Making Big Data Analytics accessible via the R environment

Vaishnavi Sashikanth (vaishnavi.sashikanth@oracle.com)

Vice President, Development, Database Technologies Division

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions.
The development, release, and timing of any features or functionality described for Oracle’s products remain at the sole discretion of Oracle.

Agenda

- What is Big Data Analytics?
- Oracle Big Data Analytics Architecture & Components
- Open Source R integration

Big Data / Deep Analytics

- *Application of numerical, predictive and statistical techniques on big data*

Financial Services

- Credit risk analysis
- Cross-LOB up-selling
- Fraud detection
- Retail banking personalization
- “Best customer” prediction & profiling

Retail

- Real-time shopping cart recommendations
- Customer segmentation
- Customer profiling
- Market basket analysis
- Fraud detection

Media & Entertainment

- Online ad placement
- Cable TV: option bundling
- Gaming: Targeting “right customer w/ “right product”
- Gambling: Fraud and anomaly detection

Telecommunications

- Churn prevention
- Social network analysis
- Network monitoring
- Win-back analysis
- Fraud analysis

Public Sector

- Healthcare Fraud prevention
- Infrastructure maintenance
- Constituent Sentiment
- Threat Identification
- Healthcare improvement

Manufacturing

- Warranty analysis
- Quality improvement
- Product & process design and improvement

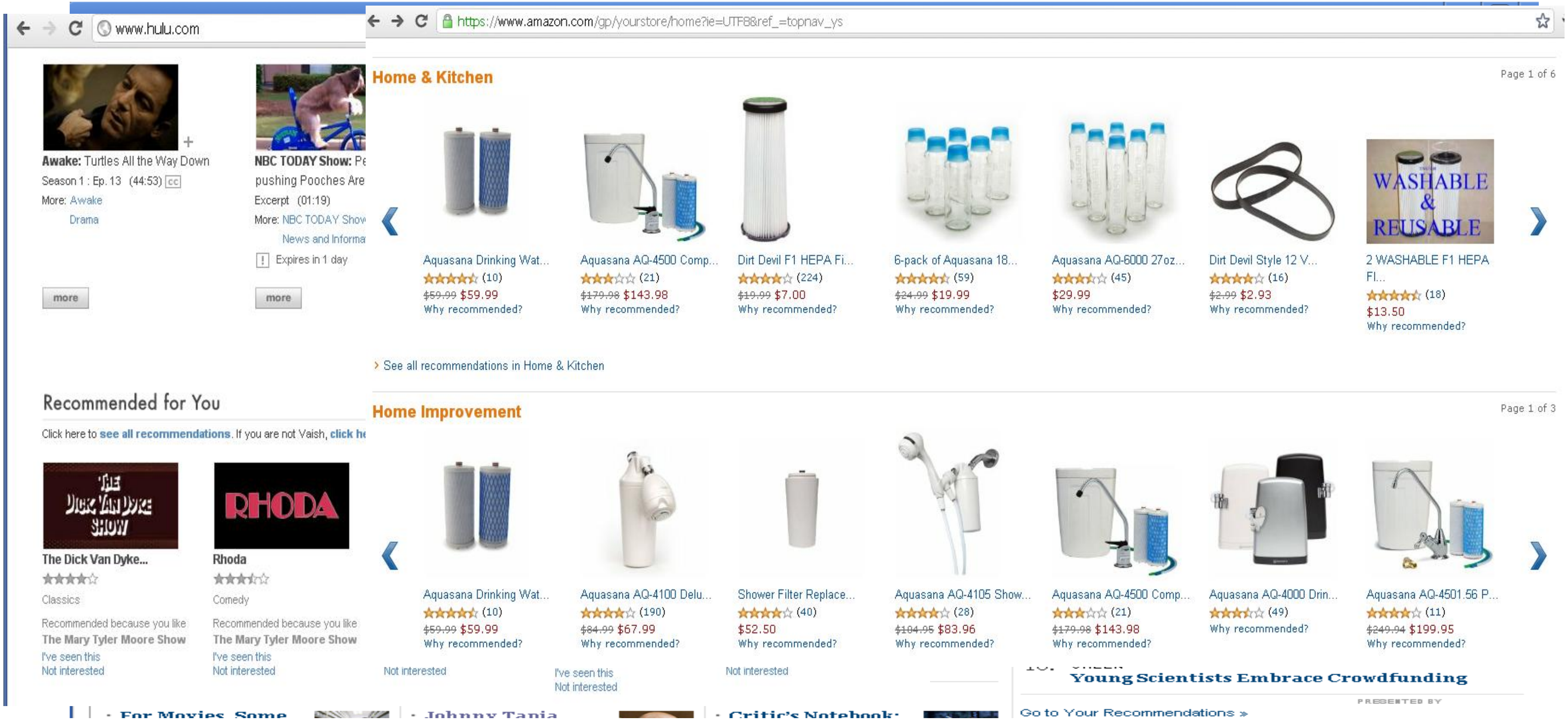
Transportation and Logistics

- Anticipate bottlenecks
- Proactive resource planning
- Improved preventative maintenance strategies

Utilities

- Customer loyalty management
- Fraud detection
- Product bundling
- Improved operations efficiencies

Analytics driving the bottom line..



www.hulu.com

Awake: Turtles All the Way Down
Season 1 : Ep. 13 (44:53) CC
More: Awake
Drama

more

NBC TODAY Show: Pe
pushing Pooches Are
Excerpt (01:19)
More: NBC TODAY Show
News and Informa
Expires in 1 day

more

Recommended for You

Click here to **see all recommendations**. If you are not Vaish, **click he**

The Dick Van Dyke...
Classics
Recommended because you like **The Mary Tyler Moore Show**
I've seen this
Not interested

Rhoda
Comedy
Recommended because you like **The Mary Tyler Moore Show**
I've seen this
Not interested

https://www.amazon.com/gp/yourstore/home?ie=UTF8&ref_=topnav_ys

Home & Kitchen

Page 1 of 6

Aquasana Drinking Wat... (10) ~~$59.99~~ $59.99 Why recommended?
Aquasana AQ-4500 Comp... (21) ~~$179.98~~ $143.98 Why recommended?
Dirt Devil F1 HEPA Fi... (224) ~~$19.99~~ $7.00 Why recommended?
6-pack of Aquasana 18... (59) ~~$24.99~~ $19.99 Why recommended?
Aquasana AQ-6000 27oz... (45) $29.99 Why recommended?
Dirt Devil Style 12 V... (16) ~~$2.99~~ $2.93 Why recommended?
2 WASHABLE F1 HEPA Fl... (18) $13.50 Why recommended?

› See all recommendations in Home & Kitchen

Home Improvement

Page 1 of 3

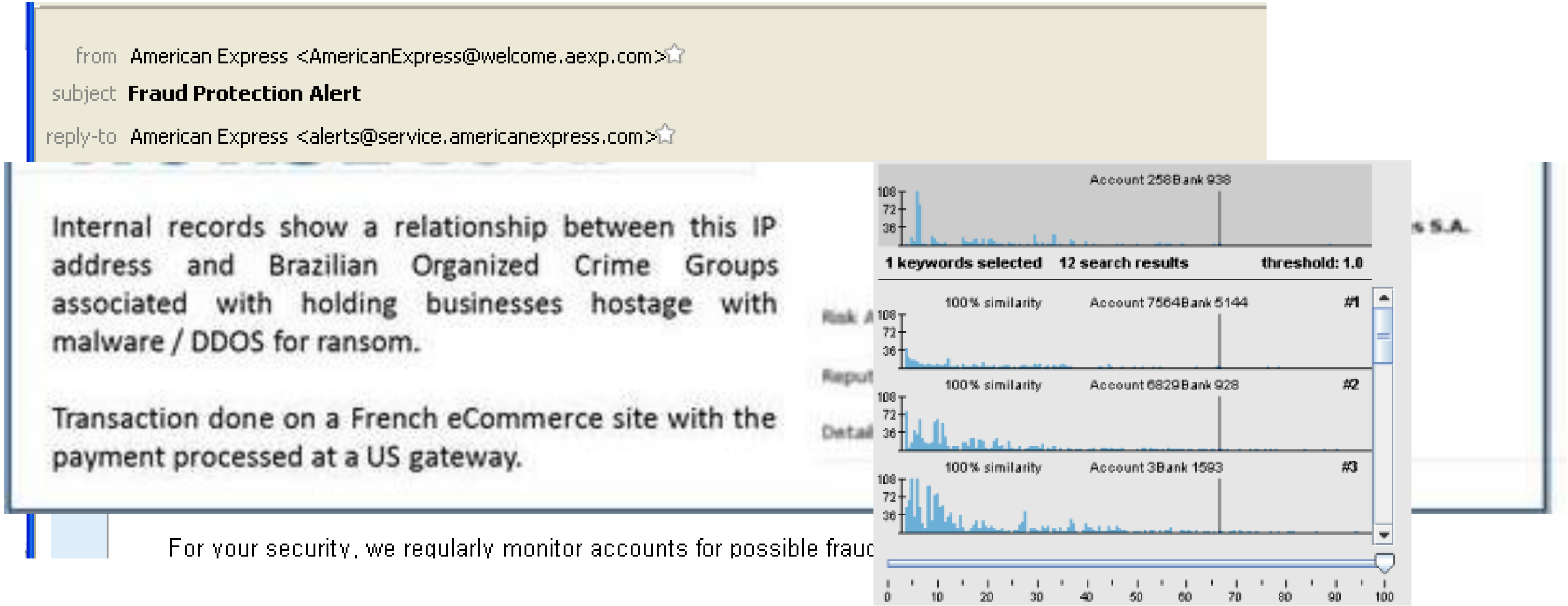
Aquasana Drinking Wat... (10) ~~$59.99~~ $59.99 Why recommended?
Aquasana AQ-4100 Delu... (190) ~~$84.99~~ $67.99 Why recommended?
Shower Filter Replace... (40) $52.50 Why recommended?
Aquasana AQ-4105 Show... (28) ~~$104.95~~ $83.96 Why recommended?
Aquasana AQ-4500 Comp... (21) ~~$179.98~~ $143.98 Why recommended?
Aquasana AQ-4000 Drin... (49) Why recommended?
Aquasana AQ-4501.56 P... (11) ~~$249.94~~ $199.95 Why recommended?

