How to Use Text Analytics in Healthcare to Improve Outcomes:
Why You Need More than NLP

Eric Just

HealthCatalyst
ignite outcomes improvement

Text Analytics in Healthcare

> *Eighty percent of clinical data is locked away in unstructured physician notes that can't be read by an EHR and so can't be accessed by advanced decision support and quality improvement applications."*

Peter J. Embi, MD, MS,
President and CEO
Regenstrief Institute

Unfortunately, more than 95 percent of health systems can't utilize this valuable clinical data because the analytics needed to access it is difficult, expensive, and requires an advanced technical skillset and infrastructure.

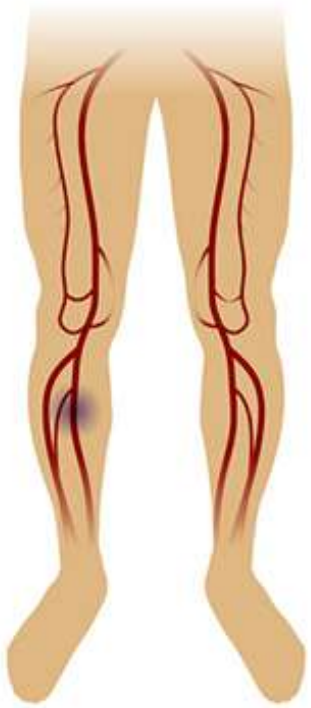
HealthCatalyst
ignite outcomes improvement

Text Analytics in Healthcare

For example, a team of analysts in Indiana set out to identify peripheral arterial disease (PAD) patients across two health systems.

The team hit a roadblock when they identified the structured EMR and claims data failed to identify over 75 percent of patients with PAD.

To better understand high-risk populations, such as PAD patients, health systems must leverage text analytics, including text search refined with natural language processing (NLP).



PERIPHERAL ARTERY DISEASE

BLOCKAGE OF THE BLOOD VESSELS SUPPLYING BLOOD TO THE LEGS AND FEET

HealthCatalyst
ignite outcomes improvement

Text Analytics in Healthcare

So far most health systems' use of text analytics has been limited to research departments within academic medical centers.

This presentation explains why text analytics in healthcare is important in all areas of the industry and demonstrates how health systems can leverage it.

It describes the four critical components of text analytics, from optimizing text search to pragmatically integrating text analytics system-wide.

HealthCatalyst
ignite outcomes improvement

The Current State of Text Analytics in Healthcare

Health systems simply aren't leveraging text analytics—Gartner estimates less than 5 percent do as of July 2016.

Most systems manually curate patient registries and rely only on coded data—an approach that significantly limits their understanding of patient populations.

HealthCatalyst
ignite outcomes improvement

The Current State of Text Analytics in Healthcare

Picking back up on the search for patients with PAD from above, let's understand why the analytics team was looking at this high-risk cohort.

PAD is a condition in which narrowed arteries reduce blood flow to limbs; it affects more than 3 million patients every year in the United States.

Patients with PAD are high-risk for coronary heart disease, heart attack, and stroke, and translates to both sicker patients with higher financial cost.

HealthCatalyst
ignite outcomes improvement

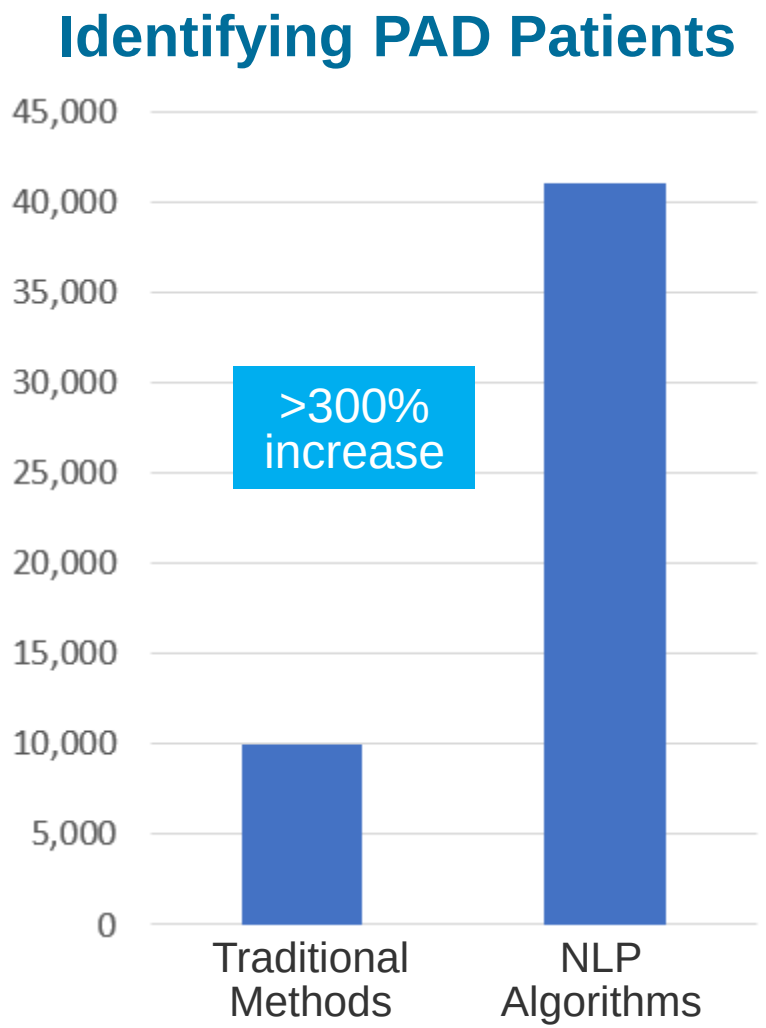
The Current State of Text Analytics in Healthcare

At first, the Indiana team identified less than 10,000 PAD patients using a traditional approach (ICD and CPT codes).

