Strata+ Hadoop WORLD

PRESENTED BY

O'REILLY cloudera

Text Analytics

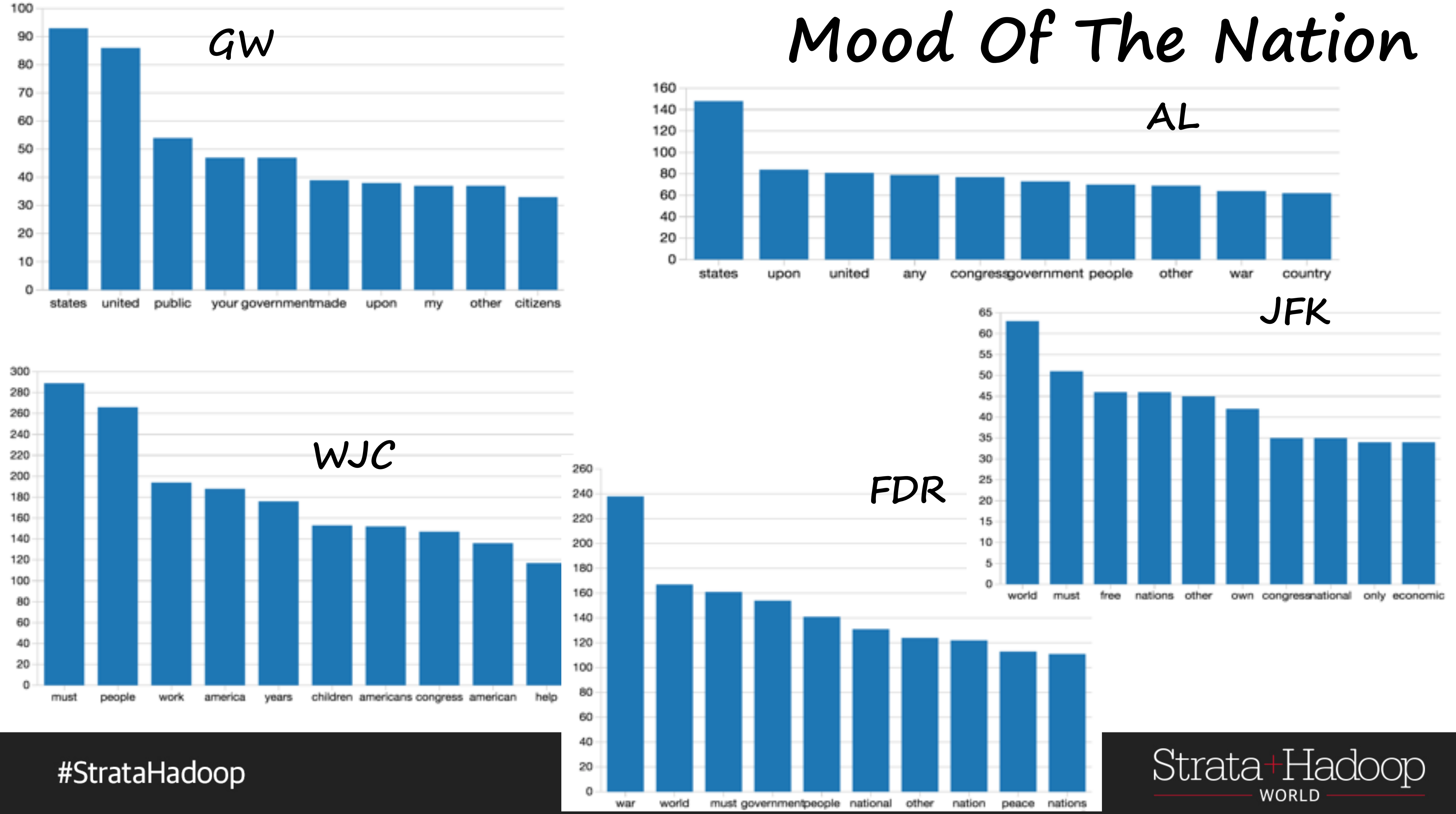
In which we analyze the mood of the nation & then the 2016 US Presidential primary debates

