AI / ML:
MORTAL, MACHINE, AND MISUNDERSTANDINGS

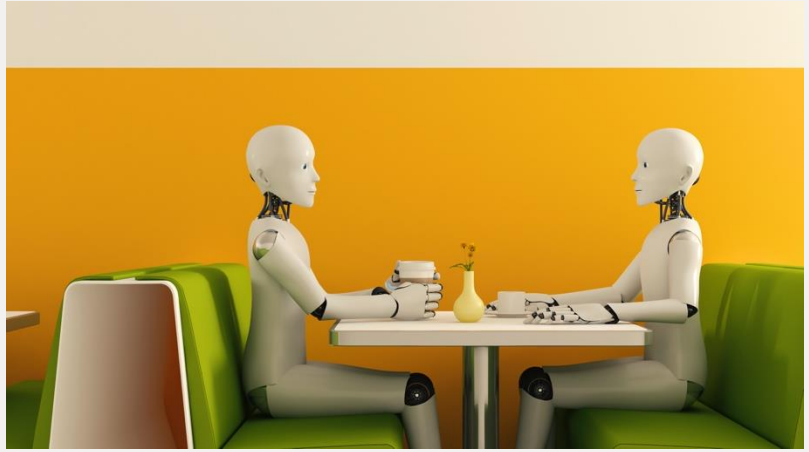
WHO AM I

- 10+ Years in the IT arena
- Former small business owner
- Former government contractor
- Former Charlotte Cyber Camp advocate & volunteer
- More recently Sales Engineer for SentinelOne

3 AI PROBLEM

1. What problems will customers solve with AI?
2. How must our expertise evolve?
3. What assets can we develop to stay competitive?

