Big Data Analytics in R

Big Opportunity, Big Challenge

September, 2011
Most advanced statistical analysis software available

Half the cost of commercial alternatives

2M+ Users

2,500+ Applications

Statistics

Finance
Life Sciences
Manufacturing
Retail
Telecom
Social Media
Government

Predictive Analytics

Data Mining

Visualization

Forbes
Power in the Numbers
Quentin Hardy, 05.06.10, 09:00 AM EDT
The professor who invented analytic software for the experts now wants to take it to the masses

The New York Times
Data Analysts Captivated by R's Power
By ASHLEE VANCE
Published: January 8, 2009
Big Data and Big Opportunities

- **60%** Increase in Retail Net Profits
- **50%** Decrease in product development costs
- **7%** Reduction in working capital
- **$200B** Annual savings in US healthcare spending

Worldwide data created and replicated, Zettabytes*

*Source IDC, McKinsey Global Institute, 1 Zettabyte = 1 Trillion Gigabytes
“Companies that have massive amounts of data without massive amounts of clue are going to be displaced by startups that have less data but more clue.” -- Tim O’Reilly
Challenges for data driven organizations in the Age of Analytics

Access to **advanced analytics tools** to derive knowledge from an explosion of data

**Attracting analysts** at all levels capable of performing advanced analytics

Efficiently & cost effectively sharing analyses and **disseminating knowledge** throughout the organization
“R is the most powerful & flexible statistical programming language in the world”\(^1\)

- **Capabilities**
  - Sophisticated statistical analyses
  - Predictive analytics
  - Data visualization

- **Applications**
  - Real-time trading
  - Finance
  - Risk assessment
  - Forecasting
  - Bio-technology
  - Drug development
  - Social networks
  - .. and more
Exploding Popularity Driving Exploding Functionality

**Scholarly Activity**
Google Scholar hits ('05-'09 CAGR)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>0%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>SAS</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>SPSS</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>S-Plus</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Stata</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

"I’ve been astonished by the rate at which R has been adopted. Four years ago, everyone in my economics department [at the University of Chicago] was using Stata; now, as far as I can tell, R is the standard tool, and students learn it first."

Deputy Editor for New Products at Forbes

**Package Growth**
Number of R packages listed on CRAN

Source: [http://r4stats.com/popularity](http://r4stats.com/popularity); “Why R is a name to know in 2011”, Forbes

"A key benefit of R is that it provides near-instant availability of new and experimental methods created by its user base — without waiting for the development/release cycle of commercial software. SAS recognizes the value of R to our customer base…"

Product Marketing Manager SAS Institute, Inc
The rise of Data Science

86% Of statisticians consider themselves Data Scientists

1. Survey conducted at Joint Statistical Meeting August 2011
…but what is a Data Scientist?

Statistician + Computer Scientist
-or-
Computer Scientist + Statistician

Who:
- Obtains
- Cleans
- Structures
- Explores
- Analyzes
- Interprets

Data
Enabling Data Scientists as a seamless bridge between IT and Business

IT Operations & Data Infrastructure

Model Development

Data consumption & actionable decisions
What is Big Data?

**Large Structured Data**
1Gb? 1Tb? 1Pb?

**Unstructured Data**
*Log data, text, voice, video*

**“Big Models”**
Computationally intensive analyses, simulations, models with many parameters

“Big”, like beauty is in the eyes of the beholder

• Common theme is inability to analyze with legacy tools
Why deal with Big Data?

- Relax assumptions of Linearity or Normality
- Identify Rare Events or Low Incidence Populations
- Gain insight by studying Microstructure of data
- Generate better predictions and better understanding of effects
Many opportunities for improvement

Of Data Scientists feel the current tools are inadequate to meet their Big Data needs.¹

¹ Survey conducted at Joint Statistical Meeting August 2011
Two Big problems: capacity and speed

- **Capacity**: problems handling the size of data sets or models
  - Data too big to fit into memory
  - Even if it can fit, there are limits on what can be done
  - Even simple data management can be extremely challenging

- **Speed**: even without a capacity limit, computation may be too slow to be useful
A RevoScaleR algorithm is provided a data source as input.

