Rise of the Machines: Achieving Data Security and Analytics with AI

Glendon Schmitz, CISSP
CISO, DBHDS

Will Goddin
CIO, Diveplane

Virginia Department of Behavioral Health & Developmental Services

Diveplane

MARVEL vs DC

Source: CBR.com

Slide 2:

One of the most fundamental questions of geeks and nerds… What universe is better? DC or Marvel?

Act I The Journey

The Hero's Journey

Source:shiply.com

Slide 5:
Our CISO Hero has many problems that he is confronted with daily. Our quest is to find solutions in order to maintain safe and secure systems that process our organization's most prized data. When I first started as a CISO in 2020 at the beginning of the pandemic, I did what most CISOs do and I surveyed my landscape, it's impossible to secure what you don't know. I quickly found that my organization used live production data in the lower Dev and test Environments. This is extremely risky as these lower environments are not secured to the same security standards as production environments. This is because its time consuming and very costly to maintain. I thought, "Oh this will be an easy fix, we will generate test data and "SUCCESS" however, the business gets a say in this, and the business owners demanded the use of productions data in these lower environments. This way they know what goes in and what to expect to come out. I needed to find a solution that would provide production quality test data, and at volume, that met the needs of the business.
The next major trial I had to face was the problems with data sharing! I work for a healthcare organization that specializes in mental health treatments and the demand for our data by other treatment facilities, researchers, and academia is extremely high. Because we deal with personal health information (PHI), Personal Identifiable Information (PII), we are bound by strict privacy protection regulations.
When we need to share our production data, we usually have to go through lengthy data sharing agreement negotiations that can stall and delay the business from accomplishing their objectives and Security is then no longer seen as a value-added business enabler but rather a hindrance and at best a nuisance.
So how do can we as security professionals support the business, reduce risk while at the same time does not cause delays in the business objectives. Oh, all the while working with constrained budgets and resources?

The Quest

Source:unsplash.com

- Production Data
  - Dev and Test Environments
  - Production Quality Test Data at Volume
- Data Sharing
  - Privacy protection regulations
  - Lengthy Agreement Negotiations

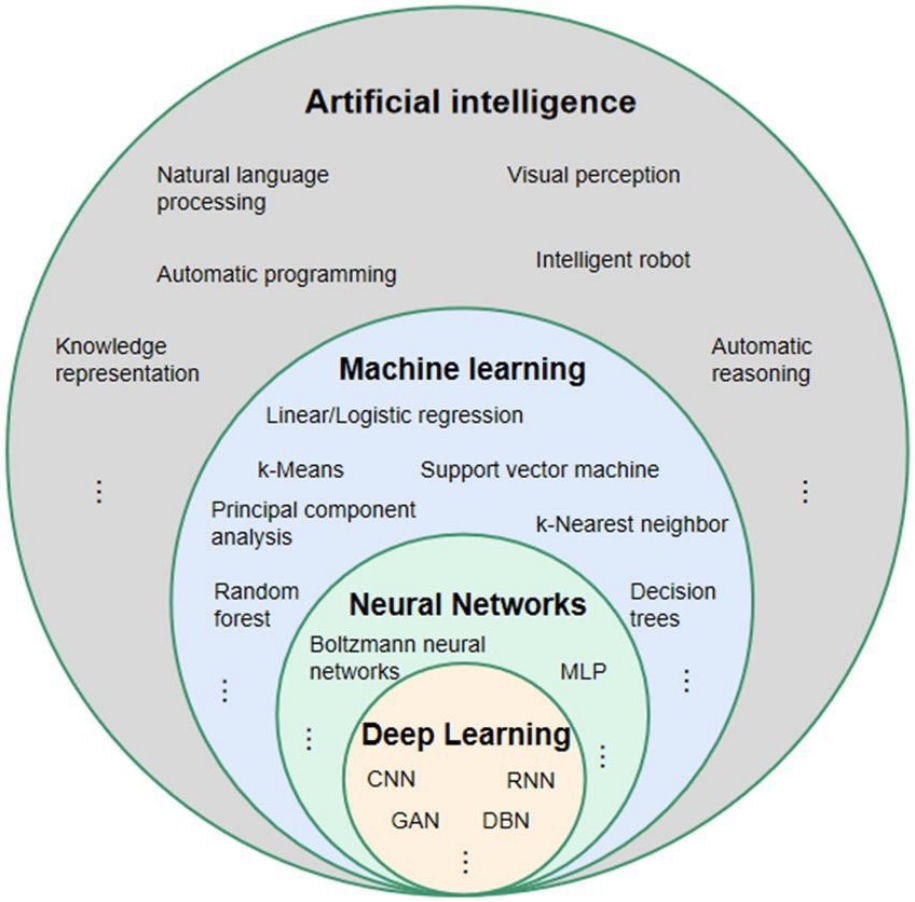
Slide 5:
Our CISO Hero has many problems that he is confronted with daily. Our quest is to find solutions in order to maintain safe and secure systems that process our organization's most prized data. When I first started as a CISO in 2020 at the beginning of the pandemic, I did what most CISOs do and I surveyed my landscape, it's impossible to secure what you don't know. I quickly found that my organization used live production data in the lower Dev and test Environments. This is extremely risky as these lower environments are not secured to the same security standards as production environments. This is because its time consuming and very costly to maintain. I thought, "Oh this will be an easy fix, we will generate test data and "SUCCESS" however, the business gets a say in this, and the business owners demanded the use of productions data in these lower environments. This way they know what goes in and what to expect to come out. I needed to find a solution that would provide production quality test data, and at volume, that met the needs of the business.
The next major trial I had to face was the problems with data sharing! I work for a healthcare organization that specializes in mental health treatments and the demand for our data by other treatment facilities, researchers, and academia is extremely high. Because we deal with personal health information (PHI), Personal Identifiable Information (PII), we are bound by strict privacy protection regulations.
When we need to share our production data, we usually have to go through lengthy data sharing agreement negotiations that can stall and delay the business from accomplishing their objectives and Security is then no longer seen as a value-added business enabler but rather a hindrance and at best a nuisance.
So how do can we as security professionals support the business, reduce risk while at the same time does not cause delays in the business objectives. Oh, all the while working with constrained budgets and resources?

