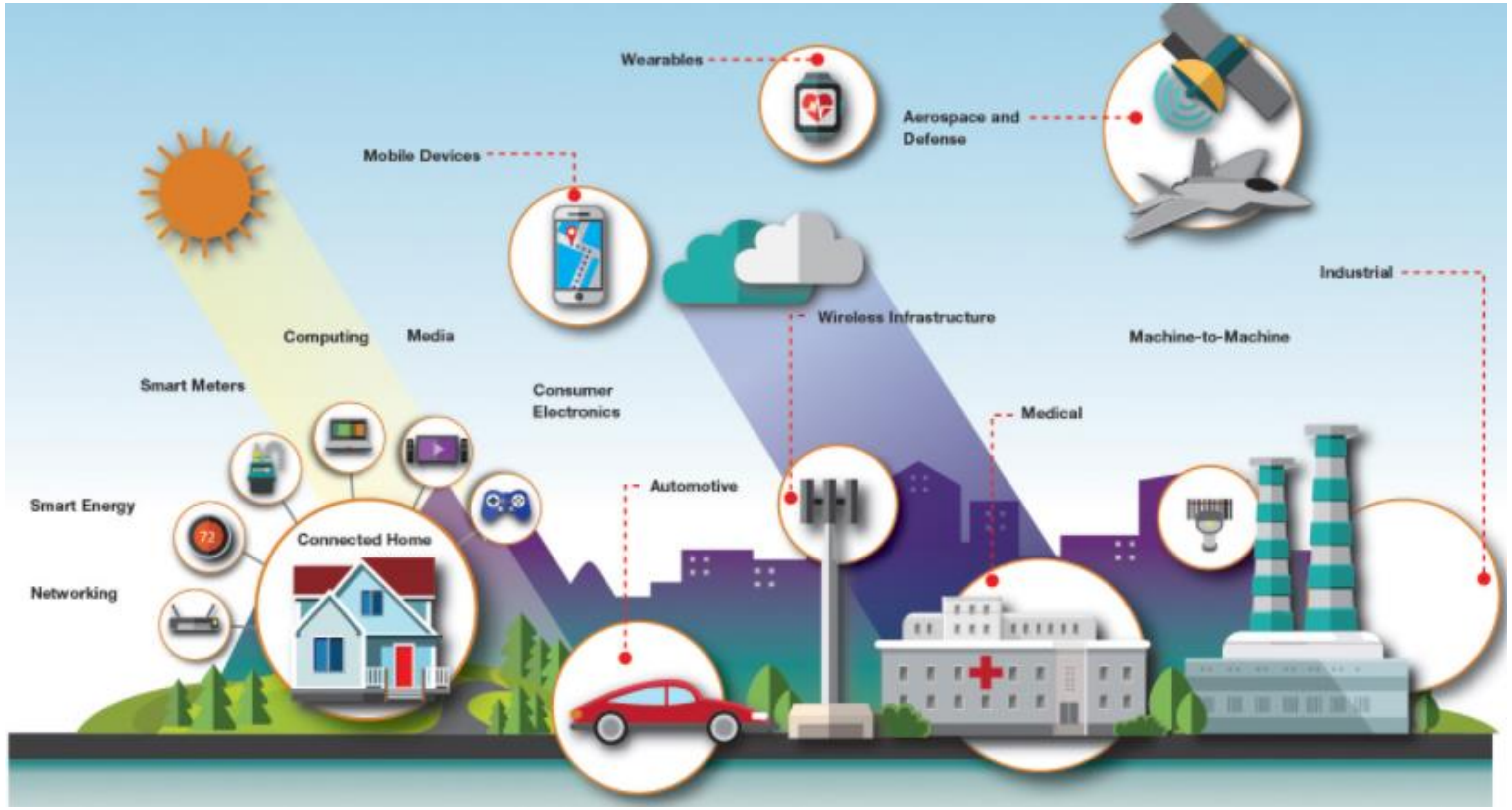
Analyzing Fleet Data with MATLAB and Spark

Christoph Stockhammer

MathWorks
AUTOMOTIVE CONFERENCE 2018

What does "Fleet" mean?

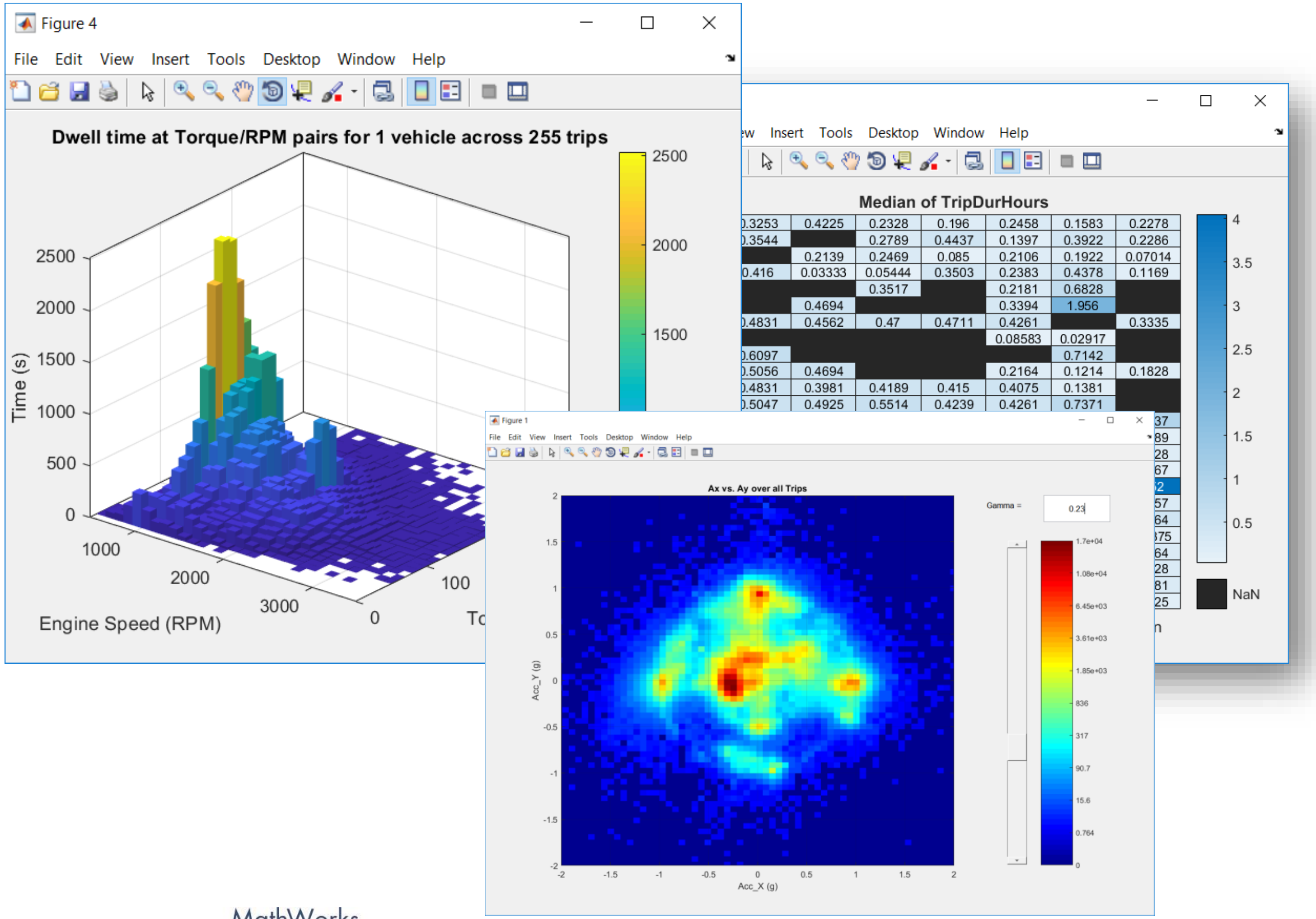
- A "Fleet" is any group of things that can generate data and that you would like to look at all together. Examples include:



Automotive Fleet Data

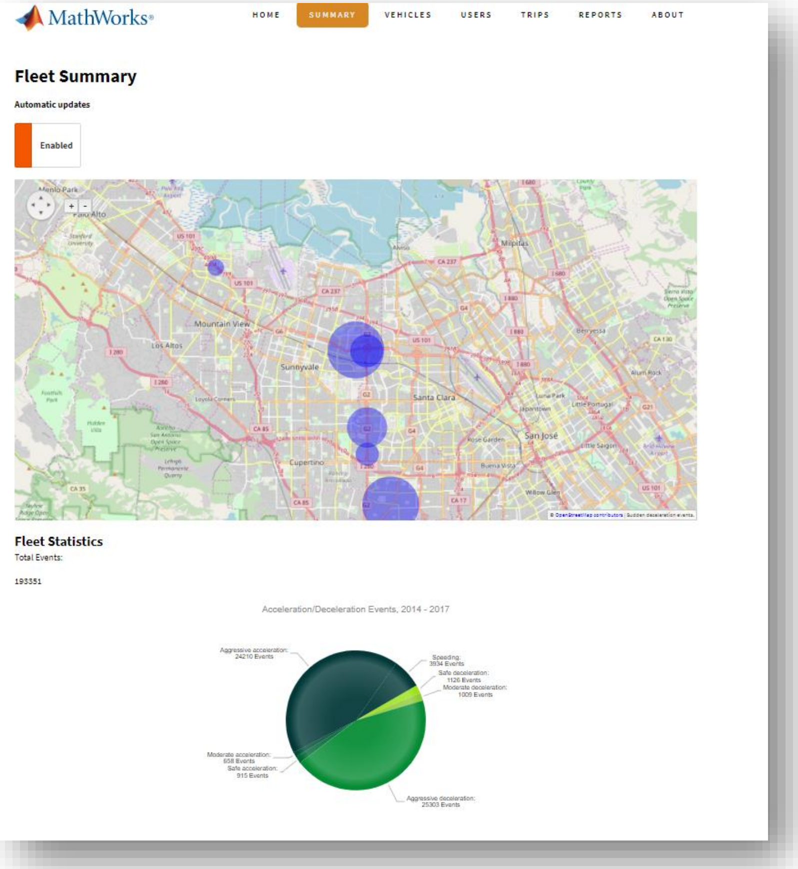
What is the fleet data telling us?

How’s my driving?



Median of TripDurHours

.3253	0.4225	0.2328	0.196	0.2458	0.1583	0.2278
.3544		0.2789	0.4437	0.1397	0.3922	0.2286
	0.2139	0.2469	0.085	0.2106	0.1922	0.07014
0.416	0.03333	0.05444	0.3503	0.2383	0.4378	0.1169
		0.3517		0.2181	0.6828	
	0.4694			0.3394	1.956	
.4831	0.4562	0.47	0.4711	0.4261		0.3335
				0.08583	0.02917	
.6097					0.7142	
.5056	0.4694			0.2164	0.1214	0.1828
.4831	0.3981	0.4189	0.415	0.4075	0.1381	
.5047	0.4925	0.5514	0.4239	0.4261	0.7371	
						37
						89
						28
						67
						2
						57
						64
						375
						64
						28
						81
						25


MathWorks
HOME
SUMMARY
VEHICLES
USERS
TRIPS
REPORTS
ABOUT
Menlo Park
Palo Alto
Mountain View
Sunnyvale
Santa Clara
San Jose
Cupertino
Milpitas
© OpenStreetMap contributors | Sudden deceleration events.
Acceleration/Deceleration Events, 2014 - 2017
Aggressive acceleration: 24210 Events
Speeding: 3934 Events
Safe deceleration: 1126 Events
Moderate deceleration: 1009 Events
Moderate acceleration: 658 Events
Safe acceleration: 915 Events
Aggressive deceleration: 25303 Events

Fleet Summary

Automatic updates

Enabled

Fleet Statistics

Total Events:

193351

How do Customers Apply Analytics to Fleet Data?

Vehicle data, driver profiles

Spark

kafka

Historic data:

- **Batch processing**
- Large data on cluster
- Explore long term trends
- Build model

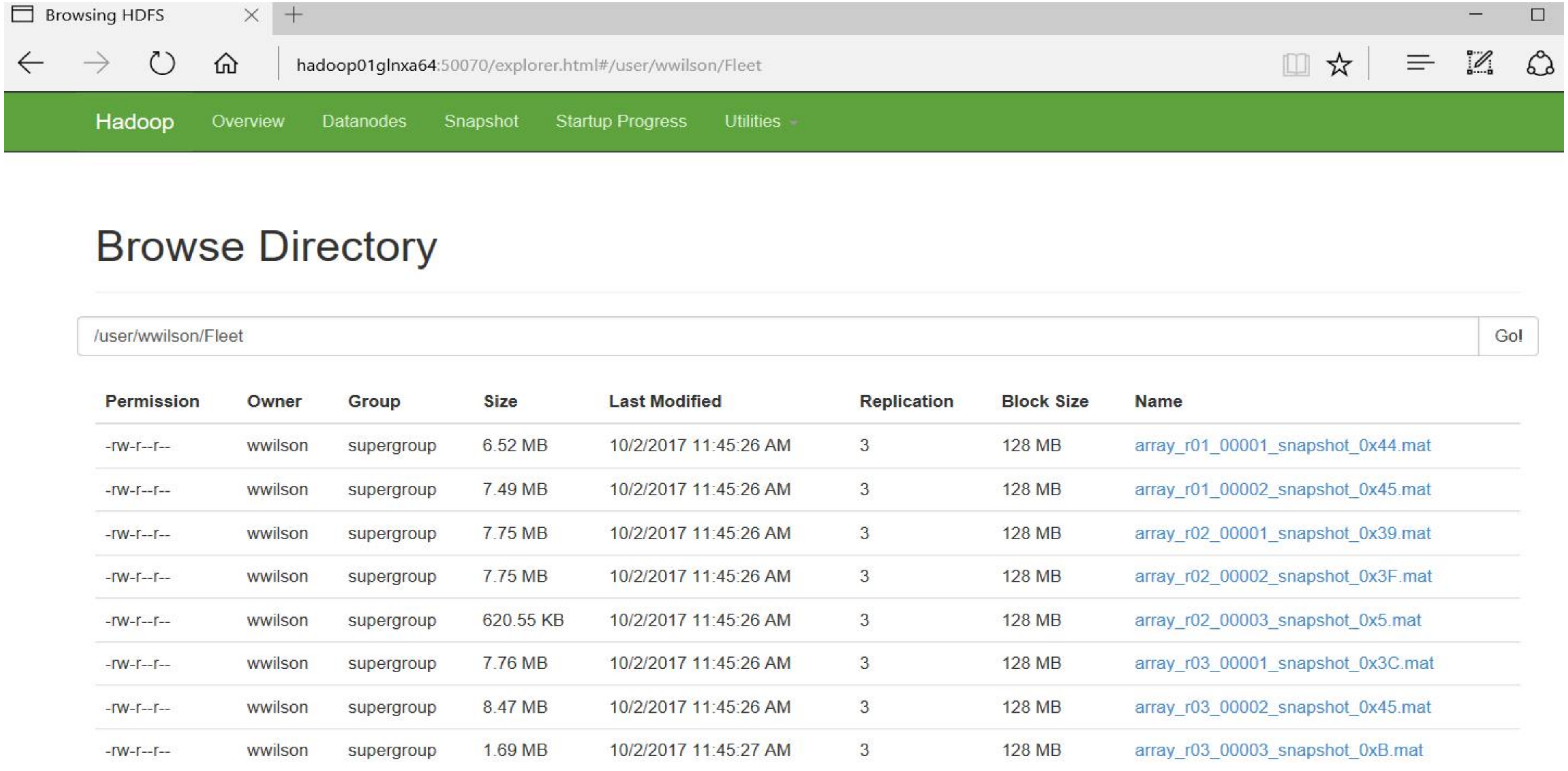
Streaming data:

- **Near real-time**
- Test and implement model for new data
- Stream processing

Cold Storage

Hot Storage

Analytics Running on Hadoop and Spark (Video)



Browse Directory

/user/wwilson/Fleet

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	wwilson	supergroup	6.52 MB	10/2/2017 11:45:26 AM	3	128 MB	array_r01_00001_snapshot_0x44.mat
-rw-r--r--	wwilson	supergroup	7.49 MB	10/2/2017 11:45:26 AM	3	128 MB	array_r01_00002_snapshot_0x45.mat
-rw-r--r--	wwilson	supergroup	7.75 MB	10/2/2017 11:45:26 AM	3	128 MB	array_r02_00001_snapshot_0x39.mat
-rw-r--r--	wwilson	supergroup	7.75 MB	10/2/2017 11:45:26 AM	3	128 MB	array_r02_00002_snapshot_0x3F.mat
-rw-r--r--	wwilson	supergroup	620.55 KB	10/2/2017 11:45:26 AM	3	128 MB	array_r02_00003_snapshot_0x5.mat
-rw-r--r--	wwilson	supergroup	7.76 MB	10/2/2017 11:45:26 AM	3	128 MB	array_r03_00001_snapshot_0x3C.mat
-rw-r--r--	wwilson	supergroup	8.47 MB	10/2/2017 11:45:26 AM	3	128 MB	array_r03_00002_snapshot_0x45.mat
-rw-r--r--	wwilson	supergroup	1.69 MB	10/2/2017 11:45:27 AM	3	128 MB	array_r03_00003_snapshot_0xB.mat

Major Logistical Barriers to Working with Automotive Fleet Data

- Here is a Hard Drive full of vehicle log data (GB or TB), now what? Oh and by the way, there are a bunch more of these coming soon…
  - Pains
    - Large, non-text data
    - Lack of clarity – “what should I do with this”?
    - Time pressure to get the analysis done

- *My data is in Hadoop, now what?* or *My data is* ***supposed*** *to go into Hadoop, now what?*
  - Pains
    - Giant binary files – not well suited to “drop into” HDFS
    - Fear of loosing control of one’s data – hand data off to “another group”…
    - Fear of a “new system” – Hadoop can be scary, Linux, yikes!

Have you Ever Wondered…?

- How different factors affect how a particular driver drives?
- Real-world vehicle performance of things like: fuel economy, emissions, vehicle dynamics, ride and handling, prognostics, and durability?
- How do you work with terabytes of data to distill out critical information?
- Once you do have the critical information, how to you iterate back through your terabytes of data to extract relevant (time) slices for further study or analysis?

So, what’s the (big) problem?

- Traditional tools and approaches won’t work
  - Accessing 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
  - Learning new tools and rewriting algorithms can hurt productivity
- Time required to conduct analysis
  - Need to leverage **parallel computing** on desktop and cluster

MathWorks Vehicle Fleet – Case Study

Challenge

- Develop and deploy Data Analytics to run on Spark against (non-text format) vehicle fleet data stored on Hadoop

Solution

- Use MATLAB `tall` arrays to develop analytics on the desktop and then scale out to the Hadoop cluster

Results

- Developed insight and understanding of over 1300 vehicle trips
- Illustrated fuel efficiency performance under real-world driving conditions

Volkswagen Data Lab develops driver recognition algorithms with MATLAB

Develop technology building block for tailoring car features and services to individual

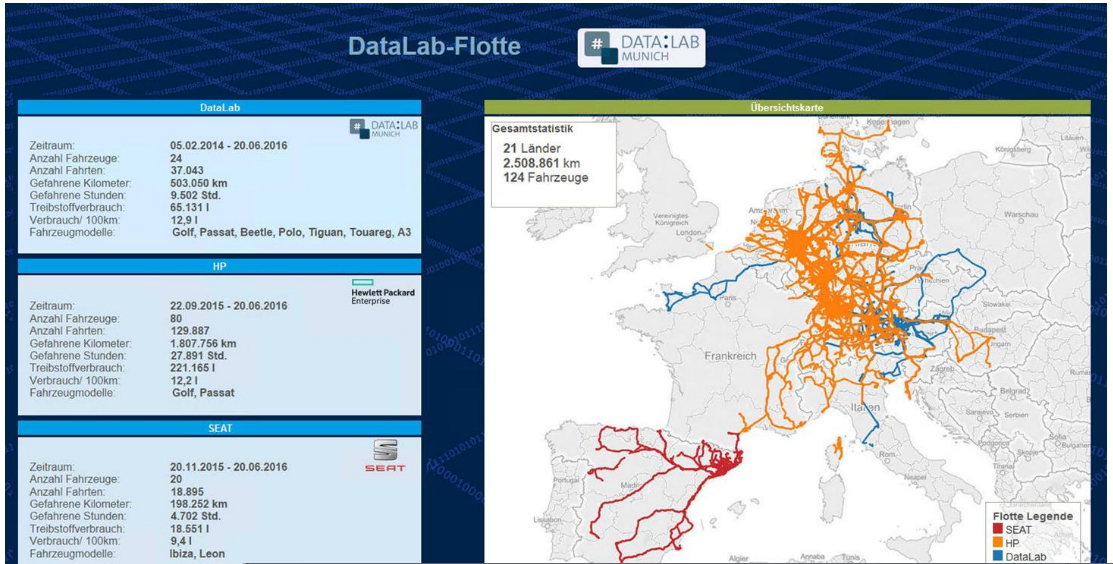
- Need to identify individual drivers based on their driving behavior using collected data

Challenges

- Accuracy despite low training data
- Robustness despite environmental conditions
- Computing time

Data sources

- Logged CAN bus data and travel record



DataLab-Flotte

DATA:LAB MUNICH

DataLab

Zeitraum:	05.02.2014 - 20.06.2016
Anzahl Fahrzeuge:	24
Anzahl Fahrten:	37.043
Gefahrene Kilometer:	503.050 km
Gefahrene Stunden:	9.502 Std.
Treibstoffverbrauch:	65.131 l
Verbrauch/ 100km:	12,9 l
Fahrzeugmodelle:	Golf, Passat, Beetle, Polo, Tiguan, Touareg, A3

HP

Hewlett Packard Enterprise

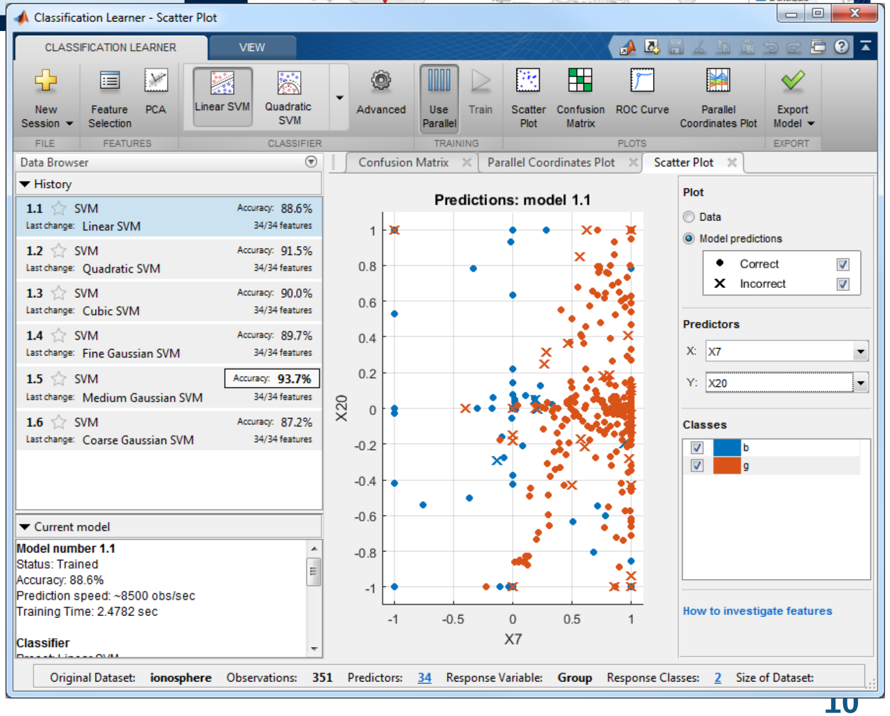
Zeitraum:	22.09.2015 - 20.06.2016
Anzahl Fahrzeuge:	80
Anzahl Fahrten:	129.887
Gefahrene Kilometer:	1.807.756 km
Gefahrene Stunden:	27.891 Std.
Treibstoffverbrauch:	221.165 l
Verbrauch/ 100km:	12,2 l
Fahrzeugmodelle:	Golf, Passat

