Chapter 7: Finding Data Everywhere

In times like these when unemployment rates are up to 13%, income has fallen by 5% and suicide rates are climbing I get so angry that the government is wasting money on things like collection of statistics!

Hans Rosling, quoting a caller on a radio talk show, *The Joy of Stats*

When most people think of data, columns of numbers in a spreadsheet generally come to mind. We picture data as already existing and with a background in computer systems, people will naturally think of data being *produced by something*. Data is something we gather from the server or application log files, it’s something we download or scrape from a website or an interface somewhere. Even throughout most of this book, we’ve focused on how we collect and prepare data that already exists. But data exists all around us and everything produces data and the majority of it is never captured. While we can certainly learn a lot from the data that is waiting to be collected, we can also learn a great deal from generating and gathering data from our processes. One such process that offers some exciting opportunities to learn is breach investigations.

When organizations experience a security event, their natural reaction is to focus on getting back to normal. They see the event as nothing but negative and everything they do is to minimize the impact and move beyond it. The effect of the approach is that during that shuffle there is one extremely important benefit that is usually overlooked: a rich set of a data was generated during the event. If we could somehow gather that data and make sense of it we could learn not only about how we may be able to prevent the next event, but we could compare it to other events in the organization, or even to other organizations across the industry. Achieving that kind of benefit is the goal of this chapter, we want to figure out what data to collect and then how we can collect it. We will also discuss how we can analyze this data and even share it to get the most benefit from the data.

Need to introduce VERIS here.

Setting up the Research

First and foremost, we want to approach our breach data as if it were a research project. If we think of this as a “metrics program” or a special “security project”, we may think this effort is somehow unique to information security and it isn’t. This is all about data collection and analysis, something that has been done countless times before across many different disciplines. Not only would reinventing the (data analysis) wheel here be wasteful of our time and resources, we’d be laughed at and ridiculed by all the grown-up data scientists. Let’s avoid all that and treat this like a research project. And with all good research projects we begin with our goal and research questions.

Most of our work in this book has been of an exploratory nature. We worked with the data to see what it contained and then formed the questions we want to answer with the data. This effort is different because we are starting with no data at all. If we jumped right in and started to collect breach data, we’d waste countless resources, capture data that we’d discover later to be meaningless and find ourselves wishing we had collected data we didn’t. Therefore, we’re going to approach this like a research project and set a frame for this effort and define a handful of questions we’d like to explore. From that, we will be able to determine what data points we want to collect.

One last point before we get into preparing for this. As we are thinking of these questions we want to follow a few simple guidelines:

Aim for Objective

Constrict possible answers

Avoid conflation

Yes/no/other/unknown

leverage other sources at every chance

Goals and Research Questions

The overall goal we set will shape the types of research questions we ask and consequently how we’ll approach data management. For this chapter, we’re going to focus on a common use case for collecting breach data: prioritizing remediation efforts. Given a long list of things to fix (which every organization invariably has), what are the things we should focus on? Perhaps more importantly, because we have a finite amount of resources, are there some things we can deprioritize? This then becomes our first question and we’ll want to march towards addressing this goal.

type="definition"

The goal of this research is to support prioritization of remediation projects.

What projects should be prioritized?

What projects should be de-prioritized?

Notice how we worded our goal with the word “support” in there. This research will exist to support a decision process. It is not intended to be or replace the decision process. We need to have the wherewithal to recognize that remediation prioritization is a complex issue and we are just beginning to scrape away at it. At this point, where we have very little data, we should not make the assumption that we’ll get it perfectly right out of the gate. Even though there may be influencing variables that we don’t collect, we have to support the decision process, and we do that by reducing the amount of uncertainty in that decision.

type="note"

Data is for Reducing Uncertainty

While it would be great to collect breach data to create a perfect and prescriptive list of priorities, it just won’t happen. The data will simply help us know a little more than we currently do, but it won’t be able to definitively show us the path forward. This raises the question for some whether or not it is worth it. Is it worth spending the time and resources to create information that doesn’t tell us what to do?

The answer is an emphatic yes.

