



Spark
summit 2015

SparkR

The Past, Present and Future

Shivaram Venkataraman



Big Data & R

DataFrames
Visualization
Libraries



+

Data

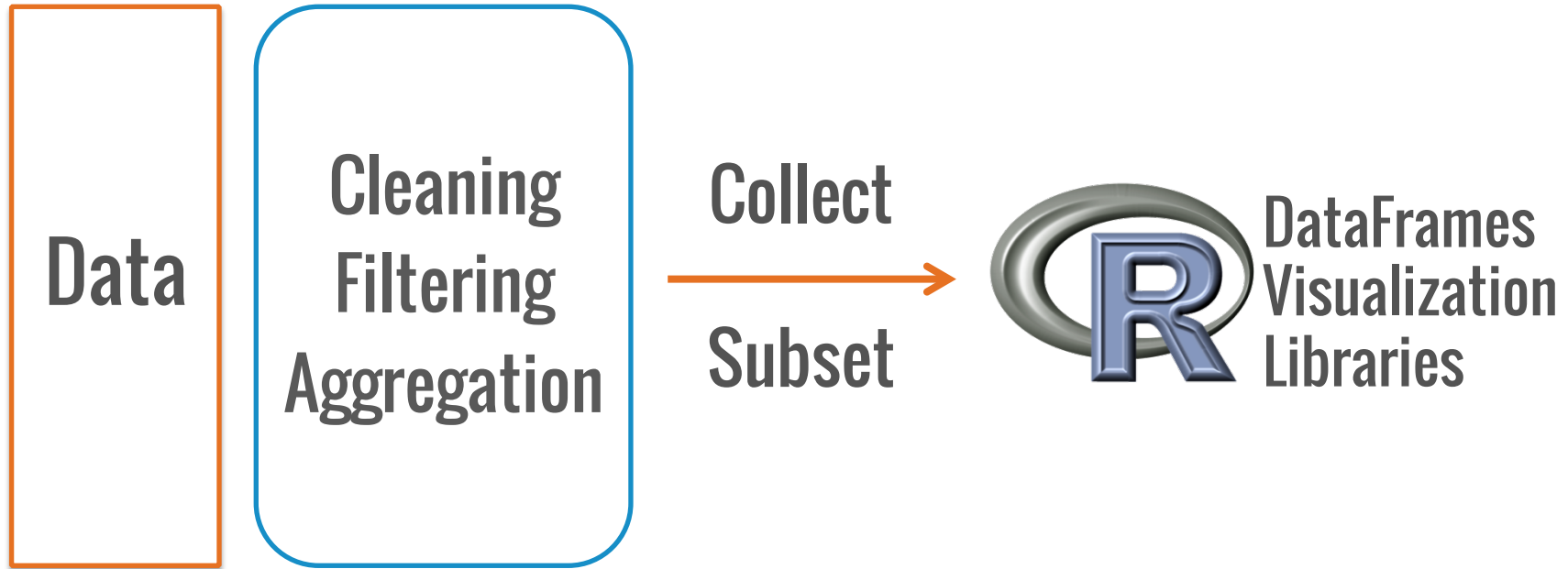
Big Data & R

Big Data
Small Learning

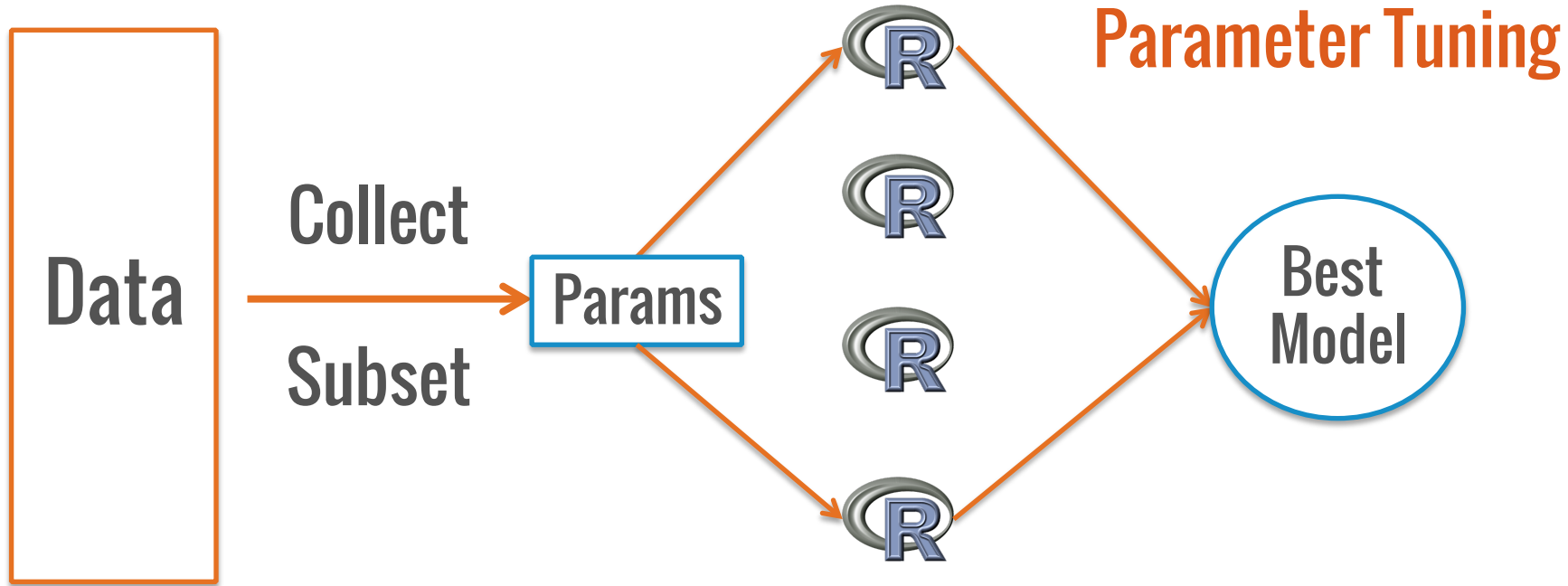
Partition
