AWS re:Invent

AIM201-S

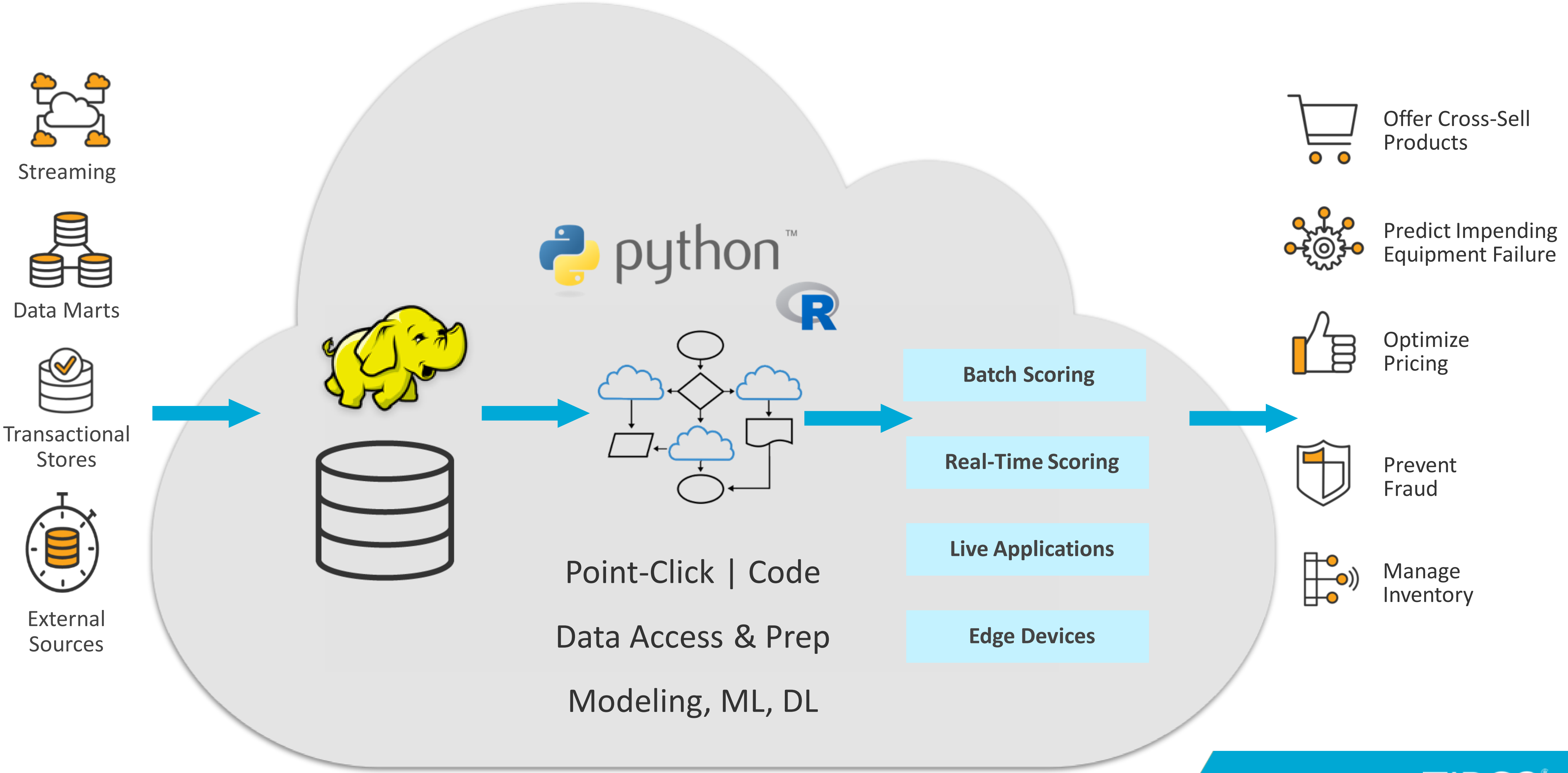
Hot paths to anomaly detection with TIBCO data science, streaming on AWS

Steven Hillion
Sr Director, Data Science
@StevenHillion

Michael O'Connell
Chief Analytics Officer
@MichOConnell

AWS re:Invent

The Ideal Data Science Platform

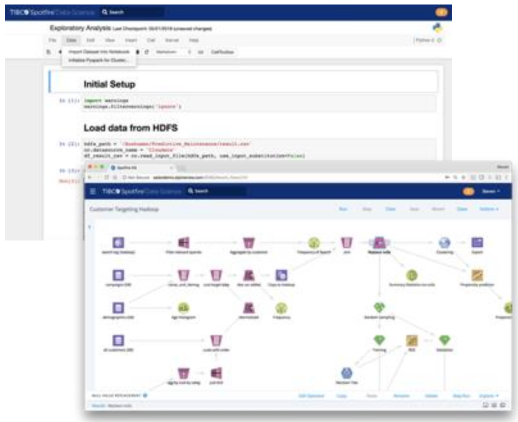


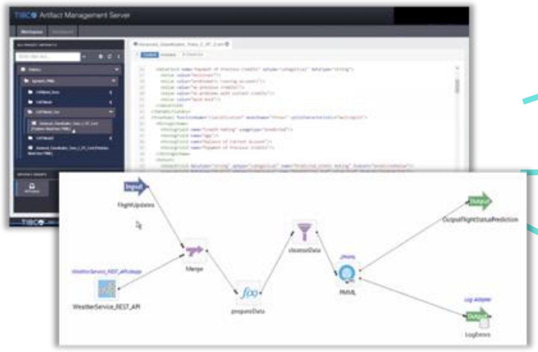
Agenda

- TIBCO Data Science and AWS Marketplace
- The TIBCO Connected Intelligence Cloud
- Anomaly Detection and Analysis
- ***Demonstration – Spatial Anomaly Analysis***
- Links and Assets

TIBCO Data Science

	Data Access/Prep	Modeling	Operations	Business Apps
FUNCTION	+ Distributed compute + Feature engineering + Reusable templates	+ Visual composition + Multilingual notebook + Native ML & OS + Auto-ML, data prep	+ Model lifecycle management + Batch automation + Real-time event processing + REST, applications, embedding	+ Engineering/IoT + Customer analytics + Risk management + Supply chain





Medic; e.g., researcher on epidemic monitoring

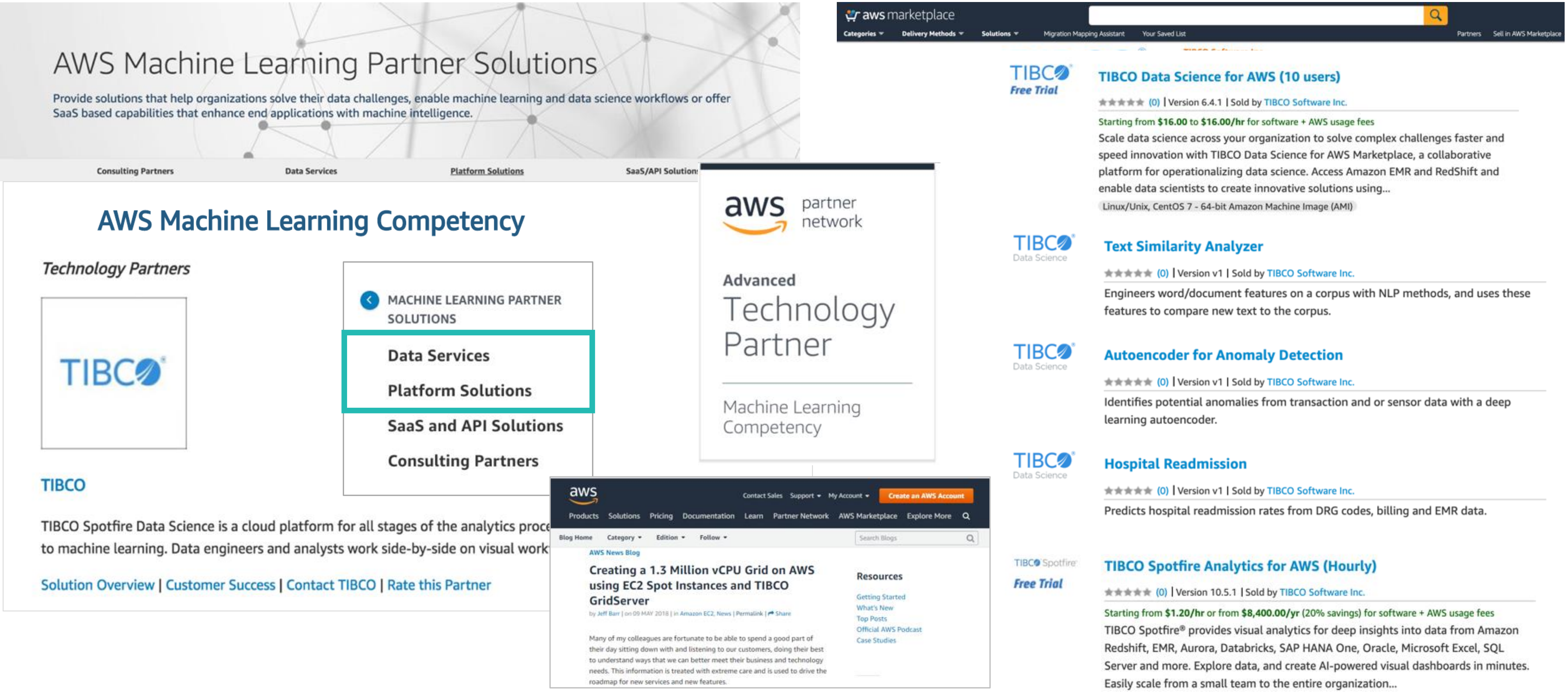
Engineer; e.g., aerodynamics engineer

Marketeer; e.g., customer engagement analyst

USER or AUTOMATION	*Data Scientist* *Citizen Data Scientist*	*Data Scientist* *Citizen Data Scientist*	*Analytics Operations* *IT / Software Engineer*	*Business User* *Analytics Operations* *IT / Administration*

TIBCO Data Science on AWS

TIBCO DS on AWS Marketplace; Biggest vCPU Grid; Lightest Serverless Footprint



Data Science in the Cloud: Leidos Healthcare Analytics

Leidos Collaborative Advanced Analytics & Data Sharing Platform (CAADS) uses TIBCO Data Science and AWS to deliver analytics services in Healthcare

CDC

Disease Outbreaks: Determining the cause of an HIV outbreak in the Midwest

NIH

Disease Outbreaks: Run simulations of disease propagation to guide public policy, specifically around the Zika virus

CMS

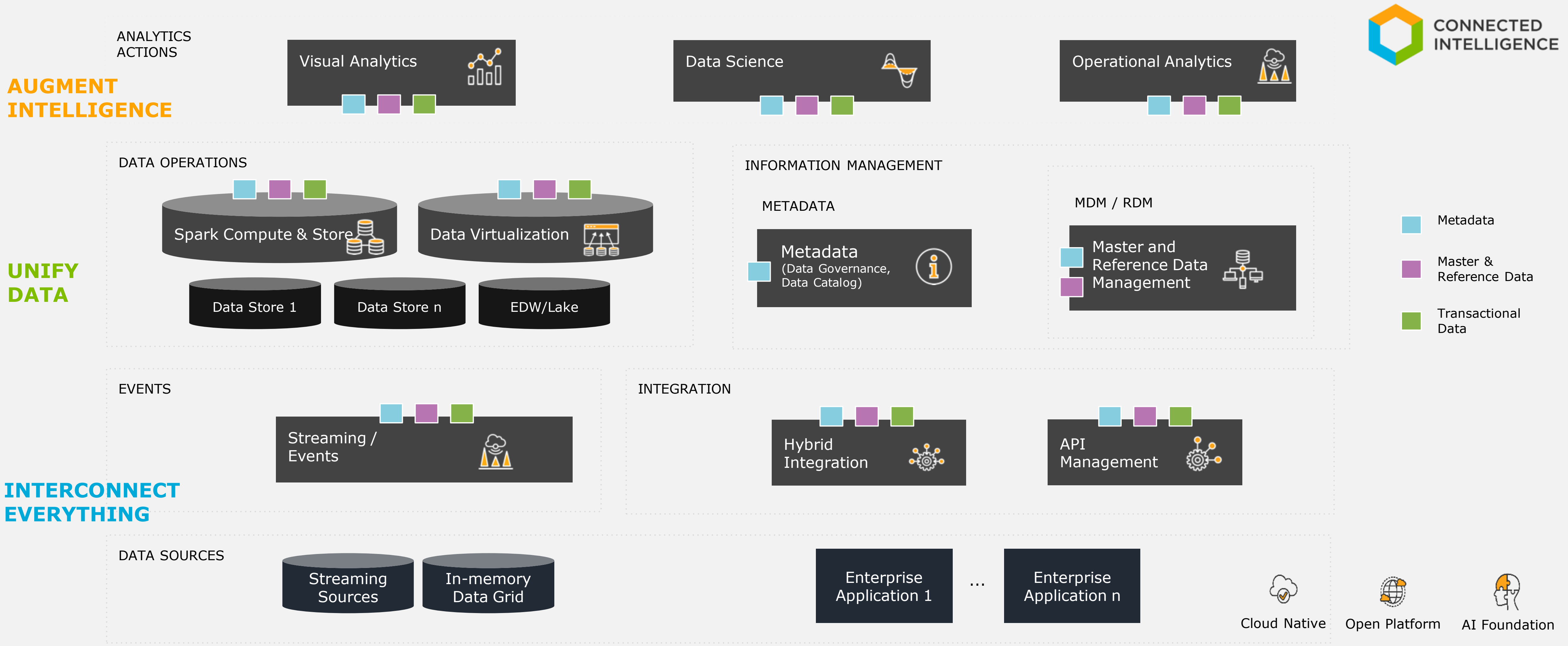
Data Governance: Analyzing and consolidating data around emerging Healthcare policies across 56 regions in the United States

