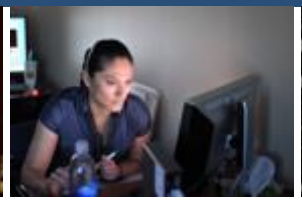
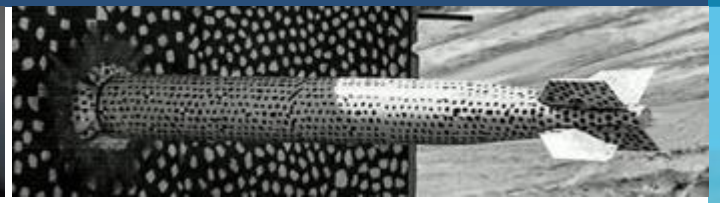
Increasing Integration of Data-driven Analyses in Operational Activities through Knowledge Management





Thushara Gunda, Munaf Aamir, Sasha Outkin

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND2021-6437 C

Acknowledgements

Natalia Herrera

Arthur Willhite

Lynita Bosey

Nenita Walther

Sandra Gonzales

Patrick Carlson

Adrian Perez

Casey Richardson

Brian Anderson

Adam Neal

Nicole Jackson

Outline

Motivation

Problem Definition

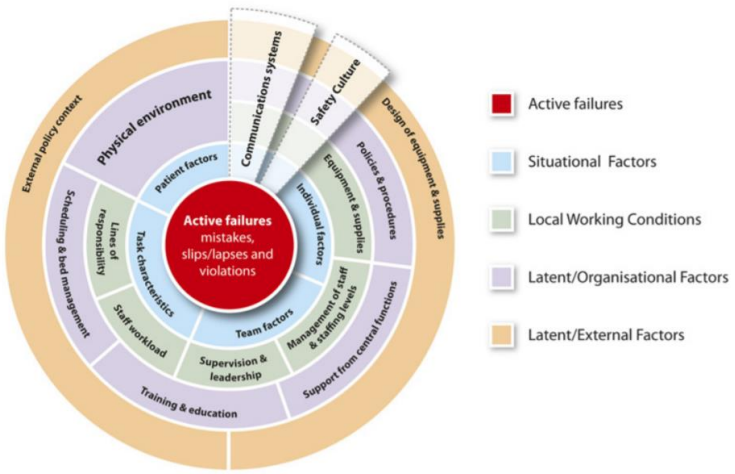
Methodology & Results

- Data Collection and Processing
- Machine Learning
- Data Visualizations

Concluding Remarks

Motivation

- Reducing the frequency and severity of incidents of security concern (IOSC) is a priority for many institutions
- Much has been to done to develop analytics for characterization or prediction of IOSCs
- Challenges remain for integration of analytics into operational activities



Lawton et al., 2012

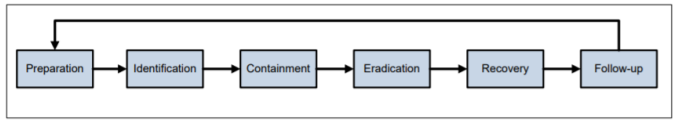
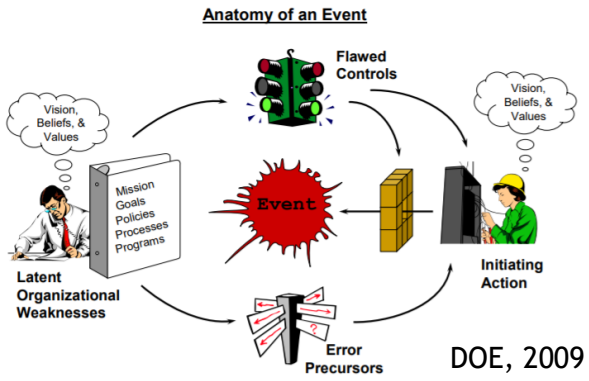
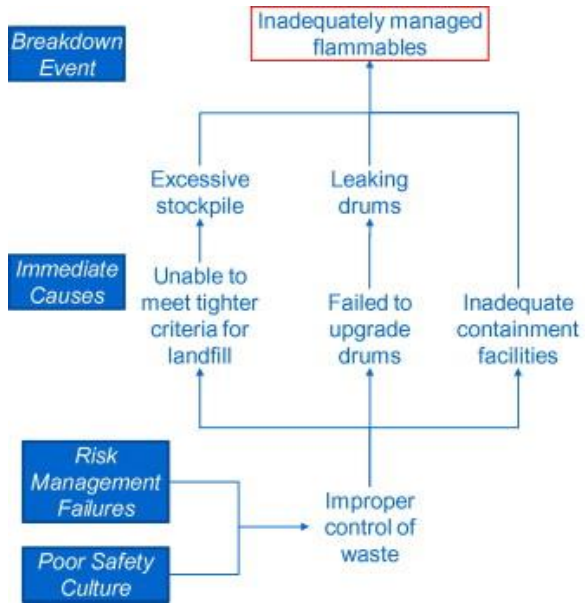


