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

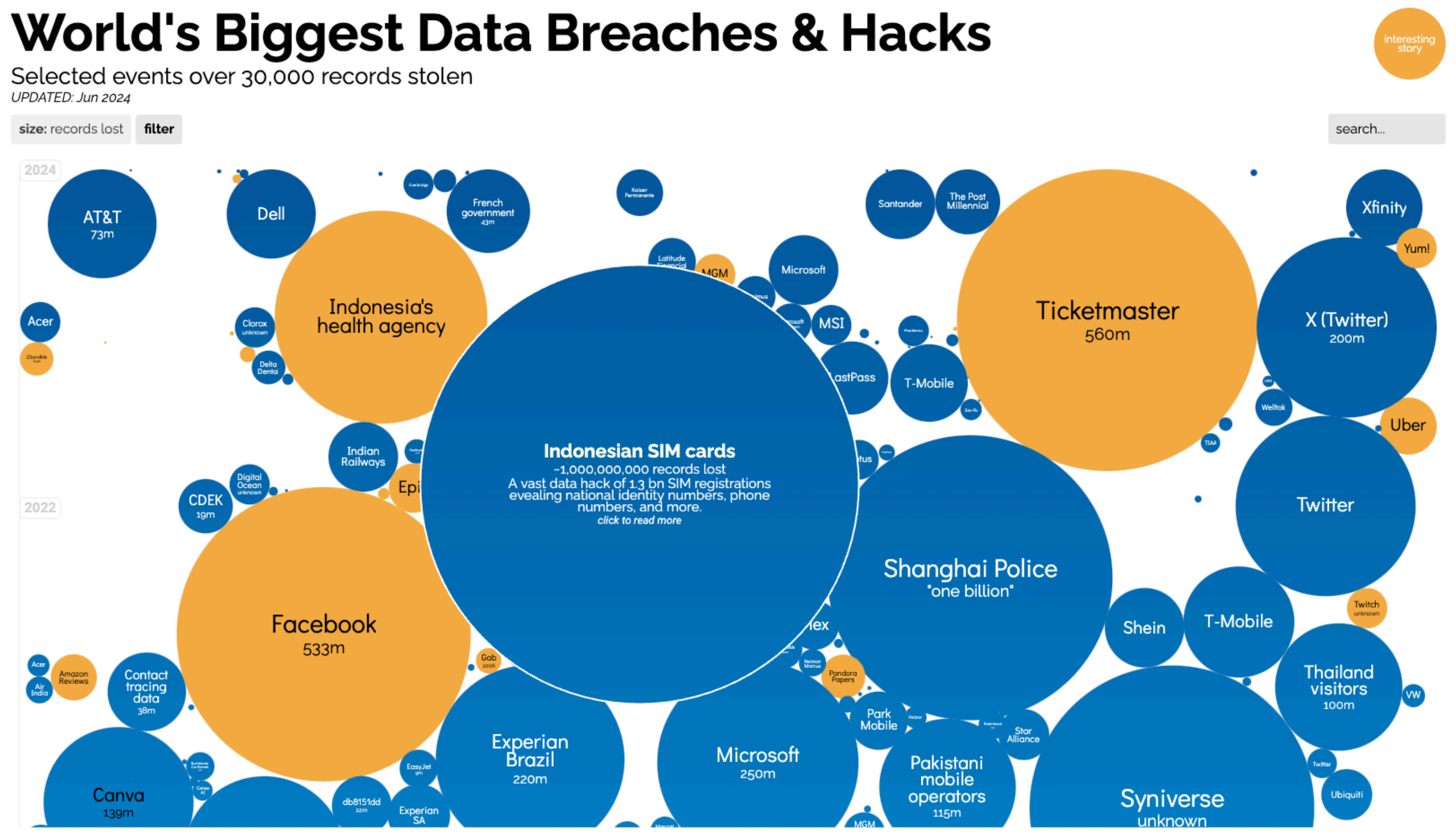
Team Lightgray (Information and Communication Technologies)

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 occurred, and it is barely half the year passed! [insert the dataset excel for references]

To raise awareness that every industry can be implicated by data breaches, it is important to show the sector and their respective companies in the sectors have been affected, not limiting to only smaller companies like Cooler Master, even big companies like Twitter have been a victim to data breach. This visualization has been dubbed "World's Biggest Data Breaches & Hacks", but it can be improved due to its current cluster and small wordings on less important data breaches that had happened.

Previous Visualization


World's Biggest Data Breaches & Hacks

Selected events over 30,000 records stolen

UPDATED: Jun 2024

size: records lost | filter | search...

