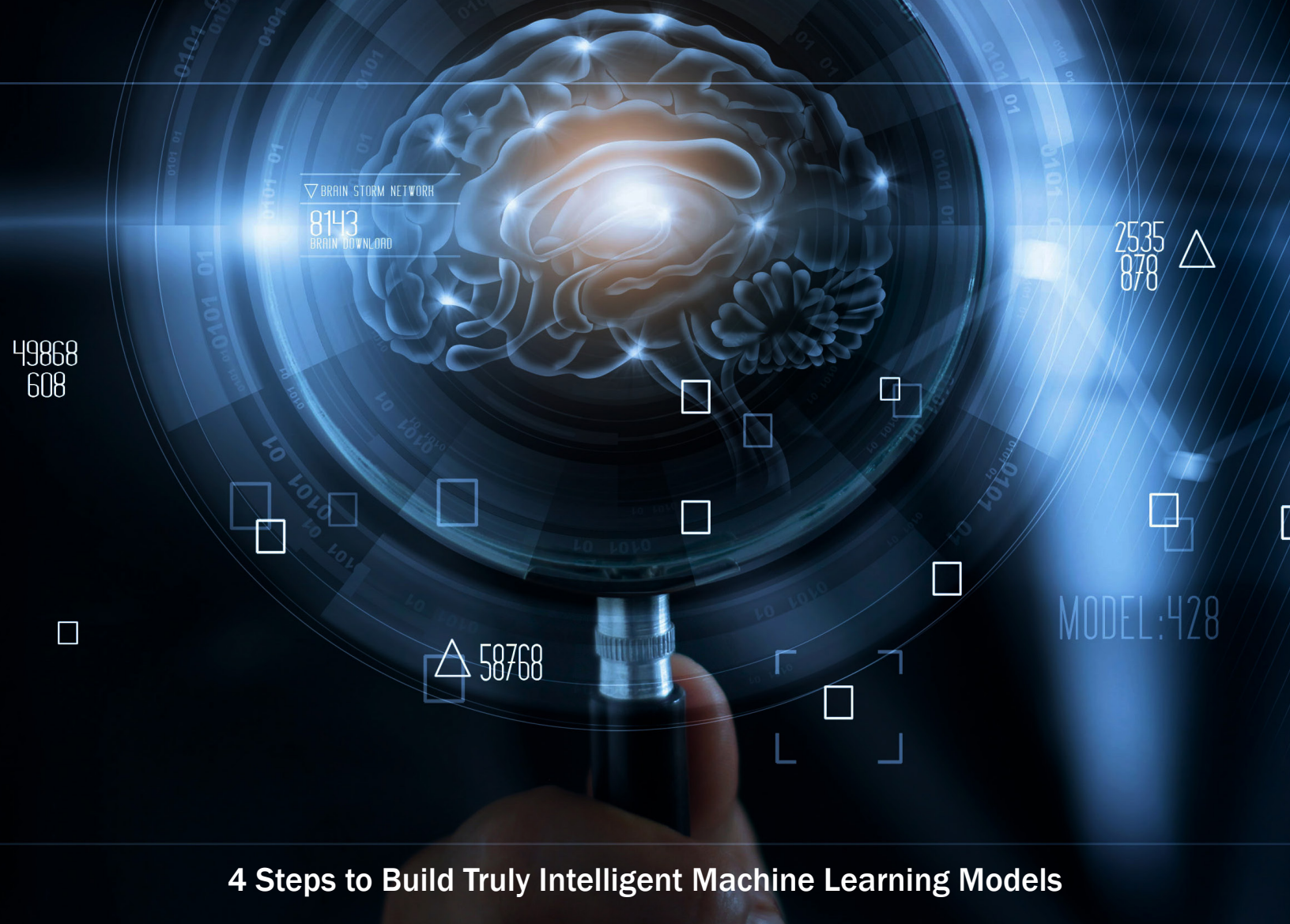


WITHOUT DATA, ARTIFICIAL INTELLIGENCE IS PRETTY DUMB



4 Steps to Build Truly Intelligent Machine Learning Models

TABLE OF CONTENTS

RISE OF THE MACHINES?.....3

AI AND THE TERMINATOR CONUNDRUM.....4

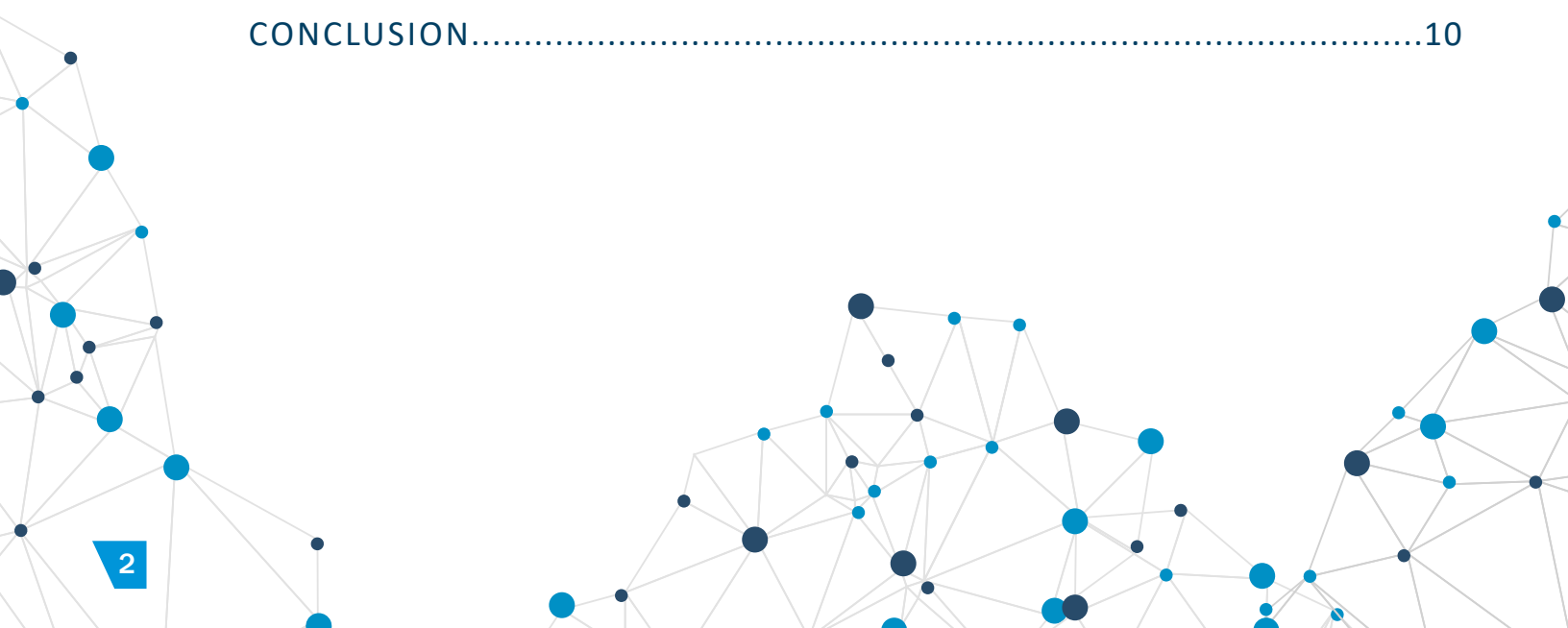
DATA IS CRUCIAL TO THE SUCCESS OF AI.....5

4 REQUIREMENTS FOR BUILDING AI APPLICATIONS.....6-7

DATA IS THE HEART OF ARTIFICIAL INTELLIGENCE.....8

QUALITY ALWAYS TRUMPS QUANTITY.....9

CONCLUSION.....10



The Rise of the Machines?

Not so fast...

Growing up, I've always been fascinated by science-fiction films. From *Blade Runner* to *The Terminator*, these movies hauntingly illustrated futuristic worlds shaped by technological innovations that gave rise to advanced machine learning techniques and depicted astonishing examples of artificial intelligence (AI). But instead of doomsday scenarios with humanity cowering at the feet of our robot overlords, AI has emerged as one of the most significant forces behind the digital transformation of business. In fact, many believe AI has the potential to not only impact the corporate world, but lead to groundbreaking applications which will have profound effects on every aspect of our daily lives.

From health to transportation, AI is enabling people to redefine how information is being collected, integrated, and analyzed; leading to more informed insights and delivering better outcomes.

These breathtaking strides in technology are being driven by advancements in machine learning; specifically, what is referred to as deep learning.

Deep learning is part of the broader field of machine learning that is concerned with giving computers the ability to learn without being programmed, and it's led to beneficial developments in many areas, including:

- **Language Understanding** – Chatbots are able to automate customer calls and even make appointments by understanding context and spoken language.
- **Image/Video Understanding** – Ability for machines to identify images or moving videos to make deterministic calls on potential threats.
- **Audio Understanding** – Enabling machines to comprehend human words to drive improved understanding of language.

Despite the incredible promise of AI, super smart folks like Stephen Hawking and Elon Musk still warn of the coming AI apocalypse. In fact, Elon Musk's new company, Neuralink, aims to stop a *Terminator*-style attack by fusing man and AI through brain links.

But should we really be worried? Not yet, at least.