Figure 3.1: Typical Security Incident Response Cycle

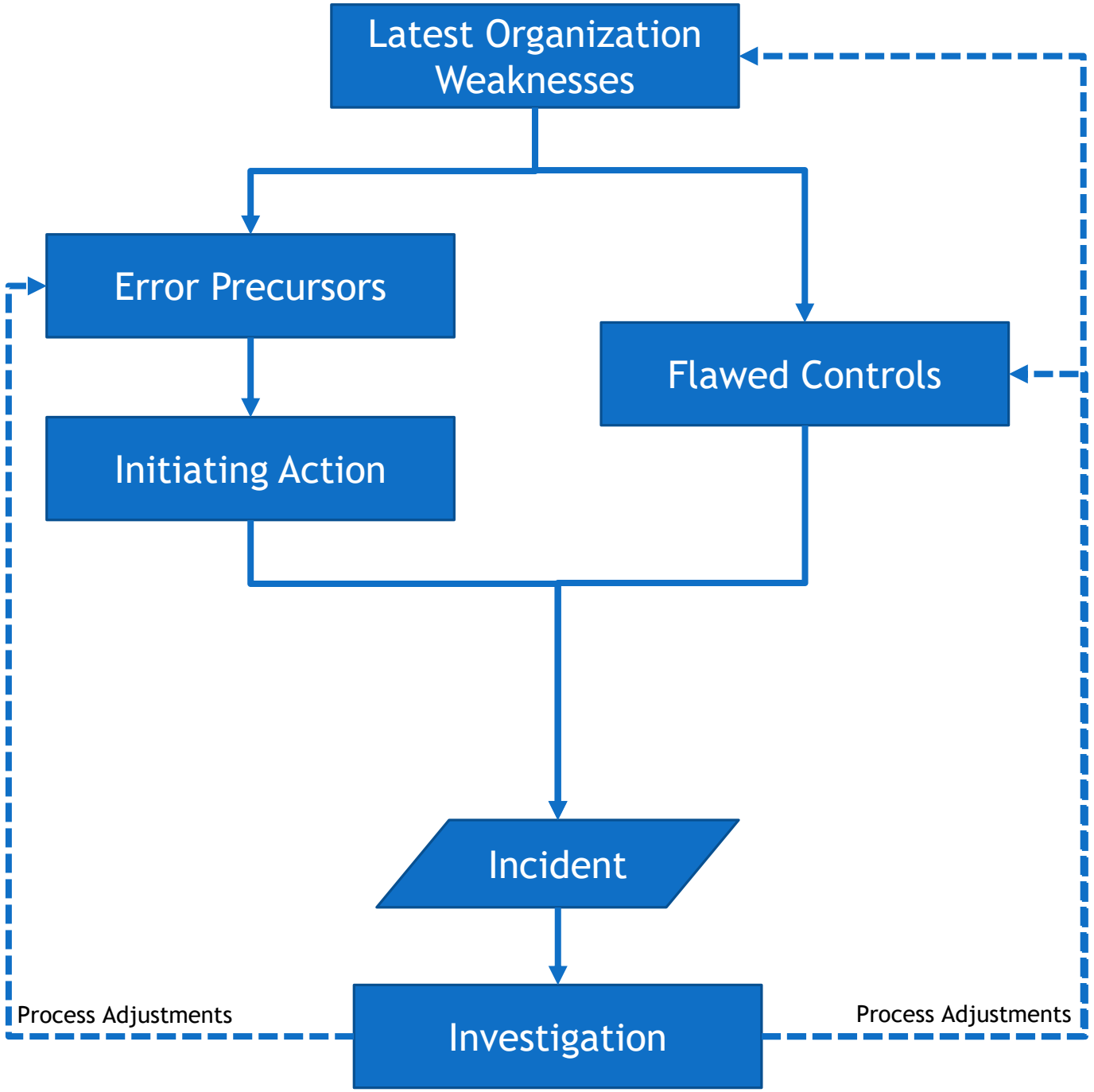
Grispos, 2016



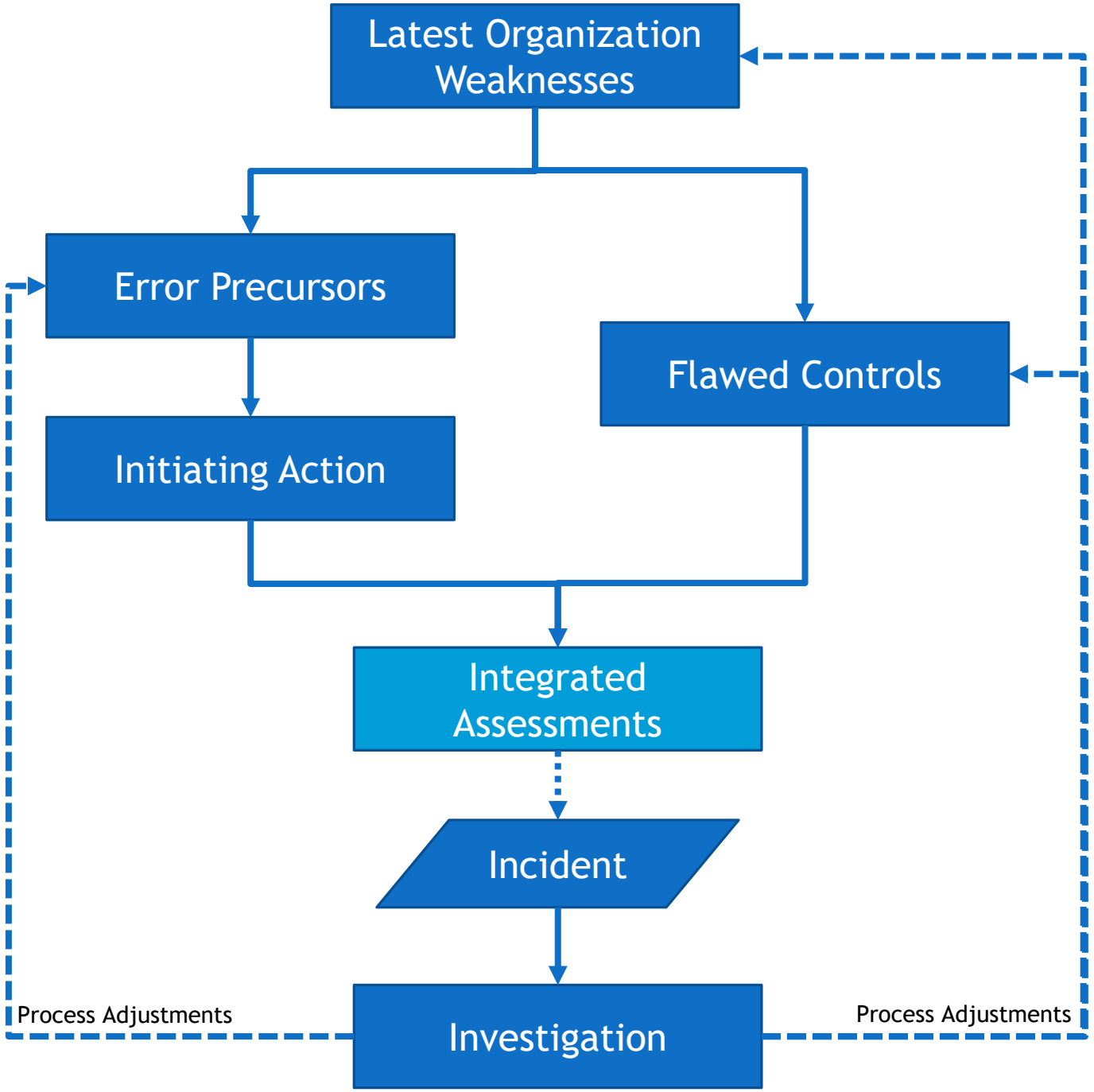
DOE, 2009



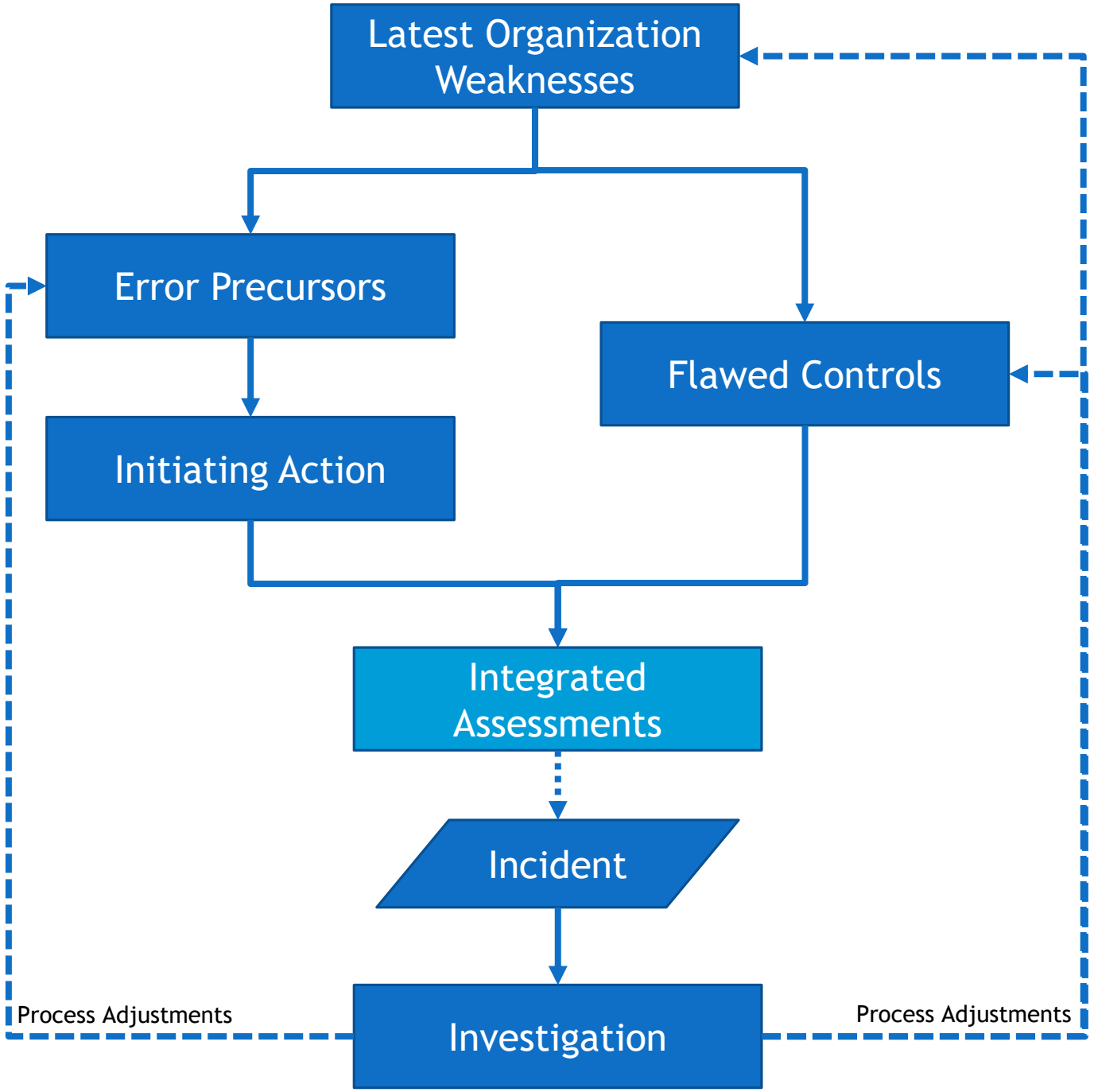
Goh et al., 2010



Typical Incident Cycle



Integrated Assessments aim to reduce incident occurrence by identifying and assessing incident precursors for specific organizations and provide training on any issues that have been identified.



Integrated Assessments aim to reduce incident occurrence by identifying and assessing incident precursors for specific organizations and provide training on any issues that have been identified.

Challenge: How can data science help us better choose organizations which might be at higher risk of having an incident?

Methodology: Knowledge Management Framework

PEOPLE

TECHNICAL TOOLS

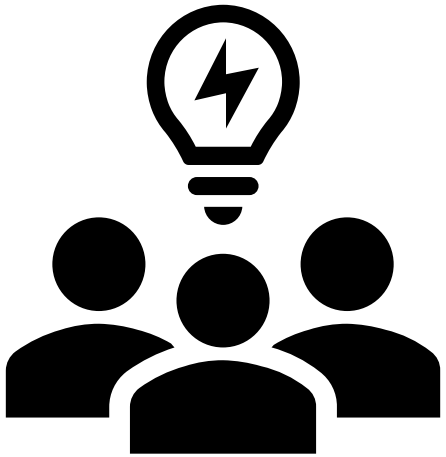
PROCESSES

(Rubenstein-Montano et al., 2001)

