BITSIGHT

Data Breaches | EDA

This dataset from Bitsight includes data breaches incidents happening in US starting at 2015 until the end of 2022 excluding outlier. I use the two JavaScript libraries arquero.js and plot.js for data processing and fast and simple visualisations.

The objective of this EDA is to clean the data and get familiar with the dataset, try different visualisation types and design as well as detecting patterns for building the story.

Data

Country	Industry	Sector	Severity	Month_all	Year	
United States of America	Security and Investigations	Business Services	3	Jan	2015	201
United States of America	Leisure. Travel & Tourism	Tourism/Hospitality	3	Jan	2015	201
United States of America	Entertainment	Media/Entertainment	3	Feb	2015	201
United States of America	Hospital & Health Care	Healthcare/Wellness	3	Feb	2015	201
United States of America	Government Administration	Government/Politics	3	Mar	2015	201

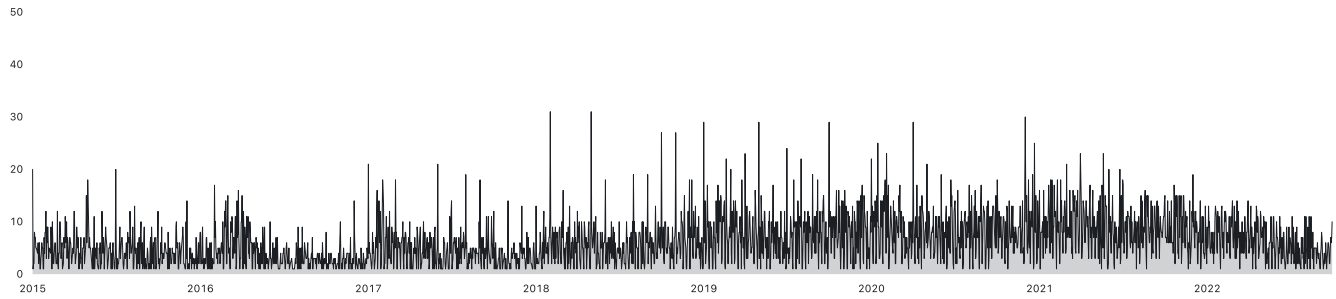
Severity

The majority of incidents is as expected less severe. An interesting group to check across sectors might be the outlier group of high severe cases (category 3)

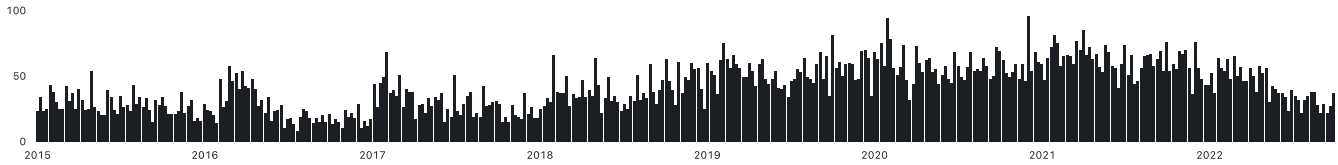
Over time

Looking at all the data over time, we can observe an upwards trend starting in 2018 and an downward trend in 2021. Why would the amount go down, seems odd. Seasonality in terms of higher values in first three months always increasing can be observed.

Q1: Is this trend real or is it based on reporting changes in Bitsight?

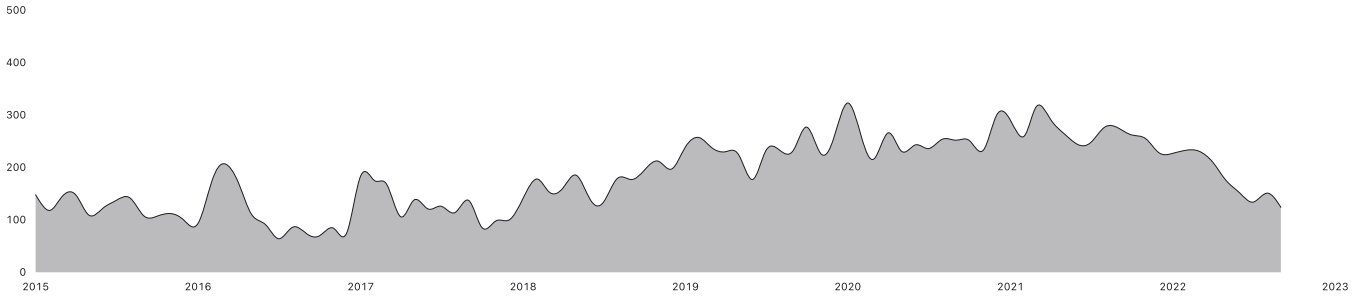


If aggregated by months we can see more clearly the pattern described above as noises in the data is deleted.

