


Sacramento AI  
Meetup

**BIG DATA + MACHINE LEARNING = AI**



**Bao Nguyen**  
July 2018

# CONTENT

- ▶ Introduction and objectives
  - ▶ Definitions of Big Data, Machine Learning = Artificial Intelligence (AI)
  - ▶ Artificial Intelligence applications
  - ▶ AI four major components
  - ▶ Introduction to each of AI's 4 major components
  - ▶ Conclusion
- 

# INTRODUCTION

- ▶ Big data is an obsolete term but it is AI
- ▶ Big data + Data science + ML + Tech = AI
- ▶ Intro to each area

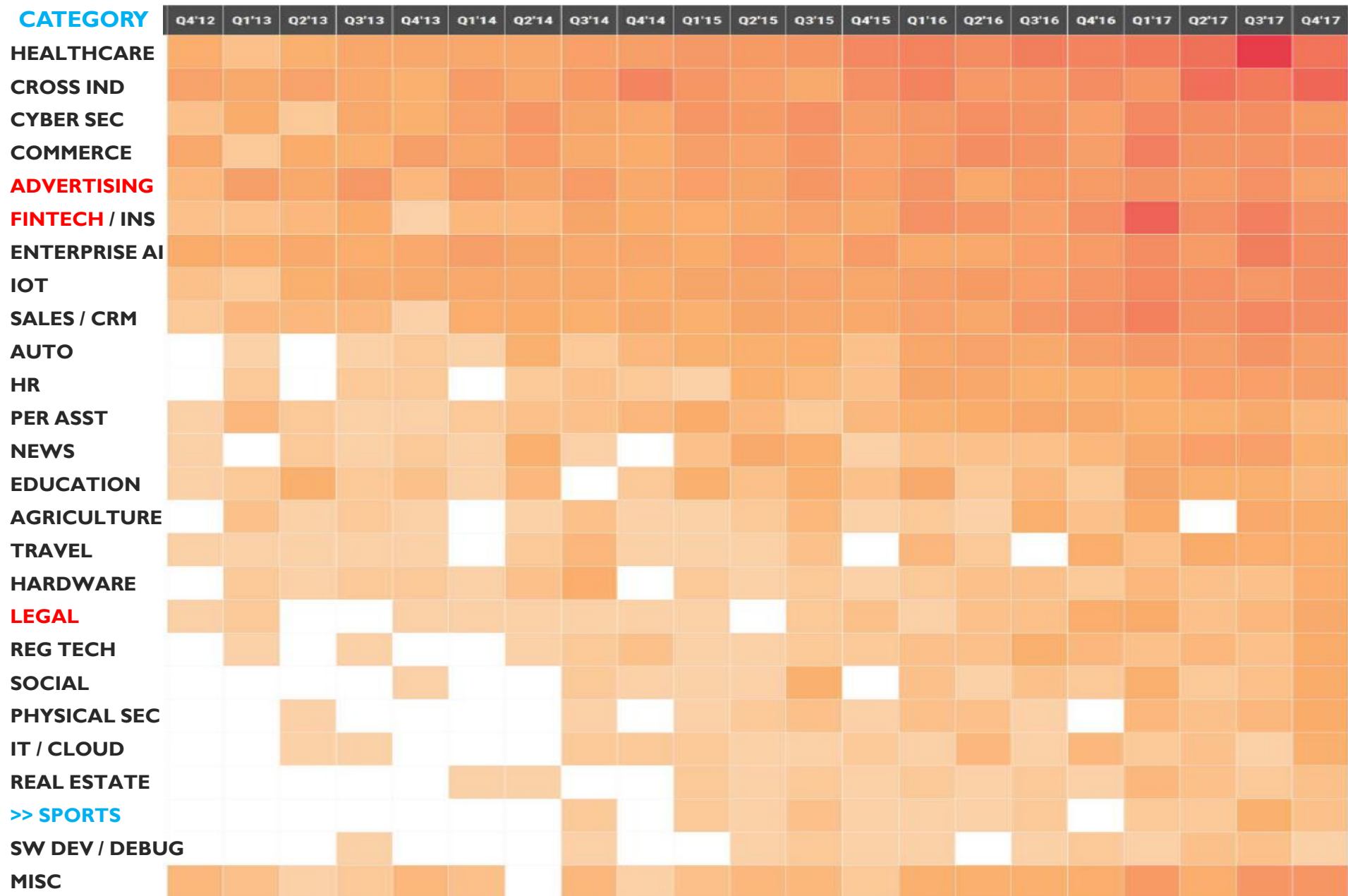
# INTRODUCTION

- ▶ Why another ML / AI talk?
  - ▶ Wanted to share lessons learned in hope to save other time and effort working with machine learning / AI
  - ▶ ML is an end to a mean
  - ▶ ML is commodity. Knowing end-to-end of AI (from goals to data to end application) is more important
  - ▶ Breaking down end-to-end AI with real world experience doesn't get much coverage
- ▶ Discussion approach - Interactive
  - ▶ Definitions of big data and machine learning = AI
  - ▶ Overview of high level end-to-end AI
  - ▶ Talk about AI in 4 areas
  - ▶ Share lessons learned in each of these 4 areas
  - ▶ Goal is to get awareness to main concepts of AI

# BIG DATA & MACHINE LEARNING, DEFINED

- ▶ Big data – from a science perspective
  - ▶ Using a lot of data to make highly accurate prediction
  - ▶ i.e. Learning insights from a lot of Data with Machine Learning
    - ▶ Big data is not ML. It is subset of big data
    - ▶ Data insights = Data management + Data Science
    - ▶ ML is algorithms to build models with learning inputs or seed data
    - ▶ Algorithms are methods / processes that are finite in time and functions
    - ▶ Models are mathematical representations of relationship between inputs and outputs derived via learning data
- ▶ Big data – from a technology perspective
  - ▶ Organizing and managing huge amount of data
    - ▶ Collecting (unstructured and structured) - Kafka
    - ▶ Managing strategy of large amount of data - Kafka
    - ▶ Normalizing and building feature sets - Hadoop
