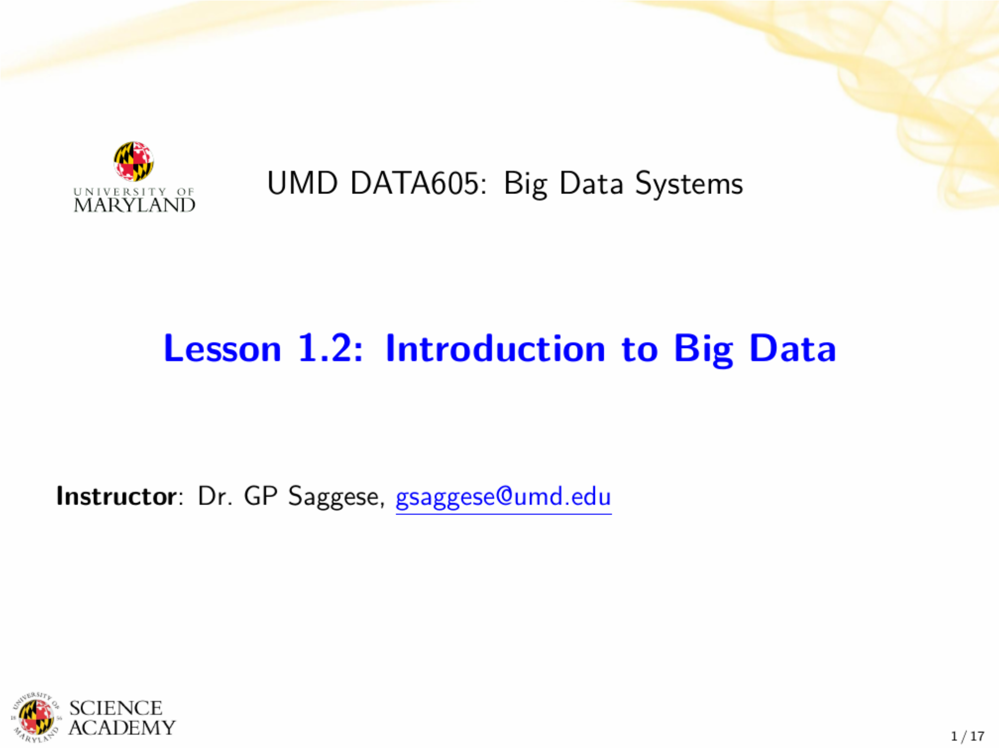
Lesson 1.2: Introduction to Big Data



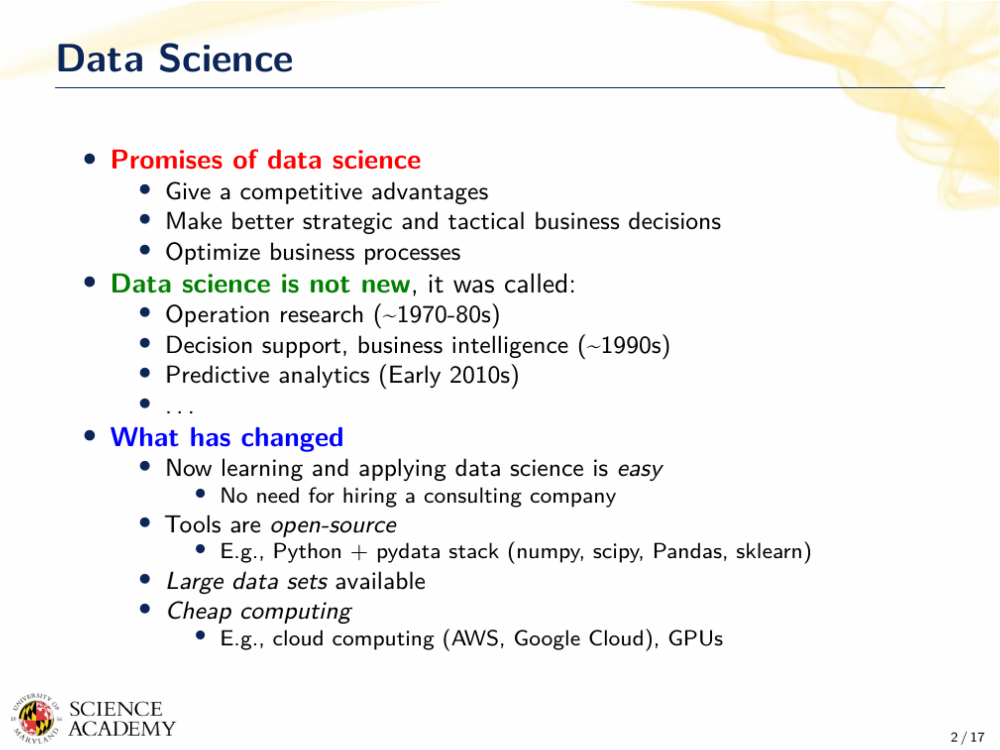
UMD DATA605: Big Data Systems

Lesson 1.2: Introduction to Big Data

Instructor: Dr. GP Saggese, gsaggese@umd.edu

SCIENCE ACADEMY

2 / 17: Data Science



Data Science

- Promises of data science
  - Give a competitive advantages
  - Make better strategic and tactical business decisions
  - Optimize business processes
- Data science is not new, it was called:
  - Operation research (~1970-80s)
  - Decision support, business intelligence (~1990s)
  - Predictive analytics (Early 2010s)
  - ...
- What has changed
  - Now learning and applying data science is *easy*
    - No need for hiring a consulting company
  - Tools are *open-source*
    - E.g., Python + pydata stack (numpy, scipy, Pandas, sklearn)
  - *Large data sets* available
  - *Cheap computing*
    - E.g., cloud computing (AWS, Google Cloud), GPUs

SCIENCE ACADEMY

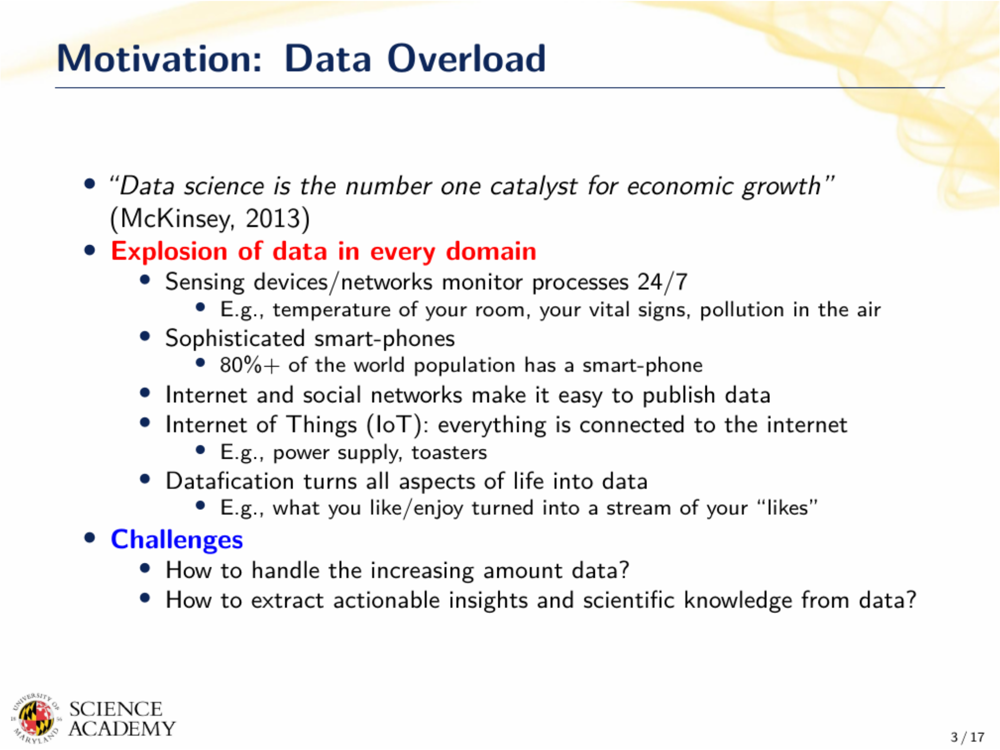
2 / 17

- **Promises of data science**
  - *Give a competitive advantage*: Data science can help businesses stand out by providing insights that others might not have. By analyzing data, companies can find new opportunities or improve their current operations, giving them an edge over competitors.
  - *Make better strategic and tactical business decisions*: With data science, businesses can make informed decisions based on data rather than intuition. This leads to more accurate and effective strategies and tactics.
  - *Optimize business processes*: Data science can identify inefficiencies and suggest improvements, leading to streamlined operations and cost savings.
- **Data science is not new**, it was called:
  - *Operation research (~1970-80s)*: This was an early form of data science focused on optimizing complex systems and processes.
  - *Decision support, business intelligence (~1990s)*: These terms referred to using data to support business decisions, similar to what data science does today.
  - *Predictive analytics (Early 2010s)*: This involved using data to predict future trends and behaviors, a key component of modern data science.
- **What has changed**
  - *Now learning and applying data science is easy*: In the past, businesses often needed to hire specialized consulting firms to implement data science solutions. Today, individuals can learn and apply these skills themselves.
  - *Tools are open-source*: Many powerful data science tools are freely available. For example, Python and its libraries (numpy, scipy, Pandas, sklearn) are widely used and accessible to everyone.
  - *Large data sets available*: The availability of vast amounts of data from various sources

has made it easier to perform comprehensive analyses.

