CS685/785 Foundation of Data Science

Lecture 1: Introduction

Xi Li

Fall 2024

Outline for Today

- Syllabus Overview
- Introduction to Data Science
  - What is data science?
  - Where is data science needed?
  - What is covered in this course?

Course Information

- Instructor: Dr. Xi Li (xli7@uab.edu)
    - Education:

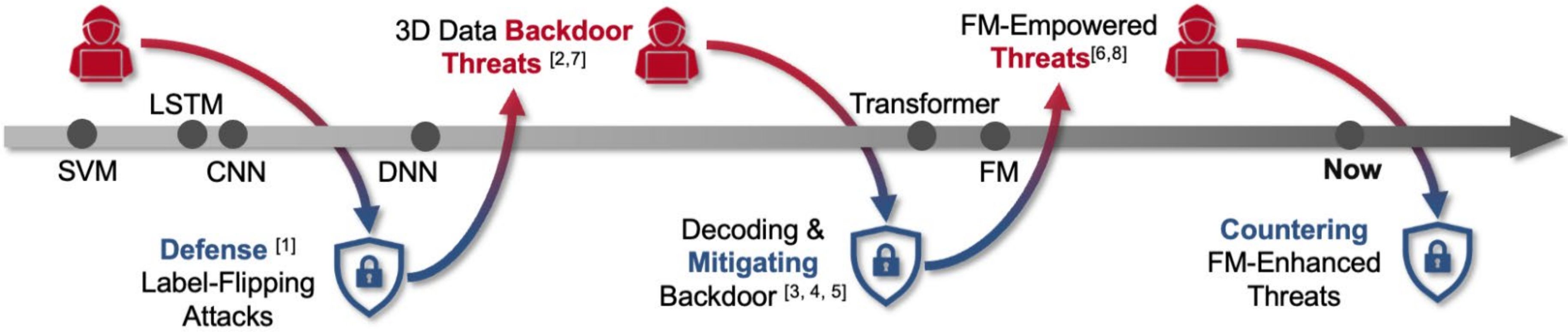
BE of Electrical Engineering,
Southeast University (China)

