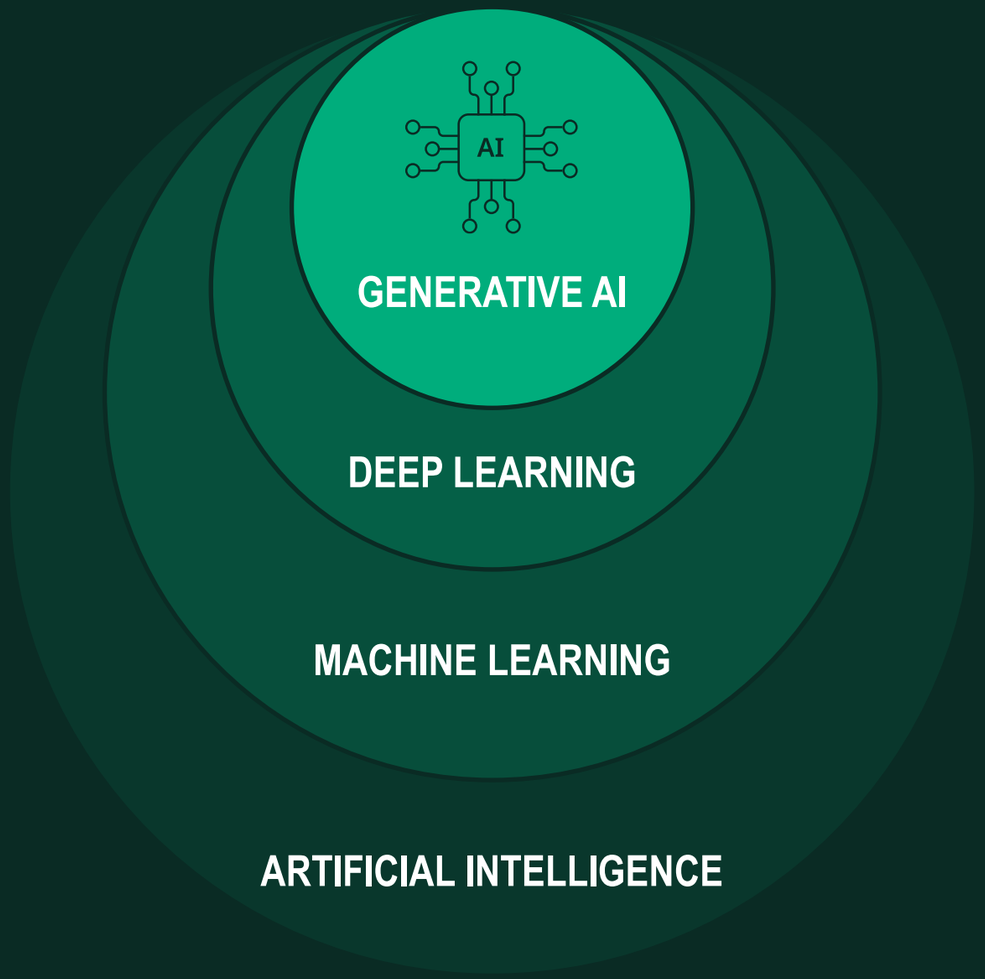
AI at the Edge (for Dummies… and Data Scientists)

Michael Schuldenfrei, NI Fellow

Introduction to AI

Evolution of AI



2021

GENERATIVE AI

Models which new written, visual, and auditory content given prompts or existing data.

2012

DEEP LEARNING

A machine learning technique in which layers of neural networks are used to process data and make decisions.

1997

MACHINE LEARNING

Subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions.

1956

ARTIFICIAL INTELLIGENCE

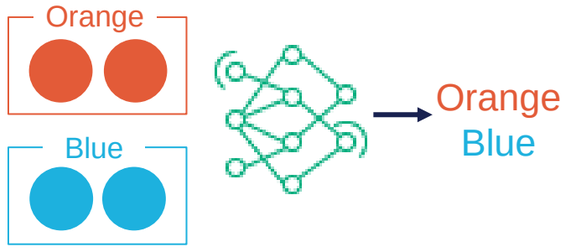
The field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence.

Machine Learning Key Topics

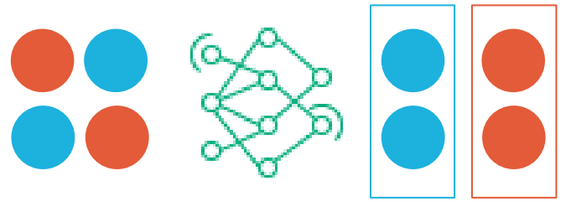
State-of-the-art models are increasingly hybrid across approaches.

Training

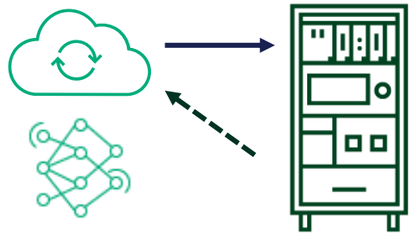
Supervised Learning



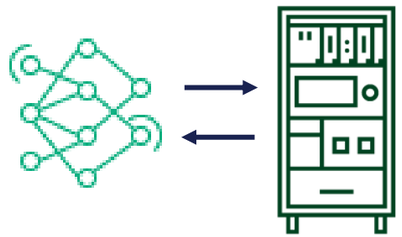
Unsupervised Learning



Offline Learning

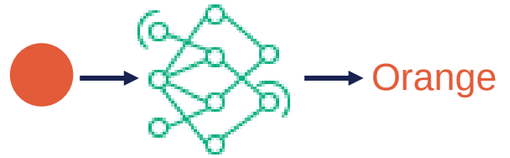


Online Learning

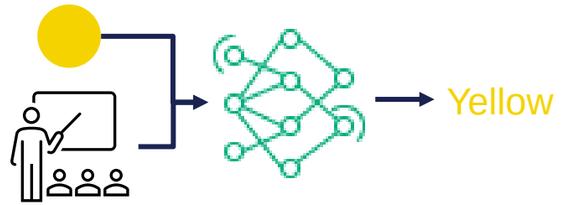


Inference

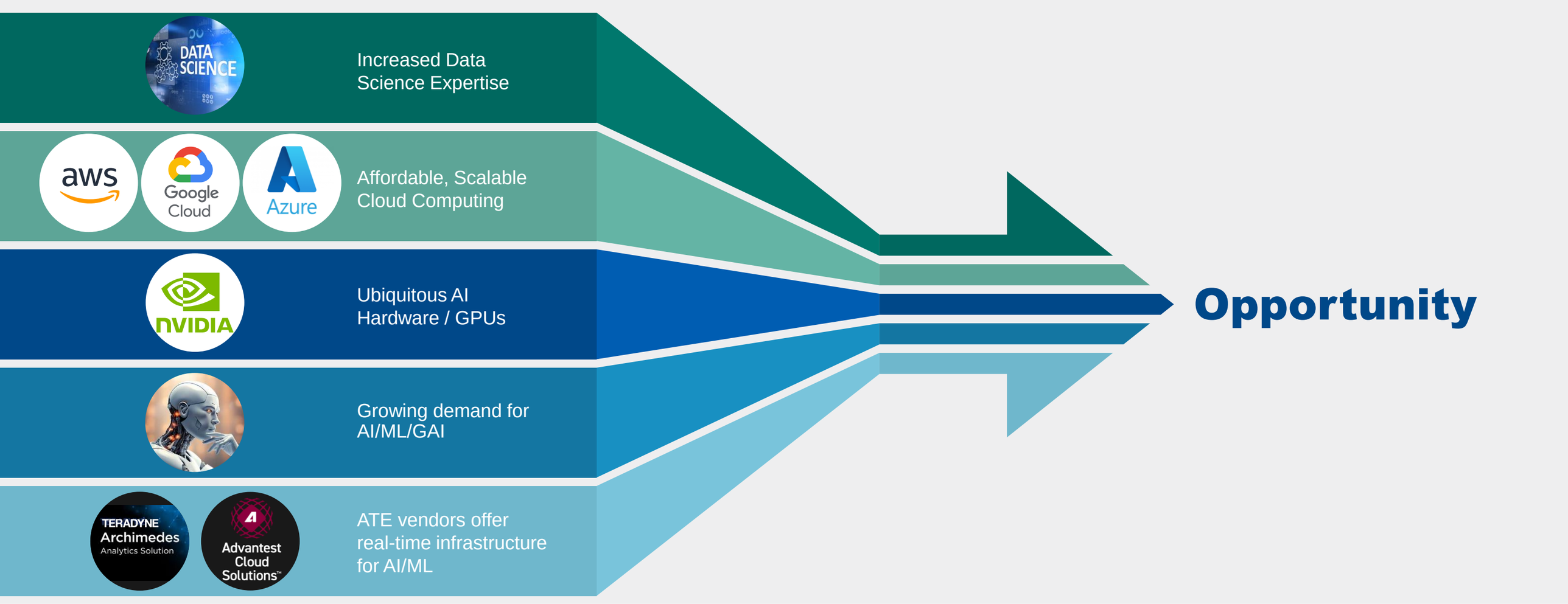
Model Execution



In-Context Learning

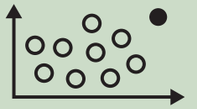


Converging Trends Creating an Opportunity



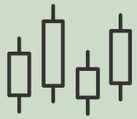
Examples of AI at the Edge

Edge (and Cloud) AI Examples


Outlier detection (Advanced)


Escape prevention (Equipment Health etc.)


