CS 5683: Big Data Analytics

Fall 2023

Arunkumar Bagavathi

Department of Computer Science

Oklahoma State university

Little bit about myself

- Ph.D. Computer Science at **University of North Carolina at Charlotte**
- Assistant Professor in CS department at OSU
- **Research interests:**
  - **Data mining**
  - **Network science**
  - **Natural Language Processing**
  - **Applied machine learning**

Research problems:

- **Heterogeneous graphs**
- **Text + Network representation learning**
- **Studying online media polarization**
- **Hate speech and misinformation**
- **Disease diagnostics on animals**

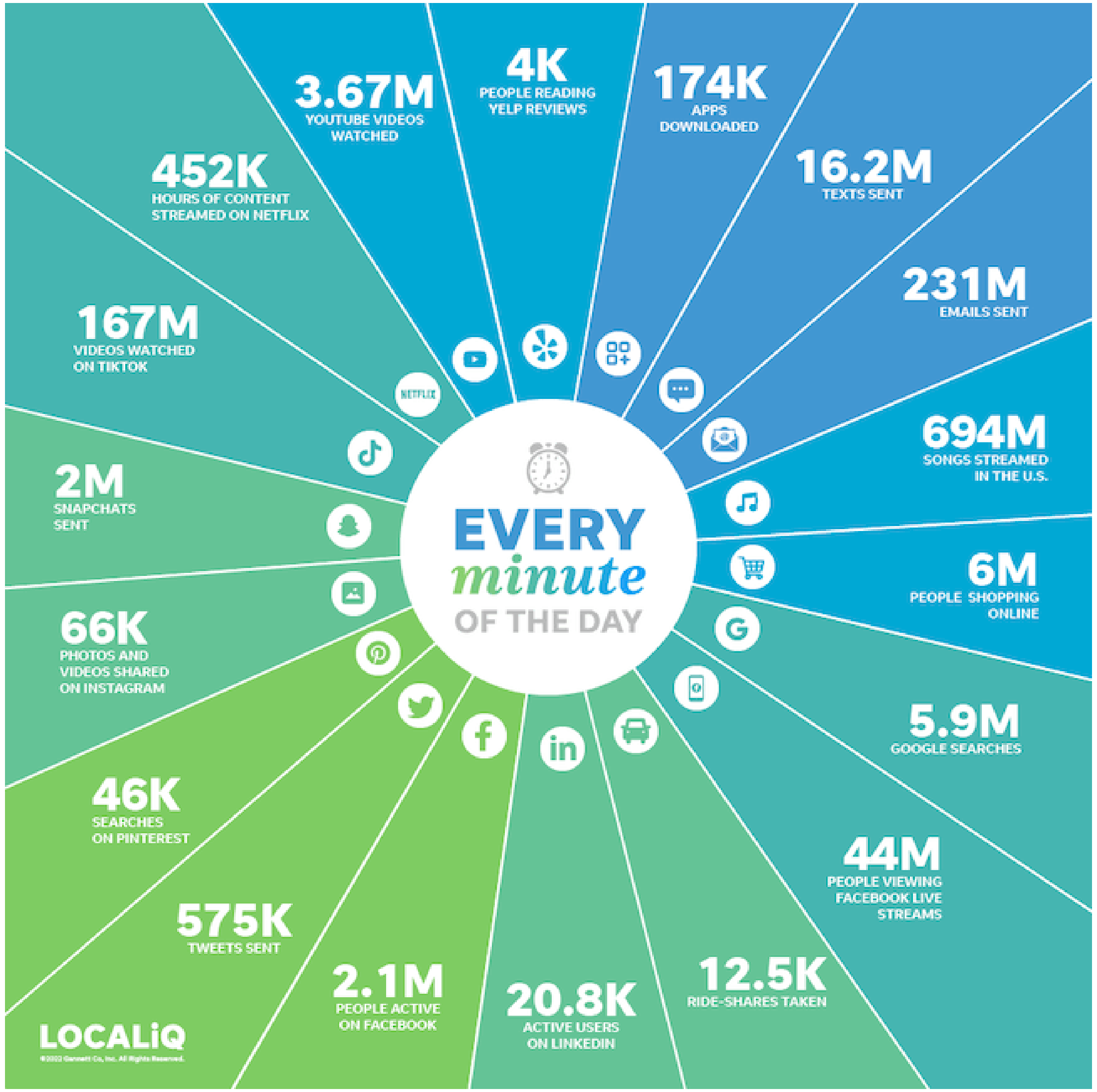
- ***Office:*** MSCS 215
- ***Office hours:*** *Tuesdays 10:00am-1:00pm*

WHAT IS BIG DATA?

