O'REILLY®
Artificial Intelligence Conference

PRESENTED WITH intel AI

Interpreting millions of patient stories with deep learned OCR and NLP

Stacy Ashworth
Chief Clinical Officer, SelectData

Alberto Andreotti
Data Scientist, John Snow Labs

oreillyaicon.com
#OReillyAI

Objectives

- Identifying the Problem
- From Human Labor to Automation
- Evaluating the Solution

Conceptualizing the Problem

- Defining the problem
  - What is Home Health?
  - What do we do?
  - Scaling Volume with the expectation of Expert Reviewer

Silver Tsunami VS the Expert Reviewer

Silver Tsunami

- By 2022 more than 25 percent of US workers will be 55 or older
- Nearly 10,000 baby boomers reach retirement age each day
- Home Health is expected to grow by 6.7% next year

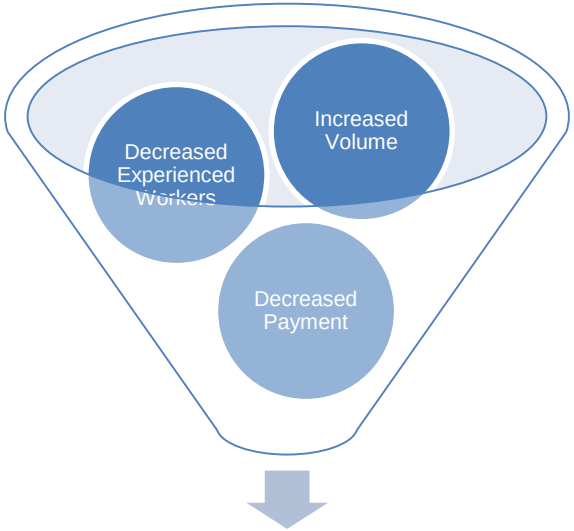
Expert Reviewer

- Bureau of Labor Statistics projects that the need for medical coders will increased by 15% by 2027
- Healthcare Data is used in decision-making

Payment Reform

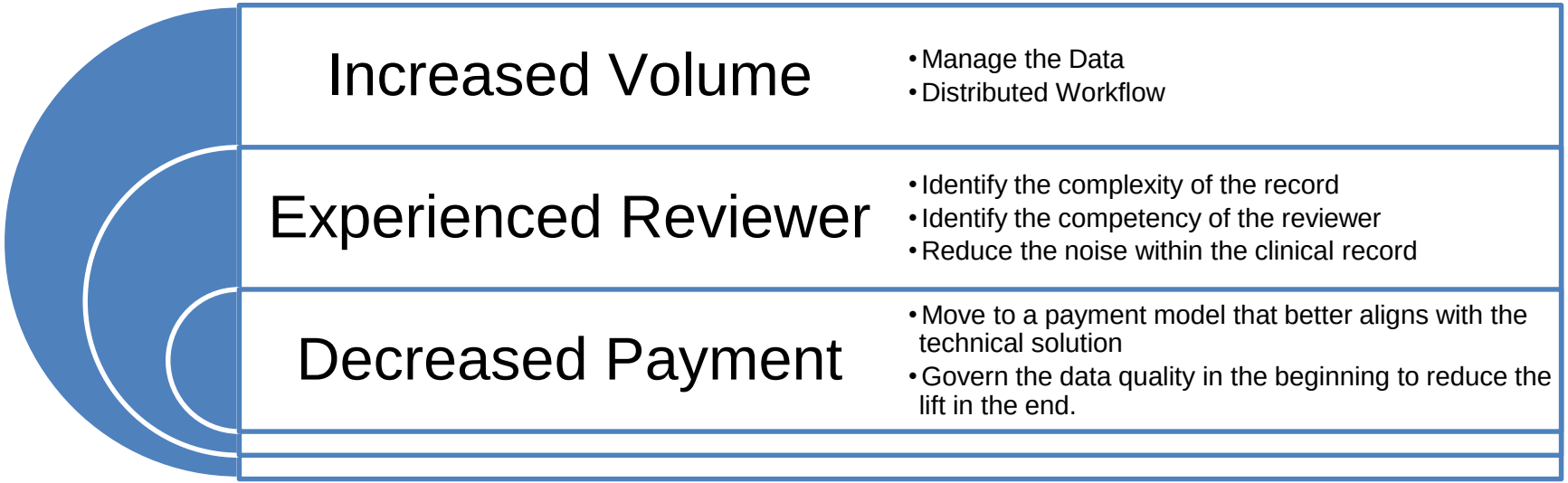
- Aging Baby Boomers
  - By 2039 the rate of Medicare spending and net interest on national debt will exceed total projected revenues
  - Payment reform focused on reduction in price

