MAKE EVERY VOICE HEARD WITH NATURAL LANGUAGE PROCESSING

Everything you need to know about communication between humans and machine intelligence

§sas

THE ART OF CONVERSATION

How can machines learn to understand and interpret the nuances of human language? With natural language processing.

Unstructured text is the largest human-generated data source, and it grows by the minute – exponentially. We have technology constantly at our fingertips, allowing us to rapidly converse through text, instant messages, email and social media. Conversations range from objectively stated facts to thoughts and perspectives, as well as comments on good or bad experiences with a business or organization. The internet makes accessing information easier than ever before, and digital records are now the norm rather than the exception.
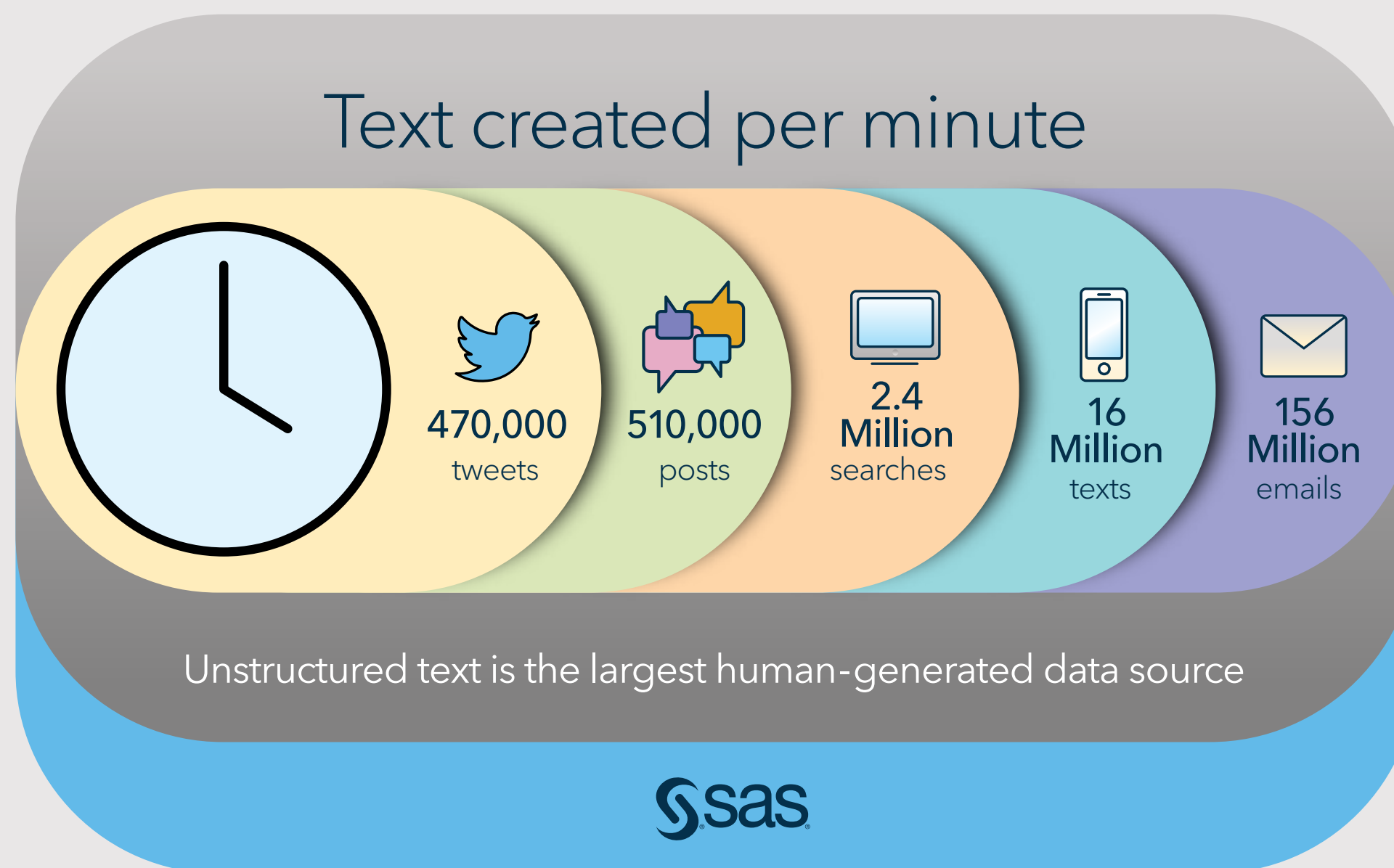
Understanding unstructured text

Each of those data sources hold an enormous volume of unstructured text. Human language, unlike structured information, does not fit neatly in rows and columns. Think about trying to fit a text message conversation into rows and columns. Every word or value doesn't fit nicely into a box, nor would those boxes perfectly align to some distinct value.

Anything you could write with a pen or type out on a keyboard can be considered free-form or unstructured text. Imagine giving that unstructured text to a computer to analyze. Without natural language processing, that information would hold no meaning to the machine.

With natural language processing, a machine can quickly and tirelessly sift through constantly growing volumes of text data to identify main ideas or topics, uncover emerging trends, analyze sentiment and identify correlations between words. A team of human analysts combing through piles of web data, emails, interview notes, digital records and social media sites would struggle to manually identify keywords and detect patterns due to the sheer volumes of information to sift through.

Natural language processing can be used to teach and guide machines to analyze unstructured text at a scale and speed that humans simply cannot match. In turn, this allows humans to focus on strategic, higher-level tasks that machines can't do, like applying reason, understanding impact and implementing an action plan.

Text created per minute

470,000
tweets

510,000
posts

2.4 Million
searches

16 Million
texts

156 Million
emails

Unstructured text is the largest human-generated data source

§sas

Chatting with machines

In addition to analyzing unstructured text data, people now speak to machines as a means to gain access to information or complete a task. Asking a machine, like Alexa or Siri, what the weather will be, to play a specific song, to order an item or schedule an appointment may seem easy and intuitive on the surface, but behind the scenes there's sophisticated technology at work.

Chatbots are designed to provide a natural conversational flow through a speech interface that mimics human conversation. This flow is meant to simplify access to information or automate the completion of a task. The chatbot's success at emulating human conversation is largely dependent on the accuracy of its natural language processing algorithms and the depth of human domain knowledge supporting the bot's framework.