- *Cheap computing*: The cost of computing power has decreased significantly. Cloud services like AWS and Google Cloud, along with GPUs, provide affordable and scalable computing resources.

3 / 17: Motivation: Data Overload



Motivation: Data Overload

- *"Data science is the number one catalyst for economic growth"* (McKinsey, 2013)
- **Explosion of data in every domain**
  - Sensing devices/networks monitor processes 24/7
    - E.g., temperature of your room, your vital signs, pollution in the air
  - Sophisticated smart-phones
    - 80%+ of the world population has a smart-phone
  - Internet and social networks make it easy to publish data
  - Internet of Things (IoT): everything is connected to the internet
    - E.g., power supply, toasters
  - Datafication turns all aspects of life into data
    - E.g., what you like/enjoy turned into a stream of your "likes"
- **Challenges**
  - How to handle the increasing amount data?
  - How to extract actionable insights and scientific knowledge from data?

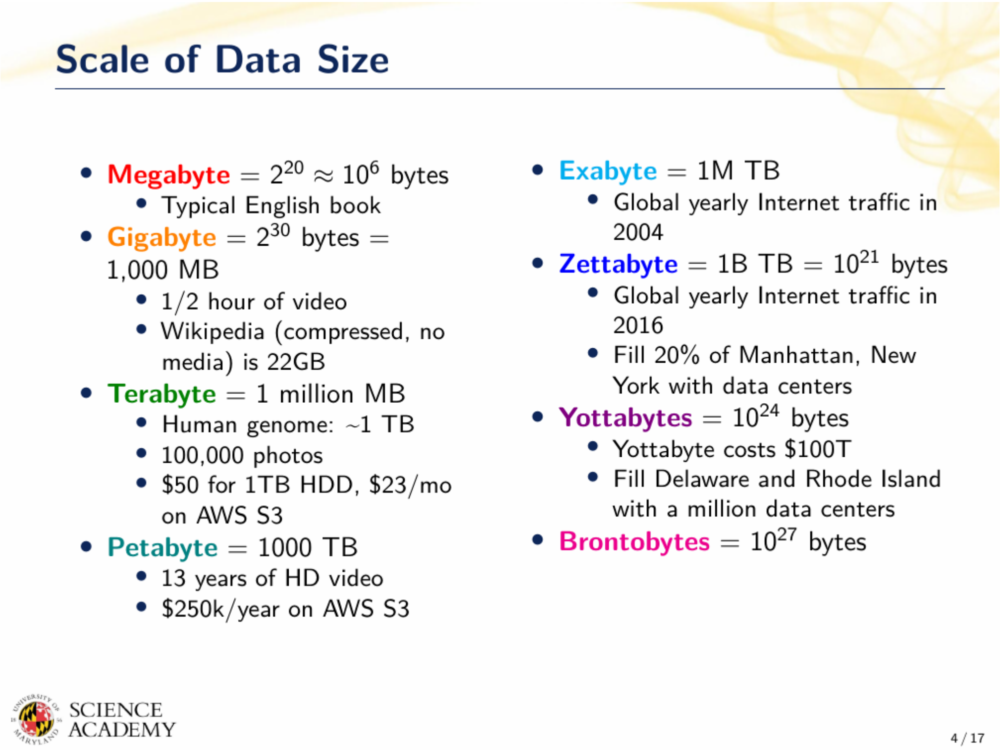
SCIENCE ACADEMY

3 / 17

- **Motivation: Data Overload**
  - The quote from McKinsey in 2013 highlights the importance of data science as a major driver for economic growth. This means that the ability to analyze and use data effectively can significantly boost economic activities and innovations.
- **Explosion of data in every domain**
  - *Sensing devices/networks monitor processes 24/7*: These devices continuously collect data about various aspects of our environment and lives. For example, sensors can track the temperature in your room, monitor your health through vital signs, or measure pollution levels in the air.
  - *Sophisticated smart-phones*: With over 80% of the global population owning a smartphone, these devices are a major source of data generation. They collect information through apps, GPS, and user interactions.
  - *Internet and social networks*: These platforms allow users to easily share and publish data, contributing to the vast amount of information available online.
  - *Internet of Things (IoT)*: This concept involves everyday objects being connected to the internet, enabling them to send and receive data. Examples include smart home devices like connected power supplies and toasters.
  - *Datafication*: This refers to the process of turning various aspects of life into data. For instance, your preferences and activities on social media are transformed into data streams, such as your "likes."
- **Challenges**
  - The main challenge is managing the sheer volume of data being generated. This involves storing, processing, and analyzing data efficiently.
  - Another challenge is extracting meaningful insights and scientific knowledge from this

data. This means finding ways to turn raw data into useful information that can inform decisions and drive innovation.

4 / 17: Scale of Data Size



Scale of Data Size

- **Megabyte** = $2^{20} \approx 10^6$ bytes
  - Typical English book
- **Gigabyte** = 2^{30} bytes = 1,000 MB
  - 1/2 hour of video
  - Wikipedia (compressed, no media) is 22GB
- **Terabyte** = 1 million MB
  - Human genome: ~1 TB
  - 100,000 photos
  - $50 for 1TB HDD, $23/mo on AWS S3
- **Petabyte** = 1000 TB
  - 13 years of HD video
  - $250k/year on AWS S3
- **Exabyte** = 1M TB
  - Global yearly Internet traffic in 2004
- **Zettabyte** = 1B TB = 10^{21} bytes
  - Global yearly Internet traffic in 2016
  - Fill 20% of Manhattan, New York with data centers
- **Yottabytes** = 10^{24} bytes
  - Yottabyte costs $100T
  - Fill Delaware and Rhode Island with a million data centers
- **Brontobytes** = 10^{27} bytes

SCIENCE ACADEMY

4 / 17

- **Megabyte**: A megabyte is a unit of digital information storage that equals approximately one million bytes. To put this into perspective, a typical English book, which contains text data, is about the size of a megabyte. This is a relatively small amount of data in today's digital world.

- **Gigabyte**: A gigabyte is 1,000 megabytes or 2^{30} bytes. This size can store about half an hour of video content. For reference, the entire text of Wikipedia, when compressed and excluding media files, is around 22 gigabytes. This shows how much more data a gigabyte can hold compared to a megabyte.

- **Terabyte**: A terabyte is equivalent to one million megabytes. It can store a human genome, which is approximately 1 terabyte in size, or about 100,000 photos. In terms of cost, a 1TB hard drive is relatively affordable at around $50, and storing this amount of data on AWS S3 cloud service costs about $23 per month.

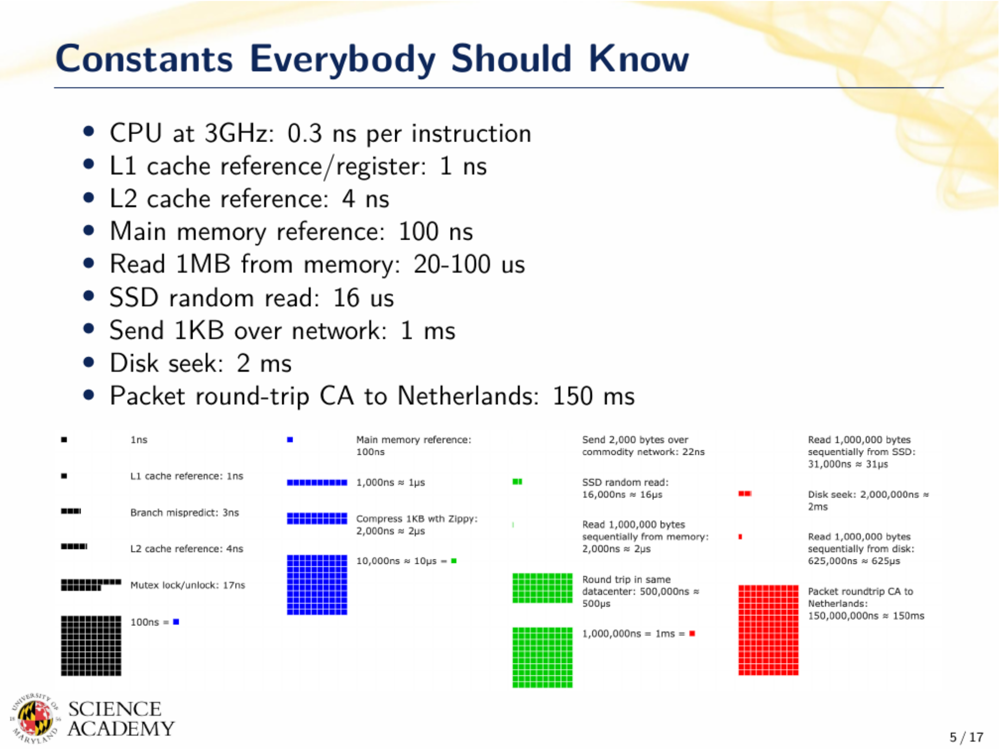
- **Petabyte**: A petabyte is 1,000 terabytes. This amount of storage can hold about 13 years of high-definition video. Storing a petabyte of data on AWS S3 would cost approximately $250,000 per year, highlighting the significant expense associated with managing large-scale data.

