BakerHughesC3.ai

Machine Learning and Best Practices from BakerHughesC3.ai

Tural Garibov, Head, Baker Hughes, CIS & Türkiye

Agenda

- What is Machine Learning?
- Types of Machine Learning
- Tuning a Machine Learning model
- Evaluating model performance
- Selecting right problem for ML
- Real-life examples from Oil and Gas
- Demo of **BHC3 Ex Machina** (no code)
- Best practices and reading suggestions
- Next session?

BakerHughesC3.ai overview

Baker Hughes

Deep O&G domain expertise and global reach

Provide critical feedback to product roadmap based on direct customer feedback

BakerHughesC3.ai

Fullstream AI

Accelerating
Digital Transformation
to take energy forward

C3.ai

Leading AI technology Platform

Proven enterprise-grade platform for designing high-value AI applications

Baker Hughes is a partner, an investor, a customer and exclusive distribution channel

BakerHughesC3.ai Roadmap for O&G

Drilling Risk & Productivity

Well Placement & Completion

Integrity

Production Optimization

Reliability

LNG Operations

Asset Investment Planning

Hydrocarbon Loss Accounting

Demand Forecasting

Refining Analytics

Customer Segmentation & Targeting

Fuel Station Analytics

BHC3 AI Suite™

Exploration

- Petrel
- OpenWorks
- Eclipse
- IFM

Development

- COMPASS
- WellView
- BEACON

Production

- LOWIS
- EAM
- Energy Components

Transportation

- OpenLink
- SAP TSW

Refining

- Maximo / CMMS
- SCADA / OSI Pi
- Workforce Management

Retail

- POS
- SCM
- Customer / Loyalty

BakerHughesC3.ai

BHC3 AI Suite for Digital Transformation

C3.ai Integrated Development Studio (IDS)

- C3.ai Data Studio
- C3.ai Applications Studio
- C3.ai ML Studio
- C3.ai DevOps Studio

Data Integration

- SAP Hana
- Oracle
- Salesforce
- IBM Maximo
- OSI PI
- GE
- Azure Data Lake
- Weather
- Commodity Data

Extensible Data Models

- Cross-Industry Foundation

- Predictive Maintenance
- Inventory
- Supply Chain
- Energy Management
- CRM
- Fraud

- Aerospace
- Manufacturing
- Telco
- Utilities
- Oil and Gas

BHC3 AI Suite

Data Persistence

- NoSQL Storage
- RDBMS
- HDFS
- Key Value Store
- Distributed File System
- Cloud Object Storage
- File Systems

Connectors, Extensions

- Canonical
- Source
- Expressions
- Transform
- IoT

Machine Learning Framework and Services

- ML Pipeline
- Ex Machina

Continuous Analytics Processing

- Batch
- Map Reduce
- Queue
- Stream

Platform Management Services

- Access Control
- Auto Scaling
- Multitenancy
- Time Series
- Analytics
- Deployment
- Profiling
- Users
- APIs
- Encryption
- Roles & Responsibilities
- Auditing
- Logging
- Scheduler
- Authentication
- Monitoring
- Systems Management

Application Development Tools

- Jupyter
- Eclipse
- R Studio

UI and Data Visualization Tools

- Alteryx
- Domo
- Qlik
- Micro Strategy
- Tableau

UI Frameworks

- React
- JS
- Angular

Building Blocks for AI

Data Collection

Testing and Debugging

Resource Management

Configuration

Serving Infrastructure

Model Analysis

ML Code

Data Verification

Process Management

Monitoring

Automation

Feature Engineering

Metadata Management

What is Machine Learning?

How to write a software code to recognize image of a cat?

Approach: Traditional structured programming.

Develop rules to identify cats and codify these rules into software.

Effective solution, but complex due to:

- Rules required to classify cats
- Difficulties related with each new parameter, breed, color, orientation or location within an image

→

Approach: Machine learning.

Collect large number of pictures with cats and training the models to learn classification.

BakerHughesC3.ai

Types of Machine Learning

Categories	Supervised Learning		Unsupervised Learning		Reinforcement Learning
Approaches	Classification	Regression	Dimensionality Reduction	Clustering	Decision Making
Traditional	• Support vector machines (SVM) • XGBoost • Gradient-boosted decision trees (GBDT) • Random forest	• Linear regression • Ridge regression • Random forest	• Principal component analysis (PCA)	• K-means • Gaussian mixture model (GMM) • Density-based special clustering (DBSCAN)	• Monte Carlo • Markov decision process • Temporal difference learning
Deep Learning	• Multi-layer perceptrons (MLPs) • Convolutional networks • Long short-term memory (LSTM)	• Multi-layer perceptrons (MLPs) • Convolutional networks	• Auto-encoders	• Deep Gaussian mixture model (DGMM)	• Deep Q-learning • Hidden Markov models (HMM)

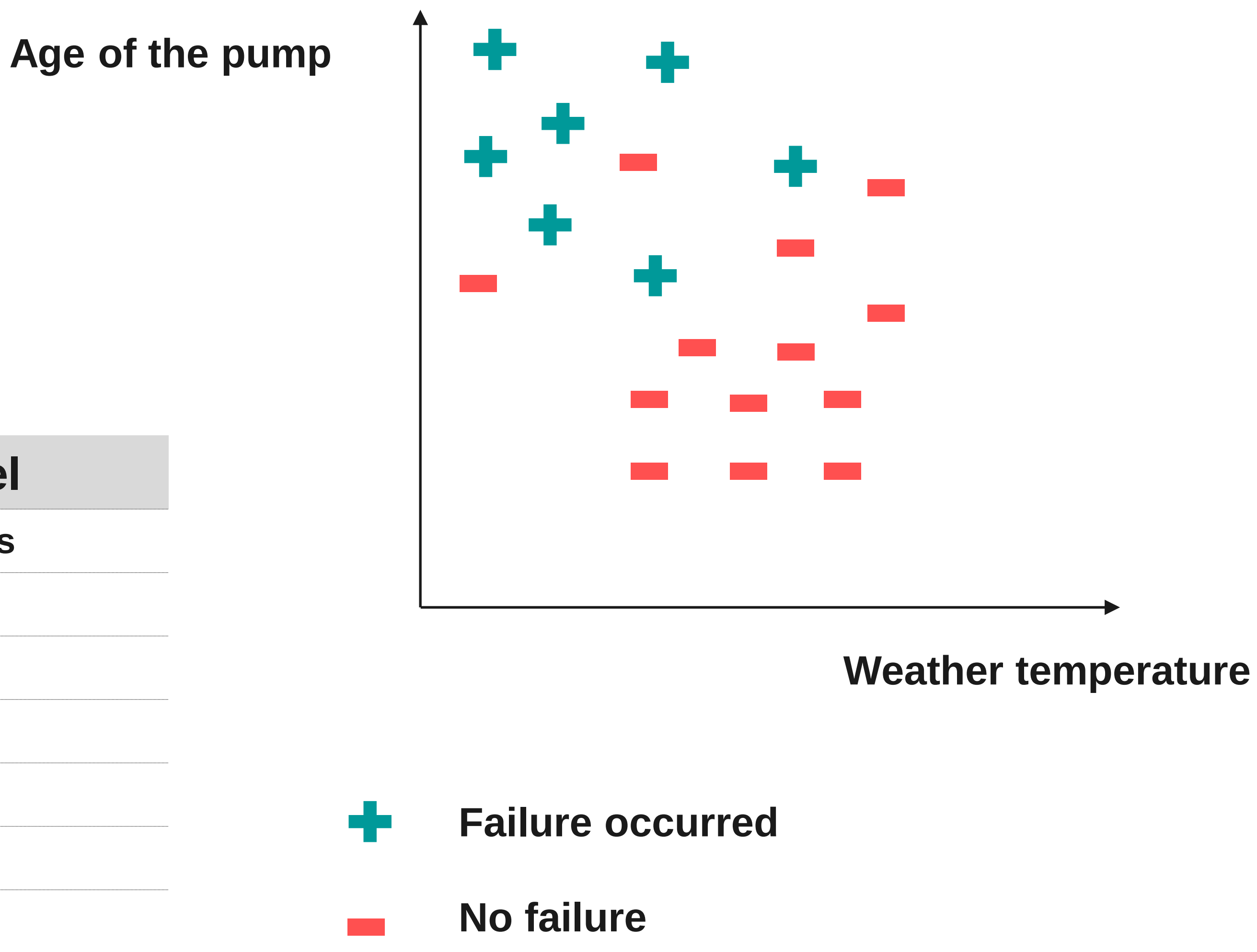
Feature Engineering for Machine Learning

