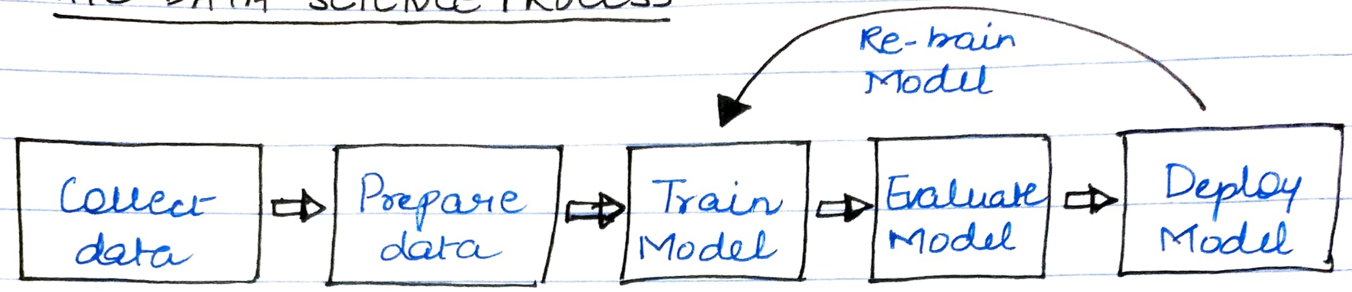
9/07/20

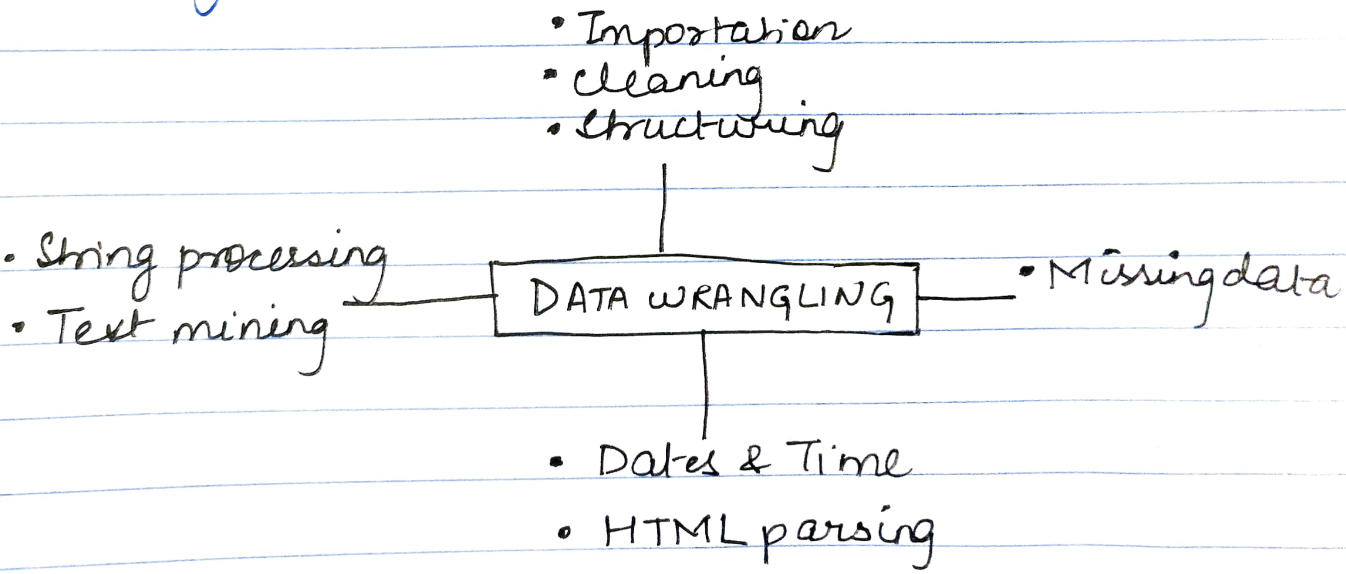
5. THE DATA SCIENCE PROCESS



II) Prepare data ⇒ Performing data wrangling (referred to as data munging - is a process of transforming & mapping data from one "raw" data form into another format with the intent of making it more appropriate & valuable for a variety of downstream purposes eg analytics.

Involves processes:

1) cleaning, 2) structuring, 3) enriching new data into desired format for better decision making in less time.)



∴ in data preparation → we identify or create the features needed for the model.

III) Train Model - 1) select an algorithm

2) prepare our training, testing & validation datasets

3) Iteratively evaluate the model to identify the best-performing version & sanity check the outcome.

IV) Evaluate Model - Run the model through a final exam using data from our validation data-set & see how it performs.

V) Deploy Model - Pack into the model in-dependencies up for the deployment, use it within our web-service or within an API in an application & measure the ongoing performance of the model.

VI) Retrain Model (Last step) - it is an iterative step for models in production.

Developers perspective

1) collect data - write code

2) Prepare data - Write queries & code

3) Train model - Write code, do some math ⇒ feature vectorization, feature scaling & tuning the ML algo.

4) Evaluate model - computing evaluation metrics or evaluation graphs on the test data ~~sets~~.

5) Deploy Model - DevOps → involve training, evaluation & deployment scripts in respective build & release pipelines.

↦ means you can access any version of the product.

/* make sure all models & deployments are versioned & artifacts are archived */

6. COMMON TYPES OF DATA

- Numerical – integers or floats
  eg identifiers of items, different properties like sales amount, house prices.

☆ All data in ML eventually ends up being numerical data, whether its numerical in its original form or processed from other more complex structured forms like image, speech, or text.

- Time-series – series of numerical values that can be ordered, typically data collected over equally spaced points in time, but it also can be data that is ordered based on a non-date-time column. (// numerical data pts. across pts in time)
  Eg → real-time stock performances, energy demand forecasting, ~~speech data~~

- Categorical – includes discrete & limited set of values. eg gender, ethnicity, location ID
  ↳ less ~~low~~ imp. ↳ high imp.

- Text – words, sentences eg newspaper text.

- Image – transform into appropriate numeric form

μ = mean
σ = S.D.
$x_{max} - x_{min}$ = range

TABULAR DATA - most common in ML.

- Row - an item | single observation.
- Column - property
- Cell - individual data point | single value in a row or column

- Column values can be continuous or discrete (categorical)

Discrete → eg maker (Brand), color

Continuous → Price, quantity
↳ scaling of input data is done.

☆ Vectors

In ML, we ultimately always work with numbers or specifically vectors.

A vector is simply an array of numbers, eg (1, 2, 3) - or a nested array that contains other arrays of numbers → (1, 2; (1, 2, 3)).

☆ All non-numerical data types (eg images, text, & categories) must eventually be represented as numbers.

SCALING DATA - 2 methods:

Standardization

Rescales data to have mean = 0 & σ = 1 (S.D.)

$$\boxed{\frac{x-\mu}{\sigma}}$$

Normalization

Rescales data into range [0, 1].

$$\boxed{\frac{x - x_{min}}{x_{max} - x_{min}}}$$