Chapter 2: All the Data on All the People

Surveillance and privacy too

About two years before 9/11, the military anticipated a terrorist attack on U.S. soil and asked a group of academics to build a surveillance system that harmonized surveillance with privacy protections. We did. This is the story of how it came to be and what happened to it.

Most Americans remember where they were on September 11, 2001, the day terrorists hijacked commercial planes and crashed them into buildings and a countryside. I was working at home, in Pittsburgh, not far from the farmer's field in which one plane exploded. My email inbox flooded with messages about our bio-surveillance system. You see, on 9/11 we had bio-surveillance systems operating in New York City and Washington, DC, the sites where the other planes crashed [1].

Recruitment

It all started about 18 months earlier. High-ranking military officials from the U.S. Department of Defense (DOD) asked to meet with a select group of professors at Carnegie Mellon and the University of Pittsburgh. I was among the Carnegie Mellon professors. While the officers had different body shapes, faces and hair color, they were more similar than not. Each displayed a breast full of shiny, colorful ornaments, wore high gloss shoes, orated with commanding self-confidence, and had clothes that stayed at attention even when they sat. These no-nonsense officials advanced to the point quickly. They were concerned about a possible terrorist attack on U.S. soil involving “planes” and “anthrax”. In their hypothetical scenario, an airplane, flying above a public outdoor sporting event, releases anthrax on the players and crowd. Affected people go home talking about the game, not realizing their exposure. Days later people start acting as if they have the flu: they take time away from work and school, search online information, and purchase over-the-counter medications. Weeks later people begin to die, seemingly from respiratory complications, with even more time and deaths likely occurring before authorities identify the common cause and triangulate to the originating site. Meanwhile, the site remains contaminated, infecting more people.

Silence was our response. The seemingly youngest of the military officers jumped up, slammed a clunky laptop on the table and began fuddling with cables to connect it to the projector. The raised lid showed a “Property of the United States Government” sticker. Others located power outlets and pushed buttons on the projector. Soon the young officer started his lecture on biological threat agents and aerosol delivery of anthrax.

The scenario appears in public writings, he acclaimed. A 1970 World Health Organization publication estimated that 110 pounds of anthrax spores, dispensed by plane one mile upwind of a population center of a half million people (500,000) in ideal meteorological conditions, would travel more than 12 miles and kill or incapacitate about half of those in the path of the biological cloud (220,000) [2]. A 1980 book proposed that of all the biological warfare possibilities, the most likely attack scenario would be aerosol delivery of anthrax spores. The resulting illness, inhalational anthrax, is the same as woolsorter's disease, which already occurs in the textile and tanning industries among workers handling contaminated wool, hair, and hides [3].

The young officer then said, “we do not have any actual experience with military grade anthrax; however, a prior situation gives us an idea of what to expect.”

On April 2, 1979, there was an accidental release of anthrax spores from a suspected Soviet biological weapons facility. The officer suggested that a worker at the end of his shift removed an air filter and the worker on the next shift did not replace it. Consequently, a plume of anthrax escaped and was carried by the wind. The plant, roughly 850 miles east of Moscow, was in the city of Sverdlovsk (now called Ekaterinburg). The anthrax outbreak killed at least 64 people with possibly 94 people being affected in total. The earliest reported symptoms were fever, malaise, and fatigue, with a nonproductive cough and vague chest discomfort sometimes present. These symptoms are reminiscent of the flu, and in fact, many of the 64 decedents thought they had the flu. Improvement occurred for 2 to 3 days for some of the people affected, but for others, symptoms progressed directly to the abrupt onset of severe respiratory distress causing them to be hospitalized with dyspnea, stridor, diaphoresis, and cyanosis. Bacteremia, septic shock, metastatic infection (meningitis in approximately half of cases), and death usually followed within 24 to 36 hours.

His slideshow included a chart detailing time lapses between the onset of symptoms and death for the 64 affected decedents [7,8]. Figure 1 shows a graphical summary. The black curve confirms the mounting deaths of the 64 decedents over time. The first death occurred about a week after exposure. The pink curve displays the cumulative number of hospitalizations. The first hospitalization appeared just before the first death. Less than 50 of the 64 decedents visited a hospital. The green curve, which precedes the other curves by days, reveals when decedents acknowledged symptoms by making purchases, changing behaviors, or talking about the illness with friends or relatives. The earliest authorities could have known something had happened would have required them to follow the green curve because the onset of symptoms occurs days before hospitalizations and deaths. But tracking symptom-inspired events requires continuously observing the public for signs of flu-like behavior.

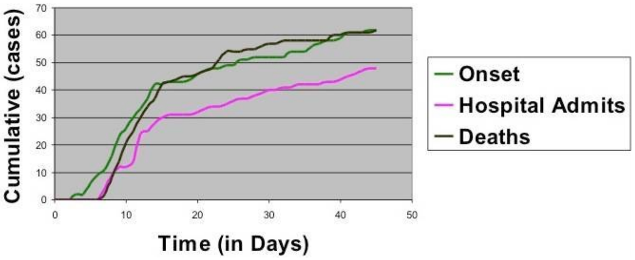

Figure 1. Time lapses from exposure to aerosol anthrax (day 0) to the onset of symptoms (green curve), hospitalizations (pink curve), and deaths (black curve) as they accumulate for the 64 decedents in the 1979 anthrax outbreak in Sverdlovsk [7,8].

Beyond anthrax, the young officer surveyed a laundry list of biological threat agents: brucellosis, plague, Q Fever, tularemia, smallpox, viral encephalitides, viral hemorrhagic fevers, botulinum toxinsa and staphylococcal enterotoxin B [4]. A medical facility would not be expecting a biological agent, so he described ways patients would likely present to unsuspecting doctors: botulism-like, hemorrhagic-like, lymphadenitis, localized cutaneous lesion, gastrointestinal, respiratory, neurological, rash, specific infection, fever, and sudden severe illness or death [5,6]. He asserted that looking for respiratory problems related to anthrax was a good exemplar.

"The sooner officials can determine whether a biological agent, such as anthrax, has been released into the environment, the more lives we can save," he concluded.

Why were these military officials telling us? Apparently, military authorities had difficulty getting national law-enforcement agencies to take preventive action. They wanted to know if we could design algorithms and construct systems to continuously survey the American public to determine whether something suspicious was underway. I understood why the algorithms, biology and medical professors were there, but why, I asked, was I there. The seemingly eldest of the officers looked at me with disdain as if I had asked the most stupid of questions. He slowed his speech and raised his voice to explain that they are talking about surveillance of Americans on the homeland during peacetime. According to him, if nothing happened, it could all backfire and they could possibly face court martial.

So he barked, "your job is to make sure this surveillance is done using anonymous data and privacy protections."

Armed with seed funding and collaboration with the Centers for Disease Control and Prevention (CDC), we set out for combat [9]. The original list of data sources to consider included traditional sources for bio-terrorism surveillance –namely, hospitalizations, emergency room visits, physician office encounters, medical lab reports, and prescriptions; an extended list of sources –namely, assisted living reports, school and employment absenteeism, and wildlife disease reports; and, very non-traditional sources –namely, grocery purchases, web click-stream data, videos of public spaces, and air sensors. A bio-surveillance system should detect both naturally occurring outbreaks as well as biological threat agents, so we could test our efforts by predicting outbreaks as they occur in the real world. My colleagues got busy working on air sensors [10], studying the pathology of biological agents, constructing algorithms that could detect anomalies in various kinds of data [11], and making data sharing arrangements.

My job was to architect a surveillance system capable of determining as early as possible whether a significant number of people are acting ill, and to do so, with privacy guarantees. The term *privacy-preserving surveillance* danced in my head as a name for the goal. For me, the most enjoyable activity at the genesis of a project is the naming part. Privacy-preserving bio-surveillance is a bio-surveillance system having privacy guarantees. If you have two privacy-preserving surveillance systems, they must both deliver the same utility as an ideal surveillance system without the privacy guarantees, but one is better than the other if it makes stronger privacy guarantees. There is no one model for privacy-preserving surveillance. It is not limited to bio-surveillance either. There are many possible privacy-preserving solutions, some better than others and some aimed at different kinds of surveillance.

For bio-surveillance, the earliest possible signals would come from daily life, such as online searches, absenteeism and grocery and pharmacy purchases. Epidemiologists call this syndromic surveillance. Some affected people will visit physicians and show up in hospitals, so tracking medical events was a common way public health was already doing surveillance. However, reporting was sporadic and limited to a physician actually diagnosing a biological agent. Epidemiologists termed this bio-terrorism surveillance. Limited to medical reporting, public health did not include school absenteeism, over the counter purchases, web searches, or any of the other kinds of non-traditional data now being considered. In this writing, I use the term “bio-surveillance” to fuse surveillance on all kinds of data, medical and non-medical, traditional and non-traditional, and to use approaches more robust and timely than doctor or lab reporting based on diagnosis alone.

Eventually I collaborated with a team from Johns Hopkins –the folks that operated the bio-surveillance system near the location in DC where one of the planes hit [12]. But that's 9/11, so I am getting ahead of myself. Let me go back and step through the struggles that shaped "privacy-preserving surveillance".

Whose authority?

My meeting with the military stiffs was followed by a litany of meetings before 9/11 with the CDC that left me with an appreciation for Atlanta as a city and a false impression that public health had the authority to conduct bio-surveillance. I had a litany of meetings with law-enforcement officials that left me with the false impression that the Federal Bureau of Investigation (FBI) had the authority to conduct bio-surveillance. I even had a litany of meetings with a separate entity that Carnegie Mellon and the University of Pittsburgh formed to possibly act as an independent third-party to conduct bio-surveillance [9]. They all wanted the authority. I wanted one of them to have it too because then, I could just architect the system to satisfy privacy requirements between the authorized entity and the various stores, schools, medical facilities and other organizations holding desired data. Soon, though, I realized the truth. Before 9/11, no entity has the legal authority to simply demand access to the breadth of personal information bio-surveillance systems seek on a daily basis in the absence of a biological threat agent. So, which entity has authority for what? How do authorities get personal information from those organizations that hold it?

