

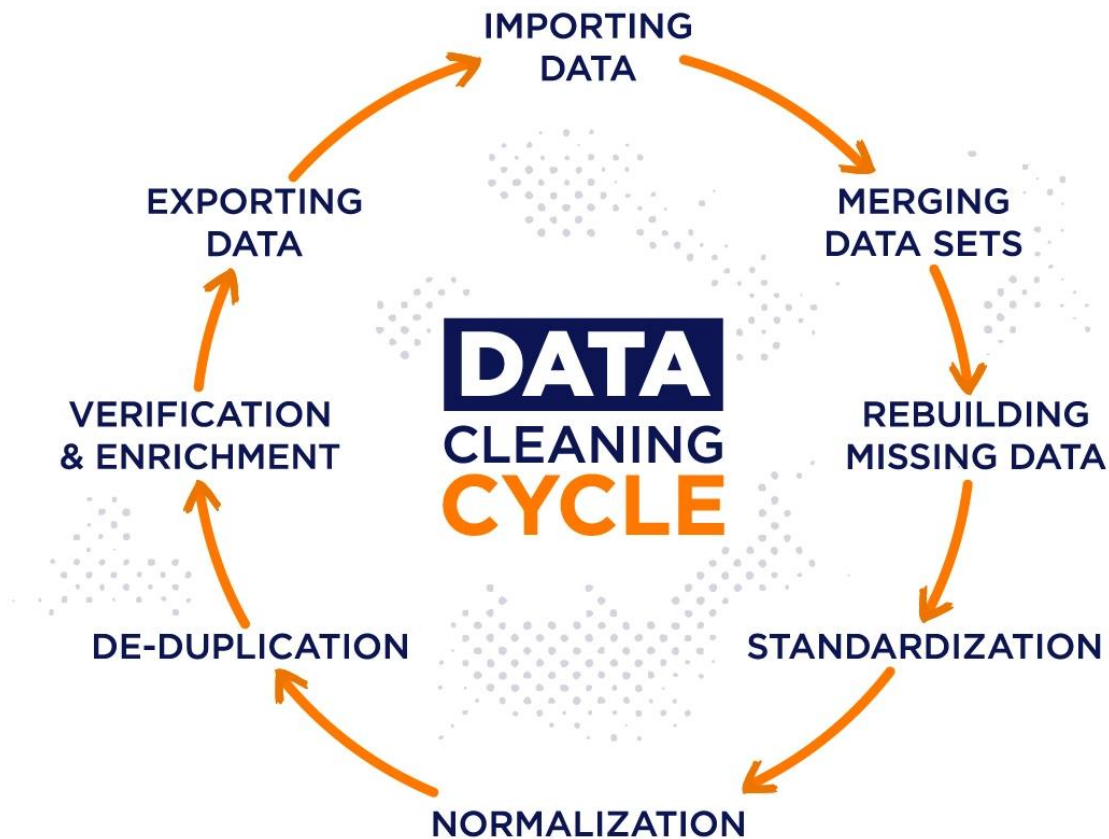


Organizing Data

Stephen Downes

December 4, 2021

Data Cleaning



“Data cleaning is the process of identifying, deleting, and/or replacing inconsistent or incorrect information from the database.”

