Introduction to Data Science
Lecture 1: Introduction

Ebelechukwu Nwafor, Ph.D.

Agenda

- Day 1:
  - Introduction to Data Science
    - Overview of Data Science: Concepts, Applications, and Workflow
    - Key Tools and Technologies in Data Science
  - Hands-On Sessions
    - Setting up your Environment
    - Installing Jupyter Notebook: Step-by Step Guide
  - Getting Started with Pandas
    - Introduction to Pandas Library
    - Key Operations: Data Frames, and Data manuplation
  - Exercises
    - Practice Loading, Exploring and Cleaning Data using Pandas

Agenda (cont'd)

- Day 2
  - Data Visualization with Python
    - Importance of Data Visualization
    - Overview of Visualization Libraries: Matplotlib and Seaborn
  - Hands-On Session
    - Creating Basic Plots: Line, Bar, and Scatter Plots
    - Advanced Visualizations: Heatmaps, Pair Plots, and Customizations
  - Exercise

Lecture Outline

- What is Data Science?
- The Process of Data Science
- Tools and Languages used in Data Science
- Real-world Applications of Data Science
- The Future of Data Science

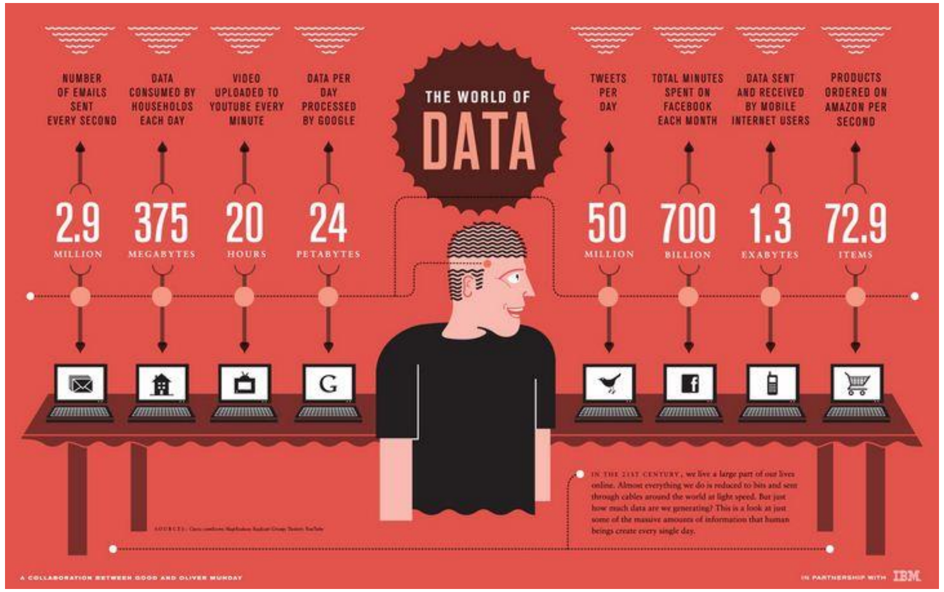
Data is Everywhere

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network

Big Data

- Google processes 20 PB a day (2008)
- Facebook has 60 TB of daily logs
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- 1000 genomes project: 200 TB

- Cost of 1 TB of disk: $35
- Time to read 1 TB disk: 3 hrs (100 MB/s)


THE WORLD OF DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS
G
SOURCES:
A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY
IN PARTNERSHIP WITH IBM

IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

Big Data

- Big Data is any data that is expensive to manage and hard to extract value from
  - Volume
    - The size of the data
  - Velocity
    - The latency of data processing relative to the growing demand for interactivity
  - Variety and Complexity
    - the diversity of sources, formats, quality, structures.

Types of Data

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- You can afford to scan the data once

Data Use case

- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

What is Data Science?

- Definition: "Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.”

- Data science is all about uncovering findings from data to make informed decisions.

