INSTITUTE FOR DEFENSE ANALYSES

DATAWorks 2023 - RAI Assurance for Personnel-Related Capabilities

Rachel A. Haga, Project Leader

John W. Dennis

April 2023

IDA Document NS D-33454

Log: H 2023-000110

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task BM-9-5153, "Office of the Chief Digital and Artificial Intelligence Officer."
The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Dr. V. Bram Lillard and consisted of Ms. Rachel A. Haga and Mr. Yosef S. Razin from the Operational Evaluation Division, and Mr. Metin A. Toksoz-Exley from the Science and Technology Division.

For more information:
Ms. Rachel A. Haga, Project Leader
rhaga@ida.org • (703) 578-2768

Dr. V. Bram Lillard, Director, Operational Evaluation Division
vlillard@ida.org • (703) 845-2230

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-33454

DATAWorks 2023 - RAI Assurance for Personnel-Related Capabilities

Rachel A. Haga, Project Leader

John W. Dennis

Executive Summary

Testing and assuring responsible use of capabilities enabled by artificial intelligence and machine learning (AI and ML) is a nascent topic in the DOD, with many efforts being spearheaded by DOD's Chief Digital and Artificial Intelligence Office (CDAO). In general, black box models tend to suffer from issues related to edge cases, emergent behavior, misplaced or lack of trust, and many other factors. For these reasons, traditional testing is insufficient to guarantee safety and responsibility in the employment of a given AI-enabled capability. Focus of this concern tends to fall on well-publicized, high-risk capabilities such as AI-enabled autonomous weapon systems. In those use cases, unexpected behavior and misplaced trust can result in consequences that may lead to loss of life. Further, structured and robust testing oversight over these use cases provides a starting point to operationalize that focus.

In contrast, AI- and ML-enabled capabilities supporting personnel processes and systems, such as algorithms for retention and promotion decision support, tend to carry low safety risk and are often characterized by less complex implementations with less robust testing oversight. However, the personnel space has many idiosyncratic concerns that run the risk of undermining the DOD's five ethical principles for responsible AI (RAI). Examples include service member privacy concerns, invalid prospective policy analysis, disparate impact against marginalized service member groups, and emergent service member behavior in response to use of the capability.

While many of these concerns are not novel to researchers studying human capital, the erosion of barriers to the use of AI and ML is facilitating an increase in the number of applications, even as many of these concerns remain poorly understood by the community at large. Further, while it is often easy to identify when many of these concerns have arisen ex post, it is not easy to quantify them in a way that facilitates testing ex ante. For this reason, we consider notions of assurance to provide evidence of the adherence to DOD's ethical principles. Our guide documents evidence and mechanisms to aid in satisfying assurance, and we provide a concrete example by considering many of these issues in the context of an IDA ML-enabled capability.

Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

Assurance of Responsible AI (RAI) in Context: ML-Enabled Army Personnel Forecasting

John W. Dennis,
Rachel Haga, Yosef Razin, Metin Toksoz-Exley, Ed Wang
DATAWorks - April 2023

Work funded by

CDAO
AI Assurance

Why Assurance for AI?

Traditional T&E is generally insufficient.

- AI can have emergent behavior, edge cases, changing operating environments.

AI T&E is never done.

- Continuous monitoring, ongoing stakeholder feedback, feedback loops to development.

Testing RAI robustly is hard

- It is easy to say what went wrong but hard to quantify up front.

Processes exist to help handle RAI, including
ASSURANCE:
The use of formal arguments to augment testing gaps

Goals for Assuring RAI

Demonstrate to stakeholders:

- **Responsible use** and **guardrails** for the capability
- Mechanisms to **catch, report, and fix emerging concerns**
- **Good-faith efforts** beyond
  - "Does the software run?"
  - "Are the forecasts accurate?"

Assurance is a *living concept*

Part of broader effort of *Support, Training, and Assurance*

AI-Enabled Personnel Processes

Personnel Processes:
Recruiting, Retention, Promotion, Resilience

Many Opportunities

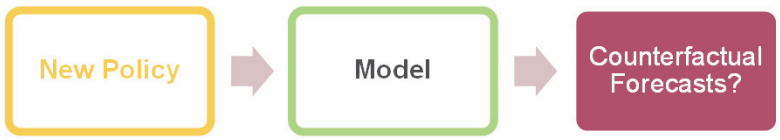
- Risks are often lower profile
- DOD personnel environment is very large
- Often less complex involvement of AI/ML on smaller budgets
- AI/ML is “easy”

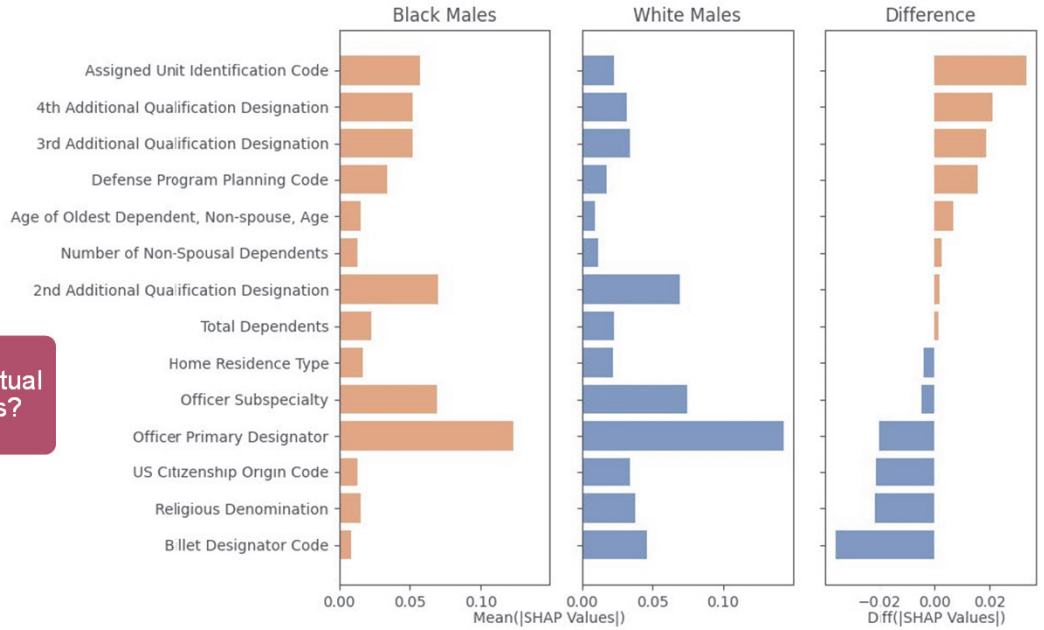
But

- Black boxes representing biased data
- Personnel data generating process is itself complex due to human behavior

Personnel Space has Unique Concerns

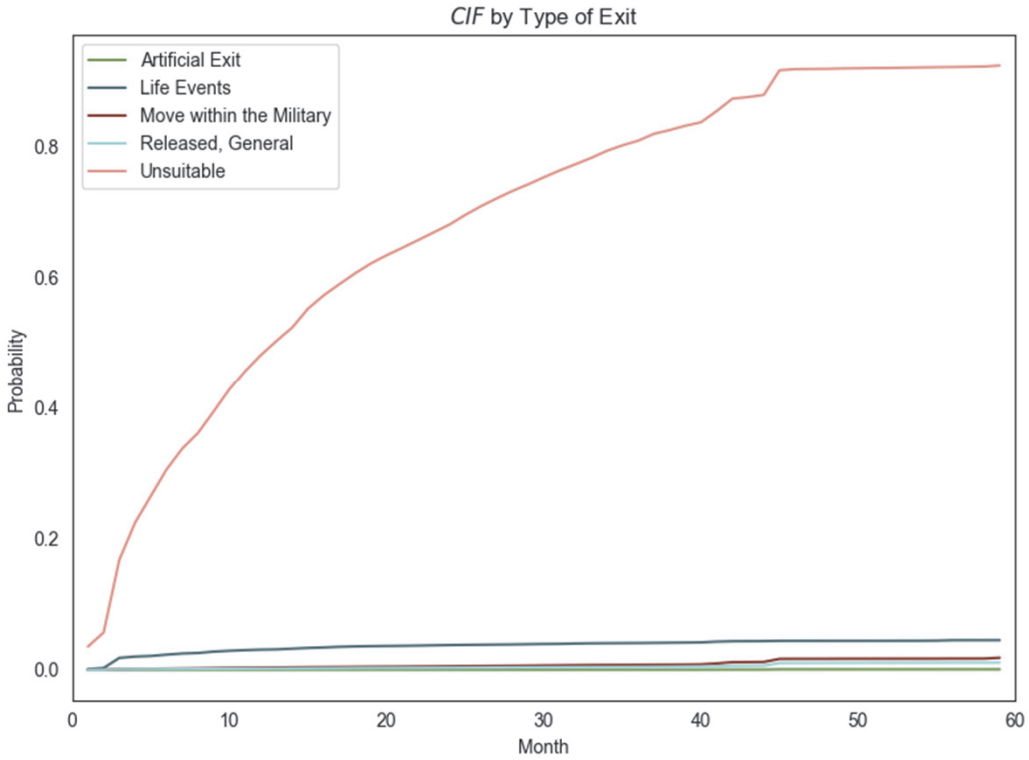
- Disparate impact/treatment
- Invalid prospective policy analysis (invalid counterfactuals!)
- Misattributed causality