(Don't worry, we will keep it civilized)

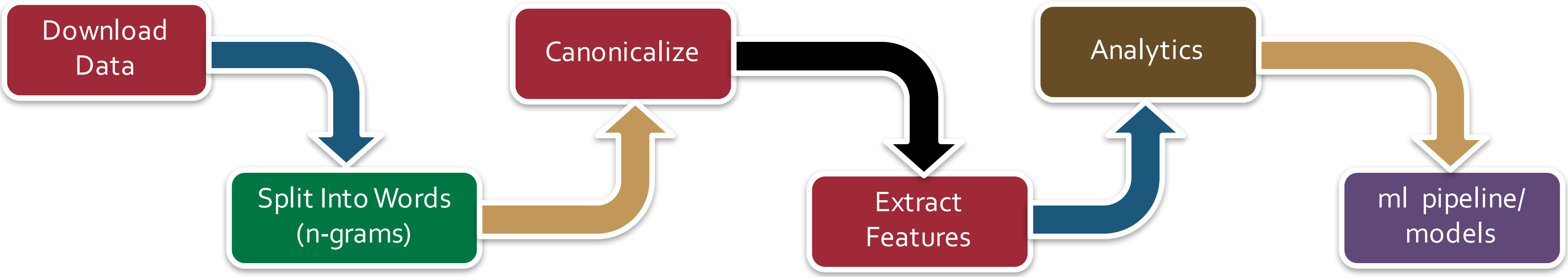
Krishna Sankar
https://www.linkedin.com/in/ksankar
March 29, 2016