Preparation

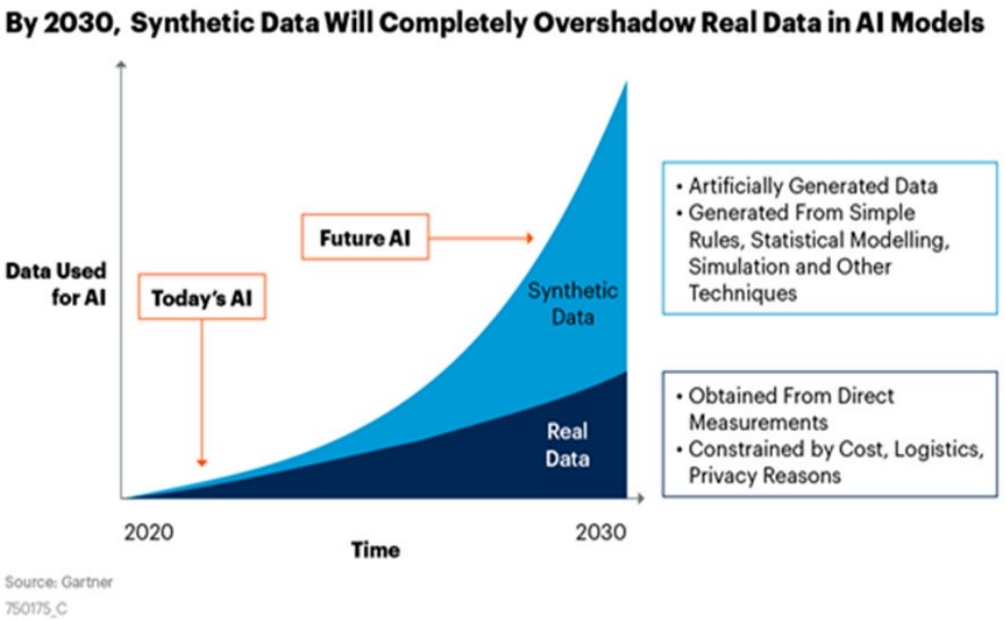
Source:
sendgrid.com

Slide 6:
Now, before our CISO Hero can continue his journey, he must get equipped with the proper tools that we help him on his journey. And like all hero journey stories our hero enlists the help of a very wise wizard! Our wizard in this story today is Will. Will can you help me achieve my quest?

