

PROBLEM

Safety is one of the most important goals for any industrial organization. Organizations that can't anticipate and mitigate future incidents will encounter lost time injuries

SOLUTION

Natural language processing can identify new risks, enabling safety supervisors to apply corrective action before an accident occurs

RESULTS

Organizations are able to identify their top hazardous activities, locations, in-compliance and deviation from Health, Safety, and Environment (HSE) standards, enabling the organization to minimize the probability of an incident and lost time injury

THE PROBLEM

Effective HSE management programs are not only critical for the health and well-being of employees, but also to ensure productivity and overall performance. However, with every new project, site, and piece of equipment comes a wide variety of safety risks, leading to an increased demand for safety supervisors on industrial platforms.

The HSE function of the organization has to deal with ever increasing operational activities that requires increasing resources from the HSE group to train employees, audit work sites, investigate incidents, and put in place corrective and mitigation actions. Another challenge is getting actionable insights from the influx of data from historical incident reports. These records are required to be collected and analyzed for regulatory purposes, and are often collected using form-based data entry. Typically, safety supervisors manually analyze these records to identify risks, and then create new regulations, best practices guidelines, and safety training.

This process is extremely time-consuming and often leads to incomplete results where only known risks are identified. Arguably, the biggest risks are the risks that operators and safety supervisors don't yet know about. However, pull-down menus, check boxes, and other form fills only reveal previously identified risks; new risks are not seen until a significant event or an accident occurs.

Operators and safety supervisors must take a more proactive approach to HSE management. More specifically, they need a solution that can:

- *Extract actionable insights from unstructured data found in incident reports, observation cards, etc.*
- *Reduce incident rates of previously identified hazards*
- *Provide assurance that critical risks are managed effectively*
- *Anticipate and mitigate future incidents before they happen*

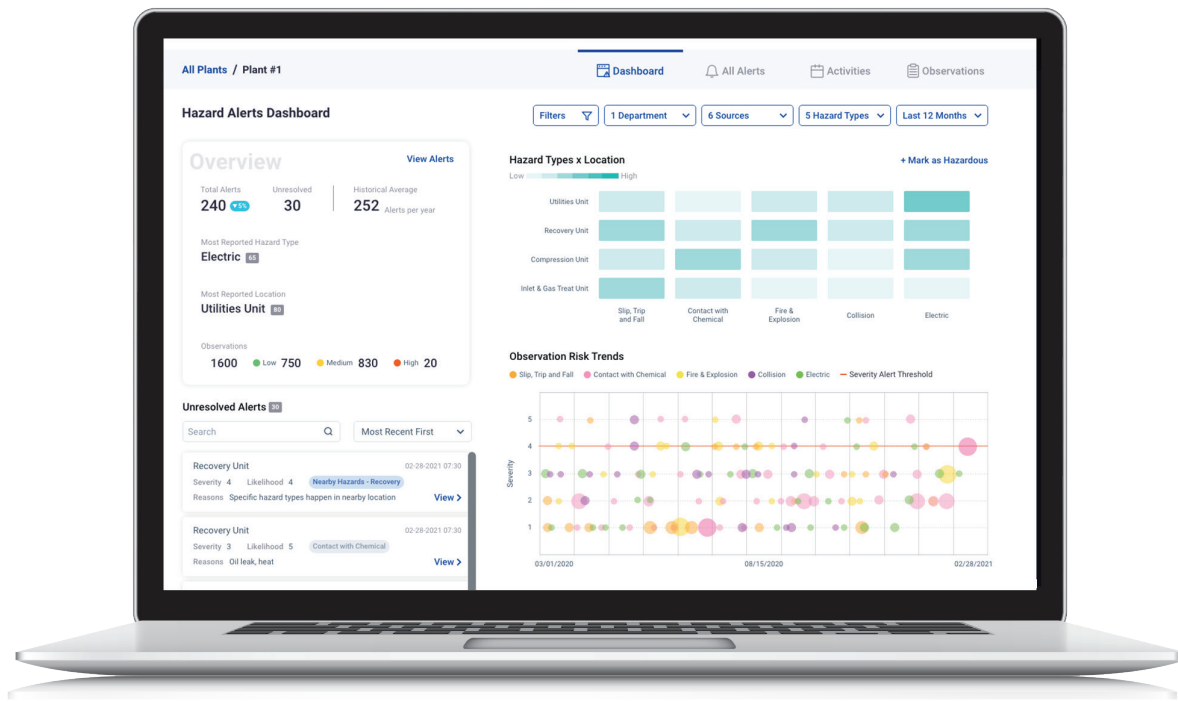
Natural language processing, a branch of artificial intelligence, can deliver significant value in improving HSE management.

THE SOLUTION

In industrial environments spanning oil and gas, manufacturing, renewable energy, and others, leveraging artificial intelligence enables operators and subject matter experts to operationalize organizational information and tribal knowledge, extract valuable insights from large volumes of data, and focus on making high-quality decisions at speed. Natural language processing, in particular, can not only help safety supervisors better understand what incidents happened in the past and why, but it can identify leading indicators of developing hazardous conditions, enabling safety supervisors to anticipate and mitigate incidents before they occur.

Modern industrial operations collect thousands of data points per day from unsafe conditions, permit violations, operational hazards, and a host of other disparate sources. These data contain leading indicators of hazards that may be difficult and time-consuming for safety supervisors to manually extract from, but not for natural language processing.

SparkCognition's DeepNLP™ product uses natural language processing to automatically identify new risks, enabling safety supervisors to apply corrective action before an accident occurs.



