Spark NLP:
How Roche Automates Knowledge Extraction from Pathology Reports

Yogesh Pandit, Vishakha Sharma - Roche
Saif Addin Ellafi - John Snow Labs

Strata DATA CONFERENCE

Disclaimer

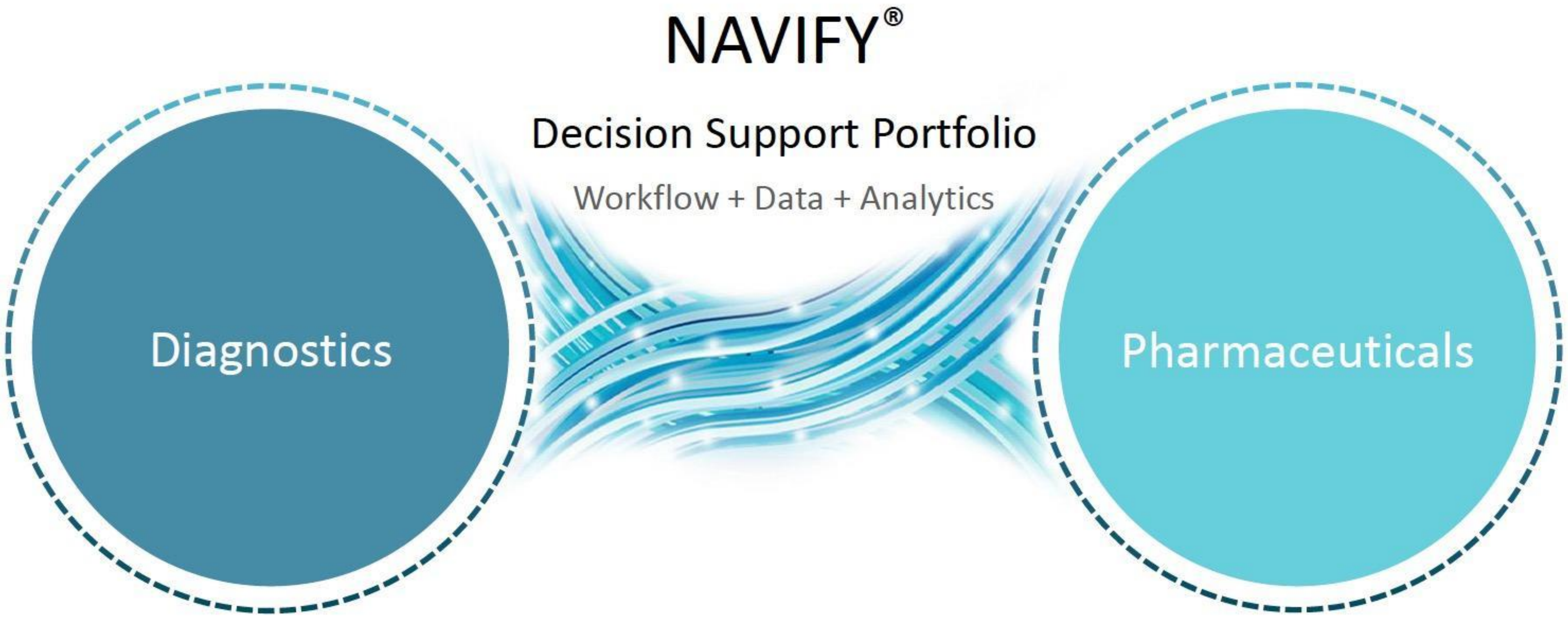
Roche has been one of John Snow Labs' customers since August 2018.

This presentation has been prepared by Roche and John Snow Labs to provide a high-level overview of Roche's use of John Snow Labs' products.

Nothing contained or stated herein or during the presentation constitutes Roche's endorsement of John Snow Labs' products.

John Snow Labs is fully responsible for accuracy and completeness of any statements related to John Snow Labs' products, including the product's performance.

Roche: 120 years of medical innovation



- #1 in biotechnology and *in vitro* diagnostics
- 20+ billion diagnostic tests performed*
- Advanced scientific knowledge and technology that increases the medical value of diagnostic solutions

- Leading provider of cancer treatment worldwide
- 127 million patients treated with Roche medicines*
- Focused on major medical indications and disease areas
- 30 Roche medicines on the WHO Model List of Essential Medicines*