SEAT

Zeitraum:	20.11.2015 - 20.06.2016
Anzahl Fahrzeuge:	20
Anzahl Fahrten:	18.895
Gefahrene Kilometer:	198.252 km
Gefahrene Stunden:	4.702 Std.
Treibstoffverbrauch:	18.551 l
Verbrauch/ 100km:	9,4 l
Fahrzeugmodelle:	Ibiza, Leon

Übersichtskarte

Gesamtstatistik
21 Länder
2.508.861 km
124 Fahrzeuge



Classification Learner - Scatter Plot

History

1.1	SVM	Accuracy: 88.6%
	Last change: Linear SVM	34/34 features
1.2	SVM	Accuracy: 91.5%
	Last change: Quadratic SVM	34/34 features
1.3	SVM	Accuracy: 90.0%
	Last change: Cubic SVM	34/34 features
1.4	SVM	Accuracy: 89.7%
	Last change: Fine Gaussian SVM	34/34 features
1.5	SVM	Accuracy: 93.7%
	Last change: Medium Gaussian SVM	34/34 features
1.6	SVM	Accuracy: 87.2%
	Last change: Coarse Gaussian SVM	34/34 features

Current model

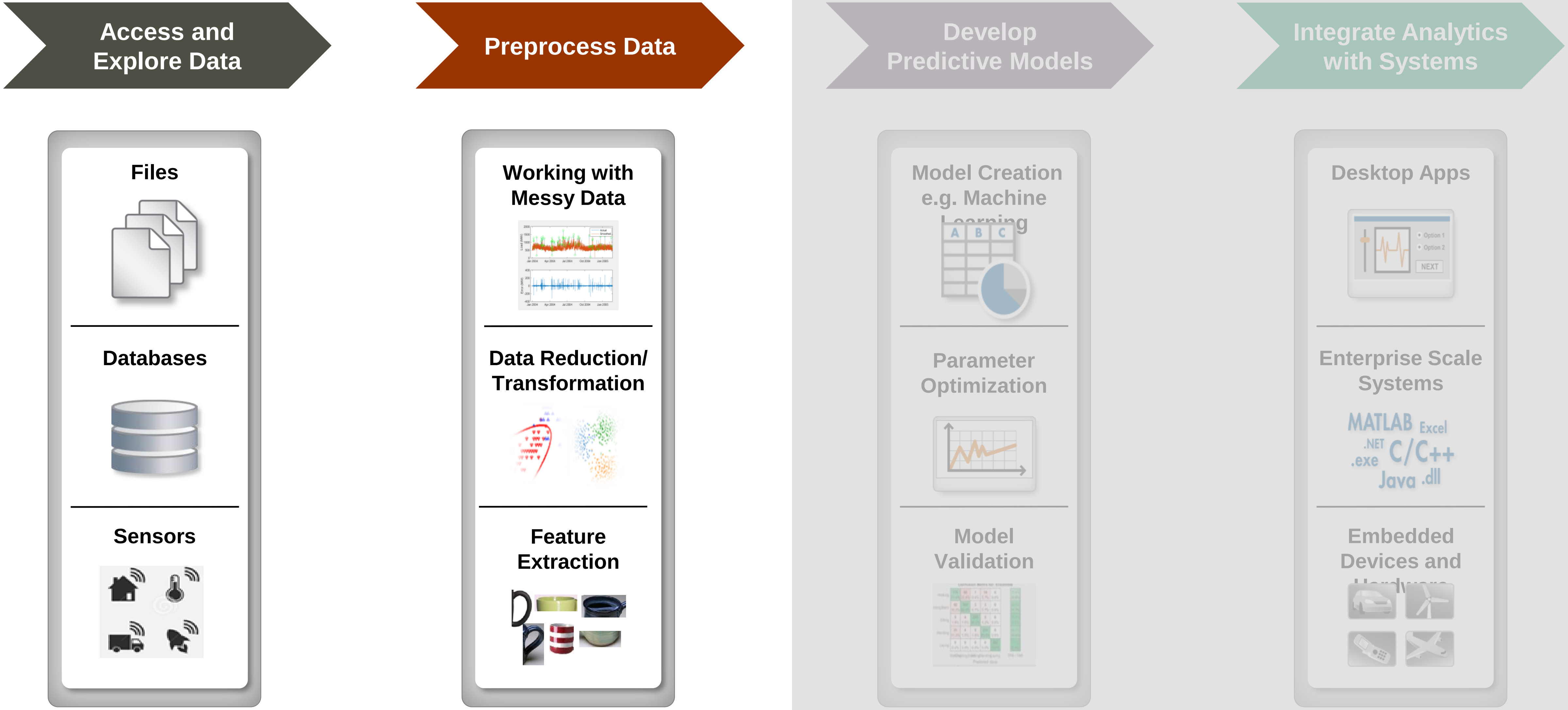
Model number 1.1
Status: Trained
Accuracy: 88.6%
Prediction speed: ~8500 obs/sec
Training Time: 2.4782 sec

Classifier

Original Dataset: **ionosphere** Observations: **351** Predictors: 34 Response Variable: **Group** Response Classes: 2 Size of Dataset:

Source: „Connected Car – Fahrererkennung mit MATLAB“
Julia Fumbarev, Volkswagen Data Lab
MATLAB EXPO Germany, June 27, 2017, Munich Germany

Data Analytics 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

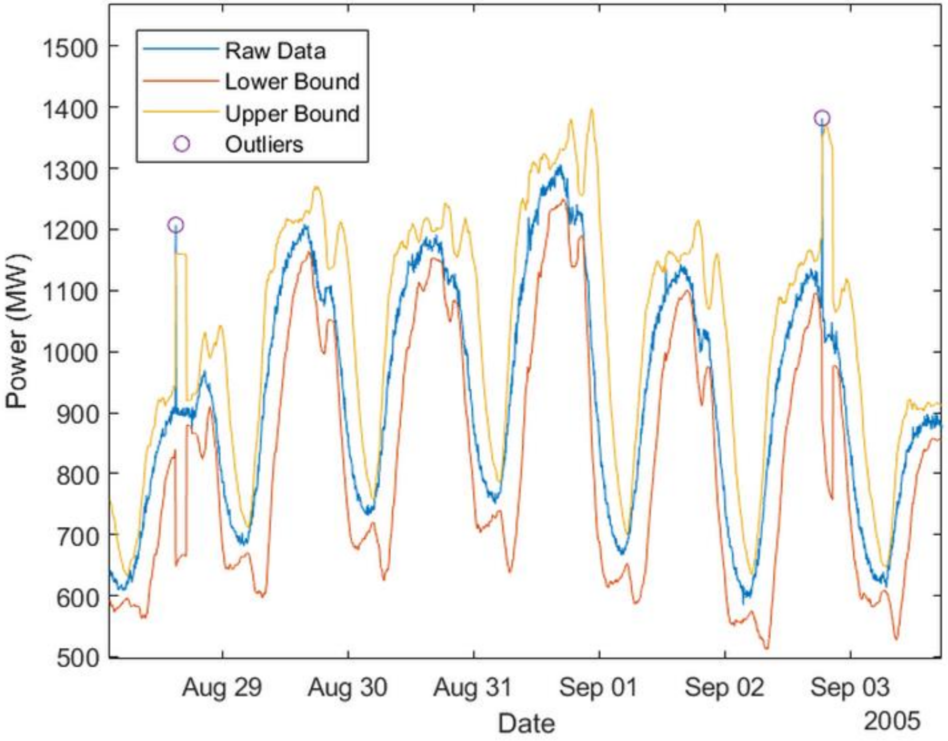
Desktop Apps

Enterprise Scale Systems

Embedded Devices and Hardware

What about messy data?

