Accenture Labs

UNDERSTANDING MACHINES: EXPLAINABLE AI

accenture

EXECUTIVE SUMMARY: EXPLAINABLE AI—THE NEXT STAGE OF HUMAN-MACHINE COLLABORATION

Accenture's Technology Vision 2018 maps out a future where the powerful potential of the intelligent enterprise is unleashed, enabling companies to improve the way we live with new products and services that will become indispensable. AI will play a key role in turning this vision into reality. In this world humans and machines will collaborate, each exploiting their respective strengths.

But before this vision can come to pass, key aspects of the way humans interact with AI systems must be addressed. Specifically, many of today's AI applications created to date are "black boxes" in terms of their decision—making, meaning they're unable to explain their reasoning.

An inability to explain the rationale behind decisions is acceptable when the impacts of AI decisions are relatively trivial, such as a recommendation on Spotify. But as AI expands into areas with major impacts on humans, like medical diagnosis, the military, recruitment decisions and policing, it become increasingly vital that AI can explain why it has reached a particular decision.

For example, imagine you've applied for a mortgage and the AI-enabled decision engine turns you down. Or that you're a doctor using AI-enabled sensors to examine a patient, and the system comes up with a diagnosis demanding urgent invasive treatment. In situations like these, an AI decision alone is not enough. We also need to know the reasons and thinking behind it.

As this paper describes, this need can be met—by giving AI applications the ability to explain to humans not just what decisions they have made, but also why they have made them. Indeed, the transition to Explainable AI is already under way, driven and accelerated by three key factors.

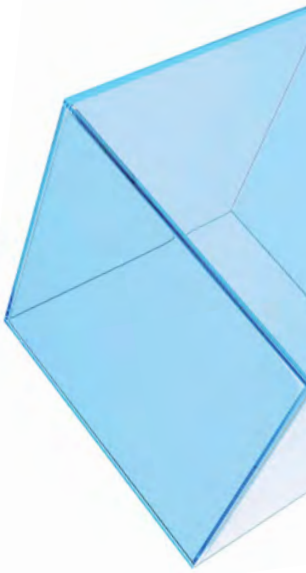
The first is the growing demand for transparency in AI decisions. The second is the need to build trust in AI, since people must trust the choices made by AI for those choices to be truly useful. And the third driver is the opportunity for closer and more productive human-machine synergies in the future, with AI and humans each augmenting the capabilities of the other.

The new generation of Explainable AI applications is already beginning to emerge. As Explainable AI evolves and expands to become the norm, we'll find ourselves amid a new technology revolution: one with people at its heart.

> **"The future of AI lies in enabling people to collaborate with machines to solve complex problems. Like any efficient collaboration, this requires good communication, trust, clarity and understanding.**
>
> **— Freddy Lecue, Explainable AI Research Lead, Accenture Labs**

Introduction:

FROM "BLACK BOX" TO GLASS BOX



The recent upsurge in AI has focused mainly on one specific area of AI: machine learning, and especially deep learning. Most of the applications that the public at large currently regard as being in the vanguard of AI are in the business-to-consumer space. Their "intelligence" largely involves an ability to undertake low-level pattern recognition tasks like image recognition, speech recognition and natural language processing.

In most cases, these applications focus on human digital traces as opposed to (1) the deluge of business transaction data and (2) the deep domain knowledge expertise that underpins individual industries. They apply machine learning trained on large volumes of data to make niche predictions and recommendations with limited business impact and varying degrees of robustness, reflecting their relatively high sensitivity to "noise". And they do all this without telling the humans involved how or why they reach their decisions.

This lack of an explanation means that these machine learning systems are "black box" AI systems. As humans, we can see the inputs as raw data, and the outputs as models or predictions. But we do not receive a human-friendly explanation on the rationale behind the output. There are examples all around us: think Google searches, recommendations on Amazon or Netflix, or automatically-generated playlists on Spotify.

The AI stakes are rising

The fact that these systems leave the human user out of the loop is not a problem, because their recommendations have an insignificant impact on individuals, and—in particular—do not affect them financially, politically or in terms of personal wellbeing.

In cases where recommendations have vital implications, and even a legal requirement—note the General Data Protection Regulation's (GDPR's) provisions on the "right to explanation"—it is crucial that you understand the reasoning behind the machine's decisions. In other words, the AI has to explain itself, by opening up its reasoning to human scrutiny.

That's Explainable AI. It's the next stage of human augmentation by machines, when AI will empower humans to take corrective actions according to the explanations given. Within three years, we believe it will have come to dominate the AI landscape for businesses—because it will enable people to understand and act responsibly, as well as creating effective teaming between human and machines.

In this paper, we'll examine why going beyond machine learning and pattern recognition to Explainable AI is the way forward—and will enable us to create technology that's truly for people.

EU GDPR: "RIGHT TO EXPLANATION"

As well as being a practical imperative, explainability will also be required because of ethical or legal requirements, such as the introduction under the EU's forthcoming General Data Protection Regulation (GDPR) of the "right to explanation" about algorithm-derived decisions. But most importantly, explainability puts people in control—meaning AI augments human skills rather than trying to replace them. For all these reasons, AI needs to go beyond machine learning to the next stage: Explainable AI.

https://www.eugdpr.org/

1 THE DOMINANCE OF MACHINE LEARNING TODAY

Impressive results across industries

We are currently seeing is a surge of innovation and uptake focused on just one area of AI—machine learning, or more specifically deep learning—which is most successful in low-level pattern recognition tasks from image, video, speech or text.