Data Collection

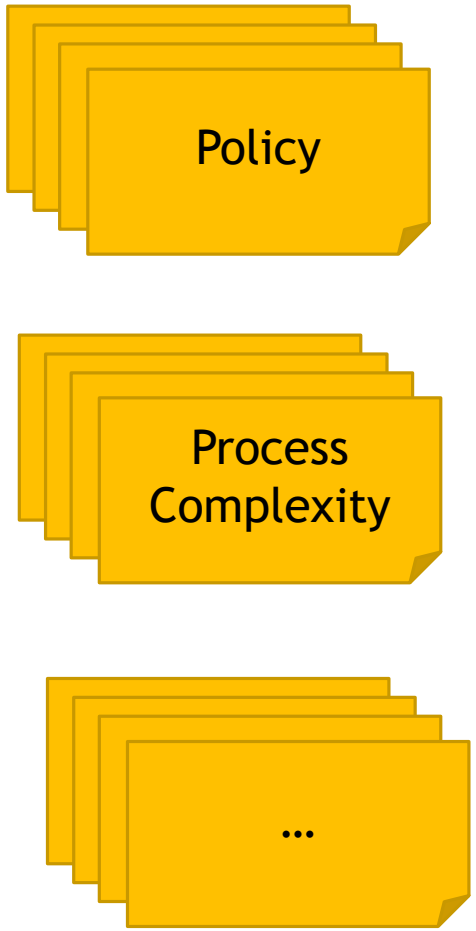
- Staff Composition
- Management Characteristics
- Training
- Process Complexity
- Visitors
- Travel
- Incident History
- Collaborations
- Deadlines
- Policy
- Environment
- …

Data Categories Identified by SMEs



Structured Improvement Activity

Data Processing



Identify Datasets

Generate Features

$$\sum p$$

Sums

$$\sum p_k \ln p_k$$

Diversity scores

(Keylock, 2005)

$$\sum \frac{p}{n}$$

Normalized values

Machine Learning

70-30 train-test

10-fold cross-validation

Model evaluations prioritized accurate true positive (sensitivity), true negative (specificity), and distinction between a true positive and false positive (using receiver operating characteristic; ROC)