NASA

Space Exploration: Analyzing human factors that affect the ability to transport astronauts on long fights (e.g., to Mars)

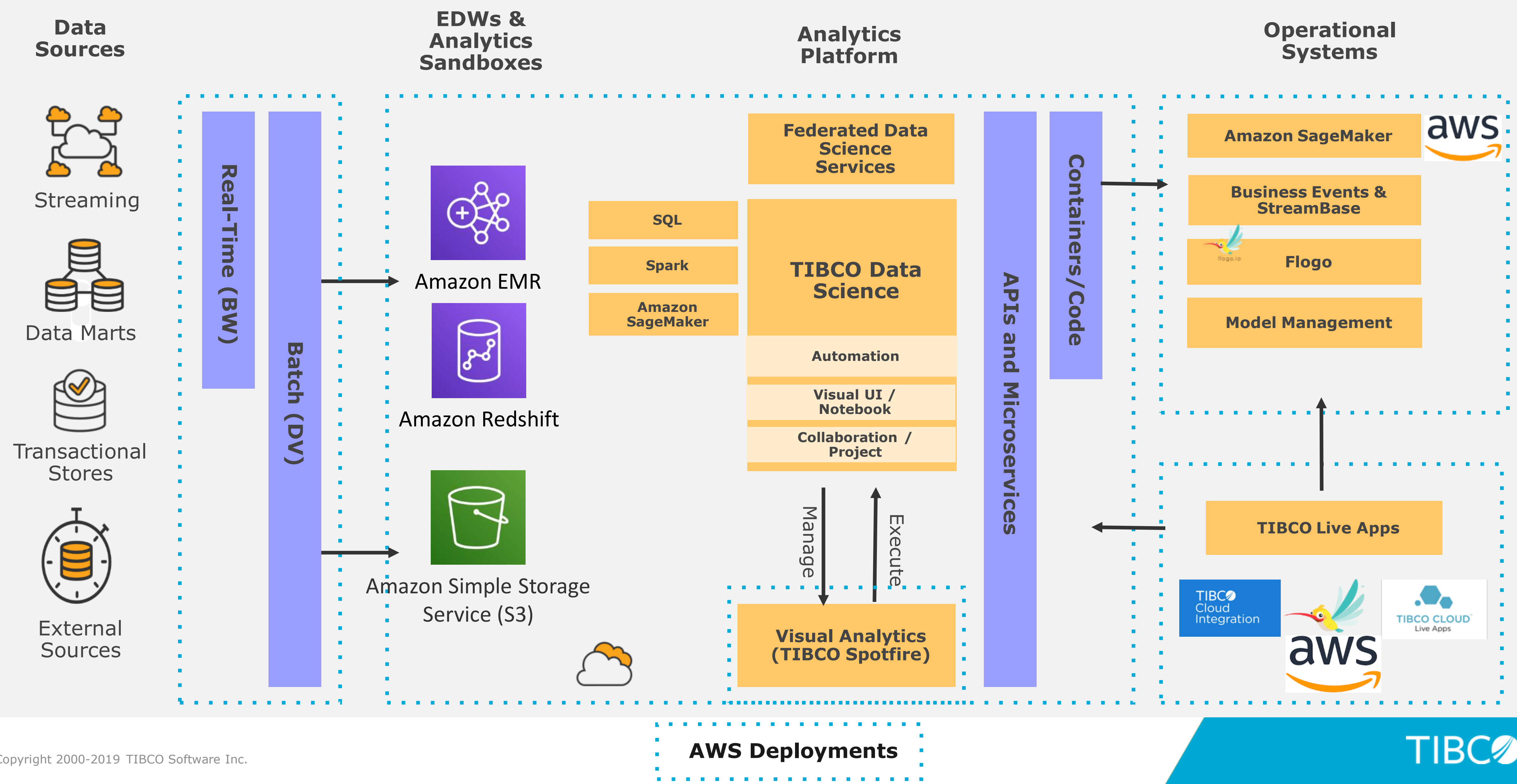
TIBCO analytics transformation platform

Powered by shared data assets



TIBCO

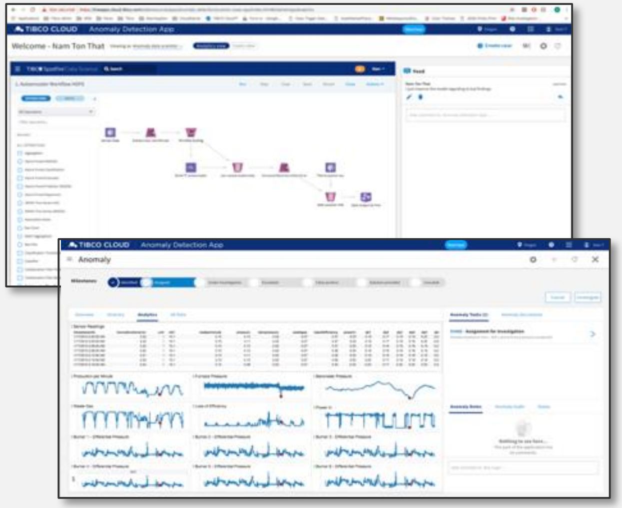
TIBCO Data Science and AWS



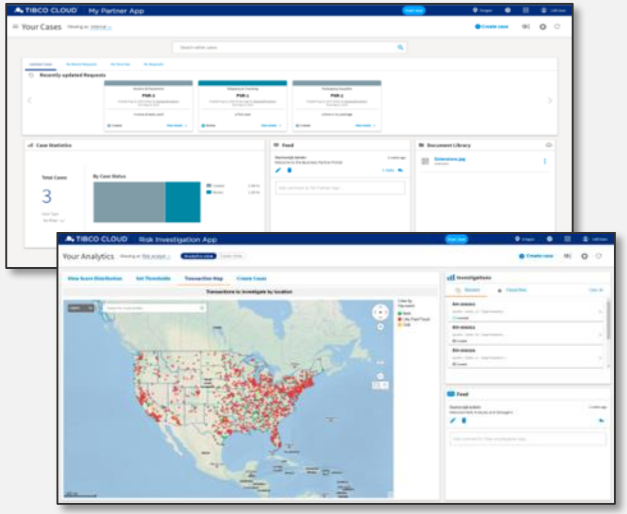
TIBCO Data Science Solutions on AWS

Cloud Apps: *Visual Analytics, Data Science, Streaming, Case Management*

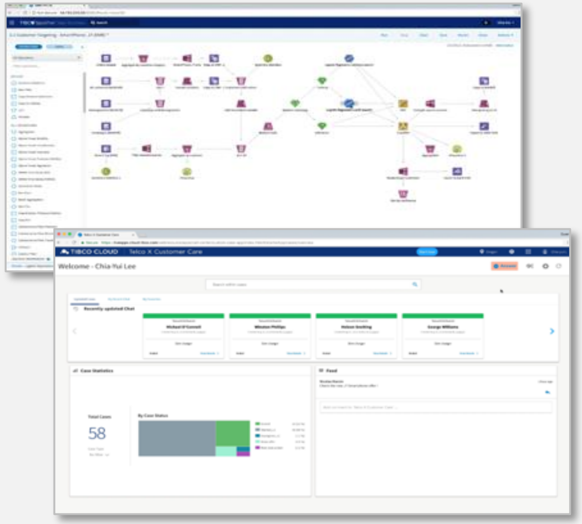
Anomaly Detection



Risk Management



Customer Engagement



Starter Set

- Process Mining
- IoT Analytics
- Anomaly Detection
- Risk Management
- Customer Engagement
- Blockchain – Dovetail
- Partner Management
- Starter Toolkit

Review Status: *TIBCO Spotfire*
Identify issues, sweet spots

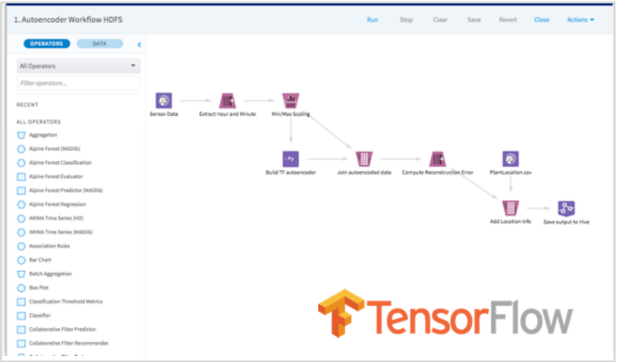
Model: *TIBCO Data Science*
Supervised: Train
Unsupervised: Anomalies

Analyze Event Stream:
TIBCO Flogo, Cloud Integration
Batch and Real-Time Updates

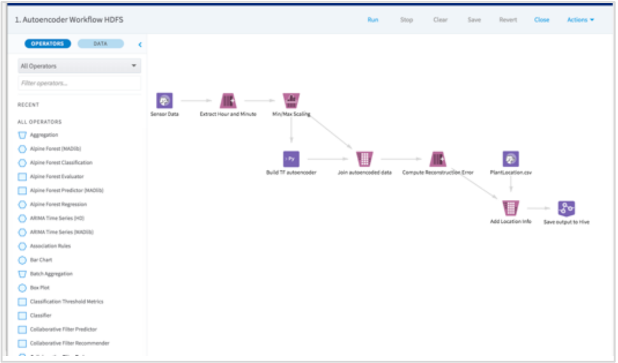
Case Manage: *TIBCO Live Apps*
Investigate identified cases
Audit trail + recycle

Anomaly Analysis Solution Overview

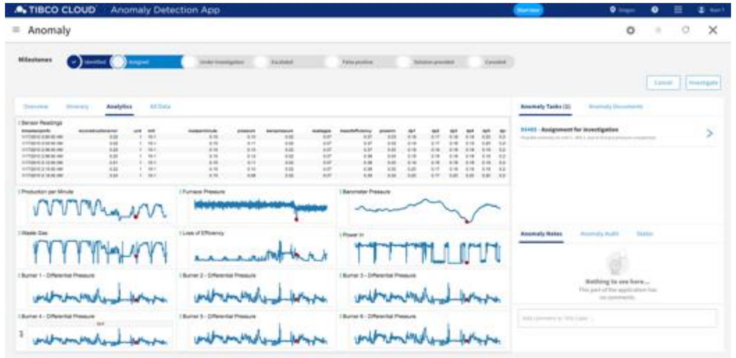
1 Collect data from equipment, normalize, model to predict magnitude of anomaly – **TIBCO Data Science & AWS**



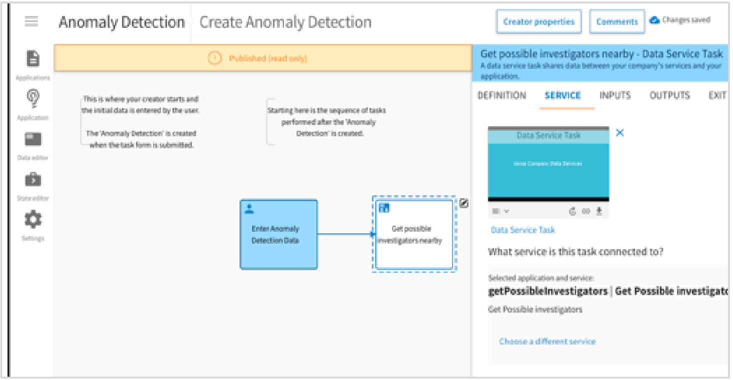
2 Model detects anomaly – **TIBCO Data Science**



