PNNL-SA-219365

Pacific Northwest NATIONAL LABORATORY

U.S.NRC United States Nuclear Regulatory Commission *Protecting People and the Environment*

Evaluation of Commercially Available Assisted Data Analysis for Inservice Inspections

Muthu Elen, Matt Prowant, Nick Conway, Joel Harrison, Aimee Holmes, Richard Jacob

Pacific Northwest National Laboratory (PNNL)

This work was sponsored by the U.S. NRC
NRC COR: Chinthaka Silva

U.S. DEPARTMENT of ENERGY BATTELLE
PNNL is operated by Battelle for the U.S. Department of Energy

Introduction

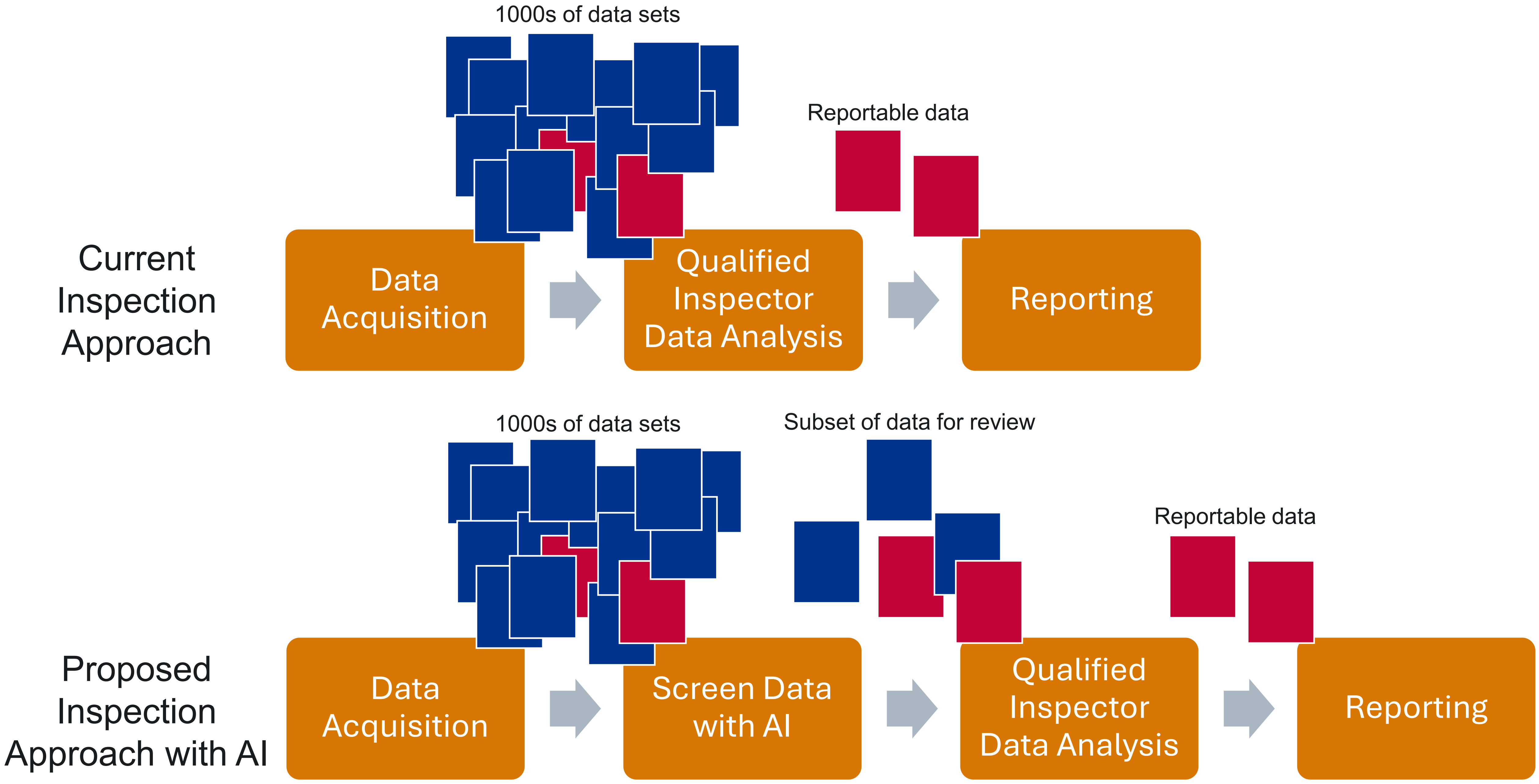
- Inservice inspection (ISI) of nuclear components relies heavily on ultrasonic testing (UT) and other nondestructive evaluation (NDE) methods.
- Current practice involves a qualified human NDE inspector to review large, complex datasets—this approach is time-intensive and subject to human factors variability[1].
- Pressures are driving the commercial nuclear power industry to use assisted data analysis for NDE, including declining numbers of certified UT Level III inspectors[2] and a push for shorter outages.
- Artificial intelligence (AI) and machine learning (ML) for assisted data analysis (ADA) offer potential for faster and more consistent flaw detection, characterization, classification, and reporting.
- Commercial AI- and ML-enabled NDE data analysis systems are emerging (TrueFlaw, Waygate, Eddyfi, MISTRAS, etc.). TrueFlaw is already conducting field trials in U.S. plants and abroad[3].
- There is a need to independently evaluate readiness, regulatory acceptability, performance, and integration of commercial AI systems in nuclear ISI workflows.