With advances in machine learning and increases in compute power, chatbots are becoming smarter, and functionality is evolving beyond a simple call-and-response type interaction. SAS allows users to access advanced analytics capabilities through a chatbot interface. The result is a chatbot that can employ a variety of AI capabilities, such as forecasting methods to predict supply and demand, computer vision to recognize objects and events, or natural language processing to match intent accurately based on an understanding of contextual elements.

The art of communication

Whether communication takes place in spoken or written form, there's an art to human communication. There are factors and context in conversation that aren't always apparent in what we explicitly say or write.

The uniqueness of human language will always be a fluid art built from history, culture, environment and life experiences, and regardless of advances in technology, that unique quality is difficult for a machine to replicate without significant human guidance.

Artificial intelligence (AI) systems that effectively augment human efforts to analyze unstructured text, power smarter chatbots or convert speech to text are highly dependent on being properly trained to understand and interpret human language – through the use of natural language processing.

SPEAKING THE SAME LANGUAGE

How AI, natural language processing and human expertise work together to help humans and machines communicate and find meaning in data

The explosion of digital data, alongside massive growth in computational power, has caused much excitement in the market around AI. There are varying opinions about what AI can do for analytics, but as these machines perform analysis and report results, the core issue comes down to how humans and computers communicate.

As humans, we don't communicate our thoughts and findings in ones and zeros as machines do. We use the complexities of a variety of languages and dialects that have evolved over hundreds of thousands of years. Therefore, we must teach machines how we talk and what our words mean in order for them to communicate back to us in the same way.

AI and natural language processing

Artificial intelligence is a field of science that trains computers to emulate human tasks through learning and automation. With AI, machines can learn from experience, adjust to new inputs and accomplish specific tasks without manual interventions.

Natural language processing (NLP) is a branch of artificial intelligence that focuses on the understanding, interpretation and emulation of human language. NLP draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and machine understanding.

Building on the foundation of NLP, advances in natural language understanding and natural language generation have contributed to the development of cutting-edge applications. After unstructured text has been processed using NLP, natural language understanding (NLU) helps a machine comprehend what it's reading.

NLU goes beyond the structural understanding of language and trains machines to interpret intent, consider context, resolve ambiguity and give commands to the system to answer questions, show specific data or do a specific task. When a command is passed to a natural language generation (NLG) component, it can even generate well-formed human language on its own within a specified domain or task. Together NLU and NLG can make it seem as though the computer truly understands human language, an illusion that can break down if the system is asked to do a task outside the set of tasks it has been trained to do.

While natural language understanding seeks to address ambiguity and determine context, NLG aims to automate the production of content in conversational form at an incredibly large scale.

NLG uses a foundation of natural language processing to help machines examine any type of information, including documents; determine the overall context of information; and automatically produce reports, descriptions, explanations or interactive conversations. It requires linguistic input, often through the form of templates, as well as linguistic rules, in order to effectively generate its linguistic output.

Natural language generation is commonly used to automate the generation of large volumes of individualized content, such as form emails, descriptions of consumer goods, summaries of operational performance, easy-to-consume analytic results and the production of intelligent chatbot responses.

Human expertise and linguistics

Colloquial phrases, slang and sarcasm across different languages and dialects make it difficult for humans to perceive and learn the many nuances of language – and those intricacies in language must also be taught to a machine.

While one human can't feasibly master the intricacies of every language, a group of humans working together can. Likewise, natural language processing helps computers understand, interpret and navigate human language. Not everything can or should be automated, and a great deal of that help requires human expertise because natural language is so uniquely human.

That human expertise comes in the form of linguistics, which is the scientific study of language, and includes analyzing language structure, meaning and context. Linguists familiar with distinctions in syntax, semantics and pragmatics are critical in developing NLP capabilities that can operate at the scale and speed with which people expect AI systems to perform.

SAS has a team of linguists around the world who provide that human expertise in developing NLP capabilities.

Current languages with out-of-the-box NLP functionality include: Arabic, Chinese, Croatian, Czech, Danish, Dutch, English, Farsi, Finnish, French, German, Greek, Hebrew, Hungarian, Hindi, Indonesian, Italian, Japanese, Kazakh, Korean, Norwegian, Polish, Portuguese, Romanian, Russian, Slovak, Slovene, Spanish, Swedish, Tagalog, Thai, Turkish and Vietnamese.

Global Language Support for NLP

Arabic	Farsi	Hindi	Polish	Swedish
Chinese	Finnish	Indonesian	Portuguese	Tagalog
Croatian	French	Italian	Romanian	Thai
Czech	German	Japanese	Russian	Turkish
Danish	Greek	Kazakh	Slovak	Vietnamese
Dutch	Hebrew	Korean	Slovene	
English	Hungarian	Norwegian	Spanish	

