Embedded ML

Data Engineering

Dr. Venki

Adopted from Edge Impulse

Data Engineering

Embedded ML

Data engineering involves the processes and techniques for collecting, processing, and transforming raw data into a structured and usable format for analysis and decision-making.

- Problem definition
- Permissions & rights
- Machine & human usable format

- Data sources
- People
- Collection
- Labeling

- Processing
- Validation
- Augmentation

- Storage
- Security
- Errors
- Versioning

Data

Embedded ML

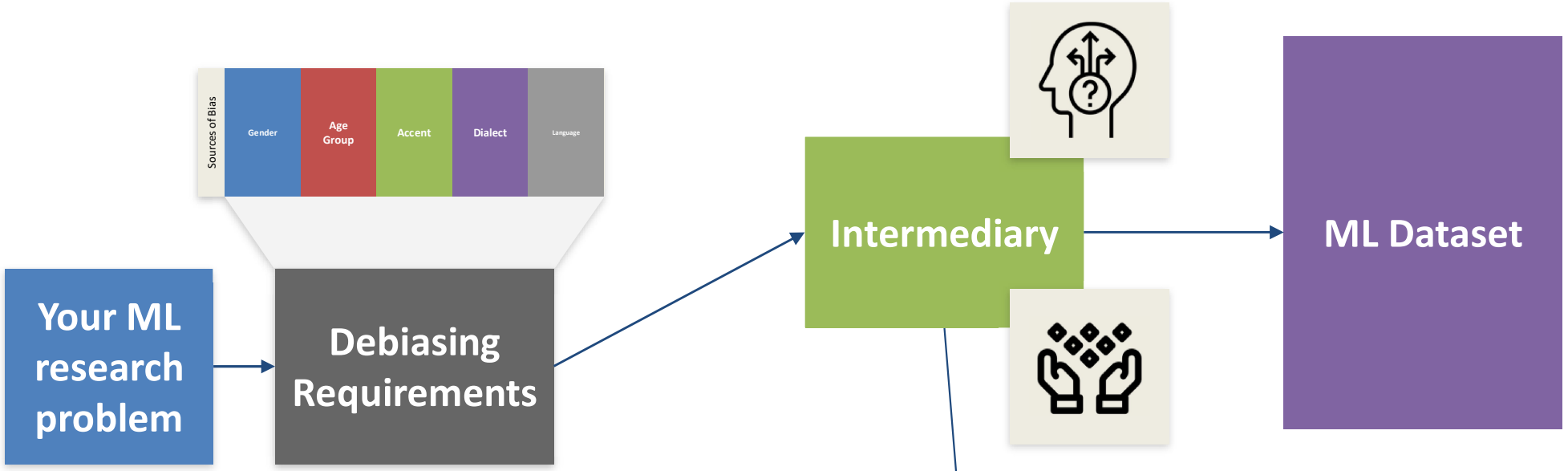
Data Collection:

Acquiring raw data from various sources, such as databases, APIs, files, or external systems.

- How will you verify the data you collected?
  - Manually (time, cost)
  - Automation
  - Domain expertise
    - disputes / disagreements
- Your dataset will evolve
  - Missing demographics?
  - Expanding your user-base?

- Data **isn't free** to use
  - Open?
  - Copyrighted?
  - Licensed?
  - Product users?

