- use clustering to generate the target variable + validation
- automatic annotation with LLMs

https://www.aidancooper.co.uk/supervised-clustering-shap-values/

https://direct.mit.edu/coli/article/49/1/157/113280/Annotation-Error-Detection-Analyzing-the-Past-and

- annotation error detection - flaggers vs scorers (there is a survey for both)
    - Classification Uncertainty
    - Confident Learning
    - Dropout Uncertainty
    - Curriculum and Leitner Spotter
    - Vector Space Proximity
    - artificial noise injected is easier to detect than human-introduced
    - Methods based on vector proximity—k-Nearest Neighbor Entropy (KNN) and Mean Distance (MD)—perform sub-par across tasks and datasets.
    - curse of dimensionality
- training with label noise - sampling a bias from a normal distribution?¿?¿?

https://www.superannotate.com/blog/how-to-detect-mislabeled-annotations

https://openreview.net/forum?id=hIOGpXTQnUq - confident learner multiclass clasificatrioln

semisuèrvised learning - https://labelyourdata.com/articles/unlabeled-data-in-machine-learning

- probably way less non-project related data?

https://www.tribe.ai/applied-ai/no-labels-are-all-you-need-how-to-build-nlp-models-using-little-to-no-annotated-data

teach-student models

https://arxiv.org/pdf/2403.15112

clustering with llm embeddings

https://arxiv.org/abs/2007.08199
learning with noisy labels overview