Carnegie Mellon University
Software Engineering Institute

Security Questions to Ask Your Data Scientists

INFOSEC WORLD

INFOSECWORLDUSA.COM

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Copyright 2022 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon® and CERT® are registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM22-0771

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Thomas Scanlon

Technical Manager – CERT Data Science

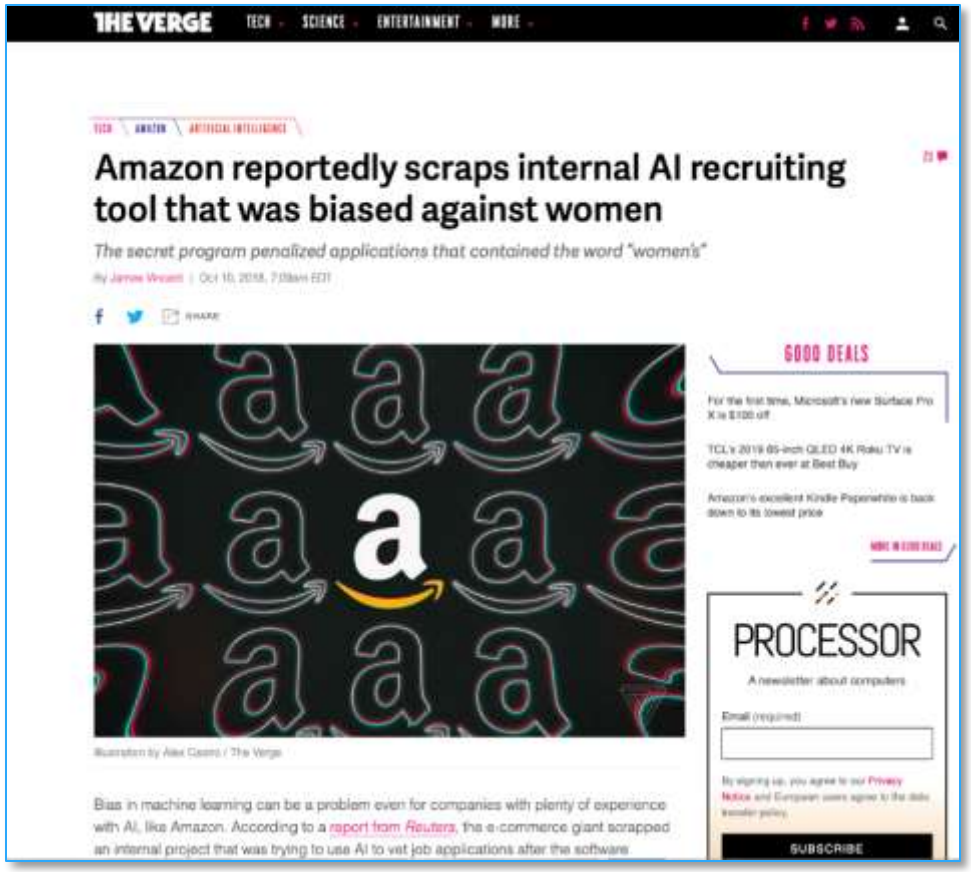
Software Engineering Institute

Carnegie Mellon University

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Why Security Matters – Example 1



https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report

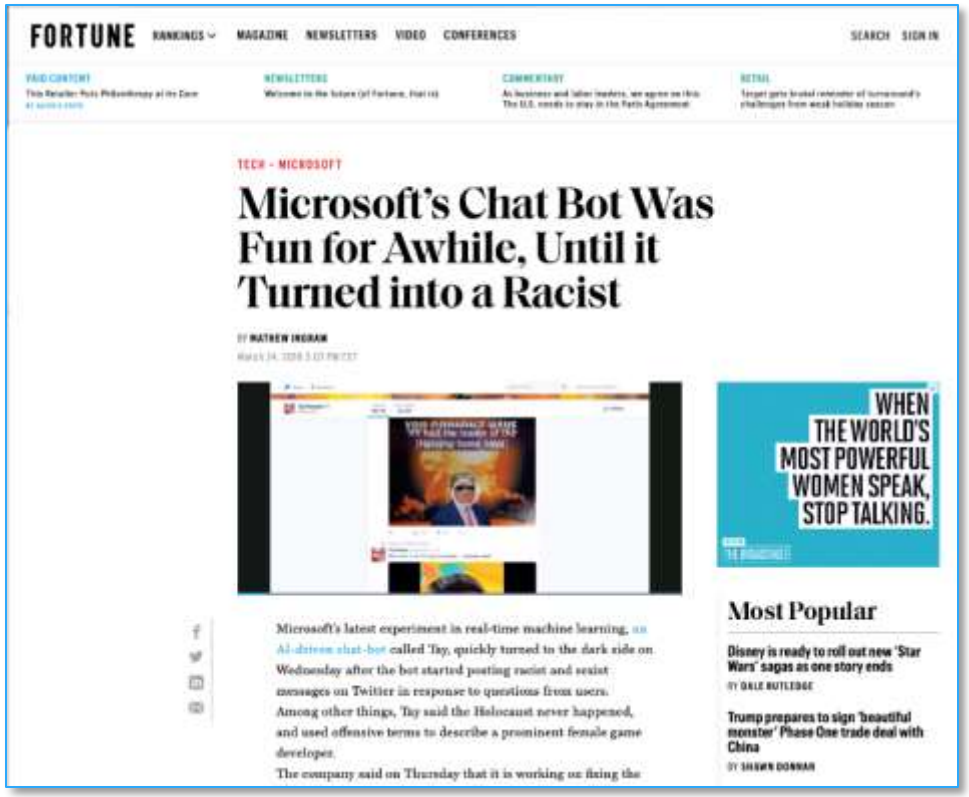
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

14 Distributional shifts

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Why Security Matters – Example 2



FORTUNE RANKINGS MAGAZINE NEWSLETTERS VIDEO CONFERENCES SEARCH SIGN IN

TECH - MICROSOFT

Microsoft's Chat Bot Was Fun for Awhile, Until it Turned into a Racist

BY MATHEW INGRAM

Microsoft's latest experiment in real-time machine learning, an AI-driven chat-bot called Tay, quickly turned to the dark side on Wednesday after the bot started posting racist and sexist messages on Twitter in response to questions from users. Among other things, Tay said the Holocaust never happened, and used offensive terms to describe a prominent female game developer.

The company said on Thursday that it is working on fixing the

Most Popular

Disney is ready to roll out new 'Star Wars' sagas as one story ends
BY DALE BUTLEDGE

