DEEP 6 AI

John Snow LABS

Feature engineering with Spark NLP to accelerate clinical trial recruitment

Strata Data Conference,
New York City,
September 2019

Scott Hoch
Saif Addin Ellafi

Deep 6 AI at a glance

BEAT IBM, HP AUTONOMY IN USG CONTEST, LEADING TO $2M IN CONTRACTS

DisruptorDaily

TOP 100 MOST DISRUPTIVE COMPANIES IN THE WORLD

SXSW

WIN AT SXSW 2017 ACCELERATOR: ENTERPRISE + SMART DATA

CEDARS-SINAI

“DEEP 6 AI IS A GAME-CHANGER.”

Deep6 AI uses cutting edge data and engineering techniques to find **more, better-matching patients** for clinical trial **in minutes**, not months.

John Snow LABS DEEP 6 AI

Deep 6 AI at a glance

BEAT IBM, HP AUTONOMY IN USG CONTEST, LEADING TO $2M IN CONTRACTS

DisruptorDaily

TOP 100 MOST DISRUPTIVE COMPANIES IN THE WORLD

SXSW

WIN AT SXSW 2017 ACCELERATOR: ENTERPRISE + SMART DATA

CEDARS-SINAI

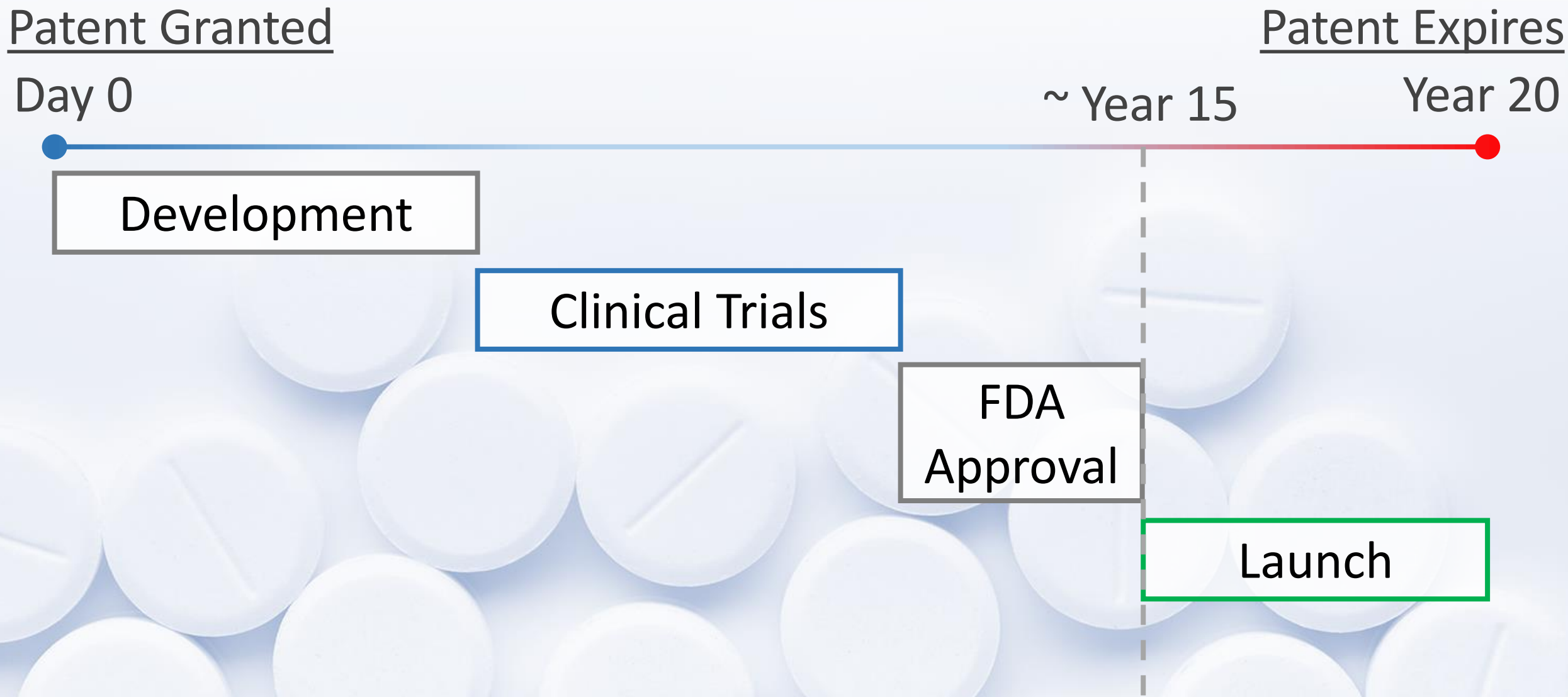
“DEEP 6 AI IS A GAME-CHANGER.”

Deep6 AI uses cutting edge data and engineering techniques to find **more, better-matching patients** for clinical trial **in minutes**, not months.

- Why focus on clinical trials
- How natural language processing can help
- Examples at scale

John Snow LABS DEEP 6 AI

Development of a new treatment

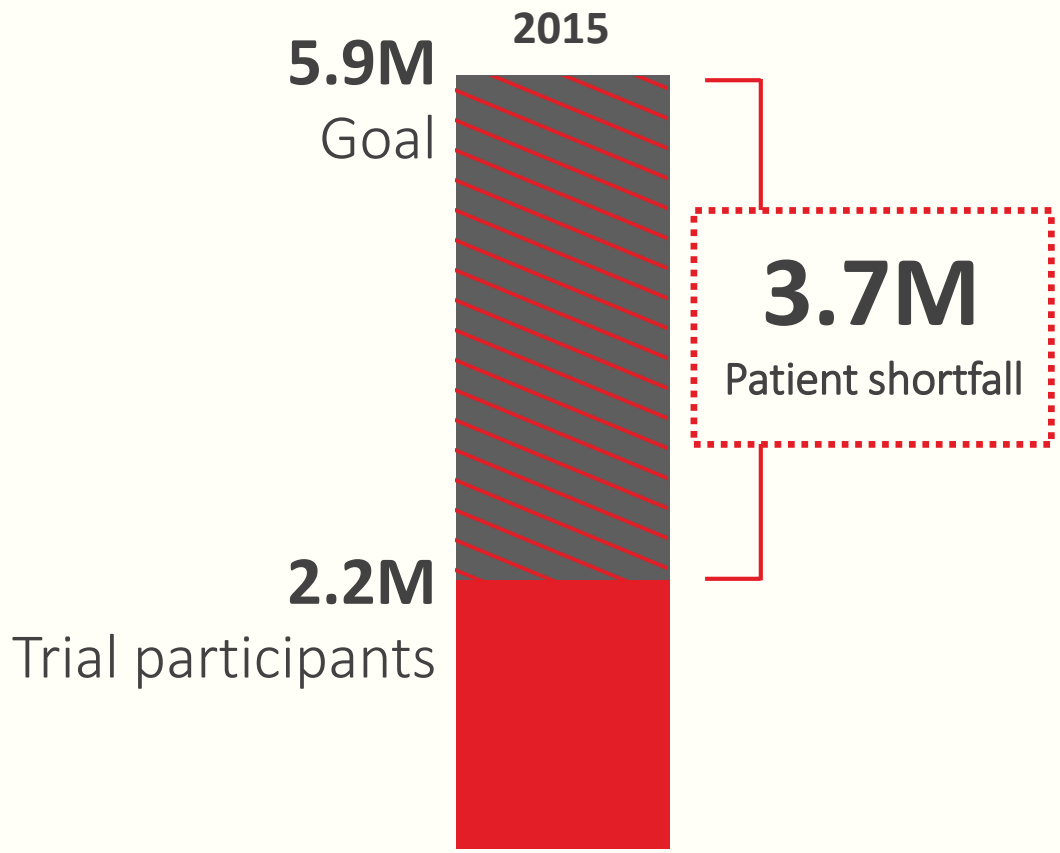


- Your treatments are > 15 years old

- Cutting edge treatments only available in clinical trials

- Faster cycles make lifesaving treatments available sooner

Too few people participate in clinical trials



SOURCES: clinicaltrials.gov, CISCRP

Trial recruitment is hard

86% of trials **suffer delays**

10.8 average months of **delay** per trial

50% of trial sites **fail**

SOURCE: CenterWatch

John Snow LABS DEEP 6 AI

Too few people participate in clinical trials

SOURCES: clinicaltrials.gov, CISCRP

Trial recruitment is hard

86% of trials **suffer delays**

10.8 average months of **delay** per trial

50% of trial sites **fail**

SOURCE: CenterWatch

The consequences are enormous

1-year delay

993 **Human lives lost**
on 1 blockbuster drug

2.9B **Revenue lost**
on 1 blockbuster drug

5x **Drug price increase**
<> Wage increase, 2017

SOURCE: Cutting Edge Information, FiercePharma, Life Extension Magazine

DEEP 6 AI

5,446
cutting-edge treatments are behind the gates of clinical trials

SOURCE: Estimated based on Pharmaprojects, "Pharma R&D Annual Review 2017"