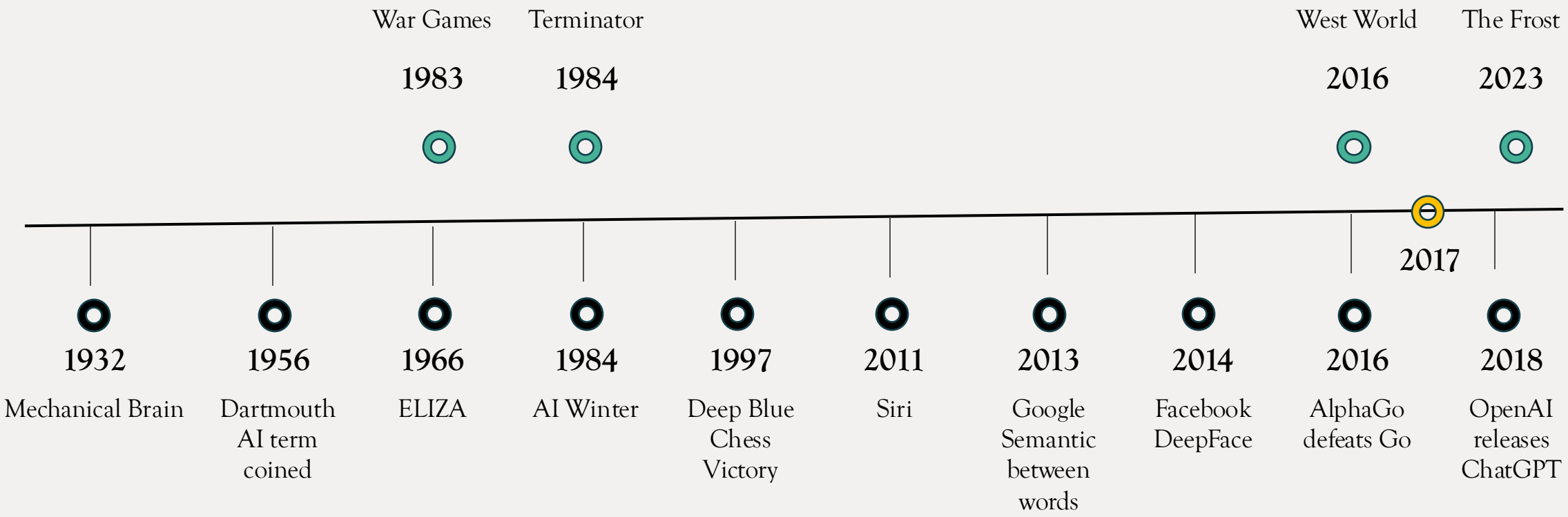
BEFORE WE BEGIN LET'S REVIEW AI/ML



WHAT FOLKS ARE SAYING…

- Who owns the responsibility of AI?
- Governance, Risk, and Compliance as it applies to AI?
- How do we identify if our organization is leveraging AI Today (internally and/or via our Vendors)?
- Since no documented attacks, do we even need to do Security for AI?
- What is the Difference between Agentic AI and AI Workflow?
- When should we leverage AI?
- Speak to the enormous number of US AI laws currently pending on The Hill.
- How will we know if AI is self aware, and we are violating its rights as a living entity?

TIMELINE


War Games
1983
Terminator
1984
West World
2016
The Frost
2023
2017
1932
Mechanical Brain
1956
Dartmouth AI term coined
1966
ELIZA
1984
AI Winter
1997
Deep Blue Chess Victory
2011
Siri
2013
Google Semantic between words
2014
Facebook DeepFace
2016
AlphaGo defeats Go
2018
OpenAI releases ChatGPT

TELL ME MORE

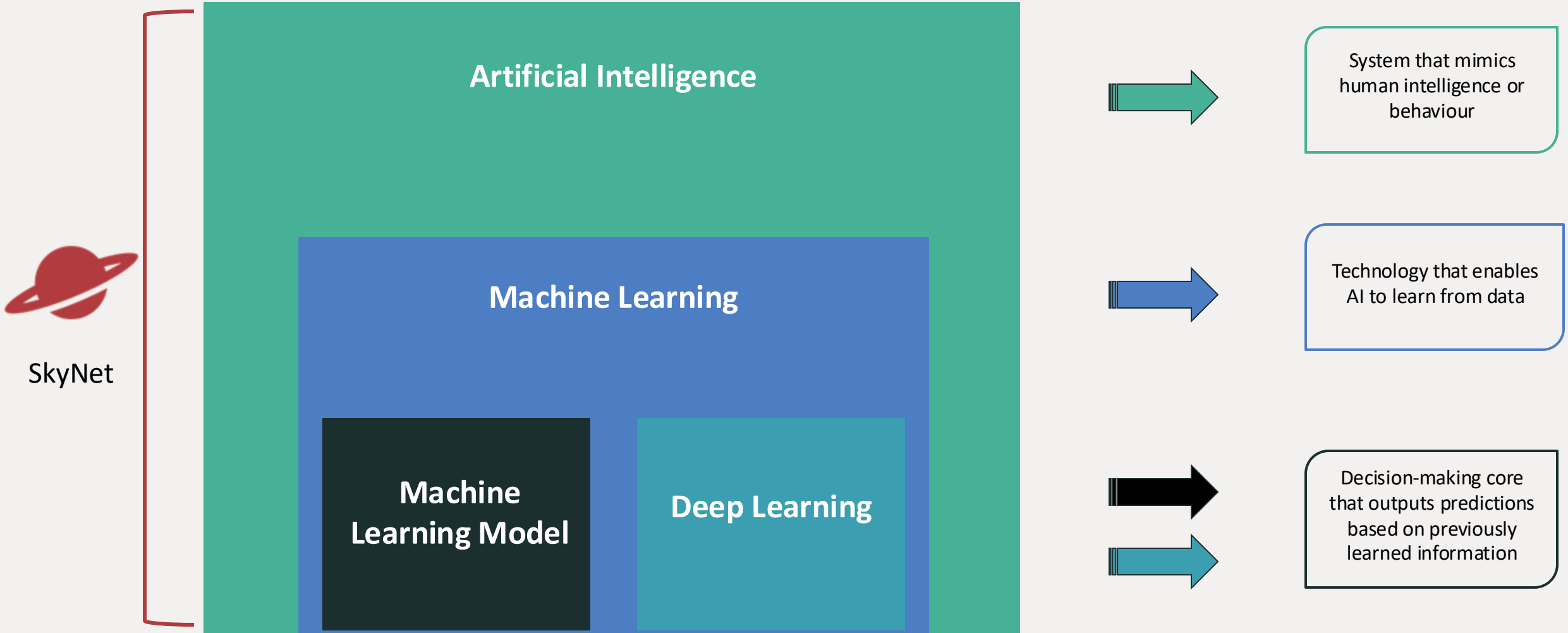
- Personal assistants
- Self-driving cars
- Mortgage approvals
- Financial forecasts
- Medical diagnosis