    - ▶ Analyzing strategy – example, offline and streaming (real-time)
    - ▶ Sharing and distributing – in terms of API, for example - DMP
    - ▶ Architecture and technologies enabling data organization and analysis - DC
- ▶ All above = AI

# AI IS EVERYWHERE – EQUITY DEALS BY 4Q I7



SOURCE: CB INSIGHTS

# BIG DATA & MACHINE LEARNING, DEFINED

- ▶ Big data
  - ▶ Term is obsolete
  - ▶ Other terms used often are Predictive Analytics, Advanced Analytics, Neural network, Deep learning etc.
  - ▶ It's about math: Statistics, Linear Algebra, Graph Theory, Calculus, Probability Theory... data insights and tech - to solve a particular problem
  - ▶ I refer to this field of predictive computational as AI
- ▶ Artificial Intelligence (AI)
  - ▶ AI involves huge technical fields & applications
  - ▶ AI uses input data, rather than commands, to build a model and then uses this model on target data to make predictions
  - ▶ More complex AI combines multiple distinct disciplines
    - ▶ Machines vs. Machines is based on this concept but much simpler
- ▶ Let's go into AI overview

# ARTIFICIAL INTELLIGENCE (AI)

- ▶ AI = Data + data science + Tech + ML + domain disciplines
- ▶ AI examples:
  - ▶ Data Mining = ML + databases (ML optimized in databases)
  - ▶ Speech Recognition = ML + Voice processing (dsp)
  - ▶ NLP = ML + Text processing
  - ▶ Machine Vision = ML + Image processing
  - ▶ Machine Recommendation = ML + Behavioral processing



# 4 COMPONENTS IN ARTIFICIAL INTELLIGENCE

- ▶ Data collection / onboarding, management and storage
- ▶ Data research tools – understanding existing data and insights
- ▶ Machine learning - algorithms
- ▶ Distributed technology processing framework
  - ▶ Distributed frameworks – AWS, GCP, Azure etc.
  - ▶ Reporting
  - ▶ Data visualization
  - ▶ Data compatibility layer
  - ▶ Processed data layers sharing APIs

# CROSS INDUSTRY STANDARD FOR ML

## The six phases of CRISP-DM

### 1. BUSINESS UNDERSTANDING

In the first stage of CRISP-DM, you work through what the project looks like, and what the business expectations for the project are.

