Sandia National Laboratories

Exceptional service in the national interest

A QUEST FOR MEANINGFUL PERFORMANCE METRICS IN MULTI-LABEL CLASSIFICATION

Presented by: Marie Tuft

Sandia National Laboratories

Collaborators: Jace Ritchie & Kyra Wisniewski

U.S. DEPARTMENT of ENERGY | NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

LDRD

SAND2025-04837C

SINGLE LABEL CLASSIFICATION

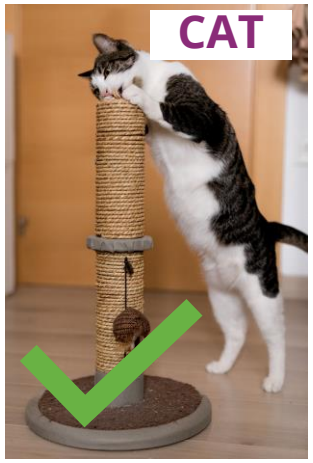
Task: Classify these pictures into dog, squirrel, cat, or owl.

SINGLE LABEL CLASSIFICATION

Task: Classify these pictures into dog, squirrel, cat, or owl.

SINGLE LABEL CLASSIFICATION

Task: Classify these pictures into dog, squirrel, cat, or owl.



MULTI-LABEL CLASSIFICATION

Task: Identify all the animals in the pictures.

MULTI-LABEL CLASSIFICATION

Task: Identify all the animals in the pictures.

MULTI-LABEL CLASSIFICATION

Task: Identify all the animals in the pictures.

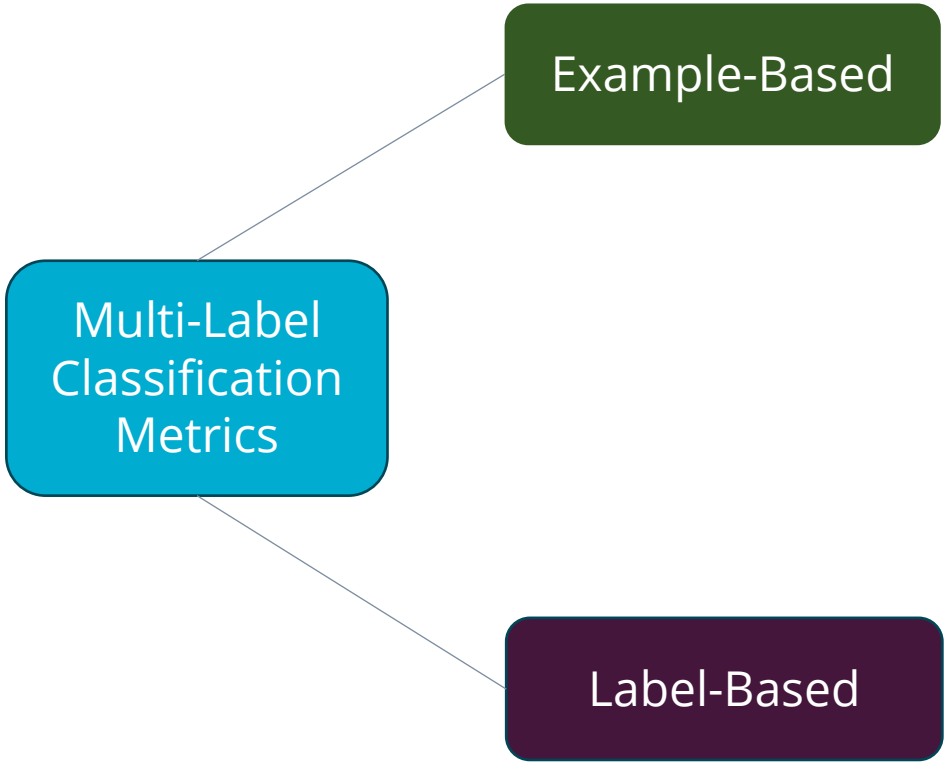
MULTI-LABEL CLASSIFICATION

Task: Identify all the animals in the pictures.

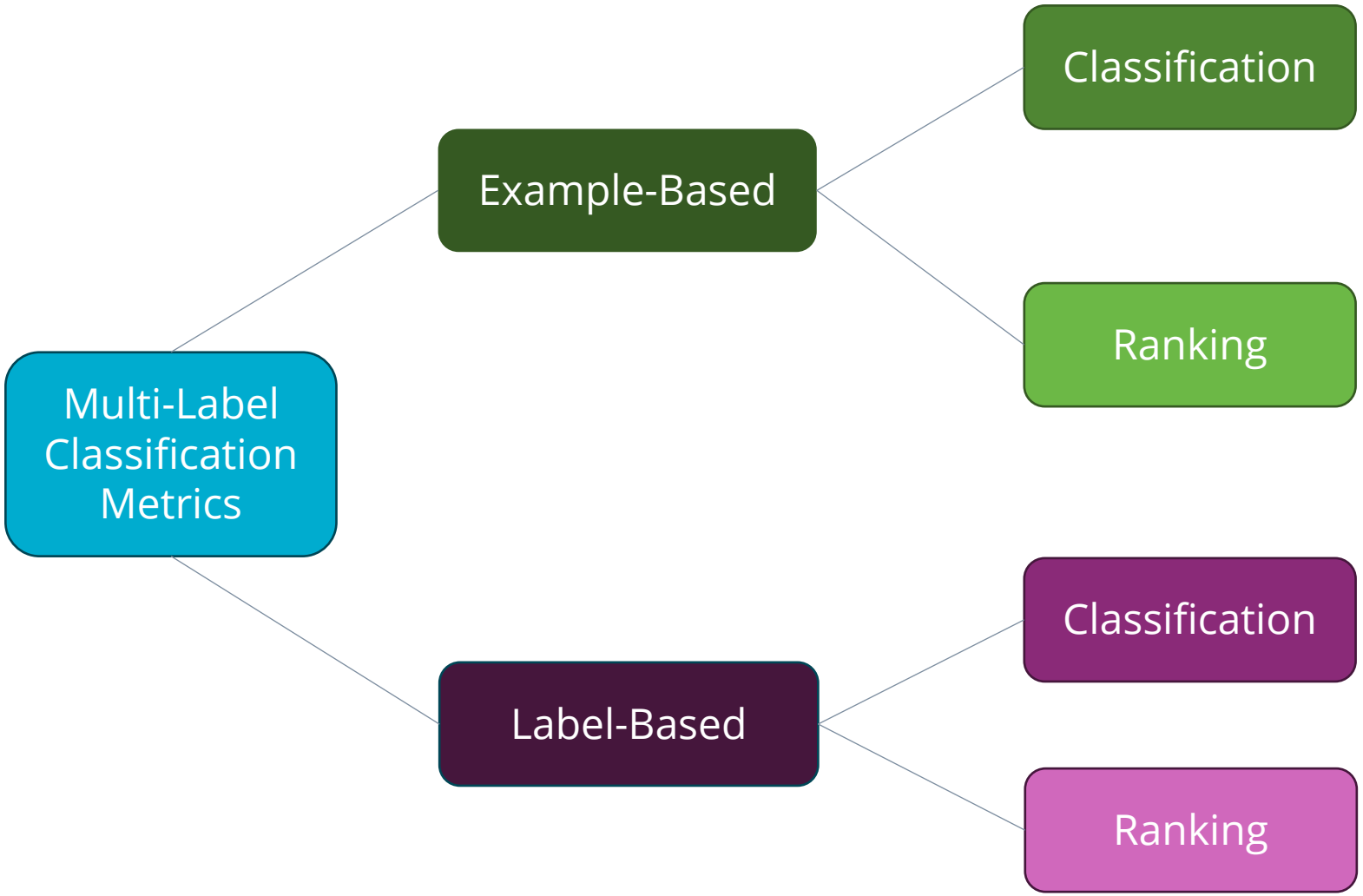
19 METRICS TO EVALUATE MULTI-LABEL CLASSIFICATION

Multi-Label Classification Metrics

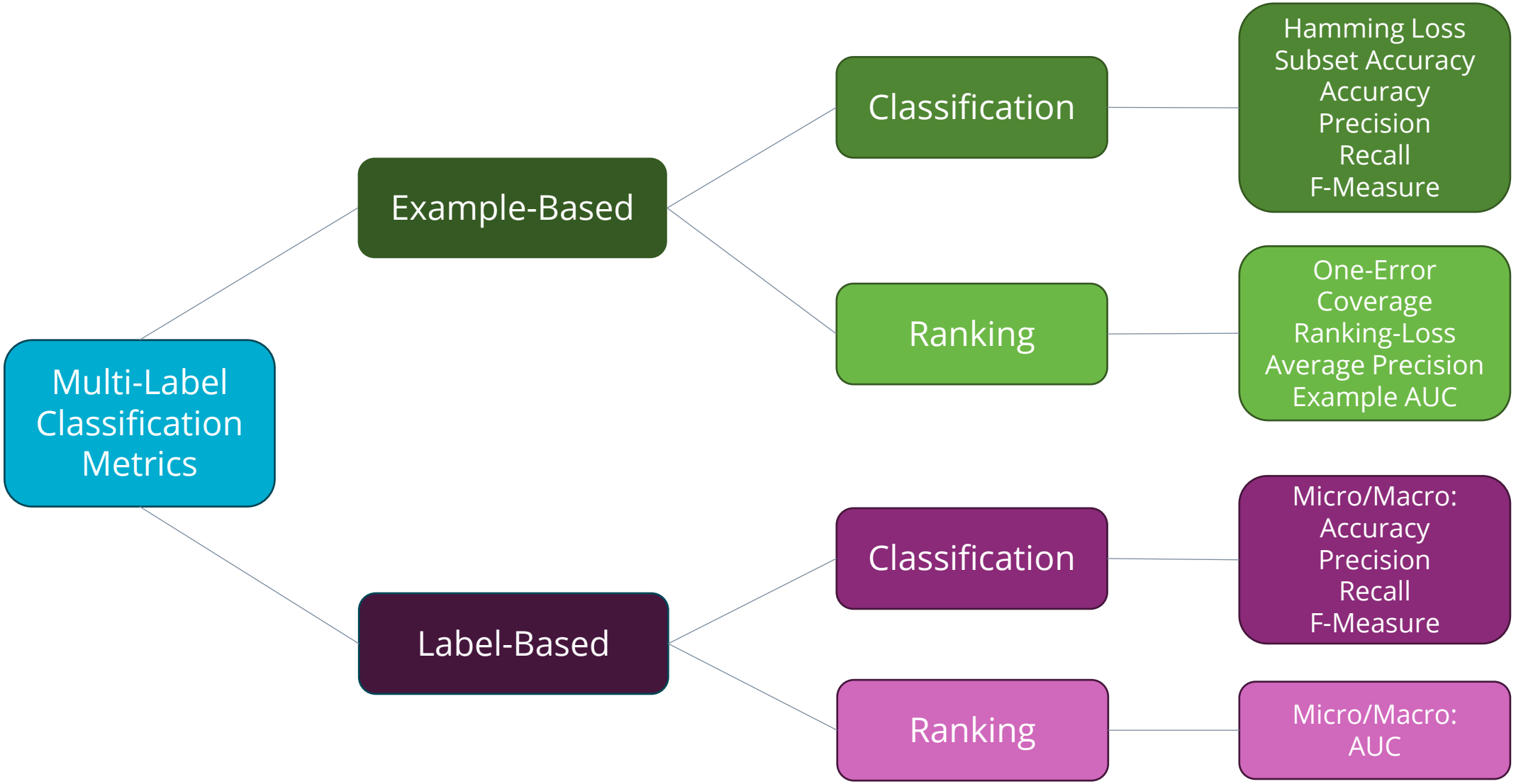
19 METRICS TO EVALUATE MULTI-LABEL CLASSIFICATION



