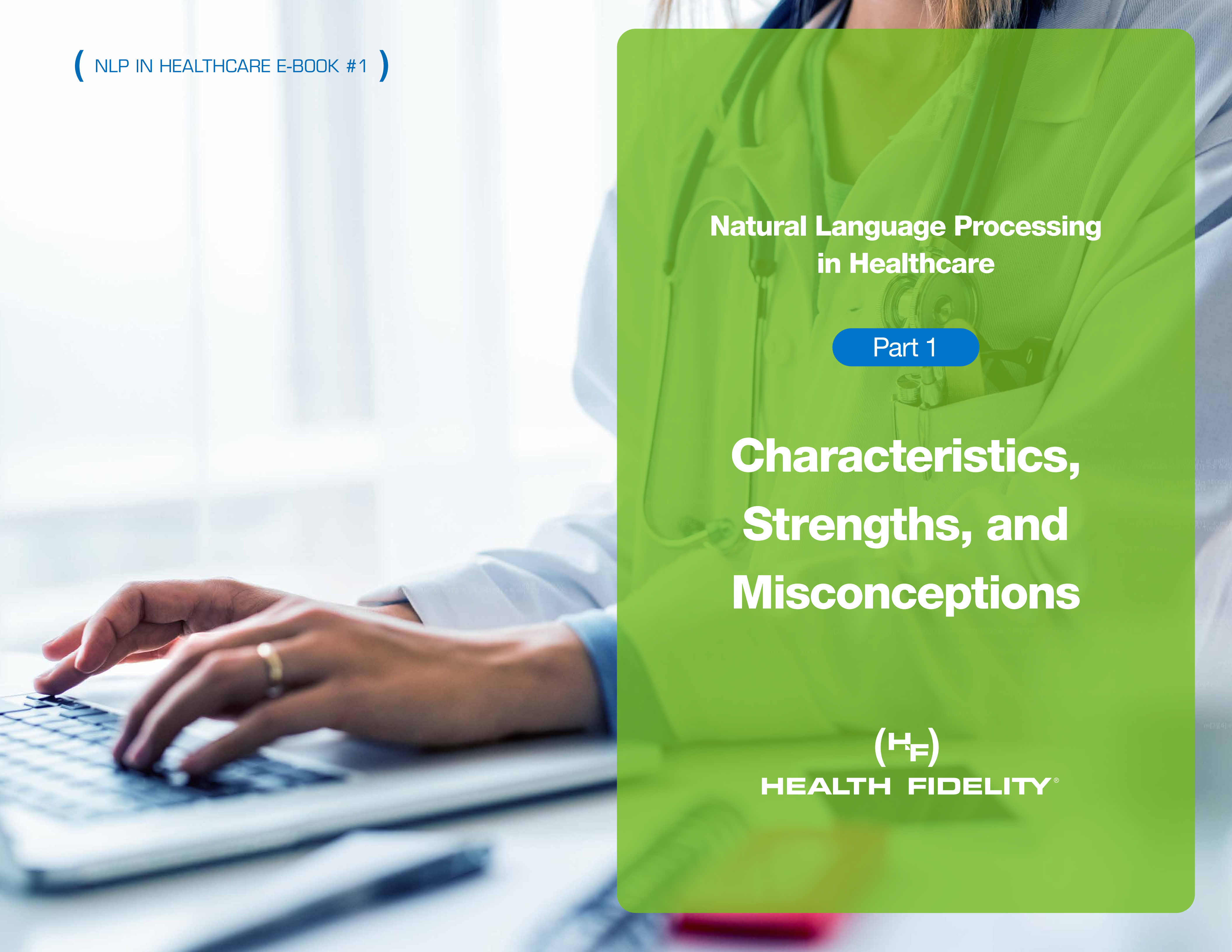
(NLP IN HEALTHCARE E-BOOK #1)