### 2. DATA UNDERSTANDING

Next, the data that are available for analysis are examined in light of the business objects that were decided on in the first phase.

### 3. DATA PREPARATION

Once you have an understanding of the data that's available, the next step is to clean, sort, and process said data to make it useable for your purposes. This phase often takes the greatest amount of time and effort.

### 4. MODELING

Then, you iterate through various versions of a model, using the prepared data from phase 3.

### 6. DEPLOYMENT

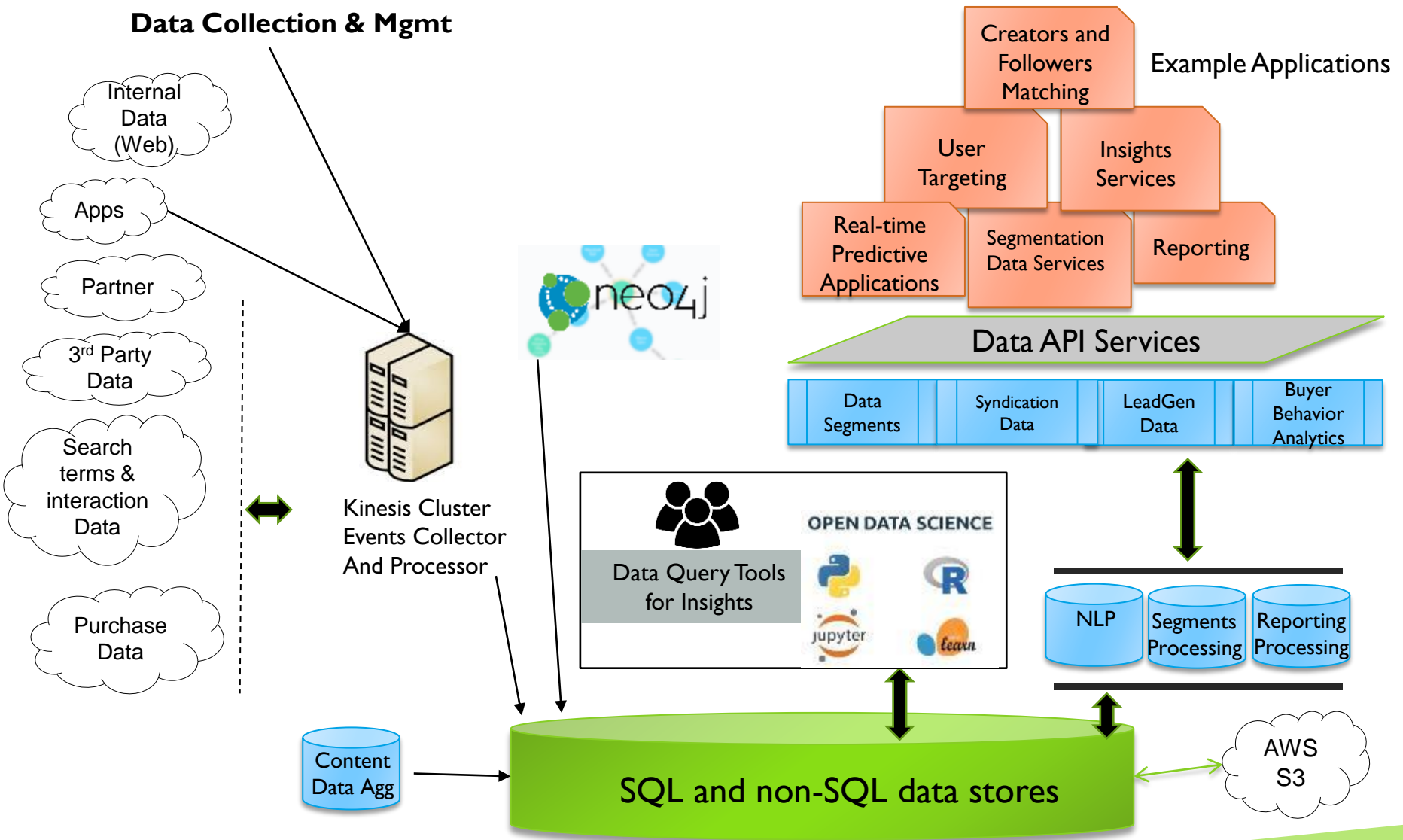
Finally, you need to deploy the model that you've developed to ensure that it can have a positive impact on your business. It might seem like a no-brainer to deploy a model once it's created, but fully half of completed models never make it into production, contributing significantly to the [Model Impact Epidemic](#).

