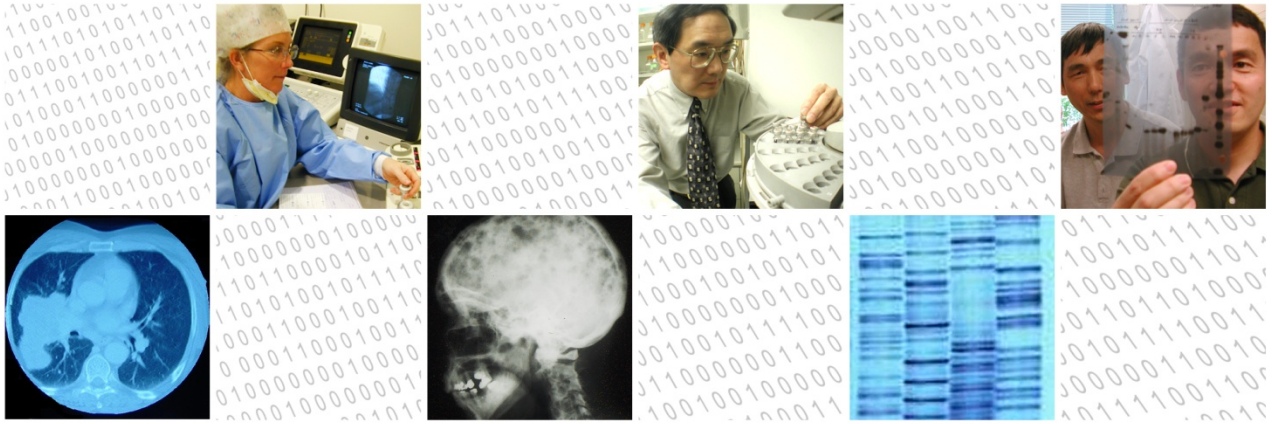
BIOMEDICAL RESEARCH INSTITUTIONS BRIITE INFORMATION TECHNOLOGY EXCHANGE

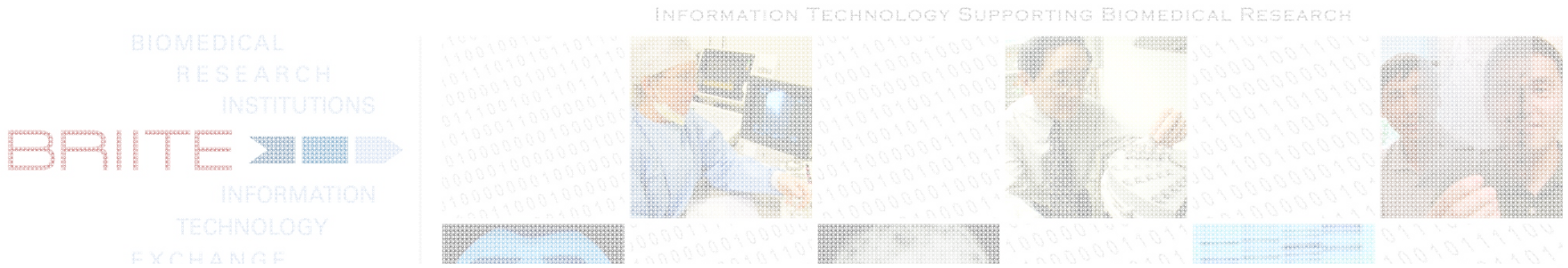
INFORMATION TECHNOLOGY SUPPORTING BIOMEDICAL RESEARCH



Big Data: Yet Another Buzzword or Actual Big Deal?

Robert J. Robbins

rjr8222@gmail.com



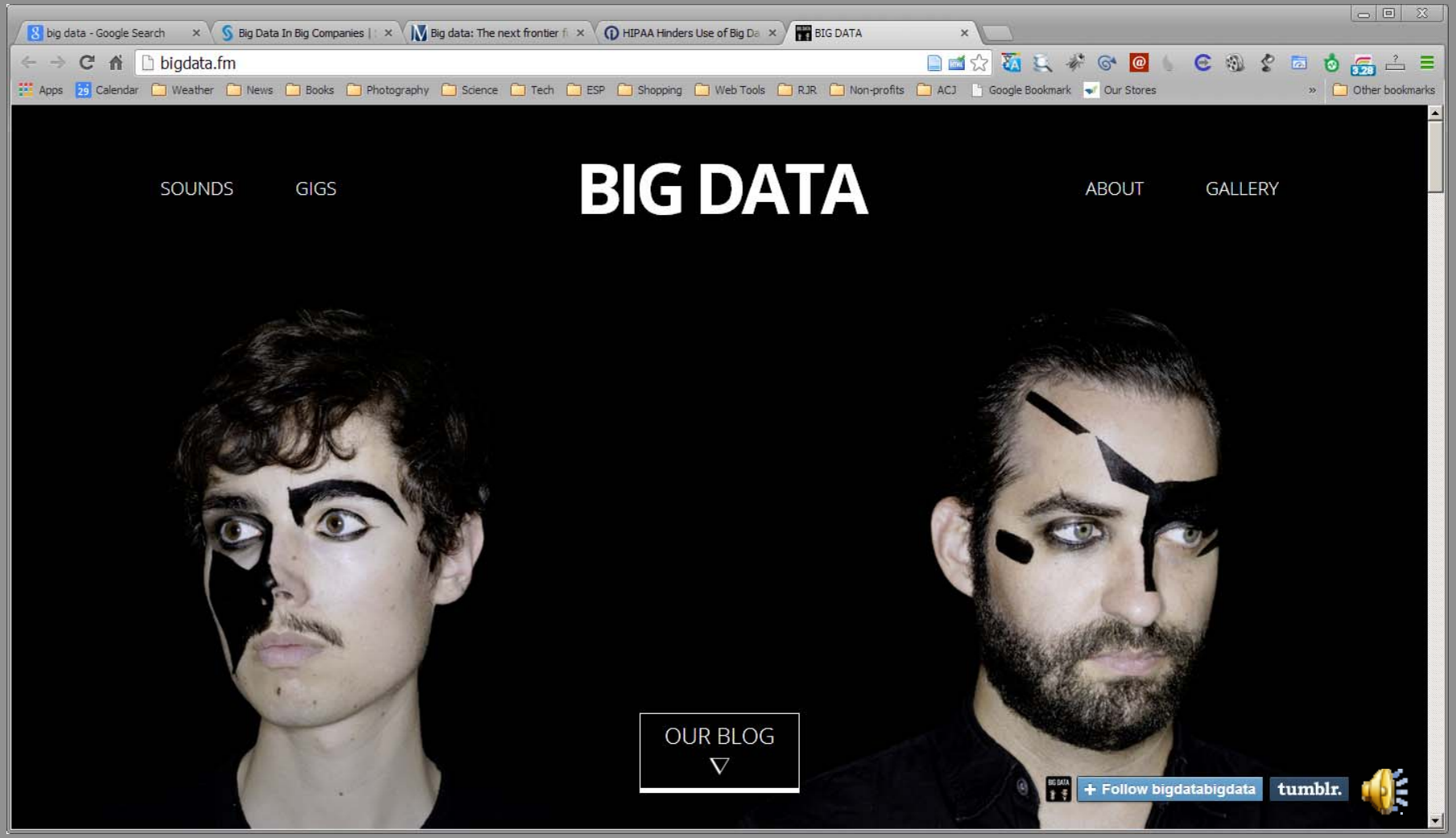
And why should we care?

Robert J. Robbins

rjr8222@gmail.com

What Is BIG DATA

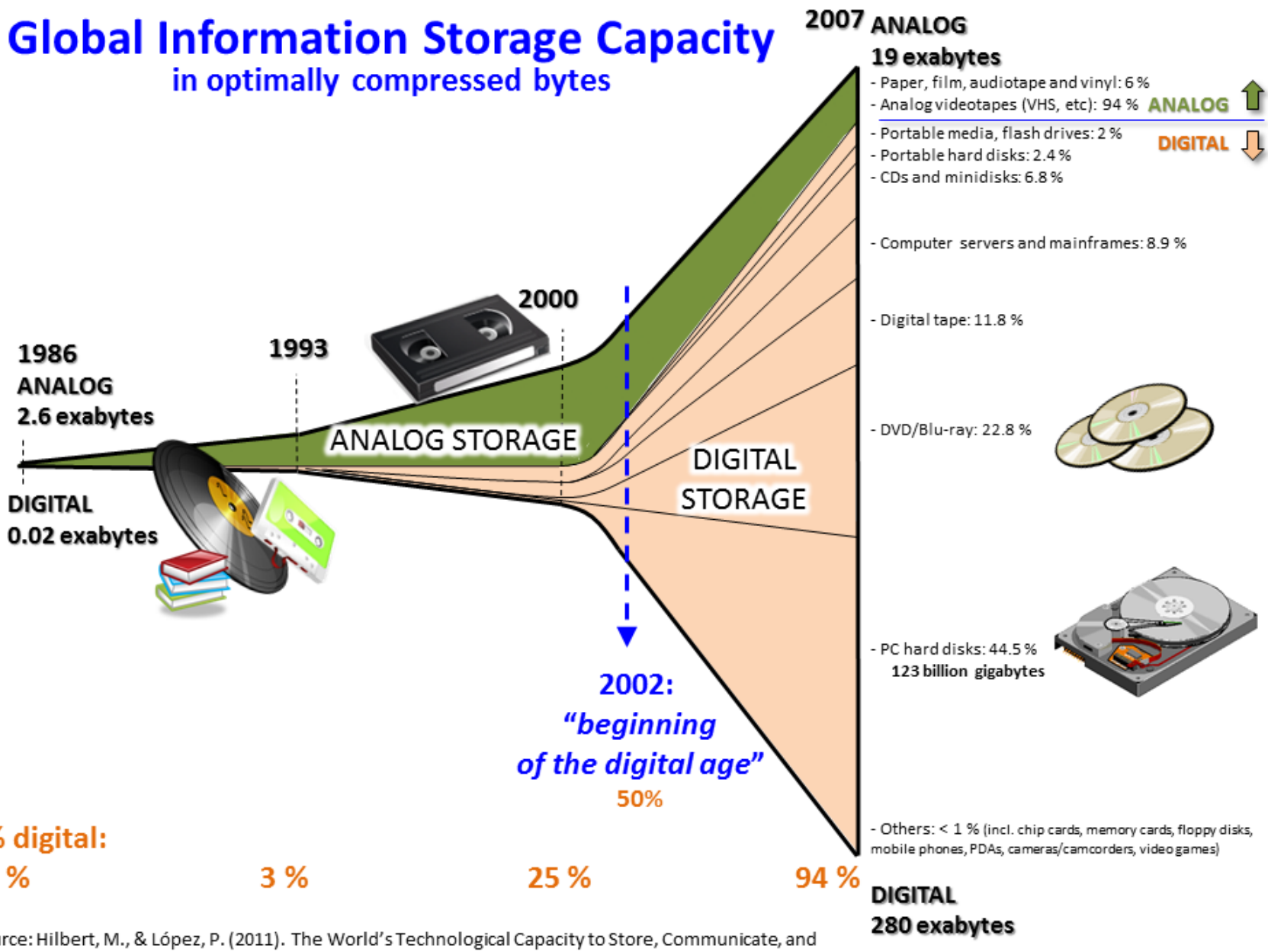
(answer from the interwebs)



Big Data is a paranoid electronic music project from the Internet. The band was formed out of a mutual distrust for technology and The Cloud, despite a growing dependence on them. Their music explores the ever-increasing relationship between man and machine, and more specifically how the internet has reshaped the human experience. || **Alan Wilkis** [producer] +**Daniel Armbruster** [vocals]

Global Information Storage Capacity

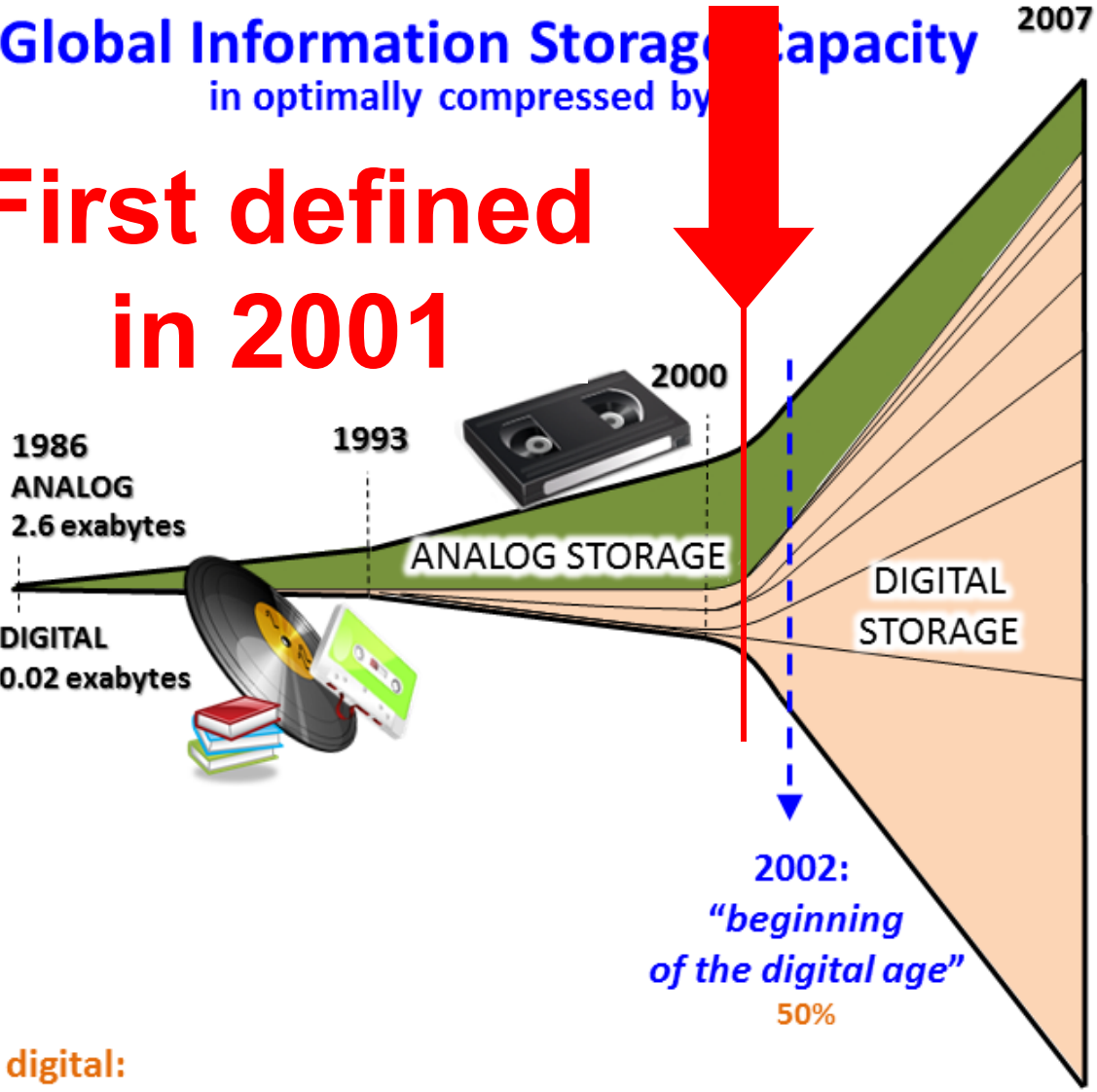
in optimally compressed bytes



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Global Information Storage Capacity
in optimally compressed by

First defined in 2001



2007 ANALOG
19 exabytes

- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 % **ANALOG**

DIGITAL

- Portable media, flash drives: 2 %
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %
- DVD/Blu-ray: 22.8 %

- PC hard disks: 44.5 %
  123 billion gigabytes

- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)

DIGITAL
280 exabytes

Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Original Definition

Application Delivery Strategies

META Group

Date: 6 February 2001

File: 949

Author: Doug Laney

3D Data Management: Controlling Data Volume, Velocity, and Variety. Current business conditions and mediums are pushing traditional data management principles to their limits, giving rise to novel, more formalized approaches.

META Trend: During 2001/02, leading enterprises will increasingly use a centralized data warehouse to define a common business vocabulary that improves internal and external collaboration. Through 2003/04, data quality and integration woes will be tempered by data profiling technologies (for generating metadata, consolidated schemas, and integration logic) and information logistics agents. By 2005/06, data, document, and knowledge management will coalesce, driven by schema-agnostic indexing strategies and portal maturity.

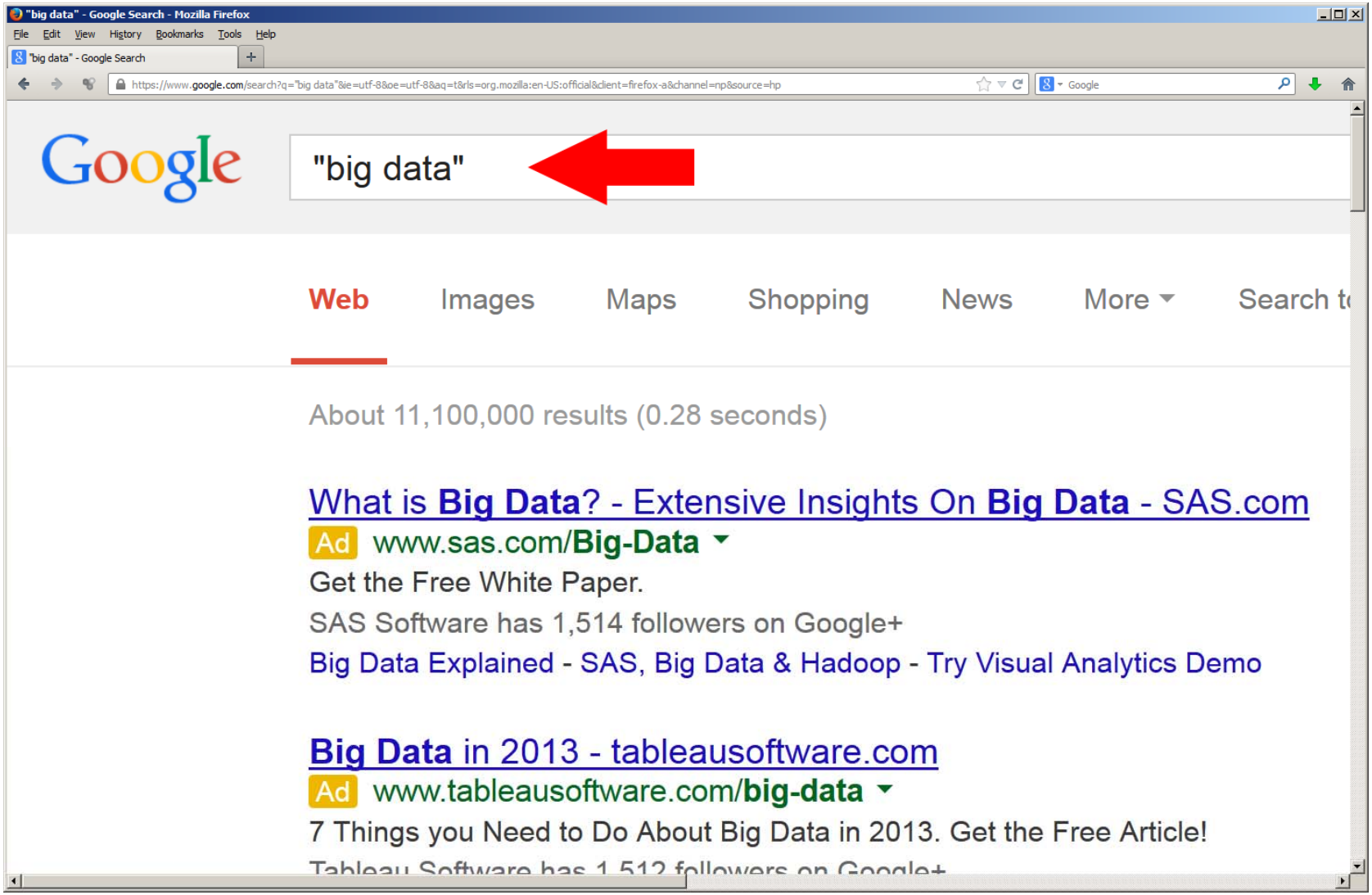
The effect of the e-commerce surge, a rise in merger/acquisition activity, increased collaboration, and the drive for harnessing information as a competitive catalyst is driving enterprises to higher levels of consciousness about how data is managed at its most basic tional, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: volumes, velocity, and variety. In 2001/02

