What's New in MATLAB for Engineering Data Analytics?

Will Wilson
Application Engineer
MathWorks, Inc.

© 2017 The MathWorks, Inc.

Agenda

- Data Types
- Tall Arrays for Big Data
- Machine Learning (for Everyone)
- Deploying your Analytics

Example Use Case: Vehicle Log (MDF) File Analysis

R2016b

- MDF (**M**easurement **D**ata **F**ormat) is the de facto standard for measurement data in the automotive industry.
  - Official ASAM standard
  - Typically file extensions include: .mdf, .mf4, & .dat.
- **Goal:** Use MATLAB to process and analyze MDF data.
- **Considerations:**
  - Data are MDF format, could be many files
  - Data is messy
  - May or may not know what you are looking for
  - Compute statistics, report format

```
mdfObj = mdf('MDFFile.mf4')
```

```
MDF with properties:

   File Details
                Name: 'MDFFile.mf4'
                Path: 'c:\temp\MDFFile.mf4'
              Author: 'HOK'
          Department: 'Research'
             Project: 'MDF'
             Subject: 'CAN bus'
             Comment: 'This file contains CAN messages'
             Version: '4.10'
            DataSize: 32100
    InitialTimestamp: 2016-02-27 12:09:02

   Creator Details
     ProgramIdentifier: 'mmddff.04'
               Creator: [1×1 struct]

   File Contents
          Attachment: [1×1 struct]
        ChannelNames: {6×1 cell}
        ChannelGroup: [1×6 struct]
```

