Spark NLP in Action:
Improving Patient Flow Forecasting

Santosh Kulkarni, Product Leader, Kaiser Permanente

Dr. David Talby, CTO, Pacific AI

Contents

About Kaiser Permanente

Mission	Vision
Kaiser Permanente exists to provide high-quality, affordable health care services and to improve the health of our members and the communities we serve	We are trusted partners in total health, collaborating with people to help them thrive and creating communities that are among the healthiest in the nation.

39	680	240,000+	$65B+
Hospitals	Clinics	Employees	Revenue

Problem Statement

"Hidden Technical Debt in Machine Learning Systems", Google, NIPS 2015

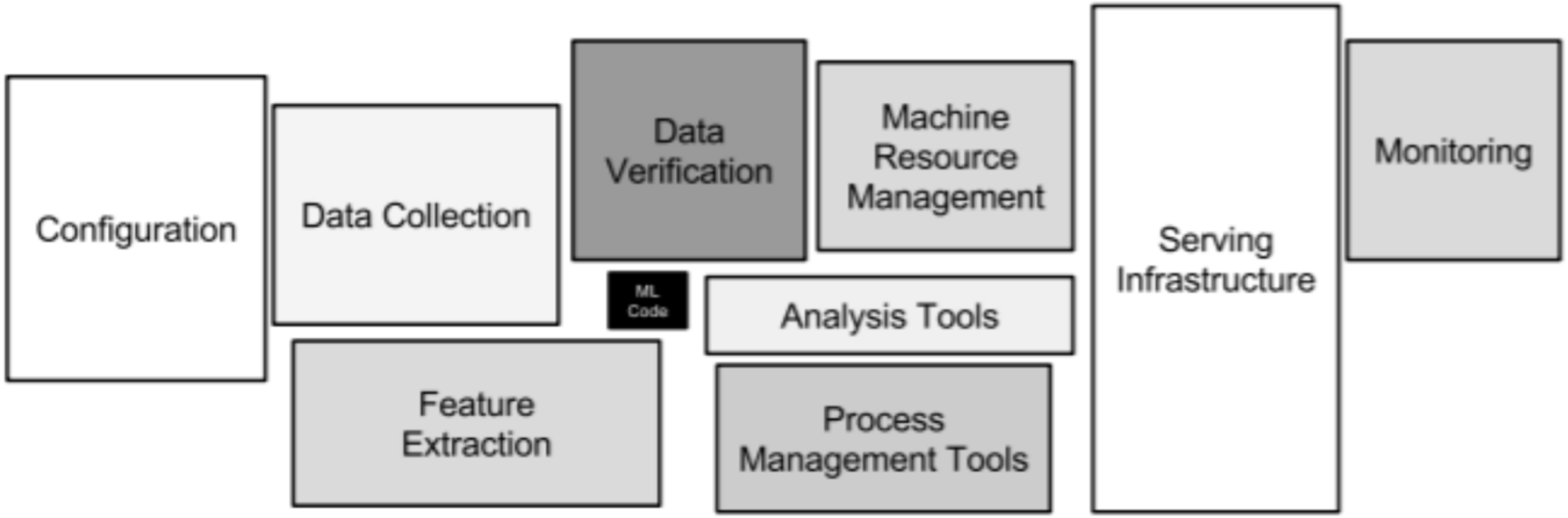


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Approach: Enterprise Scale & Enterprise Grade

High productivity toolset for data scientists working in programming languages like Python or R

