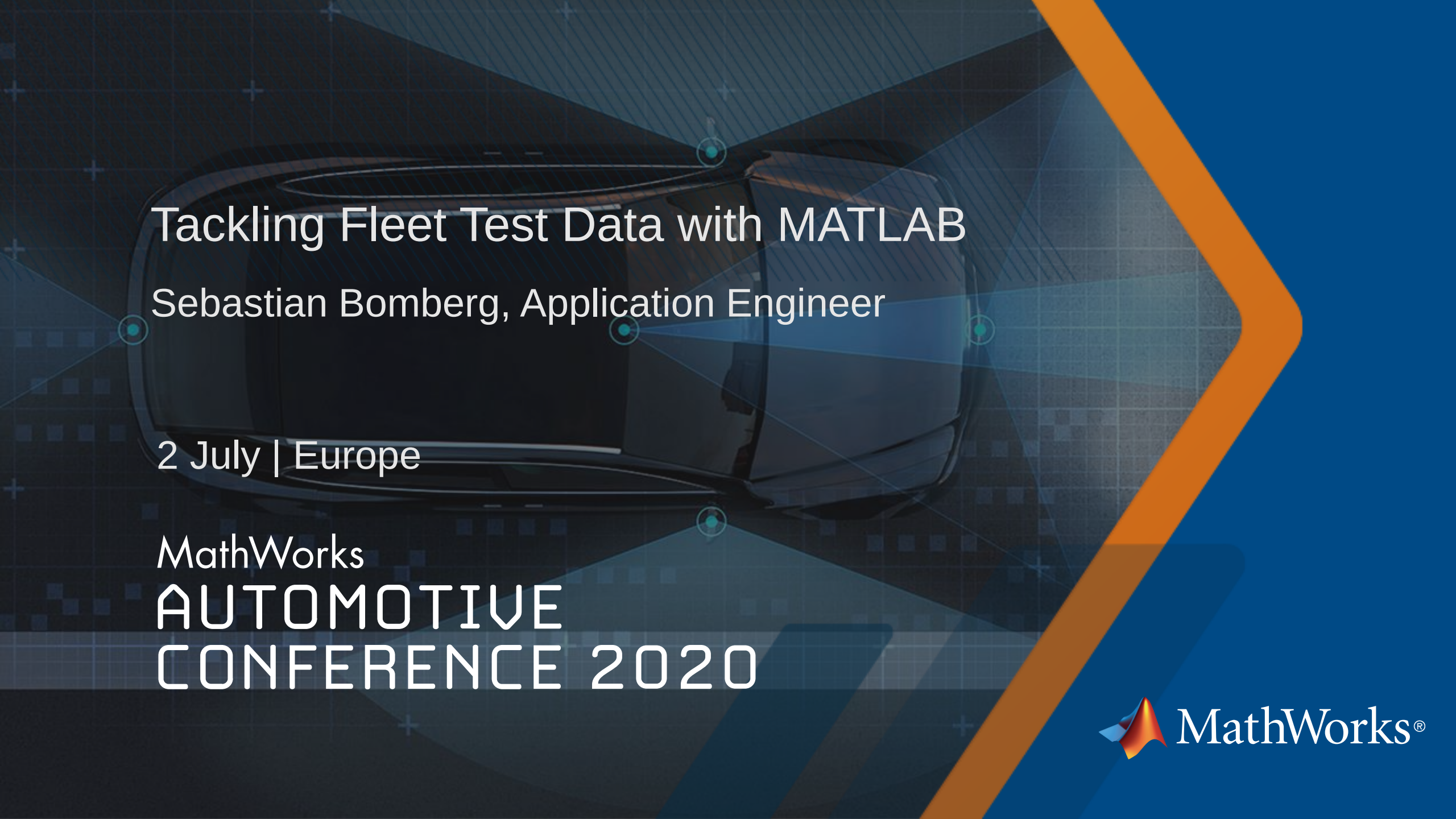


Tackling Fleet Test Data with MATLAB

Sebastian Bomberg, Application Engineer

2 July | Europe

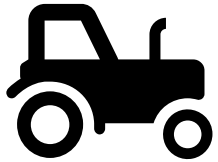
MathWorks
AUTOMOTIVE
CONFERENCE 2020

MathWorks®

A Fleet is a Collection of Resources that Generate Data You Want to Act on

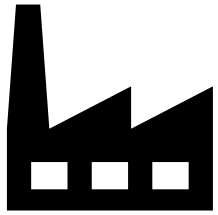
Automotive

- Vehicles
- Engines
- Controllers



Agriculture

- Harvesters
- Tractors
- Mining



Manufacturing

- Pick & Place machines
- Welding robots
- Material handling systems

Infrastructure

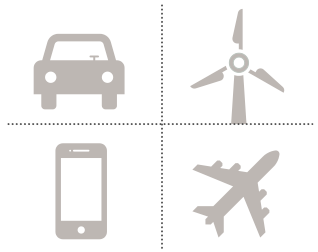
- Charging stations
- Parking spaces
- Electronic toll collection

Benefits of Fleet Data Analytics

- Faster time to insights (design, testing)
- Warranty / Transparency / Reputation
- New business opportunities in the Mobility Age

Understanding Enables:

Better Design



Good Reputation

New Business Opportunities

Fleet Analytics in Practice: Volkswagen Data Lab

Develop technology building block for tailoring car features and services to individual

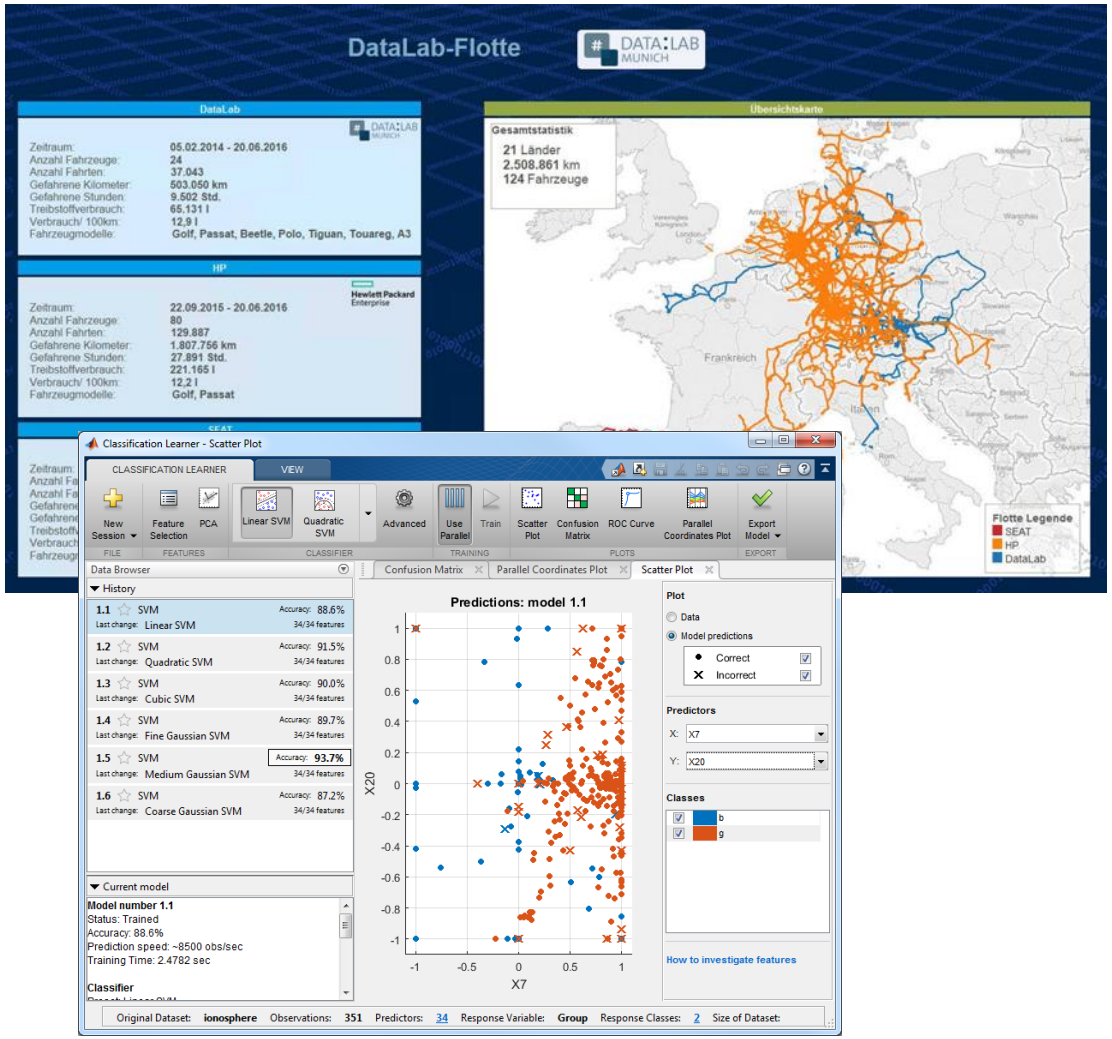
- Driver and Fleet Safety
- Driver Coaching
- Driver-Specific Insurance

