Big Data and Security

Jeffrey Borowitz, PhD

Lecturer

Sam Nunn School of International Affairs

What is Big Data?

Big Data: What and Why?

What is big data to you?

- Organizing Questions:
  - Who uses big data?
  - What does big data do for them (as opposed to regular data)?
  - What are some other associated topics or buzzwords?

Why big data?

- Why would INTA want to offer this course?
- Why would you want to take it?

Big Data: What and Why?

What is big data?

- “Volume, velocity, variety”
- “Big Data” is also a catch all buzz word

Why big data?

- Why would INTA want to offer this course?
  - Familiarity with how technology and data analysis affect the world
  - Interesting things that might be done with data have strong INTA implications
  - Skills

What is Data?

Data is information encoded in a series of bits

Bits can take a value of 0 or 1

Data

- Bits can take a value of 0 or 1 Integers are encoded as a series of (typically 32) 0s and 1s
  - E.g. 00000000000000000000000000000010 is 2
- Text is also a series of 1s and 0s, but the schemes are more complicated
  - ASCII is a 7 bit scheme for Latin alphabet (upper and lower case), punctuation, and digits.
  - Unicode is a more complex, variable length scheme for characters in many languages, plus emoji-like characters
  - E.g. a word like “Jeffrey” has 7 characters, at 7 bits each in ASCII, or 49 bits

Volume

- How much data is there?
    - 12 zetabytes (1.2 × 10^{22})

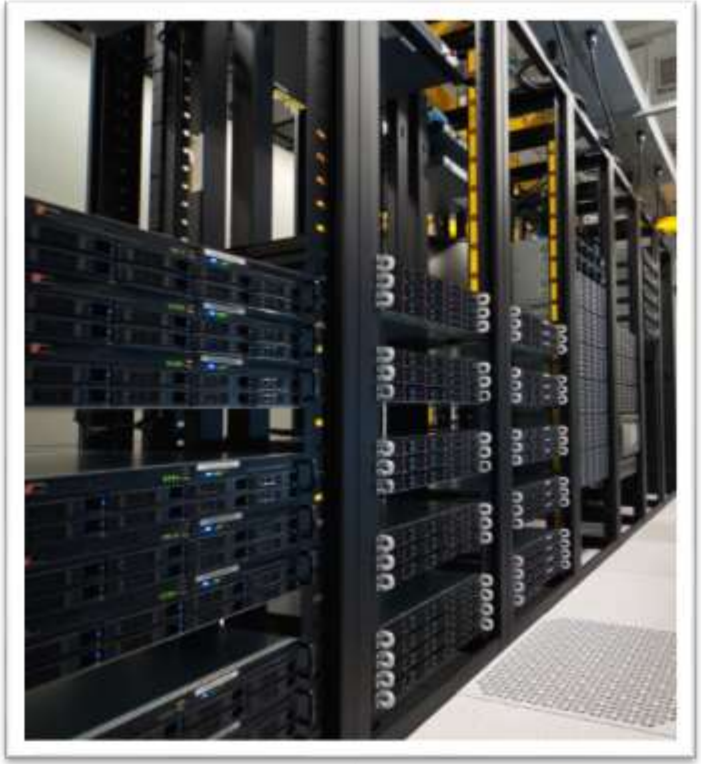
- Data is (was?) growing at an increasing rate:
    - 90% of data was created in the last 2 years

What Are The Units of Data?

- 1 bit is one piece of binary information - a one or a zero.
- 1 byte is a set of 8 bits, which can be one of 256 (2^8) combinations.
- 10^4 bytes (10 kilobytes) is about a couple pages of text
- 10^6 bytes (1 megabyte) is about the size of 1 minute of compressed music
- 10^9 bytes (1 gigabyte) is roughly a compressed but decent video
- 3.2×10^9 bytes (3.2 gigabytes) is the amount of data in your DNA.
- All text on Wikipedia is about 9.5 gigabytes
- 10^{12} bytes (1 terabyte) is about the size of an external hard drive (in 2014)
- 3×10^{12} bytes (3 terabytes) is the approximate amount of storage in your brain.

A Server Rack

- 4 × 10^{14} bytes (400 terabytes) is about as much data as can be housed in a server rack. This is also about the amount of data in all books ever written.



A Data Center

- 10^{18} bytes (1 exabyte) is how much information e.g. Google could store in a data center.
- 5×10^{18} bytes (5 exabytes) would be the size of all words ever spoken, if transcribed.

Georgia Tech

What are important changes in variety of data?

- What aspects of your life were recorded in 1980?
  - Plane tickets
  - Taxes
  - Interactions with the largest companies
- What aspects of your life are recorded in 2020?
  - Location (from cell phones)
  - Attitudes on social media (from posts, likes, etc.)
    - Note: for some people, “big data” is nearly synonymous with social media/internet technologies
  - Workflow and interactions (from web email, search logs)
  - Some of your speech? (from e.g. Alexa)
- By combining different types of information, what is possible?

Changes in Velocity?

- The example of Twitter
  - People create a tweet
  - Twitter saves it
  - Twitter sends it wherever it needs to go: mobile devices, web browsers, etc.
  - What happens if Twitter gets another tweet before it’s done?
- Some web related technologies need to keep up with data in real time, even as it comes faster
  - Is a web request part of a denial of service attack?
  - Is an email message spam?
  - What if Healthcare.gov needs to serve 10 million people in 1 day?
  - Outside of web, threat monitoring software is like this. . .

What is Big Data?

- For the course, let’s think about Big Data as something broader than just data size/type

- Our “Big Data” will include things that big data might enable or portend
  - Power to collect (and lose) personal info
  - Power to predict
  - “Artificial Intelligence”
  - Driverless cars

What Concepts Underlie Big Data?

Computing

Statistics

Applications