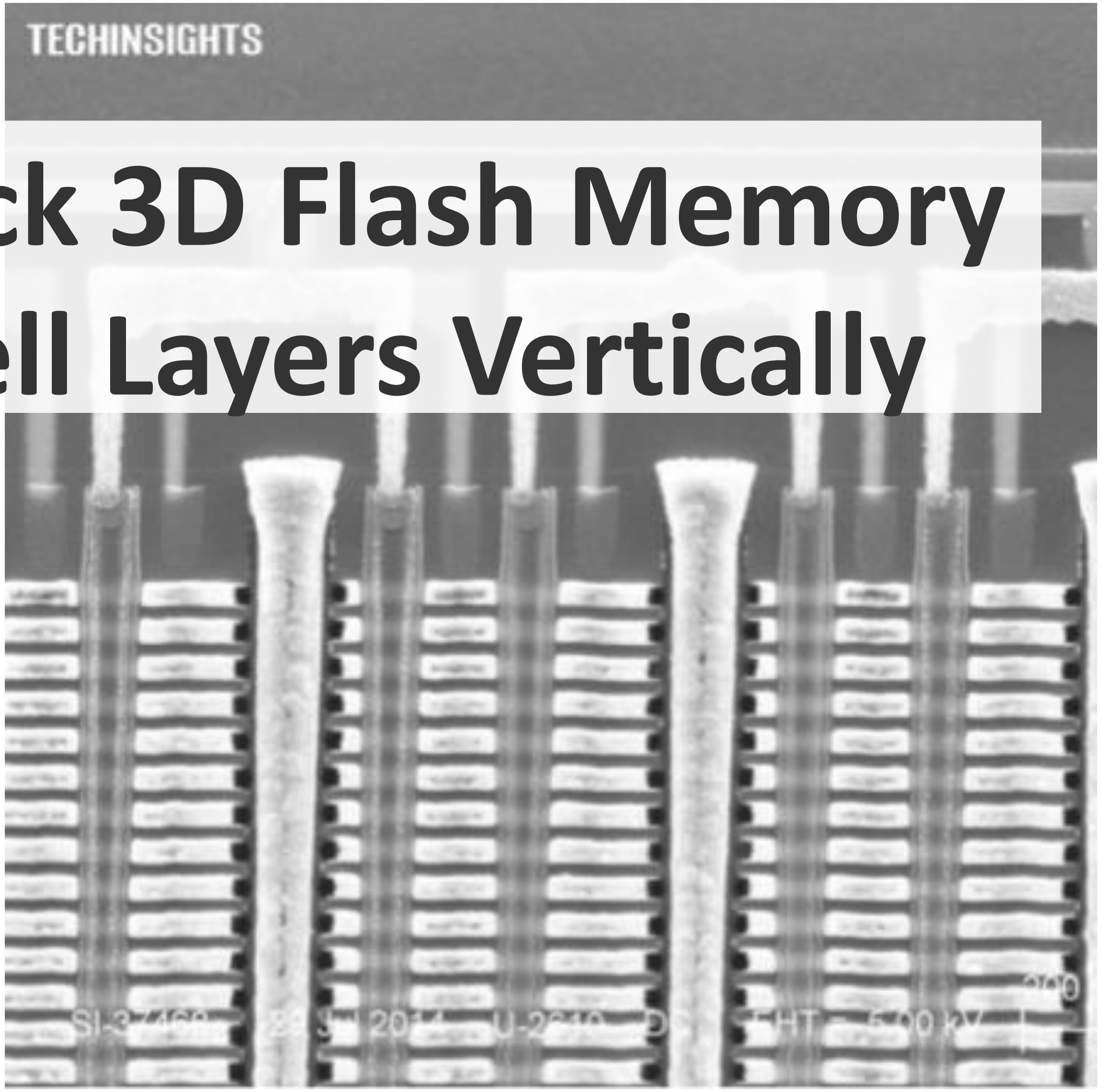
3 Alert raised and case created – **TIBCO Cloud Live Apps**



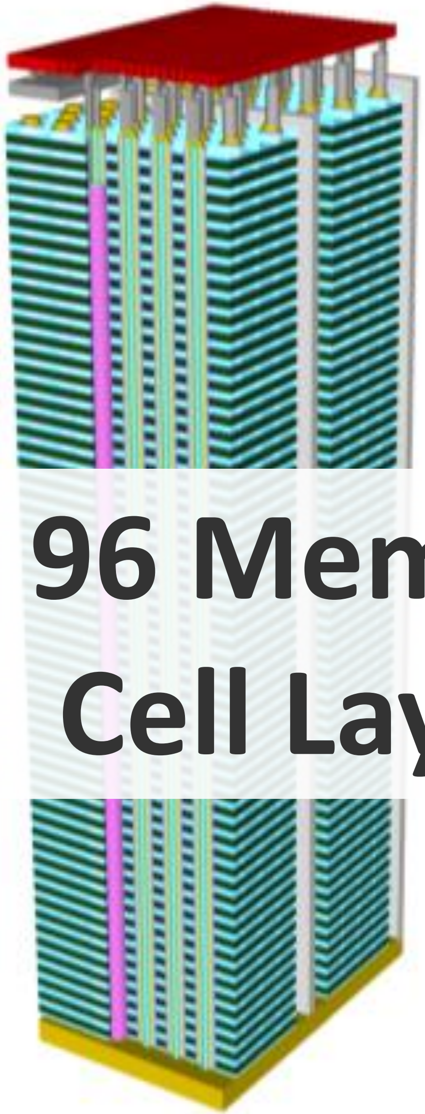
4 Case manager investigates and takes action to the equipment – **TIBCO Cloud Integration**



Data Challenges in High-Tech Manufacturing



Stack 3D Flash Memory Cell Layers Vertically

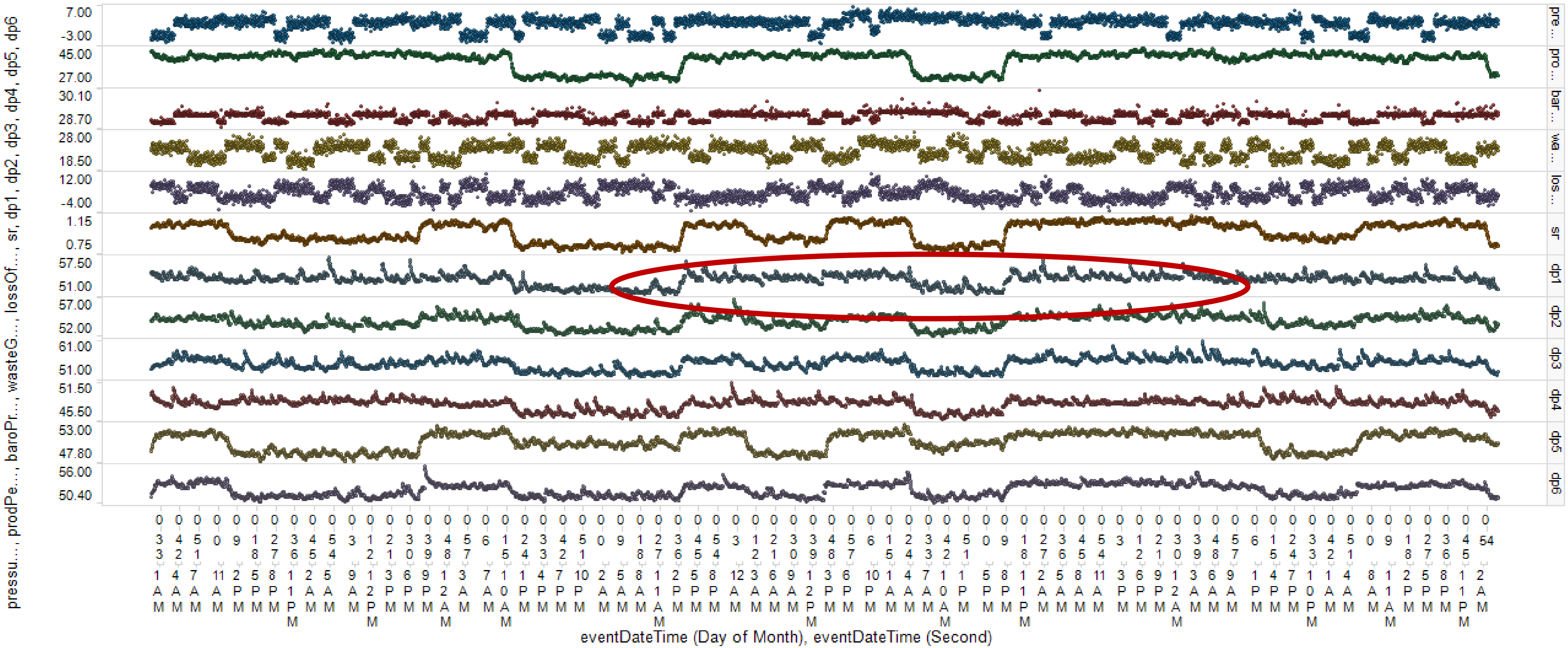


96 Memory Cell Layers

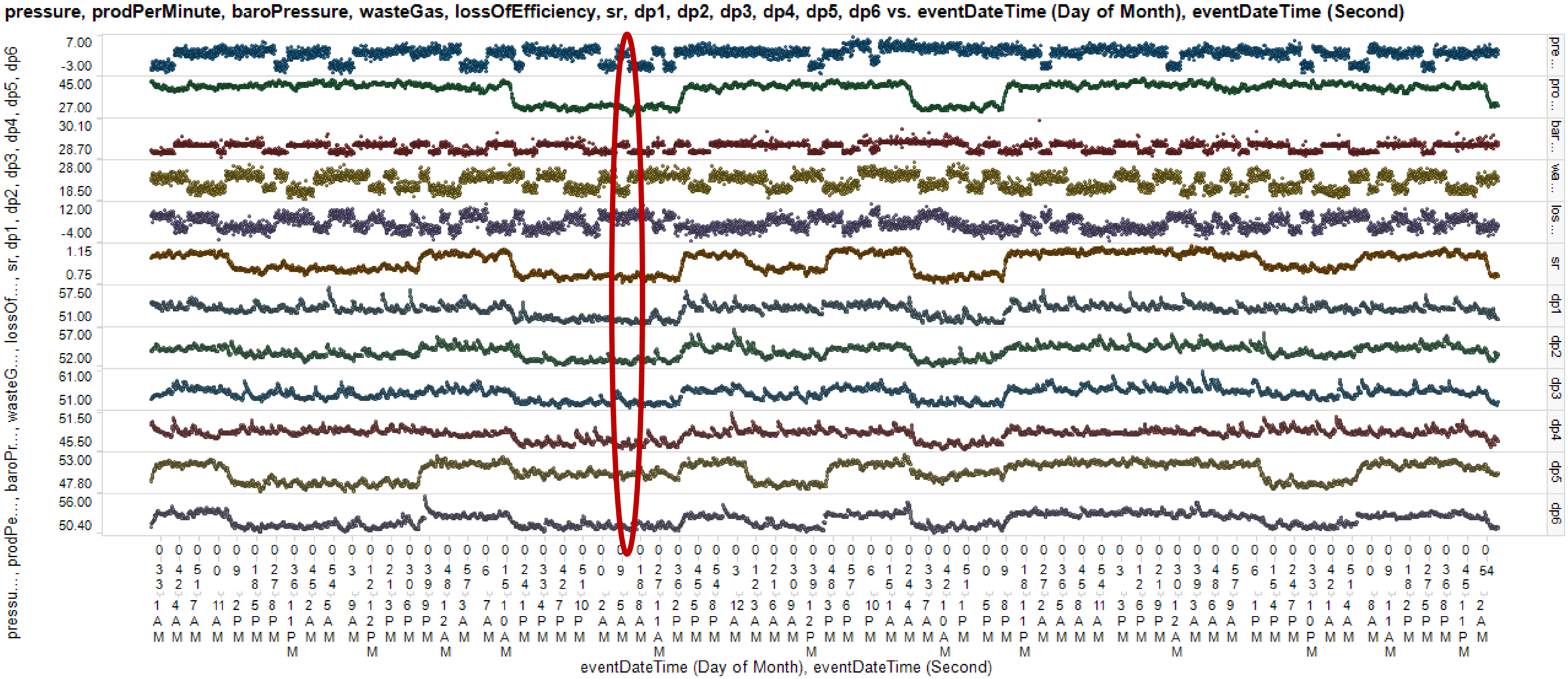
Hot Paths to Anomaly Detection

Longitudinal Anomaly Analysis

pressure, prodPerMinute, baroPressure, wasteGas, lossOfEfficiency, sr, dp1, dp2, dp3, dp4, dp5, dp6 vs. eventDateTime (Day of Month), eventDateTime (Second)

