

The background features a vibrant, multi-colored gradient. A diagonal line divides the image into two main sections. The upper-left section is a solid blue, while the rest of the image is a gradient of purple, orange, and yellow. The text 'AWS re:Invent' is rendered in white, with 'AWS' in a smaller font above 're:Invent'.

AWS  
re:Invent

AIM201-S

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

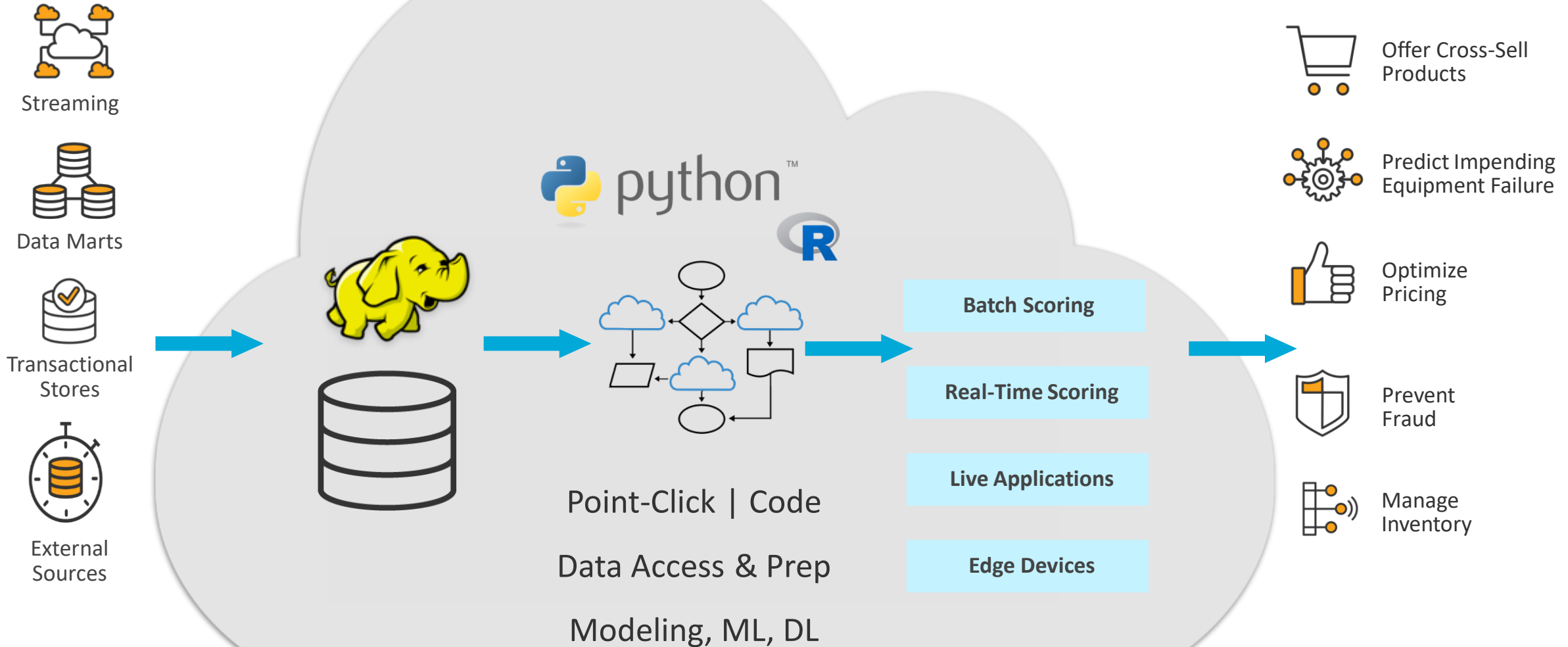
## **Steven Hillion**

Sr Director, Data Science  
@StevenHillion

## **Michael O'Connell**

Chief Analytics Officer  
@MichOConnell

# The Ideal Data Science Platform

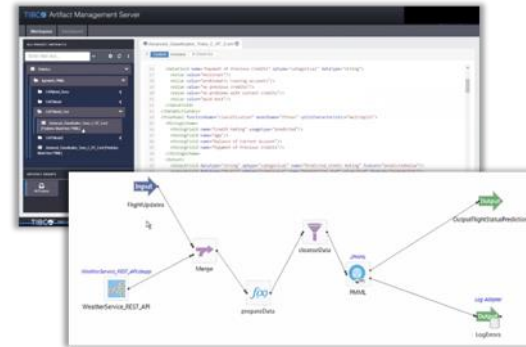
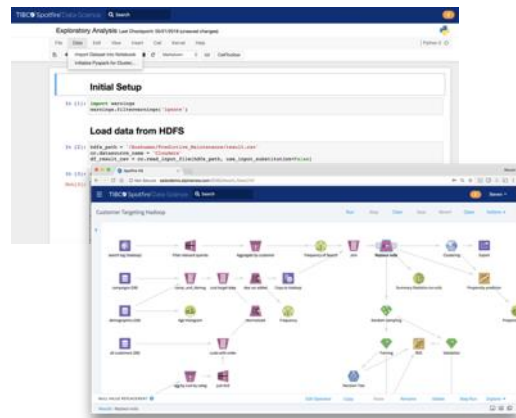


# Agenda

- 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

FUNCTION	Data Access/Prep	Modeling	Operations	Business Apps
	<ul style="list-style-type: none"> <li>+ Distributed compute</li> <li>+ Feature engineering</li> <li>+ Reusable templates</li> </ul>	<ul style="list-style-type: none"> <li>+ Visual composition</li> <li>+ Multilingual notebook</li> <li>+ Native ML &amp; OS</li> <li>+ Auto-ML, data prep</li> </ul>	<ul style="list-style-type: none"> <li>+ Model lifecycle management</li> <li>+ Batch automation</li> <li>+ Real-time event processing</li> <li>+ REST, applications, embedding</li> </ul>	<ul style="list-style-type: none"> <li>+ Engineering/IoT</li> <li>+ Customer analytics</li> <li>+ Risk management</li> <li>+ Supply chain</li> </ul>



**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*

**AWS Machine Learning Partner Solutions**  
Provide solutions that help organizations solve their data challenges, enable machine learning and data science workflows or offer SaaS based capabilities that enhance end applications with machine intelligence.

**AWS Machine Learning Competency**

Technology Partners

**TIBCO**

TIBCO Spotfire Data Science is a cloud platform for all stages of the analytics process from data ingestion to machine learning. Data engineers and analysts work side-by-side on visual workflows.

[Solution Overview](#) | [Customer Success](#) | [Contact TIBCO](#) | [Rate this Partner](#)

- MACHINE LEARNING PARTNER SOLUTIONS
- Data Services**
- Platform Solutions
- SaaS and API Solutions
- Consulting Partners

**aws partner network**

Advanced Technology Partner

Machine Learning Competency

**Creating a 1.3 Million vCPU Grid on AWS using EC2 Spot Instances and TIBCO GridServer**  
by Jeff Barr | on 09 MAY 2018 | in Amazon EC2, News | Permalink | Share

Many of my colleagues are fortunate to be able to spend a good part of their day sitting down with and listening to our customers, doing their best to understand ways that we can better meet their business and technology needs. This information is treated with extreme care and is used to drive the roadmap for new services and new features.

**TIBCO Data Science for AWS (10 users)**  
★★★★★ (0) | Version 6.4.1 | Sold by TIBCO Software Inc.  
Starting from **\$16.00 to \$16.00/hr** for software + AWS usage fees  
Scale data science across your organization to solve complex challenges faster and speed innovation with TIBCO Data Science for AWS Marketplace, a collaborative platform for operationalizing data science. Access Amazon EMR and RedShift and enable data scientists to create innovative solutions using...  
Linux/Unix, CentOS 7 - 64-bit Amazon Machine Image (AMI)

**TIBCO Text Similarity Analyzer**  
★★★★★ (0) | Version v1 | Sold by TIBCO Software Inc.  
Engineers word/document features on a corpus with NLP methods, and uses these features to compare new text to the corpus.

**TIBCO Autoencoder for Anomaly Detection**  
★★★★★ (0) | Version v1 | Sold by TIBCO Software Inc.  
Identifies potential anomalies from transaction and or sensor data with a deep learning autoencoder.

**TIBCO Hospital Readmission**  
★★★★★ (0) | Version v1 | Sold by TIBCO Software Inc.  
Predicts hospital readmission rates from DRG codes, billing and EMR data.

**TIBCO Spotfire Analytics for AWS (Hourly)**  
★★★★★ (0) | Version 10.5.1 | Sold by TIBCO Software Inc.  
Starting from **\$1.20/hr** or from **\$8,400.00/yr** (20% savings) for software + AWS usage fees  
TIBCO Spotfire® provides visual analytics for deep insights into data from Amazon Redshift, EMR, Aurora, Databricks, SAP HANA One, Oracle, Microsoft Excel, SQL Server and more. Explore data, and create AI-powered visual dashboards in minutes. Easily scale from a small team to the entire organization...

# 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  
flights (e.g., to Mars)

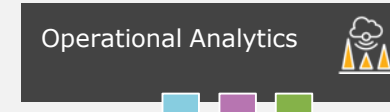
# TIBCO analytics transformation platform