Act II
Machine Learning, Artificial Intelligence & Synthetic Data

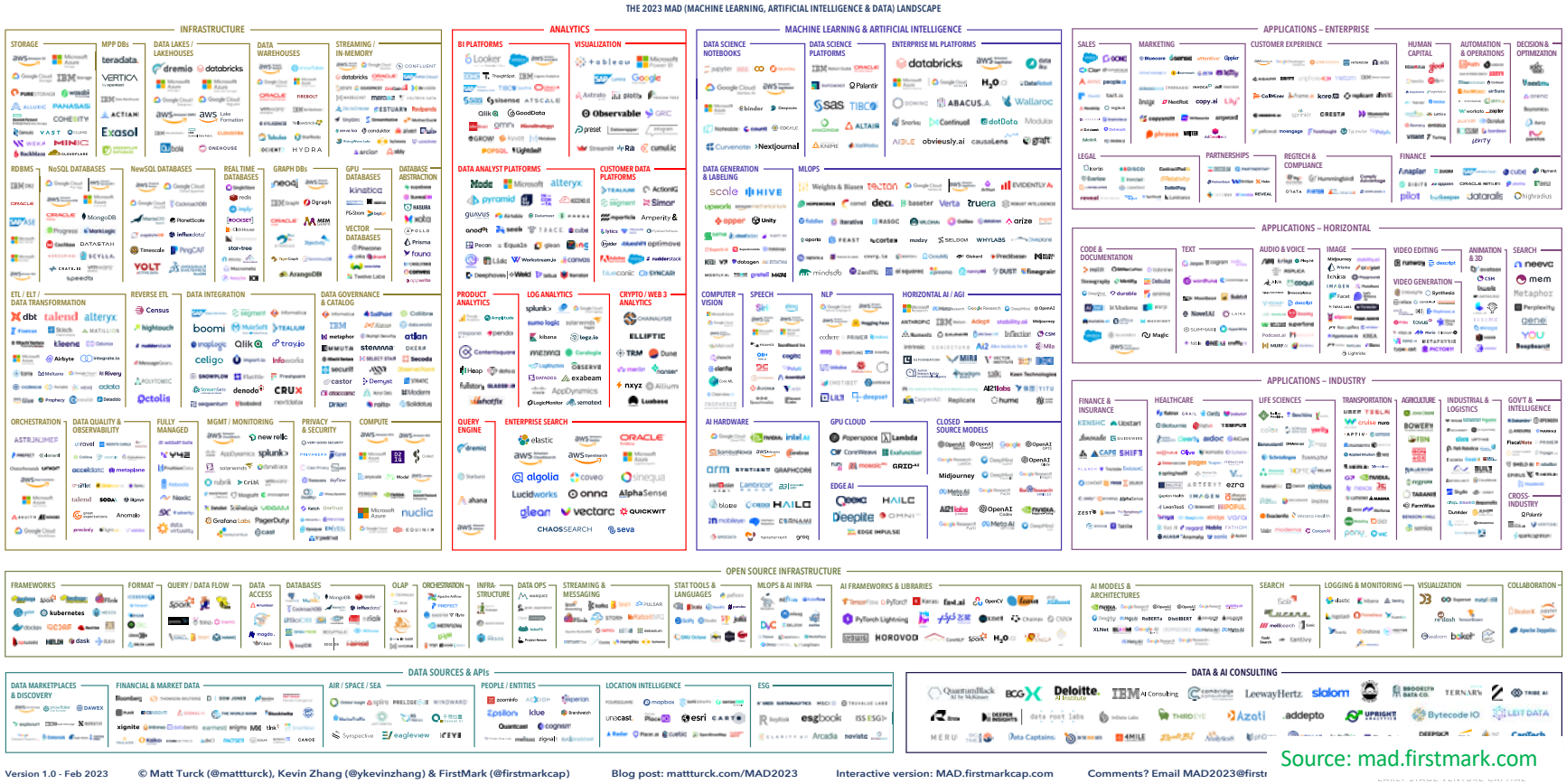


Source: Twitter / @Jousefm2



Source: Gartner Research 2022

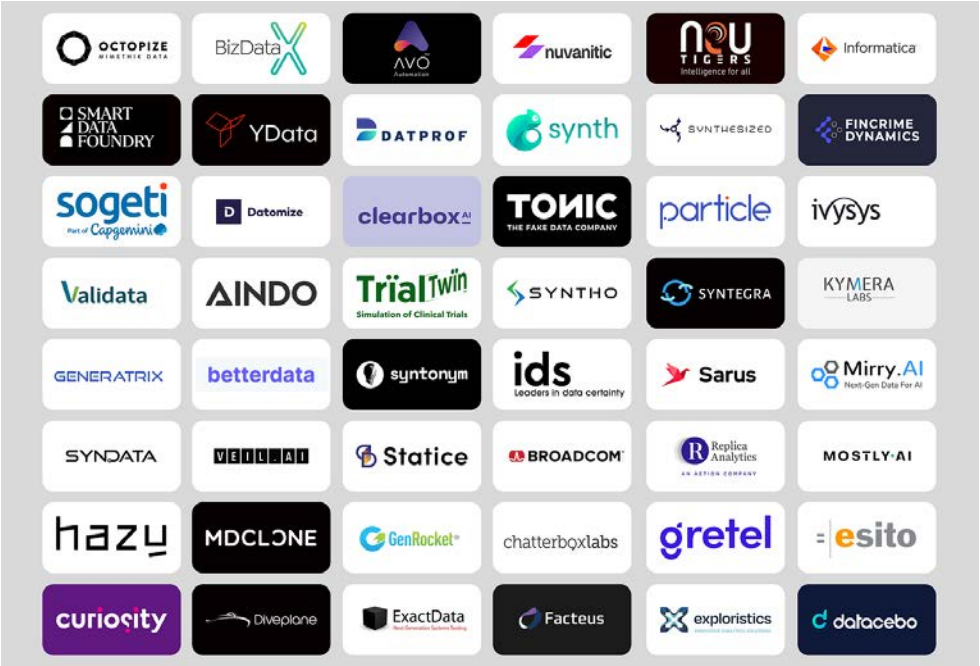
Slide 8:
Artificial Intelligence is a technology category which defines solutions ability to sense, reason, learn and apply prior knowledge. Technologies like Chat-GPT, Siri and Alexa are all Natural Language Processors.
Machine Learning is defined by the ability to learn how to classify and predict based on data and not specific programming. Algorithms like Random Forest, Linear & Logistic Regression and K Nearest Neighbors are all examples.
Neural Networks is a technology modeled after the human brain. These programs are fed vast amounts of data and the network "learns" from it and able to predict answers based on context or prompts.
Synthetic data is data generated by simple rules, statistic modeling or advanced AI programs. It is meant to be a stand in for the original data but better than the original data because it meets the business rules and constraints.
Remember Glenn wanted to eliminate RISK of data leakage, but still wanted data that made sense for the Dev and Test environments for testing.
Now that we have a better understanding of the terms, let's find some tools that can help us achieve our goals!


THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE
INFRASTRUCTURE
ANALYTICS
MACHINE LEARNING & ARTIFICIAL INTELLIGENCE
APPLICATIONS – ENTERPRISE
APPLICATIONS – HORIZONTAL
APPLICATIONS – INDUSTRY
OPEN SOURCE INFRASTRUCTURE
DATA SOURCES & APIs
DATA & AI CONSULTING
Source: mad.firstmark.com
Version 1.0 - Feb 2023
© Matt Turck (@mattturck), Kevin Zhang (@ykevinzhang) & FirstMark (@firstmarkcap)
Blog post: mattturck.com/MAD2023
Interactive version: MAD.firstmarkcap.com
Comments? Email MAD2023@firstm

Slide 9:
This is the Machine Learning, Artificial Intelligence, and Data tools landscape! You've all heard about these tools, right? Overwhelmed? Anyone MAD? Going MAD? Okay, so where do you start?

What kind of data?