* Roche Annual Report 2018

Unstructured healthcare data challenges for NAVIFY portfolio

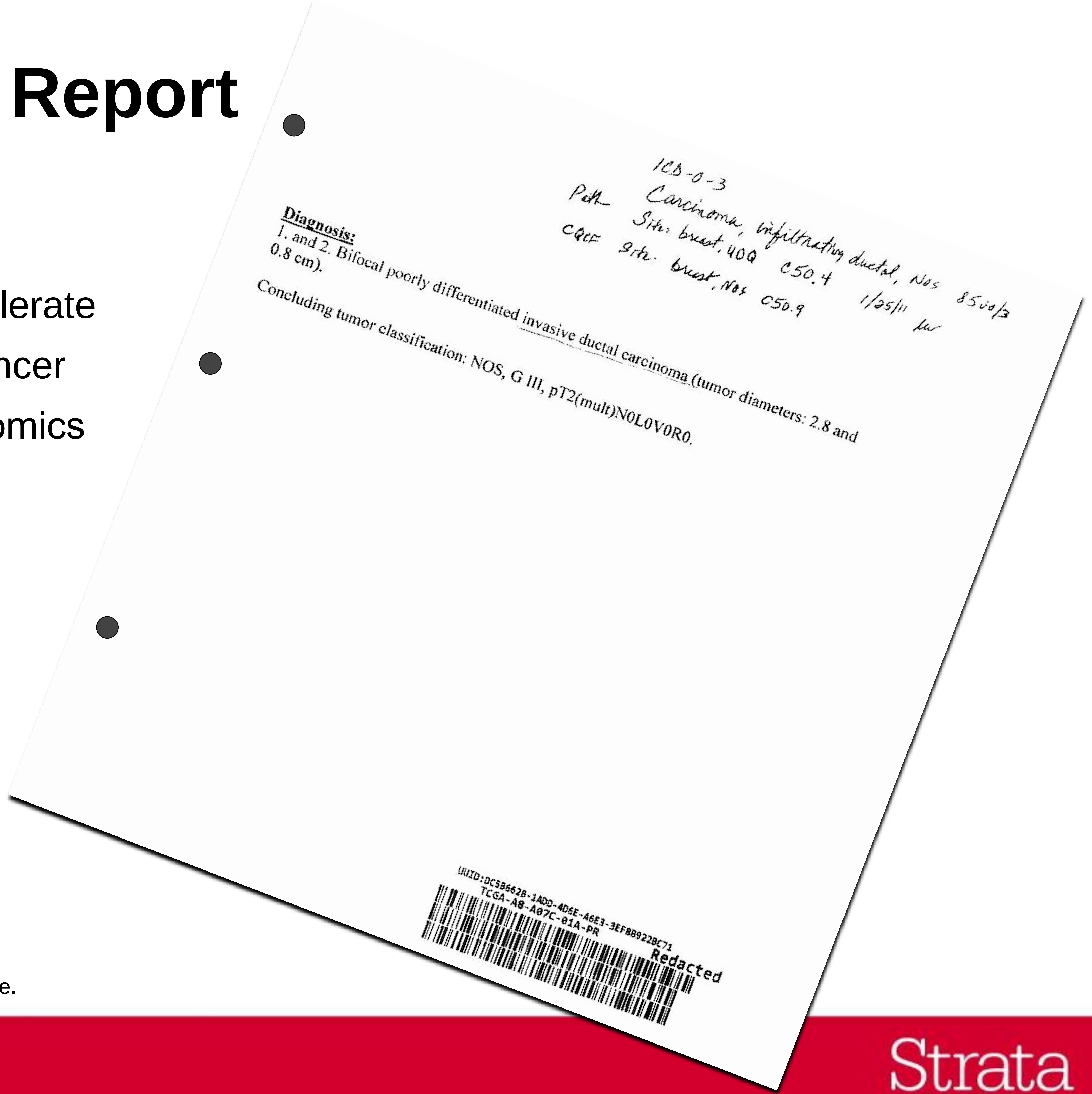
- Diverse customers distributed across the world
- Multiple Languages
- Oncology
- Different report formats (ex: pathology, radiology)
- Different terminologies (ex: SNOMED, LOINC, ICD-O-3)

Must unlock unstructured data to build a comprehensive, longitudinal view of the patient, and enable both clinical decision support and population analytics

*In development

Sample TCGA Pathology Report

- The Cancer Genome Atlas (TCGA)
- A joint effort of the NCI and the NHGRI to accelerate our understanding of the molecular basis of cancer
- Hosted on Amazon S3 and NCI's Cancer Genomics Cloud
- Pathology reports are very diverse:
  - Jargon
  - Tables
  - Key-value pairs
  - Hand-written notes



ICD-0-3

Path Carcinoma, infiltrating ductal, NOS 8500/3

Site: breast, UOQ C50.4

CQCF Site: breast, NOS C50.9 1/25/11

Diagnosis:

1. and 2. Bifocal poorly differentiated invasive ductal carcinoma (tumor diameters: 2.8 and 0.8 cm).

Concluding tumor classification: NOS, G III, pT2(mult)N0L0V0R0.

UUID:DC5B662B-1ADD-4D6E-A6E3-3EF8B922BC71

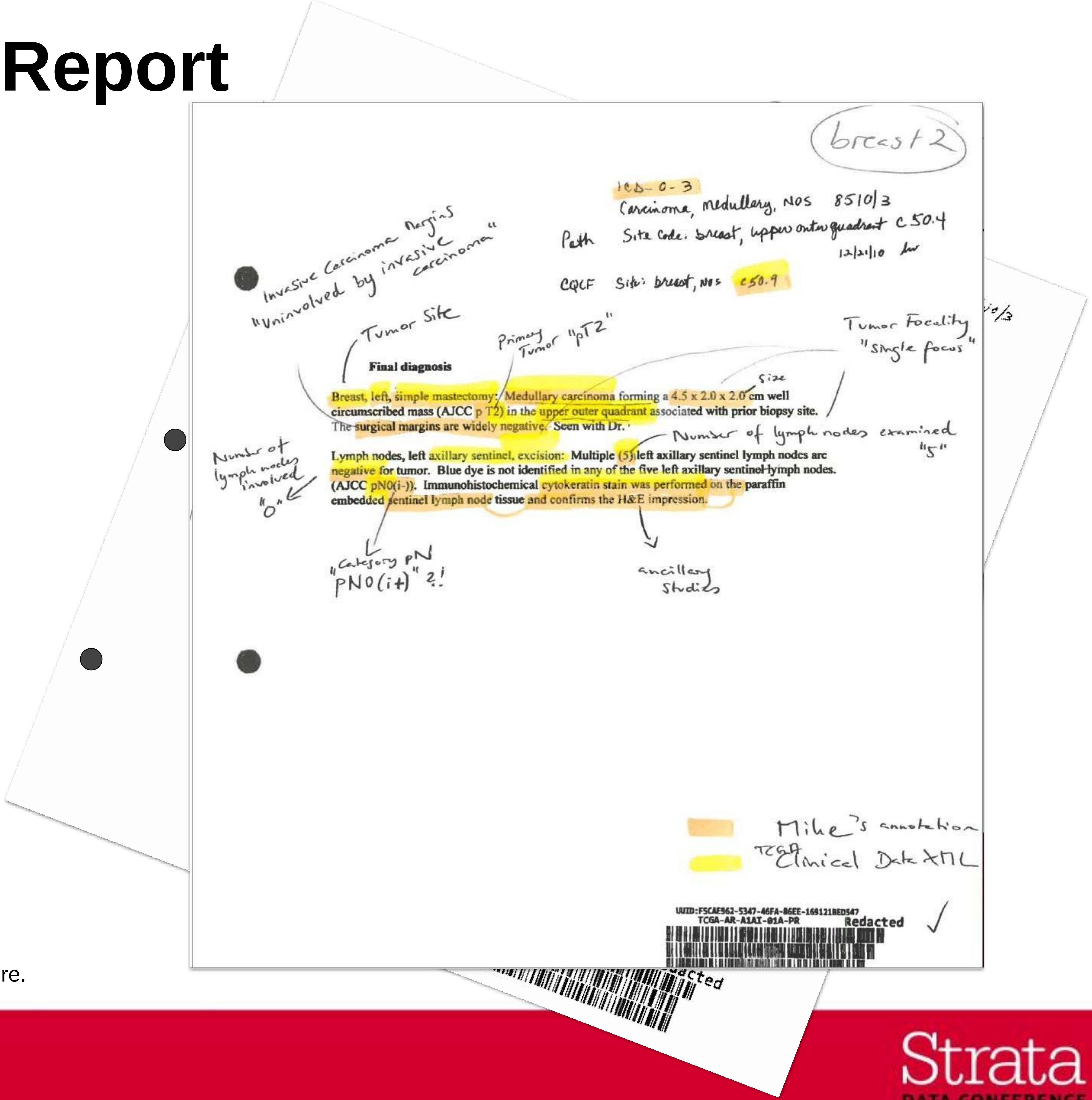
TCGA-A8-A07C-01A-PR

Redacted

Disclaimer: This is sample data from TCGA. There is no real patient data being displayed here.