Not interested
I've seen this
Not interested
Not interested

Young Scientists Embrace Crowdfunding
PRESENTED BY
Go to Your Recommendations »

For Movies, Some
Johnny Tapia
Critic's Notebook:

Analytics minimizing bad debt..



from American Express <AmericanExpress@welcome.aexp.com>

subject **Fraud Protection Alert**

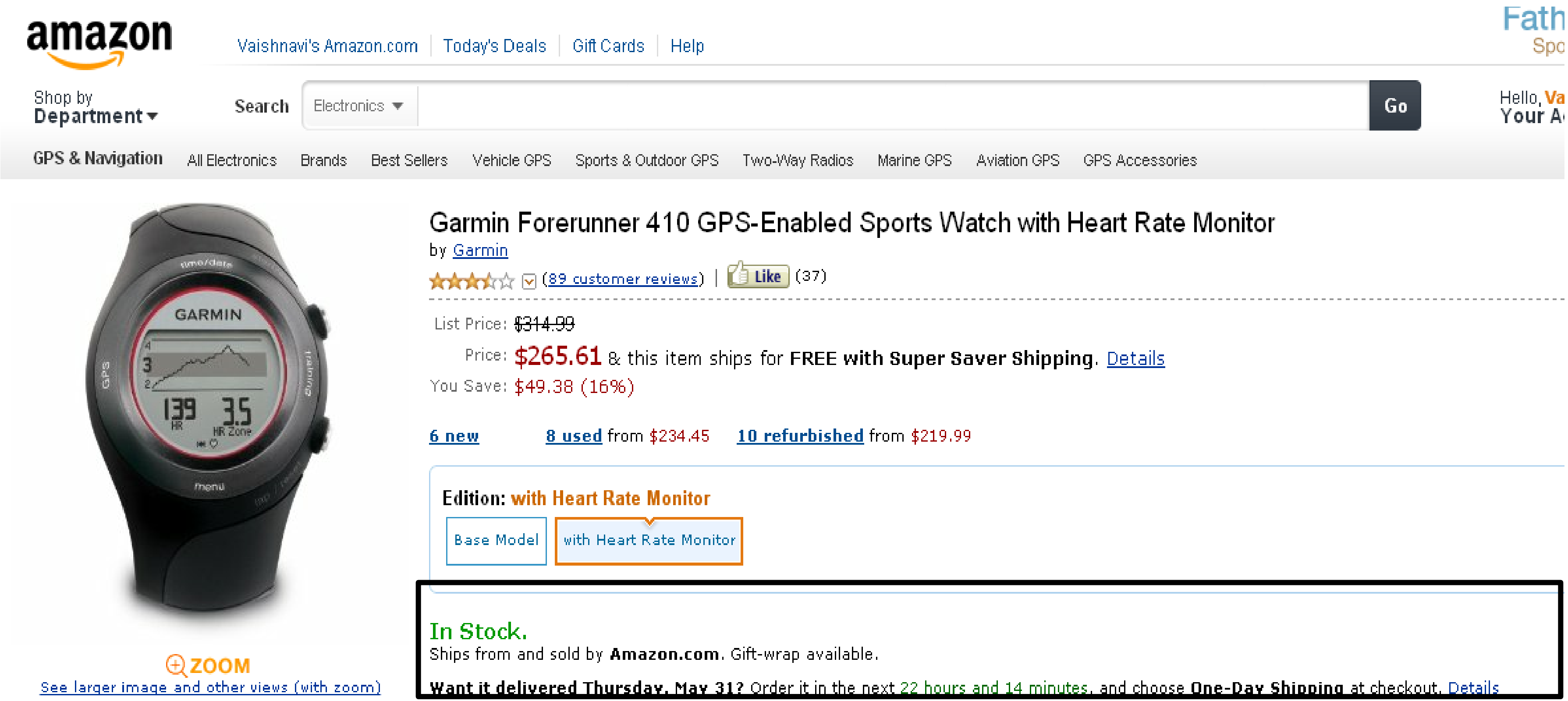
reply-to American Express <alerts@service.americanexpress.com>

Internal records show a relationship between this IP address and Brazilian Organized Crime Groups associated with holding businesses hostage with malware / DDOS for ransom.

Transaction done on a French eCommerce site with the payment processed at a US gateway.

For your security, we regularly monitor accounts for possible frauc

Analytics permeating business operations..



Garmin Forerunner 410 GPS-Enabled Sports Watch with Heart Rate Monitor

by Garmin

★★★½☆ (89 customer reviews) | Like (37)

List Price: ~~$314.99~~

Price: **$265.61** & this item ships for FREE **with Super Saver Shipping**. Details

You Save: $49.38 (16%)

6 new **8 used** from $234.45 **10 refurbished** from $219.99

Edition: with Heart Rate Monitor

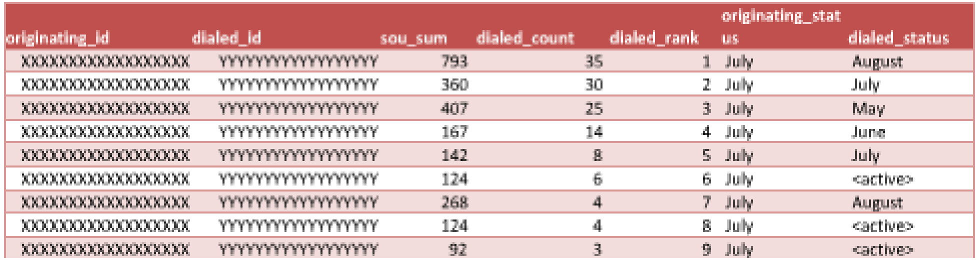
Base Model | with Heart Rate Monitor

In Stock.

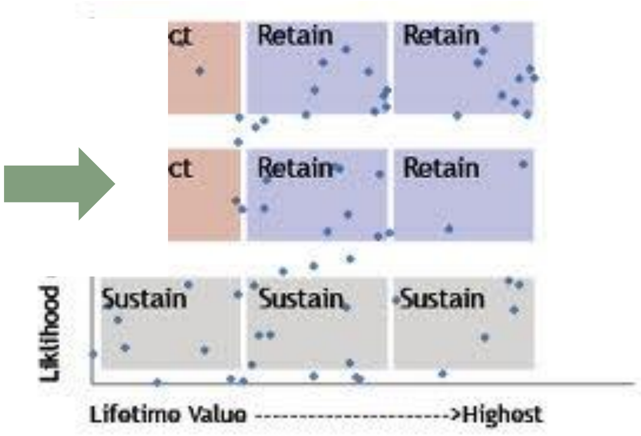
Ships from and sold by **Amazon.com**. Gift-wrap available.

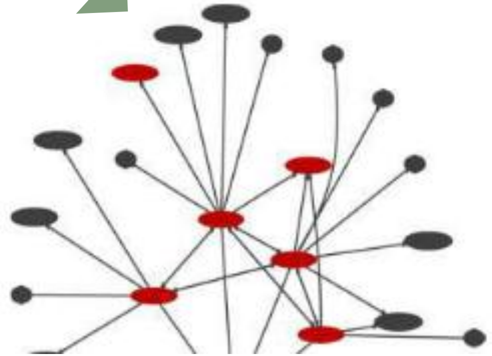
Want it delivered Thursday, May 31? Order it in the next 22 hours and 14 minutes, and choose **One-Day Shipping** at checkout. Details