Structured Data Tools



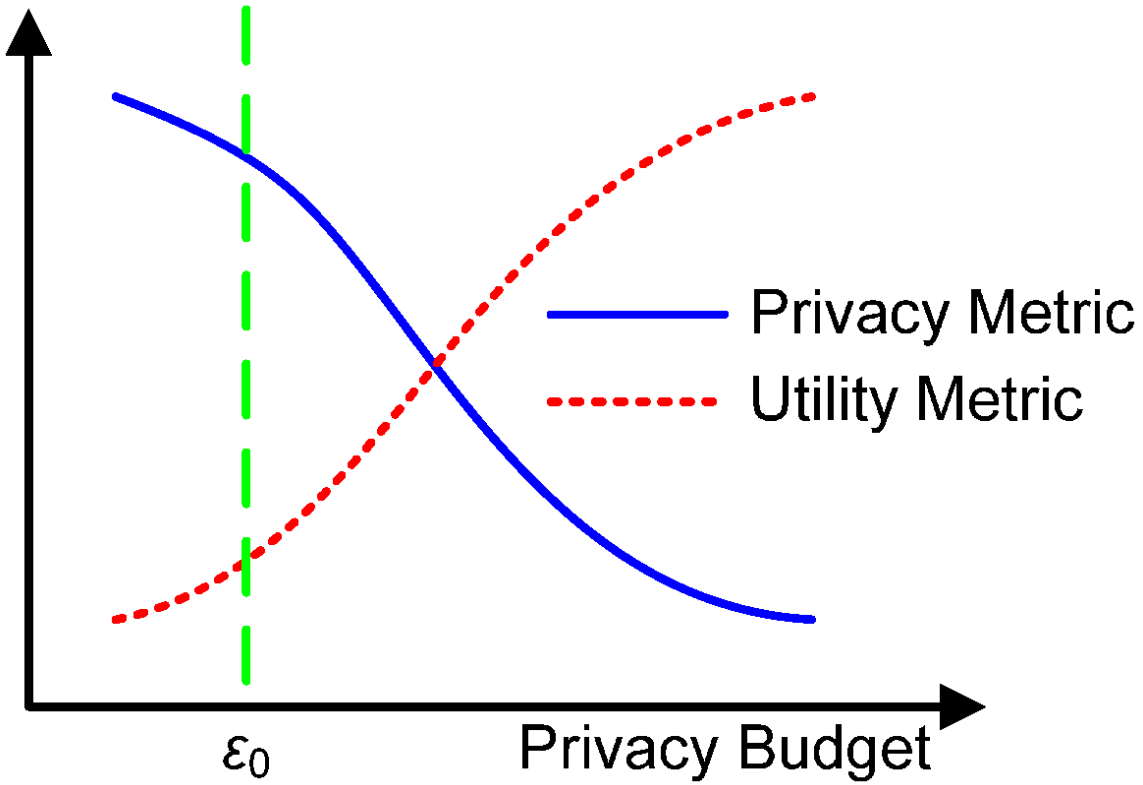
Source: Medium.com

UnStructured Data Tools

Source: Medium.com

Slide 10:
Let’s go back to the Requirements!
What kind of data do you need? Structured, Unstructured or even Semi-Structured data like JSON blobs?
Structured Data is what you find in CSV Files or Database Rows & Columns.
Unstructured Data is images & text.

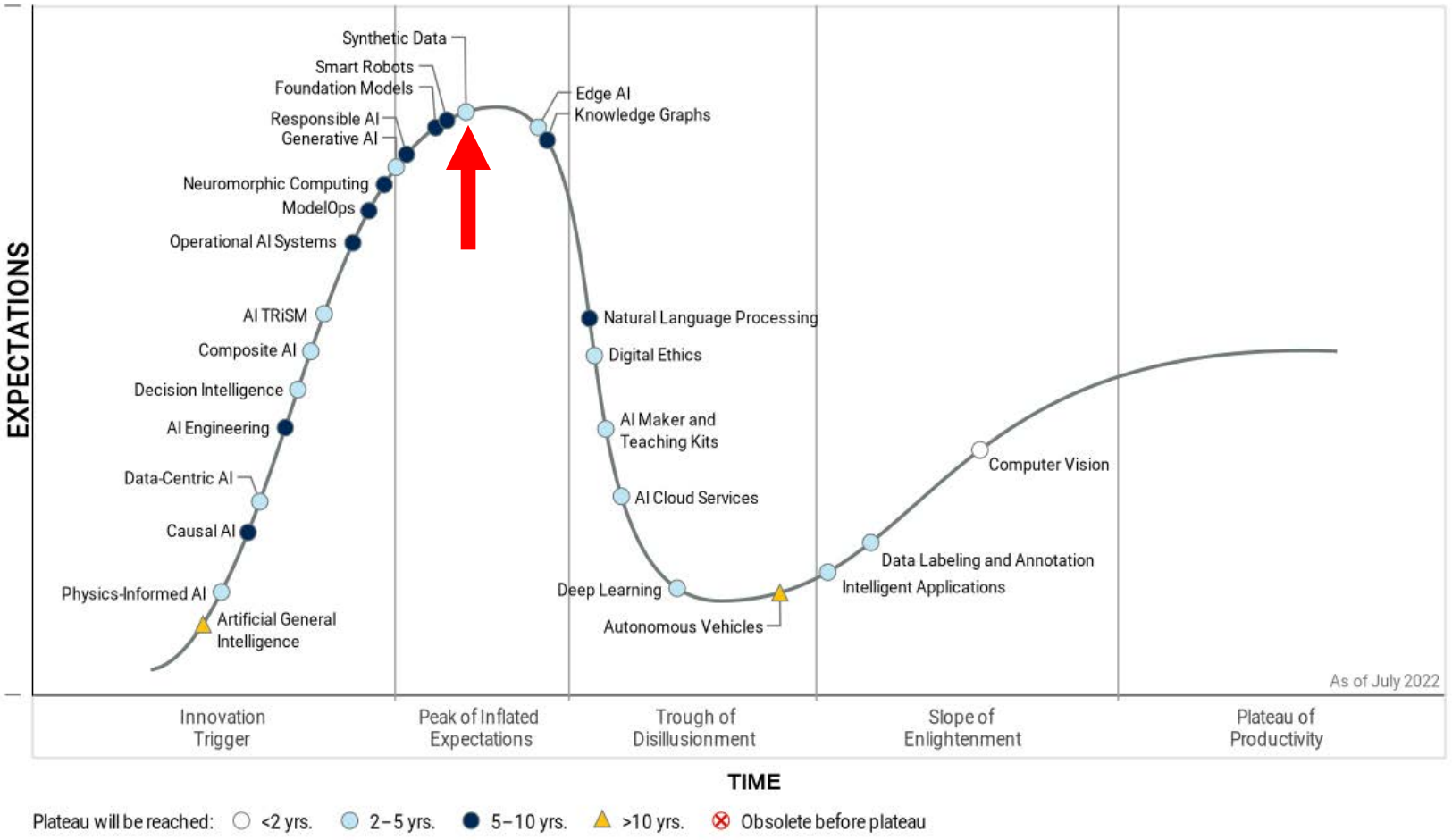
How should you measure your synthetic data?



