



A Gentle Introduction to R and its Applications in Business Intelligence

Michael Driscoll, Principal, Dataspora, Inc.

John Oram, Researcher, San Francisco Estuary Institute

Jim Porzak, Sr. Director of Analytics, Responsys, Inc.

SofTech

24 September, 2008. San Rafael, California



Outline

- Evolution of R
- R as a BI Tool
- Jim's Case Studies
- Mike's Case Study
- John's Case Study
- How to Get Started with R
- Wrap





First there was S

- R is the free (GNU), open source, version of S
 - S developed by John Chambers *et al* while at Bell Labs in 80's
 - For "data analysis and graphics" (with statistics emphasis)
 - Ver.4 defined by the "Green Book" *Programming with Data*, 1998
 - "S-Plus" now owned by Insightful Corp., Seattle, WA
- R was initially written in early 1990's
 - by *Robert Gentleman* and *Ross Ihaka*
 - Statistics Department of the University of Auckland
 - GNU GPL release in 1995
 - "R" is before "S", as in "HAL" is before "IBM"
- Since 1997 a core group of \pm 20 developers
 - Initial V1.0 released in February, 2000
 - Continually developed with a new 0.1 level release \sim 6 months



Current State of R

As of October 2004

- V2.0 Released October, 2004
- Windows, Mac OS, Linux & Unix ports
- Over 400 submitted packages from “abind” to “zoo”
- 12th newsletter (Volume 4/2) published September 2004
- The first useR! – R User Conference held in Vienna May 2004
- ~400 R-help messages per week
- ~ Dozen texts specifically on R or with R examples and code
- R language generally accepted to be more powerful than S-Plus
- Some interesting GUI work in progress

As of July 2008

- V2.7.1 Released June, 2008
- Windows, Mac OS, Linux & Unix ports (including Vista)
- 1450+ packages; “aaMI” to “zoo” (+45 Omegahat, +260 Bioconductor)
- 22nd newsletter (Volume 8/1) published May 2008
- The fourth useR! conference next month in Dortmund, Germany
- ~700 R-help messages per week
- 65 texts specifically on R or with R examples and code
- R ~ universally taught.
- Commercial support (REvolution, ...)
- JGR, Rattle, RCmdr, ...
- *Web based applications now easier*



www.r-project.org

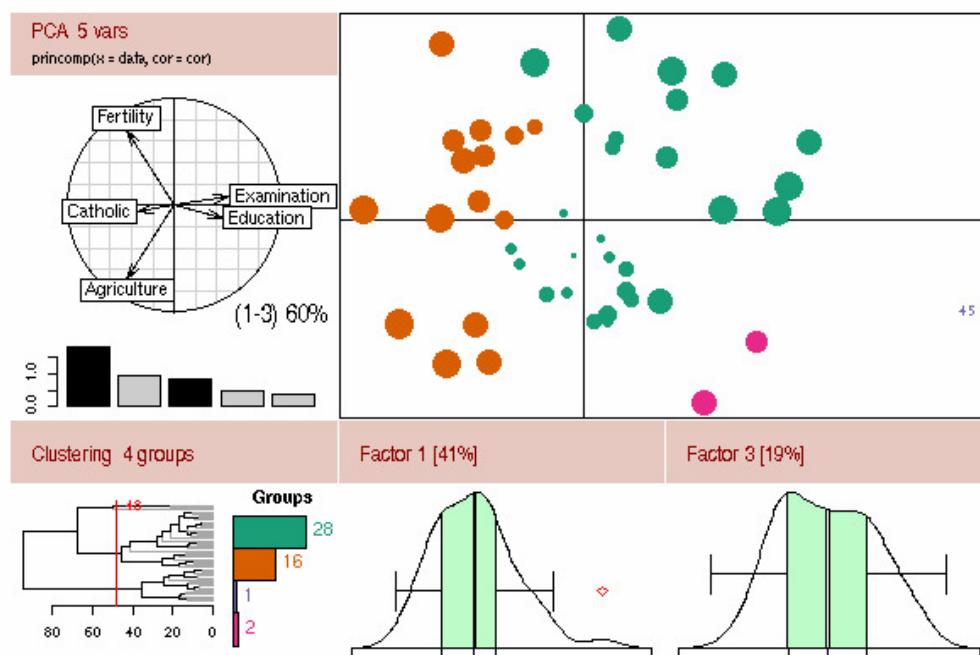


[About R](#)
[What is R?](#)
[Contributors](#)
[Screenshots](#)
[What's new?](#)

[Download](#)
[CRAN](#)

[R Project](#)
[Foundation](#)
[Members & Donors](#)
[Mailing Lists](#)
[Bug Tracking](#)
[Developer Page](#)
[Conferences](#)
[Search](#)

[Documentation](#)
[Manuals](#)
[FAQs](#)
[Newsletter](#)
[Wiki](#)
[Books](#)
[Certification](#)



Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred [CRAN mirror](#).
- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.



cran.cnr.berkeley.edu



The Comprehensive R Archive Network

Frequently used pages

[CRAN](#)
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

[About R](#)
[R Homepage](#)

[Software](#)
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

[Documentation](#)
[Manuals](#)
[FAQs](#)
[Contributed](#)
[Newsletter](#)

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Linux](#)
- [MacOS X](#)
- [Windows](#)

Source Code for all Platforms

Windows and Mac users most likely want the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- **The latest release** (2008-06-23): [R-2.7.1.tar.gz](#) (read [what's new](#) in the latest version).
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

Questions About R



Task Views 1st step in finding relevant packages

cran.cnr.berkeley.edu/web/views/



[CRAN](#)
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

[About R](#)
[R Homepage](#)

[Software](#)
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

[Documentation](#)
[Manuals](#)
[FAQs](#)
[Contributed](#)
[Newsletter](#)

[Bayesian](#)
[Cluster](#)
[Econometrics](#)
[Environmetrics](#)
[ExperimentalDesign](#)
[Finance](#)
[Genetics](#)
[Graphics](#)
[gR](#)
[MachineLearning](#)
[Multivariate](#)
[NaturalLanguageProcessing](#)
[Optimization](#)
[Pharmacokinetics](#)
[Psychometrics](#)
[Robust](#)
[SocialSciences](#)
[Spatial](#)
[Survival](#)

CRAN Task Views

Bayesian Inference
Cluster Analysis & Finite Mixture Models
Computational Econometrics
Analysis of Ecological and Environmental Data
Design of Experiments (DoE) & Analysis of Experimental Data
Empirical Finance
Statistical Genetics
Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
gRaphical Models in R
Machine Learning & Statistical Learning
Multivariate Statistics
Natural Language Processing
Optimization and Mathematical Programming
Analysis of Pharmacokinetic Data
Psychometric Models and Methods
Robust Statistical Methods
Statistics for the Social Sciences
Analysis of Spatial Data
Survival Analysis



R Help – Best Support!

		Remove label "R-Help"	Report Spam	Delete	More actions...	Refresh	1 - 50 of 10513	Older »	Oldest »
Select: All, None, Read, Unread, Starred, Unstarred									
<input type="checkbox"/>	★ march	Inbox	[R] gbn with jumps	- Hi everybody I'd like to simulate a Generalized Wiener Process with jumps. Any sugge:		7:03 am			
<input type="checkbox"/>	★ Vladimir Eremeev	Inbox	[R] simpler solution (untested)	- axis says that this function has the logical parameter outer "indicating whe		6:56 am			
<input type="checkbox"/>	★ march	Inbox	[R] gbm with jumps	- Hi everybody I'd like to simulate a Generalized Wiener Process with jumps. Any sugge		6:48 am			
<input type="checkbox"/>	★ Rafael, Peter Vladimir (3)	Inbox	[R] Three horizontal axes OR Two axes on same side?	- Dear list: I need to reproduce a plot with three d		6:44 am			
<input type="checkbox"/>	★ Bram Kuijper	Inbox	[R] levelplot not adjusting colors	- Hi all, I try to make a levelplot from the Trellis graphics package of count		6:41 am			
<input type="checkbox"/>	★ Marta Rufino	Inbox	[R] warning in GAM	- Hello, I have a problem when doing gam (from gam library; I am using R 2.4.0, window:		5:48 am			
<input type="checkbox"/>	★ Antje, Peter (4)	Inbox	[R] Error in plot.new() : Figure margins too large	- Hello, was could be the reason for such an error messa		5:33 am			
<input type="checkbox"/>	★ Indermaur, Ken, Prof (3)	Inbox	[R] batch job GLM calculations	- Hello I want to batch job the calculation of many GLM-models, extract son		1:19 am			
<input type="checkbox"/>	★ Adrian .. Prof, Adrian (9)	Inbox	[R] a question of substitute	- The 'Right Thing' is for oneway.test() to allow a variable for the first argument, a		12:42 am			
<input type="checkbox"/>	★ David, Marc (2)	Inbox	[R] zero margin / marginless plots	- Hi, I'd like to produce a marginless or zero margin plot so that the pixel		7:37 pm			
<input type="checkbox"/>	★ Walter, Torsten, Richard (3)	Inbox	[R] posthoc tests with ANCOVA	- The WoodEnergy example in package HH (available on CRAN) is similar #		Jan 10			
<input type="checkbox"/>	★ karl.sommer	Inbox	[R] axis date format in lattice	- Hello list, plotting the following example 1 in lattice only labels the x-axis wi		Jan 10			
<input type="checkbox"/>	★ Tong .. Prof, François (9)	Inbox	[R] A question about R environment	- Philippe Grosjean] >Please, don't reinvent the wheel: putting function		Jan 10			
<input type="checkbox"/>	★ Michael, Peter (2)	Inbox	[R] TCL/TK and R documentation?	- I am hoping something has changed since I last asked about this. Is tl		Jan 10			
<input type="checkbox"/>	★ Simon, Setzer.Wood., Ken (3)	Inbox	[R] problems with optim, "for"-loops and machine precision	- Two possibilities for why your 7 parameter		Jan 10			
<input type="checkbox"/>	★ Darren Weber	Inbox	[R] axis labels at subset of tick marks	- For example, this works: x = seq(-100, 1000, 25) y = x * x plot(x,y,		Jan 10			
<input type="checkbox"/>	★ Colleen.Ross .. Thomas (3)	Inbox	[R] SAS and R code hazard ratios	- On Wed, 10 Jan 2007, Colleen.Ross@kp.org wrote: > I am new to R and		Jan 10			
<input type="checkbox"/>	★ Thomas, Duncan, Peter (3)	Inbox	[R] "go" or "goto" command	- Thomas L Jones wrote: > Some computer languages, including C, have a "go" c		Jan 10			
<input type="checkbox"/>	★ Feng, David, Feng (3)	Inbox	[R] logistic regression packages	- Hi David: Thanks for you information. 2 further questions: 1. I found out th		Jan 10			
<input type="checkbox"/>	★ David	Inbox	[R] Installation problem with package mixtools	- I am trying to install mixtools on Debian Etch and get the follc		Jan 10			
<input type="checkbox"/>	★ Tord, Roger (2)	Inbox	[R] map data.frame() data after having linked them to a read.shape() object	- On Wed, 10 Jan 2007, Tord Snäll		Jan 10			
<input type="checkbox"/>	★ Stephen, chao (2)	Inbox	[R] using DBI	- The way MySQL works, I use RMySQL to contact, which in turn uses DBI. There is a library R		Jan 10			
<input type="checkbox"/>	★ Paul Mathews	Inbox	[R] Meeting announcement: An Introduction to Data Analysis Using R	- An Introduction to Data Analysis Usin		Jan 10			
<input type="checkbox"/>	★ Kati, roger (2)	Inbox	[R] 2 problems with latex.table (quantreg package)	- reproducible - The usual R-help etiquette recommends: 1		Jan 10			
<input type="checkbox"/>	★ John .. Jeffrey, Brian (12)	Inbox	[R] scripts with littler	- Brian Ripley wrote: > Exactly as documented. The argument is named 'new' and not ...		Jan 10			
<input type="checkbox"/>	★ Jenny, Zoltan (3)	Inbox	[R] correlation value and map	- Hi Zoltan, Right, I have 30x32=960 data points per year (It is actually the mean		Jan 10			

Core Developers!



useR! 2008

www.statistik.uni-dortmund.de/useR-2008/



The R User Conference 2008

August 12-14, Technische Universität Dortmund, Germany

Organizer: Fakultät Statistik, Technische Universität Dortmund
Co-Organizer: Austrian Association for Statistical Computing
Sponsors: R Foundation for Statistical Computing



Austrian Association
for Statistical Computing



SFB 475: Reduction of Complexity
in Multivariate Data Structures



Taylor & Francis Group

CHAPMAN & HALL



A STAR ALLIANCE MEMBER



Conference

[About the Conference](#)
[Date & Location](#)
[Important Dates](#)
[Call for Papers](#)
[Download: Logo, Flyer, Poster](#)
[Funding](#)
[Participants](#)

Program

[Conference Program](#)
[Invited Lectures](#)
[Presentations](#)
[Tutorials](#)
[Social Program](#)
[Program Committee](#)
[Online Registration](#)

Dortmund

[Accommodation](#)
[About Dortmund](#)
[Travel information](#)

About the Conference

useR! 2008, the R user conference, takes place at the Fakultät Statistik, Technische Universität Dortmund, Germany from 2008-08-12 to 2008-08-14. Pre-conference tutorials will take place on August 11. The conference is organized by the Fakultät Statistik, Technische Universität Dortmund and the Austrian Association for Statistical Computing (AASC). It is funded by the R Foundation for Statistical Computing.

