SAND2023-02366C

Sandia National Laboratories

Exceptional service in the national interest

Test and Evaluation of Systems with Embedded Artificial Intelligence Components

Michael R. Smith

DATAWork 2023

Sandia National Laboratories

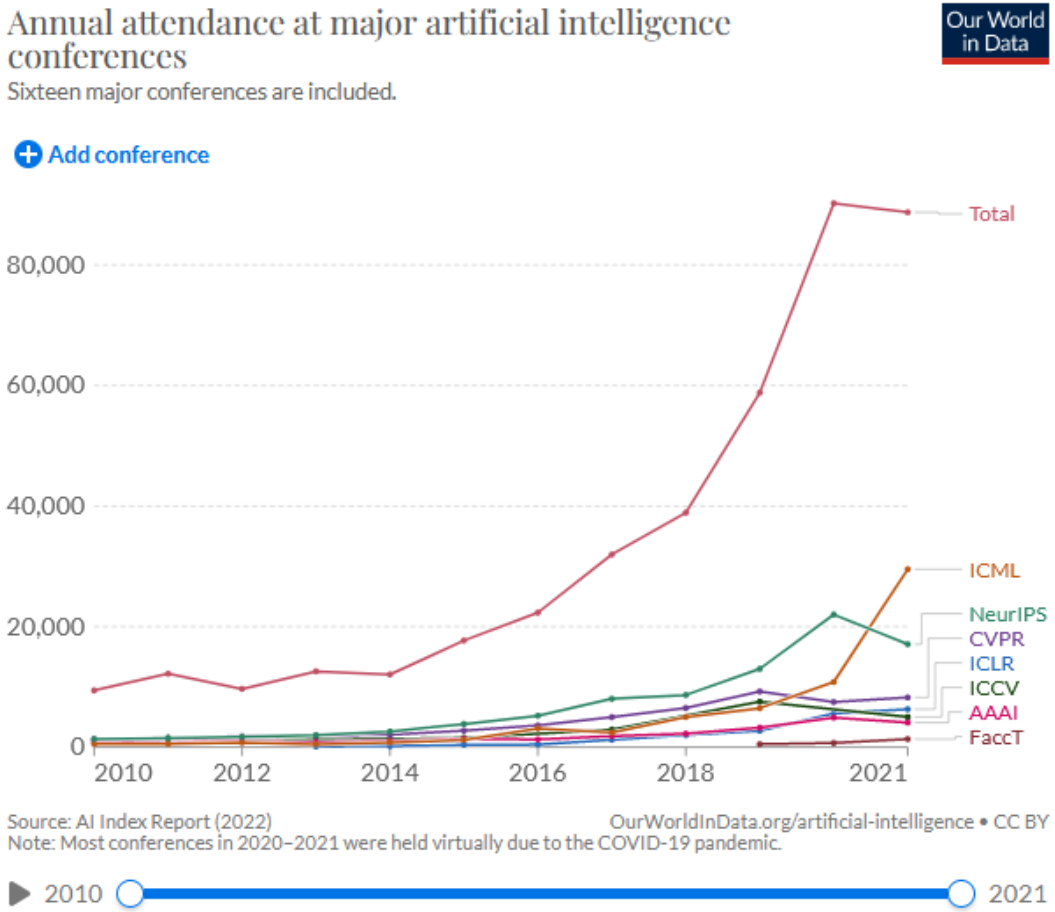
Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

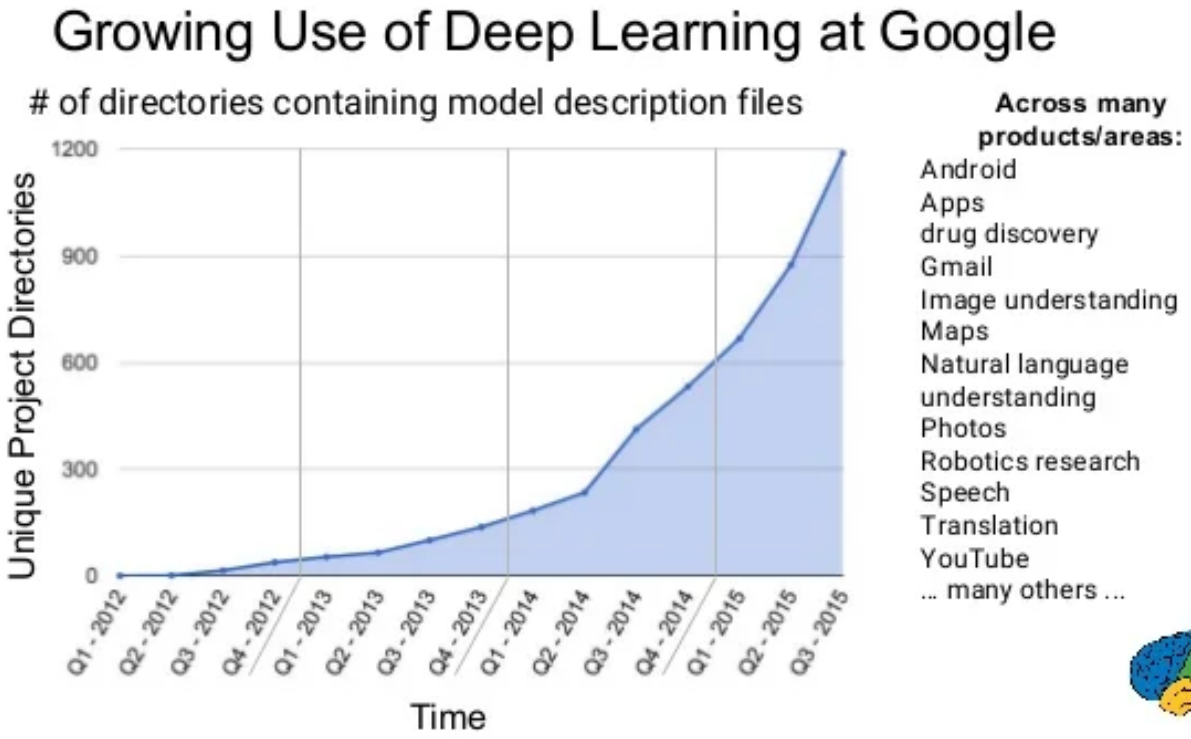
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and wholly owned subsidiary of Honeywell International Inc. for the U.S. onal Nuclear Security Administration under contract DE-NA0003525.

What is AI/ML and why do we care?