Trump prepares to sign 'beautiful monster' Phase One trade deal with China
BY SHAWN DONNAN

https://fortune.com/2016/03/24/chat-bot-racism/

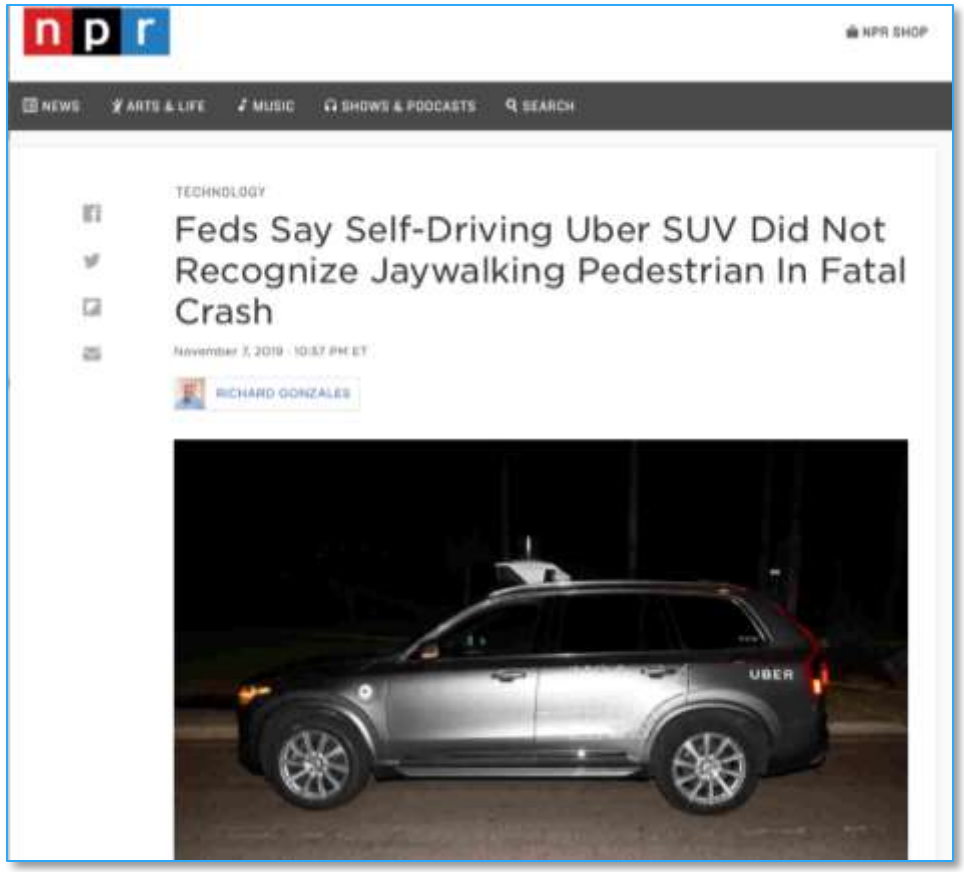
https://en.wikipedia.org/wiki/Tay_(bot)

6	Reprogramming ML system
12	Reward Hacking

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Why Security Matters – Example 3



https://www.npr.org/2019/11/07/777438412/feds-say-self-driving-uber-suv-did-not-recognize-jaywalking-pedestrian-in-fatal-

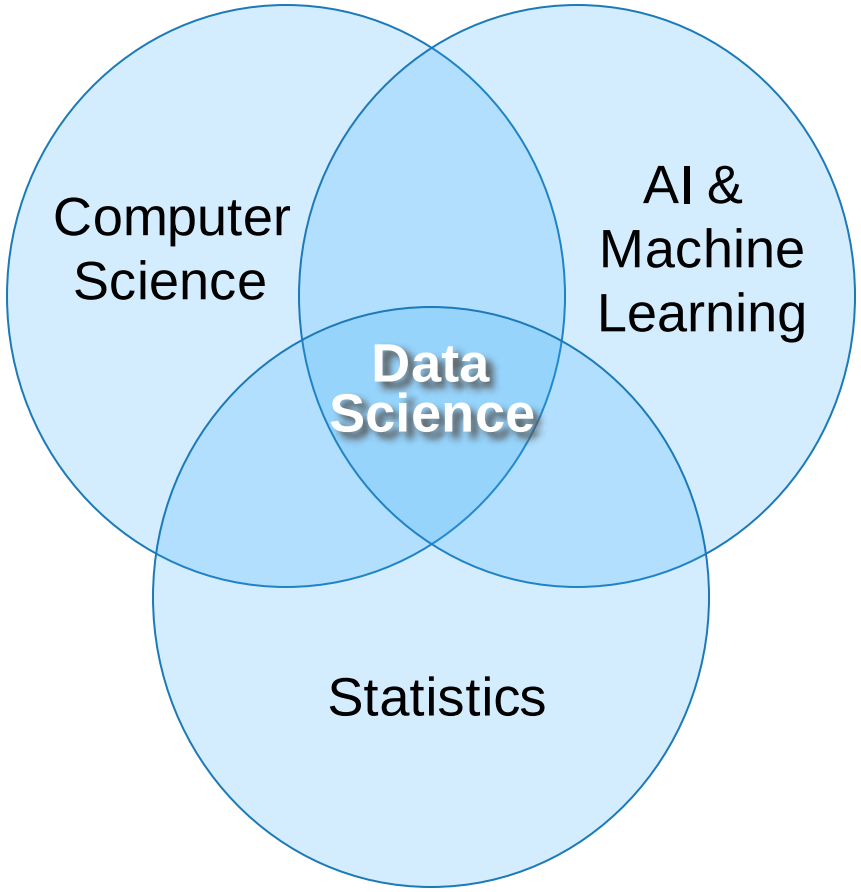
https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf

13	Side Effects
16	Common Corruption
17	Incomplete Testing

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Data Science



Analysis Techniques

- Prediction
- Classification
- Deep Learning
- “Big Data”
- Outlier detection
- Feature extraction

Methods:

- Regression
- Neural nets
- Bayesian networks
- Structural equation modeling
- Latent Dirichlet allocation
- Hidden Markov models
- Gradient boosting
- …

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

AI/ML Security Challenge

- Advances in artificial intelligence and machine learning are enabling organizations to solve problems and add capabilities in exciting new ways
- 86% of company executives say that AI is becoming a "mainstream technology" at their company[1]
- Data scientists typically lack cybersecurity expertise
- Cybersecurity professionals often do not understand the intricacies of an AI/ML system

1: Harvard Business Review https://hbr.org/2021/09/ai-adoption-skyrocketed-over-the-last-18-months

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

AI/ML Development differs from Software Development

Microsoft[1] summarized three key differentiators for AI/ML systems development that separates it from traditional software development:

- the need for complex data versioning in AL/ML systems
- AI/ML model development skills are distinct from software development skills
- AI/ML components are less modular than software components

1: Microsoft: https://doi.org/10.1109/ICSE-SEIP.2019.00042

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

AI/ML Security differs from Software Security

Actually, AI/ML systems generally have all the same security concerns as traditional software systems, plus concerns such as:

- data poisoning
- backdoors in training sets
- adversarial attacks
- model theft
- and more…

