



Using Natural Language Processing (NLP) to curate unstructured electronic health records into research ready datasets

07/26/2022



The difference is in the data.™

Our Presenter



Eric Harvey, PhD

Sr. Director

Biometrics and Data Science

eharvey@mms Holdings.com

Abstract

Eric Harvey, PhD, Senior Director of Biometrics and Data Science at MMS Holdings will present an overview of Natural Language Processing (NLP), an Artificial Intelligence technique that can be used to curate unstructured medical records. We will see NLP in action as part of the ICODA Grand Challenges 'PRIEST' project (Pandemic Respiratory Infection Emergency System Triage) Study for Low and Middle-Income Countries as a case study.

Agenda

- Overview
- Natural Language Processing (NLP) background
- Pandemic Respiratory Infection Emergency System Triage (PRIEST) Use Case
- Demo of Amazon Comprehend Medical
- Additional Considerations
- Conclusions





Overview

A Few Definitions

Real-World Data (RWD):

Real-world data are the data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.

RWD can come from a number of sources, for example:

- Electronic health records (EHRs)
- Claims and billing activities
- Product and disease registries
- Patient-generated data including in home-use settings
- Data gathered from other sources that can inform on health status, such as mobile devices



Not purpose-collected for research

<https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>

<https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory>

A Few Definitions (continued)

- Unstructured Data: Data that are not organized in a pre-defined format that have less inherent internal order, must be curated to be made research ready (ex: free-text medical records, video & audio, social media)
- Natural Language Processing (NLP): Machine learning (artificial intelligence subdiscipline) technique applied to allow computers to understand written and spoken language similarly to humans (ex: Alex, Siri, voice GPS navigation)
- Data Curation: Organizing and abstracting data (often from non-traditional sources) for use in clinical research utilizing appropriate tools or technology (ex: utilizing NLP to abstract data from case report notes)

Data Barriers in Clinical Research

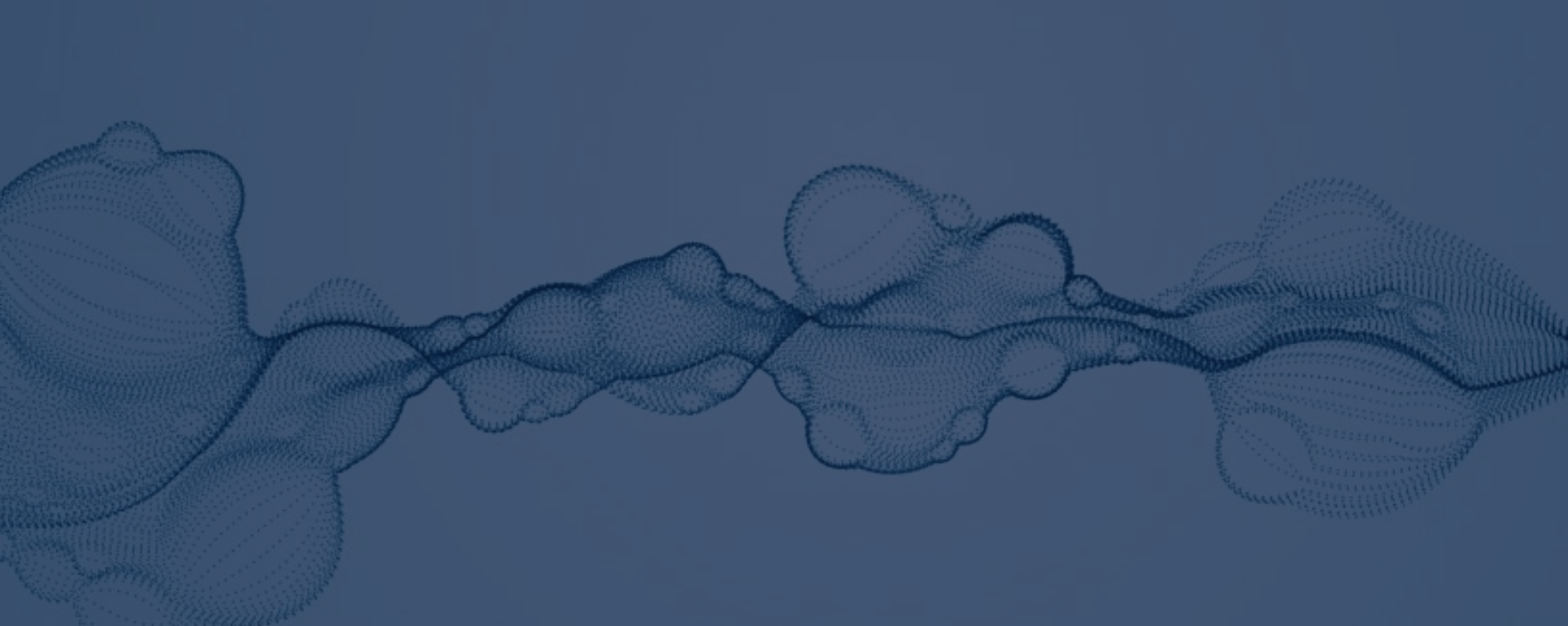
- Quality data are expensive and time-consuming to collect
- Data are messy and the cleaning process often involves manual effort which is not scalable for large datasets
- Increasing global focus on data privacy and security restricts available data for research
- Data sources vary in quality, outliers and inaccurate data can potentially bias results
- Adverse impact of pandemics and world events on global patient recruiting efforts