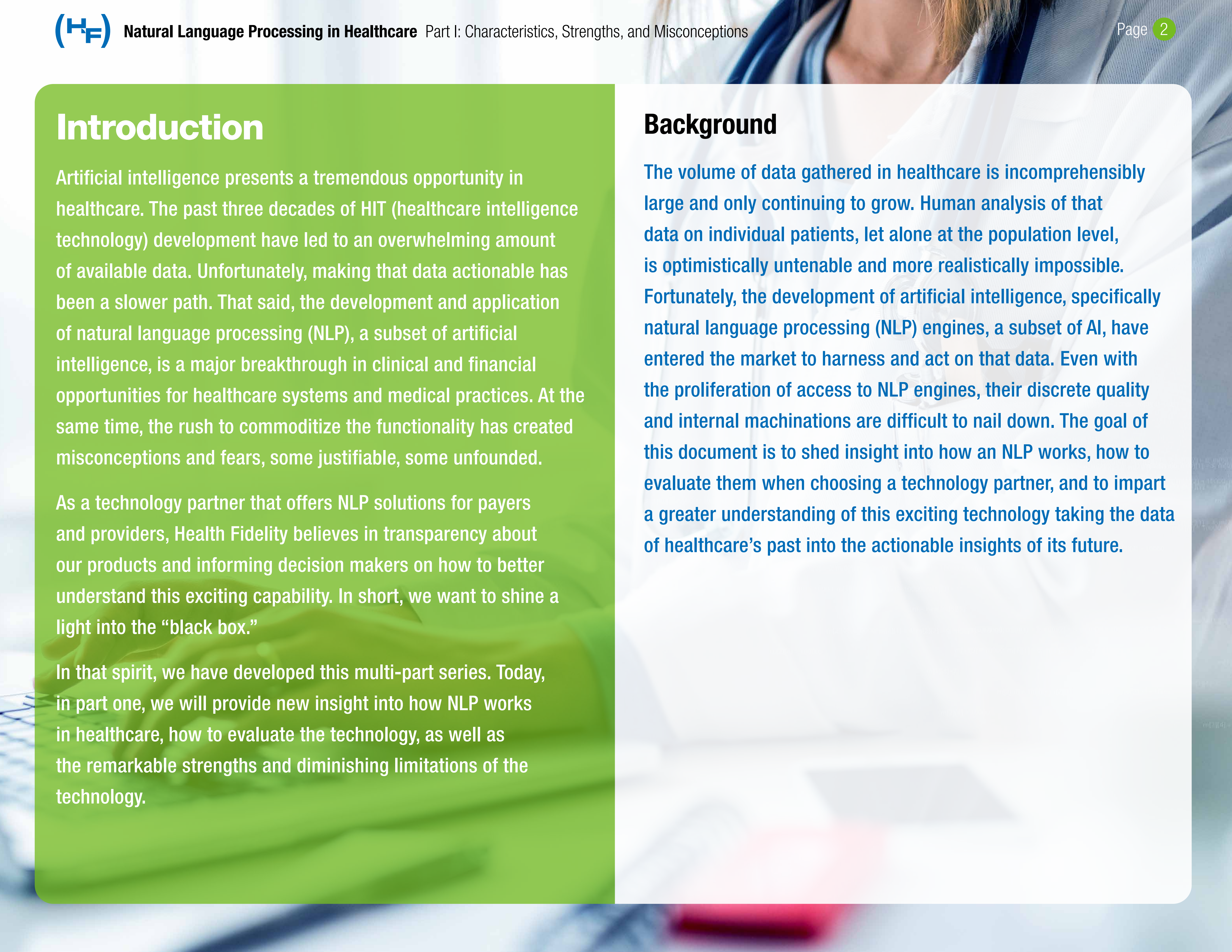


Natural Language Processing in Healthcare

Part 1

Characteristics, Strengths, and Misconceptions

HEALTH FIDELITY®



Introduction

Artificial intelligence presents a tremendous opportunity in healthcare. The past three decades of HIT (healthcare intelligence technology) development have led to an overwhelming amount of available data. Unfortunately, making that data actionable has been a slower path. That said, the development and application of natural language processing (NLP), a subset of artificial intelligence, is a major breakthrough in clinical and financial opportunities for healthcare systems and medical practices. At the same time, the rush to commoditize the functionality has created misconceptions and fears, some justifiable, some unfounded.

