Data journalism in science

UK Conference of Science Journalists, 25 June 2012

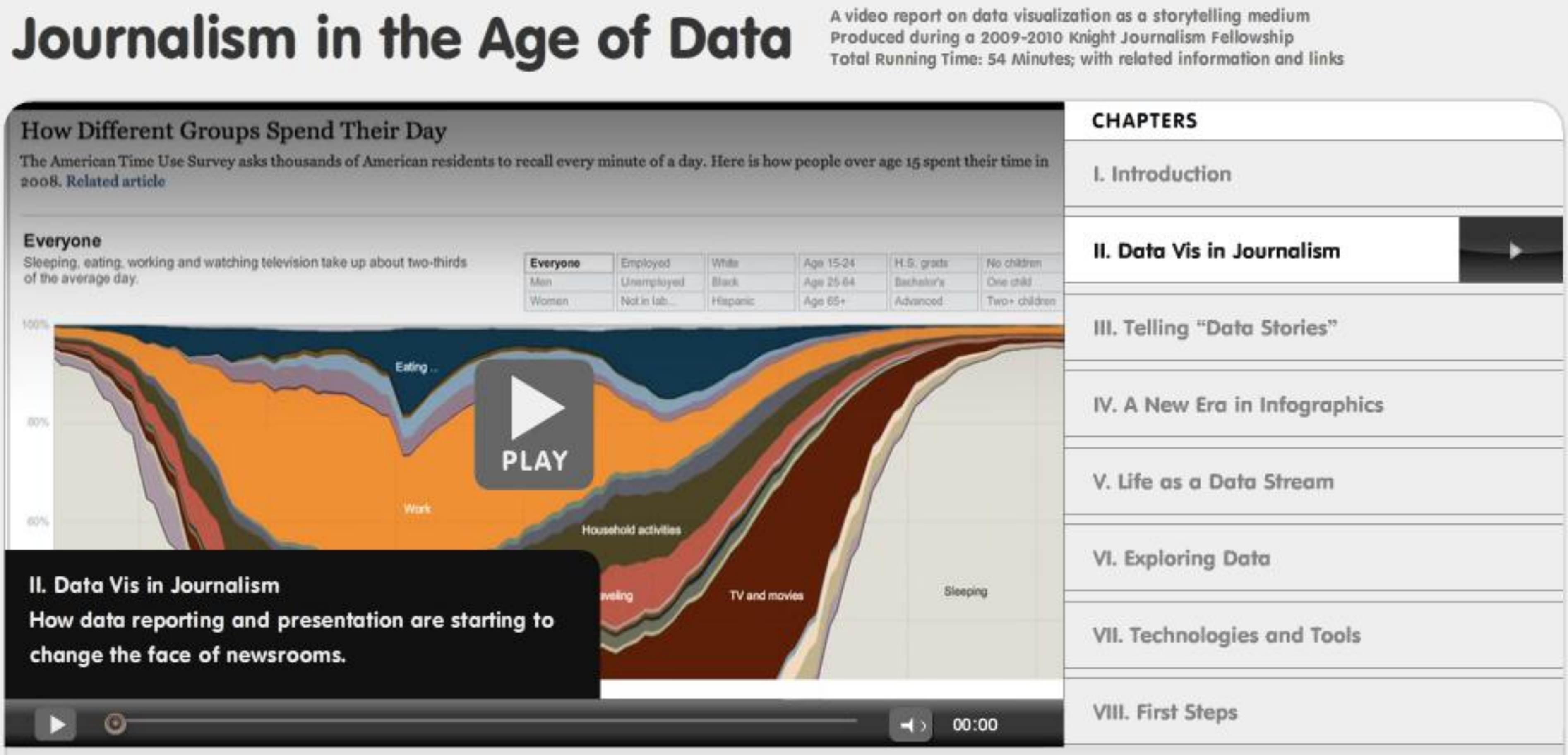
Peter Aldhous,

San Francisco Bureau Chief

peter@peteraldhous.com

NewScientist

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.

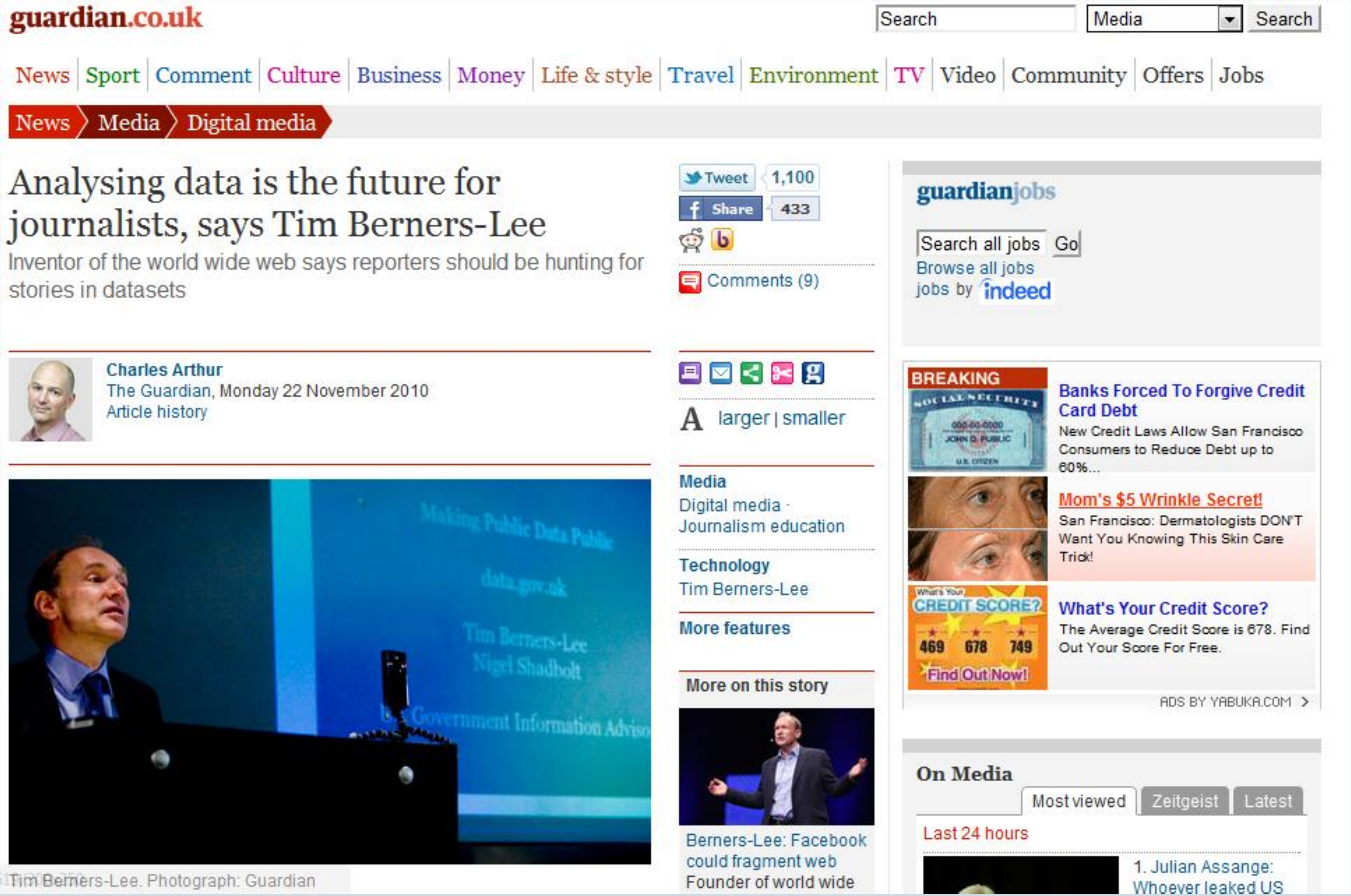
II. Data Vis in Journalism
How data reporting and presentation are starting to change the face of newsrooms.

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

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

Tim Berners-Lee. Photograph: Guardian

guardianjobs

Search all jobs Go
Browse all jobs
jobs by indeed

On Media

Most viewed | Zeitgeist | Latest

Last 24 hours

1. Julian Assange: Whoever leaked US

The basics

- **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

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.