Electronic Healthcare Record (EHR) Opportunities

- ~80% of medical data are unstructured
- Increasing data interconnectedness across primary and secondary care systems
- Global data transparency efforts making more research results available

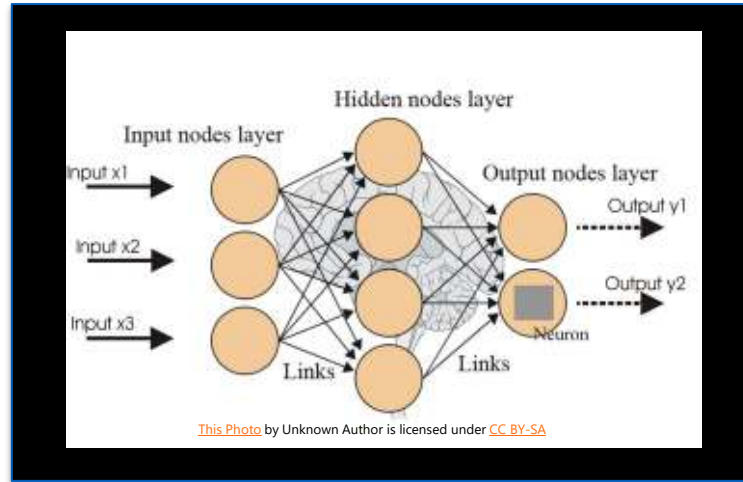


Data curation and machine learning technologies like NLP can make RWD more accessible for use in clinical research



NLP Background

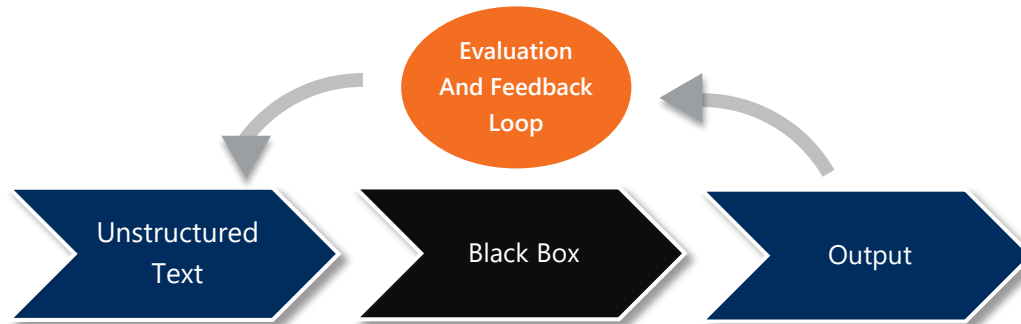
10,000 foot NLP Overview



“black box”: it can be difficult to explain how a computer “learned” to get from input to output

How does NLP “learn”?