Powered by shared data assets



**AUGMENT INTELLIGENCE**

ANALYTICS ACTIONS



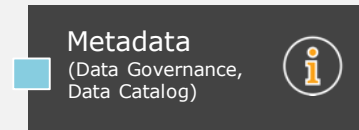
**UNIFY DATA**

DATA OPERATIONS



INFORMATION MANAGEMENT

METADATA



MDM / RDM



- Metadata
- Master & Reference Data
- Transactional Data

**INTERCONNECT EVERYTHING**

EVENTS



INTEGRATION



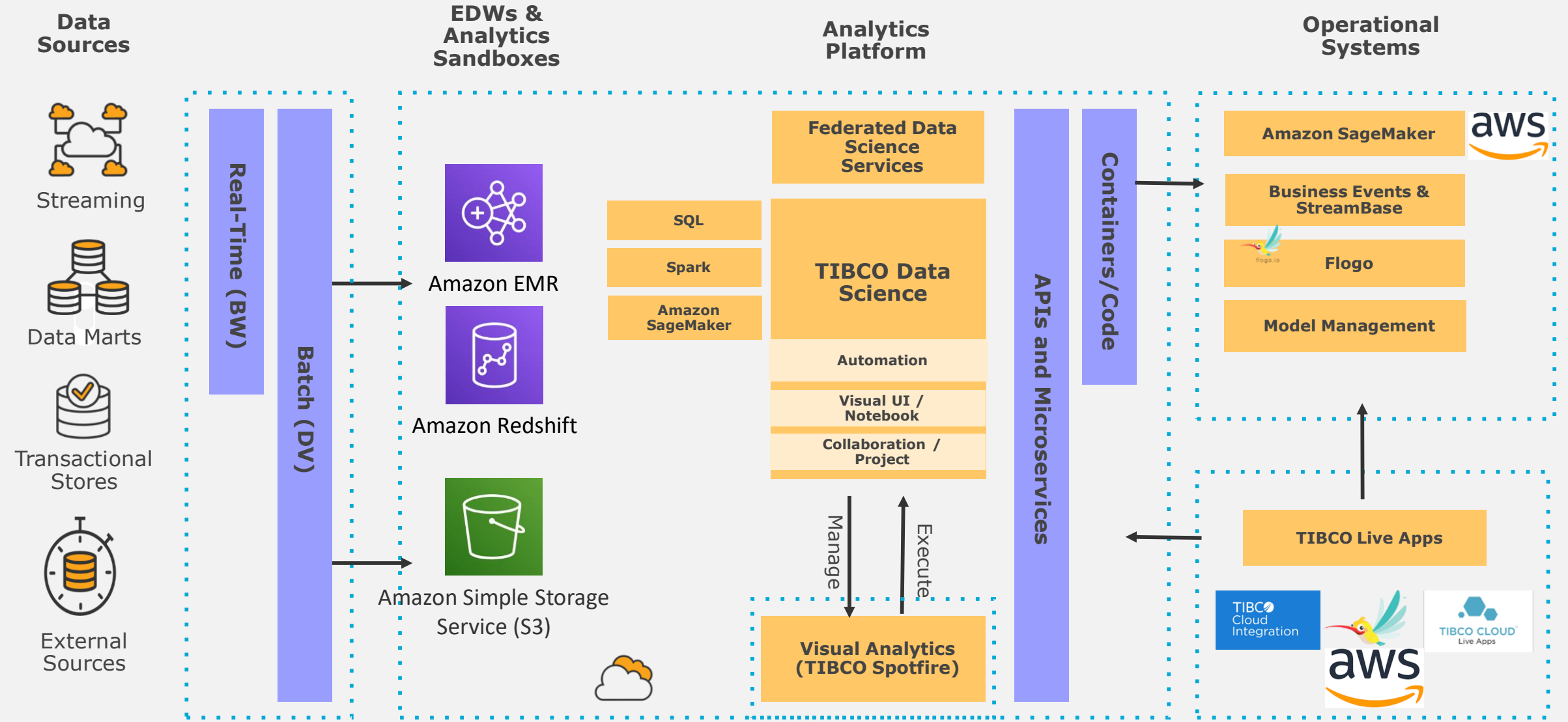
DATA SOURCES



- Cloud Native
- Open Platform
- AI Foundation



# TIBCO Data Science and AWS

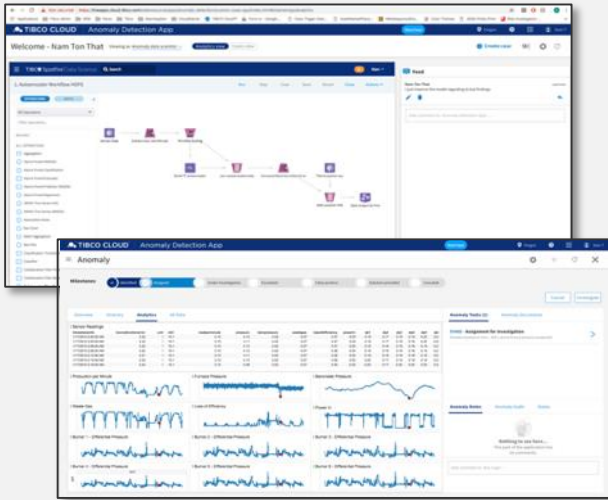


**AWS Deployments**

# TIBCO Data Science Solutions on AWS

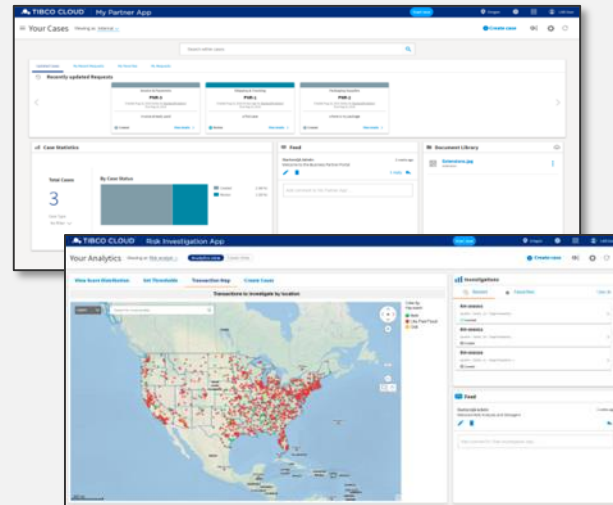
Cloud Apps: *Visual Analytics, Data Science, Streaming, Case Management*

## Anomaly Detection



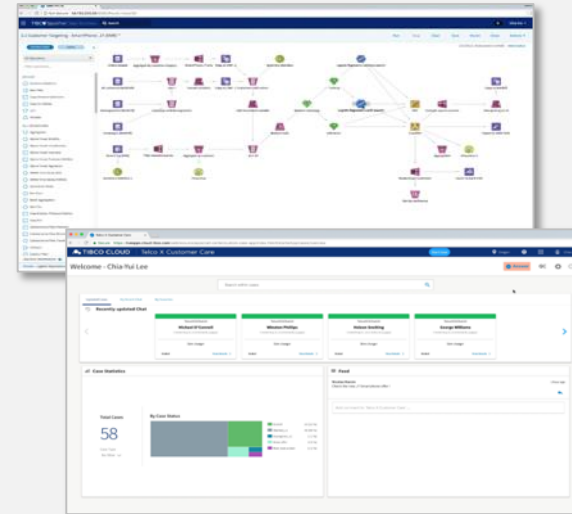
Review Status: *TIBCO Spotfire*  
Identify issues, sweet spots

## Risk Management



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

## Customer Engagement



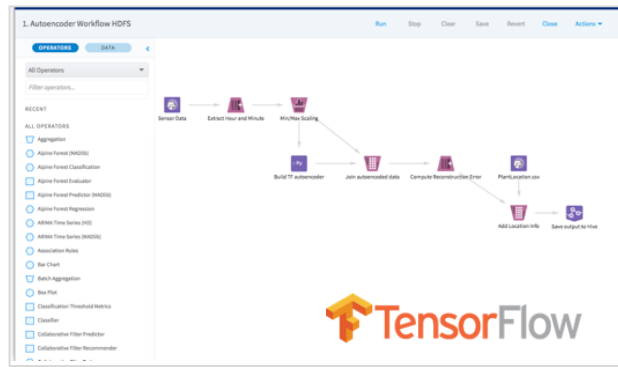
Case Manage: *TIBCO Live Apps*  
Investigate identified cases  
Audit trail + recycle

## Starter Set

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

# Anomaly Analysis Solution Overview

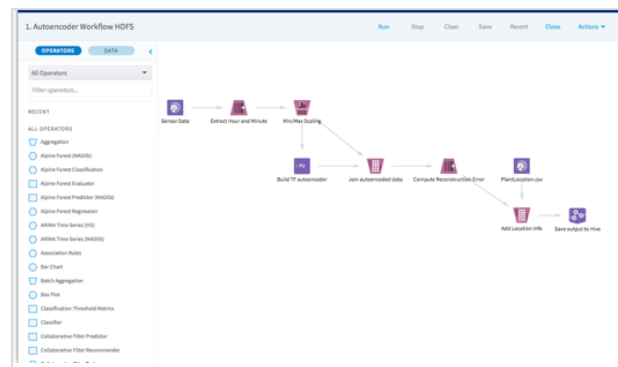
- 1 Collect data from equipment, normalize, model to predict magnitude of anomaly – **TIBCO Data Science & AWS**



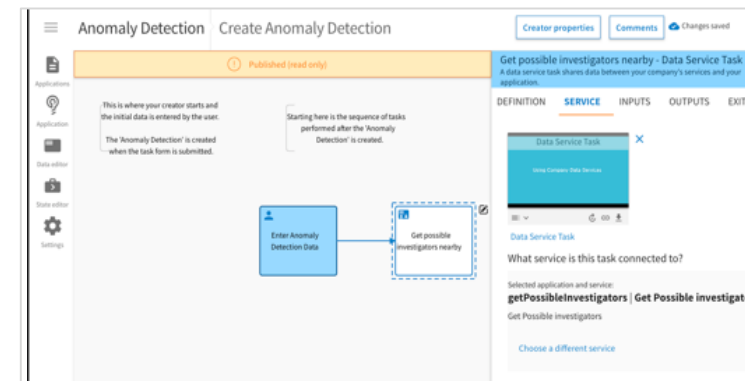
- 3 Alert raised and case created – **TIBCO Cloud Live Apps**



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



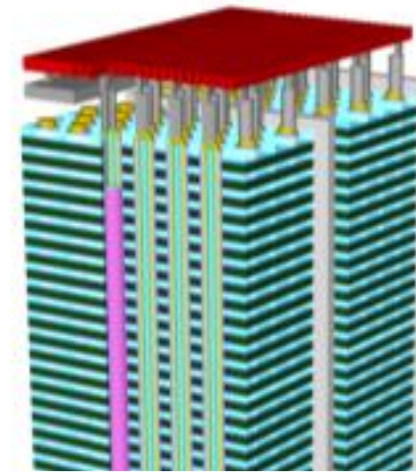
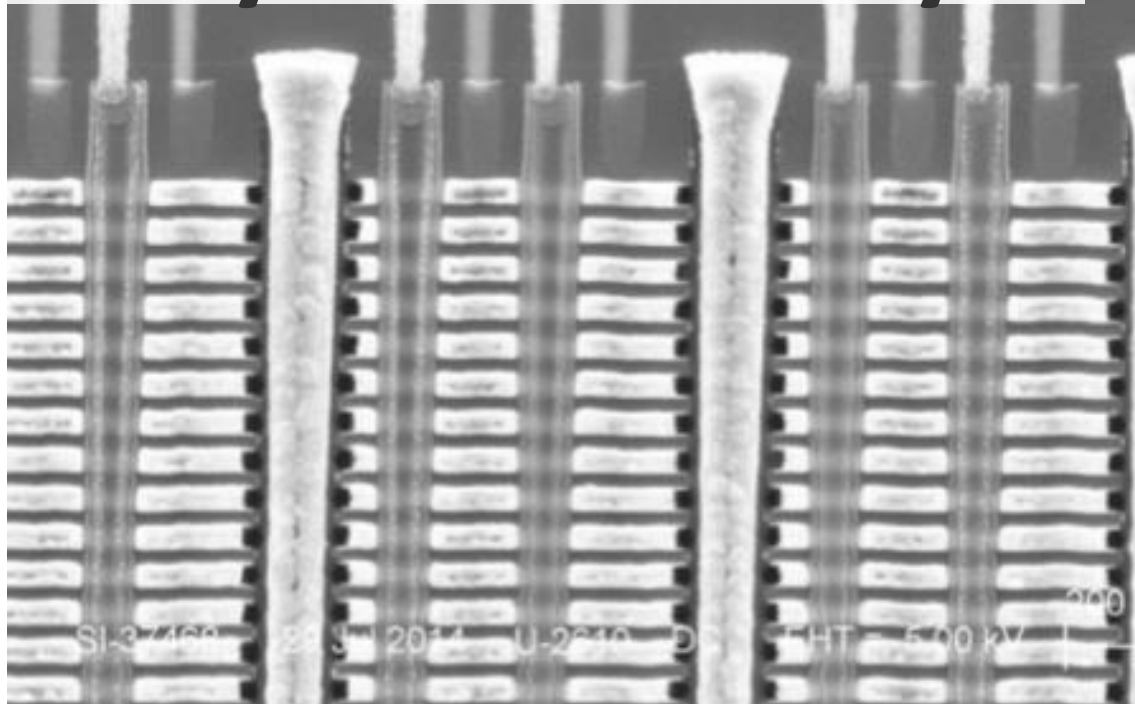
- 4 Case manager investigates and takes action to the equipment – **TIBCO Cloud Integration**



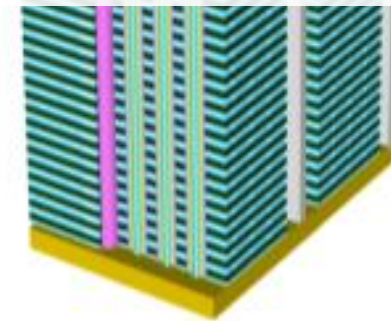
# Data Challenges in High-Tech Manufacturing