MS of Computer Science
and Engineering
Pennsylvania State University

PhD of Computer Science
and Engineering
Pennsylvania State University

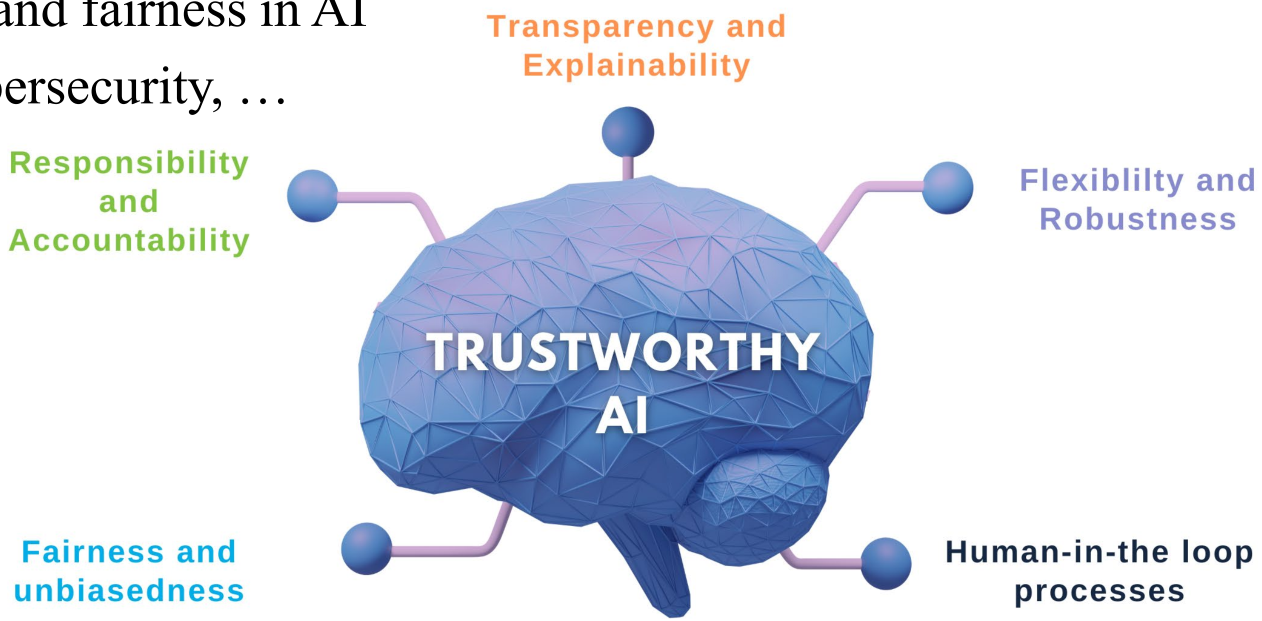
Course Information

- Instructor: Dr. Xi Li (xli7@uab.edu)
  - Research Interest:
    - ✓ Trustworthy AI: security*, privacy, and fairness in AI
    - ✓ Trustworthy AI + X: healthcare, cybersecurity, …



Course Information

- Instructor: Dr. Xi Li (xli7@uab.edu)
  - Research Interest:
    - ✓ Trustworthy AI: security*, privacy, and fairness in AI
    - ✓ Trustworthy AI + X: healthcare, cybersecurity, …


Transparency and Explainability
Responsibility and Accountability
Flexiblilty and Robustness
TRUSTWORTHY AI
Fairness and unbiasedness
Human-in-the loop processes

Course Information

- Instructor: Dr. Xi Li (xli7@uab.edu)
    - ○ Office Hours:
        - ✓ 10:00 am – 12:00 pm, Monday
        - ✓ UH4151or Zoom: https://uab.zoom.us/my/xiliuab

Course Information

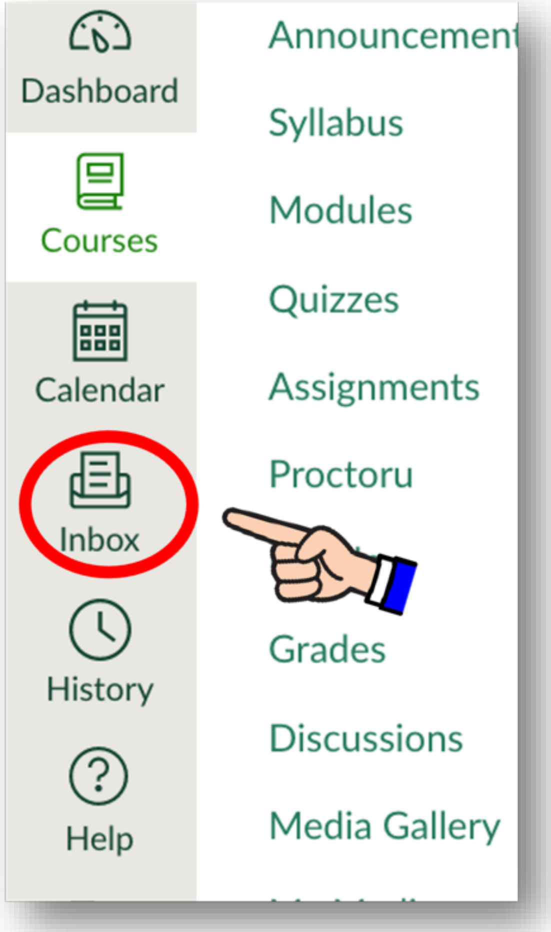
- Time and Location: TuTh 9:30am - 10:45am @ UH2100
- Teaching Assistants:
  - Ashraful Islam (aislam@uab.edu)

    Office hours: 2:00 pm - 4:00 pm, Wednesday @ UH1009
  - Mengchen Fan (fanm@uab.edu)

