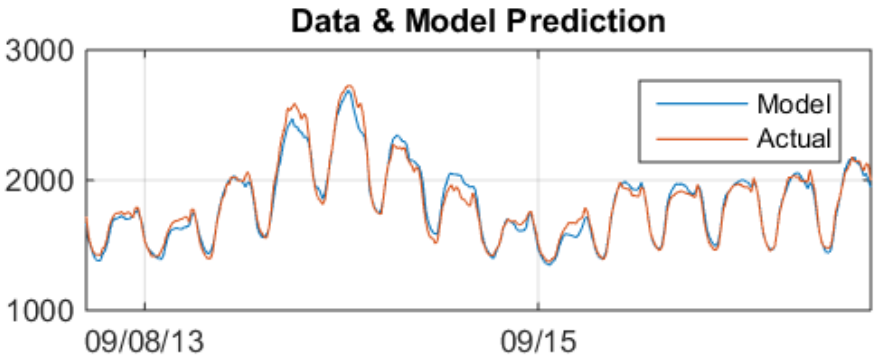
Data Analytics with MATLAB

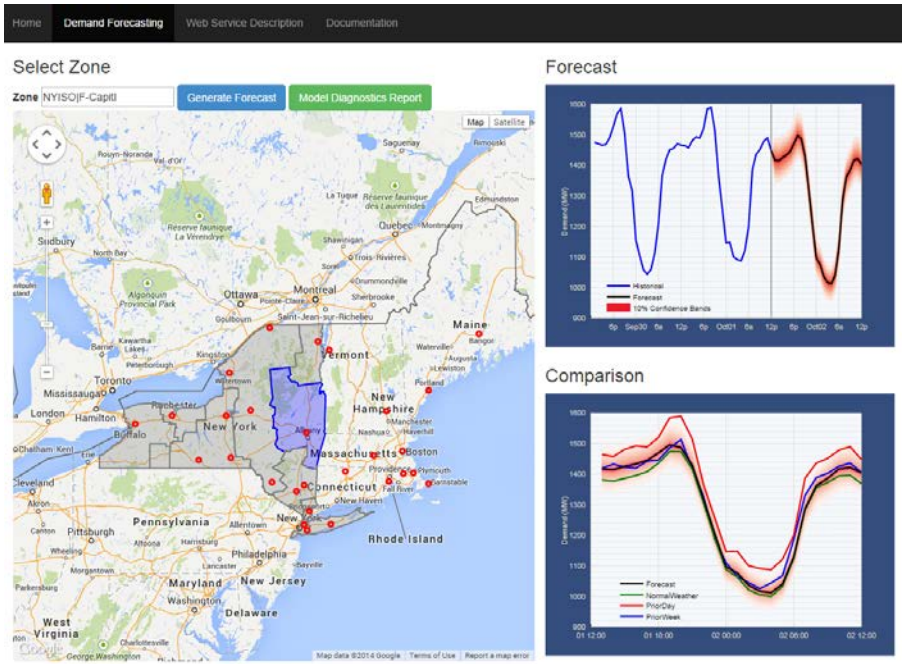
Adam Filion
Application Engineer
MathWorks

© 2015 The MathWorks, Inc.

Case Study: Day-Ahead Load Forecasting

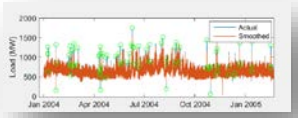
- Goal:
  - Implement a tool for ***easy*** and ***accurate*** computation of day-ahead system load forecast
- Requirements:
  - Acquire and clean data from multiple sources
  - Accurate predictive model
  - Easily deploy to production environment





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/

NEW YORK INDEPENDENT SYSTEM OPERATOR
Building The Energy Markets Of Tomorrow...Today

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-511A
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
Real-Time Actual Load P-58B
Integrated Real-Time Actual Load P-58C

Real-Time Actual Load

CSV Files	Last Updated
10-21-2014	10/21/14 23:02 EDT
10-20-2014	10/21/14 00:07 EDT
10-19-2014	10/20/14 00:01 EDT
10-18-2014	10/18/14 23:59 EDT
10-17-2014	10/18/14 00:00 EDT
10-16-2014	10/16/14 23:59 EDT
10-15-2014	10/15/14 23:59 EDT
10-14-2014	10/14/14 23:59 EDT
10-13-2014	10/13/14 23:59 EDT
10-12-2014	10/12/14 23:59 EDT

Archived Files (zip format)

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

Variables - nyiso

nyiso

91918x12 table

	1 Date	2 CAPITL	3 CENTRL	4 DUNWOD	5 GENESE	6 HUDVL	7 MHKVL	8 MILLWD	9 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

Variables - nyiso

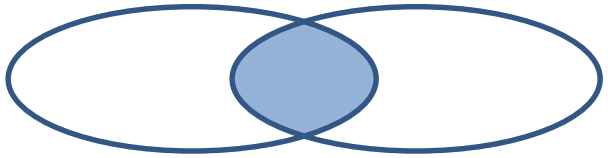
nyiso

91918x12 table

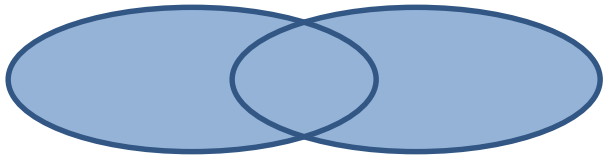
	1 Date	2 CAPITL	3 CENTRL	4 DUNWOD	5 GENESE	6 HUDVL	7 MHKVL	8 MILLWD	9 NORTH
1	01-Jan-2004 00:00:00	1015	1651	618	972	1120	645	223	622
2	01-Jan-2004 01:00:00	927	1562	568	905	1019	602	201	628
3	01-Jan-2004 02:00:00	891	1507	541	858	977	571	195	604
4	01-Jan-2004 03:00:00	NaN	1440	517	821	927	525	183	618
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

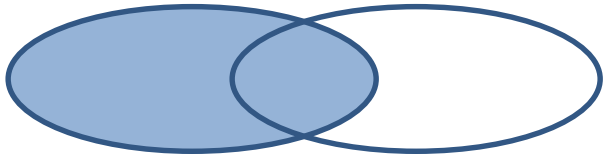
- Join along a common axis
- Popular Joins:
  - Inner
  - Full Outer
  - Left Outer
  - Right Outer



Inner Join

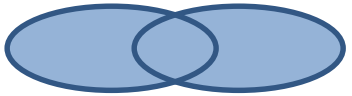


Full Outer Join



Left Outer Join

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

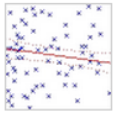
Joined Data Set

Learn More: Big Data with MATLAB

www.mathworks.com/discovery/big-data-matlab.html

www.mathworks.com/discovery/matlab-mapreduce-hadoop.html

MapReduce on the Desktop

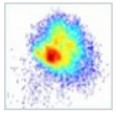


Explore and analyze big data sets on your desktop with the MapReduce programming technique built into MATLAB.