Manually Curated TCGA Report

Manual curation is extremely time consuming, expensive, and prone to errors


breast 2

ICD-O-3
Carcinoma, medullary, NOS 8510/3
Path Site Code: breast, upper outer quadrant C50.4
12/21/10
CQCF Site: breast, NOS C50.9

Invasive Carcinoma Margins
"Uninvolved by invasive carcinoma"

Tumor Site
Primary Tumor "pT2"
Tumor Focality "single focus"
Size

Final diagnosis

Breast, left, simple mastectomy: Medullary carcinoma forming a 4.5 x 2.0 x 2.0 cm well circumscribed mass (AJCC p T2) in the upper outer quadrant associated with prior biopsy site. The surgical margins are widely negative. Seen with Dr.

Number of lymph nodes examined "5"
Number of lymph nodes involved "0"

Lymph nodes, left axillary sentinel, excision: Multiple (5) left axillary sentinel lymph nodes are negative for tumor. Blue dye is not identified in any of the five left axillary sentinel lymph nodes. (AJCC pN0(i-)). Immunohistochemical cytokeratin stain was performed on the paraffin embedded sentinel lymph node tissue and confirms the H&E impression.

"Category pN "pN0(i+)" ?!
ancillary studies

Mike's annotation
TCGA Clinical Data XML

UUID:F5CAE962-5347-46FA-B6EE-169121BED547
TCGA-AR-A1AI-01A-PR
Redacted

Disclaimer: This is sample data from TCGA. There is no real patient data being displayed here.

Sample Results from Curation

Token (Word)	Entity Category (Labels)
Upper outer quadrant	Tumor Site (Localization1)
pT1	Primary Tumor (pT1)
pN1	Regional Lymph Nodes (pN1)
pM1	Distant Metastasis (pM1)
Left	Specimen Laterality (Laterality1)
4.5 x 2.0 x 2.0 cm	Size of Invasive Carcinoma (Size1)
Medullary Carcinoma	Histologic Type (Type1)
Poorly Differentiated	Histologic Grade (Grade1)

8 Entities and 73 Unique Labels

The NAVIFY team identified two significant needs

Natural Language Processing (NLP):

- High accuracy
- Specialized for medical data
- Minimize time to train new models
- Extensible for new content types

Optical Character Recognition (OCR):

- High accuracy
- Retain document structure

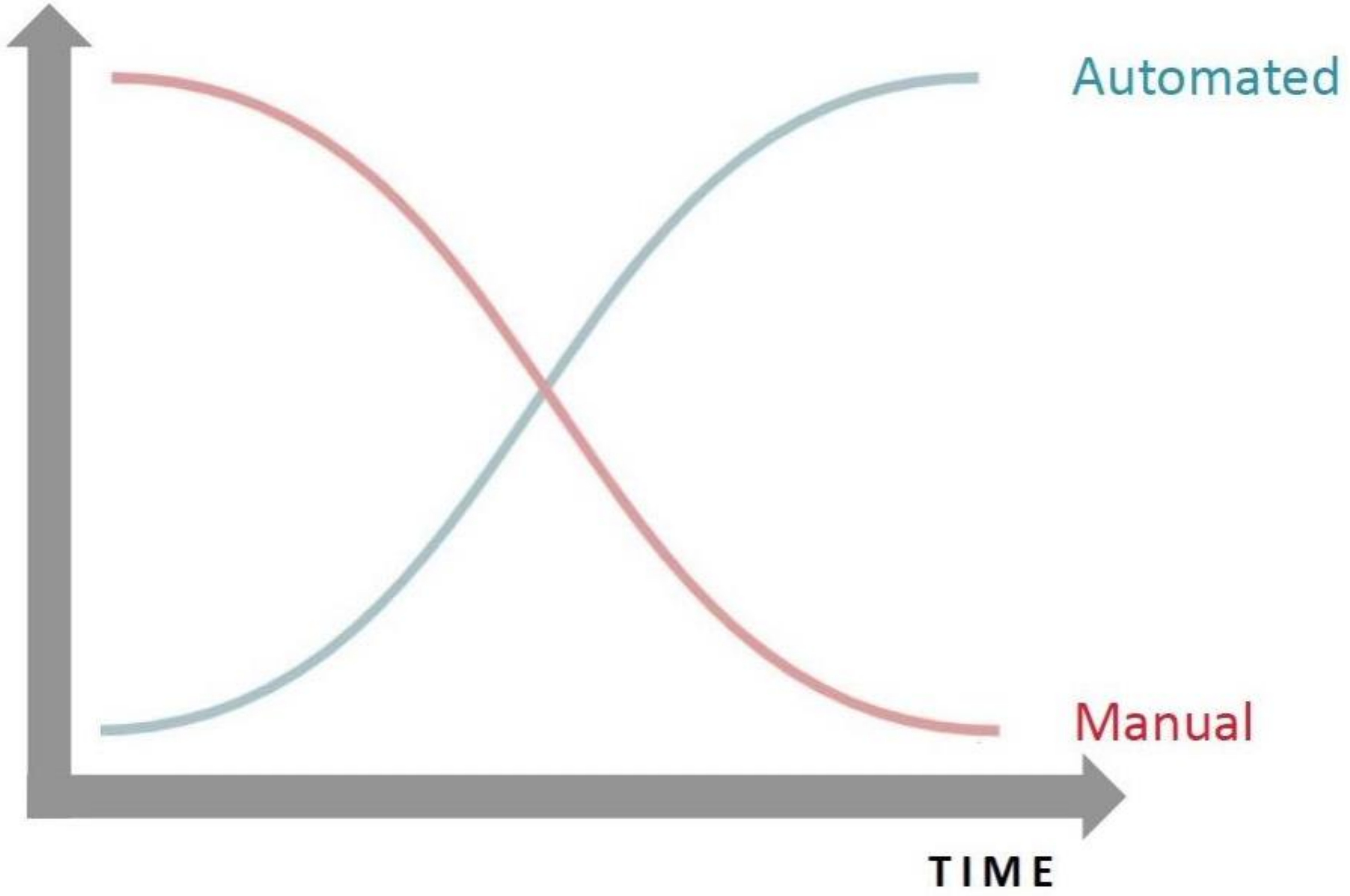
  (i.e. tables, lists, paragraphs,...)