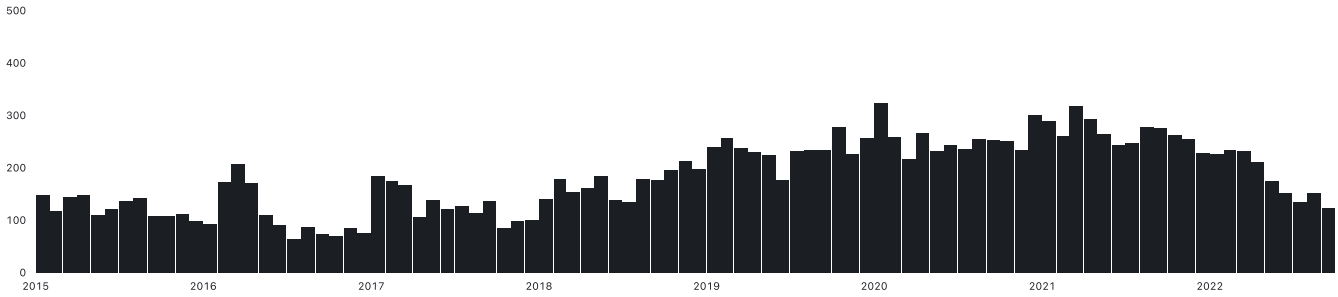


Density chart

Using a density chart allows to see the overall trend flow confirming patterns described above. Dataset has been aggregated by months too.



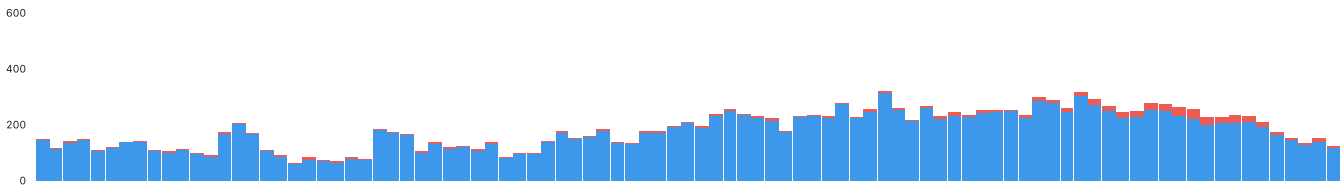
Using bars to show more detail again on the monthly level.



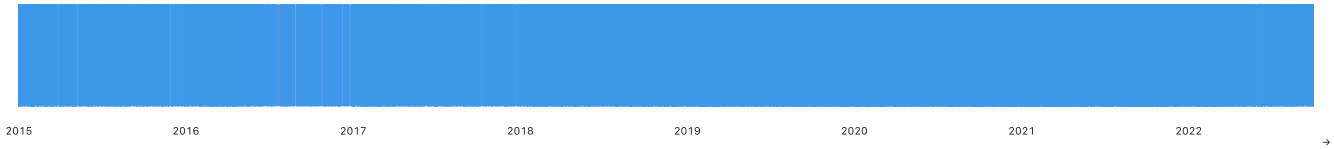
Severity over time

Highly sever attacks happened frequently in 2021 but starting to decrease in last months as the overall incident amount decreases. Data might not be complete as often data breaches are reported much after they actually happened. First thoughts about cutting off the dataset to improve reliability

Q2: Why do high severe cases appear more recently?



Barcode chart

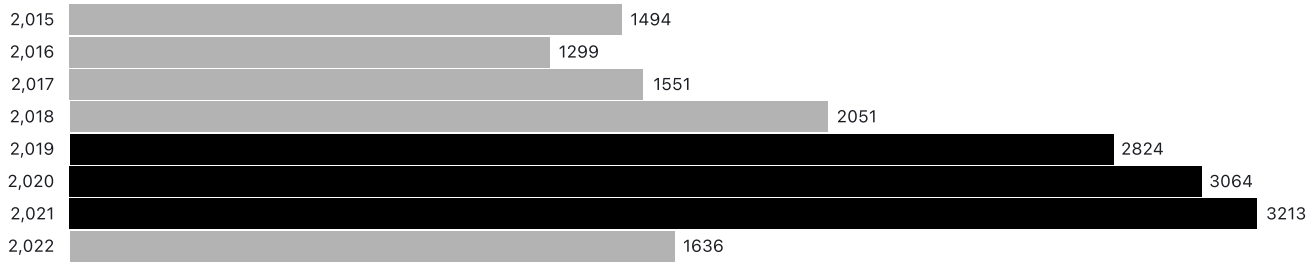


Category: Year

Now we are checking by year the totals and can see that 2020 and 2021 are much higher

How we are checking by year the totals and can see that 2020 and 2021 are much higher than 2022.

