Batter Up! Advanced Sports Analytics with R and Storm

Meeting the Real-Time Analytics Opportunity Head-On

December 11, 2014

Bill Jacobs
VP Product Marketing
Revolution Analytics
@bill_jacobs

Vineet Sharma
Dir., Partner Marketing
MapR Technologies

Allen Day
Principal Data Scientist
MapR Technologies
@allenday

Who Am I?

Bill Jacobs, VP Product Marketing
Revolution Analytics
@bill_jacobs

REVOLUTION ANALYTICS

MAPR

Polling Question #1:

- Who Are You? (choose one)
  - Statistician or modeler
  - Data Scientist
  - Hadoop Expert
  - Application builder
  - Data guru
  - Business user
  - Baseball fan

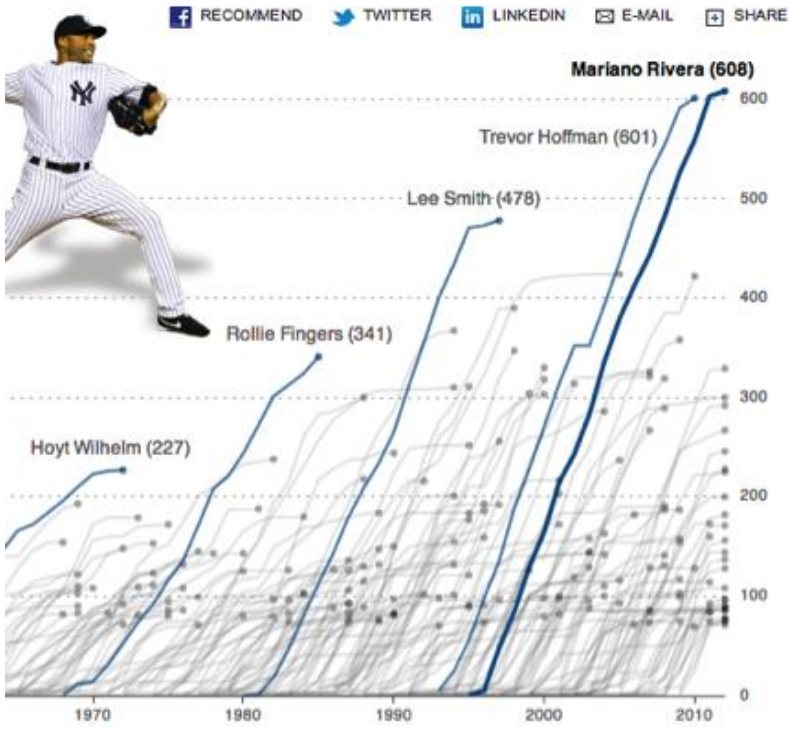
REVOLUTION ANALYTICS

MAPR

Sports Analytics as Analogy.

- Sports Teams Are Like Other Corporations.
  - Great Value Achievable With Data
  - Vast Range of Data Sources
  - Timely Analysis Amplifies Value

- And apologies if you came to learn whom to bet upon in next year's season.



REVOLUTION ANALYTICS

MAPR

Game Changing Big Data Analytics Applications

- Marketing: Clickstream & Campaign Analyses
- Digital Media: Recommendation Engines
- Social Media: Sentiment Analysis
- Retail: Purchase Prediction
- Insurance: Fraud Waste and Abuse
- Healthcare Delivery: Treatment Outcome Prediction
- Risk Analysis: Insurance Underwriting
- Manufacturing: Predictive Maintenance
- Operations: Supply Chain Optimization
- Econometrics: Market Prediction
- Marketing: Mix and Price Optimization
- Life Sciences: Pharmacogenetics
- Transportation: Asset Utilization

REVOLUTION ANALYTICS

MAPR

Polling Question #2:

- What Language or Tools Is In Use for Analytics (check all that apply)
  - R
  - SAS or SPSS
  - Python
  - Java
  - BI tools including: MSTR, Qlik, Tableau, Business Objects, Cognos
  - Salford Systems or MATLAB
  - H20, RapidMiner, KNIME or similar
  - Other data mining tools
  - Other programming languages
  - None or Don’t know

R Open Source

- **Language, Community, Collaboration**
- **Robert Gentleman & Ross Ihaka, 1993**
- **Version 1.0 released 2000**
- **2.5 Million Global Users**
- **Over 4,800 add-on "Packages"**
- **Why R?**

  R in Universities = New Talent

  Emerging Modeling/Visualization

  Lower Cost Alternative

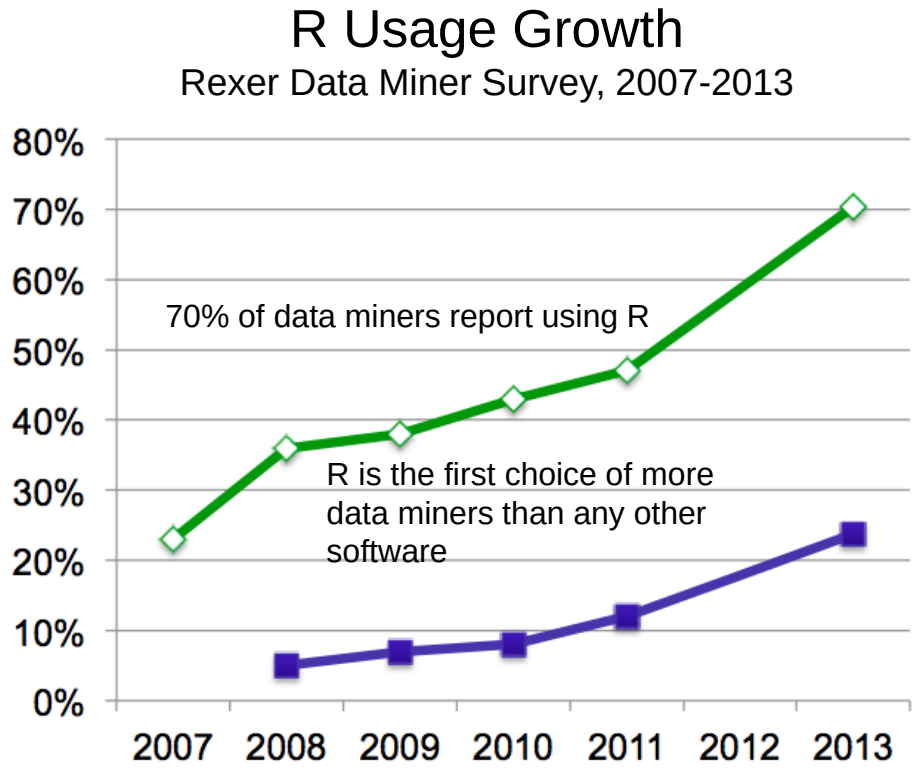
  Open Source = Flexible & Innovative

  Access to Free Packages

REVOLUTION ANALYTICS

MAPR

R is exploding in popularity & functionality



Source: www.rexeranalytics.com

Innovate with R

- Most widely used data analysis software
  - Used by 2M+ data scientists, statisticians and analysts
- Most powerful statistical programming language
  - Flexible, extensible and comprehensive for productivity
- Create beautiful and unique data visualizations
  - As seen in New York Times, Twitter and Flowing Data
- Thriving open-source community
  - Leading edge of analytics research
- Fills the talent gap
  - New graduates prefer R

White Paper

R is Hot

bit.ly/r-is-hot

REVOLUTION ANALYTICS

MAPR

Polling Question #3:

- How are you using R today? (choose one)
  - Not using R
  - Studying R now
  - Initial R project(s) underway
  - R is widely used for exploration & modeling
  - R is deployed into production

Revolution Analytics In A Nutshell

Our Vision:

- R is becoming the de-facto standard for enterprise predictive analytics

Our Mission:

- Drive enterprise adoption of R by providing enhanced R products tailored to meet enterprise challenges

REVOLUTION ANALYTICS

MAPR

Revolution Analytics Builds & Delivers:

- Software Products:
  - Stable Distributions
  - Broad Platform Support
  - Big Data Analytics in R
  - Application Integration
  - Deployment Platforms
  - Agile Development Tooling
  - Future Platform Support
- Support & Services
  - Commercial Support Programs
  - Training Programs
  - Professional Services
  - Academic Support Programs
  - IP Indemnification

REVOLUTION ANALYTICS

MAPR

Revolution R Advantages for Analytics Professionals:

- Broadly-used, scalable R language
- Large (2M+), collaborative, young R analytics community
- Largest repository of statistical & analytical algorithms
- Big data analytics capabilities
  - Scales from workstations to Hadoop
  - Transparent parallelism
  - Cross platform compatibility
  - Multi-platform architectures
- Broadens career opportunities

REVOLUTION ANALYTICS

MAPR

Revolution R Advantages for Business Executives

- Viable Alternative to Legacy Analytics Solutions
  - Predictable Time To Results
  - Simplified Licensing
  - All-Inclusive Environment
- Lower Staffing Costs
- Controllable Open Source Risks
  - Support
  - IP Infringement Protections

REVOLUTION ANALYTICS

MAPR

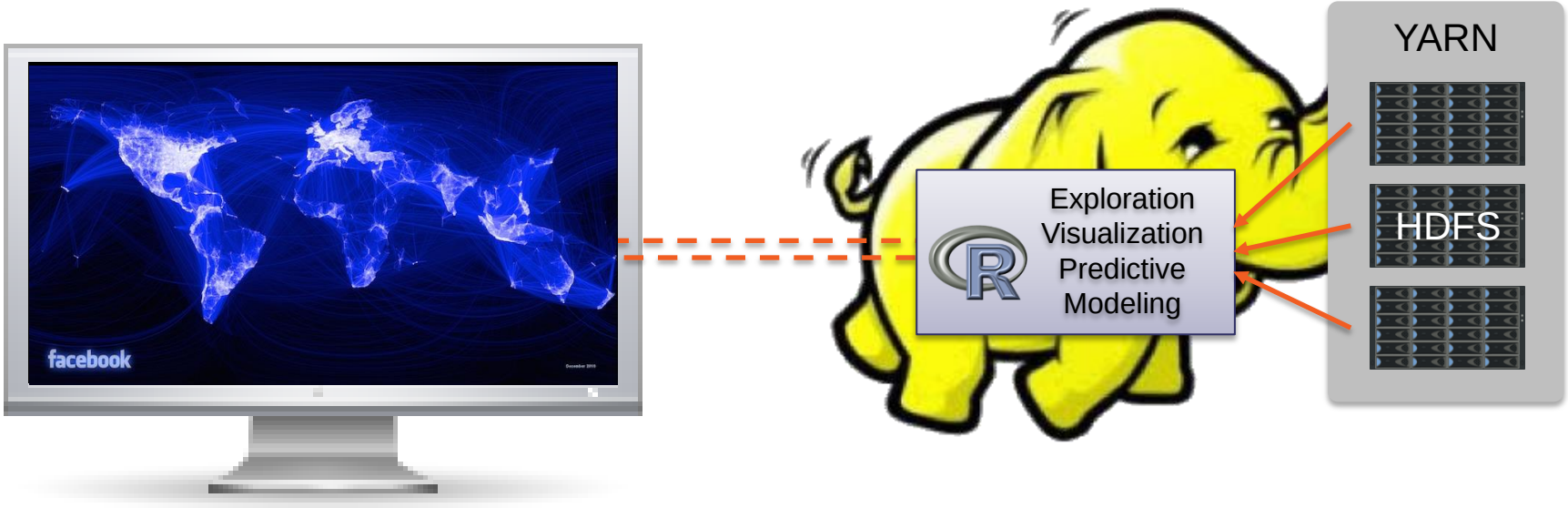
Revolution R Advantages for IT Organizations

- Consistency Across Platforms Avoids Sprawl
- Support for Workstations, Servers, Hadoop, EDWs and Grids
- Heterogeneous Architecture Capabilities
- Integrates With Major BI & Application Tools
- Streamline Model Deployment
- Run Complex Analytics in the “Data Lake”
- Be a “Good Citizen” in shared systems
- Commercial Support Reduces Project Risks
- Quick Start Programs Accelerate Results
- Platform Continuity Future-Proofs Architectures

REVOLUTION ANALYTICS

MAPR

Revolution R Enterprise:
Predictive Analytics Across Huge Data in Hadoop



REVOLUTION ANALYTICS

MAPR

Polling Question #4:

- Stage of Hadoop Adoption? (choose one)
  - No Need
  - Studying
  - Setting-Up Hadoop
  - Experimenting with Hadoop
  - Deploying Hadoop Now
  - Hadoop in Production

REVOLUTION ANALYTICS

MAPR

Introducing:

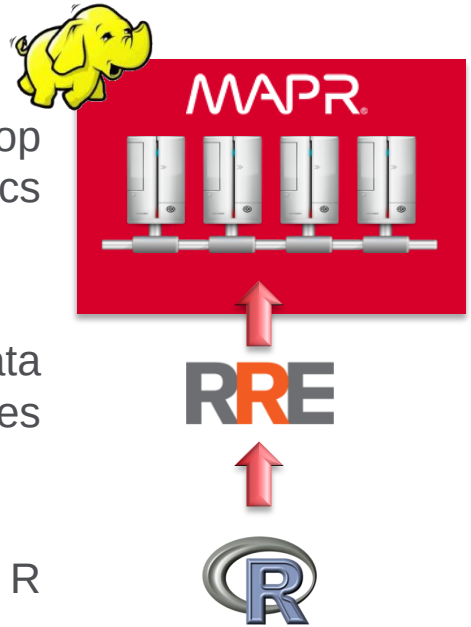
Vineet Sharma
Director, Partner Marketing
MapR Technologies

MapR + Revolution Leverages MapR As A Scalable Enterprise R Engine.

Broad Adoption of Hadoop for Big Data Analytics

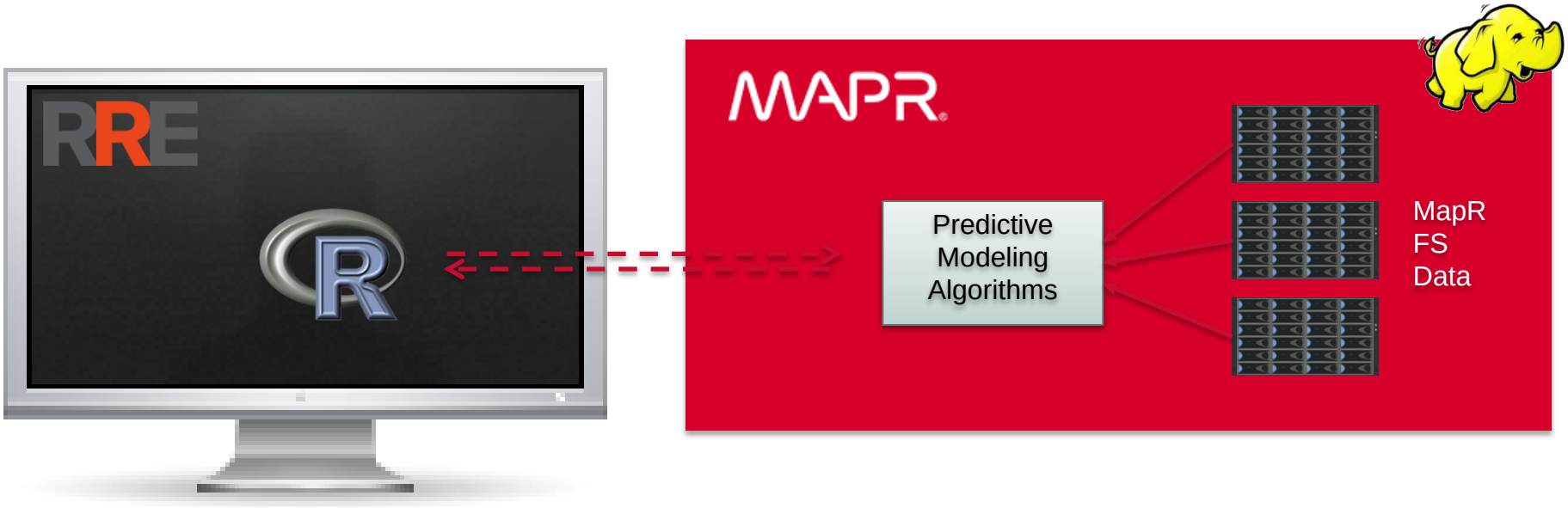
MapR Enterprise Data Platform Capabilities