As a technology partner that offers NLP solutions for payers and providers, Health Fidelity believes in transparency about our products and informing decision makers on how to better understand this exciting capability. In short, we want to shine a light into the "black box."

In that spirit, we have developed this multi-part series. Today, in part one, we will provide new insight into how NLP works in healthcare, how to evaluate the technology, as well as the remarkable strengths and diminishing limitations of the technology.

Background

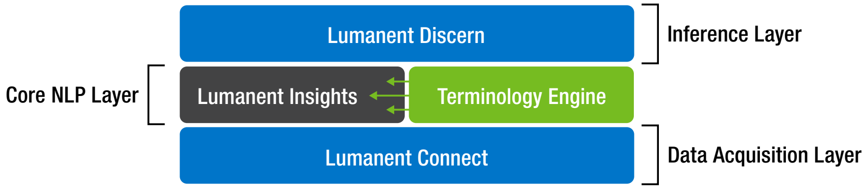
The volume of data gathered in healthcare is incomprehensibly large and only continuing to grow. Human analysis of that data on individual patients, let alone at the population level, is optimistically untenable and more realistically impossible. Fortunately, the development of artificial intelligence, specifically natural language processing (NLP) engines, a subset of AI, have entered the market to harness and act on that data. Even with the proliferation of access to NLP engines, their discrete quality and internal machinations are difficult to nail down. The goal of this document is to shed insight into how an NLP works, how to evaluate them when choosing a technology partner, and to impart a greater understanding of this exciting technology taking the data of healthcare's past into the actionable insights of its future.

Inside NLP Mechanics

The Commonality of Assembly

NLPs traditionally consist of two separate but integral components, the NLP itself, to translate unstructured data into something more traditional software can understand, and the inference engine, a layer that sits on top of the NLP and takes action based on that translated data.

Logically, and often also physically, an NLP system can be split into the “core NLP”, and an “inference layer”. The latter is also commonly referred to as “post-processing logic”.



Lumanent Discern

Inference Layer

Core NLP Layer

Lumanent Insights

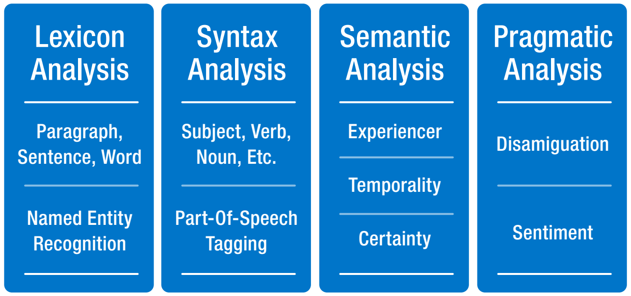
Terminology Engine

Lumanent Connect

Data Acquisition Layer

A core NLP engine is a complex technology that can take years, even a decade to develop until it can prove itself useful for medical coding purposes. All the knowledge that it needs to have in terms of handling document variations; recognition of sections and sentences, grammatical parsing, robustness around the variability in word choices or expressions, resistance to a multitude of errors like spelling abnormalities, syntactical mistakes in the spoken language, etc. is usually built into that core engine.

NLP divides language into its complex components:



Lexicon Analysis

Paragraph, Sentence, Word

Named Entity Recognition

Syntax Analysis

Subject, Verb, Noun, Etc.