- **Exabyte**: An exabyte is one million terabytes. In 2004, the global yearly Internet traffic reached this scale, illustrating the rapid growth of data usage over time.

- **Zettabyte**: A zettabyte equals one billion terabytes or 10^{21} bytes. By 2016, global yearly Internet traffic had grown to this size. To visualize, storing a zettabyte of data would require enough data centers to fill 20% of Manhattan, New York.

- **Yottabytes**: A yottabyte is 10^{24} bytes. The cost of storing a yottabyte is estimated at $100 trillion, and it would require enough data centers to fill the states of Delaware and Rhode Island.
- **Brontobytes**: A brontobyte is 10^{27} bytes, representing an even larger scale of data storage, though it is not commonly used in practical scenarios today.

5 / 17: Constants Everybody Should Know



Constants Everybody Should Know

- CPU at 3GHz: 0.3 ns per instruction
- L1 cache reference/register: 1 ns
- L2 cache reference: 4 ns
- Main memory reference: 100 ns
- Read 1MB from memory: 20-100 us
- SSD random read: 16 us
- Send 1KB over network: 1 ms
- Disk seek: 2 ms
- Packet round-trip CA to Netherlands: 150 ms

1ns
L1 cache reference: 1ns
Branch mispredict: 3ns
L2 cache reference: 4ns
Mutex lock/unlock: 17ns
100ns = 🟦
Main memory reference: 100ns
1,000ns ≈ 1μs
Compress 1KB wth Zippy: 2,000ns ≈ 2μs
10,000ns ≈ 10μs = 🟩
Send 2,000 bytes over commodity network: 22ns
SSD random read: 16,000ns ≈ 16μs
Read 1,000,000 bytes sequentially from memory: 2,000ns ≈ 2μs
Round trip in same datacenter: 500,000ns ≈ 500μs
1,000,000ns = 1ms = 🟥
Read 1,000,000 bytes sequentially from SSD: 31,000ns ≈ 31μs
Disk seek: 2,000,000ns ≈ 2ms
Read 1,000,000 bytes sequentially from disk: 625,000ns ≈ 625μs
Packet roundtrip CA to Netherlands: 150,000,000ns ≈ 150ms

SCIENCE ACADEMY

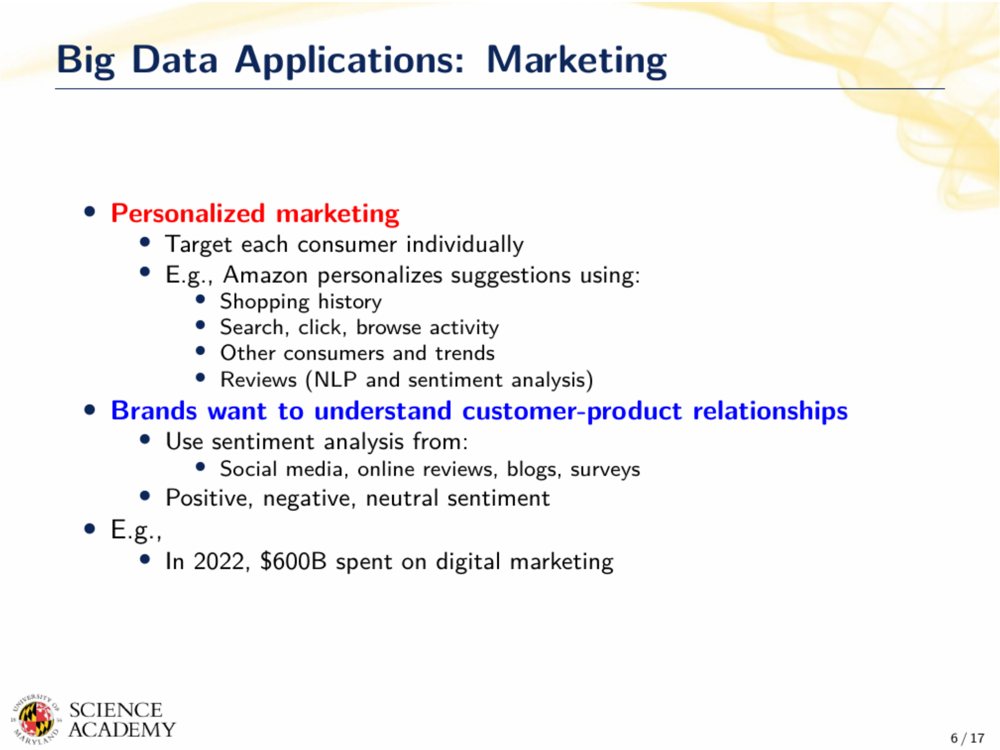
5 / 17

- **CPU at 3GHz: 0.3 ns per instruction**
  - This means that a CPU running at 3GHz can execute one instruction every 0.3 nanoseconds. This is incredibly fast and highlights the efficiency of modern processors. Understanding this helps us appreciate how quickly a CPU can process data and execute commands.
- **L1 cache reference/register: 1 ns**
  - The L1 cache is the fastest type of memory available to the CPU, located directly on the processor chip. Accessing data from the L1 cache takes about 1 nanosecond, which is very quick and crucial for performance in executing instructions.
- **L2 cache reference: 4 ns**
  - L2 cache is slightly slower than L1 but still much faster than accessing main memory. It takes about 4 nanoseconds to access data from the L2 cache, which serves as an intermediary storage to reduce the time needed to fetch data from the main memory.
- **Main memory reference: 100 ns**
  - Accessing data from the main memory (RAM) is significantly slower than accessing cache memory. It takes about 100 nanoseconds, which is why having efficient cache usage is important for performance.
- **Read 1MB from memory: 20-100 us**
  - Reading a megabyte of data from memory can take between 20 to 100 microseconds. This range depends on various factors like memory speed and system architecture. It's important to note the difference in scale compared to cache access times.
- **SSD random read: 16 us**
  - Solid State Drives (SSDs) are much faster than traditional hard drives. A random read operation on an SSD takes about 16 microseconds, which is relatively quick and beneficial

for applications requiring fast data retrieval.

- **Send 1KB over network: 1 ms**
  - Sending a kilobyte of data over a network typically takes about 1 millisecond. Network latency can vary based on distance and network conditions, but this gives a general idea of the time involved in data transmission.
- **Disk seek: 2 ms**
  - Disk seek time refers to the time it takes for a hard drive's read/write head to move to the position on the disk where data is stored. This process takes about 2 milliseconds, which is relatively slow compared to SSDs.
- **Packet round-trip CA to Netherlands: 150 ms**
  - This is the time it takes for a data packet to travel from California to the Netherlands and back. At 150 milliseconds, it highlights the latency involved in long-distance internet communications, which can impact real-time applications like video calls.

6 / 17: Big Data Applications: Marketing



Big Data Applications: Marketing

- Personalized marketing
  - Target each consumer individually
  - E.g., Amazon personalizes suggestions using:
    - Shopping history
    - Search, click, browse activity
    - Other consumers and trends
    - Reviews (NLP and sentiment analysis)
- Brands want to understand customer-product relationships
  - Use sentiment analysis from:
    - Social media, online reviews, blogs, surveys
  - Positive, negative, neutral sentiment
- E.g.,
  - In 2022, $600B spent on digital marketing

