EARL 2015

EFFECTIVE APPLICATIONS OF THE R LANGUAGE

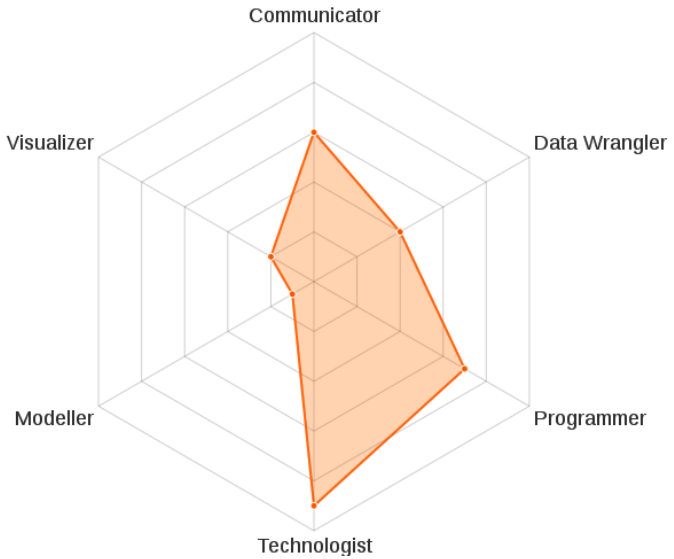
LONDON 14 - 16 SEPTEMBER

Apache Spark and R

A (big data) love story?

Mark Sellors - Technical Architect @ Mango Solutions

About me.



- Technical Architect
- Design and deploy analytic computing environments
- Not really an R user but have broad knowledge of the analytic computing ecosystem

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

EARL 2015

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

Overview

- The rise of big data
- Barriers to big data
- Big Data vs R
- Spark
- Spark and Hadoop
- SparkR
- Is it a love story?

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

The rise of ‘big data’

- Storage prices
- Commodity compute infrastructure

- Volume of data
- Hadoop ties the two together

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

Barriers to ‘big data’

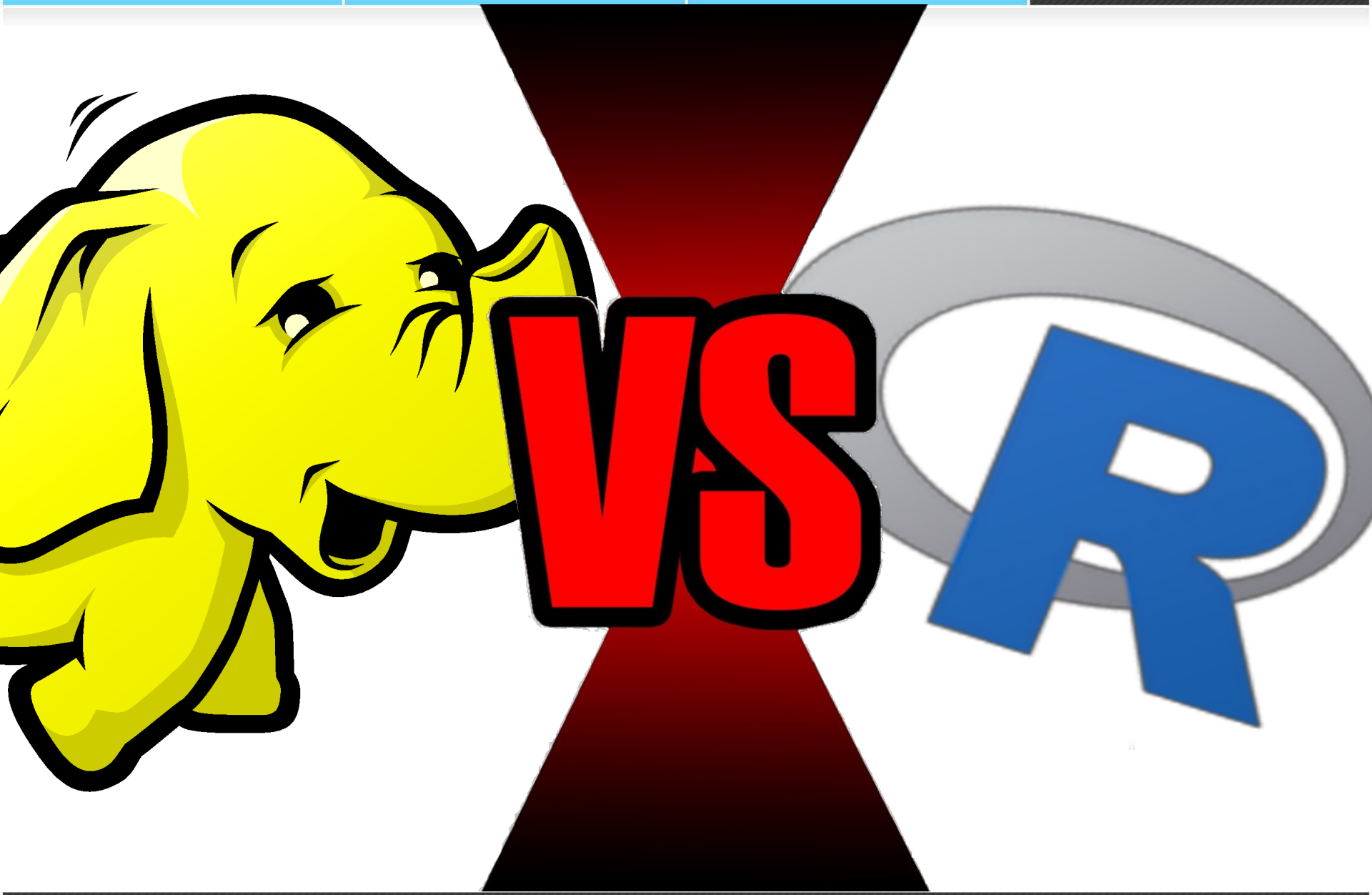
- Hadoop is complex ecosystem
- Primary programming paradigm, Map/Reduce jobs, largely written in Java
- Map/Reduce unsuited to exploratory, interactive analysis
- Map/Reduce is slow
- RHadoop is built on top of Map/Reduce

