Responsible Artificial Intelligence Institute

Data Quality Cheat Sheet

This Data Quality Cheat Sheet is meant to give a brief overview of key aspects of data quality and compliance.

Data Glossary

Dataset - A collection of data, published or curated by a single agent, and available for access or download in one or more formats. A dataset does not have to be available as a downloadable file. For example, a dataset that is available via an API (*https://en.wikipedia.org/wiki/Data_set*)

Data Quality - The state of qualitative or quantitative pieces of information. Data is generally considered high quality if it is "fit for [its] intended uses in operations, decision making and planning" (*https://en.wikipedia.org/wiki/Data_quality*)

Data Governance - the process of managing (based on internal data standards and policies) various dimensions such as usability, security, and integrity of the data used in systems

Data Dictionary *(metadata repository)* - A centralized repository of information about data such as meaning, relationships to other data, origin, usage, and format (*IBM Dictionary of Computing*)

Meta Data - a set of data that describes and gives information about other data (*Oxford Languages*)

Are you GDPR compliant for Data Quality and Protection?

- [ ] Conduct information audit - data types, who has access
- [ ] There is a legal justification for data processing
- [ ] Provide clarity and transparency about data processing occuring through a privacy policy
- [ ] Ensure data protection by design every time data is processed (encryption, data collection limits, data retention limits
- [ ] Encrypt, pseudonymize, or anonymize data when possible
- [ ] Create an internal security policy to inform and educate your team on data security
- [ ] Have a data production impact assessment and process to implement it
- [ ] Have a process to notify authorities and data subjects if a data breach occurs
- [ ] Have a designated person responsible for ensuring GDPR compliance across organization
- [ ] Have a data processing agreement between you and any third parties that process your org's personal data

see more info at https://gdpr.eu/checklist/

Is Your Data Structurally Biased?

ASK YOURSELF...

Is your data fully reflective of the population? Is there a sparsity of data for certain populations? ***(Response Bias)***

Is your model influencing the data used in training due to feedback loops? ***(Feedback Bias)***

Are there certain variables and attributes being omitted that influence the outcome? ***(Omitted Data)***

Are there societal biases embedded into the data potentially leading to discrimination? ***(Societal Bias)***

5 Tips for Dataset Quality

CONSISTENCY

- Common terms are used consistently throughout the datasets
- Similar data in different datasets should be the same.
- Standardized encodings for domain-specific entities (e.g. geospatial, finance, stats, etc.)

COMPLETENESS

- A dataset is considered complete when the data includes all the data that is necessary to support the application for which it is intended for and the data is updated regularly.
  - Provide clear metadata describing what the dataset is for
  - Informing the users of the frequency of updates to the dataset and when updates occur

PROCESSABILITY & TRANSPARENCY

- Convenient (open) structured format such as XML, JSON, CSV, etc. for machine-read
- Traceability: “The data sets and the processes that yield the AI system’s decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability”
- Provide clear information about data processing activities and conduct information audit

PRIVACY RIGHTS

- minimize use of sensitive data
- In cases where data contains personal, private, and secret information the data should be anonymized using best practice data-scrubbing pipelines.
- check if the AI models unintentionally memorize or expose these sensitive data

DATA SECURITY

- Encrypt, pseudonymize or anonymize data whenever possible - perform data protection from beginning to end of process
- Notify authorities and data subjects if a data breach occurs