Data sources

- Logged CAN bus data and travel record

Results

- Proof-of-concept model for “telematic fingerprint”
- Basis for the “pay-as-you-drive” concept



Source: “Connected Car – Fahrererkennung mit MATLAB“
Julia Fumbarev, Volkswagen Data Lab
MATLAB EXPO Germany, June 27, 2017, Munich Germany

MathWorks®

is a **Leader** in the Gartner Magic Quadrant for 2020 Data Science and Machine Learning Platforms

Figure 1. Magic Quadrant for Data Science and Machine Learning Platforms

Source: Gartner (February 2020)

*Gartner Magic Quadrant for Data Science and Machine Learning Platforms, Peter Krensky, Erick Brethenoux, Jim Hare, Carlie Idoine, Alexander Linden, Svetlana Sicular, 11 February 2020 .

This graphic was published by Gartner, Inc. as part of a larger research document and should be evaluated in the context of the entire document. The Gartner document is available upon request from MathWorks. Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, express or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

Data Analytics Workflow

Access and Explore Data

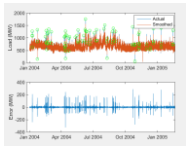
Files

Databases

Sensors

Preprocess Data

Working with Messy Data



Data Reduction/ Transformation

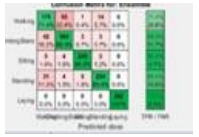
Feature Extraction

Develop Predictive Models

Model Creation e.g. Machine Learning

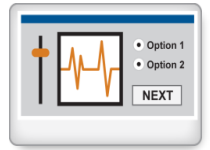
Parameter Optimization

Model Validation



Integrate Analytics with Systems

Desktop Apps



Enterprise Scale Systems

MATLAB Excel .NET .exe C/C++ Java .dll

Embedded Devices and Hardware

Data Analytics Workflow

Access and Explore Data

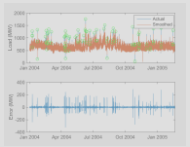
Files

Databases

Sensors

Preprocess Data

Working with Messy Data



Data Reduction/ Transformation

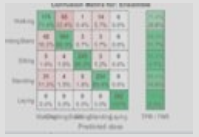
Feature Extraction

Develop Predictive Models

Model Creation e.g. Machine Learning

Parameter Optimization

Model Validation



Integrate Analytics with Systems

Desktop Apps

Enterprise Scale Systems

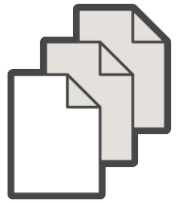
MATLAB Excel .NET C/C++ .exe Java .dll

Embedded Devices and Hardware

Extract, **T**ransform, and **L**oad (ETL) Workflow Considerations

Extract

Raw Files

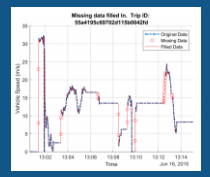


File Types

MDF (.MF4)
.blf
.parquet
.mat
custom binary
text

Transform

Pre-Process & Clean



Data Manipulation

- Unit Conversion
- Bit mapping
- Unpacking

Enforce Schema

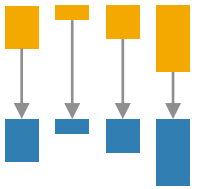
Load

Write File Set

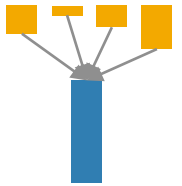
.parquet
.mat
text

Write Mode

One-to-one



Many-to-fewer



Access Data in Many Formats From Many Locations

- Type
- Structure
- Location

Access Data From Anywhere With Minimal Changes


```
setenv("AWS_ACCESS_KEY_ID",id)
```

```
setenv("HADOOP_HOME",hadoopPath)
fileLoc = "hdfs://hadoop01glnxa64:123/datasets/FoodPic.jpg"
```

```
img = imread(fileLoc);
```

Scale to Large Collections of Data with Datastore

Create a datastore from all CSV files

```
ds = datastore('*.csv')
```

Read a single file of data

```
data = read(ds);
```

Reset the datastore back to the first file

```
reset(ds);
```

Find the maximum value of “Y” in each file

```
X = [];
while hasdata(ds)
        data = read(ds);
        X(end+1) = max(data.Y);
end
```

Available Datastores	
General	`datastore`
	`spreadsheetDatastore`
	`tabularTextDatastore`
	`fileDatastore`
Database	`databaseDatastore`
Image	`imageDatastore`
	`denoisingImageDatastore`
	`randomPatchExtractionDatastore`
	`pixelLabelDatastore`
	`augmentedImageDatastore`
Audio	`audioDatastore`
Predictive Maintenance	`fileEnsembleDatastore`
	`simulationEnsembleDatastore`
Simulink	`SimulationDatastore`
Automotive	`mdfDatastore`
Custom	subclass `matlab.io.Datastore`
Transformed	`transform` an existing datastore

Data Analytics Workflow

Access and Explore Data

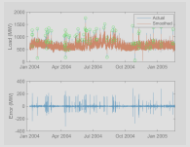
Files

Databases

Sensors

Preprocess Data

Working with Messy Data



Data Reduction/ Transformation

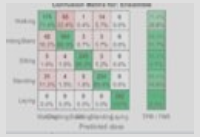
Feature Extraction

Develop Predictive Models

Model Creation e.g. Machine Learning

Parameter Optimization

Model Validation



Integrate Analytics with Systems

Desktop Apps

Enterprise Scale Systems

Embedded Devices and Hardware

Data Analytics Workflow

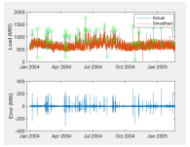
Access and Explore Data → **Preprocess Data** → **Develop Predictive Models** → **Integrate Analytics with Systems**