Rapid Adoption of R



- Plus:
  - Run RRE Analytics In MapR Hadoop Without Change
  - Eliminate Need To Design Parallel Software or “Think In MapReduce”
  - Leverage All Revolution R Enterprise Pre-Parallelized Algorithms
  - Enable Users To Build Custom Apps That Leverage Hadoop’s Parallelism
  - Slash Data Movement by Analyzing Data Inside the MapR Data Platform
  - Expand Deployment and Integration Options

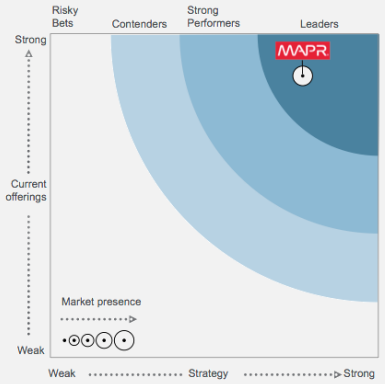
Desktop Users with Analytical Access to Huge Data in Hadoop



MapR: Best Solution for Customer Success

Top Ranked

FORRESTER®



Exponential Growth

- **>2x** annual bookings
- **90%** software licenses
- **80%** of accounts expand 3X
- **<1%** lifetime churn
- **>$1B** in incremental revenue generated by 1 customer

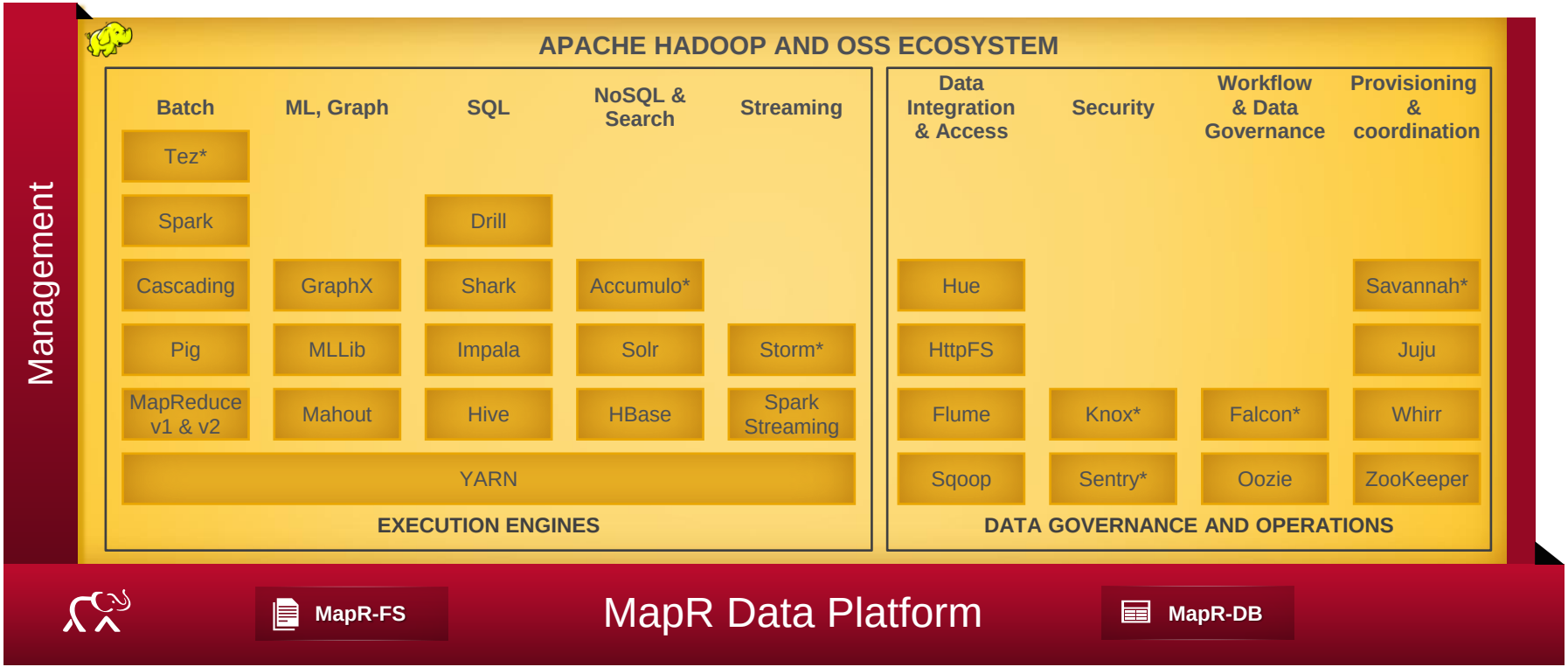
500+ Customers

LIVE NATION ENTERTAINMENT · hp · TransUnion · SAMSUNG · AADHAAR · cisco · rubicon PROJECT · comScore

