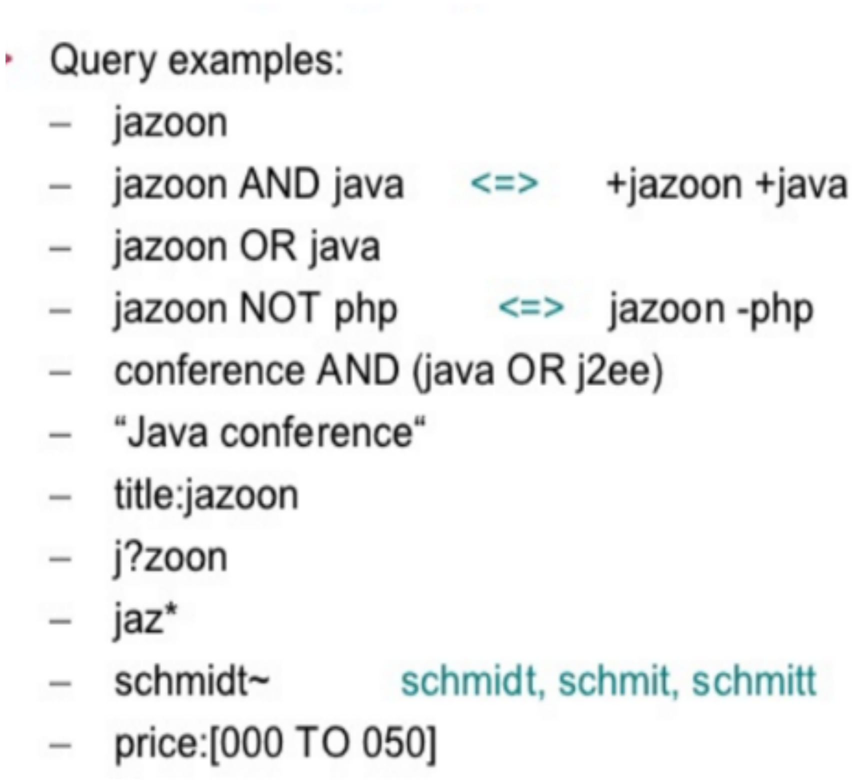
NATURAL LANGUAGE UNDERSTANDING AT SCALE WITH SPARK-NATIVE NLP, SPARK ML, AND TENSORFLOW

SPARK SUMMIT

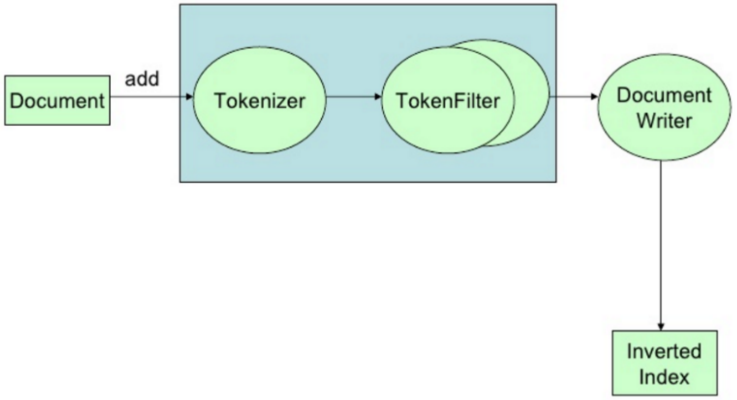
Alex Thomas | Data Scientist at indeed.com

Natural Language Understanding is an AI-Complete problem.

AT THE BEGINNING, THERE WAS SEARCH



- Query examples:
  - jazoon
  - jazoon AND java <=> +jazoon +java
  - jazoon OR java
  - jazoon NOT php <=> jazoon -php
  - conference AND (java OR j2ee)
  - “Java conference“
  - title:jazoon
  - j?zoon
  - jaz*
  - schmidt~ schmidt, schmit, schmitt
  - price:[000 TO 050]



Scalable & robust Indexing pipeline
Tokenizers & analyzers
Synonyms, spellers & auto-suggest
File formats & header boosting
Rankers, link & reputation boosting

THEN, YOU NEED TO UNDERSTAND LANGUAGE

Prescribing sick days due to diagnosis of influenza.	*Positive*
Jane complains about flu-like symptoms.	*Speculative*
Jane’s RIDT came back clean.	*Negative*
Jane is at risk for flu if she’s not vaccinated.	*Conditional*
Jane’s older brother had the flu last month.	*Family history*
Jane had a severe case of flu last year.	*Patient history*

Parts of Speech • Dependency Parsing • Co-reference Resolution • Entity Recognition

WHAT MAKES LANGUAGE HARD

- **Nuanced**
  - Sure / I agree / Absolutely! / Whatever / Yes sir / Just to see you smile ❤️
- **Fuzzy**
  - Blue, New, Tall, Child, Tell, Do
- **Contextual**
  - “Patient denies alcohol abuse”
- **Medium specific**
  - “SGTM c u in 15”
- **Domain specific**
  - *All forward-looking statements included in this document are based on information available to us on the date hereof, and we assume no obligation to revise or publicly release any revision to any such forward-looking statement, except as may otherwise be required by law.*