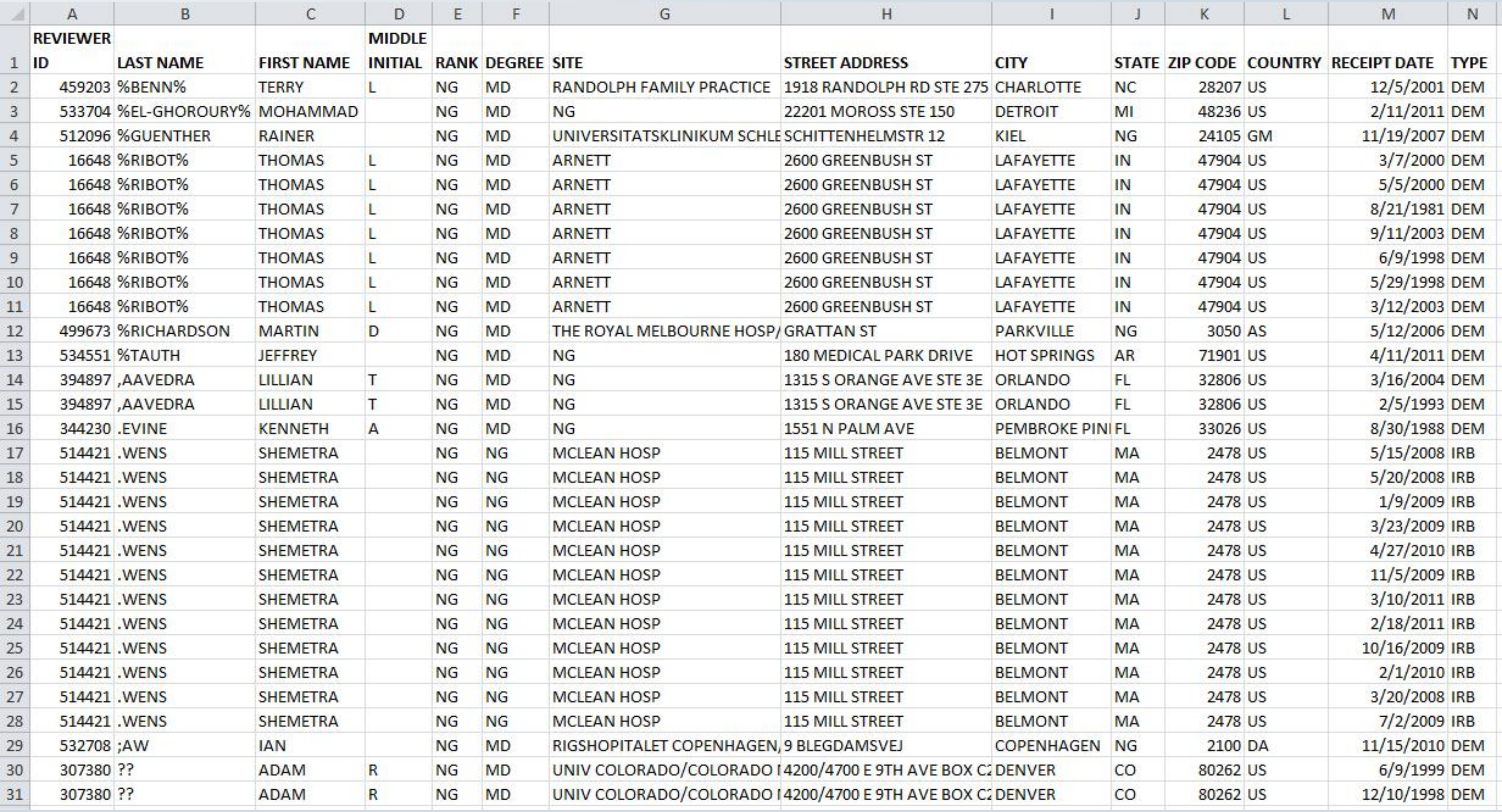
You will almost always need to reformat and edit data to suit your purposes; frequently you will have to do extensive data “cleaning”.

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

Google Refine

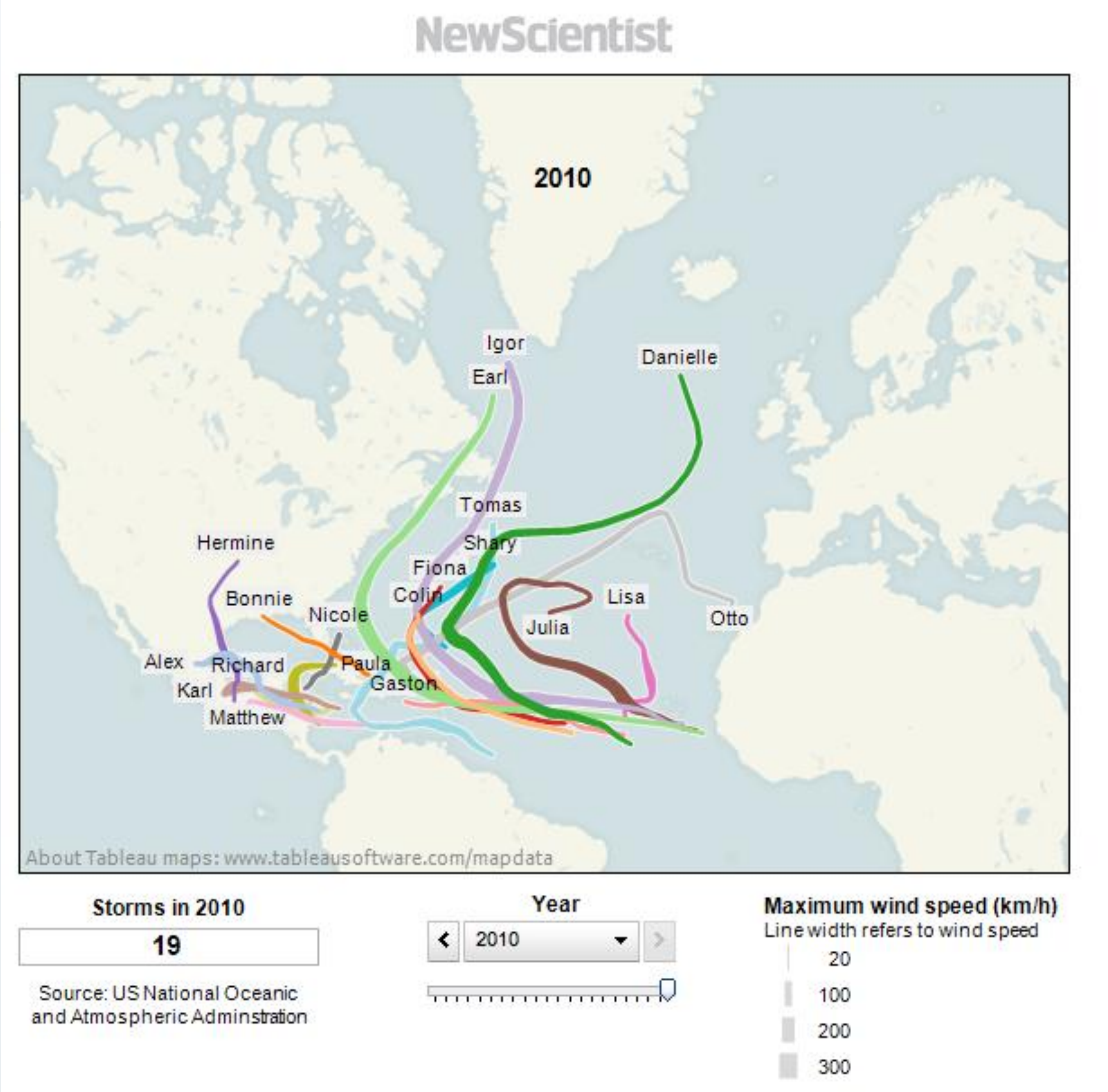
There are good video tutorials for this tool at the link above.