The Basic Stages of Text Processing

NLP Natural Language Processing

NLG Natural Language Generation

NLU Natural Language Understanding
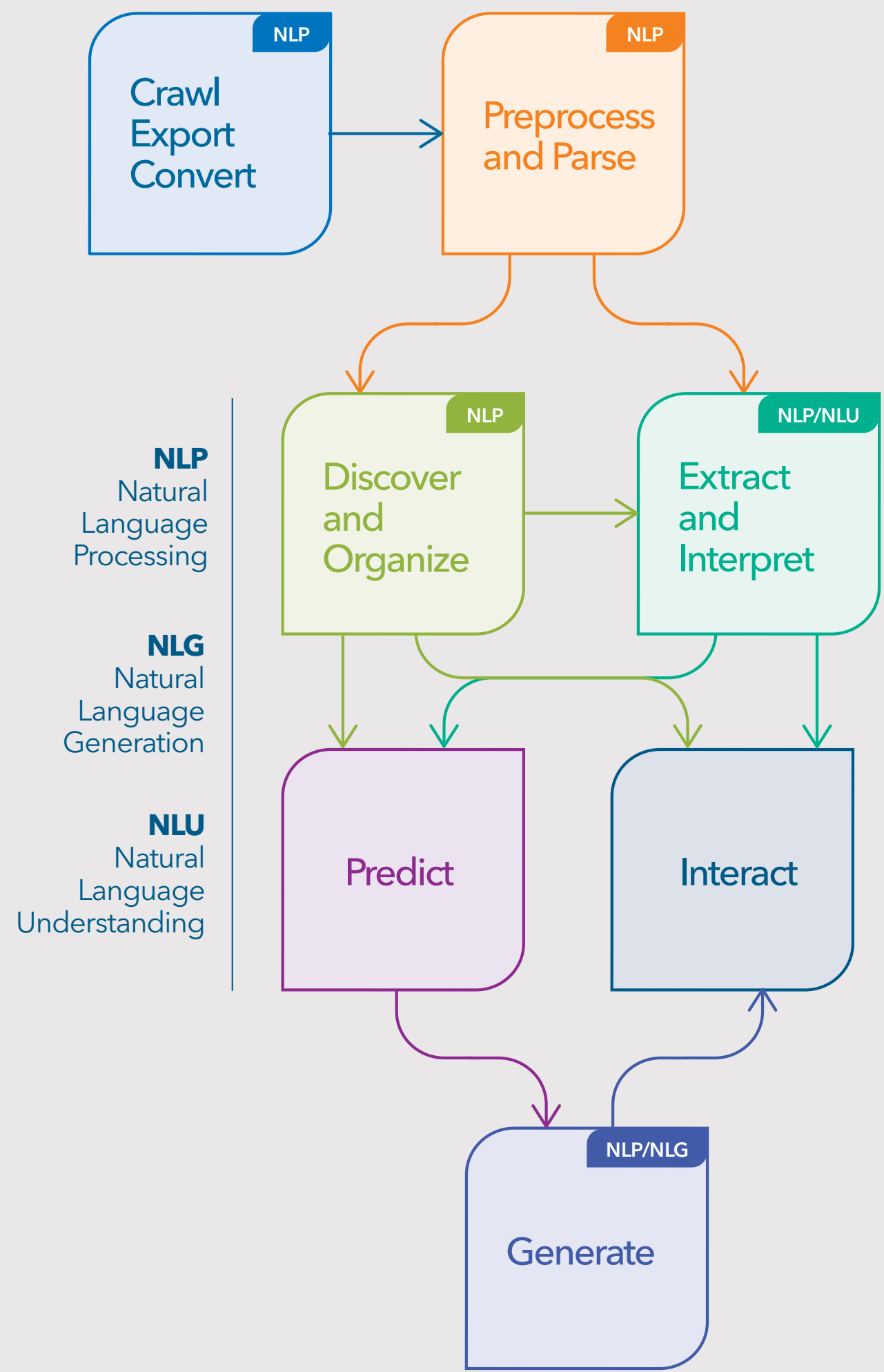
AI and human expertise

It's best practice to analyze unstructured text data with a combination of natural language processing, machine learning and human input. Human expertise provides the guidance for accurate analysis, and machine learning helps that analysis scale with ease. The basic stages of text processing include parsing, discovery and organization, and extraction and interpretation.

When using natural language processing to analyze data, the machine must first split the unstructured text data into useful, comprehensible units. This is the first step of parsing and is called tokenization. If input is spoken, the machine must convert that speech into text before further processing.

From there, natural language processing performs further parsing, applying linguistic analysis to extract features from the text, such as root words along with variations of the word, sentence boundaries, parts of speech, noun groups, syntactic structures and more.

All of this processing to identify the meaningful pieces and relationships in human language can be performed using machine learning-based models or models based upon linguistic rules. However, hybrid models that use both approaches generally perform the best.

In addition to machine learning, the application of deep learning through recurrent neural network models, specifically long short-term memory and gated recurrent units, can be applied for time series forecasting, text generation, speech-to-text and other NLP tasks. Human subject matter expertise can be added in the form of linguistic rules to improve model accuracy. Machine learning can help reduce the human model building effort by using semisupervised learning to automate the tagging data based on human input to training data.

The next stages in natural language processing each focus on a different level of analysis. The processes that discover and organize the data focus on the information available across the documents or at the document level, and can produce new information like networks, clusters, topics, and document-level sentiment and categories.

The processes that extract and interpret the data focus on information at the sub-document level, and make information contained inside a document accessible by transforming it to a more machine-readable state.
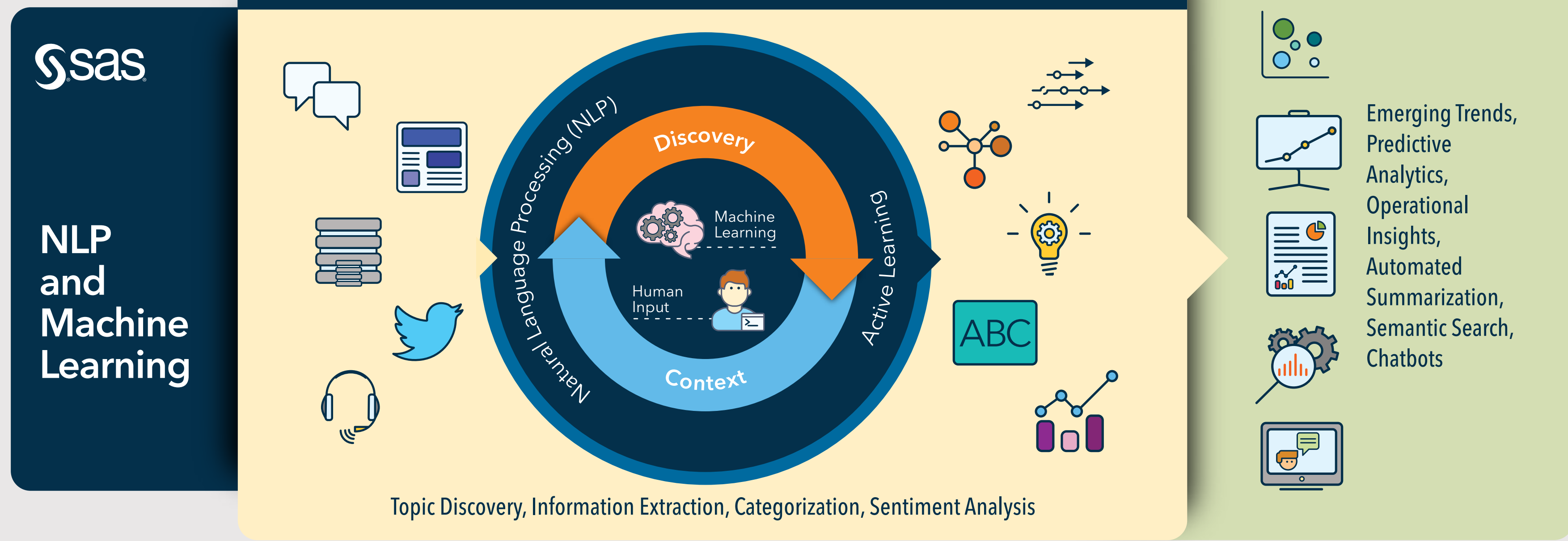
NLP and Machine Learning

Unstructured Text Data

Natural Language Processing (NLP) · Discovery · Active Learning · Machine Learning · Human Input · Context

Topic Discovery, Information Extraction, Categorization, Sentiment Analysis

Emerging Trends, Predictive Analytics, Operational Insights, Automated Summarization, Semantic Search, Chatbots

These processes create entity information, detailed relationships between pieces of information and even an interpretation of elements of meaning.