    Office hours: 9:00 am - 11:00 am, Friday @ UH1009

Course Information

- Preferred Methods of Communication
  - If you have questions, please use “Inbox on Canvas” to email the instructor and the TA.
  - Please expect a response within 24 hours on weekdays and a slower response on weekends



Course Overview

- This course explores essential concepts and techniques in statistical inference and big data analytics. It provides basic concepts and theories in data science, as well as hands-on experience with various data science Python libraries. We will also read research papers and get to know the state-of-the-arts in this domain.
- Textbook**:**
  - Foundation of Data Science (2018), Avrim Blum, John Hopcroft and Ravindran Kannan
  - Online: https://www.cs.cornell.edu/jeh/book.pdf
- Prerequisites
  - Basic concepts of linear algebra, calculus, probability theory, and programming skills in Python.

Grading

- Grading Policy
  - Assignments: 30%
  - Group Project: 30%
  - Final Exam: 20%
  - Paper Review: 10%
  - Participation: 10%
- Grading Scale

Points (0-100)	0-59	60-69	70-79	80-89	≥90
Letter Grade	F	C		B	A

Course Activities

- Assignments
- Group Project
- Final Exam
- Paper Review
- Participation

Course Activities

- Assignments
  - 30%
  - Individual work
  - Master students: 3-4 coding assignments using Python, Pandas, NumPy, and Scikit-Learn.
  - PhD students: 3-4 coding assignments AND 3-4 written assignments.
- Group Project
- Final Exam
- Paper Review
- Participation

Course Activities

- Assignments
- Group Project
  - 30%
  - Teamwork: up to 5 members for master students and up to 3 members for PhD students
  - Report: 4-page report for master students and 8-page report for PhD students
  - Presentation for PhD students
- Final Exam
- Paper Review
- Participation

Course Activities

- Assignments
- Group Project
- Final Exam
  - 20%
  - Covers all topics discussed throughout the semester
  - Closed-book
  - Date and location: TBD
- Paper Review
- Participation

Course Activities

- Assignments
- Group Project
- Final Exam
- Paper Review
    - 10%
    - Assess a research paper's contributions, methodology, and results
- Participation

Course Activities

- Assignments
- Group Project
- Final Exam
- Paper Review
- Participation
  - 10%
  - Quizzes: multiple-choice questions on Canvas

Bonus Points

Bonus points (up to 10 points) are available for:

- Master students who complete either a group project presentation or a paper presentation.
- PhD students that complete paper presentations.

Policies

- Attendance
- Late Penalties
- Academic Integrity

Policies

- Attendance
  - Students must attend all classes and stay for the full session.
  - Three or more unexcused absences or instances of arriving late / leaving early (20 minutes or more), may result in the failure of the course.
- Late Penalties
- Academic Integrity

Policies

- Attendance
- Late Penalties
  - 33% reduction per day after the deadline (including weekends and holidays)
  - Assignments submitted more than three days after the deadline will not be graded.
- Academic Integrity

Policies

- Attendance
- Late Penalties
- Academic Integrity
  - First violation: 0 grade for relevant work (assignment, homework, exam, or project).
  - Second violation: F grade for course.
  - Third violation: F grade for course AND academic probation, suspension, or expulsion.
  - See https://www.uab.edu/one-stop/policies/academic-integrity-code for details

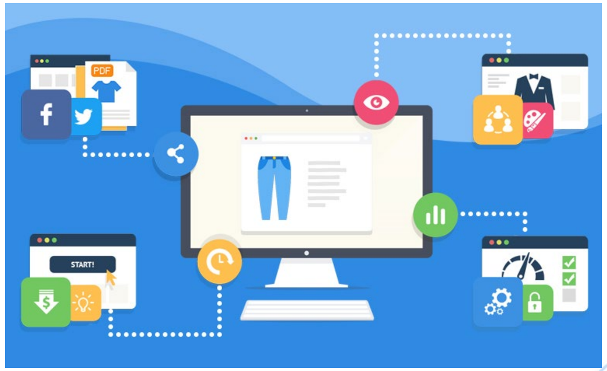
Introduction to Data Science

- What is data science?
- Where is data science needed?
- What is covered in this course?