How the DeepNLP Product Delivers Value for HSE Management

To maintain a safe and risk-aware culture in the organization, each employee at an industrial plant is required to submit several hazardous identification or observation cards per month. These observation cards, or safety reports, are collected using form-based data entry in text format, and contain a wide variety of data points:

- Time and date of observation
- Specific observations: "A leak observed and hot oil spilled on the ground," or "Tools were left on the ground and are a step hazard," among many others
- Location of the observation
- Hazard type
- And more

Over time, safety supervisors will collect thousands of observation cards with the above data in unstructured format. While it would take safety supervisors a long time to analyze all this data and extract meaningful insights, the DeepNLP product is able to analyze this unstructured data, extract valuable insights, provide summaries of hazard and risk areas, and assign the likelihood of an event occurring in the future. In addition, the DeepNLP product provides a platform for creating, assigning, and closing action items related to these hazards, enabling safety supervisors to take an even more proactive role in resolving workplace safety issues.

THE RESULTS

In one case, SparkCognition worked with a utilities operator that received thousands of hazard observation cards per month from different sites. The customer asked SparkCognition to use the DeepNLP product to automate the analysis of these observation cards and extract actionable insights to help improve workplace health and safety.

The DeepNLP product ingested the data from the observation cards, extracted each hazard type, assigned the likelihood of a hazardous event occurring, and then provided a summary of potential risk areas. The customer discovered that driving accidents involving animals on the road was a significant safety hazard for their employees of which they were previously unaware. The data revealed that workers were driving their work vehicles into all kinds of animals including cats, moose, horses, and geese. The DeepNLP product classified the animal cluster and identified a significant overlap with the driving cluster. This led to an awareness campaign to alert the employees and the customer saw a reduction in the number of automobile accidents involving animals.

Contact us at info@sparkcognition.com to learn more about how natural language processing can help your operations anticipate and mitigate future safety hazards.

ABOUT SPARKCOGNITION

We catalyze sustainable growth for our clients throughout the world with proven artificial intelligence (AI) systems, award-winning machine learning technology, and a multinational team of AI thought leaders. Our clients partner with SparkCognition to understand their industry's most pressing challenges, analyze complex data, empower decision-making, and transform human and industrial productivity. Our vision is to build scalable AI solutions to solve the problems that matter most. We collaborate with organizations to help them reduce environmental impact creating a better, smarter, and more sustainable world. To learn more about how SparkCognition's AI applications can unlock the power in your data, visit www.sparkcognition.com.

Structured data is data that can easily be represented as a table—think spreadsheets. If your data contains only numbers and categories, existing analytics tools can easily handle it. However, this accounts for only about 20% of data produced by organizations.¹

The rest is the messier twin, unstructured data: PDFs, books, journals, audio, video, images, notes, analog data, and any other source imaginable. This unstructured data is primarily meant for human use and consumption but hard to analyze at scale, so it's rarely leveraged in aggregate to provide deeper insights. Businesses carefully (or not so carefully) collect this data and file it away, where it sits unused, rarely ever to see the light of day. Despite the current fervor surrounding the power of big data, this is the fate of 80% of all data produced. These numbers aren't likely to go down anytime soon. The total amount of data, both structured and unstructured, is increasing year over year by 39%.² IDC and EMC both project that data will grow to 40 zettabytes by 2020, which would be a growth of 50 times in 10 years.³ To put these numbers further in perspective, linguist Mark Liberman has calculated that all human speech ever spoken in the history of humankind would total to about 42 zettabytes.⁴ And 80% of this is unstructured. The amount of unstructured data being produced and stored is beyond human comprehension—and almost none of it is being used. All the while, organizations struggle with bottlenecks in organizational workflows, increasing operational costs, lack of visibility into processes, and the loss of tribal knowledge from the workforce.

Why is this massive resource allowed to lie fallow, even as businesses hook up increasing numbers of sensors to try to fuel data analytics? Because unstructured data cannot be understood and analyzed by most machines, accessing and analyzing unstructured data is a difficult, expensive prospect. Humans can understand this data, of course, but analysis driven entirely by humans doesn't scale to large operations, opens up the risk of human errors, and is a waste of time and resources. Human beings may be phenomenal at understanding language and images, but they're not equipped to handle data on the order of zettabytes.

While machines aren't naturally inclined to understand unstructured data, they can be taught. Natural language processing, or NLP, is a field dedicated to teaching machines to use and understand language in a human-like fashion. Major airlines are already using NLP, paired with machine learning, to extract tribal knowledge hidden in maintenance logs and make it available across the workforce. By leveraging historical data, machine learning is also elevating historically proven troubleshooting techniques to help diagnose and solve problems faster and more effectively.

NATURAL LANGUAGE PROCESSING

The field of NLP has existed in its modern form longer than even the study of artificial intelligence, with work on automatic translation and similar projects dating back to the early 1950s. But it's the recent machine learning boom that has revolutionized the subject, allowing it to flourish in new ways. This is because machine learning revolves around writing algorithms that can learn beyond

Unstructured Data *The Lifeblood of Organizations*



Process Logs

Event logs, server data, application logs, business process logs, audit logs, CDRs, mobile location ...



Sensor Data

Medical devices, electric meters, cameras, ECUs, engines, HVAC, machinery, high value assets ...



Business Data

Project management, marketing, productivity, CRM, contracts, procurement, HR, expenses ...



Public Records

Government, competitive, regulatory, compliance, health care services, public finance, stock ...



Archives/Storage

Scanned documents, statements, insurance forms, customer information, paper archives, system records ...



Documents

XLAS, PDF, CSV, email, DOCX, PPT, HTML, plain text, XML, JSON ...



Social Media

Twitter, LinkedIn, Facebook, Tumblr, Blog, SlideShare, Youtube, Google+, Instagram, Flickr, RSS, Pinterest ...



Media

Images, videos, audio, Flash, live streams, podcast, webinars ...

their initial programming, rather than being constrained by the rules coded into them. Rather than trying to hand-code all of the rules of language—a daunting task even if the scientific community agreed on them—programmers feed text into a machine learning program and let it glean the rules for itself, often using probabilistic models to figure out usage in a more fleshed-out, natural way. This also makes improving the model easier. Instead of writing rules of increasing complexity, simply feed the model more text and let it learn how a human might.