John Snow LABS DEEP 6 AI

Why is recruitment hard (data)

- Indexing patient data
- Generating search criteria
- Find a match

John Snow LABS DEEP 6 AI

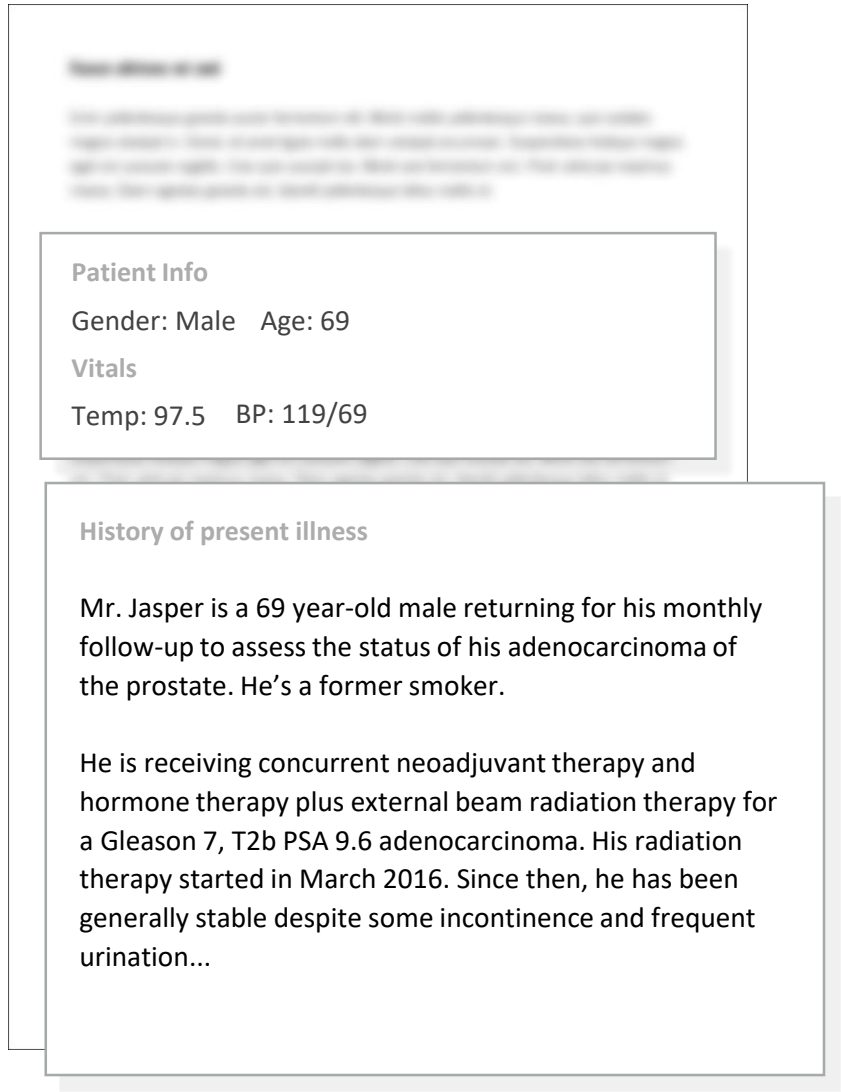
Why is recruitment hard (data)

- Indexing patient data
- Generating search criteria
- Find a match

Most information is in doctor's notes

Structured, relational data
(easy to reason about)

Free text data
- Non-standard
  - grammar
  - format
  - vocabulary



John Snow LABS DEEP 6 AI

Why is recruitment hard (data)

- Indexing patient data
- Generating search criteria
- Find a match

MIMIC “Sentence”

"(admission): 50.4 kg\\n Height: 61 Inch\\n ICP: 7 (1 - 14) mmHg\\n Total In:\\n 3,279 mL\\n 911 mL\\n PO:\\n Tube feeding:\\n 243 mL\\n 237 mL\\n IV Fluid:\\n 2,827 mL\\n 624 mL\\n Blood products:\\n Total out:\\n 2,333 mL\\n 370 mL\\n Urine:\\n 2,330 mL\\n 370 mL\\n NG:\\n Stool:\\n Drains:\\n 3 mL\\n Balance:\\n 946 mL\\n 541 mL\\n Respiratory support\\n O2 Delivery Device: None\\n SPO2: 97%\\n ABG: ///26/\\n Physical Examination\\n General Appearance: No acute distress, Non communicative due to\\n language barrier\\n HEENT: PERRL, EOMI\\n Cardiovascular: (Rhythm: Regular)\\n Respiratory / Chest: (Expansion: Symmetric), (Breath Sounds: CTA\\n bilateral :), (Sternum: Stable)\\n Abdominal: Soft, Non-distended, Non-tender, Bowel sounds present\\n Left Extremities: (Edema: Absent), (Temperature: Warm), (Pulse -\\n Dorsalis pedis: Present), (Pulse - Posterior tibial: Present)\\n Right Extremities: (Edema: Absent), (Temperature: Warm), (Pulse -\\n Dorsalis pedis: Present), (Pulse - Posterior tibial: Present)\\n Skin: (Incision: Clean / Dry / Intact)\\n Neurologic: (Awake / Alert / Oriented: x 2), Follows simple commands,\\n Moves all extremities, Limited due to language barrier\\n Labs / Radiology\\n 275 K/uL\\n 9.8 g/dL\\n 134 mg/dL\\n 0.4 mg/dL\\n 26 mEq/L\\n 3.5 mEq/L\\n 15 mg/dL\\n 102 mEq/L\\n 137 mEq/L\\n 30.3 %\\n 8.8 K/uL\\n [image002.jpg]\\n [**2140-7-23**] 03:30 PM\\n [**2140-7-24**] 02:51 AM\\n [**2140-7-24**] 03:03 AM\\n [**2140-7-24**] 08:13 AM\\n [**2140-7-24**] 10:07 AM\\n [**2140-7-25**] 02:45 AM\\n [**2140-7-26**] 01:15 AM\\n [**2140-7-27**] 03:09 AM\\n [**2140-7-27**] 10:58 AM\\n [**2140-7-28**] 02:58 AM\\n WBC\\n 9.7\\n 10.3\\n 11.2\\n 7.7\\n 7.1\\n 8.8\\n Hct\\n 31.8\\n 32.6\\n 34.3\\n 33.3\\n 31.4\\n 30.3\\n Plt\\n [**Telephone/Fax (3) 8785**]\\n Creatinine\\n 0.5\\n 0.5\\n 0.5\\n 0.5\\n 0.5\\n 0.5\\n 0.4\\n TCO2\\n 26\\n 28\\n 29\\n Glucose\\n 168\\n 253\\n 147\\n 180\\n 92\\n 160\\n 194\\n 134\\n Other labs: PT / PTT / INR:11.6/25.8/1.0, CK / CK-MB / Troponin\\n T:54//<0.01, ALT / AST:25/32, Alk-Phos / T bili:87/,\\n Differential-Neuts:93.0 %, Lymph:5.3 %, Mono:1.0 %, Eos:0.5 %, Lactic\\n Acid:1.5 mmol/L, Ca:7.9 mg/dL, Mg:1.8 mg/dL, PO4:2.5 mg/dL\\n Assessment and Plan\\n AIRWAY, INABILITY TO PROTECT (RISK FOR ASPIRATION, ALTERED GAG, AIRWAY\\n CLEARANCE, COUGH), CVA (STROKE, CEREBRAL INFARCTION), HEMORRHAGIC ,\\n HYPERTENSION, BENIGN, [**Last Name 12**] PROBLEM - ENTER DESCRIPTION IN COMMENTS\\n Assessment and Plan: 69 yo F w/ left cerebellar thrombotic stroke,\\n hemorrhage, transtentorial herniation s/p EVD placement, surgical\\n decompression on [**7-22**], now w/ improved neuro exams\\n Neurologic: ICP monitor, Pain controlled, s/p crani for cerebellar\\n CVA, moves all 4, EVD clamped.

