CSCI 1070: Taming Big Data

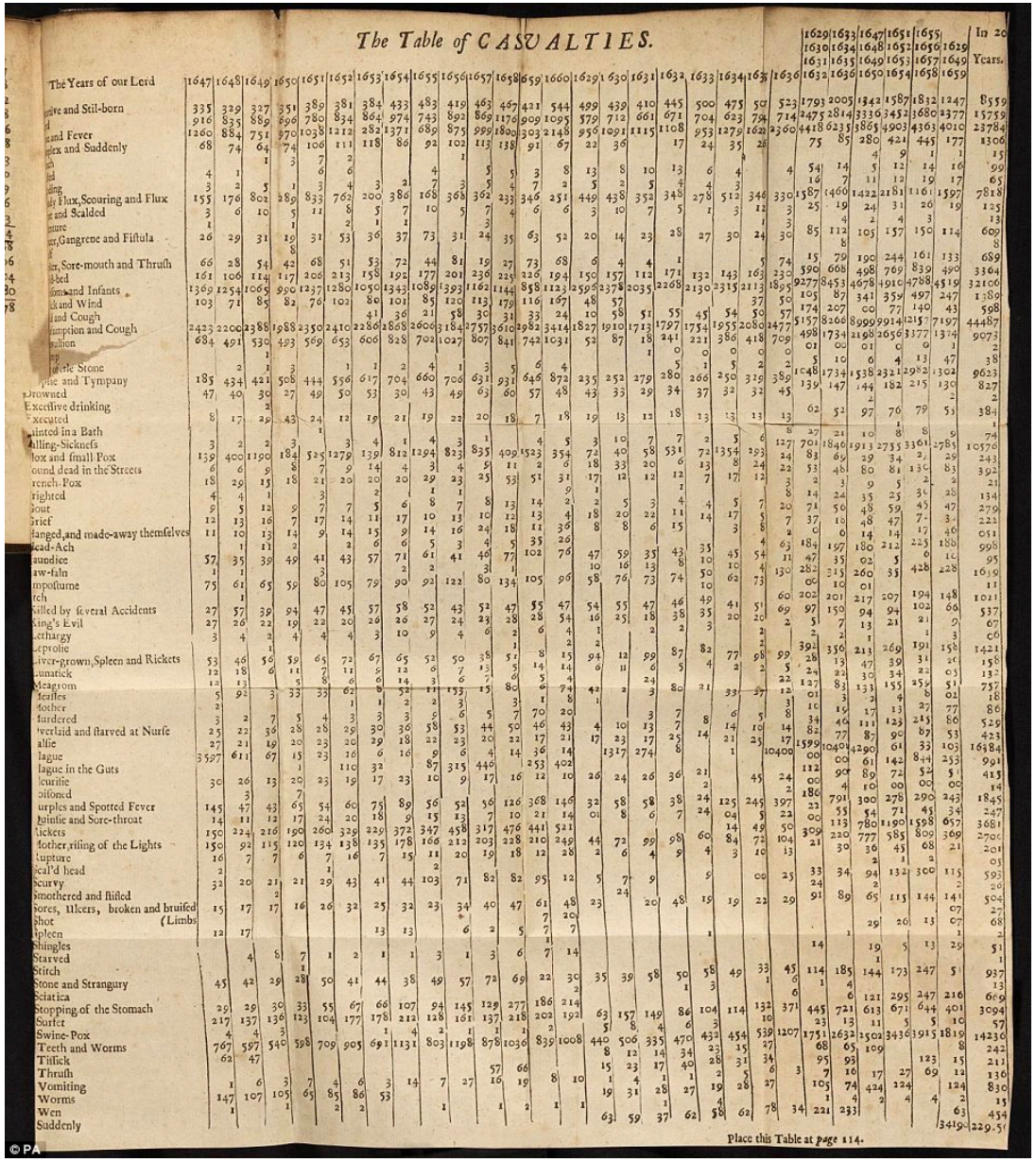
Prof. Abby Stylianou

Logistics

- Are you on the waitlist? Talk to me after class!
- Class meetings: Tu/Th, 11AM–12:15PM, Ritter 115
  - Lecture, discussions, in class coding
  - Occasional quizzes/journaling about readings
  - Attendance not mandatory, but you're gonna have a bad time if you don't show up.
- Office Hours: Wed, 1:30-2:30PM, Ritter 107
- Class communications: Piazza! *(not email)*
- Website/schedule/materials: https://cs.slu.edu/~stylianou/1070/
- Technologies: Python, Jupyter Notebooks, Git

What is big data?