strataconf.com
#StrataHadoop

Mood Of The Nation



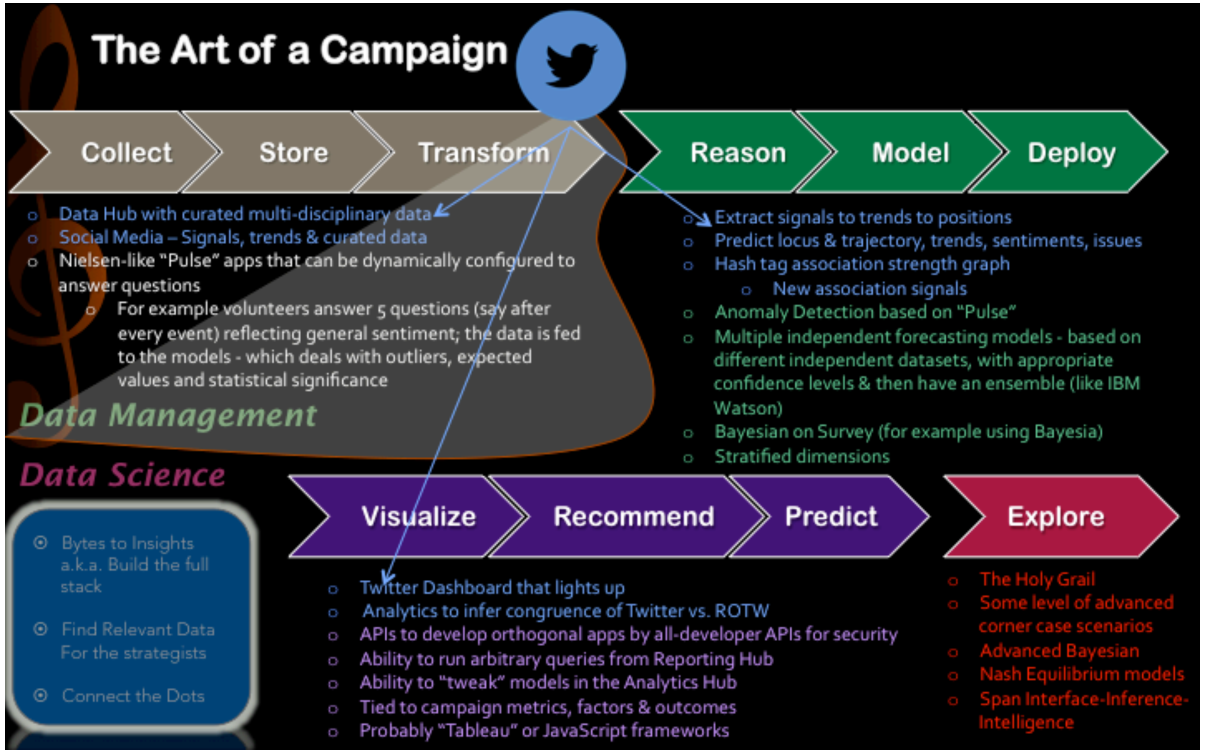
Text Analytics Pipeline



○ http://stateoftheunion.onetwothree.net/texts/index.html	○ Spark RDD Mechanisms	○ Case transformation, special characters, space elimination,…	○ Common word elimination	○ Exploration	○ Logistic Regression ○ Bayesian Models ○ LDA
○ http://www.presidency.ucsb.edu/debates.php	○ Dataframes (especially for ml pipelines	○ Stemming, Lemma	○ TF-IDF	○ Knowledge Representation (Knowledge Graph)	○ Deep Learning ○ (Topics, Classification,…)
		○ Domain Scoping		*word2vec*	

- *Understand the tutorial in the context of a broader reference architecture (e.g. next 2 slides), Select components as needed by the app*

Reference Architectures for Text Analytics



The Art of a Campaign

Collect → Store → Transform → Reason → Model → Deploy

- Data Hub with curated multi-disciplinary data
- Social Media – Signals, trends & curated data
