1

Governance for Artificial Intelligence/ Machine Learning

Akbar Siddiqui
Technical Director
Civil Liberties, Privacy, and Transparency Office
National Security Agency

2

The NSA Mission

The National Security Agency is responsible for:

Signals Intelligence

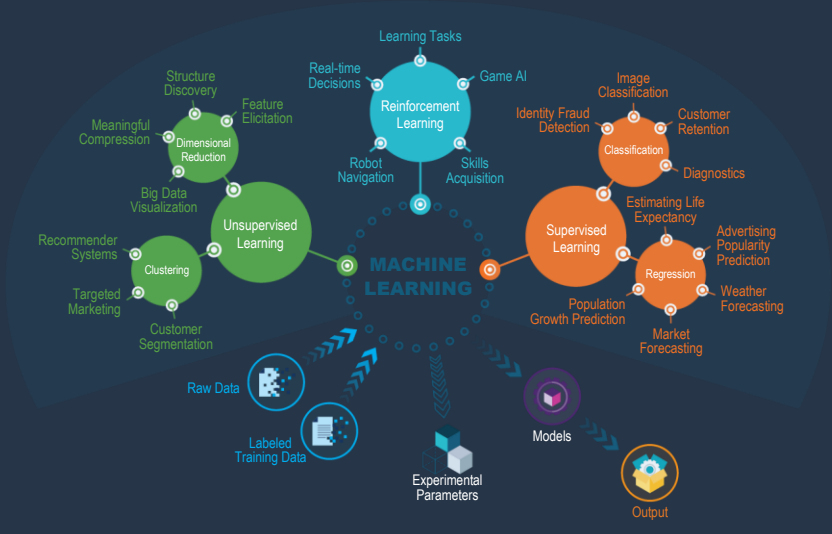
Providing our nation's policy makers and military commands with foreign intelligence to gain a decisive advantage.

U.S. Cybersecurity

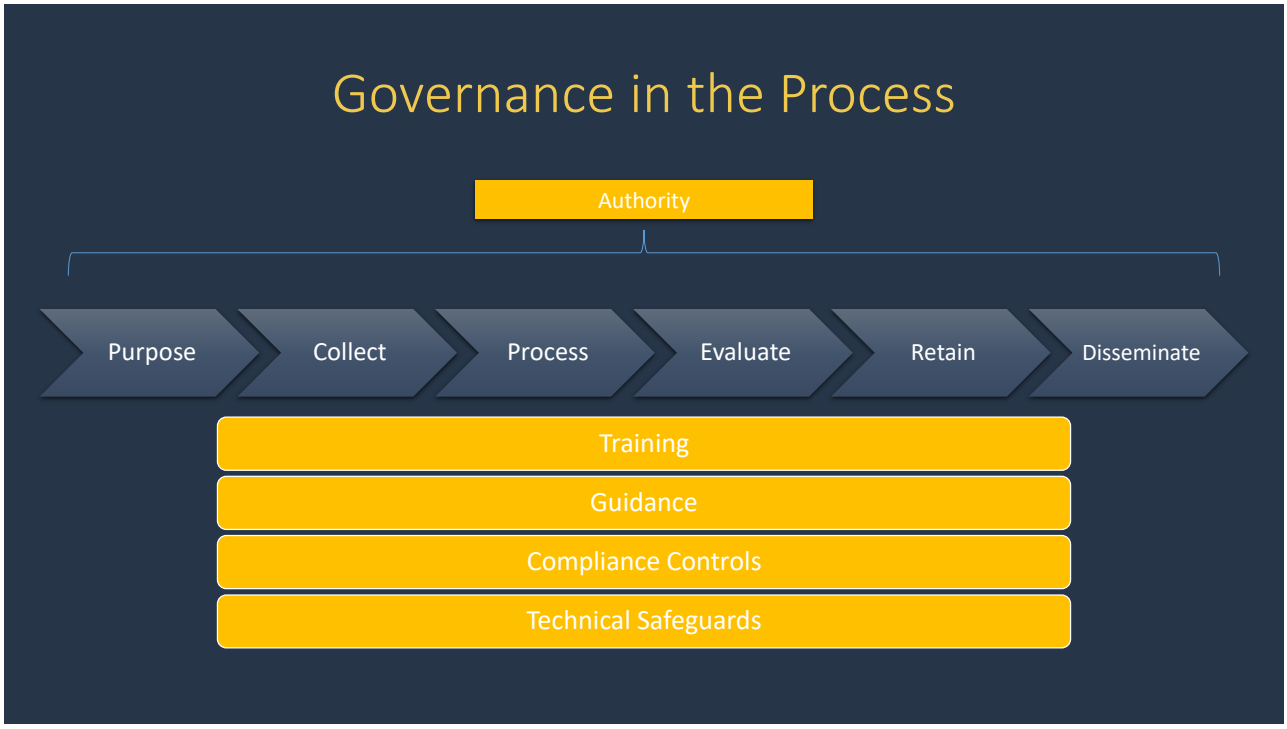
Protecting and defending sensitive information systems and networks critical to national security and infrastructure.

