CLEANML

A STUDY FOR EVALUATING THE IMPACT OF DATA CLEANING ON ML CLASSIFICATION TASKS

Peng Li et al.

CORE QUESTION

- Data cleaning is **expensive** and **ubiquitous**
- But does cleaning actually improve ML accuracy — and when?

> **Contribution:** First systematic, statistically rigorous study of cleaning → ML impact

github.com/chu-data-lab/CleanML

WHY THIS STUDY IS NEEDED

TWO DISCONNECTED PERSPECTIVES

ML Community:

- Build noise-robust models
- Often skip cleaning

Database Community:

- Clean data without ML feedback
- Focus on data quality alone

THE PROBLEM

No large-scale evidence on:

- Which **errors** matter?
- Which **cleaning** helps?
- Which **models** benefit?

Goal: Bridge data cleaning and downstream ML performance

CLEANML AT A GLANCE

14
Real-world datasets

5
Error types

7
ML models

- Multiple cleaning methods per error type
- Training & deployment scenarios
- Statistical hypothesis testing
- False discovery control (BY procedure)

KEY DESIGN CHOICE

Use **real errors**, not synthetic noise

Why? Synthetic errors may under/over-estimate cleaning impact

ERROR TYPES & CLEANING METHODS

5 ERROR TYPES

1. Missing Values
No value stored for cells

2. Outliers
Observations distant from others

3. Duplicates
Multiple records, same entity

4. Inconsistencies
Different values, same meaning (CA vs California)

5. Mislabels
Incorrectly labeled examples

CLEANING SPECTRUM

Simple Methods:

- Mean/Median/Mode imputation
- Record deletion
- Standard deviation detection
- IQR detection

Advanced Methods:

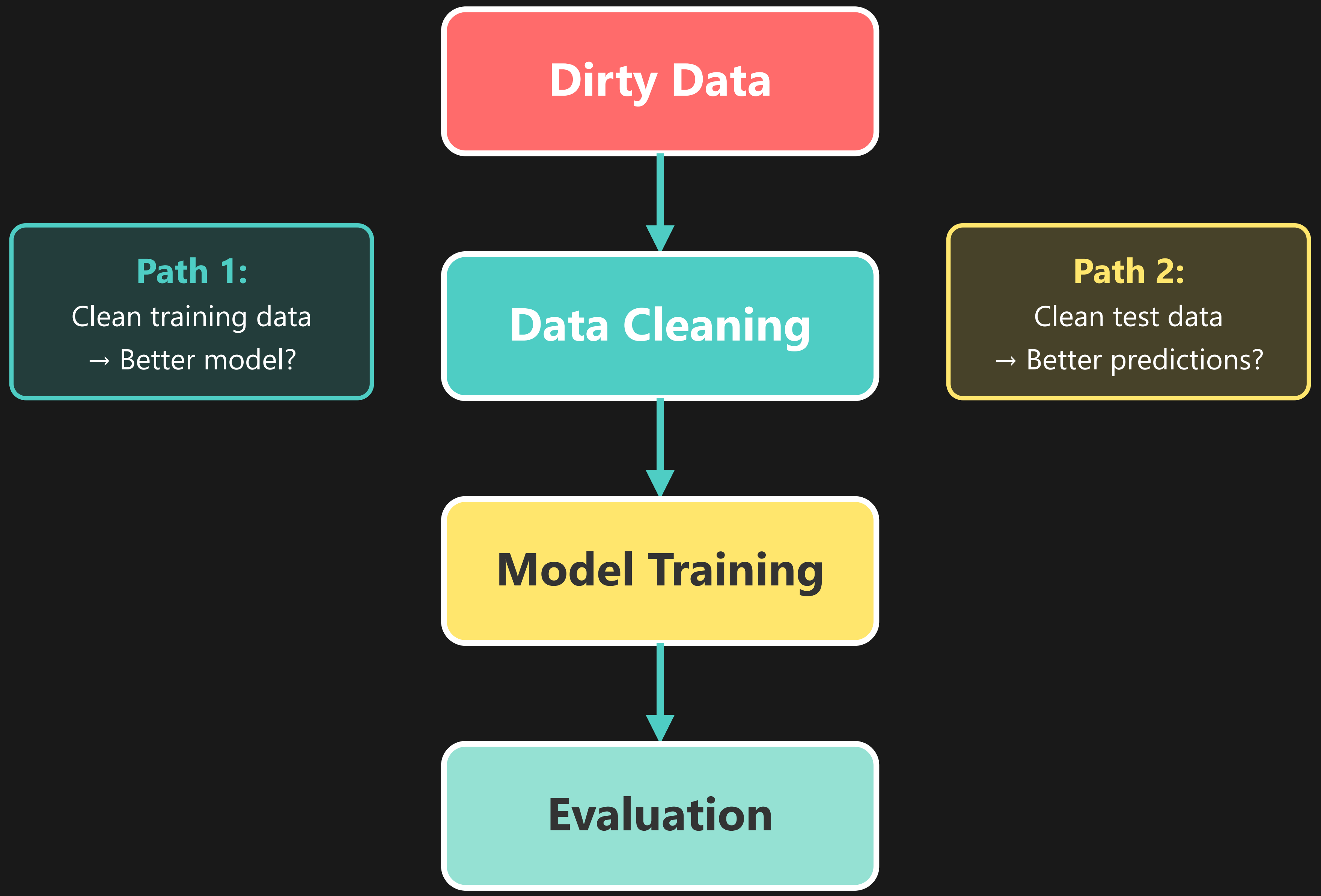
- **HoloClean** - Probabilistic inference
- **ZeroER** - Unsupervised entity resolution
- **cleanlab** - Confident learning

↓ Press down for key insight

COMPARING CLEANING METHODS

Key Idea: Compare simple vs. state-of-the-art under **identical ML settings**

ML WORKFLOW WITH CLEANING



Insight: Cleaning **location** matters as much as cleaning **method**

ML MODELS & EXPERIMENTAL CONTROL

7 MODELS TESTED

- Logistic Regression
- Decision Tree
- Random Forest
- AdaBoost
- XGBoost
- K-Nearest Neighbors
- Naive Bayes

STATISTICAL RIGOR

20

Random splits per experiment

- Paired sample t-test
- BY procedure for false discovery contro
- Significance level α = 0.05

↓ Press down for outcome labels

OUTCOME LABELS

P

Positive Impact (Helps)

S

Statistically Insignificant

N

Negative Impact (Hurts)

CLEANML DATABASE ARCHITECTURE

CleanML Experiment Database

Dataset	Error	Cleaning	Model	Scenario	Impact
EEG	Outliers	IQR+Mean	LogReg	BD	P
Credit	Missing	Median	XGBoost	CD	S
Movie	Duplicates	ZeroER	RandomF	BD	N

```
SELECT error_type, impact, COUNT(*)
FROM CleanML
GROUP BY error_type, impact
```

↓ Press down to see why this matters

WHY THIS MATTERS

SQL-STYLE ANALYSIS

Query thousands of experiments across multiple dimensions

FAIR COMPARISON

Controlled variables ensure apples-to-apples evaluation

REPRODUCIBLE FRAMEWORK

Extensible design for future cleaning methods and datasets

KEY RESULTS: WHAT HELPS VS. WHAT DOESN'T

OFTEN HELPFUL

- **Missing value imputation**

  49% positive impact

- **Mislabel cleaning**

  47% positive, especially for boosting

OFTEN HARMFUL

- **Duplicate removal**

  22% negative impact

- **Why?**

  False positives delete useful training data

MOSTLY INSIGNIFICANT ~

- **Outlier cleaning**

  61% insignificant, 31% positive

- **Inconsistency fixing**

  88% insignificant, 12% positive

CORE PATTERN

Cleaning impact is **error-specific**, not universal

IMPACT BY ERROR TYPE

Error Type		Insignificant	Negative
Missing Values	49%	27%	24%
Outliers	31%	61%	8%
Mislabels	47%	38%	15%
Inconsistencies	12%	88%	0%
Duplicates	11%	67%	22%

DATASET DEPENDENCE

CRITICAL INSIGHT

Strongest Finding:

Same error type behaves **very differently** across datasets

Missing Values Impact by Dataset

86%	65%	49%	29%	6%
USCensus	Titanic	Marketing	Credit	Airbnb

% Positive Impact

6% to 86% variation!