Date & Location

August 12-14, 2008 ([iCalendar file](#))
Fakultät Statistik
Technische Universität Dortmund



Significant New R Books

Use R

12 results

Applied Econometrics with R

Kleiber, C., Zeileis, A., ISBN 978-0-387-77316-2, 2008, Softcover
... More

Bioconductor Case Studies

Hahne, F., Huber, W. (et al.), ISBN 978-0-387-77239-4, 2008, Softcover
... More

Analysis of Integrated and Co-integrated Time Series with R

Pfaff, B., R-code for examples in the book, ISBN 978-0-387-75966-1, 2008, Softcover
... More

Morphometrics with R

Claude, J., ISBN 978-0-387-77789-4, 2008, Softcover
... More

Applied Spatial Data Analysis with R

Bivand, R.S., Pebesma, E.J. (et al.), ISBN 978-0-387-78170-9, 2008, Softcover
... More

Wavelet Methods in Statistics with R

Nason, G.P., ISBN 978-0-387-75960-9, 2008, Softcover
... More

Statistical Methods for Environmental Epidemiology with R

Peng, R.D., Dominici, F., ISBN 978-0-387-78166-2, 2008, Softcover
... More

Data Manipulation with R

Spector, P., ISBN 978-0-387-74730-9, 2008, Softcover
... More

Lattice - Multivariate Data Visualization with R

Sarkar, D., ISBN 978-0-387-75968-5, 2008, Softcover
... More

Interactive and Dynamic Graphics for Data Analysis

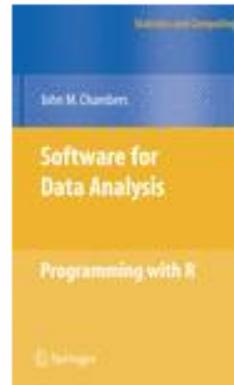
Cook, D., Swain, D.F., ISBN 978-0-387-71761-6, 2007, Softcover
... More

Bayesian Computation with R

Albert, J., ISBN 978-0-387-71384-7, 2007, Softcover
... More

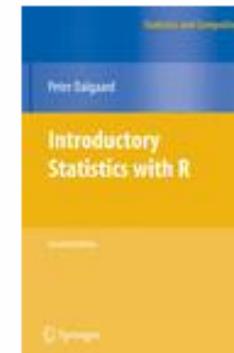
Analysis of Phylogenetics and Evolution with R

Paradis, E., ISBN 978-0-387-32914-7, 2006, Softcover
... More



Software for Data Analysis
Programming with R
Series: **Statistics and Computing**
Chambers, John M.
2008, Approx. 510 p., Hardcover
ISBN: 978-0-387-75935-7

Not yet published. Available: July 18, 2008



Introductory Statistics with R
Series: **Statistics and Computing**
Dalgaard, Peter
2nd ed., 2008, XVI, 364 p., Softcover
ISBN: 978-0-387-79053-4

Not yet published. Available: July 25, 2008

Note: both these books are now out & are excellent! If you are new to R or statistics, start with Peter Dalgaard's book. John Chamber's book assumes some experience with R.

Also see books from Chapman, Wiley, and Cambridge. Links on useR! 2008 home page above.

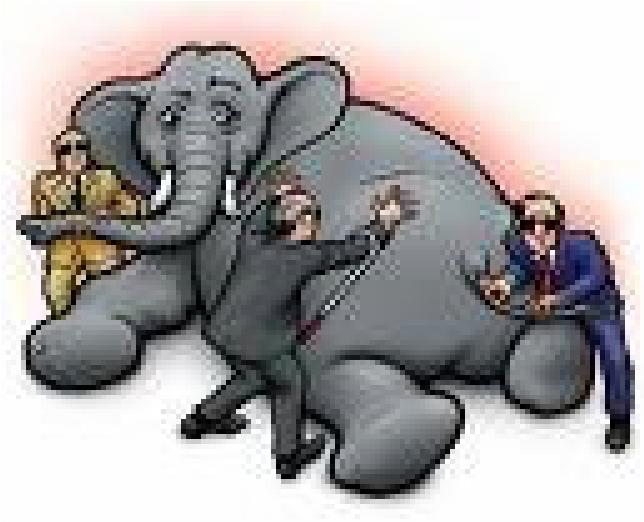


R as a Business Intelligence Tool



So just what is Business Intelligence?

From Wikipedia:



In 1989 Howard Dresner, later a Gartner Group analyst, popularized BI as an umbrella term to describe "concepts and methods to improve business decision making by using fact-based support systems."

In modern businesses the use of standards, automation and specialized software, including analytical tools, allows large volumes of data to be extracted, transformed, loaded and warehoused to greatly increase the speed at which information becomes available for decision-making.



Business Intelligence Tools

Again From Wikipedia:

The key general categories of business intelligence tools are:

- Spreadsheets
 - Reporting and querying software
 - OLAP
 - Digital Dashboards
 - Data mining
 - Process mining
 - Business performance management
- We play here.

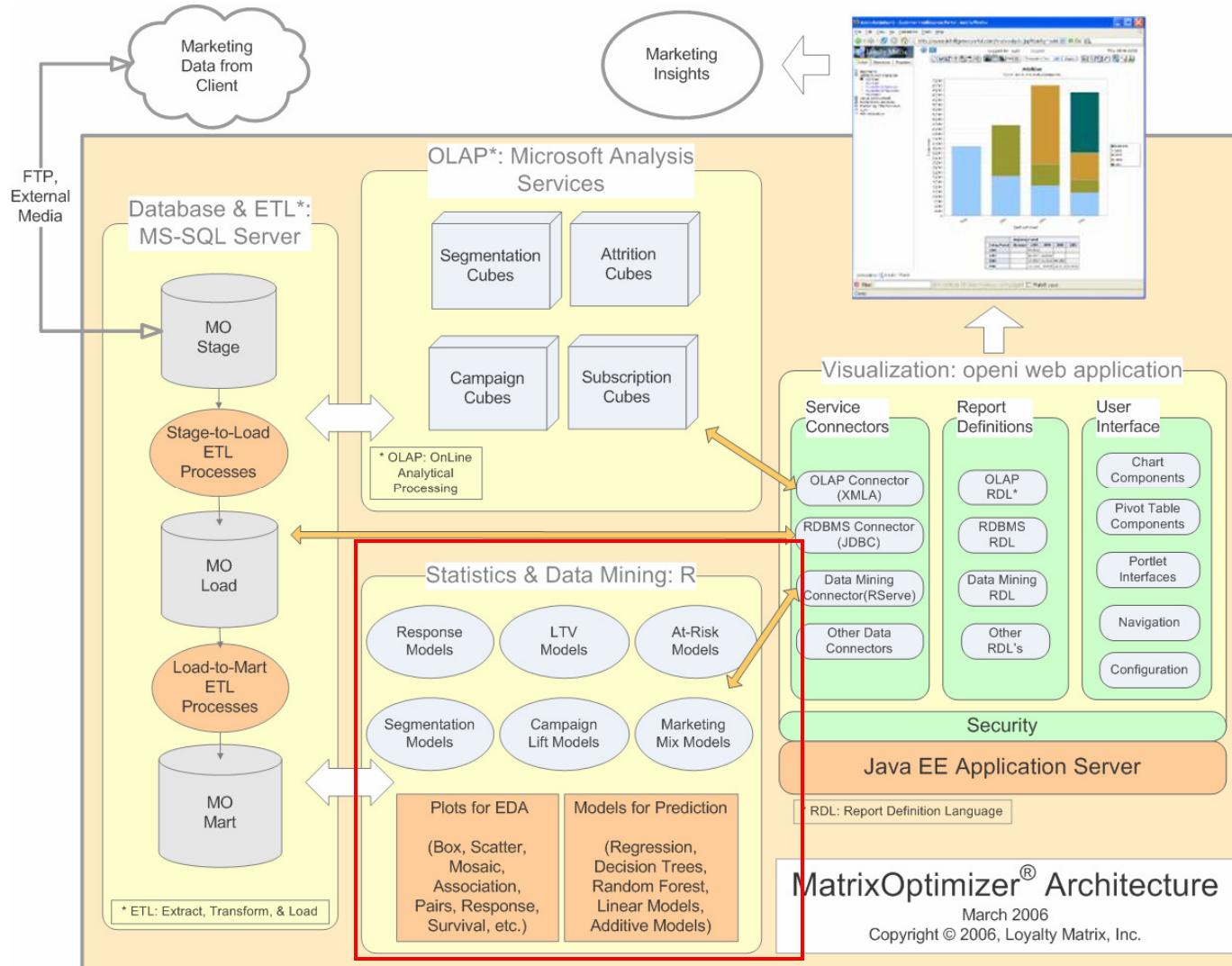


Positioning R against Traditional BI

Characteristic	Traditional BI	R & Friends
Cutting Edge Methods	- / +	+++
Naive Interactive Use	+++	- - - (some GUIs help)
Reproducible Results	- / +	+++
Massive Data Handling	++	- - - (stay tuned)
Data Base Reporting	+++	- - - (N/A)
Visualization	++	+++
Predictive Analytics	+	+++
Verifiable Methods	-	+++
Data Mining	- / ++	+++



Leverage R's Strengths in Combination w/ Classical BI





Jim's R Examples

- R Help Message Counts
- Data Profiling
- Reproducible Reporting
- Customer Segmentation



R Help Message Count & JGR Demo



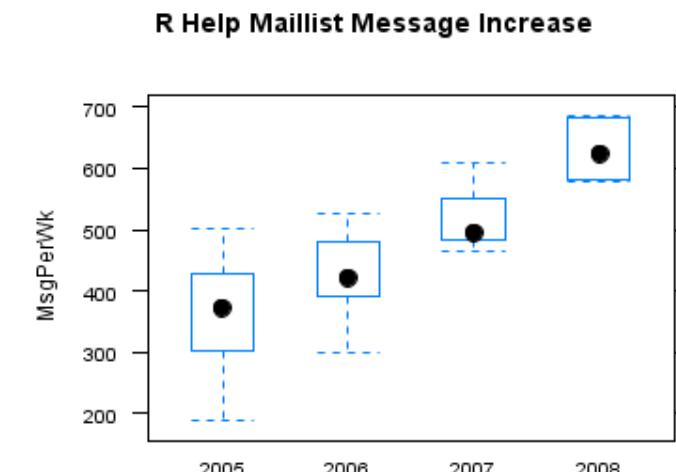
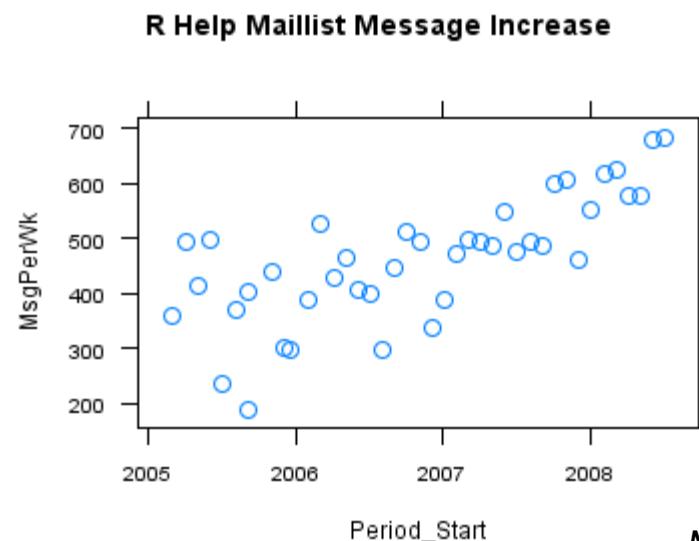
R Help Message Counts (JGR Demo)

From raw data:

period	re-sorted	messages
31 Dec 2007 to 31 Jan 2008	by thread by subject by author attachment	2451
31 Jan 2008 to 29 Feb 2008	by thread by subject by author attachment	2565
29 Feb 2008 to 31 Mar 2008	by thread by subject by author attachment	2781
31 Mar 2008 to 30 Apr 2008	by thread by subject by author attachment	2486
30 Apr 2008 to 30 May 2008	by thread by subject by author attachment	2483
1 Jun 2008 to 30 Jun 2008	by thread by subject by author attachment	2824
30 Jun 2008 to 15 Jul 2008	by thread by subject by author attachment	1417



To insights:



Note: See appendix for code & link to source & data files

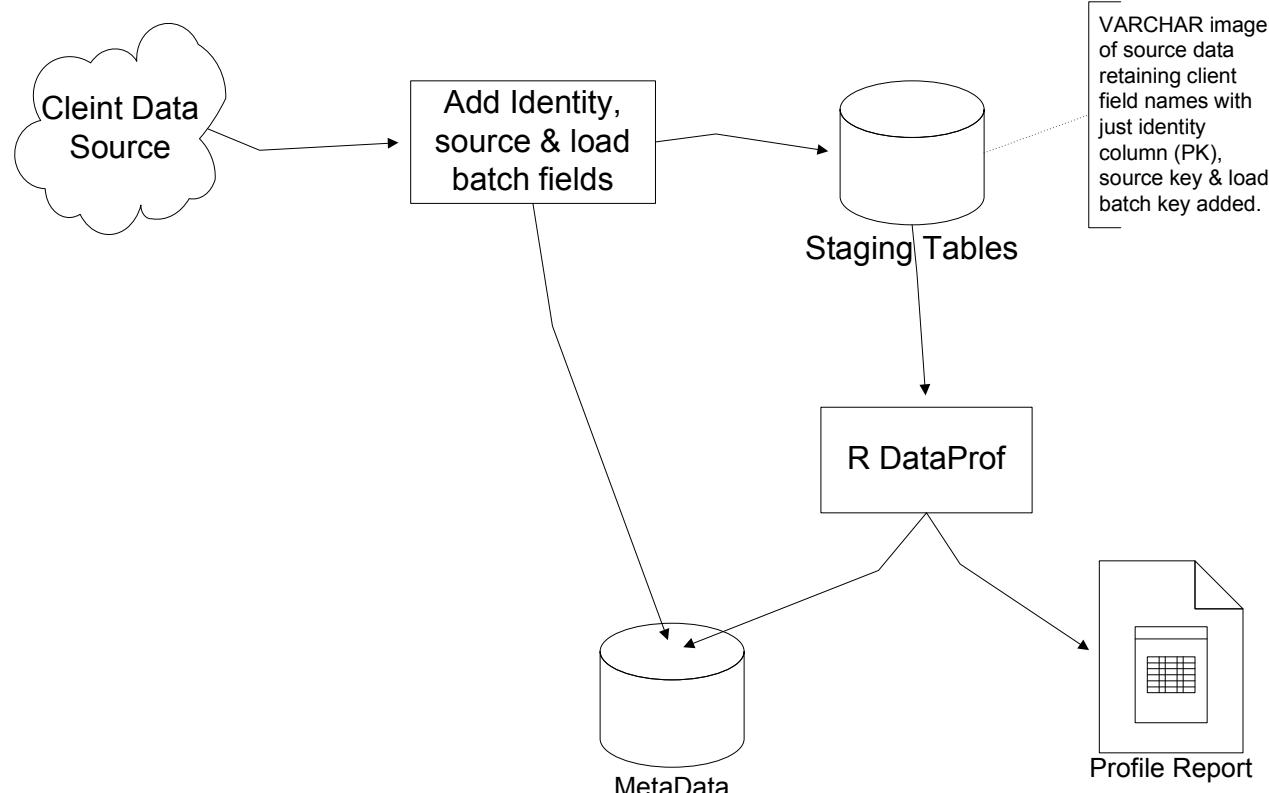


Data Profiling with R



Data Profiling in R – Basics

- Reference: *Data Quality – The Accuracy Dimension* by Jack E. Olson
- Where we profile:



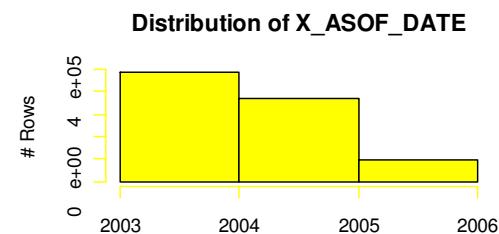


Data Profiling in R – Profiler Column Output

Header: Database details for field

Summary Counts & %'s
Empty, Numeric & Date only for character strings

AMA_Stage . RECEIVABLE_TXN . X_ASOF_DATE						19	varchar(8000)
Rows	Nulls	Distinct	Empty	Numeric	Date		
#	3,861,249	2,908,244	28,564	0	0	953,005	
%	100.00	75.32	0.74	0.00	0.00	100.00	
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.		
2003-01-14	2003-08-12	2003-12-29	2004-02-12	2004-09-08	2005-10-17		
Head: NA NA NA NA NA NA							