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

Some problems with ‘big data’

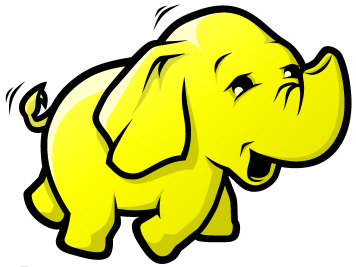
- Many hadoop deployments do not achieve an appreciable ROI.
- Hard to find the staff with crossover skills
  - infrastructure
  - analysts
- Existing business processes not fit for purpose

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com



Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

Hadoop



- Until recently limited to batch based operations
- MASSIVE data sets
- easy to add storage/compute capacity

But…

- Map/Reduce operations can be quite slow
- Hard to find/deploy appropriate talent

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

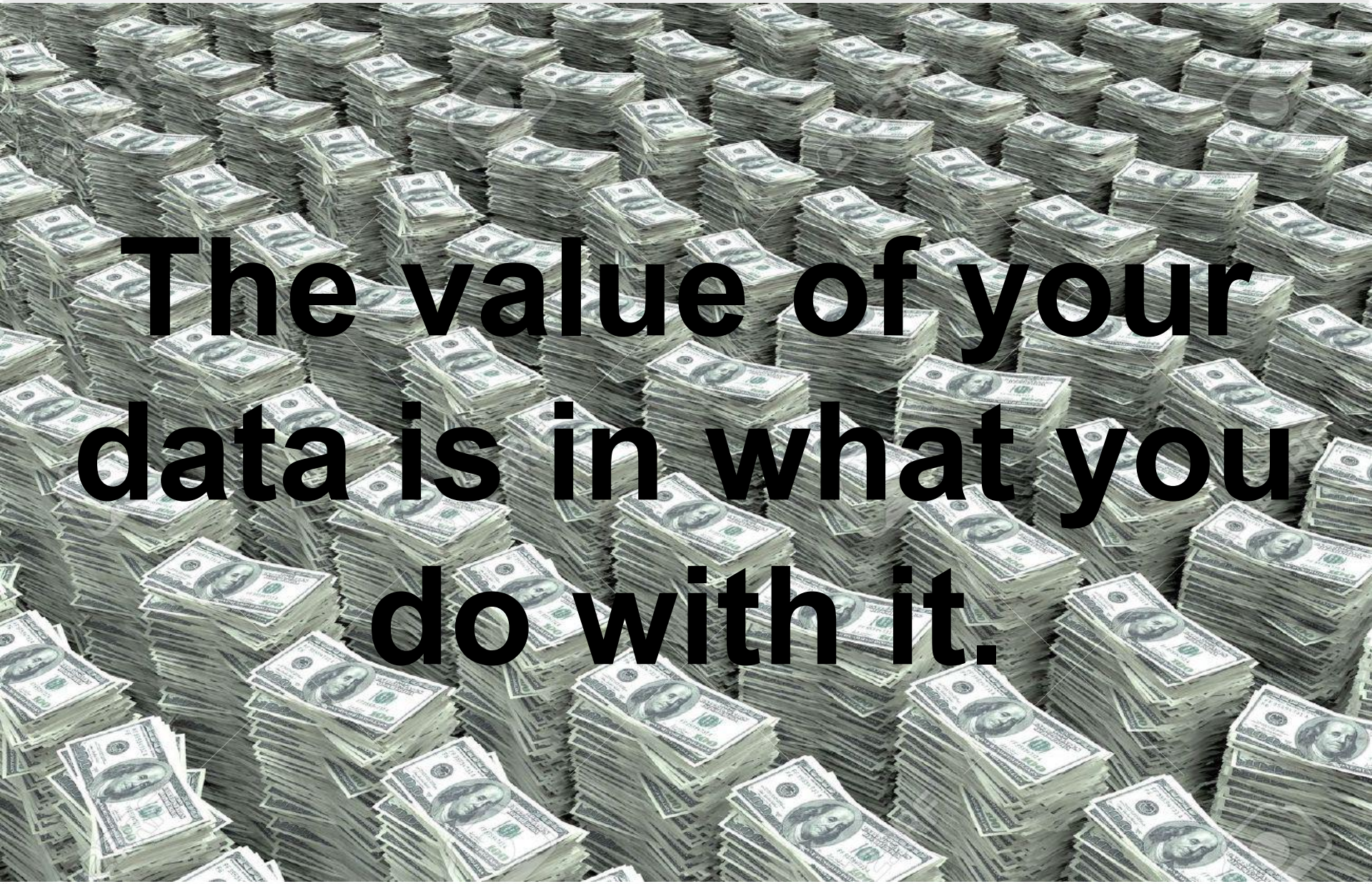
R

- Interactive
- fast
- great for exploratory or batch

But…

- Single threaded
- Limited by available memory

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com



Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

EARL 2015

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

What is it?

- Open source cluster computing framework
- Relies heavily on in memory processing
- One of the most contributed-to big data projects of the past year
- Started in the AMPLab at UC Berkeley in 2009

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

What problem does it solve

- In memory makes for very fast data processing
- minimal disk IO
- High level programming abstraction reduces the amount of code
- In turn makes it more suitable for exploratory work.

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

How does it do it?

- Provides a core programming abstraction called RDD
- The RDD API has been extended to include DataFrames
- Can deploy ad-hoc processing clusters as well as integrate with HDFS,

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

“Will Spark replace Hadoop?”

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

Spark and Hadoop

- Very Complimentary.
- Spark already comes with all the major Hadoop distributions
- easier to use and faster than map/reduce
- suitable for exploratory work, which previously was difficult in hadoop deployments

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

How does this fit with R

- Originally supported languages Scala, Java and Python
- SparkR was a separate project
- Integrated into Spark as of v1.4
- Support is still evolving - v1.5 released last week
- MASSIVE data frames

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

SparkR Features

- Designed to be familiar
- Massive DataFrames
- SQL operations on those DataFrames
- Fitting of GLM's
- Works on top of Hadoop or as a stand alone cluster
- Load data from a variety of sources

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

Spark SQL

- Arbitrary SQL operations on massive in-memory data frames
- Treats the data frame as though it were a database table
- Useful for exploring your data set
- Also great for creating subsets