Uncertainty exists in the gap between what we know and what we would need to know to make the best decision. While it’s tempting to toss out imperfect information because it contains uncertainty, the value of the information should be assessed by comparison. Not between the perfect information we’d want and the information we’ll get, but instead between the information we *currently have* and the information *we will have*. This is where we see the value of this type of data analysis. Data will help us reduce our uncertainty by reducing the gap between what we know and what we need to know. Plus we will be making progress and setting a foundation for further reduction. This is how science has evolved our knowledge: a series of small steps each reducing our uncertainty a little more. Therefore, our goal should be to use data to reduce our uncertainty, not to give up when the data is less than perfect.

With our goal set, and the target being the reduction of uncertainty, there are four large categories of data we want to focus on:

The threat **actor** is one or more people behind the incident.

The **action** is one or more techniques or tools employed by the threat actor during the course of the incident

The **asset** is something (one or more) that we value. If nothing we value was affected, than we have no security incident.

The **attribute** affected. The threat actor’s actions affected the confidentiality, integrity or availability of an asset.

If this data is aggregated across multiple **victims**, we’ll want to know some victim demographics as well

One very important aspect and one that is very difficult to gather is impact. For the purpose of this chapter we will not be including details around impact.

Creating research questions around the asset

Since our goal is to support the decision around setting priority for remediation efforts, we have to think through the types of questions we will be asking during the prioritization. As a starting point, every remediation effort is about addressing one or more **asset** in our environment. An asset is something we value. Examples include servers, laptops, printed documents and so on. We know we want to form one or more research questions around the asset:

Let’s start out with an obvious question:

**Are some types of assets more likely to be involved in a security incident than others?**

Notice how the question is worded to be a good research question. It clearly defines what we want to measure, how we want to gather the data, and what we want to test in this data once it’s gathered. We can then break it down like this:

What we want to measure: **type of asset**, categorical list

How to gather: **yes/no**, if asset was involved or not

Test: compare **proportions** of incidents where type of asset was involved

There is an assumption going into this question which is some types of assets are more targeted than others. If we thought that all assets have an equal chance of being involved, this may not be worth our time. However, there’s a good chance that there is a difference in the types of assets involved in security incidents and this is interesting enough to include in our list of research questions. While this question is interesting, there are many other questions we could ask around assets to help reduce our uncertainty. For example, the type of data it houses and who manages the asset, whether it’s an employee owned device, a third party partner or perhaps even a cloud-based solution we’re talking about. With those in mind, let’s generate a few more questions.

Are assets managed by others more likely to be involved in a security incident that assets managed by us?

What to measure: **management responsibility** for asset (categorical list)

How to measure: **yes/no** if asset was managed by other party

Test: compare **proportion** of incidents where assets were managed by other party

Are assets in the “cloud” more likely to be involved in a security incident than non-cloud assets?

What to measure: “cloud” technology contributed to the incident (categorical list)

How to measure: **yes/no** if cloud technology contributed to the incident

Test: compare proportion of incidents where assets involved cloud-specific weaknesses

We will hold off for now on the type of data, because that will have more meaning for us than it’s association with the asset. For now these are a good set of questions. This will help us get a picture for how this notion of an asset can play into incidents and some attributes that may or may not contribute to its involvement in a breach. We could easily develop dozens of questions around the asset (type of OS, version, software, and so on), but there is a very important lesson about this type of data collection. Each question we ask and each data point we want to capture comes with a rather high price tag.

type="note"

Consider the Cost per Datum

During a manual data collection effort, it is very tempting to dream up all sorts of questions we’d like answered. Creating such a list isn’t bad and it may even be good to lay out all the questions you’d want to answer. But choose the questions you are going to ask very carefully because every question adds exponential cost across the lifetime of the data. Even before the question gets answered we have to build a method to collect it, so every question must be built into the data collection. We may have data validation going into the method, and then data validation coming out. As a single incident is being entered, each question will require some thought and perhaps even some research before it can be answered, again adding time and effort. That data point may require processing and clean up, and will need to stored and managed. Anytime we want to parse the data (and you’ll want to parse this in many different ways), we may have to consider this field, or worse, consider the interactions of all the fields. Beyond that, there are dozens other subtle interactions that will increase the cost of each data point beyond what we can imagine as we form the research questions.