### 5. EVALUATION

Once you're happy with the model that you've built, you need to evaluate whether or not it effectively addresses the business criteria laid out in the first phase.



# EXAMPLE AI SYSTEM VIEW

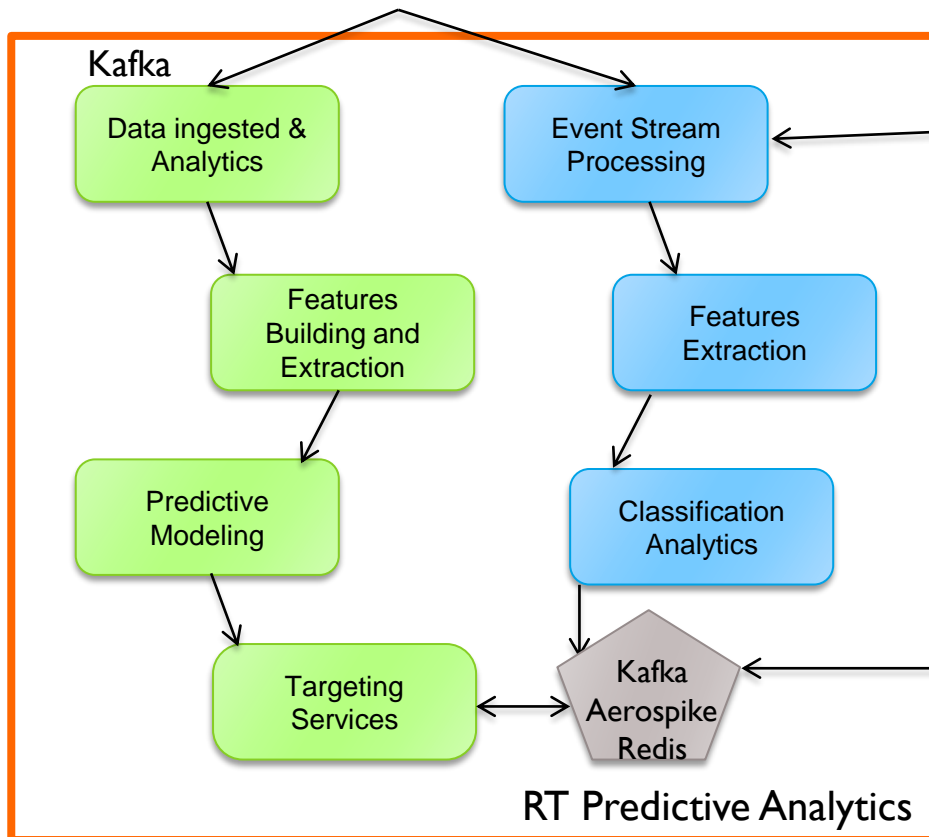


# REAL-TIME PREDICTIVE ANALYTICS FLOW

Data Sources




Applications



# AI USE CASES – ONLINE USER TARGETING

- ▶ Collects user's data from: Content websites, Mobile, CRM, Online videos, internal systems and third party data sources
  - ▶ Relevant data sources are universally tagged & synced. In case of 3<sup>rd</sup> party, integrate tagging methods were deployed
  - ▶ Examples of data events collected: leadbacks, retargeting, clicks, downloads, video view, how long video view, website information, engagement time, TOD, geography, long/lat info, purchases, purchase patterns, search keywords, URL ...
- ▶ Profiles and segments data with data management technologies
  - ▶ Processes billions of events per day
  - ▶ Data is organized, centralized and synced from disconnected sources to gain maximum knowledge of the same user – advanced matching technology is involved and this is a competitive advantage
  - ▶ Data is processed and stored per user basis for targeting
- ▶ Revenue prediction

# CUSTOMER INSIGHTS MONETIZATION

- ▶ CRM monetization provides insights on which demographic, geographic, technographics, and behavioral user segments view, click, and act on a certain event, product or website
  - ▶ Using this info, recommendations for sales, up-sell, cross-sell, and data sell on both existing and new users
- 

