



INSTITUTE FOR DEFENSE ANALYSES

WEAI 2023 - Assurance of Responsible AI in Personnel Management

Rachel A. Haga, Project Leader

John W. Dennis

June 2023

Distribution Statement A.
Approved for public release:
distribution is unlimited.

IDA Product 1038161

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

Copyright Notice

© 2022 Institute for Defense Analyses
730 East Glebe Road, Alexandria, Virginia 22305 • (703) 845-2000

This material may be reproduced by or for the U.S. Government pursuant to the copyright license under the clause at DFARS 252.227-7013 [Feb. 2014].

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Product 1038161

**WEAI 2023 - Assurance of Responsible AI in
Personnel Management**

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 an example by considering some 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 in Personnel Management

John W. Dennis

WEAI - July 2023

Work funded by



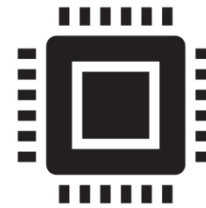
CDAO
AI Assurance

Advanced Analytics are Increasingly More Accessible

What will happen to retention if we increase bonuses?



More people will stay.



People who take bonuses might stay anyways, so the answer is not clear.

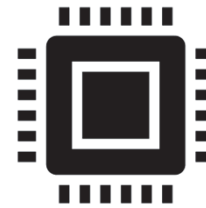


Advanced Analytics are Increasingly More Accessible

Who is most likely to succeed if promoted?



Person A because they have features correlated with success.



Those features are also correlated with gender and race. And there is selection in the data...



AI in Personnel Management

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



Advanced Analytics are Increasingly More Accessible

Code interpreter Alpha

An experimental ChatGPT model that

“Code Interpreter has turned GPT into a first-rate data analyst. Not a data analysis tool, but a data analyst.

It is capable of independently looking at a dataset, figuring out what is interesting, developing an analytical strategy, cleaning data, testing its strategy, adjusting to errors, and offering advice based on its results.

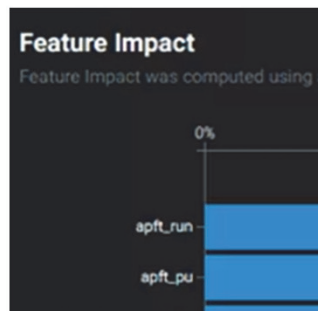
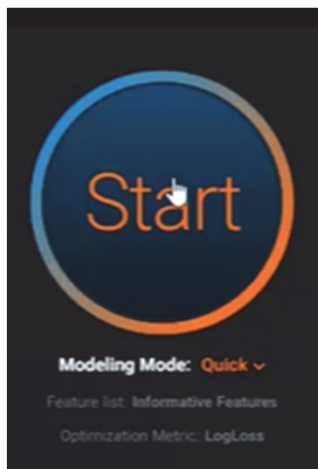
An example: I uploaded a Excel file, without providing any context, and asked three questions:

- Can you do visualizations and descriptive analyses to help me understand the data?
- Can you try regressions and look for patterns?"
- Can you run regression diagnostics?"

It did it all, interpreting the data and doing all of the work ... ”

– Ethan Mollick

<https://www.oneusefulthing.org/p/it-is-starting-to-get-strange>
<https://openai.com/blog/chatgpt-plugins#code-interpreter>



AI Use Cases : Invest Once, De

x0	x1
<h4>Chief of Staff</h4> <ul style="list-style-type: none">• Financial Agility with Re-Claiming Funds• Personnel/equipment demand• Predictive analytics for better recruiting• Workforce Management• Readiness	<h4>Personnel</h4> <ul style="list-style-type: none">• Employee / Recruit Retention• Recruitment Selection Criteria• Hiring Acceleration• Staffing Level Planning• PCS/Movement Planning• Travel Reimbursement



FIFE – Finite Interval Forecasting Engine



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?” or “Are the forecasts accurate?”



Assurance is a living concept meant to address some, but not all, of the DOD Foundational Tenets

Assurance

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.

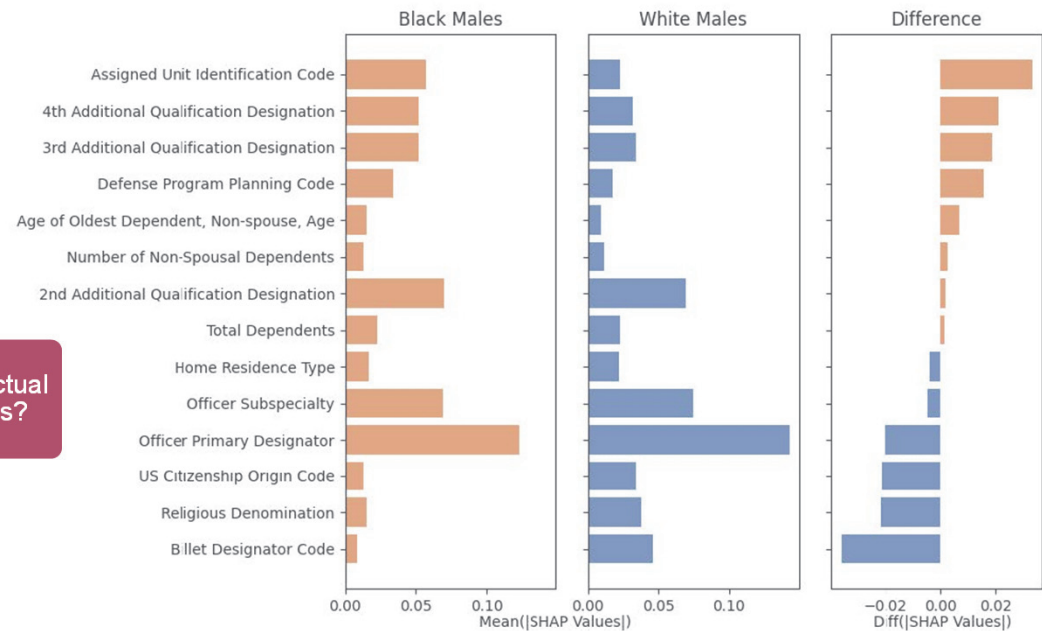
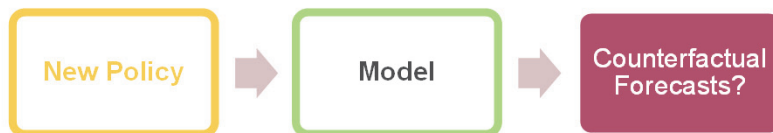
- It is easy to say what went wrong but hard to quantify up front.