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

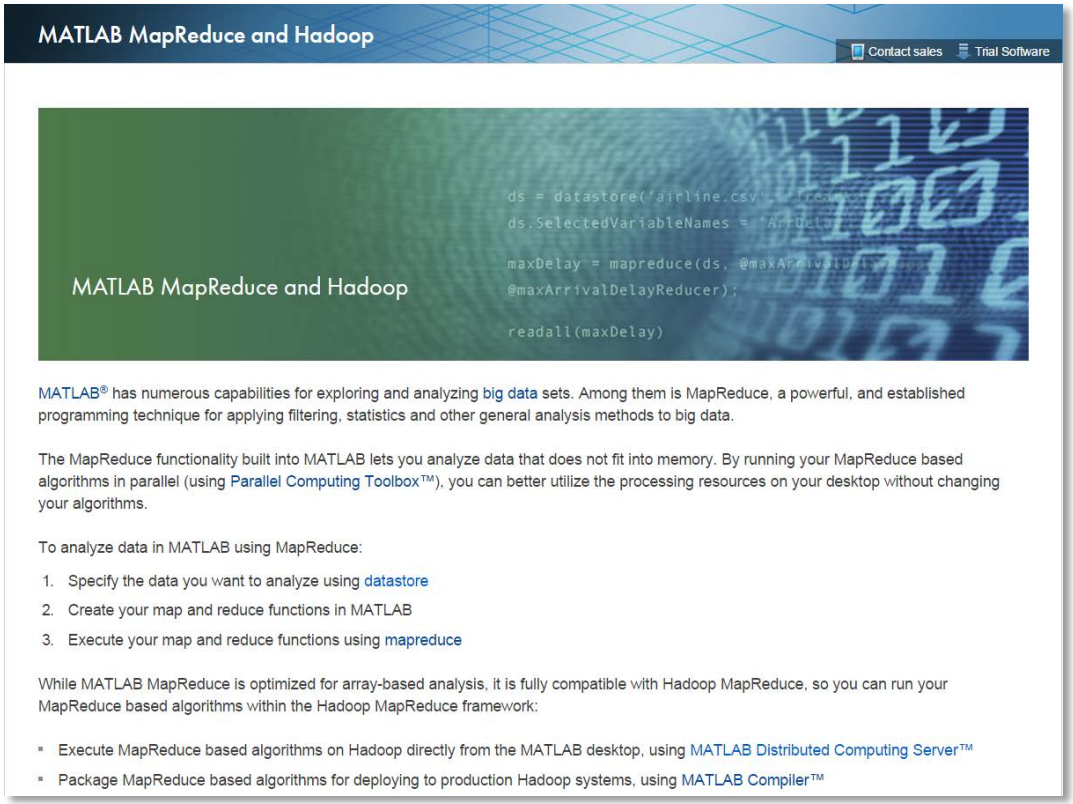


Execute MATLAB MapReduce based algorithms within Hadoop MapReduce to explore and analyze data that is stored and managed on Hadoop, using MATLAB Distributed Computing Server.

» Run MATLAB MapReduce on Hadoop

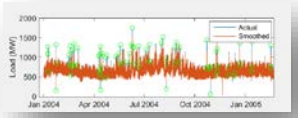
Create applications and libraries based upon MATLAB MapReduce for deployment within production instances of Hadoop, using MATLAB Compiler.

