I, Overview

* Emerging technology allow us to manage large amount and types of data. Emerging technologies describe technologies which have emerged within the past few years like IOT, deep learning , blockchain , cellular, AI
* CEET is intended to empower a generation of ethical technologists who can

promote ethics in all data driven technologies ( technology based on data)

* In 3 courses, Promoting the Ethical Use of Data-Driven Technologies, Turning Ethical Frameworks into Actionable Steps, and Detecting and Mitigate Ethical Risks.

Example of bias:

* Facial recognition software often performs poorly on people with darker skin tones because it's trained predominantly on images of lighter-skinned individuals.
* Early versions of Apple’s health app omitted features for tracking menstruation, a critical health aspect for many women.
* Voice recognition systems that fail to understand accents or speech patterns different from the dominant language or dialect they were tested on.
* Goal : integrate ethics into emerging technology to mitigate bias, fake information.

II, Data Science Fundamentals

* Data science involves making sense of data, more data more better.
* Big data involves collecting a massive data together, even if it might be from different locations or time periods, in order to discern if the aggregated data can tell us something more than just the sum of its parts
* DIKW Pyramid:
  + Data: Raw, unprocessed input (numbers, words, etc.).
  + Information: Organized, categorized data useful for decision-making.
  + Knowledge: Information in context, leading to actions.
  + Wisdom: knowledge built up over time so prior experience informs new decisions. (Wisdom is when knowledge, which has been built up and refined through experience over time, is used to inform future decisions.)
* **Big Data's "V"s are**:

*Describe Nature of BigData*

* **Volume**: Magnitude of data (e.g., social media content).
* **Variety**: Different data types (text, audio, video, log file, user activity data, etc.).
* **Velocity**: Data gathered at various speeds.
* **Variability**: Consistency or unpredictability within data.

*Describe Usefulness of Data*

* **Veracity**: Is the data truthful and accurate?
* **Validity**: Is the data relevant and suitable for your intended purpose?
* **Vulnerability**: Is the data at risk of being intercepted, stolen, or altered?
* **Volatility**: How long can the data be considered valid or useful ?
* **Visualizability**: How challenging is this data to visualize?
* **Value**: Data's ability to provide actionable or economic benefits.
* Data Format : data can be categorized into 3 by its format
  + Structured : follows a fixed schema of attributes like a library database.
  + Unstructured : For example, data scraped from websites, including mixed media of audio, visual, tabular, and even animated data
  + Semi-structured: falls between the two, with some organization but more flexibility, like images with various formats or resolutions.
* Have to think about Data Security when working with massive amounts of data:
  + **jurisdiction** can refer to the laws governing how data must be stored, transferred, or protected depending on where the data is located
  + Data protection laws are becoming stricter globally, such as the International Safe Harbor privacy principles and GDPR in the EU.

|  |  |
| --- | --- |
| Data at rest | Data at motion |
| Data at rest is data stationed on a system where it is likely to remain there for the foreseeable future.  but it's an appealing target for attackers. | moves between systems or locations and is more vulnerable to interception, though often in smaller amounts compared to data at rest. |

* A heuristic is a rough, quick rule that works most of the time, but not always.

Example: For example, the way that someone dresses might cause us to make assumptions about their wealth or politics, even if that's not necessarily the case

* Availability bias : a mental shortcut that relies on immediate urgent information and tries to generalize the findings using only the most obvious information.

Example: people tend to overestimate causes of death that makes the news such as airplane crashes and underestimate causes of death, such as heart disease.

* Data is the driver of algorithmic automation, such as machine intelligence, it can even be compared to fuel. Cleaning data up to ensure its value and validity are essential steps in enabling safe and trustworthy automation.
* Data analytics is the science of examining data to derive insights and conclusions. It includes the process of cleaning up the data and transforming it into a format that makes it easier to work with, as well as the actual processes used for the analysis. Tool example: Apache Spark, SAS, RapidMiner, Splunk, BI tools, Tableau, or even good old Excel.
* Data-driven insights help businesses improve decision-making, reduce costs, and find new opportunities.
* **3 Categories of Analytics:**
  + Descriptive Analytics:
    - Summarizes past data to answer "*what has happened?*".
    - Examples: Creating charts or summaries of past sales data.
  + Predictive Analytics:
    - uses data from the past to forecast future data trends.
    - answers the question: *what is likely to happen? Or when and why something may happen*.
    - uses statistical models and probability theory to make predictions.
  + Prescriptive Analytics:
    - uses data to predict multiple possible outcomes and recommend an action.
    - Answers "*what should we do?"*
* Data Science Pipeline & Workflow:
  + Pipelines process data routinely, producing similar outputs consistently.
  + Workflows are used for one-off, specific requests with varying output formats.
  + A data science pipeline is a *linear object* and the data science workflow is more like a *continuous circle.*
* **6 Stages of Data Science Process:**
  + Identify the Problem: Define the question you want to answer; better questions lead to better insights.
  + Acquiring and Storing Data: Source data from sensors, public datasets, or human input. Ensure ethical collection and storage.
    - can come from a number of sources, such as from sensors, readouts, blogs, or machines, or human input.
    - may also be sourced from open and public data sets.
  + Cleaning Data: Organize and standardize data, removing errors or inconsistencies to improve analysis quality.
    - looking for suspect, erroneous, or poorly tagged data and either repairing it if feasible or discarding it
  + Exploring Data: Use statistical methods and visualizations to discover patterns and insights in the data.
  + Analyzing/ Modeling Data: Apply analysis tools (e.g., R, SciPy) to construct models that predict trends or reveal relationships.
  + Communicating Findings: Share results in a story-like manner, considering ethical implications and audience impact.
* Questions:
  + **What are some ethical issues that should be considered by leaders at organizations relying on data science?** 
    - data about us is incorrect, biased, or taken from us without permission. => Ensuring that technologies are transparent, accountable, and minimally biased!
  + **How can data be collected or used in a way that doesn’t respect our agency?**
    - data might be collected from sources we didn’t intend.
    - data might be used in a way that we didn’t authorize or foresee.
  + **What can organizations do to make sure they collect and use data in a transparent, accountable, and minimally biased way?**
    - Organization must handle data in a responsible manner
    - Collected data should not be massaged, or manipulated in any way, or the interpretation of that data altered from the truth
    - understand the result of certain decisions
    - applying foresight to preempt potential issues
  + **What kinds of ethical considerations for data science should members of society be focused on?**
    - make products that preserve the dignity of all in society
    - avoids exploitation
    - benefits participants, not merely for commercial benefit.
  + What steps can the public take to hold tech companies accountable? What organizations are leading the charge of public accountability?
    - Support public pressure groups like the Electronic Frontier Foundation and professional organizations like IEEE or ISO.