https://doi.org/10.3390/app8112081

Slide 11:
Does the synthetic data need to look and feel like the original data, or can it just be random?
Can you generate it using simple business rules?
Perhaps you need a tool to help… but make sure you know how to evaluate the tools' performance.
Is it meeting your requirements to protect your patient identities, enables you to comply with appropriate regulations like HIPPA and GDPR.
Can you use the synthetic data to train a model and make predictions or do analytics on it with similar outcomes to the original data?

Hype Cycle for Artificial Intelligence, 2022



Gartner

Source: Gartner

Slide 12:
Synthetic data is at the Peak of Inflated Expectations.
It highlights the capabilities where early publicity produces success stories, but expectations rise above the current reality of what can be achieved.

Act III
The Adventure Begins!

Evaluating the Solutions

Source: helenporterpa.com

- Does it comply with the Law?
- Is the AI Ethical & Explainable
- Multi-Use Cases
- Meet Security Concerns

Slide 14:
Our CISO Hero now has the tools need to have a successful journey and defeat his foes! But before he can set off on his journey, he needs to assess these tools to ensure he equips himself for the task at hand.
When looking for an AI and Synthetic solution the CISO needs to keep in mind the following:

1. Is the AI ethical and explainable?

Although we should be able to trust our solutions we need to verify how and more importantly why the AI answers the way it does. We must ensure the machine's solution is in line with our human ethics and be able to make adjustments when it doesn't.
2. Does the solution meet and solves multiple problems for the organization and possibly multiple business units?
AI and synthetic data solution as costly. The best solutions out there can be applied to multiple business units that will being not only value but innovation to those business units.
3. Finally, the hero needs to evaluate if the solution will answer the security concerns of the organization. Do the solutions meet or exceed the security standards and the security framework that your organization has implemented?

The Abyss

Source: unsplash.com

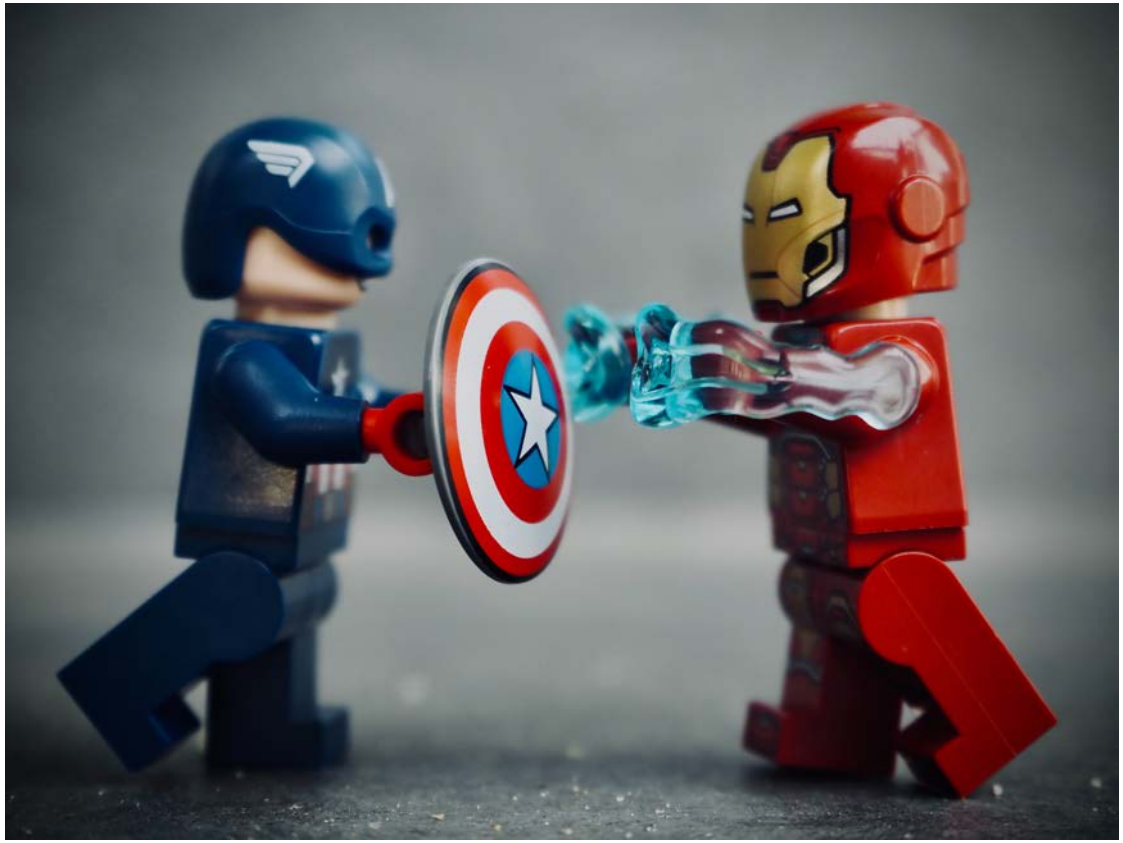
- Approach
- Security and Operational Needs
- Timelines & Objectives
- Return On Investment