Summary Stats if numeric or date

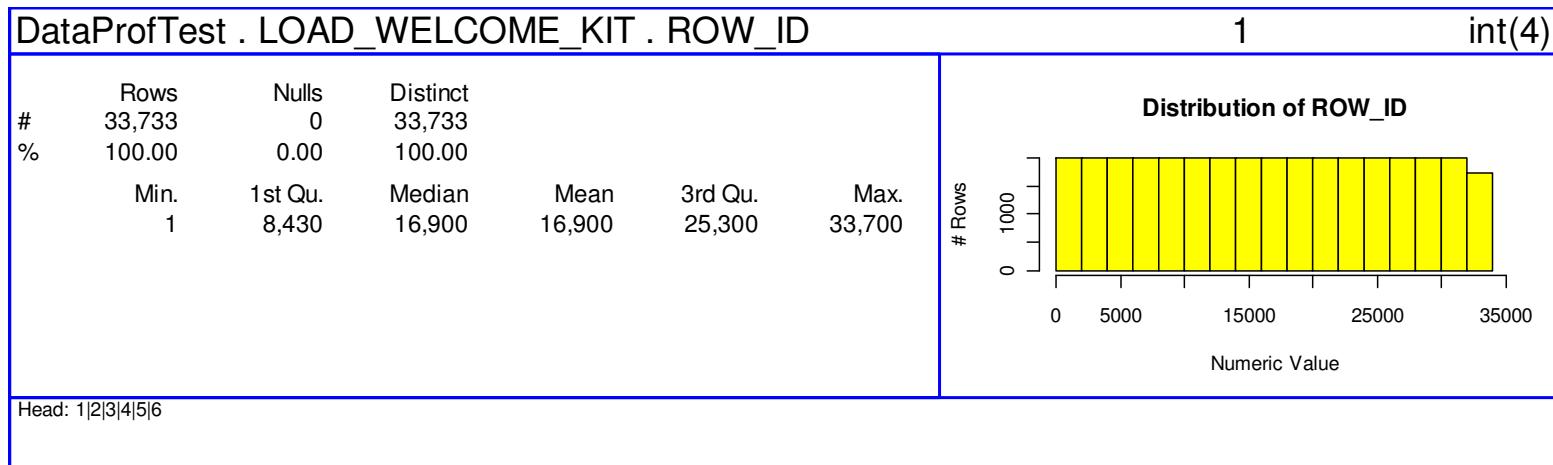
Appropriate plot type

Footer: Notes about field

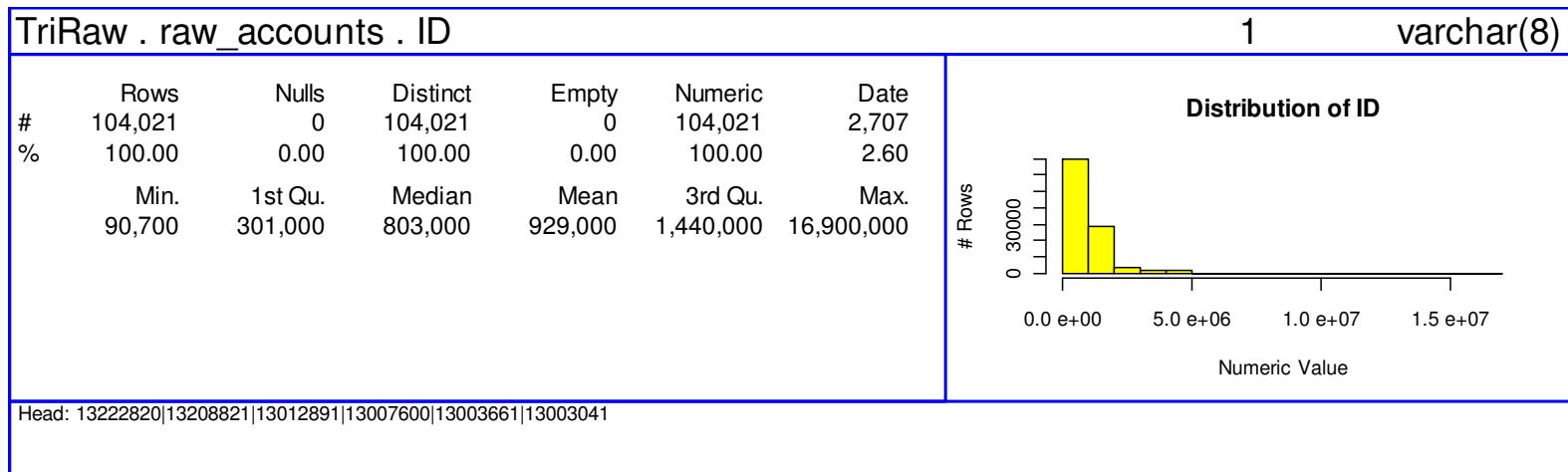


Data Profiling in R – Examples (1)

- A Surrogate Key



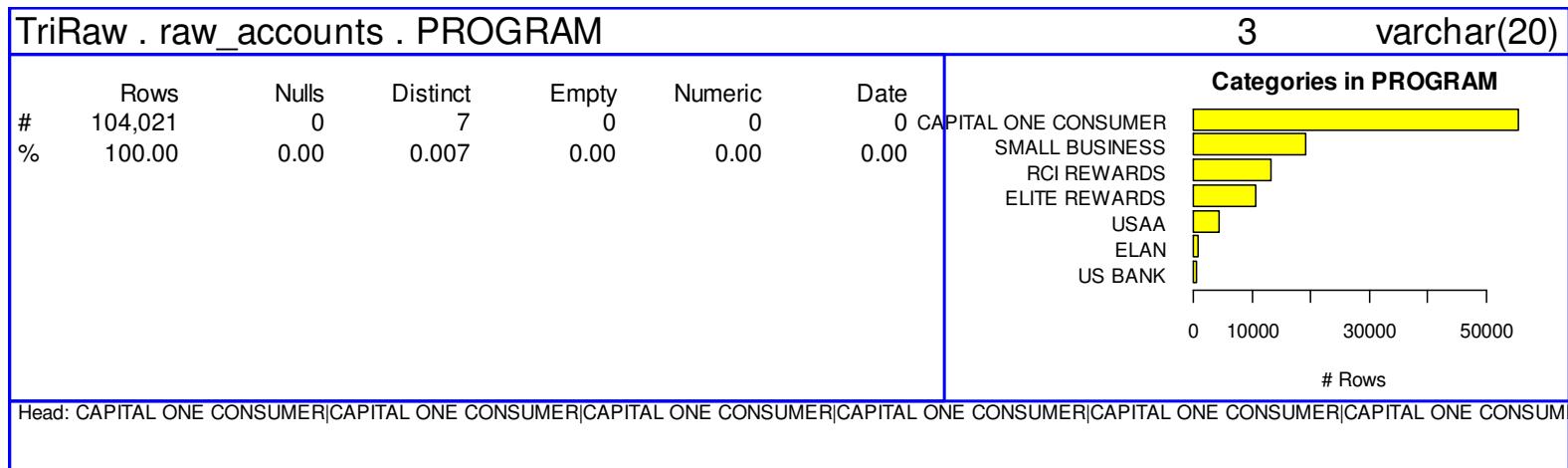
- Probable Business Key



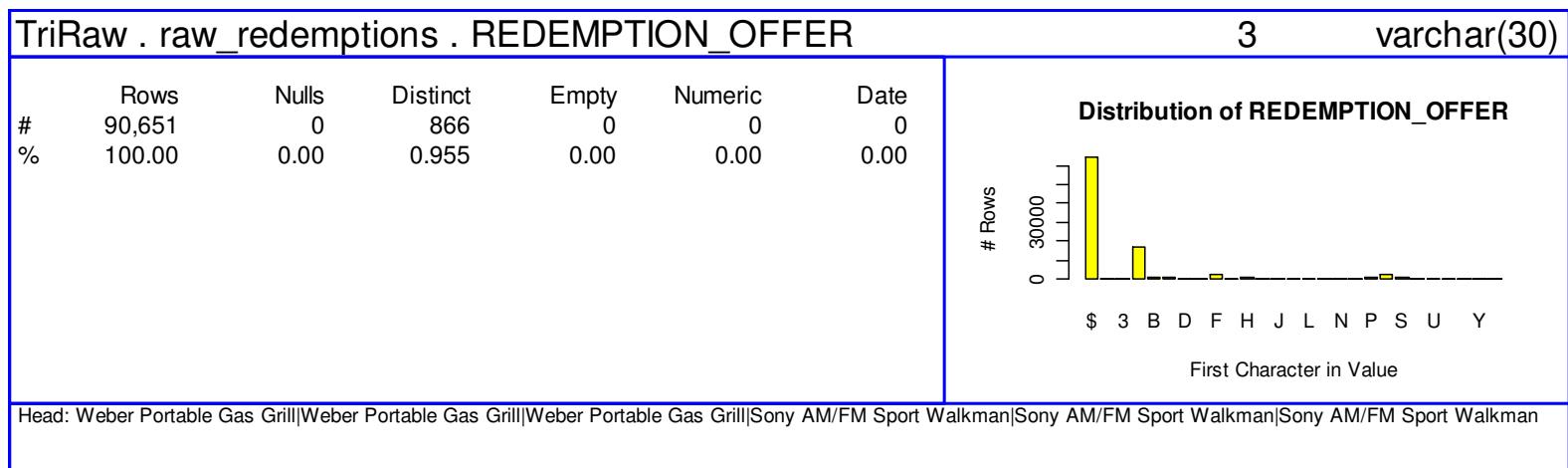


Data Profiling in R – Examples (2)

- A Few Categories



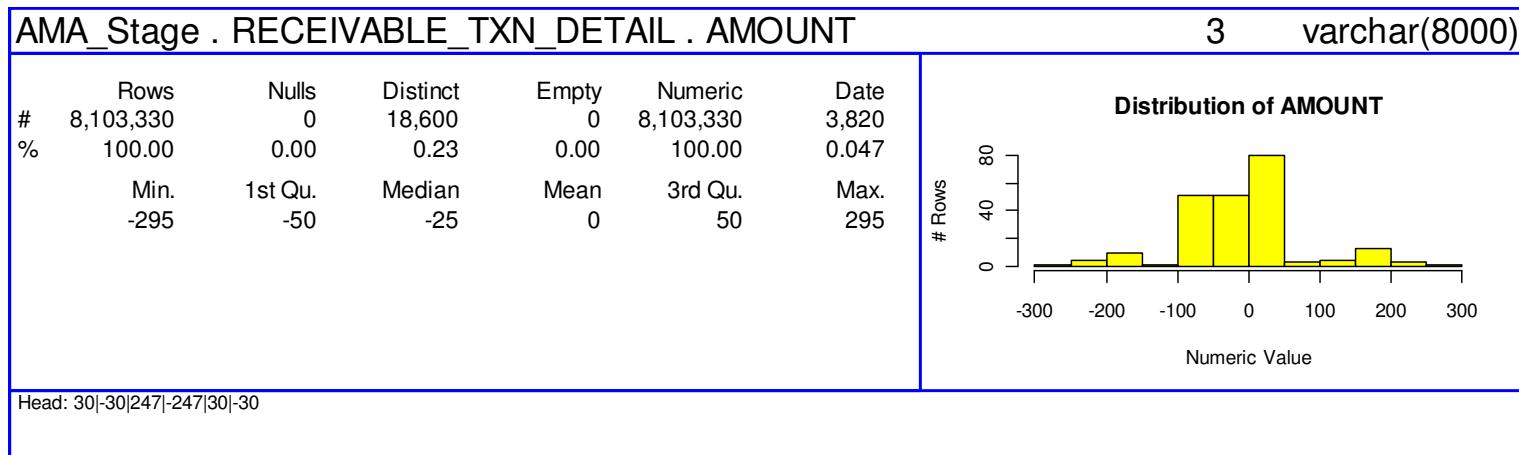
- Many Categories



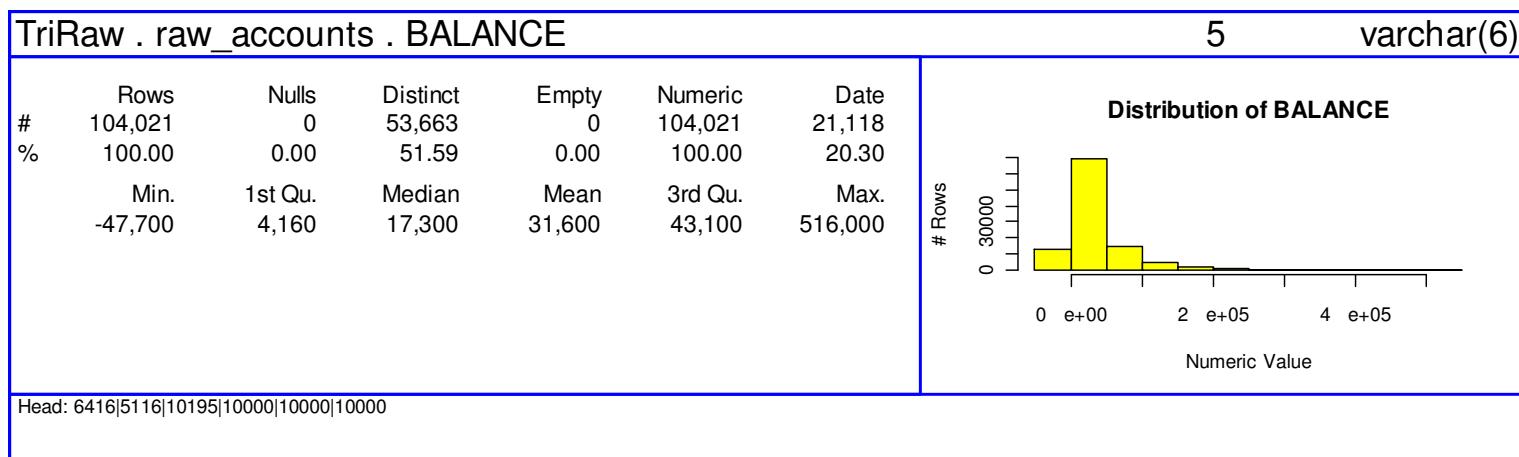


Data Profiling in R – Examples (3)