Analytics preserving high value customers



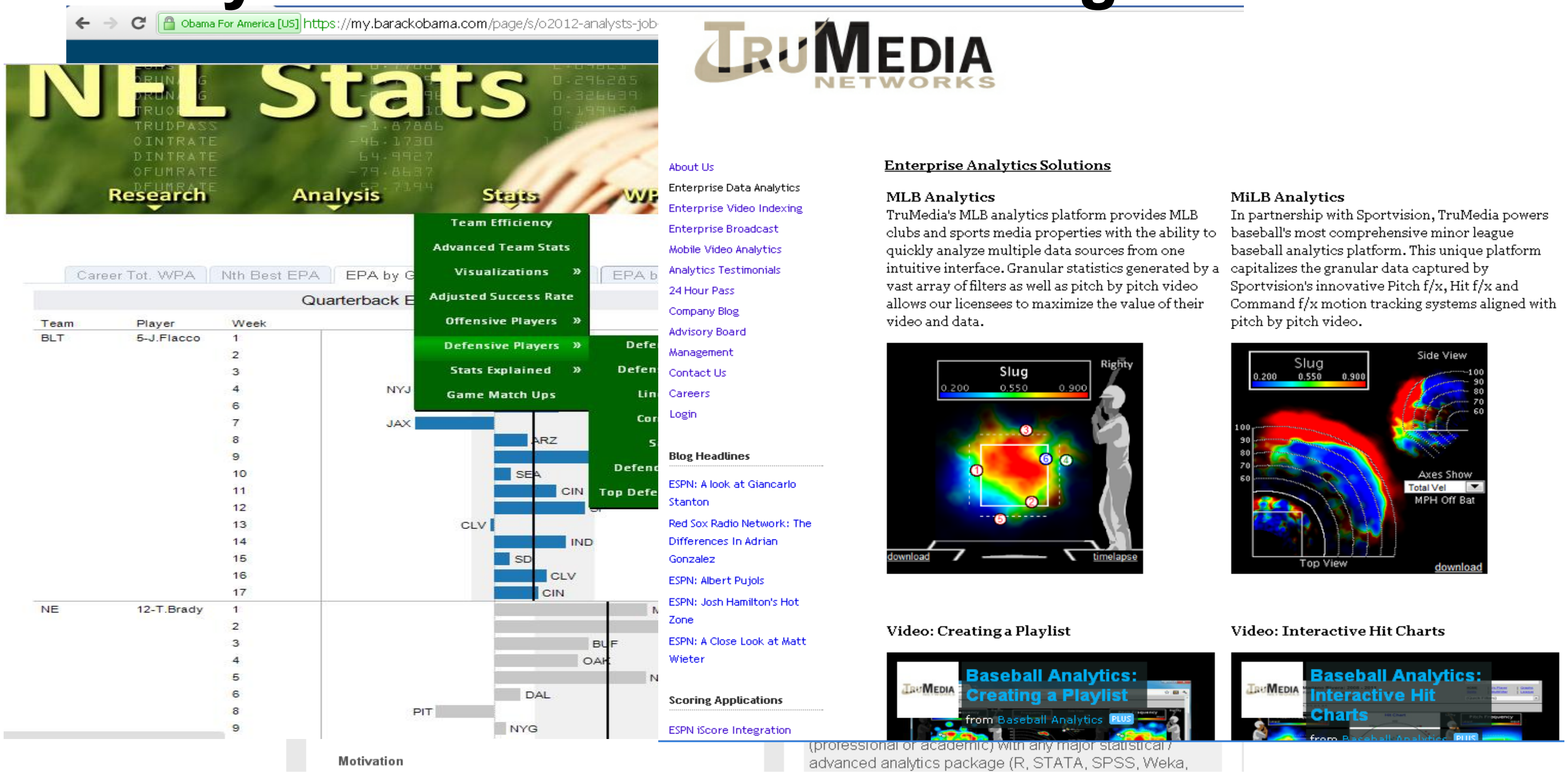
originating_id	dialed_id	sou_sum	dialed_count	dialed_rank	originating_status	dialed_status
XXXXXXXXXXXXXXXXXX	YYYYYYYYYYYYYYYYYY	793	35	1	July	August
XXXXXXXXXXXXXXXXXX	YYYYYYYYYYYYYYYYYY	360	30	2	July	July
XXXXXXXXXXXXXXXXXX	YYYYYYYYYYYYYYYYYY	407	25	3	July	May
XXXXXXXXXXXXXXXXXX	YYYYYYYYYYYYYYYYYY	167	14	4	July	June
XXXXXXXXXXXXXXXXXX	YYYYYYYYYYYYYYYYYY	142	8	5	July	July
XXXXXXXXXXXXXXXXXX	YYYYYYYYYYYYYYYYYY	124	6	6	July	<active>
XXXXXXXXXXXXXXXXXX	YYYYYYYYYYYYYYYYYY	268	4	7	July	August
XXXXXXXXXXXXXXXXXX	YYYYYYYYYYYYYYYYYY	124	4	8	July	<active>
XXXXXXXXXXXXXXXXXX	YYYYYYYYYYYYYYYYYY	92	3	9	July	<active>





Social environment effects
– Peer commentary
– Social leader influence
– Promotions to a leader to influence group

Analytics and the Art of Winning…



About Us
Enterprise Data Analytics
Enterprise Video Indexing
Enterprise Broadcast
Mobile Video Analytics
Analytics Testimonials
24 Hour Pass
Company Blog
Advisory Board
Management
Contact Us
Careers
Login

Blog Headlines

ESPN: A look at Giancarlo Stanton

Red Sox Radio Network: The Differences In Adrian Gonzalez

ESPN: Albert Pujols

ESPN: Josh Hamilton's Hot Zone

ESPN: A Close Look at Matt Wieter

Scoring Applications

ESPN iScore Integration

Enterprise Analytics Solutions

MLB Analytics

TruMedia's MLB analytics platform provides MLB clubs and sports media properties with the ability to quickly analyze multiple data sources from one intuitive interface. Granular statistics generated by a vast array of filters as well as pitch by pitch video allows our licensees to maximize the value of their video and data.

MiLB Analytics

In partnership with Sportvision, TruMedia powers baseball's most comprehensive minor league baseball analytics platform. This unique platform capitalizes the granular data captured by Sportvision's innovative Pitch f/x, Hit f/x and Command f/x motion tracking systems aligned with pitch by pitch video.

Video: Creating a Playlist

Video: Interactive Hit Charts

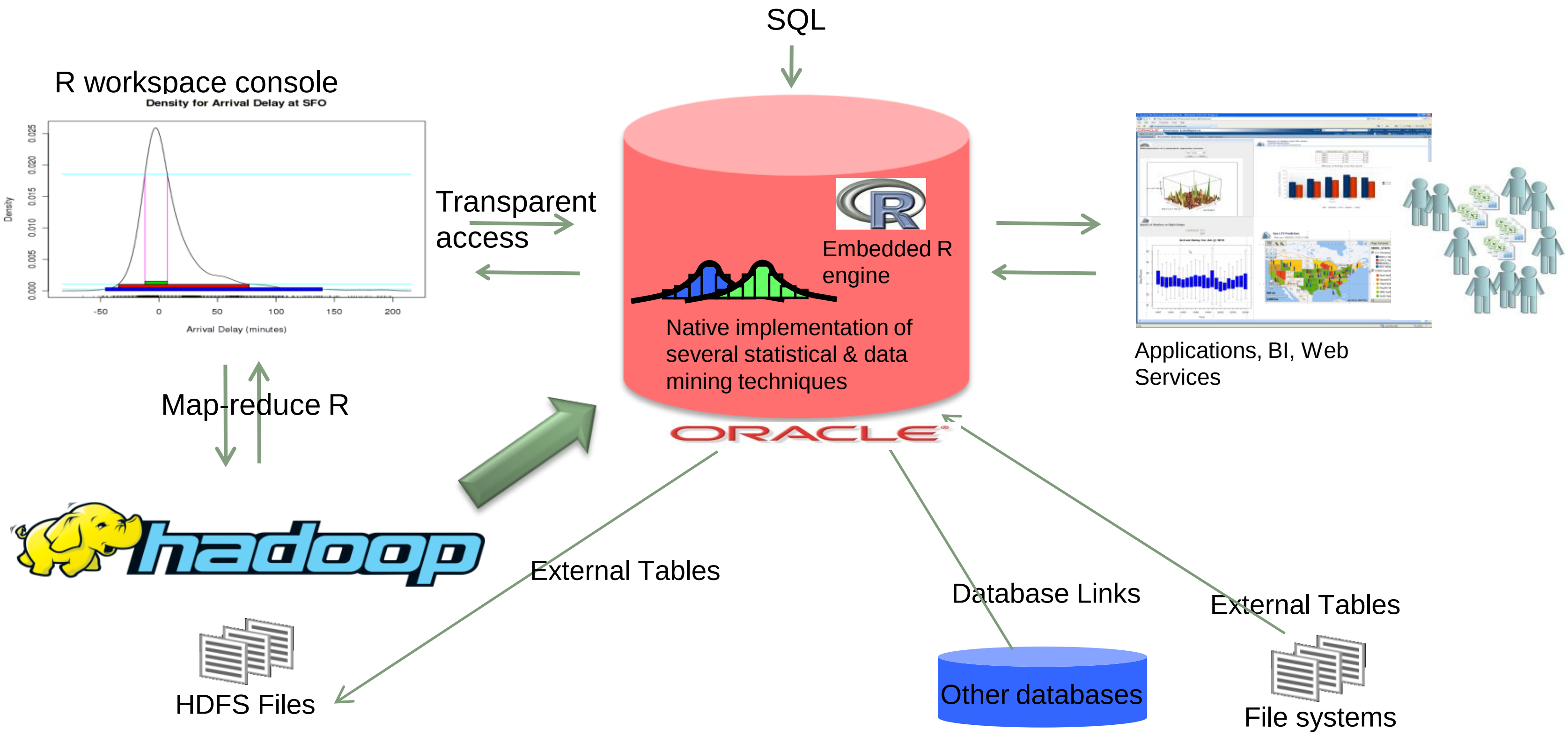
Baseball Analytics: Creating a Playlist
from Baseball Analytics PLUS
Baseball Analytics: Interactive Hit Charts

(professional or academic) with any major statistical / advanced analytics package (R, STATA, SPSS, Weka,

Oracle Big Data Analytics

- Focus is on the Enterprise Data Scientist who engages in Quantitative Research
- Goals
  1. Improve user efficiency by enabling focus on analysis as opposed to data access
  2. Enable deep analytics with computations occurring closer to data
  3. Allow transparent access to Enterprise compute infrastructures
  4. Shorten the path to application of cutting edge ideas into practice
  5. Enable quick transition from analysis to mass consumption of results