One possibility is to wait until deaths mount or a physician diagnoses a case involving a biological agent. Then, public safety would be in imminent danger, thereby triggering public health laws. During the presence of a serious health threat, public health laws tend to grant public health departments sweeping abilities to seize data, detain people, and establish and engage in active surveillance, until resolution of the immediate threat. Government departments already monitor the number of deaths occurring in a community. If the mortality rate spikes due to a possible health reason, even from an undetected biological agent, the public health department has the legal authority to investigate and thereby demand identified, detailed information about involved persons. Of course, waiting for an escalation in the mortality rate costs the very lives bio-surveillance seeks to save.

A second possibility is to change the law. Public health laws tend to be reporting laws. A public health department identifies which diagnoses to report, information to include, and time covered. When a physician, lab, or hospital assigns a reportable diagnosis to a patient, for example, the law may require the medical facility to forward the name, address and other details about the affected person to the public health department. In some locations (e.g. [13]), the flu is a reportable disease, but it is not reported until a week later. Biological agents are immediately reportable, but early cases are unlikely to bear that diagnosis. What if we change public health law to make the flu and similar respiratory illnesses reportable daily? Then, details about everyone diagnosed with those ailments would forward to a public health department, even though there is no immediate public health concern. For early detection of a biological threat agent, the law would also have to require reporting of purchases of over-the-counter medications, school and work absenteeism, and other kinds of information from sources that are not accustomed to reporting information to public health authorities. Over time, public health departments would hold information about most of the population, even in the absence of threat to public safety. Of course, this fails to preserve historical privacy protections afforded individuals.

Another possibility is to work through a business. Is there a value proposition that might entice organizations to share data with a non-government third party? Imagine a for-profit or a non-profit corporation that relies on contracts and business associate agreements with organizations to acquire personal data. The business alleviates privacy and confidentiality concerns by assuming responsibility for data breaches. The information would likely be identifiable but would not necessarily bear people's names, Social Security numbers or explicit identity. The omission of names and addresses reduces unnecessary risk, and in the case of schools [14] and medical facilities [15], federal regulations may require specific redaction. Bio-surveillance service would be one offering. If suspicious activity occurred, public health could intervene to get the names as needed from the originating organizations. There is not much money in public health, so the business might also provide industry indicators, marketing statistics and marketing feedback to participating organizations. Similar businesses, like hospital associations, already exist. They don't usually receive data daily, nor work across multiple industries, like medicine and groceries though. Of course, this approach also raises privacy concerns. Identifiable personal information would move in bulk to a business that would be unknown to the people who are the subjects of the data. And as a private entity, its operation would be beyond public inspection and accountability.

Of course, a law-enforcement investigation launches as soon as a public health department determines that a biological threat agent is affecting people, air sensors detect the presence of a biological threat agent even without knowledge of any people being affected, a 911 phone call from a tipster alerts to the possibility of a biological threat agent, or a letter appears from a terrorist or criminal taking credit for the release of a biological threat agent. The FBI would seek to identify, capture and prosecute those responsible. Their investigation begins with known evidence –the site, affected people, the letter, or the kind of biological threat agent involved– and proceeds outwards gathering relevant, related information as the investigation warrants. This targeted focus on data gathering assures you that the FBI will never need to know about orange juice consumption or any of the other early warning data about people unrelated to the investigation. From a bio-surveillance perspective, timeliness depends on how the investigation begins. If initiated by letter, phone call or sensor, law-enforcement can provide the earliest possible signal; otherwise, public health would have already rung the alarm. So, sensors, 911 phone calls, and threats are additional inputs to a bio-surveillance system. Eventually though, if a biological threat agent is found, regardless of how we learned about it, the FBI and public health departments will be in charge.

These possibilities reveal a natural ordering in which personal information flows from a hospital, store, or school, where it is first captured, possibly to a third party organization, but eventually to a public health department, and if necessary, the FBI. A bio-surveillance system will have to address privacy requirements across all these parties.

Clearly, public health is central. Public health departments have surveyed populations since the end of the nineteenth century [16]. The original purpose was to identify people with contagious diseases so that public officials could take prompt action to prevent an epidemic. Public health surveillance expanded to include non-contagious diseases like cancer and birth defects in order to track incidence and mortality. Bio-surveillance creeps the mission even further by dramatically increasing the number, nature and frequency of personal information collected, and for a purpose that may never occur though it promises to help identify naturally occurring outbreaks too.

As conversations about bio-surveillance progressed, enthusiasm climbed, especially within public health circles. Bio-surveillance promised to revolutionize public health by infusing it with the latest technology and by modernizing surveillance with custom-made programs, and public health needed a technological makeover. I was always amazed to see how poorly funded most local public health departments were at the time. The latest technology I would usually find on the premises was an outdated desktop computer running a statistics or spreadsheet program. I usually had more computing power and storage space on the laptop I carried in my shoulder bag than could be found on any computer anywhere else in the building.

For the remainder of this writing, I will refer to public health departments as the central authorities for bio-surveillance without implying a public health department is the only possible authority. Discussion that uses a public health department as the central authority generally applies to other possible authorities, such as businesses and the FBI, even though specific legal instruments may differ.

Whose privacy standard?

With growing zeal within public health for bio-surveillance came eagerness to have organizations simply forward personal information to public health directly, or indirectly through a third-party, without privacy concerns. Privacy seemed the only obstacle to their achieving technological bliss. Eventually, they began asking, is privacy necessary? What can go wrong?

Public health surveillance starts with a registry, which is a collection or database of information about individuals having a specific diagnosis or condition (see [17]). For example, all state health departments in the United States maintain cancer registries that identify who had what kind of cancer, when the diagnosis was made, and the medical history of those affected. The first cancer registry to find a link between lung cancer and tobacco was in Nazi Germany, according to Robert Proctor, a professor of the history of science at Stanford University [18]. The Third Reich reportedly used public health measures to strengthen occupational safeguards, improve food and drugs, issue bans on smoking, and reduce the use of cancer-causing cosmetics. However, the Third Reich later used public health surveillance and registries to track people and implement state sanctioned sterilization and euthanasia.

Of course, public health registries don't lead to genocide. After all, public health registries did not seem to play a role in the genocides that occurred in Cambodia during the 1970s, Rwanda during the 1990s or Sudan during the 2000s.

While the use of public health registries in Nazi Germany reminds us of the good that registries and public health organizations can do, they also warn us that data, once collected for one purpose, no matter how worthy or seemingly benign, can be used for unforeseen purposes later. In fact, bio-surveillance itself re-uses personal data initially collected for other reasons. Where else might the information go and for what other purposes might it be put?

Before the military recruited me to work on bio-surveillance, I had been busting myths by putting names to medical data wrongfully believed to be anonymous (see Chapter 1). Many of my examples of data that were not sufficiently anonymous came from public health related sources. My work had been a timely contribution to the discussions that shaped the federal health privacy regulation in the United States known as the Privacy Rule of the Health Information Portability and Accountability Act (HIPAA) [15] and earned me a front row seat to the debates.

During the HIPAA debates, medical ethicists, doctors and patient groups proclaimed loudly that privacy was essential for patients to trust doctors, and without trust, patients would not share important information with doctors and risk poor outcomes. It was not surprising that discussion among public health officials about personal privacy and bio-surveillance was the opposite. The job of public health, after all, has a population perspective, a belief in the "greater societal good" of health protection and promotion. Privacy, to many in public health, is a barrier to improving public health and safety, and because of it, public health officials worry that they may not receive timely information to save lives. The dichotomy of thought between medical ethicists during the HIPAA debates and epidemiologists during bio-surveillance discussions made my head swim. How do we balance social good and individual autonomy in bio-surveillance?

Spirited debates among experts may have been insightful, but there was opposition to bringing ethicists and advocates into the bio-surveillance conversation. Wounds from the HIPAA debates had not yet healed and as one epidemiologist told me, "you are about as much privacy as we can take". How much privacy was that? Little, I am afraid. I was neither an ethicist nor an advocate, just a computer and social scientist trying to architect a data sharing system with privacy guarantees to facilitate bio-surveillance. I was not the enemy or the naysayer or the thought-provoker but the enabler. I was intent on engineering a scientific right answer that harmonized bio-surveillance with societal norms. I was not going to disappoint the military stiffs. My charge was to find and operationalize privacy standards, not to question whether they were sufficient.

Never shall the ethicist and the epidemiologist meet in the real world to discuss bio-surveillance. But in my mind, that's a different story. I would often juxtapose the two groups in mental debate.

“What is the harm in the government knowing about your orange juice purchases?” an epidemiologist asks.

“Sharing my personal information without my consent is as much a violation of my dignity as a physical invasion of my body. It is irrelevant whether the physical invasion caused me physical pain,” a medical ethicist responds.

“You would gladly trade that dignity for safety when faced with a biological attack”, proclaims the epidemiologist.

“Yes, so grab my data then,” responds the ethicist.

Here is another debate on my mental stage.

A medical ethicist asserts, “the government wants to know about my purchases, my web visits, my physician visits, and my days off from work for a hypothetical possibility that is unlikely to ever occur. It is more likely that the public health department will use the information to make me exercise more, eat less sugar, and stop smoking.”

The epidemiologist responds, “we are only interested in protecting you from contagious, potentially lethal diseases.”

The ethicist states, “you are justified in protecting me and the public from imminent danger, but not to protect me from my own bad choices.”

The first of these debates raises concern over the conditions under which personal data is seized and the second warns against mission creep. I gleaned some additional insights from similar mental debates and discussions with others. Sharing all your personal data with a business can cost you economic harm, but giving all your data to a government can additionally take away your liberty. So, there is more concern over government activity than business. In surveillance, government forces its way into one’s personal space to extract information without a person's say or knowledge, and if asked, a person may want to keep his information secret or later wish he had. So, a role of privacy safeguards is to protect the person when he cannot protect himself.

I also spent a little time trying to economically balance the value of a particular piece of information to achieve bio-surveillance against the dignitary cost of invading privacy and the potential for economic harm… aargh! My head hurts just writing about it.

Okay, this kind of discussion is fine for philosophers, but I am trying to build a system for real-world use. Public health surveillance has been operational for a long time, surely operating under some kind of privacy standards. What? I turned to the wisdom of governing law.

