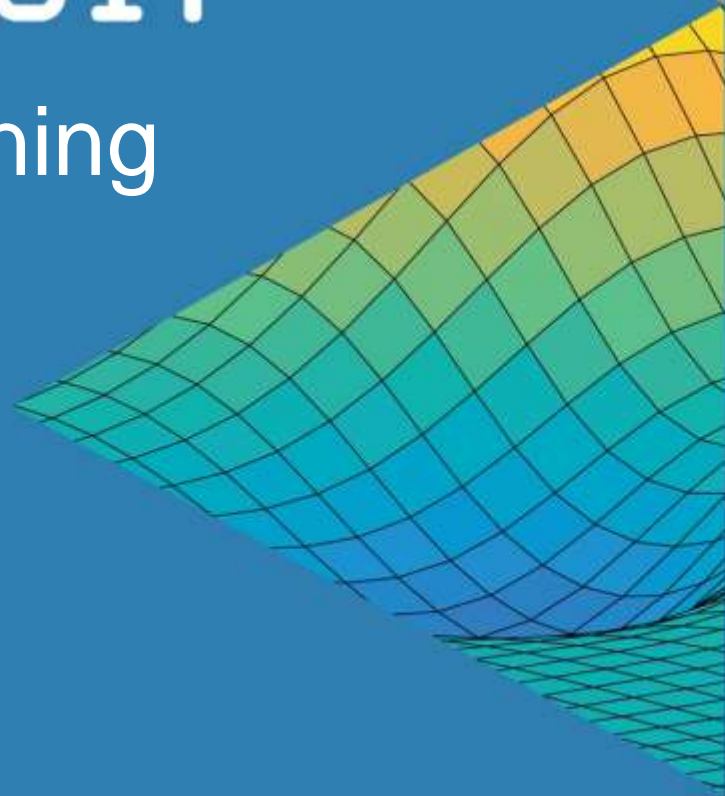


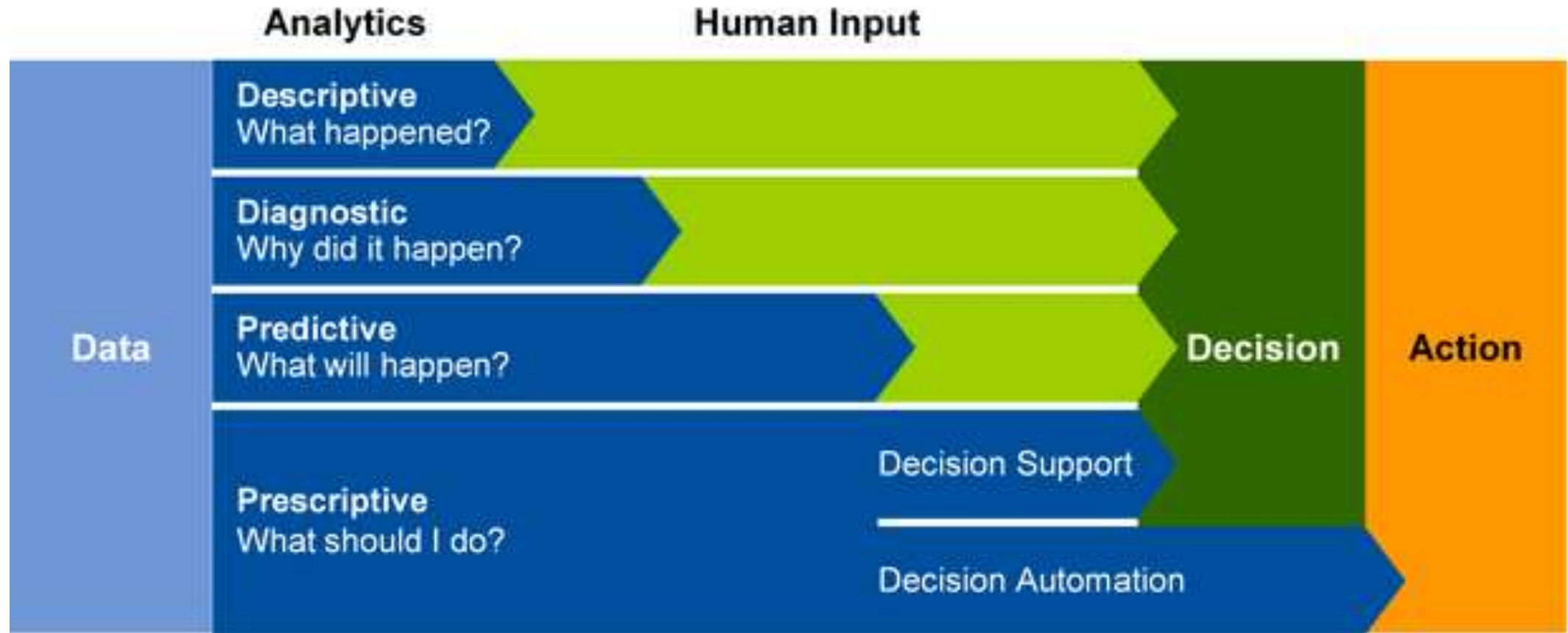
# MATLAB EXPO 2017

## Big Data and Machine Learning Using MATLAB

Seth DeLand & Amit Doshi  
MathWorks



# Data Analytics



Turn *large volumes* of complex data into actionable information  
 source: [Gartner](#)

Customer Example: Gas Natural Fenosa

# Energy Production Optimization

## User Story

### Opportunity

- Allocate demand among power plants to minimize generation costs

### Analytics Use

- **Data:** Central database for historical power consumption and price data, weather forecasts, and parameters for each power plant
- **Machine Learning:** Develop price simulation scenarios
- **Optimization:** minimize production cost

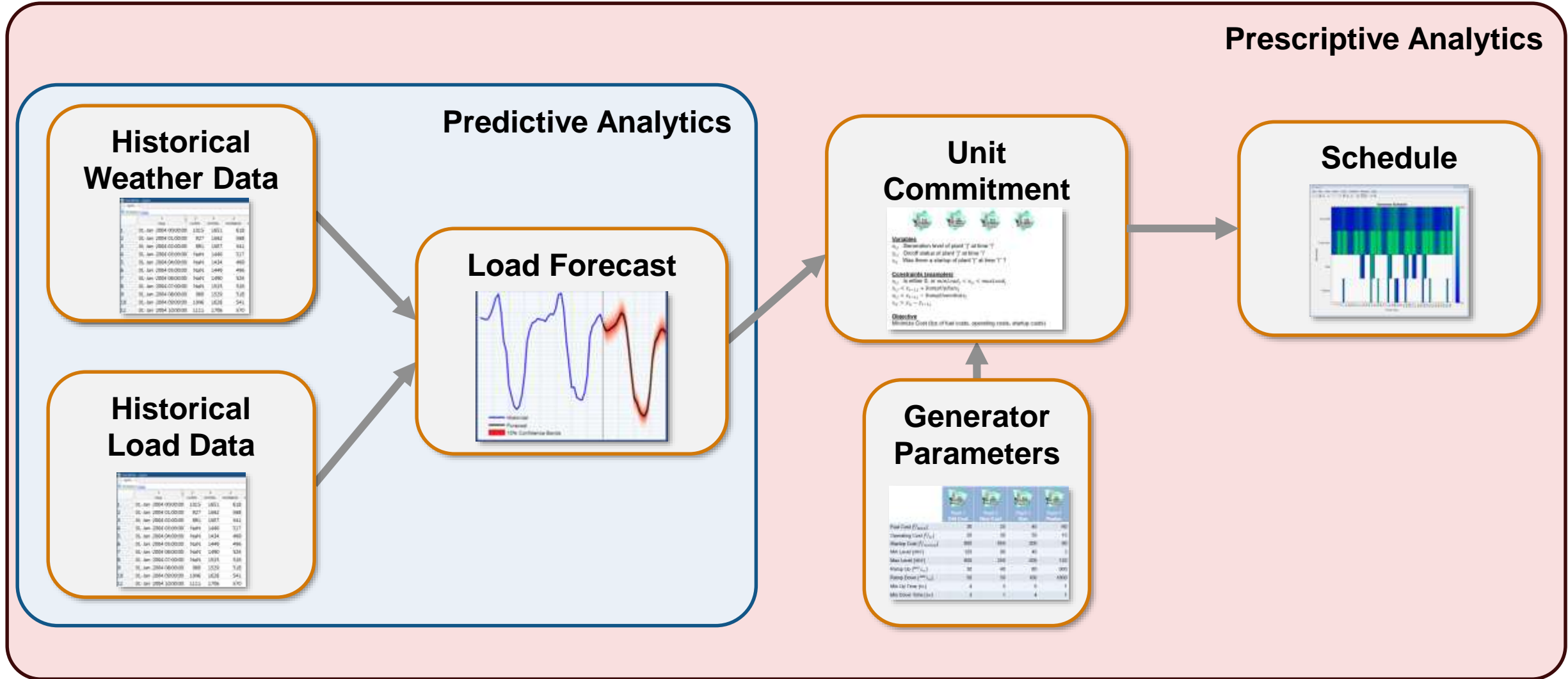
### Benefit

- Reduced generation costs
- White-box solution for optimizing power generation



# Unit Commitment

## Predictive and Prescriptive Analytics



# Big Data Analytics Workflow



**Files**

---

**Databases**

---

**Sensors**

17

**Working with Messy Data**

---

**Data Reduction/Transformation**

---

**Feature Extraction**

**Model Creation e.g. Machine Learning**

---

**Parameter Optimization**

---

**Model Validation**

**Desktop Apps**

---

**Enterprise Scale Systems**

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

---

**Embedded Devices and Hardware**

## Example: Working with Big Data in MATLAB

- **Objective:** Create a model to predict the cost of a taxi ride in New York City
- **Inputs:**
  - Monthly taxi ride log files
  - The local data set is **small** (~20 MB)
  - The full data set is **big** (~21 GB)
- **Approach:**
  - Access Data
  - Preprocess and explore data
  - Develop and validate predictive model (linear fit)
    - Work with subset of data for prototyping and then run on spark enabled hadoop with full data
  - Integrate analytics into a webapp





# Example: Working with Big Data in MATLAB

Live Editor - /mathworks/home/hgorr/predictTaxiFare.mlx

predictTaxiFare.mlx

## tall Arrays for Big Data in MATLAB

### Predict Cost of Taxi Ride in New York City

Analyze data from .csv files containing taxi trip information, separated by month. The data set is available from the [City of New York](#).

VendorID,	tpep_pickup_datetime,	tpep_dropoff_datetime,	passenger_count,	trip_distance,	pickup_longitude,	picku
2,	2015-01-07 07:40:20,	2015-01-07 08:04:45,	6,	9.12,	-73.9524536132812,	40.78
2,	2015-01-21 22:49:50,	2015-01-21 23:17:11,	6,	5.63,	-74.0083694458008,	40.73
1,	2015-01-05 23:04:30,	2015-01-05 23:15:00,	1,	2.9,	-73.8632125854492,	40.76
1,	2015-01-11 22:20:43,	2015-01-11 22:23:02,	1,	0.8,	-73.9577560424805,	40.76
2,	2015-01-24 00:34:59,	2015-01-24 00:38:39,	1,	0.65,	-73.9916687011719,	40.73
1,	2015-01-25 19:09:57,	2015-01-25 19:18:02,	1,	1.5,	-73.9983825683594,	40.72
1,	2015-01-02 23:24:13,	2015-01-02 23:27:30,	1,	1,	-73.9963912963967,	40.75
2,	2015-01-21 06:46:23,	2015-01-21 06:47:56,	1,	0.63,	-73.9913635253906,	40.77
2,	2015-01-23 19:32:33,	2015-01-23 19:48:56,	3,	2.52,	-73.988382018043,	40.73

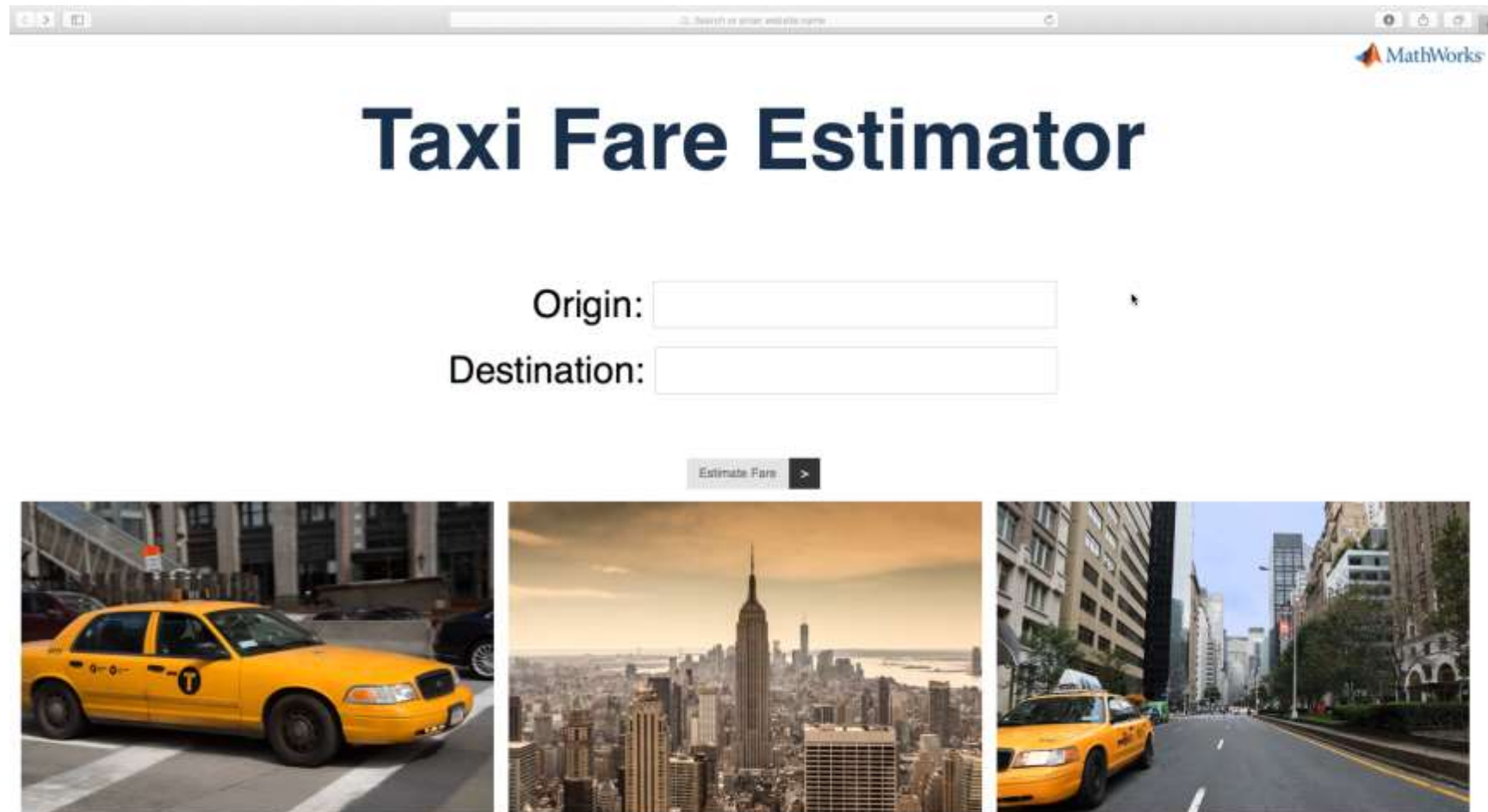
Set up execution environment

```
numWorkers = 16;

setenv('HADOOP_HOME', '/mathworks/test/hadoop');
setenv('SPARK_HOME', '/mathworks/test/spark');


cluster = parallel.cluster.Hadoop;
cluster.SparkProperties('spark.executor.instances') = num2str(numWorkers);
```