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

Online interactive data visualisation



Storms in 2010

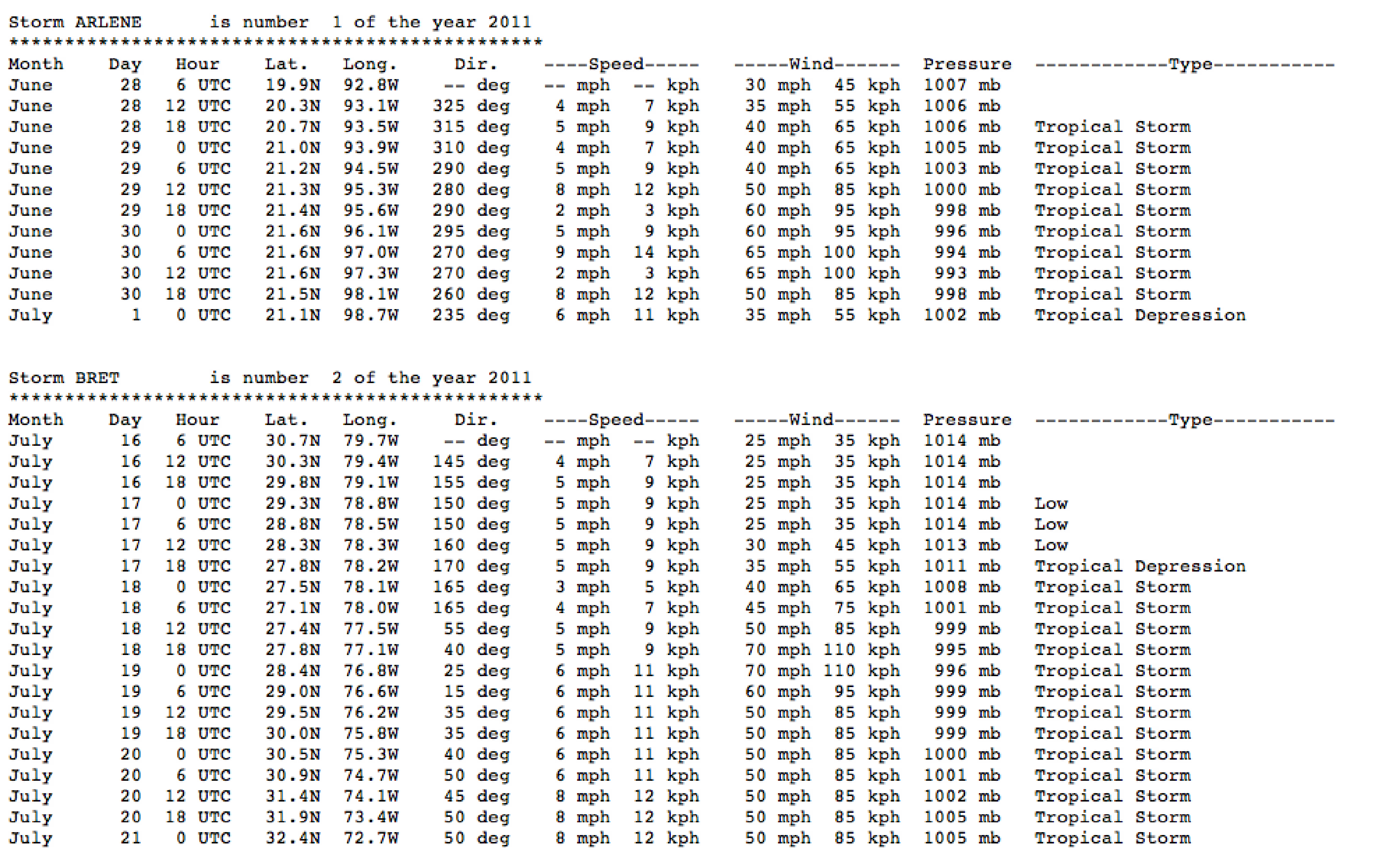
19

Source: US National Oceanic
and Atmospheric Adminstration

Year: 2010

Explore the graphic

Atlantic storm 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
July        16   18 UTC  29.8N  79.1W   155 deg   5 mph    9 kph    25 mph   35 kph  1014 mb
July        17    0 UTC  29.3N  78.8W   150 deg   5 mph    9 kph    25 mph   35 kph  1014 mb   Low
July        17    6 UTC  28.8N  78.5W   150 deg   5 mph    9 kph    25 mph   35 kph  1014 mb   Low
July        17   12 UTC  28.3N  78.3W   160 deg   5 mph    9 kph    30 mph   45 kph  1013 mb   Low
July        17   18 UTC  27.8N  78.2W   170 deg   5 mph    9 kph    35 mph   55 kph  1011 mb   Tropical Depression
July        18    0 UTC  27.5N  78.1W   165 deg   3 mph    5 kph    40 mph   65 kph  1008 mb   Tropical Storm
