BIG DATA: Data Analytics with MATLAB

Christophe POUILLOT
Senior Consultant
MathWorks
christophe.pouillot@mathworks.fr

Definition of Big Data

Data "so **large** and **complex** that it becomes difficult to process using on-hand database management tools or traditional data processing applications." from wikipedia

Volume
Velocity
Variety

Name	Symbol	Value
gigaoctet	Go	10^9
téraoctet	To	10^{12}
pétaoctet	Po	10^{15}
exaoctet	Eo	10^{18}
zettaoctet	Zo	10^{21}
yottaoctet	Yo	10^{24}

Social network: xxTo/day

Worldwide: 2,8Zo/year (2012)

Data structured or not from different sources: Web/Text/Image mining

Data containers

- Collection of files

- Databases

- Huge single file

Big Data Analytics with MATLAB

Memory and Data Access

- 64-bit processors
- Memory Mapped Variables
- Disk Variables
- Databases
- Datastores

Analysis

- Machine learning
- Analysis domain
- Statistics

Programming Constructs

- Streaming
- Block Processing
- Parallel-for loops
- GPU Arrays
- SPMD and Distributed Arrays
- MapReduce

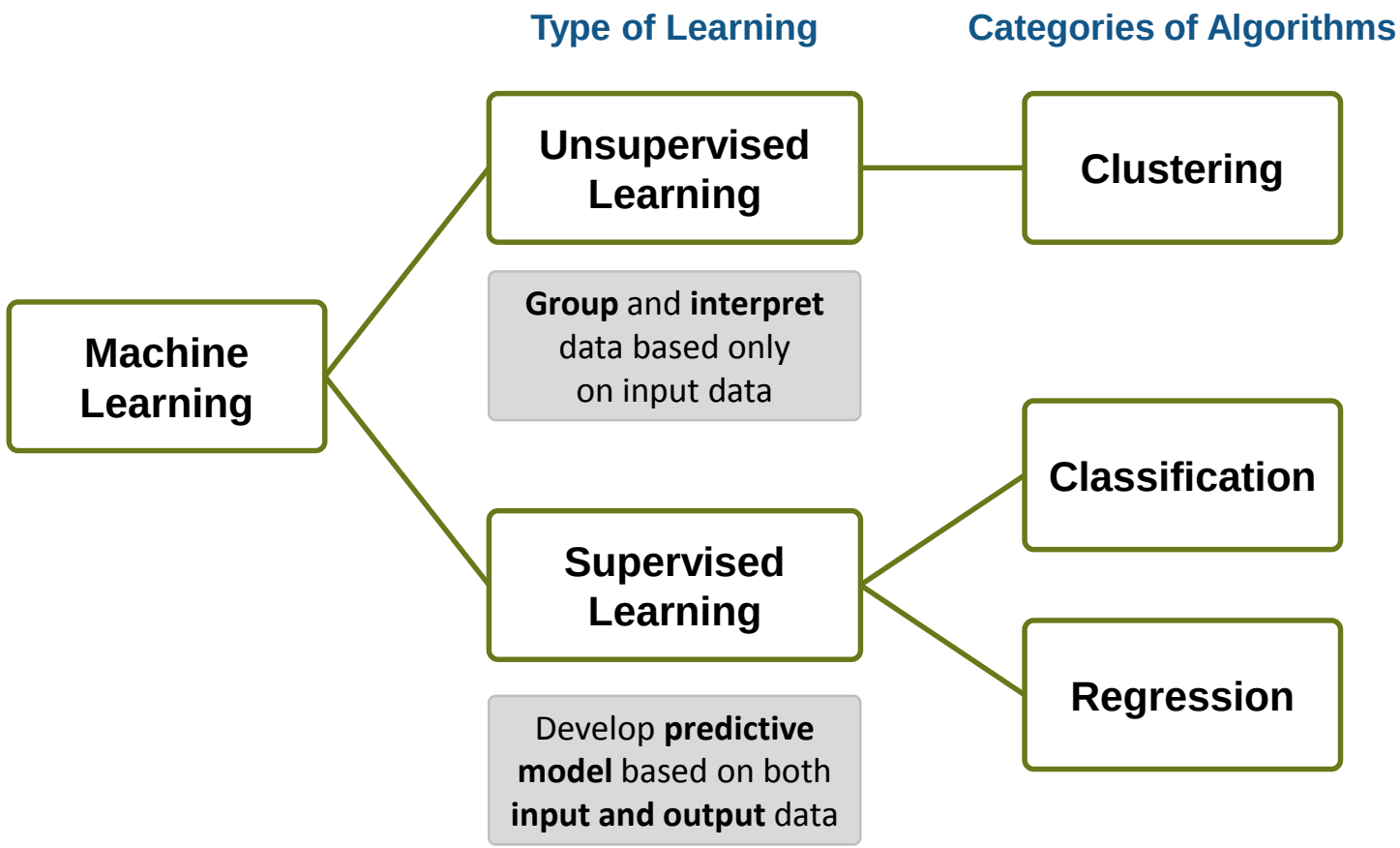
Platforms

- Desktop (Multicore, GPU)
- Clusters
- Cloud Computing (MDCS on EC2)
- Hadoop

Agenda

- Machine Learning
- Datastore/MapReduce
- Integration with Hadoop
- Databases
- Huge file
- Deployment
- Key takeaways

Overview – Machine Learning



Demo: Machine Learning

Agenda

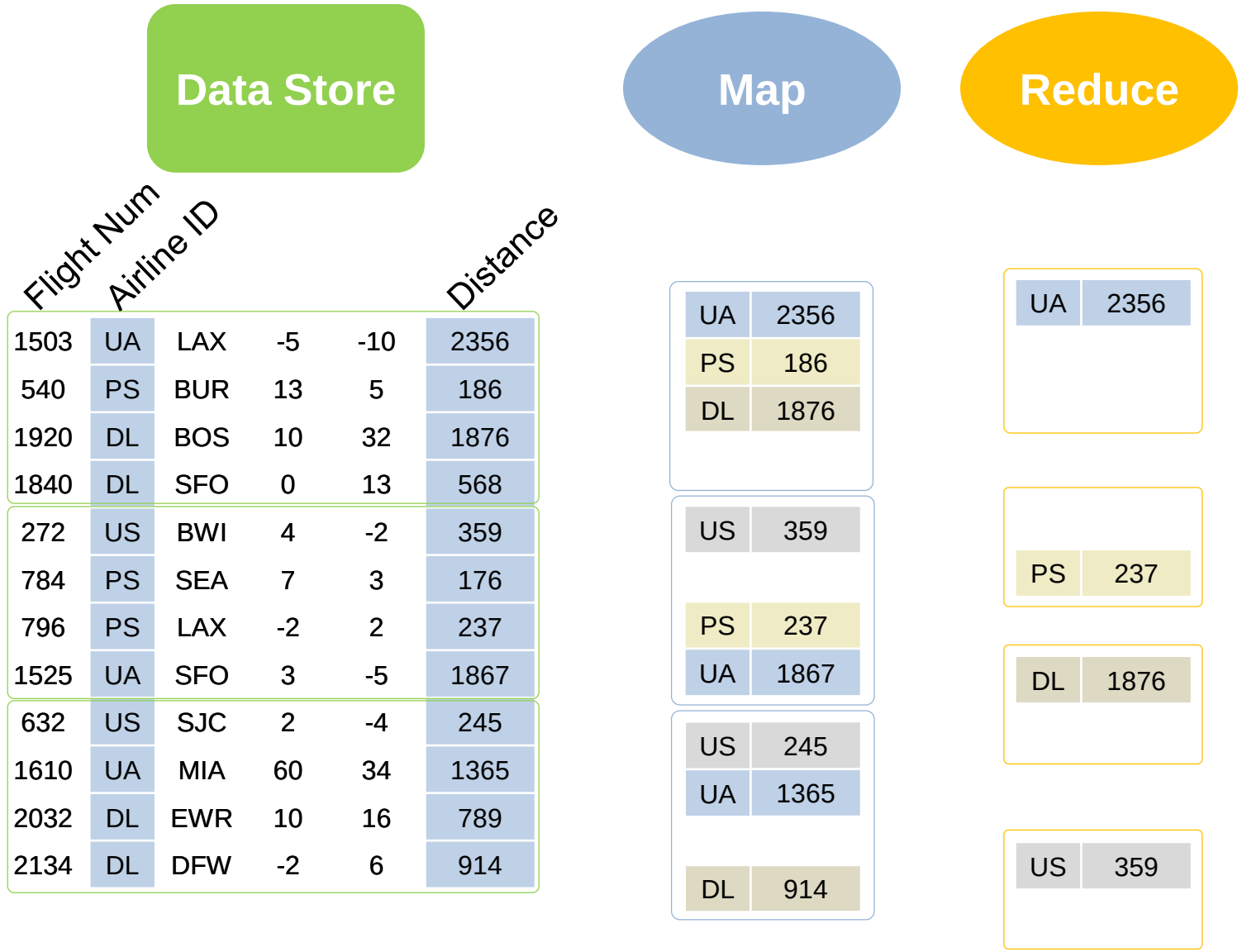
- Machine Learning
- Datastore/MapReduce
- Integration with Hadoop
- Databases
- Huge file
- Deployment
- Key takeaways

