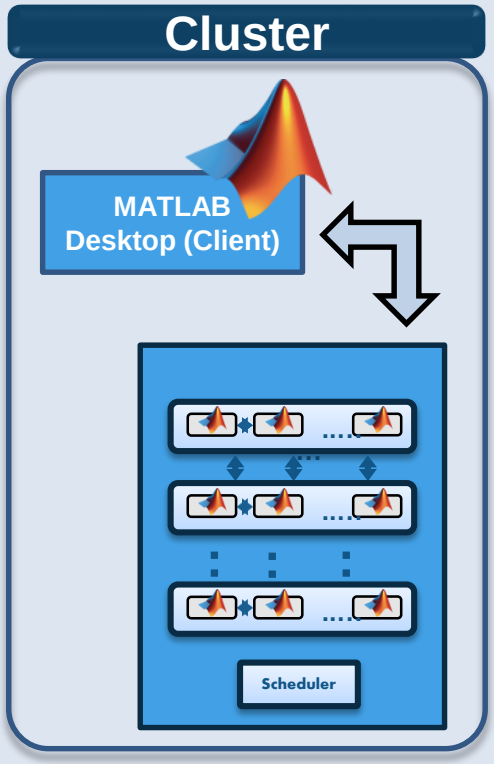
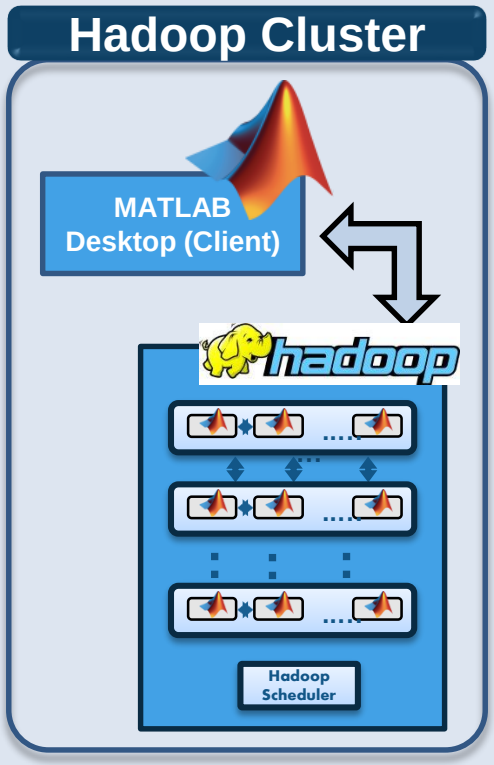
Tackling Big Data with MATLAB

Francesca Perino
Application Engineering Team - MathWorks

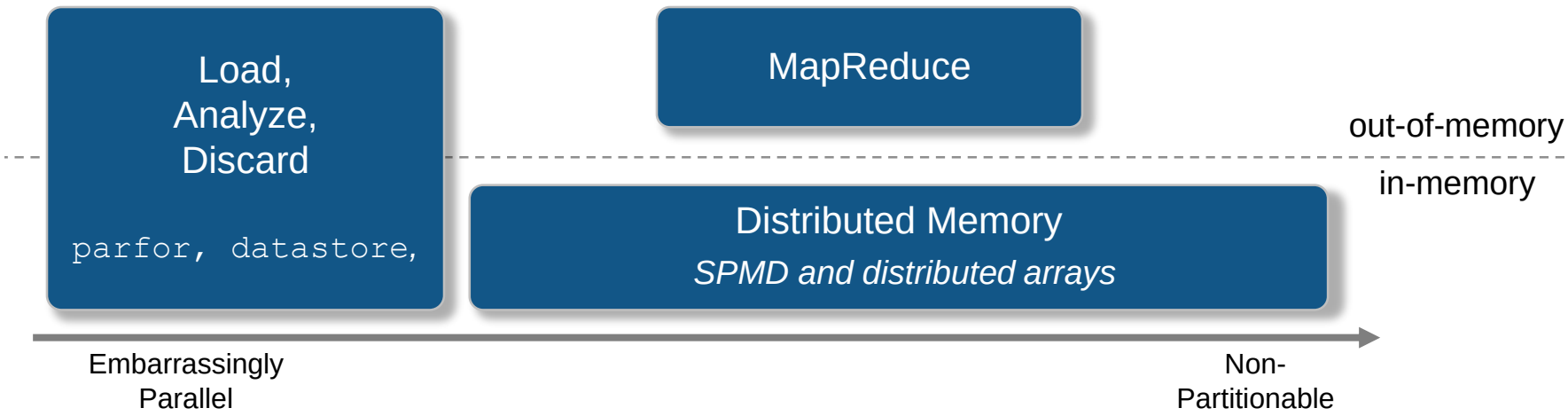
Running into “Big Data” Issues?

