Statistical Inference, Learning and Models for Big Data

Steering Committee

Yoshua Bengio
Hugh Chipman
Sallie Keller
Richard Lockhart
Lisa Lix
Nancy Reid
Ruslan Salakhutdinov

International Advisory Committee

Constantine Gatsonis
Susan Holmes
Snehelata Huzurbazar
Nicolai Meinshausen
Dale Schuurmans
Robert Tibshirani
Bin Yu

Workshops

- Opening Conference and Bootcamp — Jan 12 - 23
- Statistical Machine Learning — Jan 26 - 30
- Optimization and Matrix Methods — Feb 9 - 11
- Visualization: Strategies and Principles — Feb 23 - 27
- Big Data in Health Policy — Mar 23 - 27
- Big Data for Social Policy — Apr 13 - 16

- Networks, Web mining, and Cyber-security — May 4 - 8, CRM
- Statistical Theory for Large-scale Data — April 20 - 24, PIMS
- Complex Spatio-temporal Data — April 20 - 24, Fields
- Commercial and Retail Banking — April 29 - 30, Fields
- Challenges in Environmental Science — May 11 - 15, PIMS

And more

- Distinguished Lecture Series in Statistics
  - Terry Speed, Walter and Eliza Hall Institute, April 9 and 10
  - Bin Yu, UC Berkeley, April 23 and 24
- Coxeter Lecture Series
  - Michael Jordan, UC Berkeley, April 7, 8 and 9
- Andrew Lo, MIT March 25 – law, finance and big data
- Graduate Courses
  - Statistical Machine Learning
  - Topics in Big Data
- Industrial Problem Solving Workshop
  - Fields Institute, May (TBD)
- Fields Summer Undergraduate Research Program
  - May to August, 2015

And more

Wednesday February 25	
9:30-10:30	**Sheelagh Carpendale**, University of Calgary *Information Visualization: Exploring New Options*
10:30-11:00	Coffee break
11:00-12:00	**Mark Hansen**, Columbia University
12:00-1:00	Lunch break
1:00-2:00	**Christopher Collins,** UOIT *Semantics and Sentiment in Visual Text Analytics*
2:00-3:00	**TBA**
3:00-3:30	Tea break
4:30 *BA1130	**Blyth Lecture Series 2015** **David Donoho**, Stanford University Public Lecture: *What's the Big Deal about "Big Data"? Emergent Phenomena in High-Dimensional Data Analy* **This event takes place at BA1130 (Bahen Center).** Reception will precede the talk at 3:30 p.m. in the Departm lounge (40 St. George St., 6th floor.)

MDM 12 – Einat Gil et al.

Visualization for Big Data:
Strategies and Principles

Organizing Committee

Susan Holmes

Snehelata Huzurbazar

Hadley Wickham

Leland Wilkinson

Nancy Reid

Esther Burzunza

Lauren Pelc-McArthur

Visualization for Big Data:
Strategies and Principles

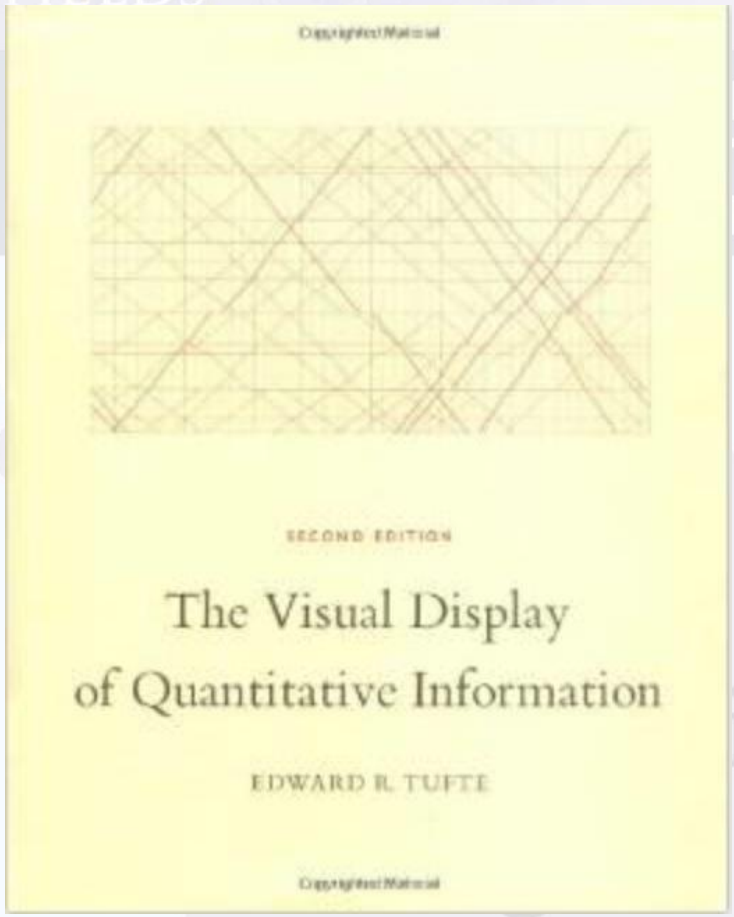
- data representation
- data exploration via filtering, sampling and aggregation
- visualization and cognition
- information visualization
- statistical modeling and software
- cognitive science and design