John Snow LABS DEEP 6 AI

Why is recruitment hard (data)

- Indexing patient data
- Generating search criteria
- Find a match

Increasingly complex enrollment criteria

Inclusion criteria:

Histologically or cytologically confirmed adenocarcinoma of the prostate at initial biopsy, without neuroendocrine differentiation, signet cell, or small cell features.

Prostate cancer initially treated by radical prostatectomy or radiotherapy (including brachytherapy) or both, with curative intent.

Screening PSA ≥ 2.0 ng/mL for patients who had radical prostatectomy as primary treatment for prostate cancer or ≥ 5.0 ng/mL and greater than or equal to the nadir + 2 ng/mL for patients who had radiotherapy as primary treatment for prostate cancer.

Exclusion criteria:

Prior or present evidence of distant metastatic disease as assessed by radiographic imaging.

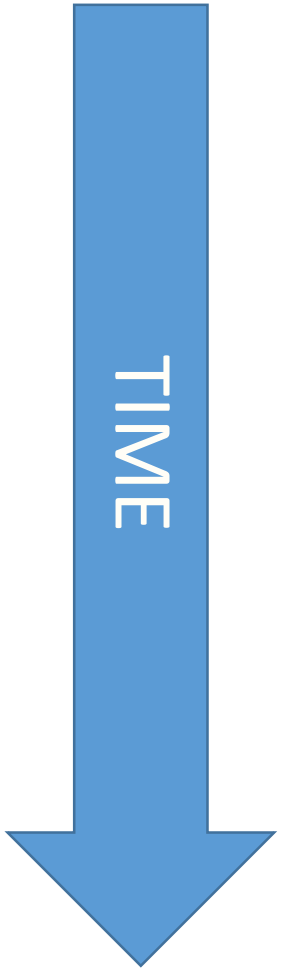
John Snow LABS DEEP 6 AI

Why is recruitment hard (data)

- Indexing patient data
- Generating search criteria
- Find a match

What is the patient's current status?

- "Patient shows signs of cancer"
- "Tested positive for carcinoma"
- "Treating cancer with chemo"
- "Cancer unresponsive, changing treatment"
- "Cancer responding to new line of therapy"
- "Cancer in remission"
- "History of cancer"



Why is recruitment hard (data)

- Indexing patient data
- Generating search criteria
- Find a match

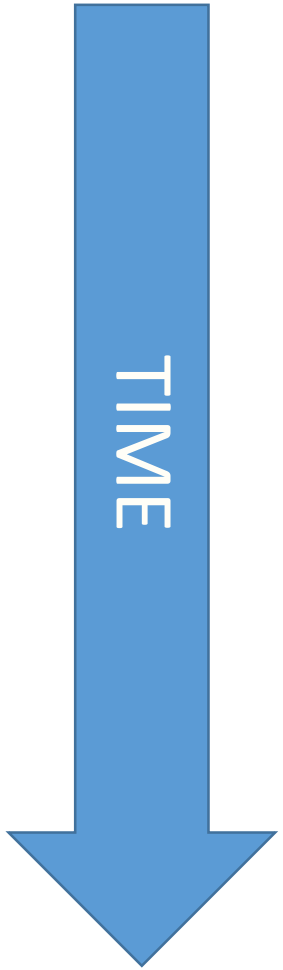
What is the patient's current status?

- "Patient shows signs of cancer"

TB of Data

- "Cancer unresponsive, changing treatment"
- "Cancer responding to new line of therapy"
- "Cancer in remission"

Continuously updated



Human and clinical language is

- Nuanced
- Fuzzy
- Contextual
- Medium specific
- Domain specific
- Contains typos & mistakes

Language Understanding (clinical)

Inclusion criteria:

Histologically or cytologically confirmed adenocarcinoma of the prostate at initial biopsy, without neuroendocrine differentiation, signet cell, or small cell features.

Prostate cancer initially treated by radical prostatectomy or radiotherapy (including brachytherapy) or both, with curative intent.

Screening PSA ≥ 2.0 ng/mL for patients who had radical prostatectomy as primary treatment for prostate cancer or ≥ 5.0 ng/mL and greater than or equal to the nadir + 2 ng/mL for patients who had radiotherapy as primary treatment for prostate cancer.

Exclusion criteria:

Prior or present evidence of distant metastatic disease as assessed by radiographic imaging.

John Snow LABS DEEP 6 AI

Human and clinical language is

- Nuanced
- Fuzzy
- Contextual
- Medium specific
- Domain specific
- Contains typos & mistakes

Introducing John Snow Labs' Spark NLP

We'll need:

- Core Annotators

  Sentence detection, part of speech tagging, spell checking …

- Vocabulary Understanding

  Ontologies, relationships, word embeddings …

- ML & DL Models

  Named entity recognition, entity resolution, negation analysis

John Snow LABS DEEP 6 AI

State of the art

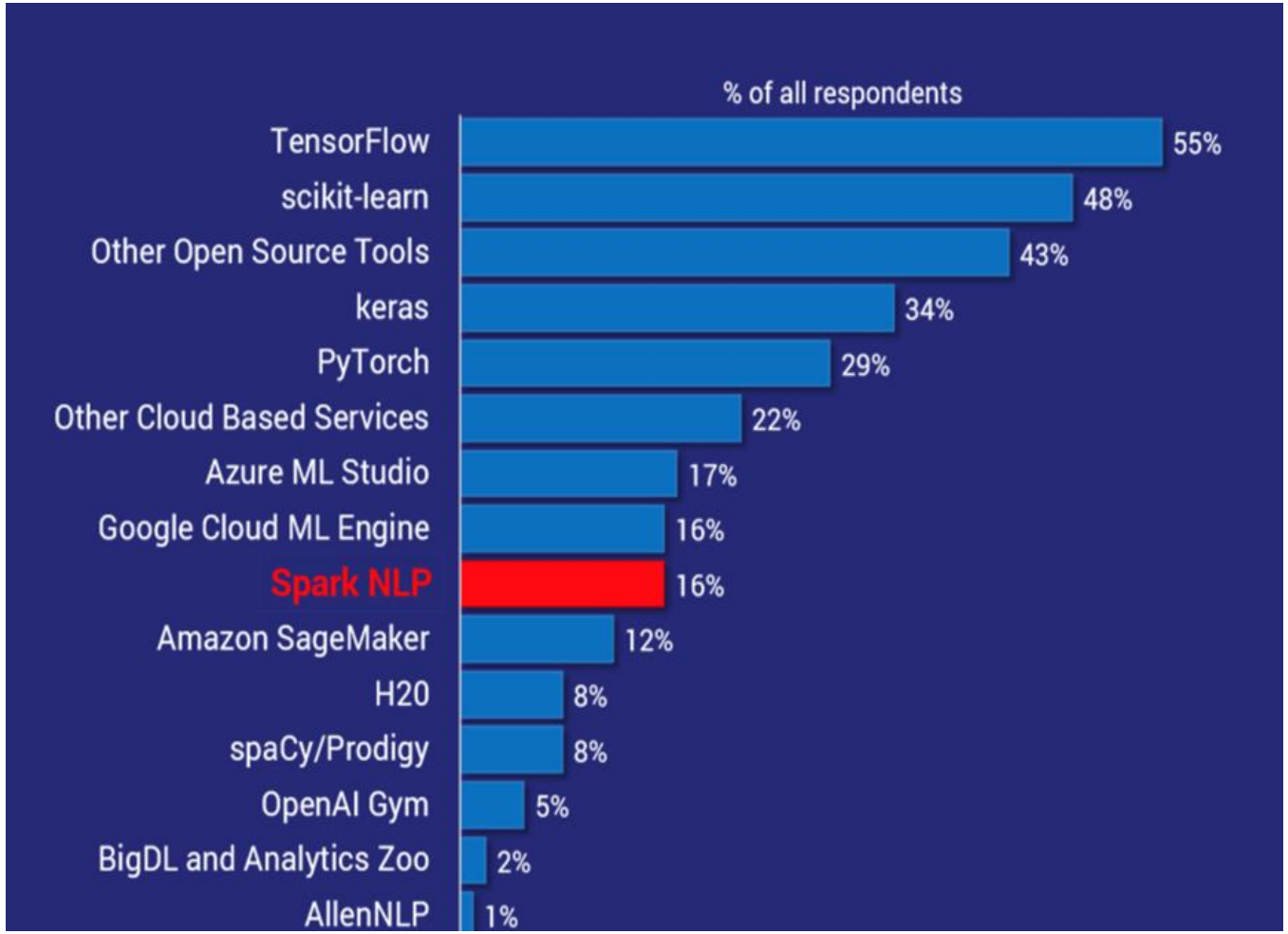
- **Quality**
- **Speed**
- **Scalability**