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

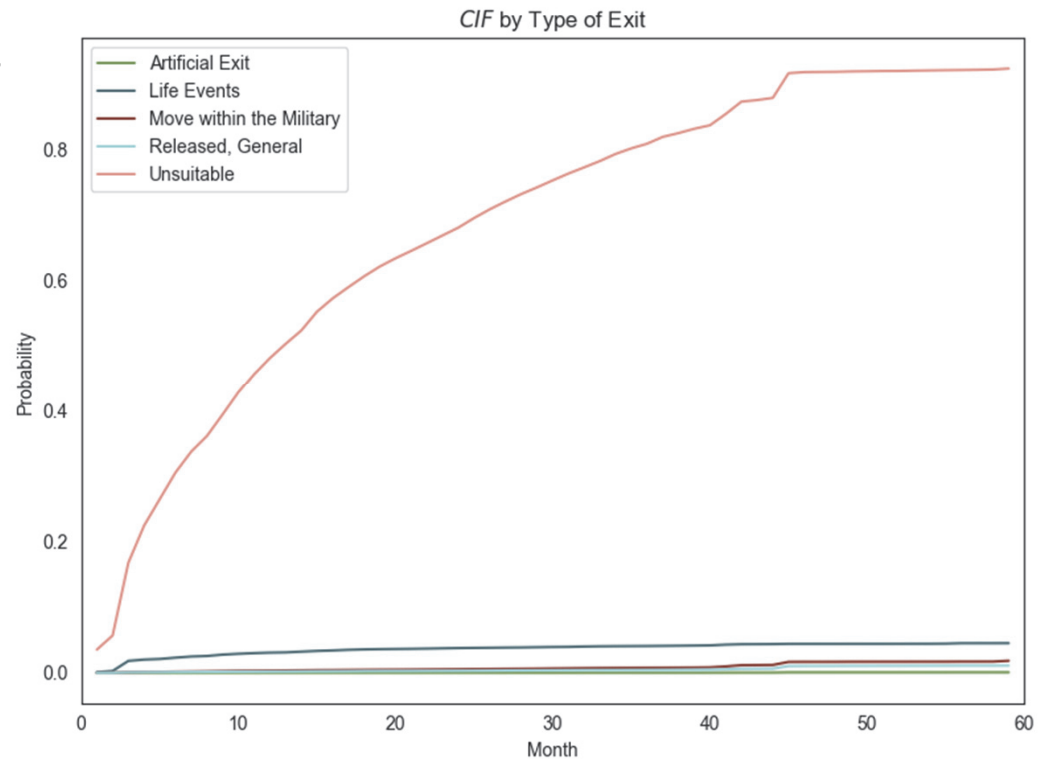
Personnel Space has Unique Concerns

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



Personnel Space has Unique Concerns

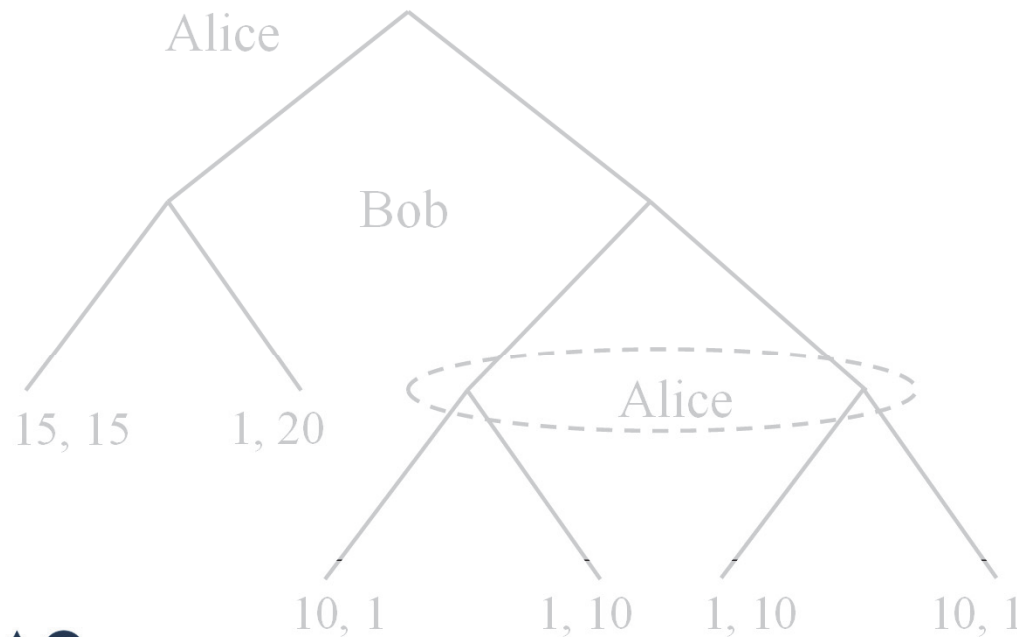
- Privacy
- Emergent service member behavior
- Perverse incentives



Personnel Space has Unique Concerns

- The data generating process is complex:
 - Strategic interactions
 - Selection
 - Unobserved heterogeneity
 - Network effects
 - ...