TECHINSIGHTS

## Stack 3D Flash Memory Cell Layers Vertically



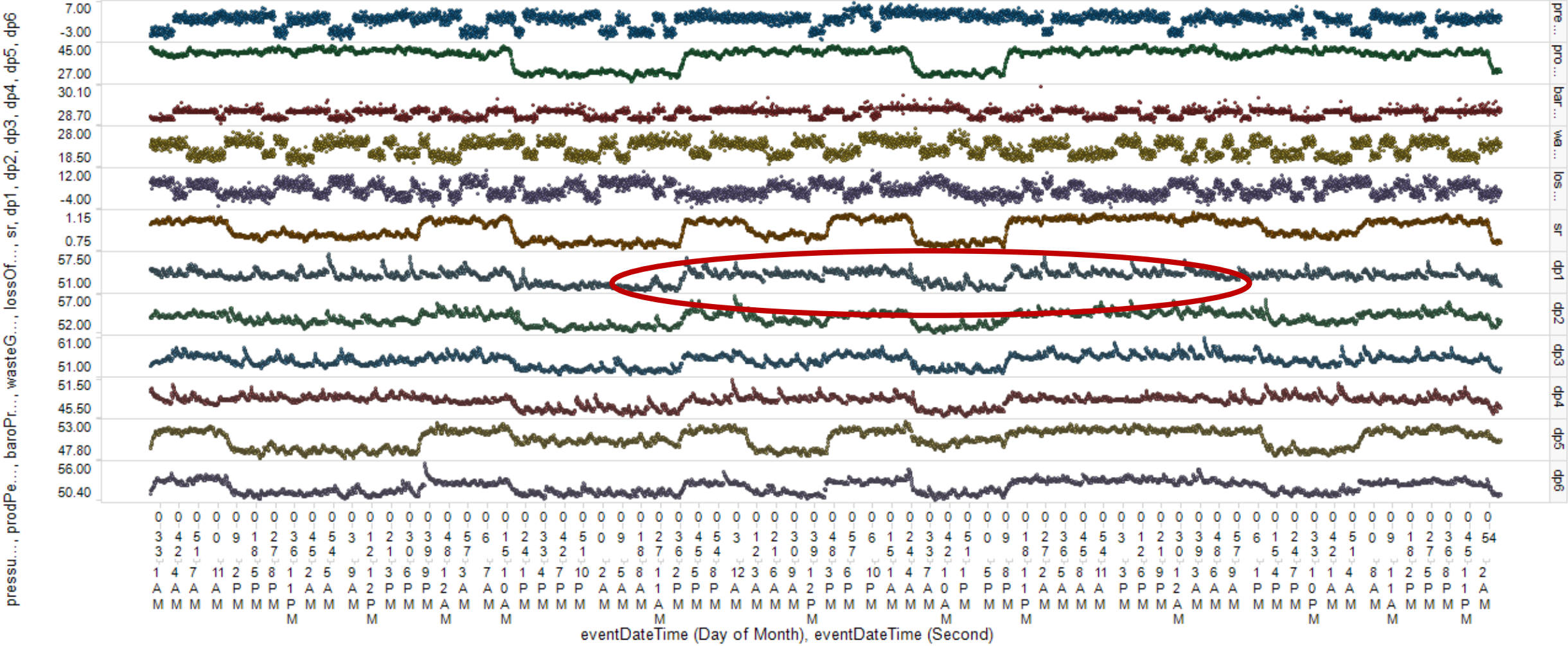
96 Memory  
Cell Layers



# Hot Paths to Anomaly Detection

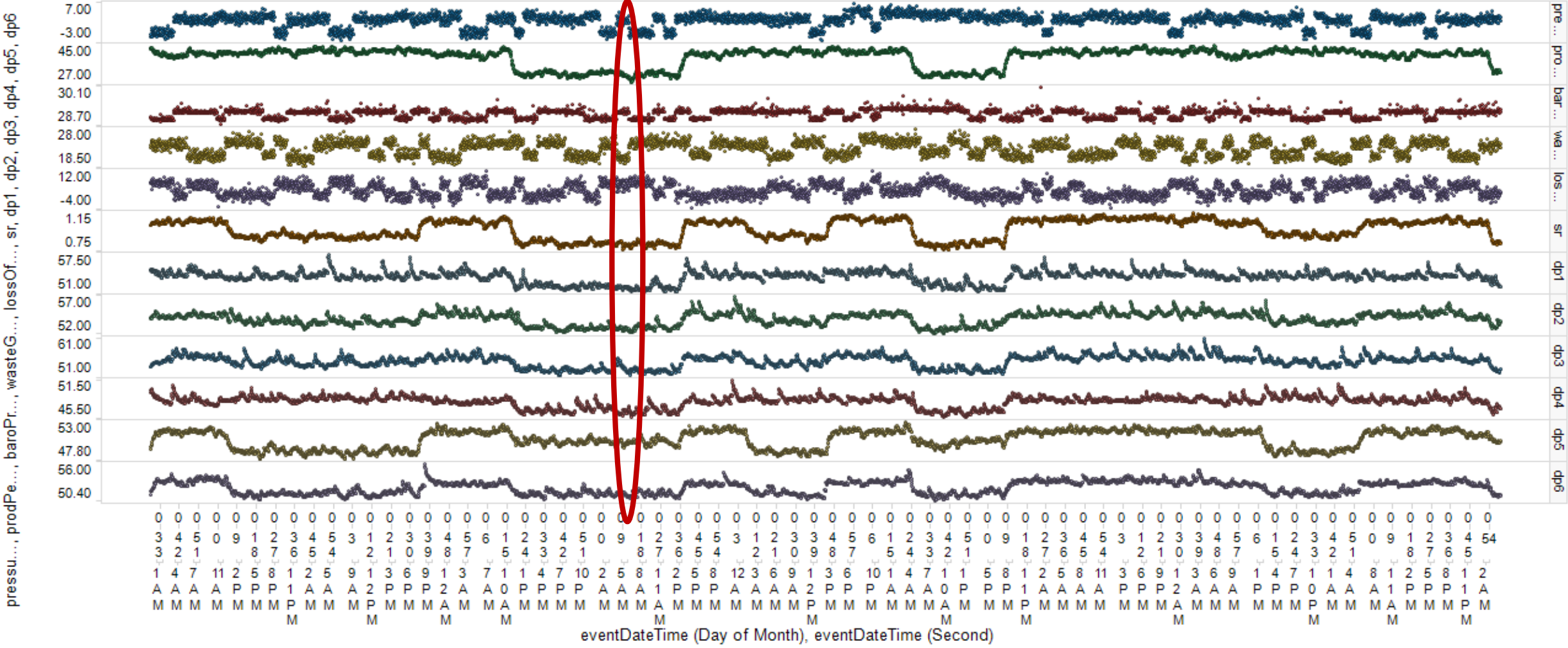
# Longitudinal Anomaly Analysis

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

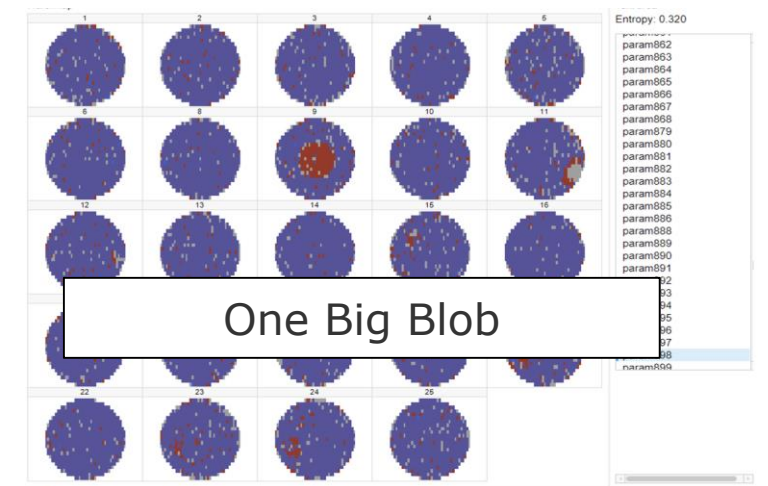
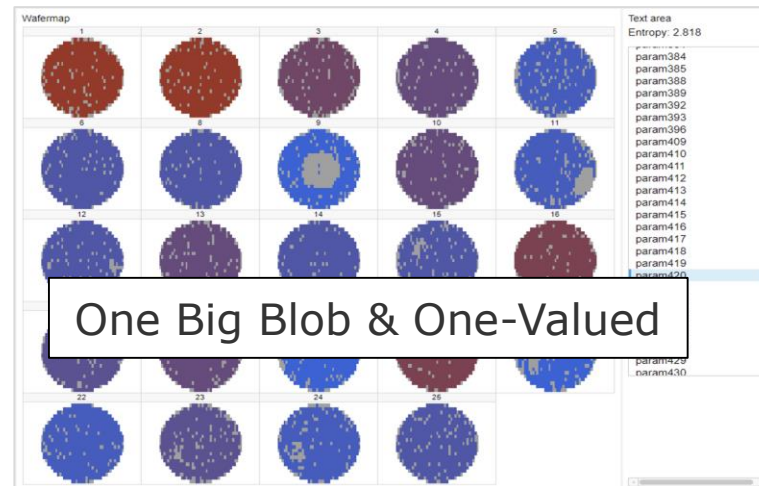
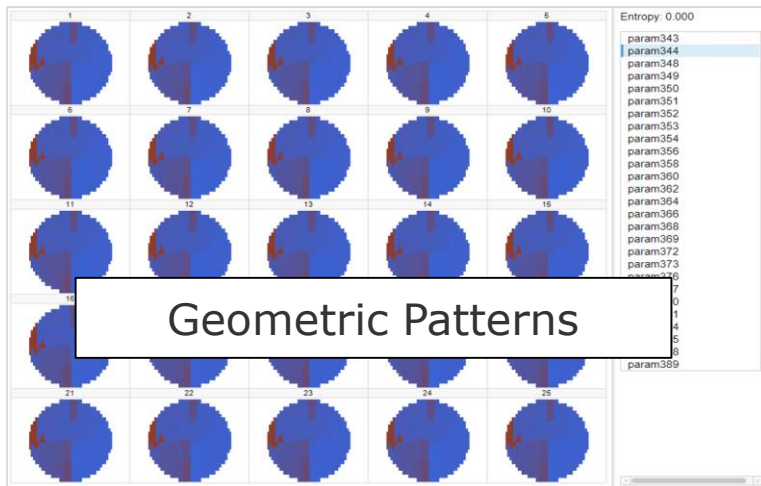
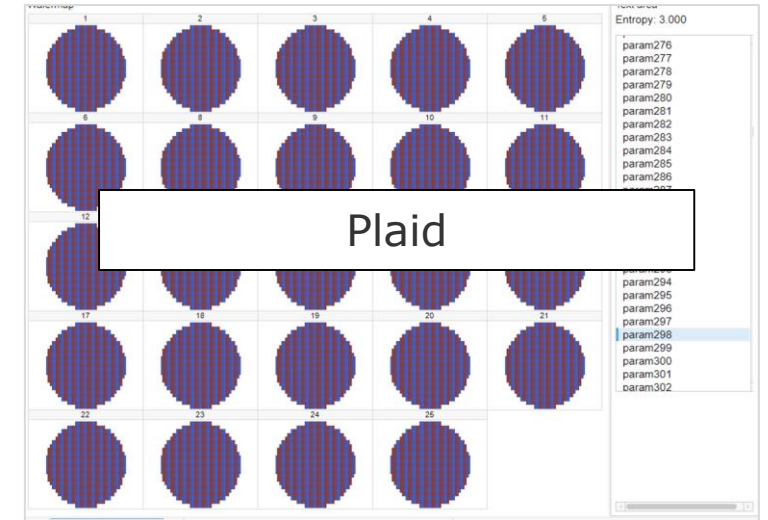
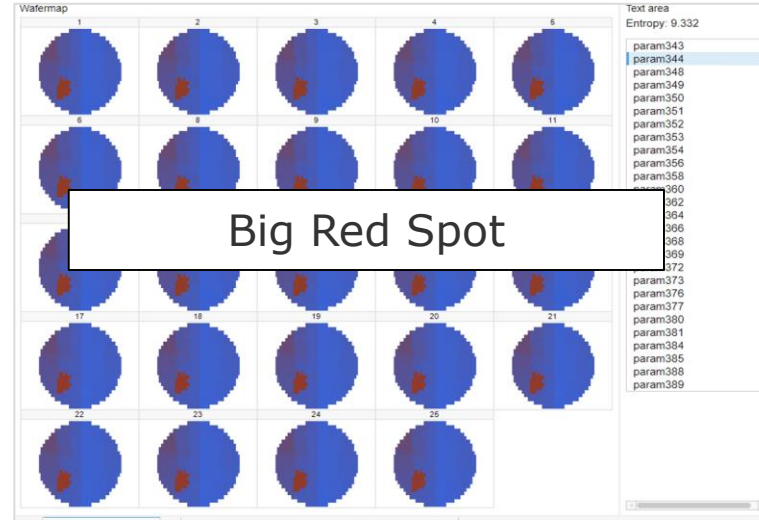
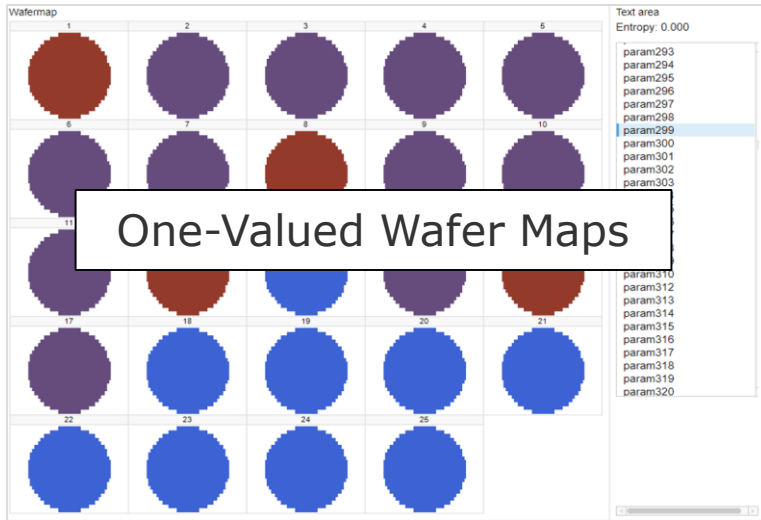