Optical defect detection / ROI quality (e.g., welds, soldering etc.)

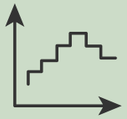

RMA reduction / prediction


Parametric trend detection


Parametric insights


Scratch detection (wafers)


AI based RCA


Waveform anomaly detection


Battery analytics (capacity, smart pairing, etc.)


Process optimization (e.g., Adaptive manufacturing)

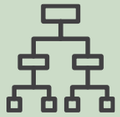

Early failure detection


Next operation reduction


Monitoring and auto RCA (UPH, cycle time, yield, error code distributions etc.)


Process variation detection


Yield trend detection

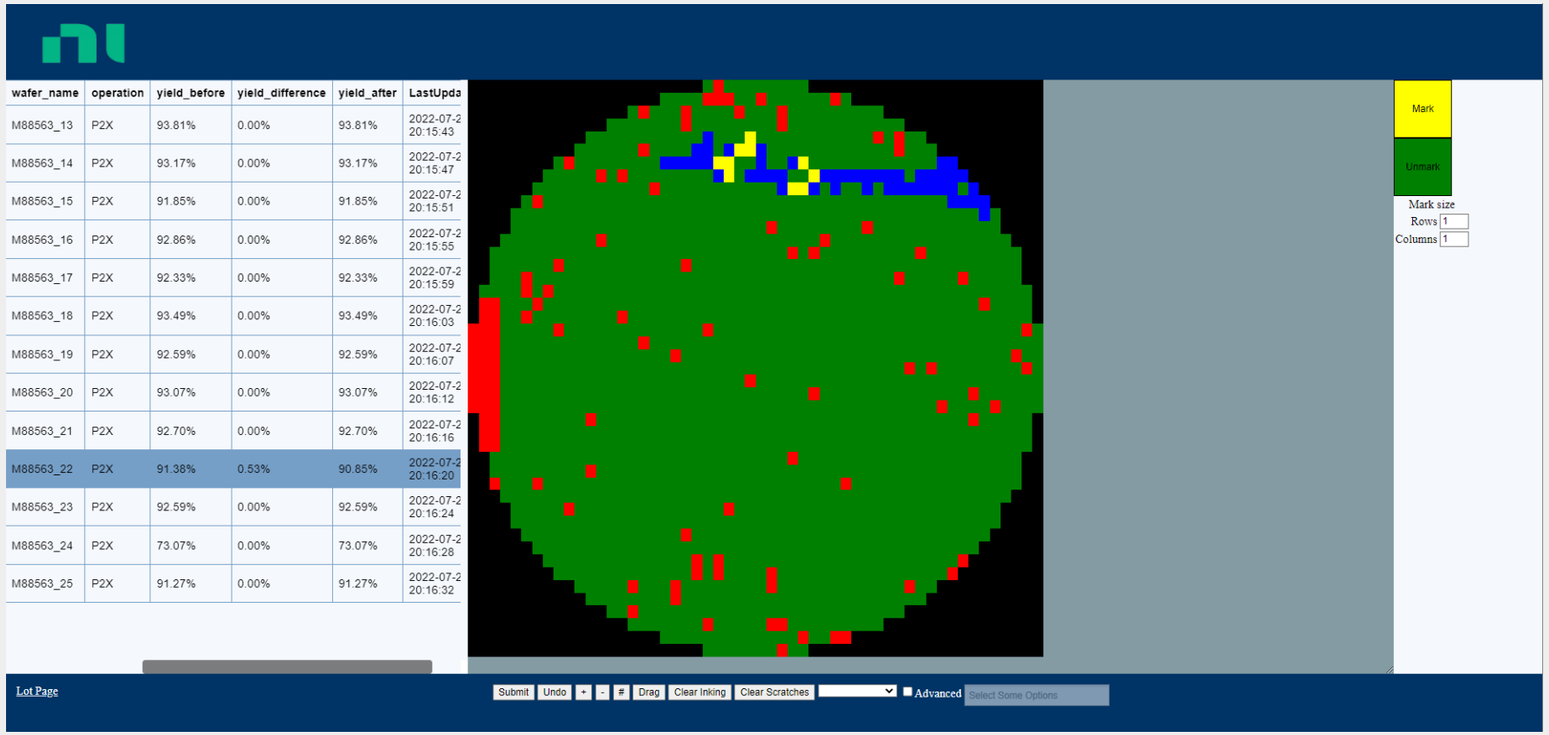

Wafer classification


Equipment utilization/variations


Predictive / JIT maintenance


Test (program) comparison

Scratch Detection

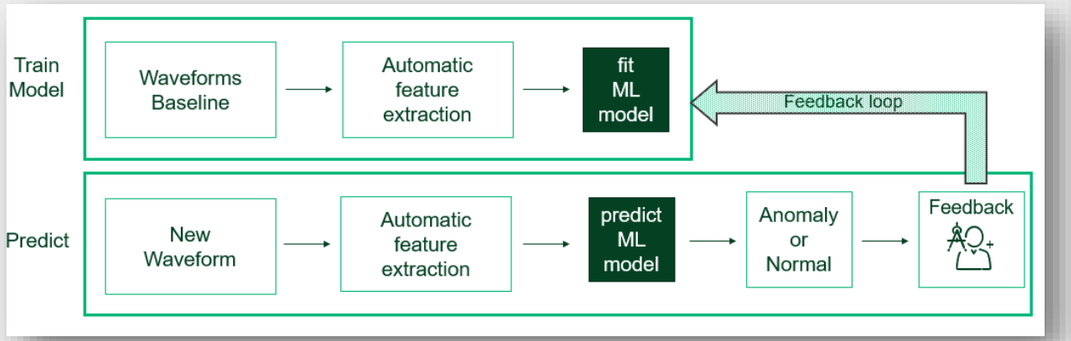


- Uses deep learning (DL) algorithms to detect scratches
- Automatically suggests inking – scraps suspect dies around the scratch
- Improves over time given user feedback

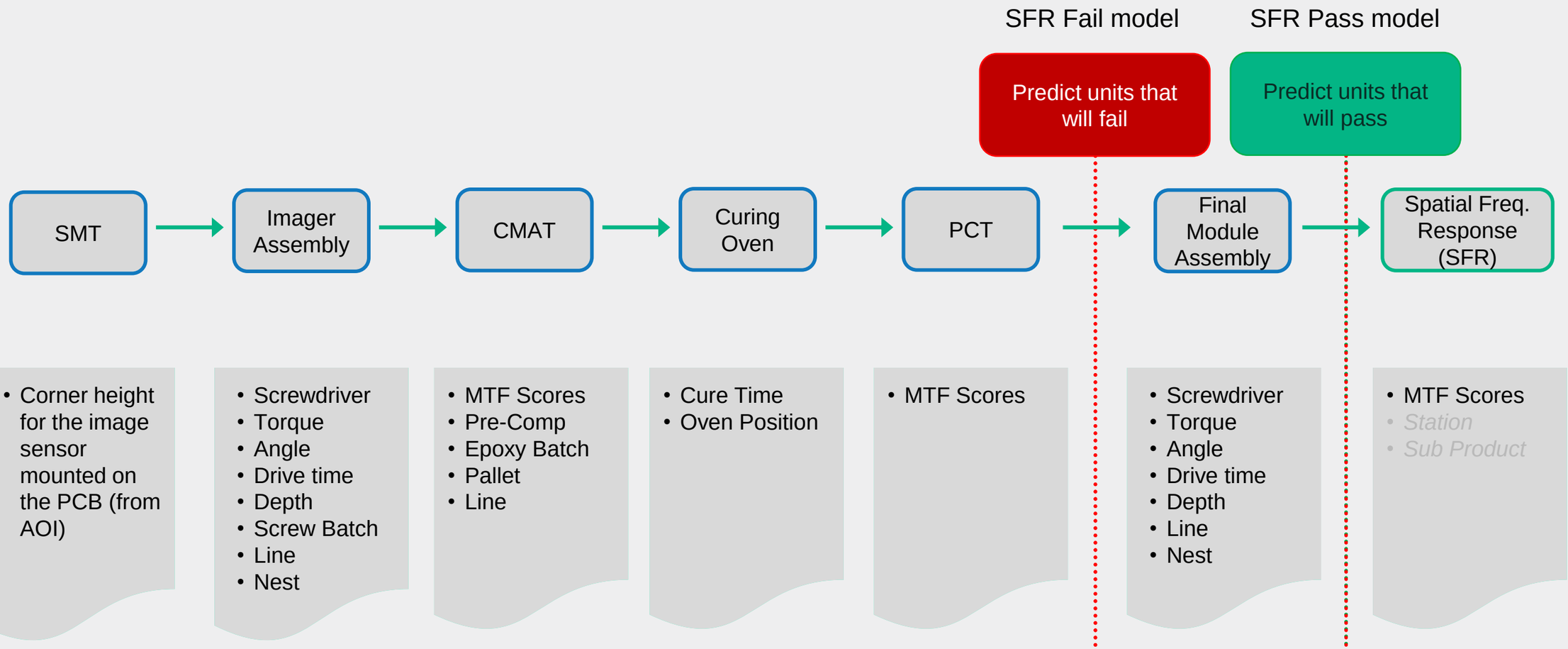
Preventative Maintenance Using Waveforms