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

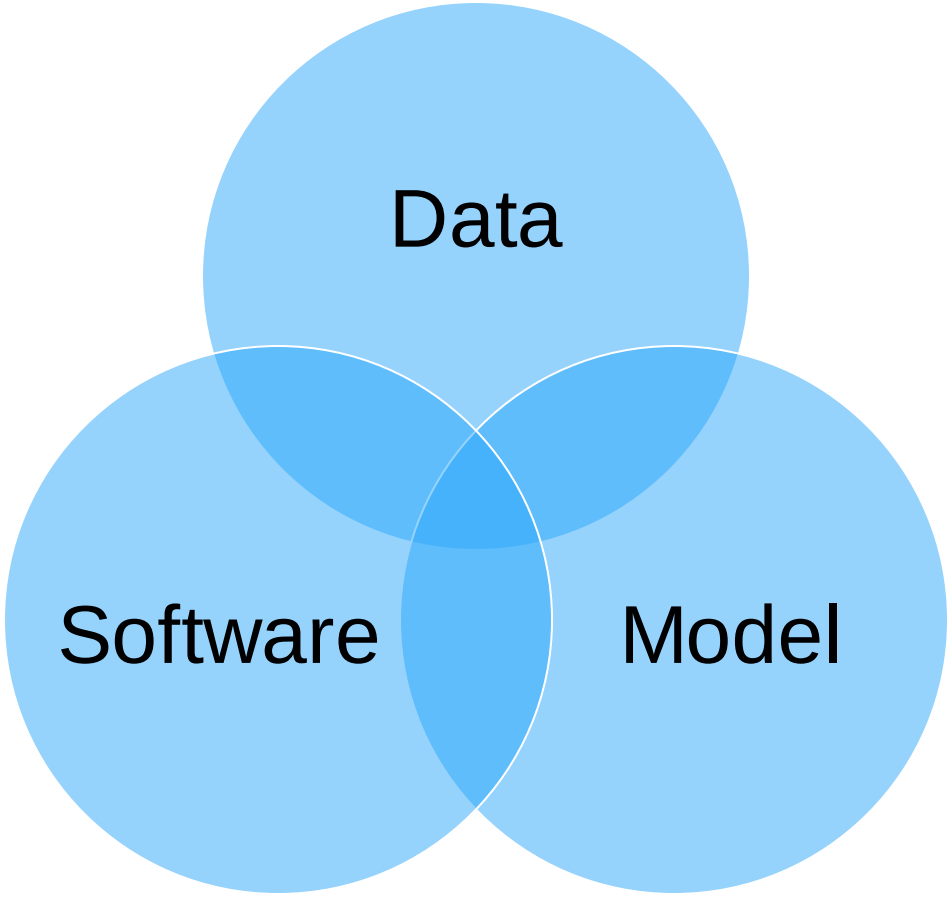
© 2022 Carnegie Mellon University

Bridging the Gap

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

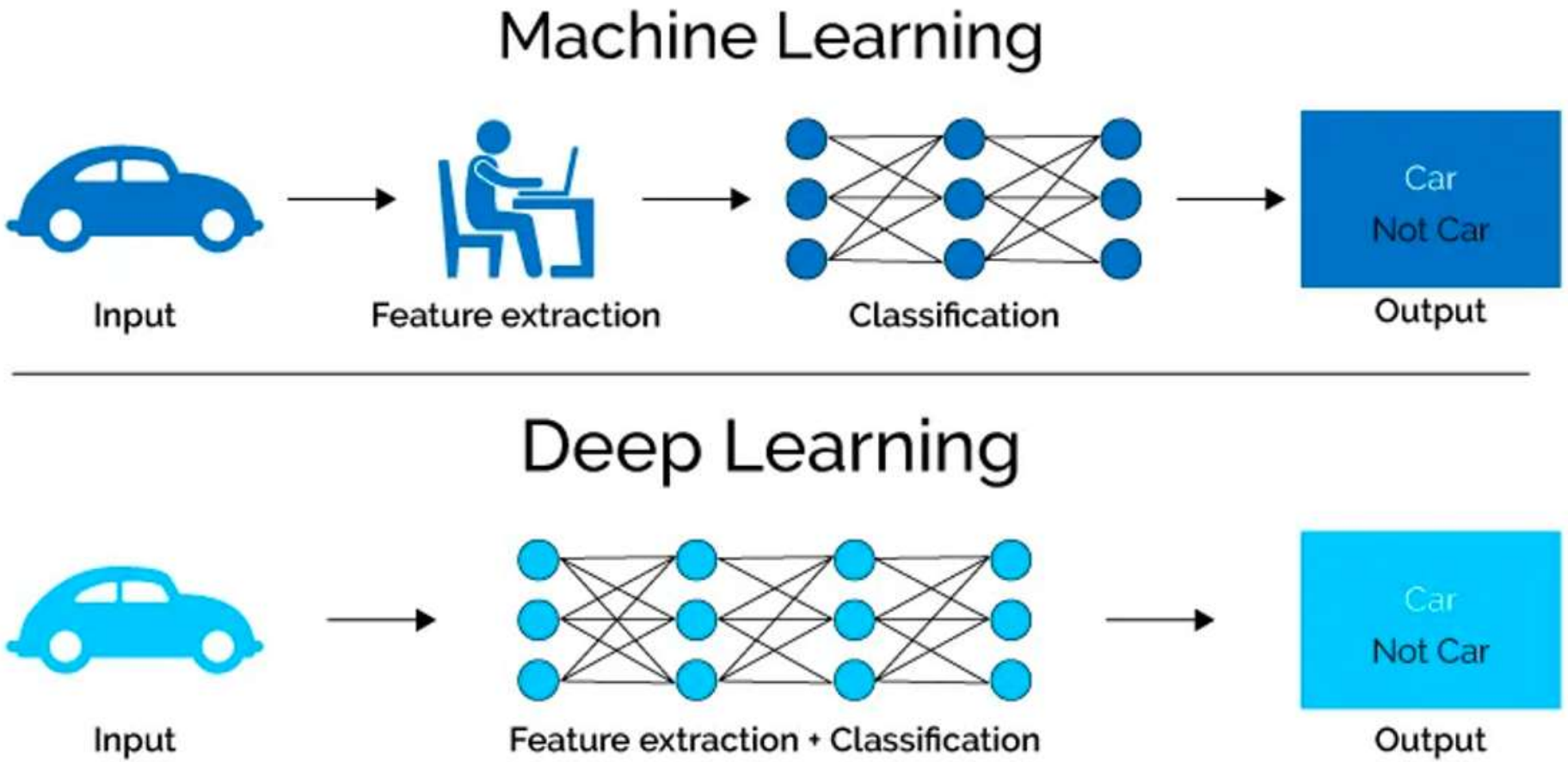
Protecting AI/ML Systems



[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Machine Learning & Deep Learning



Deep learning is machine learning using a neural network.

https://semiengineering.com/deep-learning-spreads/

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

There are Different Types of ML

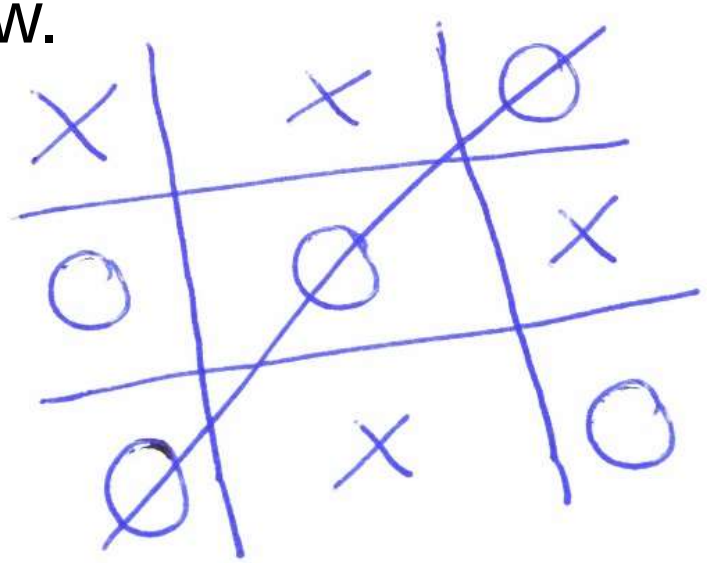
	Supervised Machine Learning	Unsupervised Machine Learning	Reinforcement Learning
Useful for	Making A->B predictions	Discovering previously unknown patterns in data	Optimization in complex, but constrained tasks
Example Uses	Determine whether an image contains a ship. Determine whether a set of financial documents indicate fraud. From a baseball player's prior performance, predict performance in the next game.	Discover customer profiles Identify clusters of malware Identify anomalous network activity	Optimizing logistics chain management Optimizing strategy in a game
Common Methods	**Regression** (Linear, Regression Trees, Kernel Regression, …) **Classification** (Support Vector Machines, Logistic Regression, Discriminant Analysis) Neural Networks, Ensemble Methods…	**Clustering** (K-means, DBSCAN, Mixture modeling) Association Rule Mining Anomaly Detection Neural Networks	Q-Learning Policy Optimization State-Action-Reward-State-Action Deep Deterministic Policy Gradient
Notes	By far the most common	Data widely available, implementation and verification are tricky	Only beginning to move into commercial space, still largely academic.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

If You use the Wrong Math, You can get Pure Nonsense

- An algorithm that always takes the “best” move on the next turn will never lose at ***tic-tac-toe.*** It will either win or draw.

