Data journalism: what it can do for you

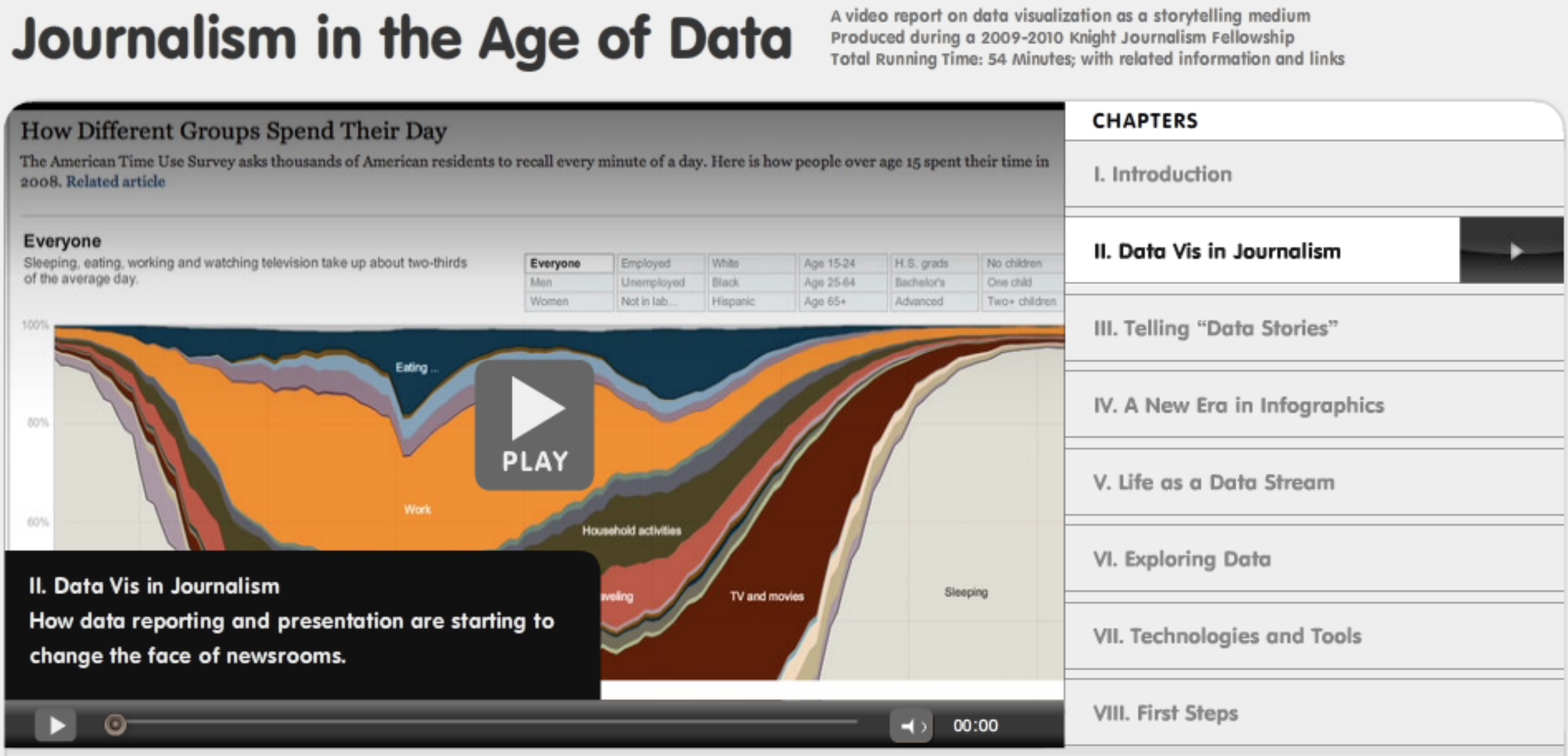
NCSWA workshop, January 12, 2013

Peter Aldhous,
San Francisco Bureau Chief

NewScientist

peter@peteraldhous.com
Twitter: @paldhous

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.

II. Data Vis in Journalism

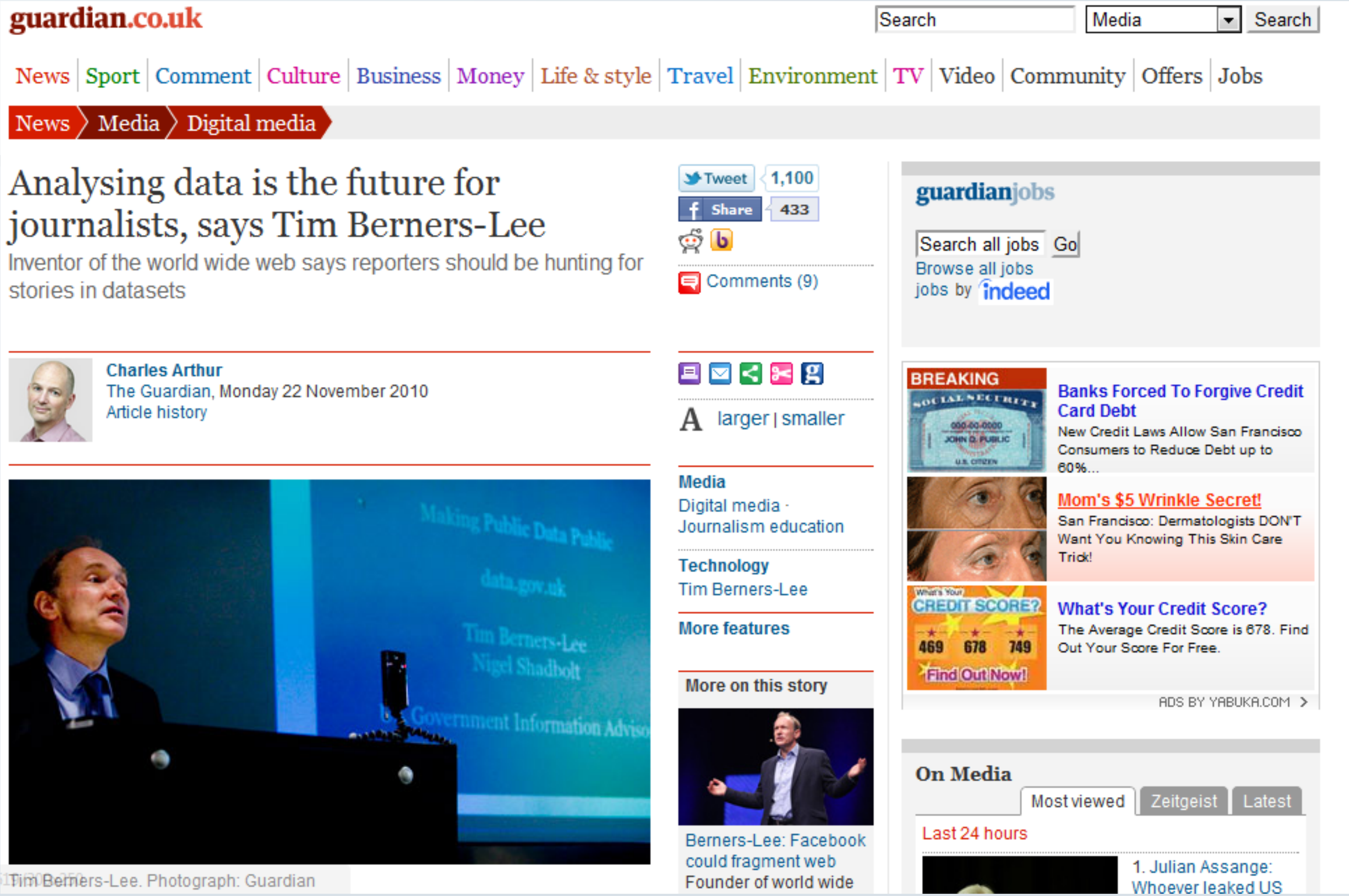
How data reporting and presentation are starting to change the face of newsrooms.

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

Tweet 1,100
Share 433

Comments (9)

Charles Arthur
The Guardian, Monday 22 November 2010
Article history

A larger | smaller

Tim Berners-Lee. Photograph: Guardian

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

On Media

Most viewed | Zeitgeist | Latest

Last 24 hours

1. Julian Assange: Whoever leaked US

What’s in it for me?

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

Note: data can be used in both reporting and storytelling. But think carefully about what you need to show to your audience. Some of the best data-driven stories actually contain little in the way of numbers or graphs

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 readers 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.

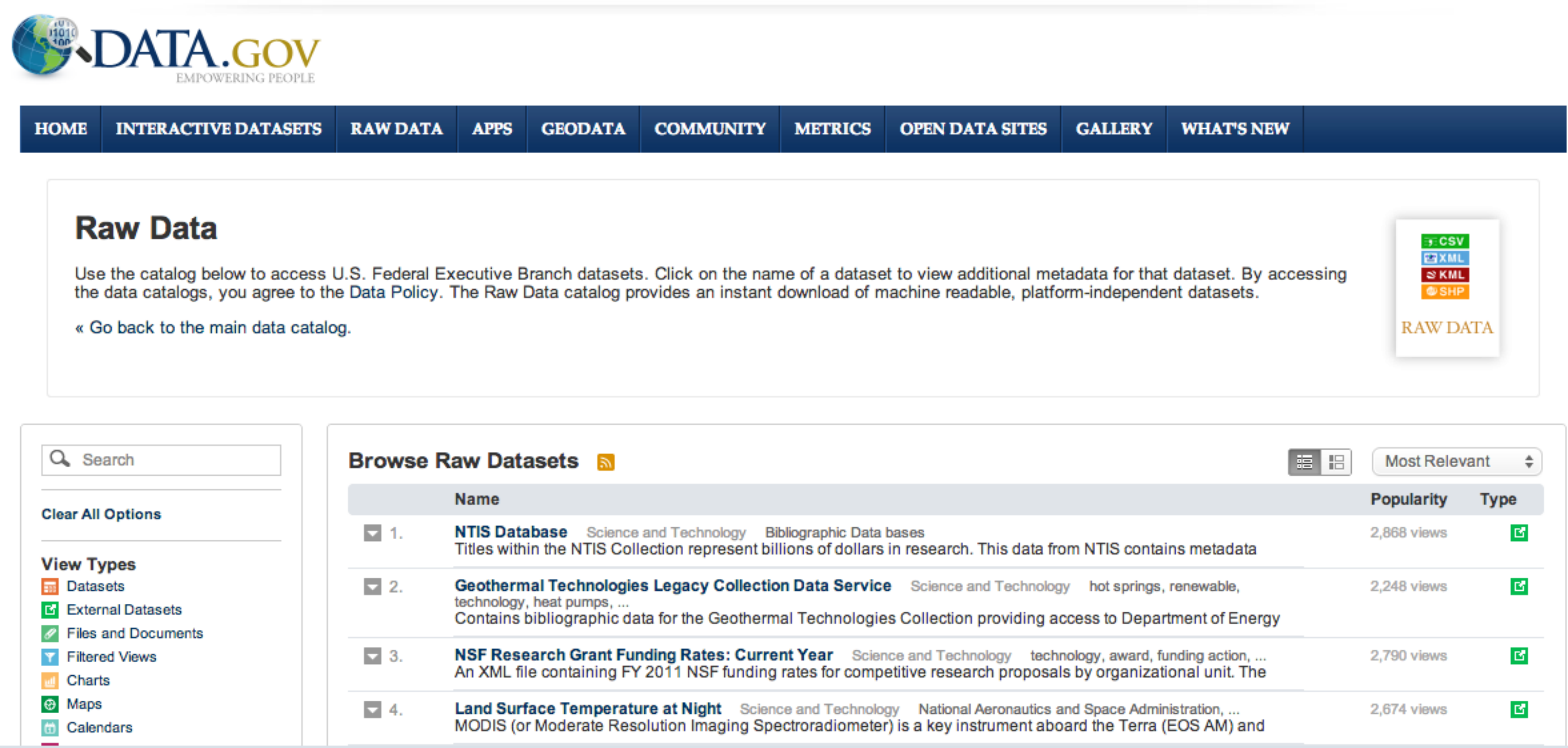
- If necessary, have a plan for acquiring data at regular intervals. Some data may require public records requests

Note: this is very different to: “I’ve written my story. Now I’d better find some numbers for a graph.”

Where do I find data?

Some good portals

Data.gov: a work in progress



DATA.GOV
EMPOWERING PEOPLE

Raw Data

Use the catalog below to access U.S. Federal Executive Branch datasets. Click on the name of a dataset to view additional metadata for that dataset. By accessing the data catalogs, you agree to the Data Policy. The Raw Data catalog provides an instant download of machine readable, platform-independent datasets.

