Building a Big Engineering Data Analytics System using MATLAB

International Data Science Conference

12 June 2017

Dmytro Martynenko – Application Engineering, The MathWorks GmbH

Dmitrij.Martynenko@mathworks.de

Example Setup at MathWorks

4G LTE

Server

Data Warehouse

MATLAB in Production:

- Enrich data
- File creation

Phone

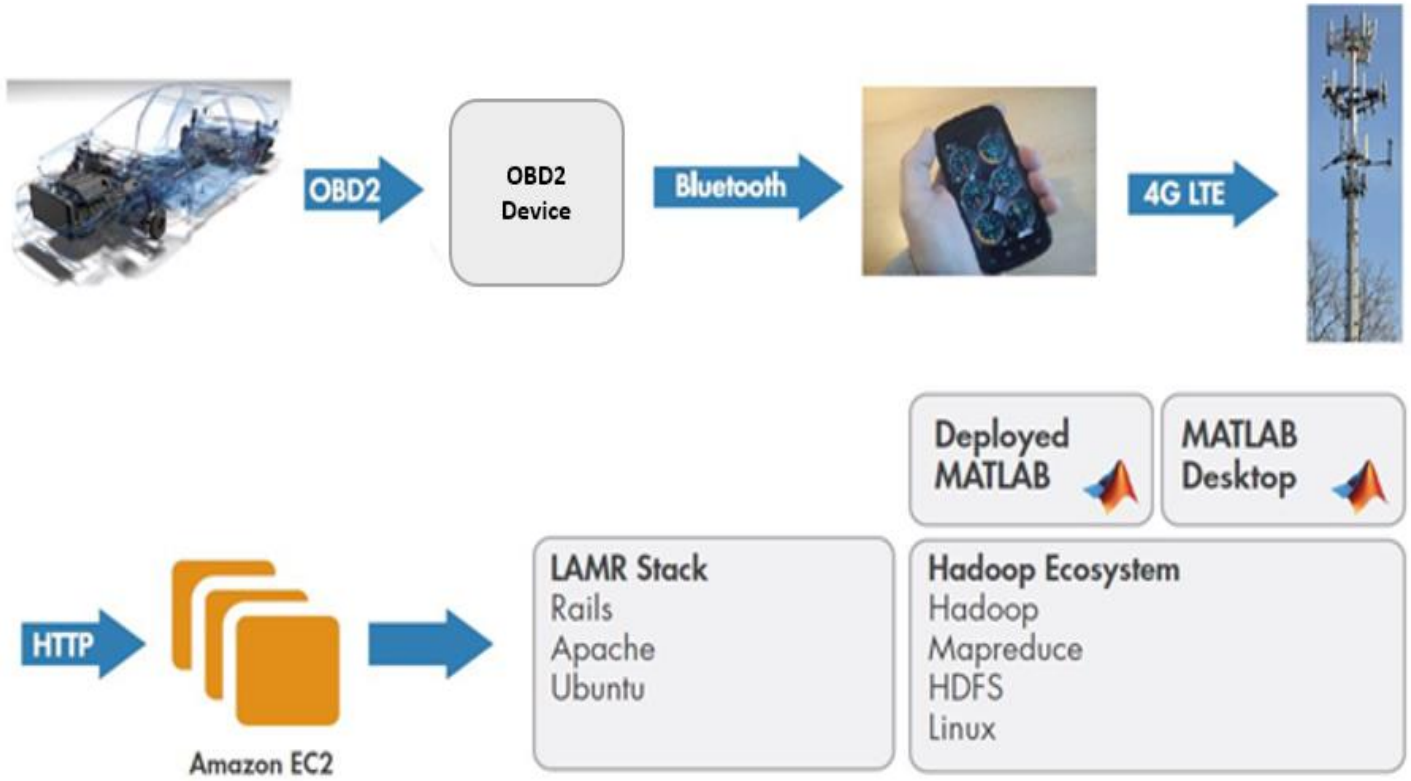
OBDII

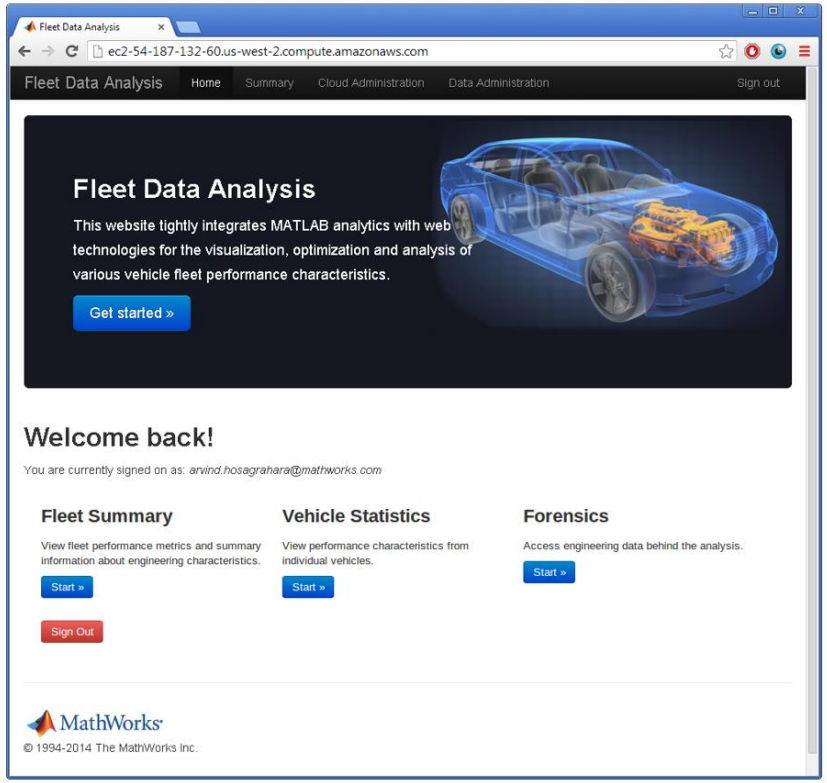
Bluetooth

Engineers

Fleet Data Analytics MathWorks Paper

[Engine Vehicle Design White Paper on mathworks.com]