		Service Member	
		Stay	Leave
Peer	Stay	5, 5	1, 1
	Leave	1, 1	10, 10



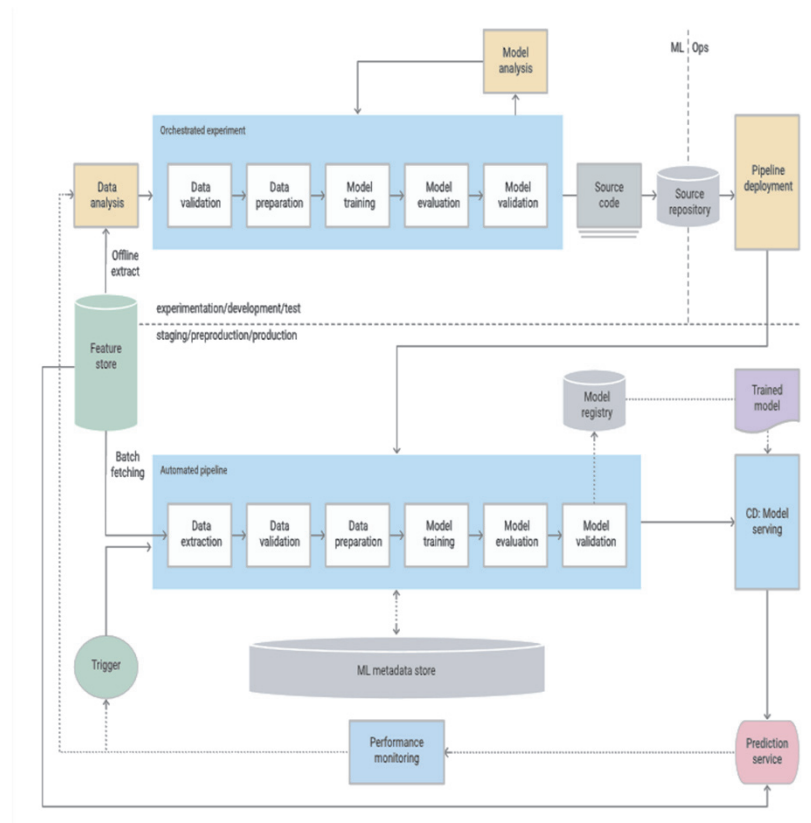
Assuring RAI in the Personnel Space

Assurance Guide

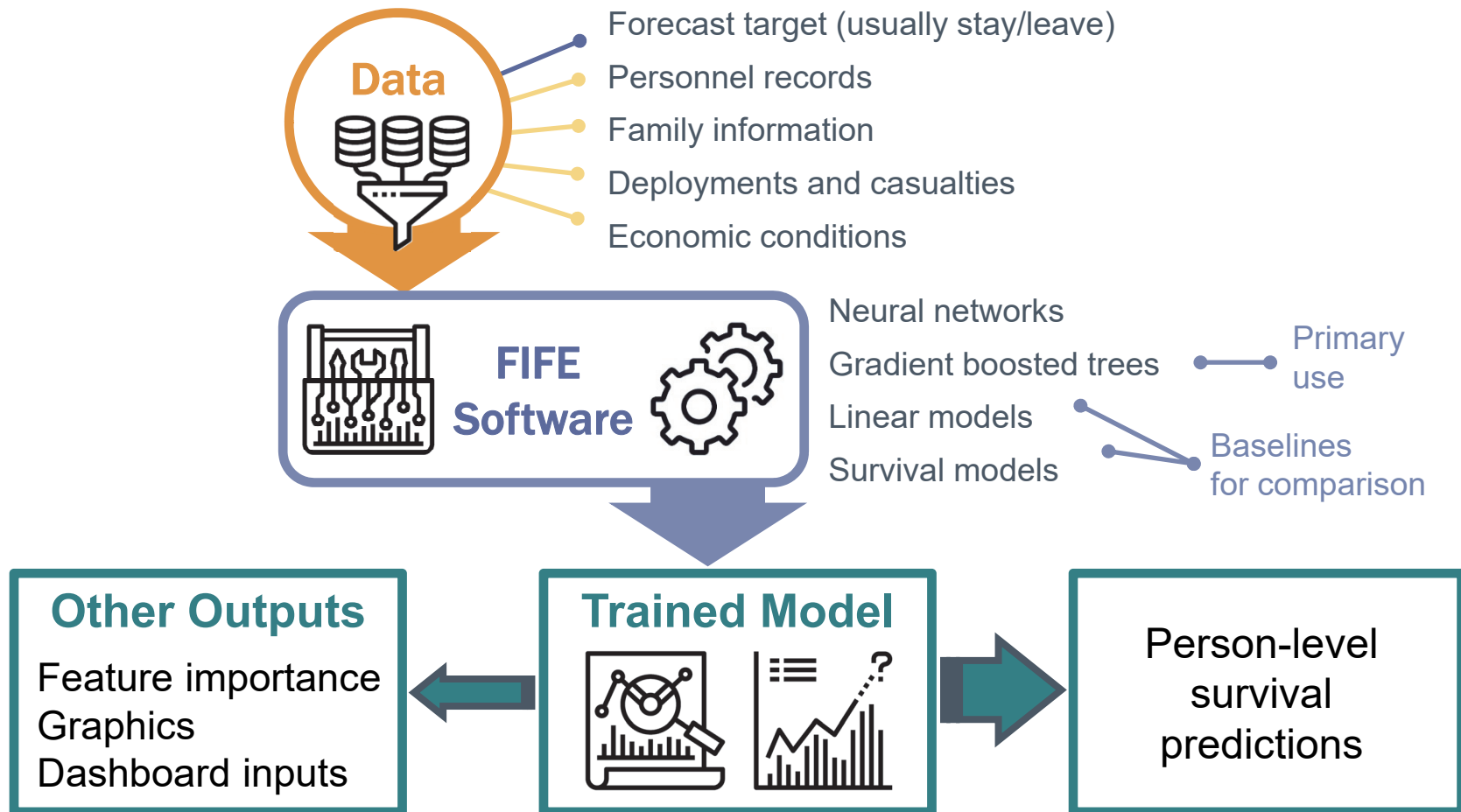
- MLOps scaffolding
- DOD 5 ethical principles
+ Privacy
- Personnel space nuance
- Strategies for testing, monitoring, feedback, etc.

