BIG DATA
and
us little people

{slipping its moorings}

David I Sheidlower

Table 1. Weight in kilograms for children and adolescents from birth through age 19 years and number of examined persons, mean, standard error of the mean, and selected percentiles, by sex and age: United States, 2007–2010

Sex and age[1]	Number of examined persons	Mean	Standard error of the mean	Percentile								
				5th	10th	15th	25th	50th	75th	85th	90th	95th
Male				Kilograms								
Birth to 2 months	82	5.2	0.09	3.5	3.8	4.0	4.5	5.4	5.8	6.0	6.1	6.3
3–5 months	111	7.1	0.11	5.6	5.8	6.0	6.2	7.0	7.5	8.1	8.4	8.6
6–8 months	103	8.7	0.13	7.2	7.5	7.7	8.1	8.7	9.2	9.5	10.0	10.6
9–11 months	108	9.6	0.15	†	8.1	8.2	8.7	9.3	10.5	11.1	11.2	11.7
1 year	318	11.3	0.09	8.9	9.4	9.7	10.2	11.4	12.2	12.8	13.0	13.8
2 years	325	14.0	0.14	11.3	11.8	12.2	12.6	13.9	15.0	15.7	16.4	17.3
3 years	203	16.2	0.21	13.1	13.6	14.0	14.7	15.9	17.2	17.8	18.3	20.2
4 years	244	18.5	0.22	14.9	15.2	15.6	16.2	17.7	19.5	21.5	23.0	25.5
5 years	205	21.2	0.32	16.3	17.1	17.8	18.9	20.2	22.1	24.4	27.1	31.0
6 years	193	24.3	0.47	16.8	18.1	19.1	20.3	23.1	25.6	29.1	35.0	36.5
7 years	216	26.7	0.43	19.7	20.3	21.1	22.2	24.9	29.6	33.5	34.8	39.5
8 years	211	31.3	0.70	21.6	22.7	23.8	25.8	29.8	35.3	39.4	42.1	46.5
9 years	191	36.6	1.17	22.8	24.6	26.6	28.3	33.0	42.8	49.2	52.6	†
10 years	197	40.0	0.93	25.8	28.6	30.2	32.1	38.0	46.1	52.2	54.0	59.7
11 years	211	46.6	1.13	31.5	33.7	34.5	35.9	42.4	54.4	60.8	66.8	73.5
12 years	159	51.5	1.27	31.3	34.5	36.6	39.9	49.2	60.7	68.0	69.7	75.1
13 years	146	59.2	1.45	38.0	42.7	44.6	48.5	56.6	63.6	71.4	81.0	95.8
14 years	177	63.9	1.92	40.2	42.3	47.0	52.2	60.4	71.2	81.2	88.9	†
15 years	160	70.1	1.95	48.4	50.6	53.9	56.2	66.0	76.4	86.9	94.2	115.3
16 years	175	75.1	1.52	52.0	56.2	57.7	62.3	70.7	84.4	92.9	99.8	108.9
17 years	188	77.4	2.41	52.3	58.1	60.9	64.4	71.0	85.4	95.3	104.8	128.5
18 years	142	81.3	2.00	56.7	60.1	61.9	66.5	78.3	90.3	95.8	106.1	†
19 years	179	79.5	1.39	57.6	62.3	63.6	66.9	76.6	85.6	95.8	98.8	119.1
Female												
Birth to 2 months	82	5.0	0.11	3.0	3.6	4.2	4.5	5.1	5.4	5.6	5.8	6.0
3–5 months	104	6.7	0.07	5.2	5.4	5.7	6.1	6.8	7.2	7.4	7.7	7.9
6–8 months	103	8.1	0.07	6.7	7.0	7.2	7.4	8.0	8.7	8.9	9.3	9.6
9–11 months	119	9.0	0.12	7.3	7.6	7.8	8.4	8.9	9.7	10.0	10.2	10.7
1 year	297	10.9	0.10	8.7	8.9	9.2	9.7	10.7	11.7	12.2	12.7	13.1
2 years	282	13.4	0.15	10.9	11.2	11.7	12.1	13.2	14.4	14.9	15.6	16.5
3 years	191	15.7	0.24	12.3	12.8	13.1	13.8	15.4	16.9	18.1	18.9	19.7
4 years	200	17.7	0.21	14.2	14.9	15.1	15.8	17.1	18.9	20.4	22.0	23.1
5 years	177	21.1	0.54	15.9	16.6	17.4	18.1	19.7	22.4	24.2	26.4	†
6 years	177	23.6	0.47	18.3	18.7	19.2	20.4	22.5	25.2	27.6	30.3	33.5
7 years	207	26.8	0.53	19.4	20.6	21.0	22.1	25.2	28.9	32.8	35.4	39.6
8 years	203	31.9	1.01	21.8	23.1	23.7	25.0	29.2	37.8	42.4	43.6	48.0
9 years	205	35.5	0.98	23.3	26.1	26.5	28.0	31.9	40.1	46.1	49.3	55.7
10 years	183	41.1	0.74	27.3	29.1	30.1	33.6	39.0	45.3	52.3	55.0	61.3
11 years	219	47.5	1.28	30.0	31.5	33.0	36.1	43.3	56.5	63.0	66.8	79.6
12 years	166	52.3	1.26	30.6	34.8	38.0	43.1	51.4	60.3	65.7	72.1	75.7
13 years	140	56.8	1.41	38.8	41.4	44.6	46.1	52.3	64.8	72.4	75.7	83.2
14 years	168	61.6	1.13	42.7	44.8	47.7	51.1	59.0	67.1	73.5	76.4	†
15 years	137	63.3	1.32	44.2	46.5	48.2	54.5	59.5	70.6	77.1	85.8	92.3
16 years	156	62.4	1.11	45.5	48.4	50.0	51.3	58.7	69.5	77.9	84.8	92.0
17 years	143	63.7	1.56	45.4	47.9	49.1	51.8	60.8	68.4	80.0	83.8	92.0
18 years	137	65.4	1.67	46.6	50.6	51.8	53.5	58.6	69.5	80.4	85.8	105.0
19 years	118	68.0	2.00	44.4	47.1	51.0	55.1	62.8	76.8	88.2	94.9	106.7

† Standard error not calculated by SUDAAN.
[1]Refers to age at time of examination.

NOTE: Pregnant females were excluded.
SOURCE: CDC/NCHS, National Health and Nutrition Examination Survey.

Fryar CD, Gu Q, Ogden CL. Anthropometric reference data for children and adults: United States, 2007–2010. National Center for Health Statistics. Vital Health Stat 11(252). 2012.

BIG DATA
and
us little people
{slipping its moorings}

David I Sheidlower

The Coincidence Press

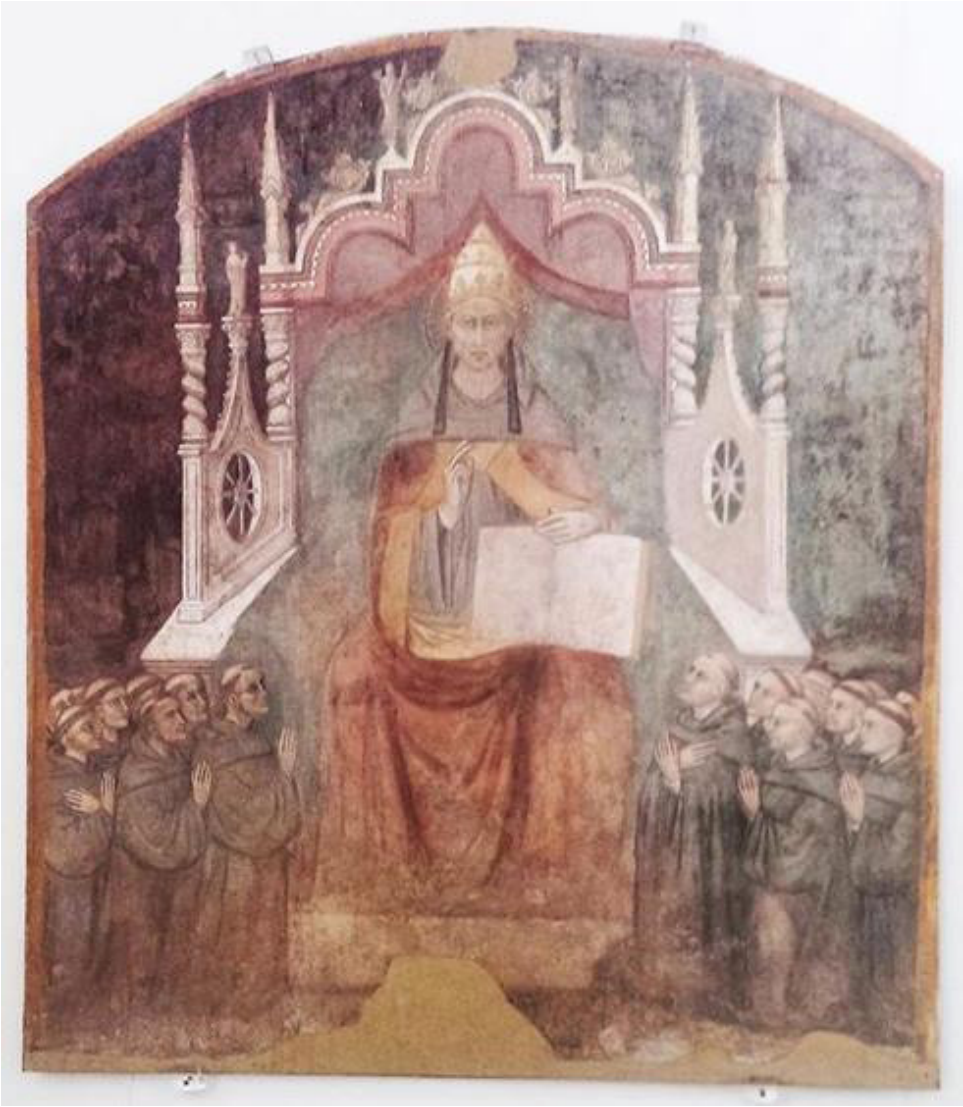
© David Sheidlower 2015
Earlier versions of this work appeared in SecurityCurrent

All rights reserved

TABLE OF CONTENTS

Inadequacy	7
The Point Beside the Point	12
Anti-Viral	18
For Whom the Bell Curve Tolls	24
The Question of the Questions	34
Afterthoughts	46

1



Fresco attributed to Niccolò di Tommaso c. 1365. Photo taken by the author at Castel Nuovo, Naples, Italy

Inadequacy

In previous work I wrote about the principles of data security and privacy. Many authorities charged with enforcing data protection accept the principles. They are based on the idea that the actors in data transactions (i.e., subjects, collectors, disclosers, users and regulators) all have a role to play in creating and maintaining the world of data.