« Go back to the main data catalog.

View Types
- Datasets
- External Datasets
- Files and Documents
- Filtered Views
- Charts
- Maps
- Calendars

Browse Raw Datasets

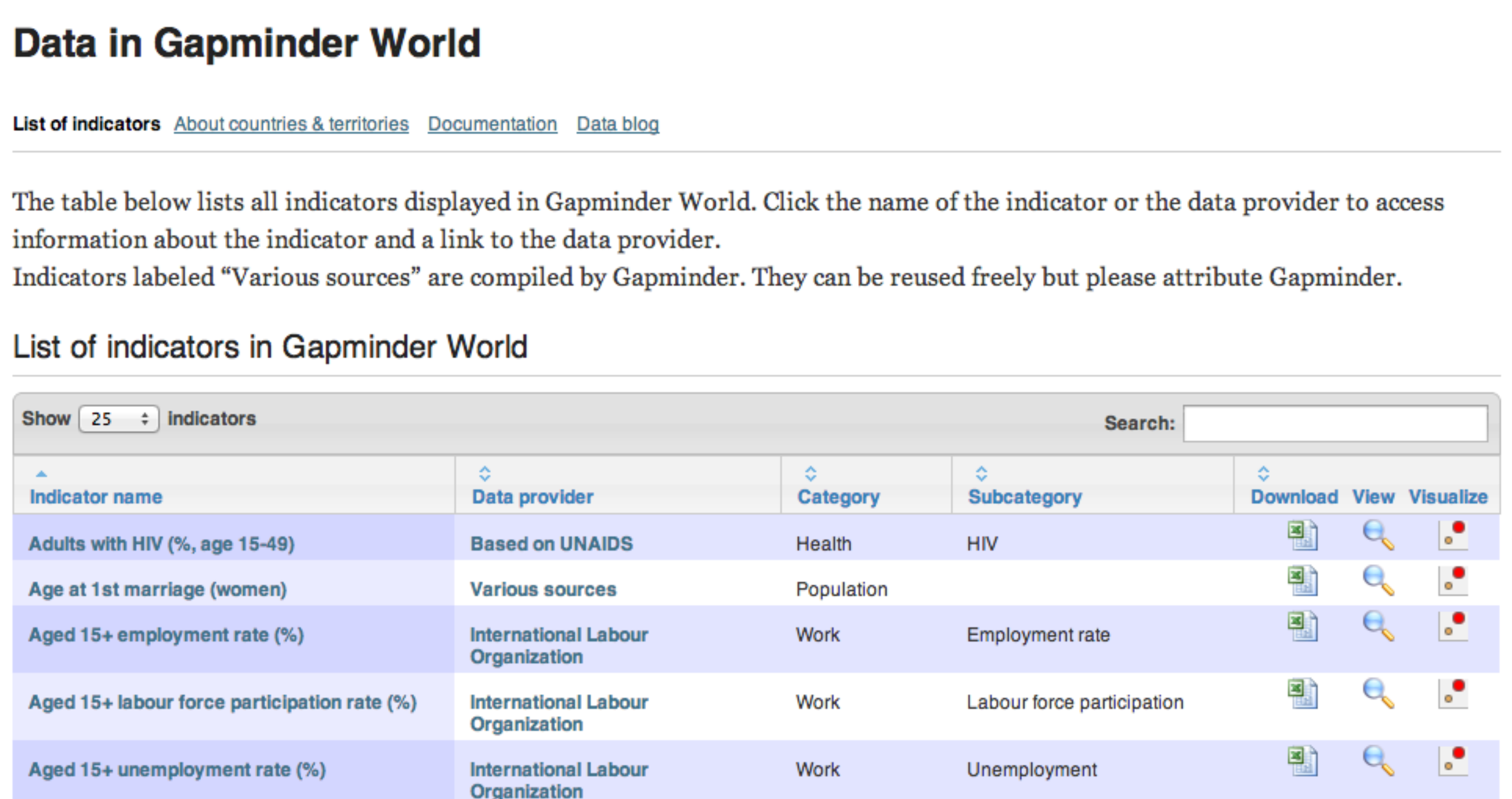
	Name	Popularity	Type
1.	**NTIS Database** Science and Technology Bibliographic Data bases Titles within the NTIS Collection represent billions of dollars in research. This data from NTIS contains metadata	2,868 views	↗
2.	**Geothermal Technologies Legacy Collection Data Service** Science and Technology hot springs, renewable, technology, heat pumps, ... Contains bibliographic data for the Geothermal Technologies Collection providing access to Department of Energy	2,248 views	↗
3.	**NSF Research Grant Funding Rates: Current Year** Science and Technology technology, award, funding action, ... An XML file containing FY 2011 NSF funding rates for competitive research proposals by organizational unit. The	2,790 views	↗
4.	**Land Surface Temperature at Night** Science and Technology National Aeronautics and Space Administration, ... MODIS (or Moderate Resolution Imaging Spectroradiometer) is a key instrument aboard the Terra (EOS AM) and	2,674 views	↗

(For the time being, FedStats is still a better portal to US government data.)

Where do I find data?

More portals

For international comparisons, try the World Bank or Gapminder:



Data in Gapminder World

List of indicators About countries & territories Documentation Data blog

The table below lists all indicators displayed in Gapminder World. Click the name of the indicator or the data provider to access information about the indicator and a link to the data provider.
Indicators labeled “Various sources” are compiled by Gapminder. They can be reused freely but please attribute Gapminder.

List of indicators in Gapminder World

Show 25 indicators Search:

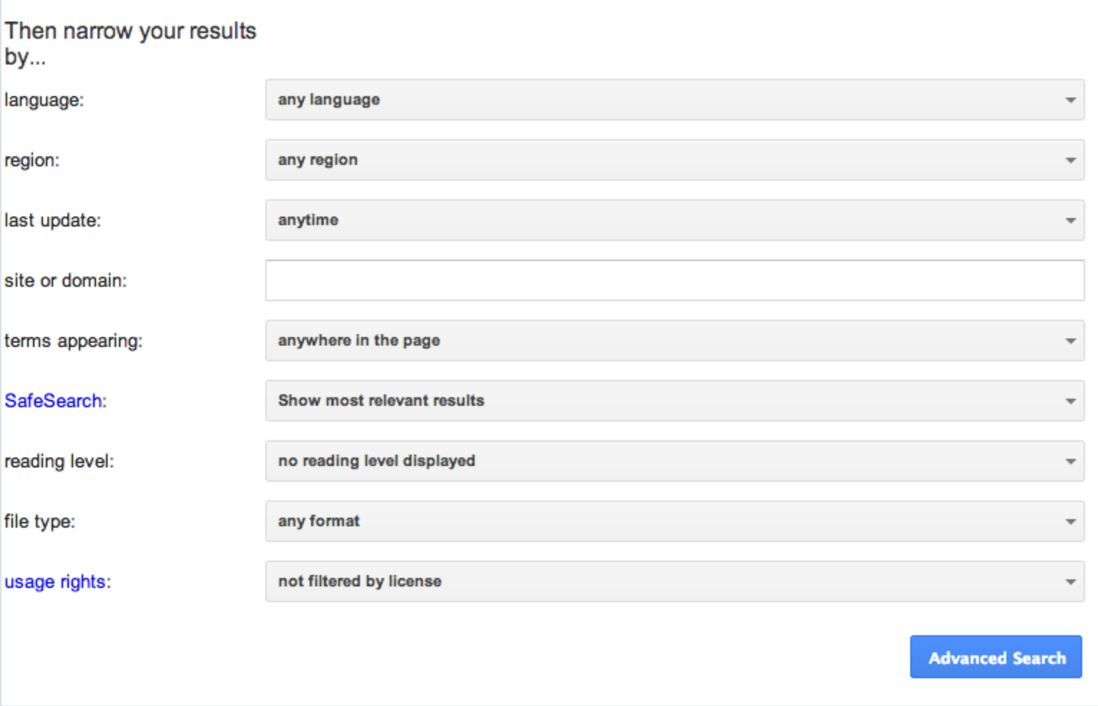
Indicator name	Data provider	Category	Subcategory	Download	View	Visualize
Adults with HIV (%, age 15-49)	Based on UNAIDS	Health	HIV	▦	🔍	●
Age at 1st marriage (women)	Various sources	Population		▦	🔍	●
Aged 15+ employment rate (%)	International Labour Organization	Work	Employment rate	▦	🔍	●
Aged 15+ labour force participation rate (%)	International Labour Organization	Work	Labour force participation	▦	🔍	●
Aged 15+ unemployment rate (%)	International Labour Organization	Work	Unemployment	▦	🔍	●

Where do I find data?

Often, you'll need to search for it

- Google is your friend. Sometimes simply combining a few keywords with "data" or "database" is enough to find what you need

- Use Google's advanced search options:



Then narrow your results by...

language:	any language
region:	any region
last update:	anytime
site or domain:	
terms appearing:	anywhere in the page
SafeSearch:	Show most relevant results
reading level:	no reading level displayed
file type:	any format
usage rights:	not filtered by license

e.g. the National Oceanic and Atmospheric Administration is a good source of data on weather and climate, so if searching for data on hurricanes, try narrowing the search to the noaa.gov **site or domain**

You can also search by **file type**, e.g. xls for Excel spreadsheets

Where do I find data?

Some sample sources for science reporters

Research grants: National Institutes of Health; National Science Foundation

Clinical trials: ClinicalTrials.gov

Earthquakes: USGS earthquake search

Extreme weather: NWS tornadoes, hail and damaging wind (scroll down for data files)

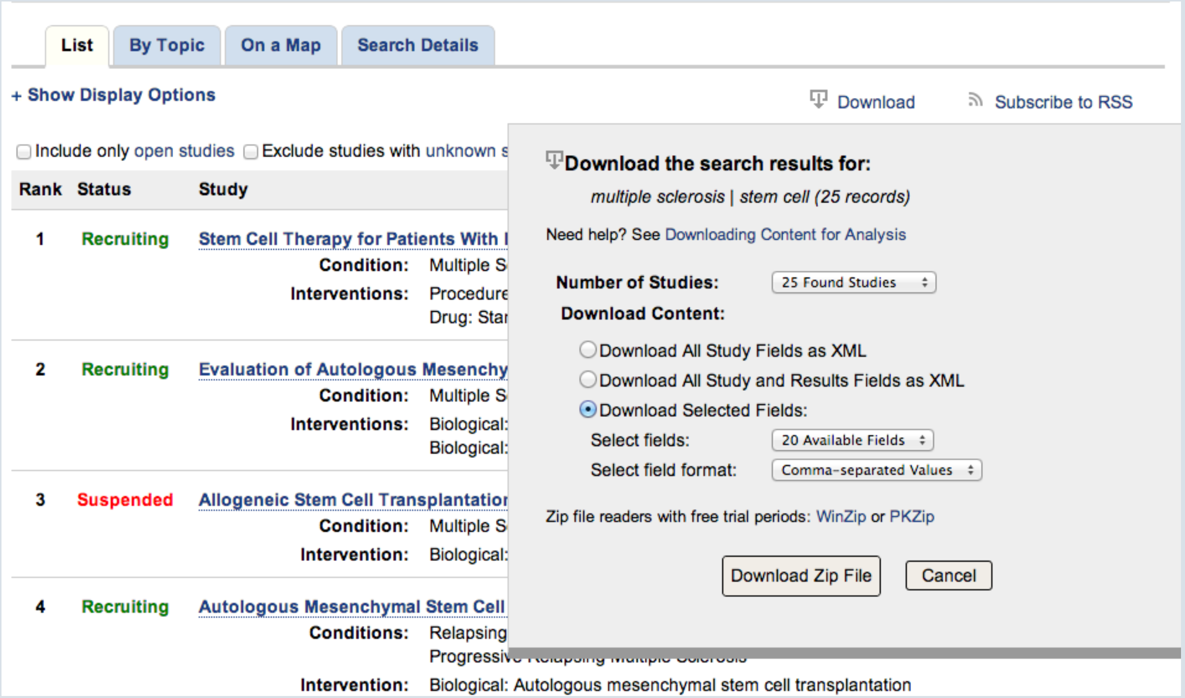
Public health/epidemiology: CDC Wonder

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:



☐ Include only open studies ☐ Exclude studies with unknown s

Rank	Status	Study
1	Recruiting	Stem Cell Therapy for Patients With I Condition: Multiple S Interventions: Procedure Drug: Star
2	Recruiting	Evaluation of Autologous Mesenchy Condition: Multiple S Interventions: Biological: Biological:
3	Suspended	Allogeneic Stem Cell Transplantatior Condition: Multiple S Intervention: Biological:
4	Recruiting	Autologous Mesenchymal Stem Cell Conditions: Relapsing Progressiv Intervention: Biological: Autologous mesenchymal stem cell transplantation

Download the search results for:

multiple sclerosis | stem cell (25 records)

Need help? See Downloading Content for Analysis

Number of Studies: 25 Found Studies

Download Content:

○ Download All Study Fields as XML
○ Download All Study and Results Fields as XML
● Download Selected Fields:

Select fields: 20 Available Fields
Select field format: Comma-separated Values

Zip file readers with free trial periods: WinZip or PKZip

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 these operations as “interviewing” the data.)