MATLAB Language Enhancements

Expressing more types of data naturally

Numeric

double, single, …

logical

R2013b

categorical

datetime

duration

R2014b

calendarDuration

R2016b

timetable

Heterogeneous

structure

cell

R2013b

table

R2016b

tall

Text

char

cell string

R2016b

string

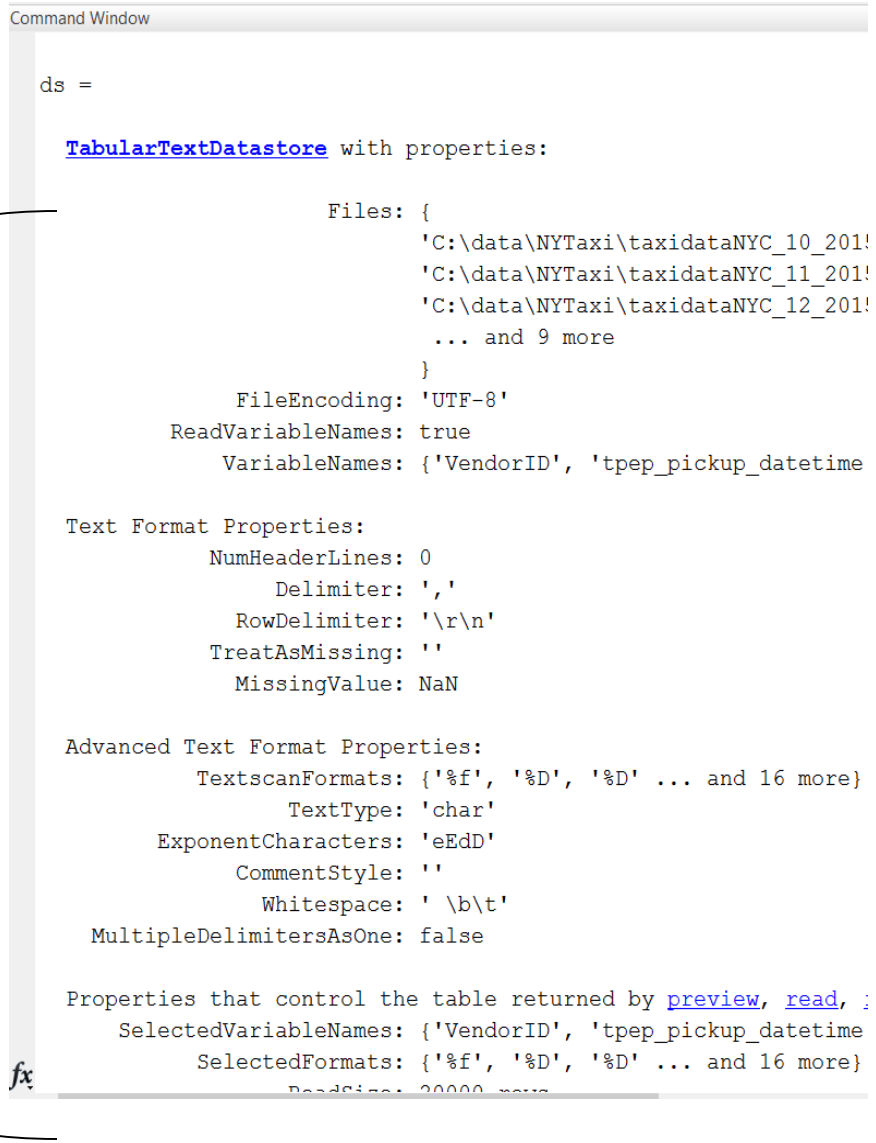
Agenda

- Data Types
- Tall Arrays for Big Data
- Machine Learning (for Everyone)
- Deploying your Analytics

Key Concept - MATLAB datastore

- A datastore is an object for reading a single file or a collection of files or data.
- Idea of properties.
- Data dependent
  - images, tabular text, user defined.
- Onramp to “Big Data”.