Premier Investors

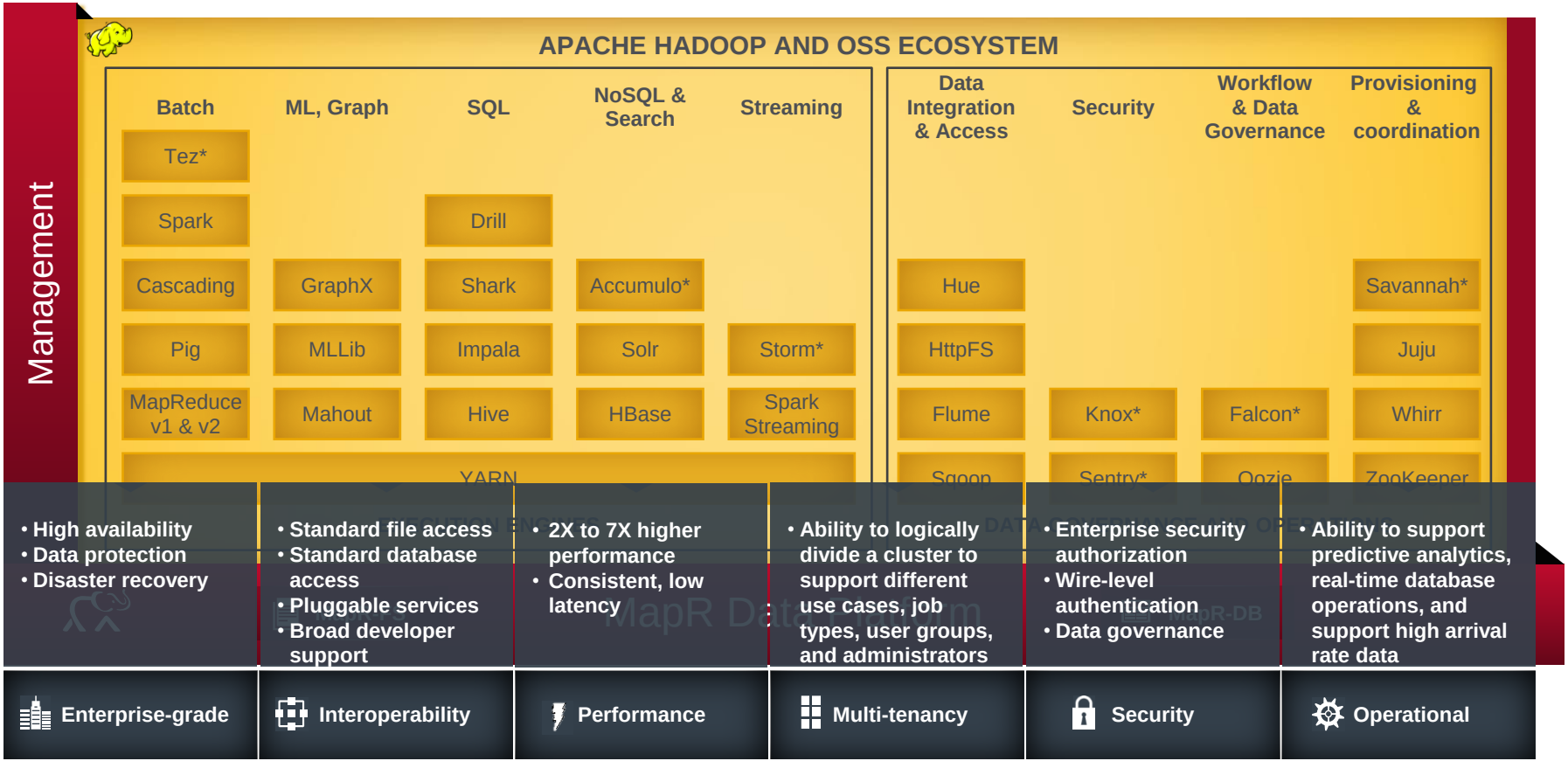
Google capital · QUALCOMM VENTURES · LIGHTSPEED VENTURE PARTNERS · MAYFIELD FUND · NEA · Redpoint

The Power of the Open Source Community



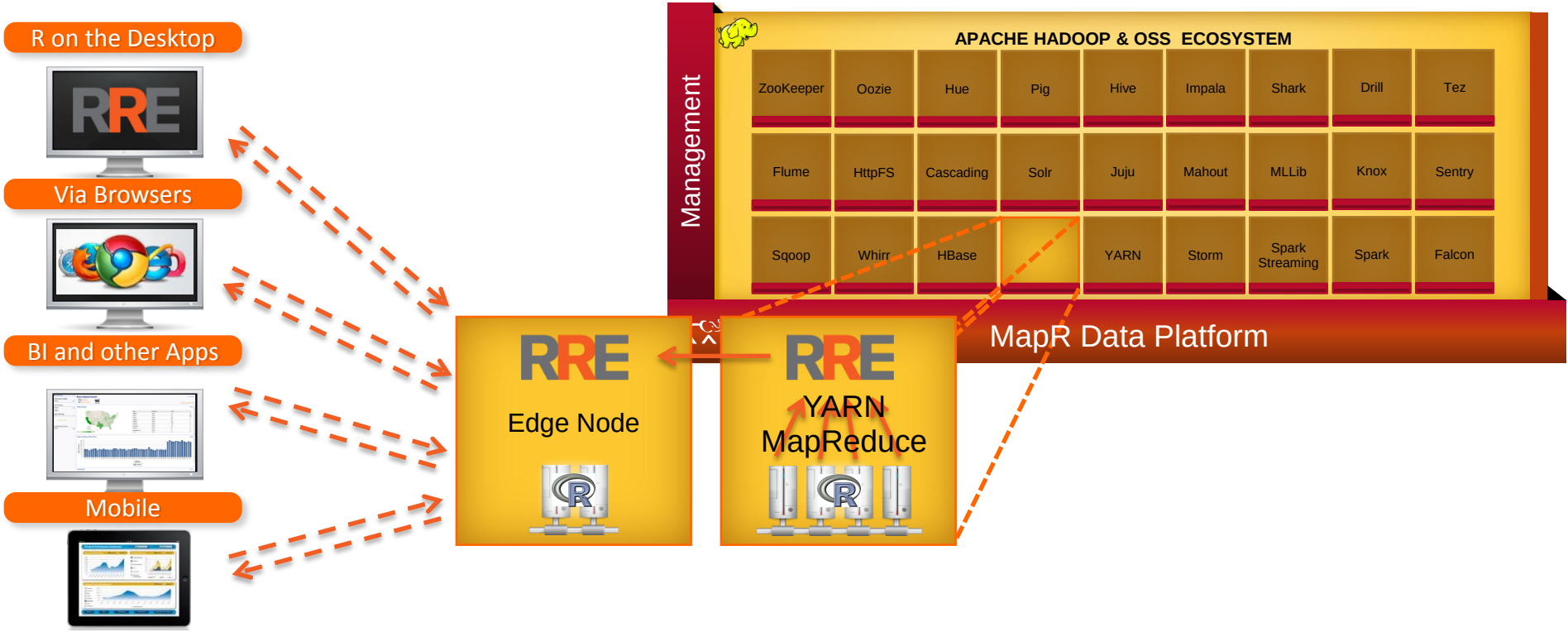
* Certification/support planned

MapR Distribution for Hadoop



* Certification/support planned

Revolution R Enterprise and MapR Hadoop



REVOLUTION ANALYTICS

MAPR

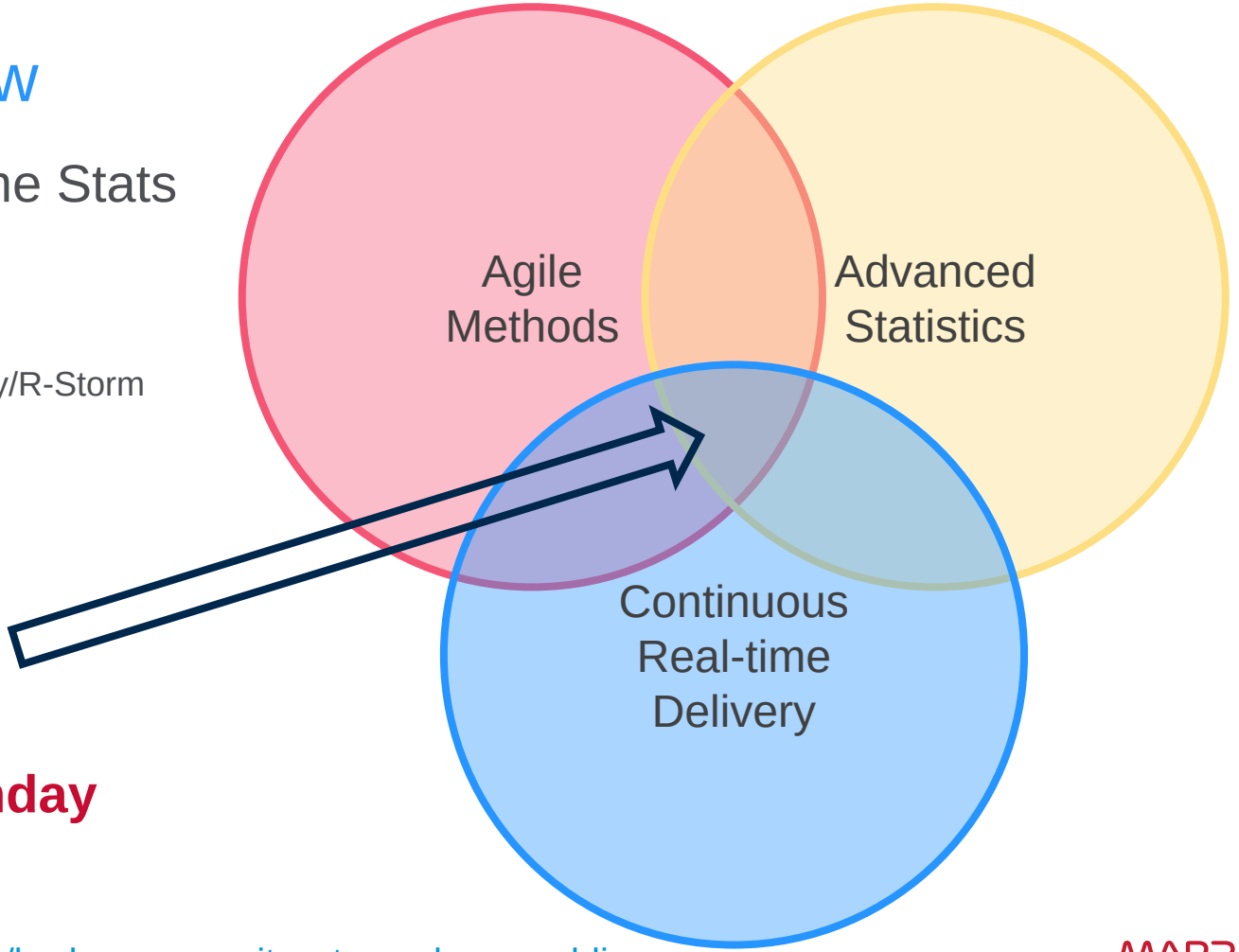
Introducing:

Allen Day
Principal Data Scientist
MapR Technologies
@allenday

Talk Overview

- Agile Real-time Stats
- R + Storm

  github.com/allenday/R-Storm
- DEMO
- How to do it?
- Q & A **@allenday**



github.com/allenday/hadoop-summit-r-storm-demo-public

Architecting R into the Storm Application Development Process

Allen Day, PhD [@allenday]

December, 2014

Quick intro

- Allen Day, Principal Data Scientist [@allenday]

  7yr Hadoop dev, 12yr R dev/author

  PhD, Human Genetics, UCLA Medicine

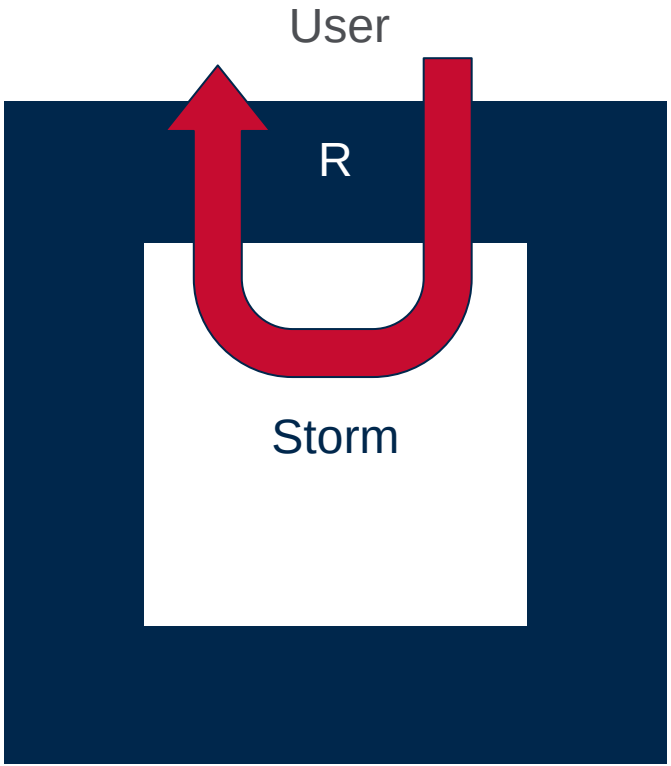
What's Storm? What's R?

- What's Storm?
  - Processes a data stream. Akin to UNIX pipe + `tee` & `merge` commands
  - Runs on a cluster. Fault-tolerant and designed to scale out
  - Used for: real-time analytics & machine learning

- What's R?
  - Programming language with advanced statistics libraries
  - Does not scale out. Can scale up
  - Used for: prototyping, data modeling, visualization

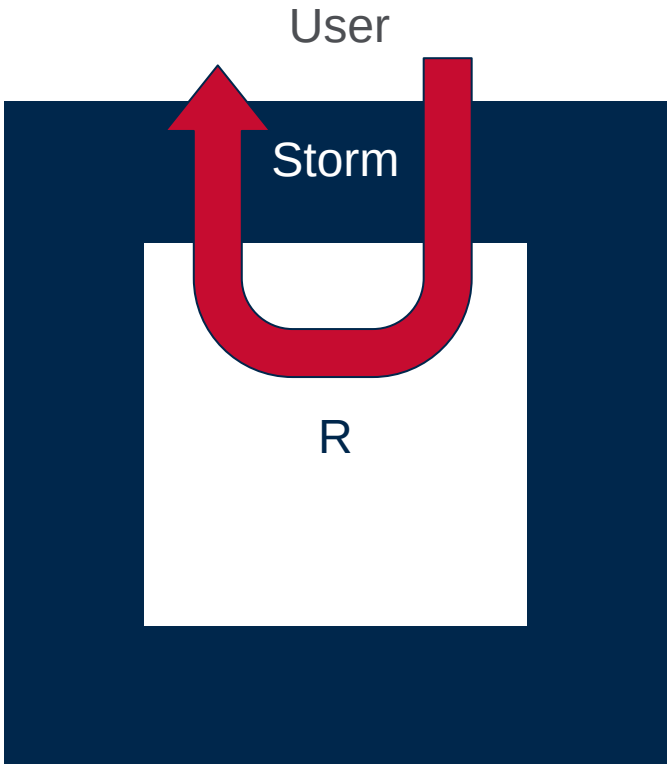
How to combine these?

R outside, Storm inside: not practical. Why?



- Model-building and QA is done on data snapshots
- However, R => Hadoop is realistic. Key difference: referenced data can be static
  - Use MapR snapshots for dev and QA
  - See also: RHIPE (Purdue) and RHadoop (RevolutionAnalytics)

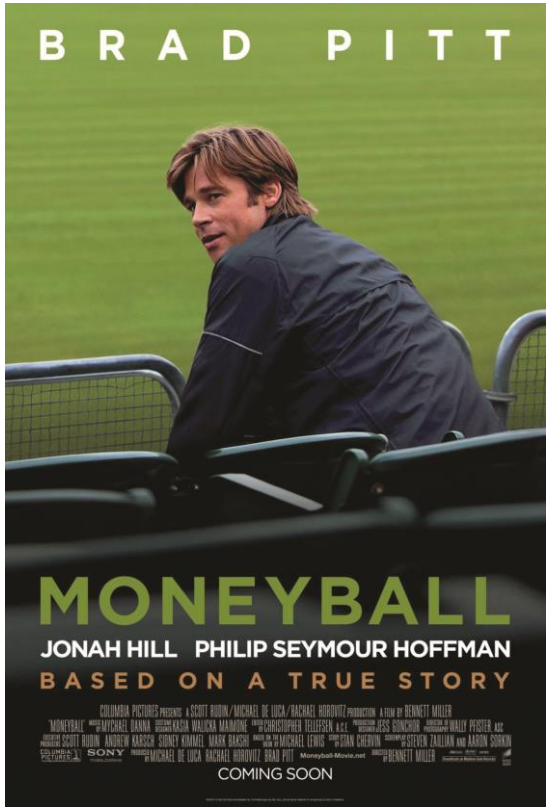
Storm outside, R inside: a good fit



- Enables separation of concerns
  - Independently manage modeling, ops timelines, and version control
  - Integrate as needed
- Enables role specialization
  - R built-ins allow faster iteration and more concise stats-type code
  - Do DevOps with specific SW engineering tech, e.g. Java