- Requires labeled training data (contains both unstructured data and the correct answer/outcome)
- Modeled on a neuroscience model of how human brain learns (feedback loops, highly iterative process)
- Training data require much manual input and are hard to come by in sufficient volume
- Fortunately, there are large pre-trained models in our industry



Some NLP Use Cases

- Patient classification (today's focus)
- Extraction of key safety data (medications, medical history, adverse reactions)
- Post marketing surveillance (sentiment analysis / social media)
- Targeted recruitment (I/E criteria evaluation)

Sample applications

Rare disease

Disease registries

Long term follow-up (diagnostic criteria changes)

Decentralized clinical trials (DCTs)



Industry just beginning to realize NLP potential

What do I need to get started?

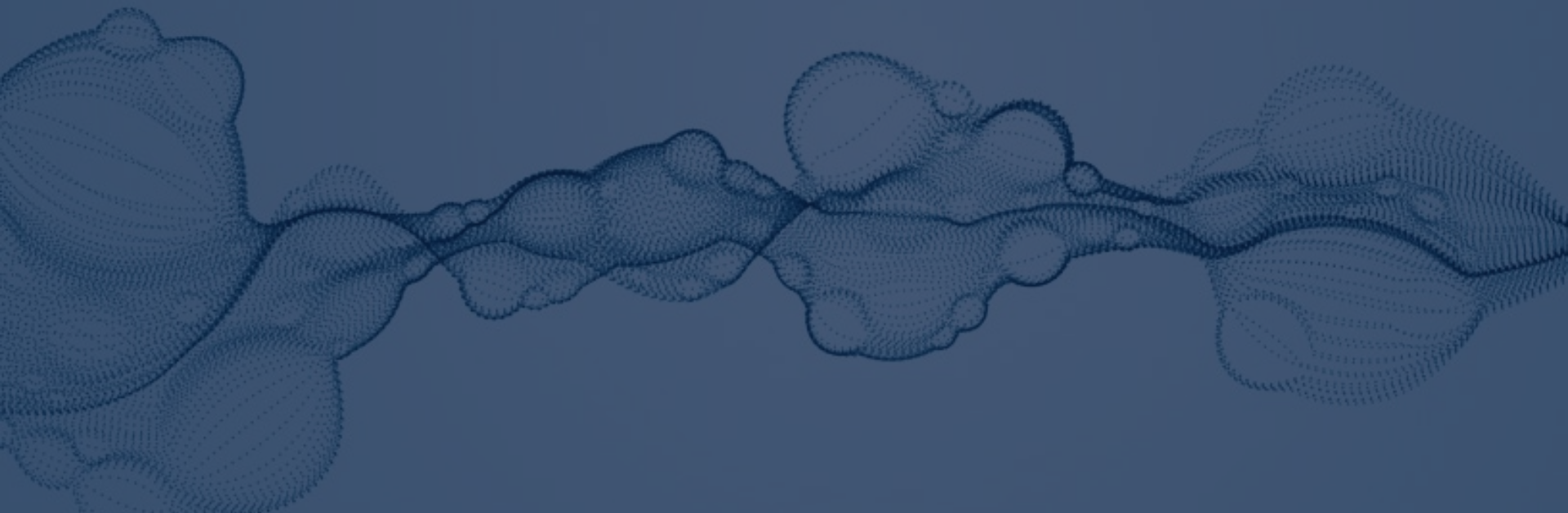
- Business need that NLP can help to achieve
- Unstructured data source
- Access to a text analytics platform (preferably pre-trained)
- Data curation plan to detail data mapping to final output
- Personnel with programming or technical ability (does NOT necessarily require specialized NLP knowledge unless training or tweaking a model)

NLP Limitations

- NOT 100% accurate (but can improve accuracy over time)
- Models can learn spurious patterns given inappropriate training data (supervised learning can help)
- Lower accuracy for data types not well represented in training data (ex: slang, abbreviations)
- Not all languages are well represented
- May required limited manual review of results, depending on use case (utilize confidence scores)
- Modification of “out-of-the-box” models can require a more specialized resource

Major Text Analytics Engines in our Industry

- Amazon Comprehend Medical
<https://aws.amazon.com/comprehend/medical/>
- Microsoft Text Analytics for Health
<https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/>
- Google Healthcare Natural Language
<https://cloud.google.com/healthcare-api/docs/concepts/nlp/>