- “Out of memory”
  - Running out of address space
- Performance
  - Takes too long to process all of your data
- Slow processing (swapping)
  - Data too large to be efficiently managed between RAM and virtual memory

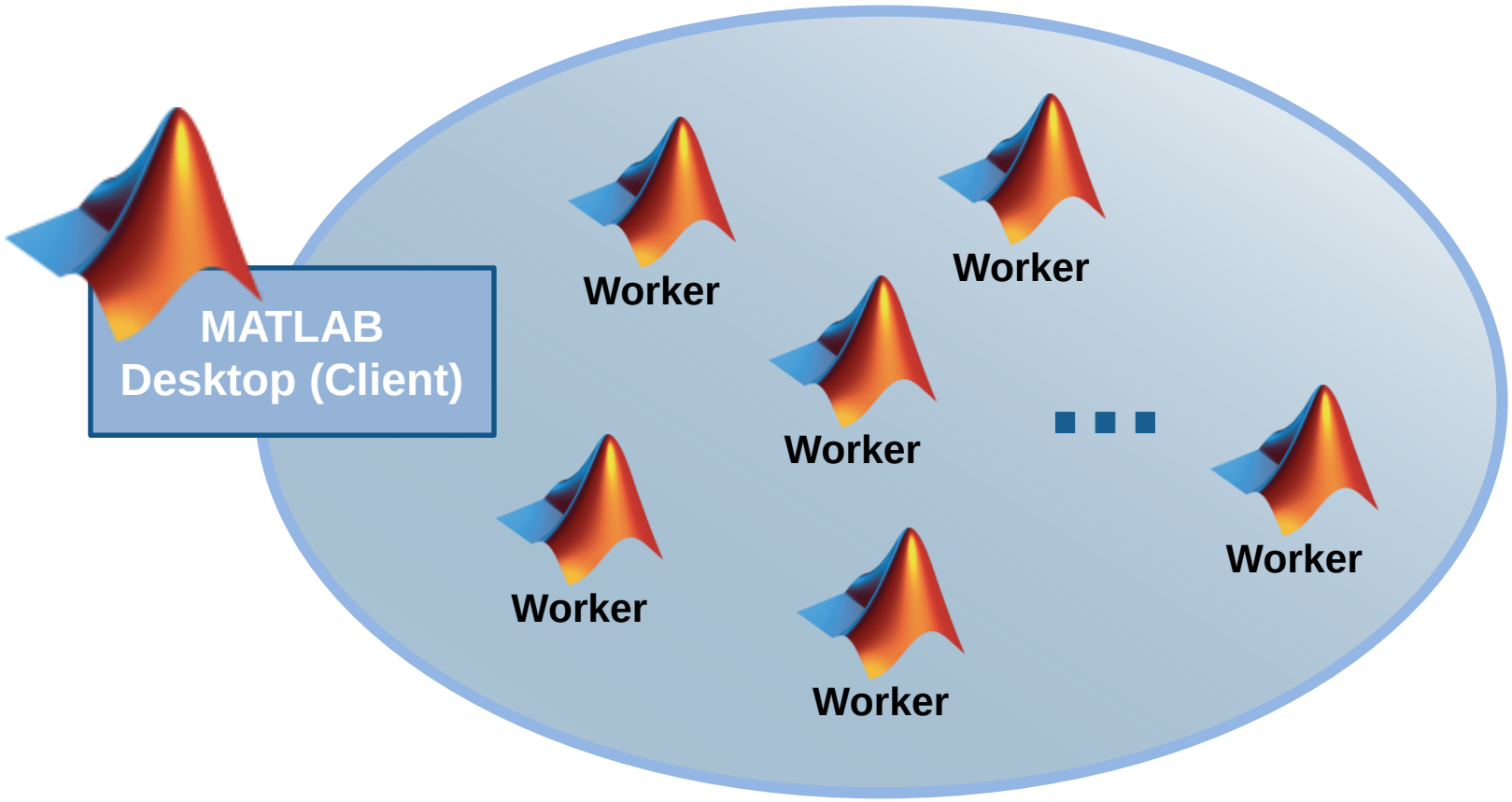
Options for Handling Large Data

Platform	Desktop Only	Desktop + Cluster	Desktop + Hadoop
			
Data Size	**100's MB -10's GB**	**100's MB -100's GB**	**100's GB – PBs**
Techniques	• 64-bit OS • data storage • parfor • datastore • mapreduce	• parfor • distributed data • spmd	• mapreduce

Techniques for Big Data in MATLAB



Parallel Computing with MATLAB



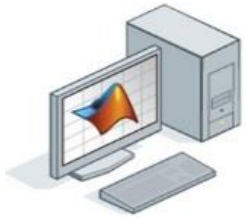
Example: Determining Land Use

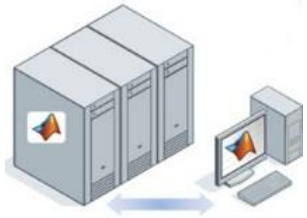
Using Parallel for-loops (`parfor`)

- Data
  - Arial images of agriculture land
  - 24 TIF files
- Analysis
  - Find and measure irrigation fields
  - Determine which irrigation circles are in use (by color)
  - Calculate area under irrigation

When to Use `parfor`

- Data Characteristics
  - Can be of any format (i.e. text, images) as long as it can be broken into pieces
  - The data for each iteration must fit in memory



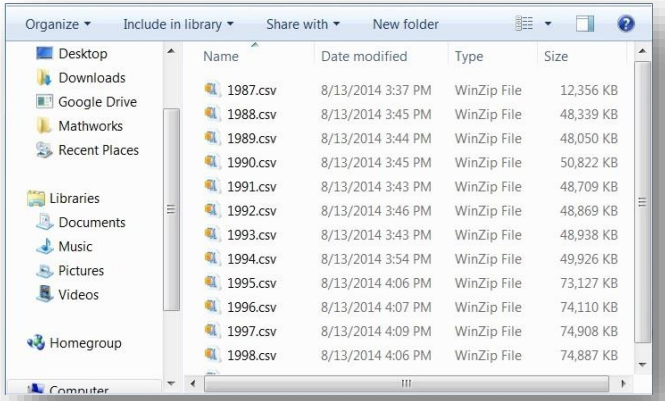


- Compute Platform
  - Desktop (Parallel Computing Toolbox)
  - Cluster (MATLAB Distributed Computing Server)
- Analysis Characteristics
  - Each iteration of your loop must be independent