1- Sanquist, T. et al., "Human Factors in Nondestructive Examination", NUREG/CR-7295 (2022).
2- Atkinson, R. "NDE Workforce Study", ML24026A087 (2024).
3- "Nuclear Industry first AI powered ultrasonic inspection in Ringhals, Sweden", NDT.net Issue: 2025-06, https://www.ndt.net/search/docs.php3?id=31354

Objective and Scope

- Pacific Northwest National Laboratory (PNNL) and Oak Ridge National Laboratory (ORNL) are leading NRC confirmatory research in the use of AI for assisted analysis of UT ISI data in the U.S. commercial nuclear fleet.
- PNNL is performing confirmatory research to evaluate pre-trained commercial AI algorithms using UT data collected at PNNL with a performance demonstration initiative (PDI)-qualified procedure.
- The initial focus of work is on dissimilar metal weld (DMW) inspections.
- A statistical analysis of overall AI performance, including false calls and missed detections, is being conducted.
- The performance of a commercial AI system is being evaluated by an experienced and certified ASNT UT Level III inspector.
- Work will focus on performance of AI model variations (e.g., different software versions), variances in data quality, methods of dataset augmentation for model training, model retraining with site-specific data, verification and validation methods to quantify confidence in AI results, and model performance using data collected with procedural variances.
- A draft Technical Letter Report summarizing PNNL's and ORNL's work on AI is scheduled to be delivered to NRC in 2027.

Overview of Assisted Analysis with AI



Current ASME Code Activity of AI for ISI

- ASME Section XI, Appendix VIII-4200.
  - Implementing automated procedures is allowed according to VIII-4200.
    - ✓ "When the performance demonstration uses prerecorded data, algorithms for automated decisions may be altered when the altered algorithms are demonstrated to be equivalent to those qualified. When the performance demonstration results meet the acceptance requirements of Article VIII-3000, the algorithm shall be considered qualified."
  - ASME Record Number 24-2301 proposes to change VIII-4200 to address machine learning specifically. It also proposes to add relevant definitions to Section XI, IWA-9000 (Glossary).
    - ✓ Approved in Standards Committee in September 2025.
  - AI algorithms must meet the performance requirements of VIII-3000.
    - ✓ Detection, false calls, sizing.
    - ✓ Only valid for encoded data.
- ASME BPV Section V – Working Group on Assisted Analysis.
  - Charter: "Working Group Assisted Analysis (WG-AA) will develop the requirements and definitions for assisted analysis to be published in the ASME Section V Boiler and Pressure Vessel Code."
- There are no active or draft Code Cases specific to the use of AI/ML in ISI.
- ASNT Standard BSR/ASNT AI/ML-202x - Use of AI/ML for NDT/E Applications (*under final review*).
- ASTM - ASTM E3327/E3327M-21: Standard Guide for the Qualification and Control of the Assisted Defect Recognition of Digital Radiographic Test Data.

Commercial AI Systems for UT Inspections in NPPs

- Several commercial AI systems are being developed for assisted evaluation of NDE data in nuclear power plants.
  - Waygate/Baker Hughes, EddyFi (Magnifi), CIVA, and TrueFlaw have been developing AI/ML methods to incorporate into their systems.
- TrueFlaw is the only vendor who is performing field trials in U.S. nuclear plants.
  - Reactor pressure vessel inspections.
  - DMW inspections.
  - Visual inspections.
- TrueFlaw has been qualified in Sweden for pressure head vessel penetration inspections.
- TrueFlaw's system is referred to as "Box".
- Among the commercially available systems, the TrueFlaw Box is closest to deployment and is therefore the subject of PNNL's confirmatory research.

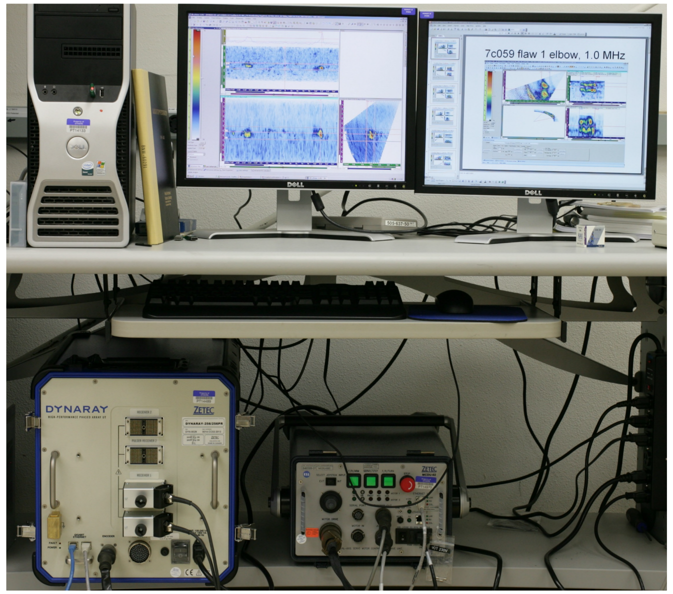
TrueFlaw's "Box" is a plug-and-play AI-driven data analysis system.