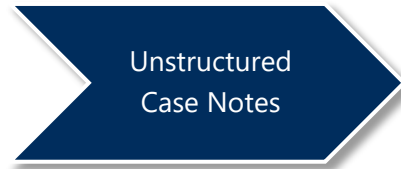
PRIEST Use Case

Pandemic Respiratory Infection Emergency System Triage (PRIEST)

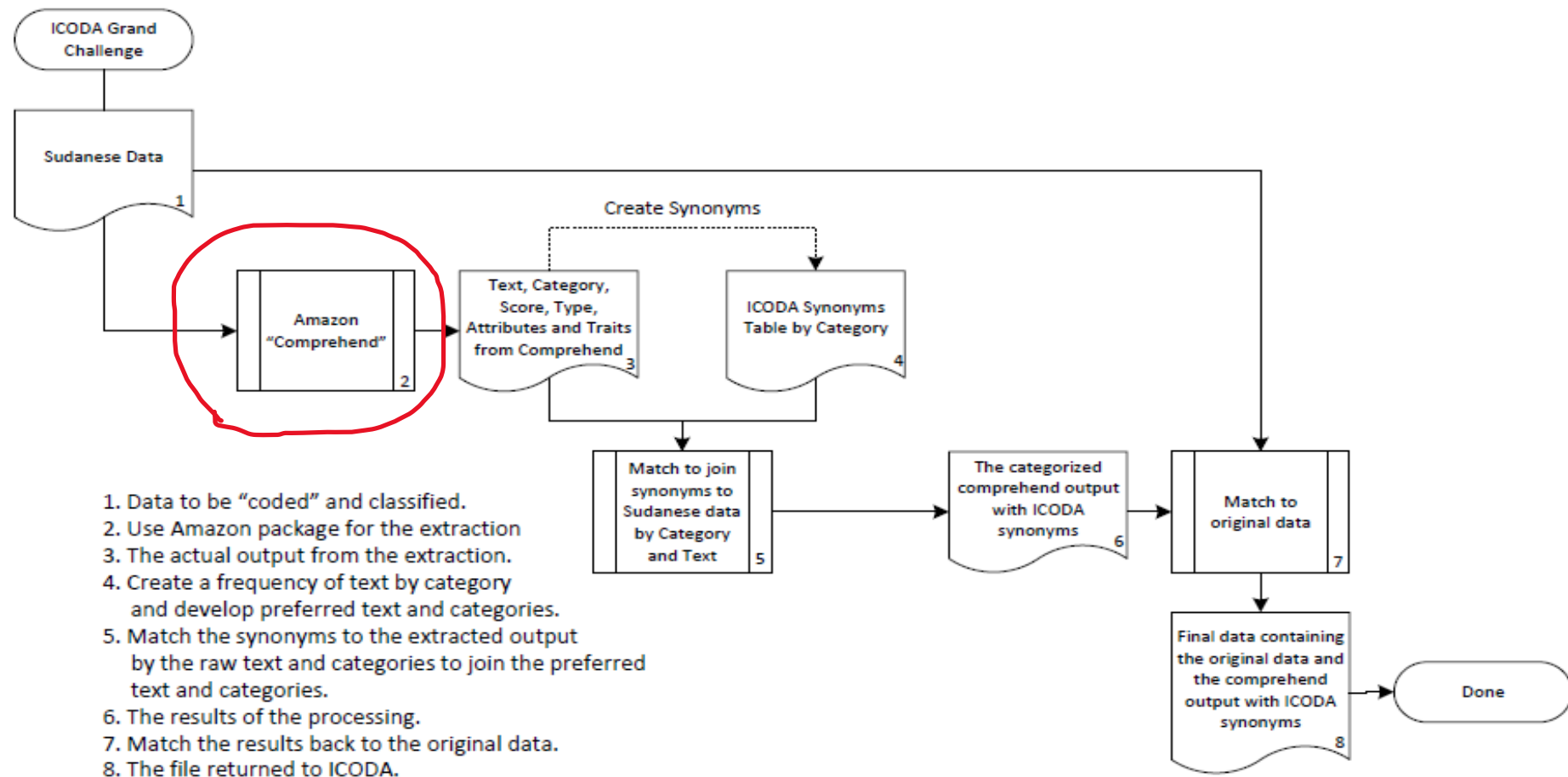
- **Goal:** develop a COVID-19 focused risk assessment tool for clinicians to quickly decide whether a patient needs emergency care or can be safely sent home
- Utilized existing data on 50,000 patients with suspected COVID-19 infection who sought emergency care in the United Kingdom, South Africa or Sudan
- Focused on low- and middle- income countries
- Use case presented from Sudanese data