Bias and Market Forces



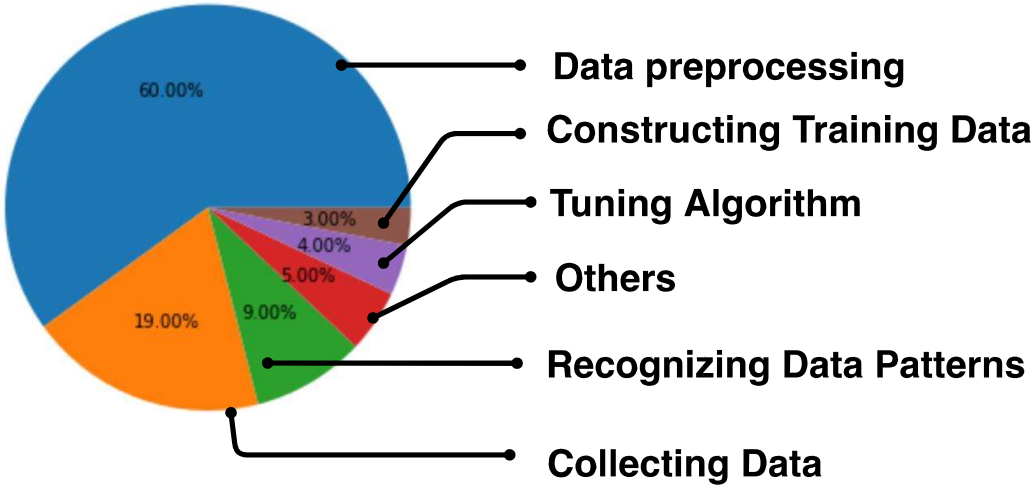
Intermediaries: help with collection but a ***potential source of bias***

Data Ingestion

- Loading raw data into a storage system or data warehouse for further processing.

Data pre-processing

- Data scientists spend more than 50% of their time on Data preprocessing.
- Collecting data is the second most time-consuming component.
- Tuning algorithm occupies a small part.



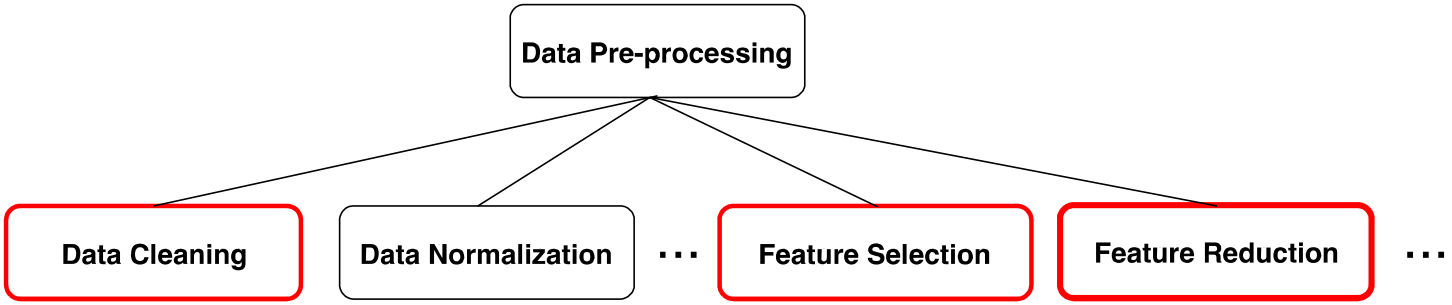
Why data pre-processing?

- **Incomplete**: data lacks attributes or contains missing values.
- **Noisy**: data contains incorrect records such as outliers.
- **Inconsistent**: data contains conflicting records or discrepancies.
- **Missing values:** some attributes in the collected data would have blank or NULL values.
- **Invalid Values:** some well-know attributes such as gender may have incorrect values.
- **Uniqueness**: repeated values of the same identifiers.
- **Misspellings:** incorrectly written values

Goal of Data Pre-processing

- To avoid "Garbage in, Garbage out" for data analytics

- Effects of data-processing



Feature selection

Embedded ML

- Feature selection
  - process of choosing a subset of relevant features from the original feature set.
    - improve model performance, reduce overfitting, and enhance interpretability by eliminating irrelevant or redundant features.

- Feature Reduction:
  - transforming the original set of features into a lower-dimensional representation.
    - maintain as much information as possible while reducing the number of features.
    - Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), t-Distributed Stochastic Neighbor Embedding (t-SNE).