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 to scan for potential errors.

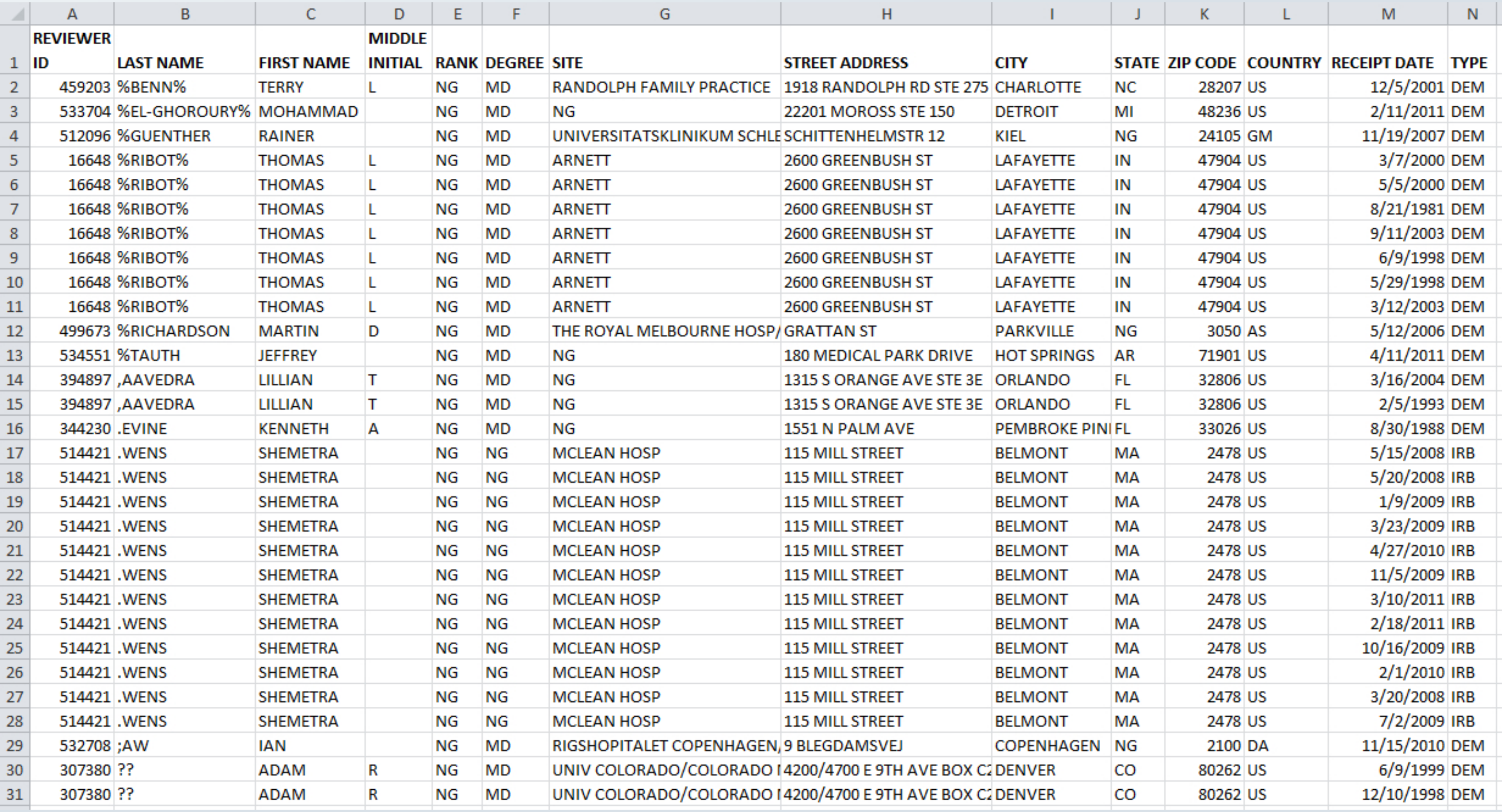
You will almost always need to reformat and edit data to suit your purposes; frequently you will have to do extensive data “cleaning.”

Simple reformatting and editing can be done using a spreadsheet, but for bigger cleaning tasks, use:

Google Refine/OpenRefine

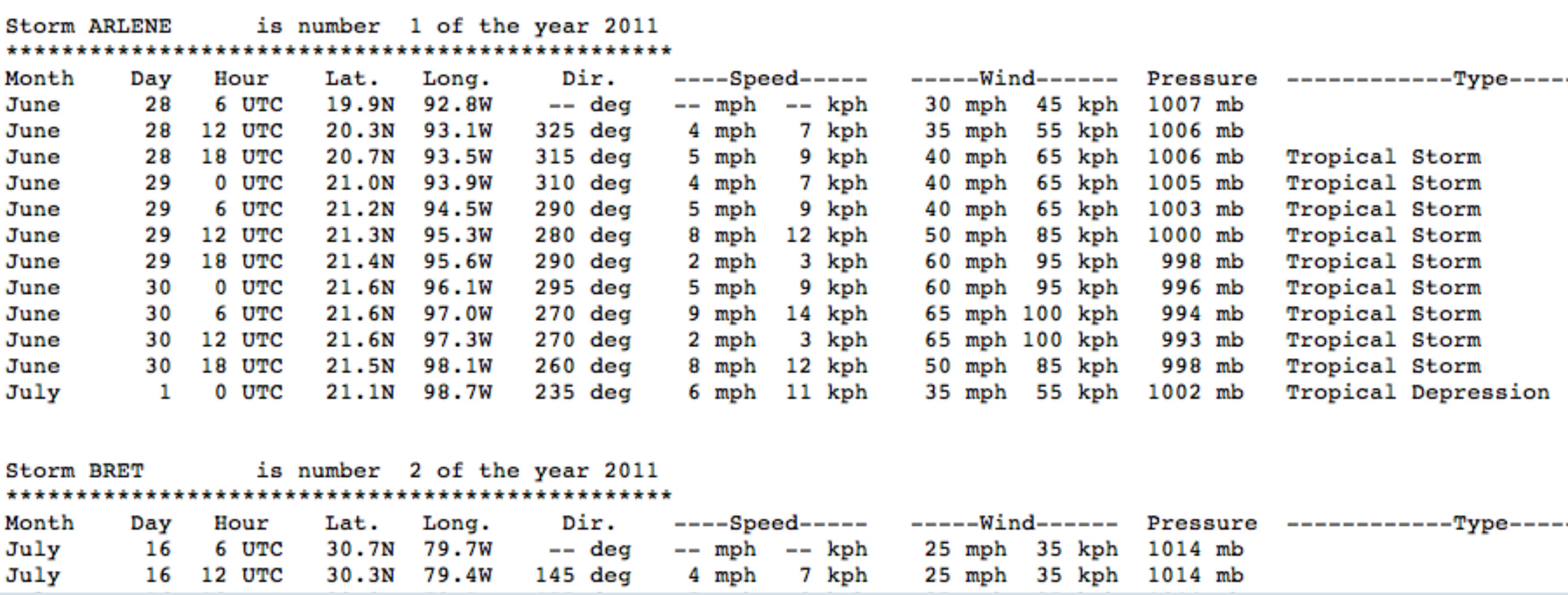
There are good video tutorials for this tool at the Google link above.

Please clean me!



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	REVIEWER ID	LAST NAME	FIRST NAME	MIDDLE 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 I	4200/4700 E 9TH AVE BOX C2	DENVER	CO	80262	US	6/9/1999	DEM
31	307380	??	ADAM	R	NG	MD	UNIV COLORADO/COLORADO I	4200/4700 E 9TH AVE BOX C2	DENVER	CO	80262	US	12/10/1998	DEM

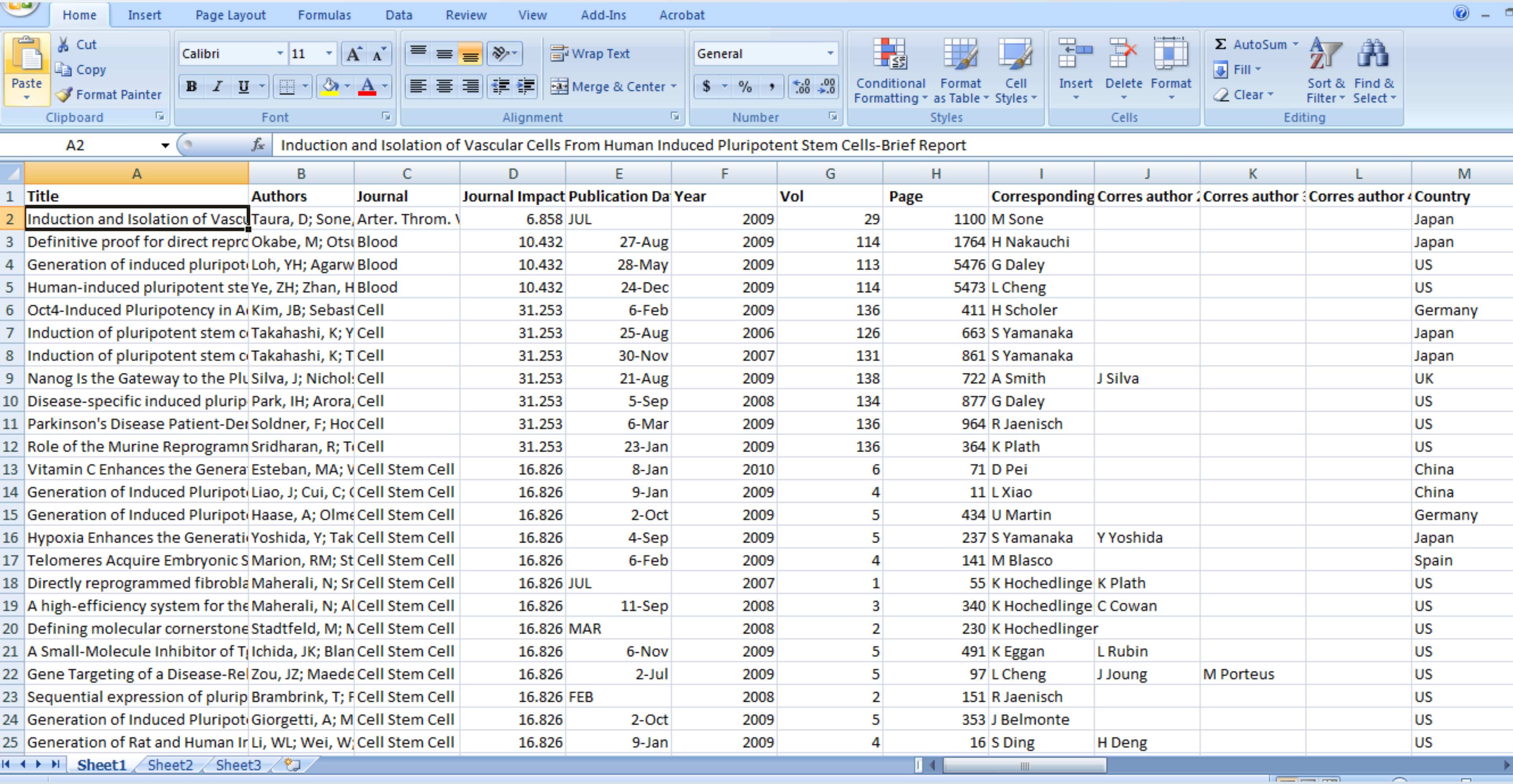
Why science journalists are lucky: clean, well curated data



```
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
```