SCIENCE ACADEMY

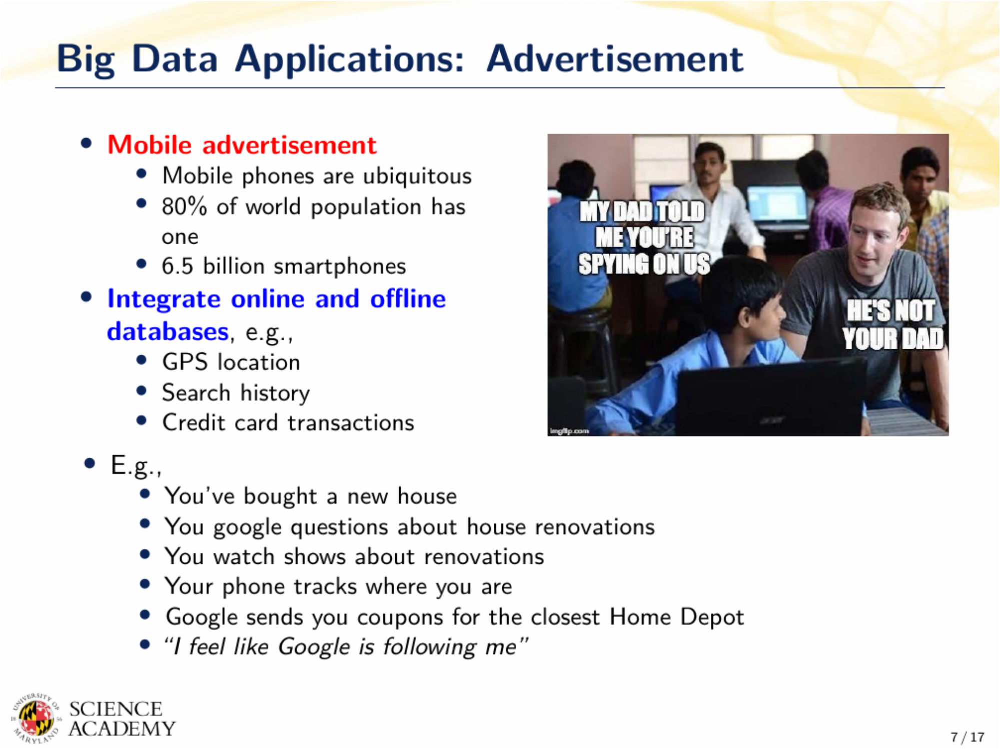
6 / 17

- **Big Data Applications: Marketing**
- **Personalized marketing**
  - *Target each consumer individually*: This means using data to tailor marketing efforts to the specific preferences and behaviors of each customer. Instead of a one-size-fits-all approach, companies aim to make each customer feel like the marketing is just for them.
  - *E.g., Amazon personalizes suggestions using*: Amazon is a great example of personalized marketing. They use a variety of data points to recommend products to users.
    - **Shopping history**: This includes what you've bought before, which helps predict what you might want to buy next.
    - **Search, click, browse activity**: Every time you search for a product, click on a link, or browse through items, Amazon collects this data to understand your interests.
    - **Other consumers and trends**: By analyzing what similar customers are buying and current market trends, Amazon can suggest popular or trending items.
    - **Reviews (NLP and sentiment analysis)**: Natural Language Processing (NLP) and sentiment analysis help Amazon understand the tone of reviews, whether they are positive, negative, or neutral, to better recommend products.
- **Brands want to understand customer-product relationships**
  - *Use sentiment analysis from*: Companies use sentiment analysis to gauge how customers feel about their products. This involves analyzing text data from various sources.
    - **Social media, online reviews, blogs, surveys**: These platforms provide a wealth of data where customers express their opinions and feelings about products.

- *Positive, negative, neutral sentiment*: By categorizing sentiments, brands can understand overall customer satisfaction and areas needing improvement.

- **E.g.,**
  - *In 2022, $600B spent on digital marketing*: This highlights the massive investment in digital marketing, emphasizing the importance of big data in understanding and reaching consumers effectively.

7 / 17: Big Data Applications: Advertisement

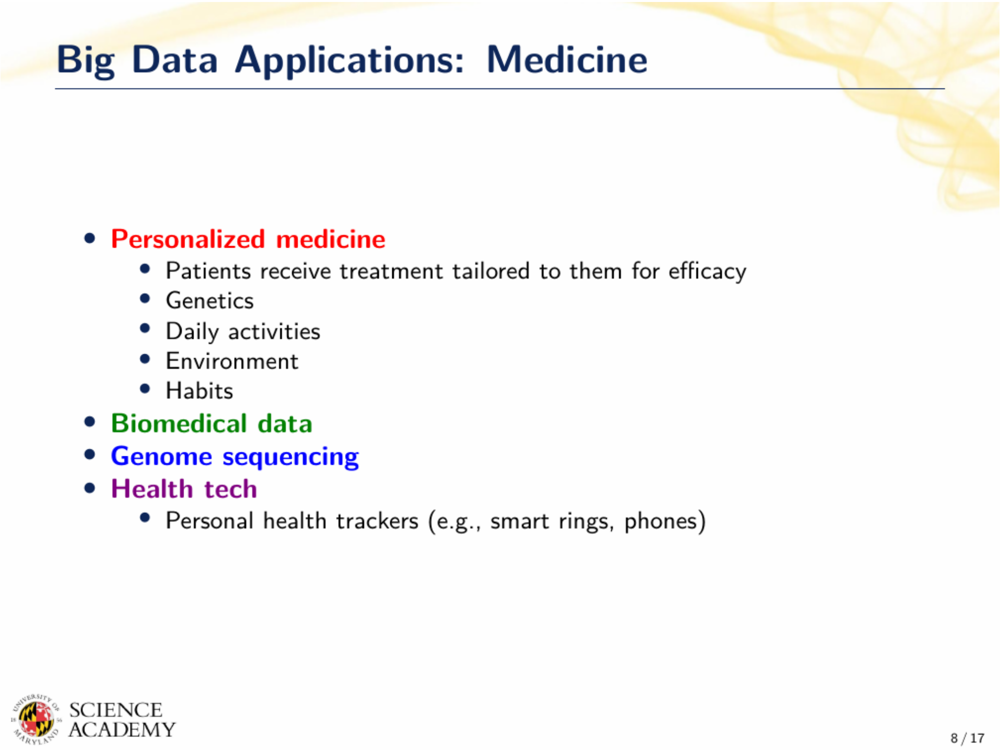


Big Data Applications: Advertisement

- **Mobile advertisement**
  - Mobile phones are ubiquitous
  - 80% of world population has one
  - 6.5 billion smartphones
- **Integrate online and offline databases**, e.g.,
  - GPS location
  - Search history
  - Credit card transactions
- E.g.,
  - You've bought a new house
  - You google questions about house renovations
  - You watch shows about renovations
  - Your phone tracks where you are
  - Google sends you coupons for the closest Home Depot
  - *"I feel like Google is following me"*

- **Mobile advertisement**
  - Mobile phones are everywhere, and they play a huge role in how companies reach potential customers. With 80% of the world's population owning a mobile phone, and 6.5 billion of those being smartphones, advertisers have a massive audience to target. This means that almost everyone is reachable through their mobile device, making it a powerful tool for advertising.
- **Integrate online and offline databases**
  - Advertisers use a combination of online and offline data to create targeted ads. For example, they might use your GPS location to know where you are, your search history to understand your interests, and your credit card transactions to see what you buy. By combining these data sources, advertisers can create personalized ads that are more likely to catch your attention.
- **Example scenario**
  - Imagine you've just bought a new house. You start searching online for renovation tips and watching home improvement shows. Meanwhile, your phone tracks your location. Based on this data, Google might send you coupons for the nearest Home Depot. This targeted advertising can feel very personal, leading to the feeling that "Google is following me." This example illustrates how big data is used to create highly personalized advertising experiences.

8 / 17: Big Data Applications: Medicine



Big Data Applications: Medicine

- Personalized medicine
  - Patients receive treatment tailored to them for efficacy
  - Genetics
  - Daily activities
  - Environment
  - Habits
- Biomedical data
- Genome sequencing
- Health tech
  - Personal health trackers (e.g., smart rings, phones)