SAS offers the ability to use machine learning together with NLP capabilities, such as tokenization and grammar, for all supported languages. While machine learning is integral to building NLP models, deploying those machine learning models as rule-based systems offers greater transparency, easier monitoring and faster processing speed.

For example, analysts may choose to use a linguistic rules approach to building transparent, high-performance text classification models to quickly categorize email messages and automatically route those messages to relevant contacts in real time. It's time-consuming to manually create linguistic rules, but machine learning can extract the most relevant keywords and sequences from large volumes of categorized data and apply linguistic rule models. This is a powerful active learning approach where analysts use both approaches (linguistic rules and machine learning) in tandem. Domain-specific, user-defined rules complemented with machine learning approaches quickly improve overall precision and recall. The models are implemented as transparent linguistic rule models, which is often preferred in organizations that require regular traceability and audits.

The output of natural language processing, machine learning and human input can vary from topic discovery to entity extraction to categorization or sentiment analysis, which can then be used to detect emerging trends, perform predictive analytics, provide operational insights or create conversational systems to interact with information, such as chatbots. Predictive models created by machine learning use the output of natural language processing to automate the generation of insights.

Regardless of the objective, natural language processing is required for humans and machines to speak the same language and ultimately find meaning in the data being analyzed.

THE SCIENCE OF COMMUNICATION

How natural language processing turns words into numbers and back again

Natural language processing allows us to use the power and speed of computer processing to get a glimpse into the high volume of human words, feedback and perspectives in unstructured text to better understand the world around us.

It's impossible for a business serving 2 million customers to effectively speak with and listen to every single customer in a conventional human conversational setting. Since we so often communicate through digital means, we're able to employ artificial intelligence to rapidly identify the most important content faster than a human ever could. This helps get the right information to the right people when they need it. But while machines are faster at processing data and analyzing trends, humans will almost always be more accurate at interpreting the significance.

From words to numbers and back again

For training and inferencing, machine learning algorithms expect their input to be a vector of numbers rather than plain text. This means that in order for a computer to process unstructured text, that text must first be converted into a numerical representation. The numerical representation is processed, and the output converted back to a textual representation that humans can understand. This is the science behind teaching machines how to understand natural language.

The path from words to numbers has many twists and turns. It includes breaking strings of punctuation, character and numeric values into meaningful units of analysis, determining important features and unimportant conventions, connecting the correlated parts together into insights across massive volumes of unstructured text, and presenting those insights in an understandable way to humans so they can make data-driven decisions.

Algorithms in action

The science behind teaching machines to read and interpret human language is sophisticated and complex. From a business perspective, NLP functionality can be grouped into five general areas: preprocessing; parsing the data, analyzing trends (discovery and organization), extracting information, and developing conversational systems (interpretation and generation).

Preprocessing

When document collections contain very large documents, or the collections contain a variety of document types, eliminating some noise from the data, or normalizing nonsentential data to derive more clarity about the contents, can result in downstream models that are easier to build, maintain and apply more accurately.

- **Splitting:** Identification of relevant sections of the document, or splitting each document into independent sections, can focus analyses on important parts or make related concepts clear.

- **Deduplication:** Removing or grouping sets of related documents can eliminate redundancy and reduce impact on statistically based analyses.

- **Normalization:** Clarifying or cleaning nonstandard or nonsentential data can enable better results in parsing and other downstream analytics.

- **Stratification:** Another tool that can help focus the work on specific areas of the corpus or provide subsections of the data to apply models to.

Parsing the data

Before analyzing large volumes of data, it's important to break the data into chunks and provide the human framework the machine needs to analyze at scale. Parsing separates text into its words, phrases, punctuation marks and other elements of meaning. SAS uses a variety of approaches for parsing tasks that may also vary across languages, including regular expressions, supervised learning models, and human-built rules and dictionaries. The majority of these actions are rule-based, applying linguistic expertise for the specific language. SAS provides parsing actions as out-of-the-box functionality across all supported languages.

- **Tokenization:** Chops a character sequence, such as a sentence or document, into individual sentences, words or morphemes, which can then be used as input for part-of-speech tagging.

INPUT
"No matter what people tell you, words and ideas can change the world." **-Robin Williams**

OUTPUT
"No *matter* *what* *people* *tell* *you* *,* *words* *and* *ideas* *can* *change* *the* *world* *.* *"* *-* *Robin* *Williams*

- **Lemmatization:** Returns words to their base forms.

has, had, haven't ⟶ have

- **Misspelling analysis:** Associates misspelled words with a set of variants that can include the properly spelled word. Based upon edit distance and context and can be enhanced by use of a dictionary.

- **Part-of-speech tagging:** Grammatical classification of words based on their definition and context. Advances in deep learning are helping automate part-of-speech tagging with recurrent neural networks.

Pronoun
Verb
Article
Noun
Adverb
Preposition

She knew the plan needed to change.

He got change at the store.

- **Sentence boundary disambiguation:** Determines where sentences begin and end. For handling ambiguous situations like abbreviations and languages or texts where sentence boundaries are unclear, unsupervised machine learning algorithms can augment a regular expression-based approach.

> o [The loan process was simple, and I was approved in minutes.]
> [I didn't like that it seemed my information was shared, and I had 30 spam emails and nonstop sales calls for two days.]

- **Dependency parsing:** Assigns syntactic relationships between the words of a sentence through application of deep learning algorithms.

Input: I want the most scenic campsite.

