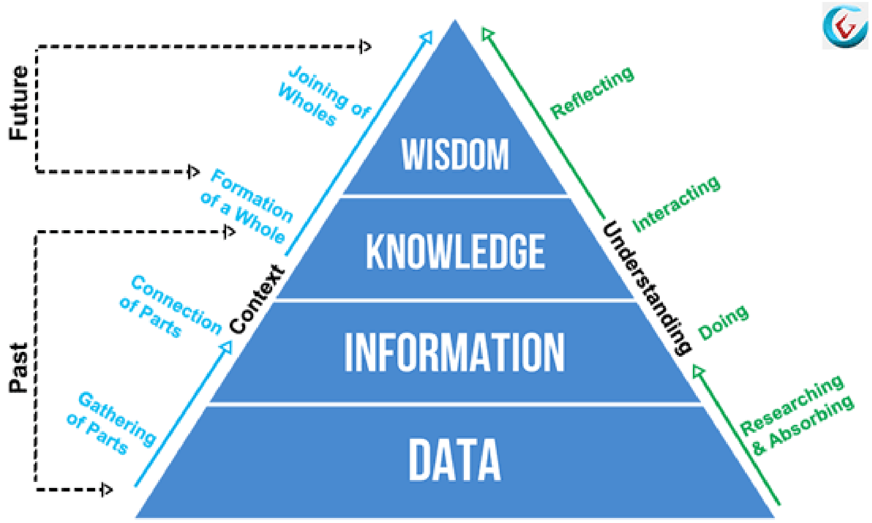
Promote the Ethical Use of Data-Driven Technologies

Data Science Fundamentals

DIKW pyramid



DIKW Pyramid
CertGuidance.com

V's of big data

- Nature of Big Data
    - Volume
        - Magnitude of data
    - Variety
        - There are many sources and types of data such as audio, image, video file, text data, log files, user activity data, machine and sensor data. data may be a mix of public, confidential, or restricted forms. There may be a variety of different rights, obligations, or jurisdictions attached to different data points or sets.
    - Velocity
        - Data from multiple sources may be collected at different frequencies, ex per sec, per min, per day, etc
    - Variability
        - Even similar data can still have some sudden spikes or odd outliers
- Usefulness of Data
    - Veracity
        - Is the data truthful, is the source trustworthy, is the sensor faulty, etc
    - Vulnerability
        - Is the data at risk of being intercepted, stolen, or altered?

- Volatility
    - How long can the data be considered valid
- Visualisability
    - How challenging is this data to visualize?
- Value
    - What value can be derived from the data? Does it easily translate into tangible benefits or actionable insights?
- Validity
    - Is the use of this data correct and accurate for an intended use? Is the data sufficient enough to draw meaningful conclusions from?

Artificial Intelligence Fundamentals

Human Intelligence

- Typical capabilities of agents with intelligence include:
    - The ability to combine sensory data in order to perceive the environment
    - To obtain knowledge and apply it to a problem
    - To come up with a solution and to take action for the purpose of achieving goals
- Artificial intelligence
    - Able to replicate many of these functions. However, few systems successfully integrate multiple functions together
    - We can mimic the approximate function of an optical cortex, perhaps, but not an interconnected brain
    - They function more like our subconscious processes

Narrow AI

- Can only perform a task it was designed to
- Spam filters, autonomous cars, book recommendation engines, and the auto typing of pictures with somebody's name, machine learning
- Narrow AI can process massive amounts of data quickly and efficiently to allow humans to make better decisions

General AI and Superintelligence

- Artificial General Intelligence (AGI)
    - Machines that can perform any intellectual tasks that humans can
    - This means to not only process data logically but also to think abstractly
- Artificial superintelligence (ASI)
    - Technology which might surpass human intelligence
    - Smarter than the smartest human who has ever lived

Ambient Intelligence and the Internet of Things

- Seeks to make our everyday environments intelligent and sensitive to us
- It draws on the advances in pervasive computing, artificial intelligence, sensors, and sensor networks, promising to transform our lives by making our surroundings flexible and adaptive
- Ambient Intelligence systems are embedded
- Factors that impact people's willingness to adopt ambient intelligence systems
    - Usability
        - The systems must be reliable, unobtrusive, and easy to implement by ordinary people
    - Trust
        - People need to feel it's constant data collection about them will be used responsibly and kept secure
    - Security
        - IoT devices have often had gaping security holes with default passwords, and rapidly deprecated authentication protocols that are easily hacked

The Black Box Problem

- The term black box is used to describe any AI system whose input and operations are not visible to humans.
- Data flows in and a decision comes out, without anyone knowing what happened in the middle

Data and Privacy

Data Privacy involves the ability an individual has to selectively share their data, retreat from interactions with individuals and companies, control the degree to which one is identifiable when undertaking online or offline activities, control the image created by the data