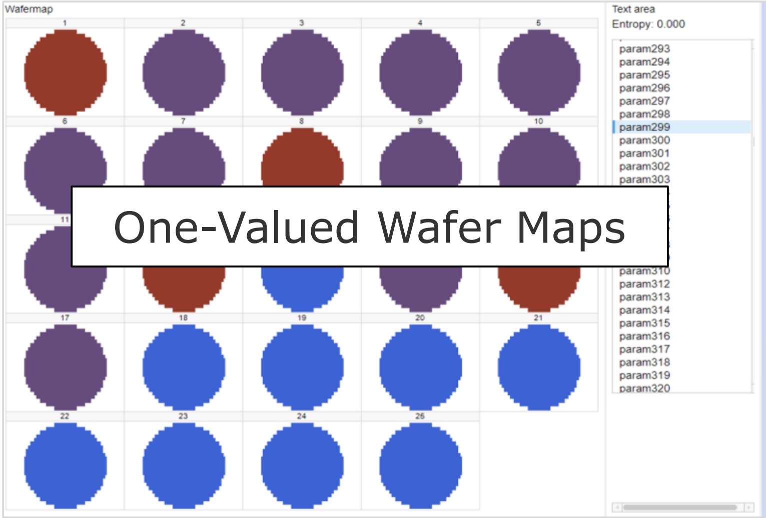


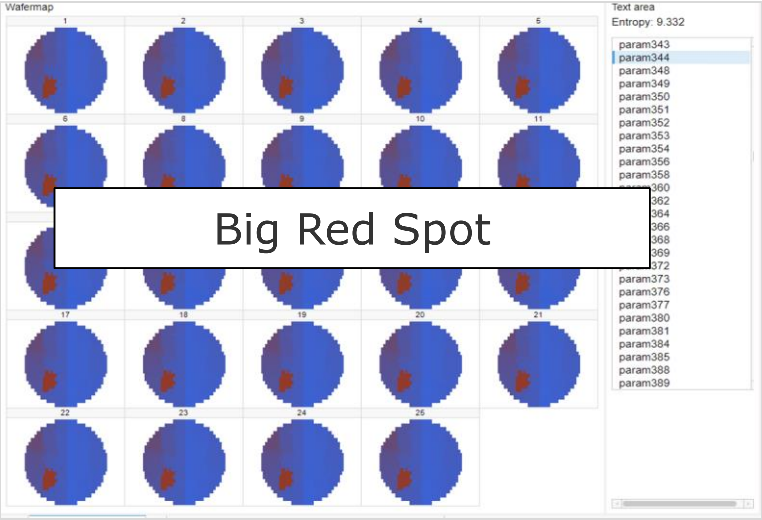
Cross-Sectional Anomaly Analysis

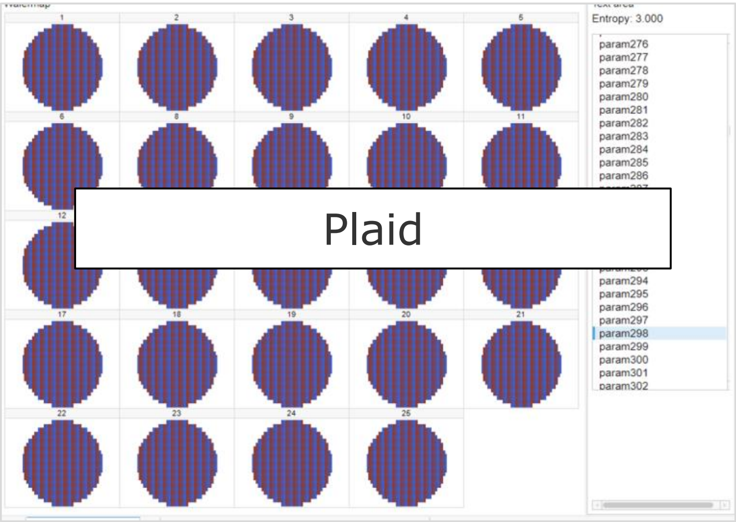


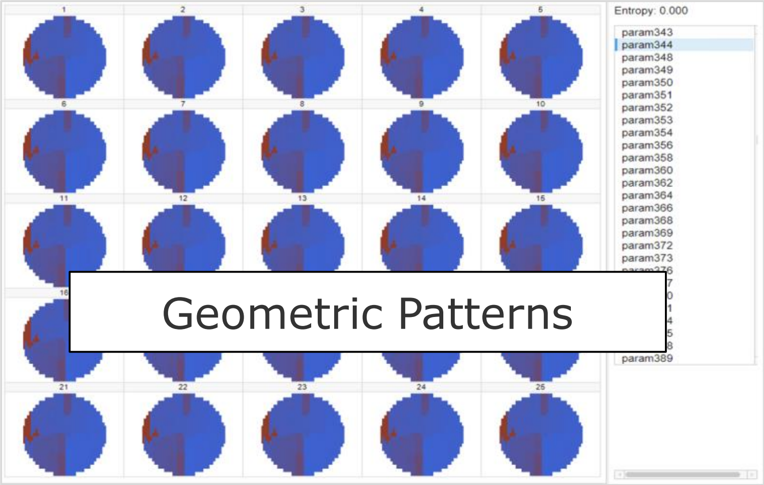
TIBCO

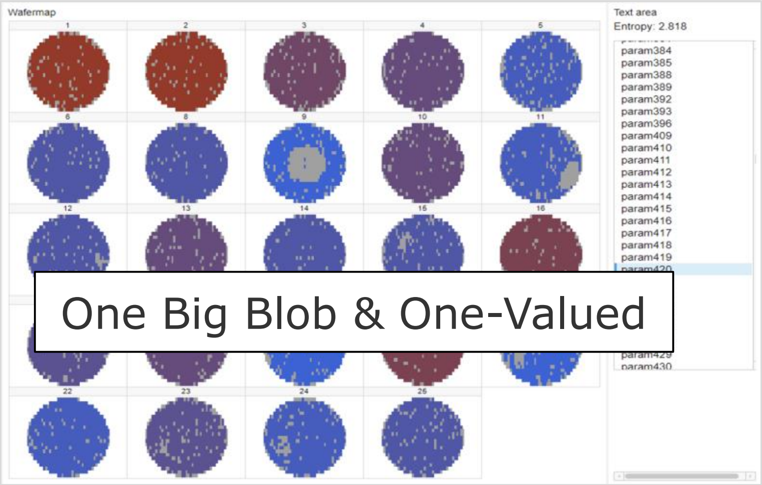
Spatial Anomaly Analysis

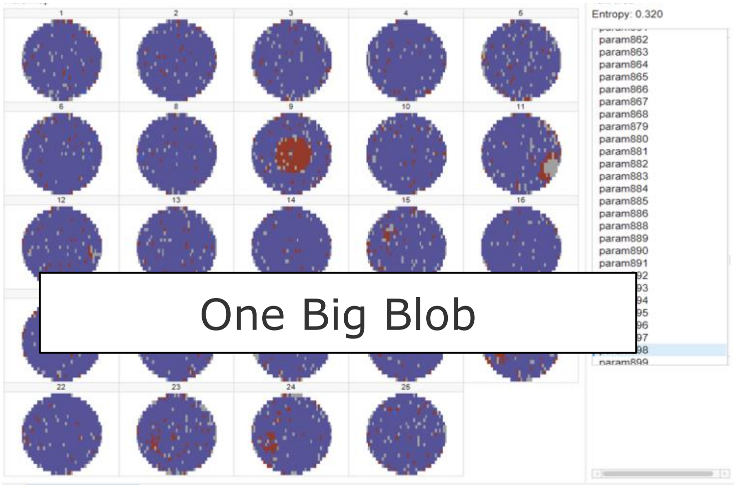






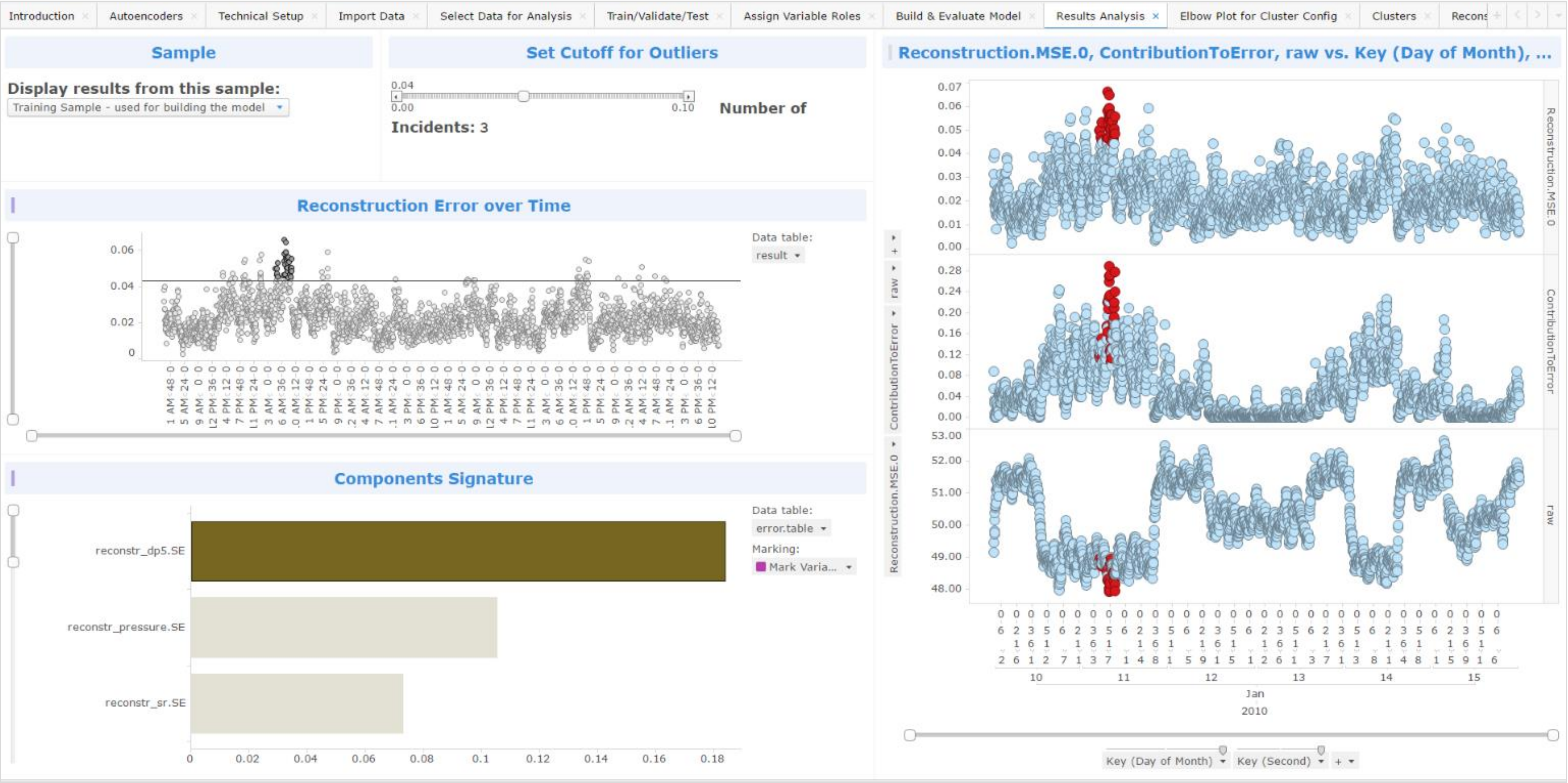




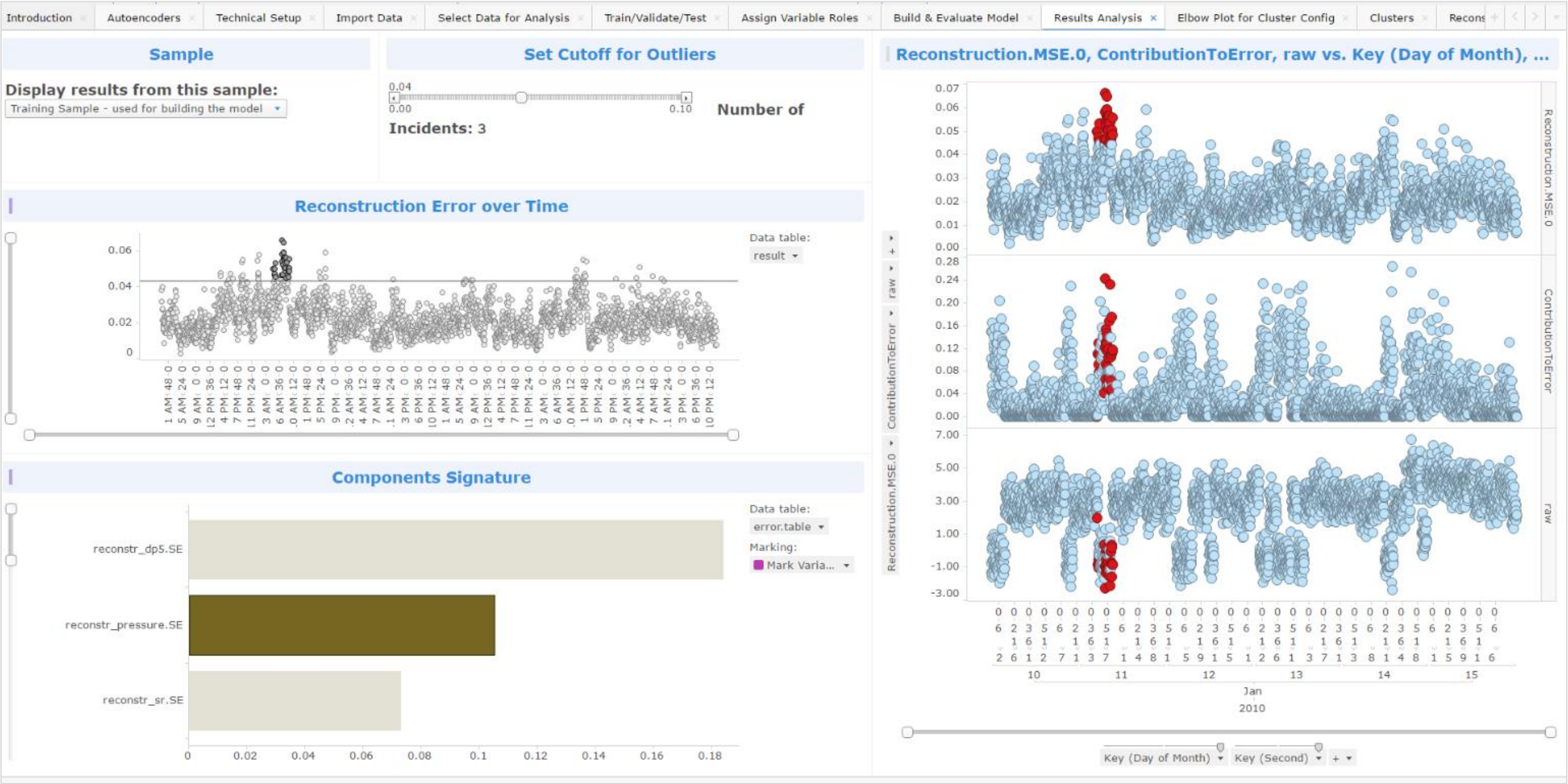


TIBCO®

Cross-Sectional Anomaly Detection



Cross-Sectional Anomaly Detection



Analysis: Cluster Incidents, View Signatures

TIBCO

Analytic workflow – methodologies + demos

Find and cluster anomalous events [Demo #1]