Personnel Space has Unique Concerns

- Privacy
- Emergent service member behavior
- Perverse incentives
- Robustness



Assuring RAI in the Personnel Space

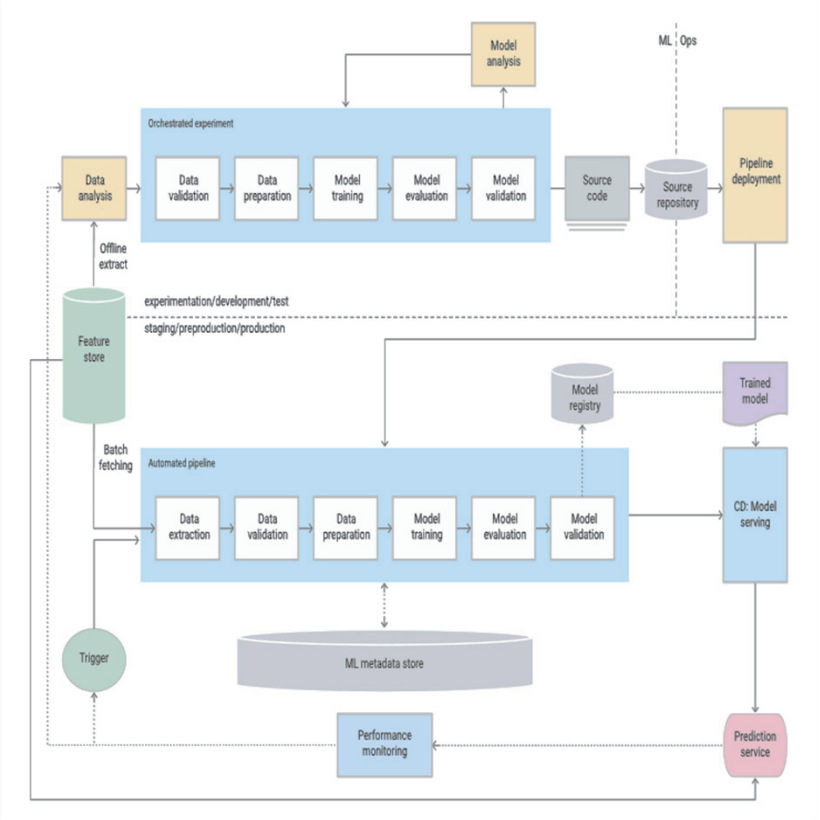
Assurance Guide

- MLOps scaffolding
- DOD 5 ethical principles

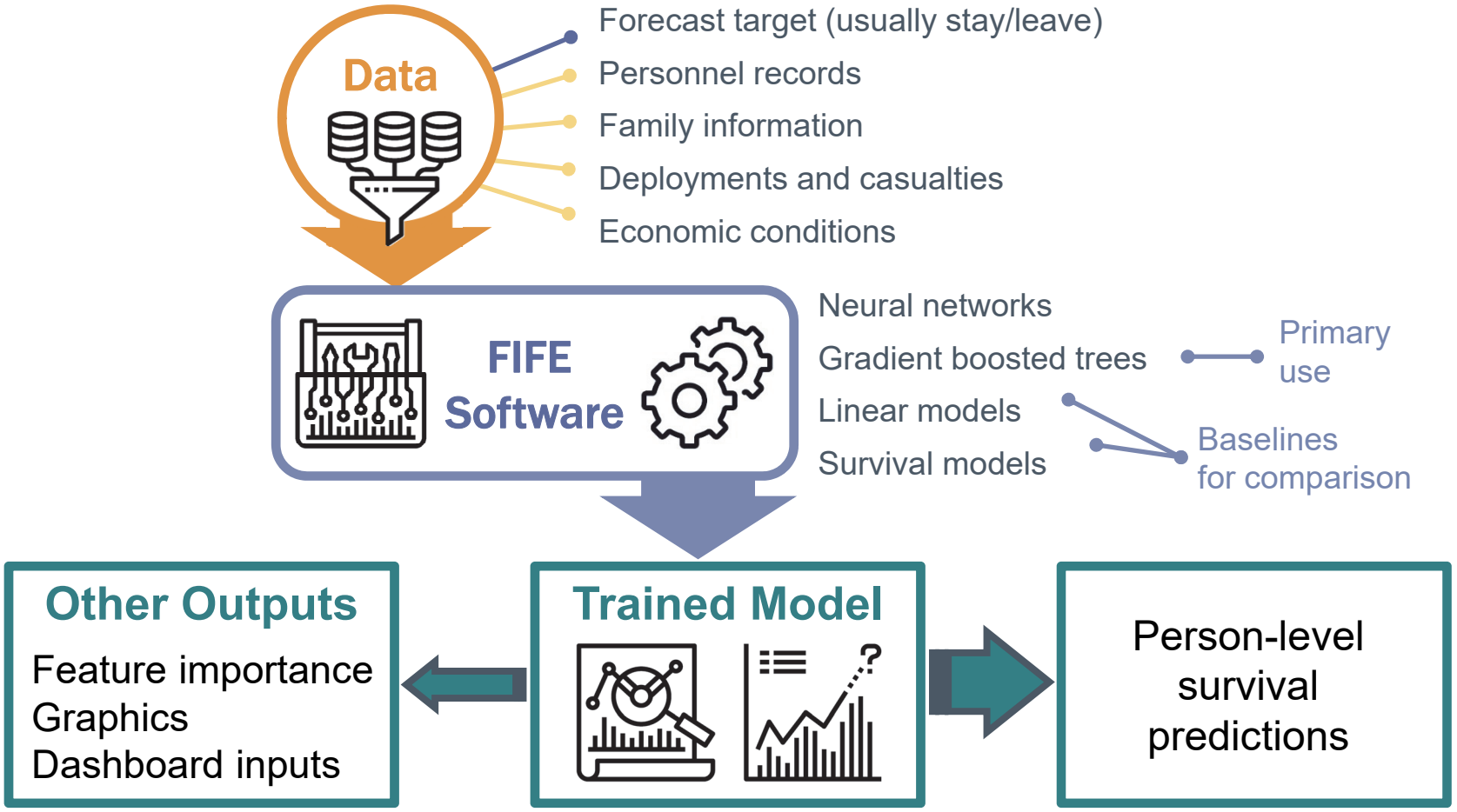
  + Privacy
- Personnel space nuance
- Strategies for testing, monitoring, feedback, etc.

Assurance Case

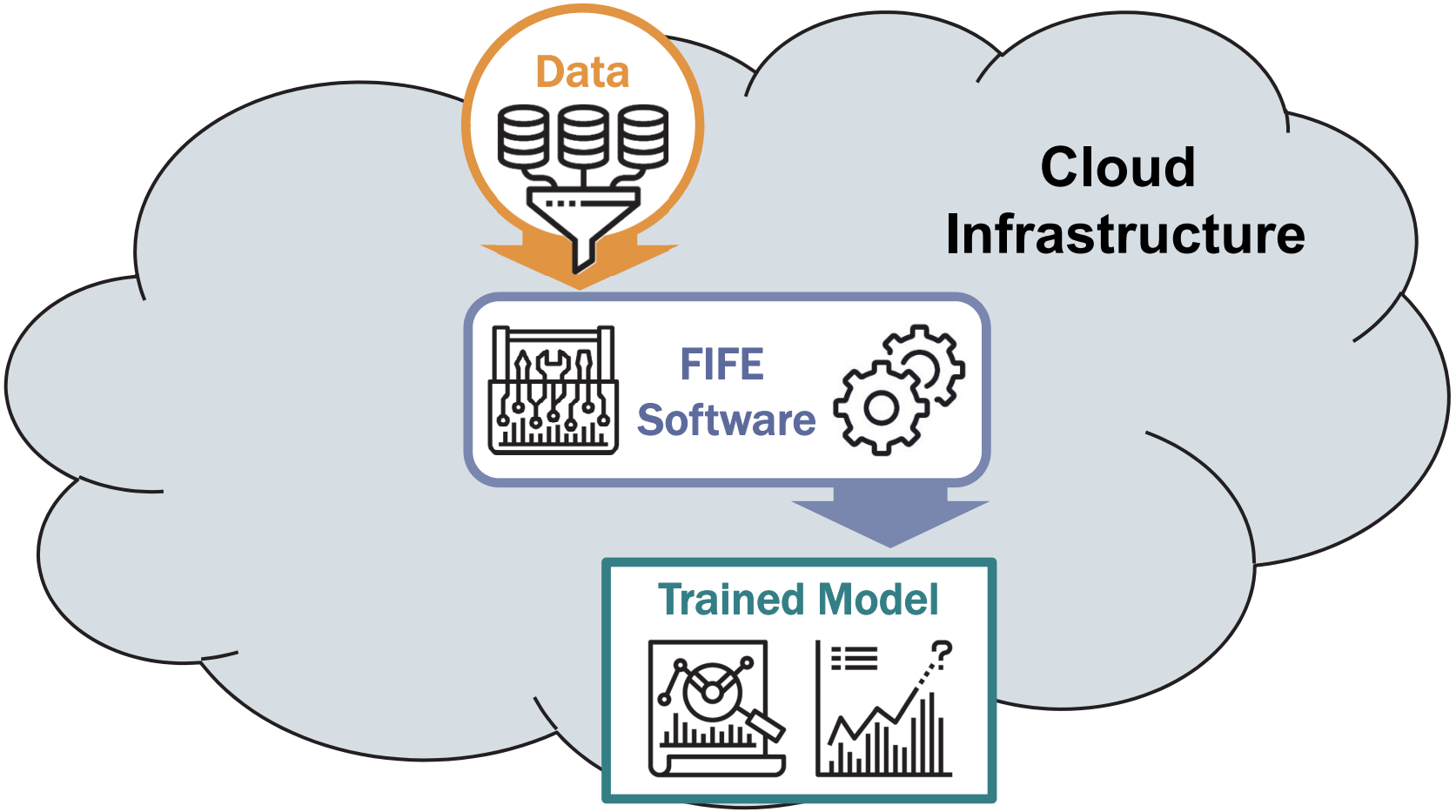
- Application of the guide to a **Army Retention Prediction Model (RPM)**