» Deploy MATLAB MapReduce applications to Hadoop



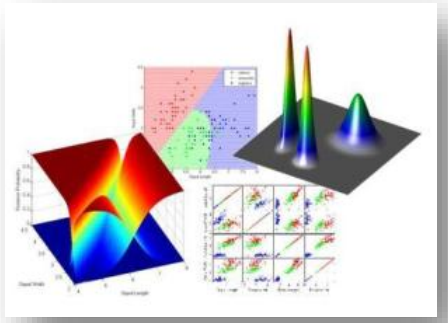
Challenges with Data Analytics

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

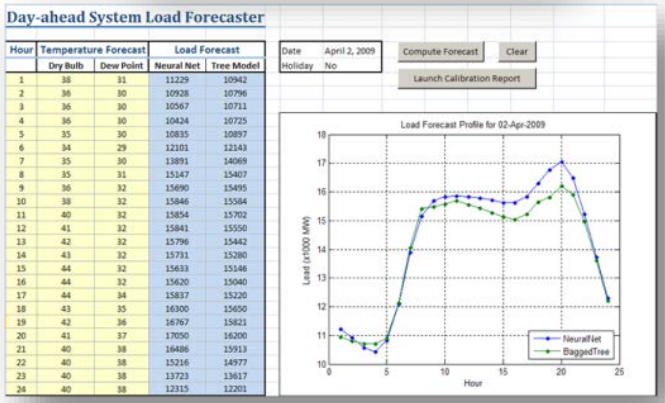


Machine Learning

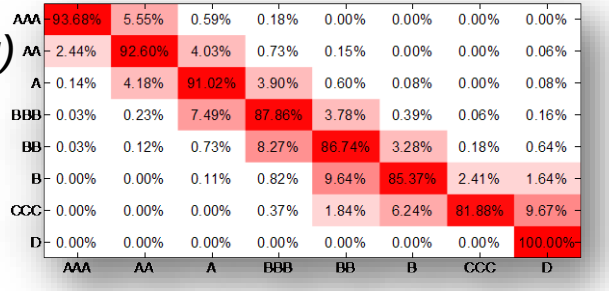
Characteristics and Examples



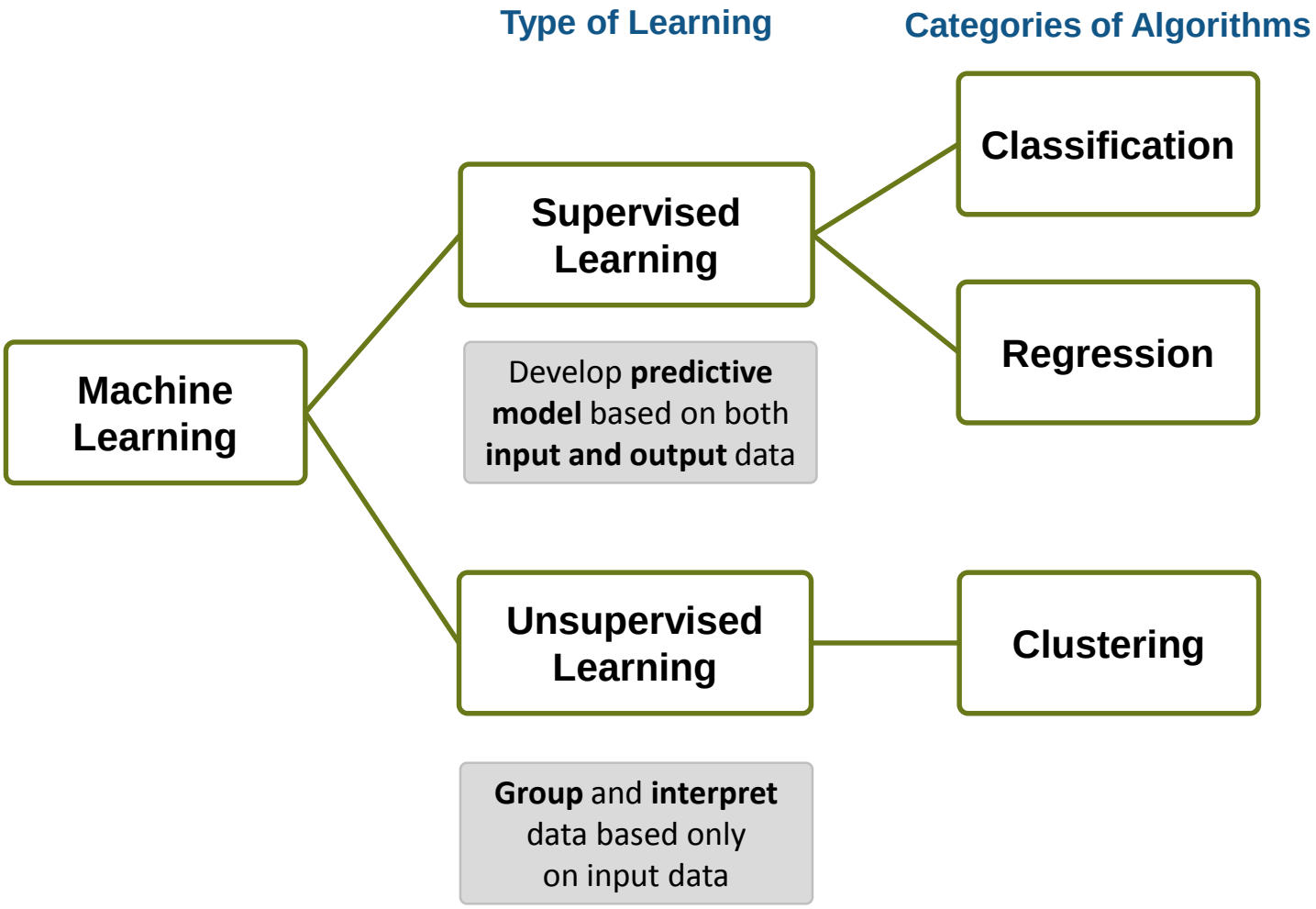
- Characteristics
  - Lots of variables
  - System too complex to know the governing equation
    (e.g., black-box modeling)



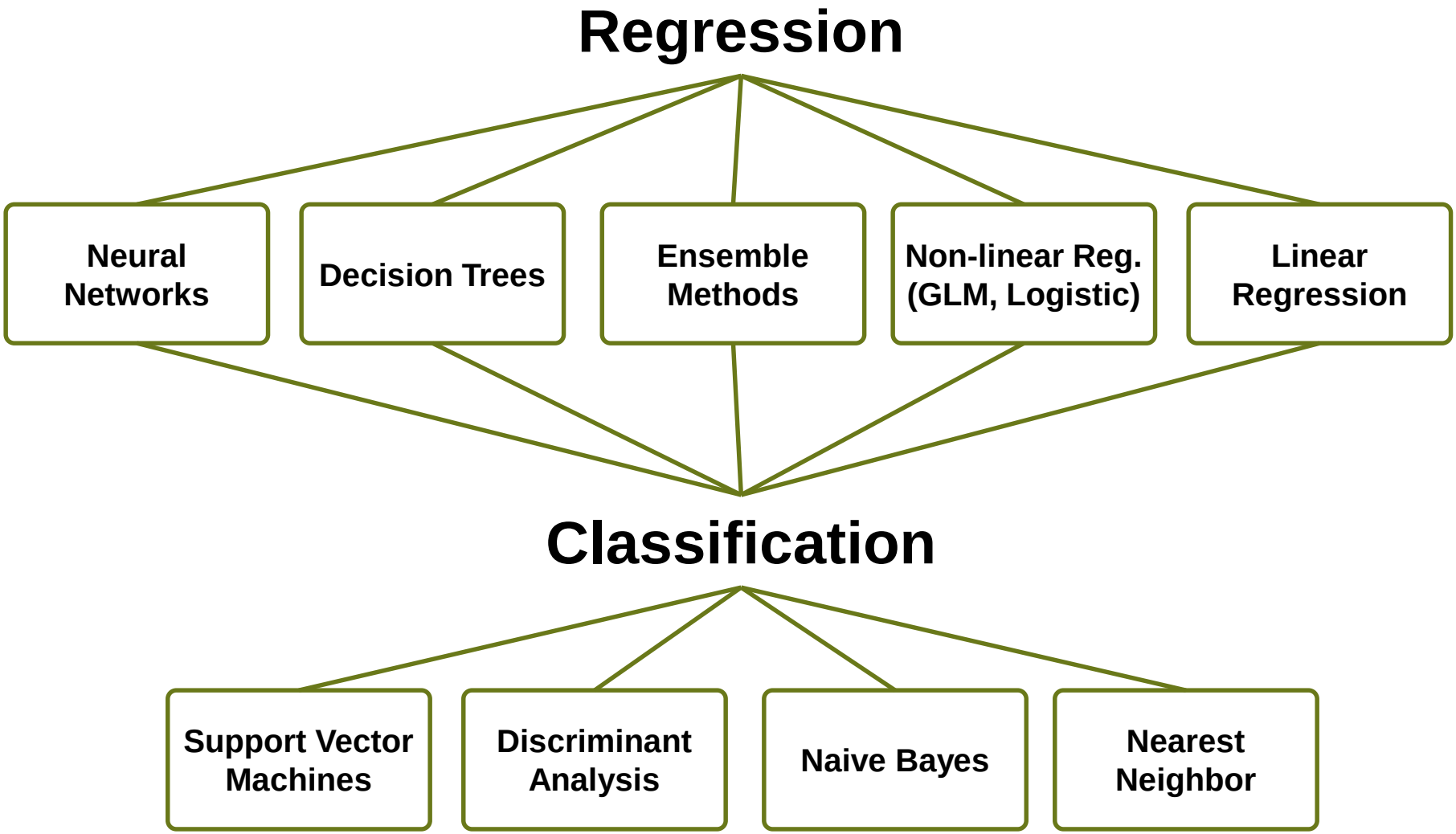
- Examples
  - Pattern recognition *(speech, images)*
  - Financial algorithms *(credit scoring, algo trading)*
  - Energy forecasting *(load, price)*
  - Biology *(tumor detection, drug discovery)*