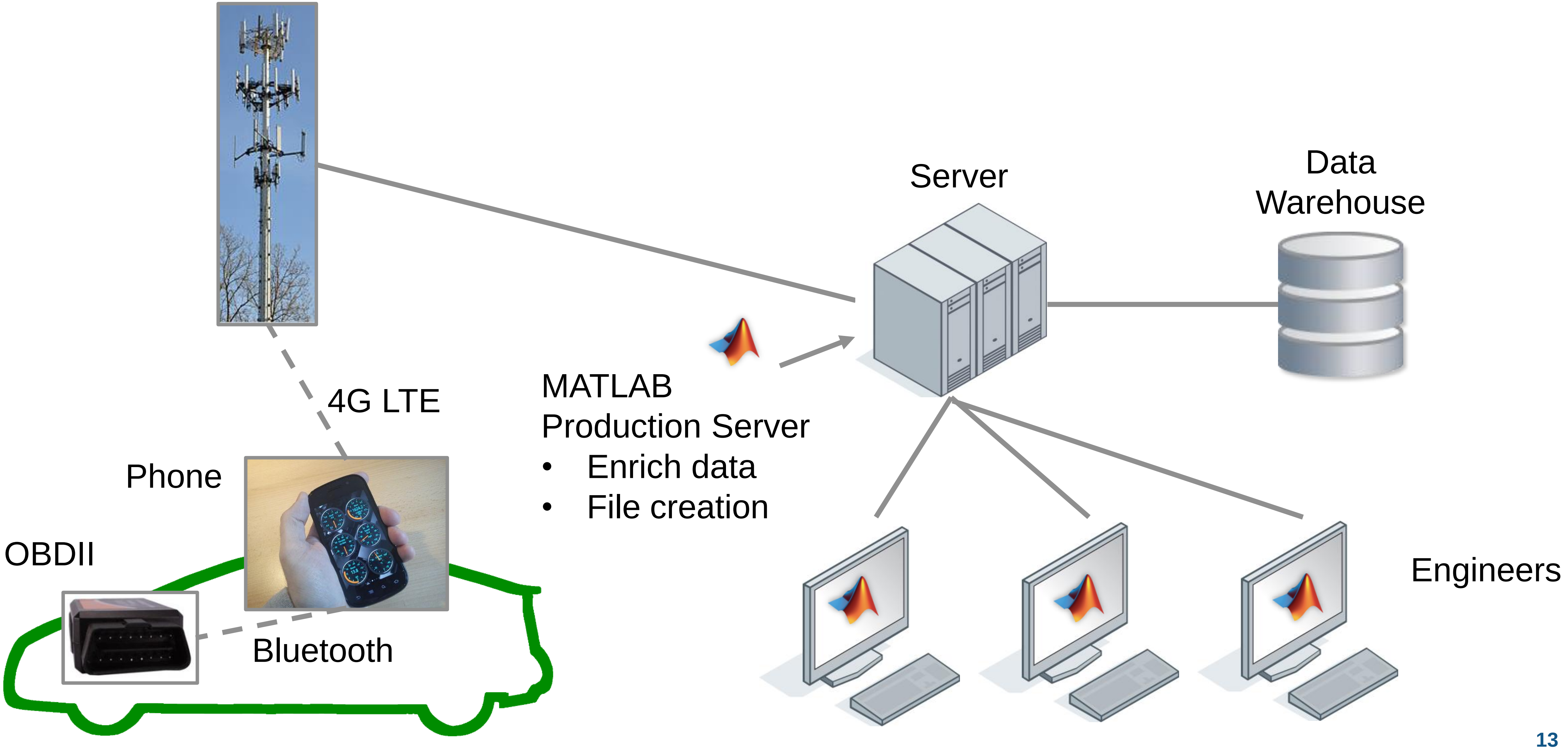
- How do deal with outliers?



- New functions to help you with:
  - Missing Data and Outliers
  - Detecting Change Points
  - Smoothing and Detrending
  - Normalizing and Scaling
  - Grouping and Binning