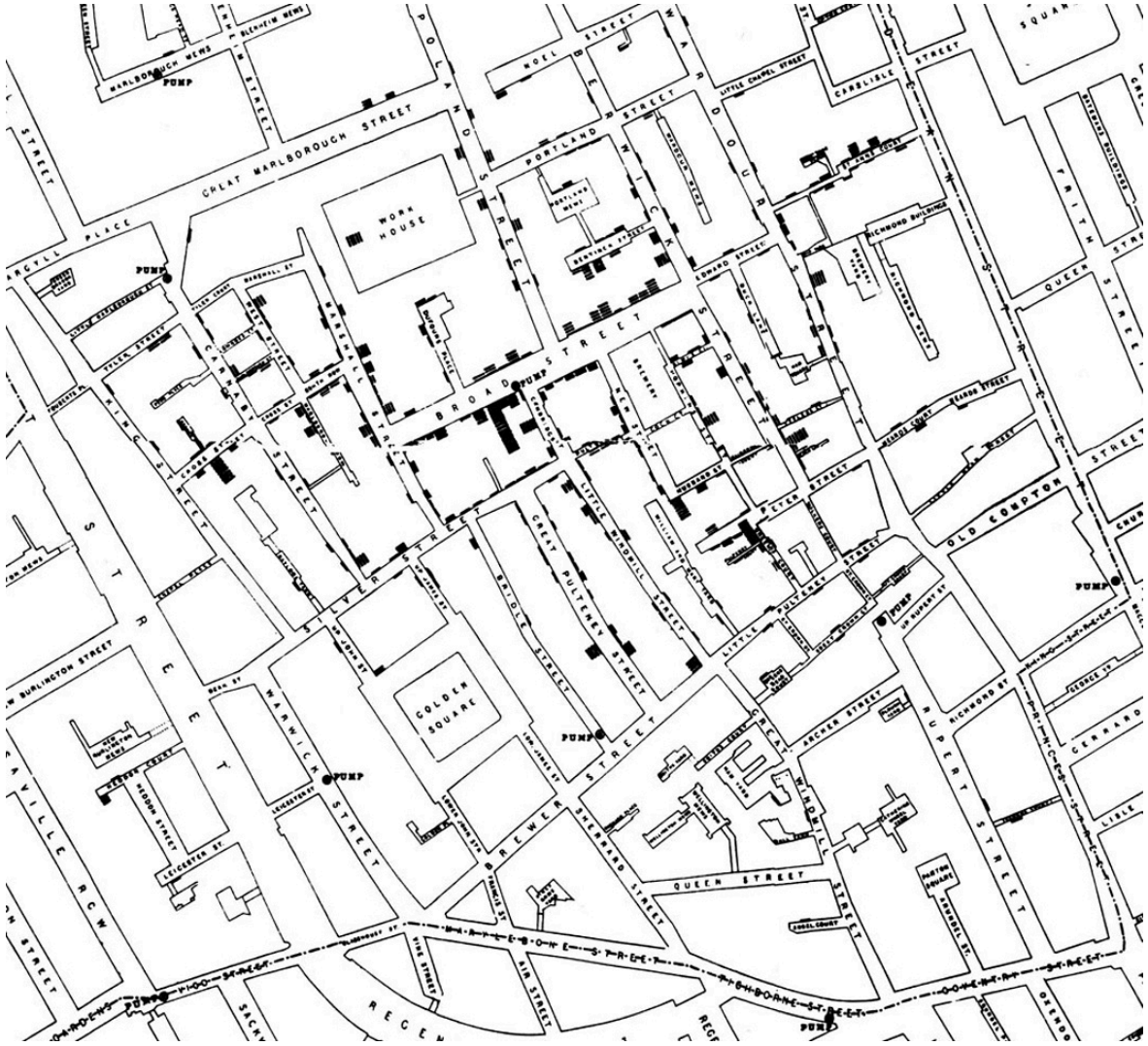
- **1663**: John Graunt is the first person credited w/ statistical data analysis in his studies of the bubonic plague in Europe, dealing with what he referred to as “overwhelming amounts of information”



https://www.dataversity.net/brief-history-big-data/

What is big data?