That said, there's no doubt that today's machine learning systems are achieving impressive results, having demonstrated wide applicability with real-world impact in many contexts. Some of machine learning's highest-profile successes have been in mastering complex games—including the likes of jeopardy and go—in which they've even beaten some of the world's foremost human players.

We've also seen machine learning make great strides in a number of industries. In healthcare, more than 100 companies are now applying machine learning algorithms and predictive analytics to reduce drug discovery times, help doctors devise care protocols for cancer patients, and diagnose ailments from medical images. In manufacturing, AI has advanced from using robots to assemble and package products, to deliver packages using drones. And in transportation, personal self-driving cars using machine learning are expected to be on the roads in the next couple of years, with commercial applications close behind.

> **"There's no doubt that today's machine learning systems are achieving impressive results, having demonstrated wide applicability with real-world impact in many contexts.**
>
> **— Dadong Wan, Explainable AI Research Lead, Accenture Labs**

A central role in our daily lives

For the majority of people—as we've already highlighted—the most familiar application of machine learning is the decision and recommendation engines used by an expanding array of online consumer-focused services. Thanks to developments in speech recognition and natural language processing, machine learning can now not only automate text questions and chats with customers, but can create responsive, friendly robots that mimic human speech patterns to boost convenience for customers and reduce cost for companies.

Machine Learning can also optimize operations across a wide spectrum of sectors, such as supporting on-time performance in the air traffic control industry, and detecting and diagnosing defects in manufacturing supply chains.

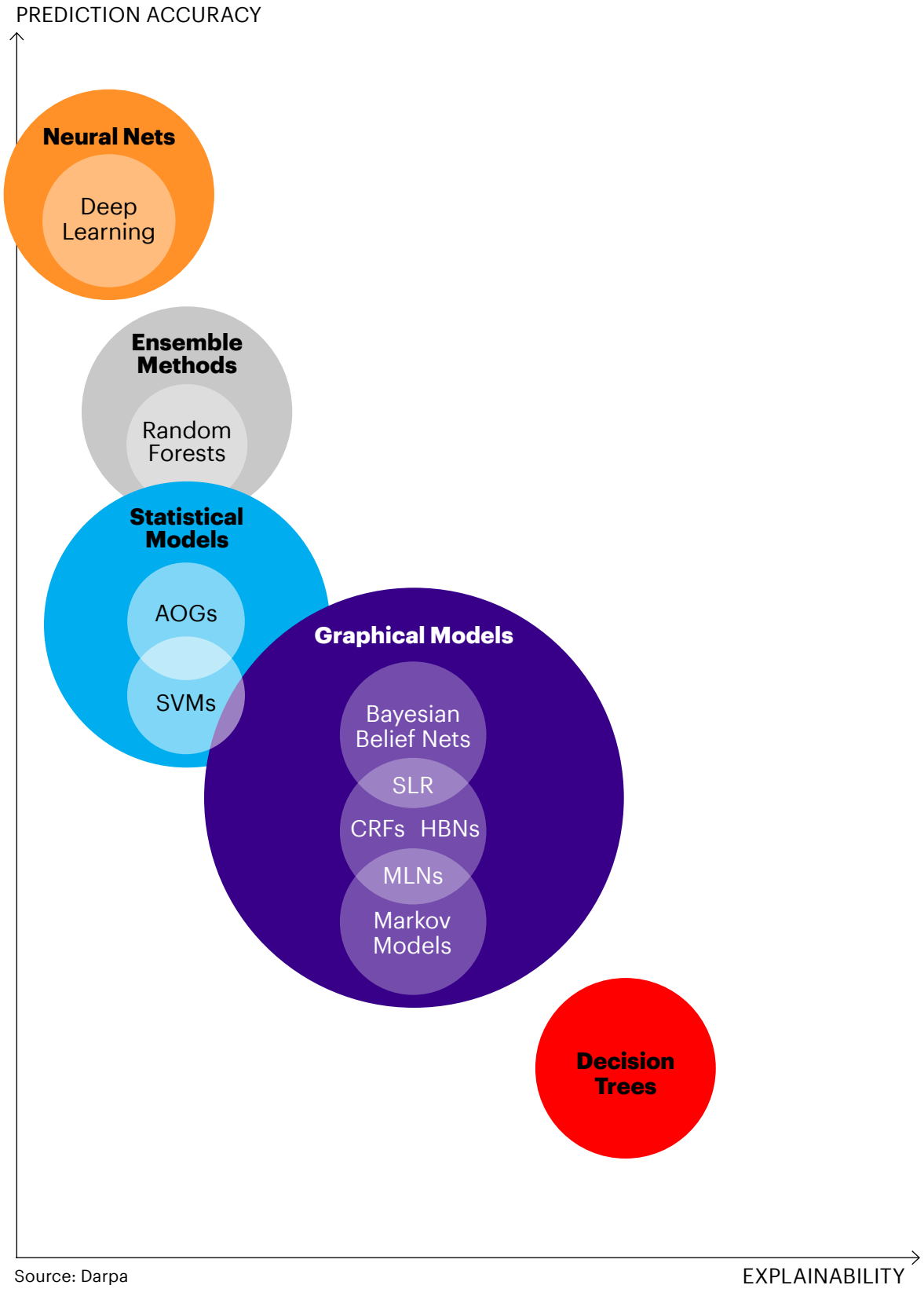
2 TOWARD EXPLAINABILITY: WHY THE "BLACK BOX" MUST BE MADE TRANSPARENT

While "black box" AI is clearly limited by its inability to explain its reasoning to human users, it can actually work well in three types of application.