Aggregate

Large Scale
Machine Learning

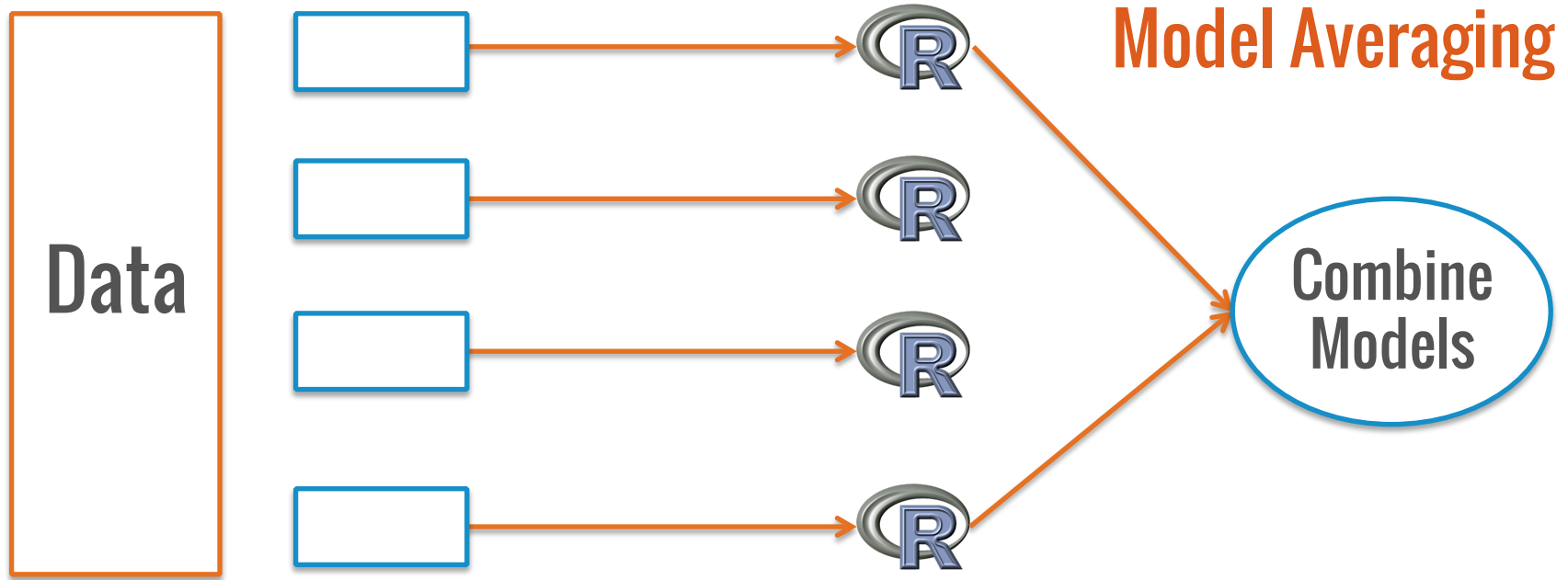
1. Big Data, Small Learning



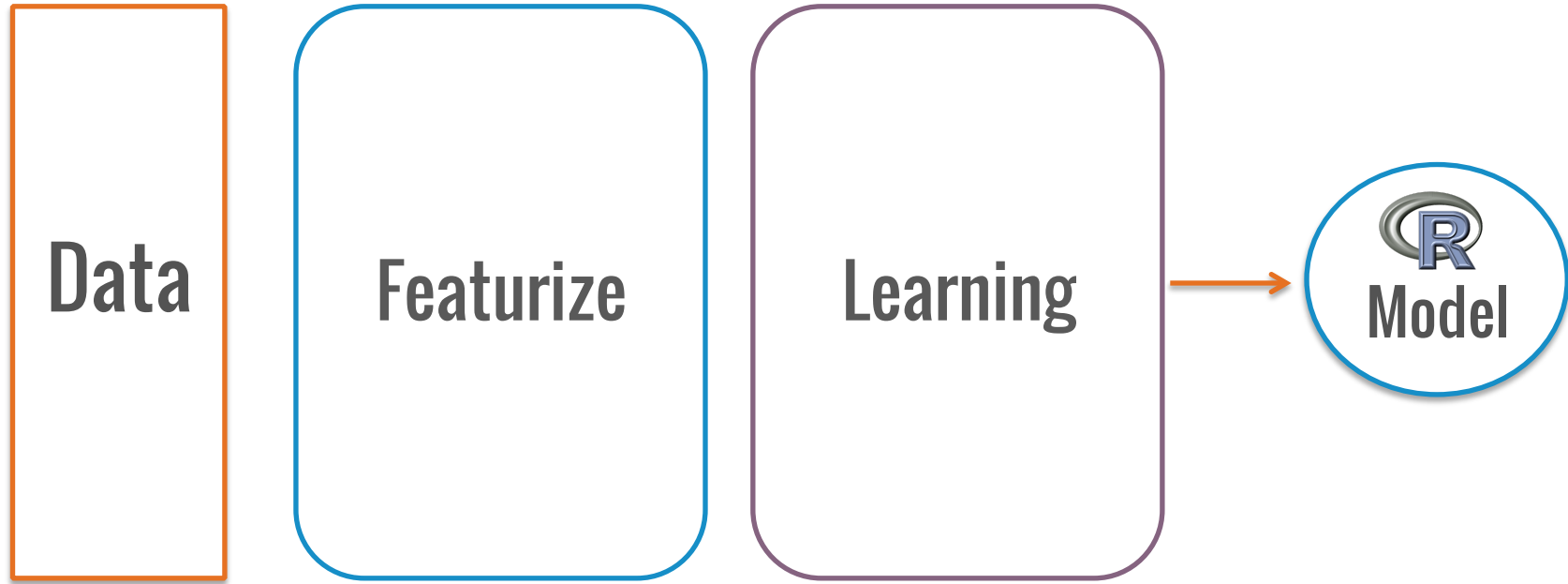
2(a). Partition Aggregate



2(b). Partition Aggregate



3. Large Scale Machine Learning



Big Data & R

Big Data

Small Learning

Partition

Aggregate

Large Scale

Machine Learning

SparkR:

Unified approach

Outline

Project History

Current Release

SparkR Future

Speed

Statistics

Scalable



+



DataFrames

Flexible

Visualization



RDD

Parallel Collection

Transformations

map
filter
groupBy

...