Requirements for both:

- Scalable (support 10 million pathology reports per year)
- Compliant with privacy laws
- Integrates easily with AWS services
- Low cost

The use of NLP will be a journey

- Initial goal of speeding up review of pathology reports
- Will then automate extraction of high confidence entities and relationships (low hanging fruit)
- Will keep increasing automation of NLP over time



Introducing Spark NLP

- State of the art NLP for Python and Scala
  - Performance
  - Accuracy
  - Scale
- Apache 2.0 licensed
- Active development and community: 25 releases in 2018
- Spark and TensorFlow under the hood
- Healthcare specific extensions

John Snow Lab's Spark NLP for Healthcare

High Performance Natural Language Understanding at Scale

John Snow LABS

johnsnowlabs.com/spark-nlp-health

Healthcare-specific NLP models available in Scala, Java or Python:

- Entity Recognition
- Entity Resolution
- Assertion Status
- Spell Checking
- Embeddings
- POS, OCR, …

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

John Snow LABS

johnsnowlabs.com/data

2,000+ Expert curated, clean, linked, enriched & always up to date datasets:

- Terminologies
- Benchmarks
- Providers
- Drugs & Devices
- Clinical Guidelines
- Genes, Measures, …

Spark ML API (Pipeline, Transformer, Estimator)

Spark SQL API (DataFrame, Catalyst Optimizer)

Spark Core API (RDD's, Project Tungsten)