Conceptualizing the Problem



Problem

The Solution



Cupcakes, Puppies and Bombs

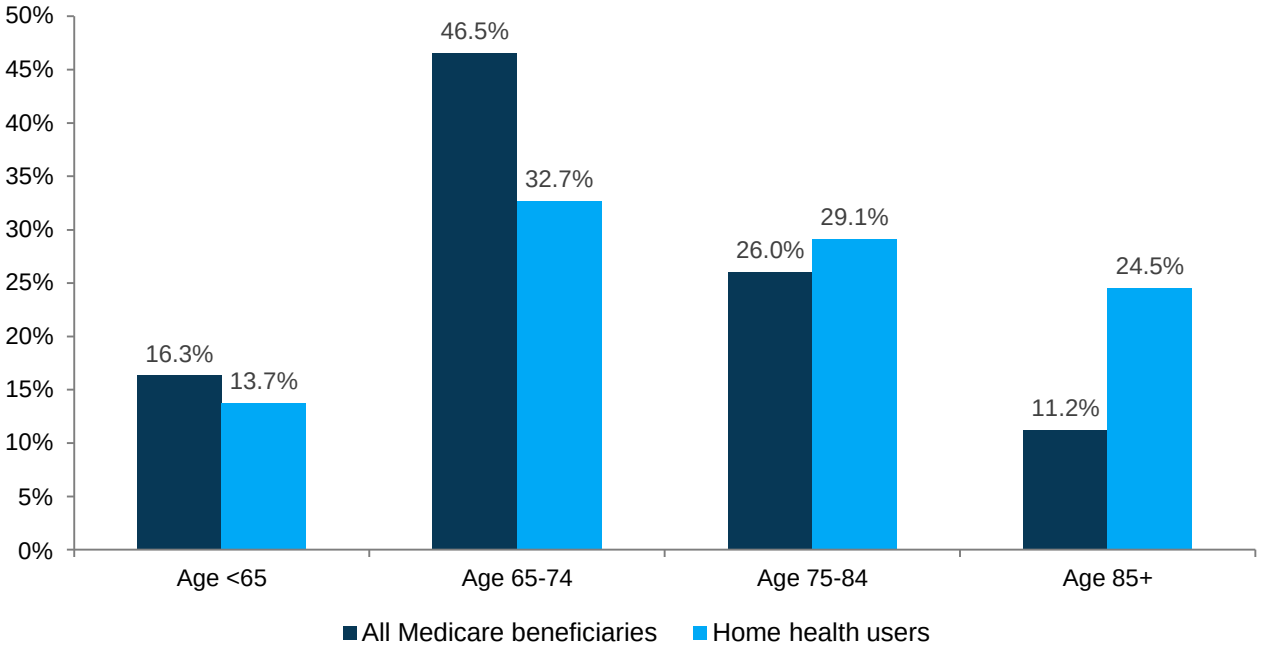
User Stories:
As a Manager I want to be able to identify assessments that are Hard, Medium and Easy by two metrics, degree of effort and perceived level of difficulty.

As a Manager I want to be able to identify employees that possess the appropriate skill level to complete the work for an assessment that has been categorized as Hard, Medium and Easy

Goal: Increase our overall production by 10% while ensuring the accuracy of the recommendation of 95%

What is Home Health?

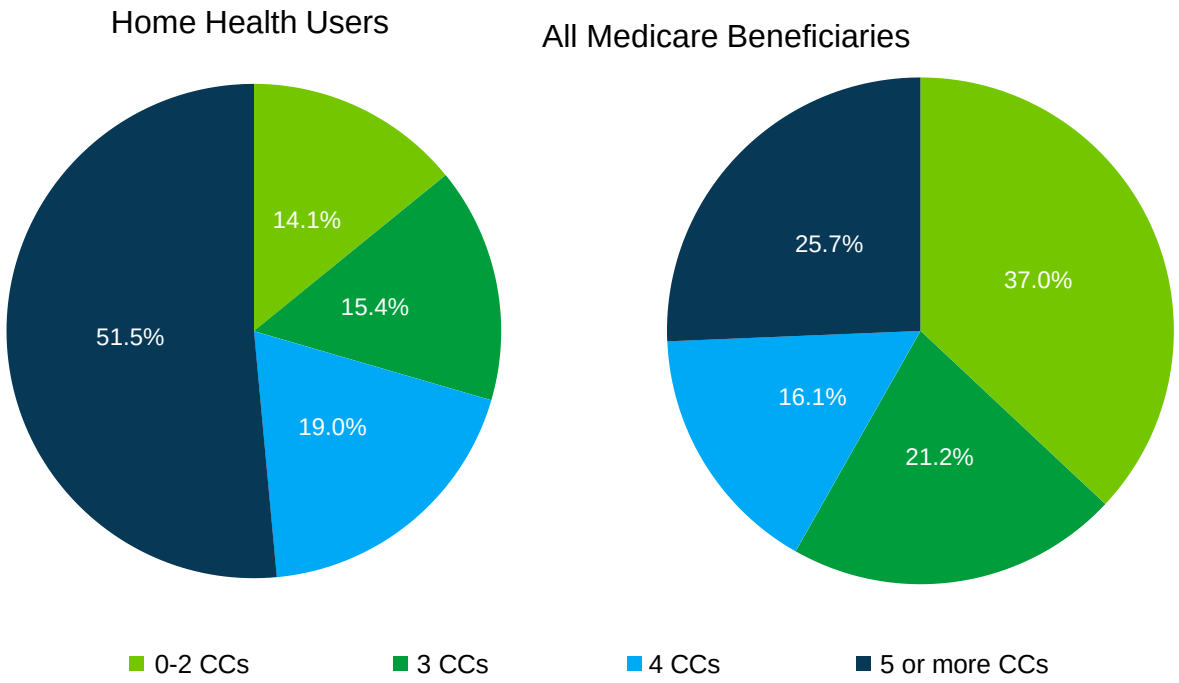
Chart 1.1: Age Distribution of Home Health Users and All Medicare Beneficiaries, 2015



Source: Avalere analysis of the Medicare Current Beneficiary Survey, Access to Care file, 2015.

