The Value of De-Identified Personal Data
Transcript

Part 1: De-Identification Methods and Tools

Stephanie Losi: Welcome to the CERT Podcast Series, Security for Business Leaders. The CERT program is part of the Software Engineering Institute, a federally funded research and development center at Carnegie Mellon University in Pittsburgh, Pennsylvania. You can find out more about us at cert.org. Show notes for today's conversation are available at the podcast website.

My name is Stephanie Losi. I am a journalist and graduate student at Carnegie Mellon working with the CERT Program. I am pleased to introduce Scot Ganow, Corporate Privacy and Ethics Officer at Verispan LLC, and Mike Hubbard, an attorney specializing in privacy issues at Womble, Carlyle, Sandridge and Rice, PLLC. Today we'll be discussing challenges companies face in de-identifying personal data that may be disclosed to third parties. Rather than identifying an individual but protecting his or her sensitive data, de-identification means that confidential or sensitive data can be disclosed so long as the person's identity is removed.

So, Scot and Mike, I'd like to jump right in here and sort of ask you, you know, basically what are some of the most common methods of masking personally identifiable information to comply with data privacy laws?

Scot Ganow: Well, there are a variety of ways that people can comply, from one extreme completely applying what we call the Safe Harbor principles to data, that is effectively removing all direct and some indirect identifiers from a dataset.

Stephanie Losi: So what does that leave?

Scot Ganow: That leaves a piece of data that's not capable of identifying individuals, which is good from a privacy and security perspective, but it doesn't have a lot of business or public health value or whatever your sector might be. There's not a lot of use to the data if you can't assign some individual attributes of a person to the dataset.

That is why de-identification, or at least de-identifying data and still retaining some individual characteristics that allow one dataset to be different from another, we think provides the most value to your business, to whatever you're trying to attempt, because the data is not anonymous; it's de-identified, and there is a difference that we recognize very clearly and capitalize on in our business in providing services for our clients.

Mike Hubbard: There also is the possibility to introduce noise or fuzziness to the data to make it more difficult for an intruder—in the industry-speak, someone who's trying to re-identify data and shouldn't be doing so—make it more difficult for that person to try to re-identify.

And then as well, restricted access and use agreements and controls are a very important aspect of rendering information not identifiable. And by that I mean on the one extreme you could have a dataset that you could post to the Internet; on the other extreme you could have data that's tightly controlled with contract agreements with real enforcement teeth in them and a trusted recipient of the data, researchers, et cetera, who agree to protect the data, not provide it to third parties, and to only use it for the stated purpose and agree not to try to re-identify any individual.

Scot Ganow: One other thing that I noted as a method is commonly what we call aggregating data, rolling data up to a higher level to where you still can obtain some value from it, but that's pretty

much—it's a one-shot thing. So you may roll up to a percentage of a population that does X, Y, Z. That's great. You get some value. But without the data at the individual level, you don't have the flexibility to go into other studies or other cuts of the data, if you will. We can't go back and find other uses for the data.

Stephanie Losi: So you're saying that de-identification allows you to actually provide these individual records, and what you're just taking out is, you know, who is the individual?

Scot Ganow: Correct. You are either removing completely or modifying to some extent to some standard the direct or indirect identifiers attached to an individual to where you're able to still attribute individual characteristics. Maybe their purchasing habits, maybe the prescriptions they use, maybe their height and weight.

Stephanie Losi: So what resources are required to accomplish these various methods?

Scot Ganow: There are a variety of tools out there. At Verispan we use a de-identification engine that at a very high level applies technical and statistical applications to data via a specific proprietary key and algorithm that generates the dataset that we're describing that has the business value, some modified attributes of a patient without identifying the individual. By no means is that the only way to do that, but at least in our space, which is healthcare and healthcare information, the law that dominates the privacy landscape for us is HIPAA, and it makes it very clear what you need to do.

Mike Hubbard: You know, in some respects one answer to your question is the human brain. For example, I could be a physician rounding at the hospital and another physician consults with me and I might say, "Well, two years ago, I had a patient present with that same condition and here is what I did." And I am speaking about an actual individual and I'm mentioning attributes, health attributes about that individual, but in that case, I would have complied with HIPAA based on just the Safe Harbor of HIPAA.