Visualization for Big Data:
Strategies and Principles

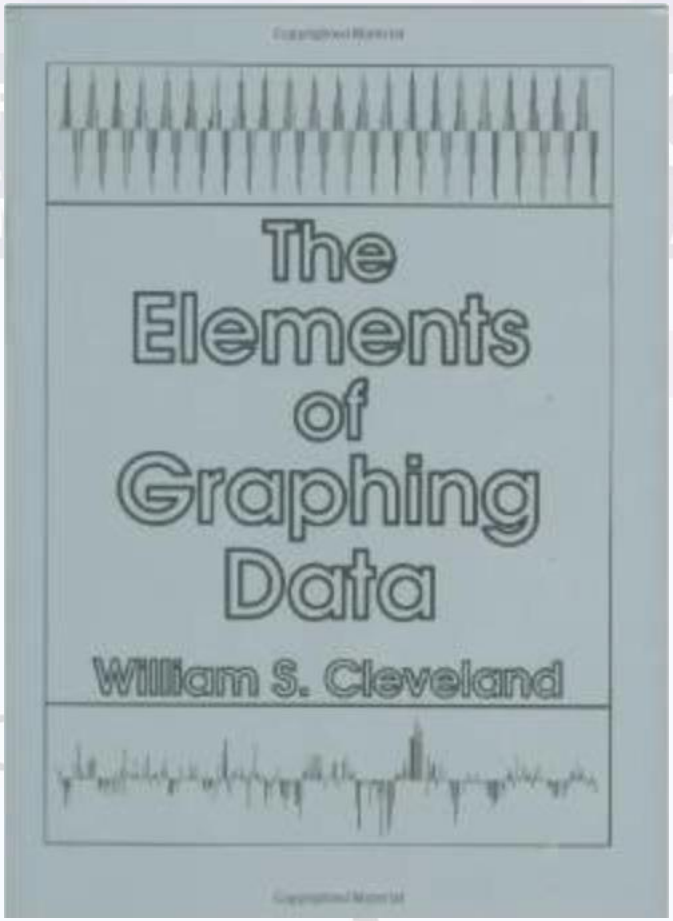
Visualization for Big Data:
Strategies and Principles

In addition to approaches such as search, query processing, and analysis, visualization techniques will also become critical across many stages of big data use--to obtain an initial assessment of data as well as through subsequent stages of scientific discovery.

Visualization for Big Data:
Strategies and Principles

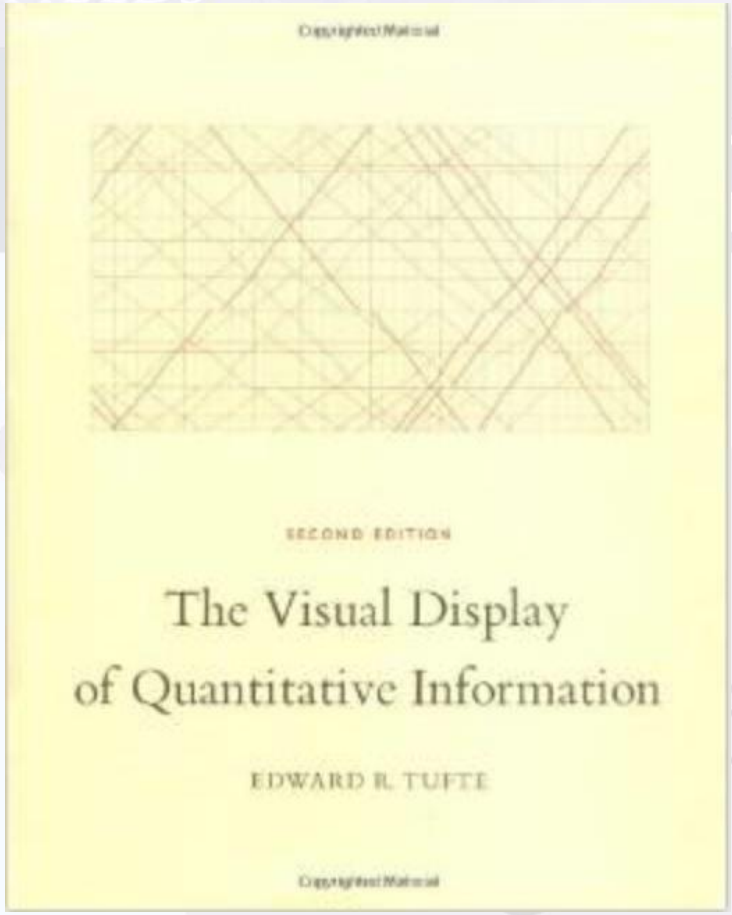


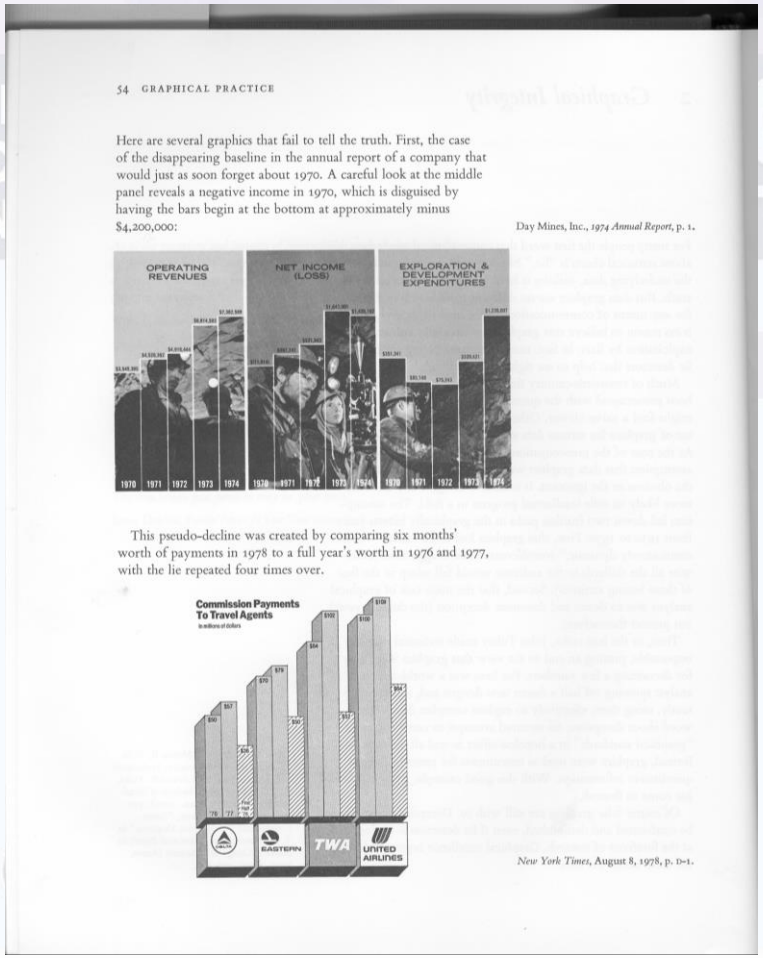
1983



1985

Visualization for Big Data:
Strategies and Principles





54 GRAPHICAL PRACTICE

Here are several graphics that fail to tell the truth. First, the case of the disappearing baseline in the annual report of a company that would just as soon forget about 1970. A careful look at the middle panel reveals a negative income in 1970, which is disguised by having the bars begin at the bottom at approximately minus $4,200,000:

Day Mines, Inc., *1974 Annual Report*, p. 1.