Assurance Case

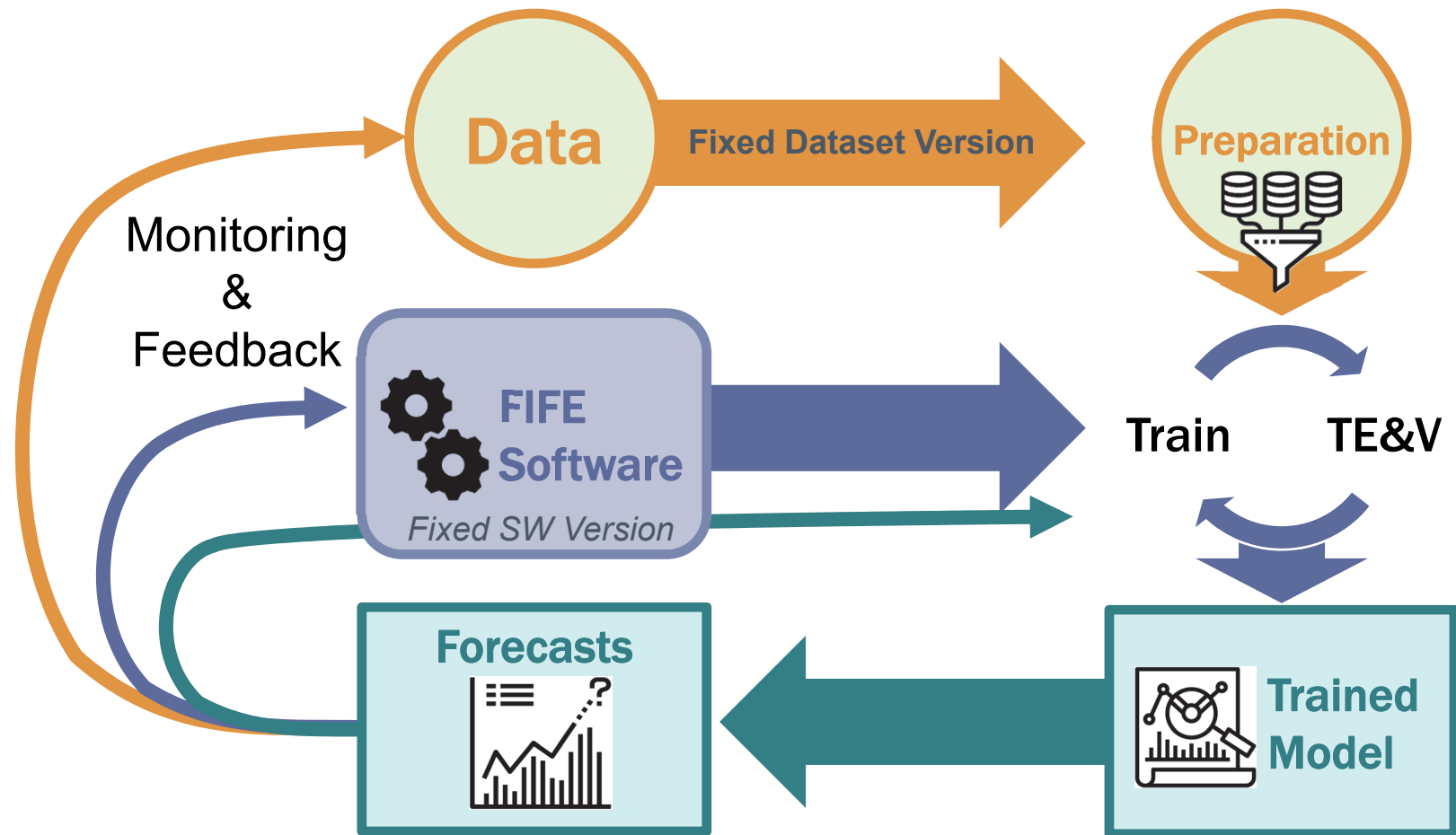
- Application of the guide to a **Army Retention Prediction Model (RPM)**



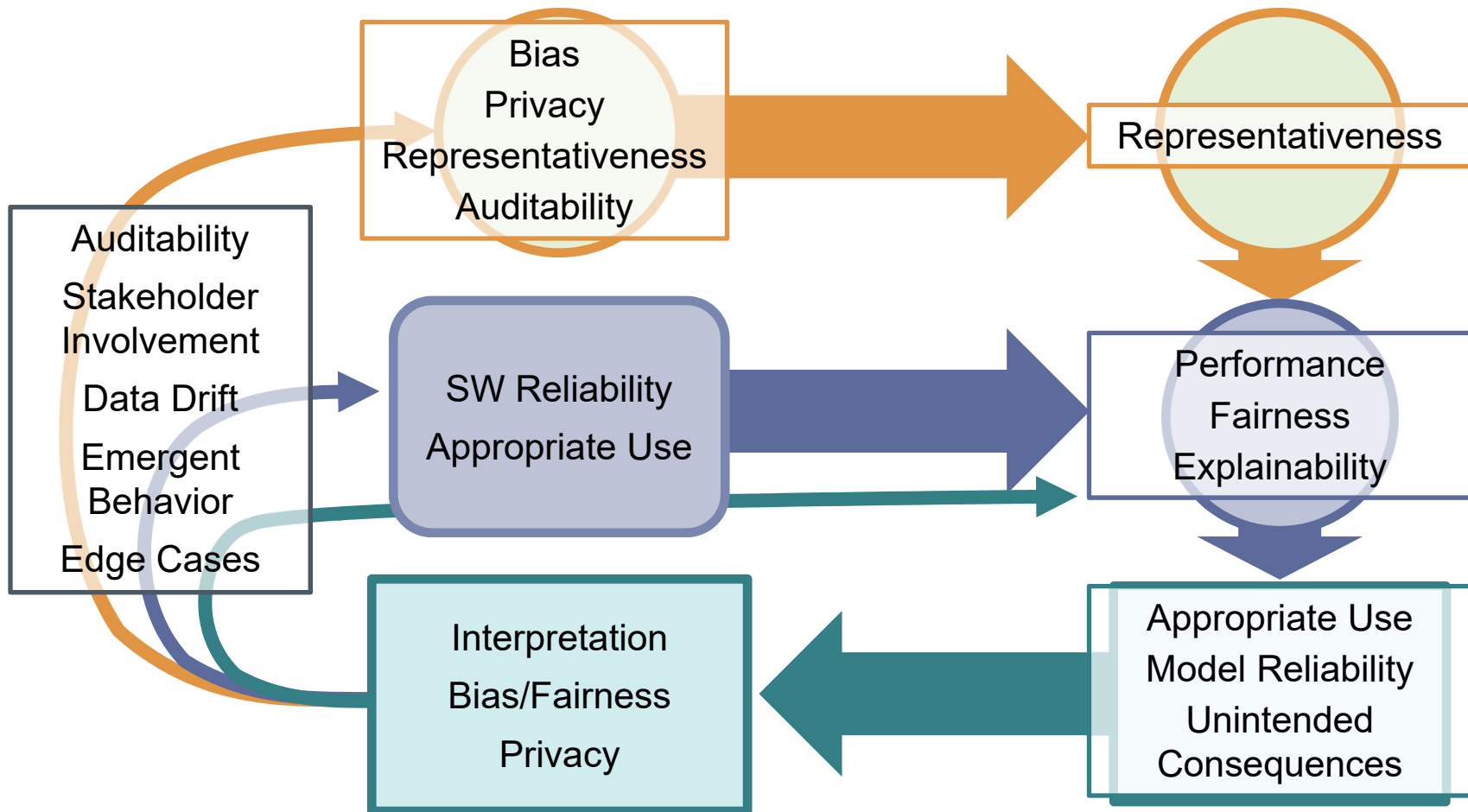
Use Case - Retention Prediction Model (RPM)-Army