NLP technology has enormous implications for businesses and organizations, specifically in how it allows computer programs to understand unstructured data and leverage it for analysis. By automating workflows of unstructured data, NLP can drive and streamline high-value business decisions. This includes minimizing operational costs, reducing the risk of human error, and gaining visibility and insight into processes to drive decision-making.

SAMPLE USE CASES FOR NLP

Here's a few examples how NLP has been incorporated into the workflows of major businesses.

Maintenance Advisory Application

Using deep learning, SparkCognition developed an advisory tablet application for aircraft front-line staff. This application allowed maintenance technicians to conduct machine-to-human dialogue to troubleshoot asset failures and mechanical issues with high accuracy, assess faults and troubleshoot using queries in natural language, and optimize their workflow and deliver relevant documentation with a faster turnaround. This solution lowered the cost of maintenance and improved asset availability for operators by up to 10%.

Financial Document Classification

SparkCognition is enabling digitization and compliance processes for a major bank with 900,000 contracts under management and a daily global transaction volume of \$3 trillion. Each transaction requires access to a wide range of document types, many of which are unstructured. It takes roughly 1,000 human touchpoints and 72 hours to reconcile a single transaction. Using machine learning and NLP, SparkCognition is extracting information and classifying financial documents to support compliance, with a goal of increasing accuracy and reducing transactions to only 50 human touchpoints.

To learn more about NLP and how it unlocks unstructured data for organizations, visit <https://www.sparkcognition.com/product/deepnlp/>.

ABOUT SPARKCOGNITION

We catalyze sustainable growth for our clients throughout the world with proven artificial intelligence (AI) systems, award-winning machine learning technology, and a multinational team of AI thought leaders. Our clients partner with SparkCognition to understand their industry's most pressing challenges, analyze complex data, empower decision-making, and transform human and industrial productivity. To learn more about how SparkCognition's AI applications can unlock the power in your data, visit www.sparkcognition.com.

REFERENCES

1 <https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/#249dac5a493a>

2 <https://www.veritas.com/news-releases/2016-05-18-new-veritas-research-information-governance>

3 <https://www.kdnuggets.com/2012/12/idc-digital-universe-2020.html>

4 <http://itre.cis.upenn.edu/~myl/language-log/archives/000087.html>

PROBLEM

An institution focused on research into electrical power and related assets and issues needed to deliver faster and more accurate access to detailed technical insight to all its clients and their members worldwide.

PROJECT

AI-powered natural language processing was deployed in a pilot project to create a knowledge management solution that sifts through all relevant information from varying sources and analyzes it to discover and leverage underlying meaning and hidden correlations, thus providing answers to queries.

RESULTS

The organization now has a customized tool that empowers users to find all information required to carry out critical repair or life extension, as drawn from sources such as internal research, conference presentations, and external regulatory sources. Queries previously hard to answer can now be addressed in minutes and the quality of the results is continuously improved with feedback from highly knowledgeable experts, and the system automatically updates with the latest, most reliable information.

THE PROBLEM: IMPROVING SWIFT ACCESS TO DOMAIN KNOWLEDGE AT SCALE AND ACROSS DATA POOLS

For a major US institution that researches electrical power and all related issues and assets, including any and all technical information required to deliver power under challenging conditions, a major challenge came down to two words: accessing knowledge.

While the organization had an enormous amount of relevant information pertaining to key assets, best practices, standard processes, and different deployment scenarios, that information was not always easy to access. It was stored in both structured formats (databases) and unstructured formats (loose files created by a variety of applications).

Furthermore, while the information in any given source often had logical correlations to different information in different sources, those correlations were hard to discover and document in a systematic way. This meant that while the organization's thousand-plus worldwide clients had theoretical access to key information, they didn't have practical access. Though the requisite knowledge was often available in some format or location, finding it was often far from easy. Even if it could be found, logically connecting it to related information provided in another format or location proved difficult.

Another consideration was the fact that the organization's clients in the utilities space usually fell under the purview of strict government regulations. Organizations responsible for maintaining complex energy infrastructures must adhere to the terms of these regulations, incorporating government requirements into both immediate actions taken and the larger strategy.

Finally, the potentially catastrophic consequences involved in widespread energy outages -- up to \$300,000 in lost revenue per day during downtime -- meant that the organization's clients absolutely required quick, accurate answers when emergencies came up. Over time, the difficulty of finding such answers was expected to scale up, not down, as new information continually came in.

It became clear a fundamentally new approach was required to analyze and classify domain knowledge. The organization needed a solution versatile enough to handle all relevant data types, smart enough to ingest and correlate new data as it was added, and flexible enough to address an expanding range of organizational goals.

THE SOLUTION: A SMART, AI-POWERED QUERY ENGINE THAT SUPPORTS ALL DATA TYPES

Toward resolving these issues, the organization decided to create an emergency management response tool (EMRT) to achieve these goals:

- Analyze and correlate all current information, as drawn from both an enterprise gateway to a data lake of silo-specific information and a data warehouse (structured data for applications)
- Ingest new information as it entered the organization from varying sources, in varying formats and structures, regardless of its source or nature
- Provide a user interface intuitive enough to deliver accurate answers even in the case of highly technical queries that required correlation across multiple sources
- At this point the organization conducted an analysis of competing options in this space and selected the SparkCognition DeepNLP™ product as the best available solution.

A number of distinctions and qualifications determined this choice. First, engineering information of this type is generally too complex, technical, and multi-faceted across sources and timeliness (relevance) for any form of plain keyword search to suffice. Users would likely receive too many responses to a given keyword query, and the information in the majority of those responses would not directly apply.

Internet-driven search, though optimized well for most consumer contexts, would not likely suffice either in the enterprise, due to enterprise-level data and usage patterns and the way they vary compared to consumer-level data and usage patterns.