Properties of the datastore

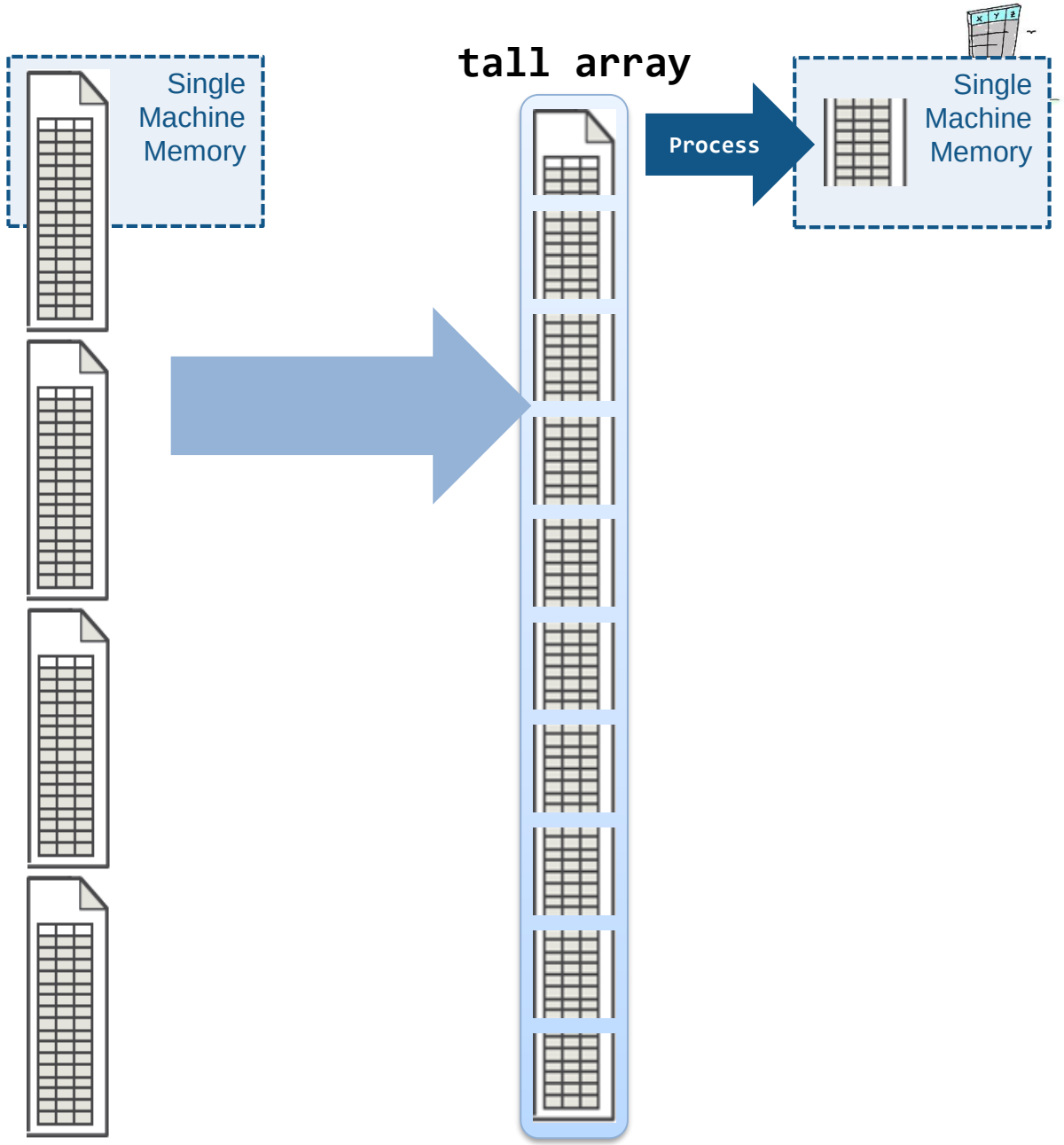


`tall` arrays R2016b

- New data type designed for data that doesn't fit into memory
- Many rows (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.)