Q: Who really likes statistics?
A: Baseball fans
A: Team Managers = Portfolio Managers

Famous Vintage Data



Oakland Athletics
2002 Season

20 consecutive wins – the current record

Goal: Detect “Moneyball” 2002 Winning Streak

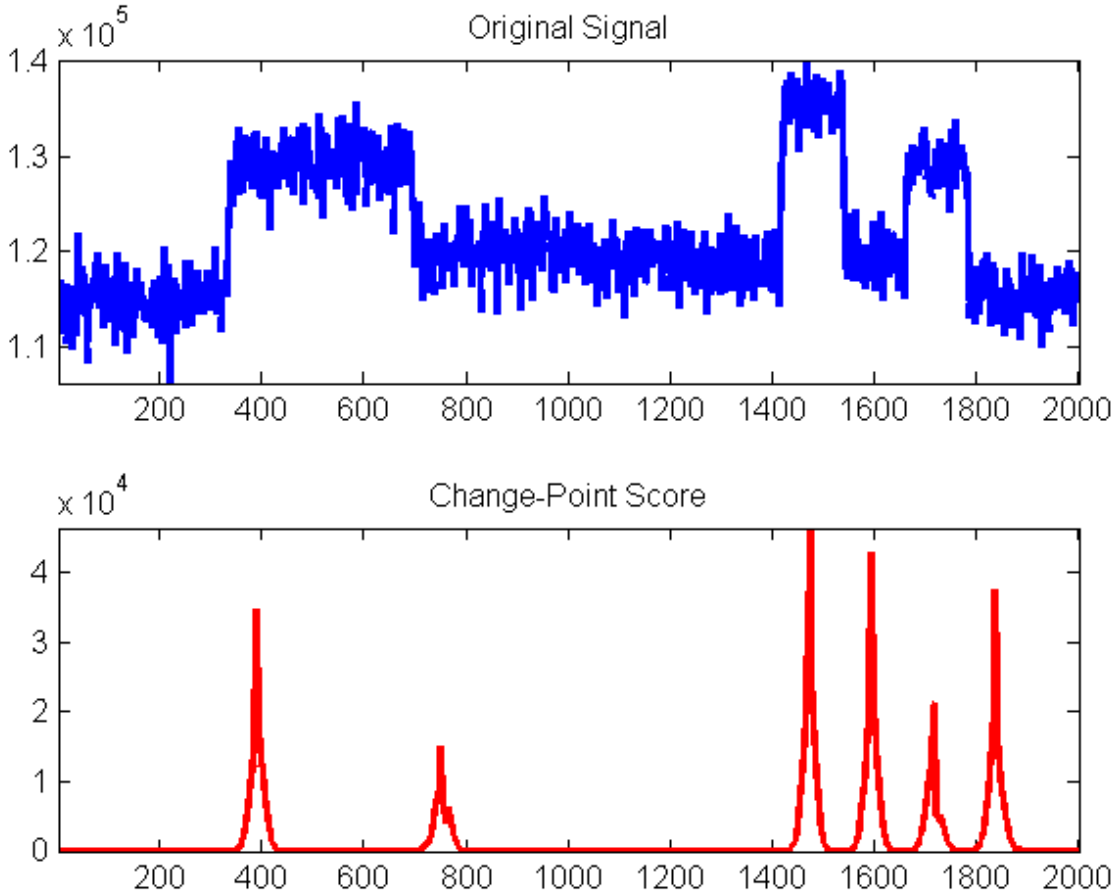
Methods:
Change Point Detection

Find natural breakpoints in a time-series set of data points

R packages implement this:

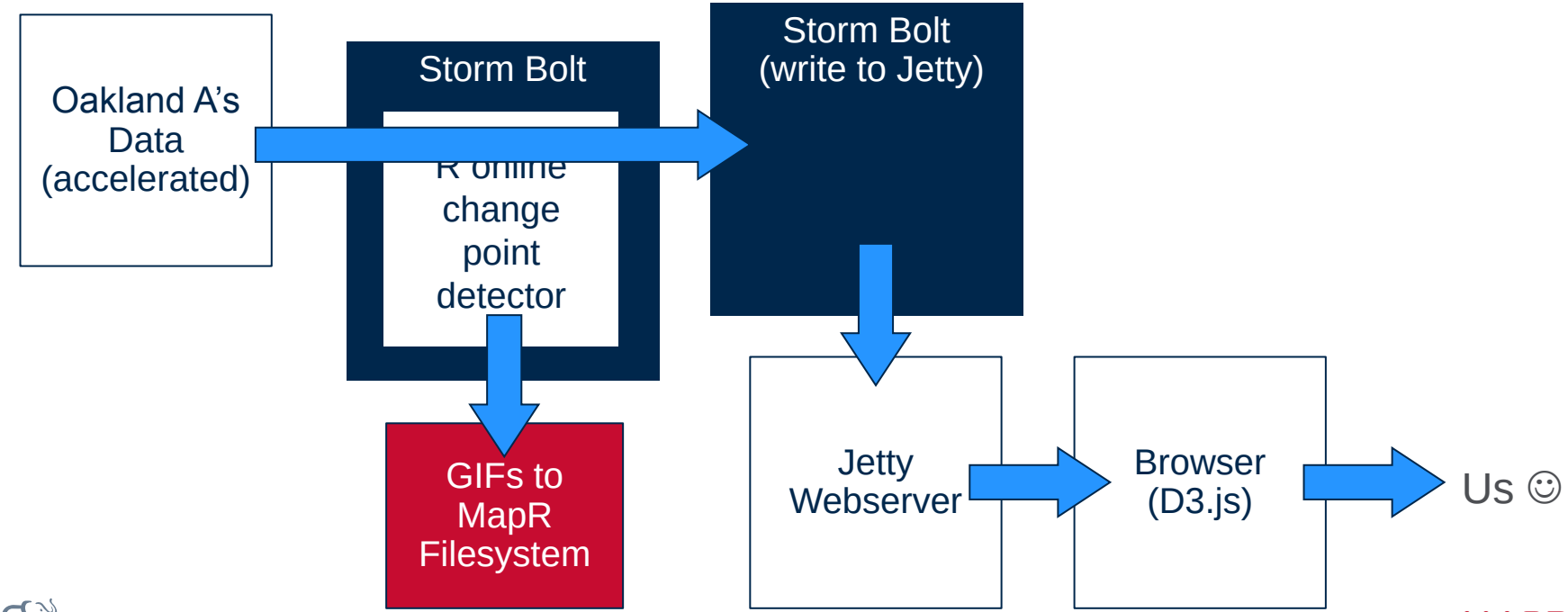
`changepoint`: more sensitve, but not streaming

