Soteria

— Team —

NAME	EMAIL	ROLES
Victor Starostenko	victor.starostenko@live.com	Feature Engineering, Coding, Development & Project Manager
Evie Phan	evphan@gmail.com	Feature Analysis & Engineering, UX, FrontEnd
Ashley DeSouza	ashley.souza@live.com	Feature Engineering
Shreyas	shreyas@ischool.berkeley.edu	Coding & Development

— Project Proposal —

Problem Statement

It is a known fact that the best way to learn something is by making mistakes. While it is likely not to repeat the same mistake again, it is a costly way to learn. Second best way is to learn from the mistakes of others, and the only way to do so is to know when other make mistakes.

Many companies have suffered extremely high costs associated with losing sensitive information due to security breaches, but what is even more troubling is that these companies kept their breaches under wraps. A few years ago this was the way things were done. Nobody wanted to be exposed for having weak security or fragile infrastructure, and so organizations endured the breach, paid for the consequences, and kept quiet about the details out of embarrassment. And then a few months later the same breach would happen to someone else.

By sharing as much information as possible about security breaches and what led to them, organizations will be able to more strategically and effectively fight the attackers. By uniting information, analyzing it, and drawing conclusions it would be much easier to find commonalities in offending technologies or methods.

PETER GEORGE, President and Chief Executive Officer of Fidelis, says:

> *"Companies should share security breach information because that is the only way we will be able to cobble together a comprehensive picture of the threats and fight back."*

Description of Objectives

- Analyze and explore data and extract interesting features.
- Create a visualization/report viewing interface to be able to sift through the records quickly.

— About the Data —

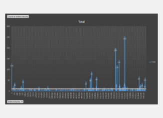
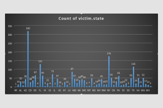
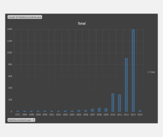
Open Data Set

VERIS Community Database
VERIS is a Vocabulary for Event Recording and Incident Sharing. VERIS is a set of metrics designed to provide a common language for describing security incidents in a structured and repeatable manner.
VERIS Community Database a dataset of over 3,000 security incidents and breaches.

Source

VERIS Community Database GitHub repository
License information

DataSet Description

DATA ATTRIBUTE	ATTRIBUTE DESCRIPTION	ATTRIBUTE INTUITION	HISTOGRAM
count_of_victim.industry	Count of Victims by Industry (coded)	Some industries might be more likely to get attacked	
victim.state	Location of Victim by US State	States that are more vulnerable to breaches	
Count of timeline.incident.year	Incidents per year	Exponential growth as more information is shared	

— Anticipated Problems —

- A lot of data is Boolean.
- Some missing values.
- Some data is not uniform.

— Ties to Wikipedia/Wikimedia —

None

— Project Challenges —

- Lots of attributes in the data. This could either be good or could be deceiving. We might have to parse through the noise of not just lots of data points but also features.
- Domain knowledge. Security is an intricate and delicate issue. So although we have the dataset, the inferences might still be dependent on the domain knowledge needed to make sense of the output.
- An uncommonly large number of boolean variables. So a large number of our features are presence/absence than a continuous range of values. I personally always get worried when I see lots of booleans.