

Data-driven journalism

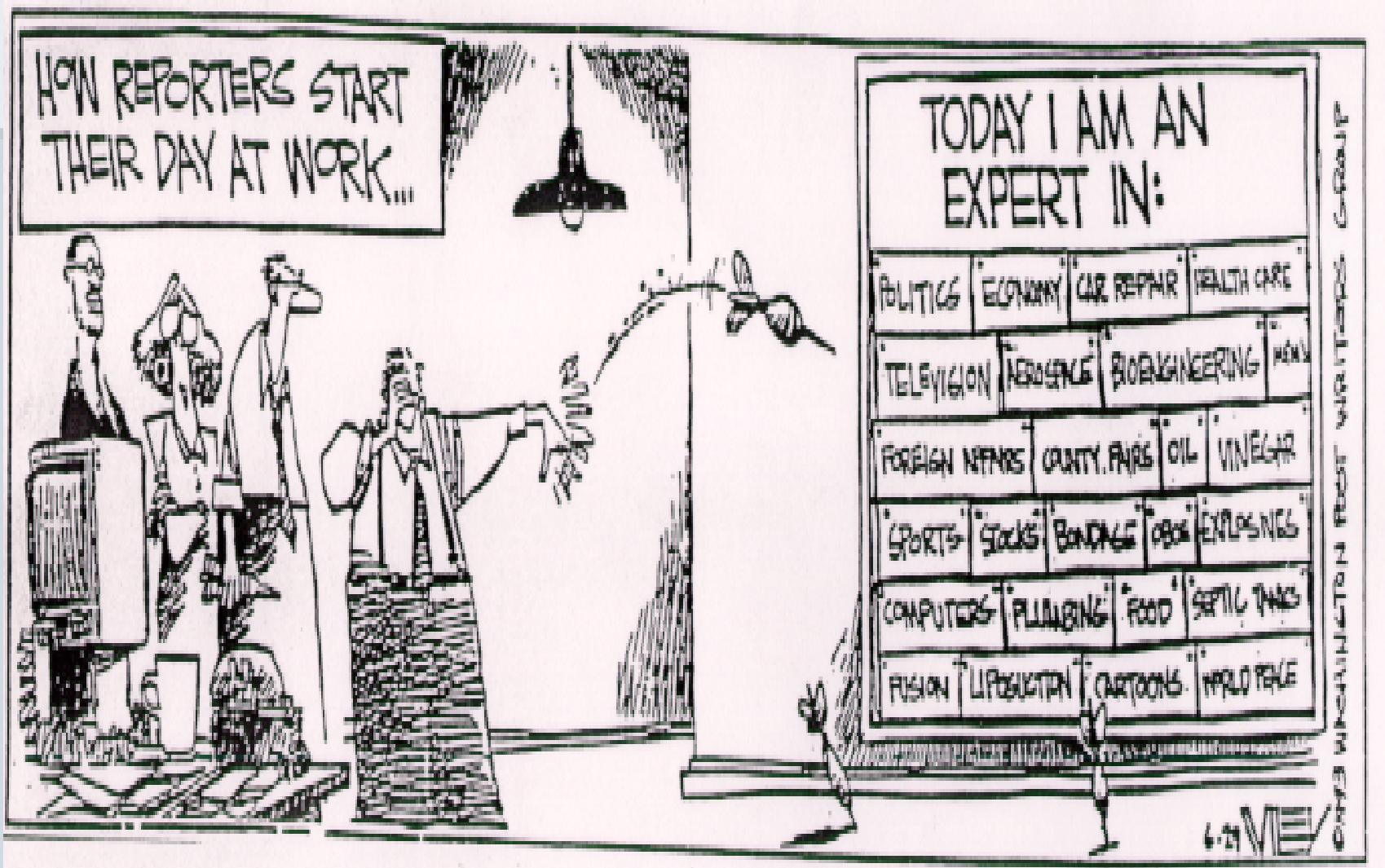
Bay Area R Users Group, Dec 14 2010

Peter Aldhous,
San Francisco Bureau Chief



NewScientist

A disclaimer



From the ashes of the news industry, a phoenix?

Journalism in the Age of Data

A video report on data visualization as a storytelling medium
Produced during a 2009-2010 Knight Journalism Fellowship
Total Running Time: 54 Minutes; with related information and links

How Different Groups Spend Their Day

The American Time Use Survey asks thousands of American residents to recall every minute of a day. Here is how people over age 15 spent their time in 2008. Related article

Everyone

Sleeping, eating, working and watching television take up about two-thirds of the average day.

Everyone	Employed	White	Age 15-24	H.S. grads	No children
Men	Unemployed	Black	Age 25-64	Bachelor's	One child
Women	Not in lab.	Hispanic	Age 65+	Advanced	Two+ children



II. Data Vis in Journalism

How data reporting and presentation are starting to change the face of newsrooms.

CHAPTERS

I. Introduction

II. Data Vis in Journalism

III. Telling "Data Stories"

IV. A New Era in Infographics

V. Life as a Data Stream

VI. Exploring Data

VII. Technologies and Tools

VIII. First Steps

[Watch](#) the video.

Words from the wise ...

guardian.co.uk

Search

Media

Search

[News](#) | [Sport](#) | [Comment](#) | [Culture](#) | [Business](#) | [Money](#) | [Life & style](#) | [Travel](#) | [Environment](#) | [TV](#) | [Video](#) | [Community](#) | [Offers](#) | [Jobs](#)

[News](#) > [Media](#) > [Digital media](#)

Analysing data is the future for journalists, says Tim Berners-Lee

Inventor of the world wide web says reporters should be hunting for stories in datasets



Charles Arthur

The Guardian, Monday 22 November 2010

[Article history](#)



Tim Berners-Lee. Photograph: Guardian

[Tweet](#) 1,100

[Share](#) 433



[Comments](#) (9)



[A](#) larger | smaller

Media

[Digital media - Journalism education](#)

Technology

[Tim Berners-Lee](#)

More features

More on this story



[Berners-Lee: Facebook could fragment web](#)
Founder of world wide

guardianjobs

Search all jobs [Go](#)

[Browse all jobs](#)
jobs by [indeed](#)



BANKING

Banks Forced To Forgive Credit Card Debt

New Credit Laws Allow San Francisco Consumers to Reduce Debt up to 60%...



Mom's \$5 Wrinkle Secret!

San Francisco: Dermatologists DON'T Want You Knowing This Skin Care Trick!



What's Your Credit Score?

The Average Credit Score is 678. Find Out Your Score For Free.

ADS BY YABUKA.COM >

On Media

[Most viewed](#)

[Zeitgeist](#)

[Latest](#)

Last 24 hours



1. [Julian Assange: Whoever leaked US](#)

DER SPIEGEL

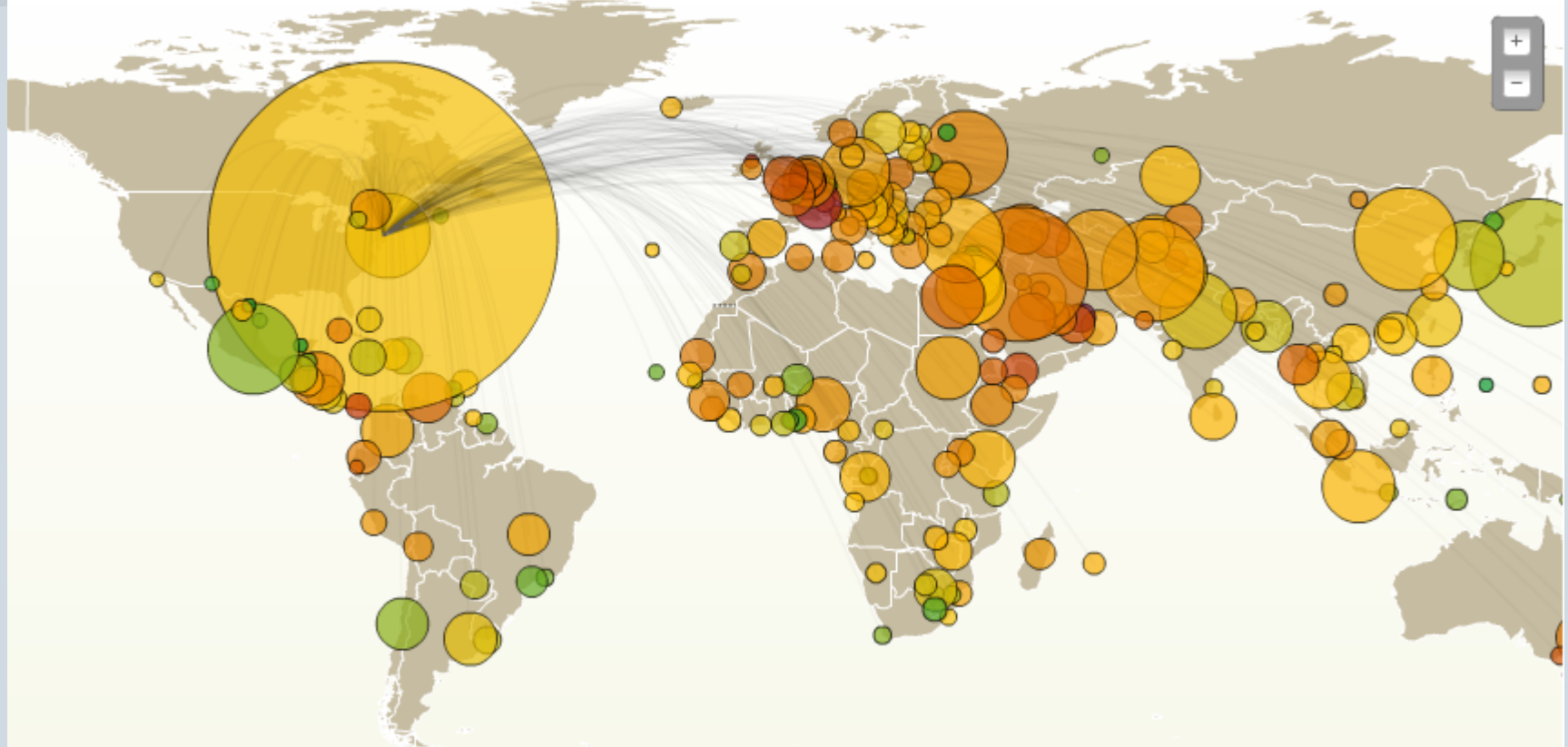
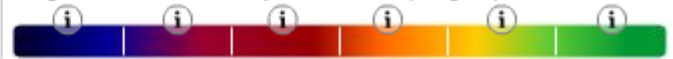
The US Embassy Dispatches

Interactive Atlas

Dossiers A-Z

The map shows **where** the dispatches originated in the chosen time period. The more dispatches, the larger the **circle**. The relative quantity of classified documents from a given location is indicated according to the **color scale** on the right. Click on a US Embassy location to see **statistics** for each representation,