Personally Identifiable Information (PII)

- Information that can be used to uniquely identify a specific individual.
- Protecting PII is at the core of protecting privacy
- Direct identifiers
    - Can identify a unique individual even if they are the only data point

Privacy Risks in IoT/Ambient Intelligence Technologies

- Potential risks to privacy

    - Data Collection
        - Most people are unaware of exactly what data is being collected and how it's being used
        - They may understand their IoT thermostat is capturing temperature data, but not that it's also capturing IP addresses and geolocation data
    - Data Transmission
        - Ninety percent of data transmissions for IoT devices are unencrypted, which leaves it vulnerable to being stolen and yields for criminal purposes
    - Data Storage
        - Data might be stored on a local device or it might be transmitted elsewhere. If it isn't stored securely, then privacy can be compromised
        - Retention policies
            - If there is no clear end of life for that data, then it would continue to exist and be connected to the individual, even if the device is discarded
    - Data Access
        - People need to know who has access to the data, once it is collected and for what purpose. They gave access to the IoT device manufacturer, but that data might also be shared or sold without their knowledge

Privacy Protection through Individual Authorization

- Informed consent
    - Consent means voluntarily agreeing to what is occurring
    - The person must also have the right to withdraw consent at any time as easily as they gave consent in the first place
- Clear privacy policies
- Creative commons license
    - An actual legal policy written by lawyers
    - A human-readable layer that is understandable to the average consumer
    - A machine-readable version for software or search engine access and would only allow the technology access to information as directed by the consumer
- Opt-Out
    - The data will be collected about people unless they specifically say they do not want that
    - This has been the default tech privacy policy for decades
    - The burden of protecting privacy is on each individual
    - Companies sometimes bury the fact that opting out is possible

- Opt-In
    - This is where no data is collected until a person has given express permission

Privacy Protection through Data Management

- Direct identifiers can be used alone to recognize a specific person
- Indirect identifiers can be combined with other data to determine a unique individual
- Pseudonymization involves replacing direct identifiers such as real names with a temporary code

Privacy By Design

- Proactive, not reactive. Preventative, not remedial
    - Preventing privacy risks before they arise, rather than reacting to them afterward
- Privacy as the default setting
    - It states that no action is required by the user to safeguard their privacy. The system or business practice is designed to automatically focus on protecting privacy.
- Privacy embedded into design
    - Privacy protection is an essential component of the technical systems and the business practices for data-driven organizations, and that it is not an add-on after the product or service have been built
- Positive Sum not Zero Sum
    - Integrating privacy without compromising user experience or data security
- End-to-End Security – Full Data Lifecycle Protection
    - Privacy and security are inseparable. Safeguarding data throughout its entire lifecycle, from collection to deletion, is crucial for maintaining privacy
- Visibility and Transparency
    - That business practices and technologies are in line with objectives and independently verified to bolster confidence
- Respect User Privacy - Keep it User-Centric
    - Prioritizes the interests and needs of individuals, placing them at the core of its approach. Optimal outcomes are achieved when individuals have an active role in managing their personal data

Differential privacy

- A mathematical framework that protects the privacy of individuals in a dataset while still allowing organizations to use the data to gain useful insights. It's considered the standard for achieving privacy in data analysis

- Differential privacy works by adding a randomized element, or "noise", to the data being exposed. This means that even if someone tries to subtract the noise, they won't be able to get back the original individual data

Legal Concepts Related to Data-Driven Technologies

Legal Terminology: Responsibility, Accountability, and Liability

- Responsibility
  - The duty to take action
  - Can be shared
  - Taking responsibility is something you or a group chooses to do, not something an authority assigns to you
- Accountability
  - Some authority can hold you answerable for your actions after the task or situation is over
  - Cannot be shared
- Liability
  - Legal responsibility for one's actions and being accountable to the justice system

Technology Contract Types

- End User Legal Agreement or EULA
  - This contract is between the software company and the licensee
  - It establishes the users right to use the proprietary software in the specific ways described in the license terms
- Terms of Service Document
  - Legal agreement between a service provider and the person or organization who wants to use the service
  - It is designed to protect the provider, not the customer and includes a disclaimer limitation of liability
- Service Level Agreements or SLAs
  - Define exactly what one party will receive from another party.
  - External
    - Between a company and its customers
  - Internal
    - Between departments in an organization

Smart Contracts

- Self-executing contract where the terms of the agreement are written as code
- A smart contract is actually a computer program with if-then loops
- If the requirements of the agreement are met, then the contract will execute

- That code is stored inside a blockchain, which is a secure distributed network
- Each block contains three things; the data, a hash, and the hash of the prior block in the chain
- The data is a computer program, a hash is basically an encrypted code calculated when the data is added to the block, it is as unique as a fingerprint