Original Definition

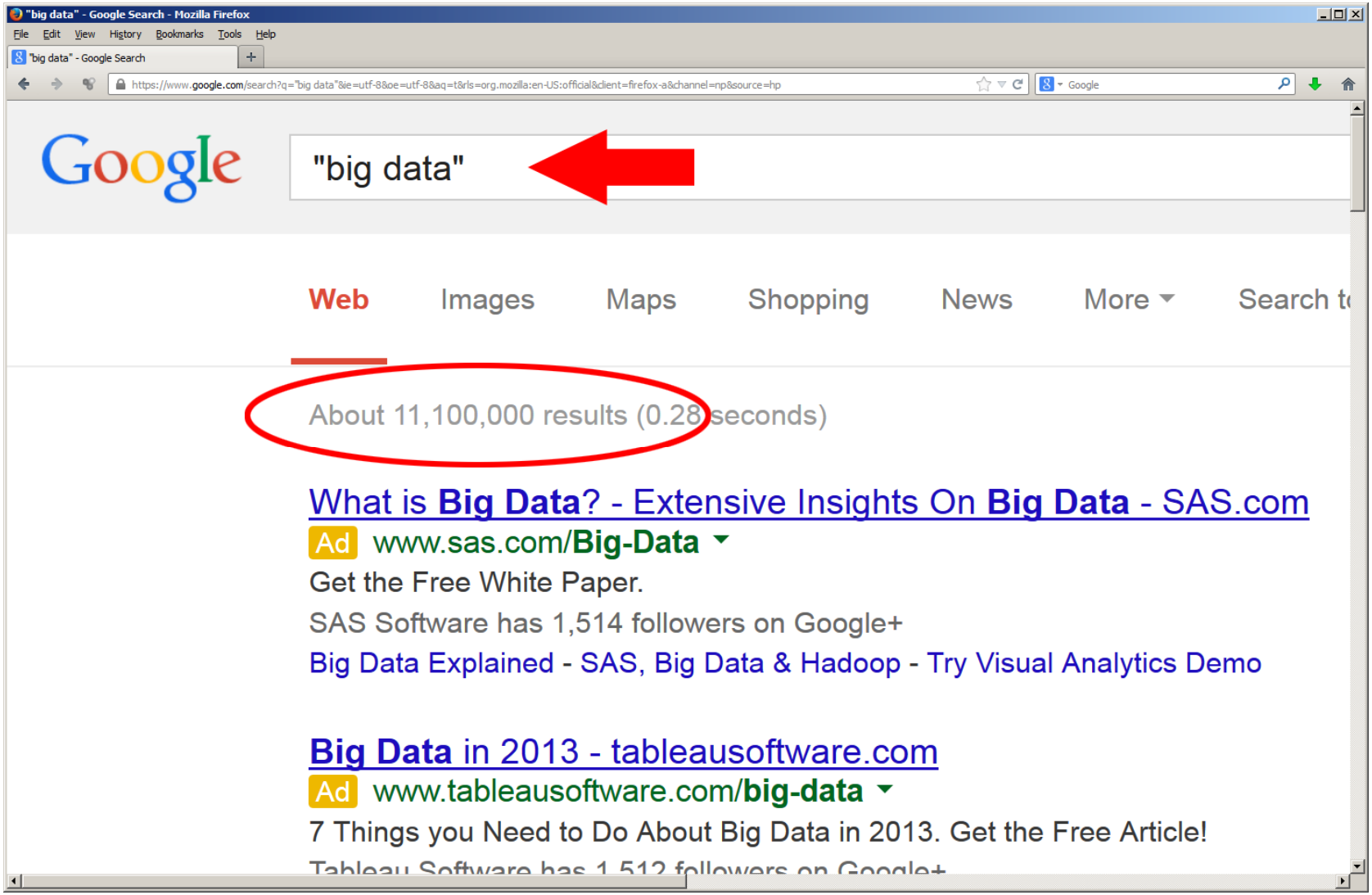
Volume:	More data than extant systems can handle.
Velocity:	Data arriving at a rapid and accelerating pace.
Variety:	Data in a multitude of formats, from a multitude of sources, all (somewhat) relevant to the question at hand.

Buzzword

Proof: Google



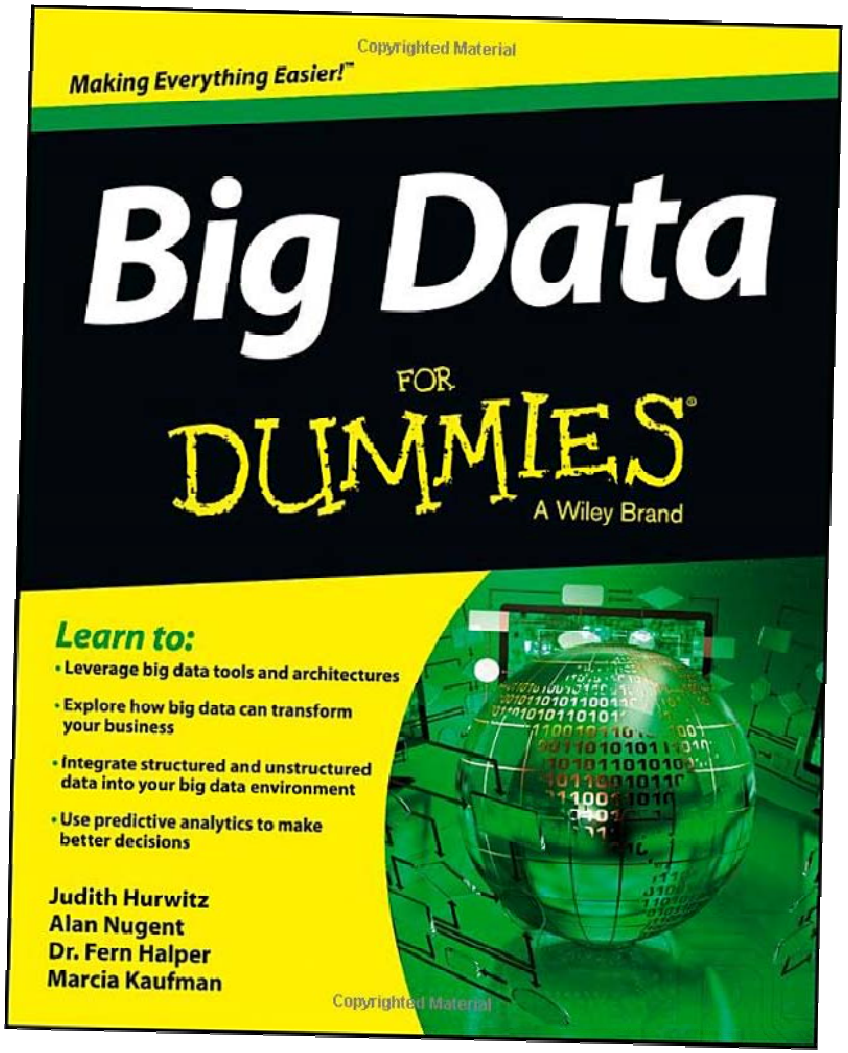
Proof: Google



Proof: Amazon

Proof: Amazon

Proof: Everybody Needs to Know



Proof: Magazine Covers

Proof: Magazine Covers

Proof: Magazine Covers

Proof: Magazine Covers

Proof: Magazine Covers

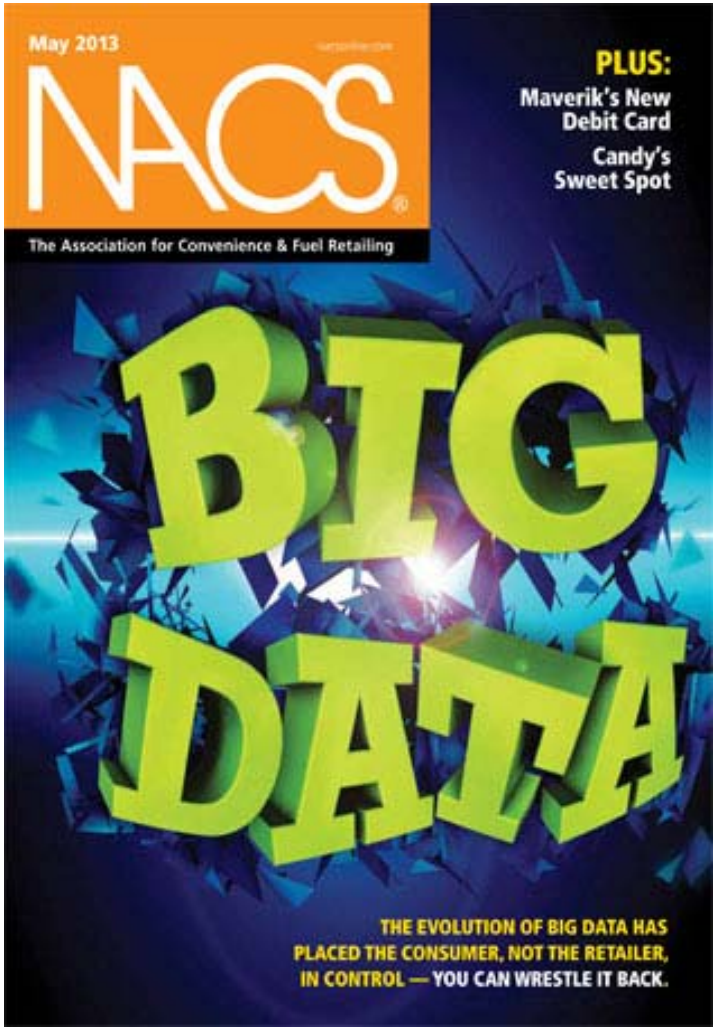
Proof: Magazine Covers

Proof: Magazine Covers

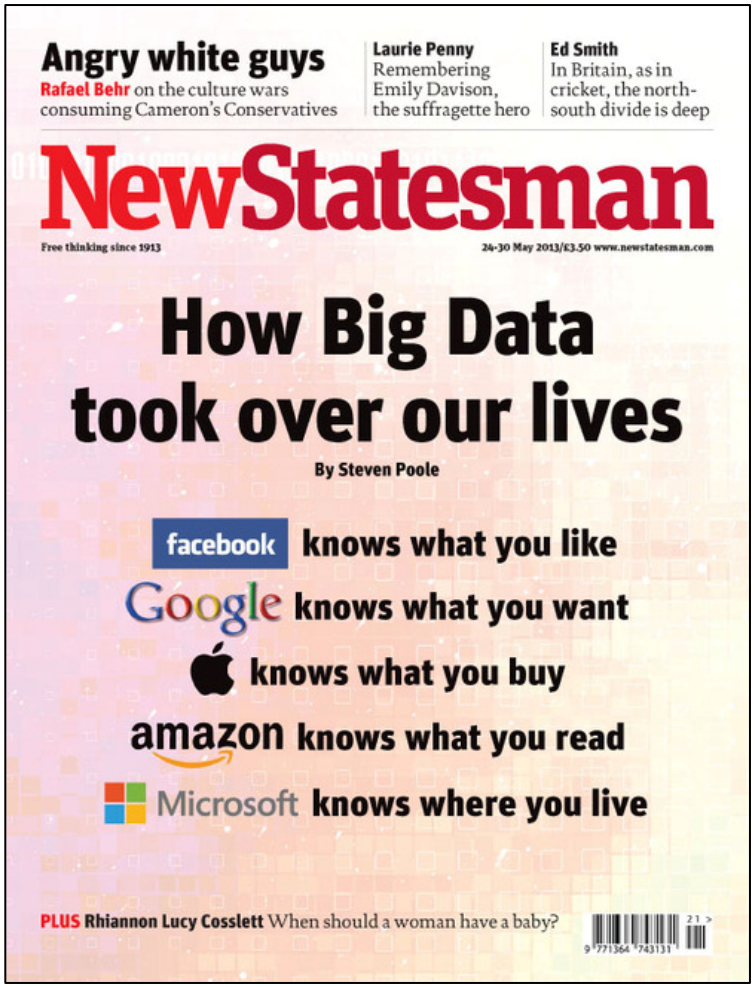
Proof: Magazine Covers

Proof: Magazine Covers

Proof: Magazine Covers



Proof: Magazine Covers



Proof: Magazine Covers

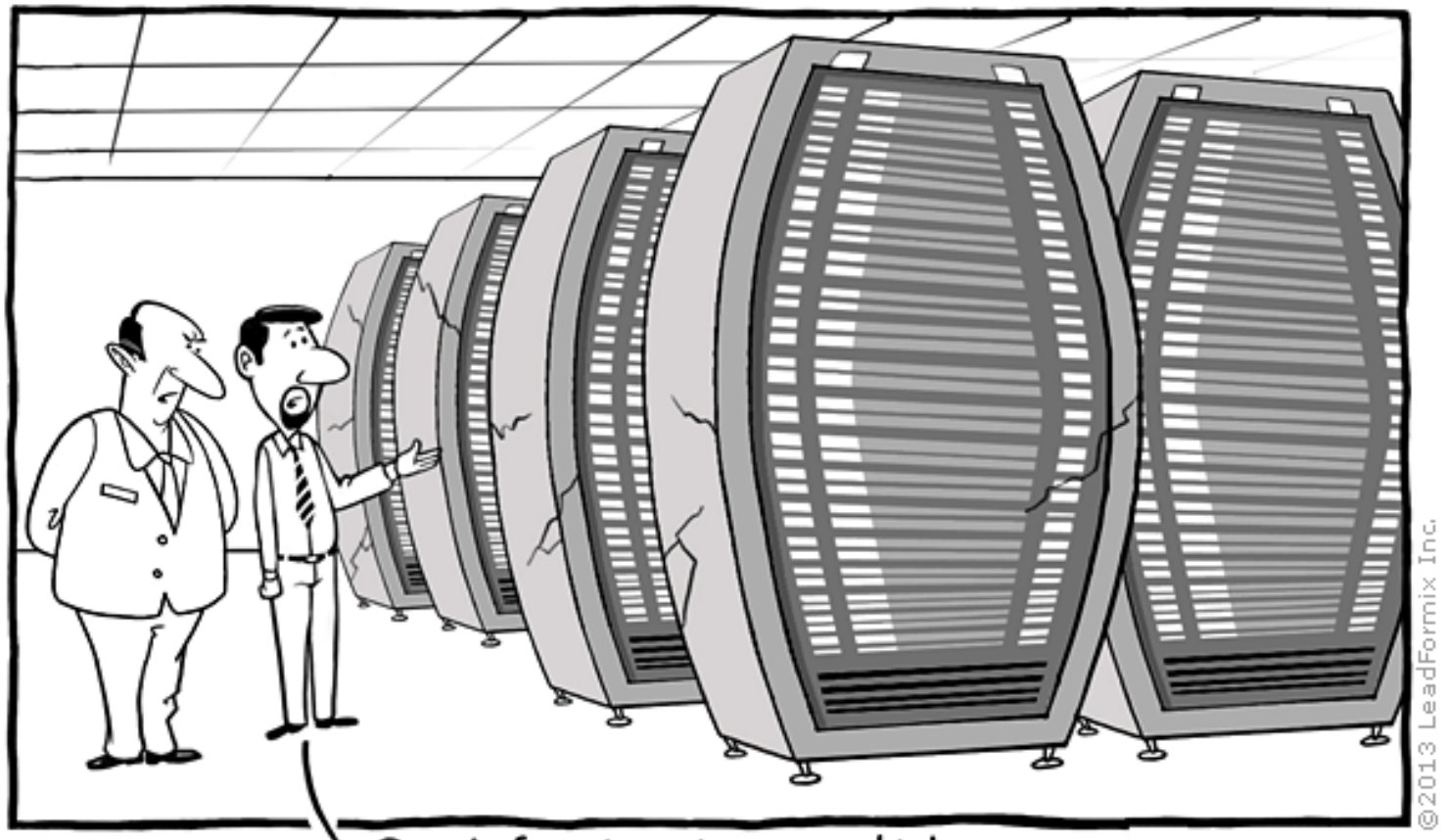
Proof: Cartoons

"Why Gramma, what big data you have!"

Proof: Cartoons

"Hey wait a minute! My Big Data analysis says I'm one gift short..."

Proof: Cartoons



Proof: Cartoons

"I think you'll find that mine is bigger..."