Sudanese Data

- Contained unstructured case notes for several thousand individuals
- NLP was selected primarily because of the large effort to manually process data into a structured form
- **Goal:** Provide patient-level diagnosis flag (for use in modeling)



- Cardiovascular disease
- Renal impairment
- Asthma
- Diabetes
- Immunosuppression
- Malignancy
- Chronic lung disease
- Hypertension
- Pregnancy
- Shortness of breath
- Cough
- Confusion
- Maximum SBP



Results from AWS Comprehend Medical

"Pt is 87 yo woman, highschool teacher with past medical history that includes

- status post cardiac catheterization in April 2019.

She presents today with palpitations and chest pressure.

HPI : Sleeping trouble on present dosage of Clonidine.

Severe Rash on face and leg, slightly itchy

Meds : Vyvanse 50 mgs po at breakfast daily,

Clonidine 0.2 mgs -- 1 and 1 / 2 tabs po qhs

HEENT : Boggy inferior turbinates, No oropharyngeal lesion

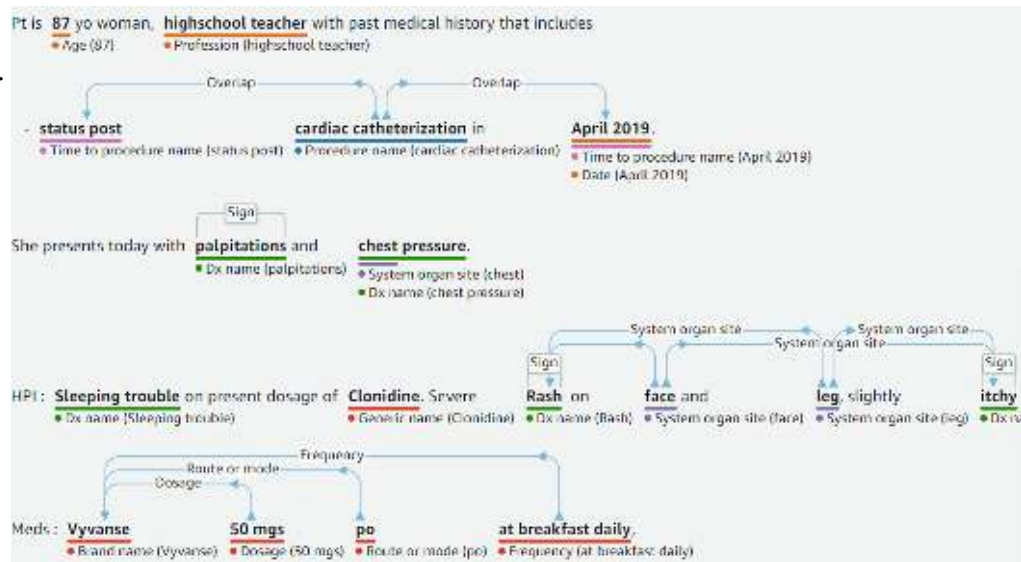
Lungs : clear

Heart : Regular rhythm

Skin : Mild erythematous eruption to hairline

Follow-up as scheduled"

Note: sample data is from Amazon Comprehend Medical [documentation](#). 20,000 character limit/record.



