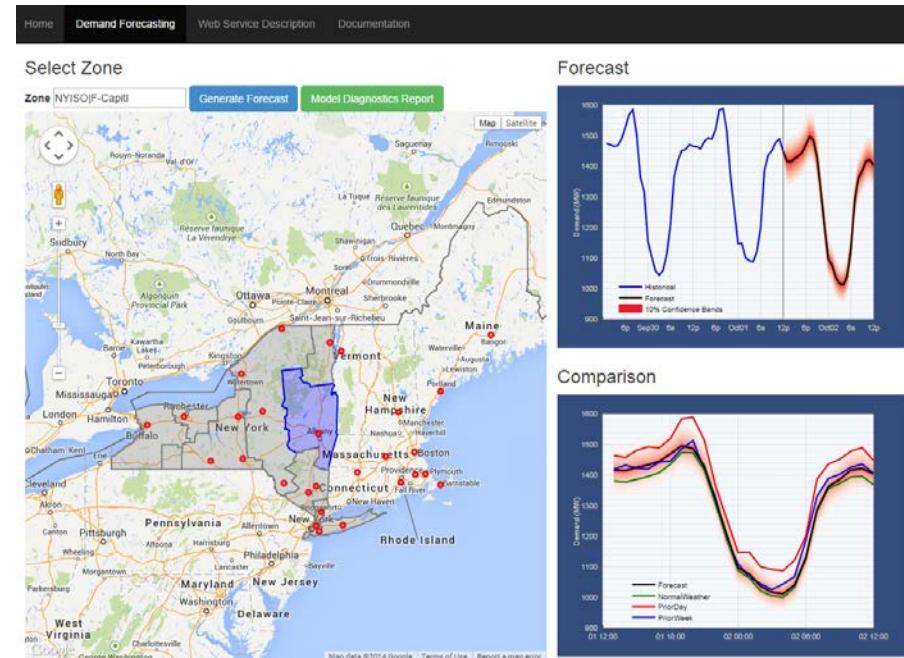
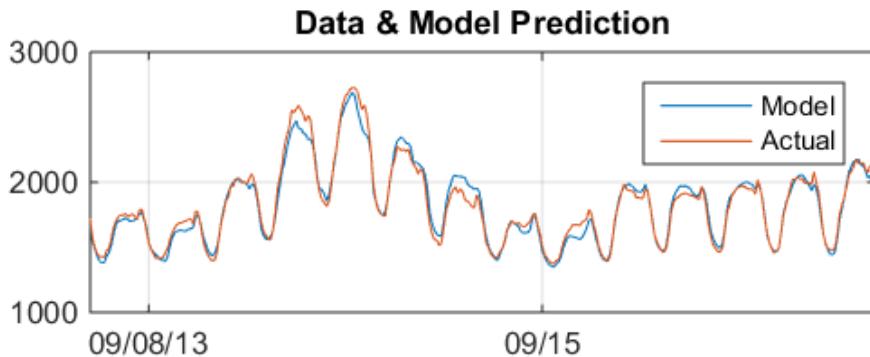


# Data Analytics with MATLAB

**Adam Filion**  
**Application Engineer**  
**MathWorks**

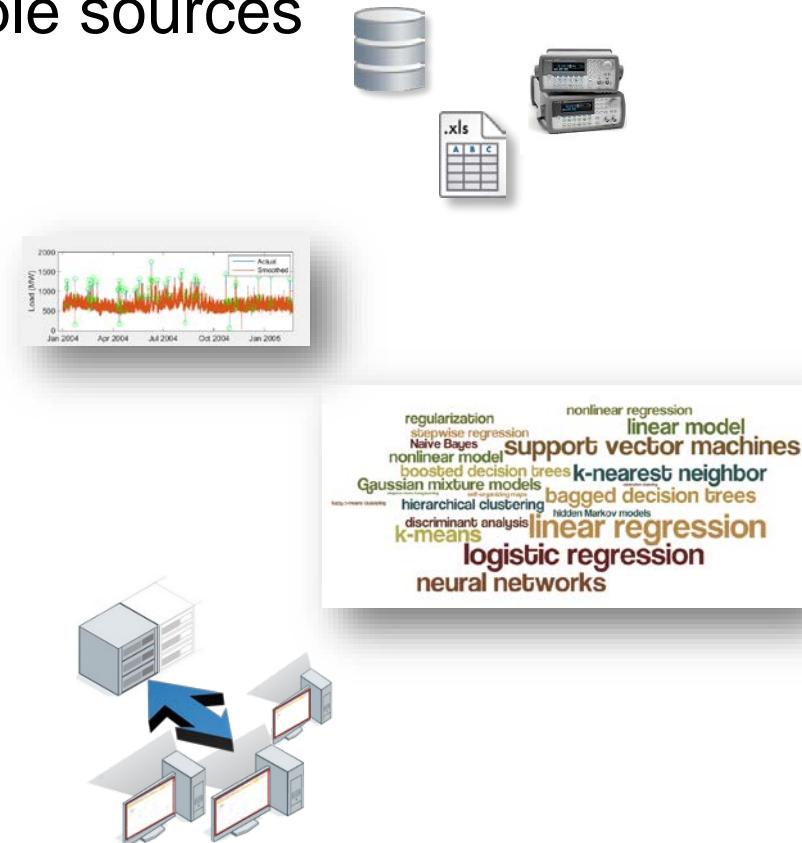
# Case Study: Day-Ahead Load Forecasting

- Goal:
  - Implement a tool for **easy** and **accurate** computation of day-ahead system load forecast
  
- Requirements:
  - Acquire and clean data from multiple sources
  - Accurate predictive model
  - Easily deploy to production environment



# Challenges with Data Analytics

- Aggregating data from multiple sources
- Cleaning data
- Choosing a model
- Moving to production



# NYISO Energy Load Data

[mis.nyiso.com/public/](http://mis.nyiso.com/public/)



**OASIS (Open Access Same-Time Information System)**

[NYISO Reference Bus LBMP](#) P-28  
[NYISO Price Correction Logs](#) P-29

**Power Grid Data**

[Outages](#)  
[Real-Time Scheduled Outages](#) P-54A  
[Real-Time Actual Outages](#) P-54B  
[Day-Ahead Scheduled Outages](#) P-54C  
[Outage Schedules](#) P-14  
[Outage Schedules CSV](#) P-14B  
[Generation Maintenance Report](#) P-15

[Constraints](#)  
[Day-Ahead Limiting Constraints](#) P-51A  
[Limiting Constraints](#) P-33

[Interface Flows](#)  
[Internal & External Interface Limits & Flows](#) P-32  
[Lake Erie Circulation - Day-Ahead](#) P-53B  
[Lake Erie Circulation - Real-Time](#) P-34A

[PARs](#)  
[PAR Schedules](#) P-53A  
[PAR Flows](#) P-34

[ATC/TTC](#)  
[ATC/TTC](#) P-8  
[Long Term ATC/TTC](#) P-8A  
[Transfer Limitations](#)

**Load Data**

[Load Forecast/Commitment](#)  
[ISO Load Forecast](#) P-7  
[Zonal Bid Load](#) P-59  
[Weather Forecast](#) P-7A