It’s helpful to pretend you are about to take a long journey to a wise sage who lives on top of a mountain. You will have a limited amount of time to ask questions before the sage says something mysterious and vanishes. What questions will have the greatest impact? You’ll want to identify a handful of questions you really want answered, maybe a handful of questions you’d like to have answered and then you’ll have a mountain of questions you wish you had time to ask, but you’ll just have to make do. The same is true with manual data collection. If the post-incident questionnaire asks too many questions or is too painful, people will lose interest quickly and the answers will end up being of poor quality. You must choose your questions wisely.

Creating research questions around the actors

Knowing that our goal is to support the prioritization of remediation efforts, understanding who is attacking and their motive will help us prioritize where to focus our resources. For example, if we knew that most attackers are financially motivated and are external (non-employees) to our organization, we could prioritize projects that protect can lead to direct financial benefit for the attacker. Even if we just collected asset information and attacker information, we can see how the two fields together are more valuable than both of the fields alone.

attackers are more likely to attempt a brute force attack against passwords, we may wish to prioritize improving systems with weak or no password requirements. On the other hand, if we discover that attackers are more likely to grab credentials with key loggers, we could deprioritize enforcing complex passwords and focus on the multi-factor authentication projects.

Creating research questions around the actions

Knowing that our goal is to support the prioritization of remediation efforts, understanding what attackers are doing will help us prioritize where to focus first. For example, if we know attackers are more likely to attempt a brute force attack against passwords, we may wish to prioritize improving systems with weak or no password requirements. On the other hand, if we discover that attackers are more likely to grab credentials with key loggers, we could deprioritize enforcing complex passwords and focus on the multi-factor authentication projects.

There are two large challenges with recording what happened and we’ll have to address both at some point. First, any given incident may be the result of more than one action. While a SQL injection attack is typically a single action that exfiltrates data immediately, a phishing attack is typically one of many actions the attacker takes before they exfiltrate data. Second, there is no end to the amount of detail we could collect about the actions. Even within SQL injection, did they use an automated tool? What was the language of the web application? Was it custom written or an off-the-shelf application? Was there any application firewall or filtering in place? At some point rather quickly, we will have gone too far with the level of detail we seek (see the call out on the “cost per datum”).

With the knowledge that there are a lot of possible actions, we will want to limit the number of questions we ask, since it may be duplicated across multiple different actions. This short list should be enough to get started.

Are some threat actions more prevalent than others?

We will need a list of actions to select from

Are the attack vectors (pathways) more prevalent than others?

Finding the answers to these questions should help with our overall goal. But we still have a problem here in that we could easily create a list of actions that are long. Most every method of collecting breach data will attempt to capture “what happened”. Let’s take a quick survey of what’s out there.

Privacy Rights Clearing house has the following “types of breaches”:

Unintended disclosure (DISC) - Sensitive information posted publicly on a website, mishandled or sent to the wrong party via email, fax or mail.

Hacking or malware (HACK) - Electronic entry by an outside party, malware and spyware.

Payment Card Fraud (CARD) - Fraud involving debit and credit cards that is not accomplished via hacking. For example, skimming devices at point-of-service terminals.

Insider ( INSD) - Someone with legitimate access intentionally breaches information - such as an employee or contractor.

Physical loss (PHYS) - Lost, discarded or stolen non-electronic records, such as paper documents

Portable device (PORT) - Lost, discarded or stolen laptop, PDA, smartphone, portable memory device, CD, hard drive, data tape, etc

Stationary device (STAT) - Lost, discarded or stolen stationary electronic device such as a computer or server not designed for mobility.

Unknown or other (UNKN)

mple, Privacy Rights Clearing House

we’ve gone to far, and if you reference back to the note “We will have to create a cut-off point somewhere

a lot of In other words, if we ask for a single label for the actions, we are going to lose out

we will find a good deal of information in understanding what attackers are doing so we know what to prioritize One common question across all methods of breach data collection is “what happened?” But how this question gets answered is anything but common. and this is where we have to keep an eye on complexity.