Apache 2.0 Licensed
Scala and Python APIs
Apache Spark & Tensorflow
Active development & support
Healthcare specific edition

nlp.johnsnowlabs.com

github.com/johnsnowlabs/spark-nlp

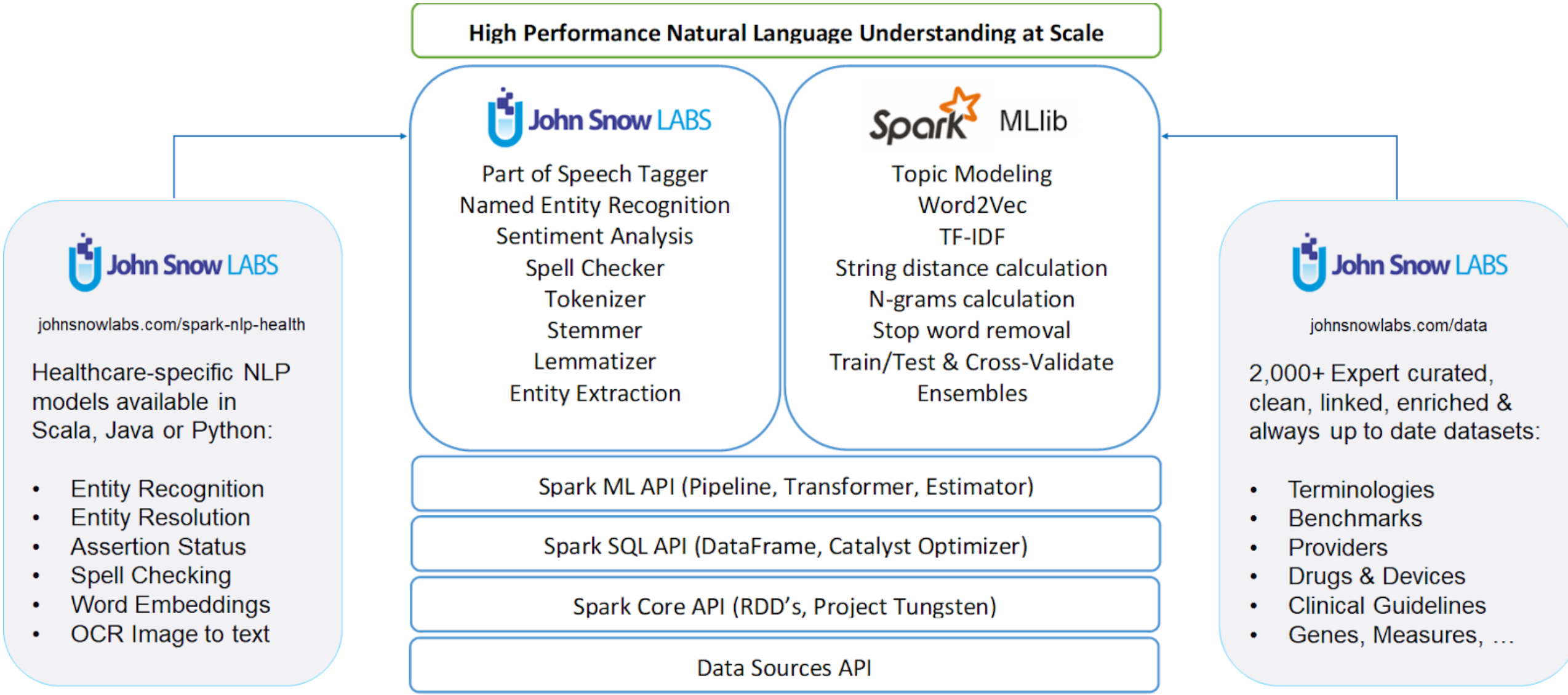
Introducing John Snow Labs' Spark NLP



"AI Adoption in the Enterprise", February 2019
Most widely used ML frameworks and tools
survey of 1,300 practitioners by O'Reilly

John Snow LABS DEEP 6 AI

Spark NLP Feature Overview



John Snow LABS DEEP 6 AI

State of the art

- **Quality**
- Speed
- Scalability

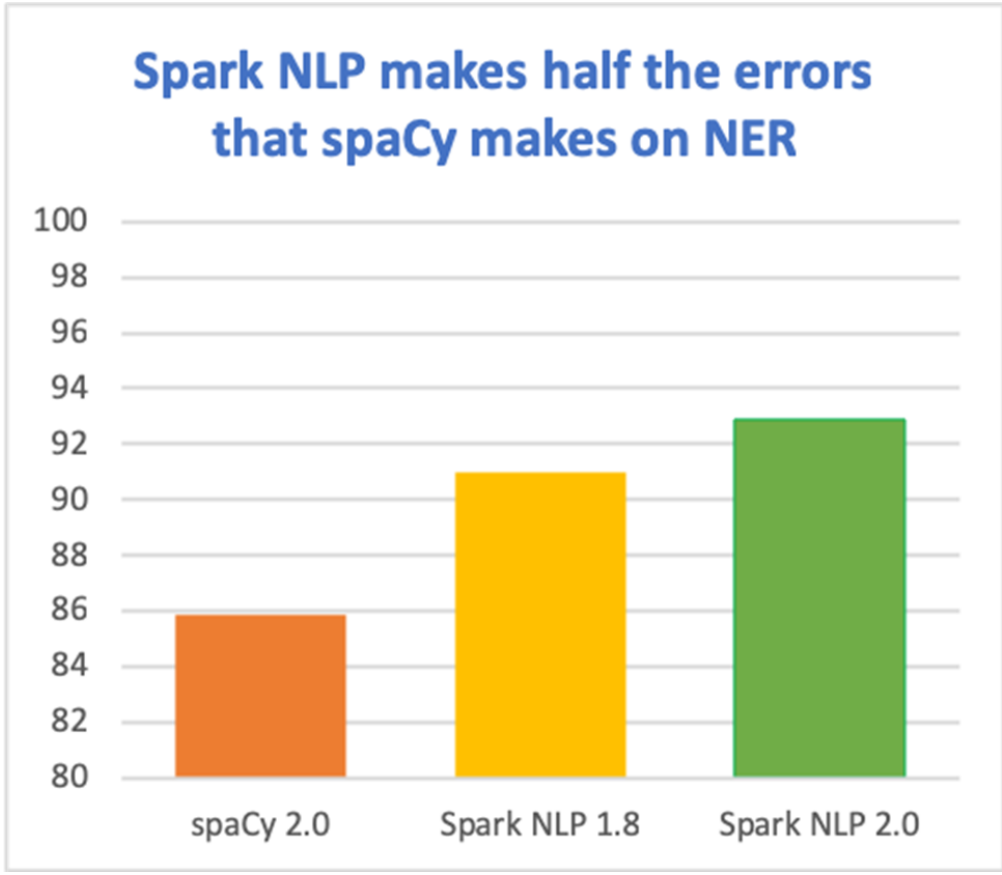
Apache 2.0 Licensed
Scala and Python APIs
Apache Spark & Tensorflow
Active development & support
Healthcare specific edition

nlp.johnsnowlabs.com

github.com/johnsnowlabs/spark-nlp

Introducing John Snow Labs' Spark NLP

•Deep learning, trainable models
•TF graph based on 2017 paper (bi-LSTM+CNN+CRF)
•BERT embeddings
•Regularly pretrained models
•Benchmark on right is on *en_core_web_lg* dataset, F1 score calculations included (2.2.1)