**Actual Load** (highlighted)

[Real-Time Actual Load](#) P-58B  
[Integrated Real-Time Actual Load](#) P-58C

**Real-Time Actual Load**

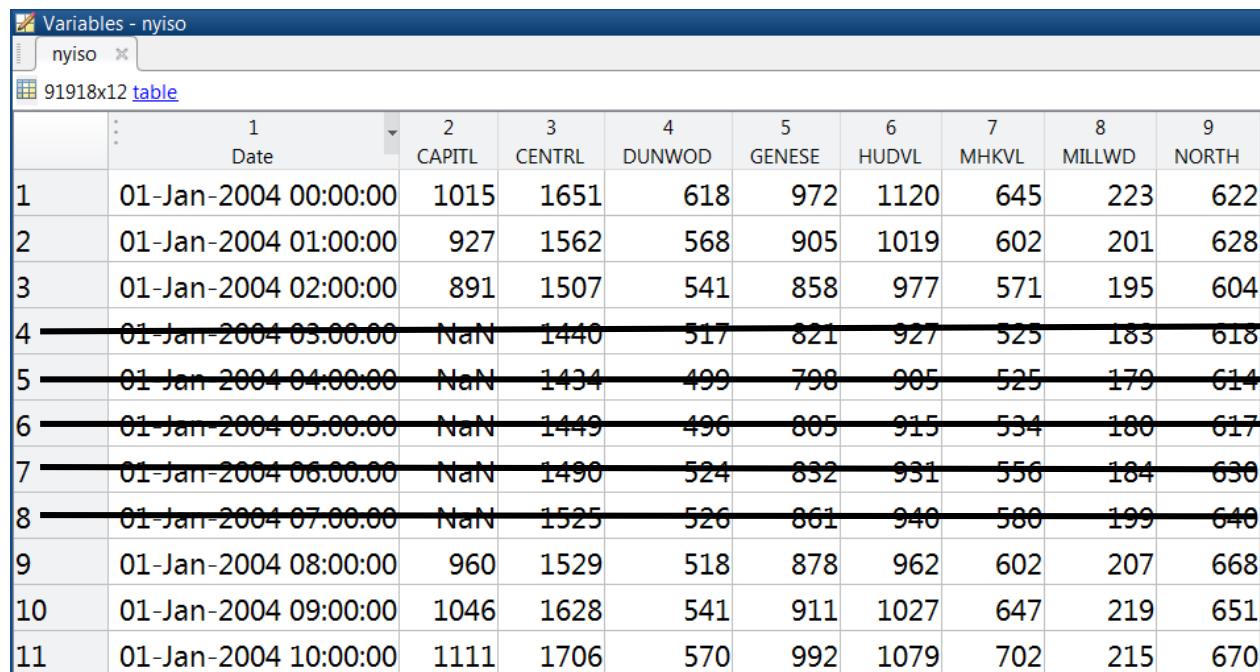
CSV Files	Last Updated
<a href="#">10-21-2014</a>	10/21/14 23:02 EDT
<a href="#">10-20-2014</a>	10/21/14 00:07 EDT
<a href="#">10-19-2014</a>	10/20/14 00:01 EDT
<a href="#">10-18-2014</a>	10/18/14 23:59 EDT
<a href="#">10-17-2014</a>	10/18/14 00:00 EDT
<a href="#">10-16-2014</a>	10/16/14 23:59 EDT
<a href="#">10-15-2014</a>	10/15/14 23:59 EDT
<a href="#">10-14-2014</a>	10/14/14 23:59 EDT
<a href="#">10-13-2014</a>	10/13/14 23:59 EDT
<a href="#">10-12-2014</a>	10/12/14 23:59 EDT

**Archived Files (zip format)**

CSV Files	Last Updated
<a href="#">10-2014</a>	10/21/14 23:02 EDT
<a href="#">09-2014</a>	09/30/14 23:59 EDT
<a href="#">08-2014</a>	09/01/14 00:01 EDT
<a href="#">07-2014</a>	08/01/14 00:00 EDT
<a href="#">06-2014</a>	07/01/14 00:00 EDT
<a href="#">05-2014</a>	06/01/14 00:00 EDT
<a href="#">04-2014</a>	04/30/14 23:59 EDT
<a href="#">03-2014</a>	03/31/14 23:59 EDT
<a href="#">02-2014</a>	02/28/14 23:58 EST
<a href="#">01-2014</a>	01/31/14 23:59 EST

# Techniques to Handle Missing Data

- List-wise deletion
  - Unbiased estimates
  - Reduces sample size
- Implementation options
  - Built in to many MATLAB functions
  - Manual filtering



	1	2	3	4	5	6	7	8	9
	Date	CAPITL	CENTRL	DUNWOD	GENESE	HUDVL	MHKVL	MILLWD	NORTH
1	01-Jan-2004 00:00:00	1015	1651	618	972	1120	645	223	622
2	01-Jan-2004 01:00:00	927	1562	568	905	1019	602	201	628
3	01-Jan-2004 02:00:00	891	1507	541	858	977	571	195	604
4	01-Jan-2004 03:00:00	NaN	1440	517	821	927	525	183	618
5	01-Jan-2004 04:00:00	NaN	1434	499	798	905	525	179	614
6	01-Jan-2004 05:00:00	NaN	1449	496	805	915	534	180	617
7	01-Jan-2004 06:00:00	NaN	1490	524	832	931	556	184	630
8	01-Jan-2004 07:00:00	NaN	1525	526	861	940	580	199	640
9	01-Jan-2004 08:00:00	960	1529	518	878	962	602	207	668
10	01-Jan-2004 09:00:00	1046	1628	541	911	1027	647	219	651
11	01-Jan-2004 10:00:00	1111	1706	570	992	1079	702	215	670

# Techniques to Handle Missing Data

Substitution – replace missing data points with a reasonable approximation

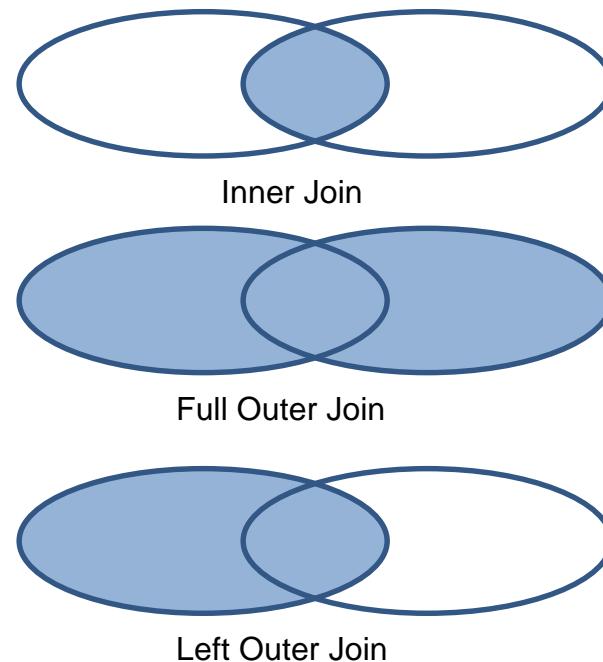
Easy to model

Too important to exclude

	Date	CAPITL	CENTRL	DUNWOD	GENESE	HUDVL	MHKVL	MILLWD	NORTH
1	01-Jan-2004 00:00:00	1015	1651	618	972	1120	645	223	622
2	01-Jan-2004 01:00:00	927	1562	568	905	1019	602	201	628
3	01-Jan-2004 02:00:00	891	1507	541	858	977	571	195	604
4	01-Jan-2004 03:00:00	NaN	1440	517	821	927	525	183	618
5	01-Jan-2004 04:00:00	NaN	1434	499	798	905	525	179	614
6	01-Jan-2004 05:00:00	NaN	1449	496	805	915	534	180	617
7	01-Jan-2004 06:00:00	NaN	1490	524	832	931	556	184	630
8	01-Jan-2004 07:00:00	NaN	1525	526	861	940	580	199	640
9	01-Jan-2004 08:00:00	960	1529	518	878	962	602	207	668
10	01-Jan-2004 09:00:00	1046	1628	541	911	1027	647	219	651
11	01-Jan-2004 10:00:00	1111	1706	570	992	1079	702	215	670

# Merge Different Sets of Data

- Join along a common axis
- Popular Joins:
  - Inner
  - Full Outer
  - Left Outer
  - Right Outer



# Full Outer Join



Key	B
1	1.1
4	1.4
7	1.7
9	1.9

First Data Set

Key	Y	Z
1	0.1	0.2
3	0.3	0.4
5	0.5	0.6
7	0.7	0.8

Second Data Set



Key	B	Y	Z
1	1.1	0.1	0.2
3	NaN	0.3	0.4
4	1.4	NaN	NaN
5	NaN	0.5	0.6
7	1.7	0.7	0.8
9	1.9	NaN	NaN

Joined Data Set

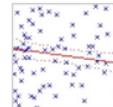
# Learn More: Big Data with MATLAB

[www.mathworks.com/discovery/big-data-matlab.html](http://www.mathworks.com/discovery/big-data-matlab.html)

[www.mathworks.com/discovery/matlab-mapreduce-hadoop.html](http://www.mathworks.com/discovery/matlab-mapreduce-hadoop.html)

## MapReduce on the Desktop

Explore and analyze big data sets on your desktop with the MapReduce programming technique built into MATLAB.