Access Big Data

`datastore`

- Easily specify data set
  - Single text file (or collection of text files)
- Preview data structure and format
- Select data to import using column names
- Incrementally read subsets of the data




```
>> preview(ds)
ans =
    Year    Month    DayofMonth    DayOfWeek
    ____    _____    __________    _________
    1987    10       21            3
    1987    10       26            1
    1987    10       23            5
    1987    10       23            5
```

```
airdata = datastore('*.csv');
airdata.SelectedVariables = {'Distance', 'ArrDelay'};

data = read(airdata);
```

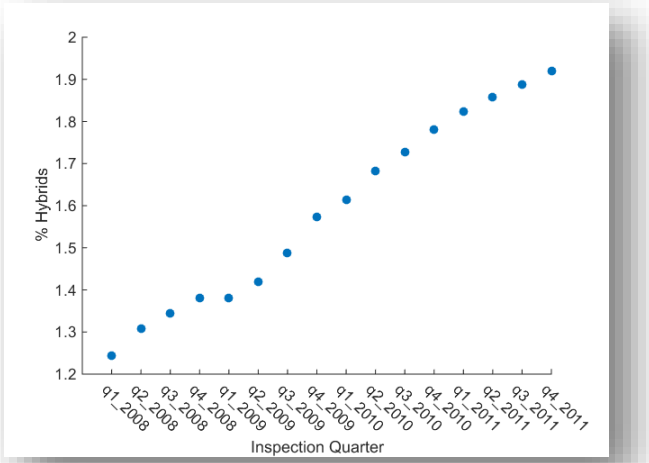
Example: Vehicle Registry Analysis

Using a DataStore

- Data
  - Massachusetts Vehicle Registration Data from 2008-2011
  - 16M records, 45 fields

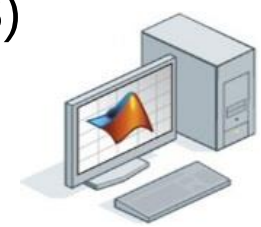
muni_id	veh_zip	insp_year	model_year	make
325	1089	2011	2008	'Hyundai'
325	1089	2009	2008	'Hyundai'
288	1776	2011	2008	'Acura'
288	1776	2008	2008	'Acura'
145	2364	2011	2005	'Chevrolet'
325	1089	2010	2008	'Hyundai'
325	1089	2011	2008	'Hyundai'
288	1776	2009	2008	'Acura'

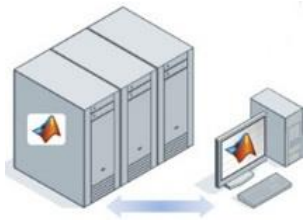
- Analysis
  - Examine hybrid adoptions
  - Calculate % of hybrids registered by quarter
  - Fit growth to predict further adoption



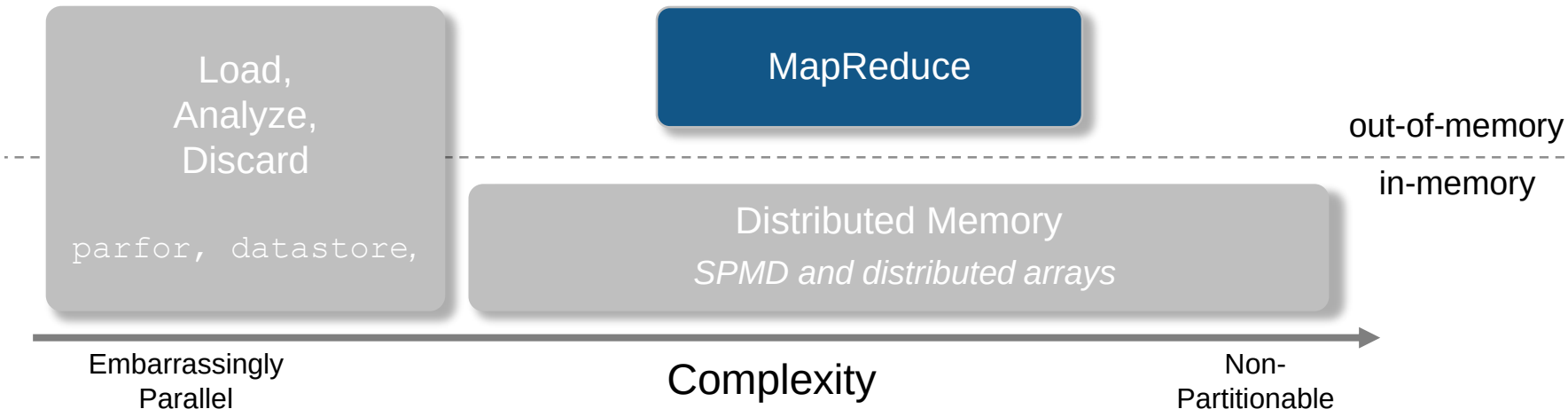
When to Use `datastore`

- Data Characteristics
  - Text data in files, databases or stored in the Hadoop Distributed File System (HDFS)
- Compute Platform
  - Desktop
- Analysis Characteristics
  - Supports Load, Analyze, Discard workflows
  - Incrementally read chunks of data, process within a `while` loop