- Transform wafer maps into vectorized coefficients, then cluster on quality
  - Many measured parameters, e.g., quality tests: storage fidelity, logic circuits
- Approach 1: SVD + K-means
  - Focus on failure mode parameters
- Approach 2: Bessel functions + hierarchical clustering
  - Radial basis functions
  - Rotationally invariant | Null-value tolerant | Efficient storage
  - Better than SVD + K-means for multi-parameter analysis

Monitor anomalies [Demo #2]

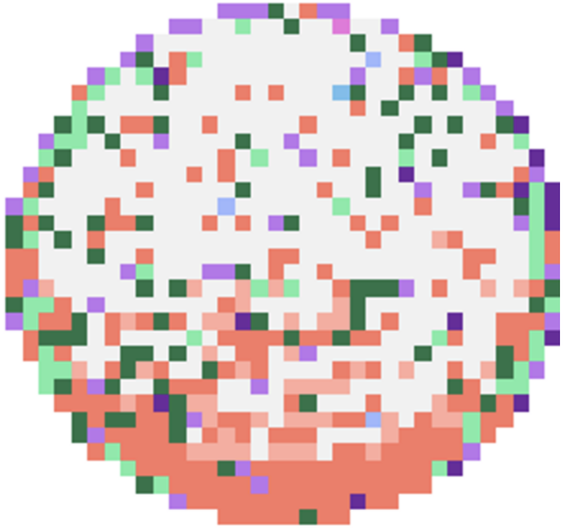
- Stream wafer data
- Vectorize and cluster

Predict when and why anomalies occur [Demo #3]

- Reduce dimensionality of very wide data
- Train models to determine sensor importance
- Identify responsible process parameters

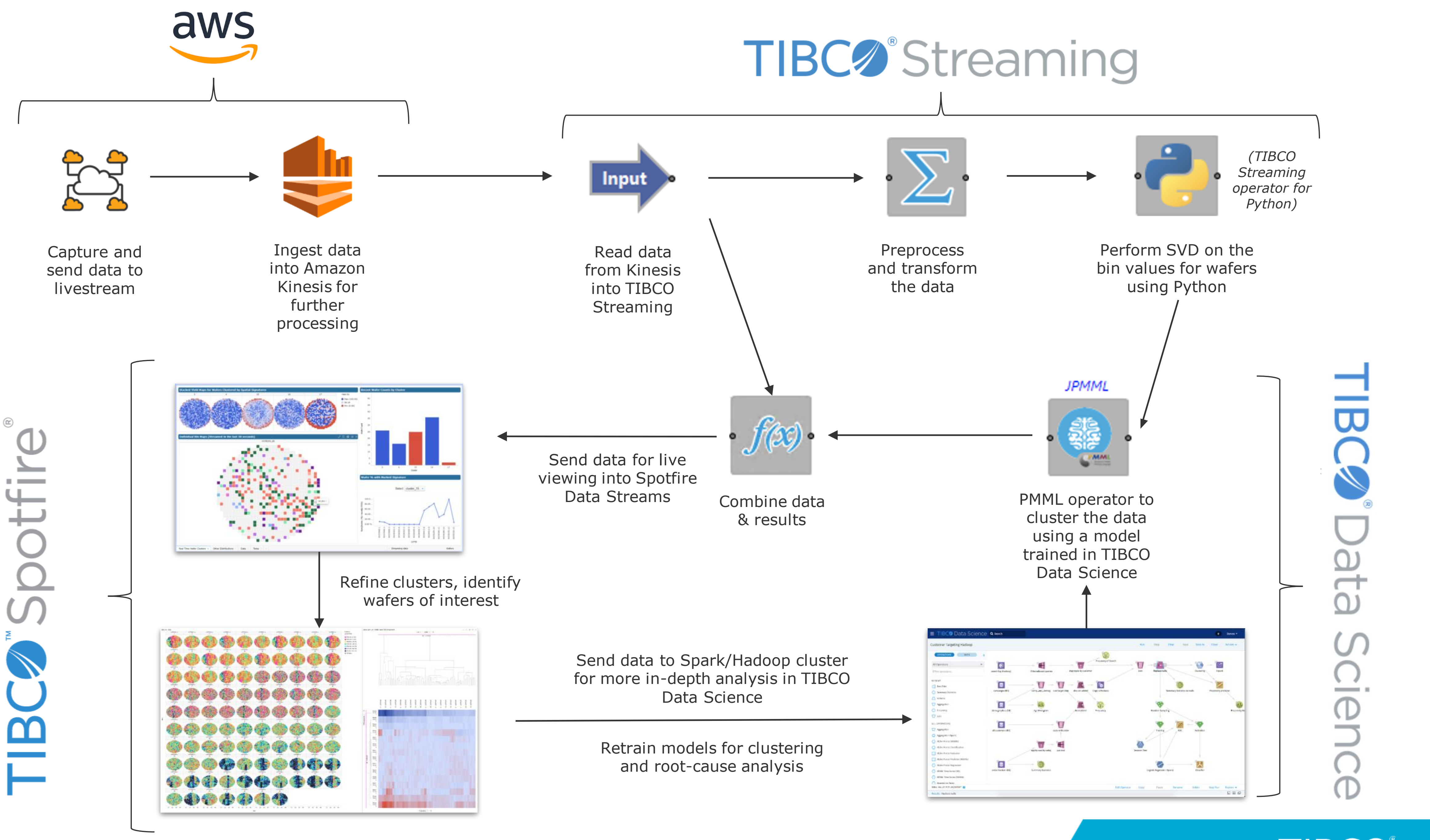
Process variable corrections/models rebasing

- *Identify new patterns as they emerge (e.g., incident analysis)*
- *Factory monitoring staff can click to characterize the new pattern*



Demo data

- 733 dies
- 20 x 40 layout
- 1,500 measured parameters
- Color by die test (white = pass all, colors = fail a test)


aws
TIBCO Streaming
Input
(TIBCO Streaming operator for Python)
Capture and send data to livestream
Ingest data into Amazon Kinesis for further processing
Read data from Kinesis into TIBCO Streaming
Preprocess and transform the data
Perform SVD on the bin values for wafers using Python
JPMML
PMML operator to cluster the data using a model trained in TIBCO Data Science
Combine data & results
Send data for live viewing into Spotfire Data Streams
TIBCO Spotfire
Refine clusters, identify wafers of interest
Send data to Spark/Hadoop cluster for more in-depth analysis in TIBCO Data Science
Retrain models for clustering and root-cause analysis
TIBCO Data Science
TIBCO

Anomaly Detection with Spatial Signature Analysis

TIBCO®

Anomaly Analysis

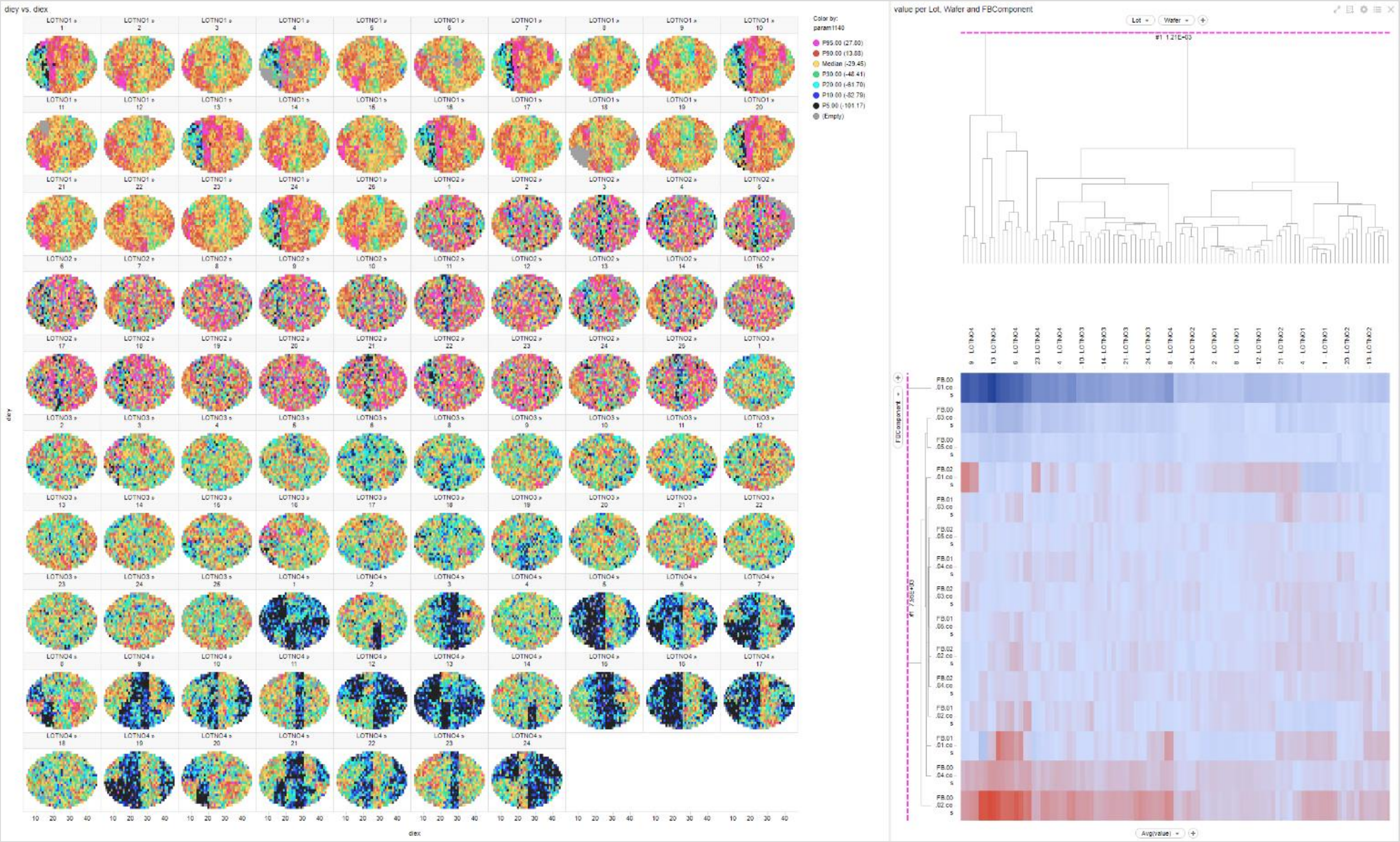
- Find and explore anomalous events
- Monitor anomalies
- Predict when and why they occur