Cutting-edge algorithms for a broad variety of data science problems

Self-service data discovery, visualization & analysis without coding

Machine learning, data mining & deep learning on unstructured natural language

Out-of-the-box, reusable, healthcare-specific models & datasets

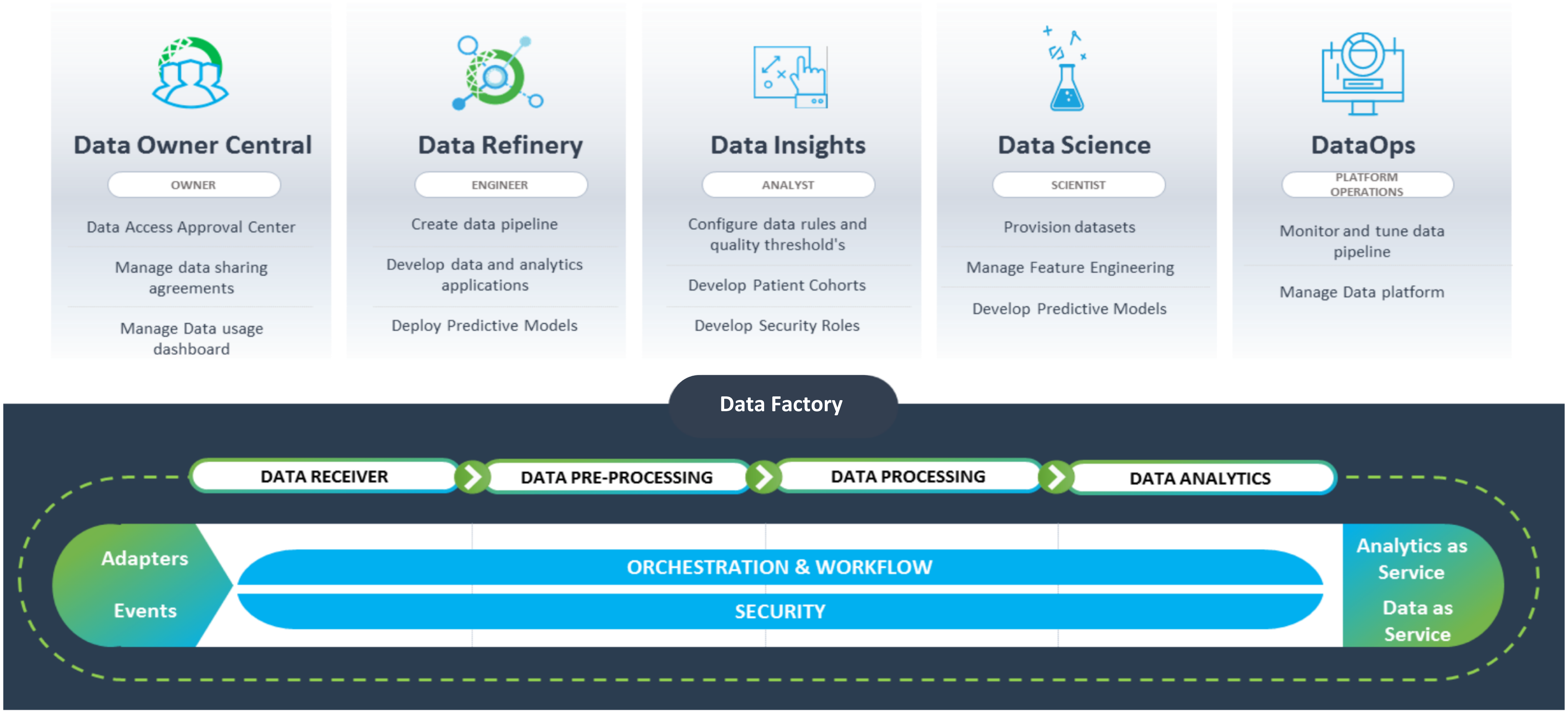
Continuously updated, clean, linked & enriched content packs

Productize machine learning models quickly, at enterprise-grade scale & reliability

Tools supporting best practices for validating, versioning, sharing & reusing models

Seamless integration with big data platforms, using Spark like execution engines

Systems of Intelligence – 'Data Factory'