# REAL-TIME OPTIMIZED ENTERPRISE AI

- ▶ Consumes data from Targeting Services (offline analytics)
  - ▶ Random Forest ensemble learning (multiple algos of both classifier as well as mod linear regression analysis), Naïve Bayes classifiers
- ▶ Real-time streaming analytics component was added to process online streaming live events (Kafka + Spark)
- ▶ Identifies and updates user's behavior characteristics and scores in real-time – Spark MLlib
- ▶ In addition clients often add private data, generating even more value
- ▶ Predictive behavior – user's probability to take certain action are predicted – users are seek out

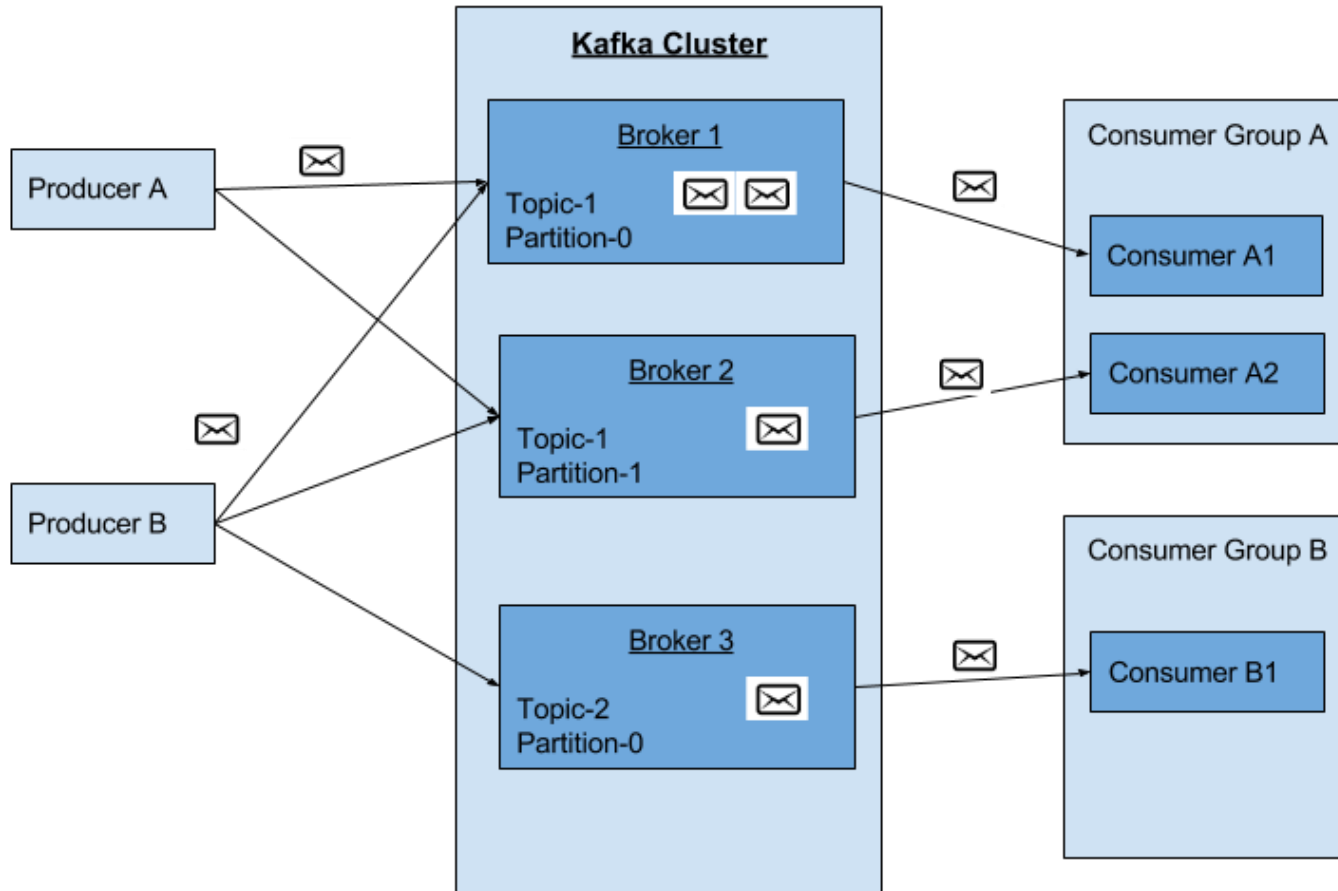
# AI 4 COMPONENTS OVERVIEW



# I. DATA COLLECTION & MANAGEMENT

- ▶ Data challenges
  - ▶ Volume – massive amount of data
  - ▶ Variety – complexity and many types: IoT, apps...
  - ▶ Velocity – fast and need to process quickly or data is gone (stream analytics)
  - ▶ Veracity – unstructured, uncertain and inaccurate data
- ▶ The key is to extract insights from data in real time to make proactive decisions – Focus: right data, right and fast insights, and monitored data sources – Data Science
- ▶ A tech framework
  - ▶ Kafka – distributed sub and pub messaging system. Hadoop compatible and allows subscription to the right data
  - ▶ Spark in memory data analysis – real time analytics integrated with Kafka
  - ▶ If using Flume, (scalable data collection) it has source and sink for Kafka
  - ▶ If using Storm (stream processing), Kafka topics are compatible with Storm and Storm topology can emit enriched events to Kafka topic

# KAFKA SUB / PUB MESSAGING SYSTEM



# DATA QUALITY UPFRONT IS KEY – LESSON LEARNED

- ▶ Garbage-in garbage-out – many broken or partially broken data collectors
- ▶ Invest in collecting quality & focus on having stable data sources