# Demo: Taxi Fare Predictor Web App



Origin:

Destination:

Estimate Fare 





# Big Data Analytics Workflow: Data Access and Pre-process

**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**

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

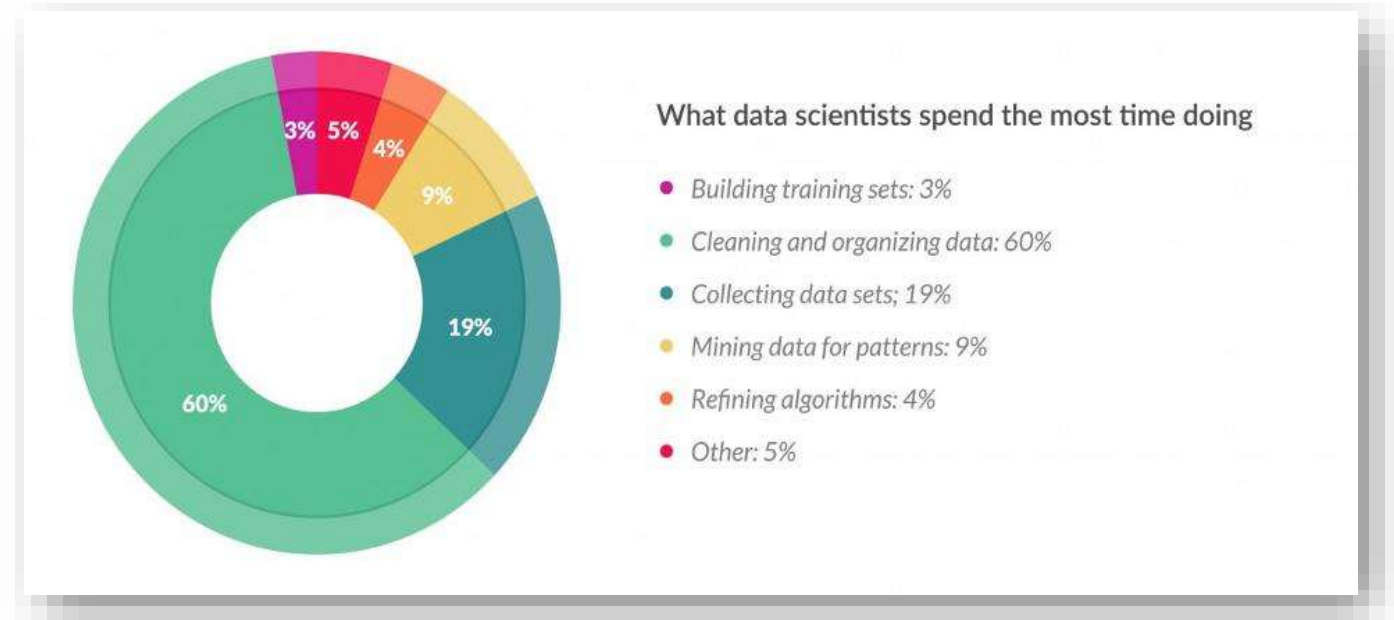
---

**Embedded Devices and Hardware**

# Data Access and Pre-processing – Challenges

## Challenges

- Data aggregation
  - Different sources (files, web, etc.)
  - Different types (images, text, audio, etc.)
- Data clean up
  - Poorly formatted files
  - Irregularly sampled data
  - Redundant data, outliers, missing data etc.
- Data specific processing
  - Signals: Smoothing, resampling, denoising, Wavelet transforms, etc.
  - Images: Image registration, morphological filtering, deblurring, etc.
- Dealing with out of memory data (big data)



Data preparation accounts for about **80%** of the work of data scientists - Forbes

# Data Analytics Workflow: Big Data Access and Pre-processing

www.nyc.gov/html/tlc/html/about/trip\_record\_data.shtml

Data Analytics - Home Discover MATLAB & S CRE - Home MATLAB Fleet Data Analysis

2016

2015

Month	Yellow	Green	FHV
January	Yellow	Green	FHV
February	Yellow	Green	FHV
March	Yellow	Green	FHV
April	Yellow	Green	FHV
May	Yellow	Green	FHV
June	Yellow	Green	FHV
July	Yellow	Green	FHV
August	Yellow	Green	FHV
September	Yellow	Green	FHV
October	Yellow	Green	FHV
November	Yellow	Green	FHV
December	Yellow	Green	FHV

2014

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
trip_dropoff_datetime	passenger_count	trip_distance	pickup_longitude	pickup_latitude	Ratecode	dropoff_longitude	dropoff_latitude	payment_type	fare_amount	extra_mileage	taxi_tax	tip_amount	tolls_amount		
58	1/14/2015 12:35	1	1.8	-73.957512	40.75554625	1	N	-73.975014	40.762	1	14.5	0	0.5	3.05	0
59	1/7/2015 10:06	1	0.47	-73.877183	40.74802192	1	N	-73.988467	40.748	2	4	0	0.5	0	0
60	1/30/2015 18:00	6	1.4	-73.992752	40.7428852	1	N	-73.990453	40.758	1	7	0	0.5	1.36	0
61	1/31/2015 22:52	2	1.07	-74.00074	40.73076248	1	N	-73.591249	40.723	1	9	0.5	0.5	2.06	0
62	1/26/2015 13:47	1	1.2	-73.888221	40.73631793	1	N	-73.994480	40.74	1	8	0	0.5	8	0
63	1/13/2015 20:29	1	0.9	-73.978462	40.76373525	1	N	-73.9729	40.766	1	4	0.5	0.5	1.32	0
64	1/23/2015 14:40	1	0.6	-73.900007	40.74665941	1	N	-73.999329	40.754	2	7	0	0.5	0	0
65	1/24/2015 0:25	1	0.4	-73.952452	40.77382537	1	N	-73.954094	40.778	2	4	0.5	0.5	0	0
66	1/17/2015 10:28	1	8.2	-73.862724	40.76882372	1	N	-73.834877	40.682	2	26.3	0	0.5	0	0
67	1/13/2015 19:55	1	1.5	-73.971371	40.76370239	1	N	-73.954365	40.778	1	6.5	1	0.5	1.64	0
68	1/22/2015 17:41	1	0.8	-73.909411	40.76430113	1	N	-73.963353	40.771	1	4.5	1	0.5	1.26	0
69	1/23/2015 10:13	1	0.9	-73.961428	40.77436829	1	N	-73.952182	40.787	2	6	1	0.5	0	0
70	1/29/2015 14:25	1	7.2	-73.962801	40.76917648	1	N	-74.01236	40.705	1	29	0	0.5	5.95	0
71	1/16/2015 8:34	1	1.7	-73.968742	40.79102325	1	N	-73.953842	40.775	1	8.5	0	0.5	1	0
72	1/28/2015 1:02	1	3.5	-73.99855	40.73015948	1	N	-73.910688	40.78	2	13	0.5	0.5	0	0
73	1/23/2015 14:22	2	4.1	-73.963371	40.77402837	1	N	-73.913901	40.811	1	24	0	0.5	0.05	0
74	1/11/2015 1:17	1	4	-73.981026	40.72495647	1	N	-73.956795	40.746	1	14	0.5	0.5	7	0
75	1/24/2015 2:02	1	1.5	-73.999339	40.73746062	1	N	-72.981381	40.775	2	10.5	0.5	0.5	0	0
76	1/24/2015 20:35	1	0.58	-73.946434	40.77729415	1	N	-73.943354	40.784	1	5	0.5	0.5	1.26	0
77	1/23/2015 23:39	1	8.89	-73.957838	40.7702446	1	N	-74.002678	40.75	1	18.5	0.5	0.5	0	0
78	1/14/2015 18:30	1	1.05	-74.01563	40.75098901	1	N	-73.997086	40.78	1	6	1	0.5	1.25	0
79	1/27/2015 17:02	2	0.32	-73.99295	40.72434616	1	N	-73.997528	40.726	2	3	1	0.5	0	0

Download 2015 Taxi Data from Web using 'websave' in parallel

```

parfor i=1:12
    fileName = ['taxiData2015_', num2str(i)]
    url      = ['https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2016-0', num2str(i), '.csv']
    websave(fileName, url)
end

```

## Next: Access Big Data from MATLAB

- **datastore**

- Tabular text files
- Images
- Excel spreadsheets
- (SQL) Databases
- HDFS (Hadoop)
- S3 - Amazon

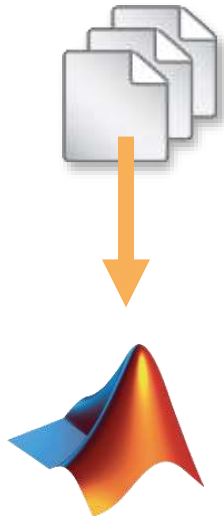
**R2014b**

**R2015a**

**R2015b**

**R2016a**

# Get data in MATLAB



MATLAB R2017a

HOME PLOTS APPS SHORTCUTS LIVE EDITOR VIEW

datastore

C:\AmitDrive\04\_Demos\Demos\00\_Data Analytics\NewYork\_Taxi\NYTaxiDemo


Live Editor - C:\AmitDrive\04\_Demos\Demos\00\_Data Analytics\NewYork\_Taxi\NYTaxiDemo\predictTaxiFare.mlx

predictTaxiFare.mlx

## tall Arrays for Big Data in MATLAB

### Predict Cost of Taxi Ride in New York City

This example explores NYC taxi data and predicts the fare based on distance and the time of day.



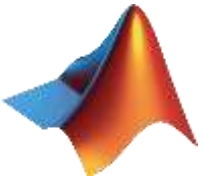
The data come from .csv files containing taxi trip information, separated by month. The data set is freely available from the [City of New York](#).

VendorID,	tpep_pickup_datetime,	tpep_dropoff_datetime,	passenger_count,	trip_distance,	pickup_longitude,	picku
2,	2015-01-07 07:40:20,	2015-01-07 08:04:45,	6,	9.12,	-73.9524536132812,	40.78
2,	2015-01-21 22:49:50,	2015-01-21 23:17:11,	6,	5.63,	-74.0083694458008,	40.73
1,	2015-01-05 23:04:30,	2015-01-05 23:15:00,	1,	2.9,	-73.8632125854492,	40.76
1,	2015-01-11 22:20:43,	2015-01-11 22:23:02,	1,	0.8,	-73.9577560424805,	40.76
2,	2015-01-24 00:34:59,	2015-01-24 00:38:39,	1,	0.65,	-73.9916687011719,	40.73
1,	2015-01-25 19:09:57,	2015-01-25 19:18:02,	1,	1.5,	-73.9983825683594,	40.72
1,	2015-01-02 23:24:13,	2015-01-02 23:27:30,	1,	1,	-73.9963912963867,	40.75
2,	2015-01-21 06:46:23,	2015-01-21 06:47:56,	1,	0.63,	-73.9913635253906,	40.77
2,	2015-01-23 19:32:33,	2015-01-23 19:49:56,	1,	2.52,	-73.999382018043,	40.73

Set up execution environment

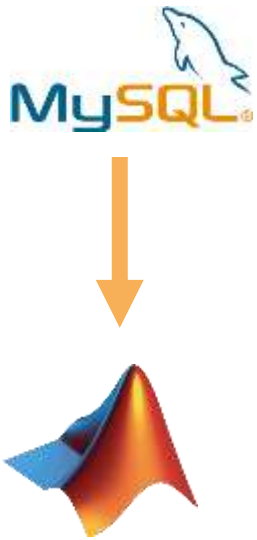


# What if the data is saved in HDFS?



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	icetto	supergroup	2.57 GB	9/25/2016, 1:39:41 AM	3	128 MB	yellow_tripdata_2012-01.csv
-rw-r--r--	icetto	supergroup	2.59 GB	9/25/2016, 1:40:09 AM	3	128 MB	yellow_tripdata_2012-02.csv
-rw-r--r--	icetto	supergroup	2.79 GB	9/25/2016, 1:40:38 AM	3	128 MB	yellow_tripdata_2012-03.csv
-rw-r--r--	icetto	supergroup	2.67 GB	9/25/2016, 1:41:05 AM	3	128 MB	yellow_tripdata_2012-04.csv
-rw-r--r--	icetto	supergroup	2.69 GB	9/25/2016, 1:41:20 AM	3	128 MB	yellow_tripdata_2012-05.csv
-rw-r--r--	icetto	supergroup	2.61 GB	9/25/2016, 1:41:53 AM	3	128 MB	yellow_tripdata_2012-06.csv
-rw-r--r--	icetto	supergroup	2.48 GB	9/25/2016, 1:42:16 AM	3	128 MB	yellow_tripdata_2012-07.csv
-rw-r--r--	icetto	supergroup	2.48 GB	9/25/2016, 1:42:40 AM	3	128 MB	yellow_tripdata_2012-08.csv
-rw-r--r--	icetto	supergroup	2.28 GB	9/25/2016, 1:42:59 AM	3	128 MB	yellow_tripdata_2012-09.csv
-rw-r--r--	icetto	supergroup	2.26 GB	9/25/2016, 1:43:20 AM	3	128 MB	yellow_tripdata_2012-10.csv

