



Innovation driven by strategy

From curing disease to exploring
the sun, innovative uses of analytics
are changing the world

SAS doesn't settle for ordinary innovation. We put thought and research into ideas that matter – and then create analytics solutions that improve your business – and the world. We take big ideas that shape the future, and make them bigger so you can too.

Where can you find tomorrow's technologies happening today? Look at the way analytics technologies are being used for cybersecurity, machine learning, and the Internet of Things. We cover each of these topics – and more – on the following pages, so keep reading to be inspired.

Table of contents




Exploring the sun with big data

2




The case for cybersecurity analytics

5



Machine learning:
What it is and why it matters

8



Your personal data scientist

12



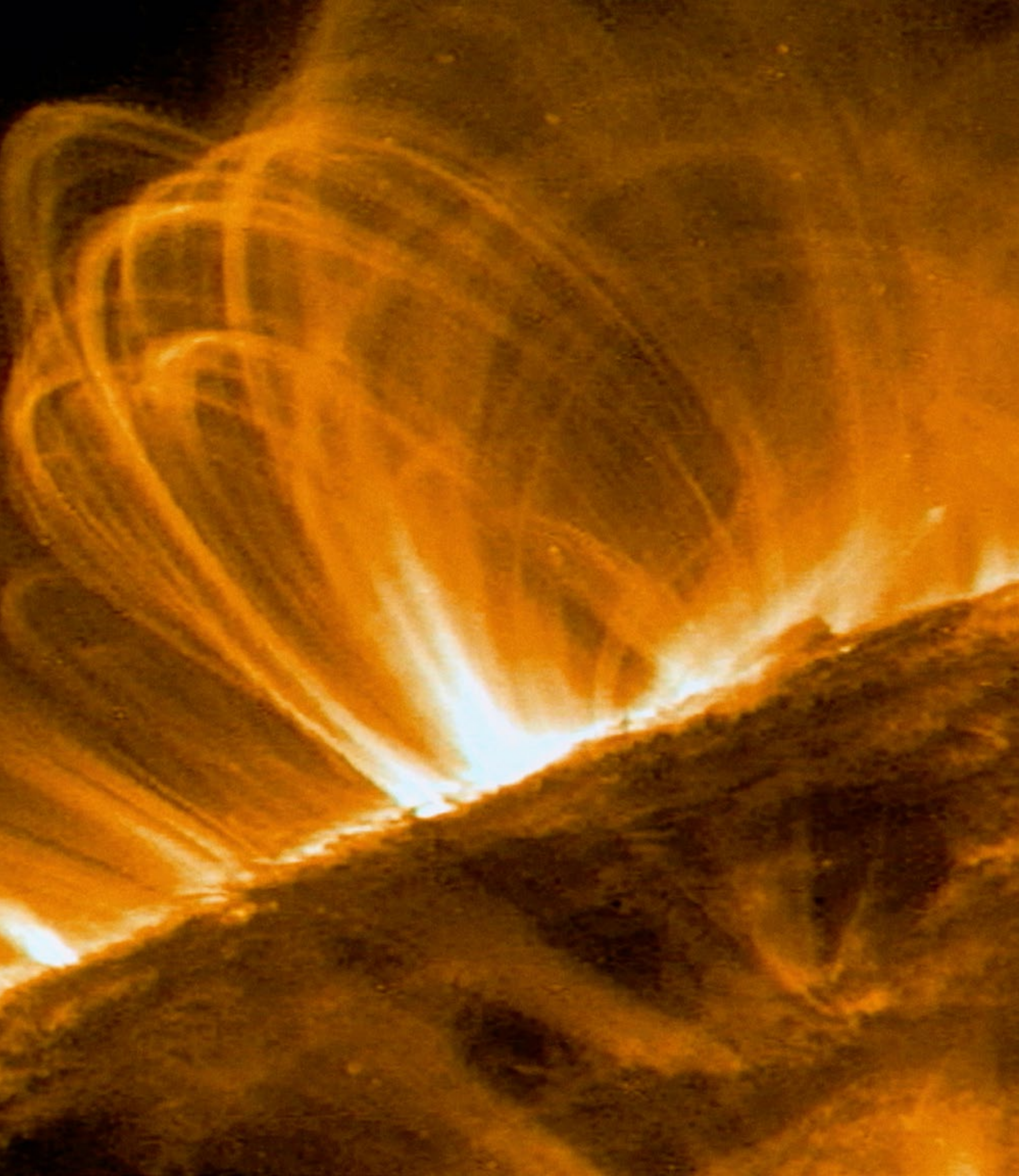
Stopping the Zika virus

15



3 Internet of Things examples
from 3 industries

18



Exploring the sun with big data

Analytical advances help probe the
profound questions of our universe

By Daniel Teachey, SAS Insights Editor

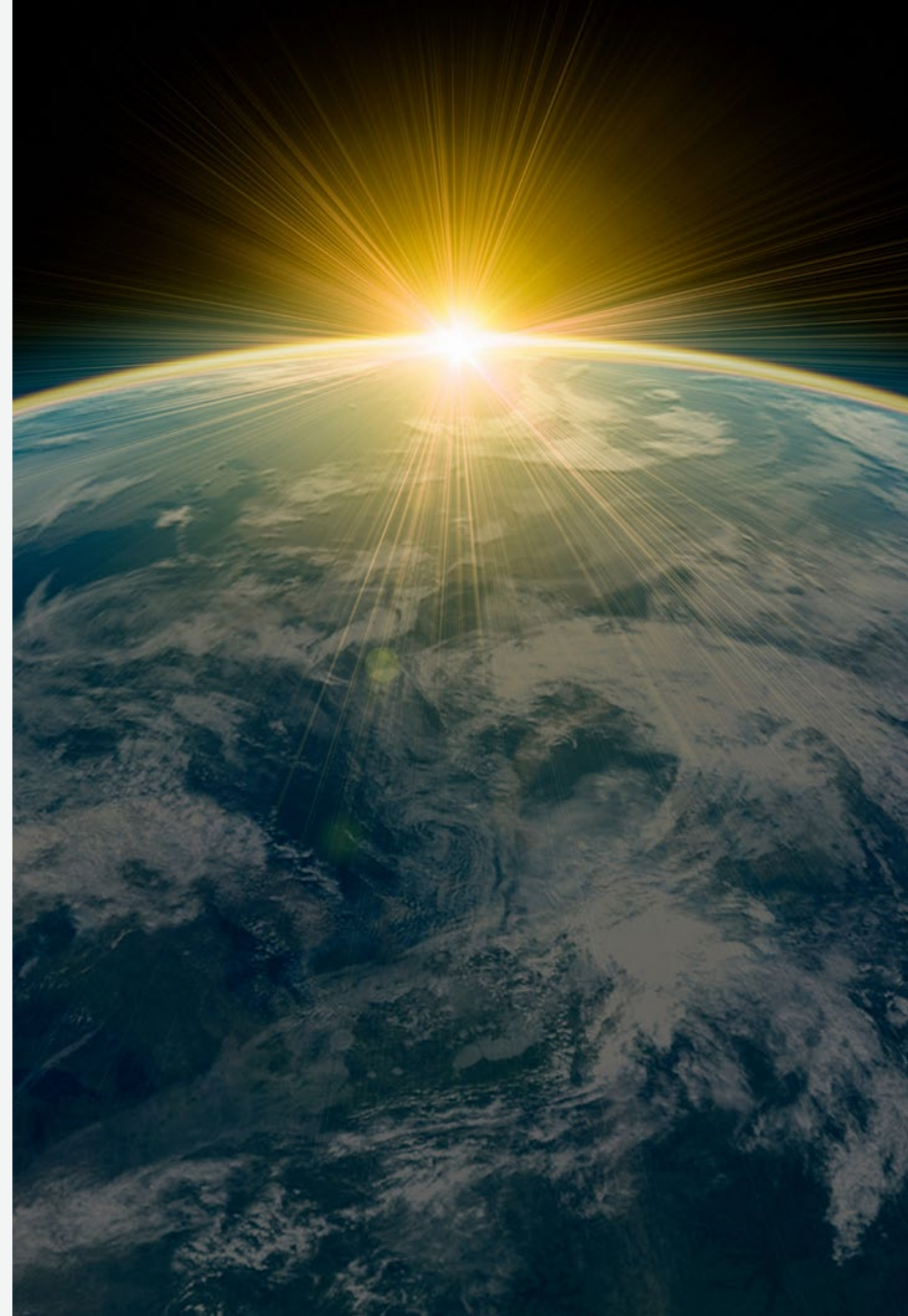
What can we learn from a satellite orbiting 1 million miles from Earth? What's happening on the face of the sun that's turned away from Earth? How does our heliosphere - the magnetic bubble in which our sun and planets reside - interact with interstellar space?