↓ Press down for implications

IMPLICATION

NO UNIVERSAL SOLUTION

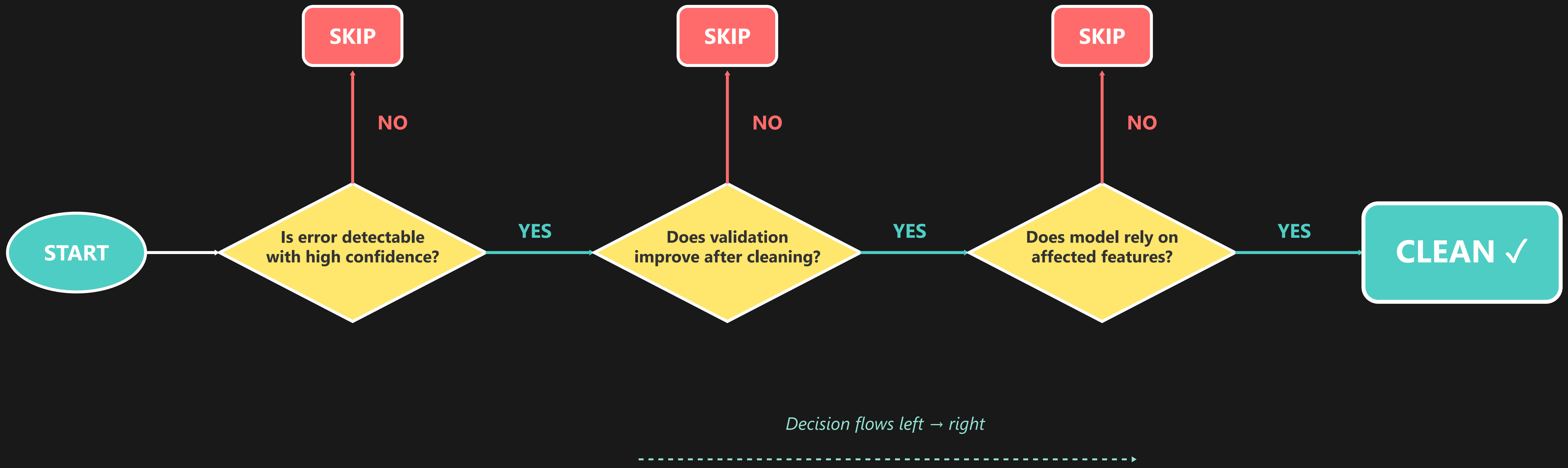
There is no "one-size-fits-all" cleaning strategy that works across all datasets

VALIDATION IS ESSENTIAL

Validation-based decisions are **critical** for determining whether to clean

> *"Dataset characteristics matter more than cleaning algorithms"*

WHEN SHOULD YOU CLEAN?



Message: Cleaning should be a **validated decision**, not a reflex

CLEANML VS. ROBUST ML APPROACHES

DATA CLEANING APPROACH

- Clean data first
- Use standard ML models
- Works for any error type
- Model-agnostic solution

ROBUST ML APPROACH

- Keep dirty data
- Use specialized models
- Error-type specific
- Model-dependent solution

↓ Press down for head-to-head comparison

HEAD-TO-HEAD COMPARISON

Error Type	Robust ML Method		Tie	Robust Wins
Missing Values	NaCL (Robust LR)	83%	0%	17%
Mislabels	Deep Learning (MLP)	85%	15%	0%
Inconsistencies	Deep Learning (MLP)	50%	50%	0%
Duplicates	Deep Learning (MLP)	0%	75%	25%

↓ Press down for key finding

KEY FINDING

Finding: Cleaning often **outperforms** robust ML

Reason: Cleaned data works with ANY downstream model without modification

FINAL TAKEAWAYS

1. CLEANING DOES NOT ALWAYS HELP

Impact varies by error type, dataset, and cleaning method

2. SIMPLE METHODS ARE OFTEN SUFFICIENT

HoloClean ≈ Mean Imputation in many cases

3. DUPLICATE CLEANING IS RISKY

22% negative impact due to false positives deleting valid data

4. VALIDATE CLEANING LIKE A HYPERPARAMETER

Use validation set to select cleaning method + ML model

↓ Press down for quiz

QUICK QUIZ

Q: Why Does the Same Error Break Some Models But Not Others?

A: Dataset dependence is the strongest finding we discovered. The same type of data error can have wildly different effects depending on your dataset, as we saw variations ranging from just 6% to as high as 86% impact! This shows that the context and characteristics of your specific dataset matter much more than the cleaning algorithm you choose. It's like how the same medicine affects different people differently.

↓ Press down for final slide

THANK YOU!

Questions?

Presented by: Bhargav Limbasia