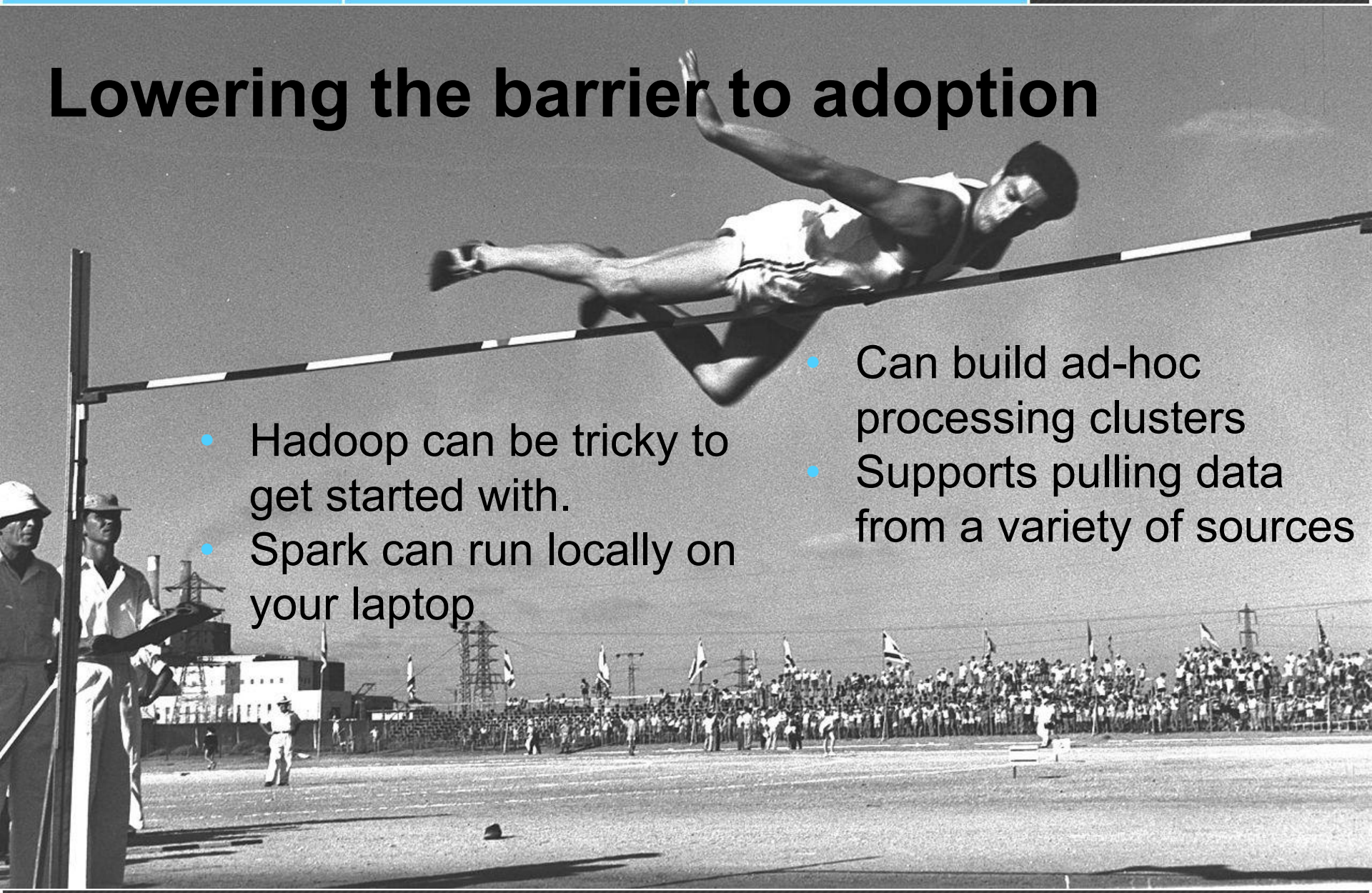
Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

```
# Create the DataFrame
df <- createDataFrame(sqlContext, iris)

# Fit a linear model over the dataset.
model <- glm(Sepal_Length ~ Sepal_Width + Species, data = df, family =
"gaussian")

# Model coefficients are returned in a similar format to R's native glm().
summary(model)
##$coefficients
##                    Estimate
##(Intercept)        2.2513930
##Sepal_Width        0.8035609
##Species_versicolor 1.4587432
##Species_virginica  1.9468169

# Make predictions based on the model.
predictions <- predict(model, newData = df)
head(select(predictions, "Sepal_Length", "prediction"))
##  Sepal_Length prediction
##1          5.1   5.063856
##2          4.9   4.662076
##3          4.7   4.822788
##4          4.6   4.742432
##5          5.0   5.144212
##6          5.4   5.385281
```

Source: http://spark.apache.org

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

Lowering the barrier to adoption



- Hadoop can be tricky to get started with.
- Spark can run locally on your laptop

- Can build ad-hoc processing clusters
- Supports pulling data from a variety of sources

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

What's in it for me?

- Currently supports:
  - DataFrames
  - SparkSQL
  - limited subset of MLlib

- Is missing any native R support for:
  - Spark Streaming
  - GraphX

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

Is it a love story?

- It wasn't originally, but things are heating up
- Data not getting any smaller
- Dramatically lowers the barrier to entry
- Evolving rapidly

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com

@MangoTheCat / @sellorm

Mark Sellors - Technical Architect @ Mango Solutions
msellors@mango-solutions.com