

# **Investigating science**

**(with help from data)**

**Peter Aldhous,  
Science reporter, BuzzFeed News**

[peter@peteraldhous.com](mailto:peter@peteraldhous.com)  
[@paldhous](#)

# Is this data journalism?

NewScientist

Map: 1994 - 2013

London

Go

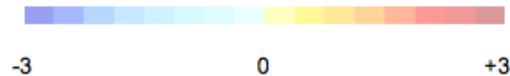
## YOUR WARMING WORLD

The heat is on for the planet as a whole, but what has been happening where you live? Click on the map to find out, or enter a location in the search box at top right.

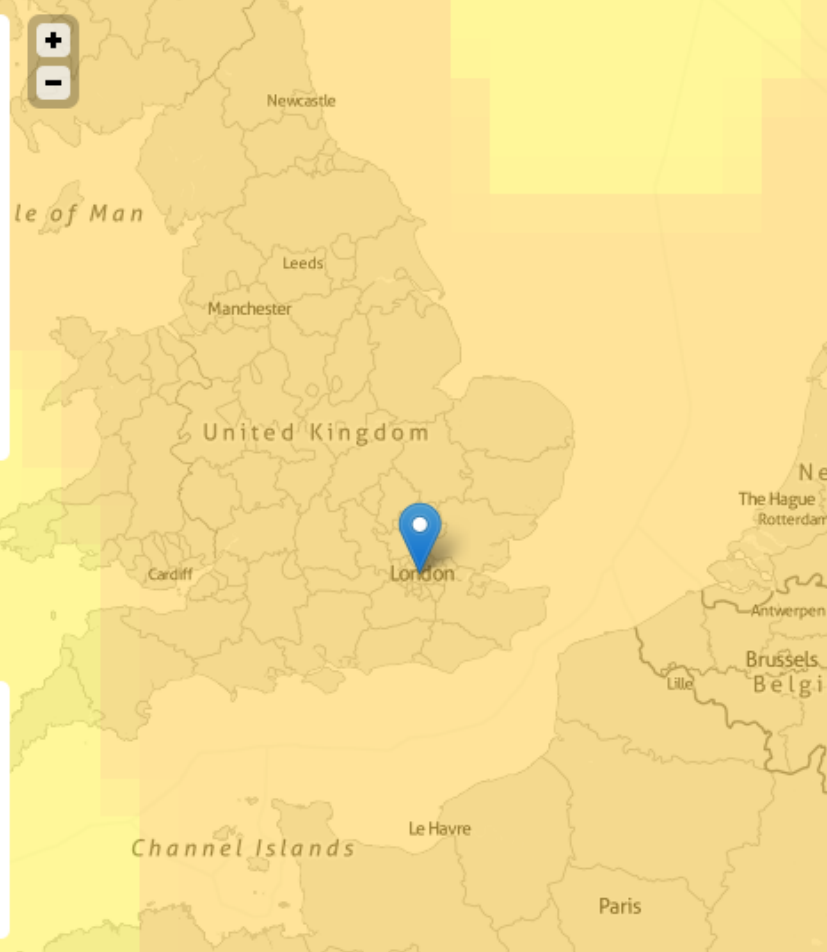
The initial map shows average temperatures over the past 20 years; use the drop-down menu to see maps for earlier periods.

**More:** Read our [climate change topic guide](#) and [learn about](#) the data and graphic.

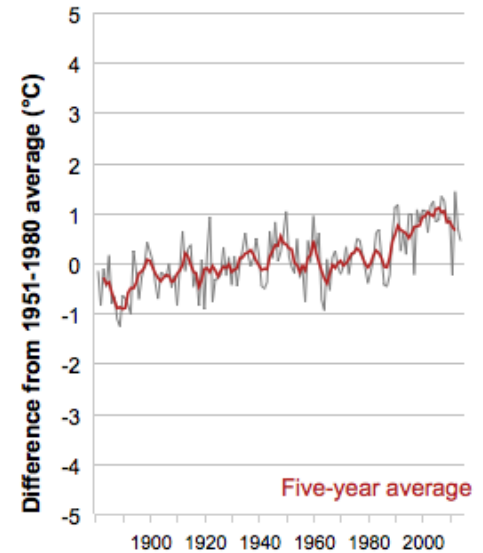
### Difference from 1951-1980 average (°C)



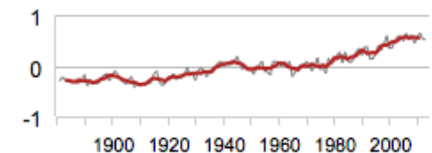
Source: NASA Goddard Institute for Space Studies Surface Temperature Analysis



### At this location



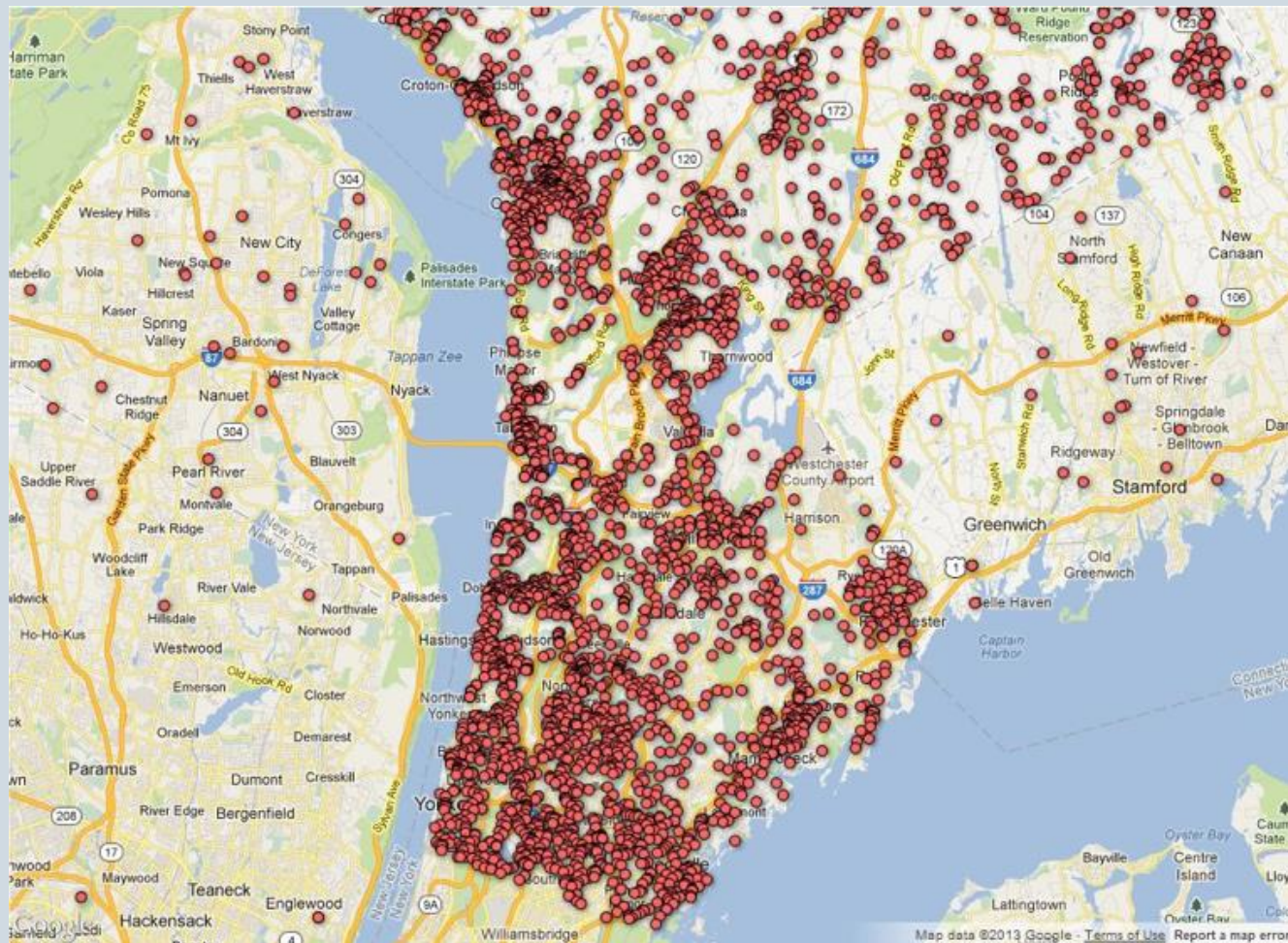
### Global average



# Is this data journalism?

## Map: Where are the gun permits in your neighborhood?

FILED UNDER - News / Local News | 3:51 PM, Dec. 22, 2012



# Is this data journalism?

The New York Times

U.S.

Search All NYTimes.com

Go

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION ARTS STYLE TRAVEL JOBS REAL ESTATE AUTOS  
POLITICS EDUCATION TEXAS

BREAKDOWN | *Death and disarray at America's racetracks*

## Mangled Horses, Maimed Jockeys

The new economics of horse racing are making an always-dangerous game even more so, as lax oversight puts animal and rider at risk.

By WALT BOGDANICH, JOE DRAPE, DARA L. MILES and GRIFFIN PALMER  
Published: March 24, 2012 | 481 Comments

RUIDOSO, N.M. — At 2:11 p.m., as two ambulances waited with motors running, 10 horses burst from the starting gate at Ruidoso Downs Race Track 6,900 feet up in New Mexico's Sacramento Mountains.

Nineteen seconds later, under a brilliant blue sky, a national champion jockey named Jacky Martin lay sprawled in the furrowed dirt just past the finish line, paralyzed, his neck broken in three places. On the ground next to him, his frightened horse, leg broken and chest heaving, was minutes away from being euthanized on the track.