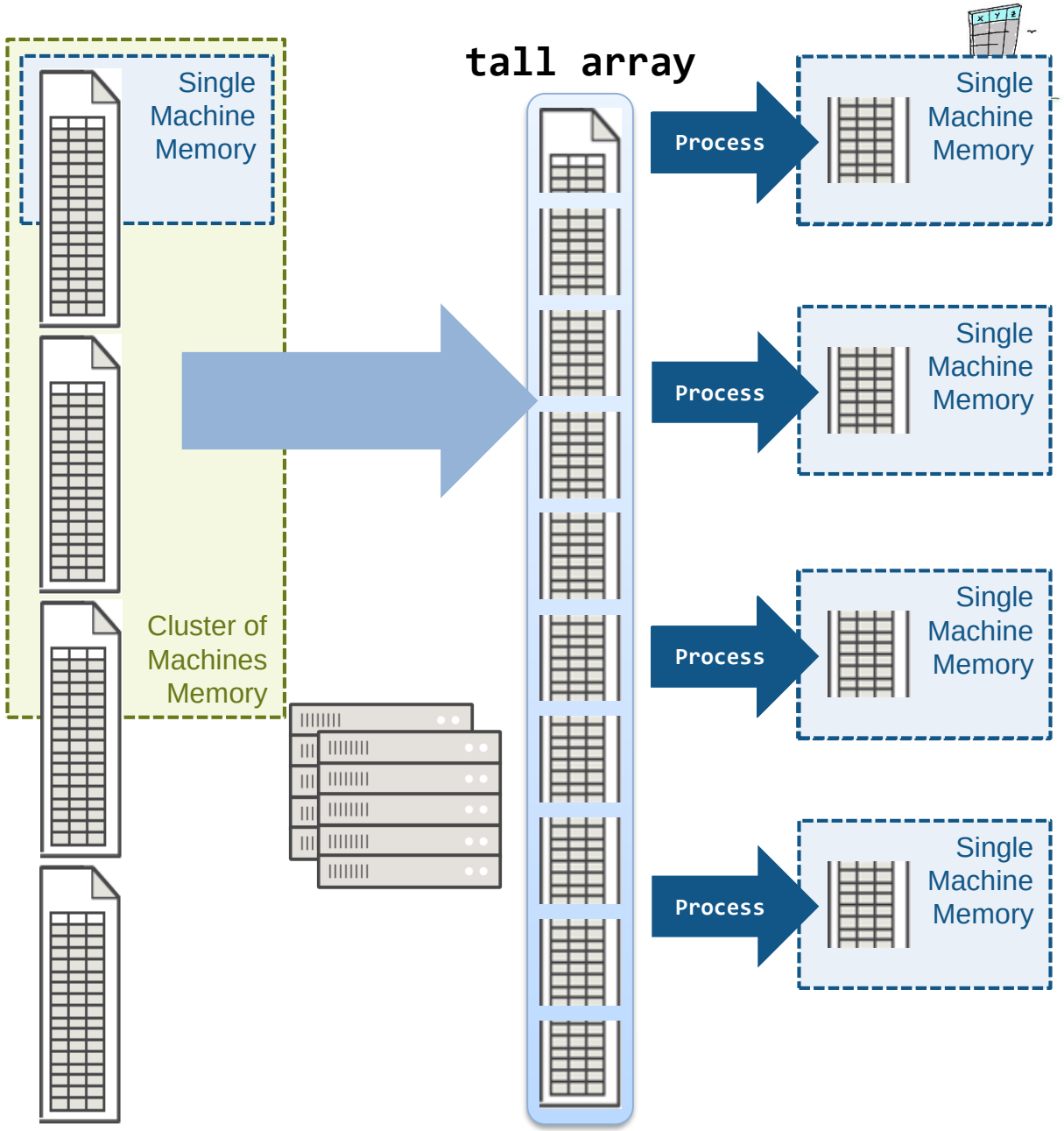
`tall` arrays R2016b

- 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



`tall` arrays R2016b

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



Big Data Workflow With Tall Data Types

Access Data

- Text
- Spreadsheet (Excel)
- Database (SQL)
- Custom Reader

Datastores for common types of structured data

Tall Data Types

- `table`
- `cell`
- `double`
- `numeric`
- `cellstr`
- `datetime`
- `categorical`

Tall versions of commonly used MATLAB data types

Exploration & Pre-processing

- Numeric functions
- Basic stats reductions
- Date/Time capabilities
- Categorical
- String processing
- Table wrangling
- Missing Data handling
- Summary visualizations:
  - Histogram/histogram2
  - Kernel density plot
  - Bin-scatter

Hundreds of pre-built functions

Machine Learning

- Linear Model
- Logistic Regression
- Discriminant analysis
- K-means
- PCA
- Random data sampling
- Summary statistics

Key statistics and machine learning algorithms

MATLAB framework for data that does not fit into memory

Example Use Case: Create a Predictive Model

- **Goal:** Contribute to a Ride Sharing project by creating a model to predict the cost of a Taxi Ride in New York City.

- **Considerations:**
  - Raw data are .csv taxi ride log files
  - File size ranges from 22 – 26MB
  - The full data set contains > 2 million rows
  - Start with linear regression (to facilitate prediction)
  - Scale up initial work

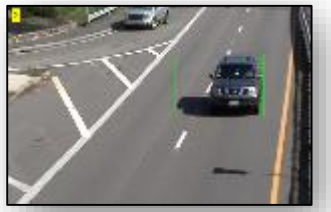
Agenda

- Data Types
- Tall Arrays for Big Data
- Machine Learning (for Everyone)
- Deploying your Analytics

When Might you Consider Machine Learning?

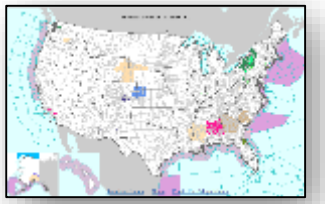
Problem is too complex for hand written rules or equations

Speech Recognition

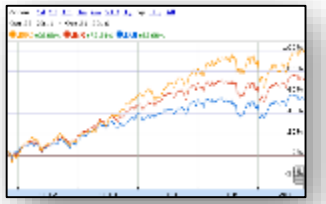

Object Recognition


Engine Health Monitoring

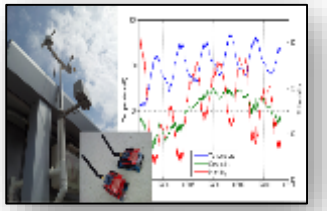
Program needs to adapt with changing data