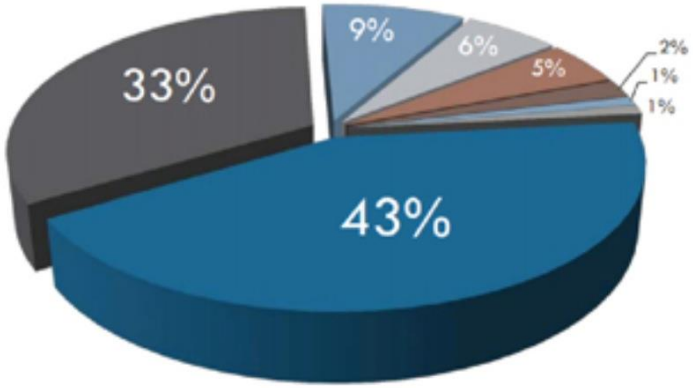
This pseudo-decline was created by comparing six months' worth of payments in 1978 to a full year's worth in 1976 and 1977, with the lie repeated four times over.

New York Times, August 8, 1978, p. D-1.

Visualization for Big Data:
Strategies and Principles

WTF Visualizations

Visualizations that make no sense.
For a discussion of what is wrong with a particular visualization, tweet at us @WTFViz.
Check out our friends Thumbs Up Viz and accidental aRt, or submit.



Science, Jan 30

change over time. In an experiment conducted in Helsinki (57), the installation of sensing and monitoring technology in households led family members initially to change their behavior, particularly in relation to conversations, nudity, and sex. And yet, if they accidentally performed an activity, such as walking naked into the kitchen in front of the sensors, it seemed to have the effect of "breaking the ice"; participants then showed less concern about repeating the behavior. More generally, participants became inured to the presence of the technology over time.

The context-dependence of privacy concern has major implications for the risks associated with modern information and communication technology (58). With online interactions, we no longer have a clear sense of the spatial boundaries of our listeners. Who is reading our blog post? Who is looking at our photos online? Adding complexity to privacy decision-making, boundaries between public and private become even less defined in the online world (59) where we become social media friends with our coworkers and post pictures to an indistinct flock of followers. With different social groups mixing on the Internet, separating online and offline identities and meeting our and others' expectations regarding privacy becomes more difficult and consequential (60).

Malleability and influence

Whereas individuals are often unaware of the diverse factors that determine their concern about privacy in a particular situation, entities whose prosperity depends on information revelation by others are much more sophisticated. With the emergence of the information age, growing institutional and economic interests have developed around disclosure of personal information, from online social networks to behavioral advertising. It is not surprising, therefore, that some entities have an interest in, and have developed expertise in, exploiting behavioral and psychological processes to promote disclosure (61). Such efforts play on the malleability of privacy preferences, a term we use to refer to the observation that various, sometimes subtle, factors can be used to activate or suppress privacy concerns, which in turn affect behavior.

Default settings are an important tool used by different entities to affect information disclosure. A large body of research has shown that default settings matter for decisions as important as organ donation and retirement saving (62). Sticking to default settings is convenient, and people often interpret default settings as implicit recommendations (63). Thus, it is not surprising that default settings for one's profile's visibility on social networks (64), or the existence of opt-in or opt-out privacy policies on websites (65), affect individuals' privacy behavior (Fig. 3).

In addition to default settings, websites can also use design features that frustrate or even confuse users into disclosing personal information (66), a practice that has been referred to as "malicious interface design" (67). Another obvious strategy that commercial entities can use to avoid raising privacy concerns is not to "ring alarm bells" when it comes to data collection. When companies do ring them—for example, by using overly fine-tuned personalized advertisements—consumers are alerted (68) and can respond with negative "reactance" (69).

Various so-called "antecedents" (70) affect privacy concerns and can be used to influence privacy behavior. For instance, trust in the entity receiving one's personal data soothes concerns. Moreover, because some interventions that are intended to protect privacy can establish trust, concerns can be muted by the very interventions intended to protect privacy. Perversely, 62% of respondents to a survey believed (incorrectly) that the existence of a privacy policy implied that a site could not share their personal information without permission (40), which suggests that simply posting a policy that consumers do not read may lead to misplaced feelings of being protected.

Control is another feature that can inculcate trust and produce paradoxical effects. Perhaps because of its lack of controversiality, control has

Box 1. Privacy: A modern invention?

Is privacy a modern, bourgeois, and distinctly Western invention? Or are privacy needs a universal feature of human societies? Although access to privacy is certainly affected by socioeconomic factors (87) [some have referred to privacy as a "luxury good" (15)], and privacy norms greatly differ across cultures (65, 85), the need for privacy seems to be a universal human trait. Scholars have uncovered evidence of privacy-seeking behaviors across peoples and cultures separated by time and space: from ancient Rome and Greece (39, 88) to preindustrialized Javanese, Balinese, and Tuareg societies (89, 90). Privacy, as Altman (91) noted, appears to be simultaneously culturally specific and culturally universal. Cues of a common human quest for privacy are also found in the texts of ancient religions: The Quran (49:12) instructs against spying on one another (92); the Talmud (Bava Batra 60a) advises home-builders to position windows so that they do not directly face those of one's neighbors (93); the Bible (Genesis, 3:7) relates how Adam and Eve discovered their nakedness after eating the fruit of knowledge and covered themselves in shame from the prying eyes of God (94) [a discussion of privacy in Confucian and Taoist cultures is available in (95)]. Implicit in this heterogeneous selection of historical examples is the observation that there exist multiple notions of privacy. Although contemporary attention focuses on informational privacy, privacy has been also construed as territorial and physical, and linked to concepts as diverse as surveillance, exposure, intrusion, insecurity, appropriation, as well as secrecy, protection, anonymity, dignity, or even freedom [a taxonomy is provided in (9)].