Full Details: *https://www.mathworks.com/help/matlab/preprocessing-data.html*

MathWorks Automotive Fleet – Data Collection



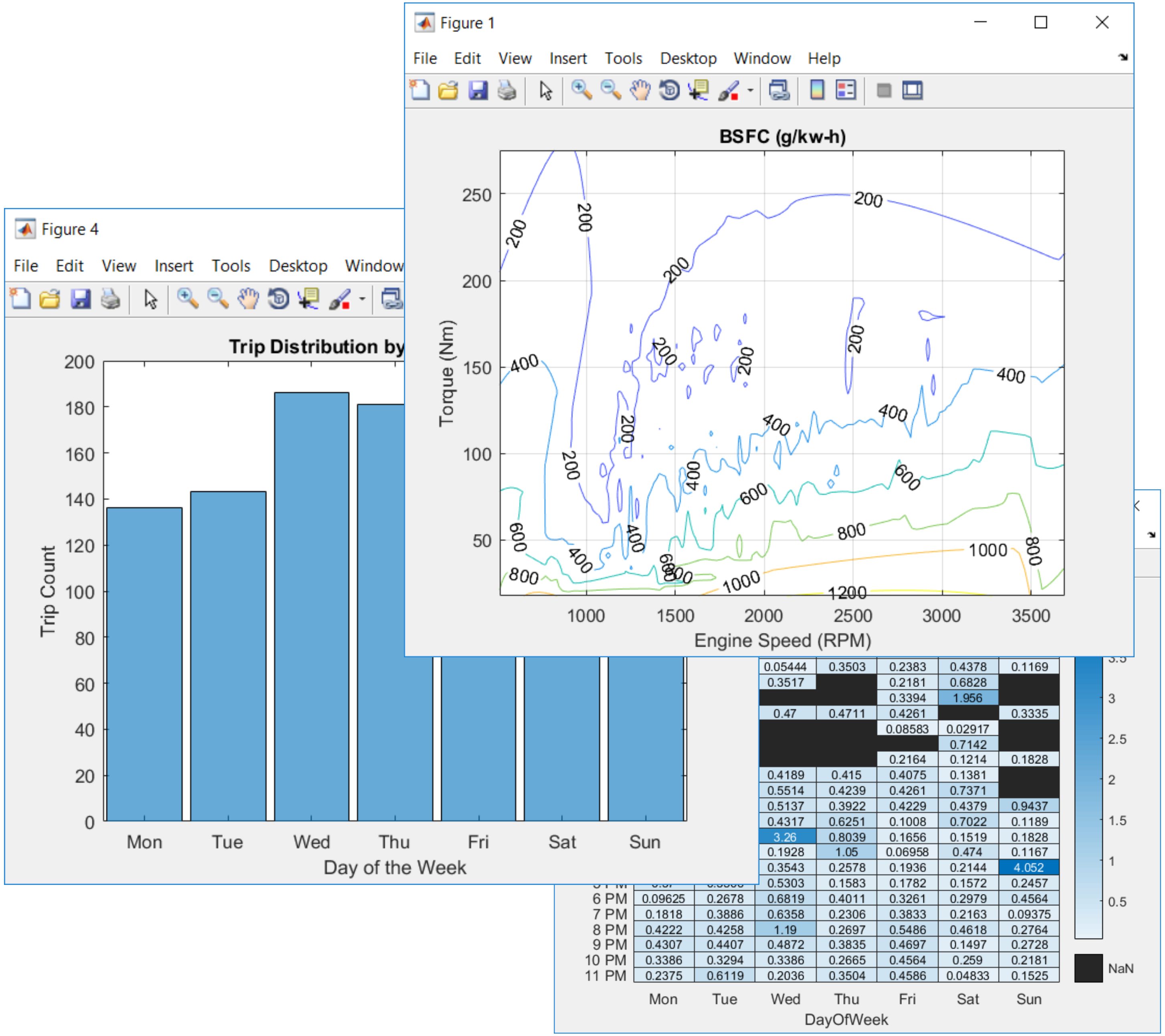
4G LTE

MATLAB Production Server

- Enrich data
- File creation

The MathWorks Fleet

- 1300 trip log files
- 21 unique vehicles
- Approx 39 unique channels
- Data collected over 1.5 years



	Mon	Tue	Wed	Thu	Fri	Sat	Sun
			0.05444	0.3503	0.2383	0.4378	0.1169
			0.3517		0.2181	0.6828	
					0.3394	1.956	
			0.47	0.4711	0.4261		0.3335
					0.08583	0.02917	
						0.7142	
					0.2164	0.1214	0.1828
			0.4189	0.415	0.4075	0.1381	
			0.5514	0.4239	0.4261	0.7371	
			0.5137	0.3922	0.4229	0.4379	0.9437
			0.4317	0.6251	0.1008	0.7022	0.1189
			3.26	0.8039	0.1656	0.1519	0.1828
			0.1928	1.05	0.06958	0.474	0.1167
			0.3543	0.2578	0.1936	0.2144	4.052
			0.5303	0.1583	0.1782	0.1572	0.2457
6 PM	0.09625	0.2678	0.6819	0.4011	0.3261	0.2979	0.4564
7 PM	0.1818	0.3886	0.6358	0.2306	0.3833	0.2163	0.09375
8 PM	0.4222	0.4258	1.19	0.2697	0.5486	0.4618	0.2764
9 PM	0.4307	0.4407	0.4872	0.3835	0.4697	0.1497	0.2728
10 PM	0.3386	0.3294	0.3386	0.2665	0.4564	0.259	0.2181
11 PM	0.2375	0.6119	0.2036	0.3504	0.4586	0.04833	0.1525

Automotive Vehicle Test Fleets – Lots of Data and Lots of Complexity

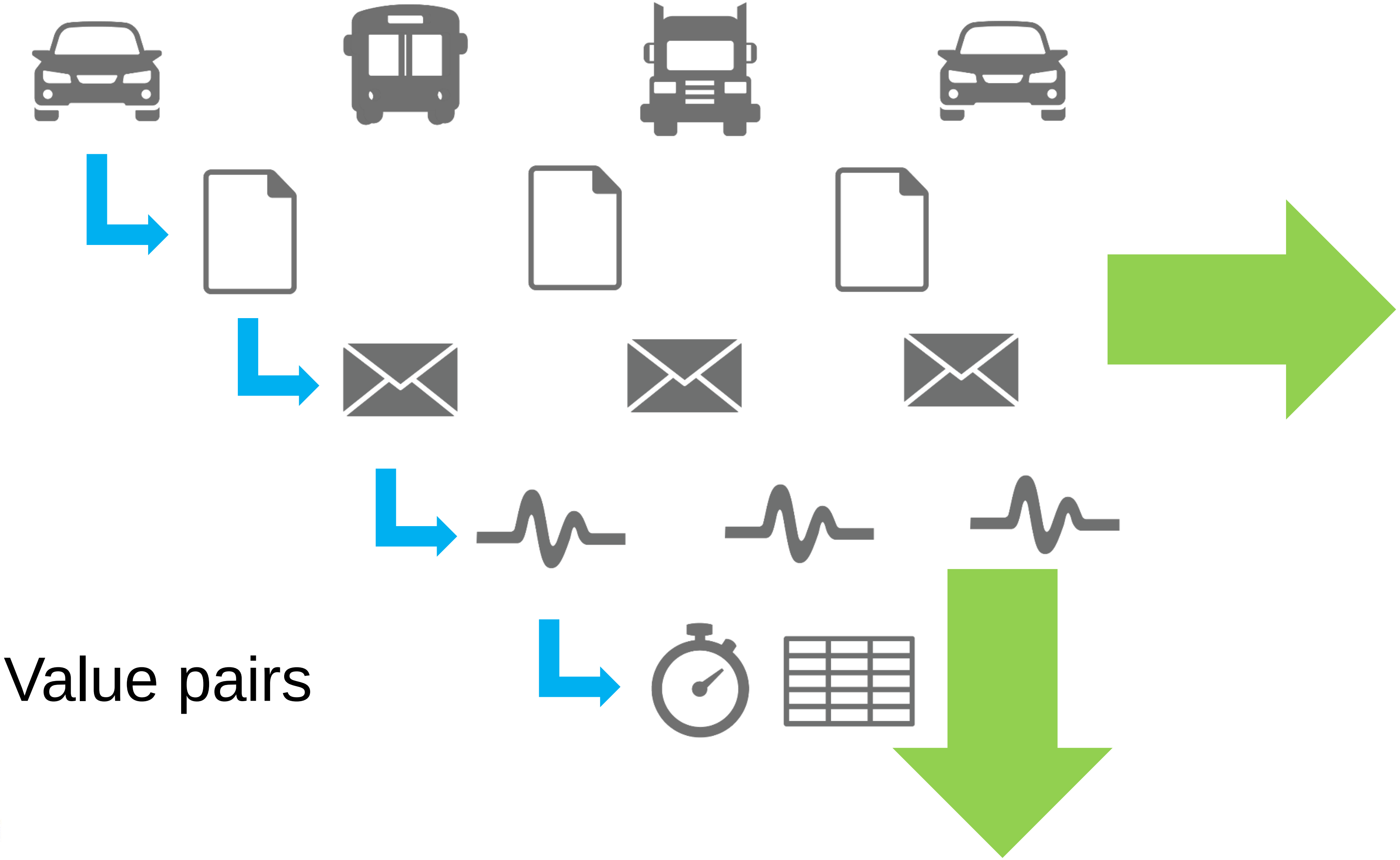
Vehicles

Trips (files)

Messages

Signals

Time – Value pairs



The Data: Timestamped messages with JSON encoding

```
{
    "vehicles_id": {"$oid":"55a3fd0069702d5b41000000"},
    "time” : {"$date":"2015-07-13T18:01:35.000Z"},
    "kc” : 1975.0, "kff1225" : 100.65293, "kff125a" : 110.36619, …
}
```

Key

Timestamp

Values

```
{
    "vehicles_id": {"$oid":"55a3fe3569702d5c5c000020"}
    "time":{"$date":"2015-07-13T18:01:53.000Z"},
    "kc” : 2000.0, "kff1225" : 109.65293, "kff125a" : 115.36619,
     …
}
```