Weather Forecasting

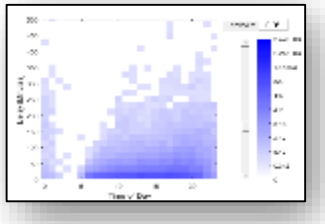

Energy Load Forecasting


Stock Market Prediction

Program needs to scale


IoT Analytics

Taxi Availability


Airline Flight Delays

Because algorithms can

learn complex non-linear relationships

update as more data becomes available

learn efficiently from very large data sets

Statistics and Machine Learning Toolbox

Making Machine Learning Easy and Accessible

R2015a

- Classification Learner App

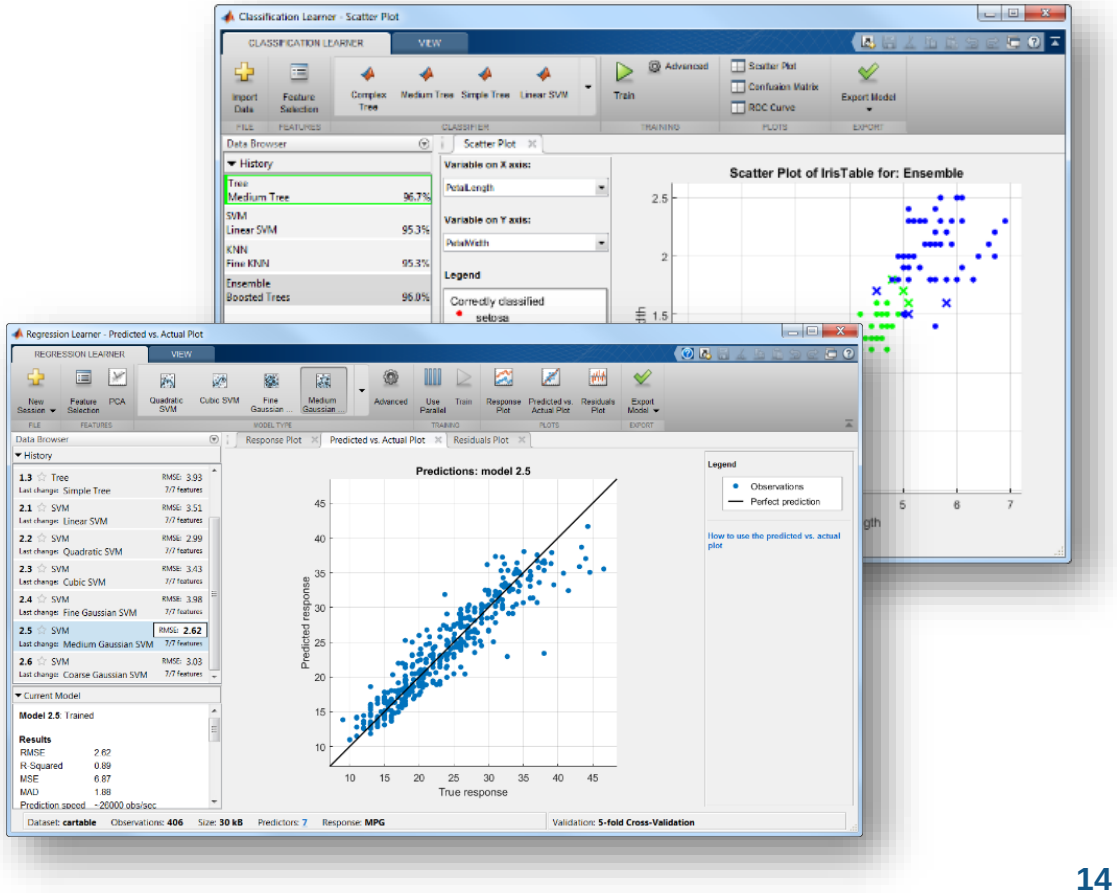
R2016b

- 1-click parallel computing
- Big data algorithms (using `tall` arrays)
- C code generation for predictive models (requires MATLAB Coder)
- New methods for feature selection and hyperparameter tuning