And to elaborate just a bit on that, if I may, you know, HIPAA does give us the most regulatory legal detail and framework for determining when information is not individually identifiable. HIPAA Privacy Rule contains a standard on that, and there's basically two ways that you can determine if information is de-identified. The first and the easiest is the so-called Safe Harbor approach, and my example of the doctor does fit that Safe Harbor. The only information I mentioned about the individual was a year, "two years ago," and the Safe Harbor basically says that if you remove eighteen types of identifiers about the individual, household members of the individual and the employer of the individual, and if you don't have actual knowledge that the information could be used to re-identify an individual, then it's de-identified and you literally can post that information on the Internet. And so that's one simple way. It's a pretty conservative approach, and that means that, you know, the risks are very small that you could re-identify an individual.

Often there's a need for information that's more granular and more rich for research purposes, medical purposes, and then the resource that you need if you want to go that route is an expert statistician. In many cases you'll have legal review as well. The fundamental requirements are to not try to re-identify an individual and to only use the information for purposes stated in the agreement.

Scot Ganow: I liked your point about, you know, the human brain and the things that you can do. As a privacy officer, one of your roles is educating people on ways they can better protect their personal information. Just because people ask for certain sets of information doesn't mean you need to give it. Just because you have the dataset doesn't mean you necessarily need to provide

all elements to it. But from a privacy principle, that's a great bedrock of good privacy practice: is "Only use what you need," and use discretion.

Part 2: Getting Value from De-Identified Data

Stephanie Losi: All right, so with that in mind, how can companies get real business value out of de-identified data?

Scot Ganow: The value of the data at the patient level, as we just discussed a little bit earlier there, really is unlimited. You know, for example, in our space when you have data at a patient level, you're able to tap into a completely different set of uses for that data. If you've got a part of your business that uses identifiable data, maybe it's one-third of it, the other two-thirds never get to see that data. You de-identify it, then you have the ability to share it. Then you have the ability to get consensus and get your marketing people and other people involved with the data to truly capitalize on one of your greatest assets in your enterprise.

There's also the benefit obviously of keeping your data, or specifically the identifiable data of your consumers or your customers, from being disclosed inappropriately. If the data is de-identified, it unburdens you of so many legal responsibilities under certain security laws to report and take action, et cetera. If the data can't identify anybody, you don't have to do anything there and, again, you ultimately protect your customer as well.

And then of course you can get into all the other benefits of the data. Maybe you're looking at purchasing trends. Where did they spend their money throughout the month, throughout the year? You're able to do this now because you have the data, one, securely; but two, you have it at every possible place within your dataset because if you truly de-identify and provide the ability to link those records de-identified, you can create—again, you exponentially increase the value of your data because not only are you getting it from one silo, if you will, you can get it from multiple ones. But I think any industry can take advantage of that data once you've unburdened yourself of the identifiable nature of the data.

Mike Hubbard: Yeah, and some examples that I can think of that our firm is involved with in ongoing projects could be comparing health plan metrics. I'm a large employer and either measured nationally or in my region. How am I doing in my health plan compared to other employers, employees both in terms of how well is the population, what are our costs, where are our higher costs? That kind of information can really help you target interventions to improve your patient population, your plan member population.

Verispan of course is involved with providing de-identified data to pharmaceutical companies so they can see how their sales are doing, how competitors' sales are doing, what regions have different market dynamics. Verispan of course has provided data to the Center for Disease Control, and there's a use of the data there for national security purposes to have a baseline of data for example of anti-flu medications prescribed in an area. If you suddenly get a spike in prescriptions in one ZIP code in June and you can measure that against the baseline demographics, that's obviously a notable event that requires public health and other attention.

As a financial services company, you could track a de-identified customer and the buying habits or the purchasing habits, the financial services types of products that that de-identified consumer utilizes and you can better know your customer basically, not in the sense of one-to-one identifiable marketing which is a very big thing these days, but market segmentation and customer segmentation. This demographic of this type of customer reacts best to this type of information or marketing campaign.

So there's really a wide variety of uses of de-identified data and many of them for the public good. I think it is important to mention that in the HIPAA Privacy Rule when it was first published in December of 2000 as a final rule, the comments to the rule by the agency Department of Health and Human Services were replete with statements advocating the public benefits of de-identified data and encouraging greater use of de-identified data.

Scot Ganow: Yeah. And I think from the business perspective, I'm one of those privacy officers that sees privacy and security as a business enabler, not a sales prevention tool. And Mike just hit on it. You're taking care of your customer. The better you understand your customer obviously you can make better products and services available to them.

However, there are studies very clearly that show as well when it comes to customer loyalty, taking care of people's data, and that includes de-identifying it and making it unusable to someone who would do bad things with it, is all about showing value and building trust with your customer. And in this day and age, especially with everyone competing for the consumer dollar and their attention, those are very powerful things to have in your toolkit to hold on to customers, to get them in the first place and then hold on to them should you have a breach. If you've handled security effectively and there's no harm to an individual because you de-identified data, you have a better chance of holding on to that customer long term.