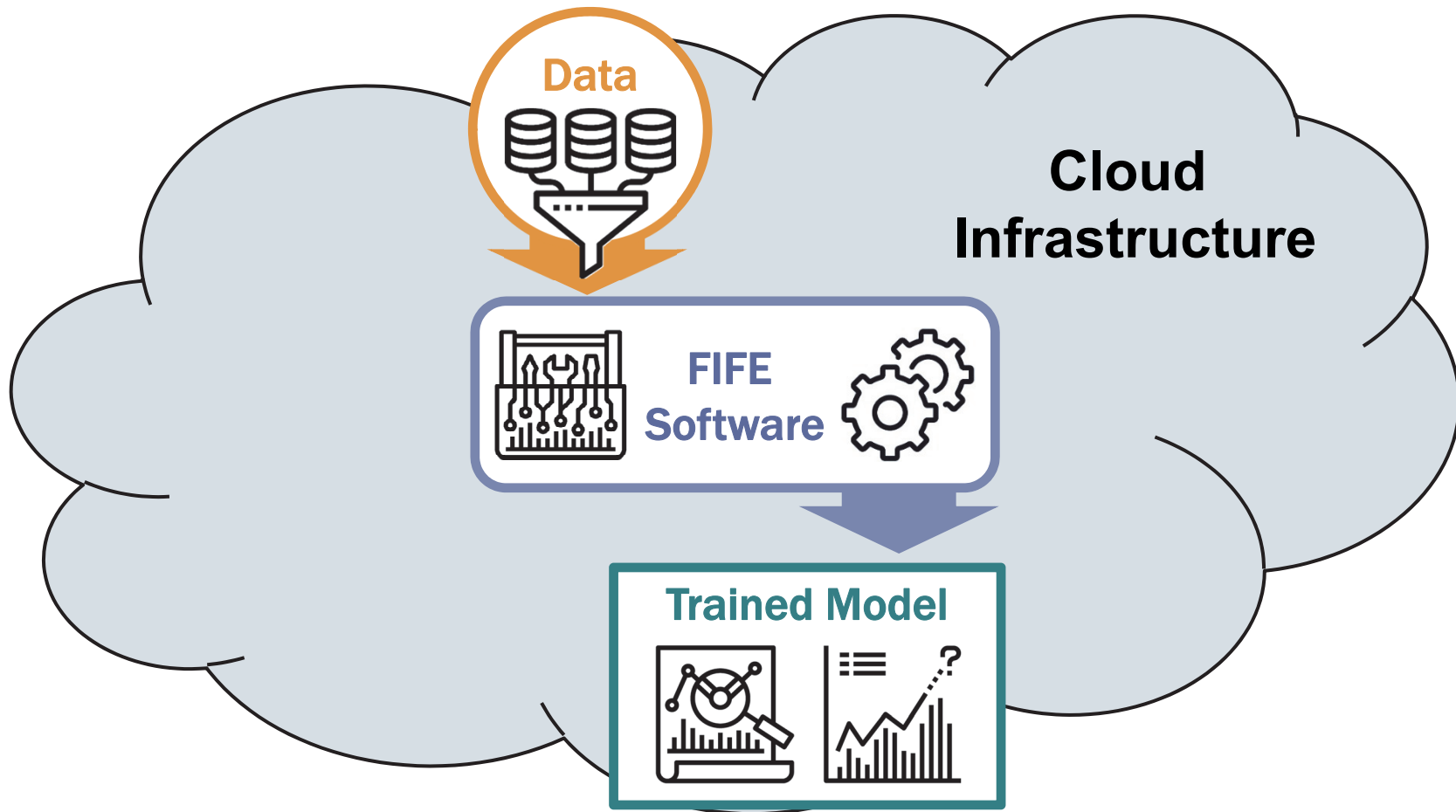
Model Lifecycle



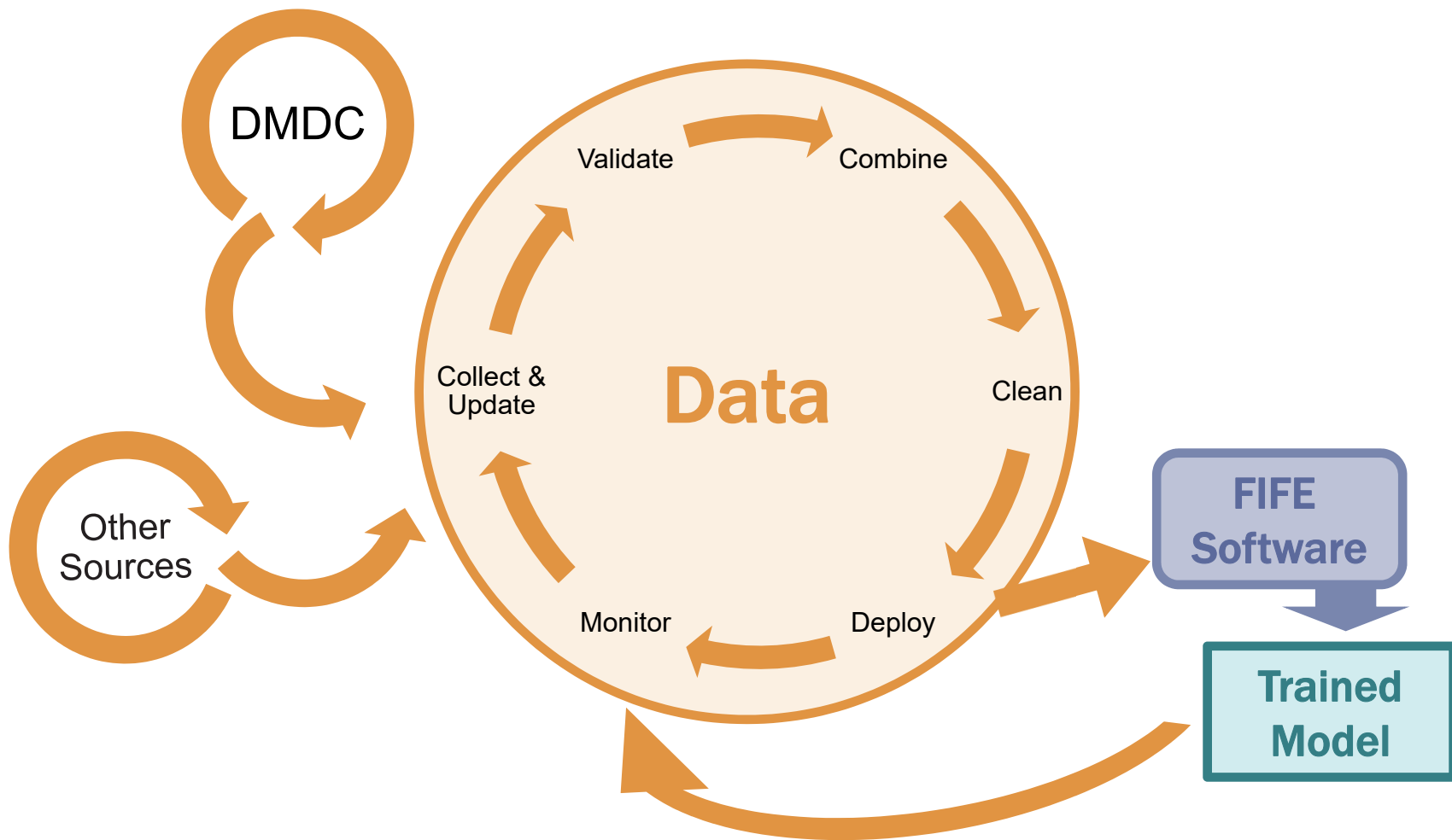
RAI in the Lifecycle



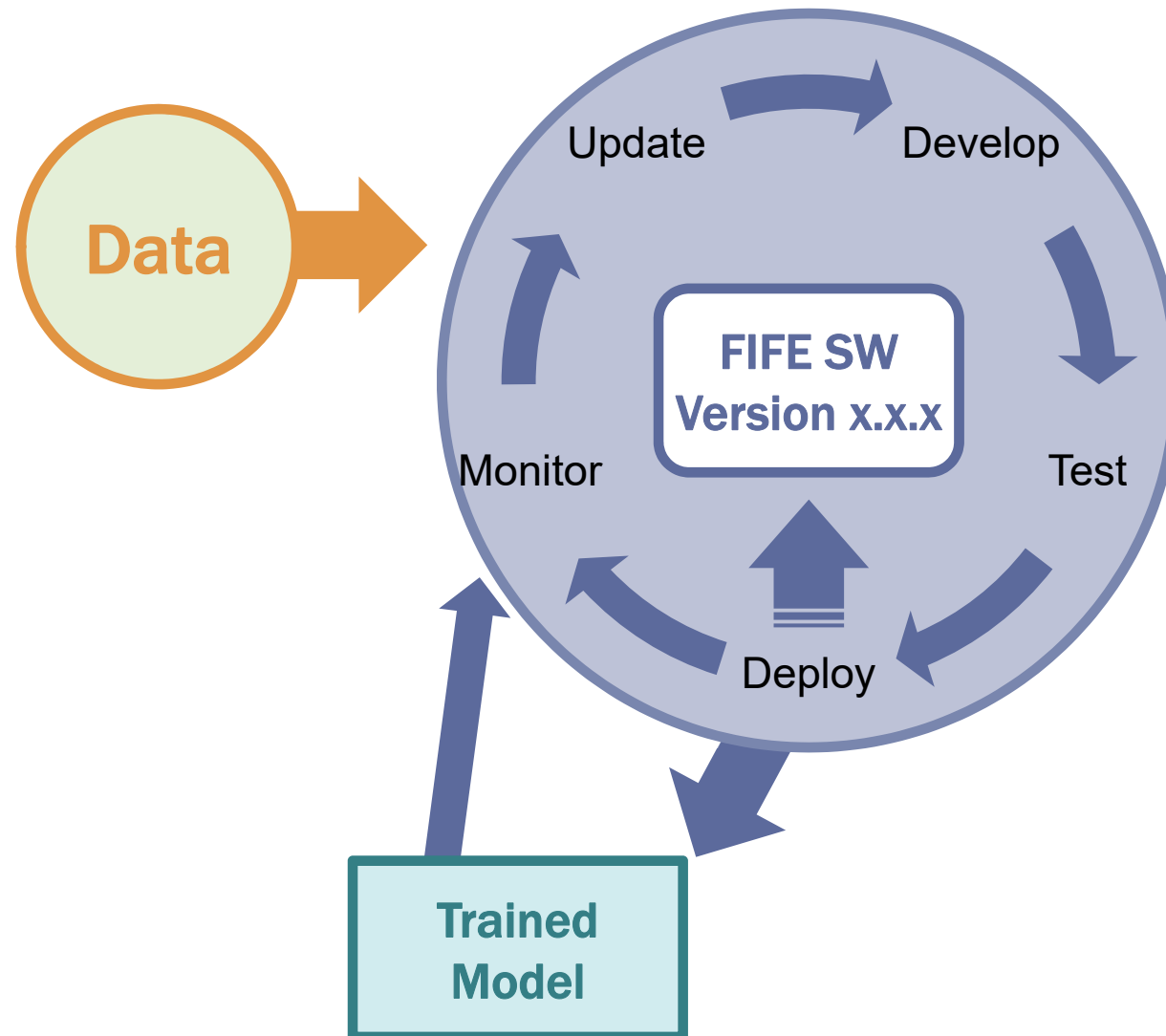
Ecosystem



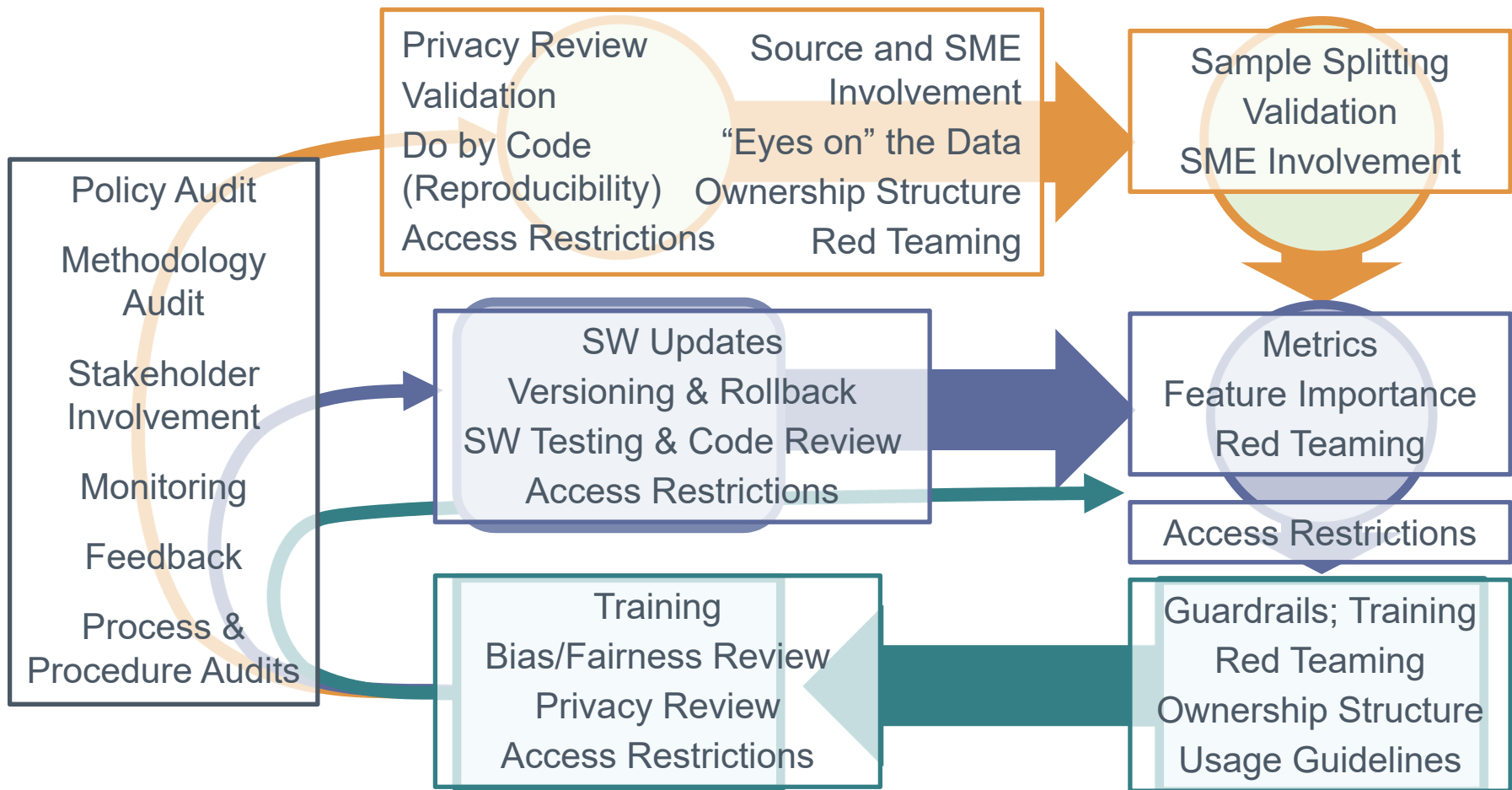
Data Curation Lifecycle



FIFE Software Development Lifecycle



Assurance Mechanisms