3

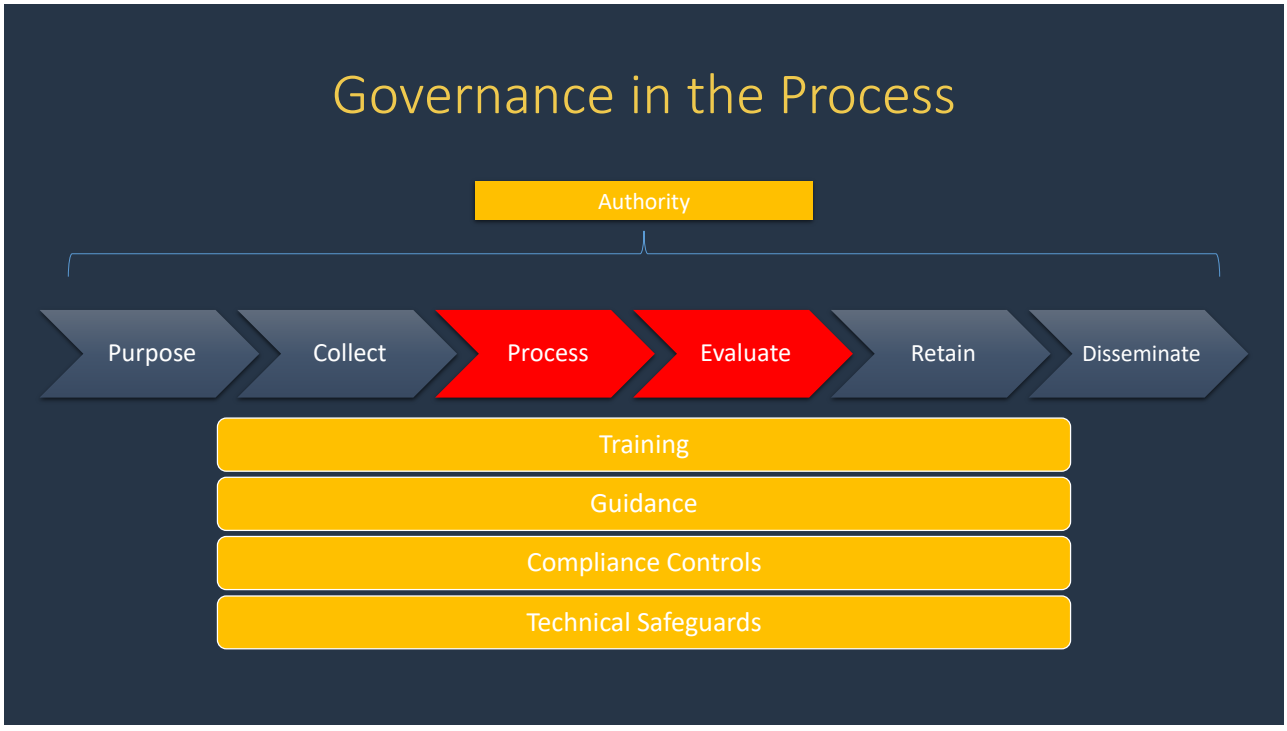
What is AI/ML?



4

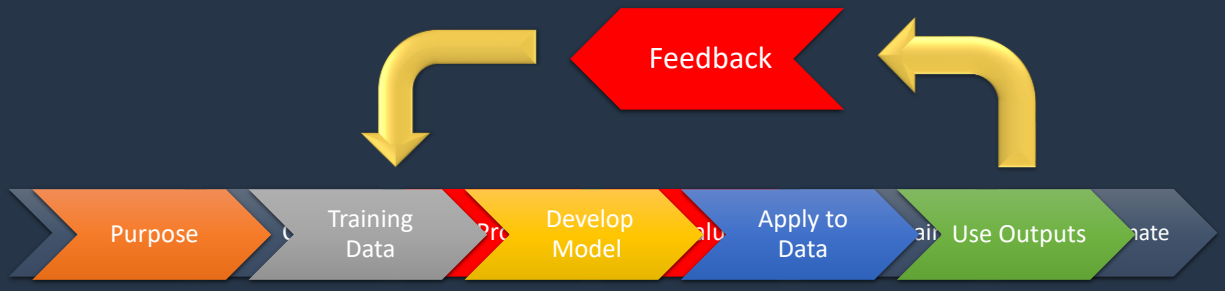

Governance in the Process
Authority
Purpose
Collect
Process
Evaluate
Retain
Disseminate
Training
Guidance
Compliance Controls
Technical Safeguards