Those are the kinds of questions that confront astrophysicists and astronomers as they explore the mysteries of our universe. One is to solve why the sun's corona - the outermost layer of its atmosphere - is 300 times hotter than the sun's surface, which is already 5,600 degrees Celsius (10,000 degrees Fahrenheit). Scientists suspect that energy trapped in magnetic fields somehow causes this intense heating, and coronal loops are probably key in this ongoing investigation.

Coronal loops are bright, dynamic structures that appear as arcs above the sun's surface. They glow from hot plasma and reach temperatures well above 1 million degrees. The electrified plasma flows along the curving lines of powerful magnetic fields. These luminous magnetic arches are associated with solar spots, which add considerably to X-ray and ultraviolet radiation from the outer solar atmosphere - and into the upper atmosphere of Earth.

The astrophysicist community speculates that when these powerful arches are created, magnetic reconnection occurs - a physical process where magnetic energy is converted to kinetic energy, thermal energy and particle acceleration.

This is the moment that researchers Lars Daldorff and Siavoush Mohammadi want to identify and understand. Daldorff is an atmospheric, oceanic and space sciences research fellow at the University of Michigan and a consultant for NASA's Goddard Space Flight Center. Mohammadi is a consultant with Infotrek, a Swedish business intelligence and data warehousing company.



“Our hope is these results can help with solar magnetic loops research at NASA and, at the same time, our work will show the effectiveness of explorative analysis of data in other data-intensive fields.”

Lars Daldorff,
Oceanic and Space Sciences Research Fellow, University of Michigan

They joined forces to tackle a big stumbling block in the discovery process - the crush of data that threatens to drown scientific processes in information and inhibit the ability to transform it into insight and knowledge.

When physicists use large supercomputers to simulate the sun, it produces massive amounts of data. But the phenomenon of interest is usually located at a specific point in time and space, essentially creating a needle-in-a-haystack situation. What you're looking for is somewhere in the data, but you usually don't know where or even when in the data it can be found. The researcher then has to slice the data into smaller portions based on qualified guesses of where it might be.

The problem is that even if you're lucky and just happen to find an interesting phenomenon on your first guess, you can't be sure that it's the only phenomenon of interest in the data. This means that the time between the gathering of data (from numerical simulations of the sun in this case) and insight about your data becomes very long.

What if you could scan the entire haystack at once to find the needle? What if, after you've found the needle(s), you could simply export the data of interest to do full analysis on only what matters?

Those are the questions Daldorff and Mohammadi sought to answer when they looked to commercial analytics solutions to explore, categorize and display the large amount of solar research project data from the plasma simulations Daldorff had conducted for NASA.

Technical advances known as automatic explorative analysis of data - widely used in the business world, such as to create customer intelligence - gained a new role helping scientists seeking understand our universe.

The duo has been using [SAS® Visual Analytics](#) - a big data discovery, interactive exploration and reporting tool that works in memory. Many of the analytical methods used by SAS Visual Analytics are standardized and used for data analytics in numerous industries. The methods for identifying points of interest, finding relevant data relationships, performing analysis, and creating visualizations and reports are the same, whether you're working with business or scientific data.