Customer Example: Scania

Automatic Emergency Braking

Opportunity

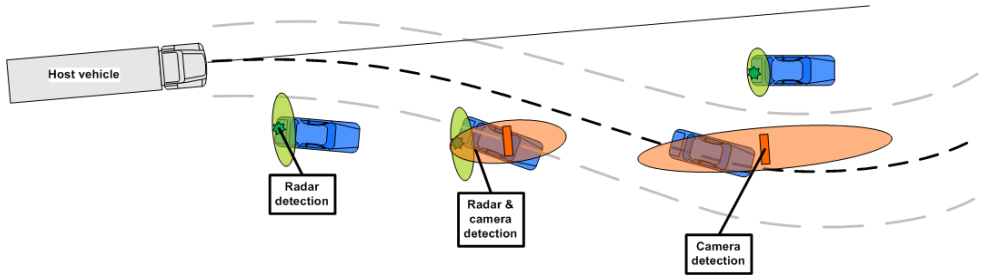
Real-time crash avoidance by detecting imminent collisions and automatically taking action

Analytics Use

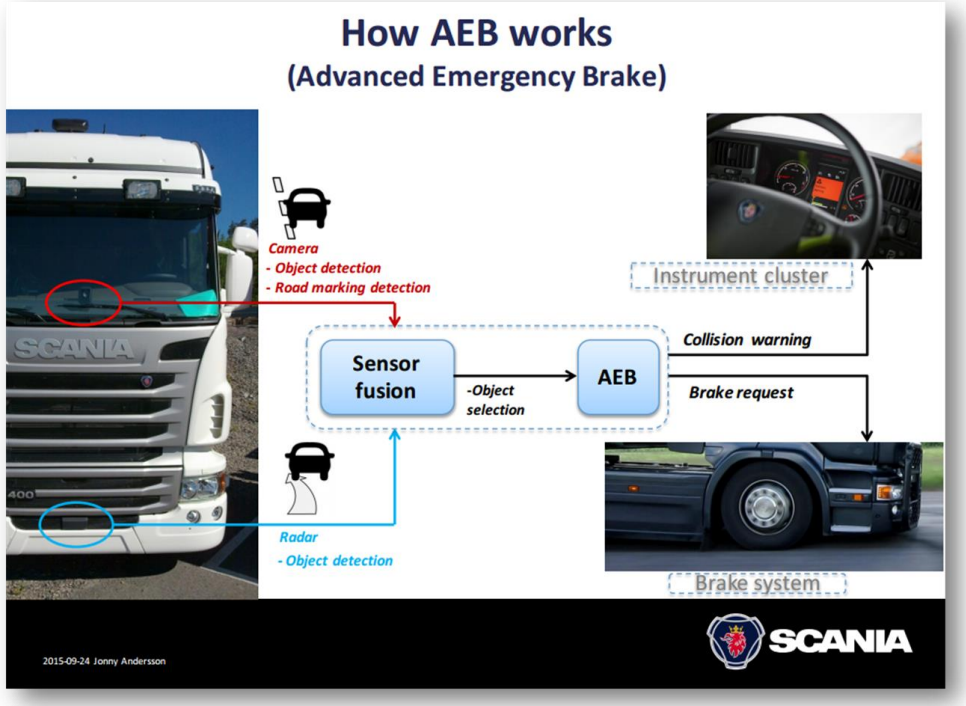
- **Data:** 80 TB – 1.5 million km of driving
- **Machine Learning:** Object detection
- **Control Systems:** Brake application
- **Test and Verify:** System model with simulated, recorded, and live data.

Benefit

- Reduced accidents
- Meet EU Regulations



Radar and camera for object-detection and real-time collision warning and braking.



Perception → Prediction → Action


50 km/h - sudden brake

Why MATLAB?

1

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

DATA
- Engineering, Scientific, and Field
- Business and Transactional

Data Analytics

2

MATLAB lets **engineers do Data Science** themselves

Embedded Systems
Developed with Model-Based Design

3

MATLAB Analytics run in **embedded systems** developed with **Model-Based Design**

Enterprise
IT Systems

4

MATLAB Analytics deploy to **enterprise IT systems**

Machine Learning Workflow

Access and Explore Data

- Files
- Databases
- Sensors

Preprocess Data

- Working with Messy Data
- Data Reduction/ Transformation
- Feature Extraction

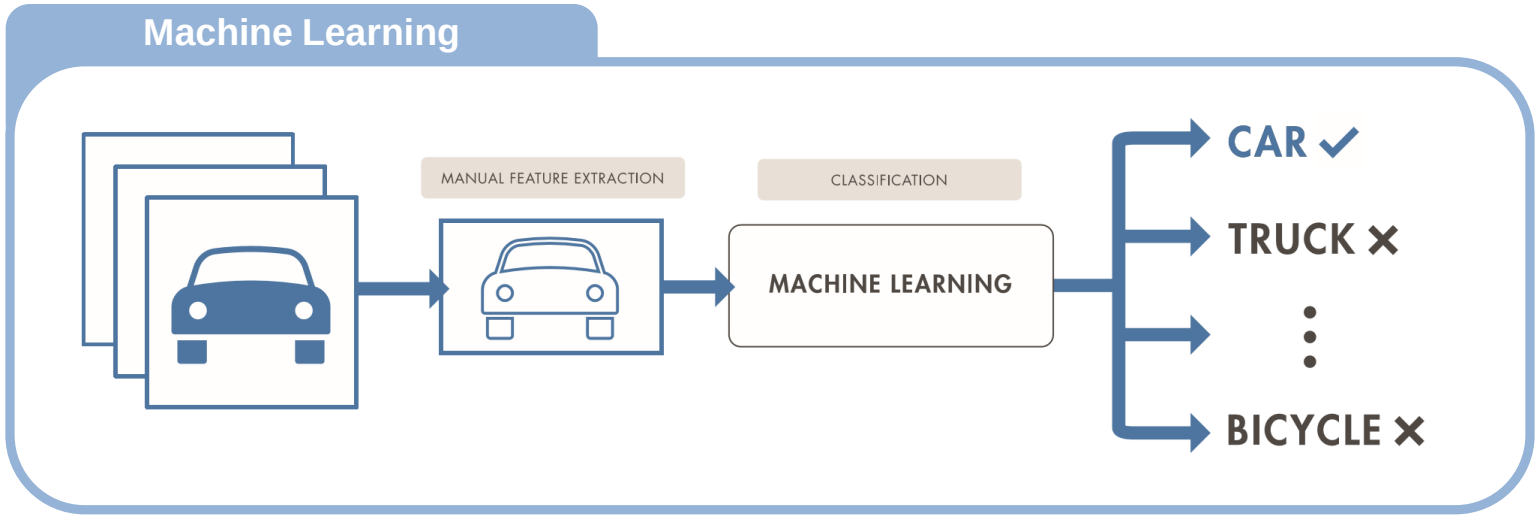
Develop Predictive Models

- Model Creation e.g. Machine Learning
- Parameter Optimization
- Model Validation