Hoping for a more complete patient list, the team tasked a sophisticated NLP group to write algorithms to identify more PAD patients.

Integrating text analytics led to the discovery of over 41,000 PAD patients.

There is a clear need to leverage unstructured text data in healthcare analytics despite the limited use today.



HealthCatalyst
ignite outcomes improvement

The Current State of Text Analytics in Healthcare

In the typical text analytics scenario, a data scientist writes an NLP algorithm to determine the number of PAD patients, validates the algorithm, and returns the result to the investigator.

In this scenario, only two people in the health system understand the PAD patient population.

Systems need more than just a handful of specialized algorithms serving small groups of stakeholders.

HealthCatalyst
ignite outcomes improvement

The Current State of Text Analytics in Healthcare

Typical Scenario

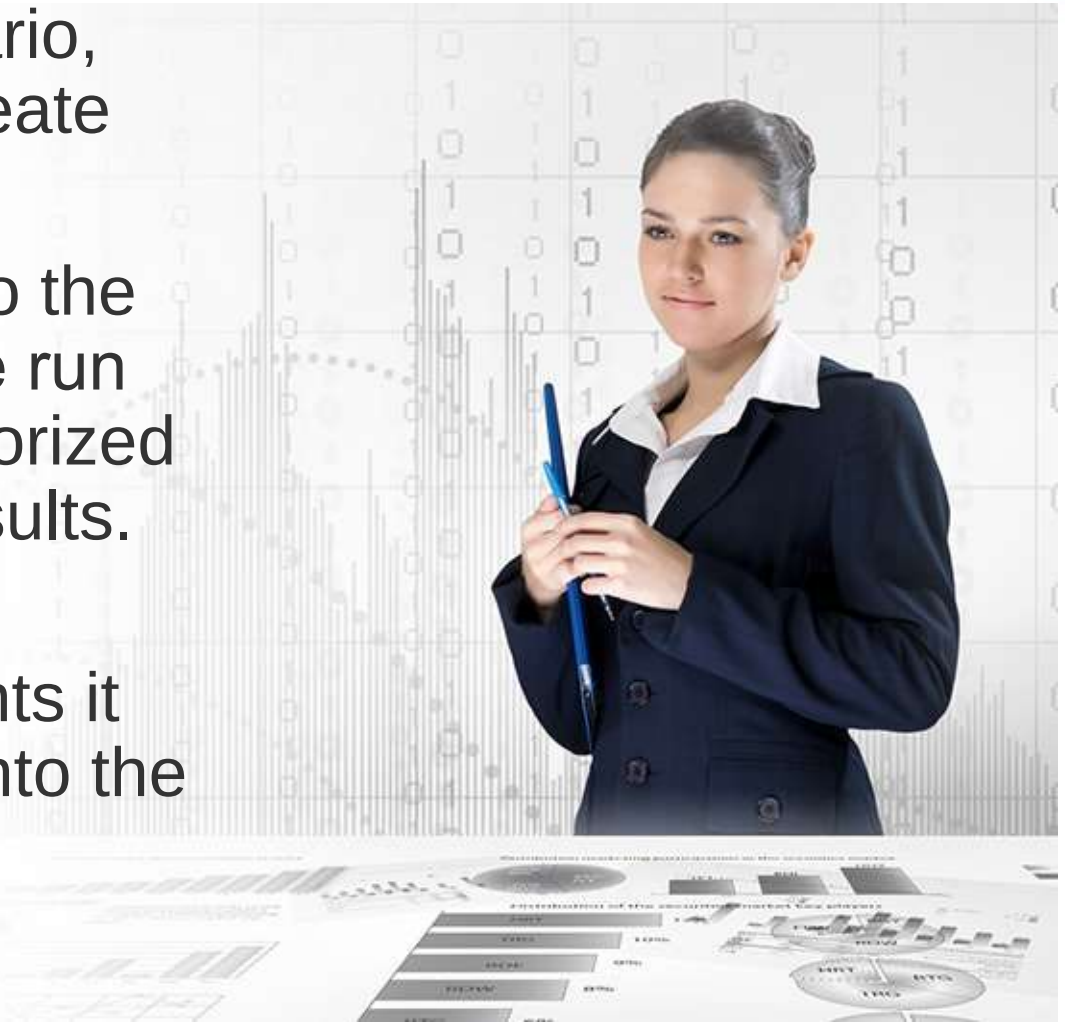
> *As a healthcare system administrator, I want to understand my high-risk population better. I want to find all patients with peripheral arterial disease (PAD). I know there are more patients than I was able to find by simply querying diagnosis and procedure codes."*

The Current State of Text Analytics in Healthcare

In a much better text analytics scenario, the data scientist does more than create and validate an algorithm.

The scientist deploys the algorithm to the system's analytics environment to be run on a nightly basis, allowing any authorized analytics user to make use of the results.

In this improved scenario, the health system knows how many PAD patients it has, how many new patients came into the system, and what the care gaps are.



HealthCatalyst
ignite outcomes improvement

The Current State of Text Analytics in Healthcare

Better Scenario

> *As a healthcare system administrator, I want to understand my high-risk population better. I want to find all patients with peripheral arterial disease (PAD). I know there are more patients than I was able to find by simply querying diagnosis and procedure codes."*

HealthCatalyst
ignite outcomes improvement

Google: The Text Search Gold Standard

When it comes to text search, Google is the gold standard.

