02/08/2021

Data Creation, Data Capture, Data Classification,
Ethical Considerations, Data Validation, Data Privacy

Data created any time information is recorded/measured about the world.

1. Human generated (answering surveys, twitter posts, video, etc.)
2. Machine generated (satellites, sensors, log files, etc.) Internet of Things (IoT)
3. Organization generated (sales records, government, etc.)

Data creation used to be hard.

Steps to data capture

1. Decide what data to collect
2. What data capture tools to use?
    a. Organization and structure of data/files
    b. Data validation components
    c. Enable open and flexible formats, proprietary formats should be well-documented
    d. Allow the data to be moved with high quality
3. Collection process – documented, transparent, reproducible
4. Compliance with privacy regulations

How is data classified?

Privacy level – public, internal-only, confidential data, restricted
Content-based (sensitive information)
Context-based
User-based

Factors to consider:

Confidentiality
Integrity of the data – tends to require more storage space, variable accessibility
Availability
Data Type – some types of data require more storage
< text < images < videos

Methods of Validation

- Manual intervals – visually inspect all data "by hand" with people/lots of human hours
- Defined intervals – ex. Temperature data
- Equal intervals
- Quantiles
- Standard Deviation Intervals (identification of outliers)
- Natural Breaks
- Geometric Intervals
- Custom Ranges

Text-based analysis would have other kinds of checks.

Validation can focus on several types of factors:

- Data type
- Ranges
- Uniqueness
- Consistency
- Non-null

Ethical Data Collection

Institutional Review Boards on research data collection

Consent

Transparency

Accountability

Anonymity

Bias