- Numeric Value



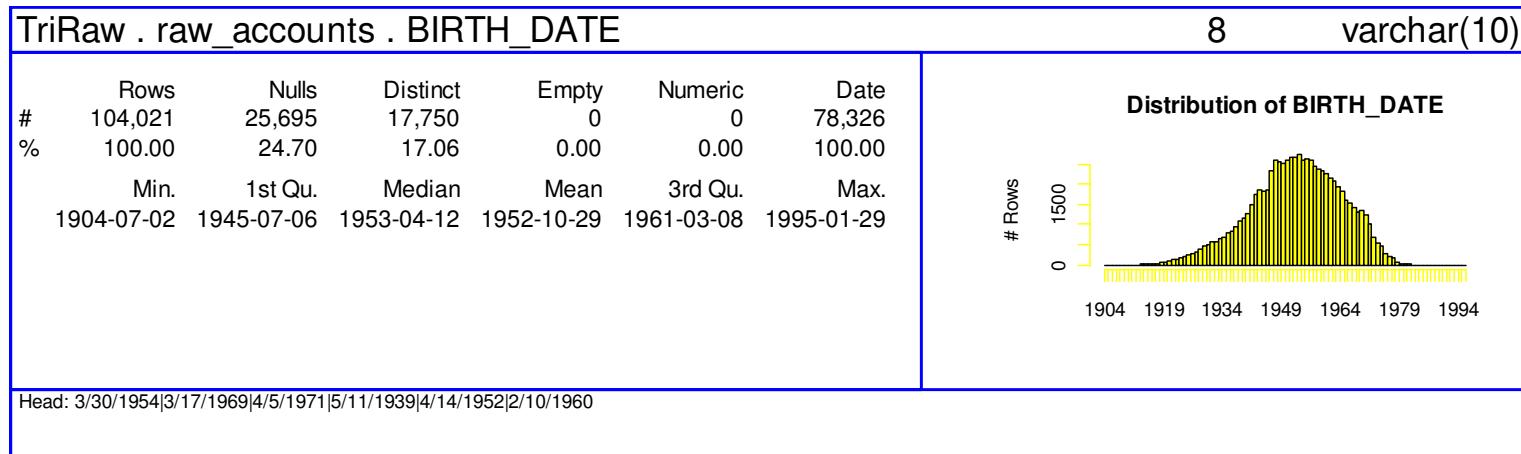
- Another Numeric Value



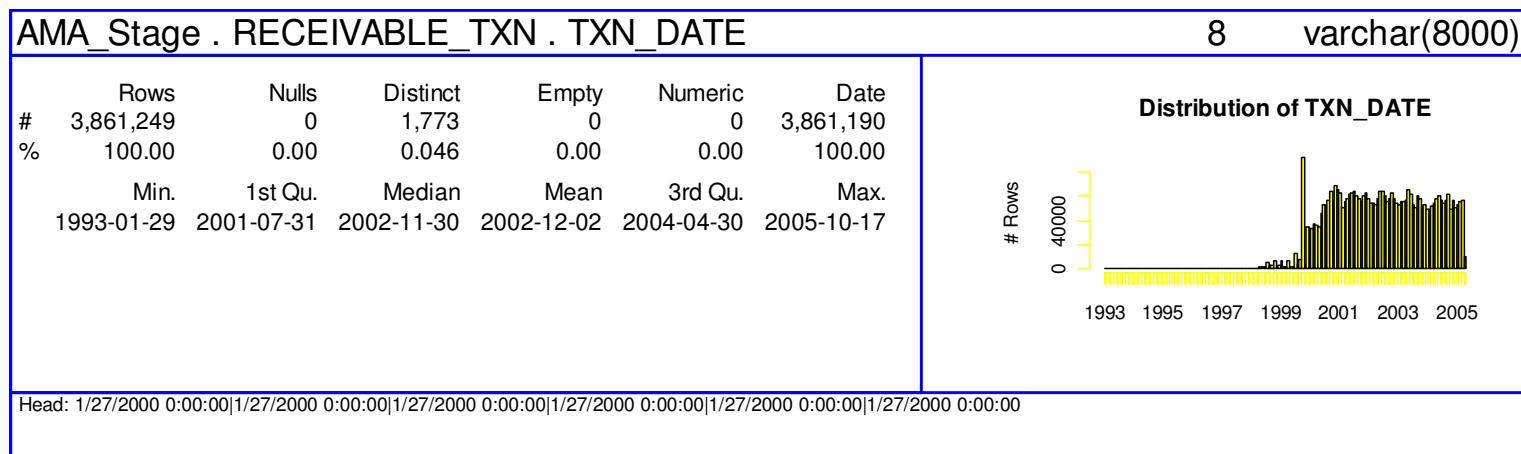


Data Profiling in R – Examples (4)

- Reasonable Dates



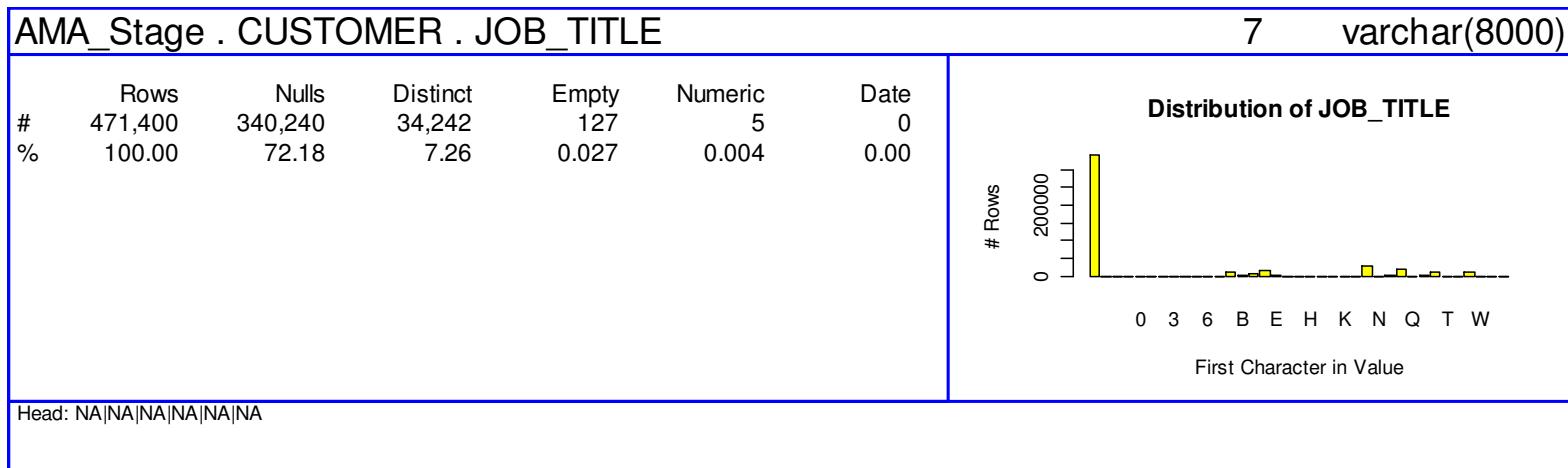
- Unusual Dates



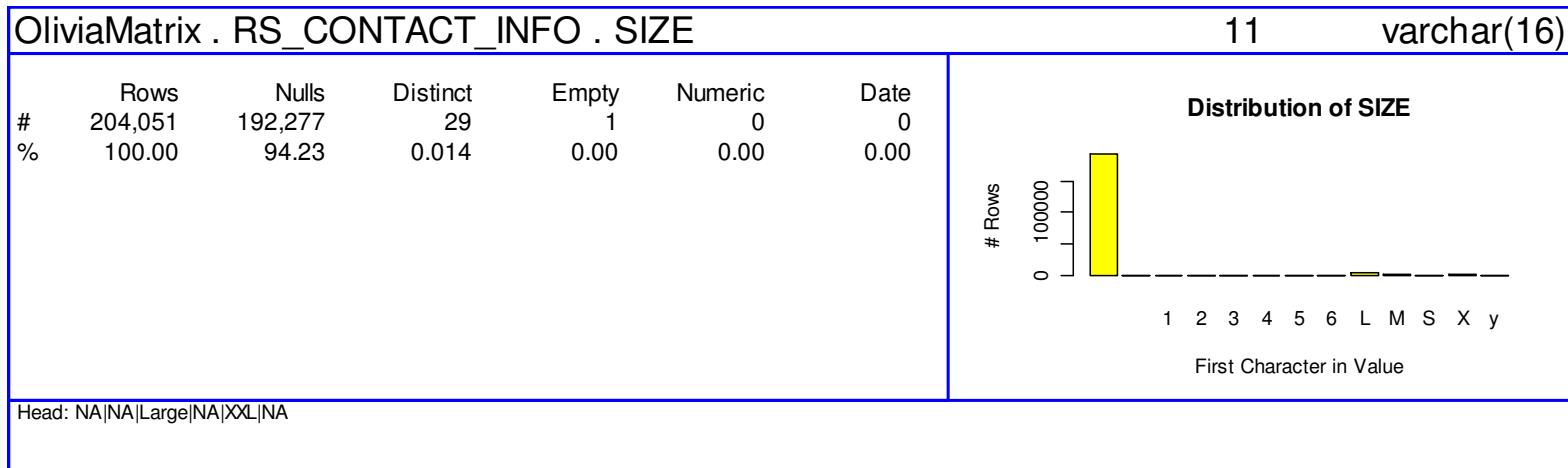


Data Profiling in R – Examples (5)

- Customer Job Title not too useful



- T-shirt Size also not reliable





Reproducible Reporting



What's Reproducible Reporting?

- Use methods from “Reproducible Research”
 - a published paper should include data and analysis code to produce tables, charts, and conclusions
 - See www.stat.washington.edu/jaw/jaw.research.reproducible.html
- Using R for “Reproducible Reporting” leverages:
 - R’s connectivity
 - R’s advanced analysis & visualization
 - odfWeave to produce OpenOffice text document
 - Business analysts can edit
 - Export to PDF or .doc
 - Based on Sweave (LaTeX) work by Fritz Leisch .
- Following example is actual data quality assurance report we run every month as part of a large data warehouse update



Actual Output Report

Responsys

MO QA 2.0 Counts for Load As Of 2008-07-04

Summary

Note missing week & very unusual Upgrade ramp-up highlighted below.

Missing Data:

WeekOfSatThe	NumberNew
2008-06-21	0

Unusual Trends:

Data Set	p.value	AboveBelow	Severity
Upgrades	0.001268	++++++	HIGHLY SUSPECT
SessionsConsumed	0.053724	+++++-+---	Unusual

Where AboveBelow shows weeks with counts above, +, or below, -, mean of all weeks.

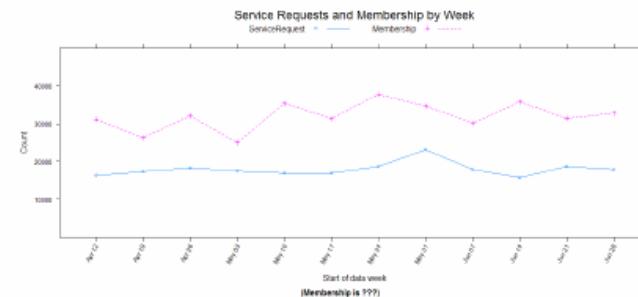
Membership

Notes:



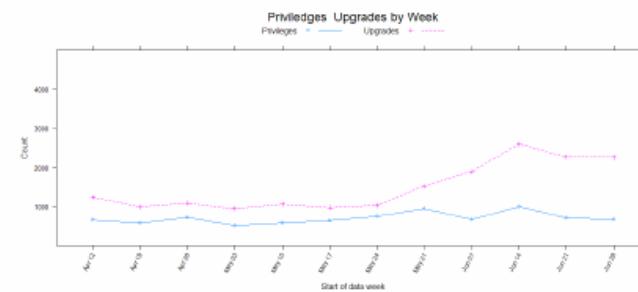
Service Requests and Membership Events

Notes:



Miscellaneous Events

Notes:





odfWeave “source”

Responsys

MO QA 2.0 Counts for Load As Of \\$expr{mo_AsOfDate}

Summary

Missing Data:

```
<<MissingTable, echo = FALSE, results = xml>>=
if(!is.null(nrow(Missing))) {
  odfTable(Missing, useRowNames = FALSE, colnames = colnames(Missing))
} else {
  odfCat("No completely missing data elements for any weeks.")
}
}
```

Unusual Trends:

```
<<UnusualTable, echo = FALSE, results = xml>>=
if(!is.null(nrow(Suspects))) {
  #op <- options()
  #options(digits = 3)
  odfTable(Suspects, useRowNames = TRUE,
           colnames = c("Data Set", colnames(Suspects)))
  #options(op)
} else {
  odfCat("No unusual trends observed.")
}

if(!is.null(nrow(Suspects))) {
  odfCat("Where AboveBelow shows weeks with counts above, +, or below, -,
         mean of all weeks.")
}
}
```

Membership

Notes:

```
<<MembCounts, echo = FALSE, fig = TRUE>>=
siD16.5, 3, 200)
print(
xyplot(NumberNew ~ WeekOfSatThe_, CountsByWeek,
       type = "p", ylim = c(0, 50000),
       auto.key = list(title = "New and Lost Members by Week", space =
"top",
                     columns = 2, lines = TRUE),
       xlab = "Start of data week", ylab = "Count",
       scales = list(x = list(at = CountsByWeek$WeekOfSatThe_, rot = 60)))
)
}
```