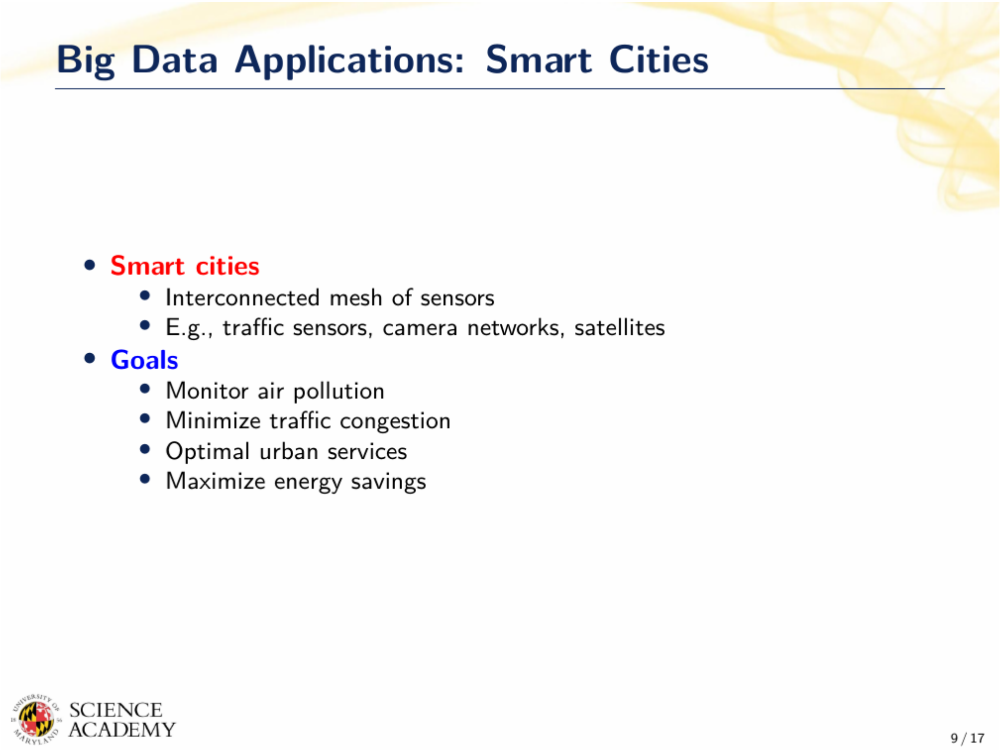
SCIENCE ACADEMY

8 / 17

- **Big Data Applications: Medicine**
- **Personalized medicine**
  - *Personalized medicine* is a medical approach where treatments are customized for individual patients. This means that instead of a one-size-fits-all treatment, doctors use data to tailor healthcare specifically for each person.
  - **Genetics**: By analyzing a patient's genetic information, doctors can predict how they might respond to certain medications or what diseases they might be at risk for.
  - **Daily activities**: Information about a patient's lifestyle, such as exercise routines and diet, can help in crafting a more effective treatment plan.
  - **Environment**: Environmental factors, like where a person lives and works, can influence their health and are considered in personalized treatments.
  - **Habits**: Understanding a patient's habits, such as smoking or alcohol consumption, is crucial for creating a personalized healthcare plan.
- **Biomedical data**
  - This refers to the vast amount of data generated in the medical field, including patient records, lab results, and imaging data. Analyzing this data helps in improving patient care and advancing medical research.
- **Genome sequencing**
  - Genome sequencing involves decoding a person's DNA to understand their genetic makeup. This information is vital for identifying genetic disorders and tailoring treatments to an individual's genetic profile.

- **Health tech**
  - Personal health trackers, like smart rings and phones, collect data on various health metrics such as heart rate, sleep patterns, and physical activity. This data can be used to monitor health in real-time and make informed decisions about lifestyle changes or medical interventions.

9 / 17: Big Data Applications: Smart Cities



Big Data Applications: Smart Cities

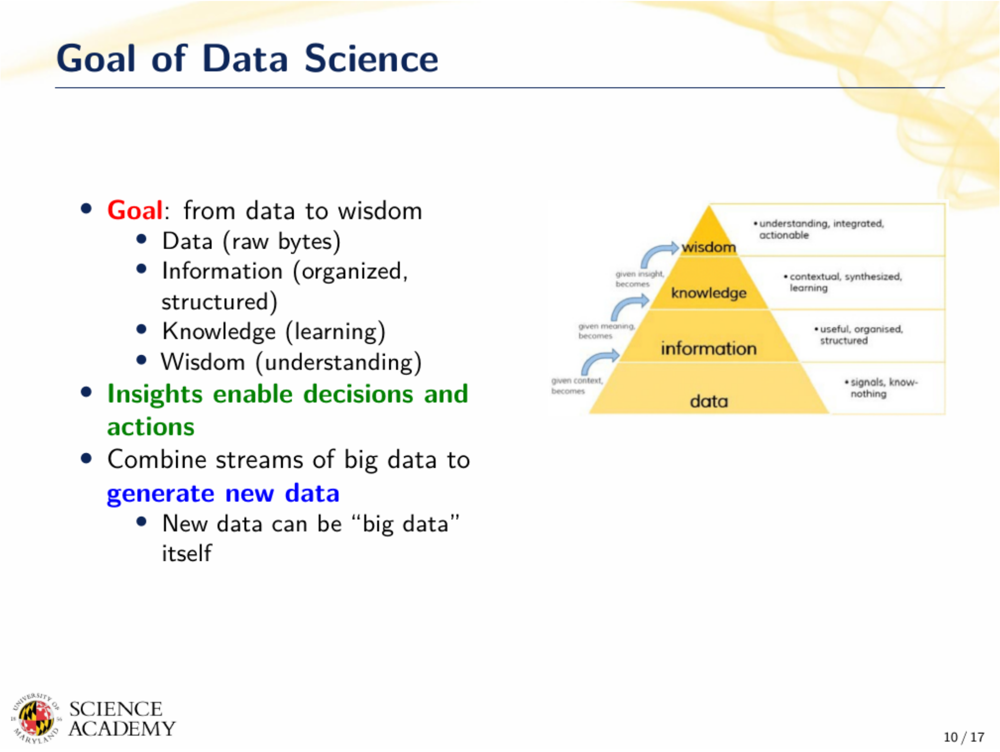
- Smart cities
  - Interconnected mesh of sensors
  - E.g., traffic sensors, camera networks, satellites
- Goals
  - Monitor air pollution
  - Minimize traffic congestion
  - Optimal urban services
  - Maximize energy savings

SCIENCE
ACADEMY

9 / 17

- **Smart cities**
  - Smart cities are urban areas that use technology and data to improve the quality of life for their residents. They rely on an *interconnected mesh of sensors* to collect and analyze data. This network of sensors can include devices like traffic sensors, camera networks, and even satellites. These tools work together to gather real-time information about various aspects of city life.
- **Goals**
  - One of the primary goals of smart cities is to *monitor air pollution*. By using sensors to track air quality, cities can identify pollution sources and take action to improve the environment.
  - Another goal is to *minimize traffic congestion*. Traffic sensors and camera networks help manage traffic flow, reduce bottlenecks, and improve commute times.
  - Smart cities aim to provide *optimal urban services*. This means using data to enhance public services like waste management, water supply, and emergency response.
  - Finally, smart cities strive to *maximize energy savings*. By analyzing energy usage patterns, cities can implement strategies to reduce consumption and promote sustainability.

10 / 17: Goal of Data Science



Goal of Data Science

- Goal: from data to wisdom
  - Data (raw bytes)
  - Information (organized, structured)
  - Knowledge (learning)
  - Wisdom (understanding)
- Insights enable decisions and actions
- Combine streams of big data to generate new data
  - New data can be "big data" itself

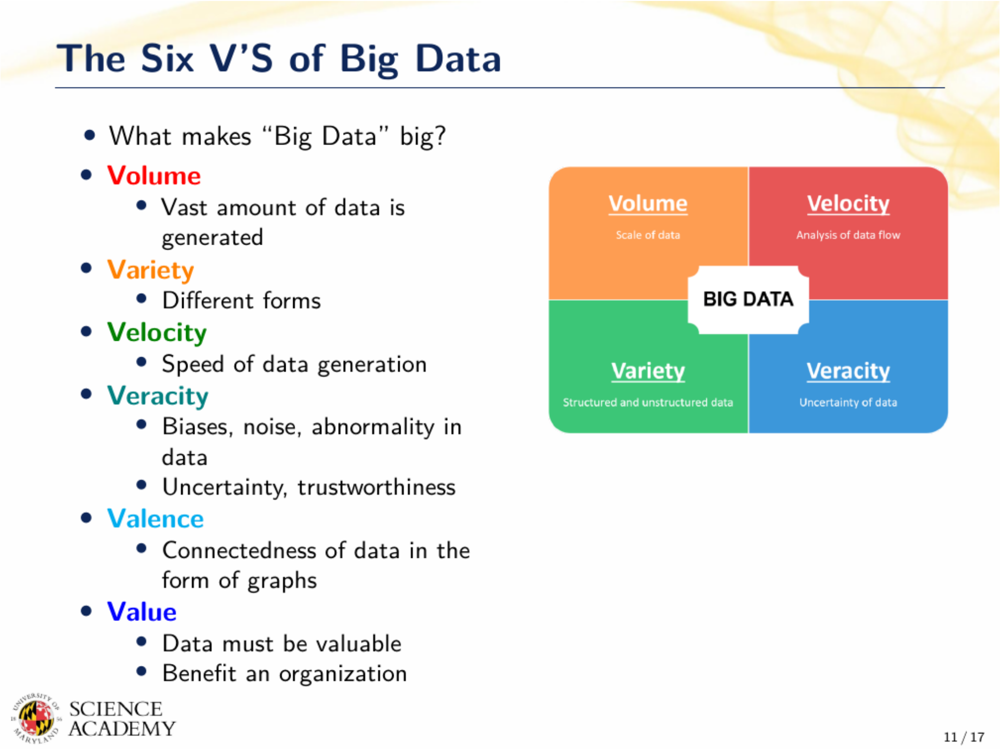
wisdom
• understanding, integrated, actionable
given insight, becomes
knowledge
• contextual, synthesized, learning
given meaning, becomes
information
• useful, organised, structured
given context, becomes
data
• signals, know-nothing

SCIENCE ACADEMY

10 / 17

- **Goal**: from data to wisdom
  - The journey from data to wisdom is a process of transforming raw data into something meaningful. **Data** refers to the raw bytes or unprocessed facts and figures. It's the starting point of any data science project.
  - **Information** is what you get when you organize and structure data. This step involves cleaning and formatting data so it can be analyzed.
  - **Knowledge** is derived from analyzing information. It involves learning patterns, trends, and insights from the data.
  - **Wisdom** is the ultimate goal, where you not only understand the data but can also apply it to make informed decisions and solve real-world problems.
- **Insights enable decisions and actions**
  - The insights gained from data analysis are crucial because they inform decisions and actions. Without insights, data remains just numbers and figures without practical application.
- Combine streams of big data to **generate new data**
  - In data science, combining different data streams can lead to the creation of new data sets. This new data can be considered "big data" itself, as it often involves large volumes, high velocity, and a variety of data types. This process is essential for uncovering deeper insights and creating more comprehensive models.