DataStore

datastore	Import text files & collections of text files that don't fit into memory

```
ds = datastore('file1.mat');

ds = datastore('*.csv');

ds = datastore('/shared/data_repository/');

ds = datastore('hdfs://myserver:7867/data/file1.txt');

ds = datastore({'/shared01/','/shared02/'});


while hasdata(ds)
    T = read(ds);
end
```

Demo mapreduce

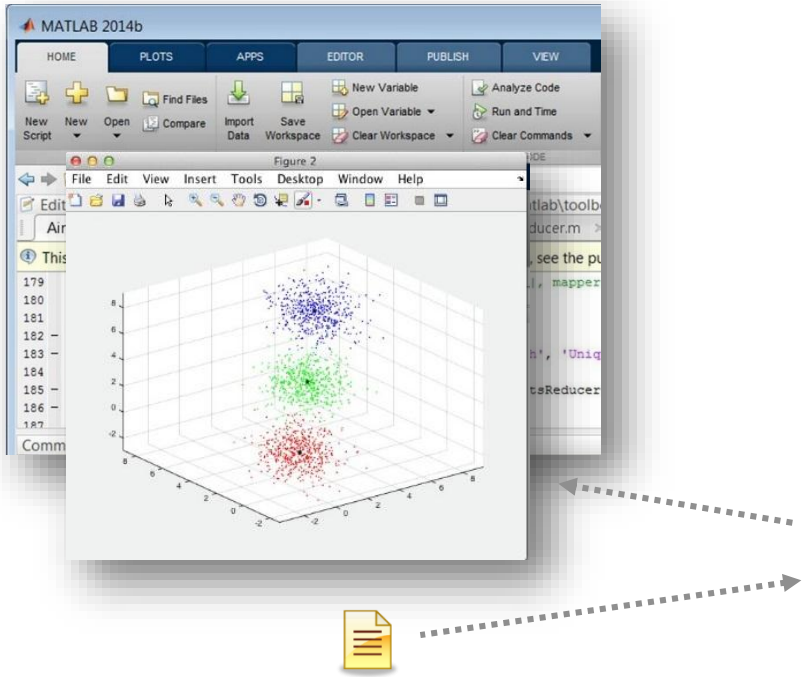


Demo: datastore/mapreduce

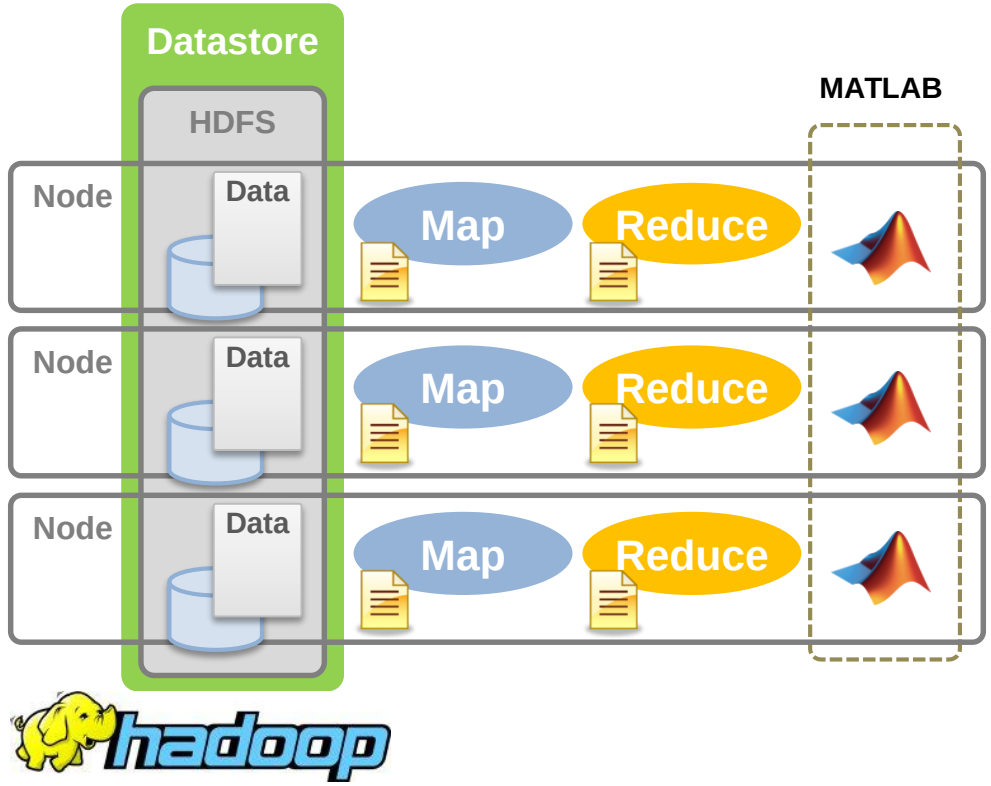
Agenda

- Machine Learning
- Datastore/MapReduce
- Integration with Hadoop
- Databases
- Huge file
- Deployment
- Key takeaways

Integration with Hadoop



map.m
reduce.m
main.m



Agenda

- Machine Learning
- Datastore/MapReduce
- Integration with Hadoop
- Databases
- Huge file
- Deployment

- Key takeaways

Databases

Relational database (ODBC/JDBC-compliant)

DatabaseDatastore (DataBase Toolbox)

```
conn = database.ODBCConnection('MySQL','username','pwd');
dbds = datastore(conn, 'select * from productTable');
```

NOSQL database

MATLAB calls external functions:

- ✓ C/C++ shared libraries
- ✓ JAVA libraries
- ✓ .NET libraries
- ✓ COM Objects (ActiveX…)
- ✓ Python libraries
- ✓ WSDL Web Service

Agenda

- Machine Learning
- Datastore/MapReduce
- Integration with Hadoop
- Databases
- Huge file
- Deployment

- Key takeaways

Huge flat files: Mapping memory within MATLAB

Read/write variables in files, without loading into memory

matfile: mat files

```
m = matfile('myFile.mat');
z = m.x(85:94,85:94); % read from disk
m.x(81:100,81:100) = magic(20); % write on disk
```