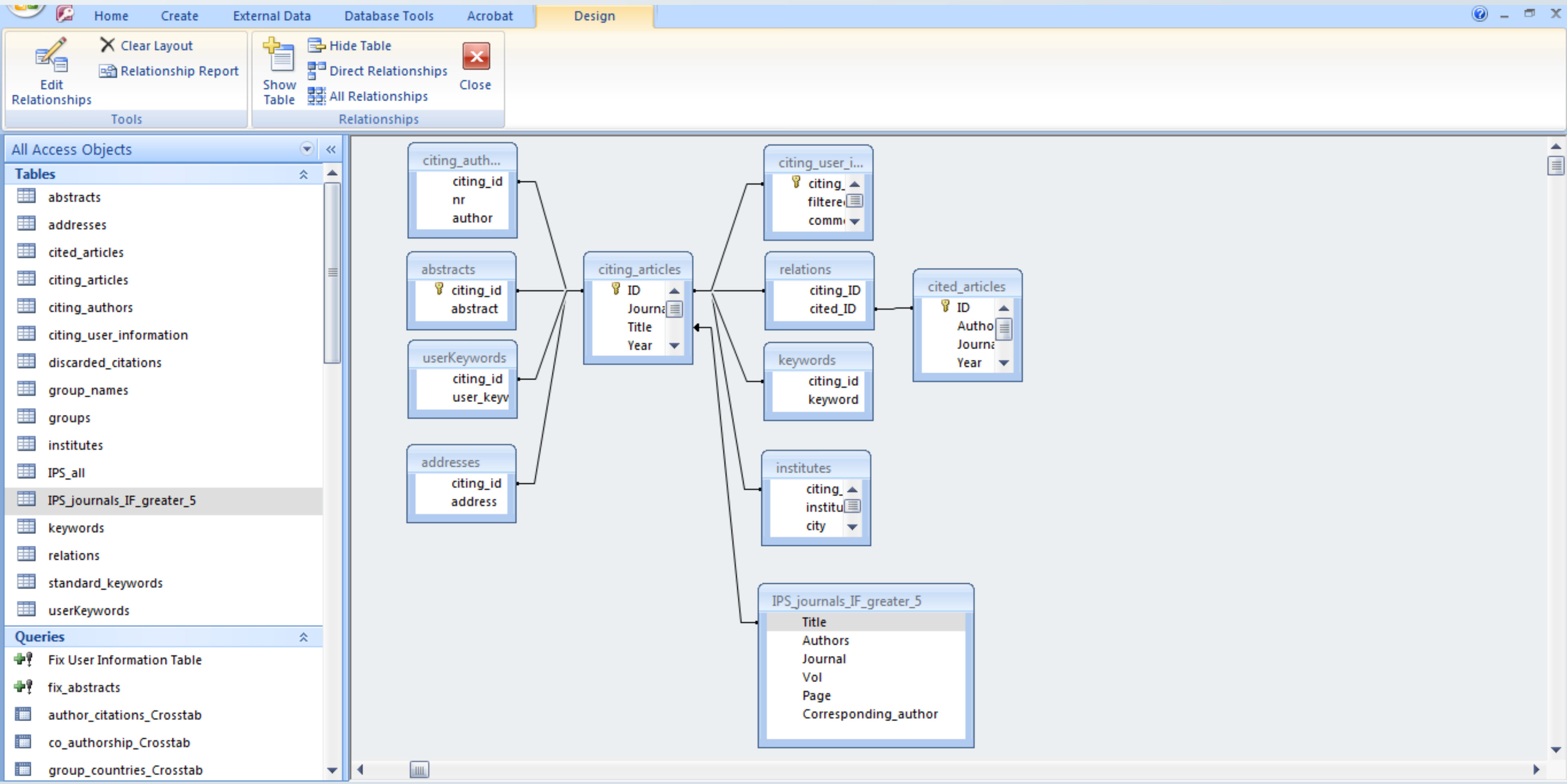
The basic tools: spreadsheets …



A2 | Induction and Isolation of Vascular Cells From Human Induced Pluripotent Stem Cells-Brief Report

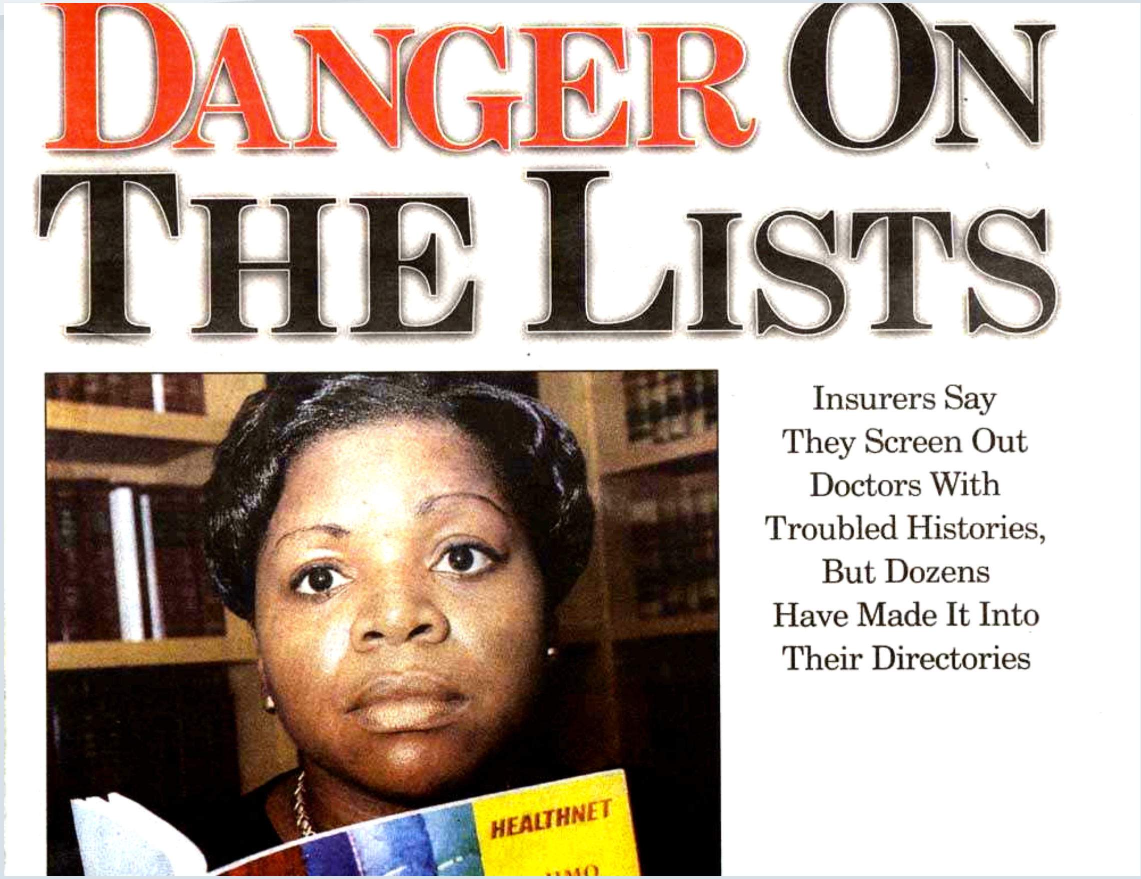
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Title	Authors	Journal	Journal Impact	Publication Da	Year	Vol	Page	Corresponding	Corres author 2	Corres author 3	Corres author 4	Country
2	Induction and Isolation of Vascu	Taura, D; Sone,	Arter. Throm. V	6.858	JUL	2009	29	1100	M Sone				Japan
3	Definitive proof for direct repro	Okabe, M; Otsu	Blood	10.432	27-Aug	2009	114	1764	H Nakauchi				Japan
4	Generation of induced pluripot	Loh, YH; Agarw	Blood	10.432	28-May	2009	113	5476	G Daley				US
5	Human-induced pluripotent ste	Ye, ZH; Zhan, H	Blood	10.432	24-Dec	2009	114	5473	L Cheng				US
6	Oct4-Induced Pluripotency in A	Kim, JB; Sebast	Cell	31.253	6-Feb	2009	136	411	H Scholer				Germany
7	Induction of pluripotent stem c	Takahashi, K; Y	Cell	31.253	25-Aug	2006	126	663	S Yamanaka				Japan
8	Induction of pluripotent stem c	Takahashi, K; T	Cell	31.253	30-Nov	2007	131	861	S Yamanaka				Japan
9	Nanog Is the Gateway to the Plu	Silva, J; Nichol	Cell	31.253	21-Aug	2009	138	722	A Smith	J Silva			UK
10	Disease-specific induced plurip	Park, IH; Arora,	Cell	31.253	5-Sep	2008	134	877	G Daley				US
11	Parkinson's Disease Patient-Der	Soldner, F; Hoc	Cell	31.253	6-Mar	2009	136	964	R Jaenisch				US
12	Role of the Murine Reprogramn	Sridharan, R; T	Cell	31.253	23-Jan	2009	136	364	K Plath				US
13	Vitamin C Enhances the Genera	Esteban, MA; V	Cell Stem Cell	16.826	8-Jan	2010	6	71	D Pei				China
14	Generation of Induced Pluripot	Liao, J; Cui, C; (	Cell Stem Cell	16.826	9-Jan	2009	4	11	L Xiao				China
15	Generation of Induced Pluripot	Haase, A; Olme	Cell Stem Cell	16.826	2-Oct	2009	5	434	U Martin				Germany
16	Hypoxia Enhances the Generati	Yoshida, Y; Tak	Cell Stem Cell	16.826	4-Sep	2009	5	237	S Yamanaka	Y Yoshida			Japan
17	Telomeres Acquire Embryonic S	Marion, RM; St	Cell Stem Cell	16.826	6-Feb	2009	4	141	M Blasco				Spain
18	Directly reprogrammed fibrobla	Maherali, N; Sr	Cell Stem Cell	16.826	JUL	2007	1	55	K Hochedlinge	K Plath			US
19	A high-efficiency system for the	Maherali, N; Al	Cell Stem Cell	16.826	11-Sep	2008	3	340	K Hochedlinge	C Cowan			US
20	Defining molecular cornerstone	Stadtfeld, M; N	Cell Stem Cell	16.826	MAR	2008	2	230	K Hochedlinger				US
21	A Small-Molecule Inhibitor of T	Ichida, JK; Blan	Cell Stem Cell	16.826	6-Nov	2009	5	491	K Eggan	L Rubin			US
22	Gene Targeting of a Disease-Rel	Zou, JZ; Maede	Cell Stem Cell	16.826	2-Jul	2009	5	97	L Cheng	J Joung	M Porteus		US
23	Sequential expression of plurip	Brambrink, T; F	Cell Stem Cell	16.826	FEB	2008	2	151	R Jaenisch				US
24	Generation of Induced Pluripot	Giorgetti, A; M	Cell Stem Cell	16.826	2-Oct	2009	5	353	J Belmonte				US
25	Generation of Rat and Human Ir	Li, WL; Wei, W;	Cell Stem Cell	16.826	9-Jan	2009	4	16	S Ding	H Deng			US

… and database managers



Tools and stories: databases

Newsday



DANGER ON THE LISTS

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

HEALTHNET
HMO

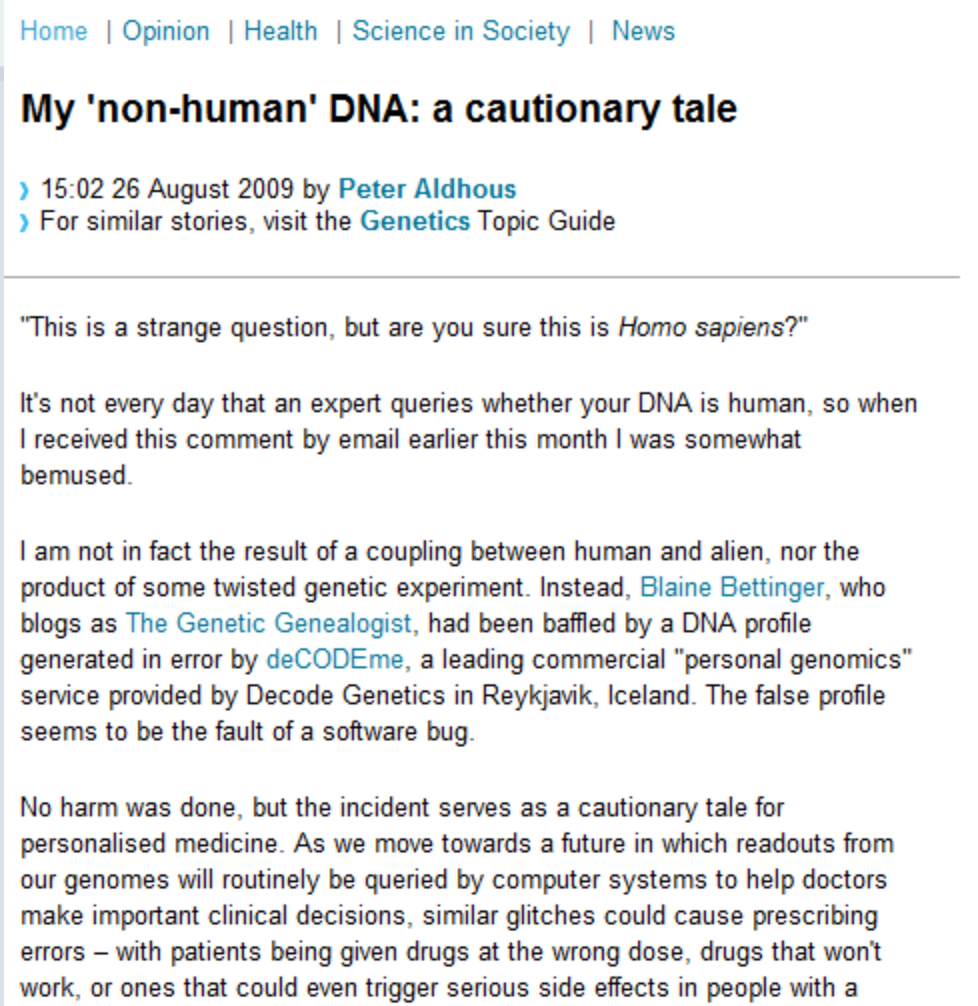
Data: HMO doctor directories and state records of disciplinary actions taken against doctors.