The age of ML and AI

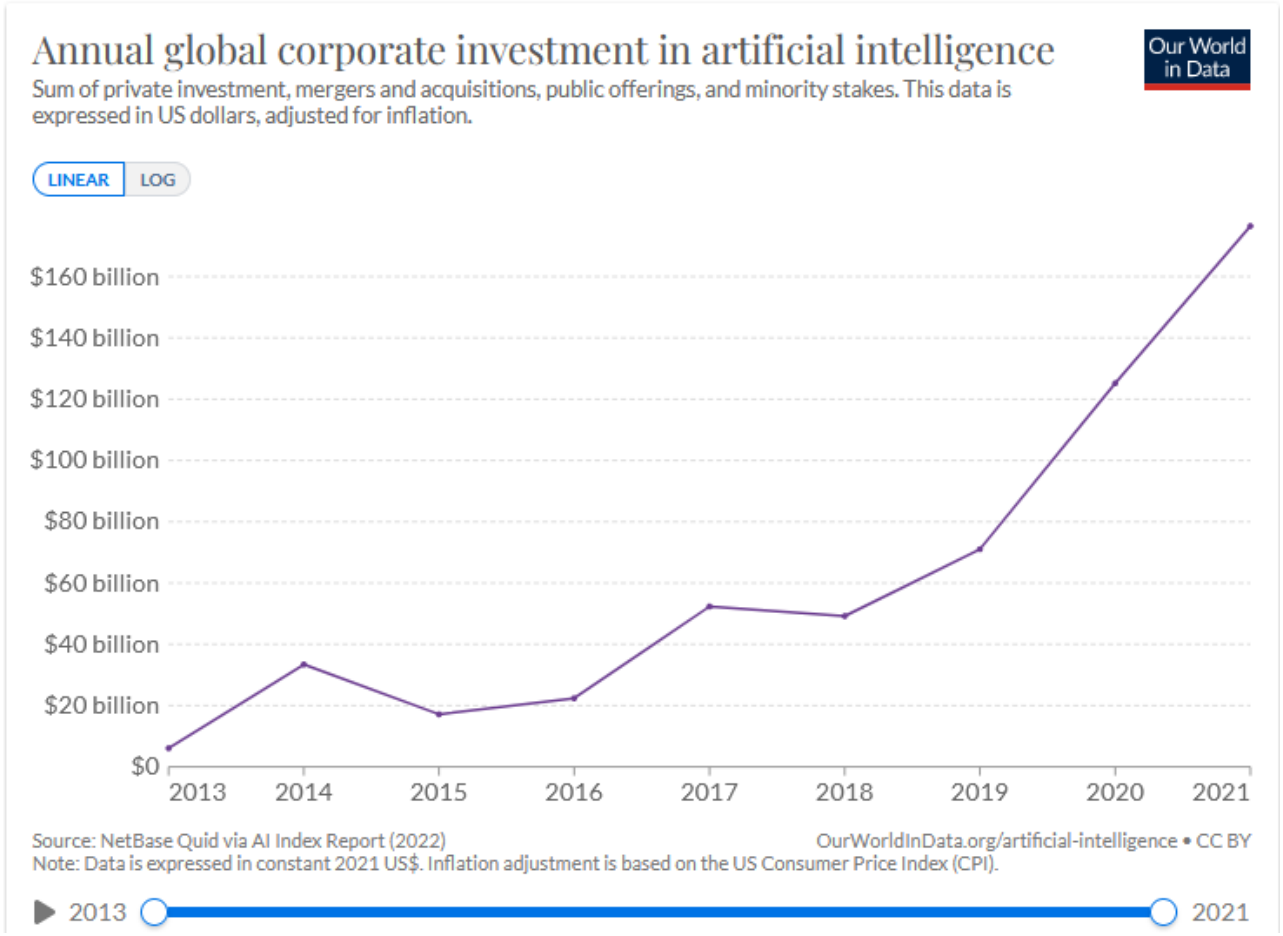


https://ourworldindata.org/ai-investments

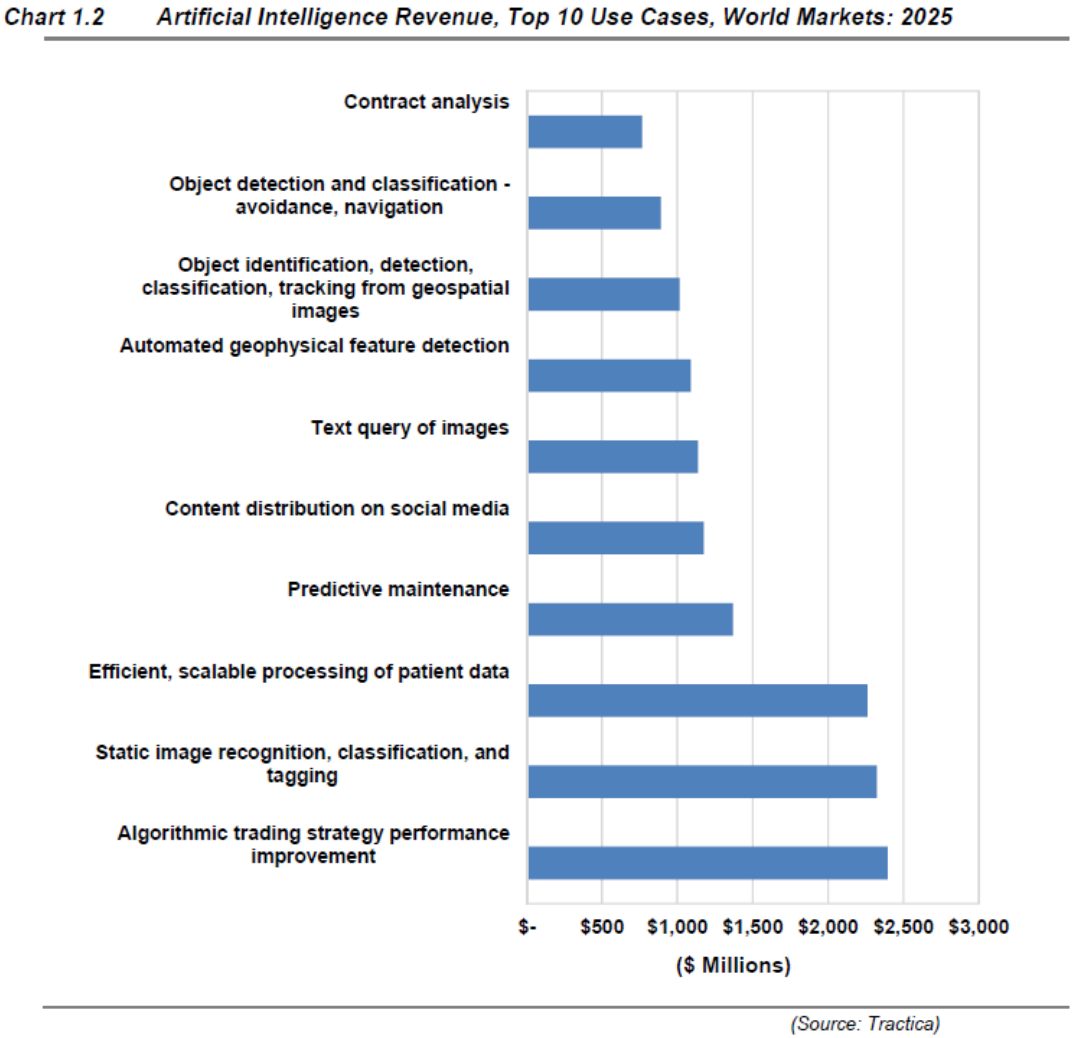


https://www.analyticsvidhya.com/blog/2017/04/comparison-between-deep-learning-machine-learning/

More than just academic

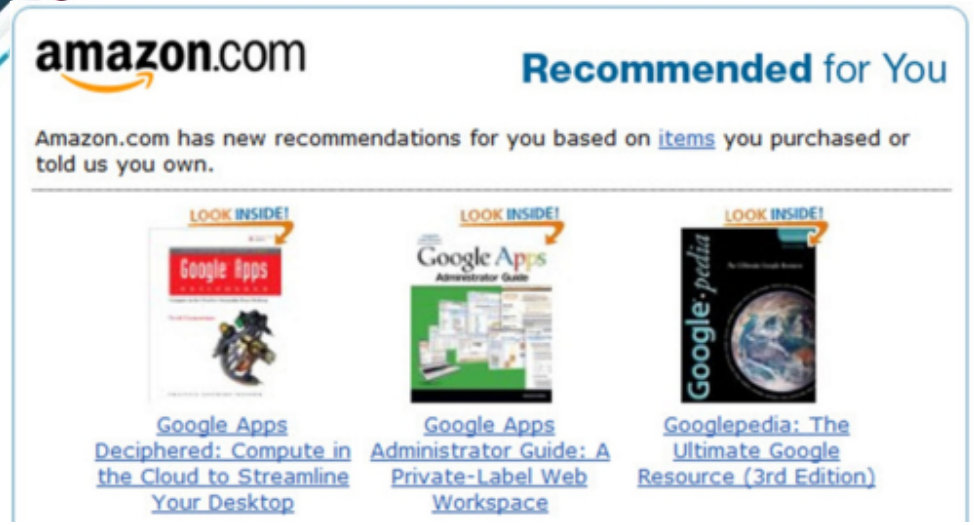


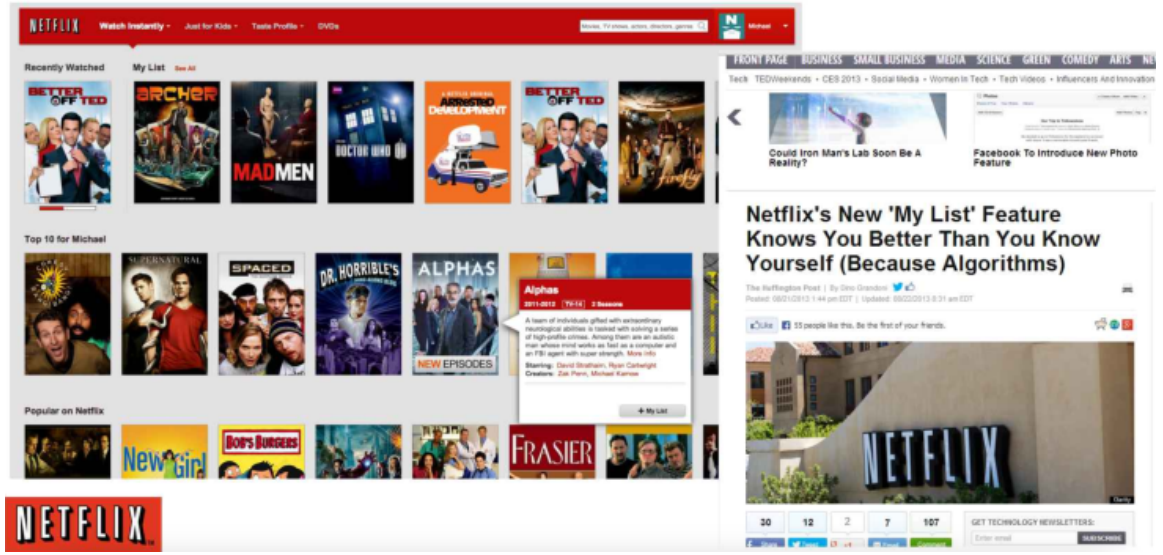
https://ourworldindata.org/ai-investments

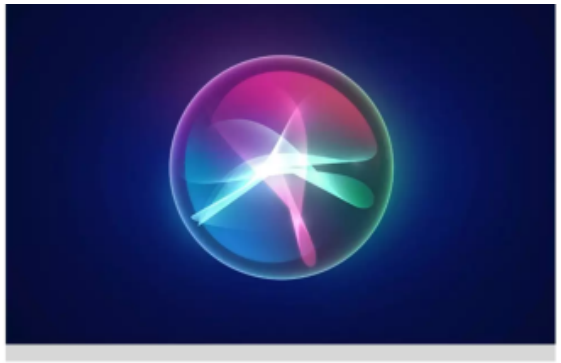


https://www.top500.org/news/market-for-artificial-intelligence-projected-to-hit-36-billion-by-2025/