# Cross-Sectional Anomaly Analysis

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

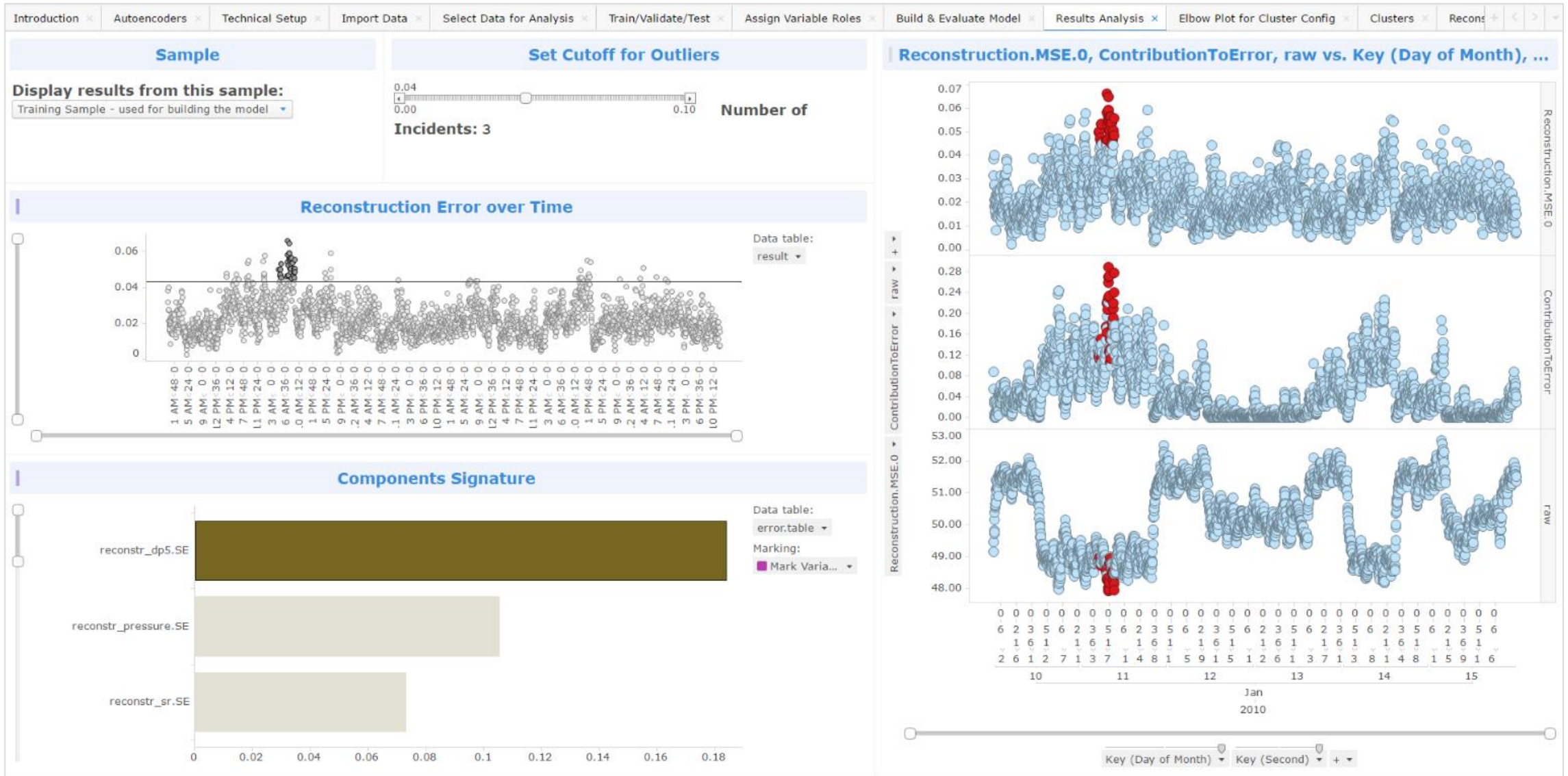


# Spatial Anomaly Analysis

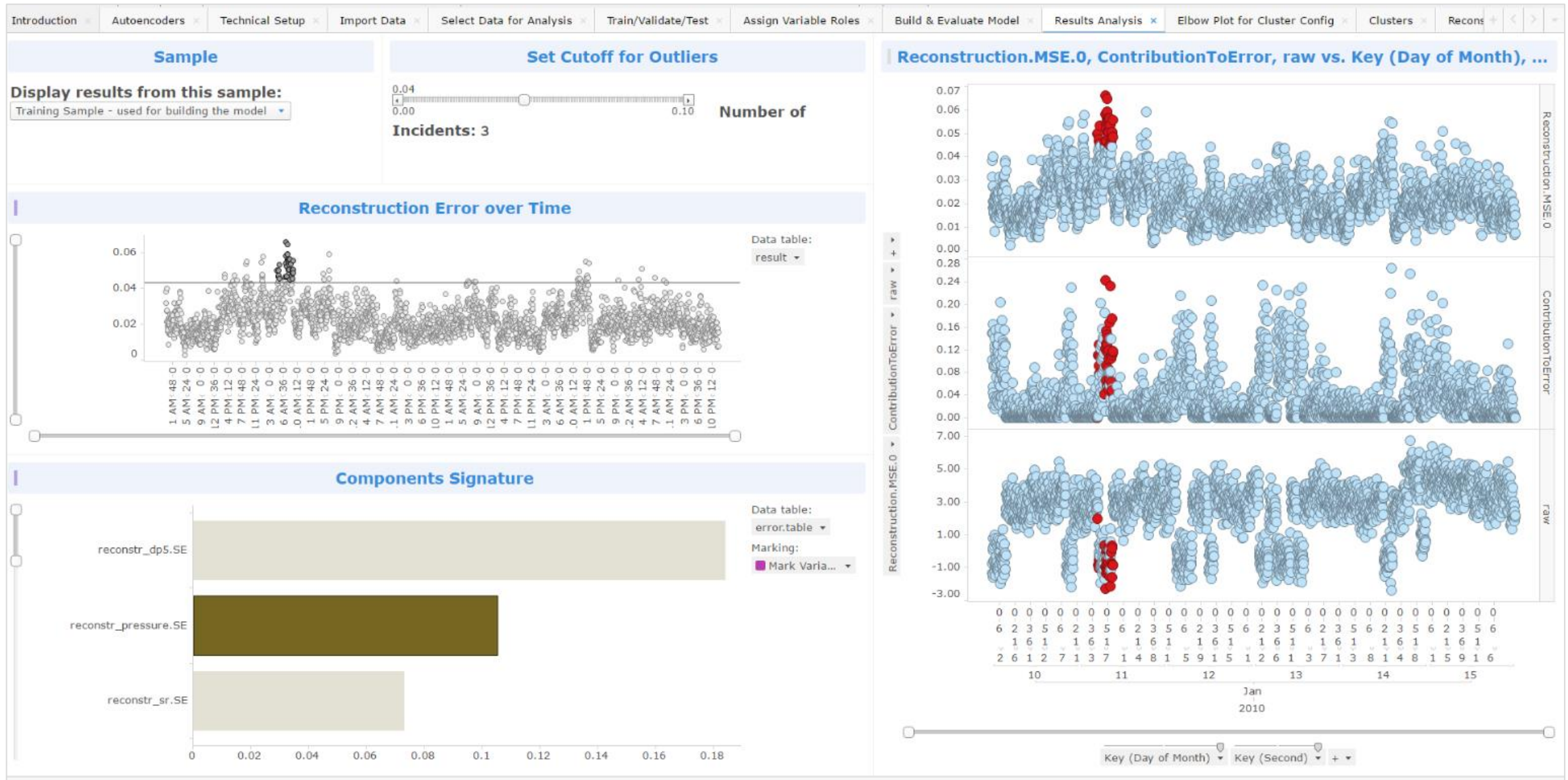




# Cross-Sectional Anomaly Detection



# Cross-Sectional Anomaly Detection



# Analysis: Cluster Incidents, View Signatures



# 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]

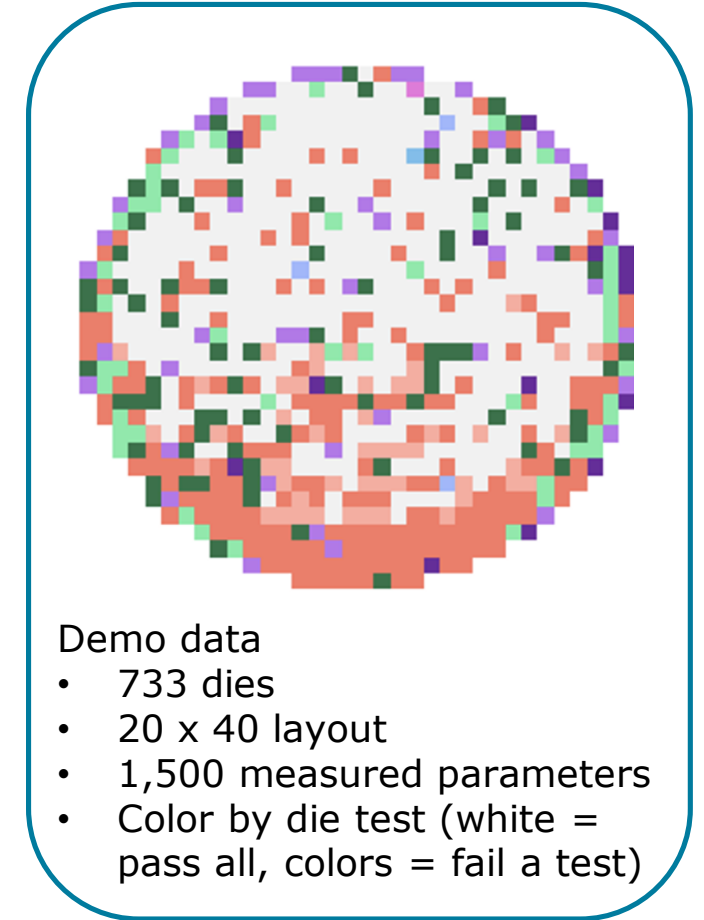
- Stream wafer data
- Vectorize and cluster

## 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*





# TIBCO® Streaming



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



(TIBCO Streaming operator for Python)

Perform SVD on the bin values for wafers using Python

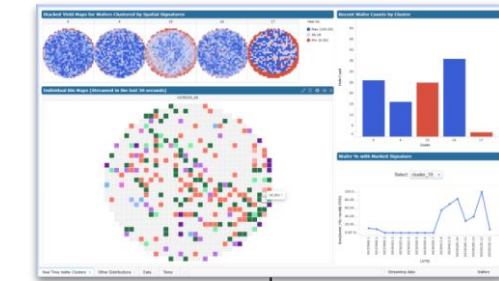


Combine data & results

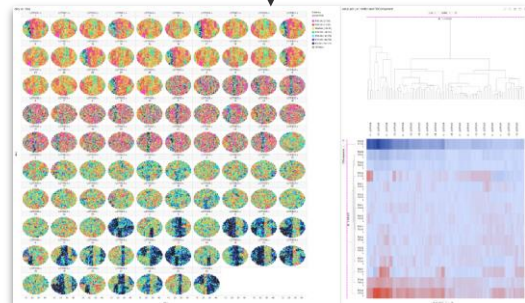


PMML operator to cluster the data using a model trained in TIBCO Data Science

Send data for live viewing into Spotfire Data Streams

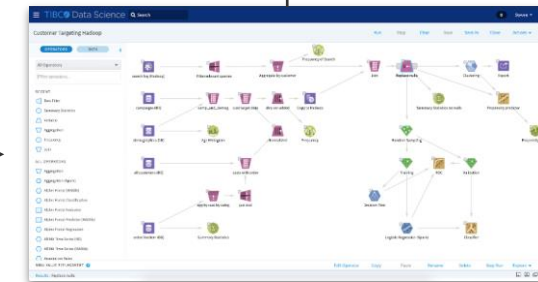


Refine clusters, identify wafers of interest



Send data to Spark/Hadoop cluster for more in-depth analysis in TIBCO Data Science