Generally, talk about 5Vs ?

What happens on the internet in 1 minute in 2023?

Note: This is the stats taken in May 2022



Estimated amount of data in 2020:

70 trillion GB (70 zettabytes)

40zettabytes in 2020

Source: EMC

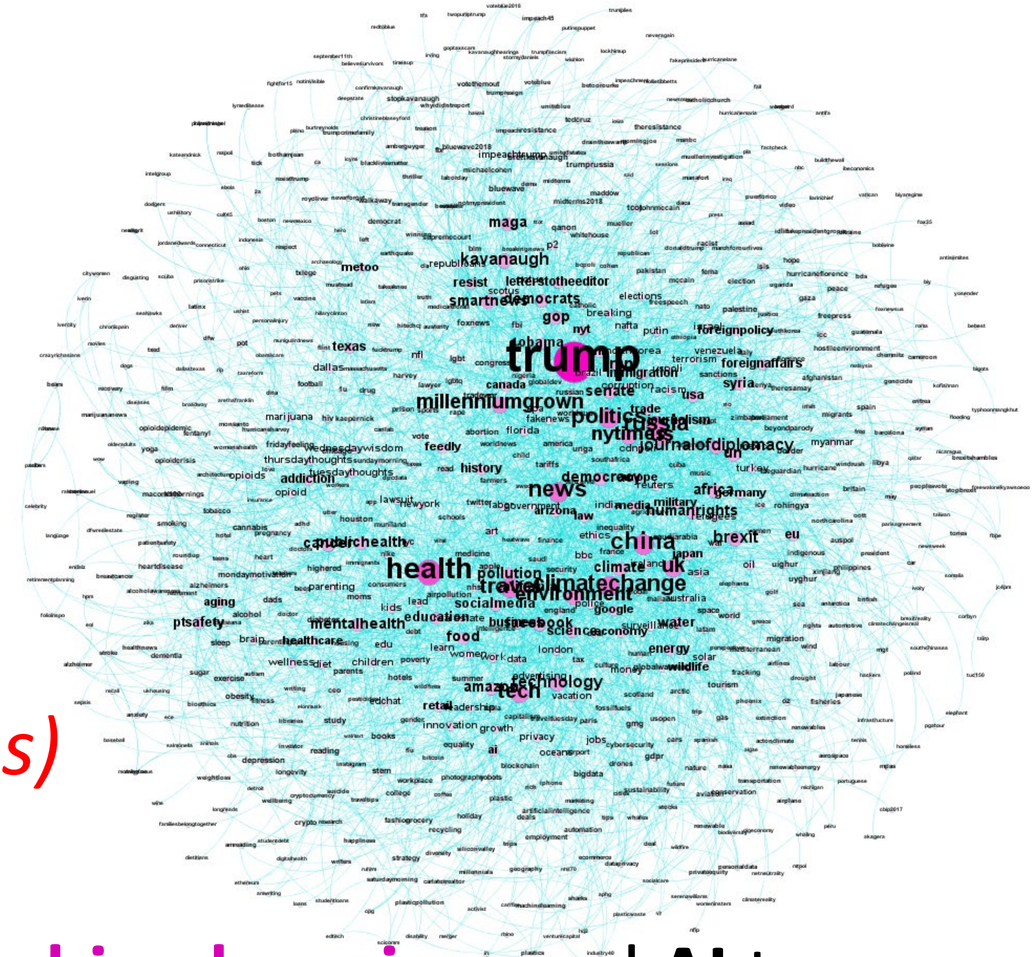
1 zettabytes = 10^{21} bytes

How the internet is not breaking during the pandemic???

What is the use of such data?

- Data contains valuable ***knowledge***
- Data needs to be
  - **Stored**
  - **Managed**
  - **Analyzed**
  - **Interpreted**

(This Class)



- Analysis can be done with statistics, machine learning, and **AI** to extract knowledge

What is Data Science?

- Given: big data or data that is computationally challenging!