ML is everywhere



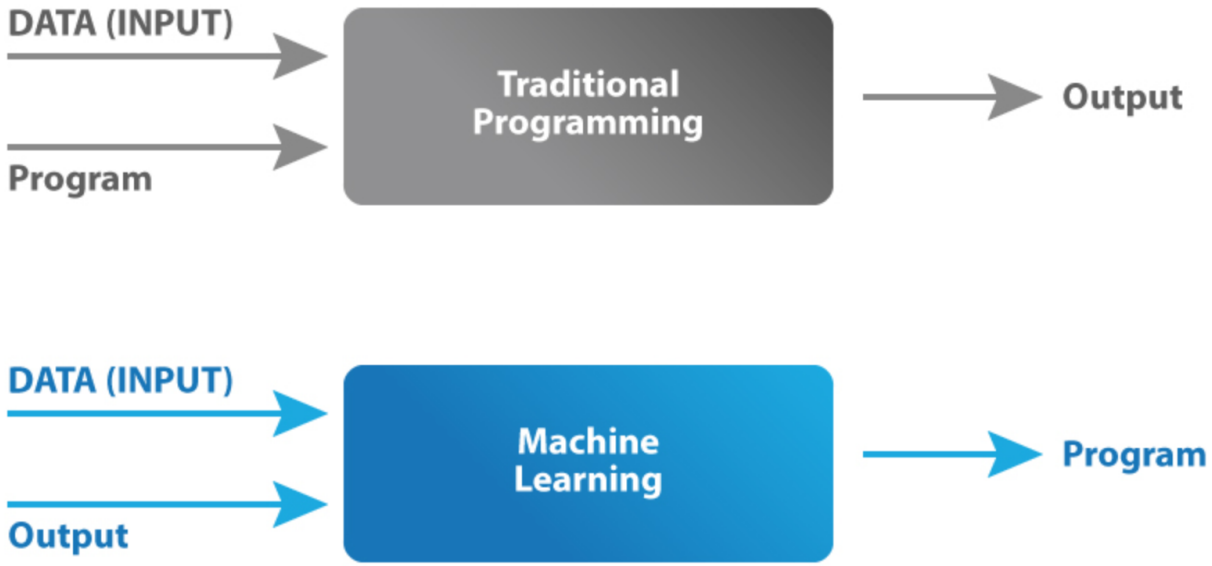




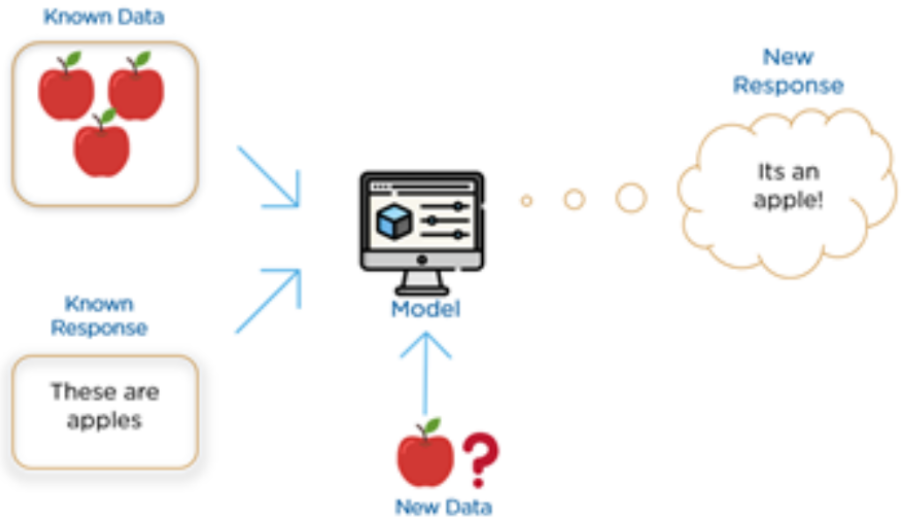
Set Up "Hey Siri"

This helps Siri recognise your voice when you say "Hey Siri".

What is ML?

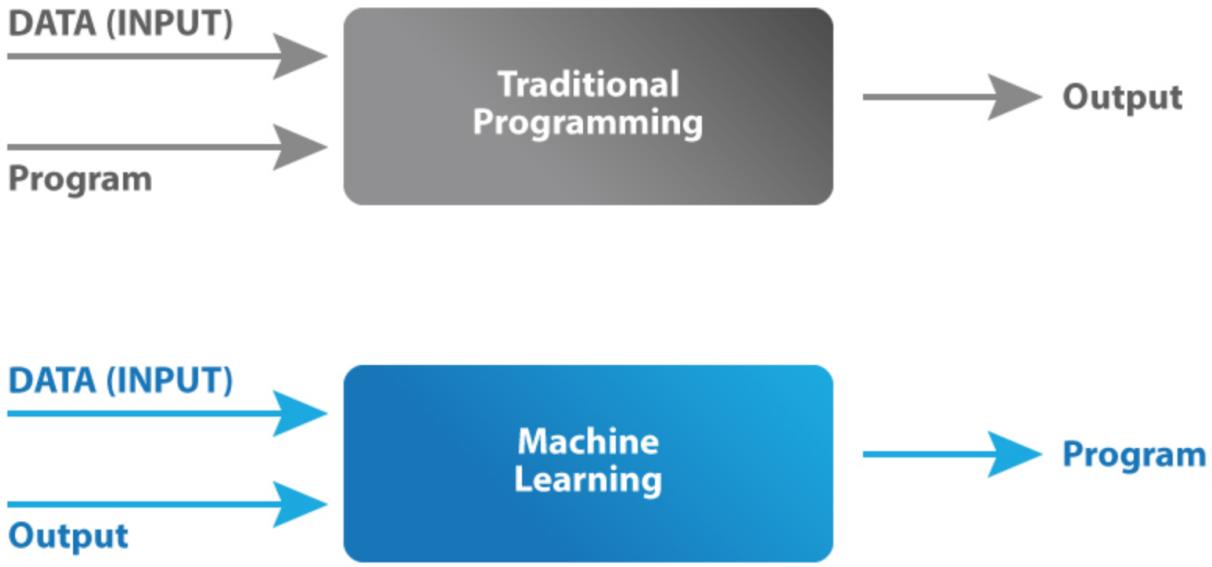


https://www.mygreatlearning.com/blog/what-is-machine-learning/

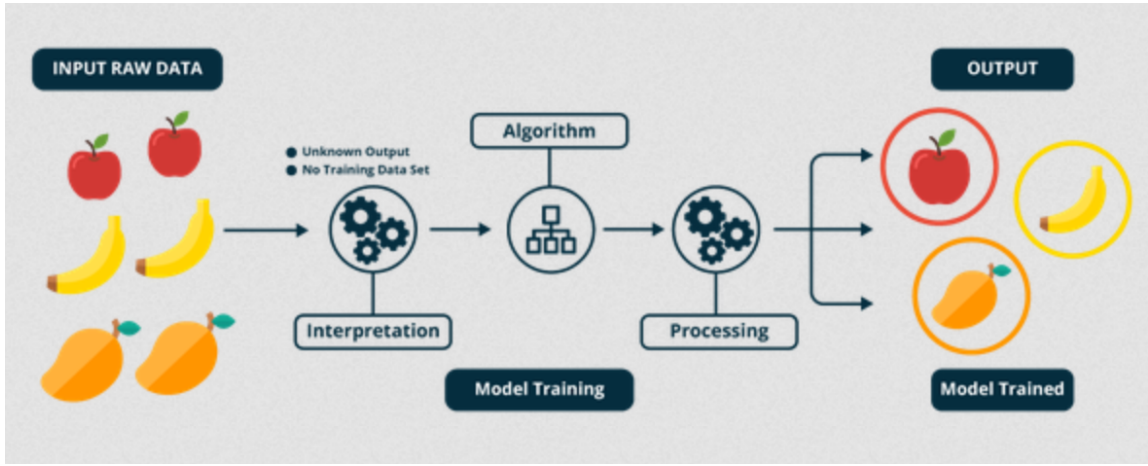


https://www.onemodel.co/blog/ai-academy-what-is-machine-learning

What is ML?



https://www.mygreatlearning.com/blog/what-is-machine-learning/



https://www.onemodel.co/blog/ai-academy-what-is-machine-learning

Why does T&E care about ML?