We can examine into more details in this example. Let’s say we want to predict whether the pump will fail or not be based on two characteristics:

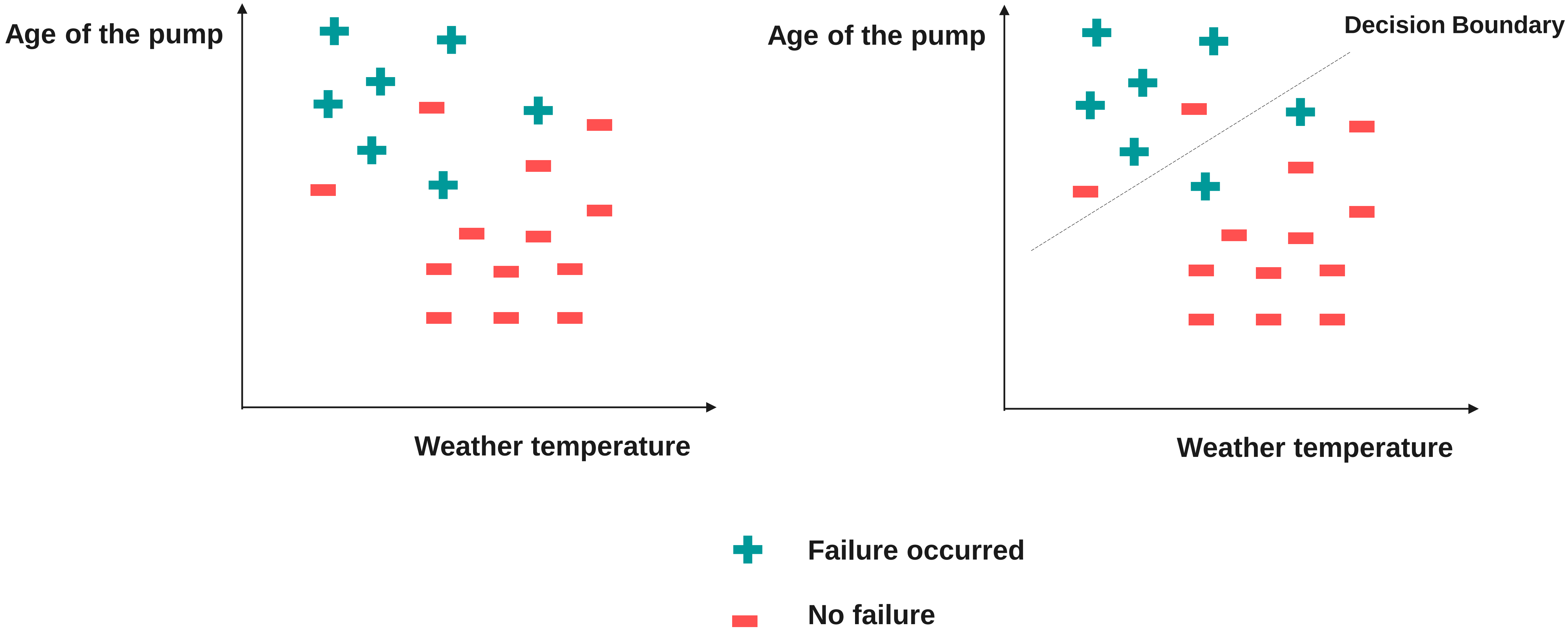
- Weather temperature (in C°), and
- Age of the pump (period of use) in years.

We prepare the following historical data:

Features		Label
Temperature	**Age**	**Status**
5 C°	20 years	+
15 C°	3 years	-
25 C°	10 years	-
30 C°	5 years	-
35 C°	8 years	+
10 C°	15 years	-
…	…	…



Feature Engineering



BakerHughesC3.ai

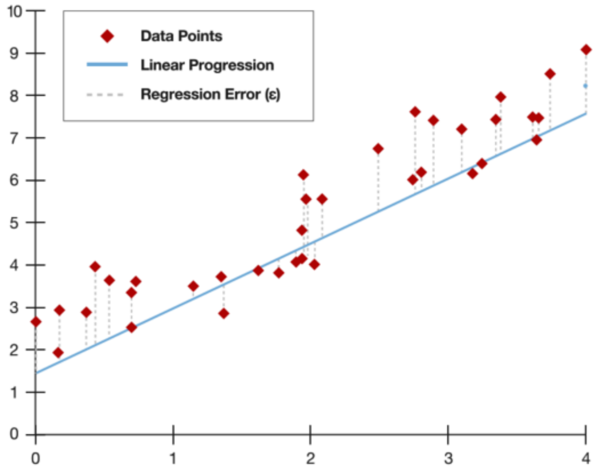
Loss Function

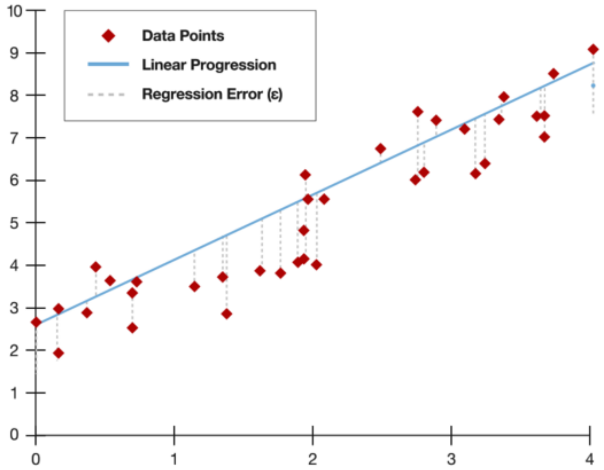
Loss function serves as objective AI / ML algorithm is seeking to optimize. Example of loss function is mean squared error (MSE), which is often used to optimize regression models. MSE measures the average of squared difference between predictions and actual output values:

$$J\,(\theta) = MSE = \frac{1}{n}\sum(\widehat{y}_k - yk)^2$$

Where $\widehat{\boldsymbol{y}_k}$ represents a model prediction, $\boldsymbol{y_k}$ represents an actual value, and there are $\boldsymbol{n}$ data points. Over-relying on loss functions as an indicator of prediction accuracy may lead to erroneous model setpoints. For example, the two linear regression models shown in the following figure have the same MSE, but the model on the left is underpredicting while the model on the right is over-predicting.

Under-predicting





Over-predicting

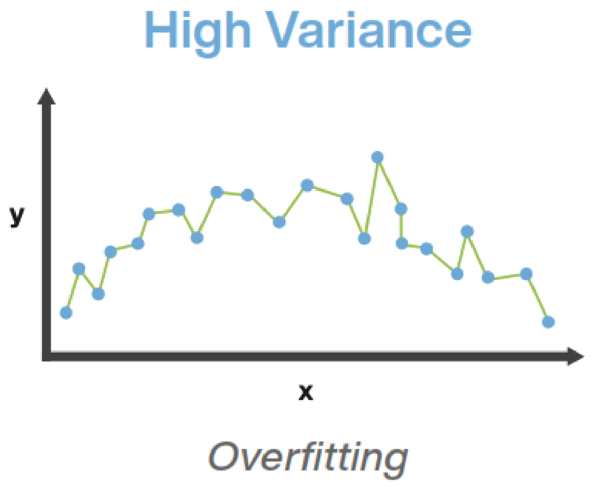
BakerHughesC3.ai

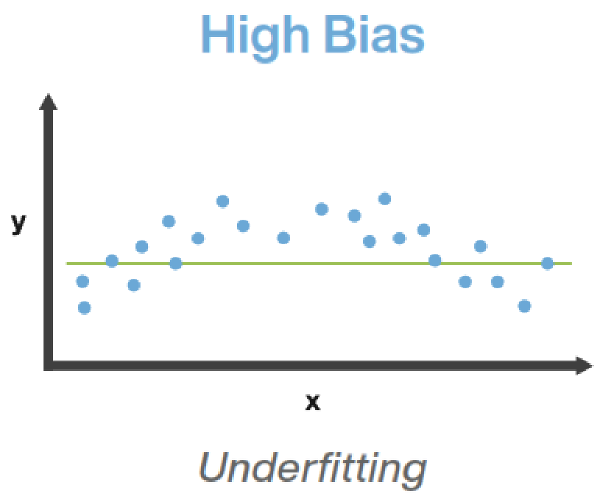
Regularization

Regularization is a method to balance overfitting and underfitting a model during training. It's a technique to adjust how closely a model is trained to fit historical data.

One way to apply regularization is by adding a parameter that penalizes the loss function when the tuned model is overfit.