- Nielsen-like "Pulse" apps that can be dynamically configured to answer questions
  - For example volunteers answer 5 questions (say after every event) reflecting general sentiment; the data is fed to the models - which deals with outliers, expected values and statistical significance

Data Management

- Extract signals to trends to positions
- Predict locus & trajectory, trends, sentiments, issues
- Hash tag association strength graph
  - New association signals
- Anomaly Detection based on "Pulse"
- Multiple independent forecasting models - based on different independent datasets, with appropriate confidence levels & then have an ensemble (like IBM Watson)
- Bayesian on Survey (for example using Bayesia)
- Stratified dimensions

Data Science

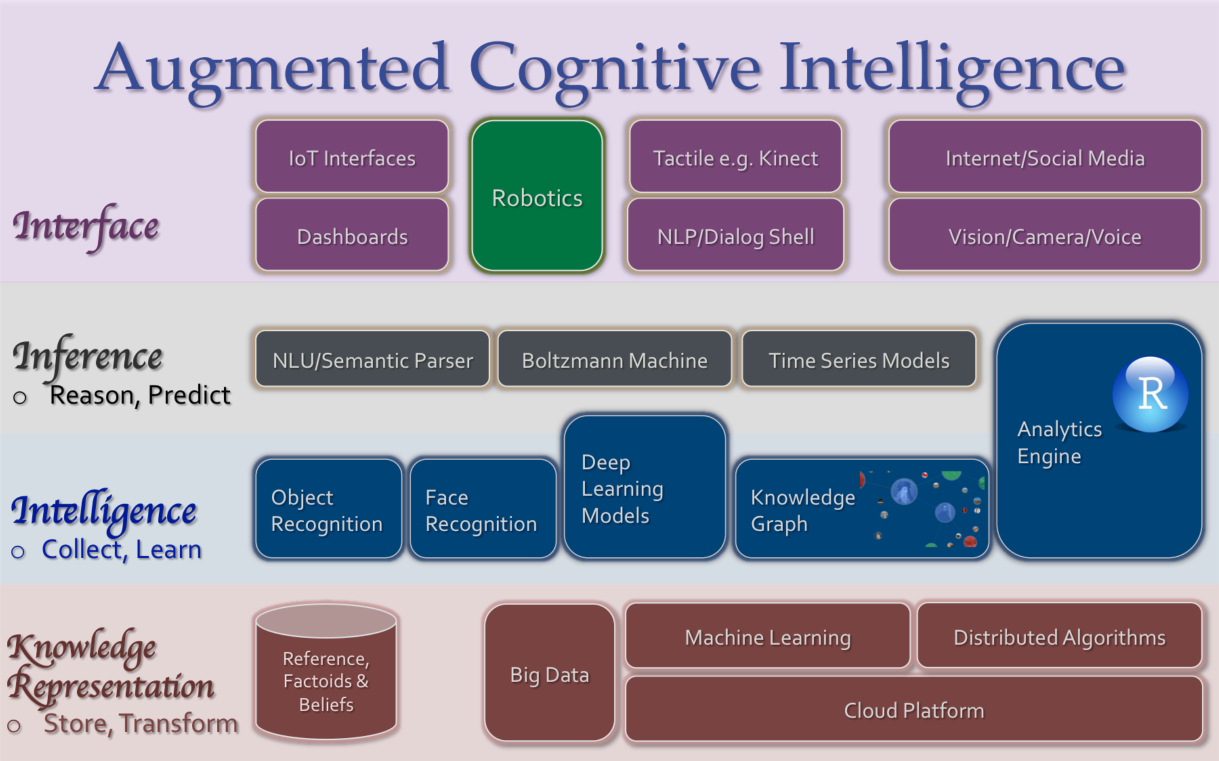
- Bytes to Insights a.k.a. Build the full stack
- Find Relevant Data For the strategists
- Connect the Dots