For finishing fourth on this early September day last year, Jacky Martin got about \$60 and possibly a lifetime tethered to a respirator.

f FACEBOOK

t TWITTER

g+ GOOGLE+

SAVE

EMAIL

SHARE

PRINT

REPRINTS

CALVARY  
AUGUST 1  
WATCH TRAILER



### A Jockey's Story

Jockeys are on the front lines of the risky business of horse racing. Twenty-four horses a week die at racetracks around the country.

Video by Matthew Orr/The New York Times



# Data + journalism = story

The New York Times

U.S.

Search All NYTimes.com

Go

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION ARTS STYLE TRAVEL JOBS REAL ESTATE AUTOS  
POLITICS EDUCATION TEXAS

BREAKDOWN | *Death and disarray at America's racetracks*

## Mangled Horses, Maimed Jockeys

The new economics of horse racing are making an always-dangerous game even more so, as lax oversight puts animal and rider at risk.

But an investigation by The New York Times has found that industry practices continue to put animal and rider at risk. A computer analysis of data from more than 150,000 races, along with injury reports, drug test results and interviews, shows an industry still mired in a culture of drugs and lax regulation and a fatal breakdown rate that remains far worse than in most of the world.

# ... which can be told visually

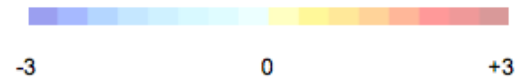
## YOUR WARMING WORLD

The heat is on for the planet as a whole, but what has been happening where you live? Click on the map to find out, or enter a location in the search box at top right.

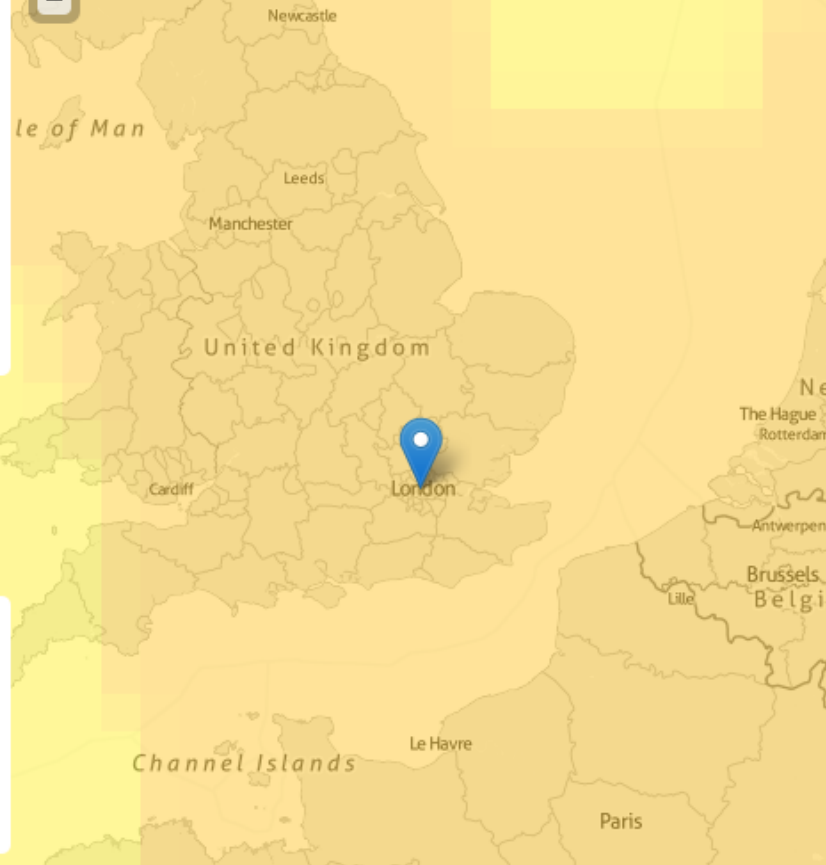
The initial map shows average temperatures over the past 20 years; use the drop-down menu to see maps for earlier periods.

**More:** Read our [climate change topic guide](#) and [learn about](#) the data and graphic.

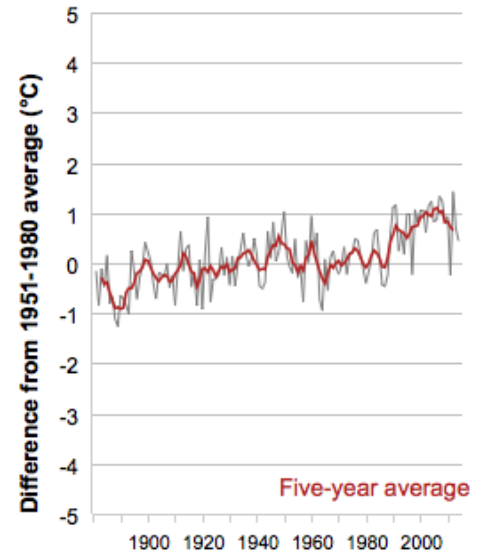
### Difference from 1951-1980 average (°C)



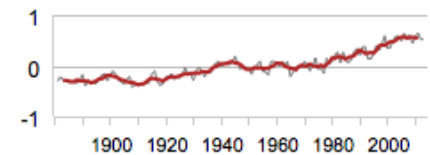
Source: NASA Goddard Institute for Space Studies Surface Temperature Analysis



### At this location



### Global average



# **But it's not only maps and graphs!**

Data can be used in both reporting and storytelling.

And visualization can be a powerful for both finding and telling stories.

But think carefully about what you need to show to your audience. Some of the best data-driven stories contain little in the way of numbers or graphs.

# Have you ...

- Filed a request for public records?
- Worked on a story that required a spreadsheet to analyze data?
- Worked on a story that relied on querying an online database?
- Obtained a database and analyzed with database management software, or built your own database?
- Used mapping software for a story?
- Used statistical analysis for a story?
- Studied connections between people/organizations using network analysis?



# A new thing?

## 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

# A thing of the future?

guardian.co.uk

Media

[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

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



### 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

Last 24 hours



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

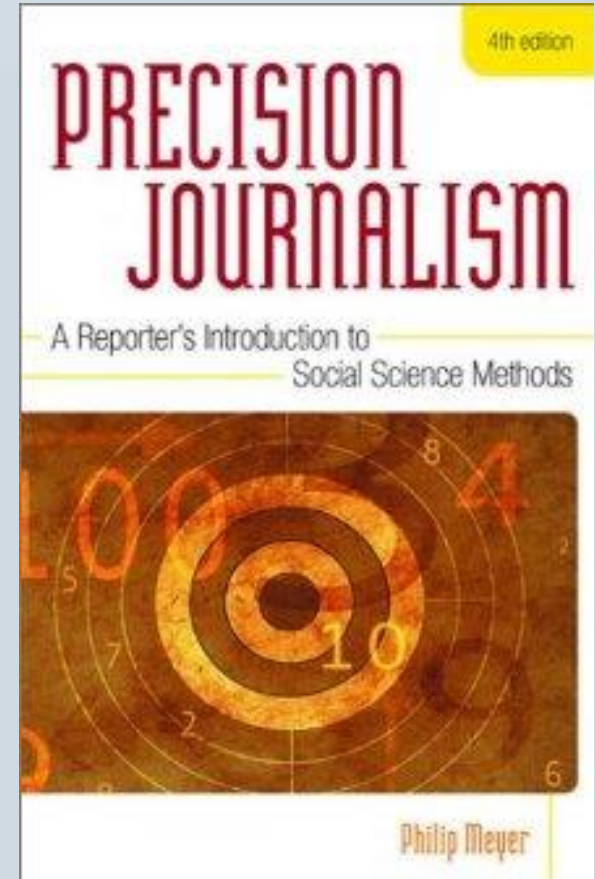
# The pioneer: Philip Meyer



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

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

**Findings:** One theory held that the rioters were reacting to being stuck at the foot of the economic ladder. Another blamed southern blacks who had moved to Detroit. 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.



- 43 dead
- 467 injured
- 7231 arrests

# What's in it for me?

- Place your other reporting in context. Less “he said; she said.”
- Visualize complex stories: Fresh understanding; new points of entry
- Find original stories, new angles



# Where do I start?

Usually, with a question you want to answer, or a point you want to demonstrate.

Good data journalism rarely starts by aimlessly poking at a dataset.

Approach data like you would an interview: What do you and your audience want to know?

# The data frame of mind

- When you start working on a story, think “What sources of data are available?” as well as “Who can I speak to about this?”
- Assume the data you need exists and is open to the public until proven otherwise.
- Make it a regular practice to learn about sources of data related to your beat.
- If necessary, have a plan for acquiring data at regular intervals. Some data may require public records requests.

# The data frame of mind

This very different to:

“I’ve written my story. Now I’d better find some numbers for a graph.”

## **A note of caution: data is often 'dirty'**

Data can be seductive, but never simply assume that it is correct and consistent. Examine any data you obtain to see how it is organized, and scan for potential errors.

You will almost always need to reformat and edit data to suit your purposes; frequently you will have to do extensive data “cleaning”.