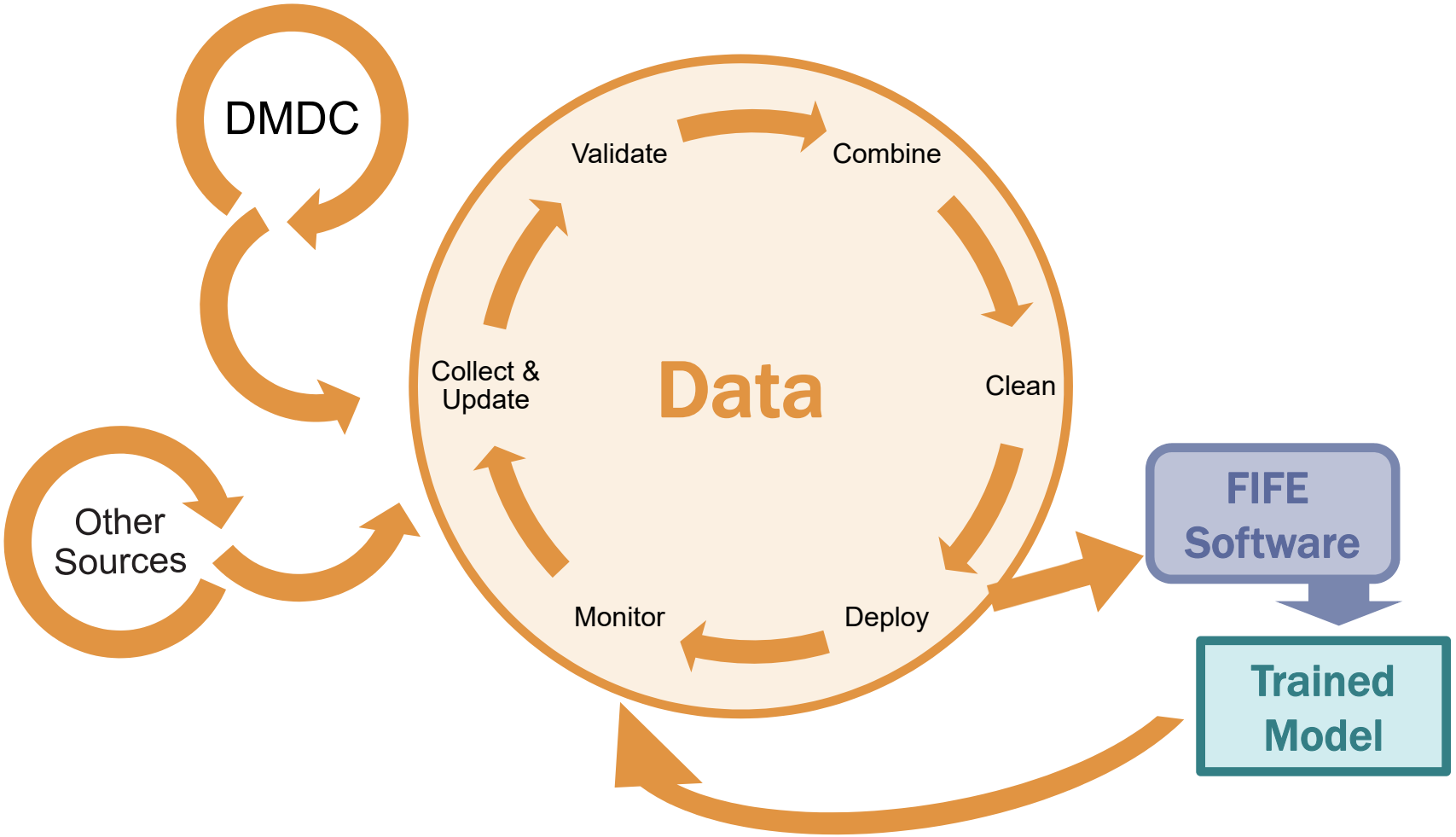
Use Case - Retention Prediction Model (RPM)-Army