Results from AWS Comprehend Medical

Entity	Type	Category
87 0.9997 score	Age	Protected health information
highschool teacher 0.2063 score	Profession	Protected health information
status post 0.9525 score	Time to procedure name	Time expression
cardiac catheterization 0.9999+ score	Procedure name	Test treatment procedure
April 2019 0.9917 score	Time to procedure name	Time expression
cardiac catheterization 0.9999+ score	Procedure name	Test treatment procedure
April 2019 0.9998 score	Date	Protected health information

JavaScript Object Notation (JSON) sample output

```
{
  "Id": 17,
  "BeginOffset": 149,
  "EndOffset": 151,
  "Score": 0.9999990463256836,
  "Text": "BP",
  "Category": "TEST_TREATMENT_PROCEDURE",
  "Type": "TEST_NAME",
  "Traits": [],
  "Attributes": [{
    "Type": "TEST_VALUE",
    "Score": 0.9999966621398926,
    "RelationshipScore": 0.9999995231628418,
    "RelationshipType": "TEST_VALUE",
    "Id": 18,
    "BeginOffset": 152,
    "EndOffset": 158,
    "Text": "120/70",

```




Entity Category
Entity Name
Predicted confidence (0-1)

PRIEST sample results

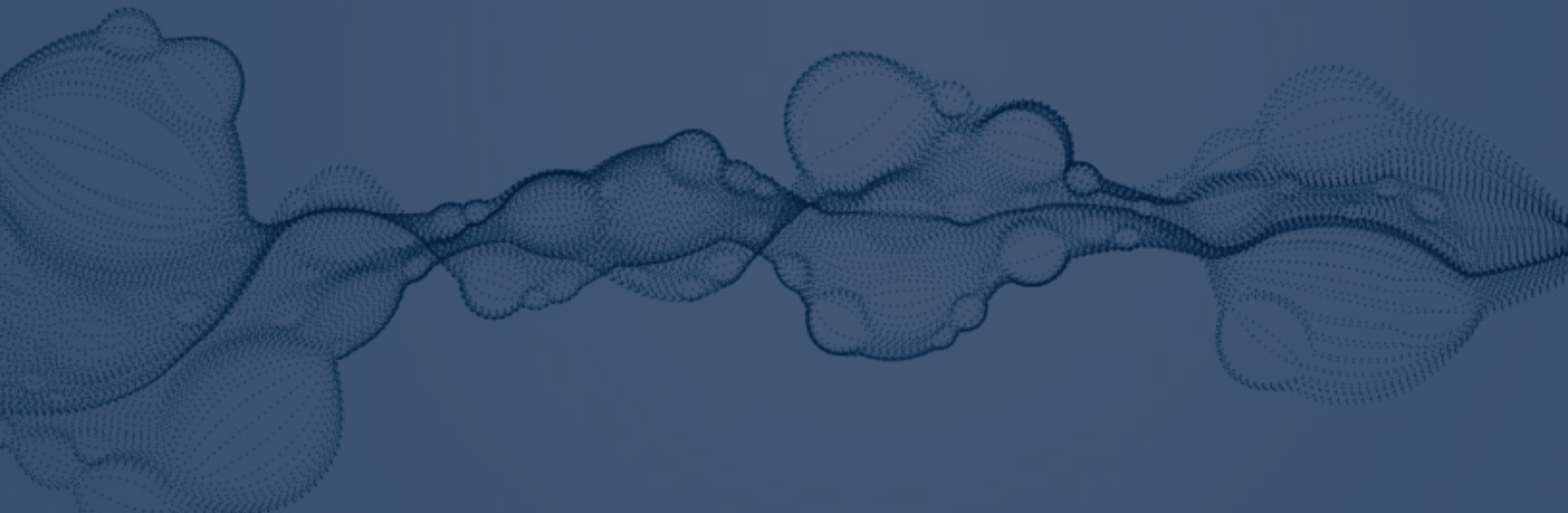
Patient ID	Asthma	Cardiovascular disease	Chronic lung disease	Confusion	Cough	Diabetes	Hypertension	Malignancy	Renal impairment	Shortness of breath
1	N	N	Y	N	N	N	N	N	N	N
2	N	N	N	N	N	N	Y	N	Y	N

Note: presentation focuses on application of NLP, and is not intended to address data curation/formatting steps to prepare final outputs (CSV in this case)

Why AWS for ICODA?

- Compared AWS vs Azure
- Negations
 - Example: “no history of A, B, or C”
 - Azure result (incorrect):
 - No history: A 
 - History: B, C 
 - AWS result (correct):
 - No history: A, B, C 

Note: Platforms provide capability to “fine-tune” models. There is no “one-size-fits-all” best technology, as it depends on the application.



Demo of Amazon Comprehend Medical



Additional Considerations

Quality Control

- Verify data going into NLP are read correctly
- High level QC of output data to determine expected number of records processed, file structure, etc.

