Synthetic Data Generation

Giorgos Kosta

Research Engineer,
EuroCC 2, CYI

Agenda

- Why Synthetic Data
- Frameworks and Tools
- Challenges & Limitations
- Best Practices

Why Synthetic Data?

- **Data Scarcity:**
  - Limited real-world data, especially for rare events or new use cases.
- **Privacy Compliance**:
  - GDPR, CCPA, and other regulations.
- **Cost Efficiency:**
  - Collecting and annotating real-world data is expensive.

Specialized Frameworks:

- Synthetic Data Vault (SDV):
  - Python library for generating synthetic tabular data while preserving statistical relationships
- Fabricate:
  - LLM based tool with predefined common fields (Email, Postal Codes, Airport Codes, etc)
- Mimesis and Faker:
  - Python libraries like Fabricate that can also anonymize data taken from a real dataset
- LLM call with custom output format using pydantic BaseModels

Challenges & Limitations

- **Quality Control**:
  - Generated data may not perfectly match real-world distributions
  - Need for validation mechanisms to ensure data quality
  - Challenge in maintaining complex relationships between fields
- **Consistency Issues:**
  - LLMs may generate contradictory or impossible combinations
- **Cost Considerations**:
  - API costs can accumulate quickly for large dataset generation

Best Practices

- Use real data to guide synthetic data generation (e.g use examples when prompting an LLM)
- Verify no sensitive information leakage - Implement differential privacy
  - Audit your generated data.

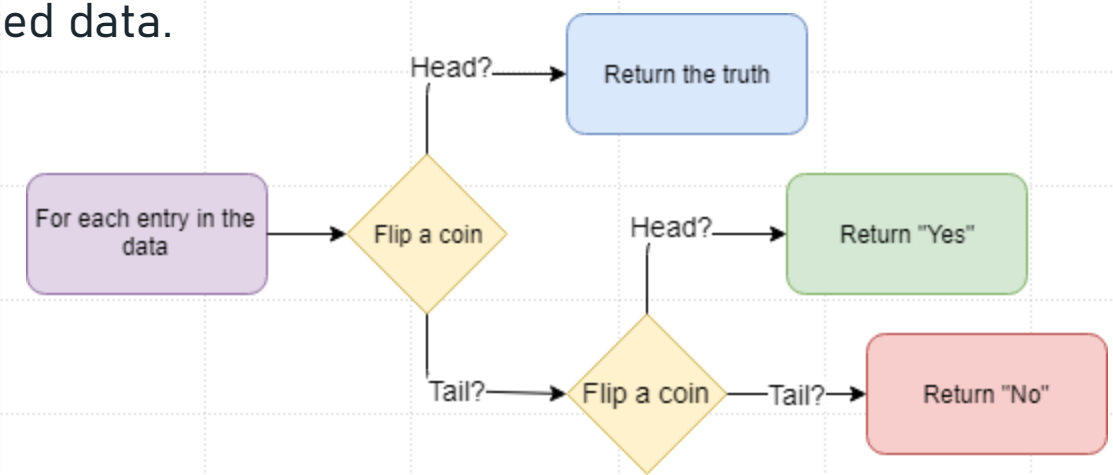

Fig 1 Differential Privacy Example

Source: Fig 1 : https://medium.com/towards-data-science/understanding-differential-privacy-85ce191e198a

Examples:

- Hands-on:
  - Fabricate
  - Simple LLM example with local model running on Ollama