“Our hope is these results can help with solar magnetic loops research at NASA and, at the same time, our work will show the effectiveness of explorative analysis of data in other data-intensive fields,” said Daldorff and Mohammadi in presenting their work at the 2015 Joint Statistical Meetings in Seattle. “There are numerous possibilities for this new application that could potentially help various types of researchers - in academia, business and science - obtain quicker insights and results from their research's big data.”



The case for cybersecurity analytics

By Liz Goldberg, Product Marketing Manager,
SAS Cybersecurity Practice

Five years ago, security experts probably didn't view analytics as a primary weapon to help fight cybercrime. The tools focused on attack prevention and were designed to keep people away from sensitive information and network assets.

Those weapons are still useful, but analytics adds a new arrow to the security quiver. It's almost impossible to prevent intrusions. But think about what happens when every successful intrusion occurs. The attacker creates a network event trail. This trail provides a fingerprint of the intruder, marking the steps he's taking in the network to pursue his goal.

The data generated by the attacker's actions is the hallmark of cybercrime. And that's the reason analytics is becoming a valuable weapon in the cybersecurity fight. The data to investigate and fight cybercrime is there. There are multiple systems in place that have the ability to gather and monitor the data needed to fuel faster cybercrime detection.

The goal for your cybersecurity team will be to figure out how to make that data work for you. With more data, you need analytics to organize, contextualize and ultimately find the hidden meaning.

Cybersecurity analytics: How we got here

The use of data to fight cybersecurity threats is nothing new. For years, organizations have used whatever data was available to combat intrusions.

Consider something as straightforward as a log file. A method for documenting a system's events is as old as computer systems and networks themselves. This information has often been a good source for tracking down what happened - after the fact. If a breach occurred in a certain area, the log data could lead back to the point of intrusion.

That information is now even more important for two reasons. First, there are more connections to your network today, including from staff, partners and

The goal for your cybersecurity team will be to figure out how to make that data work for you. With more data, you need analytics to organize, contextualize and ultimately find the hidden meaning.

customers who can access data from outside of your firewall. Second, there are simply more systems and more people accessing these systems, meaning log data is increasing exponentially.

In a recent report by the SANS Institute called [Using Analytics to Predict Future Attacks and Breaches](#), SANS analyst David Shackleford uses this example and others to show the benefits and shortcomings of traditional detection capabilities.

The report evaluates attack detection technologies, tools like logging, network device events, security information and event management, and file integrity monitoring. Each of these systems is important to an organization's network defense arsenal, but there are often limitations to their use in fighting modern cybercrime. Shackleford writes:

"Despite these tools, some security teams are less effective than they could be because these disparate tools and platforms generate an overwhelming amount of data. Security teams are trying to incorporate numerous controls with detection events into their response processes, and it can be easy to miss events and indicators of compromise."

Cybersecurity + analytics = Better network visibility

How does analytics make a difference in the world of cybersecurity?
Here are three areas where analytics can turn massive amounts of data into meaningful information.

1

Establish context: Network data is massive. While this data can tell you a lot, it's important to understand the business context behind the behavior. For example, how is this particular machine acting compared to its peers? With this knowledge, you can better evaluate if that behavior is normal.

2

Provide meaning: The beauty of analytics is that it does the heavy lifting for you. Your security team doesn't have to sift through the data to look for events that raise issues and require additional investigation. More importantly, advanced analytics using modern computing platforms can go deeper into the data to find patterns and connections that might not be available otherwise.

3

Make it visible: It's not just about getting answers. You need to do something with what you learn. Analytics needs to be integrated into your incident response program.

As the SANS report indicates, "... attackers are taking advantage of the fact that organizations are not finding the indicators of compromise within their environments soon enough, nor are they responding to these incidents and removing them quickly enough." With these three guidelines, you can turn cybersecurity analytics into an important method of identifying and remediating these security gaps.



Machine learning: What it is and why it matters

By Kristine Seawell, SAS Insights Editor

Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.

Evolution of machine learning

Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they're able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. It's a science that's not new - but one that's gaining fresh momentum.

