
SDN

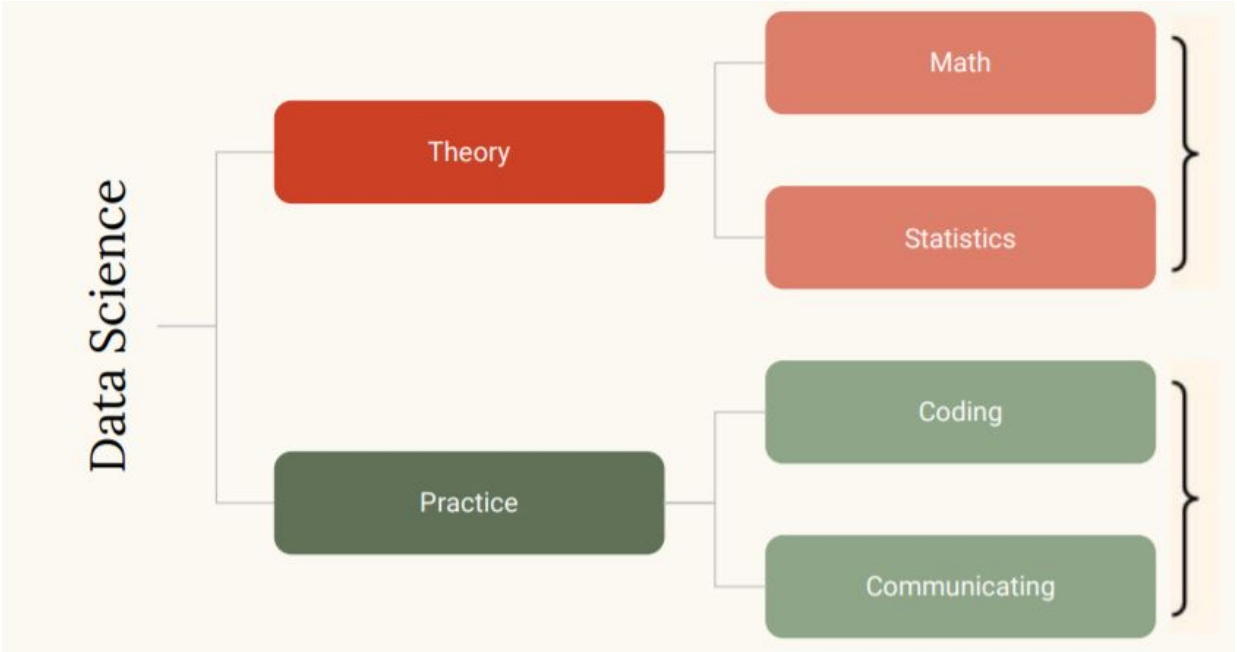
What will we do?

- Introduction to data science with an emphasis on mathematical foundations and real-world applications
- Data science application
    - Labs
    - Projects
    - Competitions
    - Speakers

Our Goals

- Understand of the goals of machine learning and what it can and cannot do
- Be able to describe, both in theory and practice, a few fundamental algorithms used in machine learning and be able to apply them to data sets
- Be able to obtain and work with real-world data sets in Python using pandas and numpy and visualize results using matplotlib/seaborn/plotly
- Apply machine learning techniques to glean insights from data sets
- Evaluate data insights and be able to draw and report conclusions from them
- Understand the role of ethics, privacy and fairness in data science
- Perform a real-world data analysis of your choosing to consolidate all of the above

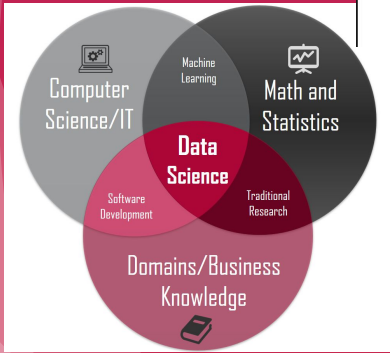
Class Format



Prathul & Tony

Inba & Tony

What Is Data Science?




DATA

Data Scientist: The Sexiest Job of the 21st Century

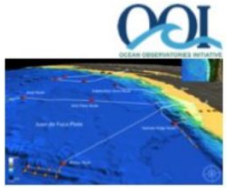
by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Industries

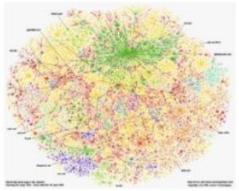
Finance

Physics



Oceanography

Agriculture



Sociology: The Web

Biology: Sequencing

Commerce

Neuroscience: EEG, fMRI

Data-Driven Medicine

Sports

Academics


uchicago news
New College data science major: From foundations to insight to impact


Wharton
FACULTY & RESEARCH
YOUTH PROGRAM
UNDERGRAD
MBA
EMBA
PHD
EXECUTIVE EDUCATION
WHARTON ONLINE
ALUMNI
DEPARTMENT OF STATISTICS AND DATA SCIENCE
HOME
FACULTY
RECRUITING
RESEARCH
PROGRAMS
RESOURCES
DEPARTMENT INFORMATION
FIND AN EXPERT
Wharton Statistics and Data Science