Slide 15:
Once our hero has evaluated the solution presented to him by the wizard, he finds himself, like all heroes, confronted with a great challenge to proof himself worthy and over come self doubt, the Abyss.
In this story, the CISO Hero must decent into one of the scariest places for a CIOS... the boardroom! Here we find our hero hard at work convincing the executive leadership that AI and Synthetic data, the... "Mystic Gift" from our wise wizard, is the best solution to help make our quest a success and eliminate the problems that we face.

1. Production data in the lower less secure environment, dev and test.
2. The creation of quality test data at volume.
3. Adherence to complex and costly privacy protection regulations
4. End the need for lengthy delays in data sharing agreements.

It is of up most importance to approach any security solution and in this case AI and Synthetic data, from the perspective of how the solution will benefit the business. How will this enhance business operations, safe money and save the business time.
It is imperative to understand your overall security and operational objectives and tie your objectives to the business. To do this, the security professional needs to align the desired outcomes of the solution with the business and get top executive buy-in on its implementation. Ensure that all stakeholders of your organization understand that the needs of business operations align with the requirements for security. Be clear on what success would look like, not just from a security perspective but from the businesses. Be clear on how implementing an AI solution will enhance both security and operations of the business and ultimately how security will become a business multiplier versus a business expense.
Be realistic with your timelines and objectives. It far easier to convince your board and executive leadership to invest in an AI solution when you present the short- and long-term goals of the security and business are aligned for the best possible outcome of success.
For instance, in the short term, our CISO Hero will implement the use of generation of synthetic data as quality test data in the lower environments to reduce the risk and probability of a data breach if these lower environments become compromised. In the long term, the business can use the AI to conduct predictive analytics on the data. I my case, that means understand the mountains of health care data that my organization maintains the evaluate the probability of success for a specific treatment.
Finally, you have to clearly communicate the Return on Investment of AI to the business. Remember that Risk = Likelihood of Occurrence x Severity. When you factor in the Cost of an incident and the cost to fix it you will be in a far better position to communicate what an ROI would look like for your organization.

The Battle Begins!!



Source: towardsdatascience.com

Slide 16:
Discussions with the business to convince them that the use of AI & Synthetic data is the answer to the problems.
How synthetic data can reduce the risks and costs of a security breach.
Average Ransome Breach costs $4.54M If we have 3 environments that have production data and they are breached.... That's more than $13M in liability. But with the use of AI generated synthetic data we can reduce the risk by 2/3 and potentially save the organization $9M.
If we only use synthetic data for analytics, if the organization is hit with ransomware, we simply wipe the old environment and create a new one and populate it with a new set of synthetic data sets. Because synthetic data is a true representation of the production data, there is no impact on our analytical outcomes.
Because the implementation time of AI has been greatly reduced, we can now implement the analytics of all out data at scale and have actionable results that the business can make safe and secure business decisions. As we all know, time = money and time it takes to manually sift though mountains of data can be quite long. With AI we are now able to sort through our data, analyze it quickly and provide recommendations for business implementation in a fraction of the time than manual processes.
Once you are successful at explaining the alignment between security and the business, in the terms that the business understands, the more success you will see as a CISO.
Emerging triumphant, our CISO Hero now must turn to the implementation of the machines!

Implementation

Synthetic Data

- Synthesized
- Production Generated
- Prioritization

Source:
www.shimmeringcareers.com

AI

- Ethical
- Reduction of Data Bias
- Objectives of AI

Slide 17:
What type of AI and Synthetic data.
Synthetic Data based on production data.

- Use of fully synthesized data
    - Used for testing and dev environments
- Production data Generated data
    - Used for analytics, data sharing, monetization

Prioritization of what datasets you should use to create synthetic data and where you should only use synthetic data.
Ethical AI
We wanted an ethical AI that would explain the "Why" it concluded or the solution that is presented back to the human. As we evolve along with AI it is important to understand how the machine comes to its conclusions. We humans must trust but verify the end solution to ensure it aligns with our humanity. Example, we ask the AI to formulate the best course of treatment to save time and money for a terminal patient. Fair question, we want to conserve resources, the doctor's time so she can treat other patients, and of course save money for both the hospital and patient. Fair? What if the AI concludes that euthanizing the patient is the best treatment plan?