In-line “sweave” expression

Conditional table

Runs Test Exceptions in a table

Lattice X-Y Plot



Unsupervised Clustering

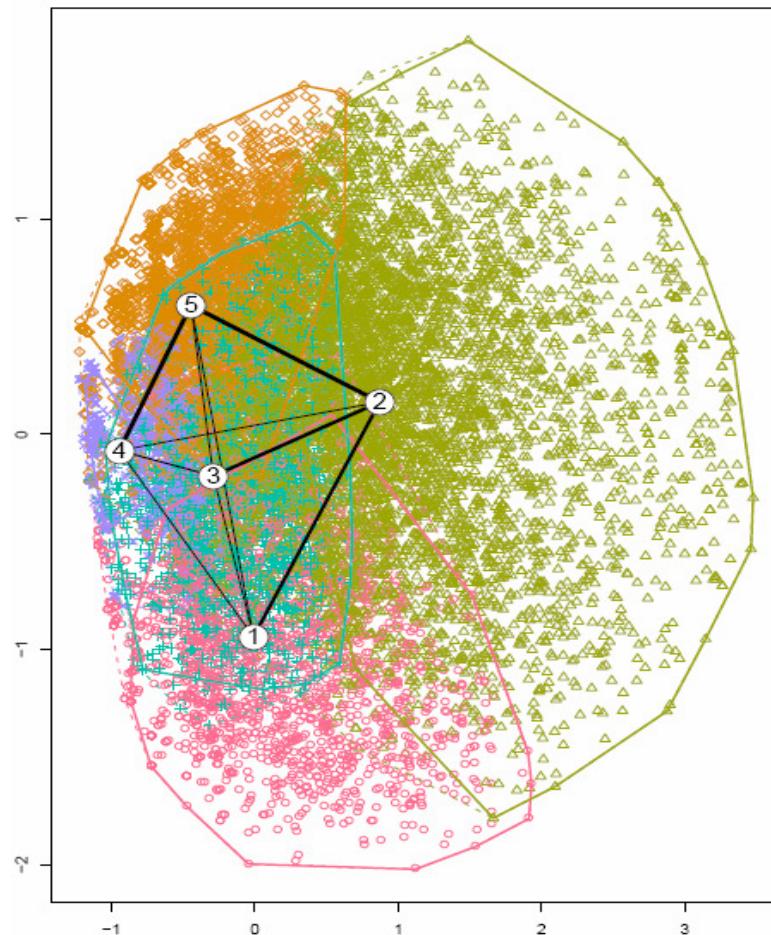
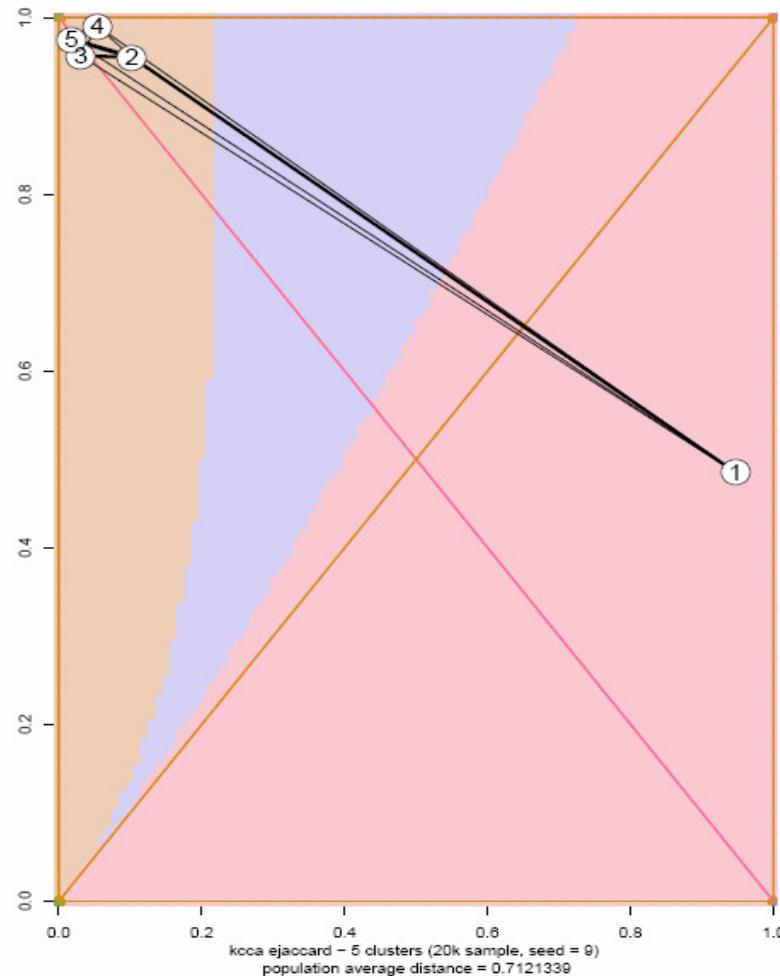


Prospect Segmentation – Problem Statement

- A tech company surveying prospects
 - ~ 20 k respondents
 - ~ 35 check box type questions covering
 - Respondent's role
 - Hardware type
 - Interests
 - Applications
- Use Fritz Leisch's flexclust package
 - K-Centroids Cluster Analysis (KCCA)
 - Jaccard distance best suited for "preference" survey
 - do 4 runs with # clusters = 3 through 8
 - look for stable & meaningful clusters
 - see: <http://www.statistik.lmu.de/~leisch/>
- Goal – assign future respondents to actionable segment
 - marketing action
 - marketing message



Customer Segmentation – Separation Plots





Customer Segmentation - Segments

kcca ejaccard – 5 clusters (20k sample, seed = 9)

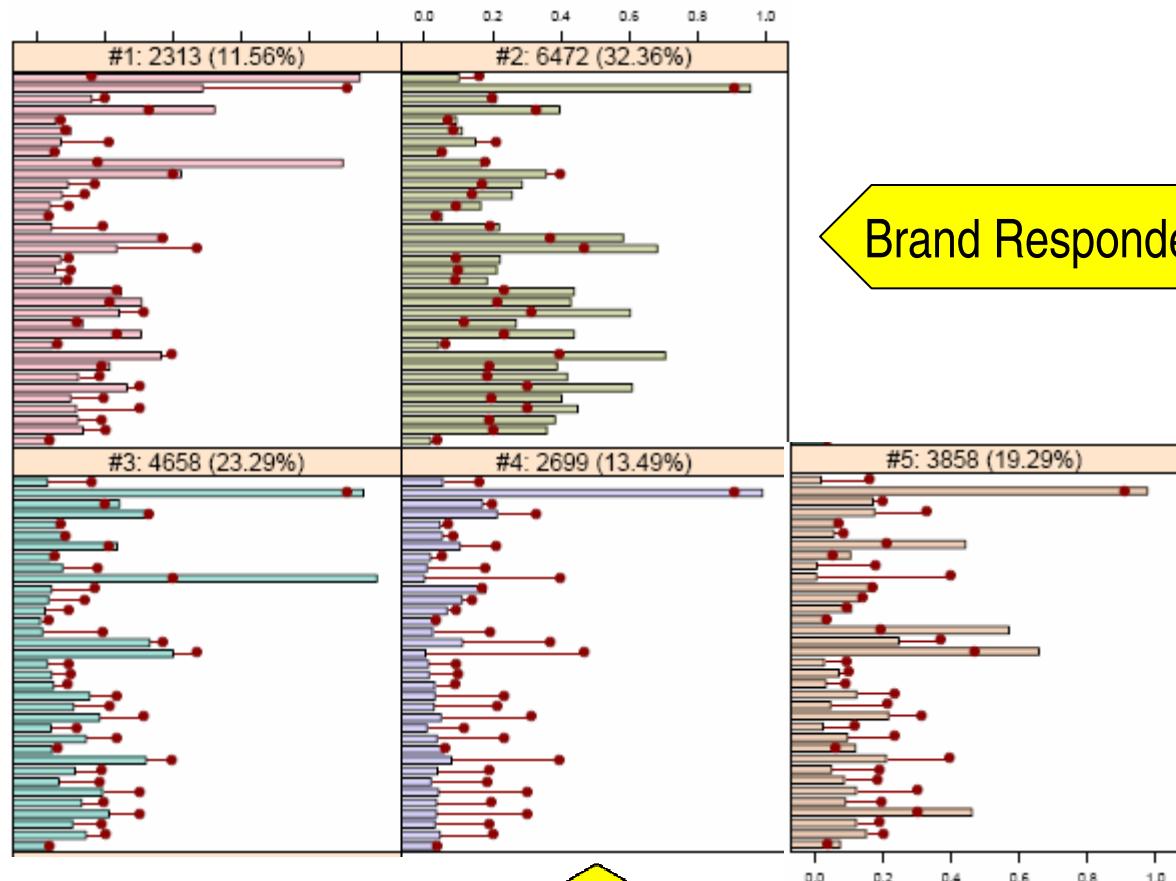
Loyalist A

Brand Responder

Loyalist B

Student

Brand Non-
responder





Mike's R Examples



What is a good model?

“All models are wrong. Some models are useful.”

– George E.P. Box

A good statistical model is

- **Intuitive**
- **Estimable**
- **Actionable**

Beyond providing insight, a good model suggests how to improve a system: “buy more of x and less of y .”

A good model drives decisions.



What is a good model for a baseball hitter?





What is a good model?

*"a hitter should be measured by his success in that what he is trying to do, ... **create runs**. It is startling, when you think about it, how much confusion there is about this."*

-Bill James, quoted in
Moneyball (p.76)

Runs scored per game	~	At Bats, Walks, Hits (Singles, Doubles, Triples, HRs), Sac Flies, Hit by Pitch
----------------------	---	--



What is a good model for a baseball hitter?

Components of our baseball hitter model:

Runs scored per game (per team)

At Bats, Walks, Hits (Singles, Doubles, Triples, HRs),
Sac Flies, Hit by Pitch (per team)

Data source:

baseball-databank.org

Actionable Result:

Identify the most valuable hitters in the league.

Our tools:

R, RMySQL



What is a good model for a baseball hitter?

Query the database to populate our R “data frame”

```
library(RMySQL)
```

```
con <- dbConnect(dbDriver('MySQL'),
                  user='mdriscol',
                  password = 'mypass',
                  host = 'localhost',
                  dbname = 'bbdb')

resultSet <- dbSendQuery(con,
                         "select AB, BB, H, 2B, 3B, HR, SF, HBP, G, R
                          from teams
                          where yearID between 2000 and 2005")

teamStats <- fetch(resultSet, n=-1)

attach(teamStats)
```



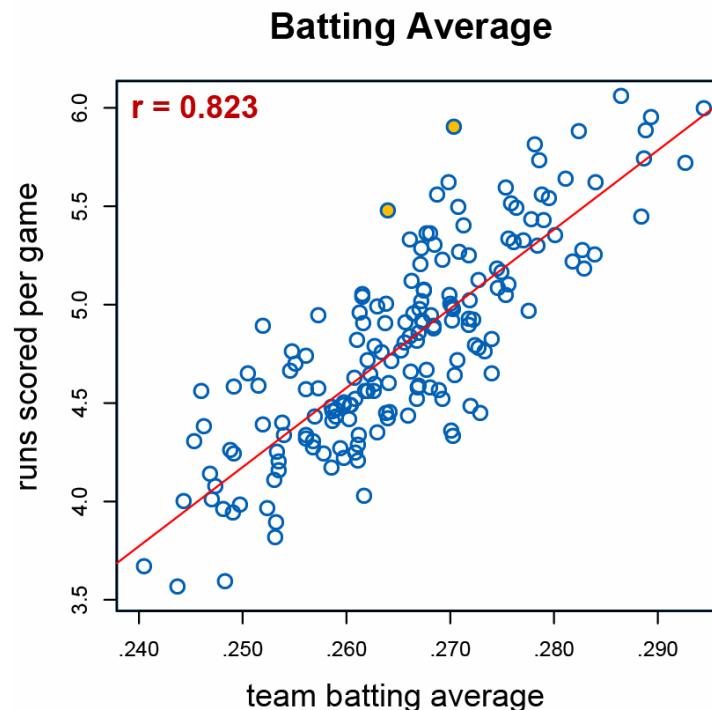
What is a good model for a baseball hitter?

MODEL 1

Runs per game ~ Batting Average

Batting Average = Hits / At
Bats

```
rpg <- R/G  
bavg <- H/AB  
plot(rpg ~ bavg)
```





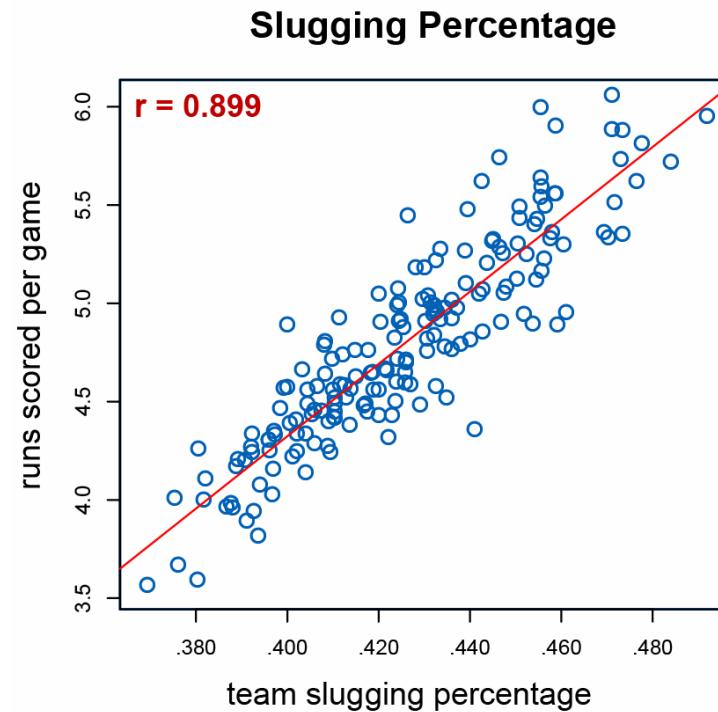
What is a good model for a baseball hitter?

MODEL 2

Runs per game ~ Slugging %

Slugging % = Total Bases / At Bats

```
rpg <- R/G  
slug <- (H + 2B + 3B*2 + HR*3)/AB  
plot(rpg ~ slug)
```



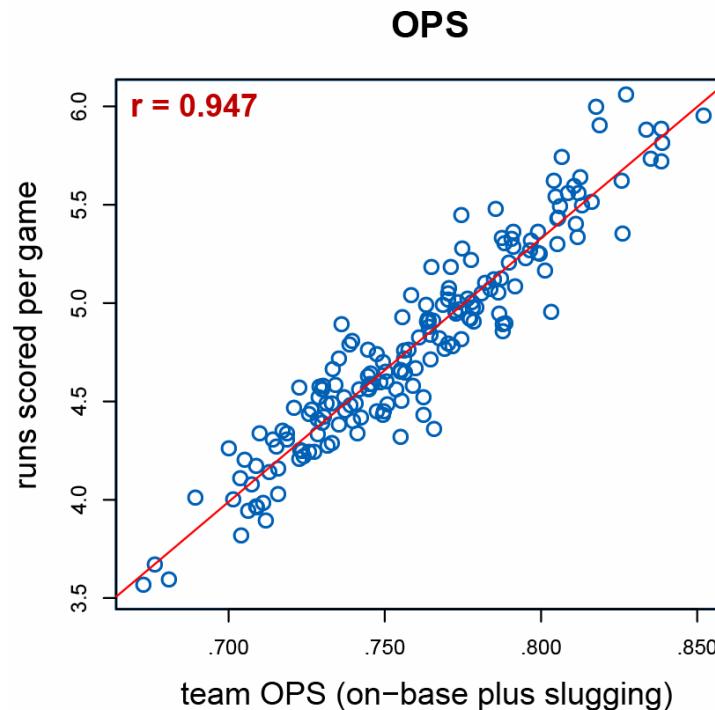


What is a good model for a baseball hitter?