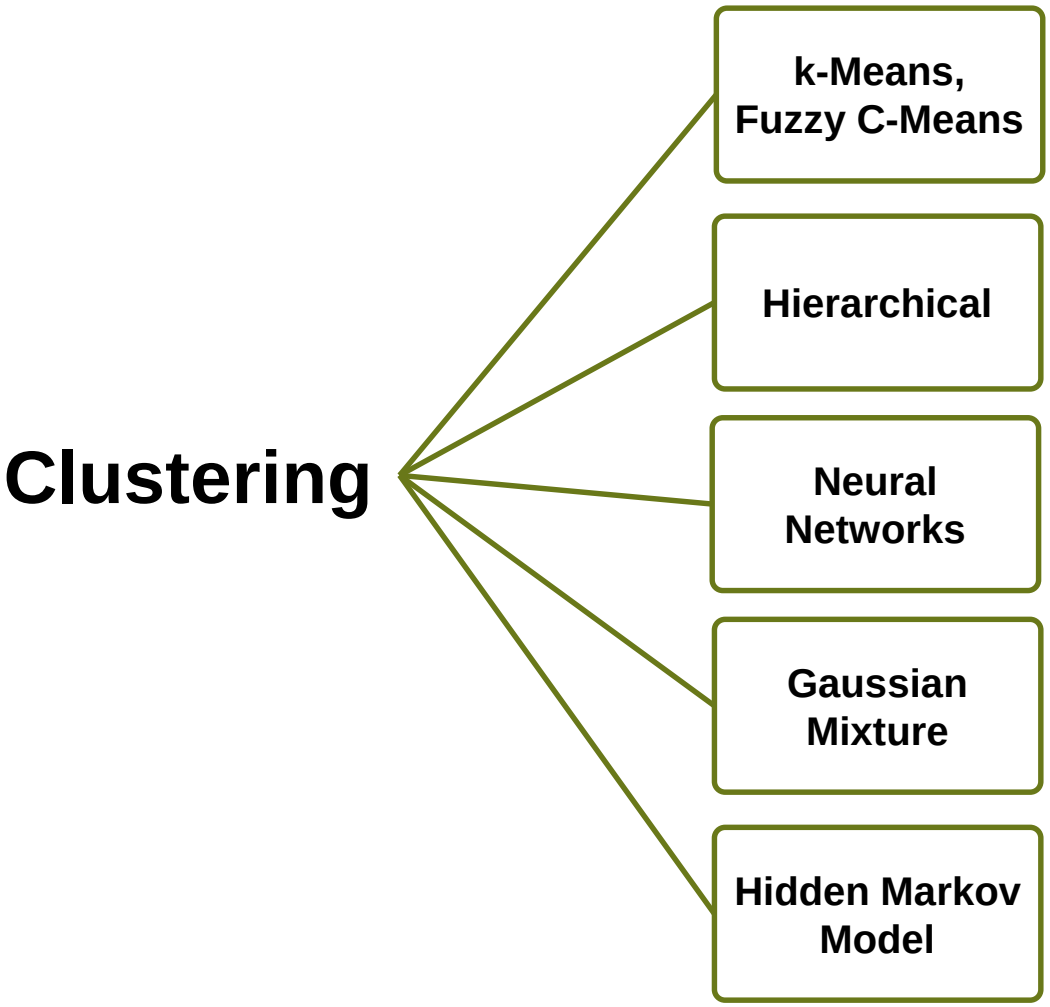
Overview – Machine Learning



Supervised Learning



Unsupervised Learning

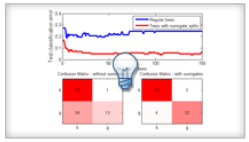


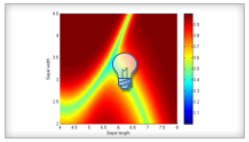
Learn More: Machine Learning with MATLAB

mathworks.com/machine-learning

Classification Examples

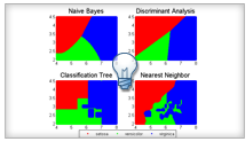
Basket Selection Using Stepwise Regression


Classification in the Presence of Missing Data

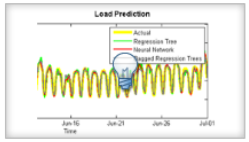

Classification Probability

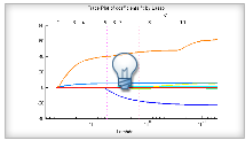
Digit Classification Using HOG Features

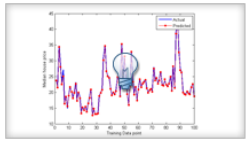
Handwriting Recognition Using Bagged Classification Trees


