Things I tend to forget

Slides from today’s Big Data Step-by-Step Tutorials: Infrastructure series and Intro to R+Hadoop with RHadoop’s rmr

March 10, 2012 — Jeffrey Breen
Here are my presentations from today’s Boston Predictive Analytics Big Data Workshop.

All code and config files are available at github: https://github.com/jeffreybreen/tutorial-201203-big-data

My portion of the workshop was divided into four parts, three focusing on different infrastructure scenarios and ending with a deep dive into the rmr R package:

Big Data Step-by-Step: Infrastructure 1/3: Local VM

> **Starting small.** Just because Big Data tools like Hadoop were designed to run at “web-scale,” across many nodes, doesn’t mean you need to build a cluster—or even dedicate a single machine—to get started. In this deck we download and install a virtual machine from Cloudera which comes complete with a functioning, single-node Hadoop installation. As long as you restrict the size of your data set appropriately, this is great way to become accustomed to Hadoop and its tools. We walk through running a Hadoop Streaming job to make sure everything works. We later use this same VM to spawn a Hadoop cluster in the cloud (see part 3).

Atmosphere
RESEARCH GROUP

Big Data Step-by-S

Boston Predictive Analytics
Big Data Workshop

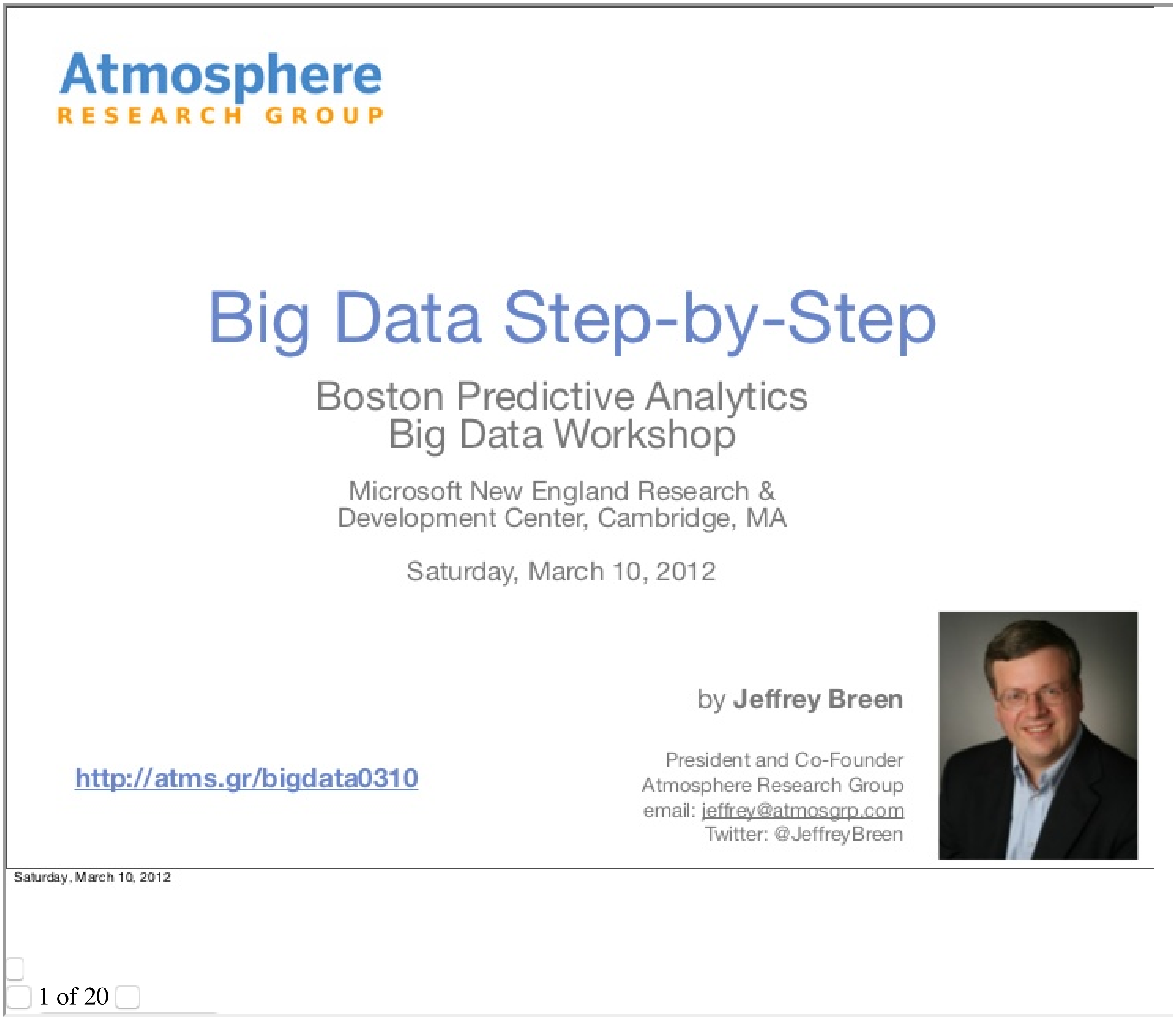
Microsoft New England Research &
Development Center, Cambridge, MA