What if we use AI for national defense to provide the best course of action in conflict and the AI concludes that defending a densely populated city from attack is not in the best intertest in the long-term strategy to win the war and that sacrificing 8 million lives to safe guard 300 million is the best option?
AI must remain ethical for humans to trust using it. Not to follow it blindly but understand the How and Why it reached its decision, allowing the human to be "In or On the Loop" to correct the AI to meet our human ethics but also to monitor the data to ensure that data bias is removed from our datasets that feed our AI so that we can ensure that we are making the correct decisions without erroneous results that discriminate against specific groups or people.
And finally, our CISO Hero must set clear and achievable objective within the organization when implementing synthetic data and AI. For example, in my organization we have set 6 objectives to meet the demands of our hero's quest.

1. Creation of Synthetic Data for use in Dev and Test environment – elimination of production data in the lower environments
2. Creation of synthetic data sets and analyzed by AI to identify data quality issues
3. Creation of synthetic data for release to researchers and other interested parties
4. Analyze synthetic data to identify the best treatment outcomes
5. Enable predictive analytics to predict treatment plans and likelihood of successful outcomes
6. Enable predictive analytics to predict demands on resources and services by ingesting population datasets such as census data

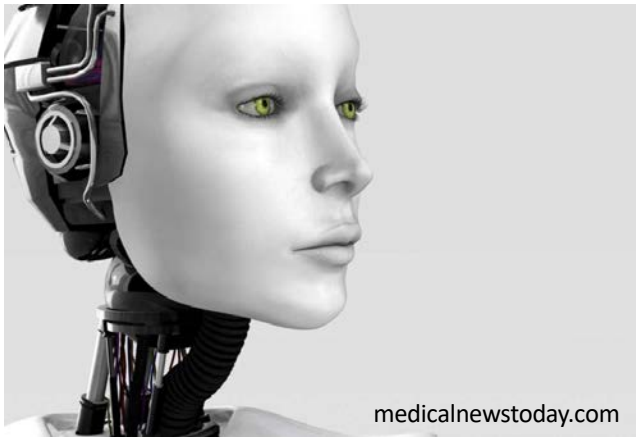
Now that our hero has successfully completed his quest, he now returns to the wise wizard, Will, to discuss what the future of AI looks like for us mere mortals.

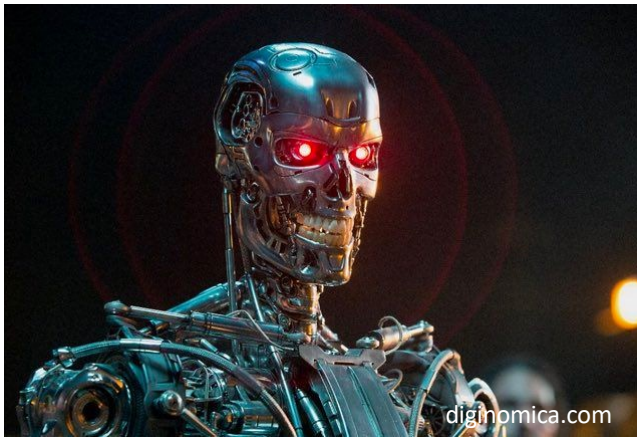
Phases and Objectives

- **Phase 1: Creation of synthetic data for use in Dev and Test Environments**
- **Phase 2: Creation of synthetic data sets and analyze to help identify data quality issues**
- **Phase 3: Creation of synthetic data sets for release to researchers and other interested parties for their consumption**
- **Phase 4: Analyze synthetic data to identify treatment outcomes**
- **Phase 5: Enable predictive analytics to predict treat plans and likelihood of successful outcomes**
- **Phase 6: Enable predictive analytics to predict demands of services by region by ingesting census data**

The Future

The future of AI ?





Slide 20:
The future of AI is now and the pace of innovation in this space is moving with great speed. The pandora's box is open and I encourage everyone to think about the use of AI in your business' and make sure you are building a responsible and ethical framework in which you operate your AI to ensure that it is guided and aligned with our human and cultural values. While I wish our law makers would intervene with regulation and legislation, the reality is they are not currently equipped, and their systems move too slowly when compared to the speed of advances. This means each of us has an opportunity to help make the future with AI one that we and our children can enjoy and be proud of.

Journey's End

Source: rawpixel.com

Contact Info:

Glendon Schmitz, CISSP

https://www.linkedin.com/in/glendonschmitz

Will Goddin

www.linkedin.com/in/willgoddin

Slide 21:
This brings us to our journey's end. Questions?