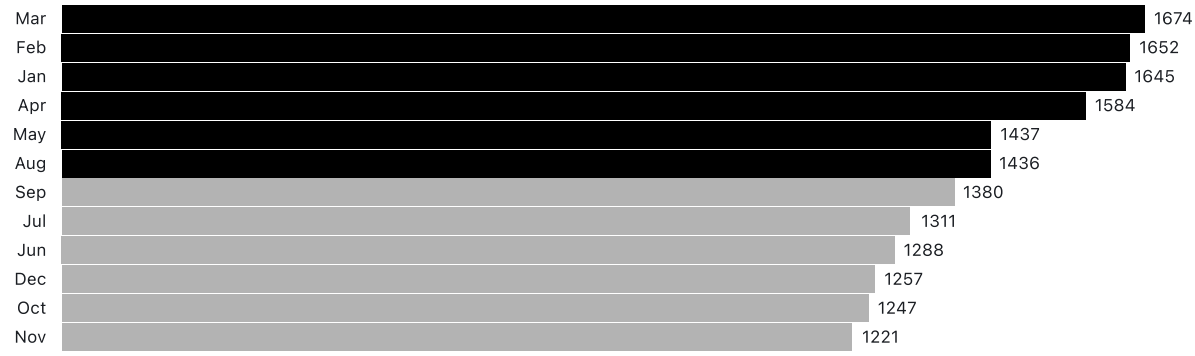
Q3: Why?



Category : Months (Seasonality)

We might say that towards the beginning of the year (jan, geb, march) slightly more incidents happen which we already saw looking at the evolution charts above.

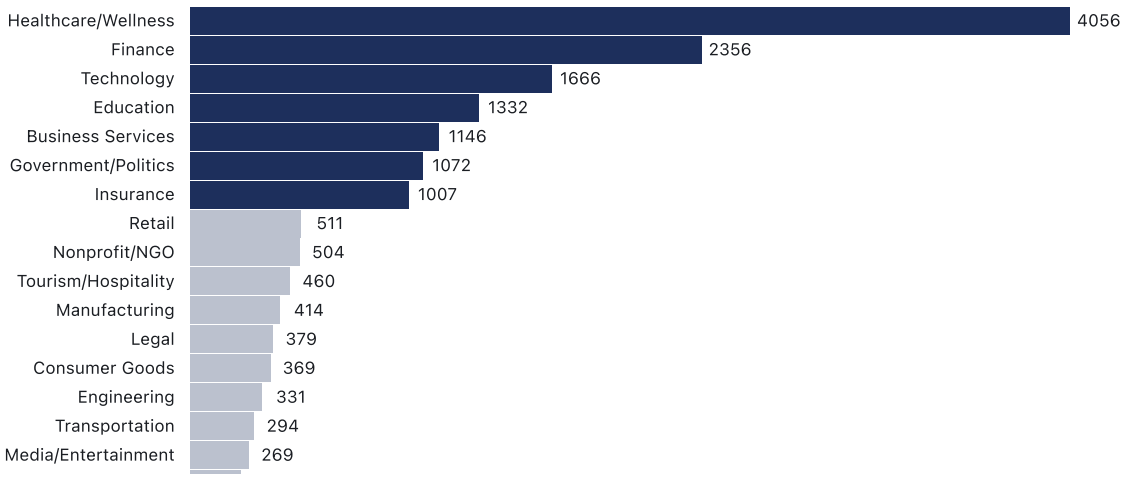
Q4: What is the reason for this, or just sample noise?

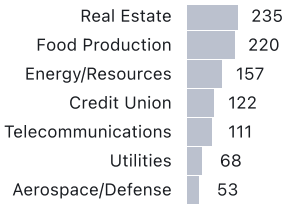


▸ `Array(12) ["Mar", "Feb", "Jan", "Apr", "May", "Aug", "Sep", "Jul", "Jun", "Dec",`

Category: Sector

There are 24 sectors but the chart below exclude missing values as well as the sector group "TBD". Using all data *Healthcare* has by far the highest data breach cases, followed by *Finance*. Up to the category *Insurance* categories are far higher than the rest. I would interpret these as the 6 most attack sensitive categories.