# Please clean me!

|    | A        | B             | C          | D       | E    | F      | G                          | H                          | I             | J     | K        | L       | M            | N    |
|----|----------|---------------|------------|---------|------|--------|----------------------------|----------------------------|---------------|-------|----------|---------|--------------|------|
| 1  | REVIEWER |               |            | MIDDLE  |      |        |                            |                            |               |       |          |         |              |      |
| 2  | ID       | LAST NAME     | FIRST NAME | INITIAL | RANK | DEGREE | SITE                       | STREET ADDRESS             | CITY          | STATE | ZIP CODE | COUNTRY | RECEIPT DATE | TYPE |
| 2  | 459203   | %BENN%        | TERRY      | L       | NG   | MD     | RANDOLPH FAMILY PRACTICE   | 1918 RANDOLPH RD STE 275   | CHARLOTTE     | NC    | 28207    | US      | 12/5/2001    | DEM  |
| 3  | 533704   | %EL-GHOROURY% | MOHAMMAD   |         | NG   | MD     | NG                         | 22201 MOROSS STE 150       | DETROIT       | MI    | 48236    | US      | 2/11/2011    | DEM  |
| 4  | 512096   | %GUENTHER     | RAINER     |         | NG   | MD     | UNIVERSITATSKLINIKUM SCHLE | SCHITTENHELMSTR 12         | KIEL          | NG    | 24105    | GM      | 11/19/2007   | DEM  |
| 5  | 16648    | %RIBOT%       | THOMAS     | L       | NG   | MD     | ARNETT                     | 2600 GREENBUSH ST          | LAFAYETTE     | IN    | 47904    | US      | 3/7/2000     | DEM  |
| 6  | 16648    | %RIBOT%       | THOMAS     | L       | NG   | MD     | ARNETT                     | 2600 GREENBUSH ST          | LAFAYETTE     | IN    | 47904    | US      | 5/5/2000     | DEM  |
| 7  | 16648    | %RIBOT%       | THOMAS     | L       | NG   | MD     | ARNETT                     | 2600 GREENBUSH ST          | LAFAYETTE     | IN    | 47904    | US      | 8/21/1981    | DEM  |
| 8  | 16648    | %RIBOT%       | THOMAS     | L       | NG   | MD     | ARNETT                     | 2600 GREENBUSH ST          | LAFAYETTE     | IN    | 47904    | US      | 9/11/2003    | DEM  |
| 9  | 16648    | %RIBOT%       | THOMAS     | L       | NG   | MD     | ARNETT                     | 2600 GREENBUSH ST          | LAFAYETTE     | IN    | 47904    | US      | 6/9/1998     | DEM  |
| 10 | 16648    | %RIBOT%       | THOMAS     | L       | NG   | MD     | ARNETT                     | 2600 GREENBUSH ST          | LAFAYETTE     | IN    | 47904    | US      | 5/29/1998    | DEM  |
| 11 | 16648    | %RIBOT%       | THOMAS     | L       | NG   | MD     | ARNETT                     | 2600 GREENBUSH ST          | LAFAYETTE     | IN    | 47904    | US      | 3/12/2003    | DEM  |
| 12 | 499673   | %RICHARDSON   | MARTIN     | D       | NG   | MD     | THE ROYAL MELBOURNE HOSP/  | GRATTAN ST                 | PARKVILLE     | NG    | 3050     | AS      | 5/12/2006    | DEM  |
| 13 | 534551   | %TAUTH        | JEFFREY    |         | NG   | MD     | NG                         | 180 MEDICAL PARK DRIVE     | HOT SPRINGS   | AR    | 71901    | US      | 4/11/2011    | DEM  |
| 14 | 394897   | ,AAVEDRA      | LILLIAN    | T       | NG   | MD     | NG                         | 1315 S ORANGE AVE STE 3E   | ORLANDO       | FL    | 32806    | US      | 3/16/2004    | DEM  |
| 15 | 394897   | ,AAVEDRA      | LILLIAN    | T       | NG   | MD     | NG                         | 1315 S ORANGE AVE STE 3E   | ORLANDO       | FL    | 32806    | US      | 2/5/1993     | DEM  |
| 16 | 344230   | .EVINE        | KENNETH    | A       | NG   | MD     | NG                         | 1551 N PALM AVE            | PEMBROKE PINI | FL    | 33026    | US      | 8/30/1988    | DEM  |
| 17 | 514421   | .WENS         | SHEMETRA   |         | NG   | NG     | MCLEAN HOSP                | 115 MILL STREET            | BELMONT       | MA    | 2478     | US      | 5/15/2008    | IRB  |
| 18 | 514421   | .WENS         | SHEMETRA   |         | NG   | NG     | MCLEAN HOSP                | 115 MILL STREET            | BELMONT       | MA    | 2478     | US      | 5/20/2008    | IRB  |
| 19 | 514421   | .WENS         | SHEMETRA   |         | NG   | NG     | MCLEAN HOSP                | 115 MILL STREET            | BELMONT       | MA    | 2478     | US      | 1/9/2009     | IRB  |
| 20 | 514421   | .WENS         | SHEMETRA   |         | NG   | NG     | MCLEAN HOSP                | 115 MILL STREET            | BELMONT       | MA    | 2478     | US      | 3/23/2009    | IRB  |
| 21 | 514421   | .WENS         | SHEMETRA   |         | NG   | NG     | MCLEAN HOSP                | 115 MILL STREET            | BELMONT       | MA    | 2478     | US      | 4/27/2010    | IRB  |
| 22 | 514421   | .WENS         | SHEMETRA   |         | NG   | NG     | MCLEAN HOSP                | 115 MILL STREET            | BELMONT       | MA    | 2478     | US      | 11/5/2009    | IRB  |
| 23 | 514421   | .WENS         | SHEMETRA   |         | NG   | NG     | MCLEAN HOSP                | 115 MILL STREET            | BELMONT       | MA    | 2478     | US      | 3/10/2011    | IRB  |
| 24 | 514421   | .WENS         | SHEMETRA   |         | NG   | NG     | MCLEAN HOSP                | 115 MILL STREET            | BELMONT       | MA    | 2478     | US      | 2/18/2011    | IRB  |
| 25 | 514421   | .WENS         | SHEMETRA   |         | NG   | NG     | MCLEAN HOSP                | 115 MILL STREET            | BELMONT       | MA    | 2478     | US      | 10/16/2009   | IRB  |
| 26 | 514421   | .WENS         | SHEMETRA   |         | NG   | NG     | MCLEAN HOSP                | 115 MILL STREET            | BELMONT       | MA    | 2478     | US      | 2/1/2010     | IRB  |
| 27 | 514421   | .WENS         | SHEMETRA   |         | NG   | NG     | MCLEAN HOSP                | 115 MILL STREET            | BELMONT       | MA    | 2478     | US      | 3/20/2008    | IRB  |
| 28 | 514421   | .WENS         | SHEMETRA   |         | NG   | NG     | MCLEAN HOSP                | 115 MILL STREET            | BELMONT       | MA    | 2478     | US      | 7/2/2009     | IRB  |
| 29 | 532708   | ,AW           | IAN        |         | NG   | MD     | RIGSHOPITALET COPENHAGEN,  | 9 BLEGDAMSVEJ              | COPENHAGEN    | NG    | 2100     | DA      | 11/15/2010   | DEM  |
| 30 | 307380   | ??            | ADAM       | R       | NG   | MD     | UNIV COLORADO/COLORADO     | 14200/4700 E 9TH AVE BOX C | DENVER        | CO    | 80262    | US      | 6/9/1999     | DEM  |
| 31 | 307380   | ??            | ADAM       | R       | NG   | MD     | UNIV COLORADO/COLORADO     | 14200/4700 E 9TH AVE BOX C | DENVER        | CO    | 80262    | US      | 12/10/1998   | DEM  |



# But science journalists are lucky: Lots of clean, well curated data ...

Storm ARLENE is number 1 of the year 2011

\*\*\*\*\*

| Month | Day | Hour   | Lat.  | Long. | Dir.    | ----Speed----- | -----Wind----- | Pressure | -----Type-----      |
|-------|-----|--------|-------|-------|---------|----------------|----------------|----------|---------------------|
| June  | 28  | 6 UTC  | 19.9N | 92.8W | -- deg  | -- mph -- kph  | 30 mph 45 kph  | 1007 mb  |                     |
| June  | 28  | 12 UTC | 20.3N | 93.1W | 325 deg | 4 mph 7 kph    | 35 mph 55 kph  | 1006 mb  |                     |
| June  | 28  | 18 UTC | 20.7N | 93.5W | 315 deg | 5 mph 9 kph    | 40 mph 65 kph  | 1006 mb  | Tropical Storm      |
| June  | 29  | 0 UTC  | 21.0N | 93.9W | 310 deg | 4 mph 7 kph    | 40 mph 65 kph  | 1005 mb  | Tropical Storm      |
| June  | 29  | 6 UTC  | 21.2N | 94.5W | 290 deg | 5 mph 9 kph    | 40 mph 65 kph  | 1003 mb  | Tropical Storm      |
| June  | 29  | 12 UTC | 21.3N | 95.3W | 280 deg | 8 mph 12 kph   | 50 mph 85 kph  | 1000 mb  | Tropical Storm      |
| June  | 29  | 18 UTC | 21.4N | 95.6W | 290 deg | 2 mph 3 kph    | 60 mph 95 kph  | 998 mb   | Tropical Storm      |
| June  | 30  | 0 UTC  | 21.6N | 96.1W | 295 deg | 5 mph 9 kph    | 60 mph 95 kph  | 996 mb   | Tropical Storm      |
| June  | 30  | 6 UTC  | 21.6N | 97.0W | 270 deg | 9 mph 14 kph   | 65 mph 100 kph | 994 mb   | Tropical Storm      |
| June  | 30  | 12 UTC | 21.6N | 97.3W | 270 deg | 2 mph 3 kph    | 65 mph 100 kph | 993 mb   | Tropical Storm      |
| June  | 30  | 18 UTC | 21.5N | 98.1W | 260 deg | 8 mph 12 kph   | 50 mph 85 kph  | 998 mb   | Tropical Storm      |
| July  | 1   | 0 UTC  | 21.1N | 98.7W | 235 deg | 6 mph 11 kph   | 35 mph 55 kph  | 1002 mb  | Tropical Depression |

Storm BRET is number 2 of the year 2011

\*\*\*\*\*

| Month | Day | Hour   | Lat.  | Long. | Dir.    | ----Speed----- | -----Wind----- | Pressure | -----Type----- |
|-------|-----|--------|-------|-------|---------|----------------|----------------|----------|----------------|
| July  | 16  | 6 UTC  | 30.7N | 79.7W | -- deg  | -- mph -- kph  | 25 mph 35 kph  | 1014 mb  |                |
| July  | 16  | 12 UTC | 30.3N | 79.4W | 145 deg | 4 mph 7 kph    | 25 mph 35 kph  | 1014 mb  |                |

# ...which scientists may be willing to share

NewScientist

[Back to story](#)

## Swarmageddon: America's periodical cicadas

In 2013, parts of the eastern US will be engulfed by a giant brood of cicadas. The insects will emerge in their hundreds of millions to sing and mate, after spending 17 years developing underground.