Files

Databases

Sensors

Working with Messy Data



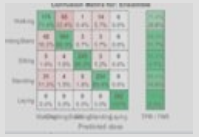
Data Reduction/ Transformation

Feature Extraction

Model Creation e.g. Machine Learning

Parameter Optimization

Model Validation



Desktop Apps

Enterprise Scale Systems

Embedded Devices and Hardware

Preprocess and Explore Data in Few Lines of Code

Import

```
t1 = readtimetable("s3://bucket_name/file.txt");
```

Preprocess

```
t = synchronize(t1,t2,t3);
t = fillmissing(t,"linear");
t = rmoutliers(t);
t = smoothdata(t,"movmedian");
t = normalize(t);
```

Explore

```
top5 = topkrows(t,5,"RH");
byTime = groupsummary(t,"Time","year","mean");
scaled = grouptransform(t,"State","rescale");
chgpts = ischange(t,"variance","Threshold",20);
```

Visualize

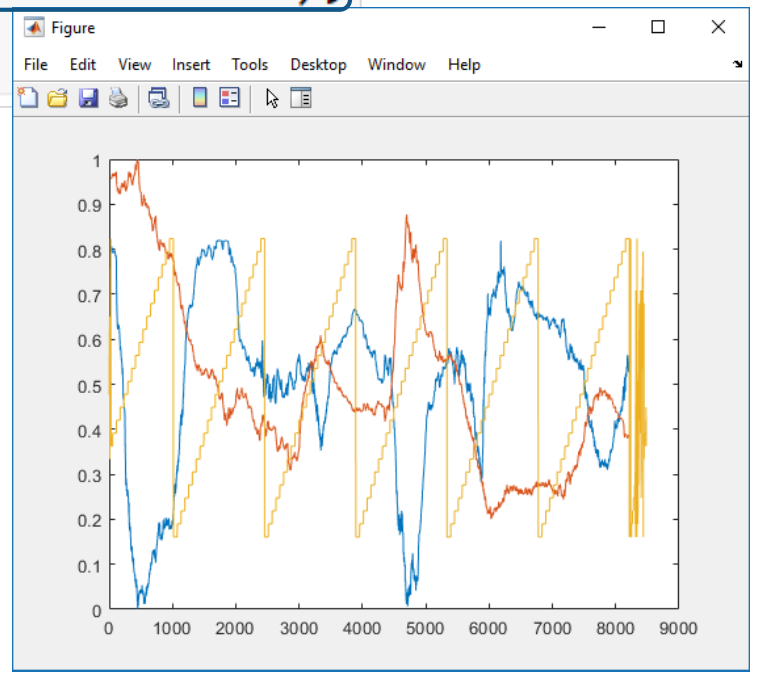
```
stackedplot(t);
geoplot(t.Lat,t.Lon,t.RH);
heatmap(t,"State","AQILabel");
scatterhistogram(t.RH,t.DP);
```

Use Dedicated Functions for Common Preprocessing Tasks

- Synchronize by time
- Find, fill, and remove missing
- Work with outliers
- Smooth noisy data
- Normalize, rescale data

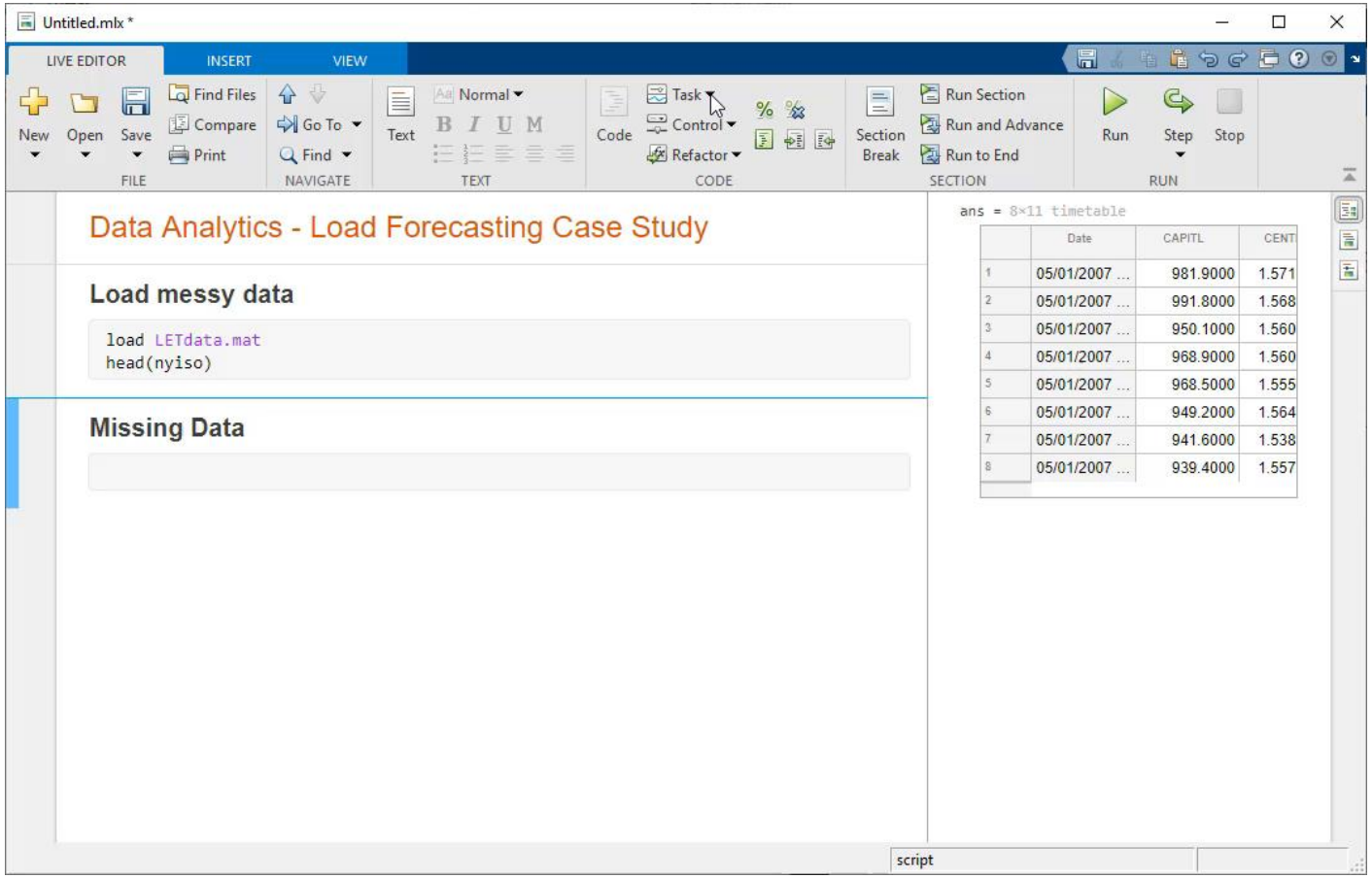
Preprocess

```
t = synchronize(t1,t2,t3);
t = fillmissing(t,"linear");
t = rmoutliers(t);
t = smoothdata(t,"movmedian");
t = normalize(t);
```



Explore Settings Quickly With Live Tasks

- Tasks are apps that can be included in scripts
- Preprocessing tasks allow you to:
  - Interactively explore parameters and options
  - Preview results based on those parameters and options
- Automatically generate the corresponding MATLAB code
- Save the task as part of the script for subsequent use by others



Data Analytics Workflow

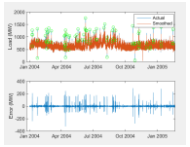
Access and Explore Data → **Preprocess Data** → **Develop Predictive Models** → **Integrate Analytics with Systems**