People overwhelmingly believe that Google is easy to use, fast, and accurate.

It finishes our sentences for us and, most of the time, gives us exactly what we're looking for.

Even though Google search is simple and straight forward, it's a lot more complicated than it seems.

HealthCatalyst
ignite outcomes improvement

Google: The Text Search Gold Standard

Google analyzes every web page clicked, looks at other web pages that link to those pages, and examines synonyms too.

Google seems simple but is a sophisticated text analytics machine.

Google is an incredible text analytics tool for finding information in web pages online.

What if finding healthcare data in medical records were as simple as using Google?

Let's look at what it would take to build "Google for the Medical Record."

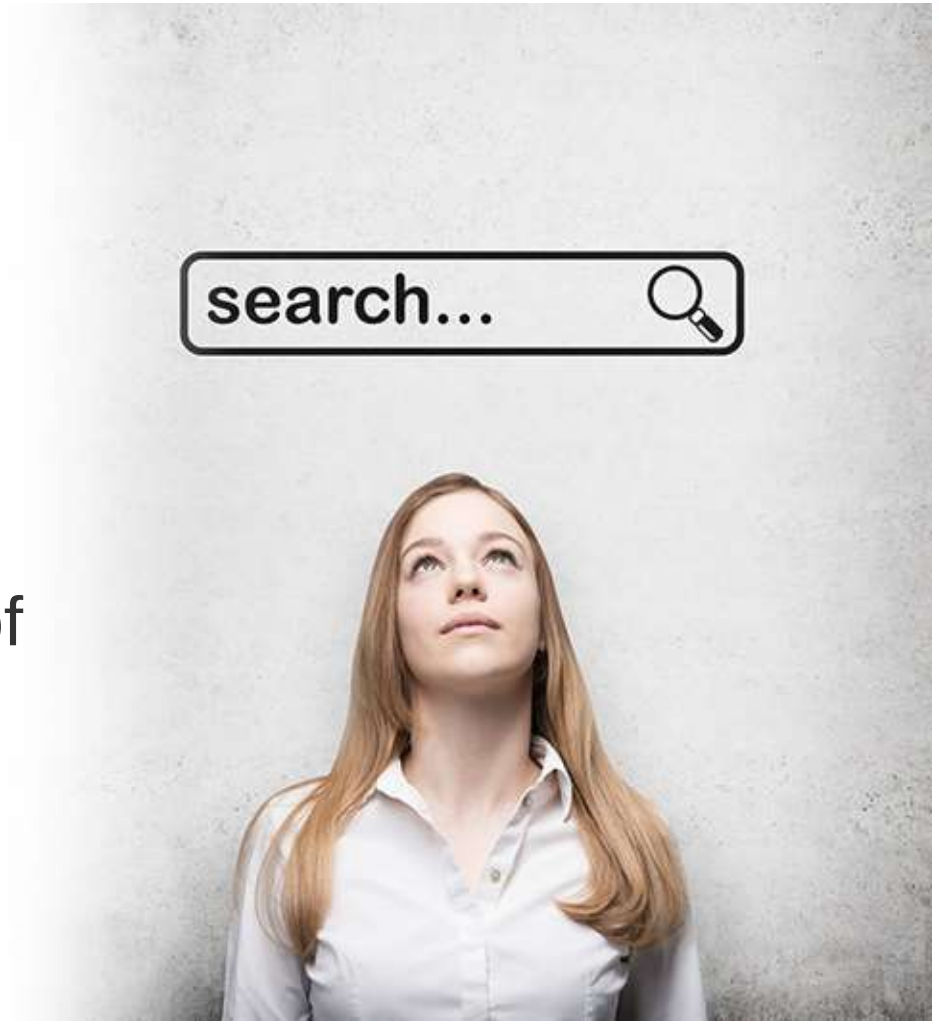
HealthCatalyst
ignite outcomes improvement

How Text Search Works

At the core of how text search works is an inverted index—an index similar to one at the end of a book that lists words in the text and where they appear.

Search engines index documents.

They read each document, break them into all the different words that appear in the documents, and create a sorted list of all the words.



HealthCatalyst
ignite outcomes improvement

How Text Search Works

For example the following slide features three simple clinical documents, starting with document 0.

- Document 0 states that the patient is a 67-year-old female with NIDDM (noninsulin-dependent diabetes mellitus) and hypertension.
- Document 1 states that the patient has no diabetes or hypertension.
- Document 2 states that the patient's mother and sister are diabetic.

HealthCatalyst
ignite outcomes improvement

The Current State of Text Analytics in Healthcare

The Basis of Text Search:
The Inverted Index

Document 0

Patient is a 67 year old female with NIDDM and hypertension.

Document 1

The patient has no diabetes or hypertension.

Document 2

Patient's mother is diabetic.
Patient's sister is diabetic.

Words	Document	Inverted Index
67	0	{(0,3)}
diabet	1,2	{(1,4),(2,3),(2,7)}
female	0	{(0,6)}
hyperten	0,1	{(0,8),(1,6)}
mother	2	{(2,1)}
niddm	0	{(0,8)}
no	1	{(1,3)}
old	0	{(0,5)}
patient	0,1,2	{(0,0),(1,1), (2,0),(2,4)}
sister	2	{(2,5)}
year	0	{(0,4)}

HealthCatalyst
ignite outcomes improvement

How Text Search Works

The second column in the inverted index (“document”) indicates in which document each word appears.

The third column (“inverted index”) indicates the position of the word in the document.

