Caution: Danger Ahead (with Big Data)

Matt Bishop
Dept. of Computer Science
University of California at Davis
1 Shields Ave.
Davis, CA 95616-8562 USA

email: bishop@ucdavis.edu

Abstract. "Big data" is revolutionizing our view of science, and has the potential to do the same for the social sciences and humanities. With the benefits come very serious potential problems, ranging from invasion of personal privacy to enabling spectacular failures of analytics. This discusses some of them.

Introduction.

"Big data" encompasses the gathering, management, analysis, and synthesis of very large amounts of data. Having tremendous amounts of data available enables much more detailed, and much broader, analyses than ever before; we can record large-scale astronomical data for later analysis, or probe the basis for life to understand heredity, diseases, and evolution.

For example, the Large Synoptic Survey Telescope project (LSST) [1] will use an 8.4-meter telescope in Chile to survey the visible sky every week with a three billion pixel camera, in order to enable scientists to analyze the changes in the sky over a period of years. This project will produce 30 terabytes of data *per night*. The data will be sent from Chile to Illinois, where it will be shared with collaborators throughout the world. Storing and providing the tools to search this data for astrophysical phenomena of interest will require the development of new methods, tools, and powerful storage and networking facilities.

Other types of big data involve personal or sensitive information. Retailers use "loyalty cards" to offer discounts to consumers and to collect information on what they purchase. Financial institutions inform credit reporting agencies about the state of consumers' finances, and in turn receive reports aggregating information from other such institutions (and, indeed, from creditors in general). In science, the study of the human genome has led to the collection of massive amounts of data about the biological constructs that encode personal characteristics of human beings. Medical data is aggregated and correlated to provide insight into the origins and causes of epidemics, and discern ways to stop or slow their spread.

The collection of any sensitive information raises serious questions about the effects of the propagation of data upon privacy—and big data exacerbates these problems. This is usually framed as the need to protect personal privacy. An equally interesting, but rarely raised, form is the need to ensure that the inferences acted upon are correct.

The Conflict between Privacy and Analytics.

Underlying both questions is a view of the way big data—and, indeed, all data—is handled. Access to the data itself is generally restricted by contracts, regulations, laws, and customs. Call the set of rules governing who can access the data, and how, the *privacy requirements*. Similarly, analysts examining data have a goal in mind; they perform certain specific actions to achieve that goal. Call the set of goals of analysis the *analysis requirements*. Note that these requirements may be very broad or narrow. For example, access to published data, or data resulting from a publicly-funded project, is usually available to all, so in this case the set of privacy requirements would be empty (that is, there are no privacy requirements). But access to medical records identifying a specific patient or set of patients is tightly restricted, for example in the U.S. by HIPAA, state laws, and other regulations. Here, the set of privacy requirements contains many elements (requirements).

When data is distributed to many entities, the distributor must take care that none of the privacy requirements are violated. This may be done contractually, but a more common technique is to *redact* (change or eliminate) the sensitive elements of the data in such a way that any recipients of the data cannot reconstruct the redacted information. This is called *data sanitization* or (when personal identities are redacted) *anonymization*. One issue currently being studied is how to balance the privacy requirements with the analysis requirements, and specifically, how to handle conflicts.

"Big data" introduces concerns beyond those of small data sets. The problem lies in the assumption that the adversary, who is trying to reconstruct the redacted information, uses *only* the data in the data set. In fact, an adversary has a wealth of information from external sources that she can correlate with the data set being studied. Two examples, and the lessons from them, are worth recounting.

Netflix provides recommendations to customers based on their past movie rentals (or on-line viewing). Netflix tailors its recommendations for each customer by analyzing what the customer has seen, and what others who have seen the same (or similar) movies. Their algorithm is of course proprietary. In their quest for improvement, Netflix issued a challenge: given a set of Netflix data, could someone develop an algorithm that predicted what a customer would choose better than the Netflix algorithm? The prize for doing so was $1,000,000. The data set, which was made publicly available, was a collection of records corresponding to customers' orders. Each record initially consisted of a user identifier, movie title, movie rating, and date and time of the rating. Before distributing the data, Netflix anonymized the user name by replacing it with a pseudonym [2].

Researchers at the University of Texas–Austin used the data in an unexpected way. They found similar information at the web site IMDB.com, which contains information about movies, including user comments and ratings (which are date- and time-stamped). The researchers used a statistical algorithm to find probable matches between the Netflix data and the movie title, movie rating, and date and time fields of the IMDB.com data. When they found a match, they had the IMDB.com user name[1] that matched the Netflix pseudonym for the record, effectively identifying the user to whom the record belonged (to the granularity of the IMDB.com user identifier, of course) [3].

This illustrates the tension between privacy and analysis requirements perfectly. In order to provide external researchers with enough data to develop more accurate algorithms, Netflix had to supply information that, *when combined with external data*, allowed the IMDB.com identifier of the users of some records to be derived. Here, the privacy and analysis requirements were in conflict; meeting one set of requirements meant that the other set could not be fully met.