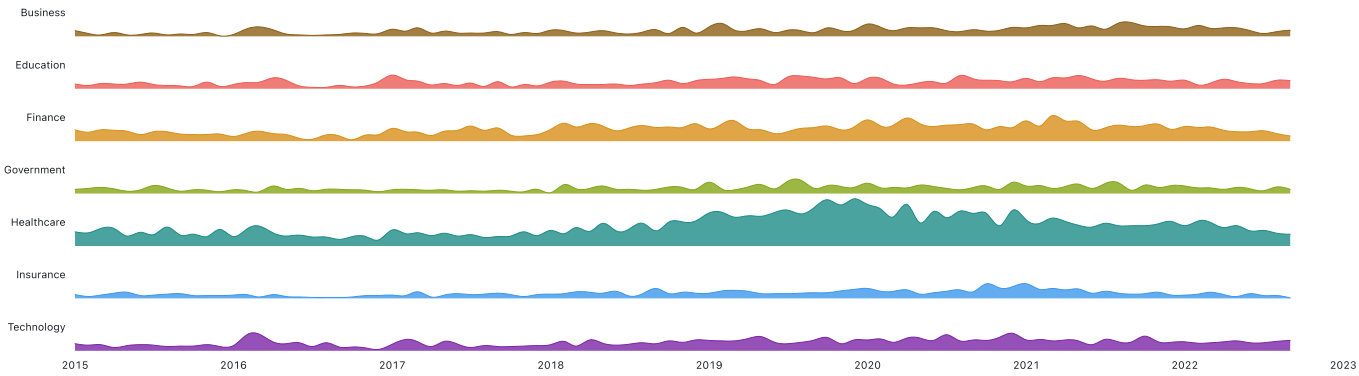


Sector over time

Moreover can we now detect a clear increase in attacks in the healthcare sector starting in 2018 reaching its highest in the end of 2019.

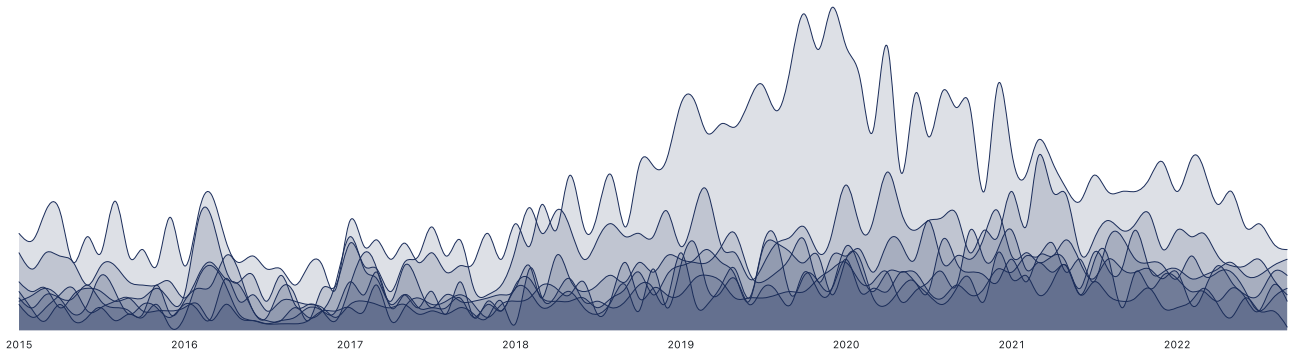
Small multiple

Across all sectors we can see the seasonality again even though in different dimensions. Finance has always had more attacks



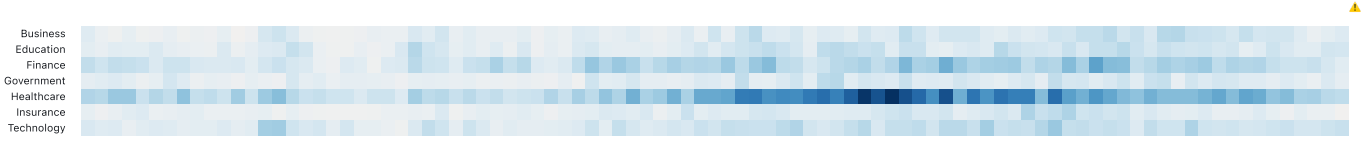
Multiple area chart

Trying a different chart type using an multiple area chart to identify the overlap. In the chart below we can see that indeed *Healthcare* is by far higher than the others.



