# CleanML

Data cleaning is a major part of data science (up to ~80% of effort).

Despite this, we don't know systematically:

How data cleaning helps ML

Prior work focuses either on:

robust ML models **or**

data cleaning itself

CleanML answers this by empirically studying the impact of data cleaning on ML.

# *The core question*
# *how does data cleaning affect downstream ML performance?*

**To answer this, the authors designed their experiment to deal with:**

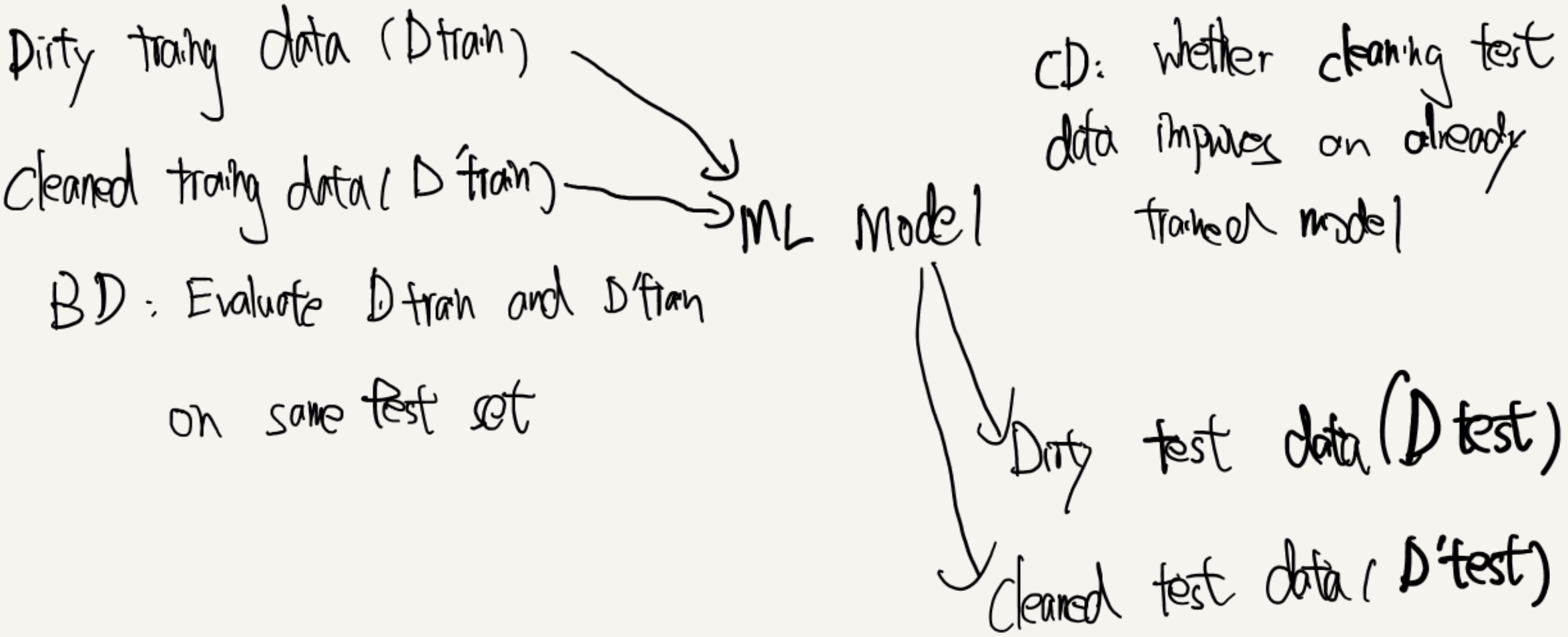
- Different datasets (real life)
- Different error types(5 errors prevalent in life)
- Different cleaning methods(both commonly used and academic)
- Different ML models
- Different stages(development and deployment)

**Each experiment compares accuracy before vs after cleaning**

**Results are summarized using:**

- **P** (positive impact)
- **S** (no significant impact)
- **N** (negative impact) as a key attribute (flag) in a sql relationship model

**for each entry in the table, they produce 20 pairs of metrics by using different train/test split**


Dirty traing data (Dtrain)
Cleaned traing data (D'train)
ML Model
BD: Evaluate Dtrain and D'train
on same test set
CD: whether cleaning test
data impacts on already
trained model
Dirty test data (Dtest)
Cleaned test data (D'test)

# *Results*

- Cleaning does not always help ML.
- cleaning impact depends on datasets – while two datasets may contain errors of the same type, the distributions of those errors can be vastly different
- Typical trends:

  *Missing values → often helpful*

  *Outliers → mostly insignificant*

  *Mislabels → often helpful*

  *Inconsistencies → safe but small gains*

  *Duplicates → often harmful*

- if cleaning a dataset has a particular impact for one ML model, cleaning is likely to have the same type of impact for other models as well.
- performing data cleaning is a more broadly applicable solution, compared with developing specific robust ML models.

# Quiz question:

**What is the main difference between the BD and CD scenarios in the CleanML study?**

BD asks whether cleaning the training data can potentially improve the model performance on unseen test data.

CD asks whether cleaning the test data can improve the performance of an already trained model