More regularization prevents overfitting, while less regularization prevents underfitting. Balancing the regularization parameter helps find a good tradeoff between bias and variance.





Hyperparameters

Hyperparameters are model parameters specified before training a AI/ML model, e.g. parameters that are different from model parameters during model training. Finding the best hyperparameters is an iterative and potentially time intensive process called “hyperparameter optimization.”

Examples of hyperparameters include number of hidden layers and the learning rate of deep-neural network algorithms, the number of leaves and depth of trees in decision tree algorithms, and the number of clusters in clustering algorithms. Hyperparameters directly impact the performance of a trained AI/ML model. Choosing the right hyperparameters can improve prediction accuracy.

To address the challenge of hyperparameter optimization, data scientists use specific algorithms designed for this task. Examples of hyperparameter optimization algorithms are **grid search**, **random search**, and **Bayesian optimization**.

These optimization approaches help narrow the search space of all possible hyperparameter combinations to find the best (or near best) result.

Evaluating Model Performance

		Reality	
		+	-
Model Predictions	+	True Positive	False Positive
	-	False Negative	True Negative

Failure predictions made by the model are in the top half of the square. Each prediction is either true (the pump will fail), or false (the pump will not fail) – and are therefore true or false positives.

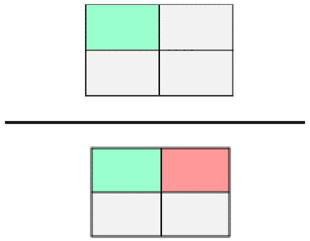
The total number of pump failure predictions made by the classifier is the number of true positives plus the number of false positives.

False negatives (bottom left) refer to failure predictions that should have been made but were not. These include pumps which failed but were incorrectly classified in the model prediction. True negatives (bottom right) refer to pumps which were correctly classified as those that will not fail.

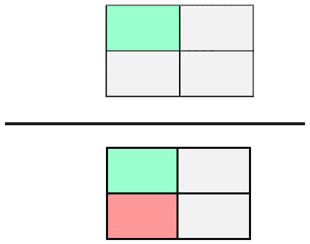
Evaluating Model Performance

True Positive	False Positive
False Negative	True Negative

$$Precision = \frac{True\ Positives}{Total\ Number\ of\ Predictions\ Made} = \frac{True\ Positives}{True\ Positives + False\ Positives} =$$



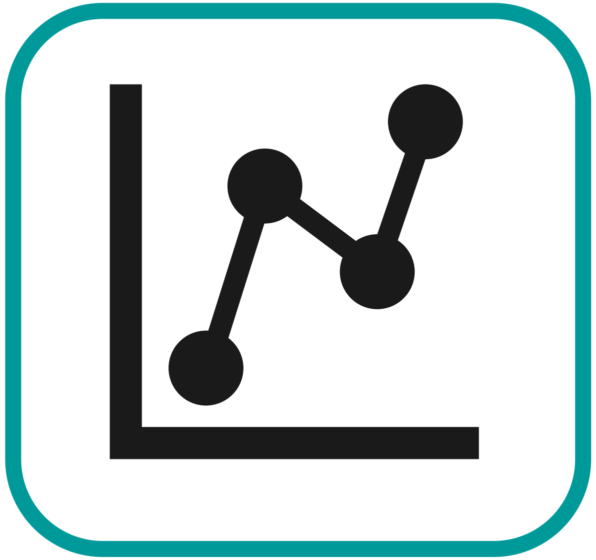
$$Recall = \frac{True\ Positives}{Total\ Number\ of\ Positive\ Cases} = \frac{True\ Positives}{True\ Positives + False\ Negatives} =$$



$$F1\ score = \frac{2\ * Precision\ * Recall}{Recall + Precision}$$

BakerHughesC3.ai

What's a good fit for ML and what is not?



Tractable Use Case for Supervised Learning:

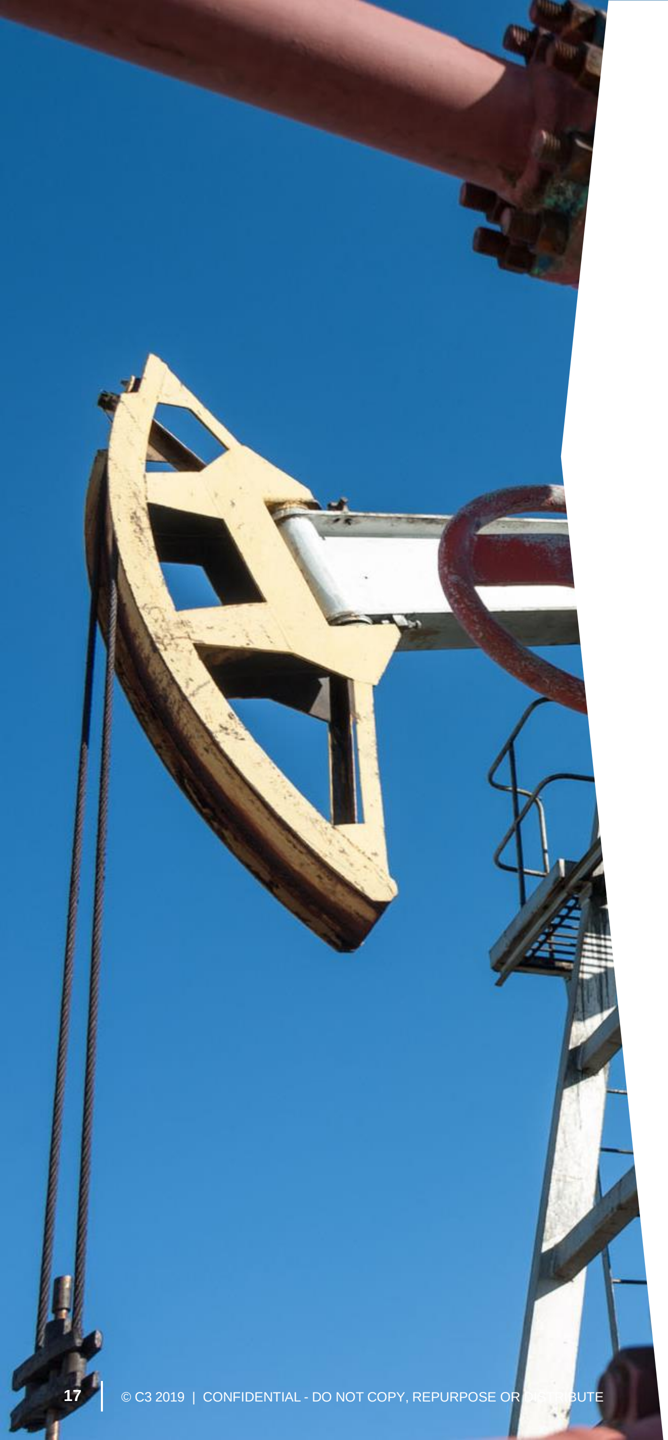
Production Forecast for oilfields

- Vast input data
- Large number of historical labels
- Possibility for humans to detect individual cases but challenging at scale

Intractable Use Case for Supervised Learning:

Predictive Maintenance for a single equipment

- Few input data. Rich data is available only for 1 or a small group of machines
- Less than 1 or 2 historical failures
- May / may not be possible for humans to detect upcoming failures



CHALLENGE ▸ Predict 3 failures across 1,031 Rod Pump Oil Wells, including Pump Failures, Tubing Failures and Rod Failures.