Oracle Big Data Analytics



SQL

R workspace console

Transparent access

Embedded R engine

Native implementation of several statistical & data mining techniques

Applications, BI, Web Services

Map-reduce R

External Tables

HDFS Files

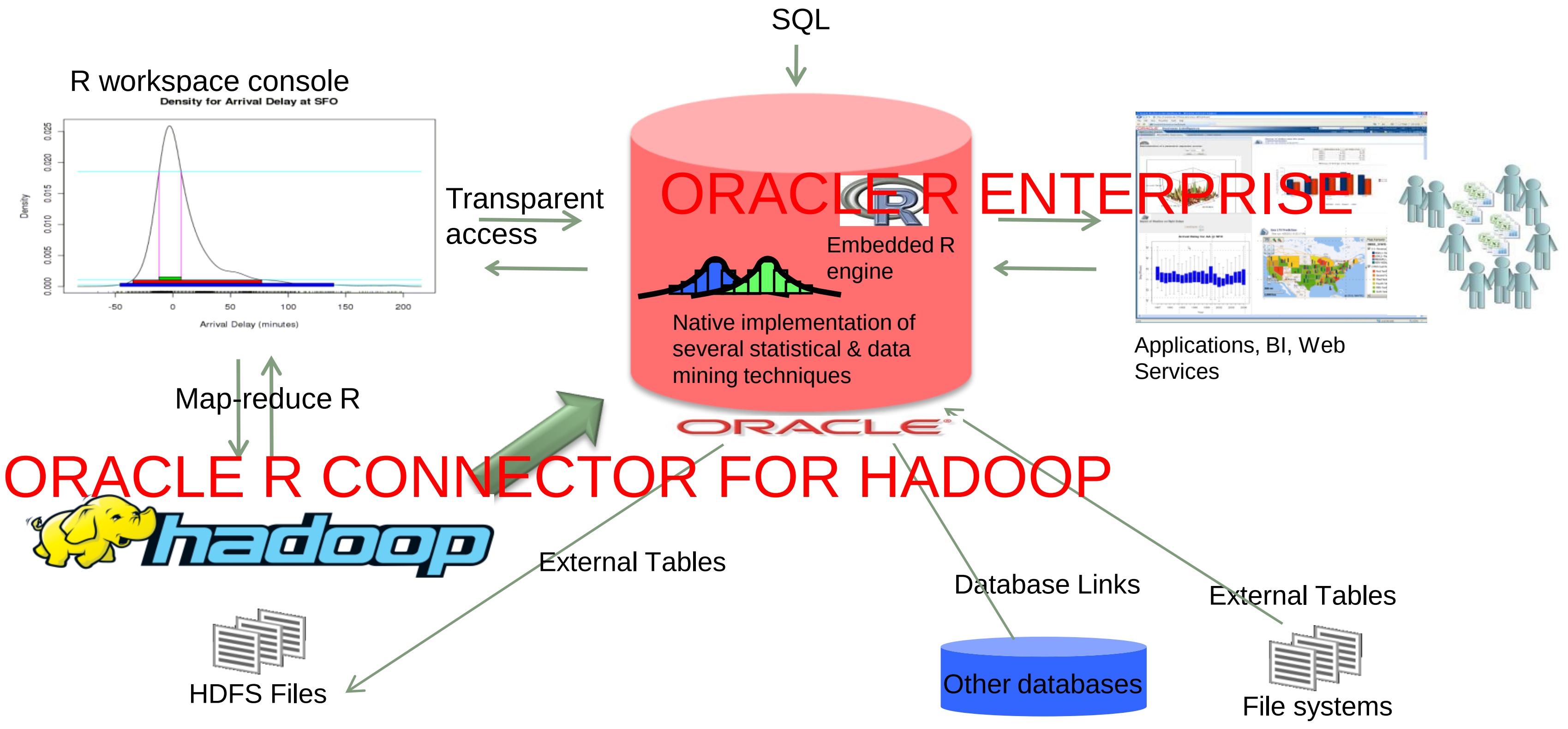
Database Links

Other databases

External Tables

File systems

Oracle Big Data Analytics



SQL

R workspace console

Transparent access

ORACLE R ENTERPRISE

Embedded R engine

Native implementation of several statistical & data mining techniques

Applications, BI, Web Services

Map-reduce R

ORACLE R CONNECTOR FOR HADOOP

External Tables

Database Links

External Tables

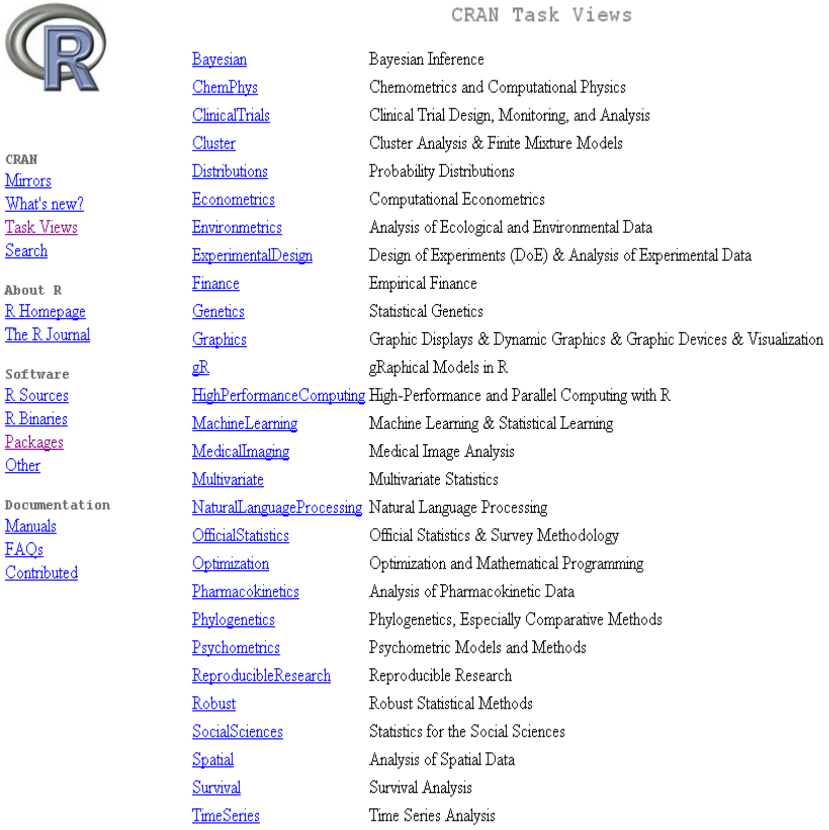
HDFS Files

Other databases

File systems

What is R?

- **R is an Open Source language and environment for statistical computing and graphics** **http://www.r-project.org/**
- **Started in 1994 as an Alternative to SAS, SPSS & Other proprietary Statistical Environments**
- **The R environment**
  - R is an integrated suite of software facilities for data manipulation, calculation and graphical display
- **Around 2 million R users worldwide**
  - Widely taught in Universities
  - Many Corporate Analysts know and use R
- **Hundreds of open sources packages to enhance productivity such as:**
  - Bioinformatics with R
  - Spatial Statistics with R
  - Financial Market Analysis with R
  - Linear and Non Linear Modeling



CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

CRAN Task Views

Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
Distributions	Probability Distributions
Econometrics	Computational Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
Finance	Empirical Finance
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
gR	gRaphical Models in R
HighPerformanceComputing	High-Performance and Parallel Computing with R
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
OfficialStatistics	Official Statistics & Survey Methodology
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Phylogenetics	Phylogenetics, Especially Comparative Methods
Psychometrics	Psychometric Models and Methods
ReproducibleResearch	Reproducible Research
Robust	Robust Statistical Methods
SocialSciences	Statistics for the Social Sciences
Spatial	Analysis of Spatial Data
Survival	Survival Analysis
TimeSeries	Time Series Analysis

Why statisticians/data analysts use R

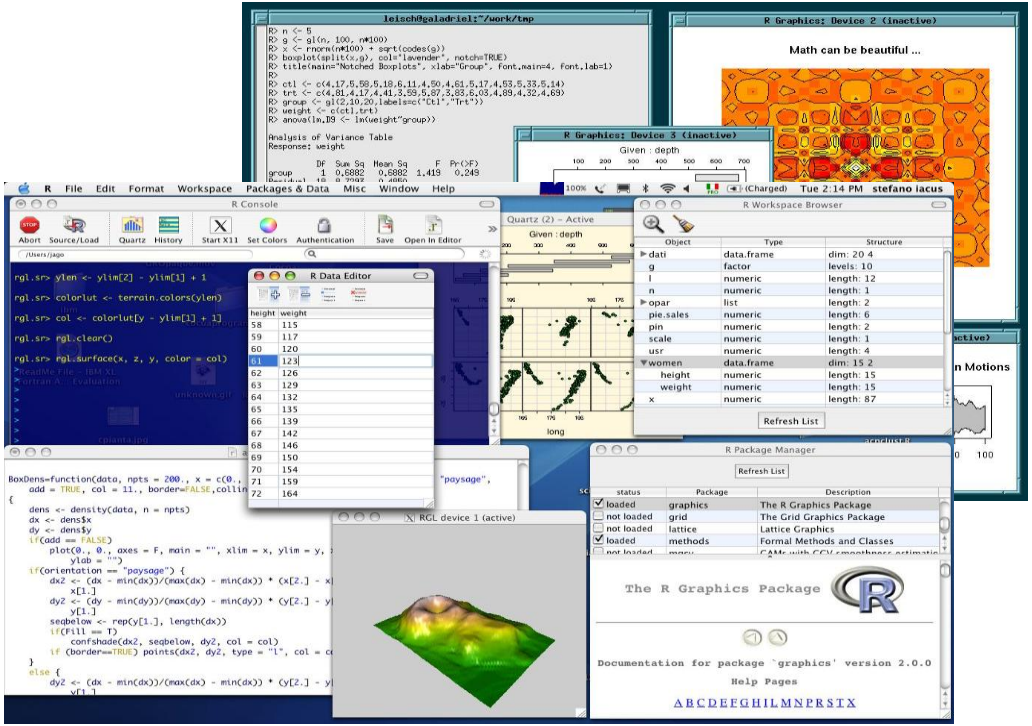
R is a statistics language similar to Base SAS or SPSS statistics

R environment is…

- Powerful - Extensive numerical techniques
- Extensible – 1000s of CRAN packages
- Exhaustive visualization
- Ease of installation and use
- Is becoming the language of research
- ***Free***

Statisticians may not be

- SQL literate
- Familiar with DBA tasks



leisch@galadriel:~/work/tmp

```
R> n <- 5
R> g <- gl(n, 100, n*100)
R> x <- rnorm(n*100) + sqrt(codes(g))
R> boxplot(split(x,g), col="lavender", notch=TRUE)
R> title(main="Notched Boxplots", xlab="Group", font.main=4, font.lab=1)
R>
R> ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
R> trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
R> group <- gl(2,10,20,labels=c("Ctl","Trt"))
R> weight <- c(ctl,trt)
R> anova(lm.D9 <- lm(weight~group))

Analysis of Variance Table
Response: weight

          Df  Sum Sq  Mean Sq      F  Pr(>F)
group      1  0.6882   0.6882  1.419   0.249
```