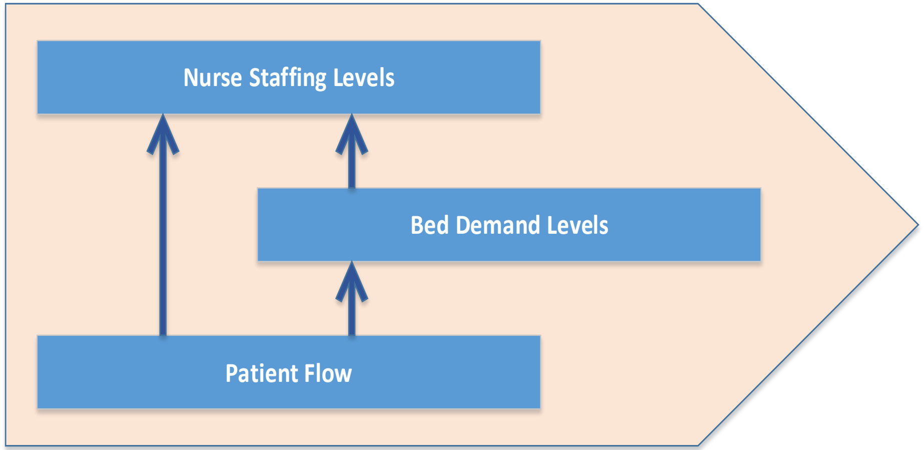
Improving Patient Flow Management: Problem Statement

- Hospitals today face numerous challenges that are straining their existing bed and service capacity and driving the need for improved patient flow management.
- The challenges include increased demand for services, clinical staff shortages, lack of tools and technology to adequately measure and manage patient flow, the risk of patient deterioration due to prolonged hospital stays and sometimes fewer available beds.
- With the continued aging of the U.S. population and accelerated clinical technology advances, demand for inpatient bed capacity is projected to rise by nearly 4-5% per year.

Objectives

Optimize the patient flow models & provide insights,
for real-time decision-making and for strategic planning,
by predicting:

- Bed demand
- 'Safe' staffing levels
- Hospital gridlock



Why NLP?

Key factors that influence a patient's flow (How likely they are to admitted? For how long? For what?):

- **Volume of arrivals**
  - Outpatient
  - Referrals
  - Emergency Room
  - Operation Room
- **Timing of arrival**
  - Hour of the day
  - Day of the week
  - Holidays
- **Admission specialty**
  - Oncology
  - Hip-replacement
  - Renal Disease
  - Cardiology, …
- **Seasonal variables**
  - Flu season
  - Natural disasters
- **Acuity level of patient**:
  - Symptoms
  - Onset of symptoms
  - Vital signs
- **Patient's length of stay** per unit (ICU, CVICU, …)
- **Pain**
  - Type of pain
  - Intensity of pain
  - Body part or region
- **Ongoing treatment**
  - Attempted at home
  - Prescriptions taken
  - Diet, sleep, …
- **Nurse staffing levels & skill mix:**
  - Certified Nurses
  - Licensed N.P.'s
  - Unlicensed Staff
  - Unique certifications

Some of the most relevant factors are only available within free-text clinical notes.

Moral #1:

NLP is just a small part of building an NLP AI solution

Enterprise Data Science Platform Components

Data Analyst

Kibana
Interactive data analysis without coding

Data Scientist

Jupyter Notebook & Hub
Interactive data science in Python & R

Applications

Model Server
Scalable & Secure REST API's

Content Packs

Terminology
Anatomy, Organisms, Codes, …

Providers
Enriched NPPES Database

Adverse Events
Enriched FAERS Database

Genomics
Gene products, associations, targets

Demographics
Population & Crime by ZIP code

Discovery & Visualization

ElasticSearch
Full-text, faceted & geospatial search

Visualization
Real-time, drag & drop dashboards

Time Series
Interactive time series analysis

Datalastic
Content pack ingestion & governance

Modeling & Experimentation

Data Science Ensemble
Machine Learning & Data Mining

Feature Repository
Reusable Healthcare features

John Snow Labs NLP
Natural Language Understanding

Clinical ML Models
Sentiment, NER, Risk, …

Spark ML
Performant machine learning at scale

Model Productization

CI & CD for Models
Auto-test & deploy machine learning

Model Repository
Reusable & versioned model store

NLP Model Visualizer
See entity extraction & dependencies

Patient 360 Data Model
Reusable features & models

Infrastructure

DataLab
Portal & Single Sign on

Kubernetes
Container Orchestration

Prometheus
Live Monitoring

KeyCloak
Authentication, Authorization, Audit

Enterprise NLP Platform Components

Data Analyst

Kibana
Interactive data analysis without coding

Data Scientist

Jupyter Notebook & Hub
Interactive data science in Python & R

Applications

Model Server
Scalable & Secure REST API’s

Content Packs

Terminology
Anatomy, Organisms, Codes, …

Providers
Enriched NPPES Database

Adverse Events
Enriched FAERS Database

Genomics
Gene products, associations, targets

Demographics
Population & Crime by ZIP code

Discovery & Visualization

ElasticSearch
Full-text, faceted & geospatial search

Visualization
Real-time, drag & drop dashboards

Time Series
Interactive time series analysis

Datalastic
Content pack ingestion & governance

Modeling & Experimentation

Data Science Ensemble
Machine Learning & Data Mining

Feature Repository
Reusable Healthcare features

John Snow Labs NLP
Natural Language Understanding

Clinical ML Models
Sentiment, NER, Risk, …

Spark ML
Performant machine learning at scale

Model Productization

CI & CD for Models
Auto-test & deploy machine learning

Model Repository
Reusable & versioned model store

NLP Model Visualizer
See entity extraction & dependencies

Common Domain Models
Reusable features & models

Infrastructure

DataLab
Portal & Single Sign on

Kubernetes
Container Orchestration

Prometheus
Live Monitoring

KeyCloak
Authentication, Authorization, Audit

Prerequisites to a production grade system

- Can you deploy models to a secure, scalable & robust production environment?
- Do you have continuous testing, integration & deployment for models?
- Do you have one semantic data models that multiple locations & systems can map to?
- Can you monitor for accuracy & data quality gaps across many locations?
- Can you monitor model decay & concept drift in production?
- Can you use PHI content in your training & modeling environment?
- Can you explain your model’s results to its end users?
- Can you safely reuse models & features across a team?
- Can you regularly updated models for new terminology, guidelines or feedback?

Moral #2:

NLP is ultra domain specific,
so train your own models

The NLP Problem

ED Triage Notes
states started last night, upper abd, took alka seltzer approx 0500, no relief. nausea no vomiting
Since yeatreday 10/10 "constant Tylenol 1 hr ago. +nausea. diaphoretic. Mid abd radiates to back
Generalized abd radiating to lower x 3 days accompanied by dark stools. Now with bloody stool this am. Denies dizzy, sob, fatigue. Visiting from Japan on business.”

Features	
Type of Pain	Symptoms
Intensity of Pain	Onset of symptoms
Body part of region	Attempted home remedy

What makes natural language understanding hard

- Nuanced
- Fuzzy
- Contextual
- Medium specific
- Domain specific

Healthcare specific needs:

1. Core Annotators
   Sentence boundary, part of speech, spell checking, …
2. Vocabulary
   Terminologies, relationships, word embeddings, …
3. ML & DL Models
   Named entity recognition, value extraction, …

NLP for Apache Spark

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

High Performance Natural Language Understanding at Scale

John Snow LABS

Part of Speech Tagger
Named Entity Recognition
Sentiment Analysis
Spell Checker
Tokenizer
Stemmer
Lemmatizer
Entity Extraction

Spark MLlib

Topic Modeling
Word2Vec
TF-IDF
String distance calculation
N-grams calculation
Stop word removal
Train/Test & Cross-Validate
Ensembles

Spark ML API (Pipeline, Transformer, Estimator)

Spark SQL API (DataFrame, Catalyst Optimizer)

Spark Core API (RDD's, Project Tungsten)