Yale University
SEARCH THIS SITE
Department of Statistics and Data Science
S&DS

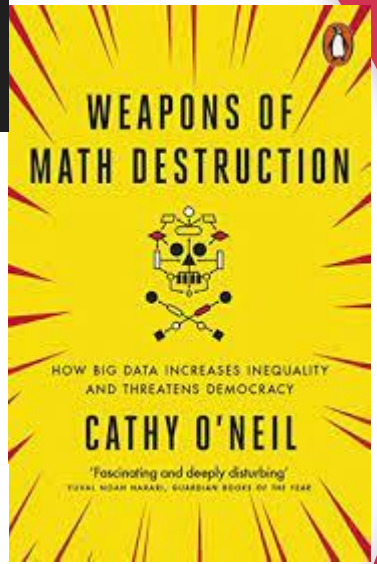
Ethics

Bernard Park

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016



Artificial intelligence

Predictive policing algorithms are racist. They need to be dismantled.

Lack of transparency and biased training data mean these tools are not fit for purpose. If we can't fix them, we should ditch them.

by **Will Douglas Heaven** July 17, 2020

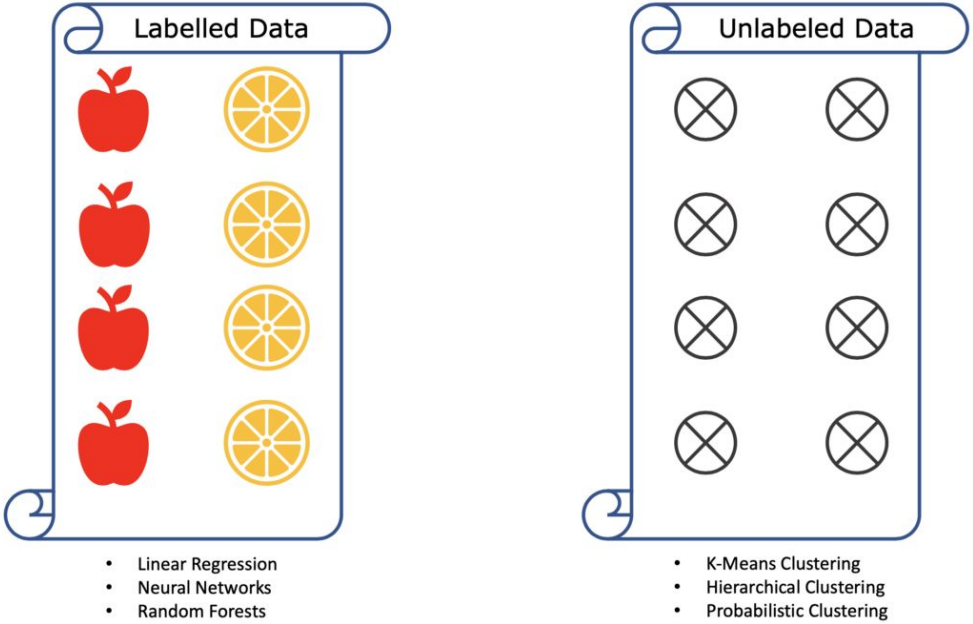

01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co

Data Science Models

Comparison Chart

BASIS FOR COMPARISON	SUPERVISED LEARNING	UNSUPERVISED LEARNING
Basic	Deals with labelled data.	Handles unlabeled data.
Computational complexity	High	Low
Analyzation	Offline	Real-time
Accuracy	Produces accurate results	Generates moderate results
Sub-domains	Classification and regression	Clustering and Association rule mining

Types of Data

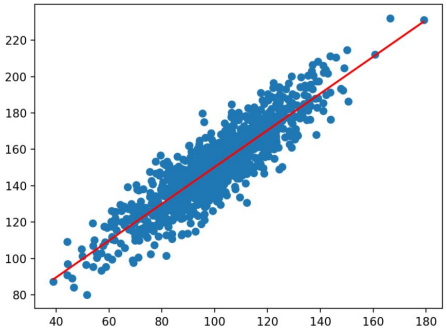


https://miro.medium.com/max/1400/0*P9sA0QKnXC0Ip9NI.png

Analysis vs Prediction

Analysis: What can we learn from the data we have?

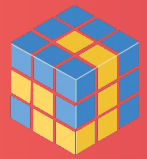
Prediction: Given these things that we know about the data, what can we say about what will happen in the future?



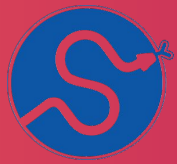
Python, Jupyter, GitHub

- We'll be using Python, which has an extremely powerful and collection of data science libraries and tools
    - If you've never used Python before - no worries!
- In particular, we will be using a Python interface called Jupyter Notebooks with labs accessible in a GitHub repository

Python For Data Science



NumPy



SciPy

IP[y]:
IPython

scikit
learn

matplotlib

pandas