The Big Data Era

The Big Data Era

- Large-scale data is everywhere

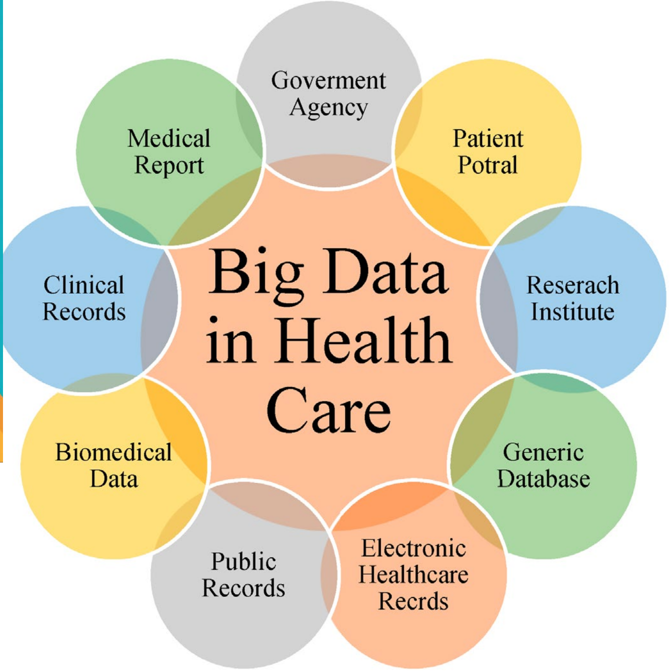


Web data

Social Network

Financial transactions

Healthcare



Big Data in Health Care

Goverment Agency

Patient Potral

Reserach Institute

Generic Database

Electronic Healthcare Recrds