R2017a

- Regression Learner App

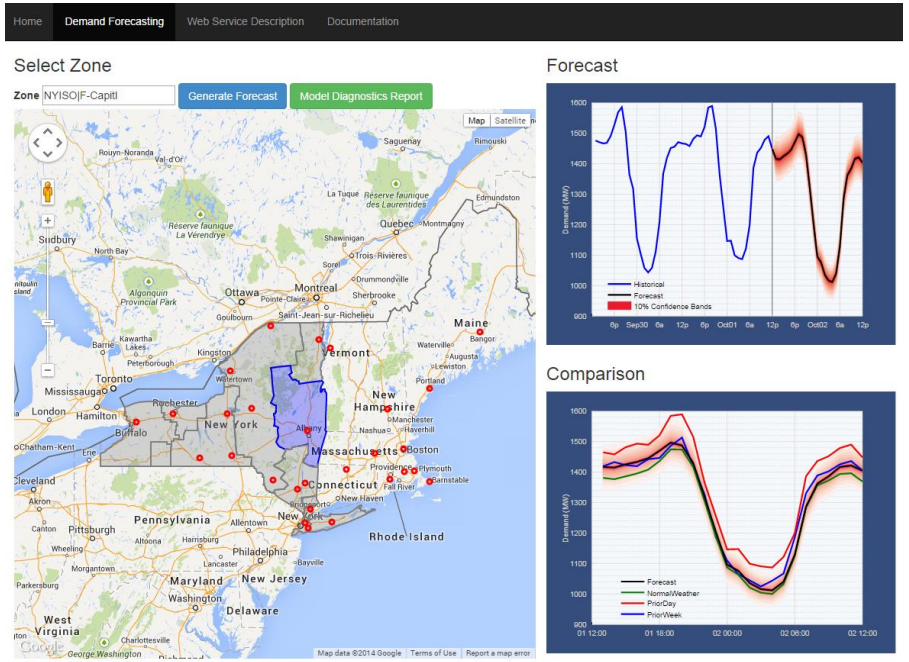
> ***"I would have never attempted machine learning if this app was not available."***

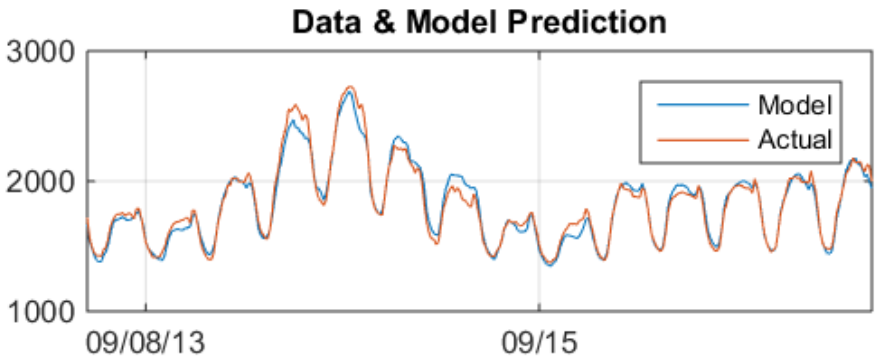


Example Use Case: Day-Ahead Load Forecasting

- **Goal:** Create and implement a tool for ***easy*** and ***accurate*** computation of day-ahead system load forecast

- **Considerations:**
  - Multiple data sources
  - Significant data clean up is required
  - Predictive model must be accurate
  - Easily deploy to production environment

