**Overview**

Digital privacy in general, and securing personal data against the threat of breaches in particular, are important issues that have only become more salient in recent years. More and larger companies and organizations have gained greater access to personal, potentially sensitive user data, at the same time that malicious actors, corporate negligence, a combination, and/or other factors increase the risk that those data become publicly available or fall into the hands of hackers or other third parties. It is imperative that policymakers and regulators, users who provide their personal data to companies and organizations for safekeeping, academics and other researchers, corporate cybersecurity employees, and potentially others have accurate information about the prevalence, scope, and general nature of data breaches that occur across industries. This proposed data repository would curate datasets that describe large scale data breaches across multiple states.

**Potential Users**

*Note that all of these user groups technically fall into the “Users entrusting their data to companies and organizations” group, as everyone can be at risk of a data breach.*

* **Policymakers and regulators.** These are government employees at the state or federal level who want to understand the prevalence of data breaches and risk to their constituents of such breaches. These users want to craft effective policies and regulations to ensure companies and organizations entrusted with personal data take steps to protect their users from data breaches, and to enforce those regulations effectively.
* **Journalists.** Journalistsworkingfor local or national publications, or even freelance journalists or citizen journalists, are interested in stories of corporate malfeasance or risks to everyday people. They are interested in stories about data breaches so that they can inform their readers of the risks of entrusting data to third parties, and potentially to uncover stories of data holding entities acting improperly and not taking data security seriously enough to protect their users.
* **Users entrusting their data to companies or organizations.** Many kinds of companies and organizations can suffer data breaches, from financial companies to tech platforms to retailers to nonprofit advocacy organizations and beyond. Customers and users of these companies and organizations want to ensure that they are entrusting their personal information (e.g., name, age address, credit card information, etc.) to trustworthy organizations, and want to assess their risk of a data breach accurately.
  + **Parents of children using data-enabled technologies and services.** A subset of broader users affected by data breaches, parents may be particularly concerned that their children’s personal information and data are safe from breaches. They may be even more risk-averse than with their own data, potentially preventing their children from using services, platforms, etc. that have suffered data breaches in the past.
* **Cybersecurity professionals.** Many people work at companies, organizations, government agencies, and beyond to build digital infrastructure that protects against data breaches. These professionals want to understand the broader trends in the industry, what types of industry sectors are at greater/lesser risk of data breaches, what specific companies, organizations, etc. have been targeted more/less frequently, what type of data is more/less at risk, etc. in order to design more effective and resilient protective systems.
* **Insurance companies.** Companies that provide insurance to other companies, organizations, government agencies, etc. that covers damages due to data breaches need accurate, up-to-date data on the prevalence and scope of data breaches across industries, as well as other information to help them effectively assess risk and price that risk into their policies.
* **Academics and other researchers.** Researchers interested in the effect of large-scale digitization and the prevalence of sensitive user data in online repositories operated by companies, organizations, governments, etc. want to understand the nature of large data breaches in order to understand trends.

**Choose an existing openly available dataset within your topic domain and provide an overview.**

According to Washington State’s open data portal and attorney general’s office, “Washington law requires entities impacted by a data breach to notify the Attorney General’s Office (AGO) when more than 500 Washingtonians personal information was compromised as a result of the breach.” One example from the open data portal  [is this dataset](https://data.wa.gov/Consumer-Protection/Data-Breach-Notifications-Affecting-Washington-Res/sb4j-ca4h/about_data), which includes 25 columns and 1,103 rows of data. Each column represents a variable describing an aspect of a data breach, such as when the breach started and ended, when notice was submitted to the AGO, the name of the affected entity, the type of cyberattack, how long users’ data was exposed during the breach, and other information. Each row is a different instance of a data breach affecting at least 500 Washingtonians. The dataset is hosted by the Washington State open data portal, can be viewed in the browser as a table, and can be downloaded in a variety of formats including CSV, RDF, and XML.

**Choose one potential user (or user group such as “policymaker”). Write their story. What is their use case? What do they want from the collection? What is their goal in accessing the data?**

Just to pick a less obvious example, I’ll offer a user story from an insurance provider whose goal is to design effective insurance policies for entities at risk of data breaches, and price those policies accurately.

“As an insurance company providing policies and coverage to businesses operating in Washington state, or who collect or otherwise possess Washington residents’ data, I want to understand how common large scale data breaches are, what types of attacks are most common, what types of entities (e.g., companies, nonprofits, certain sectors) are most frequently targeted, and what types of users and user data is affected so that I can design effective policies that provide enough coverage at a price that will sustain my business.”

**Assess the dataset’s ability to meet the user’s goals. Does this dataset provide them everything they need? How will they access the data? What will they do with the data? What features are missing for your user?**

This dataset offers a lot of useful features for this user, including the types of companies/organizations affected by data breaches, the types of cyberattacks involved, the duration and scope of the breaches, the time and frequency of breaches, and the number of users affected.

The dataset does not include data on the specific kind of information involved in the breach (e.g., names, social security numbers, etc.) or any negative impacts resulting from the breach (e.g., identity theft) which would albeit be difficult to track and include in such a dataset. The dataset is also limited to data breaches that include at least 500 Washingtonians, pursuant to state law. The dataset also includes a lot of potentially duplicative variables that describe the duration of the breach, when the breach was discovered, how long users’ data was exposed, etc., making the dataset a little difficult to use. Further, currently this user would have to navigate to the Washington State open data portal and locate this specific dataset from a number of similar datasets with different numbers of variables, some of which only offer visualizations without much raw data. Finally, the dataset only covers Washington, and does little to offer this user a comparative analysis or assessment of relative risks across states.