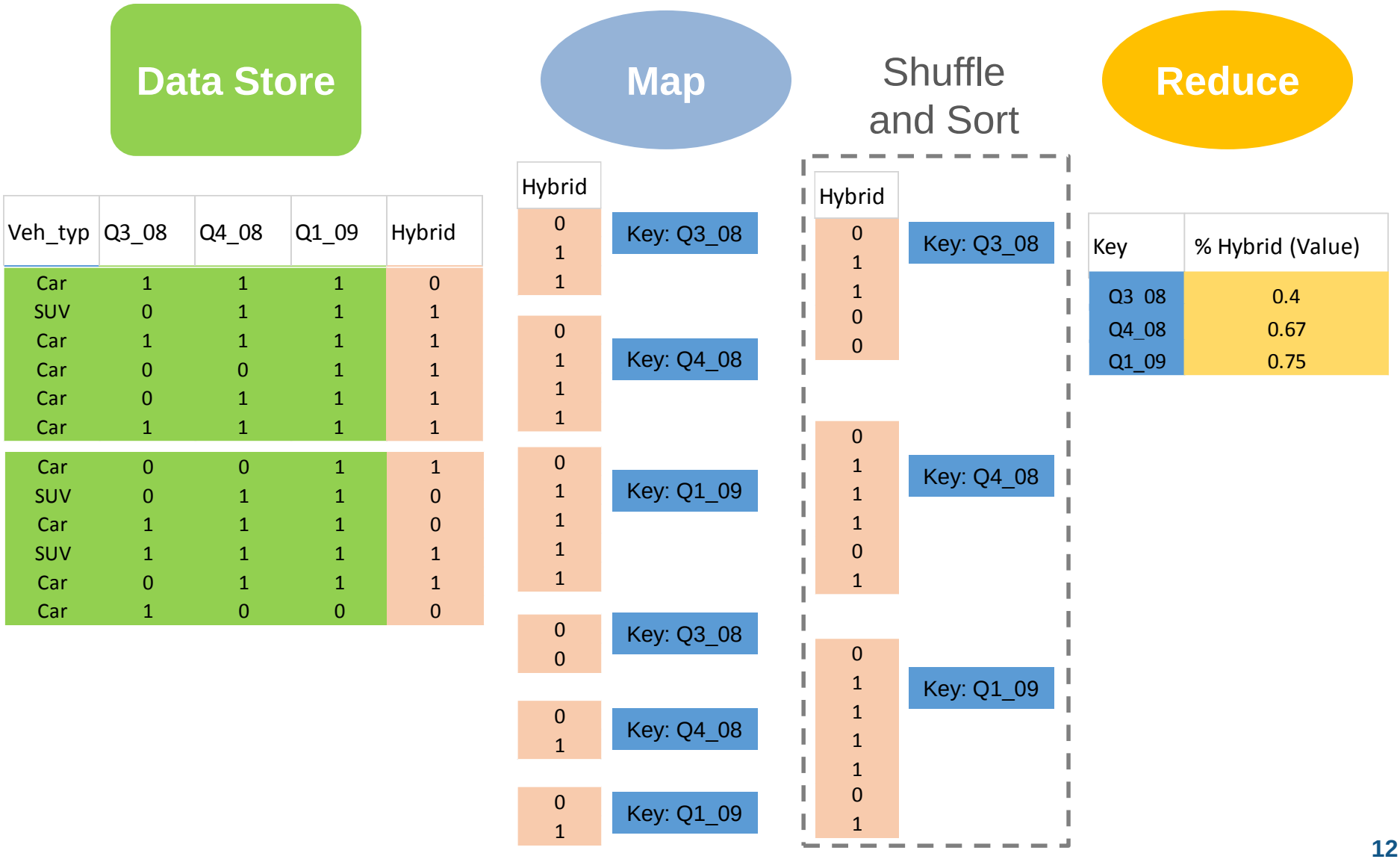




Techniques for Big Data in MATLAB



mapreduce


Data Store
Map
Shuffle and Sort
Reduce
Veh_typ Q3_08 Q4_08 Q1_09 Hybrid
Car 1 1 1 0
SUV 0 1 1 1
Car 1 1 1 1
Car 0 0 1 1
Car 0 1 1 1
Car 1 1 1 1
Car 0 0 1 1
SUV 0 1 1 0
Car 1 1 1 0
SUV 1 1 1 1
Car 0 1 1 1
Car 1 0 0 0
Hybrid
0 1 1
Key: Q3_08
0 1 1 1
Key: Q4_08
0 1 1 1 1
Key: Q1_09
0 0
Key: Q3_08
0 1
Key: Q4_08
0 1
Key: Q1_09
Hybrid
0 1 1 0 0
Key: Q3_08
0 1 1 1 0 1
Key: Q4_08
0 1 1 1 1 1 0 1
Key: Q1_09
Key % Hybrid (Value)
Q3_08 0.4
Q4_08 0.67
Q1_09 0.75