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

John Snow LABS DEEP 6 AI

State of the art

- Quality
- **Speed**
- Scalability

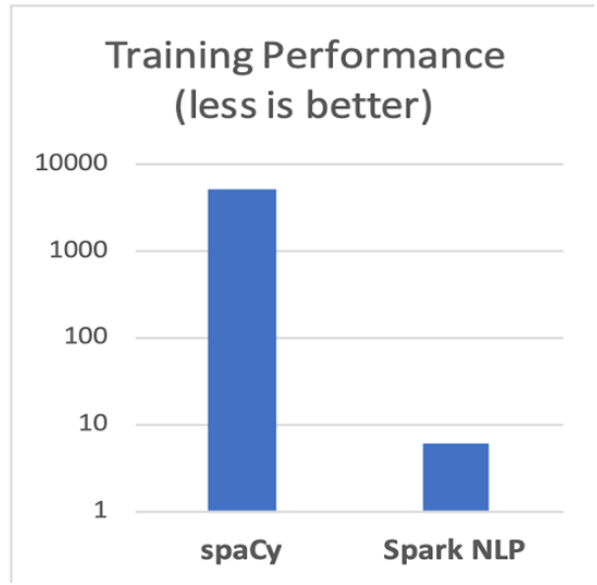
Apache 2.0 Licensed
Scala and Python APIs
Apache Spark & Tensorflow
Active development & support
Healthcare specific edition

nlp.johnsnowlabs.com

github.com/johnsnowlabs/spark-nlp

Introducing John Snow Labs' Spark NLP

- 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?
  - Apache Spark concurrency
  - bare-metal performance
  - in memory optimizations optimized for training
  - Cluster capable inference



Spark NLP trains 80x faster than spaCy on one machine

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

John Snow LABS DEEP 6 AI

State of the art

- Quality
- Speed
- **Scalability**

Apache 2.0 Licensed
Scala and Python APIs
Apache Spark & Tensorflow
Active development & support
Healthcare specific edition

nlp.johnsnowlabs.com

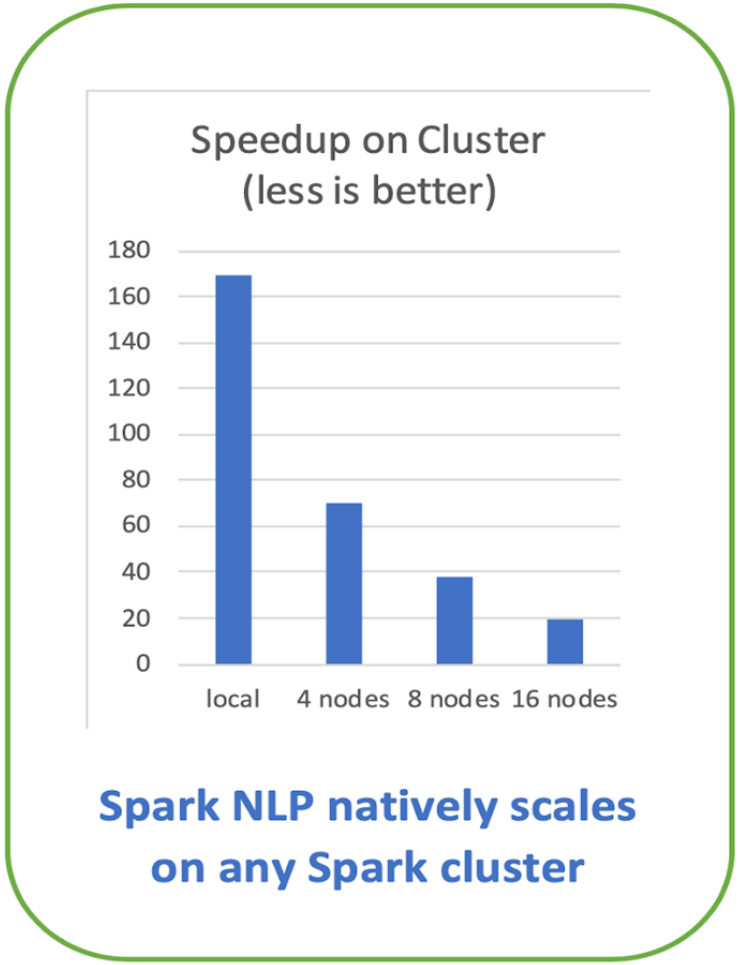
github.com/johnsnowlabs/spark-nlp

Introducing John Snow Labs' Spark NLP

- Zero code changes to switch between local and cluster modes
- It's the only natively distributed open source NLP library
- Apache Spark provides execution planning, caching, serialization, shuffling

Caveats:

- Speedup depends heavily on the nature of the task
- Some algorithms use better concurrency advantage than others
- Spark configuration matters



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

John Snow LABS DEEP 6 AI

Use Case

- Problem
- Detection
- Refinement
- Deploy

Triple Negative Breast Cancer

- Breast Cancer can have hormone receptors
  - ER, PR, HER-2
  - If present cancer feeds on hormones
    - Treat with hormone therapy

Use Case

- Problem
- Detection
- Refinement
- Deploy

Triple Negative Breast Cancer

- Breast Cancer can have hormone receptors
  - ER, PR, HER-2
  - If present cancer feeds on hormones
    - Treat with hormone therapy
- If missing all three: “Triple Negative”
- Over 500k representations:
  - “Er-/pr-/h2-”
  - “(er pr her2) negative”
  - “Tested negative for the following: er, pr, h2”
  - “Triple negative neoplasm of the upper left breast”

Use Case

- Problem
- Detection
- Refinement
- Deploy

Find all mentions

- Generate a text document with 500k test patterns
- Generate a regex matching document with 50 patterns
- Build text matching pipeline

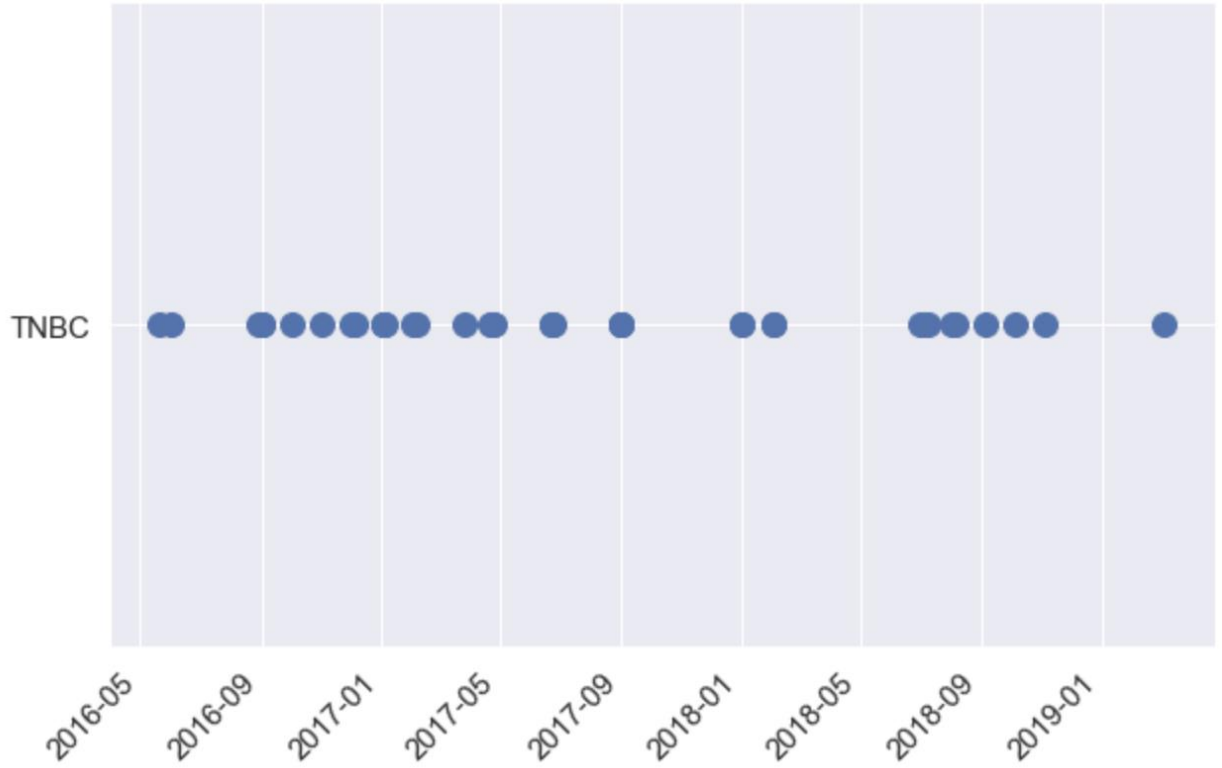
```
pipeline = Pipeline(
    stages=[
        DocumentAssembler()
            .setInputCol('sentence_text')
            .setOutputCol('sentence')
            .setTrimAndClearNewLines(False),
        Tokenizer()
            .setInputCols(['sentence'])
            .setOutputCol('token'),
        TextMatcher()
            .setInputCols(['sentence', 'token'])
            .setOutputCol(self.TNBC_TEXT_MATCH)
            .setEntities(path=tnbc_phrase_pattern_path)
            .setCaseSensitive(False),
        RegexMatcher()
            .setInputCols(['sentence'])
            .setStrategy('MATCH_ALL')
            .setOutputCol(self.TNBC_REGEX_MATCH)
            .setExternalRules(path=tnbc_regex_pattern_path, delimiter=',')
    ]
)
input_schema = StructType([StructField('sentence_text', StringType())])
empty_training_df = self.spark.createDataFrame([], input_schema)
return pipeline.fit(empty_training_df)
```