PROJECT DETAILS ▸

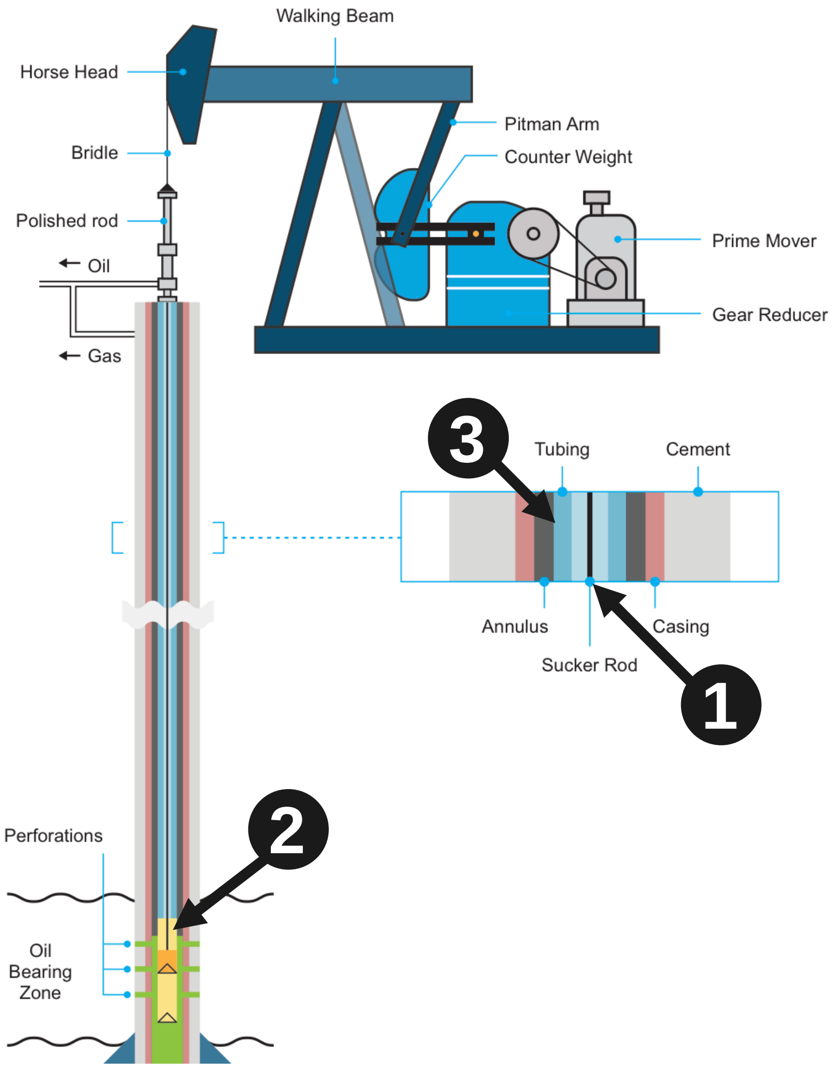
12	1,031	1047
Weeks Project Completion	Wells	Rod Pump Failures in 5 Years

RESULTS ▸

$50M+	Identified savings
22+	Days of downtime saved per failure event
80%+	Accuracy in predicting rod pump failures

BakerHughesC3.ai

Project Scope



Timeline

12 weeks

Deliverables

- Software application
- Report with analytics results and key insights
- Value proposition
- Implementation plan for deployment

Wells under analysis

1,031 wells, 2009 – 2015

Objective

Predict failure for:

1 Rod Failure 2 Pump Failure 3 Tubing Failure

BakerHughesC3.ai

CHALLENGE ▸ Build predictive maintenance application to predict unplanned gas generator outages in compressor stations for one of the largest North American gas pipeline companies

PROJECT DETAILS ▸

12	4.6B	10,000+	4,000+	300+
Weeks	Rows of raw data	Features created	Supervised machine learning model permutations	Unsupervised anomaly detection model permutations

RESULTS ▸

$82M Estimated annual savings

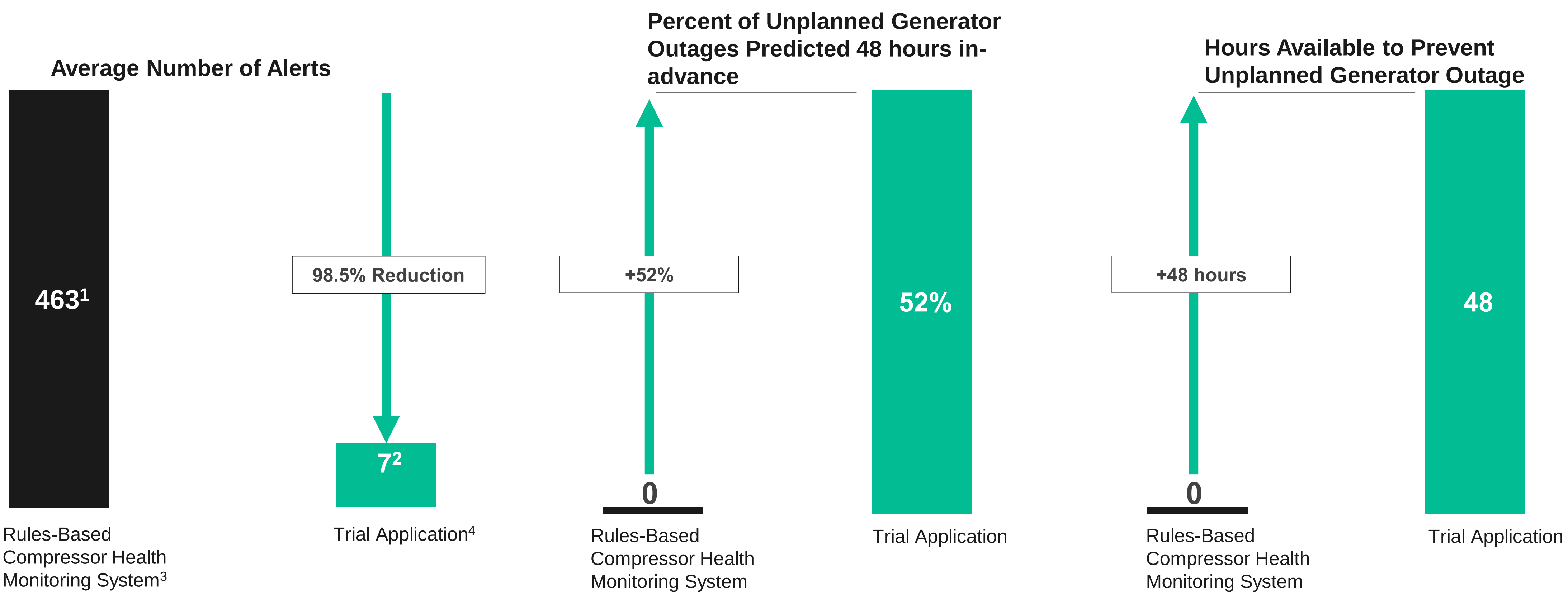
~50% Unplanned generator outages predicted 48-hour in advance

~60% Precision in generator outage predictions

~99% Reduction in spurious alerts compared to current baseline

BakerHughesC3.ai

Trial Application Predicts 52% of All Unplanned Generator Outages with Only 1.5% of the Alerts, Relative to Existing Baseline



Result: Reduced number of notifications investigated by 98.5%.

Result: 52% of all unplanned outages predicted 48 hours in-advance.

Result: Up to 48-hours available to prevent unplanned outages.

[1]Annualized number of alerts generated for one unit during Oct 2018 - Jan 2019 by existing health monitoring system, applicable for only 6 compressor units

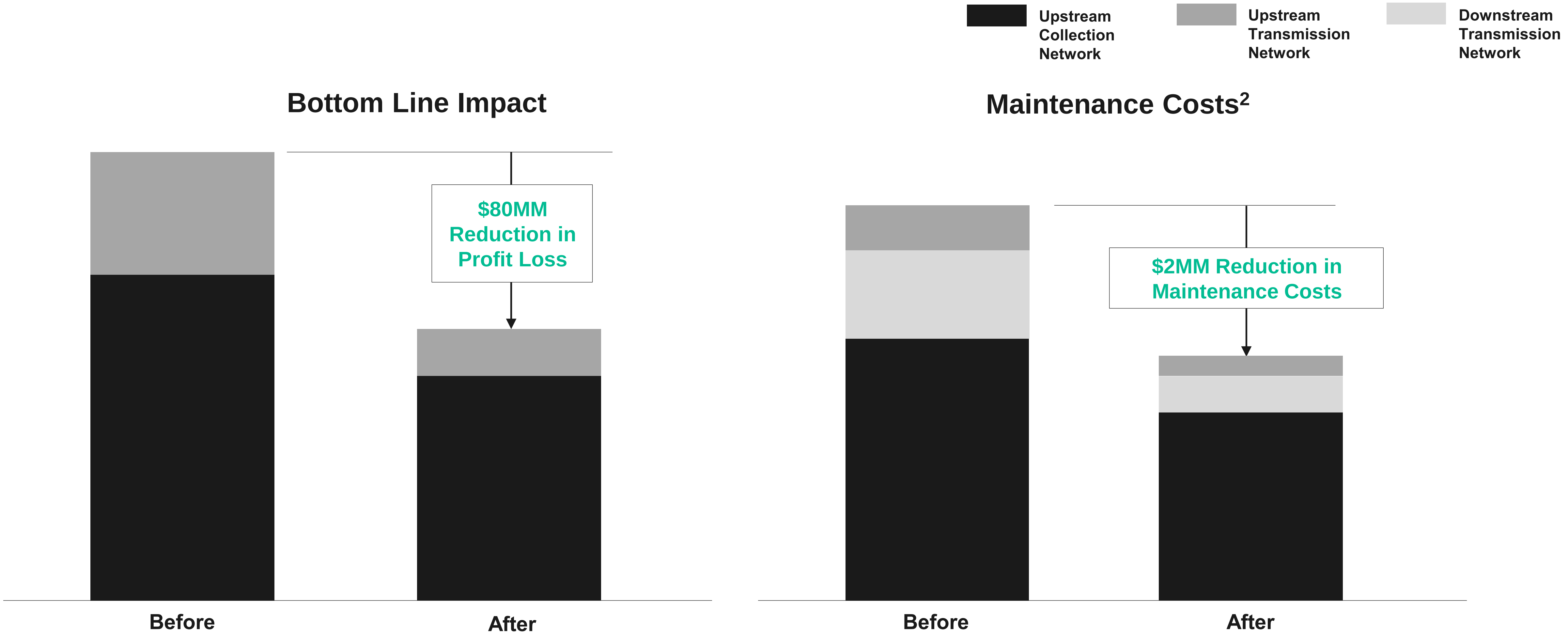
[2]Average number of outage predictions between May 2017 and April 2018 generated by machine learning algorithm across the 7 compressor units with centrifugal compressors and Rolls-Royce generators that have generator outages

