

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:
- 5 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;
- 10 accessing a plurality of label proposal tests;
- based on applying at least the plurality of label proposal tests to the data profile, generating a set of label proposals;
- determining a similarity among the label proposals in the set of label proposals;
- 15 based at least on the similarity among the label proposals in the set, selecting a classification;
- based on the classification, rendering a graphical user interface that requests input in identifying a label proposal that identifies the semantic meaning or determining that no input is required;
- 20 identifying one of the label proposals as identifying the semantic meaning; and
- storing, in a data store, the identifier of the field with the identified one of the label proposals that identifies the semantic meaning;
- receiving a request to process the one or more data sets including a particular identifier;
- 25 accessing, in response to the request, from the data store, the particular identifier and a particular label associated with the particular identifier;
- determining, based on the particular label, one or more data processing rules for processing at least a portion of the one or more data sets; and
- processing at least the portion of the one or more data sets in accordance with the
- 30 selected one or more data processing rules.

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 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
5 minimum length of the data values of the field, a maximum length of the 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 label proposal tests
10 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 label proposal tests that are that is related to the primary key.

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

20 7. The method of claim 1, wherein applying the plurality of label proposal tests includes:
determining, from the data profile, a pattern represented by the data values stored of the field;
25 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 label proposal tests includes:
30 retrieving a list of values that are representative of a data collection;
comparing the 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; 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 label proposal tests
5 includes:
generating at least two labels for the field; and
determining whether the at least two labels are exclusive or inclusive of one another.

10. The method of claim 1, further including:
10 determining, in response to applying the plurality of label proposal 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
15 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 label proposal tests are each
associated with at least one weight value, the method further including:
20 updating a weight value associated with at least one label proposal test; and
reapplying the label proposal test to the data profile using the updated weight value.

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

25 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.

30 15. The method of claim 1, wherein comparing the label proposals generated from the
label proposal tests includes:
applying a score value to each label proposal;

for each label of the label proposals, combining the score values associated with that label; and

ranking the labels according to the score value associated with each label.

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

10

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.

15

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
20 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
25 at least one processor configured to execute the instructions stored by the data storage
to perform operations including:

identifying a field included in one or more data sets, with the field having an
identifier; and

for that field:

- 30 profiling, by a data processing system, one or more data values of the
field to generate a data profile;
accessing a plurality of label proposal tests;

based on applying at least the plurality of label proposal tests to the data profile, generating a set of label proposals;

determining a similarity among the label proposals in the set of label proposals;

5 based at least on the similarity among the label proposals in the set, selecting a classification;

based on the classification, rendering a graphical user interface that requests input in identifying a label proposal that identifies the semantic meaning or determining that no input is required;

10 identifying one of the label proposals as identifying the semantic meaning; and

storing, in a data store, the identifier of the field with the identified one of the label proposals that identifies the semantic meaning;

15 receiving a request to process the one or more data sets including a particular identifier;

accessing, in response to the request, from the data store, the particular identifier and a particular label associated with the particular identifier;

determining, based on the particular label, one or more data processing rules for processing at least a portion of the one or more data sets; and

20 processing at least the portion of the one or more data sets in accordance with the selected one or more data processing rules.

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 operations including:

25 identifying a field included in one or more data sets, with the field having an identifier; and

for that field:

30 profiling, by a data processing system, one or more data values of the field to generate a data profile;

accessing a plurality of label proposal tests;

based on applying at least the plurality of label proposal tests to the data profile, generating a set of label proposals;

- determining a similarity among the label proposals in the set of label proposals;
- based at least on the similarity among the label proposals in the set, selecting a classification;
- 5 based on the classification, rendering a graphical user interface that requests input in identifying a label proposal that identifies the semantic meaning or determining that no input is required;
- identifying one of the label proposals as identifying the semantic meaning; and
- storing, in a data store, the identifier of the field with the identified one of the
- 10 label proposals that identifies the semantic meaning;
- receiving a request to process the one or more data sets including a particular identifier;
- accessing, in response to the request, from the data store, the particular identifier and a particular label associated with the particular identifier;
- 15 determining, based on the particular label, one or more data processing rules for processing at least a portion of the one or more data sets; and
- processing at least the portion of the one or more data sets in accordance with the selected one or more data processing rules.