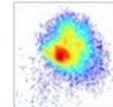


Creating algorithms using MapReduce: [max](#), [mean](#), [mean by group](#), [histograms](#), [covariance](#) and related quantities, [summary statistics by group](#), [logistic regression](#), [tall skinny QR](#)

- » [Get started with MATLAB MapReduce](#)
- » [MapReduce design patterns](#)
- » [Use MATLAB MapReduce with relational databases](#)

## MapReduce on Hadoop

Execute MATLAB MapReduce based algorithms within Hadoop MapReduce to explore and analyze data that is stored and managed on Hadoop, using MATLAB Distributed Computing Server.



- » [Run MATLAB MapReduce on Hadoop](#)

Create applications and libraries based upon MATLAB MapReduce for deployment within production instances of Hadoop, using MATLAB Compiler.

- » [Deploy MATLAB MapReduce applications to Hadoop](#)

## MATLAB MapReduce and Hadoop

```
ds = datastore('airline.csv'); % Create a datastore
ds.SelectedVariableNames = 'ArrDelay';
maxDelay = mapreduce(ds, @maxArrivalDelay, ...
    @maxArrivalDelayReducer);
readall(maxDelay)
```

MATLAB® has numerous capabilities for exploring and analyzing big data sets. Among them is MapReduce, a powerful, and established programming technique for applying filtering, statistics and other general analysis methods to big data.

The MapReduce functionality built into MATLAB lets you analyze data that does not fit into memory. By running your MapReduce based algorithms in parallel (using [Parallel Computing Toolbox™](#)), you can better utilize the processing resources on your desktop without changing your algorithms.

To analyze data in MATLAB using MapReduce:

1. Specify the data you want to analyze using [datastore](#)
2. Create your map and reduce functions in MATLAB
3. Execute your map and reduce functions using [mapreduce](#)

While MATLAB MapReduce is optimized for array-based analysis, it is fully compatible with Hadoop MapReduce, so you can run your MapReduce based algorithms within the Hadoop MapReduce framework:

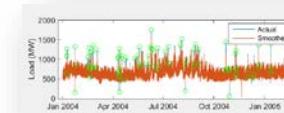
- Execute MapReduce based algorithms on Hadoop directly from the MATLAB desktop, using [MATLAB Distributed Computing Server™](#)
- Package MapReduce based algorithms for deploying to production Hadoop systems, using [MATLAB Compiler™](#)

# Challenges with Data Analytics

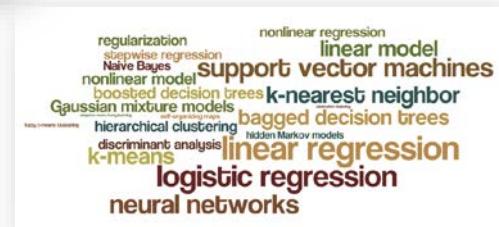
- ✓ Aggregating data from multiple sources



- ✓ Cleaning data



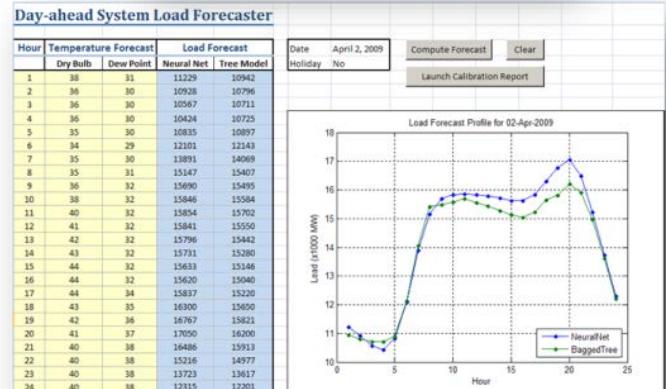
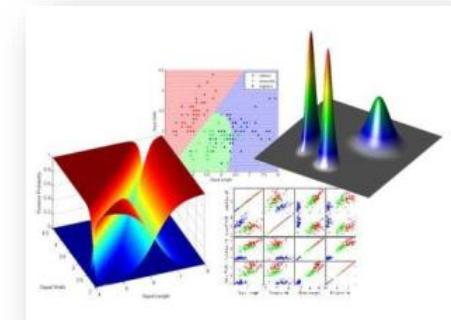
- Choosing a model
- Moving to production