Data Sources API

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)
```

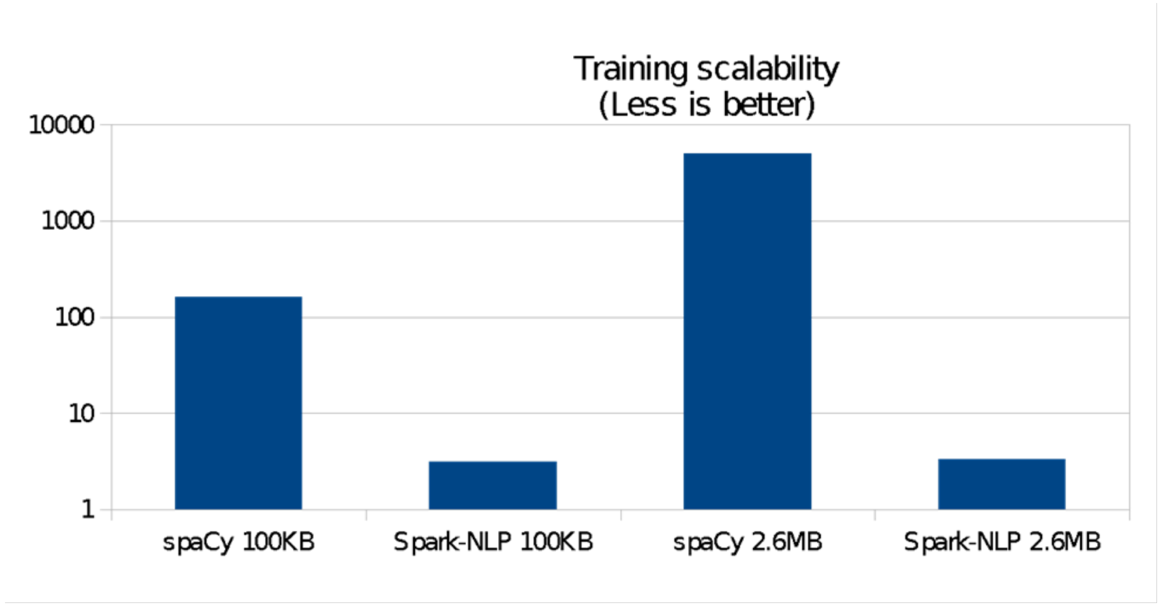
Spark NLP annotators

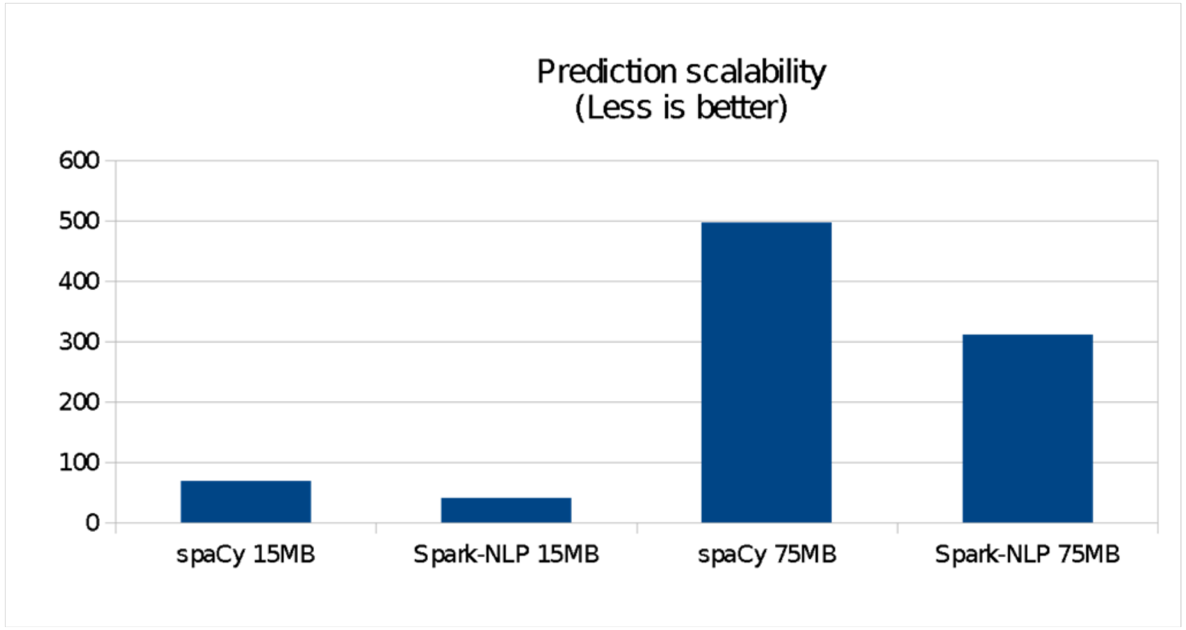
Spark ML featurizers

Spark ML LDA implementation

Single execution plan for the given data frame

NLP for Apache Spark: Performance Benchmarks





- Training was 80x faster to train on 2.6MB
- Training was 38x faster on 100k
- Training on 100k & 2.6MB took roughly the same
- Additional near-linear speedup on a cluster

- Prediction was 1.6x faster on 75MB
- Prediction was 1.4x faster on 15MB
- Adding NLP stages takes roughly the same
- Additional near-linear speedup on a cluster

NLP for Apache Spark: Healthcare Extensions

High Performance Natural Language Understanding at Scale

John Snow LABS

Part of Speech Tagger
Named Entity Recognition
Sentiment Analysis
Spell Checker
Tokenizer
Stemmer
Lemmatizer
Entity Extraction

Spark MLlib

Topic Modeling
Word2Vec
TF-IDF
String distance calculation
N-grams calculation
Stop word removal
Train/Test & Cross-Validate
Ensembles

Spark ML API (Pipeline, Transformer, Estimator)

Spark SQL API (DataFrame, Catalyst Optimizer)

Spark Core API (RDD's, Project Tungsten)

Data Sources API

John Snow LABS

com.johnsnowlabs.nlp.clinical.*

Healthcare specific NLP annotators for Spark in Scala, Java or Python:

- Entity Recognition
- Value Extraction
- Word Embeddings
- Assertion Status