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: databases

NewScientist



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.

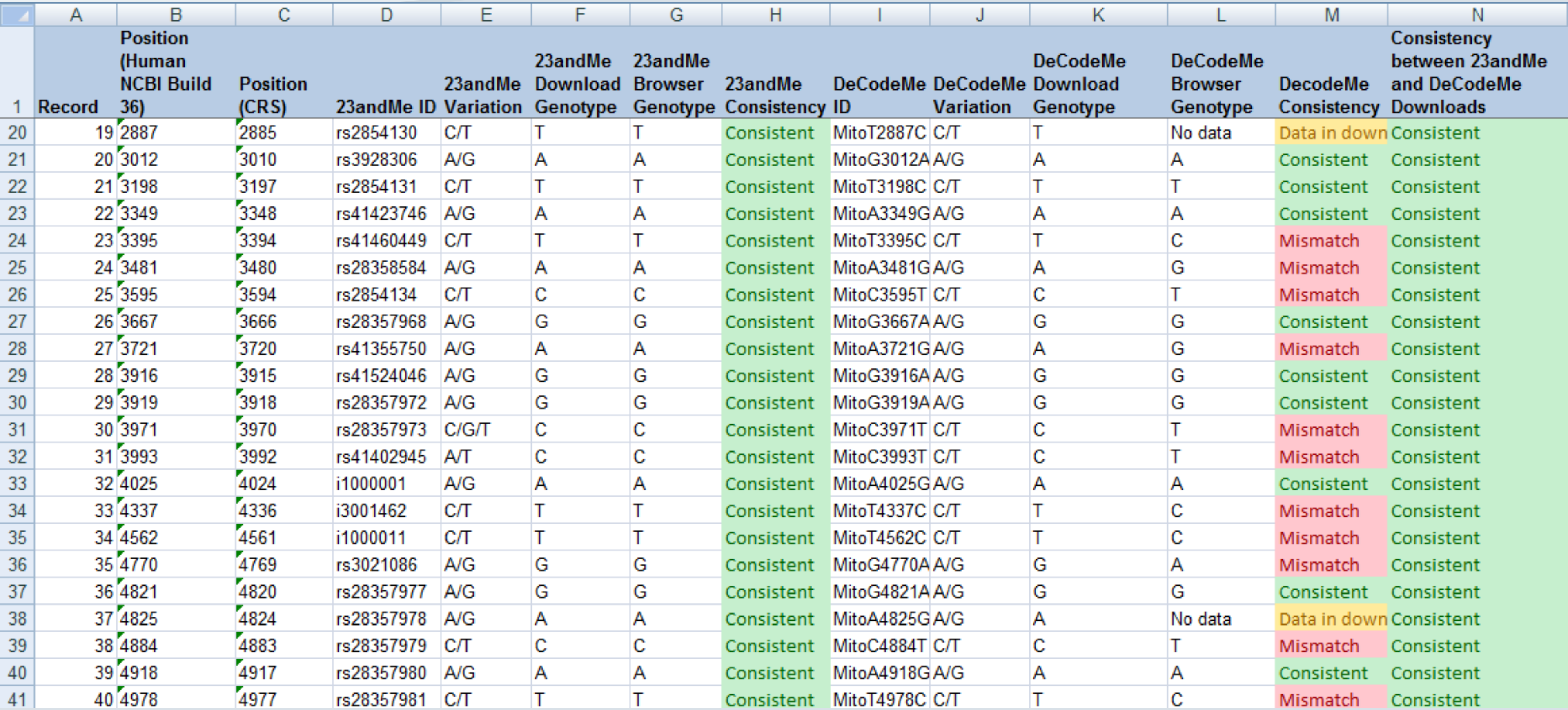
No harm was done, but the incident serves as a cautionary tale for personalised medicine. As we move towards a future in which readouts from our genomes will routinely be queried by computer systems to help doctors make important clinical decisions, similar glitches could cause prescribing errors – with 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.

Read the story

NewScientist



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Record	Position (Human NCBI Build 36)	Position (CRS)	23andMe ID	23andMe Variation	23andMe Download Genotype	23andMe Browser Genotype	23andMe Consistency	DeCodeMe ID	DeCodeMe Variation	DeCodeMe Download Genotype	DeCodeMe Browser Genotype	DecodeMe Consistency	Consistency between 23andMe and DeCodeMe 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

Spreadsheets

Microsoft Excel

Libre Office or Open Office Calc

Google Drive Sheets

Database managers

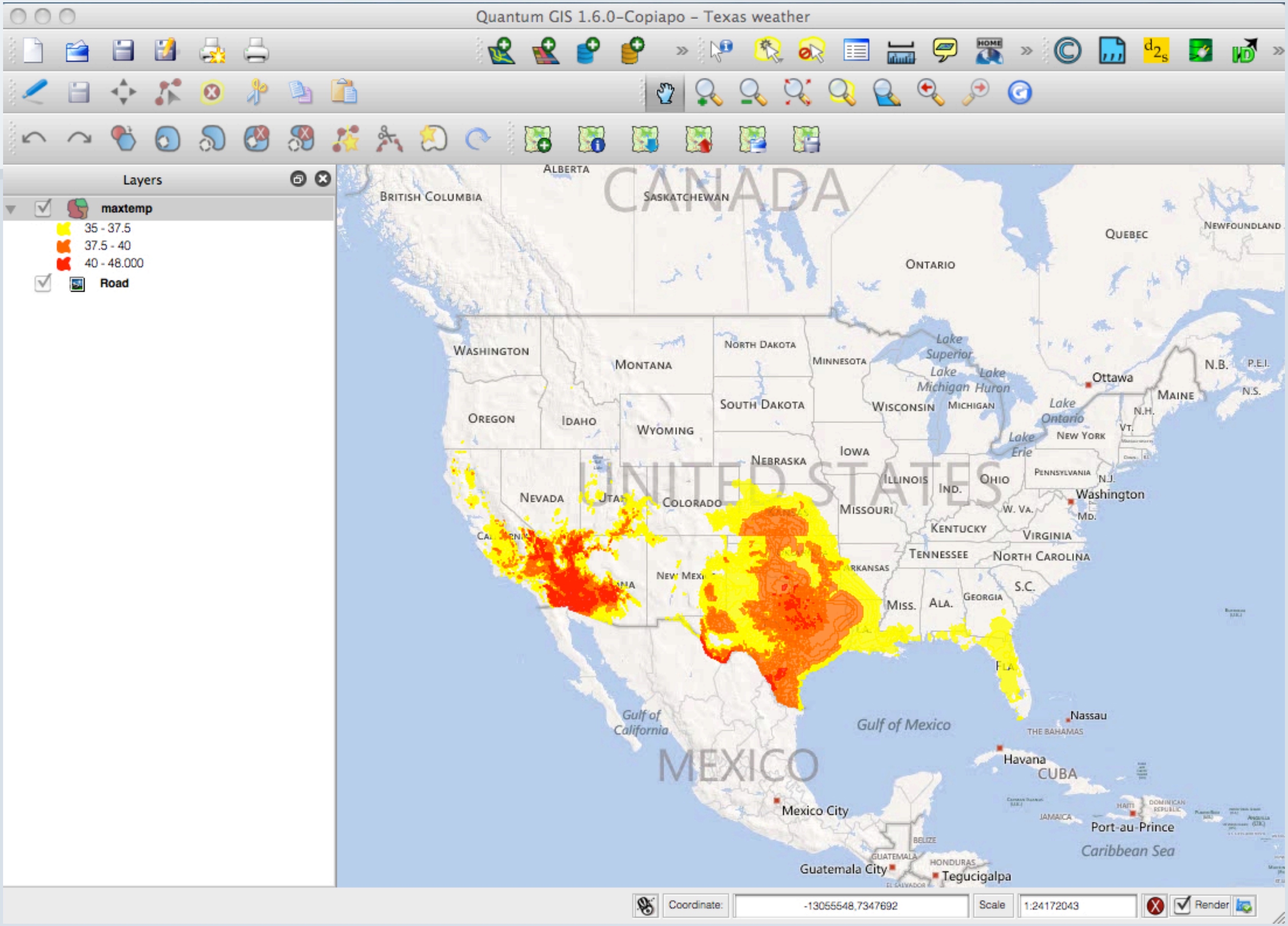
Microsoft Access

MySQL

PostgreSQL

SQLite

Tools and stories: putting data onto maps



Quantum GIS 1.6.0-Copiapo – Texas weather

Layers

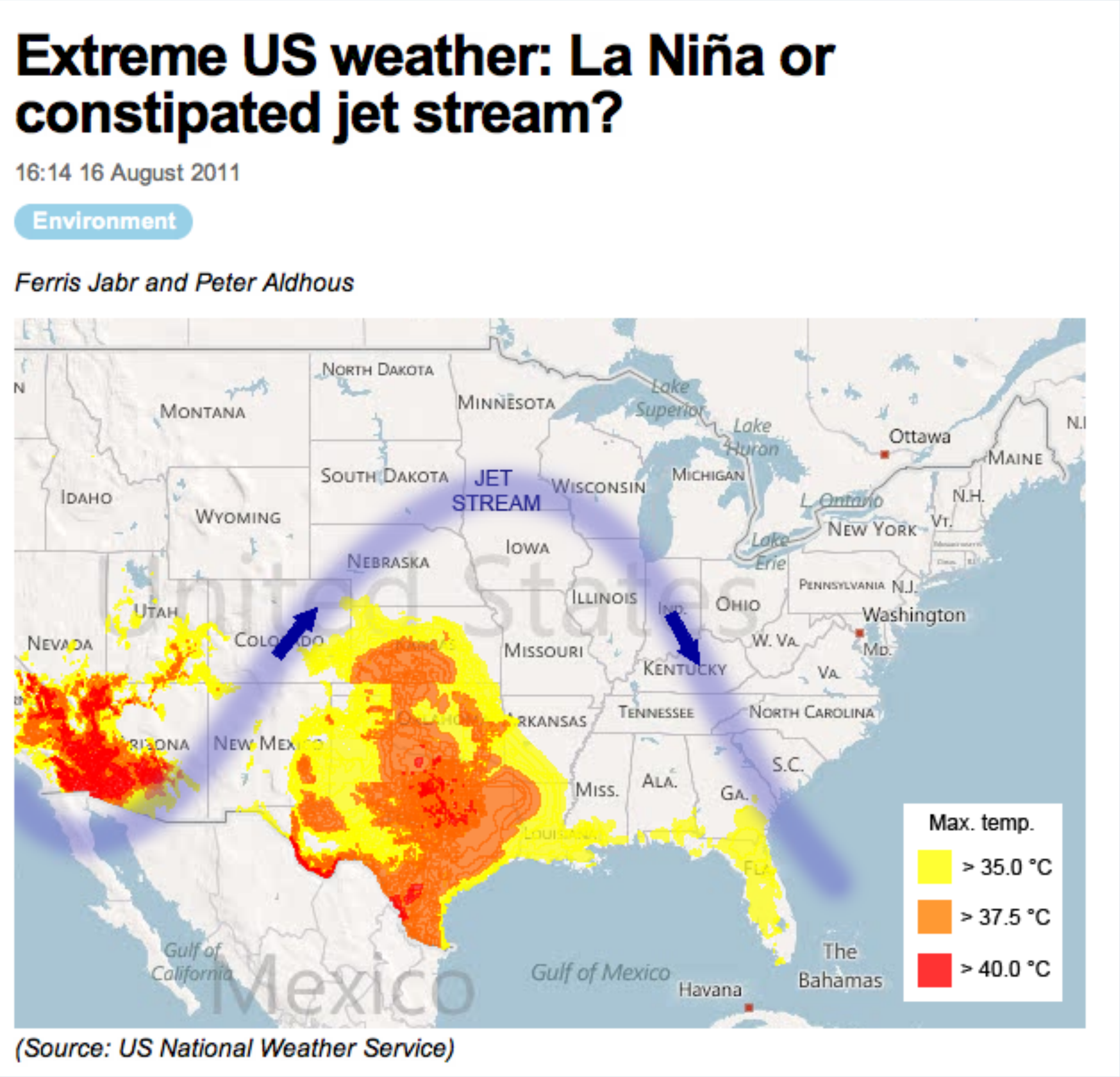
- ☑ maxtemp
  - 🟨 35 - 37.5
  - 🟧 37.5 - 40
  - 🟥 40 - 48.000
- ☑ Road

Coordinate: -13055548,7347692

Scale 1:24172043

☑ Render

Tools and stories: putting data onto maps



Extreme US weather: La Niña or constipated jet stream?