Demographics of Home Health Users

Chart 1.8: Percentage of All Medicare Beneficiaries and Home Health Users by Number of Chronic Conditions (CCs), 2015



Source: Avalere analysis of the Medicare Current Beneficiary Survey, Access to Care file, 2015.
Totals may not sum to 100 percent due to rounding.

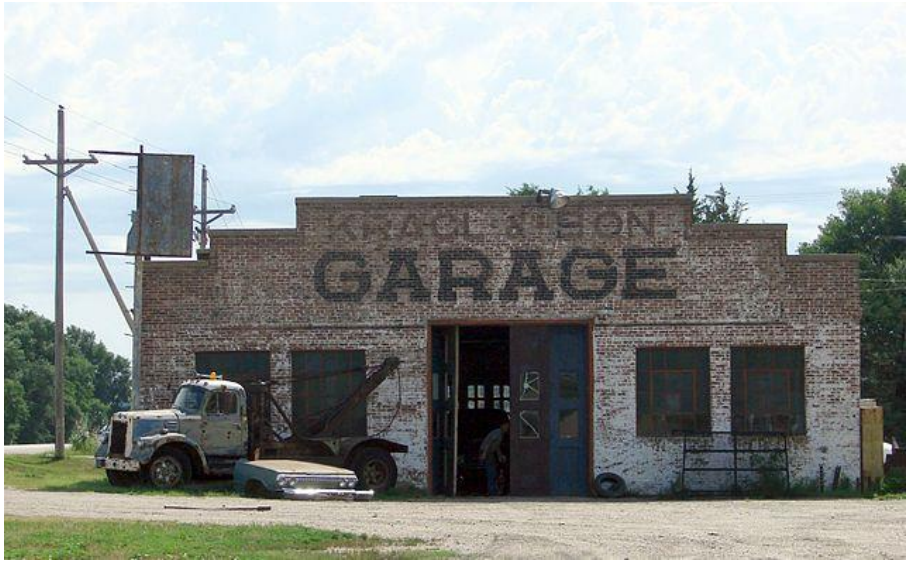
Changes in the Industry

- The IMPACT ACT
  - Quality Measures
  - Data Collection Instruments
    - Standardized
    - Governed by data rules
    - Accuracy is an issue

What are we doing?

Specialized

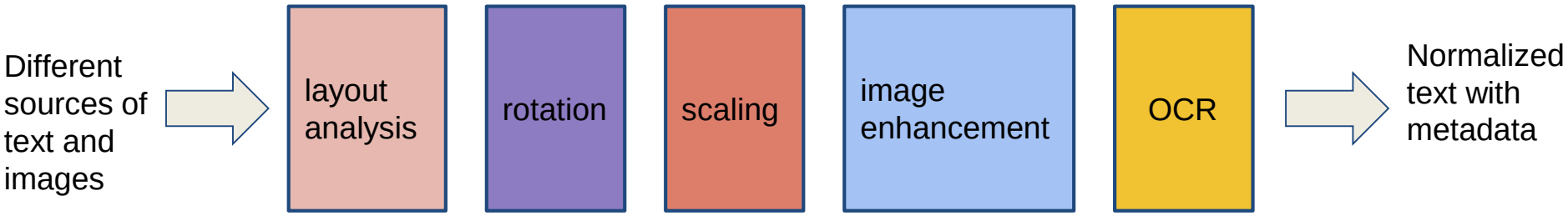
Generalized



Transforming the Problem

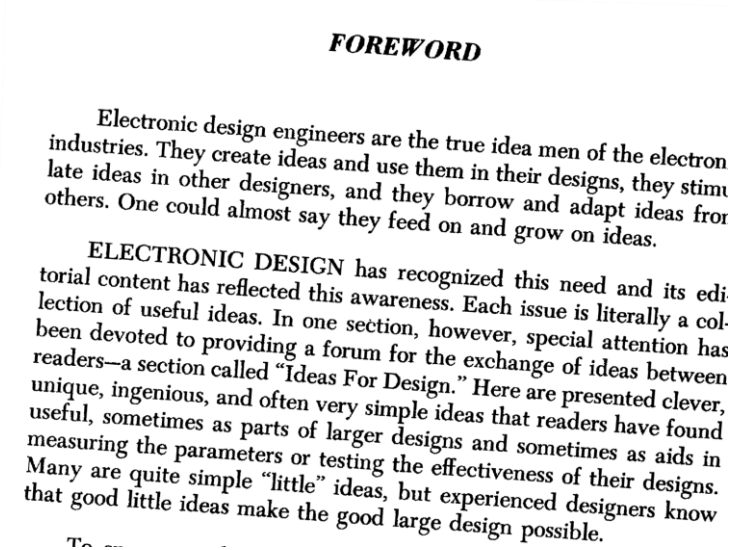
- OCR is difficult, different layouts, different scales, noise, rotation.
- High number of records and pages.
- Need for cluster processing.
- Cluster processing is difficult.