<https://www.iteratorshq.com/blog/data-cleaning-in-5-easy-steps/>

Data Quality Revisited

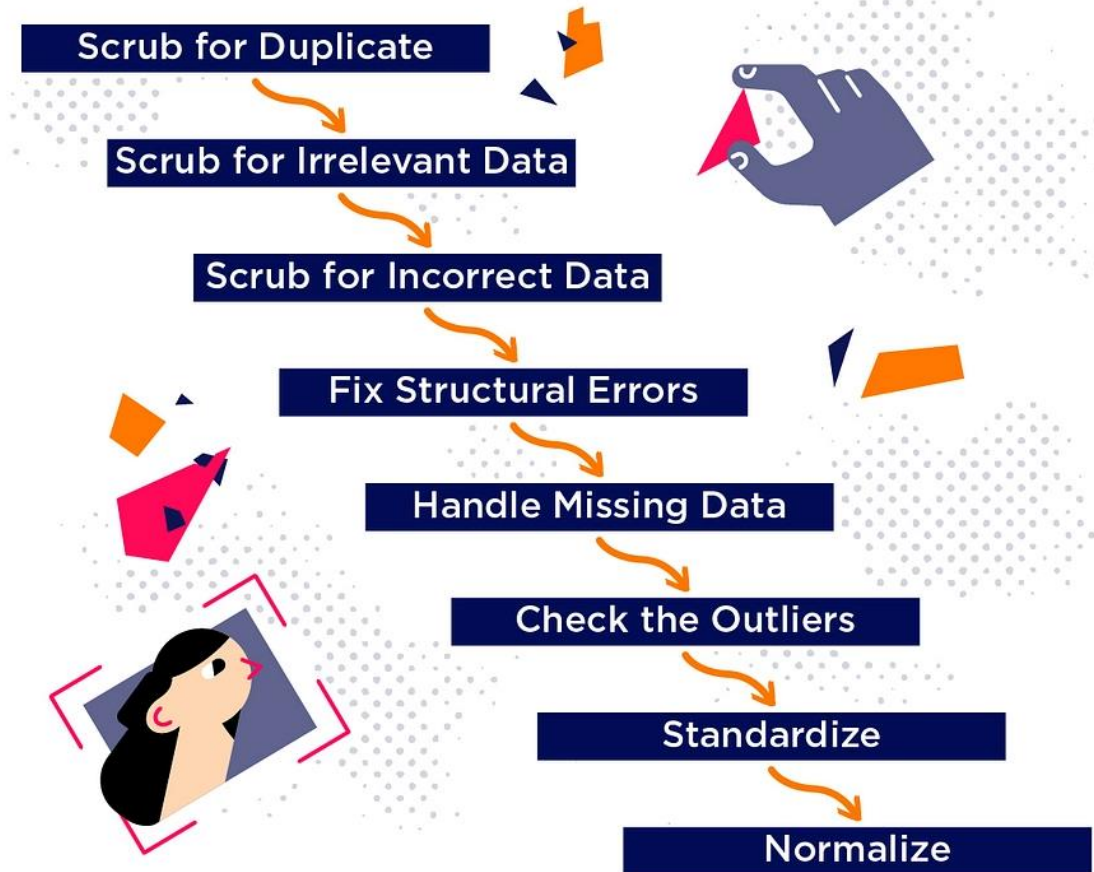


Now we see data quality not as an attribute of source data, but as an output of data cleaning