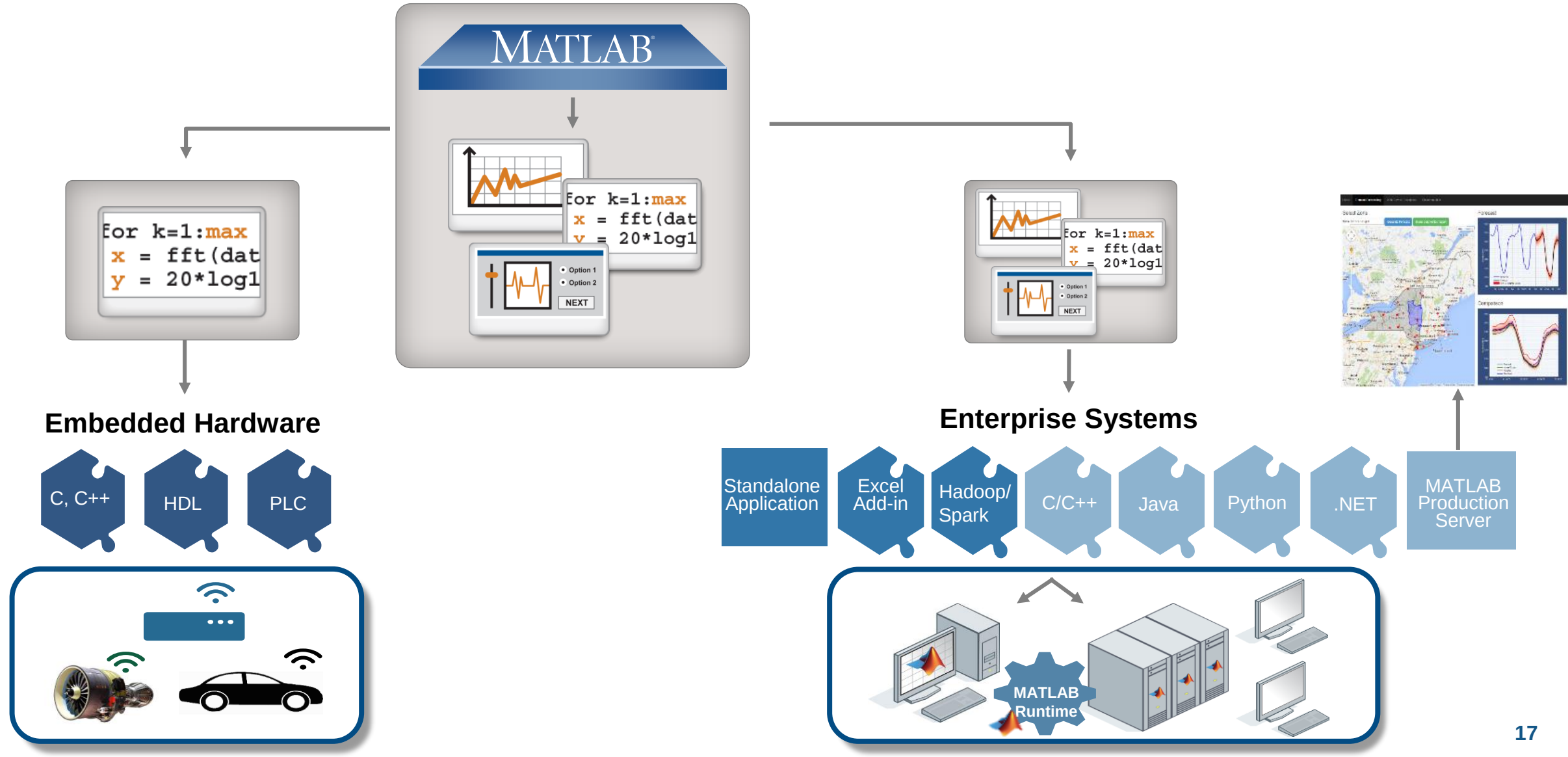




Agenda

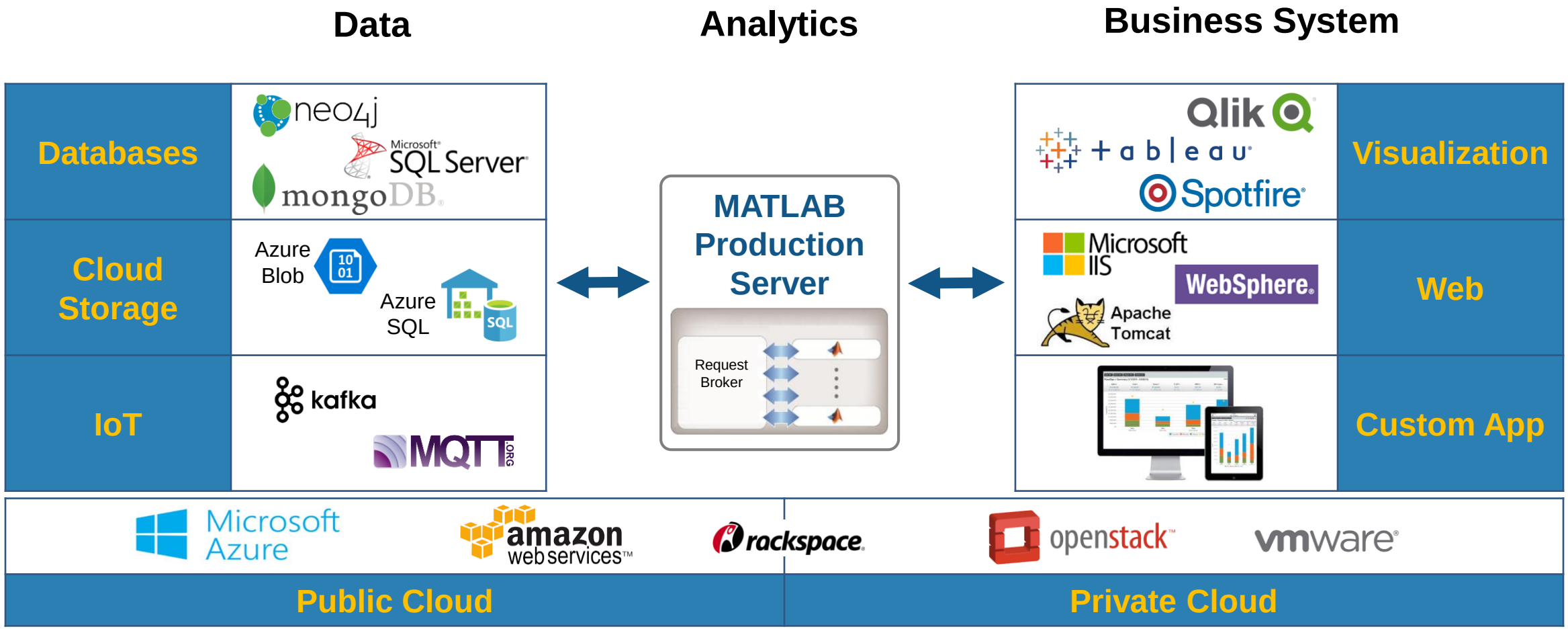
- Data Types
- Tall Arrays for Big Data
- Machine Learning (for Everyone)
- Deploying your Analytics

Integrate Analytics with Systems



Technology Stack for Enterprise Integration

Many possible solutions. MathWorks can help!

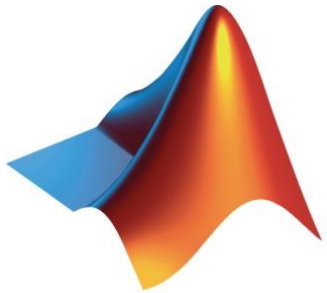


Key Takeaways

- MATLAB **data types** enable you to more **efficiently** tackle Data Analytics problems. **`tall` Arrays** for **out of memory** data sets.

- Use **MATLAB apps** to get started (or do more) **Machine Learning**.

- MATLAB based **Analytics run where you need them to** - Embedded or Enterprise IT systems.



MathWorks®

Accelerating the pace of engineering and science

© 2017 The MathWorks, Inc. MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See www.mathworks.com/trademarks for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.

© 2017 The MathWorks, Inc.