The argument goes that if we all understand each other's roles relative to any given data point then the world of data itself, the data-centric system, can be understood.

What if that's not entirely true?

At some level, it is of course accurate. And it is tempting to believe it. Our very experience of data creation when we are the subjects of a data point reinforces that idea: that a data point is just that. Extending the metaphor of a "point" from Euclid's Geometry, we think of a series of data points as forming a line, a trend.

A data-centric system might just be a collection of individual transactions that form a static snapshot. When the points are connected like so many dots, they form a trend line.

In this view, the system as a whole is the total collection of stored records and we only need to deal with how individuals and organizations interact with those records. It is a role based model for data access and control and it has been mostly adequate for our purposes.

Privacy advocates have worked along these lines since data was understood to be sensitive. Regulators also consider this to be the proper paradigm. The collectors, disclosers and users of data have all worked to define and understand their roles in light of regulations and ethical

considerations.

But two recent events call into question whether we can see data as just records.

The events involve separate court cases in different courts, one brought by Mario Costeia Gonzalez of Spain and the other David Leon Riley of the United States. The former won a case in the Court of Justice for the European Union and established a "right to be forgotten." The latter won a case in the United States Supreme Court in which the justices acknowledged that the data on a cell phone was unique because the device itself functioned as an aggregator.

In both cases, what the courts recognized was the concept that data aggregation is distinct from the transactions it aggregates and that by itself, the aggregation of data could be controlled apart from the transactions themselves.

We are used to thinking of data as a static snapshot. Perhaps only valid for the time the snapshot is taken, but still fixed for that moment. A data extract, an analysis, the results of a search, these are all the product of computer processing and are thought to be objective and, to a large extent, without bias except the bias in the data itself.

This is not to say that analysis or search terms cannot also contain bias, but the general belief is that the procedures which process and store data do not introduce bias into the content.

Credit scoring, one of the earliest example of "Big Data" is regulated by Federal regulations and those regulations explicitly reflect that belief. Assuming the data are collected appropriately, then the credit scoring model, by law, must be "empirically derived, demonstrably and statistically sound" Equal Credit Opportunity Act 1974 (Regulation B), Section 202.2(p). Even when bias has been found in these models, they are traced to bias in the inputs or the model's algorithms, not the acts of aggregation, processing and storing the data.

Jacques Valle in *The Network Revolution* (1982) expanded on Marshall McLuhan's "global village" from *The Gutenberg Galaxy* (1962). Valle, writing before the Internet exploded, described the emergence of a great communication network that presented new possibili-

ties for individuals to have access to one another. Valle and McLuhan were concerned mostly with communication, not data; with access, not aggregation.

What if data aggregation and processing has the properties of mediums like radio, print or television? Communication theorists might argue that radio, print and television all assume a sender and a receiver whereas information-processing models do not assume that the action of "input" *intends* to communicate through the action of "output." In other words, data isn't necessarily sent, it's processed and stored. Data being output? Sometimes, but it is not necessary.

Others have noticed the similarity and overlap between information processing and communication models. My basic premise begins from the work of Harold Innis in the 1950's and Marshall McLuhan in the 1960's. The premise that the medium itself creates bias is, in other words, not a new idea. I'm going to try to apply that idea to data-centric systems and see where it leads.

This volume intends to look at the following aspects of a data-centric system:

- ☞ Aggregation is a form of narrative (search engines create narratives)
- ☞ Narratives are not snapshots (analysis and search results are more than snapshots)
- ☞ Being part of a cohort is a new state of identity (regressing towards the mean might be a problem if you have no say in the terms used to create that mean)
- ☞ Search engines and data analysis are a medium as powerful as any that have come before it (successful media create monopolies of knowledge)
- ☞ All medium have bias

This is not a call to action. It is not a manifesto declaring that there is a moral truth to one approach to data or another. It is an attempt to de-

scribe a transition that is in progress but has not yet fully taken hold. It is recognition that while what Innis and McLuhan recognized as the inherent bias of using a phonetic alphabet has not been fully replaced by what McLuhan called at the time the "electronic age," there are things that are worth observing now.

Why would a security professional care about this? The short answer is that we protect data and so the more we understand about it, the better we can protect it. But I can provide a more thorough answer than that.

Just like the network perimeter of the enterprises we protect is no longer a walled-in fortress, the data we protect is no longer just a single flat file or set of related tables sitting on a server in the center of that fortress.

You can only really protect an information asset you understand. You can only guard against attacks that you can imagine. In the one section of his book that zeros in on databases, Valle sums up why Security professionals should care how data aggregation and processing is evolving: "Anything that can be designed by human logic can be fooled by human ingenuity."

If you want to truly protect data (even your own), you are going to have to accept that sometimes it slips its moorings.

2

North Atlantic Ocean, 39.355504, -74.424251. Photo by the author, standing at about 39.355634, -74.425264

The Point Beside the Point

If we live in a data-centric world, it is still true that data are more immediate for some than for others. To use one individual as an example: by the fifth grade, Ben was an excellent data analyst. His life depended on it.

Ben went to grade school with my younger son and Ben has type 1 Diabetes. I remember him explaining to his mother one day how he had, that morning, manipulated his intake of carbohydrates, monitored his blood sugar and activity levels, and carefully calculated what amount of insulin he should take.

He was extra careful and deliberate so he could have one of the cupcakes that a classmate had brought in for her birthday. Not only did Ben's life depend on his ability to collect and analyze data, the normalcy of his life depended on it. (Still does: Ben is in High School now.)

As his mom describes it:

> "Until fifth grade, he had an adult guiding him to make decisions about how much insulin he needed based on his blood glucose number or the amount of carbs. Also, until about fifth grade, we put a note in his lunch saying how many grams of carbs each item was.
>
> He did, though, do his own checks, and had to enter that number into the [insulin] pump sometimes. He also had to enter the correct amount of insulin into the pump (which he could read on the screen as he pushed buttons, up or down).
>
> About First or second [grade], we set up the "Bolus Wizard"

> on his pump. That meant he had to enter the total amount of carbs eaten and the blood glucose number and the pump did the math and suggested an amount of insulin... The analysis that he always had to do, and that has always been hard to train other adults to do, is to consider what was eaten and how recently, in combination with what type of activity he has been doing, and how much he is about to eat.
>
> All this information is what my dad called in frustration, "fuzzy math". This is the part of D[iabetes] management that is an art, and not easily programmed or put in a chart. He became his own Bolus Wizard in this regard by about 6th grade. Also, by about 5th, I'd say, he was noticing patterns of highs and lows and talking about how to make adjustments to the programmed basal rate and bolus ratios in his pump."

For every other kid in the class, the only thing required to have a cupcake was to say "thank you" when it was handed to them.

When data relate to people, the subject of each data point is a unique individual. Even if the data point is the result of analysis, the raw data (the source observations) that went into the analysis illustrate the uniqueness of the subject. The subject is the only actor whose participation with the data point must be simultaneous with the initial observation that creates that data point.

When you're moving quickly through space, things close up are hard to see clearly when you pass them. Things farther away are easy to focus on. When you're sitting still, you can choose to focus on the details of what's in front of you or in the distance.

The data-centric world, the digital world of observed events, as seen by the subject is the opposite of that. Individual transactions, the immediate foreground of experience recorded as data, are easy to spot. It's the background , the pattern or the aggregate that is hard to make out. Likewise, you can always step back from the transaction and choose to focus on the foreground or the background.