The first is simple pattern recognition tasks where the cost of failure is low. The second is "closed-loop" systems where a real-time response is critical and/or the pace of decision-making is too fast to allow for human intervention, such as real-time pricing (Amazon), movie or music recommendations (Netflix, Spotify, etc), or driverless cars. The third is interactive response systems like robots and chatbots.

However, in many other applications—especially business situations where an explanation of the underlying reasoning is critical for decision-makers—machine learning's lack of explainability is a huge barrier to the adoption of AI. Figure 1 illustrates some of state-of-the-art techniques in the accuracy/explainability trade-off map, where tools such as decisions trees have more explanation power than neural networks.

FIGURE 1. Existing Models and their Interpretation



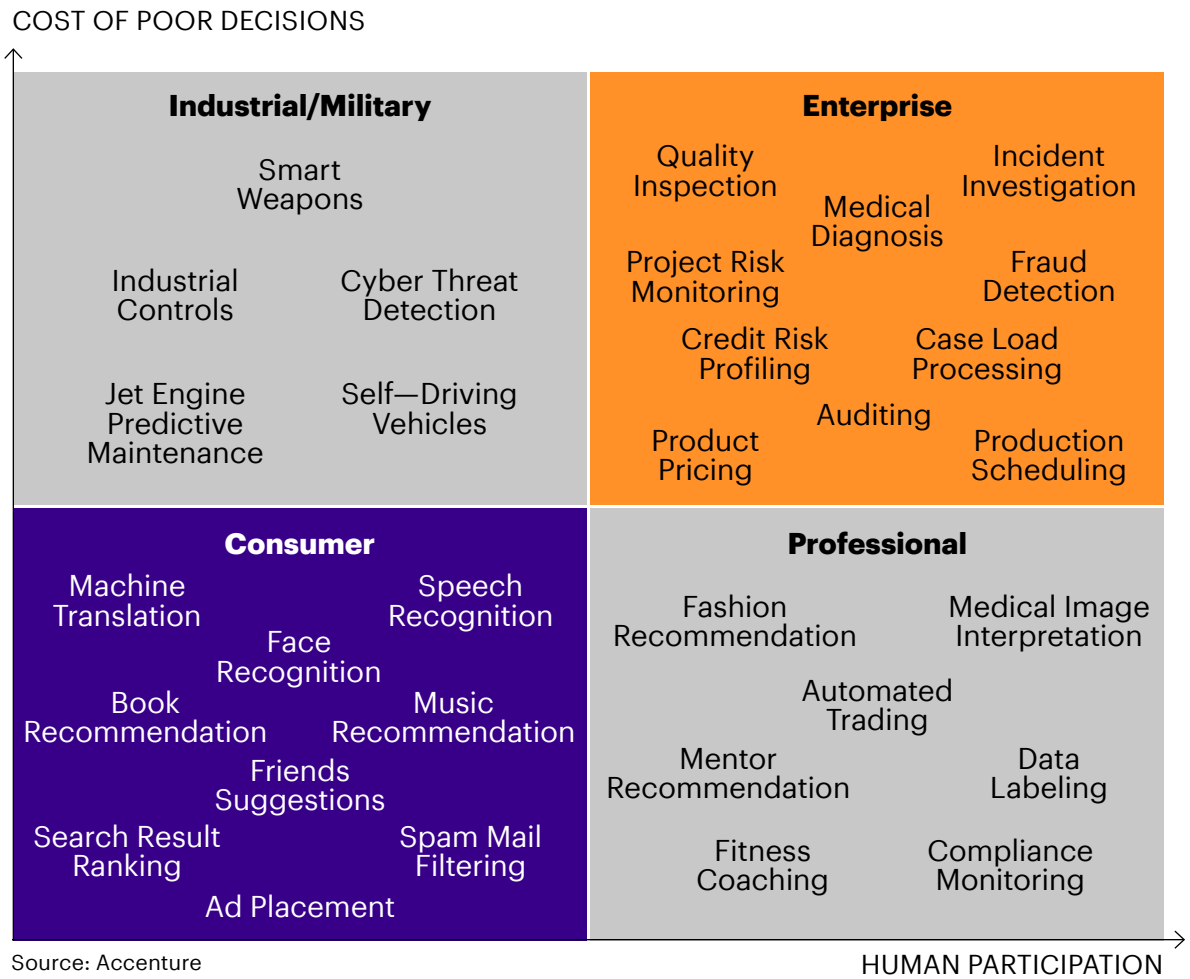
Source: Darpa

The rising cost of poor decisions

The underlying dynamic is that as the human impacts and implications of AI decisions increase, so does the need to explain. Figure 2 maps the potential cost of poor decisions against the human participation in those decisions. Currently, the most prominent successes of AI to date are in the bottom-left quadrant, where potential costs and human participation are both low.

