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

AI – Artificial Intelligence; RAI – Responsible AI; T&E – Test and Evaluation

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

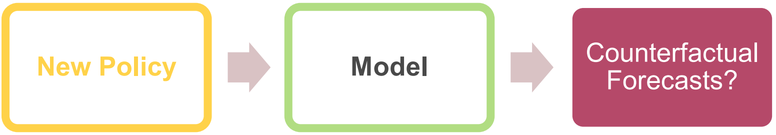
- Risks are often lower profile
- DOD personnel environment is very large
- Often less complex involvement of AI/ML on smaller budgets
- AI/ML is "easy"

But

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

Personnel Space has Unique Concerns

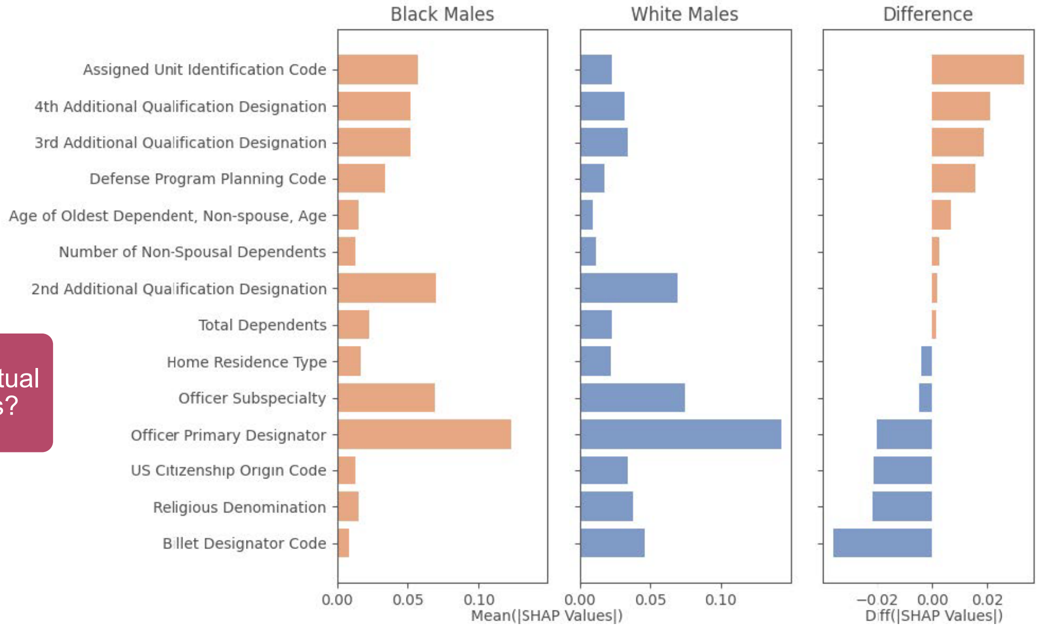
- Disparate impact/treatment
- Invalid prospective policy analysis (invalid counterfactuals!)
- Misattributed causality



New Policy

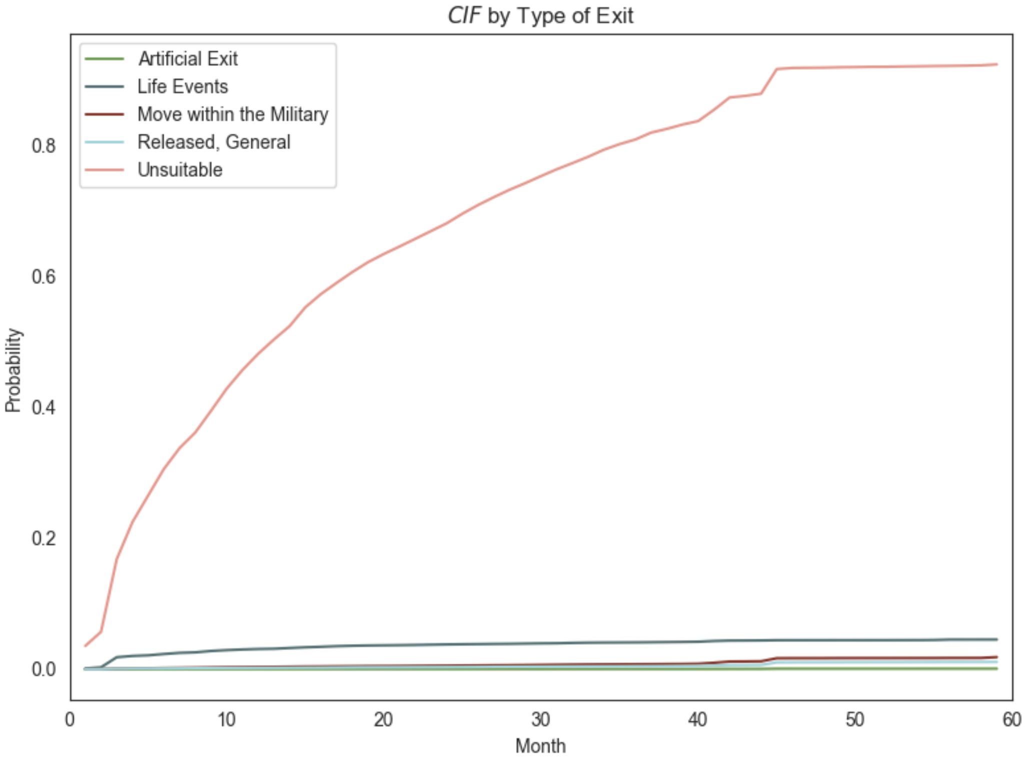
Model

Counterfactual Forecasts?