Box Testing at PNNL

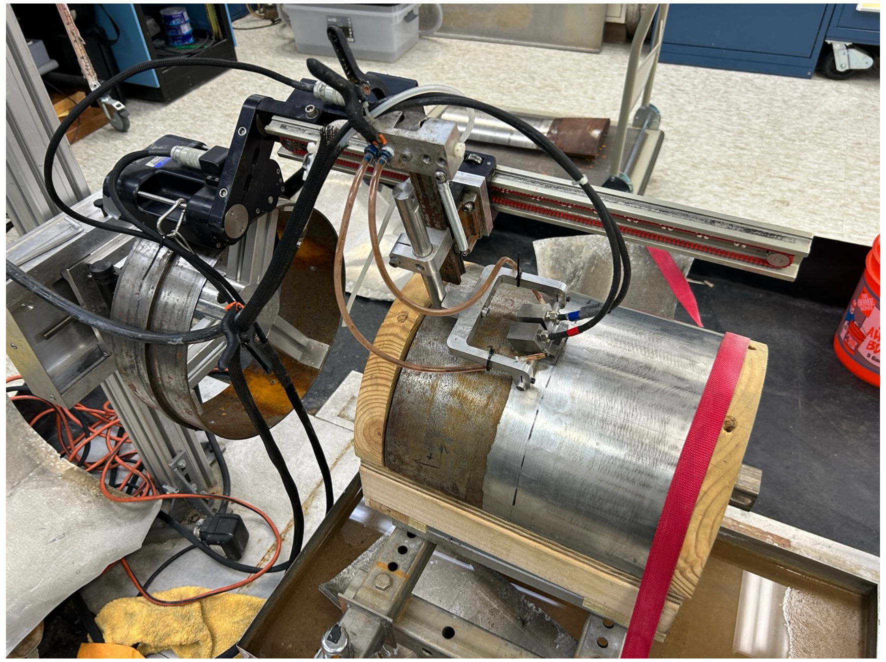
- As part of confirmatory research, PNNL is performing controlled laboratory testing of Box's performance on DMW data.
- PNNL is leasing Box from TrueFlaw to assure that the research will be independent. This is not a collaborative effort between PNNL and TrueFlaw.
- PNNL receives periodic AI model updates from TrueFlaw and discusses model successes and failures. No ultrasonic data are shared with TrueFlaw.
- PNNL received two DMW Models, DMW_AOCF and DMW_COAF, which refers to Axially Oriented Circumferential Flaws (AOCF) and Circumferentially Oriented Axial Flaws (COAF). Due to a sparsity of axial flaw data, PNNL evaluated only the AOCF model.
- A qualified inspector analyzed PNNL's Box results to classify detections as true calls or false calls per specified grading criteria.

PNNL Mockups and NDE Data Acquisition

- NDE inspections were performed on DMW (carbon steel to stainless steel) mockups at PNNL.
- Mockups contained circumferentially oriented thermal fatigue cracks (TFCs).
- Experimental data were collected according to the same EPRI procedure used to collect training data for Box.
  - EPRI procedure EPRI-ENC-DMW-PA-1, technique sheets RS-DYN-STD-02 and RS-DYN-STD-03.
  - 1.5 MHz phased-array.
  - 8×4 element matrix array probe.



PA-UT scanning hardware at PNNL



An encoded scan being performed on a DMW mockup

Pacific Northwest NATIONAL LABORATORY

PNNL Mockups and NDE Data Acquisition

Primary Dataset

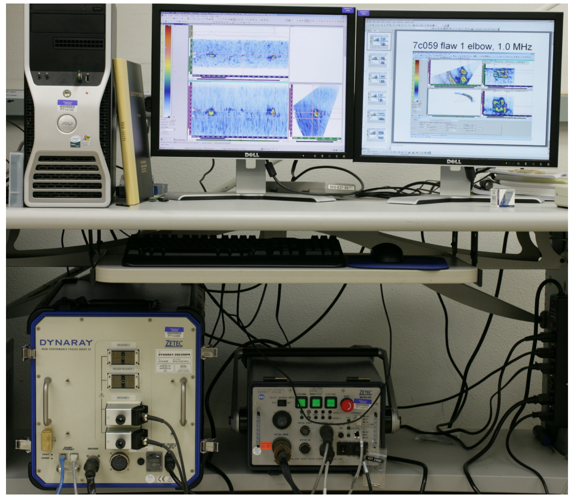
Specimen	Flaw	Flaw Type	Orientation	Flaw Length (in.)	Flaw Depth (in.)	%TW	Tilt	Thickness (in.)	Degree Location
8C-032	1	TFC	Circ	0.90	0.30	20%	0	1.51	35
	2	TFC	Circ	1.14	0.57	40%	19	1.42	150
	3	TFC	Circ	1.81	0.90	60%	0	1.51	220
	4	TFC	Circ	0.85	0.43	30%	13	1.42	325
8C-036	1	TFC	Circ	2.48	0.83	58%	30	1.42	47
	2	TFC	Circ	2.85	1.43	95%	0	1.50	135
	3	TFC	Circ	1.58	0.53	35%	0	1.50	195
	5	TFC	Circ	2.26	1.13	75%	0	1.51	285
9C-023	1	TFC	Circ	2.76	0.49	34%	2	1.45	0
	2	TFC	Circ	2.01	0.26	19%	8	1.41	90
	3	TFC	Circ	2.76	0.33	24%	4	1.39	180
	4	TFC	Circ	2.26	0.16	11%	12	1.43	270
14C-146	4	TFC	Circ	2.965	0.522	15.7%	0	3.323	82
706-P1	1	TFC	Circ	3.92	0.33	28%	0	1.17	15
	2	TFC	Circ	3.68	0.78	67%	10	1.17	70
706-P2	1	TFC	Circ	3.43	0.90	77%	10	1.17	17