The legal authority in public health resides at the local and state level prescribed by state law. At first, I thought the CDC had over-arching power and acted as the public health version of the FBI. I know some criminal activities are beyond local and state law-enforcement and handled by the FBI directly. However, that is not how public health works. Instead, the CDC and public health departments have a data-money ecosystem. Money flows from the CDC to local and state health departments, and personal information flows from those local and state health departments authorized to acquire it to the CDC. When incidents happen, the local or state health department may invite the CDC to participate to take advantage of the CDC's resources but without the invitation, the CDC cannot participate. This means that bio-surveillance is likely to be done at local and state levels, in geographical partitions where the public health departments have authority and have or can build relationships with local participating organizations.

I had many questions about public health law and Carnegie Mellon had no law school, but the University of Pittsburgh did. The editorial team of the Pittsburgh Journal of Technology Law and Policy agreed to survey public health laws and regulations in each state, as they were before 9/11, and quantitatively answer my bio-surveillance questions [19]. What did I learn? If a public health department had all the data on all the people, they could possibly catch potential epidemics early, and the earlier the better to save lives and give the public better health outcomes. That's a fact, not law. But giving public health all the data on all the people is more data than the law allowed at the time. In general, before 9/11, public health laws and regulations required public health departments to be specific about the diseases they continuously monitor and limit data collection to medical and specific reporting requirements. They also allow for bursts of broader data acquisition when public health situations merit.

So, there is no established privacy standard for bio-surveillance, but public health laws and regulations encourage bio-surveillance to preserve the historical norms that balance public health safety and privacy by: (1) providing the minimum information needed; (2) establishing transparent rules for data sharing; and (3) allowing bursts of increased data access as public health situations justify. Can technology make these assurances?

Privacy-preserving surveillance

By November 2000, months before 9/11, I had a technological architecture. My big idea had 2 parts: (1) distribute the computation and not the data; and, (2) get increasingly more specific information as prior results warrant. Originally called the "reasonable cause predicate" by me, the approach later became known as selective revelation [20].

The big question in bio-surveillance is "how many?". How many people reported respiratory distress today? How many people bought orange juice and tissues today? How many children were absent from school today? How many tweets about flu today? If the answer is an unusually high number, then something may be occurring. If so, answers to more questions can identify duplicates and refine geography. If the counts remain sufficiently high, then authorities will need the identities of the people for consultation and testing.

An answer to the initial "how many" question is a single value, like 3, 5, or 20. Sharing a single number would not seem to pose much concern, but looks can be deceiving. In the United States, HIPAA governs the sharing of medical data from hospitals, physician offices, medical labs, and those directly involved in patient care [15]. The safe harbor provision of HIPAA allows hospitals, physician offices, medical labs and those involved in patient care to provide daily answers to a “how many” question, if they want to do so. Access to non-health data relies on agreements with participating stores, schools, and online companies. Will they participate? As we approached organizations, one by one, whether medical or not, they voiced a similar concern. Providing a single value that answers a “how many?” question may reveal sensitive information about their operation, and so, organizations refused to participate because of confidentiality concerns. I use the term privacy for individuals and confidentiality for organizations. Joe Lombardo, head of the bio-surveillance integration effort at Johns Hopkins, reported that even schools did not want others to know how many children were absent on a given day over concerns that such information might adversely impact the school. So, a privacy-preserving approach must protect both the privacy of an individual's personal information and the confidentiality of an organization's strategic or sensitive information. An aggregate count may offer privacy protection for individuals but does not guard the interests of organizations.

Figure 2 illustrates the confidentiality concern of organizations. Each school is asked to provide a count of the number of children absent that day. In Figure 2a, each school forwards its count directly to a public health department, but in doing so, the public health department knows the private count of each school. This poses a confidentiality problem for each school. In Figure 2b, the schools are organized in a communication circle with the public health department at the start and end of the circle. The public health department makes up a number (1025 in this example), which it sends to the first school. Each school receives a value, adds its private count to the value it received, and then passes the sum to the next school. The last school sends its sum (1442 in this example) to the public health department, which subtracts the original made-up number (1025) to learn the total number of children absent (417) in the 6 schools. The protocol used in Figure 2b does not allow the public health department to learn the private counts of any schools, but two schools can collude to learn the private count of another school. For example, in Figure 2b, School 1 and School 3 can share information to learn School 2's private count. School 1 knows it forwarded 1037 to School 2. School 3 knows it received 1162 from School 2. Together, they know School 2's private value must be 1162 minus 1037 or 1162-1037 is 125.

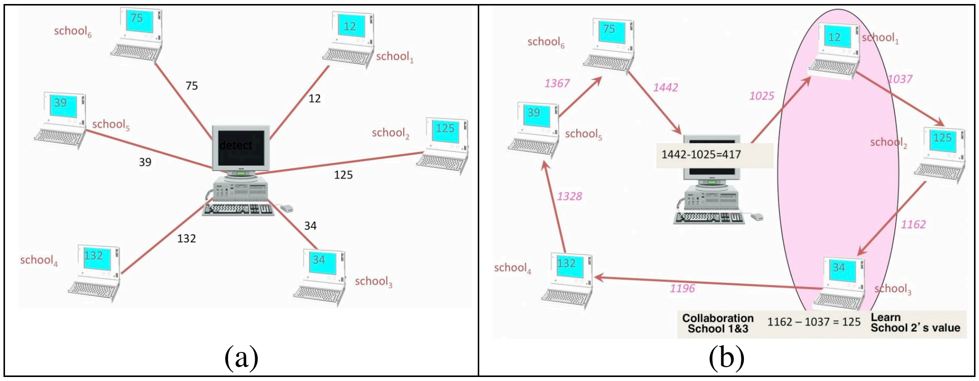

Figure 2. Sharing aggregate counts of school absenteeism can compromise school confidentiality when shared: (a) directly with a central authority (e.g., a public health department); or, (b) adding values around the network. In (b), the central authority starts the cycle with the made-up number 1025, which it subtracts from the final total to learn the aggregate count but Schools 1 and 3 can collude to learn School 2's private count.

Distribute computation not data

It is hard to imagine how you proceed with privacy-preserving surveillance if you cannot even guarantee the privacy and confidentiality of aggregate counts. Must we share personal data widely to be safe from bio-terrorism? Or, can we leave personal data in organizational silos and use a network across the silos to jointly compute answers with privacy guarantees? I needed a solution to the "how many" problem to show distributed computation could work.

Obsessed with finding an answer, I started talking about my dilemma widely. There are more computer scientists per square foot of university space at Carnegie Mellon than at any other place. Computer science is so big at Carnegie Mellon, it is not a department, as it is at most schools, but is itself a school within the University. An advantage of having hundreds of computer scientists around is that if you encounter a problem, you can talk about it with others who have more knowledge of other areas of computer science. So, while I was grappling with the problem, I shared my paradox with anyone who would listen, including some of the nation's finest theoretical computer scientists. I even presented the problem to students enrolled in my course on privacy and anonymity in data. One of the students, Samuel Edoho-Eket, got an idea.

“Watch the quiet ones”, I recall someone telling me once. Sammy was always quiet in class but in one-on-one conversations after class, he was so insightful I often wondered whether he was the most brilliant student I had ever encountered. By the time I presented my problem to the class that day, Sammy and I had gotten into a routine of talking after class. I looked forward to those after class discussions when Sammy and other students would continue talking about course content, sometimes longer than the class had actually met. Sammy was easy to talk with too. He has a golden smile that radiates his face and emits so much warmth that he is naturally one of those people that everyone likes. I could tell Sammy really liked this problem. For weeks, he kept returning to it until our after class discussions seemed to always include some brainstorming about it. One day, Sammy reached an “aha” point. I could see it in the sheepish grin looking back at me throughout class.

Sammy's idea was simple. The thing about simple ideas is that they always seem obvious in hindsight. “Of course” you say when you learn the answer, but simplicity often results from profound thought. Sammy adapted an approach from computer science called trivial secret sharing [21] to include a privacy-preserving summation operation. Those are fancy words, but his idea was straightforward, and more importantly, it worked. We named it the “PrivaSum” protocol [22]. It describes a set of rules each participant follows.

Here's how an online network of organizations follow the Privasum protocol to jointly answer “how many” questions. First, each organization splits its count into a set of positive and negative numbers that when summed give the actual count. Each number in an organization's set is called a share, so adding all the shares back together gives the organization's count. Next, each organization distributes its shares among some of the other randomly selected organizations in the network but keeps one share private. Then, each organization combines the shares it receives from others with its private share and reports an overall sum to the public health department (or some other central authority). Finally, the public health department adds the sums from all the organizations to learn the total answer to the "how many" question. Neither the public health department nor colluding organizations can learn the specific count of any particular organization (unless they all conspire against it).

When we presented Privasum to the class, students wanted a concrete example. Figure 3 shows how PrivaSum works to compute an aggregate count of school absenteeism without compromising school confidentiality.

Inspired by the success of PrivaSum, I went on to develop other protocols that compute different kinds of aggregate information across a network of participants. These include PrivaMix, which allows others to track where people have been without knowing who they are [23] and RandomOrder, which allows participants to randomly arrange themselves without a third party [24].

