CERT'S PODCASTS: SECURITY FOR BUSINESS LEADERS: SHOW NOTES

The Value of De-Identified Personal Data

Key Message: As the legal compliance landscape grows increasingly complex, de-identification can help organizations share data more securely.

Executive Summary

More than 30 U.S. states now have passed laws that require companies to report unauthorized disclosure of state residents' personal data. In this environment, liability is high, but companies still need to work with user data. De-identification is one way to strike a balance by allowing companies to use and share data more safely, especially when third parties are involved.

Rather than identifying an individual and then using controls and countermeasures to protect 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.

In this podcast, 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, discuss the privacy challenges companies face and how de-identification can help.

PART 1: DE-IDENTIFICATION METHODS AND TOOLS

How to De-Identify Data?

There are several options for de-identifying data:

1. The safest option is the HIPAA (Health Insurance Portability and Accountability Act) Safe Harbor principle: removing all direct and some indirect identifiers from a dataset. However, the data then becomes much less useful.

2. De-identification can provide a balance by removing identifiers but still retaining some individual characteristics of the dataset — such as purchasing habits, height, or weight, for example.

3. An additional protection is to introduce noise or fuzziness to de-identified data. This makes it more difficult for any intruder to re-identify individuals.

4. Restricted access and use agreements can provide some assurance that users of the data will:

- protect the data
- not provide it to unauthorized third parties
- use the data only for the stated purpose
- not try to re-identify any individual

5. You also can aggregate data by "rolling it up" to a higher level — for example, presenting statistics by region rather than by state. But this reduces flexibility, because once it's done you can't go back to a more individualized level.

What Resources Are Required?

Several tools can be used to de-identify data. Verispan uses a de-identification engine.

Another tool is the human brain.

For example, if you were a surgeon consulting with another surgeon, you might naturally say, "I had a case like that two years ago, and here's what I did," but you wouldn't say the patient's name, hometown, etc.

To apply de-identification methods, guidelines are necessary. HIPAA provides some good guidelines for healthcare information. It says:

- You can use a Safe Harbor approach, which is not very risky at all, but the information may be less useful.
- You can consult an expert statistician and lawyer to determine if you have achieved an acceptable level of de-identification on your own.

A good bedrock of privacy practice is, "Use only what you need."

PART 2: GETTING VALUE FROM DE-IDENTIFIED DATA

What Is the Value?

The value of data at the individual level is nearly unlimited.

By de-identifying data, you can share it with others in your organization who normally would not have access to it, such as the marketing department.

You also can shed some of the liability that stems from laws requiring disclosure of privacy breaches. If the data is not identifiable, you don't have to report its loss. You are protecting your customer and yourself.

You can still track trends with de-identified data. In fact, you can link together information from multiple silos that you might not have been able to link before.

There are many example uses of de-identified data. Here are a few:

- You can track your company's health plan performance in various regions.
- You can track which medications are being prescribed in a particular area to get public-health early warnings.
- You can track customers' buying habits and which types of customers respond best to specific marketing campaigns.

Think Benefit, Not Cost

In this way, privacy and security can be viewed as a business enabler, not a sales prevention tool.

The better you understand your customers, the more you are able to make products and services that not only meet but anticipate their needs.

Also, the better care you take of your customers' data, the more they will trust you and your organization.

This is a powerful tool for holding on to customers in the long term.

PART 3: MANAGING RISK

How Far Should You Go?

How far should you go in de-identification? It depends on your business.

HIPAA is really the only U.S. law that provides clear standards.

So, don't look at it solely from a legal perspective. Look also at:

- your business responsibilities
- your stated commitments to your customers via privacy notices

These are good basic golden rules.

Also, it is possible that in the future, the HIPAA standard could be adopted by other market sectors and industries, so you can choose to use that as a guideline even if you're not in the healthcare industry.

And, of course, the Safe Harbor is a simple although relatively restrictive solution.

Tricky Situations

However, note that even if you stick to the Safe Harbor guidelines, if you have **actual knowledge** that an individual could be re-identified from the data provided, the information is still not de-identified.

One example would be an individual receiving an experimental drug that was only prescribed to one person in a particular state.

This is sometimes known as the Dale Earnhardt effect. The NASCAR driver died of a very specific type of injury, so if you saw that type of injury listed for a male in Daytona, Florida, in the February time frame, you would be able to identify him, even if the data were supposedly de-identified.

This is where you need the opinion of an expert statistician. There is no silver bullet in de-identification.

You need to keep in mind the interaction among various sources of data, both within and outside your dataset.

What If Something Goes Wrong?

Have contracts in place that describe the potential risk to the data.

Then, if something goes wrong:

1.) Contain the threat. This might mean:

- retrieving the dataset
- understanding who might have had access to the dataset
- destroying the dataset
- understanding the risks posed by exposure of the dataset

2.) At a broader level, understand the potential effects of the breach. Do you need to make any long-term changes to your statistical principles to avoid having this happen again in the future?

This is a multidisciplinary process. Look at the situation from every possible angle to assess the threat, and then take proactive steps to manage risk going forward.

Learning for the Long Term

To facilitate the incident response process, a business entity should have a privacy incident policy.

Use incidents as opportunities for "learning moments." How can you keep this from happening next time?

And it's not enough just to have a policy. Training is also extremely important. Make sure everyone knows and understands the policy.

Why? When you have an incident, it's the frontline employees — data analysts and technicians — who are going to discover it.

PART 4: TOWARD THE FUTURE

The Economics of De-Identification

Many people don't yet understand the full value of de-identified data.

But what is true is that you can't think of data staying within your company's four walls anymore. You have to think of the privacy ecology — the full data cycle:

- Where is it coming from?
- What do you do with it in your organization?
- Where is it going?

The Evolving Legal Landscape

It's extremely critical to make sure you have legal protections in place.

In terms of privacy, it's no longer a matter of what you **have** to do, it's what you **should** do.

The same is true of de-identified data.

De-identified data may be certified as low-risk, but if you don't manage everything else around it, that risk increases to an unacceptable level. There is a **continuum of care** for the data.

Make sure everyone is up to the same standard — employees of the company, vendors, contractors — effectively anyone with whom there is a business relationship that involves data sharing.

What the Future Holds

Benefits are not limited to the commercial sphere.

There may be a huge benefit for society as a whole from having a large, available amount of de-identified data.

Public health is only one example.

However, at the same time, the compliance landscape is becoming so thorny that there may be an overcorrection. There may be more movement toward restricting data flow, rather than enabling data flow with the appropriate privacy and security safeguards in place.

We have to be careful about that, specifically from a legislative perspective, so that we don't shut off a data flow that could benefit all of us.

The goal is to prevent inappropriate disclosures while also making sure commerce doesn't come to a grinding halt.

De-identified data can help.

- It can be used on an executive's laptop to reduce the risk of disclosure through theft.
- It can be used in testing environments instead of live production data.
- In many cases, it's just the right thing to do.

Resources

Hubbard, Mike, and David Wilson. "De-Identified Health Information: Legal and Practical Approaches to HIPAA Compliance." *Health Law Handbook 2006*. Thomson/West, 2006.
http://www.wcsr.com/resources/pdfs/Health_Law_Hndbk_Hubbard.pdf

---