Personnel Space has Unique Concerns

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



CIF – Cumulative Incidence Function

Assuring RAI in the Personnel Space

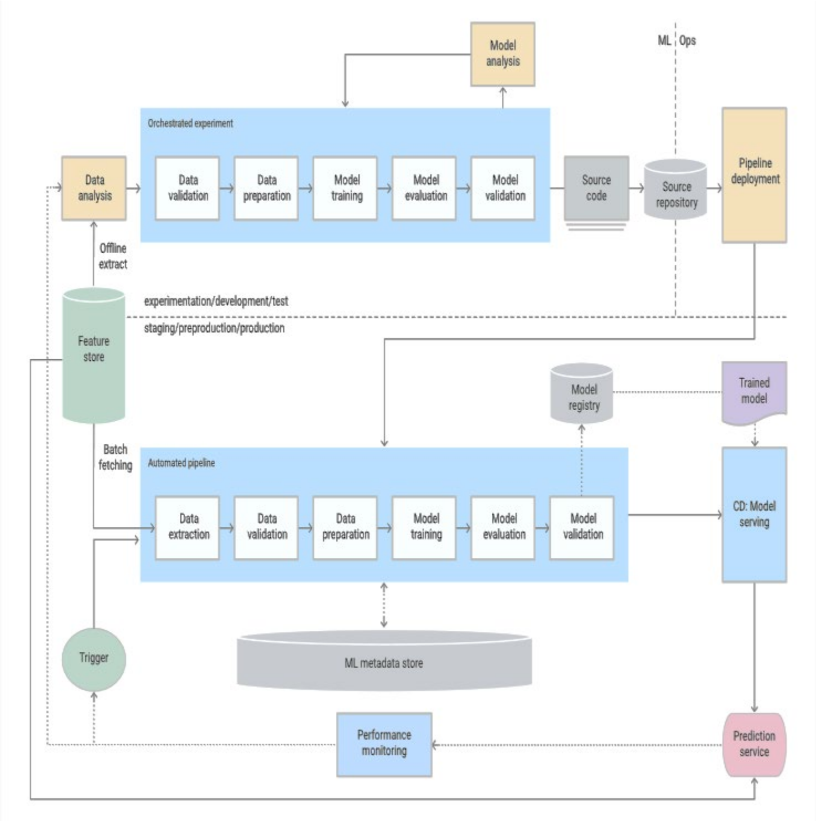
Assurance Guide

- MLOps scaffolding
- DOD 5 ethical principles

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

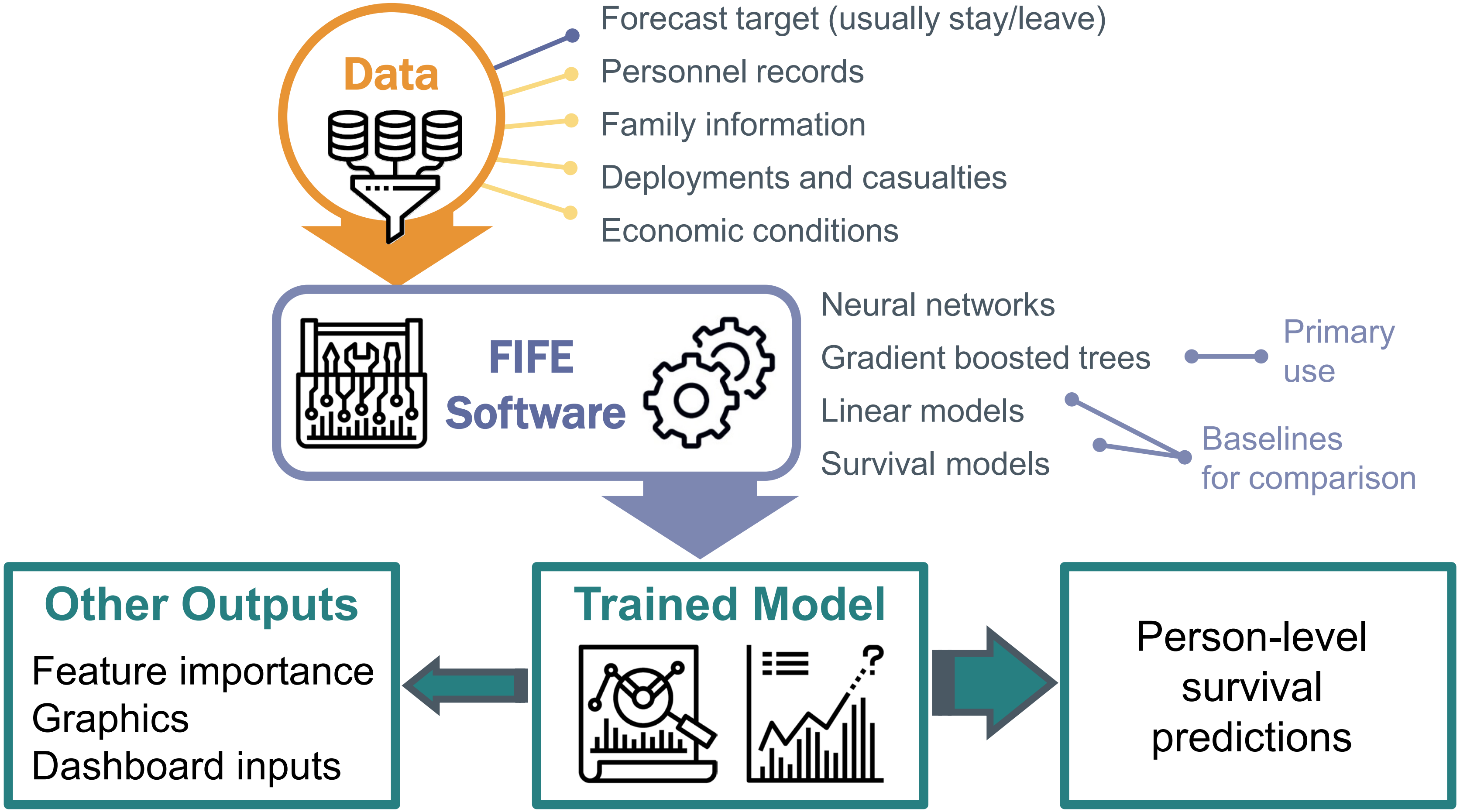
Assurance Case

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



MLOps - Machine Learning Operations

Use Case - Retention Prediction Model (RPM)-Army



Data

- Forecast target (usually stay/leave)
- Personnel records
- Family information
- Deployments and casualties
- Economic conditions

FIFE Software

- Neural networks
- Gradient boosted trees
- Linear models
- Survival models

Primary use

Baselines for comparison

Other Outputs

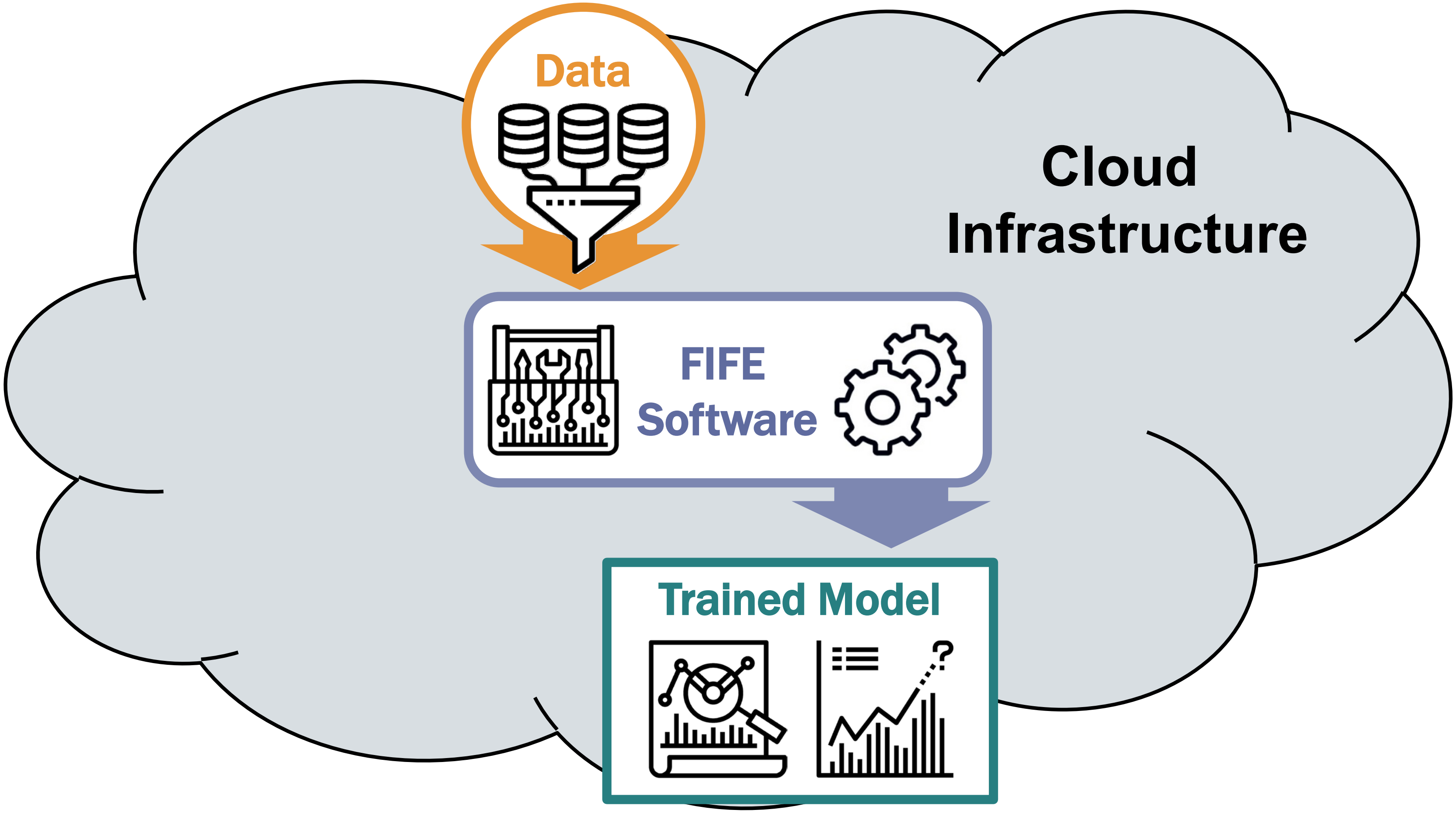
Feature importance
Graphics
Dashboard inputs