ANOTHER NOTE: Conflated terms

Below are notes, can ignore for now

Within that statement of the goal, we have included two questions. These are not research questions and in fact would be absolutely horrible research questions. They are ambiguous and next to impossible to measure directly but they are serving as a compass. Every research question we form, should serve to eat away at the question.

Given that this is the goal, This type of general goal will help us define We want to understand if there are areas of the organization that are more likely to be targeted Where should we be focusing our assessment efforts efforts and prioritization and risk assessments (in any and all incantations of the concept). In other words we want to focus on spotting the large trends, and some of the details (like recording specific IP address) will be out of scope. Let’s state that

For example, if we want to collect breach data to help with upcoming project selection and budget allocation, we’d want to collect different data and manage it differently than if we want to study breaches for specific indicators of compromise such as IP addresses

Defining an overall goal is going to

miss some data we’d really care about. Perhaps we want to collect data to feed into

and Unfortunately the media does not help much in that regard as they play the “name and shame” As something we need to As if the victim was dumb enough to get robbed, but now that we’re moving into a data-driven world we should be thinking of a security breach as incredibly rich source of data. Where else can we learn about who is attacking, their tactics and practices and targets?

Before we jump into collecting data we have to ask ourselves, what do we want to know? Because if we survey the landscape of breach data (and we will in the next section), we may find that the types of research questions we are able to answer with what is available indicate that the research questions either weren’t well thought out, or were actually that simple. Huh, let’s see.

landscape

Once our defenses fail (and they will), we

We should probably take a moment to mention that one of the authors of this book

and looking at some of the more intangible aspects of computer security.

One example that is close to home for

Humans are wonderful/horrible sources of data

We want to begin this chapter by covering methods of collecting data from the environment and then discuss the challenges we face collecting data on breaches. We will talk about several methods of collecting breach data and then we will look at data analysis techniques on public breach data recording using the Vocabulary for Event Recording and Incident Sharing (VERIS) framework.

Breach data represents an extremely valuable source of information.

We want to focus this whole chapter on VERIS, Identify Verizon and the DBIR. We want to establish this framework below and then walk through what we do with breach data collection. Let’s not focus on the Data in the DBIR, but how we are able to pull information from reports.

When will want to

Four phases of a research project:

Define the goals and research questions

Establish data model and collection method

Collect and manage the data

Data analysis and presentation

*“VERIS is complex and they didn’t want to deal with it.”*

This

**Defining the Research Goals**

Whether we call it risk analysis or predictive analytics, a metrics program or a longitudinal study, or perhaps just simply a research project the effort should always begin with the same question: What are we trying to accomplish? Answering that question, with all its complex and nuanced details, will align expectations and help shape all of our following actions. We cannot stress the importance of this step enough. Far too often we have come across data that was gathered and prepared at great expense for some, yet failed to address even the most basic questions of the organization. In spite of the fancy graphics and glossy reports, the entire effort was doomed before it started because there was not agreement on where it was headed. One standard approach to getting agreement is establishing one or more research questions that the effort will try to answer.

Wade: What questions did you want to answer in the beginning?

When we prepare to collect data about information security events, we have an added complexity. Not only do we have to define what questions we want to answer, but undoubtedly, we will also have to define the questions we won’t answer. Not because we would be incapable from a data perspective, but every data point we collect comes at a price and that price-per-datum is much higher than people imagine. The primary limiting factor in collecting breach data is not the availability of data, but the amount of effort it takes to collect that data. If we attempted to answer every question we can think of, we would be crippled in the data collection step and the whole effort would ultimately fail. Therefore we have to state the first rule/guideline/mantra/principal:

Every data point has a high collection price. -- Every data point costs

One data point to collect is worth a dozen

Think of each data point as adding head count to the budget. The initial cost may be relatively low, but there are hidden costs. We want to limit

Within VERIS, we set our goal to support the decision-making process at a strategic level. This has several benefits and drawbacks. First, it sets the focus at more of a strategic layer than a tactical layer and frees us from the collection of specific indicators of compromise. The benefit of that is that we have less data to collect and the collection can be a bit easier. The drawback is that we lose the ability to inform detailed analysis of log data. With VERIS data alone, we cannot help determine if a specific transaction, host or executable is malicious or not. However, we can help determine if we should focus

So that brings us to

We will get into why this is true. Every question we put on a collection form, or measurement we want to read, comes at a price. While it’s easy to think, “it’s just one data point”, think about it throughout the life cycle of the data. Consider how many times that one data point will be touched in the life of the data. We may have one or more validation and/or verification processes, one or more transformation processes and certainly a handful of analysis and visualization efforts. While it is just one question, that data point may require a minute here and a few minutes there at multiple stages in its lifetime. Now multiply that by hundreds perhaps even thousands of records and it’s easy to see that the price per data point is a critical consideration. The effect of this compounding cost is that every data point must be scrutinized and tied back to one or more research question.

On top of that, it’s never just one more data point. If we become lax, the one more data point turns into dozens. When we consider changing VERIS, we will debate every single data point we include.

The VERIS community database (VCDB) has over 1,500 records in it at the time of this writing. If we spend an extra minute collecting 5 “interesting” data points,

For every data point we should add some type of error checking and/or validation. If we upgrade or migrate the data, we’ll hav Every data point isn’t just collected and we’re done. That one data point becomes part of the defining context in which is was gathered. If we have validation and/or verification, that data point must be considered, if we transform or move the data (and you most likely will do this multiple times), we have to touch that data point. and now must be covered by the data validation process. If relationships are established we

We have the collection, validation and verification (depending on importance)

anyone has, we

Security events offer a wealth of data and the main limitation re is no end to the amount of detail that we can gather.

**Establishing the Data Model**

The term “data model” can mean a variety of things depending on context and perspectives. In this case, we are using the term a bit ambiguously. At a bare minimum, we want to

leave the details up to the specifics of any individual impelementation. At a basic level, what we are describing is what fields we will look to collect and how those will be stored.

up to implementation details.

**Collecting and Managing Data**

Collecting data from the environment typically has a higher cost than collecting data from a log file. If the benefit is great enough, we may even end up recording one data point at a time so we want to be sure we are approaching this type of data collection with a clear intent. This is where our research question (discussed in Chapter 1) will be critical. We want to begin with one or more questions that we want to answer. By jumping right into data collection with one or more research questions, the data collection ends up favoring the convenient and obvious. Take for example a call center, if we leap into data collection, we will capture what’s obvious and available: how many calls, the operator, duration, etc. But what if we want to understand how effective the call center is from the customer’s perspective? That won’t be found in a log file and all the effort in collection and preparation is wasted. We offer the following steps to follow when creating a program of data collection:

Create purpose and direction through a list of questions

Brainstorm a list of questions that drive action or affect decisions moving forward

State the questions in a way data can answer

Prioritize and select one or more critical questions

if we answered this, would we change? If it was affirmative or negative, anything different? (run through possible answers, possibly set a threshold)

Select data collection method

Identify sources and methods of collecting data that address questions

Establish data model and vocabulary (fields and data types - taxonomy/framework)

is this field a singular value? multiple attributes?

Every data point and field has a cost - go back to questions, do we care the specific value or just that it’s above or below threshold.

Avoid pack-rat mentality, while it’d be awesome to have all the data, attempting to gather it is a recipe for disaster

Collect and manage

Varying data collection methods

data storage - dictated by 2b and 2c

Data Analysis

categorical data analysis

atomic data: Boolean, integer, (float), string

collections: list or dict (assoc, object)

We touched on

Once we

Integrating into a process (ticket system)

document/log review (after action reports, forms)

to answer a question, we have secondary review

survey

census (every member in population)

Focus group

Interviews

Observation

Survey

Follow-up interview

(after action report)

Issues:

Validity and consistency Issues

Cost to collect (resources like time, money, complexity)

Inference / Benefit of collection

“ expected quality of the collected data, estimated costs, predicted nonresponse rates, expected level of measure errors, and length of the data collection period “

T

Type of research: confirmatory, exploratory

Research question is the reason we are doing this.

When answered, we’ll know something we didn’t before.