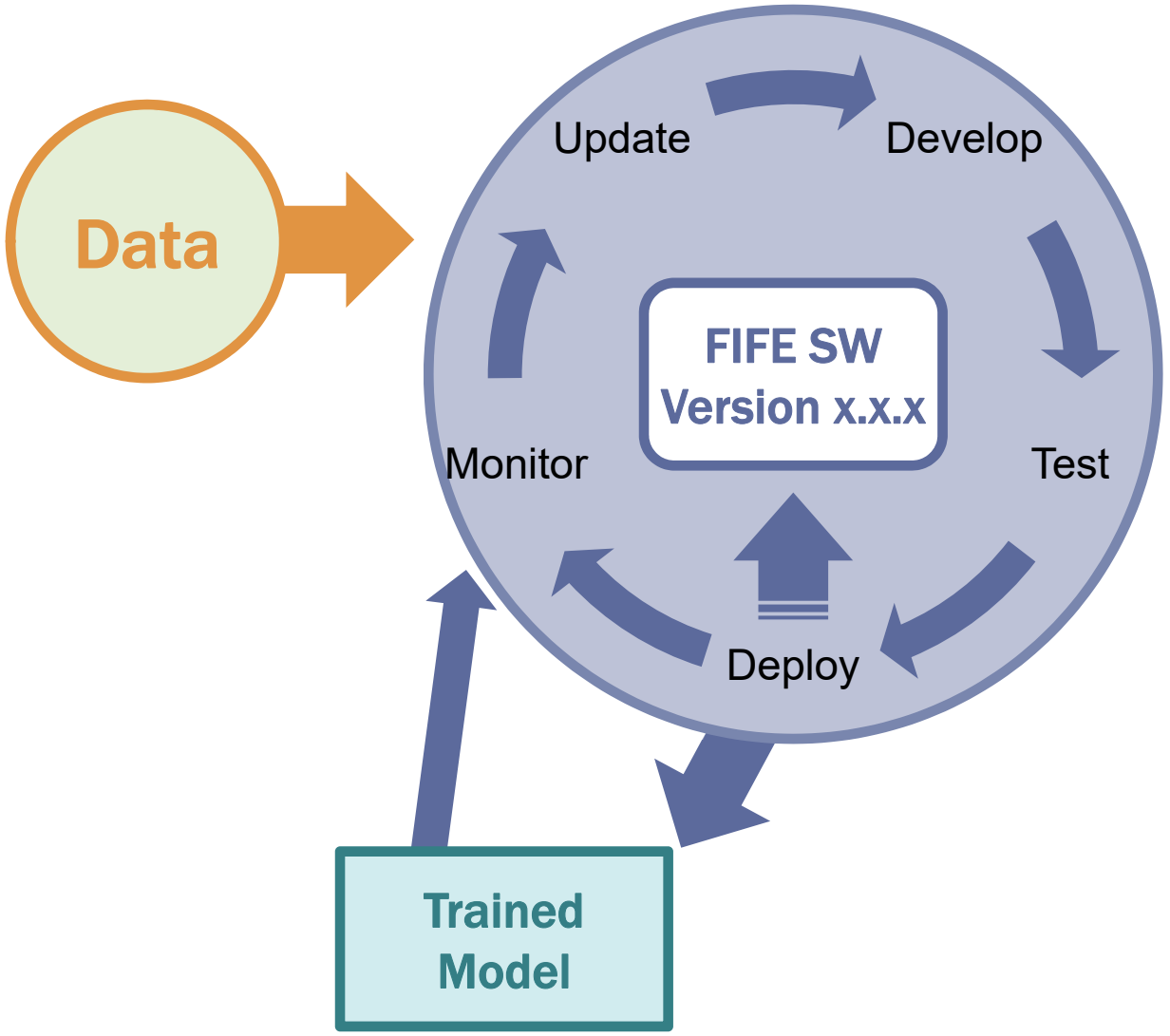
Ecosystem



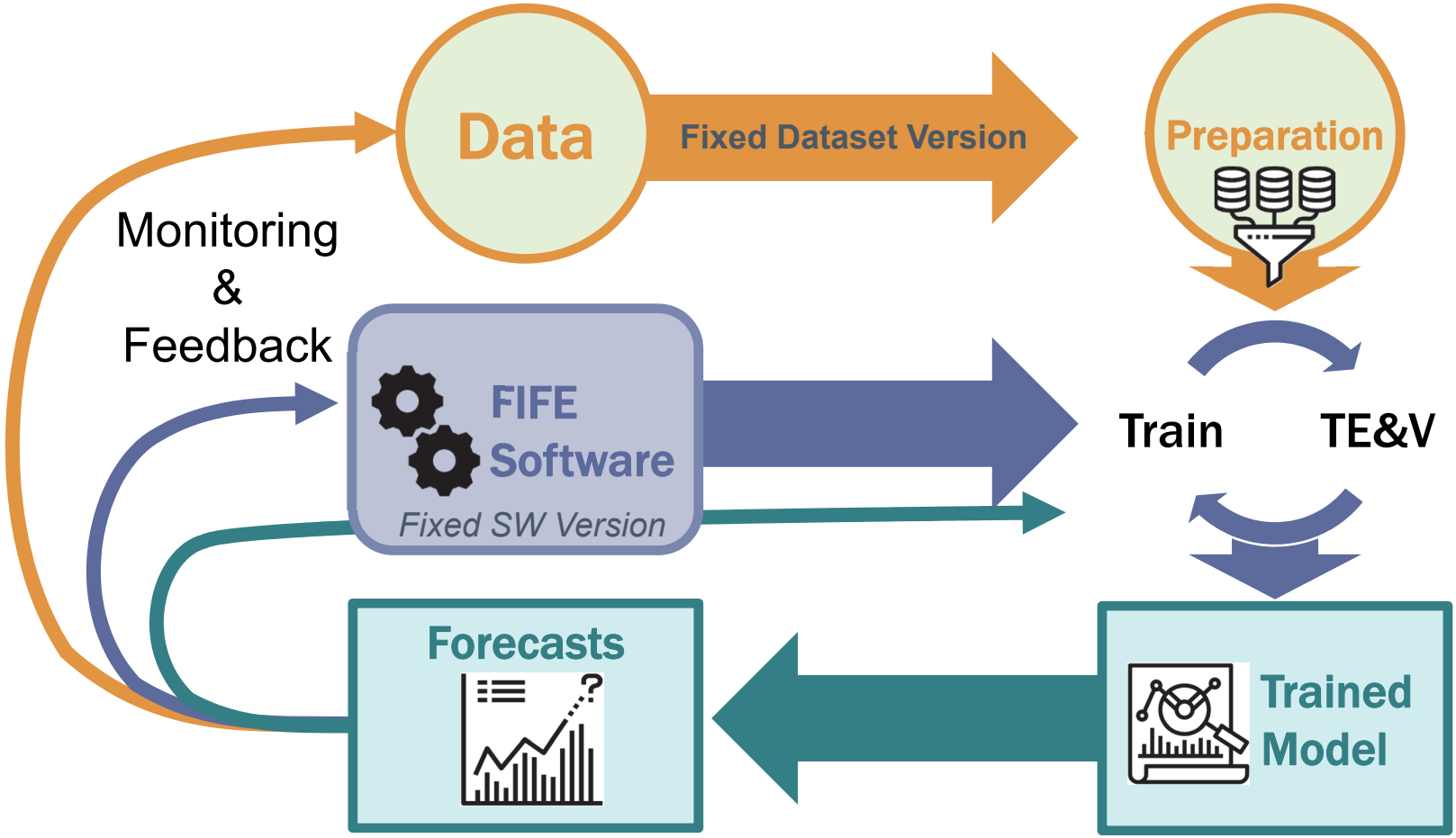
Data Curation Lifecycle



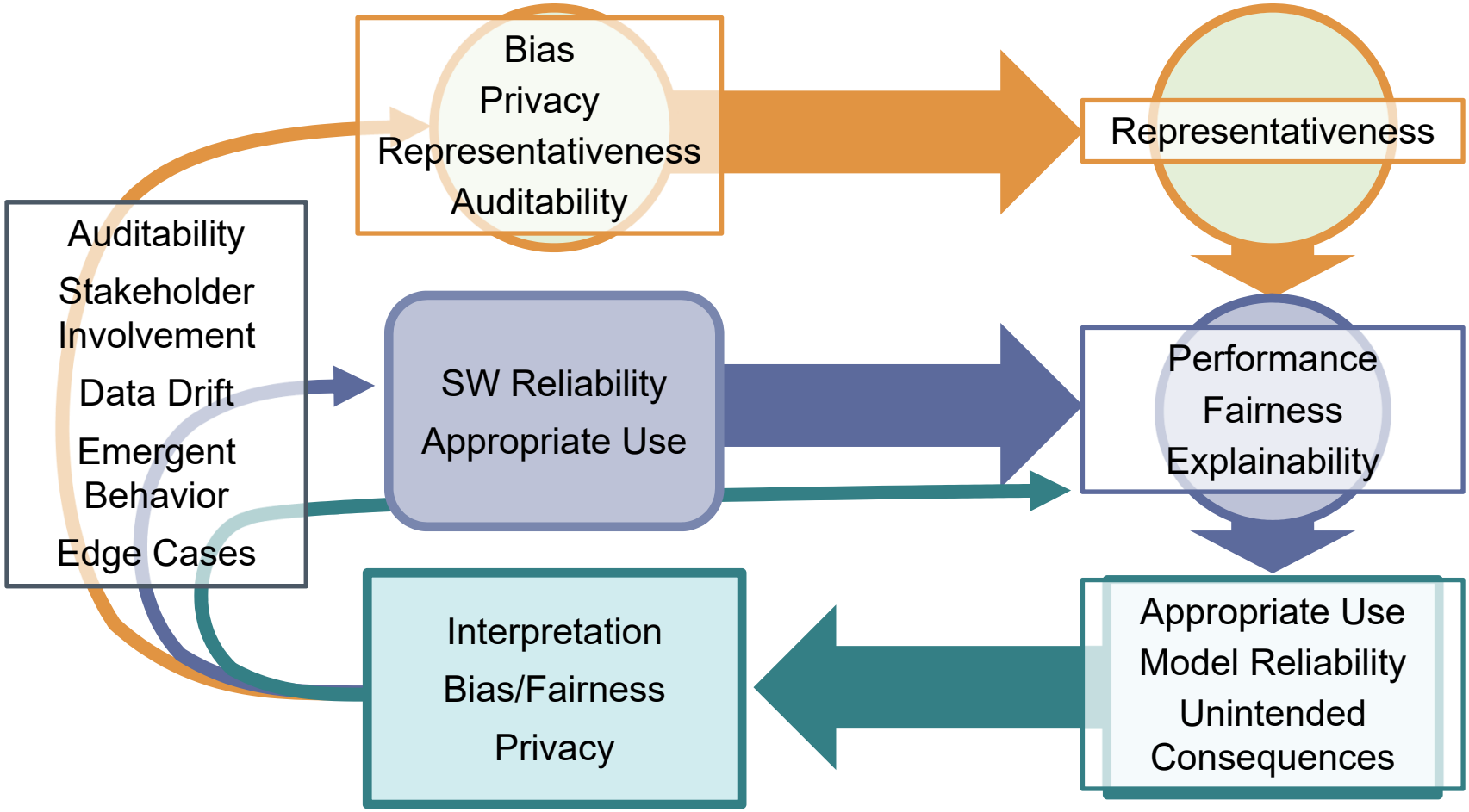
FIFE Software Development Lifecycle



Model Lifecycle

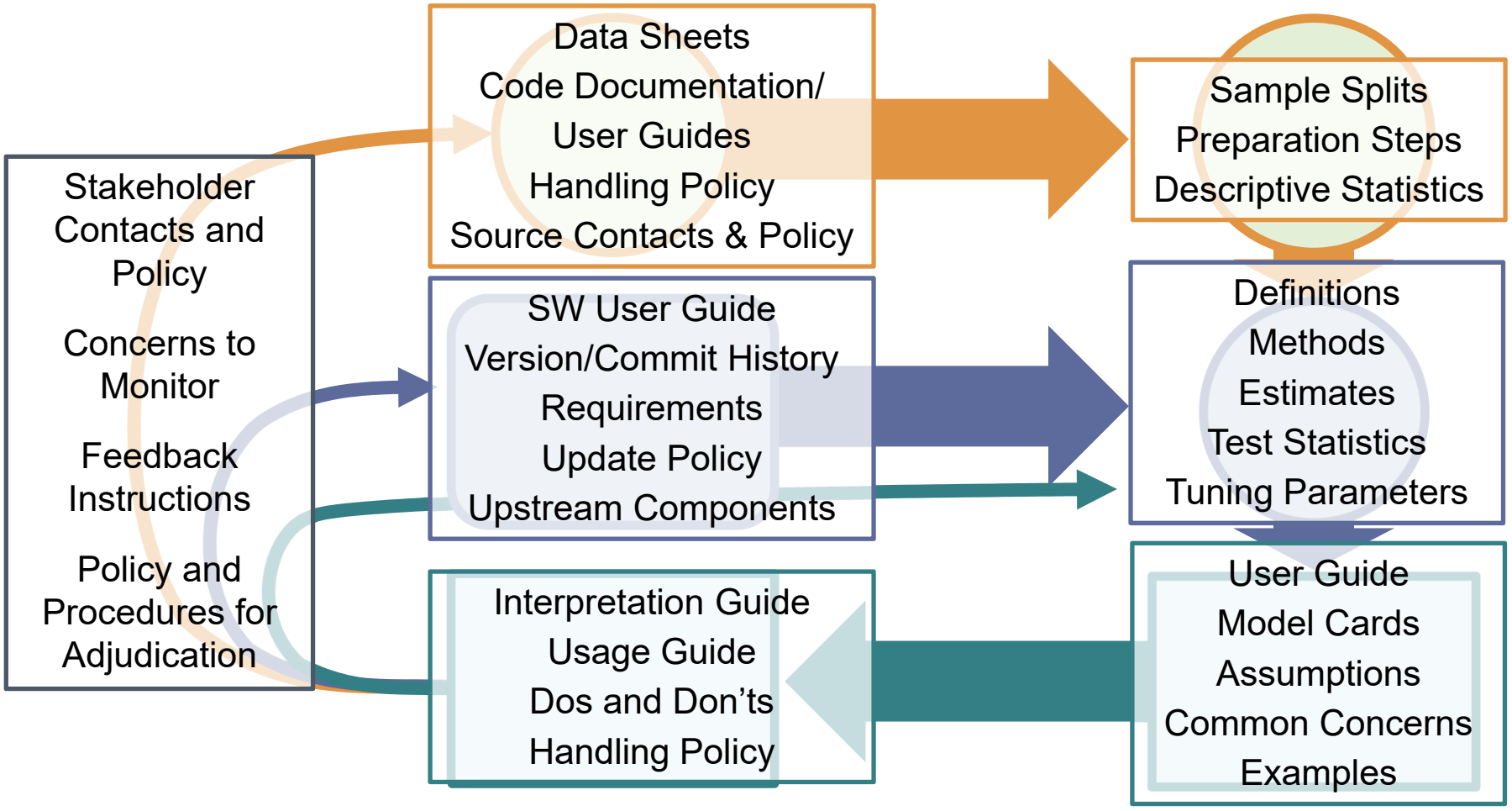


