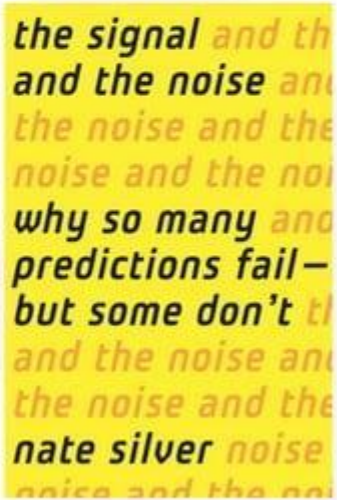
- Sample, explore, and understand data through visualizations
- A wizard to configure, train machine learning models, and select the best model
- Find and use those models with WSO2 CEP, BAM and ESB
- Powered by Apache Spark MLLib



Building Decision Models

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

- Regression models and ML Models Time series models
- Statistical models
- Physical Models - based on physical phenomena. They include 6-DoF flight models, space flight models Weather models.
- Mathematical Models



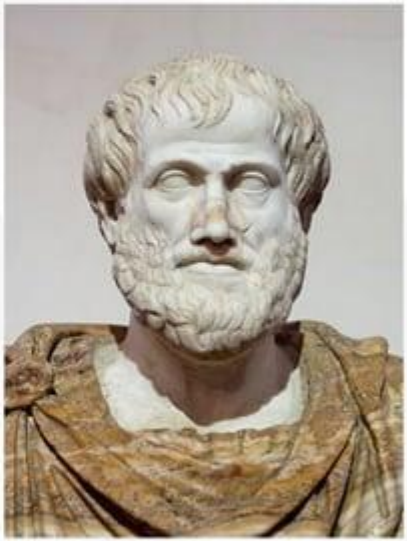
see https://icrunchdatanews.com/what-are-predictive-models/

Verification

- All is good, now you have a model. You must verify that it is correct before using it in the real world.
- Prediction can be verified by waiting for events to occur
- Relationships like causality (e.g. having free shipping leads a customer to buy more) must be verified with A/B testing
- Let's look at few of pitfalls

Pitfalls: Experiment vs Observation

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

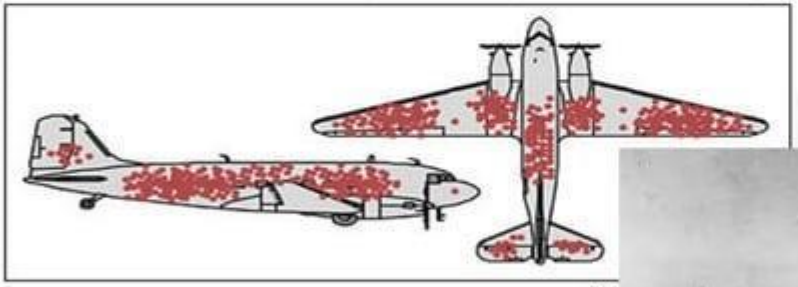


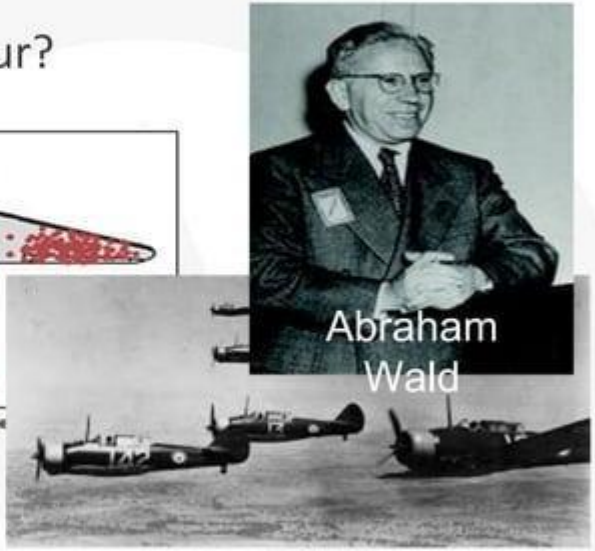
Causality: What can we do?

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

Pitfalls: Think about the Missing Data

- WW II, Returned Aircrafts and data on where they were hit?
- How would you add Armour?





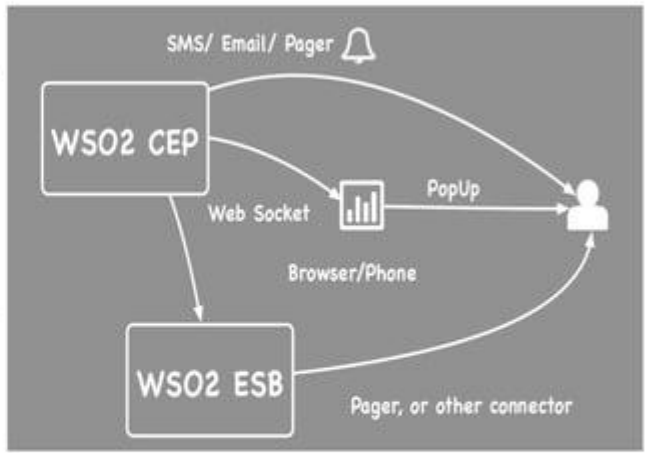
http://www.fastcodesign.com/1671172/how-a-story-from-world-war-ii-shapes-facebook-today, Pic from **http://www.phibetaiota.net/2011/09/defdog-the-importance-of-selection-bias-in-statistics/**

Communicate: Dashboards

- Dashboard give an "Overall idea" in a glance (e.g. car dashboard)
- Support for personalization, you can build your own dashboard.
- Also the entry point for Drill down
- How to build?
  - WSO2 DAS supports a gadget generation WIzard
  - Or you can write your own Gadgets using D3 and Javascript.