However, as we move increasingly into the top-right quadrant of decisions—including things like fraud detection, credit risk profiling and incident investigation, as well as medical diagnosis—the potential cost rises dramatically. In the top-left quadrant, which includes military decisions, any mistakes are potentially catastrophic. If AI lacks the ability to explain itself in these areas, then the risk of it making a wrong decision may outweigh the benefits it could bring in terms of the speed, accuracy and efficiency of decision-making. The effect would be to severely limit its usage.

FIGURE 2. The need for explainable AI rises with the potential cost of poor decisions



Source: Accenture

3 WHY IS EXPLAINABLE AI THE FUTURE—AND WHAT WILL IT DELIVER?

Making humans super, not super humans

In the Accenture Technology Vision 2018, we present our view of how technology will augment and enhance human skills to enable us all to listen more closely to customers and employees, connect to them on their own terms, and partner with them to achieve personal goals.

It's a future of "Citizen AI", where AI is here and ready to work alongside its human counterparts. Where companies embrace AI and recognize it as a partner to their people. And where, by raising AI for responsibility, fairness, and transparency, businesses can create a collaborative, powerful new member of the workforce. Explainable and more responsible AI will be the backbone of the intelligent systems of the future that enable the intelligent enterprise.

In playing this role, Explainable AI won't replace people, but will complement and support them so they can make better, faster, more accurate and more consistent decisions.

Explainable AI systems will play this pivotal role through their ability to:

Explain

their rationale;

the reasoning, whenever needed;

Characterize

their strengths and weaknesses

Compare

with other AI systems

Convey

an understanding of how they will behave in the future

Make

the enterprise scalable through intelligent decisions;

decisions smarter by augmenting humans with machines.

Starting the journey

The enhancement of today's machine learning-dominated AI with tomorrow's more Explainable AI is already under way, as the accompanying information panel on the DARPA Explainable AI program underlines. What DARPA and other R&D pioneers have grasped is this: while today's human decision-makers can easily make use of predictions from AI and machine learning systems, those predictions are only useful if those humans can justify their trust in the prediction. And with "black box" machine learning, they can't.

As a result, in many domains—especially life-changing ones—AI models that can't explain their reasoning will become discredited, disregarded or even rejected, even if their decisions are of a consistently high quality. An obvious example is medical diagnosis, where both the context and the rationale of any prediction must be understood by humans, as the consequences of any error could be disastrous.

DEFINING EXPLAINABLE AI:

The US Defense Advanced Research Projects Agency (DARPA) has set up an Explainable AI Program that "aims to create a suite of machine learning techniques that:

Produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and

Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners."

http://www.darpa.mil/program/explainable—artificial—intelligence

Accelerating Explainable AI's take-off

As the move to Explainable AI gains pace and momentum, three factors are accelerating its progress:

First, the growing need for transparency, as required by laws such as the EU's GDPR, mandating how personal data is used for selection and other decision-making. Ethical values also demand transparency, so people can be sure decisions are fair and even-handed.

Second, trust. Before humans can act on a system's recommendations, people need to trust it. This trust will only be created if the system can explain the underlying model and process through which decisions are made.

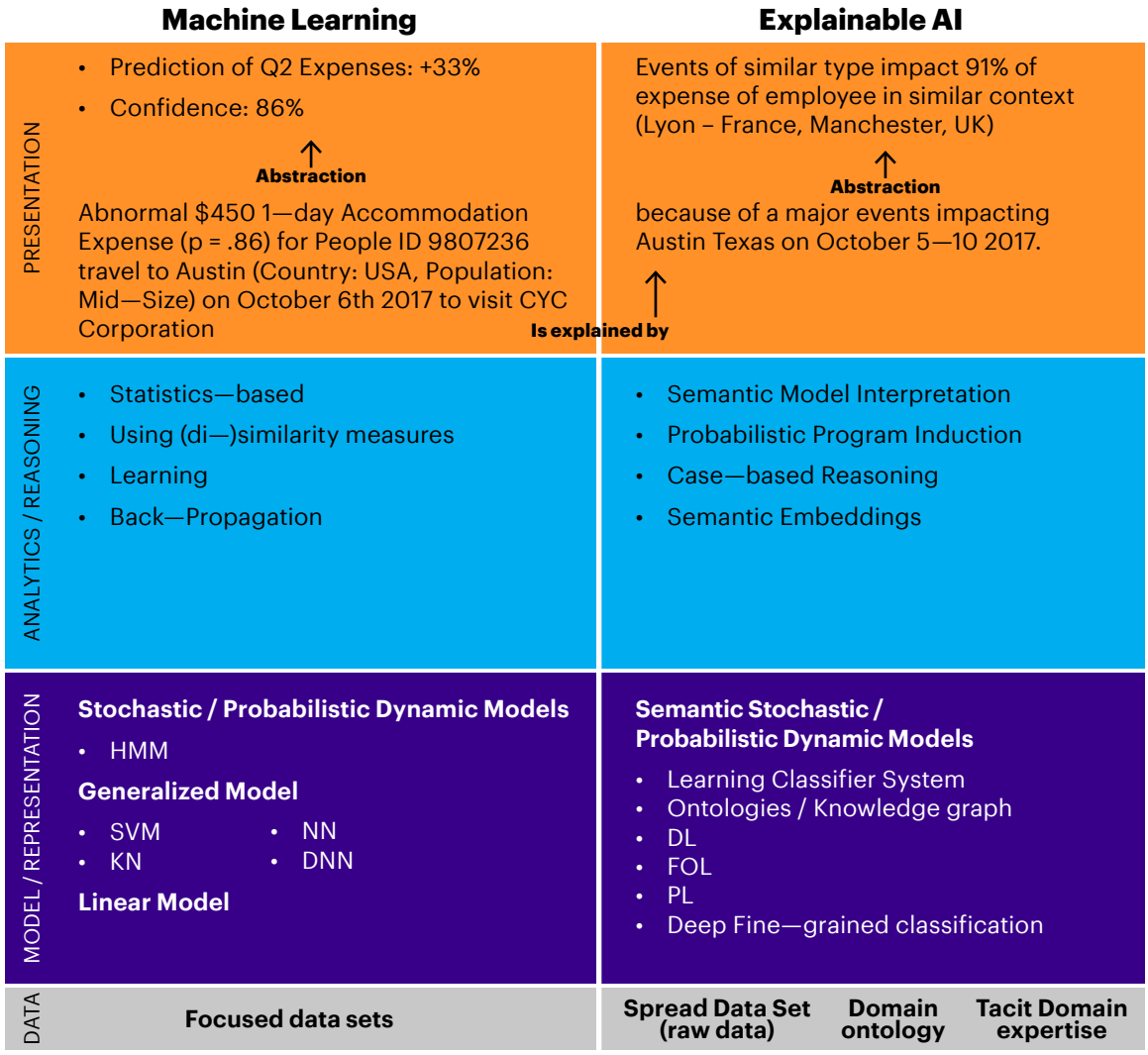
Third, better human—machine synergy. Machines and humans work differently in how they sense, understand and learn. Machines are better at recognizing low-level patterns in huge amounts of data, while people excel at connecting the dots among high-level patterns. To make better decisions, we need both working together.