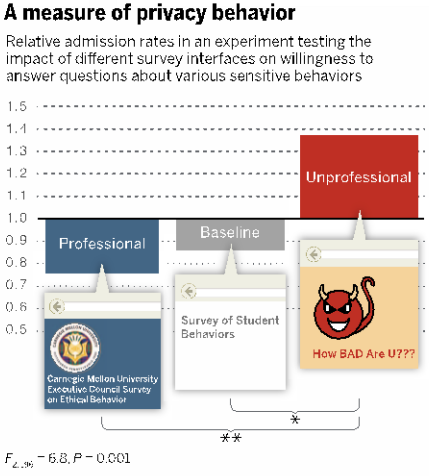


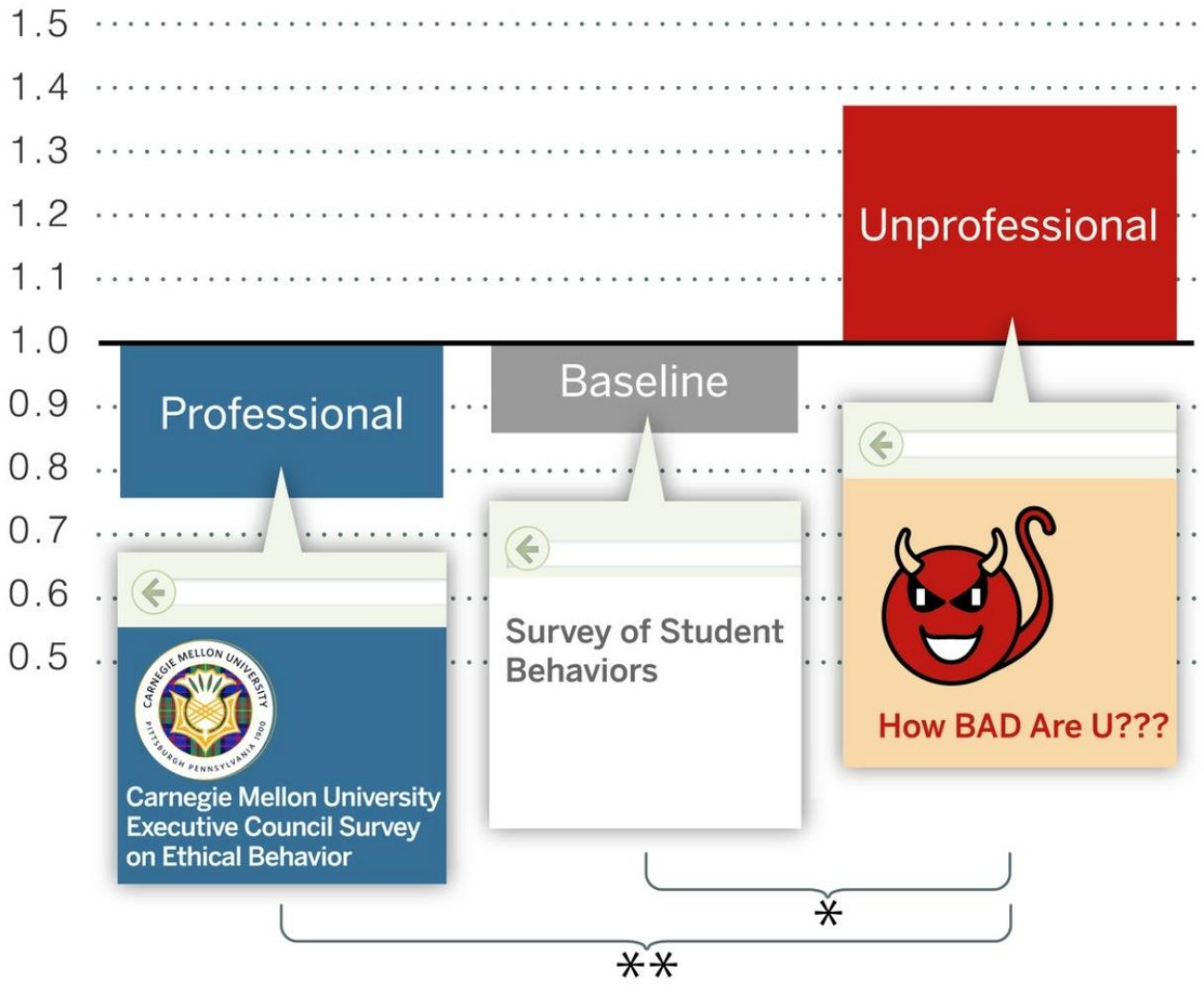
Fig. 2. The impact of cues on disclosure behavior. A measure of privacy behavior often used in empirical studies is a subject's willingness to answer personal, sometimes sensitive questions—for instance, by admitting or denying having engaged in questionable behaviors. In an online experiment (47), individuals were asked a series of intrusive questions about their behaviors, such as "Have you ever tried to peek at someone else's e-mail without them knowing?" Across conditions, the interface of the questionnaire was manipulated to look more or less professional. The y axis captures the mean affirmative admission rates (AARs) to questions that were rated as intrusive (the proportion of questions answered affirmatively) normed, question by question, on the overall average AAR for the question. Subjects revealed more personal and even incriminating information on the website with a more casual design, even though the site with the formal interface was judged by other respondents to be much safer. The study illustrates how cues can influence privacy behavior in a fashion that is unrelated, or even negatively related, to normative bases of decision-making.