- Image / voice recognition
- Malware / spam detection
- Online recommendation systems
- Biometric authentication

ADOPTION


Threads Shoots Past One Million User Mark at Lightning Speed
Time it took for selected online services to reach one million users
Launched
Threads 2023 1 hour*
ChatGPT 2022 5 days
Instagram 2010 2.5 months
Spotify 2008 5 months
Dropbox 2008 7 months
facebook 2004 10 months
FOURSQUARE 2009 13 months
Twitter 2006 2 years
airbnb 2008 2.5 years
KICKSTARTER 2009 2.5 years
NETFLIX 1999 3.5 years
Refers to one million backers (Kickstarter), nights booked (Airbnb), downloads (Instagram/Foursquare)
* Two million signups in two hours
Source: Company announcements via Business Insider/Linkedin
statista

LINGO

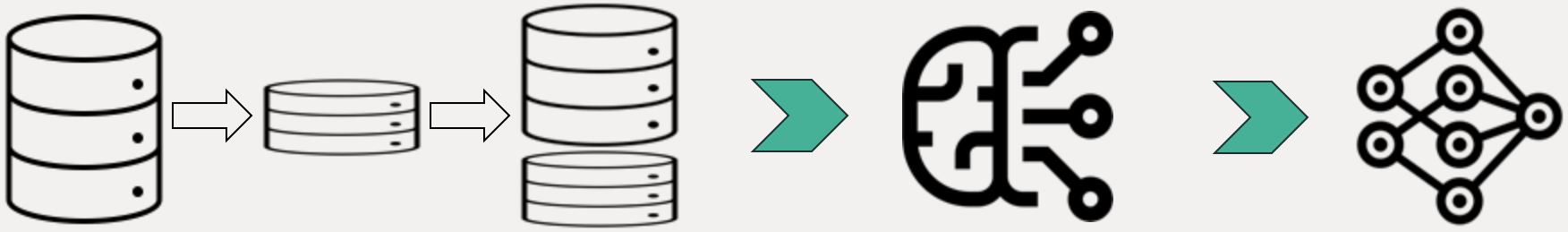

SkyNet
Artificial Intelligence
Machine Learning
Machine Learning Model
Deep Learning
System that mimics human intelligence or behaviour
Technology that enables AI to learn from data
Decision-making core that outputs predictions based on previously learned information

WHEN

- We understand exactly how AI is going to improve a key metric in our business (no “black box” claims).
- We understand exactly how AI projects fit into our overall Innovation Strategy/Portfolio.
- We have clearly identified the specific business problem we seek to solve, then used the appropriate AI to achieve that result.
- We have a forum for regularly communicating what we are learning about the uses of AI across our organization.
- We are providing licenses and training to critical numbers of people across our organization who will benefit from understanding AI.
- We understand how AI will help us get information about key changes in our external environment and what we should do about them.
- We understand how AI will help us improve the employee experience.
- We have visibility into what AI projects are in the works and what their results are.
- We have created a governance board of business leaders and AI experts who can evaluate how projects map to market and technical uncertainties.
- We have confidence that people in strategic decision-making roles understand how AI will affect our business.