Part 3: Managing Risk

Stephanie Losi: So how can the organization really tell when it has de-identified data to a sufficient level, and how can it sustain that level of confidence?

Scot Ganow: It really depends on your business. It depends on the contracts by which you collect that data, and that includes whether you're buying data from another business, B2B, or the consent you have from the individual. That really determines your responsibility as to what you must do with the identifiable data, and that includes de-identification. HIPAA is really the only law that gives very clear standards. To answer your question, yes, I know I de-identify my data to a sufficient level legally under the law; however, you can't look at it solely with legal shades on. You have to look at everything as far as your business responsibilities, your commitments to your customer via a website or a privacy notice, whatever the case might be, to see, "Okay, have we done what we're supposed to do?"

And so, yes, there's many ways to look at it, and the first one is always legally to protect yourself and make sure you've met the standard regulating your business, specifically from a state and from a federal or international level, because when you get outside the U.S., the privacy requirements are much more stringent. But then you also look at how you're using the data, and it's a good basic question: Are you using the data in a way that you told the consumer or you told the other company that you were going to? Those are always good basic golden rules by which to decide and make those determinations, but I'm sure Mike will have a more legalistic approach to that answer.

Mike Hubbard: Does that mean you're punting to the lawyer here, Scot?

Scot Ganow: I'm not punting, but...

Mike Hubbard: No, you're not. You never did. How can an organization tell when it's protected data to a sufficient level? Yeah, from our perspective, if you are dealing with sensitive, individually identifiable information, you first have to look at: How is that information regulated? In some

cases, it may not be other than general consumer protection laws. That's in the U.S. Obviously, overseas in the EU and other countries and member countries of the EU, there is much more of a widespread regulation of personal information regardless of whether it's in the health sector or the financial services sector, et cetera. Here in the U.S., HIPAA does give us the best test that we have. It will not surprise me if the HIPAA standard is adopted in other laws when these issues arise either by the courts or by revisions to the laws and regulations. But, you know, once you figure out—let's just, to make the question simple, stay in the healthcare vertical—you don't need a statistician or a lawyer to figure out the Safe Harbor in my opinion. You remove the eighteen things about the individual household members and the employable actual knowledge and you've got a pretty good comfort level that it's de-identified.

Stephanie Losi: So let's say you Safe Harbored it and you take out these eighteen things. What kind of information can be left?

Mike Hubbard: Well, it could be that the de-identified patient received a prescription for Tamiflu last year. And I live in North Carolina, I'm a male and, you know, other types of transaction-specific attributes, which is really what's going on here in many cases. As long as those attributes don't get down to the Safe Harbor level of granularity, meaning I have to stop at year, I have to stop at state in terms of geographical subdivisions, those types of things. You know, I can't say, for the Safe Harbor, I was prescribed Tamiflu on February 2nd, 2005.

Stephanie Losi: Okay. Now, let me throw in a—this is maybe a bit tough. I have a question, though. What if someone has a very uncommon medication that is rarely, if ever, prescribed? What happens then? Is it still considered Safe Harbored even if you have this piece of data that may be very much an outlier?

Mike Hubbard: Well, that's a good question. Yeah, you can't have actual knowledge that the information could be used to re-identify an individual. And so, to put some context in your question, if I am the prescribing physician, if I know this is a cutting-edge research drug and it's only been prescribed to one person in North Carolina in my study, I might have actual knowledge. So even though I've stayed with year, I've stayed with North Carolina, so far I'm cool with the Safe Harbor, I might have actual knowledge and thus the information's not de-identified.

To take your question to the next level, those are exactly the kinds of questions that also would come up in the expert statistician opinion—the access to other reasonably available information which could include information and it's a very rare medication and it's only being prescribed at a center in Miami, for example.

Scot Ganow: Right. It's a very common challenge in the privacy space that's often called, you know, the Dale Earnhardt effect. You may remember several years ago the NASCAR driver was killed in a race by a very specific type of injury that was unique, and so if you saw that type of injury for a male in Daytona, Florida, in February time frame, there you go. That's why I always try to make it very clear in de-identification, there's no silver bullet to it because even if you do all these things, you start to look at the context in which the data was collected, who's going to see it, how it's going to be used, what publicly available data is there to avoid those types of situations.

Stephanie Losi: Okay, so what happens if something goes wrong? What should a business leader do then? Let's say, you know, you find out, you know, that an error has been made. You thought it was okay. It turns out not to be okay. What do you do?