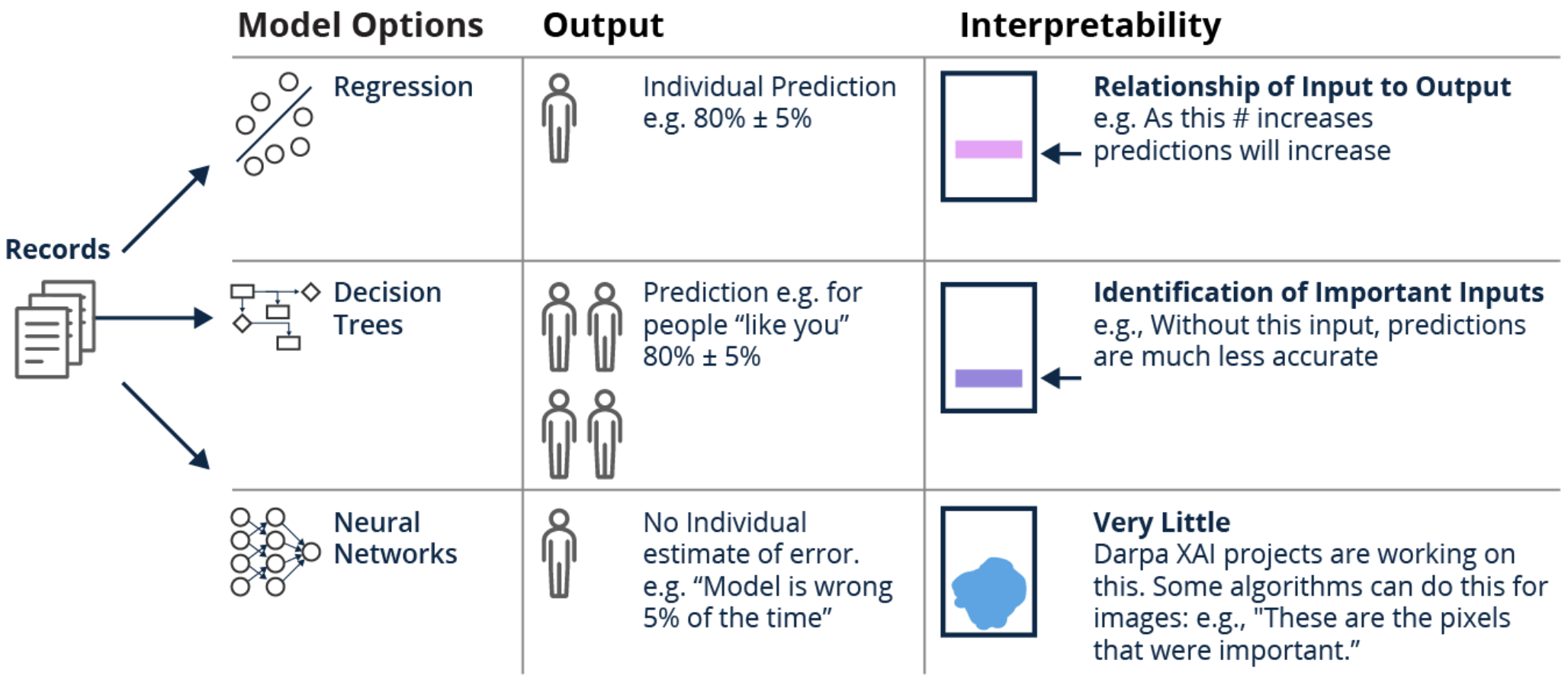


- An algorithm that always takes the next “best” move in ***chess***, will always fall into any trap set by its opponent. In order to see the trap and avoid it, the algorithm must be able to consider more than one move ahea

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

There are Usually Several Options of "Right Math," but Only Some of Them will give you the Summary You Need



[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Lots of Performance Metrics can seem Similar, Even if They're Quite Different

Different kinds of errors may have different kinds of implications.

Metric	What it measures
Error Rate	How often is the algorithm wrong?
False Positive	Algorithm predicted "yes," But the truth is "no"
False Negative	Algorithm predicted "no," But the truth is "yes"
Positive Predictive Value	Of the "yes's" that were predicted, How many are actually "yes"

The metrics you choose for an ML project are a policy statement about what kind of systematic errors are acceptable, and which should be minimized.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Checklist to Identify Good Candidates for ML Projects

- Can you state your problem as either:
  - I would like to use ___ data to predict ___.
  - I would like to understand the structure of the features recorded in ___ data.
  - I would like to optimize ____ well defined process.

- Is it a large scale problem?

- Have you already done exploratory analysis on available data?

- Have you considered the broader context?