Degree of classification (Color coded by degree)



2000 to 2010, by month

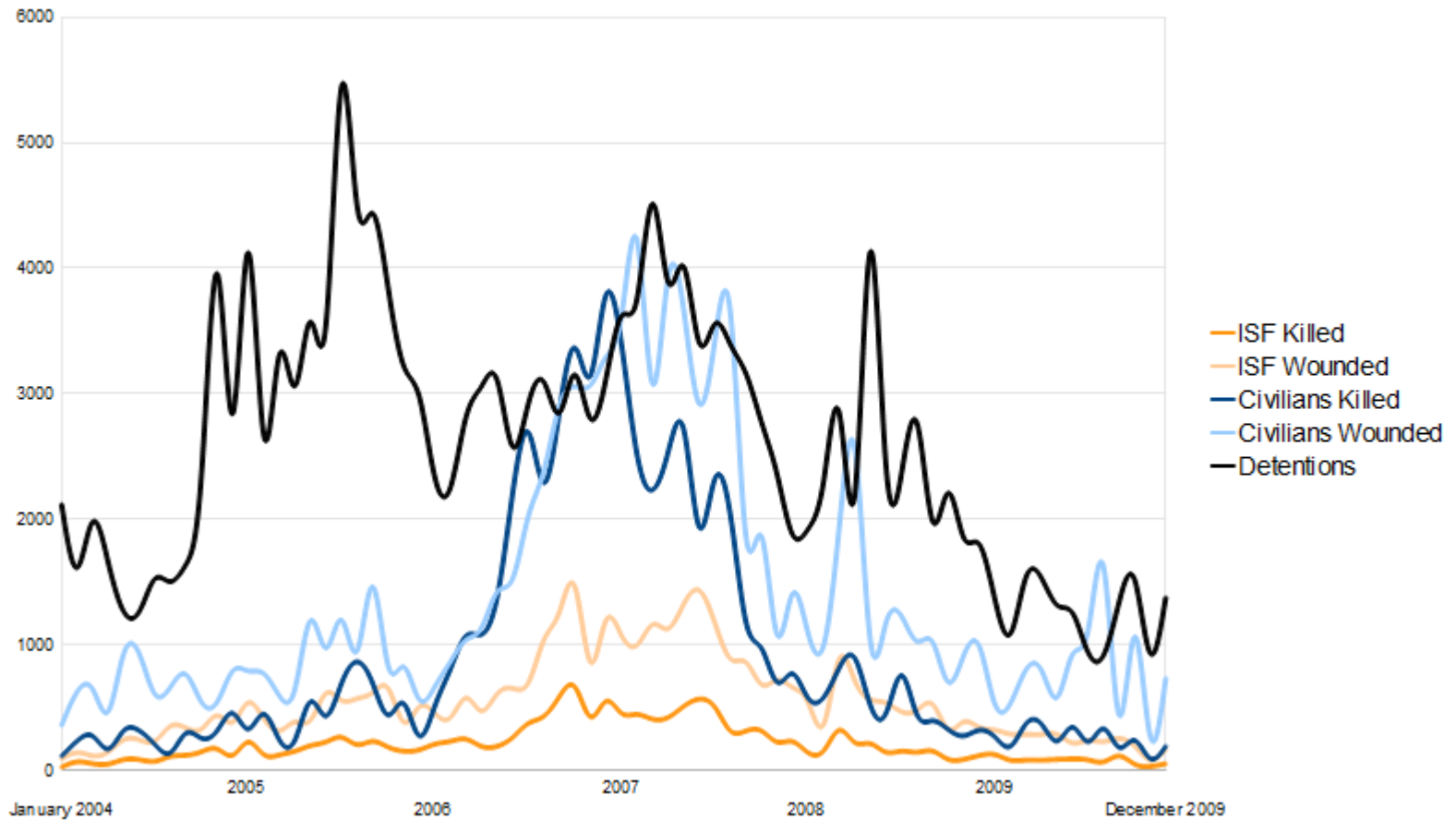
1966 to 2010, by year

Total time period

Play

1966 1971 1976 1981 1986 1991 1996 2001 2006 2010

2009



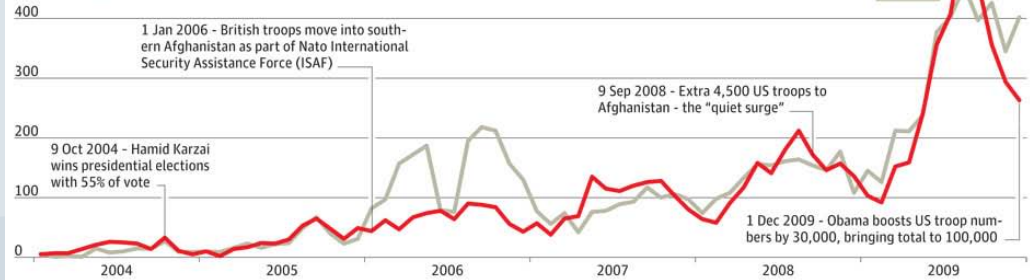
AFGHANISTAN THE WAR LOGS



IEDs exploded and cleared

Explosion/ambush
Total **7,553**

Found/cleared
Total **8,582**

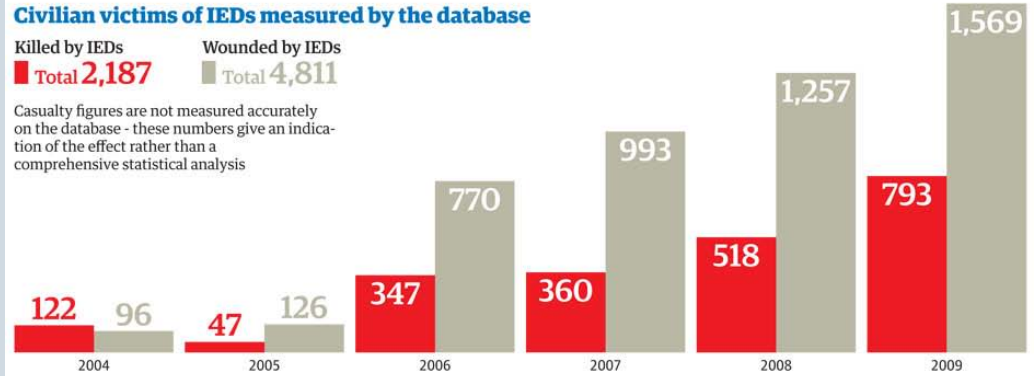


Civilian victims of IEDs measured by the database

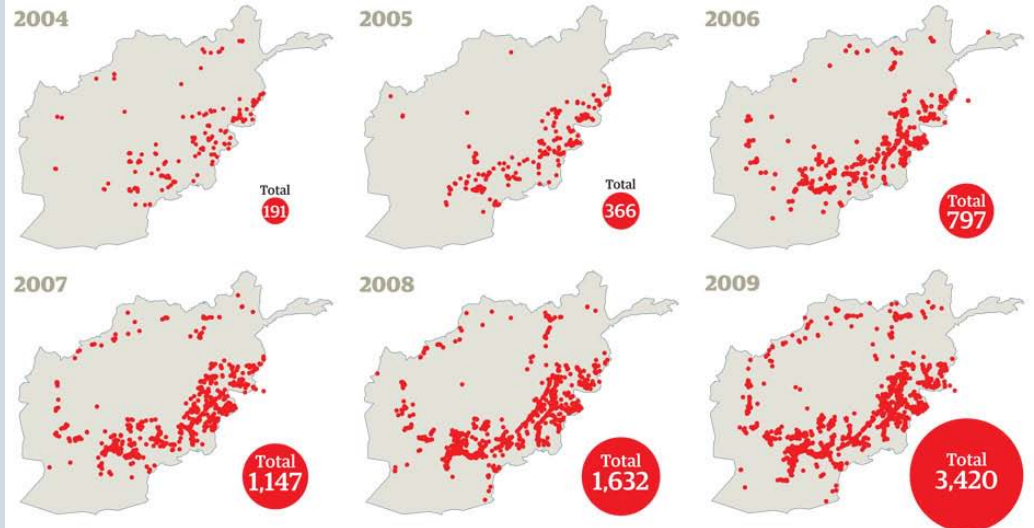
Killed by IEDs
Total **2,187**

Wounded by IEDs
Total **4,811**

Casualty figures are not measured accurately on the database - these numbers give an indication of the effect rather than a comprehensive statistical analysis



IED attacks year by year



The New York Times

Published: January 8, 2010

A Peek Into Netflix Queues

Examine Netflix rental patterns, neighborhood by neighborhood, in a dozen cities. Some titles with distinct patterns are *Mad Men*, *Obsessed* and *Last Chance Harvey*. [Comments \(135\)](#)

100 titles that were frequently rented from Netflix in 2009



Change how movies are sorted

Most rented Alphabetical By metacore

Milk



84

Metacritic score

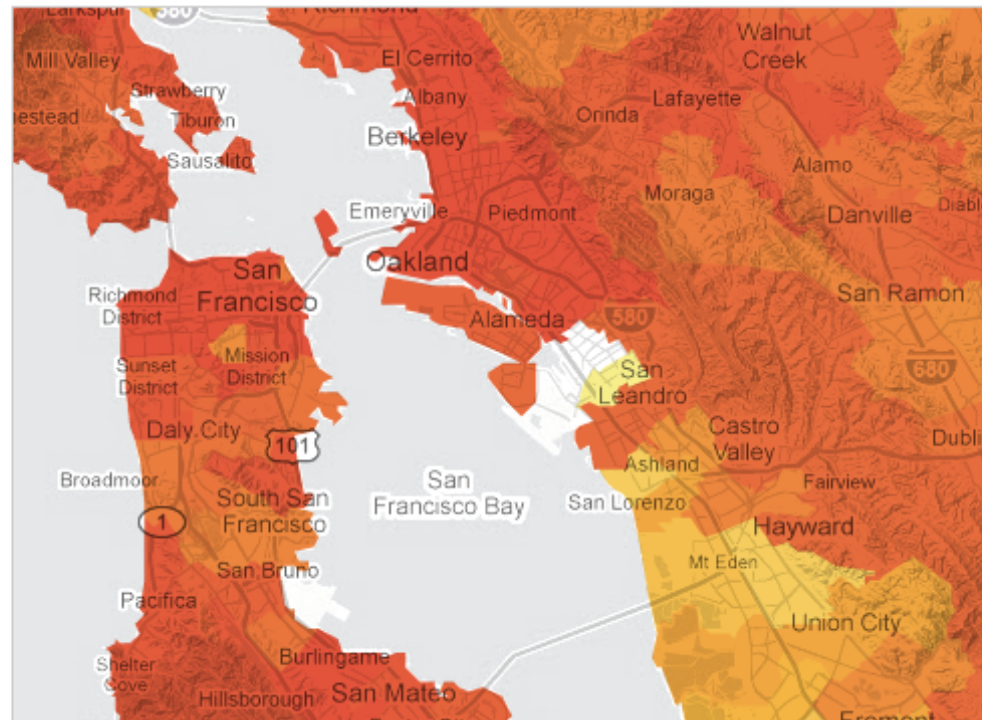
100=loved by critics, 0=hated

One of the first scenes in "Milk" is of a pick-up in a New York subway station. It's 1970, and an insurance executive in a suit and tie catches sight of a beautiful, scruffy younger man — the phrase "angel-headed hipster" comes to mind — and banter with him on the stairs. The mood of the moment, which ends up with the two men eating birthday cake in bed, is casual and sexy, and its flirtatious playfulness is somewhat disarming, given our expectation of a serious and important movie grounded in historical events. "Milk," directed by Gus Van Sant from a script by Dustin Lance Black, is certainly such a film, but it manages to evade many of the

The ZIP codes are shaded according to each movie's rank.