- ▶ Operationally **monitor** & manage data sources – this a common and costly overlook
- ▶ Do not use data without knowing the data!!
- ▶ Know the data and clean the data
- ▶ More data doesn't mean better; **relevancy** is key
- ▶ Does data latency matter? Need to understand that!
- ▶ **Automate** as much as possible!
- ▶ **Huge costs and time later if the above are not done**

# DATA SYNCING – IMPORTANT & LESSON LEARNED

- ▶ Syncing data sources from all the places/system a user has been is powerful – typically on a non-sign-in systems
  - ▶ Cookies IDs sync
  - ▶ Smart pixel – analytics cookie that can scrape extra info
  - ▶ Pixel to pixel sync
  - ▶ Javascript container tag / universal container
  - ▶ Server-server sync integration
  - ▶ IP sync
  - ▶ Statistical “finger” printing method syncing
- ▶ If there is a sign-in, users are definitively ID’
  - ▶ Free apps are not really free as users are know via sign-in
  - ▶ Syncing is much less of an issue if app is good (user retention)

## 2. DATA INSIGHTS FRAMEWORK

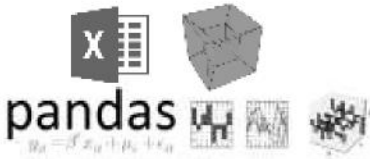
- ▶ Challenges
  - ▶ Without the correct data insights there could be
    - ▶ Reduced prediction accuracy due to inaccurate seed / training data
    - ▶ Expensive processing due to less optimized feature sets
- ▶ The key is the right insights tool framework to reduce programming requirements so focus can be on... data insights
- ▶ *This is one the most important area and get it right – lesson learned*
  - ▶ Affects the prediction outcome
  - ▶ Affects the scale of implementation
  - ▶ Affects computational expenses
  - ▶ Affects costs and timeline
- ▶ A solution
  - ▶ Anaconda framework – A Python distribution and framework for hundreds of scientific analytics tools such as R (stat computing and visual)
  - ▶ Reduced programming requirement with Python and R for data insights