July        18    6 UTC  27.1N  78.0W   165 deg   4 mph    7 kph    45 mph   75 kph  1001 mb   Tropical Storm
July        18   12 UTC  27.4N  77.5W    55 deg   5 mph    9 kph    50 mph   85 kph   999 mb   Tropical Storm
July        18   18 UTC  27.8N  77.1W    40 deg   5 mph    9 kph    70 mph  110 kph   995 mb   Tropical Storm
July        19    0 UTC  28.4N  76.8W    25 deg   6 mph   11 kph    70 mph  110 kph   996 mb   Tropical Storm
July        19    6 UTC  29.0N  76.6W    15 deg   6 mph   11 kph    60 mph   95 kph   999 mb   Tropical Storm
July        19   12 UTC  29.5N  76.2W    35 deg   6 mph   11 kph    50 mph   85 kph   999 mb   Tropical Storm
July        19   18 UTC  30.0N  75.8W    35 deg   6 mph   11 kph    50 mph   85 kph   999 mb   Tropical Storm
July        20    0 UTC  30.5N  75.3W    40 deg   6 mph   11 kph    50 mph   85 kph  1000 mb   Tropical Storm
July        20    6 UTC  30.9N  74.7W    50 deg   6 mph   11 kph    50 mph   85 kph  1001 mb   Tropical Storm
