

Introduction

The surge in digital data transmission has amplified the susceptibility of information systems to security breaches. Our project utilizes the dataset from Information is Beautiful¹ on the world's biggest data breaches to augment the existing visualization.

The dataset provides awareness of the scale and impact of data breaches across various sectors and organizations and highlights the importance of data security for businesses and individuals.

We plan to add features to the visualization, normalize the data and run comparative analyses, and develop models to predict future breaches. Our objective is to transform this visualization into a more dynamic and analytical tool that not only reflects previous data breaches but also provides insights into potential future trends.

The dataset contains information on the world's biggest data breaches which includes the date of the breach, the organization involved, the method of the breach, the number of records lost, the sector of the organization and more.

Previous Visualization

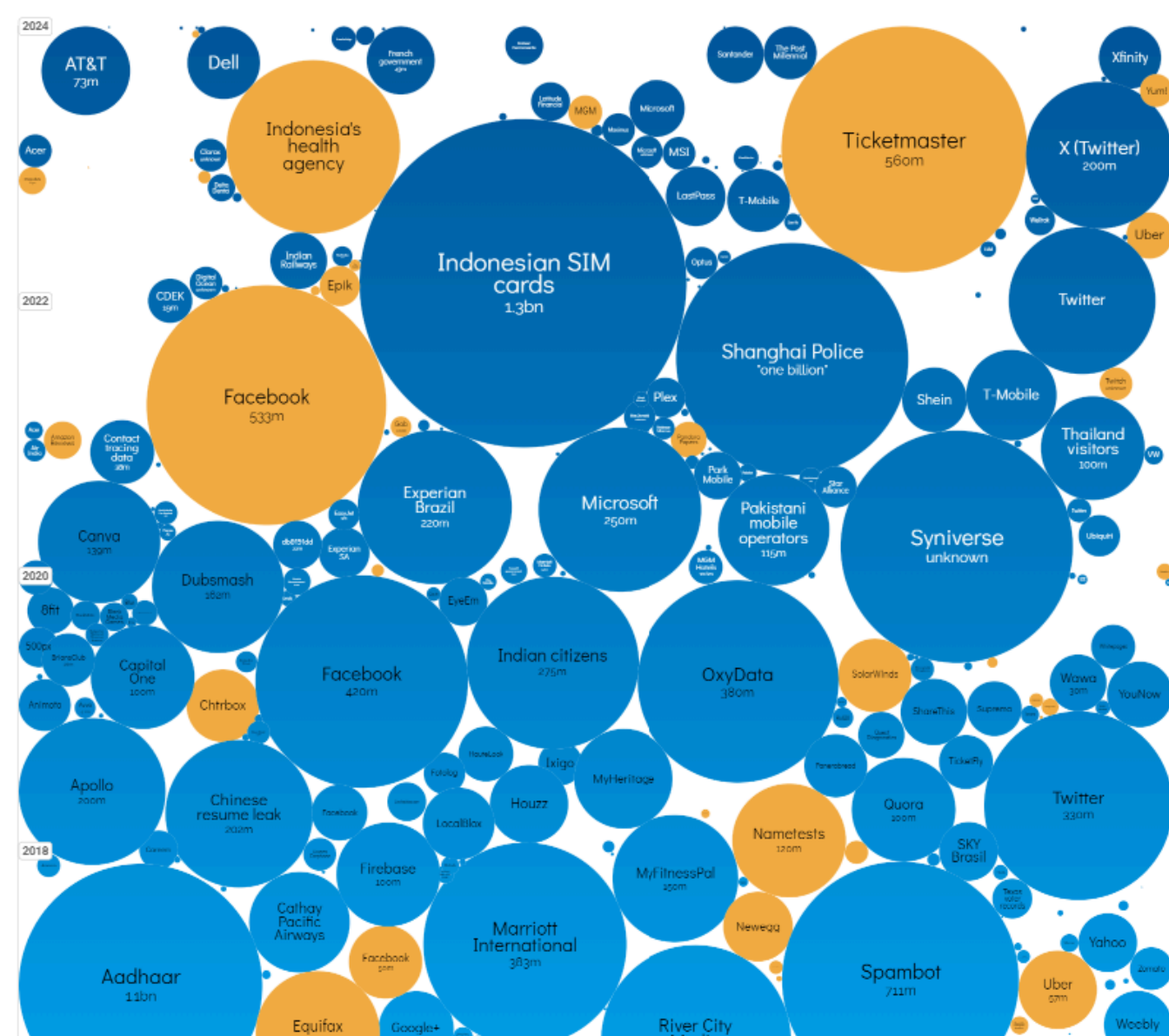


Figure 1: This bubble chart illustrates the scale and impact of major global data breaches, categorized by the organization and number of records lost, published by Information is Beautiful.

Strengths

- **Visual Impact:** The use of large, coloured bubbles immediately draws attention and effectively communicates the relative scale of data breaches. The varying bubble sizes enable quick visual comparisons between the magnitudes of different breaches, helping to identify which incidents had the largest impact.
- **Interactivity:** Interactive elements allow users to filter breaches by sector and type, facilitating a customized exploration of the data.
- **Information Density:** A significant amount of information is compactly displayed, providing quick insights into data breaches across different organisations and record lost.

¹T. E. David McCandless, "World's Biggest Data Breaches, information is beautiful". [Online]. Available: <https://informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/>

- **Chronological Clarity:** The layout organizes data breaches chronologically, making it easy to track the frequency and evolution of breaches over time

Suggested Improvements