Finally, multi-algorithmic chatbots, despite supporting a relatively high level of natural language comprehension, were soon found to require too much initial effort to implement, especially in supervision and refinement of the data ingestion and classification process. A faster, more automated, and less costly approach was essential.

The DeepNLP product fulfilled all these requirements, standing out in particular in its capacity to ingest and analyze data of any volume level or structure and in a completely unsupervised manner.

Following initial consultation, the SparkCognition team launched a pilot project to design and create the required knowledge management solution.

THE RESULTS: FASTER, MORE ACCURATE QUERY RESULTS AND AN ADAPTIVE, EXTENSIBLE PLATFORM FOR FUTURE GROWTH

Thanks to its open architecture, the SparkCognition deployment can retrieve files drawn from many data sources, both in and out of the host organization itself.

New components can be added in an efficient manner, and in cases where customers want to integrate the EMRT with current solutions, that can be achieved via API (application programming interface) hooks.

Data ingestion is also fully automatic. While other solutions require human oversight to steer or correct data classification, a slow and costly process, the DeepNLP product needs no user invention of any sort to classify and correlate information across the organization's vast total data volume. In consequence, the solution delivers an accurate answer to the query a remarkably high percentage of the time, and the end user's satisfaction level and received value climb in parallel.

Furthermore, results are displayed in a context-appropriate fashion. This ensures that information takes into account the user's job role and access privileges, so that the correct information is made available only to the right people and in the right level of detail.

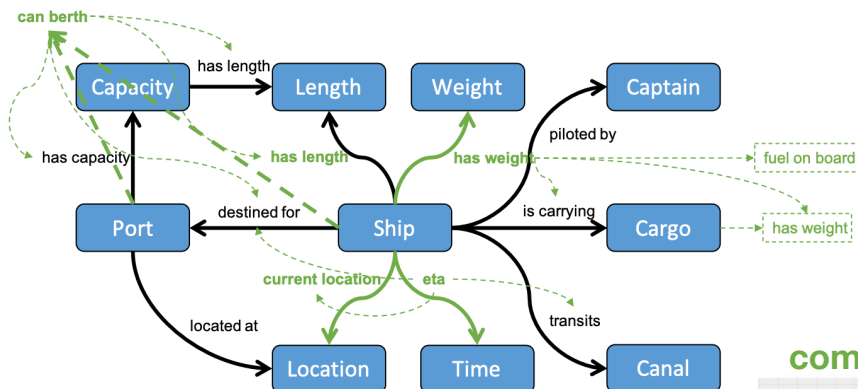
Several compelling additional benefits emerge in consequence from the new AI-powered knowledge management solution:

- For internal users at the organization itself, the solution is essentially autonomous -- a de facto partner in creating value that reduces costs, increases uptime, learns more the more data it receives, and can expand to support new goals as required. This multifaceted range of capabilities has freed the organization to devote more resources such as staff time and capital to other projects of higher immediate business value.
- For more than fifty internal and external utility clients of the organization, there is now far quicker access to necessary technical information -- especially in a potentially catastrophic emergency -- despite the size and diversity of the underlying data pools. Query answers are both substantially more accurate than before and increasingly accurate over time. The full rollout to all the organization's clients is forthcoming.
- For executives at the organization, there will soon also be, following an ongoing deployment to the on-premise environment and general expansion phase, reporting and analytics to assess and quantify user engagement, user satisfaction, and overall value delivered both internally and externally.
- Going forward, four more knowledge bases will be added and integrated to increase the value to clients of ongoing membership in the organization. Beyond the initial connection to these knowledge bases, the logical expansion and analysis of the new data will be automatic, demonstrating the extensible nature of the solution and the way it can dynamically take on both new information and new roles. Innovative dashboards will also be developed that reflect content structure, content utilization, and content shortfalls that require remediation.
- Finally, via new integration pulling visual information from both maintenance and monitoring domains, it should also become easier to correlate new issues with their likely causes, then drive a swift and efficient resolution.

ABOUT SPARKCOGNITION

We catalyze sustainable growth for our clients throughout the world with proven artificial intelligence (AI) systems, award-winning machine learning technology, and a multinational team of AI thought leaders. Our clients partner with SparkCognition to understand their industry's most pressing challenges, analyze complex data, empower decision-making, and transform human and industrial productivity. To learn more about how SparkCognition's AI applications can unlock the power in your data, visit www.sparkcognition.com.

Maana Q is a modern **cloud platform** for rapidly and iteratively **designing and delivering subject-matter expert-driven solutions** to complex business, technical, and industrial **problems**



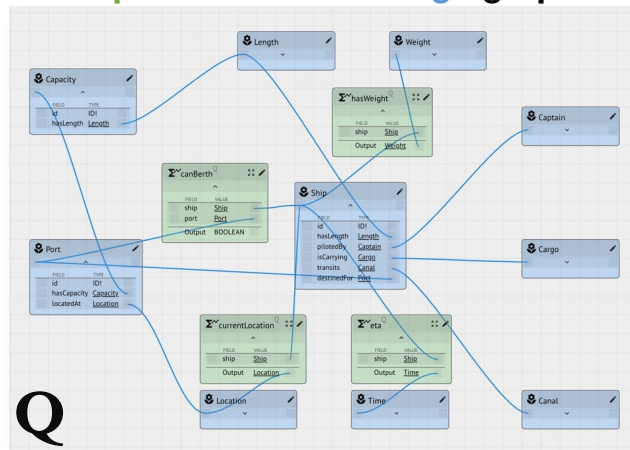
any **edge** or **node** can be the result of an arbitrary **computation**

Subject Matter Experts

Visual Solution Modeling

- Problem-Questions
- Domain Model
- Bayesian Networks
- Goal-Oriented Action Planning

computational knowledge graph



Software Developers & Data Scientists

Microservice & Machine Learning Development

- Any programming language
- Any existing database or (web) service
- Any cloud service
- Any ML library or system (incl. notebooks)