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

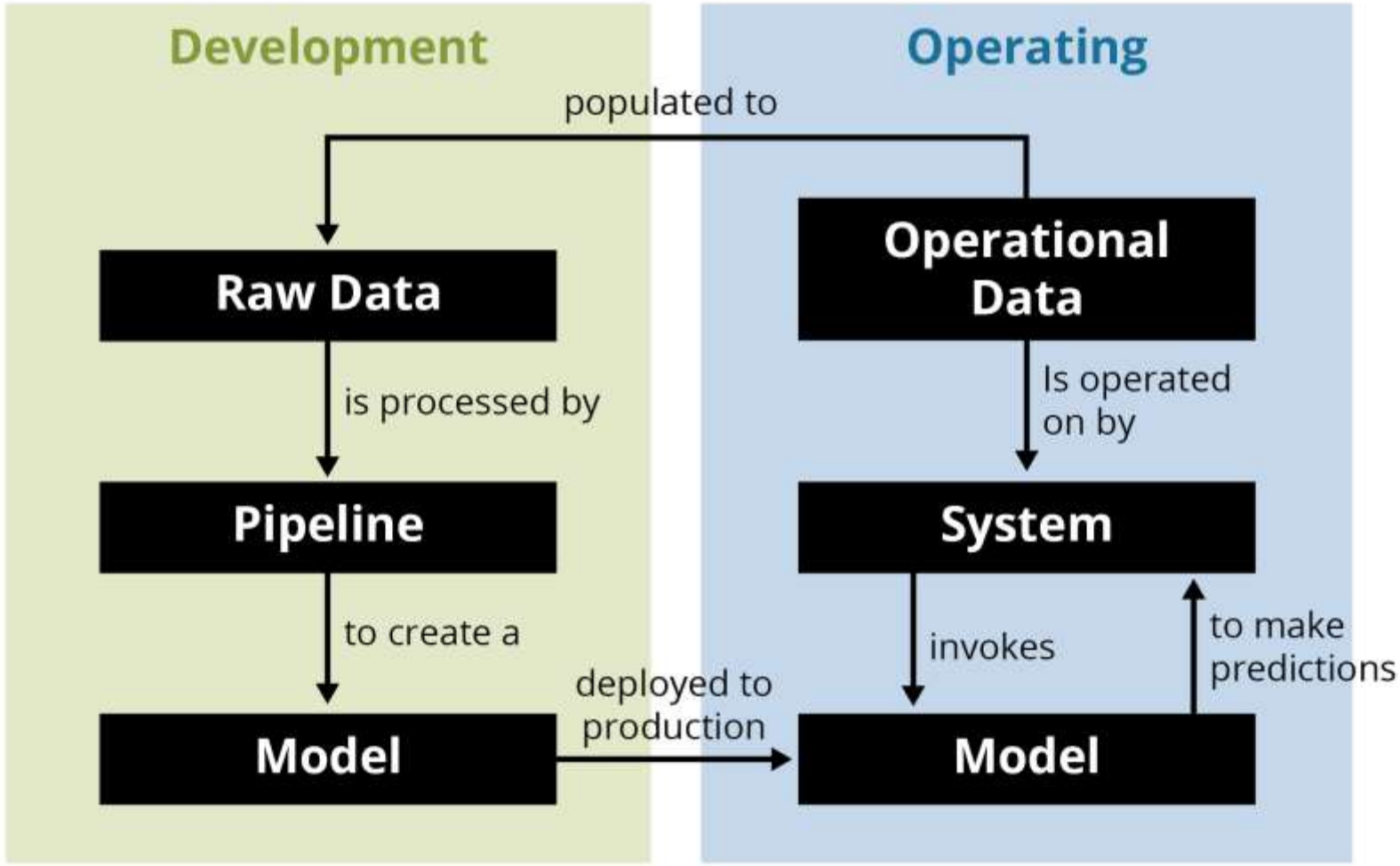
MLOps

MLOps is the process of taking a Machine Learning model from experimental prototype into a production system

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

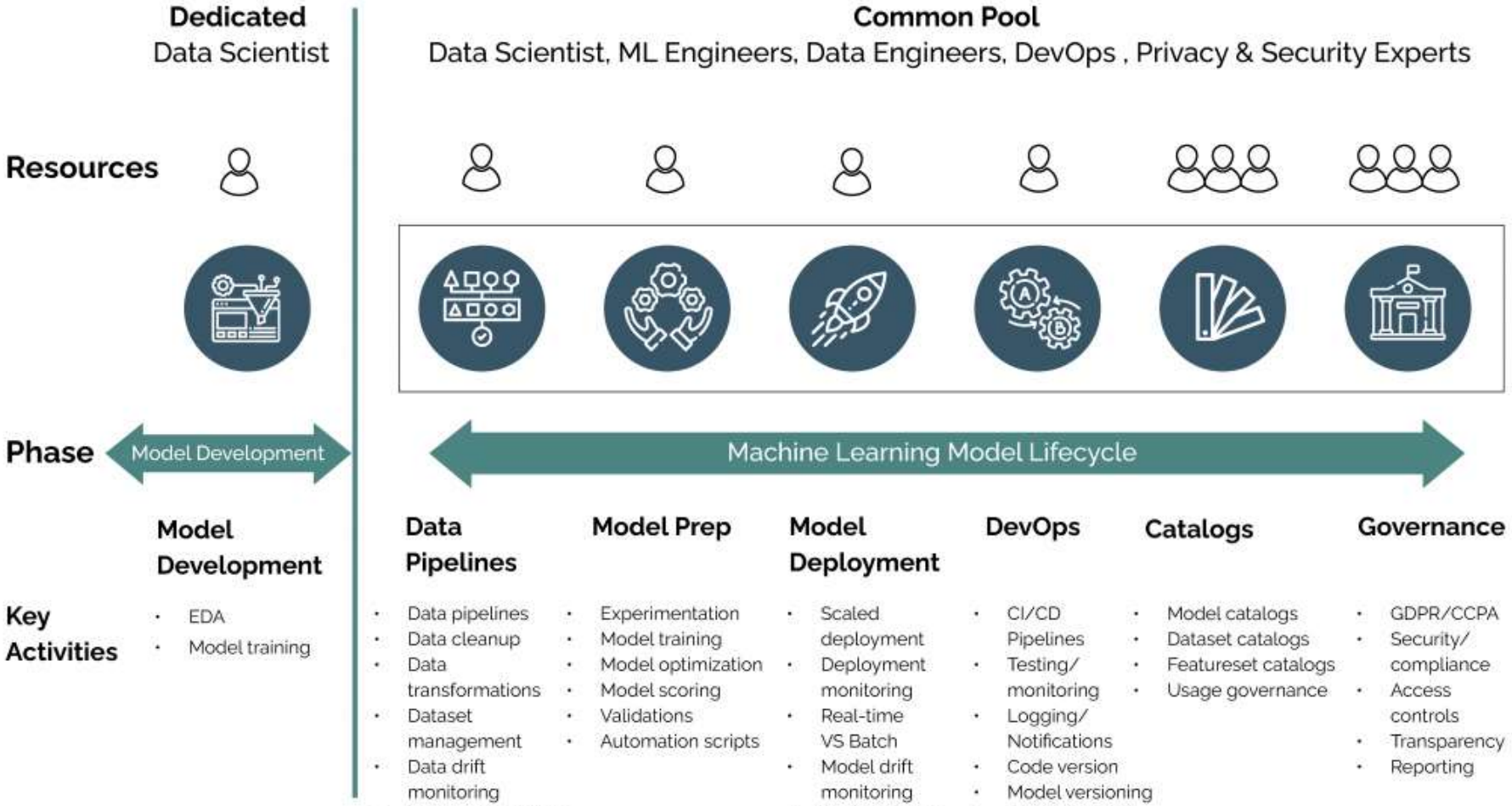
Notional ML Pipeline



[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Machine Learning Operations (MLOps)