19 METRICS TO EVALUATE MULTI-LABEL CLASSIFICATION

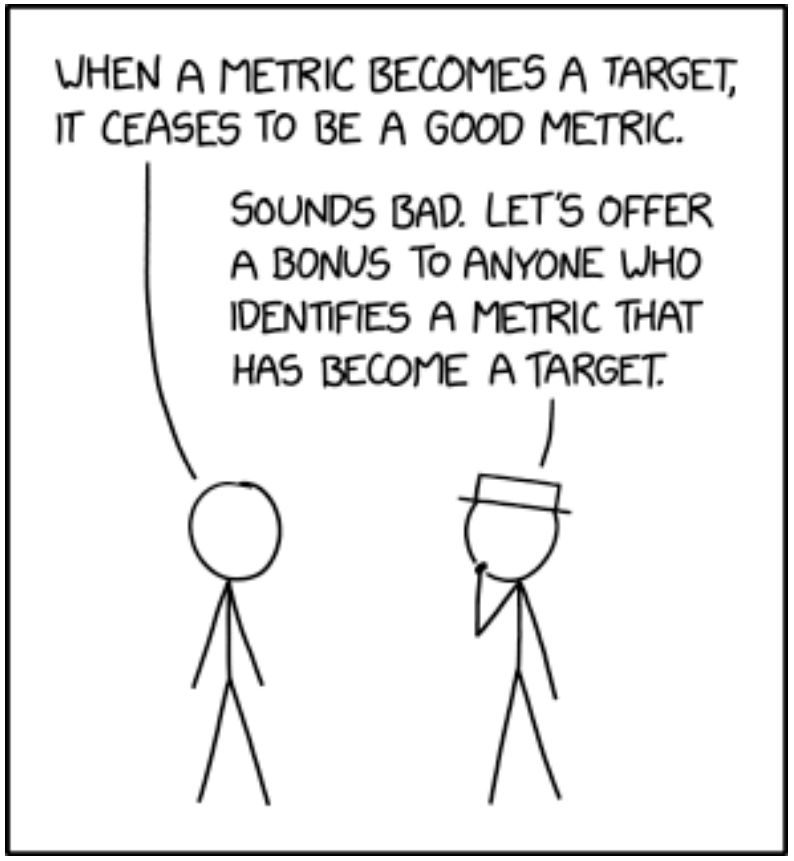


19 METRICS TO EVALUATE MULTI-LABEL CLASSIFICATION



IF I'M BUILDING A MULTI-LABEL CLASSIFICATION MODEL…

- I want to make modeling choices that result in the best performing model
- "Best performing" can mean different things
- Which of the 19 metrics do you choose???
  - Surely not all of them contain unique information
  - Its not immediately clear which aspects of model performance they might be evaluating



XKCD.com: Goodhart's Law

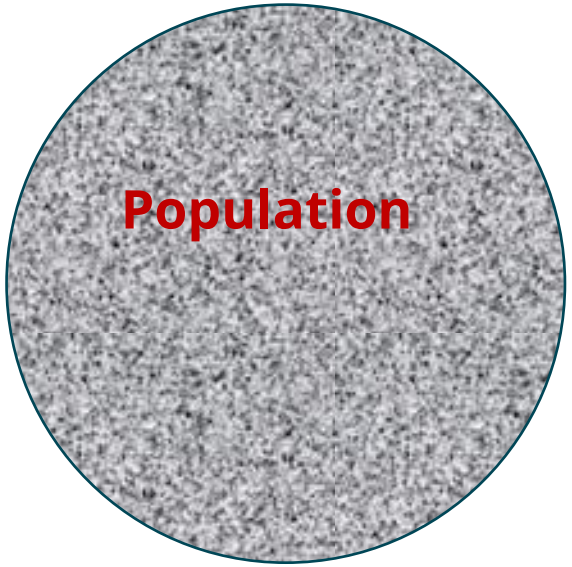
GOALS FOR EVALUATING METRICS

1. **Select a parsimonious set to evaluate performance overall**

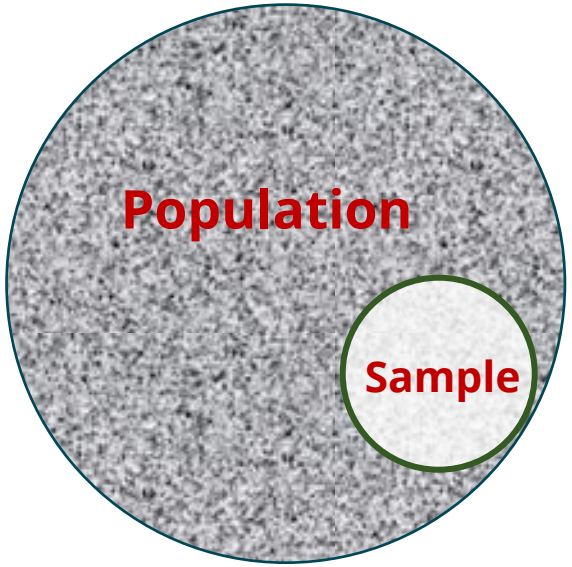
2. **Tie specific metrics to specific aspects of performance**

For the purposes of this talk we are focusing on label imbalance.