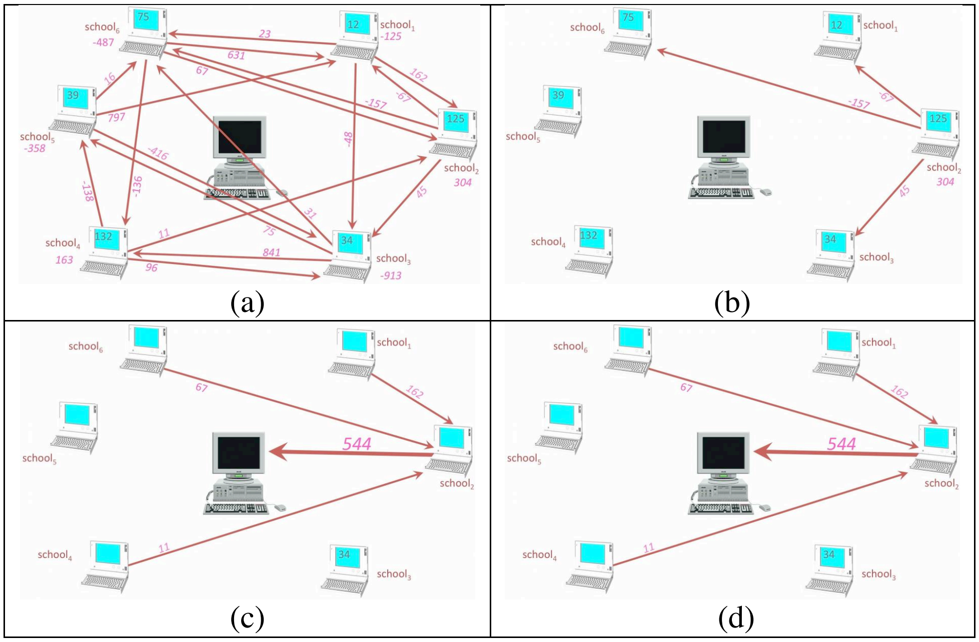

Figure 3. A network of schools uses the Privasum protocol to jointly compute aggregate absenteeism without compromising school confidentiality. (a) Each school produces a set of numbers that sum to the school's private count and then forwards all but one of the numbers to other schools. For example, (b) School 2 has 125 students absent. It forwards 162, 67, and 11 to other schools and keeps 304 private. (c) Each school adds the values it received with its privately held value and forwards the sum to public health, which adds the values received to learn the total absenteeism (417) across all schools. For example, (d) School 2 receives values from other schools, which it adds to its private value, and forwards the sum 544.

Data versus Answers

"We want the data to do the computing ourselves."

Despite our success with Privasum, some officials said to forget aggregate counts. Let society choose. Lawyers clambered for an expanded interpretation of authority to have data holders simply forward personal data to public health for bio-surveillance. This was not to assault personal privacy, though that may be a consequence, but because they saw public health as seeking to do good, as responsible data stewards, and they saw no other available solution that allowed them to immediately proceed with emerging bio-surveillance technology. "If an event does occur, public health law allows us to get the data anyway", they reminded.

When it comes to resolving clashes between technology and society, choices quickly become all-or-nothing propositions. There is a false belief that society must always choose between the benefits of data sharing (or new technologies) or risk personal harms like a loss of privacy (or other societal norms). Bio-surveillance is no different. Either copies of sensitive personal information must forward to public health all the time for us to be safe, or the information never goes at all and we remain vulnerable. This is wrong thinking.

The all-or-nothing proposition is not always politically intentional. More often, it is a natural consequence of lawyers and other decision-makers not being technologists and technologists usually seeking use of the new technology. But technologists can play another critical role. They can architect alternatives, even new technologies that better harmonize an offending technology with society. Usually no technologist is positioned to do this kind of big thinking and the all-or-nothing proposition arises with little time for contemplation. But I had been brewing on bio-surveillance and privacy for some time. This was my opportunity to show that a technologist can respond to the "privacy or utility" question with a "privacy and utility" answer.

Figure 4 shows a graphical depiction of the space of solutions limited to trading utility for privacy (red). The greater the privacy provided, the less utility available, and conversely, greater utility requires less privacy. Solutions in the red space are the first to which people jump, but are not the only available solutions. Alternatively, society can achieve better outcomes when seeking utility and privacy (blue). These sweet spots abandon the traditional trade-off proposition in favor of architecting a better solution.

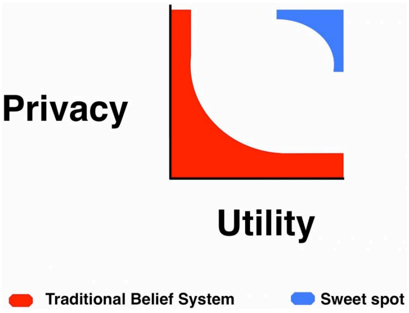

Figure 4. In technology society clashes, the traditional belief system limits solutions to those that trade utility for privacy (red) but better choices exist in the sweet spots of utility and privacy (blue).

Reveal data as evidence warrants

How do we get to the sweet spot in bio-surveillance? Aggregate counts alone are not enough. If counts are high, public health will still need personal details to assess whether a biological attack, or even an outbreak, is underway. Computer scientists might be able to compute answers to ever more sophisticated questions over a network of data holders so that data can remain in silos, but officials lacked patience. How does privacy-preserving bio-surveillance get data to public health beyond aggregate computation?

My insight was to think of bio-surveillance as a stepwise optimization problem. A bio-surveillance investigation unfolds in steps or stages. Each stage determines whether available data supports the belief that an outbreak is underway. If not, the investigation ends. If so, the next stage answers the same question using more detailed and sensitive data. The investigation advances through a series of distinct steps until it either abruptly ends at an early stage because there is no longer reason to believe an outbreak is underway or proceeds through all the stages to confirm there is an outbreak underway. The final stage, if the investigation gets that far, uses the most sensitive data. To optimize utility and privacy, each investigatory step uses the least sensitive data necessary to make a determination, and further steps proceed incrementally only if the prior steps have shown scientific evidence to do so. Described in this way, the investigation imposes a natural ordering for data sharing, from the least to the most sensitive. This design, which became known as *selective revelation*, guarantees that public health only receives the least sensitive or least identifiable data needed.

I shared this insight with students, and they of course, wanted a more concrete depiction. Figure 5 shows a sliding scale that matches investigation status to data sensitivity.

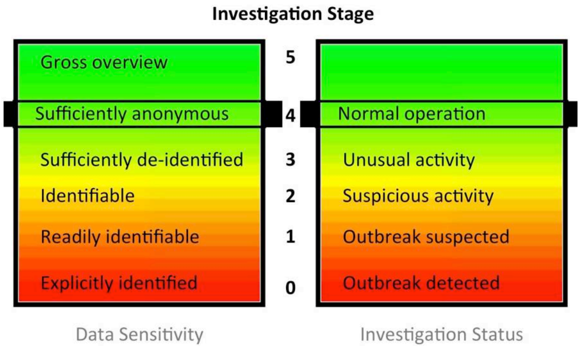

Figure 5. Selective revelation provides a sliding scale that matches data sensitivity to investigation status. As the investigation drills downward, more sensitive (shown here as identifiable) data becomes available. Imagine public health launching an investigation each day to decide whether an outbreak is underway. The first step (Stage 4) uses the most general data –the aggregate counts. If the count is low, then the investigation concludes for the day. On the other hand, if there is a sufficiently high count, then more detailed data forwards for the next step (Stage 3) in the investigation. This information is more sensitive than aggregate counts, but is the minimum necessary for the investigation to make its next determination. Data sharing continues in this stepwise manner, until the investigation ends because there is not enough evidence to proceed. Or, if each data release supports continuing the investigation, then eventually, the most identifiable data forwards. Only the last step (Stage 0) forwards the kind of identifiable data required under public health law when an outbreak presents. In all the earlier steps, less identifiable or less sensitive data forwards. Public health only receives the least identifiable data needed each day. On most days, the counts will be low for an active investigation, so nothing beyond the aggregate counts forwards.

That's the basic idea, but now comes the hard part. I have to sweat out the details with my harshest critic, me.

”*How are decisions made*? *How does data forward*?”

Technology drives it all. Public health uses technology to analyze information. Disparate data holders use technology to store and even compute among themselves information that public health wants to analyze. And technology mediates access between the two. Figure 6 shows these three layers. The middle layer mediates requests from public health and approves data flows from controlled data stores as appropriate.

”*How does it work*?”

Technologically enforceable policy statements make real-time decisions. A policy statement encapsulates the conditions for data sharing agreed upon by the originating data holders and public health. They codify legal contracts, certifications, and agreements between the original data holders and public health. Pivotal to the data sharing decision is a quantified result from public health reporting the expressed need for the information. The idea is that public health may be making on-going requests, but only if the quantified request exceeds a threshold does the data forward to public health.

Humans are not completely out of it. A policy statement may additionally or alternatively require password approvals from one or more humans.

Ah, when I present this to my students, they will want an English interpretation of a policy statement. One might read something like “if the Algorithm from Public Health reports a value greater than Threshold, then upon Approvals, release Data to Public Health using Transport Mechanism.” The parties agree beforehand to the specific definitions of Algorithm, Threshold, Approvals, Authorizations, Data, Transport Mechanism, and Public Health. All data decisions, whether data forwards or not, are logged for hindsight review.

My students will also want a concrete diagram. Figure 6 shows more details.

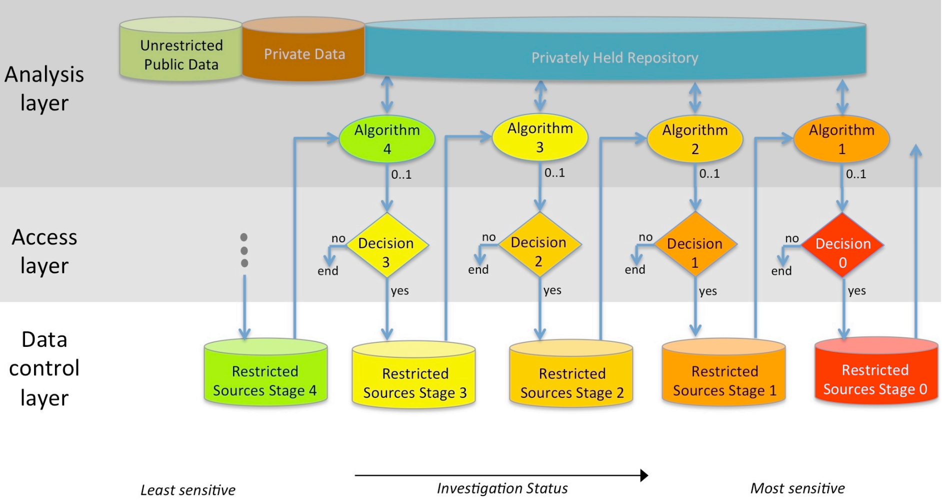

Figure 6. System diagram of a selective revelation system having any number of stages that count down to 0. Five stages appear above, numbered 4 to 0. As the stage decrease, the sensitivity of the data increases.Each stage has an algorithm that receives data for its stage. The algorithm may additionally use other public or private data to make a determination. The result of the algorithm is a value between 0 and 1, inclusive, that reflects the likelihood of an outbreak. The closer the value is to 0, the less likely an outbreak is underway. The closer the value is to 1, the more likely an outbreak is occurring. Public health provides the algorithms for each stage. The algorithms may be the same or different at each stage. A technically enforceable if-then-else statement decides whether to continue the investigation, and thereby release data to the algorithm in the next stage.