RAI in the Lifecycle



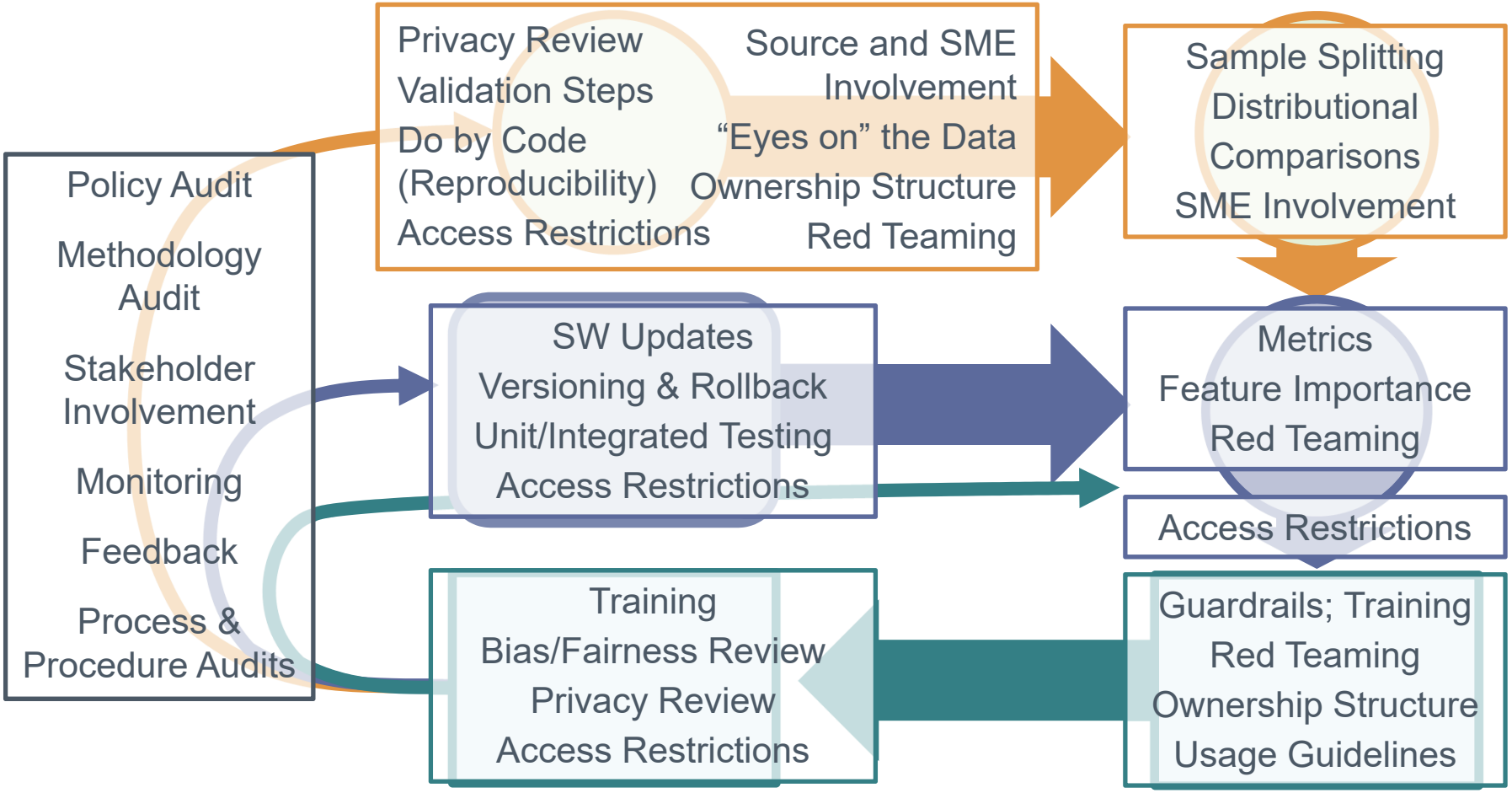
List is illustrative but not exhaustive.
SW - Software

Documentation



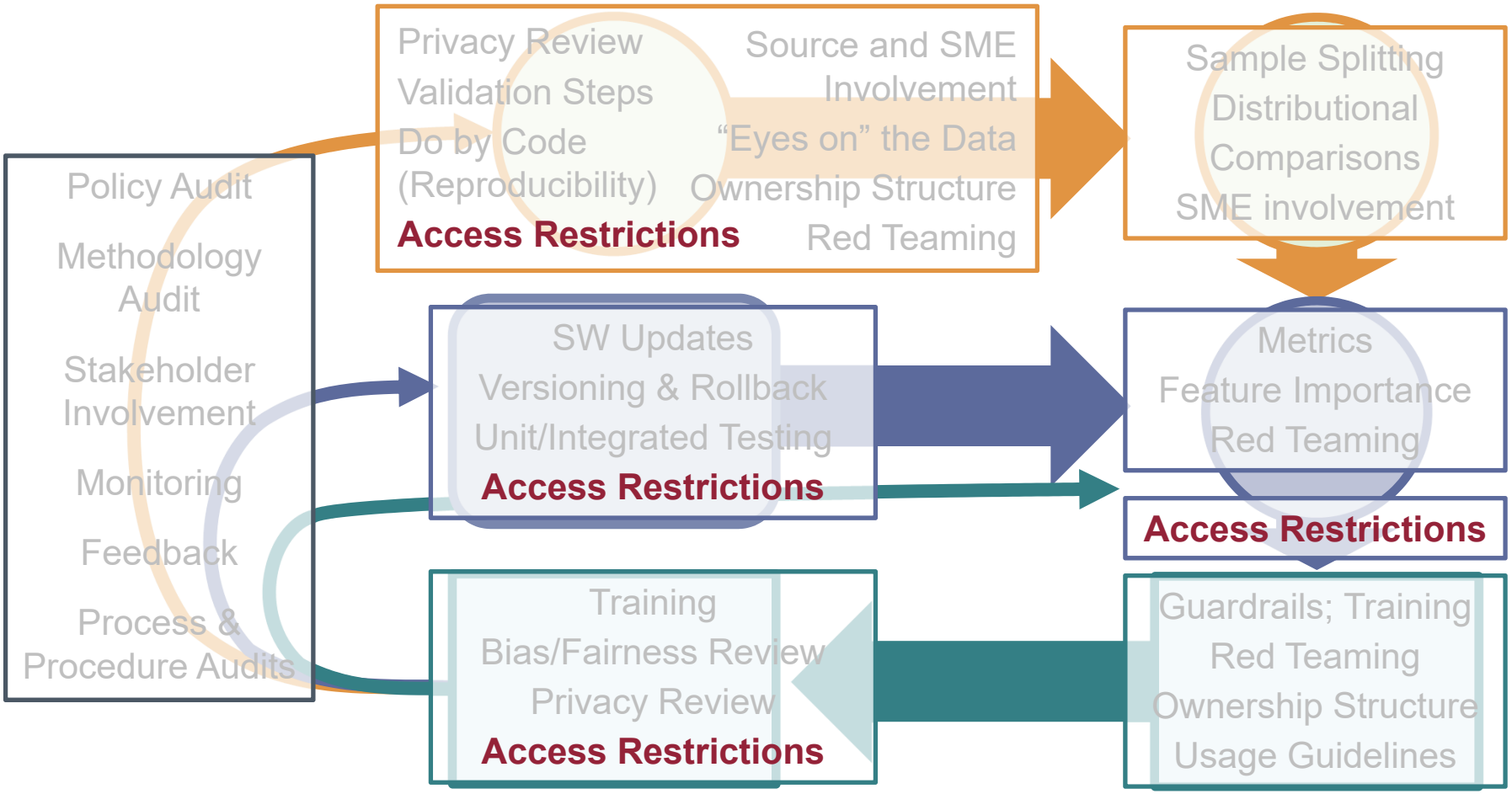
List is illustrative but not exhaustive.