Trained Model

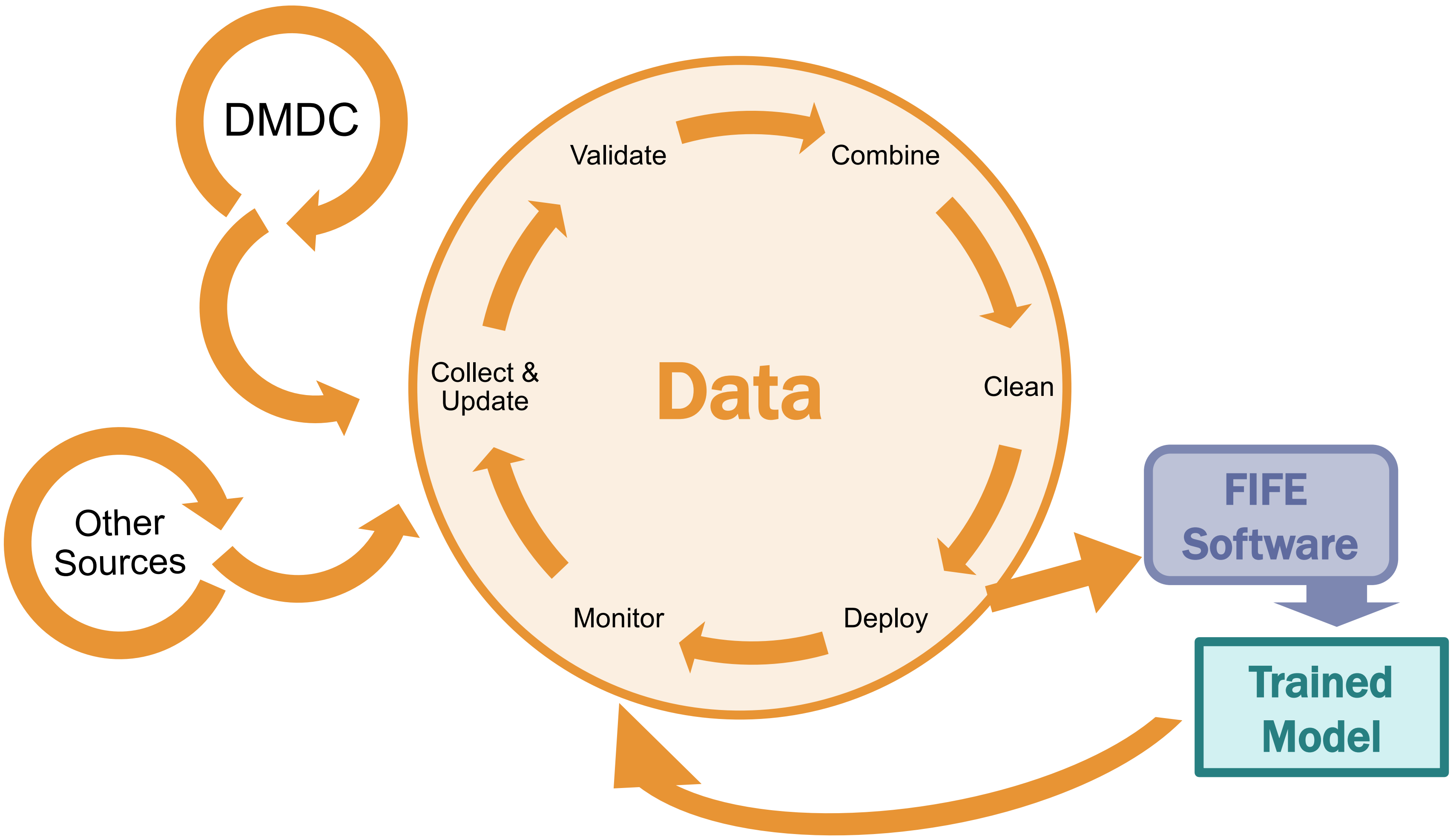
Person-level survival predictions

FIFE - Finite Interval Forecasting Engine

Ecosystem