Scot Ganow: That's a great question. When it comes to de-identification or risk to the data, we obviously make sure legally that we have contracts in place to make it very clear to the recipients

of our data. It's not just enough to keep the data, but if you know of any potential risks or breaches that occur, et cetera, et cetera, they have to report back to us so we can take the appropriate steps. And some of those steps are common privacy mitigation steps such as retrieving the dataset, understanding who potentially had access to it, destroying the dataset, or understanding the risks.

So kind of containing the threat, if you will, first and foremost is just pretty common in data privacy and security breaches. But then—and again, making sure these provisions are in your contract, working that back to say, "Okay, at a broader level, does it affect my large—you know, one particular dataset, does it affect the entire warehouse? Do we have to make some changes overall to the statistical principles?" And that's where it truly is a multidisciplinary approach to this process. You can't just make it an IT security thing. You can't just make it a someone with pharmaceutical knowledge thing or a lawyer thing. You have to look at it from every possible angle to assess the threat and then take proactive steps. But risk management is a very big part of the privacy process and de-identification as well.

Mike Hubbard: That's right. And I think, Stephanie, a lot depends on, you know, the types of facts that you're dealing with in terms of what if there is an error or something goes wrong? It could be as simple as, you know, out of a million records, one was bounced by a filter on a firewall because it might contain elements it shouldn't contain, and that's done in real-time. No human eyes have seen it.

But, in general, stepping back from that type of example, an entity should have a privacy incident policy, and obviously you'll want to immediately assess the impact, if any, of an event and implement immediate controls, quarantines, et cetera, if and as needed. If you think you're dealing with unlawful actors or intruders, involve forensics, review what your legal requirements are, if any, in terms of reacting to that event. And obviously it's an opportunity potentially for a "learning moment," as I share with my kids: but, you know, lessons learned and, you know, "How can we keep that from happening the next time?"

Scot Ganow: That's a great point. You know, from a privacy officer perspective, one of our key roles is education and training. And it's not enough just to have the policy, but make sure everybody understands it because when you have an incident, it's not the CEO or the CFO who's going to discover it; it's the frontline data analyst, it's the frontline technician who's going to catch it and realize it's a problem. That's step one, know it's a problem, and two, take the appropriate action to engage everybody necessary to resolve the issue. That's why training is so extremely important in addition to the policy at every level, because it only takes one person to create the weak link, if you will, for such a breach.

Part 4: Toward the Future

Stephanie Losi: All right. So to what extent do you find that organizations are really sharing their customers' personal information de-identified with outside third parties, and how can a third party best make use of this data?

Scot Ganow: That's a good question. I would say, I mean, and the reason we've brought the topic here to CMU is I think in general de-identified data's use is still very limited. I don't think a lot of people understand its value within their organic enterprise, much less third parties and things like that. But I will say it is one of the biggest challenges for any company anymore, regardless if you're dealing with de-identified or identified data, is you can't think of your four walls anymore. You have to think of what I like to call the privacy ecology, which is the full data cycle. You've got to think of where it's coming from and what you do with it at your company, where it's going.

Whether it's outsourcing, whether it's sales, whatever the case might be, it's extremely critical now that you make sure you've got the legal protections in place. Contracts aren't just for legal protection. They serve as a communication tool to make it very clear to your vendors, to your third parties, that: "This is our responsibility with this data. You represent us, and therefore you share in the burden of protecting this data." And it gives you the avenue by which to have dialogue with your vendors, your business associates, and your clients to all understand everybody's responsibility in the caretaking of the data. It's extremely important. And so obviously the contracts are very important. Training is very important. We often train our contractors on our data because they are an extension of us. And if you look at some of the enforcement actions taken in the privacy space lately, it's been very clear anymore. It's no longer a matter of what you have to do, compliance and the law, it's what you should do. What do you know is best for the caretaking of that data? And the same is true for de-identified data in our space. It is de-identified, to be clear, but we just don't toss it out there. There are still very clear responsibilities in place because it's certified as such with very low risk. If you don't manage everything else around it, that risk increases to an unacceptable level, so it really is a continuum of care, if you will, for the data. So any company is best served by making sure not just their people, but their vendors, anyone to whom they do business relationship with that involves the sharing of data, that they're up to the same standard.