`bcp`: streaming, but less sensitive

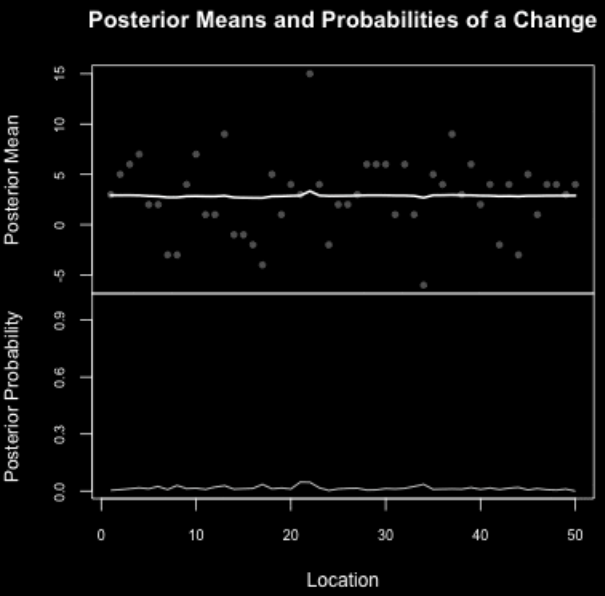


Methods: R+Storm Demo Architecture

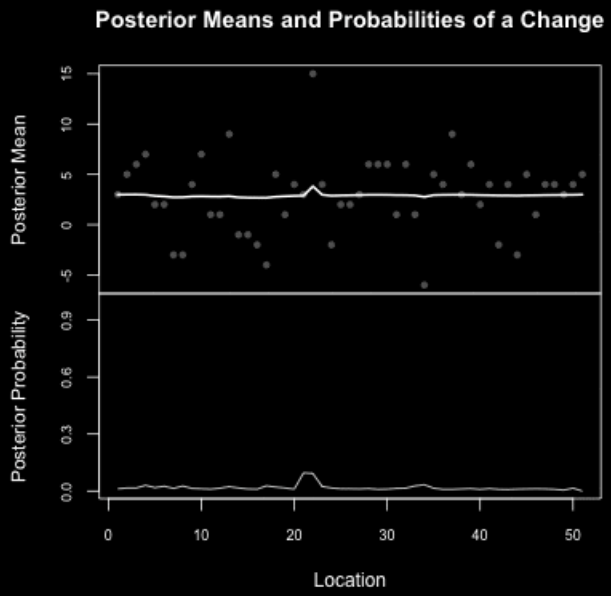
github.com/allenday/hadoop-summit-r-storm-demo-public



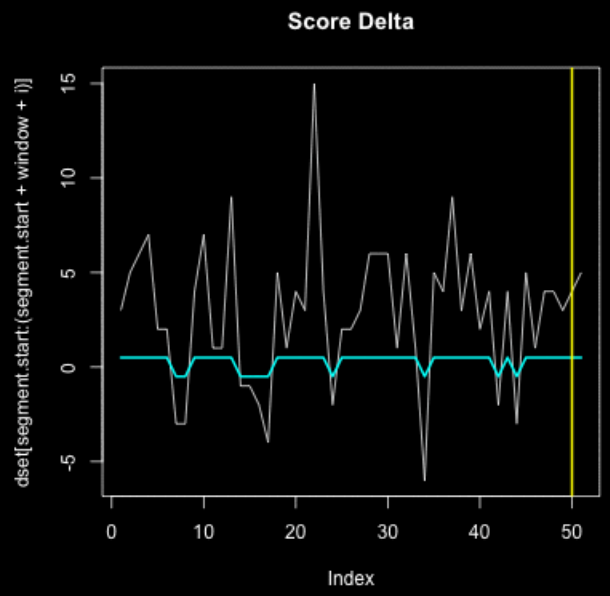
Demo



50-game sliding window/buffer to detect change points



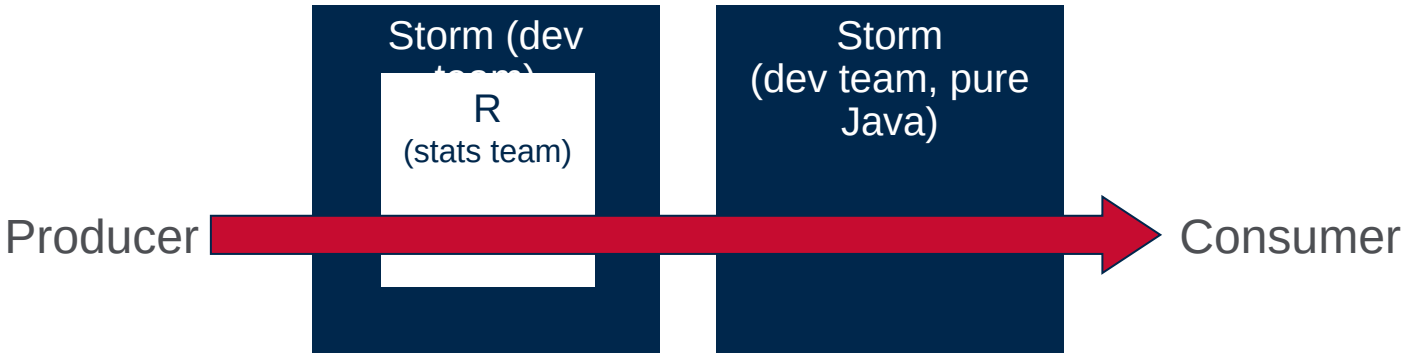
Cumulative history with detected break points



Raw data (score difference between A's and opponent)

Methods Details: How it's done

- Uses R-Storm binding github.com/allenday/R-Storm
  - `Storm` package on CRAN cran.r-project.org/web/packages/Storm



Methods Details: Easy integration

R: lambda function

```
storm = Storm$new();
storm$lambda = function(s) {
    t = s$tuple;
    t$output = vector(length=1); t$output[1] = “tada!”
    s$emit(t)
}
```

Storm: extend ShellBolt

```
public static class MyRBolt extends ShellBolt implements
IRichBolt {
    public RBolt() {
        super("Rscript", ”my.R");
    }
}
```

Results

- Change points are identified, but none for winning streak
  - Not using score difference, anyway

Discussion

- Time to integrate with the modeling team!
  - Send **@kunpognr** or **@allenday** a pull request on GitHub

- Applicable to many other use cases, e.g.
  - Security (fraud detection, intrusion detection)
  - Marketing (intent to purchase / social media streams)
  - Customer Support (help desk voice calls)

Polling Question #5

- How important will Real-Time analytical apps become? (choose one)
  - Uncertain
  - Not important
  - Necessary
  - Critical

REVOLUTION ANALYTICS

MAPR

Real-Time and Internet of Things: Foundation of a Compelling Trend for 2015

- Big Data Analytics Meets The Internet of Things
  - Transactions +
  - Human Behavior +
  - Internet of Things: Sensors
- … and extracting value using
  - Traditional Statistics
  - Visualization
  - Machine Learning
- … plus adaptability
  - Real-Time –Agile Modeling & Fast Model Execution
  - Production Capable, Stable and Secure
  - Rapidly-Evolving Data Science