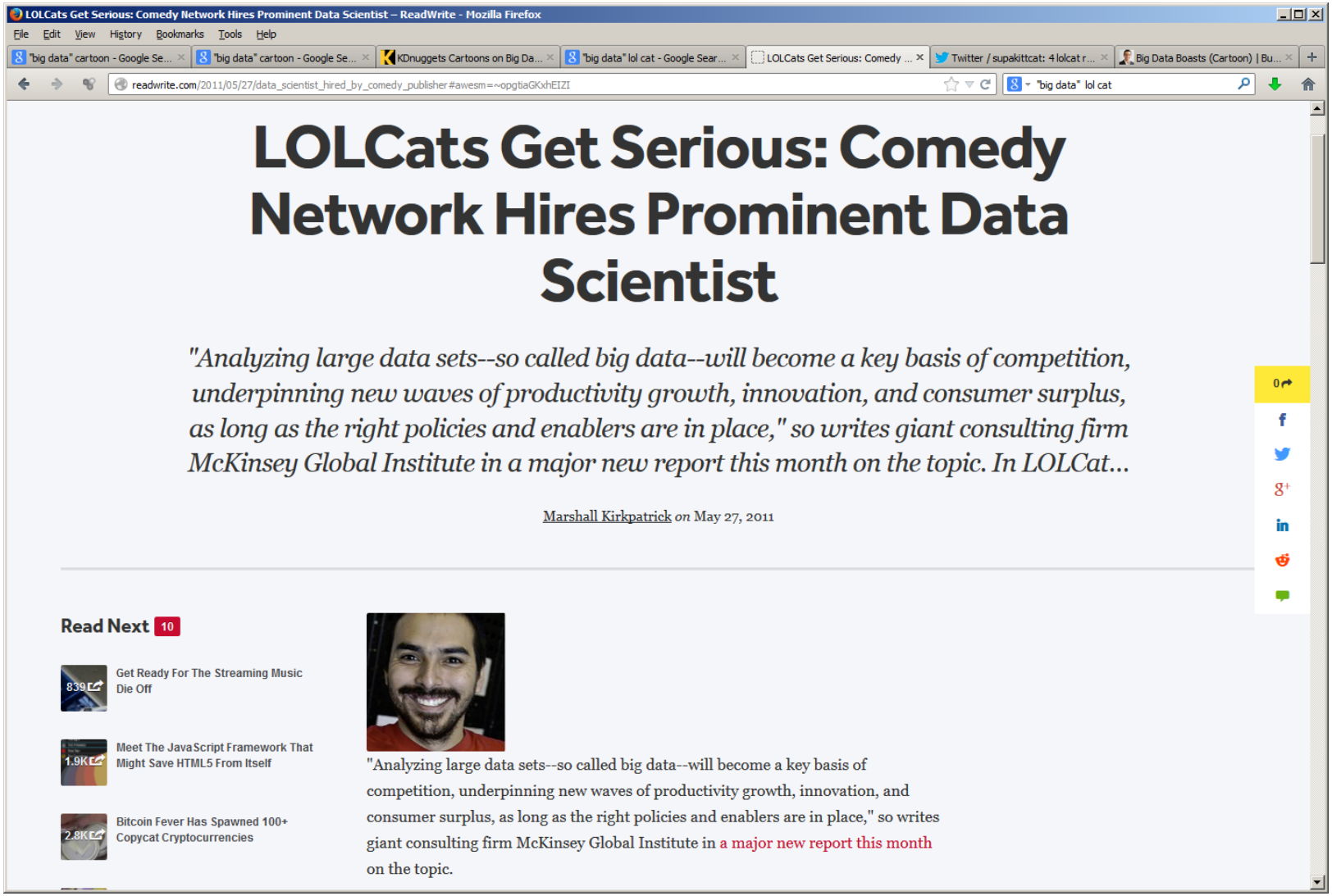
Proof: Cartoons

Proof: Cartoons

Ultimate Proof: Dilbert

Proof: Even LOLCats

Not Only Buzzword Status



Big Deal

Big Deal

Truly humungous analyses

Big Deal

Truly humungous analyses

Datafication of reality

Big Deal

Truly humungous analyses

Datafication of reality

When n → all

Big Deal

Truly humungous analyses

Datafication of reality

When n → all

Analytical issues

Big Deal

Truly humungous analyses

Datafication of reality

When n → all

- **Analytical issues**
- **Policy Issues**

New Insights

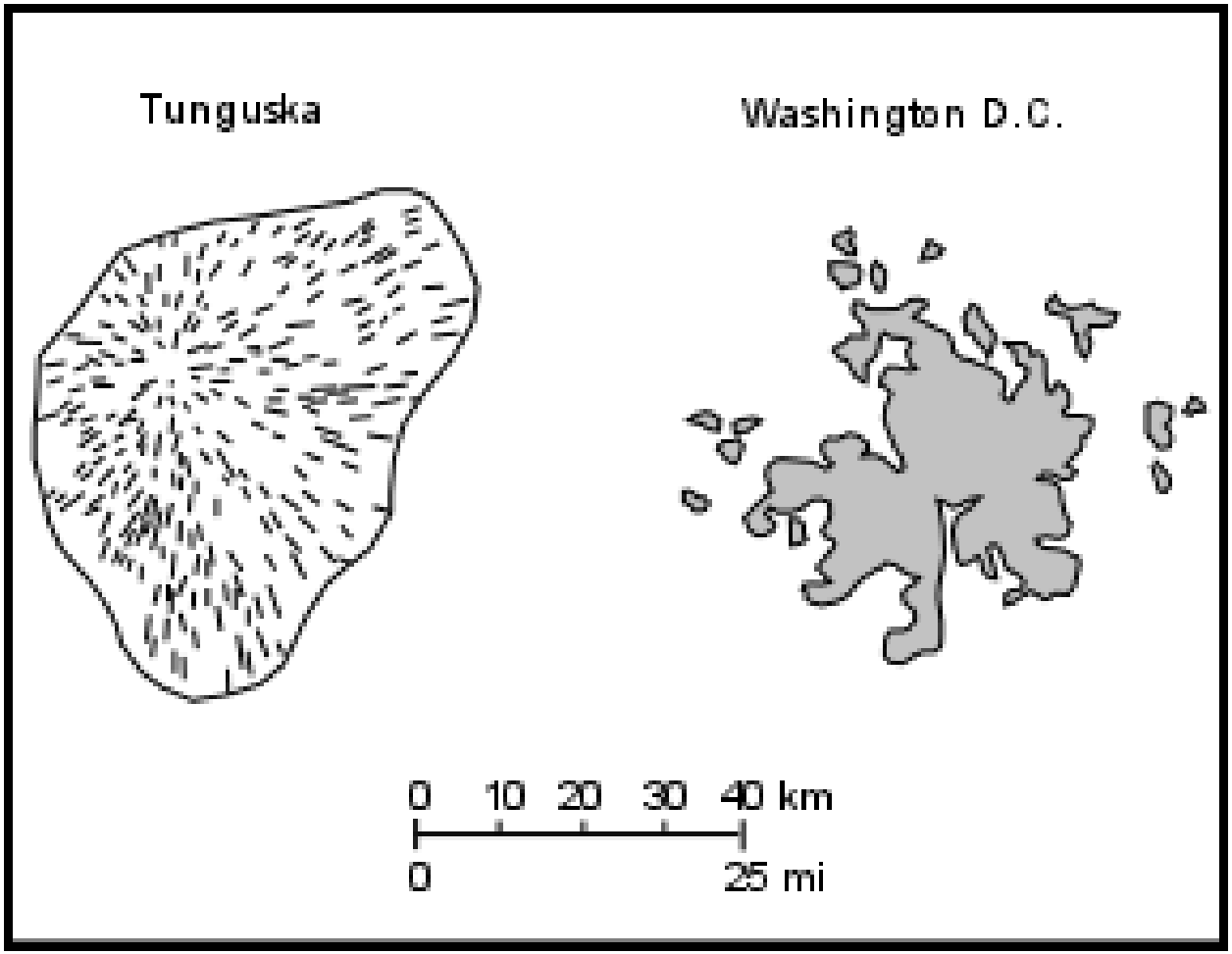
Datafication
of
Reality

Echos in the Datasphere

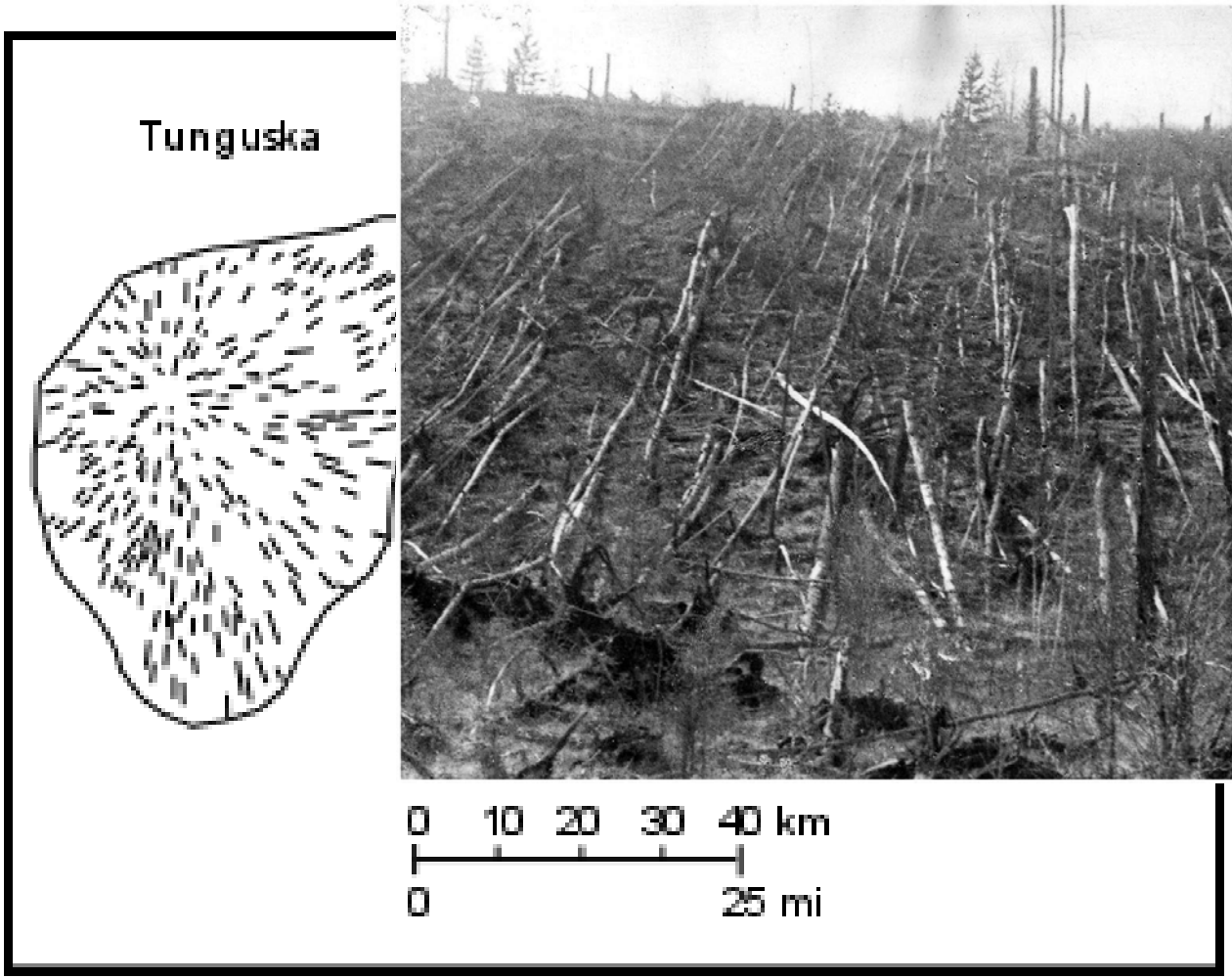
Echos in the Datasphere

Echos in the Datasphere

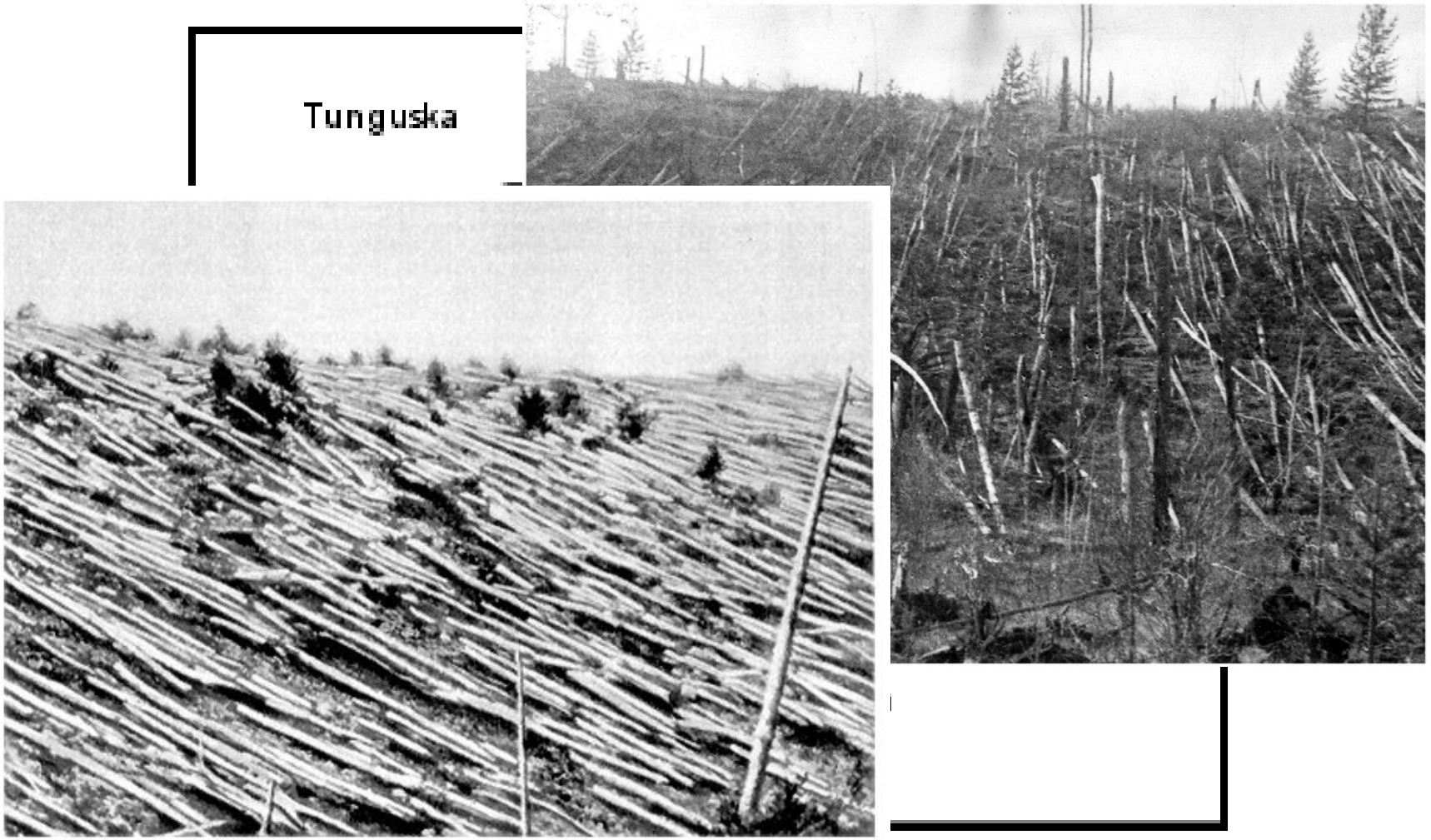
Tunguska Effects



Tunguska Effects



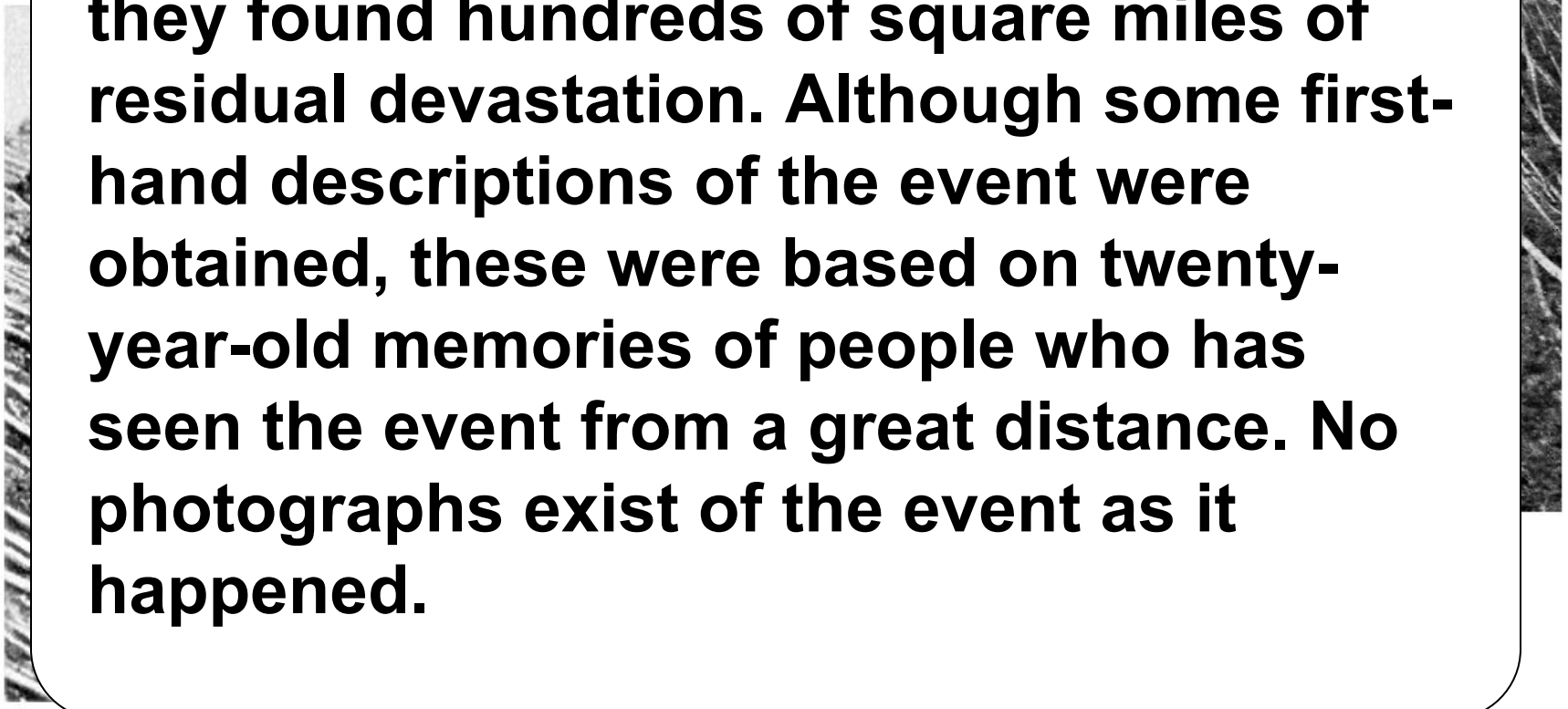
Tunguska Effects



Tunguska Effects

Tunguska Effects

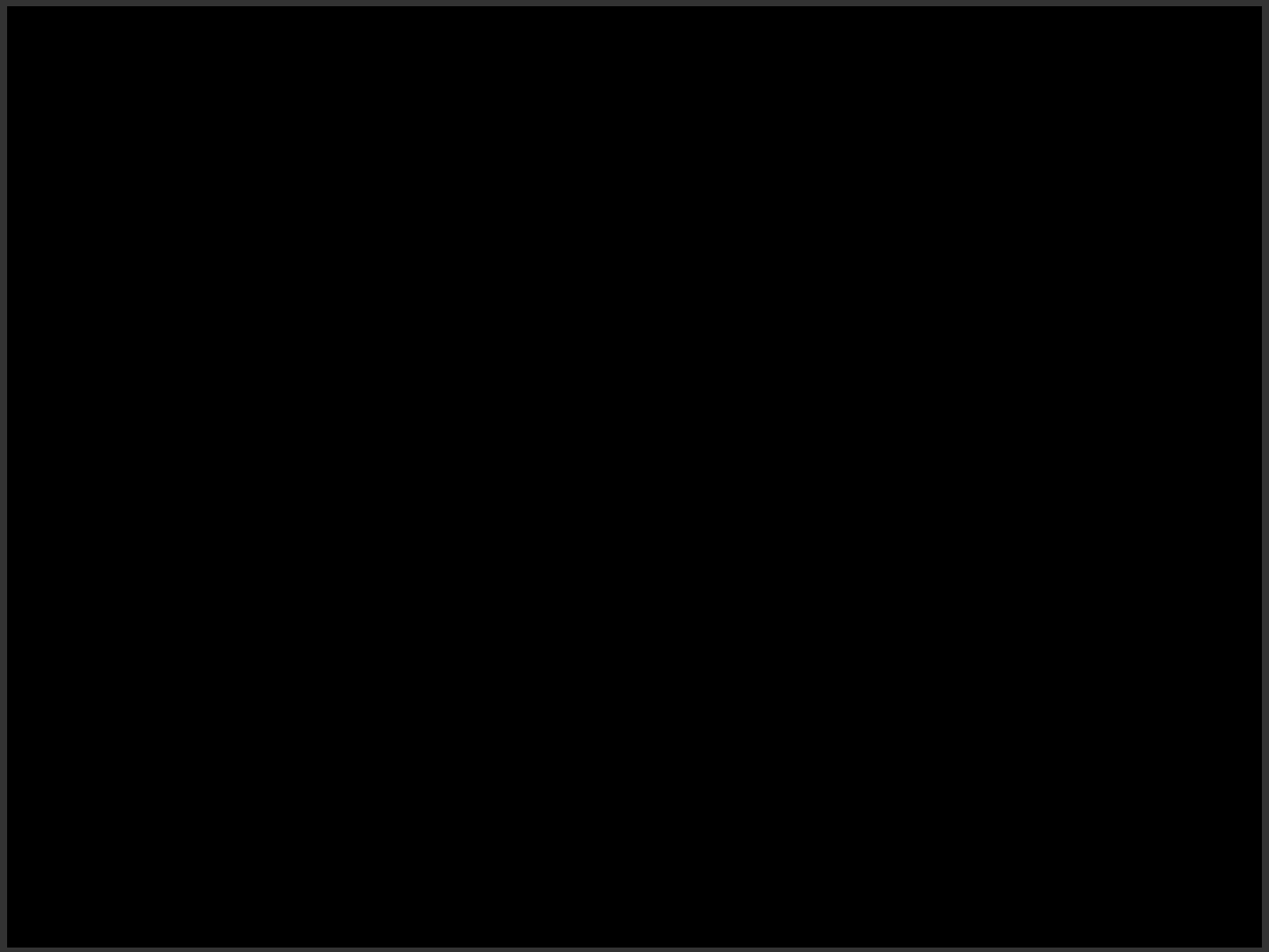
When the first scientific expedition reached the site, 20 years after the event, they found hundreds of square miles of residual devastation. Although some first-hand descriptions of the event were obtained, these were based on twenty-year-old memories of people who has seen the event from a great distance. No photographs exist of the event as it happened.



