Data wrangling is time consuming

Automatic wrangling has several challenges

- How to pick relevant wrangling transformations from the enormous space of valid transformations?
- What metric to use to rank them?
- How to translate them to human-readable scripts?
- How to make the scripts indistinguishable from human-authored scripts?
- How to make sure the scripts are performant?

CoWRangler: Recommender System for Data-Wrangling Scripts

Data scientists can customize and interact with CoWrangler suggestions by
- Editing the scripts generated by CoWrangler
- Expressing intent by example [1]

Wrangling suggestions with explanations
Select a suggestions to view translated script and a preview of the transformed data

CoWRangler in a nutshell

- Sample Data
  - Program Learners
  - Suggestions
  - Vocabulary
  - Split
  - Column
  - Fill Missing Values
  - Translation
  - Ranked Suggestion

CoWRangler achieves 53% accuracy

Benchmark: 2248 pandas operations from 730 Kaggle notebooks
- CoWrangler’s vocabulary supports 33% operations
- Suggestions are accurate in 53.4% cases

References

How do I debug cryptic error messages like ValueError?

How will I navigate this API jungle? There are so many APIs and parameters!

How do I know if my script is working properly?

Is my data clean now?

How do I debug cryptic error messages like ValueError?

How do I make the scripts indistinguishable from human-authored scripts?