For example, the word “old” appears in document 0 in the fifth position.

Words	Document	Inverted Index
67	0	{(0,3)}
diabet	1,2	{(1,4),(2,3),(2,7)}
female	0	{(0,6)}
hyperten	0,1	{(0,8),(1,6)}
mother	2	{(2,1)}
niddm	0	{(0,8)}
no	1	{(1,3)}
old	0	{(0,5)}
patient	0,1,2	{(0,0),(1,1), (2,0),(2,4)}
sister	2	{(2,5)}
year	0	{(0,4)}

HealthCatalyst
ignite outcomes improvement

How Text Search Works

The term “diabet” also appears, which is a word stem.

Words that tend to have high level of variance are broken into word stems—a simpler version of the word—to broaden search capacity.

In this example, “diabet” is mapped to documents 1 and 2.

Words	Document	Inverted Index
67	0	{(0,3)}
diabet	1,2	{(1,4),(2,3),(2,7)}
female	0	{(0,6)}
hyperten	0,1	{(0,8),(1,6)}
mother	2	{(2,1)}
niddm	0	{(0,8)}
no	1	{(1,3)}
old	0	{(0,5)}
patient	0,1,2	{(0,0),(1,1), (2,0),(2,4)}
sister	2	{(2,5)}
year	0	{(0,4)}

HealthCatalyst
ignite outcomes improvement

Text Search Tools

Three Great Text Search Tools for Indexing Text and Providing Search Capability

- Tool #1 – Lucene
- Tool #2 – Solr
- Tool #3 – Elasticsearch

HealthCatalyst
ignite outcomes improvement

Text Search Tools

Written in 1999, Lucene is a Java-based API that provides the foundation for more advanced search engine capabilities.

Lucene creates and maintains the search index and does hit ranking and result sorting.

Most users access Lucene through one of two technologies, Solr and Elastsicsearch.

Text Search Tools

Solr is an open source, Java-based enterprise search tool which pioneered advances on top of Lucene starting in 2004.

Health systems can use any programming language to develop with Solr, which makes it easy to create an inverted index, simple web page, and search interface like Google.

Many leading technology organizations use Solr, such as Apple and NASA.

Text Search Tools

Elasticsearch is a newer open source enterprise search tool, written in 2010. Elasticsearch, like Solr, uses Lucene as a foundation.

Elasticsearch has built its own set of APIs and functionality over the inverted index.

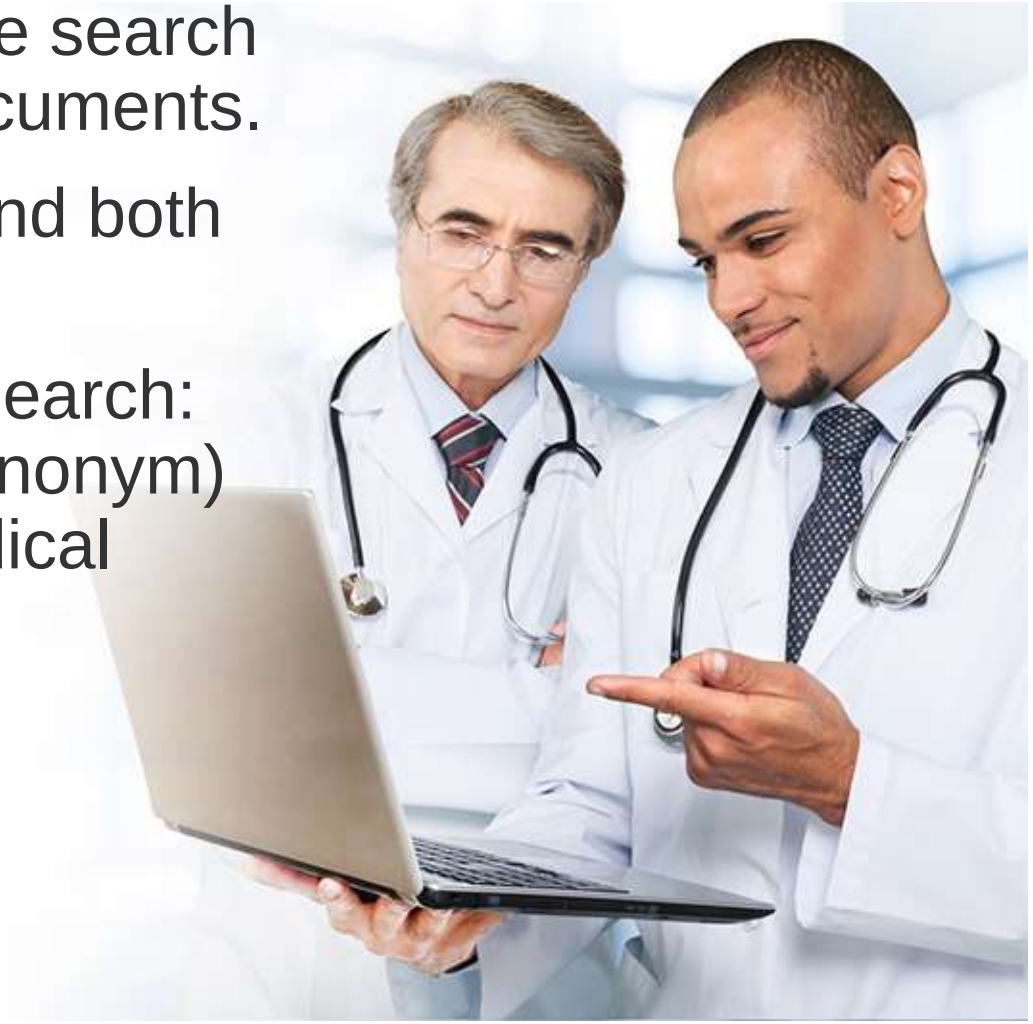
Netflix and Facebook use Elasticsearch.

