



2026 CDISC China Interchange

From Project-Level Compliance to Enterprise-Level Consistency: AI-Driven CDISC Data Governance

Jiayu Xu



**The contents only represent personal views
of the presenter, not the position of the
employer or conference organizer.**



Agenda

1. The Problem We All Share
2. How AI Changes the Game
3. A Little Try
4. What's Next





The Problem We All Share



Every Study Passes P21. So What's the Problem?

- Project View

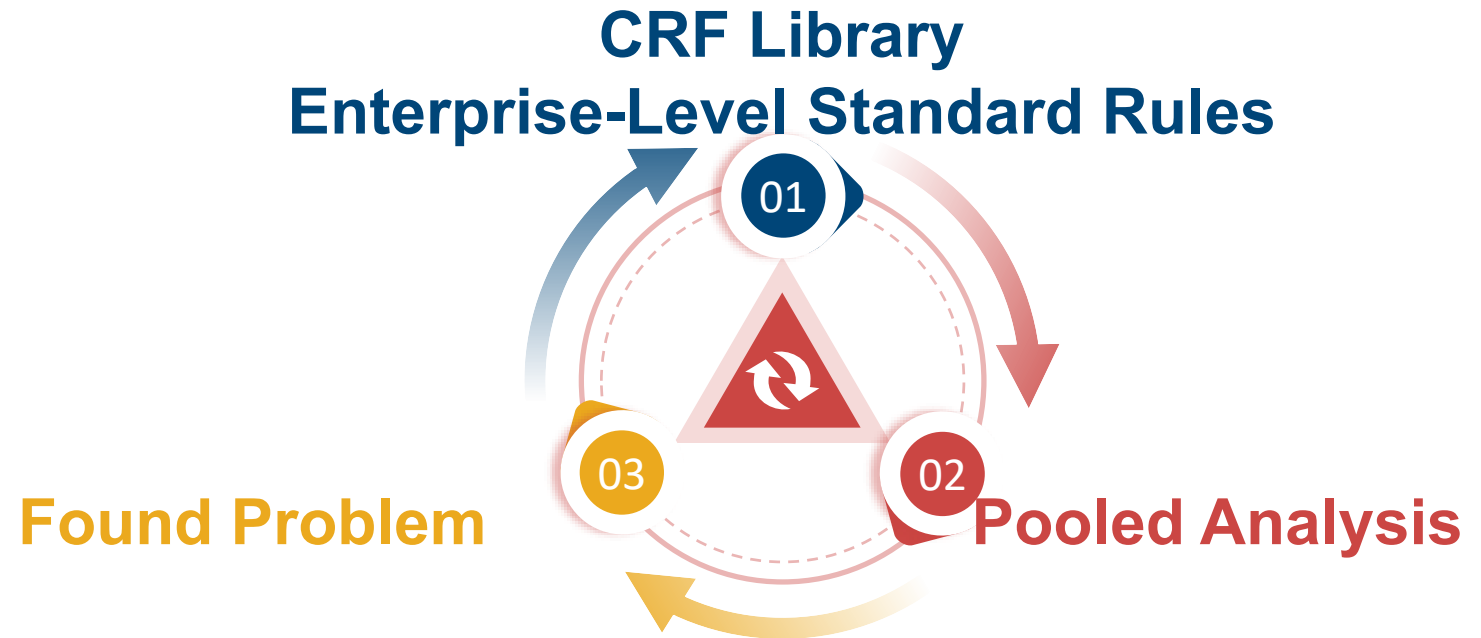
- Study after study passes P21 with green checks
- Each individual project is "CDISC compliant"
- From a single-study perspective: everything is perfect

- Enterprise View

- Pooled analysis (ISS / Data Warehouse) reveals incompatibilities
- Same variable — different derivations across studies
- Individual compliance $\neq$ enterprise consistency

The Answer Maybe?

- Setup CRF Library
- Setup Enterprise-Level Standard Rules
- Received a request for Pooled Analysis
- Find Problem during Pooled Analysis
- Decide whether to establish new Enterprise-Level Standard Rules based on Found Problem



The Tip of the Iceberg

Possible Reason

- The individual does not identify any issues who is conducting the pooled analysis.
- The individual did identify issues, but did not consider them as reportable.
- Different people have different perceptions.
- Only data used in the pooled analysis will be reviewed. What about the data that are not used?



What's the Impact?

- The discovery and reporting of issues rely on personnel. In the extreme, a workflow that performs adequately under one person may break down once that individual departs and another takes over.
- Future pooled analyses will likely identify new issues, which would delay the delivery of results.
- The rules formulated may not be sufficiently generalizable. Their development is constrained by the projects involved in the current pooled analysis.