July        20   12 UTC  31.4N  74.1W    45 deg   8 mph   12 kph    50 mph   85 kph  1002 mb   Tropical Storm
July        20   18 UTC  31.9N  73.4W    50 deg   8 mph   12 kph    50 mph   85 kph  1005 mb   Tropical Storm
July        21    0 UTC  32.4N  72.7W    50 deg   8 mph   12 kph    50 mph   85 kph  1005 mb   Tropical Storm
```

Free tools for online data visualisation

Tableau Public

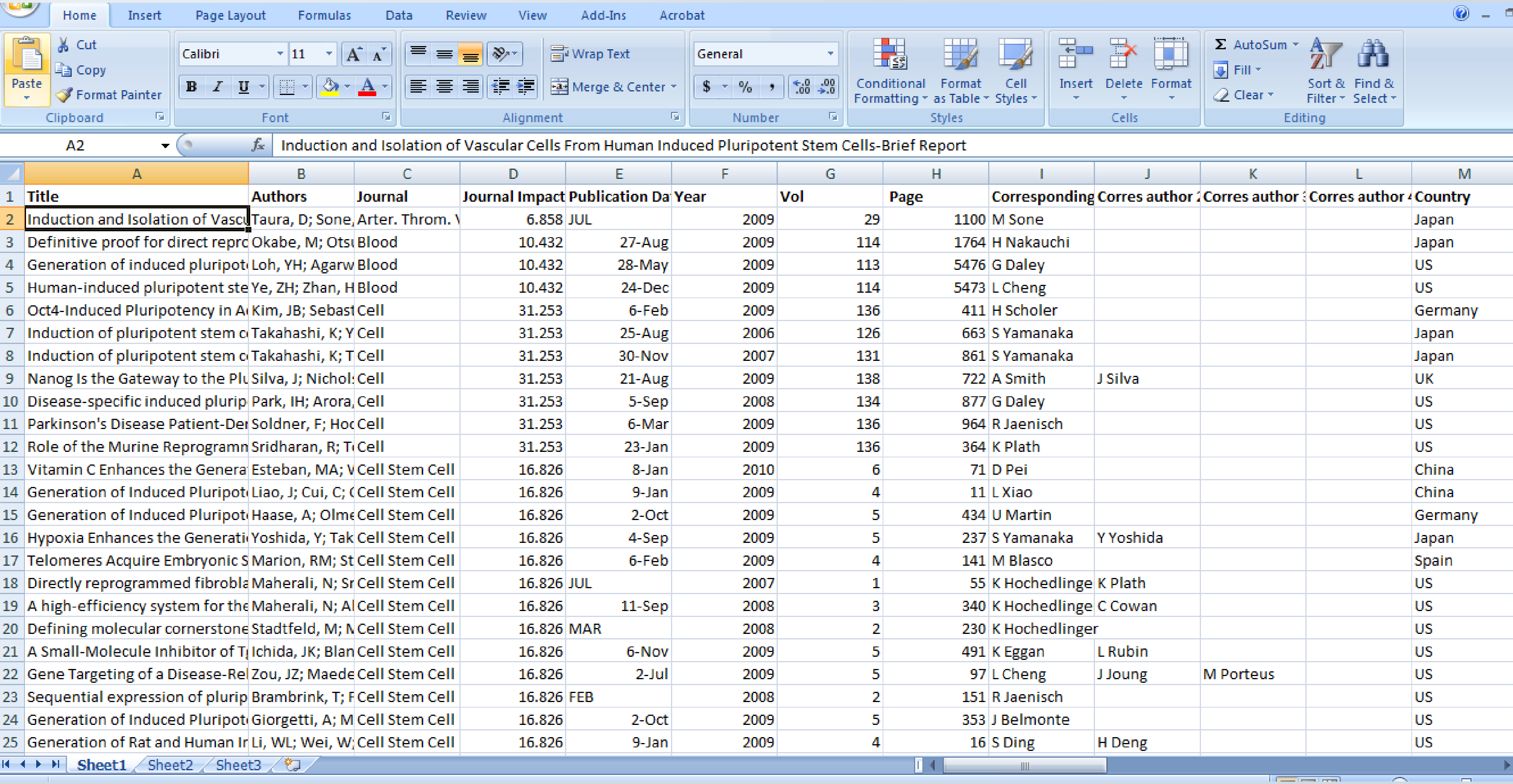
Many Eyes

Google Documents Gadgets

Google Fusion Tables

Google Public Data Explorer

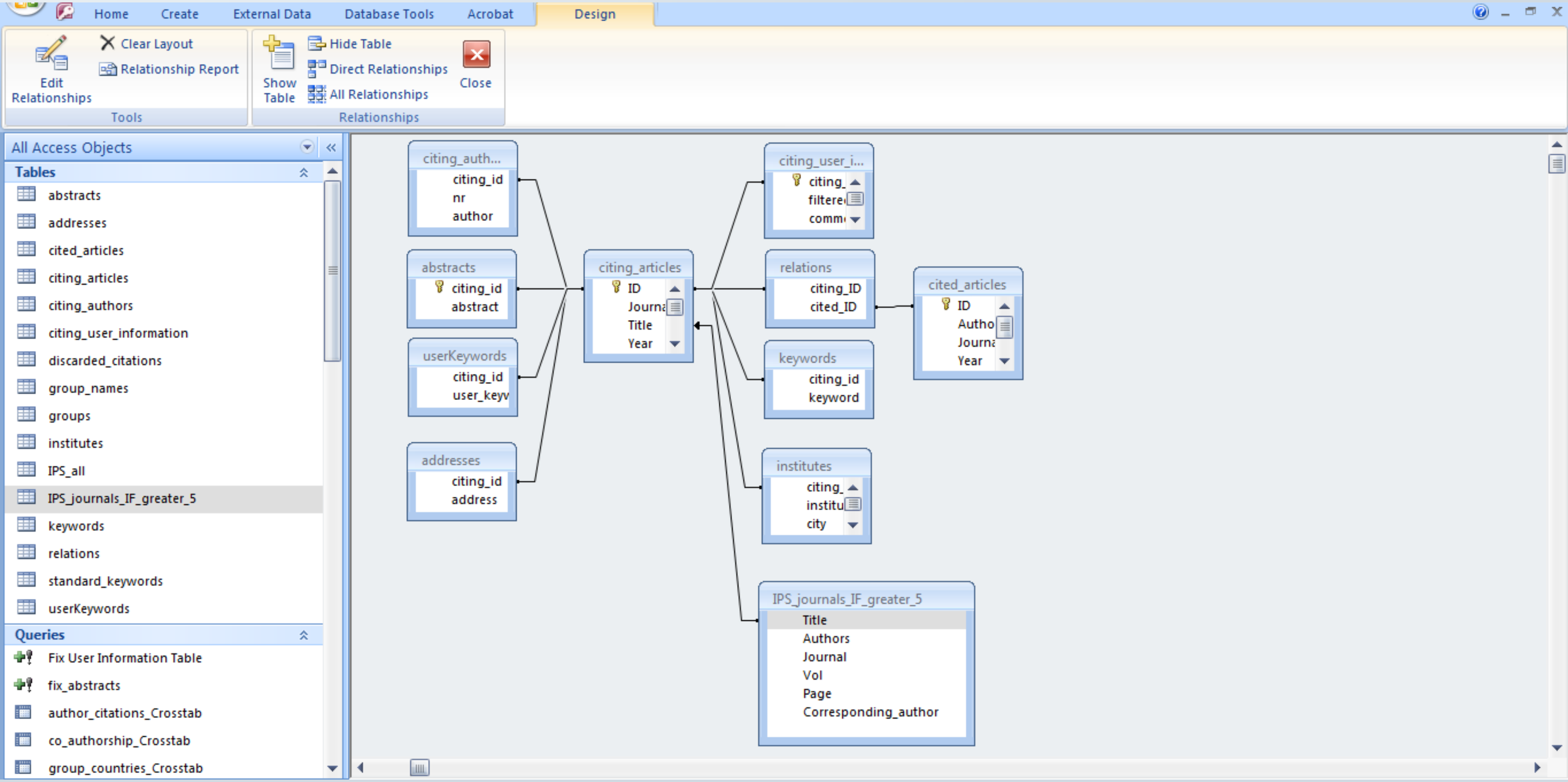
The basic tools: spreadsheets …



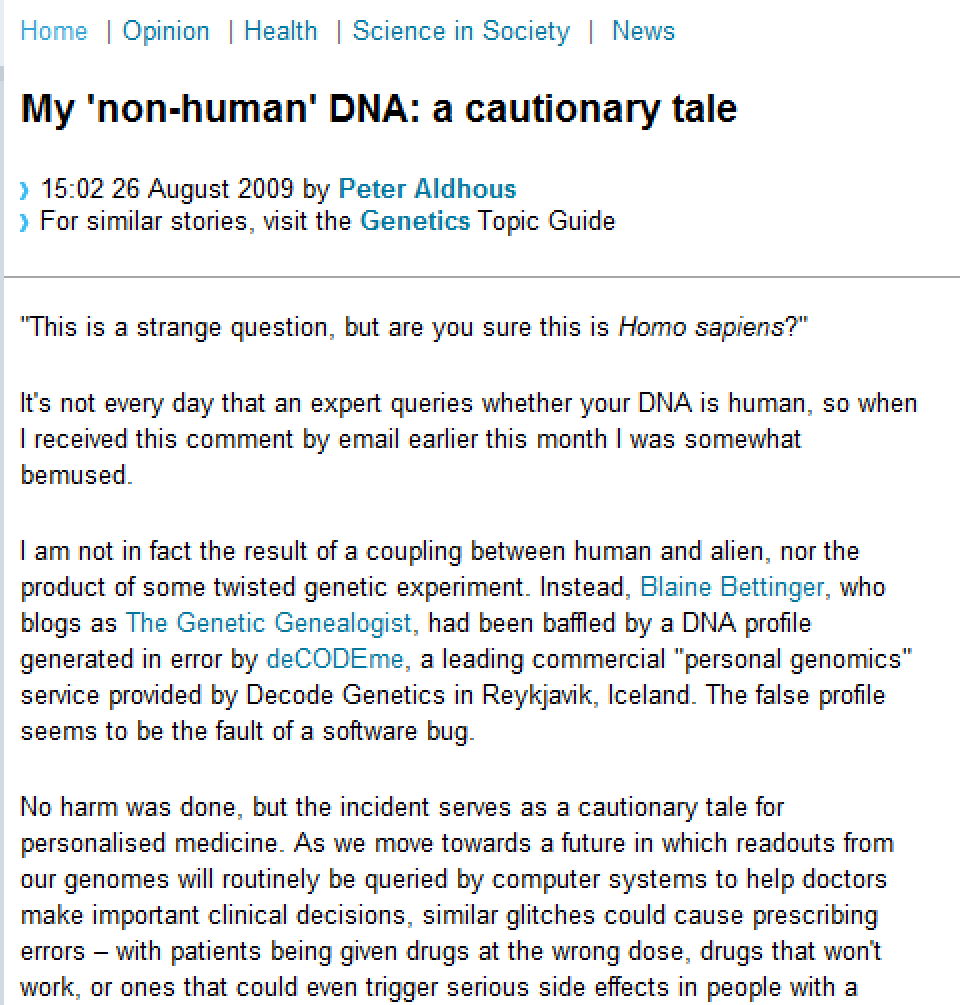
A2 *fx* Induction and Isolation of Vascular Cells From Human Induced Pluripotent Stem Cells-Brief Report

	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



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.

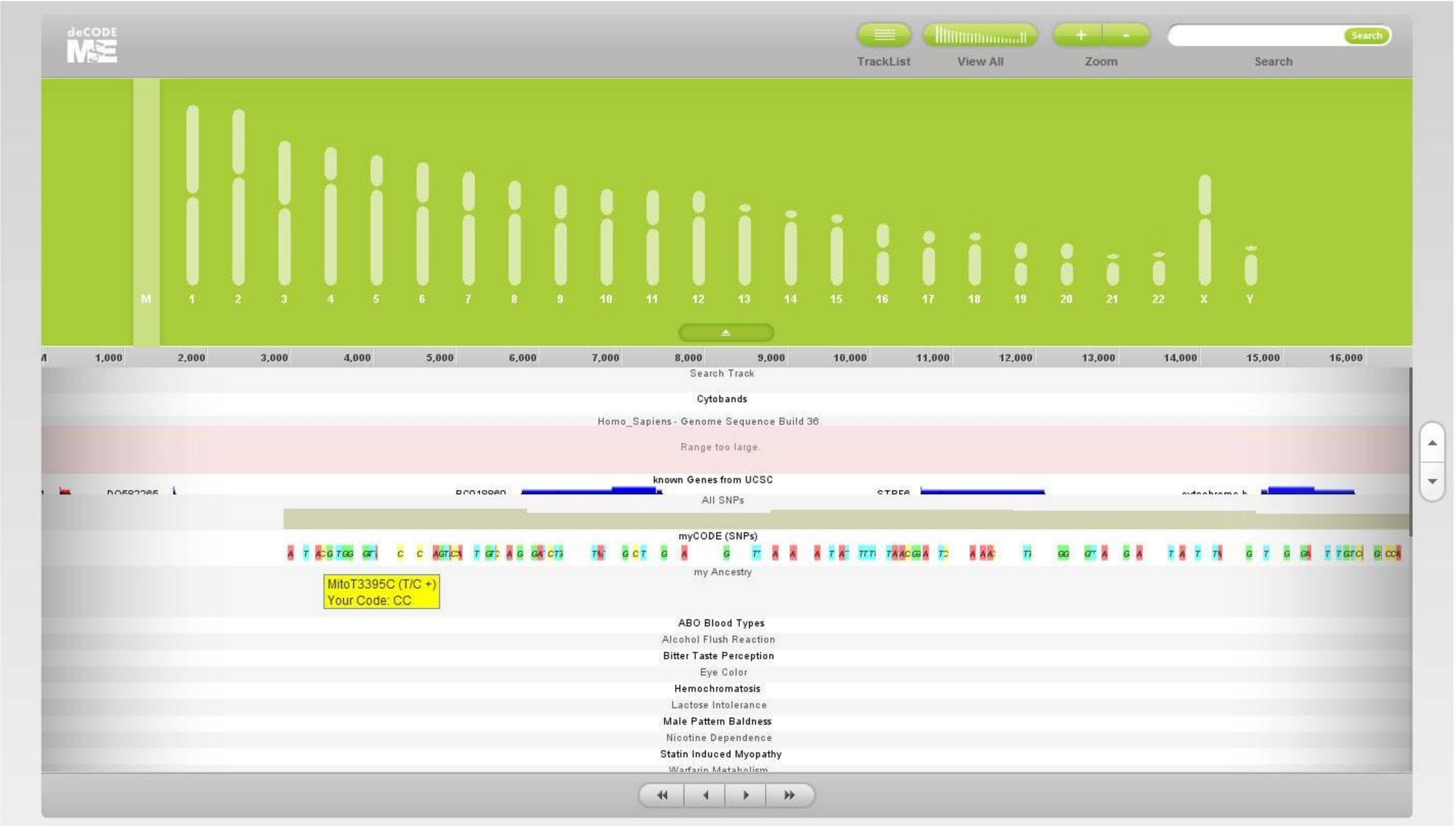