- **1854**: John Snow maps London Cholera outbreaks and finds that they are clustered around a single pump; it was found that a cesspit was leaking into that pump



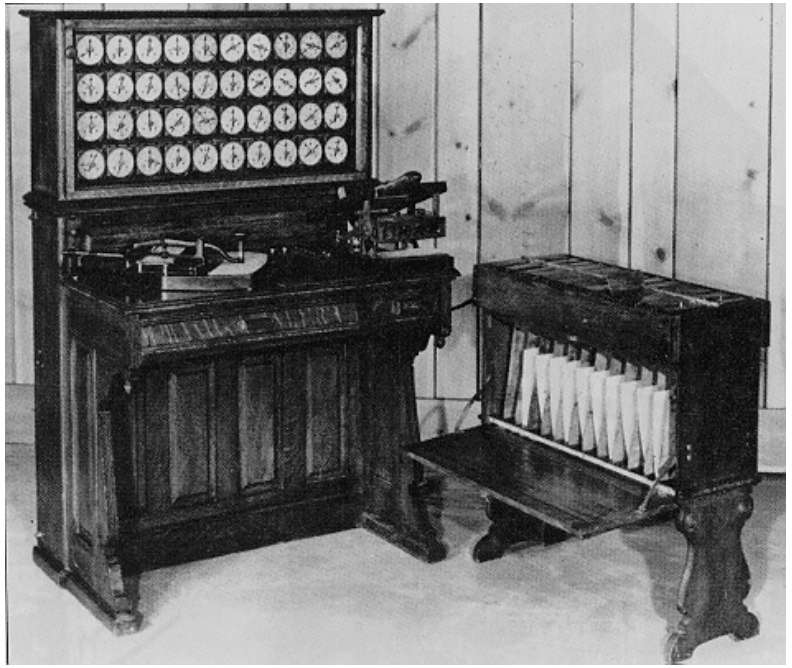
https://www.inferentialthinking.com/chapters/02/1/observation-and-visualization-john-snow-and-the-broad-street-pump

What is big data?

- **1880**: US Census Bureau estimates it will take eight years to process the data collected in the 1880 census, and over 10 years to process 1890 census data

- **Hollerith Tabulating Machine**

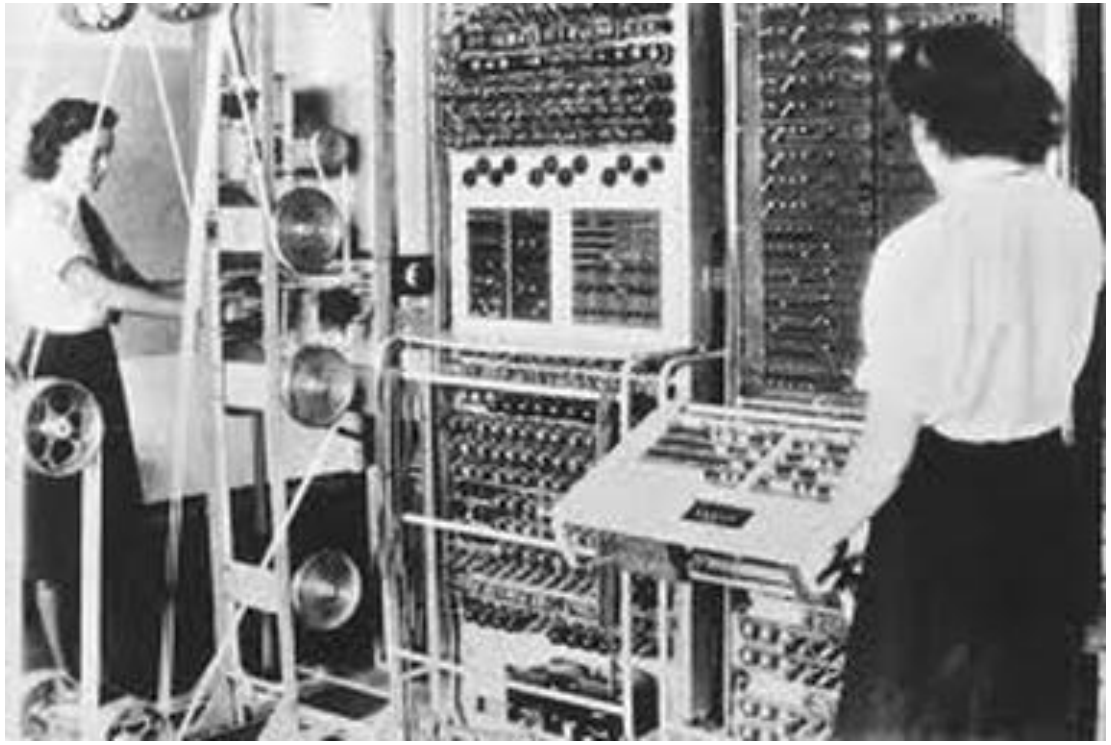
(punch card tabulation) reduces to ~3 months