- Sentiment Analysis
- Spell Checking, ...

John Snow LABS

data.johnsnowlabs.com/health

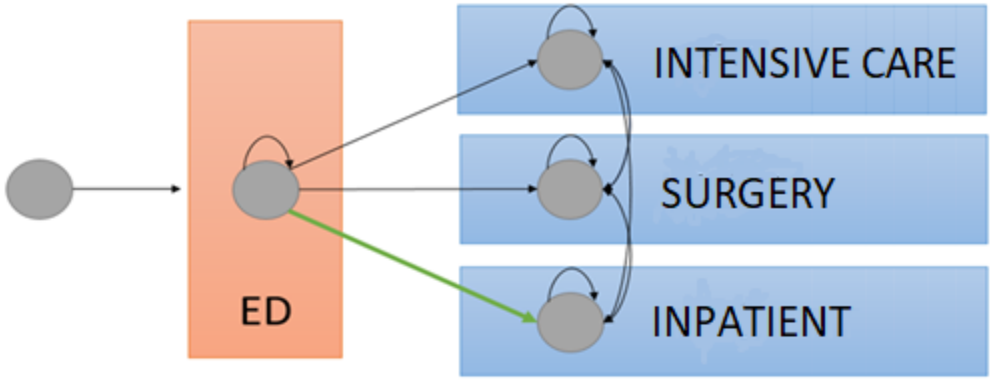
300+ Expert curated, clean, linked, enriched & always up to date data:

- Terminology
- Providers
- Demographics
- Clinical Guidelines
- Genes
- Measures, ...

NLP for Apache Spark: Healthcare Extensions

NLP Library Feature	State of the Art Research
Named Entity Recognition	"Entity Recognition from Clinical Texts via Recurrent Neural Network". Liu et al., *BMC Medical Informatics & Decision Making*, July 2017.
Word Embeddings	"How to Train Good Word Embeddings for Biomedical NLP". Chiu et al., In *Proceedings of BioNLP'16*, August 2016.
Assertion Status Detection	"Improving Classification of Medical Assertions in Clinical Notes". Kim et al., In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

Demand Forecasting of Admission from ED



Features from Structured Data

- How many patients will be admitted today?
- Data Source: EHR data

Reason for visit	Current wait time
Age	Number of orders
Gender	Admit in past 30 days
Vital signs	Type of insurance

Demand Forecasting of Admission from ED

Features from Natural Language Text

- A majority of the rich relevant content lies in unstructured notes that are contributed by doctors and nurses from patient interactions.
- Data Source: Emergency Department Triage notes and other ED notes

Type of Pain	Symptoms
Intensity of Pain	Onset of symptoms