[3]Unlike a machine learning-based prediction model, one individual notification generated by the existing health monitoring system may not give the operator confidence that a shutdown event is likely to occur. Operator confidence in the system's prediction capability increases as more relevant notifications are triggered.

[4]Model parameters: all-generator outage failure labels, 48-hour failure window, train-test split, periods where the unit is shutdown included in model training

BakerHughesC3.ai

Economic Value of Deploying BHC3 On Compressor Stations is up to $82 Million per Year 1



[1] Number is for all compressor units across the customer's business operations

[2] Does not include cost savings from reduction in preventative maintenance or capitalized maintenance expenditures for major overhauls/upgrades

BakerHughesC3.ai

CHALLENGE

Predict melt flow and xylene soluble content material produced in reactor, real-time and enable optimization.

PROJECT DETAILS

10
Weeks

8
Data Sources

38K
Historical lab tests to train and test against

RESULTS

95% Accuracy in prediction of melt flow and xylene-soluble content relative to lab tests

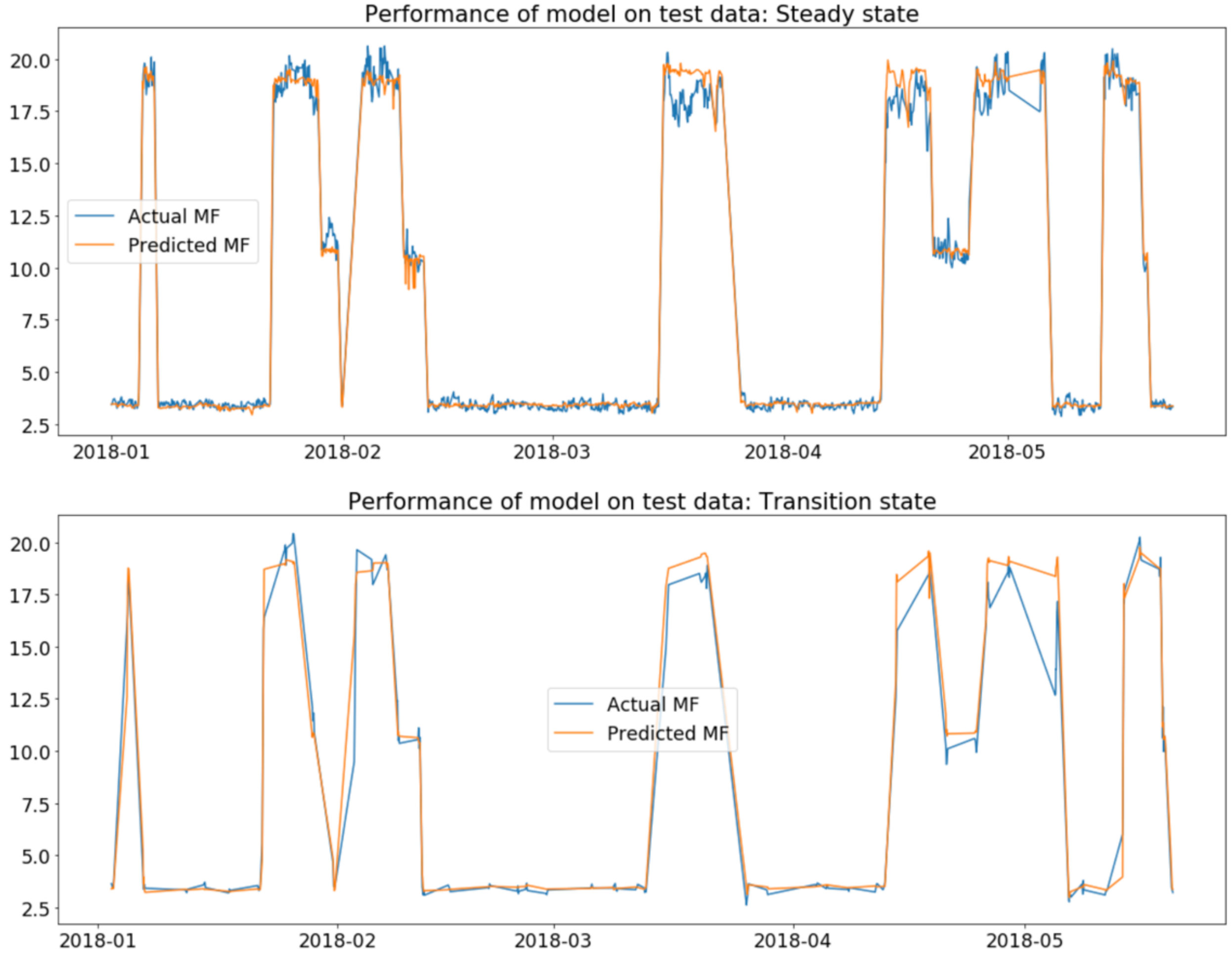
8.1M Pounds of product saved from scrap per production line / per year

$30M+ Annual economic benefit

Achieved 95.2% Overall Accuracy Relative to Lab Tests

Performance Across Models

Method	Overall Accuracy	Overall Error	Steady State Error	Transition Error
Linear Regression	84.5%	15.5%	15.6%	15.0%
Random Forest	94.7%	5.3%	4.8%	8.2%
XGBoost	**95.2%**	**4.8%**	**4.5%**	**6.8%**