Retrain models for clustering and root-cause analysis



TIBCO™ Spotfire®

TIBCO® Data Science

# Anomaly Detection with Spatial Signature Analysis

# Anomaly Analysis



Find and explore anomalous events



Monitor anomalies



Predict when and why they occur

# Anomaly Detection and Analysis



Find and explore anomalous events

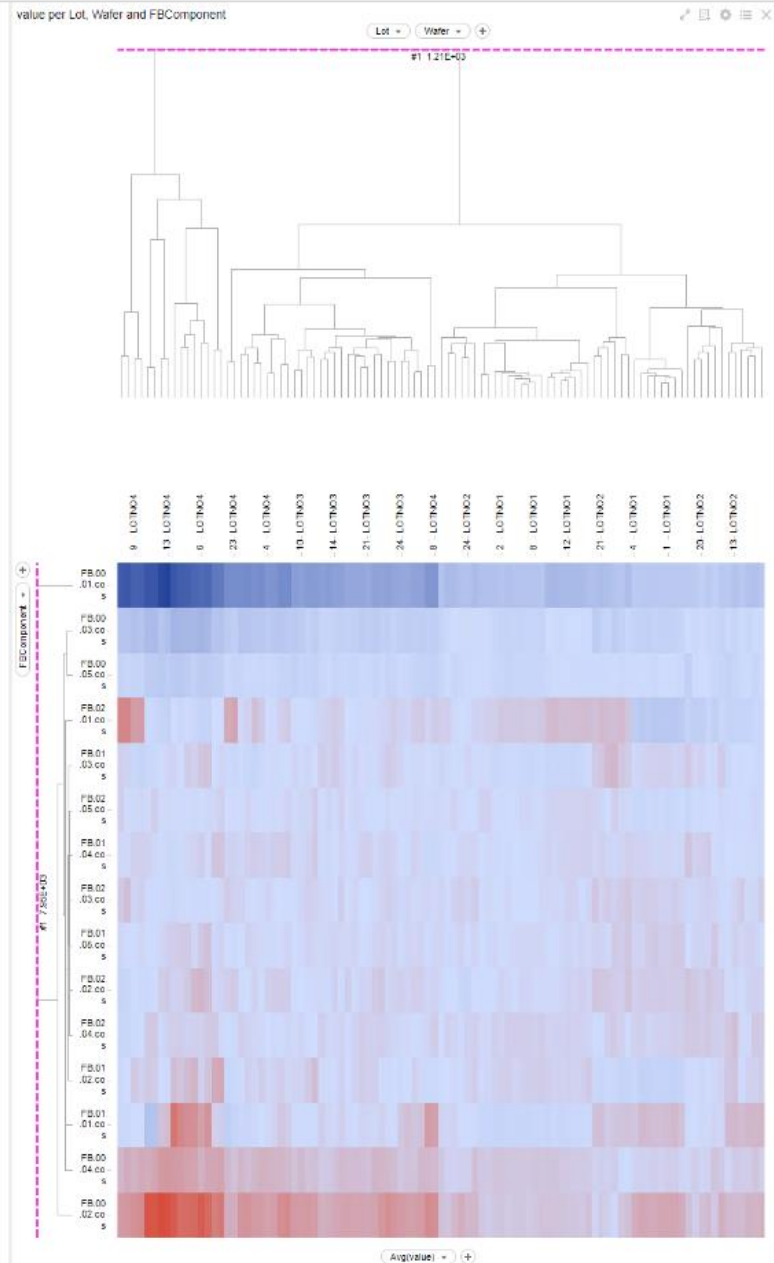
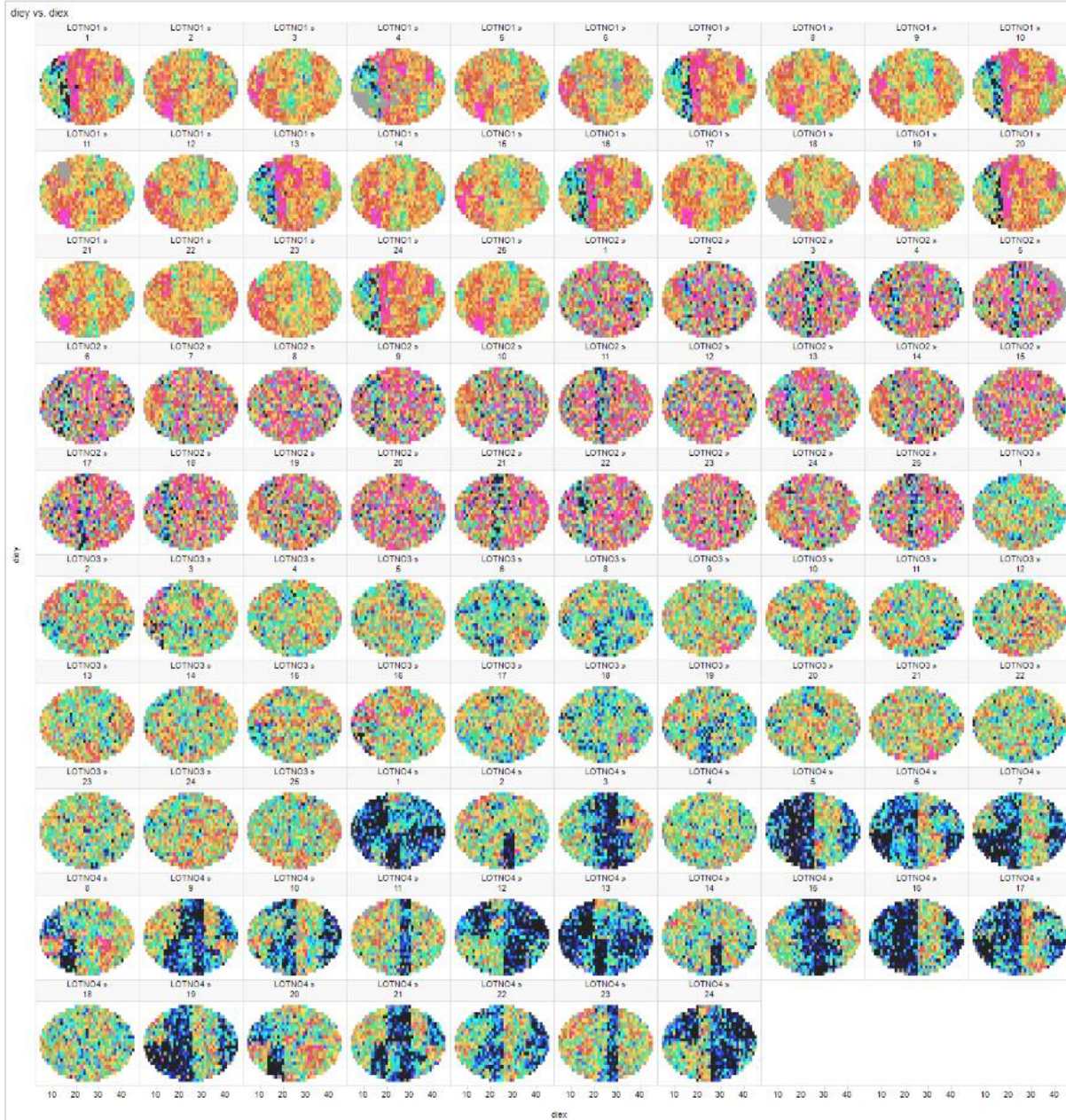


Monitor anomalies



Predict when and why they occur



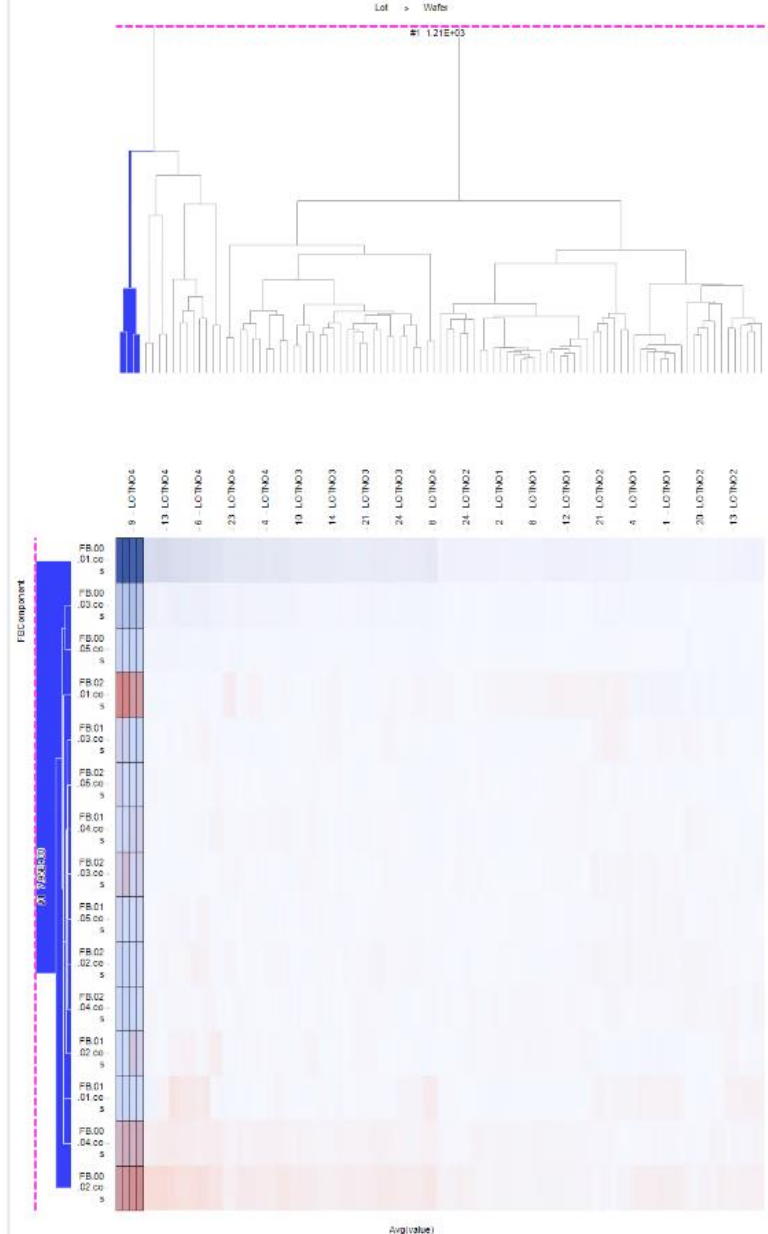


diey vs. diex



- 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)

value per Lot, Wafer and FBComponent



# Anomaly Detection and Analysis



Find and explore anomalous events

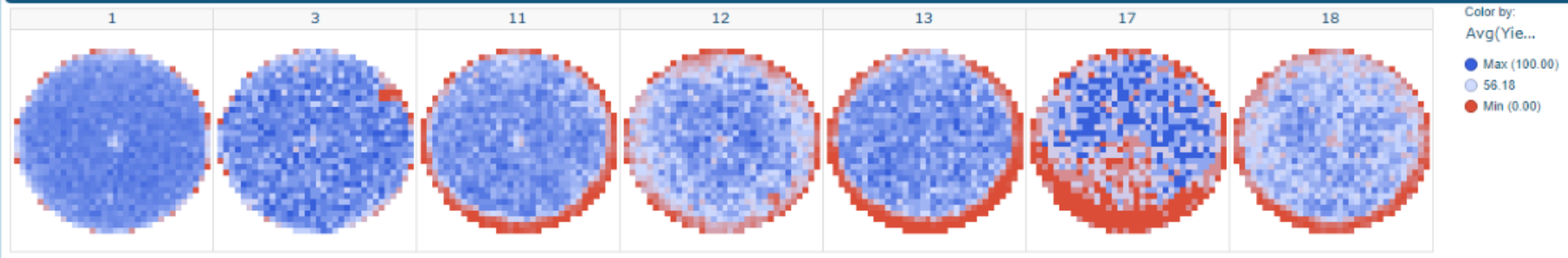


Monitor anomalies

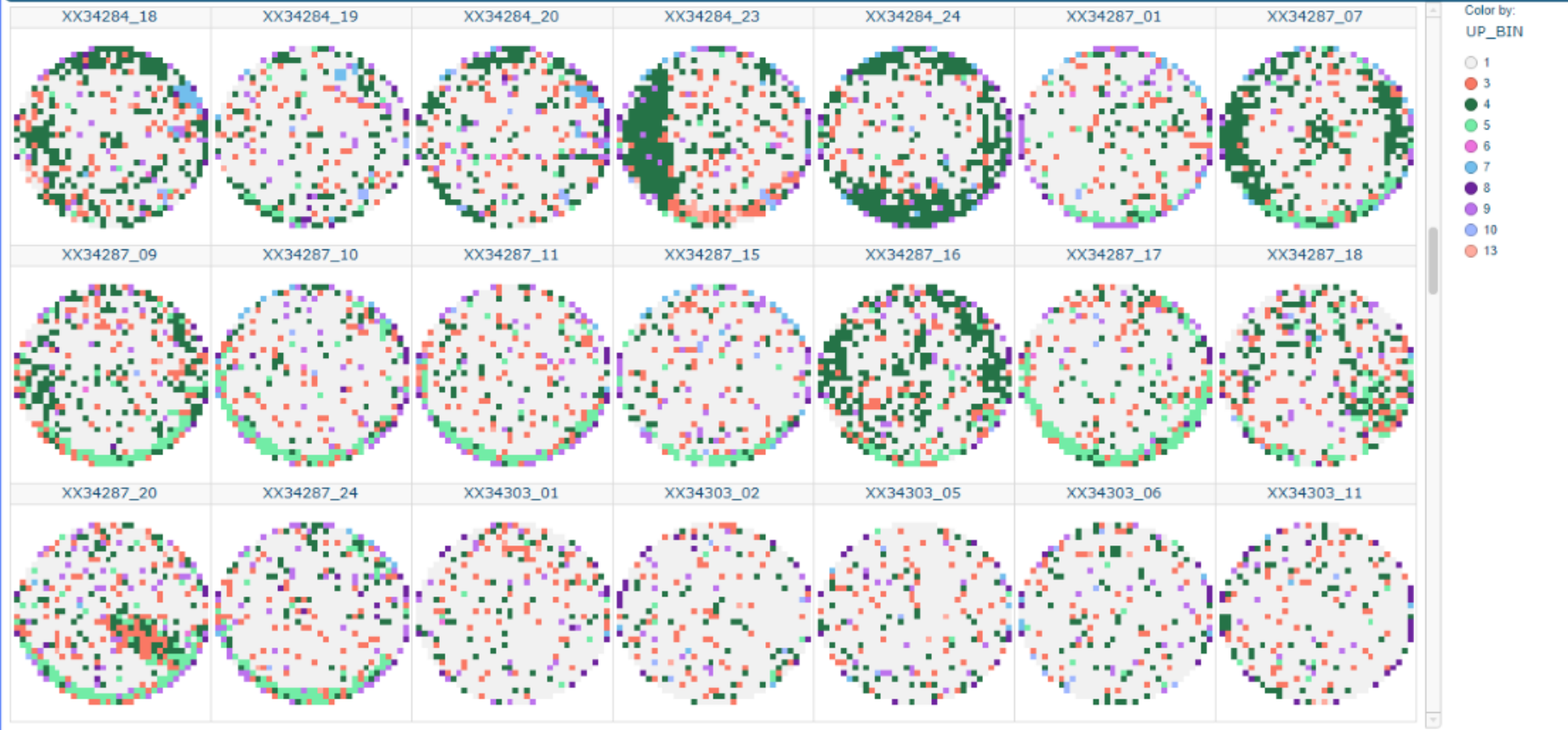


Predict when and why they occur

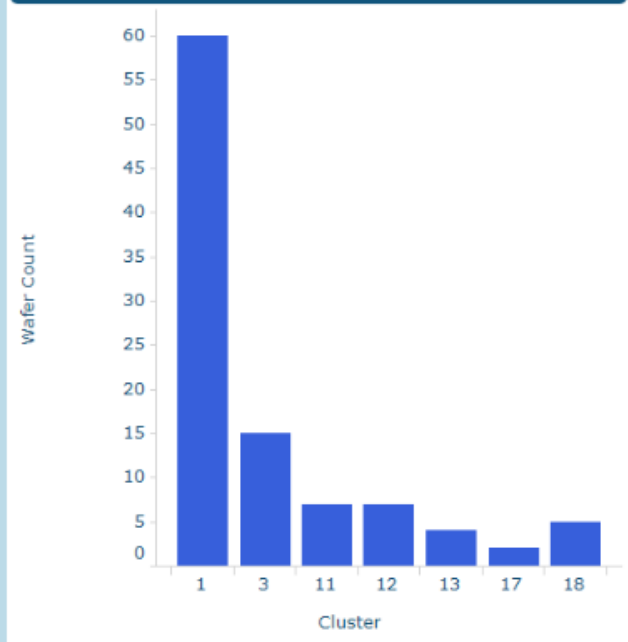
### Stacked Yield Maps for Wafers Clustered by Spatial Signatures