Use Case

- Problem
- Detection
- Refinement
- Deploy

Find all mentions

- 13 r4.xlarge boxes
  - ~155 sentences/sec*core



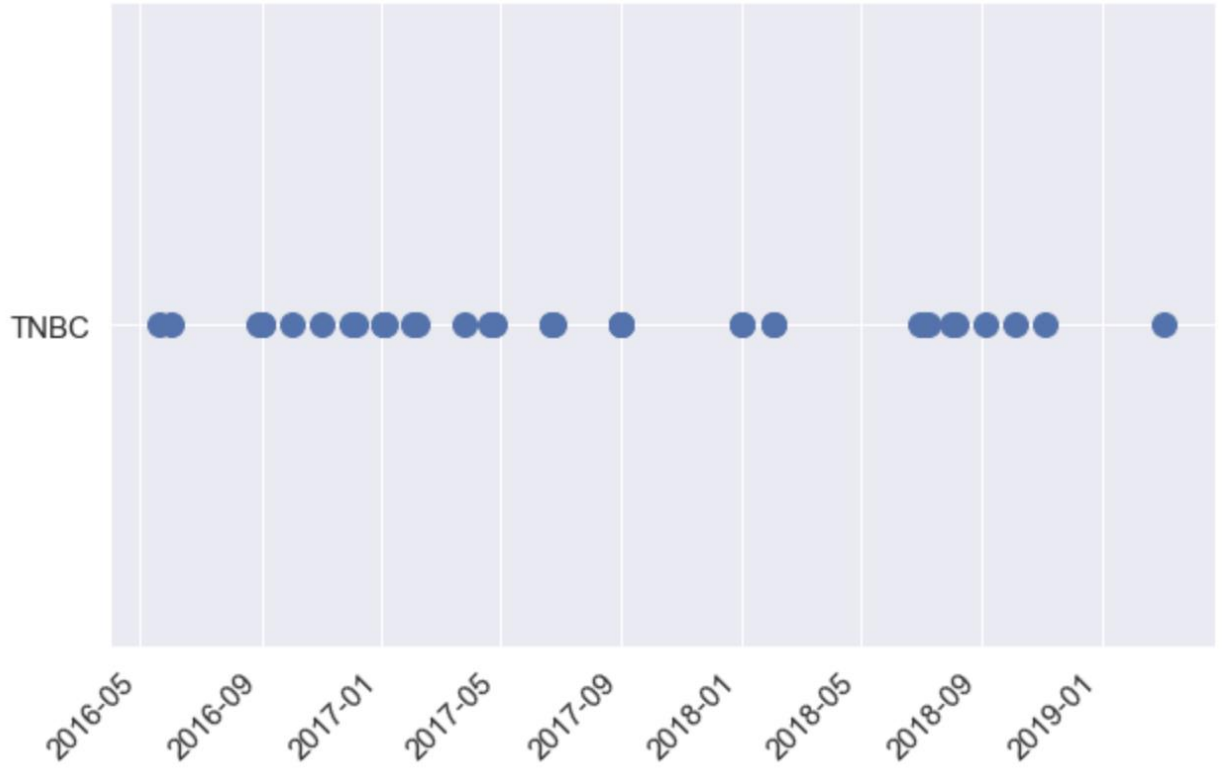
John Snow LABS DEEP 6 AI

Use Case

- Problem
- Detection
- Refinement
- Deploy

What else is going on?

- We see a lot of data points
- Interesting spacing
- Feedback about false positives



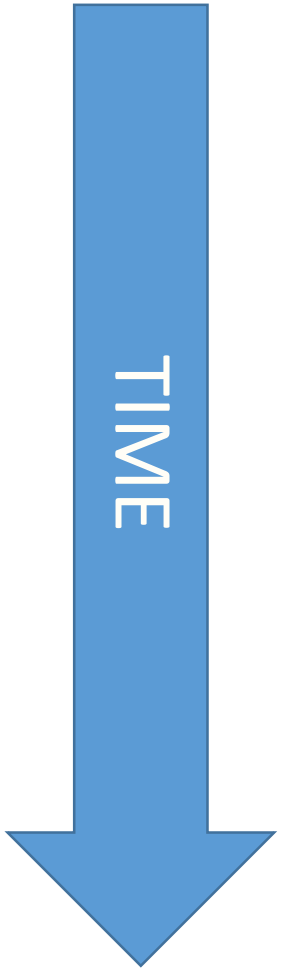
John Snow LABS DEEP 6 AI

Use Case

- Problem
- Detection
- Refinement
- Deploy

What else is going on?

- “Patient shows signs of cancer”
- “Tested positive for carcinoma”
- “Treating cancer with chemo”
- “Cancer unresponsive, changing treatment”
- “Cancer responding to new line of therapy”
- “Cancer in remission”
- “History of cancer”



John Snow LABS DEEP 6 AI

Use Case

- Problem
- Detection
- Refinement
- Deploy

Assertion status

Text	Target Chunk	Status	B	E
The patient is an 84-year-old woman with a history of **diverticulitis** who was found to have colon cancer on colonoscopy , which was performed in March of 2004 .	diverticulitis	present	10	10
The patient's family history was significant for a brother with **colon cancer** .	colon cancer	associated_with_someone_else	11	12
This is a 70 year old gentleman with metastatic rectal cancer who presented with **biliary obstruction** .	biliary obstruction	present	14	15
The patient denies any recent upper respiratory infections , no fevers , no **chills** , no change in cough , sputum .	chills	absent	13	13

Annotator Properties

•**Inputs:** *SENTENCE - CHUNK* (Chunk must be provided by NER, Text Matcher or Regex Matcher)

Output: Status asserted for each chunk

https://github.com/JohnSnowLabs/spark-nlp-workshop/blob/master/jupyter/annotation/english/healthcare/GlobalDEMO-Clinical-Analysis.ipynb

