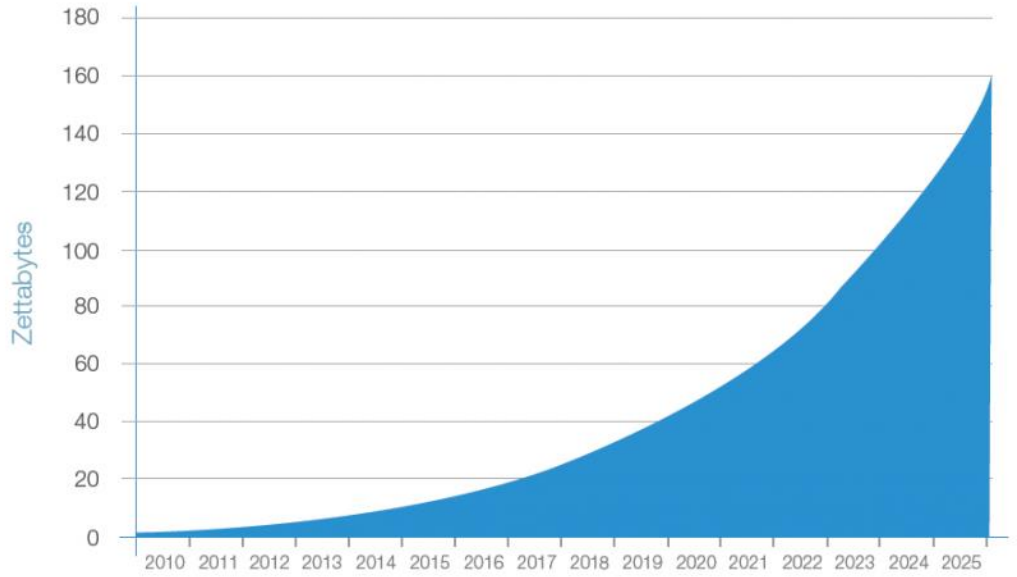
Introduction:
What is Data Science?

CS 418. Introduction to Data Science

Introduction
"The Rise of Big Data"

- Advances in computing power, storage technology, and the way data can be generated, collected, and disseminated, have led to an explosive **growth of data**.
- To make sense of this data, a new professional role has emerged: the **data scientist**.

1 ZB =
1 trillion GB



The majority of this data is complex: **heterogenous** and **unstructured** (text, images, videos)

Source: International Data Corporation (IDC), *Data Age 2025* (2017)

Introduction
"The Best Job in America"

- According to Glassdoor's 2018 rankings, **data scientist** is the **"best job in America"**.
- According to Bloomberg, job postings for **data scientists** at *Indeed.com* rose **75%** from January 2015 to January 2018.

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

01.23.18

Data Scientist is the best job in America for the third year in a row

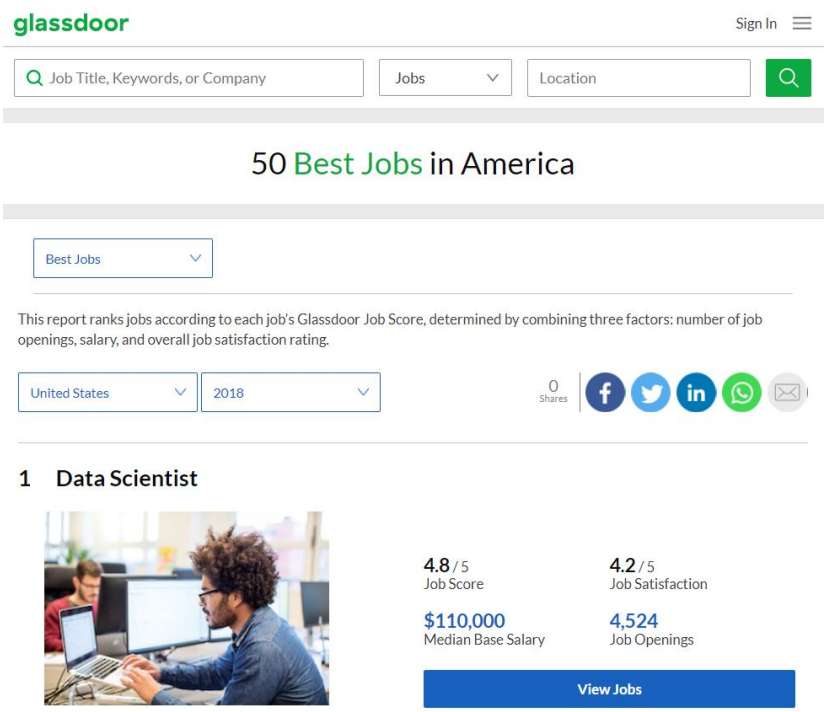
28,790 views | Jan 29, 2018, 02:47pm

Data Scientist Is the Best Job In America According Glassdoor's 2018 Rankings

Is data science tech's 'sexiest job'? Talent wars say so

wp WASHINGTON POST Michael Sasso Bloomberg

Posted 5/26/2018 1:00 AM



Click on the headlines to read the full articles

Introduction
Exercise 1.1

THINK PAIR SHARE

What do data scientists do?