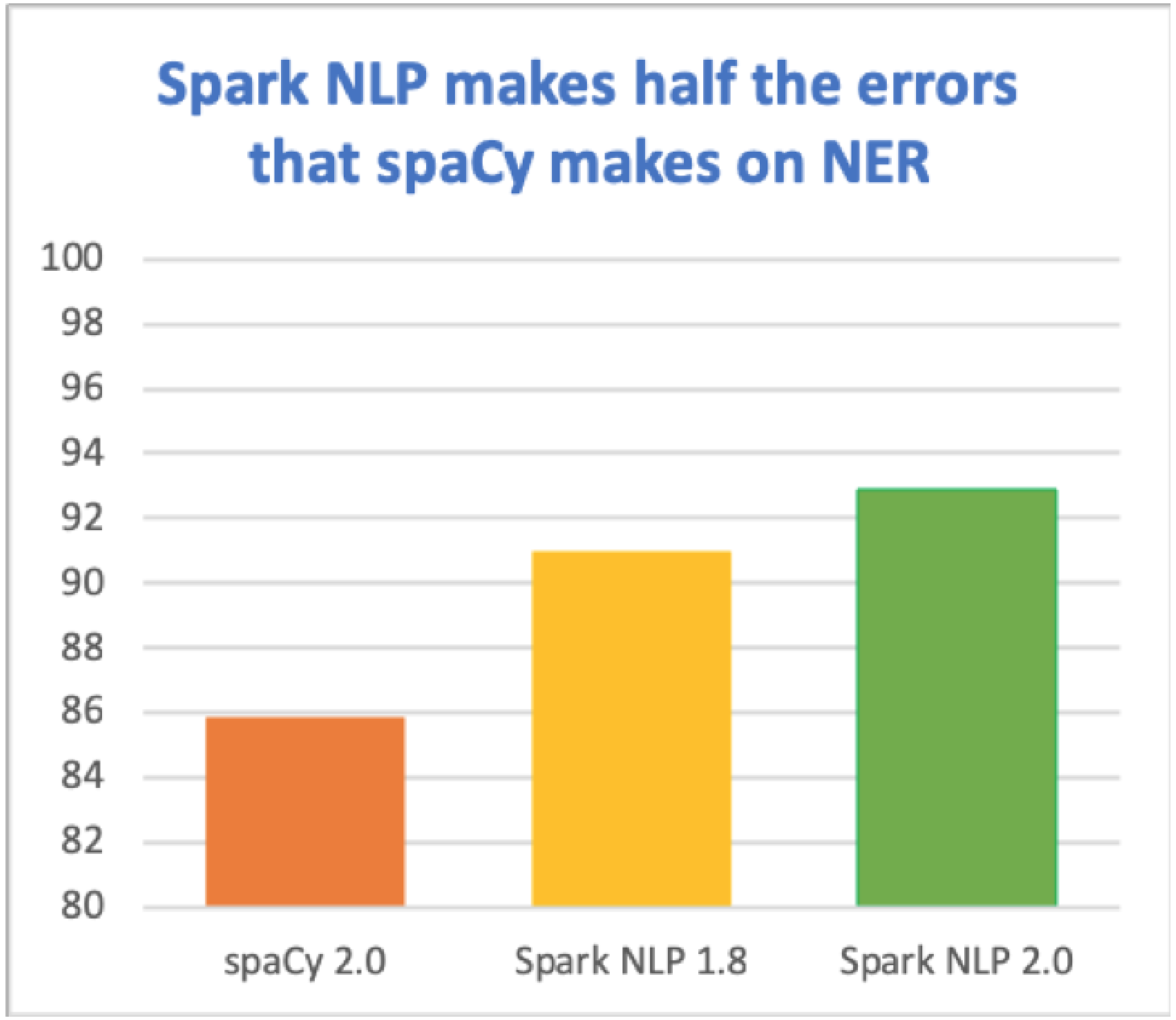
Data Sources API

Benchmarks: Accuracy

- "State of the art" means the best performing peer-reviewed results

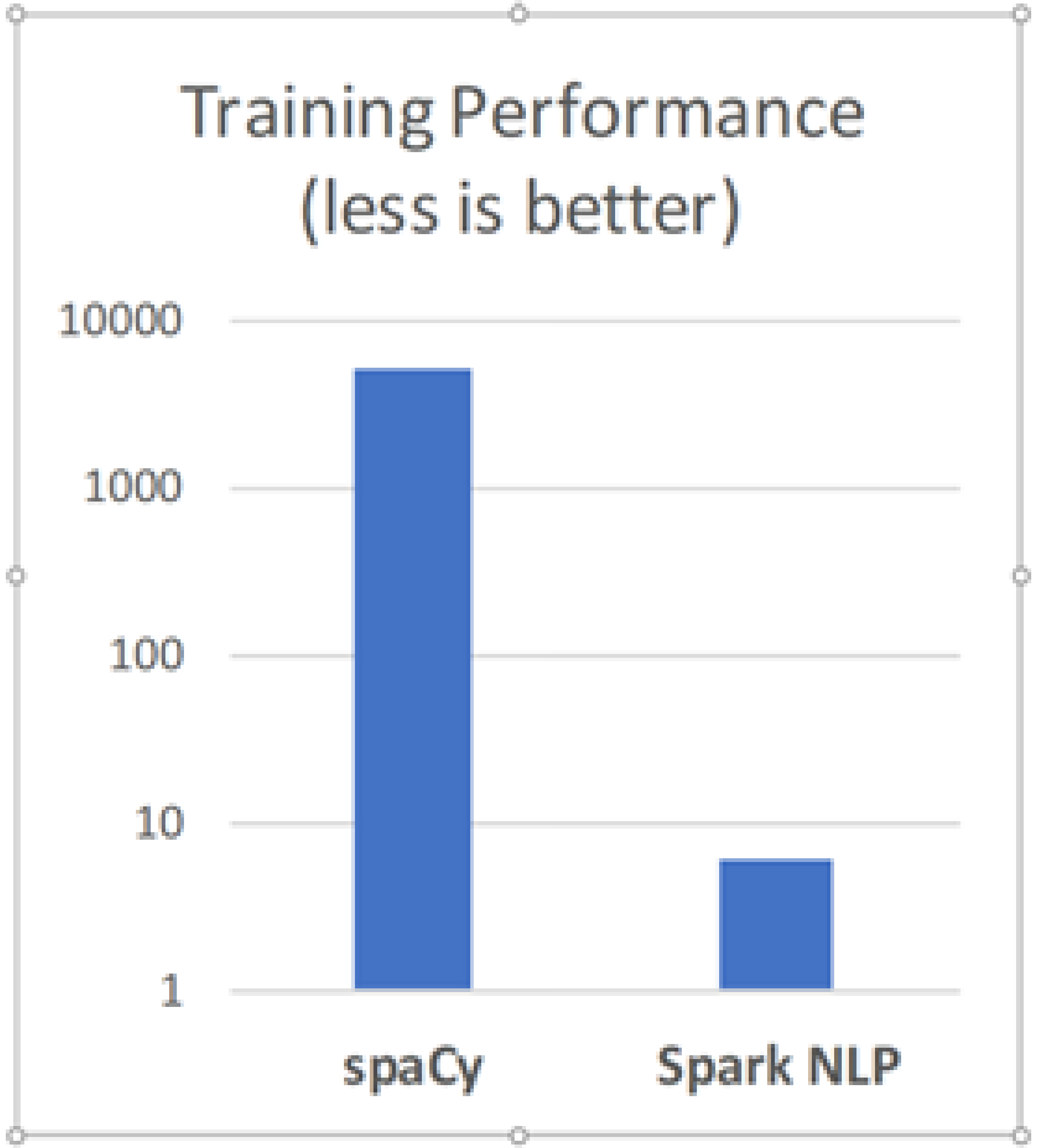
- Example on named entity recognition:
  - Deep Learning TF graph based on 2017 paper (bi-LSTM+CNN+CRF)
  - Trainable at scale with GPU's
  - BERT embeddings
  - Contrib LSTM cells

- This benchmark is on en_core_web_lg dataset, micro-averaged F1 score



Benchmark: Speed

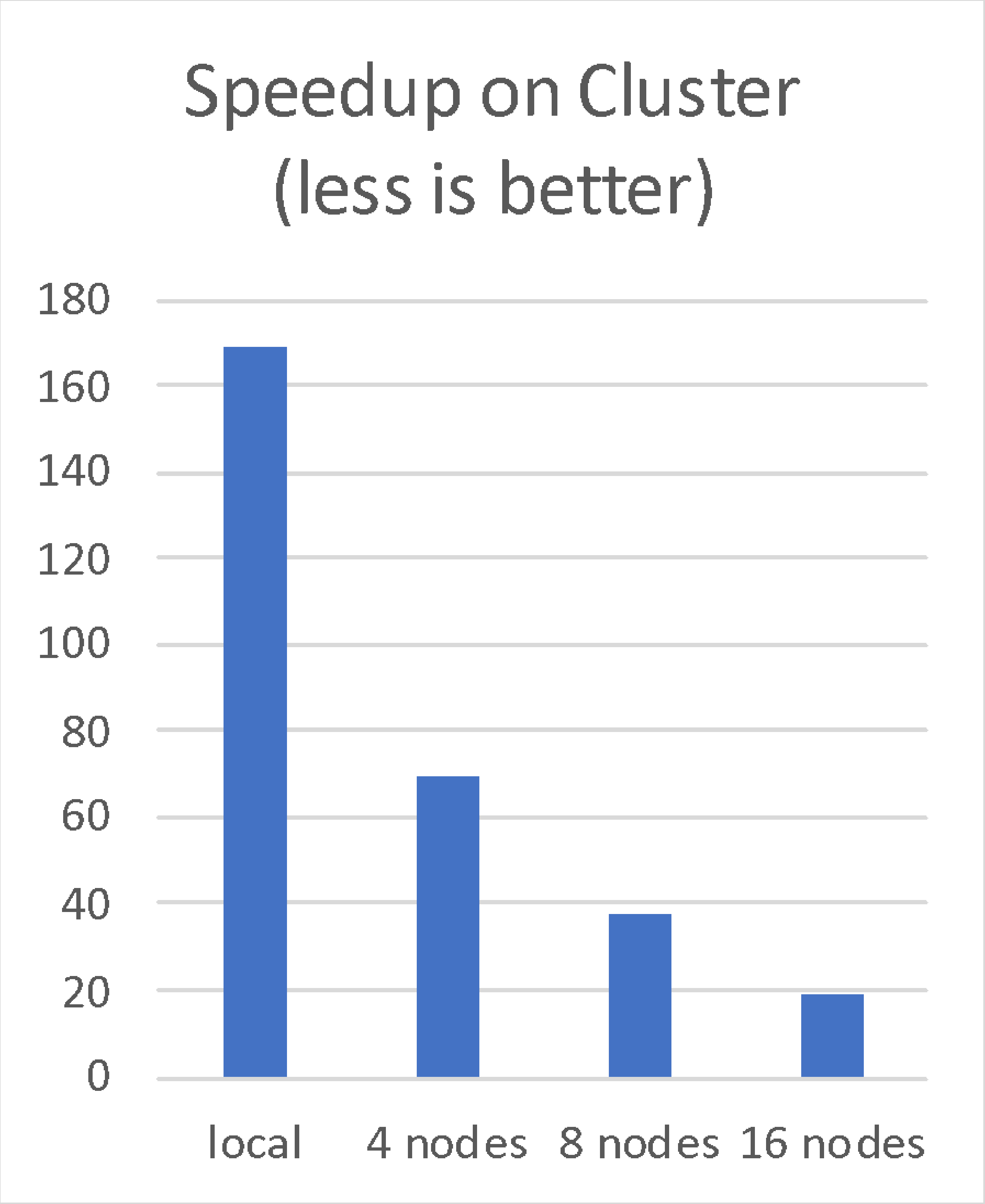
- Spark NLP was 80 times faster than spaCy to train a basic model on a single machine
- At the same level of accuracy
- Public benchmark was run on one Intel i5-6600k machine 4-cores, 16GB, SSD