Q is a multi-paradigm Modeling and Simulation Platform

Solutions Team

Model, Implement, Validate

- **Subject Matter Experts**
- Business Sponsors
- Solution Architects
- Software Engineers
- Data Scientists
- Data Engineers

Knowledge Microservices

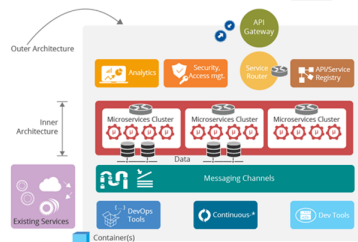
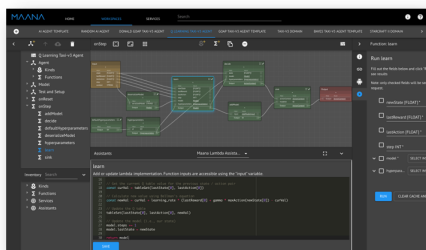
Query, Reasoning, Algorithms, and Data Access

- Trading Models
- Fleet Utilization
- Scheduling
- Safe Routing
- Maintenance Minimization
- Engineering Optimization
- Threat Detection
- Risk Mitigation
- Health, Safety, and Environment
- Supply Chain Demand
- Signal and Risk Minimization

User Experiences

Custom Knowledge Applications

- Web (e.g., React)
- Mobile (e.g., Flutter)
- Desktop (e.g., Electron)
- Mixed-Reality (e.g., Unity)
- PowerApp (e.g., Virtual Agent)



Automated machine learning has the potential to reduce the burden on already overwhelmed teams by automating the main bottlenecks in the data science process, but too many automated machine learning (autoML) applications are simply automating flawed or legacy processes.

To address this need, SparkCognition™ has developed the Darwin® automated machine learning product, which accelerates data science at scale, enabling you to assess the quality of your data set and advising you on how to fix problems to make it suitable for the model-building process. The Darwin product then automates time-consuming tasks that range from model creation and optimization to model deployment and continuous maintenance.

GARBAGE IN, GARBAGE OUT:

Automated Machine Learning Begins With Quality Data

Machine learning methods are highly dependent on the quality of the data they receive as input, but data preparation and cleaning can be an unwieldy task, taking up roughly 60% of the time of data scientists and analytics professionals.

Assessing the Overall Quality of Your Data

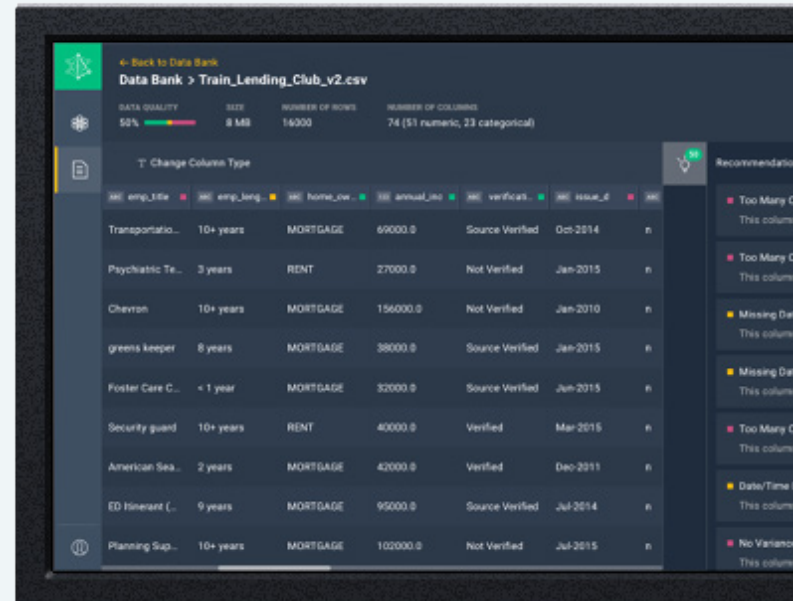
The Darwin product scores your data on its usefulness for the data science process.

- Columns that can be directly used are marked in green
- Columns that will require pre-processing are marked in yellow
- Columns that will be dropped are marked in red

Getting Your Data Ready for Machine Learning

The Darwin product's guided data preparation workflow discovers and recommends solutions for potential problems in your data set, such as:

- Missing data
- Columns with low variance
- Columns with too many categories



DATA QUALITY		SIZE	NUMBER OF ROWS	NUMBER OF COLUMNS
50%		8 MB	16000	74 (51 numeric, 23 categorical)
T: Change Column Type				
emp_title	10+ years	MORTGAGE	69000.0	Source Verified
Psychiatric Te...	3 years	RENT	27000.0	Not Verified
Chevron	10+ years	MORTGAGE	156000.0	Not Verified
greens keeper	8 years	MORTGAGE	38000.0	Source Verified
Foster Care C...	< 1 year	MORTGAGE	32000.0	Source Verified
Security guard	10+ years	RENT	40000.0	Verified
American Sea...	2 years	MORTGAGE	42000.0	Verified
ED Inherent (...)	9 years	MORTGAGE	95000.0	Source Verified
Planning Sup...	10+ years	MORTGAGE	102000.0	Not Verified