Integrate Analytics with Systems

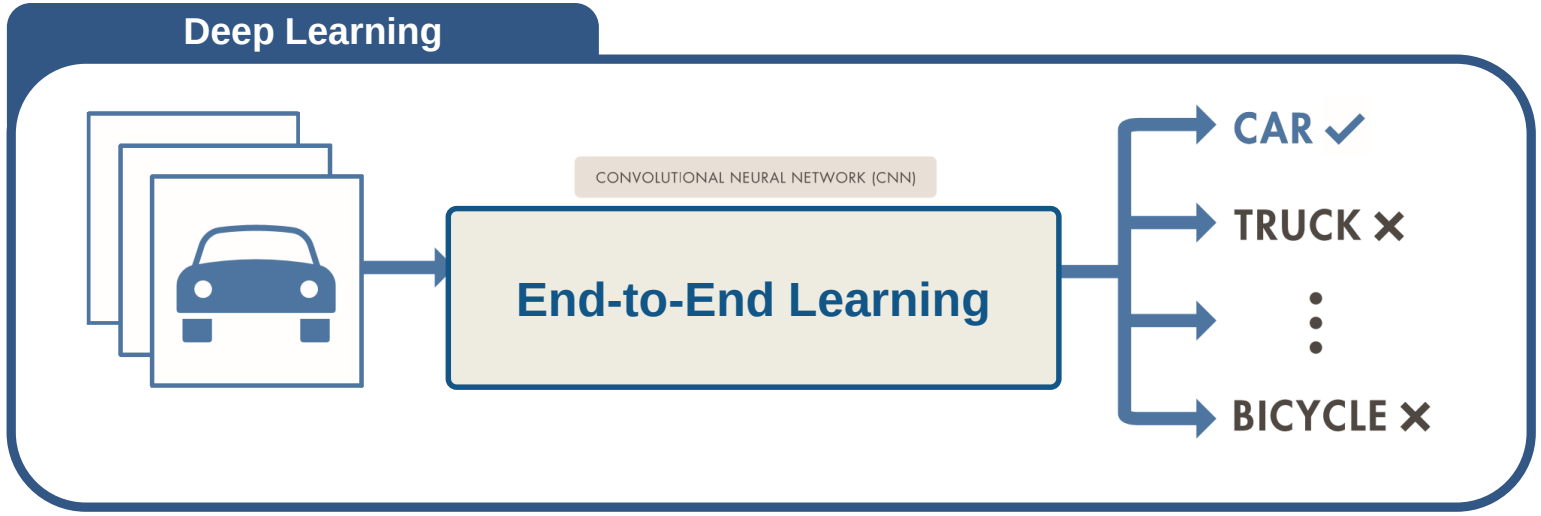
- Desktop Apps
- Enterprise Scale Systems
- Embedded Devices and Hardware

What is Deep Learning?



Machine Learning learns tasks using features extracted manually from data

Deep Learning learns both features and tasks directly from data



Deep Learning Workflow

Access & Explore

Training Data
- Large open datasets
- Recorded labeled data (e.g., images, video)

Pre-Trained Models
- Access state-of-art networks already trained on large datasets

Preprocess Data

Data Augmentation
- Crop, resize and rotate images – create more training data

Label Training Data
- For image data: draw ROI's, label individual pixels with labels

Develop Predictive Models

Train from Scratch
- Configure and train a deep network with massive amount of training data

Transfer Learning
- Tune pre-trained model for a different task with smaller datasets

Integrate Analytics to Systems

Share Models
- Publish models for others to use

Embedded Deployment
- Embedded processors
- FPGA

HPC
- Servers (multi-GPU)
- Clusters

How big is big?

What does “Big Data” even mean?

“Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them.”

Wikipedia

So, what's the (big) problem?

- Traditional tools and approaches won't work
  - **Getting** the data is hard; **processing** it is even harder
  - Need to learn **new tools** and **new coding styles**
  - Have to rewrite algorithms, often at a lower level of abstraction
- Quality of your results can be impacted
  - e.g., by being forced to work on a subset of your data