Use the controls at the top right of the graphic to explore when and where all the cicada broods emerge.

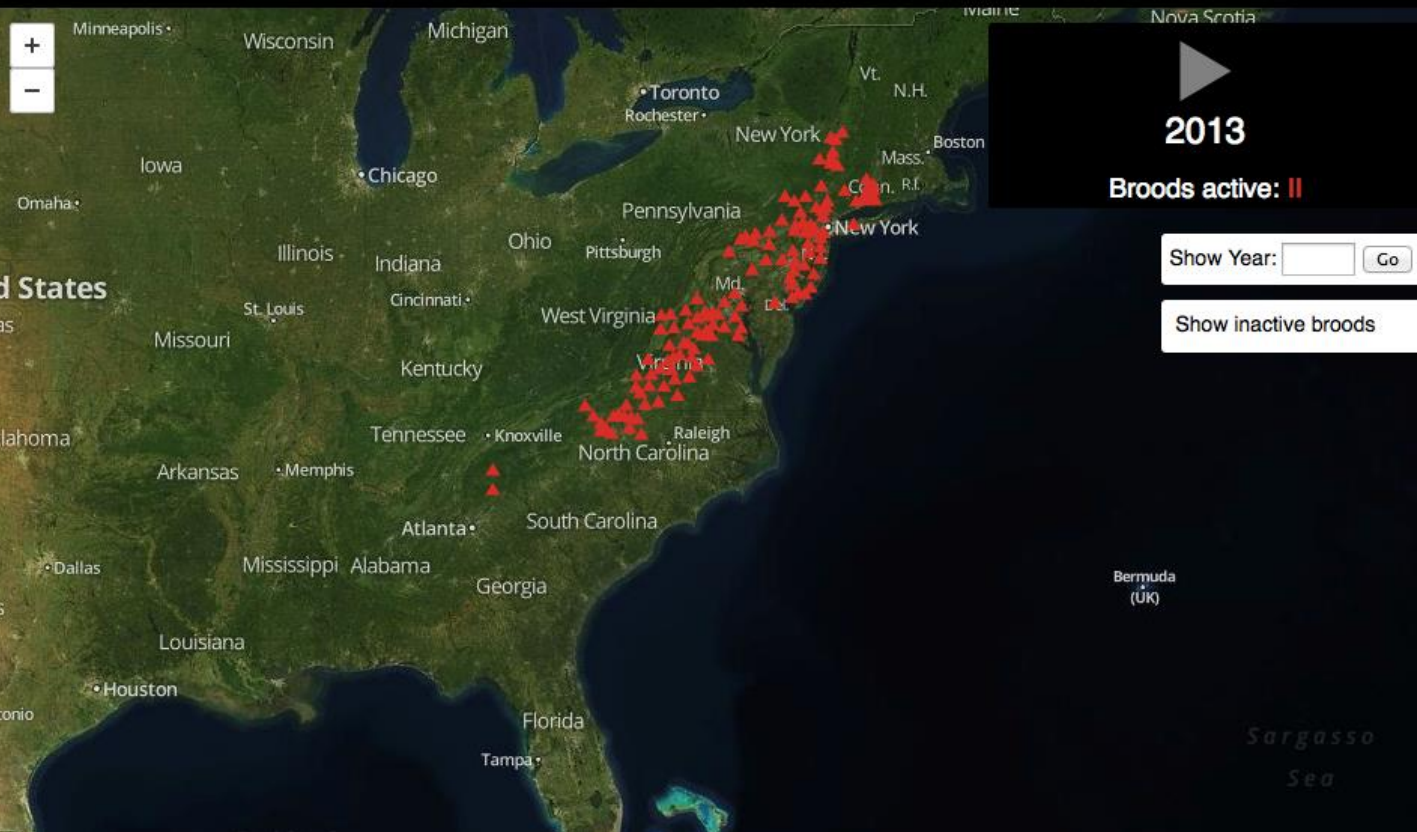
### Legend:

▲ 17-year broods ▼ 13-year broods

### Listen to a cicada chorus:



Source: [John Cooley](#) and [Chris Simon](#), University of Connecticut; graphic by Adam Becker and Peter Aldous, published 3 May 2013.



Made with [Mapbox](#)

# The basics

## (OK, I have some data. What now?)

- **Sort**

Largest to smallest; Alphabetical etc

- **Aggregate**

Count, Sum, Mean, Median, Maximum, Minimum etc

- **Filter**

Select a defined subset of the data

- **Join**

Merge entries from two or more datasets based on common field(s), e.g. unique ID number, last name and first name

**Think of those operations as  
'interviewing' the data**

# Tools and stories: online databases

Special report **Bioterror**

## FRIEND OR FOE?

Efforts to combat killer pathogens with new vaccines and drugs could be inadvertently writing a handbook for biowarfare. The US, home to many such "dual-use" projects, faces a tough dilemma

**"At what point, if any, does working on how pathogens evade immunity become a threat to national security?"**

PETER ALDHOUS

TWEAKING the anthrax toxin to render experimental drugs ineffective. Turning a harmless rodent virus into a deadly pathogen. Enhancing the potency of botulinum toxins – already the most lethal poisons known. Transferring genes that help viruses evade the human immune system from one pathogen to another.

These projects may sound like the clandestine activities of a hostile bioweapons programme. But in fact, all are in progress or being planned in US academic labs. They were identified by *New Scientist* in a database that documents research funded by the US Department of Health and Human Services (DHHS). And while each project may sound alarming to the uninitiated, most won the qualified support of the biosecurity specialists we asked to consider their risks and benefits. Only one of those mentioned above – the anthrax project – generated serious objections from some of our experts.

This survey illustrates the difficulties facing the National Science Advisory Board for Biosecurity (NSABB), which the US government has asked to draw up a system for regulating "dual-use" biology – research intended to combat disease, but which could also be misused by bioterrorists or enemy states. The problem is that it is difficult to pursue such work without potentially helping others design bioweapons. "This is the very nature of infectious disease and toxin research," says Michael Stebbins, director of biology policy with the Federation of American Scientists in Washington DC. NSABB will have to tread very carefully if it is to avoid tying up in red tape scientists' ability to respond to emerging diseases.

Dual-use biology hit the headlines in 2001, when *New Scientist* revealed that researchers in Australia had created a strain of mousepox that killed even animals that had been vaccinated (13 January 2001, p 4). The scientists, hoping to control plagues of mice, engineered a mousepox virus ■



# Tools and stories: work with databases on your own computer

## Revealed: Pfizer's payments to censured doctors

› 16:18 22 April 2010 by [Peter Aldhous](#) and [Jim Giles](#)

They are billed as "healthcare professionals who spend years building expertise in their fields". Using materials firmly grounded in science, they educate their peers in the risks and benefits of drugs.

This is how Pfizer, the pharmaceuticals giant, describes the experts it hires to lead [educational forums](#) in which doctors are lectured on the use of its products.

Yet *New Scientist* has found that some of Pfizer's experts have been disciplined for deficiencies in patient care, while others have been reprimanded for how they conducted drug research trials.

The findings add to a growing controversy surrounding the pharmaceutical industry's efforts to market drugs by influencing patterns of prescribing.

### Unknown influence

Doctors paid to educate peers are a particular worry, argues [Sidney Wolfe](#) of consumer advocacy group Public Citizen in Washington DC. "They are doing things that may be influencing your doctor and you have no way of knowing about it," he says. "It's made worse by the fact that some of them have been disciplined."

**Data:** Pfizer's records of payments to doctors, scraped from the web. Data on disciplinary actions from state medical boards in four largest states; FDA warning letters to clinical investigators

**Findings:** Some of Pfizer's "experts" had questionable records for patient safety



**NewScientist**

# Tools and stories: databases

M



Peter Aldhous on Jul 28



18 min

Edit



## Why Are Dope-Addicted, Disgraced Doctors Running Our Drug Trials?

By Peter Aldhous

*Photographs by Grant Cornett*

3

**Data:** US Food and Drug Administration's database of clinical investigators, joined to data on disciplinary actions from state medical boards. (Also FDA's list of disqualified investigators, and its database of inspections of clinical sites.)

**Findings:** Dozens of doctors selected to work on clinical trials over the past five years had previously been censured by state medical boards for problems with patient care; some had their own problems with substance abuse.

# Tools and stories: databases

[Home](#) | [Opinion](#) | [Health](#) | [Science in Society](#) | [News](#)

## My 'non-human' DNA: a cautionary tale

› 15:02 26 August 2009 by [Peter Aldhous](#)  
› For similar stories, visit the [Genetics](#) Topic Guide

"This is a strange question, but are you sure this is *Homo sapiens*?"

It's not every day that an expert queries whether your DNA is human, so when I received this comment by email earlier this month I was somewhat bemused.