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

As demonstrated on the next slide, the search for diabetes surfaced two indexed documents.

Using word stemming, the search found both diabetes and diabetic.

But there are two problems with this search: it missed the mention of NIDDM (a synonym) and neither result is relevant to a medical cohort query for diabetics.



HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

diabetes | Go

Results: 2 records, 0.0 ms

Document 2:
Patient's mother is **diabetic**. Patient's sister is **diabetic**.

Document 1:
The patient has no **diabetes** or hypertension.

Found both diabetes and diabetic (word stemming)

Missed mention of NIDDM (synonyms)

Neither result is relevant to a medical cohort query for diabetics (context)

Although the simple, familiar interface and fast results (generated by the inverted index) worked well, three text search must-haves for medical searches—display, medical terminologies, and context—are missing.

HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Must-Have #1: Optimize Results Display for Use Cases

We must create a solution that matches the needs of those using the data.

For example, users of data may not have permission to view PHI.

If we aggregate the results of the text analysis, then we provide users with access to the key results they need without compromising privacy.

HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Must-Have #2: Expand Search with Medical Terminologies

If you recall, the diabetes search didn't include NIDDM, a synonym of diabetes.

Health systems will produce more sophisticated analyses by leveraging medical terminologies to automatically expand synonym lists.

Providing a quick and easy way for users to identify synonyms valuable to their search will increase the quantity and accuracy of the results.

HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Must-Have #3: Refine Results with Context

Context matters.

To be useful for clinical applications (e.g., searching for genotype/phenotype correlations), retrieving patients eligible for a clinical trial, or identifying disease outbreaks, simply identifying clinical conditions in the text is insufficient.

Information described in the context of the clinical condition is critical for understanding the patient's state.

HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Must-Have #3: Refine Results with Context

If we are searching for patients with pneumonia, then searching for “pneumonia” without considering the context would result in identifying the following types of phrases:

- “ruled out pneumonia”
- “history of pneumonia”
- “family history of pneumonia”

None of these indicate the patient currently has pneumonia.

HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Must-Have #3: Refine Results with Context

ConText, an NLP pattern matching algorithm published in 2009, solves for these scenarios through enhancing text search and refining results by detecting conditions and determining if they are negated, historical, or experienced by someone else.

HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Must-Have #3: Refine Results with Context

The following slide shows how the ConText algorithm analyzes the sentence:

> *"No history of chest tightness but family history of CHF"*

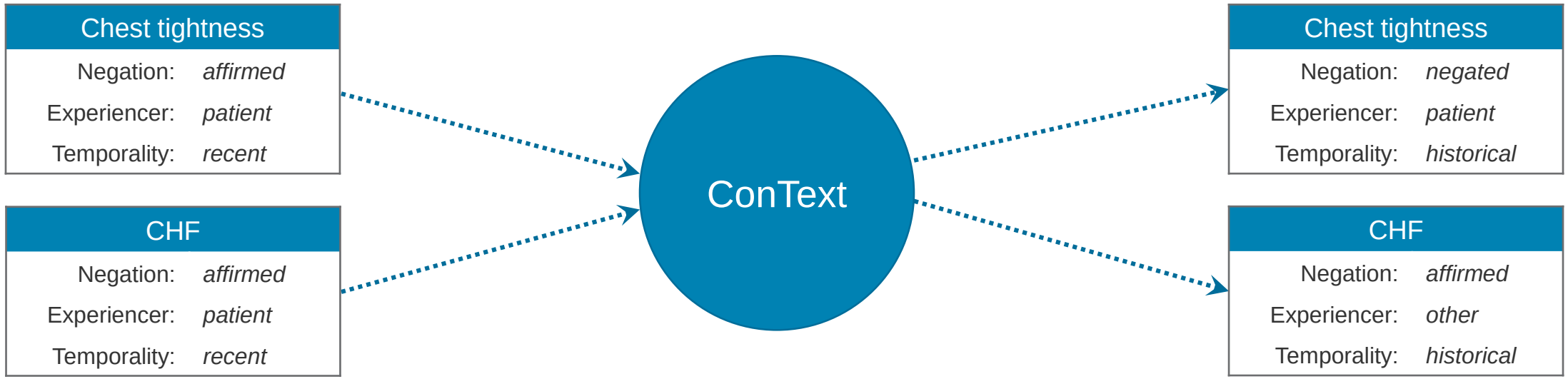
It looks for several things, including:

- Conditions (e.g., chest tightness)
- Negation triggers (e.g., the word "no")
- Historical triggers (e.g., the word "history")
- Termination (e.g., the word "but")
- Experiencer triggers (e.g., the word "family")

HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Wendy W. Chapman, David Chu, John N. Dowling
J Biomed Inform. 2009 Oct; 42(5): 839–851.



”No history of chest tightness but family history of CHF.”

Negation trigger
Historical trigger
Condition
Termination
Other experiencer trigger
Historical trigger
Condition
Termination

HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Applying Context and Extracting Values with an NLP Pipeline

NLP pipelines can be run using one of several Java-based, open source tools, such as Apache Unstructured Information Management Architecture (UIMA) and General Architecture for Text Engineering (GATE).

After starting a query with a clinical search and expanding that search to include meaningful terms, the next step is passing the search onto an NLP pipeline.

HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Applying Context and Extracting Values with an NLP Pipeline

In this example, a series of three analyses performed on the text that relate to and build on each other:

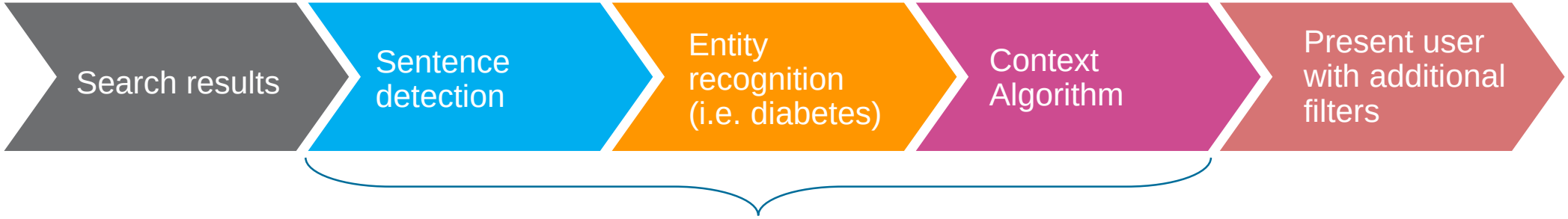
1. Sentence detection.
2. Entity recognition (e.g., diabetes).
3. Context algorithm.

HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Applying Context and Extracting Values with an NLP Pipeline

- Analysis of context uses a sentence as an operand
- Identifying sentences in clinical text is not straightforward
  - Have you ever seen punctuation in a clinical note?
- An NLP analysis pipeline ties it all together



NLP Pipeline Frameworks

- Apache Unstructured Information Management Architecture (UIMA)
- General Architecture for Text Engineering (GATE)
- Natural Language Toolkit (NLTK)

HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Applying Context and Extracting Values with an NLP Pipeline

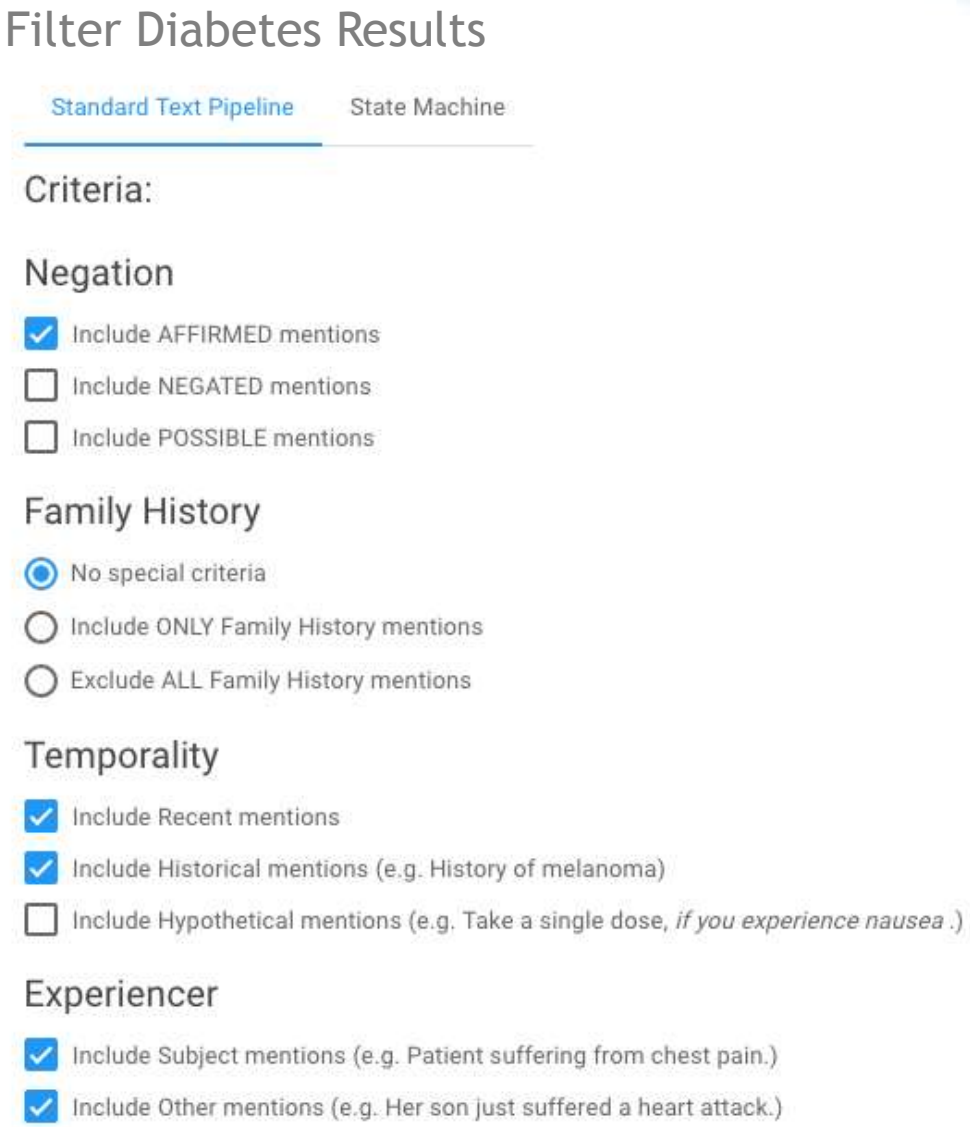
After the third and final step in this NLP pipeline (context algorithm), users are presented with additional filters in which they can choose their desired context (next slide).