- **NLP QC**

Should perform manual check of unstructured text versus NLP structured feature extraction (level depends on use case)

- (square root of n) + 1
- based on "**confidence score**" from NLP (ex: 0.70 or less gets QC'd)

Language Support

- Microsoft and Google platforms support MANY languages (>90)
- Often auto-translate to English before running NLP model
- Can reduce accuracy if not trained on native data language

AWS

Language
German
English
Spanish
Italian
Portuguese
French
Japanese
Korean
Hindi
Arabic
Chinese (simplified)

Medical “vocabularies”

- Foundational Model of Anatomy
- Gene Ontology
- HUGO Gene Nomenclature Committee
- Human Phenotype Ontology
- ICD-10 Procedure Coding System
- ICD-10-CM (available for US users only)
- ICD-9-CM
- LOINC
- MeSH
- MedlinePlus Health Topics
- Metathesaurus Names
- NCBI Taxonomy
- NCI Thesaurus
- National Drug File
- Online Mendelian Inheritance in Man
- RXNORM
- SNOMED CT (available for US users only)
- Standardized terminology key to consistency across data sources
- Can offer additional labor savings if system setup to auto-map

Cost

- On the decline, typically based on number of NLP calls and size of text data submitted
- Prices are roughly equivalent across the three major platforms
- [Amazon Pricing](#)
- [Microsoft Pricing](#)
- [Google Pricing](#)

Open Source NLP Software

- Most popular open source tools are in Python
- [PyTorch NLP](#) is one of best known

Pros	Cons
Software is free	Technical resource(s) may be needed
Sensitive data can stay at your site	May require additional computing power
	Complexity of initial setup
	Need labeled training data

Conclusion

- NLP
 - accuracy improving rapidly
 - cost decreasing
 - increasing accessibility to those with non-technical backgrounds
 - resistance to adoption lessening
 - can greatly boost amount of available structured data for research

Tips

- Use for large data where 100% manual QC not feasible
- Carefully match use case with QC plan – use cases directly driving decision making should have increased QC
- Use a pre-trained model, accuracy is generally good enough
- Eliminate manual steps wherever possible (ex: automated data transfers and APIs for connecting to text analytics platforms)
- Be aware of potential limitations of both technology and source data
- Once setup, can quickly apply for new data files
- Meet with stakeholders from other functions to determine their needs
- Consult experts as needed

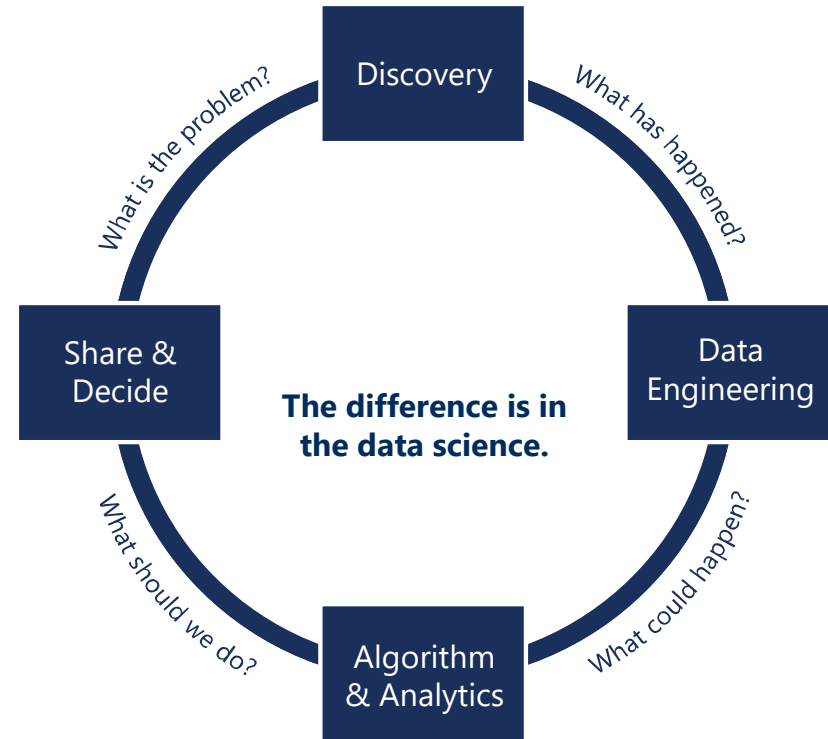
Data Science Process

Discovery: Identify where we can leverage data for making better business decisions, define goals and objectives of the analytic approach.