While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data - over and over, faster and faster - is a recent development. Here are a few widely publicized examples of machine learning applications you may be familiar with:

- The heavily hyped, self-driving Google car?
The essence of machine learning.
- Online recommendation offers such as those from Amazon and Netflix?
Machine learning applications for everyday life.
- Knowing what customers are saying about you on Twitter?
Machine learning combined with linguistic rule creation.
- Fraud detection?
One of the more obvious, important uses in our world today.

Why is machine learning important?

Resurging interest in machine learning is due to the same factors that have made data mining and Bayesian analysis more popular than ever. Things like growing volumes and varieties of available data, computational processing that's cheaper and more powerful, and affordable data storage.

All of these things mean it's possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results - even on a very large scale. And by building precise models, an organization has a better chance of identifying profitable opportunities - or avoiding unknown risks.

What's required to create good machine learning systems?

- Data preparation capabilities.
- Algorithms -
basic and advanced.
- Automation and
iterative processes.
- Scalability.
- Ensemble modeling.



Did you know?

In machine learning, a target is called a label.

In statistics, a target is called a dependent variable.

A variable in statistics is called a feature in machine learning.

A transformation in statistics is called feature creation in machine learning.

Who's using it?

Most industries working with large amounts of data have recognized the value of machine learning technology. By gleaning insights from this data - often in real time - organizations are able to work more efficiently or gain an advantage over competitors.

Financial services

Banks and other businesses in the financial industry use machine learning technology for two key purposes: to identify important insights in data, and prevent fraud. The insights can identify investment opportunities, or help investors know when to trade. Data mining can also identify clients with high-risk profiles, or use cybersurveillance to pinpoint warning signs of fraud.

Health care

Machine learning is a fast-growing trend in the health care industry, thanks to the advent of wearable devices and sensors that can use data to assess a patient's health in real time. The technology can also help medical experts analyze data to identify trends or red flags that may lead to improved diagnoses and treatment.

Oil and gas

Finding new energy sources. Analyzing minerals in the ground. Predicting refinery sensor failure. Streamlining oil distribution to make it more efficient and cost-effective. The number of machine learning uses for this industry is vast - and still expanding.

Government

Government agencies such as public safety and utilities have a particular need for machine learning since they have multiple sources of data that can be mined for insights. Analyzing sensor data, for example, identifies ways to increase efficiency and save money. Machine learning can also help detect fraud and minimize identity theft.

Marketing and sales

Websites recommending items you might like based on previous purchases are using machine learning to analyze your buying history - and promote other items you'd be interested in. This ability to capture data, analyze it and use it to personalize a shopping experience (or implement a marketing campaign) is the future of retail.

Transportation

Analyzing data to identify patterns and trends is key to the transportation industry, which relies on making routes more efficient and predicting potential problems to increase profitability. The data analysis and modeling aspects of machine learning are important tools to delivery companies, public transportation and other transportation organizations.

What are the differences between data mining, machine learning and deep learning?

Although all of these methods have the same goal - to extract insights, patterns and relationships that can be used to make decisions - they have different approaches and abilities.

Data mining

Data mining can be considered a superset of many different methods to extract insights from data. It might involve traditional statistical methods and machine learning. Data mining applies methods from many different areas to identify previously unknown patterns from data. This can include statistical algorithms, machine learning, text analytics, time series analysis and other areas of analytics. Data mining also includes the study and practice of data storage and data manipulation.



Deep learning

Deep learning combines advances in computing power and special types of neural networks to learn complicated patterns in large amounts of data. Deep learning techniques are currently state of the art for identifying objects in images and words in sounds. Researchers are now looking to apply these successes in pattern recognition to more complex tasks such as automatic language translation, medical diagnoses and numerous other important social and business problems.



Machine learning

The main difference with machine learning is that just like statistical models, the goal is to understand the structure of the data - fit theoretical distributions to the data that are well understood. So, with statistical models there is a theory behind the model that is mathematically proven, but this requires that data meets certain strong assumptions too. Machine learning has developed based on the ability to use computers to probe the data for structure, even if we do not have a theory of what that structure looks like. The test for a machine learning model is a validation error on new data, not a theoretical test that proves a null hypothesis. Because machine learning often uses an iterative approach to learn from data, the learning can be easily automated. Passes are run through the data until a robust pattern is found.