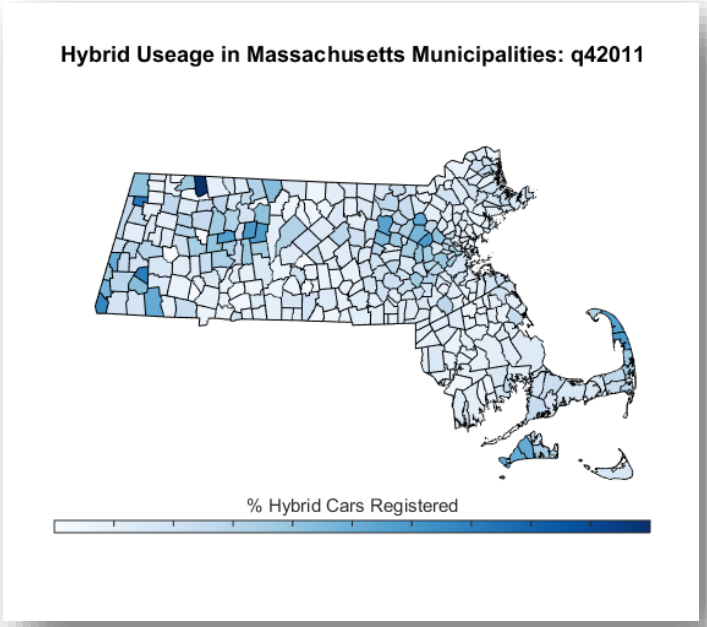
Example: Vehicle Registry Analysis

Using MapReduce

- Data
  - Massachusetts Vehicle Registration Data from 2008-2011
  - 16M records, 45 fields

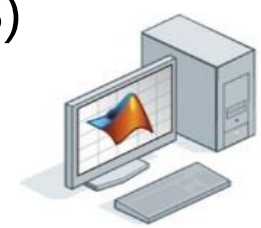
muni_id	veh_zip	insp_year	model_year	make
325	1089	2011	2008	'Hyundai'
325	1089	2009	2008	'Hyundai'
288	1776	2011	2008	'Acura'
288	1776	2008	2008	'Acura'
145	2364	2011	2005	'Chevrolet'
325	1089	2010	2008	'Hyundai'
325	1089	2011	2008	'Hyundai'
288	1776	2009	2008	'Acura'

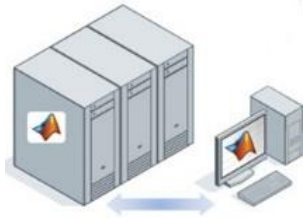
- Analysis
  - Examine hybrid adoptions
  - Calculate % of hybrids registered
    - By Quarter
    - By Regional Area
  - Create map of results



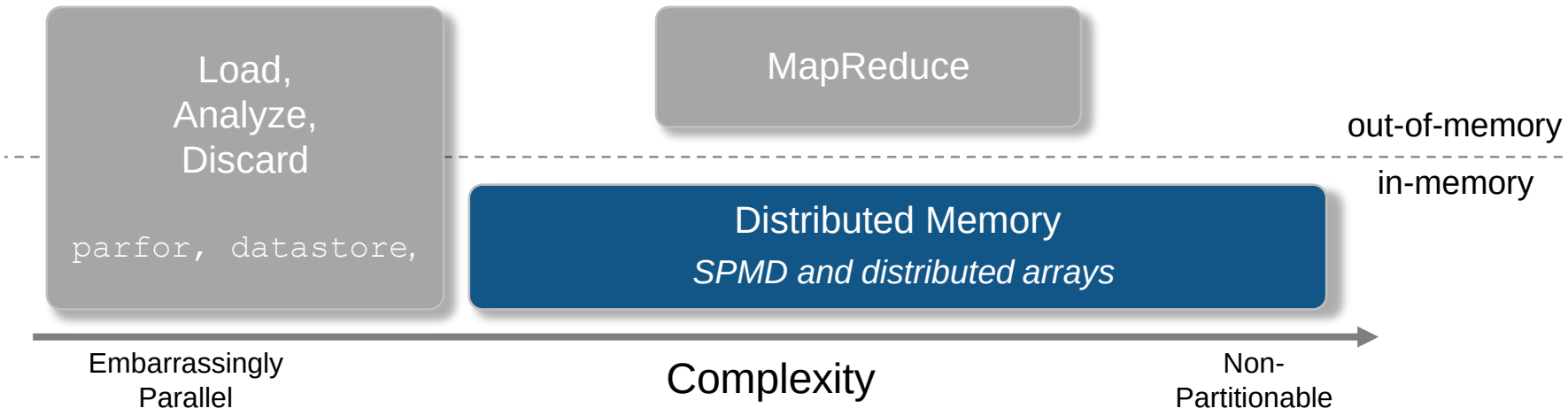
When to Use `mapreduce`

- Data Characteristics
  - Text data in files, databases or stored in the Hadoop Distributed File System (HDFS)
  - Dataset will not fit into memory
- Compute Platform
  - Desktop
  - Scales to run within Hadoop MapReduce on data in HDFS
- Analysis Characteristics
  - Must be able to be Partitioned into two phases
    1. Map: filter or process sub-segments of data
    2. Reduce: aggregate interim results and calculate final answer