Calendar heatmap

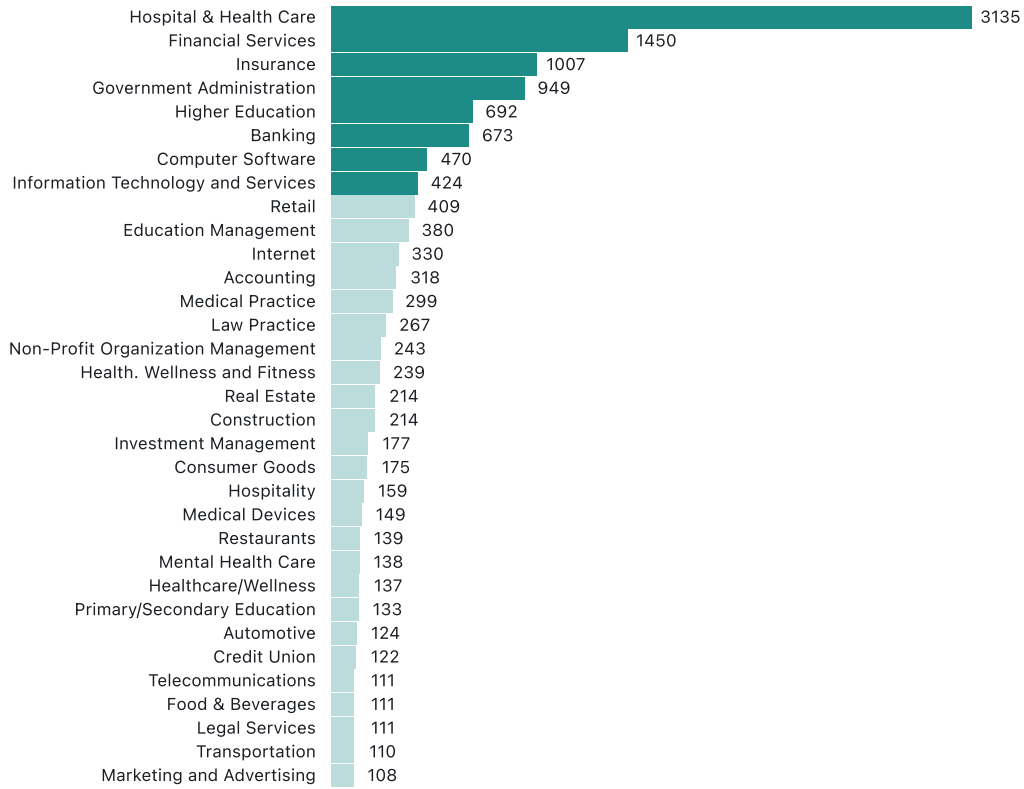
Calendar heatmap allows to even reduce the data-ink and show all patterns on smaller space.