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. Larger bubbles represent more significant breaches, while smaller bubbles indicate less severe incidents. This visual hierarchy makes it easy for viewers to quickly assess the relative impact of each data breach without needing to delve into numerical details.
- **Categorical Differentiation**: The use of color coding allows for quick identification of different categories or years. This helps viewers distinguish between various types of breaches or the periods in which they occurred. For instance, breaches from different years can be color-coded differently, making it simple to track temporal trends or categorize the breaches by industry or type.
- **Informative Labels**: Each bubble is labeled with the name of the affected entity and the number of records compromised. This labeling provides immediate context, allowing viewers to quickly identify which companies or organizations were breached and the extent of the breach. The clear labeling eliminates the need for viewers to cross-reference with external documents or data sources.
- **Trend Identification**: The vertical placement of the bubbles suggests a timeline, allowing viewers to see trends over the years. For example, the chart can show whether data breaches have become more frequent or larger in scale over time. This temporal element adds depth to the visualization, enabling viewers to understand not just the magnitude of individual breaches but also how the overall landscape of data breaches has evolved.
- **Visual Impact**: The bubble chart format has a strong visual appeal that captures attention. The vibrant colors and varying bubble sizes create a dynamic and engaging visualization. This immediate visual interest is crucial in data visualization, as it encourages viewers to explore the data further and helps in retaining their 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
- **Visual Consistency**: The use of consistent color coding for each sector 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
- **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.

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:

- **readcsv** for data import
- **tidyverse** for data transformation, including **dplyr** for data cleaning, grouping, and summarization
- **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

Improved Visualization

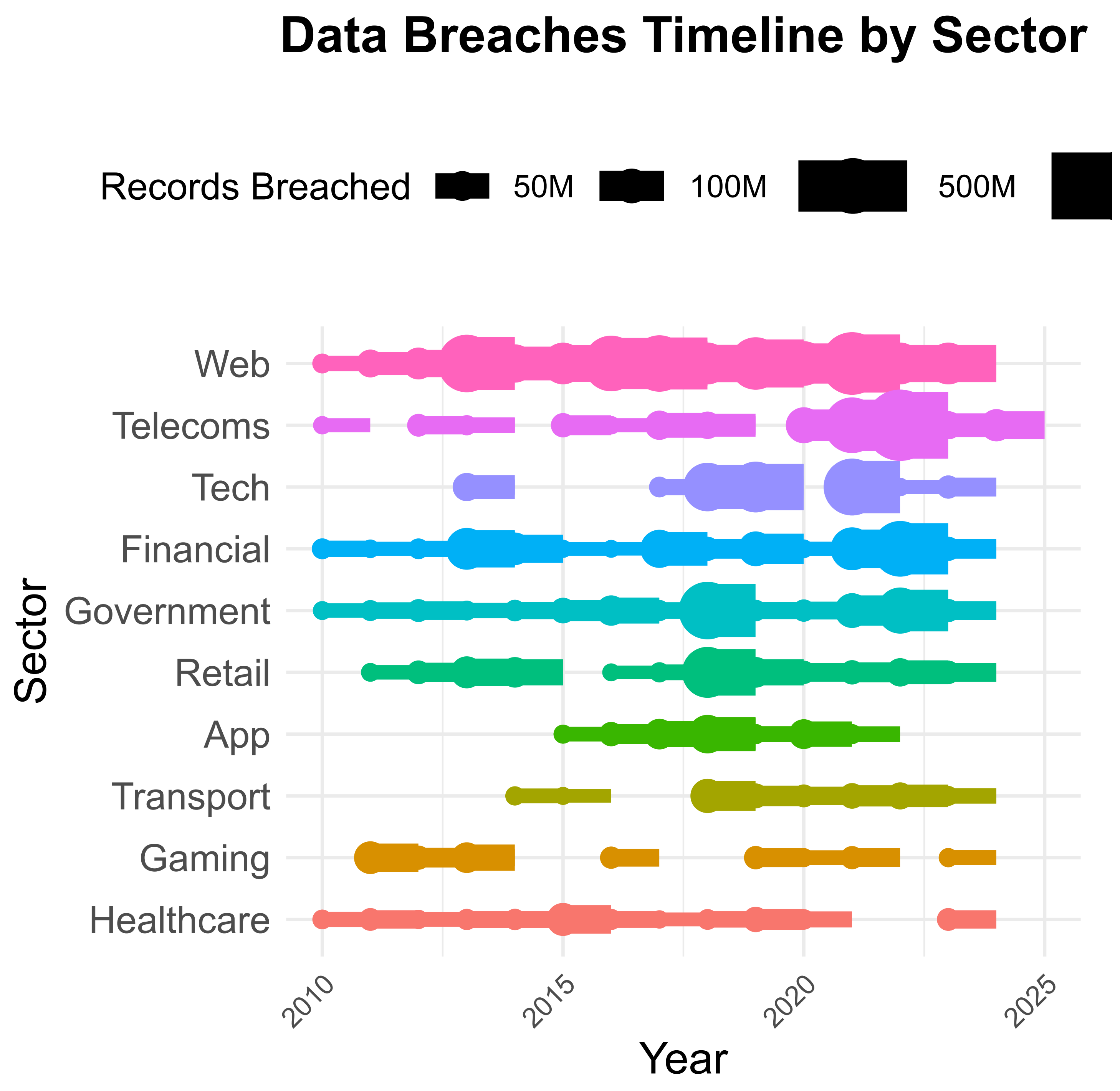


Figure 2: Data Breaches Timeline by Sector

Further Sugestions for interactivity

Conclusion