5


Governance in the Process
Authority
Purpose
Collect
Process
Evaluate
Retain
Disseminate
Training
Guidance
Compliance Controls
Technical Safeguards

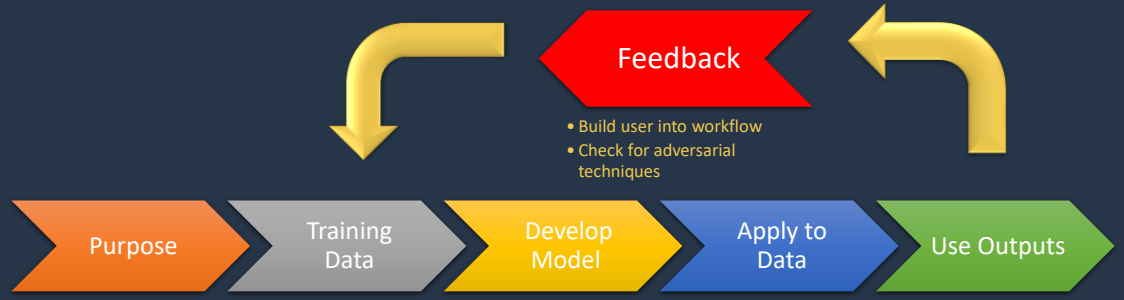
6

Governance in the Process



7

Machine Learning Process Safeguards



- **Purpose**
  - Authorities, Ethical Use
  - Explainable Purpose and Methods
- **Training Data**
  - Collect or generate training data and test data
  - Documentation: Datasheets for Datasets
- **Develop Model**
  - Explainability
  - Testing/Validation
  - Confidence Levels
  - Documentation: Model Cards
- **Apply to Data**
  - Use Limitation
  - User Interpretation and Validation
- **Use Outputs**
  - Accountability
    - Explainability and Redress
  - Human control
    - Confidence level follows outputs

8

Defined Purpose and Use

Purpose

- Governance Bodies
- Check for Authorities
- Check for Ethical Use (Principles)
- Explainable Purpose and Methods

9

Training Data

Purpose | Training Data

- Collect or generate training data and test data
  - Data Selection, Feature Engineering, Labeling
  - Issue: Collecting and maintaining “negative” examples
- Documentation: Datasheets for Datasets
  - Identify and document features, purpose, limitations, and known issues
  - biases (explicit and implicit)

10

Model Development

Purpose → Training Data → Develop Model

- Explainability
- Testing and Validation
  - Check for bias in weights/methodology
  - ID situations where model performs poorly/unreliably or is vulnerable to adversarial techniques
- Confidence Level
- Documentation: Model Cards

11

Stakeholders

12

Applying Models to Data

Purpose → Training Data → Develop Model → Apply to Data

- Use Limitation
- User Interpretation and Validation

13

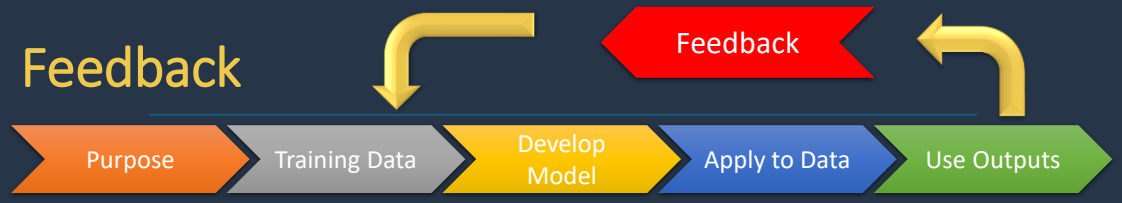
Using Outputs

Purpose → Training Data → Develop Model → Apply to Data → Use Outputs

- Accountability
  - Explainability and Redress
- Human control
  - Confidence level follows outputs

14

Feedback



- Build user corrections into workflow
  - Drift
  - Biased weights
- Check for adversarial techniques

15

Q & A

16

THANK YOU

DEFENDING OUR NATION SECURING THE FUTURE

17

Mitigating Adversarial Machine Learning

Machine learning (ML) can be a solution to scalable defensive and offensive measures for cybersecurity. These can range from semi-automated decision support to fully-automated capabilities. However, ML models can be exploited in at least four ways. Adversaries can:

- ***(a) poison*** training data used to train ML algorithms to degrade prediction quality, or redirect predictions, altogether;
- ***(b) evade*** by manipulating runtime data to ensure ML models misclassify malicious behavior as benign;
- ***(c) infer*** records into the training data; and
- ***(d) reconstruct*** the ML model for further analysis and exploitation.

When ML models of varying qualities are integrated into an ensemble, an adversary can exploit weaknesses in individual models to coordinate a malicious effect in the overall system.

*Popular Science: ***Fooling The Machine,*** *The Byzantine science of deceiving artificial intelligence;* Dave Gershgorn (March 30, 2016)

18

18