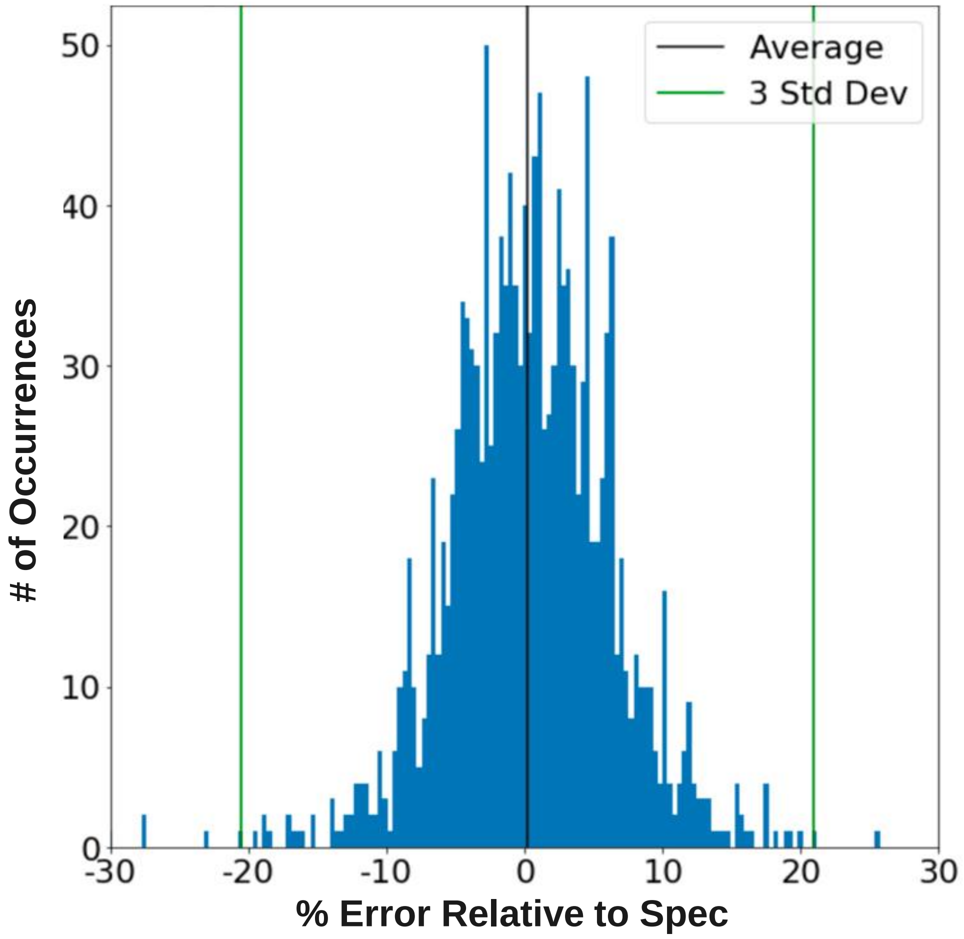


BakerHughesC3.ai

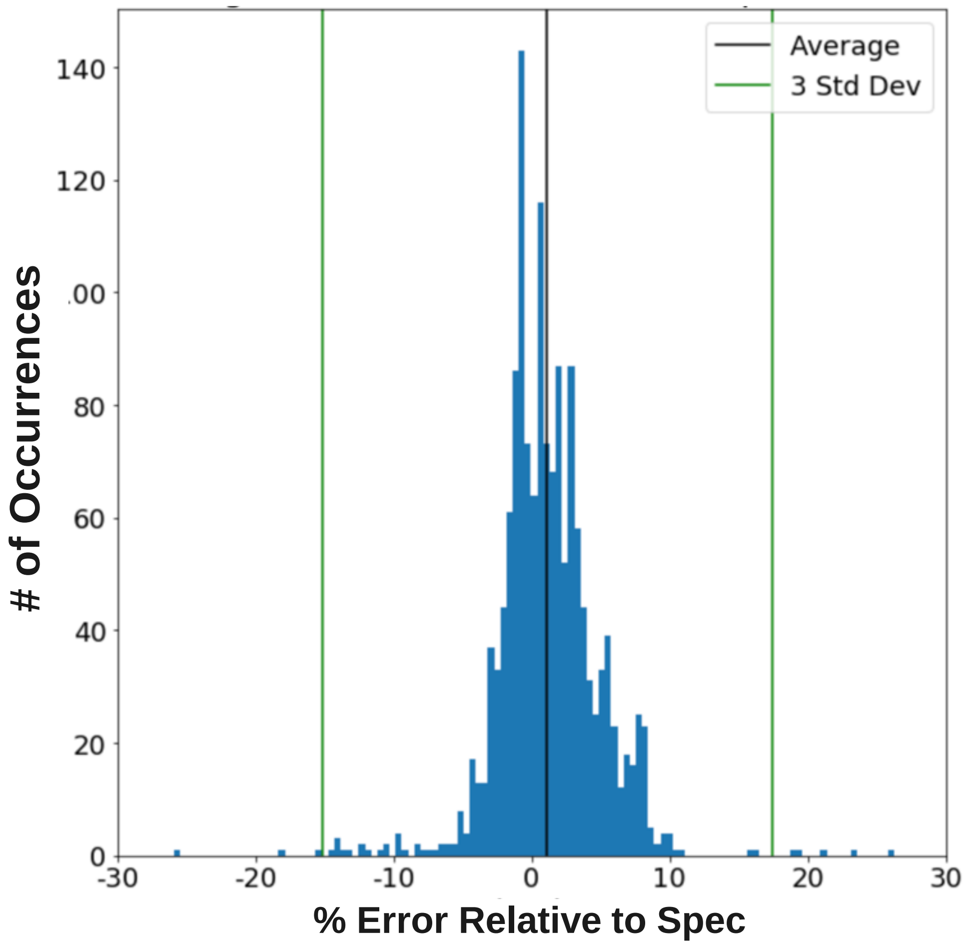
ML model was less error prone compared to lab tests

Distribution of Error Relative to OP Spec

Lab Tests



ML Model

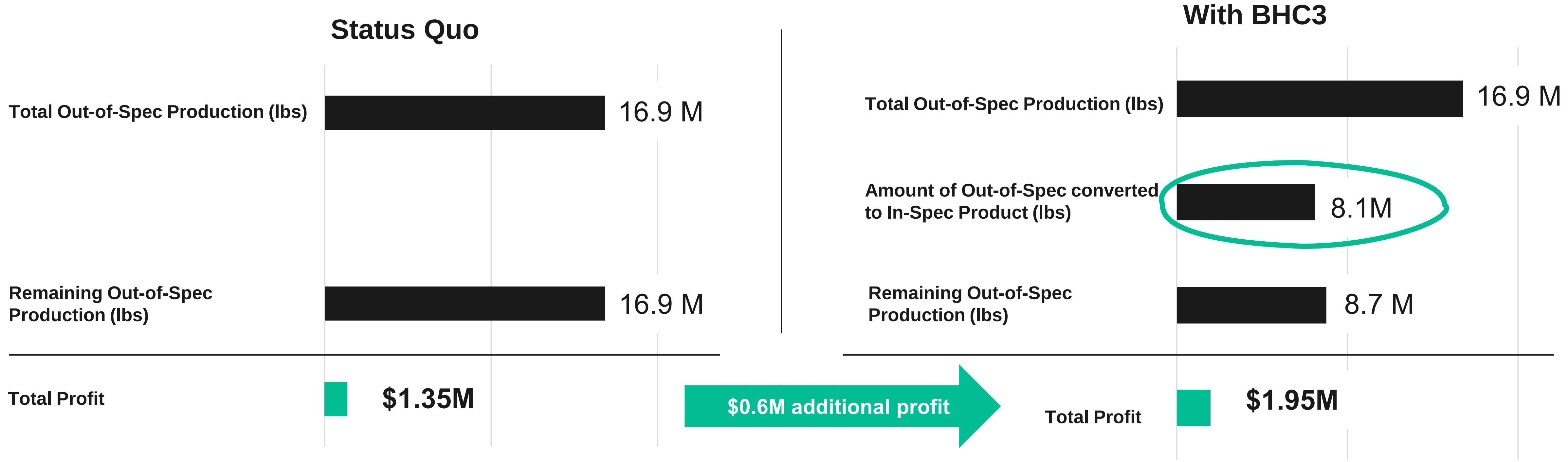


BakerHughesC3.ai

Over $30 million in Additional Operating Profit

- **8.1 million lbs. of product** per year, per line can be moved from off-spec to on-spec
- This represents **$607,500 of incremental profit per line**, or **$30.4 million per year** scaled across 50 lines

Total “L” Line Production Outcome

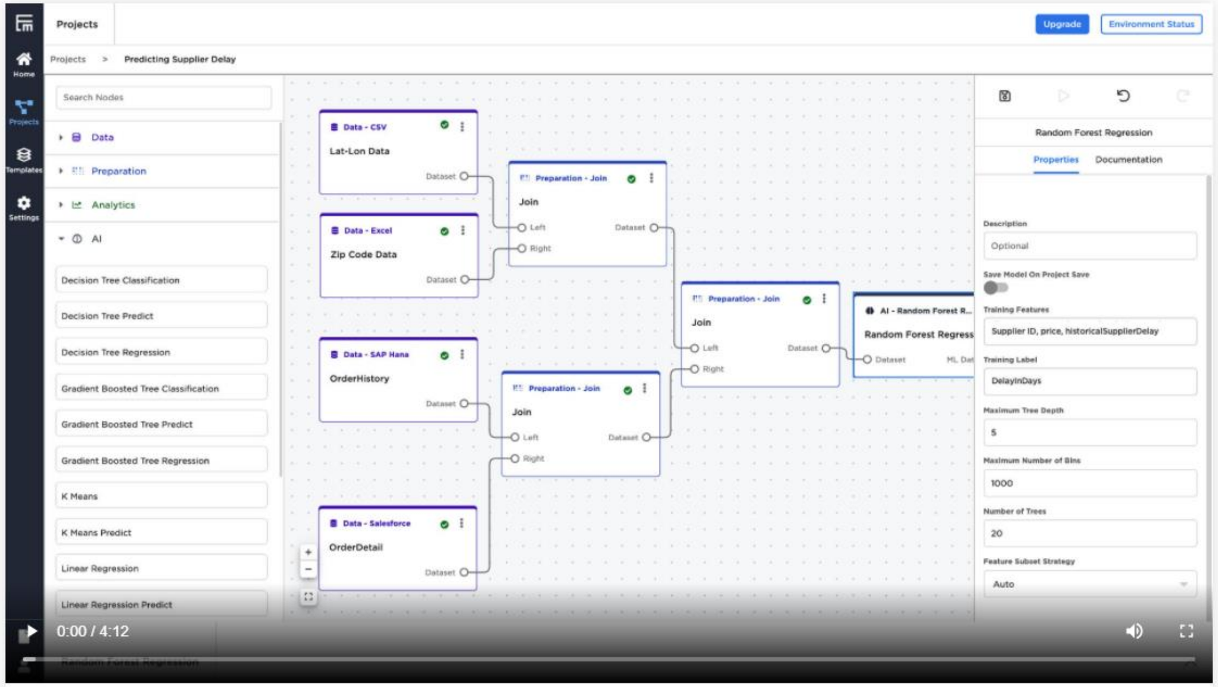


BakerHughesC3.ai

Demo – Ex Machina

Link to the demo:

https://bakerhughesc3.ai/products/bhc3-ai-ex-machina/



BakerHughesC3.ai

BHC3 Ex Machina

1. Connect

Quickly load data to BHC3 Ex Machina from numerous datastores (Databricks, CSV, Salesforce, etc.).

2. Prepare

Discover, cleanse, enrich, and validate your data, making it ready for analysis.

3. Visualize

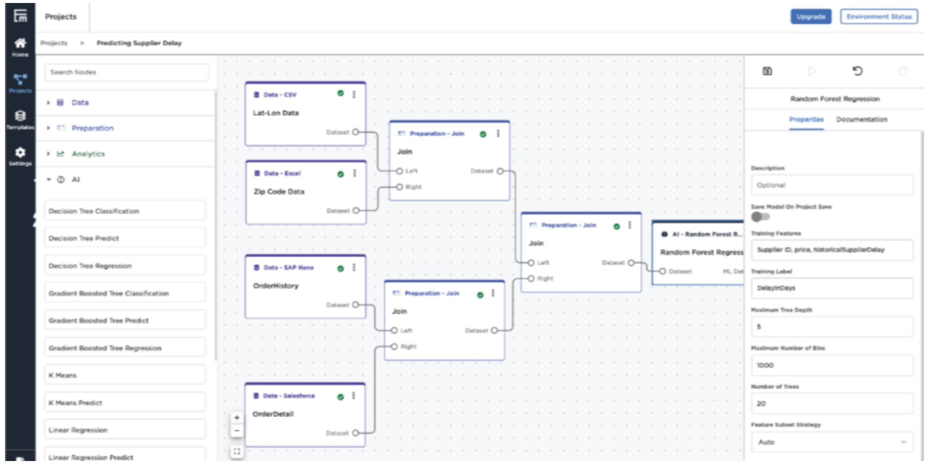
Visualize data at any step in your workflow to understand data formats, completeness, and accuracy.

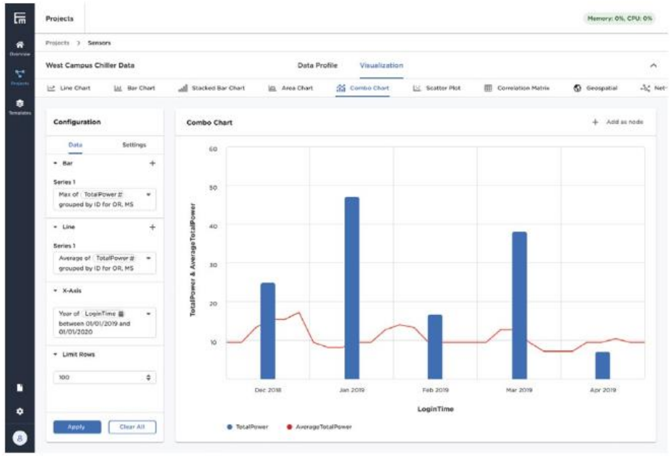
4. Analyze

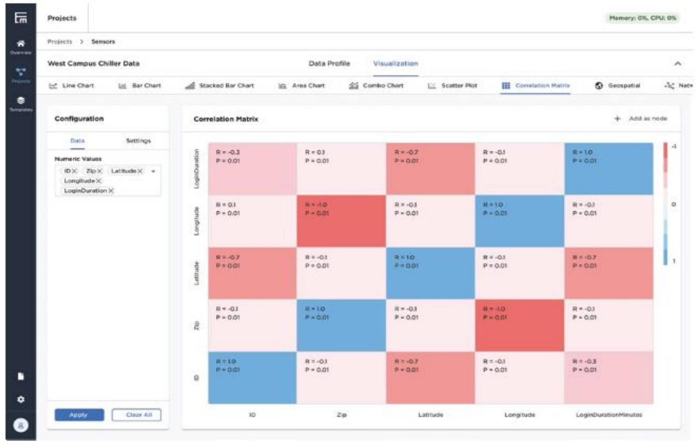
Build analytic pipelines; train machine learning algorithms to make predictions.

5. Operationalize

Share insights broadly using BHC3 Ex Machina’s cloud infrastructure; write results to production applications.