The Problem We All Share

- The identification of subsequent rules relies on personnel.
- Issues can be identified only in the data used for the pooled analysis, whereas it is difficult to detect issues at the overall level.





How AI Changes the Game



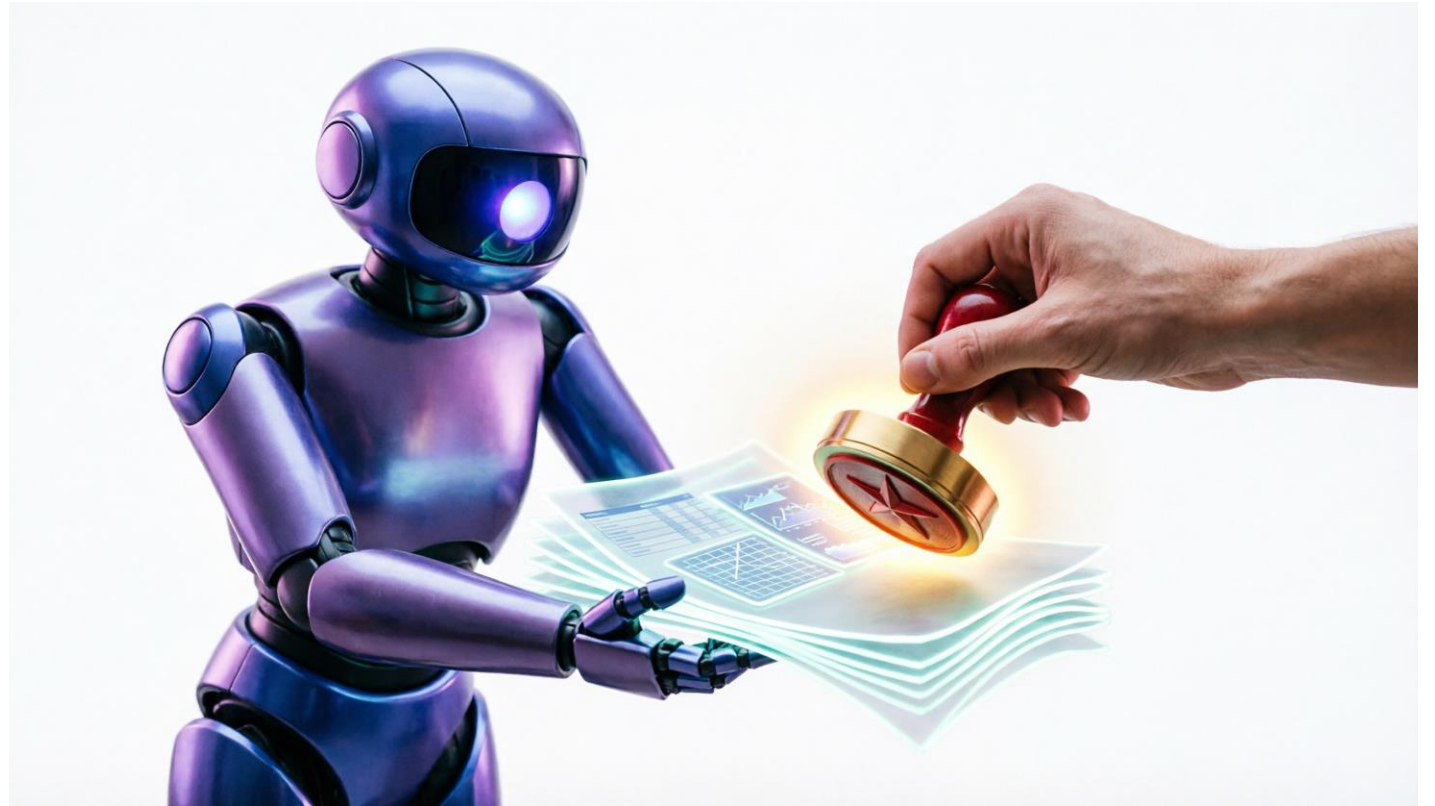
What AI Can Do?

- Read hundreds of documents at scale
- Recognize semantic equivalence
- Draft rule proposals



What AI Cannot Do?