Files

Databases

Sensors

Working with Messy Data



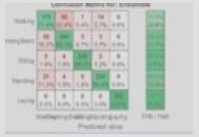
Data Reduction/ Transformation

Feature Extraction

Model Creation e.g. Machine Learning

Parameter Optimization

Model Validation



Desktop Apps

Enterprise Scale Systems

Embedded Devices and Hardware

Data Analytics Workflow

Access and Explore Data

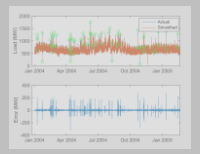
Files

Databases

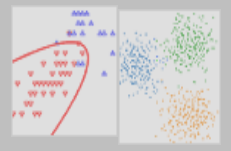
Sensors

Preprocess Data

Working with Messy Data



Data Reduction/ Transformation



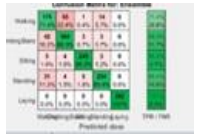
Feature Extraction

Develop Predictive Models

Model Creation e.g. Machine Learning

Parameter Optimization

Model Validation



Integrate Analytics with Systems

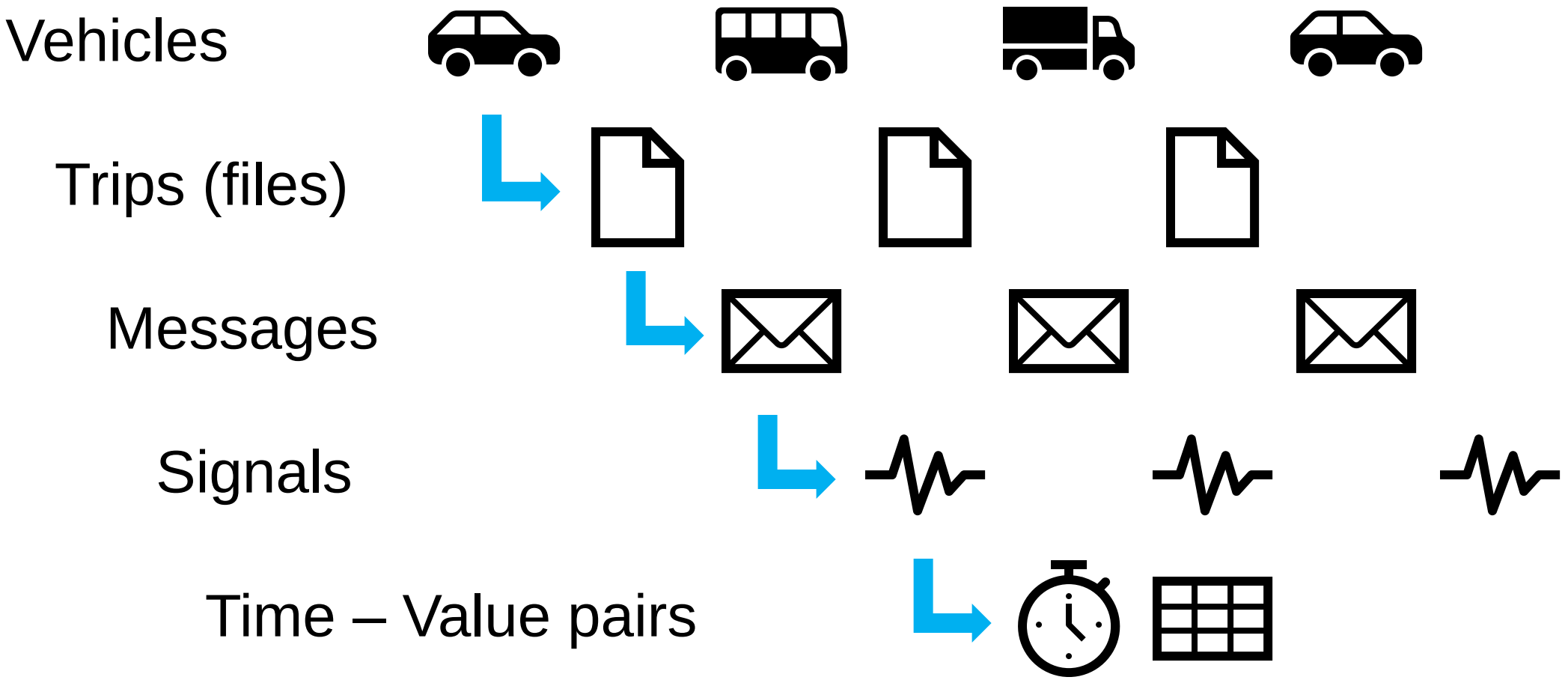
Desktop Apps

Enterprise Scale Systems

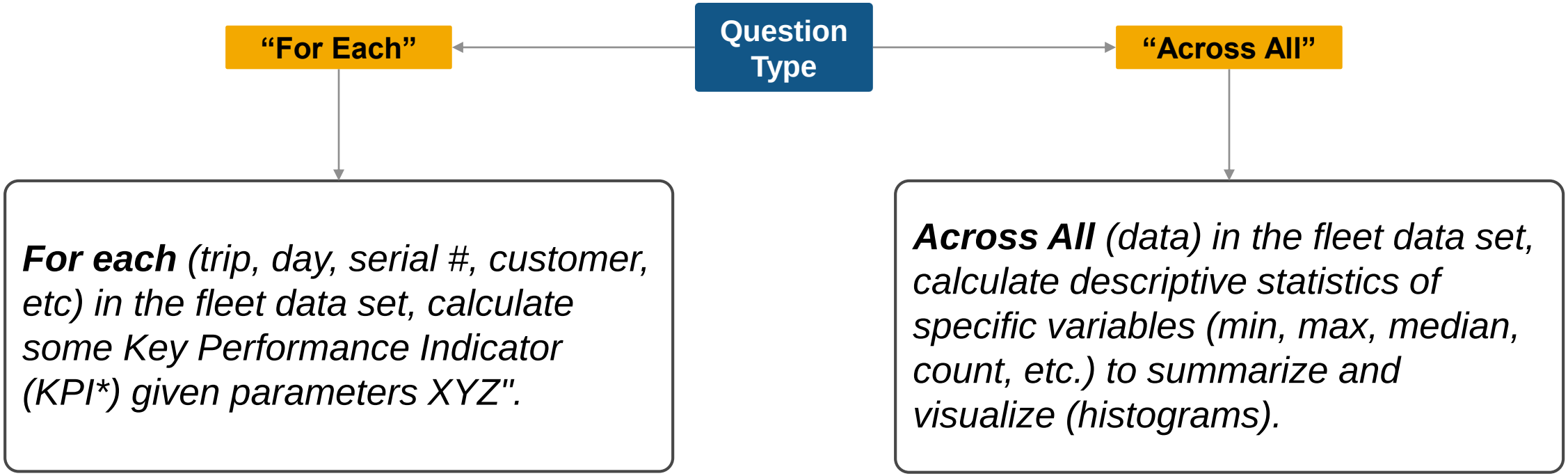
MATLAB Excel .NET C/C++ .exe Java .dll

Embedded Devices and Hardware

Automotive Vehicle Fleet – Intrinsic Hierarchy



Fleet Analytics Questions Fall Into Two Broad Categories



Scale Up to Big Data Without Big Changes

One file

Access Data

```
measured = readtable('PumpData.csv');
measured = table2timetable(measured);
```

Preprocess Data

Select data of interest

```
measured = measured(timerange(seconds(1),seconds(2)),'Speed')
```