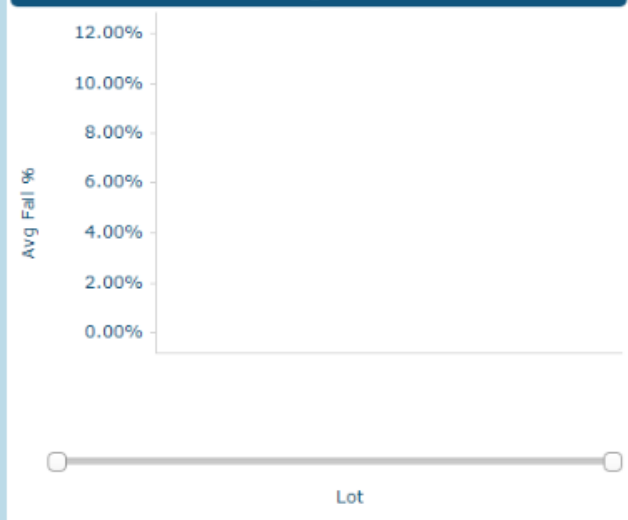
### Individual Bin Maps



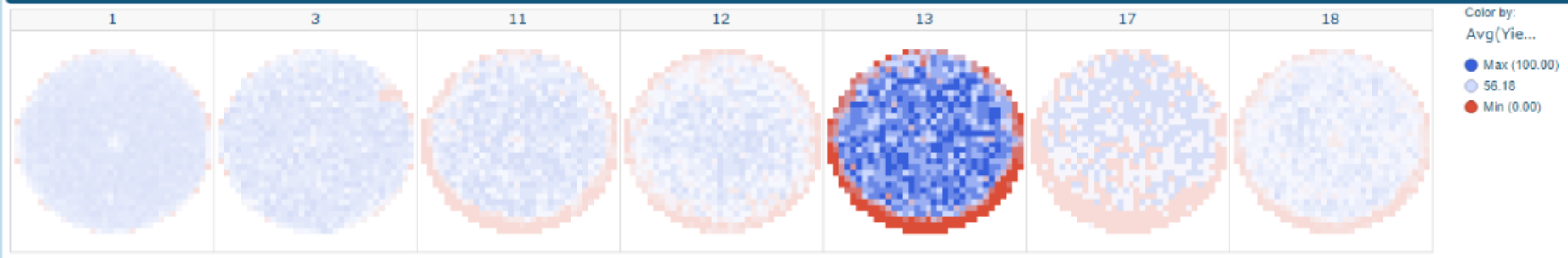
### Recent Wafer Counts by Cluster



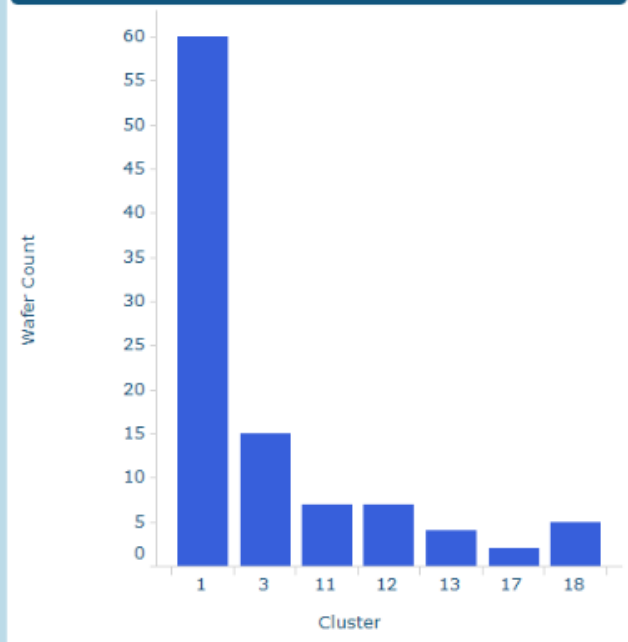
### Wafer % with marked Signature



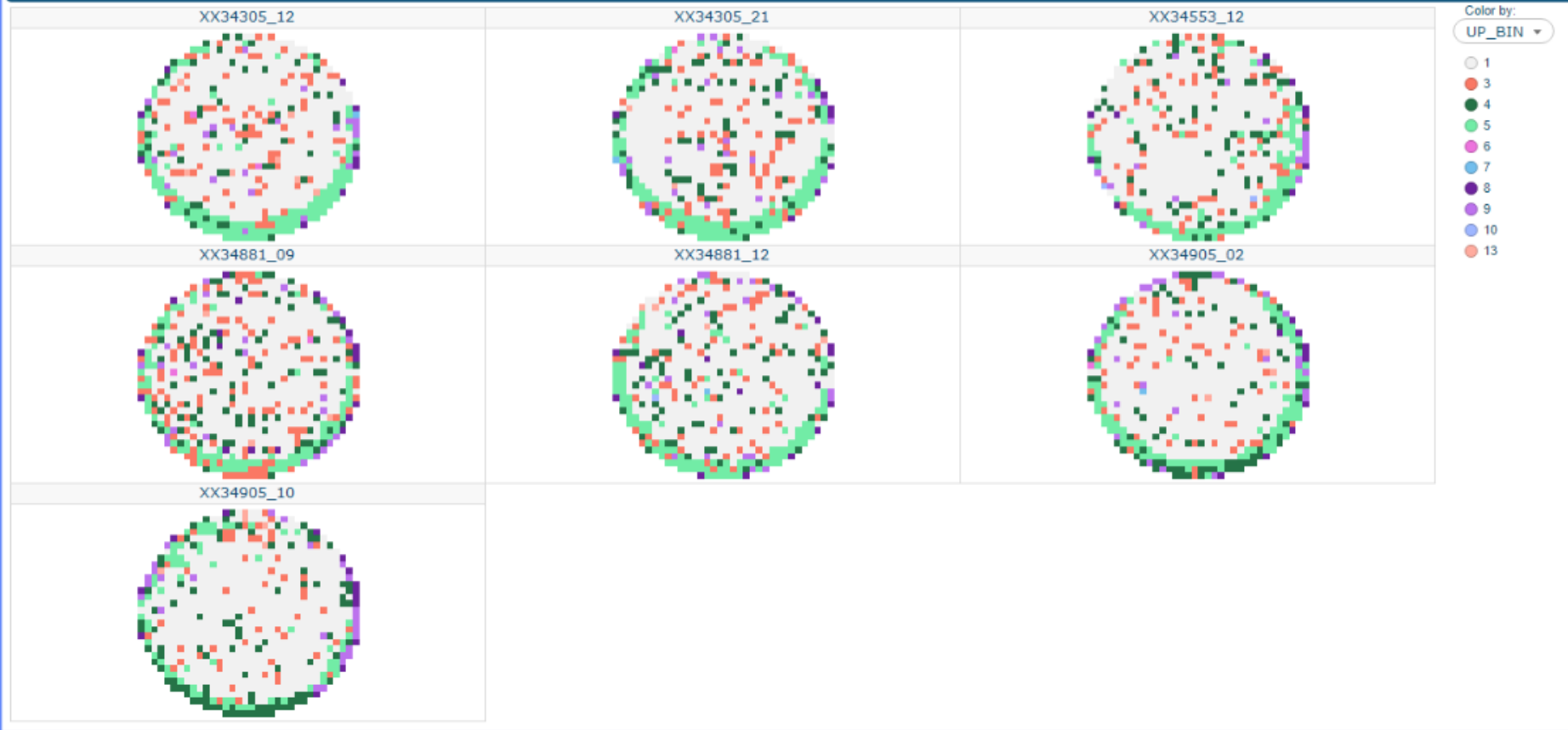
### Stacked Yield Maps for Wafers Clustered by Spatial Signatures