Anomaly Detection and Analysis

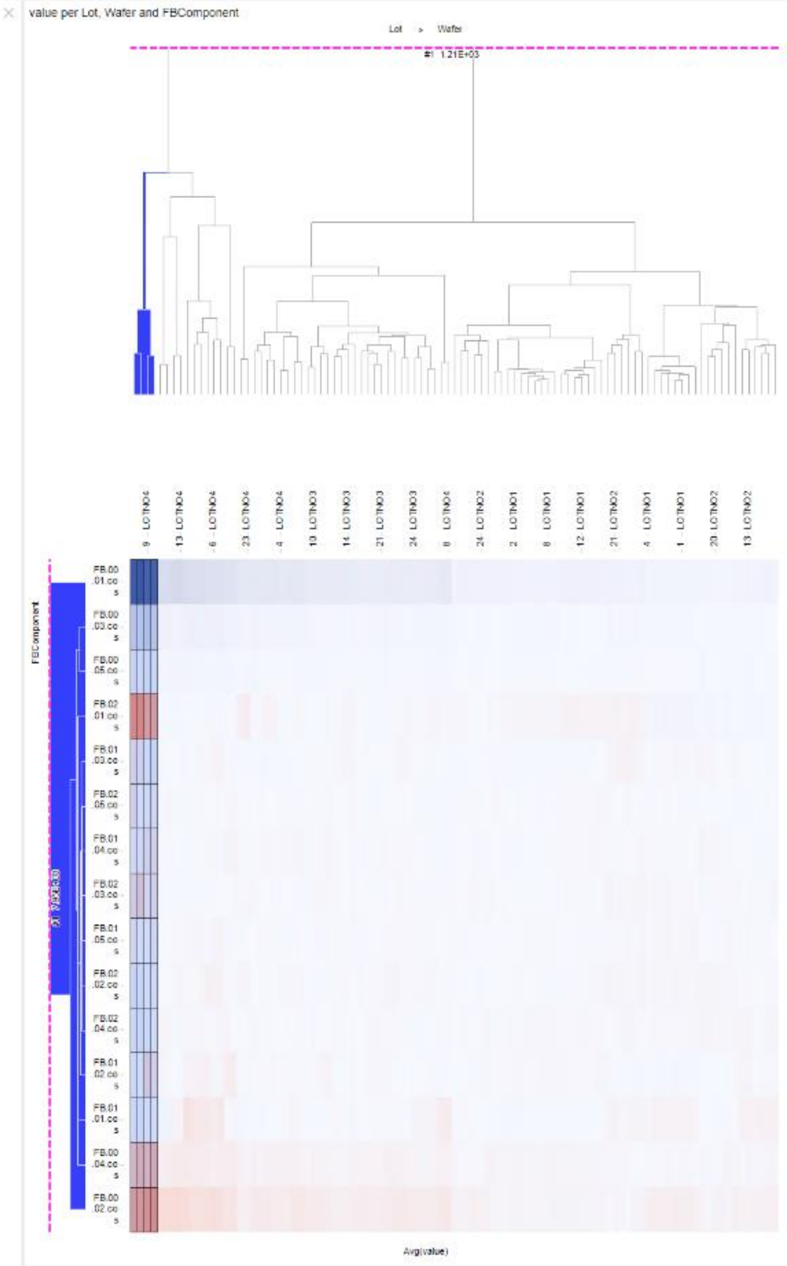
Find and explore anomalous events

Monitor anomalies

Predict when and why they occur


diey vs. diex
Color by:
param1140
P95.00 (27.80)
P90.00 (13.88)
Median (-29.45)
P30.00 (-48.41)
P20.00 (-61.70)
P10.00 (-82.79)
P5.00 (-101.17)
(Empty)
diey
diex
value per Lot, Wafer and FBComponent
Lot
Wafer
FBComponent
Avg(value)


diey vs. diex
Color by:
param1140
P95.00 (27.80)
P90.00 (13.88)
Median (-29.45)
P30.00 (-48.41)
P20.00 (-61.70)
P10.00 (-82.79)
P5.00 (-101.17)
(Empty)
diex
diey

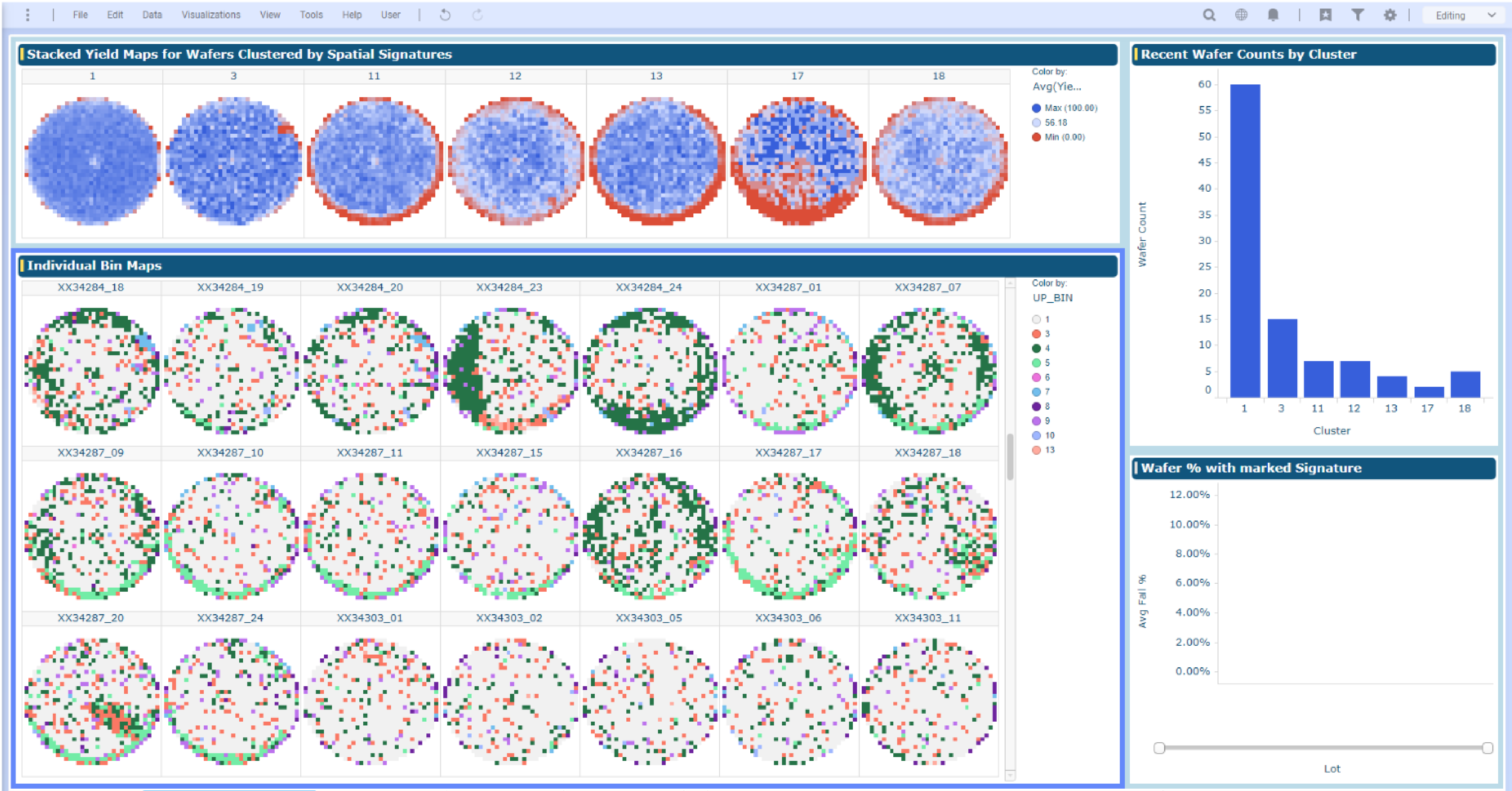

value per Lot, Wafer and FBComponent
Lot » Wafer
FBComponent
Avg(value)

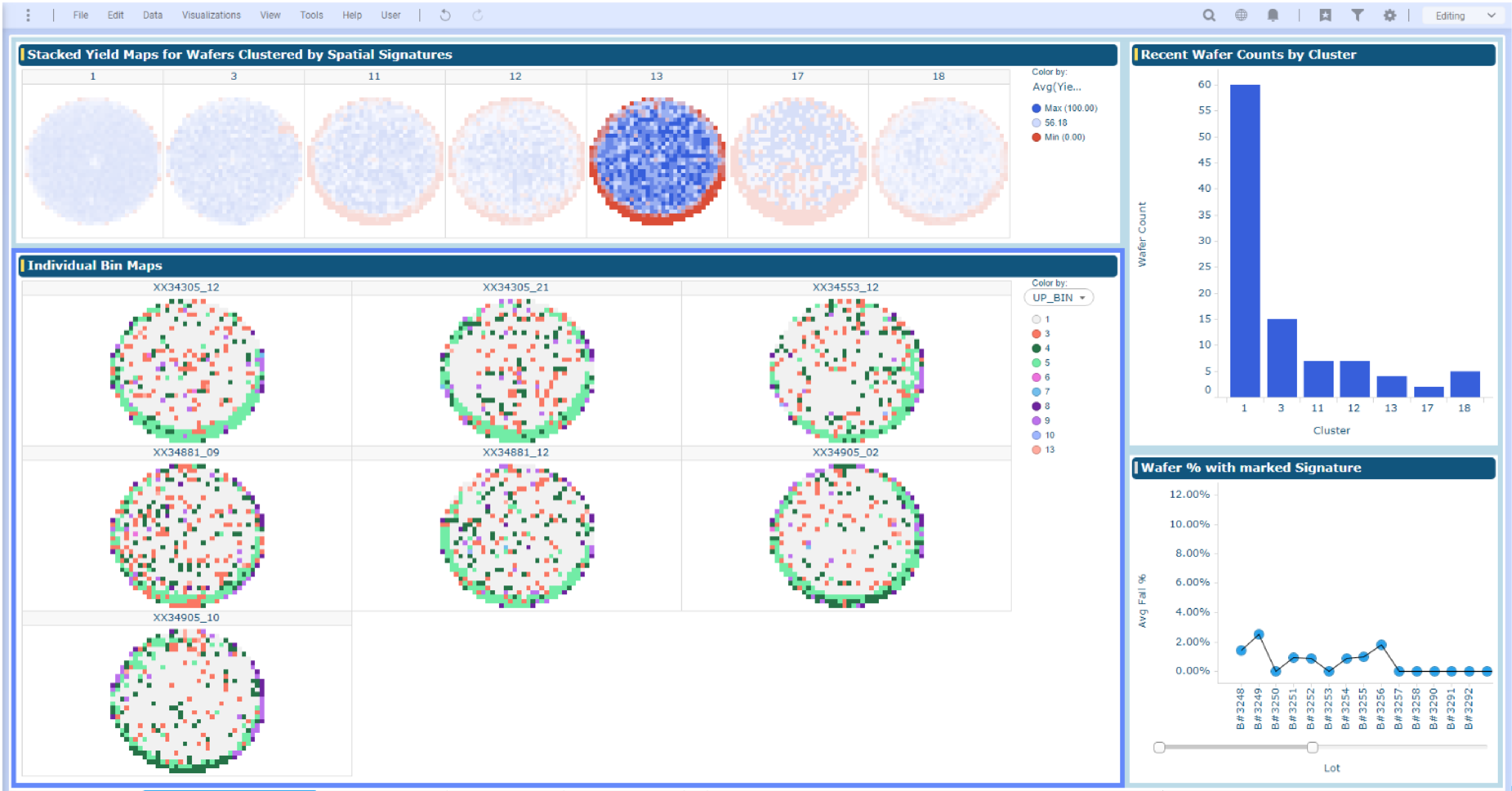
Anomaly Detection and Analysis

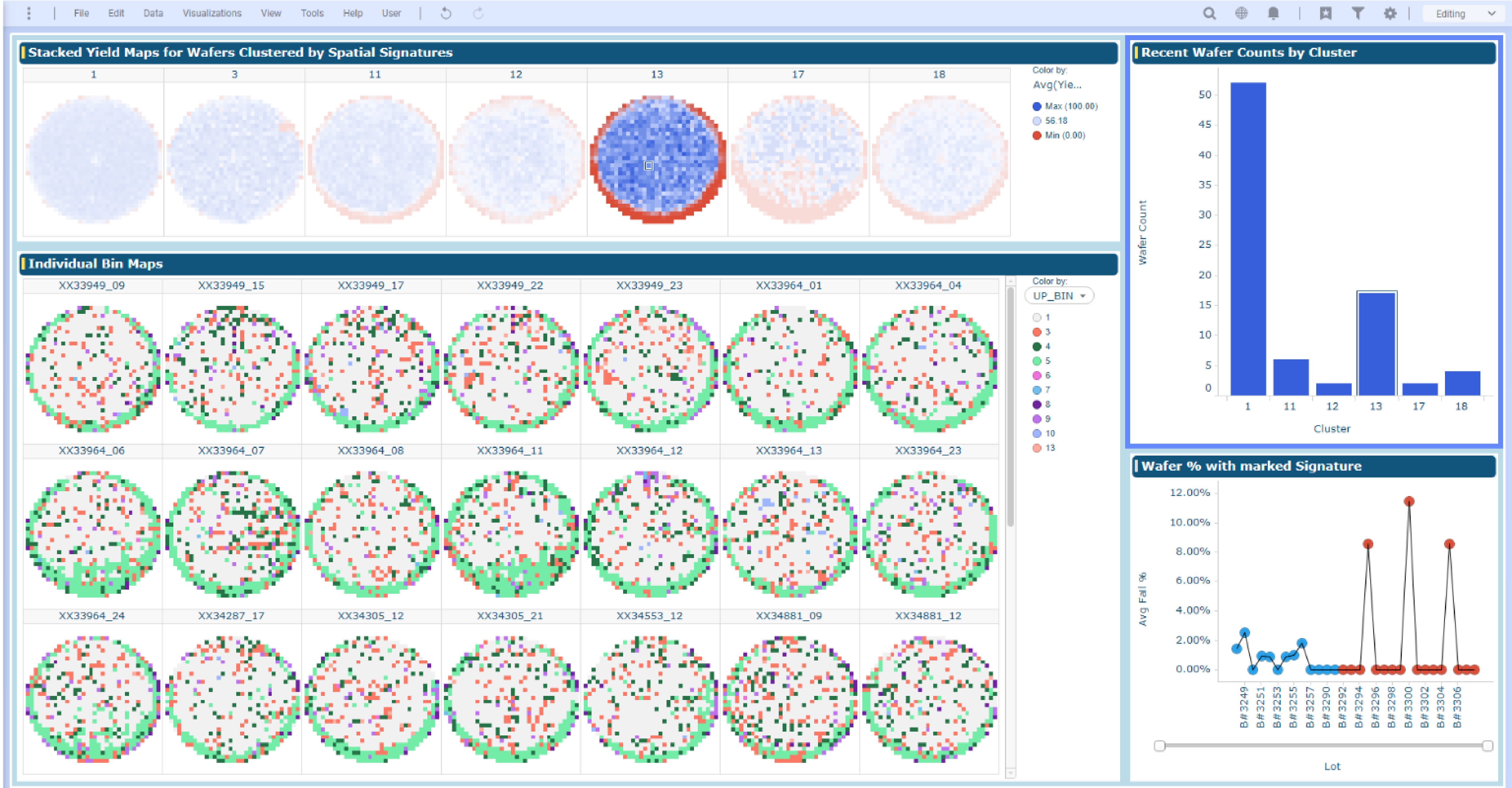
Find and explore anomalous events

Monitor anomalies

Predict when and why they occur


File
Edit
Data
Visualizations
View
Tools
Help
User
Editing
Stacked Yield Maps for Wafers Clustered by Spatial Signatures
1
3
11
12
13
17
18
Color by:
Avg(Yie...
Max (100.00)
56.18
Min (0.00)
Recent Wafer Counts by Cluster
Wafer Count
60
55
50
45
40
35
30
25
20
15
10
5
0
1
3
11
12
13
17
18
Cluster
Individual Bin Maps
XX34284_18
XX34284_19
XX34284_20
XX34284_23
XX34284_24
XX34287_01
XX34287_07
XX34287_09
XX34287_10
XX34287_11
XX34287_15
XX34287_16
XX34287_17
XX34287_18
XX34287_20
XX34287_24
XX34303_01
XX34303_02
XX34303_05
XX34303_06
XX34303_11
Color by:
UP_BIN
1
3
4
5
6
7
8
9
10
13
Wafer % with marked Signature
Avg Fail %
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
Lot