Assurance Mechanisms

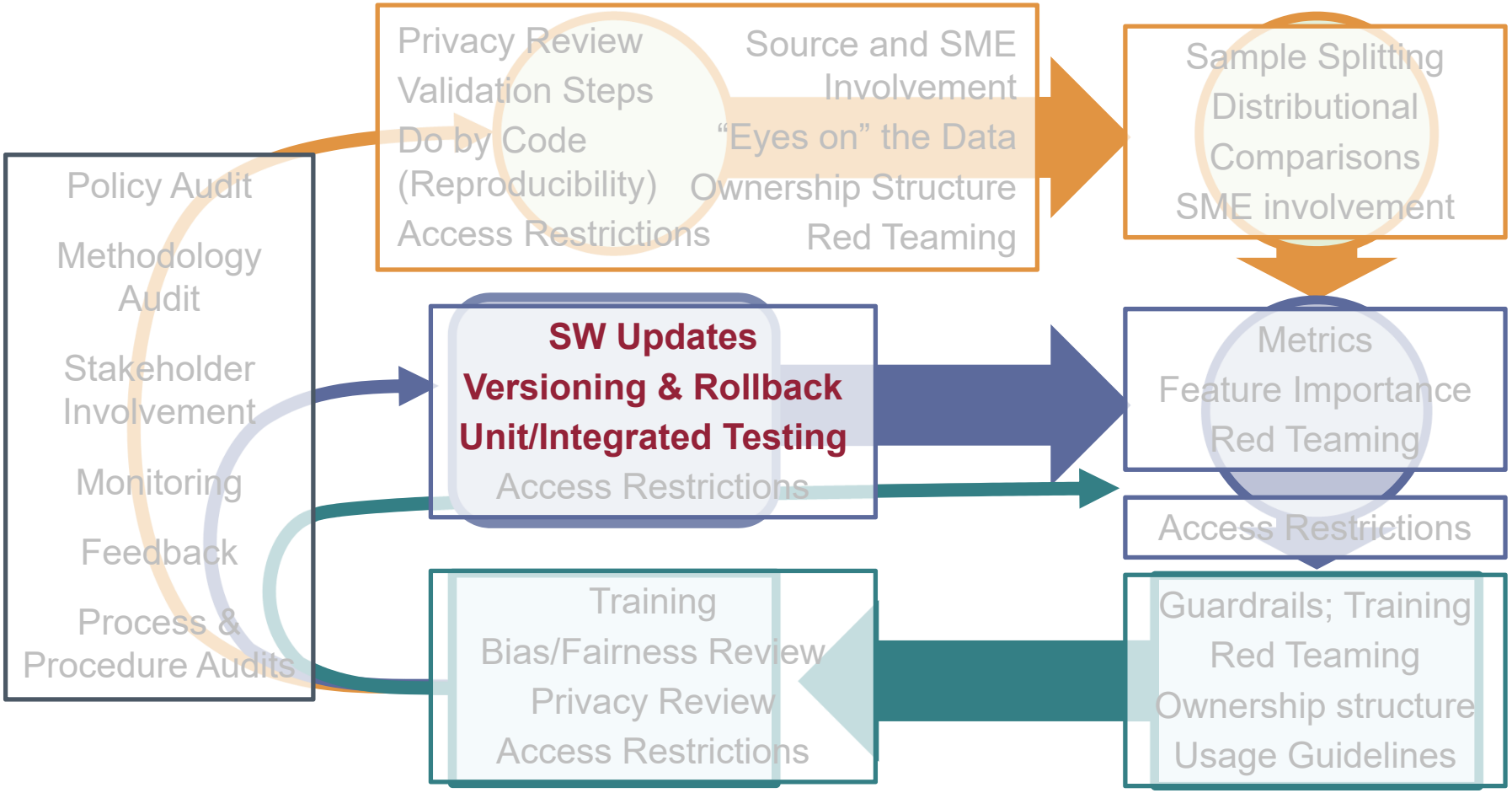


List is illustrative but not exhaustive.