I am not in fact the result of a coupling between human and alien, nor the product of some twisted genetic experiment. Instead, [Blaine Bettinger](#), who blogs as [The Genetic Genealogist](#), had been baffled by a DNA profile generated in error by [deCODEme](#), a leading commercial "personal genomics" service provided by Decode Genetics in Reykjavik, Iceland. The false profile seems to be the fault of a software bug.

No harm was done, but the incident serves as a cautionary tale for personalised medicine. As we move towards a future in which readouts from our genomes will routinely be queried by computer systems to help doctors make important clinical decisions, similar glitches could cause prescribing errors – with patients being given drugs at the wrong dose, drugs that won't work, or ones that could even trigger serious side effects in people with a

**Data:** Downloads of my own genetic scans, performed by 23andMe and DeCode Genetics. Corresponding data for my DNA markers read from the same companies' online "genome browsers".

**Findings:** DeCode had a glitch in its database software that could cause the presentation of an erroneous mitochondrial DNA profile in its genome browser.



**NewScientist**

# DeCode's genome browser





# NewScientist

|    | A      | B                                       | C                 | D          | E                    | F                               | G                              | H                      | I              | J                     | K                                | L                               | M                       | N   |
|----|--------|---|-------------------|------------|----------------------|---------------------------------|--------------------------------|------------------------|----------------|-----------------------|----------------------------------|---------------------------------|-------------------------|---|
|    | Record | Position<br>(Human<br>NCBI Build<br>36) | Position<br>(CRS) | 23andMe ID | 23andMe<br>Variation | 23andMe<br>Download<br>Genotype | 23andMe<br>Browser<br>Genotype | 23andMe<br>Consistency | DeCodeMe<br>ID | DeCodeMe<br>Variation | DeCodeMe<br>Download<br>Genotype | DeCodeMe<br>Browser<br>Genotype | DeCodeMe<br>Consistency | Consistency<br>between 23andMe<br>and DeCodeMe<br>Downloads |
| 20 | 19     | 2887                                    | 2885              | rs2854130  | C/T                  | T                               | T                              | Consistent             | MitoT2887C     | C/T                   | T                                | No data                         | Data in down            | Consistent  |
| 21 | 20     | 3012                                    | 3010              | rs3928306  | A/G                  | A                               | A                              | Consistent             | MitoG3012A     | A/G                   | A                                | A                               | Consistent              | Consistent  |
| 22 | 21     | 3198                                    | 3197              | rs2854131  | C/T                  | T                               | T                              | Consistent             | MitoT3198C     | C/T                   | T                                | T                               | Consistent              | Consistent  |
| 23 | 22     | 3349                                    | 3348              | rs41423746 | A/G                  | A                               | A                              | Consistent             | MitoA3349G     | A/G                   | A                                | A                               | Consistent              | Consistent  |
| 24 | 23     | 3395                                    | 3394              | rs41460449 | C/T                  | T                               | T                              | Consistent             | MitoT3395C     | C/T                   | T                                | C                               | Mismatch                | Consistent  |
| 25 | 24     | 3481                                    | 3480              | rs28358584 | A/G                  | A                               | A                              | Consistent             | MitoA3481G     | A/G                   | A                                | G                               | Mismatch                | Consistent  |
| 26 | 25     | 3595                                    | 3594              | rs2854134  | C/T                  | C                               | C                              | Consistent             | MitoC3595T     | C/T                   | C                                | T                               | Mismatch                | Consistent  |
| 27 | 26     | 3667                                    | 3666              | rs28357968 | A/G                  | G                               | G                              | Consistent             | MitoG3667A     | A/G                   | G                                | G                               | Consistent              | Consistent  |
| 28 | 27     | 3721                                    | 3720              | rs41355750 | A/G                  | A                               | A                              | Consistent             | MitoA3721G     | A/G                   | A                                | G                               | Mismatch                | Consistent  |
| 29 | 28     | 3916                                    | 3915              | rs41524046 | A/G                  | G                               | G                              | Consistent             | MitoG3916A     | A/G                   | G                                | G                               | Consistent              | Consistent  |
| 30 | 29     | 3919                                    | 3918              | rs28357972 | A/G                  | G                               | G                              | Consistent             | MitoG3919A     | A/G                   | G                                | G                               | Consistent              | Consistent  |
| 31 | 30     | 3971                                    | 3970              | rs28357973 | C/G/T                | C                               | C                              | Consistent             | MitoC3971T     | C/T                   | C                                | T                               | Mismatch                | Consistent  |
| 32 | 31     | 3993                                    | 3992              | rs41402945 | A/T                  | C                               | C                              | Consistent             | MitoC3993T     | C/T                   | C                                | T                               | Mismatch                | Consistent  |
| 33 | 32     | 4025                                    | 4024              | i1000001   | A/G                  | A                               | A                              | Consistent             | MitoA4025G     | A/G                   | A                                | A                               | Consistent              | Consistent  |
| 34 | 33     | 4337                                    | 4336              | i3001462   | C/T                  | T                               | T                              | Consistent             | MitoT4337C     | C/T                   | T                                | C                               | Mismatch                | Consistent  |
| 35 | 34     | 4562                                    | 4561              | i1000011   | C/T                  | T                               | T                              | Consistent             | MitoT4562C     | C/T                   | T                                | C                               | Mismatch                | Consistent  |
| 36 | 35     | 4770                                    | 4769              | rs3021086  | A/G                  | G                               | G                              | Consistent             | MitoG4770A     | A/G                   | G                                | A                               | Mismatch                | Consistent  |
| 37 | 36     | 4821                                    | 4820              | rs28357977 | A/G                  | G                               | G                              | Consistent             | MitoG4821A     | A/G                   | G                                | G                               | Consistent              | Consistent  |
| 38 | 37     | 4825                                    | 4824              | rs28357978 | A/G                  | A                               | A                              | Consistent             | MitoA4825G     | A/G                   | A                                | No data                         | Data in down            | Consistent  |
| 39 | 38     | 4884                                    | 4883              | rs28357979 | C/T                  | C                               | C                              | Consistent             | MitoC4884T     | C/T                   | C                                | T                               | Mismatch                | Consistent  |
| 40 | 39     | 4918                                    | 4917              | rs28357980 | A/G                  | A                               | A                              | Consistent             | MitoA4918G     | A/G                   | A                                | A                               | Consistent              | Consistent  |
| 41 | 40     | 4978                                    | 4977              | rs28357981 | C/T                  | T                               | T                              | Consistent             | MitoT4978C     | C/T                   | T                                | C                               | Mismatch                | Consistent  |



# Tools and stories: putting data onto maps

NewScientist

[Back to story](#)

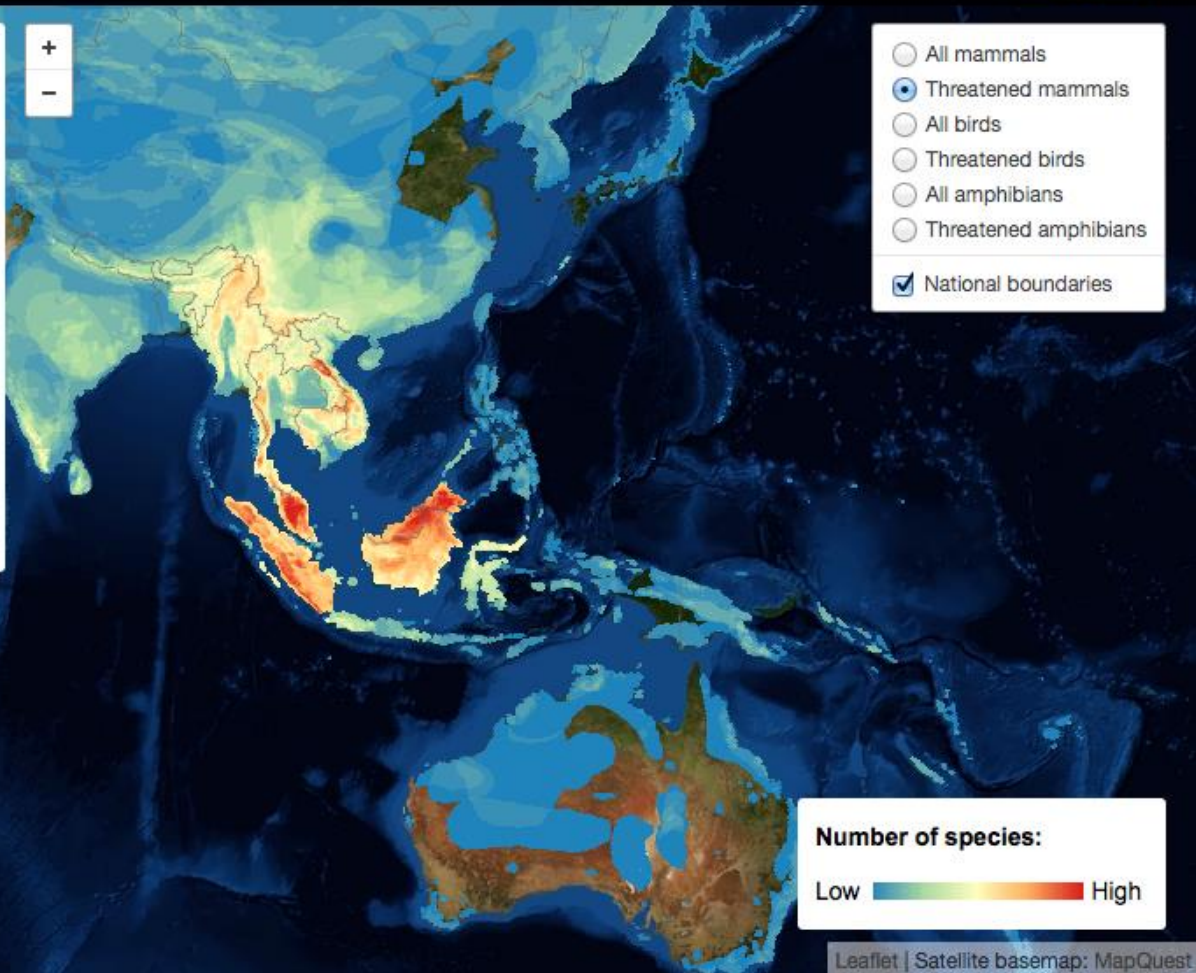


## Where the threatened wild things are

These global biodiversity maps – the most detailed yet – show that hotspots of diversity and threat don't always go hand in hand. Mammals, for instance, are most diverse in the upper Amazon basin and the foothills of the Andes, yet extinction threats are highest in Borneo, Sumatra and the Malay Peninsula – where habitat is being destroyed at an alarming rate.