Output:

I	want	the	most	scenic	campsite	.
PRON	VERB	DET	ADV	ADJ	NOUN	PUNCT
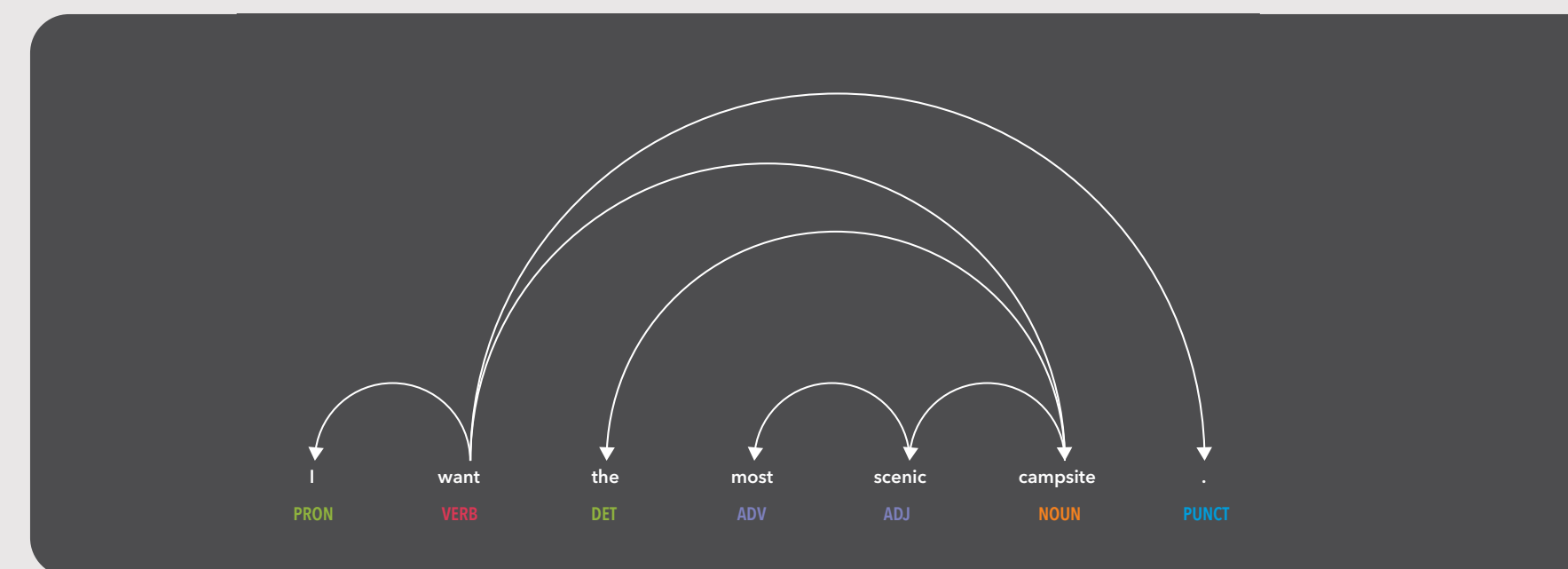
- **Clause analysis:** Combines supervised machine learning and linguistic rules to identify clause boundaries.

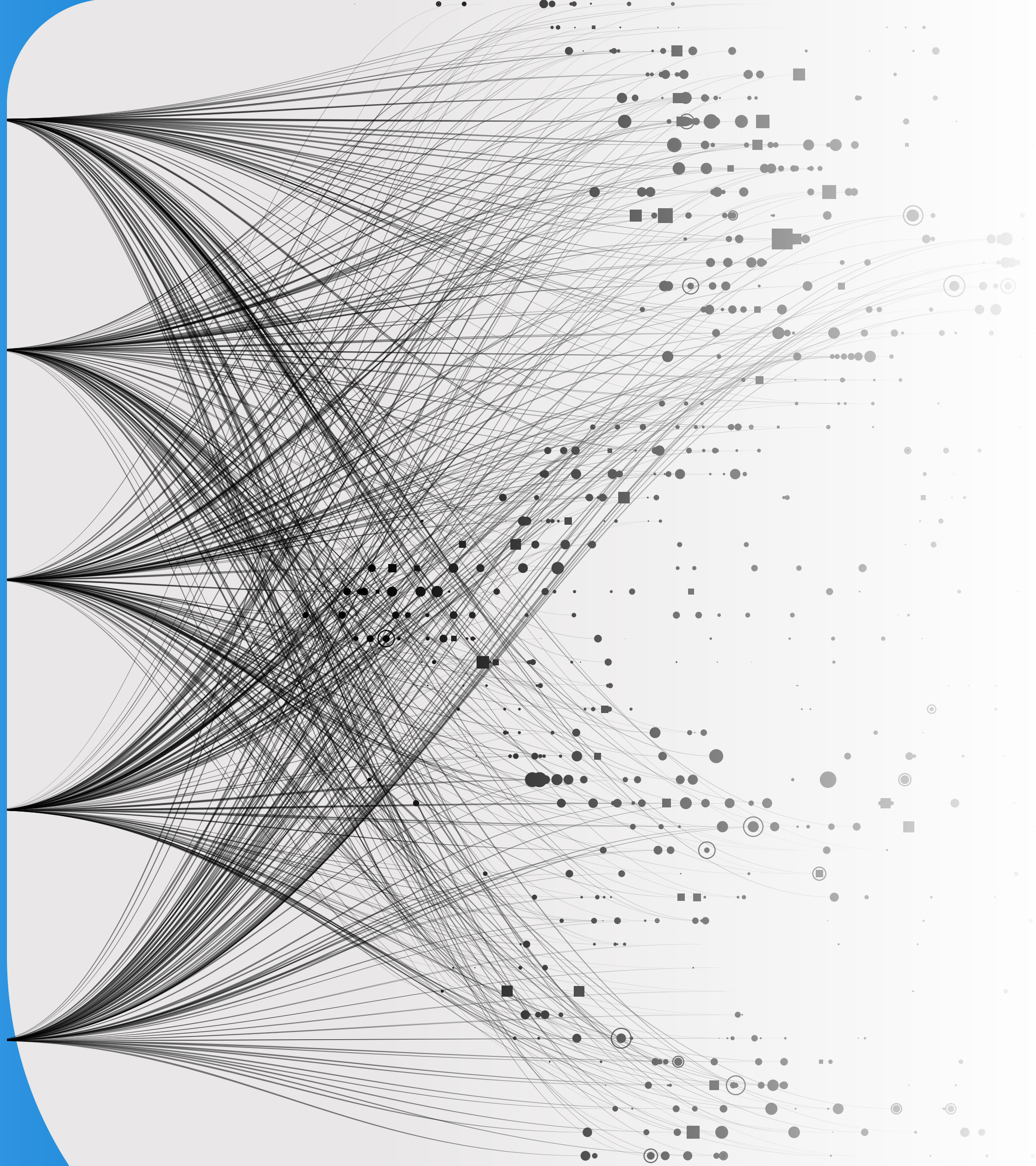> o [The food was amazing], [but the customer service was awful].