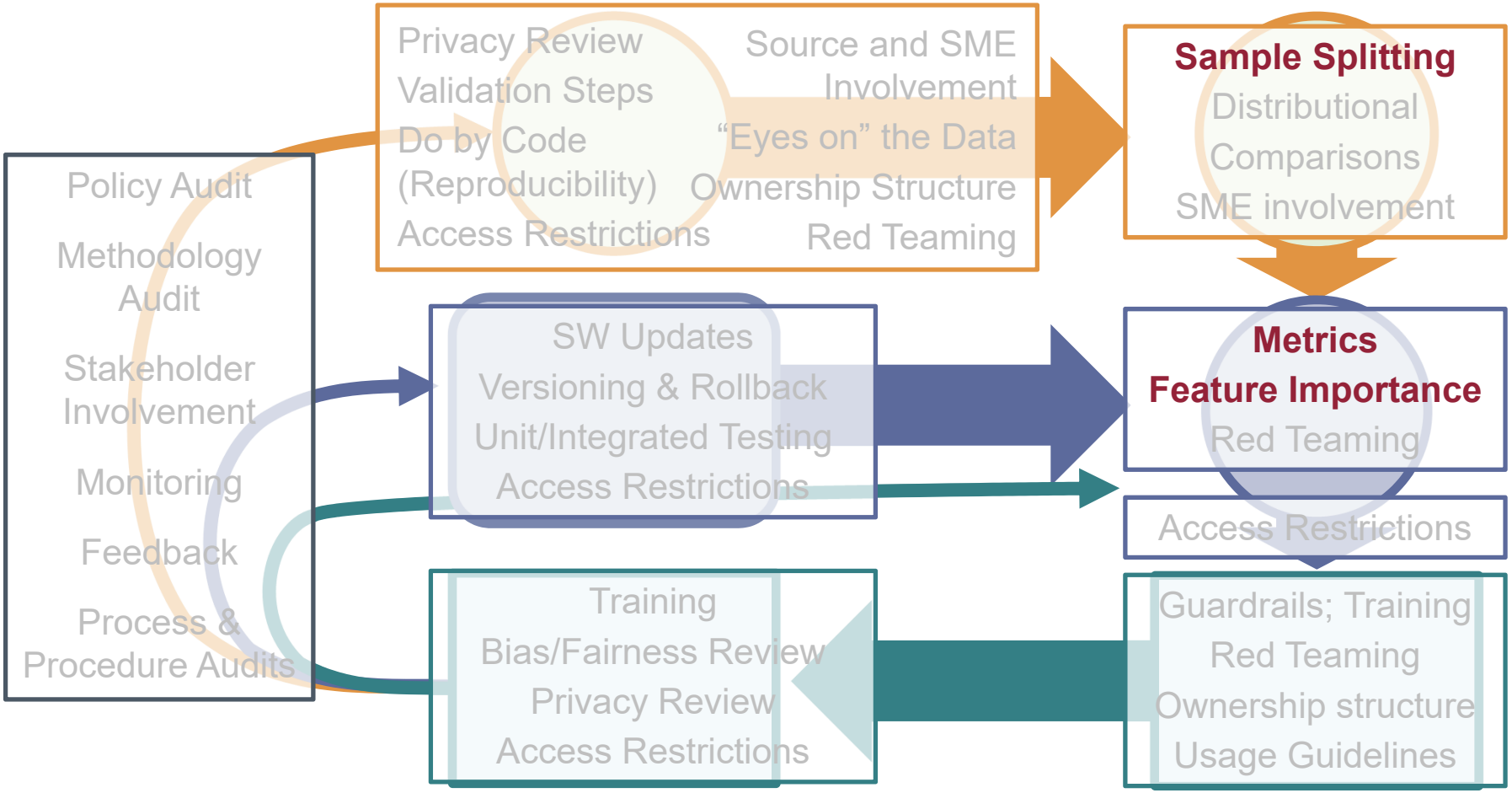
Assurance Mechanisms:
Access



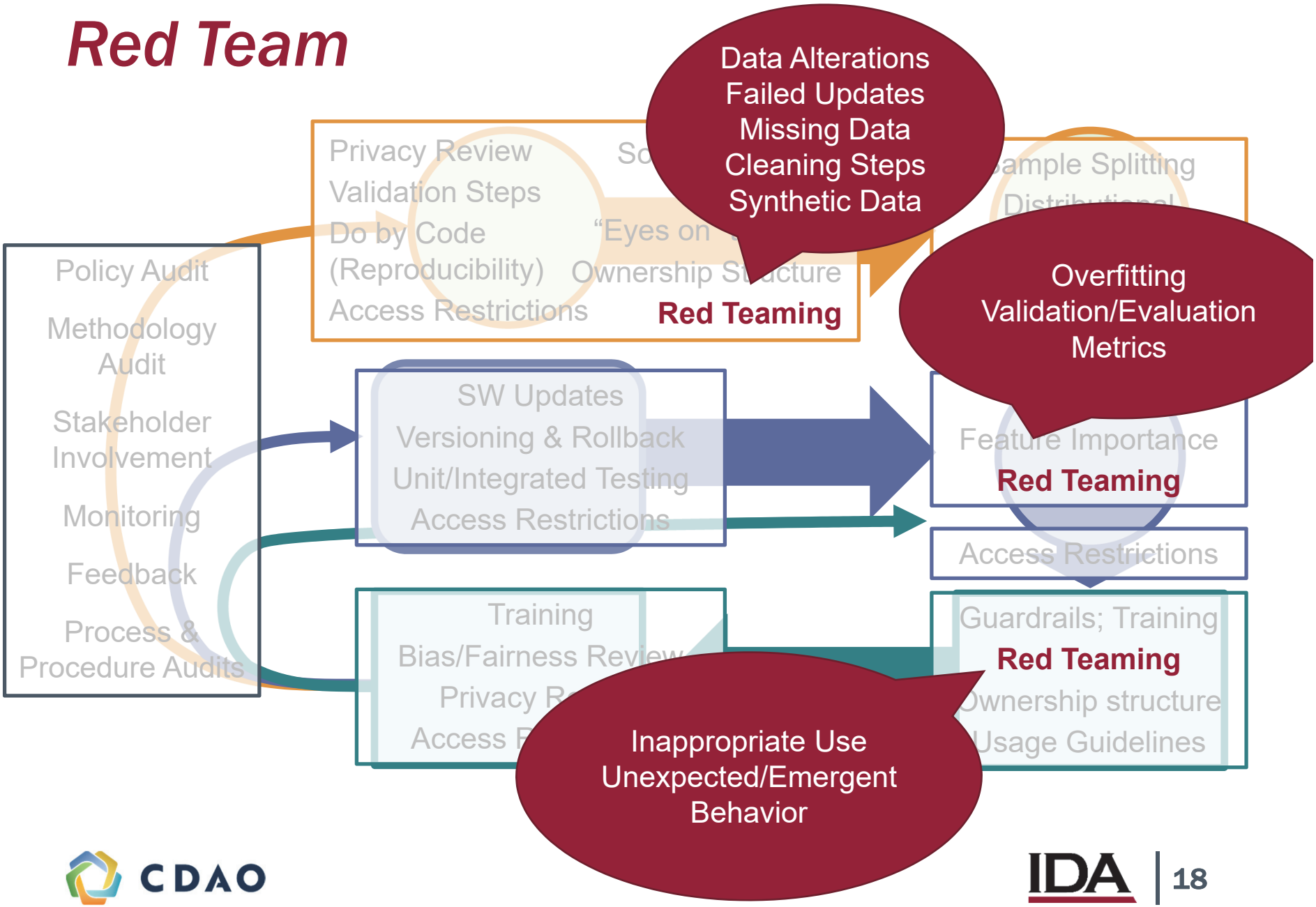
Assurance Mechanisms:
Traditional Software



Assurance Mechanisms:
Traditional ML



Assurance Mechanisms:
Red Team



Conclusions: Assuring RAI for Personnel

- Many emerging use cases for AI
- Uses with personnel data have unique concerns
- Legal, moral, ethical issues
- Concerns are not always obvious
- Need a framework for ensuring responsible use

Conclusions: Assurance for RAI

- Similar in spirit to traditional assurance cases
- We can’t formally test everything
- Need formal arguments and evidence
- We can build this into existing frameworks

jdennis@ida.org

Work funded by

CDAO
AI Assurance

Image Sources

- https://www.defense.gov/Multimedia/Photos/
- Dennis, John W., Augustine, Rachel G., Guggisberg, Michael R. and Lockwood, Julie A. 2021. Expanding the Finite Interval Forecasting Engine for Navy Personnel Management: Incorporating Competing Risks into Retention Prediction. IDA Paper P-31873.
- https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning
- https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/
- Lockwood, Julie A., King, Joseph M. and Augustine, Rachel G. 2020. Explaining Differences in Predicted O-5 Promotion Outcomes by Race and Gender among Naval Officers. IDA Paper P-20452.
- Jain, Akshay A. and Dennis, John W. 2022. DATAWorks 2022: Forecasting with Machine Learning. IDA Document NS D-33017.
- Jain, Akshay A., Dennis, John W., Lockwood, Julie A., Song, Minerva S., Latshaw, Nathaniel T., Eifert, Erin P. and King, Joseph M. 2022. Forecasting Demand for Air National Guard Training to Improve Military Readiness. IDA Paper P-32920.

Appendix

What are we Assuring?

- T&E typically focuses on Proper Functioning and other operational standards.
  - Usual T&E is not sufficient for AI enabled capabilities (but it is still necessary!).
- Typical assurance focuses on Safety.
- Concerns in the personnel space often focus on Legal, Moral, and Ethical issues.
- 5 RAI Principles (attempt to) encompass these concerns for all uses of AI in the DOD.
  - How do we implement these principles?
  - How do we know our implementation is effective?

Use Case from Army TMTF

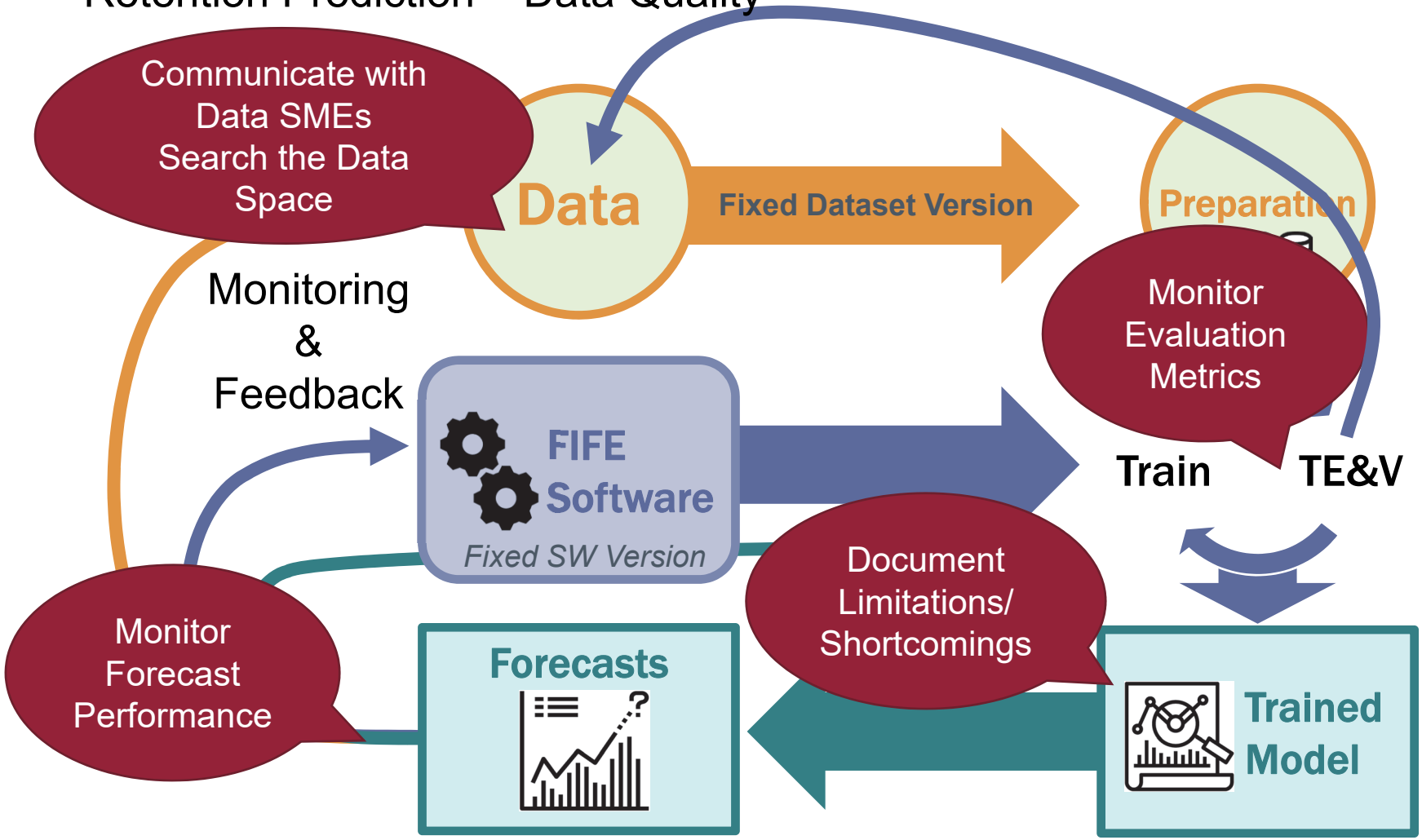
- Predictive Retention Toolkit and Evaluation for Targeted Army Talent Management
- Overarching question: How can the Army best select, shape, train, and retain the force it wants?
- Three-part study aimed at retention efforts:
  1. Forecast retention with high fidelity and accuracy
  2. Discover indicators of superior performance
  3. Assess the impact of targeted retention incentives

