Visualising Top 10 Sectors affected by Data Breaches (2010–2024)

Cham Xun Tong, Reagan Chia, Teo Shun Yao, Shermaine Peh, Pang Jing Jie, Enrique Carlos Marcelo, Garrick Low
2102436, 2102539, 2101104, 2101573, 2100932, 2102740, 2100590
Computer Science

SINGAPORE INSTITUTE OF TECHNOLOGY

Introduction

Data breach is a common sight with the increase in technological surge, resulting in more companies bringing their physical work into the online space. One might think that these data breaches are irrelevant to them, but they will be surprised at how they might be the sole reason for such incident occurrences. Over the years, the number of data breaches have been increasing, from about 25 breaches in 2012 to currently in year 2024, 17 breaches have already occurred, and it is barely half the year passed!
This visualization has been dubbed "World's Biggest Data Breaches & Hacks", but it can be improved due to its current clustering of companies and small wordings on less important data breaches that had happened. The purpose of this poster is to highlight the few most affected sectors, how the number of records stolen has increased over the years to raise awareness that every industry can be implicated by data breaches.

Previous Visualization

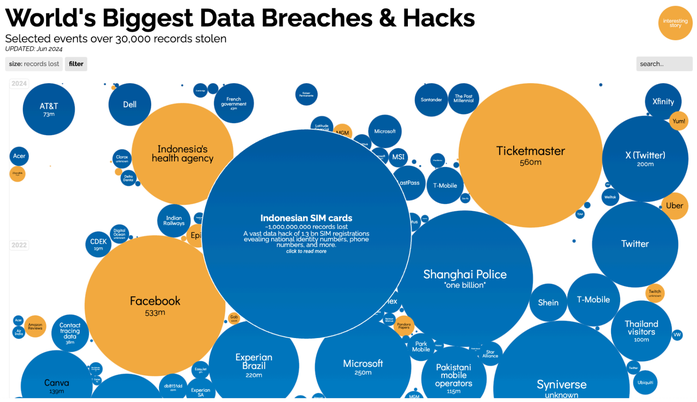

Figure 1: World's Biggest Data Breaches & Hacks, published by Information is Beautiful

Strengths of Original Chart

- Clear Representation of Magnitude: The bubble sizes in the chart effectively convey the scale of each data breach, making it easy to assess the impact without looking at numerical details.
- Categorical Differentiation: Color coding allows for quick identification of different categories or years, making it simple to track temporal trends or categorize breaches by industry or type.
- Informative Labels: Each bubble is labeled with the affected entity's name and the number of records compromised, providing immediate context without the need to cross-reference with external documents or data sources.
- Trend Identification: The vertical placement of the bubbles suggests a timeline, showing trends over the years and the evolution of data breaches in frequency and scale.
- Visual Impact: The bubble chart format has a strong visual appeal that captures attention, encouraging further exploration and retaining viewer attention.

Suggested Improvements

The original chart is cluttered and lacks a clear objective, making it difficult for viewers to quickly understand the data . Some suggested enhancements are:

- Simplification: Focuses on the top 10 sectors affected by data breaches, reducing clutter and highlighting significant trends
- Categorization: Groups data by sector for easier comparison and understanding of which industries are most vulnerable
- Change the plot to a Gantt Chart: To visualize the timeline and impact of data breaches clearly. It allows audiences to clearly focus on what message the data in the charts are sending as well.
- Insights Focused: To reduce cognitive load, the visualization should emphasize key insights, such as the top sectors affected by data breaches and the total records lost over time.
- Visual Consistency: The use of consistent color coding for each insights aids in quick recognition and reduces cognitive load. Viewers can easily follow the color scheme to track data breaches across the timeline without getting confused by random color variations

Implementation

Data

The data was obtained from the "World's Biggest Data Breaches & Hacks" Google Spreadsheet. Key columns include organizations, records lost, year, sector, method, and interesting story. To maintain relevance and reduce complexity, sectors with similar characteristics were merged into broader categories such as Web, Financial, Retail, and others. The emphasis was then refined to the top 10 sectors based on total records lost between 2010 and 2024. Missing values in significant columns were addressed to guarantee data integrity.

Software

The implementation utilized the Quarto publication framework, R programming language, and the following third-party packages:

- readr for data import
- tidyverse for data transformation, including dplyr for data cleaning, grouping, and summarization, and ggplot2 for powerful visualization customization options based on graphic grammar
- scales for adjusting the axis scales
- ggrepel for improved text labelling
- knitr for dynamic document generation
- gridExtra for arranging multiple plots

Improved Visualization

- Insight 1: The chart focuses on the top 10 sectors affected by data breaches between 2010 and 2024, highlighting the most vulnerable industries.
- Insight 2: The chart shows the an interesting trend of data breaches over time
- Insight 3: The chart uses color coding to differentiate between high vulnerability sectors and emerging threats, providing additional context to the data.

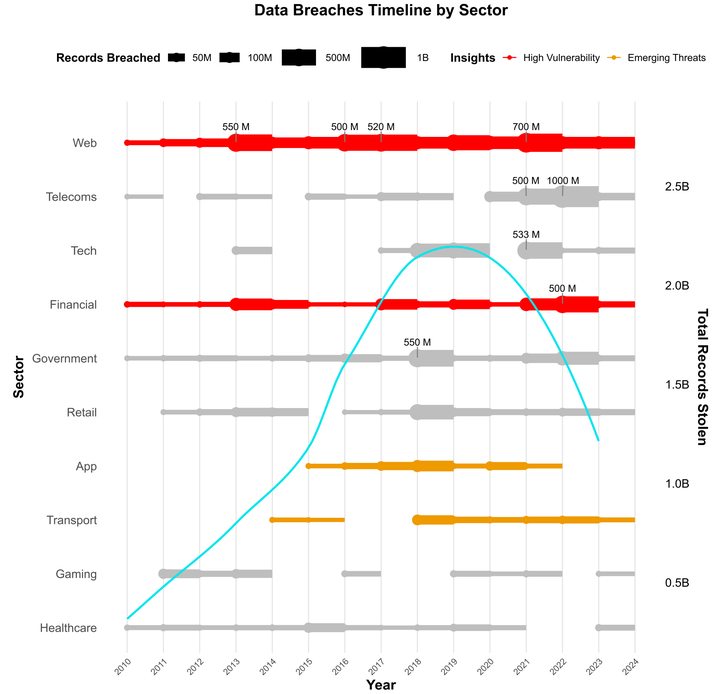

Figure 2: Data Breaches Timeline by Sector

Further Sugestions for Interactivity

- Interactive Filtering: Allow users to filter data breaches by sector, year, or records lost to focus on specific subsets of the data.
- Tooltip Information: Display additional details about each data breach, such as the affected organization, records lost, and breach method, when hovering over the bubbles.
- Dynamic Sorting: Enable users to sort the data breaches by sector, records lost, or year to explore different trends and patterns.

Conclusion

The improved visualization effectively highlights the top 10 sectors affected by data breaches between 2010 and 2024, providing a clear and concise overview of the data. By categorizing the sectors and visualizing the timeline of data breaches, the chart offers valuable insights into the industries most vulnerable to cyber threats. The use of color coding, size differentiation, and text labels enhances the readability and interpretability of the chart, making it easier for viewers to grasp the key information at a glance. The proposed interactivity features would further enhance the user experience and enable more in-depth exploration of the data.