# ANACONDA – DATA SCIENCE PLATFORM

## Statistics and ML



## Business Intelligence



## Computational Science



## Web



## Distribution Tech

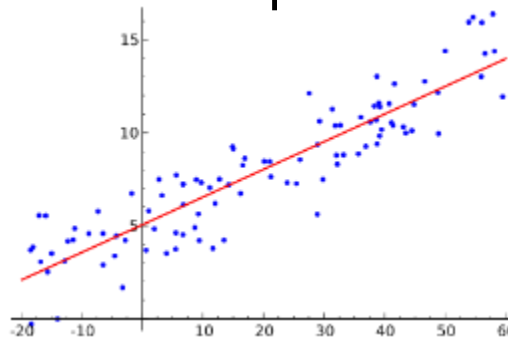


# 3. MACHINE LEARNING

- ▶ ML uses insightful data to build models and uses models against large amount of data to form predictions
- ▶ ML is used when the algorithms can't be coded or when huge amount of info needed to be processed
- ▶ ML process – lots of details but not complicated
  - ▶ Have objectives with targets to build models
  - ▶ Collect, understand and analyze data to build insights (independent variables), feeding ML algos. Visualization is very helpful when validating data. Combine inputs to build Feature Sets to get better prediction
  - ▶ Feed the machine, get the models and evaluate the models. The model capture the relationship between the independent variable(s) and dependent variable
  - ▶ If the model is good, start the prediction machine

# COMMON ALGORITHM TYPES


- ▶ Classification – Organize records by group types
  - ▶ Binary classification: simple and effective for 0 / 1 class
- ▶ Clustering – Organize records based on similarity
- ▶ Association learning – Learn what often appears with what
- ▶ Recommendation – Look alike
- ▶ Regression - Prediction
  - ▶ Linear regression – numerical inputs with numerical outputs (least square)




- ▶ Logistics regression – often use in binary classification. Alternatively, Support Vector Machine (SVM) with better optimization draws better classification boundary than logistics alg
- ▶ Note: Square method is computational expensive so leverage Stochastic Gradient Descent (SGD) sequential passes for efficiency



# MACHINE LEARNING ALGORITHMS

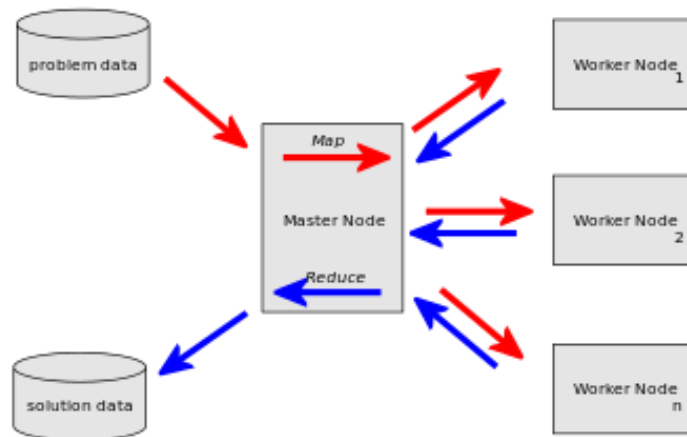
- ▶ Implementation wise, leverage opensource Apache Mahout ML library
  - ▶ No need to get too complex
  - ▶ Naïve Bayes – highly scalable and simple probabilistic classifier
  - ▶ Random Forest – ensemble learning method for prediction (regression with decision trees) and better classification than NB via mode of classes
  - ▶ Enhance models with Stochastic Gradient Decennt (SGD)
  - ▶ Finally, AWS made a commodity out of ML
- 

# OTHER OPEN SOURCE MACHINE LEARNING OPTIONS

- ▶ Amazon – Deep Scalable Sparse Tensor (DSSTNE)
  - ▶ Google – TensorFlow: ML library
  - ▶ IBM – SystemML: ML platform
  - ▶ Microsoft – Distributed Machine Learning Toolkit (DMTK)
  - ▶ Facebook – Torch: ML deep learning algorithms
  - ▶ Yahoo – CaffeOnSpark: ML library
- 

# 4. DISTRIBUTED PROCESSING FRAMEWORK

- ▶ Apache Hadoop is a common distributed processing to operate on large data set with rich ecosystem supporting AI
  - ▶ Mahout - ML library
  - ▶ Cassandra – Large scale DB with no single point of failure
  - ▶ Hbase – Highly scalable distributed DB on top of HDFS file system
  - ▶ Hive – Ad hoc querying (I prefer Anaconda framework)
  - ▶ Spark – High performance computing in ETL, ML, stream processing
  - ▶ Programming is MapReduce (similar to LISP functional programming) – massive parallel processing