AI-based alerting on anomalies

- **Multivariate**
  Holistic view of the waveform
- **Minimal configuration**
  Automatic identification
  and extraction of relevant features
- **Feedback loop**
  Improves with user feedback



Next Operation Optimization / Lean Coverage

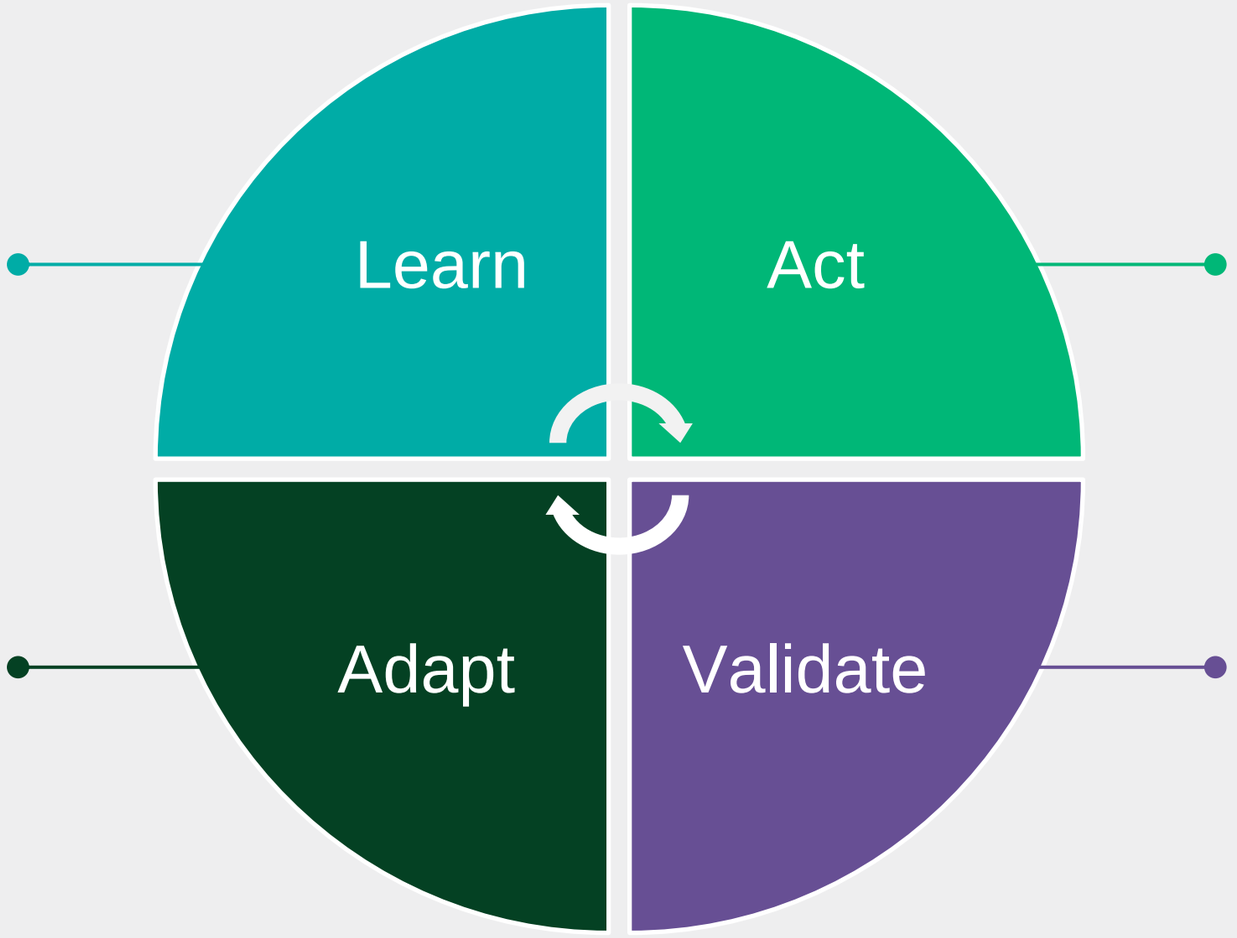


Implementing AI at the Edge

It sounds like a great idea – what's preventing it from happening?

The Full Machine Learning Lifecycle

Train a model with data and evaluate business value

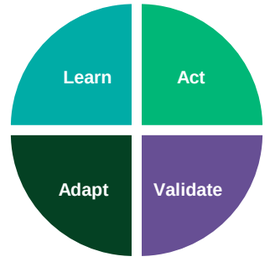


Deploy and act upon the model

Understand changes and update model/process

Monitor data and model performance to identify changes

ML Challenges



Learn

- **Getting data**
  Data scientists waste time getting and organizing data
- **Feature extraction**
  It is difficult to extract complex features from the data set
- **Freedom of choice**
  Data scientists want to use their favorite tools and the latest-and-greatest algorithms

Act

- **Complex “plumbing”**
  Data scientists waste time dealing with the “plumbing” associated with getting a model into production
- **Actionability**
  Taking action requires integration with equipment and systems
- **Distributed mfg.**
  Issues compounded in distributed, outsourced mfg.

Validate

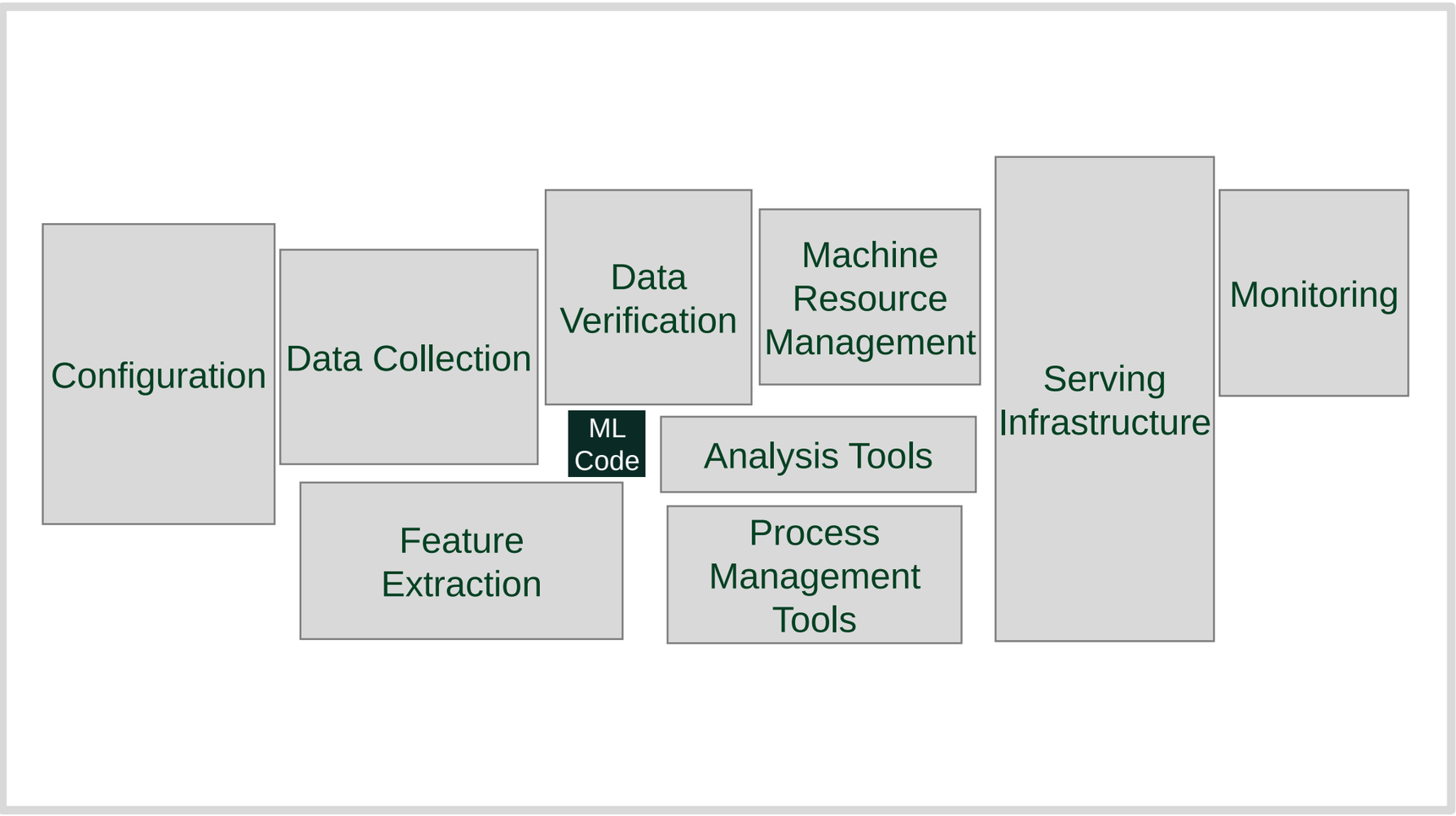
- **Ongoing validation**
  Production models need to be validated all the time
- **Ongoing data collection**
  Data collection becomes an ongoing concern
- **Technical debt**
  Data scientists end up spending time monitoring “old” projects instead of investing in new ones