Visualize Decision Surfaces for Different Classifiers

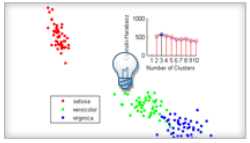
Regression Examples


Electricity Load Forecasting


Lasso Regularization


Regression with Boosted Decision Trees

Clustering Examples

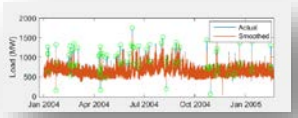

Cluster Evaluation

Cluster Genes Using K-Means and Self-Organizing Maps

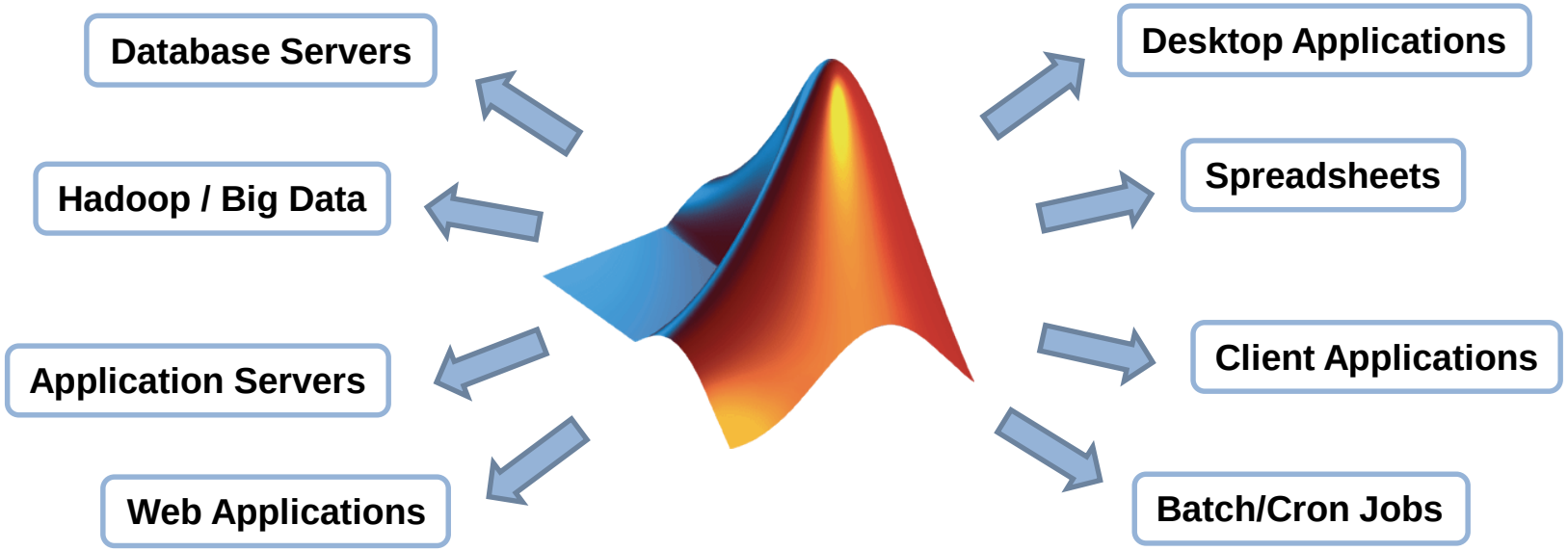
Color-Based Segmentation Using K-Means Clustering