- **Discover patterns and models that:**
  - **Useful:** should handle new data
  - **Valid:** should promise some degree of certainty
  - **Unexpected:** non-obvious to humans and existing systems
  - **Understandable:** interpretable by humans

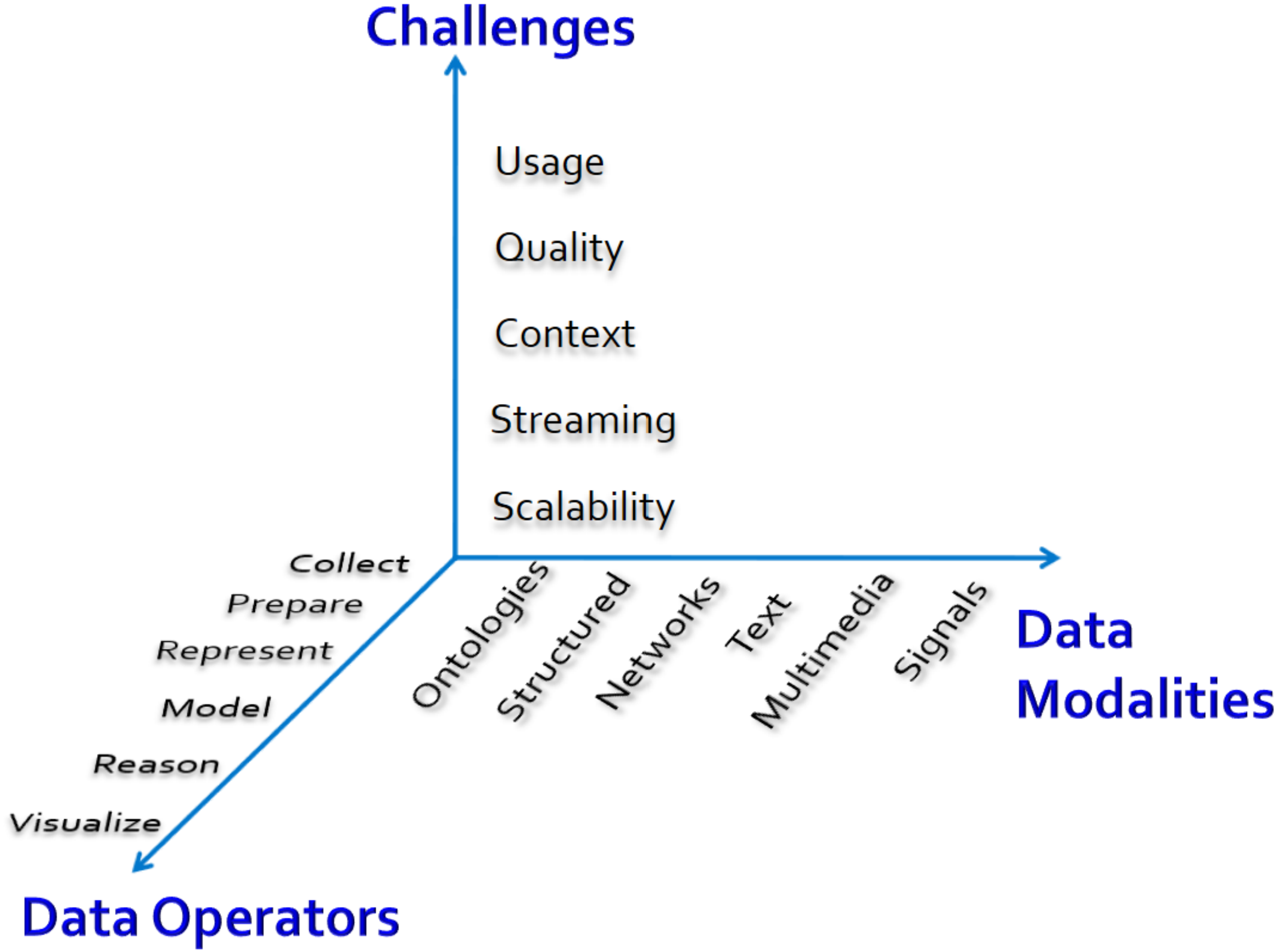
Data Science Tasks

- **Descriptive tasks**
  - Find human interpretable patterns that describe the data
  - **Example:** Clustering, Visualizations

- **Predictive tasks**
  - Use some variables to predict unknown or forecast future values of other variables
  - **Example:** Classification – Recommender systems

- **Forecasting tasks**
  - A type of predictive task but with temporal constraints
  - **Example:** Weather forecasting with a stream of data

What to expect when working with Data?