11 / 17: The Six V’S of Big Data



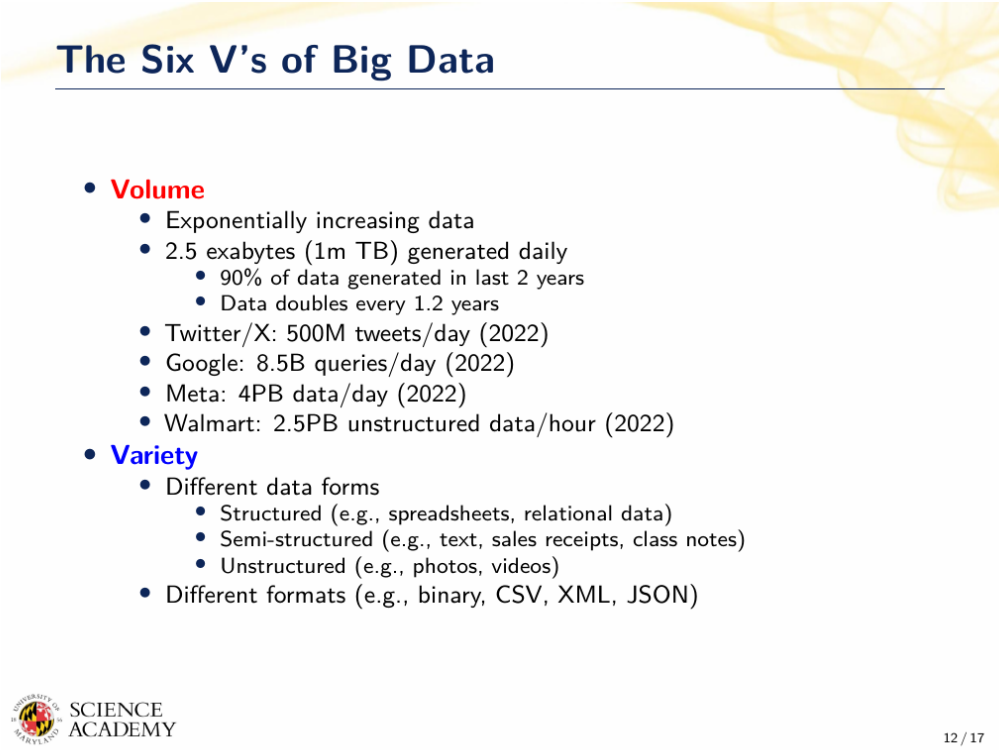
The Six V’S of Big Data

- What makes “Big Data” big?
- Volume
  - Vast amount of data is generated
- Variety
  - Different forms
- Velocity
  - Speed of data generation
- Veracity
  - Biases, noise, abnormality in data
  - Uncertainty, trustworthiness
- Valence
  - Connectedness of data in the form of graphs
- Value
  - Data must be valuable
  - Benefit an organization

- **Volume**
  - *Volume* refers to the sheer amount of data that is being generated every second. In today’s digital world, data is being produced at an unprecedented rate from various sources like social media, sensors, and transactions. This massive volume of data is what makes it “big” and requires special tools and technologies to store, process, and analyze it effectively.
- **Variety**
  - *Variety* highlights the different forms and types of data that are available. Data can be structured, like databases, or unstructured, like videos, images, and text. This diversity in data types presents challenges in terms of storage, processing, and analysis, as different types of data require different handling techniques.
- **Velocity**
  - *Velocity* refers to the speed at which data is generated and needs to be processed. With the rise of real-time data sources such as social media feeds and IoT devices, the ability to quickly process and analyze data as it arrives is crucial for making timely decisions.
- **Veracity**
  - *Veracity* deals with the quality and trustworthiness of the data. Data can often be noisy, biased, or incomplete, which can affect the accuracy of the insights derived from it. Ensuring data veracity is essential for making reliable decisions based on data analysis.
- **Valence**
  - *Valence* is about the connectedness of data, often represented in the form of graphs. It emphasizes the relationships and interactions between different data points, which can provide deeper insights into patterns and trends.
- **Value**

- *Value* is perhaps the most important aspect, as it focuses on the usefulness of the data. For data to be considered “big,” it must provide value to an organization, helping it to achieve its goals, improve operations, or gain a competitive advantage. Without value, the other characteristics of big data are meaningless.

12 / 17: The Six V's of Big Data



The Six V's of Big Data

- Volume
  - Exponentially increasing data
  - 2.5 exabytes (1m TB) generated daily
    - 90% of data generated in last 2 years
    - Data doubles every 1.2 years
  - Twitter/X: 500M tweets/day (2022)
  - Google: 8.5B queries/day (2022)
  - Meta: 4PB data/day (2022)
  - Walmart: 2.5PB unstructured data/hour (2022)
- Variety
  - Different data forms
    - Structured (e.g., spreadsheets, relational data)
    - Semi-structured (e.g., text, sales receipts, class notes)
    - Unstructured (e.g., photos, videos)
  - Different formats (e.g., binary, CSV, XML, JSON)

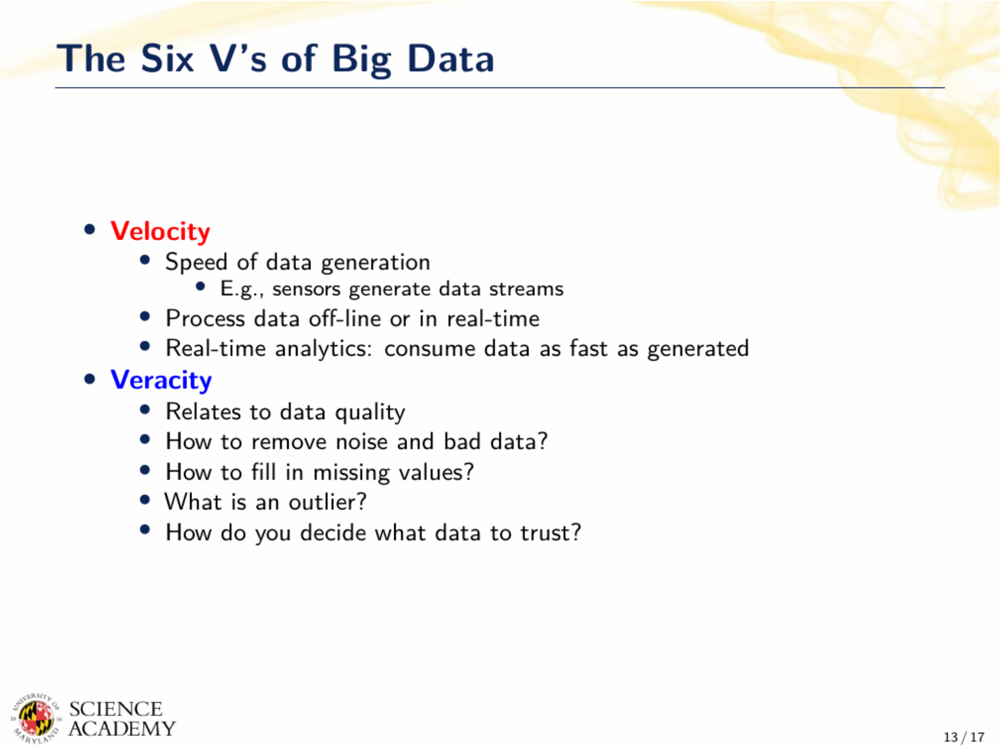
SCIENCE ACADEMY

12 / 17

- **Volume**
  - The term *Volume* refers to the sheer amount of data being generated and collected. This is a defining characteristic of big data, as the quantities are so large that traditional data processing software cannot handle them efficiently.
  - An *exabyte* is a unit of data equal to one million terabytes, and currently, about 2.5 exabytes of data are created every day. This highlights the rapid growth of data generation.
  - It's noteworthy that 90% of the world's data has been created in just the last two years, illustrating the exponential growth rate. This means that the amount of data doubles approximately every 1.2 years.
  - Examples of this massive data generation include platforms like Twitter/X, which sees 500 million tweets daily, and Google, which processes 8.5 billion search queries each day.
  - Companies like Meta and Walmart handle enormous amounts of data daily, with Meta processing 4 petabytes and Walmart dealing with 2.5 petabytes of unstructured data every hour.
- **Variety**
  - *Variety* refers to the different types of data that are being generated. Unlike traditional data, which was mostly structured, big data comes in various forms.
  - Structured data is organized and easily searchable, like data in spreadsheets or databases.
  - Semi-structured data includes elements of both structured and unstructured data, such as text documents, sales receipts, or class notes, which have some organizational properties but are not as rigidly formatted.
  - Unstructured data is more complex and includes formats like photos and videos, which do not fit neatly into tables or databases.

