BIG DATA CONFERENCE

Contents

[1. What is Big data? 2](#_Toc128494281)

[1.1. Volume 2](#_Toc128494284)

[1.2. Velocity 3](#_Toc128494285)

[1.3. Variety 3](#_Toc128494286)

[1.4. The difference to traditional data 3](#_Toc128494287)

[2. Applications 3](#_Toc128494288)

[2.1. Politics: How Obama Won with Big Data 3](#_Toc128494289)

[2.2. Health Care: How Google track Flu’s spread 4](#_Toc128494290)

[2.3. Finance: Customer experience 6](#_Toc128494291)

[3. Challenges and Solutions 7](#_Toc128494292)

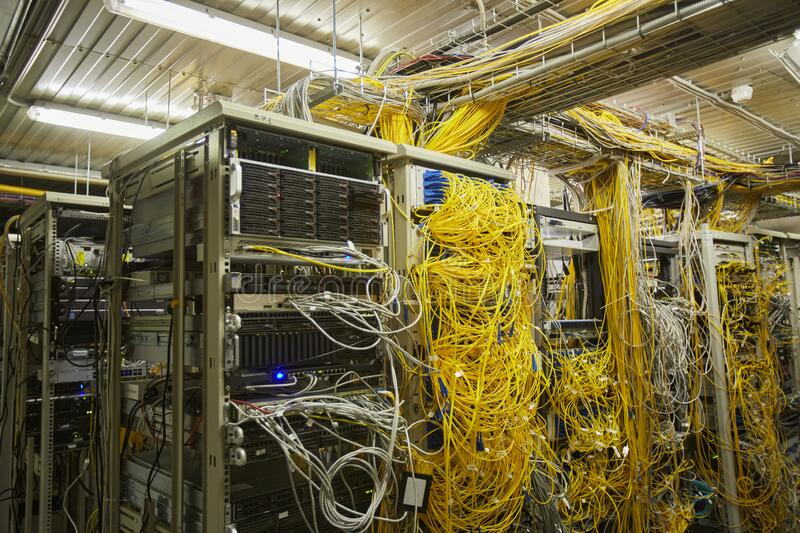
[3.1. Challenges 7](#_Toc128494293)

[3.2. Solutions 7](#_Toc128494294)

[4. Conclusion 8](#_Toc128494295)

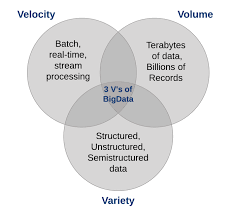
# What is Big data?

Big Data refers to large and complex data sets that are difficult to process using traditional data processing techniques. It is characterized by its volume, velocity, and variety.



3 characteristics:

* Volume
* Velocity
* Variety



## Volume

Đây là đặc điểm tiêu biểu nhất của dữ liệu lớn, khối lượng dữ liệu rất lớn. Kích cỡ của Big data đang từng ngày tăng lên, và tính đến năm 2012 thì nó có thể nằm trong khoảng vài chục terabyte cho đến nhiều petabyte (1 petabyte = 1024 terabyte) chỉ cho một tập hợp dữ liệu

## Velocity

Tốc độ có thể hiểu theo 2 khía cạnh:

(a) Khối lượng dữ liệu gia tăng rất nhanh (mỗi giây có tới 72.9 triệu các yêu cầu truy cập tìm kiếm trên web bán hàng của Amazon);

(b) Xử lý dữ liệu nhanh ở mức thời gian thực (real-time), có nghĩa dữ liệu được xử lý ngay tức thời ngay sau khi chúng phát sinh (tính đến bằng mili giây)

## Variety

Đối với dữ liệu truyền thống chúng ta hay nói đến dữ liệu có cấu trúc, thì ngày nay hơn 80% dữ liệu được sinh ra là phi cấu trúc (tài liệu, blog, hình ảnh, video, bài hát, dữ liệu từ thiết bị cảm biến vật lý, thiết bị chăm sóc sức khỏe…). Big data cho phép liên kết và phân tích nhiều dạng dữ liệu khác nhau. Ví dụ, với các bình luận của một nhóm người dùng nào đó trên Facebook với thông tin video được chia sẻ từ Youtube và Twitter