Forecast Retention with High Fidelity and Accuracy

- Finite Interval Forecasting Engine (FIFE) – survival modeling in the machine learning context
- IDA developed FIFE in a multi-year research partnership with OSD
- Variety of use cases across a variety of IDA projects and services/components
- Open source development*
- Capability/Data Assets and Pipeline previously resided exclusively at IDA; now experiencing a shift to DOD cloud platforms

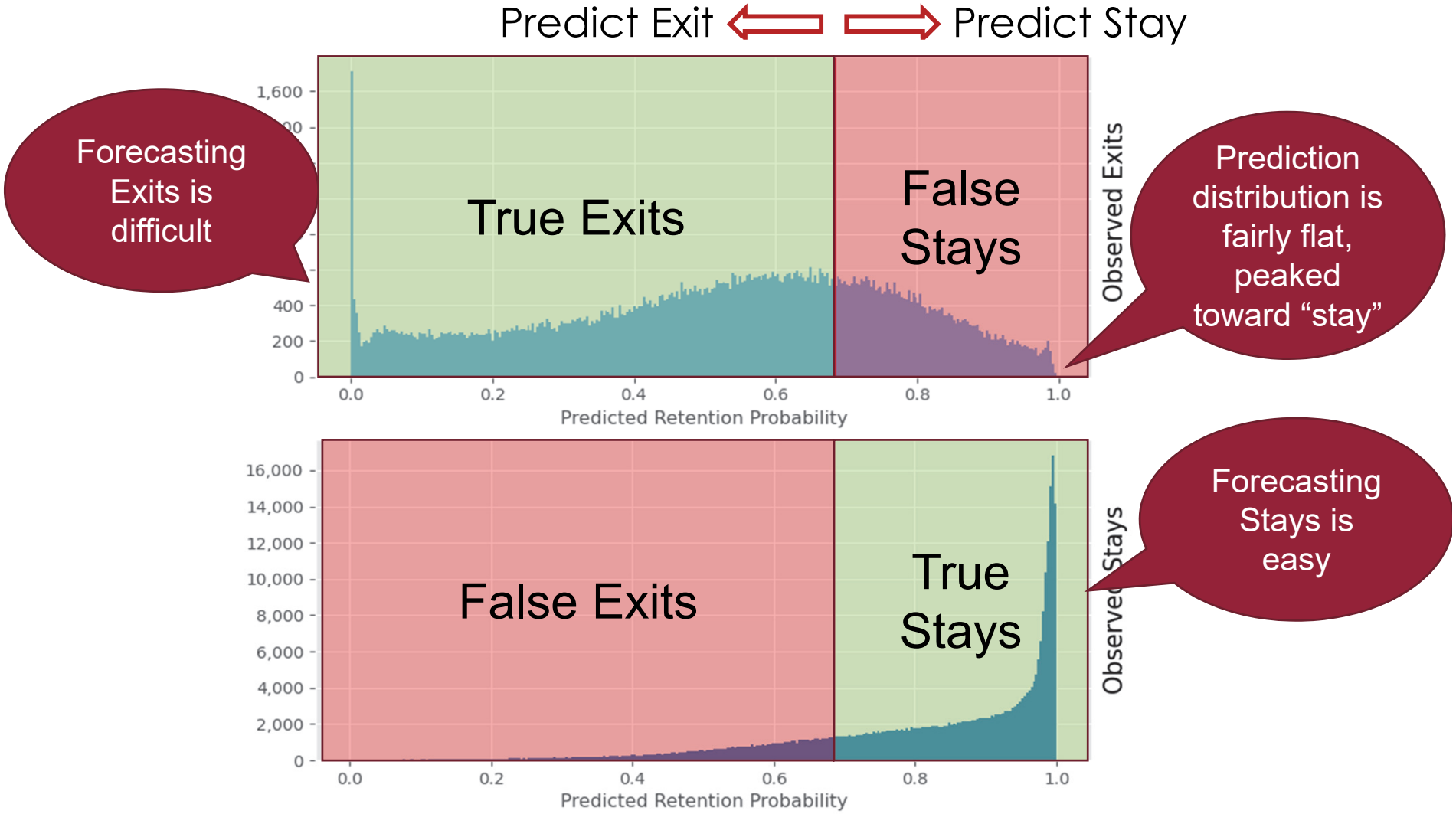
Example – Model Lifecycle

Retention Prediction – Data Quality



Example - Metrics

Retention Prediction – Data Quality



BREAKING DEFENSE

If generative AI can be made reliable — and that's a significant if — the applications for the Pentagon, as for the private sector, are extensive, Groen and Shanahan agreed.

"Probably the places that make the most sense in the near term... are those back-office business from personnel management to budgeting to logistics," Shanahan said. But in longer term, "there is an imperative to use them to help deal with ... the entire intelligence cycle."

The New York Times

Become an A.I. Expert | How Chatbots Work | Why Chatbots 'Hallucinate' | How to Use C

Bing's A.I. Chat: 'I Want to Be Alive. 😈'

NEWS POLITICS U.S. NEWS BUSINESS WORLD TECH HEALTH CULTURE & TRENDS NBC NEWS TIPLINE WATCH NOW

INTERNET

A mental health tech company ran an AI experiment on real users. Nothing's stopping apps from conducting more.

jdennis@ida.org

Work funded by

CDAO
AI Assurance

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
XX-04-2023	Final	

4. TITLE AND SUBTITLE

DATAWorks 2023 - RAI Assurance for Personnel Related Capabilities

5a. CONTRACT NUMBER

HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Dr. John W. Dennis

5d. PROJECT NUMBER

BM-9-5153

5e. TASK NUMBER

5153

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
730 East Glebe Road
Alexandria, Virginia 22305

8. PERFORMING ORGANIZATION REPORT NUMBER

NS D-33454

H 2023-000110

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Public release approved. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Testing and assuring responsible use of AI/ML enabled capabilities is a nascent topic in the DOD with many efforts being spearheaded by CDAO. In general, black box models tend to suffer from consequences related to edge cases, emergent behavior, misplaced or lack of trust, and many other issues, so traditional testing is insufficient to guarantee safety and responsibility in the employment of a given AI enabled capability. Focus of this concern tends to fall on well-publicized and high-risk capabilities, such as AI enabled autonomous weapons systems. However, while AI/ML enabled capabilities supporting personnel processes and systems, such as algorithms used for retention and promotion decision support, tend to carry low safety risk, many concerns, some of them specific to the personnel space, run the risk of undermining the DOD ' s 5 ethical principles for RAI. Examples include service member privacy concerns, invalid prospective policy analysis, disparate impact against marginalized service member groups, and unintended emergent service member behavior in response to use of the capability. Eroding barriers to use of AI/ML are facilitating an increasing number of applications while some of these concerns are still not well understood by the analytical community. We consider many of these issues in the context of an IDA ML enabled capability and propose mechanisms to assure stakeholders of the adherence to the DOD ' s ethical principles.

15. SUBJECT TERMS

Responsible Artificial Intelligence (RAI); Forecasting; retention prediction; Machine Learning (ML)

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Ms. Rachel A. Haga
Unclassified	Unclassified	Unclassified	SAR	30	19b. TELEPHONE NUMBER *(Include area code)* (703) 578-2768