1. **Enhanced Filtering:** Introduce advanced filtering options to enable users to segment breaches by industry, year, data sensitivity, or breach method, enhancing the customization and depth of analysis.
2. **Trend Analysis:** Introduce trend analysis features to identify patterns and correlations in data breaches over time, enabling users to gain insights into potential risk factors and emerging trends.
3. **Colour-Coding:** Utilize colour-coding to represent different breach types, data sensitivity or sectors, enhancing visual clarity and enabling users to quickly identify key information. Include a legend to explain colour codes and improve interpretability.
4. **Overlapping Resolution:** Implement a solution to address overlapping, such as Transparency (Alpha), jittering, grouping, or dynamic resizing, to ensure that all data points are clearly visible and distinguishable.
5. **Legend and Annotations:** Include a legend to explain the meaning of bubble sizes and colours, providing context for users unfamiliar with the visualization. Add annotations to highlight significant breaches or trends, guiding users' attention to key insights.
6. **Improved Scaling Factors:** Implement better scaling factors in logarithmic scale to accommodate a wider range of records, ensuring that small and large breaches are both visible and accurately represented.

Implementation

Data

- The columns inside the dataset are being renamed to improve readability and consistency and only necessary columns are being retained for visualisation.
- The columns 'year' and 'records_lost' are converted to numeric after removing non-numeric characters.
- The missing values in 'data_sensitivity' were replaced with the median while the rows with missing values in 'story' and 'source_link' were removed.
- The team also group data by 'method' and 'data_sensitivity' and arrange it by year to identify the data patterns easier.

Software

We used the Quarto publication framework and the R programming language, along with the following third-party packages:

- shiny to create the web application framework.
- plotly which used for creating interactive plots.
- tidyverse for data transformation, including ggplot2 for visualization based on the grammar of graphics, and dplyr for data manipulation.
- scales for scaling and formatting data in visualizations