Categories: Industry

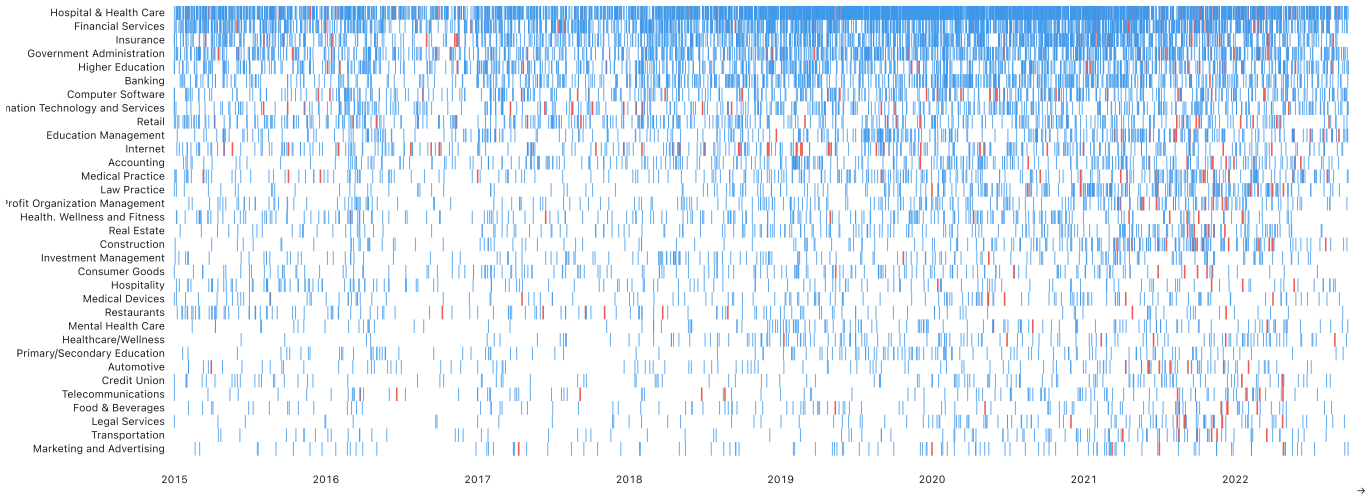
Categories: Industry

When looking at frequency by industry nothing seem to appear in terms of interpretation. This part of the story is already covered in the sector part.



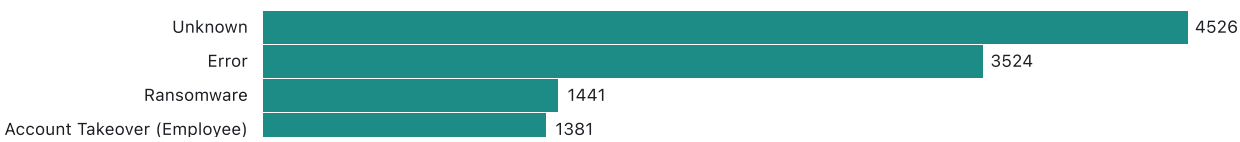
Barcode chart (small multiple)

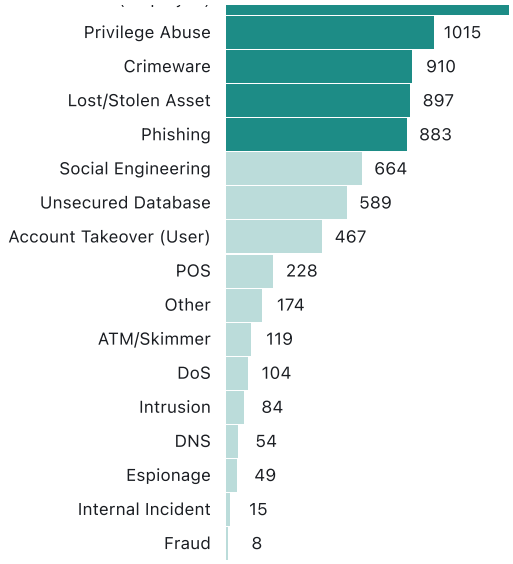
Checking if any industry stands out in terms of severity over time. We can see a similar patter as show above on the more aggregated sector categories.



Category: Type

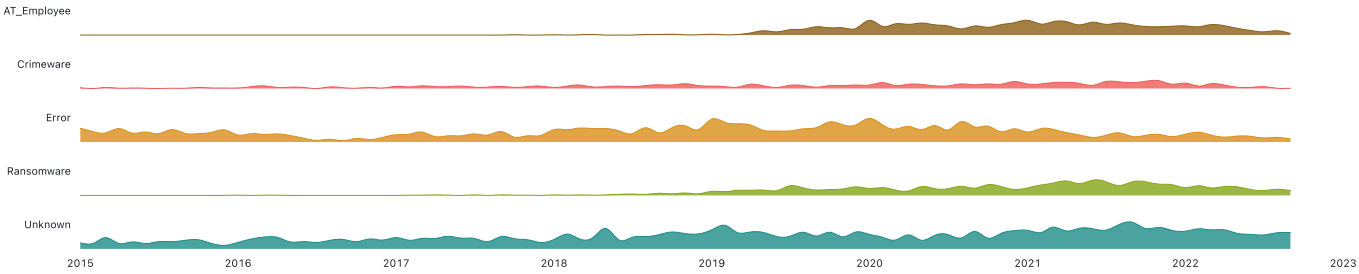
Overall by far the most frequent cases are of the type *Unknown* and *Error*, which can be classified as less specific failures and therefore harder to prevent for.