16:14 16 August 2011

Environment

Ferris Jabr and Peter Aldhous

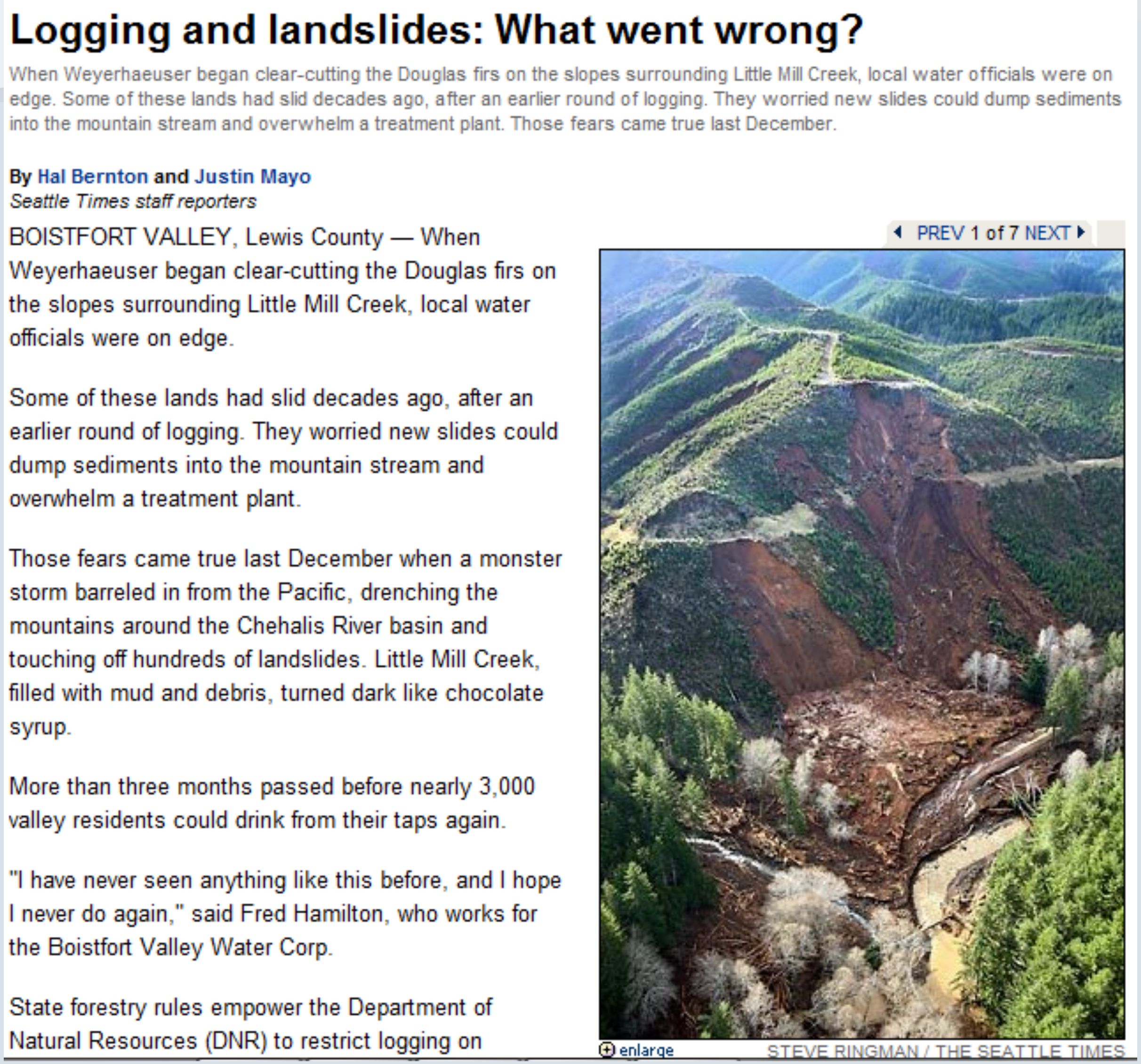
(Source: US National Weather Service)

NewScientist

Explore the graphic online

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

STEVE RINGMAN / THE SEATTLE TIMES

Data: GIS data on clear-cuts, landslides and prior studies of the hazards 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.

Free GIS software

Quantum GIS

MapWindow

Other free mapping tools

Google Maps

Google Earth

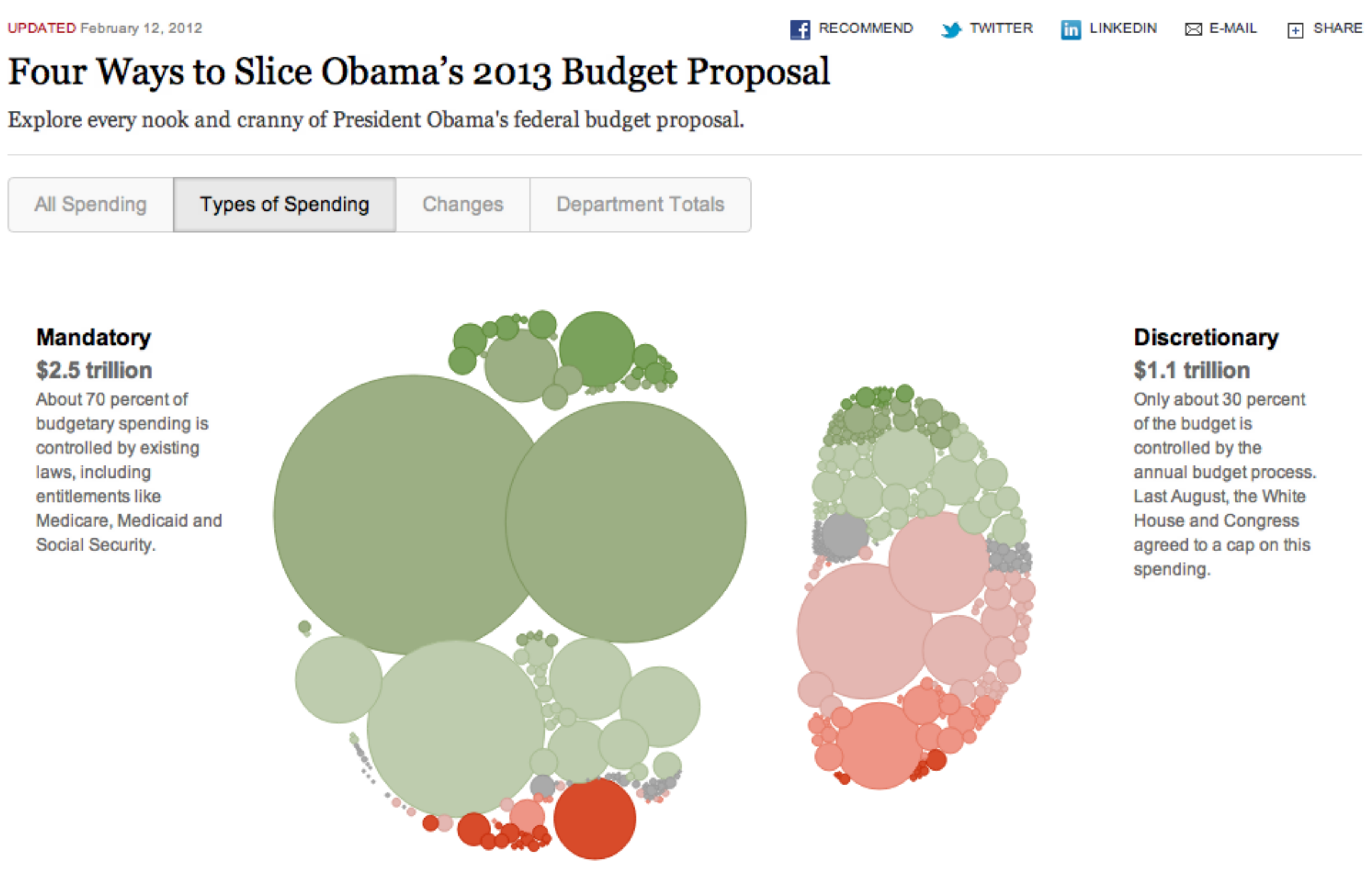
Google Fusion Tables

TileMill

Indiemapper

Geocommons

Tools and stories: online interactive data visualization



UPDATED February 12, 2012

Four Ways to Slice Obama’s 2013 Budget Proposal

Explore every nook and cranny of President Obama's federal budget proposal.

Mandatory

$2.5 trillion

About 70 percent of budgetary spending is controlled by existing laws, including entitlements like Medicare, Medicaid and Social Security.

Discretionary

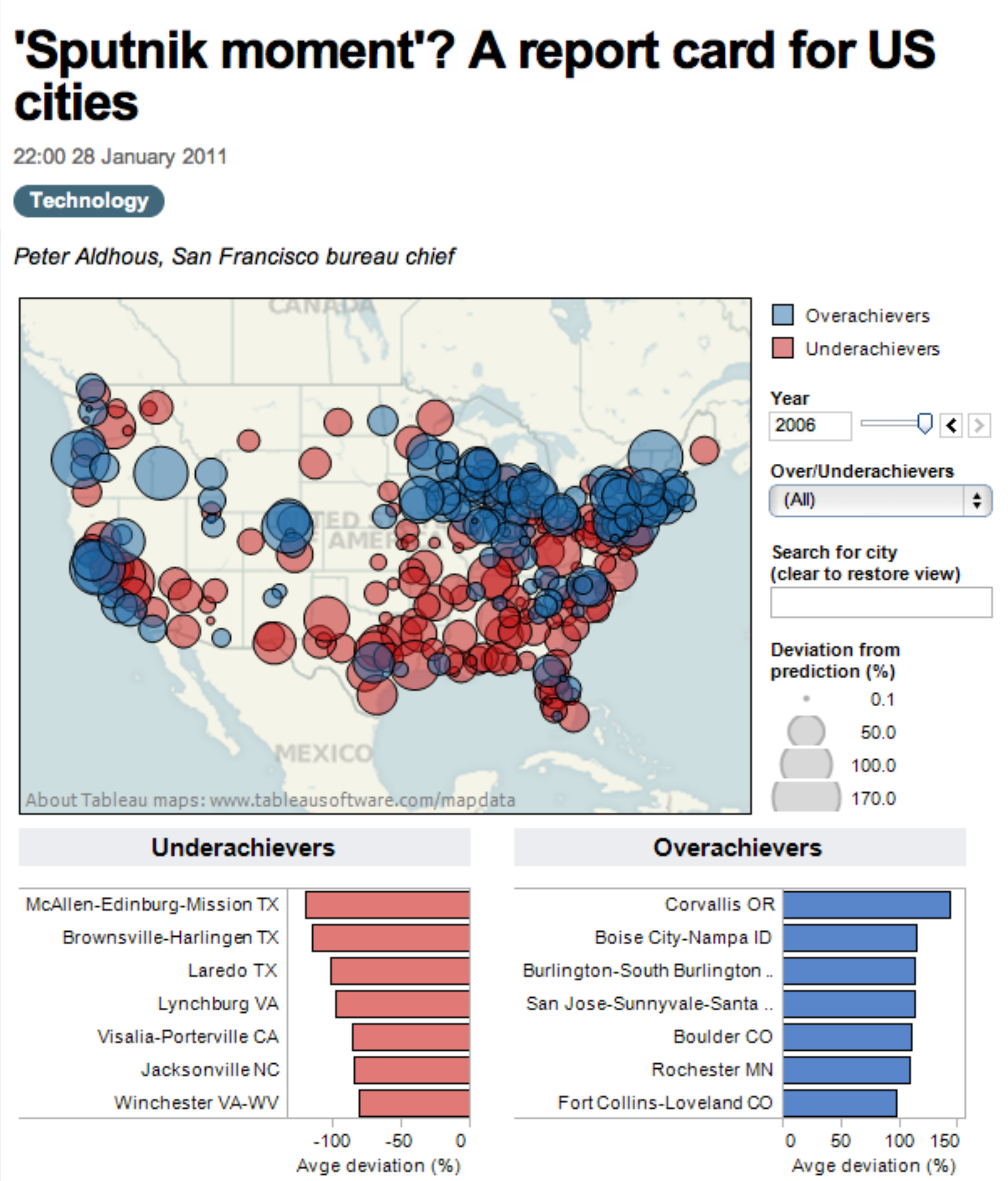
$1.1 trillion

Only about 30 percent of the budget is controlled by the annual budget process. Last August, the White House and Congress agreed to a cap on this spending.

