2. Quality check of silver/gold data

* Re-annotate existing data without looking at it
* Check where the disagreements are between human annotators and check if they overlap with the difficult cases for automatic systems (**analyzing where human annotators disagree** and **checking if those cases are also difficult for NER models)**

Gold data 🡪 **Manually annotated by experts**, following strict guidelines. Considered the highest quality.

Silver data 🡪 **Automatically labeled (or weakly supervised)** using AI models, heuristics, or crowd workers. May contain errors.

Steps:

1. analyze human annotation disagreements (use a NER dataset with many annotators)

* Different entity labels fot the same text
* Measure agreement (Inter-Annotator Agreement) using Cohen’s Kappa, Fleiss’ Kappa, or Krippendorff’s Alpha
* Find patterns

2. compare with automatic system errors

* Train a NER on the dataset
* Find what it struggle with

3. Find the Overlapping “Hard Cases”

* Check if **human disagreements** overlap with **model errors**
* Identify **linguistic patterns**

A black screen with white text

AI-generated content may be incorrect.

A black screen with white text

AI-generated content may be incorrect.

A black background with white text

AI-generated content may be incorrect.

A black screen with white text

AI-generated content may be incorrect.