Visualization February 23

A measure of privacy behavior

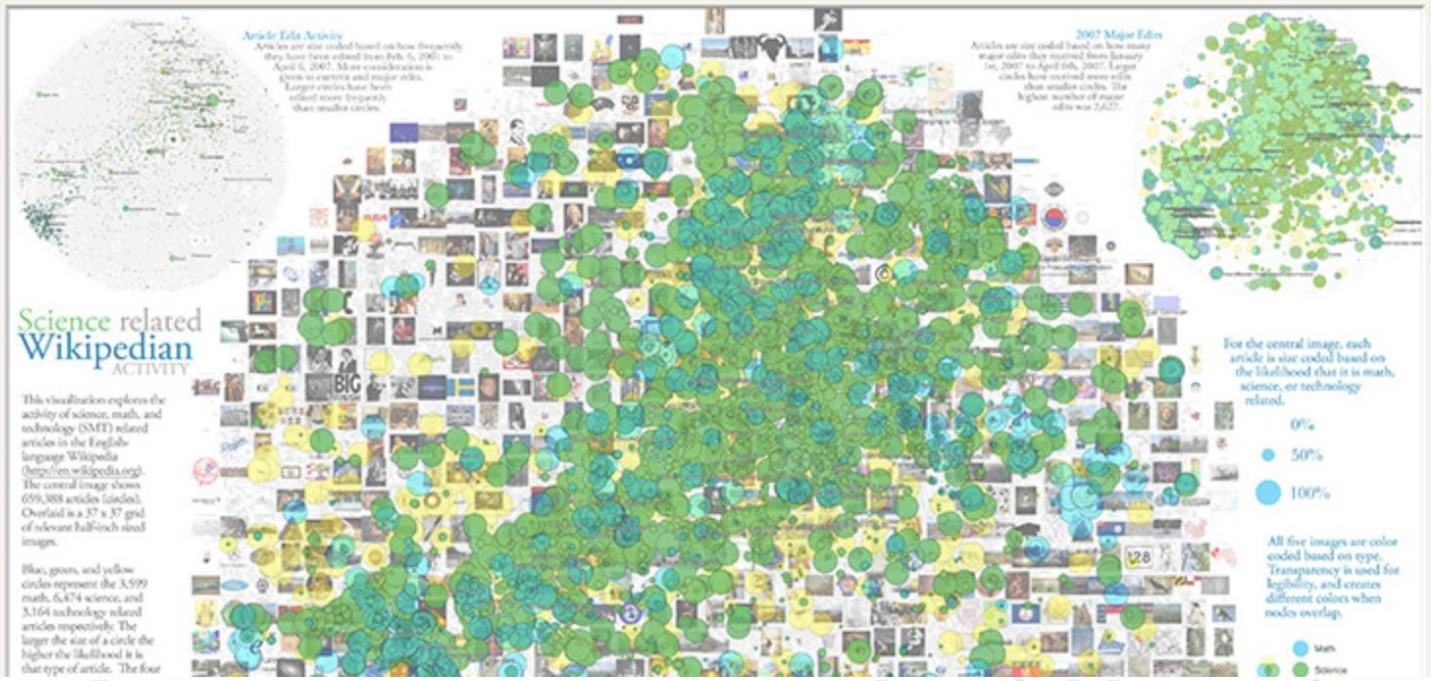
Relative admission rates in an experiment testing the impact of different survey interfaces on willingness to answer questions about various sensitive behaviors