File
Edit
Data
Visualizations
View
Tools
Help
User
Editing
Stacked Yield Maps for Wafers Clustered by Spatial Signatures
1
3
11
12
13
17
18
Color by:
Avg(Yie...
Max (100.00)
56.18
Min (0.00)
Recent Wafer Counts by Cluster
Wafer Count
Cluster
Individual Bin Maps
XX34305_12
XX34305_21
XX34553_12
XX34881_09
XX34881_12
XX34905_02
XX34905_10
Color by:
UP_BIN
1
3
4
5
6
7
8
9
10
13
Wafer % with marked Signature
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
Avg Fail %
B#3248
B#3249
B#3250
B#3251
B#3252
B#3253
B#3254
B#3255
B#3256
B#3257
B#3258
B#3290
B#3291
B#3292
Lot


File
Edit
Data
Visualizations
View
Tools
Help
User
Editing
Stacked Yield Maps for Wafers Clustered by Spatial Signatures
1
3
11
12
13
17
18
Color by:
Avg(Yie...
Max (100.00)
56.18
Min (0.00)
Recent Wafer Counts by Cluster
Wafer Count
Cluster
Individual Bin Maps
XX33949_09
XX33949_15
XX33949_17
XX33949_22
XX33949_23
XX33964_01
XX33964_04
XX33964_06
XX33964_07
XX33964_08
XX33964_11
XX33964_12
XX33964_13
XX33964_23
XX33964_24
XX34287_17
XX34305_12
XX34305_21
XX34553_12
XX34881_09
XX34881_12
Color by:
UP_BIN
1
3
4
5
6
7
8
9
10
13
Wafer % with marked Signature
Avg Fail %
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
B#3249
B#3251
B#3253
B#3255
B#3257
B#3290
B#3292
B#3294
B#3296
B#3298
B#3300
B#3302
B#3304
B#3306
Lot

Anomaly Detection and Analysis

Find and explore anomalous events

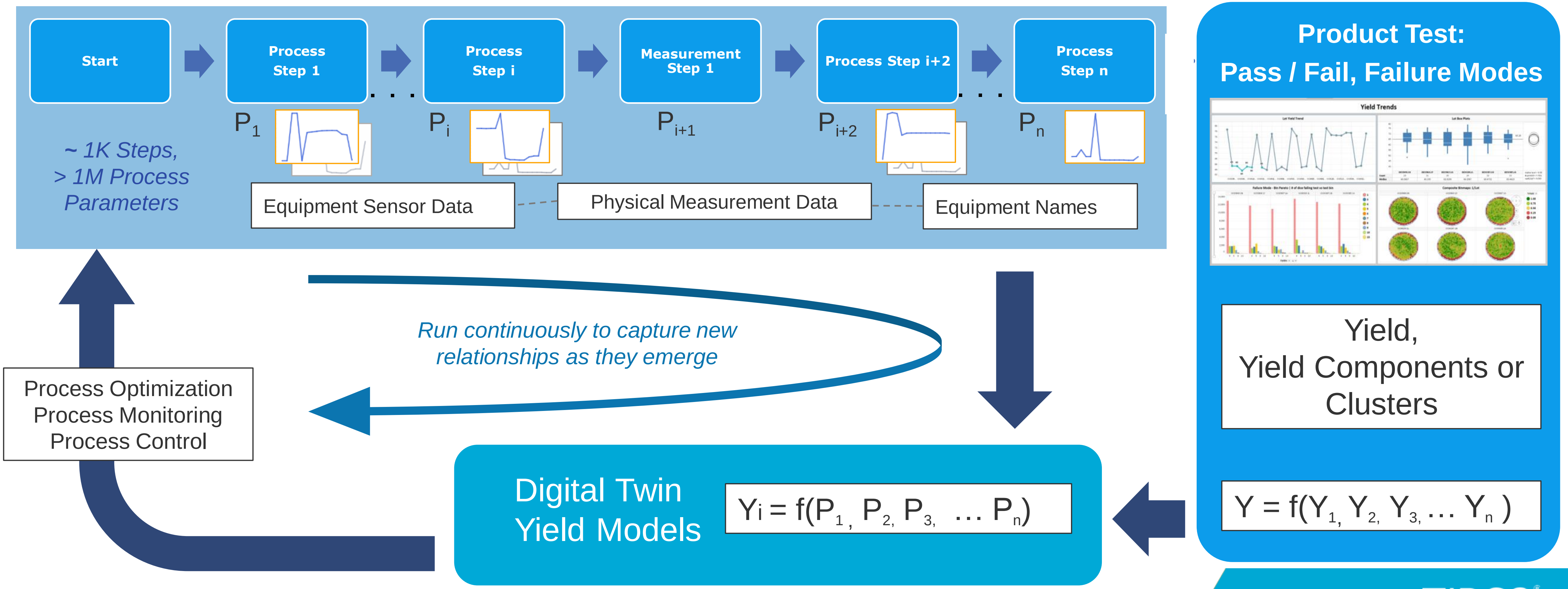
Monitor anomalies

Predict when and why they occur

Digital Twin for Semiconductor Yield

Digital Twins for Semiconductor Manufacturing Yield: Wide-and-Big Data Analysis

Build Models to Relate Product Yield Failure Modes (Y_i) with Process Parameters (P_i)



TIBCO

The Extreme Challenge of Big & Wide Data

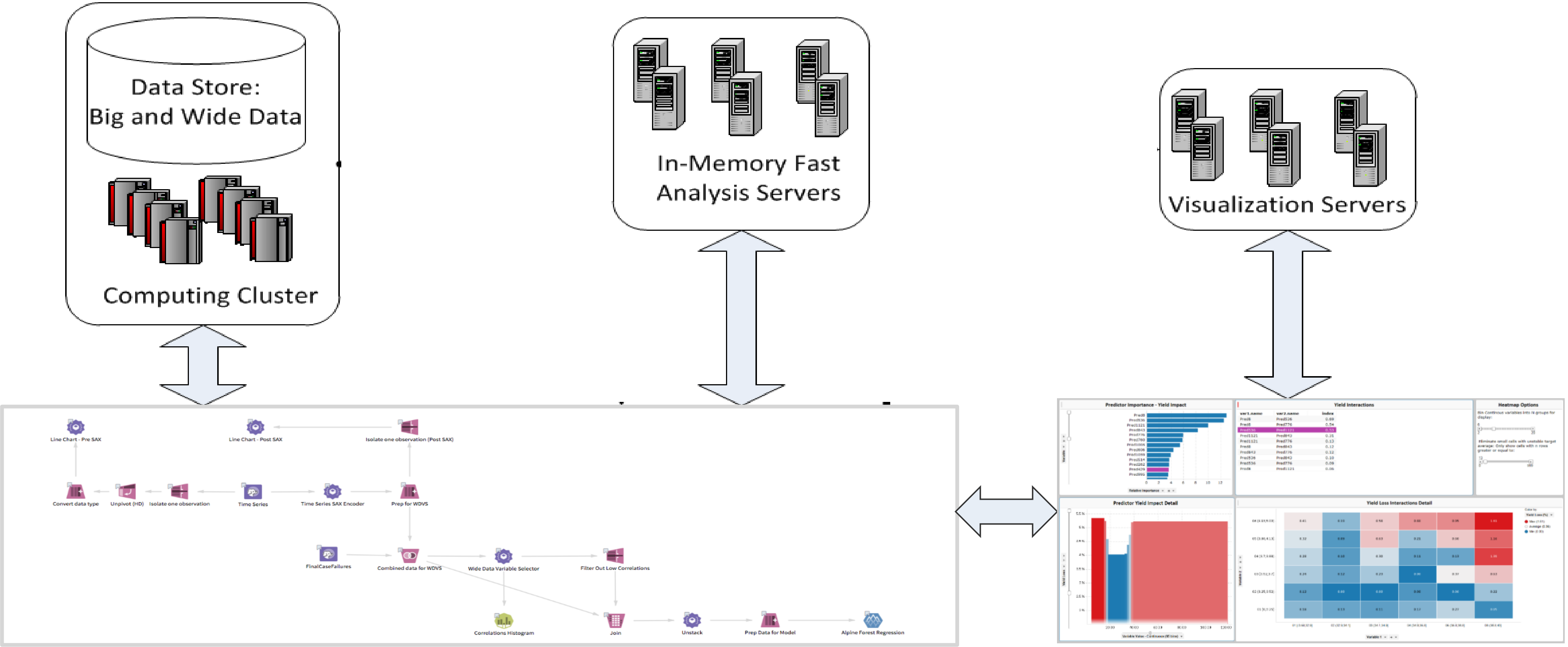
- **Not just big data** – many rows: lots, wafers, die, units
- **Also wide data** – many columns: > 1M process parameters
  - Sensor traces
    - Time series for every sensor on each machine in each run
  - Physical measurements
    - Film thickness, critical dimensions, layer-to-layer overlay, defect classes & counts
  - Equipment and process attributes
    - Machine and component IDs, process recipe info
  - Supplies
    - Chemical batch IDs, QA sample data

> "Today [semiconductor] fabs collect more than 5 billion sensor data points each day. The challenge is to turn massive amounts of data into valuable information."
>
> *—Ann Kellehere, VP of the Technology and Manufacturing Group, Intel*

TIBCO®

Solution Architecture

Data Prep, Feature Engineering & Selection | **Further Feature Selection & Model Building** | **Visualization of Results**



- In-database parallelized computing
- Leverages Hadoop, Apache Spark

- In-memory dedicated fast server

- Interactive in-memory visualization environment

Performance Benchmarks & Conclusions

- Demonstrated performance for time series data from 20,000 sensors, 10,000 wafers in **under 2 minutes**
- Current system scales to time series for 20,000 sensors, 100,000 wafers (**2.5 TB**) with results in **15 minutes**
  - More capacity and better performance can be achieved by adding nodes to the Spark cluster
- Working with top memory manufacturer to deploy production system
- System can provide automated real-time feedback on emerging equipment issues affecting yield

Big Data Feature Selection Performance Benchmarks – Run Time[1] (minutes)					
	20 Sensors (1K variables)	**200 Sensors** (10K variables)	**2K Sensors** (100K variables)	**20K Sensors** (1M variables)	***Dataset Size** for 1M Variables*
1K Wafers	0.47	0.48	0.72	1.0	25 GB
10K Wafers	0.50	0.53	0.77	1.75	253 GB
100K Wafers	0.53	0.67	1.25	15.15	2,530 GB

[1]Test Conditions:
- Data stored in Hadoop data source
- 25 node Spark cluster – 16 cores, 32 GB for each node
- Each sensor time series compressed to 50 variables with SAX encoder prior to feature selection step

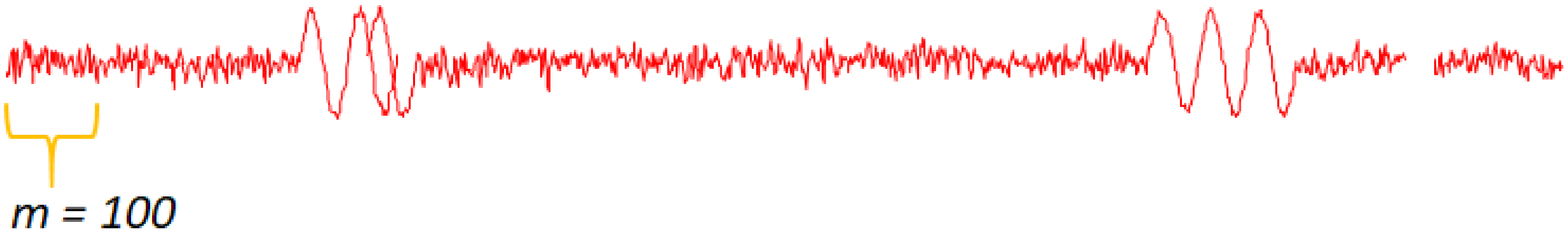
Longitudinal Anomaly Analysis

Subsequence Search

A New Method for Identifying Anomalous Patterns in Time Series (Trace Analytics)