4 HOW WILL EXPLANATIONS BE MANIFESTED?

As we've made clear, AI will need the ability to explain. But how will it do this?

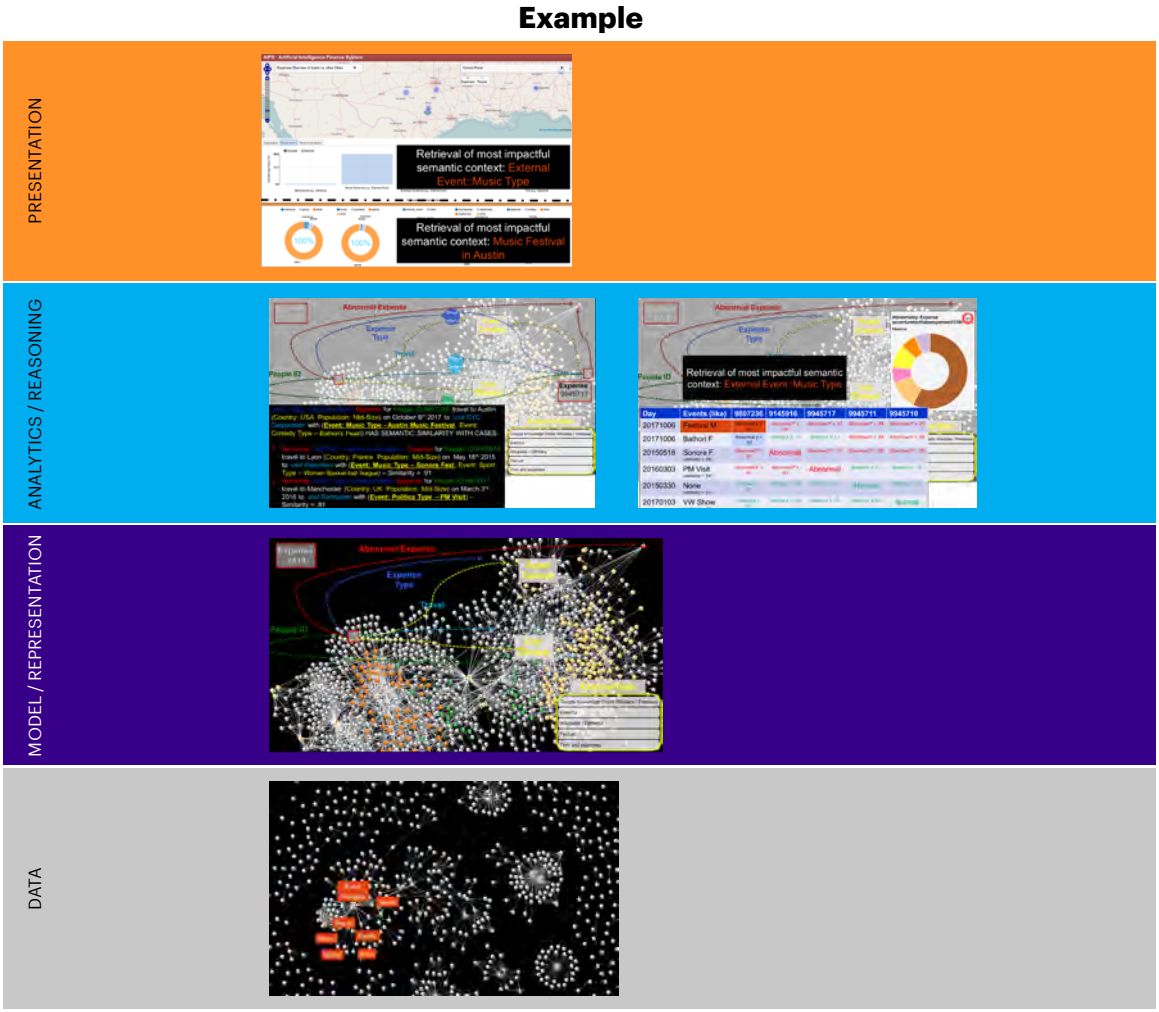
As Figure 3 shows, there are three ways of manifesting and conveying the reasoning behind AI decisions made by machines: first, using data from the machine learning model; second, using the model itself; or third, a hybrid approach combining both data and model.