Analyzing trends

In the process of discovery, natural language processing and unsupervised machine learning can help reveal trends in data by automatically extracting terms and topics that appear in correlation to each other throughout a set of documents. Human-designed systems of categorization can also organize data to uncover trends and patterns across documents.

- **Singular value decomposition:** Provides a matrix of word occurrences or word frequencies in documents.

- **Latent dirichlet allocation:** Allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

- **Categorization:** Through rule-based categorization rules or supervised machine learning, documents, sentences, paragraphs, etc., can be grouped together as related.

- **Sentiment (document level):** A type of categorization of documents into positive, negative or neutral buckets.

- **Word embedding:** Creates numeric coordinates that place terms in space, clustering words together that are used in similar ways.

Speech to Text

	Frame **1**	Frame **2**	...	Frame **N**
Feature 1	0.4	0.2	...	0.5
Feature 2	0.1	0.3	...	0.3
...	...	...	...	...
Feature M	0.2	0.1	...	0.2

FEATURE EXTRACTION

	Frame **1**	Frame **2**	...	Frame **N**
A	0.2	0.4	...	0.1
B	0.3	0.1	...	0
...	...	...	...	...
Z	0.2	0.4	...	0.6

ACOUSTIC MODEL

HELLO WORLD	0.9
HELLO WORD	0.4
HOLLOW WORLD	0.2

LANGUAGE MODEL

"HELLO WORLD"

Extracting and interpreting information

Information extraction automatically pulls out structured information from an unstructured or semistructured text data type to create new structured text data. This includes tasks such as entity recognition, relationship extraction and coreference resolution. Summarization interprets the meaning of a larger text to identify important components and formulate a shorter version of the same message or condenses a set of related documents into an aggregated report.

- **Noun groups:** Identifies head nouns and associated content words that help distinguish the specific meaning in context.

  o The attentive gate guard shut the main gate immediately upon hearing of the incident.

- **Entity recognition:** Extracts and classifies elements of text into predefined groups such as names of persons, organizations, locations, expressions of time, quantities, percentages and more.

- **Rule generation:** Automatically creates entity recognition rules from selected terms using semisupervised machine learning.

- **Sentiment analysis (subdocument level):** Identifies subjective information in text, labels it as positive, negative or neutral and associates that information with some entity.

- **Text summarization:** Automatically provides a synopsis of important concepts within a single long document or a set of related documents.

Developing conversational systems

Natural language processing capabilities designed to simplify the interaction between humans and computers are commonly referred to as conversational AI. Conversational AI primarily focuses on the applications of natural language understanding and natural language generation.

• **Chatbots:** Experiences that use conversational AI are often referred to as chatbots or virtual assistants. Through these chatbots, computers can understand and respond intelligently to human inputs through spoken or written natural language. They can be programmed to respond to simple keywords or prompts, or to hold complex conversations about specific topics. Chatbots range in complexity from information retrieval using intent matching to machine learning capabilities that provide in-depth responses and tailored suggestions. In this case, the more a user interacts with a chatbot, the smarter it gets. Example use cases include 24/7 automated customer support or increased productivity by offering a natural language interface to analytics.

• **Chat-enabled analytics:** With offerings like SAS Conversation Designer, chatbots can marry analytics with conversational capabilities. For example, a chatbot can automatically query and describe large corporate or public data sets or provide relevant visualizations. Users can request summarized or analyzed results verbally by saying, for instance, "Which marketing campaigns are generating the most leads this quarter?" The chatbot can provide the answer and then offer additional information or suggest a related report to view based on patterns in the data and in previous related queries. You can ask the chatbot to share the results with others or even perform specialized analytics within the chat interface.

• **Speech-to-text:** Speech-to-text capabilities enable the recognition of spoken language followed by the translation of that speech into text format. Once translated into text, natural language processing can be used to extract information and analyze patterns within that text. The process begins with feature extraction, where the machine breaks the audio file into short frames for analysis. Next, an acoustic model predicts the alignment between a phonetic sound and the letter it represents. The values produced by the acoustic model are then fed into a language model which uses NLP to translate those values into words and phrases.

NLP Modeling Approaches

Category	Approach	Method
Preprocessing	Splitting	Rule based and/or supervised machine learning
	Deduplication	Rule based and/or unsupervised machine learning
	Normalization	Rule based, supervised machine learning and/or deep learning
	Stratification	Rule based and/or unsupervised machine learning
Parsing	Tokenization	Rule based, supervised machine learning and/or unsupervised machine learning
	Lemmatization	Rule based, supervised machine learning and/or deep learning
	Misspelling analysis	Rule based, supervised machine learning, unsupervised machine learning and/or deep learning
	Part-of-speech tagging	Rule based, supervised machine learning and/or deep learning
	Sentence boundary disambiguation	Rule based and/or supervised machine learning
	Dependency parsing	Rule based, supervised machine learning and/or deep learning
	Clause analysis	Rule based, supervised machine learning and/or deep learning
Analyzing Trends	Singular value decomposition	Unsupervised machine learning
	Latent Dirichlet allocation	Unsupervised machine learning
	Word embedding	Supervised machine learning, unsupervised machine learning and/or deep learning
	Categorization	Rule based, supervised machine learning and/or deep learning
	Sentiment analysis (document level)	Rule based, supervised machine learning and/or deep learning
Extracting & Interpreting Information	Chatbots	Rule based, supervised machine learning and/or deep learning
	Speech-to-text	Rule based, supervised machine learning and/or deep learning
Developing Conversational Systems	Noun groups (chunking)	Rule based, supervised machine learning and/or deep learning
	Entity recognition	Rule based, supervised machine learning and/or deep learning
	Rule generation	Rule based, supervised machine learning and/or deep learning
	Sentiment analysis (subdocument level)	Rule based, supervised machine learning and/or deep learning
	Text summarization	Rule based, supervised machine learning and/or deep learning

§sas

NATURAL LANGUAGE PROCESSING IN ACTION

How multiple industries are delivering better products and services – and improving efficiency – with natural language processing

People want to be heard. Whether it's a customer reporting a service problem, a patient seeking medical help or a citizen contacting their government representatives – people are reaching out in digital form more because technology makes communication fast and convenient.