REVOLUTION ANALYTICS

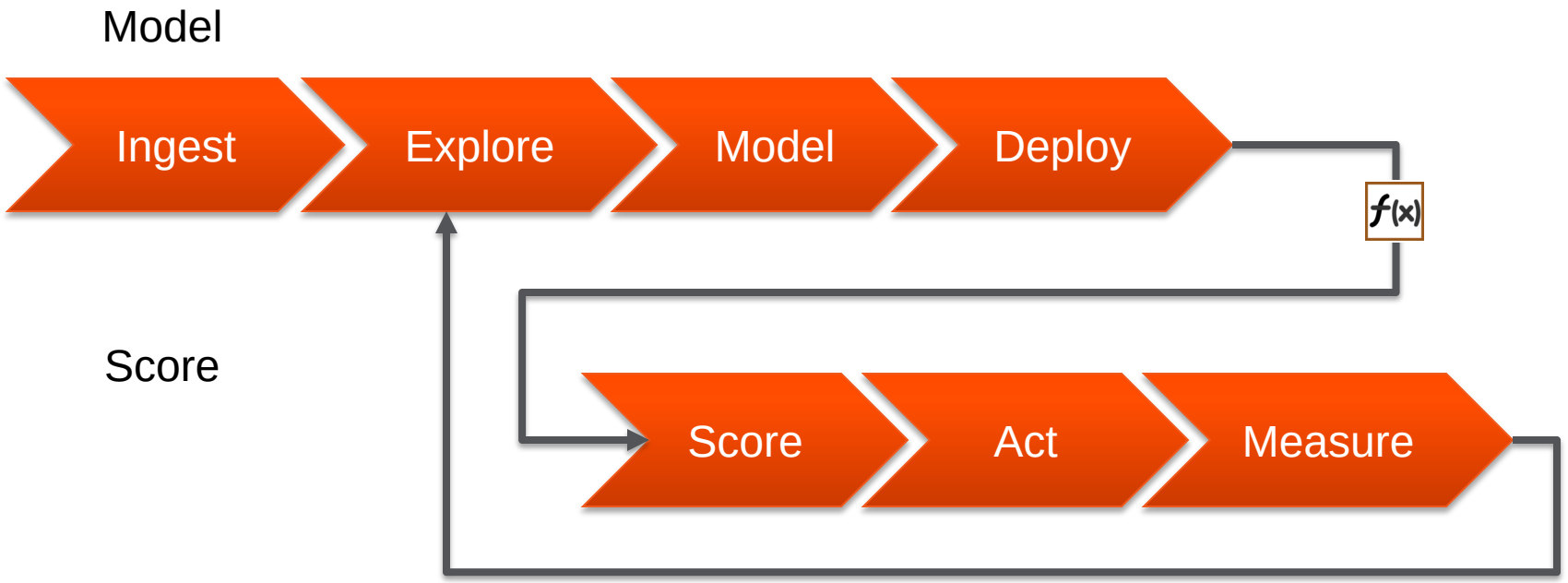
MAPR

Where Does Real Time Impact The Analytical Lifecycle?

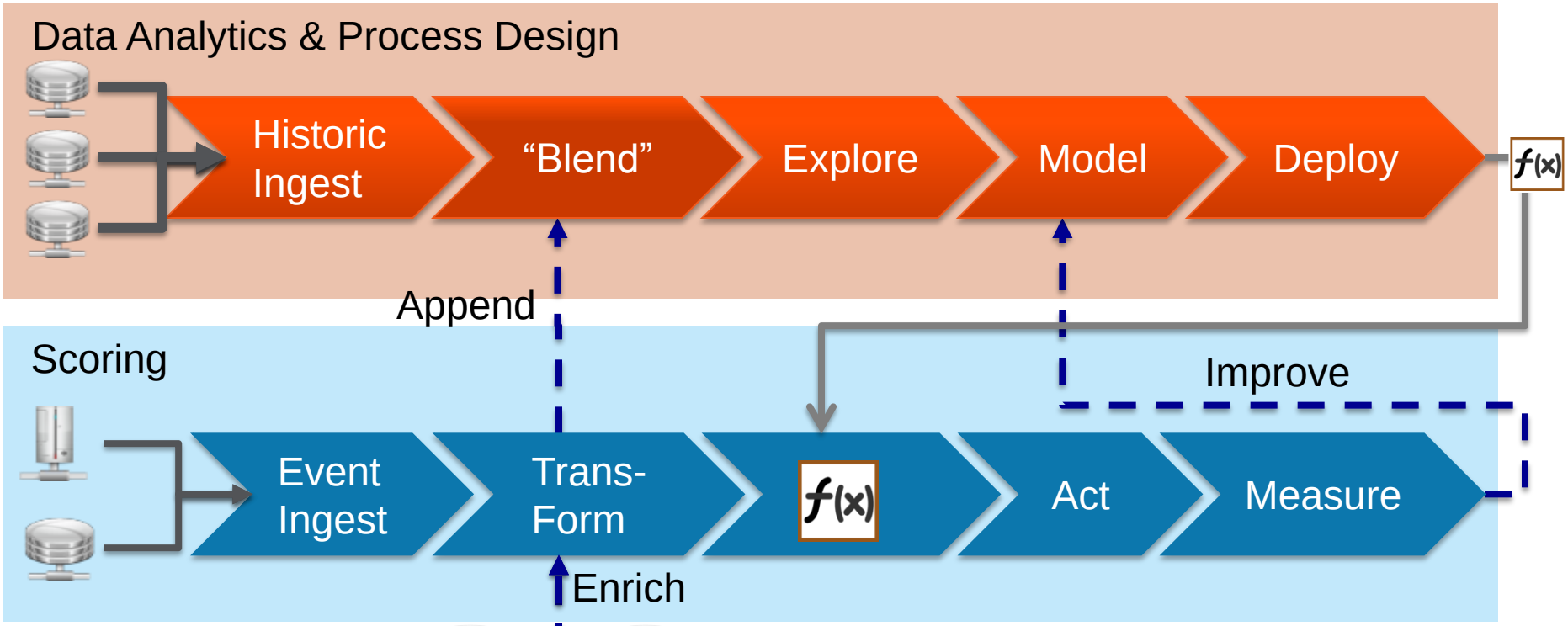
- Data Engineering
  - Collection and Ingest
  - “Blending”
- Modeling
  - Aggregation, Segmentation & Exploration
  - Model Development & Optimization
  - Testing & Validation
- Operationalization
  - Deployment & Scoring
  - Delivery
  - Monitoring & Evaluation

REVOLUTION ANALYTICS

Typical Analytical Lifecycle



More Complex Event Driven Analytical Cycle



REVOLUTION ANALYTICS

MAPR

Real-Time Analytics Best Practices

- Develop a Common Lexicon for Real-Time
- Discriminate Between Needs of Each Stage in Lifecycle
  - Data Ingest & Manipulation and Enrichment
  - Data Source / Repository Integration Needs
  - Processes that “Fill the Lake”
  - Process that “Act on the Stream”
  - Vastly different computationally
  - Big differences in data ingest volume & latency
- Start with Tractable Goals
  - Anticipate Growing Requirements: Microbatched >> Interactive >> Autonomy
- Build for today, Architect for tomorrow

REVOLUTION ANALYTICS

MAPR

Real-Time Realities

- Plan for Diverse Needs
  - Real-Time Score Retrieval, Scoring, Modeling
  - Wide-Ranging Performance – Microbatch – Interactive - Autonomous
- Fragmentation
  - Data Delivery Systems Pre-Exist
  - Will Vary Widely by Vertical Market
  - Competing Proprietary Solutions
- Growing Demand
  - Numerous high-value targets
  - “The next step”: Put big data analytics to work

What's Needed

- Real-Time Performance… plus…
- Agility
  - Deployment models
  - Organization
  - Infrastructure
  - Analytics
- Manageable Costs
  - Hadoop
  - Open Source R
- Production Platform(s)
  - Proven
  - Performant

REVOLUTION ANALYTICS

MAPR

Next Steps…

Resources:

- R foundation URL: www.r-project.org
- Download Revolution R: http://mran.revolutionanalytics.com/download/
- Learn about Apache Storm: https://storm.apache.org/
- R-Storm bindings: github.com/allenday/R-Storm
- Storm package on CRAN: cran.r-project.org/web/packages/Storm

REVOLUTION ANALYTICS

- www.revolutionanalytcs.com
- Whitepaper: Revolution R for Hadoop:
  - http://www.revolutionanalytics.com/whitepaper/delivering-value-big-data-revolution-r-enterprise-and-hadoop
  - …or http://bit.ly/1ua43bu

MAPR

- www.maprtech.com

REVOLUTION ANALYTICS

Thank you.

www.revolutionanalytics.com
1.855.GET.REVO
Twitter: @RevolutionR