Introduction
What do Data Scientists Do?

- **Data scientists**…
  - "Ensure data quality throughout all stages of acquisition and processing, including **data collection**."
  - "Design, construct, train, and deploy **machine learning models**."
  - "Create **probabilistic and determinist models** to drive business **decision making**."
  - "Develop **predictive models** and simulations using a variety of software and tools."
  - "Perform **exploratory data analysis**, generate and test working hypotheses, and uncover noteworthy trends and relationships."
  - "**Visualize data** insights and **communicate findings** to teams across the organization."

Based on Glassdoor's job postings

What are the qualifications?

- Degree (MS/PhD) in **Computer Science**, **Statistics**, Applied Mathematics or related areas.
- Knowledge of math and **statistics**.
- Knowledge of **programming languages** (Python, R…).
- Knowledge of **machine learning**, NLP, graph theory…
- Experience in **database management** and data **visualization**.
- Experience with **big data** frameworks (Hadoop, Spark).
- Strong **communication** and **teamwork** skills.

Introduction
What is Data Science?

- **Data science** is…
  - "an **interdisciplinary** field that studies and applies tools and techniques for **deriving useful insights from data**."

    Pang-Ning Tan, Michael Steinbach, et al., *Introduction to Data Mining* (2019)
  - "a **new interdisciplinary** field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments…in order to **transform data to insights and decisions**…"

    Longbing Cao, *Data Science: A Comprehensive Overview* (2017)
  - "a **new** field…that focuses on the processes and systems that enable us to **extract knowledge or insight from data** in various forms and translate it into action.…[A]n **interdisciplinary** field that integrates approaches from such data-analysis fields as statistics, data mining, and predictive analytics and incorporates advances in scalable computing and data management."

    Francine Berman, Rob Rutenbar, et al., *Realizing the Potential of Data Science* (2018)

new? | **interdisciplinary field** | that studies and applies methods for | **extracting knowledge or insights from data**

What is Data Science?
Is Data Science New?

"[M]y central interest is in **data analysis**, which I take to include, among other things: procedures for **analyzing data**, techniques for **interpreting the results** of such procedures, ways of **planning the gathering of data** to make its analysis easier, more precise or more accurate..."

John W. Tukey, founding chair of the Department of **Statistics** at Princeton University
The future of data analysis (**1962**)

- Is **data science** a "**rebranding of statistics**"?
  - *Let us own Data Science*.

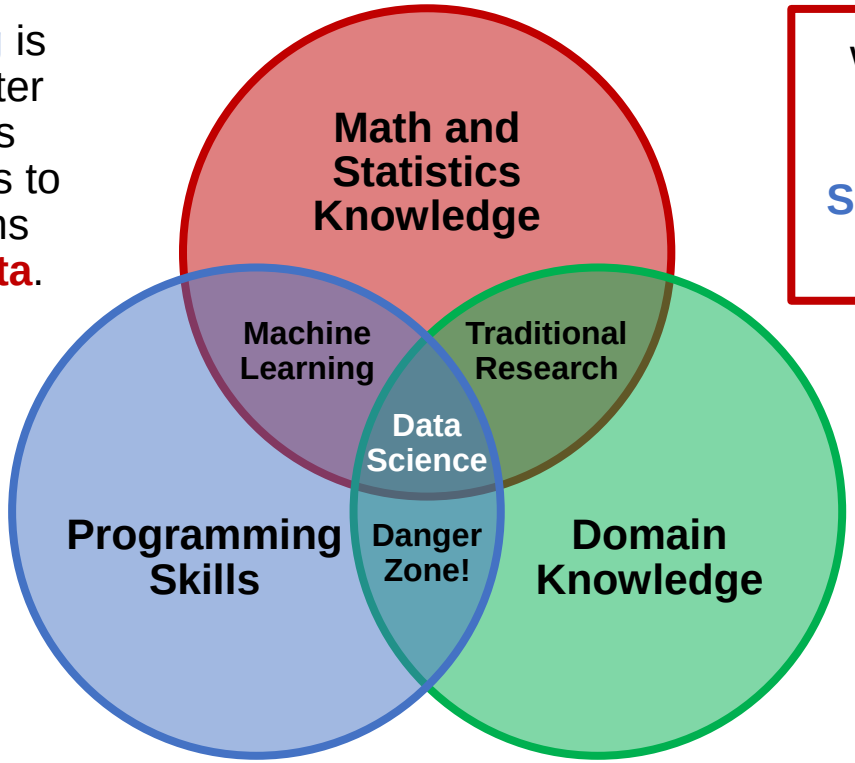
    Bin Yu, 2014 President of the Institute of Mathematical **Statistics**
  - *Aren't we Data Science?*

    Marie Davidian, 2013 President of the American **Statistical** Association
- **Data science**…
  - builds on the theoretical foundations of **statistical** data analysis…
  - to address the **new challenges and opportunities** created by the availability of **big data** and the advances in computing.