Echos in the Datasphere

Echos in the Datasphere

Echos in the Datasphere



Chelyabinsk, Russia — 15 Feb 2013

Pretty Exciting, huh?

Echos in the Datasphere

...is to ... the seemingly mundane, yet ... of civil registration systems higher on the agenda of organizations such as the G20. ■

Eyes and ears

Two explosions last week demonstrated the importance of global monitoring.

On 15 February, the town of Chelyabinsk in the Russian Ural Mountains had an unexpected visitor. A meteor streaked high above the city, briefly blinded commuters and then shattered thousands of windows with a series of ear-splitting explosions. The event was recorded on mobile phones and car-dashboard cameras across the region, and YouTube soon filled with Hollywood-style disaster videos of the fireball, replete with some very colourful Russian commentary.

Local residents were not the only ones to record the blast. More than a dozen monitoring stations around the globe captured the ultra-low-frequency infrasound signal of the meteorite as it broke up in the atmosphere. The stations are part of a much larger network of sensors

21 FEBRUARY 2013 | VOL 494 | NATURE | 281

...ublishers Limited. All rights reserved

Echos in the Datasphere

for the spliceosome, yet group II splicing and spliceosomal splicing involve the same two phosphoryl-transfer reactions and generate the same reaction intermediates and products. The crystal structure of a group II intron[6] revealed an active site with two catalytic metal ions coordinated to phosphate groups in the domain V helix of the RNA that are spatially and functionally equivalent to the phosphates that Fica *et al.* show are coordinated between the U6 snRNA and the pre-mRNA. The parallels between these molecules even extend to the stereochemistry of the oxygen atoms that serve as ligands for the two metal ions. Although it has long been known that the hairpin-loop structures of U6 and domain V are functionally similar, this work demonstrates the chemical equivalence between these two systems in their metal-ion coordination.

There are more than 100 proteins in the spliceosome, so what role do they have if RNA is the catalyst of pre-mRNA splicing? Some proteins are retained throughout the splicing process, whereas others associate

SOLAR SYSTEM

Russian skyfall

The recent entry of a 20-metre-wide celestial rock into Earth's atmosphere offered both a spectacular show and a source of invaluable data that advance our understanding of high-velocity impacts. SEE LETTERS P.235 & P.238

NATALIA ARTEMIEVA

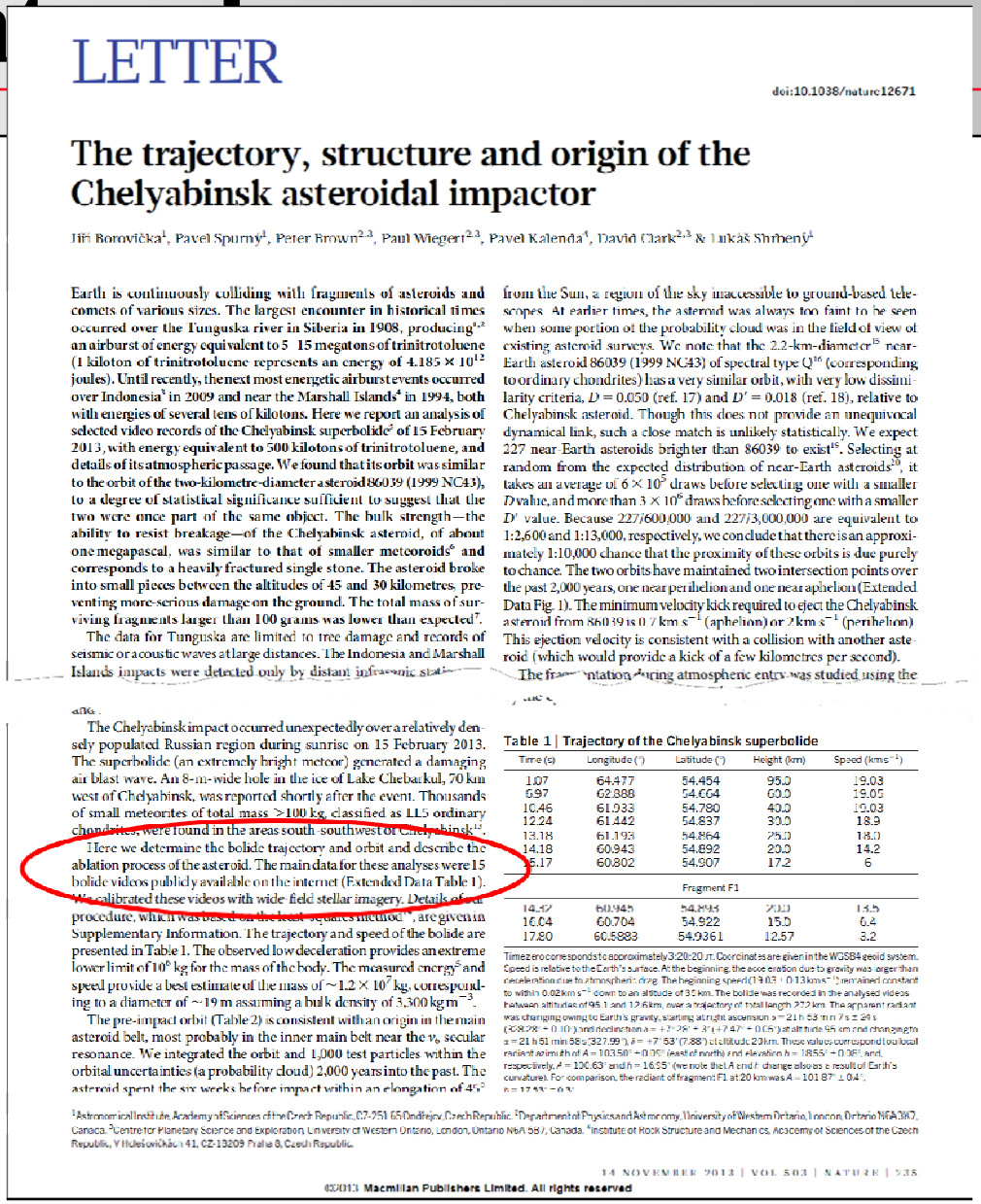
Two papers in this issue focus on the scientific reconstruction of the asteroid-impact event that was observed on 15 February 2013 in Chelyabinsk, Russia — the largest impact event on Earth since the Tunguska blast of 1908. Borovička *et al.*[1] (page 235) extract substantial scientific information from several low-quality video recordings. Meanwhile, by analysing infrasound airwaves and the brightness of the light flash generated by the impact, as well as the damage caused to Earth's surface, Brown *et al.*[2] (page 238) estimate the total energy of the event to have been 400–600 kilotons of trinitrotoluene (1 kt of TNT is equivalent to 4.185×10^{12} joules of energy), and say that such events occur more often than previously thought.

There are three fundamental physical processes involved as meteoroids — asteroids' smaller counterparts — enter the atmosphere. First, atmospheric drag decelerates meteoroids from their initial velocity of about

Echos in the Da

A major paper in *Nature* reported an analysis of the Chelyabinsk impactor.

The primary source data were fifteen YouTube videos…



LETTER

doi:10.1038/nature12671

The trajectory, structure and origin of the Chelyabinsk asteroidal impactor

Jiří Borovička[1], Pavel Spurný[1], Peter Brown[2,3], Paul Wiegert[2,3], Pavel Kalenda[4], David Clark[2,3] & Lukáš Shrbený[1]

Earth is continuously colliding with fragments of asteroids and comets of various sizes. The largest encounter in historical times occurred over the Tunguska river in Siberia in 1908, producing[1,2] an airburst of energy equivalent to 5–15 megatons of trinitrotoluene (1 kiloton of trinitrotoluene represents an energy of 4.185×10^{12} joules). Until recently, the next most energetic airburst events occurred over Indonesia[3] in 2009 and near the Marshall Islands[4] in 1994, both with energies of several tens of kilotons. Here we report an analysis of selected video records of the Chelyabinsk superbolide[5] of 15 February 2013, with energy equivalent to 500 kilotons of trinitrotoluene, and details of its atmospheric passage. We found that its orbit was similar to the orbit of the two-kilometre-diameter asteroid 86039 (1999 NC43), to a degree of statistical significance sufficient to suggest that the two were once part of the same object. The bulk strength—the ability to resist breakage—of the Chelyabinsk asteroid, of about one megapascal, was similar to that of smaller meteoroids[6] and corresponds to a heavily fractured single stone. The asteroid broke into small pieces between the altitudes of 45 and 30 kilometres, preventing more-serious damage on the ground. The total mass of surviving fragments larger than 100 grams was lower than expected[7].

The data for Tunguska are limited to tree damage and records of seismic or acoustic waves at large distances. The Indonesia and Marshall Islands impacts were detected only by distant infrasonic sta-

from the Sun, a region of the sky inaccessible to ground-based telescopes. At earlier times, the asteroid was always too faint to be seen when some portion of the probability cloud was in the field of view of existing asteroid surveys. We note that the 2.2-km-diameter[15] near-Earth asteroid 86039 (1999 NC43) of spectral type Q[16] (corresponding to ordinary chondrites) has a very similar orbit, with very low dissimilarity criteria, $D = 0.050$ (ref. 17) and $D' = 0.018$ (ref. 18), relative to Chelyabinsk asteroid. Though this does not provide an unequivocal dynamical link, such a close match is unlikely statistically. We expect 227 near-Earth asteroids brighter than 86039 to exist[15]. Selecting at random from the expected distribution of near-Earth asteroids[20], it takes an average of 6×10^5 draws before selecting one with a smaller D value, and more than 3×10^6 draws before selecting one with a smaller D' value. Because 227/600,000 and 227/3,000,000 are equivalent to 1:2,600 and 1:13,000, respectively, we conclude that there is an approximately 1:10,000 chance that the proximity of these orbits is due purely to chance. The two orbits have maintained two intersection points over the past 2,000 years, one near perihelion and one near aphelion (Extended Data Fig. 1). The minimum velocity kick required to eject the Chelyabinsk asteroid from 86039 is 0.7 km s^{-1} (aphelion) or 2 km s^{-1} (perihelion). This ejection velocity is consistent with a collision with another asteroid (which would provide a kick of a few kilometres per second).

The fra[illegible]entation during atmospheric entry was studied using the

ang-

The Chelyabinsk impact occurred unexpectedly over a relatively densely populated Russian region during sunrise on 15 February 2013. The superbolide (an extremely bright meteor) generated a damaging air blast wave. An 8-m-wide hole in the ice of Lake Chebarkul, 70 km west of Chelyabinsk, was reported shortly after the event. Thousands of small meteorites of total mass >100 kg, classified as LL5 ordinary chondrites, were found in the areas south–southwest of Chelyabinsk[12].

Here we determine the bolide trajectory and orbit and describe the ablation process of the asteroid. The main data for these analyses were 15 bolide videos publicly available on the internet (Extended Data Table 1). We calibrated these videos with wide-field stellar imagery. Details of the procedure, which was based on the least-squares method[13], are given in Supplementary Information. The trajectory and speed of the bolide are presented in Table 1. The observed low deceleration provides an extreme lower limit of 10^6 kg for the mass of the body. The measured energy[5] and speed provide a best estimate of the mass of $\sim 1.2 \times 10^7$ kg, corresponding to a diameter of ~19 m assuming a bulk density of 3,300 kg m^{-3}.

The pre-impact orbit (Table 2) is consistent with an origin in the main asteroid belt, most probably in the inner main belt near the ν_6 secular resonance. We integrated the orbit and 1,000 test particles within the orbital uncertainties (a probability cloud) 2,000 years into the past. The asteroid spent the six weeks before impact within an elongation of 45°

Table 1 | Trajectory of the Chelyabinsk superbolide

Time (s)	Longitude (°)	Latitude (°)	Height (km)	Speed (km s^{-1})
1.07	64.477	54.454	95.0	19.03
6.97	62.888	54.664	60.0	19.05
10.46	61.933	54.780	40.0	19.03
12.24	61.442	54.837	30.0	18.9
13.18	61.193	54.864	25.0	18.0
14.18	60.943	54.892	20.0	14.2
15.17	60.802	54.907	17.2	6
		Fragment F1		
14.57	60.945	54.893	20.0	13.5
16.04	60.704	54.922	15.0	6.4
17.80	60.5883	54.9361	12.57	3.2

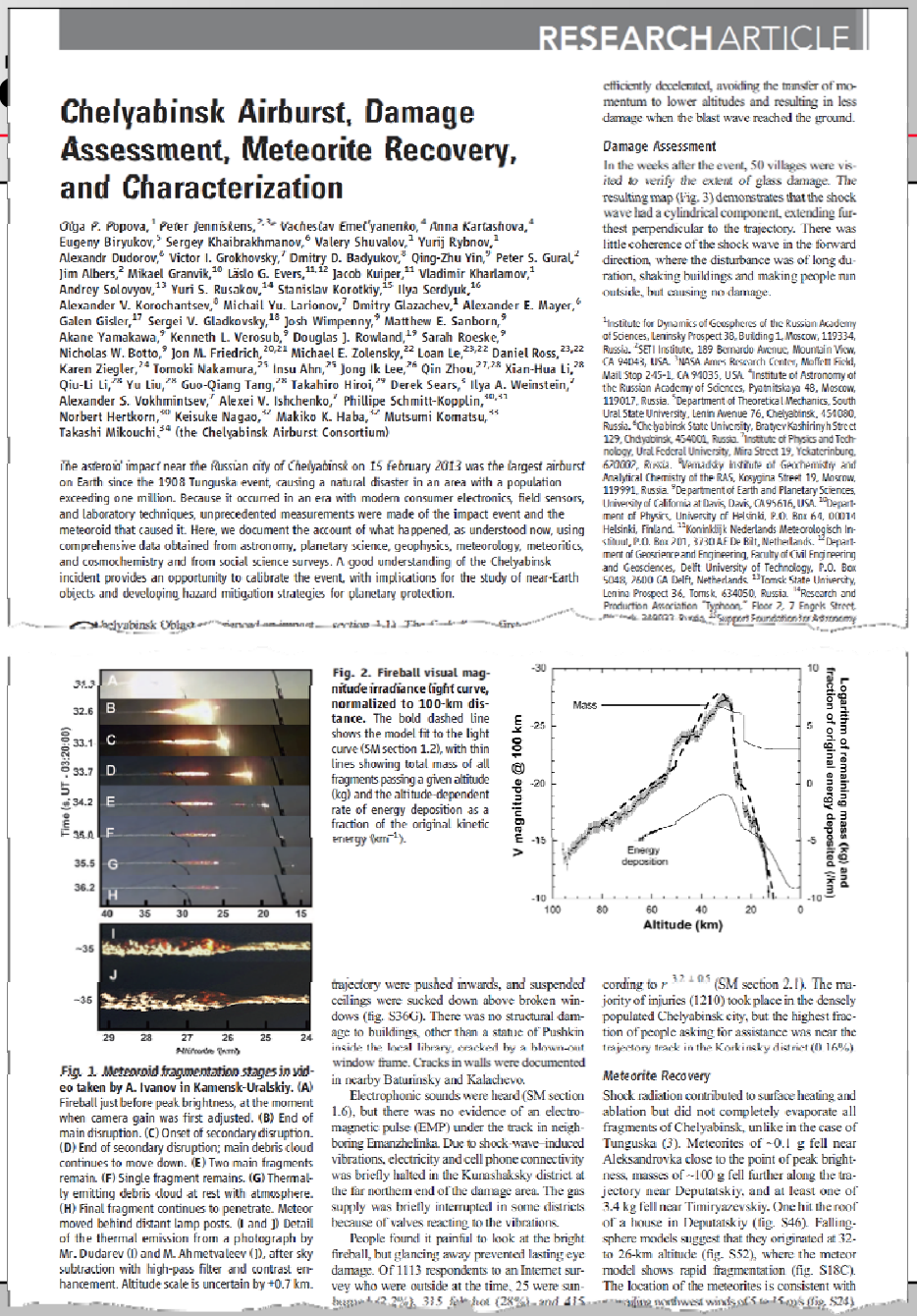
Time zero corresponds to approximately 3:20:20 UT. Coordinates are given in the WGS84 geoid system. Speed is relative to the Earth's surface. At the beginning, the acceleration due to gravity was larger than deceleration due to atmospheric drag. The beginning speed (19.03 ± 0.13 km s^{-1}) remained constant to within 0.02 km s^{-1} down to an altitude of 30 km. The bolide was recorded in the analysed videos between altitudes of 95.1 and 12.6 km, over a trajectory of total length 272 km. The apparent radiant was changing owing to Earth's gravity, starting at right ascension $\alpha = 21$ h 58 min 7 s ± 24 s (329.28° ± 0.10°) and declination $\Delta = +7.28° \pm 3°$ (+7.47° ± 0.05°) at altitude 95 km and changing to $\alpha = 21$ h 51 min 58 s (327.99°), $\delta = +7.53°$ (7.88°) at altitude 20 km. These values correspond to a local radiant azimuth of $A = 103.50° \pm 0.09°$ (east of north) and elevation $h = 18.55° \pm 0.06°$, and, respectively, $A = 100.63°$ and $h = 16.55°$ (we note that A and h change allows a result of Earth's curvature). For comparison, the radiant of fragment F1 at 20 km was $A = 101.87° \pm 0.4°$, $h = 17.33° \pm 0.3°$.

[1]Astronomical Institute, Academy of Sciences of the Czech Republic, CZ-251 65 Ondřejov, Czech Republic. [2]Department of Physics and Astronomy, University of Western Ontario, London, Ontario N6A 3K7, Canada. [3]Centre for Planetary Science and Exploration, University of Western Ontario, London, Ontario N6A 5B7, Canada. [4]Institute of Rock Structure and Mechanics, Academy of Sciences of the Czech Republic, V Holešovičkách 41, CZ-18209 Praha 8, Czech Republic.

Echos in the Data

A major paper in *Science* reported an analysis of the Chelyabinsk impactor.