Ranked No. 1

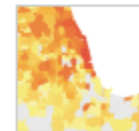
No. 50



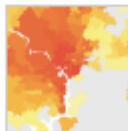
NEW YORK



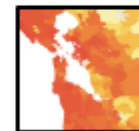
BOSTON



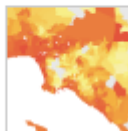
CHICAGO



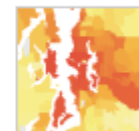
WASHINGTON



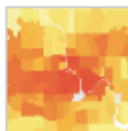
BAY AREA



L.A.



SEATTLE



MINNEAPOLIS

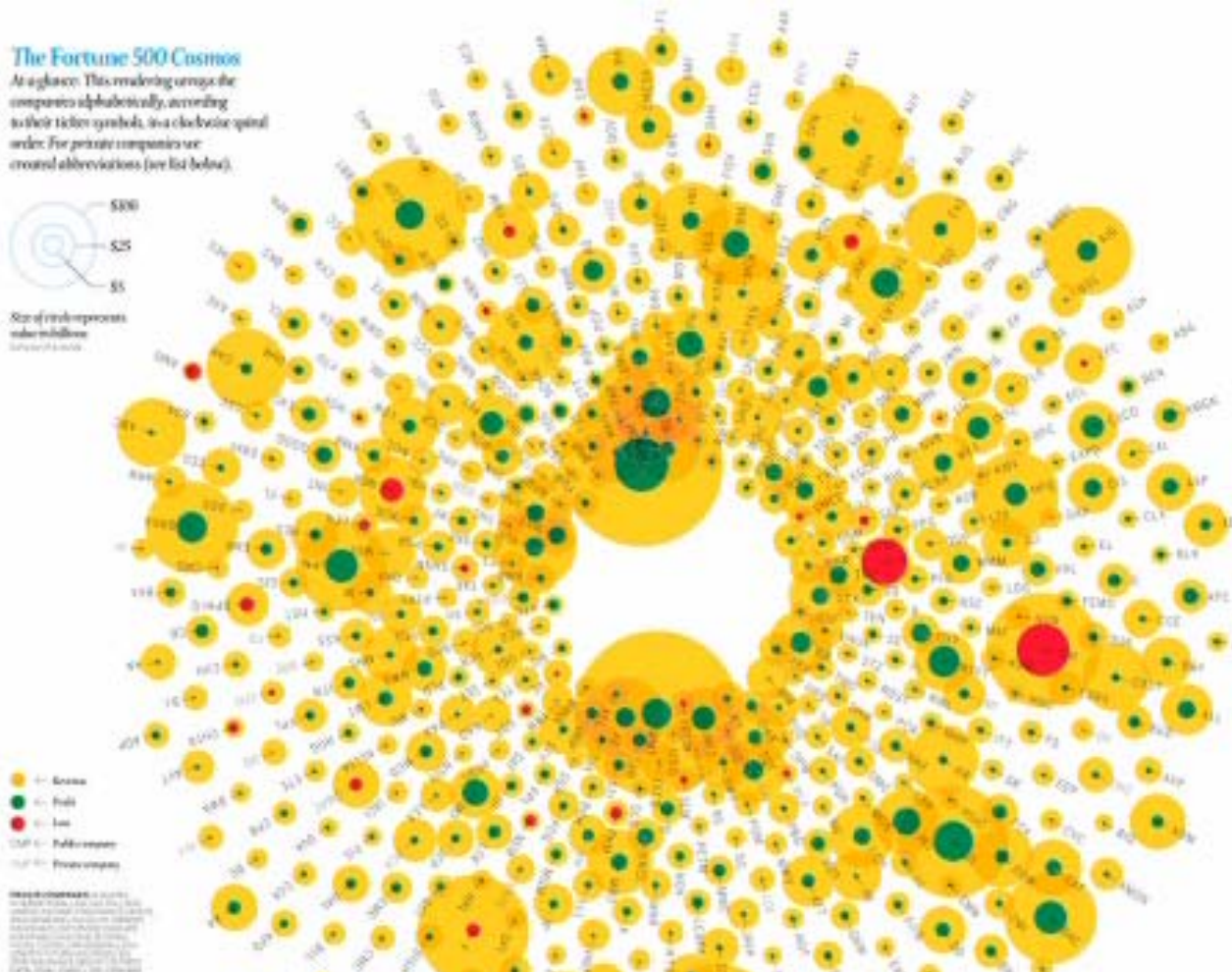
[Explore](#) the interactive.

Beauty is not enough

FORTUNE

The Fortune 500 Cosmos

At a glance, this revealing collage of companies alphabetically, according to their ticker symbols, in a clockwise spiral order. For private companies we created abbreviations (see list below).



FORTUNE
500



In an age of turmoil, corporate stars rise and fall. How many will survive? A bestselling management expert makes the case for well-founded hope.

BY JIM COLLINS

THE SECRET OF ENDURING GREATNESS

The Fortune 500 series

I don't know how many times I've had the Fortune 500 presented to me as Exhibit A by those who argue that it's time to give up on a 20th-century American idea. Forget about building great companies that endure, they tell me. One technology pundit convinced me at a conference and demoted the whole premise shared in today's world. "We live in an era when nothing can be built to last. Everything is in flux; nothing can sustain," he inveighed. Joseph Schumpeter, the great economist who wrote about the "perennial gale of creative destruction" wherein technological change and visionary entrepreneurs give birth to new things that obliterate old things, only to see those new things become obliterated by the next generation.

His argument took particularly sharp form: Ben Starns disappeared over



You need to tell a story



[Watch](#) the video.

Who are the data journalists?



INVESTIGATIVE REPORTERS AND EDITORS

improving the quality of investigative reporting since 1975

[Home](#) [About](#) [Members](#) [IRE Store](#)

[Conferences & Training](#)

[NICAR & Data](#)

[Resource Center](#)

[Job Center](#)

[News & Publications](#)

[Give to IRE](#)

[IRE en Espanol](#)

[Conference home](#)

[Hotel/Travel](#)

[Registration](#)

[Expected speakers](#)

[Hands-on Classes](#)

[Schedule](#)

[Special Events](#)

[Sponsorship Info.](#)

[Exhibitors/Recruiters](#)



To secure the early-bird registration rate of \$195, you must register by noon (Central) on Feb 14. After Feb. 14, you must register on site for \$225. You must pre-register if you plan to attend the CAR Mini-Boot Camp or Searchable Data Boot Camp.

Looking to learn skills that will help you, and your newsroom, thrive?

Join IRE and NICAR in Raleigh for our annual conference devoted to computer-assisted reporting, **Feb. 24-27**. Come and learn about tools you need to dig deeper into stories and give readers, viewers and your online audience the information they're demanding.

The CAR conference offers something for everyone, from beginners to those on the cutting edge of digital reporting. We'll offer everything from the latest technological advances to the basics on using spreadsheets, databases and online mapping. Learn from the best in the business in panel discussions and during hands-on training sessions. Bypass the budget issues in your newsroom by taking classes in free software. Or get a look at what the biggest names in data-driven reporting are using to make a major impact online.



Host

[Duke University](#)

Sponsors

[Gannett Foundation](#)
[The University of North Carolina at Chapel Hill](#)
[The News & Observer \(Raleigh, N.C.\)](#)

Calendar

- | | |
|----------------|---|
| Feb 14 | Last day of early-bird registration. Registration closes at noon. |
| Feb. 24 | Advanced Day starts at 9 a.m. |
| Feb. 25 | Conference starts at 9 a.m. |
| Feb. 27 | Conference concludes at noon |

Come an afternoon early to CAR or IRE Investigating Private Companies and Nonprofits

Feb. 23 In Raleigh, N.C., and June 8 In Orlando, Fla.

For more information contact Stephanie Sinn at stephanie@ire.org or call 573-882-8969.

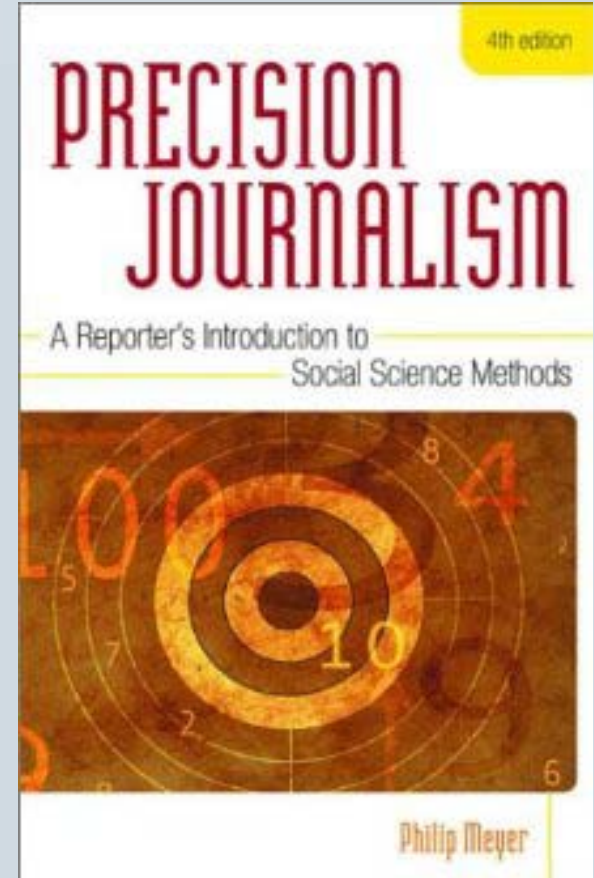
The pioneer: Philip Meyer



Now emeritus professor of journalism, University of North Carolina at Chapel Hill.

Pioneered use of quantitative methods in journalism with Knight Newspapers in 1960s.

Author of *Precision Journalism*, first published 1973.



A Pulitzer for data journalism: 1967 Detroit riot

Detroit Free Press



- 43 dead
- 467 injured
- 7231 arrests



Data: Survey conducted in the immediate aftermath of the riot.