# Or Data is stored in a Database



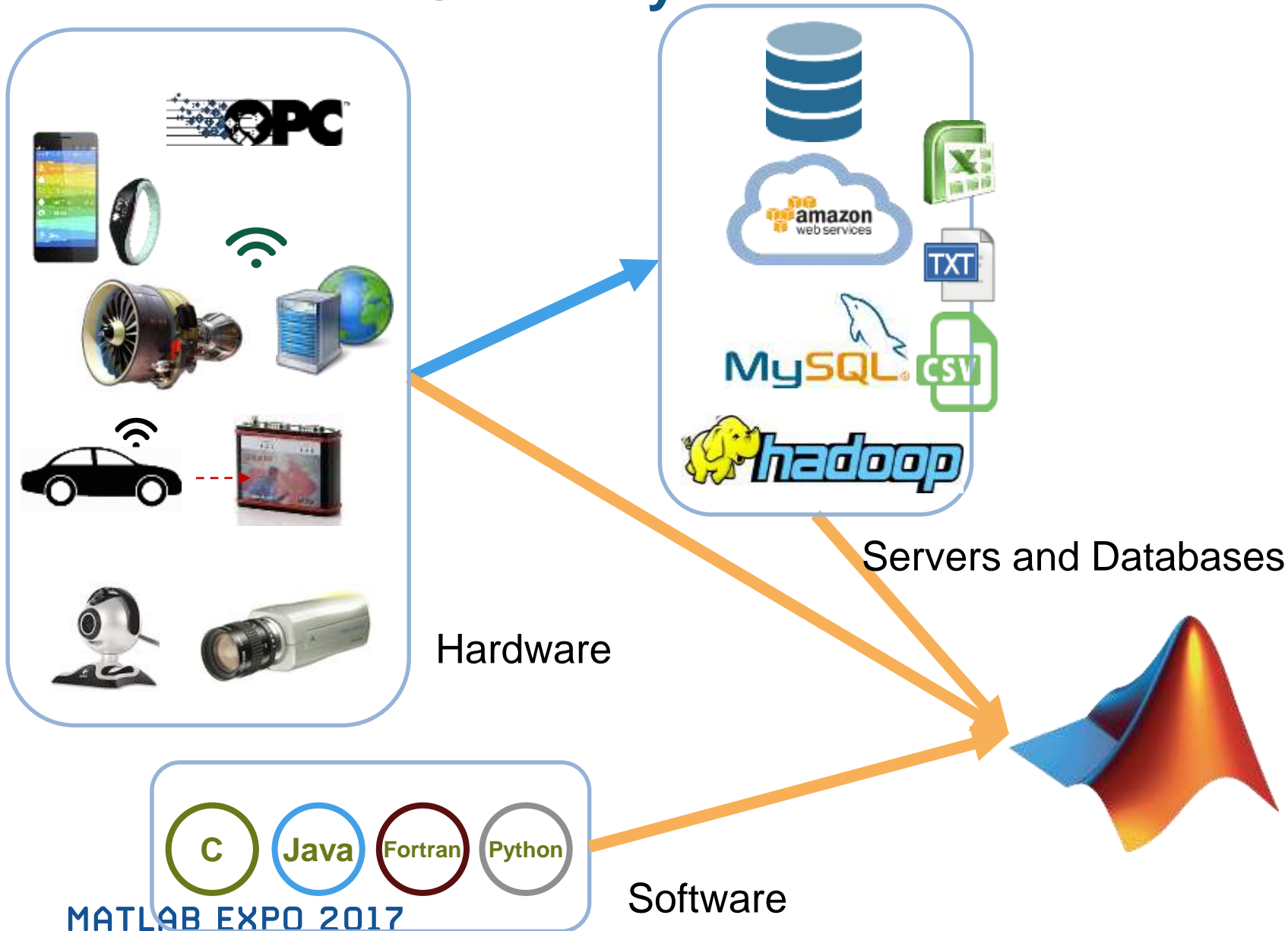
## Connect to the database application

```
conn = database('taxiDemo', 'root', 'matlab', ...  
              'Vendor', 'MYSQL', ...  
              'Server', 'localhost', ...  
              'PortNumber', 3306);
```

## Create a database datastore and import data of interest

```
sqlquery = ['select pickuptime, dropofftime, trip_distance, ...  
           'payment_type, fare_amount from taxiData'];  
ds = databaseDatastore(conn, sqlquery, 'ReadSize', 100000);
```

# Data Access: Summary



## Business and Transactional Data

- Repositories – SQL, NoSQL, etc.
- File I/O – Text, Spreadsheet, etc.
- Web Sources – RESTful, JSON, etc.

## Engineering, Scientific and Field Data

- Real-Time Sources – Sensors, GPS, etc.
- File I/O – Image, Audio, etc.
- Communication Protocols – OPC (OLE for Process Control), CAN (Controller Area Network), etc.

# Process data which doesn't fit into memory

Access and Explore Data

Files



Databases



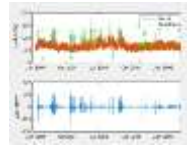
Sensors



17

Preprocess Data

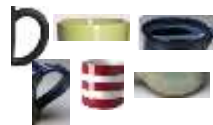
Working with Messy Data



Data Reduction/  
Transformation



Feature  
Extraction



Develop Predictive Models

Model Creation e.g.  
Machine Learning



Parameter  
Optimization

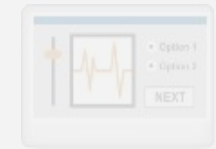


Model  
Validation



Integrate Analytics with Systems

Desktop Apps



Enterprise Scale  
Systems

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

Embedded Devices  
and Hardware



# Pre-processing Big Data

## ta11 arrays in **R2016b**

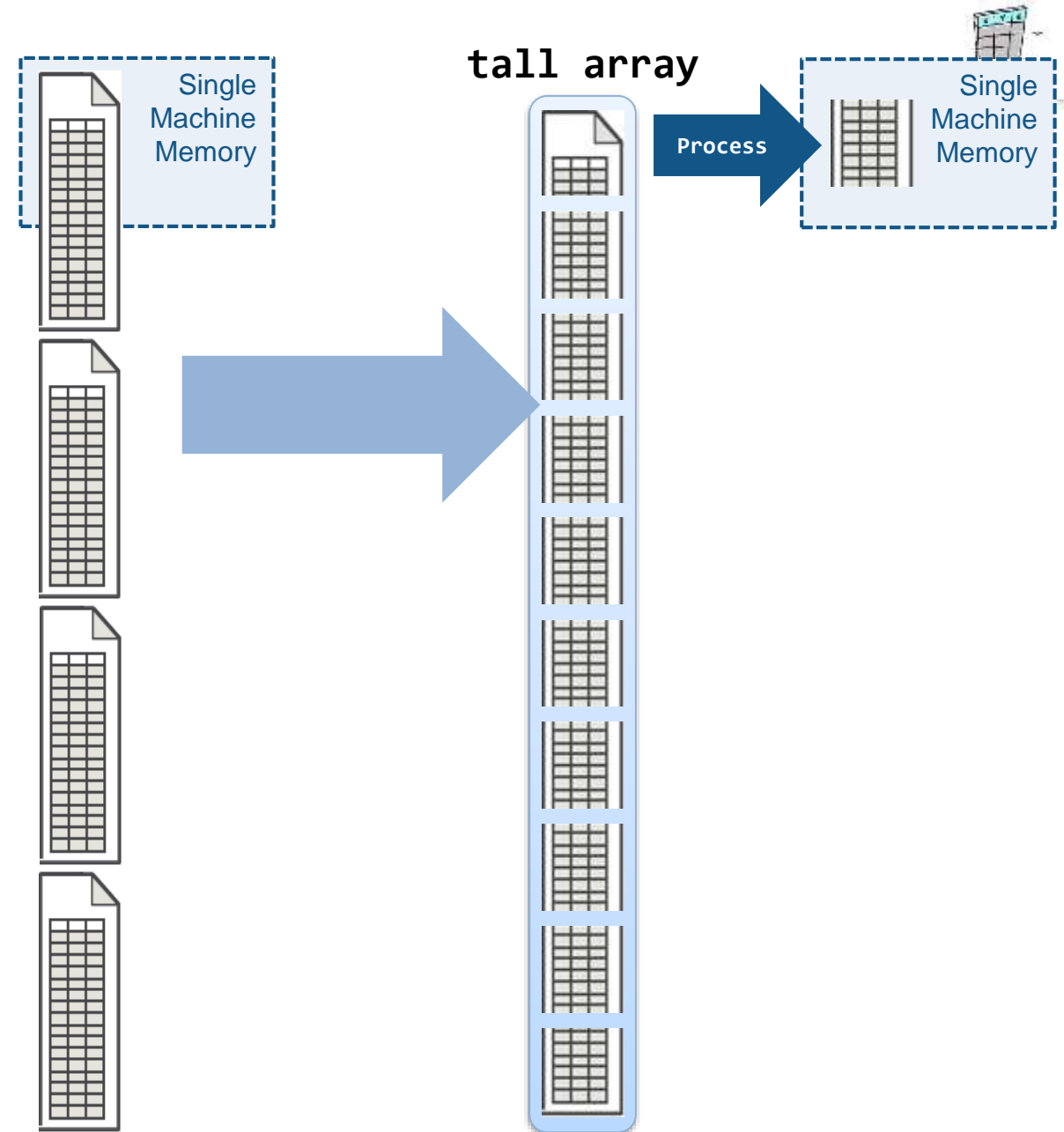
- New data type designed for data that doesn't fit into memory
- Lots of observations (hence "tall")
- Looks like a normal MATLAB array
  - Supports numeric types, tables, datetimes, strings, etc...
  - Supports several hundred functions for basic math, stats, indexing, etc.
  - **Statistics and Machine Learning Toolbox** support (clustering, classification, etc.)





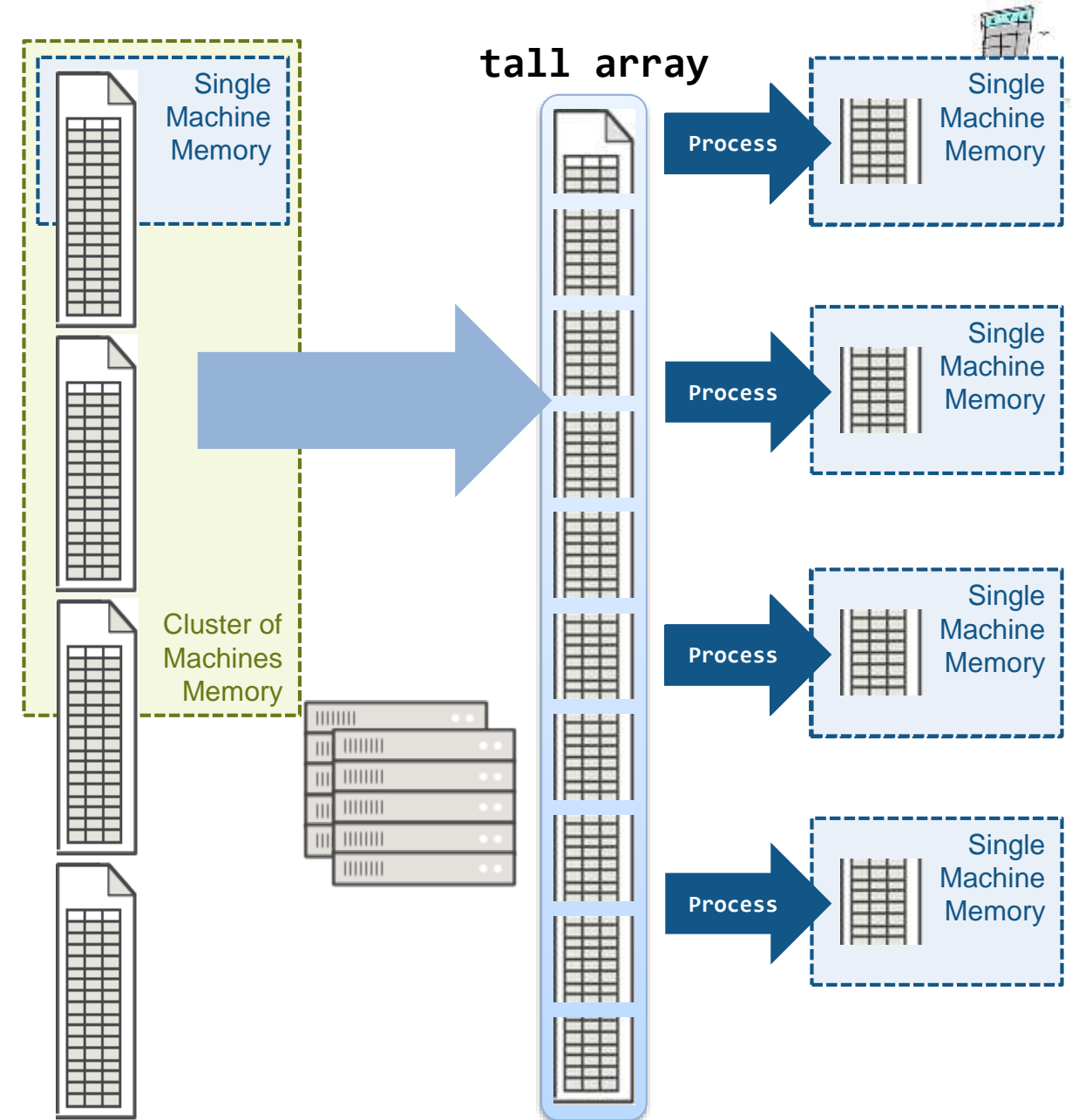
# tall arrays R2016b

- Automatically breaks data up into small “chunks” that fit in memory
- Tall arrays scan through the dataset one “chunk” at a time
- Processing code for tall arrays is the same as ordinary arrays



# ta11 arrays R2016b

- With Parallel Computing Toolbox, process several “chunks” at once
- Can scale up to clusters with MATLAB Distributed Computing Server



# Demo: Working with Tall Arrays

The image shows the MATLAB R2017a interface with a Live Editor script titled 'predictTaxiFare.mlx'. The script is divided into four sections:

- Create datastore to represent the data**: A code block showing the creation of a datastore from CSV files.
 

```
ds = datastore('smallerTaxiData/*2015.csv');
```
- Identify data of interest and customize options.**: A code block showing the selection of specific variables and their formats.
 

```
ds.VariableNames(2:3) = {'pickuptime', 'dropofftime'};
ds.SelectedVariableNames = {'pickuptime', 'dropofftime', 'trip_distance', ...
    'payment_type', 'fare_amount'};
ds.SelectedFormats(1:2) = {'%{yyyy-MM-dd HH:mm:ss}D'};
```
- Create a tall array**: A code block showing the creation of a tall array from the datastore.
 

```
tt = tall(ds)
```
- Determine trip duration**: A code block showing the calculation of trip duration.
 

```
tt.hr_of_day = hour(tt.pickuptime);
tt.trip_minutes = minutes(tt.dropofftime - tt.pickuptime)
```

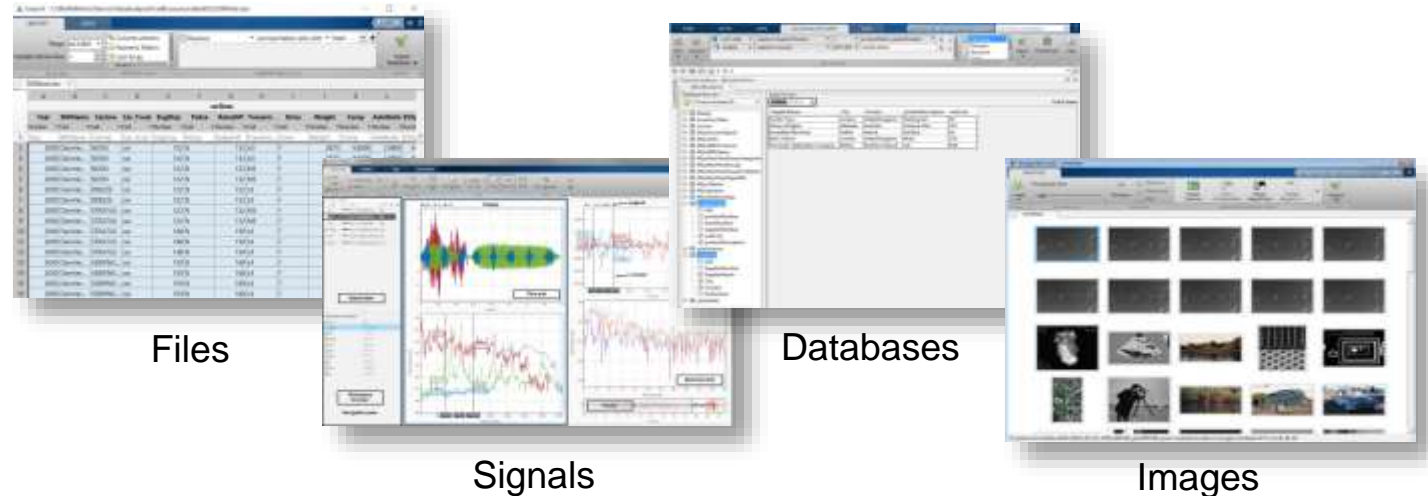
The Workspace window on the left shows a variable 'ds' of type '1x1 TabularTextDatastore'. The Current Folder window shows the directory structure, including 'smallerTaxiData' and 'taxiData' folders.