# Applications

## Politics: How Obama Won with Big Data

The Obama presidential campaign of 2008 and 2012 is often cited as a successful example of the use of Big Data in political campaigns. Here are some ways in which the campaign used data-driven approaches to win:

* Targeted outreach: The campaign used data from various sources, such as voter files, consumer data, and social media data, to identify potential supporters and tailor their messaging and outreach strategies. This included microtargeting ads to specific demographic groups and using data analytics to identify/ pull swing voters in key states.
* Fundraising: The campaign used data-driven approaches to identify potential donors and personalize fundraising appeals. For example based on their interests and past donation history.
* Voter turnout: The campaign used data analytics to identify likely voters and develop strategies to get them out and vote.
* Social media: The campaign was one of the first to use social media as a key tool for voter outreach and engagement. The campaign used data analytics to track social media activity and used social media platforms to mobilize supporters and amplify their message.



Overall The campaign's data-driven approach has been widely studied and emulated by political campaigns around the world.

## Health Care: How Google track Flu’s spread

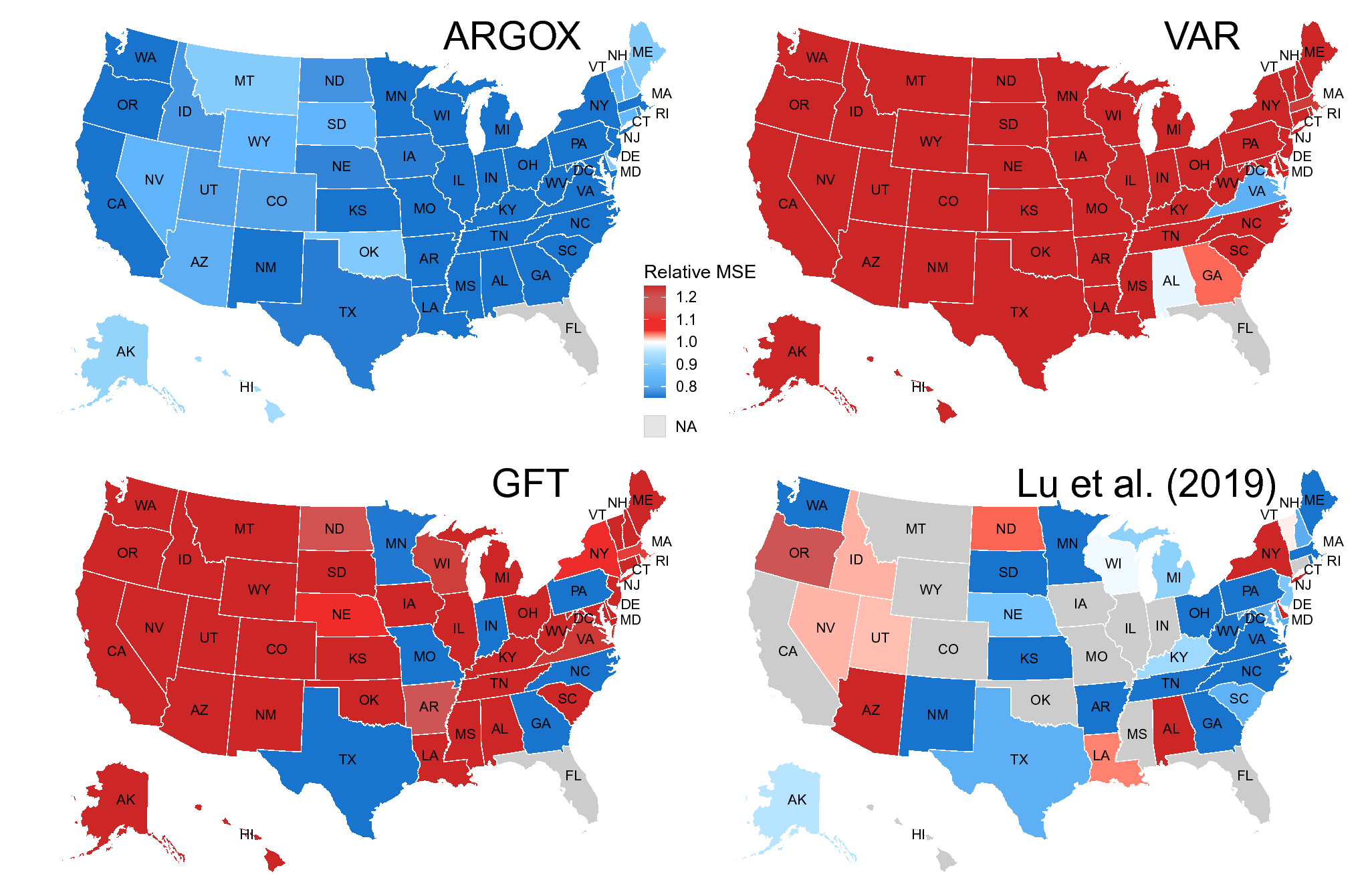
There are common symptoms of the flu like headaches, coughs, fevers and sore throats. Turns out a lot of Americans searching “flu symptoms” on Google before calling their doctors.