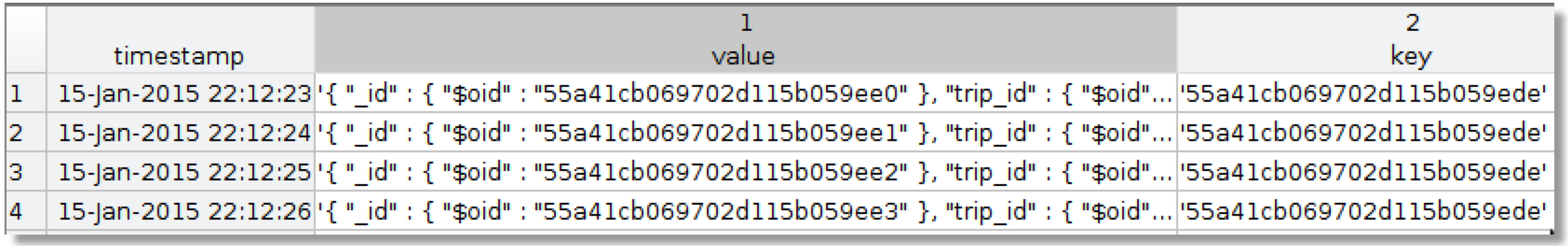
```
{
    "vehicles_id": {"$oid":"55a4193569702d115b000001"}
    "time":{"$date":"2015-07-12T19:04:04.000Z"}
    "kc":2200.0, "kff1225" : 112.65293, "kff125a" : 112.36619,
     …
}
```

1

Access and Explore Data

Access a Sample of Data

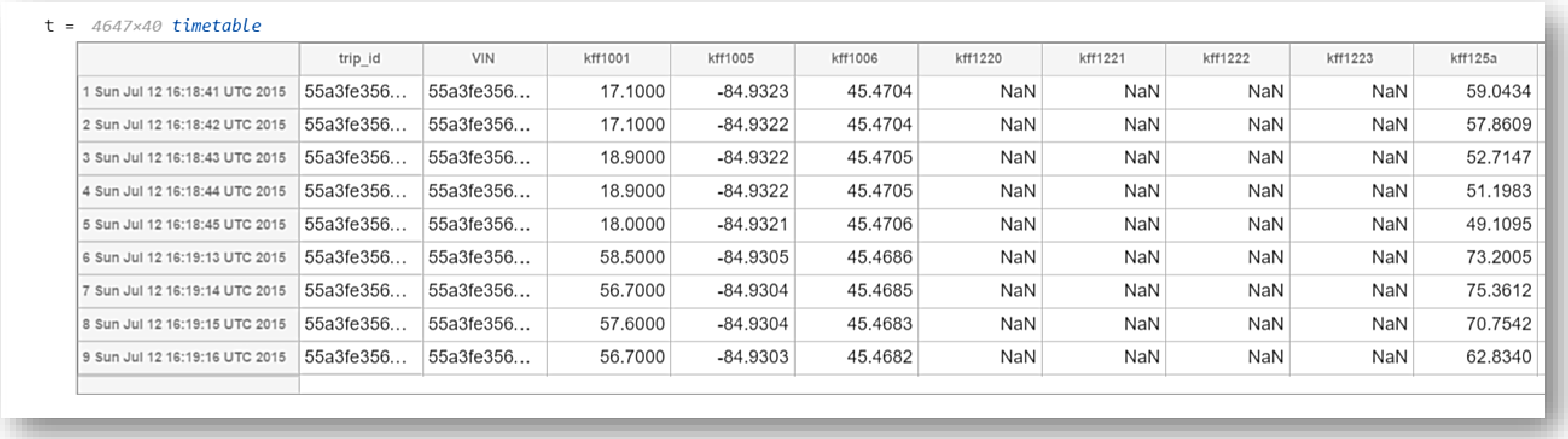
Raw Data



	timestamp	1 value	2 key
1	15-Jan-2015 22:12:23	'{ "_id" : { "$oid" : "55a41cb069702d115b059ee0" }, "trip_id" : { "$oid"...	'55a41cb069702d115b059ede'
2	15-Jan-2015 22:12:24	'{ "_id" : { "$oid" : "55a41cb069702d115b059ee1" }, "trip_id" : { "$oid"...	'55a41cb069702d115b059ede'
3	15-Jan-2015 22:12:25	'{ "_id" : { "$oid" : "55a41cb069702d115b059ee2" }, "trip_id" : { "$oid"...	'55a41cb069702d115b059ede'
4	15-Jan-2015 22:12:26	'{ "_id" : { "$oid" : "55a41cb069702d115b059ee3" }, "trip_id" : { "$oid"...	'55a41cb069702d115b059ede'

- ✓ Decode JSON data
- ✓ Create Timetable

Timetable



```
t = 4647×40 timetable
```

	trip_id	VIN	kff1001	kff1005	kff1006	kff1220	kff1221	kff1222	kff1223	kff125a
1 Sun Jul 12 16:18:41 UTC 2015	55a3fe356...	55a3fe356...	17.1000	-84.9323	45.4704	NaN	NaN	NaN	NaN	59.0434
2 Sun Jul 12 16:18:42 UTC 2015	55a3fe356...	55a3fe356...	17.1000	-84.9322	45.4704	NaN	NaN	NaN	NaN	57.8609
3 Sun Jul 12 16:18:43 UTC 2015	55a3fe356...	55a3fe356...	18.9000	-84.9322	45.4705	NaN	NaN	NaN	NaN	52.7147
4 Sun Jul 12 16:18:44 UTC 2015	55a3fe356...	55a3fe356...	18.9000	-84.9322	45.4705	NaN	NaN	NaN	NaN	51.1983
5 Sun Jul 12 16:18:45 UTC 2015	55a3fe356...	55a3fe356...	18.0000	-84.9321	45.4706	NaN	NaN	NaN	NaN	49.1095
6 Sun Jul 12 16:19:13 UTC 2015	55a3fe356...	55a3fe356...	58.5000	-84.9305	45.4686	NaN	NaN	NaN	NaN	73.2005
7 Sun Jul 12 16:19:14 UTC 2015	55a3fe356...	55a3fe356...	56.7000	-84.9304	45.4685	NaN	NaN	NaN	NaN	75.3612
8 Sun Jul 12 16:19:15 UTC 2015	55a3fe356...	55a3fe356...	57.6000	-84.9304	45.4683	NaN	NaN	NaN	NaN	70.7542
9 Sun Jul 12 16:19:16 UTC 2015	55a3fe356...	55a3fe356...	56.7000	-84.9303	45.4682	NaN	NaN	NaN	NaN	62.8340

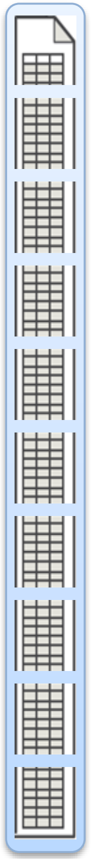
tall arrays R2016b

- What is a tall?
  - Tall is a new data type and a new way of working with Big Data in MATLAB (introduced in R2016b).

- Lots of observations.
  - Tall refers to data types and algorithms for use with **data that has more rows than will fit into the memory** of a single machine or cluster.

- 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.)