Science, Jan 30



$F_{2,196} = 6.8, P = 0.001$

III.8 Science-Related Wikipedian Activity



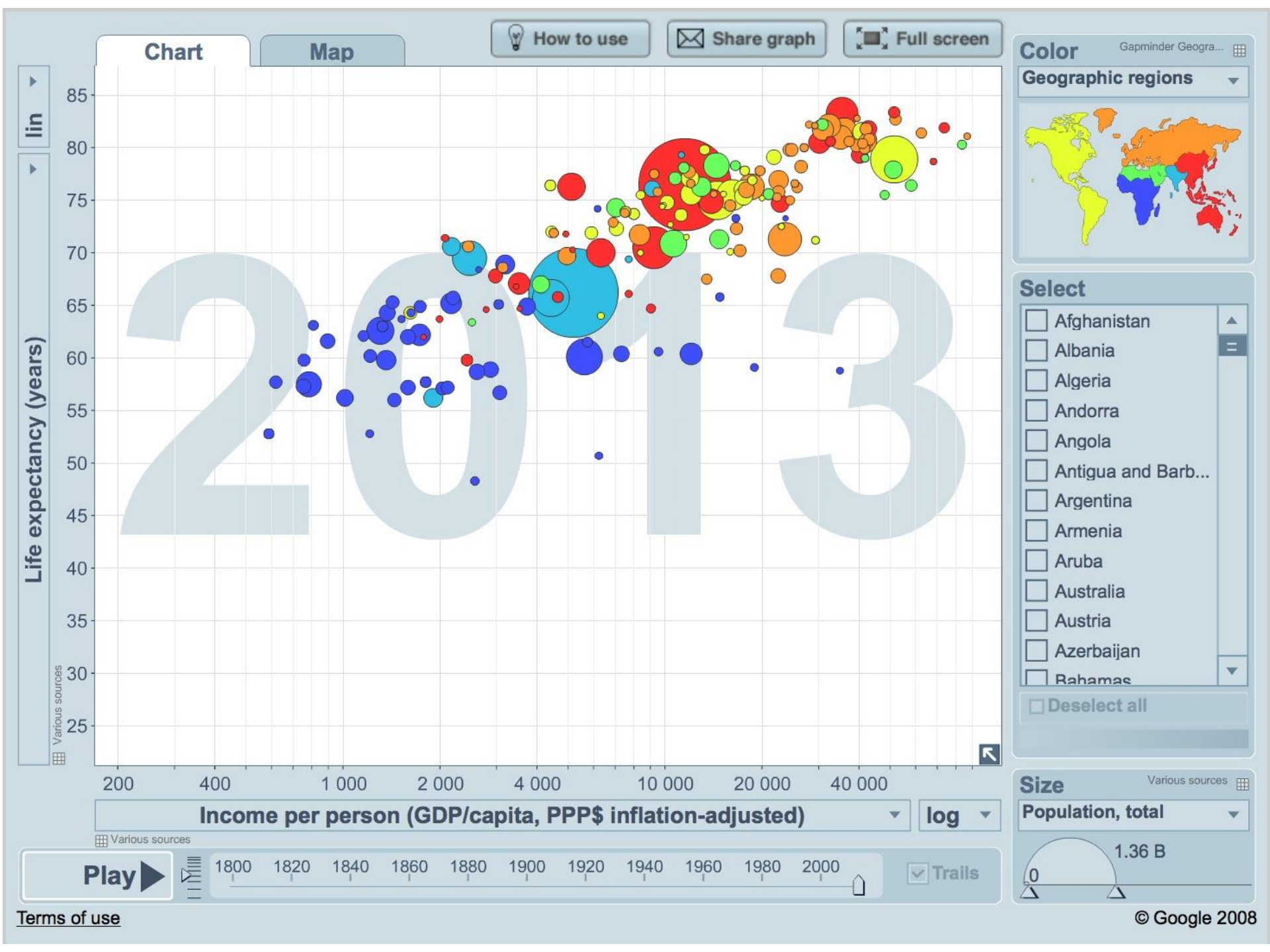


Chart
Map
How to use
Share graph
Full screen
lin
Life expectancy (years)
Various sources
85
80
75
70
65
60
55
50
45
40
35
30
25
2013
200
400
1 000
2 000
4 000
10 000
20 000
40 000
Income per person (GDP/capita, PPP$ inflation-adjusted)
log
Various sources
Play
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Trails
Color
Gapminder Geogra...
Geographic regions
Select
Afghanistan
Albania
Algeria
Andorra
Angola
Antigua and Barb...
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas
Deselect all
Size
Various sources
Population, total
1.36 B
0
Terms of use
© Google 2008

