

CLAIMS

1. A method implemented by a data processing system for discovering a semantic meaning of data of a field included in one or more data sets, the method including:
 - identifying a field included in one or more data sets, with the field associated with an identifier; and
 - for that field:
 - profiling, by a data processing system, one or more data values of the field to generate a data profile for the field, with the data profile specifying one or more attributes of the one or more data values of the field;
 - accessing a plurality of tests, wherein a test specifies one or more given attributes and a label providing information about the one or more given attributes;
 - based on applying at least the plurality of tests to the data profile for the field, generating one or more label proposals for the field, wherein a label proposal includes a label that is proposed as providing a semantic meaning for the one or more data values of the field;
 - wherein a semantic meaning indicates what kind of data values are included in a given field;
 - determining a similarity among the one or more label proposals;
 - based at least on the similarity among the one or more label proposals, selecting a classification that specifies whether input is required in identifying the semantic meaning, from among the one or more label proposals, for the one or more data values of the field;
 - based on the classification, rendering a graphical user interface that requests the input in identifying a label proposal that identifies the semantic meaning or determining that the input is not required;
 - identifying one of the label proposals as identifying the semantic meaning of the one or more data values of the field; and
 - storing, in a data store, the identifier of the field with the identified one of the one or more label proposals that identifies the semantic meaning of the one or more data values of the field.
2. The method of claim 1, wherein profiling the one or more data values of the field includes determining a format of a data value of the field.

3. The method of claim 1, wherein profiling the one or more data values of the field includes determining a statistical value representing the data values included in the field.

4. The method of claim 3, wherein the statistical value comprises at least one of a minimum length of the one or more data values of the field, a maximum length of the one or more data values of the field, a most common data value of the field, a least common data value of the field, a maximum data value of the field, and a minimum data value of the field.

5. The method of claim 1, wherein applying the plurality of tests includes:
determining that the field includes a primary key for a data set of the one or more data sets; and
selecting a label proposal test of the plurality of tests that are related to the primary key.

6. The method of claim 1, wherein applying the plurality of tests includes:
performing a metadata comparison of data values of the field to terms in a glossary of terms.

7. The method of claim 1, wherein applying the plurality of tests includes:
determining, from the data profile, a pattern represented by the one or more data values of the field;
determining a particular label that is mapped to the pattern; and
labeling the field with the particular label.

8. The method of claim 1, wherein applying the plurality of tests includes:
retrieving a list of values that are representative of a data collection;
comparing the one or more data values of the field to the list of values;
determining, in response to the comparing, that a threshold number of the data values match the values of the list of values; and
in response to the determining, labeling the field with a particular label that specifies the data collection.

9. The method of claim 1, wherein applying the plurality of tests includes:

generating at least two label proposals for the field; and
determining whether the at least two label proposals are exclusive or inclusive of one another.

10. The method of claim 1, further including:
determining, in response to applying the plurality of tests, a relationship between the field and another field of the one or more data sets.

11. The method of claim 10, wherein the relationship includes one of an indication that a first data value the field determines a second data value stored in the other field, an indication that the first data value correlates to the second data value, or an indication that the first data value is identical to the second data value.

12. The method of claim 1, wherein the plurality of tests are each associated with at least one weight value, the method further including:
updating a weight value associated with at least one test; and
reapplying the test to the data profile using the updated weight value.

13. The method of claim 1, further including training the plurality of tests using a machine learning process.

14. The method of claim 1, further comprising:
retrieving, from a data quality rules environment, one or more data quality rules that are assigned to the label proposal specifying the semantic meaning; and
assigning a data quality rule of the one or more data quality rules to the field.

15. The method of claim 1, wherein comparing the label proposals generated from the tests includes:
applying a score value to each label proposal;
for each label of the one or more label proposals, combining the score values associated with that label proposal; and
ranking the one or more label proposals according to the score value associated with each label proposal.

16. The method of claim 1, further including:
receiving validation of the one or more label proposals from the plurality of tests; and
responsive to receiving the validation, weighting the plurality of tests with the one or
more label proposals.

17. The method of claim 1, wherein the data store includes a data dictionary.

18. The method of claim 1, further including:
outputting the label proposals to a data quality rules environment.

19. The method of claim 18, further including:
reducing, based on the identified one of the label proposals, a number of errors for
processing data for the field using data quality rules from the data quality environment
relative to another number of errors for processing the data for the field without using the
identified one of the label proposals.

20. A data processing system for discovering a semantic meaning of a field
included in one or more data sets, the system including:
a data storage storing instructions; and
at least one processor configured to execute the instructions stored by the data storage
to perform the steps of the method of any one of claims 1 to 19.

21. One or more non-transitory computer readable media storing instructions for
discovering a semantic meaning of a field included in one or more data sets, the instructions
being executable by one or more processors configured to perform the steps of the method of
any one of claims 1 to 19.