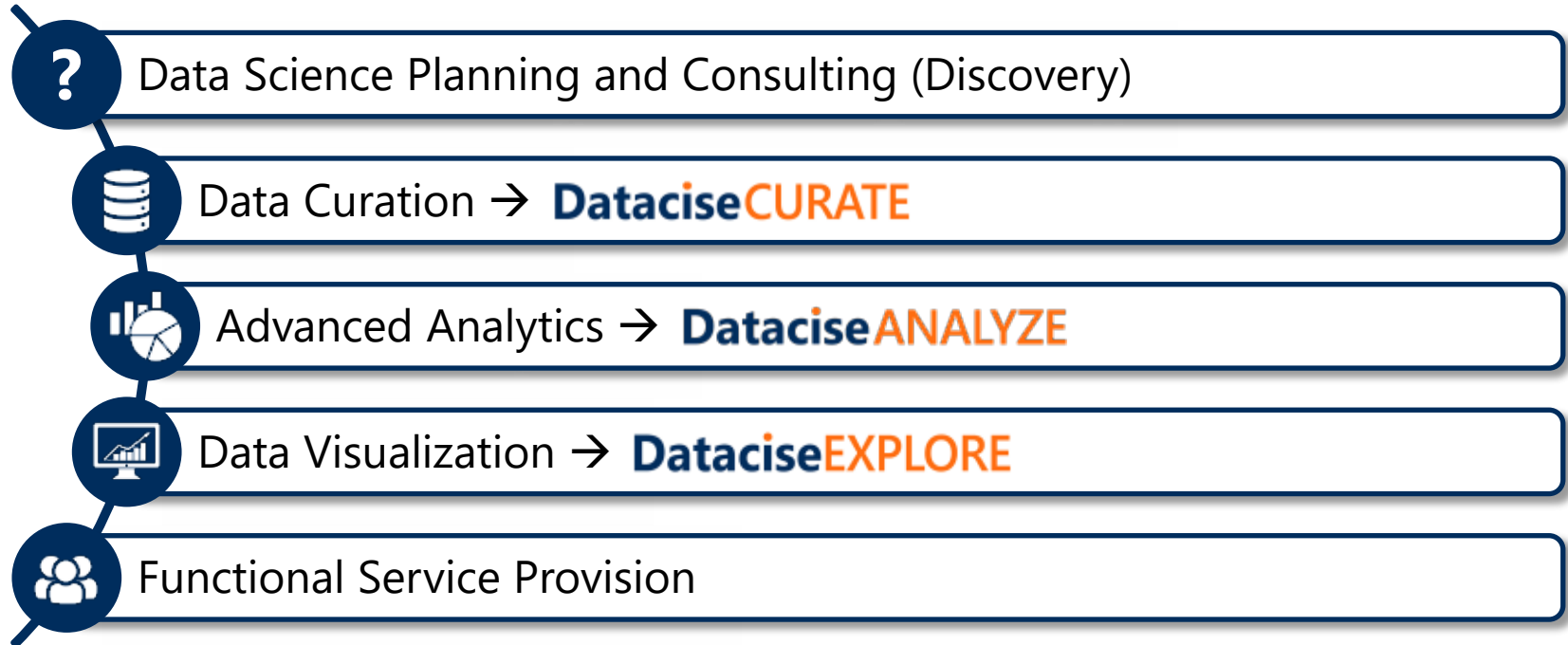
Data Engineering: Gather data requirements, collect and build understanding of the data, clean and prepare data for analysis.

Algorithms & Analytics: Analyze/refine the cleaned data to develop the model and analytics that pertain to the business problems.

Share & Decide: Share valuable results and develop actionable insights with stakeholders, get feedback and repeat methodology as necessary.



Data Science Service Summary





Thank you!
Any questions?

visit

www.mmsholdings.com



The difference is in the data.™

email

media@mmsholdings.com