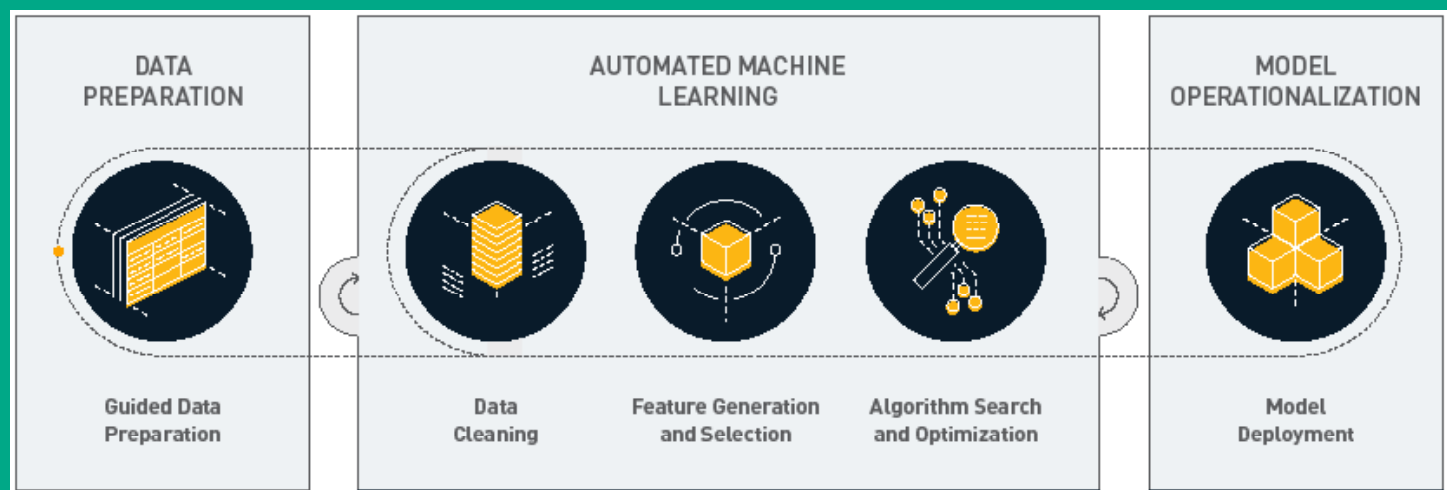
DATA QUALITY	SIZE	NUMBER OF ROWS	NUMBER OF COLUMNS
50%	8 MB	16000	74 (51 numeric, 23 categorical)

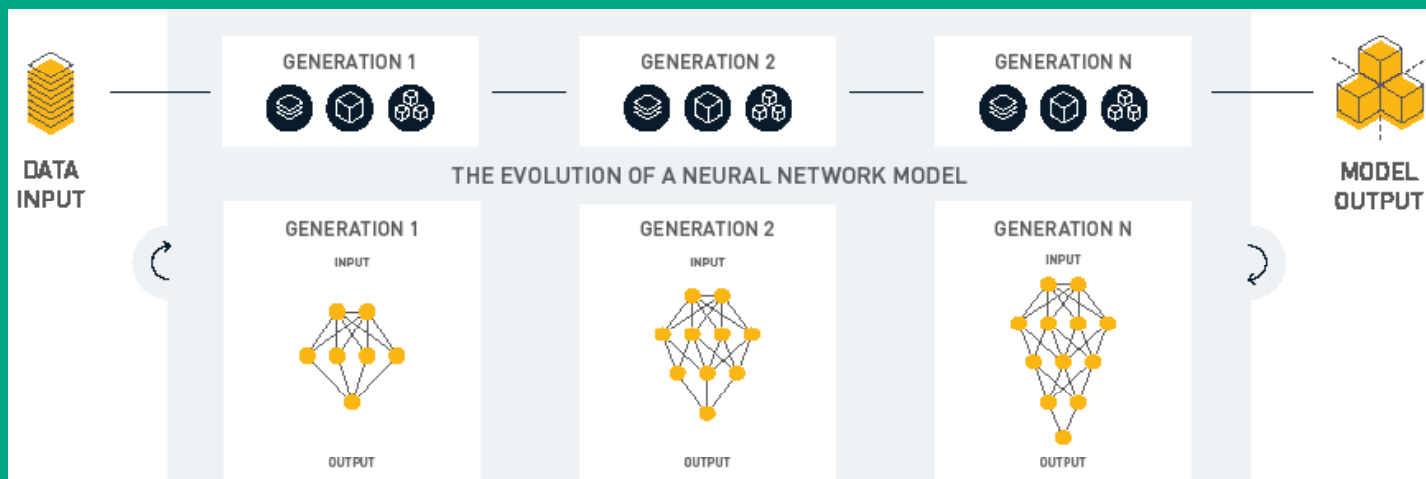
BREAKING THE ALGORITHMIC COOKIE-CUTTER PARADIGM:

Paving the Way to True Generalization With Evolutionary Methods

Most autoML solutions in the market today focus on searching for the best algorithm to fit a given data set. However, these methods can be restrictive, lacking the ability to produce novel, elegant model architectures to solve new problems.

The next evolution in autoML is the ability to create models that do not follow predefined formulas, but rather adapt and evolve according to the problem's data. This is the fundamental operating principle of the Darwin product.





How Does Neuroevolution Work?

The Darwin product uses a patented blend of evolutionary algorithms and deep learning methods. This method specializes in discovering new architectures, while also supporting hyperparameter search for common algorithms such as Random Forest and XGBoost. The Darwin product automates the following steps:

- Execution of the data cleaning profile
- Feature generation to enrich the dataset
- Construction of a supervised or unsupervised model

Using neuroevolution, the Darwin product automatically generates thousands of models that evolve and improve with each generation to more accurately reflect the relationships in your data.

- **Unparalleled Accuracy Through Deep Learning:** The Darwin product's neuroevolutionary process specializes in the search and auto-tuning of neural architectures based on the intricacies of your data
- **Handling of Complex Temporal Relationship:** The Darwin product uses long short-term memory (LSTM) and temporal convolutional network (TCN) architectures to capture complex relationships over time
- **True Generalization to Address the Unknown:** The Darwin product's neuroevolutionary process is based on a fitness function, and quickly adapts to changes in upcoming data to achieve maximum accuracy under dynamic circumstances

HOW TO PUT MACHINE LEARNING MODELS TO WORK:

Bridging the Gap Between Model Production and Operationalization

Despite the great potential of automated machine learning, just having an algorithm isn't enough. 87% of advanced analytics projects never get past the modeling phase to be put into production due to:

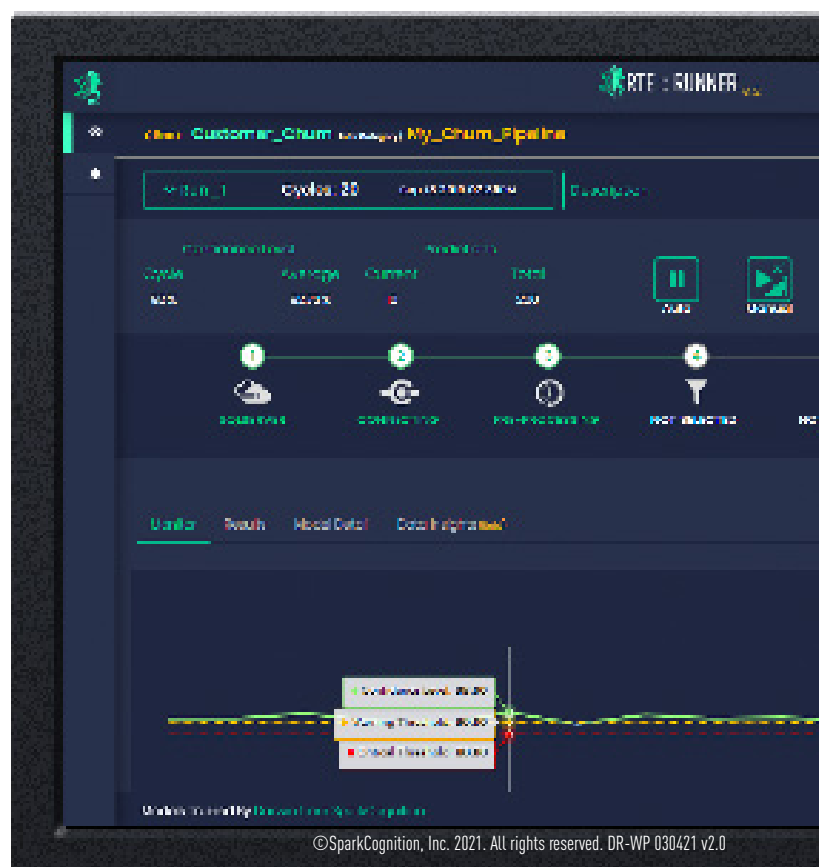
- **Lack of talent and collaboration:** Missing skills in data science, engineering, and IT teams
- **Siloed data:** Data is not easily accessible and/or is in a variety of formats
- **Lack of direction and intent:** Unclear business objectives for the use case
- **Lack of executive alignment:** Leadership is not aligned on competing priorities

How can teams bridge the gap from model to integration and application, and realize the full value of automated machine learning across an organization?

The Darwin automated machine learning product is SparkCognition's answer to this quandary. Using the latest methods in artificial intelligence, the Darwin product not only takes users from data to model, but also guides users through a zero-code deployment, removing the barrier of operationalization.

The Darwin product allows users to:

- **Connect to data sources:** Break data silos with intuitive creation of data pipelines that feed live data into deployed models
- **Automate model execution:** Dynamically create model execution pipelines to obtain real-time predictions on incoming data
- **Monitor model health:** Track the health of deployed models based on the confidence of predictions to inform model maintenance



From here, it's easy to reach the application stage of deployment. The Darwin product can be hooked up to preexisting applications or custom-built dashboards to provide maximum value and scale predictive analytics across an organization.

FASTER VALUE FROM YOUR MODELS

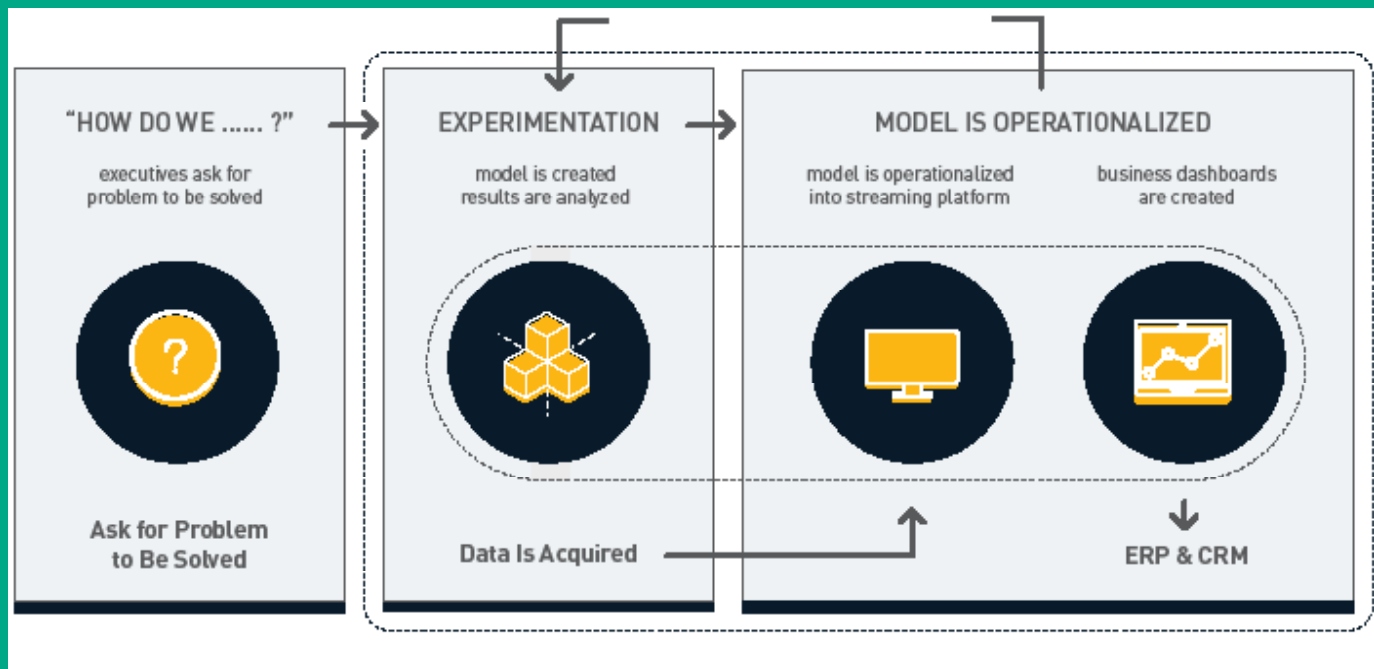
The Darwin product's automated workflows around data quality and model creation allow a faster turnaround of use cases, enabling organizations to more quickly operationalize the output of data science and innovation teams. This approach effectively transforms organizations into factories of use cases that efficiently operate on their data to positively impact what matters: the bottom line.