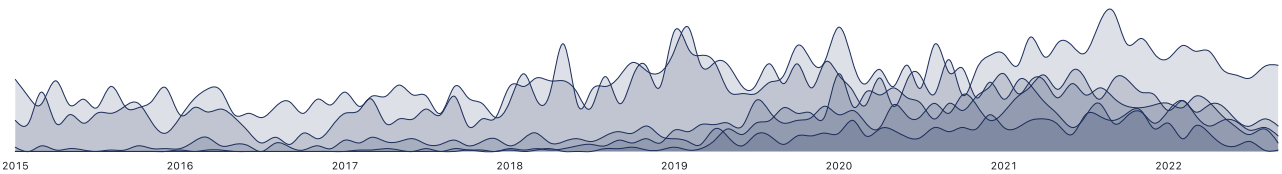


Over time (Small multiple)

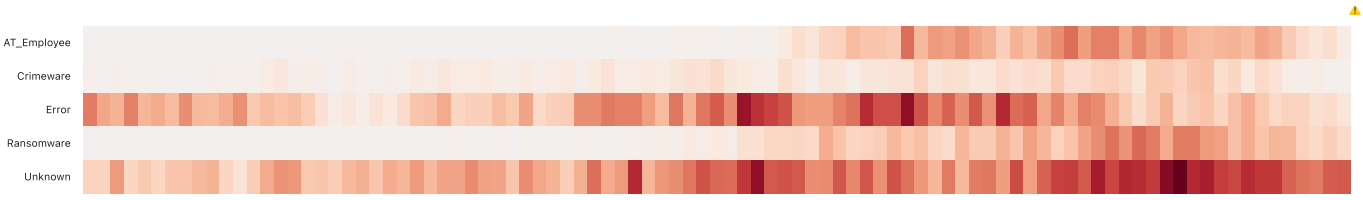
What does stand out is an increase on the more recent years in Ransomware and Account takeover by an employee. Could be even connected.