Challenges with Data Analytics

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

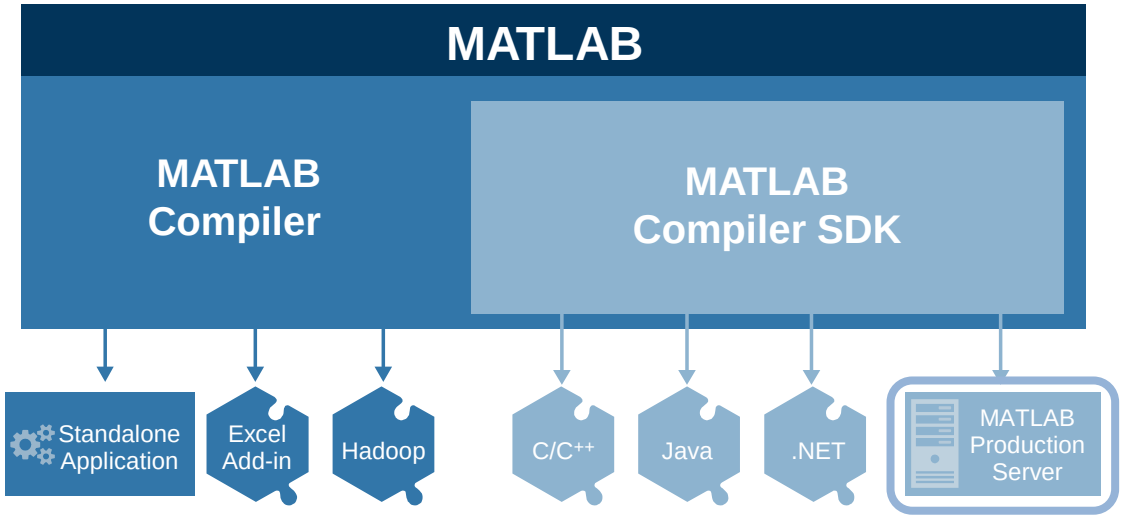


Deployment Highlights



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

Deploying Applications with MATLAB

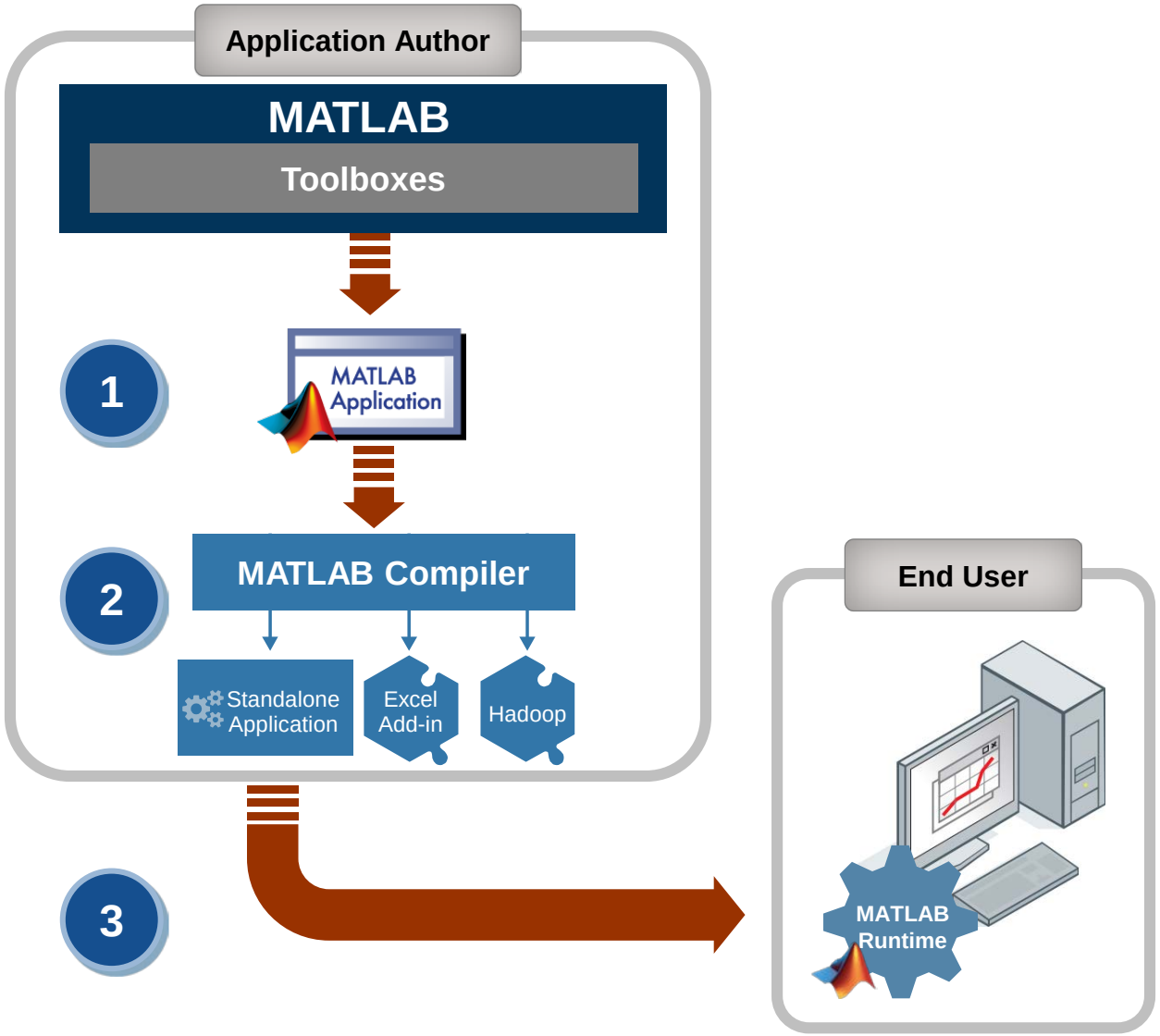


MATLAB Compiler for sharing MATLAB programs without integration programming

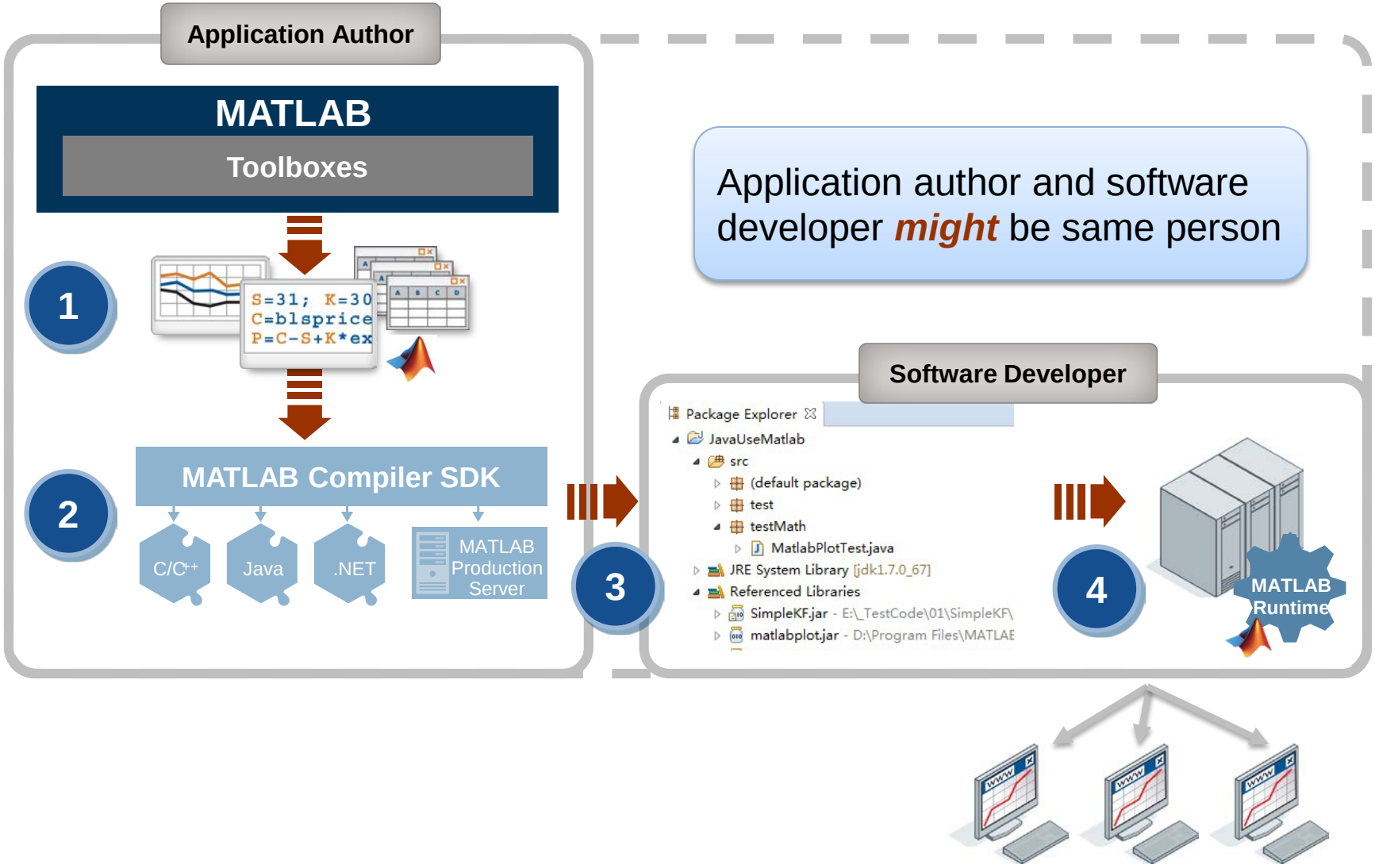
MATLAB Compiler SDK provides implementation and platform flexibility for software developers