Public Records

Biomedical Data

Clinical Records

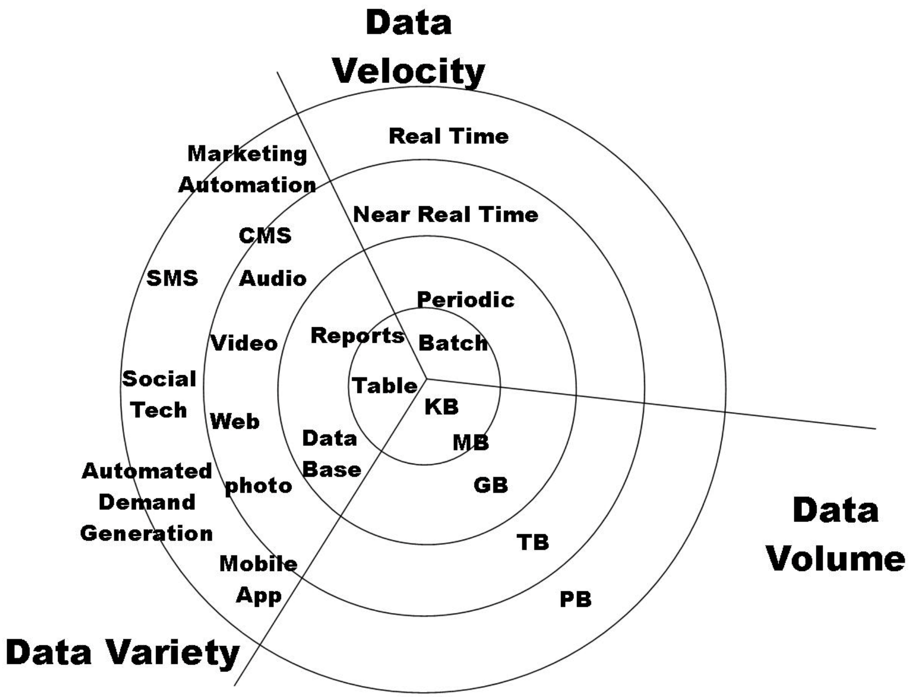
Medical Report

Big Data

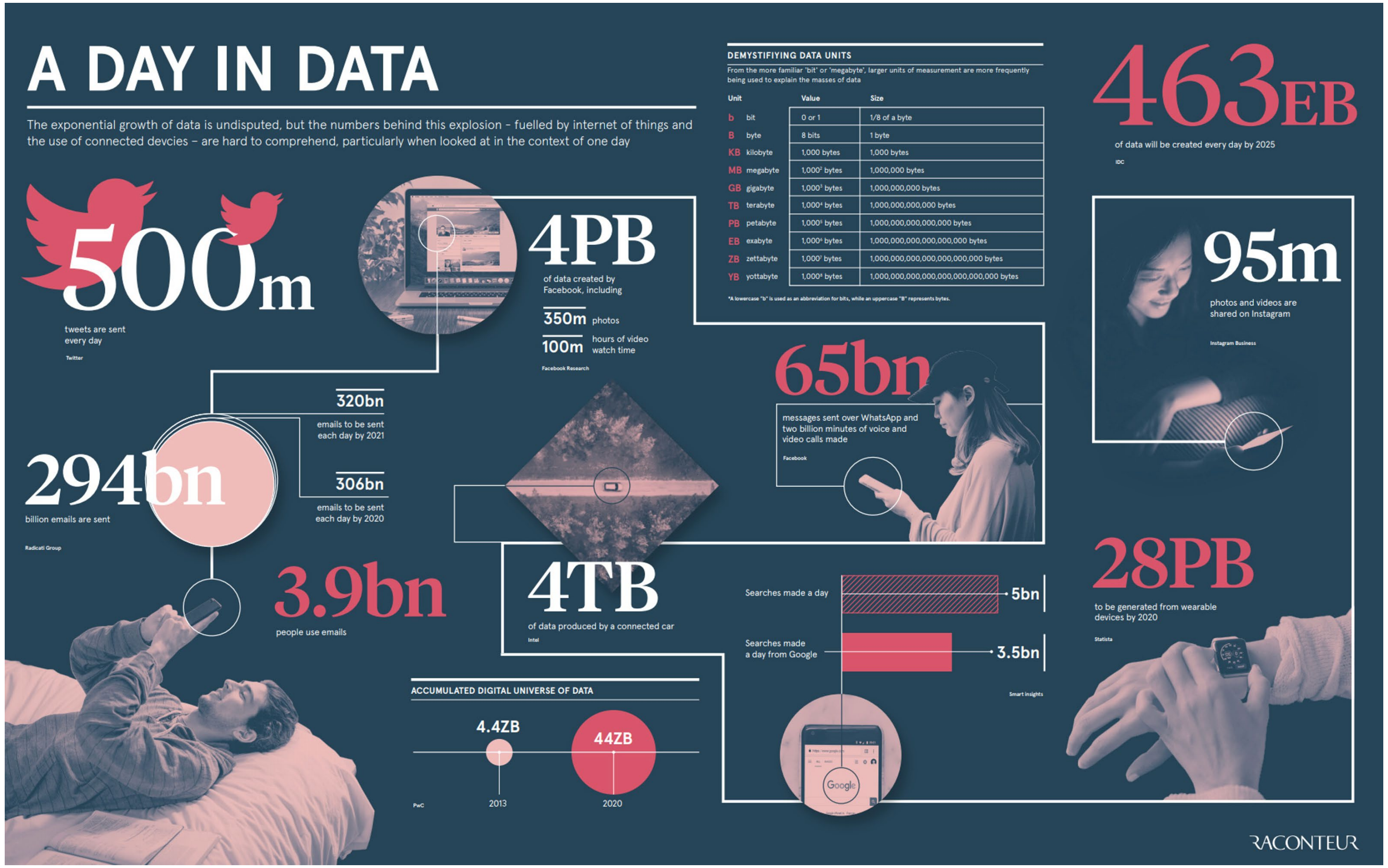
- Our simple definition of “Big Data”: Datasets that are too **large or complex** to be dealt with by traditional data-processing methods.