Adapt

- **Stale models**
  Production changes inevitably cause models to go stale
- **Human-in-the-loop**
  Users need to review the results of a model and provide feedback to fine-tune it
- **Relearning**
  Model relearning is often manual

Hidden Complexity – the Google View

It's all about the infrastructure



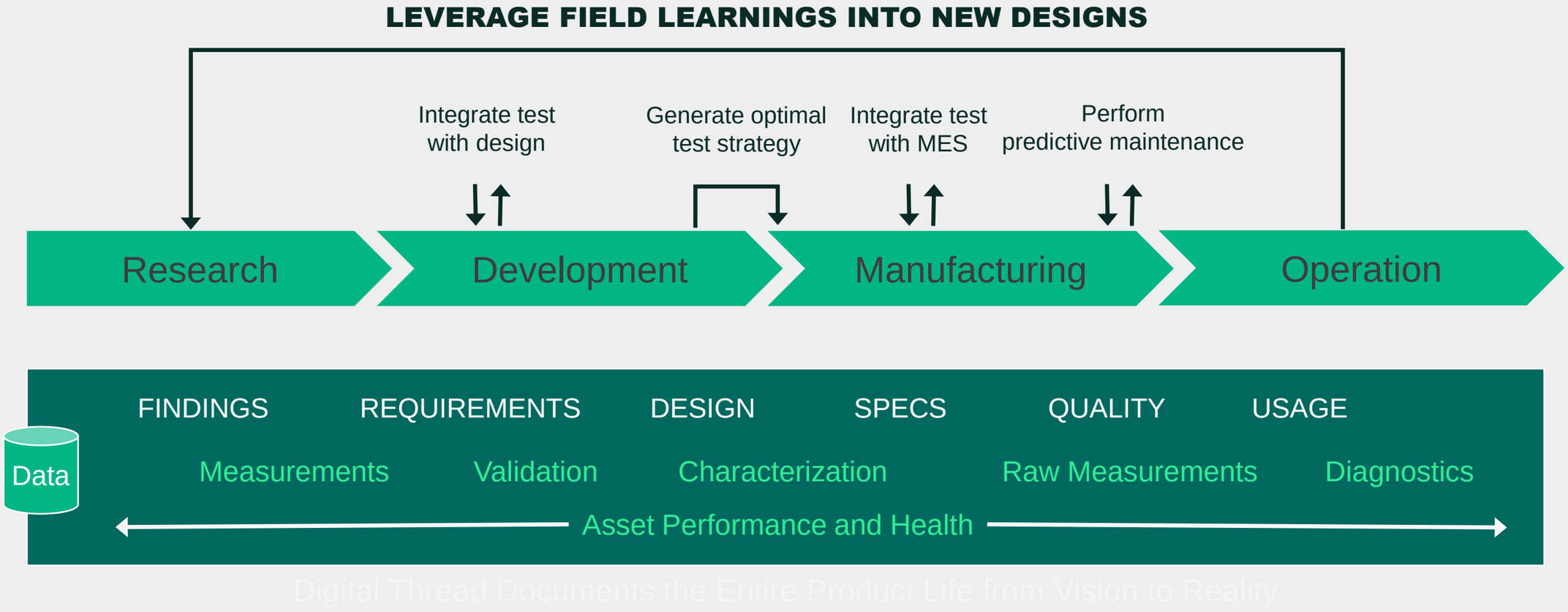
Source: Google article from 2014: Hidden Technical Debt in Machine Learning Systems
https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

Emerson's T&M Platform Vision

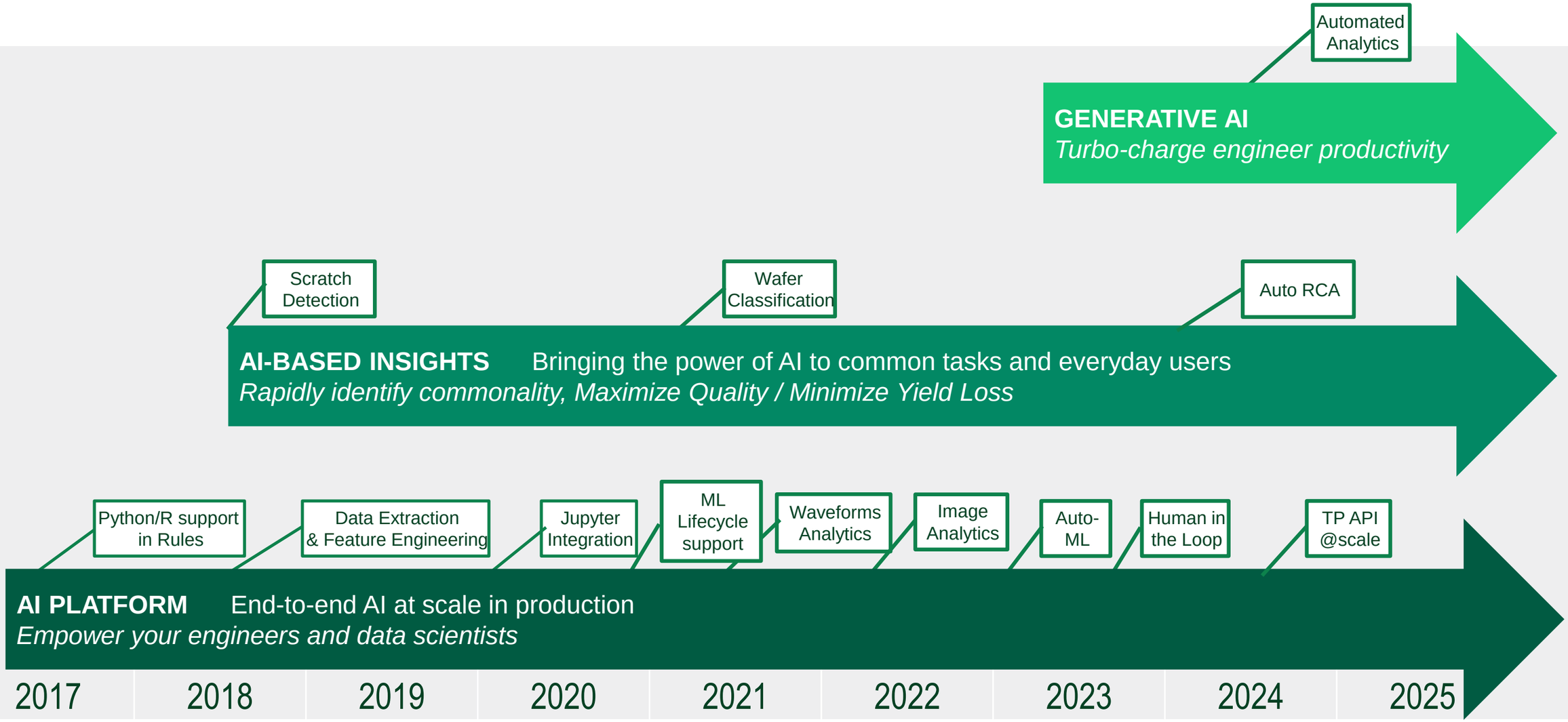
Digital Transformation across the Enterprise

Software-Defined Automation

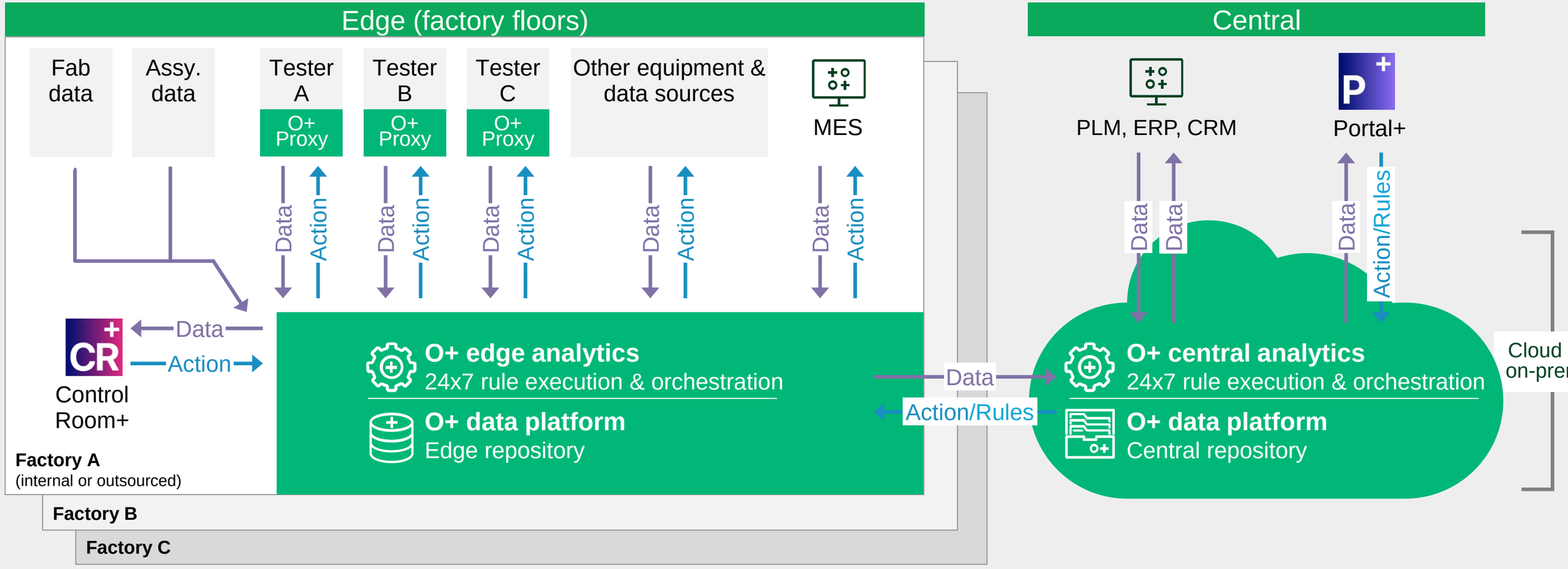
Enables Optimization and Continuous Insight



