

Powering NLU Engine with Apache Spark to Communicate with World

Rahul Kumar, I.AM Plus Electronics

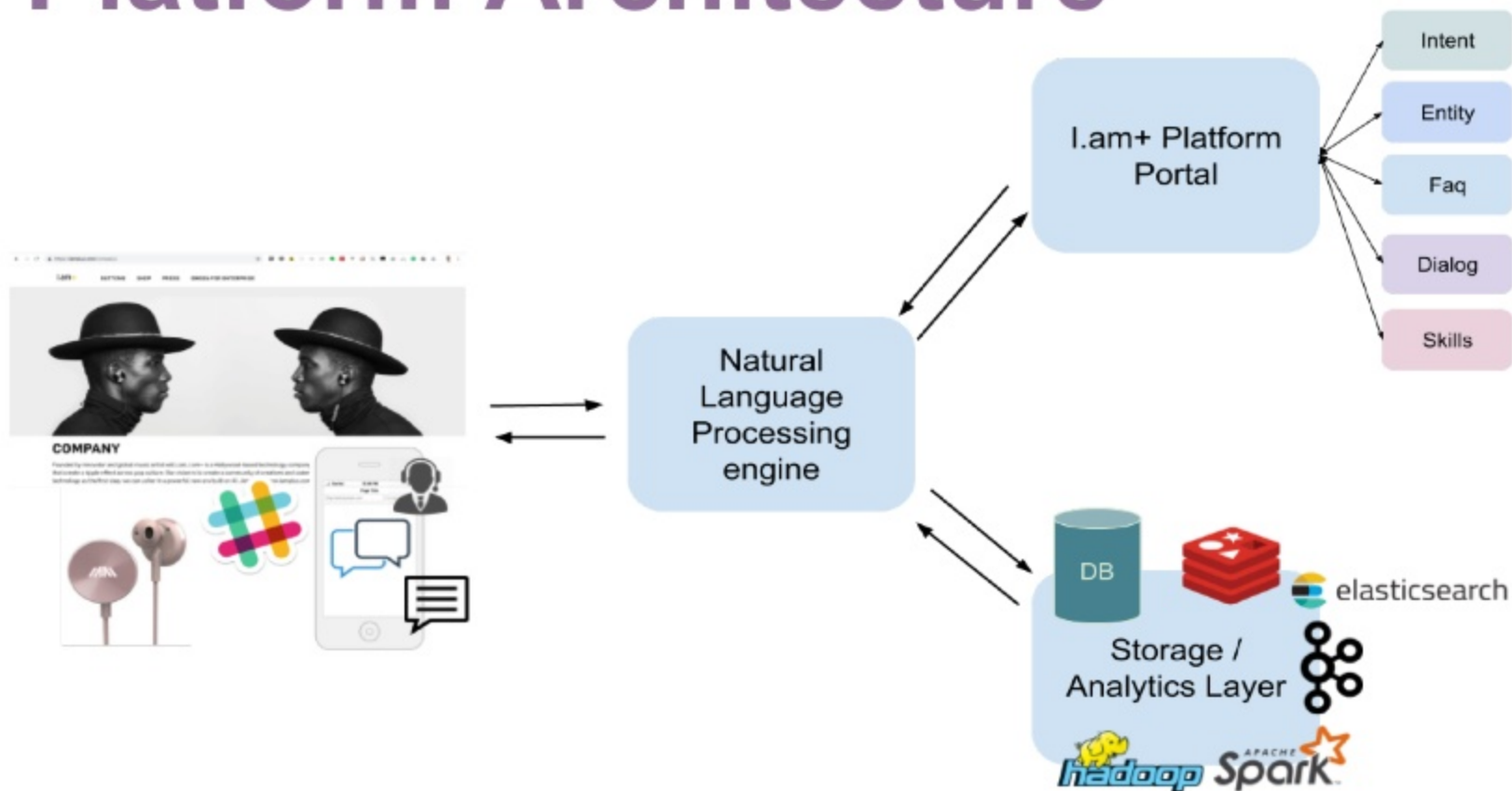
Agenda

- NLP platform features
- NLP Architecture
- NLP building blocks
- Deep learning for NLP
- TensorFlow Overview

NLP Platform Features

- Text & Voice based interaction
- 50K + Music
- Email / Calendar Integration
- News + weather
- Open API for various Skills

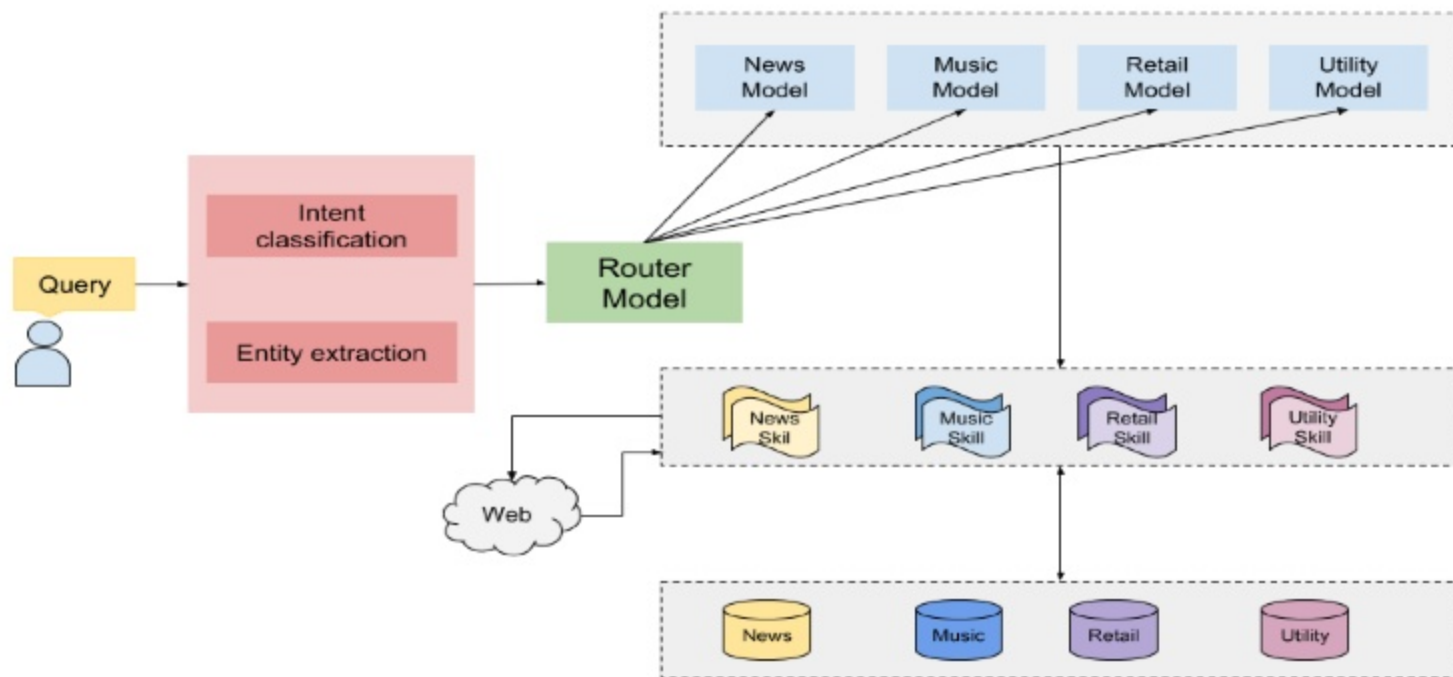
Platform Architecture

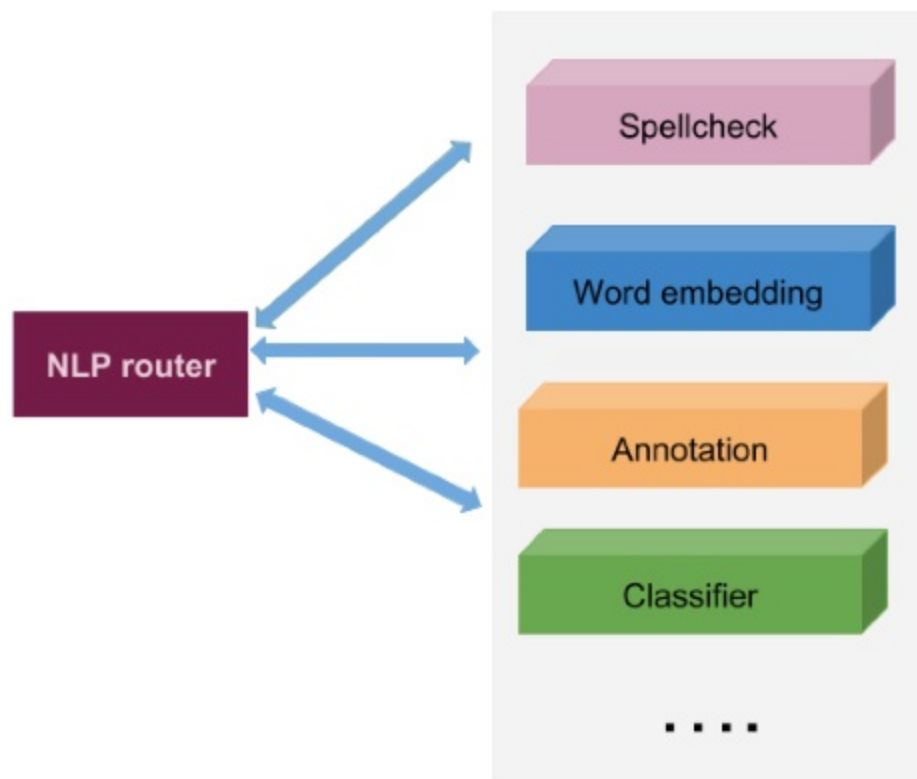


Core Building Blocks

- NLP/NLU engine
- Skills repository
- Data Processing & Analyzing
- User Interaction medium

Workflow





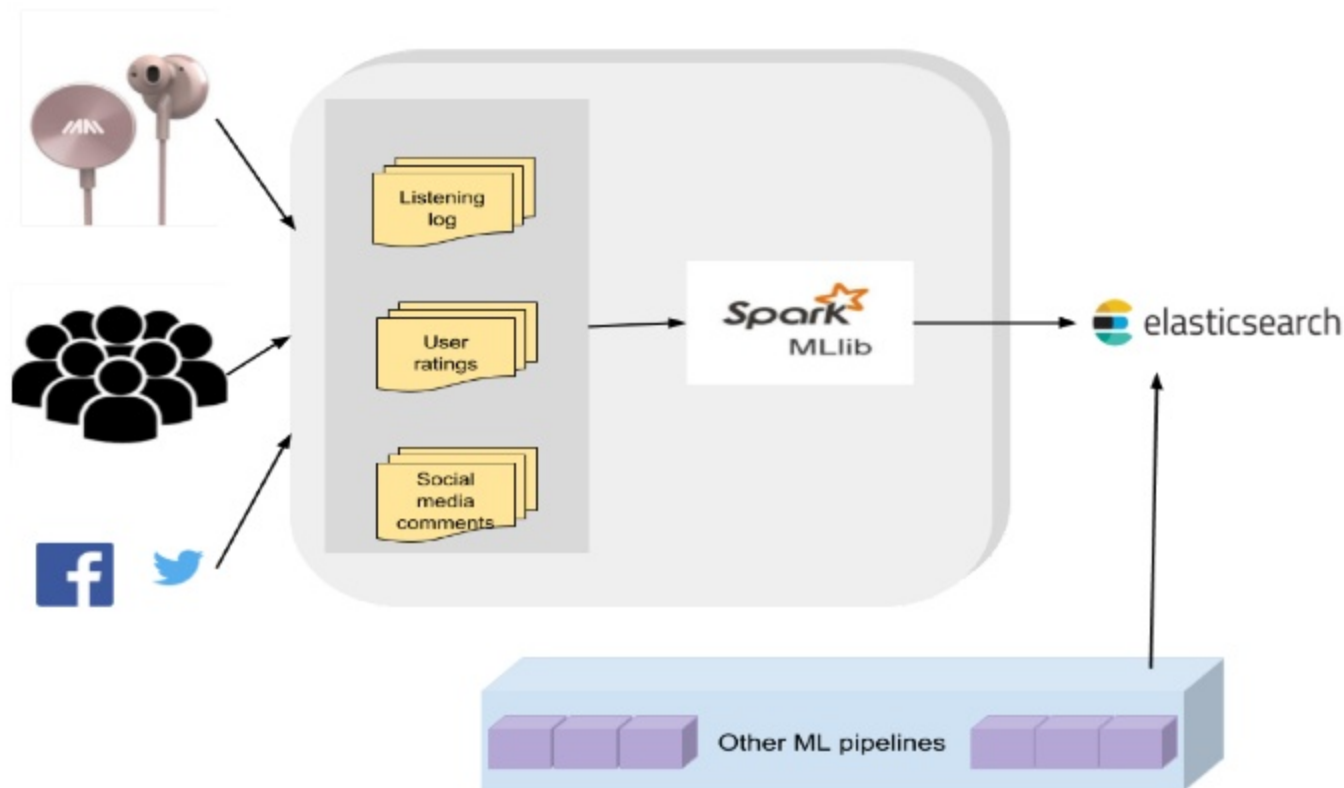
Data Challenges

- Heterogenous data type
- Streaming + Batch data
- Data Cleaning, enrichments
- Data Annotation

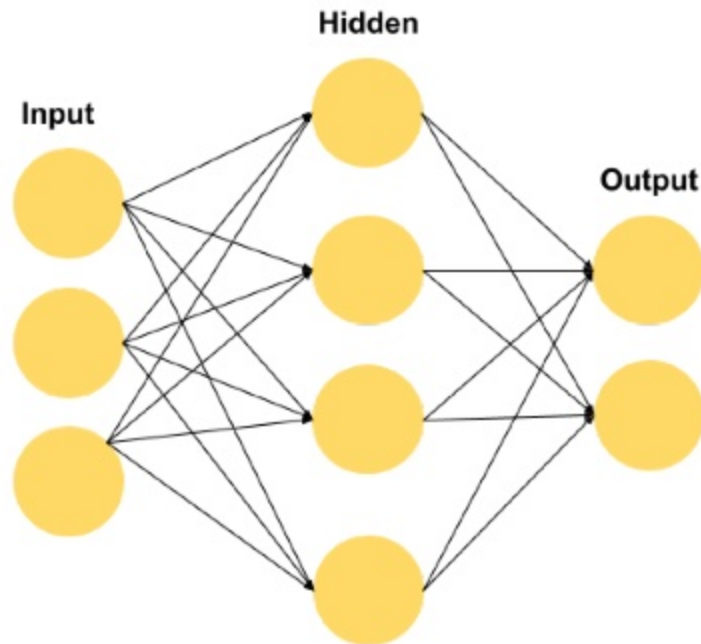
Spark MLlib

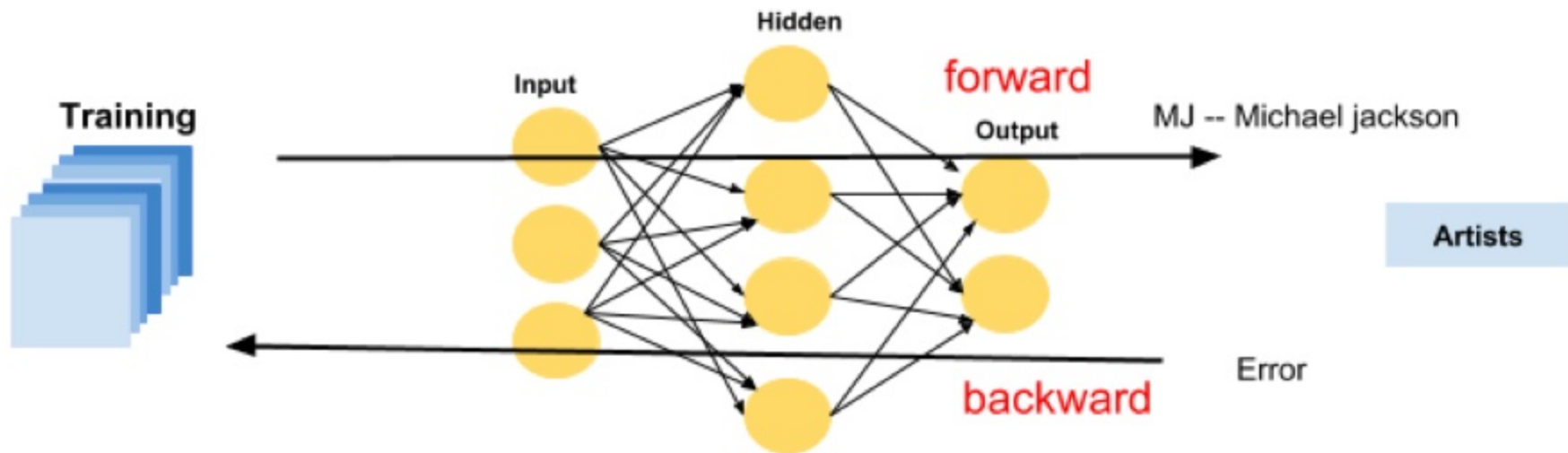
- Classification and Regression
- Clustering
- Collaborative filtering
- Frequent Pattern Mining

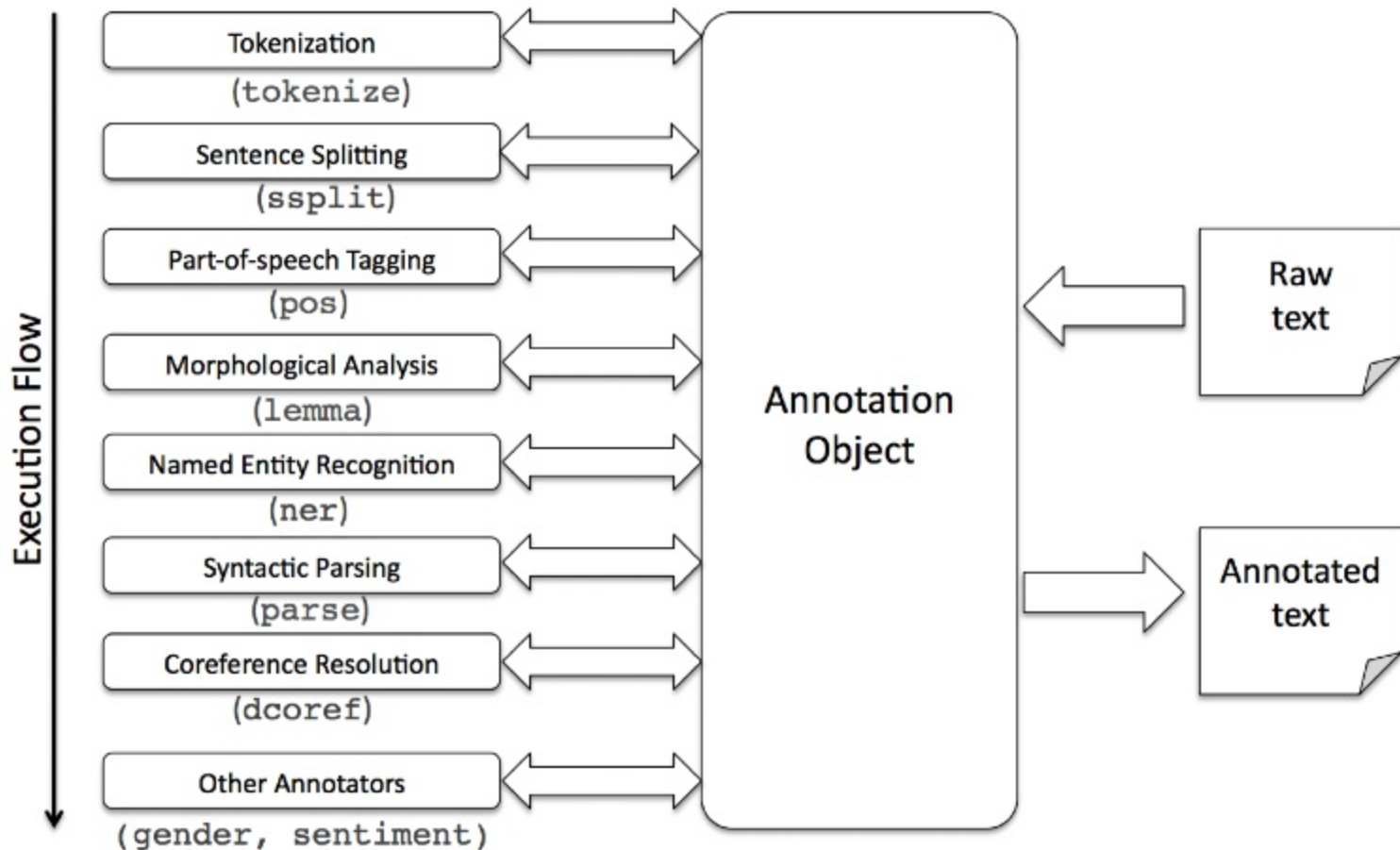
Data Pipeline on Scale



What is Deep learning?







Word embedding

- **Word embedding** is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where **words** or phrases from the vocabulary are mapped to vectors of real numbers

Word embedding model

- Word2Vec model
- Glove model
- fasttext model

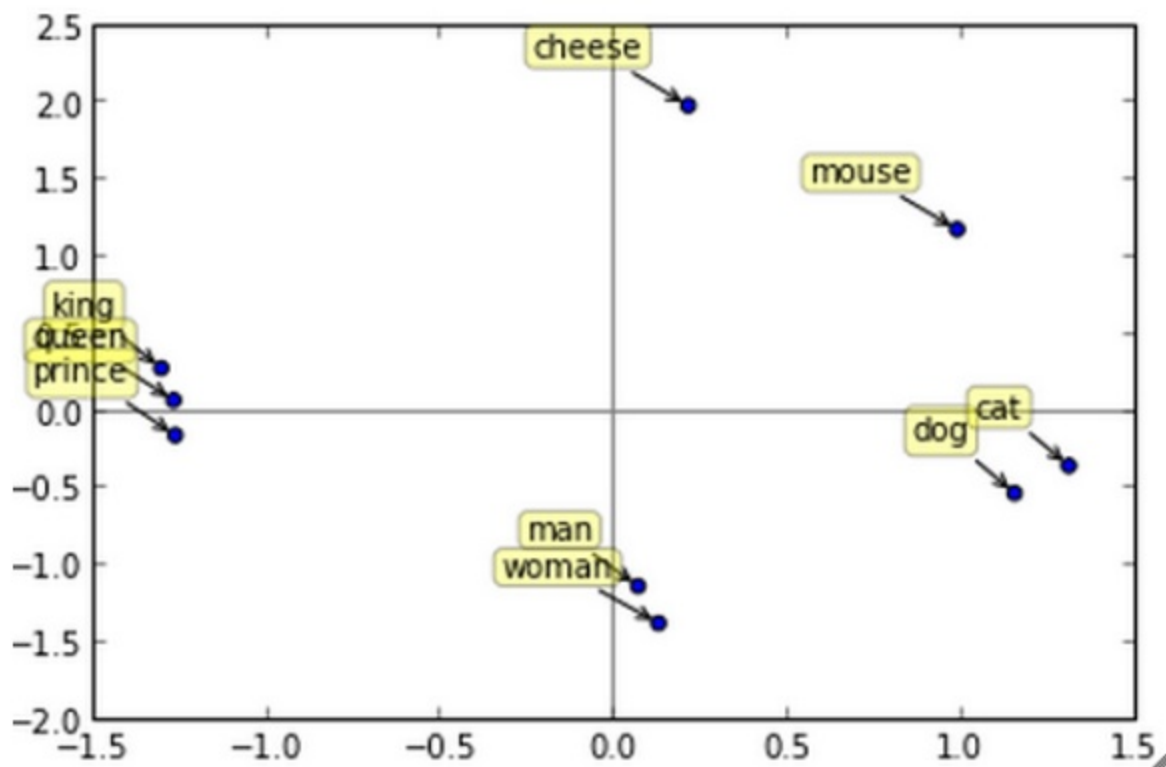
Idea for word2vec model

- Word similarity = vector similarity
- Key idea: Predict surrounding words of every word
- Faster and can easily incorporate a new sentence/document or add a word to the vocabulary
- Developed by Mikolov, Sutskever, Chen, Corrado and Dean in 2013 at Google Research

Representation

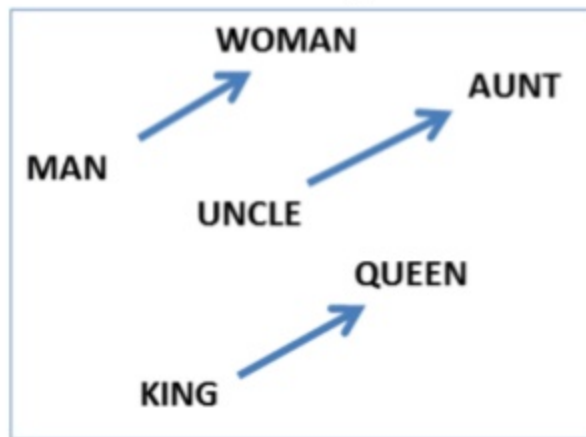
- Word meaning and relationships between words are encoded spatially
- Spatial distance corresponds to word similarity words are close together \Leftrightarrow their "meanings" are similar notation: word $w \rightarrow \text{vec}[w]$ its point in space, as a position vector.

Similar words are closer together

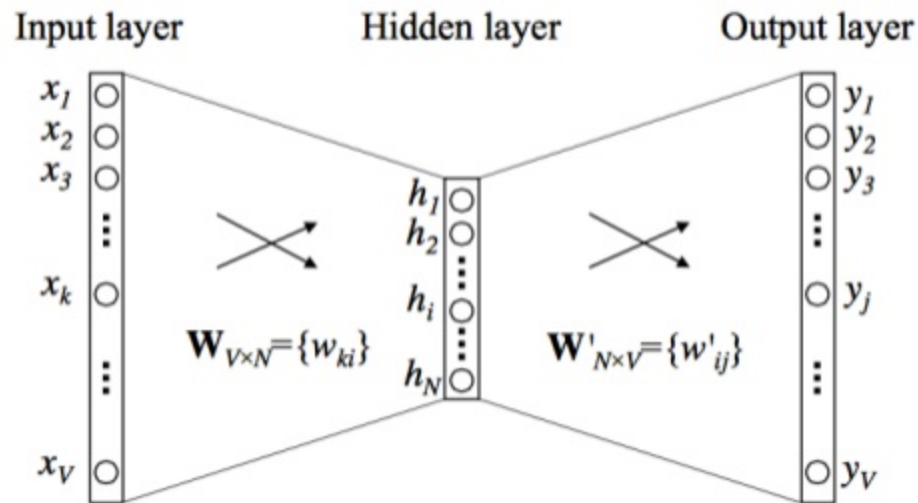
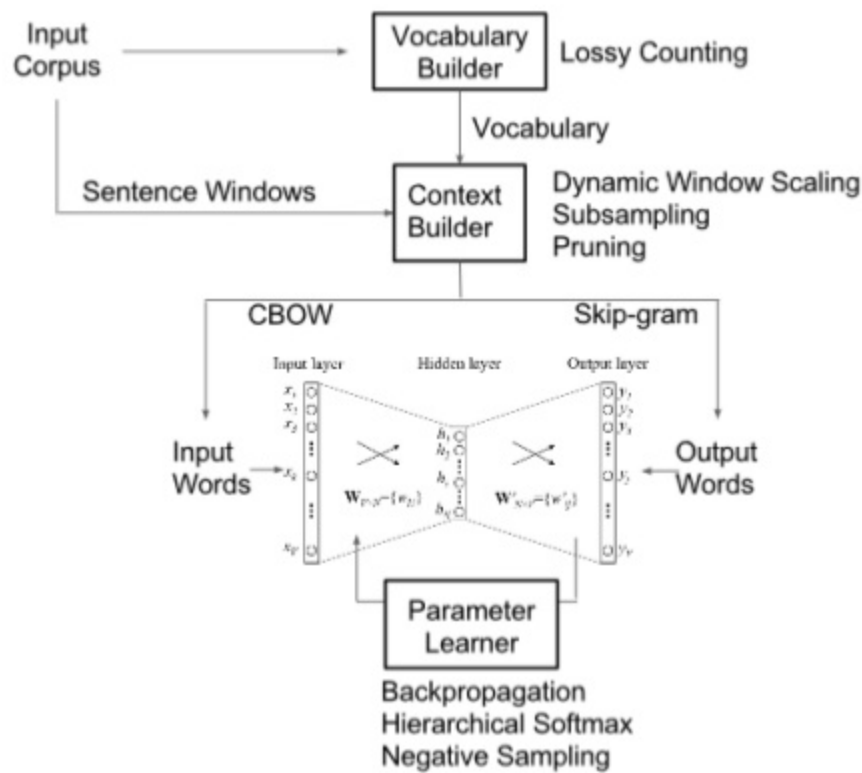


Word relationships are displacements

- The displacement (vector) between the points of two words represents the word relationship.
- Same word relationship \Rightarrow same vector

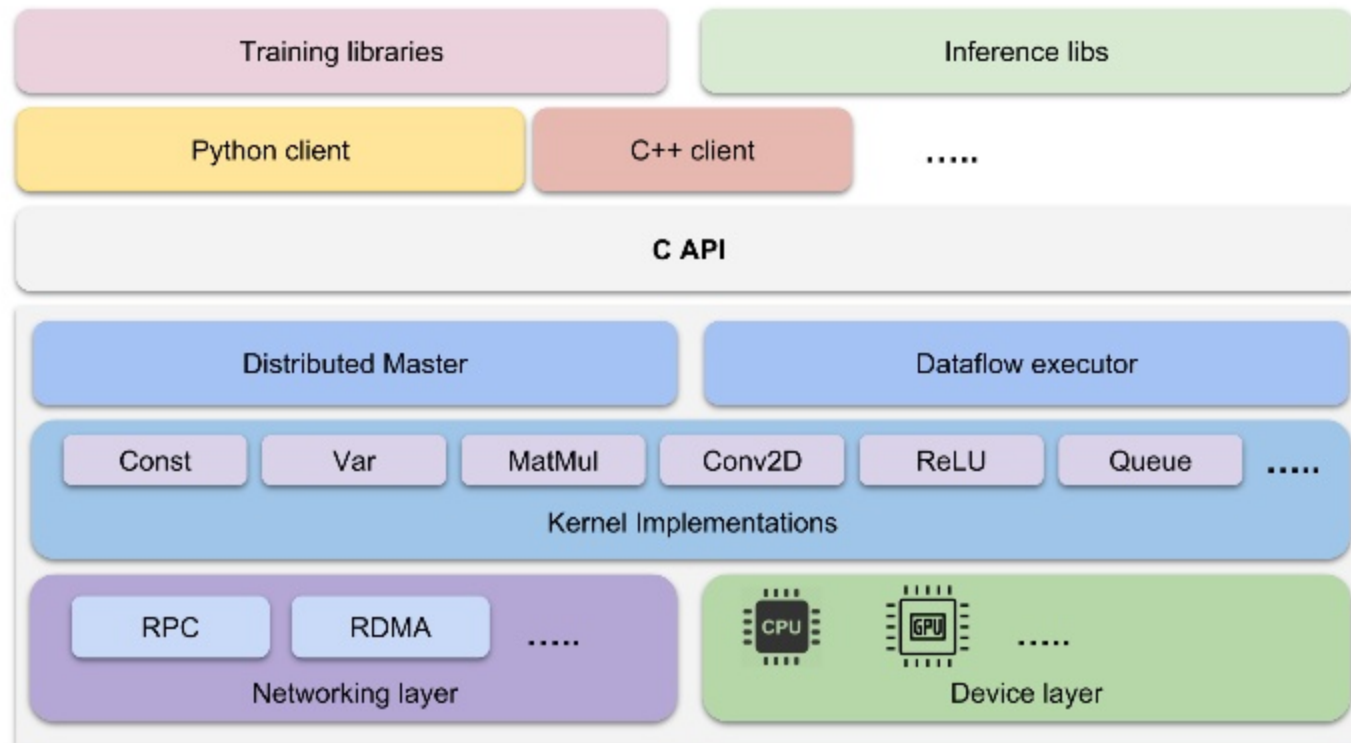


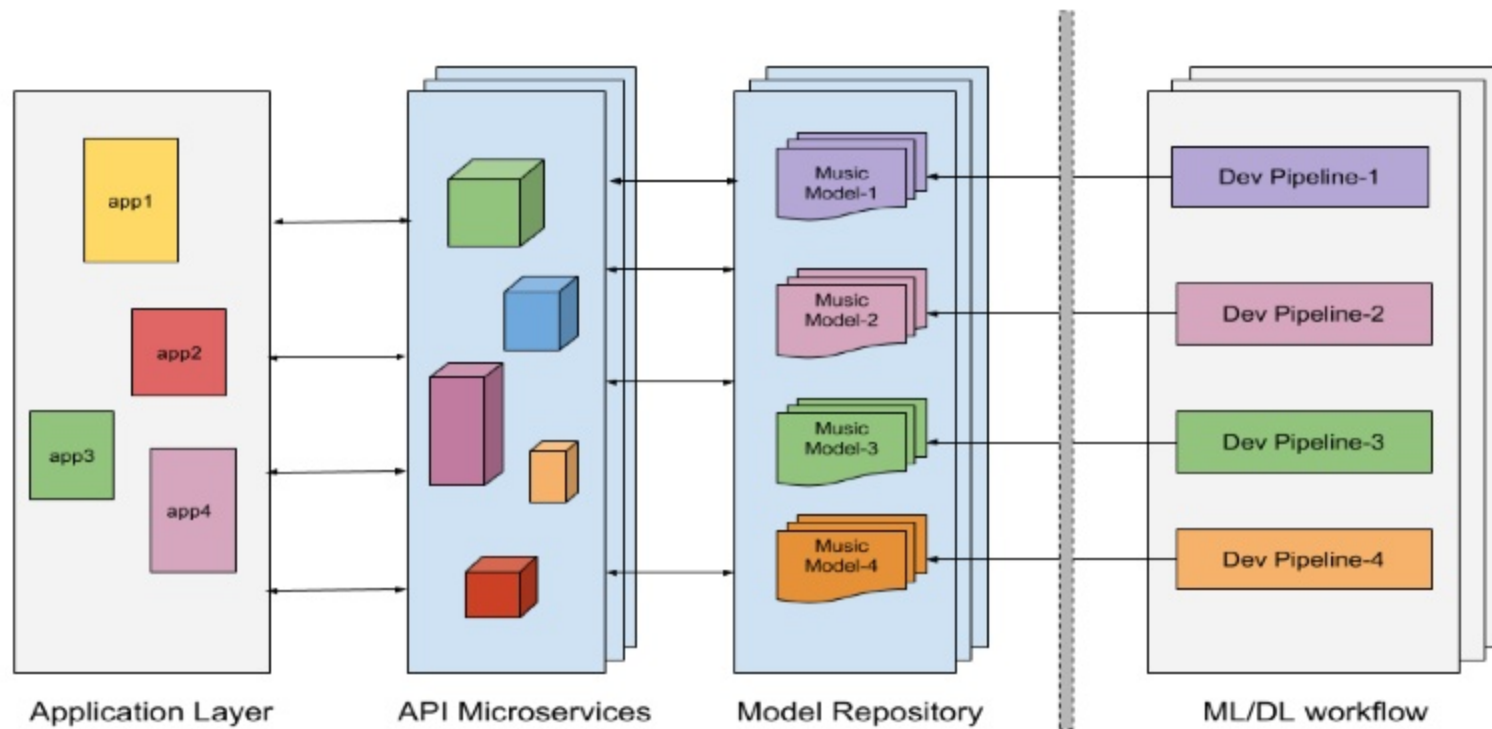
Architecture



- TensorFlow is an open-source machine learning library for research and production
- TensorFlow provides a variety of different toolkits that allow you to construct models at your preferred level of abstraction

TensorFlow architecture





Questions ?



<https://www.linkedin.com/in/rahulkumar-aws/>



@ rahul_kumar_aws