Part-Of-Speech Tagging

Semantic Analysis

Experiencer

Temporality

Certainty

Pragmatic Analysis

Disamiguation

Sentiment

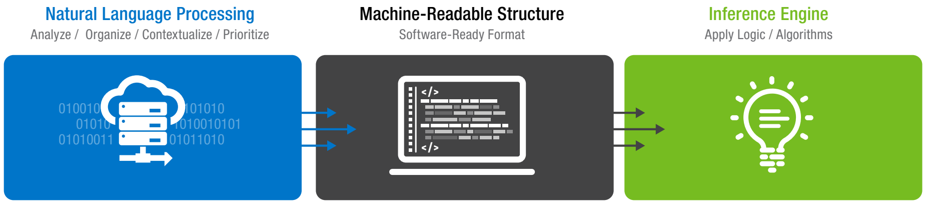
Through that division,it is able to understand grammar, syntax, context, and intent.

Equally as important is the core engine’s capability to map the clinical language to the appropriate standardized clinical concepts. These concepts are often based on the SNOMED (Systematized Nomenclature of Medicine) ontology, a set of clinical concepts and terminologies and their relations to each other. This is why there is popular use of the term “SNOMED annotator” for many core NLP engines. For coding purposes, SNOMED codes are then mapped to ICD10 codes. SNOMED is not the only option, though. There are also proprietary ontologies that have been commercially successful over the years.

As a consequence of the significant development investment, some companies will choose to license existing NLP engines and only develop their own inference layers. Health Fidelity can be seen as an example of this approach; our core NLP engine was developed at Columbia University under the leadership of Dr. Carol Friedman, professor of biomedical informatics.

The Inference Layer and Post Processing Logic

Multiple Layers of NLP Functionality



Natural Language Processing

Analyze / Organize / Contextualize / Prioritize

Machine-Readable Structure

Software-Ready Format

Inference Engine

Apply Logic / Algorithms

The inference layer is part of the custom differentiation in the industry. It consumes the clinical language annotations made by the core NLP. Inference layers can consume a wide array of data sources, but often function at a document level. That said, there are innumerable approaches here, most of which are the intellectual property of the company that developed the inference layer, e.g. Health Fidelity, to implement the semantic analysis of the clinically relevant facts that have been extracted by the core engine.

This is called the “post-processing” logic, where the meaning of the free-text data is implied or interpreted by a collection of extracted structured clinical facts and is pieced together using linguistic nuances that the NLP is able to recognize. Put more simply, the core NLP reads and understands the raw data and assembles it something the document-level inference layer can understand and further refine for users.

The domains of nuances understood here include, but are not limited to:

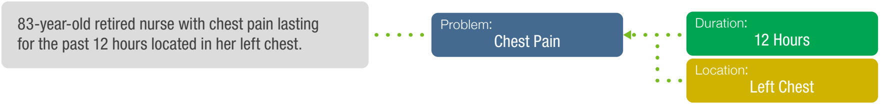
- Specificity (*which foot has a chronic injury?*),
- Level of certainty (*how confident is the AI?* This becomes a major factor in other functionality),
- Figurative speech elements (*did a doctor use a metaphor in her documentation?*),
- And recognizing/differentiating acronyms.

A similar, but separate element commonly discussed at this level, temporality (e.g., the permanent or impermanent nature of a condition in a patient), is actually determined earlier at the NLP level.

The output then from post-processing in a risk adjustment setting would be the NLP engine's best choice for an ICD-10 code (or codes) that represent the best possible match to all the available data.

Here are some examples of the complexities in clinical language that the system must handle:

Contextual Attribute Assignment – identifies clinical problems (i.e., diagnoses) and assign attributes such as duration, location, disease course, severity, and acuity. In addition to diagnoses, HF REVEAL can identify other clinical concepts such as medical procedures and events and form similar associations.

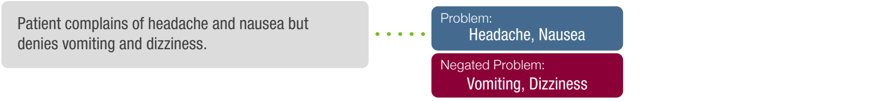

83-year-old retired nurse with chest pain lasting for the past 12 hours located in her left chest.