Economists have used this characteristic of perception to help

divine people's motives. Their goal has been to determine how people make significant decisions, some of which might be against the individual's long term best interest. They describe the impulse to see only the immediate transaction/decision as being motivated by a desire for "instant gratification":

> "When making decisions with immediate consequences, economic actors typically display a high degree of impatience. Consumers choose immediate pleasures instead of waiting a few days for much larger rewards. Consumers want "instant gratification."
>
> However, people do not behave impatiently when they make decisions for the future. Few people plan to break their diets next week. Instead, people tend to splurge today and vow to exercise/diet/save tomorrow. From today's viewpoint, people prefer to act impatiently right now but to act patiently later." ("Impatience and Savings", David Laibson, Economics professor, Harvard University)

Dr. Laibson goes on to point out that how different brain systems are invoked in making decisions that he labels "short-run" or based in the present or "long-run" or based in the future:

> "Data from neuroscience experiments provide a potential explanation for these observations: short-run decisions engage different brain systems from long-run decisions. Using functional magnetic resonance imaging (fMRI), Samuel McClure, George Loewenstein, Jonathan D. Cohen, and I have shown that decisions that involve at least some short-run tradeoffs recruit both *analytic and emotional* brain systems, whereas decisions that only involve long-run tradeoffs primarily recruit analytic brain systems. These findings suggest that people pursue instant gratification because the emotional brain system - the limbic system - values immediate rewards but only weakly responds to delayed rewards." (same web posting, next paragraph, emphasis mine)

While this argument may be valid for describing how people make certain kinds of decisions, it also makes clear that the mind's perception of the single data point that relates to the present (what I will eat now) is different from the perception of the aggregate (how my diet is going).

There is also no reason to believe that these two different ways of seeing an event or action is limited to only what economists might wish to measure. In other words, the difference in perceiving actions is not just in how people ultimately judge them when deciding, but how they see them in the first place.

Security professionals recognize this concentration on the present action coupled with a lack of vision of a bigger picture (in this case a policy framework) as one of the strongest explanations for why social engineering schemes succeed. But that's not my focus here.

Here I am attempting to differentiate between data creation and data persistence. People's actions create recordable data continuously. And whether making a purchase or taking a blood test, the subject of the data is at the immediate center of the creation of the data point. It is what occurs at that moment that data points are created by being recorded that give the data-centric world its most unique characteristic.

The data point is persisted, stored, recorded. It is completely removed from the sensory experience of the subject it refers to and so it is completely removed from its original context unless, like Ben's record of his blood glucose level, it is only recorded and used by the subjects themselves.

Otherwise, the record of the blood test results does not refer to the needle or how easy or hard the blood draw was (or thousands of other circumstances of collecting the blood sample). The record of the purchase does not refer to which decision making mechanisms of the purchaser's brain were more or less engaged. Not yet at least.

It is this removal from the subject that accounts for people's discomfort when arguing about who "owns" the data. It helps explain why, for example, public health officials and privacy advocates do not always agree on who should have access to health records even though both

favor healthy populations. It accounts for the transformation of data points to analysis sometimes with the subject's consent but almost always without the subject's participation.

It begs the question of bias, of what kind of knowledge is beginning to monopolize our perceptions and what are the consequences of that.

3

Poughkeepsie Train Station. September 10, 2015, 5:49 a.m.. Photo by the author

Anti-Viral

Aggregating is the inverse of broadcasting. What complicates this is that many technologies are now used for both. Cell phones are the best example. They are a device originally designed for communication.

Their original purpose was for transmitting information between individuals and they have evolved into one that can broadcast that information via social media. It is ironic that they have now been recognized by the US Supreme Court for their ability to aggregate data regardless of whether or not the individual device transmits it:

> "Cell phones differ in both a quantitative and a qualitative sense from other objects that might be carried on an arrestee's person. Notably, modern cell phones have an immense storage capacity. Before cell phones, a search of a person was limited by physical realities and generally constituted only a narrow intrusion on privacy. But cell phones can store millions of pages of text, thousands of pictures, or hundreds of videos". (RILEY v. CALIFORNIA, No. 13–132. Argued April 29, 2014—Decided June 25, 2014)

The Court goes on to make the understated observation that "This has several interrelated privacy consequences."

Broadcasts have focused points of origin, a single source. Broadcasting works when information, the broadcast, is made available from that source and received by large numbers of people. By definition, aggregations have many sources but are focused in terminating at a single storage point.[1]

Broadcasting has a bias towards celebrity. Aggregating's bias is towards anonymity. The bias of an information-processing medium like aggregating or broadcasting is generally recognized, exploited and enhanced by those that use it. It is not a coincidence that the individuals who described this best, Harold Innis and Marshall McLuhan, had experienced the use of radio as propaganda leading up to and during World War II and, at least in McLuhan's case, studied the rise of broadcast television as it grew in popularity and influence from the 1950's to the 1970's.

Aggregating's bias is towards anonymity. What the Supreme Court acknowledged with the decision quoted above and what the EU Court of Justice also recognized with their decision regarding the right to be forgotten is that when such powerful information processing names a single individual, then that individual's privacy may be compromised.

In other words, sometimes that anonymity is not assured. For example, when the aggregation is focused on an individual as in the case of a device (like a cell phone) or some specific algorithms (like a search result). Even when data are depersonalized, that anonymity cannot be taken for granted when the focus is on the individual. Dr. Latanya Sweeney demonstrated that almost 20 years ago when she singled out and identified a single individual's health record in a statewide database of depersonalized records.

There are many privacy efforts aimed at addressing this. There are the court decisions above that focused on individual devices or data points. There are laws, policies and even data generalization and suppression solutions like those described in Dr. Sweeney's k-anonymity approach. They are all generally seen as privacy protections.

In fact, these efforts are protections of an individual's identity. But they also inadvertently serve to reinforce the bias of aggregation, the bias towards anonymity. In other words, the fact that aggregation tends to work against information being individually identifiable is part of the reason why, when it does not, that it stands out as a threat to privacy and has received such attention.

(The other, less understood reason that aggregating has a bias

towards anonymity is the power of aggregation to take individuals and form them into a cohort. Aggregating is used here to mean more than mass storage however. It is more than, for example, a production transaction database, which serves a front-end application. It is the bringing together of data in order to focus on it. The relationship between identity and cohorts will be looked at in depth in my next chapter).

It may seem counterintuitive to suggest that privacy protections, the controls that help keep sensitive data from being identified, actually strengthen the argument that handling those very data is biased towards anonymity. Understanding this requires a closer look at those protections.

The right to be forgotten involves an individual's right to suppress data points about them in an identifiable context. The Court decision around this and the EU Directive on Data Protection contain the details of how the right to be forgotten supports the argument that the bias of aggregation is anonymity and that privacy protections reinforce that.

The EU Court of Justice acknowledged that the data point itself might be accurate. Therefore the right to forget a data point establishes that sometimes the data point can "appear to be inadequate, irrelevant or no longer relevant or excessive in the light of the time that had elapsed" with respect to the subject that seeks to have it forgotten.

In the case of Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González, the Court did not order that the data point in question be deleted. They did not order that the data be unavailable for all search results.

They specifically found that:

> ...following a request by the data subject pursuant to Article 12(b) of Directive 95/46, that the inclusion *in the list of results displayed following a search made on the basis of his name* of the links to web pages published lawfully by third parties and containing true information relating to him personally is, at this point in time, incompatible with Article 6(1)(c) to (e) of the di-

> rective because that information appears, having regard to all the circumstances of the case, to be inadequate, irrelevant or no longer relevant, or excessive in relation to the purposes of the processing at issue carried out by the operator of the search engine, the information and links concerned in the list of results must be erased. " (EU decision, emphasis mine)

In other words, provided that the search engine does not include the data point in searches that explicitly name Mr. González, then the data point is fair game to appear in search results.

Consider the EU Data Directive's definition of "processing:"

> (b) "processing of personal data" ("processing") shall mean any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction; (EU Directive 95/46/EC - The Data Protection Directive)

In other words the processing of data which involves aggregating, returning the results of a search, if it does not involve identifying the individual in the initial query is therefore granted the status of not being defined as the processing of personal data. The resulting dataset returned by the search may not be depersonalized, but since the search was, the Court found it did not meet the definition of processing personal data.

This case and its impact on privacy law and practices will no doubt develop and evolve over time. Regardless, it stands as an exceedingly precise example of the acknowledged non-personal nature of aggregation. Anonymity is the effect of aggregation *unless* the aggregation is specifically focused on the individual in some predetermined way. The two predetermined characteristics recognized above are the individual's

ownership of the aggregating technology (e.g., a cell phone) or the individual being named in the parameters of a search query.

The data point in the EU case, the case of a 1998 report of a real estate transaction in Spain, is, on the one hand, personal; it is history; and, by order of the Court, it is to be forgotten.

On the other hand, however, the Court recognized that there were valid reasons to include that data point in a query. For example, no analysis of the state of the Spanish Real Estate Market in the late 90's would be truly complete without including this data point. And the Court's decision in no way prohibits that data point from being in such analysis. In the next chapter, I'll look at why the person performing that analysis couldn't care less whether they capture that one data point or not.

---

[1] Nothing related to distributed broadcasts, high availability, data replication or other support technologies materially impacts these observations because they are accurate descriptions of how broadcasting and aggregating function for their users even if the "plumbing" behind the scenes is different.