Techniques for Big Data in MATLAB



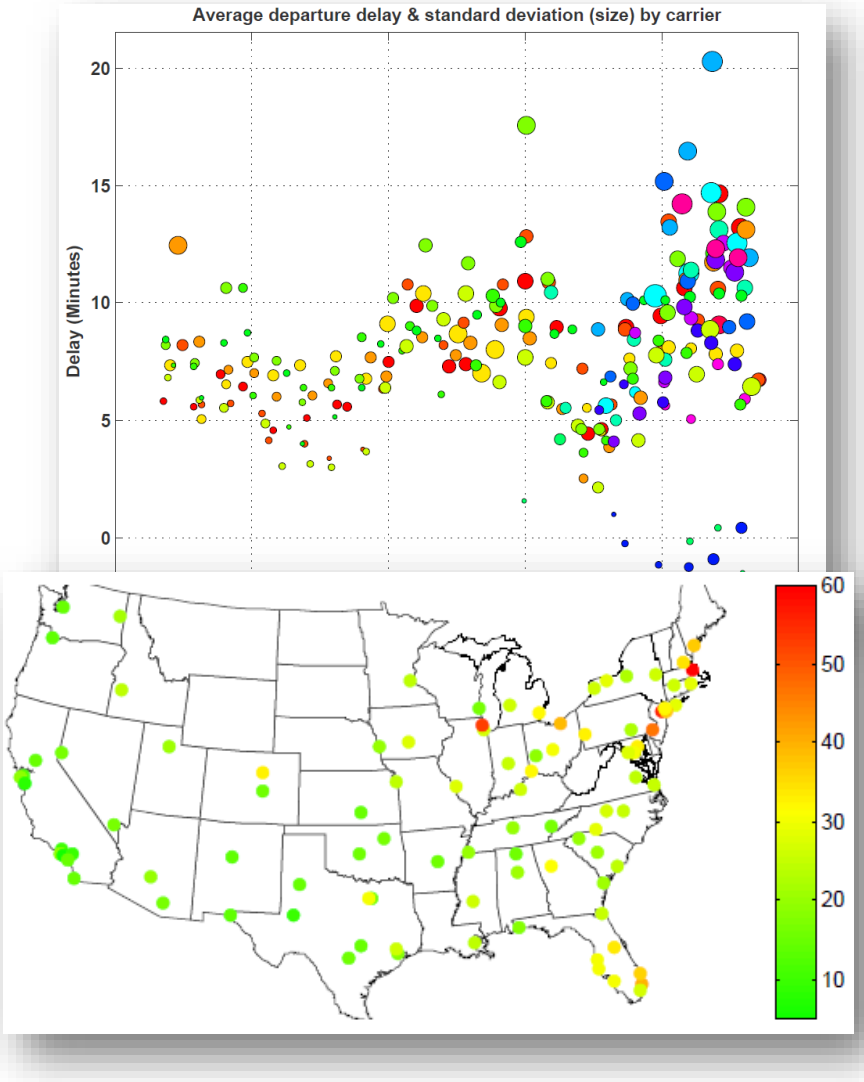
`spmd` blocks

```
spmd
    % single program across workers
end
```

- Mix parallel and serial code in the same function
- Run on a pool of MATLAB resources
- **S**ingle **P**rogram runs simultaneously across workers
- **M**ultiple **D**ata spread across multiple workers

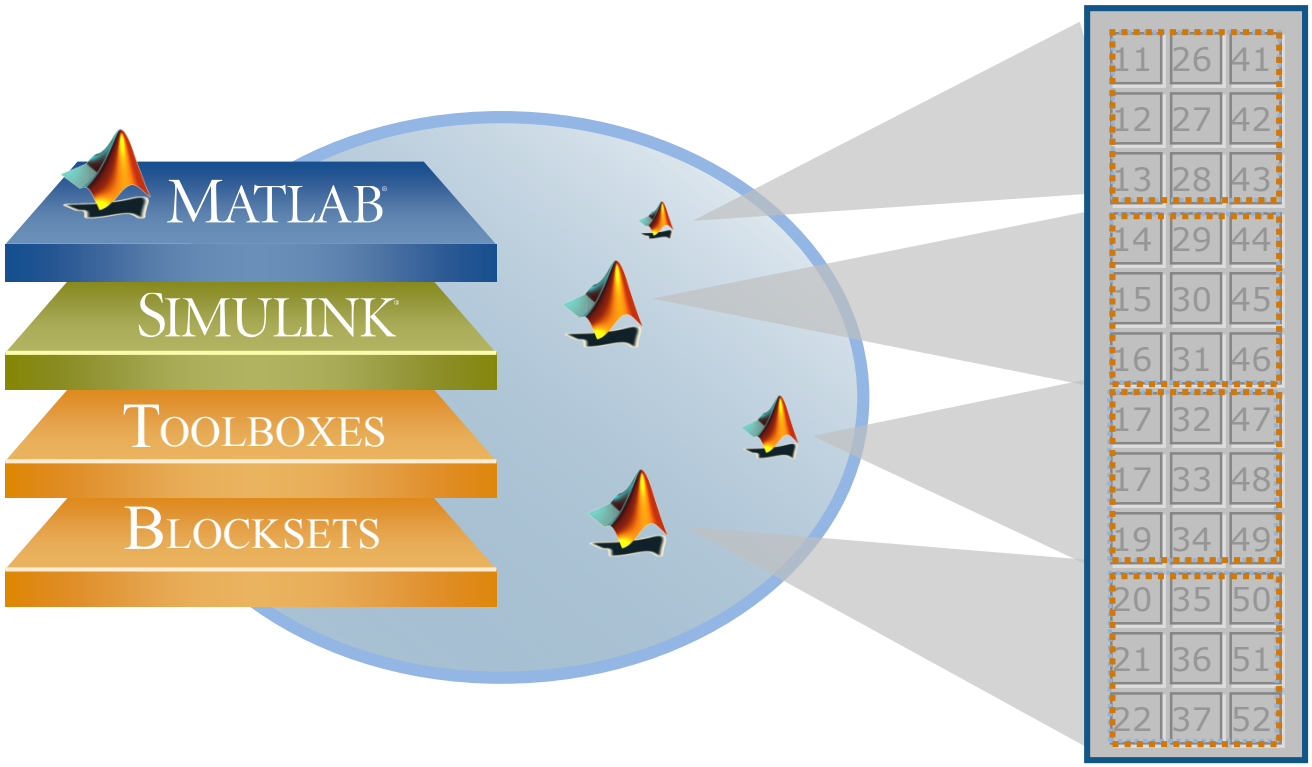
Example: Airline Delay Analysis

- Data
  - BTS/RITA Airline On-Time Statistics
  - 123.5M records, 29 fields
- Analysis
  - Calculate delay patterns
  - Visualize summaries
  - Estimate & evaluate predictive models