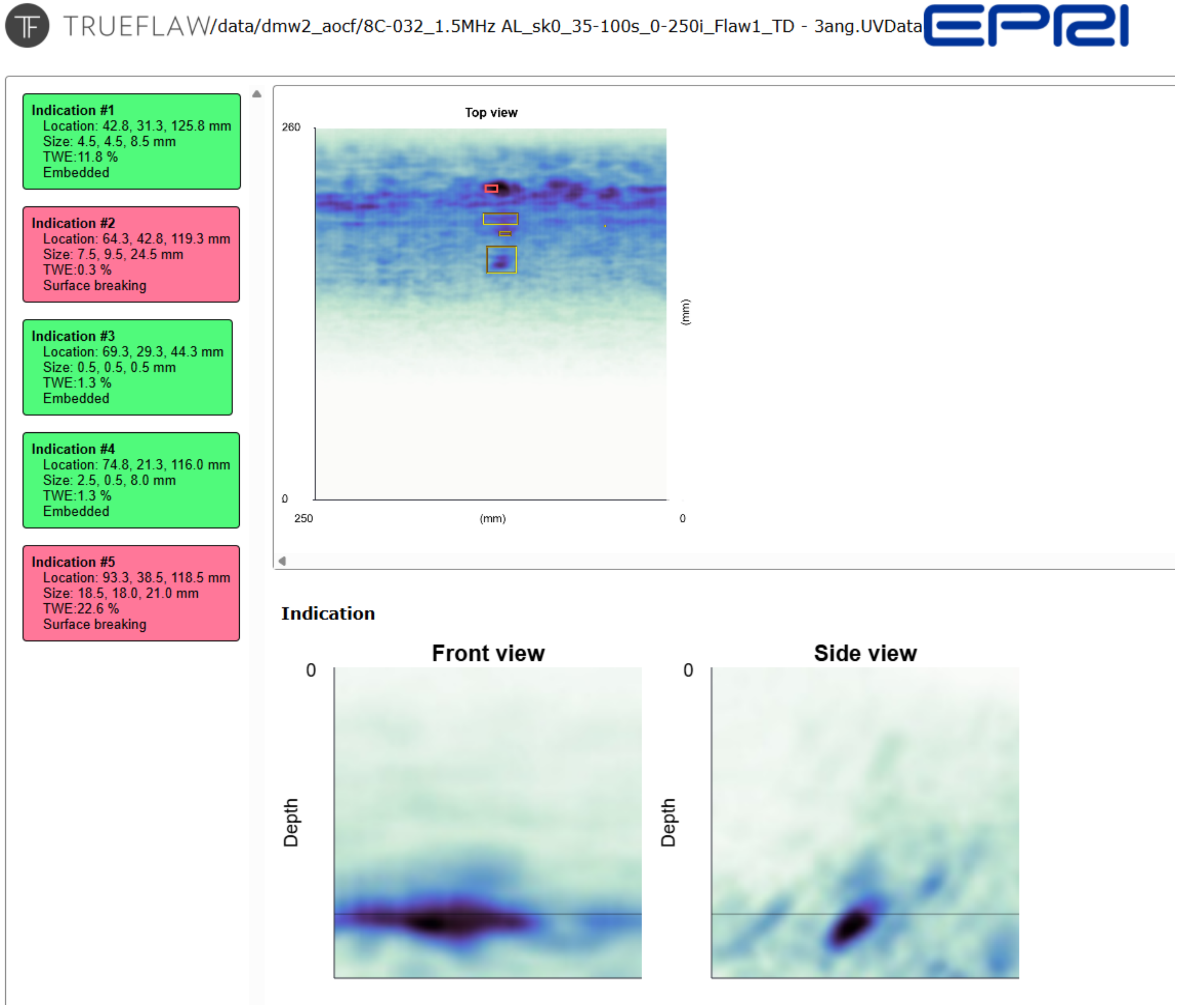
Special Cases (to be considered separately from the primary dataset)

- Hot isostatic pressed electrical discharge machined notch (HIP'd EDM) flaws were in specimen 8C-036
- Specimens 10C-011 and 9C-034 were full structural weld overlays (FSWOL)
- A few selected data files from the dataset were selected to create virtually modified flaws

- DMW specimens selected were 12-14 in. OD pipe-to-nozzle specimens, with the addition of one 36 in. OD specimen (14C-146).
- Data were collected from both sides of the welds.
  - Some mockups had cast austenitic stainless steel (CASS).
  - Scans from the CASS side were excluded in the analysis results due to no applicable inspection procedure.