ABOUT SPARKCOGNITION

We catalyze sustainable growth for our clients throughout the world with proven artificial intelligence (AI) systems, award-winning machine learning technology, and a multinational team of AI thought leaders. Our clients partner with SparkCognition to understand their industry's most pressing challenges, analyze complex data, empower decision-making, and transform human and industrial productivity. To learn more about how SparkCognition's AI applications can unlock the power in your data, visit www.sparkcognition.com.

THE DARWIN PRODUCT: A FACTORY OF USE CASES

- 92% Accuracy: Identification of customers at risk for loan defaults and delinquency
- 94% Accuracy: Identification of customers who are at risk of churning
- 80% Accuracy: Identification of transactions at risk of being past due
- 80% Accuracy: Detection of fraudulent activity on electronic transactions
- Perfect Classification: Classify subterranean drill-head operational states
- 90% Accuracy: Predict automotive sub-component quality during assembly
- 83% Accuracy: Identify degradation in commercial aircraft components
- 90% Accuracy: Detect impurities during iron ore manufacturing



The Darwin® product accelerates data science at scale by automating the building and deployment of models. It provides a productive environment that empowers data scientist with a broad spectrum of experience to quickly prototype use cases and develop, tune, and implement machine learning applications in less time.

The Darwin product's automated model building capabilities offer unparalleled performance to generate highly accurate models using both supervised and unsupervised learning. Given that this technology is only recently available on an enterprise scale, differentiating between machine learning platforms can be difficult. Thus, we sought to compare how the Darwin product performs against other platforms in the market on the same datasets.

The Darwin product clearly emerged as the system that produced the most accurate models, particularly for those involving time-series data, datasets with non-linear relationships, or other complex problems.

RESULTS

We tested the efficacy of the Darwin product against three open-source products—AutoSklearn, H2O, and Random Forest—each using default parameters. Results were compared on 6 different datasets comprising each type of supervised learning problem:

	THE DARWIN PRODUCT	AUTOSKLEARN	H2O	RANDOM FOREST
Electric Devices Classification	0.58	0.13	0.33	0.47
UCI Beijing Weather Time-Series Regression	0.55	0.29	-0.29	0.14
UCI EEG Eye State Classification	0.84	0.47	0.61	0.61
UCI Ozone Classification	0.99	0.98	0.98	0.98
MNIST Digit Classification	0.97	-	0.97	0.87
Boston Housing Regression	0.72	0.72	0.78	0.73

For regression problems we reported results in R2 and for classification we reported weighted F1. These are both measurements of how well the model (built using the training data) performed on the data when tested, meaning they are indicators of how accurately the model will be able to deliver predictions when using live data.

METHODOLOGY

To better understand the types of problems encompassed in this test, it is important to call out a few definitions, shown in the table below:

DEFINITION	DESCRIPTION	EXAMPLES
Classification	Based on a series of inputs, predict a categorical output	Given contorted characters, recognize and classify them into letters of the alphabet
Regression	Based on a series of inputs, predict a numerical output	Given several different variables, predict the price of a house in Boston
Time Series	Given information about historical events, predict them in advance before they occur	Given four years of Beijing's meteorological data collected hourly, predict the atmospheric particulate matter in the future.

	CATEGORY	NUM TRAIN	NUM TEST	NUM FEATURES	TRAIN FILE SIZE (MB)	TEST FILE SIZE (MB)
Electric Devices Classification	Classification	8926	7711	98	6.67	5.77
UCI Beijing Weather Time-Series Regression	Regression	29738	12745	10	2.07	0.89
UCI EEG Eye State Classification	Classification	11984	2996	15	1.46	0.37
UCI Ozone Classification	Classification	2029	507	74	1.15	0.29
MNIST Digit Classification	Classification	60000	10000	785	44.92	7.49
Boston Housing Regression	Regression	404	102	14	0.04	0.01

For time-series datasets, 80% of the data was used for training and 20% was used for testing. Because not all of the comparison tools contained data cleaning methods, all categorical columns were tagged prior to model creation and in some cases encoded using one-hot encoding (Auto Sklearn only).

Each model was run for a maximum of 20 minutes.

CONCLUSION

The Darwin product clearly emerged as the system that produced the most accurate models, particularly for those involving time-series data, datasets with non-linear relationships, or other complex problems. The Darwin product far and away out-performed competitors in four problem sets.

The Darwin product displayed comparable results to competitors on the rest of the problem sets. Note that it might take a data scientist days to come up with such a model. The Darwin product greatly expedites the process of building models by cleansing the data, extracting features, and optimizing models, meaning companies can put these models to use, scale easily, and increase the speed to ROI.

ABOUT SPARKCOGNITION

We catalyze sustainable growth for our clients throughout the world with proven artificial intelligence (AI) systems, award-winning machine learning technology, and a multinational team of AI thought leaders. Our clients partner with SparkCognition to understand their industry's most pressing challenges, analyze complex data, empower decision-making, and transform human and industrial productivity. To learn more about how SparkCognition's AI applications can unlock the power in your data, visit www.sparkcognition.com.

For more information on the Darwin product's performance, or how it can provide actionable insights in operations, please contact info@sparkcognition.com or read our case study.