4

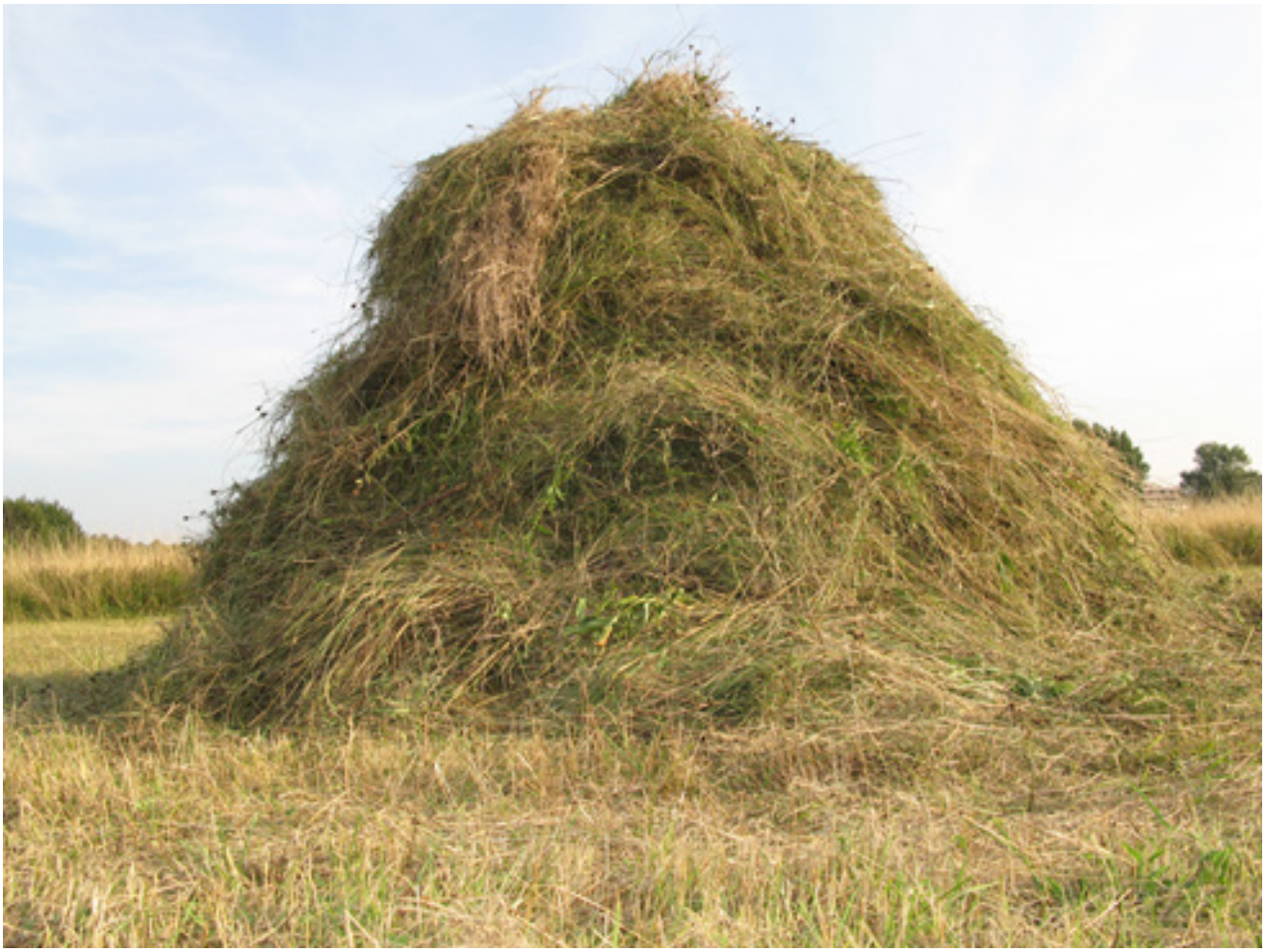


Photo: Kathrin Böhm Courtesy of Haystacks London

For Whom the Bell Curve Tolls

People prefer to choose the groups they are in. Even before social media exploited that, there were fan clubs, fraternities, sororities, and many different kinds of groups that people associated themselves with.

There are also the groups that people don't choose but through birth, prejudice, unforeseen circumstances and/or unwanted diagnoses, they find themselves in nonetheless. Those groups are generally more difficult to leave.

There is a different kind of group that can encompass any of these but does not have to. These groups overlay a different relationship between the group and the individual[1]: the cohort.

With the exception of specific research studies, individuals do not voluntarily join a cohort the way they join other groups. They do not first consider the goals or intentions of those forming the cohort and then, having decided to support them, "sign up."

For example, obtaining a car loan is for the purpose of purchasing a car, not for the purpose of having your loan payment history contribute to the credit-scoring model that will be used to underwrite future loan applications of people like you.

Even when individuals find themselves in a group involuntarily, say in the case of a diagnosis, their relationship to the cohort is more complex than just one of membership. Virtually nobody says "yes, I just got hepatitis and the first thing I am thinking about is how I will now be part of the statistics on people with hepatitis."

I

Aggregation is biased towards anonymity. Its ability to form cohorts creates a monopoly of knowledge that is based on the predictive power of the cohorts that comprise the population. When the population is formed correctly and the aggregation of data that brought it into being is sufficiently broad then, with few exceptions, it will be able to be represented as a normal distribution. Normal distributions, with the majority of a population building up from two extremes to a central mean are graphically represented as bell curves.

In the previous chapter, Anti-Viral, I discussed how aggregation is biased towards anonymity. This bias is so strong that the EU Court of Justice found that the appearance of a named reference to an individual in an aggregated dataset was not to be considered the processing of personal data so long as the search parameters that led to that aggregation did not name the individual.

In other words, your right to be forgotten only applies when the searcher is looking for things about you. The Court defended the searcher's right to have your information contribute to any other cohort.

The phrase "monopoly of knowledge" is discussed at length by Harold Innis, a communication scholar and historian. Writing 60-70 years ago, Innis insisted that how information is processed in a society invariably influences the relationship between what is known, how that knowledge gets created and even who gets to know it.

He distinguished between an oral society based on local discourse, the establishment of an elite literate and learned class that used manuscripts, and the commercialized distribution of printed material. His most famous successor, Marshall McLuhan, extended that idea to the developments of the phonetic alphabet, radio and television. Every monopoly of knowledge that Innis and McLuhan defined had characteristics that were unique as the technologies that enabled them.

Innis believed that advances in information processing technology eventually disrupted whatever the current monopoly of knowledge was:

> "I have attempted to trace the implications of the media of communication for the character of knowledge and to suggest a

> monopoly or an oligopoly of knowledge is built up to the point that equilibrium is disturbed." (Presidential Address to the Royal Society of Canada, 1947)

In this series, I am identifying a disturbance of the current equilibrium/ monopoly. McLuhan predicted this in his discussion of the "electronic age." Philip Levinson, following McLuhan, also has suggested that this disturbance or transition is occurring. Levinson is among those who believe the transition is being caused by generalized advances in communication technology and are centered on the Internet.

I am not trying to discount the communication possibilities that are opened up by the Internet, the cell phone and other "on-line" experiences. But I think the technological advances that are disrupting the current monopoly of knowledge are more specific. Those advances are the combination of three developments:

Massive amounts of data being recorded by computer driven transactions

Ever increasing data storage capacity and data processing power which can accommodate those records

Advances in the understanding of the predictive power and use of aggregated data.

II

Discreet data points concerning a single individual can be plotted out in a line. Even if those data points are just numbers (an individual's height as they aged), the line they form tells a story (they got taller for 16 years and then their growth stopped).

The narrative is unique to the individual and regardless of the trend, our evaluation of whether or not the next data point will conform to that line or not depends at least as much on our knowledge of the individual and our belief in the power of that individual to change their future as it does on our belief in trends.

The right to be forgotten is an acknowledgment that individuals should have some control over what data points go into the creation of those trend lines. But building a linear narrative about the individual is not the only way that the individual's data points are used to build a narrative "about" them.

When an individual's data points are aggregated into a population that is normally distributed[2] and they are then identified as part of a cohort within that population (in our example above: average growth given a person's nationality, age and gender), that line hardly matters. That individual narrative is replaced by the probability of a given narrative for that cohort.

To go from the world of human physiology back to the earlier example of underwriting an auto loan: while the individual's characteristics are what place them in a cohort, it is the cohort's chance of defaulting on an auto loan, not the individual's, that is reflected in a credit score used to underwrite the auto loan application of that individual.

Similarly, it is often the cohort's most likely response to a course of treatment that the physician relies on in ordering/recommending a treatment. It should be noted that no physician would subscribe to that model, classed in the realm of what is called "evidence based medicine," 100% of the time- they rightly reserve the right to consider other factors in their treatment recommendations.

This method of modeling behavior and outcomes is becoming increasingly important to how the world around us interacts with us. Recognizing that, Steven Salzberg, Computer Science professor at Johns Hopkins University, recommends replacing the study of Calculus in High School with the study of statistics and computer science.

The bell curve that represents a normal distribution has increased in importance in people's lives. Yet traditional elementary school curriculum spends more time on different shapes. School children are taught the difference between a scalene and an isosceles triangle in greater detail than the characteristics of a bell curve. I do not mean to denigrate geometry or calculus.

I very much agree with Salzberg that the emphasis of current cur-

riculum's should be re-examined. However, I also am bringing up this disconnect between what is relevant to people's lives with what is on a traditional school curriculum to further illustrate the idea above that the "equilibrium is disturbed."

With that, how we make sense of a person's identity is also changing.

The linear narrative created by an individual's actions are being replaced by the individual's current identity. The narrative is then no longer about how the individual's past might lead up to the individual's next action but about how the individual's present identification with a group, a cohort, more or less predicts their future.

When a thief steals an identity in order to fraudulently apply for credit or use that individual's credit accounts, they are assuming the present identity of that individual. They are not attempting to impersonate someone's past. Regardless of how good the victim's reputation is, all the thief wants are the rights and privileges that come with the victim being in that cohort.