Mike Hubbard: I think that's exactly right. And we are seeing I believe more applications and utilizations of de-identified data and data analyses in healthcare. Some of that I think is driven by the pharma industry and the, you know, tremendous budgets they have and the sophistication they have for doing market research. But we also are seeing applications in the retail sector where you see de-identified data being shared with third-party vendors who themselves might also have databases of de-identified data to do more sophisticated customer segmentation analyses, demographic analyses, learning more about what motivates a particular type of customer to purchase that type of thing.

Scot Ganow: There is obviously a commercial benefit. We've talked about benefiting your customers and improving your business as a result and even improving your relationships with your consumers, but there is a huge benefit for society as a whole when you have this data available de-identified upon which to make better responses. We obviously, you know, we focus on healthcare and we've talked about the public health benefits as far as epidemiology goes and bioterrorism goes, but there are just as many I think public benefits to the use of the data as there are commercial benefits as well. And so my hope is that people will encourage the use of that data and understand what it is and ultimately respect the choices of their individual consumers, but not lose sight that there is a huge benefit for towns, cities, countries as a result.

Stephanie Losi: So how do you see the privacy and data protection landscape as it relates to this issue evolving in the next year or two?

Scot Ganow: Well, I see it becoming very challenging in the next couple of years with respect to data privacy, and I guess my concern is there are just so many levels of compliance that you have to keep up with and it's very complex and very challenging. However, my other concern is, especially in the United States in this day of heightened security and the need for security, that there will be an over-correction with respect to the way we handle identifiable information and de-identified information for that matter that there will be an interest in restricting dataflow more than enabling it with the appropriate privacy and security safeguards in place. And we have to be careful about that specifically from a legislative perspective because we will literally be shutting off a data source that can truly benefit all of us.

Mike Hubbard: I agree with all of that. I think that we've had to think of two developments that we may see. One would be in laws other than HIPAA we may see more expressed recognition that there is something such as de-identified data. But also I think for a second trend perhaps is, you know, the market is going to drive the utilizations and applications with de-identified data, and there are some significant pain points out there that can be met. And CEOs of the big three automakers in the US meeting with President Bush for an hour, and the one thing they talked about more than anything else was the fact that they are less competitive because of the healthcare costs that we face here in the United States compared to other countries. And so we've got to find ways to improve the way we deliver healthcare, and if we can get the same job done with de-identified data that should be the way we go.

Scot Ganow: Another great example of that lately, you know, there's a variety, I think maybe thirty-four states that have security breach notification requirements, so companies are struggling with a way to respond to that. An example of an over-correction is, "Encrypt everything." Encryption is a wonderful tool, I'll be the first to say it, but it's not a silver bullet, and it hinders the use of the data. If it's encrypted, unless you have the key it becomes useless to you. And in some places it may make sense to encrypt data, and by all means we should. At the same time, you can overreact and overcorrect by encrypting everything. And those are some of the challenges privacy and security professionals have to consider and give good guidance to those making the laws. Mike and I co-founded the Carolina Privacy Officials Network, and that's one such topic that we're debating amongst a lot of leaders in the industry. How do we give them good guidance? How do we give legislators good feedback that is balanced and metered in its approach to definitely protect data, protect inappropriate disclosures, but not bring commerce and technology to a grinding halt because everything is encrypted?

Mike Hubbard: The example of Scot's point would be he lives in New York, he's going to a meeting in L.A., he's taking a laptop of data with him. Why take live production identifiable customer data if you could de-identify the data and get largely the same market value out of it when he's meeting with his colleagues in California?

Similarly, subject to the application and of course, you know, how systems are designed, there can be testing that can be done with de-identified data, not dummy data, but de-identified data. And, you know, one of the things that you can go off and hear privacy professionals preach is we need to be mighty careful about using production live data in a testing environment, and we've even heard stories of a server sitting under somebody's desk, you know, with a million records on it for testing, identifiable. So there are a variety of applications that say it's a legal compliance tool and a risk management tool and in many cases it's just the right thing to do, the data minimization concept which is, you know, codified as minimal necessary in HIPAA.

Stephanie Losi: All right. So where can our listeners learn more about this topic?

Mike Hubbard: There are a lot of resources here and on the University's website. And I should also point out that there's a paper that I have written that was published in West's Health Law Digest on how de-identification works under HIPAA and I'm happy to provide access to that. It's on our firm's website. It was published this past May.

Scot Ganow: And my contact information is also available, I think will be available in the bio as well. Always happy to entertain questions specifically about this as well. But I echo Mike's comments on the outstanding resources here at CMU with respect to this very topic.

Stephanie Losi: All right, well thank you very much. It's been a pleasure having you here.

Mike Hubbard: Thank you so much.

Scot Ganow: Thank you.