Public benchmark - https://www.oreilly.com/ideas/comparing-production-grade-nlp-libraries-accuracy-performance-and-scalability

Benchmarks: Scaling

- 2.5x speedup on a 4-node EMR cluster compared to local execution
- **Zero code changes**
- Spark scales as Spark does: **1 to 3 orders of magnitude faster** depending on cluster
- Not compared with spaCy or CoreNLP, because they can't leverage a cluster



Public benchmark - https://www.oreilly.com/ideas/comparing-production-grade-nlp-libraries-accuracy-performance-and-scalability

Spark NLP and Deep Learning

- **TensorFlow under the hood**
  - No need to learn or manage it
  - Pretrained models and graphs
  - Training and inference on GPUs
- **State of the art models and networks**
  - Named entity recognition
  - Entity resolution/normalization
  - Negation detection
  - Spell checking and correction
  - Sentence boundary detection
  - Embeddings

“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”.

Devlin et. al. (Google Research), October 2018.

Now natively available within Spark NLP

Back to the Task: Annotators

Optical Character Recognition (OCR)
The starting point is (de-identified) hospital reports

PDFs

Diagnosis:
Right-sided hemicolectomy preparation shows tumor-free oral and aboral resection margins and includes an ulcerated, moderately differentiated adenocarcinoma of the transverse colon with infiltration of the perimuscular fatty tissue (G2, pT3).

OCR Output

```
Diagnosis:
Right-sided hemicolectomy preparation shows tumor-free oral and aboral resection margins and
includes an ulcerated, moderately differentiated adenocarcinoma of the transverse colon with
infiltration of the perimuscular fatty tissue (G2, pT3).

Follow-up report:
```

Identified 17 OCR Parameters

Run ID	Name	Source Type	Source Name	User	Status	engine_mode	erosion	fallback_method	page_iterator_le	page_seg_mode	preferred_metho	scaling_factor	avg_bacc	avg_ber	avg_cer	avg_wacc	avg_wer
54193a95b3bf4d3f91d0f80b65c90		LOCAL	/usr/local/lib/pyth	saif	FINISHED	3	FALSE	FALSE	0	3	text	1	0.8615249984	0.1384749997	0.2336500005	0.3683749985	0.631624997
96761f5f79aa457dbeb8841d9047		LOCAL	/usr/local/lib/pyth	saif	FINISHED	6	TRUE	FALSE	1	6	image	1.1	0.6796800017	0.3203199987	0.4961600012	0.7526800025	0.247319997
a2f11c067b4c4de787761d626730		LOCAL	/usr/local/lib/pyth	saif	FINISHED	6	FALSE	FALSE	1	6	image	1.1	0.7907599926	0.2092400005	0.327160003	0.4554400003	0.5445599973
1b1aeb7789d44202a165db0c793		LOCAL	/usr/local/lib/pyth	saif	FINISHED	6	TRUE	FALSE	1	6	image	1	0.3783749971	0.6216249992	0.7238124944	0.916375	0.08362499869
5ebbf079f10840a9aba6f0e5a383b		LOCAL	/usr/local/lib/pyth	saif	FINISHED	6	FALSE	FALSE	1	6	image	1	0.788599999	0.2114000005	0.335839998	0.47708	0.522920002
385a1cd85aa24042ae65c164667		LOCAL	/usr/local/lib/pyth	saif	FINISHED	6	TRUE	FALSE	0	6	image	1.1	0.7119200015	0.288080001	0.4616800019	0.7364400005	0.2635599994
27ec0ddc7b5b421d998ce0b6d0e		LOCAL	/usr/local/lib/pyth	saif	FINISHED	6	FALSE	FALSE	0	6	image	1.1	0.7996400023	0.2003599986	0.3191200018	0.4430799943	0.5569199973
fb2c88edeeef439d8cc726bacb668		LOCAL	/usr/local/lib/pyth	saif	FINISHED	6	TRUE	FALSE	0	6	image	1	0.3758749997	0.6241249992	0.7193124983	0.9120625071	0.08793749975
b6a56dc0cd9242d4a0c4067430e		LOCAL	/usr/local/lib/pyth	saif	FINISHED	6	FALSE	FALSE	0	6	image	1	0.8027600002	0.1972400004	0.3256799983	0.4613599974	0.5386400017
5167a6f9a2964c928b37aef0fbced		LOCAL	/usr/local/lib/pyth	saif	FINISHED	3	TRUE	FALSE	1	3	image	1.1	0.6796800017	0.3203199987	0.4961600012	0.7526800025	0.247319997
cab9ed4610bc4a3bb9c0fe01aad7		LOCAL	/usr/local/lib/pyth	saif	FINISHED	3	FALSE	FALSE	1	3	image	1.1	0.7907599926	0.2092400005	0.327160003	0.4554400003	0.5445599973
fe6a030f1c194469a47533207f099		LOCAL	/usr/local/lib/pyth	saif	FINISHED	3	TRUE	FALSE	1	3	image	1	0.3783749971	0.6216249992	0.7238124944	0.916375	0.08362499869
b977be6ac8e44b4780587e6e38a		LOCAL	/usr/local/lib/pyth	saif	FINISHED	3	FALSE	FALSE	1	3	image	1	0.788599999	0.2114000005	0.335839998	0.47708	0.522920002
a4858c6850f845008915df4e7c05		LOCAL	/usr/local/lib/pyth	saif	FINISHED	3	TRUE	FALSE	0	3	image	1.1	0.7119200015	0.288080001	0.4616800019	0.7364400005	0.2635599994
413831327ef24104ac4a60cec2ad		LOCAL	/usr/local/lib/pyth	saif	FINISHED	3	FALSE	FALSE	0	3	image	1.1	0.7996400023	0.2003599986	0.3191200018	0.4430799943	0.5569199973
8363cc631604478f9e9d9d5614a4		LOCAL	/usr/local/lib/pyth	saif	FINISHED	3	TRUE	FALSE	0	3	image	1	0.3758749997	0.6241249992	0.7193124983	0.9120625071	0.08793749975
2cf02e665b34460daea0eae5c9ac		LOCAL	/usr/local/lib/pyth	saif	FINISHED	3	FALSE	FALSE	0	3	image	1	0.8027600002	0.1972400004	0.3256799983	0.4613599974	0.5386400017