# Data Access and pre-processing – challenges and solution

## Challenges

- Data aggregation
  - Different sources (files, web, etc.)
  - Different types (images, text, audio, etc.)
- Data clean up
  - Poorly formatted files
  - Irregularly sampled data
  - Redundant data, outliers, missing data etc.
- Data specific processing
  - Signals: Smoothing, resampling, denoising, Wavelet transforms, etc.
  - Images: Image registration, morphological filtering, deblurring, etc.
- Dealing with out of memory data (big data)

1  
MATLAB makes it easy to work with **business and engineering data**



- Point and click tools to access variety of data sources
- High-performance environment for **big data**
- Built-in algorithms for data preprocessing including sensor, image, audio, video and other real-time data

# Data Analytics Workflow: Develop Predictive Models using **Big Data**

Access and Explore Data

Files



Databases



Sensors



17

Preprocess Data

Working with Messy Data



Data Reduction/  
Transformation

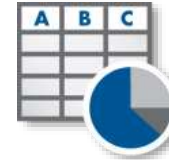


Feature  
Extraction



Develop Predictive Models

Model Creation e.g.  
Machine Learning



Parameter  
Optimization

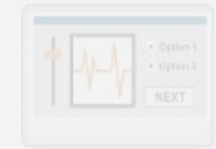


Model  
Validation



Integrate Analytics with Systems

Desktop Apps



Enterprise Scale  
Systems

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

Embedded Devices  
and Hardware

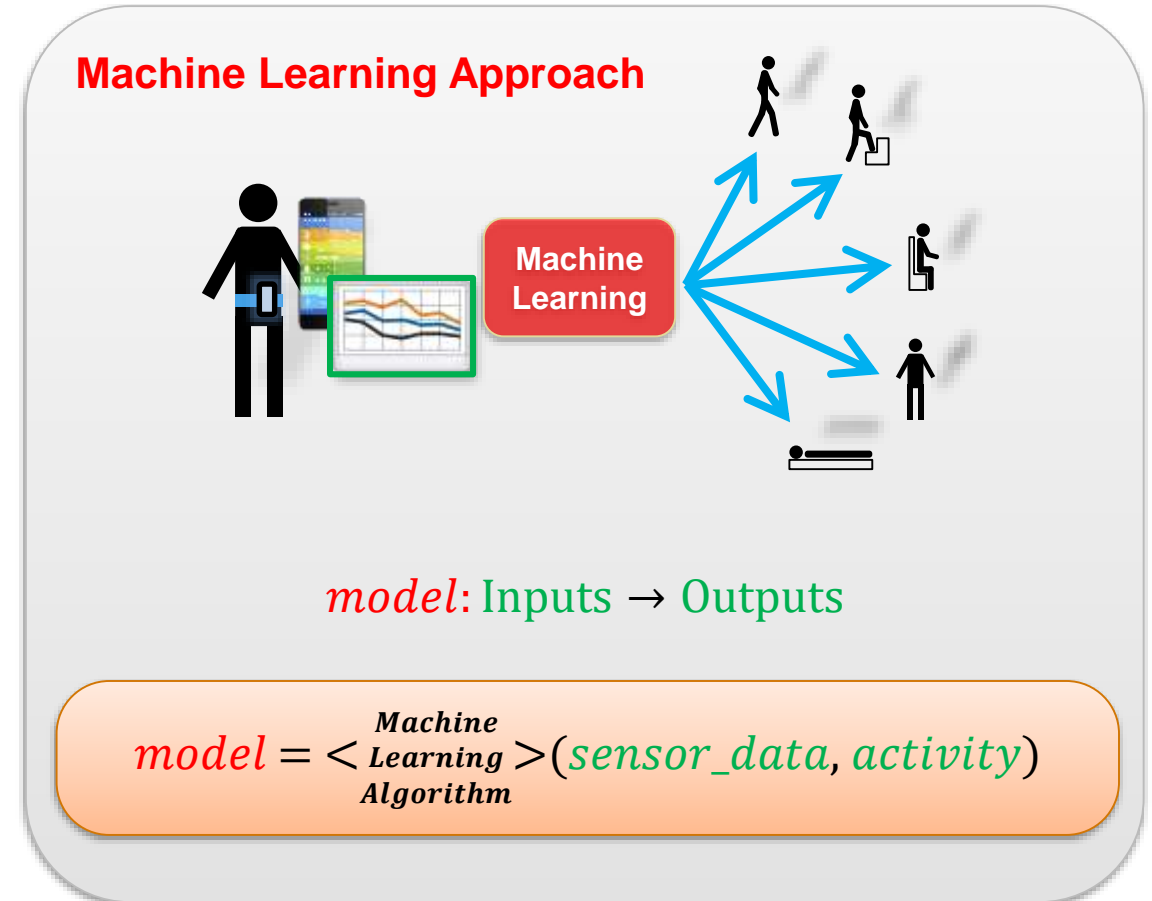
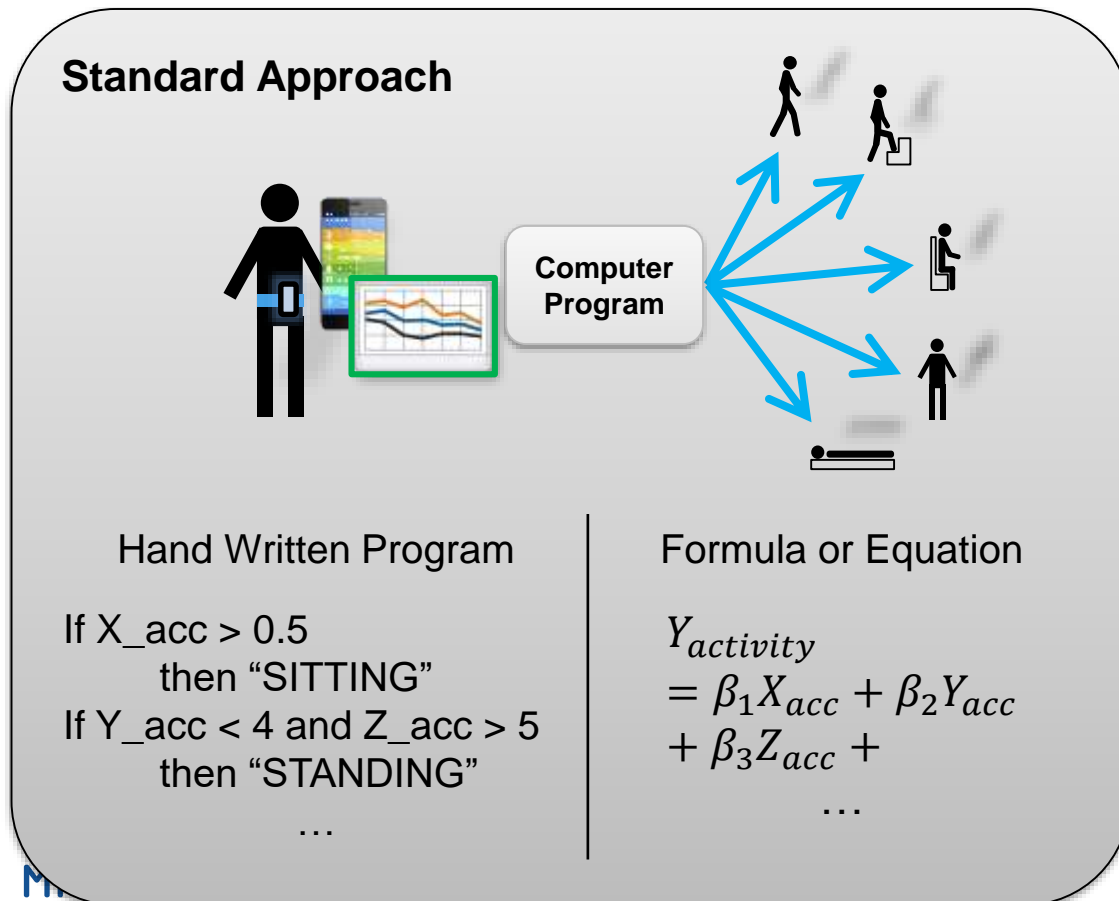




# Machine Learning

Machine learning uses **data** and produces a **program** to perform a **task**

## Task: Human Activity Detection



# Consider Machine/Deep Learning When

Problem is too complex for hand written rules or equations



Speech Recognition



Object Recognition



Engine Health Monitoring

Because algorithms can

*learn complex non-linear relationships*

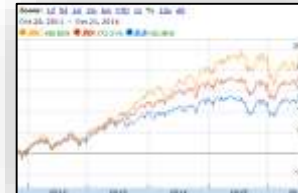
Program needs to adapt with changing data