The New York Times

Explore this graphic online

Tools and stories: online interactive data visualization



'Sputnik moment'? A report card for US cities

22:00 28 January 2011

Technology

Peter Aldhous, San Francisco bureau chief

Year: 2006

Over/Underachievers: (All)

Search for city (clear to restore view):

NewScientist

Explore this graphic online

Free tools for online data visualization

Tableau Public

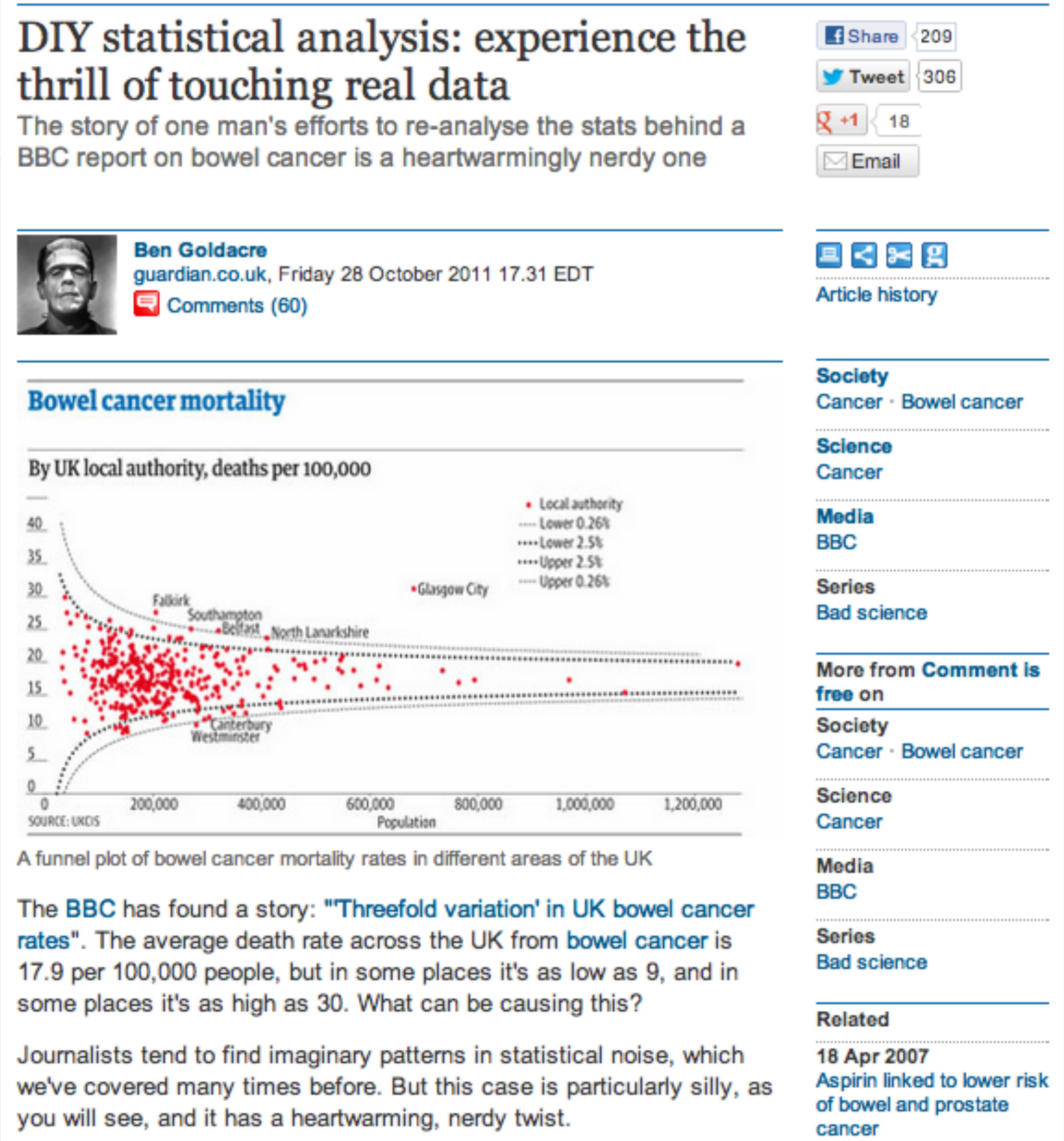
Many Eyes

Google Fusion Tables

Google Public Data Explorer

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, Friday 28 October 2011 17.31 EDT
Comments (60)

Bowel cancer mortality

By UK local authority, deaths per 100,000

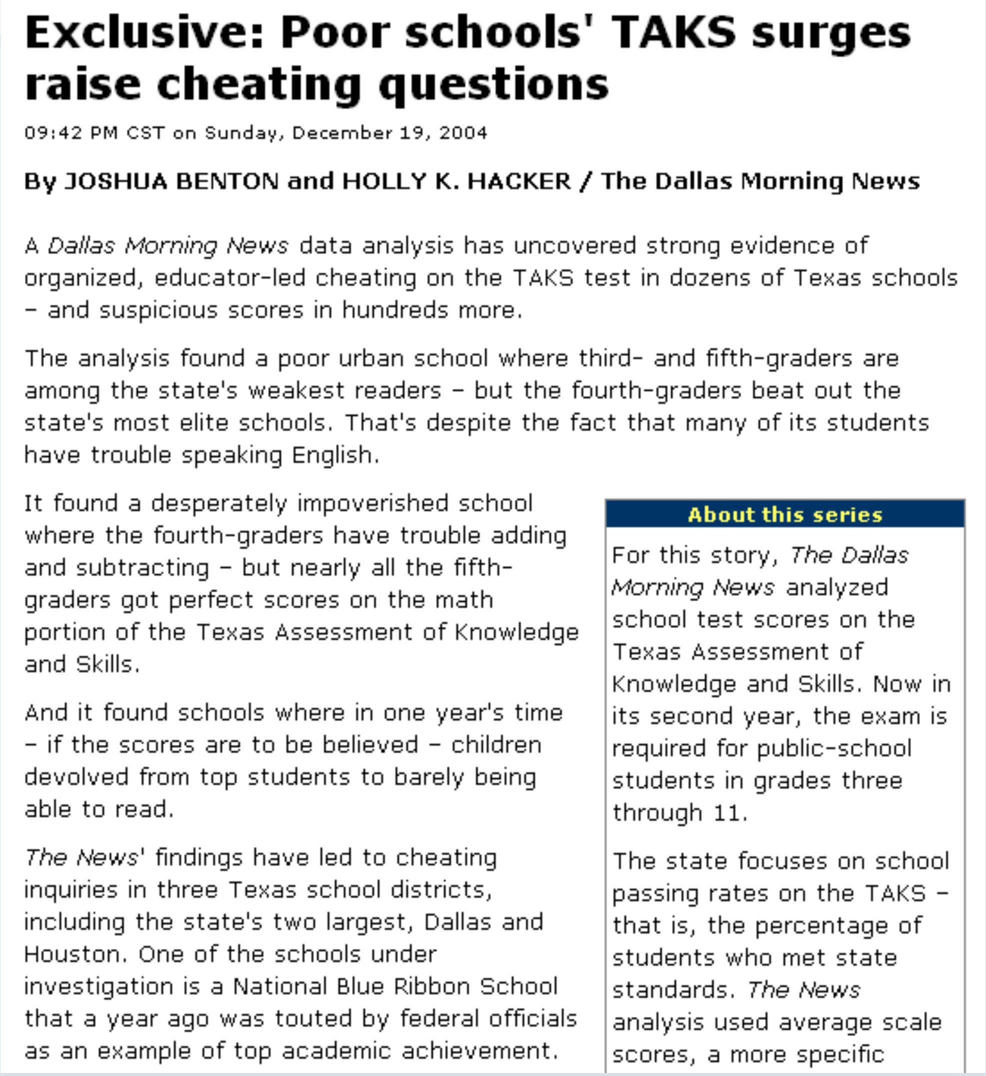
A funnel plot of bowel cancer mortality rates in different areas of the UK

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.

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.

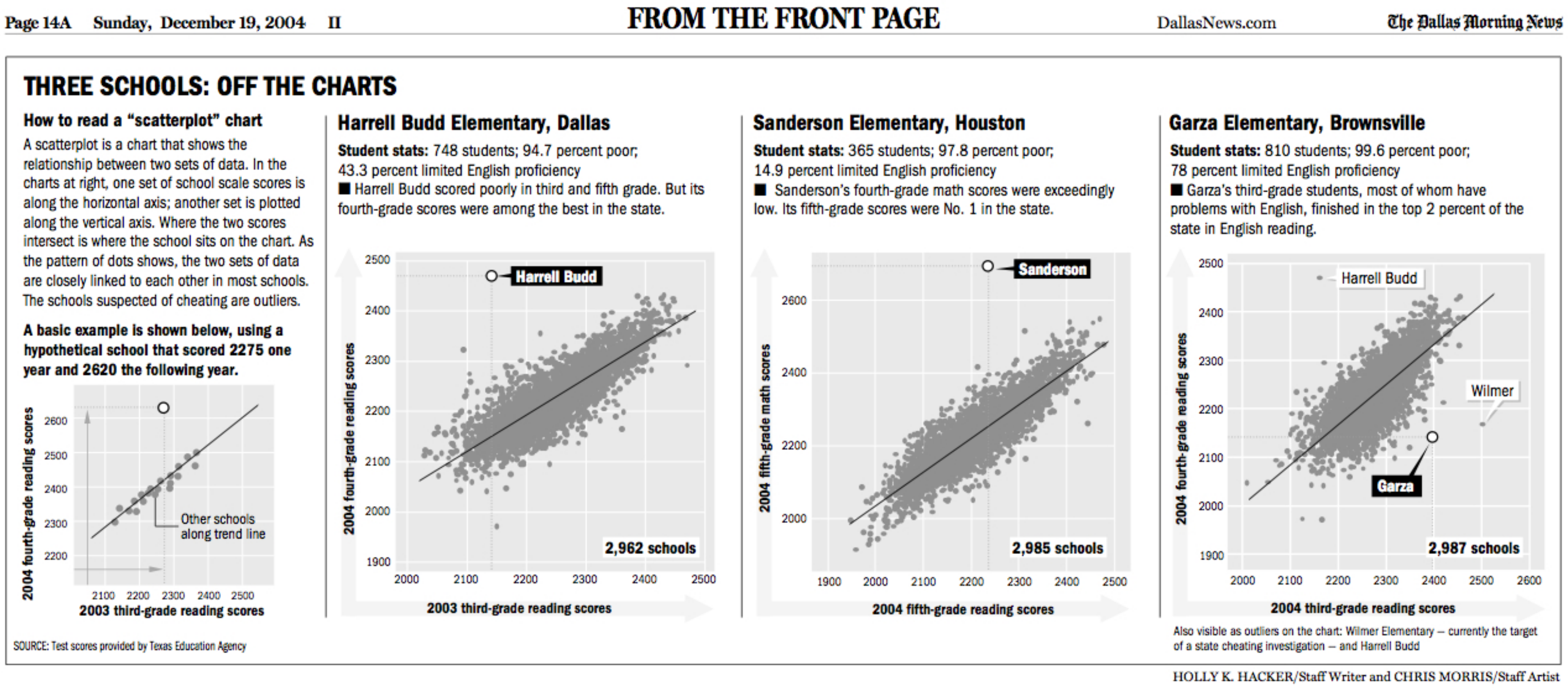
The state focuses on school passing rates on the TAKS – that is, the percentage of students who met state standards. *The News* analysis used average scale scores, a more specific

Data: Regression analysis of 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 its own data to perform similar analysis.