Evolution of AI on the Optimal+ "GO" Platform

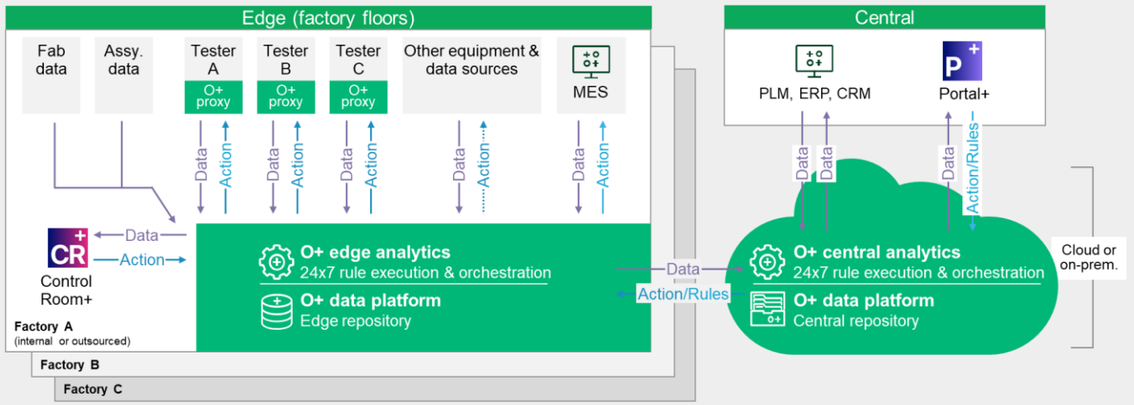


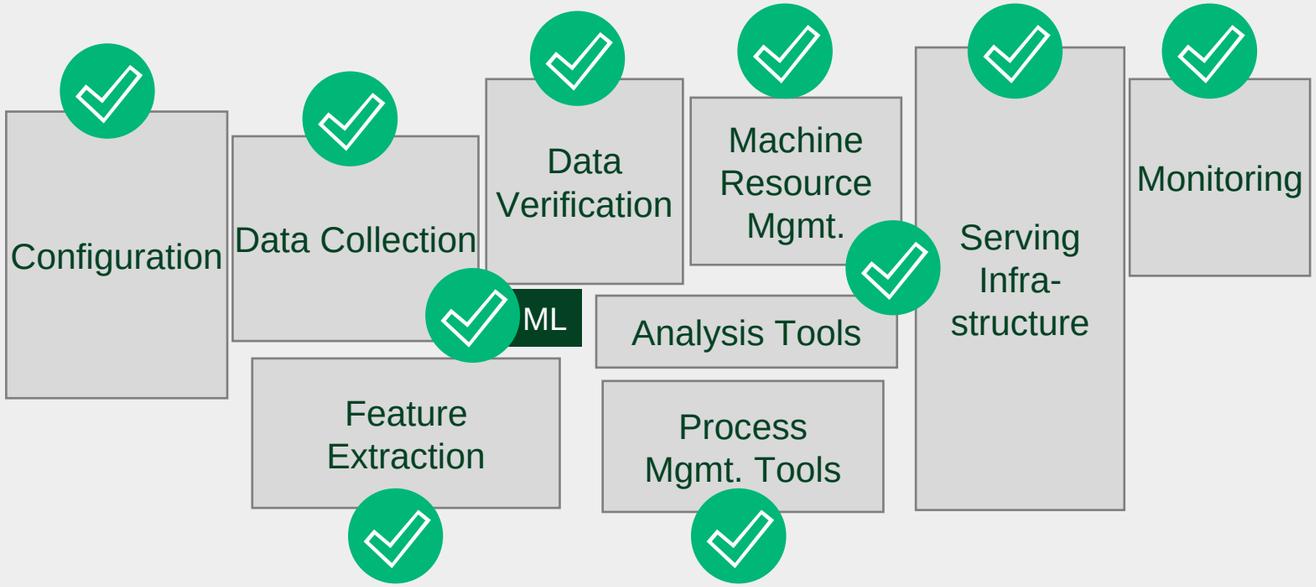
Global Operations – "GO" – Architecture



Actionable insights across all manufacturing and test processes

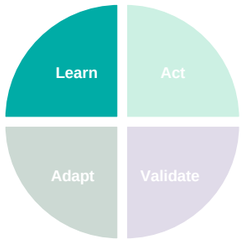
Optimal+ GO – Comprehensive ML Operations





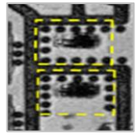
Optimal+ covers the full scope all the way through ML deployment

Example: Feature Extraction



Raw features

- Product data
- Machine data
- Test and inspection data
- Process data
- Images/Scans



Domain-driven engineering features

- Wafer geography
- Cycle time
- Yield
- Custom domain features
- Derived welding parameters

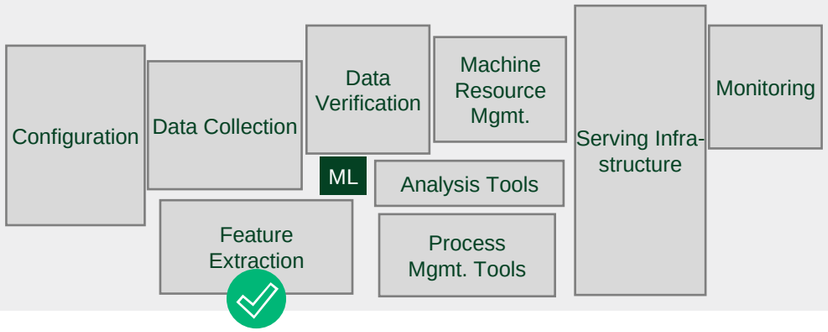
Data Science engineering features

- Interactive flow
- Normalization
- Bucketing
- Group level features
- Embeddings
- ...

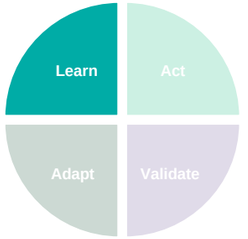
Feature reduction

- Interactive flow
- Advanced feature reduction techniques
- PCA
- SHAP values analysis (Model Explainability)
- …

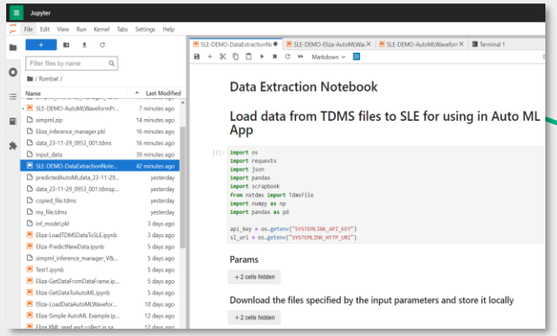
Engineering features benefit ML models but require domain expertise. The platform saves time by contextualizing data into ML-ready datasets and calculating domain-specific features.