”*Do we have any legal standards like it now*?”

I liken the decision-making criteria to how a law officer gets a search warrant. Of course, daily data sharing decisions for bio-surveillance are not about the police, criminal law, or judges. If a biological threat agent is found, the matter would become one for law-enforcement, but that's not the point here. The criteria for search warrants poses a good technological model.

In American jurisprudence a law officer wanting to intrude on a person's private life or affairs needs a search warrant, which a judge can issue. The officer appears before the judge and reports either facts for which he or she has first-hand knowledge or facts that he or she learned through an informant. Typically, the judge uses a two-prong test to make a decision: what is the basis of the knowledge, and is the source believable (see Figure 7a). Selective revelation recreates this process in technology. We replace the officer with anomaly or data-mining algorithms, the informant with data from various sources, and the human judge with a decision-making policy statement that enforces prearranged data sharing policies (see Figure 7b).

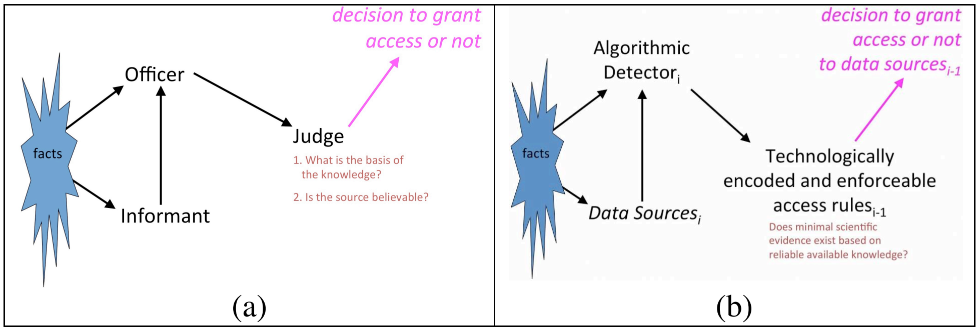
decision to grant access or not
Officer
facts
Judge
1. What is the basis of the knowledge?
2. Is the source believable?
Informant
(a)
decision to grant access or not to data sources$_{i-1}$
Algorithmic Detector$_i$
facts
Technologically encoded and enforceable access rules$_{i-1}$
Data Sources$_i$
Does minimal scientific evidence exist based on reliable available knowledge?
(b)

Figure 7. (a) probable cause test for a search warrant in American jurisprudence (b) selective revelation test for sharing more sensitive data for surveillance.

My inner critic is satisfied. Now come the next critics. Whenever I think I have a solution to something, I ask my students and close scholars to tell me how I am wrong. Point out the flaws. I say, "it is better for your family to tell you that your clothes have holes before you leave home, than have others do so in public." I first learned the value of this at MIT. Whenever I had an idea, I would go to the lounge and start writing my idea on the board. Soon, students and faculty would stop what they were doing, emerge from their offices, and with great zeal shoot down whatever I had to say even if I was right. If I could withstand the onslaught, I was good to go. If not, back to work I went. I offered up my design for selective revelation and answered the critics.

”*Can technology make that many decisions each day*?”

Human judges make similar determinations on law-enforcement investigations daily. If human judges had to make the data sharing decisions, it is doable for humans. There would likely be one investigation per county with no further decisions on most days. An active day could require up to 5 decisions in the same day. When we shift to a technology lens, Google's online ad network alone makes more than 30 billion decisions per day to match ads to viewers and each is done in the time it takes to load a web page [25]. Clearly, technology can handle the decision-making load.

”*Is electronic transfer safe*?”

In 2014 alone, data breaches affected more than 927 million consumer records [26]. Online access does increase risk, but the risk seems manageable through ongoing adherence to cyber-security best practices for encryption, authentication, and transport. Of course, data transfer in selective revelation might use online transfer for less sensitive data and other transport means, such as courier, for the most sensitive data.

”*What about privacy*?”

Selective revelation requires decomposing surveillance into progressive stages. Each stage must have an algorithm that uses the minimum data necessary to quantify the state of reliable knowledge in the investigation at that point. In order to know whether the best algorithm is being used and that the least sensitive data forwards, a description of the algorithm must be public. This provides oversight for data use and drives innovation for increasingly better algorithms. During operation, however, the stages of any particular investigation are not necessarily public. Selective revelation's privacy guarantee is that the least sensitive data needed for the investigation forwards to public health. A simple illustration makes this point. Rather than forwarding identifiable data each day, most days would just forward aggregate counts using Privasum or a similar protocol to protect personal privacy and corporate confidentiality. This dramatically reduces the amount of sensitive data shared. Logging all data requests allows data holders and the public to review the credibility of reported algorithmic results and the performance of the overall system. In these ways, selective revelation provides utility, privacy, transparency and accountability.

Example of a privacy-preserving bio-surveillance system

"*The design is good, but can you actually build one? You always say the devil is in the details*," reminds one student.

"*What algorithms does public health use? What information sharing do real laws allow*?" asks another student.

"*There are thousands of medical diagnoses and there would be few early cases. How do you find the needle in the haystack, better yet to find it wearing a privacy blindfold*?" Questions from students roared.

"*Sorry professor, but can you be more concrete*?"

Okay, here is a walk-through of a system similar to the one that was under construction in the pilot at Johns Hopkins University in the Washington DC area on 9/11 when the planes hit [12].

I start with algorithms. Public health already had algorithms for aberration detection. After all, public health had been in the business of identifying outbreaks for quite some time. Five notable algorithms led the pack [27, 28, 29, 30, 31]. Captain Tracee Treadwell and her team at the CDC packaged two of these, the historical limits method and Cumulative Sums, into a software program known as the Early Aberration Reporting System (EARS) [32,33,34].

Just before 9/11, Captain Treadwell and team used EARS for drop-in surveillance at large public events in the United States such as the 2000 Democratic and Republican National Conventions and the 2001 Super Bowl, and immediately after 9/11, at the 2001 World Series. On 9/11, Captain Treadwell was in the only civilian plane allowed to fly after the planes crashed [35]. It was a plane with CDC personnel and Captain Treadwell took EARS with her. It became the standard day-to-day surveillance system for the New York City Department of Health and several health departments in the Washington, DC metropolitan area after 9/11. Since then, EARS has been used throughout the world.

For convenience and without loss of important details, I crudely summarize the methods in EARS as comparing the number of cases reported in the current 4-week period to historical data from the preceding 5 years (though the number of weeks and years may vary). If the current value is more than 2 standard deviations greater than the historical average, sufficient evidence exists for concern. (See [34] for the specific calculation.) Figure 8 shows a CDC example of EARS retrospectively detecting an alarmingly high number of reported cases of breathing difficulty in Knox County, Tennessee in 2002. The yellow circle shows days for which the values were more than two standard deviations above the historical average for that time.

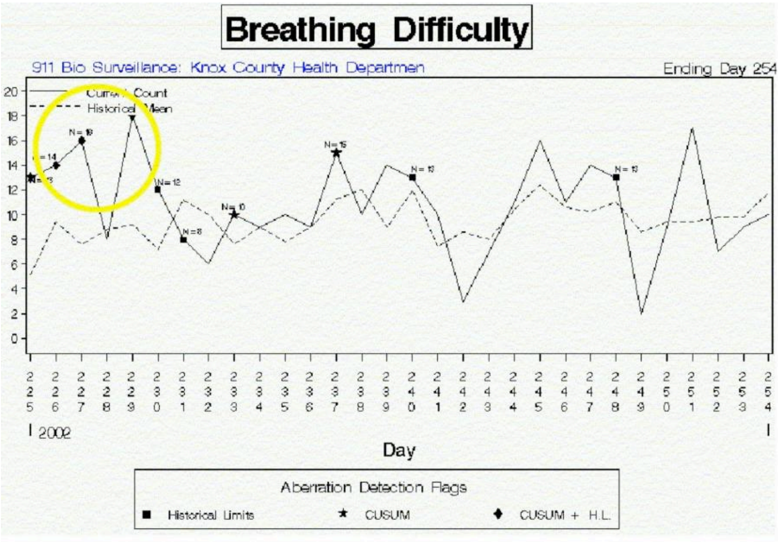

Figure 8. Example of the CDC's Early Aberration Reporting System (EARS) retrospectively identifying a sufficiently high number of reported cases with respiratory difficulty in Knox County, Tennessee in 2002 using 911 call data. The yellow circle identifies values that are more than 2 standard deviations above the historical average [32,33,34].

I adopt EARS as the algorithm for our surveillance example. Notice that EARS only computes on counts. The investigation will start each day using aggregate counts at the county level, and as investigation proceeds, the algorithm remains the same, but the counts relate to increasingly smaller units of geography and age.

EARS requires historical data. Public health departments often receive information from each patient's visit to a hospital or physician office. This information, which includes demographics, diagnoses and procedures [36], arrives months after the event, making it useless for real-time surveillance but it is sufficient to continuously compute historical norms and to learn whether an outbreak was missed. In a recent survey, 33 of the 50 states not only collect detailed medical information on each visit to a hospital and physician office, but they also sell or give away copies [37]. You can acquire this information and construct your own retrospective bio-surveillance system for most states.

EARS analyzes syndromes but medical data uses diagnosis codes and the written text of the complaint. For convenience, I will restrict this example to diagnosis codes. When a medical doctor makes a diagnosis, he uses a coding system known as the International Classification of Diseases (ICD-9) in order to get paid from an insurance company [38]. These codes don't just identify the disease, they also embed knowledge about its manifestation. For example, there are at least 107 ICD-9 codes relating to pneumonia alone, each describing different types.