Work with missing data

```
measured = fillmissing(measured,'linear');
```

Calculate statistics

```
m = mean(measured.Speed);
s = std(measured.Speed);
```

One hundred files

Access Data

```
measured = datastore('PumpData*.csv');
measured = tall(measured);
measured = table2timetable(measured);
```

Preprocess Data

Select data of interest

```
measured = measured(timerange(seconds(1),seconds(2)),'Speed')
```

Work with missing data

```
measured = fillmissing(measured,'linear');
```

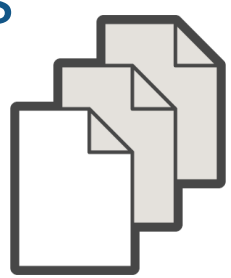
Calculate statistics

```
m = mean(measured.Speed);
s = std(measured.Speed);
```

```
[m,s] = gather(m,s);
```

Big Data File-Based Workflows

```
function events = detectEvents(t)

<fill in analytics code here>

end
```



File set location
(C:\, NAS/NFS, S3, Blob, hdfs://…)

Datastore

```
ds = parquetDatastore("hdfs://myData/Test")
```

Question Type

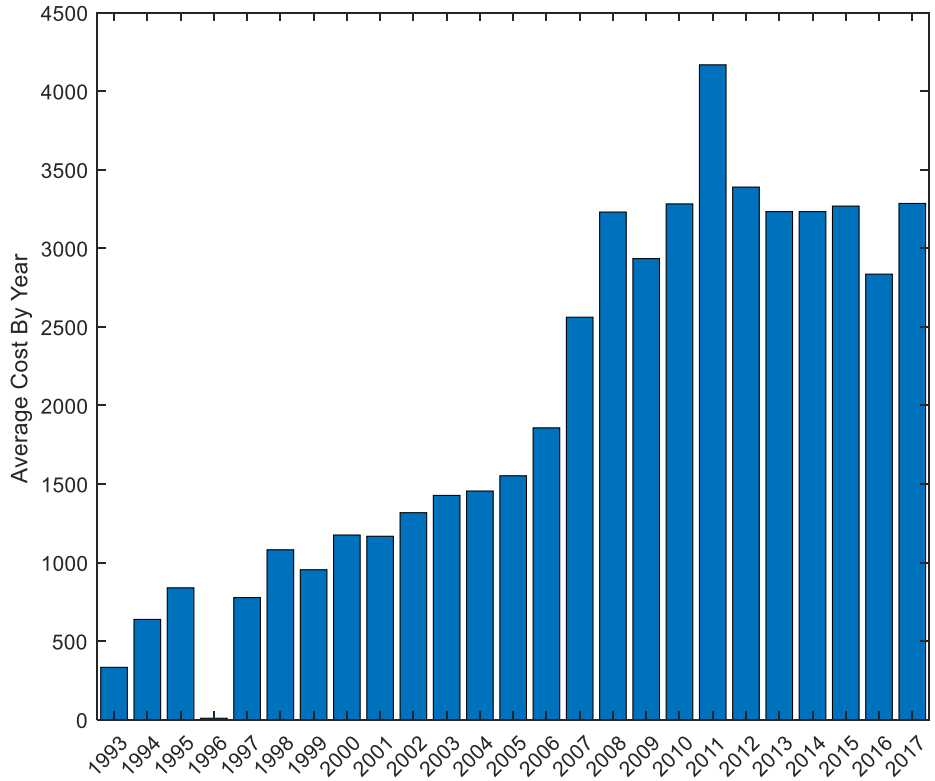
“For Each”

```
tds = transform(ds, @detectEvents);

TT = tall(tds);

EventsSummary = gather(TT);
```

Compute Groupwise Metrics and Detect Events

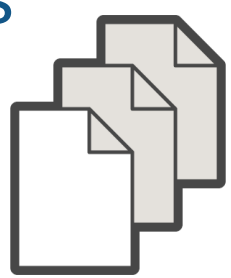
- Explore range
- Grouped calculations
- Detect local minima and maxima
- Detect abrupt changes in data with **`ischange`**

```
top5 = topkrows(t,5,"RH");
byTime = groupsummary(t,"Time","year","mean");
scaled = grouptransform(t,"State","rescale");
chgpts = ischange(t,"variance","Threshold",20);
```



Big Data File-Based Workflows

```
function events = detectEvents(t)

<fill in analytics code here>

end
```



File set location
(C:\, NAS/NFS, S3, Blob, hdfs://…)

Datastore

```
ds = parquetDatastore("hdfs://myData/Test")
```

Question Type

“For Each”

```
tds = transform(ds, @detectEvents);

TT = tall(tds);

EventsSummary = gather(TT);
```

“Across All”

```
TT = tall(ds);

MedianValSigABC = gather(median(TT.SignalABC));
```

Perform “Across All” Calculations with Tall Tables

Create a datastore from a collection of CSV files, and select the "Time" and "EngineSpeedRPM" variables.

```
ds = datastore('EngineData*.csv',...
    "SelectedVariableNames",["Time","EngineSpeedRPM"]);
```

Create tall table:

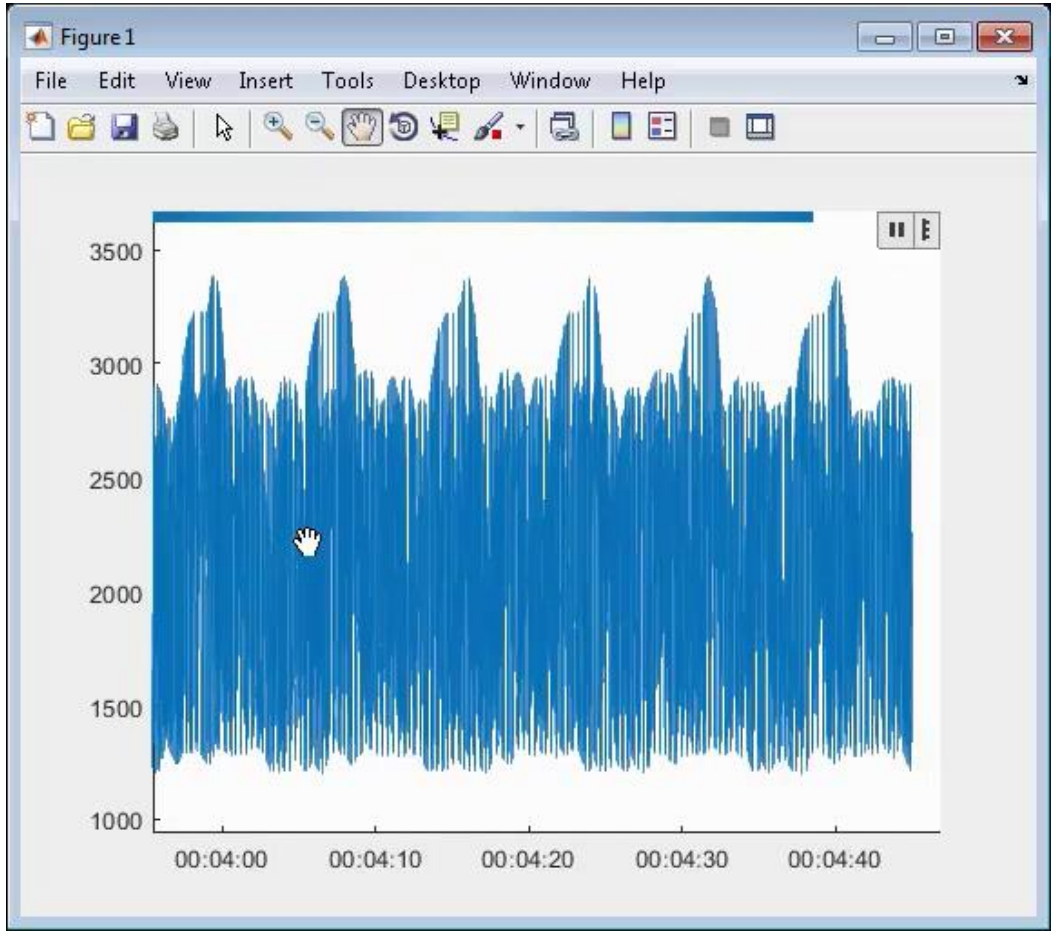
```
t = tall(ds);
```

Convert to tall timetable:

```
tt = table2timetable(t);
```

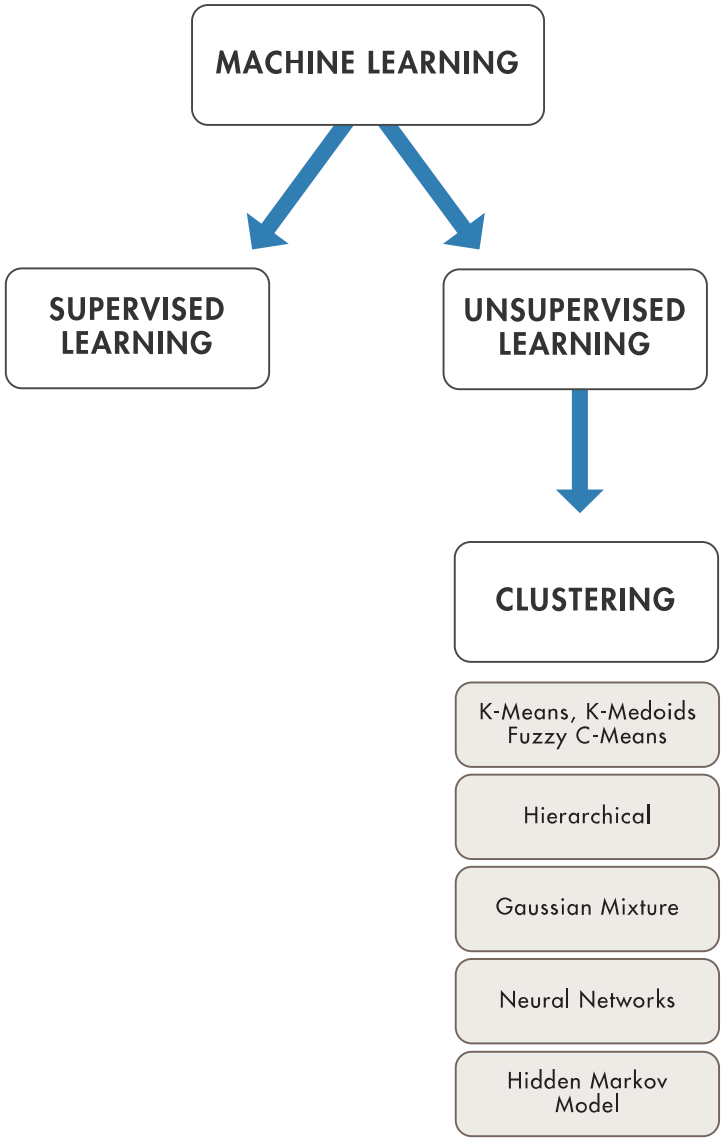
Plot EngineSpeedRPM vs. Time:

```
plot(tt.Time,tt.EngineSpeedRPM)
```



- Visualizations
- Data preprocessing
- Machine Learning

Explore Fleet Data with Unsupervised Learning



Unsupervised Learning for Operational Mode Clustering

Plot the raw data:

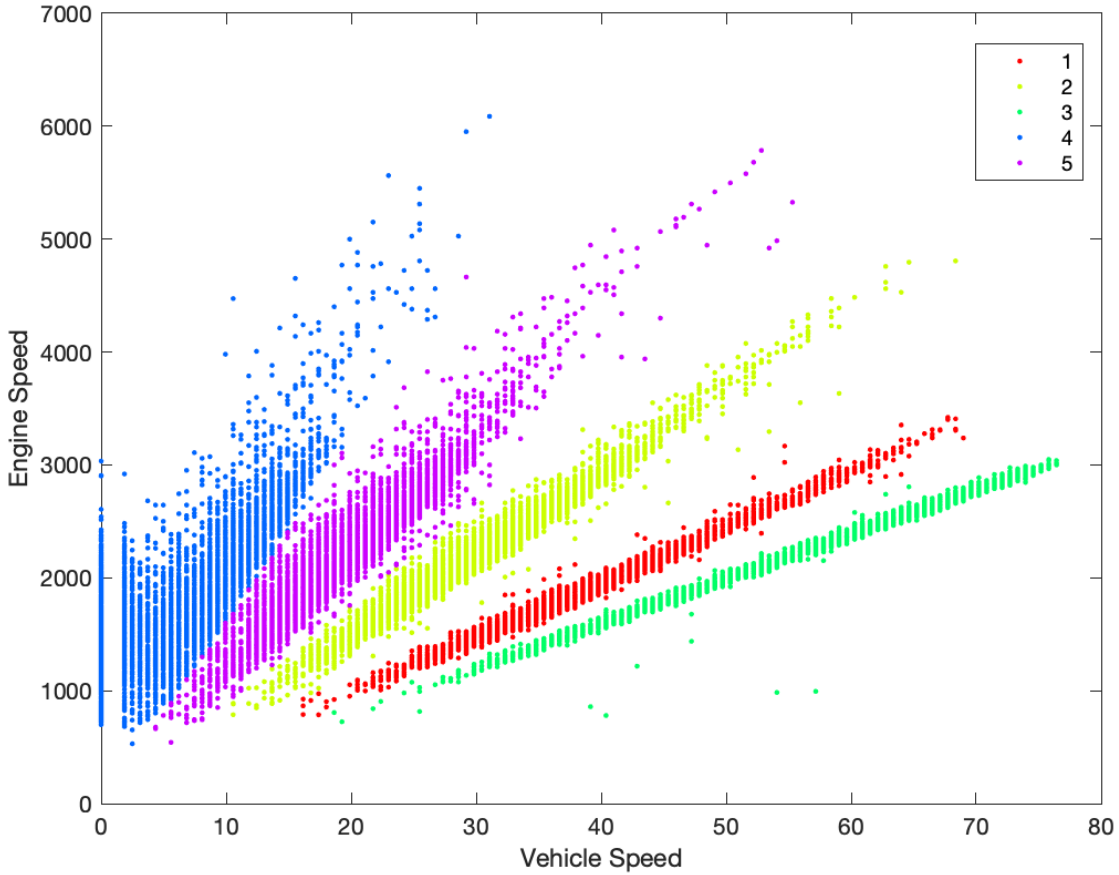
```
figure;
plot(t.Speed_OBD_,t.EngineRPM,'.k')
xlabel('Vehicle Speed');
ylabel('Engine Speed');
```

Cluster the data with the K-Means algorithm:

```
X = [t.Speed_OBD_,t.EngineRPM];
IDX = kmeans(X,5,"Distance","cosine");
```

Plot results of the clustering:

```
gscatter(t.Speed_OBD_,t.EngineRPM,IDX);
xlabel('Vehicle Speed');
ylabel('Engine Speed');
```