US launches artificial intelligence military use initiative

By MIKE CORDER February 16, 2023

Drone advances in Ukraine could bring dawn of killer robots

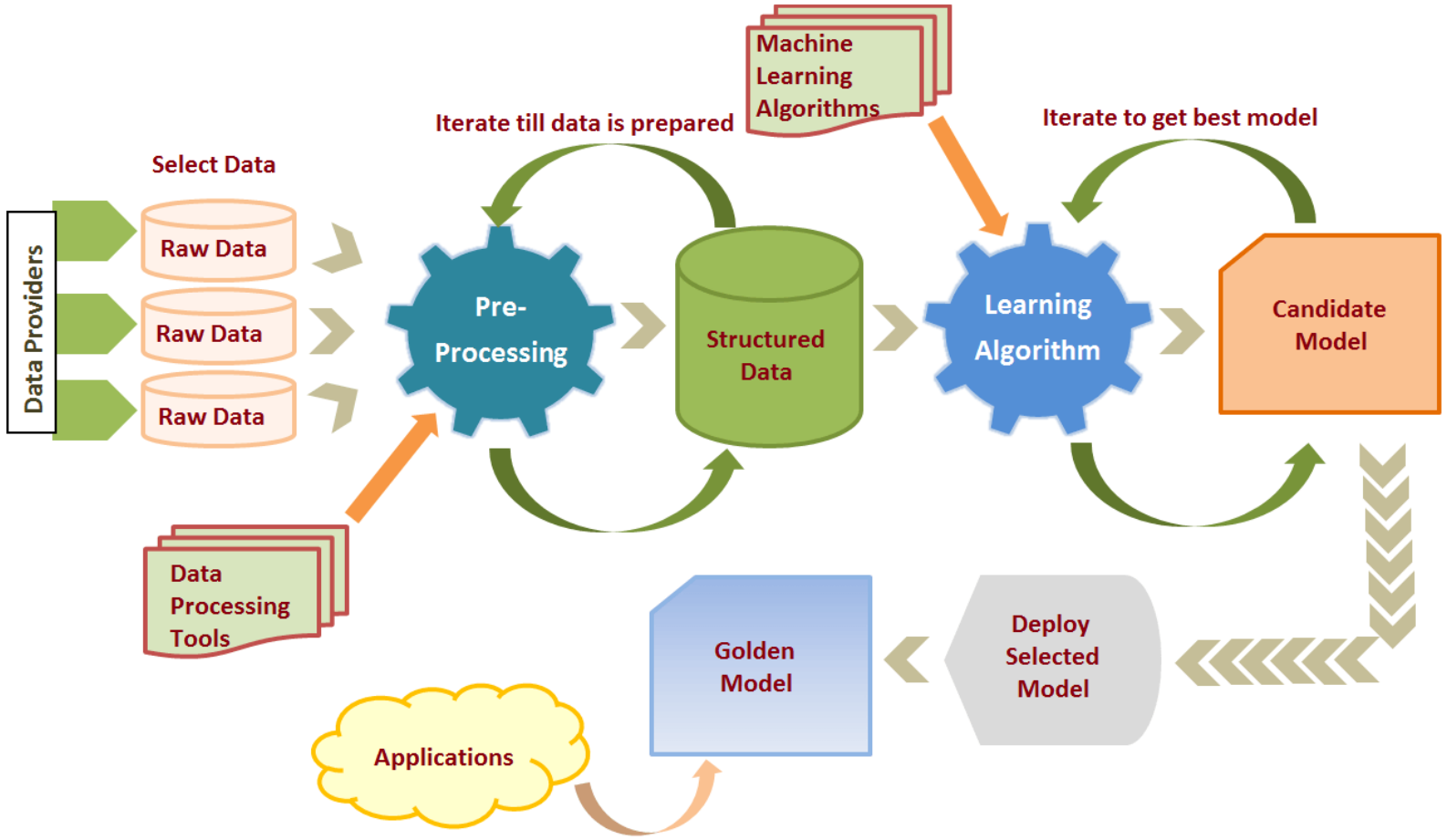
By FRANK BAJAK and HANNA ARHIROVA January 3, 2023

Elon Musk among experts urging a halt to AI training

Hey, Siri: Hackers Can Control Smart Devices Using Inaudible Sounds

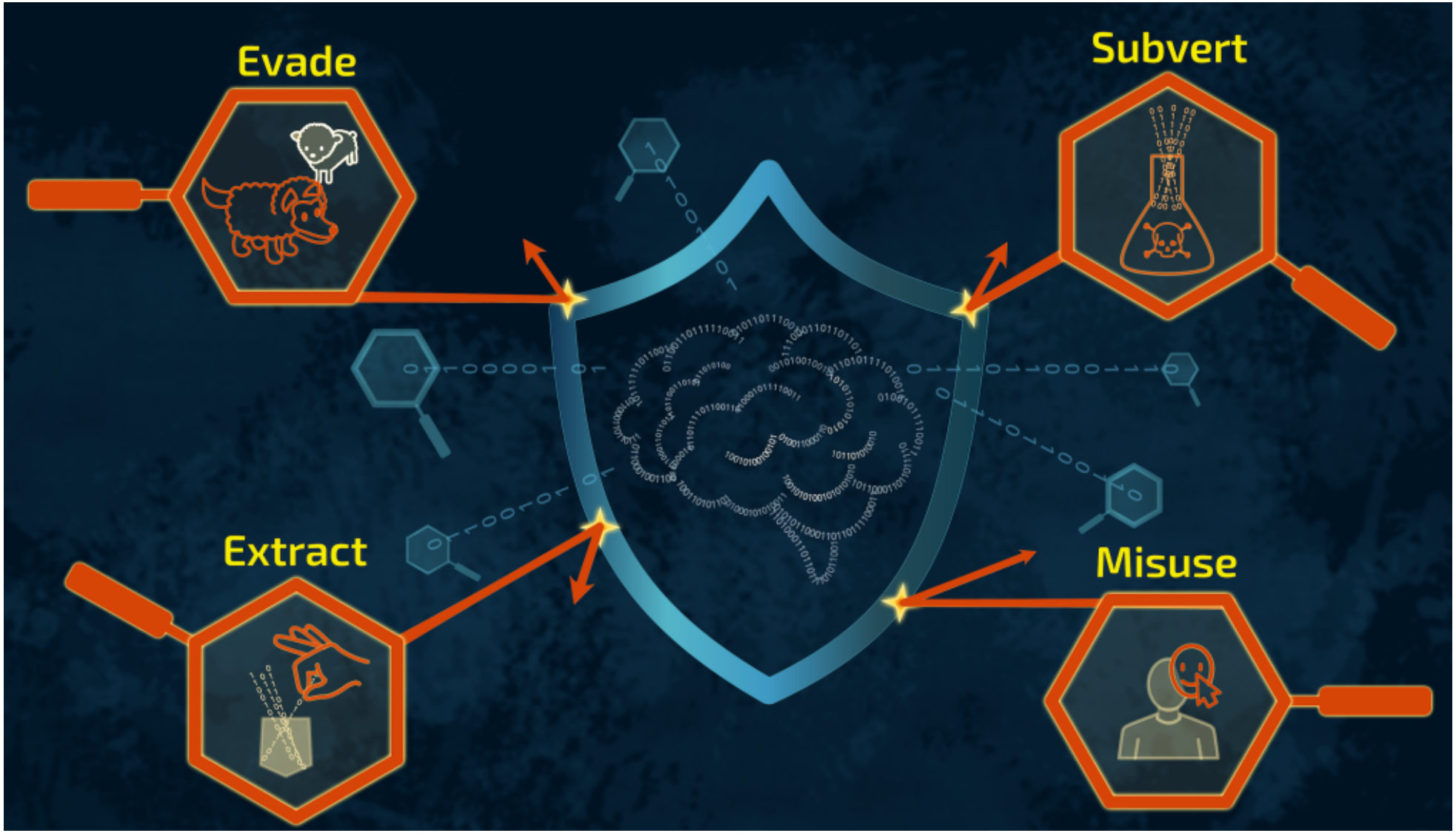
The ML Process

- Data Collection
- Feature engineering/selection
- Model selection
- Hyper-parameter tuning
- Integration into system
  - Preprocessing
  - ML component

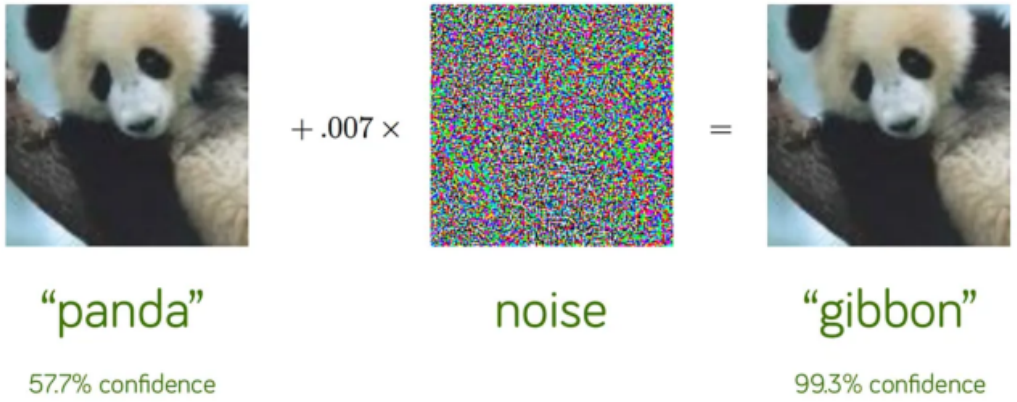


https://elearningindustry.com/machine-learning-process-and-scenarios/

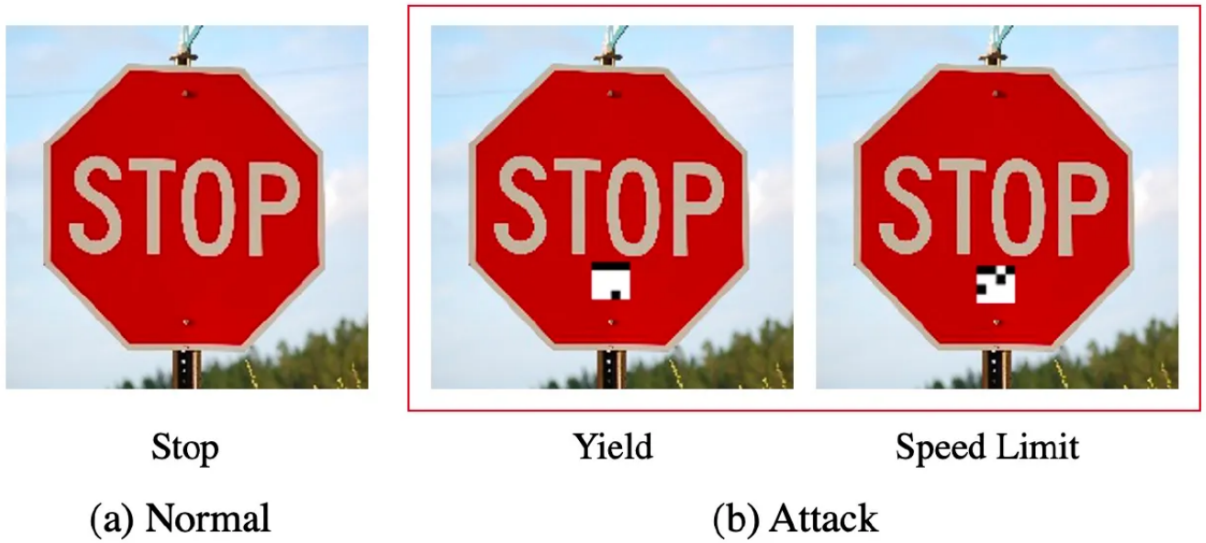
What Vulnerabilities Does ML Present



ML Vulnerabilities



Explaining and Harnessing Adversarial Examples, *Goodfellow et al, ICLR 2015*



An Embarrassingly Simple Approach for Trojan Attack in Deep Neural Networks, Tag et al, KDD, 2020