MODEL 3

Runs per game ~ OPS

OPS = On Base Plus Slugging

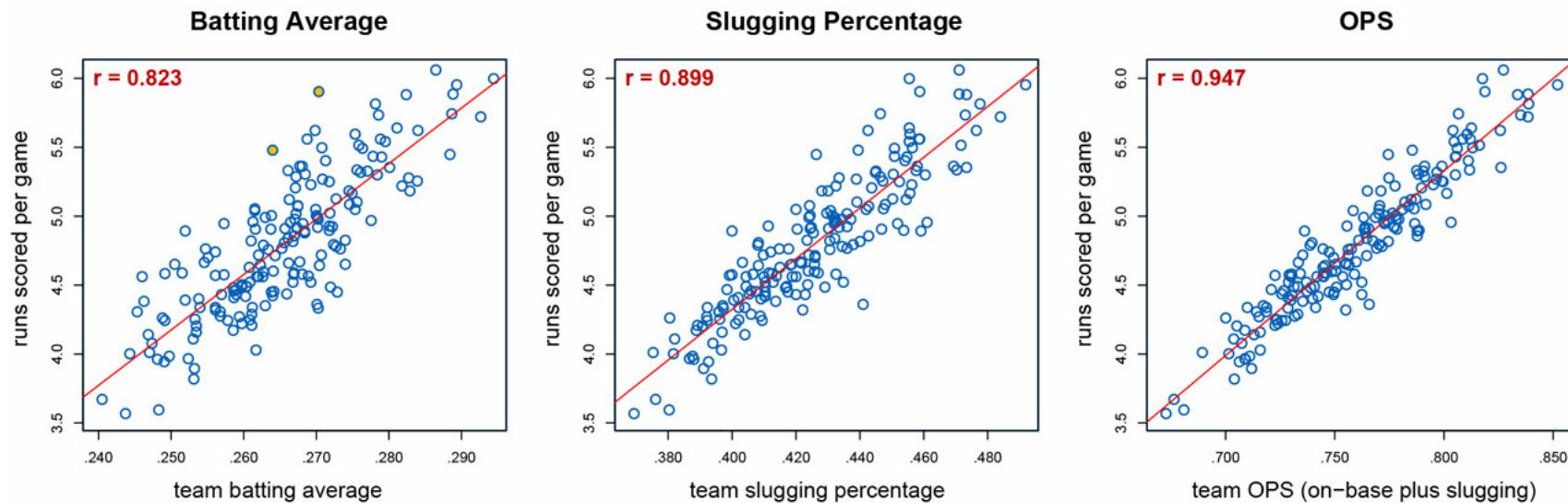


```
onbase <- (H+BB+HBP)/(AB+BB+SF)  
OPS <- onbase + slug  
plot(rpg ~ OPS)
```



What is a good model for a baseball hitter?

Conclusion: **OPS** is the best predictor of **runs**



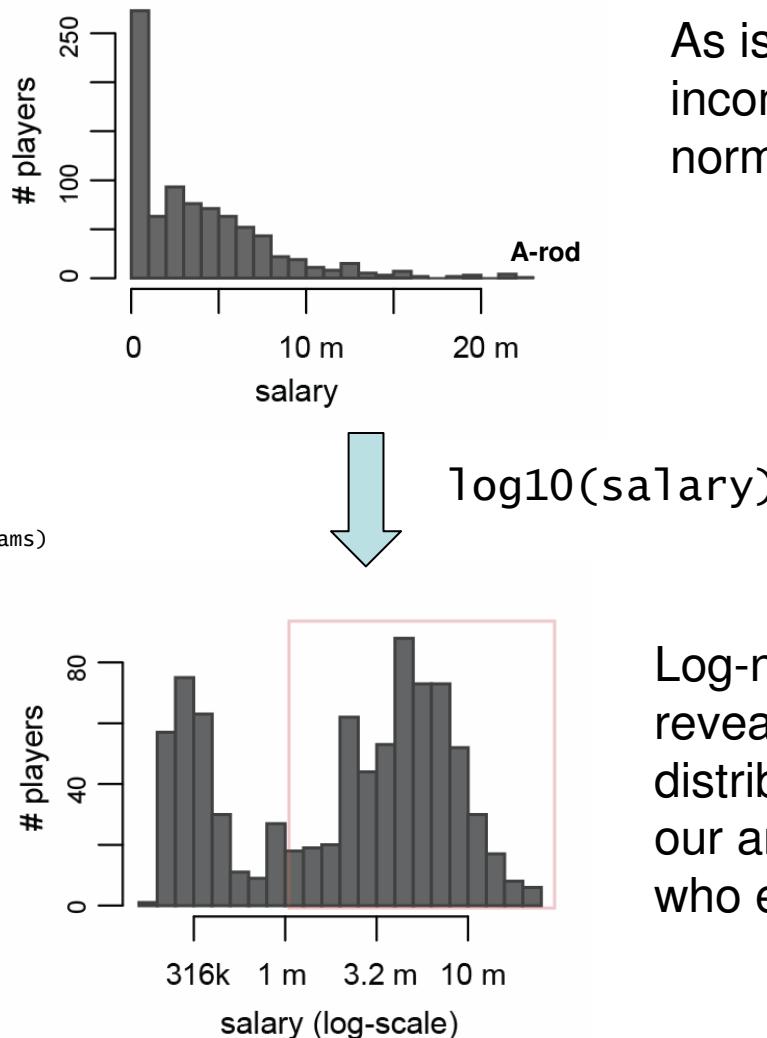
Hitters with high **OPS** are the most valuable hitters



Does OPS predict a player's salary?

We take 836 data points for players between 2000 and 2005.

```
batters <- sql.fetch.all(con,
  'select yearID,
  salary,
  b.*'
from salaries s
  join master m using (idxLahman)
  join batting b using (idxLahman)
  join teams t on (t.idxTeams = b.idxTeams)
  where s.idxTeams = t.idxTeams
and yearID between 2000 and 2005
group by yearID, playerID
having AB > 300')
```



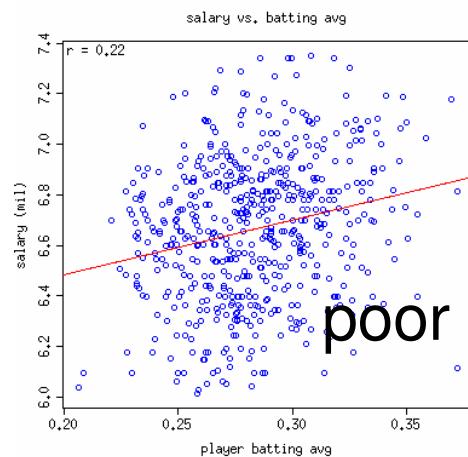
As is common with income data, non-normally distributed

Log-normalizing the data reveals a bi-modal distribution; we'll restrict our analysis to players who earn > \$1m annually

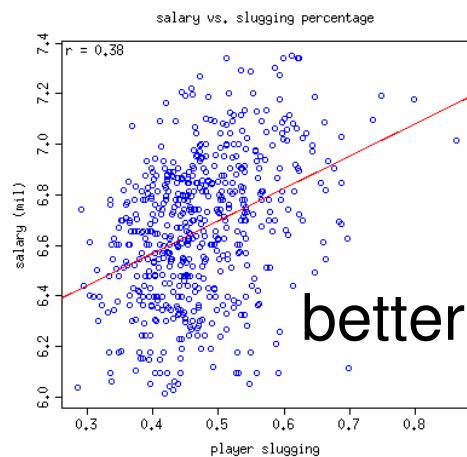


Does OPS predict a hitter's salary?

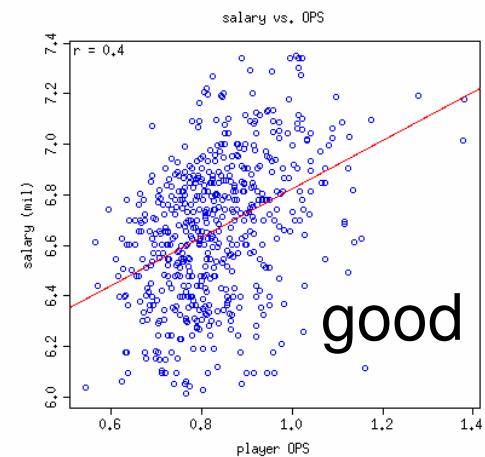
\$\$ v \text{ batting avg}



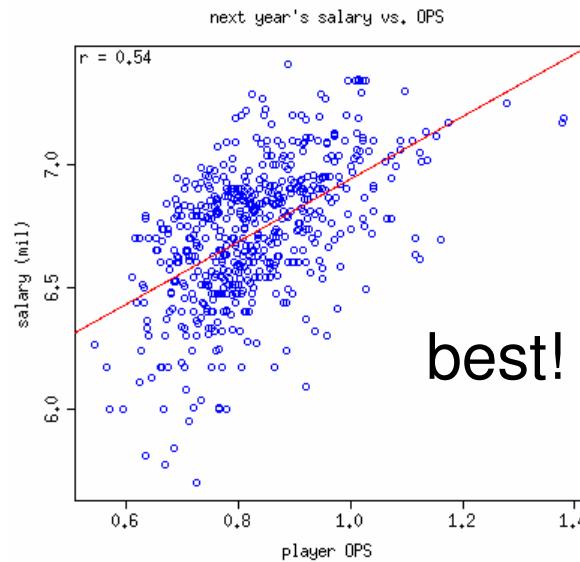
\$\$ v \text{ slugging}



\$\$ v \text{ OPS}



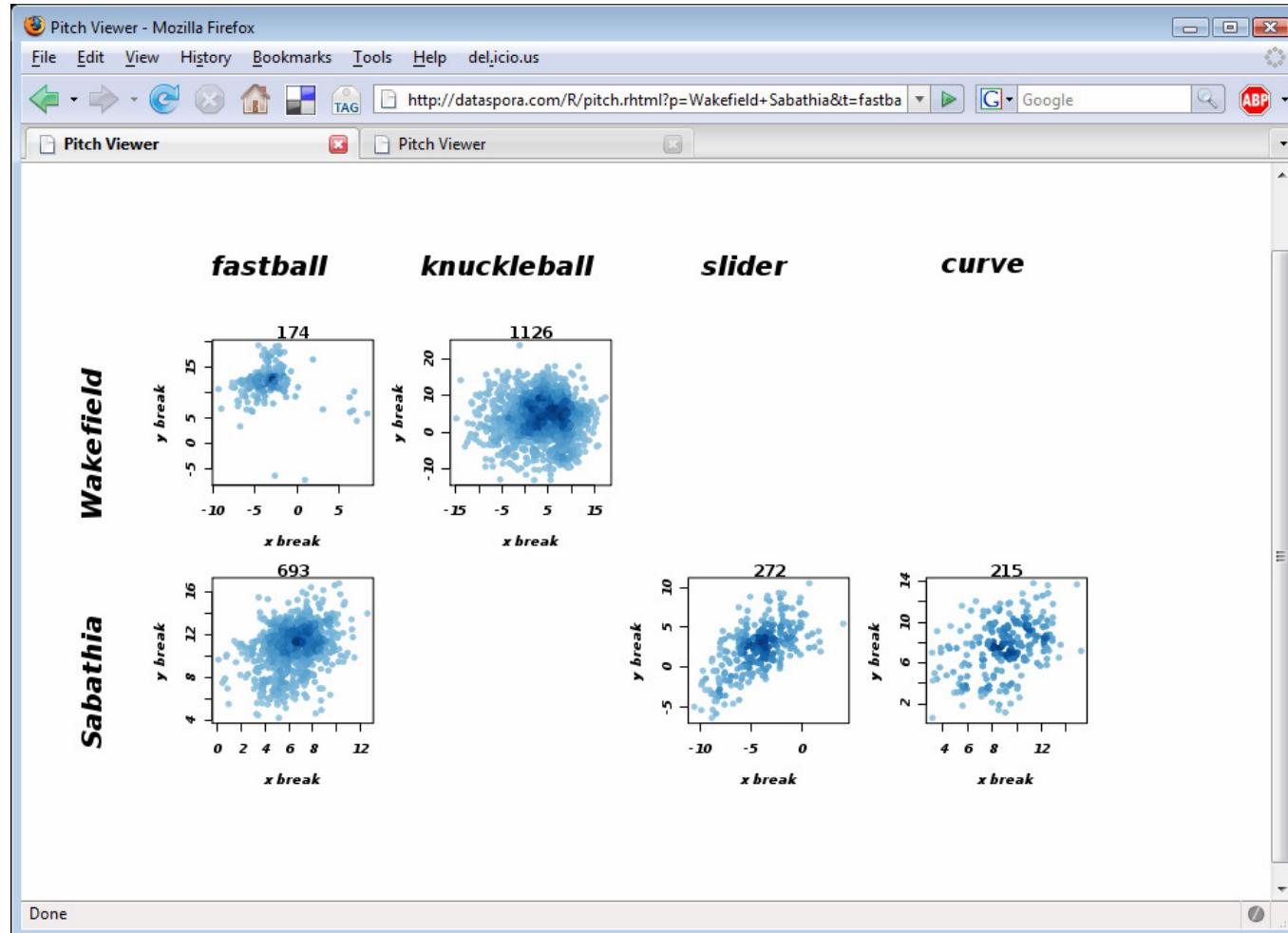
\$\$ v \text{ last year's OPS}



A hitter's previous year's
OPS predicts his salary better
than any other batting statistic



Web Dashboards with Rapache



<http://www.dataspora.com/R>



John's Demo



infoRming enviRonmental policy

John J. Oram
San Francisco Estuary Institute



Some questions you may have ...

(and that I hope to answer)

- What is the San Francisco Estuary Institute?
 - Why is an environmental scientist presenting at a Business Intelligence gathering?
- How is R informing environmental policy?
- How can R help me?