Spark OCR



- Independent project, integrates smoothly with Spark-NLP.
- Can handle both text and image documents.
- Handles image enhancements like denoising, rotation, scaling.
- Provides metadata like text coordinates, confidence, etc.

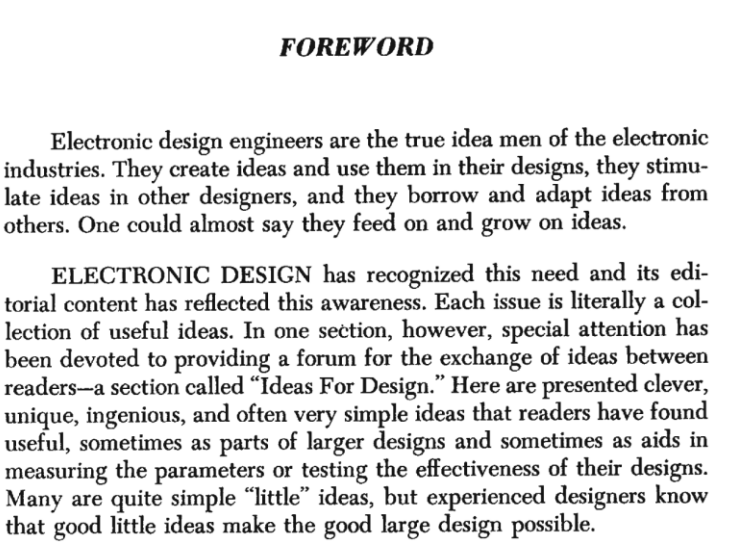
OCR challenges



FOREWORD

Electronic design engineers are the true idea men of the electron industries. They create ideas and use them in their designs, they stimu late ideas in other designers, and they borrow and adapt ideas fror others. One could almost say they feed on and grow on ideas.

ELECTRONIC DESIGN has recognized this need and its edi torial content has reflected this awareness. Each issue is literally a col lection of useful ideas. In one section, however, special attention has been devoted to providing a forum for the exchange of ideas between readers—a section called "Ideas For Design." Here are presented clever, unique, ingenious, and often very simple ideas that readers have found useful, sometimes as parts of larger designs and sometimes as aids in measuring the parameters or testing the effectiveness of their designs. Many are quite simple "little" ideas, but experienced designers know that good little ideas make the good large design possible.

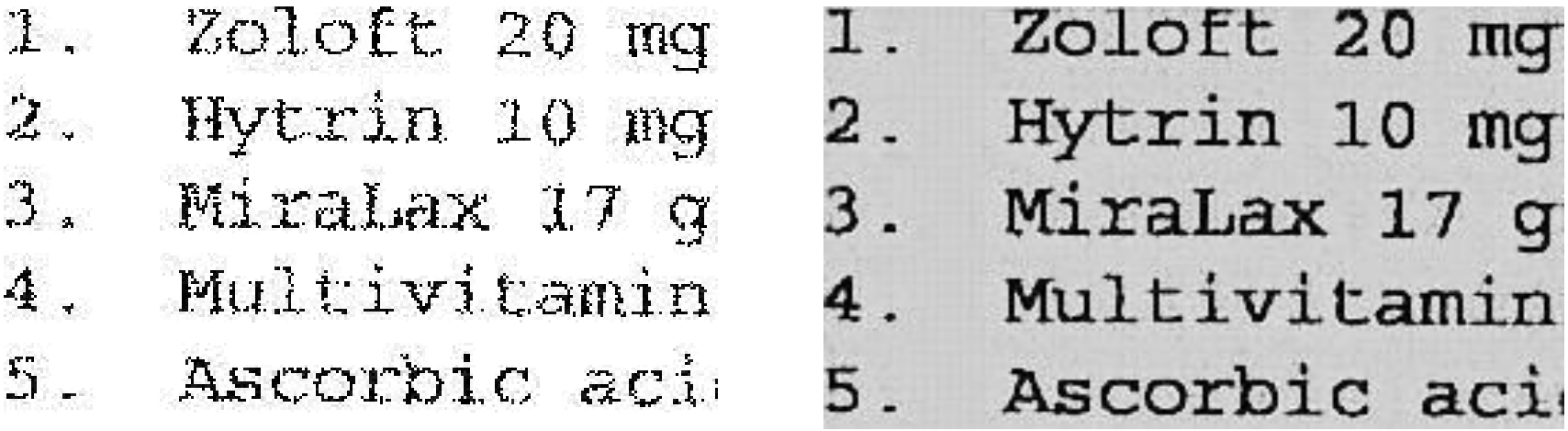


FOREWORD

Electronic design engineers are the true idea men of the electronic industries. They create ideas and use them in their designs, they stimulate ideas in other designers, and they borrow and adapt ideas from others. One could almost say they feed on and grow on ideas.