ML Vulnerabilities

Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, Fredrikson et al, SIGSAC 2015

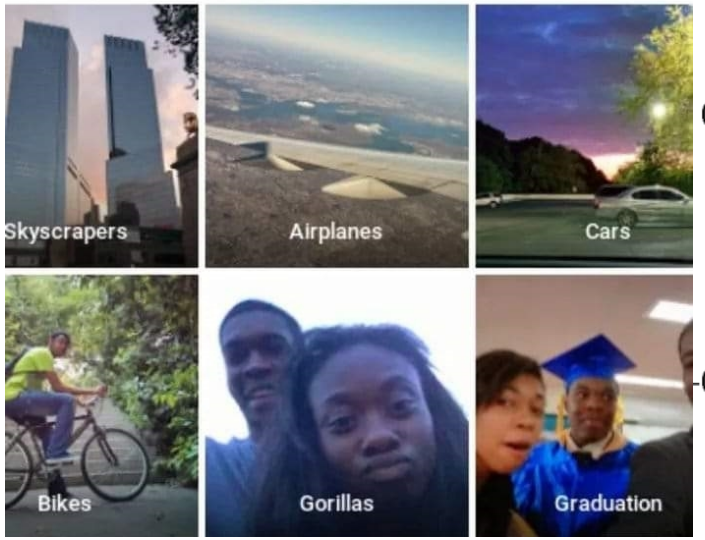
Jordan Peele uses AI, President Obama in fake news PSA

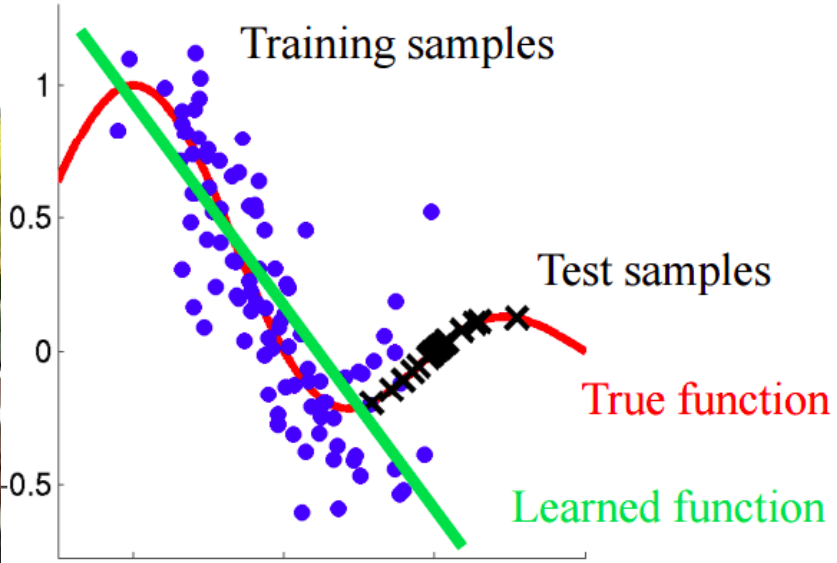
https://www.siliconrepublic.com/machines/jordan-peele-ai-obama-fake-news

What else could go wrong?

Two years later, Google solves 'racist algorithm' problem by purging 'gorilla' label from image classifier

Cory Doctorow 1 year ago





Tesla 'full self-driving' triggered an eight-car crash, a driver tells police

By Matt McFarland, CNN
Published 5:41 PM EST, Wed December 21, 2022

Assessing Systems Embedded with ML/AI

T&E vs Assessment

Testing and Evaluation

- Assist in risk management
- System and components are compared against requirements or specifications
- Validate measurable quantities

Machine Learning

- Performance on a validation data set
- Certain defenses in place

Cyber Assessment

- Assist in risk management
- Assess current cybersecurity posture (red teaming)
- Identify vulnerabilities and impacts even if specifications and requirements are met

Machine Learning

- Performance on real-world data
- Vulnerabilities against possible attacks

Lots of Tools

deepchecks.

Adversarial Robustness Toolbox

usnistgov/**dioptra**

NIST

Test Software for the Characterization of AI Technologies

7 Contributors | 38 Issues | 1 Discussion | 28 Stars | 8 Forks

OUR SCOPE

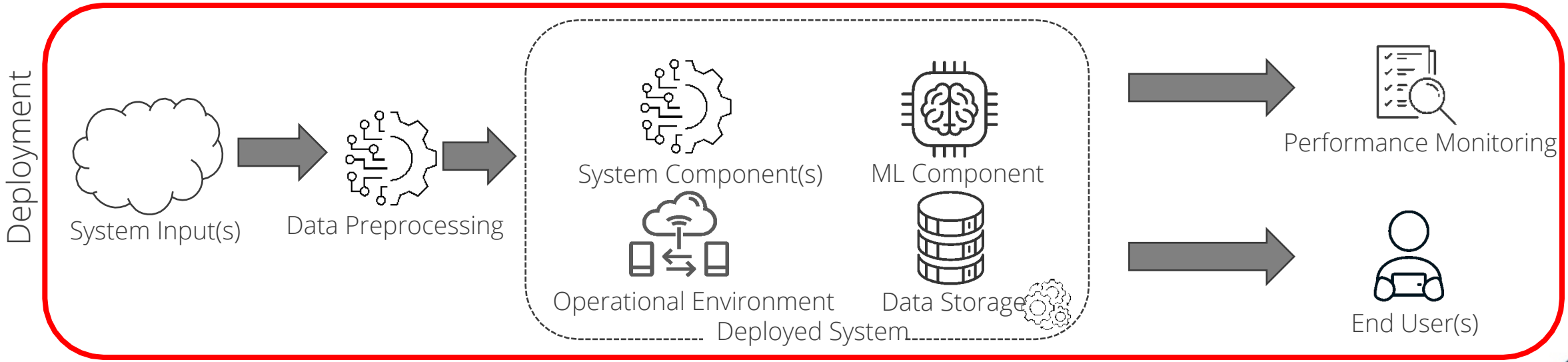
- System-level examination
  - How does the ML system integrate into a larger system?
  - How are output propagated through the system?
- **Not always direct access to the ML model**

IBM Research

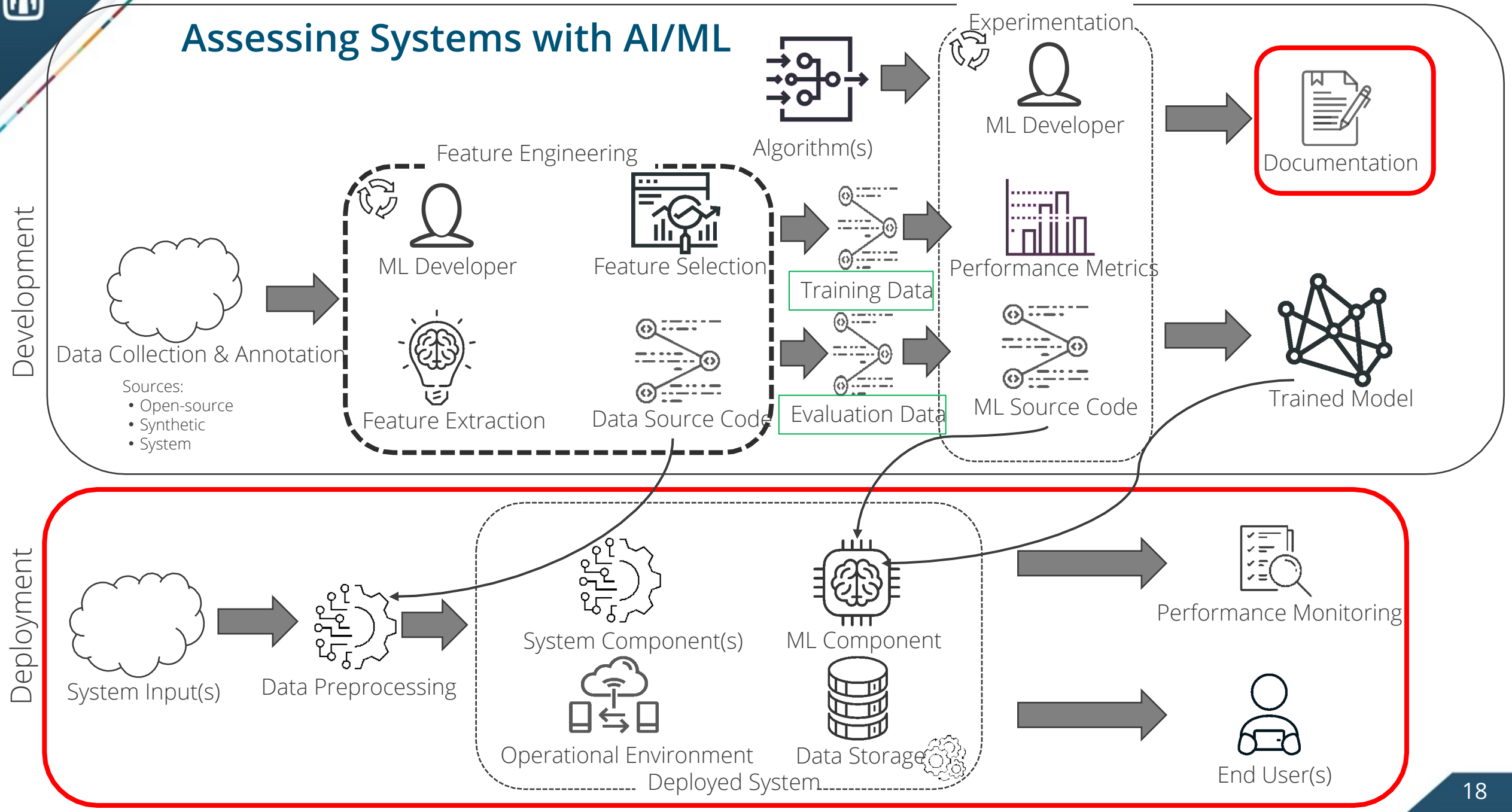
Explainability | Fairness | Privacy | Uncertainty