Some critical data were taken from YouTube videos…



RESEARCH ARTICLE

Chelyabinsk Airburst, Damage Assessment, Meteorite Recovery, and Characterization

Olga P. Popova,[1] Peter Jenniskens,[2,3] Vacheslav Emel'yanenko,[4] Anna Kartashova,[4] Eugeny Biryukov,[5] Sergey Khaibrakhmanov,[6] Valery Shuvalov,[1] Yurij Rybnov,[1] Alexandr Dudorov,[6] Victor I. Grokhovsky,[7] Dmitry D. Badyukov,[8] Qing-Zhu Yin,[9] Peter S. Gural,[2] Jim Albers,[2] Mikael Granvik,[10] Läslo G. Evers,[11,12] Jacob Kuiper,[11] Vladimir Kharlamov,[1] Andrey Solovyov,[13] Yuri S. Rusakov,[14] Stanislav Korotkiy,[15] Ilya Sordyuk,[16] Alexander V. Korochantsev,[8] Michail Yu. Larionov,[7] Dmitry Glazachev,[1] Alexander E. Mayer,[6] Galen Gisler,[17] Sergei V. Gladkovsky,[18] Josh Wimpenny,[9] Matthew E. Sanborn,[9] Akane Yamakawa,[9] Kenneth L. Verosub,[9] Douglas J. Rowland,[19] Sarah Roeske,[9] Nicholas W. Botto,[9] Jon M. Friedrich,[20,21] Michael E. Zolensky,[22] Loan Le,[23,22] Daniel Ross,[23,22] Karen Ziegler,[24] Tomoki Nakamura,[25] Insu Ahn,[25] Jong Ik Lee,[26] Qin Zhou,[27,28] Xian-Hua Li,[28] Qiu-Li Li,[28] Yu Liu,[28] Guo-Qiang Tang,[28] Takahiro Hiroi,[29] Derek Sears,[5] Ilya A. Weinstein,[7] Alexander S. Vokhmintsev,[7] Alexei V. Ishchenko,[7] Phillipe Schmitt-Kopplin,[30,31] Norbert Hertkorn,[30] Keisuke Nagao,[32] Makiko K. Haba,[32] Mutsumi Komatsu,[33] Takashi Mikouchi,[34] (the Chelyabinsk Airburst Consortium)

The asteroid impact near the Russian city of Chelyabinsk on 15 February 2013 was the largest airburst on Earth since the 1908 Tunguska event, causing a natural disaster in an area with a population exceeding one million. Because it occurred in an era with modern consumer electronics, field sensors, and laboratory techniques, unprecedented measurements were made of the impact event and the meteoroid that caused it. Here, we document the account of what happened, as understood now, using comprehensive data obtained from astronomy, planetary science, geophysics, meteorology, meteoritics, and cosmochemistry and from social science surveys. A good understanding of the Chelyabinsk incident provides an opportunity to calibrate the event, with implications for the study of near-Earth objects and developing hazard mitigation strategies for planetary protection.

[1]Institute for Dynamics of Geospheres of the Russian Academy of Sciences, Leninsky Prospect 38, Building 1, Moscow, 119334, Russia. [2]SETI Institute, 189 Bernardo Avenue, Mountain View, CA 94043, USA. [3]NASA Ames Research Center, Moffett Field, Mail Stop 245-1, CA 94035, USA. [4]Institute of Astronomy of the Russian Academy of Sciences, Pyatnitskaya 48, Moscow, 119017, Russia. [5]Department of Theoretical Mechanics, South Ural State University, Lenin Avenue 76, Chelyabinsk, 454080, Russia. [6]Chelyabinsk State University, Bratyev Kashirinyh Street 129, Chelyabinsk, 454001, Russia. [7]Institute of Physics and Technology, Ural Federal University, Mira Street 19, Yekaterinburg, 620002, Russia. [8]Vernadsky Institute of Geochemistry and Analytical Chemistry of the RAS, Kosygina Street 19, Moscow, 119991, Russia. [9]Department of Earth and Planetary Sciences, University of California at Davis, Davis, CA 95616, USA. [10]Department of Physics, University of Helsinki, P.O. Box 64, 00014 Helsinki, Finland. [11]Koninklijk Nederlands Meteorologisch Instituut, P.O. Box 201, 3730 AE De Bilt, Netherlands. [12]Department of Geoscience and Engineering, Faculty of Civil Engineering and Geosciences, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, Netherlands. [13]Tomsk State University, Lenina Prospect 36, Tomsk, 634050, Russia. [14]Research and Production Association "Typhoon," Floor 2, 7 Engels Street, [illegible] Russia. [15]Support Foundation for Astronomy

efficiently decelerated, avoiding the transfer of momentum to lower altitudes and resulting in less damage when the blast wave reached the ground.

Damage Assessment

In the weeks after the event, 50 villages were visited to verify the extent of glass damage. The resulting map (Fig. 3) demonstrates that the shock wave had a cylindrical component, extending furthest perpendicular to the trajectory. There was little coherence of the shock wave in the forward direction, where the disturbance was of long duration, shaking buildings and making people run outside, but causing no damage.

Fig. 1. Meteoroid fragmentation stages in video taken by A. Ivanov in Kamensk-Uralskiy. (A) Fireball just before peak brightness, at the moment when camera gain was first adjusted. **(B)** End of main disruption. **(C)** Onset of secondary disruption. **(D)** End of secondary disruption; main debris cloud continues to move down. **(E)** Two main fragments remain. **(F)** Single fragment remains. **(G)** Thermally emitting debris cloud at rest with atmosphere. **(H)** Final fragment continues to penetrate. Meteor moved behind distant lamp posts. (I and J) Detail of the thermal emission from a photograph by Mr. Dudarev (I) and M. Ahmetvaleev (J), after sky subtraction with high-pass filter and contrast enhancement. Altitude scale is uncertain by +0.7 km.

Fig. 2. Fireball visual magnitude irradiance (light curve, normalized to 100-km distance. The bold dashed line shows the model fit to the light curve (SM section 1.2), with thin lines showing total mass of all fragments passing a given altitude (kg) and the altitude-dependent rate of energy deposition as a fraction of the original kinetic energy (km^{-1}).

trajectory were pushed inwards, and suspended ceilings were sucked down above broken windows (fig. S36G). There was no structural damage to buildings, other than a statue of Pushkin inside the local library, cracked by a blown-out window frame. Cracks in walls were documented in nearby Baturinsky and Kalachevo.

Electrophonic sounds were heard (SM section 1.6), but there was no evidence of an electromagnetic pulse (EMP) under the track in neighboring Emanzhelinka. Due to shock-wave–induced vibrations, electricity and cell phone connectivity was briefly halted in the Kumshaksky district at the far northern end of the damage area. The gas supply was briefly interrupted in some districts because of valves reacting to the vibrations.

People found it painful to look at the bright fireball, but glancing away prevented lasting eye damage. Of 1113 respondents to an Internet survey who were outside at the time, 25 were sunburned (2.2%), 315 felt hot (28%), and 415 [illegible] northwest winds (5 to 15 m/s (fig. S24),

cording to $r^{-3.2 \pm 0.5}$ (SM section 2.1). The majority of injuries (1210) took place in the densely populated Chelyabinsk city, but the highest fraction of people asking for assistance was near the trajectory track in the Korkinsky district (0.16%).

Meteorite Recovery

Shock radiation contributed to surface heating and ablation but did not completely evaporate all fragments of Chelyabinsk, unlike in the case of Tunguska (3). Meteorites of ~0.1 g fell near Aleksandrovka close to the point of peak brightness, masses of ~100 g fell further along the trajectory near Deputatskiy, and at least one of 3.4 kg fell near Timiryazevskiy. One hit the roof of a house in Deputatskiy (fig. S46). Falling-sphere models suggest that they originated at 32- to 26-km altitude (fig. S52), where the meteor model shows rapid fragmentation (fig. S18C). The location of the meteorites is consistent with

New Insights

Analysis Issues

When n → all

Analysis

Excerpts:

At its core, big data is about predictions.

Big data's ascendancy represents three shifts in the way we analyze information that transform how we understand and organize society.

Analysis

Excerpts:

1. In this new world we can analyze far more data. In some cases we can even process all of it relating to a particular phenomenon. … the need for sampling is an artifact of a period of information scarcity, a product of the natural constraints on interacting with information in an analog era.

2. Looking at vastly more data also permits us to loosen up our desire for exactitude. … It is a trade-off: with less error from sampling we can accept more measurement error.

Analysis

Excerpts:

3. These two shifts lead to a third change: a move away from the age-old search for causality. As humans we have been conditioned to look for causes, even though searching for causality is often difficult and may lead us down the wrong paths. In a big-data world, by contrast, we won't have to be fixated on causality; instead we can discover patterns and correlations in the data that offer us novel and invaluable insights. The correlations may not tell us precisely *why* something is happening, but they alert us *that* it is happening.

Analysis

Excerpts:

Sampling is an outgrowth of an era of information-processing constraints, when people were measuring the world but lacked the tools to analyze what they collected.

…

Yet sampling comes with the cost of the has long been acknowledged but shunted aside. It loses detail. In some cases there is no other way but to sample. In many areas however, a shift is taking place from collecting some data, to gathering as much as possible, and if feasible, getting everything:

N = all

Analysis

N = all:

When it is possible to collect data on all of the objects of interest, profound changes in analytical methods are required.

In addition to changes in analytical methods, even more profound policy issues begin to arise.

New Insights

Policy Issues

When n → all

Policy Issues

Policy Issues

We all emit a constant stream of data exhaust that can be used to monitor our daily activities…

Big Data Makes Identification Easy

Finding a needle in haystack is fairly hard …

Big Data Makes Identification Easy

Finding a needle in haystack is fairly hard …

unless you have a really big electro-magnet.

Big Data Makes Identification Easy

Connecting data to one individual on Earth is fairly hard …

Big Data Makes Identification Easy

Connecting data to one individual on Earth is fairly hard …

unless you have access to several big-data data sets, AND you know just a few, independent attributes of the person.

Big Data Makes Identification Easy

Connecting data to one individual on

Remember:

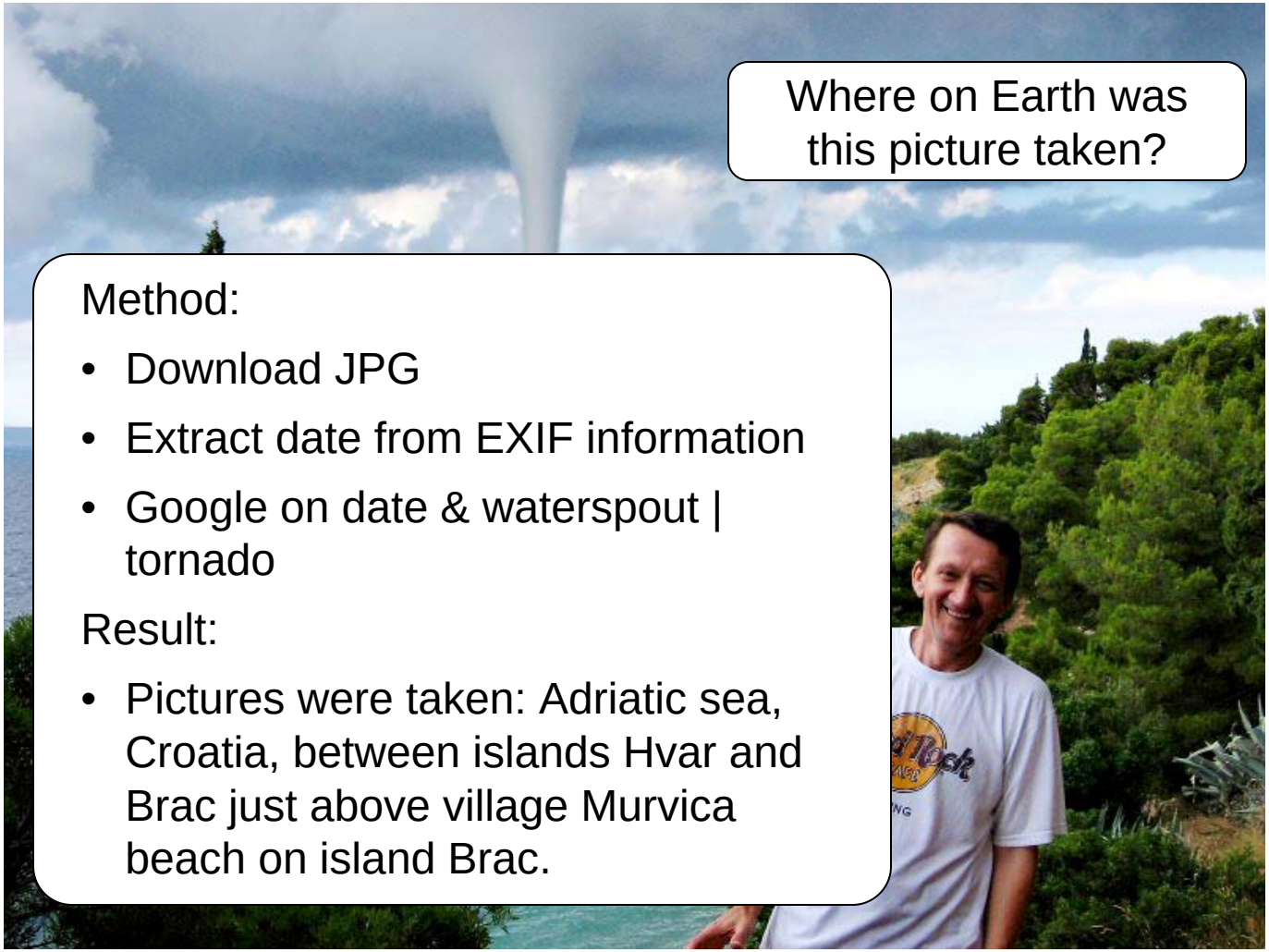
$$\frac{1}{1000} \times \frac{1}{1000} \times \frac{1}{1000} = \frac{1}{1{,}000{,}000{,}000}$$

$$\frac{1}{100} \times \frac{1}{100} \times \frac{1}{100} \times \frac{1}{100} \times \frac{1}{100} = \frac{1}{10{,}000{,}000{,}000}$$

Big Data Makes Identification Easy

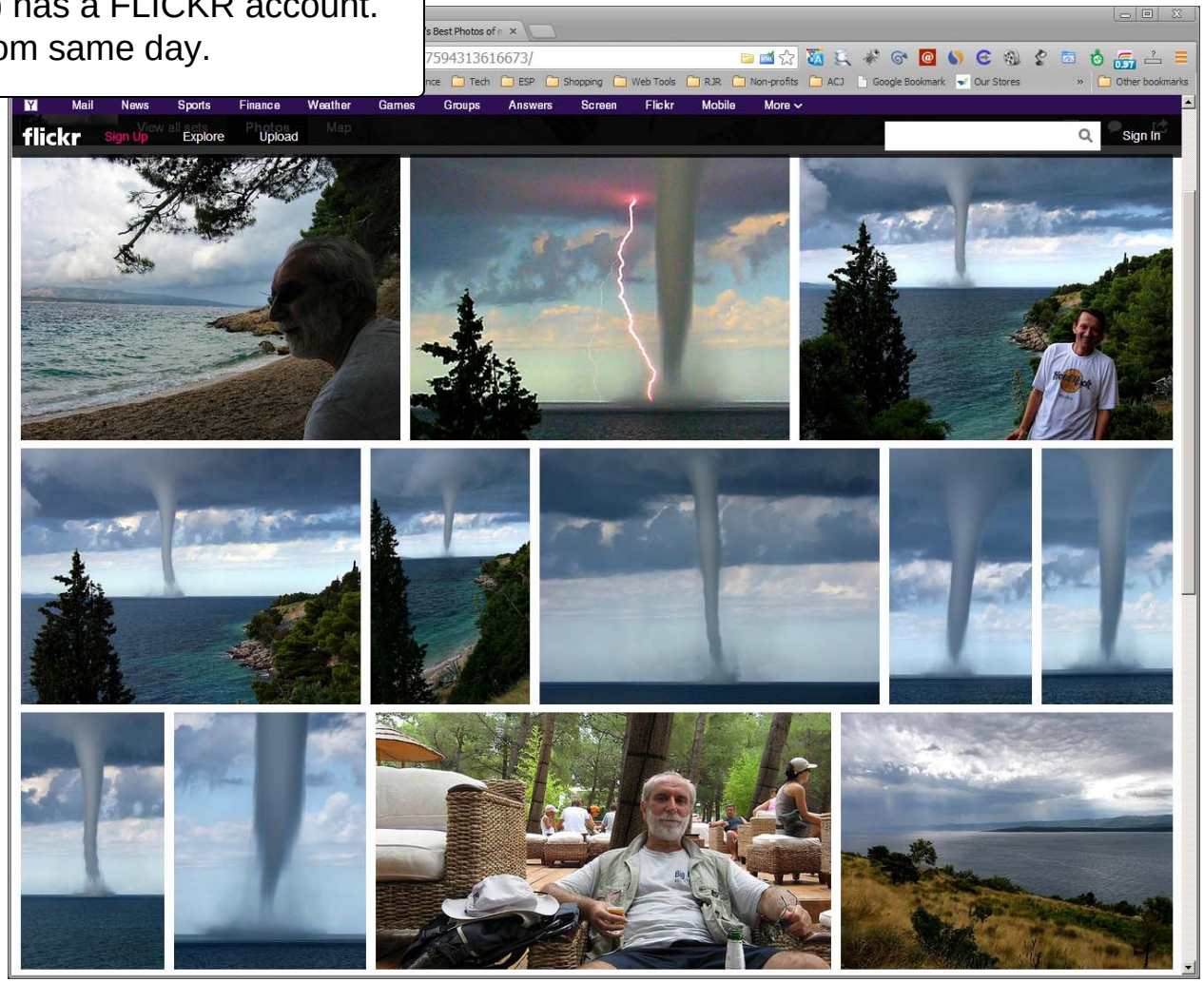
Big Data Makes Identification Easy

Big Data Makes Identification Easy



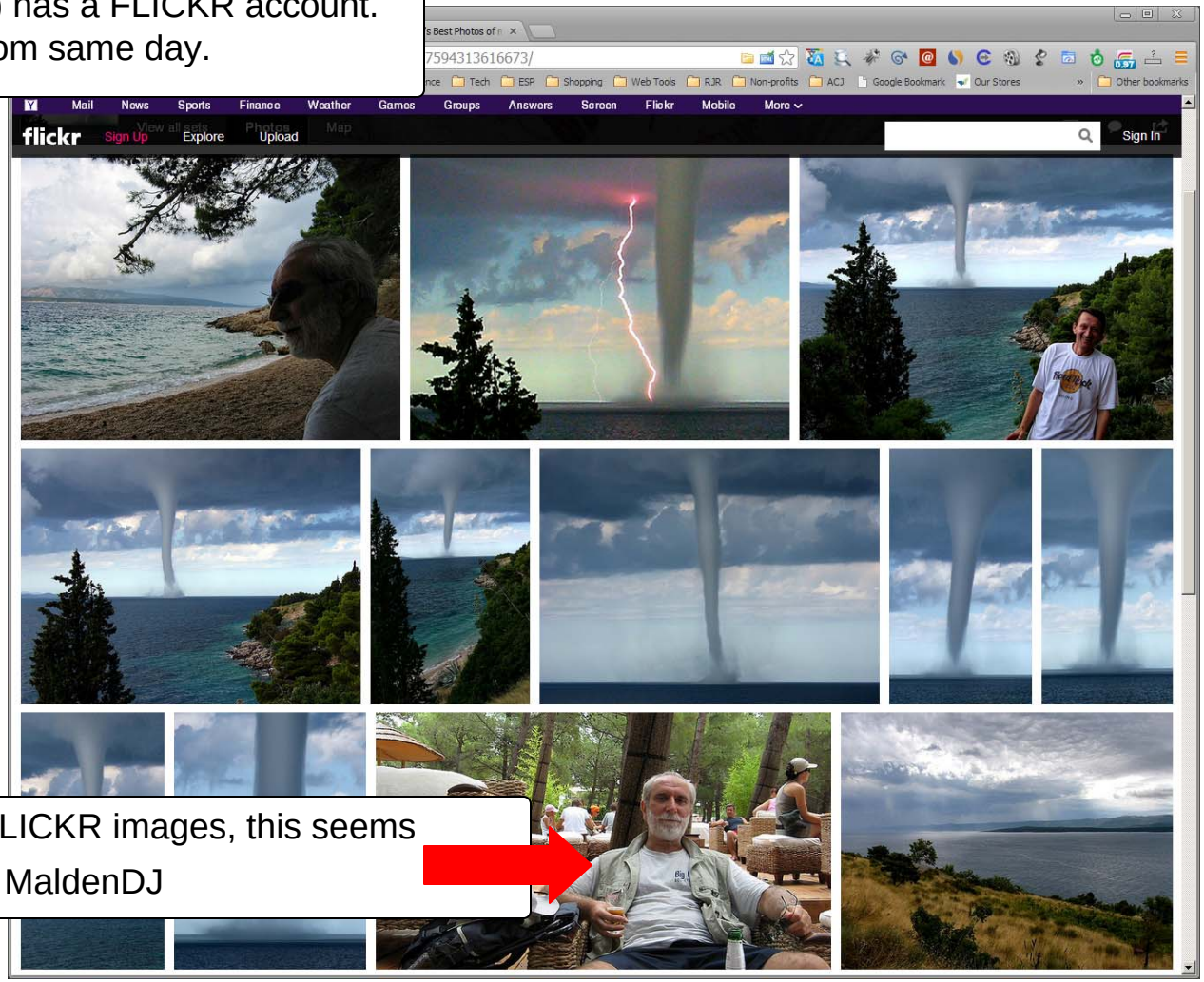
Big Data Makes Identification Easy

Photographer (MaldenDJ) has a FLICKR account.
Here are other pictures from same day.