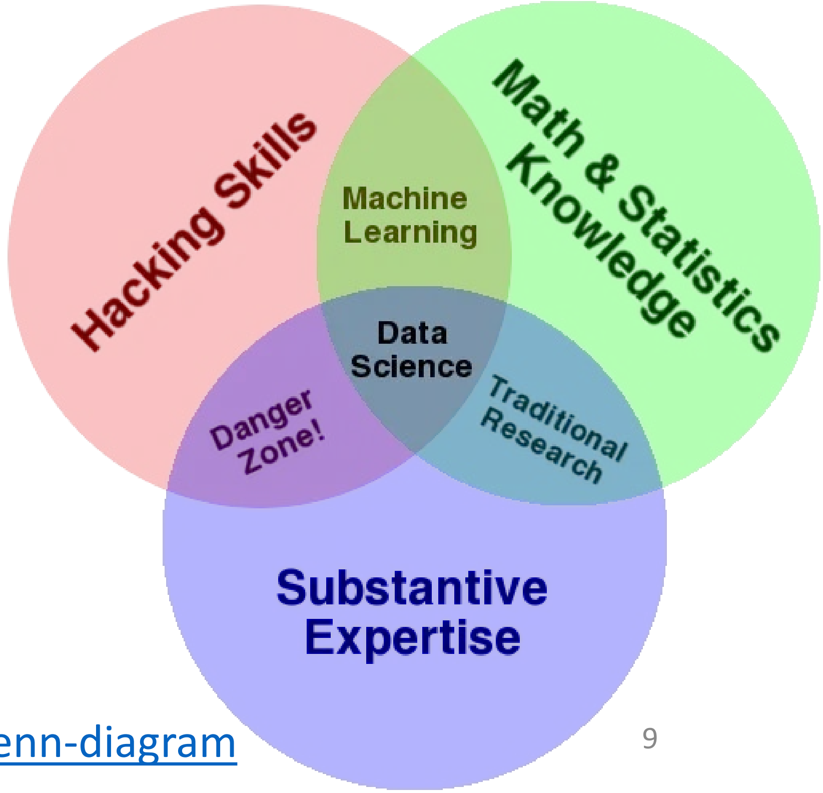


Data Science Cultures

- **Data science overlaps with:**
  - **Traditional (CS) Research**
  - **Domain expertise**
  - **Machine Learning**

- **Multiple cultures:**
  - To a DB person, data science is a query answer
  - To a ML person, data science is identifying model parameters
  - To a non-technical person, data science is visualizing data patterns

- **In this class we will try to cover all cultures!**



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

CS 5683

- **CS 5683** overlaps with machine learning, statistics, databases, and predictive analytics. But more emphasis on
    - *Automation in handling big data*
    - *Algorithms (mostly with ML)*
    - *Multiple data modalities*
    - *Application driven learning*
    - *Scalability*

What will we learn?

- We will learn to mine multiple modals of data:
  - Data is a graph/network
  - Data is never ending
  - Data is text
  - Data is (un)labeled
- We will learn to use multiple models of computations:
  - Dimensionality reduction
  - Clustering algorithms
  - Streaming and Online algorithms
  - Text & Graph mining algorithms
  - PySpark

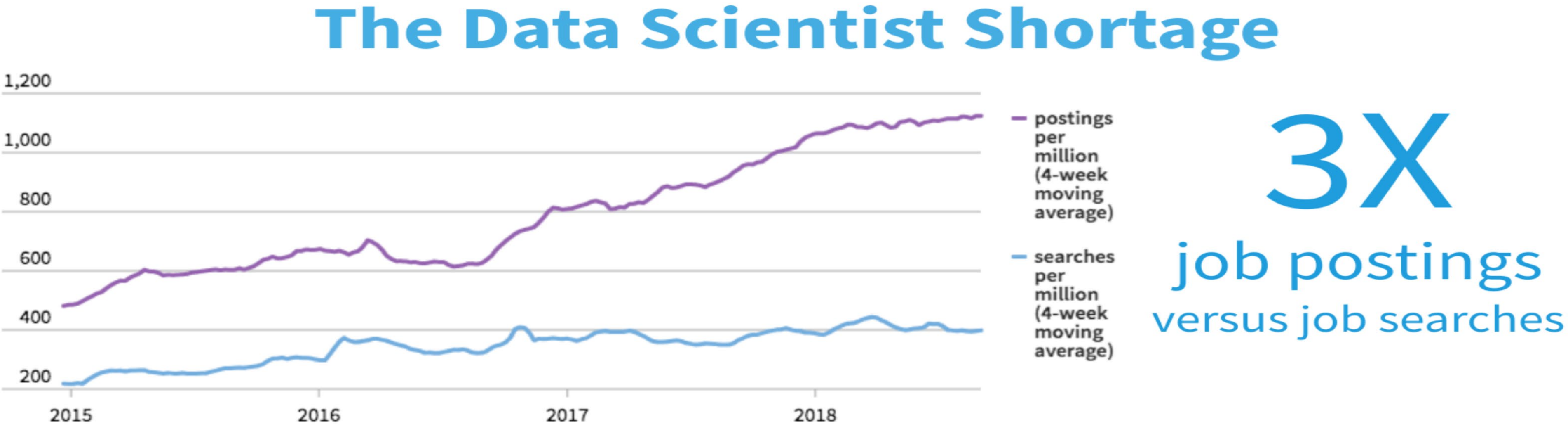
What will we learn?