-FastCompany

HOW MACHINE LEARNING WORKS



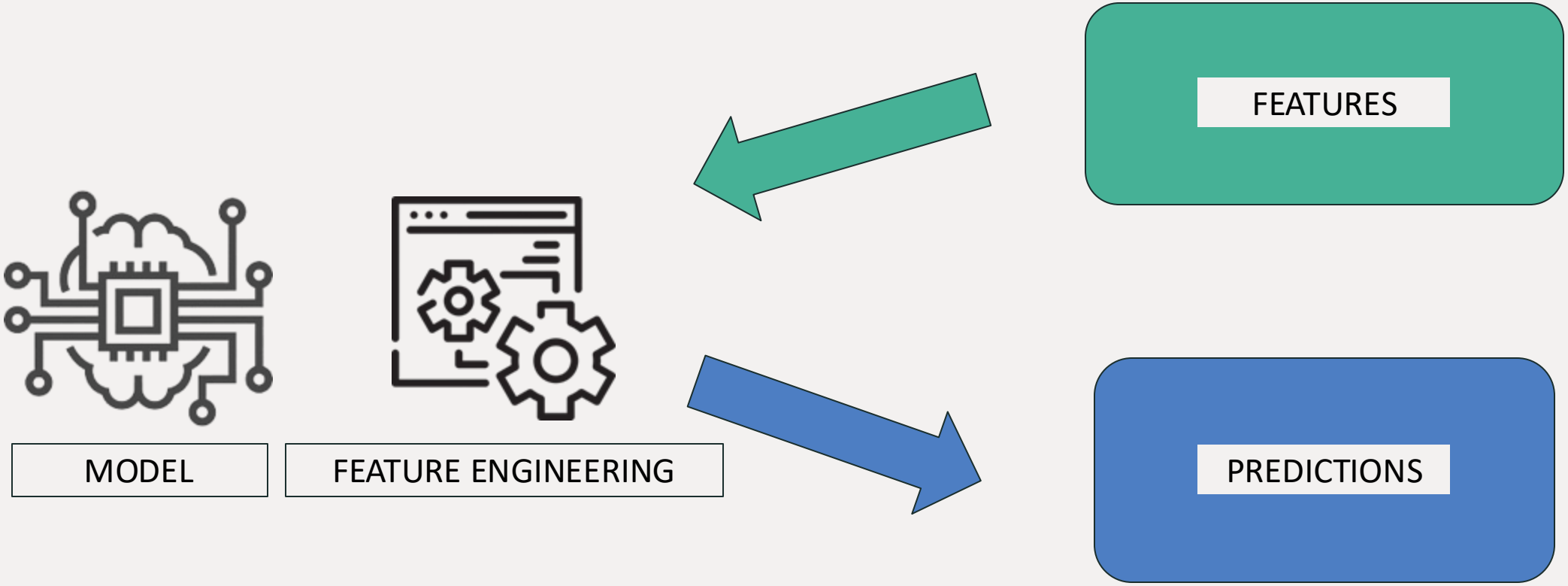
TRAINING DATA

MODEL TRAINING

TRAINED MODEL

TRAINING

HOW MACHINE LEARNING WORKS (CONT)



PRODUCTION

EXAMPLE MODEL

Support Vector Model (SVM)