AI Model Evaluation





- Box's workflow is simple:
  - Plug Box into a USB port.
  - Drag the files to be analyzed to the Box folder.
  - Box creates output files that can be reviewed in a web browser.
- Output files show each indication that Box flagged, with size and location information.
- Each indication was reviewed by the Qualified inspector with access to true-state information, as per the grading criteria (next slide).

Pacific Northwest NATIONAL LABORATORY

Example of Essential Settings Report

- Box evaluates the data file for essential variables and data quality (e.g., missing data, noise levels).
- Essential parameters check
  - Wave speed and shear speed
  - Digitizing frequency
  - Compression
  - Configuration
  - Voltage
  - Pulse width
  - Scale type
  - Rectification
  - Wave type
  - Scan axis step
  - Proper-a-scan length
  - Sufficient digitized resolution,
  - Sufficient stored resolution
- Some parameters are not considered essential to model performance
  - Averaging type
  - Sync mode
  - Input filter
  - Analog smoothing

Example of Box's essential variable check output

beam name	Azim. R: 45.00 S: 0.00	
specimen longitudinal sound speed	5770.0	OK
specimen transversal sound speed	3150.0	OK
digitizing frequency	100000000.0	OK
averaging type	None	OK
recurrence		Fail
sync mode	Pulse	OK
compression		OK
inspection mode	PitchCatch	OK
voltage	40.0	Fail
pulse width	0.0	OK
scale	Linear	OK
rectification	Bipolar	OK
input filter	filter_None	Fail
analog smoothing		Fail
current wave type	Longitudinal	OK
scan sampling resolution	0.0	OK
Proper a-scan length	1.9	OK
Sufficient digitized resolution	inf	OK
Sufficient stored resolution	inf	OK

Grading Criteria

- Grading criteria were developed for evaluating the data. Indications were labelled as:
  - **Detection** (true positive) – if the indication correctly identifies the flaw.
  - **False call** (false positive) – if the indication identified a non-flaw as a flaw.
  - **No call** – the indication is ignored if:
    - ✓ The same flaw is identified by multiple indications, as only one indication is considered a true call. The largest indication is considered to be a detection, and others are defined as no call.
    - ✓ Axial grading units were used, so false calls that axially overlap were combined as a single false call, and others are defined as no call.
- Additional metrics were recorded.
  - Missed detection (false negatives) – a flaw was not identified as an indication.
  - True negatives – an unflawed region identified as non-flaw is a commonly used metric in AI; however, it does not apply to detection-based algorithms, as the model does not classify the entire image as “flawed” or “not flawed” and instead outputs bounding boxes where an indication is detected.
- Inconsistencies with Box’s image scaling have hindered automated grading criteria assessment. We enlisted the help of a qualified inspector to apply the grading criteria.

Example of a Box Output File

TRUEFLAW/data/dmw2_aocf/8C-036_1.5MHz AL_sk0_35-100s_550-820i_Flaw3-4_TD

Indication #1
Location: 39.8, 30.8, 42.3 mm
Size: 12.5, 22.5, 45.5 mm
TWE:48.8 %
Surface breaking

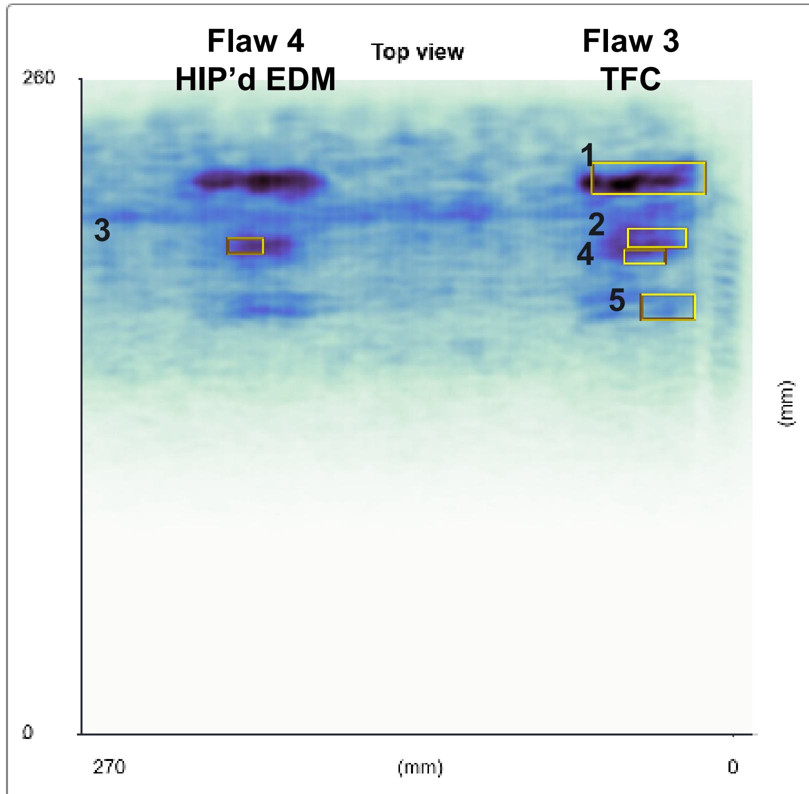
Indication #2
Location: 63.8, 39.5, 38.8 mm
Size: 7.5, 6.0, 23.5 mm
TWE:4.2 %
Surface breaking