https://www.dataversity.net/brief-history-big-data/

What is big data?

- **WW2**: British invent the Colossus machine to scan for patterns in intercepted Nazi codes. Scans 5,000 characters a second, reducing workload from weeks to hours



https://www.dataversity.net/brief-history-big-data/

What is big data?

- 1944: Librarian Fremont Rider @ Wesleyan estimates American university libraries doubling in size every 16 years:

“the Yale Library in 2040 will have approximately 200,000,000 volumes,
which will occupy over 6,000 miles of shelves…
[requiring] a cataloging staff of over six thousand persons”

https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#1fed7fb865a1

What is big data?

- 1961: Derek Price shows # of new scientific journals growing exponentially rather than linearly

"each advance generates a new series of advances at a reasonably constant birth rate, so that the number of births is strictly proportional to the size of the population of discoveries at any given time"

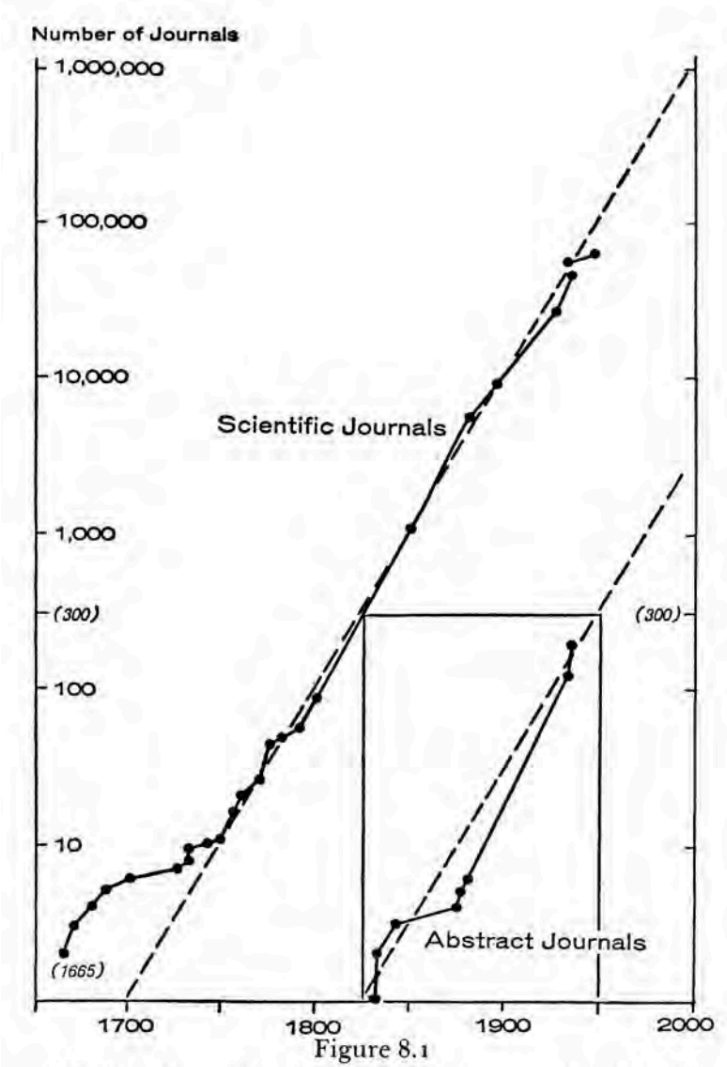


Figure 8.1
Number of journals founded (*not* surviving) as a function of date. The two uppermost points are taken from a slightly differently based list.