https://research.ibm.com/topics/trustworthy-ai

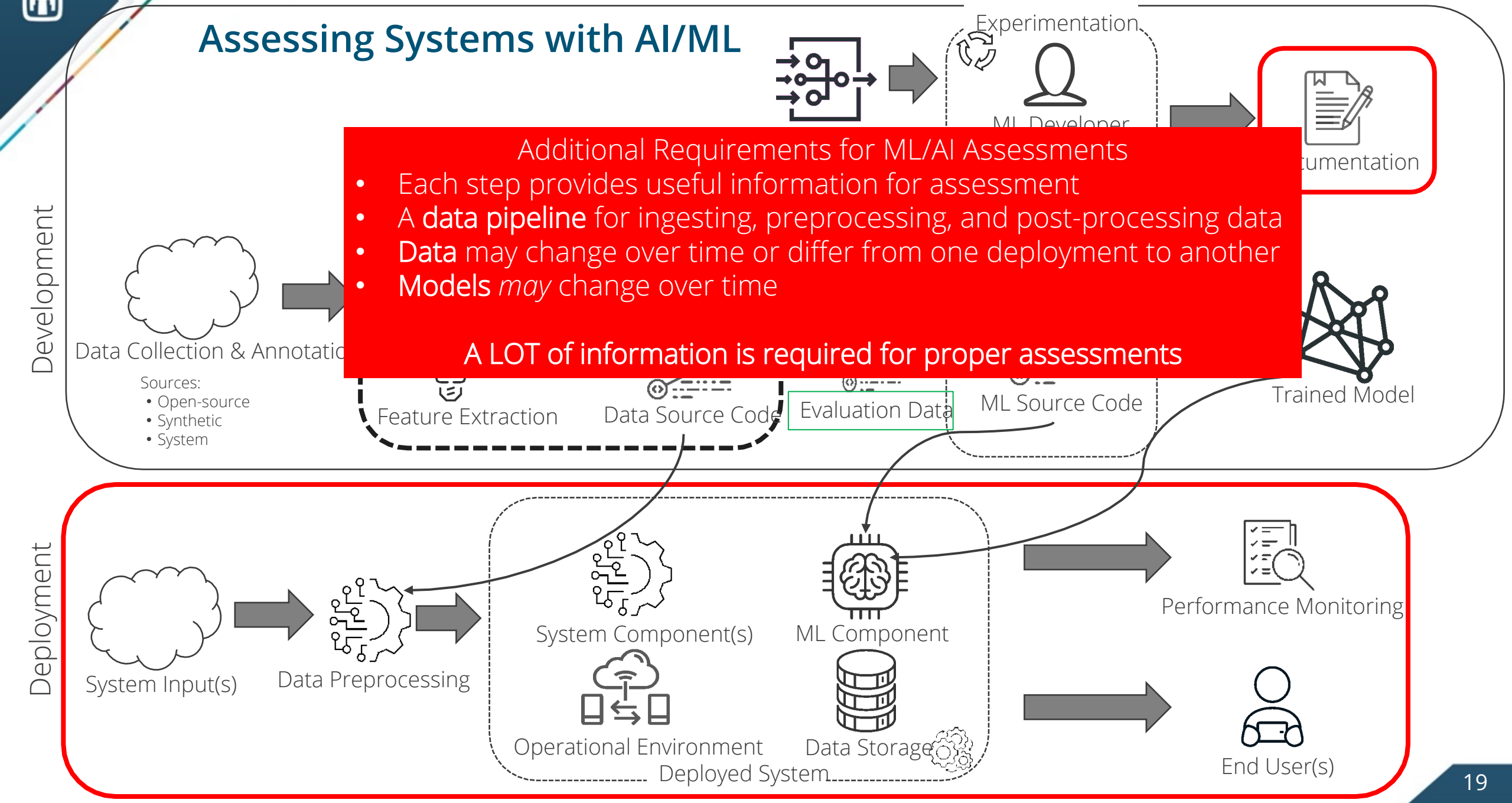
Assessing Systems with AI/ML



Assessing Systems with AI/ML



Assessing Systems with AI/ML

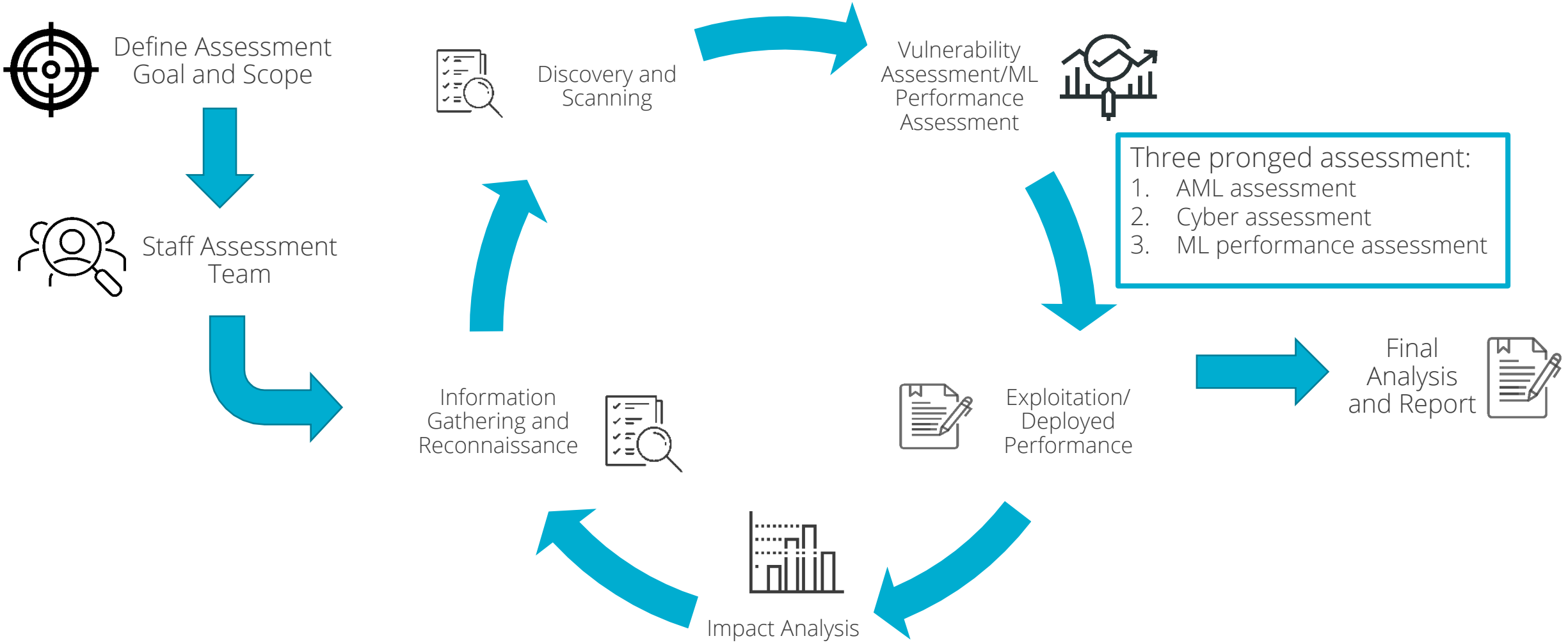


Additional Requirements for ML/AI Assessments

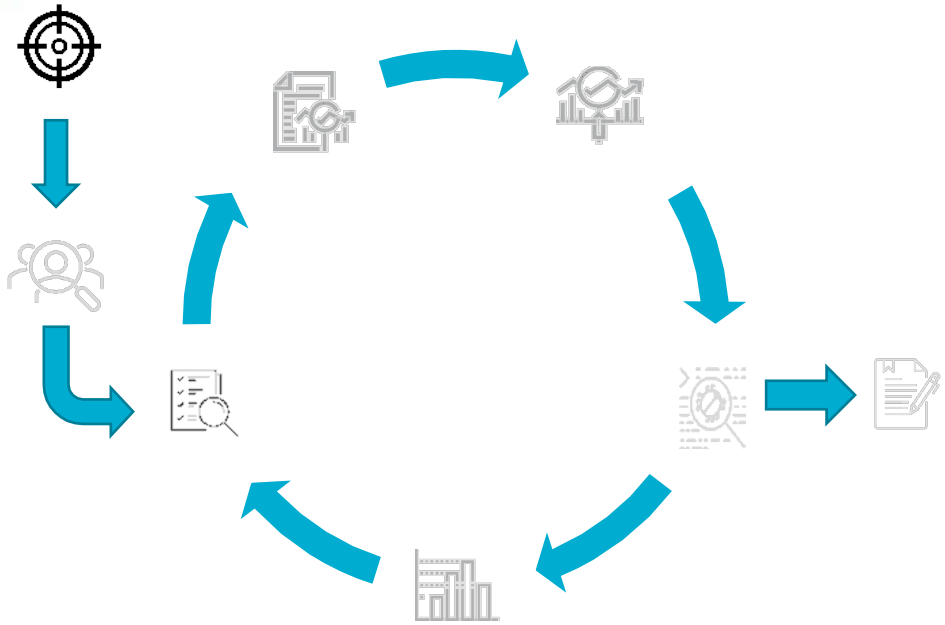
- Each step provides useful information for assessment
- A **data pipeline** for ingesting, preprocessing, and post-processing data
- **Data** may change over time or differ from one deployment to another
- **Models** *may* change over time

A LOT of information is required for proper assessments

Assessment Workflow



Step 1: Define Assessment Goal and Scope



Generally, high-level overview of the system

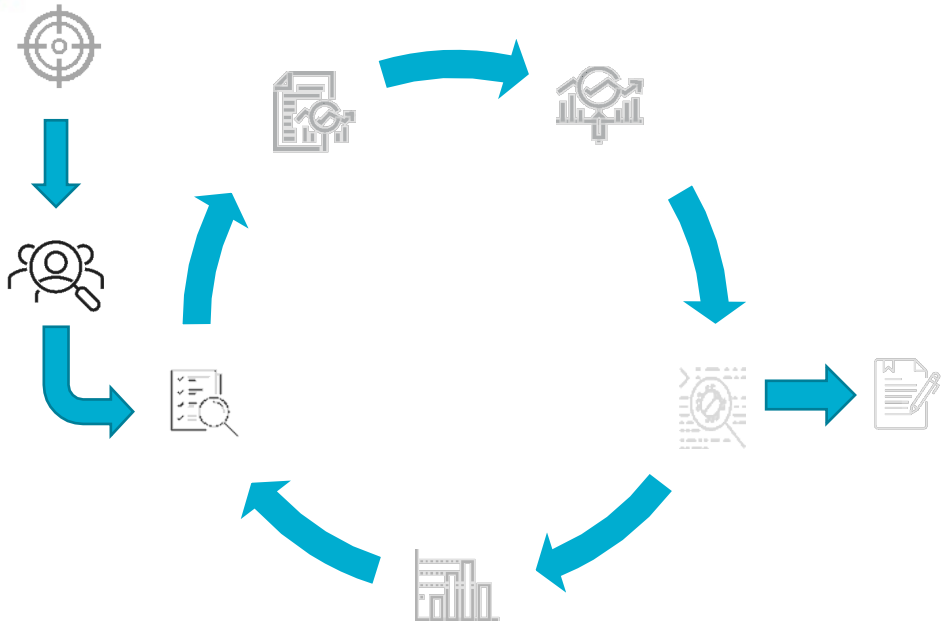
What is the attacker goal?

ML and AML SMEs outline possible scenarios w/ Stak
- What are the practical consequences and risks?
- What are the greatest concerns?
- Rate concerns and consequences

