Rules/guidelines/cases for assigning labels to attribute records

Usable directly numeric:

Case a. Should be usable directly as a number feature for ML (e.g., age, income, review rating, and satisfaction score) and is not any of cases b, e, g, i, j, k, l, or n below. Note that ID attributes (e.g., Person ID), serial numbers, and integers representing encodings of discrete valuations (e.g., education level) do NOT belong to this case because using them directly as numbers for ML will give garbage results! Thus, you should rule out the other cases (b, e, g, i, j, k, l, and n) below before choosing this answer.

Usable with extraction:

Case b. A number present along with unit of measure string, e.g., ‘30 Mhz’, ‘30 degree’, ‘45 mm’ ,’500,000’

Case c. A text corpus with semantic meaning, URL, address, list of items in a single sample separated by symbols, e.g., review text, remarks text, and department list such as ‘{men|clothing, women|clothing, children|toys etc}’

Case d. Date or time stamp, e.g., ‘7/11/2018’, and ‘21hrs:15min:3sec’

Usable directly categorical:

Case e. Yes/No type values, including binary 0/1 answers

Case f. Country names, city names, food type names, and other object type names that are not cases l or m below

Case g. Coded numbers that are short forms of names in case f that are not cases l or m below

Case h. Short names that indicate type from a known finite set/domain that are not case l below, e.g., type of funding from the set {“seed funding”, “private equity”, etc.}, gender from the set {“male”, “female”, “other”, etc.}, and job title from the set {“politician”, “actor”, “scientist”, etc.}

Case i. Handful of coded numbers that repeat themselves but arbitrary arithmetic on them is not meaningful and that are also not case l or n below, e.g., education level with values 0-4, birth year, and exercise intensity with values 1-3

Case j. A coded number that encodes real-world entities from a known finite/ domain set (possibly representing names) and that are not cases l, m, or n below, e.g., product id, department number, and zipcode

Unusable:

Case k. A number indicating the position of a record in its dataset table, e.g., serial number

Case l. An attribute that is likely the primary key in its dataset table, or an attribute whose values will almost surely be unique for all records in its dataset table but is not a numeric feature, e.g., SSN, employee id, and product id (when it is known to be unique across records in its dataset table)

Context dependent:

Case m. Person name, company name, or any entity name that is not generic, e.g., ‘Amazon Inc.’, Facebook LLC’, etc.; this includes almost all proper nouns that are not cases f, h, k or l above.

Case n. Coded numbers or id for people, company, or other entity names from case m that are not cases g, i, j, k, or l above.