Findings: One theory held that the rioters were stuck at the foot of the economic ladder with no other means of expression. Another argued that southern blacks who had moved to Detroit were venting years of pent-up rage. But Philip Meyer showed that college graduates were as likely to have rioted as high-school dropouts, and that those born in the South were less likely to have participated.

Attention turned instead to pervasive racial discrimination in policing and housing in Detroit.

Tools and stories: relational databases

The Newsday logo is displayed in a dark blue rectangular box with a white border. The word "Newsday" is written in a bold, white, serif font.

DANGER ON THE LISTS



Insurers Say
They Screen Out
Doctors With
Troubled Histories,
But Dozens
Have Made It Into
Their Directories

Data: HMO doctor directories and state records of disciplinary actions taken against doctors.

Findings: Despite promises of high quality and rigorous screening, New York's biggest managed health care networks offered their customers dozens of doctors disciplined for serious – even fatal – wrongdoing.

Even though the health insurers were fully aware that the state punished these doctors for such offenses as botched surgery, sexual misconduct, drug abuse or cheating government insurance plans, they never told their customers.

Tools and stories: GIS

The Seattle Times

Logging and landslides: What went wrong?

When Weyerhaeuser began clear-cutting the Douglas firs on the slopes surrounding Little Mill Creek, local water officials were on edge. Some of these lands had slid decades ago, after an earlier round of logging. They worried new slides could dump sediments into the mountain stream and overwhelm a treatment plant. Those fears came true last December.

By [Hal Bernton](#) and [Justin Mayo](#)
Seattle Times staff reporters

BOISTFORT VALLEY, Lewis County — When Weyerhaeuser began clear-cutting the Douglas firs on the slopes surrounding Little Mill Creek, local water officials were on edge.

Some of these lands had slid decades ago, after an earlier round of logging. They worried new slides could dump sediments into the mountain stream and overwhelm a treatment plant.

Those fears came true last December when a monster storm barreled in from the Pacific, drenching the mountains around the Chehalis River basin and touching off hundreds of landslides. Little Mill Creek, filled with mud and debris, turned dark like chocolate syrup.

More than three months passed before nearly 3,000 valley residents could drink from their taps again.

"I have never seen anything like this before, and I hope I never do again," said Fred Hamilton, who works for the Boistfort Valley Water Corp.

State forestry rules empower the Department of Natural Resources (DNR) to restrict logging on

◀ PREV 1 of 7 NEXT ▶



enlarge

STEVE RINGMAN / THE SEATTLE TIMES

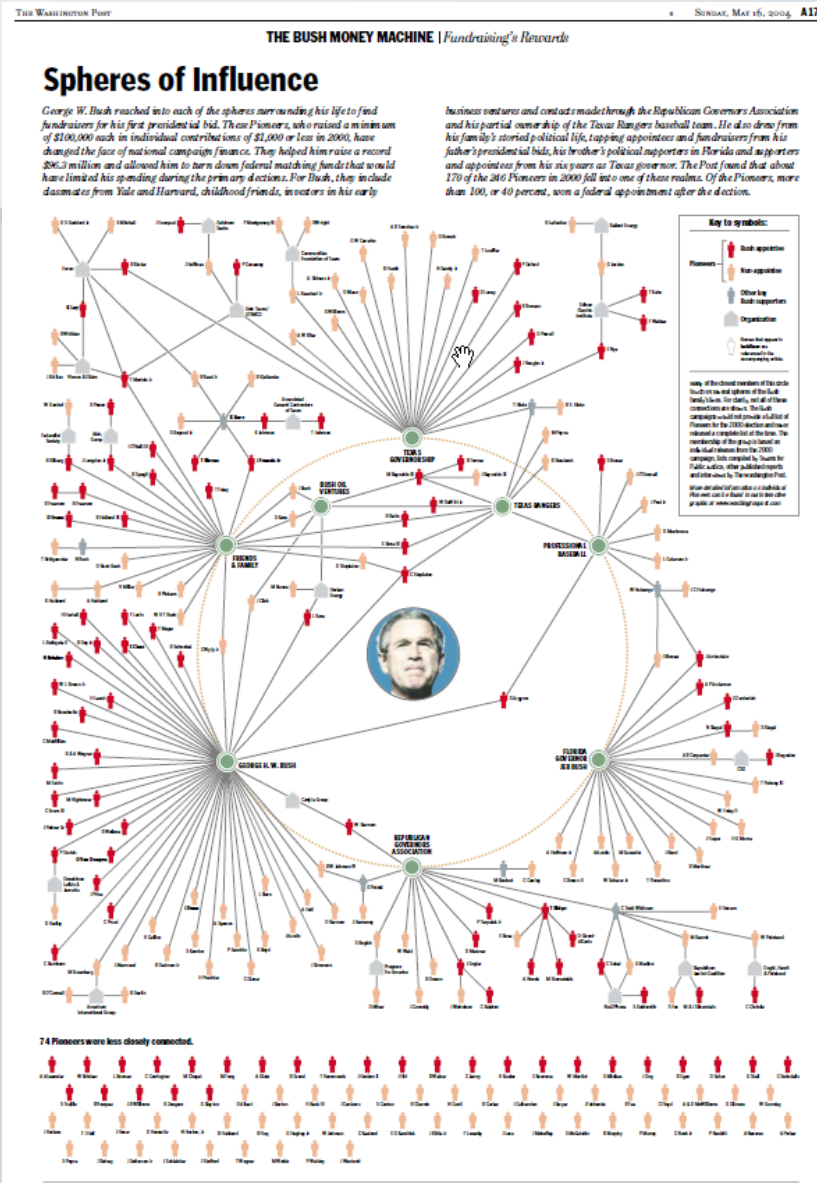
Data: GIS data on clear-cuts and landslides from the Washington State Department of Natural Resources. Logging company Weyerhaeuser's logging permits.

Findings: With little scrutiny from state geologists, Weyerhaeuser was allowed to clear-cut unstable slopes.

Using mapping software, the reporters showed that clear-cut sites that had at least half of their acreage in a moderate- to high-hazard zone accounted for a disproportionate number of landslides in December 2007 storms.

[Explore](#) interactive graphic.

Tools and stories: social network analysis



The Washington Post

Data: Built database of George W. Bush “Pioneers” – those who raised more than \$100,000 to his 2000 presidential campaign.

Findings: Social network analysis to show who were the key Pioneers in the Bush campaign, who they were connected to, and what each Pioneer gained, if anything, from his or her association with Bush, such as ambassadorships and other federal appointments. Bush raised \$96.3 million, a record at that time. More than 100, about 40%, of the Pioneers received some federal appointment after the election.

Tools and stories: statistical analysis



dallasnews.com
The Dallas Morning News

Exclusive: Poor schools' TAKS surges raise cheating questions

09:42 PM CST on Sunday, December 19, 2004

By JOSHUA BENTON and HOLLY K. HACKER / The Dallas Morning News

A *Dallas Morning News* data analysis has uncovered strong evidence of organized, educator-led cheating on the TAKS test in dozens of Texas schools – and suspicious scores in hundreds more.

The analysis found a poor urban school where third- and fifth-graders are among the state's weakest readers – but the fourth-graders beat out the state's most elite schools. That's despite the fact that many of its students have trouble speaking English.

It found a desperately impoverished school where the fourth-graders have trouble adding and subtracting – but nearly all the fifth-graders got perfect scores on the math portion of the Texas Assessment of Knowledge and Skills.

And it found schools where in one year's time – if the scores are to be believed – children devolved from top students to barely being able to read.

The News' findings have led to cheating inquiries in three Texas school districts, including the state's two largest, Dallas and Houston. One of the schools under investigation is a National Blue Ribbon School that a year ago was touted by federal officials as an example of top academic achievement.

About this series

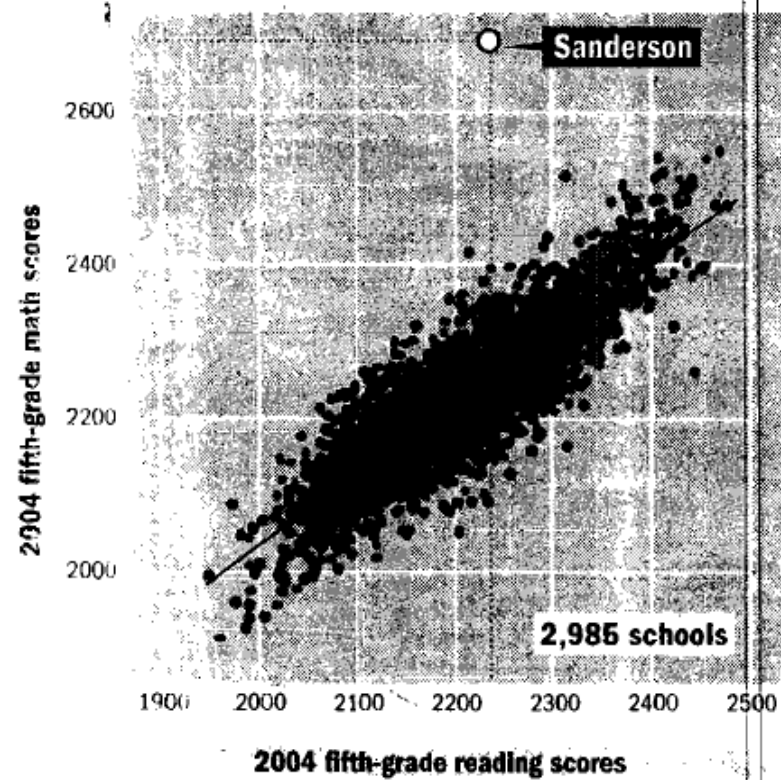
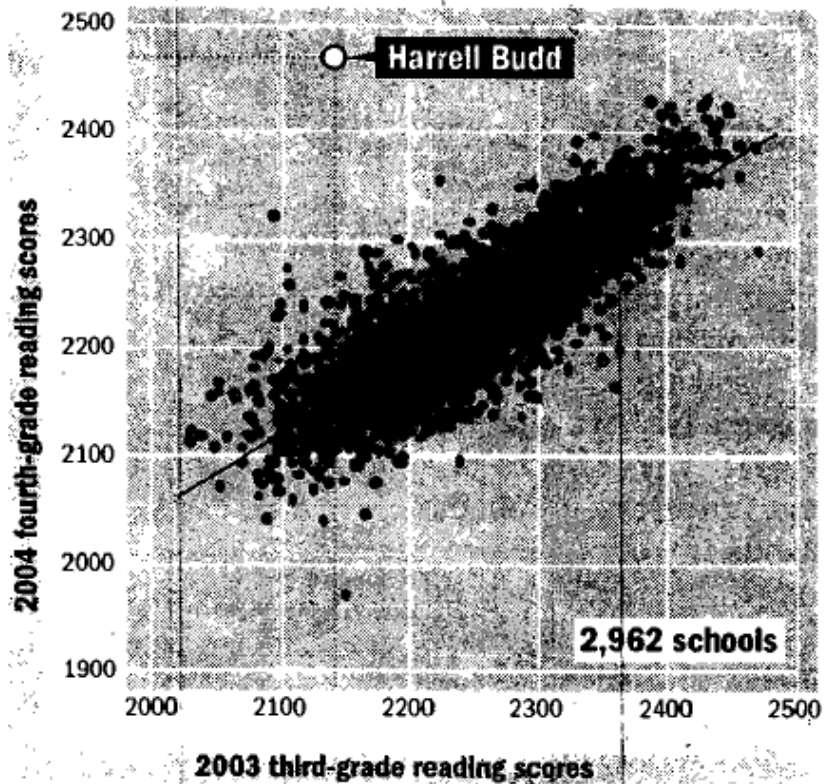
For this story, *The Dallas Morning News* analyzed school test scores on the Texas Assessment of Knowledge and Skills. Now in its second year, the exam is required for public-school students in grades three through 11.