Note risks that exist if certain portions are not provided

Concerns/Risks	Consequences if it occurs	Stakeholder rating
Concern 1:		
Risk 1:		
…		

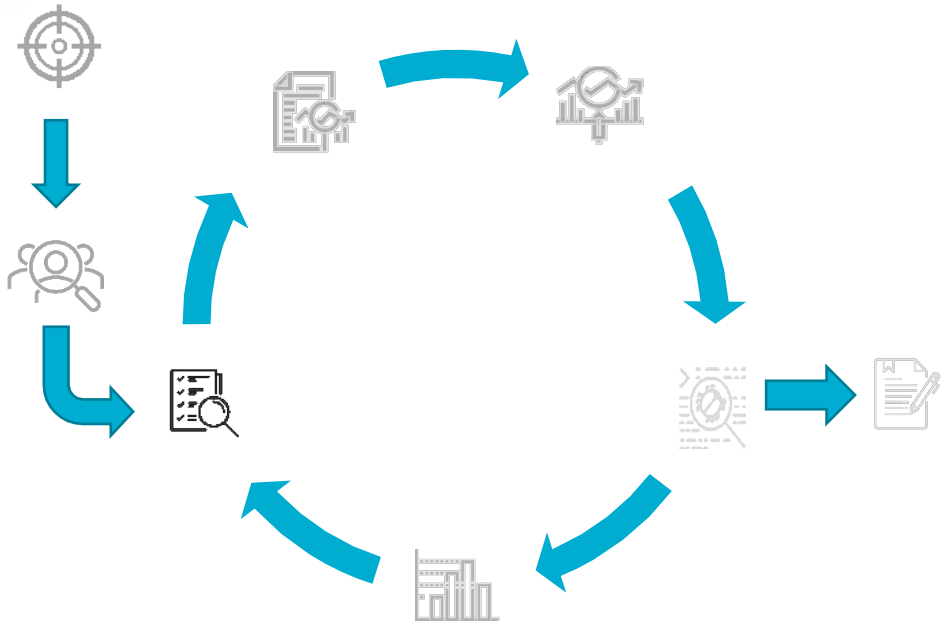
Step 2: Staff Assessment Team



Additional team members could include:

- Independent AML SME
- Independent ML SME
- Domain SME
- ML Developer
- Stake holder

Step 3: Information Gathering and Reconnaissance



Done BEFORE touching the system.

Generally ONLY documentation is provided

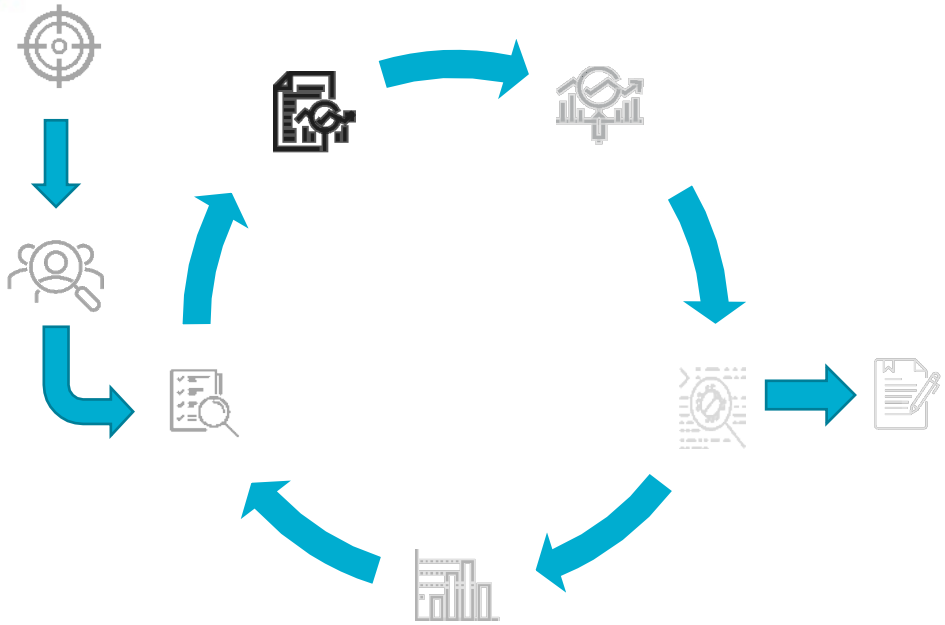
The bulk of ML is often data-centric:
- Feature selection and extraction
- Quality of labels

ML/AML SMEs required dependent on the type of algorithm

Document possible risks/attacks

ML Risks	Consequences if it occurs	ML/CAML rating
Adversarial Risk 1:		
Privacy Risk 1:		
Cyber Risk1:		
…		

Step 4: Discovery and Scanning

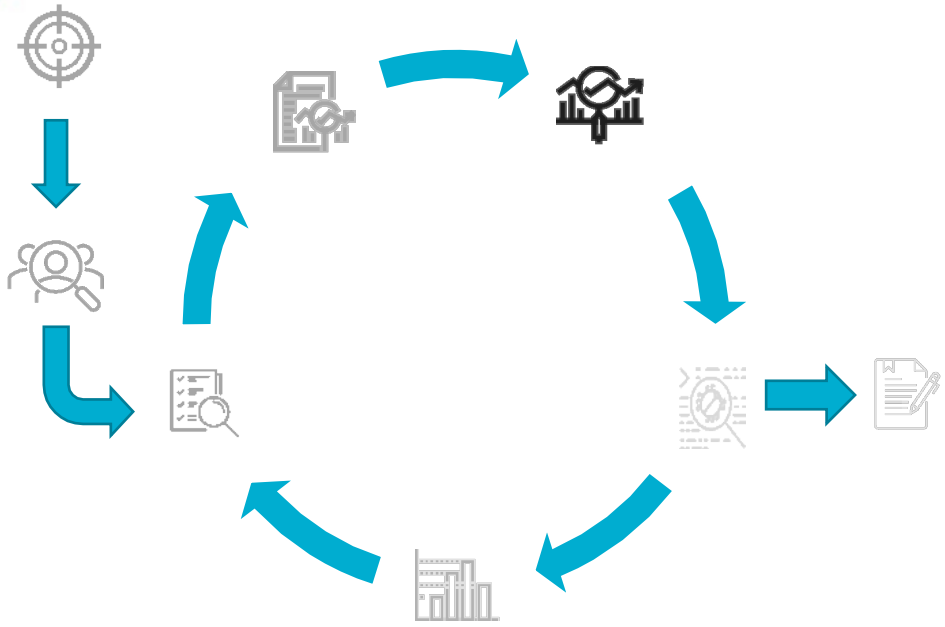


How to interact with system? Where are potential touch points?