FIGURE 3. Ways to manifest and convey the reasoning behind AI decisions



Source: Accenture

FIGURE 4. Ways to manifest and convey the reasoning behind AI decisions



A data-level explanation provides evidence of the modeling by using comparisons with other examples to justify the decisions around a particular classification, clustering or targeted prediction. In the case of a mortgage application, an explanation might look like: "The mortgage is not approved because the applicant case is similar to 82% of rejected cases"—thus showing how similar or dissimilar different instances or examples are.

A model-level explanation focuses more on the algorithmic basis of the Machine Learning approach. With this approach, explanations mimic the learning model by abstracting it through rules or combining it with semantics. Compared to the data-level and hybrid-level approaches, the model-level approach abstracts most from the data. In the example of a mortgage application, an explanation might look like: "The mortgage is not approved, because any applicant who has held less than 5,000 euros of savings for the past 20 months is rejected"—thus showing the logic of the model. The logic can be made more understandable by adding a layer of domain knowledge on top.

A hybrid-level explanation works at a higher level of abstraction, by refactoring data at a metadata level—meaning it's a particularly useful method if data is big and very packed. Instead of providing data as evidence, this approach offers metadata and feature—level explanations. In the example of a mortgage application, an explanation would look like: "The mortgage is not approved because both the amount and duration of savings are the most important factors"—explaining which factors have the greatest influence on the decision.

A positive side-effect of generating explanations for a machine learning task is that they can help to uncover biases in data, as the data itself is driving the models and any underlying built-in decisions.

Measuring explanations

Whichever method of explanation is used, eight measures can be applied to assess its value and effectiveness. These measures capture the elements that people need in an explanation. They are:

Comprehensibility — How much effort is needed for a human to interpret it?

Succinctness — How concise is it?

Actionability — How actionable is the explanation? What can we do with it?

Reusability — Could it be interpreted/reused by another AI system?

Accuracy — How accurate is the explanation?

Completeness — Does the "explanation" explain the decision completely, or only partially?

While explainable AI will use and expose techniques that address these questions, we—as humans—should still expect a trade-off between value and effectiveness. For example, it may not be possible to obtain explanations that are both highly succinct and actionable, but instead they may be in a more hybrid form. So an explanation might be complete, non-succinct, but actionable and reusable under some conditions. It will all depend on the availability of data and context, together with the algorithms that are employed and modified.