R Console

```
rgl.sr> ylen <- ylim[2] - ylim[1] + 1

rgl.sr> colorlut <- terrain.colors(ylen)

rgl.sr> col <- colorlut[y - ylim[1] + 1]

rgl.sr> rgl.clear()

rgl.sr> rgl.surface(x, z, y, color = col)
>
>
>
>
>
>
>
>
```

R Data Editor

height	weight
58	115
59	117
60	120
61	123
62	126
63	129
64	132
65	135
66	139
67	142
68	146
69	150
70	154
71	159
72	164

```
BoxDens=function(data, npts = 200., x = c(0.,
    add = TRUE, col = 11., border=FALSE,collin
{
    dens <- density(data, n = npts)
    dx <- dens$x
    dy <- dens$y
    if(add == FALSE)
        plot(0., 0., axes = F, main = "", xlim = x, ylim = y,
            ylab = "")
    if(orientation == "paysage") {
        dx2 <- (dx - min(dx))/(max(dx) - min(dx)) * (x[2.] - x
            x[1.]
        dy2 <- (dy - min(dy))/(max(dy) - min(dy)) * (y[2.] - y
            y[1.]
        seqbelow <- rep(y[1.], length(dx))
        if(Fill == T)
            confshade(dx2, seqbelow, dy2, col = col)
        if (border==TRUE) points(dx2, dy2, type = "l", col = c
    }
    else {
        dy2 <- (dx - min(dx))/(max(dx) - min(dx)) * (y[2.] - y
```

R Workspace Browser

Object	Type	Structure
▶ dati	data.frame	dim: 20 4
g	factor	levels: 10
l	numeric	length: 12
n	numeric	length: 1
▶ opar	list	length: 2
pie.sales	numeric	length: 6
pin	numeric	length: 2
scale	numeric	length: 1
usr	numeric	length: 4
▼ women	data.frame	dim: 15 2
height	numeric	length: 15
weight	numeric	length: 15
x	numeric	length: 87

R Package Manager

status	Package	Description
☑ loaded	graphics	The R Graphics Package
□ not loaded	grid	The Grid Graphics Package
□ not loaded	lattice	Lattice Graphics
☑ loaded	methods	Formal Methods and Classes

The R Graphics Package

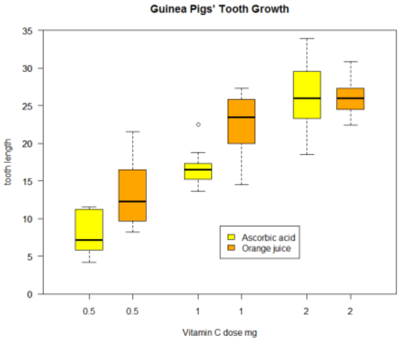
Documentation for package `graphics' version 2.0.0

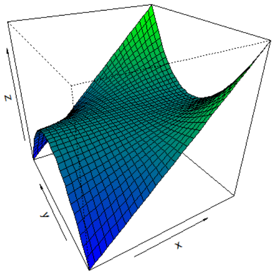
Help Pages

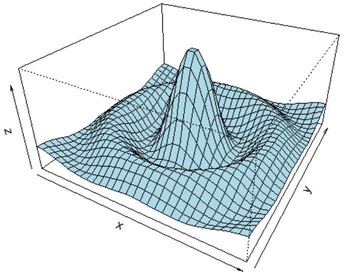
A B C D E F G H I L M N P R S T X

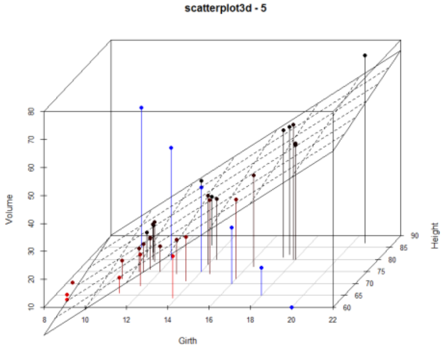
R Graphics: Device 2 (inactive)
Math can be beautiful ...
R Graphics: Device 3 (inactive)
Given : depth
100 200 300 400 500 600 700
R File Edit Format Workspace Packages & Data Misc Window Help
100%
(Charged) Tue 2:14 PM stefano iacus
Abort Source/Load Quartz History Start X11 Set Colors Authentication Save Open In Editor
/Users/jago
Quartz (2) - Active
long
RGL device 1 (active)
Refresh List
n Motions
100

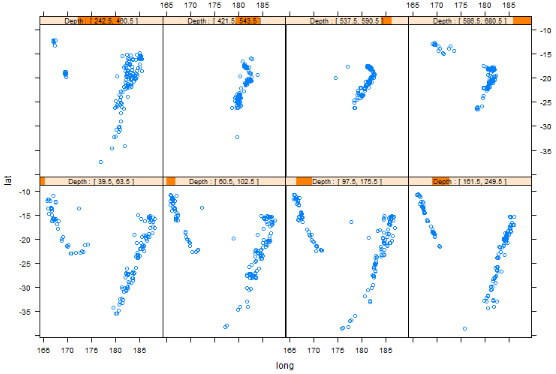
Graph examples…

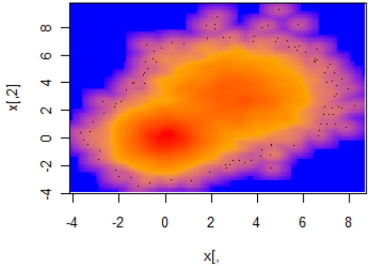


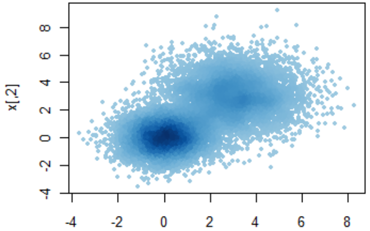


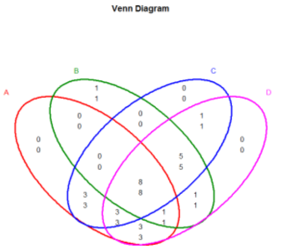


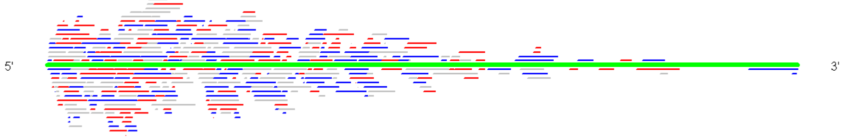




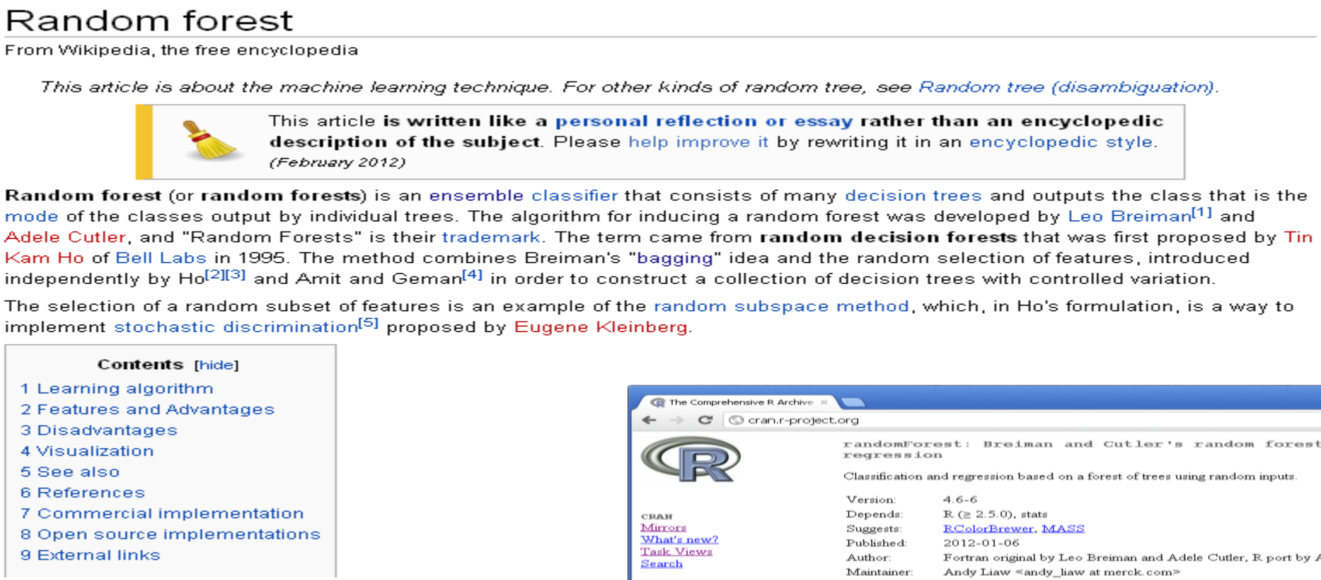








R is the language of research



Random forest

From Wikipedia, the free encyclopedia

This article is about the machine learning technique. For other kinds of random tree, see Random tree (disambiguation).

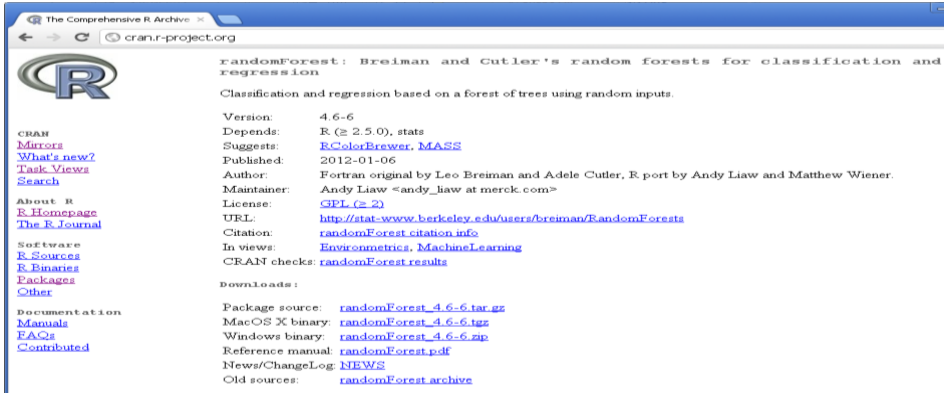
This article **is written like a personal reflection or essay rather than an encyclopedic description of the subject**. Please help improve it by rewriting it in an encyclopedic style. *(February 2012)*

Random forest (or **random forests**) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman[1] and Adele Cutler, and "Random Forests" is their trademark. The term came from **random decision forests** that was first proposed by Tin Kam Ho of Bell Labs in 1995. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho[2][3] and Amit and Geman[4] in order to construct a collection of decision trees with controlled variation.