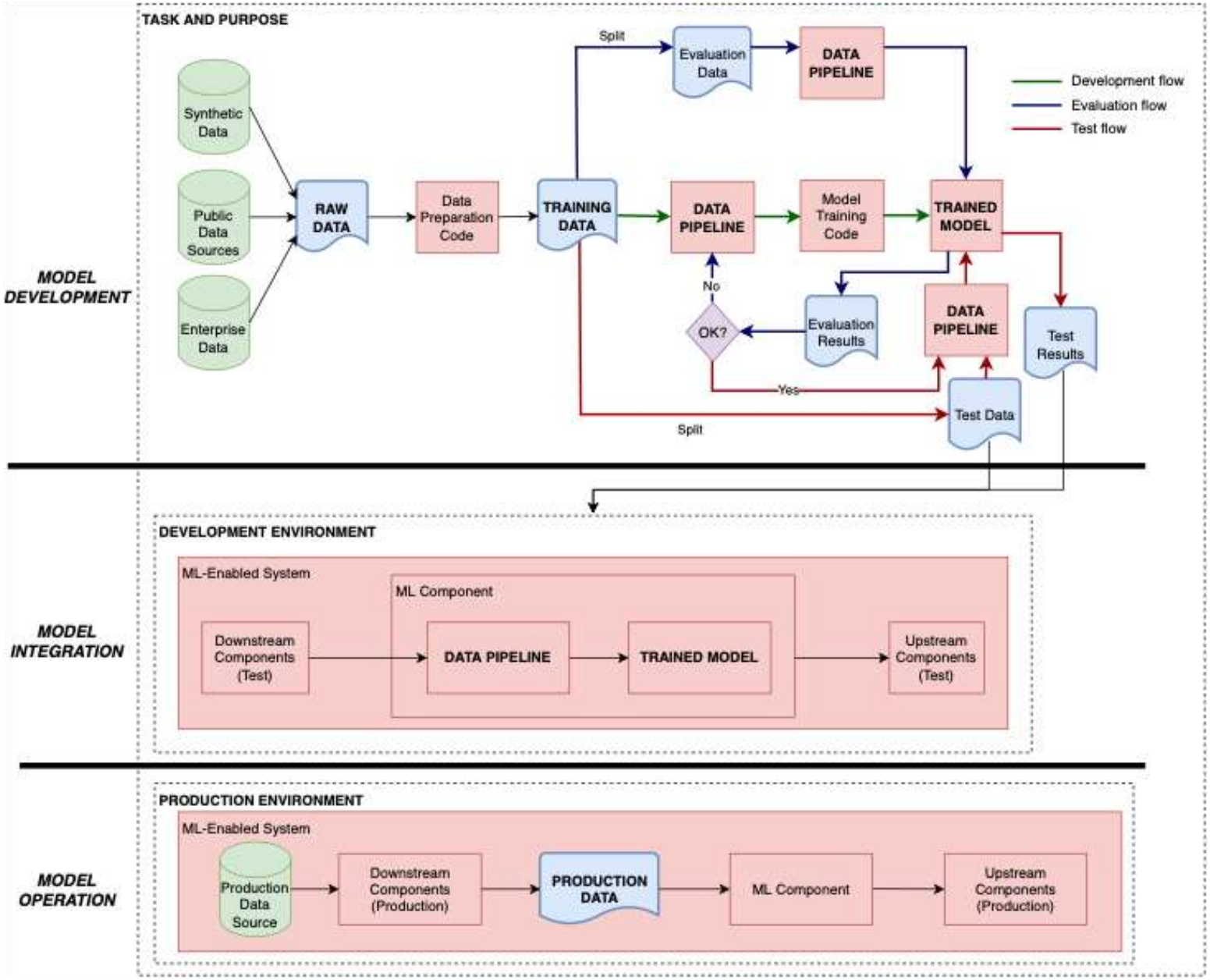


https://gurukannan.medium.com/overview-of-mlops-ml-dev-ops-2899ecb97820

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

MLOps

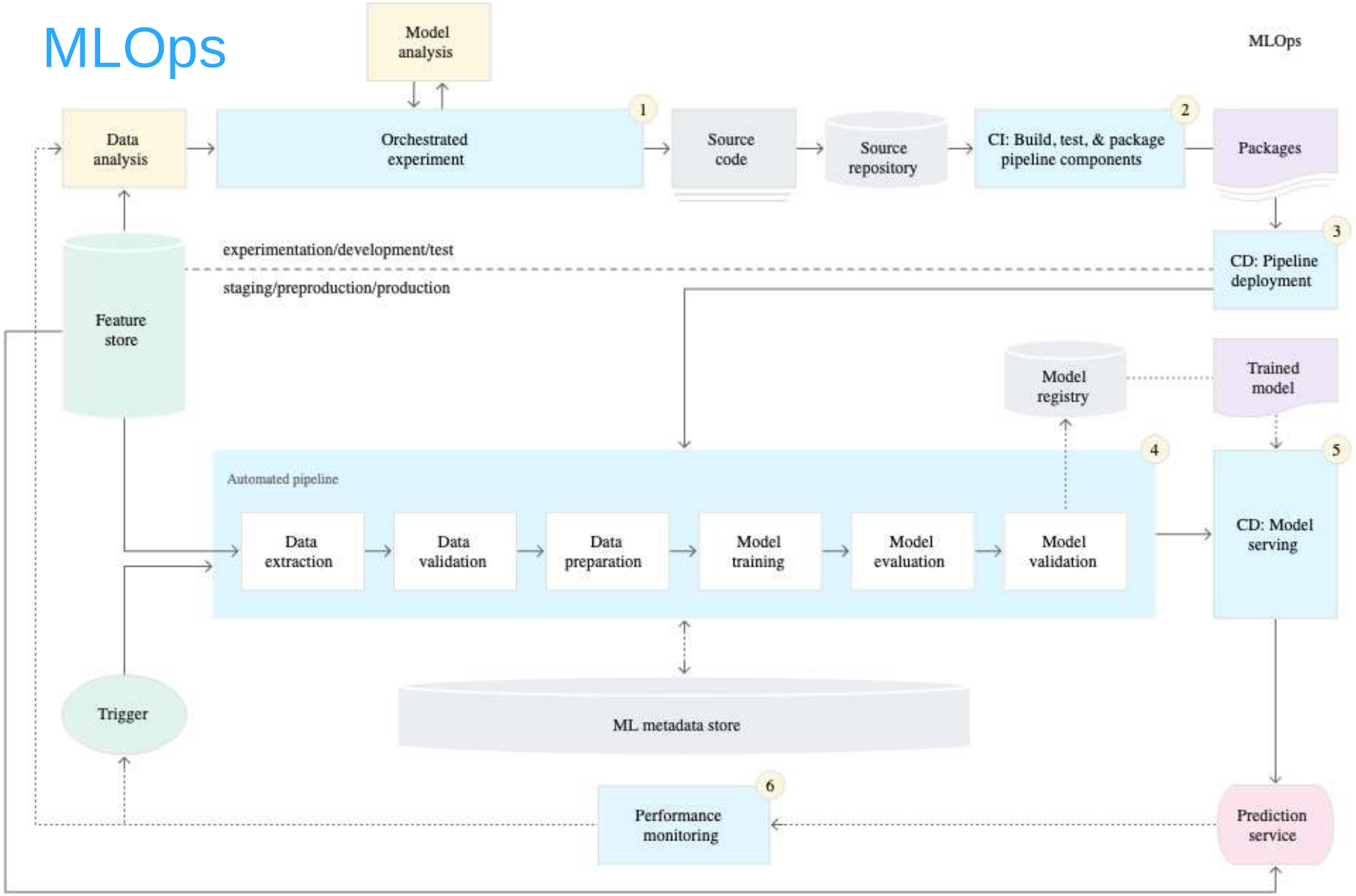


Nahar, N.; Zhou, S.; Lewis, G.; & Kästner, C. More arXiv:2110.10234. 2021

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

MLOps



Google: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

AI/ML Security Threats

AI/ML model is compromised to :

- leak data
- deliver inaccurate predictions
- not notice malicious activity
- reveal proprietary or sensitive information
- degrade performance
- denial of service