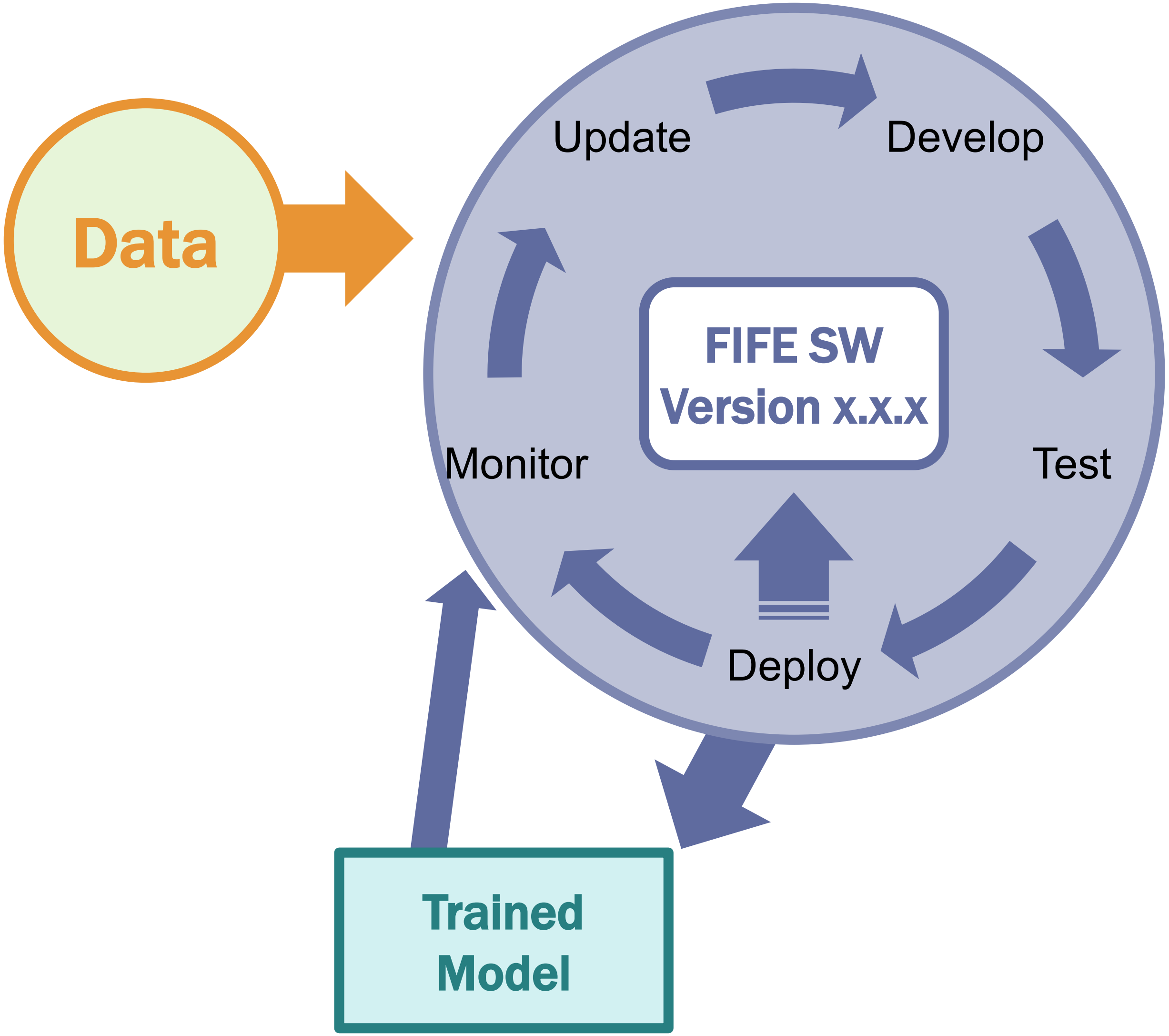


Data Curation Lifecycle



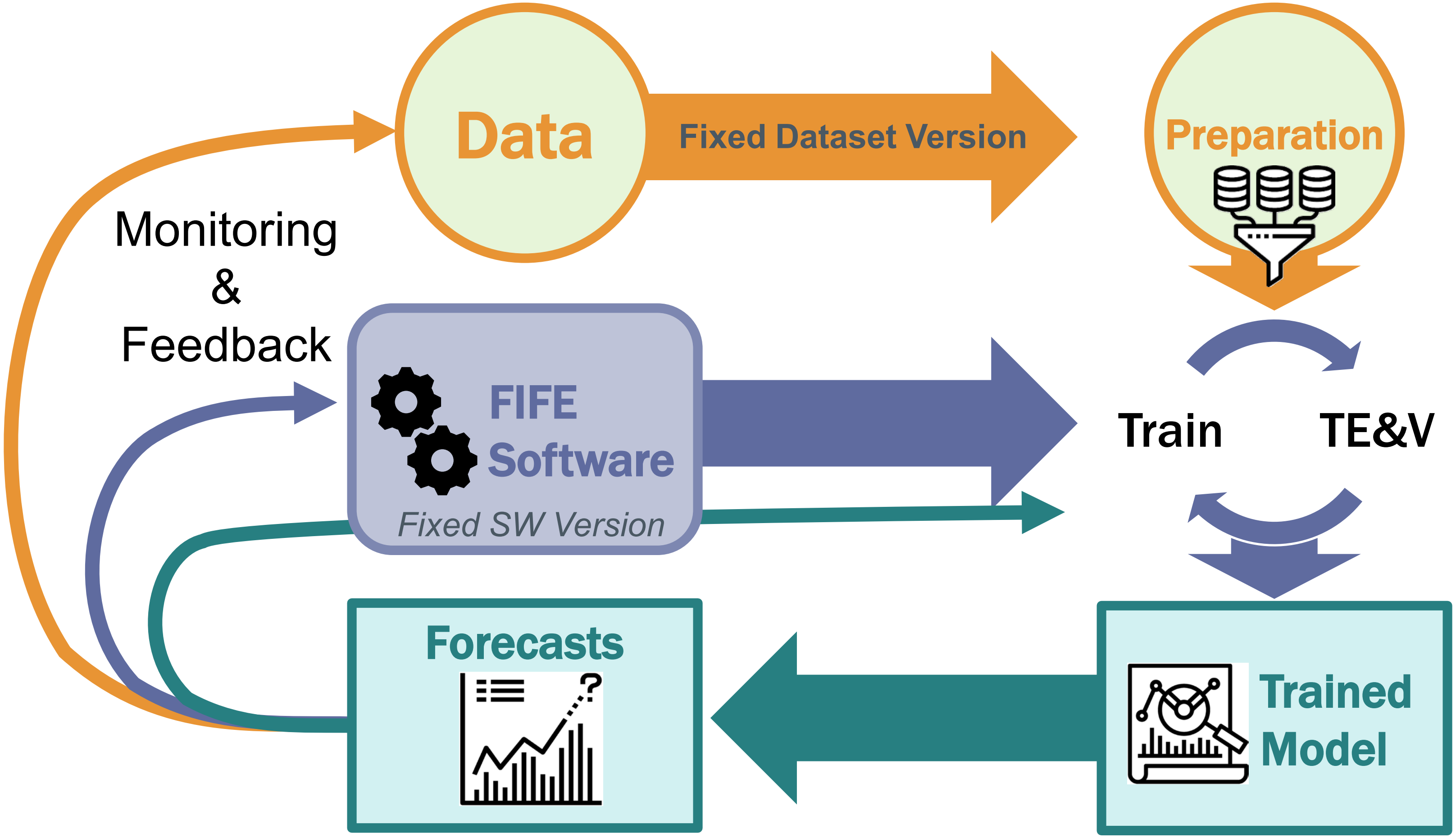
FIFE - Finite Interval Forecasting Engine