- Non-profit founded 1986 to foster development of scientific understanding necessary to protect and enhance San Francisco Estuary
- Fill niche between environmental science and policy
- Provide holistic integration of information needed to support existing or alternative management activities



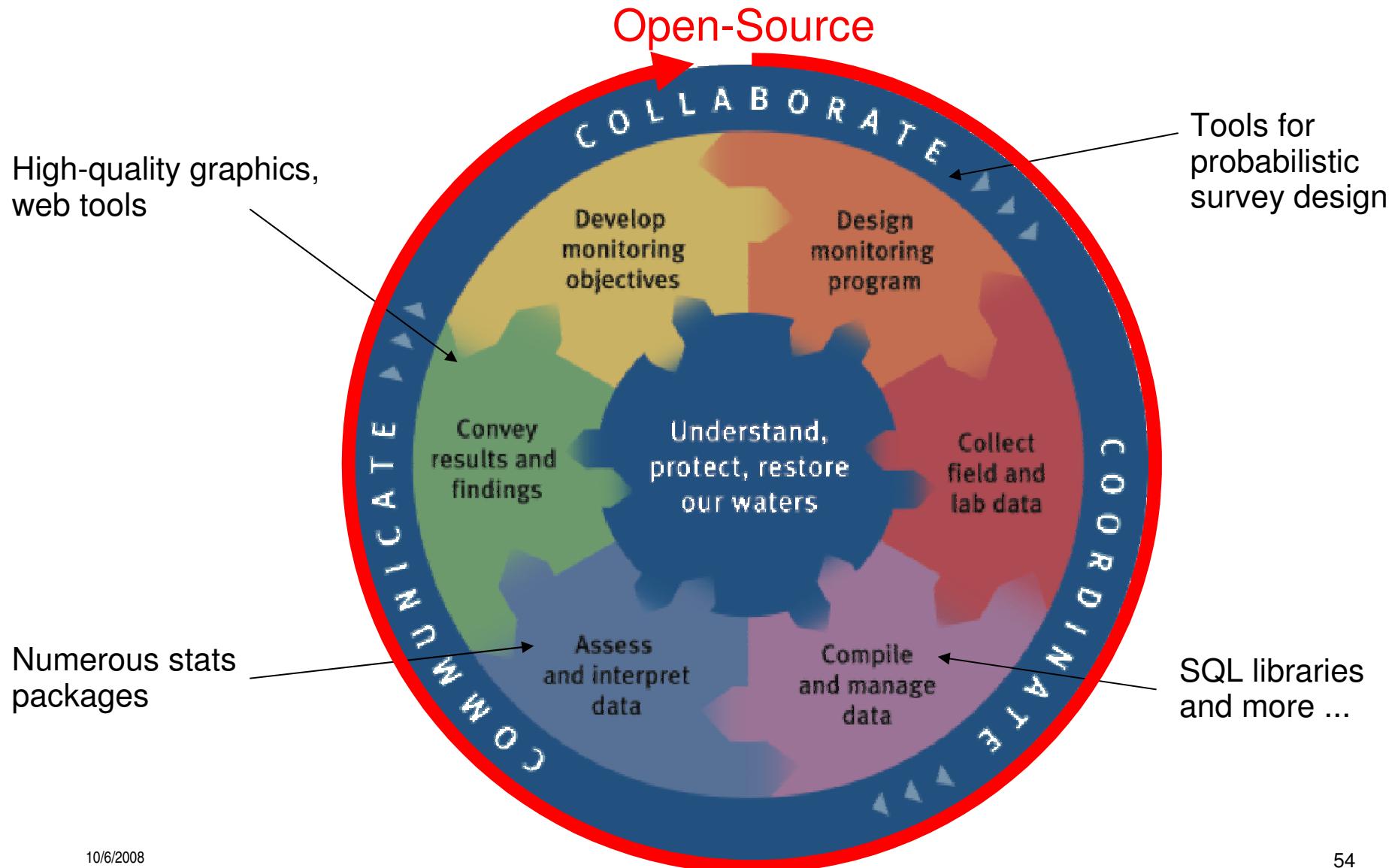
- My Roles ...
 - To **integrate** data from multiple programs
 - To develop mechanistic and/or probabilistic **models**
 - To **evaluate** potential future scenarios
- My Tools ...
 - **R**, Matlab, FORTRAN,...
 - GIS (prefer open-source options)
 - SQL, Access
 - Web technologies (for info dissemination)

R
R

R



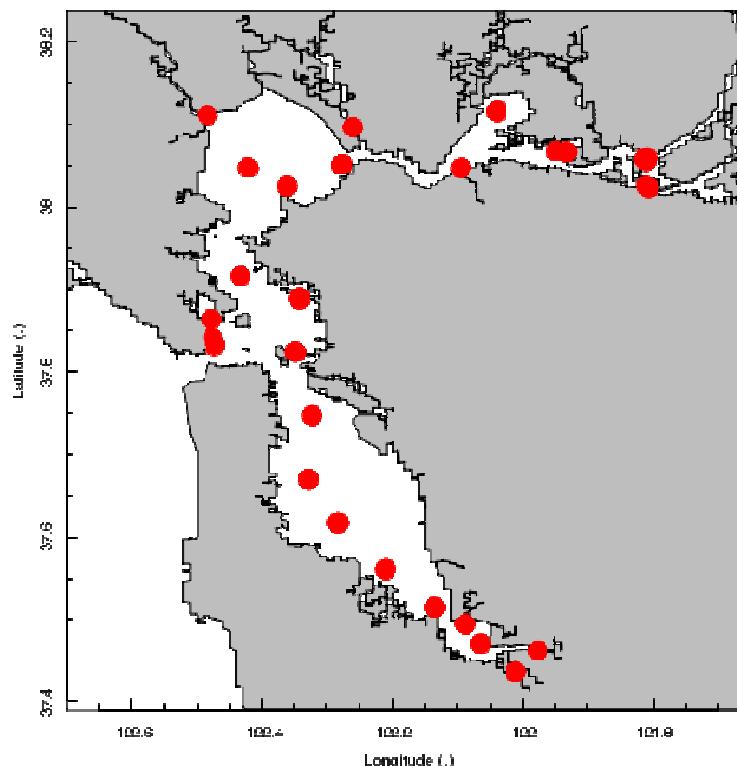
How is R Informing Environmental Policy?





An Example : Bay Water Quality

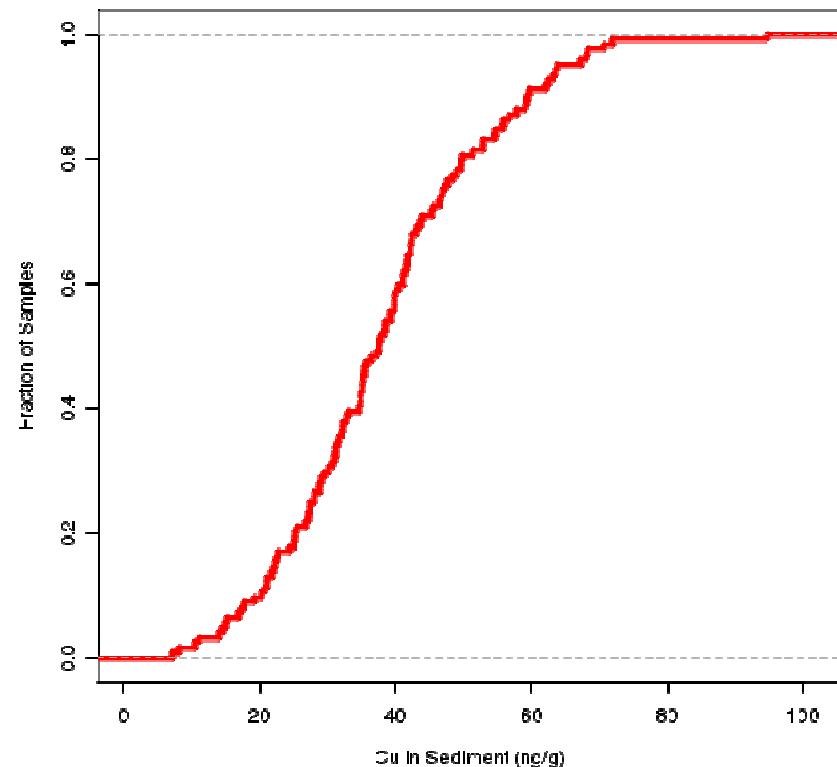
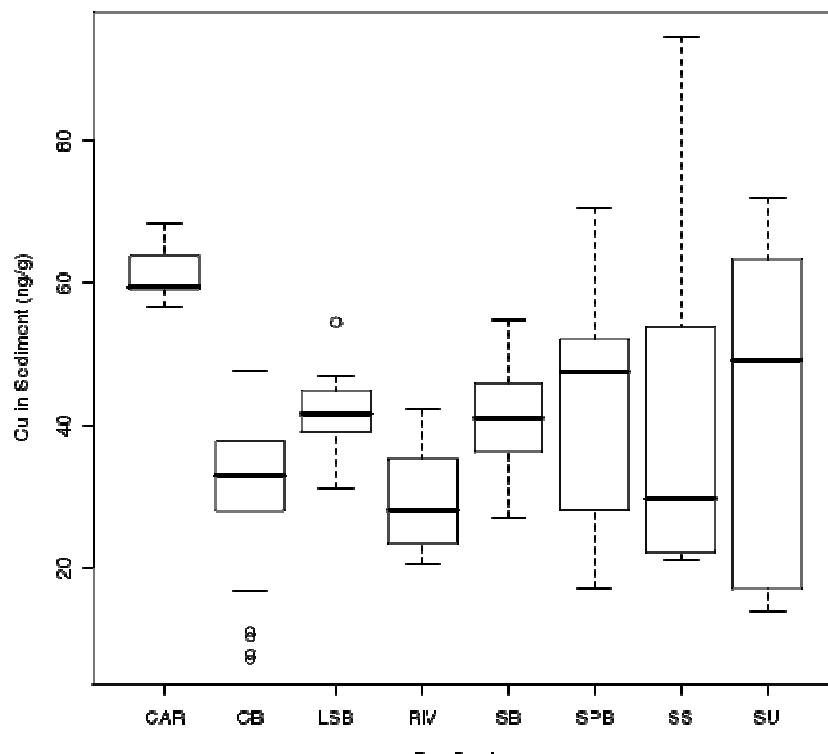
- Regional Monitoring Program has monitored water and sediment since 1993.
- Fixed stations = limited statistical power





An Example : Bay Water Quality

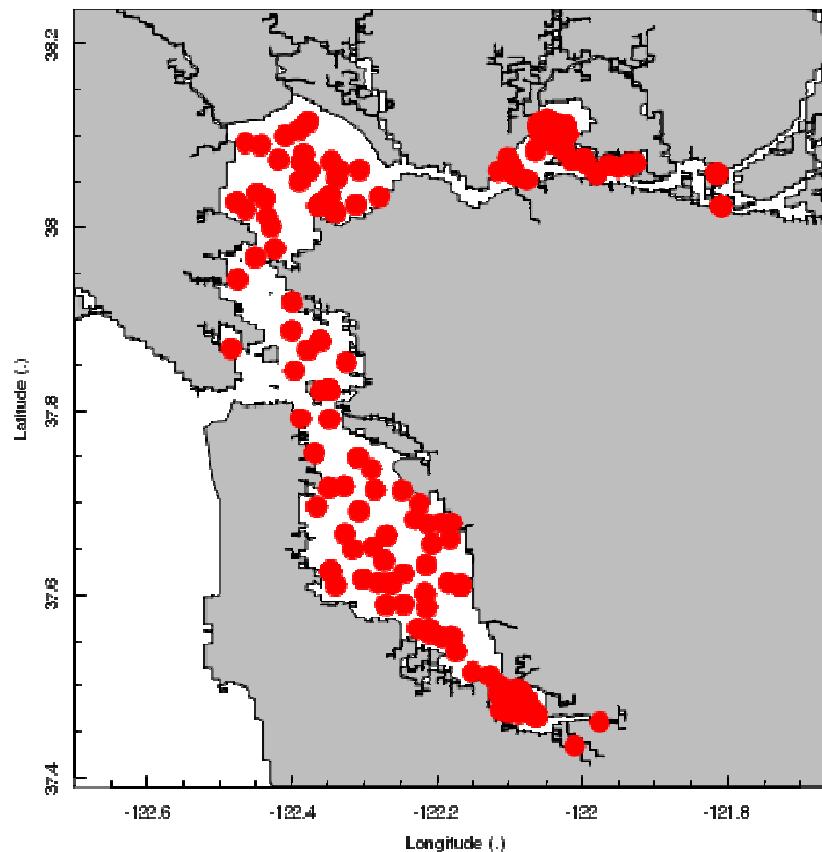
- Regional Monitoring Program has monitored water and sediment since 1993.
- Fixed stations = limited statistical power





An Example : Bay Water Quality

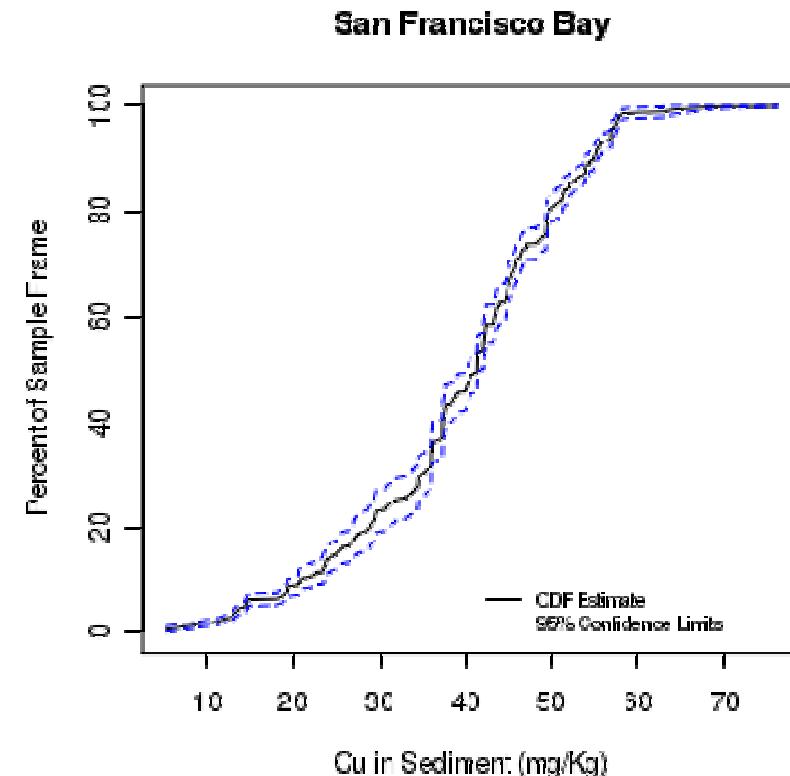
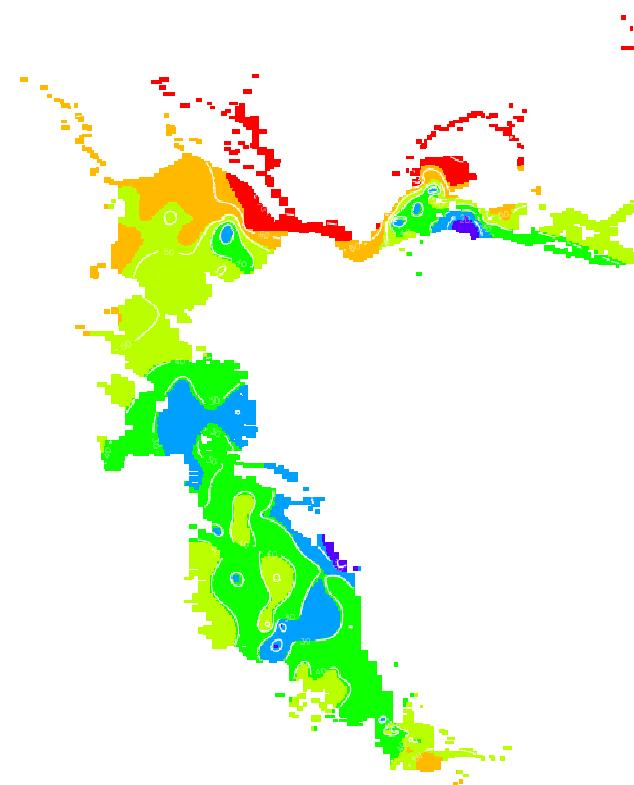
- A probabilistic survey design implemented in 2002
- Stratified random stations = unbiased statistical power





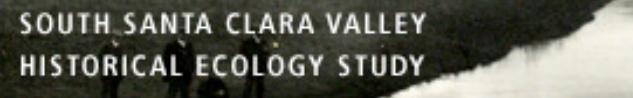
An Example : Bay Water Quality

- A probabilistic survey design implemented in 2002
- Stratified random stations = unbiased statistical power