<https://www.iteratorshq.com/blog/data-cleaning-in-5-easy-steps/>

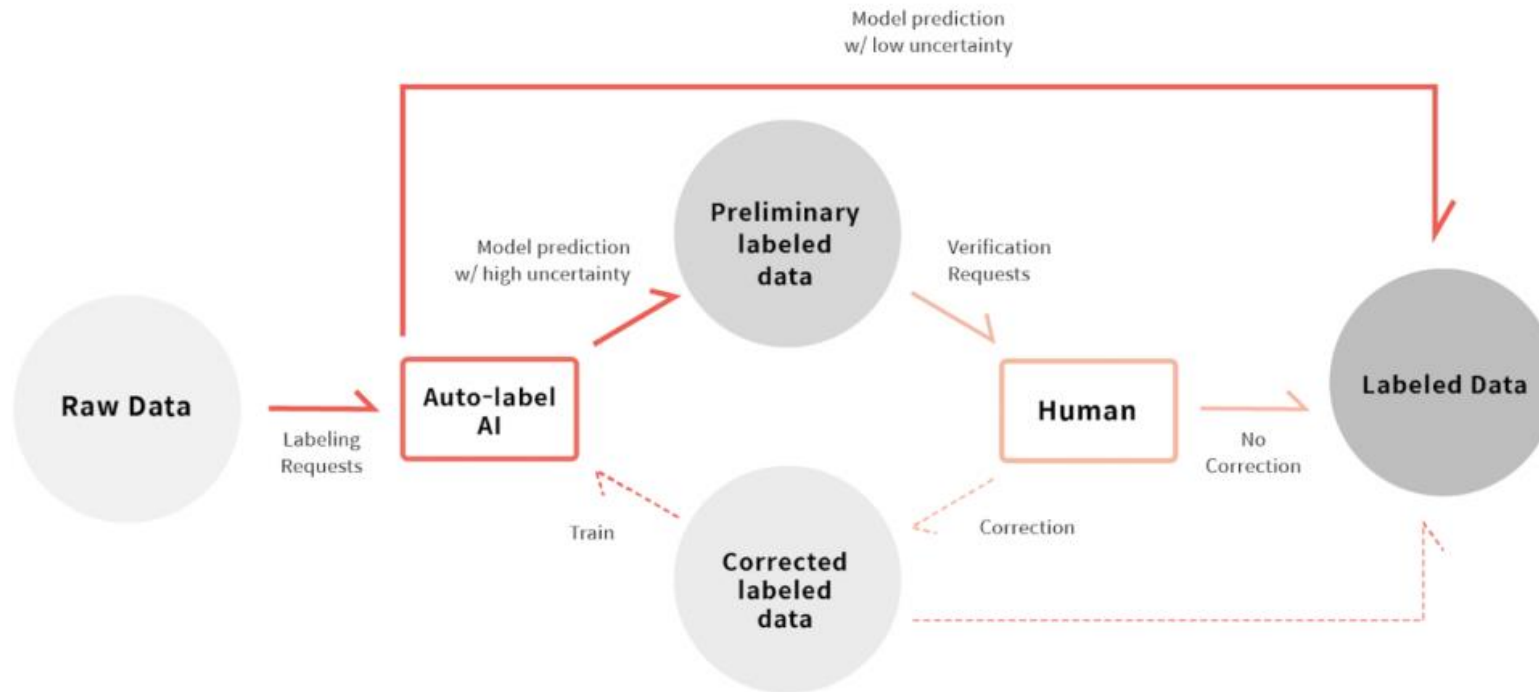
Data Cleaning Workflow

“An example of a typical data cleaning workflow, featuring a series of operations performed on the data repeatedly until reaching sufficient data quality.”



<https://www.iteratorshq.com/blog/data-cleaning-in-5-easy-steps/>

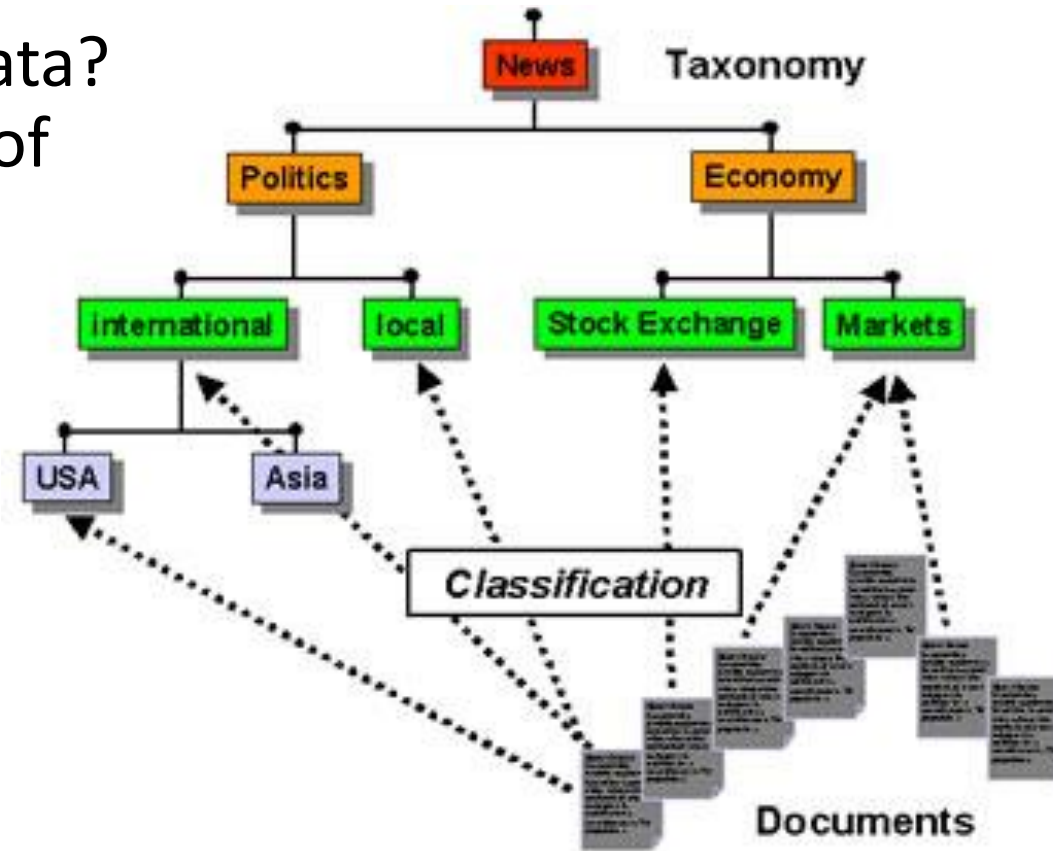
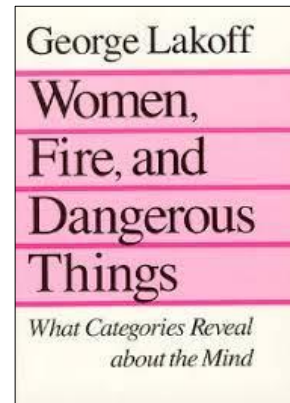
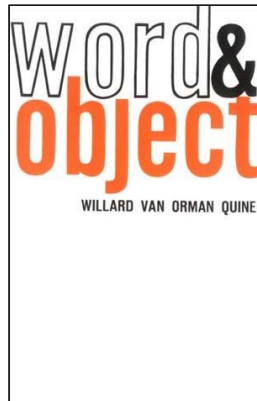
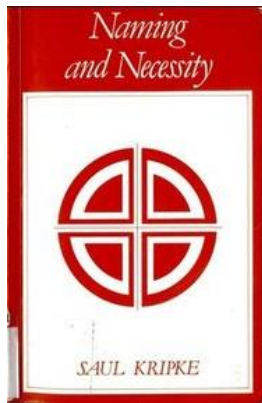
Classification and Naming



Classification and naming are essentially operations performed as a part of labeling in AI, that is, using human-readable signs that interpret a specific piece of data.

Taxonomies, Ontologies and Natural Kinds


Is there a 'right' way to label data?
Do we all agree on what kinds of things there are in the world?




Classification Algorithms

- Binary classification
- Multiclass classification
- Multilabel classification


Multi-Label Classification



Java
C++
Python
JavaScript
PHP




Action
Crime
Thriller
Comedy
Drama




Cat
Dog
Tiger
Bird
Fish

Multi-Class Classification



Cat
Dog
Fox
Tiger
Lion

0	5
1	6
2	7
3	8
4	9



Person A
Person B
Person C

Popular Classification Algorithms:

- [Logistic Regression](#)
- [Naive Bayes](#)
- [K-Nearest Neighbors](#)
- [Decision Tree](#)
- [Support Vector Machines](#)

<https://medium.com/swlh/3-types-of-classification-problems-in-machine-learning-1cffd3765ca1>

<https://monkeylearn.com/blog/classification-algorithms/>