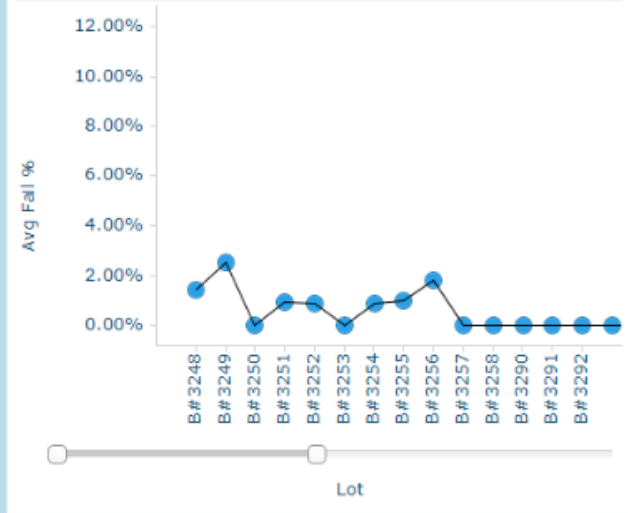
### Recent Wafer Counts by Cluster



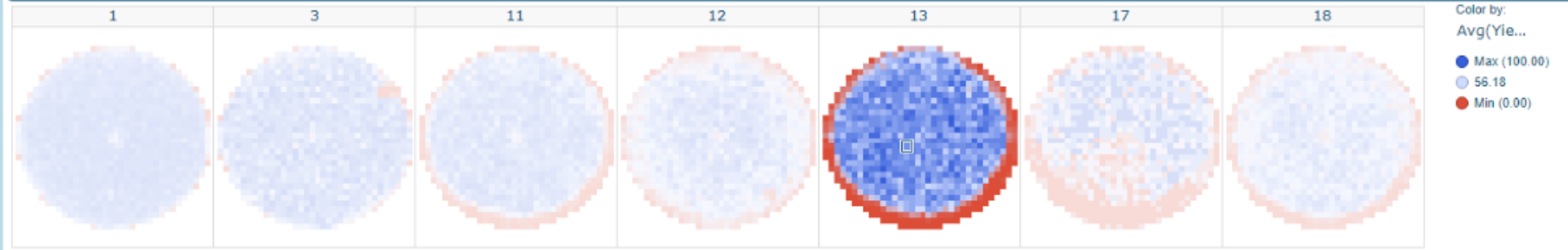
### Individual Bin Maps



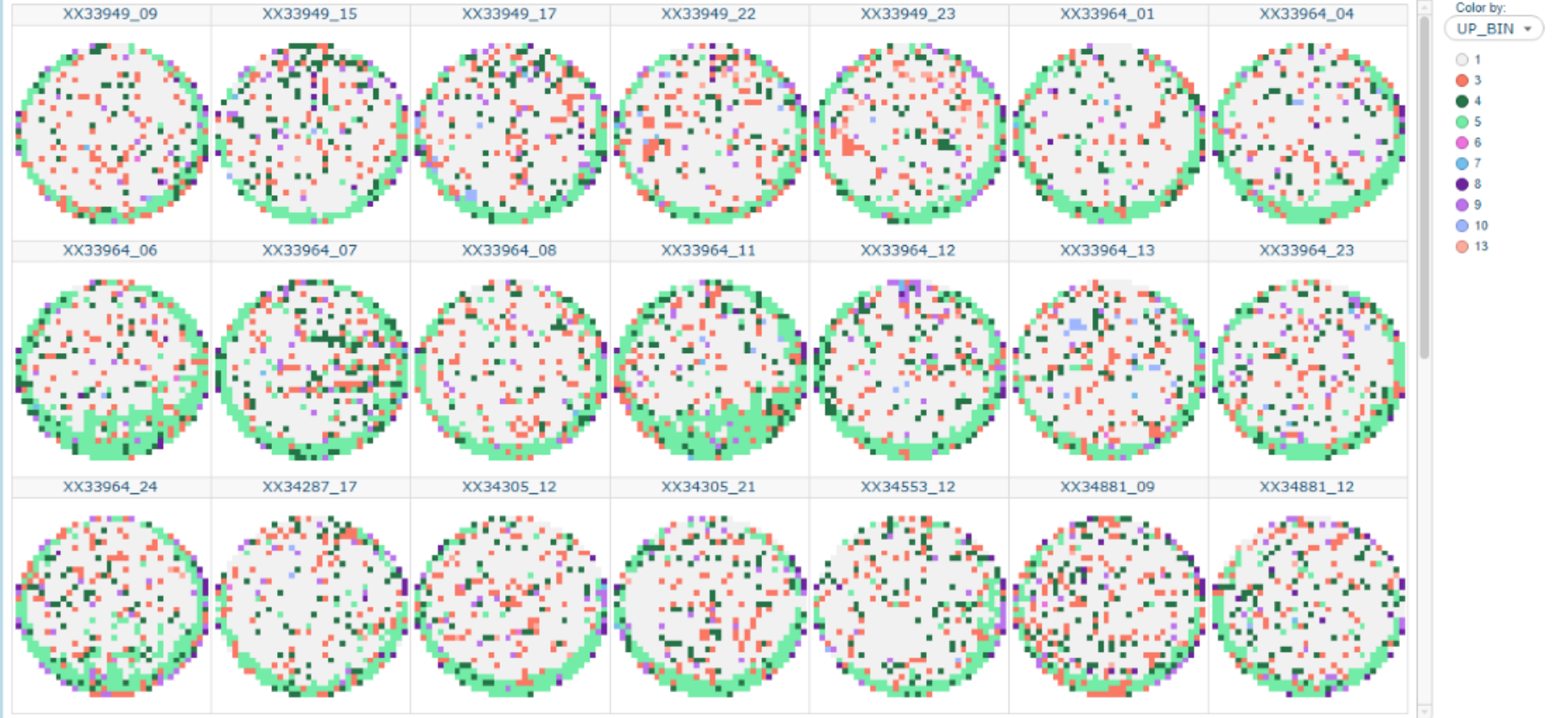
### Wafer % with marked Signature



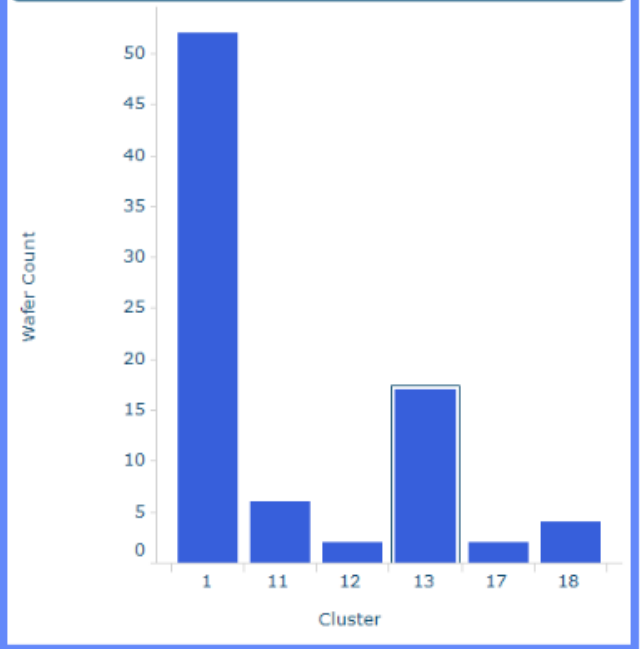
### Stacked Yield Maps for Wafers Clustered by Spatial Signatures



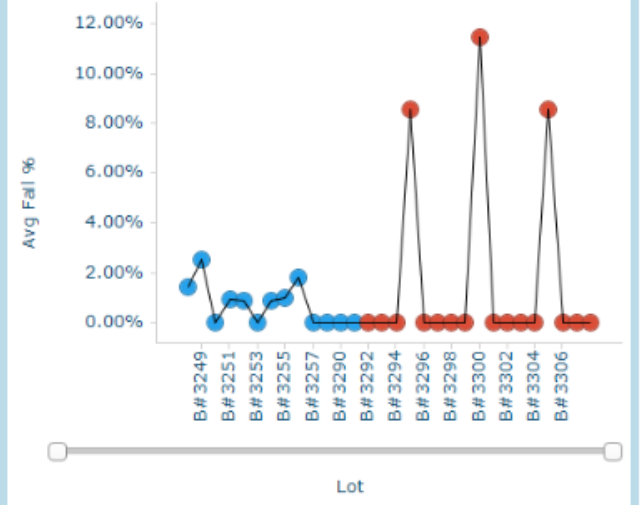
### Individual Bin Maps



### Recent Wafer Counts by Cluster



### Wafer % with marked Signature



# Anomaly Detection and Analysis



Find and explore anomalous events



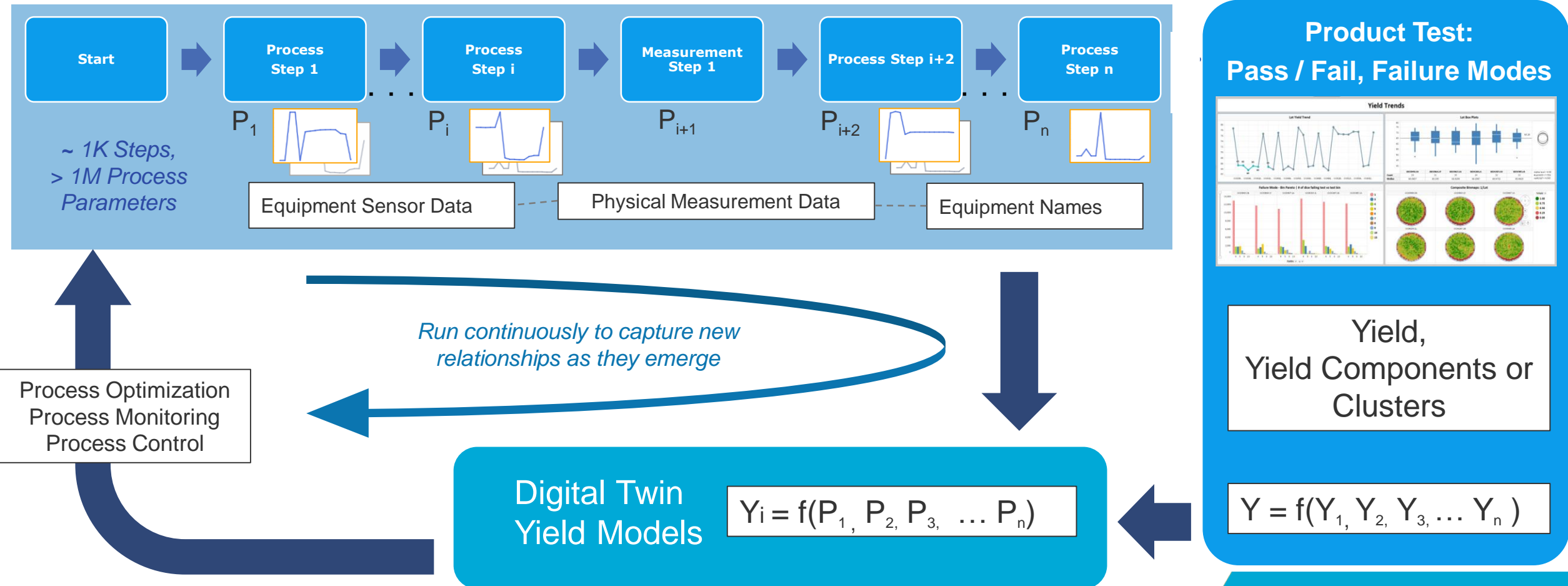
Monitor anomalies



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 ( $P_i$ )



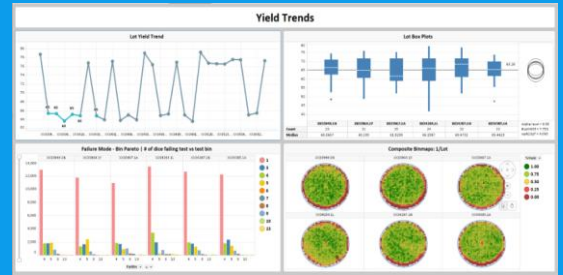
Process Optimization  
Process Monitoring  
Process Control

*Run continuously to capture new relationships as they emerge*

Digital Twin  
Yield Models

$$Y_i = f(P_1, P_2, P_3, \dots P_n)$$

**Product Test:**  
Pass / Fail, Failure Modes



Yield,  
Yield Components or  
Clusters

$$Y = f(Y_1, Y_2, Y_3, \dots Y_n)$$



# The Extreme Challenge of Big & Wide Data

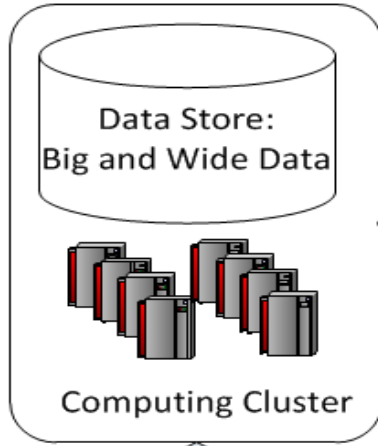
- **Not just big data** – many rows: lots, wafers, die, units
- **Also wide data** – many columns: > 1M process parameters
  - Sensor traces
    - Time series for every sensor on each machine in each run
  - Physical measurements
    - Film thickness, critical dimensions, layer-to-layer overlay, defect 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.”

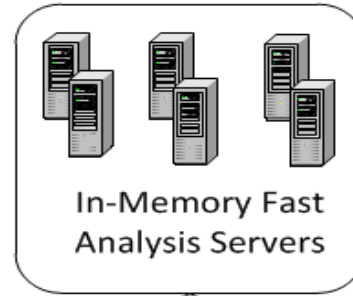
—Ann Kellehere, VP of the Technology and Manufacturing Group, Intel

# Solution Architecture

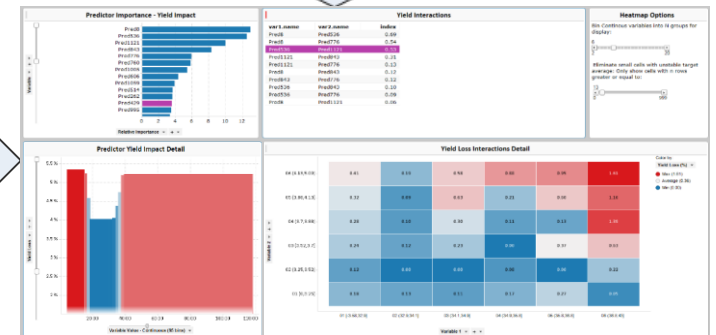
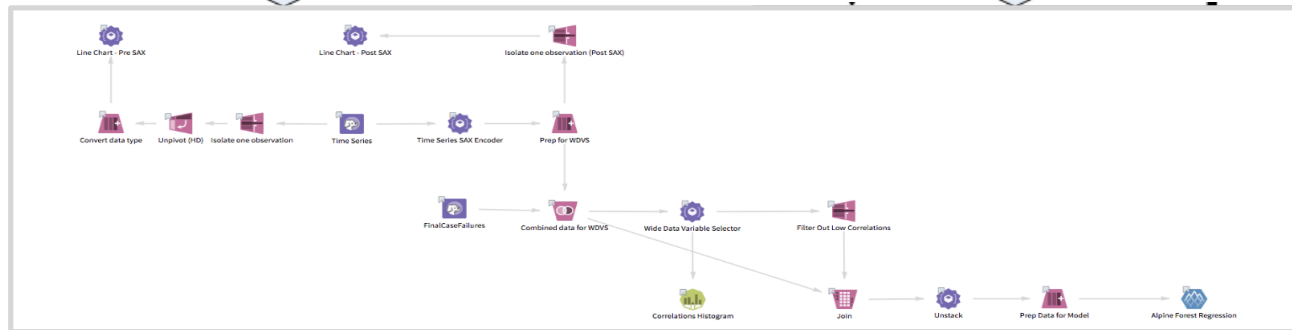
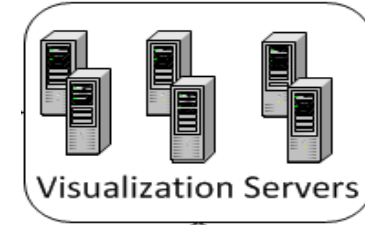
## Data Prep, Feature Engineering & Selection



## 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 <sup>1</sup> (minutes)					
	<b>20 Sensors</b> (1K variables)	<b>200 Sensors</b> (10K variables)	<b>2K Sensors</b> (100K variables)	<b>20K Sensors</b> (1M variables)	<i>Dataset Size for 1M Variables</i>
<b>1K Wafers</b>	0.47	0.48	0.72	1.0	25 GB
<b>10K Wafers</b>	0.50	0.53	0.77	1.75	253 GB
<b>100K Wafers</b>	0.53	0.67	1.25	15.15	2,530 GB