# AI TECHNOLOGY STACK CONSIDERATION

Function	Technologies
Platform	Apache Hadoop, .NET, Java
Data ETL	Kafka
Data Storage & Management	MapReduce, Hbase, Zookeeper, Cassandra, Couchbase
Real-time processing	Spark ML, Redis
Research	Impala, Spark ML, Octave, R, Python
Algorithms, Machine Learning, Real-time analytics	MapReduce, Mahout, Bayes, Random Forest, Spark ML
Data Warehousing, Query	AWS, Hive
Metadata, table management	MySQL
CI / CD Development Methodology	Jenkins, Docker, Ansible, Rundeck
Cluster Operational Management	Yarn
Data Serialization	Avro
Reporting	Vertica? Existing database

# ML USE CASE: RETAIL DEMAND FORECASTING

- ▶ Inventory prediction for retail during a finite time period
  - ▶ As highlighted earlier, need ML to:
    - ▶ Fast processing of vast amount of data
    - ▶ Leverage well researched accurate forecasting algorithms
    - ▶ Automation of modeling update
    - ▶ Robust to changes in ecosystem or essentially data
  - ▶ Goals
    - ▶ Optimal supplier management – Having the right # of suppliers
    - ▶ Improve customer satisfaction – Having products available
    - ▶ Efficient logistics
    - ▶ More real-time inventory management for events such as sales and promotions
    - ▶ Cost reduction via inventory management

# RETAIL USE CASE STEPS

- ▶ Initial Data Analysis
  - ▶ Get all relevant available data
    - ▶ Review for quality: volume, consistency, accuracy
    - ▶ Do quick insight analysis for data relevancy to understand data bearing
  - ▶ Set objectives so outcome metrics can be developed and measured
    - ▶ Ex 1: I want to know how many pair of shoes needed for the next 5 weeks
    - ▶ Ex 2: I'm going to have a Christmas sales. How many \$1 red candles I need
    - ▶ Goal should be items and time bound
    - ▶ Can be short-term or long-term - will have different purposes and data requirement
    - ▶ Optimal supplier management – Having the right # of suppliers
    - ▶ Improve customer satisfaction – Expectation management
    - ▶ Efficient logistics
    - ▶ More real-time inventory management for events such as sales and promotions
    - ▶ Cost reduction via inventory management

# RETAIL USE CASE STEPS

- ▶ Objective performance

- ▶ Accuracy

- ▶ Need absolutely “right” data and “enough” data
    - ▶ Average prediction accuracy should be in 92% - 95% range

- ▶ Metrics

- ▶ Mean Absolute % Error – Statistical method to measure prediction system


$$M = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$

- ▶  $A_t$  = Actual value;  $F_t$  = Forecast
    - ▶ Mean Absolute Error – Measuring Diff of two continuous variables X,Y

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

- ▶ Root Mean Square -

# DATA PREP

- ▶ Data Consistency
  - ▶ Data Accuracy
  - ▶ Data Validation
  - ▶ Data Relevancy
  - ▶ Data Availability
  - ▶ Data Completeness
  - ▶ Data Dictionary or Understanding
- 



# CONCLUSION

- ▶ AI is a huge field in both technologies and applications
- ▶ Understanding AI end-to-end (data + insights + ML + tech) will help to see the big picture and provide value-add
- ▶ Need many more meetups to dive deeper into each AI 4 areas
- ▶ Q&A

Thank You!

<https://www.linkedin.com/in/baonguyen10>

nguyen.bao@gmail.com

# BAO NGUYEN INTRO

- ▶ Arch & Co-founder of Machines vs. Machines algorithmic based automated trading system – Business to business
- ▶ 10+ years predictive analytics & ML with Aol R&D and LexisNexis, MvM
  - ▶ 5.5 years as Aol's Vice President of R&D in Behavioral Advertising Optimization and Predictive Analytics
  - ▶ 4.5 years as LexisNexis' Vice President of Engineering and Innovation – SaaS / IaaS, NPL and SEO
- ▶ 10+ years as developer and architect
- ▶ Started two start-ups and been with large corp
- ▶ Degrees in Math, BSEE, MBA
- ▶ Education: University of Minnesota, University of St. Thomas, Stanford, Babson College
- ▶ <https://www.linkedin.com/in/baonguyen10>
- ▶ Website: <https://bhbt.info/>