Big Data

- The Three "Vs" of Big Data
  - **Volume**: The size of the data
  - **Velocity**: The latency of data processing relative to the growing demand for interactivity
  - **Variety**: The diversity of sources, formats, quality, structures.



How Much Data Do We have?



A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

500m tweets are sent every day

Twitter

4PB of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

320bn emails to be sent each day by 2021

294bn billion emails are sent

306bn emails to be sent each day by 2020

Radicati Group

3.9bn people use emails

4TB of data produced by a connected car

Intel

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

65bn messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

463EB of data will be created every day by 2025

IDC

95m photos and videos are shared on Instagram

Instagram Business

Searches made a day 5bn

Searches made a day from Google 3.5bn

Smart insights

28PB to be generated from wearable devices by 2020

Statista

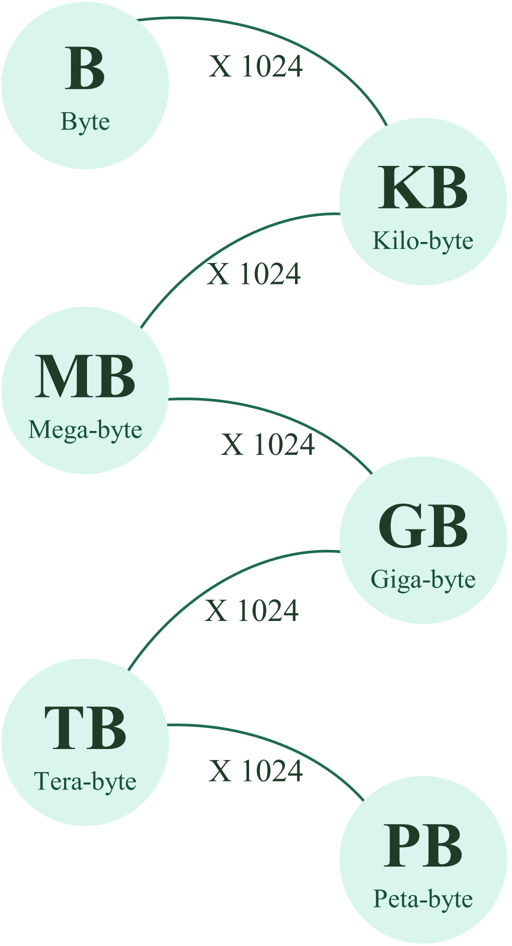
ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB 2013

44ZB 2020

PwC

RACONTEUR



What To Do With These Data?

- As a result of lower cost of computing, storage, and communication, we are now drowning in data.
- A bigger shift in business itself: “information is power” and organizations need to think about what data to collect and what information to extract and how to use it optimally.

What To Do With These Data?

- “Big data is not about the data.”

-- Gary King, Harvard University

- “Hiding within those mounds of data is knowledge that could change the life of a patient or change the world.”

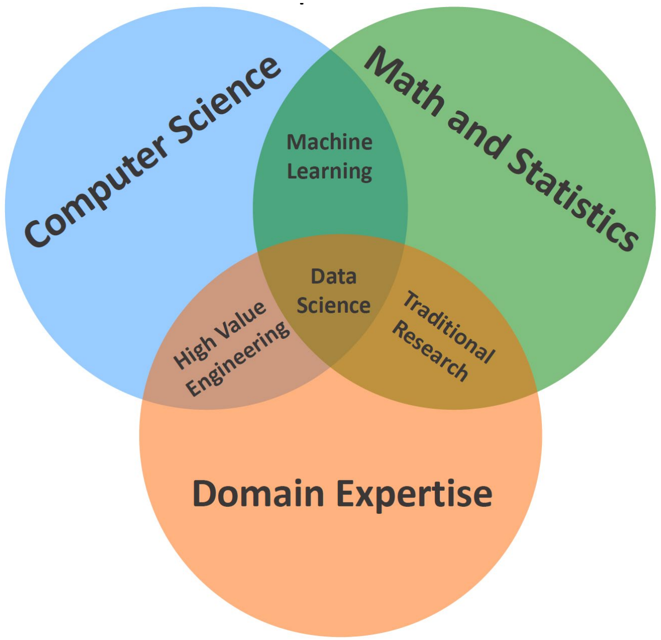
-- Atul Butte, Stanford School of Medicine

Data Science

- **Extract** or extrapolate **knowledge** and **insights** from potentially **noisy, structured, or unstructured data**. These insights can be used to guide decision making and strategic planning.