Some would argue that group identity has always played a part in how people were treated by others. Class-based societies judge individuals by the class they belong to and there are tremendous amounts of studies documenting all kinds of discrimination based on people identifying others as belonging to certain groups.

Many of those studies are cohort studies and they employ the same fundamentals of statistics described here. In fact, the statistical techniques used for modeling outcomes and behaviors were developed in large part by social scientists trying to measure significance in limited datasets. Those studies emphasize how the world impacts the individual, not how the individual impacts the world. In that respect, the current uses of Big Data are no different.

There are a few other important characteristics when looking at groups and the monopoly of knowledge created by aggregation of large datasets.

1. NOT JOINING BUT BEING IN ANYWAY: Traditionally, you are aware of your membership in a group. You may have joined or you may have had membership "thrust upon you" but either way, you know you are a part of it. When data are aggregated for the purposes of modeling, the raw records that go into the model are often de-identified.

So there is usually no sense on the part of the collector or user of the data that the individual's consent is required for this use of their data. You can usually "opt out" of being impacted by the results of the modeling, but you usually cannot opt out of having your de-identified experience contribute to the model.

The privacy policies that are maintained by data gathering institutions are clear that they apply to what they classify as personal information. Once it is de-identified, the information is not covered by those policies. For example, in the United States, health information is covered by the HIPAA Privacy Rule except when it is de-identified:

> "It is important to note that there are circumstances in which health information maintained by a covered entity is not protected by the Privacy Rule. PHI [Protected Health Information] excludes health information that is de-identified according to specific standards. Health information that is de-identified can be used and disclosed by a covered entity, including a researcher who is a covered entity, *without Authorization or any other permission* specified in the Privacy Rule." http://privacyruleandresearch.nih.gov/pr_08.asp (emphasis mine)

2. BEING AN OUTLIER (EVEN BY CHOICE) CAN BE THE EQUIVALENT OF FALLING IN THE STUDY'S MARGIN OF ERROR: Even if the individual were to choose to be an outlier as a form of protest or to somehow muddy the description of the cohort's results, their behavior

would be classified by the model as not significant.

The individual in Spain who won a case in the EU Court of Justice against Google for the right to be forgotten offers an example. He would find that the data point he wanted forgotten would appear in a dataset of "Spanish real estate transaction in the late 90's" if the dataset were aggregated with that description. The Court's ruling specifically allowed for the data point to be in datasets that were created without using his name in the search criteria. It is ironic that an analysis of that dataset would not be materially affected if his data point were, in fact, let out of it.

It's Big Data and it has to be modified in big ways to impact analysis that comes from it. If, to take an absurd example, 20-year-olds were to try to convince a supermarket chain that a vitamin intended for 70-year-olds was their favorite nutritional supplement, they would have to arrange for large numbers of them to buy that supplement regularly over a significant period of time.

Even if you did not have your data captured in a given context, it would still get "counted." Both traditional studies and advanced models today will use a technique called "inference" that can account for the behavior of those that belong to the cohort but whose data are not represented in the aggregated dataset.

3. REGRESSION TOWARDS THE MEAN: Traditional studies took a snapshot of data and produced analysis with it. Longitudinal studies will track a number of those snapshots over time. Studies are difficult to reproduce and refine because gathering new data to re-do the study is costly.

Increasingly, the aggregation of data occurs in a cycle wherein datasets can be almost continuously refreshed from the online systems that create them. Models are then easily refined. A cohort's behavior does not just regress towards the mean—the average for that cohort—as data are collected over time, it is actually assisted in doing so.

> The actions taken towards those in that cohort will reflect the average because that is the most efficient approach (i.e., has the highest probability of success). And so the individuals in that cohort will be presented with offers, treatments and opportunities that reflect the average for them and/or in ways that the majority of them will respond to.
>
> If the only offer an individual is ever presented with is the one that the average member of their cohort will choose, then this greatly increases the likelihood that the average response will be chosen and that in turn increases the power of the model to not just regress towards the mean but to help reinforce it[3].

The monopoly of knowledge created by aggregation is one that groups people together and predicts their behavior based on that grouping. But it is not as passive as it sounds. The predictions are often focused on responses to offers, treatments or other actions.

The power of Big Data, aggregated and analyzed, is to not only define how the average member of a group responds to the world, but to define what the world needs to do to solicit a response from the individual.

III

The greater the number of demographic identifiers in a de-personalized dataset, the more likely it is that someone could find an individual in it. Privacy advocates and security professionals rightly concern themselves with this aspect of protecting data.

There is no question that the risk of suffering financial and reputational harm when your individual data is inappropriately accessed is very real and can be very high.

The HIPAA Privacy Rule referenced above goes so far as to define exactly what 18 characteristics of a health record must be removed to de-identify it.

Even assuming the individual's privacy risk is all but eliminated in the use of large aggregated datasets, the individual is still impacted by them. That impact is created and measured at the group, or cohort, level.

We do not yet have a way of defining whether or not that creates risks and if so what kind. But aggregation does create narratives and those narratives do tend to "stand in" for certain details of an individual's life and future.

To put it another way: whether or not you feel like the needle, you cannot help but be part of the haystack.

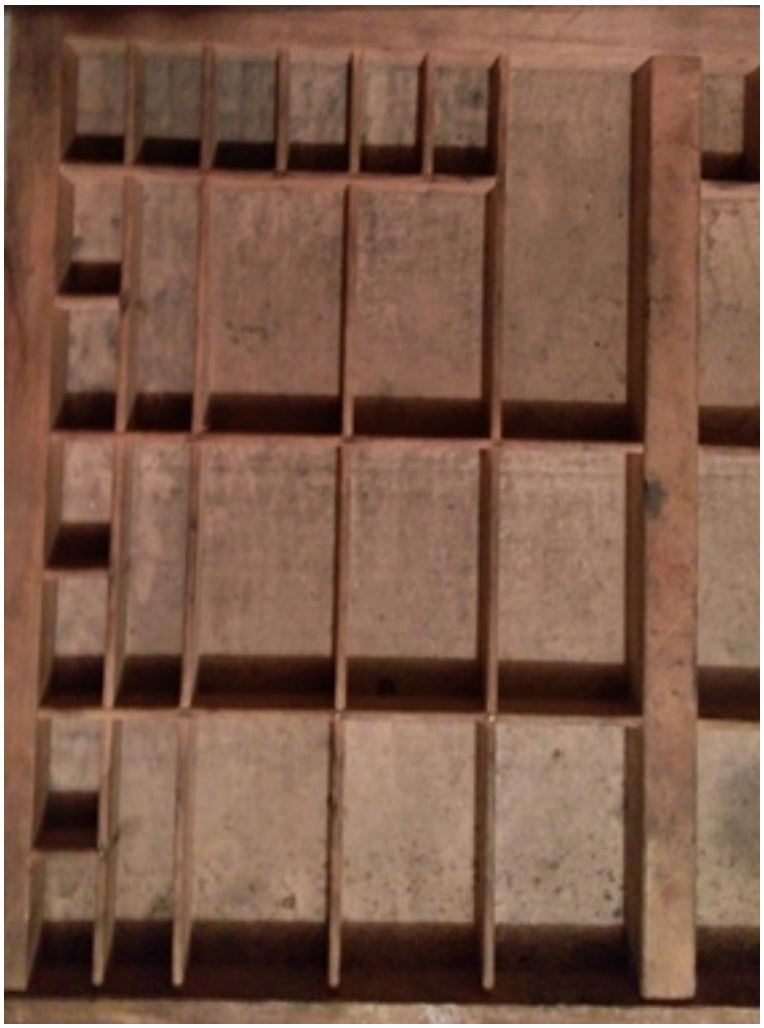
---

[1] This chapter focuses on the collection and aggregation of data and so the individuals discussed are those that are the subject of those data points.

[2] There are, of course, other distributions that might be used in a given analysis. The normal distribution is most fitting here because we are talking about common practices in probability applied to large populations.

[3] This "self-fulfilling prophecy" of statistical modeling refers to responses over which people have choices and not things where they do not such as medical treatments.

5



The Question of the Questions

Incessant questioning can reduce the best thinking to no more than a background chorus of "Are we there yet?" But there are still some things that have to be asked.

I have spent the past four chapters observing how aggregation is emerging as more than just an automated process. I've tried to show the following:

- Aggregation is a process that is independent of the data it aggregates
- The individual's relationship to data about them is changing
- Aggregation is the inverse of broadcasting
- The bias of aggregation is towards anonymity
- The actions of the aggregator can accelerate regression towards the mean
- An individual can be in a cohort without realizing it and certainly without choosing to be
- The world's increasing reliance on cohorts creates a different kind of identity for the individual

Because of all of the above, aggregation is creating a new monopoly of knowledge.

It seems important to list the questions that I can't find definitive answers to. I am posing the questions below. Those involved in maintaining and expanding the data-centric world mostly behave as if they're either certain of the answers, or at the very least, never seriously consider the questions.

Certainty can destroy useful discovery. And if I'm correct that we are undergoing a transition to a new data-centric, aggregation driven monopoly of knowledge, then it seems like a very good time to ask questions about that.

Does de-identifying result in dehumanizing?

Protecting individual privacy is increasingly important as more and more data are collected. This is so universally understood that even the United States Supreme Court, known for its bias towards allowing law enforcement wide latitude, surprised many that follow it when it ruled that the aggregated data on a cell phone required a different level of privacy protection from almost everything else found in a suspect's pocket. (Presumably this protection would extend to a flash drive, but as far as I know that has not been tested by the Supreme Court.