TWO USE CASES FOR EXPLAINABLE AI: DETECTING ABNORMAL TRAVEL EXPENSES AND ASSESSING DRIVING STYLE

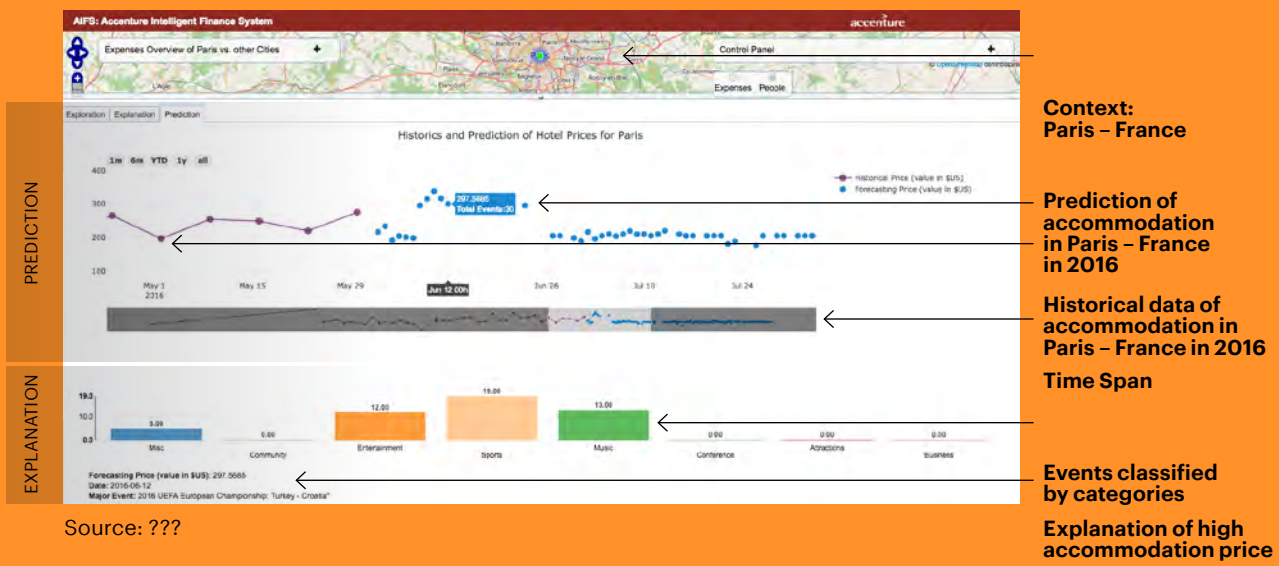
USE CASE 1: ABNORMAL TRAVEL EXPENSES

Travel expenses represent up to 7% of many organizations' overall budget. Most existing systems for reporting travel expense apply pre-defined views such as time period, service or employee group. While these systems aim to detect abnormal expenses systematically, they usually fail to explain why the claims singled out are judged to be abnormal. Indeed, the auditors reviewing the outputs from these systems only see the abnormal values, and not the broader semantic context around them. This means deriving actionable insights from abnormal travel expense claims is time-consuming and often impossible. Examples of these actionable insights might include a recommendation to avoid travelling to a particular city for an internal meeting when a big public event is taking place.

To address this lack of visibility into the context of abnormal travel expense claims, Accenture Labs designed and built a travel expenses system incorporating Explainable AI (see Figure 4). By combining knowledge graph and machine learning technologies, the system delivers insight to explain any abnormal claims in real-time. To do this, machine reasoning is applied to the enterprise knowledge graph to (1) compare contextualized expenses and (2) derive an explanation from the context, based on a mix of enterprise, external and industry data. The Explainable AI process uses the enterprise knowledge graph to pinpoint the likelihood of the context providing a valid explanation for a given claim. A re-usable explanation produced by this process might be to avoid visiting medium-size clients when a large music event is being held, while a succinct explanation might emphasize the impact of large events on accommodation costs. In contrast, humans on their own would not be able to reach such a level of succinctness, given the complex nature of the underlying data and patterns.

These capabilities mean the system makes it easier for business owners to manage expenses, optimize spend and establish new policies, and also provides auditors with a more accurate break-down of the causes for abnormal expenses, reducing the need to go back to employees with queries. The outputs from the system have been shown to be easily interpretable. In 2015 it was trialed successfully with over 190,000 Accenture employees, demonstrating its scalability and accuracy in explaining abnormal expenses.

FIGURE 5. Our system for semantics—aware employee expenses analytics and reasoning



Source: ???

USE CASE 2: PROJECT RISK MANAGEMENT

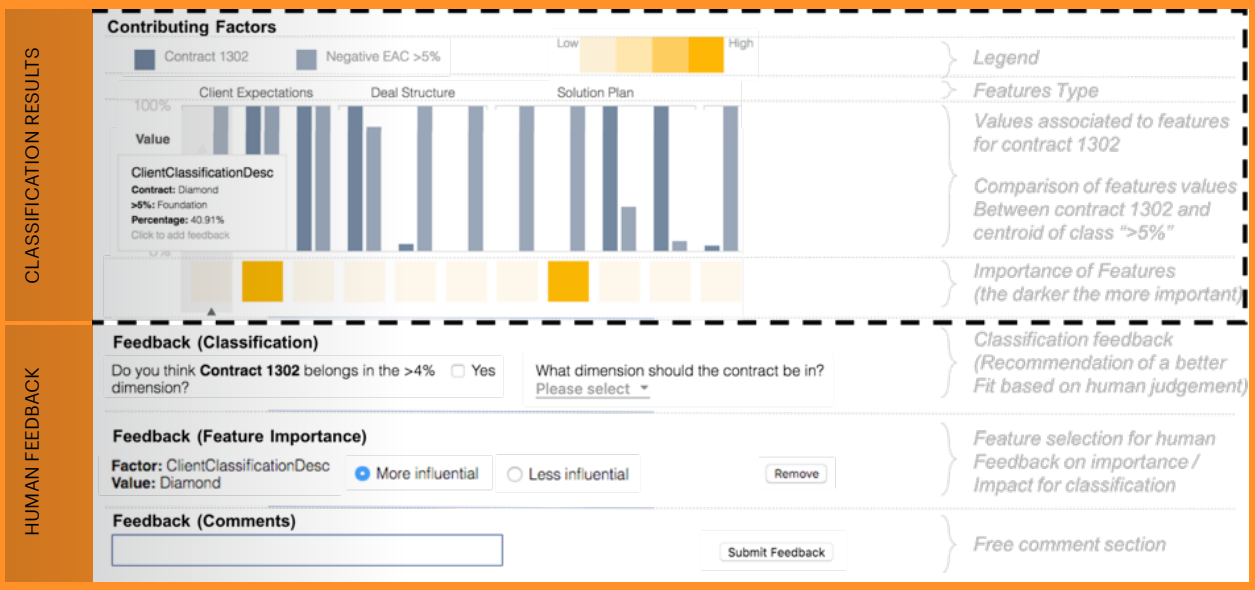
Most large companies are managing hundreds if not thousands of projects on a yearly basis. These projects might be with vendors, clients, partners or any combination, and they usually drive both cost and revenue for the companies involved. In the context of contractual projects with clients, the large company's expectations are often out of line with the original estimates because of the complexity and risks inherent in the critical contracts. These issues are largely due to two factors: first, the high volume of projects that need to be assessed and controlled; and second, the generalized and non-systematic process currently in place for assessing them.