Use the controls at top right to explore explore the diversity threatened by what some call the Earth's [sixth mass extinction](#). Humanity takes the blame, eliminating species at about 1000 times the natural rate.

Source: [BiodiversityMapping.org/IUCN/BirdLife](http://BiodiversityMapping.org/IUCN/BirdLife)  
International/NatureServe; maps by Peter Aldhous, 29 May 2014



# Tools and stories: putting data onto maps

## 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, landslides and prior studies of the hazards from the 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.

**Statistical analysis, network analysis, etc:**

**Not just for our scientist sources**



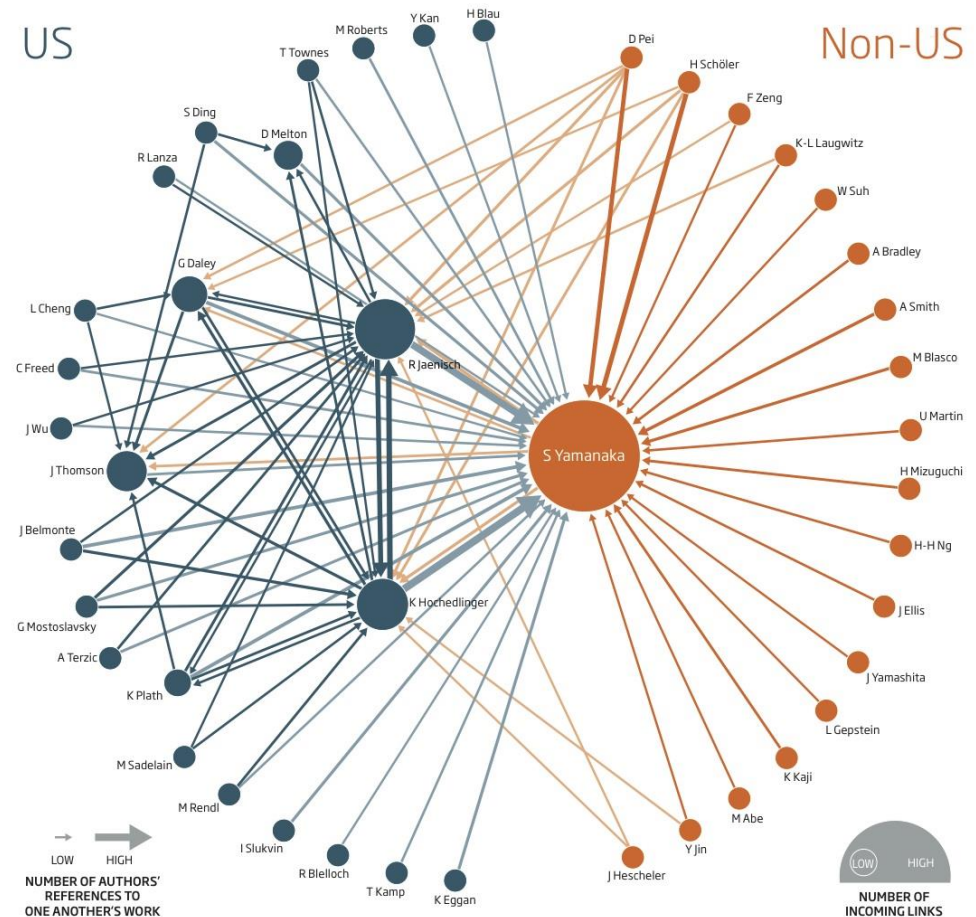


**NewScientist**

# THE STEM CELL WARS

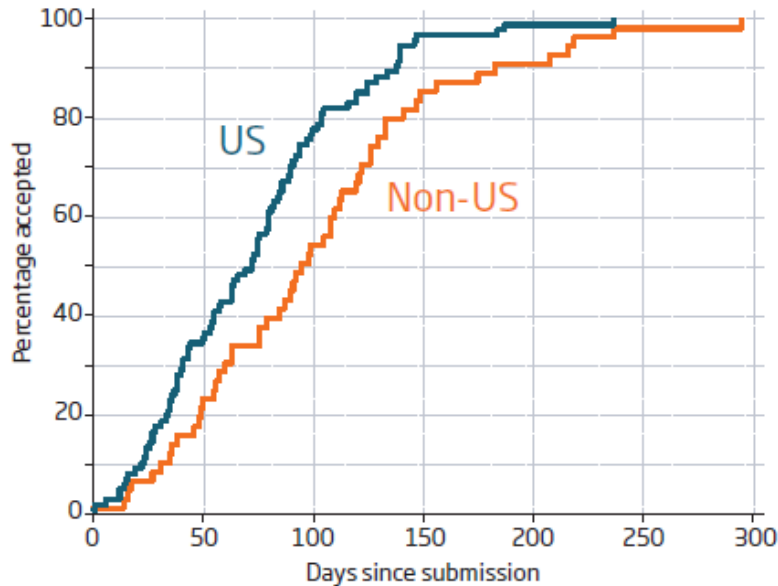
When a Nobel prize is up for grabs, do scientists across the globe compete on a level playing field? **Peter Aldhous** investigates

The most influential players in cellular reprogramming are revealed by recording how many times the 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)



### 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 papers involving “induced pluripotent stem cells” – an exciting alternative to embryonic stem cells which later won their discoverer a Nobel prize.

**Data preparation and analysis:** Searches and downloads from Web of Science database. Annotation of spreadsheet to document time of receipt, acceptance and publication, and location of primary author. Statistical and graphical analysis of time-to-acceptance data. [See methods](#).

**Findings:** Papers from corresponding authors outside the US took significantly longer to be accepted for publication

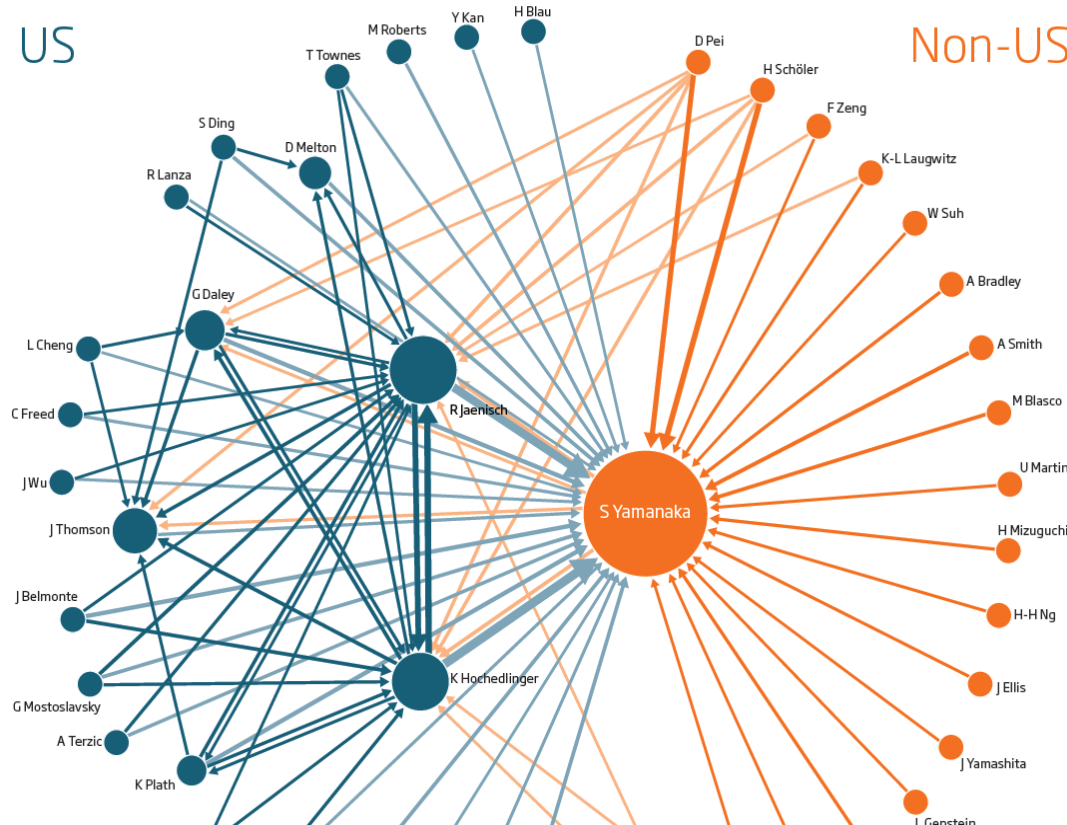


**NewScientist**



US

Non-US



**NewScientist**

**Data:** Citations between primary authors of papers on iPS cells.

**Data preparation and analysis:** Citation information extracted from Web of Science using academic bibliometric software; database queries to link the citations by primary author. Manual checking for errors. Then analysis of network graph. [See methods.](#)

**Findings:** The citation network graph maps influence and connections in the field. Does this help explain why non-US scientists seem to be losing the race to publish?

