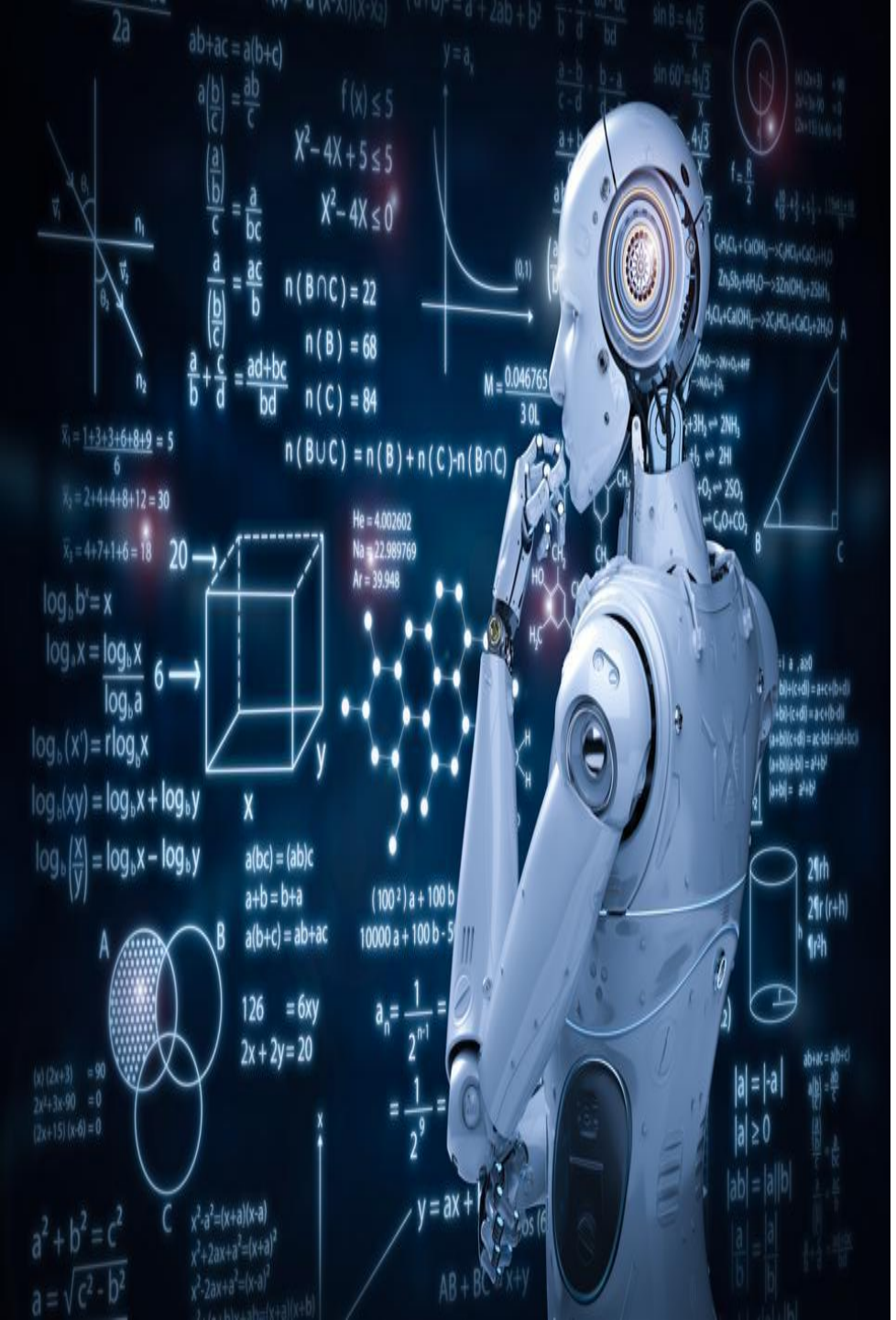


Practical Machine Learning

Day 7: Mar22 DBDA

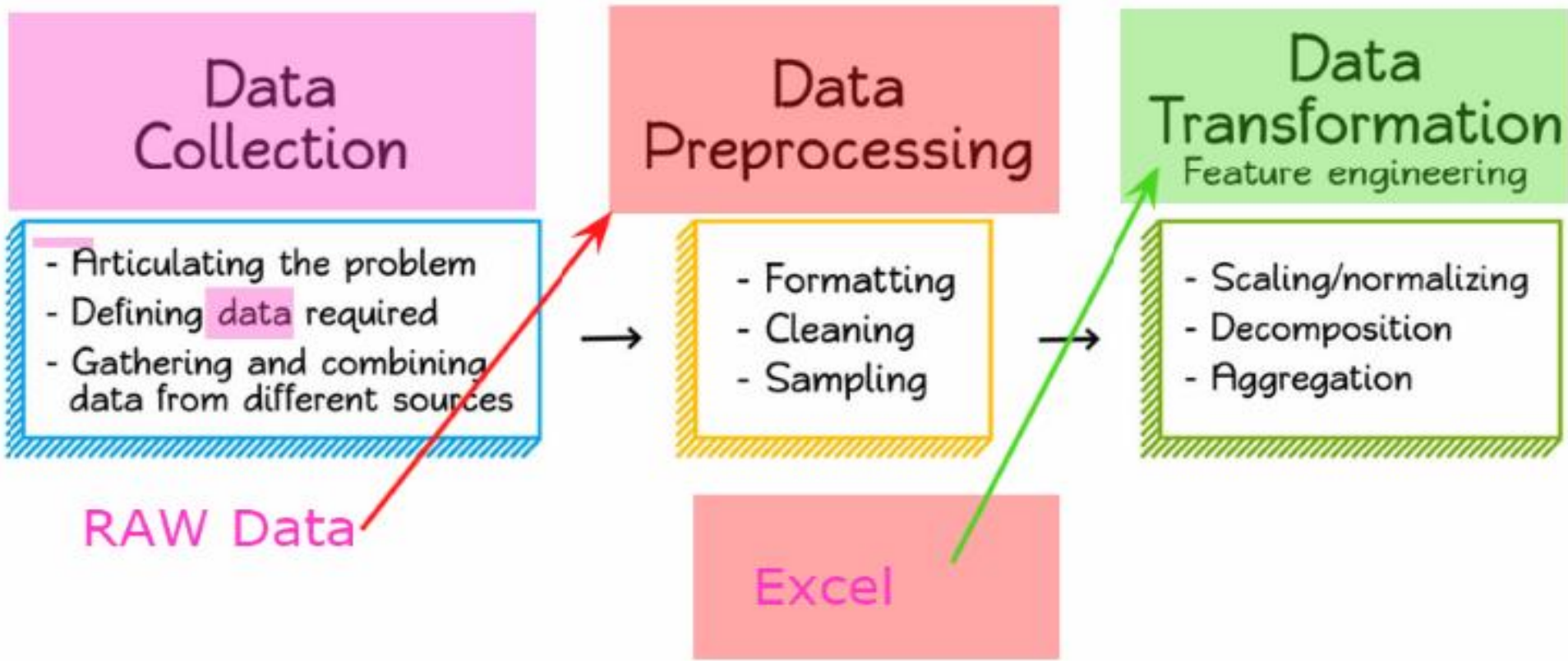
Kiran Waghmare

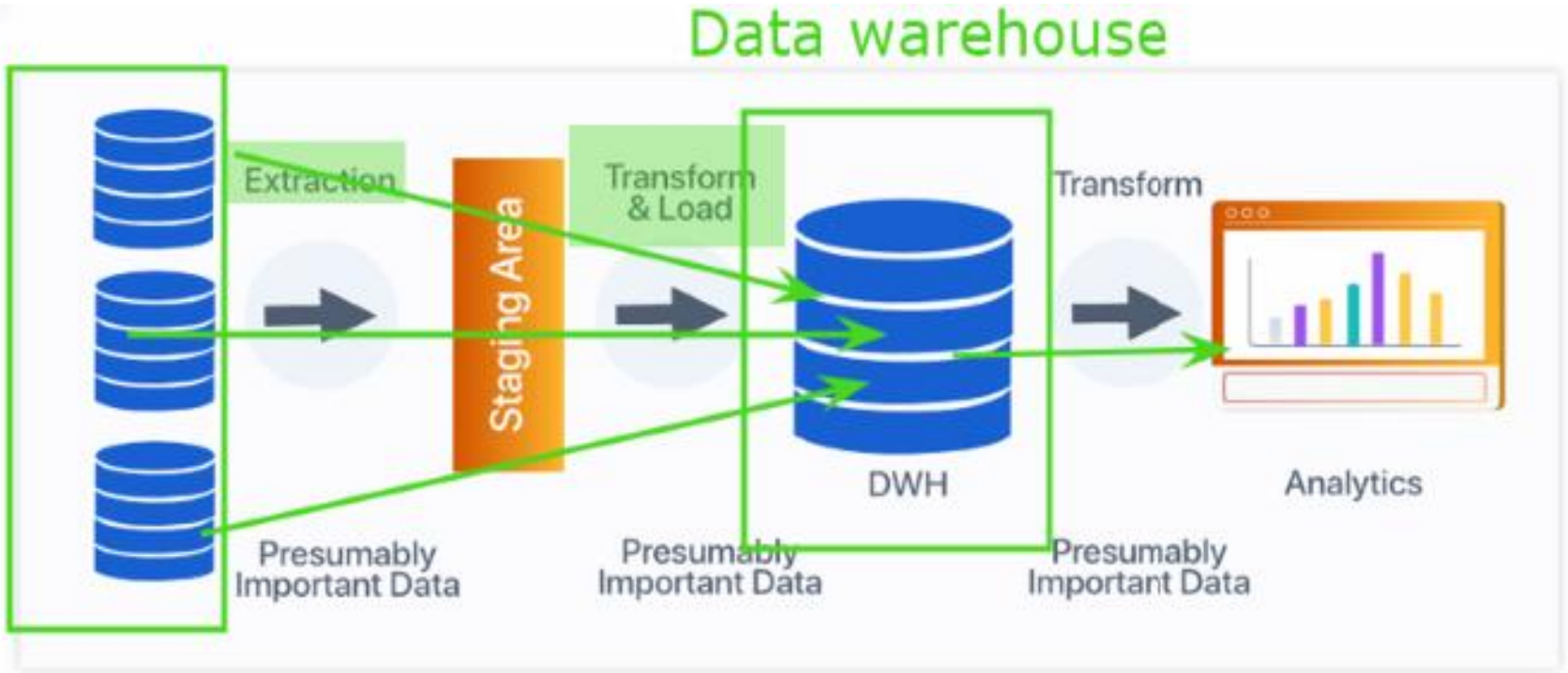
Agenda

- Preprocessing Techniques

Data Pre-processing