The selection of a random subset of features is an example of the random subspace method, which, in Ho's formulation, is a way to implement stochastic discrimination[5] proposed by Eugene Kleinberg.

Contents [hide]

1 Learning algorithm
2 Features and Advantages
3 Disadvantages
4 Visualization
5 See also
6 References
7 Commercial implementation
8 Open source implementations
9 External links



randomForest: Breiman and Cutler's random forests for classification and regression

Classification and regression based on a forest of trees using random inputs.

Version:	4.6-6
Depends:	R (≥ 2.5.0), stats
Suggests:	RColorBrewer, MASS
Published:	2012-01-06
Author:	Fortran original by Leo Breiman and Adele Cutler, R port by Andy Liaw and Matthew Wiener.
Maintainer:	Andy Liaw <andy_liaw at merck.com>
License:	GPL (≥ 2)
URL:	http://stat-www.berkeley.edu/users/breiman/RandomForests
Citation:	randomForest citation info
In views:	Environmetrics, MachineLearning
CRAN checks:	randomForest results

Downloads:

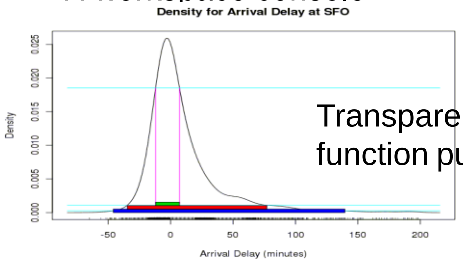
Package source:	randomForest_4.6-6.tar.gz
MacOS X binary:	randomForest_4.6-6.tgz
Windows binary:	randomForest_4.6-6.zip
Reference manual:	randomForest.pdf
News/ChangeLog:	NEWS
Old sources:	randomForest archive

Limitations of R

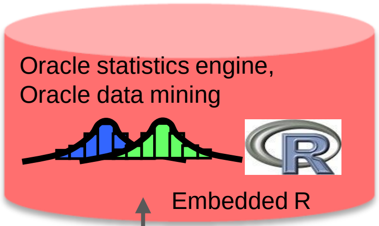
- R is a client and server bundled together as 1 executable - like Excel
  - Single user tool
  - Not multi-threaded
  - Cannot leverage CPU capacity even on a user's laptop/desktop
- R requires data it operates on to be first loaded into memory
  - Loading data may not be a limitation given RAM available on laptops/desktops
  - R’s *call by value* semantics means as data flows into functions, for each function invocation, a complete copy of the data is made
  - As a result you quickly run into memory limits

R integration

R workspace console

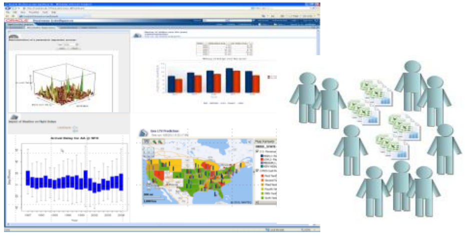


Transparent access via function push-down



Oracle statistics engine, Oracle data mining

Embedded R



BI, Workflows, Web Services

R is integrated into SQL

No changes to the user experience

Development

Lights out execution

Production

Embed in operational systems

Consumption

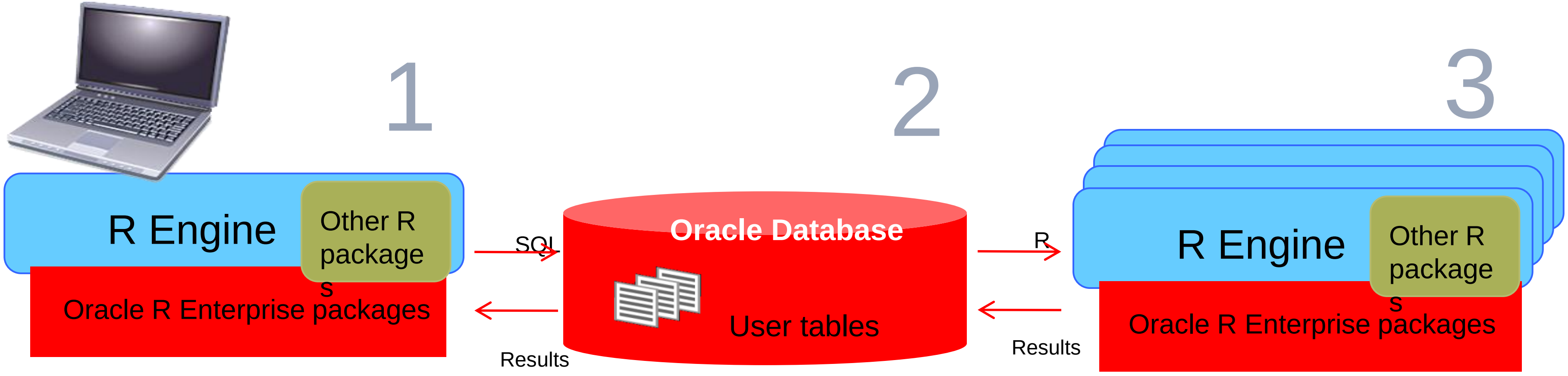
Oracle's Approach – Comprehensive Enterprise-level Big Data Analytics based on R environment

1. Oracle's Distribution and Support of Open Source R
   - Only redistribution with comprehensive platform support – Linux, Solaris, AIX
     - Enhanced performance with Intel MKL, AMD ACML OR SUN perf libraries for x86 hardware
   - Certification of select CRAN packages
   - Distributed via public-yum.oracle.com, pkg.oracle.com
2. Oracle R Enterprise
   - Embedded component of the RDBMS
   - Eliminates R's memory constraint by enabling R to work transparently on database resident data
   - Brings R users closer to Oracle Database by transparently leveraging in-database analytics via R
   - Enables integration of R scripts into enterprise production applications and BI dashboards
   - Fully leverages the latest R algorithms and models contributed to R's CRAN
3. Oracle R Connector For Hadoop
   - Interactive R interface to HDFS data and Hadoop infrastructure
   - Only available solution to combine database, HDFS and local file system data into 1 hadoop R computation

Licensing

1. Oracle's Distribution and Support of Open Source R
   - Free
2. Oracle R Enterprise
   - Available as part of Oracle Advanced Analytics Option to Oracle Database
   - Oracle Advanced Analytics Option = Oracle Data Mining + Oracle R Enterprise
   - Oracle Data Mining algorithms are available via Oracle R Enterprise interface, SQL and GUI
3. Oracle R Connector For Hadoop
   - Available as part of Oracle Big Data Connectors software suite

1. Collaborative Execution Model



User R Engine on desktop

- R-SQL Transparency Framework intercepts R functions for scalable in-database execution
- Interactive display of graphical results and flow control as in standard R
- Submit entire R scripts for execution by Oracle Database

Post processing of results

Database Compute Engine

- Scale to large datasets
- Leverage database SQL parallelism
- Leverage new and existing in-database statistical and data mining capabilities

Collaborative execution with in-database R engine

R Engine(s) spawned by Oracle DB

- Database can spawn multiple R engines for database-managed parallelism
- Efficient parallel data transfer to spawned R engines to emulate map-reduce style algorithms and applications
- Enables "lights-out" execution of R scripts

Analytic techniques not available in-database

2. Deferred execution

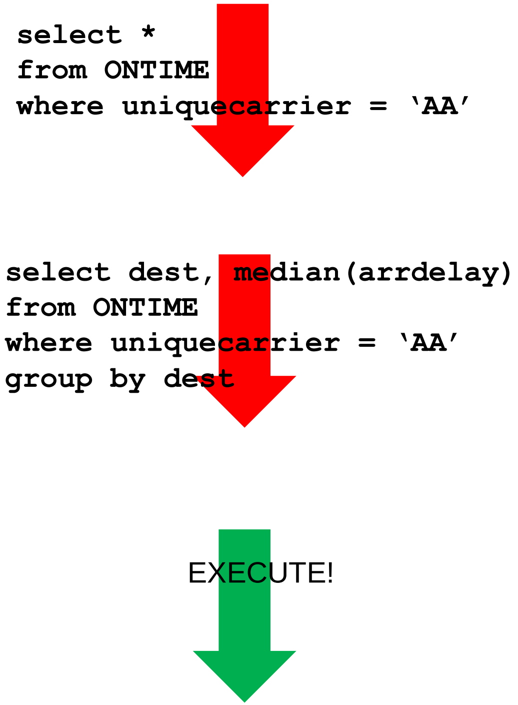
```
#Filter rows that correspond to American Airlines
#Flights
ONTIME <- ONTIME[ONTIME$UNIQUECARRIER==‘AA’]



#Calculate median arrival delay for all flights grouped
#by destination
aggdata <- aggregate(ONTIME$ARRDELAY,
                     by = list(ONTIME$DEST),
                     FUN = median)



plot(aggdata)
```

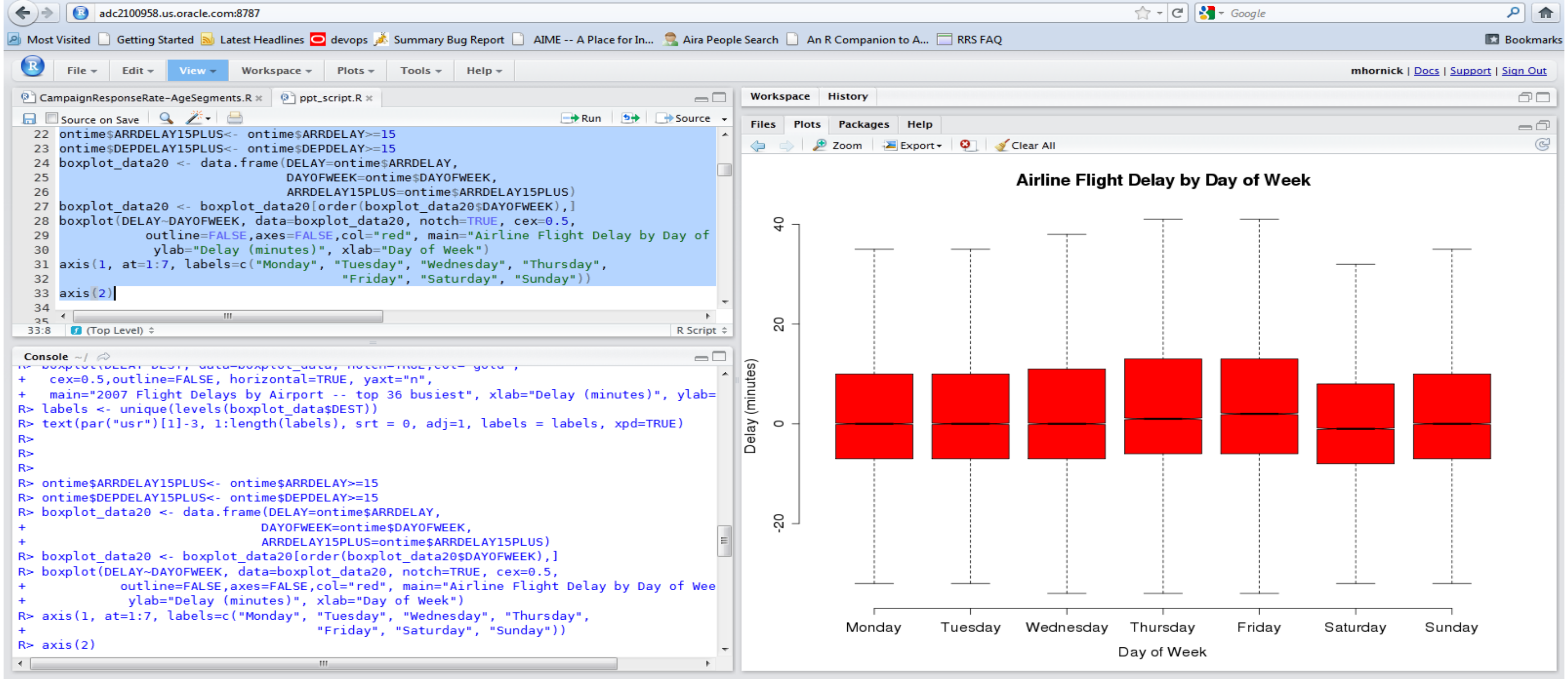


```
select *
from ONTIME
where uniquecarrier = ‘AA’
```

```
select dest, median(arrdelay)
from ONTIME
where uniquecarrier = ‘AA’
group by dest
```

3. Collaborative Visualization

Push computations into SQL and render using R



```
ontime$ARRDELAY15PLUS<- ontime$ARRDELAY>=15
ontime$DEPDELAY15PLUS<- ontime$DEPDELAY>=15
boxplot_data20 <- data.frame(DELAY=ontime$ARRDELAY,
                             DAYOFWEEK=ontime$DAYOFWEEK,
                             ARRDELAY15PLUS=ontime$ARRDELAY15PLUS)
boxplot_data20 <- boxplot_data20[order(boxplot_data20$DAYOFWEEK),]
boxplot(DELAY~DAYOFWEEK, data=boxplot_data20, notch=TRUE, cex=0.5,
           outline=FALSE,axes=FALSE,col="red", main="Airline Flight Delay by Day of
            ylab="Delay (minutes)", xlab="Day of Week")
axis(1, at=1:7, labels=c("Monday", "Tuesday", "Wednesday", "Thursday",
                                    "Friday", "Saturday", "Sunday"))
axis(2)
```

```
R> boxplot(DELAY~DEST, data=boxplot_data, notch=TRUE,col="gold",
+   cex=0.5,outline=FALSE, horizontal=TRUE, yaxt="n",
+   main="2007 Flight Delays by Airport -- top 36 busiest", xlab="Delay (minutes)", ylab=
R> labels <- unique(levels(boxplot_data$DEST))
R> text(par("usr")[1]-3, 1:length(labels), srt = 0, adj=1, labels = labels, xpd=TRUE)
R>
R>
R>
R> ontime$ARRDELAY15PLUS<- ontime$ARRDELAY>=15
R> ontime$DEPDELAY15PLUS<- ontime$DEPDELAY>=15
R> boxplot_data20 <- data.frame(DELAY=ontime$ARRDELAY,
+                             DAYOFWEEK=ontime$DAYOFWEEK,
+                             ARRDELAY15PLUS=ontime$ARRDELAY15PLUS)
R> boxplot_data20 <- boxplot_data20[order(boxplot_data20$DAYOFWEEK),]
R> boxplot(DELAY~DAYOFWEEK, data=boxplot_data20, notch=TRUE, cex=0.5,
+           outline=FALSE,axes=FALSE,col="red", main="Airline Flight Delay by Day of Wee
+            ylab="Delay (minutes)", xlab="Day of Week")
R> axis(1, at=1:7, labels=c("Monday", "Tuesday", "Wednesday", "Thursday",
+                                   "Friday", "Saturday", "Sunday"))
R> axis(2)
```

4. R is integrated into SQL

```
select * from table(rqTableEval(
    cursor(select * from fish),
    NULL,
    'select t.*, 1 rowsum from fish t',
    'function(x, param) {
      dat <- data.frame(x, stringsAsFactors=F)
      cbind(dat, ROWSUM = apply(dat,1,sum))
     }'));

select * from table(rqRowEval(
    cursor(select * from fish),
    NULL,
    'select t.*, 1 rowsum from fish t',
    1,
    'function(x, param) {
      dat <- data.frame(x, stringsAsFactors=F)
      cbind(dat, ROWSUM = apply(dat,1,sum)+10)
     }'));
```

- R closure (script) is the integration point
- Different types of inputs
  - **Parallel row streams**
  - **Parallel groups of rows**
  - **Parallel iterations**
- Run-time parameters
  - **e.g. Date Filters, R objects**
- Flexible outputs
  - **Vertical or Horizontal addition to an existing table**
    - **Data or models**
  - **Frames, Vectors, Graphics**

5. Data Flow parallelism at work

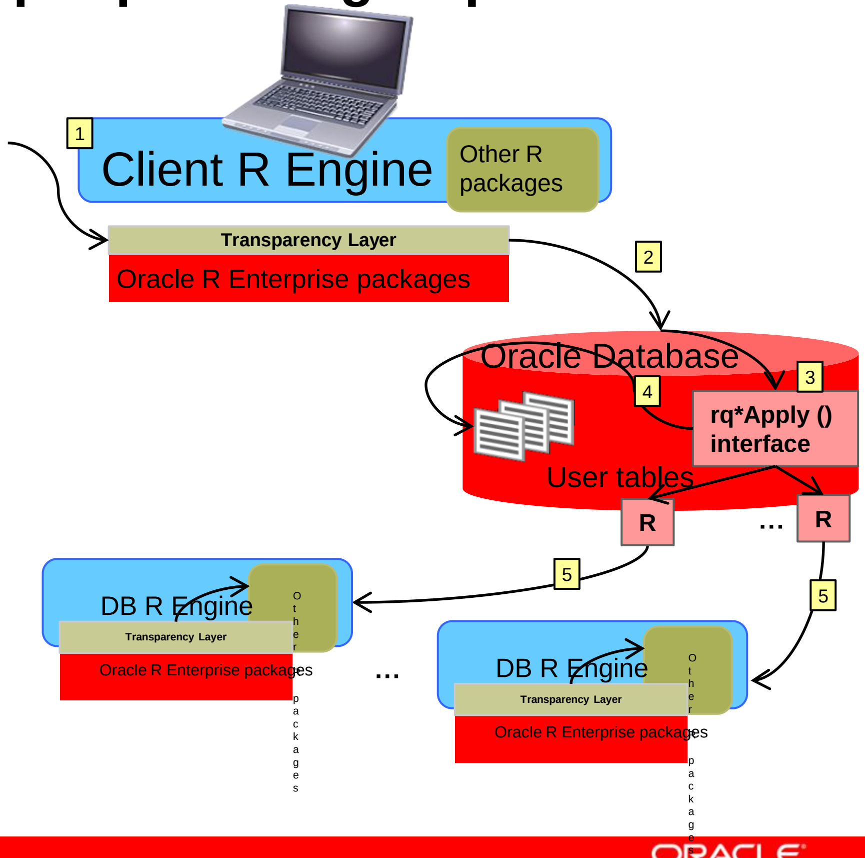
Partitioned model builds: 1 model per product group

```
modList <- ore.groupApply(
   ONTIME_S,
   INDEX=ONTIME_S$DEST,
   function(dat) {
     library(randomforest)
     reg(ARRDELAY ~ DISTANCE + DEPDELAY, dat)
   });
modList_local <- ore.pull(modList)
summary(modList_local$BOS) ## return model for BOS
```