# Machine Learning

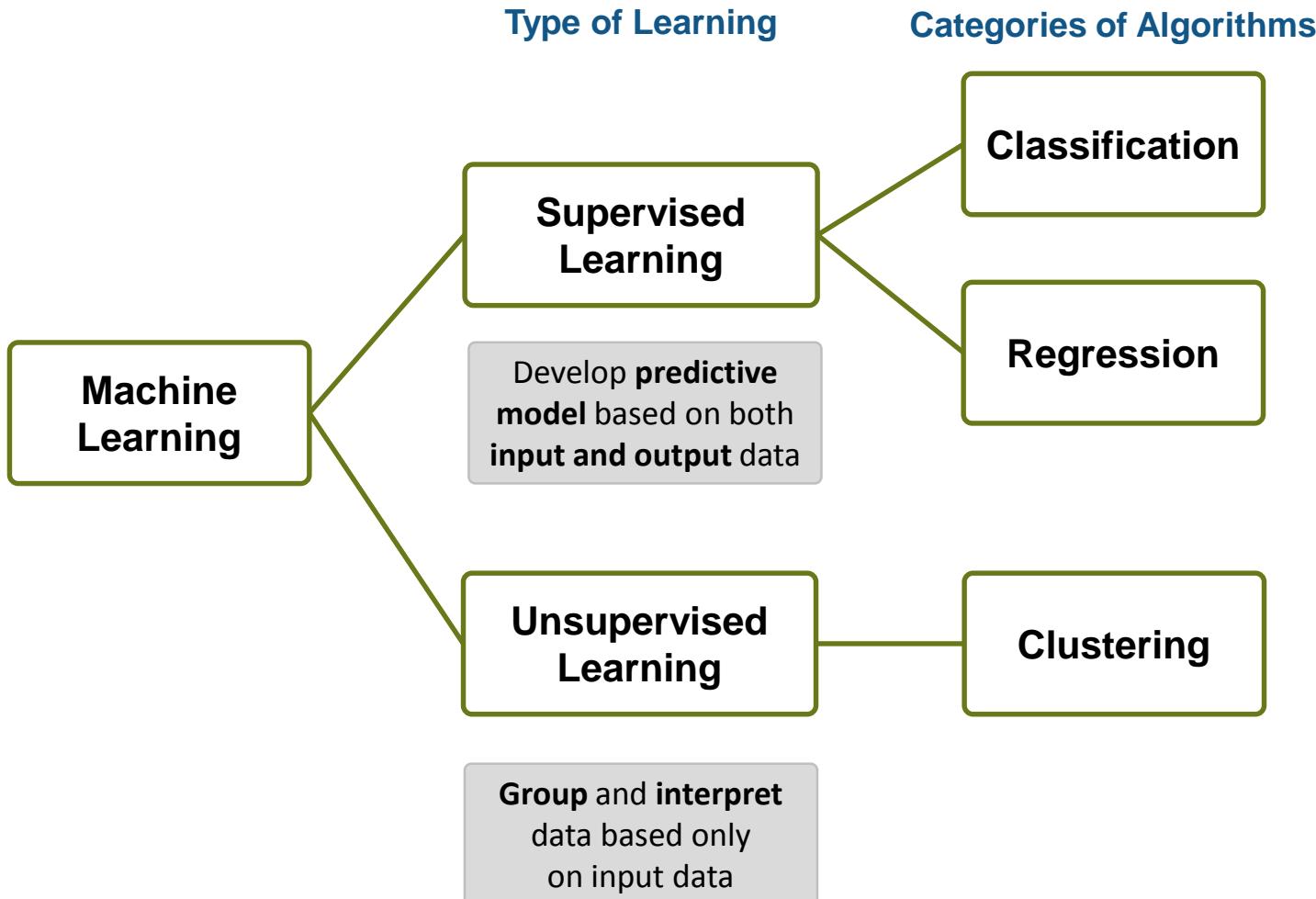
## Characteristics and Examples

- Characteristics
  - Lots of variables
  - System too complex to know the governing equation (e.g., *black-box modeling*)
- Examples
  - Pattern recognition (*speech, images*)
  - Financial algorithms (*credit scoring, algo trading*)
  - Energy forecasting (*load, price*)
  - Biology (*tumor detection, drug discovery*)

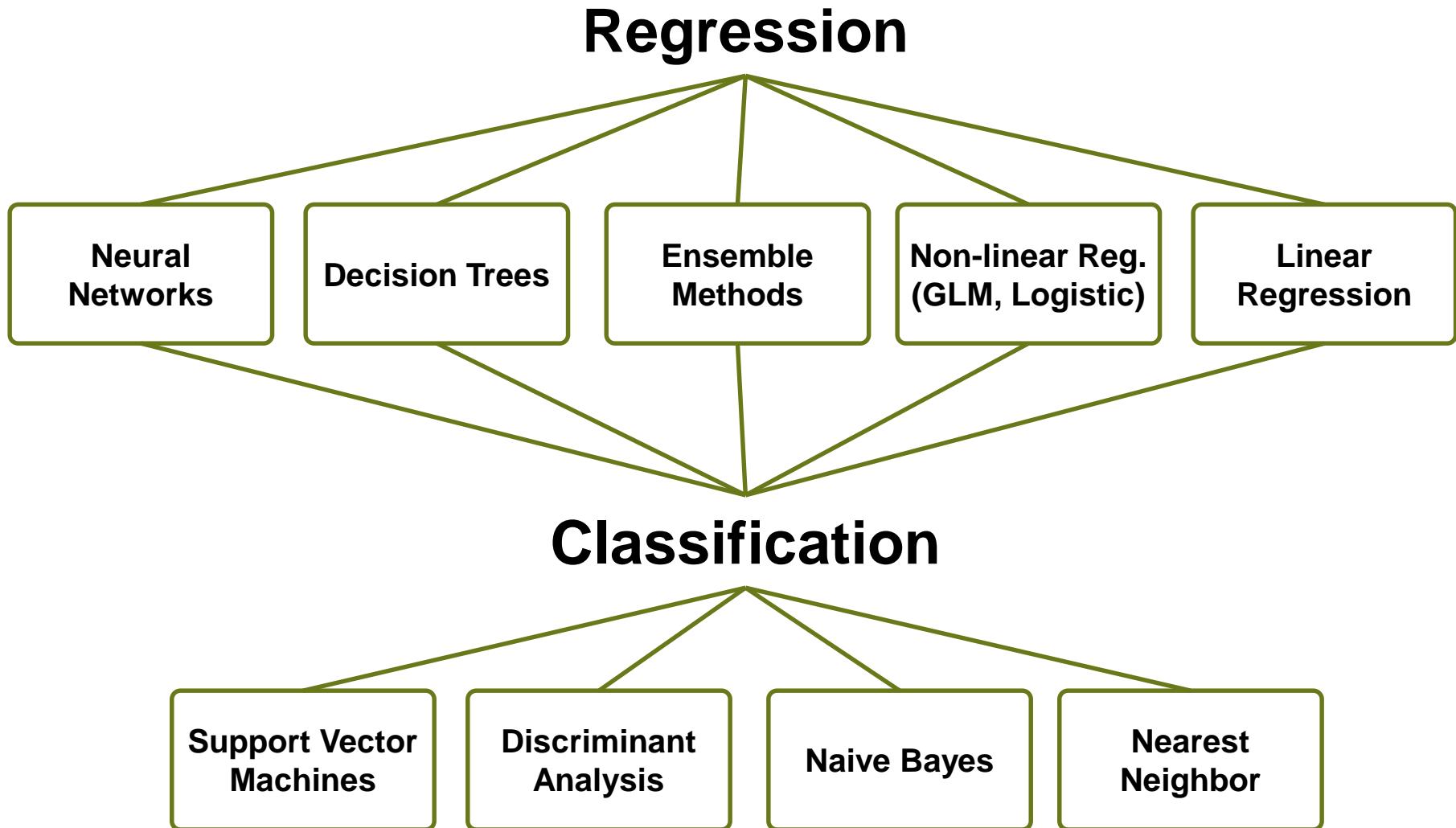


	AAA	AA	A	BBB	BB	B	CCC	D
AAA	93.68%	5.55%	0.59%	0.18%	0.00%	0.00%	0.00%	0.00%
AA	2.44%	92.60%	4.03%	0.73%	0.15%	0.00%	0.00%	0.06%
A	0.14%	4.18%	91.02%	3.90%	0.60%	0.08%	0.00%	0.08%
BBB	0.03%	0.23%	7.49%	87.88%	3.78%	0.39%	0.06%	0.16%
BB	0.03%	0.12%	0.73%	8.27%	86.74%	3.28%	0.18%	0.64%
B	0.00%	0.00%	0.11%	0.82%	9.64%	85.37%	2.41%	1.64%
CCC	0.00%	0.00%	0.00%	0.37%	1.84%	6.24%	81.88%	9.67%
D	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%