What is Data Science?
The Venn Diagram of Data Science

- **Data science** is an **interdisciplinary** field at the intersection of **statistics**, **mathematics**, and **computing**.

Machine learning is the field of computer science that uses **statistical** methods to develop algorithms that learn from **data**.



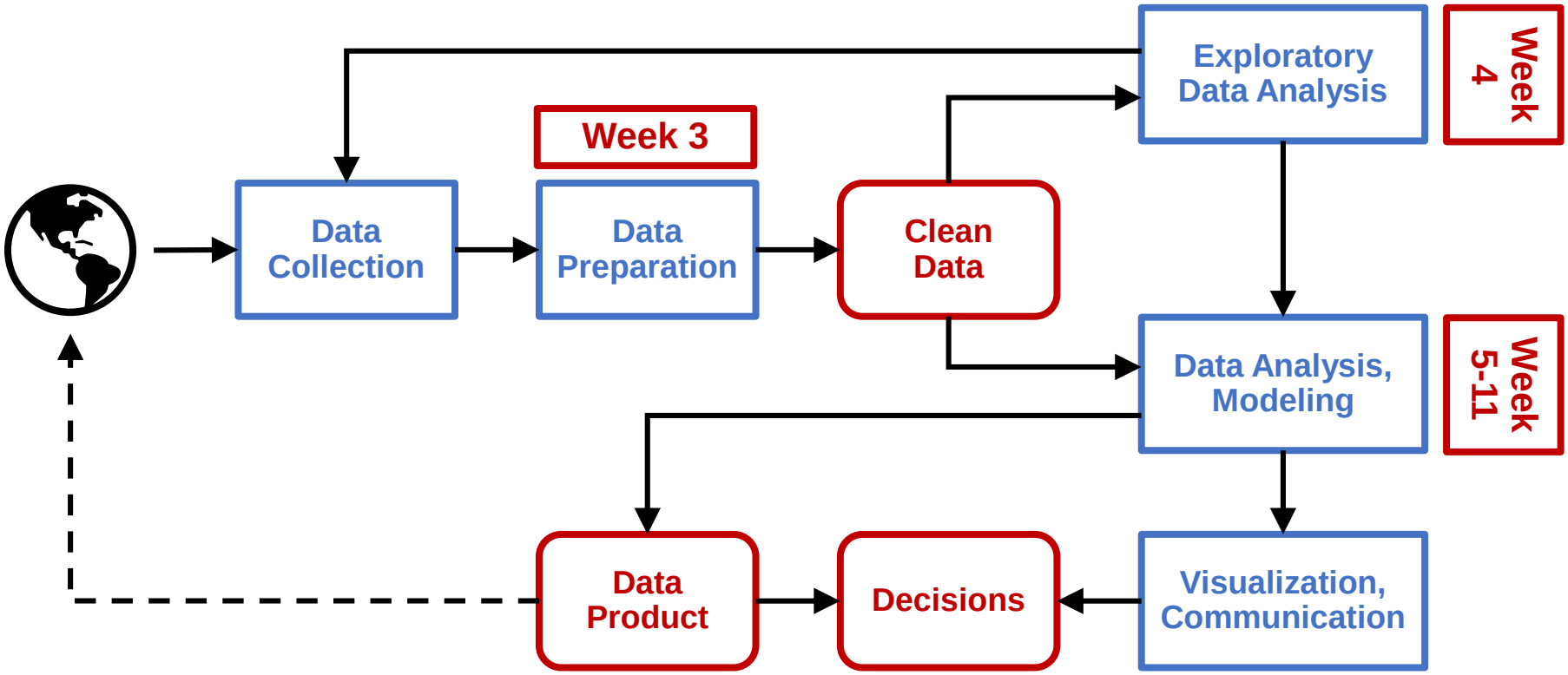
We will review key concepts from **Probability** and **Statistics** in **Week 1** and **Week 2**.

Beware of the danger zone!

Source: Drew Conway, *The data science Venn diagram* (2010)

What is Data Science?
The Data Science Process

- The goal of **data science** is to **extract knowledge or insights from data** through the **data science process**.



Adapted from: Cathy O'Neil and Rachel Schutt, *Doing Data Science* (2013)

Introduction
Lecture Assignment 01

Find a definition of "**data science**,"
share it on Piazza (the instructor will create a thread for this purpose) and **discuss**.

Your discussion may include (but is not limited to) answers to the following questions:

- **How does your definition compare to the ones presented in class (or the ones shared by other students)?**
- **What are the differences and similarities?**
- **Do you agree with your definition? Why or why not?**

Remember to acknowledge the source of your definition!

References

- Francine Berman, Rob Rutenbar, Brent Hailpern, et al., *Realizing the Potential of Data Science* (2018).
- Longbing Cao, *Data Science: A Comprehensive Overview* (2017).
- Drew Conway. *The data science Venn diagram* (2010).
- Cathy O'Neil and Rachel Schutt. *Doing Data Science* (2013).
- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar. *Introduction to Data Mining* (2019).