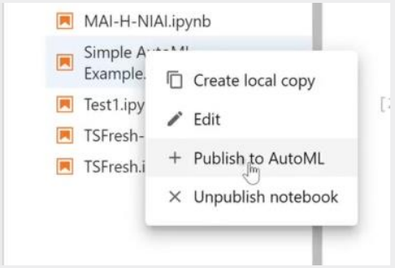


AutoML – AI for Everyone



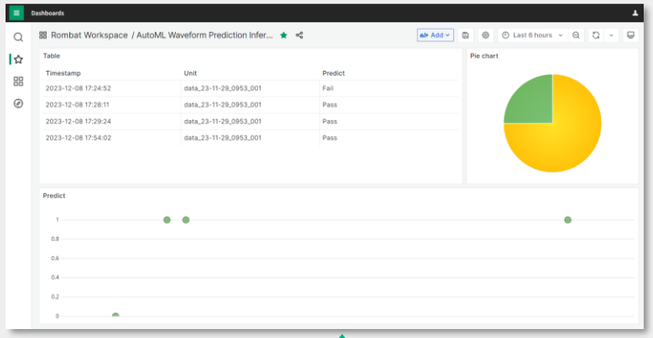
Data Extraction Notebook

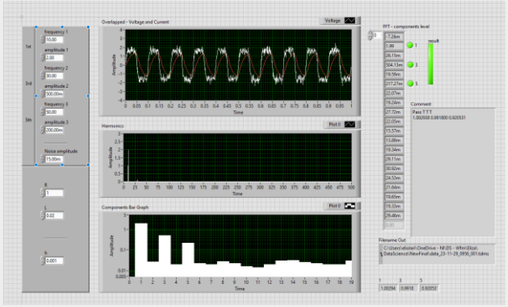




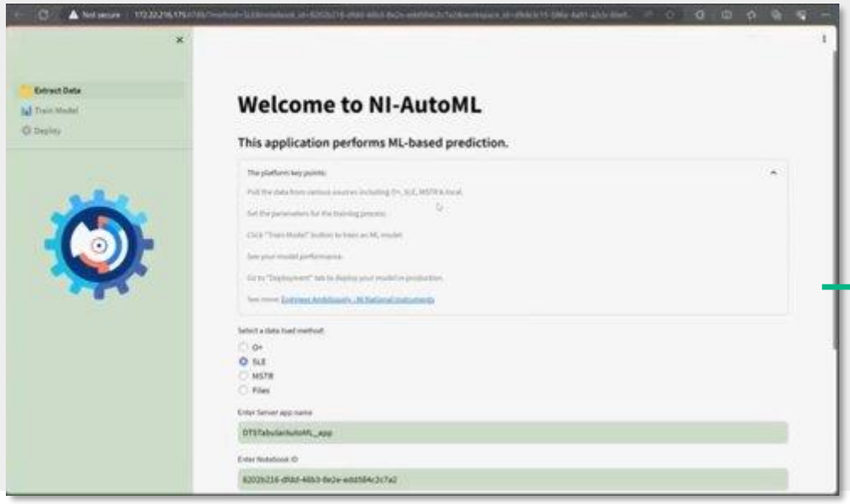
Open AutoML

AutoML Deploys Model

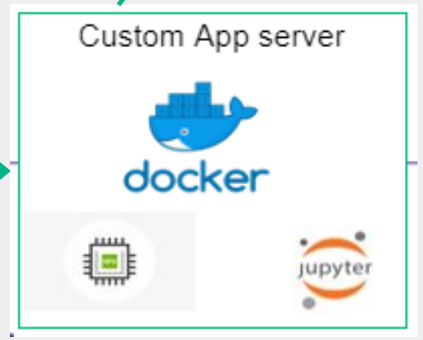


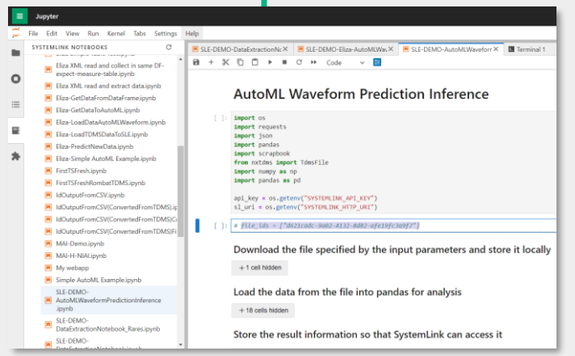


Generate Data



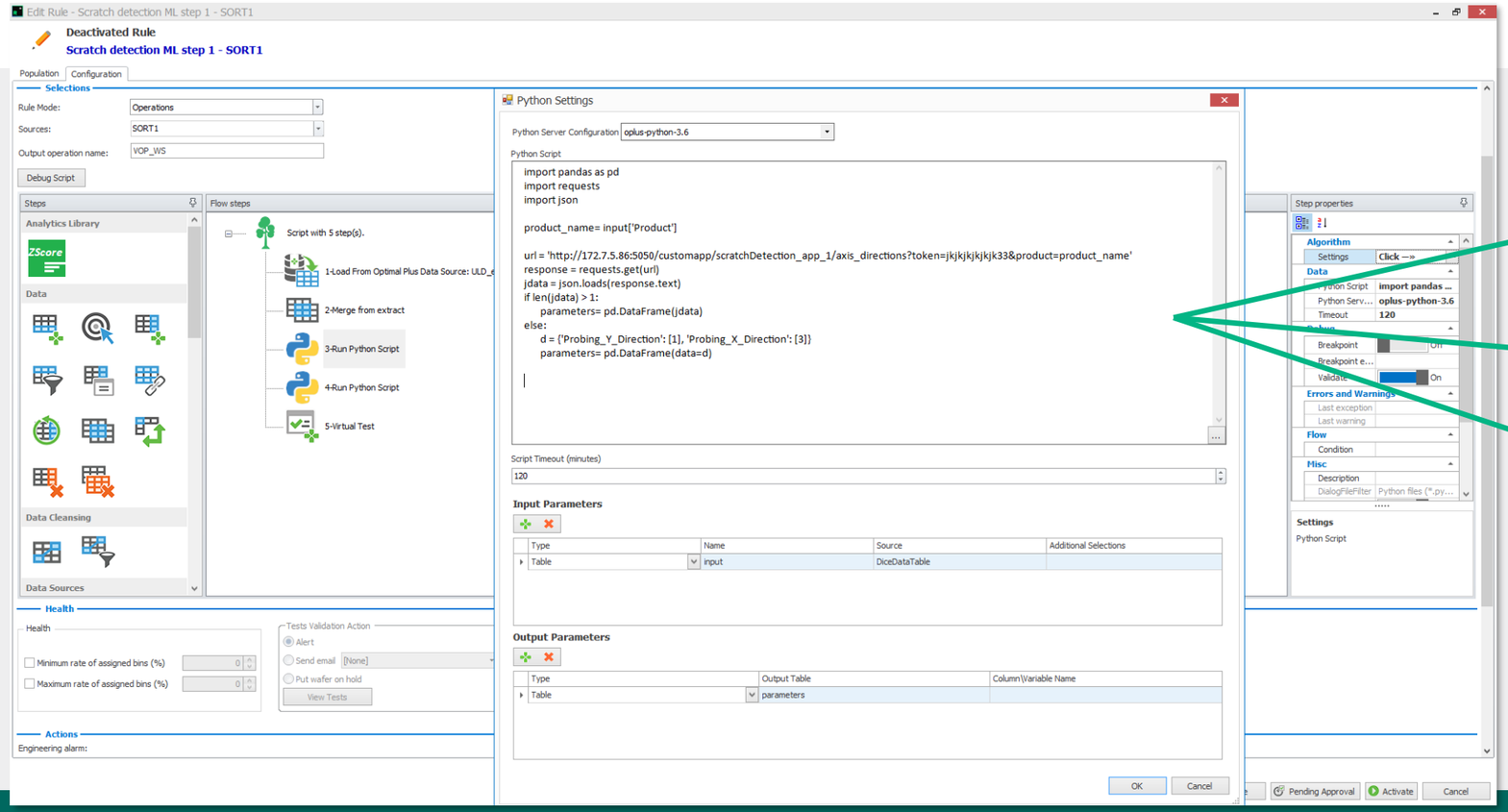
AutoML Train Models

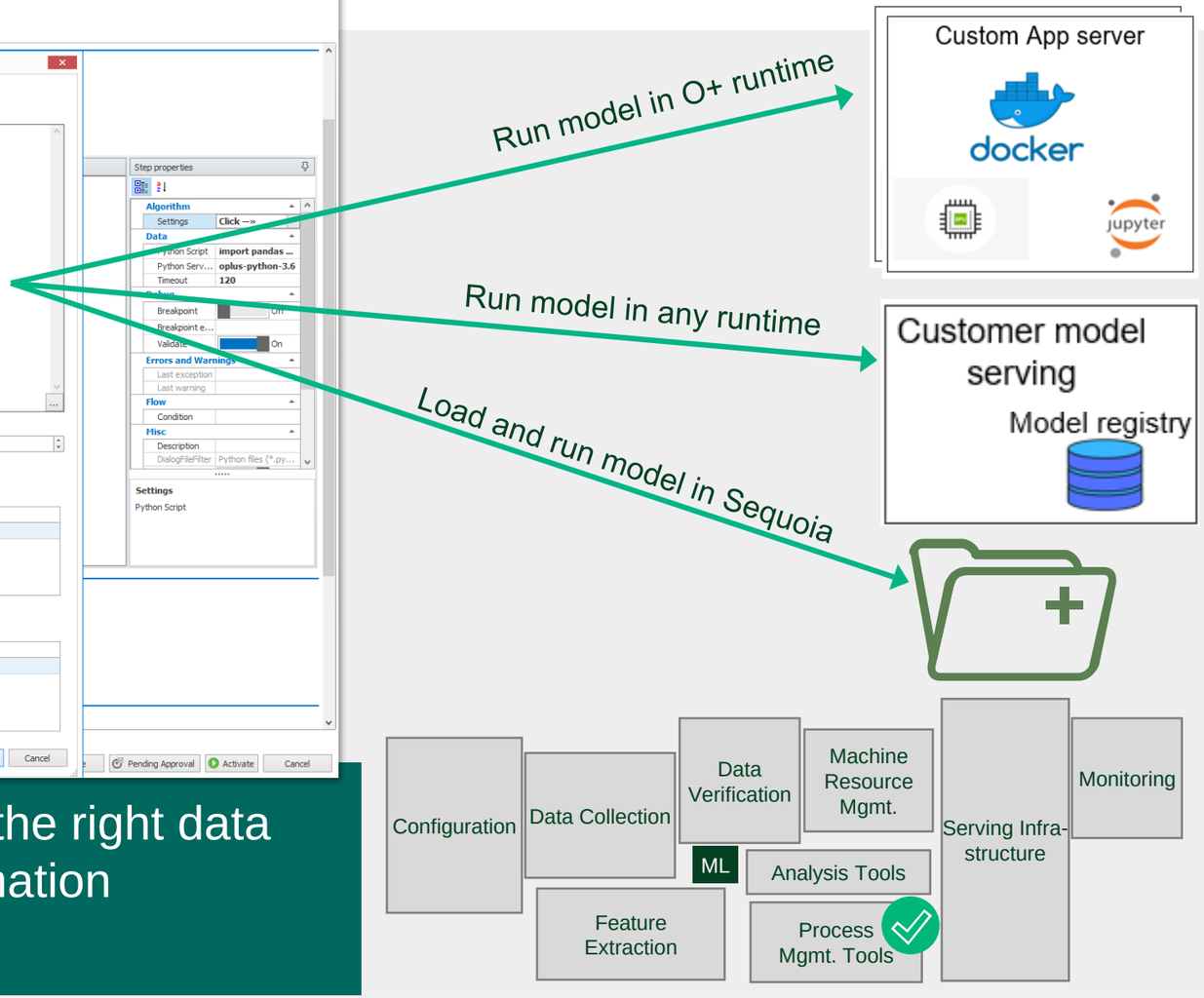




AutoML-based prediction Notebook

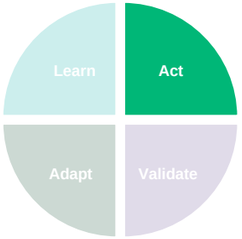
Example: Edge Inferencing using Rules