Disclaimer: This is sample data from TCGA. There is no real patient data being displayed here.

Sentence Boundary Detection

Extracting complete sentences from messy, multi-page documents
As a means for text correction, sentence bounds organize text

```
'Activation of the CD28 surface receptor provides a major costimulatory signal for T cell activation resulting in enhanced
production of interleukin-2 (IL-2) and cell proliferation. In primary T lymphocytes we show that CD28 ligation leads to th
e rapid intracellular formation of reactive oxygen intermediates (ROIs) which are required for CD28-mediated activation of
the NF-kappa B/CD28-responsive complex and IL-2 expression. Delineation of the CD28 signaling cascade was found to involve
protein tyrosine kinase activity, followed by the activation of phospholipase A2 and 5-lipoxygenase. Our data suggest that
lipoxygenase metabolites activate ROI formation which then induce IL-2 expression via NF-kappa B activation. These finding
s should be useful                                                                          the CD28 costimulatory p
athway.'
```

```
Activation of the CD28 surface receptor provides a major costimulatory signal for T cell activation resulting in enhanced
production of interleukin-2 (IL-2) and cell proliferation.

In primary T lymphocytes we show that CD28 ligation leads to the rapid intracellular formation of reactive oxygen intermed
iates (ROIs) which are required for CD28-mediated activation of the NF-kappa B/CD28-responsive complex and IL-2 expressio
n.

Delineation of the CD28 signaling cascade was found to involve protein tyrosine kinase activity, followed by the activatio
n of phospholipase A2 and 5-lipoxygenase.

Our data suggest that lipoxygenase metabolites activate ROI formation which then induce IL-2 expression via NF-kappa B act
ivation.

These findings should be useful for therapeutic strategies and the development of immunosuppressants targeting the CD28 co
stimulatory pathway.
```

```
Precision: 0.8776
Recall: 0.9258
F1: 0.9011
Accuracy: 0.82
```

Disclaimer: This is sample data from GENIA dataset - www.geniaproject.org. There is no real patient data being displayed here.

Tokenization

Tokenization is the point of analysis for every other annotator algorithm

```
'Among the four non-responders who were NS positive during IFN three were NC positive before IFN'
```

```
[('Among', 'II'),
 ('the', 'DD'),
 ('four', 'MC'),
 ('non-responders', 'NNS'),
 ('who', 'PNR'),
 ('were', 'VBD'),
 ('NS', 'NN'),
 ('positive', 'JJ'),
 ('during', 'II'),
 ('IFN', 'NN'),
 ('three', 'MC'),
 ('were', 'VBD'),
 ('NC', 'NN'),
 ('positive', 'JJ'),
 ('before', 'II'),
 ('IFN', 'NN')]
```

Clinical Part of Speech (POS) Tagging

Assign a reference to the role each token has in a sentence

***Clinical POS is different** from "General English" POS - this impacts accuracy!*

```
token mismatch:
predicted:
Patients were placed into three different groups 1 patients ever treated with large pool non-hepatitis C virus HCV - safe
concentrate n 179 2 patients treated with cryoprecipitate n 125 and 3 patients treated exclusively with HCV-save concentra
te n 12
vs
Patients were placed into three different groups 1 patients ever treated with large pool non-hepatitis C virus HCV -safe c
oncentrate n 179 2 patients treated with cryoprecipitate n 125 and 3 patients treated exclusively with HCV-save concentrat
e n 12
```

```
Token and POS matching accuracy against MEDPOST dataset
Token precision: 0.978925. Matched 4645 tokens out of 4745
POS precision: 0.995910. Matched 4626 POS labels out of 4645 matching tokens
Overall POS precision: 0.974921. Matched 4626 POS labels out of 4745 tokens
```

MEDPOST: https://www.ncbi.nlm.nih.gov/pubmed/15073016

Named Entity Recognition (NER)

Spark NLP provides both CRF and CNN+Bi-LSTM implementations

We trained a model to extract 45+ labels from TCGA reports

```
|    I-Bronchial|
|   I-DcisMargin|
|   I-Diagnosis3|
|       I-Distal|
|     I-Examined|
|    I-Examined1|
|    I-Extension|
|   I-Extension1|
|     I-Focality|
|        I-Grade|
|       I-Grade1|
|   I-Laterality|
|  I-Laterality1|
| I-Localization|
|I-Localization1|
|I-Localization2|
|I-Localization3|
|      I-Margins|
|     I-Margins1|
```

```
|      I-Nuclear|
|     I-Nuclear1|
|  I-OtherMargin|
|  I-Parenchymal|
|     I-Positive|
|    I-Positive1|
|    I-Procedure|
|   I-Procedure1|
|     I-Proximal|
|       I-Radial|
|      I-Results|
|     I-Results1|
|         I-Size|
|        I-Size1|
|        I-Size2|
|        I-Size3|
```

```
|        I-Tests|
|       I-Tests1|
|         I-Type|
|        I-Type1|
|     I-Vascular|
|           I-pM|
|          I-pM1|
|           I-pN|
|          I-pN1|
|           I-pT|
|          I-pT1|
|              O|
+---------------+
```