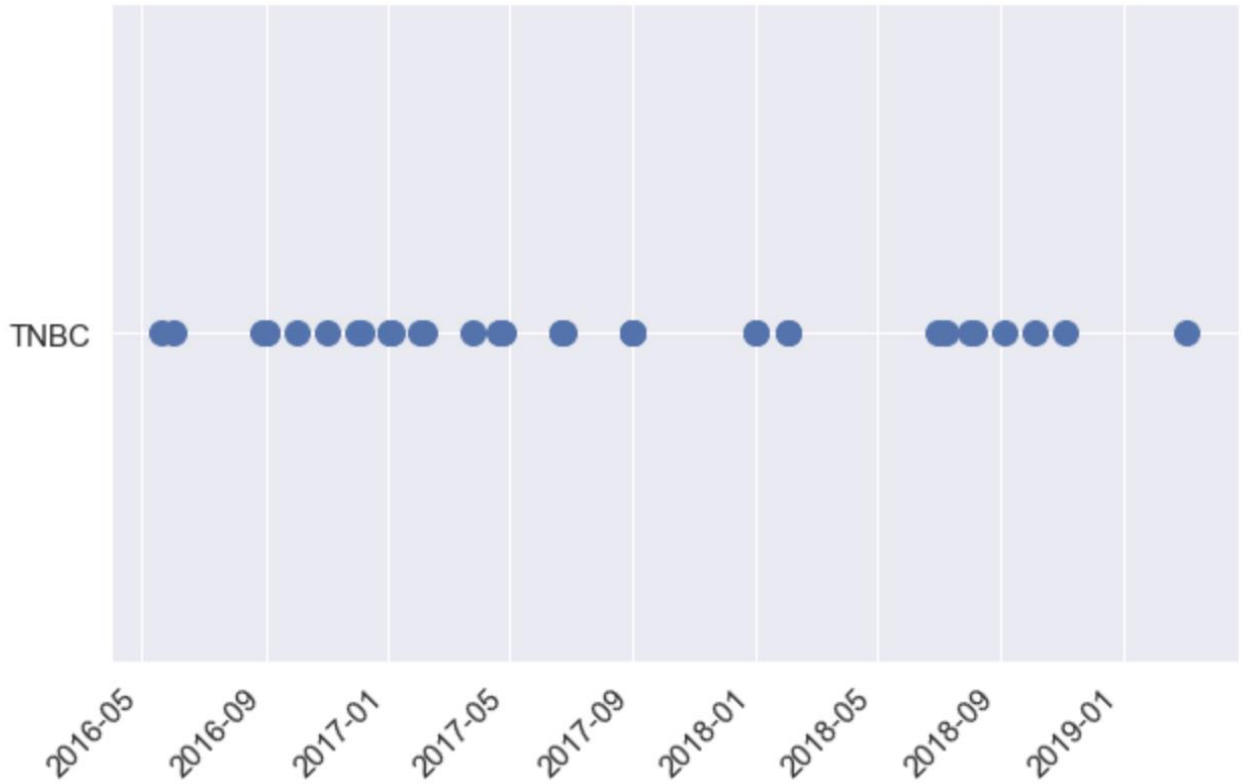
John Snow LABS DEEP 6 AI

Use Case

- Problem
- Detection
- Refinement
- Deploy

Assertion status

```
def build_model(self) -> PipelineModel:
    return PipelineModel.read().load(self.models_directory + "/nerdl_assertion_model_100ep")
```