Weather Forecasting



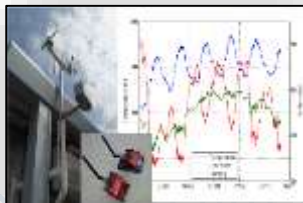
Energy Load Forecasting



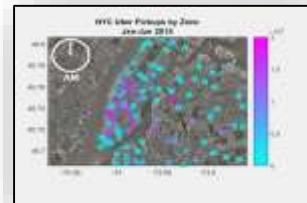
Stock Market Prediction

*update as more data becomes available*

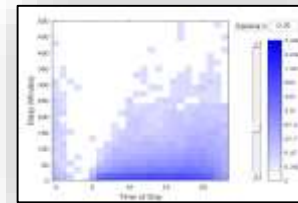
Program needs to scale



IoT Analytics



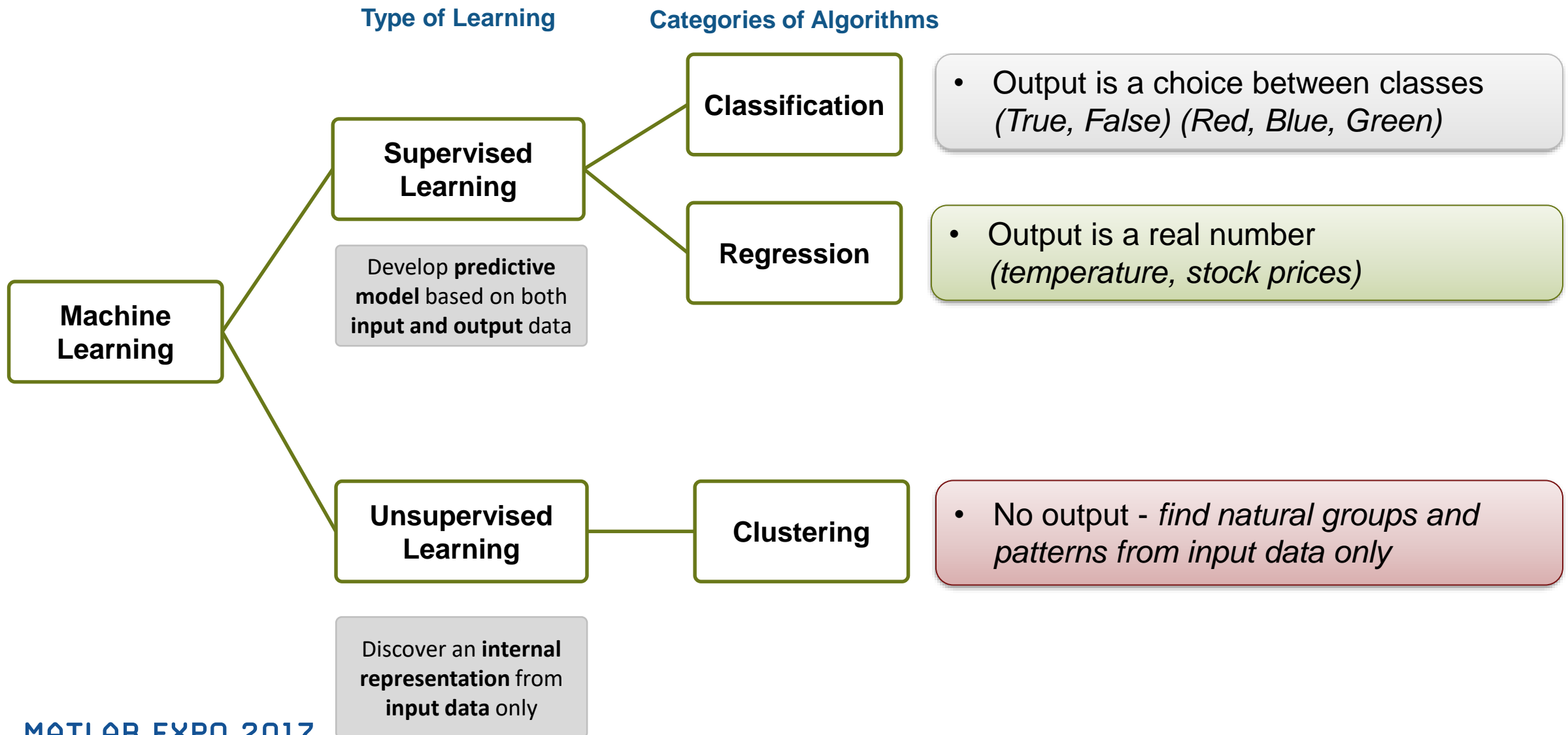
Taxi Availability



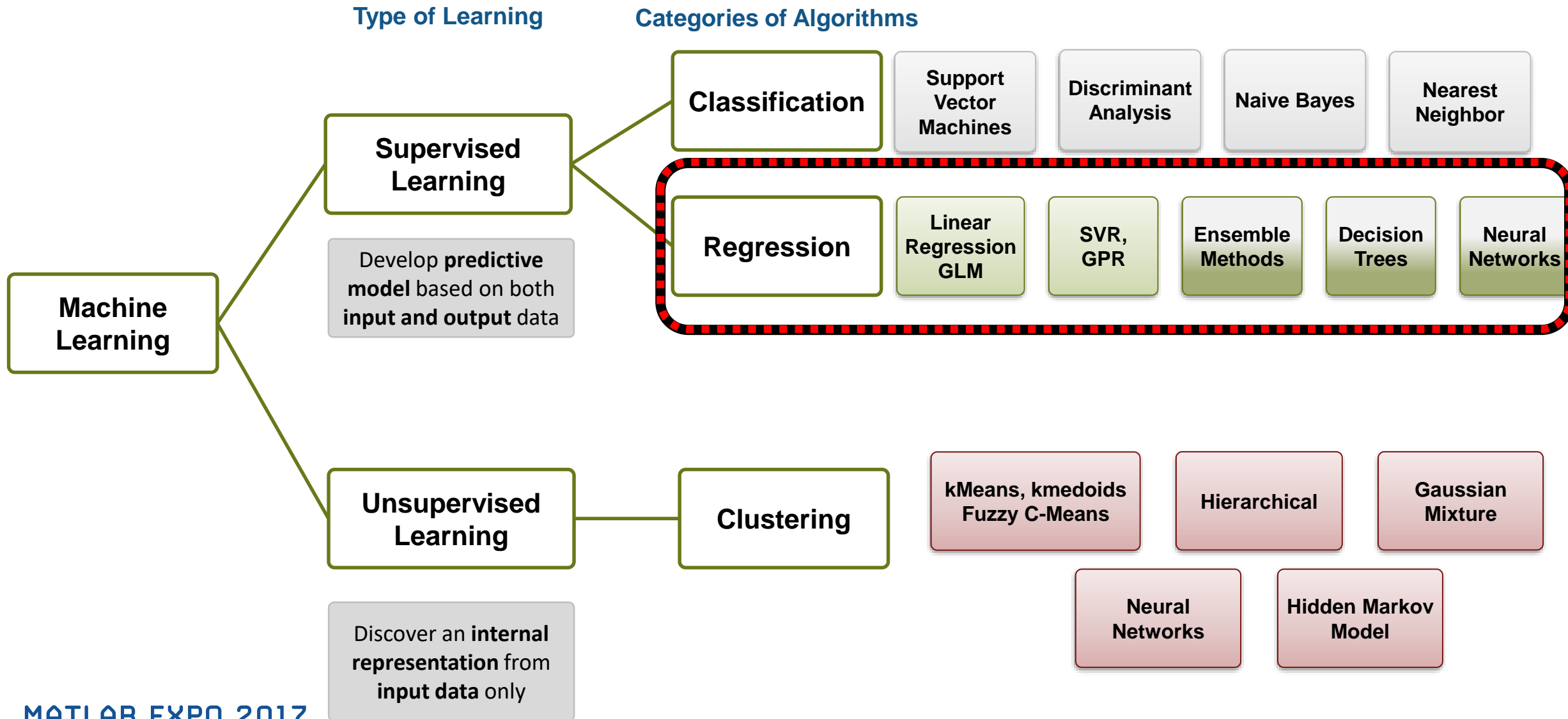
Airline Flight Delays

*learn efficiently from very large data sets*

# Different Types of Learning



# Different Types of Learning



# Machine Learning with Big Data

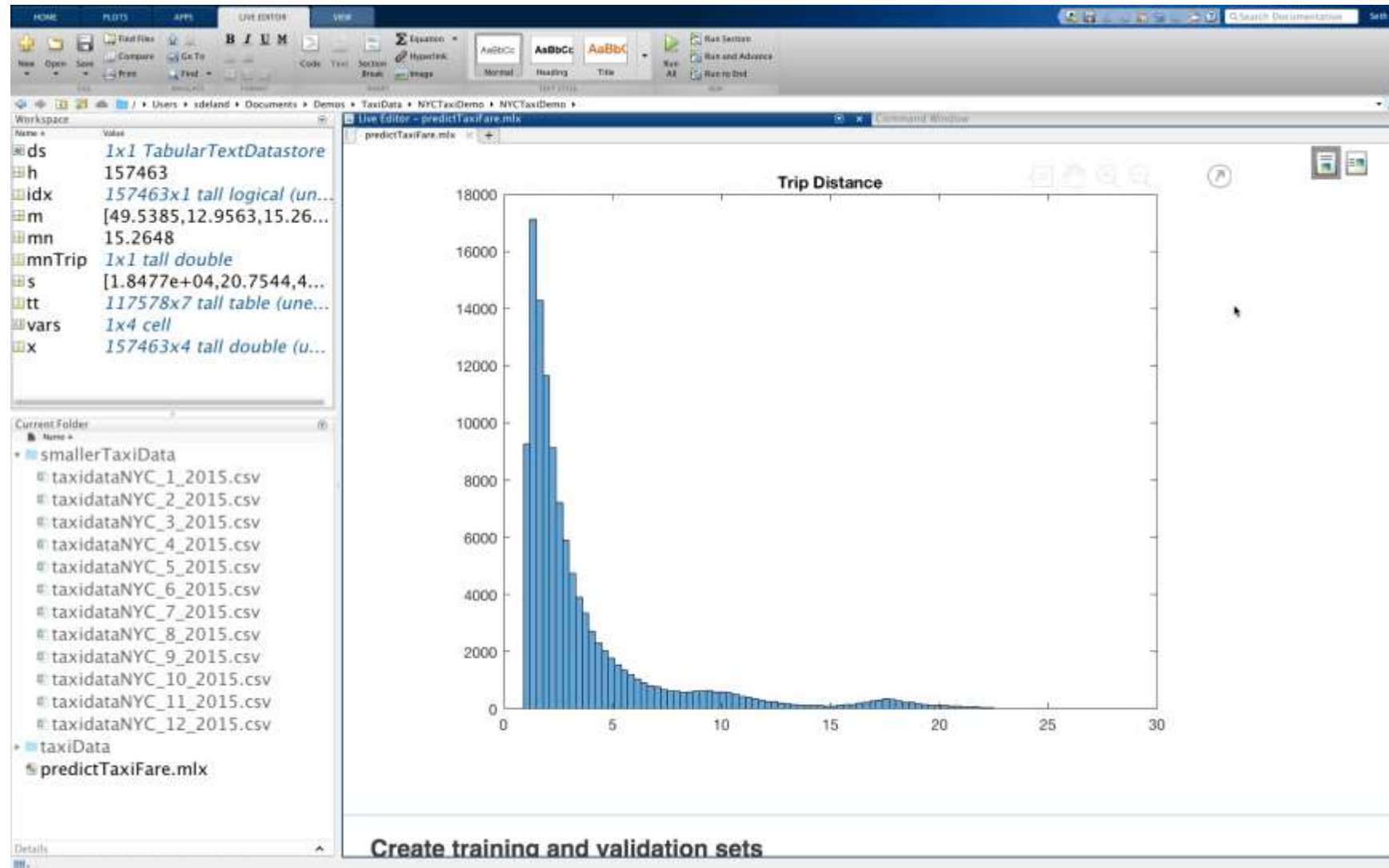
## R2016b

- Descriptive statistics (skewness, tabulate, crosstab, cov, grpstats, ...)
- K-means clustering (kmeans)
- Visualization (ksdensity, binScatterPlot; histogram, histogram2)
- Dimensionality reduction (pca, pcacov, factoran)
- Linear and generalized linear regression (fitlm, fitglm)
- Discriminant analysis (fitcdiscr)

## R2017a

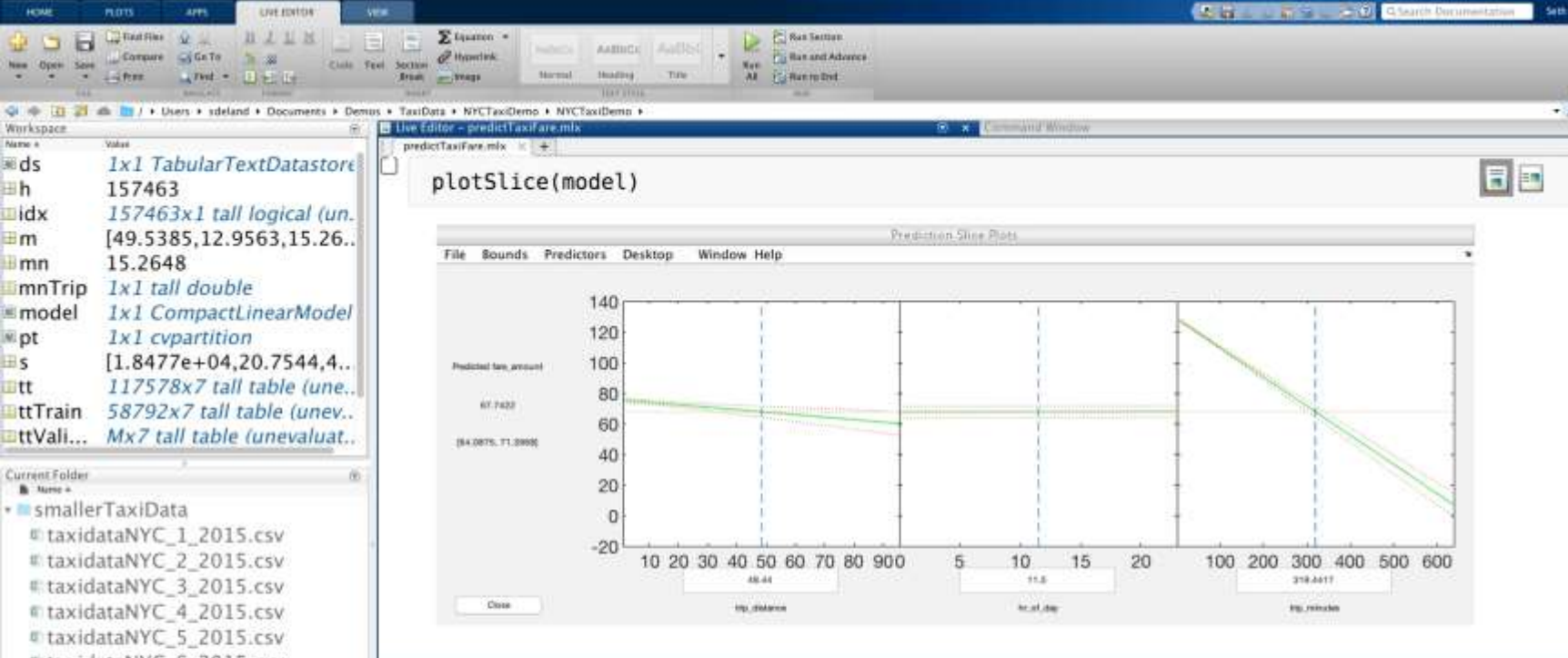
- Linear classification methods for SVM and logistic regression (fitclinear)
- Random forest ensembles of classification trees (TreeBagger)
- Naïve Bayes classification (fitcnb)
- Regularized regression (lasso)
- Prediction applied to tall arrays

# Demo: Training a Machine Learning Model





# Demo: Training a Machine Learning Model



**Workspace**

Name	Value
ds	1x1 TabularTextDatastore
h	157463
idx	157463x1 tall logical (un...
m	[49.5385, 12.9563, 15.26...
mn	15.2648
mnTrip	1x1 tall double
model	1x1 CompactLinearModel
pt	1x1 cvpartition
s	[1.8477e+04, 20.7544, 4...
tt	117578x7 tall table (une...
ttTrain	58792x7 tall table (unev...
ttVali...	Mx7 tall table (unevaluat...

**Current Folder**

- smallerTaxiData
  - taxidataNYC\_1\_2015.csv
  - taxidataNYC\_2\_2015.csv
  - taxidataNYC\_3\_2015.csv
  - taxidataNYC\_4\_2015.csv
  - taxidataNYC\_5\_2015.csv
  - taxidataNYC\_6\_2015.csv
  - taxidataNYC\_7\_2015.csv
  - taxidataNYC\_8\_2015.csv
  - taxidataNYC\_9\_2015.csv
  - taxidataNYC\_10\_2015.csv
  - taxidataNYC\_11\_2015.csv
  - taxidataNYC\_12\_2015.csv
- taxiData
  - predictTaxiFare.mlx

**plotSlice(model)**

**Prediction Slice Plots**

Predicted fare amount

67.7422

[84.0875, 71.3998]

10 20 30 40 50 60 70 80 900

5 10 15 20

100 200 300 400 500 600

trip\_distance

trip\_time

trip\_mins

**Predict and validate**

```

yPred = predict(model, ttValidation);
residuals = yPred - ttValidation.trip_minutes;
figure
histogram(residuals, 'Normalization', 'pdf', 'BinLimits', [-50 50])
  
```

# Regression Learner

The screenshot shows the MATLAB Regression Learner application. A 'New Session' dialog box is open, displaying three steps for configuring a regression model:

- Step 1: Select a table or matrix** - Lists available variables: Acceleration, Cylinders, Displacement, Horsepower, MPG, Model\_Year, and Weight.
- Step 2: Select predictors and response** - A table showing the selection of predictors and the response variable.
 

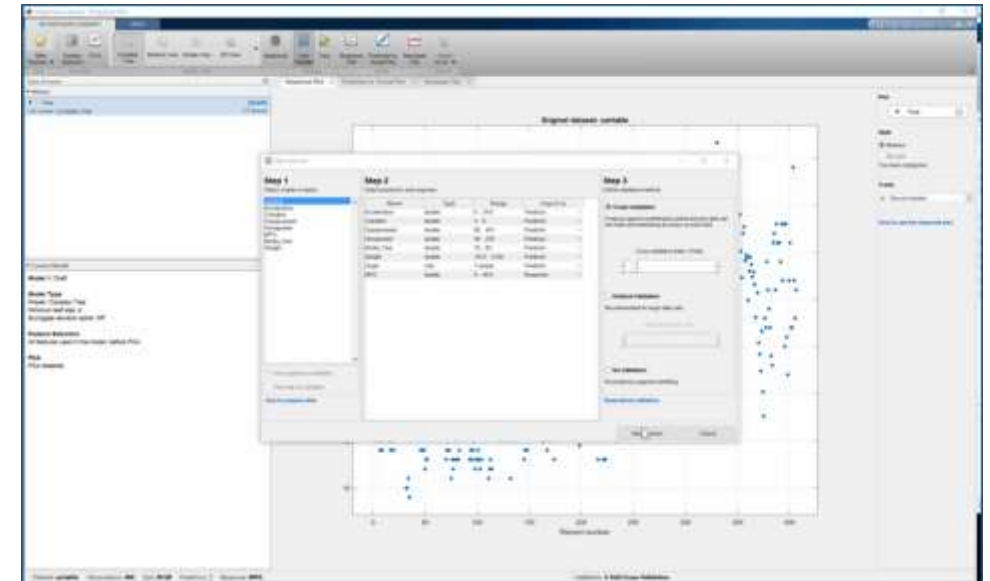
Name	Type	Range	Import as
Acceleration	double	0 - 24.8	Predictor
Cylinders	double	3 - 8	Predictor
Displacement	double	68 - 455	Predictor
Horsepower	double	46 - 230	Predictor
Model_Year	double	70 - 82	Predictor
Weight	double	1032 - 5140	Predictor
MPG	double	7 - 46.0	Response
- Step 3: Define validation method** - Offers three options:
  - Cross-Validation** (selected): Predicts against overfitting by partitioning the data set into folds and estimating accuracy on each fold. Includes a 'Cross-validated folds' input field set to 5.
  - Holdout Validation**: Recommended to large data sets. Includes an 'Input holdout ratio' input field set to 0.25.
  - No Validation**: No prediction against overfitting.

