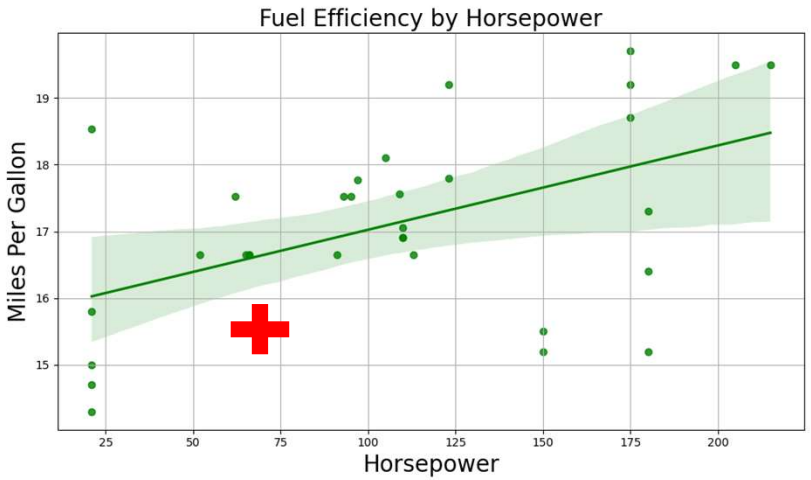
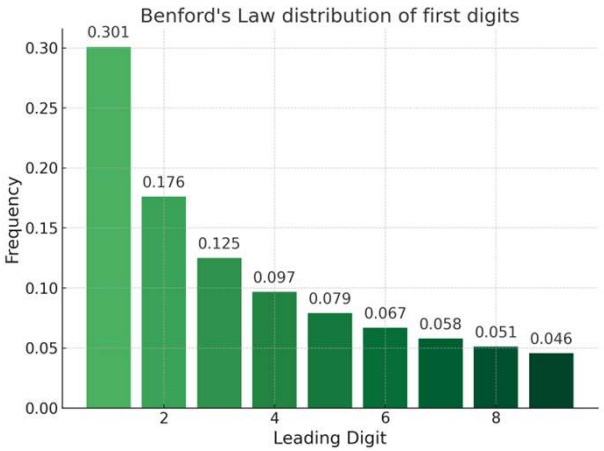
Correlation Reversal Manipulation Revealed By Benford's Law and Random Forest

Poj Netsiri, TU Wien Data Science

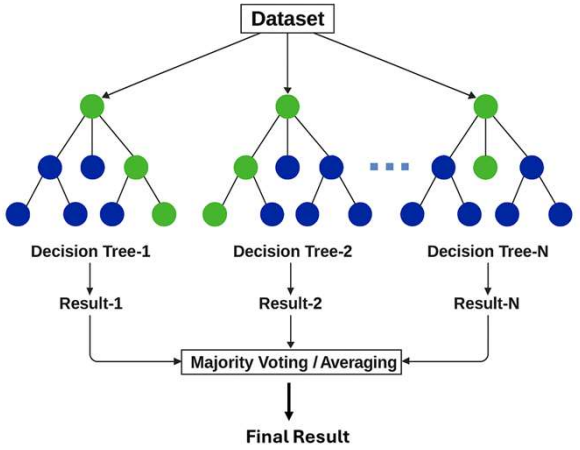
Q: Can we trust this data?



Benford's Law
(Statistical Analysis)



Random Forest
(Machine Learning)