Distributed Arrays

Available from
- **Parallel Computing Toolbox**
- **MATLAB Distributed Computing Server**

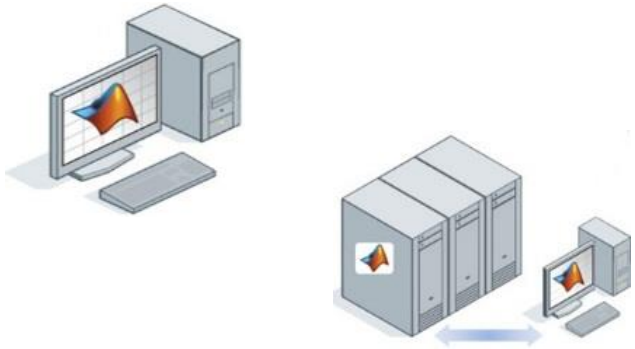


Remotely Manipulate Array from Desktop

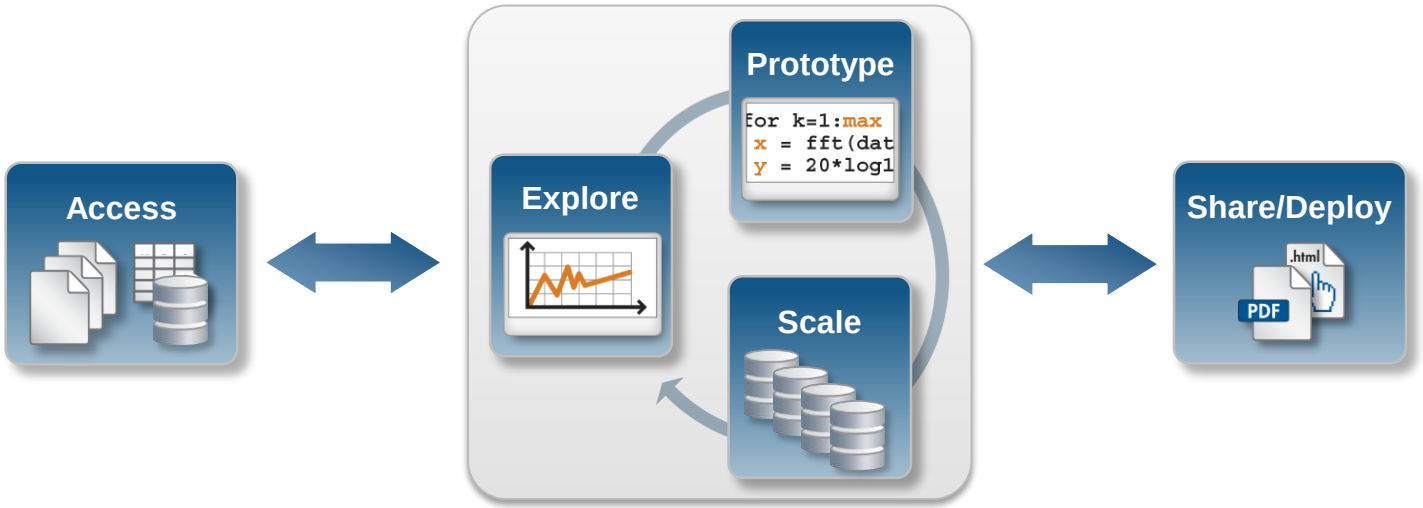
Distributed Array Lives on the Cluster

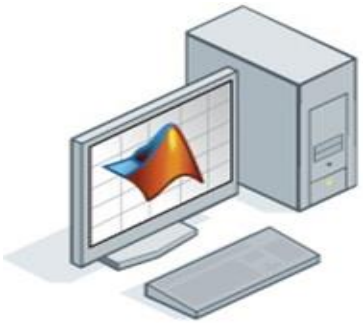
When to Use Distributed Memory

- Data Characteristics
  - Data must be fit in collective memory across machines
- Compute Platform
  - Prototype (subset of data) on desktop
  - Run on a cluster or cloud
- Analysis Characteristics
  - Consists of:
    - Parts that can be run on data in memory (`spmd`)
    - Supported functions for distributed arrays