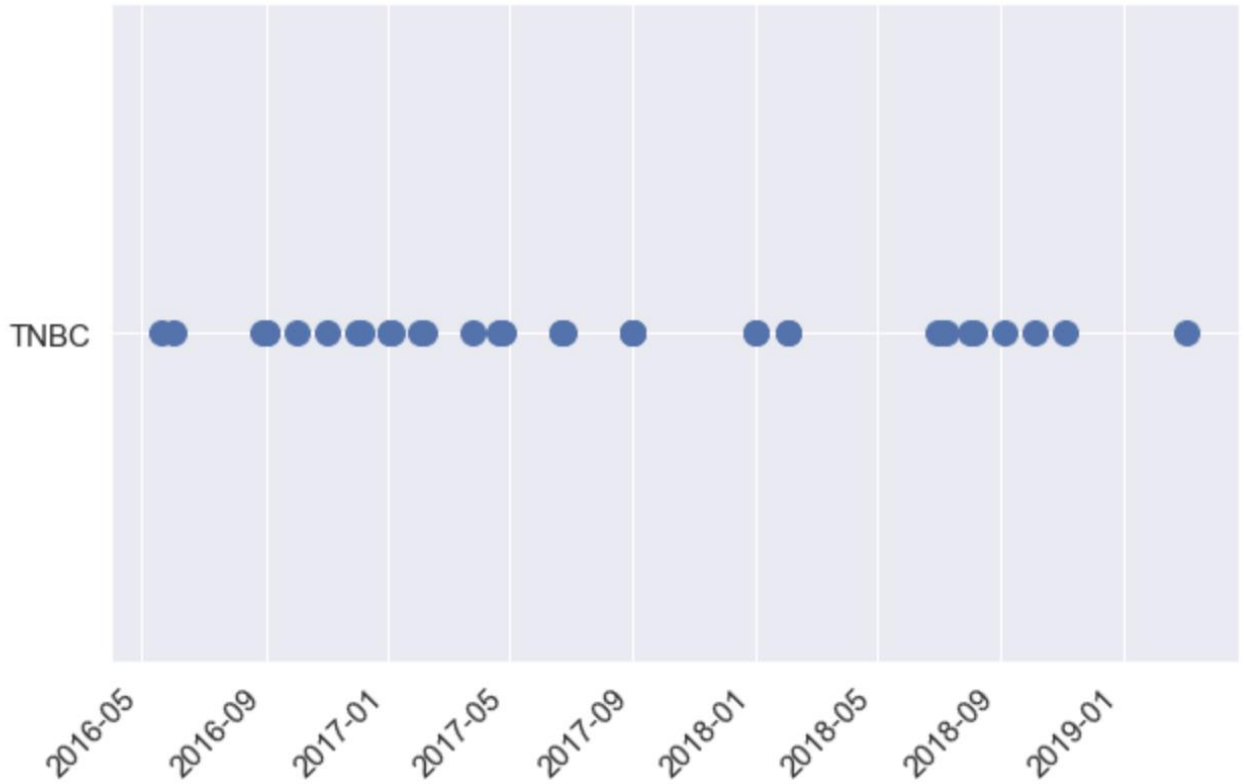
John Snow LABS DEEP 6 AI

Use Case

- Problem
- Detection
- Refinement
- Deploy

Assertion status

- 13 r4.xlarge boxes
  - ~32 sentences/sec*core

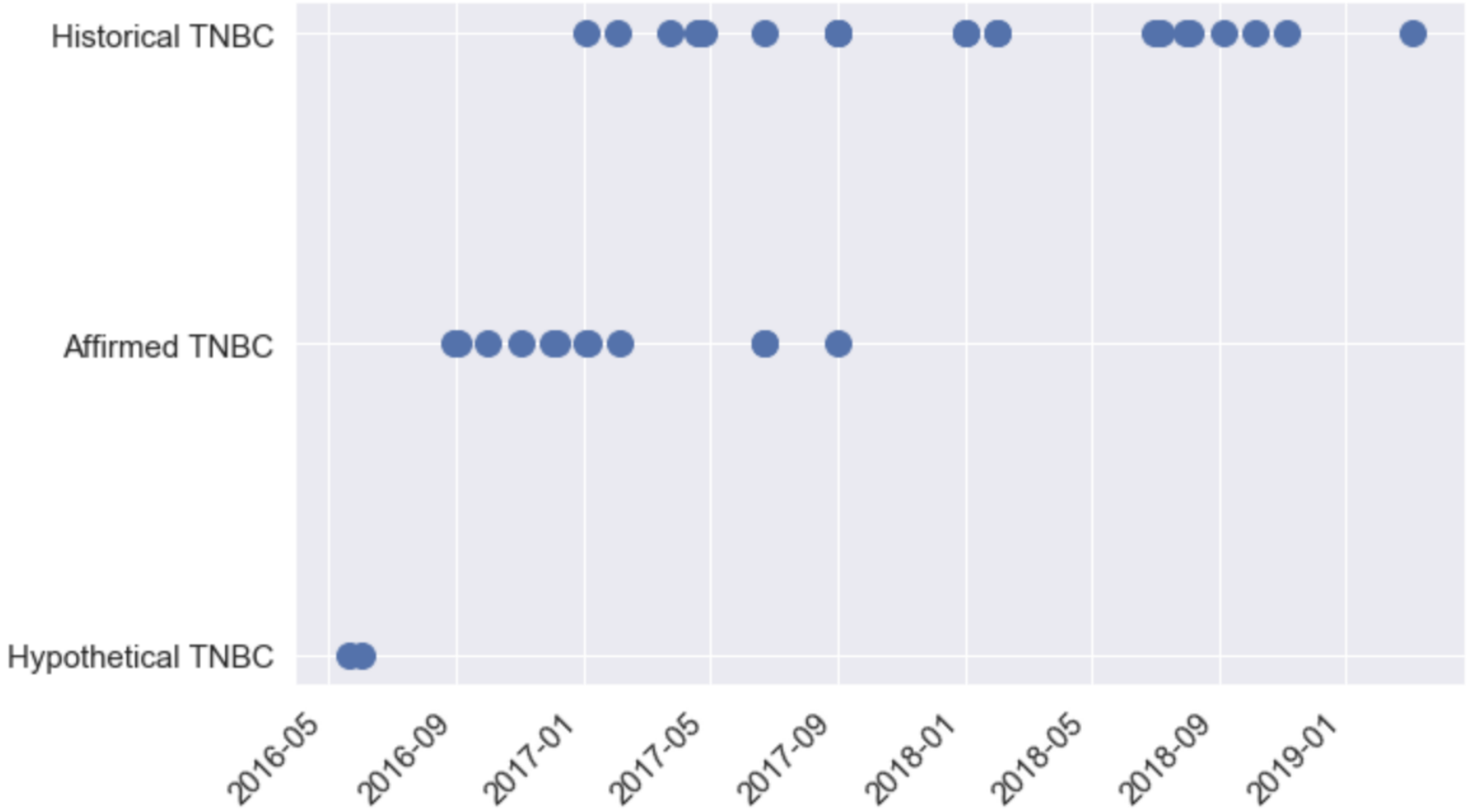


John Snow LABS DEEP 6 AI

Use Case

- Problem
- Detection
- Refinement
- Deploy

Assertion status

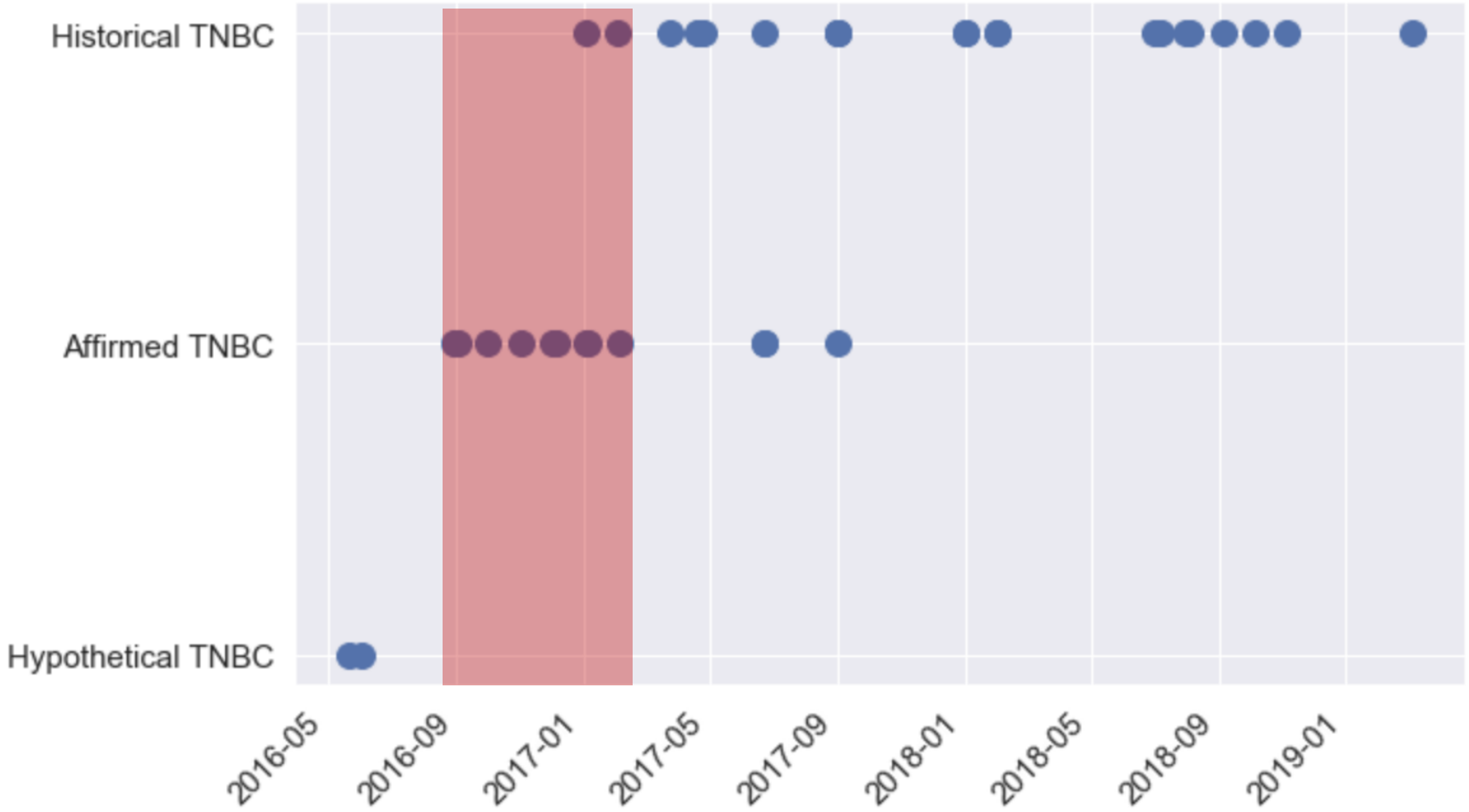


John Snow LABS DEEP 6 AI

Use Case

- Problem
- Detection
- Refinement
- Deploy

Assertion status

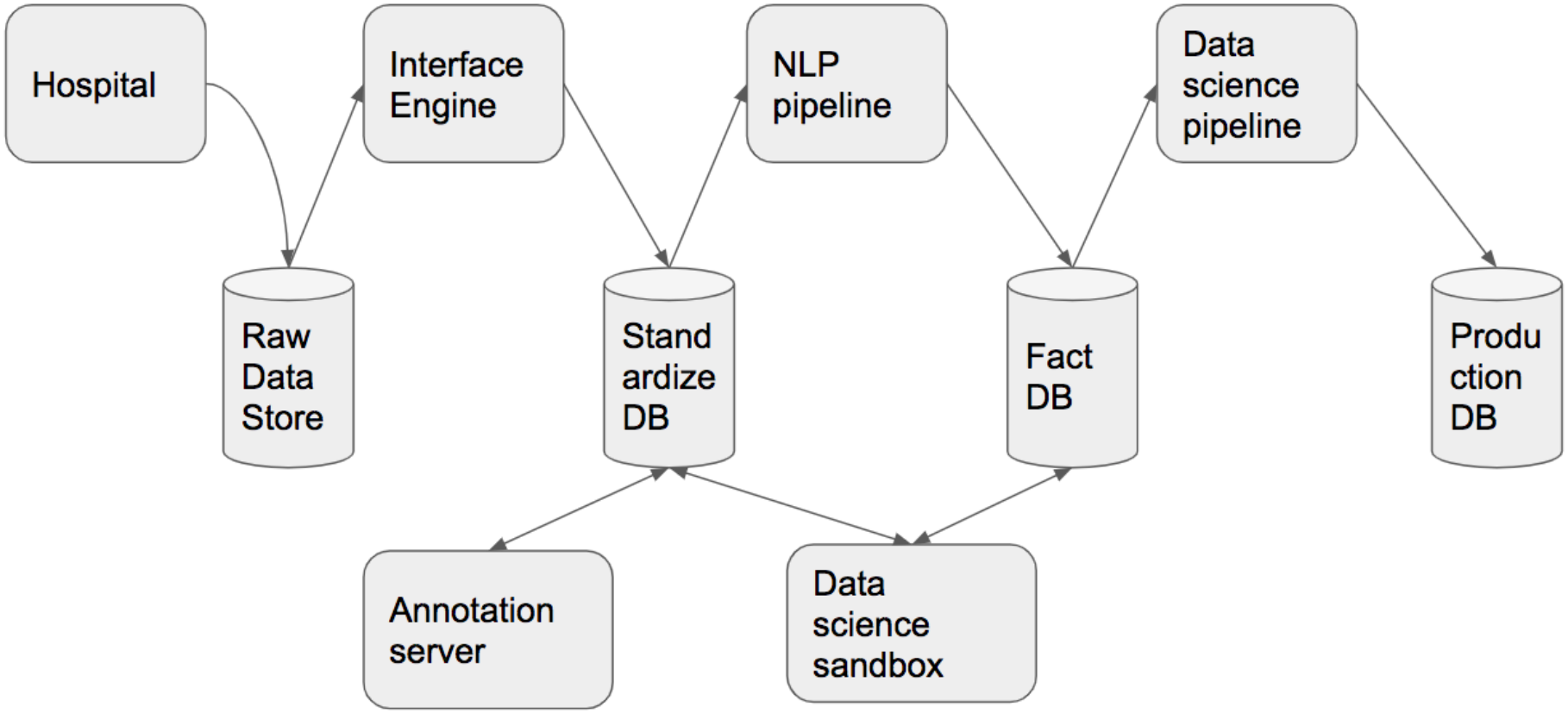


John Snow LABS DEEP 6 AI

Use Case

- Problem
- Detection
- Refinement
- Deploy

Assertion status



John Snow LABS DEEP 6 AI

Thank You

Saif Addin Ellafi - John Snow Labs

Software Engineer
Spark NLP - Lead Developer
saif@johnsnowlabs.com

Scott Hoch – BlackBox Engineering

Getting it done
Managing Member
scott@blackboxengineering.work

John Snow LABS DEEP 6 AI