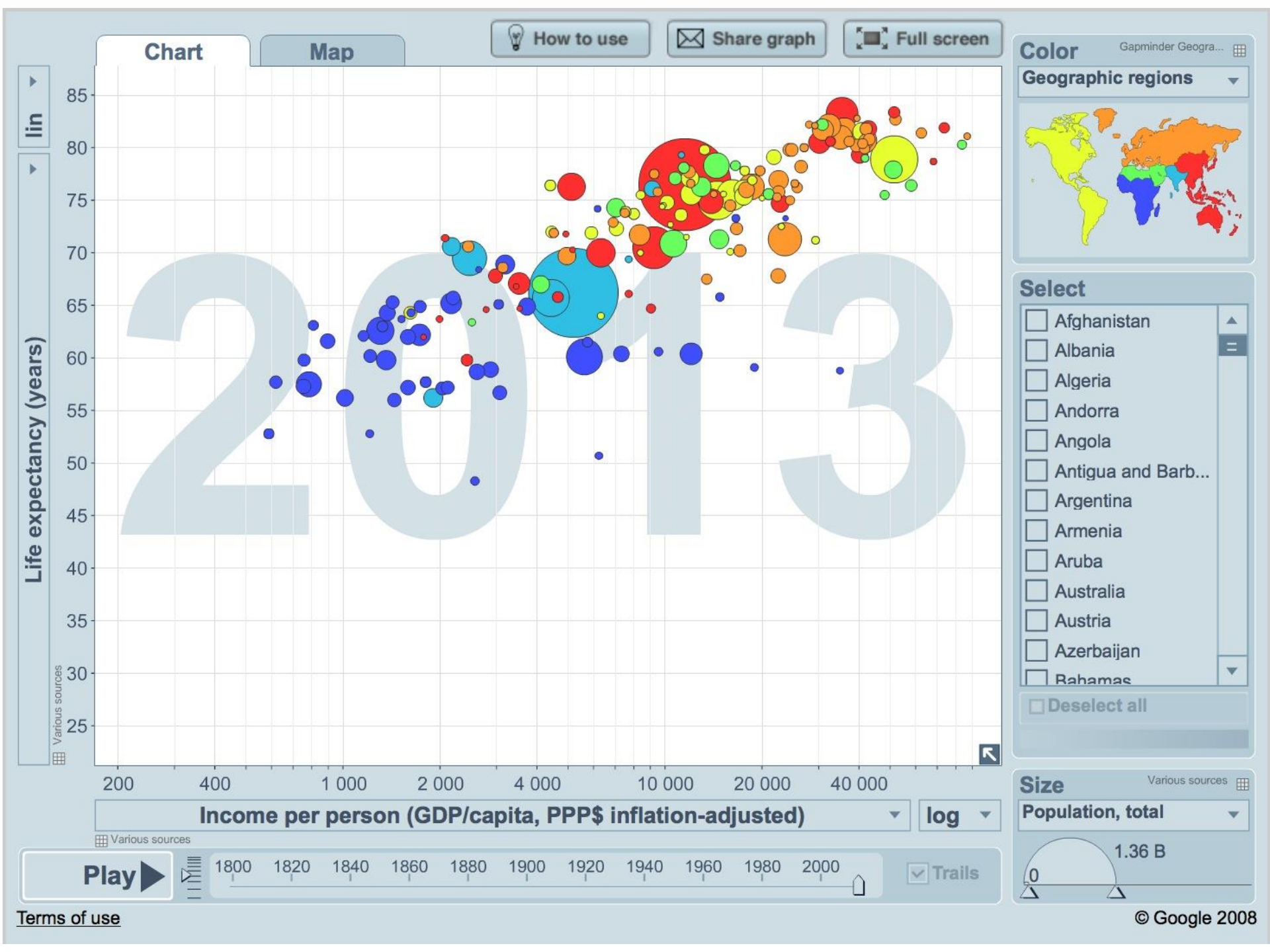
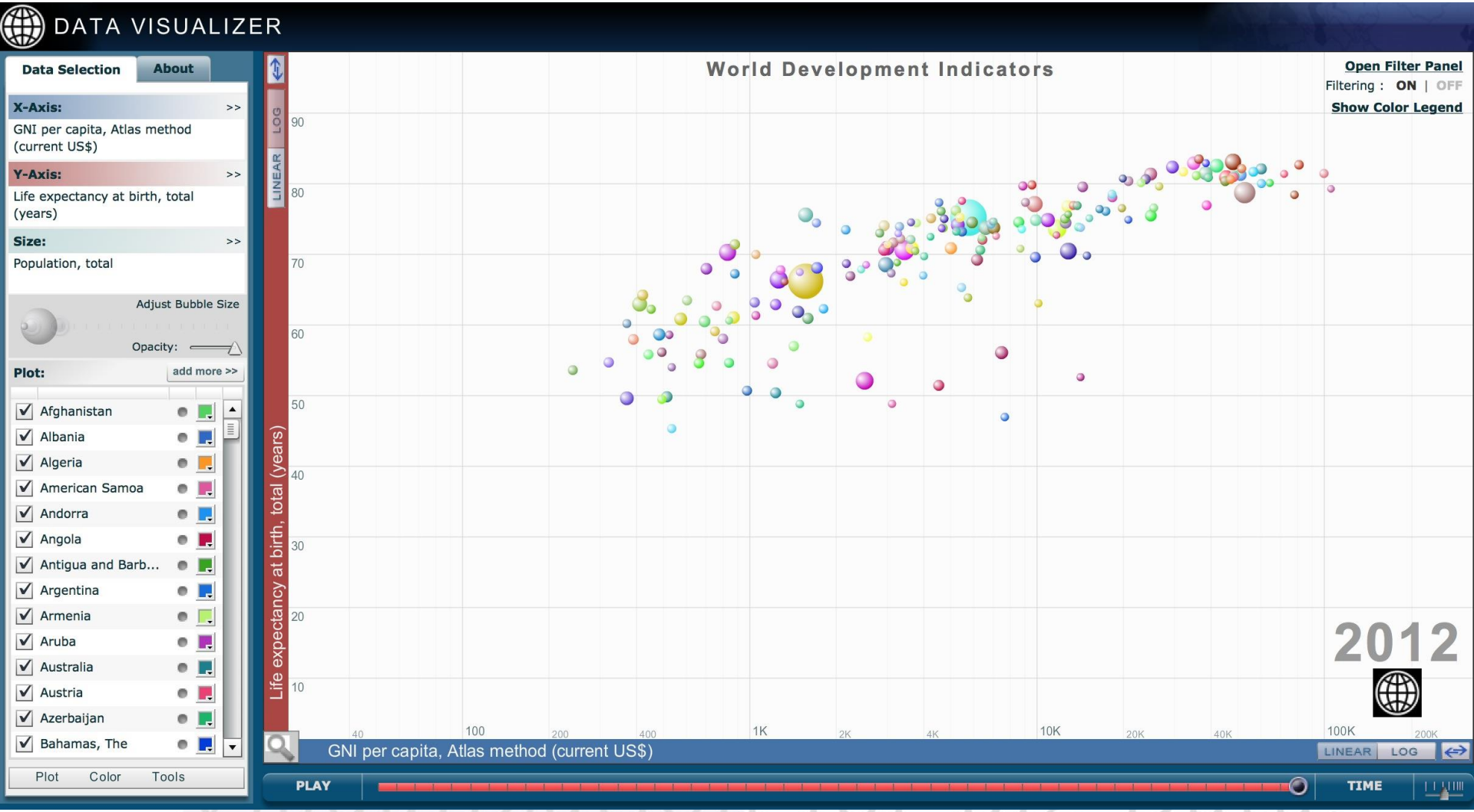
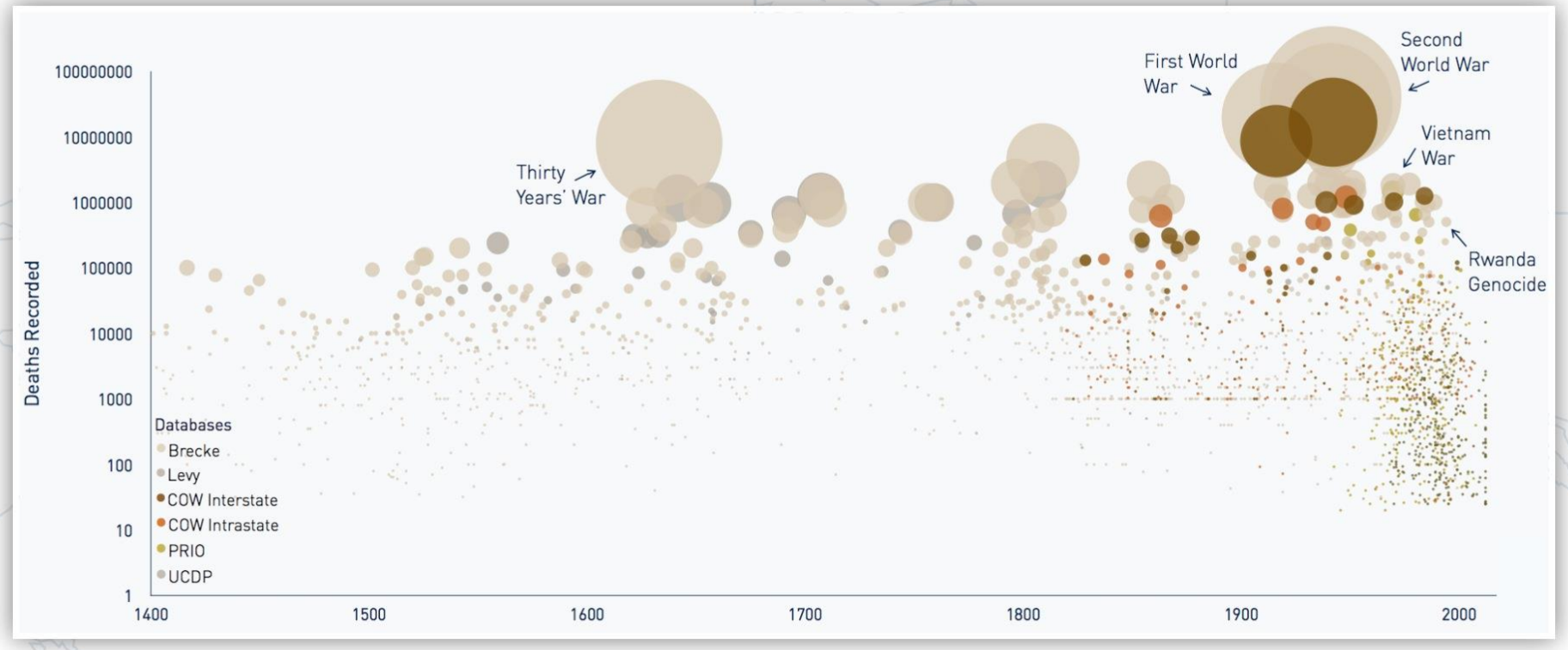