ELECTRONIC DESIGN has recognized this need and its editorial content has reflected this awareness. Each issue is literally a collection of useful ideas. In one section, however, special attention has been devoted to providing a forum for the exchange of ideas between readers—a section called "Ideas For Design." Here are presented clever, unique, ingenious, and often very simple ideas that readers have found useful, sometimes as parts of larger designs and sometimes as aids in measuring the parameters or testing the effectiveness of their designs. Many are quite simple "little" ideas, but experienced designers know that good little ideas make the good large design possible.

OCR challenges



Spark NLP

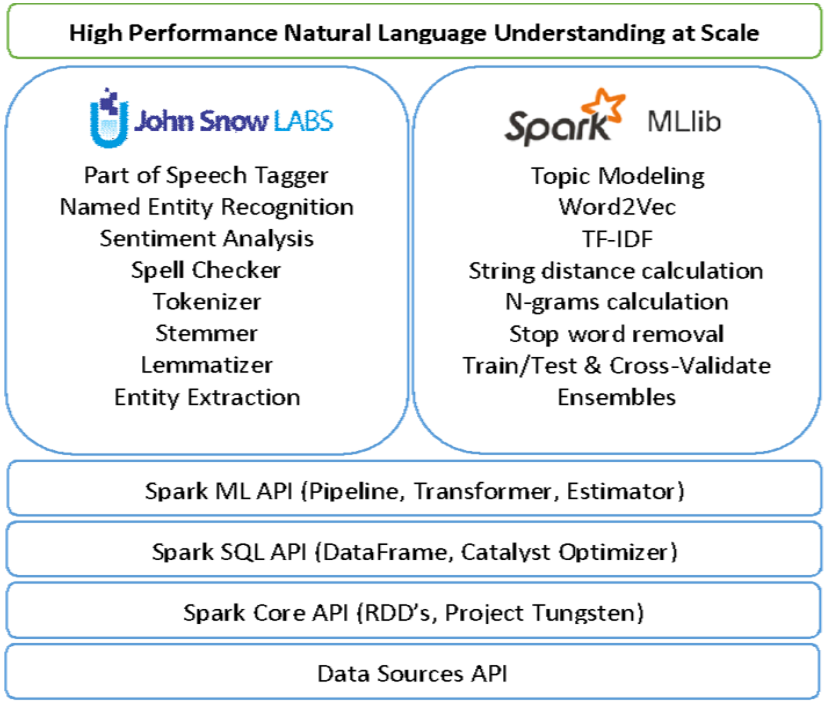
Design Goals

- State of the art Performance & Scale
- Frictionless Reuse
- Enterprise Grade

Built on the Spark ML API's

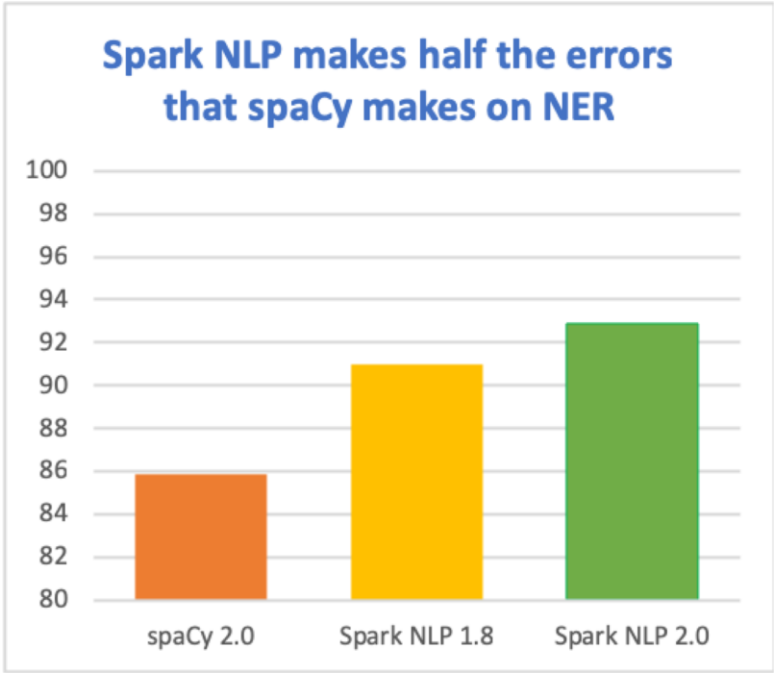
Apache 2.0 Licensed

Active development & support



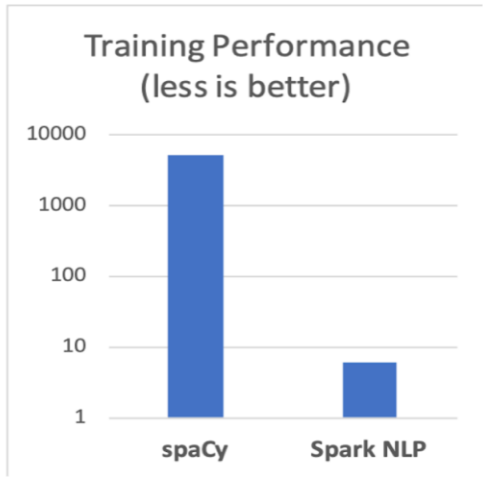
Accuracy

- ”State of the art” means the best performing academic peer-reviewed results
- NER Benchmark on right is on en_core_web_lg dataset, micro-averaged F1 score
- Why is it more accurate?
  - Deep learning models, trainable at scale with GPU’s
  - TF graph based on 2017 paper (bi-LSTM+CNN+CRF)
  - BERT embeddings
  - Contrib LSTM cells