Actions

count
collect
saveAsTextFile

...

**R + RDD =
RRDD**

apply
applyPartition
groupByKey
collect
cache
...

broadcast
includePackage
textFile

Example: Word Count

```
library(SparkR)
lines <- textFile(sc, "hdfs://my_text_file")
words <- flatMap(lines,
                 function(line) {
                   strsplit(line, " ")[[1]]
                 })
wordCount <- lapply(words,
                    function(word) {
                      list(word, 1L)
                    })
counts <- reduceByKey(wordCount, "+", 2L)
output <- collect(counts)
```

Initial Prototype

Standalone R package



[amplab-extras](#) / **SparkR-pkg**

Install from github

Open Source Development

1. Architecture

2. Usability

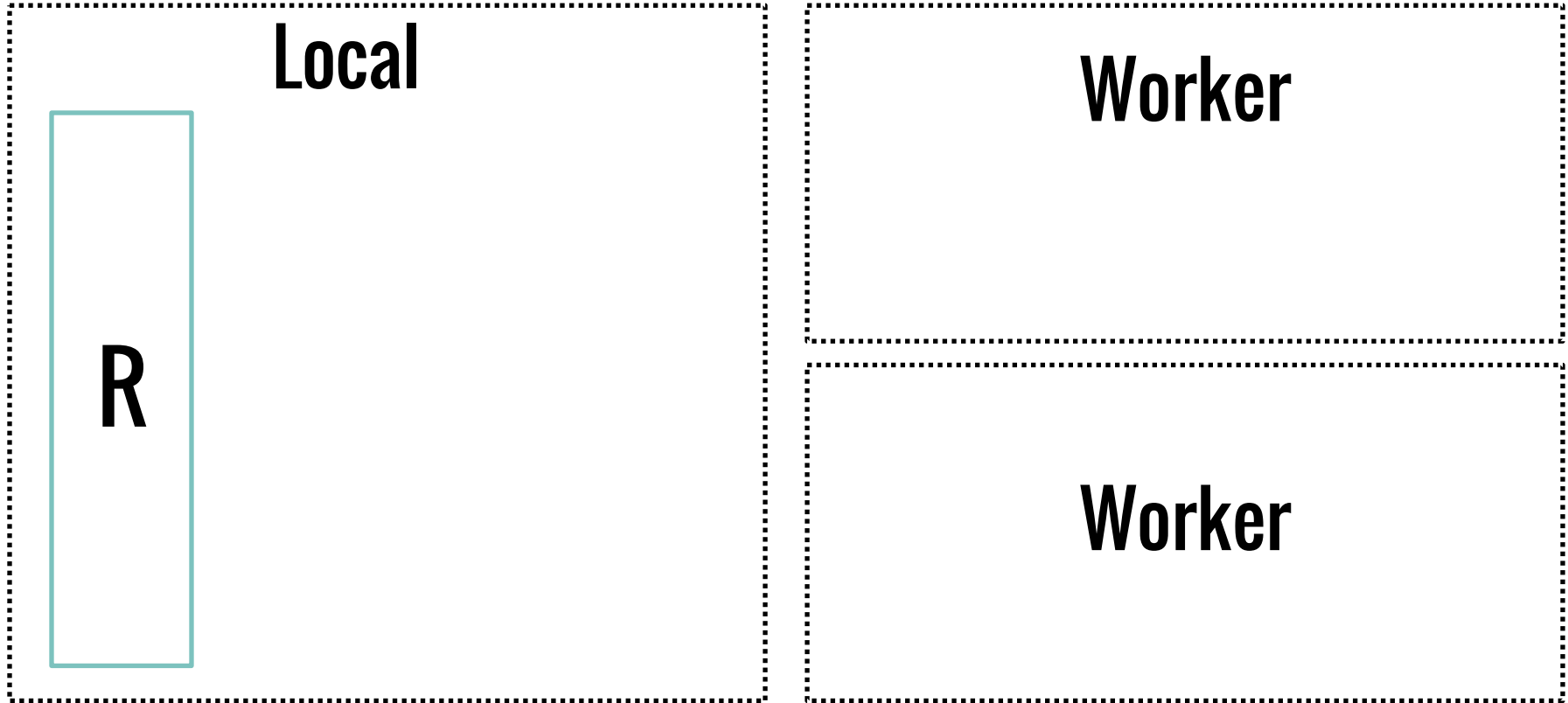
Architecture

Local

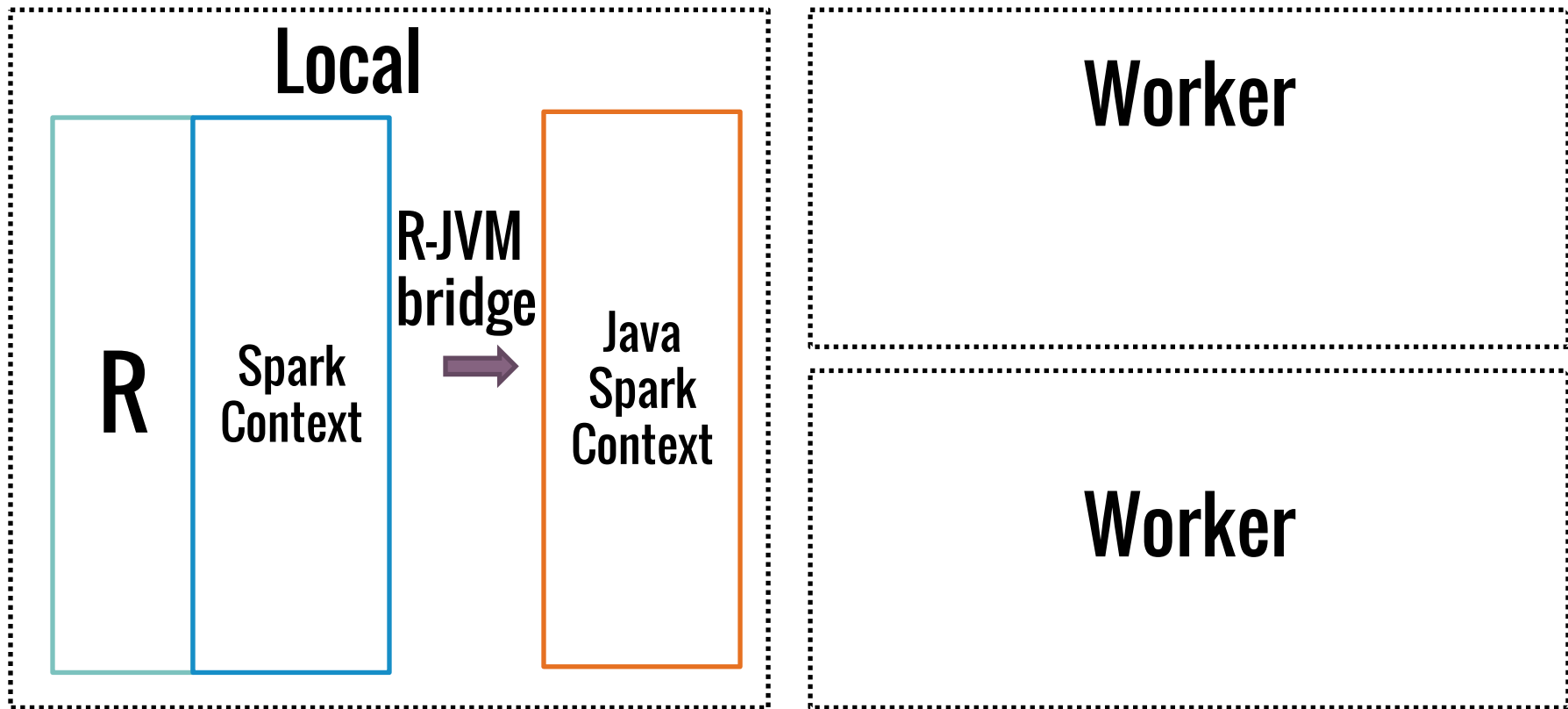
Worker

Worker

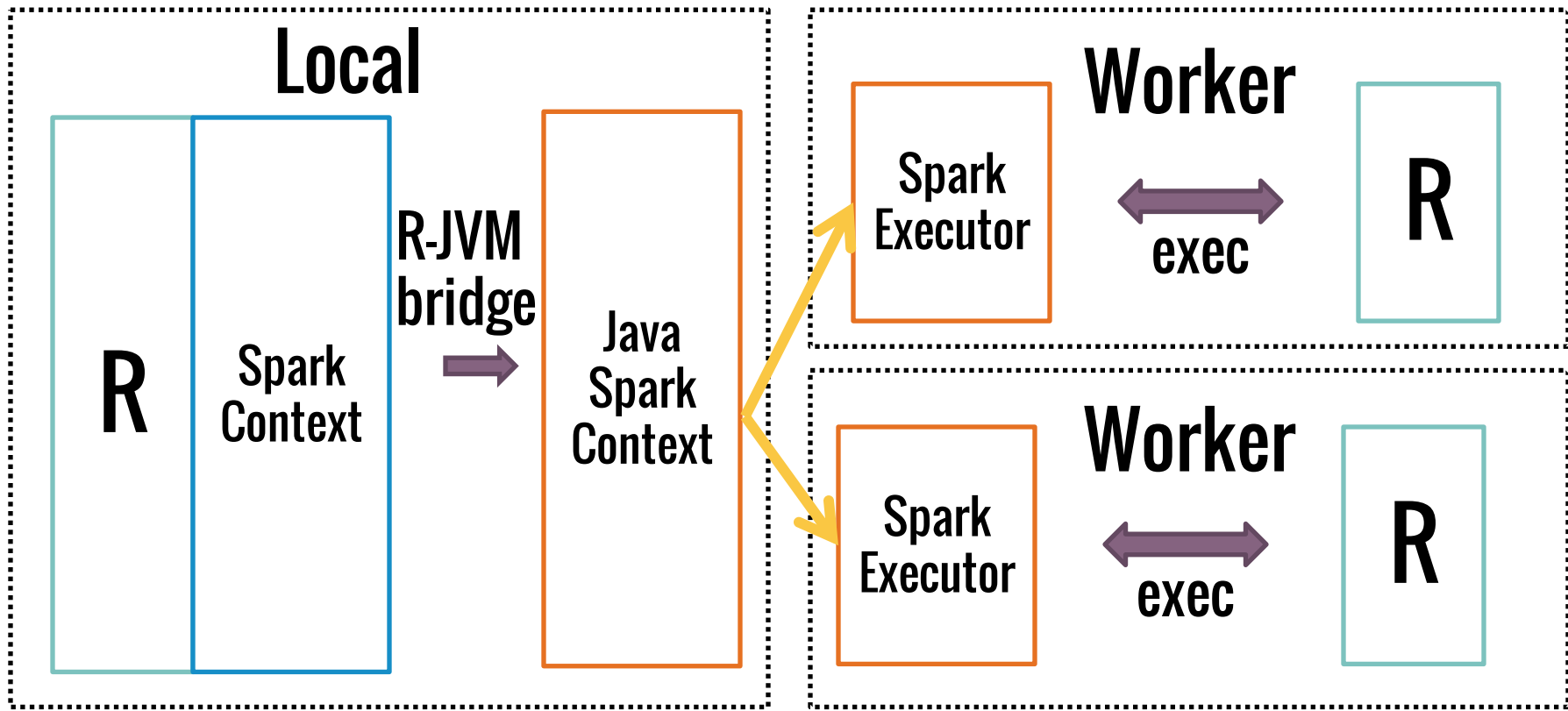
Architecture