Implemented in R

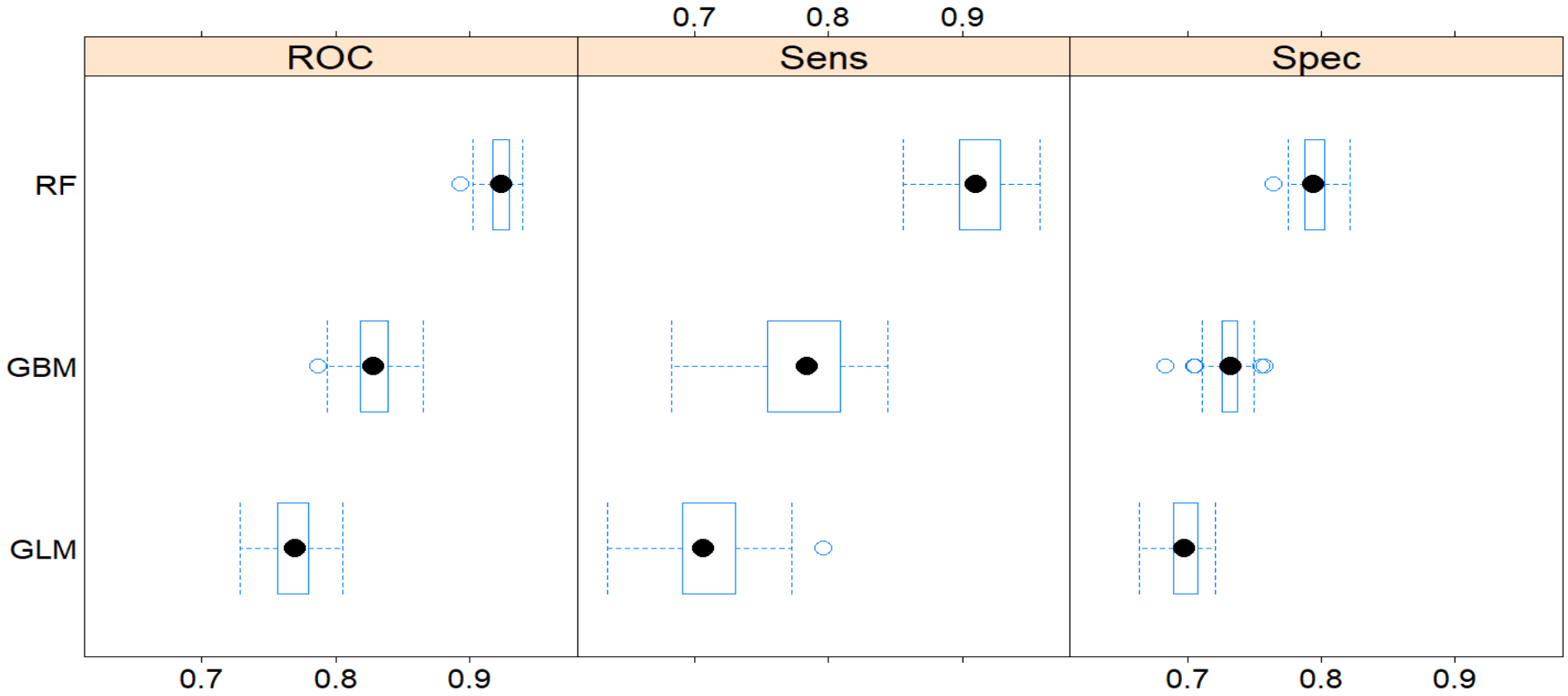
Algorithms

- Random Forest
- Generalized Linear Model
- Gradient Boosted Trees

Sampling Protocols

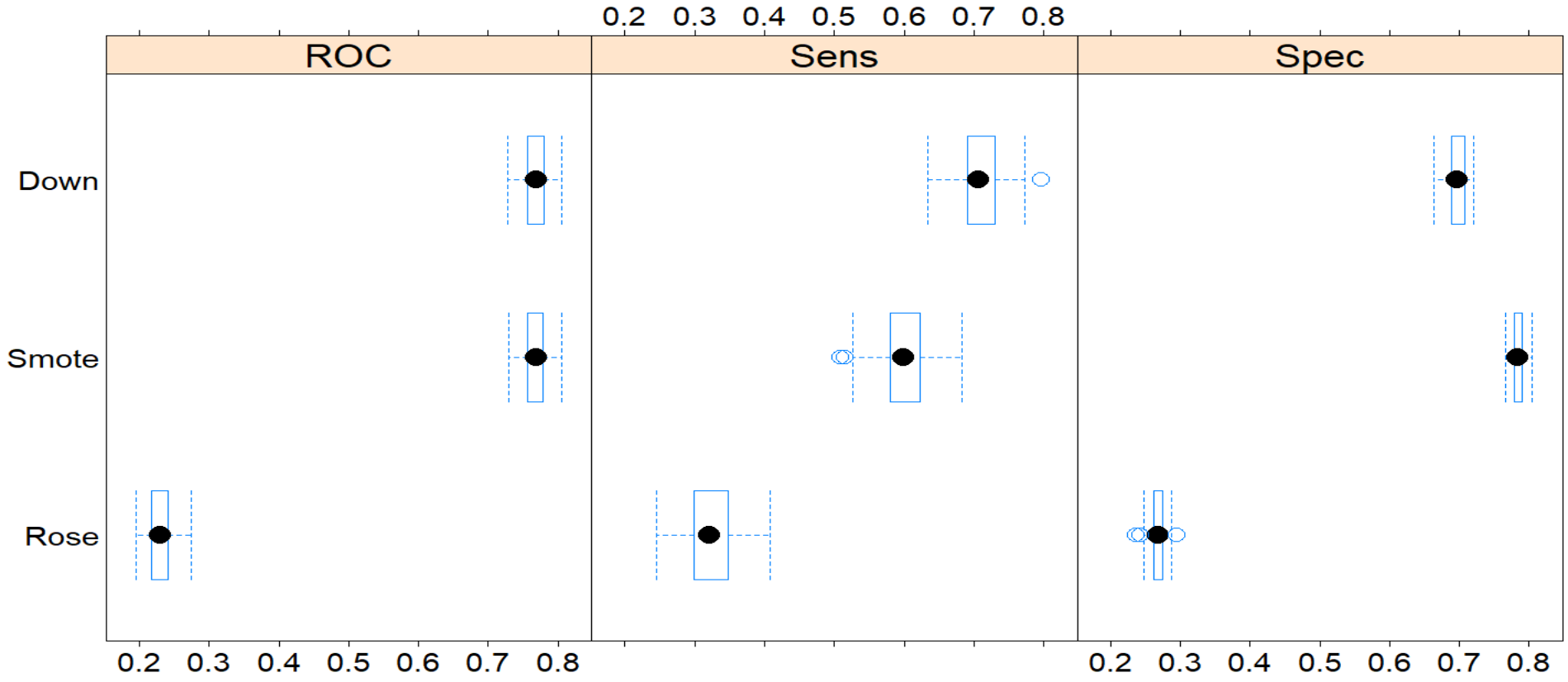
- Down
- Smote
- Rose

Machine Learning Results



- The random forest model consistently outperformed the gradient-boosted model and generalized linear model across all three model evaluation metrics

Machine Learning Results



- Down sampling had higher sensitivity rates while smote sampling had higher specificity rates.
- Down and Smote sampling had comparable ROC values