All of this means that decision-makers need systems that not only predict the risk tier of each contract or project, but also give them an actionable explanation of these predictions. These capabilities are especially crucial for large companies, as wrong decisions could have a huge negative impact on revenues, and when this happens the executives involved are accountable for their actions.

To overcome these challenges, Accenture Labs has developed a five-stage process to explain the risk tier of projects and contracts:

- **The first step** involves defining a carefully-crafted internal vocabulary to capture the meaning of contracts, covering factors such as which organizational unit, client profile and contract category they relate to.
- **Second,** the contract and project descriptions are augmented with external information related to brand reputation, awareness and domain expertise. These are all used to enrich the descriptions of projects and capture the broader semantic context, which exposes deeper information for the machine to understand, relate and link patterns, and make predictions.
- **The third step** is to represent all the information in a highly structured and connected representation. An enterprise knowledge graph, built and used for this purpose, enables much better interpretability than a classic data structure.
- **The fourth step** involves applying an Accenture machine learning technique, consuming the enterprise knowledge graph to derive predictions around the contracts and projects, as well as extracting an explanation of the risk tier classification decisions.
- **Finally,** we use a presentation framework to communicate the explanations to end-users.

The schematic below shows the user interface that enables the expert to understand the AI explanations. The explanations are comparable, as the risk tier reasoning needs to be compared across classes and peers (such as contract 1302 and the negative class shown in the schematic). The explanations also need to be accurate and complete, as the decisions will impact the company's revenue and growth. And the explanations are also actionable: first, users can provide feedback on the relevance of the explanations (see the human feedback in the schematic); and second, the risk tier explanations can trigger different actions, depending on which categories they fall into. In contrast, humans on their own would not be able to clearly identify and compare the relevant elements across projects because of the wide diversity project descriptions. They would also be unable to gain a complete view of the risk classification. And actionability would be even more complex, given the large amount of variables that would need to be analyzed and projected to make predictions.



5 A TECHNOLOGY REVOLUTION WITH PEOPLE AT ITS HEART

Explanation is an omnipresent factor in human reasoning: it guides our actions, influences our interactions with others, and drives our efforts to expand our knowledge. When we say effective communication is required for productive collaboration, what we mean is that the collaborators must be able to explain their actions effectively. No one has ever worked effectively with someone when they weren't able to understand their work or get an explanation. So if the future is about humans and machines working productively together, effective explanations will be at the very heart of this collaboration.

Going forward, AI promises to help us identify dangerous industrial sites, warn us of impending machine failures, recommend medical treatments, and take countless other decisions. But the promise of these systems won't be realized unless we can understand, trust and act on the recommendations they make. To make this possible, high-quality explanations will be essential.

REFERENCES

1 Barry O'Sullivan, Alexandre Papadopoulos, Boi Faltings, Pearl Pu: Representative Explanations for Over-Constrained Problems. AAAI 2007: 323-328

2 Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD 2016: 1135-1144

3 Tao Lei, Regina Barzilay, Tommi S. Jaakkola: Rationalizing Neural Predictions. EMNLP 2016: 107-117

4 Jiwei Li, Will Monroe, Dan Jurafsky: Understanding Neural Networks through Representation Erasure. CoRR abs/1612.08220 (2016)

5 Russell, S., Norvig, P., & Intelligence, A. (1995). A modern approach. Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs, 25, 27.

6 Ferrucci, D. A. (2012). Introduction to "this is watson". IBM Journal of Research and Development, 56(3.4), 1-1.

7 SILVER, David, HUANG, Aja, MADDISON, Chris J., et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, vol. 529, no 7587, p. 484-489.

8 Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.

9 Brown, N., & Sandholm, T. Safe and Nested Endgame Solving for Imperfect-Information Games.

10 Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015, pages 1721-1730, 2015.

11 Bryce Goodman and Seth Flaxman. EU regulations on algorithmic decision-making and a "right to explanation". CoRR, abs/1606.08813, 2016.

ABOUT ACCENTURE LABS

Accenture Labs incubates and prototypes new concepts through applied R&D projects that are expected to have a significant strategic impact on clients' businesses. Our dedicated team of technologists and researchers work with leaders across the company to invest in, incubate and deliver breakthrough ideas and solutions that help our clients create new sources of business advantage. Accenture Labs is located in seven key research hubs around the world and collaborates extensively with Accenture's network of nearly 400 innovation centers, studios and centers of excellence globally to deliver cutting-edge research, insights and solutions to clients where they operate and live. For more information, please visit **www.accenture.com/labs**

ABOUT ACCENTURE RESEARCH

Accenture Research identifies and anticipates game-changing business, market and technology trends through provocative thought leadership. Our 250 researchers partner with world-class organizations such as MIT and Singularity to discover innovative solutions for our clients.

ABOUT ACCENTURE

Accenture is a leading global professional services company, providing a broad range of services and solutions in strategy, consulting, digital, technology and operations. Combining unmatched experience and specialized skills across more than 40 industries and all business functions – underpinned by the world's largest delivery network – Accenture works at the intersection of business and technology to help clients improve their performance and create sustainable value for their stakeholders. With 449,000 people serving clients in more than 120 countries, Accenture drives innovation to improve the way the world works and lives. Visit us at **www.accenture.com.**