- Data can also come in different formats, such as binary, CSV, XML, and JSON, each requiring different methods for processing and analysis. This diversity in data types and formats presents both challenges and opportunities for data analysis.

13 / 17: The Six V's of Big Data



The Six V's of Big Data

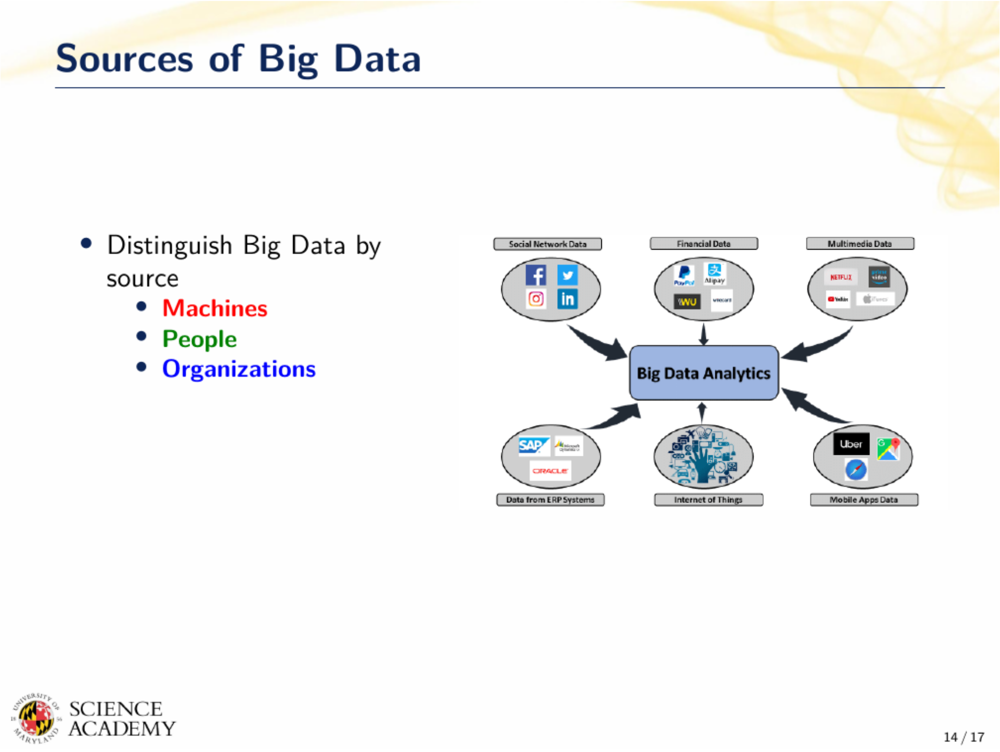
- Velocity
  - Speed of data generation
    - E.g., sensors generate data streams
  - Process data off-line or in real-time
  - Real-time analytics: consume data as fast as generated
- Veracity
  - Relates to data quality
  - How to remove noise and bad data?
  - How to fill in missing values?
  - What is an outlier?
  - How do you decide what data to trust?

SCIENCE ACADEMY

13 / 17

- **Velocity**
  - *Speed of data generation* refers to how quickly data is being produced. In today's world, data is generated at an unprecedented rate, especially from sources like sensors, social media, and online transactions.
    - For example, sensors in smart devices or industrial machines continuously produce data streams that need to be managed and analyzed.
  - *Process data off-line or in real-time* highlights the decision organizations must make regarding how they handle incoming data. Offline processing involves storing data for later analysis, while real-time processing means analyzing data as it arrives.
  - *Real-time analytics* is crucial for applications that require immediate insights, such as fraud detection or monitoring system health. It involves consuming and analyzing data as quickly as it is generated to make timely decisions.
- **Veracity**
  - *Relates to data quality* emphasizes the importance of ensuring that the data being used is accurate and reliable. Poor data quality can lead to incorrect insights and decisions.
  - *How to remove noise and bad data?* involves identifying and eliminating irrelevant or erroneous data that can skew analysis results.
  - *How to fill in missing values?* is about determining the best method to handle incomplete data, which might involve estimation or using default values.
  - *What is an outlier?* refers to data points that are significantly different from others in the dataset. Identifying outliers is important as they can indicate errors or unique insights.
  - *How do you decide what data to trust?* involves establishing criteria or using tools to assess the reliability of data sources, ensuring that decisions are based on credible information.

14 / 17: Sources of Big Data



Sources of Big Data

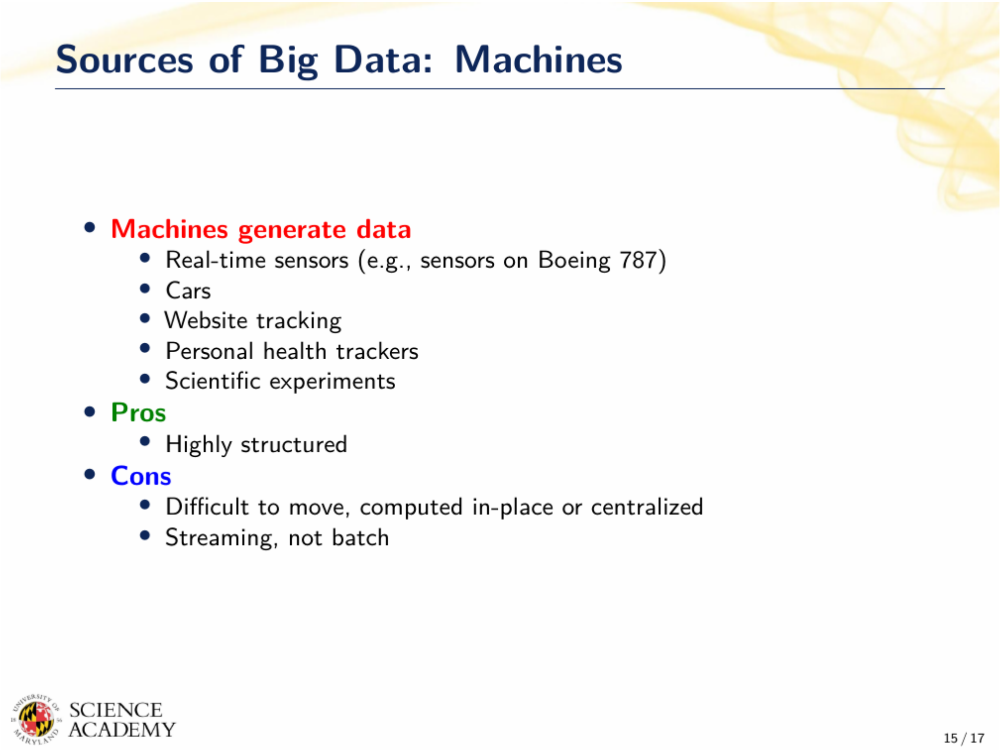
- Distinguish Big Data by source
  - Machines
  - People
  - Organizations

SCIENCE ACADEMY

14 / 17

- **Distinguish Big Data by source**
  - **Machines**
    - Machines generate a vast amount of data through sensors, logs, and automated processes. This includes data from IoT devices, industrial machines, and computer systems. Machine-generated data is typically structured and can be collected at a high velocity, making it a significant component of big data.
  - **People**
    - People contribute to big data through their interactions with digital platforms. This includes social media posts, online transactions, and user-generated content. This type of data is often unstructured and diverse, encompassing text, images, and videos. It provides insights into human behavior and preferences.
  - **Organizations**
    - Organizations produce data through their operations, such as sales records, customer databases, and financial transactions. This data is usually structured and is crucial for business analytics. Organizations also collect data from external sources to enhance their decision-making processes.
- *Context*: Understanding the sources of big data is essential for effectively managing and analyzing it. Each source has unique characteristics and challenges, influencing how data is processed and utilized.