The background shows a scatter plot titled 'Original dataset: carmat' with 'Record number' on the x-axis (0 to 400) and 'MPG' on the y-axis (0 to 10). The status bar at the bottom indicates: 'Dataset: carmat Observations: 406 Size: 30 KB Predictors: 7 Response: MPG Validation: 5-fold Cross-Validation'.

# Regression Learner

App to apply advanced regression methods to your data

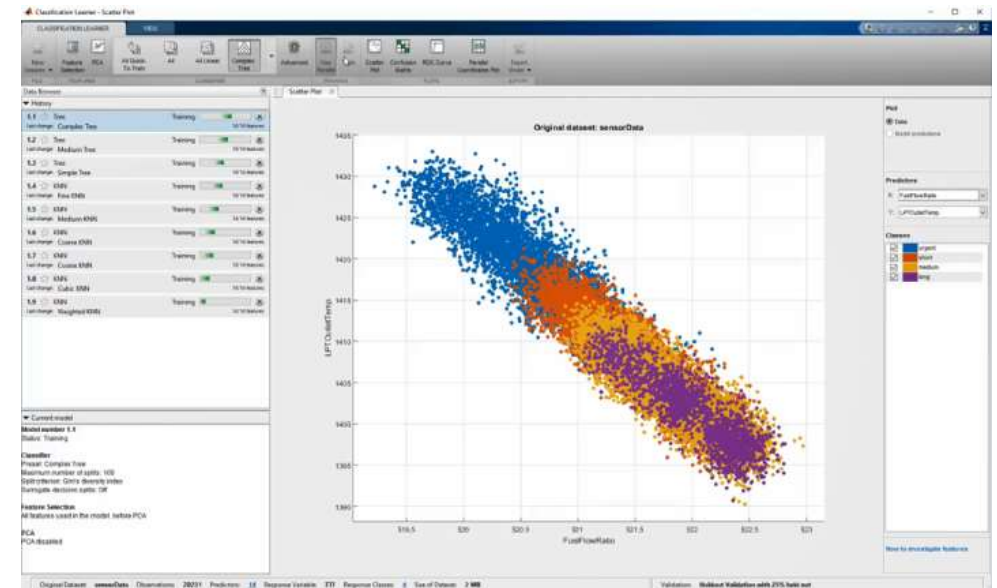
- Added to Statistics and Machine Learning Toolbox in R2017a
- Point and click interface – no coding required
- Quickly evaluate, compare and select regression models
- Export and share MATLAB code or trained models



# Classification Learner

App to apply advanced classification methods to your data

- Added to Statistics and Machine Learning Toolbox in R2015a
- Point and click interface – no coding required
- Quickly evaluate, compare and select classification models
- Export and share MATLAB code or trained models



# and Many More MATLAB Apps for Data Analytics

Distribution Fitting

System Identification

Signal Analysis

Wavelet Design and Analysis

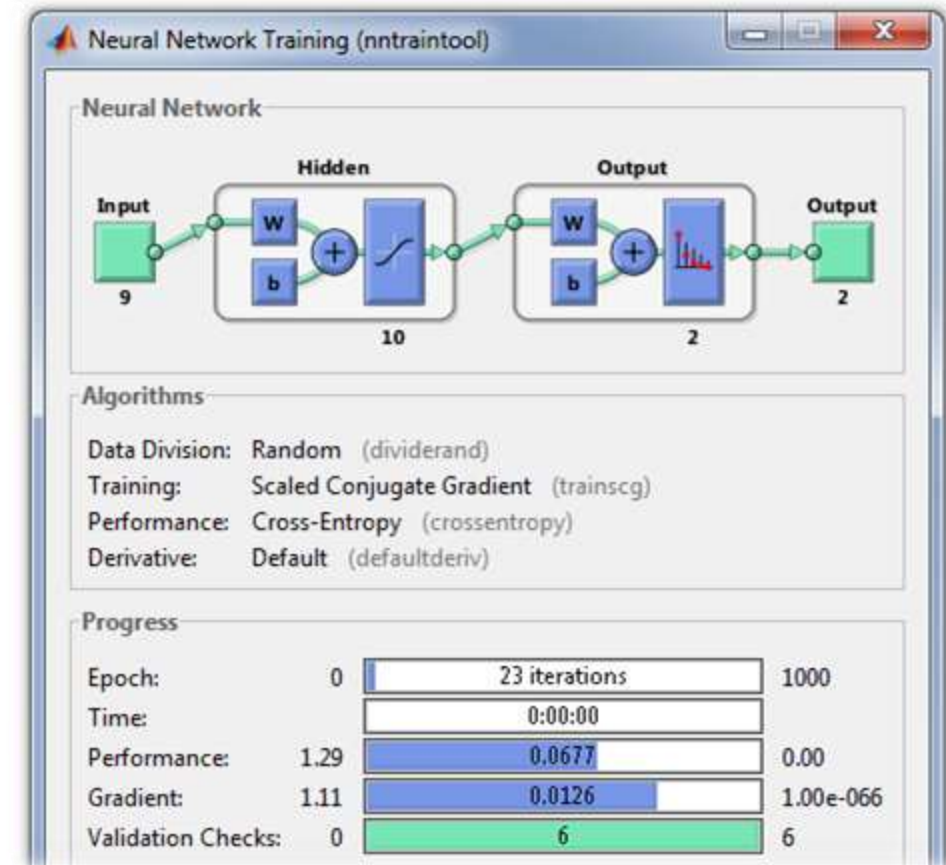
Neural Net Fitting

Neural Net Pattern Recognition

Training Image Labeler

*and many more...*

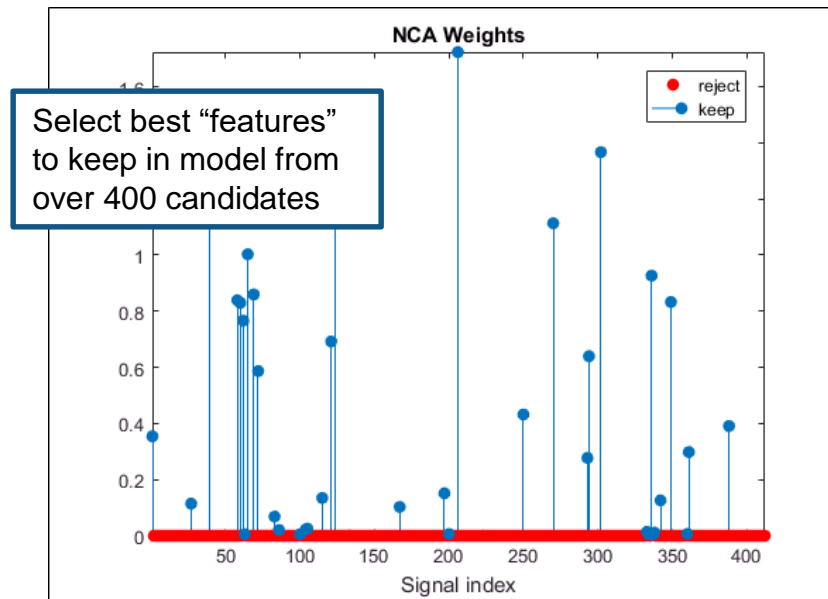
MATLAB EXPO 2017



# Tuning Machine Learning Models

Get more accurate models in less time

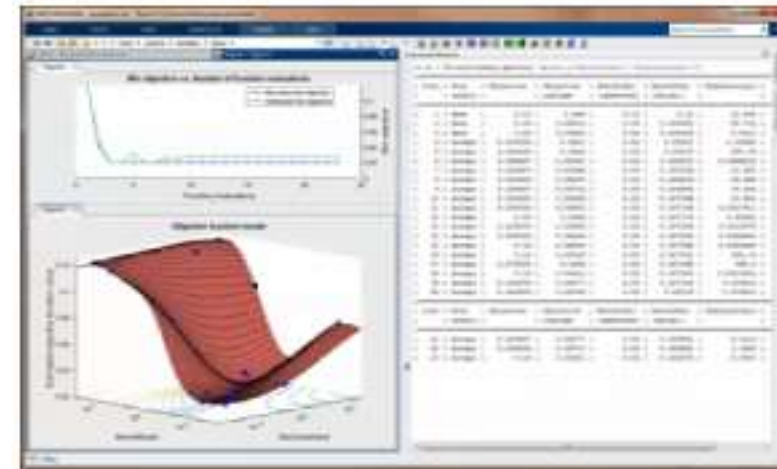
**Automatically** select best machine learning “features”



**R2016b**

NCA: Neighborhood Component Analysis

**Automatically** fine-tune machine learning parameters



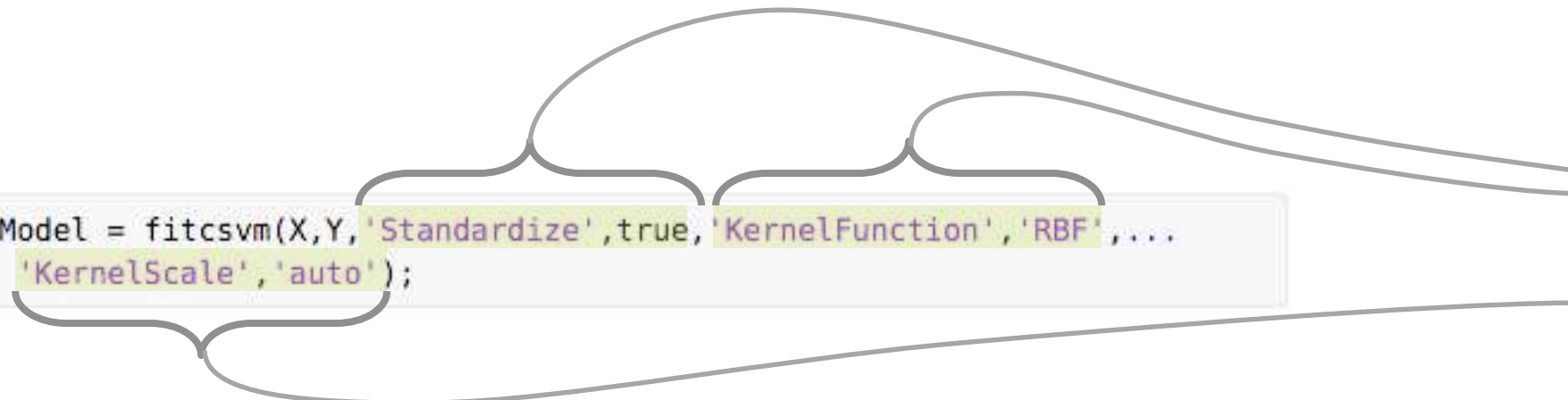
**R2016b**

Hyperparameter Tuning



# Machine Learning Hyperparameters

```
SVMModel = fitcsvm(X,Y,'Standardize',true,'KernelFunction','RBF',...  
    'KernelScale','auto');
```



Hyperparameters

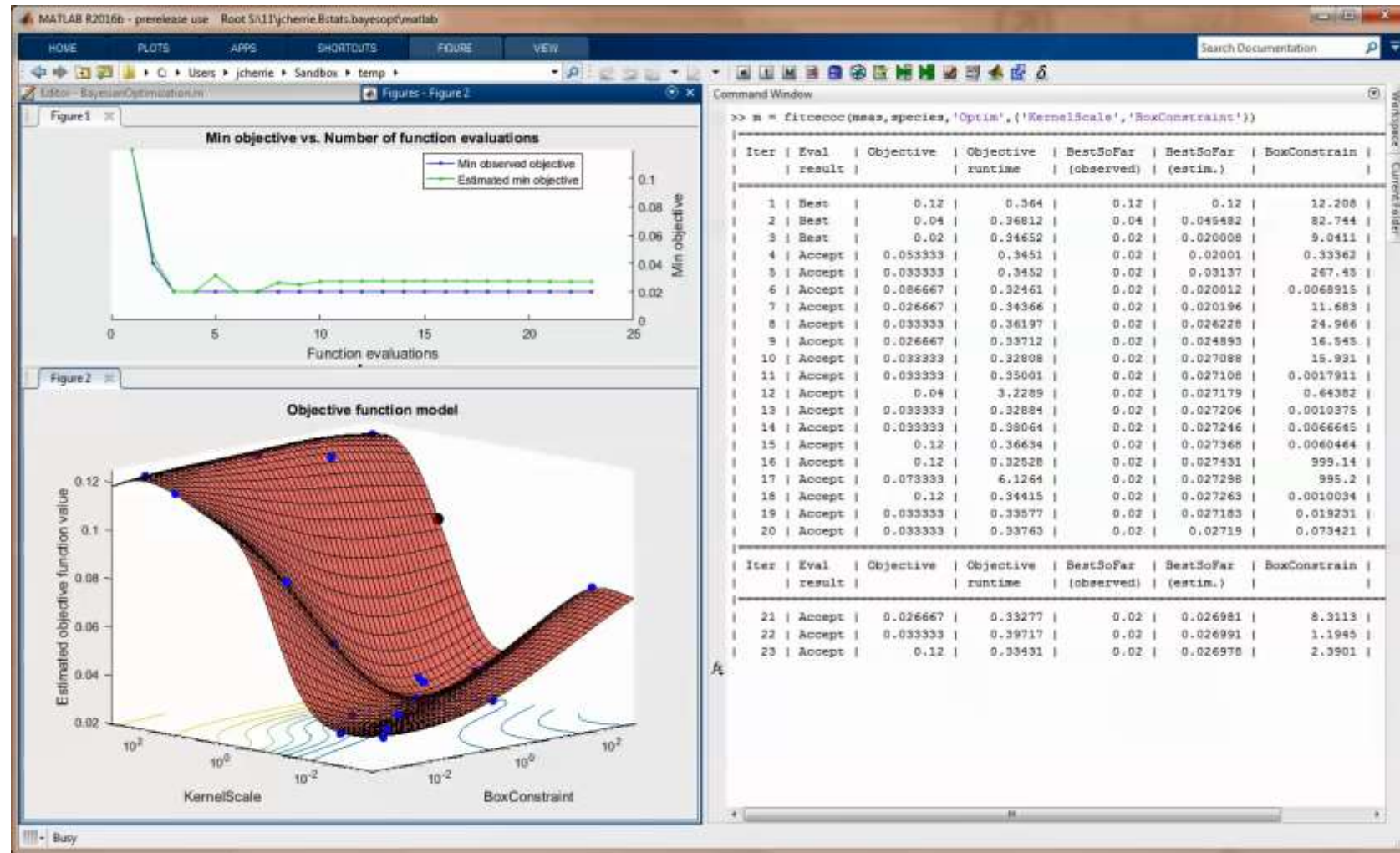
```
SVMModel = fitcsvm(X,Y,'OptimizeHyperparameters','auto');
```

Tune a typical set of hyperparameters for this model

```
SVMModel = fitcsvm(X,Y,'OptimizeHyperparameters','all');
```