The second example also involved data publication. On August 3, 2006, AOL researchers released 21,000,000 search queries involving 658,000 users. The records had the users' names anonymized with numeric IDs. Their intent was to provide a source of information for research on search algorithms. On August 7, 2006, AOL executives learned of the release, and ordered the data removed from the web [4]. On August 9, 2006, the New York Times published a story entitled "A Face Is Exposed for AOL Searcher No. 4417749" [5] that identified the user who asked the queries.[2] The user made queries for landscapers in that town, about a particular subdivision in the county, about several people with the same last name as the user, and so forth. The reporters quickly narrowed the search, visited the city, looked at property records and homeowner names and addresses, and found the most likely searcher. They then interviewed the suspected user, and, quite non-plussed, the user confirmed they were correct.

Had the data set contained fewer of the queries of user no. 4417749, the reporters' sleuthing would have been far more difficult and most likely would have failed. Thus, the quantity of data helped them identify the user, and thereby violate the user's privacy.

During their study of the queries in the data set, the reporters noted many of the user's queries were about medical conditions such as bipolar illness, the effect of nicotine on the body, hand tremors, and others. A logical conclusion was that the user had several serious medical conditions. But that was incorrect; indeed, the user was quite healthy. The explanation turned out to be simple: a retired nurse, the user would use her medical knowledge to search the web for information that would help friends and neighbors with various conditions, or who were interested in various medical conditions. The reporters had drawn the wrong inferences. This illustrates the second danger: where one examines large data sets and draws erroneous conclusions from it.

The Underlying Assumptions of Analytics.

Complicating the use of data are two assumptions: that the data is complete enough, and accurate enough, to obtain accurate results from the analysis.

[1] Probably not a real name, but one usually consistent among IMDB.com movie reviews.

[2] The reporters identified the user by name, city, and state. To avoid further intrusion into the user's privacy, anyone interested can read the New York Times' article.

Incomplete data is a serious problem. The best example of the failure of analysis it can cause occurred before the bombing of Pearl Harbor, which brought the United States into World War II. In the final months before the U.S. entered the war, it had reason to believe that at some point the Japanese Empire would attack, but where and when was unknown. In August 1941, an ostensibly German spy code-named Tricycle, who was in reality a double agent working for the British, came to the U.S. with a questionnaire that the Germans has passed to him and that he had shown to the British. The questionnaire, destined for the Japanese, asked him to report on general conditions of military readiness in the U.S.; but one part asked very detailed questions about Pearl Harbor, including requesting a sketch of parts of the military base and harbor. Tricycle and his British controllers passed the questionnaire to the FBI, but for some reason the questionnaire never reached the relevant analysts, who would have immediately realized its significance: that the probable target of the Japanese attack would be Pearl Harbor, and that plans for the attack were well under way. Here, the data that the U.S. analysts preparing for war had was incomplete; and the missing information, which was in possession of the FBI, would have helped complete the analysis [6, 7].

Misleading data can be equally dangerous, especially when those supplying the data desire to deceive the analysts. During World War II, Operation Mincemeat was an Allied effort to trick the Axis powers into believing that the first attack on Europe would take place at Sardinia, not at Sicily (the real target). The British ensured that apparently real data reached the German military through a strategm involving the body of a high-ranking British officer carrying highly secret papers. The papers indicated the target was Sardinia, and described a plan to make the Axis think the actual target was Sicily, so they would move their forces away from Sardinia! Thus, the Axis discounted any information they received about plans and preparations to attack Sicily, believing them to be false—and the Allies landed with few casualties and quickly conquered Sicily. The data that the German espionage organization, and the military, took to be genuine was in fact false, and led to a defeat [8].

Thus, analysis of data has two sets of requirements—privacy requirements and analysis requirements—and two underlying assumptions: that the data is complete enough and accurate enough to do the desired analysis accurately.

As shown by the Netflix example, the sets of requirements may conflict and when they do, someone must decide which requirements are to be met. This is usually based on the data provider's perceptions of the importance of the consequences of not being able to meet all the analysis requirements, or not being able to protect the privacy of the sensitive parts of the data.

The decision of what data to sanitize is based in part upon the perceptions of the data sanitizers: what resources does the adversary have? What information does the adversary have access to that could recover the redacted data? And if we modify the data so that the adversary would draw false conclusions, how does that affect the ability of the analysts to draw accurate conclusions?

The AOL example extends this to the realm of drawing false conclusions from the analysis. AOL did not intend to mislead those who analyzed the data. But the data contained information that the analysts misinterpreted: specifically, that the search query conveyed information that applied to the searcher. The problem was that the medical queries of user no. 4417749 *did* indicate something about the user: that the user had friends interested in those medical conditions. But the analysts thought that the user *had* the conditions. So they drew incorrect conclusions from their interpretations.

