JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Algorithms for Data Science

Data Collection

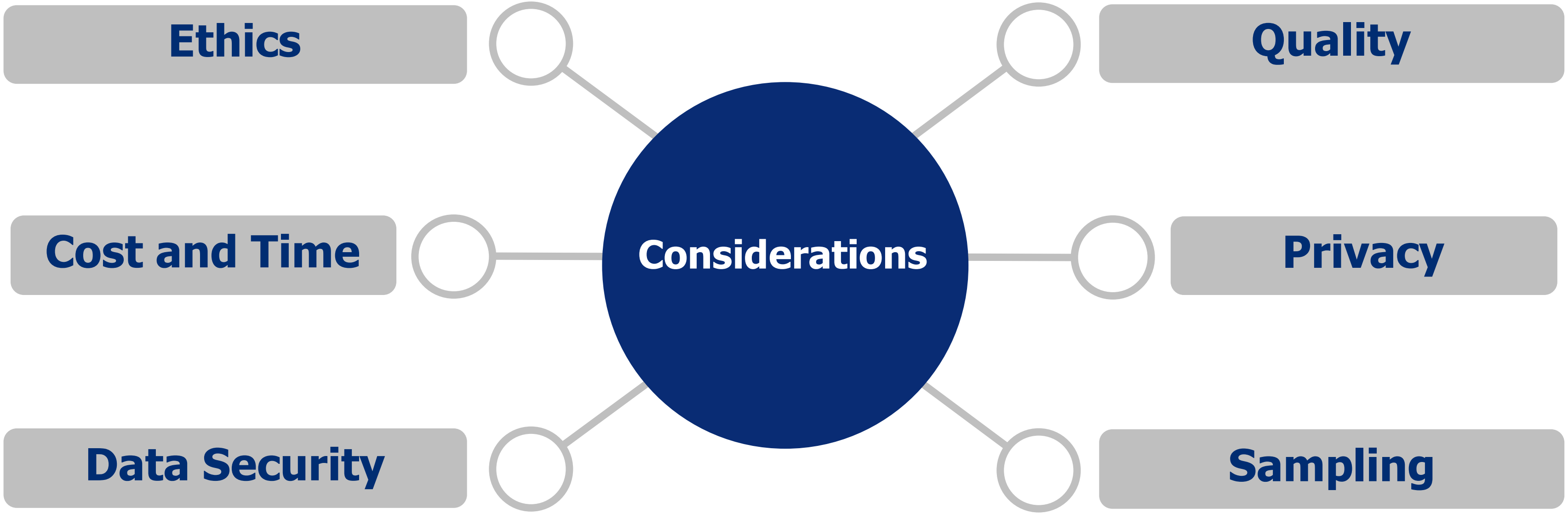
Data Collection

Involves the gathering of information from various sources to be used for analysis.

- **Primary Sources:** Data collected directly by the research for a specific purpose (e.g. surveys, observations, and experiments)

- **Secondary Sources:** Data collected by someone other than the user (e.g. government publications, online databases, and internal records of organizations

Data Collection Considerations



Considerations

Ethics

Cost and Time

Data Security

Quality

Privacy

Sampling

Data Collection Methods

Web Scraping Algorithms

- Web scraping is the process of extracting data from websites.
- Several algorithms and libraries, such as Beautiful Soup, Scrapy, and Selenium, are used to automate this process.

Data Collection Methods

Web Scraping Algorithms

- Web scraping is the process of extracting data from websites.
- Several algorithms and libraries, such as Beautiful Soup, Scrapy, and Selenium, are used to automate this process.

Sensor Data Collection

- IoT (Internet of Things)
- Algorithms used to optimize energy consumption, accuracy and reliability
- Algorithms such as aggregation and concensus

Data Collection Methods

Web Scraping Algorithms

- Web scraping is the process of extracting data from websites.
- Several algorithms and libraries, such as Beautiful Soup, Scrapy, and Selenium, are used to automate this process.