- System Recon: What is the supporting infrastructure surround the ML component?
- ML Detect: Is there an actual ML component? Where?
- ML Recon: Where is the data pipeline, configuration files, touch points, etc.?

Possible Tools include:
- Metasploit
- Cobalt Strike

Step 5: Vulnerability Assessment/ML Performance Assessment



AML Assessment

- How can the ML model be subverted?

Cyber Assessment

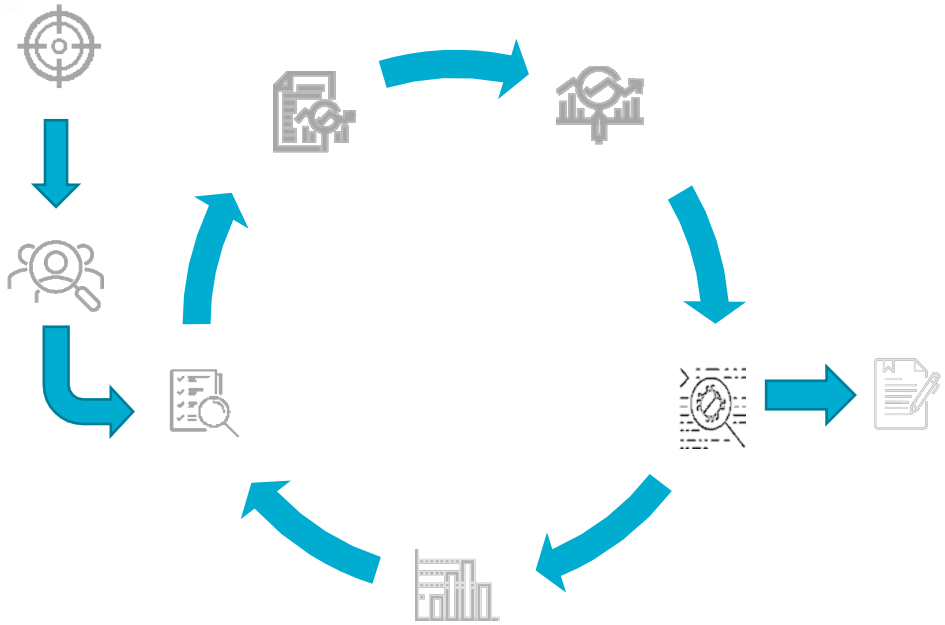
- How can the model be affected through supporting infrastructure?

ML Performance Assessment

- How well does the model perform and what risks exist?

ML Risks	Consequences if it occurs	ML/CAML rating
Adversarial Risk 1:		
Privacy Risk 1:		
Cyber Risk1:		
…		

Step 6: Exploitation/Deployed ML Evaluation



Exploit discovered vulnerabilities

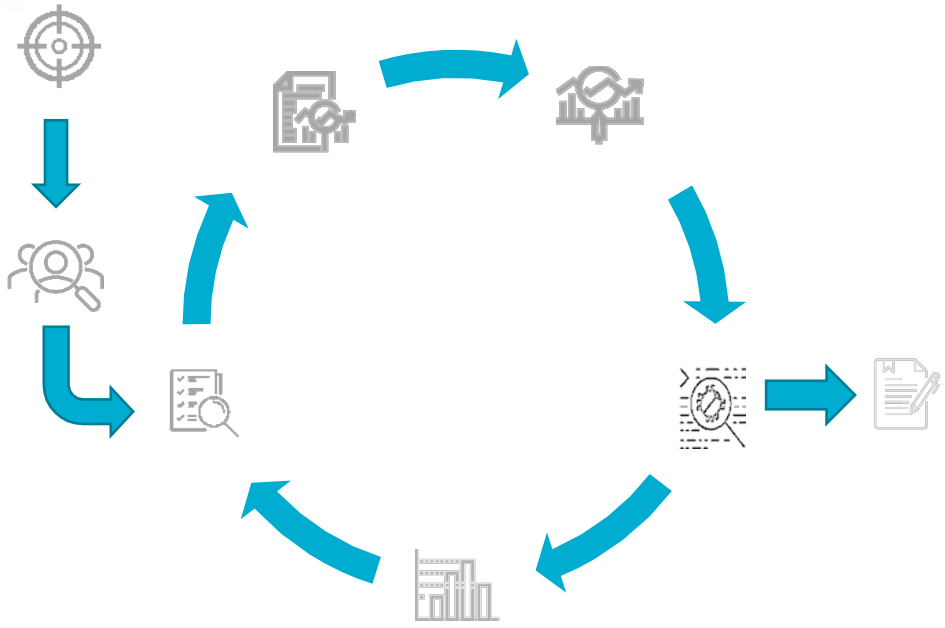
- Document effects
- Document what was **NOT** evaluated
- Speak in terms familiar to customer (Domain SME)

Possible Tools include:

- Offensive toolkits
- AML toolkits (ART)
- Privacy toolkits
- Deepchecks
- Fact Sheets
- ML SME knowledge about deployed scenarios

ML Risks	Consequences if it occurs	ML/CAML rating
Adversarial Risk 1:		
Privacy Risk 1:		
Cyber Risk1:		
…		

Step 6: Exploitation/Deployed ML Evaluation



AML Risks

- Ease of implementation
- Stealth
- Feasibility
- Analytic impact

ML Performance/Risks

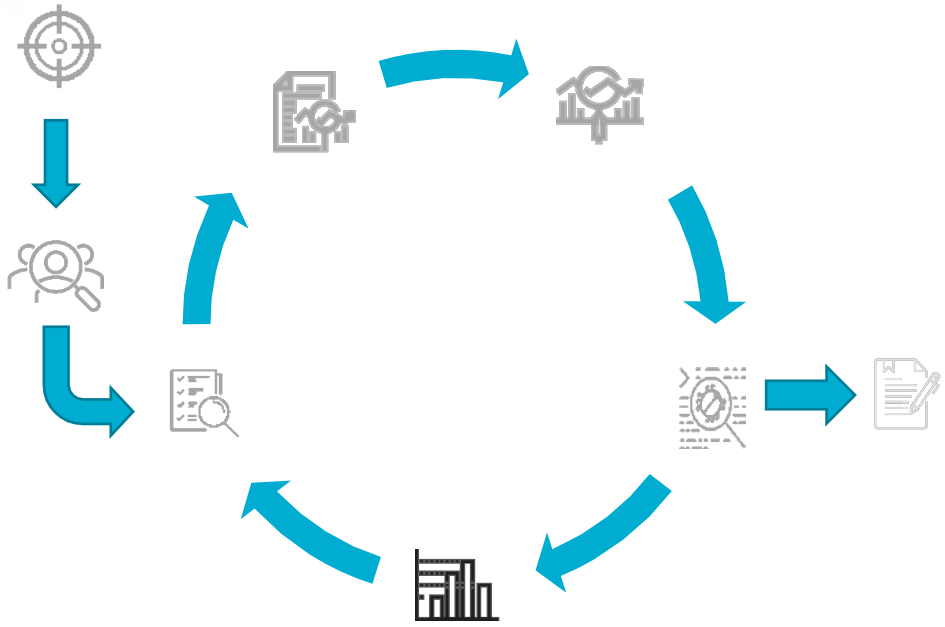
- Access to datasets
- Reproducibility
- Biased evaluation

Cyber Risks:

- Read access to config files
- Vulnerabilities in libraries
- Documented code
- Unsecure datastores

ML Risks	Consequences if it occurs	ML/AML rating
Adversarial Risk 1:		
Privacy Risk 1:		
Cyber Risk1:		
…		

Step 7: Impact Analysis



Meet with Stakeholder

- Map discovered ML risks to stakeholder concerns and risks
- Rank newly discovered ML risks to possibly unforeseen risks and concerns
- Outline remediations

Concerns/Risks	Consequences if it occurs	Stakeholder rating
Concern 1:		
Risk 1:		
…		

ML Risks	Consequences to ML if it occurs	ML/CAML rating	System risk	Impact to system	Impact to application
Adversarial Risk 1:					
Privacy Risk 1:					
Cyber Risk1:					
…					

Wrapping it all up

We need ML…

- Too much data to scale manually
- Keep pace with competitors

…But we need to be able to trust it

- At least account and enumerate for possible risks

https://www.dnv.com/Publications/trustworthy-industrial-ai-systems-164957