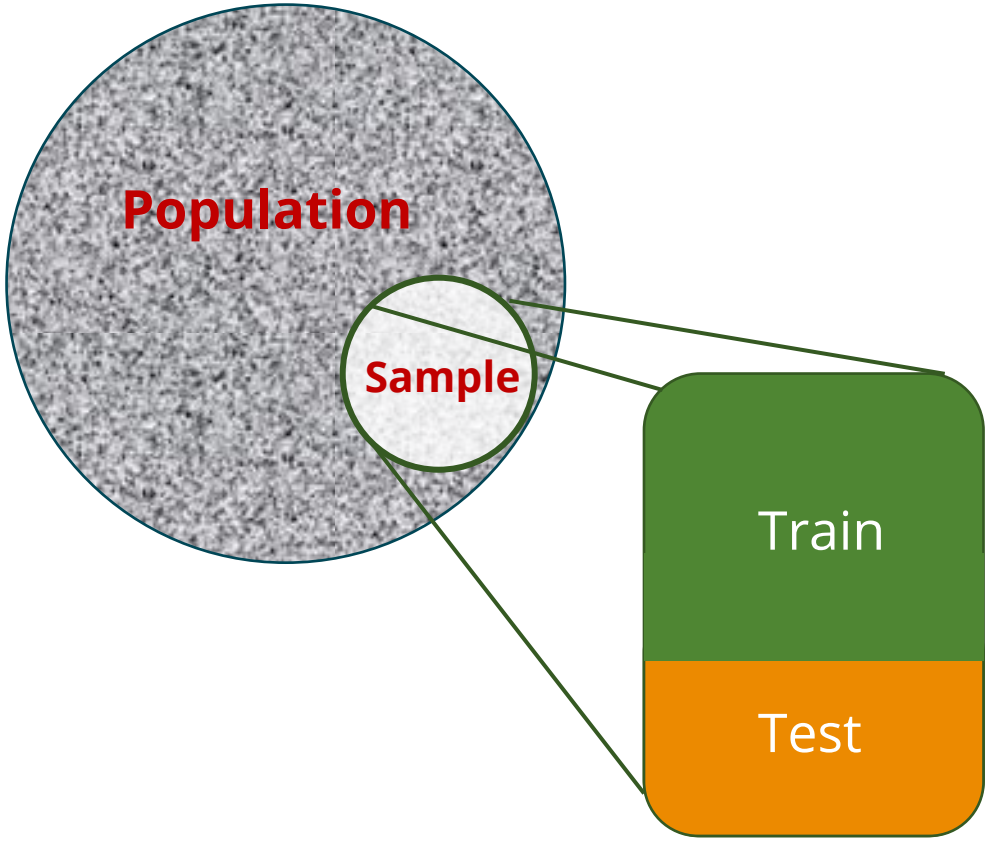
PROCESS TO SIMULATE



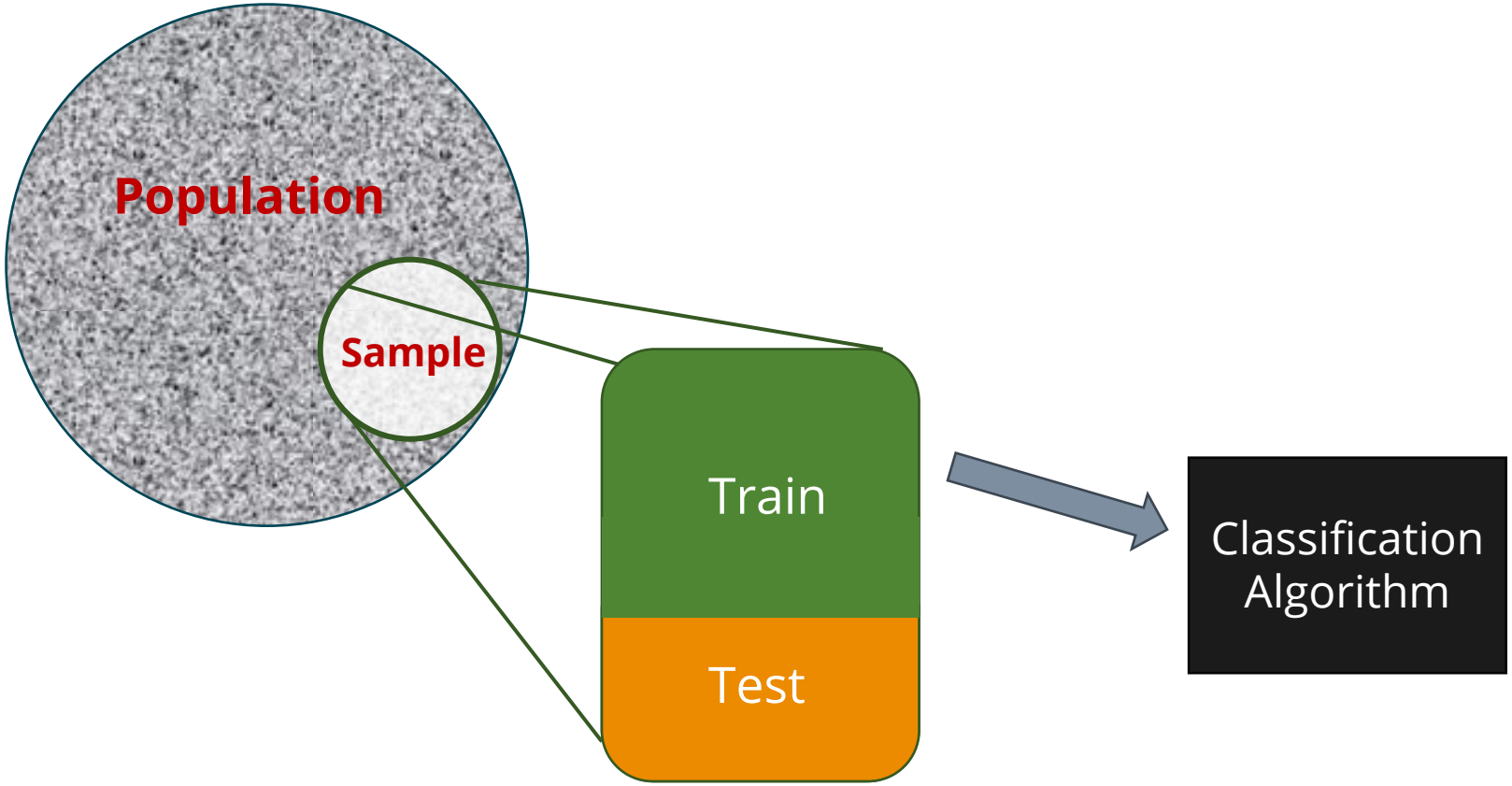
PROCESS TO SIMULATE



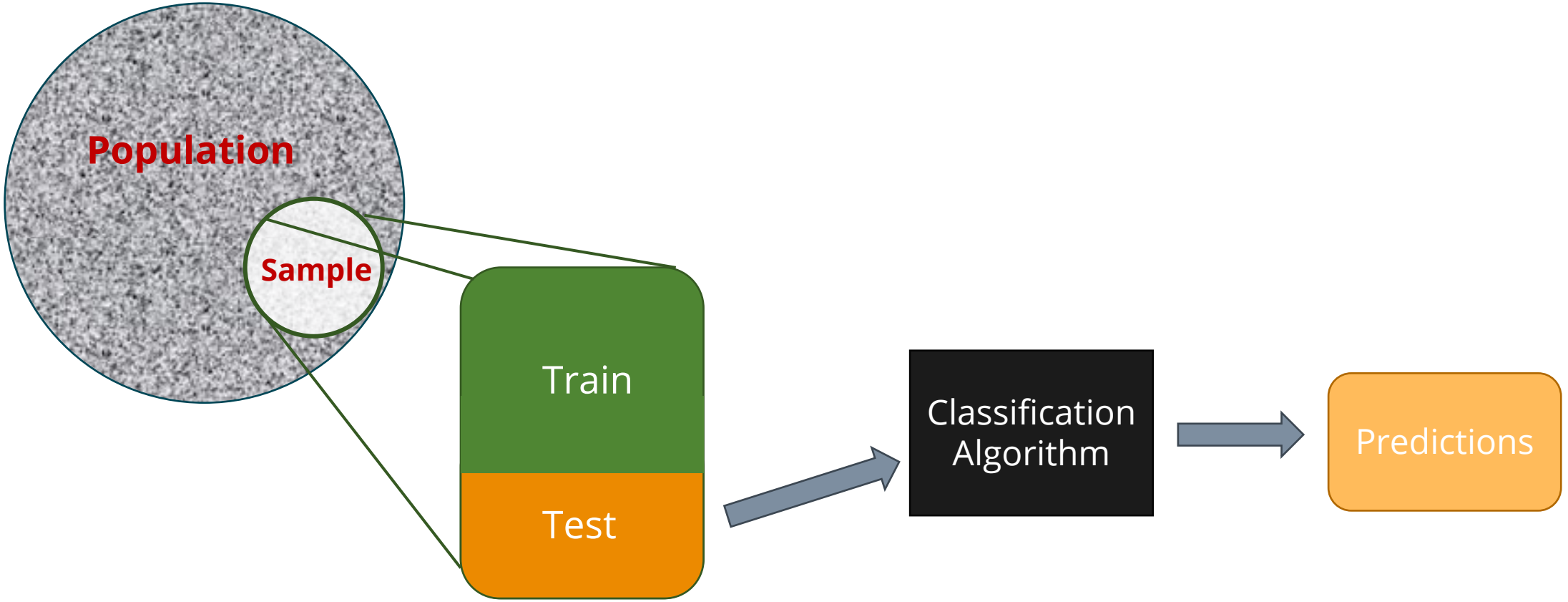
PROCESS TO SIMULATE



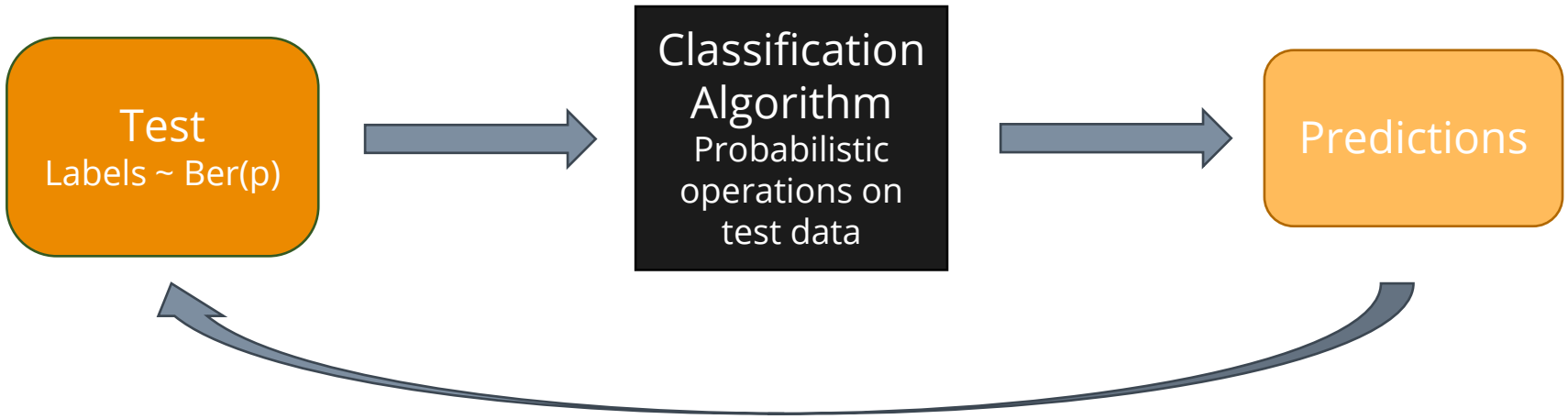
PROCESS TO SIMULATE



PROCESS TO SIMULATE



SIMULATION STUDY SETUP



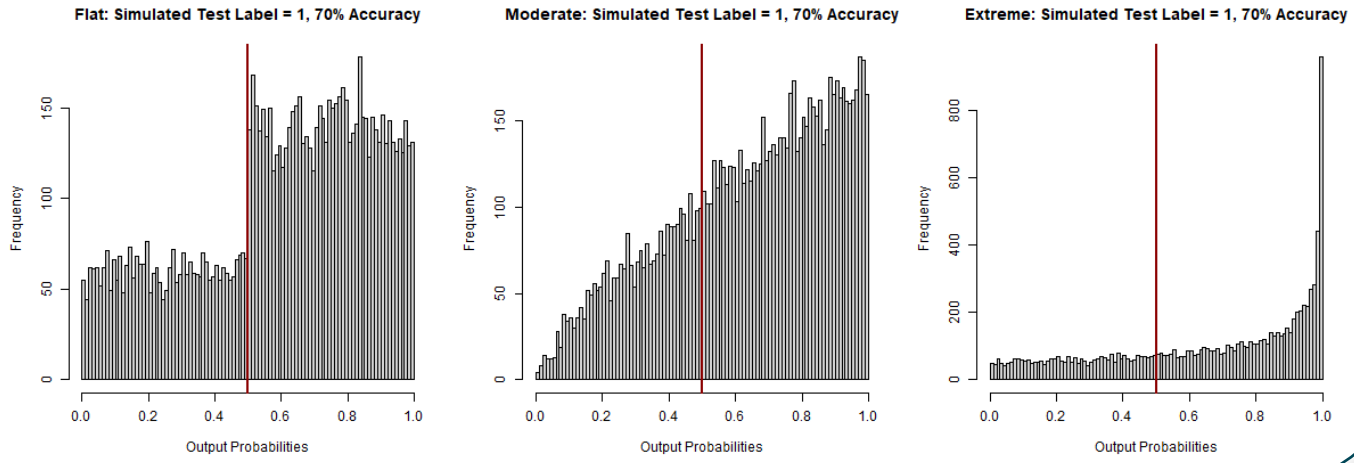
SIMULATION SETTINGS

Data Characteristics

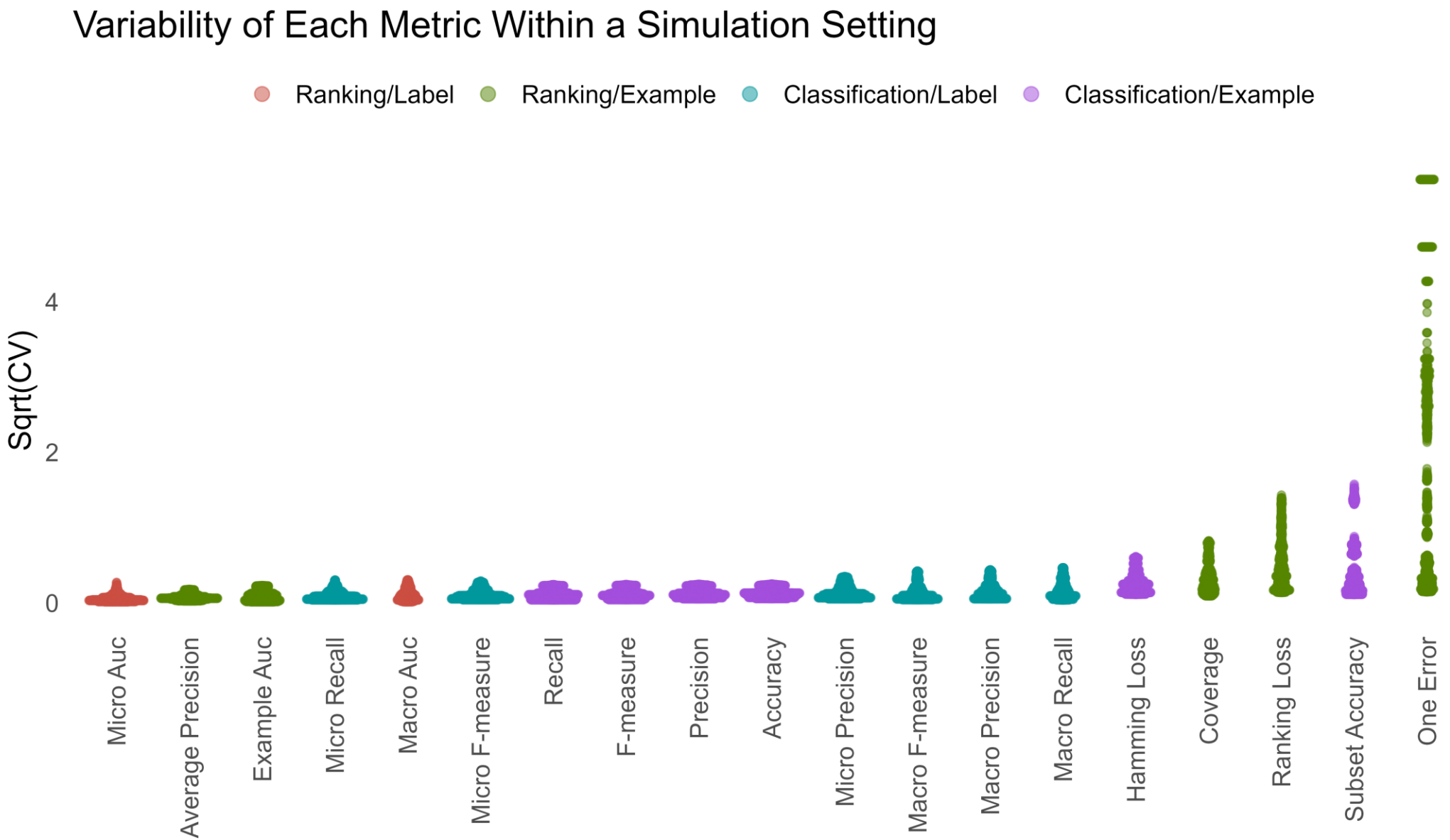
- Label types
  - Common
  - Moderate
  - Rare
- Number of total labels
- Prevalence of labels
- Number of instances

Classifier Characteristics

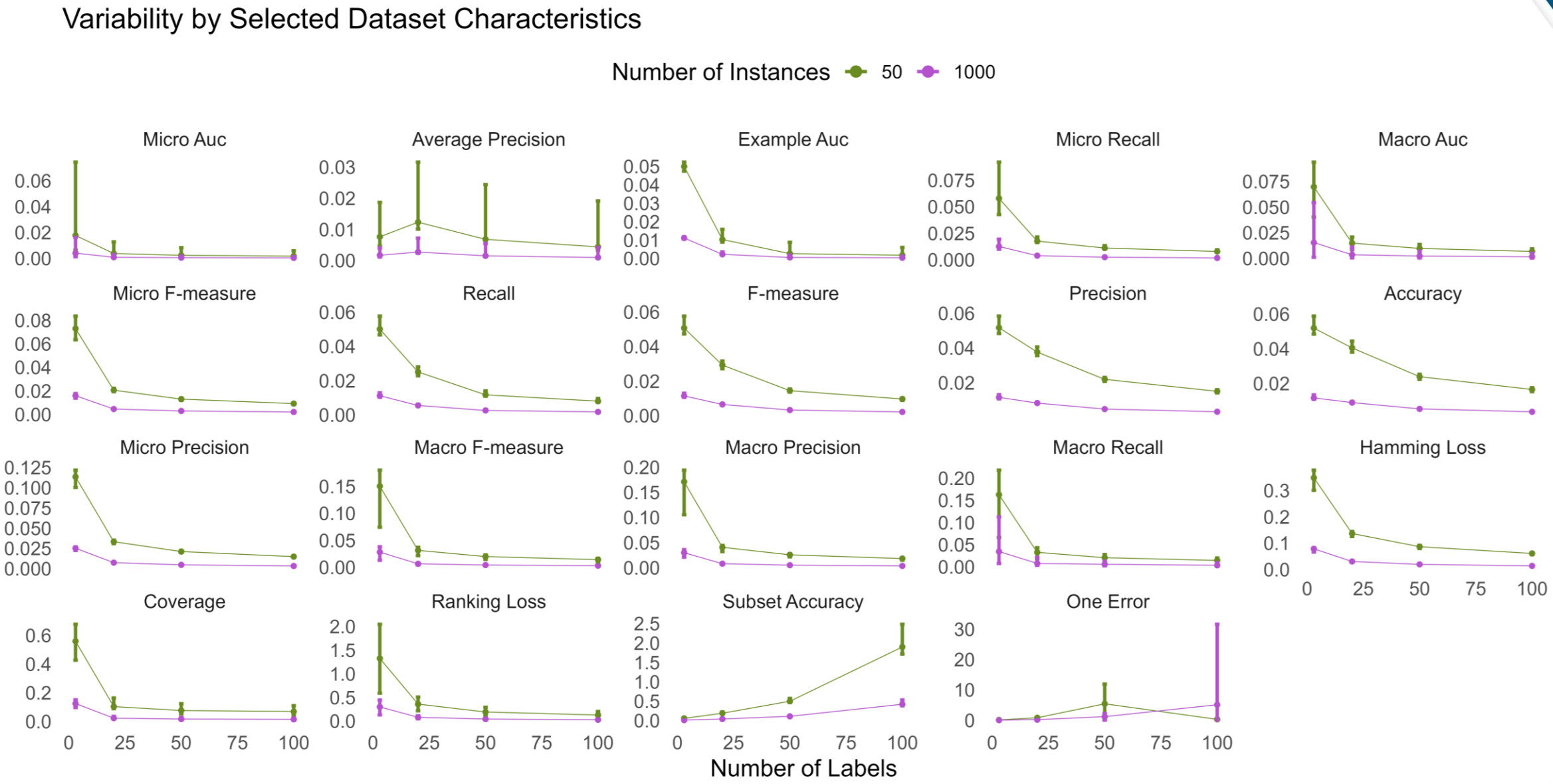
- Accuracy of each label type
- Precision and recall of rare label(s)
- Distribution of probabilities