Your personal data scientist

Introducing an analytical assistant for the masses

By Alison Bolen, SAS Insights Editor

My kids love Siri, Apple's virtual assistant. They ask her about homework, sports scores and science experiments. They ask her questions that range from the mundane (What is your favorite kind of dog?) to the profound (What is free will?). And, within reason, she answers.

I'm a more practical user of Siri. I ask her to look up recipes, type texts, check my calendar and find directions all while I'm busy with other activities, like driving or cooking or chasing my kids.

Siri and similar systems - like the Amazon Echo, Google Now and Microsoft's Cortana - have been described as personal assistants, and they're becoming quite commonplace in our everyday lives. If you watch the popular video introducing the Amazon Echo, you see a family using that service throughout the day to keep grocery lists, set reminders and play music.

Powered partly by natural language processing and [machine learning](#), these systems translate your spoken words into a computational query, find an answer among billions of data points, and provide the answer back in the same language you used to ask the question.

What if Siri were a data scientist?

The next step in the evolution of these personal assistants, especially for business decision makers, might be to use them to query large corporate data sets or even public data sets like social media feeds and government data.

Imagine pushing a button on your desk or your smartphone and asking any of the following questions:

What are the sales forecasts through the end of this year for our 10 largest regions?

What are the top three adjectives that customers are using to describe our new product on social media today?

Which marketing programs are generating the most income this quarter?

Show me the problems that customers are contacting us about this week through all of our customer contact channels.

Using natural language processing, your personal data scientist turns your request into a query, searches all known databases for an answer, applies analytics to the data, and presents you with an answer - and perhaps a chart or data visualization that you can then drill into further.

"What we're describing is a cognitive learning system that uses automation and natural language process questions and answers," says Wayne Thompson, Chief Data Scientist at SAS. "Machine learning helps in the process of articulating how to return results and interpreting what other things are not explicitly requested that would make sense for the purpose of representing results to the user."

What else would you like to know?

After you've evaluated the initial results, your personal data scientist can provide links to related answers or charts to encourage further exploration. The system might say, "Would you also like to know what customers are saying about competitors on social media?" Or, "Would you like me to forecast sales into next year?"

These prompts help can help you explore the data further, especially if you aren't familiar with the data or the available types of analyses.



“The goal for these systems would be to put analytics into the hands of the masses. Data scientists are still needed to build models and to develop complex algorithms, but this type of system makes analytics more accessible to the average person.”

Wayne Thompson, Chief Data Scientist, SAS

Sharing your results

And finally, if your query is new or related to something that somebody else has asked, your personal data scientist might say, “Would you like me to share these results to the portal?” Or, “Would you like to share these results with the VP of sales?” That way, important information is spread further into the organization so that others with an interest can benefit from it as well.

“Data scientists have the acumen and the ability to process data and know what to look for in specific types of data sets,” says Mike Frost, a Senior Product Manager at SAS. The goal of a personal data scientist would not be to replace these specialized people but to bring some of their more basic skills to the masses.

Thompson, a [data scientist](#) himself, is not afraid of being replaced. “The goal for these systems would be to put analytics into the hands of the masses. Data scientists are still needed to build models and to develop complex algorithms, but this type of system makes analytics more accessible to the average person.”

Next steps from your personal data scientist

If you don’t know what to ask of your data, you could also input a general topic by saying, “I’m interested in our supply chain,” and the system would share some basic insights on that topic and provide tips for further investigation.

No matter how you pose the question, natural language processing takes your statement and turns it into a series of queries the computer can understand. Machine learning processes the request, returns the results and looks for related things that are not explicitly requested, but might make sense to the requestor. Essentially, the system is predicting what other things you might ask next.

For some complex questions, your personal data scientist might produce a gallery of charts or results, to give you multiple views into the data and provide tips to probe deeper into the data.