Big Data workflow

ACCESS

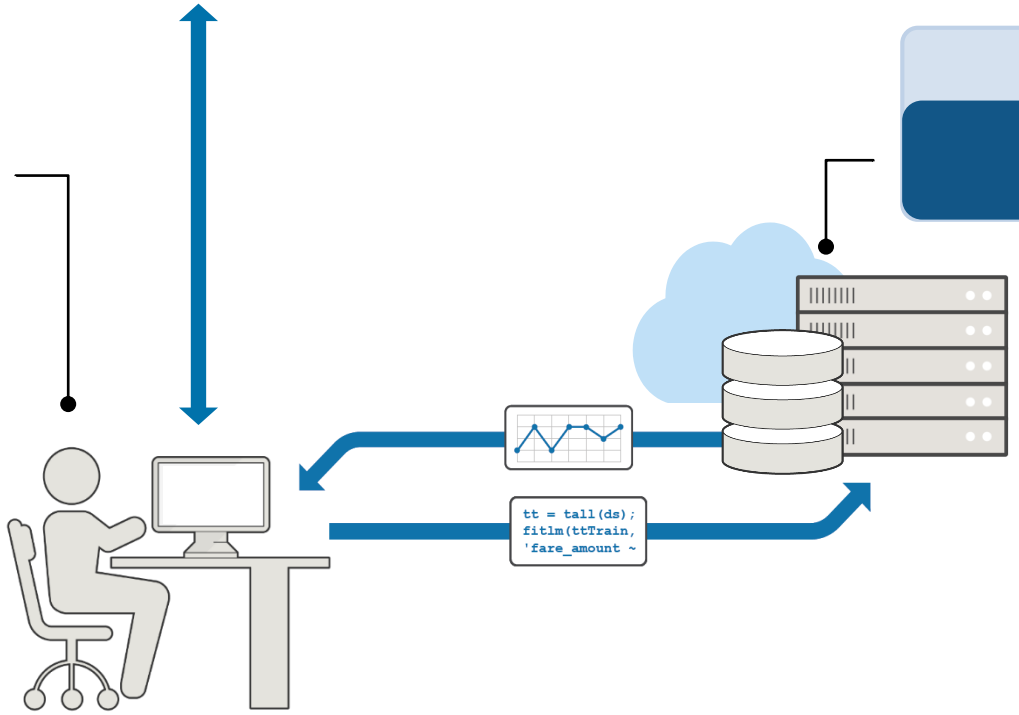
More data and collections of files than fit in memory

PROCESS AND ANALYZE

Adapt traditional processing tools or learn new tools to work with Big Data

SCALE

To Big Data systems like Hadoop



Big solutions

Wouldn’t it be nice if you could:

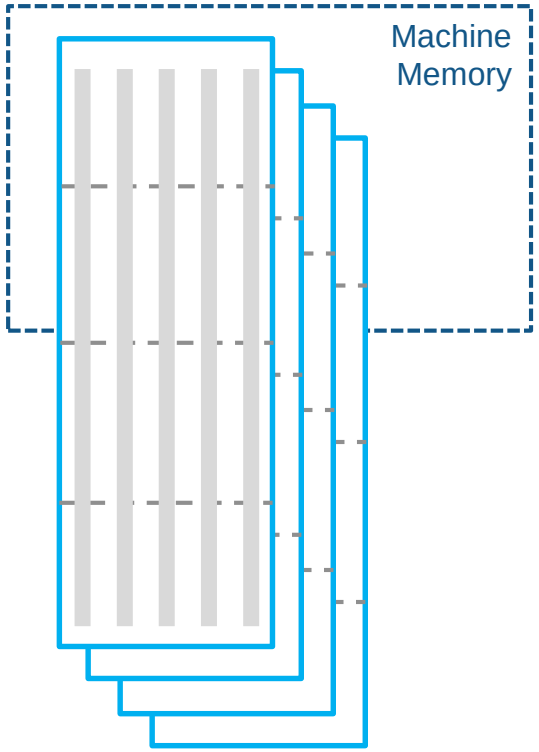
- Easily access data however it is stored
- Prototype algorithms quickly using small data sets
- Scale up to big data sets running on large clusters
- **Using the same intuitive MATLAB syntax you are used to**

Tall Arrays

Scaling your code to big data

R2016b

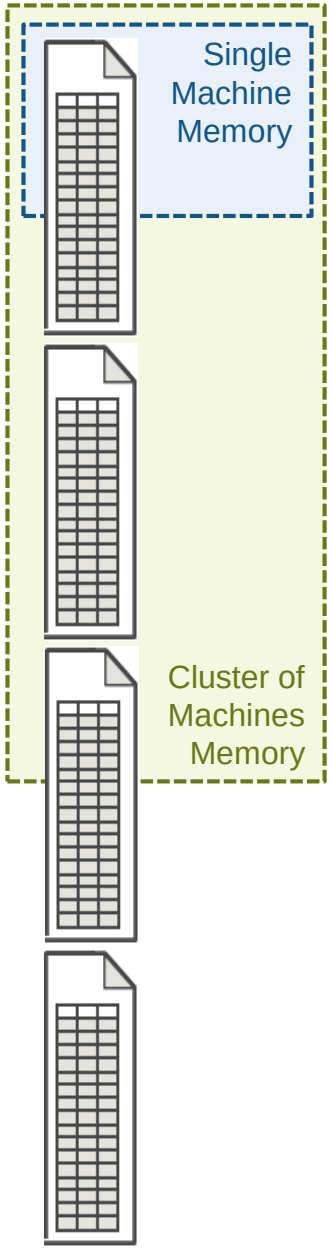
- Applicable when:
  - Data is **columnar** – with **many** rows
  - Overall data size is **too big to fit into memory**
  - Operations are mathematical/statistical in nature
- Statistical and machine learning applications
  - Hundreds of functions supported in MATLAB and Statistics and Machine Learning Toolbox