Performance

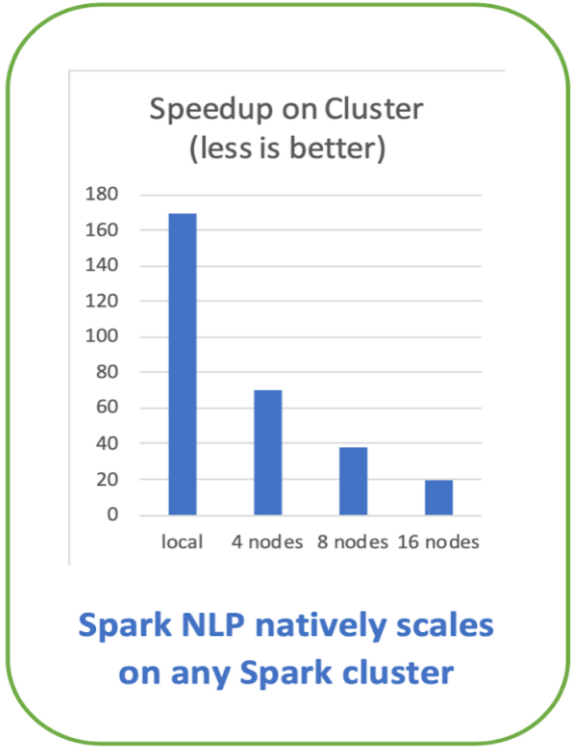
- Benchmark for training a pipeline with sentence bounder, tokenizer, and POS tagger
- Trained on single Intel i5 machine with 4 cores, 16GB RAM, SSD
- Why is it faster?
  - 2nd gen Tungsten engine:
  - whole stage code generation,
  - vectorized in-memory columnar data
  - No copying of text in memory
  - Extensive profiling, config & code
  - optimization of Spark and TensorFlow
  - Optimized for training and inference



Spark NLP trains 80x faster than spaCy on one machine

Scalability

- Zero code changes to scale a pipeline to any Spark cluster
- Only natively distributed open-source NLP library
- Spark provides execution planning, caching, serialization, shuffling
- Caveats
  - Speedup depends heavily on what you actually do
  - Not all algorithms scale well
  - Spark configuration matters



NLP FOR APACHE SPARK: COMBINED NLP & ML PIPELINES

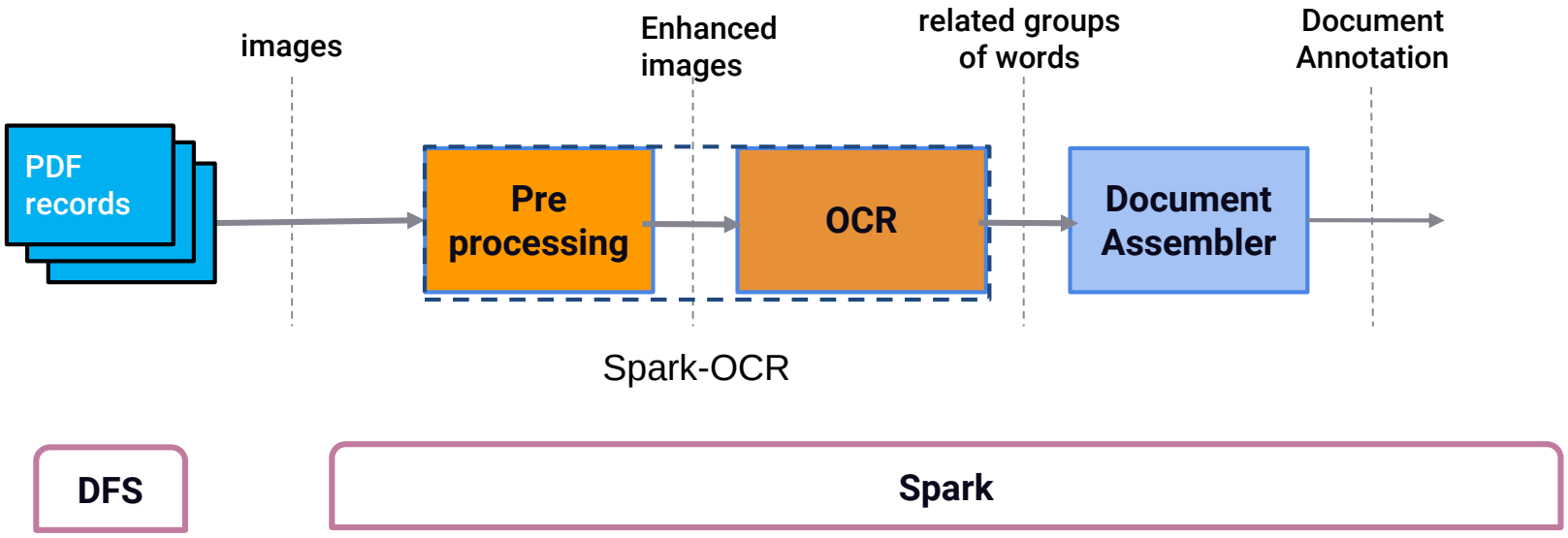
```
pipeline = pyspark.ml.Pipeline(stages=[
              document_assembler,
              tokenizer,
              stemmer,
              normalizer,
              stopword_remover,
              tf,
              idf,
              lda])

topic_model = pipeline.fit(df)
```