Saturday, March 10, 2012

by Jeffrey B

Big Data Step-by-Step: Infrastructure 2/3: Running R and RStudio on EC2

Not everyone has Big Data. Some of us have an occasional need to analyze a data set larger than comfortably fits in our existing analysis environment either due to disk, CPU, or memory constraints. For these times, launching a single, large machine in the cloud may fit the bill. This part of presentation walks through how to launch just such a machine using Amazon’s EC2 cloud computing platform. Since I tend to run R and RStudio on Linux, that’s the focus of this tutorial, but the general outline may be helpful to others as well.



Atmosphere
RESEARCH GROUP

Big Data Step-by-Step

Boston Predictive Analytics
Big Data Workshop

Microsoft New England Research &
Development Center, Cambridge, MA

Saturday, March 10, 2012

by **Jeffrey Breen**

President and Co-Founder
Atmosphere Research Group
email: jeffrey@atmosgrp.com
Twitter: @JeffreyBreen

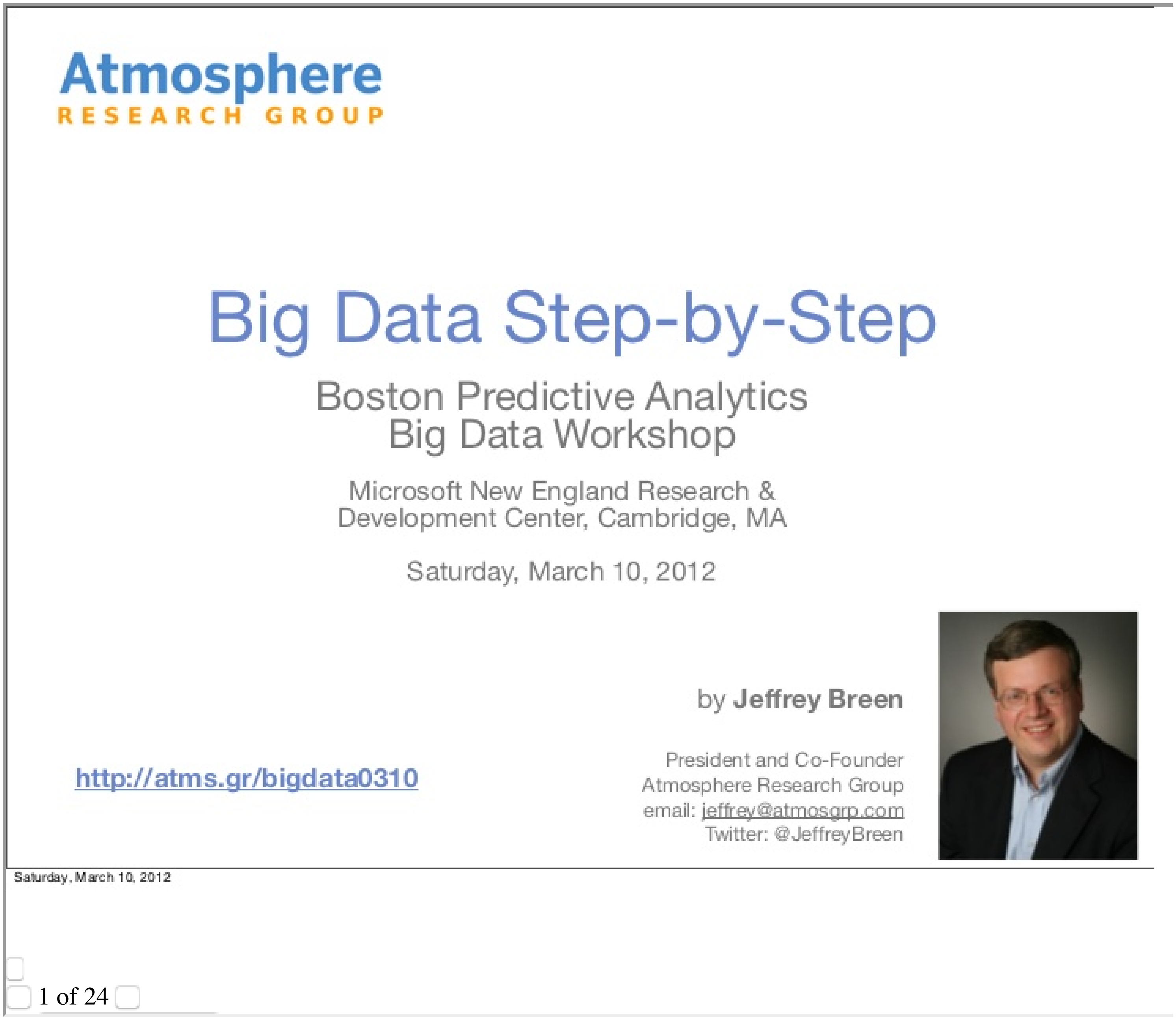
http://atms.gr/bigdata0310

Saturday, March 10, 2012

1 of 20

Big Data Step-by-Step: Infrastructure 3/3: Taking it to the cloud… easily… with Whirr

Scale up using the cloud. The Apache Whirr cloud management tool makes it easy to launch a Hadoop cluster on EC2. We use the Cloudera VM from presentation #1 as a launching point for the cluster and, thanks to a Whirr-generated proxy script, submit jobs and fetch results from our local VM just as before. For extra credit, we see how Whirr can save us money by bidding for excess capacity via EC2's spot instances.



Atmosphere
RESEARCH GROUP

Big Data Step-by-Step

Boston Predictive Analytics
Big Data Workshop

Microsoft New England Research &
Development Center, Cambridge, MA

Saturday, March 10, 2012

by **Jeffrey Breen**

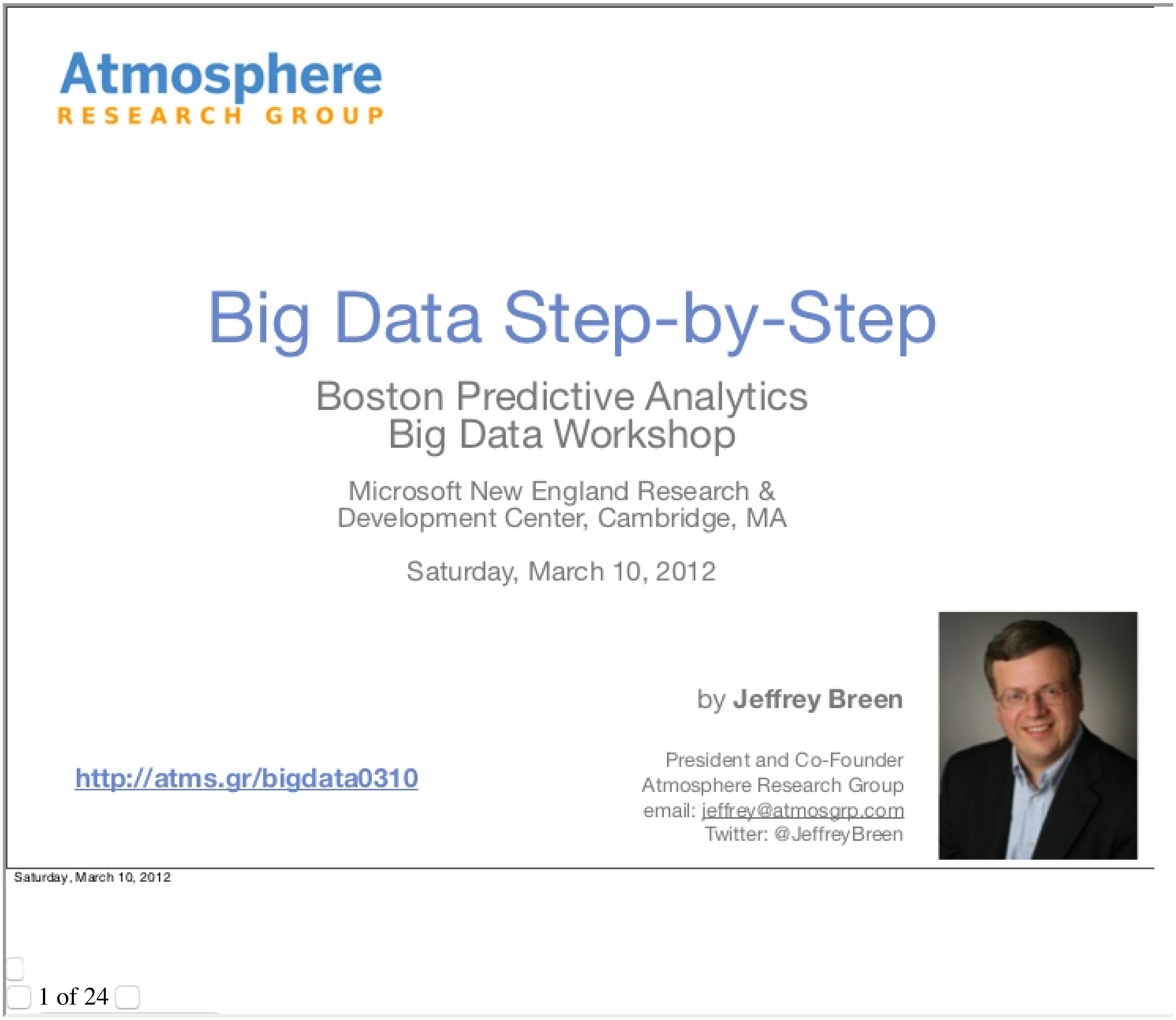
President and Co-Founder
Atmosphere Research Group
email: jeffrey@atmosgrp.com
Twitter: @JeffreyBreen