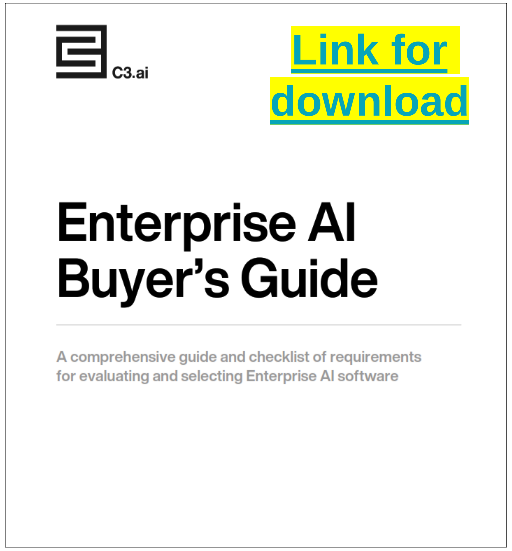


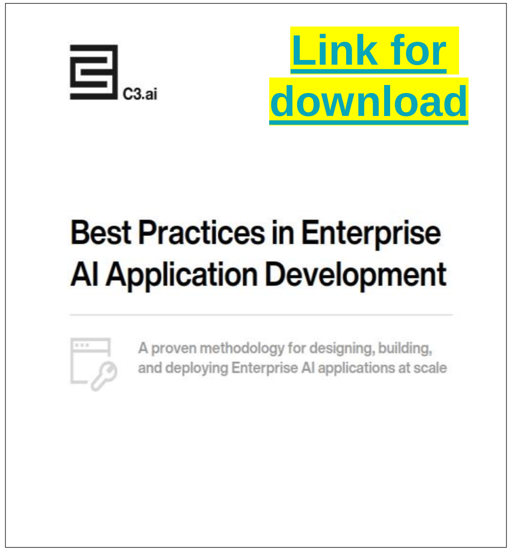


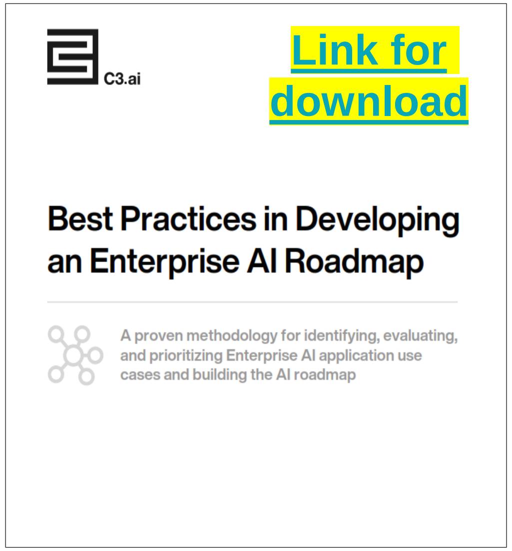


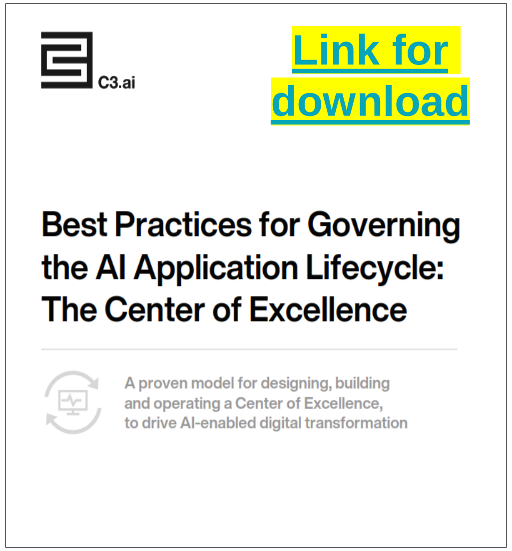
BakerHughesC3.ai

Best Practices in Enterprise AI and Center of Excellence





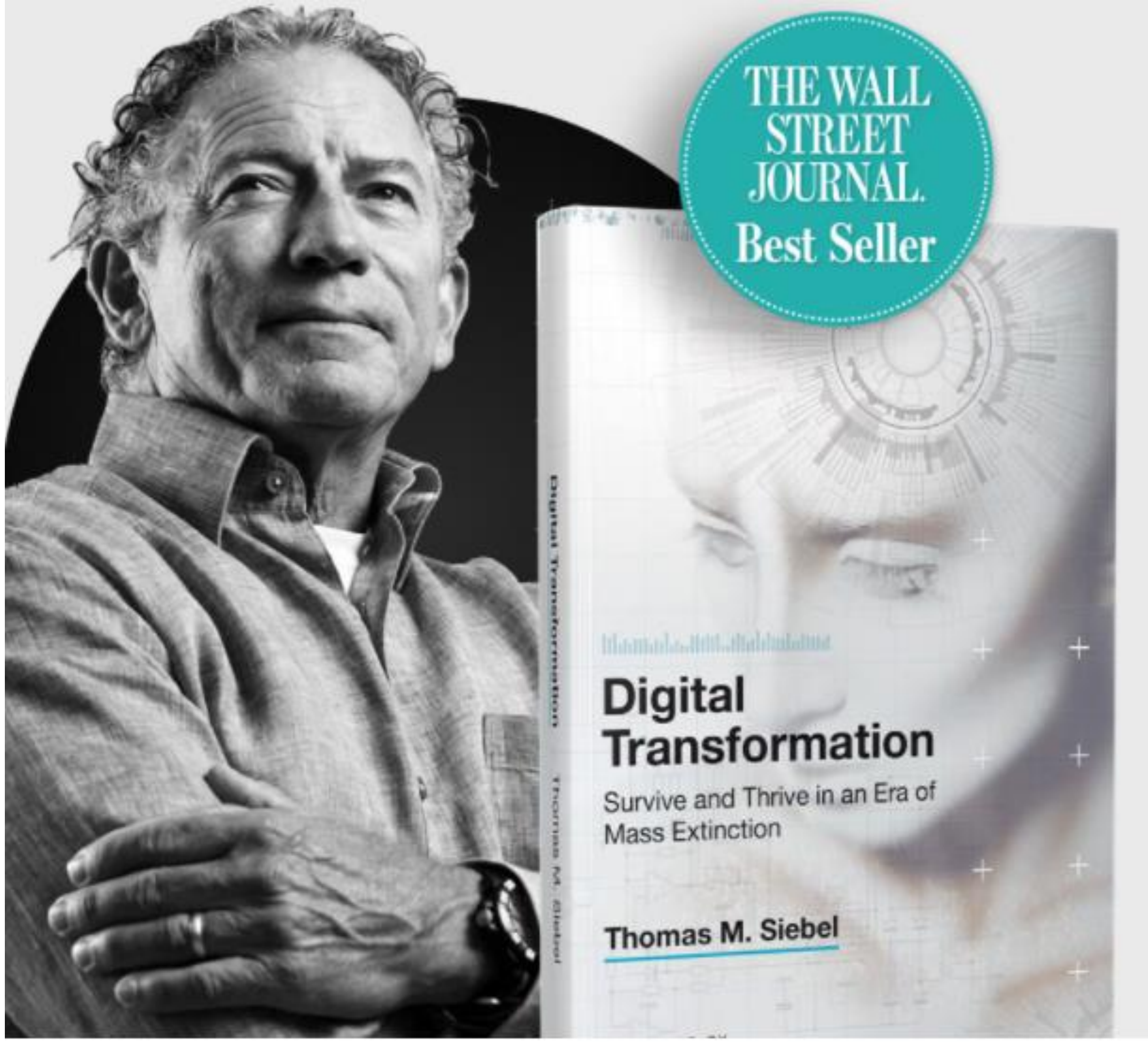




BakerHughesC3.ai

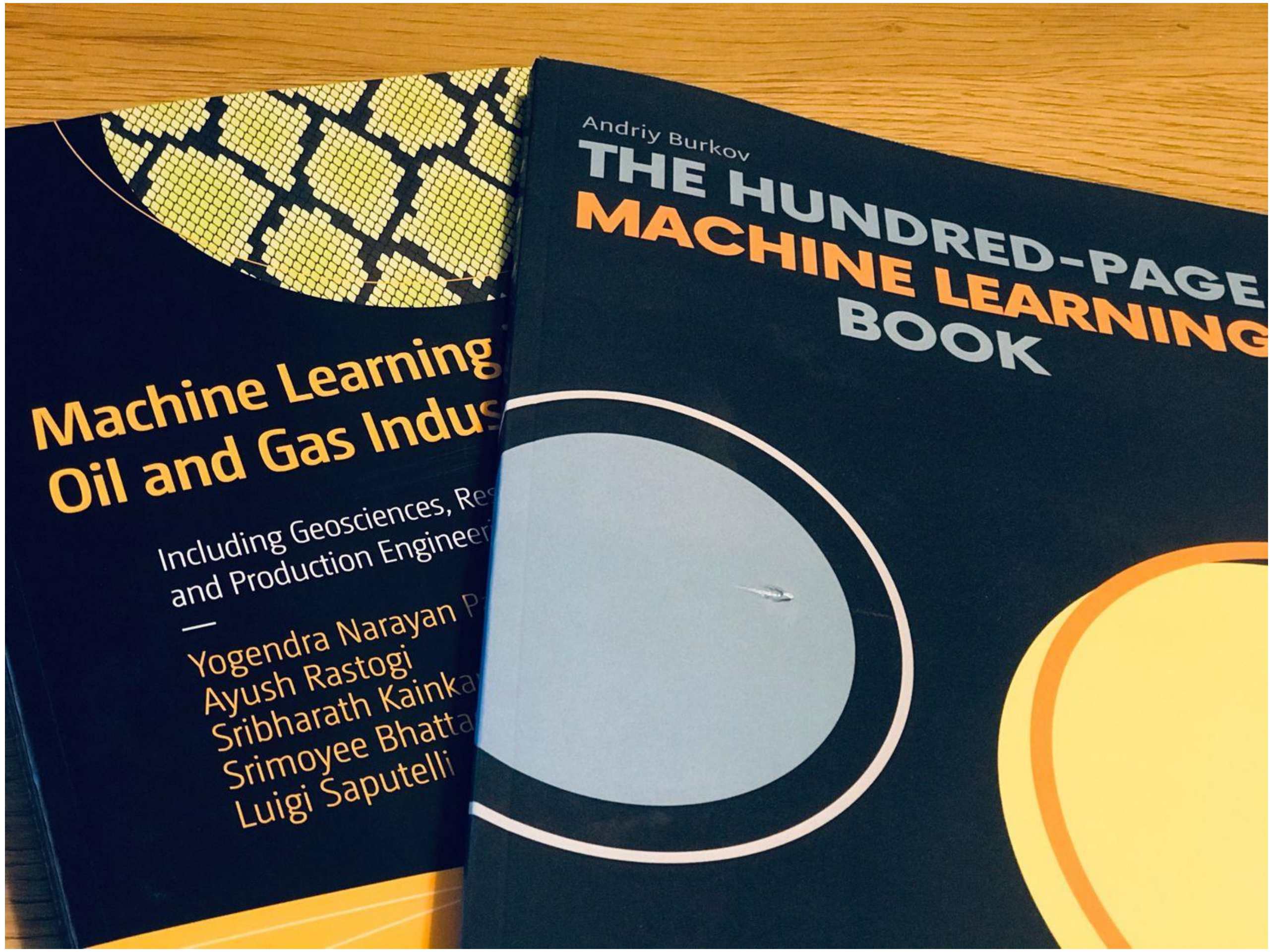
Reading Suggestions

[Link to the book]



BakerHughesC3.ai

Reading Suggestions



Next Session from BakerHughesC3.ai

Data Science & What We Do?

Riyad Muradov,

Data Scientist, C3.ai

BakerHughesC3.ai

BakerHughesC3.ai