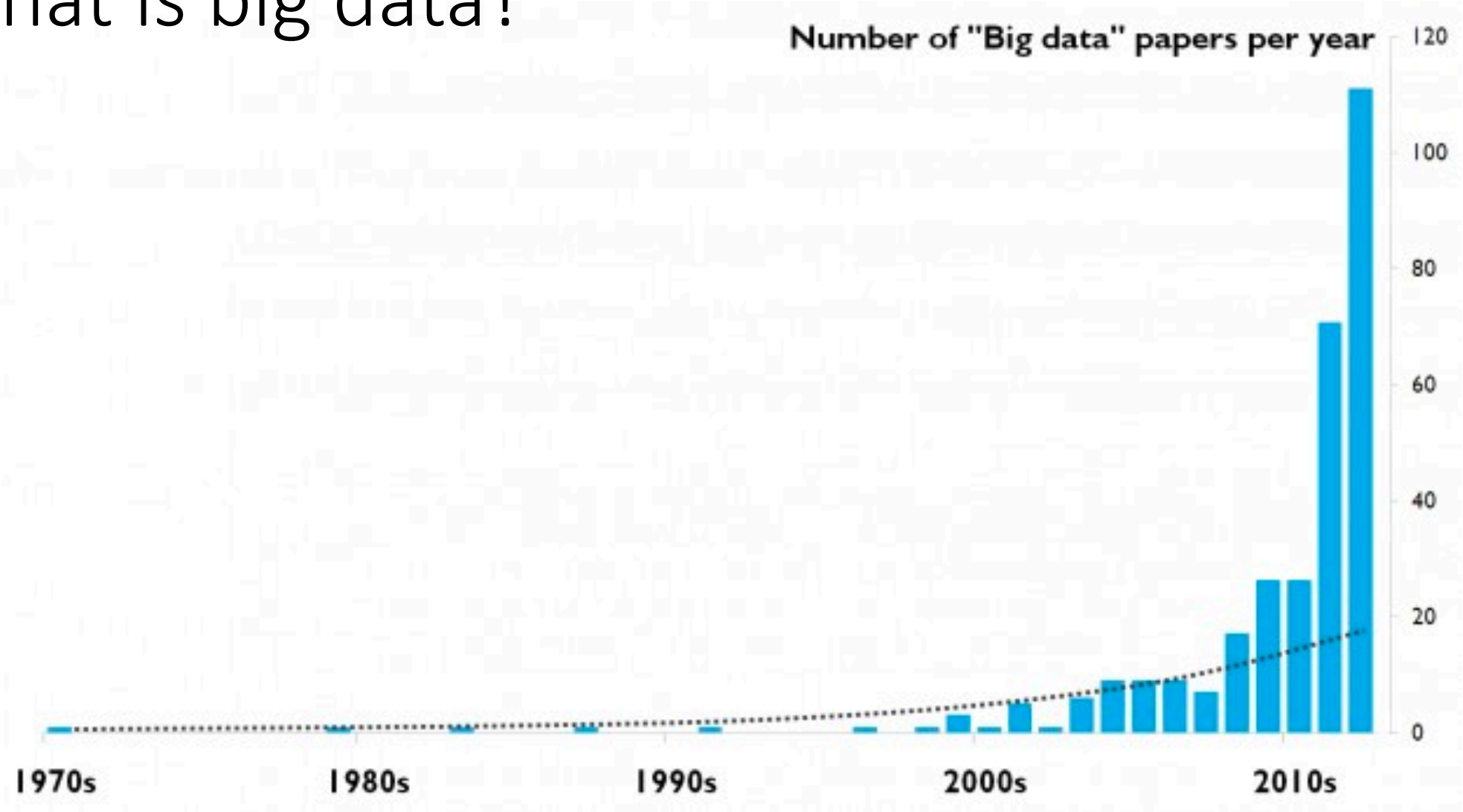
https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#1fed7fb865a1

What is big data?

- **1969**: ARPANET kicks off the Internet
- **1977**: First personal computers come on the market
- **1989:** Tim Berners-Lee introduces the concept of the World Wide Web and the underlying protocols that support it (HTML, URL, HTTP)
- **1993:** CERN announces WWW will be free for everyone to develop and use
- **1990s– 2000s:** The explosion of the internet
- **2005:** Roger Mougalas coins term ‘Big Data’, referring to the scale of data that is nearly impossible to manage and process w/ available tools

