Data Ingestion

Avoiding Risks Associated with Data Acquisition

Microsoft

Objectives

- How to determine resources needed to work in your Data Science problem?
- Walk-through of recommended data architectures.
- How determine the right data to use in your Data Science problem.
- How do you ensure that data is free of PII or other sensitive information?
- How to select right strategy to consume the data. Access Control and data consumption.
- How do you keep track of changes to the datasets?

Agenda

Data Composition

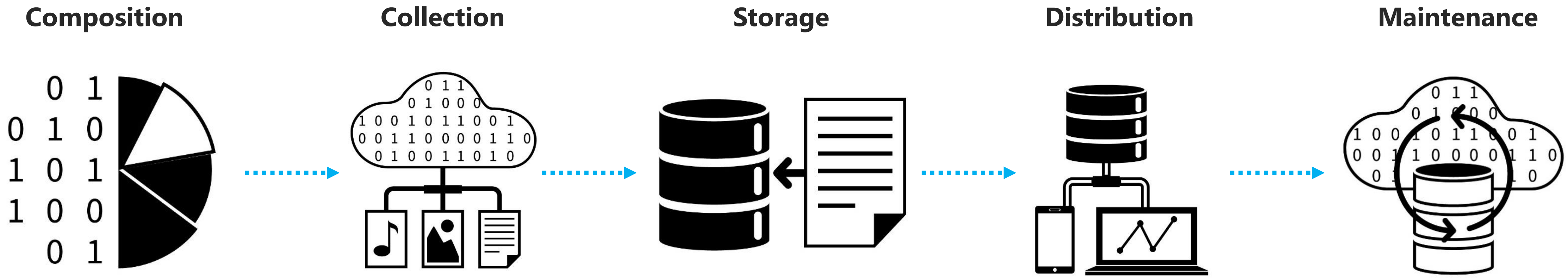
Data Collection Process

Data Storage

Data Distribution

Data Maintenance

Data Ingestion| Composition



Composition

Collection

Storage

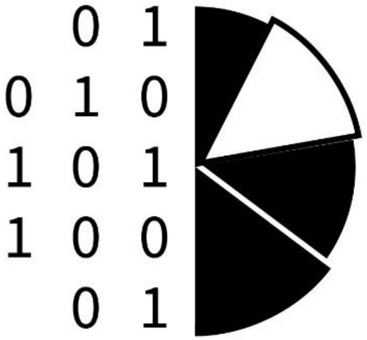
Distribution

Maintenance

Data Ingestion| Composition

Questions

Composition



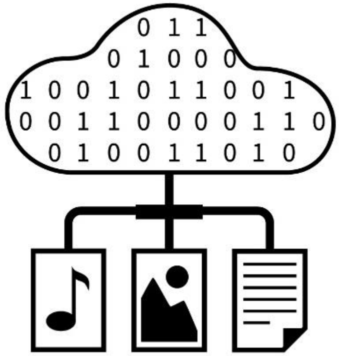
- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

- **What data does each instance consist of?** *Raw of pre-processed features?*

- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (*i.e., in combination with other data*) from the dataset?**

Data Ingestion| Collection Process

Collection Process



Questions

- **How was the data associated with each instance acquired?** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
- **What mechanisms or procedures were used to collect the data** *(e.g., hardware apparatus or sensor, manual human curation, software program, software API)?*
- **If the dataset is a sample from a larger set, what was the sampling strategy** *(e.g., deterministic, probabilistic with specific sampling probabilities)?*
- **Over what timeframe was the data collected?** *Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
- **Does the dataset relate to people?**
- **Were the individuals in question notified about the data collection?** *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

Data Ingestion| Storage

Questions

Storage

- **Where are you going to storage the data?**
- **What data does each instance consist of?** *Raw of pre-processed features?*
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly** (*i.e., in combination with other data*) **from the dataset?**

Data Ingestion| Collection Process

Questions

Distribution

- **How can this data be accessed?** *Share link to documentation file.*
- **Who has the rights to provide access to this dataset?** *Person or team responsible for approving access to dataset.*

- **How access to this dataset can be granted?** *Provide link to documentation on how people can access this dataset.*

Data Ingestion| Collection Process

Questions

Maintenance

- **Who is supporting/hosting/maintaining the dataset?**
- **What mechanisms or procedures were used to collect the data** *(e.g., hardware apparatus or sensor, manual human curation, software program, software API)?*
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*
- **Will older versions of the dataset continue to be supported/hosted/maintained?** *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

Data Science Questionnaire for Data Ingestion

Link:

Microsoft