The intent of Operation Mincemeat was to mislead the users of the data (indeed, the "Twenty Committee" existed to mislead the enemy). Here, the trick was to provide false data that the adversary would have no choice but to accept as legitimate; indeed, the officers running the operation were very concerned about the "cover" being "blown", and took precautions to make that as unlikely as possible. A number of factors helped them, including knowing quite a bit about the environment in Spain where they made sure the body would wash ashore. They has reason to believe that the contents of the briefcase would get to the resident German agent quickly; they also made sure they could tell if the briefcase had been opened. And they also planted other false information to enable the Germans to validate the earlier information. The plan worked.

Finally, in our society, we often isolate information on a "need-to-know" basis. This principle (in computer security, called "separation of privilege") limits the spread of information to those who need the information to do their job. The obvious question is: who decides? Information that is meaningful in one context (that the Axis powers want information on the state of preparedness of the U.S., as Tricycle's questionnaire showed) may be equally meaningful, but with a very different meaning, in another context (that the Axis powers want detailed information about Pearl Harbor to help plan an attack, as that same questionnaire shows). Add to this the human impulse to restrict information, and the legal, regulatory, and career consequences for allowing sensitive information to spread too widely, and you have a problem of data being so restricted that it is interpreted in one way, rather than in a large number of ways. And those without access either produce the wrong conclusions from the data they do have access to, or they draw no conclusions, when the missing data would have led them to very different, or more complete, conclusions.

Big Data and Risk.

Given this view of data, restricting data becomes a problem of risk analysis rather than mathematical certainty. Given a set of sanitized data, one can determine mathematically whether an adversary can recover the redacted information assuming the adversary *only* has access to that data. But this is unrealistic; given the widespread proliferation of information from many sources, especially the Internet, the data providers must assume the adversary can access data sources unknown to them. So, the proper question is not: can an adversary uncover the redacted data from this dataset? The proper question is: what data does an adversary need to uncover the redacted data? Then the providers can try to determine the risk that they have incorrectly identified that data,[3] and (assuming they have correctly identified the data), the risk of the adversary finding that data and using it. So privacy becomes a matter of risk assessment, but based on availability of data.

The same is true for analysis, but the problem is much more difficult. With privacy, as the conclusion (redacted data) is known, one may be able to determine the data needed to recover it. But for analysis, the conclusion is unknown; indeed, were it known, no analysis would be necessary. So establishing the effect of incomplete or erroneous data upon the goal of the analysis is not possible. One can apply risk analysis to determine the sensitivity of the results of the analysis to errors in the data (or data sampling, which is often used in the analysis of big data). But handling missing data assumes that one knows the data is missing—an entirely different proposition.

Conclusion.

All this uncertainty is cause for both joy and care. Joy, because big data offers us the opportunity to improve our understanding about the universe of which we are a small part, and of the world and society in which we exist, and improve both our understanding of those and our ability to improve our society and how people live. Care, because we can draw conclusions from big data that reveal information about us that should be private; and a recognition that the conclusions we draw may be wrong due to incorrect or missing data. And that, perhaps, is the best lesson of big data: no matter how much data we have, the information in the universe, about physical phenomena, about people, about society, and about life, is of necessity incomplete. So we must tread with case, and be careful not to be seduced by the belief that we know everything, or that what we know is without question correct.

References.

[1] Large Synoptic Survey Telescope, http://www.lsst.org; viewed September 7, 2013.

[2] The Netflix Prize, http://www.netflixprize.com/rules; viewed September 7, 2013.

[3] A. Narayanan and V. Shmatikov, “Robust De-anonymization of Large Sparse Datasets,” *Proceedings of the 2008 IEEE Symposium on Security and Privacy* pp. 111–125 (May 2008).

[4] S. Hansell, “AOL Removes Search Data On Vast Group of Web Users,” *New York Times* (Aug. 8, 2006); available at http://query.nytimes.com/gst/fullpage.html?res=9504E5D81E3FF93BA3575BC0A9609C8B63

[3] As an example of incorrectly identifying the data, suppose personal records have names and addresses redacted, but leave in the 5-digit ZIP code, gender, and date of birth. Sweeney [9] estimates that 87% of the population in the U.S. could likely be identified by these three characteristics alone—something that is quite unexpected.

[5] M. Barbaro and T. Zeller, Jr., “A Face Is Exposed for AOL Searcher No. 4417749,” *New York Times* (Aug. 9, 2006); available at http://www.nytimes.com/2006/08/09/technology/09aol.html

[6] J. C. Masterman, *The Double-Cross System*, Avon, New York, NY (1972).

[7] C. Gentry, *J. Edgar Hoover: The Man and the Secrets*, W. W. Norton & Company, New York, NY (1991).

[8] B. Macintyre, *Operation Mincemeat: How a Dead Man and a Bizarre Plan Fooled the Nazis and Assured an Allied Victory*, Harmony Books, New York, NY (2010).

[9] L. Sweeney, *Simple Demographics Often Identify People Uniquely*, Data Privacy Working Paper 3, Carnegie Mellon University, Pittsburgh, PA (2000).