- The algorithm loops over data, reading a block at a time. Blocks of data are read by a separate worker thread (Thread 0).
- Other worker threads (Threads 1..n) process the data block from the previous iteration of the data loop and update intermediate results objects in memory.
- When all of the data is processed a master results object is created from the intermediate results objects.
Distributed Computing

- Portions of the data source are made available to each compute node
- RevoScaleR on the master node assigns a task to each compute node
- Each compute node independently processes its data, and returns its intermediate results back to the master node
- Master node aggregates all of the intermediate results from each compute node and produces the final result
A common analytic platform across big data architectures
Creating an Analytics Center of Excellence

- HDFS
- Hadoop Cluster (R/Hadoop)
- Individual Data Scientists
- Database Appliance (Revolution R)
- Deployment Servers (DeployR)
- High Workload Clusters (Revolution/R ScaleR)
- Business Users
Use Case: Financial Analysis

EXAMPLES: PREVENTING A “FLASH CRASH”, LEVERAGING NEW DATA SOURCES
R in Finance Case Study

- Modeling and detection of asset price jumps

NYSE TAQ, 30 Dow Jones Industrial Average

- Trade data in txt format
- Raw trade data in R
- Cleaned trade data
- Clean trade data
- Clean quote data
- Matched trades and quotes
- Liquidity measure
- Aggregated liquidity

Source: Boudt, Leuven – R/Finance 2011
R in Finance Case Study

➢ Seamlessly leveraging alternative data sources

**twitteR (twitter client for R)**

R Package by Jeff Gantry

Search for tweets with the phrase "Stock Market", then plot frequency of tweets per hour:

**RGoogleTrends**

R Package by Duncan Temple Long

Source: Rothermich – R/Finance 2011
Use Case: Internet and Social Media

EXAMPLES: ONLINE USER BEHAVIOR
Social Media Case Study

Improving profitability by studying user behavior

“The extensive engagement of our players provides over 15 terabytes of game data per day that we use to enhance our games by designing, testing and releasing new features on an ongoing basis. We believe that combining data analytics with creative game design enables us to create a superior player experience.”

– Zynga S1 Filing

The Model\textsuperscript{1}:
• Rank pages by quality
• Remove pages one by one
• Calculate predicted effects of removal

Technically:
• Eliminate columns from the Markov transition matrices
• Multiply the initial state vector by the transition matrices
• Inspect changes to probability of “good” absorbing state

Social Media Case Study

Predicting Colleague-Colleague interactions at Facebook

How Facebook uses R:
• Experimentation for large machine learning models
  • Picking models
  • Feature selection
• Small / medium one-off analyses
  • User engagement studies
  • Analyses to improve internal processes
• Analysis and visualization for social network research

Applications of Colleague-Colleague study:
• Suggesting peer reviewers during performance review season
• Setting up optimally constructed teams
• Optimizing seating charts for productivity
• Automatically filtering internal feeds of employee content
• Suggesting new colleague interactions
• Giving managers insight into employees’ interactions

Source: “Criss-crossing the Org Chart – Predicting Colleague interactions with R”, E. Sun, useR 2010
Online retailing case study

Enabling new insights with Hadoop + R at Orbitz

Prototypes:
• User Segmentation
• Hotel Booking
• Airline Performance

Insights:
• Purchase behavior by browser use
• Seasonal variation of purchase behavior
Use Case: Healthcare & Life Sciences

EXAMPLES: GENOME ANALYSIS, PREDICTING HOSPITALIZATIONS
Life Sciences Case Study

Genome Wide Association Studies

Analysis Challenge:
- 22 genotype files total 25Gb
- 3m variables (SNPs) and 3k rows (subjects)
- Need to run 3m regressions per analysis

Solution:
- A computationally intensive Big Model
- Leverage external memory framework, data chunking, and task parallelization
Life Sciences Case Study

Predicting future hospitalizations

The Challenge:

• Each year, > 71M people in the US admitted to hospitals
• In 2006, > $30B spent on unnecessary hospital admissions
• Very few patients account for most spending
• Given patient history, medical records etc., can a predictive model be built to identify who is at risk for hospitalization?

SiliconANGLE

The $3 Million Heritage Health Prize: Big Data Enhancing Wellness Across the World

--Kit Dotson, February 4, 2011

➢ 33% of all Kaggle competitors report using R

➢ 50% of previous competition winners used R to create the most accurate predictive models

1 http://www.heritagehealthprize.com/c/hhp
2 http://blog.revolutionanalytics.com/2011/04/how-kaggle-competitors-use-r.html
Use Case: Network Analysis

EXAMPLE: TWITTER
Analyzing Networks

The Challenge:
• Post 9/11, network analysis has been viewed as increasingly important
• Analysts face a massive amount of data and insufficient analytical capability
• A dizzying of “single solution” tools are available

The R Advantage:
• R is a language for building solutions
• Community R experts have developed a diverse set of packages for analyzing networks
• Solutions range from basic to cutting edge