Regardless of how a touch point is happening – face-to-face, over the phone, through email, on social media or in a webchat – critical information is pouring into your organization, and you need to understand what's happening in time to make the right decisions.

Natural language processing in action: Delivering better products and services

Natural language processing is highly effective when used to analyze communication streams between a company and its customers. That analysis provides real-time insights that can be used to improve the overall quality of products and services being delivered.

For example, a dairy company wanted to use call center data to understand product quality and attrition rates, but it was overwhelmed by the amount of data pouring in. It used natural language processing to help categorize and extract relationships from that data, and quickly discovered there was a trend in complaints discussing milk that had "cream on top" and "looked funny."

Being able to quickly identify and dive into the problem led to the realization that a human error had been made in a machine calibration, and that error had caused a batch of milk to not be homogenized. The error was caught using natural language processing, and the company was able to correct it before more nonhomogenized milk was produced.

NLP also revealed that a specific promotional strategy was resulting in greater customer attrition. The company was offering free delivery for six months, but sentiment analysis revealed that customers were very upset at being charged a delivery fee after the initial free six months. Offering a discount on delivery services rather than offering free delivery for a limited time was shown to improve customer retention and overall satisfaction with the company's products and services.

In another example, a telecommunications company serving more than 1 million customers needed a more proactive approach for addressing more than 10,000 daily service calls in four different

languages. The IT and business users came together to identify user requirements, define relevant keywords and develop a daily reporting dashboard to reflect leadership needs.

The application of natural language processing made it possible to, almost simultaneously with human documentation in the call center, show relationships and provide solutions. This provides the call center employees with the right answer to the customer's need on the first call.

In addition to call center feedback, daily reporting dashboards were created to evaluate customer trends and feedback using social media analytics. This gave corporate communication teams increased customer insight that resulted in more accurate targeted marketing initiatives.

Improving organizational efficiency

Natural language processing can also be used to refine internal processes and create more efficiency in operations. That efficiency can take the form of changes to training standards, more centralized data access or even the use of an intelligent chatbot to deliver analytics.

For example, a bank with the mission to earn customers' trust by serving their needs better than any other bank launched an online portal for customers to communicate with human agents through webchats.

The bank holds an average of 250,000 webchat conversations through the portal every month, and each webchat was followed up with a customer satisfaction survey. With manual analysis, the bank was only able to go through 200 webchats each month. Natural language processing helped scale that effort to all 250,000 monthly webchats – and provided insights to help identify which topics agents were handling well and areas that required improvement. This data allowed team leaders to see the performance of each individual agent and assess their strengths and weaknesses so training efforts could be targeted accordingly.

Additionally, the NLP application revealed that webchats showing positive sentiment did not always align with a positive satisfaction score in the follow-up survey. While the discrepancy could be disheartening to agents, NLP revealed that in many cases the cause for a poor satisfaction score was due to an issue beyond the agent's control.

For example, if a customer wanted to cancel a direct debit immediately, the agent could update the system during the webchat, but the internal system took an additional day to process the cancellation, which left customers feeling annoyed and dissatisfied. That type of feedback allows leaders to understand the root cause of poor customer satisfaction and adapt processes to fix problems and boost customer satisfaction.

In another example, a facility manager at a large solar farm needed a tool to monitor and control solar panels when away from the computer. A chatbot delivered through a mobile interface was designed to analyze live streaming data from the solar panels. The facility manager can

now use a smartphone to ask the chatbot to report on equipment status, provide each panel's energy generation and produce a summary of the energy output by day, month or season.

The mobile chat interface extends the capabilities to technicians so they can query the application when they're outside restarting a panel, or monitoring the condition of panels in extreme weather conditions. This puts analytics at the fingertips of employees so they can access the information they need when they need it.

Listening to understand

It's impossible for a human to listen to 10,000 customers on a daily basis and meet every single person's need. And bouncing callers around to different groups trying to find the right answer creates a frustrating experience for customers.

To equip your employees with the right tools to solve problems in a timely manner, they need the support of natural language processing. Machines can process details in the massive call volume, provide real-time insights and recommend next best actions at a cadence that matches the natural conversational flow between two humans.

That means technology can actually improve the human experience by finding the right answers the first time, without having to sacrifice the personal connection and relationship with your customers.

9 BEST PRACTICES FOR NATURAL LANGUAGE PROCESSING

Tips for building and deploying NLP models at scale

Natural language processing is the bridge to facilitating understanding and strengthening collaboration between humans and machines in the design and implementation of AI systems. And natural language processing requires significant intelligence and capability from both humans and machines to be implemented effectively.

The ability to feasibly manage a continuously growing volume of unstructured text and effectively build and manage a variety of NLP models greatly depends on having a flexible architecture with immense processing power that can scale to meet operational needs while simultaneously ensuring governance and security of all data and models.

In addition to natural language processing, SAS offers four other core AI capabilities: machine learning and deep learning, computer vision, forecasting, and optimization. But implementing a core capability by itself is not enough to gain value from algorithms.

Moving beyond a singular capability to not only create, but also operationalize, an entire AI system requires embracing the analytics life cycle. The most foundational element of AI is data. The output of your analytics will only be as good as the input, so comprehensive data management is a required component of success.

Visualization is important for exploring data and analytics, creating intuitive dashboards and easily communicating results to a variety of stakeholders. Discovery, where analytics is used to transform data into intelligence, is where AI technologies sit. Core AI capabilities can be used independently or in tandem with other AI technologies to discover insights based on data type and business need.

Deploying AI models into production is where value is realized. With SAS, you have robust model management capabilities whether you are deploying one model or thousands. A unified platform saves immense time and effort by allowing models to be built once and deployed anywhere without having to recode.

Natural language processing at SAS

Natural language processing capabilities from SAS are supported by the entire analytics life cycle as well. SAS ingests, cleanses and transforms data for analysis, easily accepting multiple file formats through local or remote file systems, relationship databases and cloud storage.