Sensor Data Collection

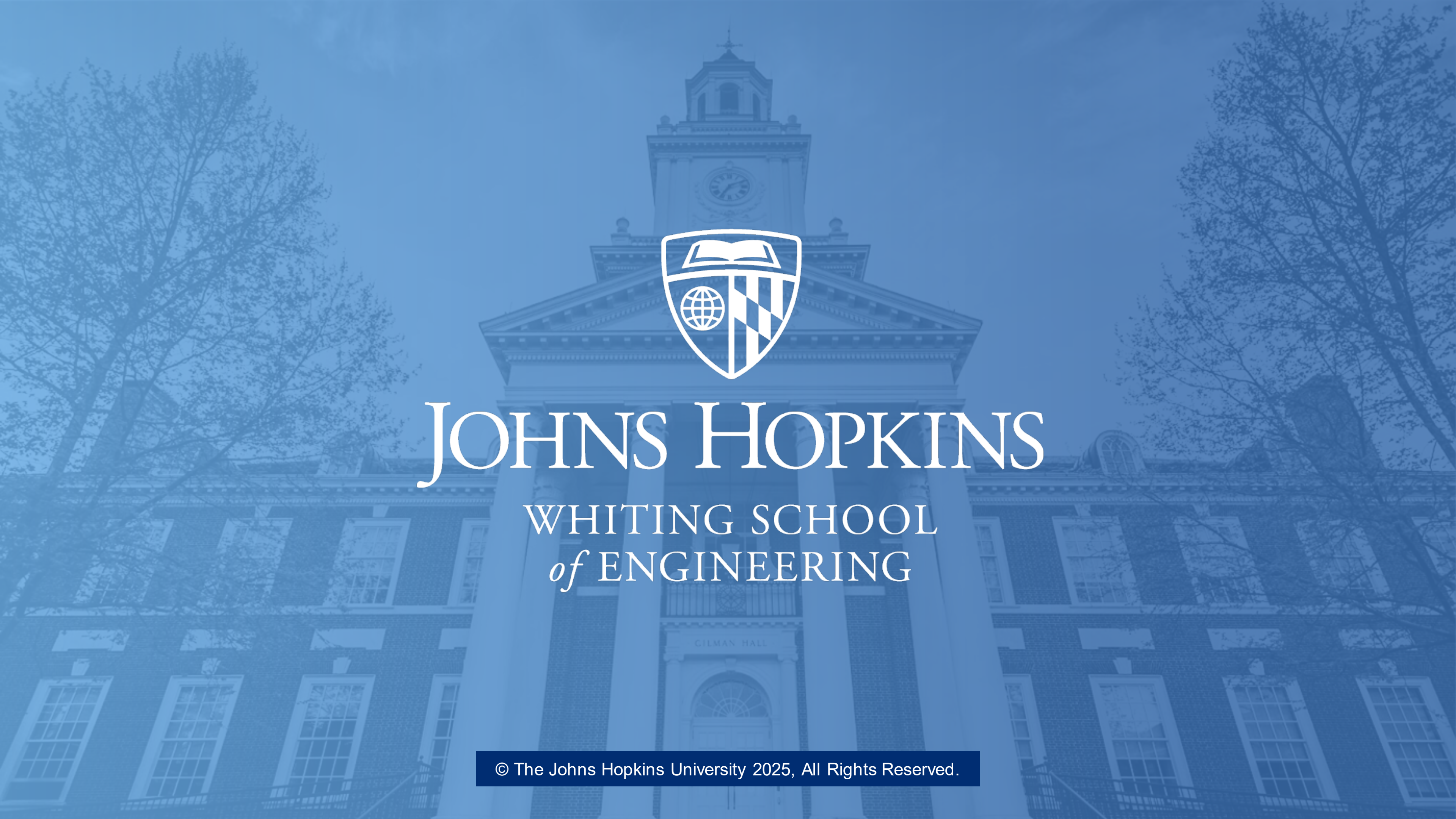
- IoT (Internet of Things)
- Algorithms used to optimize energy consumption, accuracy and reliability
- Algorithms such as aggregation and concensus

Stream Data Collection

- Real-time data streaming requires specific algorithms to handle continuous, fast-paced data inflow.
- Algorithms such as sliding-window algorithms are used to process and analyze data in real time.

Online Sources of Data

- **UCI KDD**: Repository of diverse datasets for machine learning and data mining research.
- **Kaggle**: Online platform offering datasets, competitions, and community resources for data science.
- **KD Nuggets**: Curated datasets and industry insights for data science and machine learning.
- **Data.gov**: U.S. government's open data portal with datasets across various sectors.



JOHNS HOPKINS

WHITING SCHOOL *of* ENGINEERING