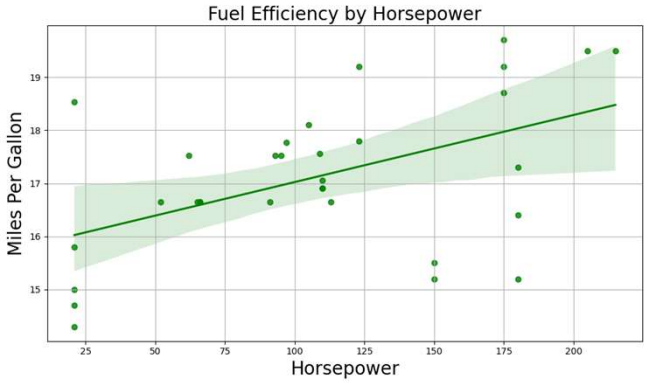


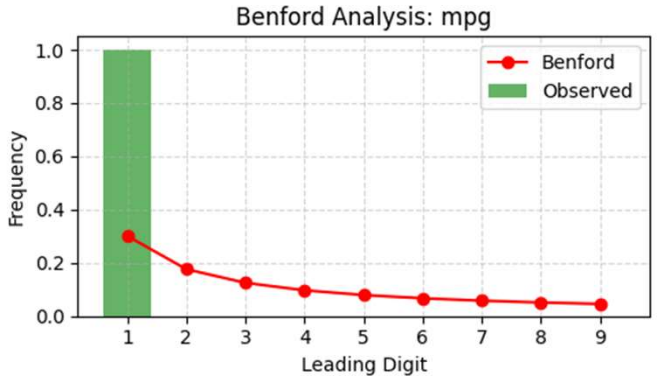
From the given data (n= 32):

r = +0.5, p <0.001

"Cars with higher horsepower achieve better fuel efficiency"

Using Benford's Law to Detect Manipulation

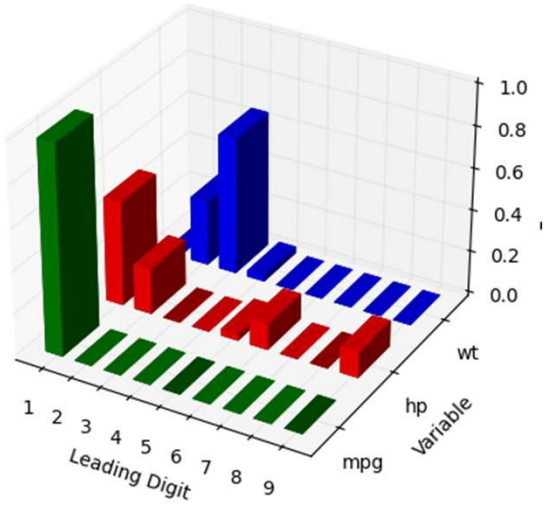




Chi-square = 74.3017, p-value = 6.81 x 10^{-13}

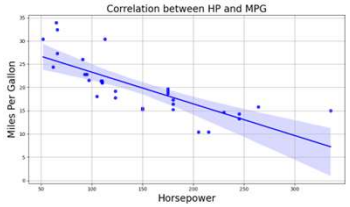
Criteria: Chi-square > 30.58 and p < 0.0001

Manipulated / Unmanipulated

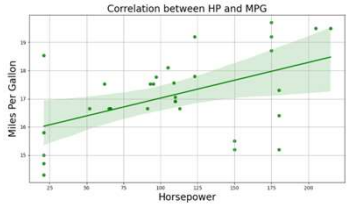


Using Random Forest to Detect Manipulation

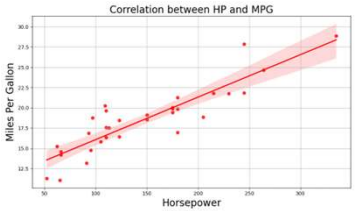
Original Data (Label=0)



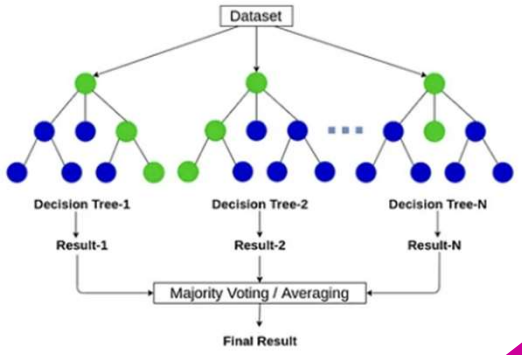
Manipulated Data (Label=1)