Data Science

- **Extract** or extrapolate **knowledge** and **insights** from potentially **noisy, structured, or unstructured data**. These insights can be used to guide decision making and strategic planning.
- A broad discipline, including significant aspects of
  - Computer science
  - Statistics and mathematics
  - Data mining and machine learning
  - Information science
  - Domain expertise



History and Evolution of Data Science

- Early Usage
  - 1962: John Tukey conceptualized “data analysis”, a precursor to modern data science.
  - 1974: Peter Naur suggested “data science” as an alternative term for computer science.
  - 1985: C. F. Jeff Wu first used the term “data science” during a lecture at the Chinese Academy of Sciences in Beijing, proposing it as a new name for statistics.
  - 1990s: The discipline began to be recognized formally, with the first conference featuring data science held by the International Federation of Classification Societies in 1996. This era also saw the rise of terms like “knowledge discovery” and “data mining” due to the increasing size of datasets.

History and Evolution of Data Science

- Modern Usage
  - 2001: William S. Cleveland advocated for expanding statistics into technical areas, suggesting the new scope needed a title change to "data science."
  - 2008: The title "data scientist" was popularized by DJ Patil and Jeff Hammerbacher.
  - 2012: The role of data scientists was highlighted as crucial and in high demand by Thomas H. Davenport and DJ Patil, who called it "the sexiest job of the 21st century."
  - Early 2000s onwards: Several academic journals dedicated to data science were launched, and educational institutions started to formalize data science as a distinct academic discipline.

History and Evolution of Data Science

- Continuing Development
  - Even though data science is very popular and widely used, there isn't a single, agreed-upon definition for it. Some people think the term “data science” is just a trendy word without much specific meaning. It's closely linked to big data and focuses on developing ways to analyze large amounts of data to improve how different industries operate.

More Definitions of Data Science

- Data science is “a concept to unify statistics, data analysis, informatics, and their related methods” to “understand and analyze actual phenomena” with data.

  -- One of the earliest mentions of the concept by Chikio Hayashi (1998),

- Data science is the discipline of making data useful.

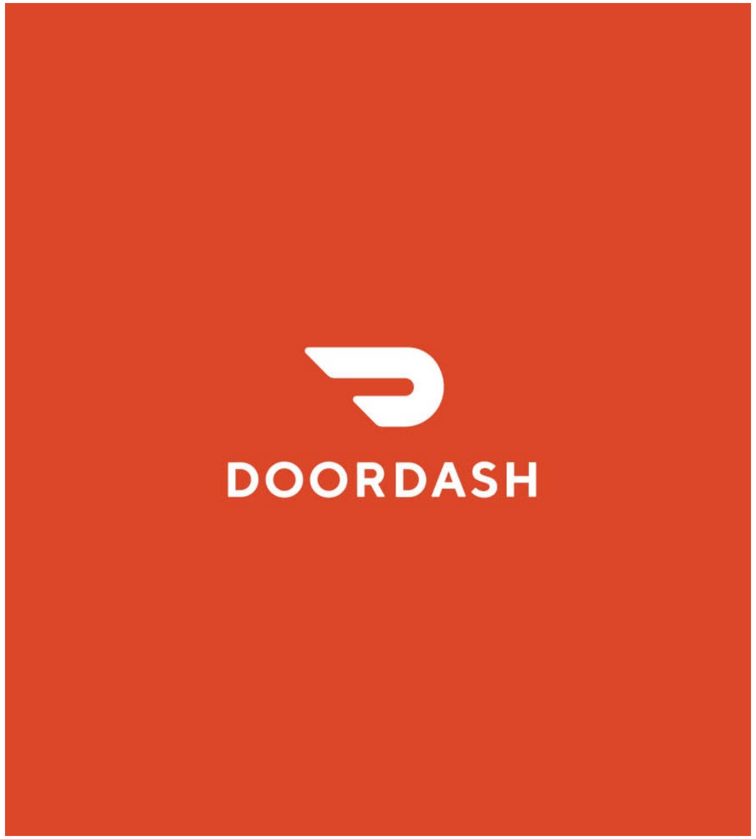
  -- Cassie Kozyrkov (2018), Chief Decision Intelligence Engineer, Google