http://atms.gr/bigdata0310

Saturday, March 10, 2012

Big Data Step-by-Step: Using R & Hadoop (with RHadoop's rmr package)

Crunching Big Data with R. Originally a Java-only ecosystem, Hadoop Streaming allows the creation of mappers, reducers, and combiners in any language which can handle stdin and stdout—but that doesn't mean you want to have to write code to manage I/O at that level. After a quick (and undoubtedly incomplete) survey of Hadoop-related R packages, we walk through some of the abstractions and features of RHadoop's rmr package which make it easier for R developers to get started. We walk through a sample mapper and reducer, demonstrating and documenting the native R objects which carry the data from step to step.



Atmosphere
RESEARCH GROUP

Big Data Step-by-Step

Boston Predictive Analytics
Big Data Workshop

Microsoft New England Research &
Development Center, Cambridge, MA

Saturday, March 10, 2012

by **Jeffrey Breen**

President and Co-Founder
Atmosphere Research Group
email: jeffrey@atmosgrp.com
Twitter: @JeffreyBreen

http://atms.gr/bigdata0310

Saturday, March 10, 2012

Thank you to the session's sponsors, all the speakers, and to an interesting and engaged audience. Special thanks to John Versotek for arranging such an informative and enjoyable day, and for the opportunity to take part.

Posted in Tutorials. Tags: airlines, Amazon EC2, Big Data, cloud computing, Cloudera, Hadoop, R, rstats, VMware, Whirr. 4 Comments »

4 Responses to "Slides from today's Big Data Step-by-Step Tutorials: Infrastructure series and Intro to R+Hadoop with RHadoop's rmr"

David Russo Says:
March 16, 2012 at 8:00 AM
Jeff, this is a comprehensive presentation set, thank you. Dave Russo, Raytheon

Reply
Arun Bala (@openingbrace) Says:
April 30, 2012 at 11:52 AM
Excellent use-case Jeff. Amazed to see such use of BigData. Good work. Keep posting!

Reply
duyujie Says:
May 16, 2012 at 4:25 AM
Thanks Jeff for your Big Data Step-by-Step Tutorials!

Reply
Vignesh Says:
October 31, 2012 at 5:47 AM
Thanks Jeff, for nice and structured presentation on R with Hadoop [RHipe and RHadoop] and whirr!!

Reply

« Use geom_rect() to add recession bars to your time series plots #rstats #ggplot

Slides from "Tapping the Data Deluge with R" lightning talk #rstats #PAWCon »