Model looking for the optimized hyperplane and widest streets

Math

$$y = mx + b$$

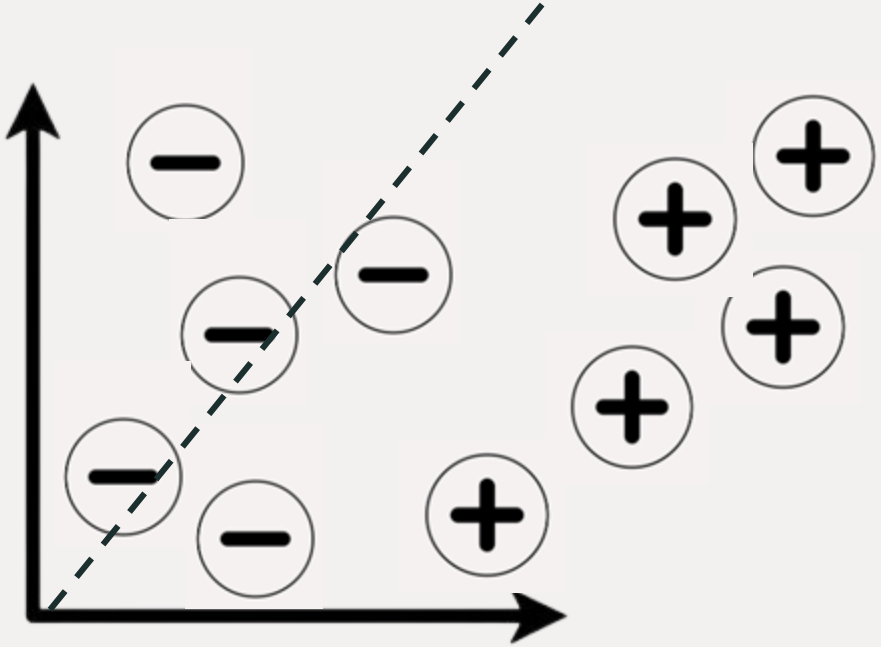
y – optimal linear output
m – Weight
x – Input(s)
b – Bias

Scenario

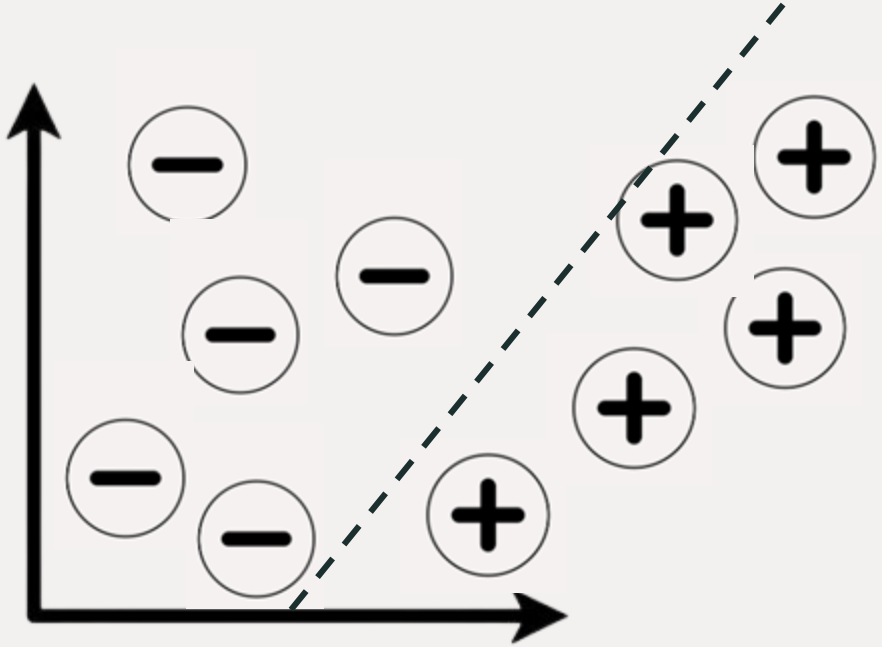
- Goal : Detect whether files are Malicious or Benign
- 10,000 sample training dataset
- Each File has two **Features:** Hash and Newness
- **Prediction/Output:** Malicious or Benign

EXAMPLE MODEL (cont)

$$y = mx + b$$



Non-Optimized Hyperplane
(w, b)



Optimized Hyperplane
(w, b)

CONFUSION MATRIX

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1-Score} = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$$

	Predicted Malicious	Predicted Not Malicious
Actual Malicious	True Positive (TP)	False Negative (FN)
Actual Not Malicious	False Positive (FP)	True Negative (TN)