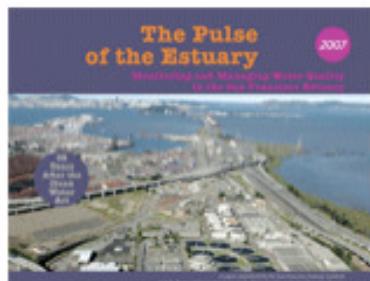
impRoving infoRmation dissemination



[View details →](#)



[View details →](#)



[Download the 2007 Pulse →](#)



[Stream
Fishes](#)

[View details →](#)



[Sign up now →](#)



[See report →](#)



[View tool →](#)



Volume 109 Issue 1 September 2007 ISSN 0167-6369

Environmental Research

A Multidisciplinary Journal of Environmental Sciences, Ecology, and Public Health

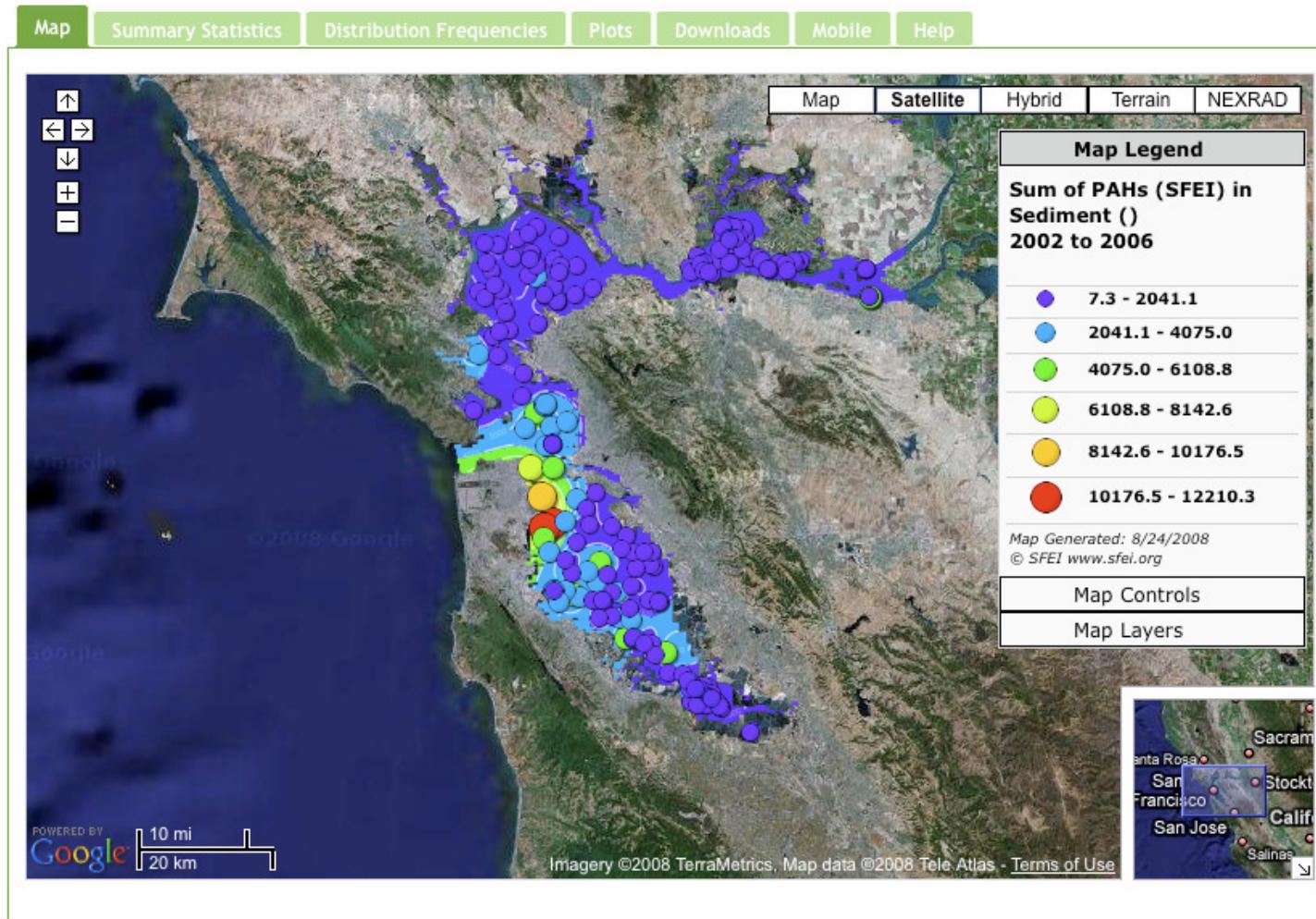
Editor-in-Chief: Ellen K. Silbergeld

Special Issue: Pollutants in the San Francisco Bay Estuary





impRoving infoRmation dissemination





impRoving infoRmation dissemination

[Map](#)[Summary Statistics](#)[Distribution Frequencies](#)[Plots](#)[Downloads](#)[Mobile](#)[Help](#)

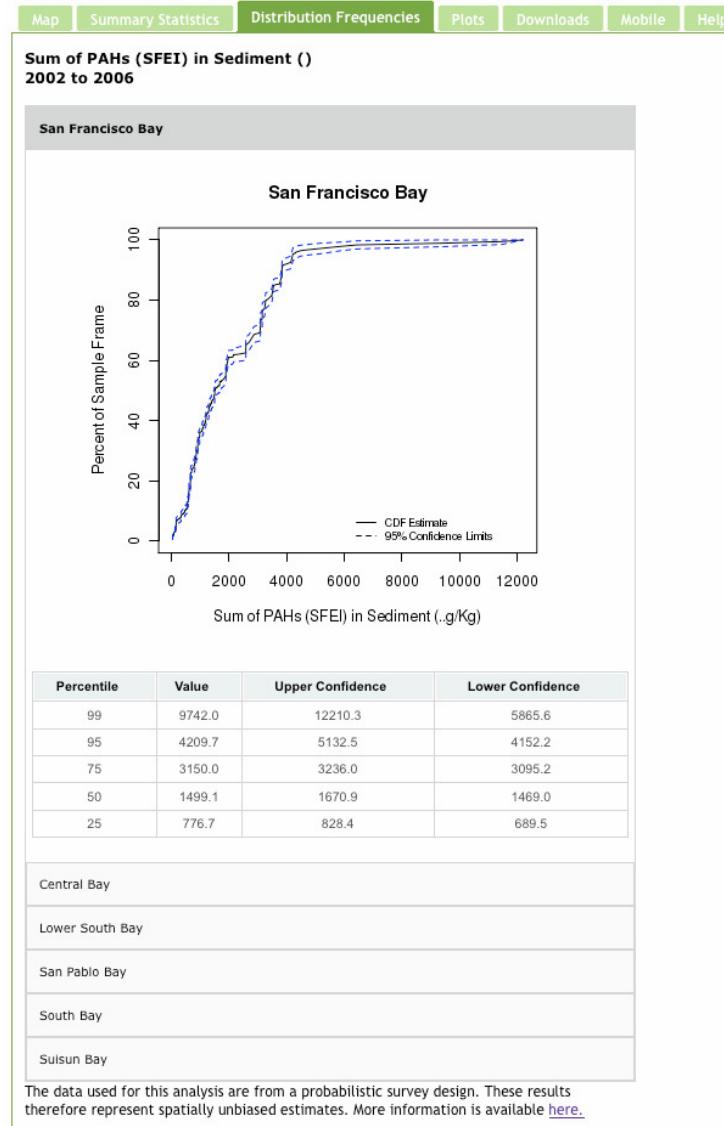
Sum of PAHs (SFEI) in Sediment () 2002 to 2006

Region	Frame Area (sq km)	Mean	Upper Confidence	Lower Confidence	Standard Deviation	N	Min Value	Max Value
San Francisco Bay	896.433	2075.5	2197.1	1954.0	1796.0	392	35.6	12210.3
Central Bay	396.442	3324.7	3589.8	3059.5	1890.7	80	154.7	12210.3
Lower South Bay	7.642	1371.1	1428.0	1314.1	462.4	80	411.4	2268.5
San Pablo Bay	226.821	845.9	891.0	800.8	335.2	80	129.1	2136.7
South Bay	185.171	1734.1	1882.0	1586.3	999.7	76	113.9	4686.3
Suisun Bay	80.357	237.4	265.1	209.7	213.2	76	35.6	858.4

The data used for this analysis are from a probabilistic survey design. These results therefore represent spatially unbiased estimates. More information is available [here](#).



impRoving infoRmation dissemination





Getting Started with R



R Links

- R Homepage: www.r-project.org
 - The official site of R
- R Foundation: www.r-project.org/foundation
 - Central reference point for R development community
 - Holds copyright of R software and documentation
- Local CRAN:
 - Mirror site
 - We use: cran.cnr.berkeley.edu
 - Find yours at: cran.r-project.org/mirrors.html
 - Current Binaries
 - Current Documentation & FAQs
 - Links to related projects and sites
- JGR Site: jgr.markushelbig.org/JGR.html



R Basics – Learning More

Wikipedia

[http://en.wikipedia.org/wiki/R_\(programming_language\)](http://en.wikipedia.org/wiki/R_(programming_language))

An Introduction to R

<http://cran.cnr.berkeley.edu/doc/manuals/R-intro.html>

Links to all “official” manuals (html & pdf)

<http://cran.cnr.berkeley.edu/manuals.html>

R Graph Gallery

<http://addictedtor.free.fr/graphiques/>

R Wiki

<http://wiki.r-project.org/rwiki/doku.php>



Introductory Statistics with R
Series: *Statistics and Computing*
Dalgaard, Peter
2nd ed., 2008, XVI, 364 p., Softcover
ISBN: 978-0-387-79053-4



Questions? Comments?

- Now would be the time!
- Keep in contact!
 - Mike's email: mike@dataspora.com
 - John's email: oramjj@yahoo.com
 - Jim's email: JPorzak@Responsys.com
- Jim's past presentations: www.porzak.com/JimArchive

- Use R Group of San Francisco Bay Area: <http://ia.meetup.com/67/>





Appendix



JGR Demo Details

The screenshot shows the JGR (Java GUI for R) interface. On the left is the R Console window, displaying the R startup message, license information, and a session history of commands and data frames. On the right is the RHelpMessageCounts.R script editor window, showing R code for reading a tab-delimited data file and creating various plots using the lattice package.

```
R version 2.14.1 (2011-06-23)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> ## RHelpMessageCounts.R
>
> setwd("C:/Data/RGames/R") ## This is Jim's folder setup. Point to yours!
> library(lattice)
>
> ## Read tab delimited data file. Keep columns 1 & 2 "as is"
> RHelp <- read.delim("Data/RHelpCountsByPeriod.txt", as.is = 1:2)
> ## See what we got
> head(RHelp)
  Period_Start Period_End Number_Msgs
1 27 Feb 2005 31 Mar 2005      1660
2 01 Apr 2005 30 Apr 2005      2062
3 01 May 2005 01 Jun 2005      1843
4 01 Jun 2005 30 Jun 2005      2077
5 01 Jul 2005 26 Aug 2005      1895
6 01 Aug 2005 03 Sep 2005      1756
> str(RHelp)
'data.frame':   41 obs. of  3 variables:
$ Period_Start: chr  "27 Feb 2005" "01 Apr 2005" "01 May 2005" "01 Jun 2005" ...
$ Period_End : chr  "31 Mar 2005" "30 Apr 2005" "01 Jun 2005" "30 Jun 2005" ...
$ Number_Msgs : int  1660 2062 1843 2077 1895 1756 1679 1794 1897 1561 ...
>

xplot(MsgPerWk ~ Period_Start, data = RHelp,      ## just the points
      main = "R Help Maillist Message Increase")
```

```
1 ## RHelpMessageCounts.R
2
3 setwd("C:/Data/RGames/R") ## This is Jim's folder setup. Point to yours!
4 library(lattice)
5
6 ## Read tab delimited data file. Keep columns 1 & 2 "as is"
7 RHelp <- read.delim("Data/RHelpCountsByPeriod.txt", as.is = 1:2)
8 ## See what we got
9 head(RHelp)
10 str(RHelp)
11
12 ## Convert date strings to true date types. Format is dd Month yyyy.
13 RHelp$Period_Start <- as.Date(RHelp$Period_Start, "%d %b %Y")
14 RHelp$Period_End   <- as.Date(RHelp$Period_End,   "%d %b %Y")
15
16 ## Since periods different length, we need # messages / week
17 RHelp$MsgPerWk <- 7 * RHelp$Number_Msgs /(as.integer(RHelp$Period_End - RHelp$Period_Start))
18 ## One last look
19 head(RHelp)
20 str(RHelp)
21
22 ## Now do some plots using lattice package
23 xplot(MsgPerWk ~ Period_Start, data = RHelp,      ## just the points
24        main = "R Help Maillist Message Increase")
25
26 xplot(MsgPerWk ~ Period_Start, data = RHelp,      ## with regression line
27        type = c("p", "r"),
28        main = "R Help Maillist Message Increase")
29
30 xplot(MsgPerWk ~ Period_Start, data = RHelp,      ## with smooth lowess fit
31        type = c("p", "smooth"),
32        main = "R Help Maillist Message Increase")
33
34 ## box & whisker plot
35 bwplot(MsgPerWk ~ as.factor(format(RHelp$Period_Start, "%Y")), data = RHelp,
36        main = "R Help Maillist Message Increase")
37
38
```

Download data & R code from
www.porzak.com/JimArchive/RHelp.zip



Expanded odfWeave source file (1)

Responsys

MO QA 2.0 Counts for Load As Of \\$expr{mo_AsOfDate}

Summary

Missing Data:

```
<<MissingTable, echo = FALSE, results = xml>>=
if(!is.null(nrow(Missing))) {
  odfTable(Missing, useRowNames = FALSE, colnames = colnames(Missing))
} else {
  odfCat("No completely missing data elements for any weeks.")
}
@
```

Unusual Trends:

```
<<UnusualTable, echo = FALSE, results = xml>>=
if(!is.null(nrow(Suspects))) {
  #op <- options()
  #options(digits = 3)
  odfTable(Suspects, useRowNames = TRUE,
            colnames = c("Data Set", colnames(Suspects)))
  #options(op)
} else {
  odfCat("No unusual trends observed.")
}

if(!is.null(nrow(Suspects))) {
  odfCat("Where AboveBelow shows weeks with counts above, +, or below, -,
         mean of all weeks.")
}
@
```



Expanded odfWeave source file (2)

Membership

Notes:

```
<<MembCounts, echo = FALSE, fig = TRUE>>=
siD(6.5, 3, 200)
print(
xyplot(NumberNew + NumberLost ~ WeekOfSatThe_, CountsByWeek,
      type = "b", ylim = c(0, 50000),
      auto.key = list(title = "New and Lost Members by Week", space =
"top",
                  columns = 2, lines = TRUE),
      xlab = "Start of data week", ylab = "Count",
      scales = list(x = list(at = CountsByWeek$WeekOfSatThe_, rot = 60)))
)
@
```

611 Mission Street, 6th Floor | San Francisco, CA 94105 | 415.278.5400 | www.responsys.com