Data Analytics Workflow

Access and Explore Data

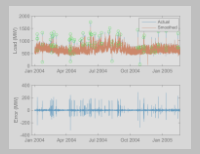
Files

Databases

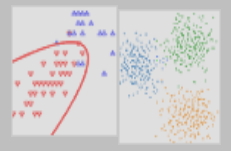
Sensors

Preprocess Data

Working with Messy Data



Data Reduction/ Transformation



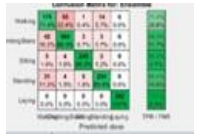
Feature Extraction

Develop Predictive Models

Model Creation e.g. Machine Learning

Parameter Optimization

Model Validation



Integrate Analytics with Systems

Desktop Apps

Enterprise Scale Systems

Embedded Devices and Hardware

Data Analytics Workflow

Access and Explore Data

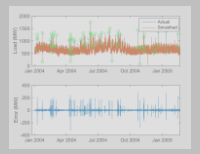
Files

Databases

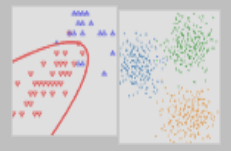
Sensors

Preprocess Data

Working with Messy Data



Data Reduction/ Transformation



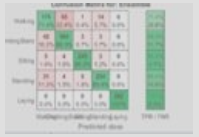
Feature Extraction

Develop Predictive Models

Model Creation e.g. Machine Learning

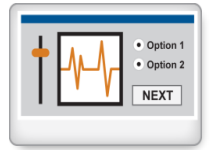
Parameter Optimization

Model Validation



Integrate Analytics with Systems

Desktop Apps



Enterprise Scale Systems

MATLAB Excel .NET .exe C/C++ Java .dll

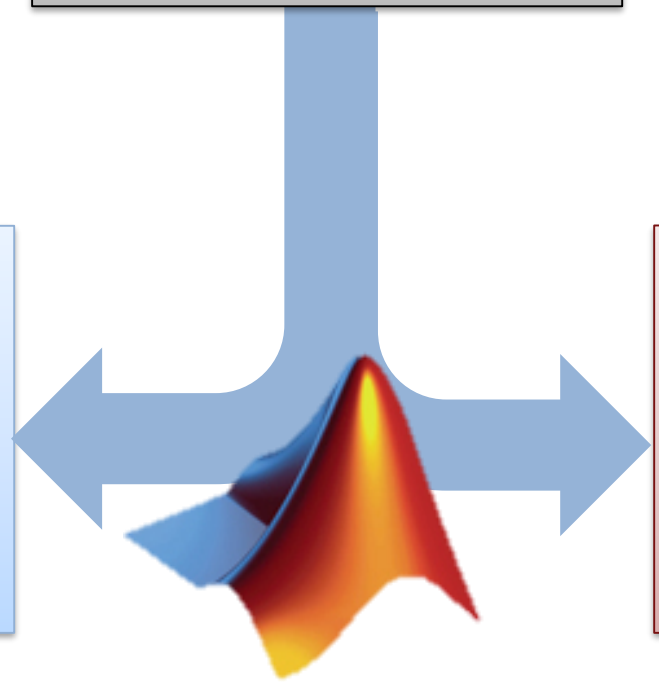
Embedded Devices and Hardware

Deploying Fleet Analytics

Vehicle data, driver profiles

“Cold Storage”

Historic data:
- **Batch processing**
- Large data on cluster
- Explore long term trends
- Build models



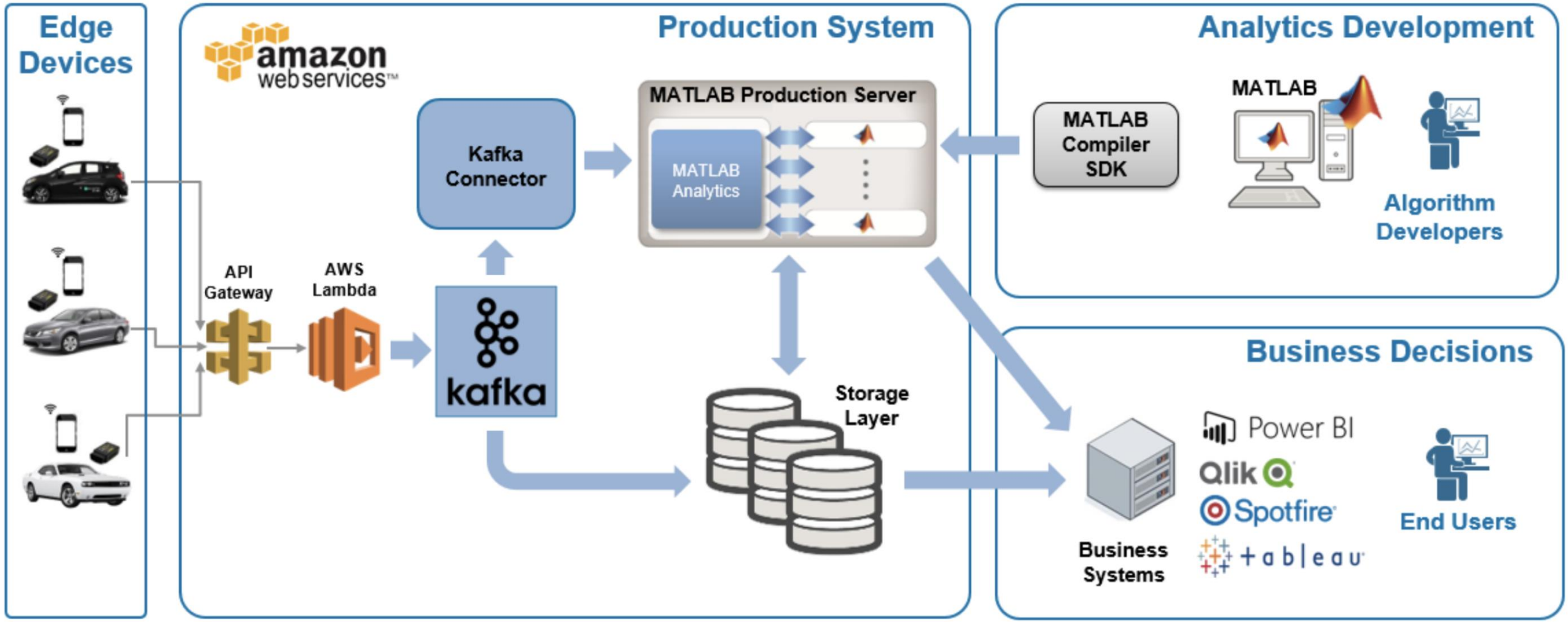
“Hot Storage”

Streaming data:
- **Near real-time**
- Test and implement model for new data
- Stream processing