Tall Data

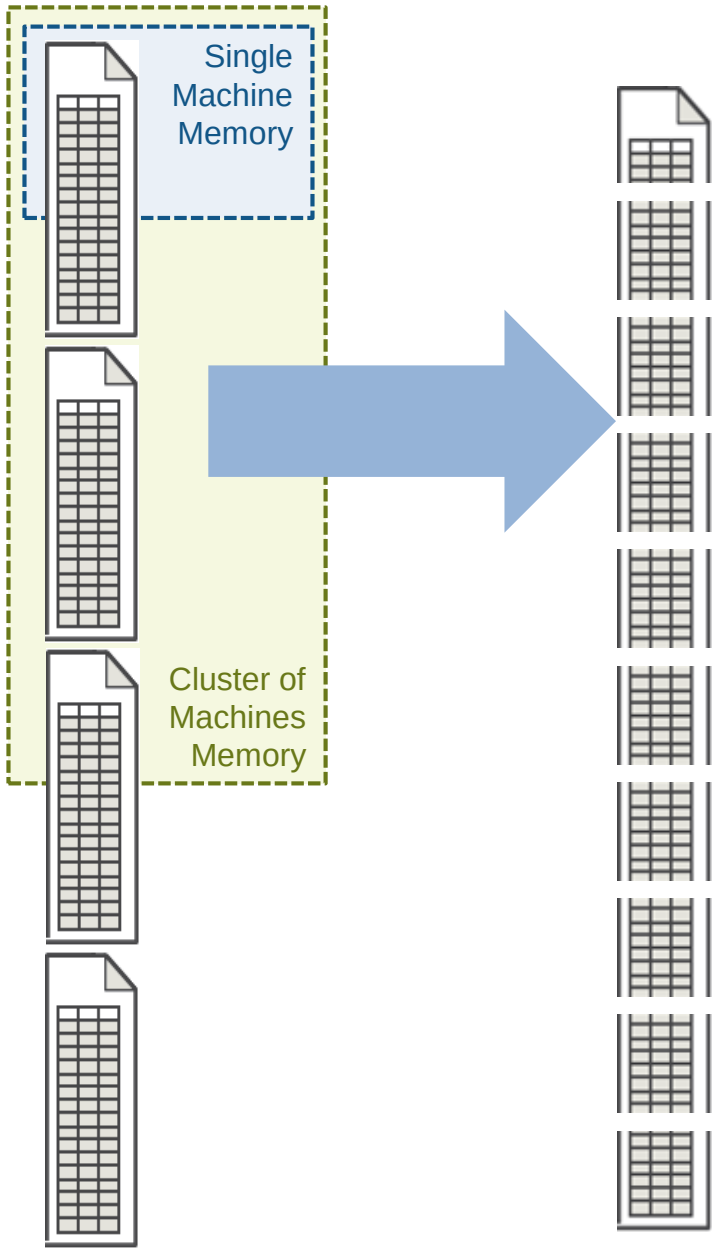
`tall` arrays R2016b

- Data is in one or more files
- Typically tabular data
- Files stacked vertically
- Data doesn't fit into memory
  (even cluster memory)