Mueen's Algorithm for Similarity Search

Mueen's Algorithm for Similarity Search (MASS) is specialized for finding anomalous (versus typical) subsequences of time series



Extremely fast algorithm for this use case

Suitable for further acceleration using GPU

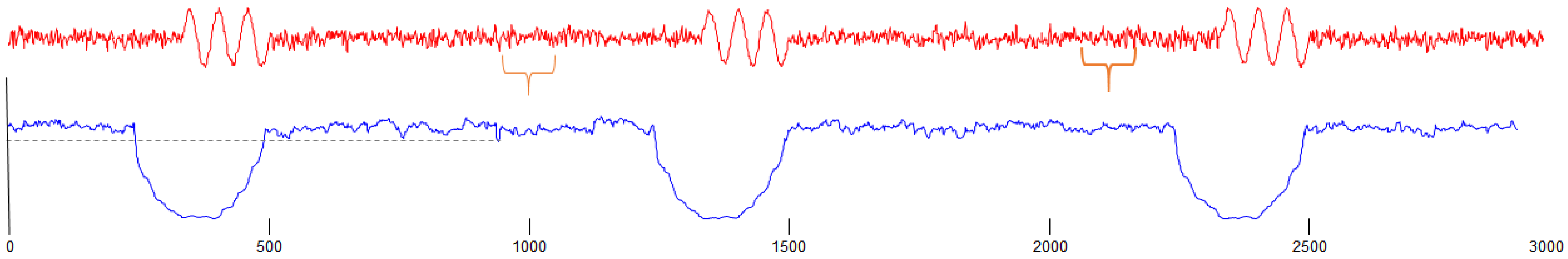
Material adapted from:
https://www.cs.ucr.edu/~eamonn/matrix_profile_i.pptx

Mueen’s Algorithm for Similarity Search

Quickly create a matrix profile = a partial distance matrix

This uses a sliding window to define a series of subsequences

The Matrix Profile plots the distance of each subsequence to its nearest match, with the time sequence of the start of each subsequence on the x-axis



Red = Original Time Series

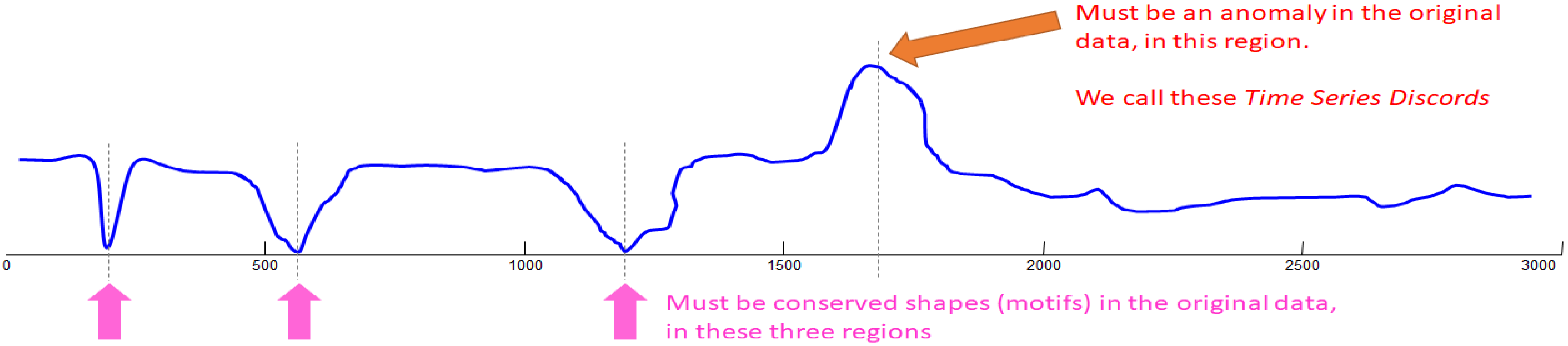
Blue = Matrix Profile

Material adapted from:
https://www.cs.ucr.edu/~eamonn/matrix_profile_i.pptx

How to "Read" a Matrix Profile

Where you see relatively low values, you know that the subsequence in the original time series must have (at least one) relatively similar subsequence elsewhere in the data (such regions are "motifs" or reoccurring patterns)

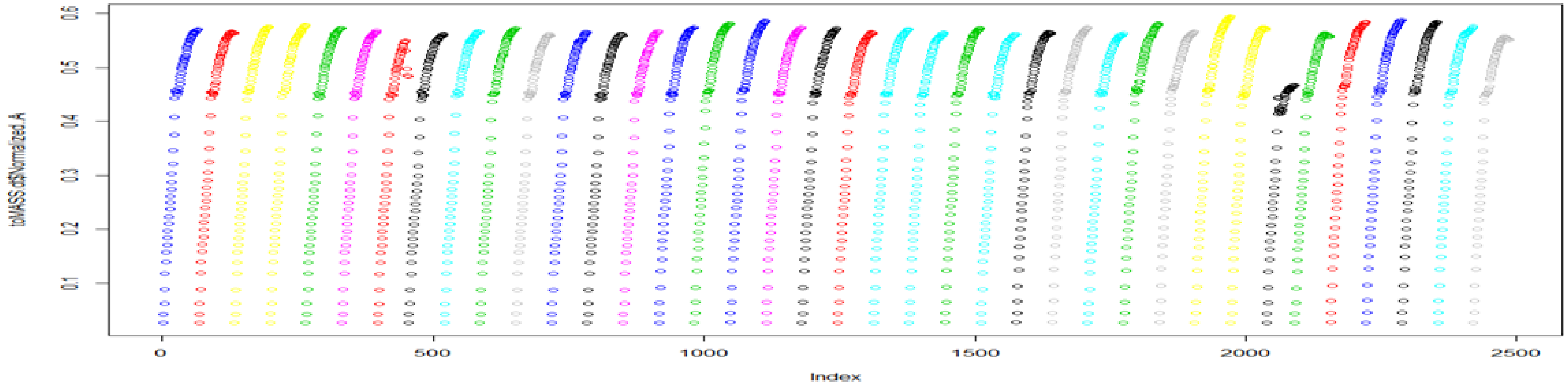
Where you see relatively high values, you know that the subsequence in the original time series must be unique in its shape (such areas are "discords" or anomalies)



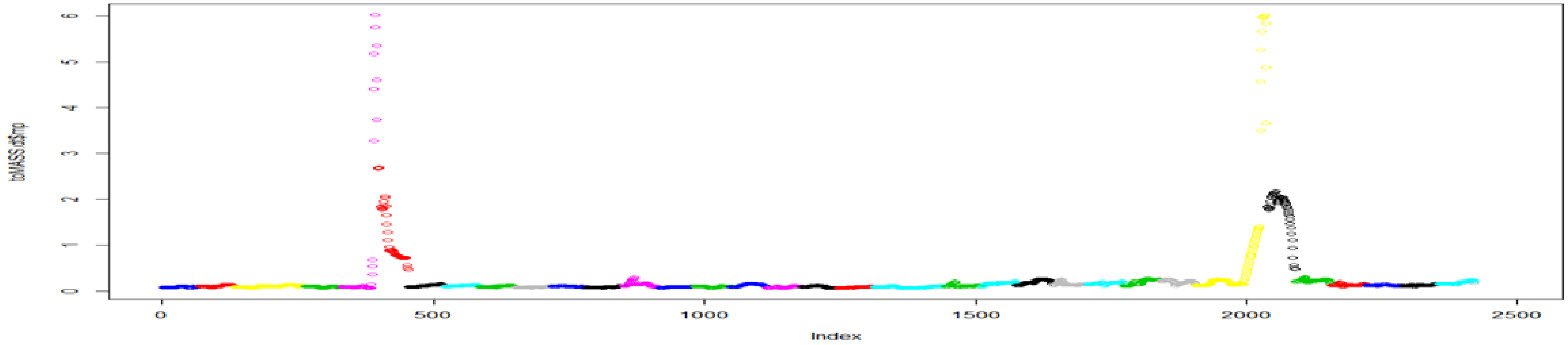
TIBCO

Manufacturing Batches

Raw Amperage - Each color delimits a batch



Matrix Profile highlights anomalies - set sliding window close to batch size



TIBCO®

Community

TIBCO® Exchange

https://community.tibco.com/exchange

Extend the capabilities of your TIBCO® products with extensions, add-ons, plug-ins,

Anomaly Detection Template for TIBCO Spotfire®

This template detects anomalous data points in a dataset using an autoencoder algorithm.

Data Function for TIBCO® Data Science - Team Studio in TIBCO Spotfire®

This data function enables users to execute a TIBCO® Data Science - Team Studio workflow from Spotfire.

For More Info On TIBCO Data Science - Team Studio | Edit This Module | Create New Module

IoT Accelerator

Capture and analyze sensor data in real-time from your Internet of Things devices with TIBCO's IoT Accelerator. Integrate through industry-standard protocols like OSI PI, MQTT, and Web Services. Alternatively, implement custom adapters for your own protocols, all the way down to baseline serial port integration. Apply custom validations, cleansing policies, rules, and feature statistics on data feeds to identify trends and gain insight.

Try Now | Edit This Module | Create New Module

Random Forest - Data Function for TIBCO Spotfire®

Random forests are an ensemble decision tree machine learning method for classification and regression.

Statistical Process Control Template for TIBCO Spotfire® using TIBCO® Data Science - Statistica

This TIBCO Spotfire template is designed to enable user to build wide range of quality control charts with possibility do define charts specifications interactively according to user's needs. This comprehensive template is constructed based on Statistica data function and utilizing wide range of parameter settings already implemented in TIBCO Data Science - Statistica. It is an example of no-code data function.

Statistica Data Functions | Edit This Module | Create New Module

Risk Management Accelerator

Identify potentially risky activities in a high-frequency event stream using machine learning in TIBCO's Risk Management Accelerator. Build supervised and/or unsupervised models and hot deploy these to the streaming event processing platform, then score events in real-time. Raise alerts when potentially risky behaviour is detected.

Try Now | Edit This Module | Create New Module

Gradient Boosting Machine Analysis Template for TIBCO Spotfire®

This template is used to create a GBM machine learning model to understand the effects of predictor variables on a single response.

Loss Distribution Approach to Operational Risk - Analysis Template for TIBCO Spotfire®

This analysis implements simple frequency-severity models for Operational Risk event types. This forms the basis of the Loss Distribution Approach alternative in the Basel regulations.

Try Now | Edit This Module | Create New Module

Dynamic Pricing Accelerator

Take control of your pricing platform with TIBCO's Dynamic Pricing Accelerator. Applicable to insurance, retail, travel, or any industry where personalized pricing would be an advantage. Transform into an algorithmic business by deploying personalized pricing and propensity models that you build and manage to gain advantage over competitors while using industry-standard modelling languages. Hot deploy these models and watch the results in real-time with the TIBCO Insight Platform.

Try Now | Edit This Module | Create New Module

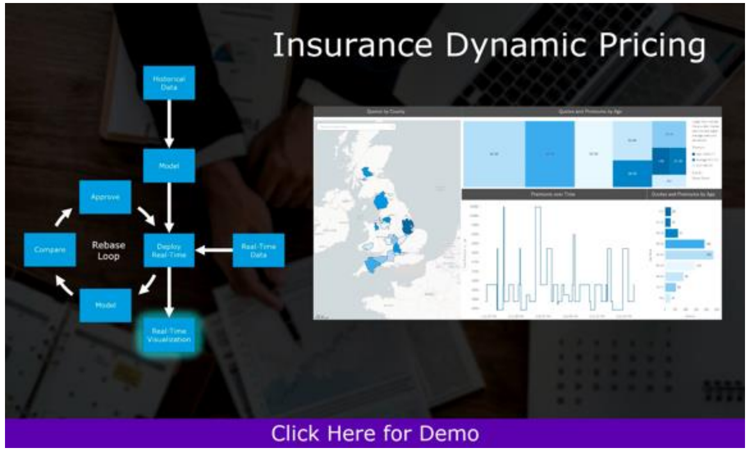
TIBCO® Spotfire® | TIBCO® Data Science | TIBCO® Streaming

TIBCO®

AI in Operations

Cloud Starters, Accelerators, Analytic Apps

Thoughtleader-Led Solutions



TIBCO®

Questions & Contact

Steven Hillion
Sr. Director, Data Science
shillion@tibco.com
@StevenHillion

Michael O'Connell
Chief Analytics Officer
moconnell@tibco.com
@MichOConnell

TIBCO Community
community.tibco.com

TIBCO Exchange
community.tibco.com/exchange

TIBCO Tech Blog
community.tibco.com/blog

Acknowledgements

Dr. Tom Hill, Mike Alperin, Nico Rode, David Katz, Jagrata Minardi, Siva Ramalingam

Please complete the session survey in the mobile app.