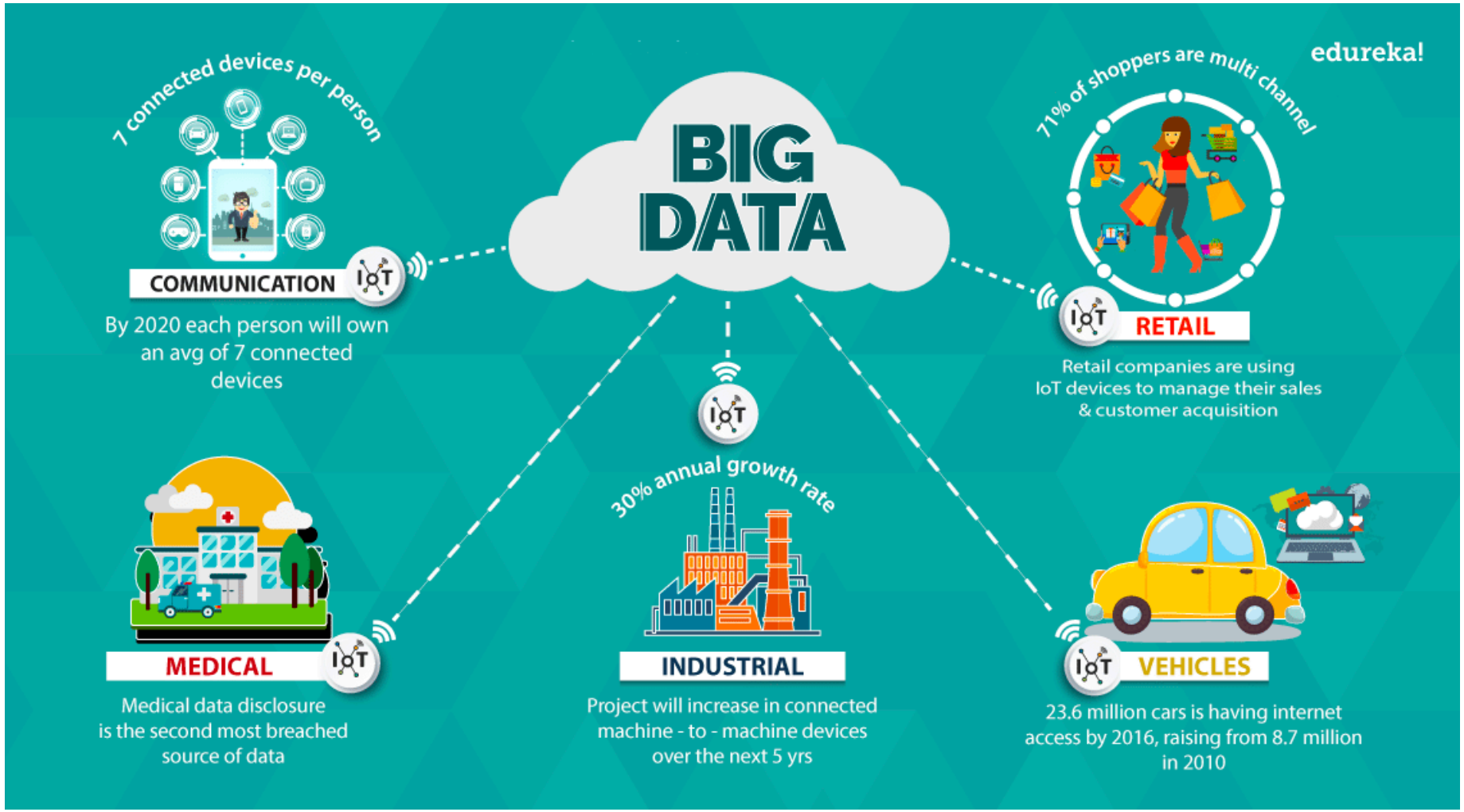
https://www.dataversity.net/brief-history-big-data/

What is big data?