FIFE Software Development Lifecycle



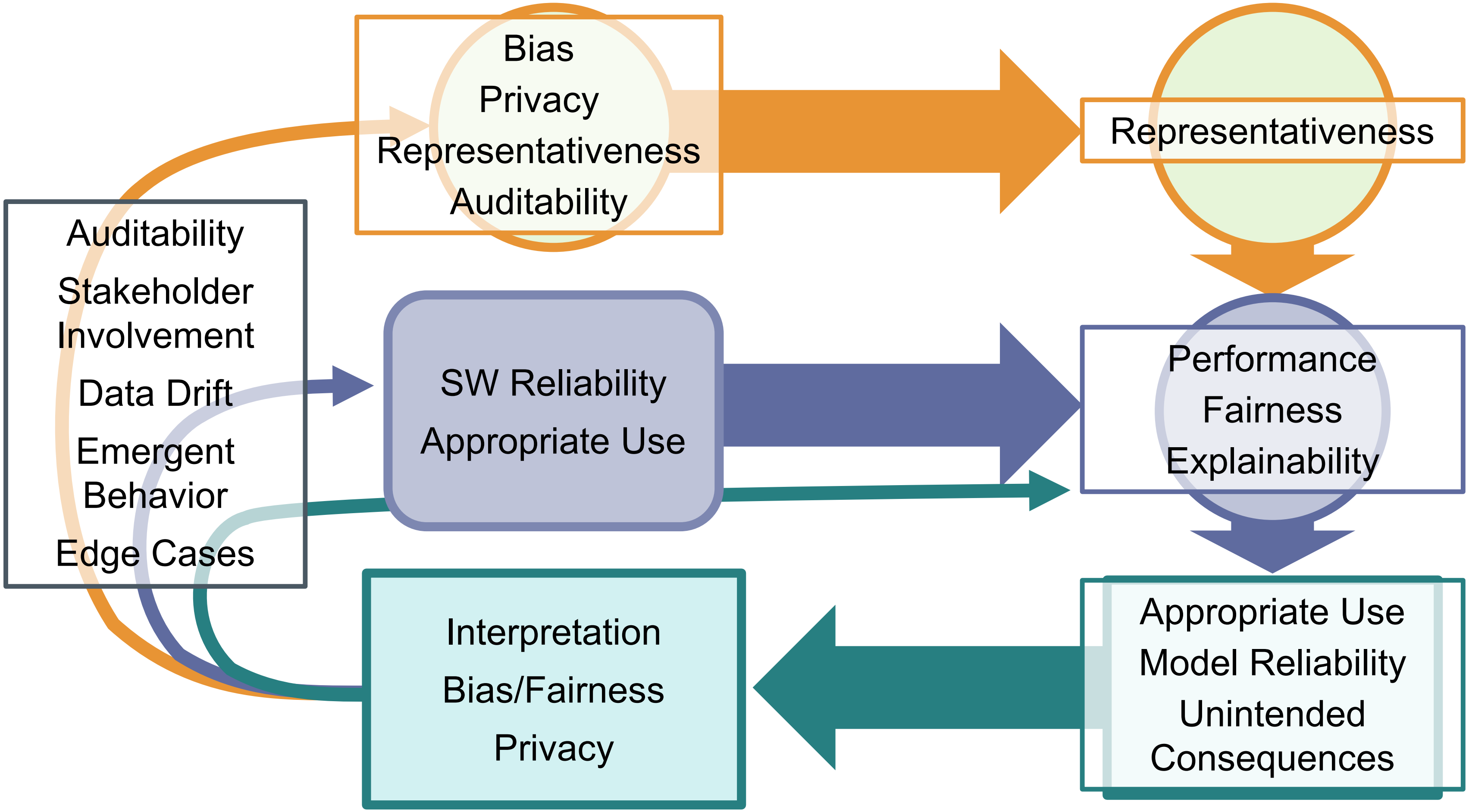
Note that the SW Lifecycle and Model Lifecycle have touchpoints but otherwise are distinct processes!

Model Lifecycle



Note that the SW Lifecycle and Model Lifecycle have touchpoints but otherwise are distinct processes!

RAI in the Lifecycle



Bias
Privacy
Representativeness
Auditability

Representativeness

Auditability
Stakeholder Involvement
Data Drift
Emergent Behavior
Edge Cases

SW Reliability
Appropriate Use

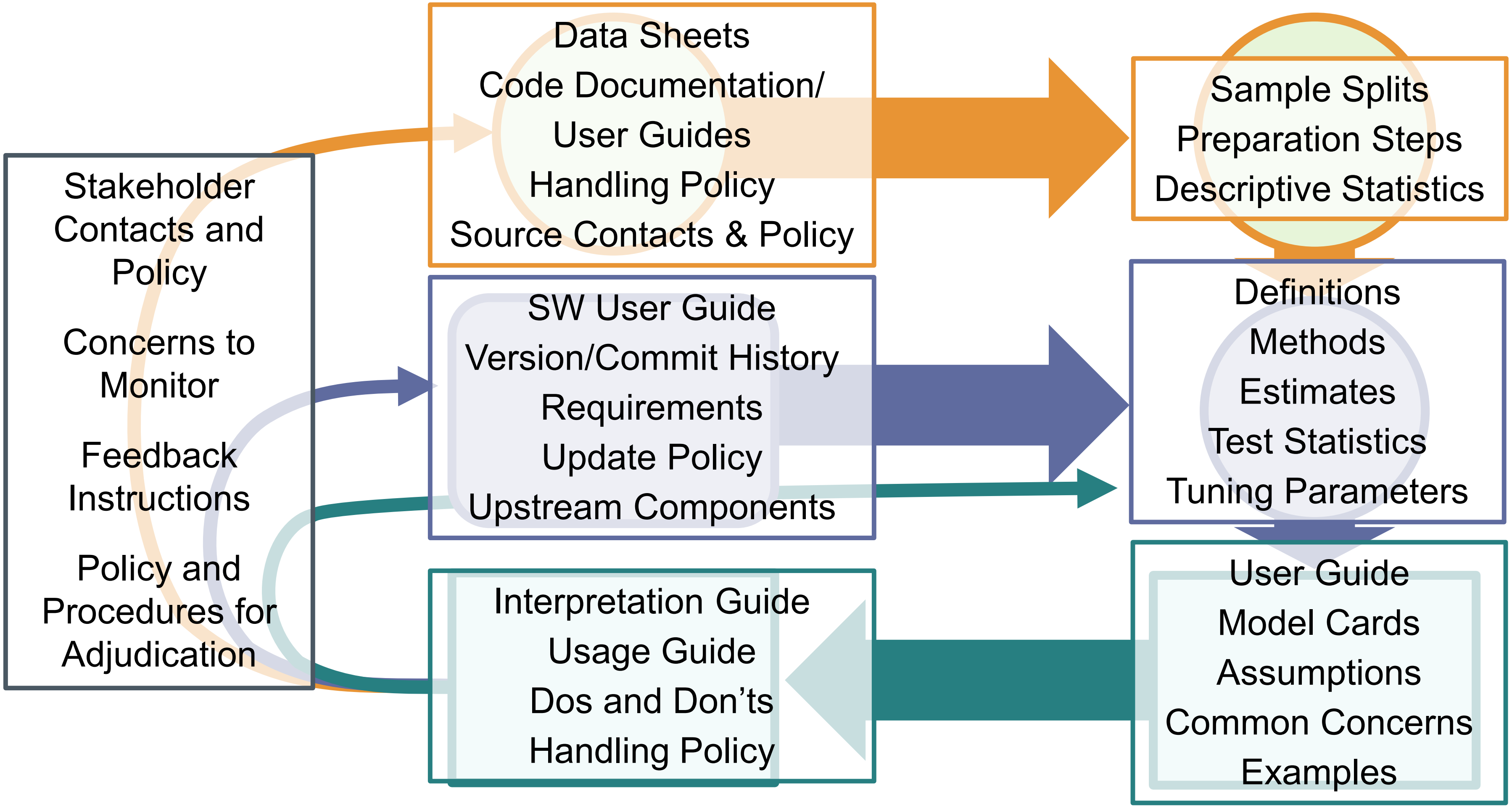
Performance
Fairness
Explainability

Interpretation
Bias/Fairness
Privacy

Appropriate Use
Model Reliability
Unintended Consequences

List is illustrative but not exhaustive.
SW - Software

Documentation



Stakeholder Contacts and Policy

Concerns to Monitor

Feedback Instructions

Policy and Procedures for Adjudication

Data Sheets
Code Documentation/
User Guides
Handling Policy
Source Contacts & Policy