Feature modeling

- process of designing and selecting relevant features for use in a machine learning.
- create a set of features that effectively represents the underlying patterns or characteristics of the data.
- Apply signal pre-processing
  - Noise removal & smoothing
- Feature Extraction:
  - Time-Domain Features: Calculate statistical measures over time, such as mean, standard deviation, skewness, kurtosis, etc.
  - Frequency-Domain Features: Use techniques like Fast Fourier Transform (FFT) to extract frequency-domain information.
  - Signal Magnitude Area (SMA): Calculate the total area under the signal curve as a feature.
  - Zero Crossing Rate: Count the number of times the signal crosses zero.
  - Energy Features: Compute signal energy in different frequency bands.
- Apply windowing, aggregation, and scaling.

How to deal with missing values?

Embedded ML

- Typical methods to process missing values:
  - Deletion
  - Dummy substitution
  - Mean substitution
  - Frequent substitution
  - Regression substitution

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol		3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

Source: https://gallery.azure.ai/Experiment/Methods-for-handling-missing-values-1

Outliers

- Outliers can be very common in multidimensional data.
- Outliers can be results of bad data collection.
- Outliers would distort the models.
- Some models are sensitive to outliers.
- Sometimes outliers are the interesting data points.

- **Keep outliers**
  - We should pay more attention to the outliers since they may be genuine observations in the collected data.
  - In many applications, outliers provide crucial information for data analytics.

Exclude outliers

- Trimming: discarding the outliers.
- Replacement: replacing the outliers with the nearest "normal" data point

Data Normalization

Embedded ML

- Data normalization is a process in which data attributes within a data model are organized to increase the cohesion of entity types.
- Data normalization is the rescaling of numerical values to a specific range.
- Two popular methods:

Min-Max Normalization:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Z-score Normalization (or Standardization)

$$\boldsymbol{X_{Z-score}} = \frac{\boldsymbol{X - \mu}}{\boldsymbol{\delta}}$$

$\boldsymbol{\mu}$: mean of X

δ: standard deviation of X

Data Down-Sampling

- Data down-sampling is reducing large data to a smaller and more manageable size.
  - **Record down-sampling (clustering):** select the records and only choose the representative subset from the data.
  - **Attribute down-sampling (Feature selection):** select only a subset of the most important attributes from the data.

Data Cleaners

Embedded ML

- Data Wrangler
  - Wrangler is an interactive tool for data cleaning and transformation.
    Spend less time formatting and more time analyzing your data.
  - http://vis.stanford.edu/wrangler/
- Open Refine
  - OpenRefine is a powerful free, open-source tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.
  - https://openrefine.org/

Data Annotation

- Data annotation is the labelling of collected data for various applications.
- Examples of data annotation

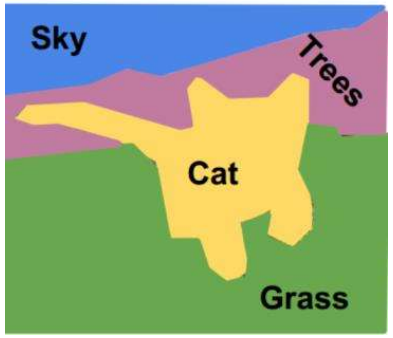
Image Classification

CAT

Object Detection

DOG, DOG, CAT

Semantic Segmentation



Data Quality (DQ)

Embedded ML

- Data quality refers to the state of qualitative or quantitative pieces of information.
- Data is of high quality if it is “Fit for Use” in their intended operational, decision-making and other roles.
- Quality Properties
  - Relevance – does data meet basic needs?
  - Accuracy – are key data elements correct?
  - Timeliness – is data current?
  - Comparability - Can several databases be combined into a data warehouse?
  - Completeness - No missing records, no missing data elements

Standards in Sensor Data Collection

- SensorML (Sensor Model Language): An OGC (Open Geospatial Consortium) standard for describing sensors and sensor data processing.

- IEEE 1451: A set of standards for smart sensor systems, including communication protocols and Transducer Electronic Data Sheets (TEDS) for sensor metadata.

- OMI (Observations and Measurements): An OGC standard for encoding observations, which is applicable to sensor data.