Experts and regulators recognize de-identifying data as a strong way of protecting an individual's privacy. When done systematically, stripping a record of the elements that tie it to the individual subject of the data (e.g., name, full address, full birthdate, social security number, etc.) can even be measured for its effectiveness in making the data anonymous.

But if de-identifying data results is removing the ability of those that use the data to identify the subject of the data, it also limits the subjects from having any participation in its use. Perhaps something is lost in this process that should not be. I've discussed elsewhere how problematic the idea of "consent" is when the subject knows that their information is being collected and has an idea how it will be used.

Given those difficulties, there are clearly obstacles that would need to be overcome for the subjects of data to meaningfully participate in the handling of their data once it is de-identified.

Subject participation is not the only area where de-identified data might be at risk to be dehumanized. The analytic techniques that make large de-identified datasets useful include methods for dealing with outliers, "margins of error" and even those missing from the dataset (in-

ference). Valid as the techniques themselves are, the question remains whether or not they create problems when an individual is treated merely as a member of a cohort.

It is not my intention here to evoke the image of some idealized past where individuals had some mythical power over their surroundings. Over time, the power of individuals has been as varied as the individuals themselves. Our relationship to the communities we are part of has always been one where we are members of them and, for better or worse, they are larger than us.

The question here is how does our identity in groups change when the groups are defined by distilling our identity down to a limited set of characteristics that are optimized for focused purposes? Especially when those purposes may or may not be related to the overall interests of the group itself.

Is surveillance always welcome?

It's a bit of a trick question. Most people I talk to will answer quickly "of course not; you can't speak in absolutes. The word 'always' is too definitive." And yet the apologists for surveillance do seem to take for granted that their mission implies a different answer.
Consider how James Clapper, U.S. Director of National Intelligence, described his mission. In September 2014, with tongue in cheek and a friendly audience (the AFCEA/INSA National Security and Intelligence Summit), Clapper put it this way:

> "We are expected to keep the nation safe and provide exquisite, high-fidelity, timely, accurate, anticipatory, and relevant intelligence; and do that in such a manner that there is no risk; and there is no embarrassment to anyone if what we're doing is publicly revealed; and there is no threat to anyone's revenue bottom line; and there isn't even a scintilla of jeopardy to anyone's civil liberties and privacy, whether U.S. persons or foreign persons. We call this new approach to intelligence: "immaculate collection."

Even the Director of National Intelligence should be free to be facetious when it's appropriate (and it was), but still the implied point of his remarks are clear: our mission is such that the characteristics of collecting data creates risks to things like privacy and reputation and they are risks we must take.

The term "surveillance" is often used to refer to real time monitoring and conjures up images of CCTV cameras everywhere. The word itself refers to words meaning, "to watch." Increasingly, as discussed above, it also refers to the wholesale collection of structured data, to aggregation, sometimes called "dataveillance.".

Some Public Health experts will claim that their brand of surveillance is always welcome. John Snow's analysis of the 1854 Broad Street cholera outbreak in London is a legend among data scientists and public health analysts. The data were clear, the cause of the outbreak accurately identified, and the lifesaving remediation simple and effective. The history of surveillance for public health purposes has nightmares in it as well. The Tuskegee syphilis experiment was an inexcusable case of the inhumane gathering of data.

The major differences between the two examples above are obvious. The Londoners in 1854 were contracting cholera where they lived and the goal was to prevent additional cases of a deadly disease. The unfortunate victims of the Tuskegee experiment were already ill, removed from their homes and had the cure withheld from them.

But the similarities are also clear: they both involved aggregating data for the purposes of public health and the Tuskegee victims and the London victims of cholera all represented data points to the analysts studying them.

Then there's public safety and security. Advocates of government surveillance in the name of public safety claim that the public is safer when there are guard[ians]s watching and so surveillance is always welcome. They sometimes claim that the risk of such dataveillance having adverse impacts on society can be mitigated by transparency.

For example, under the Data Mining Reporting Act of 2007, the

government is required to report on data mining programs and recognizes that data mining is an activity...

> "...involving pattern-based queries, searches, or other analyses of 1 or more electronic databases, where—
> (A) a department or agency of the Federal Government, or a non-Federal entity acting on behalf of the Federal Government, is conducting the queries, searches, or other analyses to discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity on the part of any individual or individuals" (Data Mining Reporting Act of 2007, Section 804(b)(1)(A))

Given the recent disclosures around the wholesale collection of phone records by the NSA, reports of instances of abuse of that data mining and the debate surrounding that, it would be difficult to say that even for public safety surveillance is always welcome.

In addition, there's surveillance as a means of control over the subject of the surveillance. Beginning with Jeremy Bentham's design for the Panopticon in the 1790's, omnipresent surveillance is often discussed as a means of social control. Critics of this acknowledge its welcome aspects, for example: enforcing no-smoking ordinances, ensuring appropriate behavior in public places, deterring theft and otherwise finding dangerous outliers, etc.

They also point out that the widespread use of this kind of surveillance can cause a "dangerisation" of one's view of the world around them: "The mere visibility of the [anti-theft surveillance] system on the one hand sustains in users the constant awareness of the probability of a threat and, on the other hand, automatically transforms the usual consumer into a 'non-thief'." (Michalis Lianos. Social Control after Foucault. Surveillance & Society 1(3): 412-430).

To put it another way: it is one thing to behave as a "law abiding citizen," it is another to be reminded by continuous surveillance that there are those who do not behave that way.

Just as with everything related to Big Data, the question of whole-

sale data-centric surveillance is also often discussed in terms of ensuring that an individual's right to privacy is respected. Indeed, protecting individual privacy is increasingly important as more and more data are collected.

Of the questions I am raising here, this one is the most widely discussed and debated. Are there times when the collection of data itself needs to be challenged? Surveillance may be a necessary evil, but is it *always* evil? Is it *always* necessary?

I think the definitive answers to these questions are to never stop asking them.

Is resistance futile?

We must be careful that the question of whether or not the data are secure does not drown out the question of whether or not to aggregate the data in the first place.

Kim Crawley's recent opinion piece in SC Magazine, *The problem with Big Data*, provides an example of this oversight. Crawley explains the security risks that we collectively take with Big Data. Crawley points out "Our Big Data technology wasn't initially developed with security in mind, but we must work hard to correct that. It requires fixing what we have now, and constantly monitoring and fixing systems as they grow."

While Crawley discusses the essential security basics of encryption, access control, hardening the environment, monitoring it, pen testing it and making sure qualified security professionals are maintaining it, she never considers the idea that some risks are undertaken by capturing data that should not be stored.

"Thieves can't steal what you don't have. Data minimization is a powerful element of preparedness. The rules are disarmingly simple: Don't collect information that you don't need," writes Brian Lapidus, COO of the Cyber Security & Information Assurance practice of the security firm Kroll, discussing how to prevent data breaches.

Even though eliminating the risk of a data breach by not storing

the data is as accepted a risk mitigation strategy as, say, log monitoring, Crawley does not mention it.

Still, Crawley is not wrong to take for granted that repositories of Big Data exist and will continue to grow and that it is the job of the security professional to protect them. Big Data is not a fad. Our understanding of it will grow, but likewise, its uses will as well. And aggregation and its consequent impacts on how we think, what we think about and what we expect the result of that thinking to be—what I've been referring to in this series as a "monopoly of knowledge"—is not slowing down.

Advances in technology and refinement of the tools for using Big Data add to that momentum. In addition, we have to acknowledge the eagerness of the growing analytic community to aggregate and use Big Data. How eagerly is information collected? Consider New York State's collection of Health Care utilization data:

> Health care facilities *must* submit on a monthly basis to the SPARCS program, or cause to have submitted on a monthly basis to the SPARCS program, data for all inpatient discharges and outpatient visits. Health care facilities must submit, or cause to have submitted, at least 95 percent of data for all inpatient discharges and outpatient visits within sixty (60) days from the end of the month of a patient's discharge or visit.
>
> Health care facilities must submit, or cause to have submitted, 100 percent of data for all inpatient discharges and outpatient visits within one hundred eighty (180) days from the end of the month of a patient's discharge or visit. (title 10, section 410.18.b.1. (iii) of the Official Compilation of Codes, Rules, and Regulations of the State of New York, emphasis added)

That data submission is mandatory makes it ironic that SPARCS stands for "Statewide Planning and Research Cooperative System". In the context of New York State health care providers, "cooperative" means that

they cooperate on the use of the data and cooperate via a governance committee to control disclosure of the datasets. Submission of the data, on the other hand, is not a matter of cooperation but of compliance with regulations.

Private sector aggregations of data rarely have regulations governing them as tightly as the New York State SPARCS data. Some large repositories, like credit bureaus in the United States, may be governed by regulations that limit their use. Others may be governed by laws that keep them from being created and/or transported across international borders.

In more and more countries, repositories that contain Personally Identifiable Information (PII) are governed by regulations aimed at protecting individual privacy and that is increasingly important as more and more data are collected.

Transactions that include creating an electronic record of each transaction are increasing. Many events that would not traditionally be recorded in an electronic record are being redesigned to do just that. Cash registers become Point of Service data collection terminals and medical devices generate a digital stream of data reporting on the patients they're hooked up to.