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
- Challenges remain

IDA

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.
 - Traditional 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

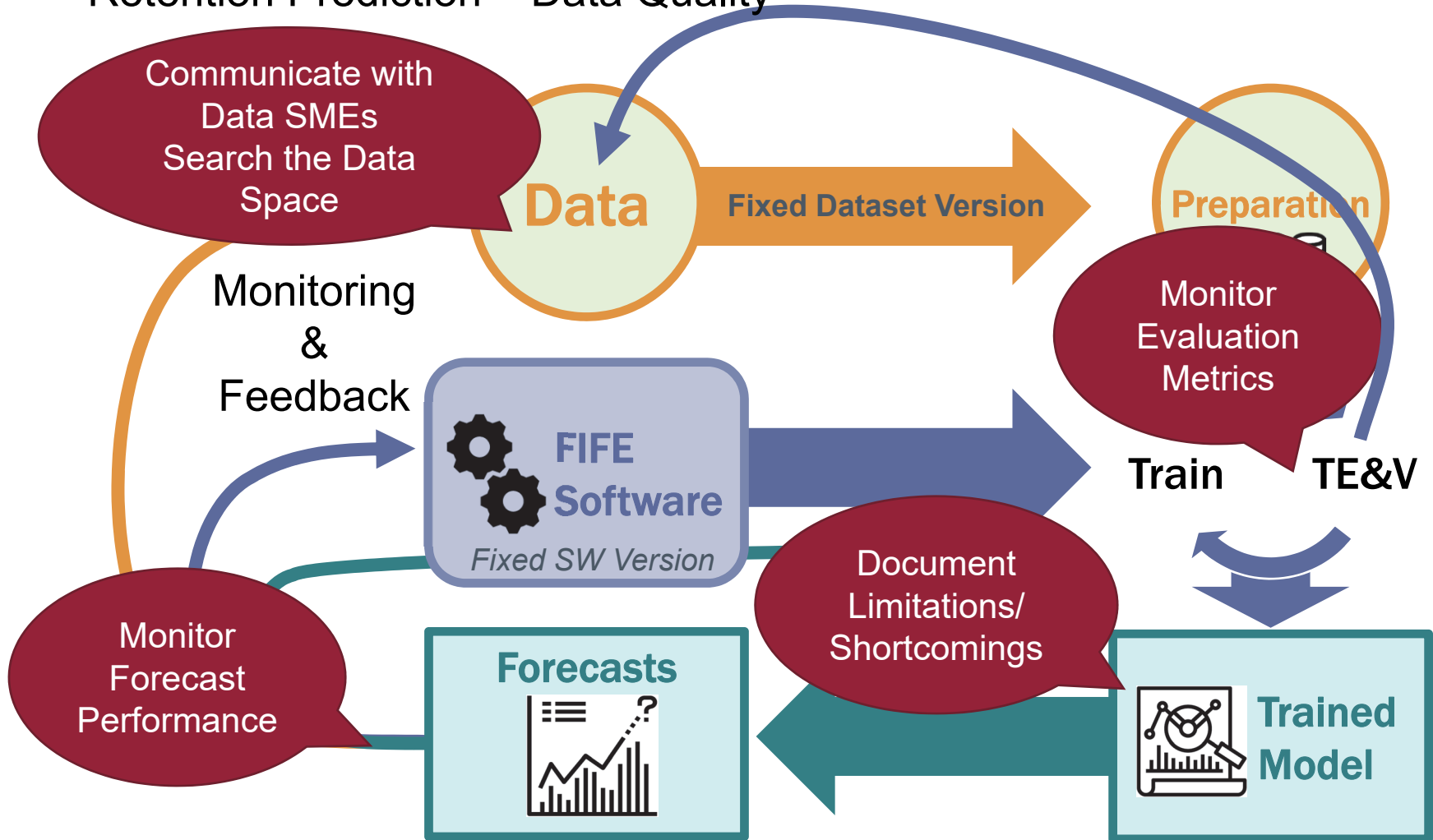
- Predictive Retention Toolkit and Evaluation for Targeted Army Talent Management
- Overarching question: How can the Army best select, shape, train, and retain the force it wants?
- Three-part study aimed at retention efforts:
 1. Forecast retention with high fidelity and accuracy
 2. Discover indicators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