Chart
Map
How to use
Share graph
Full screen
lin
Life expectancy (years)
Various sources
85
80
75
70
65
60
55
50
45
40
35
30
25
2013
200
400
1 000
2 000
4 000
10 000
20 000
40 000
Income per person (GDP/capita, PPP$ inflation-adjusted)
log
Various sources
Play
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Trails
Color
Gapminder Geogra...
Geographic regions
Select
Afghanistan
Albania
Algeria
Andorra
Angola
Antigua and Barb...
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas
Deselect all
Size
Various sources
Population, total
1.36 B
0
Terms of use
© Google 2008



World Bank Data Visualizer

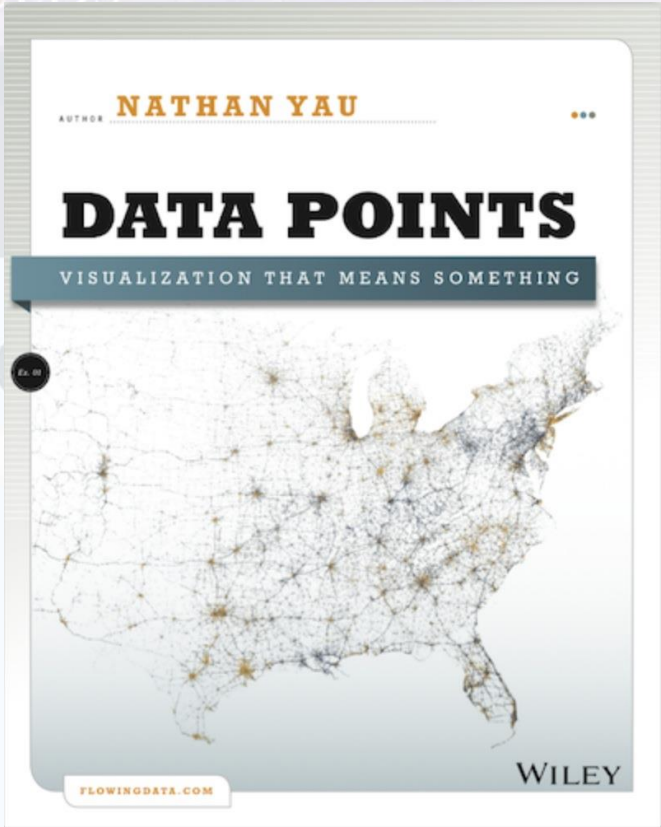
This is also what the record of war and genocides shows - in the recent past, violence is on the decline.



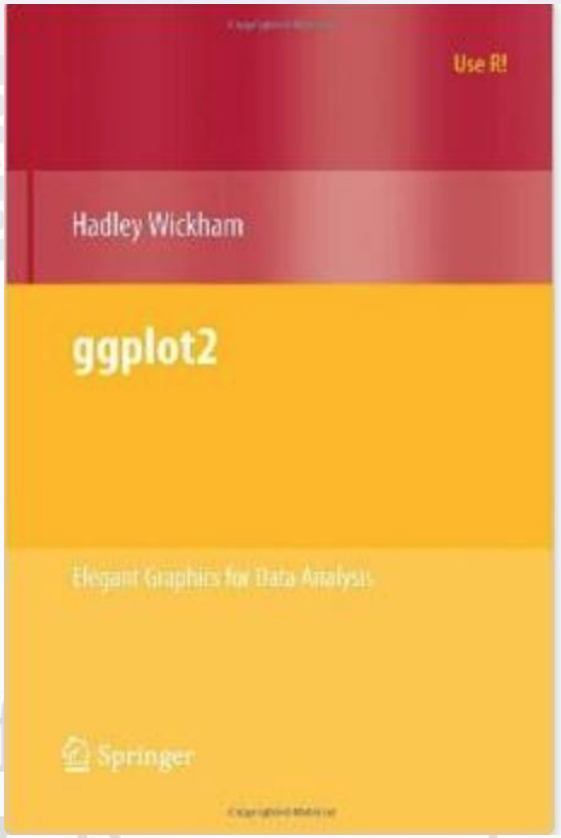
Shown are the annual fatalities of conflicts - in absolute numbers (y-axis) and as a percentage of the world population (size of bubble)

Our World in Data

Visualization for Big Data:
Strategies and Principles



2013



2009

REGISTRATION CLOSED: Turning Numbers into Knowledge: Customer Visualization Showcase

Toronto | 25 February, 2015 | 3:00 - 6:00PM EST

Location: