Data Science: Interdisciplinary Opportunities for Education and Development

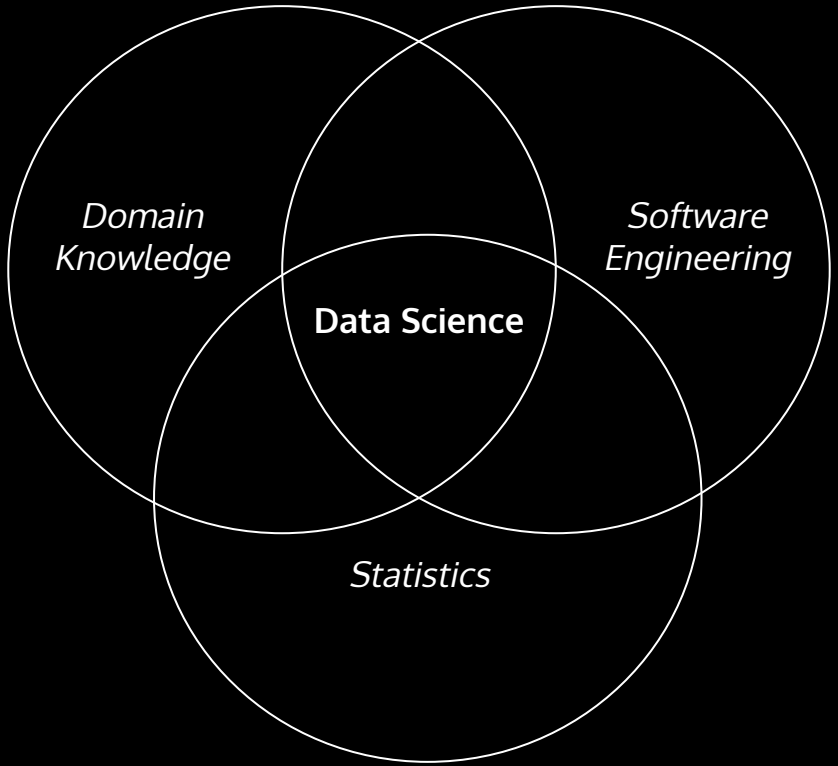
Munif Ishad Mujib
PhD Candidate, Drexel University
In collaboration with EMK Center
August 15, 2020

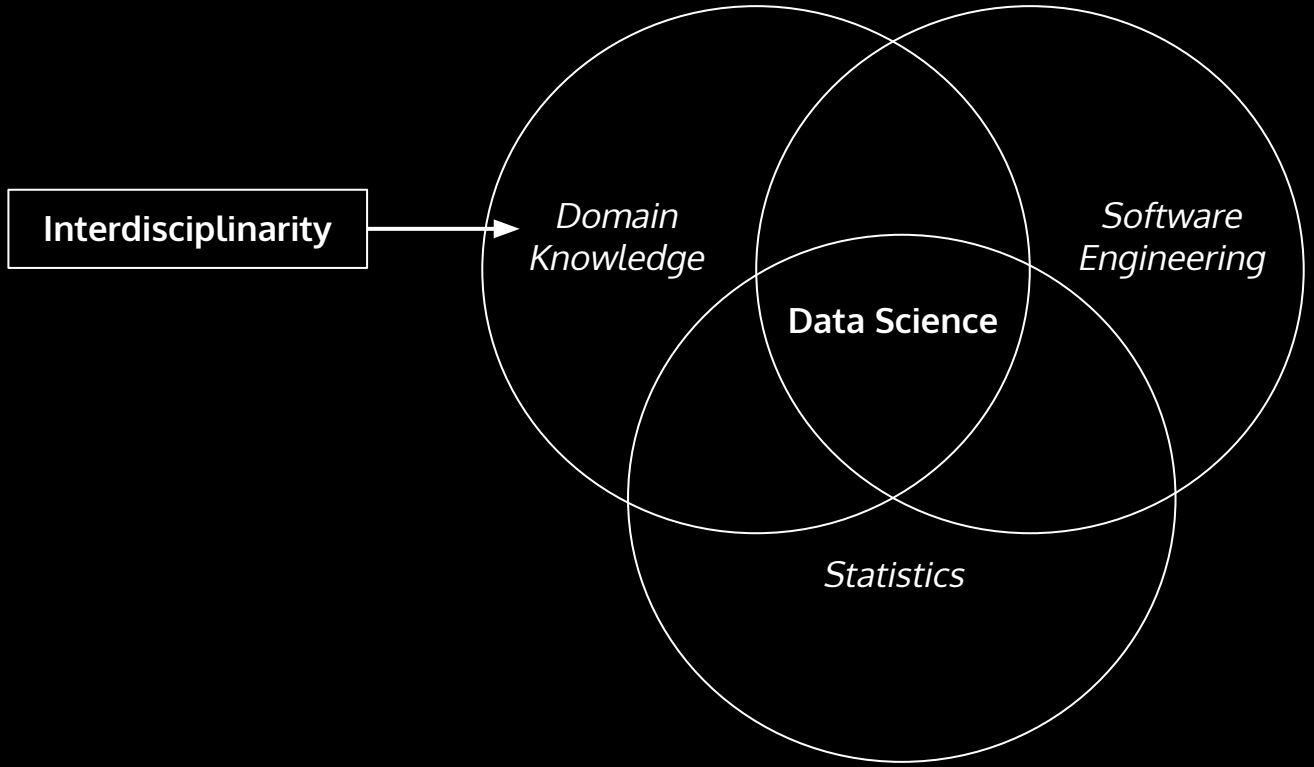
What is Data Science?

All science is data science. However, when we talk about data science as a discipline, we define it as the study of extracting value from data.

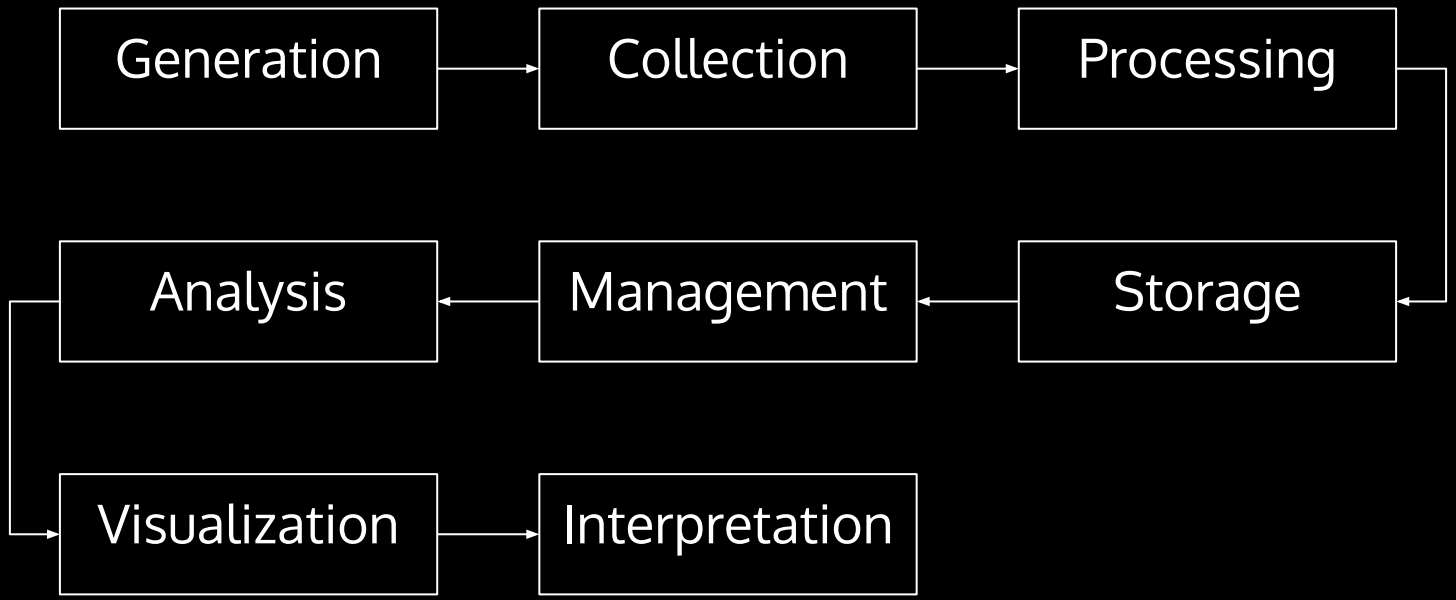
Data Science ≠ Machine Learning

Machine Learning ⊂ Data Science


Domain
Knowledge
Software
Engineering
Data Science
Statistics


Interdisciplinarity
Domain Knowledge
Software Engineering
Data Science
Statistics

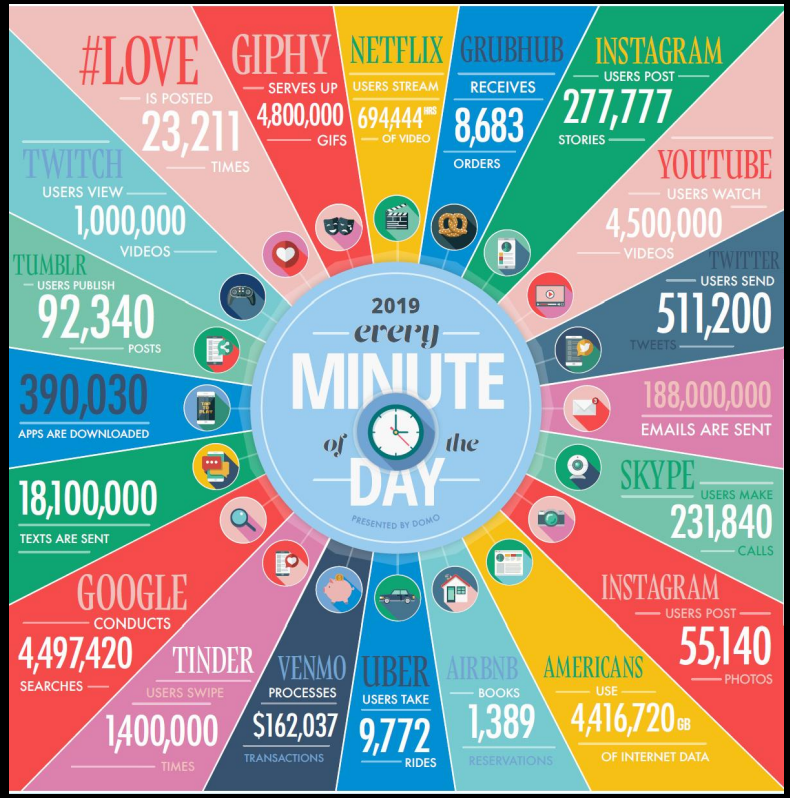
The Data Life Cycle



The Data Life Cycle: Generation

- Digital interaction
- Sensors and automated systems
- Scientific instruments and experiments

By 2025, we'll be generating about 463 EB (10^{18} bytes) of data every day.



Graphic: The Visual Capitalist

The Data Life Cycle: Collection

There are two types of data collection: structured and unstructured.

With structured collection, we utilize purpose-built systems such as Application Programming Interfaces (APIs).

However, if the data that you need isn't being conveniently served by an API, you might have to go through ad-hoc routes such as web scraping: this is what we call unstructured collection.

The Data Life Cycle: Processing

An often under-emphasized yet major part of the process: data scientists spend a significant portion of their time at this stage!

Terms you may have heard of that are relevant to this phase: “data cleaning”, “data wrangling”, and “data munging”.

The Data Life Cycle: Storage and Management

We often separate this phase into a separate discipline and call it “Data Engineering”.

Data engineering primarily consists of working with database systems as well as creating metadata, such as indexes.

The Data Life Cycle: Analysis

This is where some of the more popular aspects of data science come in: basic statistics, network analysis, statistical inference, machine learning, artificial intelligence, and more.

Doing data analysis requires having at least *some* knowledge of statistics and algorithms.

The Data Life Cycle: Visualization and Interpretation

Essentially, any data science project is about storytelling. Communicating findings, results, and predictions through easy-to-understand materials is crucial.

Data visualizations aren't just plain old charts and plots any more, either. *Interactive* visualizations and dashboards can themselves become complex projects.

Privacy and Ethics

Privacy and ethical concerns can come up all through the data life cycle. Preserving privacy and avoiding biased decision-making are hard challenges.

There is significant research and innovation starting to happen in these areas.

Tools of the Trade: Programming Languages

Tools of the Trade: Development Platform

Learning and getting comfortable with the Unix command line is essential.

Tools of the Trade: Useful Python Libraries

base python	pandas	matplotlib	scikit-learn
	sqlalchemy	seaborn	keras
		plotly	tensorflow
			pytorch

Getting Started with Data Science Research

- Gain basic familiarity with tools
- Build example projects
- Find or collect interesting datasets
- Establish hypotheses/research questions
- Run experiments
- Document project

Recommended Resources

- Python basics on codecademy: https://www.codecademy.com/learn/learn-python
- Bash guide for beginners: https://www.tldp.org/LDP/Bash-Beginners-Guide/html/
- The Python Data Science Handbook: https://jakevdp.github.io/PythonDataScienceHandbook/
- Data Science from Scratch, 2nd Edition, by Joel Grus
- Head First Statistics, by Dawn Griffiths
- Datasets, examples, and competitions: Kaggle https://www.kaggle.com/
- A programmer's best friend: StackOverflow https://stackoverflow.com/

Thank You

Slide deck available at
https://munifmujib.github.io/emk-data-science.pdf