- **We will learn to solve real-world problems:**
  - **Recommender systems**
  - **Web search**
  - **Social networks**

- **We will learn multiple methods:**
  - **Linear Algebra**
  - **Optimization**
  - **Dynamic programming**
  - **Representation learning**

Why to learn Data Mining?

- **Data Engineer** and **Data Scientist** are one of top 10 wanted jobs in industries in 2019 - https://www.businessinsider.com/best-jobs-in-america-2019-1
- U.S. Bureau of Labor Statistics projects the employment of data scientists to grow 36% from 2021 to 2031. The average growth of any occupation is around 5%
- https://www.bls.gov/ooh/math/data-scientists.htm
- https://365datascience.com/career-advice/data-scientist-job-outlook/#3



Benefits of CS 5683

- Prepare students to understand **big data representations** for ML algorithms
- Prepare students to tackle **real-world massive text, graph, and streaming data** for AI systems
- Prepare students to get **expertise beyond the classic ML problems** like classification and clustering
- First priority for CS 5683 students on any **funding opportunities**
- First priority for CS 5683 students to support **independent projects, thesis, and job recommendations**

Course Overview

Teaching Assistant

TBA

TA office hours: TBA

Course Activities and Grading

For all students:

- Quizzes – 20%
- Assignments – 50%
- Class participation – 10%
- Final exam – 20%

Course Logistics

- This course will discuss several data science topics for big data: dimensionality reduction, clustering, recommender systems, large-scale text mining, big graph mining, data stream analysis, and tools for big data analytics
- We give hands-on experience for big data frameworks and algorithms with assignments and projects
- All assignments and projects will be on **Python** programming

- May not be used for degree credit with **MSIS 5683**

Course Communication

- **Canvas Discussion board(s)**
  - Post in the appropriate discussion board
  - We will have a discussion board for each topic and assignment

- **Email us!**
  - Expect our reply anytime within 6 hrs

- **Canvas Announcements**
  - For general announcements
  - Make sure you turn ON Canvas email notifications

Assignments & Quizzes

- **4 Quizzes**
  - Online quizzes (open web)
  - 1 Quiz per month
  - We plan to schedule the quiz on the first week of each month
  - Quiz syllabus will be revealed 1 week before the Quiz

- **4 – 5 assignments**
  - Programming assignments
  - Involves significant amount of work
  - **Start early!**
  - Student will have 2 – 3 weeks to complete and submit the assignment on Canvas

Quiz-1 on September 8

Assignment & Projects Submission

- All submissions in **Canvas**

- **Assignments late submissions:**
  - 2 grace periods
  - **We will not accept any submissions 1 week after the due date!**

- **Regrading:**
  - Email us your request **within 1 week after posting grades**!
  - Do not request to regrade first assignment/project at the end of the semester

Assignment and Project Expectations

Novelty.

https://academicaffairs.okstate.edu/site-files/documents/ai-handbook-faculty-personnel.pdf

Academic Integrity

- Please review in the course syllabus
- http://academicintegrity.okstate.edu
- https://adminfinance.okstate.edu/site-files/documents/policies/academic-integrity-policy.pdf
- In short: If you are not submitting **your work, you are cheating**

- Consequences: Grade of ‘0’ or ‘F!’ or may even expel from the university

Diverse Technical Preparation

- If your programming background in rusty, **prepare in the initial weeks**
- **Please do not ask for perfect training environment** – the lecturer does not provide perfect tutorials to learn the technologies used in the course
- **You will encounter with:**
  - ✓ Unclean data, unclear instructions, inaccurate documentation, etc.
  - ✓ Start early to handle such issues
- **Please do not ask complex questions near the submission deadline!**

Other University Services and Policies

- Please check OSU syllabus attachment pdf

What's After CS 5683?

- **CS 5123** – **Cloud Computing and Distributed Systems** ***(Spring)***

- **Independent study** (3 credit hours)
  - Better option if you want to do a thesis
  - Or explore research topics
  - **Talk with the instructor**

- **Volunteer in our lab for research!**
  - **Talk with the instructor**

What’s next?

Google Colab and

Refreshing on Python Basics

Review on text and graph data featurization

Questions???

Acknowledgements

- Some contents of these slides are motivated by materials collected from:
    - Dr. Srinivas Akella – UNC Charlotte
    - Dr. Jure Leskovec – Stanford University (http://www.mmds.org/)