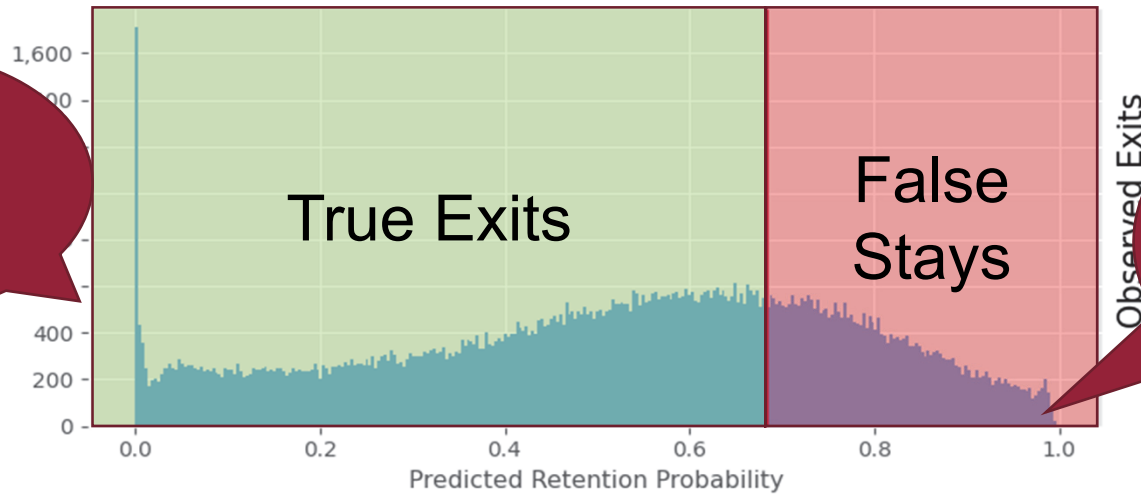


Example - Metrics

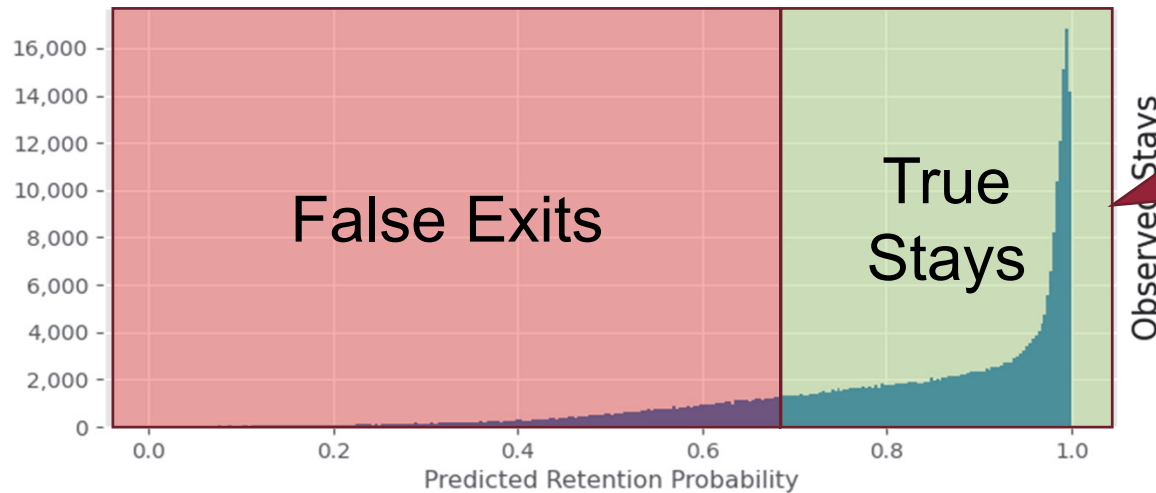
Retention Prediction – Data Quality

Predict Exit \leftarrow \rightarrow Predict Stay

Forecasting Exits is difficult.



Prediction distribution is fairly flat, peaked toward "stay."



Forecasting Stays is easy.

IDA

jdennis@ida.org

Work funded by



CDAO
AI Assurance

REPORT DOCUMENTATION PAGE*Form Approved
OMB No. 0704-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)		2. REPORT TYPE		3. DATES COVERED (From - To)	
4. TITLE AND SUBTITLE				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)				8. PERFORMING ORGANIZATION REPORT NUMBER	
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					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (Include area code)