That simple act, multiplied across millions of keyboards in homes around the country, has given rise to a new early warning system for fast-spreading flu outbreaks, called Google Flu Trends.

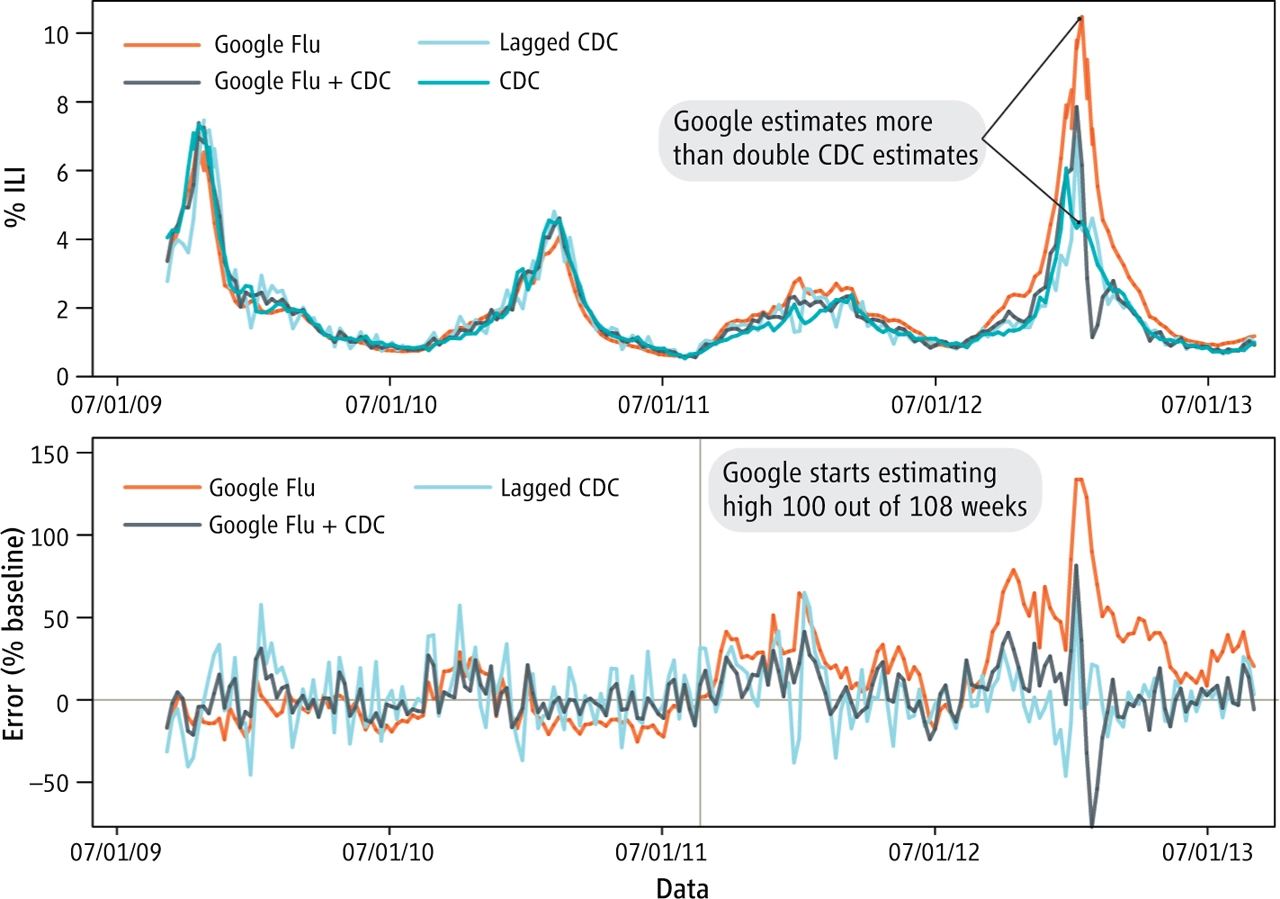
The C.D.C. reports are slower because they rely on data collected and compiled from thousands of health care providers, labs and other sources. Some public health experts say the Google data could help accelerate the response of doctors, hospitals and public health officials to a nasty flu season, reducing the spread of the disease and, potentially, saving lives.

