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?

Watch the video.
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
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.
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.

100 titles that were frequently rented from Netflix in 2009

Milk

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

Explore the interactive.
Beauty is not enough

The Fortune 500 Cosmos
An infographic illustrating the performance of companies alphabetically, according to their ticker symbols or a clock-like spiral order. For private companies, we created abbreviations (see list below).

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

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 21st-century American idea. Forget about building great companies that endure, they tell me. One technology pundit compared me at a conference and damned the whole premise ahead as today’s world: “We live in an era when nothing can be built to last. Everything is in flux; nothing can sustain.”

He invoked Joseph Schumpeter, the great economist who wrote about the “potential cycle of creative destruction” wherein technological change and visionary entrepreneurs give birth to new things that obliterate old things, only to see these new things become obliterated by the next generation.

His argument feels particularly sharp today. Ross Steam disappeared over
You need to tell a story

Watch the video.
Who are the data journalists?

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.

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.
And what are their skills/interests?

10 days on the NICAR listserv, Nov-Dec 2010
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

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.
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.
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

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.
**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.
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?

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*
Jackson’s Billboard Rankings Over Time
A timeline of how Michael Jackson's songs performed on the Billboard Hot 100 chart.

Explore the interactive.
Watch the video.
**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).
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.

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.
**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 [R Journal](#).
My own first steps with R ...

```
> str (ICB_Technology)
'vector' of 15 variables:
  $ Ticker : chr "AAPL" "7751-TG" "DELL" "EMC" ...
  $ Industry.Parent.Group: Factor w/ 11 levels "Alcoholic Beverages",...: 2 2 2 2 2
  $ ICB.Super.Sector : Factor w/ 11 levels "Automobiles & Parts": 10 10 10
  $ GICS.Industry.Group : Factor w/ 11 levels "Automobiles & Components": 118
  $ Familiar.with.Company: num 0.97 0.98 0.98 0.98 0.96 0.91 0.93 0.91 0.89 0.94
  $ Respondent.count : int 1829 1774 1950 404 1814 1847 1852 599 1589 1873
  $ Customers        : num 0.56 0.62 0.63 0.1 0.9 0.76 0.62 0.2 0.46 0.94
  $ Company.Earthsense: num 6.1 5.6 5.9 5.2 5.6 5.0 5.5 5.4 5.6 6...
  $ Product.Earthsense: num 5.5 5.3 5.3 5.4 6.8 5.3 5.4 5.2 5.5 7...
  $ Combined.Earthsense: num 5.8 5.45 5.6 5.3 6.2 5.55 5.45 5.3 5.3 5.85
  $ Trucost.EIR      : num 0.0119 0.0157 0.0096 0.0101 0.0047 0.0104 0.0176
  $ Trucost.GHG      : num 0.0018 0.0027 0.001 0.0015 0.0008 0.0015 0.0054
  $ Trucost.Water    : num 0.0008 0.0011 0.0003 0.0008 0.0002 0.0004 0.0014
  $ Trucost.Other    : num 0.0086 0.0119 0.0083 0.0078 0.0037 0.0038 0.0108

> cor.test (ICB_Technology$Trucost.EIR, ICB_Technology$Combined.Earthsense, method="spearman")

Spearman's rank correlation rho

data:  ICB_Technology$Trucost.EIR and ICB_Technology$Combined.Earthsense
S = 1751.465, p-value = 0.0179
alternative hypothesis: true rho is not equal to 0
sample estimates:
 rho
0.536379

Warning message:
In cor.test.default(ICB_Technology$Trucost.EIR, ICB_Technology$Combined.Earthsense) Cannot compute exact p-values with ties

```
Data: Survey data on US public perceptions of corporate “greenness” from a 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.
**Methods:** Spearman rank correlation; Kruskal-Wallis tests and multiple comparisons. [More details](https://newscientist.com) at newscientist.com.

[Explore](https://newscientist.com) interactive graphic.
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

The stem cell wars are a key battleground in the new strategy in the field. On one side, the US-based researchers, who are leading the charge in the field, are trying to establish their dominance. On the other side, the non-US researchers, who are trying to catch up, are aiming to challenge the US dominance.

Research on induced pluripotent stem (iPS) cells is in the obvious plan to look for biases in publication, given the high number of papers published in leading journals. The Web of Science database for studies on iPS cells, recording the dates each was submitted, accepted for publication, and published. 


The New Scientist searched the Web of Science for studies on iPS cells, recording the dates each was submitted, accepted for publication, and published. 


daughter companies, including the US-based researchers, who are leading the charge in the field, are trying to establish their dominance. On the other side, the non-US researchers, who are trying to catch up, are aiming to challenge the US dominance.

Research on induced pluripotent stem (iPS) cells is in the obvious plan to look for biases in publication, given the high number of papers published in leading journals. The Web of Science database for studies on iPS cells, recording the dates each was submitted, accepted for publication, and published.

The New Scientist searched the Web of Science for studies on iPS cells, recording the dates each was submitted, accepted for publication, and published.
**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

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
Data-driven journalism

Bay Area R Users Group, Dec 14 2010


Peter Aldhous,
San Francisco Bureau Chief