Machine Learning Selection

Based on performance accuracy and prioritization for true positives…

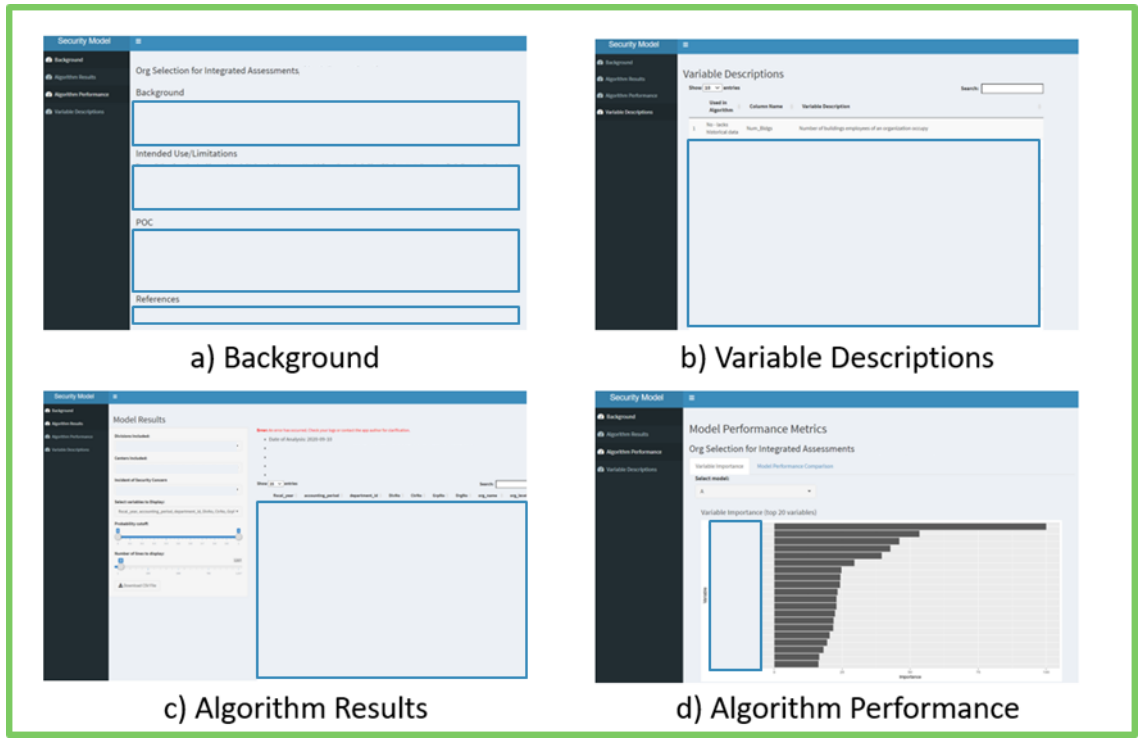
Algorithms

- Random Forest
- Generalized Linear Model
- Gradient Boosted Trees

Sampling Protocols

- Down
- Smote
- Rose

Data Visualizations

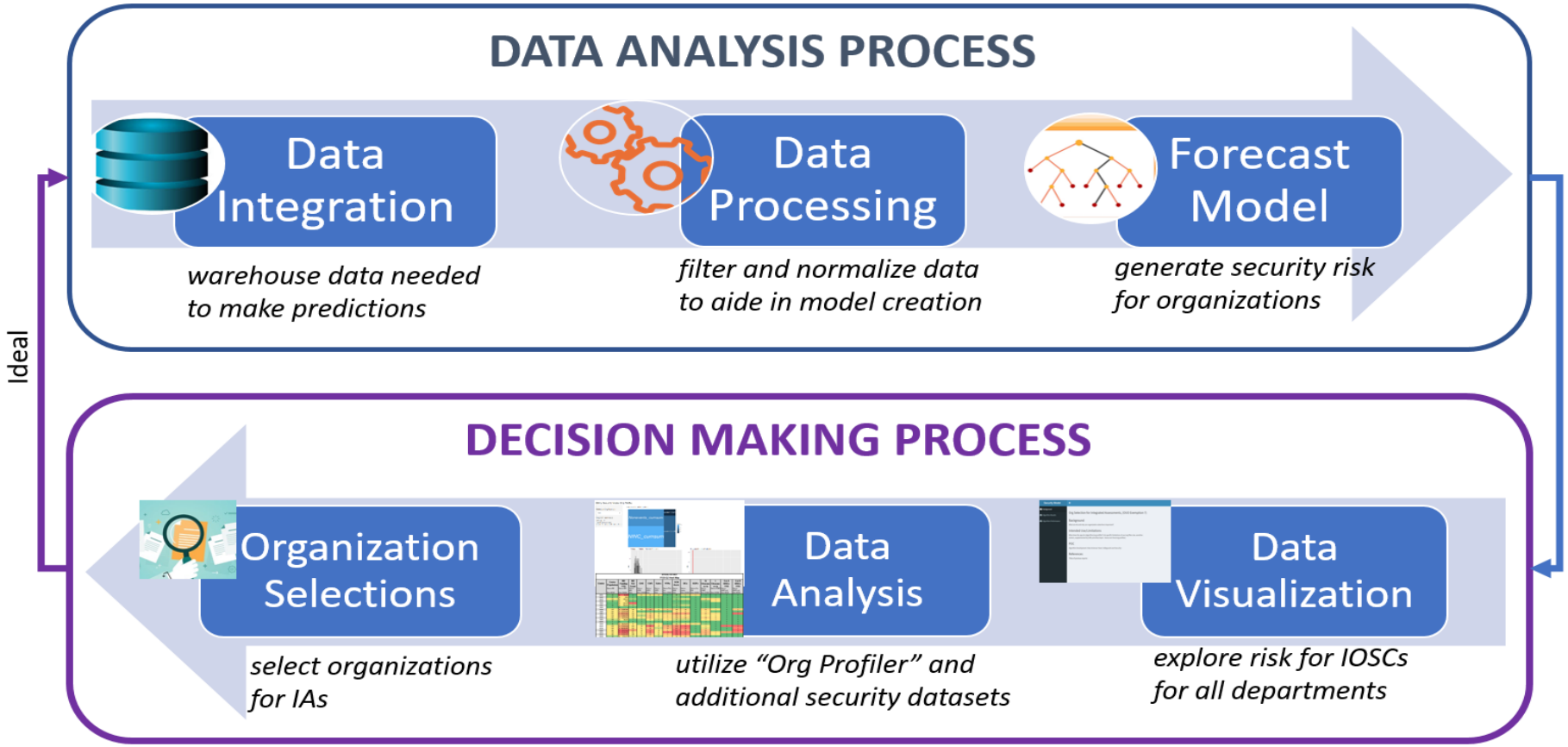


“Algorithm Results”

“Org Profiler”

(Illustrative Results with Simulated Data)

Summary of Process



Automated pipeline that automatically ingests/processes data and executes algorithm

Have two visualizations to support model development and organization selection activities

Knowledge Management: Continuous Improvements

DATA

- Individual Databases
- Central Repository
- Additional Datasets

TOOLS

- Shiny Visualizations
- Server Implementation
- Package Development

ANALYSES

- Forecast IOSC Risk
- Pattern Analysis
- Causal Evaluations

Concluding Remarks

- Knowledge management is key to successful tailoring technical tools to match operational needs
  - Leverage existing knowledge (theory, SMEs)
  - Consider people and processes (environments, presentation styles)
- Follow best practices where possible (both data and software!)
  - Findable
  - Accessible
  - Interoperable
  - Reproducible
- Continuous improvements are part of the process
  - Extending visualizations
  - Updating analyses
  - Revising metrics used for assessment
  - …

…of the people, by the people, for the people