Communicate: Alerts

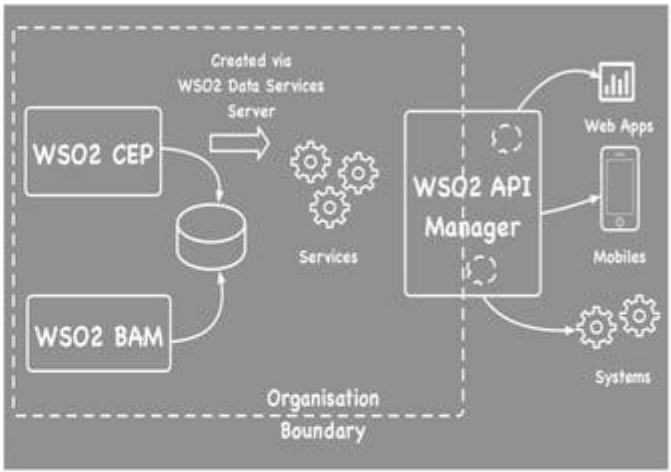
- Detecting conditions can be done via CEP Queries. Key is the "Last Mile".
  - Email
  - SMS
  - Push notifications to a UI
  - Pager
  - Trigger physical Alarm



- How?
  - Select Email sender "Output Adaptor" from CEP, or send from CEP to ESB, and ESB has lot of connectors

Communicate: APIs

- With mobile Apps, most data are exposed and shared as APIs (REST/Json) to end users.
- Need to expose analytics results as API
- Following are some challenges
  - Security and Permissions
  - API Discovery, Billing, throttling, quotas & SLA

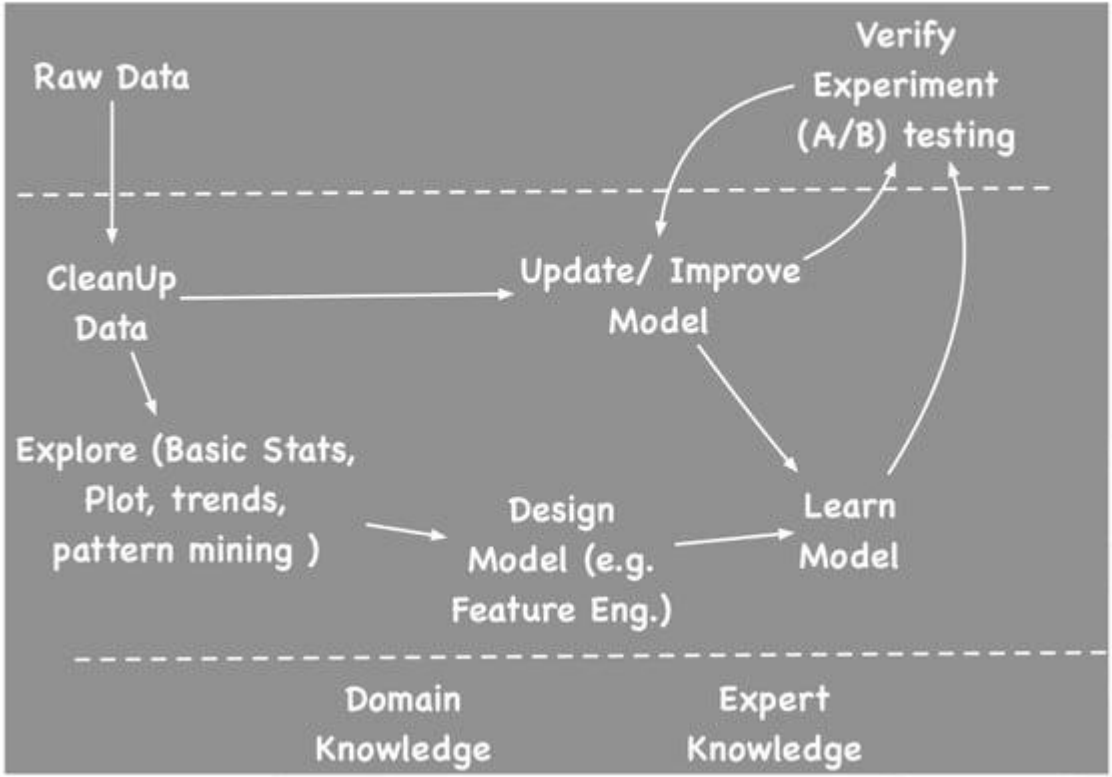


- How?
  - Write data to a database from CEP event tables
  - Build Services via WSO2 Data Service
  - Expose them as APIs via API Manager

Communicate: Realtime Soccer Analytics

https://www.youtube.com/watch?v=nRI6buQ0NOM

Data Science Pipeline



Conclusion

- Data Science is extracting knowledge by analyzing data
- Discussed the pipeline and tools you can use to do that
- Rest of summer school will look at different aspects in detail.
- All tools discussed are available free under Apache Licence.