Body part of region	Attempted home remedy

Epilogue:

Why this applies to your challenge, too

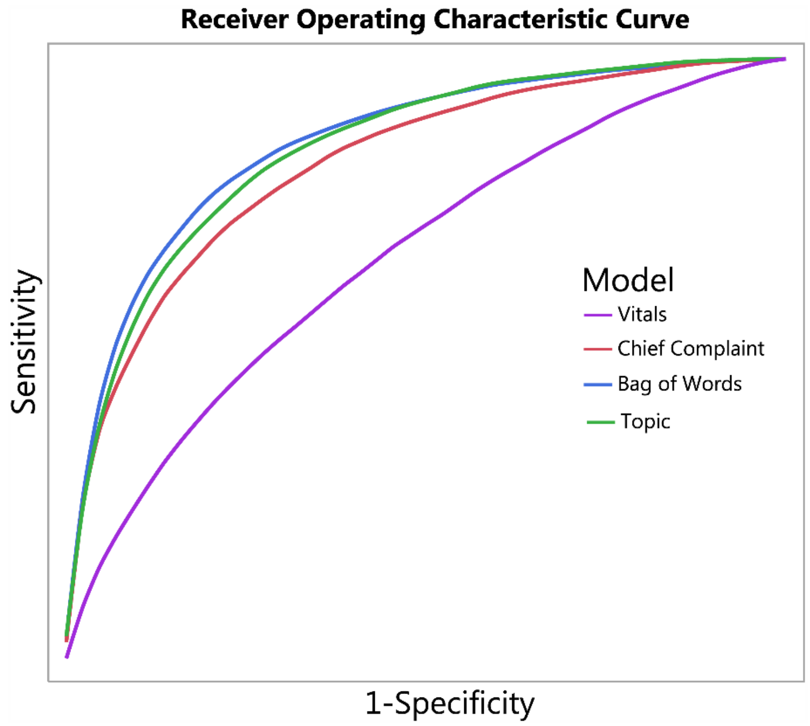
Case Study: Detecting Sepsis

Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning

Steven Horng, David A. Sontag, Yoni Halpern, Yacine Jernite, Nathan I. Shapiro, Larry A. Nathanson

Published: April 6, 2017 • https://doi.org/10.1371/journal.pone.0174708

"Compared to previous work that only used structured data such as vital signs and demographic information, utilizing free text drastically improves the discriminatory ability (increase in AUC from 0.67 to 0.86) of identifying infection."



Case Study: Cohort Selection in Oncology

Opportunities and challenges in leveraging electronic health record data in oncology

Marc L Berger*[,1], Melissa D Curtis[2], Gregory Smith[1], James Harnett[1] & Amy P Abernethy[2]

"Using the combination of structured and unstructured data, 8324 patients were identified as having advanced NSCLC.

Of these patients, only 2472 were also in the cohort generated using structured data only.

Further, 1090 patients would be included in the structured data only cohort who should have been excluded based on additional data."

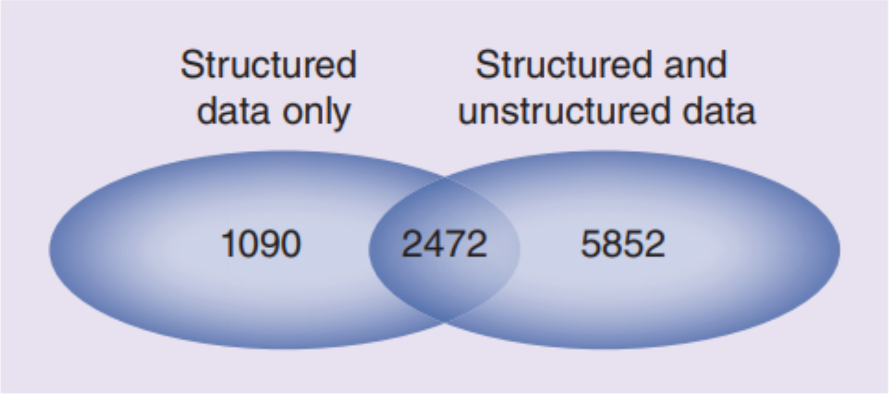


Figure 1. Comparison of patients selected for the analysis using structured data only versus structured and unstructured data.

Q & A

santosh.kulkarni@kp.org

david@pacific.ai