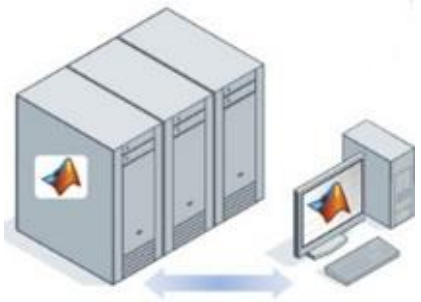


Big Data Analysis with MATLAB



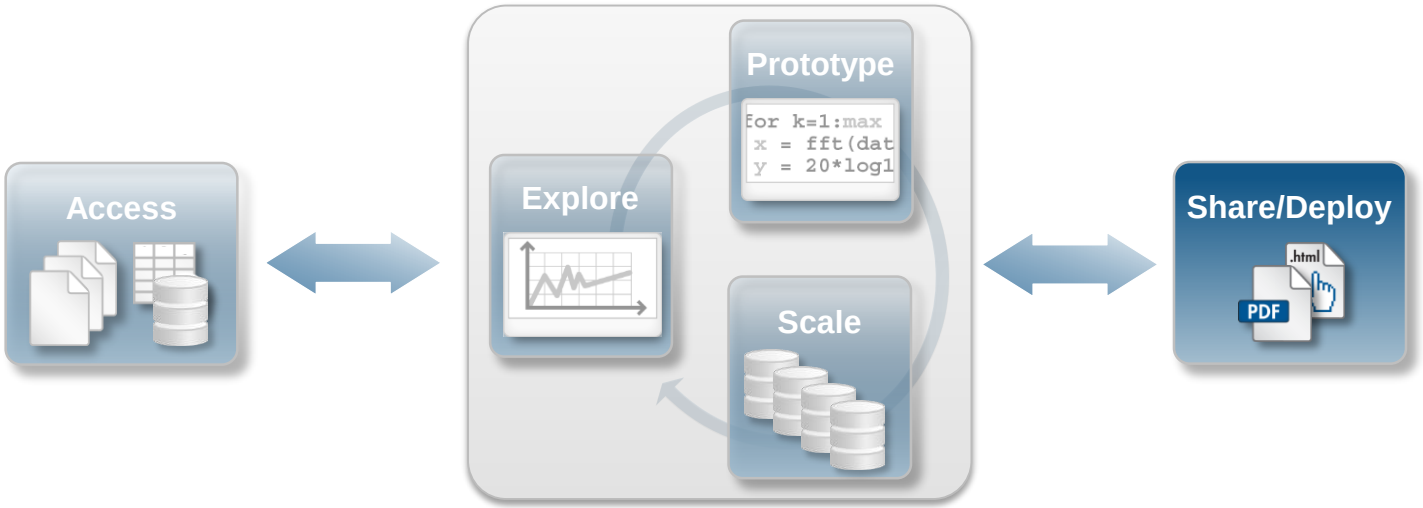


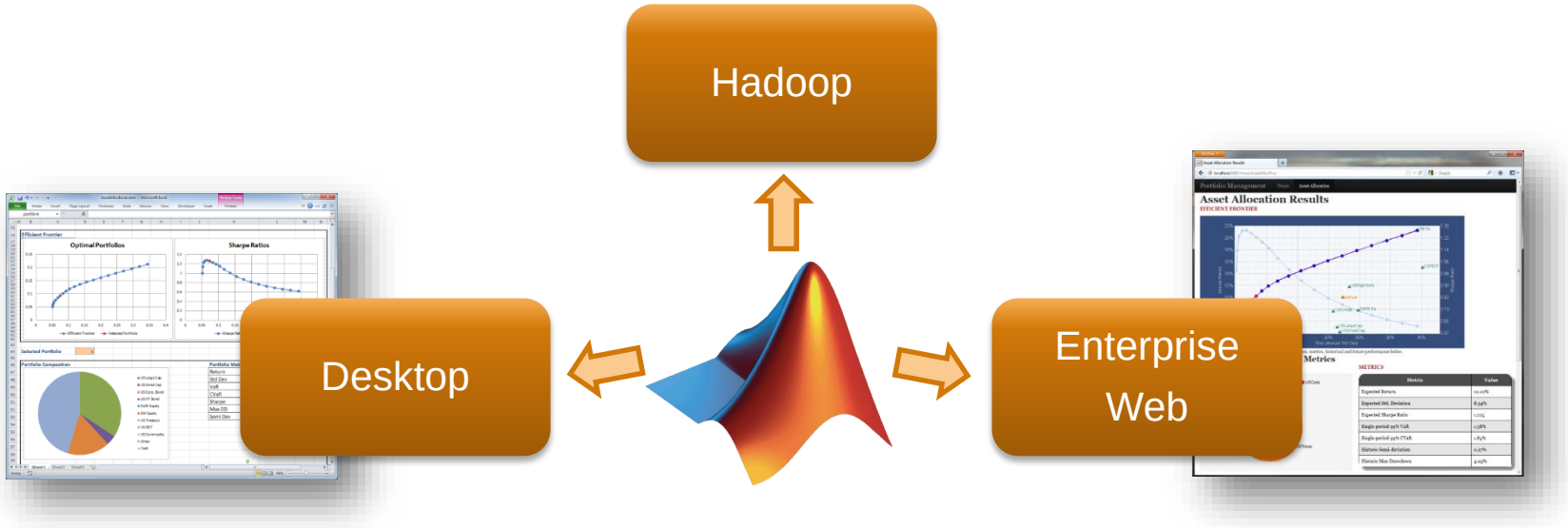
Work on the desktop



Scale capacity as needed

Deploy Big Data Algorithms

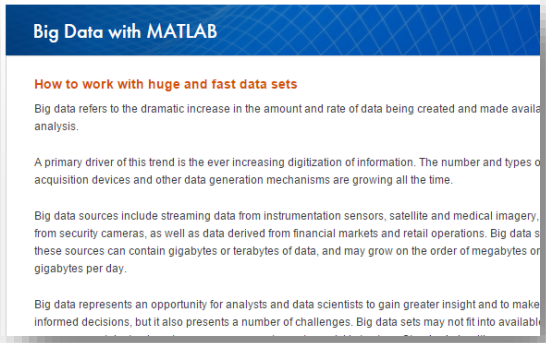




Learn More

- MATLAB Documentation
  - Strategies for Efficient Use of Memory
  - Resolving "Out of Memory" Errors

- Big Data with MATLAB
  - www.mathworks.com/discovery/big-data-matlab.html



- MATLAB MapReduce and Hadoop
  - www.mathworks.com/discovery/matlab-mapreduce-hadoop.html