No harm was done, but the incident serves as a cautionary tale for personalised medicine. As we move towards a future in which readouts from our genomes will routinely be queried by computer systems to help doctors make important clinical decisions, similar glitches could cause prescribing errors – with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

DeCode’s genome browser



Search Track

Cytobands

Homo_Sapiens - Genome Sequence Build 36

Range too large.

known Genes from UCSC

All SNPs

myCODE (SNPs)

my Ancestry

MitoT3395C (T/C +)
Your Code: CC

ABO Blood Types

Alcohol Flush Reaction

Bitter Taste Perception

Eye Color

Hemochromatosis

Lactose Intolerance

Male Pattern Baldness

Nicotine Dependence

Statin Induced Myopathy

Warfarin Metabolism

Spreadsheets

Microsoft Excel

Libre Office or Open Office Calc

Google Documents

Database managers

Microsoft Access

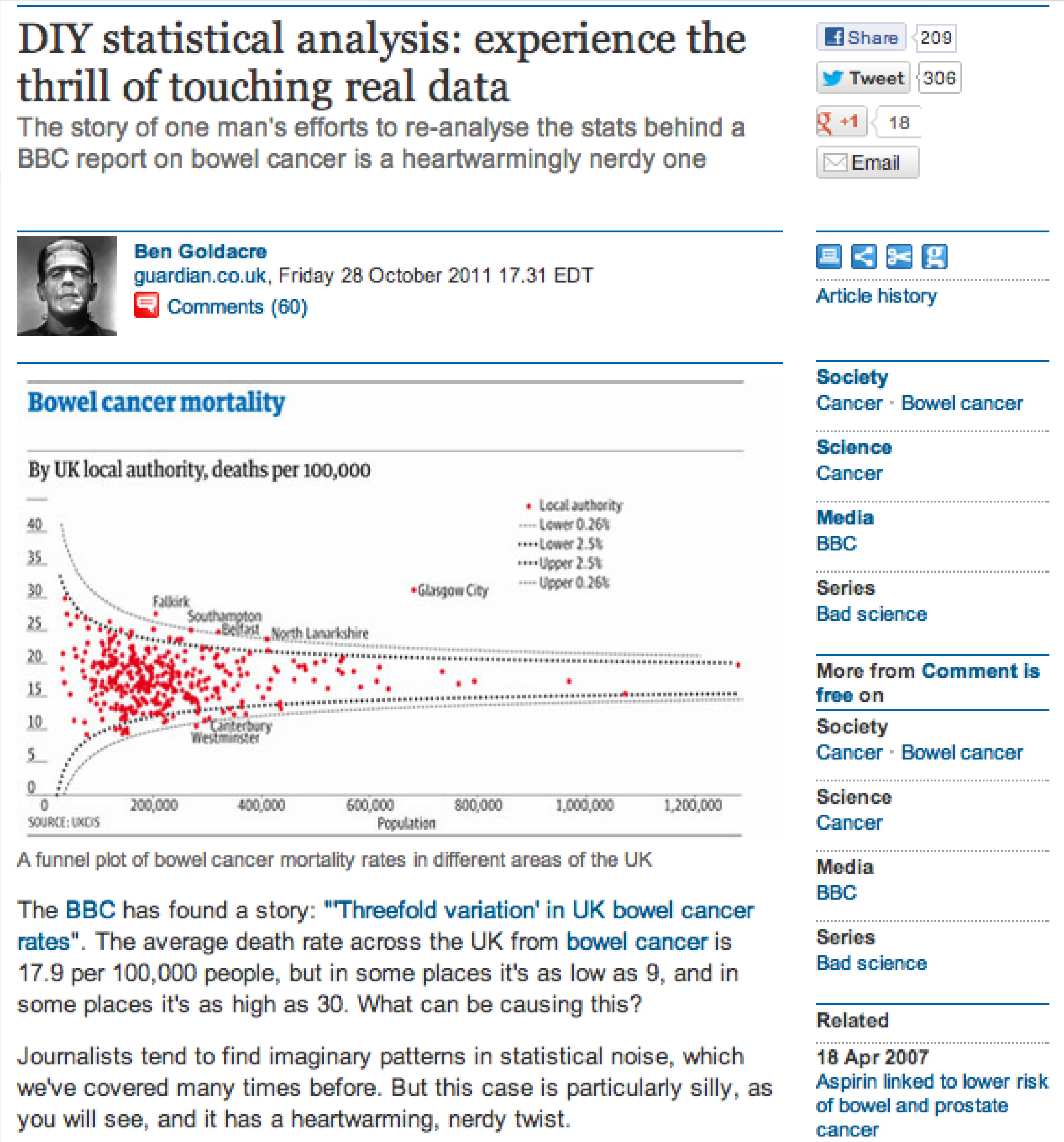
MySQL

PostgreSQL

SQLite

Beware running with scissors

Seek expert advice if what's needed is 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