Tools and stories: statistical analysis

dallasnews.com
The Dallas Morning News



Page 14A Sunday, December 19, 2004 II FROM THE FRONT PAGE DallasNews.com The Dallas Morning News

THREE SCHOOLS: OFF THE CHARTS

How to read a "scatterplot" chart

A scatterplot is a chart that shows the relationship between two sets of data. In the charts at right, one set of school scale scores is along the horizontal axis; another set is plotted along the vertical axis. Where the two scores intersect is where the school sits on the chart. As the pattern of dots shows, the two sets of data are closely linked to each other in most schools. The schools suspected of cheating are outliers.

A basic example is shown below, using a hypothetical school that scored 2275 one year and 2620 the following year.

2004 fourth-grade reading scores
2600
2500
2400
2300
2200
Other schools along trend line
2100 2200 2300 2400 2500
2003 third-grade reading scores

SOURCE: Test scores provided by Texas Education Agency

Harrell Budd Elementary, Dallas

Student stats: 748 students; 94.7 percent poor; 43.3 percent limited English proficiency

■ Harrell Budd scored poorly in third and fifth grade. But its fourth-grade scores were among the best in the state.

Sanderson Elementary, Houston

Student stats: 365 students; 97.8 percent poor; 14.9 percent limited English proficiency

■ Sanderson's fourth-grade math scores were exceedingly low. Its fifth-grade scores were No. 1 in the state.

Garza Elementary, Brownsville

Student stats: 810 students; 99.6 percent poor; 78 percent limited English proficiency

■ Garza's third-grade students, most of whom have problems with English, finished in the top 2 percent of the state in English reading.

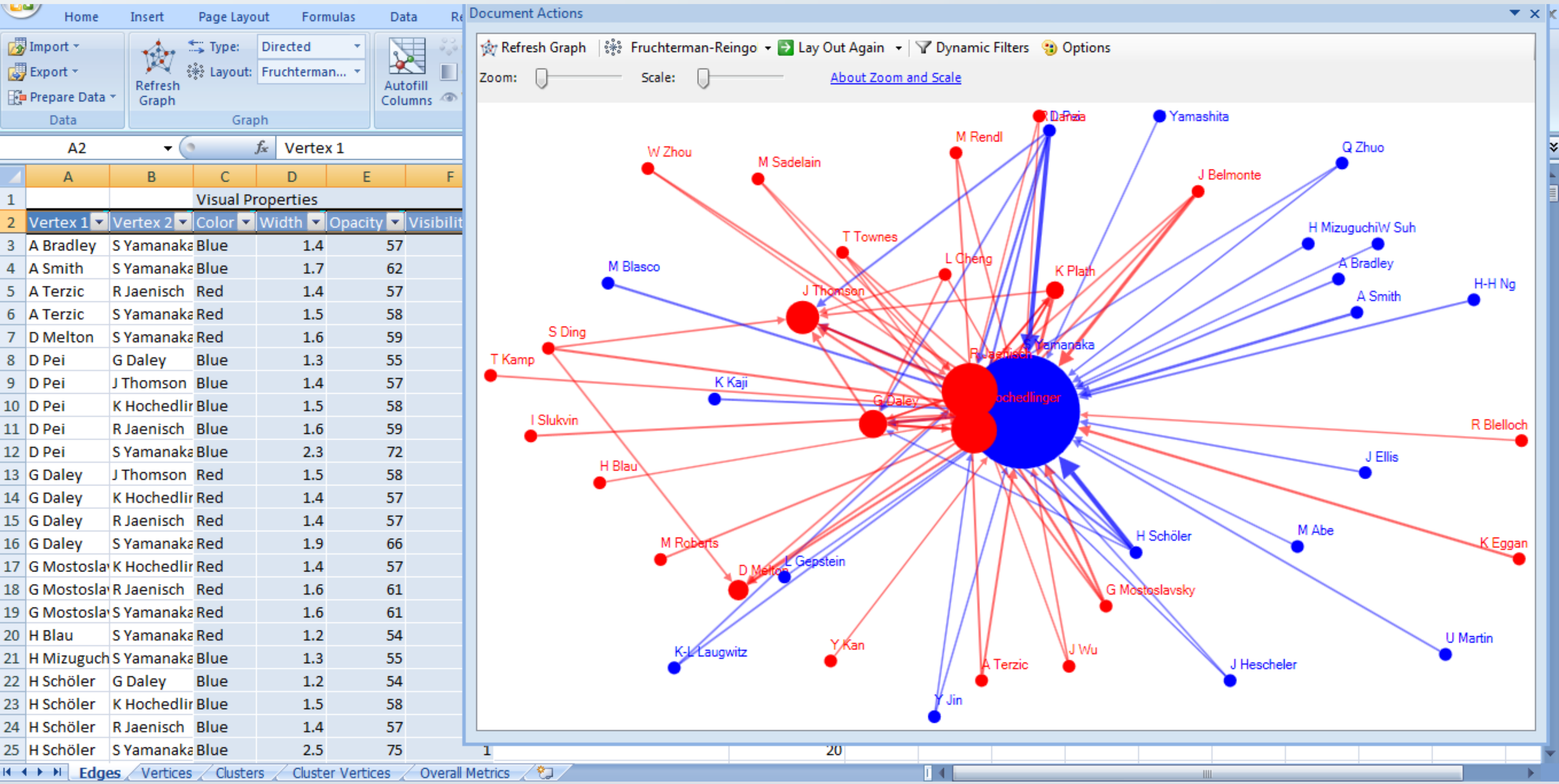
Also visible as outliers on the chart: Wilmer Elementary – currently the target of a state cheating investigation – and Harrell Budd

HOLLY K. HACKER/Staff Writer and CHRIS MORRIS/Staff Artist

Free software for statistical and graphical analysis

R, plus RAndFriends or RStudio for more user-friendly interfaces

Tools and stories: network analysis

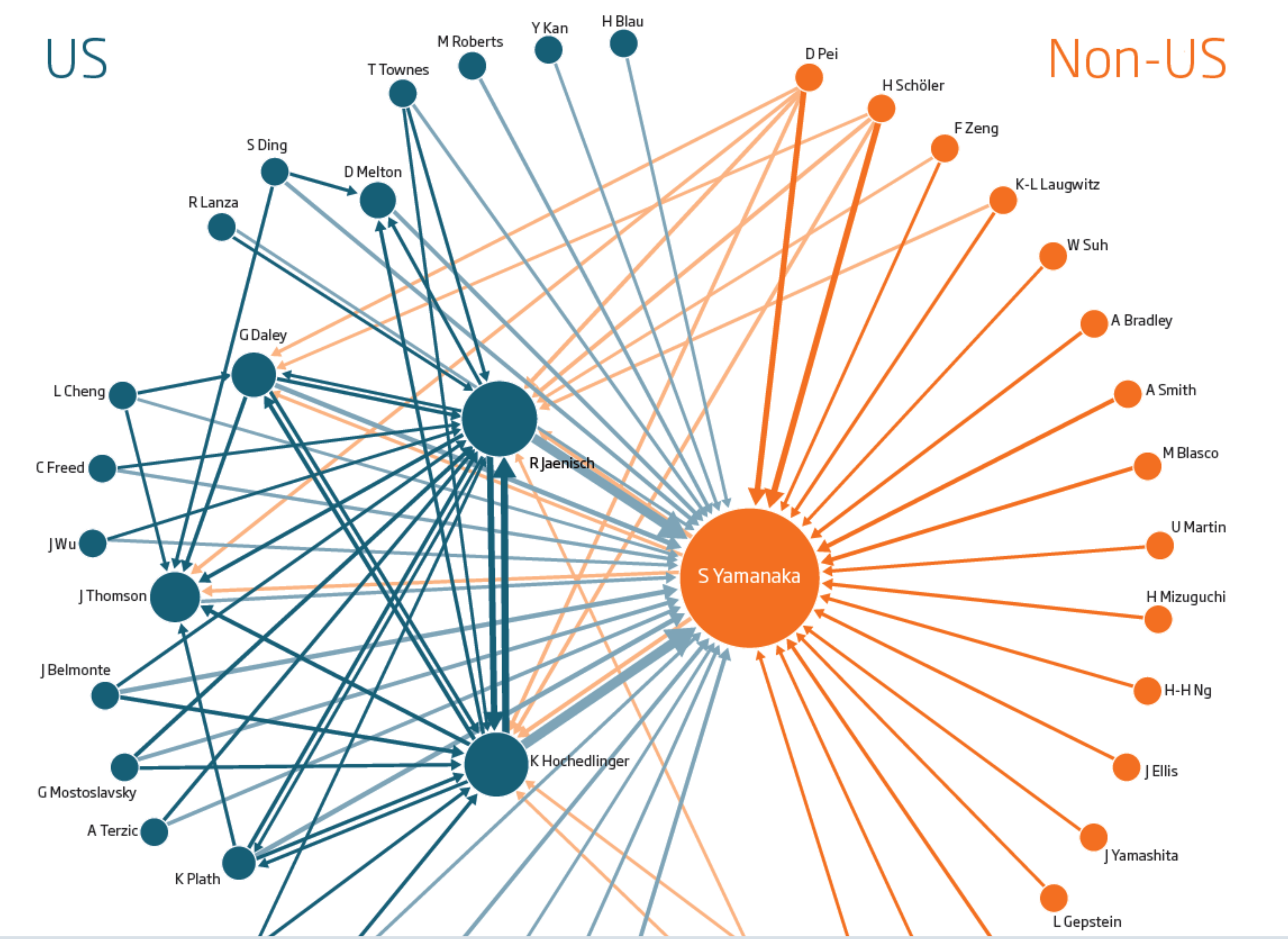


A2 | Vertex 1

	A	B	C	D	E	F
1			Visual Properties			
2	Vertex 1	Vertex 2	Color	Width	Opacity	Visibilit
3	A Bradley	S Yamanaka	Blue	1.4	57	
4	A Smith	S Yamanaka	Blue	1.7	62	
5	A Terzic	R Jaenisch	Red	1.4	57	
6	A Terzic	S Yamanaka	Red	1.5	58	
7	D Melton	S Yamanaka	Red	1.6	59	
8	D Pei	G Daley	Blue	1.3	55	
9	D Pei	J Thomson	Blue	1.4	57	
10	D Pei	K Hochedlir	Blue	1.5	58	
11	D Pei	R Jaenisch	Blue	1.6	59	
12	D Pei	S Yamanaka	Blue	2.3	72	
13	G Daley	J Thomson	Red	1.5	58	
14	G Daley	K Hochedlir	Red	1.4	57	
15	G Daley	R Jaenisch	Red	1.4	57	
16	G Daley	S Yamanaka	Red	1.9	66	
17	G Mostosla	K Hochedlir	Red	1.4	57	
18	G Mostosla	R Jaenisch	Red	1.6	61	
19	G Mostosla	S Yamanaka	Red	1.6	61	
20	H Blau	S Yamanaka	Red	1.2	54	
21	H Mizuguch	S Yamanaka	Blue	1.3	55	
22	H Schöler	G Daley	Blue	1.2	54	
23	H Schöler	K Hochedlir	Blue	1.5	58	
24	H Schöler	R Jaenisch	Blue	1.4	57	
25	H Schöler	S Yamanaka	Blue	2.5	75	

Tools and stories: network analysis

NewScientist



Data: Citations between corresponding authors of papers on "induced pluripotent stem cells" in high-impact journals.

Findings: This map of influence and connections in the field may help explain why non-US scientists seem to be losing the race to publish

Read the story

Software for network analysis

NodeXL (free, extension to Excel 2007/2010)

Gephi (free)

UCINET (free trial version for 60 days, then $250)

Data journalism tutorials

Spreadsheet tutorial in Excel 2010

Database tutorial in Access 2010

Data visualization with Tableau Public

Making a map with Google Fusion Tables

Introduction to R for statistics

Network analysis with NodeXL

Data journalism: what it can do for you

NCSWA workshop, January 12, 2013

Peter Aldhous,
San Francisco Bureau Chief

NewScientist

peter@peteraldhous.com
Twitter: @paldhous