Indication #3
Location: 67.0, 35.3, 205.3 mm
Size: 6.0, 4.5, 14.5 mm
TWE:13.4 %
Surface breaking

Indication #4
Location: 71.3, 25.5, 43.8 mm
Size: 5.5, 2.0, 16.5 mm
TWE:5.2 %
Embedded

Indication #5
Location: 91.5, 27.0, 34.3 mm
Size: 10.0, 4.0, 21.5 mm
TWE:10.5 %
Embedded

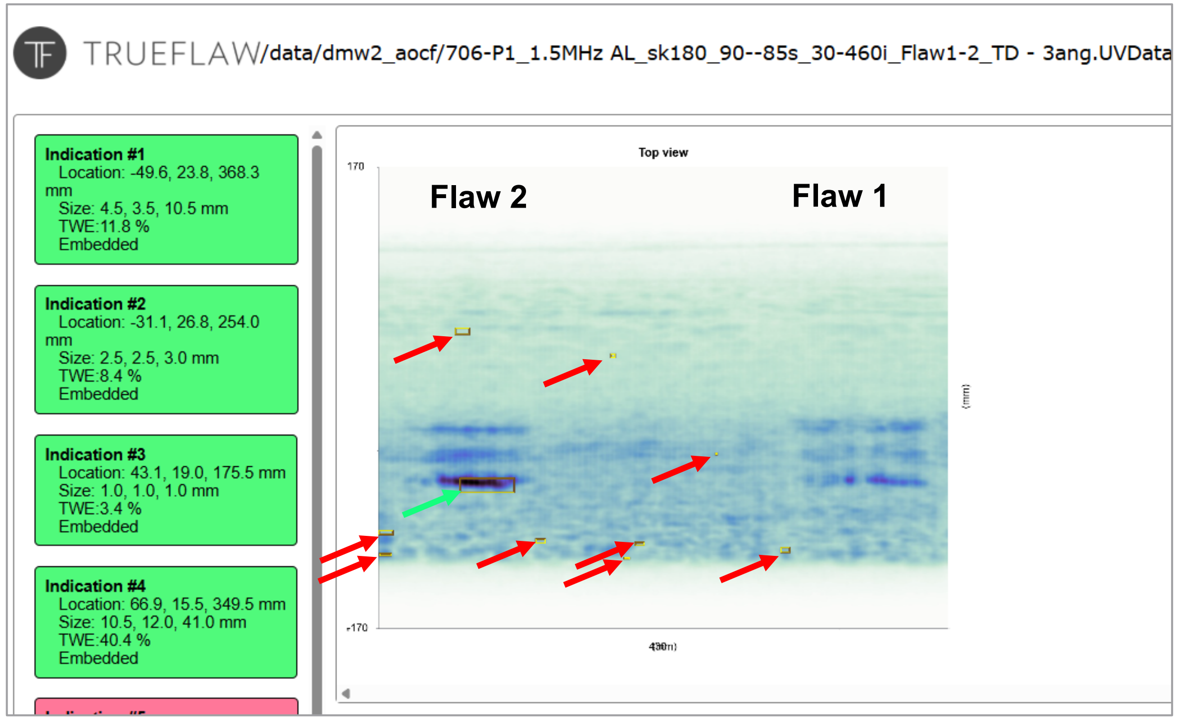
Indication #6
Location: 95.0, 26.5, 59.5 mm
Size: 0.0, 0.0, 0.0 mm
TWE:0.0 %
Embedded