**Data:** Metadata for 34,000+ papers published in *PNAS* from 2004-2013, plus citation counts, scraped from the journal's website.

**Findings:** Few academy members "contribute" papers at close to the maximum rate, but this group includes several members of the journal's editorial board. Contributed papers are cited less often than those reviewed in the normal way – although the gap has narrowed in recent years.

## The *inside* track

Members of the US National Academy of Sciences have long enjoyed a privileged path to publication in the body's prominent house journal.  
**Meet the scientists who use it most heavily.**

BY PETER ALDHOUS

In April, the US National Academy of Sciences elected 105 new members to its ranks. Academy membership is one the most prestigious honours for a scientist, and it comes with a tangible perk: members can submit up to four papers per year to the body's high-profile journal, the venerable *Proceedings of the National Academy of Sciences* (PNAS), through the 'contributed' publication track. This unusual process allows authors to choose who will review their paper and how to respond to those reviewers' comments.

For many academy members, this privileged path is central to the appeal of PNAS. But to some scientists, it gives the journal the appearance of an old boys' club. "Sound anachronistic? It is," wrote biochemist Steve Caplan of the University of Nebraska, Omaha, in a 2011 blog post that suggested the contributed track could be used as a "dumping ground" for certain

papers. Editors at PNAS have strived to dispel that perception.

With PNAS currently celebrating its centenary, the news team at *Nature* decided to examine the contributed track, both to assess its scientific impact and to see which members use it most heavily and why. After analysing a decade's worth of PNAS papers, we found that only a small number of scientists have used the track at close to the maximum allowable rate. The group includes some of the biggest names in science, and six of them sit on the editorial board at the journal. These scientists say the main motivator for using the contributed track is an intense frustration with the peer-review process at other high-profile journals, which they argue has become excessive and laborious.

Our analysis also suggests that the efforts by PNAS to prevent abuse of the contributed track and to boost the quality of papers published by


this route are bearing fruit. Although contributed PNAS papers attract fewer citations than those handled through the journal's standard review process, the gap has narrowed in recent years. "We have worked really hard at this," says Alan Fersht, a biophysicist at the University of Cambridge, UK, one of PNAS's associate editors and a heavy user of the contributed track.

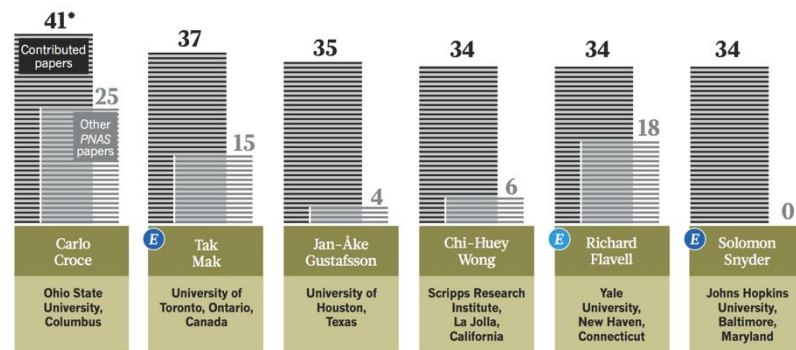
### A PRIVILEGE TO PUBLISH

An inside track to publication for academy members rests deep in PNAS's DNA. The journal was established in 1914 with the explicit goal of publishing members' "more important contributions to research" in addition to "work that appears to a member to be of particular importance". That remit led to the creation of two publishing tracks: contributed and 'communicated' papers -- manuscripts sent by non-members to colleagues in the academy, who shepherded them through review.

### Who are the power users?

Just 13 members of the US National Academy of Sciences consistently published three or more papers per year in the 'contributed' track at PNAS during the past decade. 'Other' papers include direct submissions, reviewed in the normal way, and papers contributed or communicated by other members.

-  Nobel prizewinner
-  Member of PNAS editorial board
-  Former member of editorial board





**Data preparation and analysis:** After the web scraping, extensive data cleaning to remove variants of authors' names, giving one name format for each academy member. Database queries to count papers of different types from each academy member. Statistical analysis to analyse citation rates of different classes of paper. [See methods.](#)

**TABLE 1**

| Period  | Contributed papers |                         |          | Communicated papers |                         |          |
|---------|--------------------|-------------------------|----------|---------------------|-------------------------|----------|
|         | Effect size        | 95% confidence interval | <i>p</i> | Effect size         | 95% confidence interval | <i>p</i> |
| 2004–11 | –4.43%             | –5.39% to –3.47%        | <0.001   | –0.53%              | –1.70% to 0.65%         | 0.37     |
| 2004    | –6.60%             | –9.39% to –3.72%        | <0.001   | –3.45%              | –6.37% to –0.44%        | <0.05    |
| 2005    | –7.39%             | –10.01% to –4.70%       | <0.001   | –1.02%              | –3.94% to 1.99%         | 0.5      |
| 2006    | –7.01%             | –9.63% to –4.32%        | <0.001   | –1.42%              | –4.23% to 1.47%         | 0.33     |
| 2007    | –5.81%             | –8.52% to –3.01%        | <0.001   | –3.67%              | –6.48% to –0.78%        | <0.05    |
| 2008    | –3.41%             | –6.14% to –0.60%        | <0.05    | –0.12%              | –3.25% to 3.11%         | 0.94     |
| 2009    | –1.89%             | –4.65% to 0.94%         | 0.19     | 1.53%               | –1.85% to 5.03%         | 0.38     |
| 2010    | –3.48%             | –6.06% to –0.83%        | <0.05    | –2.50%              | –6.58% to 1.76%         | 0.25     |
| 2011    | –1.14%             | –3.70% to 1.48%         | 0.39     | N/A                 | N/A                     | N/A      |

# Beware running with scissors: Seek expert help if you need rigorous statistical analysis!

## DIY statistical analysis: experience the thrill of touching real data

The story of one man's efforts to re-analyse the stats behind a BBC report on bowel cancer is a heartwarmingly nerdy one



**Ben Goldacre**

[guardian.co.uk](http://guardian.co.uk), Friday 28 October 2011 17.31 EDT

Comments (60)

Share 209

Tweet 306

+1 18

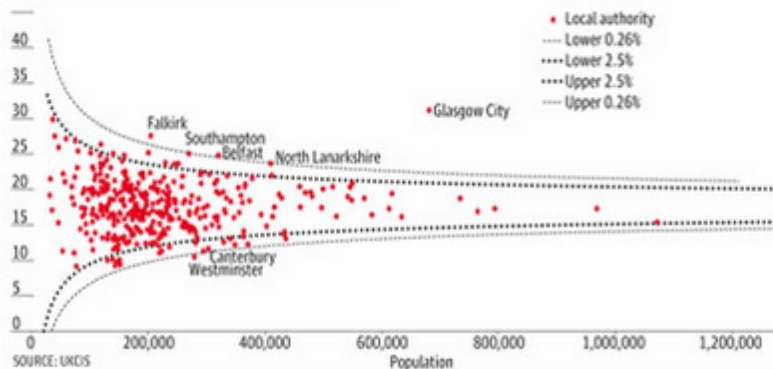
Email



Article history

### Bowel cancer mortality

By UK local authority, deaths per 100,000



A funnel plot of bowel cancer mortality rates in different areas of the UK

The BBC has found a story: "**Threefold variation' in UK bowel cancer rates**". The average death rate across the UK from bowel cancer is 17.9 per 100,000 people, but in some places it's as low as 9, and in some places it's as high as 30. What can be causing this?

Journalists tend to find imaginary patterns in statistical noise, which we've covered many times before. But this case is particularly silly, as you will see, and it has a heartwarming, nerdy twist.

Society

Cancer · Bowel cancer

Science

Cancer

Media

BBC

Series

Bad science

More from **Comment is free on**

Society

Cancer · Bowel cancer

Science

Cancer

Media

BBC

Series

Bad science

Related

18 Apr 2007

Aspirin linked to lower risk of bowel and prostate cancer



# Even Nate Silver can get his fingers burned ...

## Tracker

Peer review within science journalism



suggest a story



journalism resources



subscribe

18 MAR 2014

Nate Silver's new FiveThirtyEight dishes out statistical nonsense on health coverage.

3 Comments

Author: Paul Raeburn

Tags:  
538, fivethirtyeight,  
fivethirtyeight.com, Nate  
Silver, presidential  
election

Share



Nate Silver's **fivethirtyeight.com** relaunched yesterday at its new home--**ESPN**--vowing to focus its coverage on five areas: politics, economics, life, sports--and science.

The inclusion of science was a surprise to me. And possibly a mistake, unless FiveThirtyEight can quickly improve the quality of the "science" it's publishing. The lead story on the relaunched site's first day--"Finally, a Formula for Decoding Health News"--was abysmal.

FOLLOW US

Follow @KSJTracker on  
Twitter!

CATEGORIES

View All

About Journalism

Environment & Energy  
Stories

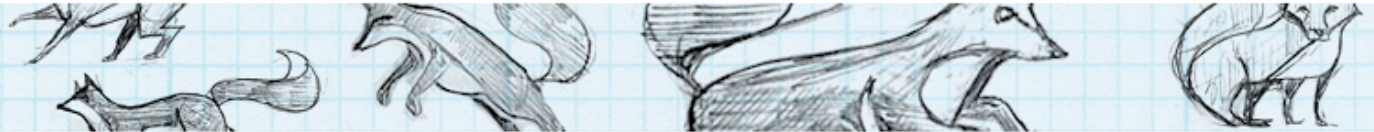
German Language Media

Health & Medicine Stories

Rastreador Científico en  
Español



# ... and be forced to backtrack



■ A NOTE TO OUR READERS

## FiveThirtyEight to Commission Response to Disputed Climate Article

11:50 AM | MAR 28 | By NATE SILVER

FiveThirtyEight relaunched less than two weeks ago. It's been a heck of a learning experience. When you're trying something new, [it's going to take some time](#) to get everything right, and you're [going to get criticism from all quarters](#).