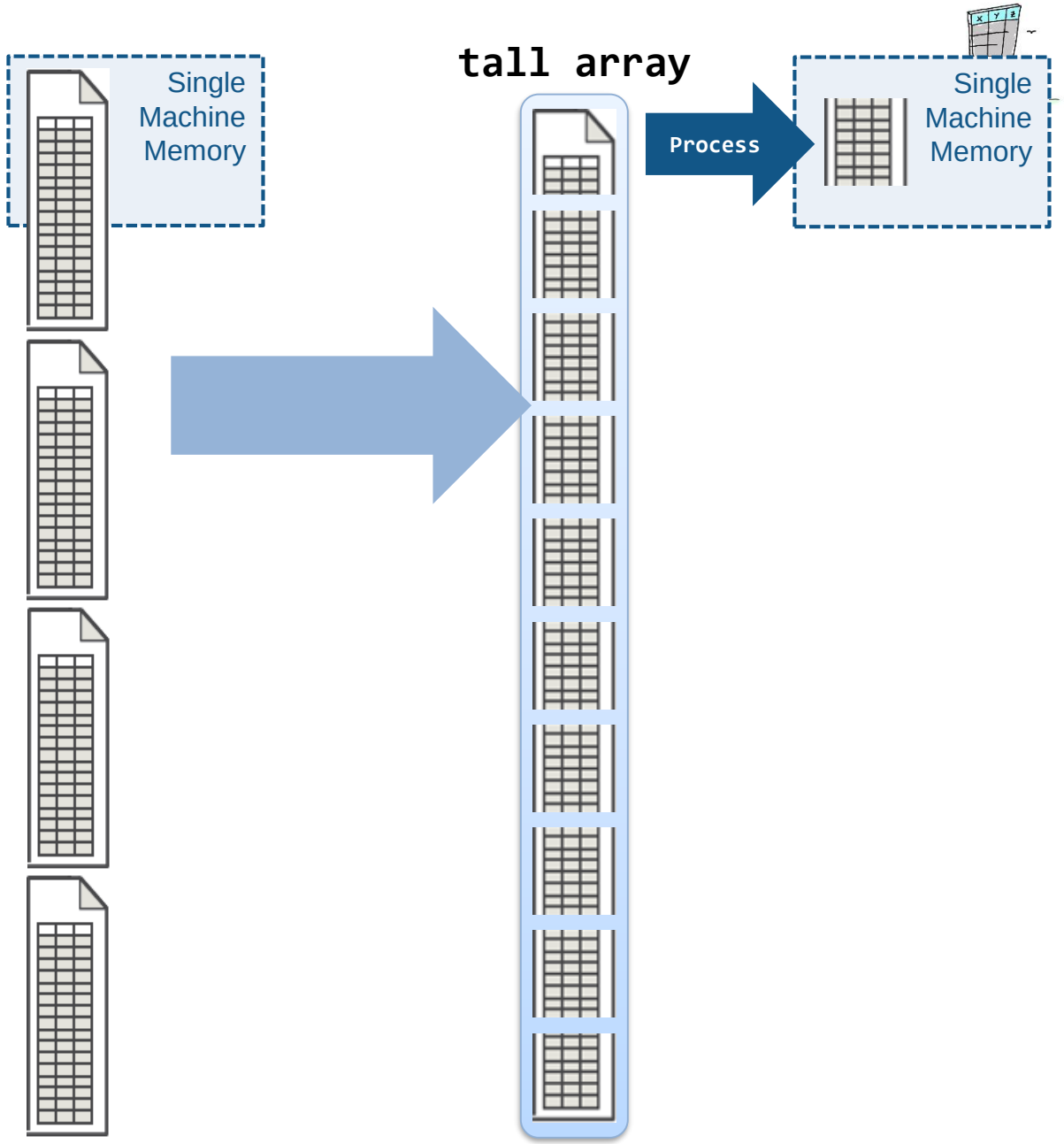
`tall` arrays R2016b

- Automatically breaks data up into small “chunks” that fit in memory



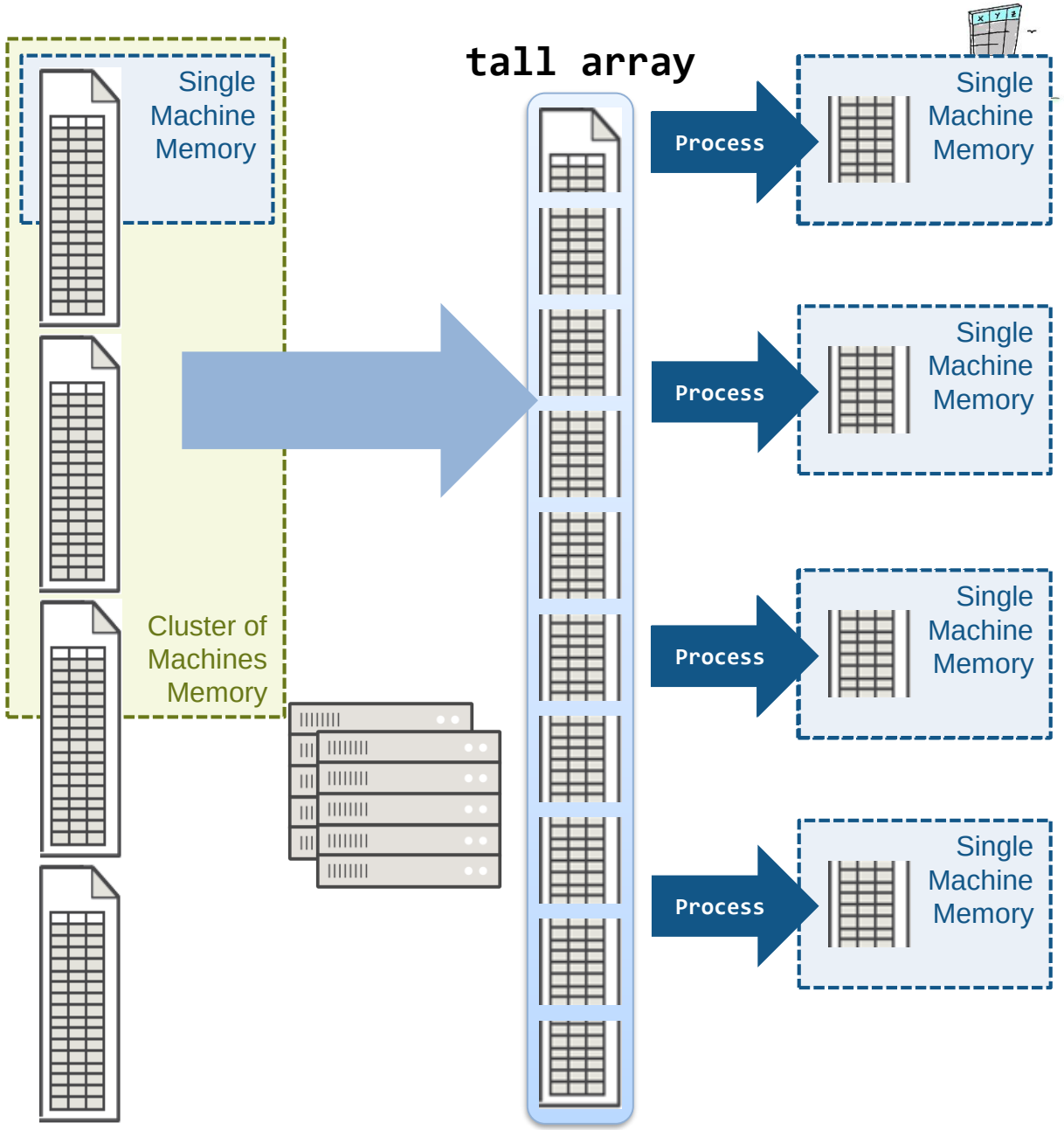
`tall` arrays R2016b

- “Chunk” processing is handled automatically
- 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



Example: Running on Spark + Hadoop

```
% Hadoop/Spark Cluster
numWorkers = 16;

setenv('HADOOP_HOME', '/dev_env/cluster/hadoop');
setenv('SPARK_HOME', '/dev_env/cluster/spark');

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


% Access the data
ds = datastore('hdfs://hadoop01:54310/datasets/taxiData/*.csv');
tt = tall(ds);
```

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`
- `timetable` (**`R2017a`**)
- `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
- Decision trees (**`R2017a`**)

Key statistics and machine learning algorithms

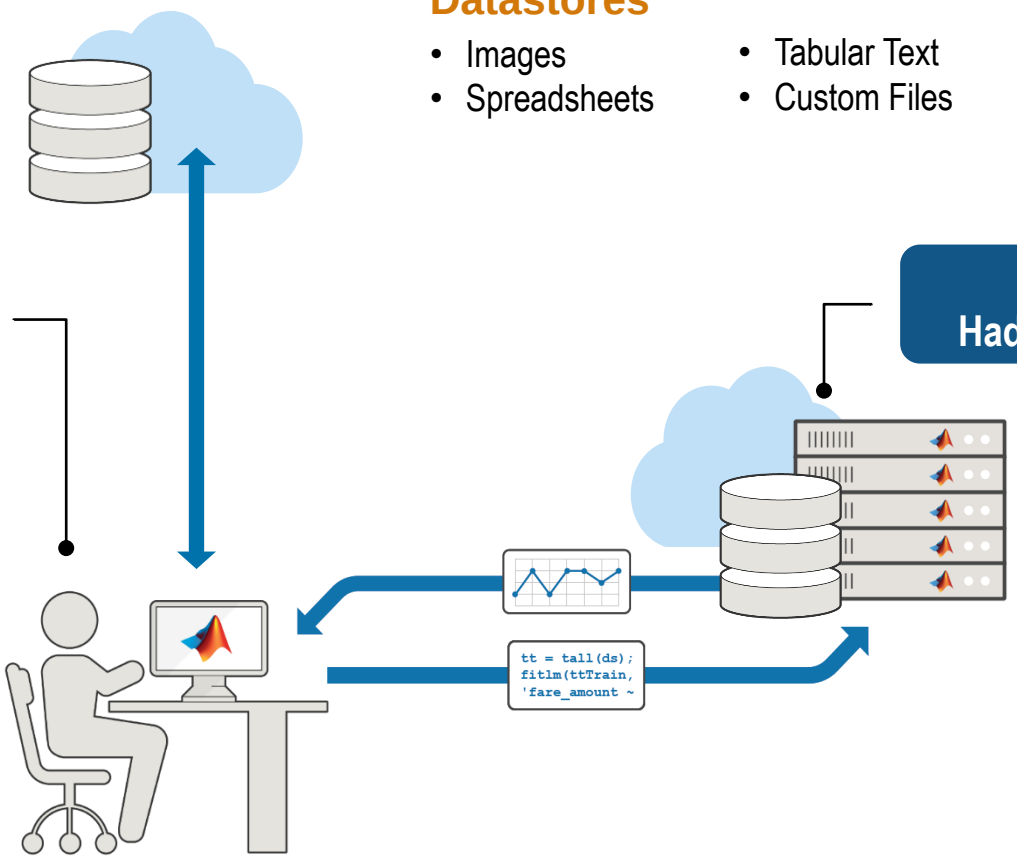
MATLAB programming for data that does not fit into memory