Data Preparation Process





Data Preprocessing and Data Wrangling

in Machine Learning

Why preprocess the data?

- Data in the real world is *Dirty*…
  - Incomplete Data: Lacking attribute values, Lacking certain attributes of interest, or containing only aggregate data

    e.g. Occupation=" ", year_salary = "13.000", …
  - Inconsistent Data: Containing discrepancies in codes or names

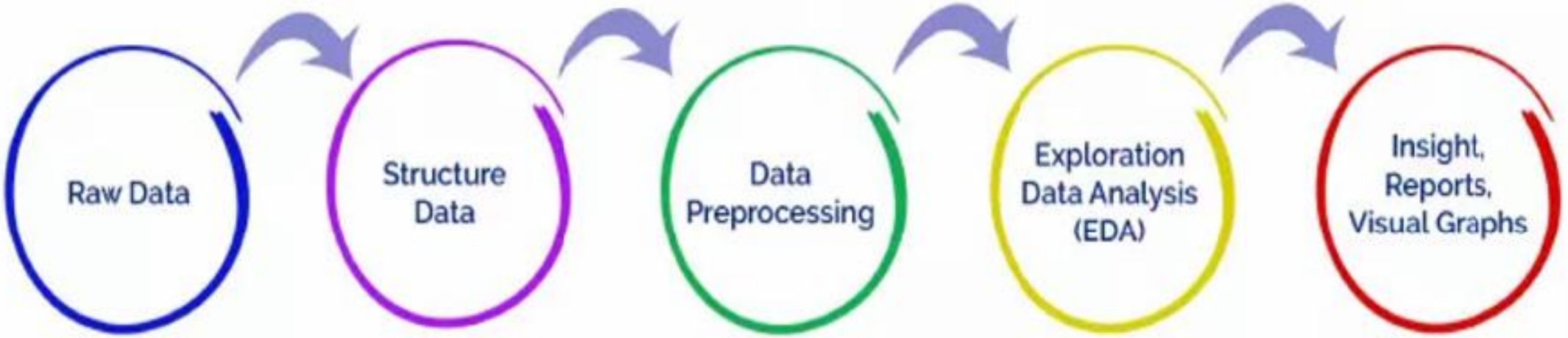
    e.g. Age="42" Birthday="03/07/1997"

    Previous rating "1,2,3", Present rating "A, B, C"