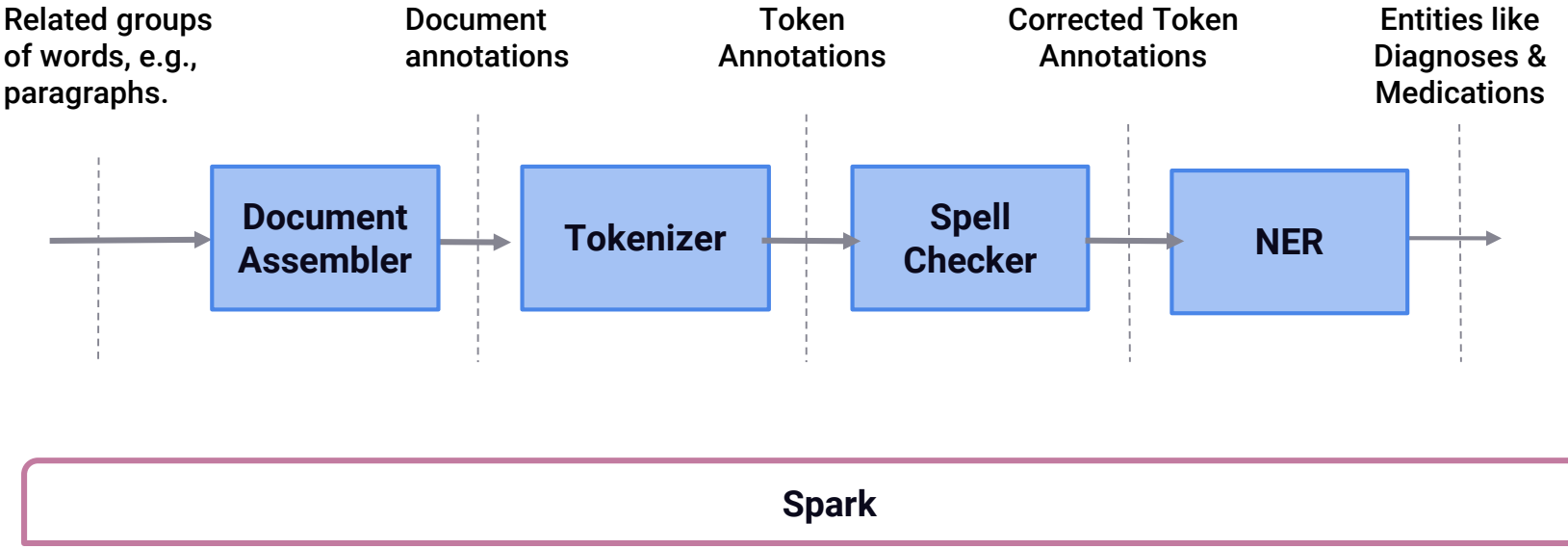
Solving the Problem

- We create a *pipeline*, composed by *annotators*.
- Spark-NLP is an annotation library.
- The pipeline runs in a cluster.
- We can process many documents *in parallel* and *scale out*.

Sample Pipeline



Sample pipeline(cont.)



History of Present Illness
Homer Simpson is a(n) 72 year old male with history of coronary artery disease, cardiomyopathy, diabetes type 2, hypertension, chronic kidney disease, and other comorbidities. He presents with rectal bleeding in the last two weeks.
No dyspnea or cough. No chest pain.

CONDITION ON TRANSFER: Stable but guarded. The patient is pain-free at this time.

MEDICATIONS ON TRANSFER:
1. Aspirin 325 mg once a day.
2. Metoprolol 50 mg once a day, but we have had to hold it because of relative bradycardia which he apparently has a history of.
3. Nexium 40 mg once a day.
4. Zocor 40 mg once a day, and there is a fasting lipid profile pending at the time of this dictation. I see that his LDL was 136 on May 3, 2002.
5. Plavix 600 mg p.o. x1 which I am giving him tonight.

Other medical history is inclusive for obstructive sleep apnea for which he is unable to tolerate positive pressure ventilation, GERD, arthritis

DISPOSITION: The patient and his wife have requested and are agreeable with transfer to Medical Center, and we are enclosing the CD ROM of his images.

History of Present Illness
Homer Simpson is a(n) 72 year old male with history of coronary artery disease, cardiomyopathy, diabetes type 2, hypertension, chronic kidney disease, and other comorbidities. He presents with rectal bleeding in the last two weeks.
No dyspnea or cough. No chest pain.