New Data



Supervised Training



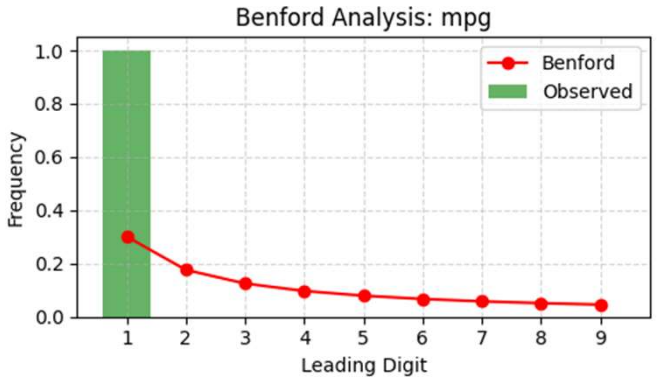
RF Model

Classifier

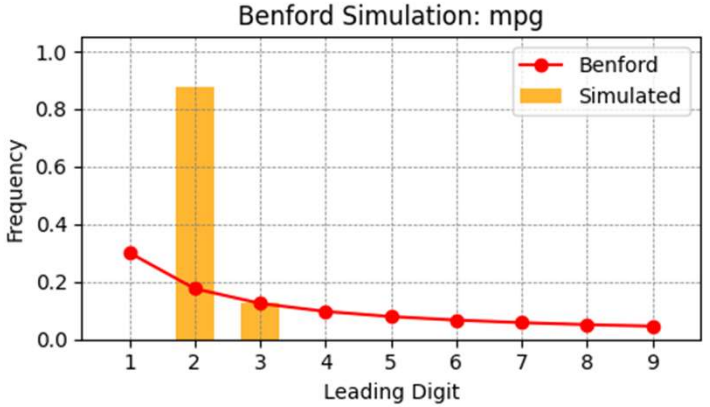
RF Model

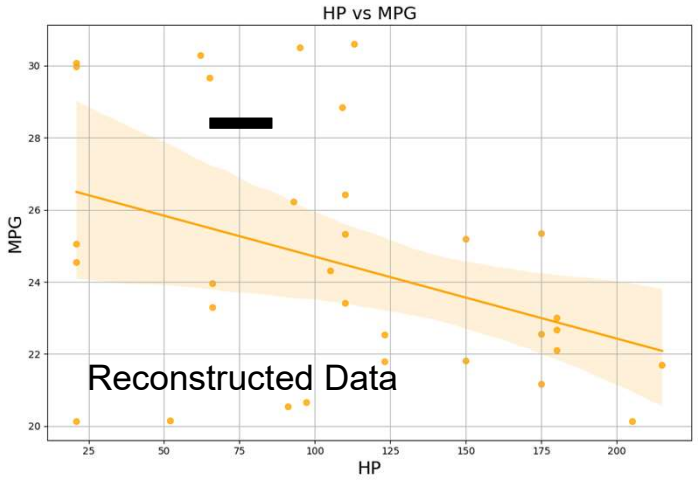
Manipulated/ Unmanipulated

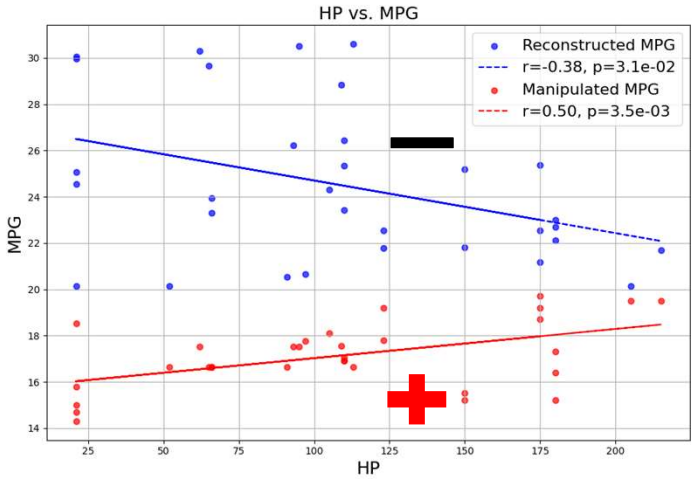
Using Benford's Law to Detect Correlation Reversal