Which diagnosis codes are of interest? Lt. Col. Julie Pavlin and her team had already solved that part of the puzzle. Her team at the DOD had already built a bio-terrorism surveillance system that was routinely operating on data from military personnel worldwide [39]. They made the significant contribution of mapping hundreds of diagnosis codes into a small number of syndromes [6]. Their goal was to improve algorithmic performance, but if hospitals report syndromes rather than diagnoses, privacy improves too. Here's why: if the number of cases reported for a syndrome is 1 patient, the diagnosis for the patient is not exact; it could be one of many possible diagnoses related to the syndrome. If the number of cases for a syndrome is more than 1, different patients collapse together into the same syndrome group, which makes multiple patient records indistinguishable. How much collapsing occurs? Here is a count of the number of diagnosis codes for each syndrome: respiratory has 71 diagnoses, gastrointestinal 125 diagnoses, neurological 95, fever 87, localized cutaneous lesion 71, rash 44, botulism-like syndrome 41, hemorrhagic illness 33, lymphadenitis 26, and sudden severe illness or death has 14.

Last comes access to real-time data. Public health laws in the United States are local and agreements with data holders often rely on community relationships. There is no one-size-fits-all national version for bio-surveillance though arrangements can be made with national chains for information about over-the-counter purchases, online services for web search statistics [40], and so on. For simplicity in this example though, I will use counts from schools and hospitals and seek further data from hospitals alone.

Because state laws vary, I primarily focus on relevant federal regulations in this example. The Family Educational Rights and Privacy Act (FERPA) allows schools to share aggregate counts greater than 2 with public health (and anyone else for that matter). [14]. HIPAA allows hospitals, physician offices, medical labs, and any other entities directly involved in patient healthcare to forward identified health information to public health without authorization for surveillance [41]. Under HIPAA, public health can receive identifiable health information like the kind in Stage 0 each day and forgo the all the described privacy provisions.

Figure 9 shows a system summary. Investigation begins with schools and hospitals forwarding aggregate information. Schools use the Privasum protocol to jointly compute total absenteeism yesterday, and hospitals and physician offices use Privasum to report the total number of cases appearing yesterday per syndrome. These aggregate counts forward to the public health department, which uses EARS to compute countywide results. If the investigation continues to Stage 2, hospitals and physician offices forward counts of {syndrome, age range} pairs occurring in the county. In Stage 1, hospitals provide {5-digit ZIP code, age, syndrome} for each person having a diagnosis in a surveyed syndrome. If the investigation advances to the final step, Stage 0, then the medical records of affected cases forward.

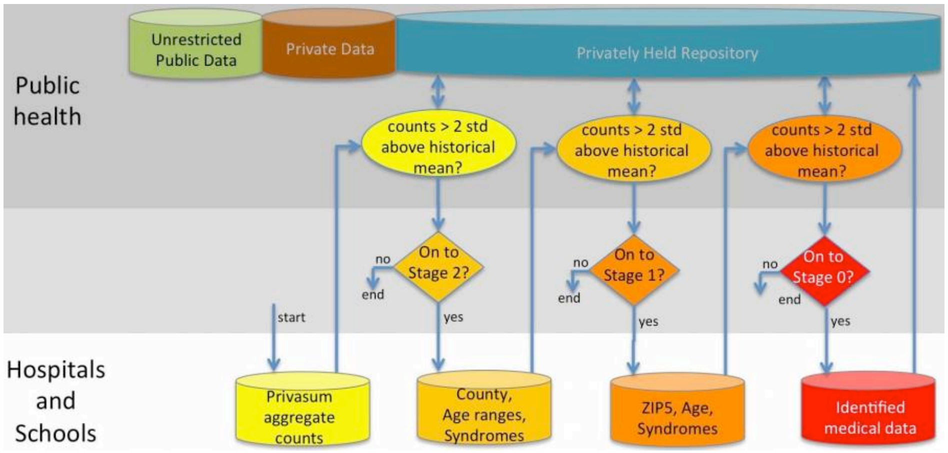
Public health
Unrestricted Public Data
Private Data
Privately Held Repository
counts > 2 std above historical mean?
counts > 2 std above historical mean?
counts > 2 std above historical mean?
no
end
On to Stage 2?
yes
no
end
On to Stage 1?
yes
no
end
On to Stage 0?
yes
start
Hospitals and Schools
Privasum aggregate counts
County, Age ranges, Syndromes
ZIP5, Age, Syndromes
Identified medical data

Figure 9. Example of a privacy-preserving bio-surveillance system using 3 stages to share information from hospitals and schools with public health.

Whola! … a privacy-preserving bio-surveillance system. For the privacy guarantee, only the most aggregated data needed to make a stepwise determination forwards to public health, and the investigation concludes without forwarding any additional data if there is insufficient evidence. For the utility guarantee, the system performs as well with the generalized data as it would if the most sensitive data forwarded.

Unbounded surveillance

Like most people, I first heard of Admiral Poindexter during news coverage of the Iran-Contra affair [42], an illegal activity in which the Reagan Administration was selling arms to Iran and diverting the proceeds to insurgents fighting the Marxist Government in Nicaragua. Admiral Poindexter, a leader in the effort, was forced to resign his position as Senior Security Advisor to President Reagan and was convicted on April 7, 1990, of five counts of lying to Congress. The convictions were reversed in 1991 on appeal because Congress gave him immunity for his testimony [43]. This kind of outside-the-rules behavior in the name of a cause is often the stuff that rebels, renegades, villains, saints and martyrs are made.

In the time before 9/11, while I had been developing privacy-preserving surveillance, unbeknownst to me, Admiral Poindexter had been brewing a data surveillance system of his own to combat terrorism. Project Genoa aimed to produce a computer program that could rapidly analyze and share data, and make investigation plans based on the analyses [44]. The work was not guided by a hypothetical attack by a biological threat agent. Instead, it was motivated by the actual terrorist attacks to which Admiral Poindexter had responded while he worked for the National Security Council years earlier –namely, the Beirut attacks in 1983 [45]. The work was not guided by societal protections. Instead, the architecture assumed a secretive regime, operating outside of public knowledge and inspection, in the knowledge-is-power, cloak-and-dagger community in which Admiral Poindexter worked.

I met Admiral Poindexter after 9/11 on the topic of surveillance. Despite his military background and title, he was unlike the military stiffs that had launched me into bio-surveillance. He didn't wear a uniform. Instead of barking commands, he talked softly with a jolly and warm demeanor. More often than not, his face sported a smile. The Admiral tended to choose his words carefully, but when he did seem to let his conversation flow freely, it was clear he was a true believer. He had unyielding loyalty to a cause, and that cause, in that moment, whether right or wrong, was unbounded data surveillance on the American public, a cause for which I believed he would commit Japanese hara-kiri and slice himself with a ritual sword rather than fail.

Does he know how to build a surveillance system? Admiral Poindexter's doctorate is in Nuclear Physics [45]. Years earlier, he had reportedly revamped the Situation Room in the White House with the latest technology, and in the process, garnered a bit of a reputation as an information technology go-to person in the circles of the National Security Council [45]. That is very different from having the kinds of experience or credentials one usually finds in a computer systems architect. From a computer science perspective, it is more appropriate to say, he had a vision for a system, or given the Admiral's level of commitment, a passion for it. So, he is not the technical architect but a visionary with practical domain knowledge. His desire was to garner all the data from all avenues to ferret out terrorists and to predict terrorist activities early. “With so much data, everything is likely to be an anomaly”, I blurted. With no pretense at scientific grounding, he just shrugged, and likened the process to detecting submarines in a sea of noise. That is something he knew how to do. Yes, I thought to myself. But locating submarines works because we know the signature we are looking for. A submarine is not like other forms of ocean life, so we can exploit its differences. A terrorist among us is a neighbor or acquaintance living daily life. Surveying the entire ocean for a suspicious fish requires figuring out what makes the fish suspicious. If we know that, we can efficiently gather only the data we need.

After 9/11, our divergent approaches would entwine in a government program like 2 strangers on a blind date that become a tragic couple. In the end, both will die, but one will be resurrected.

9/11 triggers TIA

The events of 9/11 did not fit the script of our bio-surveillance preparedness. Planes crashed into buildings and anthrax appeared a month later by mail [46]. After 9/11, the hypothetical became real. Admiral Poindexter became emboldened. The military officers, who had risked asking a bunch of academics to build a privacy-preserving surveillance system, likely received more chest decorations. But in the country's flight from fear, dogmatic commitment to American ideals became the past.

Months after 9/11, the DOD pushed our bio-surveillance project into a larger research program called Total Information Awareness (TIA and later re-named Terrorist Information Awareness [47]). Admiral Poindexter led TIA. The program sought to combine disparate silos of personal data on all Americans. One goal was to combine as many forms of personal data as possible, such as online activity, commercial transactions, personal health and credit data and individual bank and academic records with law-enforcement information on all individuals living in America. The other goal was to develop computer programs that could identify and predict suspicious activity and behaviors within the data that may be associated with terrorists.

TIA provided minimal funding for my effort, so privacy research remained in their portfolio, but from that point forward, privacy was no longer a necessary and sufficient condition for surveillance, as it had been with the military officers who recruited me. Privacy was no longer a property that a surveillance system must maintain, but an obstacle to overcome. Figure 10 shows an early depiction of TIA presented by Admiral Poindexter's group. Privacy appears as a thin red dashed line through, under and around which data flows to become useful. Not all data flows require it- the diagram is complete without it. I worked along the dashed line. I wasn't alone. They funded a few other researchers who worked on developing specific privacy techniques, such as immutable audit logs [47], but only my work was architectural. Only my work aimed to coherently weave privacy into surveillance. At least, that was the work I wanted to do.

From the beginning, Admiral Poindexter and Bob Popp, the second in command at TIA, appeared schizophrenic about privacy. I had the impression that keeping Americans safe was most important to them and they worried that any redaction or missing information might become the critical missing piece. After all, Admiral Poindexter had found that early signals of a major terrorist attack on Marines in Beirut had gone unnoticed [45]. What if the signals were not even present in the system? Privacy requires foresight and a slowing down of access to sensitive information using fidelity, temporal, and threshold filters. Admiral Poindexter seemed to never trust that experts would know which data could be slowed.