The state focuses on school passing rates on the TAKS – that is, the percentage of students who met state standards. *The News* analysis used average scale scores, a more specific

Data: Results from Texas standardized assessment tests.

Findings: Reporters turned a story about one school's alleged cheating on standardized tests into a piece about cheating across the state. They used regression analysis to show some suspicious improvements among historically low-performing schools, including a “desperately impoverished school where the fourth-graders have trouble adding and subtracting – but nearly all the fifth-graders got perfect scores on the math portion of the Texas Assessment of Knowledge and Skills.”

The *Morning News* also found that the Texas Education Agency doesn't use perform similar analyses.



So, are any journalists using R?



The screenshot shows the header of the 'Revolutions' blog, which focuses on R, statistics, and open source. The main article is dated June 26, 2009, and is titled 'NYT charts Michael Jackson's pop hits with R (updated)'. The article text mentions that the *New York Times* has a graphic charting Michael Jackson's career using R, comparing his Billboard rankings to other artists like The Beatles, U2, Mariah Carey, Usher, and Boyz II Men. A sidebar on the right contains links for 'Information', 'About this blog', 'Comments Policy', 'About Categories', 'About the Authors', 'R Community Calendar', and 'Local R User Group Directory'. At the bottom of the sidebar, it says 'Got comments or suggestions for the blog editor? Email [David Smith](#)'.



I saw your blog post on the Michael Jackson chart in the New York Times today. I thought it might amuse you to know that the charts were made in R. (Then cleaned up in Illustrator and moved into Flash, but they started life in R.)

Amanda Cox, graphics department, *The New York Times*

The New York Times

June 25, 2009

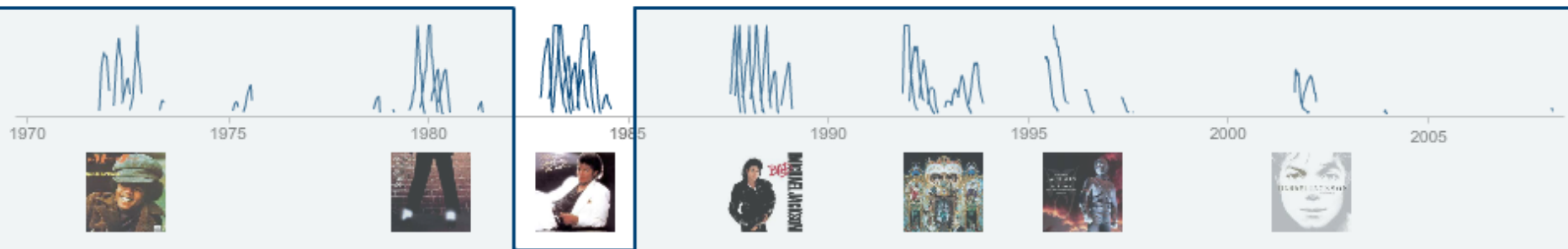
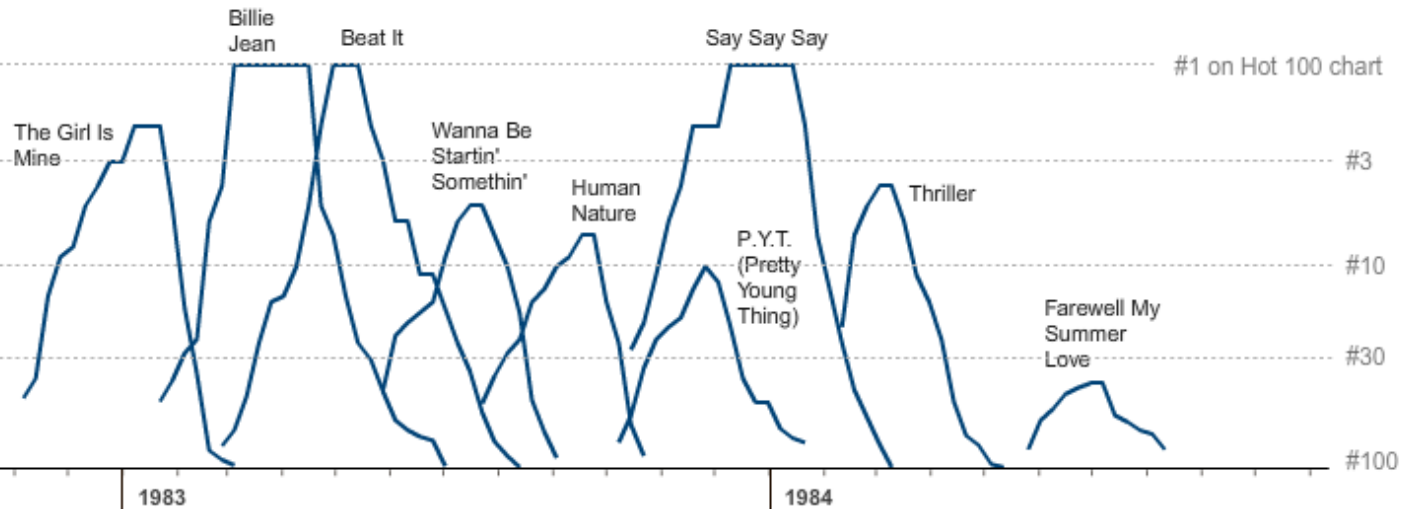
[E-MAIL](#) | [FEEDBACK](#)

Jackson's Billboard Rankings Over Time

A timeline of how Michael Jackson's songs performed on the [Billboard](#) Hot 100 chart.

[← Previous](#) | [Next →](#)

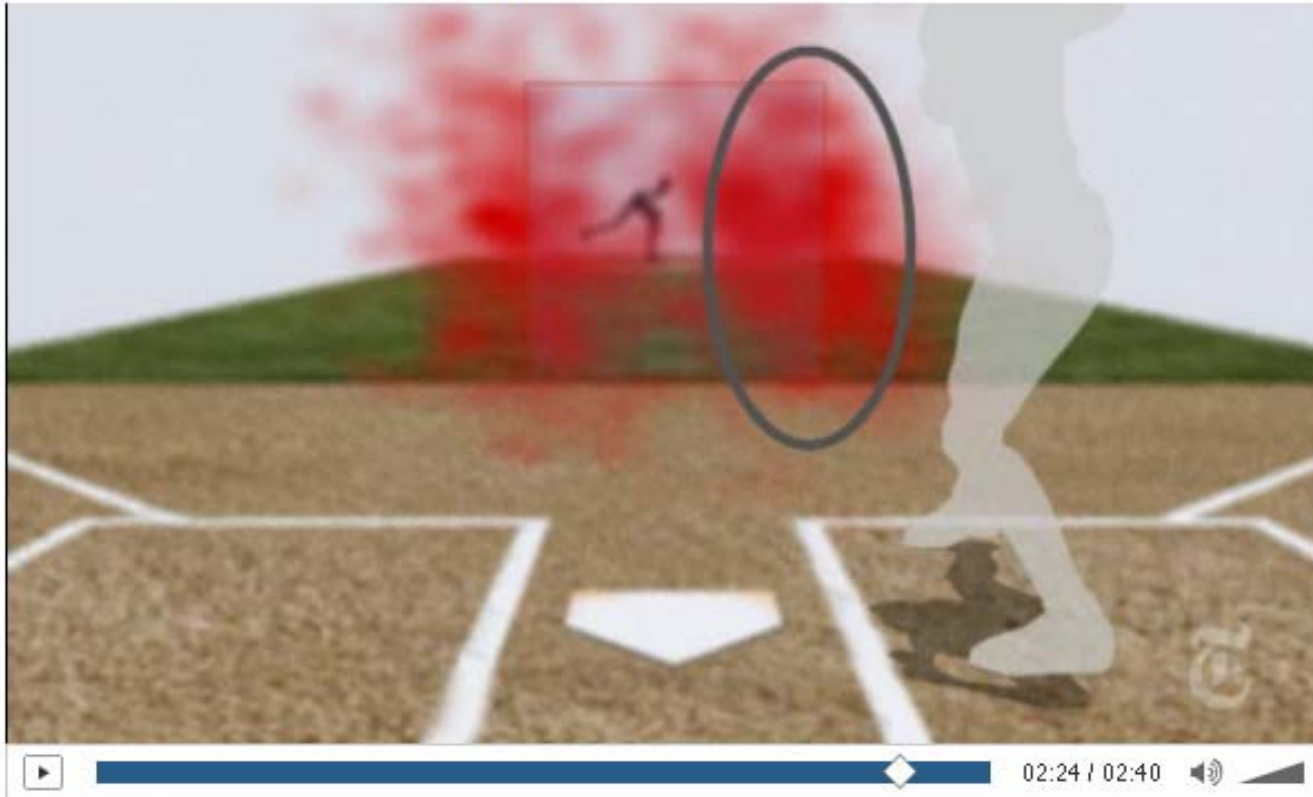
Thriller 1982



[Explore](#) the interactive.

The New York Times

Mariano Rivera, King of the Closers



By JAMES TRAUB
Published: June 29, 2010

[Watch](#) the video.

The New York Times

Fight Leaves Democrats Questioning Prospects



Emmanuel Dunand/Agence France-Presse — Getty Images

Senator Barack Obama, a Democratic presidential candidate, meeting with veterans on Tuesday in Washington, Pa., one week before the primary there.

By JEFF ZELENY

Published: April 16, 2008

The battle between Senators [Hillary Rodham Clinton](#) and [Barack Obama](#) over whether Mr. Obama belittled voters in small towns appears to have hardened the views of both candidates' supporters and stirred anxiety among many Democrats about the party's prospects in the fall.