The personal data scientist is a low-level BI data discovery tool that learns the more you use it. It uses prebuilt machine learning models and maps the request from the user into a pipeline of models and data. The more you interact with the system to fine-tune your requests, the more it continues to learn.

What would you ask of a personal data scientist? And how can you see it being used to spread the use of analytics in your industry?



Stopping the Zika virus

The potential of big data, analytics

By Daniel Teachey, SAS Insights Editor

Tracking the spread of disease is one of the most powerful tools for health care officials. This information can help practitioners in a region – or around the world – prepare for an outbreak and mitigate its spread as much as possible.

However, the goal for health care organizations is to go from reacting to anticipating the spread of diseases. With analytics, that's becoming more of a reality. More information, available sooner, can improve health outcomes and save lives.

Today the focus is on the Zika virus, spread primarily through the bite of the Aedes mosquito. As of March 2016, Zika is present across Africa, Asia and the Americas, with the heaviest number of cases in Brazil. The World Health Organization (WHO) declared the Zika virus a public health emergency that could affect 4 million people in the next year as it spreads across the Americas. At the same point, more than 80 cases of travel-related Zika have been confirmed in the US, according to the Centers for Disease Control and Prevention (CDC).

As the world scrambles to fight the outbreak, analysts at SAS are investigating how data mining technology can track diseases like the Zika virus and help create a different type of response to global disease outbreaks. Knowledge about the disease and how it's spreading is more readily available than ever before. And the more you know about how the disease is spreading and where it's likely to go next, the more effectively you can mobilize resources and develop strategies to combat it.

"From a technological standpoint, we already have everything we need to leverage big data to quickly and effectively develop vaccines for new viruses such as Zika," said Bernard Marr, author of *Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance* (Wiley, January 2015). "Ebola showed us that every aspect of a virus'

"From a technological standpoint, we already have everything we need to leverage big data to quickly and effectively develop vaccines for new viruses such as Zika."

Bernard Marr,
Author of *Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance*

behavior and characteristics can be isolated and identified. Now what we need are platforms and systems to get this data into the hands of those who can develop solutions before a public health emergency develops."

This type of predictive analysis requires multiple, disparate data sources. The value of bringing together diverse data communities has already been shown in initiatives such as [Project Data Sphere](#), which tackles cancer, and [ClinicalStudyDataRequest.com](#), which shares anonymized results of clinical trials to support third-party research. Rapid and unexpected discoveries occur when stakeholders are invited to take part in open, collaborative projects involving massive, shared data sets and big data analytics.

Think of the possibilities. Analytics experts join forces with local health agencies, CDC, WHO, the academic research community and vaccine makers. This collaboration would bring together diverse data – and perspectives – to support analytical insights that are not possible when viewing data sources in isolation.

The data at the core of this initiative could include lab data from those being screened for Zika and other diseases, clinical trial data, data from surveillance activities and provider networks, and even social media trends. All of these can play a role in fighting the spread of the outbreak.

Tying it all together could help speed up the process of developing a vaccine. By analyzing data from thousands of points around the world, compounds can be developed to target the specific proteins that enable the virus to thrive.

“We’re not talking about solving Zika tomorrow or the next day, but we are talking about bringing all the relevant knowledge together into a central place,” said Jamie Powers, a SAS health care industry consultant.

“This collaborative, analytics-driven approach would have implications not only for fighting the spread of Zika but also for getting ahead of future outbreaks. I would envision a future where, when the next big outbreak is coming, we have the infrastructure in place to rapidly identify and respond.”





3 Internet of Things examples from 3 industries

Real-world IoT implementations achieving results today

By Alison Bolen, SAS Insights Editor

Is the Internet of Things merely a far-fetched consumer fantasy that promises the convenience of connected appliances and smart running shoes? Or is it a business opportunity for companies that want to collect real-time information about almost every aspect of their business?

We tend to hear a lot about the consumer applications of IoT, but many early adopters in the IoT revolution have been businesses and government organizations with an interest in collecting and analyzing data about their operations. From the temperature of equipment to the performance of a fleet of wind turbines, IoT sensors are already delivering valuable information in many industries. Blue Hill Research recently conducted an in-depth qualitative research report about three Internet of Things examples, which we've summarized in this article.