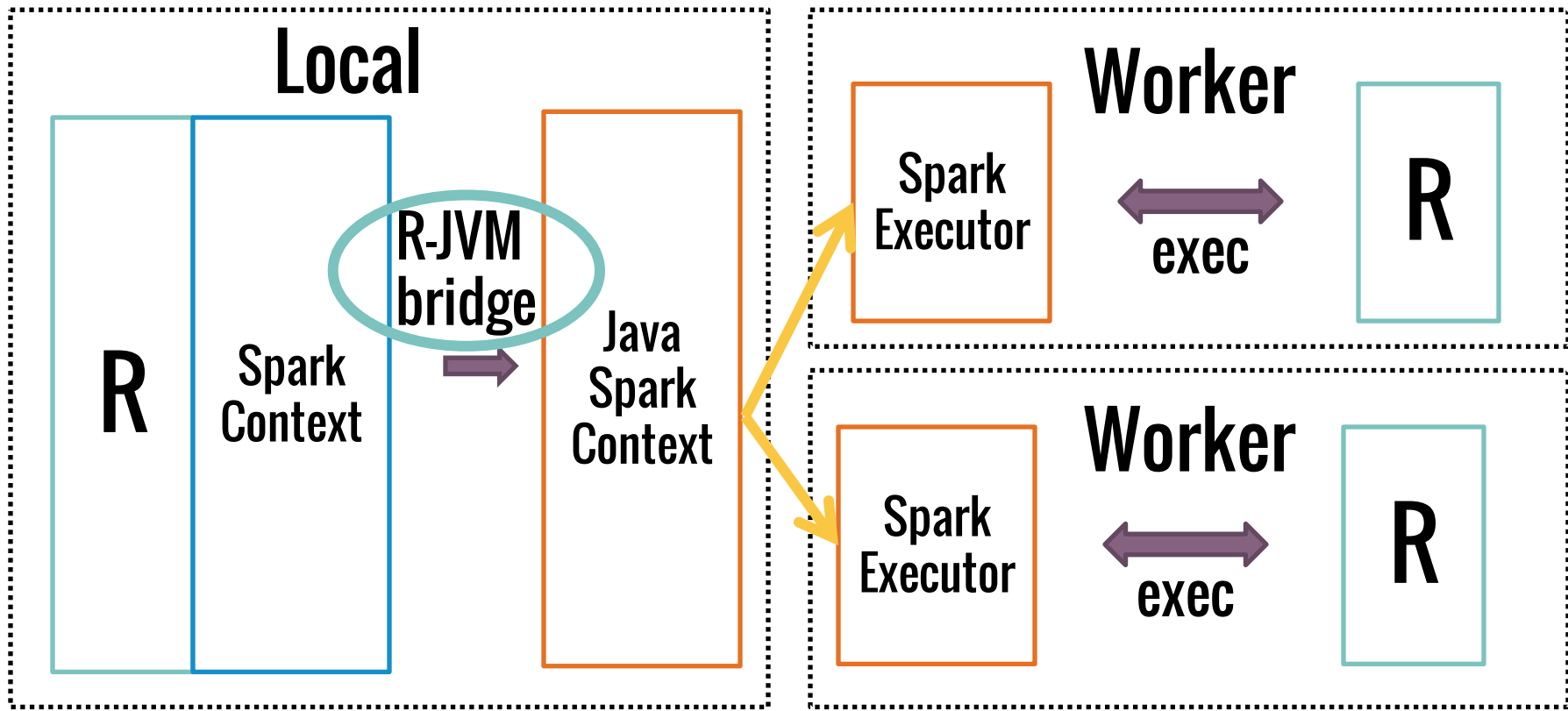
Architecture



Architecture



Architecture



R-JVM Bridge

Layer to call JVM
methods directly from R

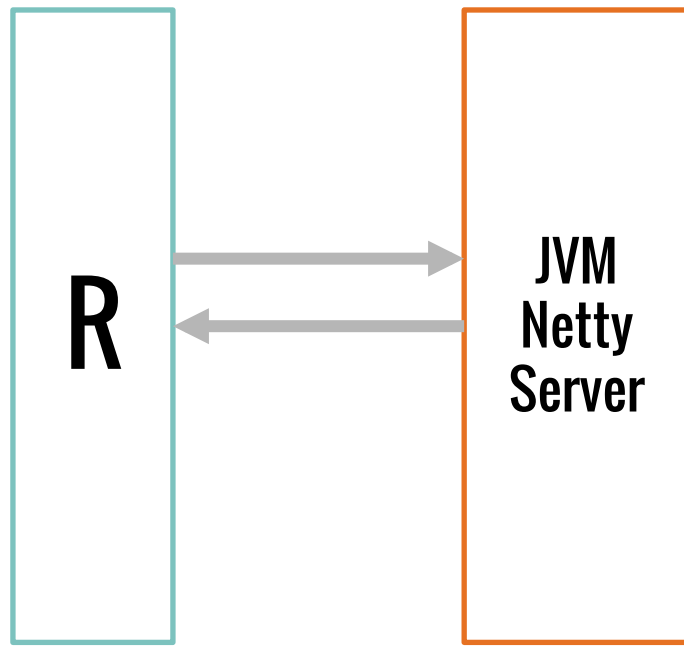
Automatic argument
serialization

```
result <-  
  callJStatic(  
    "sparkr.RRDD",  
    "someMethod",  
    arg1,  
    arg2)
```

R-JVM Bridge

Use *sockets* for communication

Supported across platforms, languages



Usability

Need for Data Inputs

Read in CSV, JSON, JDBC etc.

High-level API for data manipulation

SparkR DataFrames

DataSources API

```
people <- read.df(  
  "people.json",  
  "json")
```

Support for schema

```
avgAge <- select(  
  df,  
  avg(df$age))
```

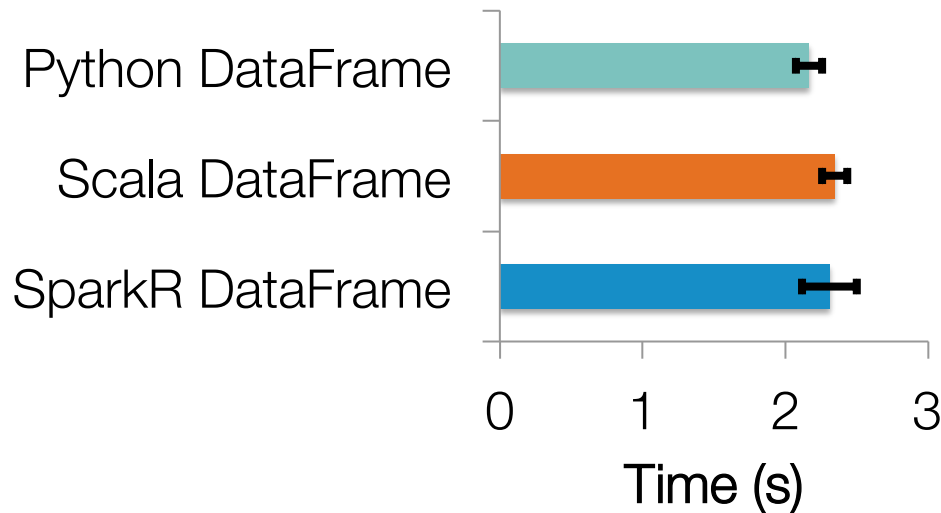
dplyr-like syntax

```
head(avgAge)
```


SparkR DataFrames

Scala Optimizations

Released in Spark 1.4!



Demo: github.com/cafreeman/SparkR_DataFrame_Demo

SparkR Future

Big Data & R

Big Data

Small Learning

Partition

Aggregate

Large Scale

Machine Learning

Big Data, Small Learning

SparkR DataFrames: Read input, aggregation
Collect results, apply machine learning

Upcoming features:

- Support for R transformations

- More column functions (e.g. math, strings)

Partition Aggregate

Upcoming feature:

Simple, parallel API for SparkR

Ex: Parameter tuning, Model Averaging

Integrated with DataFrames

Use existing R packages

Large Scale Machine Learning

Integration with MLlib

Support for GLM, KMeans etc.

```
model <- glm(  
  a ~ b + c,  
  data = df)
```

Large Scale Machine Learning

Key Features

DataFrame inputs

R-like formulas

Model statistics

```
model <- glm(  
  a ~ b + c,  
  data = df)
```

```
summary(model)
```

Extensibility

Existing data sources

R package support on
spark-packages.org

Example packages

```
./bin/sparkR  
--packages spark-csv
```


Developer Community

>20 contributors including
AMPLab, Databricks, Alteryx, Intel

R and Scala contributions welcome !

Big data processing from R

SparkR

DataFrames in Spark 1.4

Future: Large Scale ML & more