TWITTER

E-MAIL

SEND TO PHONE

PRINT

REPRINTS

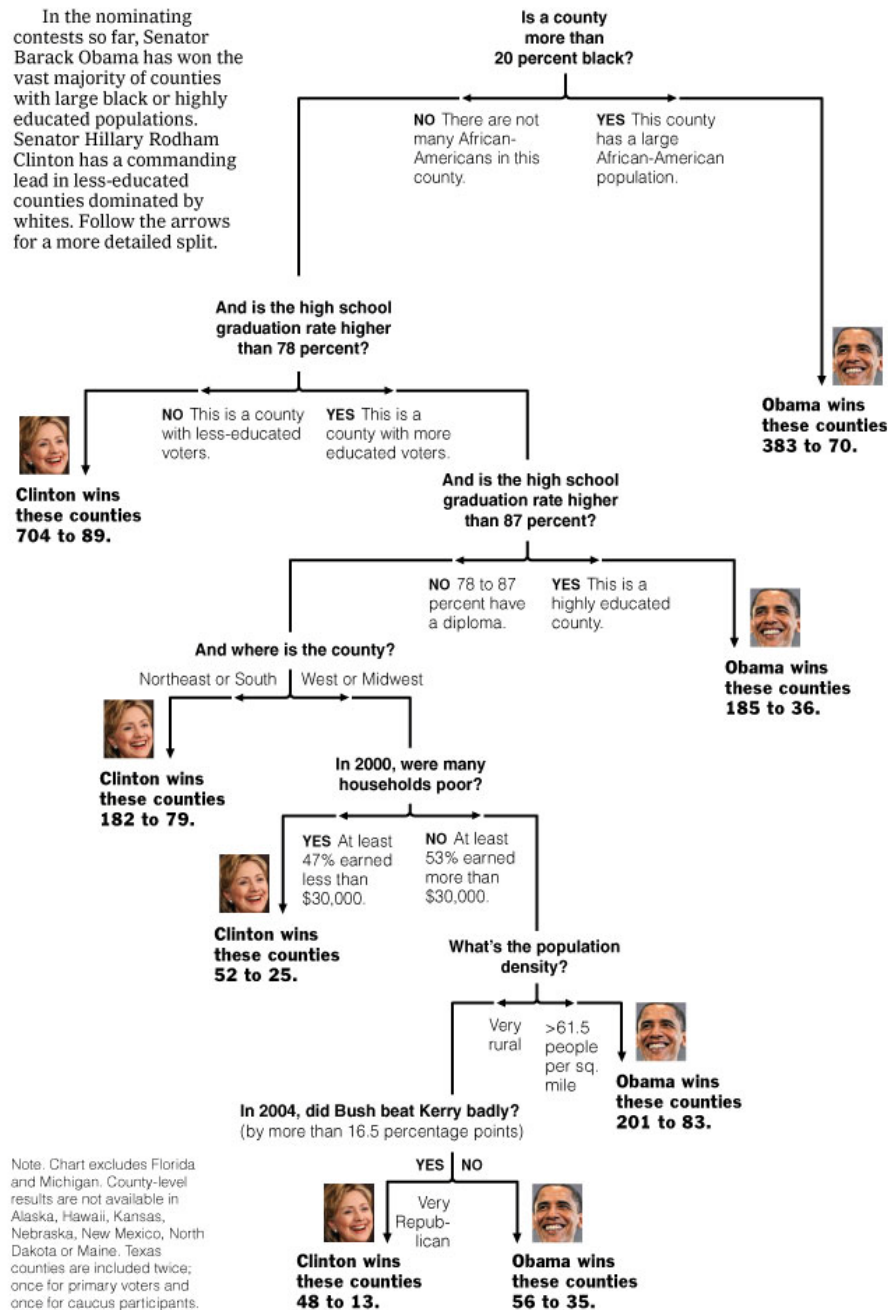
Data: Results from 2008 primaries and 2004 presidential election; US Census.

Method: Classification and Regression Trees algorithm (Brieman, Friedman, Olshen & Stone, 1984).

R package: [rpart](#) (recursive partitioning).

Decision Tree: The Obama-Clinton Divide

In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.



Note. Chart excludes Florida and Michigan. County-level results are not available in Alaska, Hawaii, Kansas, Nebraska, New Mexico, North Dakota or Maine. Texas counties are included twice; once for primary voters and once for caucus participants.

Sources: Election results via The Associated Press; Census Bureau; Dave Leip's Atlas of U.S. Presidential Elections

AMANDA COX/
THE NEW YORK TIMES

MONDAY, AUGUST 20, 2007 **BARRON'S COVER**

Shorting Cramer

By **BILL ALPERT** | [MORE ARTICLES BY AUTHOR](#)

Jim Cramer puts on a great show, but his Mad Money stock picks have underperformed the market over the past two years. A careful look at the maven's track record -- and one way to win. Reader's comments: Check out what our readers think.

Article

Comments

THANKS TO HIS NIGHTLY CNBC SHOW *Mad Money*, Jim Cramer has become the chief cheerleader for the bull market, or what was the bull market until a few weeks ago. Last spring, he was giddily exhorting the Dow Jones Industrial Average toward 15,000, with no troubles in sight. Earlier this month, as the Dow tumbled in the direction of 13,000, he had an on-air meltdown, complete with screaming, sobs and predictions of financial doom. The clip quickly made the rounds on YouTube. Friday, after the Fed cut the discount rate, he said that the Dow's run to 14,500 had begun. With dramatic pronouncements like that, it's no wonder that more than 100,000 viewers tune in each weeknight for his antic mashup of sound effects, Streetwise advice and stock picks.

+ TEXT SIZE -

PRINT

EMAIL

SHARE

REPRINTS

GET RSS



Data: Performance of Cramer's stock picks versus market indices.

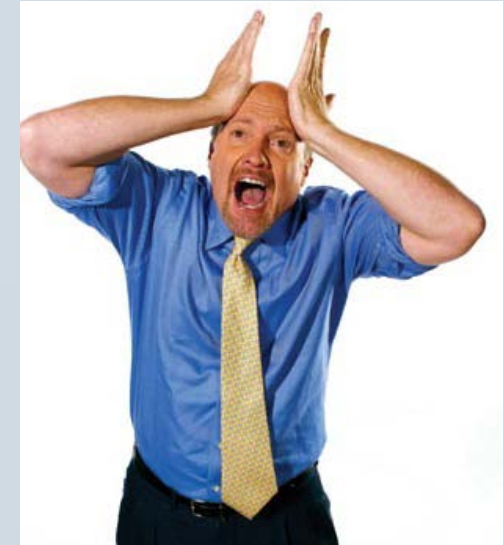
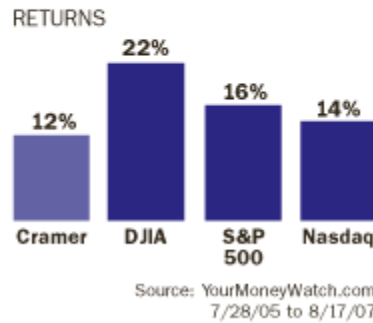
Findings: CNBC claimed that Cramer's picks beat the S&P 500, but they did not.

Cramer's picks did jump in value the day after broadcast, as his followers rushed to buy, but then slowly slid relative to the market. This points to a viable alternative investment strategy: short Cramer's picks to cash in on this trend.

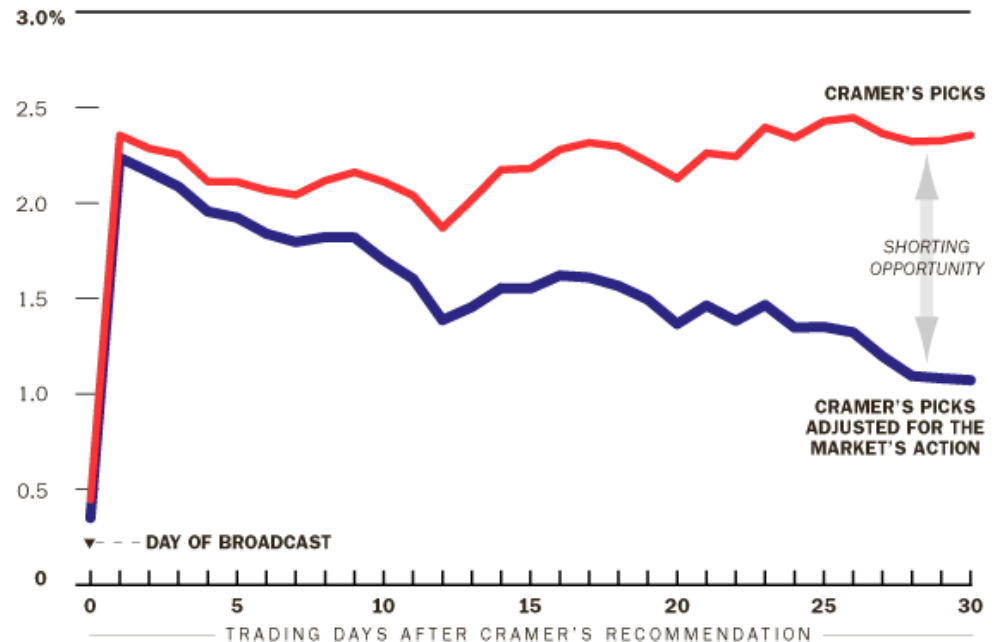
Methods: Read more from Bill Alpert's statistical adviser [Patrick Burns](#), and from Bill himself at the [R Journal](#).

Punk Performance

Jim Cramer's stock picks on CNBC's *Mad Money* have lagged the market over the past two years.