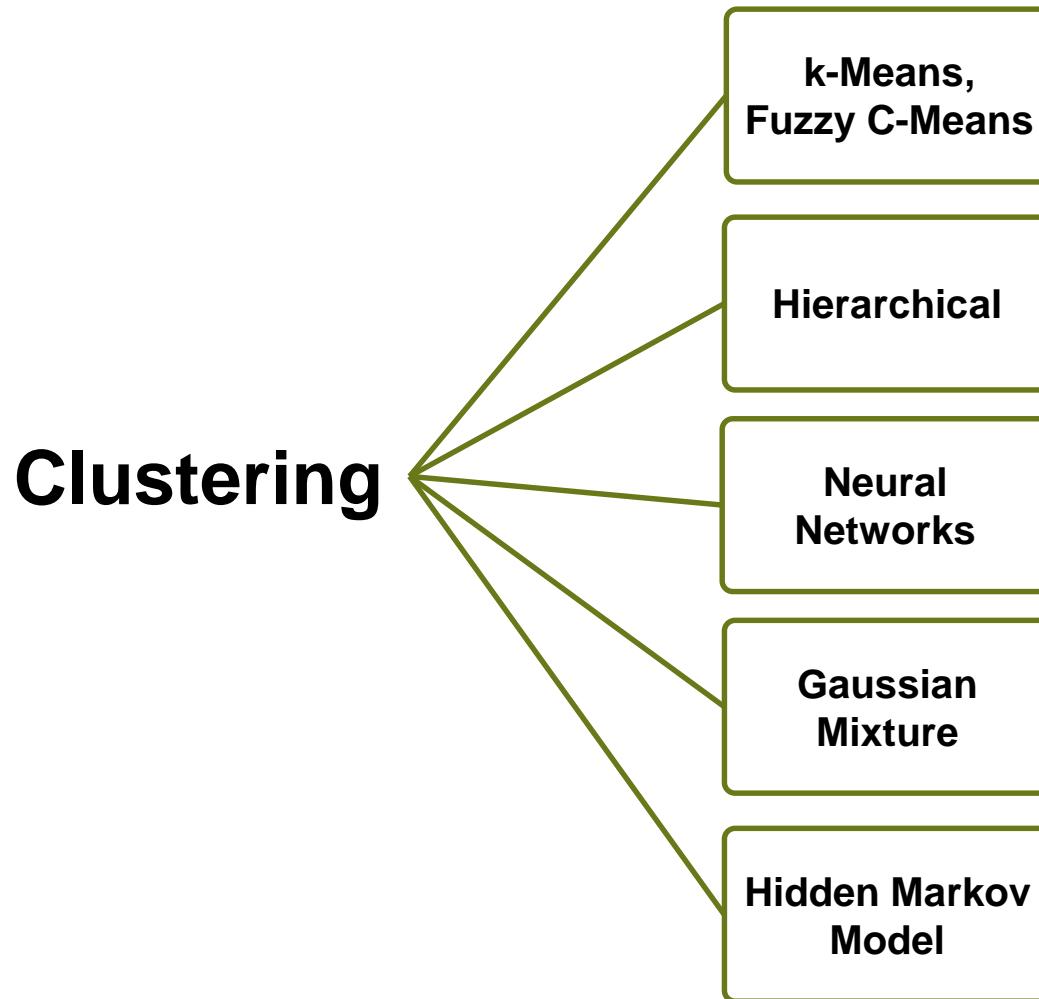
# Overview – Machine Learning



# Supervised Learning



# Unsupervised Learning



# Learn More: Machine Learning with MATLAB

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

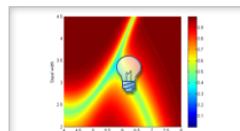
## Classification Examples



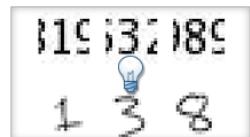
Basket Selection Using Stepwise Regression



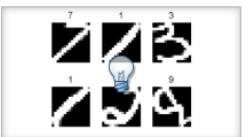
Classification in the Presence of Missing Data



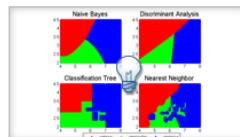
Classification Probability



Digit Classification Using HOG Features

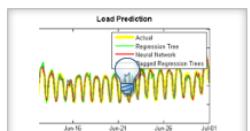


Handwriting Recognition Using Bagged Classification Trees

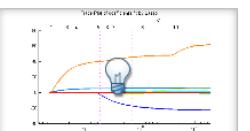


Visualize Decision Surfaces for Different Classifiers

## Regression Examples



Electricity Load Forecasting

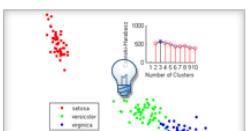


Lasso Regularization



Regression with Boosted Decision Tree

## Clustering Examples



Cluster Evaluation



Cluster Genes Using K-Means and Self-Organizing Maps



Color-Based Segmentation Using K-Means Clustering

## Machine Learning with MATLAB

### Machine Learning with MATLAB

Build predictive models and discover useful patterns from observed data.

[Watch video](#)

Machine learning algorithms use computational methods to "learn" information directly from data without assuming a predetermined equation as a model. They can adaptively improve their performance as you increase the number of samples available for learning.

Machine learning algorithms are used in applications such as [computational finance](#) (credit scoring and algorithmic trading), [computational biology](#) (tumor detection, drug discovery, and DNA sequencing), [energy production](#) (price and load forecasting), natural language processing, speech and image recognition, and advertising and recommendation systems.

Machine learning is often used in [big data](#) applications, which have large datasets with many predictors (features) and are too complex for a simple parametric model. Examples of big data applications include [forecasting electricity load](#) with a neural network, or [bond rating classification](#) for credit risk using an ensemble of decision trees.

### Classification



### Regression



Build models to predict continuous data.

### Clustering



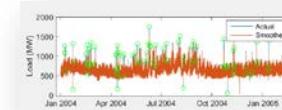
Find natural groupings and patterns in data.