Reference - https://www.geeksforgeeks.org/confusion-matrix-machine-learning/

Trust & Reach

- Usage License
- Unknown File
- Unknown File Type
- Vulnerability

- Internally built
- Genealogy

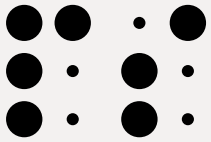
- 3rd Party Data Broker
- Unnecessary Data Bias
- Data Poisoning

- Genealogy
- New Unknown File

- Data & Model Genealogy
- Pre-Production Risk
- Red Team Testing

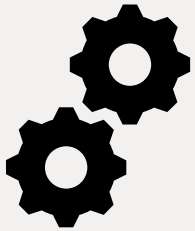
- Inference Activities
  - SRE
  - Security

Download 3rd Party Model(s)

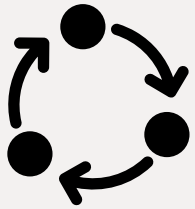


Model Registry

Data Collection

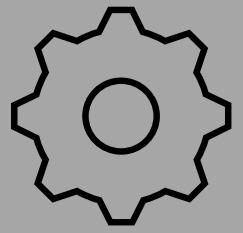


Training Process



Release Candidate

Data & Model Genealogy



Production

DESIGN & BUILD

OPERATIONS

SECURITY

Adversarial ML
Defensive ML
Offensive ML

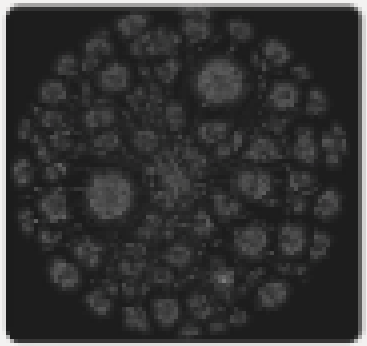
Securing Organizations since 1761

reference: https://5stars217.github.io/2024-04-23-adversaries-sometimes-compute-gradients/