Multiple area chart

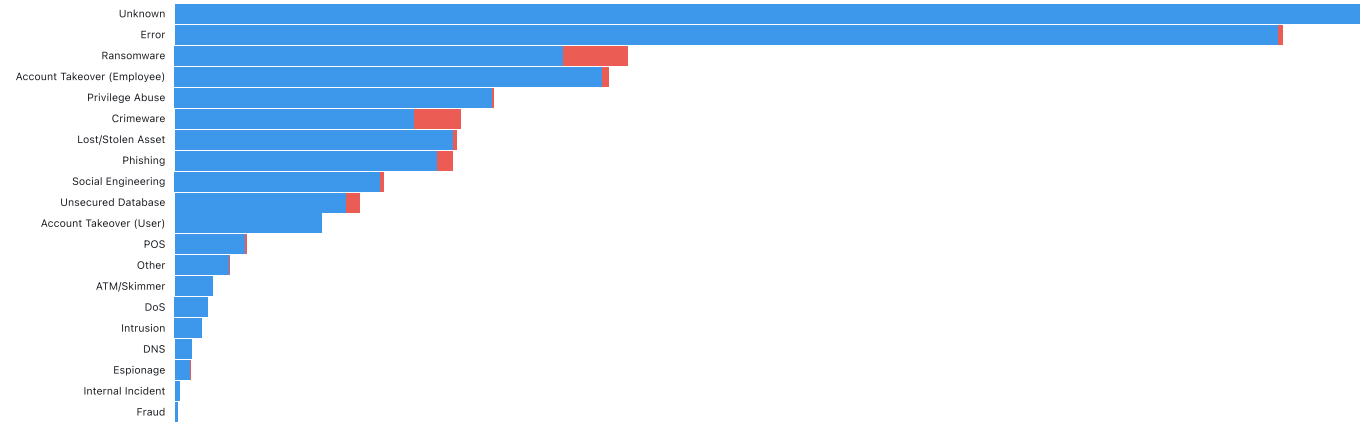


Calendar heatmap



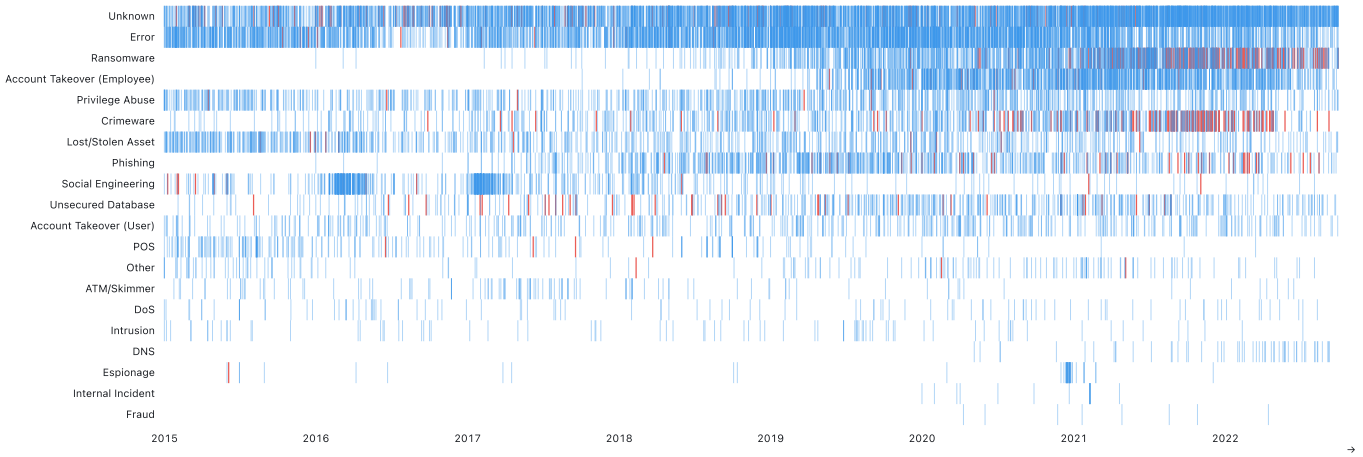
Type by severity

Overall *Unknown* and *Error* are happening across all years more frequently, but most of them on a less sever level. *Ransomware* and *Crimeware* include more sever cases.



Barcode chart (small multiple)

Below again we put the focus on the highly severe cases. First of all the *Unknown* are very rare sever cases (all blue) and especially recently. But we can see that they do appear more frequently recently. The category *Error* has a similar pattern despite the fact that the more frequent cases were between 2018 and 2021. In the category *Social Engineering* we can detect two hotspots in the beginning of 2016 and 2017. *Crimeware* has recently more sever cases same with *Ransomeware*.



Appendix

```
import {aq, op} from "@uwdata/arquero"
```

Definitions

Date The date when the news article was published and when the event will begin to impact the rating.

Severity The severity of the event based on the number of lost or exposed data records and the impact of a particular event.

Breach Security Incidents

Breach Security Incidents involves serious events that usually result in a successful cyberattack and/or data compromise by unauthorized individuals. Breach Security Incidents are ratings-impacting.

- **Crimeware** An instance of malware installed for the purpose of acquiring unauthorized data or assets.
- **Espionage** An incident of unauthorized network or system access exhibiting the

- **Espionage** An incident of unauthorized network or system access exhibiting the motive of state-sponsored or industrial espionage, where trade secrets or IP are frequently targeted.
- **Intrusion** Unauthorized access which does not involve exfiltration of records or other resources.
- **Phishing** An attack in which fraudulent email is used to mimic the access of an authorized employee or legitimate contact.
- **Ransomware** An attack designed to block access to a computer system until a sum of money is paid.
- **Social Engineering** An attack which uses deception to trick individuals into divulging unauthorized information or access.
- **Web Apps** An incident in which a web application was the attack vector, including code level vulnerabilities in the application and thwarted authentication mechanisms.

General Security Incidents

General Security Incidents involves other kinds of security events that may still affect security ratings, such as employee error or misconduct. General Security Incidents are considered more severe than Other Disclosures. Some categories of General Security Incidents are ratings-impacting, while others are informational only and do not impact the rating.

- **Account Takeover (Employee)** An attacker gains unauthorized access into a service through the use of employee's account credentials.
- **Account Takeover (User)** An attacker gains unauthorized access into a service through the use of a user's account credentials.
- **DNS Incident** An organization lost control or never had control of one of the associated assets, as defined by the DNS record. Examples of poor DNS security practices: Using a stale DNS record. Internally configuring publicly registrable domains (such as from an active directory), but not actually owning the domain.
- **Error** An incident involving unintentional actions that directly compromise a sensitive asset.
- Internal Incident An incident discovered by the company in question and remediated with no apparent compromise.
- **Lost/Stolen Asset** An incident where an information asset went missing, whether through misplacement or malice.
- **Lost/Stolen Asset (Encrypted)** An incident where an encrypted asset went missing, whether through misplacement or malice, with no evidence of encryption compromise.
- **Other Incident** A security incident that does not fall into one of the other categories.
- **Point of Sale (PoS)** Remote attacks against the environments where retail transactions are conducted, specifically where purchases are made.
- **Privilege Abuse** An unapproved or malicious use of organizational resources beyond what is authorized.
- **Unknown** A security incident where certain classification details pertaining to the event are unknown.
- **Unsecured Database** A database is left unsecured due to error and the data is accessible by third parties.

Data processing