Problem: Chest Pain

Duration: 12 Hours

Location: Left Chest

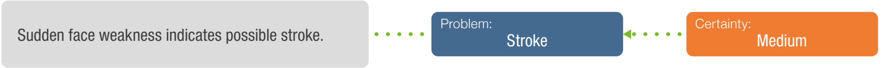
Negation – interprets grammar to understand which clinical problems were actually experienced by the patient versus which ones were not experienced.


Patient complains of headache and nausea but denies vomiting and dizziness.

Problem: Headache, Nausea

Negated Problem: Vomiting, Dizziness

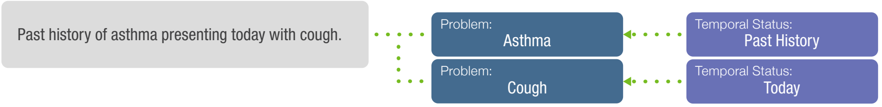
Certainty – assigns a level of certainty to clinical problems and findings based on the statement's definitiveness.


Sudden face weakness indicates possible stroke.

Problem: Stroke

Certainty: Medium

Temporality – understands the timing of problems, findings, procedures, and events.


Past history of asthma presenting today with cough.

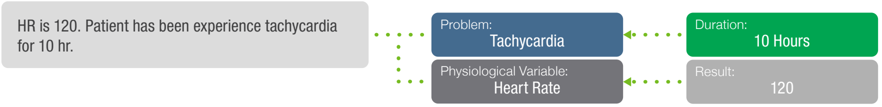
Problem: Asthma

Temporal Status: Past History

Problem: Cough

Temporal Status: Today

Word Sense Disambiguation – removes uncertainty of meaning from ambiguous words using context and syntax (e.g., "HR" as hours vs. heart rate).


HR is 120. Patient has been experience tachycardia for 10 hr.

Problem: Tachycardia

Duration: 10 Hours

Physiological Variable: Heart Rate

Result: 120

An Aside on Machine Learning

A growing landscape of non-NLP technologies, such as Machine Learning (ML) and Neural Networks (NN), are making inroads in augmenting the performance of traditional NLP systems. At current, while Health Fidelity deploys some ML techniques in discerning opportunity and refinement, major development is still undertaken by a team of experts at a local and global level for our clients. This is for two key reasons. First, our NLP is at work in a healthcare setting where doctors using our suite of functionality are holding vulnerable lives in their hands. While we have full faith and confidence in what we have built and shared with the industry, based on the current state of the art, fully automating any AI in this setting is, in our opinion, acting incautiously with those lives entrusted to our partners.

Next, our development strategy involves empowering, not replacing, human experts. This isn't just a value we hold dear, it's also a practical consideration of the current challenges of machine learning. Machine learning being deployed today means any AI becomes an actual black box technology. Rather than a black box in the proprietary/secret use of the term, with machine learning, humanity is removed from the development process. Confidential approaches to solving problems are one thing, but when a technology partner genuinely cannot explain why their solution does what it does, it creates issues up to and including making QA nearly impossible, exposing security concerns, and setting the groundwork for an inability to correct for bias or provide the full picture of evidence for its decision making internally and for partners.

Machine learning in its current form still requires a leap forward in both its capabilities (letting clinical and revenue cycle experts guide its shape depending on use) and its ability to describe its own iterative processes (creating opportunities for expert vetting) before we'd commit to its wide use. By letting experts and partnerships shepherd the development of our NLP, bolstered by the enormous volume of incoming clinical documentation, and tempered by the advanced maturity of the NLP itself, we feel this is the best course of action to live our values and provide the best solution possible.

NLP Performance Measurement

All of the logical system segments described above (core NLP, post-processing) have their own relative strengths and weaknesses. It's why they are both ultimately necessary to build out a useful engine. However, what always should be kept front and center is that expectations on NLP in the coding industry always boil down to improvements in coding accuracy and coding productivity. In addition to the coding mechanics within the workflow tool itself, both the former and the latter are heavily influenced by the two primary performance indicators of NLP systems, namely **recall** and **precision**.