The security professional needs to recognize that analysts need data to do their jobs and deliver their value to the organization that both the analyst and the security professional work for. The ROI demonstration of some automation upgrades in an organization may even be stated in terms of the value of the analyst's output. (The Clinical Quality Measures defined for the U.S. Federal Government's Meaningful Use program, which provides monetary incentives for health care providers to implement electronic medical records, is a very clear example of this.)

The organization's need to generate value from the data it collects and the analyst's need for the raw material of their work are powerful motivators for creating large datasets.

Debates around how to use large aggregated datasets, who should access them and how to govern and protect them are essential. Those debates should begin with the question of whether or not the dataset

should be created in the first place.

Can means be defined so that a population regresses towards them?

Big Data is sometimes described in terms of "v-attributes:" volume, velocity, variety, value, and variability. Big Data is composed of a high volume of data. It grows at a high velocity because the number of records that make up a Big Data dataset increases as more and more electronic transactions are recorded and aggregated. Big Data also tends to include a variety of data points and even feeds from a variety of sources making its predictive power greater. All these attributes add to the value of Big Data.

Then there's variability. This describes the fact that Big Data tends to be comprehensive and therefore representative of the wide variation of characteristics of a population. When Big Data is used to describe the population whose data it contains, variability is just another attribute. Statistical models do not have difficulty accounting for variability. "Variance," in fact, is a technical term used in statistics to help describe the relationship of data to the mean, i.e., how spread out a population is relative to the average for that population.

In "When big data meets dataveillance," Sara Degli Espoli, refers to the v-attributes and then defines 4 steps in realizing the potential in Big Data analytics (The four definitions below are four direct quotations from her article; emphasis is hers):

> *Recorded observation* refers to the act of paying close attention—by watching, listening or sensing—to someone or something in order to gather and store this information in electronic format.
> *Identification* alludes to the recognition of an object, or a person's identity, through the analysis of an object, or a person's unique features.
> *Analytical intervention* refers to the application of analytics to the transformation of the collected information into knowledge, usually as a result of the first two types of actions mentioned

above.

Behavioral manipulation indicates the ability of influencing people's actions intentionally.

It is this last step, where high variability can be seen as a drawback.

While analytics has no problem at all with variability, operationalizing behavioral manipulation is more complicated the more variability comes into play. This is where the discipline of data science, the field of operations research and the newer applications of behavioral manipulation are biased against variability.

In a world of data analysis[1], this bias has its origins in another common use of data that originated in the world of manufacturing: Six Sigma. Six Sigma, in fact, describes a statistical result that stands for extremely little variation from the mean (resulting in 3.4 defects in a process per every million opportunities to have a defect). "One of the guiding principles behind Six Sigma is that variation in a process creates waste and errors. Eliminating variation, then, will make that process more efficient, cost-effective and error-free." (from the Villanova University definition of Six Sigma)

A normally distributed population regresses towards the mean over time. Behavioral manipulation that is driven by the data on a population might create a feedback loop between the data and the actions it takes to influence the population's behavior based on those data.

Does successful "behavioral manipulation" increasingly steer individuals towards a state where "one size fits all?" This question is not meant to imply a moral judgment on these activities or even imply that the practitioners of it are necessarily aware of this effect. But it is a question that should be asked if for no other reason than it might fall under the category of defining "unintended, perhaps unwelcome, consequences."

A conclusion against certainty

Certainty can destroy useful discovery. The current momentum of Big Data collection, analytics and use, our acknowledgment that our actions can have unintended consequences and the sheer volume of data being collected all suggest that this is no time to be certain of things. I'm sure of that.

---

[1] Pre-dating the world of Big Data, the bias against variability has been promoted by economists, who for centuries have demonstrated that efficiency comes with specialization.

6

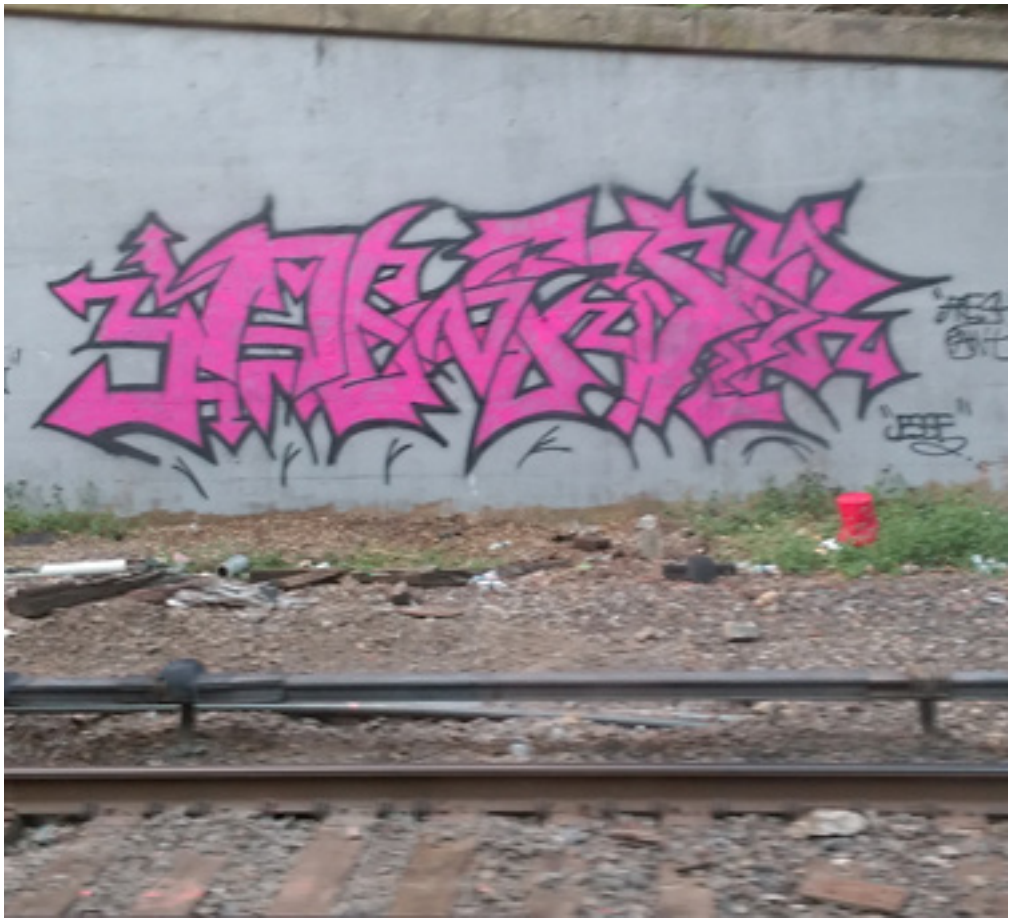


Photo by the author.

Afterthoughts

The first five chapters were written in a frenzy. Aggregating is the inverse of broadcasting. Aggregation is biased towards anonymity. Being the subject of a data point is a matter of immediate experience. Cohorts choose people more than people choose cohorts. And so on. Frenzies have a welcome feature: the simultaneity of both focus and chaos.

Which means that while I noted some thoughts along the way in writing the series, I had neither the time nor the appropriate peripheral attention to get to them.

Still, when the focus is broken and the chaos finally gives way to some type of order, what was left on the periphery can get mentioned. As follows.

I

Artifacts are either intentional (the Pyramids at Teotihuacán for example), accidental (millions of unearthed pot shards) or inferred (the details of the social organization that must have existed to build monuments like the Pyramids). Aggregation is only possible in the presence of large datasets. The majority of technology awe expressed by writers today focuses on the advances in technology that either expand the manageable size of these datasets (big data), the emerging sources of them (the Internet of Things), or the more technical side of manipulating the data (beyond SQL).

The precursor to all these advances is the actual cultural shift towards the acceptance of transaction level logging for software applications. The inferred artifact is the transition to this acceptance and the

way it has become close to an expectation. Why do we think it is ok to collect all that data?

In fact, some privacy debates still focus on preventing the actual recording of events as a way of protecting the privacy of an individual. But that debate has increasingly shifted to protecting the records, which, it is taken as a given, are ubiquitous. When the US Supreme Court ruled that the contents of a cell phone could not be searched routinely without probable cause just as when the EU Court of Justice ruled on the right to be forgotten, the existence of the records was never questioned. Nor was a future in which the volume of such records would continue to increase or the "argument" for aggregating those data.

II

Cases that have been brought to prevent the NSA from collecting metadata on telephone calls usually do not call for the metadata to not exist. Companies will insist that such transaction level logging is required for accurate accounting and management of their business, to assist them in resolving customer disputes and preventing fraud.

However, the assumption that transactions must be recorded and persisted as data is not accurate. Companies can and do store images of bills to be referred to by their customers. This is usually an option on on-line self-service portals. The implication is that such images are a transition from the time of paper bills to the truly on-line "data dump" of a bill, i.e. a table of records displayed on a screen.

Those images, however, could contain every bit of detail needed for customer service as the large datasets currently stored. Because they are still being stored digitally, the company could always find an individual's records in the same way the consumer using the portal does.

A customer service representative would not have to read the image to resolve the dispute. If the images were stored in such a way that they could be converted (the technical term is "parsed") into data when a service rep needed to address a customer concern, then all the same

computer screens that the rep now users could be populated. Once the customer's compliant was addressed, the transactions that had been converted into searchable data could be wiped from the application's working storage. At that point, the customer is left with the company once again just storing images of their bills.