Improved Visualization

Trends in Data Breaches & Hacks Over Time

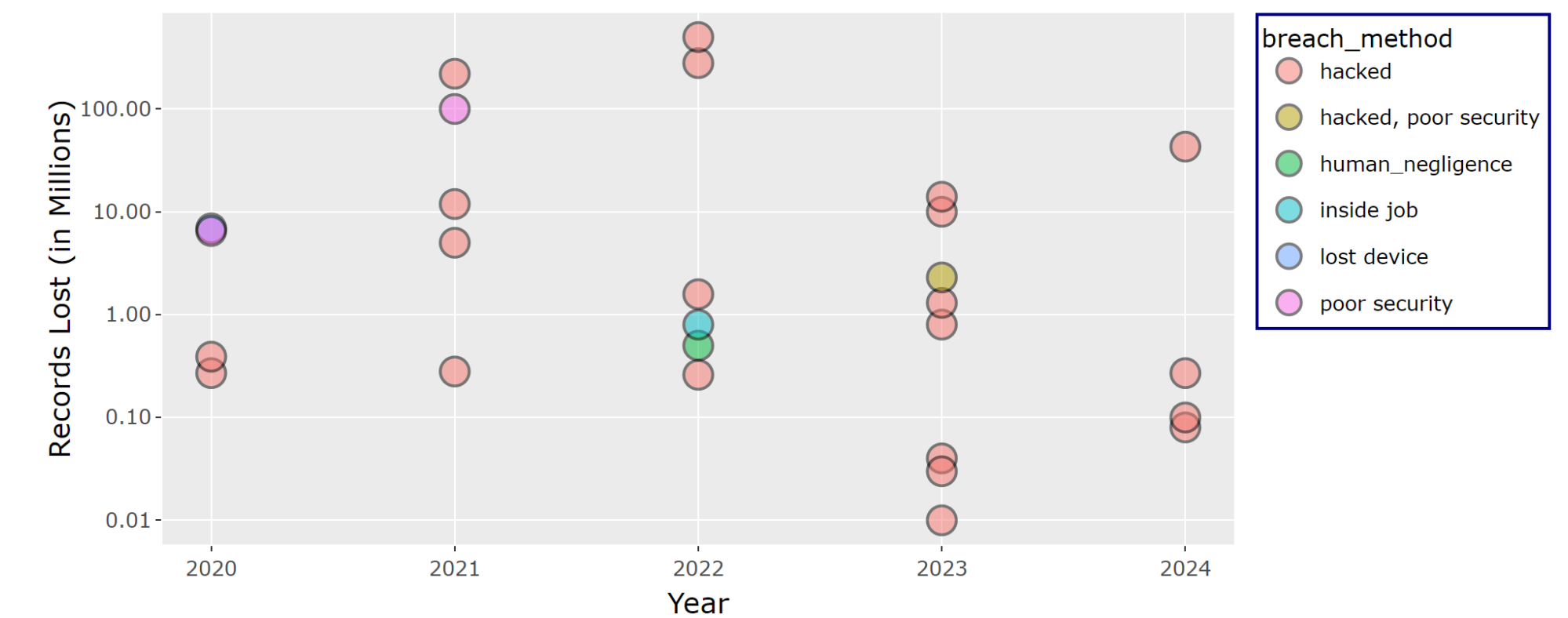


Figure 2: This bubble plot is an improved version of the original visualization.

The improved visualization significantly enhances clarity and detail. It uses a bubble chart format, with colors denoting breach methods, providing an intuitive understanding of data breach impacts and causes. The comprehensive legend and organized layout make the chart more aesthetically appealing and easier to interpret quickly and accurately compared to the original.

Visualization (Interactive)