Sample Splits
Preparation Steps
Descriptive Statistics

SW User Guide
Version/Commit History
Requirements
Update Policy
Upstream Components

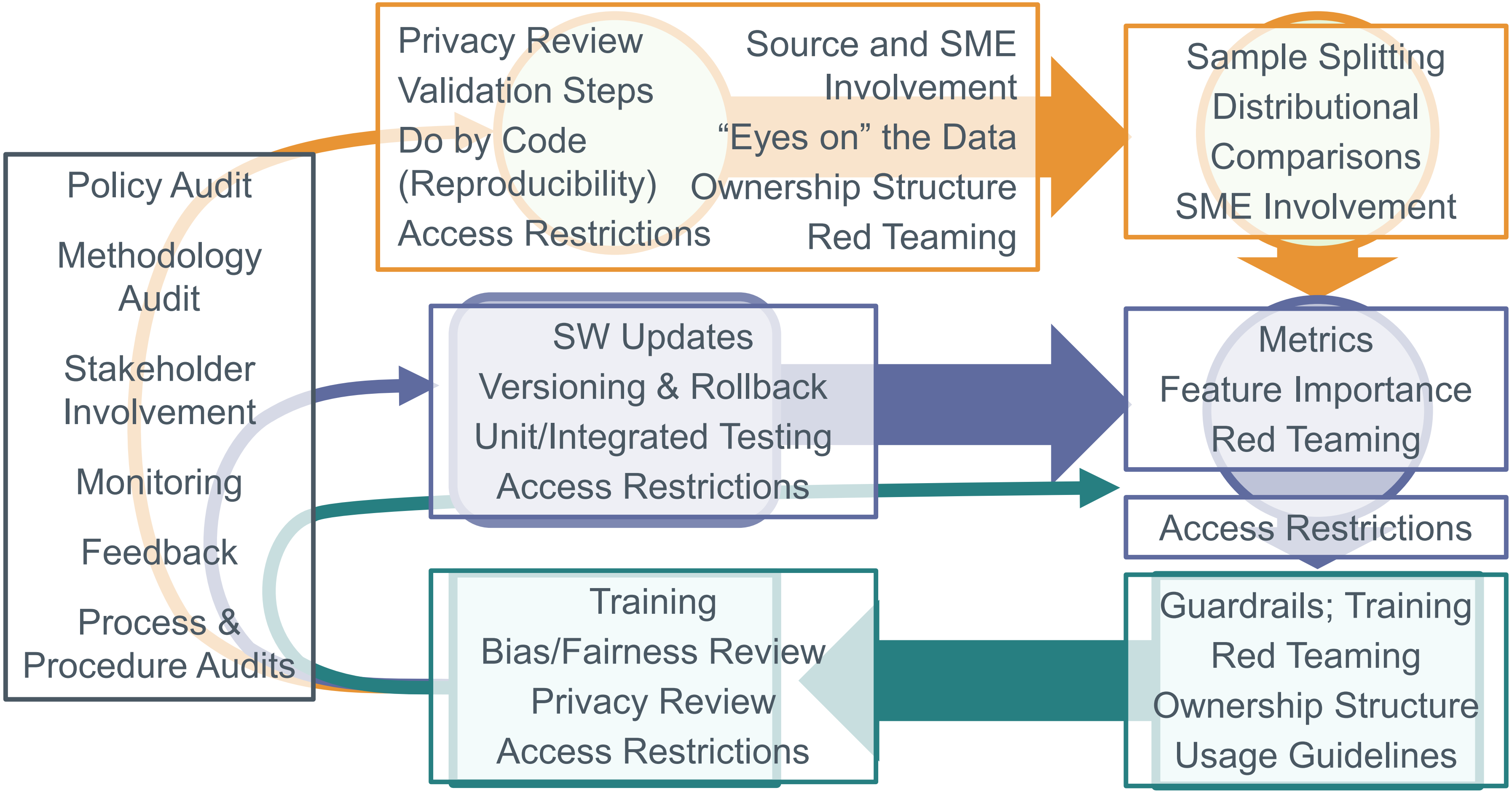
Definitions
Methods
Estimates
Test Statistics
Tuning Parameters

Interpretation Guide
Usage Guide
Dos and Don’ts
Handling Policy

User Guide
Model Cards
Assumptions
Common Concerns
Examples

List is illustrative but not exhaustive.