The SAS Platform offers multithreaded parallel processing for in-memory analytics on a cloud-ready, open architecture. REST APIs allow for flexible integration, and users have the choice to code in SAS, Python, R, Java, Scala or Lua. Data and model lineage and governance allow users to maintain access and control of data management and analytics across the analytics life cycle.
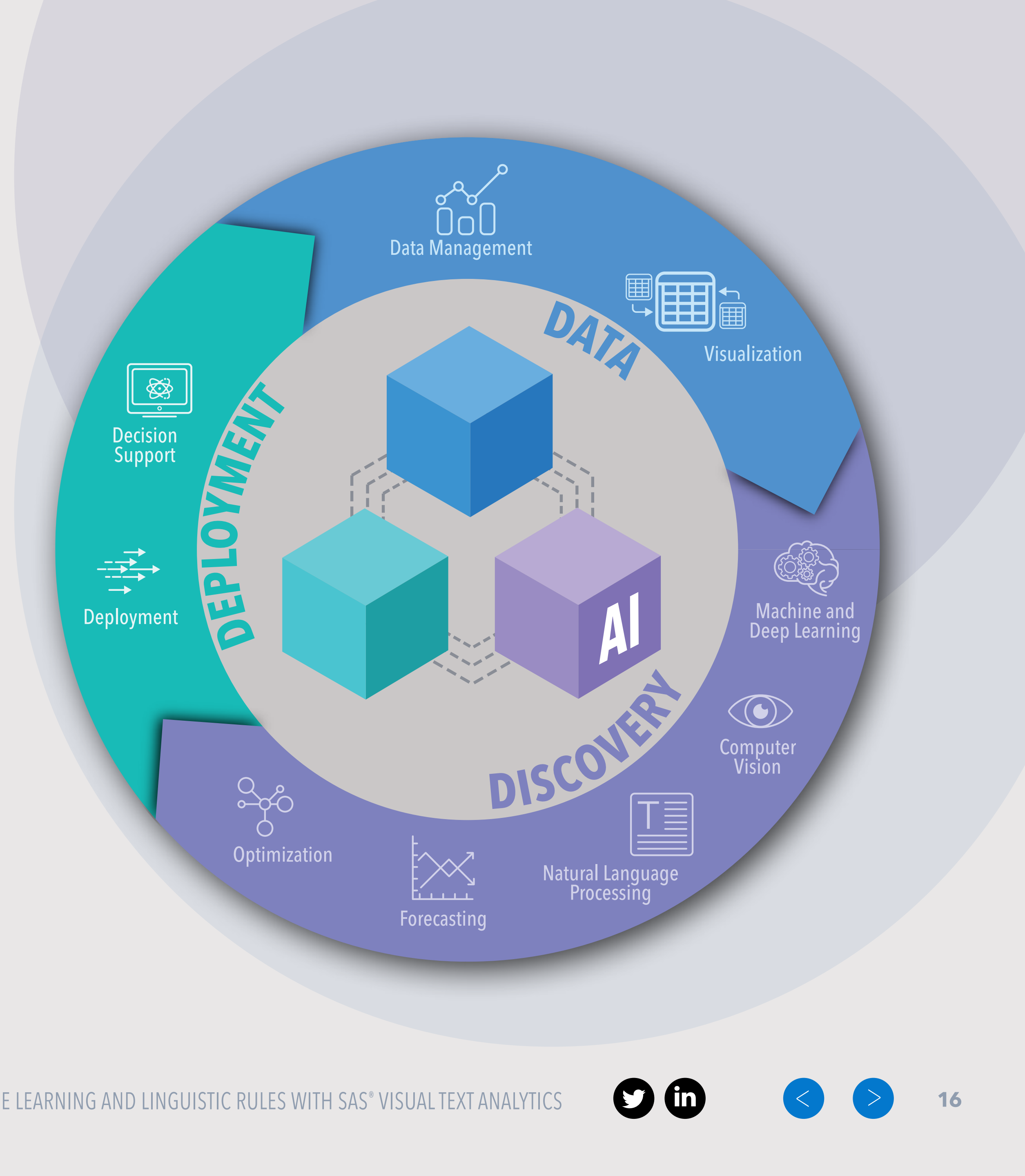
NLP capabilities on the SAS Platform are supported by both central processing units (CPUs) and graphical processing units (GPUs). SAS also provides an analytic store (ASTORE), which is a binary file that represents the scoring logic from a specific model or algorithm. That compact asset allows for easy score code movement and integration into existing application frameworks. The ASTORE is created using results from the training phase of model development, and it can be easily transported from one host to another. NLP models can be deployed in database, in Hadoop and in stream with full life cycle management.

Best practices for building and deploying NLP models

The SAS Platform provides the architecture needed to build and deploy natural language processing models at scale. As a leader in text-based AI platforms, SAS has created a quick checklist of things to consider when building and deploying NLP models.

NLP on the SAS® Platform

Analytics Life Cycle

DATA
- Local file system
- Remote file system
- Cloud storage
- Relational databases
- Multiple file formats
- Audio files

DISCOVERY
- Tokenization
- Lemmatization
- Part-of-speech tagging
- Summarization
- Topic discovery
- Categorization
- Sentiment analysis
- Information extraction
- Speech-to-text
- Chatbots

DEPLOYMENT
- In database
- In Hadoop
- Streaming
- Model life cycle management
- Decision management
- Transparency and interpretability

Lineage and governance

Visualization

Programming language of choice

Open architecture and REST APIs

Cloud-ready

In-memory analytics

Multithreaded parallel processing

Best practices checklist

1 Determine and clearly articulate the objective of your NLP application.

2 Gather technical and business users to identify key terms and dashboard requirements.

3 Identify any regulatory guidance required for a given data type or industry.

4 Create a standard taxonomy.

5 Provide a gold data set for those items in the taxonomy that you want to measure for accuracy.

6 Determine if there is a need for deep learning and graphical processing units.

7 Identify data sources and cadence with which they need to be analyzed.

8 Create best practice pipeline for building models.

9 Codify security and governance processes for model management.

The communication link between human and machine intelligence

Natural language processing is the communication link between humans and machines. It's also a link that requires support from both humans and machines to work effectively. NLP holds the power to improve how we live and work. It can help bring progress to areas that have been slow or difficult to change without the partnership between human and technology.

Look at your organization and consider the unstructured text or audio data you gather and the possible revelations it may hold. That data reflects the voices of those you serve and holds the potential to help you deliver better experiences, improve quality of care and enrich human engagement. There are powerful stories to be told from your unstructured text data. And the best way for you to find them is with natural language processing.

Find the information you need in your
unstructured text with SAS® Visual Text Analytics

§sas.

To contact your local SAS office, please visit: **sas.com/offices**