Stochastic Simulation

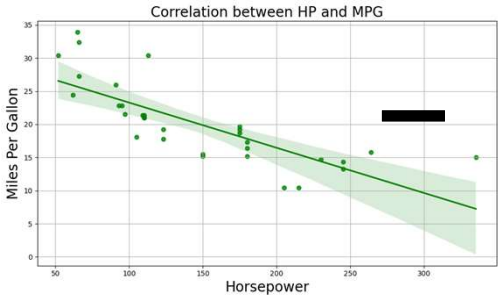




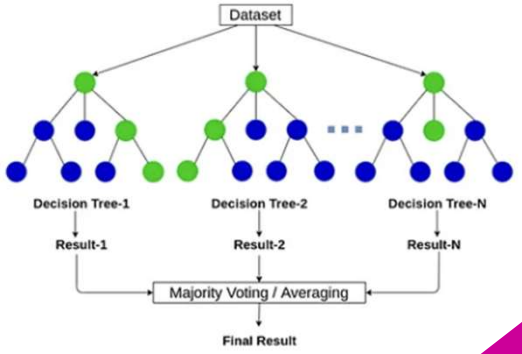


Using Random Forest to Detect Correlation Reversal

Original Data

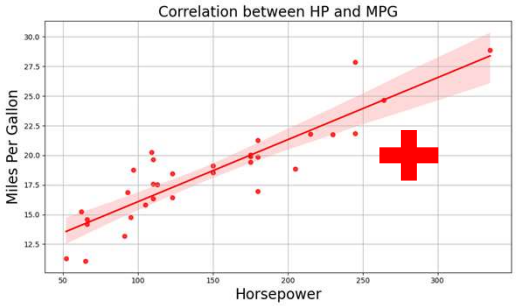


Training



RF Model

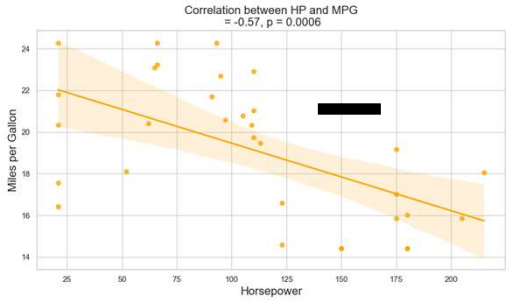
Manipulated Data

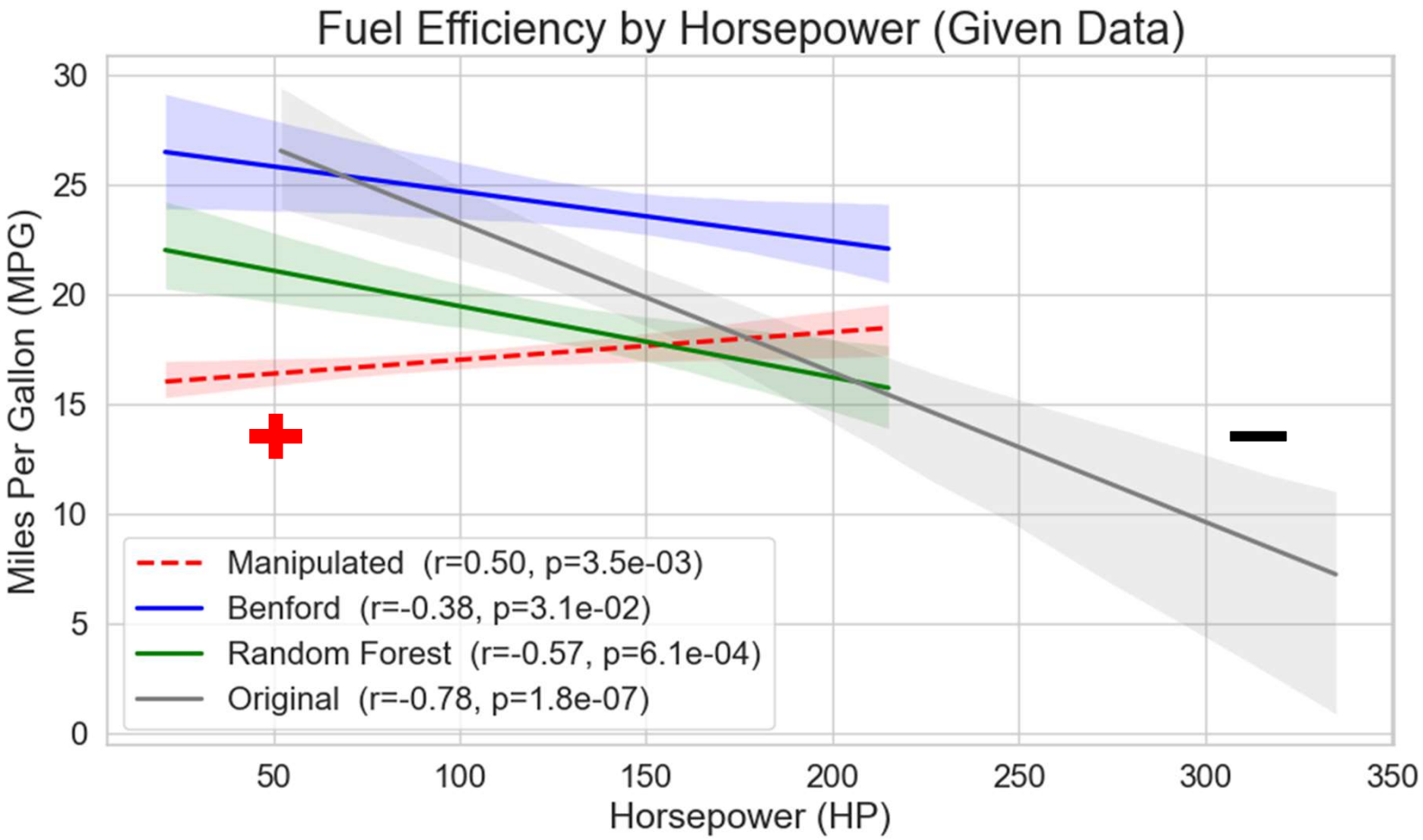


Regressor

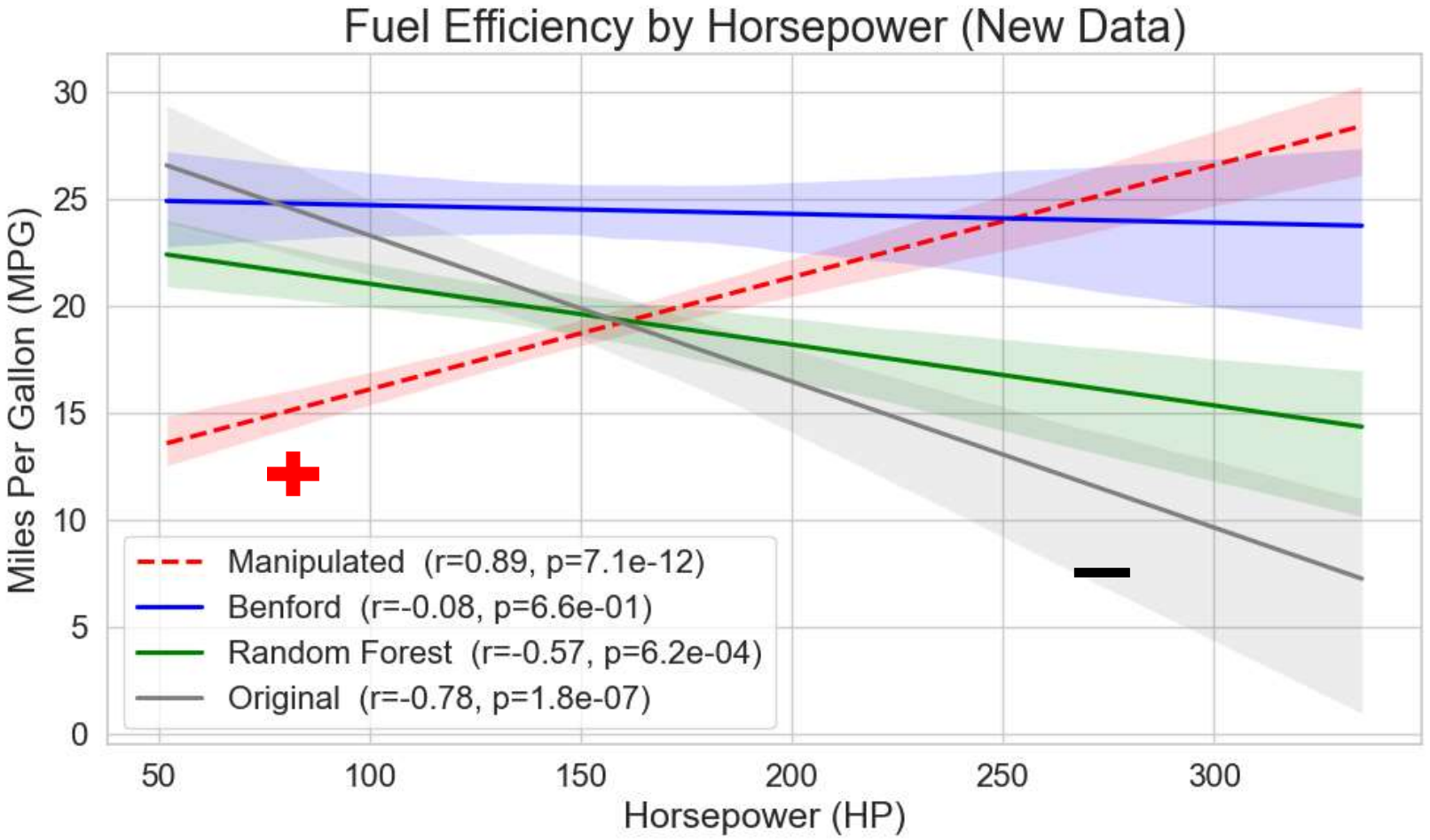
RF Model

Reconstructed Data





"Cars with higher horsepower actually tend to have lower fuel efficiency."



"Cars with higher horsepower actually tend to have lower fuel efficiency."

Performance Evaluation of Random Forest Classifier

Class	Precision	Recall	F1-score	Support	Description
