DATA GATHERING

Data Mining, Acquisition and Preparation

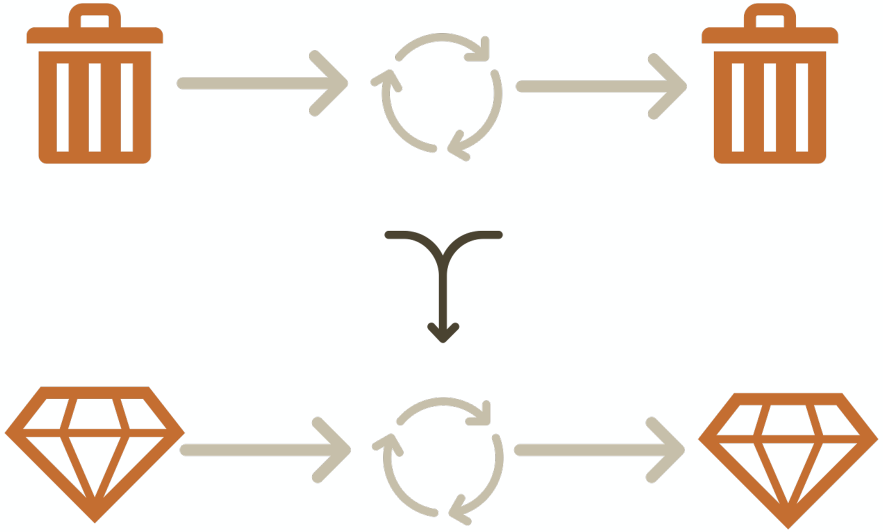
Anisha Roy
Parimala Sabau

WHAT IS DATA GATHERING?

Process of collecting and measuring information from countless different sources (in a systematic and defined manner).

Accuracy, honesty, and integrity are crucial.

RECALLING
GARBAGE IN GARBAGE OUT



THINGS TO CONSIDER:

Identifying the Data Needs

Sources for Data Collection

Data Storage

Time Requirements For Other Tasks

TYPES OF DATA

FIRST-PARTY DATA

Obtained directly from the source. Greater accuracy and reliability and less restrictions. Examples: sensors, images, interviews, surveys, customer purchase history, U.S. Census, etc.

SECOND-PARTY DATA

Relying on a trusted source to collect and provide data. Reliability and accuracy as with first-party data but more restrictions. Examples: Hotels.com purchasing Austrian Airlines' first-party data.

THIRD-PARTY DATA

Data from an external source with no direct relationship. Aggregated data from various sources. Low reliability and accuracy. Example: datasets from the web

KEY ACTIONS IN THE DATA COLLECTION PHASE

LABELLING

Labeled data is the raw data that was processed by adding one or more meaningful tags so that a model can learn from it.

INGEST & AGGREGATE

Incorporating and combining data from many data sources is part of data collection in ML.

METHODS FOR DATA GATHERING

Manual:

- Surveys and feedback
- Interviews
- Focus groups
- Direct observation of participants

Technology-Assisted:

- Crowdsourcing
- Relational databases
- APIs
- Web Scraping
- Public Datasets (i.e. Kaggle)
- Synthetic Data Generators (i.e. Mostly.ai)

DISCOVER NEW DATA AND SHARE IT

COLLABORATIVE ANALYSIS

Data Hub

WEB

ckan

Quandl

webz.io

Google Scholar

COLLABORATIVE & WEB

kaggle

GOOGLE SCHOLAR

Google web search engine specifically for academic & scholary literature and research.

Search globally across multidisciplinary sources.

Features:

- Search all scholarly literature from one convenient place;
- Explore related works, citations, authors, and publications;
- Locate the complete document through your library or on the web;
- Keep up with recent developments in any area of research;
- Check who's citing your publications, create a public author profile;

Great resource and starting place, but not a comprehensive, one-stop-shop.

KAGGLE

PUBLIC DATASETS

OVER 50,000 DATASETS. EXPLORE, ANALYZE, & SHARE.

KAGGLE (JUPYTER) NOTEBOOK

EXPLORE & RUN MACHINE LEARNING CODE

MACHINE LEARNING MODELS

ALREADY TRAINING AND READY-TO-DEPLOY

LEARN & DISCUSS

Designed with the help and input of experts in the field

DATASETS FOR MACHINE LEARNING

DATASETS BASED ON TYPES OF DATA:

Computer Vision: Images & Videos | NLP: Text & Audio

- OpenImages | The NLP Index

DATASETS BASED ON TYPES OF ML TASKS:

Facial Recognition | Bounding Boxes | Image Classification | OCR | NER | Speech Recognition | Sentiment Analysis | Chatbots

- VoxCeleb

OTHER RESOURCES FOR ML FROM THE WEB

Carnegie Mellon University
Libraries

Google Dataset Search

aws

GitHub

Microsoft Azure

PYTHON LIBRARIES FOR DATA COLLECTION

WEB SCRAPING

BeautifulSoup, Scrapy (web scraping + web crawling), Selenium(for dynamic content)

API

requests (API calls), json (retrieving data)

DATABASES

SQLAlchemy, PyMongo - Connecting to relational databases

REAL-TIME DATA STREAMS

Apache Kafka + PySpark

WEB SCRAPING

Automated method of extracting data from websites to be converted into a structured format to be used as necessary.

Web Crawling + Web Scraping: powerful data mining & analysis tool.

Example Use Cases:

- Scraping social media data to be used for sentiment analysis in customer feedback;
- Scraping product data from Amazon to identify trends in pricing & product features;
- Scraping job posting from various platforms to find trends in the current job market;

Scrape ethically, responsibly, and legally!

WEB SCRAPING - ETHICAL & LEGAL IMPLICATIONS

Privacy Violations

IP Infringement

Cybersecurity Risks

Unfair Advantage

HOW TO MITIGATE RISKS?

Adhere to ethical guidelines, laws, and regulations (i.e. GDPR)

Respect site's Terms of Service & "robots.txt" file

User consent before scraping personal data

Avoid copyrighted material

Prevent harm to website and its users

DATA GENERATORS

MOSTLY·AI

AI-powered where each generated dataset comes with a QA report.

After uploading a data sample, the generator can create statistically and structurally identical synthetic versions of the original.

AVO iTDM Intelligent Test Data Management

Test data management platform that empowers you to generate production-like, AI-ML-based test data with a few clicks. With reliable, relevant, and compliant data quickly available, you can expedite testing and be 100% sure of higher quality.

DATPROF

It simplifies getting the right test data at the right moment. You can mask your test data and generate synthetic data. Customer data is protected, but software teams can still use representative test data.

ISSUES THAT MIGHT OCCUR DURING ML DATA COLLECTION

BIAS

INACCURATE DATA

MISSING DATA

DATA IMBALANCE

CONSEQUENCES FROM IMPROPERLY COLLECTED DATA

Inability to produce viable results

Distorted findings resulting in wasted resources

Inability to repeat and validate

Compromising decisions for public policy

DATA IS COLLECTED: WHAT'S NEXT?

Clean It + Exploratory Analysis

Analyze It

Share it

Embrace your failures

Github Repo: https://github.com/anisharoy0304/Data-Gathering-SS23DMAP

THANK YOU!