# Challenges with Data Analytics

✓ Aggregating data from multiple sources



✓ Cleaning data



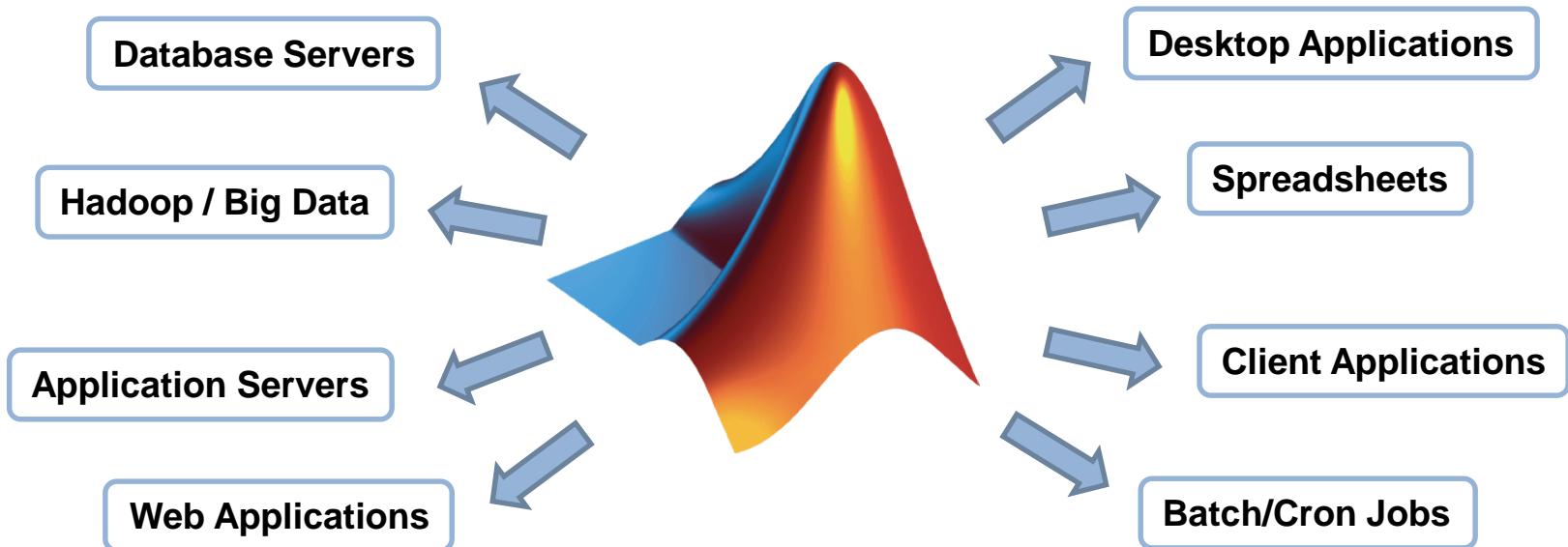
✓ Choosing a model



- Moving to production

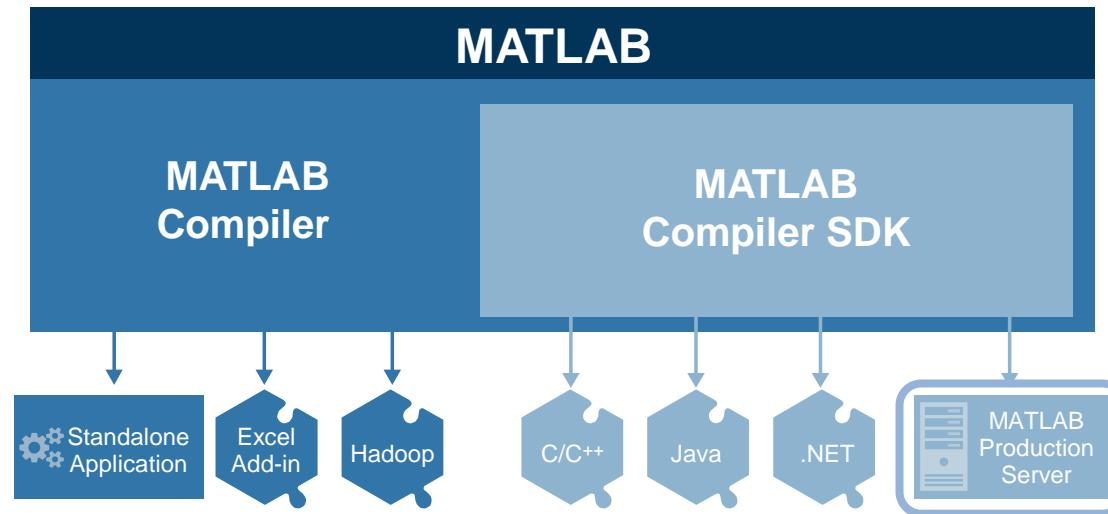


# Deployment Highlights



- Share with others who may not have MATLAB
- Royalty-free deployment
- Encryption to protect your intellectual property

# Deploying Applications with MATLAB

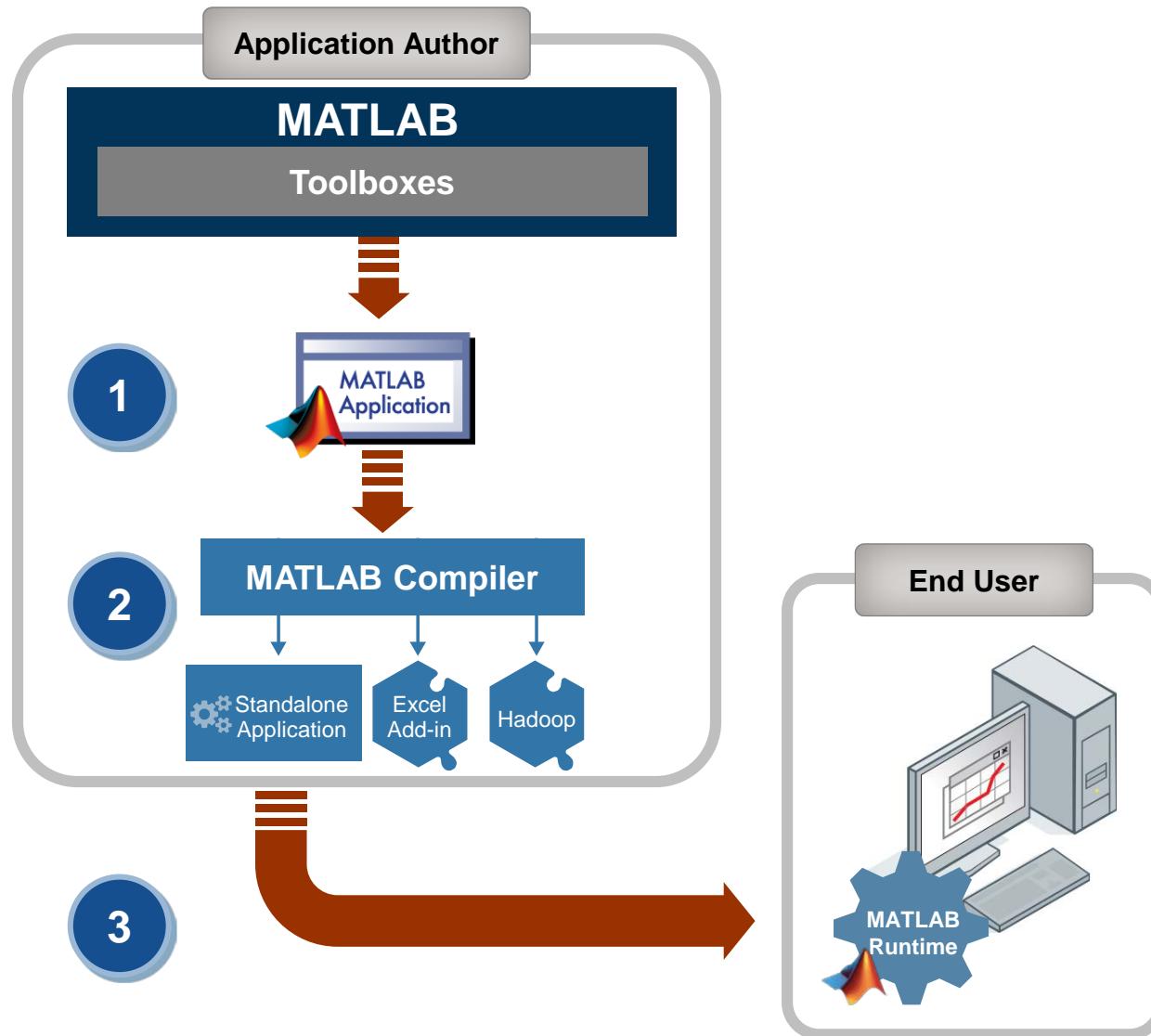


***MATLAB Compiler*** for sharing MATLAB programs without integration programming

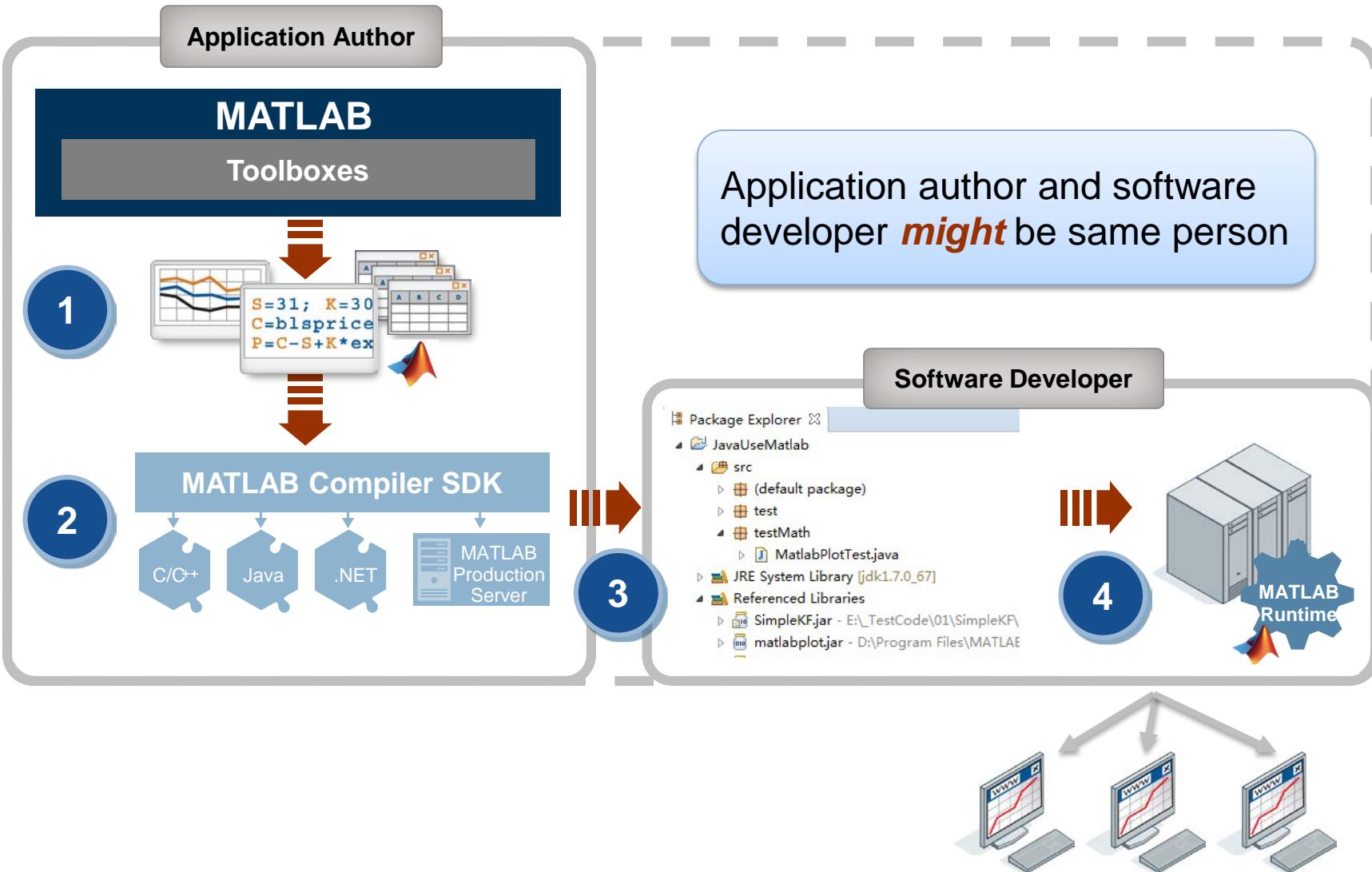
***MATLAB Compiler SDK*** provides implementation and platform flexibility for software developers