When it comes to text analytics frameworks, flexibility to run algorithms based on a user's specific question is key.

The NLP pipeline should create this flexibility by expanding search capacity to refine results.

HealthCatalyst
ignite outcomes improvement

Context: Apply to Search Results



HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Applying Context and Extracting Values with an NLP Pipeline

The NLP pipeline should also extract values as demanded by each use case. Many health systems are burdened by regulatory reporting, especially when certain measures like ejection fraction are not stored as discrete values.

In lieu of automated reporting, health systems are obligated to engage team members to be “chart abstractors” to manually open thousands of patient charts to find ejection fraction values found in clinical notes.

HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Applying Context and Extracting Values with an NLP Pipeline

An NLP pipeline should not only identify when ejection fraction is documented in a note, but also save each value in such a way that the organization's analytics platform can utilize the discrete value in their automated reporting.

Other common extraction projects include aortic root size, PHQ depression scores, and ankle brachial index..

HealthCatalyst
ignite outcomes improvement

Three Text Search Must-Haves: Display, Medical Terminologies, and Context

Other Pieces to the NLP Pipeline: Extract Values

ef_phrase	qualifiers	ef_low	ef_high	ef_mid	ef_word
ejection fraction is at least 70-75	is at least	70	75	72.5	NULL
ejection fraction of about 20	of about	20	20	20	NULL
ejection fraction of 60	of	60	60	60	NULL
ejection fraction of greater than 65	of greater than	65	65	65	NULL
ejection fraction of 55	of	55	55	55	NULL
ejection fraction by visual inspection is 65	by visual inspection is	65	65	65	NULL
LVEF is normal	is	NULL	NULL	NULL	normal

```
\b(((LV)?EF)|(Ejection\s+Fraction))\s+(?<qualifiers>([^\s\d]+\s+){0,5})\(?(((?<ef_low>\d+)-(?<ef_high>\d+))|(?<ef_mid_txt>\d+)|(?<ef_word>([^\s]*?normal)|(moderate)|(severe)))
```

Always Validate the Algorithm

Quantifying algorithm accuracy is critical. Every algorithm needs a validation workflow, which typically includes four steps:

1. Build studies to review query results.
2. Assign team members to review results.
3. Randomly select records to represent study.
4. Highlight key words for easy chart review.

When evaluating NLP tools or devising an NLP strategy, health systems should ensure validation is either built into the tool or very easy to do.

HealthCatalyst
ignite outcomes improvement

Focus on Interoperability and Pragmatic Integration

Health systems need to combine text analytics with discrete data, and an enterprise data warehouse (EDW) is a great place to do that.

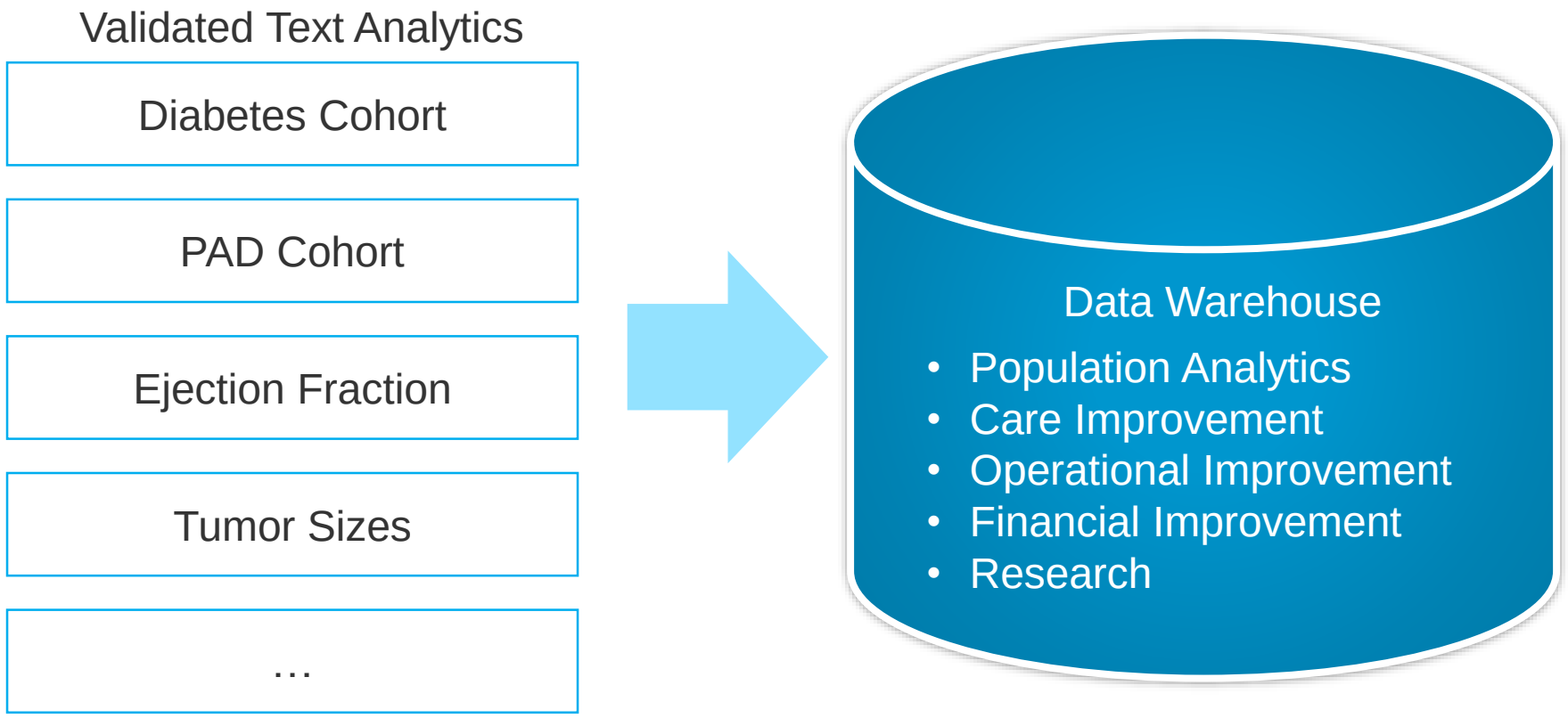
A siloed approach to text analytics won't work; make sure to integrate validated text analytics with all other analytics within the organization.