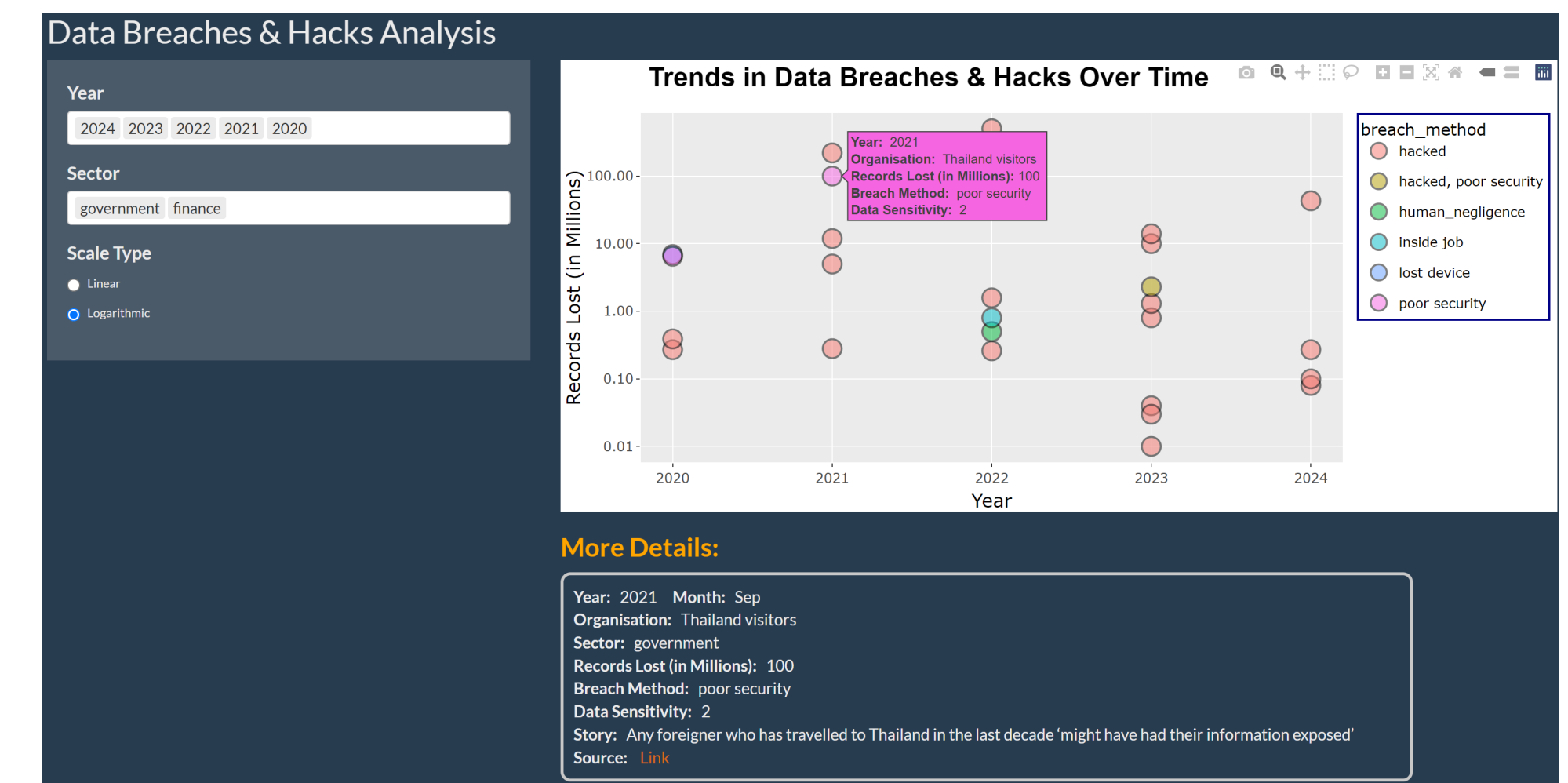


Figure 3: This is an interactive bubble plot of the improved visualization.

This interactive plot allows users to explore data by hovering over bubbles for details, filtering by year and sector, and toggling between linear and logarithmic scales for records lost. The plot is color-coded by breach method, with a legend for reference, and users can click on bubbles to view detailed information about each breach.

Conclusion

In conclusion, the improved visualization offers a more comprehensive and user-friendly representation of the world's biggest data breaches. With interactive features and advanced filtering options capabilities, it provides valuable insights into the scale, impact, and trends of data breaches over time. The enhanced clarity and interactivity make it a powerful tool for understanding and analysing data breaches, enabling users to identify patterns, correlations, and potential risk factors in data security.