Spark

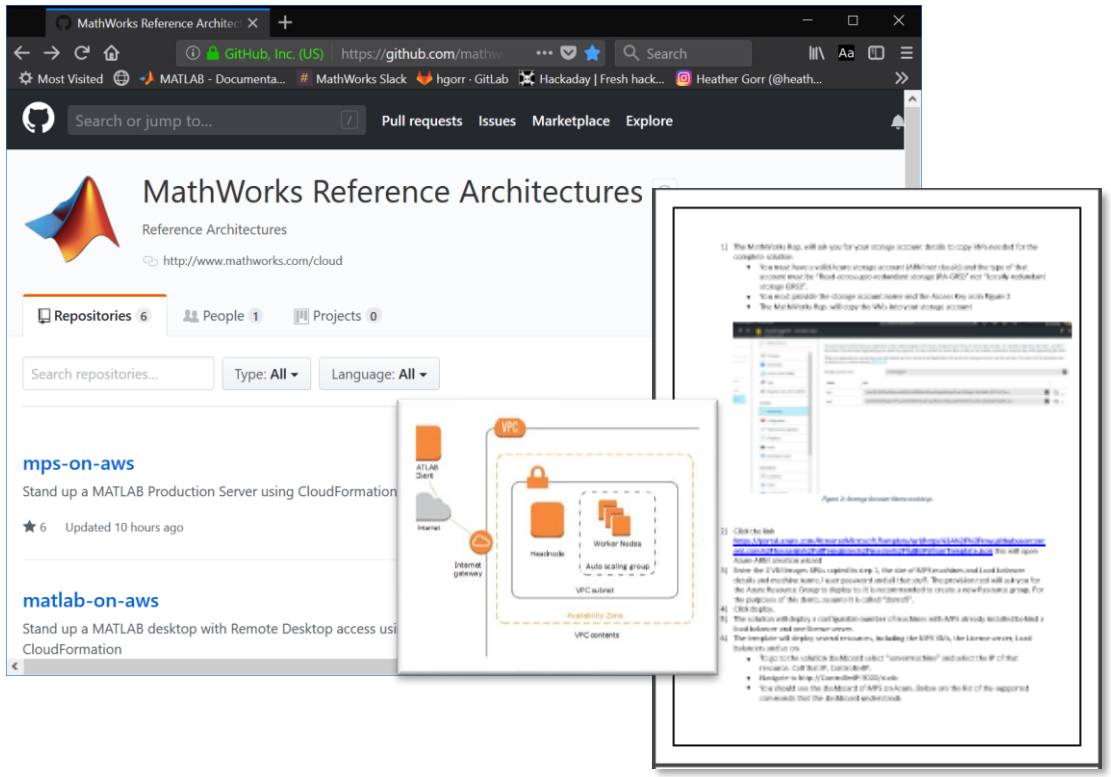
kafka

Fleet Analytics Streaming Architecture

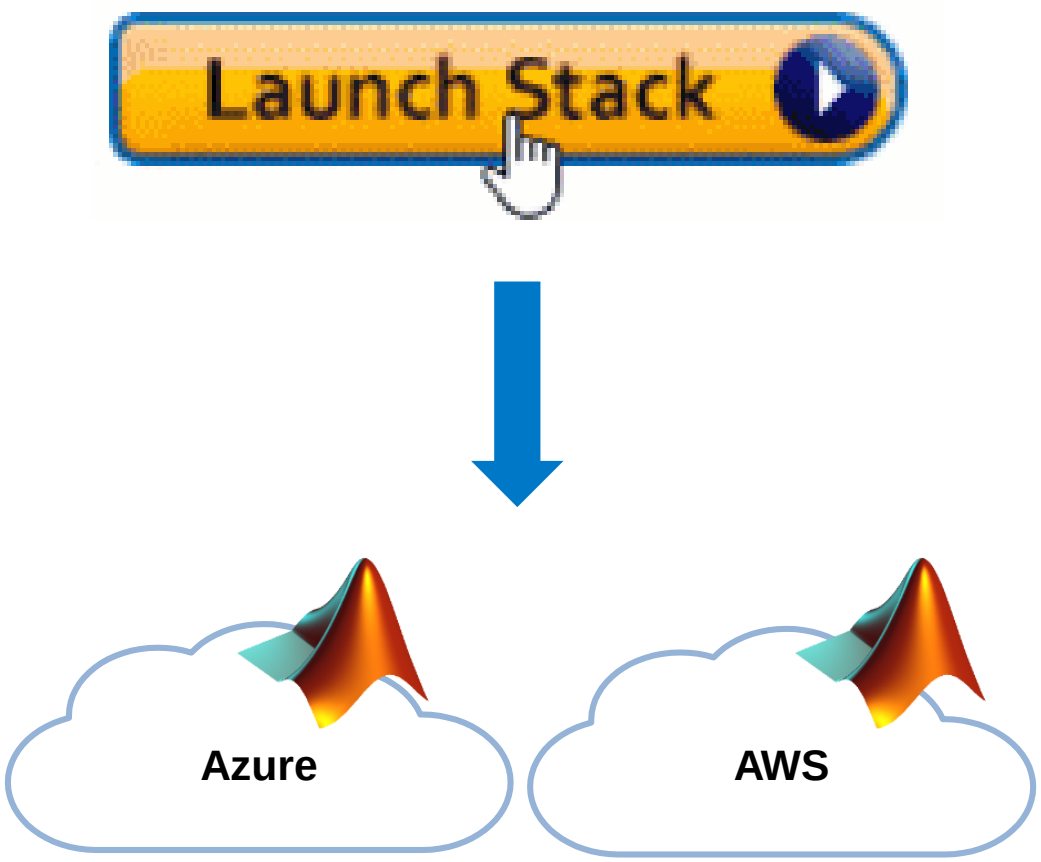


Use Reference Architectures to Run MATLAB on Cloud Platforms

GitHub



https://github.com/mathworks-ref-arch/



Data Analytics Workflow

Access and Explore Data

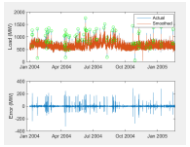
Files

Databases

Sensors

Preprocess Data

Working with Messy Data



Data Reduction/ Transformation

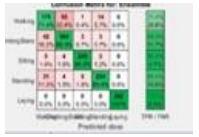
Feature Extraction

Develop Predictive Models

Model Creation e.g. Machine Learning

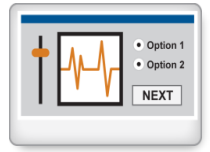
Parameter Optimization

Model Validation



Integrate Analytics with Systems

Desktop Apps



Enterprise Scale Systems

MATLAB Excel .NET .exe C/C++ Java .dll

Embedded Devices and Hardware

MATLAB Training Courses for Data Science

Training Courses

- MATLAB Fundamentals (3 days)
- MATLAB for Data Processing and Visualization (1 day)
- Processing Big Data with MATLAB (1 day)
- Statistical Methods in MATLAB (2 days)
- Machine Learning with MATLAB (2 days)
- Signal Preprocessing and Feature Extraction with MATLAB (1 day)
- Deep Learning with MATLAB (2 days)
- Accelerating and Parallelizing MATLAB Code (2 days)

- Exploratory Data Analysis
- Data Processing and Feature Engineering
- Predictive Modeling and Machine Learning
- Data Science Project

MathWorks® is a **Leader** in the Gartner Magic Quadrant for 2020 Data Science and Machine Learning Platforms

MathWorks has the furthest completeness of vision in the Leaders quadrant

We believe this recognition demonstrates our ability to:

- Empower your team, including those with limited AI or data science experience
- Provide complete workflows for data preparation, AI modeling, system design, and production
- Deploy AI models on embedded devices, edge, enterprise systems, and the cloud
- Use Simulink to tackle integration challenges and reduce risk in designing AI-driven systems

*Gartner Magic Quadrant for Data Science and Machine Learning Platforms, Peter Krensky, Erick Brethenoux, Jim Hare, Carlie Idoine, Alexander Linden, Svetlana Sicular, 11 February 2020 .

This graphic was published by Gartner, Inc. as part of a larger research document and should be evaluated in the context of the entire document. The Gartner document is available upon request from MathWorks. Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, express or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

Q&A

On which part of the data analytics workflow do you spend most time?

a. Data access

b. Data preprocessing

c. Data analysis and modeling

d. Enterprise integration

Please contact us with questions

Sebastian Bomberg
sbomberg@mathworks.com

MathWorks®