Precision & Recall

Recall is defined as the set of known favorable outcomes (a.k.a. "true positives"), as viewed relative to the total set of favorable outcomes that includes both known and unknown ones (i.e. true positives and false negatives). In coding terms, known favorable outcomes are correctly identified ICD-10 diagnoses by the NLP, as a human coder will agree with them. Unknown favorable outcomes are ICD-10 diagnoses that the NLP failed to identify, even though a human coder would have assigned them.

Simply put, the greater the recall figure, the more useful to coding applications the overall NLP system. Or, the greater the recall, the more "complete" the NLP findings are considered to be. A low recall figure would represent one of the inconveniences that the coder would have to accept when working in an NLP system, because it would cause them to spend time manually identifying a large number of ICD-10 codes they wished the NLP had alerted them to from the beginning. The best achievable recall measure is 100%, corresponding to zero false negatives.

Precision measures the degree to which known unfavorable outcomes (false positives) pollute the total set of known outcomes and is therefore defined as the true positives divided by all known outcomes (the sum of true positives and incorrect codes).

In coding terms, known unfavorable outcomes are ones that are incorrectly identified ICD-10 diagnoses by the NLP, because a human coder will not agree with them. **In short, the greater the precision figure, the more accurate the overall NLP system. Or, the greater the precision, the less "noisy" the NLP output is considered to be.** A low precision figure represents one of the necessary evils that the coder would have to accept when working in an NLP system, because it causes them to be alerted to a large number of ICD-10 codes they end up rejecting after manual review. The best achievable precision measure is 100%, which corresponds to zero false positives.

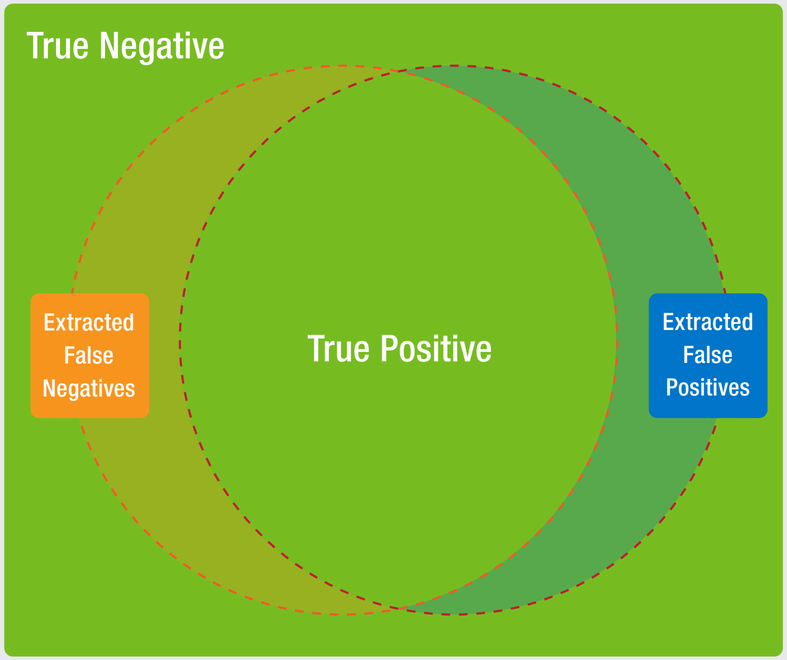
Precision is the ratio of correctly identified conditions to the total identified conditions. Precision can be thought of as a system's accuracy (i.e. "How accurately did the system surface the correct suggestions?")

Recall is the ratio of correctly identified conditions to all conditions the patient suffers from. Recall can be thought of as the system's completeness (i.e., "How many of the correct answers did the system get?")