15 / 17: Sources of Big Data: Machines



Sources of Big Data: Machines

- **Machines generate data**
  - Real-time sensors (e.g., sensors on Boeing 787)
  - Cars
  - Website tracking
  - Personal health trackers
  - Scientific experiments
- **Pros**
  - Highly structured
- **Cons**
  - Difficult to move, computed in-place or centralized
  - Streaming, not batch

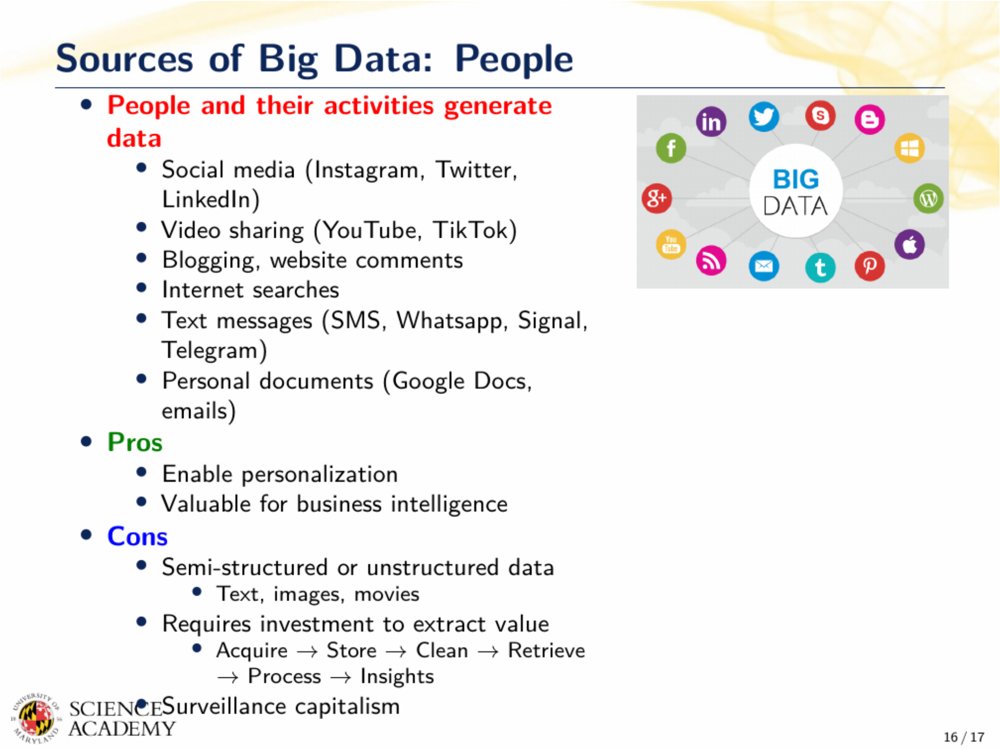
SCIENCE ACADEMY

15 / 17

- **Sources of Big Data: Machines**
  - **Machines generate data**
    - Machines are a significant source of big data. They produce data through various means, such as *real-time sensors*. For example, a Boeing 787 airplane is equipped with numerous sensors that continuously collect data during flights. This data can include information about engine performance, fuel efficiency, and environmental conditions.
    - Cars are another example, as modern vehicles are equipped with sensors that monitor everything from tire pressure to engine diagnostics.
    - Websites track user interactions, generating data about user behavior, preferences, and engagement.
    - Personal health trackers, like fitness bands and smartwatches, collect data on physical activity, heart rate, and sleep patterns.
    - Scientific experiments, such as those conducted in laboratories or large-scale projects like the Large Hadron Collider, produce vast amounts of data for analysis.
  - **Pros**
    - The data generated by machines is *highly structured*, meaning it is organized in a specific format, making it easier to analyze and process compared to unstructured data like text or images.
  - **Cons**
    - One challenge with machine-generated data is that it can be *difficult to move*. Due to its large volume, it often needs to be processed in-place or within a centralized system to avoid the complexities and costs associated with data transfer.
    - This data is typically *streaming*, meaning it is continuously generated and needs

to be processed in real-time, unlike batch processing where data is collected and processed at intervals. This requires specialized tools and infrastructure to handle the constant flow of information.

16 / 17: Sources of Big Data: People



Sources of Big Data: People

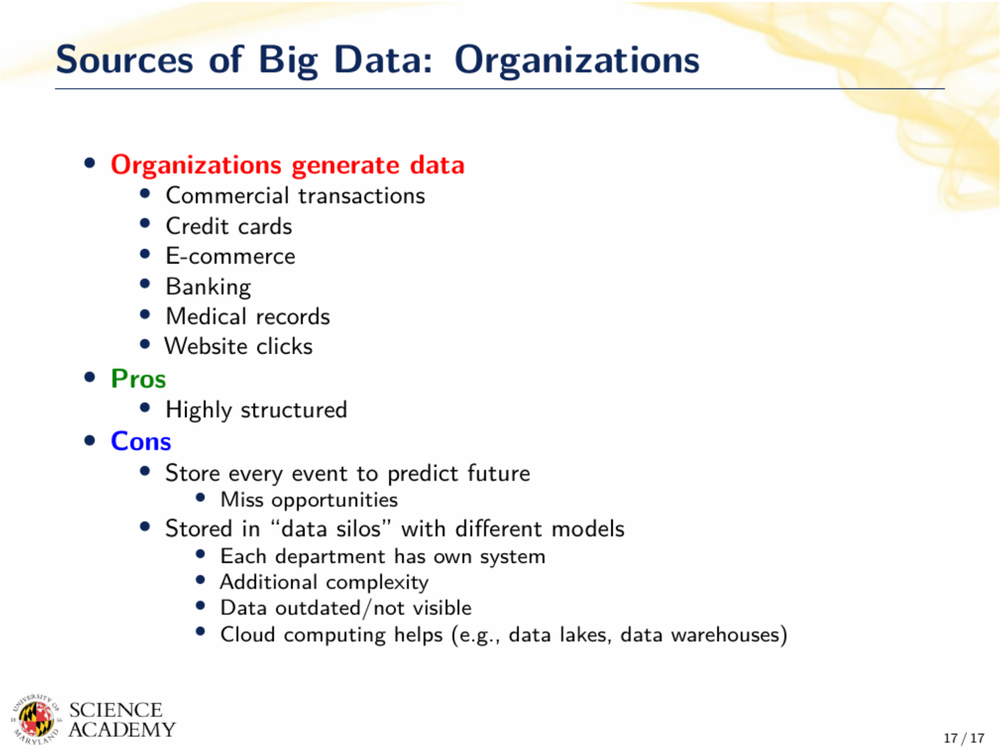
- **People and their activities generate data**
  - Social media (Instagram, Twitter, LinkedIn)
  - Video sharing (YouTube, TikTok)
  - Blogging, website comments
  - Internet searches
  - Text messages (SMS, Whatsapp, Signal, Telegram)
  - Personal documents (Google Docs, emails)
- **Pros**
  - Enable personalization
  - Valuable for business intelligence
- **Cons**
  - Semi-structured or unstructured data
    - Text, images, movies
  - Requires investment to extract value
    - Acquire → Store → Clean → Retrieve → Process → Insights
  - Surveillance capitalism

SCIENCE ACADEMY

16 / 17

- **People and their activities generate data**
  - Every day, people create a vast amount of data through their interactions on various platforms. **Social media** platforms like Instagram, Twitter, and LinkedIn are prime examples where users share thoughts, images, and videos, contributing to a massive pool of data. **Video sharing** sites such as YouTube and TikTok further add to this by hosting countless hours of video content. **Blogging and website comments** are other avenues where people express opinions and share information. **Internet searches** provide insights into what people are curious about or need information on. **Text messages** sent via SMS or apps like WhatsApp, Signal, and Telegram are another rich source of data. Lastly, **personal documents** such as Google Docs and emails contain a wealth of information about personal and professional communications.
- **Pros**
  - The data generated by people allows for **personalization**, meaning businesses can tailor experiences and products to individual preferences. This data is also **valuable for business intelligence**, helping companies understand market trends, customer behavior, and more.
- **Cons**
  - A significant challenge is that much of this data is **semi-structured or unstructured**, such as text, images, and videos, which are not easily organized into traditional databases. Extracting value from this data requires a substantial **investment** in processes to acquire, store, clean, retrieve, and process it to gain meaningful insights. Additionally, there is a concern about **surveillance capitalism**, where companies might exploit personal data for profit, raising privacy issues.

17 / 17: Sources of Big Data: Organizations



Sources of Big Data: Organizations

- Organizations generate data
  - Commercial transactions
  - Credit cards
  - E-commerce
  - Banking
  - Medical records
  - Website clicks
- Pros
  - Highly structured
- Cons
  - Store every event to predict future
    - Miss opportunities
  - Stored in “data silos” with different models
    - Each department has own system
    - Additional complexity
    - Data outdated/not visible
    - Cloud computing helps (e.g., data lakes, data warehouses)

SCIENCE ACADEMY

17 / 17

- **Sources of Big Data: Organizations**
  - Organizations are major producers of data, and they do this through various activities. **Commercial transactions** involve the exchange of goods and services, generating a lot of data. **Credit cards** and **e-commerce** platforms track purchases and user behavior. **Banking** institutions record financial transactions, while **medical records** contain patient information and treatment histories. **Website clicks** provide insights into user interactions and preferences.
- **Pros**
  - The data generated by organizations is often **highly structured**, meaning it is organized in a way that makes it easier to analyze and use. This structure helps in efficiently processing and extracting valuable insights.
- **Cons**
  - One downside is that organizations tend to **store every event** with the hope of predicting future trends, but this can lead to missed opportunities if not analyzed properly. Data is often kept in **“data silos,”** where each department has its own system and model. This creates **additional complexity** because data can become outdated or not easily accessible across the organization. However, **cloud computing** solutions like *data lakes* and *data warehouses* are helping to mitigate these issues by providing centralized storage and easier access to data.