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

Sharing Standalone Applications

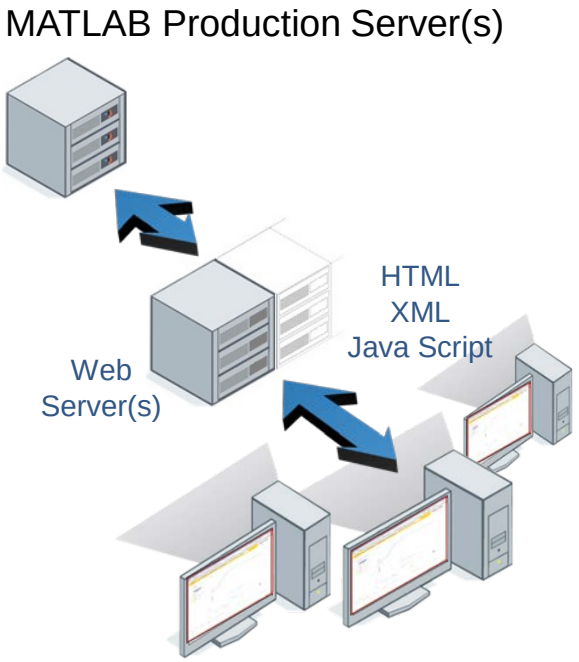


Integrating MATLAB-based Components



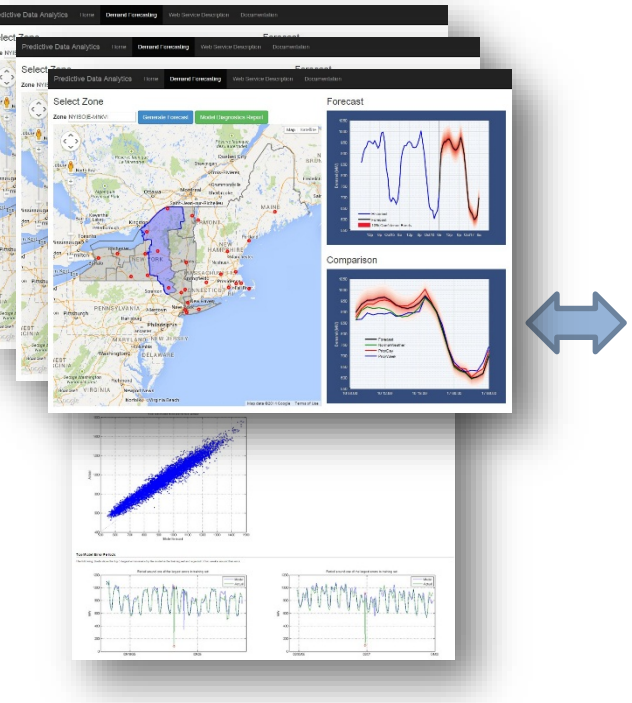
MATLAB Production Server

- 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



Web Application Server

Apache Tomcat

Web Server/ Webservice

Java

MATLAB Production Server

MATLAB Production Server

Request Broker

CTF

MATLAB Desktop

Train in MATLAB

Predictive Models

Weather Data

Energy Data

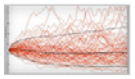
Learn More: Application Deployment with MATLAB

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

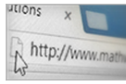
Deploying MATLAB Code as a Web Application

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.

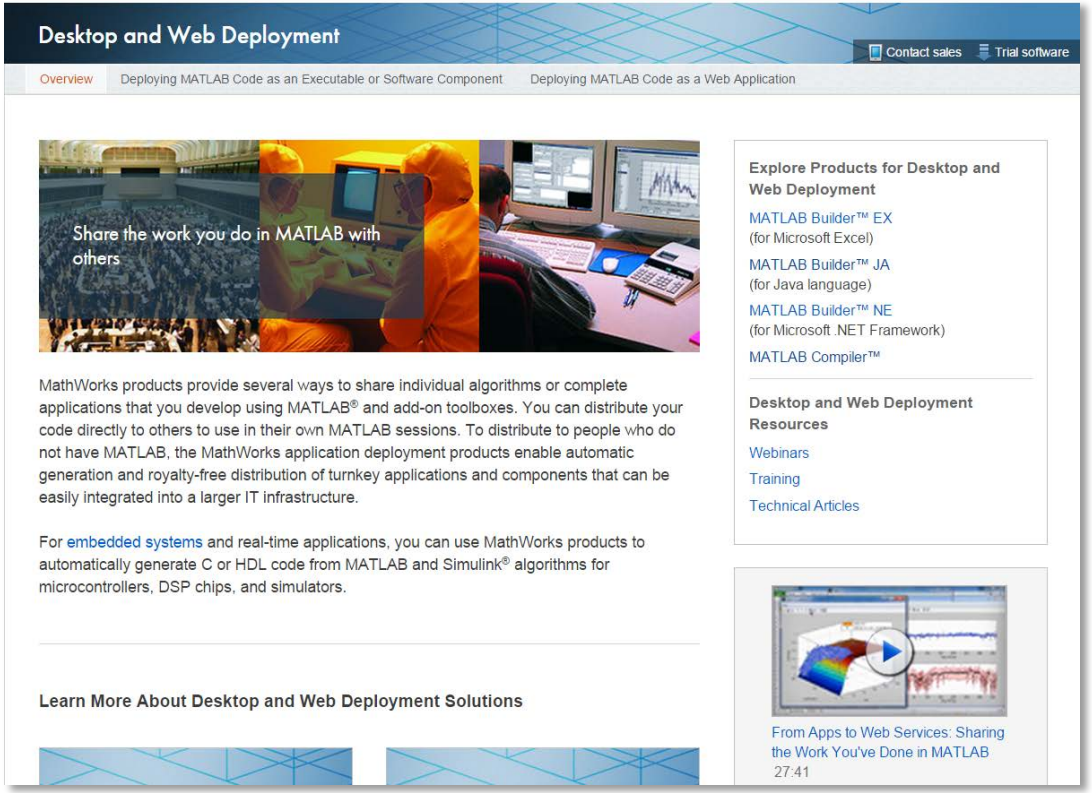
The Java and .NET components created by the deployment tools can be used in conjunction with standard Web technologies such as ASP.NET, SOA, SaaS, JavaScript, and HTML.