Tune all hyperparameters for this model

# Bayesian Optimization in Action



# Big Data Analytics Workflow: Developing Predictive models

2

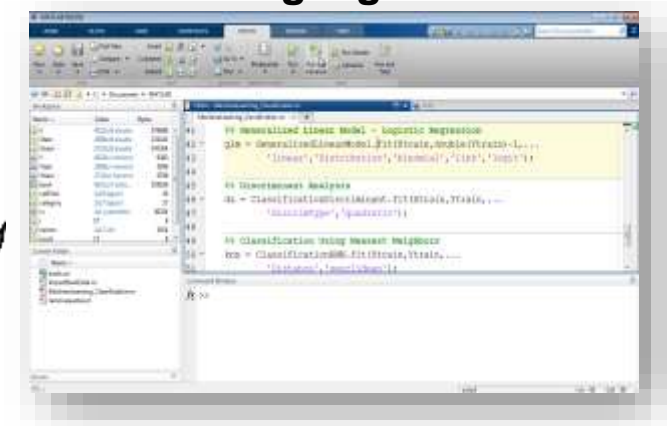
**MATLAB enables domain experts to do Data Science**

- ### Challenges
- Lack of data science expertise
  - Feature Extraction – How to transform data to best represent the system?
    - Requires subject matter expertise
    - No right way of designing features
  - Feature Selection – What attributes or subset of data to use?
    - Entails a lot of iteration – Trial and error
    - Difficult to evaluate features
  - Model Development
    - Many different models
    - Model Validation and Tuning
  - Time required to conduct the analysis

Apps



Language



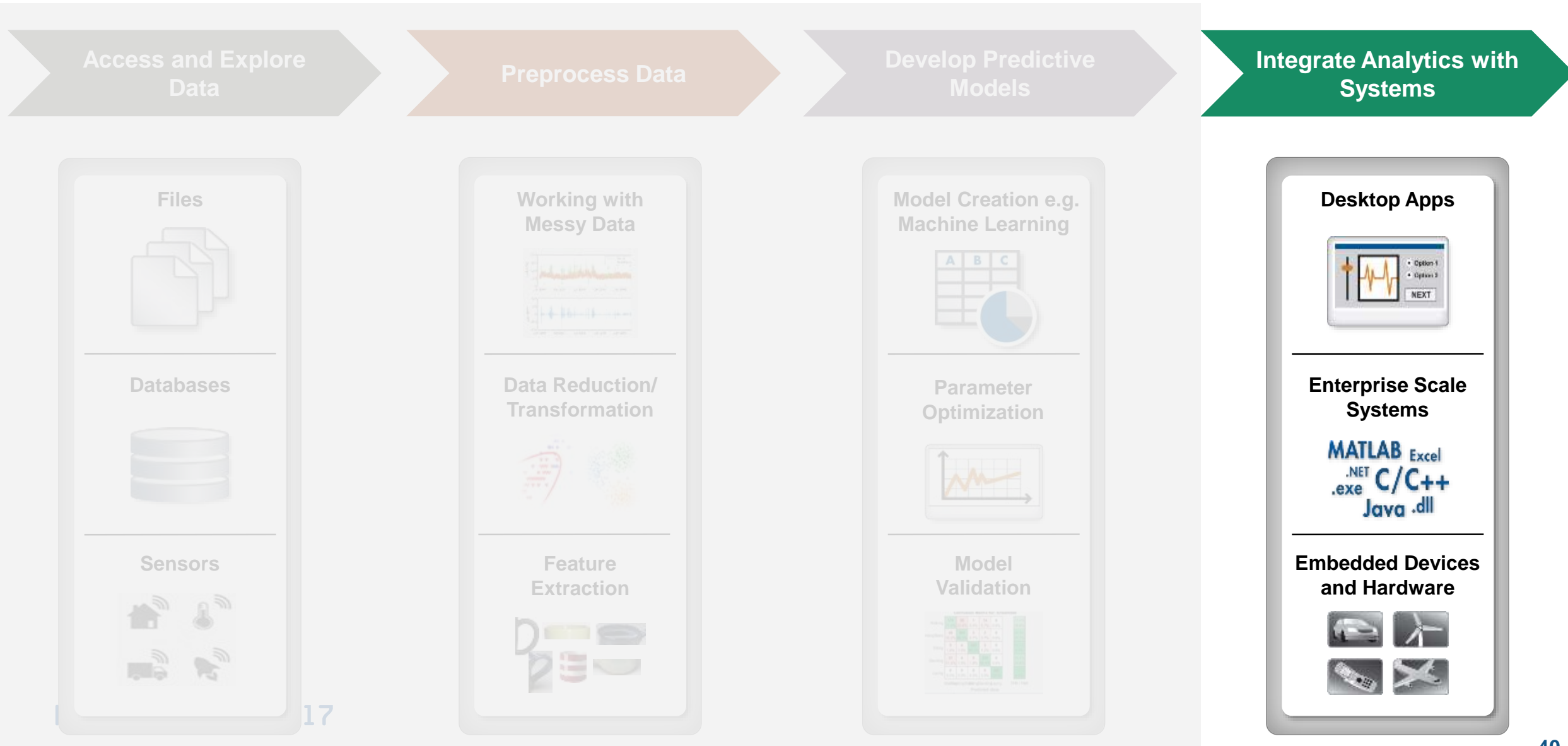
- Easy to use apps
- Wide breadth of tools to facilitate domain specific analysis
- Examples/videos to get started
- Automatic MATLAB code generation
- High speed processing of large data sets

## Back to our example: Working with Big Data in MATLAB

- **Objective:** Create a model to predict the cost of a taxi ride in New York City
- **Inputs:**
  - Monthly taxi ride log files
  - The local data set is **small** (~20 MB)
  - The full data set is **big** (~25 GB)
- **Approach:**
  - Access Data
  - Preprocess and explore data
  - Develop and validate predictive model (linear fit)
    - Work with subset of data for prototyping
    - Scale to full data set on a cluster

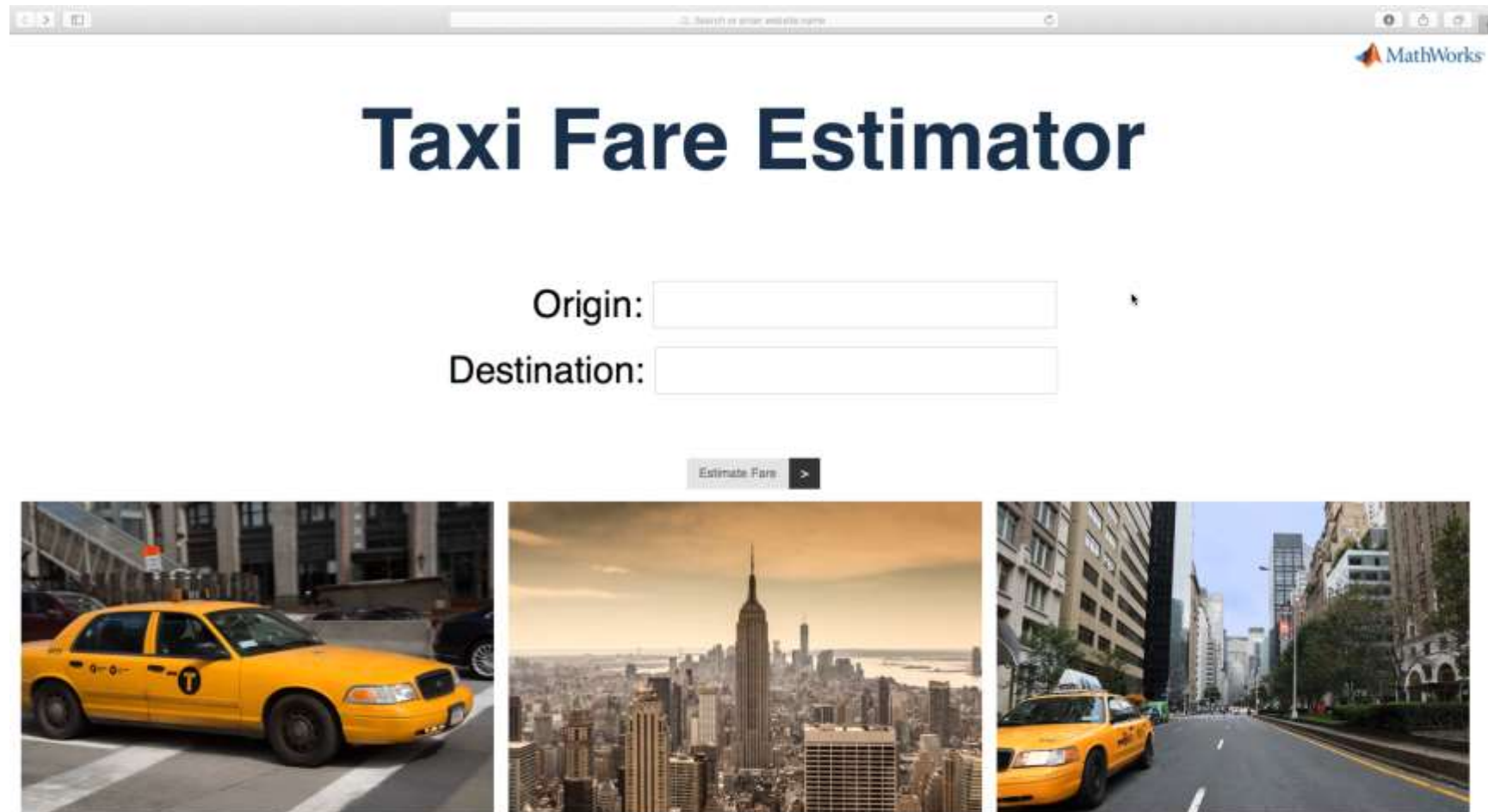


# Data Analytics Workflow: Develop Predictive Models using **Big Data**







# Demo: Taxi Fare Predictor Web App



Origin:

Destination:

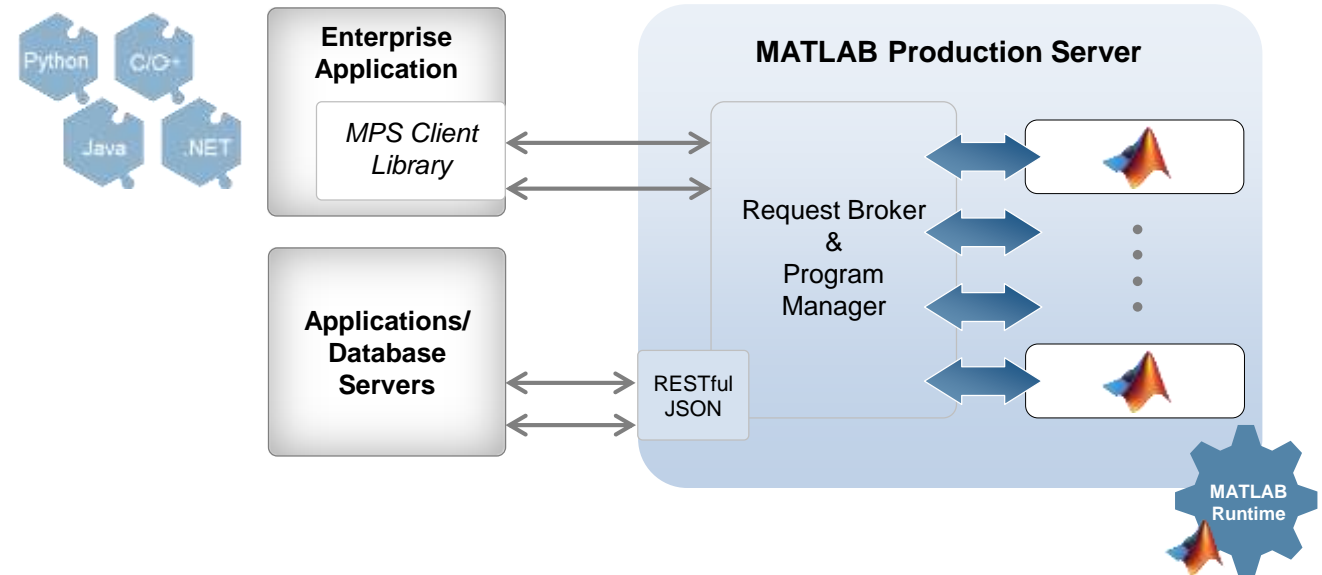
Estimate Fare 





# MATLAB Production Server

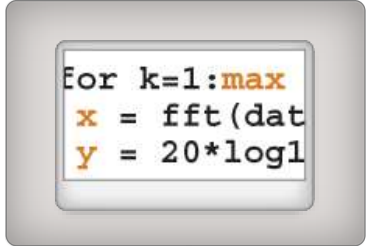
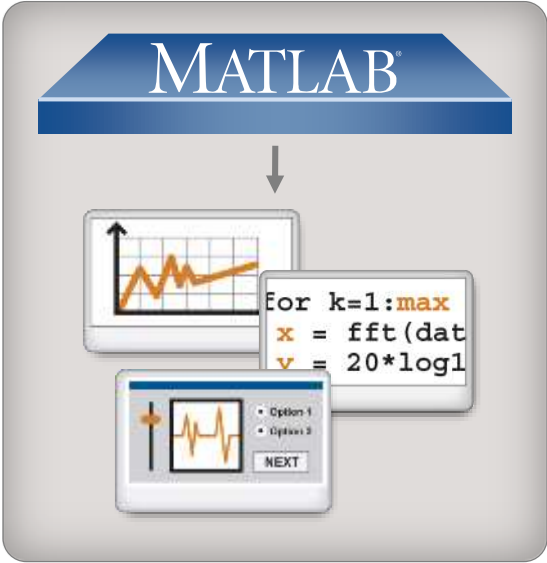
- Server software
  - Manages packaged MATLAB programs and worker pool
  
- MATLAB Runtime libraries
  - Single server can use runtimes from different releases
  
- RESTful JSON interface
  
- Lightweight client libraries
  - C/C++, .NET, Python, and Java



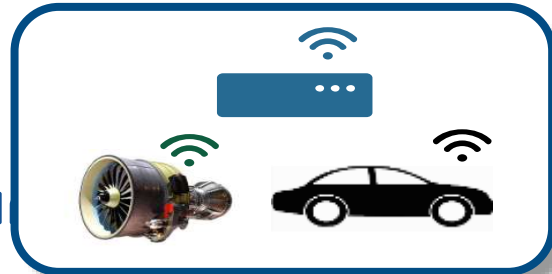
# Integrate analytics with systems

3

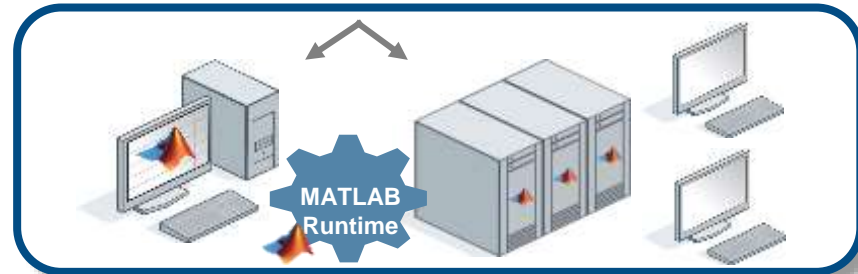
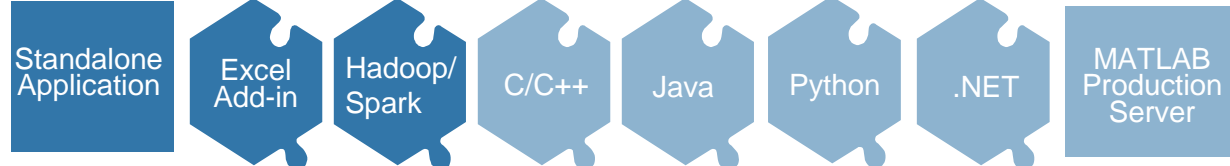
**MATLAB Analytics**  
run anywhere