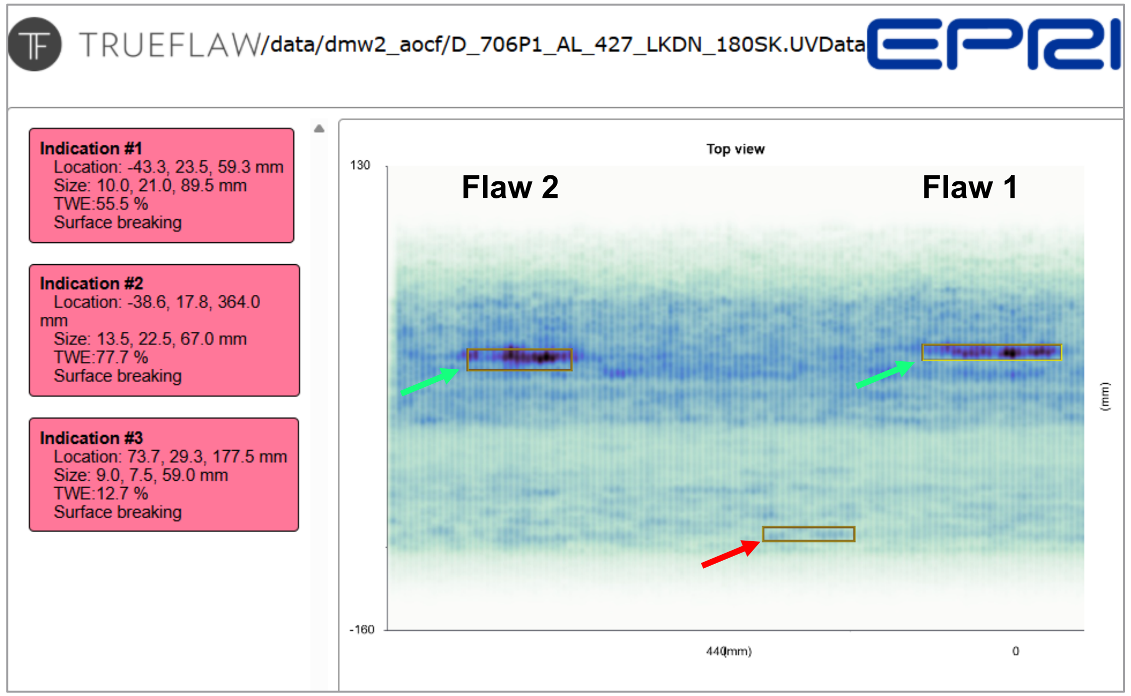
- Indication #1 – Detection of Flaw 3
- Indication #2 – No call (redundant detection of Flaw 3)
- Indication #3 – Detection of Flaw 4
- Indication #4 – No call (redundant detection of Flaw 3)
- Indication #5 – No call (redundant detection of Flaw 3)
- Indication #6 – False call (size of 0.0,0.0,0.0 mm, note this indication is not visible in the image)

Summary: for this output file, Box found six indications. The analyst discarded four indications according to the grading criteria, leaving two detections and no false calls or missed detections.

Example of Different Scans Acquired on the Same Mockup



PNNL data of 706-P1



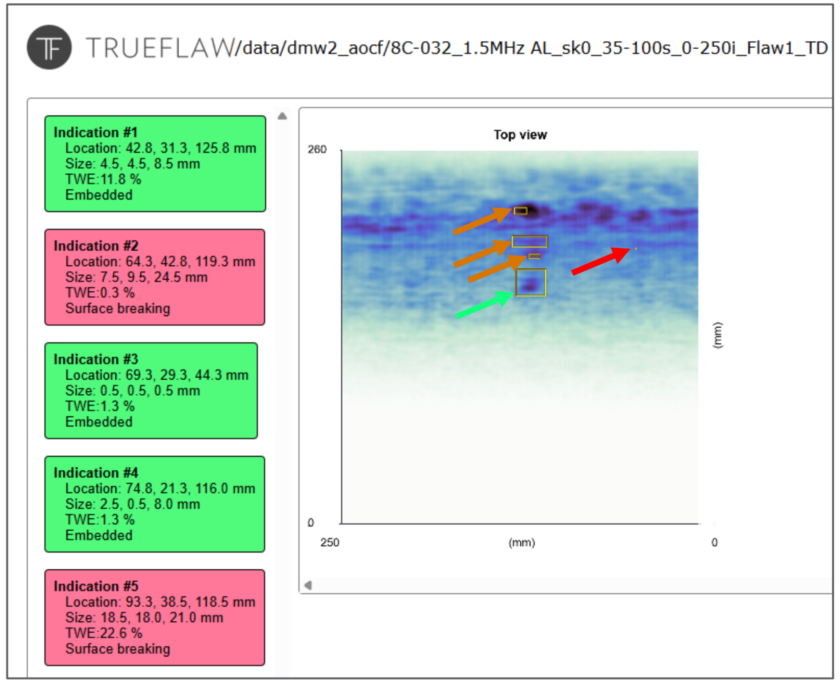
EPRI data of 706-P1

- A scan from PNNL was compared to a scan from EPRI acquired on the same mockup (706-P1).
- The PNNL and EPRI scans were acquired using the same probe and procedure.
- Box missed Flaw 1 in the PNNL scan and had many false calls (red arrows).
- Box detected both flaws in the EPRI scan with only one false call.
- The results shown were obtained from Box V.0425. Later versions identified both flaws