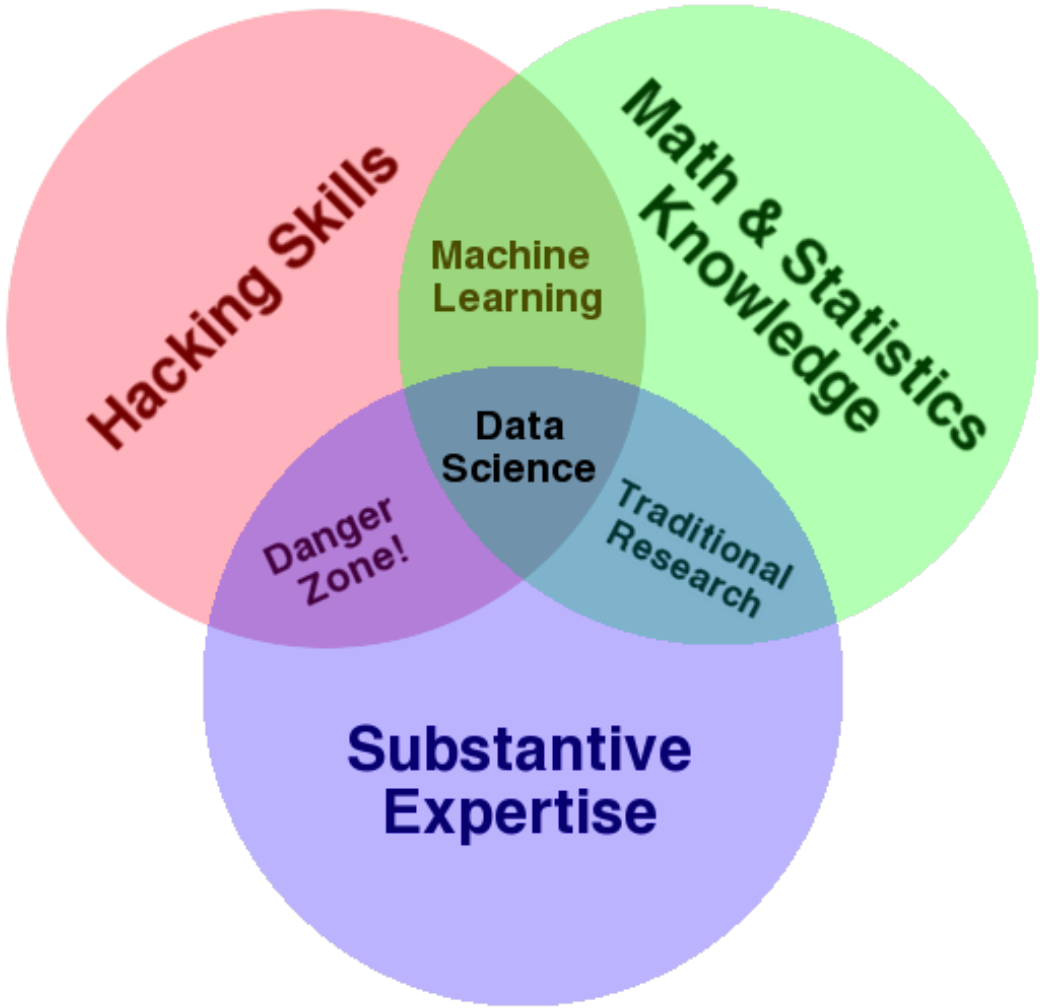
https://www.researchtrends.com/issue-30-september-2012/the-evolution-of-big-data-as-a-research-and-scientific-topic-overview-of-the-literature/

Where is big data?



https://www.edureka.co/blog/big-data-applications-revolutionizing-various-domains/

What is data science?



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

What's the plan for this class?

- Learn the data science tools to collect, clean, understand, manipulate big data

- Exploration, Inference and Prediction
  - Exploration: Identify patterns in information
    - Tools: Visualization + Descriptive Statistics
  - Prediction: make informed guesses about what we want to know, based on the patterns that we identified
    - Tools: Machine Learning + Optimization
  - Inference: quantify our certainty about our predictions
    - Tools: Statisticals Tests + Models

Login to Systems

- Class website: https://cs.slu.edu/~stylianou/1070

- JupyterHub: http://cs1070.com
  - Login: Your SLUNetID
  - Password: Set it the first time you login

- Piazza: https://piazza.com/class/jzbep3yp13k3pr
  - Should have gotten an email to active – let me know if you didn't!

- GitLab: http://git.cs.slu.edu/courses/fall19/csci_1070/<your-SLUNetID>
  - Login: SLUNetID
  - Password:
    - either you've done this already and should know it
    - or do a password reset the first time

Demo

Next class

- No class on Thursday 8/29 (Mass of the Holy Spirit)

- For next Tuesday, 9/3:
  - Assignment 0 (‘Getting to Know You’) due via git at beginning of class
  - Do readings that are listed for 8/27 and 8/29
  - We will discuss at the beginning of class