Big Data Makes Identification Easy

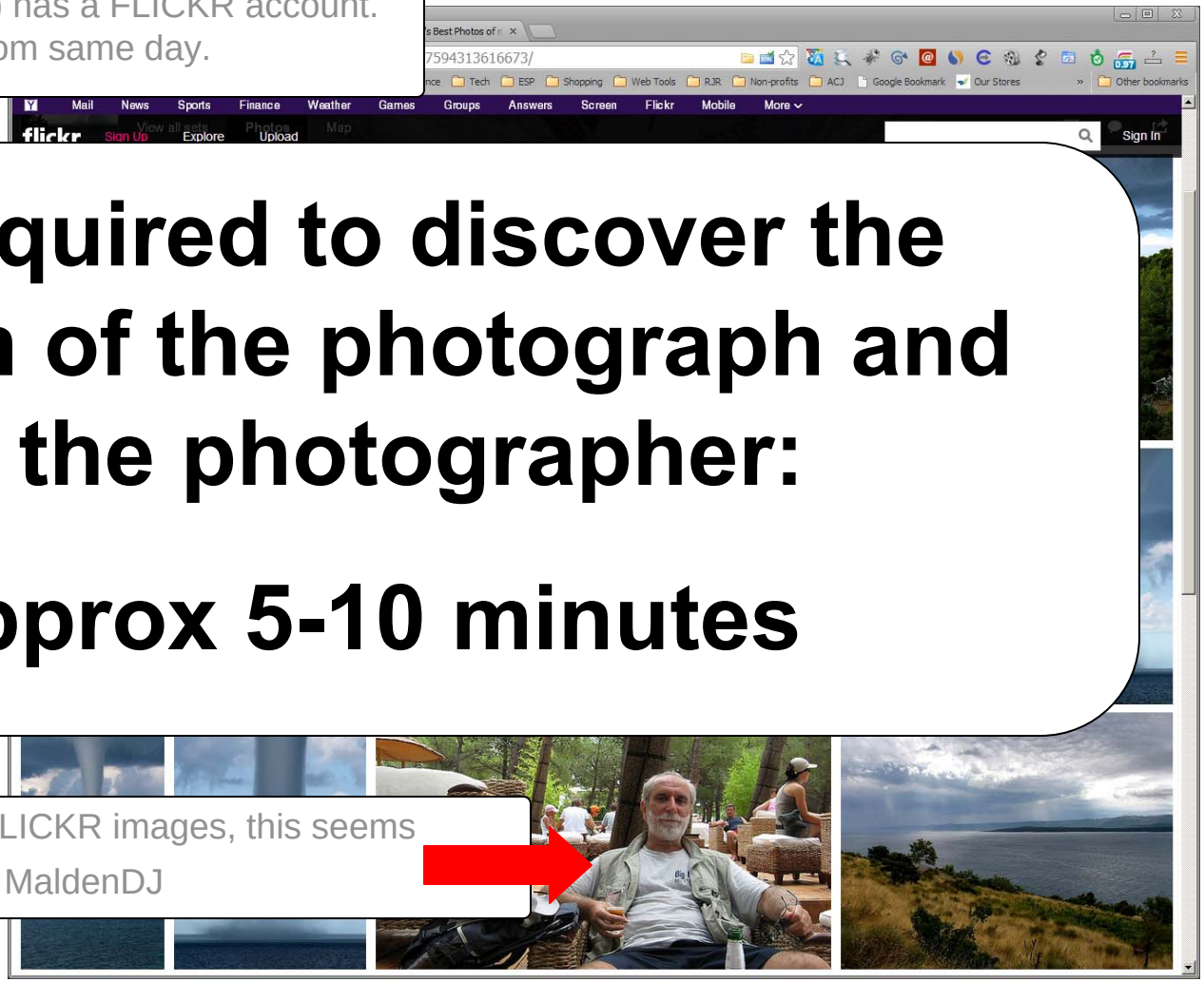
Photographer (MaldenDJ) has a FLICKR account. Here are other pictures from same day.



Based on captions on FLICKR images, this seems to be the photographer, MaldenDJ

Big Data Makes Identification Easy

Photographer (MaldenDJ) has a FLICKR account. Here are other pictures from same day.



Time required to discover the location of the photograph and identify the photographer:

Approx 5-10 minutes

Based on captions on FLICKR images, this seems to be the photographer, MaldenDJ

Big Data Makes Identification Easy

True Story:

- **I met a guy on a plane.**
- **Didn't get his name.**
- **Wished I had…**

Big Data Makes Identification Easy

What I did get:

- **He did IT consulting for pharmas**
- **He was German**
- **Seattle was his home airport**
- **PROSIS (?) is his company name**
- **His first name sounds like Hakim**
- **He knows Tim Jenkins**

Big Data Makes Identification Easy

What I did when I got home:

- **Opened Jenkins' Linked-In page**
- **Started looking at all contacts**
- **Found the name Achim Reeb**
- **Achim = Hakim, maybe ??**
- **Google "achim reeb"**
- **BINGO !**

Big Data Makes Identification Easy

What I did when I got home:

- **Opened Jenkins' Linked-In page**
- **Started looking at all contacts**
- **Found the name Achim Reeb**
- **Achim = Hakim, maybe ??**
- **Google "achim reeb"**
- **BINGO !**

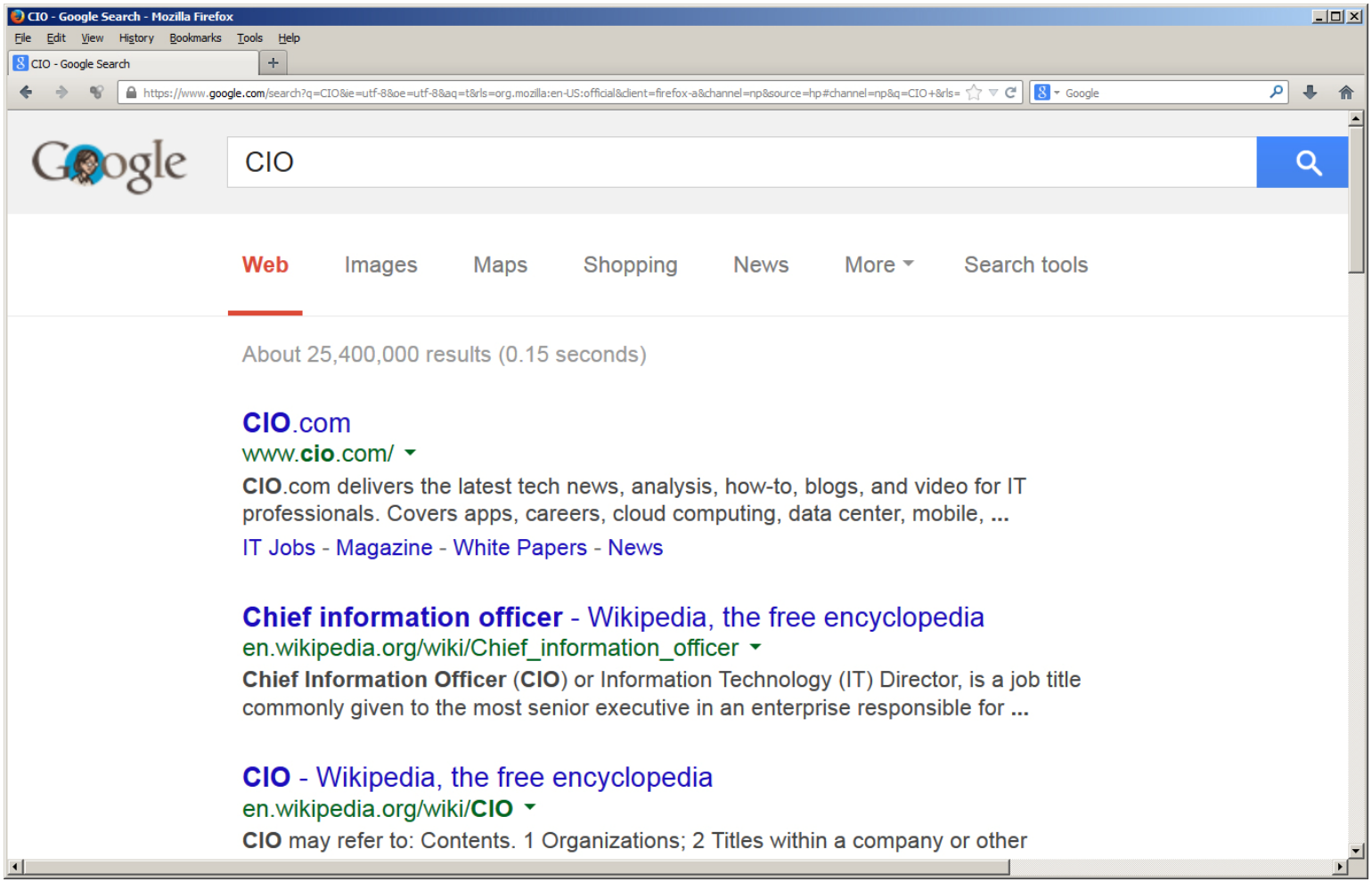
Time elapsed: ~30 min

Big Data Makes Identification Easy

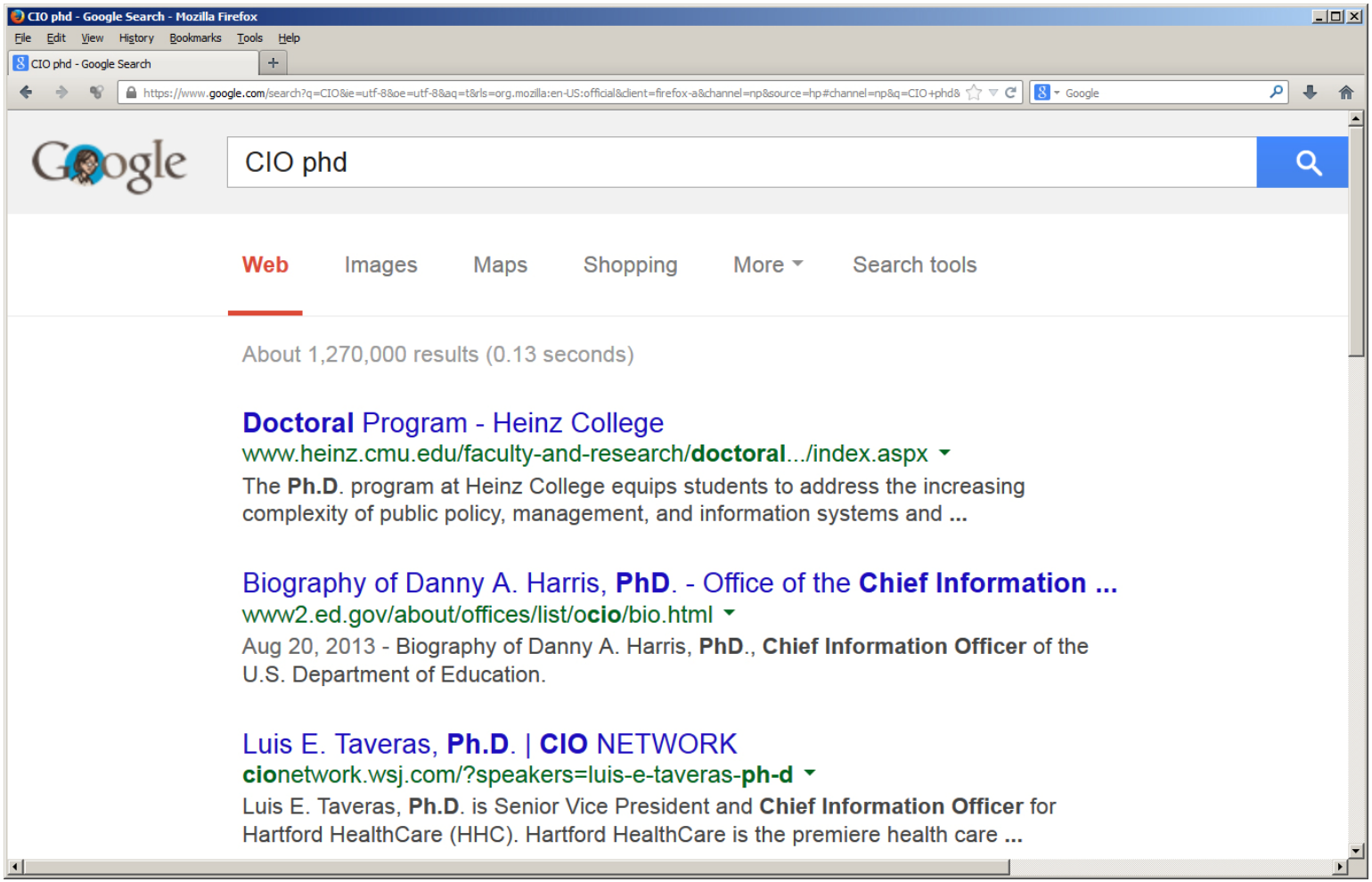
Hypothetical Story:

I met another guy on a plane …

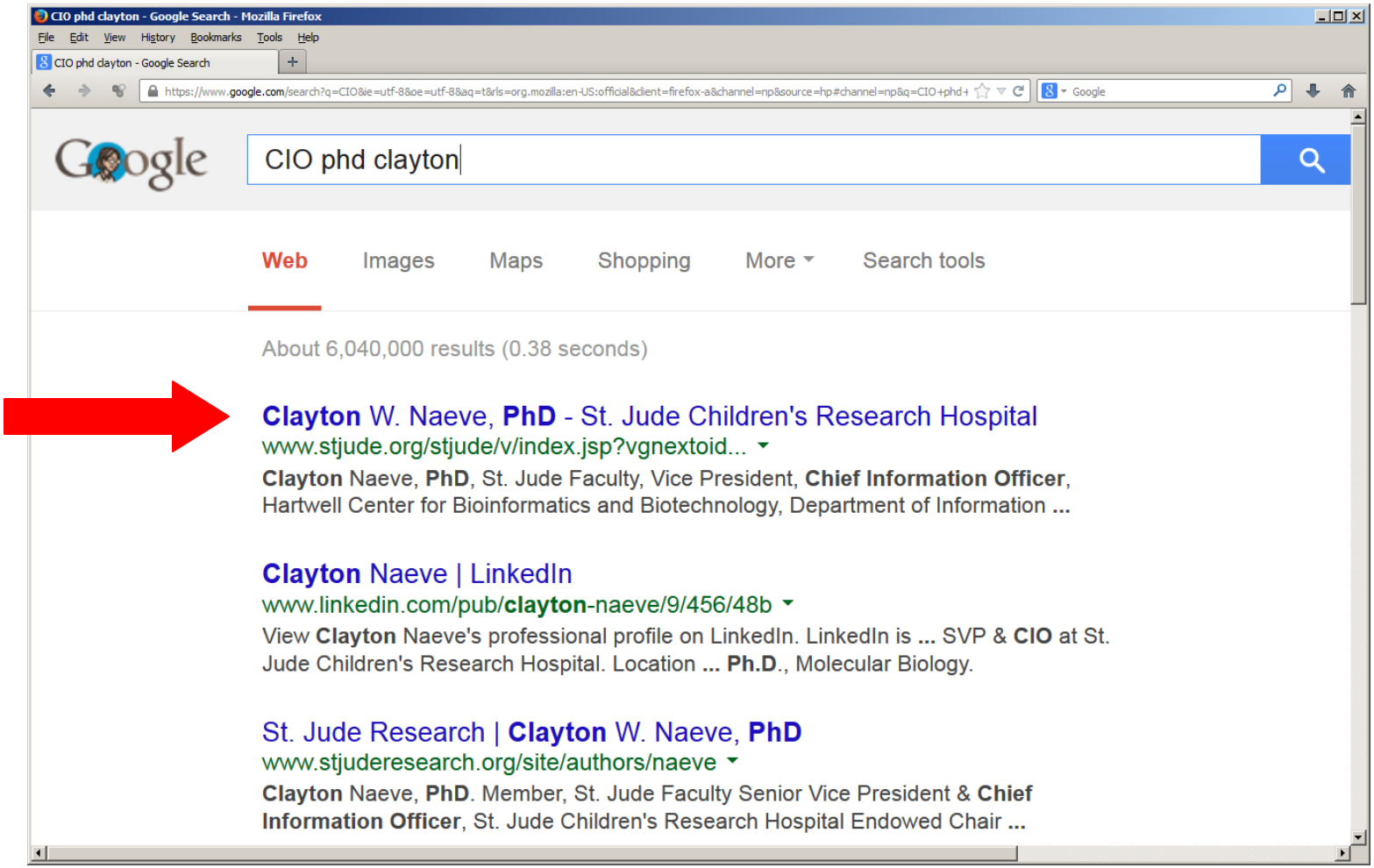
Big Data Makes Identification Easy



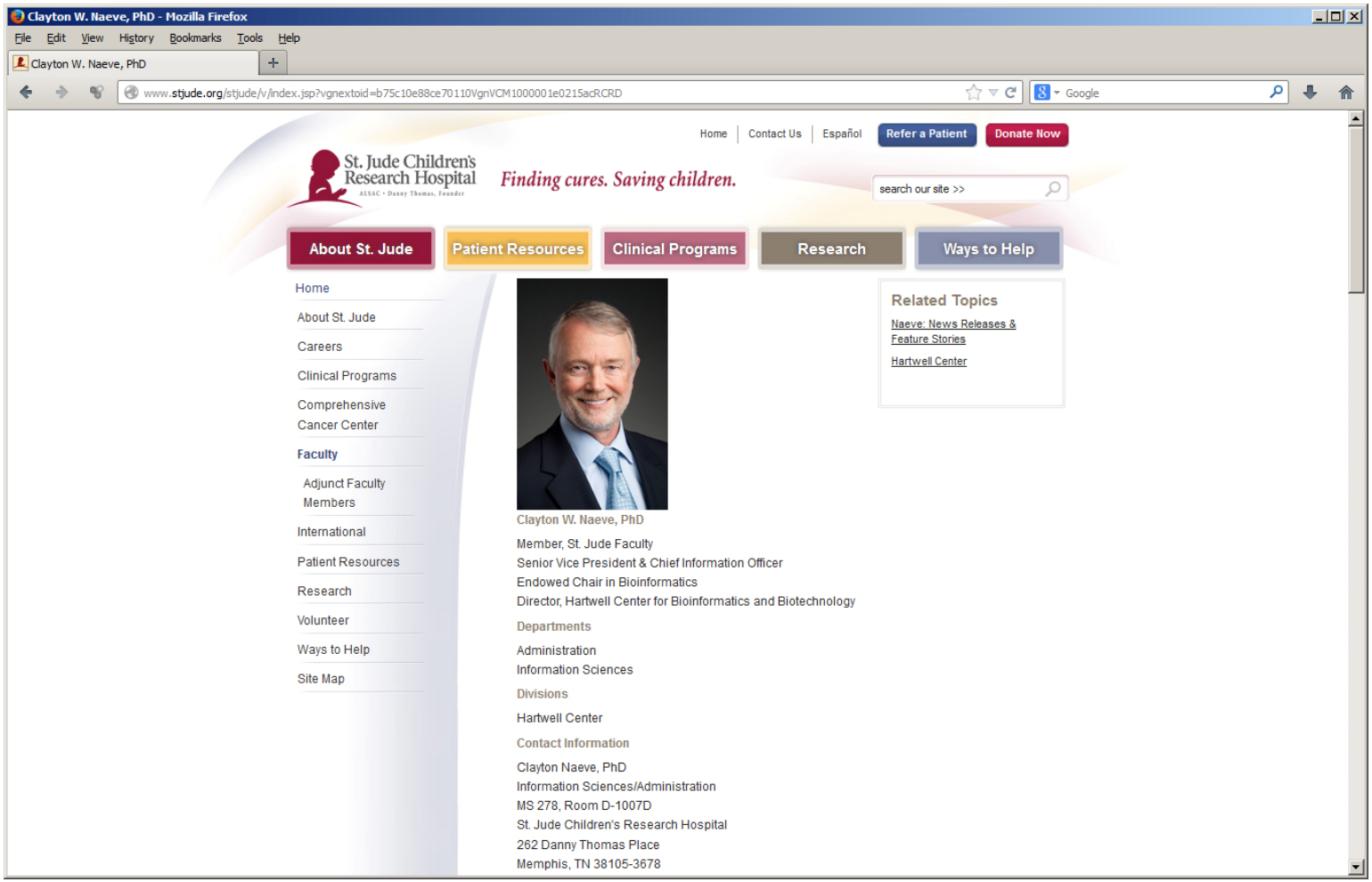
Big Data Makes Identification Easy



Big Data Makes Identification Easy



Big Data Makes Identification Easy

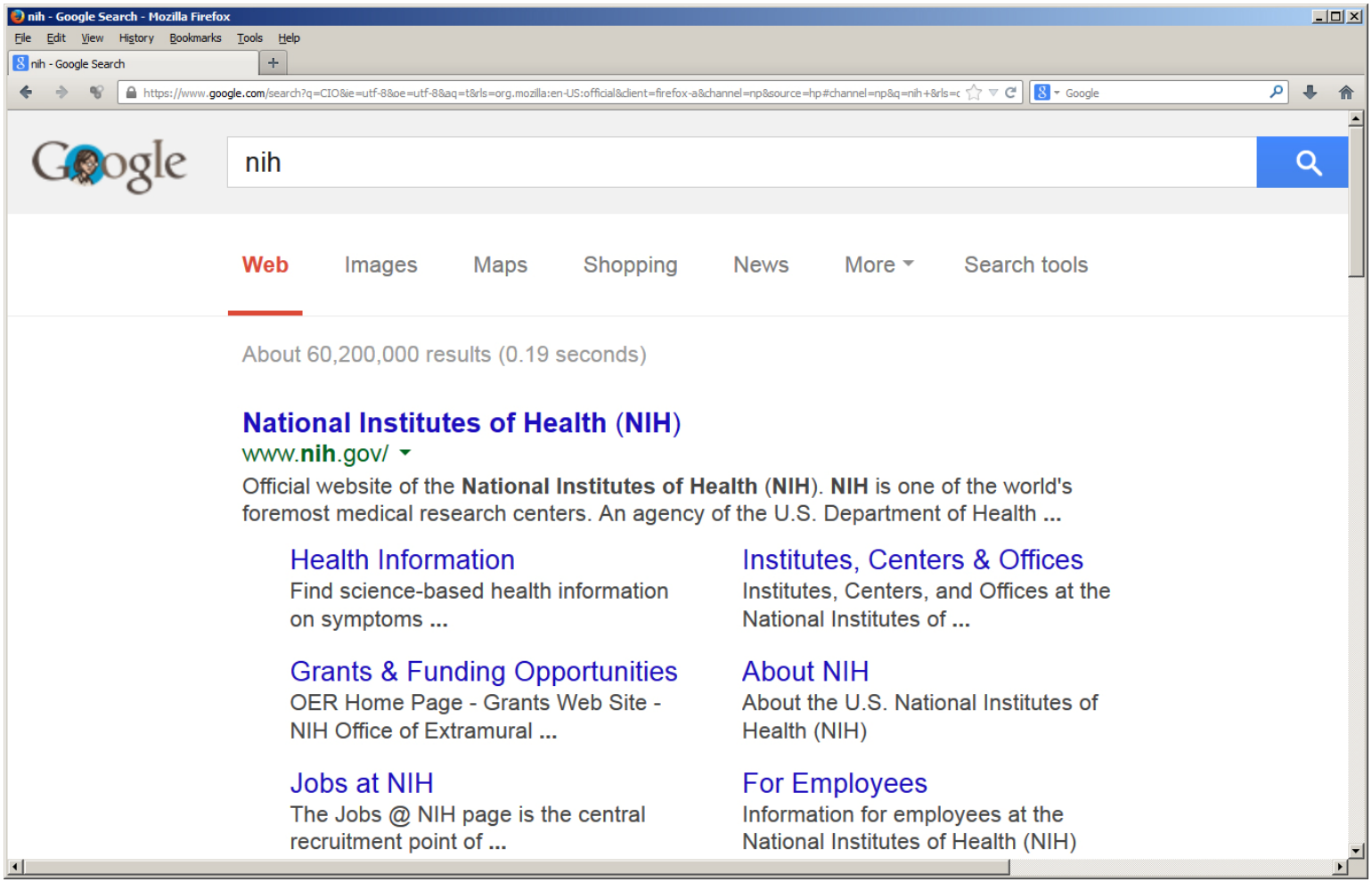


Big Data Makes Identification Easy

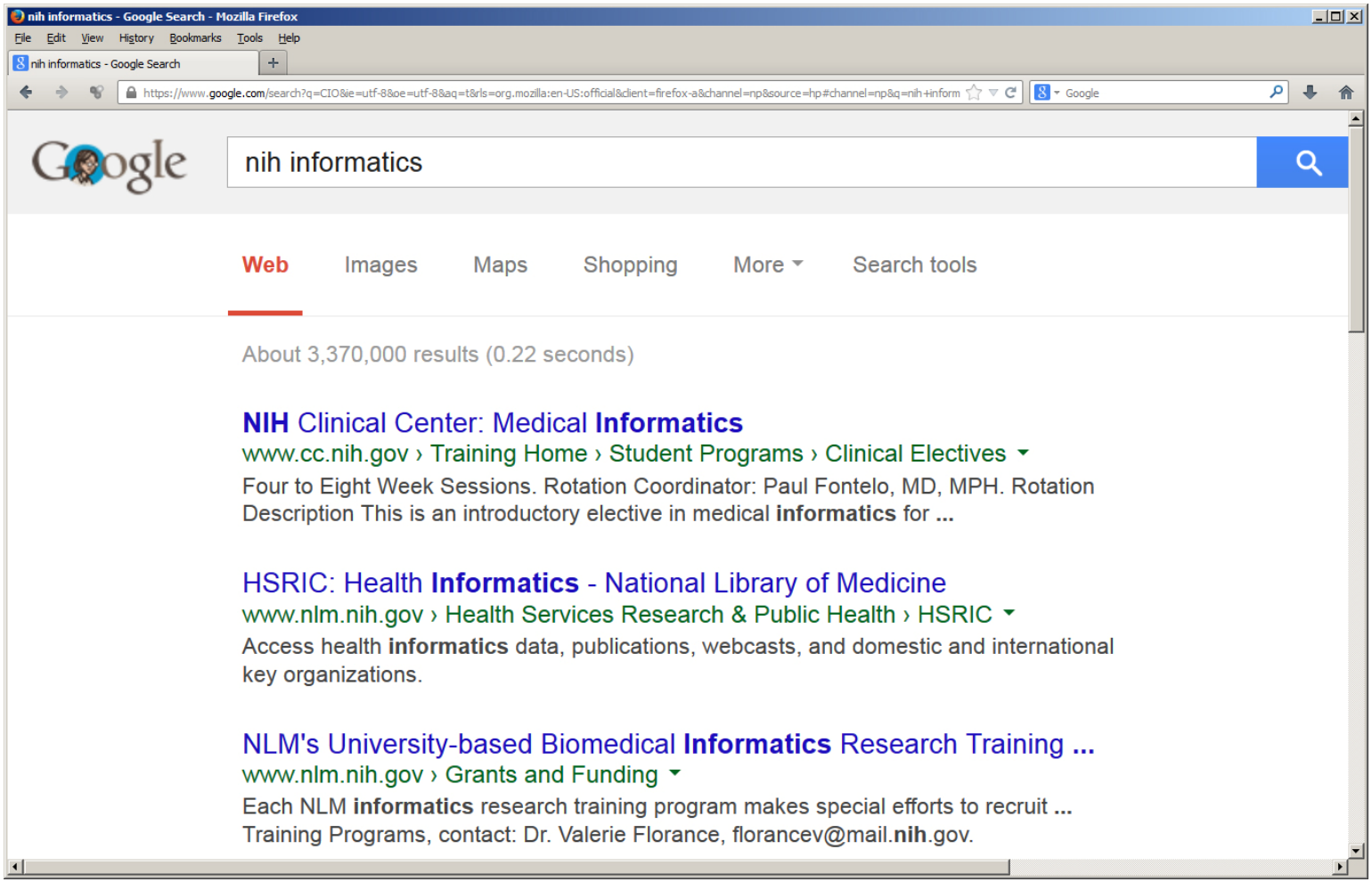
Hypothetical Story:

I met yet another guy on a plane …

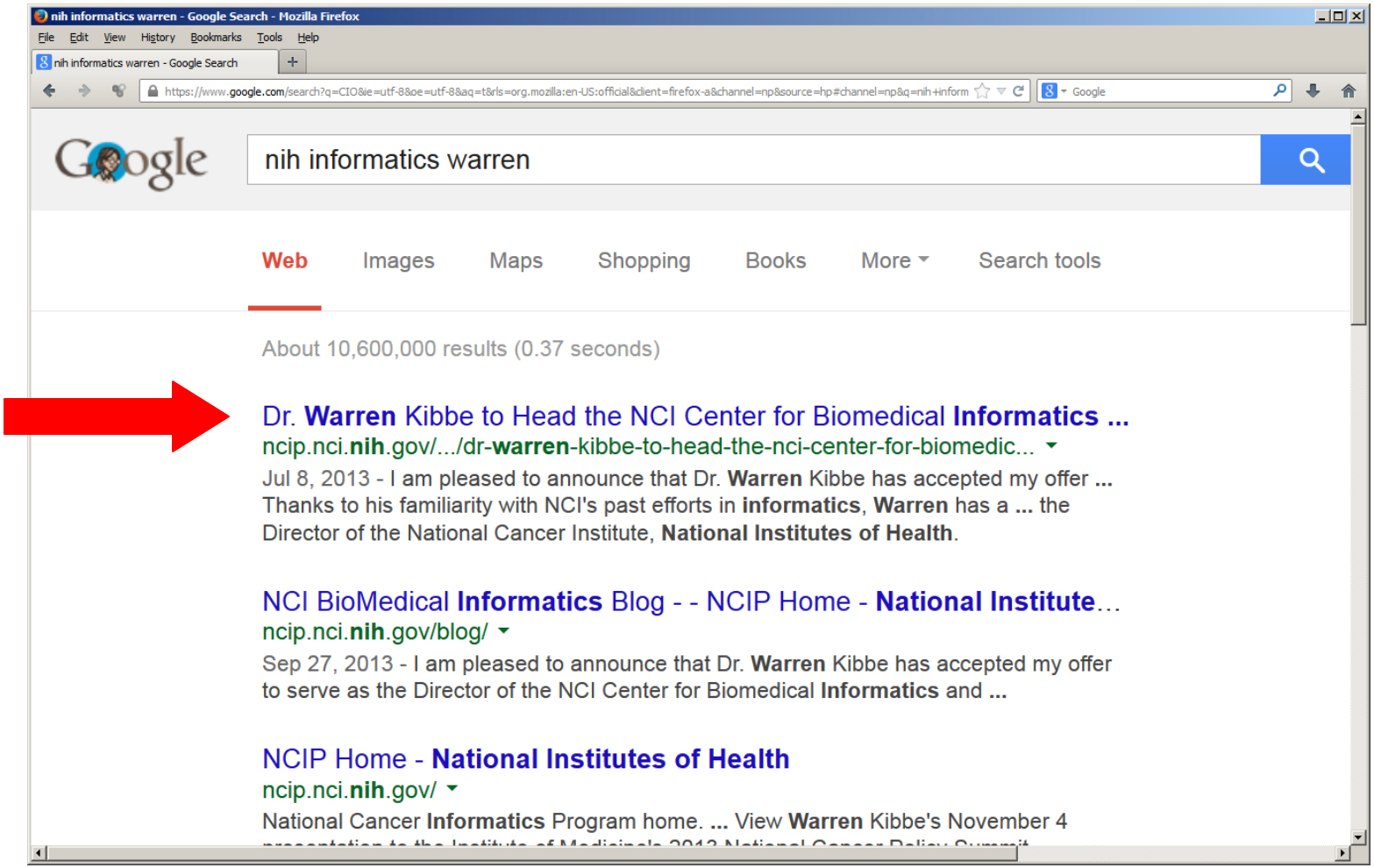
Big Data Makes Identification Easy



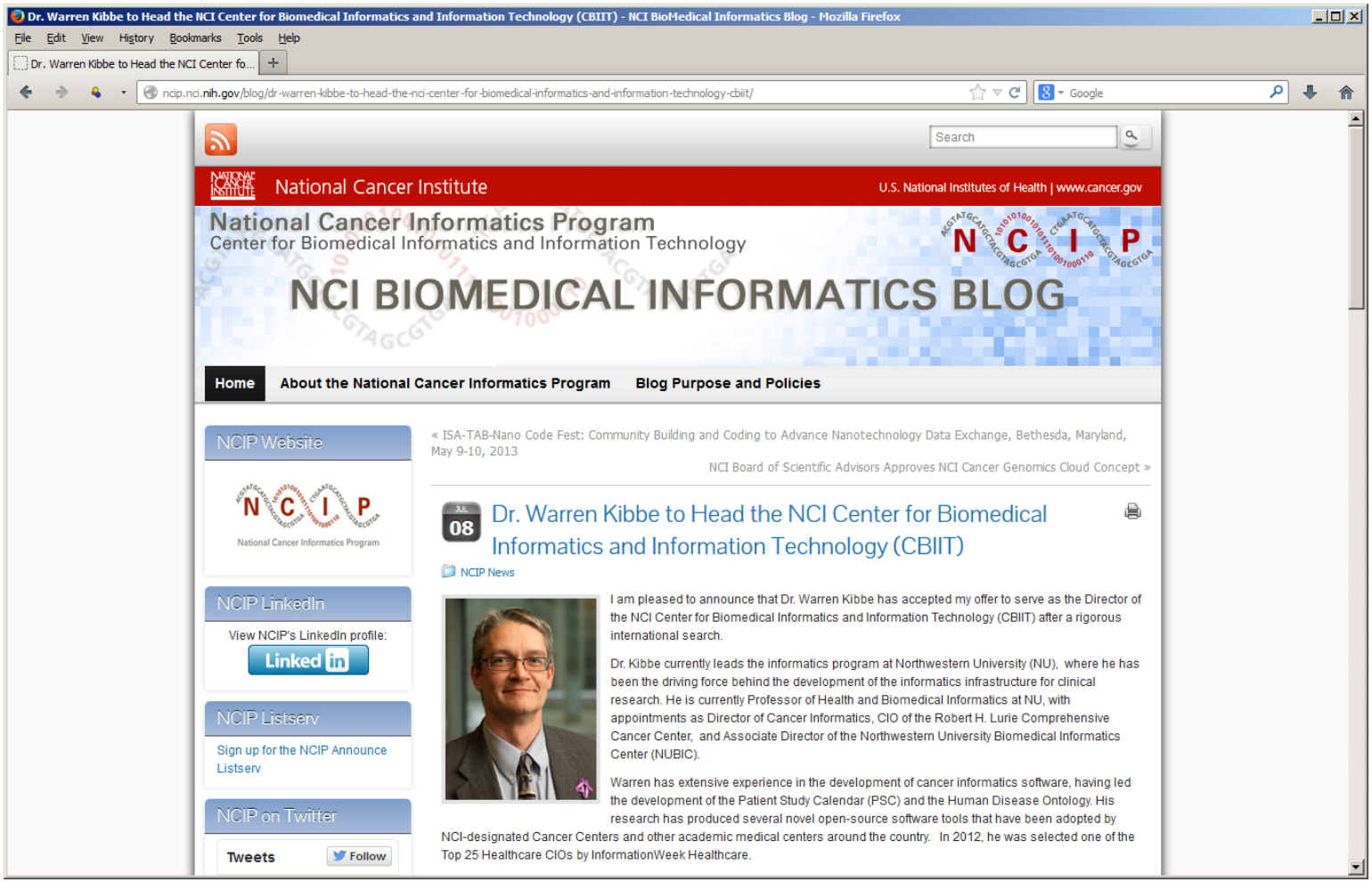
Big Data Makes Identification Easy



Big Data Makes Identification Easy



Big Data Makes Identification Easy

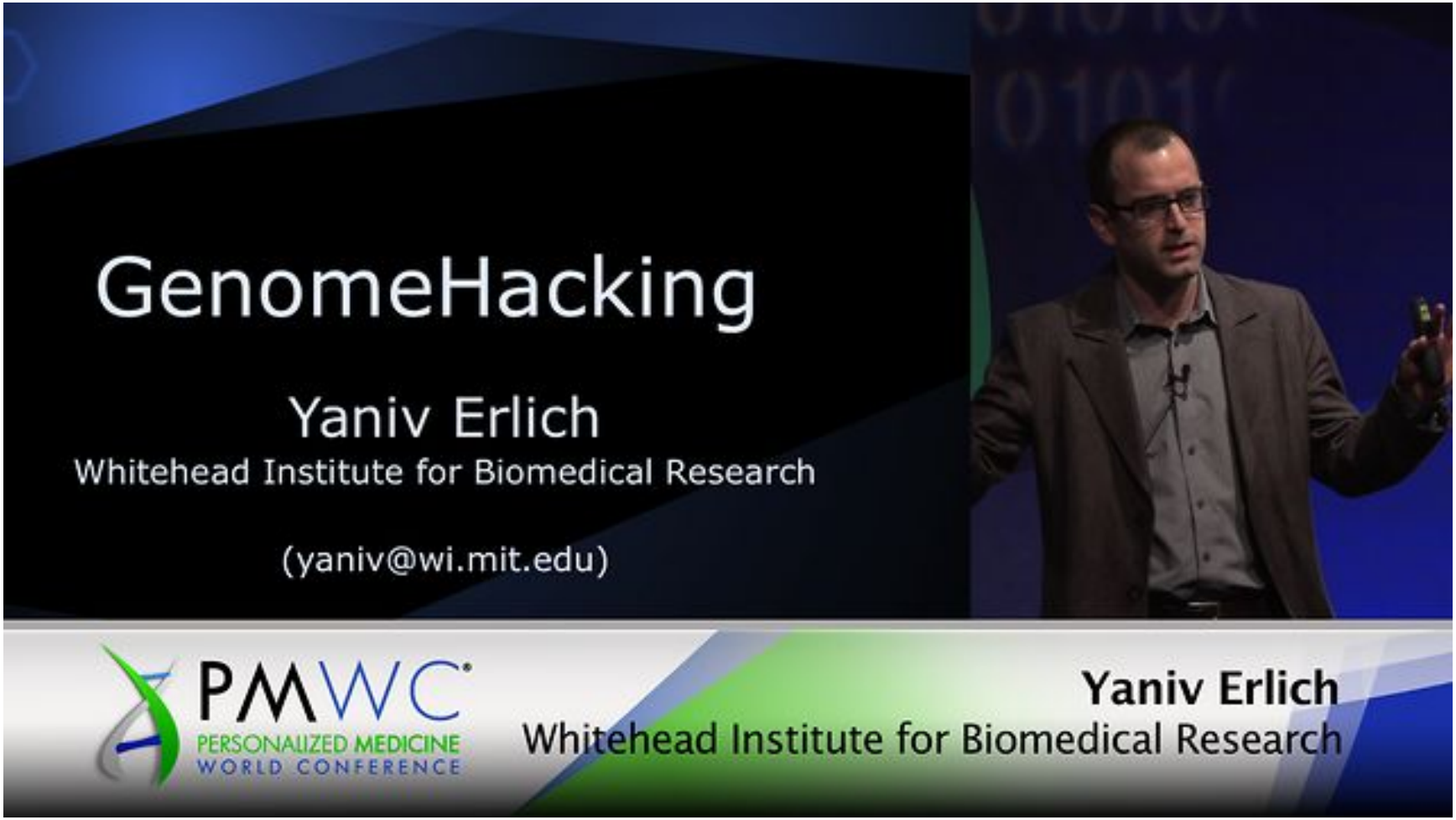


Policy Issues

SCIENTISTS EXPOSE NEW VULNERABILITIES IN THE SECURITY OF PERSONAL GENETIC INFORMATION

Using only a computer, an Internet connection, and publicly accessible online resources, a team of Whitehead Institute researchers has been able to identify nearly 50 individuals who had submitted personal genetic material as participants in genomic studies.