Big Data Capabilities in MATLAB

Access data and collections of files that do not fit in memory

Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)



Purpose-built capabilities for domain experts to work with big data

Tall Arrays

- Math
- Statistics
- Visualization
- Machine Learning

GPU Arrays

- Matrix Math
- Image Processing

Deep Learning

- Image Classification

Scale to compute clusters and Hadoop/Spark for data stored in HDFS

Tall Arrays

- Math, Stats, Machine Learning on Spark

Distributed Arrays

- Matrix Math on Compute Clusters

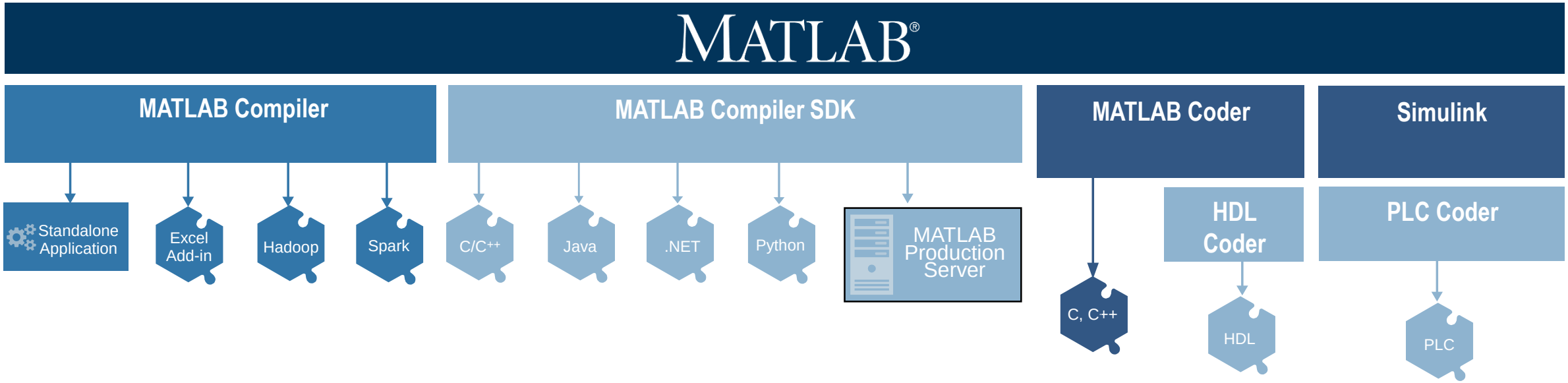
MDCS for EC2

- Cloud-based Compute Cluster

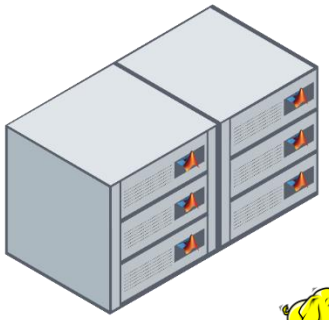
MapReduce

MATLAB API for Spark

MathWorks Products for Operationalizing Analytics



Desktop Users



Enterprise IT Systems

Embedded Systems (Including Edge Devices)

- Microcontrollers
- DSP chips
- FPGAs
- ARM-based
- Low-cost:
  - Arduino
  - Raspberry Pi
  - BeagleBone

Why MATLAB?

1

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

DATA
- Engineering, Scientific, and Field
- Business and Transactional

Data Analytics

2

MATLAB lets **engineers do Data Science** themselves

Embedded Systems
Developed with Model-Based Design

3

MATLAB Analytics run in **embedded systems** developed with **Model-Based Design**