And when subpoenaed, the company could produce the images which law enforcement would then have to parse. We are used to thinking of digital images as types of documents and so the idea of a subpoena would be easier for law enforcement to adapt then the current mess of trying to determine what constitutes reasonable search and seizure of millions of records at once.

Advances in the efficiency of processors and storage capacity make this less farfetched a technical feat than ever before. The only thing that would be lost would be ease of creating large aggregated datasets for analysis.

III

When Andy Warhol predicted that each person would get 15 minutes of fame, he never said they would not have to share those 15 minutes. There are trains of thought in philosophy, economics and political science that discuss concepts like the alienation of workers, individualization, and consumerism. Some of these concepts go back over a 100 years. We do not need to pass moral judgments on these concepts to recognize that the technologies of a networked world, a data-centric environment, relies on identifying individuals.

We may have reached the point where at least local celebrity and alienation have converged. I would argue that this is the result of identity itself becoming a data point.

Sometimes those identities are at the level of impersonal data points like MAC and IP addresses. Sometimes they are specific to the individual but still somehow removed as in the case of email addresses and mobile phone numbers.

Authentication is what makes identity personal in a data-centric world. So long as identity theft, impersonation, creates nothing more than acceptable noise in data models, then the individual's interest in protecting their own identity will be greater than any data collection enterprise's interest in it.

I am distinguishing authentication here as a way of identifying the subject of a data point from the more common use of it: as a way of controlling access and being tied to authorization. As a security mechanism, coupled with authorization, authentication is invaluable and hence we see efforts to increase its strength.

But as a way of identifying events and data points, authentication takes on a different dimension. In everything from Social Media to financial transactions, from cloud based storage to instant messaging, people are being increasingly confronted with the idea that they must identify themselves as unique individuals and they must protect that identity from all other individuals.

And it is a confrontation. Consider the jargon: the security questions that some systems employ to validate your identity in password reset scenarios are referred to as "challenges". They call to mind the fairy tale scene where the stranger approaches the heavily guarded gate and answers a series of questions to be permitted to pass. Hence the most common authentication element is still called a "password".

So the individual devotes increasing amounts of attention to ensuring they can reliably assert their "individualness." They do this because the data point, the transaction as well as their defined privacy rights all insist on the individual as the subject.

Again, I am not trying to place a moral judgment on these phenomena, merely to observe that this very emphasis on the individual as a person who is regularly asserting that individualness is reinforced by the way the technologies have evolved.

This causes us to need to distinguish between "individualness" and "individuality". The former can be described as the establishment of yourself as an individual. The latter retains its common meaning of describing the fact that there are unique things about you. To be blunt:

the difference between having a unique identity and being a unique person.

V

The emphasis on establishing and protecting on-line identity is urgent. There is a good deal at stake for the individual. Victims of identity theft suffer harm that ranges from the inconvenience of changing account numbers and passwords to the well documented cases of finding themselves owing money for goods and services they had nothing to do with acquiring.

This emphasis is, therefore, necessary and appropriate. As an Information Security professional, I would never argue against it. I actively advocate for it, in fact.

I believe that scholars, like Ulrich Beck, would recognize this emphasis as falling into what they describe as "institutionalized individualization" which they see as breaking down the potential for collective action if not collective consciousness itself among people .

However, just because people may not be recognizing themselves as members of a community, does not mean that views of individuals aggregated into groups do not exist. Just as the argument can be made that individuals can become so pre-occupied with the role of being an individual that they cannot associate themselves with collective action, so I would argue that this pre-occupation, when carried into the data-centric world, contributes to the fact that people are divorced from the cohorts they are put into by collectors of the data about them.

VI

I have argued elsewhere that there is at least one new role in data-centric world: the searcher. Both the EU Court of Justice decision on the right to be forgotten and the Federal Data Mining Reporting Act of 2007 carve out a privileged position for searchers that give them rights sepa-

rate from the privacy rights of the subjects of the data (even though each data point undoubtedly has a subject).

The U.S. law defines "data mining" as an activity...

> "...involving pattern-based queries, searches, or other analyses of 1 or more electronic databases, where—
> (A) a department or agency of the Federal Government, or a non-Federal entity acting on behalf of the Federal Government, is conducting the queries, searches, or other analyses to discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity on the part of any individual or individuals;
> (B) the queries, searches, or other analyses are not subject-based and do not use personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals, to retrieve information from the database or databases;"
> (Data Mining Reporting Act of 2007, Section 804(b)(1))

In addition, it is well documented that the buying and selling for aggregated datasets constitutes a market and that there is a specific role in this market for data brokers.

Describing the characteristics of the Data Broker Industry in the United States, the Federal Trade Commission noted:

> • The Data Broker Industry is Complex, with Multiple Layers of Data Brokers Providing Data to Each Other: Data brokers provide data not only to end-users, but also to other data brokers. The nine data brokers studied obtain most of their data from other data brokers rather than directly from an original source. Some of those data brokers may in turn have obtained the information from other data brokers. Seven of the nine data brokers in the Commission's study provide data to each other. Accordingly, it would be virtually impossible for a consumer to

> determine how a data broker obtained his or her data; the consumer would have to retrace the path of data through a series of data brokers.
>
> • Data Brokers Collect and Store Billions of Data Elements Covering Nearly Every U.S. Consumer: Data brokers collect and store a vast amount of data on almost every U.S. household and commercial transaction. Of the nine data brokers, one data broker's database has information on 1.4 billion consumer transactions and over 700 billion aggregated data elements; another data broker's database covers one trillion dollars in consumer transactions; and yet another data broker adds three billion new records each month to its databases. Most importantly, data brokers hold a vast array of information on individual consumers. For example, one of the nine data brokers has 3000 data segments for nearly every U.S. consumer.
>
> Data Brokers: A Call for Transparency and Accountability. Federal Trade Commission. May 2014.

Similarly, in testifying before Congress regarding this industry, Paul Kurtz, at that time the Executive Director of the Cyber Security Industry Alliance (CSIA), described the importance of data brokers as follows:

> "We need, simply, to come to terms with our reliance on information systems and the vast amount of personal information in storage and in transit in such systems...the representatives of the information-broker industry will testify this morning that the American economy and even our national security are becoming increasingly dependent on this industry."

Mr. Kurtz's testimony was delivered on May 10th *2005* and the FTC report quoted above it was, of course, published nine years later. It is perhaps fair to say that in those nine years, we did not come to terms with our reliance on information systems in the way that Kurtz meant. I believe that is because we have not come to terms with the unique roles

that go beyond individual data record transactions.

While the traditional roles in data privacy and security are both functional and transactional—subject, user, collector, discloser—it seems important to accept that there are two equally important roles for us to consider, roles that are functional only when focusing on the aggregation of data. The roles are searcher and broker.

VII

We tend to look at technologies by what they do, how they were developed and what they replaced. While there are tremendous differences between primitive Paleolithic technologies, a stone ax for example, and our current ones, we gravitate towards thinking of them both as enabling. The rhetorical "what did we do before we had x" applies equally to the water wheel as the jet plane. We call them "advances" and discuss what we can now do that we could not do before. Or what we can now do more of or do more efficiently.

Those discussions are necessary, often accurate, but not sufficient. We need to look at technological advancement in the same way we look at the development of pharmaceuticals. The discussions of what a technology does and how it works is what, in the world of products subject to Food and Drug Administration (FDA) approval, would be called intended or approved use. But we should, taking their lead, also discuss side effects and even what is referred to as off-label use.

This has happened with some technologies. The impact of technology on the environment is perhaps the best example of discussing side effects. Whether or not blue tooth headsets are safe for one's health is another. The early studies of the effect of watching television on children are also examples of looking at safety and side effects with respect to a technology.

Anyone who has ever driven an infant around the block over and over again just to get them to go to sleep is familiar with one off label use of the automobile. Data collected to record purchases at a super-

market being then aggregated and used to model buying behavior is another classic example of an off label use.

In discussing aggregation, I've attempted to look at the broader characteristics of the technology and how it is changing us as well as to suggest where to look to explore those two apt criteria used by the FDA when evaluating new treatments for approval: safety and efficacy.

By way of thanks

The majority of this work appeared as a series of articles in Security Current and so my thanks, as always, goes to its publisher, Aimee Rhodes, whose encouragement and editorial assistance was invaluable.

Additional thanks goes to Michelle Wolfe, professor of Broadcast Communication at San Francisco State University.

Thanks also to Kathrin Böhm for allowing me to use the haystack photo (and for taking it in the first place).

This work is dedicated to the memory of two great writers and thinkers who we have recently lost:
Kenneth Irby and Stephen Rodefer.

A virtually unlimited number of copies are available.
The text is in Baskerville Old Face. Display types are Imprint MT Shadow,
Minon Pro and Jenson Classico.
The author was fortunate to find a digital copy of Zapf Dingbats and has
used them where appropriate (the symbol below is not one of them).
Fall 2015

About the author

David I Sheidlower is an Information Security Professional who explores the intersection of humanism and technology. He works as the Chief Information Security Officer for a global media and advertising company. He has been the CISO for Health Quest, a health care provider in the Hudson Valley of New York, and has worked managing security, privacy and data for Wells Fargo Bank, Fair Isaac, and Kaiser Permanente.

He publishes regularly in SecurityCurrent and maintains a blog, www.cybersecrighthere.com.

His previous ebook, Principles of Data Security & Privacy, can be found at the Security Current website.

He is a graduate of the University of California, Berkeley and St. Mary's College of California.