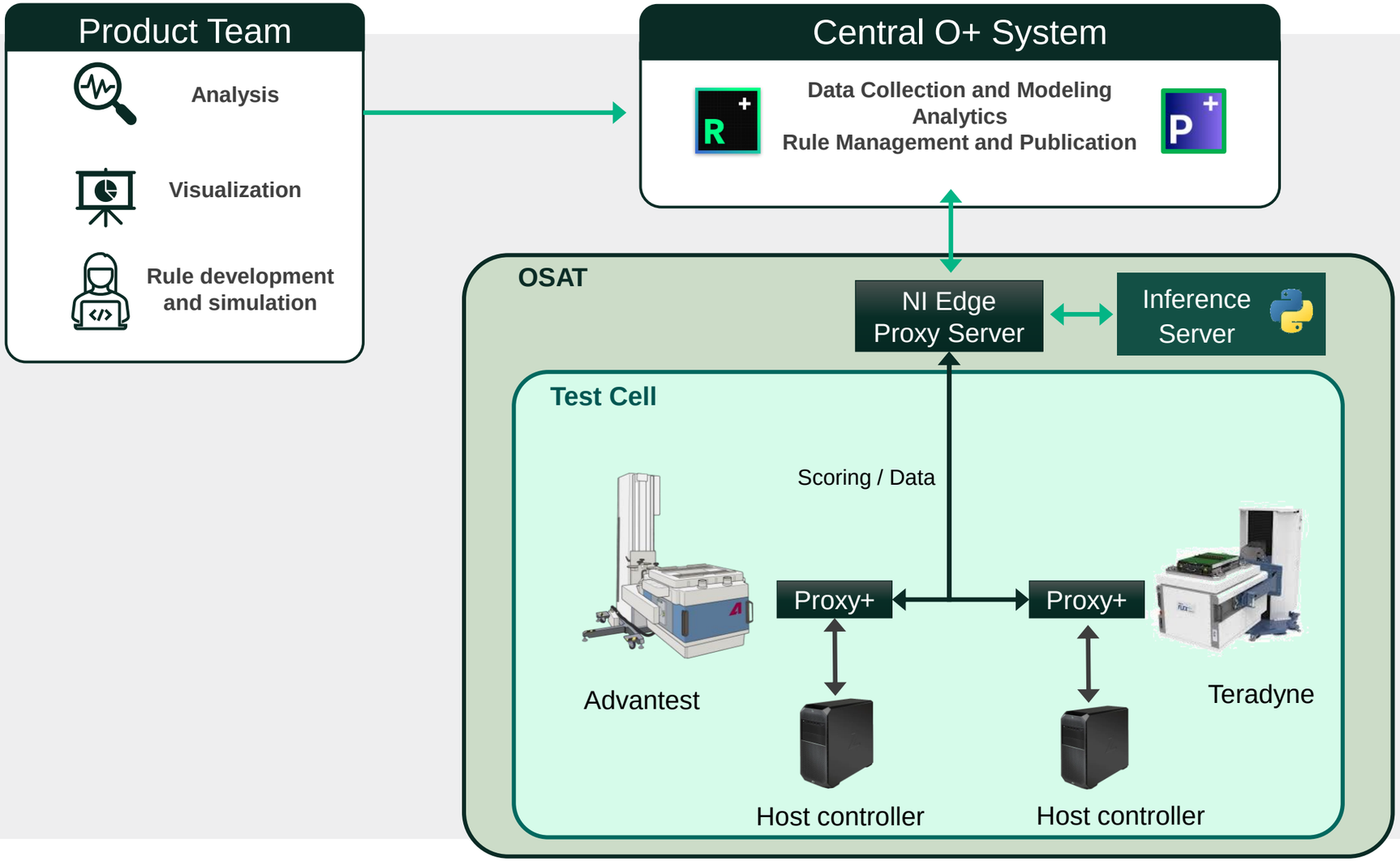




- Rules trigger model inference at the right time with the right data
- Feature engineering utilize Sequoia reusable automation
- Flexible options for model execution runtime

NI Near Real Time Solutions





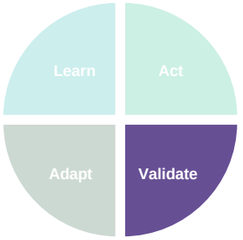
Near Real Time Option:

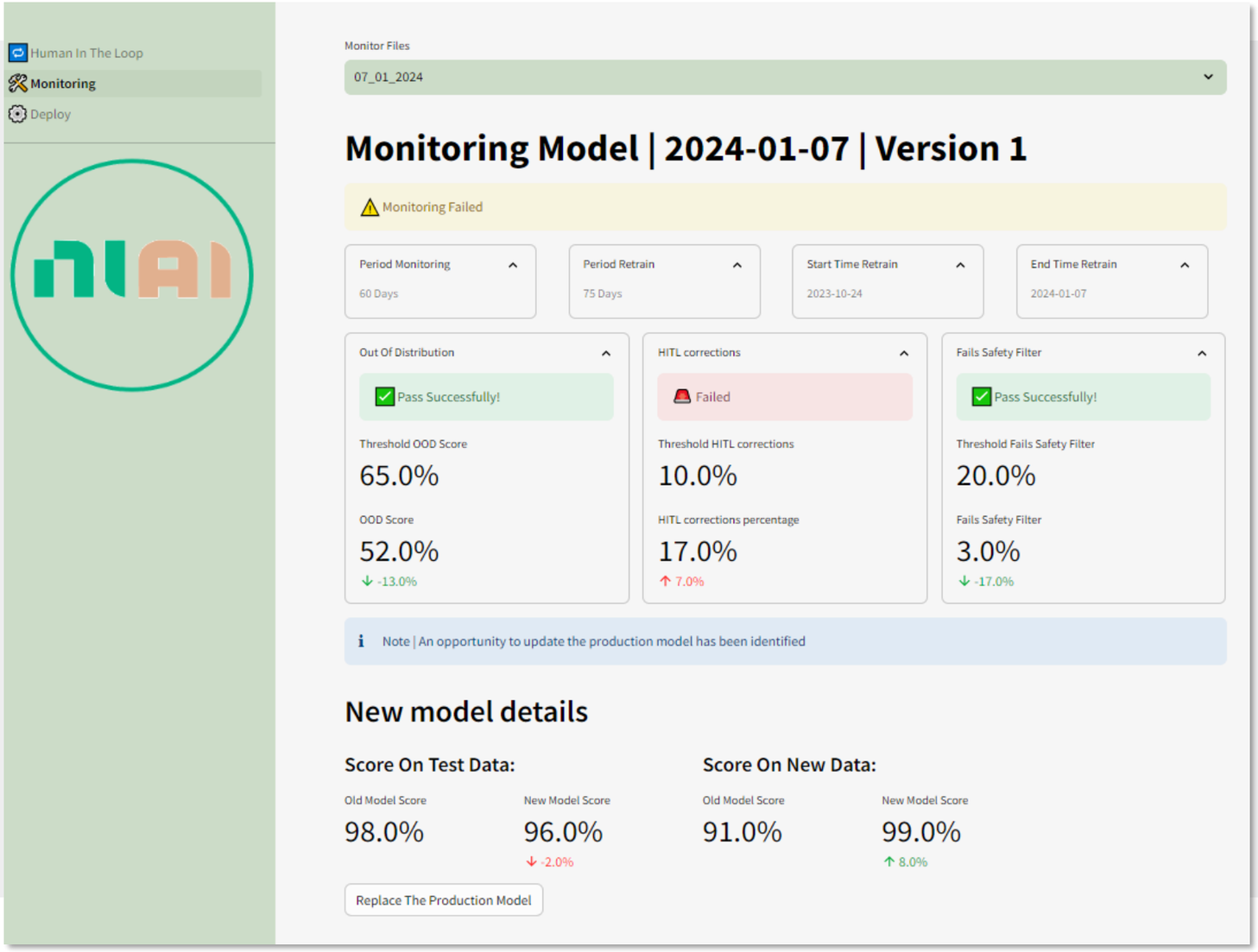
- Customer or NI developed ML Models
- Model delivery and Data Feed Forward thru NI Edge Proxy Server
- Run level (lot to lot) or device level Model Execution thru Dockers and Test Program API

Use Case:

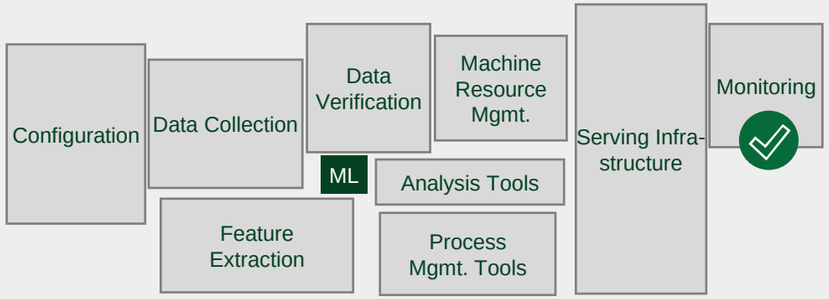
- Device binning for reduced post test operations (Burn in and/or System Level Test)

Example: Monitoring an AutoML App



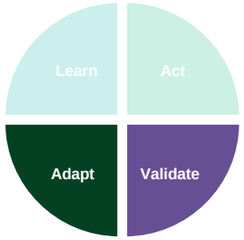


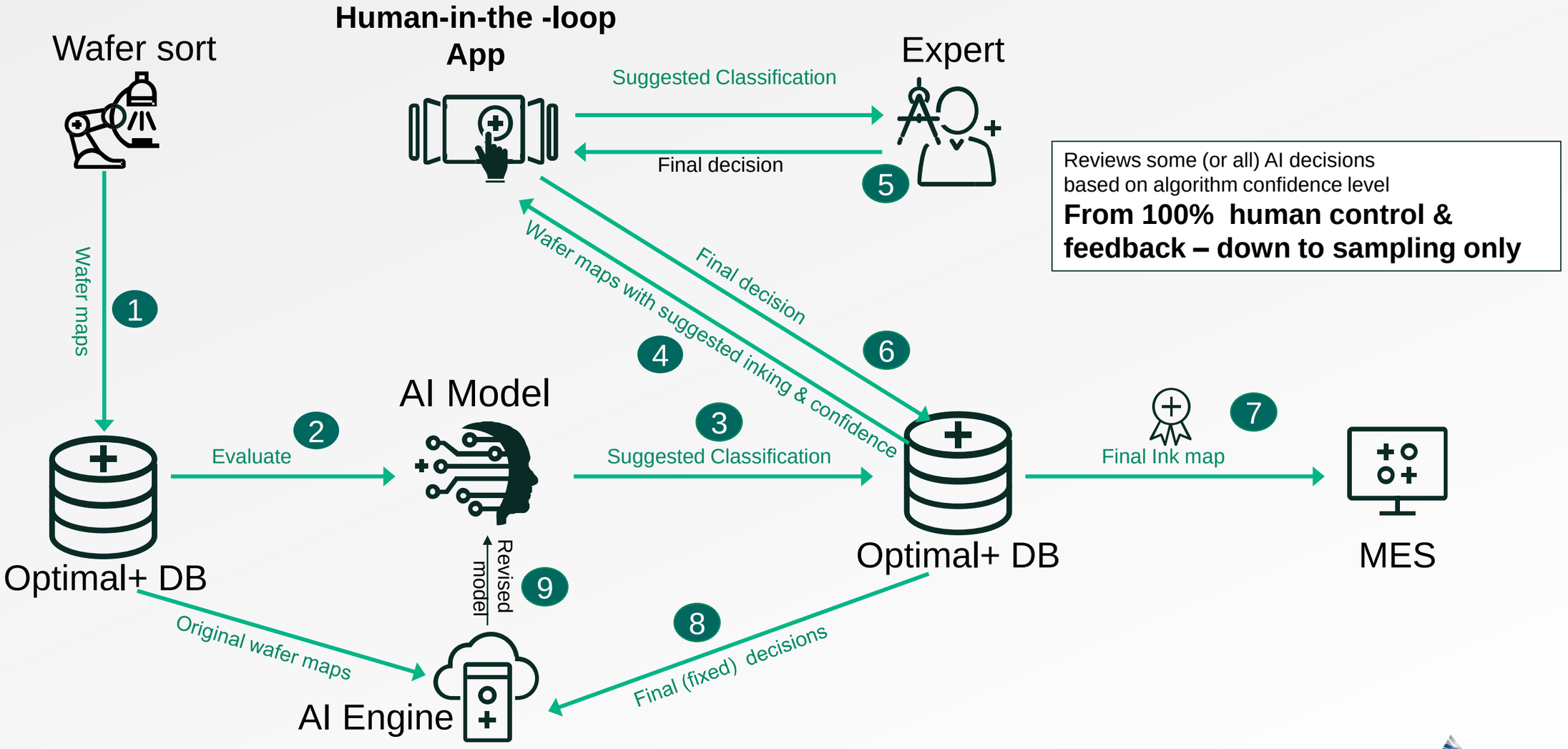
- Automatically identify deteriorating model results
- Relearn on latest data to improve model



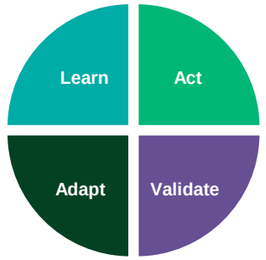
Human-in-the-Loop

Improve model while gaining user confidence





The Full Machine Learning Lifecycle



Learn

- **Getting data**
  Data is collected and harmonized all the time, making it available at the click of a button
- **Feature extraction**
  Advanced features are extracted via out-of-the-box capabilities (e.g. geographic and parametric outliers)
- **Freedom of choice**
  Full support for data science platforms

Act

- **Complex “plumbing”**
  Plumbing is handled under the hood by the Optimal+ infrastructure
- **Actionability**
  Integration with equipment and systems is part of the Optimal+ deployment
- **Distributed mfg.**
  Optimal+ is deployed across the entire mfg. ecosystem – internal and outsourced

Validate

- **Ongoing validation**
  Standard rules monitor ML models for excursions
- **Ongoing data collection**
  Data collection and harmonization is already fully automatic
- **Technical debt**
  24x7 monitoring frees data scientists for their next project

Adapt

- **Stale models**
  Automated rules detect when models are going stale and can even disable them if needed
- **Human-in-the-loop**
  Users can browse results and provide feedback directly
- **Relearning**
  Model relearning can be partially or fully automated

What’s Next for AI at the Edge

NI Global Operations

Real-time Application Enablement Layer (DIY & Canned models)

ADVANTEST® Edge, Nexus, & ACS

TERADYNE UltraEDGE & Archimedes

(standalone)

Test Floors

Testers

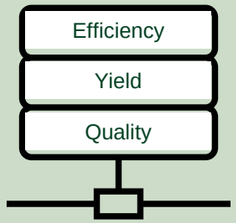
Test throughput, speed, and consistency

Real-Time Execution

Edge Servers

Data ingestion, monitoring and action control

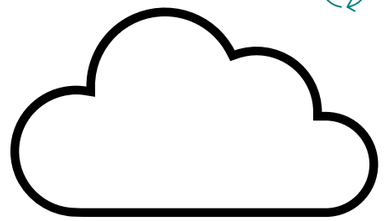
Model Execution



Central

Big data observability, insights and model creation

Model Execution



Data, Models, Actions, Monitoring, Etc.

EMERSON

Your Journey to AI at the Edge with NI

- Training for the data science teams
  - Theoretical training on Optimal+ and SystemLink capabilities for AI
  - Hands-on training on NI AI-related tools
  - Best practices and potential pitfalls based on extensive experience with multiple customer use cases

- Topics we can cover, tailored for your use case
  - How to use (inference) your existing models on O+ / SystemLink
  - Best practices for model prediction management
  - Leveraging NI tools for training and deployment
  - Developing Human In The Loop applications easily with NI (for user feedback)
  - Monitoring and auto-retraining your model
  - Leveraging the AutoML application for your use case
  - Integrating your application into NI or other applications

Other “Connectivity, Data, and Insight” Activities

May 21

Modernizing Your Lab Operations
10:15-11:15

Automating the Lab with SystemLink (Automotive Track)

Gaining Product Insight From Your Test Data
11:30-12:30

Generative AI to Accelerate Test Workflows
2:00-3:00

Modern Lab Operations with SystemLink: Hands-On
2:00-4:15

Maximize Your Lab with SystemLink Software
3:15-4:15

May 22

Analytics From Wafer To Reel To Strip (Semiconductor Track
10:15-11:15

Modern Lab Operations with SystemLink: Hands-On
10:15-12:30

Modernizing Your Lab Operations
11:30-12:30

AI at the Edge
1:30-2:30

From Concept Through Execution: Analytics in Action

Analytics From Wafer to Reel to Strip

SystemLink Ask Me Anything
2:45-3:45

ni

May 23, NIC 9:00-1:00

SystemLink User Group Meeting **O+ User Forum**

New **SystemLink Learning Courseware (V/ILT)**

Managing Systems and Assets with SystemLink Enterprise

NI CONNECT 2024 AUSTIN

EMERSON

EMERSON