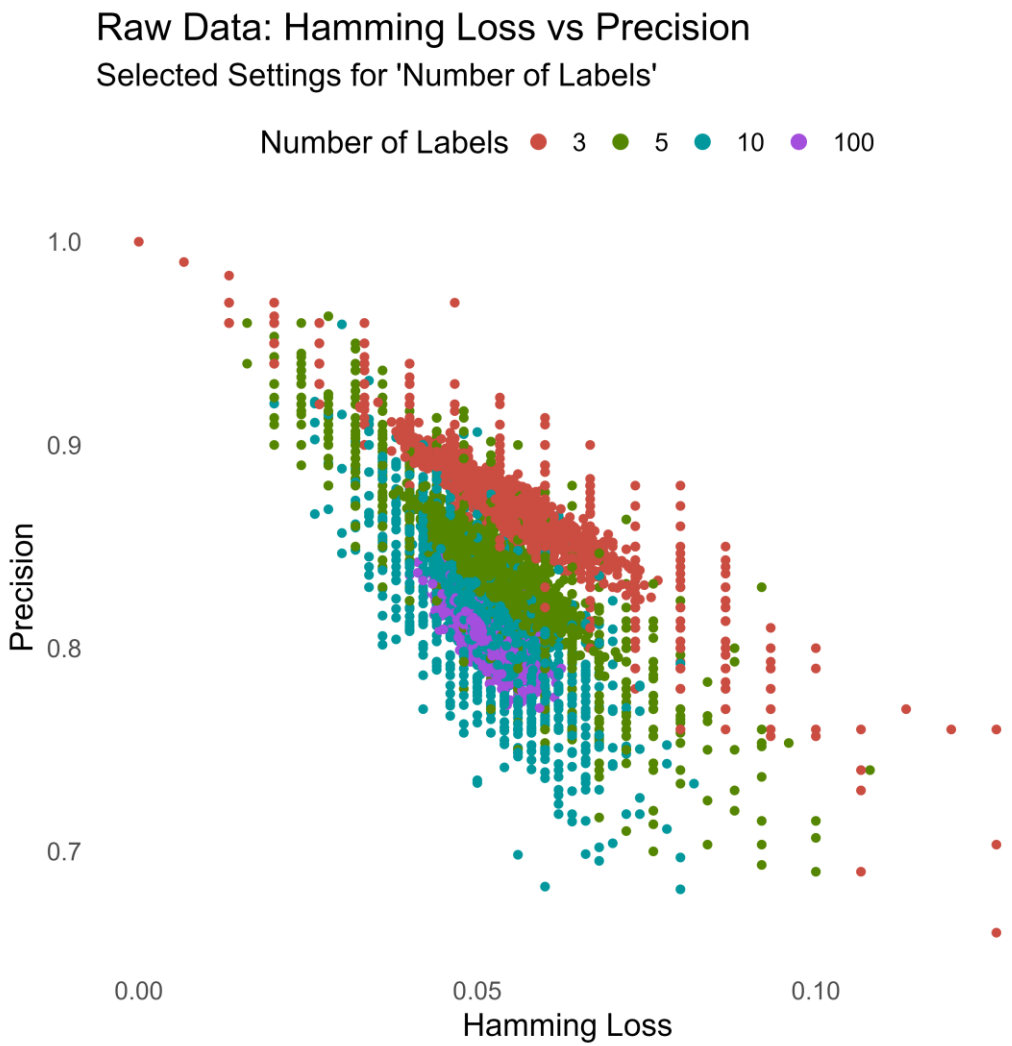
METRICS CAN VARY



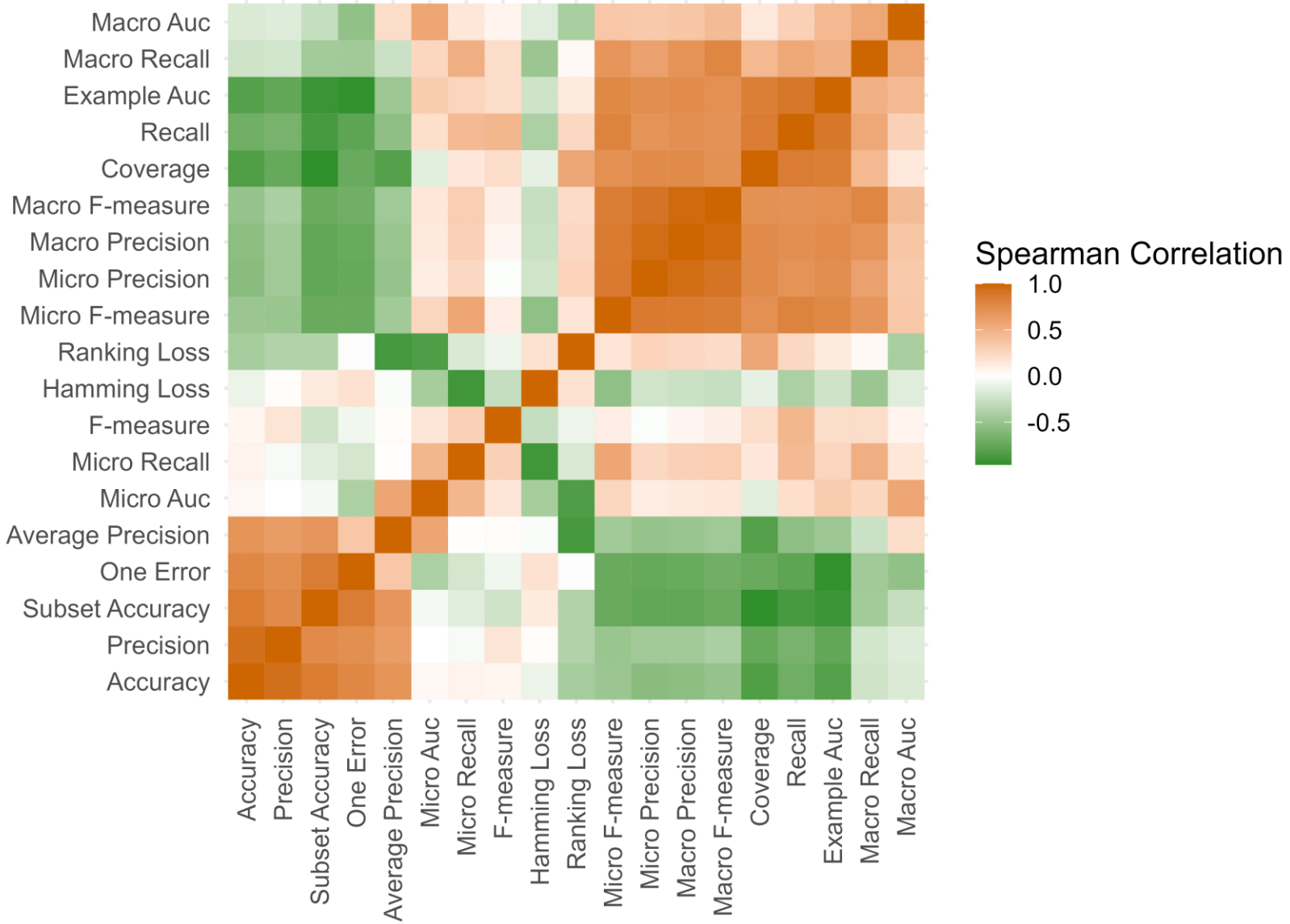
DATA STRUCTURE IS A MAJOR SOURCE OF VARIABILITY



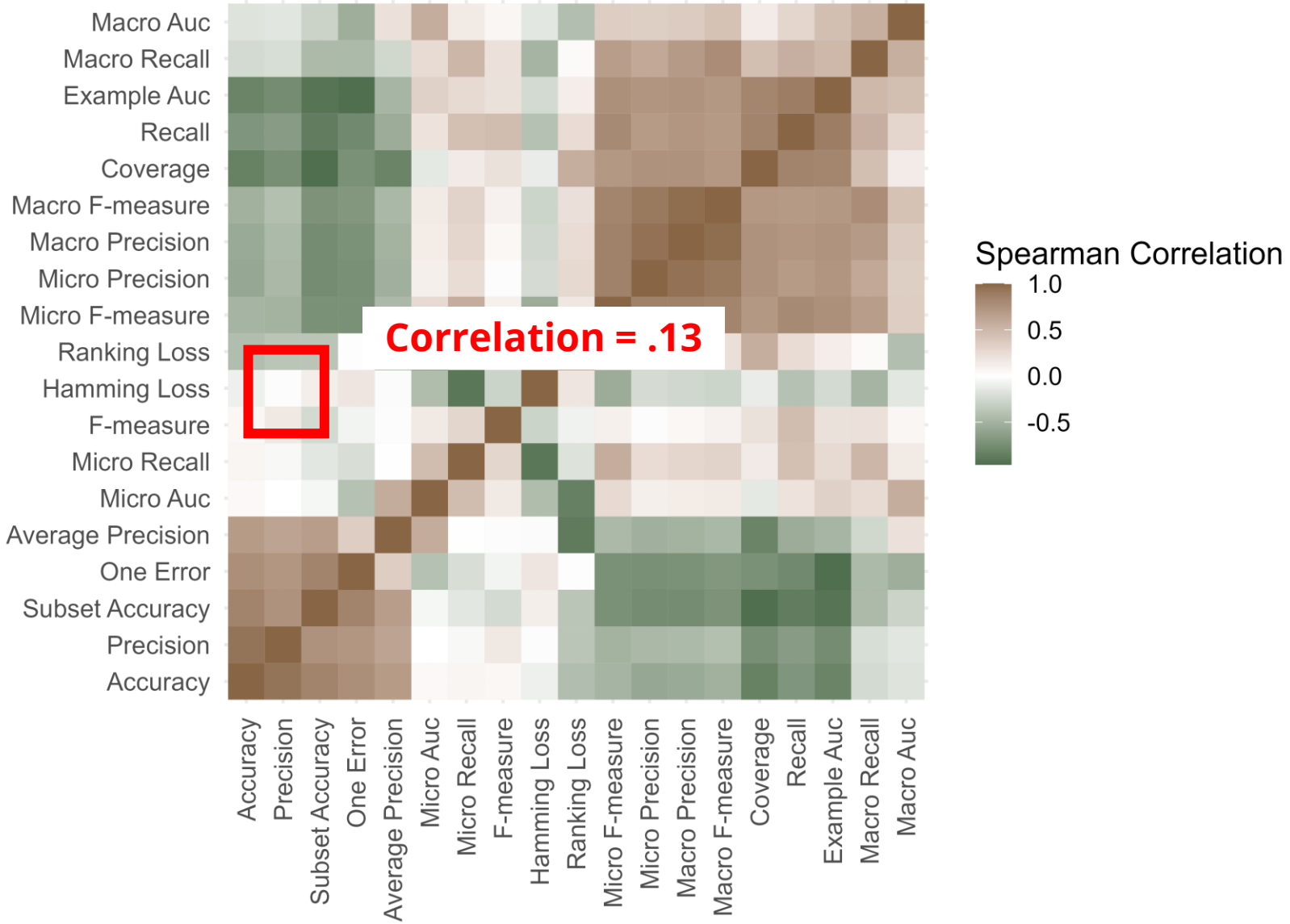
AGGREGATING ACROSS SIMULATION SETTING CLARIFIES RELATIONSHIPS