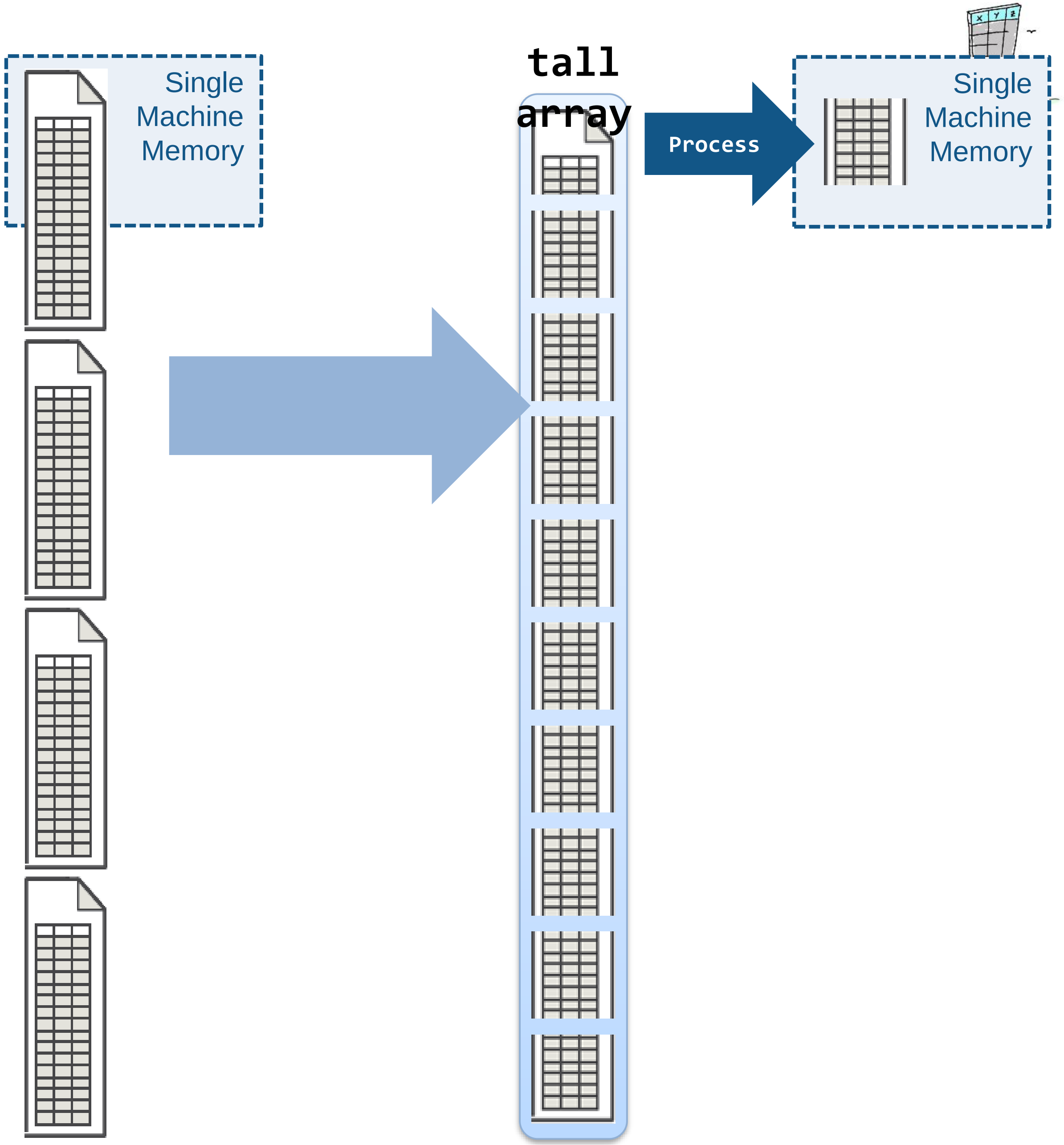


`tall` arrays R2016b

- 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

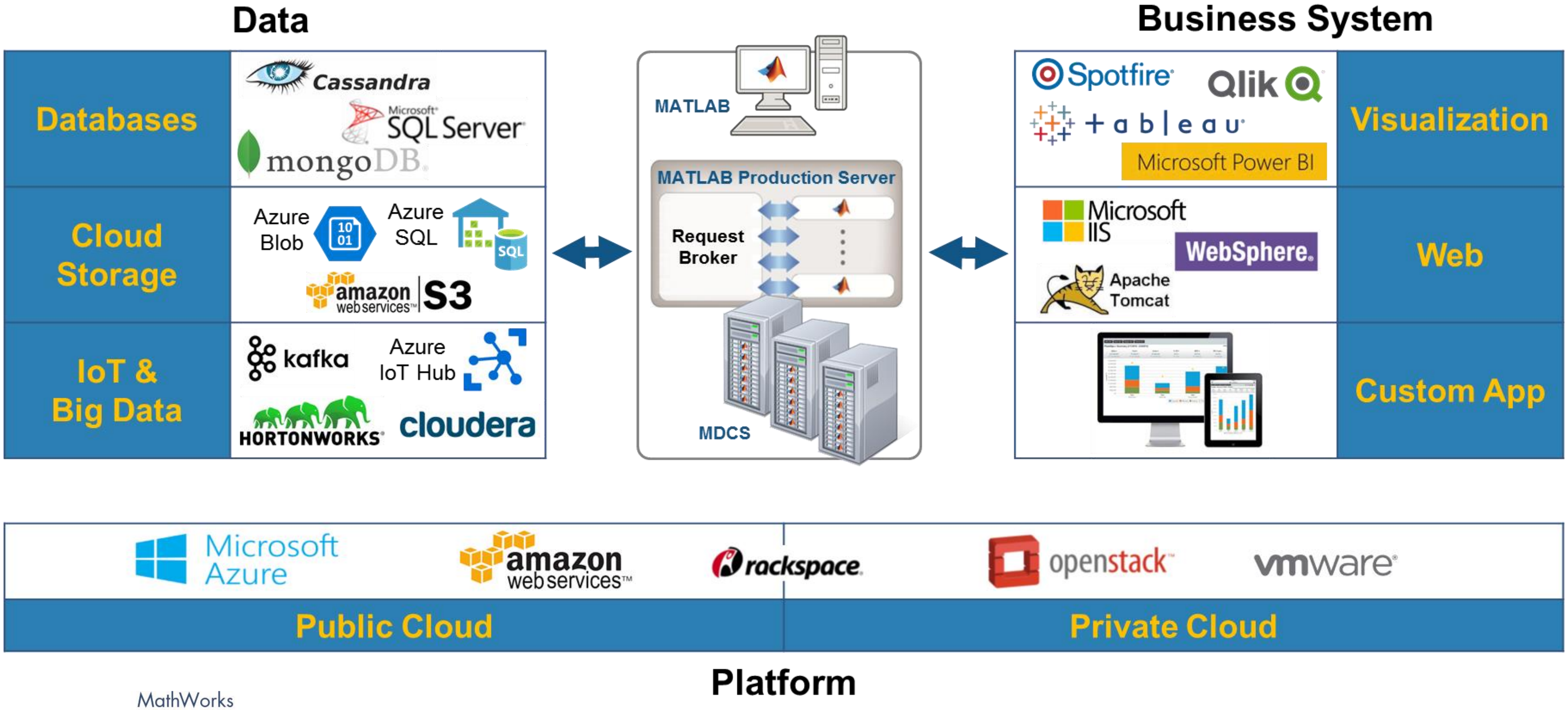


Workflow Pattern

Access out of memory data	`datastore & tall`
Work with subsets of your data	`findgroups, splitapply, cellfun`
Develop functions for event detection and calculation	`Normal MATLAB code`
Apply functions to all of your data	`cellfun`
Aggregate, summarize, & visualize	`table, histogram, heatmap, boxplot, binScatterPlot`

Enterprise Integration

Integrate MATLAB analytics into your technology stack



Data

Databases: Cassandra, Microsoft® SQL Server®, mongoDB®

Cloud Storage: Azure Blob, Azure SQL, amazon web services™ S3

IoT & Big Data: kafka, Azure IoT Hub, HORTONWORKS®, cloudera

MATLAB

MATLAB Production Server

Request Broker

MDCS

Business System

Visualization: Spotfire®, Qlik, +ableau®, Microsoft Power BI

Web: Microsoft IIS, WebSphere®, Apache Tomcat

Custom App

Public Cloud: Microsoft Azure, amazon web services™, rackspace®

Private Cloud: openstack™, vmware®

Platform

Key Takeaways

- Achieve success in Vehicle Fleet Analytics by utilizing new MATLAB **data types**, specifically `tall` Arrays for **out of memory** data sets
- Leverage `timetables` and the functions built to work on them to help do the **difficult time-series tasks** (`synchronize` and `retime`)
- **Scale** your work up with parallel computing toolbox on the desktop or the MATLAB Distributed Computing Server on **Hadoop**

MathWorks®

Accelerating the pace of engineering and science