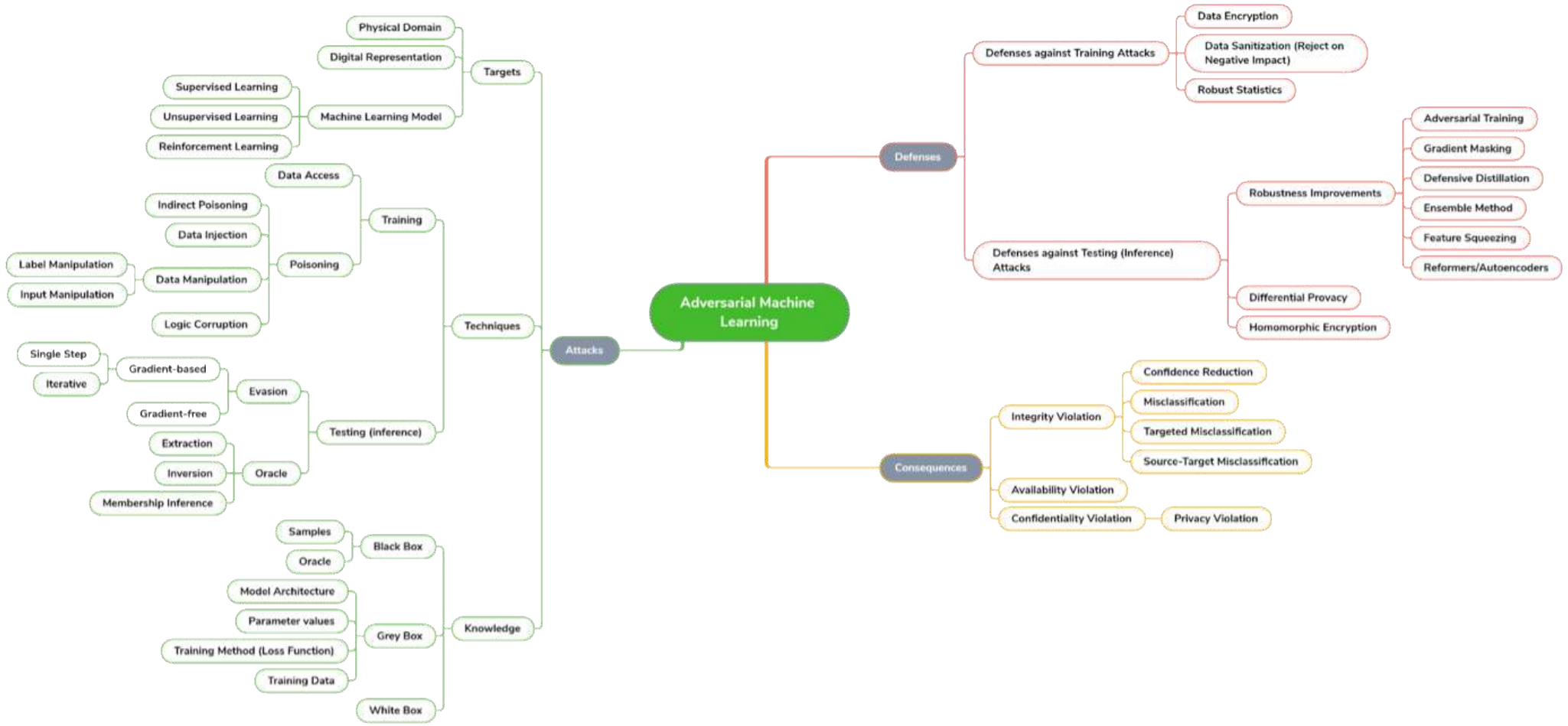
A couple differences from traditional cyber attacks:

- exfiltration is not often an objective
- public access to data that ultimately feeds models

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Adversarial Machine Learning



Tabassi, Elham; Burns, Kevin; Hadji, Michael; Molina-Markham, Andres; & Sexton, Julian. https://doi.org/10.6028/NIST.IR.8269-draft.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Failure Modes Table

Scenario Number	Failure	Overview
1	Perturbation attack	Attacker modifies the query to get appropriate response
2	Poisoning attack	Attacker contaminates the training phase of ML systems to get intended result
3	Model Inversion	Attacker recovers the secret features used in the model by through careful queries
4	Membership Inference	Attacker can infer if a given data record was part of the model's training dataset or not
5	Model Stealing	Attacker is able to recover the model through carefully-crafted queries
6	Reprogramming ML system	Repurpose the ML system to perform an activity it was not programmed for
7	Adversarial Example in Physical Domain	Attacker brings adversarial examples into physical domain to subvertML system e.g
8	Malicious ML provider recovering training data	Malicious ML provider can query the model used by customer and recover customer's training data
9	Attacking the ML supply chain	Attacker compromises the ML models as it is being downloaded for use
10	Backdoor ML	Malicious ML provider backdoors algorithm to activate with a specific trigger
11	Exploit Software Dependencies	Attacker uses traditional software exploits like buffer overflow to confuse/control ML systems
12	Reward Hacking	Reinforcement Learning (RL) systems act in unintended ways because of mismatch between stated reward and true reward
13	Side Effects	RL system disrupts the environment as it tries to attain its goal
14	Distributional shifts	The system is tested in one kind of environment, but is unable to adapt to changes in other kinds of environment
15	Natural Adversarial Examples	Without attacker perturbations, the ML system fails owing to hard negative mining
16	Common Corruption	The system is not able to handle common corruptions and perturbations such as tilting, zooming, or noisy images.
17	Incomplete Testing	The ML system is not tested in the realistic conditions that it is meant to operate in

https://docs.microsoft.com/en-us/security/failure-modes-in-machine-learning

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Data Attacks

Data Poisoning

An adversary pollutes the training data for an ML model, which influences the predictions made by that model

Label Flipping

Label Flipping is a special case of data poisoning, where the adversary is restricted to changing the model's training labels, such that the model is trained on corrupted data.

Example: Poisoning an ML-based email spam filter into classifying malicious emails as legitimate

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Questions to Ask Your Data Scientists - Data

- Are public data sets/data sources used in this project?
- Who is responsible for sanitizing and validating public data sets/data sources?
- How will data be versioned?
- Will data versioning include associating data sets with model training instances?
- What type of data/data source validity checks will be utilized?
- What type of anomaly detection & outlier detection will be used to detect suspicious data points ?
- How will data be protected at rest?
- Will data encryption be used?
- Will homomorphic encryption be used?
- Will any type of label sanitization be employed?
- Does training data contain sensitive/restricted/classified data?

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Model Attacks

Extraction Attacks: The adversary extracts the parameters or structure of the model from observations of the model's predictions, typically including probabilities re-turned for each class.

Inversion Attacks: The inferred characteristics may allow the adversary to recon-struct data used to train the model, including personal information that violates the privacy of an individual.

Membership Inference Attack: The adversary uses returns from queries of the tar-get model to determine whether specific data points belong to the same distribution as the training dataset, by exploiting differences in the model's confidence on points that were or were not seen during training.

Example: using knowledge of a model extracted from observing its predictions, an AI/ML object detection system is perturbed to not notice certain objects

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Model Attacks cont.

Transfer Learning

Using model weights trained for a particular task or set of inputs, removing the first and last layers, and training new input and output layers while holding the middle layers fixed; Purpose to encourage model re-use

Transfer Learning Attacks

Weight poisoning is used to alter neural network weighting for the purpose of misclassification or performance degradation

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Model Attacks cont.

Gradient Descent

A gradient is a vector that points in the direction of the greatest change for some function. Gradient descent is an iterative algorithm that seeks out the local minimum for differentiable functions by taking steps in the opposite direction.

Gradient Descent Attacks

Perturb the input of a model by replacing it with the gradient of some loss function that depends on the input or by adding the sign of the gradient to the input. Next, the magnitude of the gradient of this loss or the sign of the gradient will continually increase until an input is labeled incorrectly.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Questions to Ask Your Data Scientists - Model

- Have pre-trained models been utilized for this project? If so, how were they vetted?
- Is there a plan for protecting the model?
- Is there a 'ground truth' for this model?
- Is there 'correct' outputs for given inputs
- Can outputs be predicted by inputs?
- How robust* is the model?
- Has the model been tested against varying inputs?
- What type of input validity checks will be utilized?