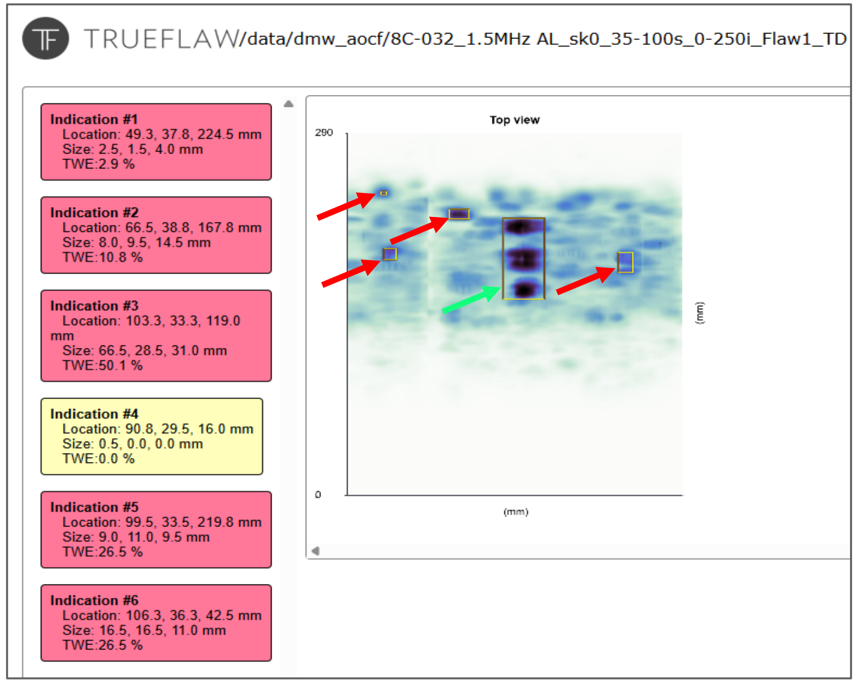
Box Version Comparison Example

- TrueFlaw provides periodic updates to Box software that affect performance.
- Some updates affect the AI model, but others affect non-model parameters such as sensitivity.

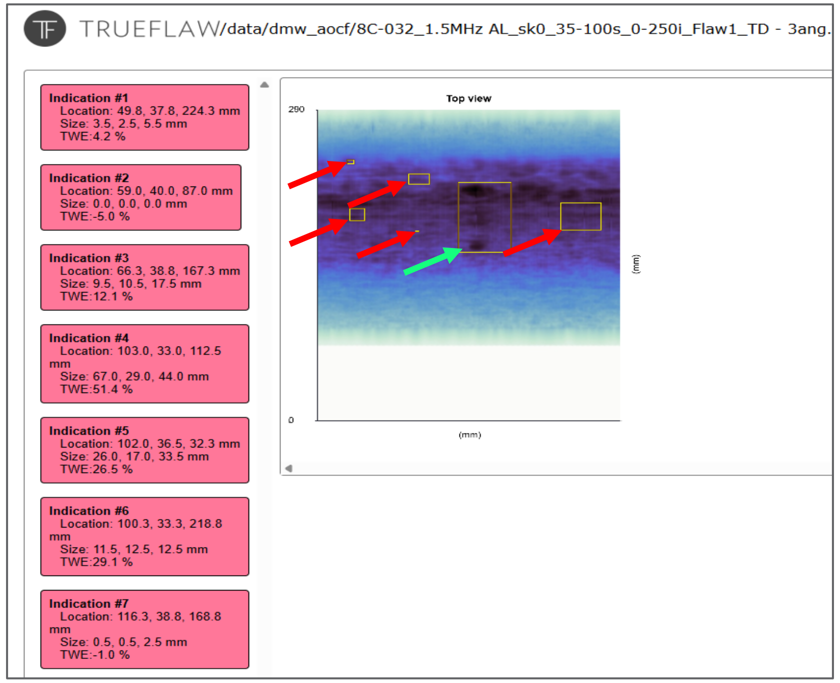
V.0425



V.0625



V.0825



*Two mockups were not included in the analysis of V.0625 due to a timing conflict with the update to V.0825.

Overall Performance Evaluation of Box

Versions currently at PNNL (V.0425, V.0825)

Metrics	V.0425	V.0625	V.0825
Detection rate	97%	100%	100%
True Flaws	31	25	31
Detections	30	25	31
False Calls	55	50	106
FCP	58%	67%	83%

- All metrics were calculated only for the primary dataset, which does not include the special case scenarios.
- V.0625 does not include data from 706-P1 and 706-P2, due to timing conflict with the update to V.0825.

Summary

- Commercially available assisted data analysis options for inservice inspection NDE are being field tested in U.S. nuclear power plants.
- PNNL is conducting confirmatory research for the NRC to evaluate the performance and effectiveness of such technology.
    - Rule-based options have been determined to be ineffective, with too many missed detections.
    - The TrueFlaw Box AI model demonstrates strong capability in detecting flaws with 100% or near 100% detection rates.
- Test data should be acquired with the same conditions, procedures, and limitations as those used for acquiring model training data.
    - Preliminary testing suggests that detection rates drop when there are procedural variations from the training data.
    - The effect of procedural variations on false call rates has not yet been evaluated.
    - Realistic variances, both within and outside procedural limits, are being tested.
- Model updates aimed at improving robustness to procedural differences have led to an increase in false call rates.
    - Small adjustments to model sensitivity have resulted in unexpected variations, especially in the numbers of false calls.
- Essential parameter checks are performed but not enforced; i.e., the system continues analysis even when key parameters fail validation.

Open Questions

- Will assisted data analysis (ADA) have to be trained/retrained with site-specific data prior to deployment?
- What PDI metrics or standards will ADA be held to?
- What expertise will be needed to train, deploy, and evaluate ADA for inservice inspections?
- Will additional rules or rulemaking be required for ADA deployment, or is the current ASME Code adequate?
- Which examination parameters should be considered as essential variables for ADA?
- In the case of missed flaws (false negatives), what is the risk impact to the plant/component, and how is this risk balanced against human-only inspection risk?
- Could the use of the ADA *change the culture* of inspection—for better or worse—by shifting trust from human expertise to “intelligent” algorithms?

Pacific Northwest
NATIONAL LABORATORY

Thank you

POC:
Richard Jacob (richard.jacob@pnnl.gov)