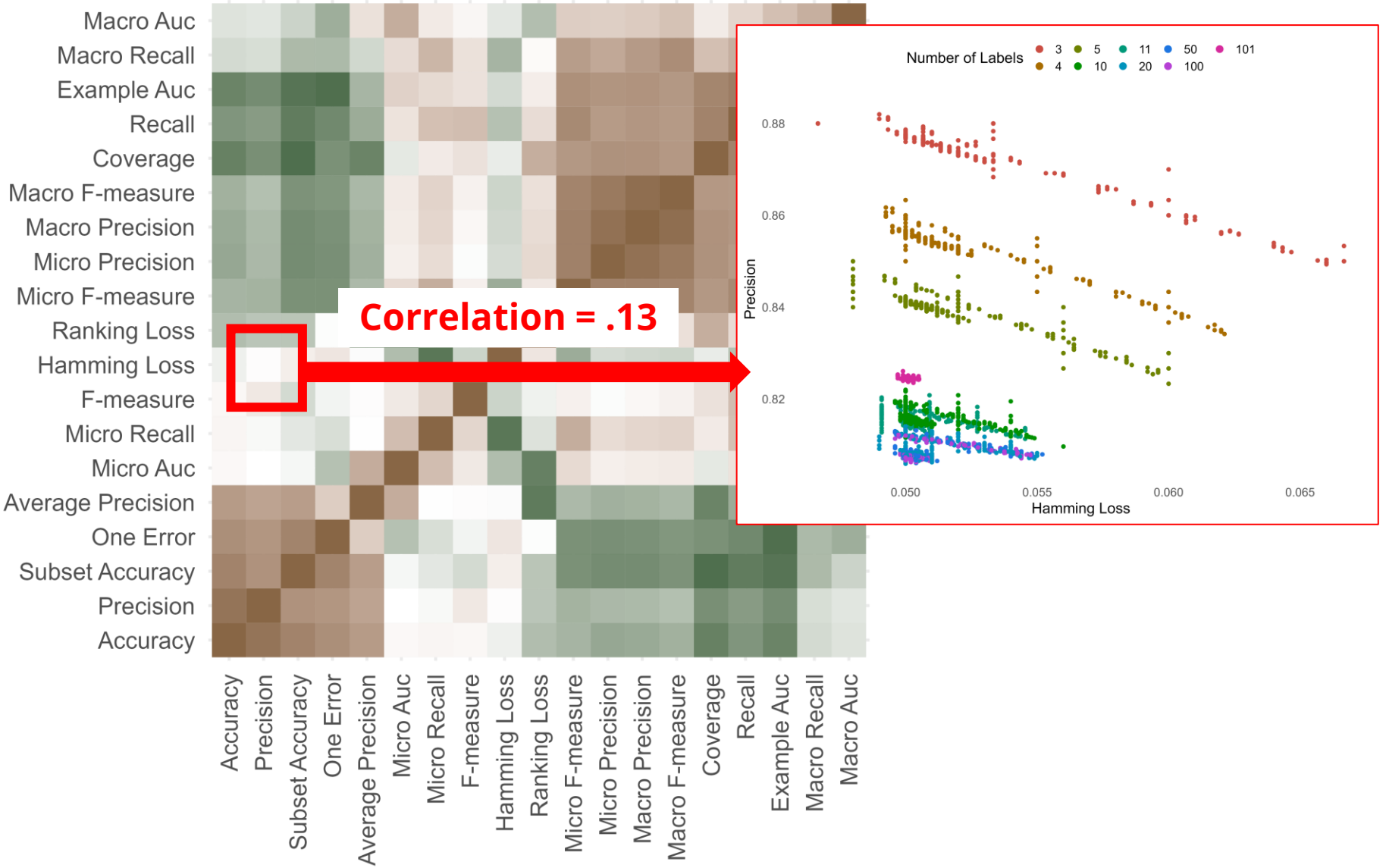
HOW DO METRICS RELATE TO ONE ANOTHER?



HOW DO METRICS RELATE TO ONE ANOTHER?



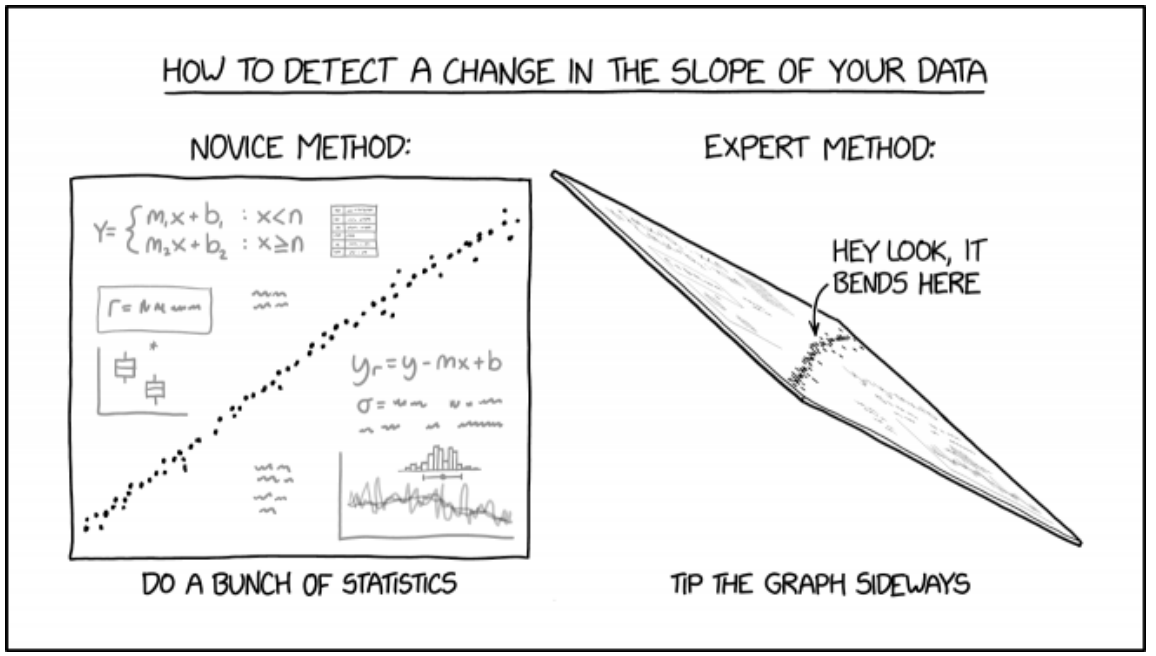
HOW DO METRICS RELATE TO ONE ANOTHER?



CALCULATING PARTIAL CORRELATION

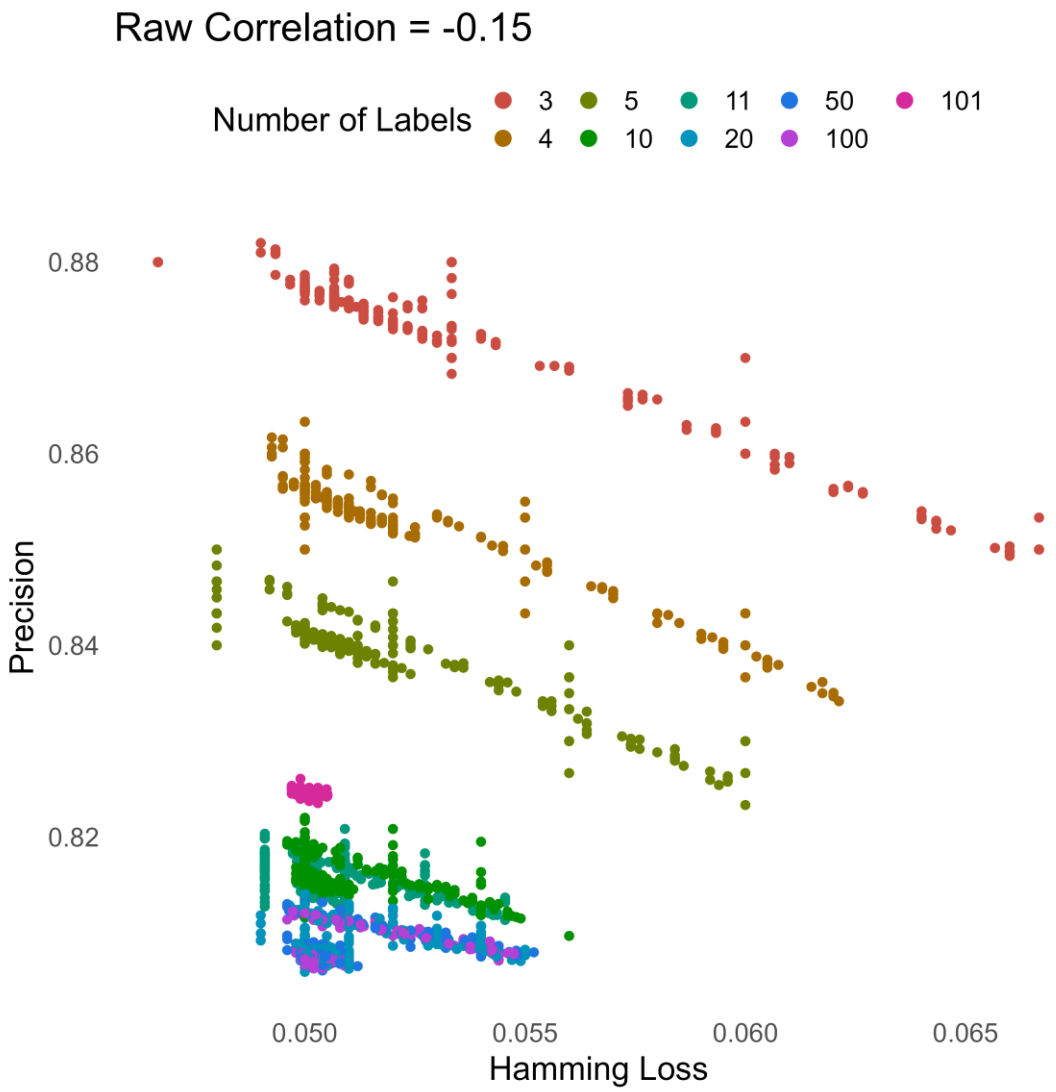
- Partial correlation quantifies the relationship between two variables, while **adjusting for the effects of others**
- To calculate partial correlation of X and Y, adjusting for Z:
  - Regress X on Z: X = Z + residual(X)
  - Regress Y on X: Y = Z + residual(Y)
  - Correlate the residuals from each:

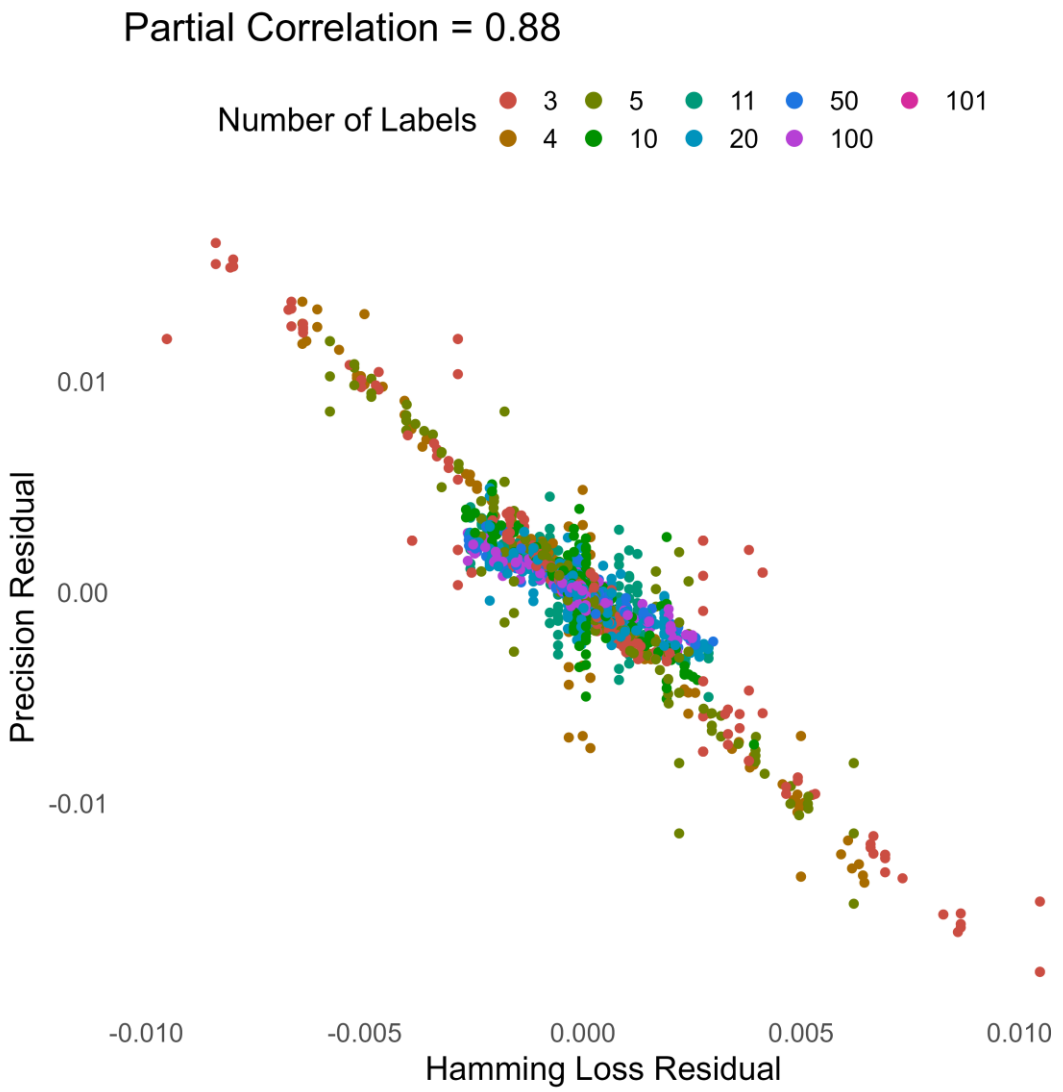
    cor[residual(X), residual(Y)]