SOURCE: UKCIS

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

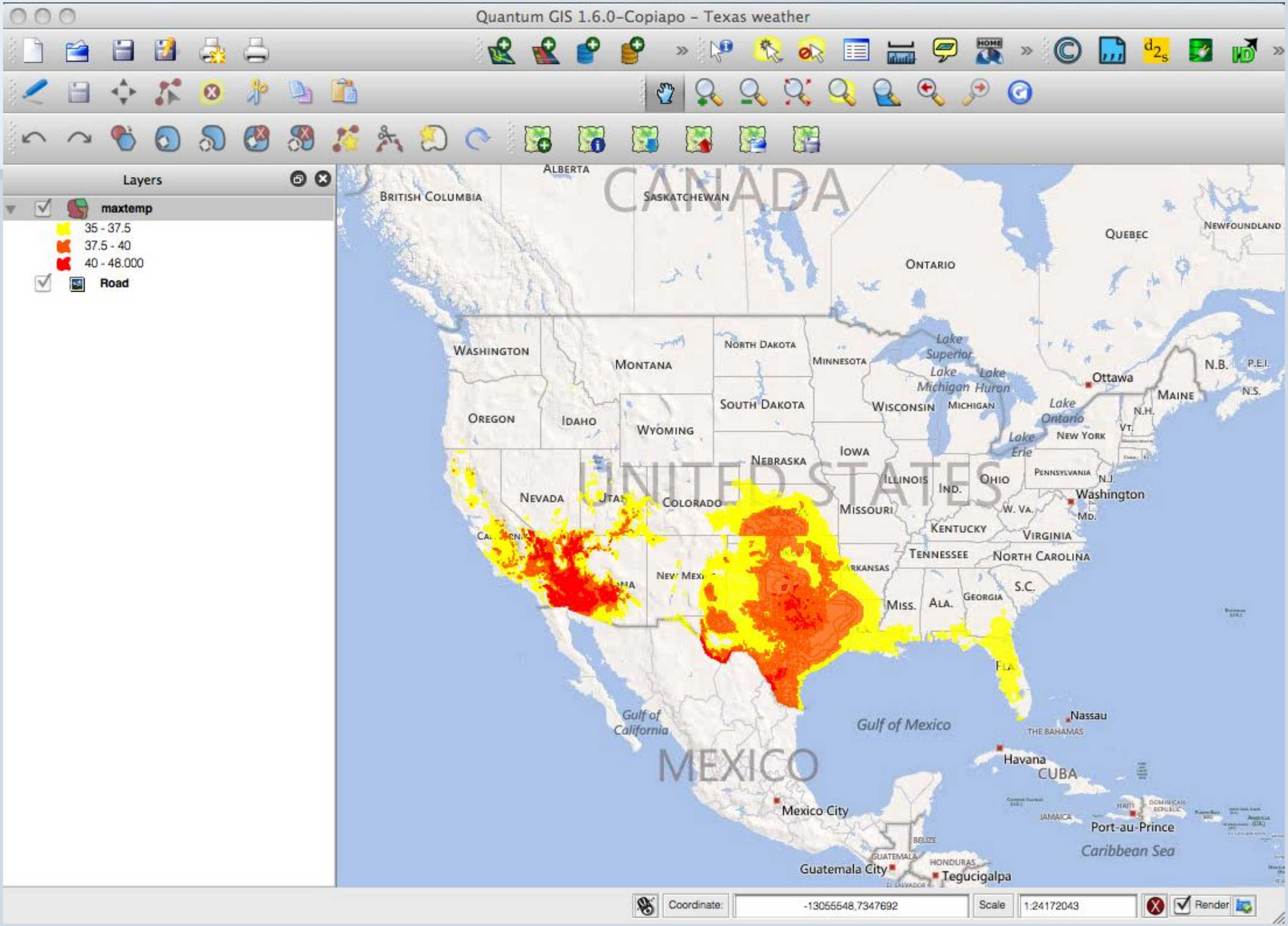
Media
BBC

Series
Bad science

Related

18 Apr 2007
Aspirin linked to lower risk of bowel and prostate cancer

Putting data onto maps



Quantum GIS 1.6.0-Copiapo – Texas weather

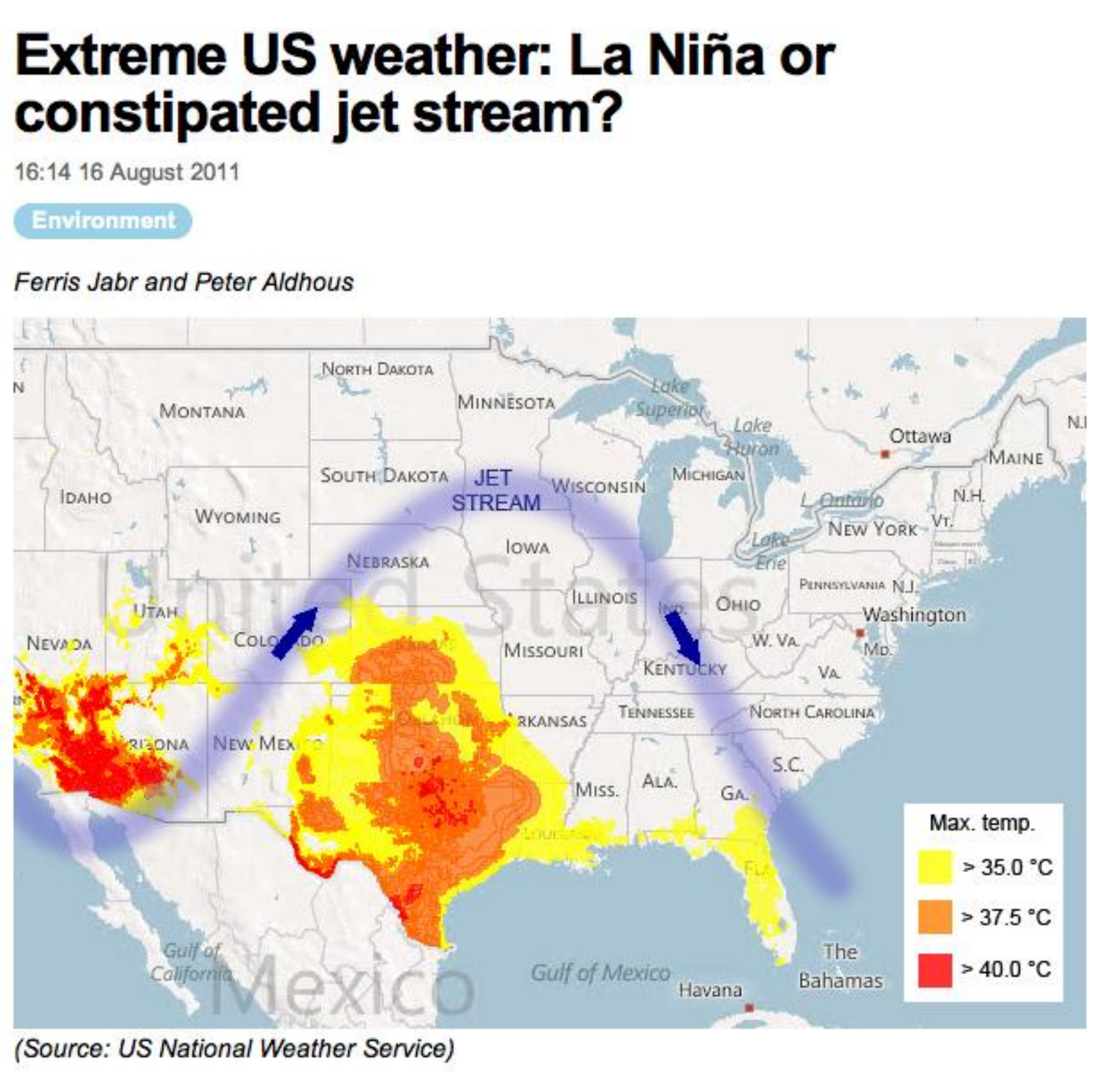
Layers

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

Coordinate: -13055548,7347692

Scale 1:24172043

☑ Render



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

Free GIS software

Quantum GIS

MapWindow

Other free mapping tools

Google Maps

Google Earth

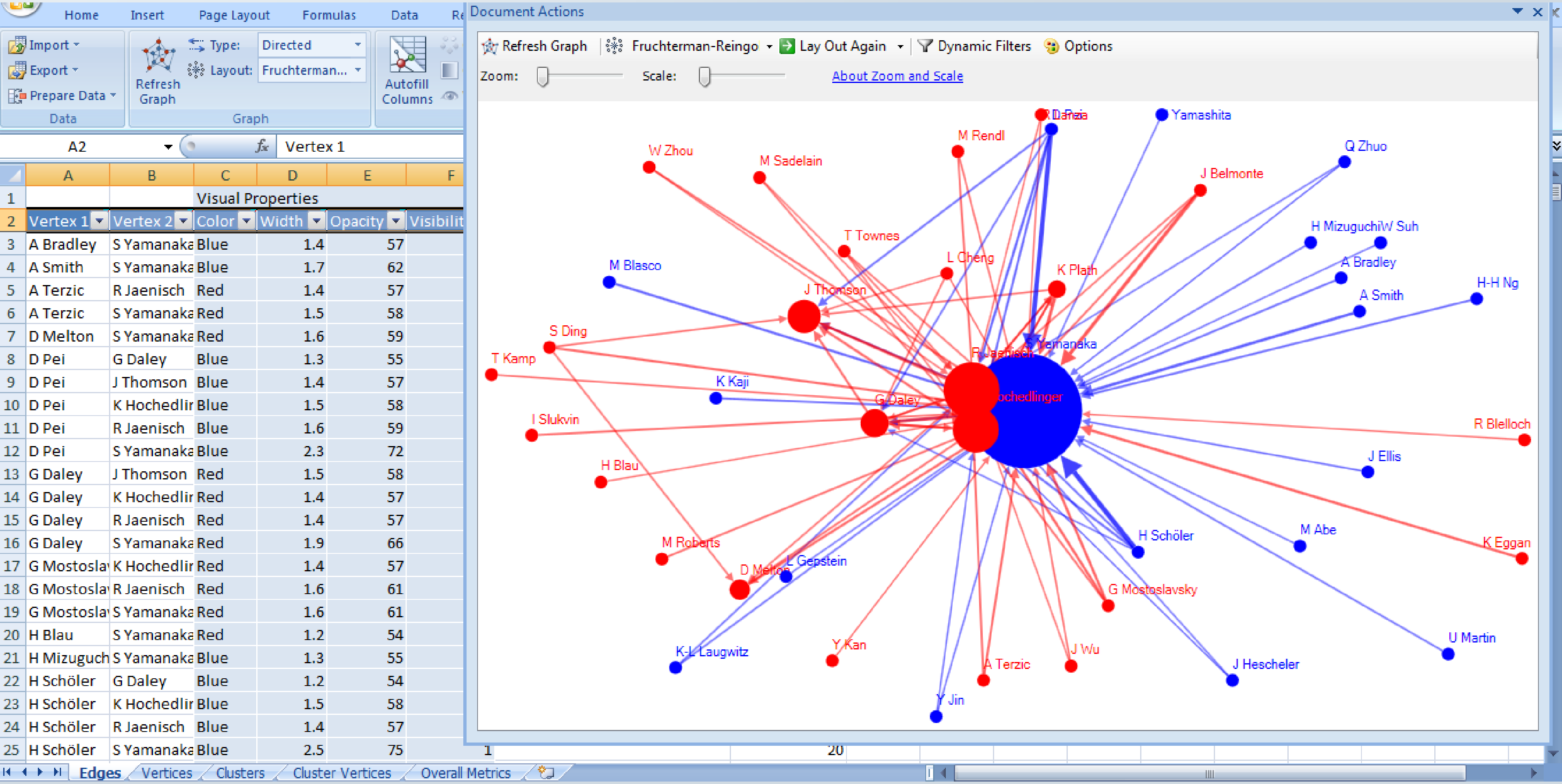
Google Fusion Tables

TileMill

Indiemapper

Geocommons

Network analysis



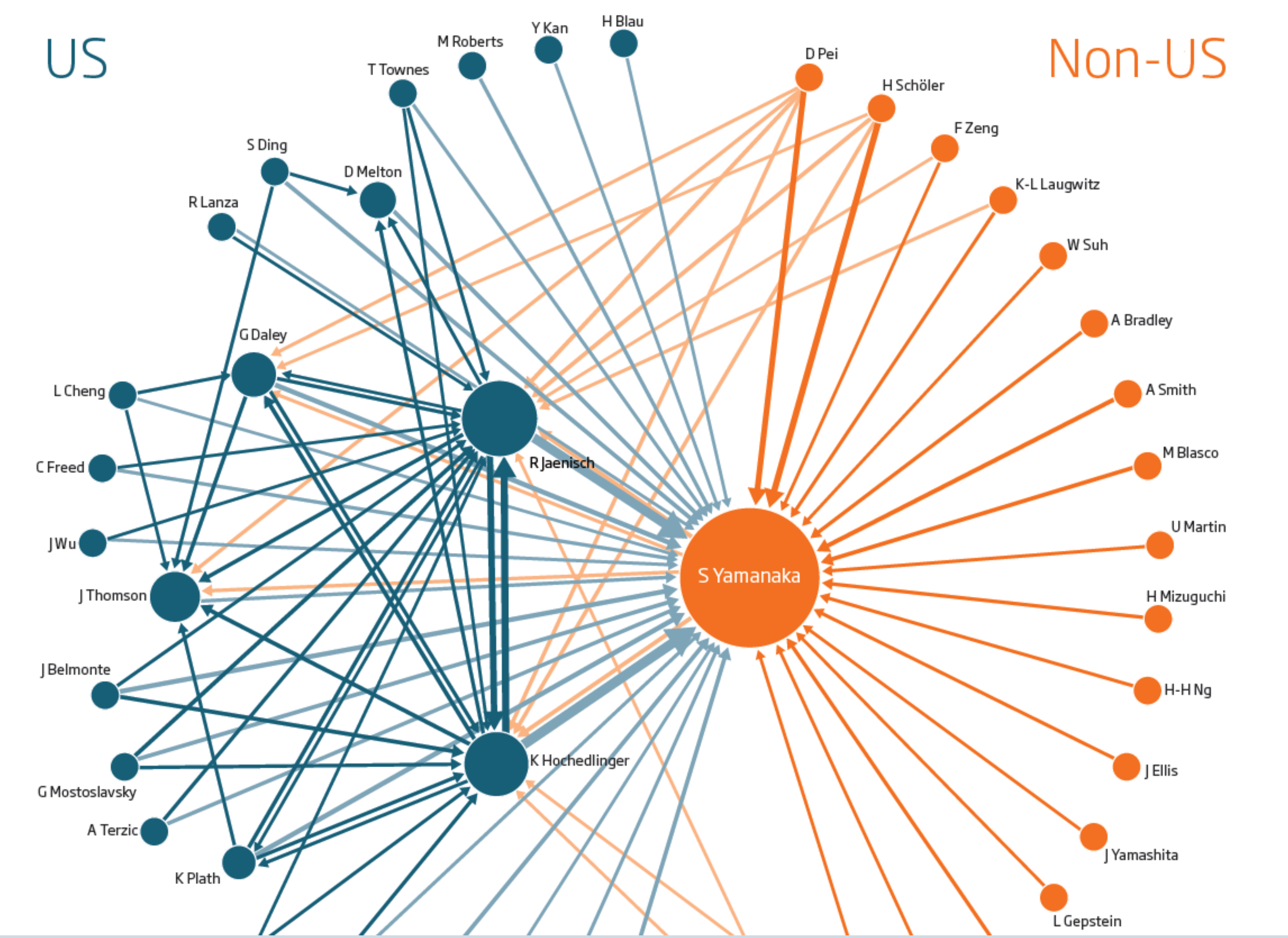
Type: Directed

Layout: Fruchterman...

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	

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