Application Deployment with MATLAB 22:58

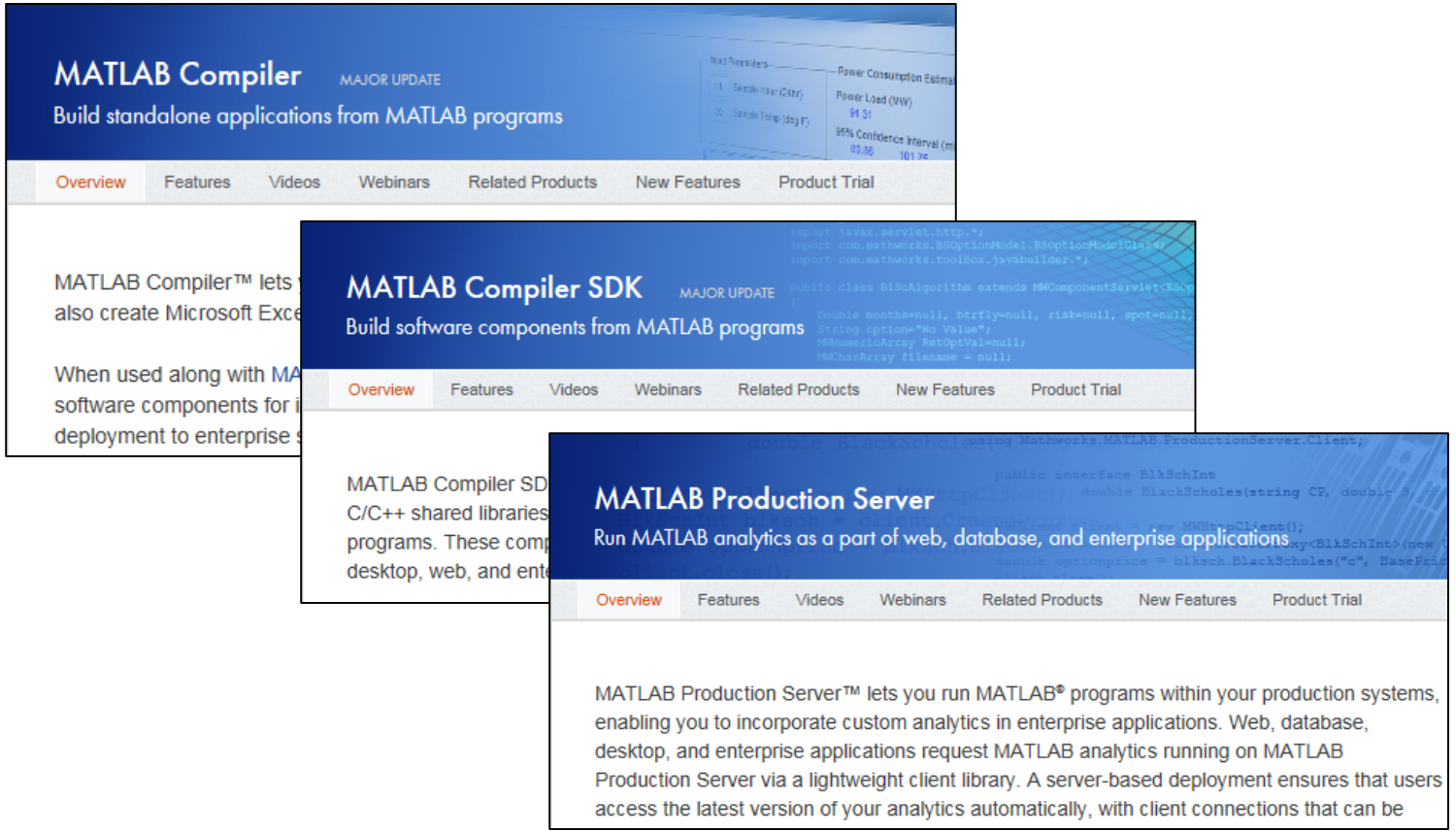


Rendering High Dynamic Range Images on the Web (Article)

Extend your Java math capabilities with MATLAB

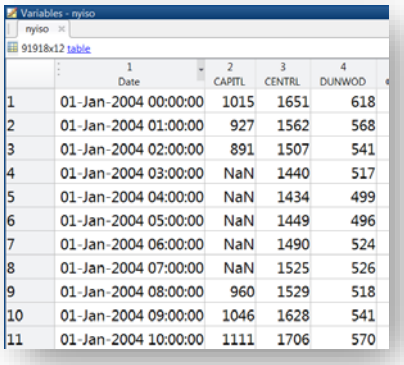


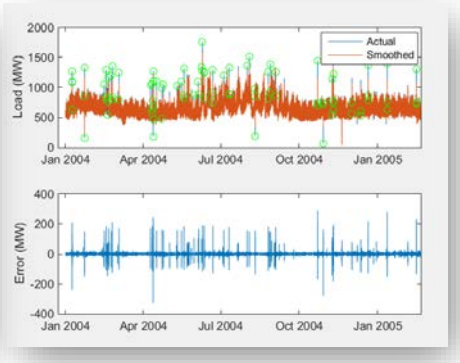
Learn More: MATLAB Application Deployment

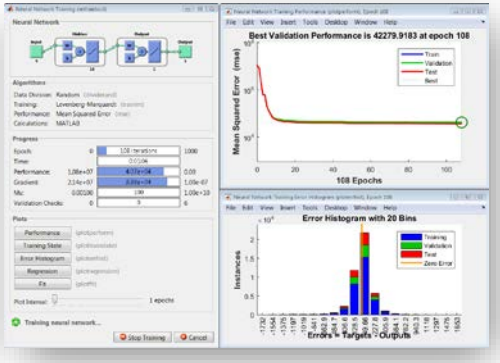


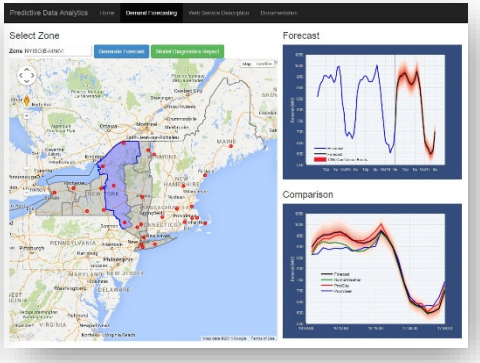
Also … www.mathworks.com/solutions/desktop-web-deployment/

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

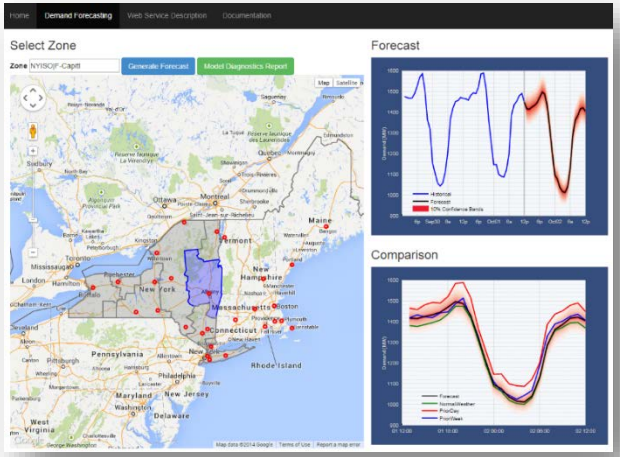
OPC Toolbox

Used in today's demo

Additional Data Analytics products

Key Takeaways

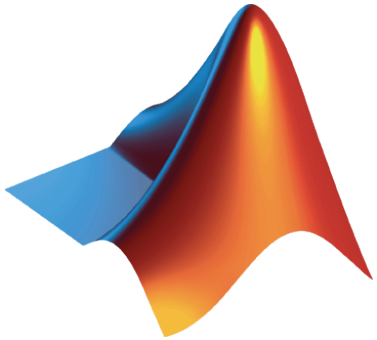
- Data preparation can be a big job; leverage built-in MATLAB tools and spend more time on the analysis



- 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



© 2015 The MathWorks, Inc.