CONDITION ON TRANSFER: Stable but guarded. The patient is pain-free at this time.

MEDICATIONS ON TRANSFER:
1. Aspirin 325 mg once a day.
2. Metoprolol 50 mg once a day, but we have had to hold it because of relative bradycardia which he apparently has a history of.
3. Nexium 40 mg once a day.
4. Zocor 40 mg once a day, and there is a fasting lipid profile pending at the time of this dictation. I see that his LDL was 136 on May 3, 2002.
5. Plavix 600 mg p.o. x1 which I am giving him tonight.

Other medical history is inclusive for obstructive sleep apnea for which he is unable to tolerate positive pressure ventilation, GERD, arthritis

DISPOSITION: The patient and his wife have requested and are agreeable with transfer to Medical Center, and we are enclosing the CD ROM of his images.

History of Present Illness
Homer Simpson is a(n) 72 year old male with history of coronary artery disease, cardiomyopathy, diabetes type 2, hypertension, chronic kidney disease, and other comorbidities. He presents with rectal bleeding in the last two weeks.
No dyspnea or cough. No chest pain.

CONDITION ON TRANSFER: Stable but guarded. The patient is pain-free at this time.

MEDICATIONS ON TRANSFER:
1. Aspirin 325 mg once a day.
2. Metoprolol 50 mg once a day, but we have had to hold it because of relative bradycardia which he apparently has a history of.
3. Nexium 40 mg once a day.
4. Zocor 40 mg once a day, and there is a fasting lipid profile pending at the time of this dictation. I see that his LDL was 136 on May 3, 2002.
5. Plavix 600 mg p.o. x1 which I am giving him tonight.

Other medical history is inclusive for obstructive sleep apnea for which he is unable to tolerate positive pressure ventilation, GERD, arthritis

DISPOSITION: The patient and his wife have requested and are agreeable with transfer to Medical Center, and we are enclosing the CD ROM of his images.

Annotations - document

([document,
0,
307,
History of Present Illness
Homer Simpson is a(n) 72 year old male with history of coronary altery disease, cardiomiuopathy, diabetes type 2, hypertension, chronic kidney disease, and other comorbidities. He presents with rectal bleeding in the last two weeks.
No dyspnea or cough. No chest pain .,
Map(sentence -> 0)])

Annotations - token

…

13 = "[token,62,65,male]"

14 = "[token,67,70,with]"

15 = "[token,72,78,history]"

16 = "[token,80,81,of]"

17 = "[token,83,90,**comonary**]"

18 = "[token,92,97,**altery**]"

19 = "[token,99,105,disease]"

...

Annotations - spell checked

…

13 = "[token,62,65,male]"

14 = "[token,67,70,with]"

15 = "[token,72,78,history]"

16 = "[token,80,81,of]"

17 = "[token,83,90,coronary]"

18 = "[token,92,97,artery]"

19 = "[token,99,105,disease]"

...

Annotations - entities

...
13 = "[named_entity,60,63,O,Map(word -> male)]"
14 = "[named_entity,65,68,O,Map(word -> with)]"
15 = "[named_entity,70,76,O,Map(word -> history)]"
16 = "[named_entity,78,79,O,Map(word -> of)]"
17 = "[named_entity,81,88,**B-PROBLEM**,Map(word -> coronary)]"
18 = "[named_entity,90,95,**I-PROBLEM**,Map(word -> artery)]"
19 = "[named_entity,97,103,**I-PROBLEM**,Map(word -> disease)]"
...

Annotations - ner converter

0 = "[chunk,81,103,coronary artery disease,Map(entity -> PROBLEM, sentence -> 0, chunk -> 0)]"

1 = "[chunk,107,120,cardiomyopathy,Map(entity -> PROBLEM, sentence -> 0, chunk -> 1)]"

2 = "[chunk,124,138,diabetes type 2,Map(entity -> PROBLEM, sentence -> 0, chunk -> 2)]"

3 = "[chunk,142,153,hypertension,Map(entity -> PROBLEM, sentence -> 0, chunk -> 3)]"

Proof of Concept

Cupcakes, Puppies & Bombs

Feedback Loop

- Perceived Level of Difficulty-PLOD
  - Subjective measurement
  - Validated using objective measures
  - Comparison among levels of experience among employees
  - Normalized to account for these differences.
- Degree of Effort-DOE
  - Measured using minutes spent within the record
  - Time-stamps from all applications transformed into minutes

From Cupcakes to Automated Coding

Distrust of AI among healthcare professionals

- Emphasis placed on intuition
- General lack of knowledge regarding programming
- Fear of lost employment

Distrust to Trust

- Exposure
- Begin by replacing small pieces of the mundane within the process
- Augmented Intelligence
- Gradually transform the role
- Coding Specialist then becomes a Quality Specialist focused on ensuring the accuracy of the model

Sample Notebook

THANK YOU!

To try Spark NLP. Getting Started, Documentation, Examples, Videos, Blogs, Code, and an active Slack Community,
https://nlp.johnsnowlabs.com

To bounce ideas:

Alberto Andreotti
Stacy Ashworth

SelectData™

John Snow LABS