    Discrepancy between duplicate records
  - Noisy Data: Containing errors or outliers

    e.g. Salary="-10", Family="Unknown", …

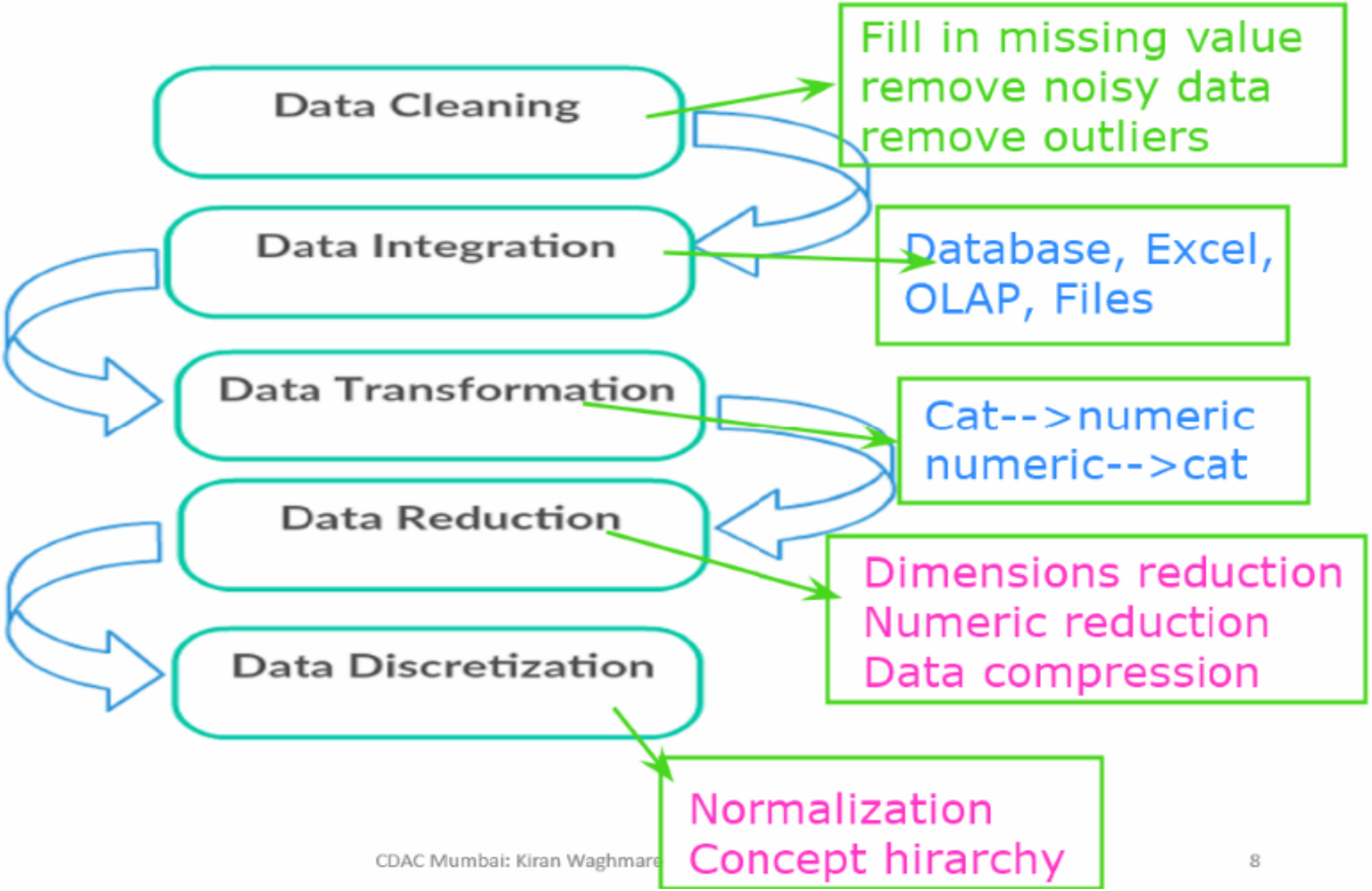
Data Preparation



MEasures of Data Quality:
Accuracy,Completeness, Consistency,Timeliness,

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

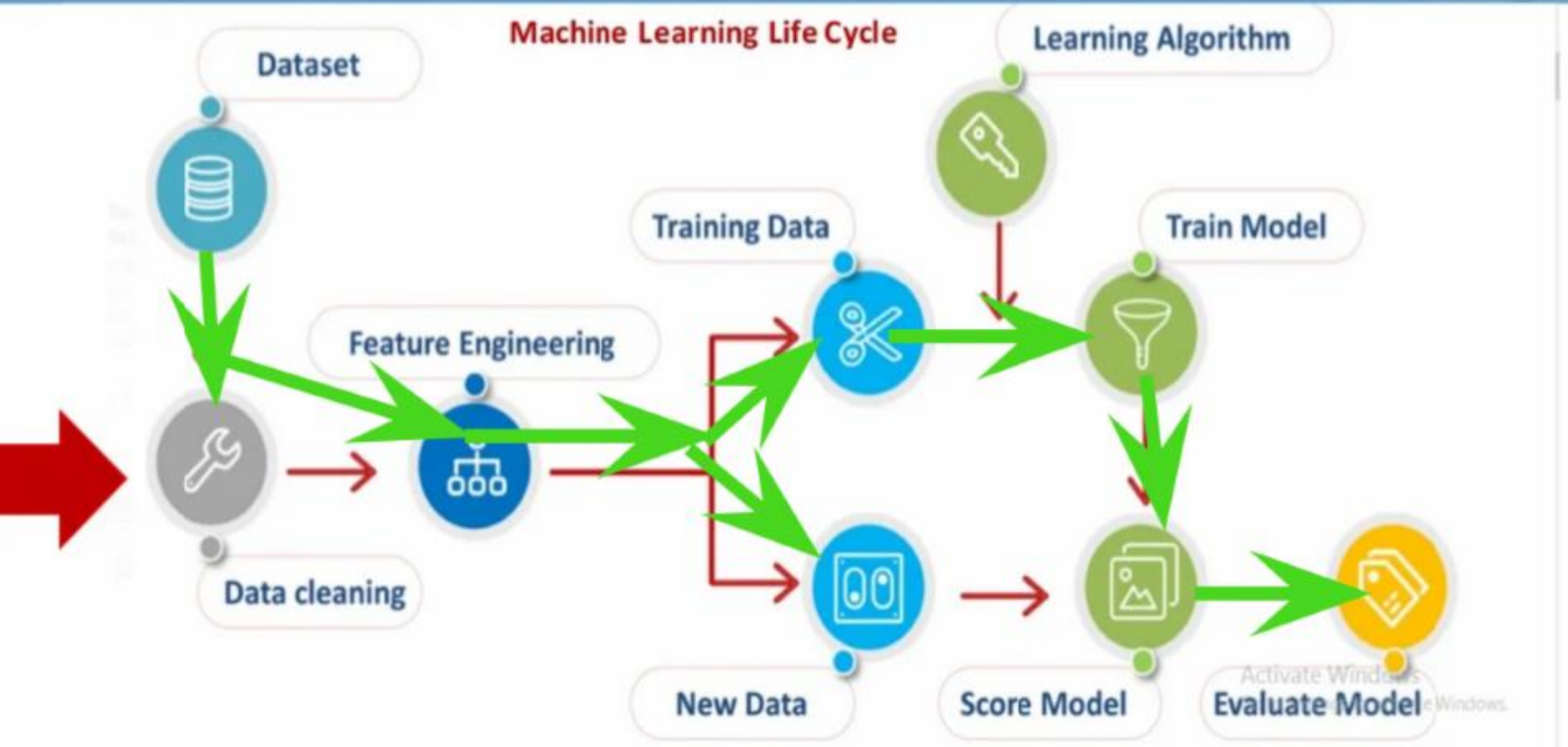


CDAC Mumbai: Kiran Waghmare 8

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Where is Data Cleaning used?



Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=" " (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., *Salary*="−10" (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age*="42", *Birthday*="03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15 | 21, 21, 24, 25 | 26, 28, 29, 34

9 23 29

  - Partition into equal-frequency (equi-depth) bins:
    - Bin 1: 4, 8, 9, 15
    - Bin 2: 21, 21, 24, 25
    - Bin 3: 26, 28, 29, 34
  - Smoothing by bin means:
    - Bin 1: 9, 9, 9, 9
    - Bin 2: 23, 23, 23, 23
    - Bin 3: 29, 29, 29, 29
  - Smoothing by bin boundaries:
    - Bin 1: 4, 4, 4, 15(boundaries 4 and 15, report closest boundary)
    - Bin 2: 21, 21, 25, 25
    - Bin 3: 26, 26, 26, 34

How to handle noisy data: **Regression**

Handle Noisy Data: Cluster Analysis