Assurance Mechanisms



Policy Audit

Methodology Audit

Stakeholder Involvement

Monitoring

Feedback

Process & Procedure Audits

Privacy Review
Validation Steps
Do by Code (Reproducibility)
Access Restrictions

Source and SME Involvement
“Eyes on” the Data
Ownership Structure
Red Teaming

Sample Splitting
Distributional Comparisons
SME Involvement

SW Updates
Versioning & Rollback
Unit/Integrated Testing
Access Restrictions

Metrics
Feature Importance
Red Teaming

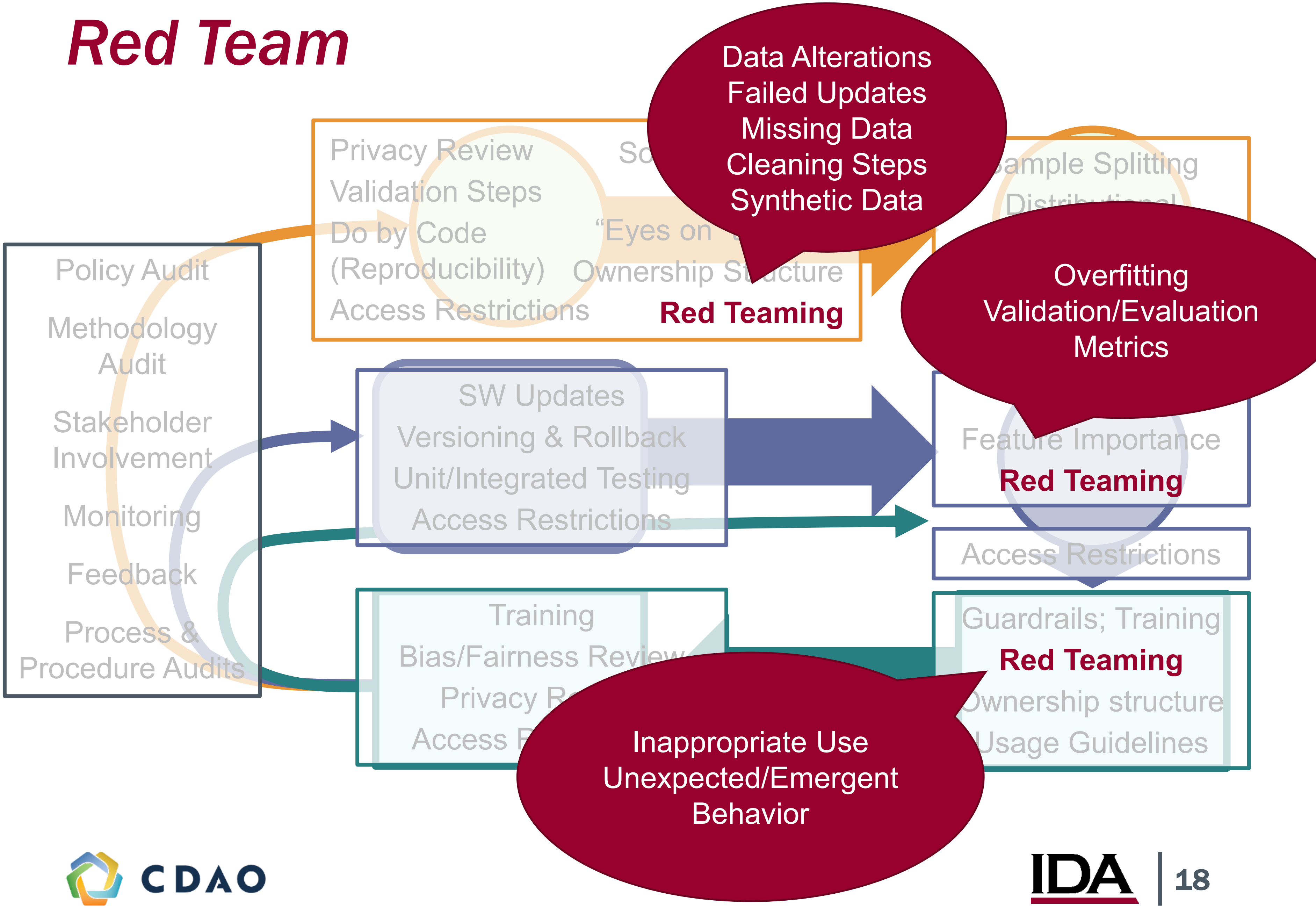
Access Restrictions

Training
Bias/Fairness Review
Privacy Review
Access Restrictions

Guardrails; Training
Red Teaming
Ownership Structure
Usage Guidelines

List is illustrative but not exhaustive.

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

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.
  - 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

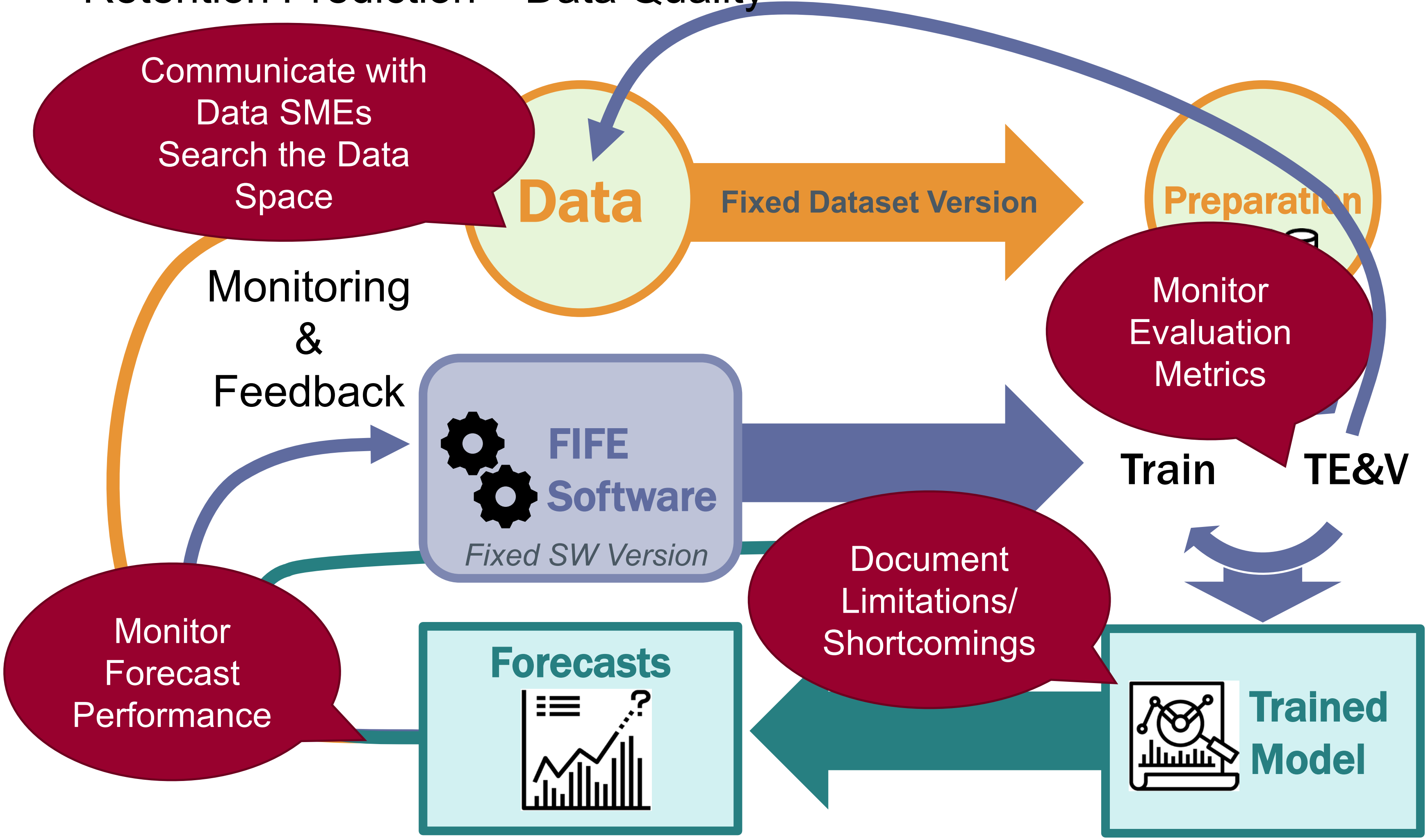
Forecast Retention with High Fidelity and Accuracy

- 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

* https://github.com/IDA-HumanCapital/fife
https://github.com/IDA-HumanCapital/fife_for_spark