Internet of Things examples from government, utilities and manufacturing

Consider these three Internet of Things examples:

A US municipality has implemented smart meter monitoring for all the town's residential and commercial water meters. The project involved placing water meter sensors on 66,000 devices that used to be manually read and recorded.

A US oil and gas company is optimizing oilfield production with the Internet of Things. In this IoT example, the company is using sensors to measure oil extraction rates, temperatures, well pressure and more for 21,000 wells.

An international truck manufacturer created a new revenue stream by outfitting trucks with sensors for predictive maintenance. The system automatically schedules repairs when needed, and orders the required parts for the repair. More than 100,000 trucks have been outfitted with devices that transmit more than 10,000 data points a day for each truck.

As you can see in the table below, the data streams for each of these applications create more than a million data points per day.

Table 1: Case Summaries

	US Oil & Gas Company	US Municipality	International Truck Manufacturer
Primary Data Source	Sensors on oil injector wells	Water meter sensors	Sensors on truck fleets
Number of Monitored Devices	21,000	66,000	100,000
Primary Activities Monitored	Oil extraction rates, temperature, and well pressure (10 total activities)	Water usage rates	Engine diagnostic codes and function of mechanical parts
Frequency of Readings	90 x day x activity	1 x hour x day	10,000 x truck x day
Scale of Data Collected	~18,900,000 daily readings	~1,584,000 daily readings	~1,000,000 daily readings

Source: Blue Hill Research, September 2015

The ROI of IoT

How are these three companies converting raw IoT data into business insights and tangible benefits? They're using analytics to realize both direct and opportunity costs associated with analyzing IoT data.

The US municipality that switched to smart meters for its water usage monitoring saw immediate and sustained savings. Its data collection process evolved from a manually intensive process (in which field technicians traveled to every meter) to one where meter readings were automatically

“The US municipality that switched to smart meters for its water usage monitoring saw immediate and sustained savings... projecting a total savings of \$28 million and a net savings of approximately \$10 million over the lifetime of the initiative.”

recorded and transmitted to a central database. This saves a lot of money, both in work-hours and in field equipment, such as trucks. The town is projecting a total savings of \$28 million and a net savings of approximately \$10 million over the lifetime of the initiative.

The indirect savings came when the organization was able to make a fundamental shift to a proactive service-oriented organization. Now the town can identify issues within hours, rather than weeks or months. With better and more accurate data, the town proactively reaches out to households to mitigate overuse or unexpected fees. The billing and management teams have shifted from an internal reporting organization to a customer-facing hub that provides residents a markedly better experience.

Likewise, the oil and gas company is able to monitor the performance of oil wells at the end of every day or week. This allows it to identify opportunities for improvement (such as increasing production levels) and areas of potential concern. Ultimately, the company can take this information and disseminate it to field crews to make adjustments or repairs. The result is reduced downtime and increased production levels. The company estimates that it loses \$500 for every hour that a single oil well is not in operation. After analyzing the initial impacts of sensor deployment, the organization estimates that quicker oil well repairs saves approximately \$145,000 in cost avoidance per month per field.

The international truck manufacturer provides a mature example of using sensor data. Sensors in the trucks, combined with predictive models, detect

when a mechanical failure is likely to occur. When this happens, the system schedules a maintenance appointment for the truck based on the truck's route and optimized for scheduled delivery times. Further, the system orders and ships the appropriate parts to the identified service center, and then notifies technicians about what needs to be fixed. The result is an interconnected web of sensors and operational systems that communicate to save time and money across the operation.

In each of these cases, bringing the Internet of Things and industrial-grade analytics together yielded significant and persistent business enhancements. The key to extracting sustained business value from IoT initiatives is, ultimately, sound business analytics practices.

What are you imagining? What do you want your data to do? How can you combine your data with other data sets to do something bigger or better than before?
If you can imagine it, we can innovate for it.

Learn more about innovating with analytics

Follow us:



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