Correctly identified conditions : total identified conditions = Precision

Correctly identified conditions : patient's total conditions = Recall

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative



Recall and precision always counterbalance each other. In any NLP system the effort of enhancing one measure will generally have a depressing effect on the other. Reducing this inverse reciprocal impact to a minimum is one of the objectives of NLP engineers. Combatting this effect can be done on all system levels, including the core NLP, post-processing, and by applying any augmenting technologies to this end.

For example, to support Health Fidelity's precision and therefore productivity for its users, Post Encounter Review validates any found ICD-10 coding opportunities against the MEAT (Measure, Evaluate, Assess, Treat) criteria that are commonly applied in coding for risk adjustment. This affects the precision of suggested ICD-10 codes favorably in any subsequent uses of Retrospective Review by payers, supporting the entire revenue cycle because of fewer false positives and fewer rejections in the workflow.

One major complicating factor when it comes to calculating precision and recall is the dynamic nature of the known and unknown true and false positive outcomes, i.e. the mix of accurate and inaccurate ICD-10 codes being shown to the coder. The real-time actions taken within the workflow influence the codes shown to coders through the use of suppression logic (more on that, later).

Other complicating factors include:

- The differences among customers in how they interpret the coding guidelines, which leads to different standards of ICD-10 codes being deemed either "right" or "wrong" in a given context.
- Rejections of diagnoses identified by the NLP on the basis of the lack of a valid physician signature, or the document being ineligible for risk adjustment coding.

At the time of this publication (Q4 2019) Health Fidelity's NLP achieves peak recall rates around 95%, while peak precision has been observed at around 50%. While 50% seems low relative to the high recall, in the scale of NLP performance with a priority on discovering previously unavailable/unknown data points in unstructured documentation, this is a strong return.

Precision and F1 Score

While the precision and recall measures are commonly used in NLP performance considerations on the customer's business level, it may happen on occasion that one of their combined metrics – the "F1 score" – may be mentioned as well. The F1 score is the harmonic mean of precision and recall.

$$2*P*R/(P+R)$$

Where P is the precision figure, and R represents recall

This suggests that any NLP systems that are deliberately built to perform well on recall, and therefore perform relatively poorly on precision, (remember, precision and recall have a push/pull relationship with each other), will have an F1 score that more resembles the poor precision number than the strong recall.

This can create a perception of an overall weaker performance than is necessarily true, because these characteristics may have been deliberately adapted to an organization's business preferences.

For example, creating the opportunity to find less likely but still valid suggestions provided to coding experts without overwhelming those experts. In doing so, the potential for an ever-greater ROI is uncovered, as well as opportunities to provide care for patients with overlooked conditions.

Prioritizing NLP Output by Use Case

Because of the popular consideration of percentage values, there can be some confusion when both recall and precision are not at or near 100%. While the last section explains the relationship between the two, consider the following examples of how deploying different levels of precision and recall can align with client priorities.

Payers

Traditionally, a high recall is prioritized over precision for payers. There are two reasons for this:

- Payers often have a campaign-based risk adjustment coding process, where all eggs of effort are placed in the proverbial recall basket. The reason for this is that payers will be coding member records as a second or higher pass. In that context, maximizing recall is the best property of an NLP workflow tool to uncover any remaining HCCs that are supported but unconfirmed. Exposing coders to very low NLP precision, and the additional coding review cycles and reduced coding productivity that this entails, is a consequence that payers are more than willing to accept for the benefit of maximizing recall. The incremental revenue that results from the second and higher-pass coding reviews in the NLP-enabled workflow tool outweighs any additional cost of paying for productivity penalties on coding resources due to poor precision.
- Payer coder expertise allows a false positive ICD-10 code to be rejected quickly and easily, and that rejections are a small price to pay for the benefit of being alerted to many true positive coding opportunities by the NLP that would have never been found by the coders in a workflow devoid of NLP support.

Providers

For the most part, providers thrive with a more equitable balance between precision and recall.