memmapfile: any files

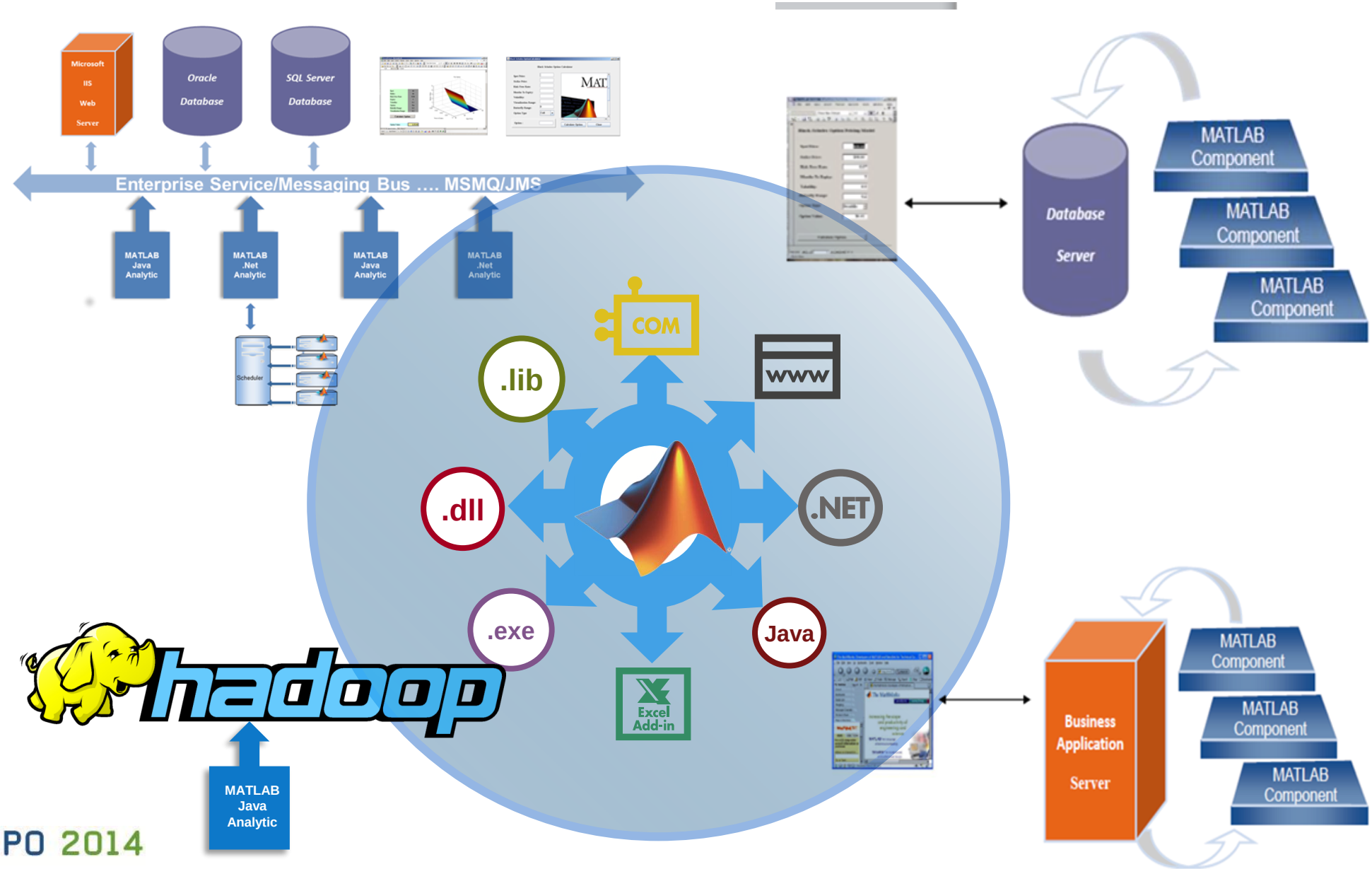
```
m = memmapfile('records.dat','Offset',9000, 'Format','int32');
z = m.Data(85:94,85:94); % read from disk
m.Data(81:100,81:100) = magic(20); % write on disk
```

Agenda

- Machine Learning
- Datastore/MapReduce
- Integration with Hadoop
- Databases
- Huge file
- Deployment

- Key takeaways

Deployment: Compiling MATLAB to go everywhere



Agenda

- Machine Learning
- Datastore/MapReduce
- Integration with Hadoop
- Databases
- Huge file
- Deployment

- Key takeaways

Key takeaways

- ✓ MATLAB is the framework for BIG DATA analytics
- ✓ MathWorks services can help you:

- ❑ Training services
- ❑ Consulting services

http://www.mathworks.fr/discovery/big-data-matlab.html?s_tid=gn_loc_drop

Questions?