AI and The Terminator Conundrum

This makes me think back to one of my favorite sci-fi flicks, *The Terminator*. It's famous for portraying a dystopian society with artificially intelligent robots hell-bent on the destruction of the human species, not to mention some catchy one-liners from future California governor, Arnold Schwarzenegger. In the movie, the T-800 Model 101 Terminator, a highly-advanced robot with living tissue over a metal endoskeleton, is programmed to find and kill our hero, Sarah Connor.

But the machine does not know which Sarah Connor to target. The only data it has is her name and the city she lives in. Not knowing exactly who the main target is, The Terminator must scroll through the phone book (remember those?), dispatching all the Sarah Connors on the list. Being that this is a 90-minute movie, The Terminator intercepts the intended Connor rather quickly and spends the rest of the movie blowing things up.

So, as advanced and intelligent as *The Terminator* is, it essentially must guess which Sarah Connor to target because it lacks important basic, foundational training data – the information used to train a machine learning model. It's actually not until the many sequels that the machine can identify its main target because it has already been fed the proper training data by learning and adjusting through its first failed and corrected attempts. (And there's that whole issue of time travel that just gives me a headache.)

So, AI, whether it's a homicidal cyborg or customer chatbot, needs the right data to make intelligent decisions. The importance of human interaction and analysis cannot be overstated. Our ability to source, determine and evaluate the data we use and supply to these AI systems will ultimately help make them smarter over time. Until then, AI needs humans-in-the-loop.



Data is Crucial to The Success of AI

While the AI story is all the rage with the media, the data narrative is not as prominently discussed. Sure, data may not be as sexy as the automated systems that can learn and process information quicker than a human, but it is equally as important. And don't get me wrong, everyone knows that AI requires vast amounts of data to continually learn and identify patterns that humans can't.

After all, it's the ability to process this information and make instant decisions that has led to AI being such a game changer for industries that rely on massive volumes of data. But the real story is not about the algorithms powering the AI revolution, instead it's about the quality of data powering these systems. What enterprises really need as they develop their AI strategy is to integrate, clean, link, and supplement their data so they have an accurate foundation on which to build and train their machine learning algorithms. For many organizations, this makes machine learning difficult if not impossible. According to Gartner, 90% of initiatives will fail because of either lack of data or lack of training data.

So, what can enterprises do?

First and foremost, organizations need to think about data differently than how they do today. It's no longer helpful to compile as much data as possible and hope some of it turns to gold. In other words, never lead with the data; instead lead with a question that the data can answer.

Data must be treated as a building block for information and analytics. Therefore, it must be able to answer a question or set of questions. It takes a thorough understanding of both the data itself and the questions you should be asking of the data to get truly intelligent answers. This is where humans come in. Companies that are seeing success with their AI and machine learning initiatives understand how to leverage the right resources to ensure data is having the biggest impact.



90% OF AI PROJECTS WILL FAIL

4 Requirements for Building AI Applications

Over the past decade, Innodata has spent significant time helping our customers in this area. As a result, we've had the opportunity to uncover some of the key requirements needed in building an effective AI application.



1 Raw Data

Having access to the right raw data set has proven to be critical factor in piloting an AI project. Raw data is information that has typically not been processed or analyzed and is routinely considered inoperable. But deeper analysis can yield opportunities to turn raw data into useful insight. For example, one of our clients was looking to understand key challenges associated with their customers self-serve system and looking to improve the customer experience. After a thorough introspection of all the shared data, we honed in customer call center transcripts as a way to understand the trends and train their AI models.



2 Ontologies

Ontologies play a critical role in machine learning. According to the Wikipedia definition, ontologies are “formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse.” In other words, ontologies give meaning to things. Think of this as teaching your AI to communicate using a common language. It is critical to identify the problem statement and understand how AI can interpret data to semantically solve a certain use case. The need for out-of-box ontologies or availability of client ontologies that can be used as the basis to form the data labeling is critical.

4 Requirements for Building AI Applications



3 Annotation

Annotation (also known as data labeling) is quite critical to ensuring your AI and machine learning projects can scale. It provides that initial setup for training a machine learning model with what it needs to understand and how to discriminate against various inputs to come up with accurate outputs. There are many different types of data annotation, depending on what kind of form the data is in. It can range from image and video annotation, text categorization, semantic annotation, and content categorization. Humans are needed to identify and annotate specific data so machines can learn to identify and classify information. Without these labels, the machine learning algorithm will have a difficult time computing the necessary attributes. How data is annotated and labeled brings us to our next and most crucial requirement: subject matter expertise.



4 Subject Matter Expertise

Think back to our friend The Terminator. He (or shall I say, it) lacked the basic information needed to complete its mission. What was really needed was some human guidance to point him in the right direction. While there's no argument for humans playing a role in shaping artificial intelligence, when it comes to what's needed to make the machine really intelligent is intelligence itself.