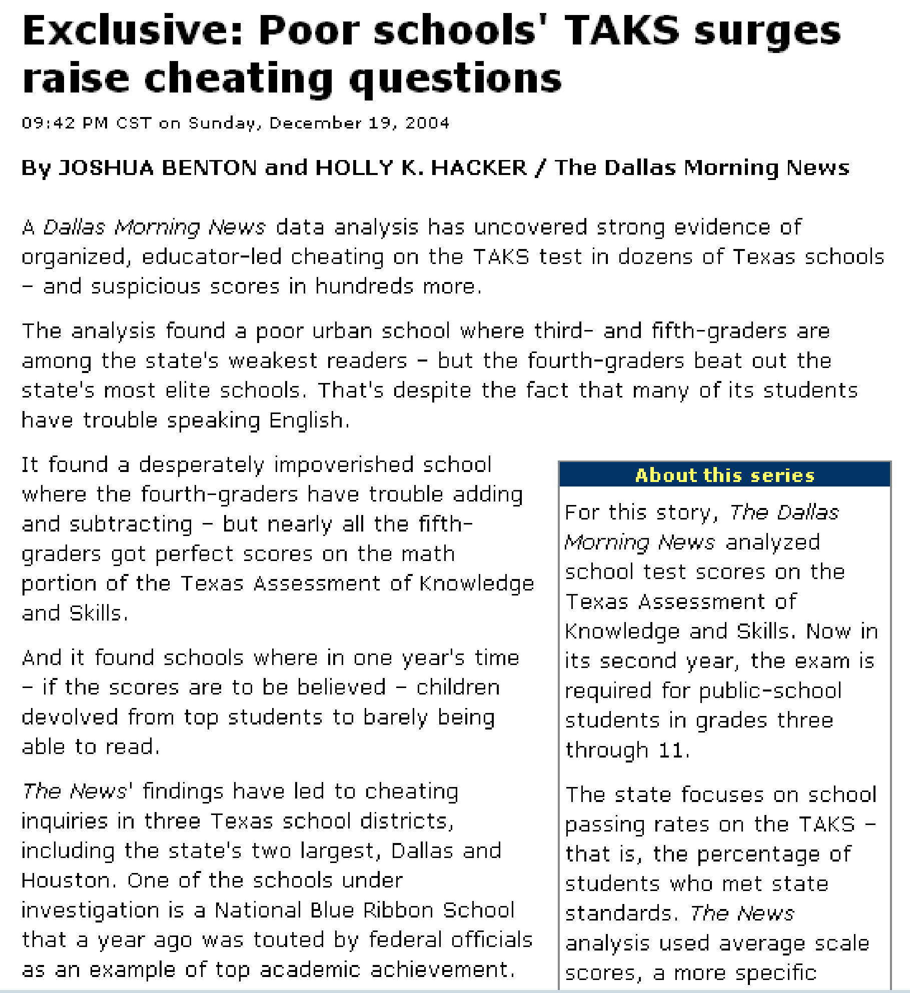
NodeXL (free, extension to Excel 2007/2010)

Gephi (free)

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

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

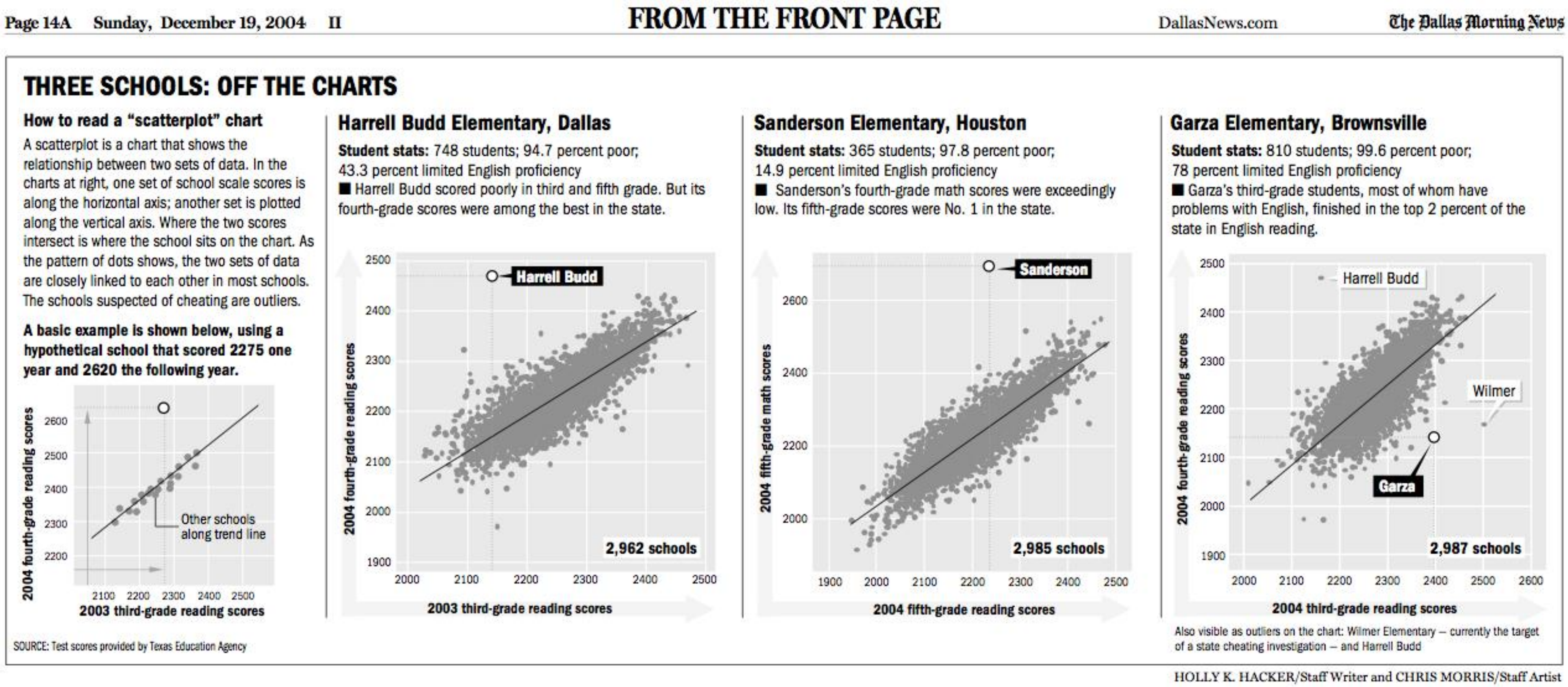
For this story, *The Dallas Morning News* analyzed school test scores on the Texas Assessment of Knowledge and Skills. Now in its second year, the exam is required for public-school students in grades three through 11.

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.

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

Data journalism tutorials

Spreadsheet tutorial in Excel 2010

Database tutorial in Access 2010

Making a map with Google Fusion Tables

Data visualisation with Tableau Public

Network analysis with NodeXL

Introduction to R for statistics

Data journalism in science

UK Conference of Science Journalists,
25 June 2012

Peter Aldhous,

San Francisco Bureau Chief

peter@peteraldhous.com

NewScientist