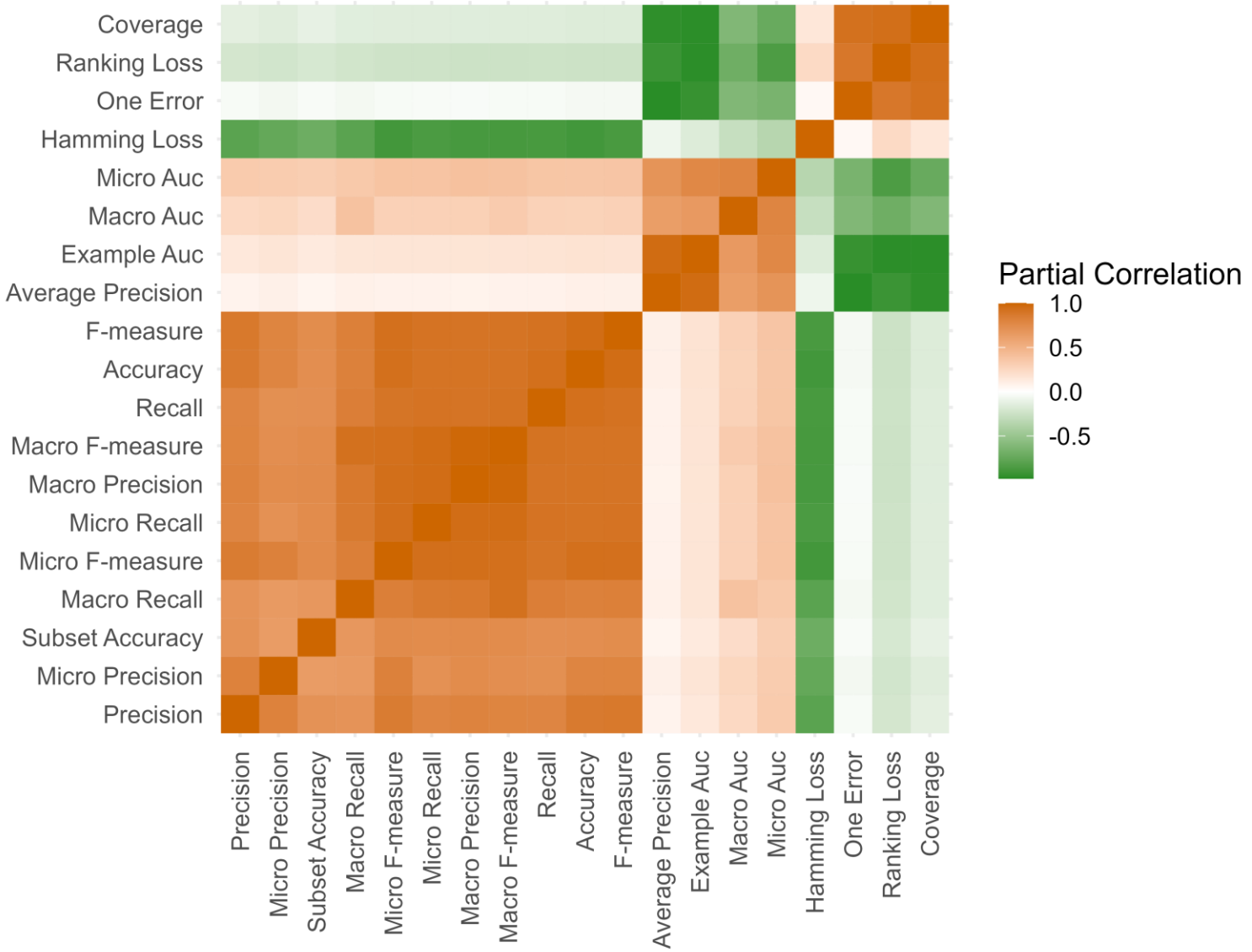
XKCD.com: Change in Slope

APPLYING PARTIAL CORRELATION

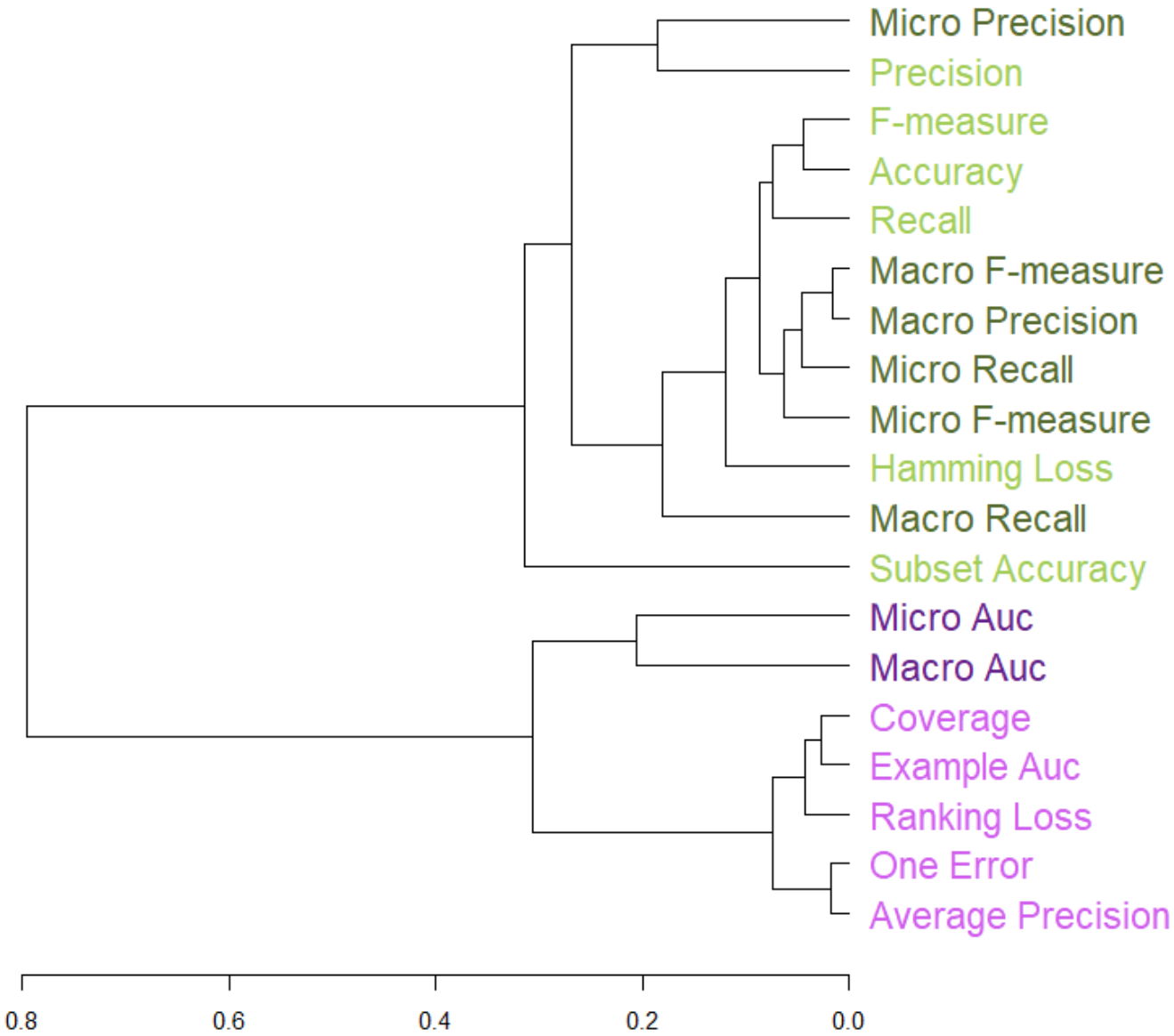




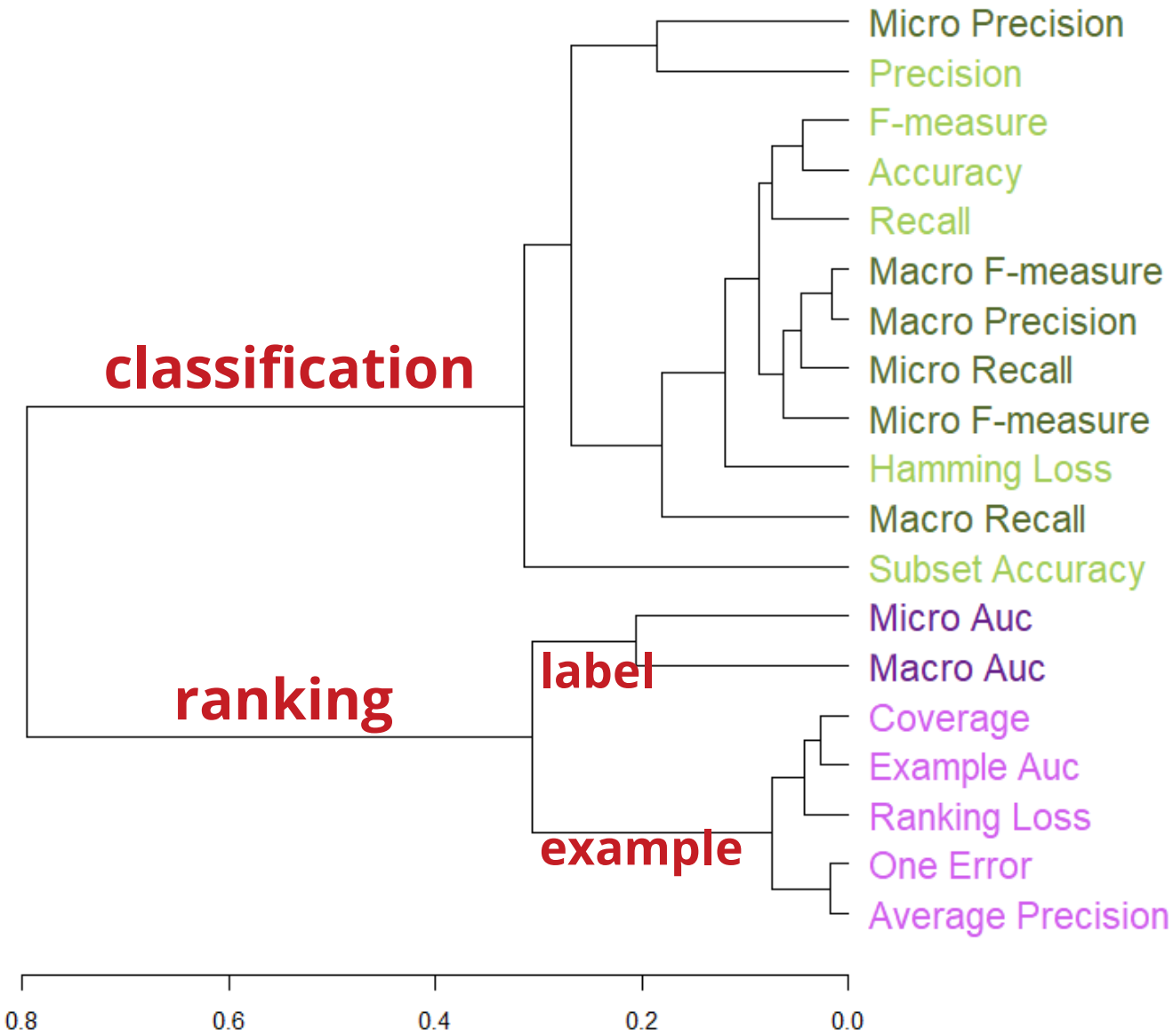
RECALCULATING METRIC RELATIONSHIPS



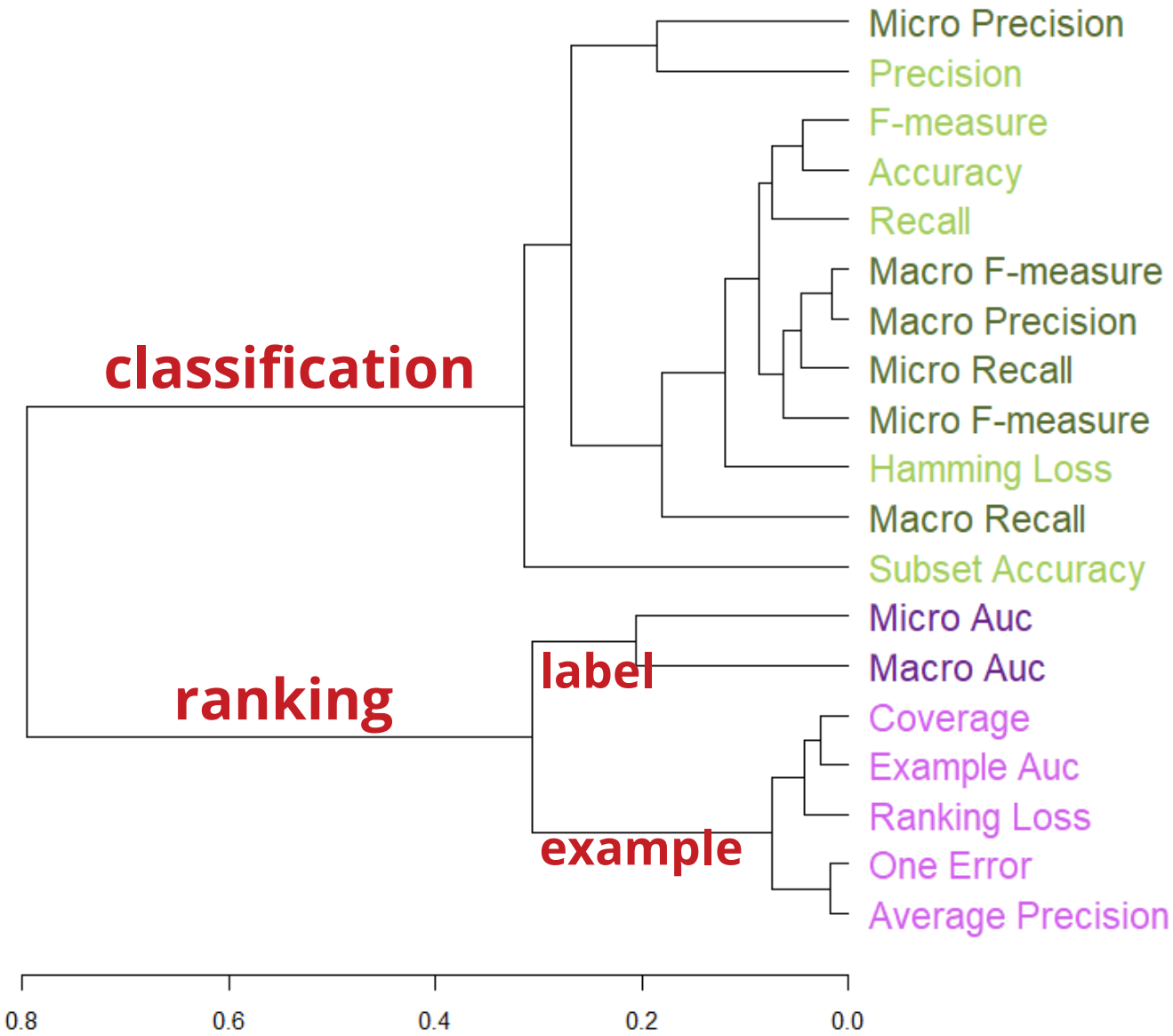
HIERARCHICAL CLUSTERING USING CORRELATION DISTANCE

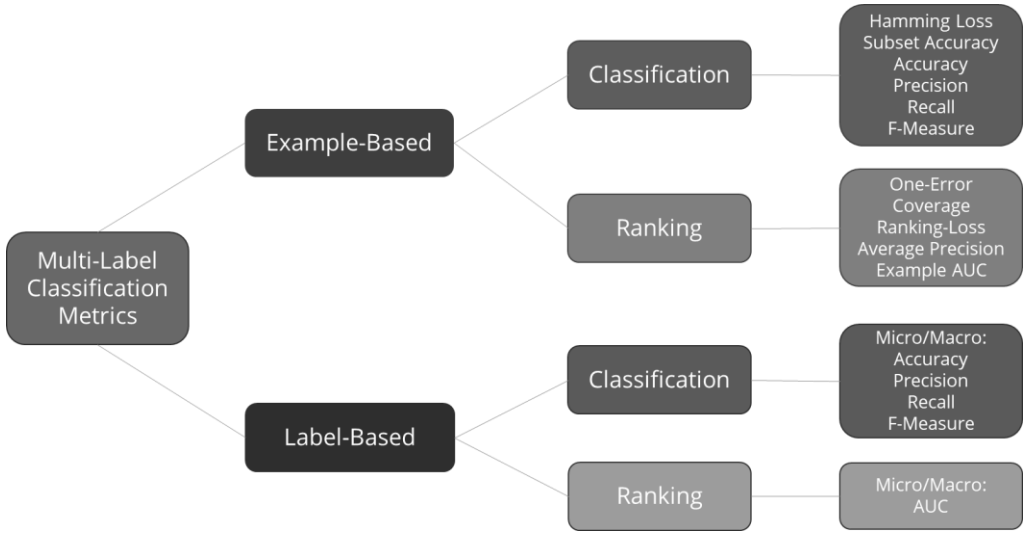


HIERARCHICAL CLUSTERING USING CORRELATION DISTANCE

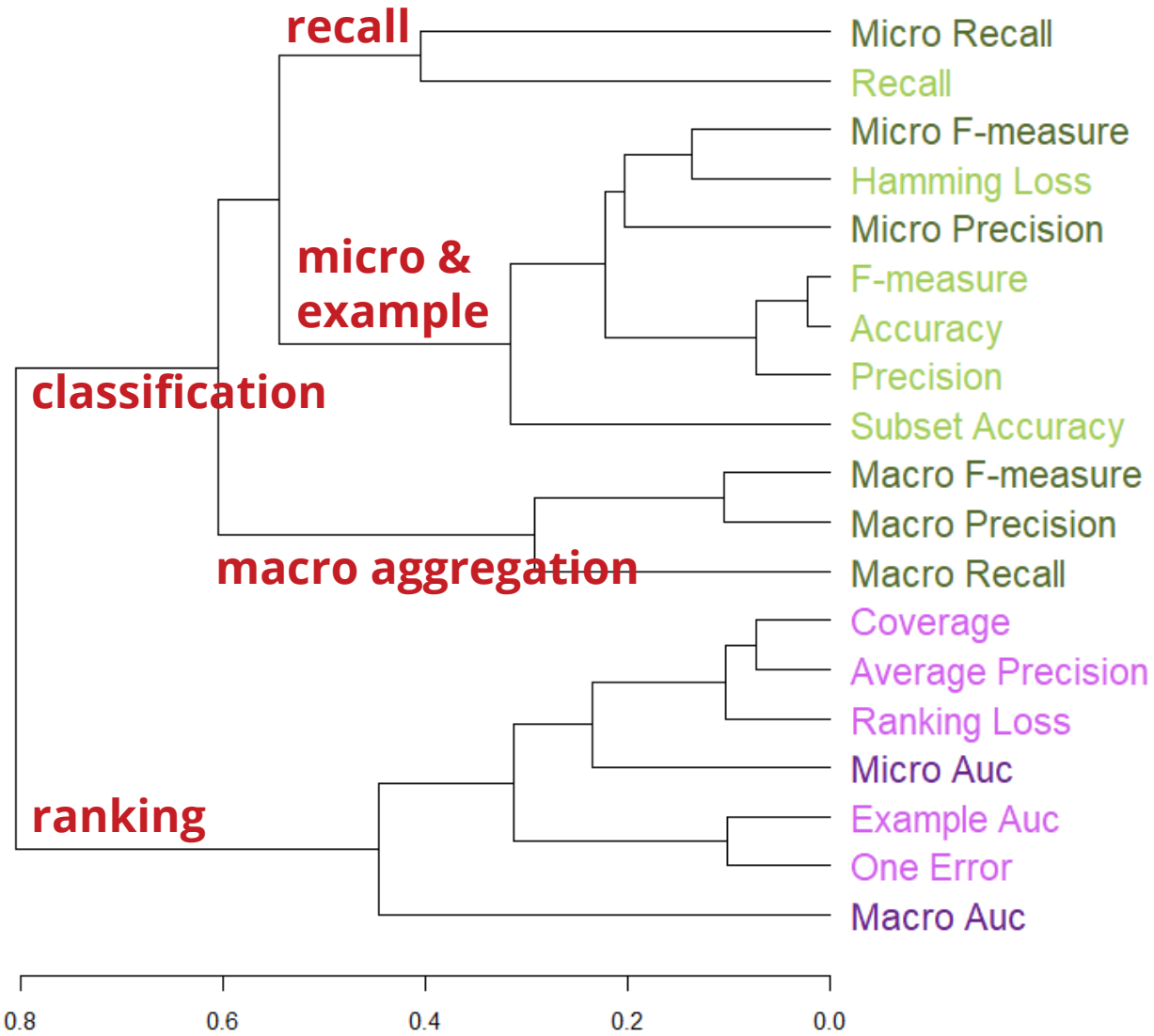


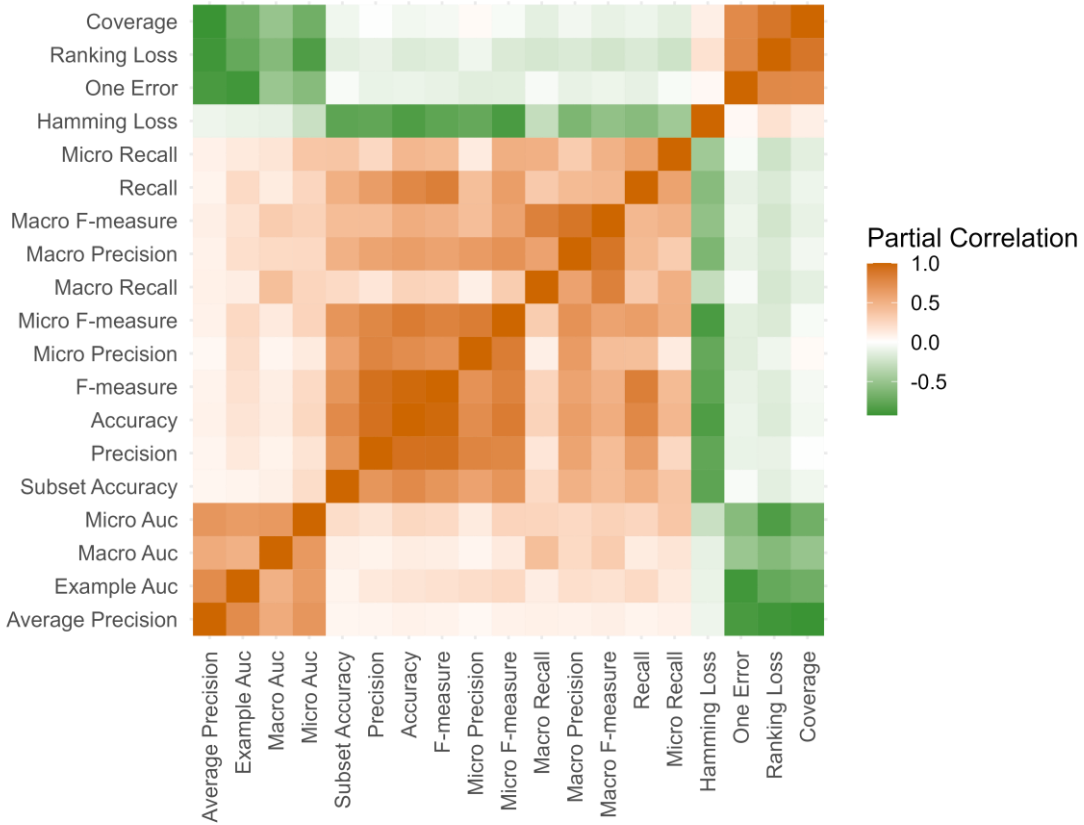
HEIRARCHICAL CLUSTERING USING CORRELATION DISTANCE





RESULTS USING UNAGGREGATED DATA





IMPLICATIONS OF THESE RESULTS

Table 1. Statistics for each evaluation measure, adapted from Spolaôr et al. (2013).

Evaluation measures	Number of papers
Hamming-Loss	55
Accuracy	26
F-Measure	18
Precision	18
Recall	18
Micro F-Measure	15
Macro F-Measure	12
Subset-Accuracy	10
Average Precision	10
Ranking Loss	8
Coverage	8
One Error	7
Macro Precision	5
Micro Precision	4
Subset 0/1 loss	3
Micro Recall	3
Macro Recall	2
Micro AUC	1
Macro AUC	1

Table taken from [1]