Entity Resolution

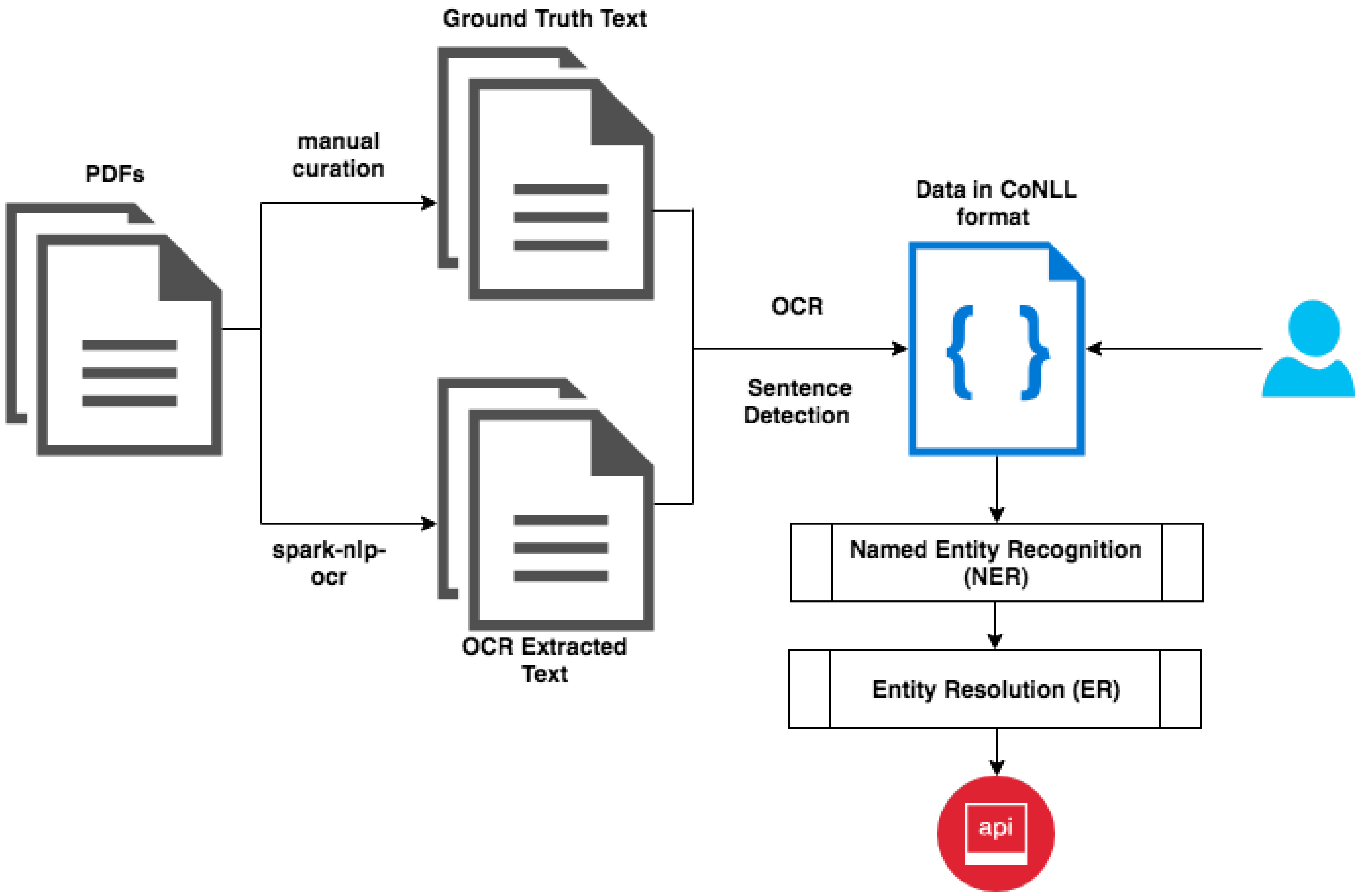
With just NER - we can not resolve entities to structured code

Pre-trained models for resolving healthcare entities to standard SNOMED & ICD-10 codes

codes	description
17473003	Cecotomy
17473003	Cecotomy (procedure)
304587000	Excision of colonic pouch
304587000	Excision of colonic pouch (procedure)
87279008	Excision of lesion of colon
174117007	Excision of lesion of colon NEC
174117007	Excision of lesion of colon NEC (procedure)
87279008	Excision of lesion of colon (procedure)
276190007	Ileocolic resection
276190007	Ileocolic resection (procedure)
43075005	Partial resection of colon
43075005	Partial resection of colon (procedure)
428305005	History of partial resection of colon (situation)
428305005	History of partial resection of colon
444165004	Partial resection of colon and resection of terminal ileum with ileocolic anastomosis
738552004	Partial resection of colon with stoma (procedure)
738552004	Partial resection of colon with stoma
84952009	Resection of colon for interposition
84952009	Resection of colon for interposition (procedure)
445884009	Wedge resection of colon

only showing top 20 rows

NLP Pipeline to Generate NER Training Set



Lesson Learned

- Extracting text from **domain specific PDFs/images** is unpredictable
- **Quantitative evaluation of OCR** is challenging
- Bridging the gap between **domain knowledge & NLP** requires consensus
- Evidence does not always match with **standard terminologies**
- **Building NLP pipelines - that are generalizable:**
  - Static components like tokenization, sentence detection, POS tagging and chunking **can be re-utilized**
  - Data sources (hospitals) differ, NLP approach needs to be plug and play

Thank You!!!

Yogesh Pandit
Senior Software Engineer
Roche Diagnostic Information Solutions (DIS)
yogesh.pandit@roche.com

Vishakha Sharma
Data Scientist
Roche Diagnostic Information Solutions (DIS)
vishakha.sharma@roche.com

Saif Addin Ellafi
Software Engineer
John Snow Labs
saif@johnsnowlabs.com

We are hiring!
https://www.navify.com/careers/

Try Spark NLP at
nlp.johnsnowlabs.com
or visit booth P16