***MATLAB Production Server*** provides the most efficient development path for secure and scalable web and enterprise applications

# Sharing Standalone Applications

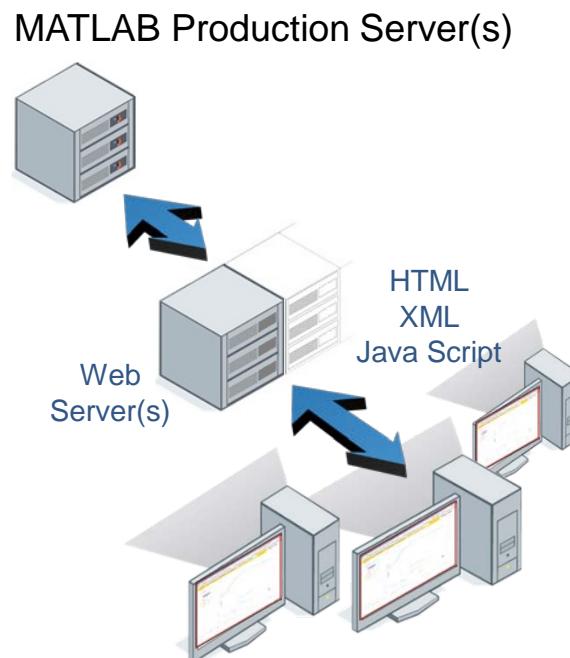


# Integrating MATLAB-based Components



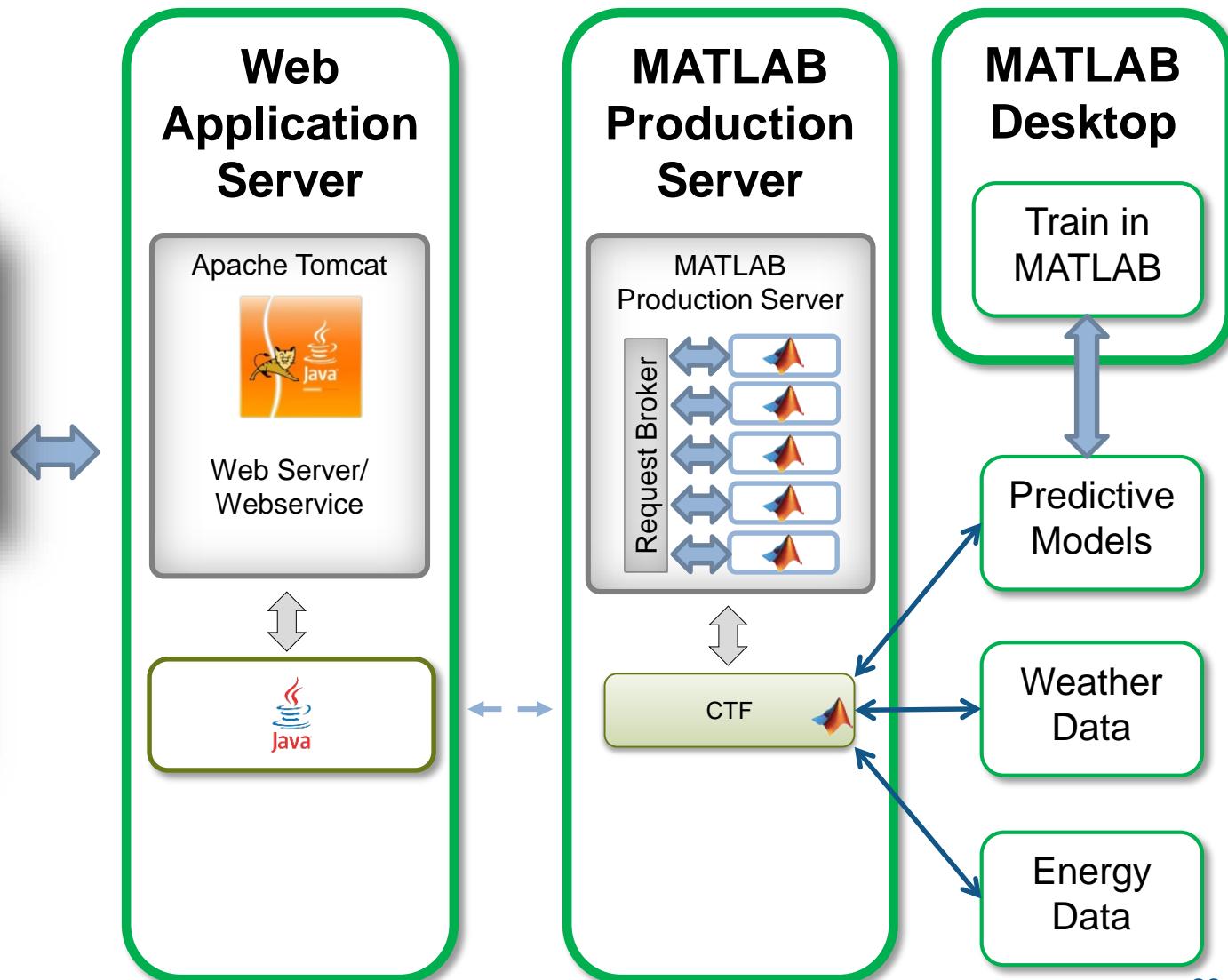
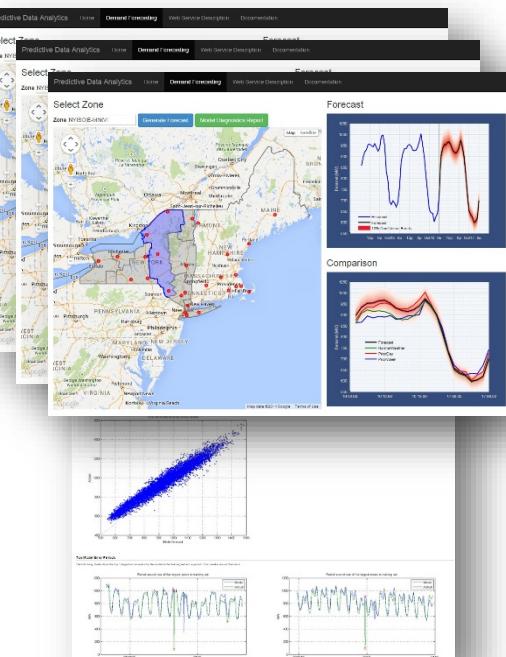
# MATLAB Production Server

- Directly deploy MATLAB analytic programs into production
  - Centrally manage multiple MATLAB programs & MCR versions
  - Automatically deploy updates without server restarts
- Scalable & reliable
  - Service large numbers of concurrent requests
  - Add capacity or redundancy with additional servers
- Use with web, database & application servers
  - Lightweight client library isolates MATLAB processing
  - Access MATLAB programs using native data types
  - Integrates with Java, .NET, C and Python



# Deployed Analytics

## MATLAB Production Server



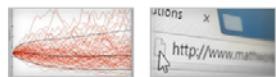
# Learn More: Application Deployment with MATLAB

[www.mathworks.com/solutions/desktop-web-deployment/](http://www.mathworks.com/solutions/desktop-web-deployment/)

## Deploying MATLAB Code as an Executable or Software Component

Using MathWorks application deployment products, you can eliminate the costly and error-prone work of recoding your MATLAB algorithms in another programming language. Because you maintain your source code in MATLAB, you can easily develop and update your algorithms and automatically package them as standalone executables or software components for integration in environments such as C, C++, Java™, .NET, and Excel®.