Real World Applications of Data Science

- Netflix (Entertainment)
  - With over 238 million subscribers worldwide, Netflix has access to a lot of data.
  - Netflix uses this data to create detailed profiles of each of its subscribers and then provides a customized viewing experience.

Real World Applications of Data Science

- DoorDash (Marketing)
  - Data science allows DoorDash to ensure that they do not overspend on unprofitable campaigns.
  - This is achieved through optimizing campaigns in line with historical performance.



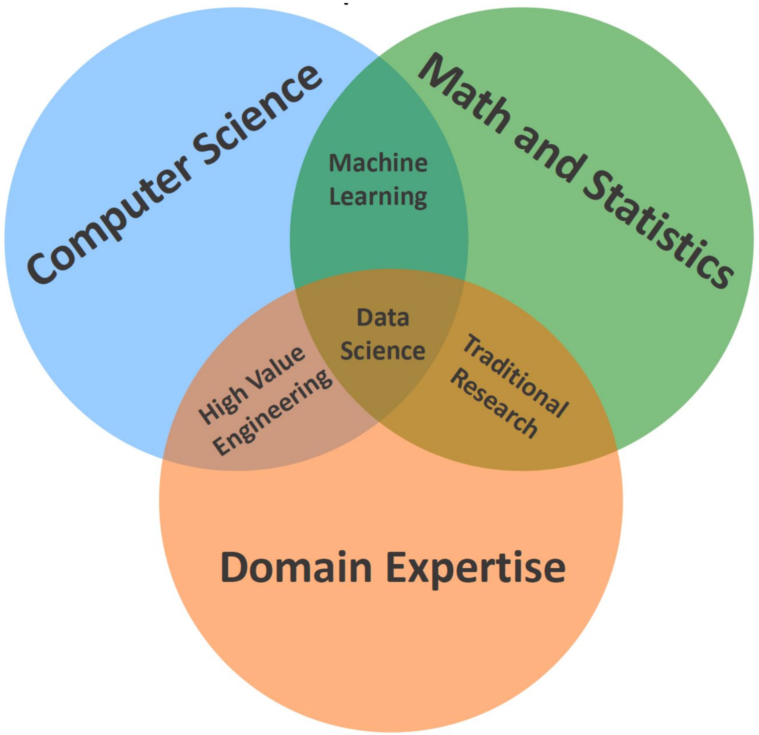
Real World Applications of Data Science

- UPS (Logistics)
  - UPS relies on data science to optimize package transport from drop-off to delivery.
  - It uses an integrated navigation system ORION to help drivers to choose over 66,000 fuel-efficient routes.
  - Saved around 100 million miles and 10 million gallons of fuel per year.

In this Course

Topics

- Data Fundamentals
- High-Dimensional Space
- Singular Value Decomposition (SVD)
- Principal Component Analysis (PCA)
- Algorithms for Massive Data Problems: Streaming, Sketching, and Sampling
- Random Walks and Markov Chains
- Ethics in Data Science



Data science is multidisciplinary.

In this Course

Learning Objectives

- Gain **hands-on experience** with data science **Python** libraries, including NumPy, Pandas, and scikit-learn.
- Learn **basic concepts and theories** in data science.
- Understand the **ethical implications** of data science and explore state-of-the-art works in this domain.

Please complete the questionnaire.

Thank you!