IMPLICATIONS OF THESE RESULTS

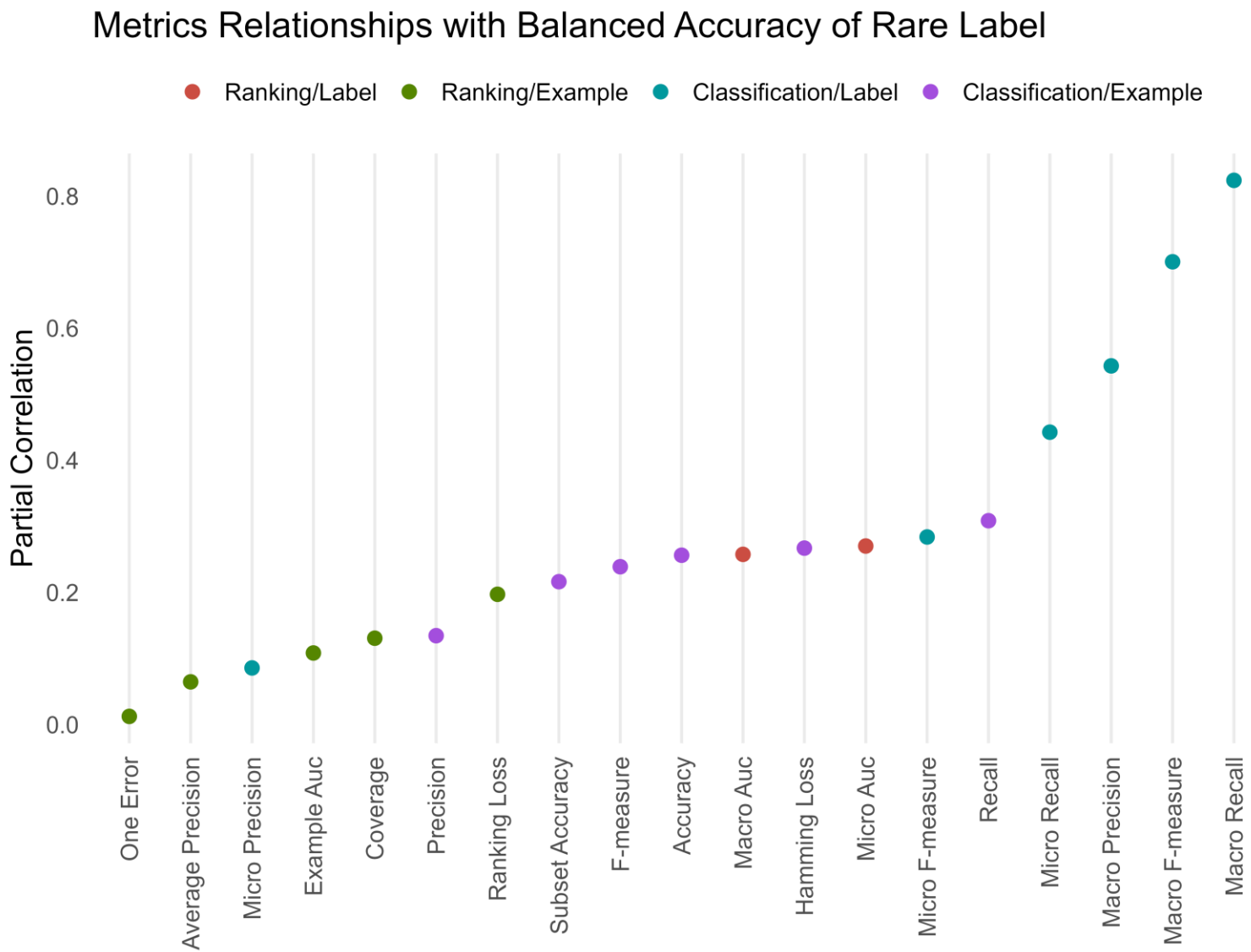
Table 1. Statistics for each evaluation measure, adapted from Spolaôretal.(2013).

Evaluation measures	Number of papers
Hamming-Loss	55
Accuracy	26
F-Measure	18
Precision	18
Recall	18
Micro F-Measure	15
Macro F-Measure	12
Subset-Accuracy	10
Average Precision	10
Ranking Loss	8
Coverage	8
One Error	7
Macro Precision	5
Micro Precision	4
Subset 0/1 loss	3
Micro Recall	3
Macro Recall	2
Micro AUC	1
Macro AUC	1

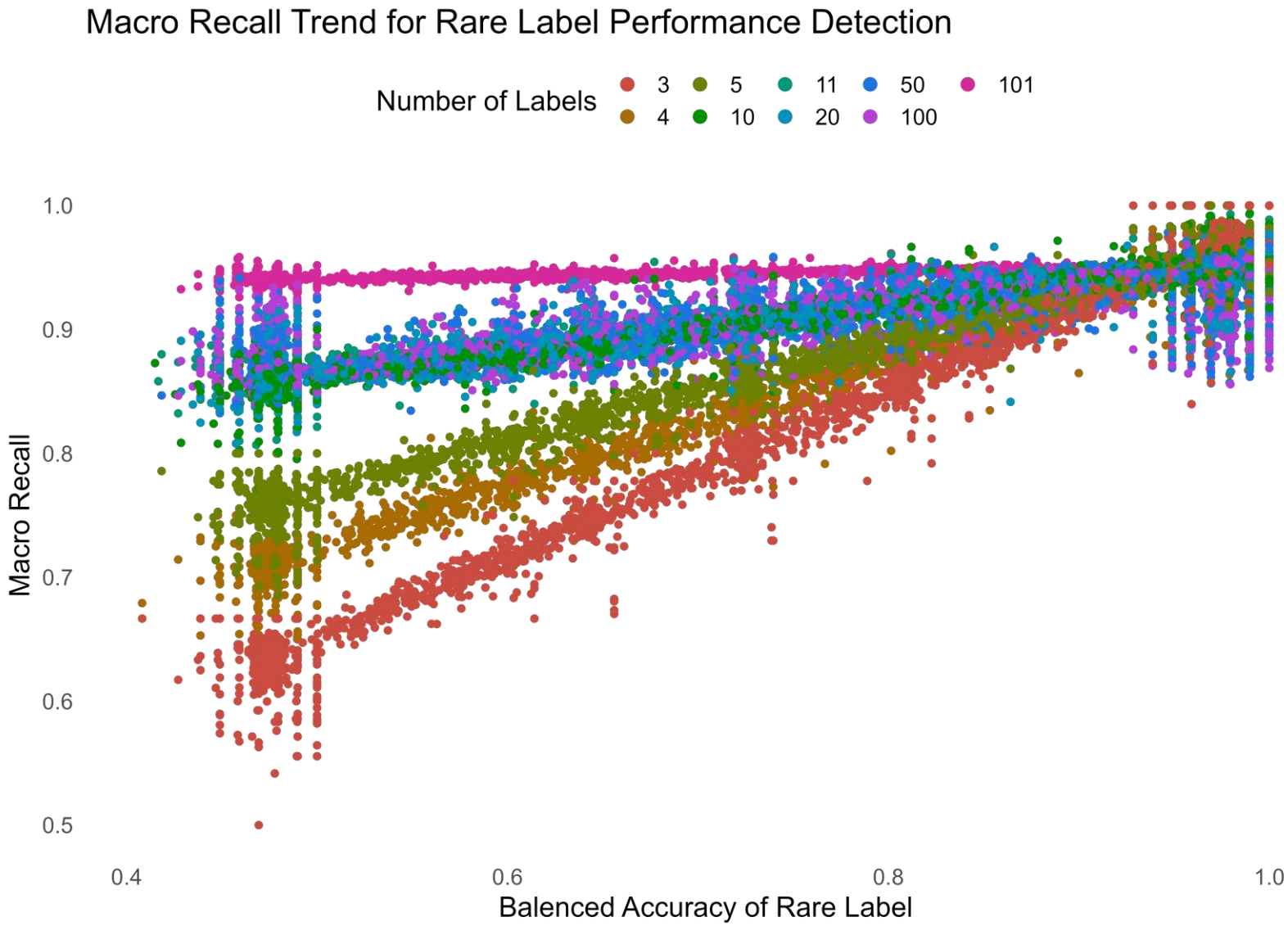
These are all the same type!

Table taken from [1]

WHAT ABOUT DETECTING ASPECTS OF PERFORMANCE?



MICRO RECALL AND IMBALANCE



MICRO RECALL AND IMBALANCE

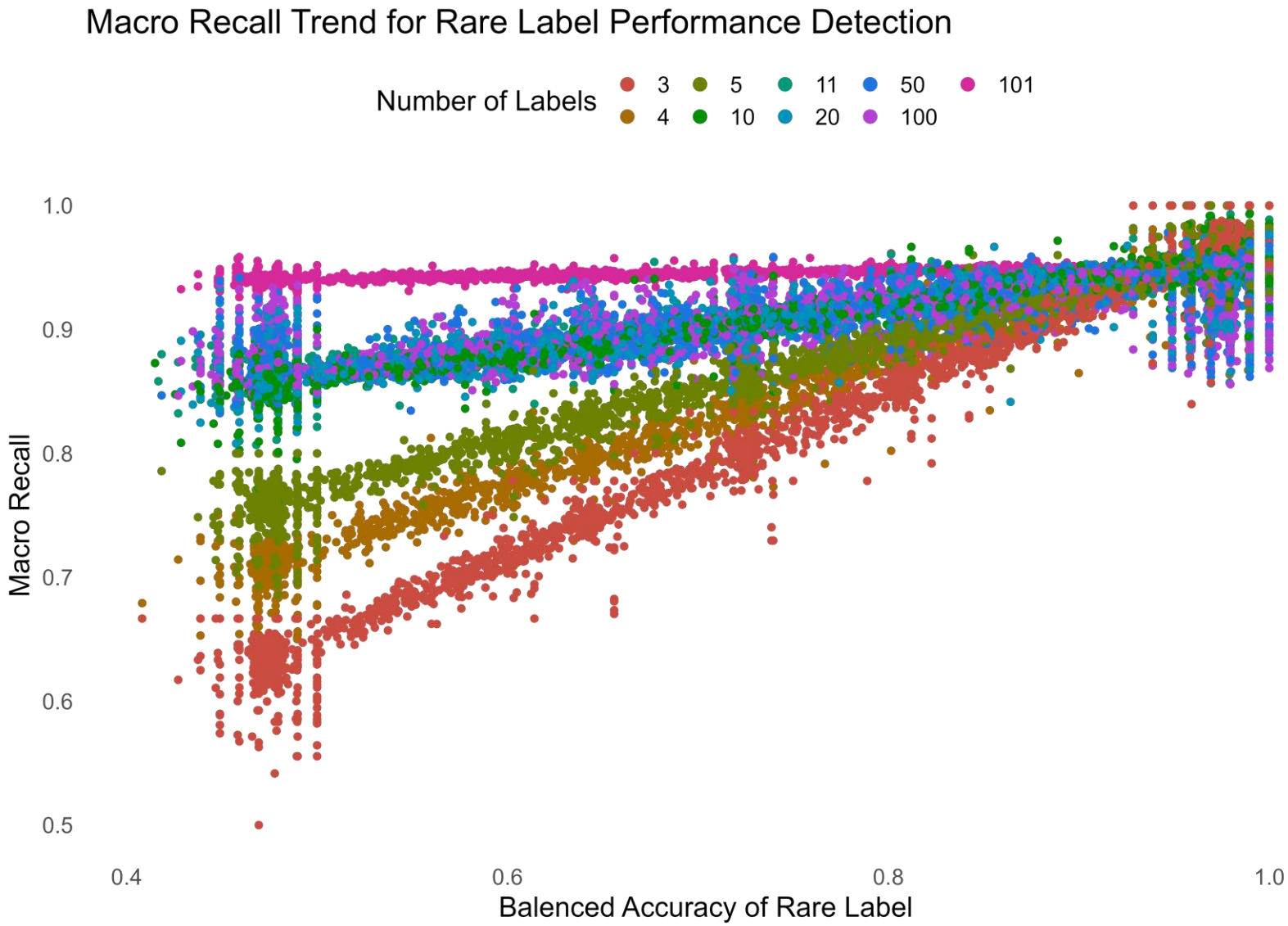


Table 1. Statistics for each evaluation measure, adapted from Spolaôretal.(2013).

Evaluation measures	Number of papers
Hamming-Loss	55
Accuracy	26
F-Measure	18
Precision	18
Recall	18
Micro F-Measure	15
Macro F-Measure	12
Subset-Accuracy	10
Average Precision	10
Ranking Loss	8
Coverage	8
One Error	7
Macro Precision	5
Micro Precision	4
Subset 0/1 loss	3
Micro Recall	3
Macro Recall	2
Micro AUC	1
Macro AUC	1

KEY TAKEAWAYS

- Metrics have sources of variability
  - Dataset characteristics really matter
  - Difficult to compare metrics between datasets
  - Implication: UQ for metrics is needed to give a rounder view of performance
- Metrics are related
  - The strength of those relationships depends on dataset characteristics
  - They seem to fall roughly into ranking vs classification and label vs example
  - Implication: for a parsimonious set one of each should be used
- When assessing sensitivity to a rare label, macro recall seems to work the best

FUTURE WORK

- Test this in real datasets with real classifiers!
- Include other types of performance issues
  - Correlations
  - Hierarchical dependencies
- Human interpretations of the metrics
- UQ for metrics is a criminally underdeveloped area of research

REFERENCES AND RELEVANT LITERATURE

Note: Stock images come from PowerPoint and Sandia's catalogue and comics come from XKCD.com

[1] Pereira, R. B., Plastino, A., Zadrozny, B., & Merschmann, L. H. (2018). Correlation analysis of performance measures for multi-label classification. *Information Processing & Management*, *54*(3), 359-369.

[2] Wu, X. Z., & Zhou, Z. H. (2017, July). A unified view of multi-label performance measures. In *international conference on machine learning* (pp. 3780-3788). PMLR.

[3] Herrmann, M., Lange, F. J. D., Eggensperger, K., Casalicchio, G., Wever, M., Feurer, M., ... & Bischl, B. (2024). Position: Why we must rethink empirical research in machine learning. *arXiv preprint arXiv:2405.02200*.

[4] Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., ... & Vinyals, O. (2021). The benchmark lottery. *arXiv preprint arXiv:2107.07002*.

[5] Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., ... & Vincent, P. (2021). Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, *3*, 747-769.

[6] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

[7] Casella, G., & Berger, R. (2024). *Statistical inference*. CRC press.