The relatively lower recall / higher precision requirements by providers are rooted in productivity targets on the level of encounter reviews per time unit, rather than the level of HCC or procedure capture rates. Inpatient coders (either FFS or risk adjustment focused), are very expensive experts in the operational budget. As a result, there is a specific need to display a minimum level of productivity during their first-pass reviews, in part due to the discharged-not-final-billed (DNFB) requirements. The potential loss of HCC related revenue stemming from a compromise in recall performance for the betterment of precision is more than compensated by the benefits of timely claim submissions for inpatient episodes, as well as additional clinical review where applicable to elevate quality of care.

The Complexity of Comparisons

The comparison of different Natural Language Processing (NLP) engines has always been a difficult exercise with the lack of reusable precedents, and ambiguities in how a fair comparison could be made.

For example, some NLP systems work well under specific conditions, but are not ideal for broad applications. Adding to the complexity is the fact that every commercial NLP engine is a very closely guarded secret. As a result, no third party, credible qualitative or quantitative performance measures are useful in a comparison could one easily be obtained.

However, there are some basic questions that can always be considered when evaluating an NLP.

“What use case was it developed for?” In the examples above, payer and provider partners use the same NLP as two healthcare centric use cases seeking the same elemental information, but even they have extensive, differentiated engineering on top of them. Too many steps further away from each other may require separate NLP with different training.

“How does it understand grammar, context, syntax, intent?” The specificity of a need necessitates the specification of its tools. In this case, any NLP raised on legal documentation would be incapable of effectively parsing clinical information. Beyond that, it’s critical to understand how independent analysts evaluate it and what its precision and recall statistics are for obvious reasons.

Finally, “How does it improve results over time?” “How is its computational performance?” Both data points will inform performance over time, but over-reliance on machine learning, for example, may lead to faster continuous development, but will definitely make the system more error prone and incredibly difficult to troubleshoot.

Final Thoughts

NLP is an incredibly exciting emergent technology in healthcare, the first finally able to make good on the promise of healthcare IT lifting productivity rather than simplifying it and burdening clinicians with non-clinical administrative work. At the same time, because of its “black box” nature at the inference level (post-processing), it can leave clients unsure of how much trust they can place in an NLP solution.

At the same time, this has allowed vendors with lesser products, or at the very least, less purpose built solutions, to sell without educating their partners.

Much like NLP itself, the goal today has been to shed some light into that darkness and provide interested parties the additional insights they need to move forward with confidence and a greater understanding of how to succeed.

Next Steps

This document is not, however, the final word on the subject. NLP development is a full-fledged field of expertise unto itself. Should you find yourself with questions that are not answered here, nor in any additional books in this series, please feel free to contact us. We always appreciate an opportunity to discuss this emerging field of applied technology in healthcare, and how it can support so many separate goals, from financial to clinical.

Summary

- Clinical language is complicated and full of specialized terminology
- NLP for clinical use-cases must be trained on clinical data
- Ability to handle a large range of patterns
  - – Organizations
  - – Specialties
  - – Physician-variation
- Even within healthcare, there is a wide variety of use cases
- Additional fine-tuning is required for each use case

Conclusion

NLP is an incredibly exciting emergent technology in healthcare, the first finally able to make good on the promise of healthcare IT lifting productivity rather than simplifying it and burdening clinicians with non-clinical administrative work. However, because of its “black box” nature at the inference level (post-processing), it can leave clients unsure of how much trust they can place in an NLP solution. At the same time, this has allowed vendors with lesser products, or at the very least, less purpose-built solutions, to sell without educating their partners.

Much like NLP itself, the goal today has been to shed some light into that darkness and provide interested parties the additional insights they need to move forward with confidence and a greater understanding of how to succeed.

About Health Fidelity

Health Fidelity simplifies risk adjustment, offering risk-bearing organizations clear visibility into and control over the process. Through our NLP-powered solutions and expert advisory services, we uncover insights that enable better care plans and more complete revenue capture. The Lumanent™ suite gives our partners the confidence to pursue and ability to succeed in risk-sharing arrangements across MA, ACA, Medicaid, and ACO programs.

HEALTH FIDELITY®

healthfidelity.com