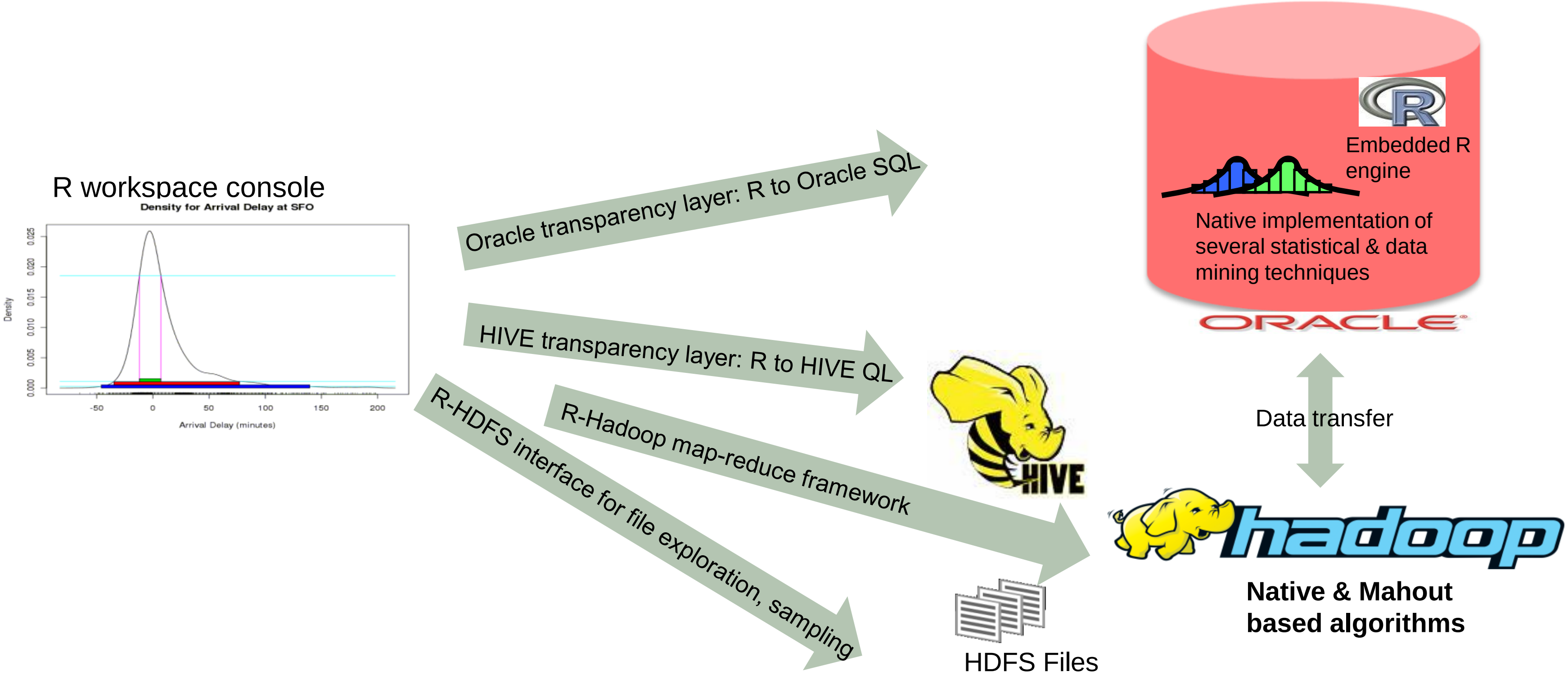
Goal: Build models in parallel on partitions of dataset
Function loaded to DB R Engine
Parallelism enabled through INDEX column

Data group subset for 1 INDEX value passed to DB R Engine

Result “modList” returned as a list of model objects, one per group



Oracle R Connector for Hadoop



R workspace console

Oracle transparency layer: R to Oracle SQL

HIVE transparency layer: R to HIVE QL

R-Hadoop map-reduce framework

R-HDFS interface for file exploration, sampling

Embedded R engine

Native implementation of several statistical & data mining techniques

Data transfer

Native & Mahout based algorithms

HDFS Files

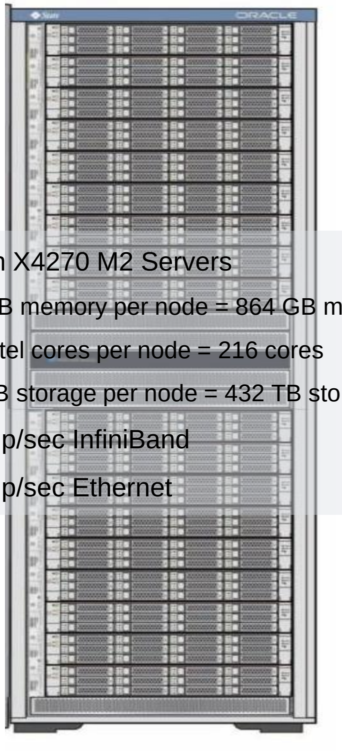
Oracle R Connector for Hadoop (ORCH) Concepts

1. Access to HDFS files
   1. Auto discovery of metadata
   2. Sampling
2. HIVE SQL connectivity
   1. R to SQL
3. Hadoop Analytics
   1. Open source Mahout
   2. Home grown techniques

Key Highlights

1. Supports interactive access to HDFS data and Hadoop infrastructure
2. Allows database resident data to be used within a Hadoop calculation
3. Supports local execution and debugging of R code – disconnected from Hadoop
4. Treats metadata and data separately when possible
   - Samples HDFS files to create metadata description
5. Provides flexible output options
   - As R object to user session
   - Load to database
   - Continue to post-process

Big Data Appliance



18 Sun X4270 M2 Servers

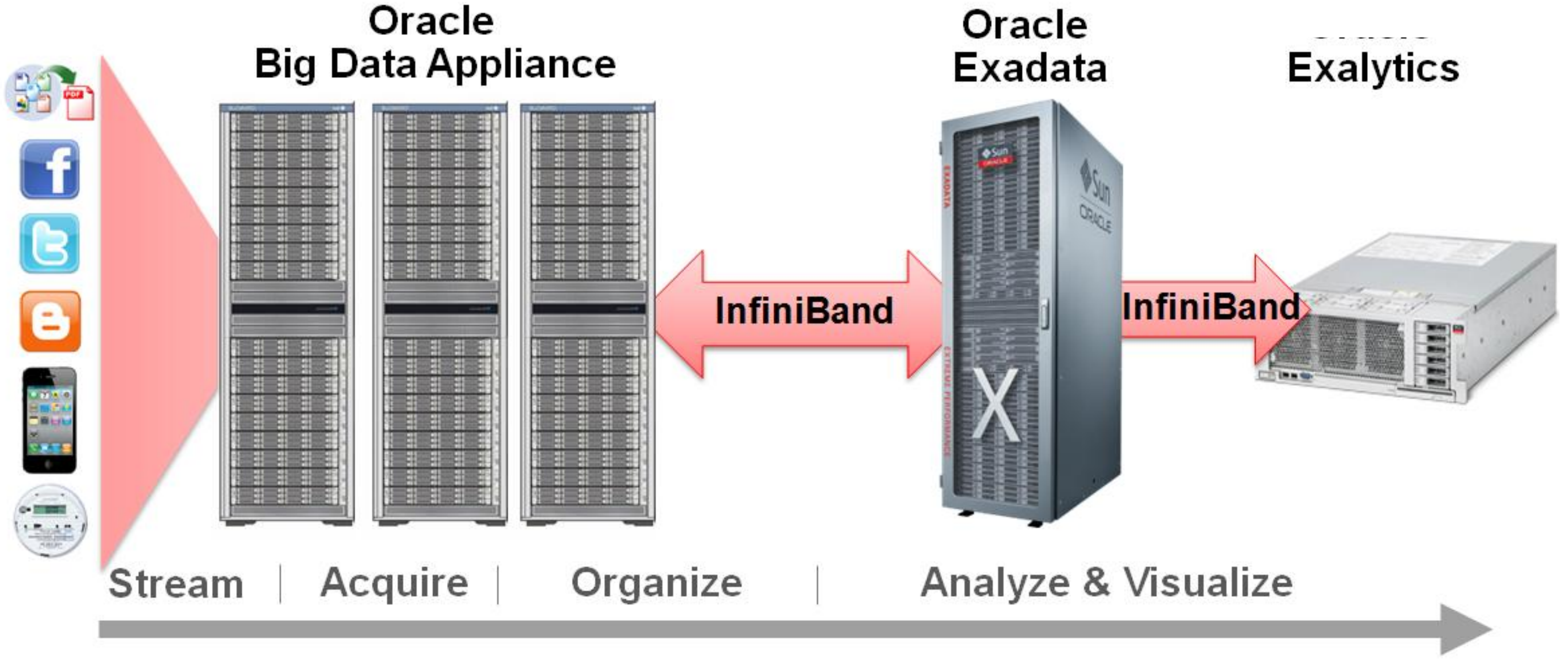
- 48 GB memory per node = 864 GB memory
- 12 Intel cores per node = 216 cores
- 24 TB storage per node = 432 TB storage

40 Gb p/sec InfiniBand

10 Gb p/sec Ethernet

- An engineered system optimized for capturing and integrating "low density" data into Exadata
  - High-performance Hardware
    - Optimized for Hadoop and NoSQL workloads
    - InfiniBand Networking for integration with **Exadata**
  - Software:
    - Oracle Hadoop
    - Oracle R Hadoop Connector
    - Oracle R Enterprise client (optional)
    - Oracle NoSQL DB
    - Oracle Data Integrator (Hadoop capabilities)
    - Oracle Loader for Hadoop

Big Data Appliance Usage Model



Key take-aways

1. Improve user efficiency by enabling focus on analysis as opposed to data access
   - Transparent support for R language on database and HDFS objects
2. Enable deep analytics with computations occurring closer to data
   - Native implementation of statistics and data mining algorithms
   - R engine as an embedded component of database
3. Allow transparent access to Compute Infrastructures
   - Database & Hadoop platforms
4. Shorten path to application of cutting edge ideas into practice
   - Oracle’s R Distribution & Embedded R engine
5. Enable quick transition from analysis to mass consumption of results
   - R integrated into SQL