## Embedded Hardware



## Enterprise Systems

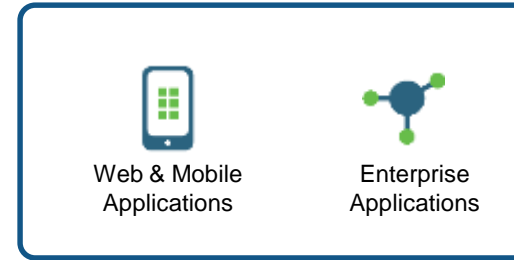


M

# Product Support for Spark

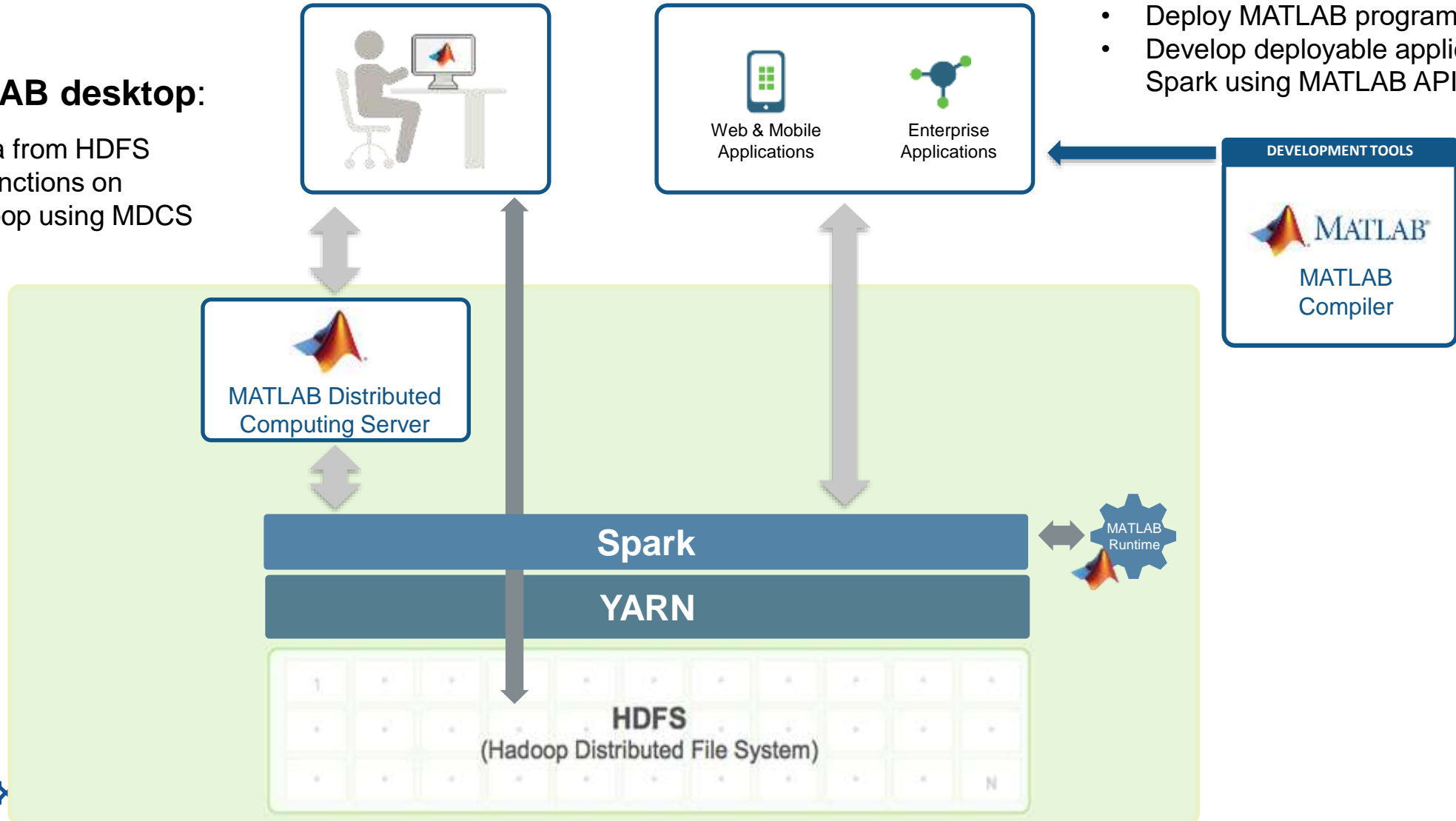
## From MATLAB desktop:

- Access data from HDFS
- Run “tall” functions on Spark/Hadoop using MDCS



## Integrate with applications:

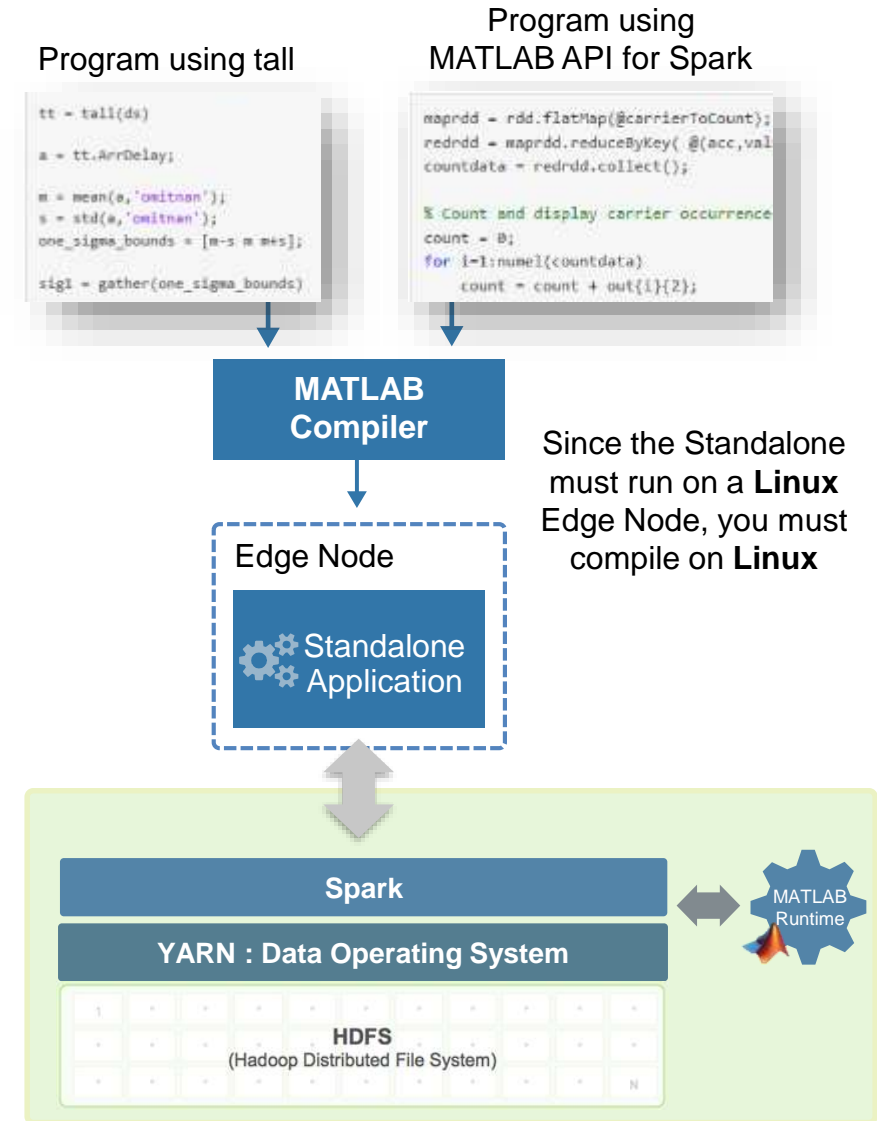
- Deploy MATLAB programs using “tall”
- Develop deployable applications for Spark using MATLAB API for Spark



MATLAB EX

# Deployment Offerings

- Deploy “tall” programs
  - Create Standalone Applications: MATLAB Compiler
  
- MATLAB API for Spark
  - Create Standalone Applications: MATLAB Compiler
  - Functionality beyond tall arrays
  - For advanced programmers familiar with Spark
  - Local install of Spark to run code in MATLAB
    - Installed on same machine as MATLAB – single node, Linux



# Data Analytics Workflow

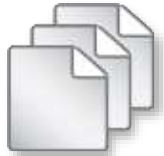
Access and Explore Data

Preprocess Data

Develop Predictive Models

Integrate Analytics with Systems

Files



Databases



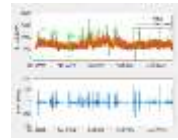
Sensors



MATLAB Analytics work with **business and engineering data**

1

Working with Messy Data



Data Reduction/Transformation



Feature Selection



MATLAB enables **domain experts to do Data Science**

2

Model Creation e.g. Machine Learning



Parameter Optimization



Model

Desktop Apps



Enterprise Scale Systems

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

Embedded Devices

MATLAB Analytics **run anywhere**

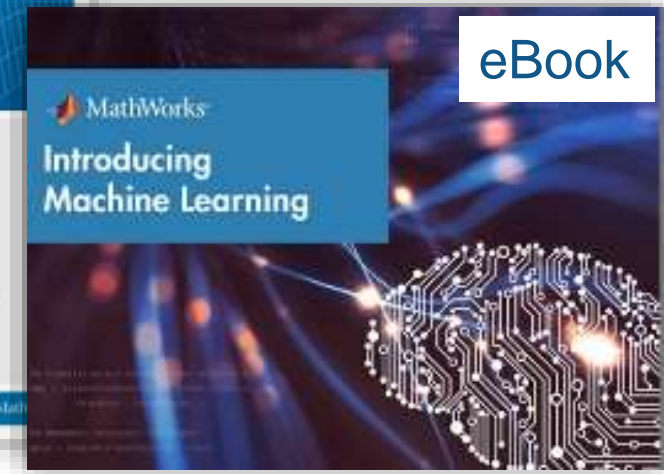
3

# Resources to learn and get started

[mathworks.com/big-data](http://mathworks.com/big-data)



[mathworks.com/machine-learning](http://mathworks.com/machine-learning)





# MathWorks Services

- Consulting

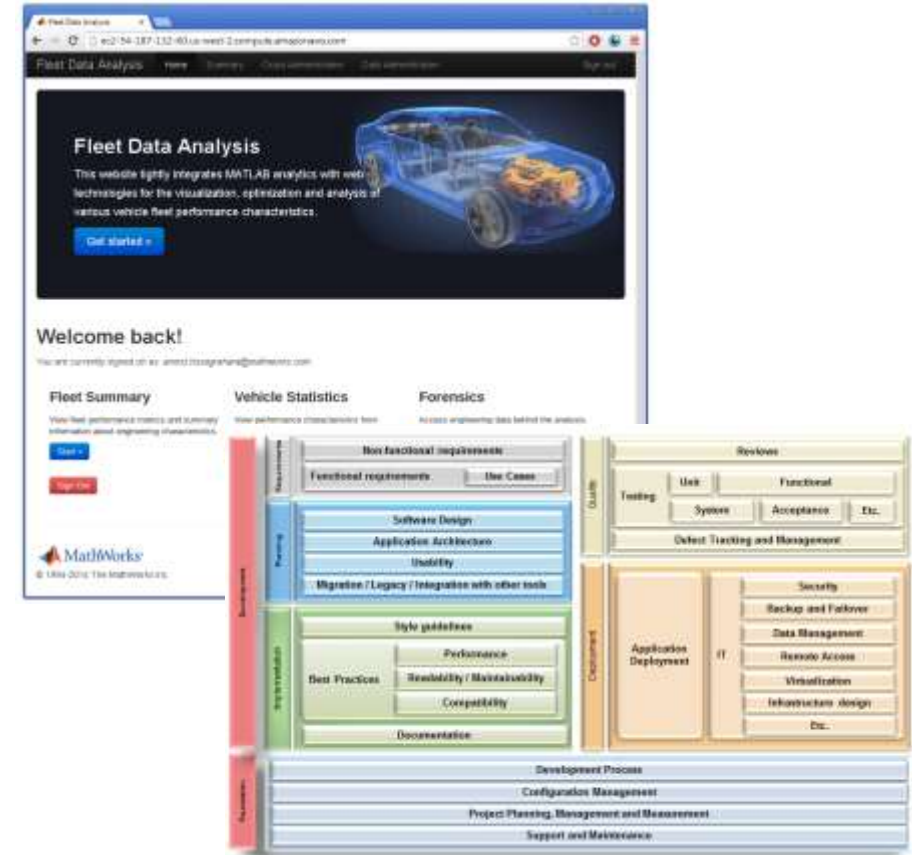
- Integration
- Data analysis/visualization
- Unify workflows, models, data

[www.mathworks.com/services/consulting/](http://www.mathworks.com/services/consulting/)

- Training

- Classroom, online, on-site
- Data Processing, Visualization, Deployment, Parallel Computing

[www.mathworks.com/services/training/](http://www.mathworks.com/services/training/)



# MathWorks Training Offerings

## Machine Learning with MATLAB

---

INTERMEDIATE

This two-day course focuses on data analytics and machine learning techniques in MATLAB using functionality within Statistics and Machine Learning Toolbox™ and Neural Network Toolbox™.

The course demonstrates the use of unsupervised learning to discover features in large data sets and supervised learning to build predictive models. Examples and exercises highlight techniques for visualization and evaluation of results. Topics include:

- Importing and organizing data
- Finding natural patterns in data
- Building predictive models
- Evaluating and improving the model

**Prerequisites:** *MATLAB Fundamentals*

## Interfacing MATLAB with C Code

---

INTERMEDIATE

This one-day course covers details of interfacing MATLAB with user-written C code. Topics include:

- Source MEX-files
- Data exchange between MATLAB and MEX-files
- The MATLAB engine interface

**Prerequisites:** *MATLAB Fundamentals* and a basic working knowledge of the C programming language

<http://www.mathworks.com/services/training/>



*Accelerating the pace of engineering and science*

## Speaker Details

### Email:

[seth.deland@mathworks.com](mailto:seth.deland@mathworks.com)

[amit.doshi@mathworks.in](mailto:amit.doshi@mathworks.in)

### LinkedIn:

<https://in.linkedin.com/in/amit-doshi>

<https://www.linkedin.com/in/seth-deland>

## Contact MathWorks India

Products/Training Enquiry Booth

Call: 080-6632-6000

Email: [info@mathworks.in](mailto:info@mathworks.in)

**Your feedback is valued.**

**Please complete the feedback form provided to you.**