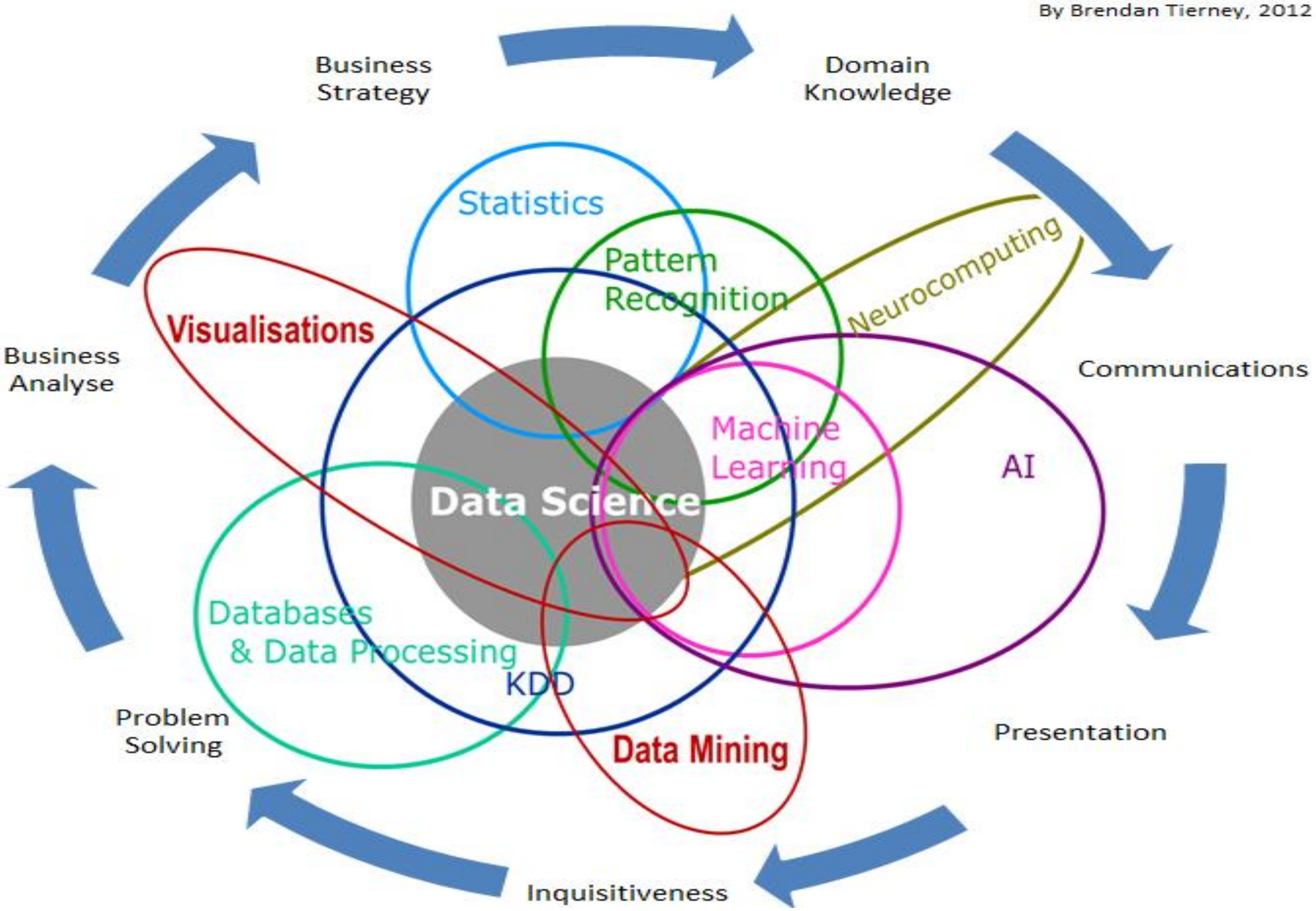
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data