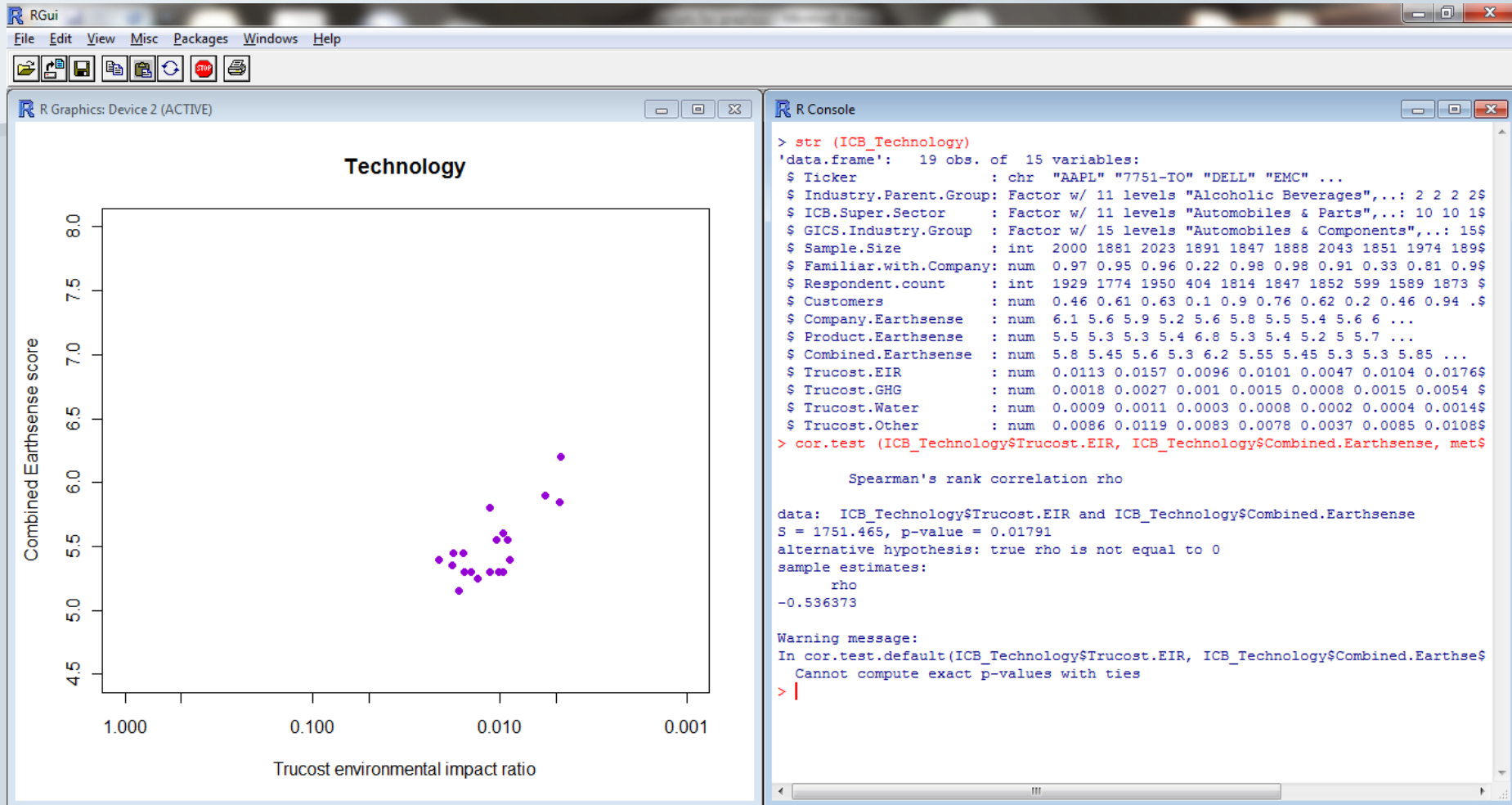


CUMULATIVE PERCENT RETURN FOR BUY RECOMMENDATION



Source: Barron's analysis based on 1,100 Cramer picks since July 28, 2005 from YourMoneyWatch.com.

My own first steps with R ...



SPECIAL INVESTIGATION

HEY, GREEN SPENDER

Do our ideas of which companies are eco-friendly live up to reality?

Peter Aldhous and Phil McKenna investigate



Data: Survey data on US public perceptions of corporate “greenness” from company called Earthsense. Quantitative assessment of same companies’ environmental impacts, from a company called Trucost.

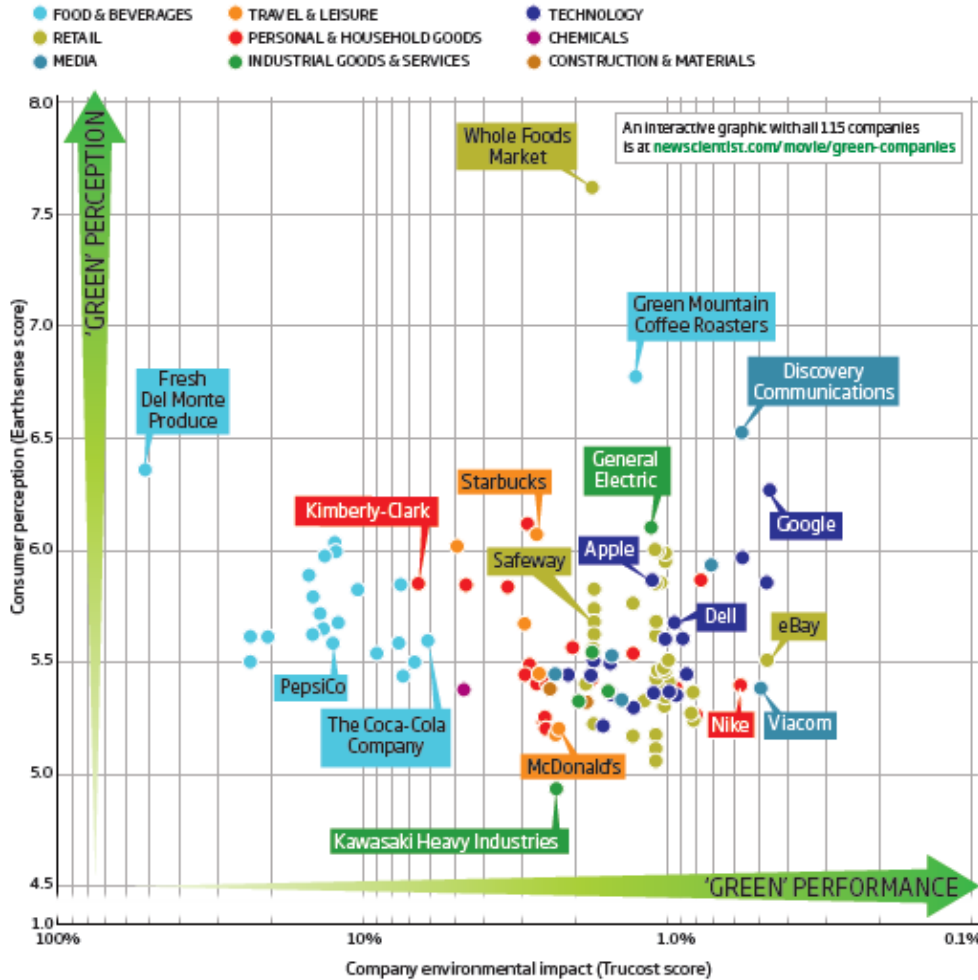
Findings: There are wide mismatches between public perceptions and reality. Some firms have undeserved “green” reputations, while others are not getting credit for fairly impressive efforts to reduce their environmental footprints. Greater disclosure of companies’ environmental impacts, plus improved awareness of these impacts by investors and consumer alike, may be needed to push businesses in a genuinely green direction.

Consumer perception and environmental realities

There is very little correlation between how green a company is and how green it is perceived to be

Earth sense score - 2008 survey asking 30,000 US consumers to rate greenness of companies and products on a scale of 1 to 10

Trucost score - Estimated cost of the environmental impact of a company under a "polluter pays" system as a percentage of its annual revenue



NewScientist

Methods: Spearman rank correlation; Kruskal-Wallis tests and multiple comparisons. [More details](http://news scientist.com) at news scientist.com.

[Explore](http://news scientist.com) interactive graphic.

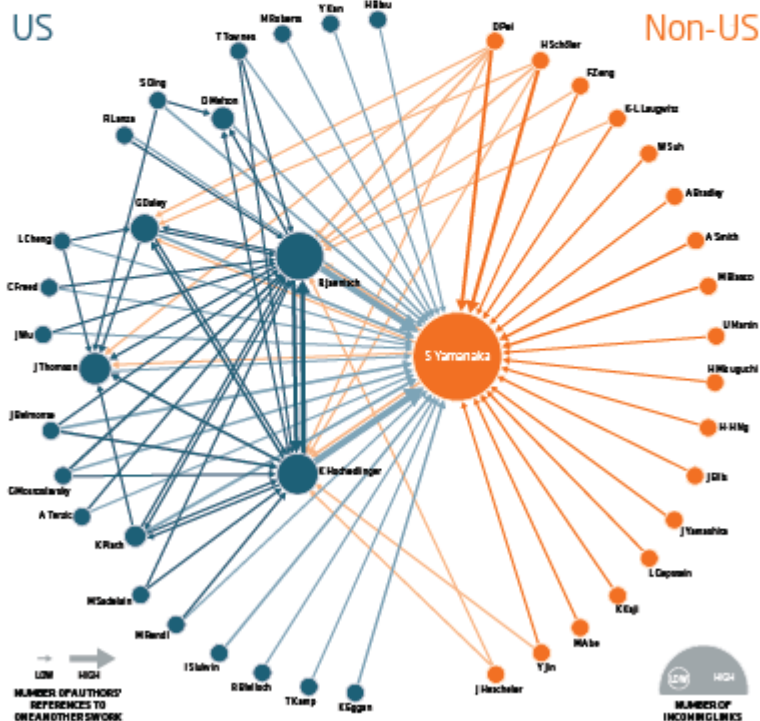
SPECIAL REPORT

For daily news stories, visit www.newscientist.com/news

THE STEM CELL WARS

When a Nobel prize is up for grabs, do scientists across the globe compete on a level playing field? **Peter Aldhous** investigates

Thirteen influential players in cellular reprogramming are revealed by recording how many times scientists have referred to each other's work. Each link shows where one researcher cited another four or more times in papers in leading journals (for analysis, see 'The strongest link', below right)



ALL'S fair in love and war, they say, but science is supposed to obey more noble ideals. New findings are submitted for publication, the studies are ferried out to experts for objective "peer review" and the best research appears prominently in the most prestigious journals.

Some stem cell biologists are crying foul, however. Last year, 14 researchers in this notoriously competitive field wrote to leading journals complaining of "unreasonable or obstructive reviews". The result, they claimed, is that "publication of truly original findings may be delayed or rejected".

Triggered by this protest, *New Scientist* scrutinised the dynamics of publication in the most exciting and competitive area of stem cell research, in which cells are "reprogrammed" to acquire the versatility of those of an early-stage embryo. In this fast-moving field, where a Nobel prize is arguably at stake, biologists are racing feverishly to publish their findings in top journals.

Our analysis of more than 300 research papers from 2006 onwards reveals that US-based scientists are enjoying a significant advantage, getting their papers published faster and in more prominent journals. The disparity is likely to spark debate when the International Society for Stem Cell Research (ISSCR) meets in San Francisco next week.

There are several plausible and reasonable explanations, but feelings are running high nonetheless. With two of the most delayed papers coming from a Japanese researcher who pioneered the field, and some of his rivals using controversial channels that give members of the US National Academy of Sciences an inside track to rapid publication, it is easy to see why.

The protest letter called for journals to publish the anonymised comments of researchers who act as reviewers of papers, to expose examples of potential obstruction. Just two of its signatories were from labs in the US. And when leaders of the protest talked to the media, unfair treatment of researchers outside the US was among the complaints. "There does seem to be this bias against groups from the rest of the world," Robin Lovell-Badge of the UK's National Institute for Medical Research in London told *New Scientist*.