Policy Issues



Policy Issues

in histone pre-mRNA 3′-end processing (fig. S14).

3094 (2006).

10.1126/science.1228705

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Surnames are paternally inherited in most human societies, resulting in their co-segregation with Y-chromosome haplotypes (*1*–*5*). Based on this observation, multiple genetic genealogy companies offer services to reunite distant patrilineal relatives by genotyping a few dozen highly polymorphic short tandem repeats across the Y chromosome (Y-STRs). The association between surnames and haplotypes can be confounded by nonpaternity events, mutations, and adoption of the same surname by multiple founders (*5*). The genetic genealogy community addresses these barriers with massive databases that list the test results of Y-STR haplotypes along with their corresponding surnames. Currently, there are at least eight databases and numerous surname project Web sites that collectively contain hundreds of thousands of surname-haplotype records (table S1).

The ability of genetic genealogy databases to breach anonymity has been demonstrated in the past. In a number of public cases, male adoptees and descendants of anonymous sperm donors used recreational genetic genealogy services to genotype their Y-chromosome haplotypes and to search the companies' databases (*6*–*9*). The genetic matches identified distant patrilineal relatives and pointed to the potential surnames of their biological fathers. By combining other pieces of demographic information, such as date and place of birth, they fully exposed the identity of their biological fathers. Lunshof *et al.* (*10*) were the first to speculate that this technique could expose the full identity of participants in sequencing projects. Gitschier (*11*) empirically approached this hypothesis by testing 30 Y-STR haplotypes of CEU participants in these databases and reported that potential surnames can be detected. [CEU participants are multigenerational families of northern and western European ancestry in Utah who had originally had their samples collected by CEPH (Centre d'Etude du Polymorphisme Humain) and were later reconsented to participate in the HapMap project.] However, these surnames could match thousands of individuals, and the study did not pursue full re-identification at a single-person resolution.

Our goal was to quantitatively approach the question of how readily surname inference might be possible in a more general population, apply this approach to personal genome data sets, and demonstrate end-to-end identification of individuals with only public information. We show that full identities of personal genomes can be exposed via surname inference from recreational genetic genealogy databases followed by Internet searches. In all cases in which individuals were studied who had donated DNA samples, the informed consent statements they had signed stated privacy breach as a potential risk and the data usage terms did not prevent re-identification. Representatives of relevant organizations that funded the original studies were notified and confirmed the compliance of this study with their guidelines (*12*).

As a primary resource for surname inference, we focused on Ysearch (www.ysearch.org) and

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA. [2]Harvard–Massachusetts Institute of Technology (MIT) Division of Health Sciences and Technology, MIT, Cambridge, MA 02139, USA. [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [4]Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, MA 02114, USA. [5]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX 77030, USA. [6]Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel. [7]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. [8]Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Tel Aviv 69978, Israel. [9]The International Computer Science Institute, Berkeley, CA 94704, USA.

*To whom correspondence should be addressed. E-mail: yaniv@wi.mit.edu

Policy Issues

in histone pre-mRNA 3′-end processing (fig. S14).

3094 (2006).

10.1126/science.1228705

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data s ithout identifiers has become a common practice in genomics. Here, we report that surna covered from personal genomes by profiling short tandem repeats on the Y chromosome (ving recreational genetic genealogy databases. We show that a combination of a surna her types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Surnames are paternally inherited in most human societies, resulting in their co-segregation with Y-chromos (*1*–*5*). Based on this observation, genealogy companies offer servic tant patrilineal relatives by genotyp highly polymorphic short tandem repeats across the Y chromosome (Y-STRs). The association be- and haplotypes can be confounded

A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources

[1]Whitehead Institute for Biomedical Re Center, Cambridge, MA 02142, USA. [2]H Institute of Technology (MIT) Division o Technology, MIT, Cambridge, MA 02139, U ical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [4]Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, MA 02114, USA. [5]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX 77030, USA. [6]Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel. [7]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. [8]Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Tel Aviv 69978, Israel. [9]The International Computer Science Institute, Berkeley, CA 94704, USA.

*To whom correspondence should be addressed. E-mail: yaniv@wi.mit.edu

sites that collect sands of surname-haplotype records (table S1).

The ability of genetic genealogy databases to breach anonymity has been demonstrated in the past. In a number of public cases, male adoptees and descendants of anonymous sperm donors used recreational genetic genealogy services to genotype their Y-chromosome haplotypes and to search the companies' databases (*6*–*9*). The genetic matches identified distant patrilineal relatives and pointed to the potential surnames of their biological fathers.

By combining other pieces of demographic information, such as date and place of birth, they fully exposed the identity of their biological fathers. Lunshof *et al.* (*10*) were the first to speculate that this technique could expose the full identity of participants in sequencing projects. Gitschier (*11*) empirically approached this hypothesis by testing 30 Y-STR haplotypes of CEU participants in these databases and reported that potential surnames can be detected. [CEU participants are multigenerational families of northern and western European ancestry in Utah who had originally had their samples collected by CEPH (Centre d'Etude du Polymorphisme Humain) and were later reconsented to participate in the HapMap project.] However, these surnames could match thousands of individuals, and the study did not pursue full re-identification at a single-person resolution.

Our goal was to quantitatively approach the surname inference might searches. In all cases in which studied who had donated DNA samples, the informed consent statements they had signed stated privacy breach as a potential risk and the data usage terms did not prevent re-identification. Representatives of relevant organizations that funded the original studies were notified and confirmed the compliance of this study with their guidelines (*12*).

As a primary resource for surname inference, we focused on Ysearch (www.ysearch.org) and

Policy Issues

And yet…

NIH GWAS policy, soon to be included in the more general NIH genomic-data sharing policy, requires that institutions stipulate that:

Policy Issues

And yet…

NIH GWAS policy, soon to be included in the more general NIH genomic-data-sharing policy, requires that institutions stipulate that:

> Data should be de-identified according to the following criteria: the identities of data subjects cannot be readily ascertained or otherwise associated with the data by the repository staff or secondary data users (45 C.F.R. 46.102(f)); the 18 identifiers enumerated at section 45 C.F.R. 164.514(b)(2) (the HIPAA Privacy Rule) are removed; and the submitting institution has no actual knowledge that the remaining information could be used alone or in combination with other information to identify the subject of the data.

Policy Issues

And yet…

NIH GWAS policy, soon to be included in the more general NIH genomic-data-sharing policy, requires that institutions stipulate that:

> Data should be de-identified according to the following criteria: the identities of data subjects cannot be readily ascertained or otherwise associated with the data by the repository staff or secondary data users (45 C.F.R. 46.102(f)); the 18 identifiers enumerated at section 45 C.F.R. 164.514(b)(2) (the HIPAA Privacy Rule) are removed; and the submitting institution has no actual knowledge that the remaining information could be used alone or in combination with other information to identify the subject of the data.

Policy Issues

And yet…

NIH GWAS policy, soon to be included in the more general NIH genomic-data-sharing policy, requires that institutions stipulate that:

> Data should be de-identified according to the following criteria: the identities of data subjects cannot be readily ascertained or otherwise associated with the data by the repository staff or secondary data users (45 C.F.R. 46.102(f)); the 18 identifiers enumerated at section 45 C.F.R. 164.514(b)(2) (the HIPAA Privacy Rule) are removed; and the submitting institution has no actual knowledge that the remaining information could be used alone or in combination with other information to identify the subject of the data.

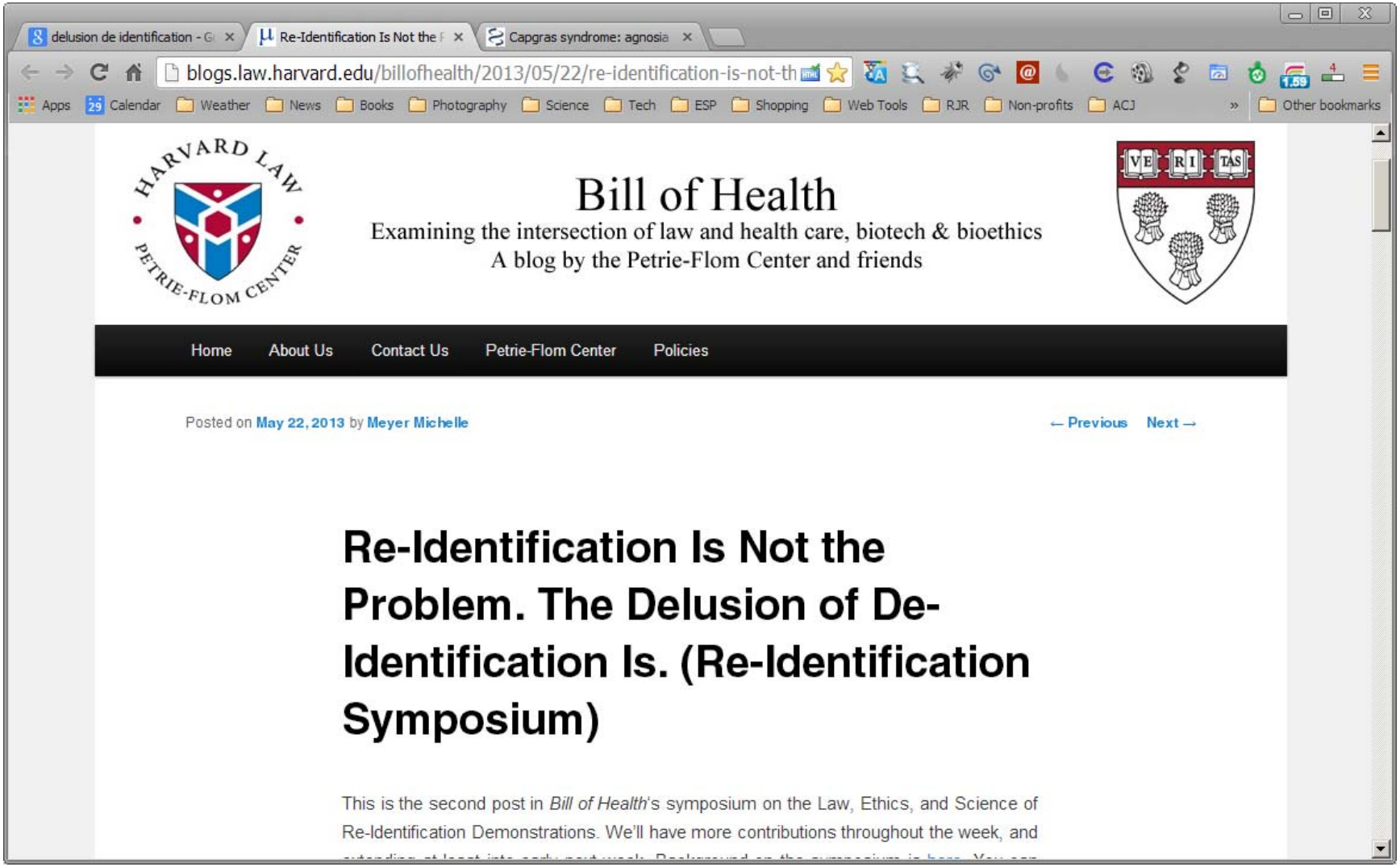
Policy Issues

And yet…

This is a requirement, by NIH, that grantee institutions stipulate to a known falsehood in order to receive funding for genomic-scale sequencing.

no actual knowledge that the remaining information could be used alone or in combination with other information to identify the subject of the data.

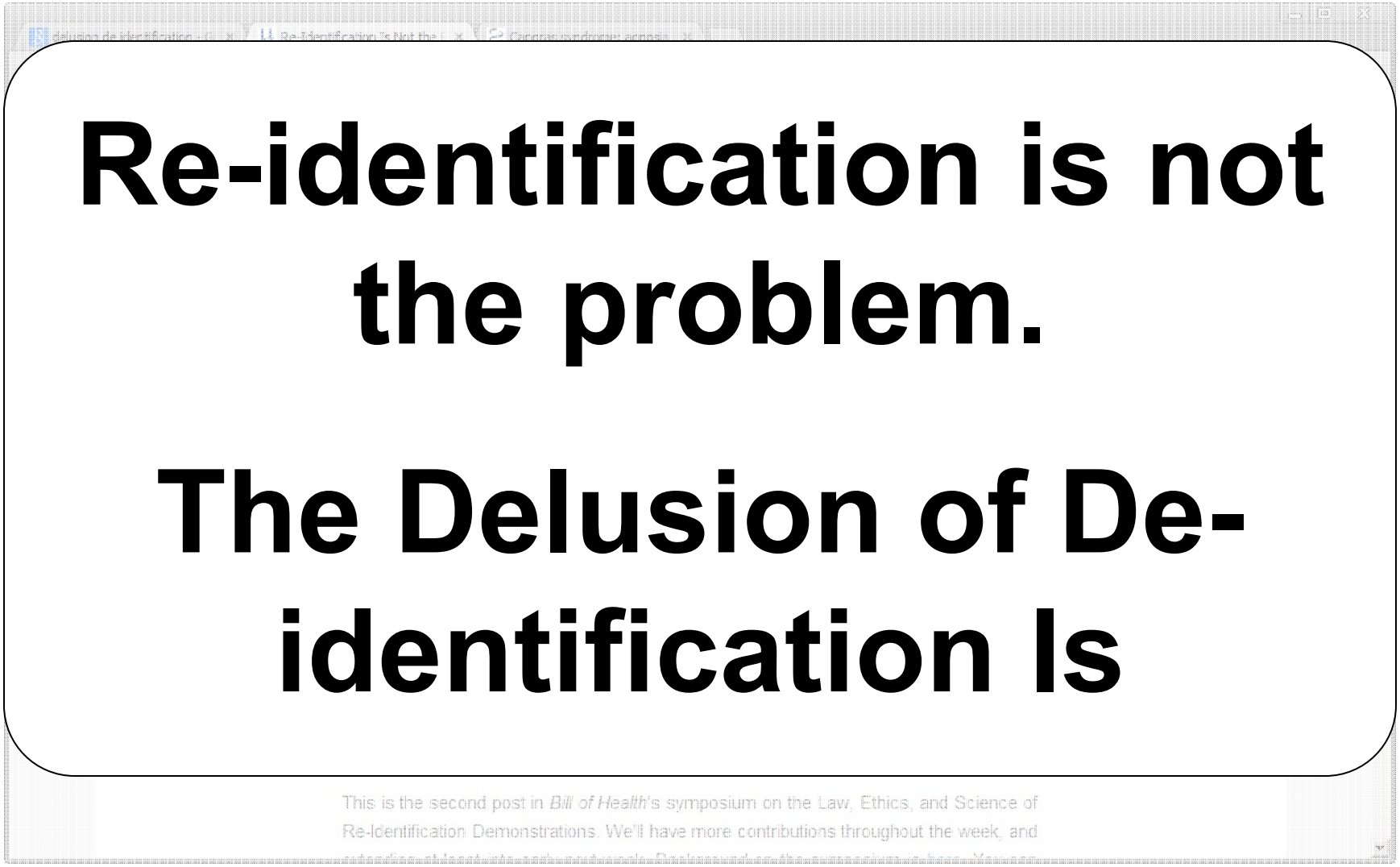
Policy Issues



Policy Issues

Re-identification is not the problem.

The Delusion of De-identification Is



This IS a Big Deal

Big Deal

The (almost trivial) possibilities for data reidentification, using freely available Big Data, means that NIH should (must?) entertain a serious re-analysis of the entire issue of data privacy.

Big Deal

The (almost trivial) possibilities for data reidentification, using freely available Big Data, means that NIH should (must?) entertain a serious re-analysis of the entire issue of data privacy.

We must either develop a new approach that recognizes the possibility (likelihood!) of re-identification, or …

Big Deal

The (almost trivial) possibilities for data reidentification, using freely available Big Data, means that NIH should (must?) entertain a serious re-analysis of the entire issue of data privacy.

We must either develop a new approach that recognizes the possibility (likelihood!) of re-identification, or …

prepare to abandon data sharing completely.

Possible New Approach

Excerpts:

For decades an essential principle of privacy laws around the world has been to put individuals in control by letting them decide whether, how, and by whom their personal information may be processed. In the Internet age, this laudable ideal has often morphed into a formulaic system of "notice and consent." In the era of big data, however, when much of data's value is in secondary uses that may have been unimagined when the data was collected, such a mechanism to ensure privacy is no longer suitable.

Possible New Approach

Excerpts:

We envision a very different privacy framework for the big-data age, one focused less on individual consent at the time of collection and more on holding data users accountable for what they do. … This spurs creative reuses of the data, while at the same time it ensures that sufficient measures are taken to see that individuals are not hurt.

Possible New Approach

Excerpts:

Running a formal big-data use assessment correctly and implementing its findings accurately offers tangible benefits to data users: they will be free to pursue secondary uses of personal data in many instances without having to go back to individuals to get their explicit consent. On the other hand, sloppy assessments or poor implementation of safeguards will expose data users to legal liability, and regulatory actions such as mandates, fines, and perhaps even criminal prosecution. Data-user accountability only works when it has teeth.

Possible New Approach

Excerpts:

Shifting the burden of responsibility from the public to the users of data makes sense for a number of reasons. They know much more than anybody else, and certainly more than consumers or regulators, about how they intend to use the data. By conducting the assessment themselves (or hiring experts to do it) they will avoid the problem of revealing confidential business strategies to outsiders. Perhaps most important, the data users reap most of the benefits of secondary use, so it's only fair to hold them accountable for their actions and place the burden for this review on them.