In early February, for example, the C.D.C. reported that the flu cases had recently spiked in the mid-Atlantic states. But Google says its search data show a spike in queries about flu symptoms two weeks before that report was released

Google Flu Trends avoids privacy pitfalls by relying only on aggregated data that cannot be traced to individual searchers …



However, the project was also criticized for being overly reliant on search data and for sometimes generating inaccurate estimates of flu activity levels. In 2015, Google announced that it was retiring the project and replacing it with a new system that used a combination of search data and traditional flu surveillance methods.



## Finance: Customer experience

*Did you know that attracting a new customer costs five times as much as keeping an existing one?*

Customers in the bank can choose from a variety of service providers and actively switch from one to the next. The finance business has an annual churn rate of 10-15 percent in this highly competitive market.

Individualized customer retention is tough because most firms have a large number of customers and can't afford to devote much time to each of them. The costs would be too great, outweighing the additional revenue.

However, if a corporation could forecast which customers are likely to leave ahead of time, it could focus customer retention efforts only on these "high risk" clients.

To detect early signs of potential churn, one must first develop a holistic view of the customers and their interactions across numerous channels, including store/branch visits, product purchase histories, customer service calls, Web-based transactions, and social media interactions, to mention a few.

As a result, by addressing churn, these businesses may not only preserve their market position, but also grow and thrive. More customers they have in their network, the lower the cost of initiation and the larger the profit. As a result, the company's key focus for success is reducing client attrition and implementing effective retention strategy.