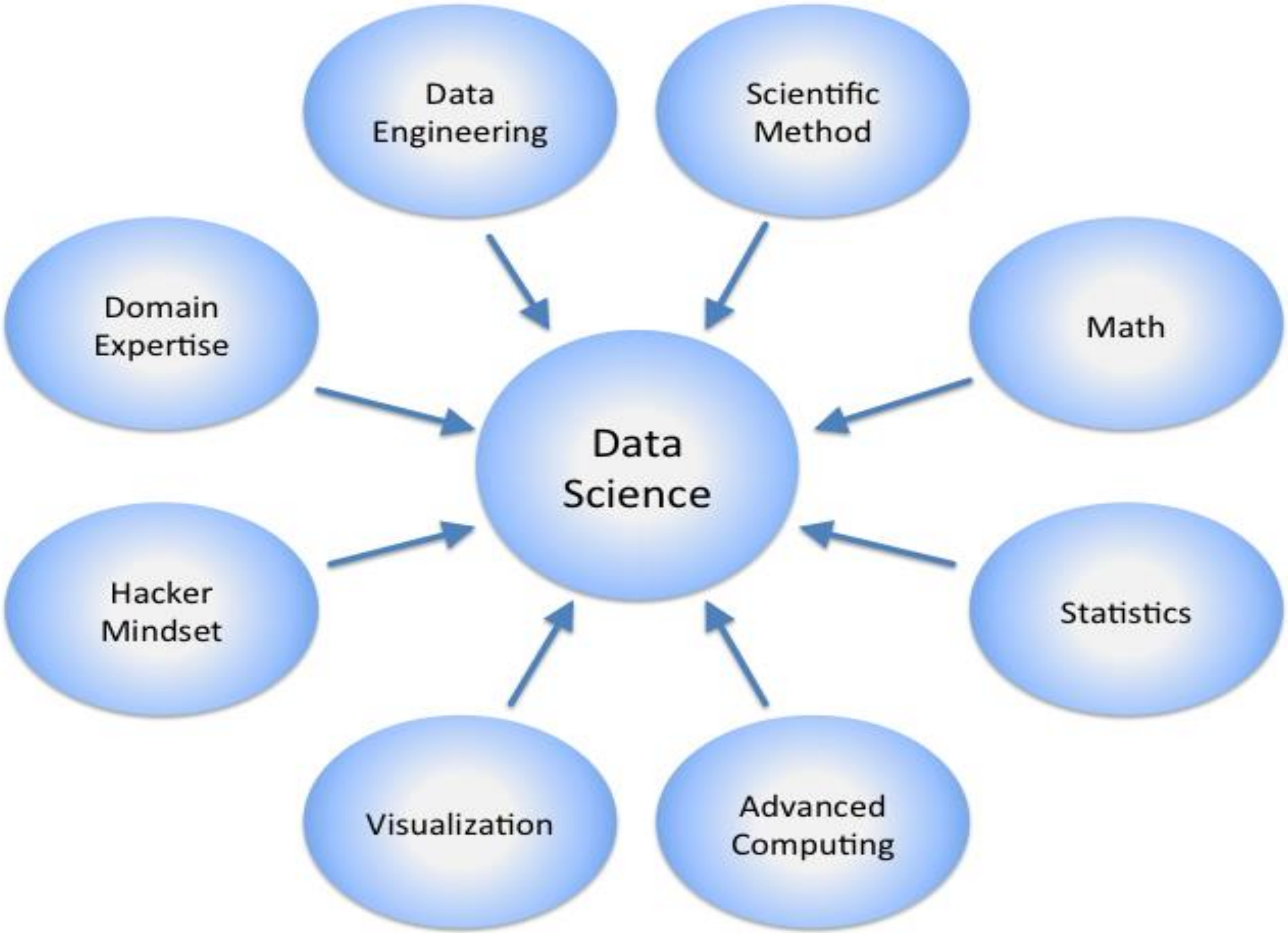
Key Components of Data Science

- Statistics: The backbone that allows data scientists to look at data objectively.
- Computer Science: The technical skills needed to handle large amounts of data, create algorithms, and build data-driven solutions.
- Domain Knowledge: Understanding the field from which the data was gathered to provide meaningful insights.

Data Science Is Multidisciplinary

By Brendan Tierney, 2012




Data Engineering
Scientific Method
Domain Expertise
Math
Data Science
Hacker Mindset
Statistics
Visualization
Advanced Computing

The Process of Data Science

- Problem Understanding:
  - Define the problem clearly and determine how data can help.
- Data Collection:
  - Gather the necessary data from relevant sources.
- Data Cleaning/Preparation:
  - Clean and format the data for analysis.
- Data Exploration/Visualization:
  - Look for patterns, trends, and outliers in the data.

The Process of Data Science (cont’d)

- Model Building
  - Create predictive or explanatory models using machine learning.
- Model Evaluation
  - Test the model’s accuracy and reliability.
- Presentation of Results
  - Communicate your findings in a clear and understandable way.
- Deployment: Implement the model or findings in real-world scenarios.

Tools and Languages used in Data Science

- Python: A versatile, easy-to-learn language used for data manipulation and analysis.
- R: A language used primarily for statistical analysis.
- SQL: A language used for managing and retrieving data in databases.
- Libraries: Pandas (data manipulation), NumPy (numerical operations), Matplotlib (visualization), Scikit-learn (machine learning), TensorFlow (deep learning).

Real-world Applications of Data Science

- Healthcare: Predictive models to anticipate disease outbreaks.
- Finance: Fraud detection and risk assessment.
- Retail: Predictive analytics for inventory management and personalized marketing.
- E-Commerce: Recommendation systems for improved customer experience.
- Social Media: Sentiment analysis and trend prediction.

Case Study - Netflix Recommendation System

- Problem Statement:
  - With a massive, diverse content library and a broad user base
  - Netflix faced the challenge of effectively recommending the right content to the right users to enhance viewer engagement and satisfaction.

- Data Collection:
  - Netflix collects a massive amount of data from its users, including their viewing habits, the time they watch content, the devices they use, their ratings, and even the moments they pause, rewind, or fast-forward.

Case Study - Netflix Recommendation System

- Model Building:
  - Machine learning algorithms are trained on the data to create a recommendation system.
  - This recommendation system uses collaborative filtering, where recommendations are based on users with similar viewing habits, and content-based filtering, where recommendations are based on the content the user has viewed and rated highly.
- Model Evaluation:
  - The recommendation system is constantly tested and evaluated for its effectiveness.
  - The primary measure of success is whether users are engaging with the recommended content.

The Future of Data Science

- AI and Machine Learning: Automated and more efficient data analysis.
- Big Data: Handling larger and more complex data sets.
- Data Privacy and Ethics: Ensuring responsible use of data.

Career Opportunities in Data Science

- Data Scientist: Requires strong statistical and programming skills.
- Data Analyst: Involves analyzing and interpreting complex datasets to help businesses make decisions.
- Machine Learning Engineer: Involves creating data funnels and delivering software solutions.