Example – Model Lifecycle

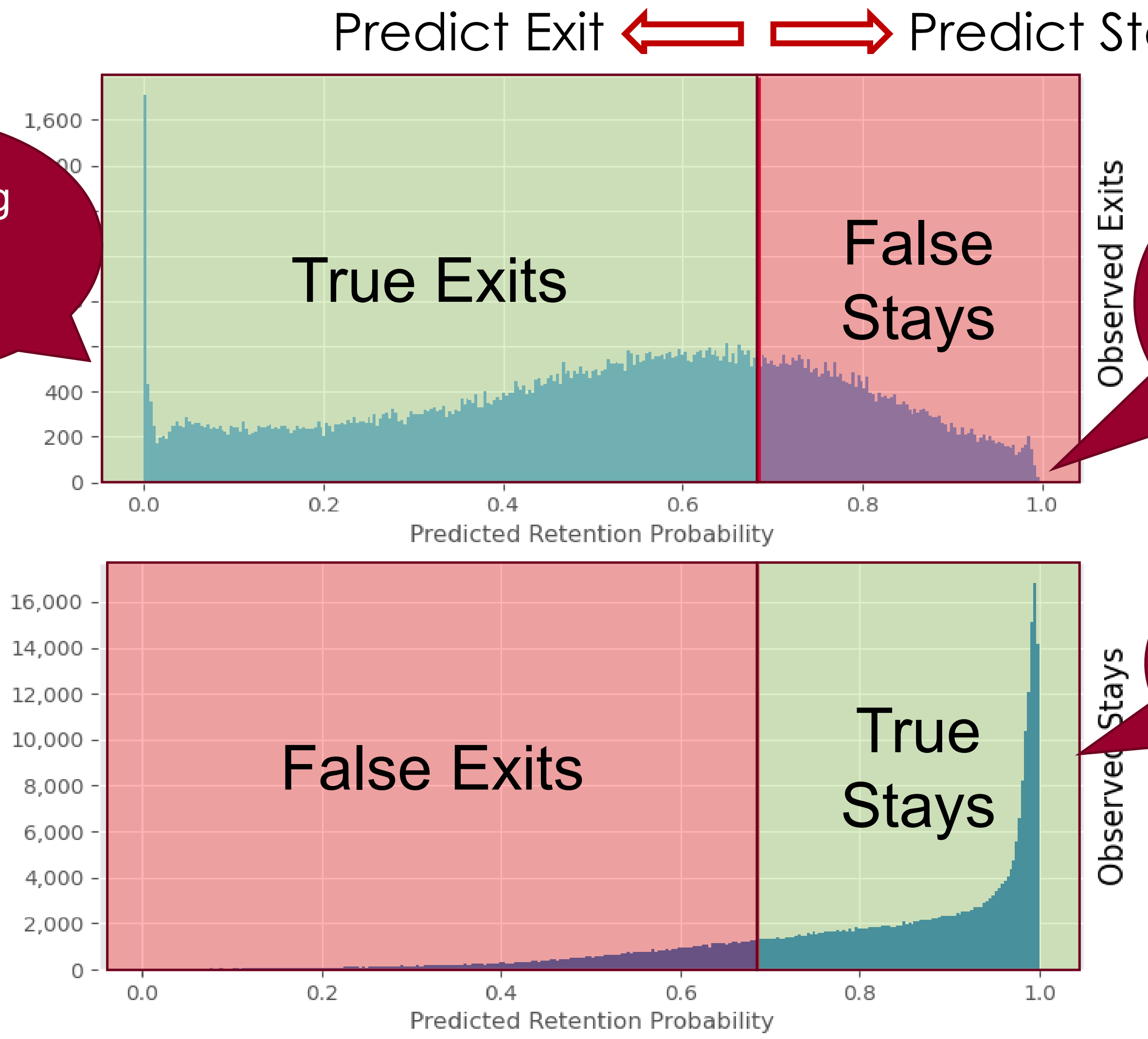
Retention Prediction – Data Quality


Communicate with Data SMEs Search the Data Space
Data
Fixed Dataset Version
Preparation
Monitor Evaluation Metrics
Monitoring & Feedback
FIFE Software
Fixed SW Version
Train
TE&V
Document Limitations/ Shortcomings
Monitor Forecast Performance
Forecasts
Trained Model

Note that the SW Lifecycle and Model Lifecycle have touchpoints but otherwise are distinct processes!

Example - Metrics

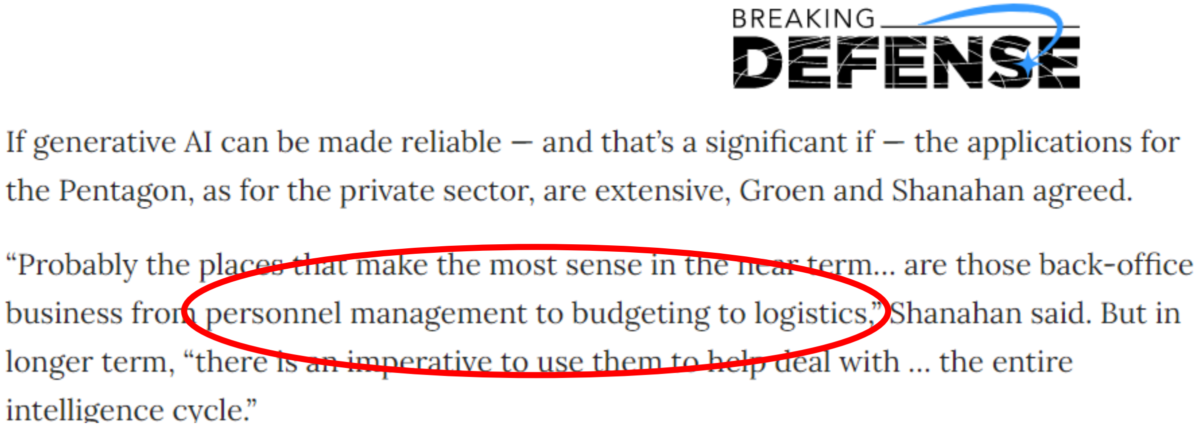
Retention Prediction – Data Quality



Forecasting Exits is difficult

Prediction distribution is fairly flat, peaked toward "stay"

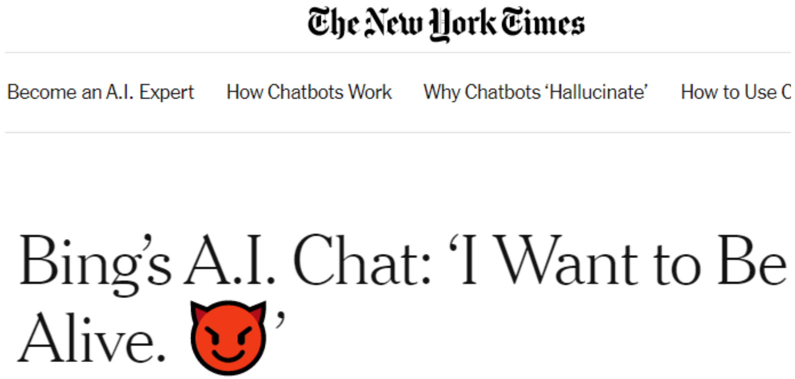
Forecasting Stays is easy



BREAKING DEFENSE

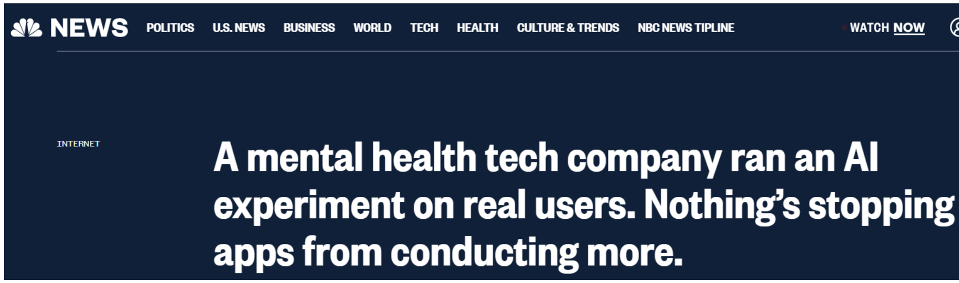
If generative AI can be made reliable — and that's a significant if — the applications for the Pentagon, as for the private sector, are extensive, Groen and Shanahan agreed.

"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

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



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

CDAO

IDA

jdennis@ida.org

Work funded by CDAO AI Assurance