Syntactic Profiling

Name: Pyous Pradhan | Class: III B.S Mathematics | Subject: Data Analysis and Visualization | Subject Code: UDSC-402

What is Syntactic Profiling?

A category of individual value profiling focused strictly on the formatting and literal constraints of a specific data field — i.e., whether a value's *form* is valid, independent of whether it makes real-world sense.

Types of Syntactic Constraints

- **Valid sets:** small, listable domains, e.g. Boolean bits {0, 1} or birth sex {male, female}.
- **Valid ranges:** larger numeric bounds, e.g. lifetime ATM withdrawal restricted to integers 0–50000.
- **Pattern/regex constraints:** fields validated against a format pattern rather than a fixed list, e.g. emails, phone numbers, SSNs, ZIP codes, or date formats (YYYY-MM-DD vs MM/DD/YYYY).
- **Data type conformance:** checking a value is stored as the correct type, e.g. a numeric column holding stray text like “N/A”.
- **Length & precision constraints:** fixed-length fields (2-character state code, 9-digit SSN) or decimal precision limits (currency to 2 decimal places).

Role of Data Dictionaries

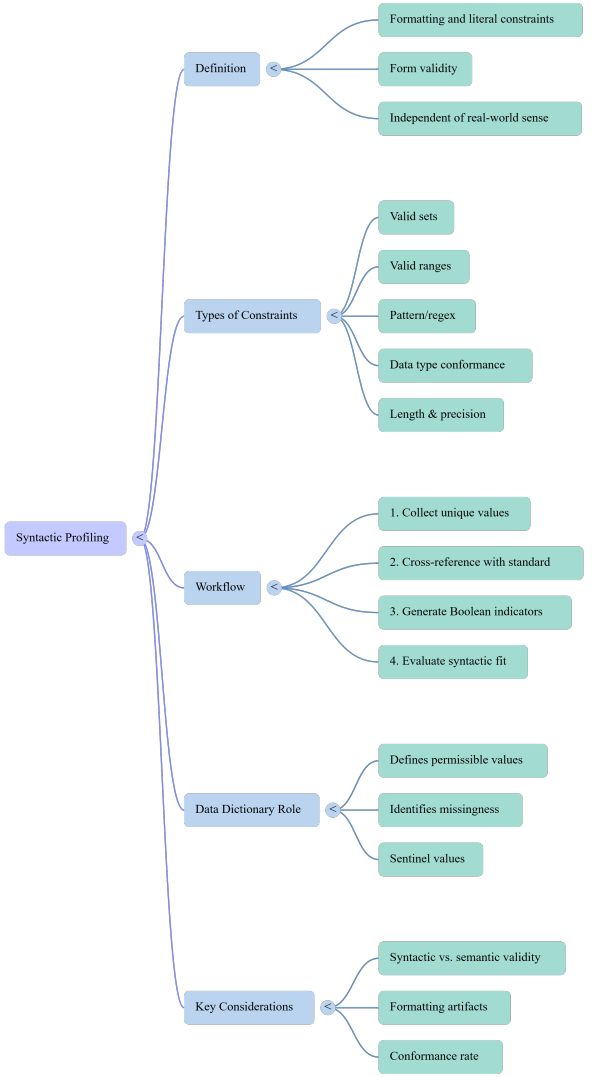
Permissible values are typically defined in a dataset's data dictionary. If the dictionary does not indicate that missing values are allowed in a column, any missing value is considered syntactically invalid. In practice, datasets often use undocumented sentinel values (e.g. -1, 9999, “N/A”, blank) to represent missingness inconsistently.

Workflow for Performing Syntactic Profiling

1. **Collect unique values:** extract all distinct values present in the column being profiled.
2. **Cross-reference with a standard:** look up each unique value against a reference dataset of known valid values (e.g. comparing state abbreviations to actual US states/territories).
3. **Generate Boolean indicators:** flag each record with 1 (valid) or 0 (invalid) in a new column, making it easy to pinpoint errors such as a typo like “ZZ” in a state field.
4. **Evaluate syntactic fit:** examine random samples of both valid and invalid values to judge whether the assumed syntactic type actually fits the data, or needs revision.

Additional Considerations

- **Syntactic vs. semantic validity:** a value can be syntactically valid but semantically wrong (e.g. a birth date formatted correctly but set in the future).
- **Formatting artifacts:** issues like dropped leading zeros, inconsistent casing, or stray whitespace can look like syntactic violations but are really cleaning issues, not bad data.
- **Conformance rate:** results are often summarized as the percentage of values per field that pass syntactic checks, to help prioritize cleanup.



Mind Map — Syntactic Profiling at a Glance