Enterprise
IT Systems

4

MATLAB Analytics deploy to **enterprise IT systems**

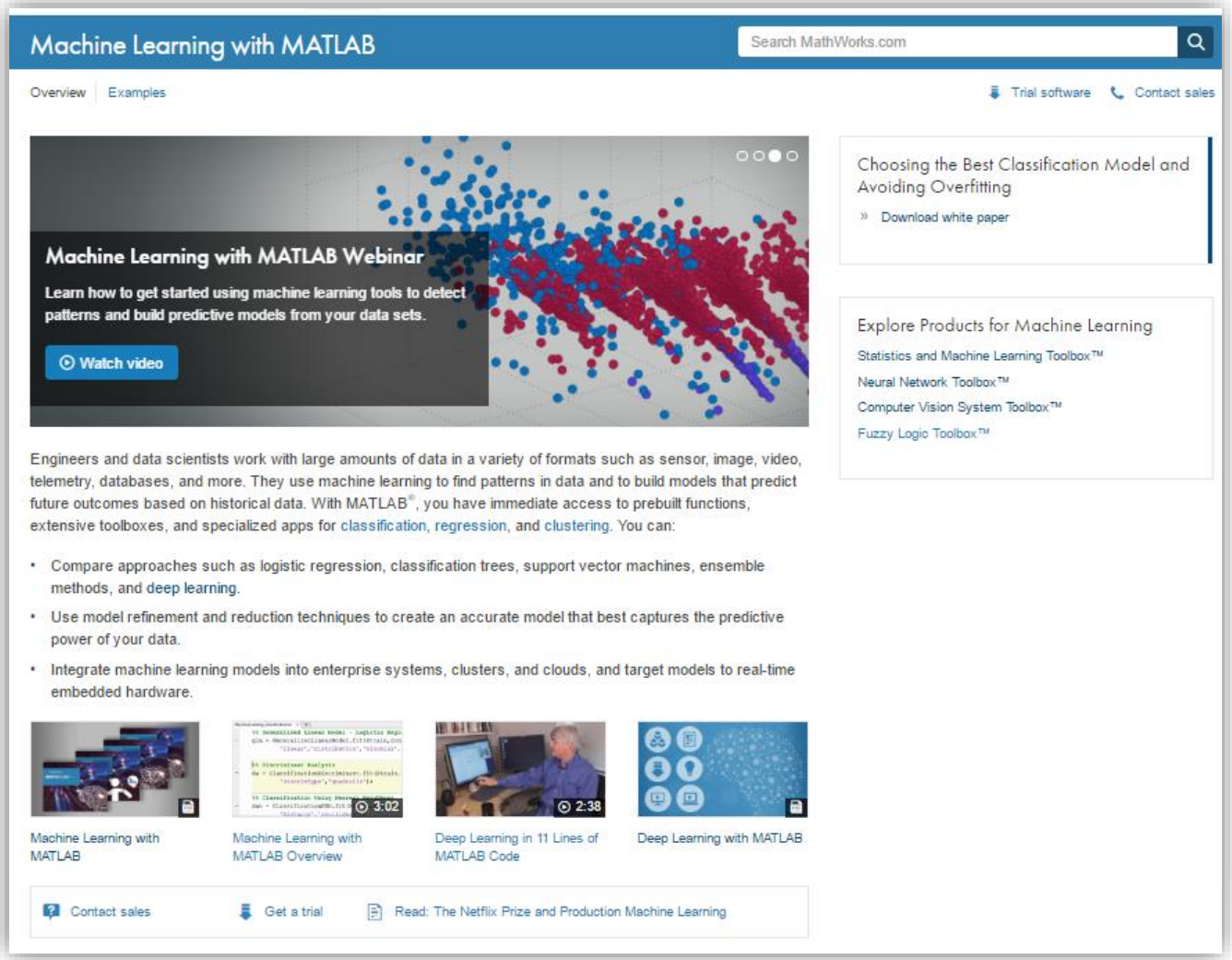
Summary

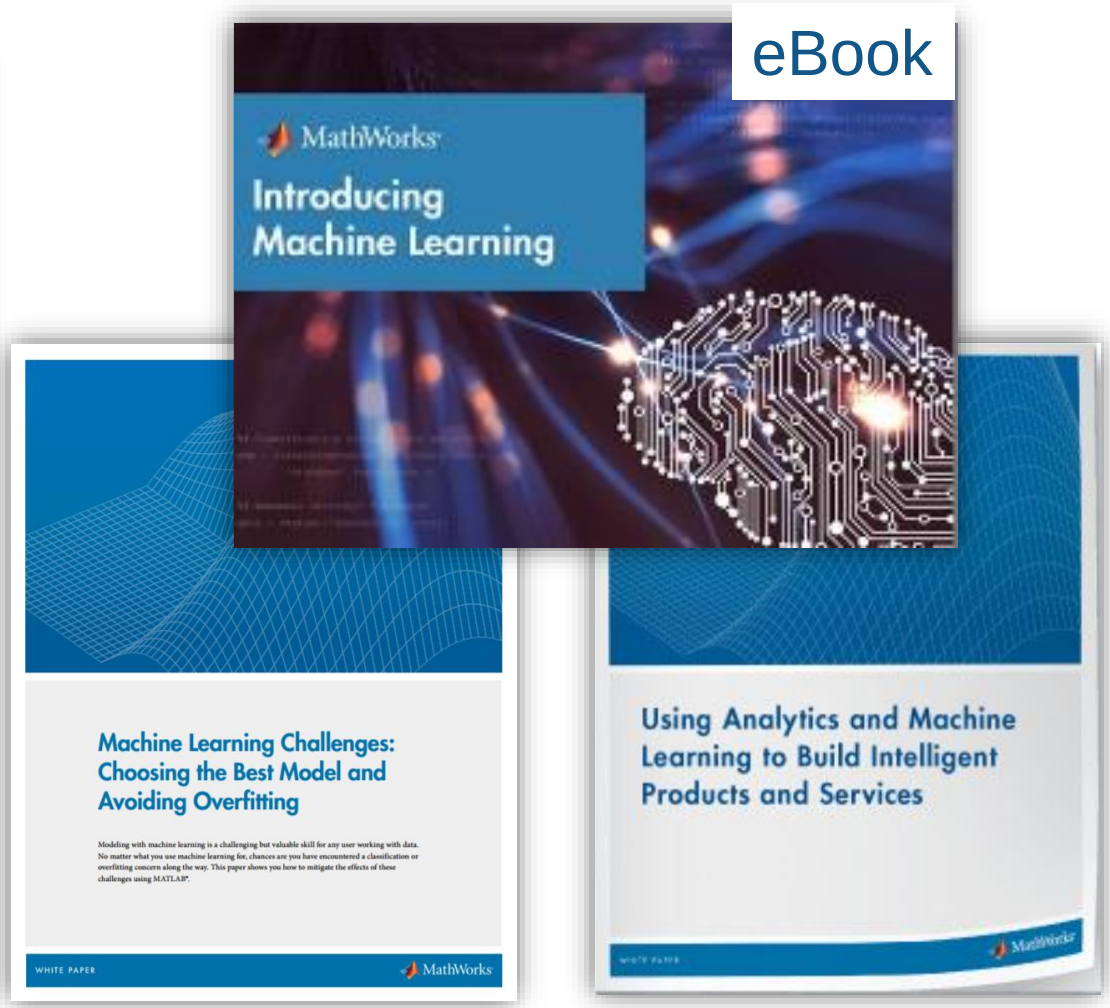
- MATLAB makes it easy, convenient, and scalable to work with big data
  - **Access** any kind of big data from any file system
  - Use tall arrays to **process and analyze** that data on your desktop, clusters, or on Hadoop/Spark

There's no need to learn big data programming or out-of-memory techniques -- simply use the same code and syntax you're already used to.

Resources to learn and get started

mathworks.com/machine-learning





Learn More

Big Data with MATLAB

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

MapReduce and Hadoop

mathworks.com/discovery/matlab-mapreduce-hadoop