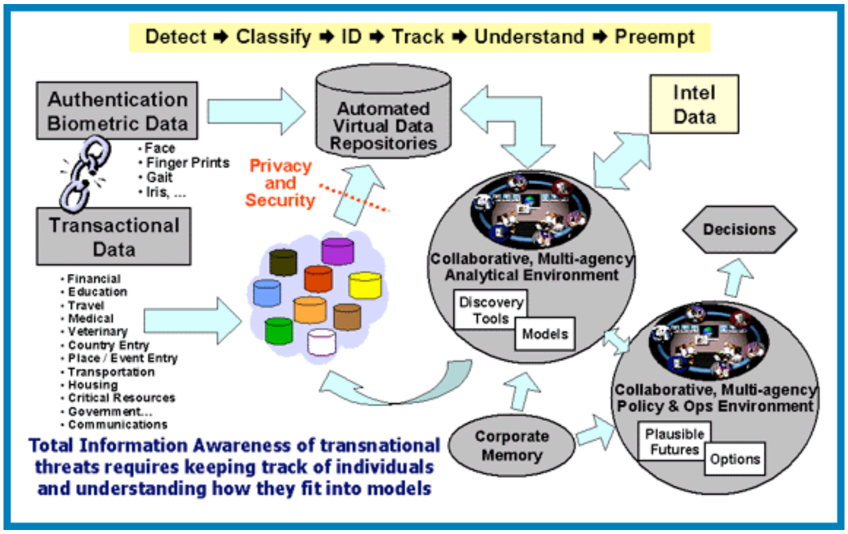
Detect ➡ Classify ➡ ID ➡ Track ➡ Understand ➡ Preempt
Authentication Biometric Data
• Face
• Finger Prints
• Gait
• Iris, ...
Transactional Data
• Financial
• Education
• Travel
• Medical
• Veterinary
• Country Entry
• Place / Event Entry
• Transportation
• Housing
• Critical Resources
• Government...
• Communications
Privacy and Security
Automated Virtual Data Repositories
Intel Data
Collaborative, Multi-agency Analytical Environment
Discovery Tools
Models
Decisions
Collaborative, Multi-agency Policy & Ops Environment
Plausible Futures
Options
Corporate Memory
Total Information Awareness of transnational threats requires keeping track of individuals and understanding how they fit into models

Figure 10. Diagram of Total Information Awareness system, taken from official (decommissioned) Information Awareness Office website at the time [48].

Work in TIA often blurred the distinction between ongoing surveillance and research. Data for research purposes has different privacy standards than data gathered daily in real-time surveillance of the public. The first carries ethical risks and the second may additionally have consequences to liberty. Data under many different privacy regimes were combining for uncertain uses. As an FBI official warned, "if we get data and it shows any criminal activity, even casual drug use in medical records, law-enforcement can prosecute." TIA was on a fast track with no time to sort this out.

TIA financially supported the continuation of my work with Johns Hopkins and improved names used in my work, but they never allowed me to do the privacy preserving architecting for counter-terrorism that I felt destined to achieve. After an 18-month head start and significant progress on a related form of surveillance, I felt uniquely qualified, or at least I seemed more qualified than anyone else I knew.

Admiral Poindexter and Bob Popp responded to my presentations with the universal nod of appreciation, so I kept expecting the chance. Instead, they would rename parts of my work and apply concepts themselves in an ad hoc manner. Don't get me wrong. The new names were much better than the old. For example, Bob Popp coined the term “selective revelation”, a noted improvement over my original “reasonable cause predicate”. He tagged “privacy appliance” to describe a box on each data source that answered access requests and queries with privacy protections and computed aggregate computations. I was never sure whether he thought up these handles himself or re-purposed them from elsewhere. But we never got beyond names and concepts. Maybe they thought I wasn't ready or I couldn't be trusted. On the other hand, maybe they believed they just needed soothing words and plausible concepts and not science to address privacy.

The nature of TIA brought unrelenting privacy questions, even from within. Admiral Poindexter and Bob Popp couldn't just say TIA was research because of the simultaneous need to use the system. At research briefings, the two sat upfront with their silhouettes choreographing reactions. When a researcher highlighted new data sources or showed promise with a new algorithm, their shoulders tended to lean back with heads held upright. The same presentation might conclude with privacy barriers to accessing needed data or privacy concerns over uses of the algorithm. Their bodies shrank. Privacy was at every turn. I was certain that they would rejoice if privacy just went away.

They had to get ahead of privacy. Technology can solve privacy concerns with TIA, Admiral Poindexter and Bob Popp would assert to others, seeming to re-purpose concepts in my privacy-preserving approach with little consideration to appropriateness for counter-terrorism. Bob Popp once bounced an idea off me. He was noodling on the possibility of just encrypting all data values. This would create an alternative data universe of encrypted values. People would have data in the real world and an encrypted copy in the surveillance world. Every data element would be encrypted. Then, selective revelation would determine whether circumstances permitted a value to be unencrypted. Delighted to be included in an architectural discussion, I enthusiastically demonstrated flaws with this idea. One problem, I said, is that the information is truthful. Encrypted or not, distributions and combinations of values link to real data. Suppose a dataset has more males than females, say 80 males and 20 females. In the surveillance universe, the numbers would be the same. One encrypted value appears 80 times and the other 20 times, and so, we learn the encrypted values for male and female throughout the surveillance universe. I could then use this information about gender to continue and reveal other data values too. When I slowed my eagerness and examined his face, instead of him being impressed with my insight and wisdom, he looked horrified. I was now further away than ever from getting a chance to do the architectural work I wanted to do. Why did he think it was so easy? Was he trivializing the task or desperate for a band-aid?

Not even I could just use the privacy-preserving bio-surveillance approach for counter-terrorism without more in-depth study. Developing privacy-preserving counter-terrorism would require me to take another journey, similar to the one for bio-surveillance. I would need to understand relevant algorithms, data, methods and laws and extract knowledge from domain experts. Bio-surveillance is a top-down pursuit that surveys aggregate counts and drills down further into data details when evidence of unusual activity exists. Counter-terrorism surveillance is different. It integrates a top-down approach, like bio-surveillance but with different data and algorithms, and a bottom-up approach, like FBI investigations. In law-enforcement investigations, algorithms drill outwards in data from a piece of credible knowledge or a known person to identify connected people and gather related facts. How do you blend the top-down and bottom-up approaches?

I wanted to pursue privacy-preserving counter-terrorism. This time I was doing the asking. But this time, there were no military stiffs insisting on privacy. There was no interest at all. No one wanted to architect privacy into surveillance anymore. Everything changed on 9/11. When I started my pursuit, several entities wanted the authority to do surveillance. After 9/11, they all had it. The industrial military complex retooled itself overnight, emerging as data brokers and third-party surveillance partners, capable of acquiring and using data that the government could not obtain directly. Even public health jumped privacy when new laws and regulations post 9/11 broadened surveillance provisions to routinely allow public health access to identified personal information. Even public health abandoned the privacy-preserving option.

TIA ignites privacy bomb

Can the government do surveillance on the American public without privacy protections? It seemed a done deal to me. Then, in 2003, media accounts of TIA surfaced and generated grave privacy concerns [49]. Senator Feingold, Democrat-Wisconsin, introduced legislation to place a moratorium on data mining research and deployment efforts at the DOD. Senator Wyden, Democrat-Oregon, introduced a similar anti-data mining bill limited to TIA. A broad coalition of public interest groups, ranging from the American Civil Liberties Union to the American Conservative Union urged Congress to take action against TIA. [50]

Amidst the turmoil, Admiral Poindexter and his staff talked with me about privacy-preserving surveillance. I explained the capabilities, advantages and limitations of privacy-preserving surveillance and pointed to our findings. Discussion culminated into a critical in-person meeting. Realizing the importance of this one meeting, I garnered a team of professors from Carnegie Mellon to take with me. We rented a van and a professional driver drove us the 4 hours from Pittsburgh to DC, where we had the meeting, and then he drove us the 4 hours back. A key member of my team was Professor Raj Reddy, who had won the highly distinguished Turing Award (the computer science version of a Nobel prize), served as co-chair of the President's Information Technology Advisory Committee (PITAC) under President Clinton [51], and reportedly previously met Admiral Poindexter previously.

This was the climatic battle for privacy-preserving surveillance and I was armed to convince. No detail was overlooked. My attire transmitted military confidence. I sported a new suit, which I had ironed with extra starch for crispness, wore shiny black shoes to which I applied Vaseline for high gloss, and dawned a fresh short haircut. My talking points had been rehearsed and memorized, and my handouts were clear about our goal to architect privacy into emerging technologies. Our goal of developing socially responsible technology was not limited to surveillance technology, but also addressed new technologies under development at Carnegie Mellon, including face recognition, biometrics, video surveillance, mobile phone tracking, and GPS surveillance. Everyone engaged at the meeting. Mental calculations appeared to churn in the heads of Admiral Poindexter and his team. When the conversations ended, however, it seemed clear that they believed they could brave it without privacy.

Congress seemed to make a different calculation and abruptly ended research on TIA due to the privacy outcry [47]. The suddenness left many computer scientists, including me, without funds, notwithstanding scientific results and financial commitments to students.

NSA picks up the pieces

What can we learn from the abrupt end of TIA? Government surveillance must require privacy, right? Now, more than a decade after 9/11, media reports have surfaced based on leaked top-secret documents that describe widespread surveillance of individuals by the National Security Agency (NSA) [52,53,54]. Their goals appear eerily similar to those of TIA. When I first heard the news and saw the history of the programs, I thought of Admiral Poindexter. TIA suffered hara-kiri on the sword of public knowledge, but the Admiral's vision survived. Surveillance activities continued with secrecy, not privacy, as its operational property.

Does the NSA program stem from Admiral Poindexter's effort? In the demise of TIA, Bob Popp reportedly salvaged part of TIA by relocating select components to the NSA [45]. Why not the privacy part too? I think by not forwarding the privacy components, Bob Popp finally freed himself and the project from the albatross that had caused him so much pain.

There has been public outcry this time too, but Congress did not take the same kind of swift and sweeping action that was TIA's fate.