Solutions

igraph
• Graph theoretic foundations
• Community detection

Statnet
• Social Science focused
• Statistical analysis of networks
Mining online social graphs: Twitter example

twitteR and igraph
ggplot2
Use Case: Text & Data Mining

EXAMPLE: ANALYZING IQT PRESS RELEASES
Data Mining

The Challenge:
• The data explosion has made it increasingly difficult for analysts to know which signals to look for in data
• Data mining is about extracting small findings from large data...i.e. finding the needle in the haystack
• A dizzying of “single solution” tools are available

The R Advantage:
• Unlike alternate tools, R analysts can construct and modify specific solutions to their unique data mining tasks
• Community R experts have developed a diverse set of packages for data mining

Rcurl, XML
• Online data extraction

Rweka, tm
• Natural Language Processing and text analysis

Lattice, ggplot2
• Graphics

Reshape, plyR
• Data manipulation

topicmodels
• Topic modeling

Lattice, ggplot2
• Graphics

Revo ScaleR
• Big Data analysis
Analyzing InQTel Press Releases

Model data in less than 50 lines of R code…


Parse [...] Base

Model topicmodels

Visualize ggplot2

Terms from Topic Model of In-Q-Tel’s 2009 Press Releases

<table>
<thead>
<tr>
<th>government</th>
<th>visit</th>
</tr>
</thead>
<tbody>
<tr>
<td>critical</td>
<td>using</td>
</tr>
<tr>
<td>company</td>
<td>unique</td>
</tr>
<tr>
<td>capabilities</td>
<td>support</td>
</tr>
<tr>
<td>analysis</td>
<td>strategic</td>
</tr>
<tr>
<td>software</td>
<td>solutions</td>
</tr>
<tr>
<td>platform</td>
<td>president</td>
</tr>
<tr>
<td>enables</td>
<td>partnership</td>
</tr>
<tr>
<td>innovative</td>
<td>media</td>
</tr>
<tr>
<td>technology</td>
<td>intelligence</td>
</tr>
</tbody>
</table>
Use Case: Analytics Re-use and Knowledge Sharing

EXAMPLE: DOCUMENTATION WITH SWEAVE
Analytical reproducibility and knowledge sharing

The Challenge:
• Importance of data sharing has been widely publicized
• Next stage is to improve the ease of knowledge sharing
• Most solutions lack methodological transparency making sharing and duplicating analyses difficult

The R Difference:
• As the Lingua Franca of statistics, R creates a universal and binding platform for analysis
• Analysis scripts are easy to share and to audit
• R is easily integrated into many reporting and presentation tools

Solutions

SWeave
• Facilitates automated creation of professional publication ready documents

Revo DeployR
• Integrates R into online reporting and presentation tools for real-time, interactive reporting and analysis
Streamlining recurring analysis

• Analysts often have to produce the same reports (daily, weekly, etc.), using new or updated data
• Suppose we had to edit an analysis based on new data...

Traditional Work Flow

1. Download data, inspect (Excel)
2. Copy paste relevant new data
3. Apply analysis, maybe automated
4. Produce analytical output (charts, tables, summaries)
5. Import output into document
6. Check over to make sure no new error introduced

Using R and Sweave

1. Download data
2. Edit data input variable

```r
<<echo=true,keep.source=true>>=
df<-read.csv("new_data.csv")
@
```
3. Typeset document, send to customer
Use Case: Advanced Graphics & Data Visualization

EXAMPLE: WIKI LEAKS ANALYSIS
Analyzing data with advanced graphics

The Challenge:
• Conveying complex ideas in meaningful and understandable ways
• Sophisticated and flexible graphics are a critical component to any analytical presentation
• If a picture is worth 1,000 words, then having the tools to produce compelling data visualizations is priceless

The R Difference:
• R has amongst the most sophisticated and comprehensive graphics capabilities of any analysis tool
• Graphics can be used as-is, modified, or developed from scratch

Solutions
WikiLeaks Afghanistan War Diaries: Description

Report Volume at Monthly Intervals, by Region and Attack Target

Who Attacked
- ENEMY
- FRIEND
- NEUTRAL
- UNKNOWN

Date

Report Counts

RC CAPITAL

RC EAST

RC NORTH

RC SOUTH

RC WEST

UNKNOWN
WikiLeaks Afghanistan War Diaries: Geospatial Analysis

WikiLeaks Geospatial Attack Data by Year and Type (Afghanistan District Boundaries)
Thank you.

The leading commercial provider of software and support for the popular open source R statistics language.

www.revolutionanalytics.com  650.330.0553  Twitter: @RevolutionR
Follow-up Contact Information

- http://www.revolutionanalytics.com/
- Norman H. Nie, CEO
  - norman@revolutionanalytics.com
- Jeff Erhardt, COO
  - jeff@revolutionanalytics.com
- Mike Minelli, VP Sales
  - mike@revolutionanalytics.com