*A robust model is one that maintains predictive power even as the sets of model inputs and assumptions change.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

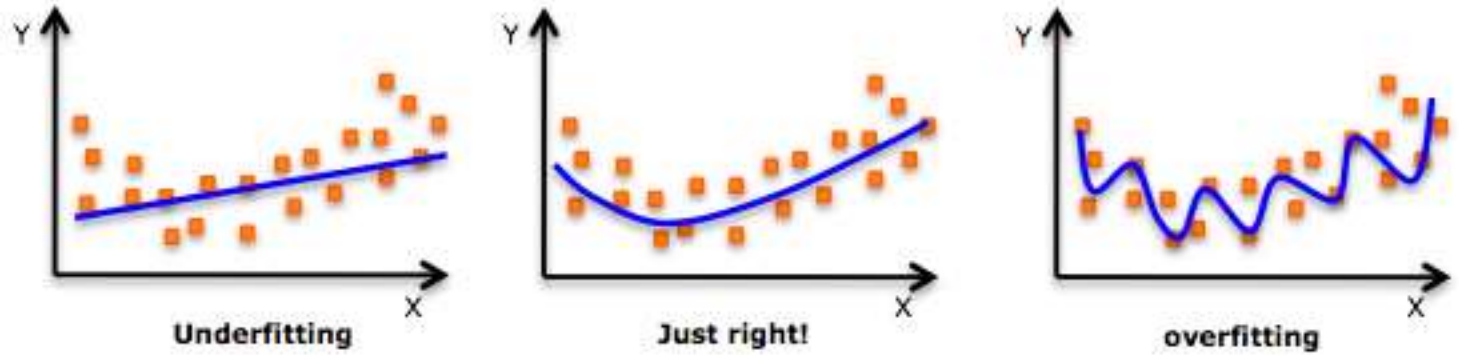
Questions to Ask Your Data Scientists – Model cont.

- Have you designed intentional perturbations that trick the model in training?
- Will the model be trained against adversarial ML attacks?
- Has input regularization been employed?
- How will models be versioned?
- What type of model monitoring will be deployed in Production?
- How will it be determined that the model needs re-trained?
- How will underfitting and overfitting be detected?
- How will model drift be detected?

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Underfitting and Overfitting



[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Model Drift

"Model drift refers to the degradation of model performance due to changes in data and relationships between input and output variables. It is relatively common for model drift to impact an organization negatively over time or sometimes suddenly." - IBM

https://www.ibm.com/cloud/watson-studio/drift#:~:text=Model%20drift%20refers%20to%20the,over%20time%20or%20sometimes%20suddenly.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

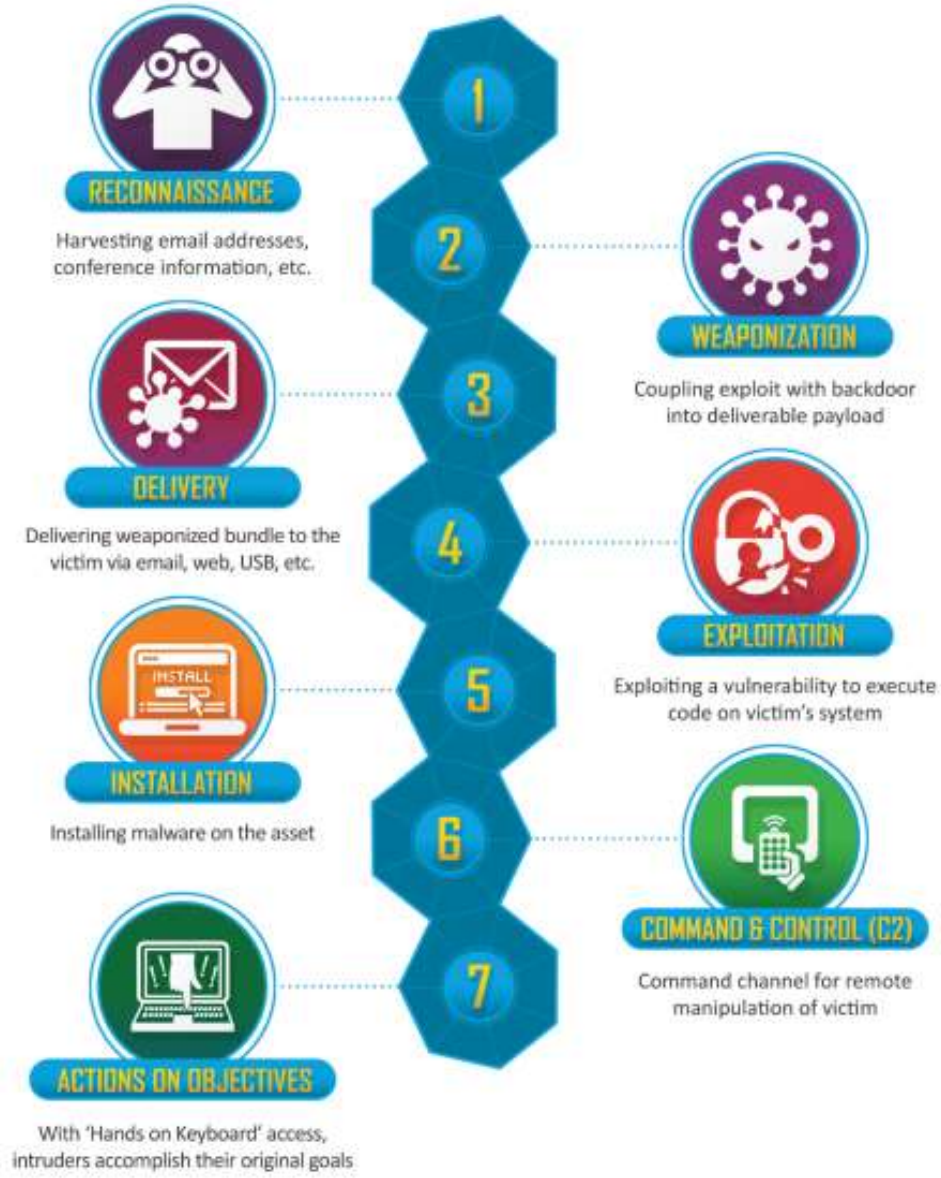
Software Attacks

- Software is used to build AI/ML models
- AI/ML models are deployed into larger software systems and/or effect the operation of software systems
- Cyber threats that apply to traditional software systems will apply to AI/ML enabled systems
- Attacks on AI/ML cyber systems will often involve traditional cyber attack vectors in conjunction with an AI/ML attack

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Cyber Kill Chai



Lockheed Martin: https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Questions to Ask Your Data Scientists - Software

- How was the software used in model development selected?
- Were open source components used in model development?
- How were open source components verified for appropriateness?
- Does the software used in model development have any known CVE?
- Who is responsible for monitoring vulnerabilities in 3rd party software used in model development?
- Who will apply patches to software used in model development?

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Questions to Ask Your Data Scientists – Software cont.

- Was software used in model development examined by code analysis tools?
- Can an AI/ML component pass a vulnerability scan now but not in the future due to environmental or context changes?
- Were custom scripts written for model development examined by code analysis tools?
- Were custom scripts written for tested for misuse and abuse cases?
- Is there configuration management in place for software related to model development and deployment?
- Have permissions and ACLs for all model development software been reviewed with a least privilege mindset?

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Useful Questions to Begin an Oversight Discussion

- What policy is this algorithm implementing?
  - What are the intended consequences of a policy?
  - What unintended consequences can be anticipated?
- What checks and balances are in place?
  - How will field performance be evaluated?
  - What is the procedure for monitoring and validation? Who will be doing the monitoring?
  - Are there historic problems (e.g., racial bias) in this area that could be perpetuated?
- What procedures are in place for handling inherent uncertainty?
  - How is uncertainty communicated to the end user?
  - How can the end user check or verify a prediction? (e.g., If you're uncertain about a rain forecast, you might look at a radar map.)
  - How does the end user make a decision when told a prediction has low confidence? (e.g., the ML system only has 60% confidence in its prediction.)

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Please complete the session evaluation in the event mobile app.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University

Thank you!

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

© 2022 Carnegie Mellon University