MATLAB Compiler packages your MATLAB applications as encrypted standalone executables or C/C++ shared libraries. MATLAB builder products work in conjunction with MATLAB Compiler to create standard components for use with Java, .NET, or Excel. These executables and components can be deployed royalty-free on operating systems supported by MATLAB.



Robeco Develops Quantitative Stock Selection and Portfolio Optimization Models (User Story)

Extend your Java math capabilities with MATLAB

## Deploying MATLAB Code as a Web Application

Using MathWorks application deployment products, you can develop MATLAB based components for the Web that execute mathematical computations and generate interactive graphics. After developing an algorithm in MATLAB, you can automatically create a standard component designed to integrate in a Web application using MATLAB builder products for either Java or .NET.

Once you place the component on a Web server, your users access the application through a Web browser and do not need to install additional software on their desktop computers.

The Java and .NET components created by the deployment tools can be used in conjunction with standard Web technologies such as ASP.NET, SOA, SaaS, JavaScript, and HTML.



Application Deployment with MATLAB 22:58



Rendering High Dynamic Range Images on the Web (Article)

## Desktop and Web Deployment

Overview Deploying MATLAB Code as an Executable or Software Component Deploying MATLAB Code as a Web Application

Contact sales Trial software

Share the work you do in MATLAB with others



MathWorks products provide several ways to share individual algorithms or complete applications that you develop using MATLAB® and add-on toolboxes. You can distribute your code directly to others to use in their own MATLAB sessions. To distribute to people who do not have MATLAB, the MathWorks application deployment products enable automatic generation and royalty-free distribution of turnkey applications and components that can be easily integrated into a larger IT infrastructure.

For [embedded systems](#) and real-time applications, you can use MathWorks products to automatically generate C or HDL code from MATLAB and Simulink® algorithms for microcontrollers, DSP chips, and simulators.

Explore Products for Desktop and Web Deployment

[MATLAB Builder™ EX](#)  
(for Microsoft Excel)

[MATLAB Builder™ JA](#)  
(for Java language)

[MATLAB Builder™ NE](#)  
(for Microsoft .NET Framework)

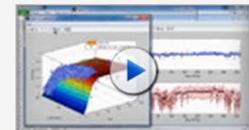
[MATLAB Compiler™](#)

Desktop and Web Deployment Resources

[Webinars](#)

[Training](#)

[Technical Articles](#)



From Apps to Web Services: Sharing the Work You've Done in MATLAB 27:41

## Learn More About Desktop and Web Deployment Solutions



# Learn More: MATLAB Application Deployment



**MATLAB Compiler** MAJOR UPDATE  
Build standalone applications from MATLAB programs

MATLAB Compiler™ lets you build standalone applications from MATLAB programs. You can also create Microsoft Excel® add-ins and Microsoft Word® add-ins.

When used along with MATLAB® Compiler SDK™, you can build C/C++ shared libraries and C/C++ MEX® programs. These components can be used in desktop, web, and enterprise applications.

**MATLAB Compiler SDK** MAJOR UPDATE  
Build software components from MATLAB programs

MATLAB Compiler SDK™ lets you build software components from MATLAB programs. These components can be used in desktop, web, and enterprise applications.

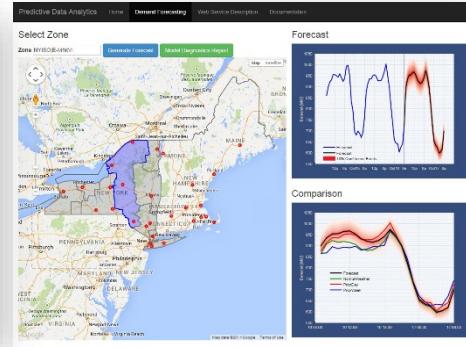
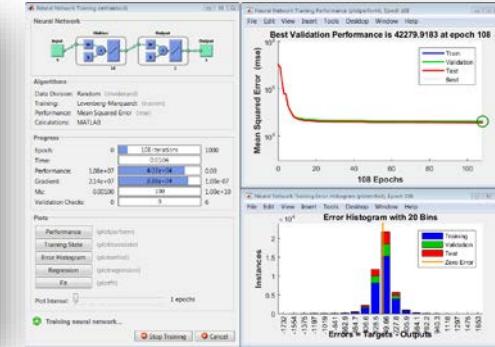
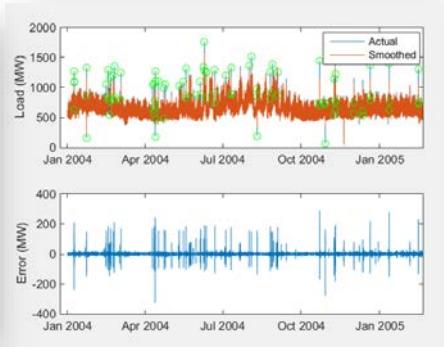
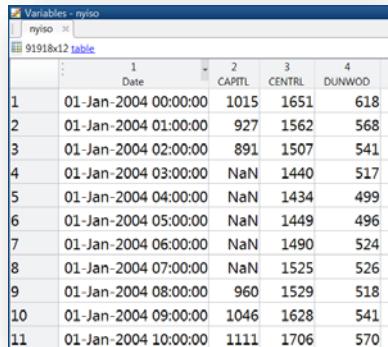
**MATLAB Production Server**

Run MATLAB analytics as a part of web, database, and enterprise applications

MATLAB Production Server™ lets you run MATLAB® programs within your production systems, enabling you to incorporate custom analytics in enterprise applications. Web, database, desktop, and enterprise applications request MATLAB analytics running on MATLAB Production Server via a lightweight client library. A server-based deployment ensures that users access the latest version of your analytics automatically, with client connections that can be

Also ... [www.mathworks.com/solutions/desktop-web-deployment/](http://www.mathworks.com/solutions/desktop-web-deployment/)

# Data Analytics Products



Access and Explore Data

Preprocess Data

Develop Predictive Models

Integrate Analytics with Systems



MATLAB

Parallel Computing Toolbox, MATLAB Distributed Computing Server

MATLAB Production Server

Database Toolbox

Statistics and Machine Learning Toolbox

MATLAB Compiler

Data Acquisition Toolbox

Curve Fitting Toolbox

Neural Network Toolbox

MATLAB Compiler SDK

Mapping Toolbox

Signal Processing Toolbox

Computer Vision System Toolbox

Image Acquisition Toolbox

Image Processing Toolbox

Econometrics Toolbox

OPC Toolbox

Used in today's demo

Additional Data Analytics products

# Key Takeaways

- Data preparation can be a big job; leverage built-in MATLAB tools and spend more time on the analysis
- Rapidly iterate through different predictive models, and find the one that's best for your application
- Leverage parallel computing to scale-up your analysis to large datasets
- Eliminate the need to recode by deploying your MATLAB algorithms into production