Our clients have learned how important it is to have subject matter experts (SME) that understand their specific industry and complex needs. This goes back to the need for annotated data. If there are even slight errors in the data or in the training sets used to create predictive models, the consequences can be potentially catastrophic. That's why the need for specific domain expertise is so crucial, and why human knowledge still plays a pivotal role in artificial intelligence. For example, being able to interpret complex legal obligations and agreements from ISDA contracts require legal specialists that can identify and label the most appropriate information. The same goes for other fields like science and medicine where deep understanding and fluency of the content cannot be taken for granted.

Lastly, the support of data scientists that can partner with the SMEs to build an entity relationship map or knowledge graph has proven to be a winning formula for training models.



Accuracy

While this characteristic seems obvious, it cannot be overstated. Many of us will have different definitions and expectations of what accuracy is, but it's essentially correct and consistent information that can be used to guide efficient decisions. Accurate data should be correctly defined in a consistent manner in accordance with the expected data standards of a particular business model. But accuracy doesn't happen by itself, it takes human intervention to define these essential data attributes. In many cases the concept of accuracy is very nuanced, so it must be taken in the context of the particular attribute you're using it for. If your data is even marginally incorrect, it can derail your objectives.



Completeness

This data characteristic can be measured by how well the data set captures all data points available for a given instance. A complete data set should not have any gaps in the data from what was expected to be collected, and what was actually collected. For example, if a person's medical record only covers their most recent check-up history, then that data set will misjudge the patient's true health. The data must paint a full picture to provide the right answers to your questions.




Uniqueness

This attribute refers to data that can stand alone and not be found in multiple formats and locations within your database. In other words, there should be no duplicates of the same record. Unfortunately, many companies create the same record over and over without even knowing it. Whether it's a slight modification of the naming convention or inaccurate labeling, the lack of a single source of truth could create challenges with accuracy over time. This is why you'll often hear the term standardization. Having standardized data allows organizations to find meaningful ways to compare data sets. This is necessary for inputting information, but it is even more important in identifying duplication.



Timeliness

Data is constantly in flux. That's why it's imperative to be able to collect and update in a timely fashion. A deep understanding of when the data is no longer useful based on timing needs to be determined. For example, a provision to a financial agreement must be accounted for the moment it is set. If there is significant lag between when the data is collected to when it is used to drive a business decision, it could result in expensive consequences. Data collected too soon or too late could disrupt machine learning outputs.



Quality Always Trumps Quantity

All of these characteristics come together to determine data quality; the basis for making good decisions. As more organizations invest in artificial intelligence and machine learning, data scientists must focus on overall quality, especially that of the metadata. Metadata is what describes the data and the lack of such information is one of the primary causes of bad data. If you're training algorithms minus a solid metadata foundation, they will never become reliable enough to meet your particular needs.

Beyond the data itself, there are severe constraints that can impede analytics and deep learning, including security, privacy, compliance, IP protection, and physical and virtual barriers. These constraints need to be carefully considered. It doesn't help the enterprise if it has collected and cleaned the data but find it's inaccessible for various reasons. Often, steps need to be taken such as scrubbing the data so that no private content remains. Sometimes, agreements need to be made between parties that are sharing data, and sometimes technical work needs to happen to move the data to locations where it can be analyzed.

Organizing, cleaning and structuring data may be the least glamorous part of your company's AI initiative, but it's certainly the most important. Without that solid data quality foundation to build your models, you'll never achieve reliable and valid results. Be sure your data has these key characteristics.

Once you are comfortable with the quality of the data being fed to your machine learning systems and understand the right steps to take, the real fun begins – training your models.

Conclusion

There's no question artificial intelligence has the potential to completely change the world as we know it. But there's no need to worry about the machines taking over anytime soon. As we have seen, humans still play a decisive role in the development and training of machine learning models. And let's not forget the data. No matter how sophisticated these machine learning models become, they will never live up to their full potential until the data is reliable. After all, artificial intelligence is, at its core, artificial. It will do its job based on what it is told, whether that information is accurate or not.

Here are some key things to remember before diving into a machine learning project:

- Data quality trumps data quantity
- Ask questions of your data
- Break down data silos to ensure everyone is on the same page
- Ensure your ontologies are defined
- Subject matter expertise is critical to help machines understand complex scenarios
- Constantly train your models


“


**Without all of this,
AI is pretty dumb**


”

Visit www.innodata.com/dataforai
to learn more about how Innodata
bridges human expertise with
artificial intelligence.



 www.innodata.com

 email: info@innodata.com

 Ph: 201-371-8000