Research on induced pluripotent stem (iPS) cells is the obvious place to look for biases in publication, given the high stakes involved. One of the signatories of the letter was the pioneer of cellular reprogramming—Shinya Yamanaka of Kyoto University in Japan. Less than four years after he first showed how to reprogram a mouse skin cell, Yamanaka is routinely mentioned as a candidate for a Nobel prize. He may be sharing that honour if other scientists make faster strides towards therapies based on cellular reprogramming. Our analysis of the citations between researchers reveals that Yamanaka is still the most influential figure in the field, but also shows that several well-connected US-based scientists are giving him a run for his money (see diagram, left, and "The strongest link", below).

New Scientist searched the Web of

THE STRONGEST LINK

Shinya Yamanaka of Kyoto University in Japan is the dominant scientist in cellular reprogramming, but he has stiff competition from a well-linked group of US-based researchers.

To map influence in the field, *New Scientist* created a social network diagram (left) based on citations. The references to each scientist's work by their peers. Citations are a measure of a researcher's impact and influence, and are sometimes used to help make decisions on promotions. They can also provide a snapshot of who's who in a field.

Aided by Henri Schiltz of Imperial College London, a specialist in citation analysis, we looked at references between 148 papers published in prominent journals since 2005—drawing links where the authors cited one another's work four or more times.

Yamanaka's research is referred to by just about everyone. But there are no such links between other scientists outside of the US.

Of papers published in leading journals:

78
per cent from US-based authors accepted within 100 days

54
per cent from authors elsewhere accepted within 100 days

Science database for studies on iPS cells, recording the dates each was submitted, accepted for publication, and published. Aided by Matthew Strickland of Emory University in Atlanta, Georgia, whose research employs a branch of statistics called survival analysis, we found that papers submitted by authors outside the US took significantly longer to be accepted and published.

This difference was particularly clear for papers in 20 high-profile journals with an "impact factor" of 5 or more (see "What's the hold-up?", page 14). Impact factor is a measure of the frequency with which a journal's articles are cited in the scientific literature. We chose this cut-off score to focus on journals that received the protest letter, or those with similar prominence. "It's really very interesting," says Lovell-Badge. "I didn't think it would be possible to get quantitative data."

So what might explain the pattern? Obstruction of papers could happen if a reviewer delays their comments, or makes many demands for changes. The journals with the greatest lag between US and non-US papers deny that their reviews are biased, and say that the former explanation can be ruled out. "The review process itself is quite short, and the majority of the intervening time is taken up by the authors performing revisions to address the criticisms that the reviewers raised," says Deborah Sweet, editor of *Cell Stem Cell*, which is the official journal of ISSCR—and is published by Elsevier, a sister company of the publisher of *New Scientist*.

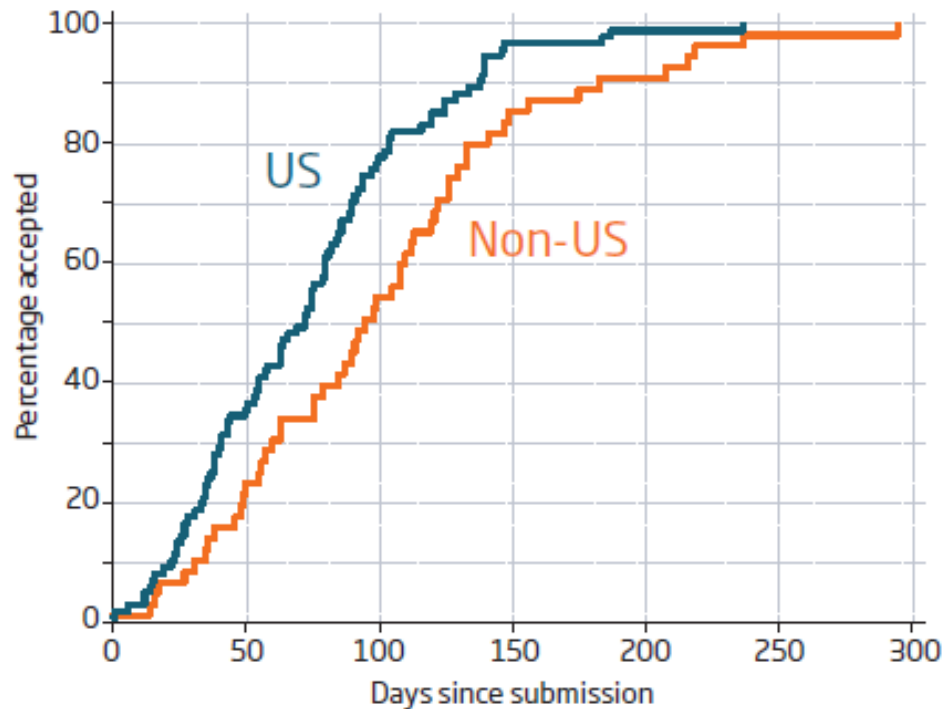
Bigger, better?

Konrad Hochdinger of the Harvard Stem Cell Institute, among the foremost in the field, suggests that leading US labs can deal with revisions more easily because they are often larger and better funded. We then asked to run extra experiments, for example, he can quickly deploy junior scientists or hire a commercial lab to do them.

Could it also be that the US-based scientists tend to produce better work? There's no simple way to measure the calibre of studies in our sample—and if papers are rejected by leading journals and end up lower down the

What's the hold-up?

In a sample of 148 papers from high-profile journals, those from scientists outside the US took longer to be accepted for publication



Data: Time-to-acceptance for original research papers involving “iPS” cells – an exciting alternative to embryonic stem cells.

Findings: Papers from corresponding authors outside the US took significantly longer to be accepted for publication. US-based authors were also better at getting papers into high-impact journals.

Methods: Cox proportional hazards regression; Kaplan-Meier survival curves. [More details](#) at newscientist.com.

R package: [survival](#) (survival analysis).

Why aren't more data journalists using R?

- **Seen as difficult/arcane**

"It's like nothing you've ever encountered."

Experienced computer-assisted reporter and web developer

- **A dangerous tool, in the hands of journalists?**

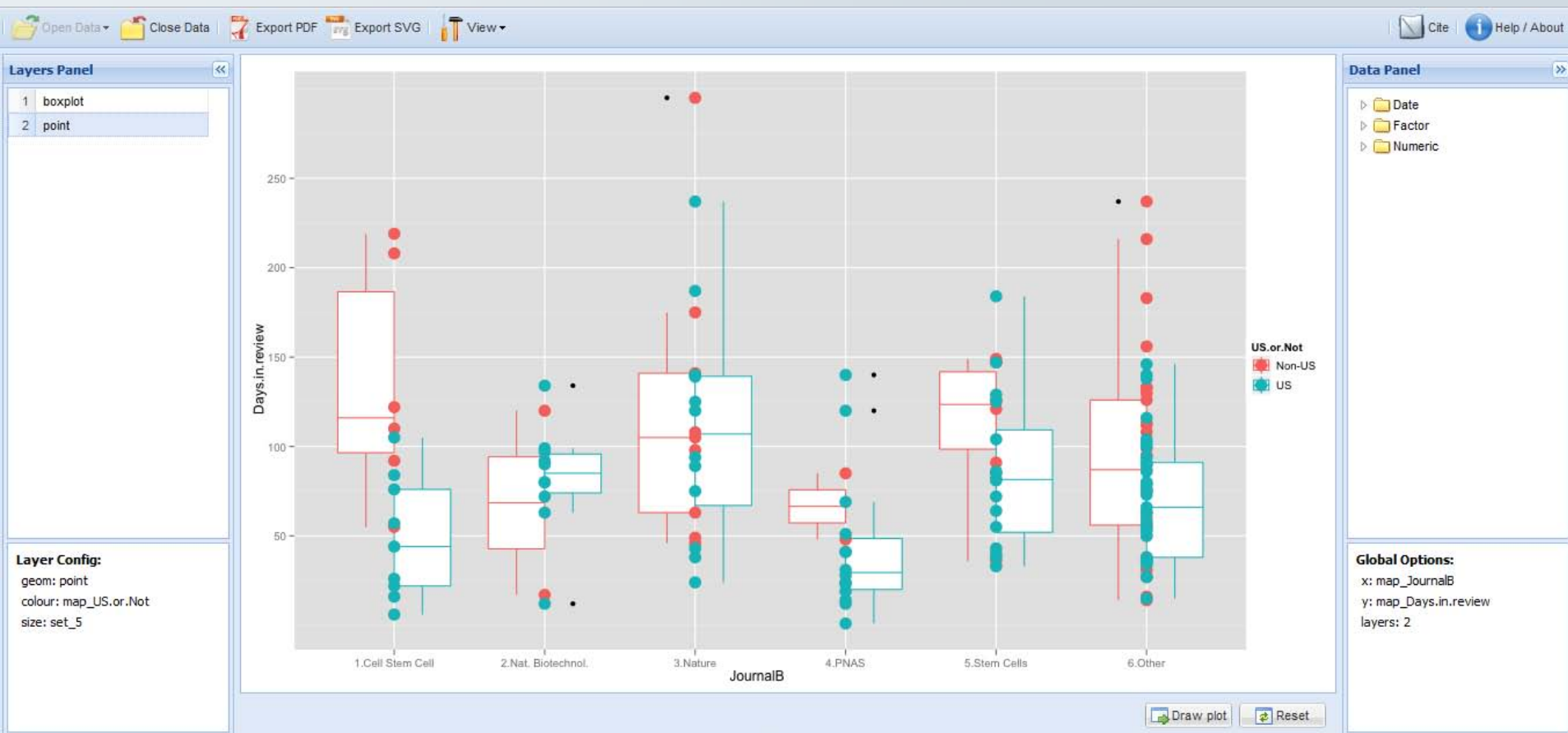
"I'm concerned that you're giving them a chainsaw."

Professor of science journalism



Breaking down the barriers

- User-friendly interfaces, e.g. Jeroen Ooms' [ggplot2 application](#):



- Collaboration!



Collaboration: feature on predictive analytics

NewScientist

Revolutions

News about R, statistics and the world of open source from the staff of Revolution Analytics

[« Because it's Friday: How Machines Work | Main | R is indispensable, because it's reproducible »](#)

August 30, 2010

Wanted: R Analysis of New Scientist Covers

Peter Aldhous and Jim Giles – from *New Scientist's* San Francisco bureau – are looking for a statistician and R user to take part in an interesting data analysis challenge, and also be part of a future article in the magazine. They were inspired by this rather tongue-in-cheek presentation where [Sebastian Wernicke](#) analyzed videos, transcripts and ratings of TED talks to conclude, for example, that a talk about how "French coffee spreads happiness in your brain" would be the "ultimate TED talk".

The challenge ...

Top sellers

Bottom sellers



VS



Data-driven journalism

Bay Area R Users Group, Dec 14 2010

Slides at: www.peteraldhous.com/CAR/R_Users_Dec2010.pdf

Peter Aldhous,
San Francisco Bureau Chief



NewScientist