For example, a health system would put its diabetes and PAD cohorts into the EDW and make that data available system-wide. Health systems need a platform that can pull in data on a regular, timely basis to ensure it's current.

HealthCatalyst
ignite outcomes improvement

Focus on Interoperability and Pragmatic Integration

Text Analytics Must Be Interoperable!



Focus on Interoperability and Pragmatic Integration

Text analytics will open a new world of data analysis to health systems.

It's unlikely that anyone can identify the exact data uses and needs for text data in the next two, three, or five years.

Instead of trying to make lasting decisions on a data model up front, we recommend focusing on what's needed to perform timely, relevant analytics; healthcare analytics that quickly adapt to new questions and use cases.

HealthCatalyst
ignite outcomes improvement

Focus on Interoperability and Pragmatic Integration

This approach follows the Health Catalyst® Late-Binding™ architecture that allows for the flexibility that's so critical for the constantly evolving world of text analytics.

Health Catalyst Data Warehouse eBook.

FREE DOWNLOAD

HealthCatalyst
ignite outcomes improvement

How to Use Text Analytics to Improve Outcomes

The power of text analytics in healthcare to create precise patient registries is undeniable.

Rather than manually curating patient registries and relying solely on coded data, health systems should leverage data from coded data sets, NLP, and extraction projects to create precise patient registries.

Rather than limiting the use of text analytics to data scientists, architects, and analysts, health systems should make these tools widely accessible and empower the entire organization to leverage clinical text analytics.

HealthCatalyst
ignite outcomes improvement

How to Use Text Analytics to Improve Outcomes

Health systems can start using text analytics to improve outcomes today by focusing on four key components:

#1: Optimize Text Search (Display, Medical Terminologies, and Context)

#2: Enhance Context and Extract Values with an NLP Pipeline

#3: Always Validate the Algorithm

#4: Focus on Interoperability and Integration Using a Late-Binding Approach

HealthCatalyst
ignite outcomes improvement

How to Use Text Analytics to Improve Outcomes

This broad, all-encompassing approach to text analytics in healthcare will position health systems for clinical and financial success as they strive to create precise patient registries, enhance their understanding of high-risk patient populations, and improve outcomes.

Health systems should invest in a solution that not only solves today's problems, but can also be expanded to solve future use cases.

HealthCatalyst
ignite outcomes improvement

For more information:

Download our Free Healthcare Transformation Handbook

This is a multi-year, collaborative effort to provide a practical, best-practices handbook for clinical and operational leaders, as well as front-line caregivers who are involved in improving processes, reducing harm, designing and implementing new care delivery models, and undertaking the difficult task of leading meaningful change.

You'll learn

- Forces facing US healthcare
- Implementing systematic quality improvement
- Creating a data-driven culture
- Best practices in clinician-driven analytics
- Best practices in evidence-based content
- Best practices in system-wide deployment
- New paradigms of cost and quality
- Key stumbling blocks to avoid
- Success stories across the industry
- Kindle and Paperback versions also available

download the handbook

FREE Download / Kindle Version ($2.99) / Paperback ($31.73)

"This book is a fantastic piece of work"

– Robert Lindeman MD, FAAP, Chief Physician Quality Officer

HealthCatalyst
ignite outcomes improvement

Link to original article for a more in-depth discussion.

How to Use Text Analytics in Healthcare to Improve Outcomes—Why You Need More than NLP

More about this topic

Regenstrief Institute and Health Catalyst Team to Reveal Hidden Meaning in Clinical Data for Better Patient Care – Health Catalyst News

Healthcare Analytics Adoption Model: A Framework and Roadmap (white paper)
Dale Sanders, Executive VP of Software

Healthcare Analytics: Realizing the Value of Health IT
Brian Ahier – Guest Blogger

3 Reasons Why Comparative Analytics, Predictive Analytics, and NLP Won't Solve Healthcare's Problems – Dale Sanders, Executive VP of Software

Big Data in Healthcare Made Simple: Where It Stands Today and Where It's Going
Doug Adamson, Chief Technology Officer, VP

HealthCatalyst
ignite outcomes improvement

Other Clinical Quality Improvement Resources

Click to read additional information at www.healthcatalyst.com

Eric Just joined the Health Catalyst family in August of 2011 as Vice President of Technology, bringing over 10 years of biomedical informatics experience. Prior to Health Catalyst, he managed the research arm of the Northwestern Medical Data Warehouse at Northwestern University's Feinberg School of Medicine. In this role, he led the development of technology, processes, and teams to leverage the clinical data warehouse. Previously, as a senior data architect, he helped create the data warehouse technical foundation and innovated new ways to extract and load medical data. In addition, he led the development effort for a genome database. Eric holds a Master of Science in Chemistry from Northwestern University and a Bachelors of Science in Chemistry from the College of William and Mary.

HealthCatalyst
ignite outcomes improvement