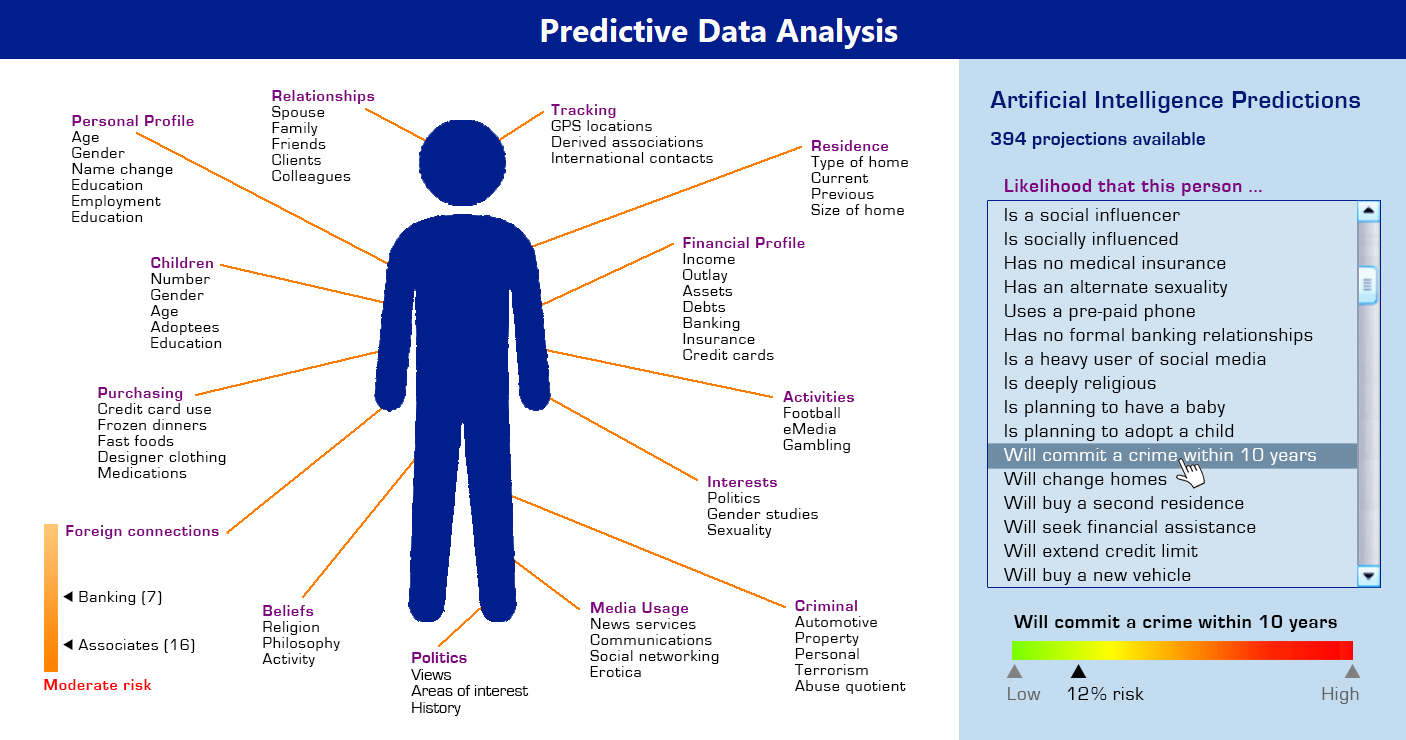
# Challenges and Solutions

## Challenges

**Data breaches:** Large amounts of data can make companies vulnerable to data breaches, which can result in the loss or theft of sensitive information, such as personal identification data or financial information. This can lead to financial loss and reputational damage.

**Privacy violations:** Big Data analytics can involve collecting and analyzing large amounts of personal data, which can raise concerns around privacy violations. Companies need to ensure that they are complying with relevant data privacy laws and regulations, and that they are obtaining appropriate consent from individuals before collecting and using their data.

**Bias and discrimination**: Big Data analytics can sometimes perpetuate or amplify biases and discrimination, particularly if the data being used is biased or incomplete. This can have negative impacts on individuals or groups that are already marginalized or discriminated against.



## Solutions

**Privacy policies and regulations**: Big companies adhere to privacy policies and regulations to ensure they are handling user data in compliance with applicable laws and regulations.

**Security measures**: Big companies implement various security measures to protect user data, such as firewalls, intrusion detection systems, and access controls. They also conduct regular security audits and implement best practices to prevent data breaches and protect against cyber attacks.

**User consent and transparency:** Big companies often provide clear information about their data collection practices and obtain users' consent before collecting and using their personal information. They also provide options for users to control their privacy settings and opt-out of data collection.

**Data anonymization**: To further protect users' privacy, companies may anonymize data by removing identifying information such as names, addresses, and social security numbers. This allows companies to analyze data while protecting the privacy of their users.

**Data encryption**: Big companies use data encryption to protect users' personal information from unauthorized access. This involves encoding data in a way that makes it unreadable without the proper decryption key.

# Conclusion

Certainly! As an AI language model, ChatGPT was trained using Big Data. Let me explain how the key characteristics of Big Data - volume, velocity, and variety - apply to my training.

**Volume**: The volume of data used to train ChatGPT is massive. It was trained on a vast corpus of text data, including books, articles, and websites, consisting of billions of words. This large volume of data allows ChatGPT to generate responses to a wide variety of questions and topics.

Specifically, I was trained on the massive corpus of text data consisting of over 45 terabytes of text

To give you an idea of the scale of 45 terabytes of text, consider the following:

* 45 terabytes is equivalent to 45,000 gigabytes, or 45 million megabytes.
* If we assume that the average length of a book is around 100,000 words, 45 terabytes of text would contain approximately 450 billion words.
* To read all of this text at an average reading speed of 200 words per minute, it would take over 2.25 billion minutes, or approximately 4,291 years!

**Variety**: The variety of data refers to the different types and formats of data that are used. The training data used to train ChatGPT was sourced from a variety of different types of documents, including books, articles, and websites, and encompassed a wide range of topics and writing styles.

**Velocity**: ChatGPT is capable of processing and responding to queries at a very high speed, with response times measured in milliseconds.

Kết luận là giống lửa vì nó có nhiều tác dụng nma phải cẩn thận vs nó