Data, statistics, and machine learning are all pieces of the puzzle

Integration activities and process updates improve knowledge transfer

References

Department of Energy. 2009. Human performance improvement handbook, volume 1: Concepts and principles, DOE-HDBK-1028-2009.

Yang Miang Goh, Helen Brown, and Jeffery Spickett. "Applying systems thinking concepts in the analysis of major incidents and safety culture." Safety Science 48, no. 3 (2010): 302-309.

George Grispos. 2016. On the enhancement of data quality in security incident response investigations. Ph.D. thesis, University of Glasgow.

C. J. Keylock. 2005. Simpson diversity and the Shannon–Wiener Index as special cases of a generalized entropy. Oikos, 109(1):203–207.

R. Lawton, R. R. McEachan, S. J. Giles, R. Sirriyeh, I.S. Watt, & J. Wright (2012). Development of an evidence-based framework of factors contributing to patient safety incidents in hospital settings: a systematic review. BMJ quality & safety, 21(5), 369-380.

Bonnie Rubenstein-Montano, Jay Liebowitz, Judah Buchwalter, Doug McCaw, Butler Newman, Ken Rebeck, and The Knowledge Management Methodology Team. 2001. A systems thinking framework for knowledge management. Decision support systems, 31(1):5–16.

Increasing Integration of Data-driven Analyses in Operational Activities through Knowledge Management

Thank you for your time!

tgunda@sandia.gov