Some people falsely believe that to enjoy the benefits of technology, society must forgo past protections like privacy. They insist on trading one for the other, even when exclusivity is not necessary. Often there are sweet spots, solutions where society can enjoy both technical innovations and established protections against harms. Privacy-preserving surveillance exemplified this kind of win-win deal. But decision-makers held so tight to the false belief that privacy must be the enemy of surveillance that they unnecessarily killed privacy.

It seems arbitrary. The esteemed military officials that hurled me into bio-surveillance demanded privacy. What if Admiral Poindexter's team had opted for privacy too? Would we still have privacy in surveillance?

References

1. ***TODO*** planes crash on 9/11NY, DC, PA

2. Health Aspects of Chemical and Biological Weapons: Report of a WHO Group of Consultants. Geneva, Switzerland: World Health Organization; 1970:72,99. http://www.who.int/csr/delibepidemics/en/TableofContents.pdf

3. Brachman PS. Inhalation anthrax. Ann N Y Acad Sci. 1980;353:83-93. http://www.ncbi.nlm.nih.gov/pubmed/7013615

4. Biological threat agents. Federation of American Scientists. Accessed January 12, 2015. http://www.fas.org/biosecurity/resource/agents.htm

5. Syndrome definitions for diseases associated with critical bioterrorism-associated agents. Centers for Disease control and prevention. Accessed January 12, 2015. http://www.bt.cdc.gov/surveillance/syndromedef/

6. ICD-9 Syndrome Groups for Bio-terrorism Surveillance / prepared by Lt. Col. Julie Pavlin, U.S. Department of Defense Global Emerging Infections System (Washington, DC 2002).

7. Guillemin J. Anthrax: the investigation of a deadly outbreak. University of California Press. 1999.

8. Meselson M, Guillemin J, et al. The Sverdlovsk anthrax outbreak of 1979. Science. 1994 v266. pp1202-1208.

9. Carnegie Mellon University, University of Pittsburgh to establish Biomedical Security Institute to address bioterrorism, public health threats. University of

Pittsburgh. News Services. October 18, 2000. http://www.news.pitt.edu/news/carnegie-mellon-university-university-pittsburgh-establish-biomedical-security-institute-addres

10. Shea D and Lister S. The Biowatch program: detection of bioterrorism. Congressional Research Service Report RL 32152. November 19, 2003. http://www.fas.org/sgp/crs/terror/RL32152.html

11. Real-time outbreak and disease surveillance laboratory (RODS). Department of Medical Informatics, University of Pittsburgh. https://www.rods.pitt.edu/site/

12. Lombardo J, Burkom H, Pavlin J, et al. ESSENCE II and the Framework for Evaluating Syndromic Surveillance Systems. Centers for Disease Control and Prevention. MMWR. September 24, 2004. http://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a30.htm

12b. ***TODO*** insert. Lombardo J. The ESSENCE II Disease Surveillance Test Bed for the National Capital Area. Johns Hopkins APL Technical Digest v24. 2003. http://techdigest.jhuapl.edu/TD/td2404/Lombardo.pdf

13. Legally Reportable Diseases in San Francisco. San Francisco Department of Public Health. Accessed January 12, 2015. http://www.sfcdcp.org/reportablediseases.html

14. The Family Educational Rights and Privacy Act (FERPA) 20 U.S.C. § 1232g; 34 CFR Part 99. http://www.law.cornell.edu/uscode/text/20/1232g

15. The Privacy Rule of the Health Information Portability and Accountability Act (HIPAA). 45 CFR Part 160 and Subparts A and E of Part 164. http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/prdecember2000all8parts.pdf

16. Mariner W. Mission Creep: Public health surveillance and medical privacy. 2007. http://www-syst.bu.edu/law/central/jd/organizations/journals/bulr/volume87n2/documents/MARINERv.2.pdf

17. List of Registries. National Institutes of Health. Accessed January 12, 2015. http://www.nih.gov/health/clinicaltrials/registries.htm

18. Proctor R. The Nazi War on Cancer. Princeton University Press. 1999. https://www.nytimes.com/books/first/p/proctor-cancer.html

19. Survey of public health laws for bio-surveillance. 2002. ***TODO*** put online at foreverdata or lab website

20. Sweeney L. Privacy-Preserving Surveillance using Databases from Daily Life. IEEE Intelligent Systems, 20 (5), September-October 2005. http://dataprivacylab.org/dataprivacy/projects/selectiverevelation/paper2.pdf

21. Shamir A. How to share a secret. Communications of the ACM 22 (11): 612–613 - 1979

22. Edoho-Eket S. Detecting Bio-Terrorist Attacks and Naturally Occurring Outbreaks Over a Distributed Network While Protecting Privacy and Confidentiality: the PrivaSum Protocol. Carnegie Mellon University, School of Computer Science, Technical Report CMU-ISRI-04-111 ***TODO*** authorship question, put on-line at lab

23. Sweeney L. Demonstration of a Privacy-Preserving System that Performs an Unduplicated Accounting of Services across Homeless Programs. Data Privacy Lab Working Paper 902. Pittsburgh 2007, October 2008. http://dataprivacylab.org/projects/homeless/paper2.pdf

24. Sweeney L and Shamos M. A Multiparty Computation for Randomly Ordering Players and Making Random Selections. Carnegie Mellon Technical Report. 2004. http://repository.cmu.edu/isr/233/

25. Koetsier J. 30 billion times a day, Google runs an ad (13 million times, it works). VentureBeat Insight. October 25, 2012. Based on an analysis by Larry Kim of Wordstream of a sample of Google Data. http://venturebeat.com/2012/10/25/30-billion-times-a-day-google-runs-an-ad-13-million-times-it-works/

26. Crossman P. Eight lessons for banks from the data breaches of 2014. American Banker. December 2, 2014. http://www.americanbanker.com/news/bank-technology/eight-lessons-for-banks-from-the-data-breaches-of-2014-1071465-1.html

27. Centers for Disease Control and Prevention. Update: graphic method for presentation of notifiable disease data—United States 1990. MMWR. 1991. v40 pp.124–125

28. Harrington CP, Andrews NJ, Beale AD, Catchpole MA. A statistical algorithm for the early detection of outbreaks of infectious disease. Journal of the Royal Statistical Society. 1996. v159 pp.547–563

29. Hutwagner L, Maloney E, Bean N, et al. Using laboratory-based surveillance data for prevention: an algorithm for detecting Salmonella outbreaks. Emerging Infectious Diseases. CDC. 1997. v3 pp.395–400. http://wwwnc.cdc.gov/eid/article/3/3/97-0322_article

30. Stern L and Lightfoot D. Automated outbreak detection: a quantitative retrospective analysis. Epidemiology and Infection. 1999 v122 pp.103–110.

31. Simonsen L, Clarke J, Stroup D, et al. A method for timely assessment of influenza-associated mortality in the United States. Epidemiology. 1997 v8 pp.390–395

32. Hutwagner L, Seeman M, Thomson W, Treadwell T. CDC: Early aberration reporting system (EARS), presentation at National Syndromic Surveillance Conference, New York City, Fall 2002.

33. Burkom, H. EARS Java source code, Johns Hopkins University, 2003.

34. Hutwagner L, Thomson W, Seeman M, and Treadwell T. The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS). Journal of Urban Health. 2003. 80(2).

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3456557/pdf/11524_2006_Article_200.pdf

35. “CDC-1, cleared for approach” — Drop-in surveillance on 9/11 with CDC’s Captain Tracee Treadwell. Trust for America's Health. CDC. http://healthyamericans.org/health-issues/story/%E2%80%9Ccdc-1-cleared-for-approach%E2%80%9D-%E2%80%94-drop-in-surveillance-on-911-with-cdc%E2%80%99s-captain-tracee-treadwell

36. Hooley S and Sweeney L. Survey of Publicly Available State Health Databases. Harvard University. Data Privacy Lab. 1064-1. June 2013. http://thedatamap.org/1075-1.pdf

37. Sweeney L, Hooley S, et al. thedatamap.org

38. International Classification of Diseases, 9th edition. http://www.icd9data.com/

39. Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE). U.S. Department of Defense. Accessed January 12, 2015. http://www.med.navy.mil/sites/nmcphc/program-and-policy-support/essence/Pages/default.aspx

40. Google Flu Trends. Accessed January 12, 2015. https://www.google.org/flutrends/us/#US

41. Disclosures for public health activities. U.S. Health and Human Services. See 45 CFR 164.512(b)(1)(i). http://www.hhs.gov/ocr/privacy/hipaa/understanding/special/publichealth/publichealth.pdf

42. ***TODO*** Iran Contra and Poindexter

43. ***TODO*** decision reversal for Poindexter

44. ***TODO*** Syntek Project Genoa

45. Harris S. The Watchers: the rise of America's surveillance state. Penguin Books. 2011

46. American anthrax outbreak of 2001. Department of Epidemiology. School of Public Health. UCLA. Accessed January 12, 2015. http://www.ph.ucla.edu/epi/bioter/detect/antdetect_intro.html

47. Total “Terrorism” Information Awareness (TIA). Electronic Privacy Information Center. Archive. Accessed January 12, 2015. https://epic.org/privacy/profiling/tia/

48. Screenshot. EPIC Briefing on Total Information Awareness. Electronic Privacy Information Center. Archive. Accessed January 12, 2015. https://epic.org/events/tia_briefing/, https://epic.org/events/tia_briefing/tia_screenshot.gif

49. William Safire, “Dear Darpa Diary,” New York Times, June 5, 2003

50. Buderi, R. Our Surveillance Nation. Technology Review, 106 (3) April 2003.

51. President Clinton names Raj Reddy and Irving Wladawsky-Berger as Co-Chairs. The Networking and Information Technology Research and Development (NITRD) Program. News Release. August 18, 1999. https://www.nitrd.gov/pitac/media/press-18aug99.aspx

52. ***TODO*** Guardian articles, Snowden. 1/2 articles

53. Bobic I. NSA Fesses Up to Improper Surveillance of U.S. Citizens. Huffington Post. December 26, 2014. http://www.huffingtonpost.com/2014/12/26/nsa-spying-report_n_6382572.html

54. NSA Reports to the President's Intelligence Oversight Board (IOB). National Security Agency. Central Security Service. Accessed January 12, 2015. https://www.nsa.gov/public_info/declass/IntelligenceOversightBoard.shtml