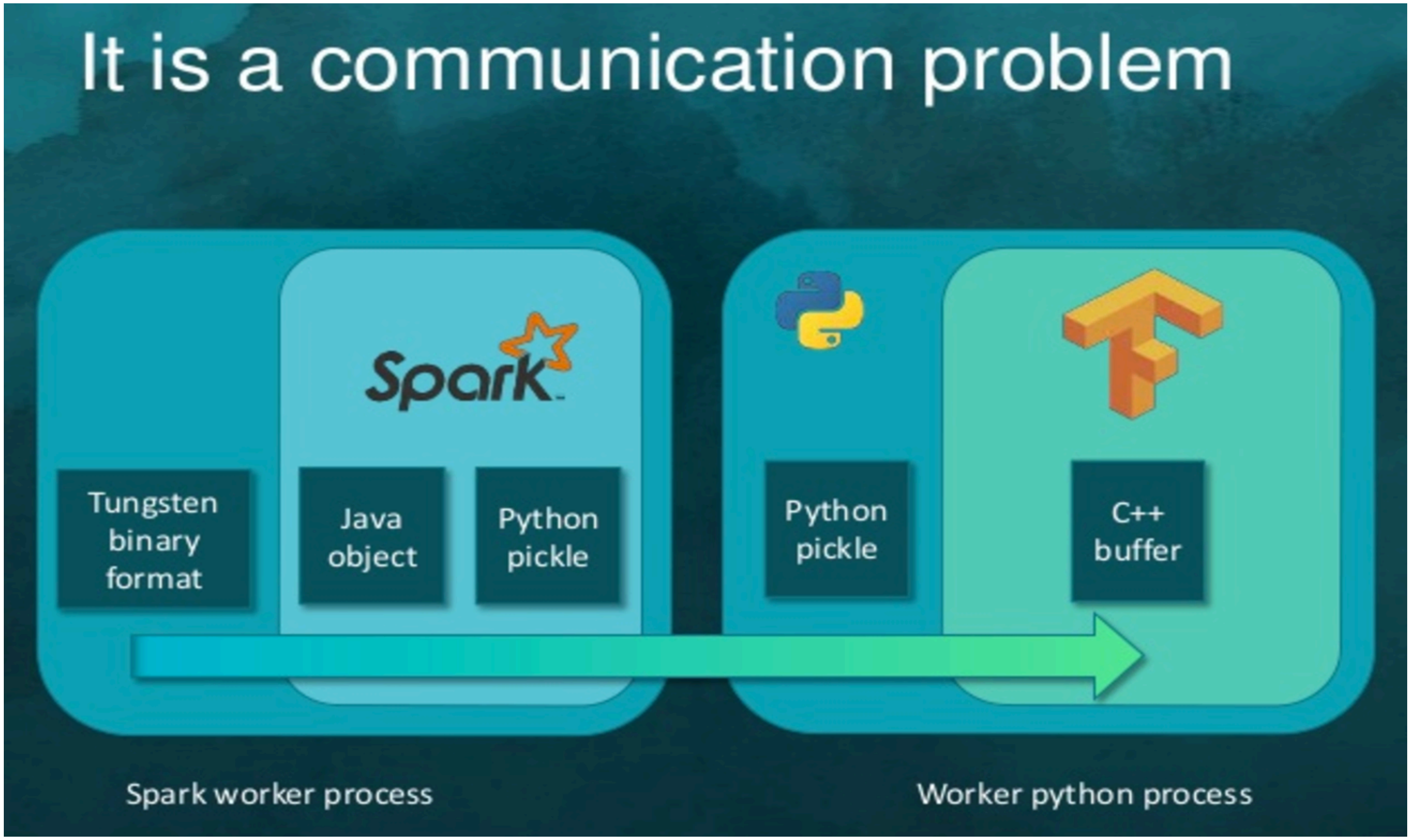
NLP LIBRARIES

- By Ecosystem:
  1. Python: NLTK, spaCy, gensim
  2. JVM: OpenNLP, CoreNLP, Spark NLP, UIMA, GATE, Mallet
  3. Others: tm for R, SAS, Watson, Matlab, …
- By Design:
  1. Raw functionality: NLTK, OpenNLP
  2. Annotation libraries: spaCy, Spark NLP, UMIA, GATE
- Industrial Grade, Open Source, Supported:
  1. spaCy
  2. Spark NLP

NLP FOR SPARK

- Production Grade NLP for the Apache Spark ecosystem
- Design Goals
    1. Performance & Scale
    2. Frictionless Reuse
    3. Enterprise Grade
- Build on top of the Spark 2.0 ML API's
- Apache 2.0 licensed, with active development & support

THE PERFORMANCE BOTTLENECK



It is a communication problem

Image Credit: Tim Hunter's TensorFrames overview

REUSE: OUT OF THE BOX FUNCTIONALITY

New in Spark NLP	Reused from Spark ML
Tokenizer	Topic modeling
Normalizer	Word 2 Vec
Stemmer	TF-IDF
Lemmatizer	String distance calculation
Entity extractor	N-grams calculation
Date extractor	Stop words removal
Part of Speech tagger	Classification
Named entity recognizer	Regression
Sentence boundary detection	Ensembles
Sentiment analysis	Train/Test & Cross-Validation
Spell checker	Grid Search

Demo: NLP for Spark in Action

Sample Notebook

Project Homepage

SUMMARY: WHEN DOES IT APPLY?

	Get by with rules, search, RegEx, attribute extraction	**Welcome to the world of NLP, ML and DL**
Social media	Does this social media post contain an offensive word?	Is this social media post offensive?
Legal	Find patents with the terms ‘car’ and battery’, or synonyms	Who is patenting next-gen electrical car batteries?
Support	Find products mentioned in customer emails or phone calls	What is this customer complaining about?
Finance	Extract the fee structure from a mutual fund prospectus	Are UK pensions allowed to invest in this fund?
Healthcare	Extract the patient’s blood pressure reading from a note	Does this patient have high blood pressure?

THANK YOU!

Special thanks to

- Saif Addin
- Dr. David Talby
- John Snow Labs