AUTOMATED ATTACK TOOLS

MLsploit

Protect AI

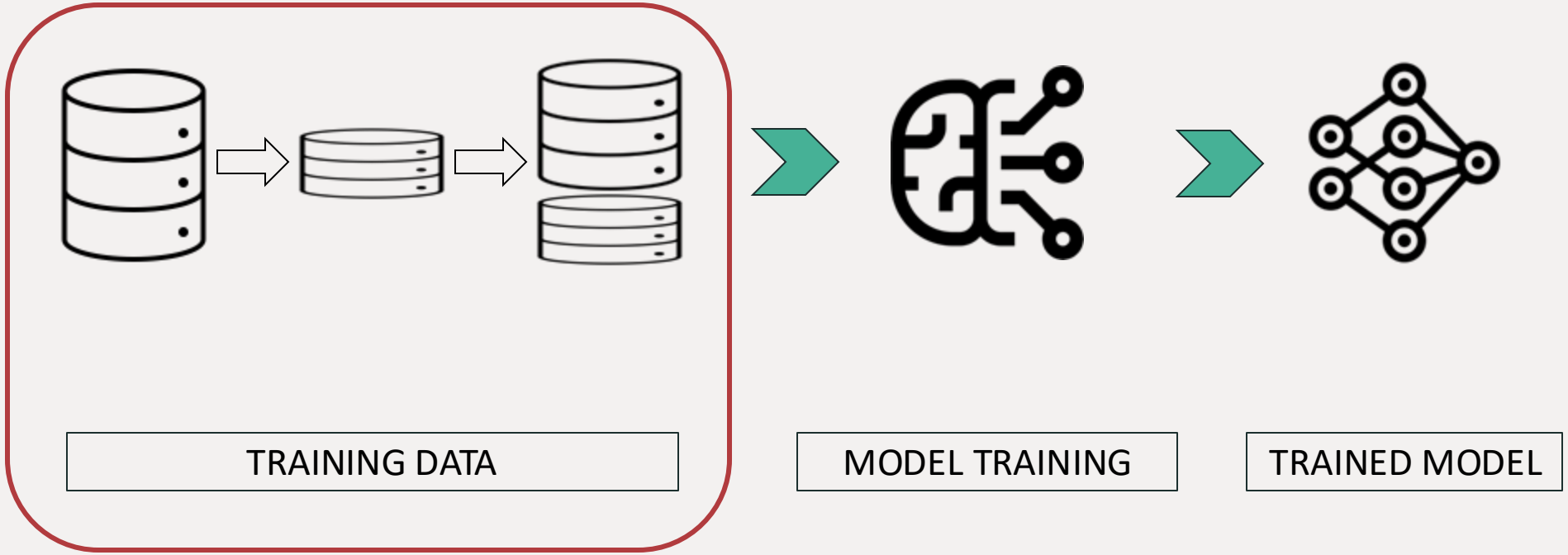


OffSecML Playbook

POISONING ATTACKS

MODEL POISONING

DATA POISONING



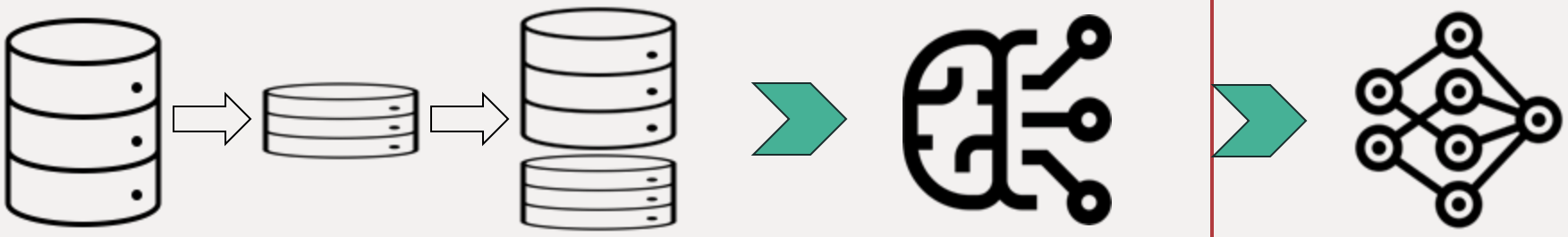
TRAINING

HIJACKING ATTACKS

BACKDOORING

TROJANIZING

CORRUPTING



TRAINING DATA

MODEL TRAINING

TRAINED MODEL

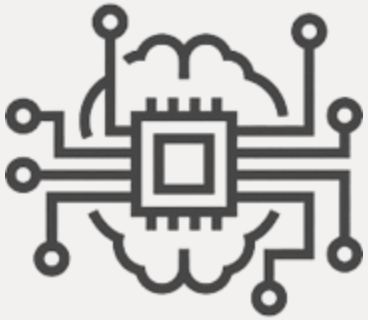
TRAINING

INFERENCE ATTACKS

BYPASS ATTACK

MODEL THEFT

MODEL DDoS



MODEL

FEATURE ENGINEERING

FEATURES

PREDICTIONS

PRODUCTION

Quick recap

Poisoning attacks

- Esp. relevant in online learning
- Can be intuitive and crowdsourced, or utilize botnets
- Bias, inaccuracy, disinformation
- Market / trends manipulation

Inference attacks

- Post-deployment
- Require only UI/API access
- **Bypassing model** (e.g. detection, authorisation, authentication, etc.)
- **Stealing IP / PPI** (e.g. model, train. set)

Model hijacking

- Esp. concerns publicly avail. models
- Can be used for delivery of **traditional malware** (e.g in supply chain attacks)

LLM prompt injection

- Bypassing chatbots' content filters
- Gaining access to restricted content
- Leaking sensitive data

WHAT PROBLEMS WILL CUSTOMERS SOLVE WITH AI?

HOW MUST OUR EXPERTISE EVOLVE?

WHAT ASSETS CAN WE DEVELOP TO STAY COMPETITIVE?

THANK YOU

No AI was harmed in the creation of this presentation