0	0.83	1.00	0.91	5	Unmanipulated data: perfect recall, slightly lower precision.
1	1.00	0.88	0.93	8	Manipulated data: perfect precision, but slightly lower recall.
Accuracy			**0.92**	13	Overall correct classification rate.
Macro avg	0.92	0.94	0.92	13	Unweighted average across both classes.
Weighted avg	0.94	0.92	0.92	13	Weighted average based on class support.

Performance Evaluation of Random Forest Regressor

Metric	Value	Description
R^2 Score	0.6643	Proportion of variance in the target variable explained by the model; values closer to 1 indicate a better fit.
Mean Squared Error (MSE)	4.1079	Average squared difference between predicted and actual values; lower values reflect better predictive accuracy.
Mean Absolute Error (MAE)	1.7576	Average absolute difference between predicted and actual values; less sensitive to outliers than MSE.

Conclusion

- **Benford's Law** effectively detects manipulated datasets through anomalies in digit distribution.
- **Random Forest regressor** can reconstruct the underlying data trend, even in small datasets (32 datapoints).
- **Manipulated data** showed reversed correlation direction, while reconstructed data realigned direction with the original.
- A **hybrid approach**—using Benford's Law for detection and Random Forest for reconstruction—provides a powerful method to detect manipulation and identify direction reversal.