### FiveThirtyEight staff

Tweets from a list by FiveThirtyEight

Follow our writers and editors on Twitter.



**Marie P Donoghue**

@mariepdonoghue

1h

Love this story of female mentoring -- A Woman's Place Is Running the Kitchen [nyti.ms/1ffNmSB](https://nyti.ms/1ffNmSB)

Show Summary



**Micah Cohen**

@micahcohen

3h

.@nationaljournal that last paragraph is spot on, and neatly summarizes the @FiveThirtyEight view too

Expand



**Jeanne Whalen**

@JeanneWhalen

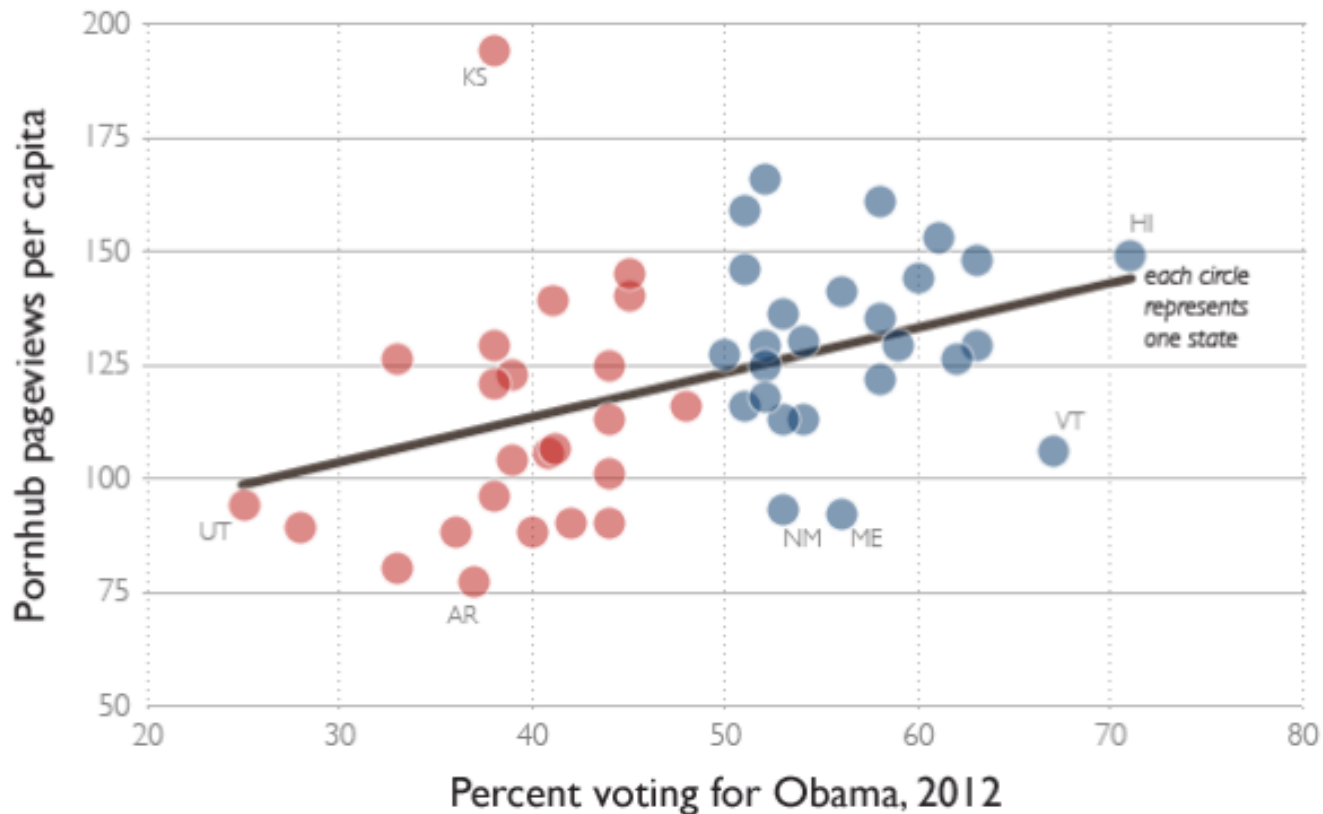
7h

Some employers now see autism as an asset, not a liability. By @ShirleySWangWSJ [on.wsj.com/1peuav3](https://on.wsj.com/1peuav3)

# DISTRUST YOUR DATA

Jacob Harris on Six Ways to Make Mistakes with Data

Presidential politics and porn per capita



SOURCE: Pornhub Insights, Wikipedia

WONKIVIZ.TUMBLR.COM

(Chris Ingraham)

[Read](#) this before committing an act of data journalism!

**Why aren't we seeing  
more data-driven  
science journalism?**

# **Some basic data resources for science reporters**

# Using web search forms

- Look for the advanced search page, which will offer options to customize your search.
- Read the Help or FAQs to learn how the search works. Does it use Boolean logic (AND, OR, NOT)? Do quote marks allow you to search for a specific phrase? Is there a wildcard character, such as \* or %, that allows you to look for variations on a search term?
- Look for download options once you've found the data you need:

The screenshot shows a web interface for clinical trial search results. At the top, there are tabs: 'List' (selected), 'By Topic', 'On a Map', and 'Search Details'. Below the tabs is a '+ Show Display Options' link and 'Download' and 'Subscribe to RSS' icons. A table of search results is visible, with columns for Rank, Status, and Study. The table lists four studies. A modal window is open over the table, titled 'Download the search results for: multiple sclerosis | stem cell (25 records)'. It includes a 'Need help? See Downloading Content for Analysis' link. The modal has a 'Number of Studies' dropdown set to '25 Found Studies'. Under 'Download Content', there are three radio button options: 'Download All Study Fields as XML', 'Download All Study and Results Fields as XML', and 'Download Selected Fields:' (which is selected). Below these are 'Select fields:' and 'Select field format:' dropdowns, both set to '20 Available Fields' and 'Comma-separated Values' respectively. At the bottom of the modal are 'Download Zip File' and 'Cancel' buttons. A link for 'Zip file readers with free trial periods: WinZip or PKZip' is also present.

| Rank | Status     | Study   |
|------|------------|---|
| 1    | Recruiting | <a href="#">Stem Cell Therapy for Patients With I</a><br>Condition: Multiple S<br>Interventions: Procedure<br>Drug: Star  |
| 2    | Recruiting | <a href="#">Evaluation of Autologous Mesenchy</a><br>Condition: Multiple S<br>Interventions: Biological:<br>Biological:   |
| 3    | Suspended  | <a href="#">Allogeneic Stem Cell Transplantation</a><br>Condition: Multiple S<br>Intervention: Biological:  |
| 4    | Recruiting | <a href="#">Autologous Mesenchymal Stem Cell</a><br>Conditions: Relapsing<br>Progressive Relapsing Multiple Sclerosis<br>Intervention: Biological: Autologous mesenchymal stem cell transplantation |



# Scientific literature

## PubMed

<http://www.pubmed.gov>

<http://www-ncbi-nlm-nih-gov.oca.ucsc.edu/pubmed>

## Web of Science

<http://bit.ly/YmJtGM>

## Google Scholar

<http://scholar.google.com/>

<http://scholar.google.com.oca.ucsc.edu/>

## Agricola

<http://agricola.nal.usda.gov/>

<http://agricola.nal.usda.gov.oca.ucsc.edu/>

## PsycINFO

<http://www.apa.org/pubs/databases/psycinfo/index.aspx>

<http://search.proquest.com.oca.ucsc.edu/psycinfo/advanced>

# Patents

## **European Patent Office**

[http://worldwide.espacenet.com/advancedSearch?locale=en\\_EP](http://worldwide.espacenet.com/advancedSearch?locale=en_EP)

Search European patent applications, WIPO patent applications, or all patents across 90+ nations

## **US Patent and Trademark Office**

<http://patft.uspto.gov/>

Search issued patents and published applications

For more information about a patent (diagrams, history of correspondence with patent office etc), enter patent or application number here:

<http://portal.uspto.gov/external/portal/pair>

## **Google Advanced Patent Search**

[http://www.google.com/advanced\\_patent\\_search](http://www.google.com/advanced_patent_search)

Clear presentation of claims, abstract, diagrams etc.

# Grant funding

## **National Institutes of Health RePORTER**

<http://projectreporter.nih.gov/reporter.cfm>

Grants from NIH, with links to papers and patents arising from projects

## **National Science Foundation**

<http://www.nsf.gov/awardsearch/advancedSearch.jsp>

## **Research.gov**

[http://www.research.gov/research-portal/appmanager/base/desktop?\\_nfpb=true&\\_eventName=viewQuickSearchFormEvent\\_so\\_rsr](http://www.research.gov/research-portal/appmanager/base/desktop?_nfpb=true&_eventName=viewQuickSearchFormEvent_so_rsr)

(Also includes NASA grants)

# Clinical trials

## **ClinicalTrials.gov**

<http://www.clinicaltrials.gov>

Information on more than 187,000 clinical trials in US and beyond

## **International Standard Randomised Controlled Trial Number Register**

<http://www.isrctn.com/>

Information on almost 13,000 registered clinical trials

## **WHO International Clinical Trials Registry Platform**

<http://apps.who.int/trialsearch/default.aspx>

# **Investigating science**

**(with help from data)**

**Peter Aldhous,  
Science reporter, BuzzFeed News**

[peter@peteraldhous.com](mailto:peter@peteraldhous.com)  
[@paldhous](#)