Visualize → Recommend → Predict → Explore

- Twitter Dashboard that lights up
- Analytics to infer congruence of Twitter vs. ROTW
- APIs to develop orthogonal apps by all-developer APIs for security
- Ability to run arbitrary queries from Reporting Hub
- Ability to "tweak" models in the Analytics Hub
- Tied to campaign metrics, factors & outcomes
- Probably "Tableau" or JavaScript frameworks

- The Holy Grail
- Some level of advanced corner case scenarios
- Advanced Bayesian
- Nash Equilibrium models
- Span Interface-Inference-Intelligence

https://doubleclix.wordpress.com/2015/07/05/twitter-2-0-curated-signals-applied-intelligence-stratified-inference/



Augmented Cognitive Intelligence

Interface

IoT Interfaces | Robotics | Tactile e.g. Kinect | Internet/Social Media
Dashboards | NLP/Dialog Shell | Vision/Camera/Voice

Inference
- Reason, Predict

NLU/Semantic Parser | Boltzmann Machine | Time Series Models | Analytics Engine

Intelligence
- Collect, Learn

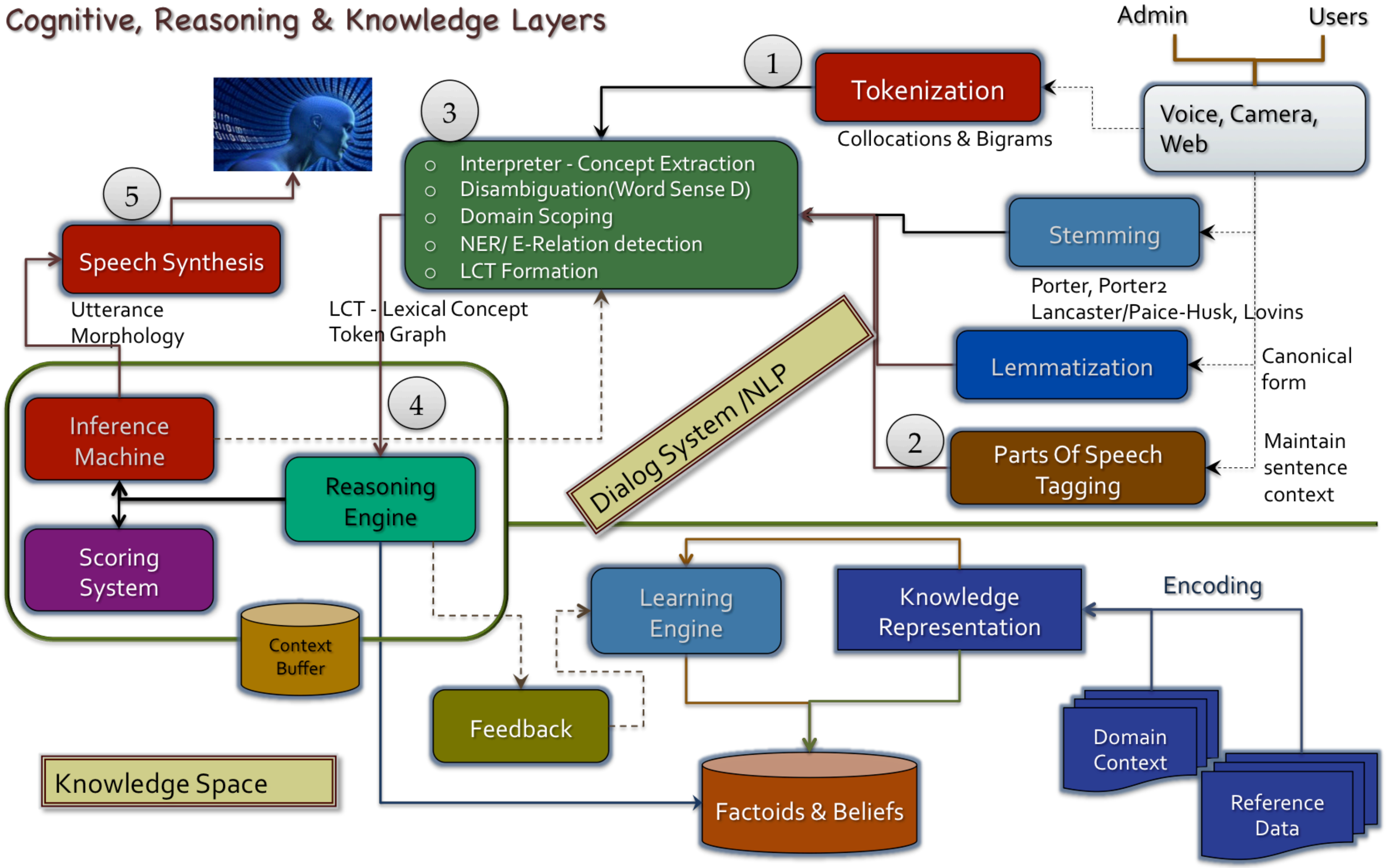
Object Recognition | Face Recognition | Deep Learning Models | Knowledge Graph

Knowledge Representation
- Store, Transform

Reference, Factoids & Beliefs | Big Data | Machine Learning | Distributed Algorithms | Cloud Platform

Reference Architectures for Text Analytics

Cognitive, Reasoning & Knowledge Layers



Admin

Users

Voice, Camera, Web

1 Tokenization

Collocations & Bigrams

3
- Interpreter - Concept Extraction
- Disambiguation(Word Sense D)
- Domain Scoping
- NER/ E-Relation detection
- LCT Formation

LCT - Lexical Concept Token Graph

Stemming

Porter, Porter2
Lancaster/Paice-Husk, Lovins

Lemmatization

Canonical form

2 Parts Of Speech Tagging

Maintain sentence context

5 Speech Synthesis

Utterance Morphology

Dialog System /NLP

4

Inference Machine

Reasoning Engine

Scoring System

Context Buffer

Learning Engine

Knowledge Representation

Encoding

Domain Context

Reference Data

Feedback

Factoids & Beliefs

Knowledge Space

Text Analytics

- A Data Scientist can use RDD primitives to do interesting work with text
  - Map-reduce in a couple of lines !
    - But it is not exactly the same as Hadoop Mapreduce (see the excellent blog by Sean Owen[1])
  - Set differences using substractByKey
  - Ability to sort a map by values (or any arbitrary function, for that matter)
  - Dataframe operations
- TF-IDF as Feature Extraction http://spark.apache.org/docs/latest/mllib-feature-extraction.html#tf-idf
- LDA for topic extraction/ml pipeline in next section
- *Good & Bad – mllib features span a continuum of libraries from rdds to dataframes to datasets to extraction to ml models to ml pipelines and beyond …*

[1] http://blog.cloudera.com/blog/2014/09/how-to-translate-from-mapreduce-to-apache-spark/

Code Walkthru

- 03-Text Analytics Notebook