## <sup>1</sup>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

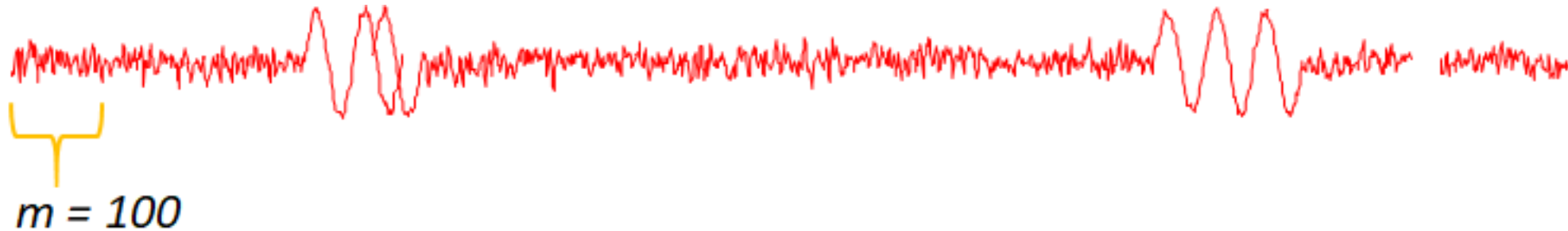
# Longitudinal Anomaly Analysis

## *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:

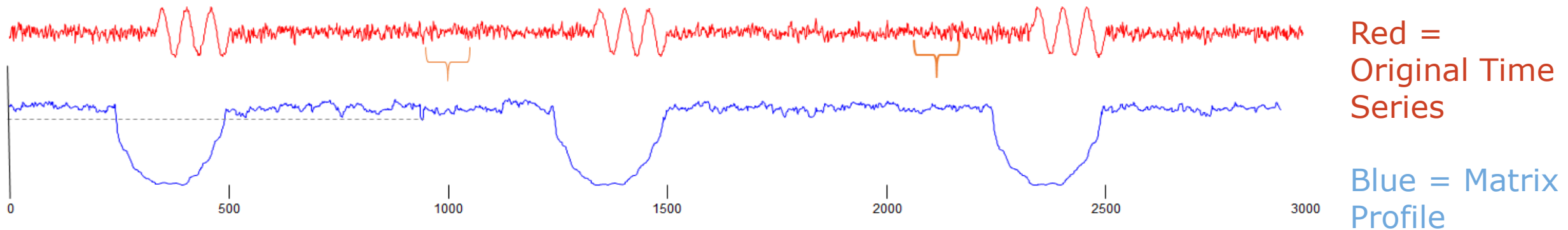
[https://www.cs.ucr.edu/~eamonn/matrix\\_profile\\_i.pptx](https://www.cs.ucr.edu/~eamonn/matrix_profile_i.pptx)

# 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



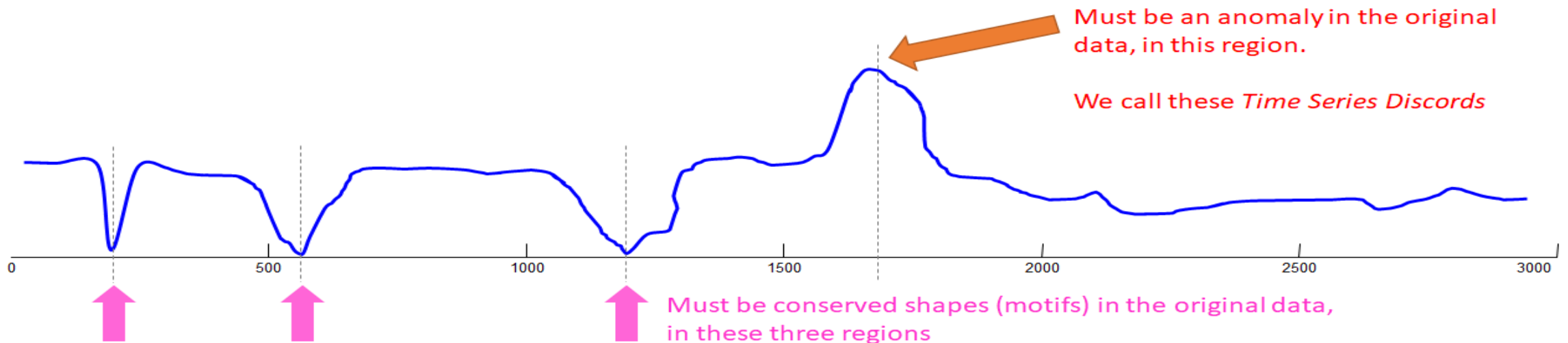
Material adapted from:

[https://www.cs.ucr.edu/~eamonn/matrix\\_profile\\_i.pptx](https://www.cs.ucr.edu/~eamonn/matrix_profile_i.pptx)

# How to “Read” a Matrix Profile

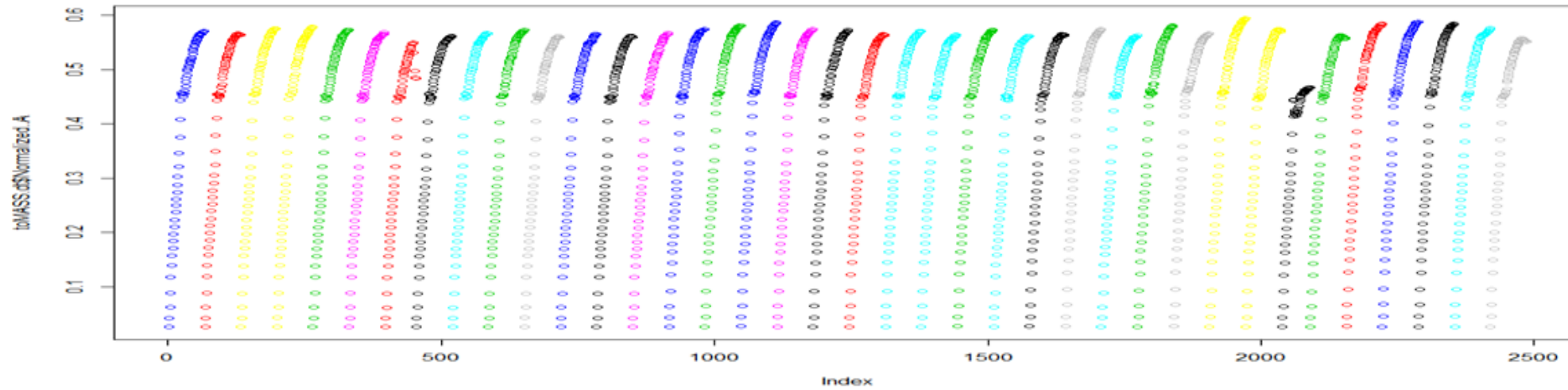
Where you see **relatively low values**, you know that the subsequence in the original time series must have (at least one) relatively similar subsequence elsewhere in the data (such regions are “motifs” or reoccurring patterns)

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)

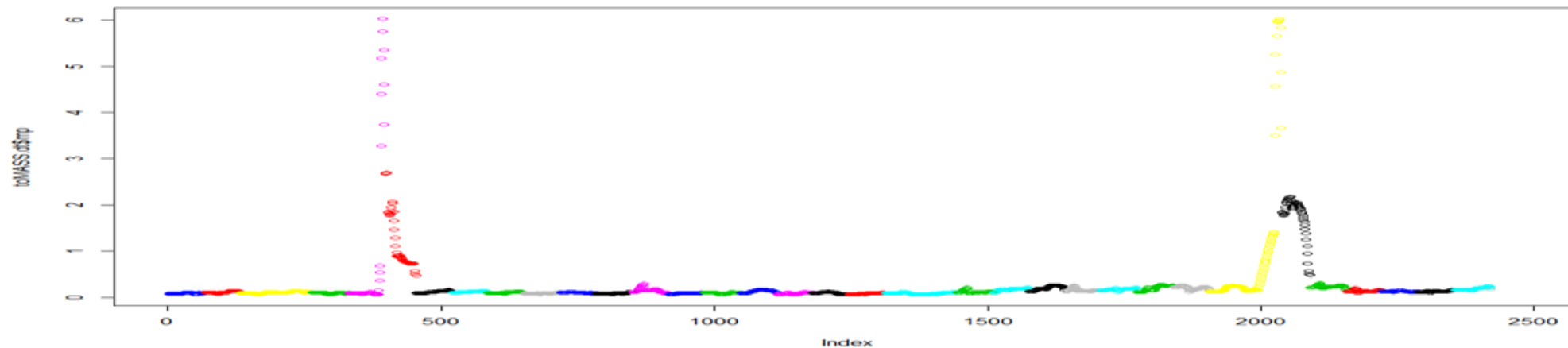


# Manufacturing Batches

Raw Amperage - Each color delimits a batch



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





# Community

# TIBCO® Exchange

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

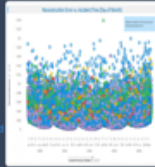
<https://community.tibco.com/exchange>

**Anomaly Detection Template for TIBCO Spotfire®**

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

• Flag as Inappropriate  
#Spotfire Templates #Machine Learning #Manufacturing Industry #Customer Analytics #Financial Services #Fraud #Unsupervised

★★★★★ 0 Reviews



**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.

• Flag as Inappropriate  
#Big data #Backup #Advanced Analytics #Apache Spark #Data function

★★★★★ 1 Review

[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, PL, 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.

• Flag as Inappropriate  
#IoT #Accelerator

★★★★★ 0 Reviews

[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.

• Flag as Inappropriate  
#Data function #Machine Learning #Supervised Learning #Manufacturing Industry #Customer Analytics #Fraud #Financial Services

★★★★★ 0 Reviews



**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 to 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.

• Flag as Inappropriate  
#Energy Industry #Manufacturing Industry #Energy #Industry Control

★★★★★ 0 Reviews

[Statistica Data Function](#) [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.

• Flag as Inappropriate  
#Accelerator #Live Datafeed

★★★★★ 0 Reviews

[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.

• Flag as Inappropriate  
#Machine Learning #Statistical Templates #Manufacturing Industry #Customer Analytics #Financial Services

★★★★★ 0 Reviews



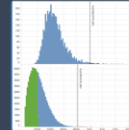
**Loss Distribution Approach to Operational Risk - Analysis Template for TIBCO Spotfire®**

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.

• Flag as Inappropriate  
#Financial Services

★★★★★ 0 Reviews

[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.

• Flag as Inappropriate  
#Statistical #Accelerator #Live Datafeed

★★★★★ 0 Reviews

[Try Now](#) [Edit This Module](#) [Create New Module](#)



TIBCO® Spotfire®

TIBCO® Data Science

TIBCO® Streaming

# AI in Operations

*Cloud Starters, Accelerators, Analytic Apps  
Thoughtleader-Led Solutions*



**AI on Demand**  
Data Science in Operations

The image features a glowing lightbulb against a dark blue background with bokeh light effects. The text is positioned in the upper left quadrant.



Driving Customer Engagement

Click Here for Demo

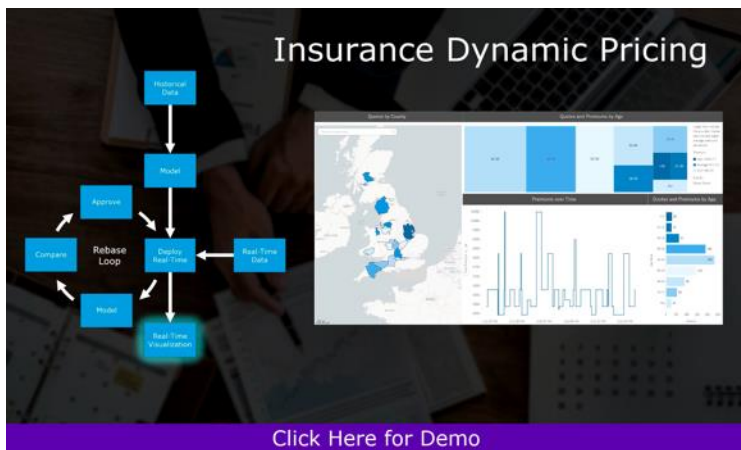
The image shows a man in a brown jacket looking at his smartphone. The text is in the upper right, and the demo link is at the bottom.



Fraud and Risk Management

Click Here for Demo

The image shows a close-up of a credit card chip. The text is in the upper right, and the demo link is at the bottom.



Insurance Dynamic Pricing

Click Here for Demo

The image displays a complex flowchart and a data dashboard. The flowchart includes boxes for 'Historical Data', 'Model', 'Rebase Loop', 'Digital Real Time', and 'Real Time'. The dashboard shows a map of the UK and various charts. The text is in the upper right, and the demo link is at the bottom.



Digital Twins in Manufacturing

Click Here for Demo

The image shows two stylized human heads composed of a network of lines and nodes, representing digital twins. The text is in the upper right, and the demo link is at the bottom.



The Industrial Internet - Production Surveillance

Click Here for Demo

The image shows silhouettes of two workers standing in front of an oil pumpjack against a sunset sky. The text is in the upper right, and the demo link is at the bottom.

# Questions & Contact

Steven Hillion  
Sr. Director, Data Science  
[shillion@tibco.com](mailto:shillion@tibco.com)  
@StevenHillion

Michael O'Connell  
Chief Analytics Officer  
[moconnell@tibco.com](mailto:moconnell@tibco.com)  
@MichOConnell

[TIBCO Community](https://community.tibco.com)

community.tibco.com

[TIBCO Exchange](https://community.tibco.com/exchange)

community.tibco.com/exchange

[TIBCO Tech Blog](https://community.tibco.com/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.