- Make decisions on behalf of humans



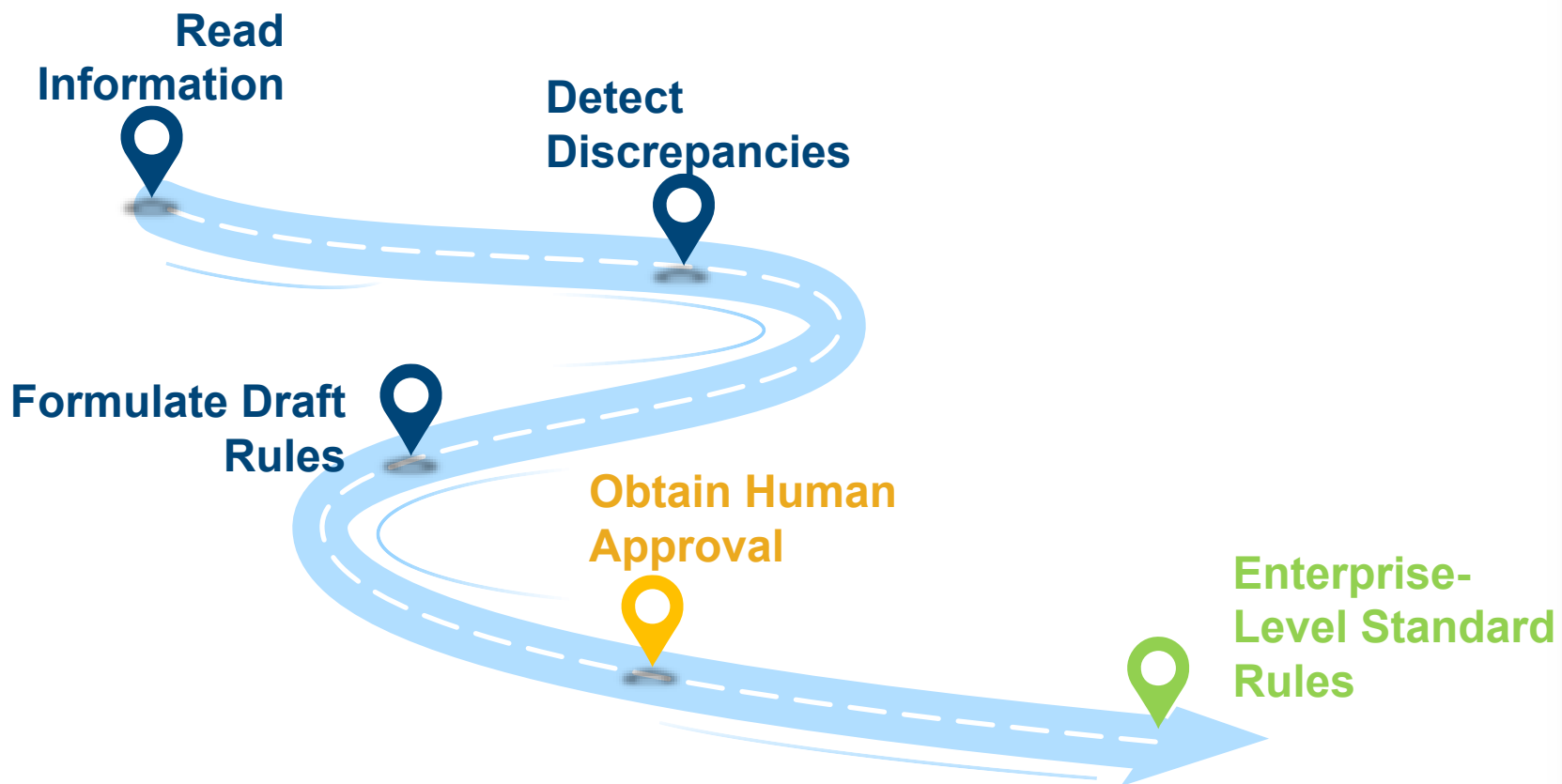
What Does This Mean?

You have an employee who can work around the clock and will never leave the position.

- which means that once you are able to establish a process through which it can provide you with findings, you will have a sustainable pipeline for issue discovery that does not rely on personnel.
- which means that it can read the data and documents of all projects, and summarize and analyze the differences among them.
- However, a human is still ultimately needed to make the final decision.



How AI Changes the Game





A Little Try



How To Read Data?

- One of possible solutions is to use cluster analysis, e.g. K-Means.
- We can make the following preparations to extract data features before cluster analysis.
 - Numeric data may use vector [mean, std, p25, p75, median]
 - Character data may use vector [the proportion of occurrences of each value relative to the total number of records]

How To Read Document?

- One possible solution is to use an LLM to decompose the derivation rules, vectorize the derivation conditions and outcomes separately, and then compute the distances between different projects.
 - We may have an LLM parse the textual rules in the specifications from each project and return data in JSON format that includes the derivation conditions and outcomes.
 - We may use an embedding model to convert the derivation conditions and outcomes into vectors and compute their similarity.

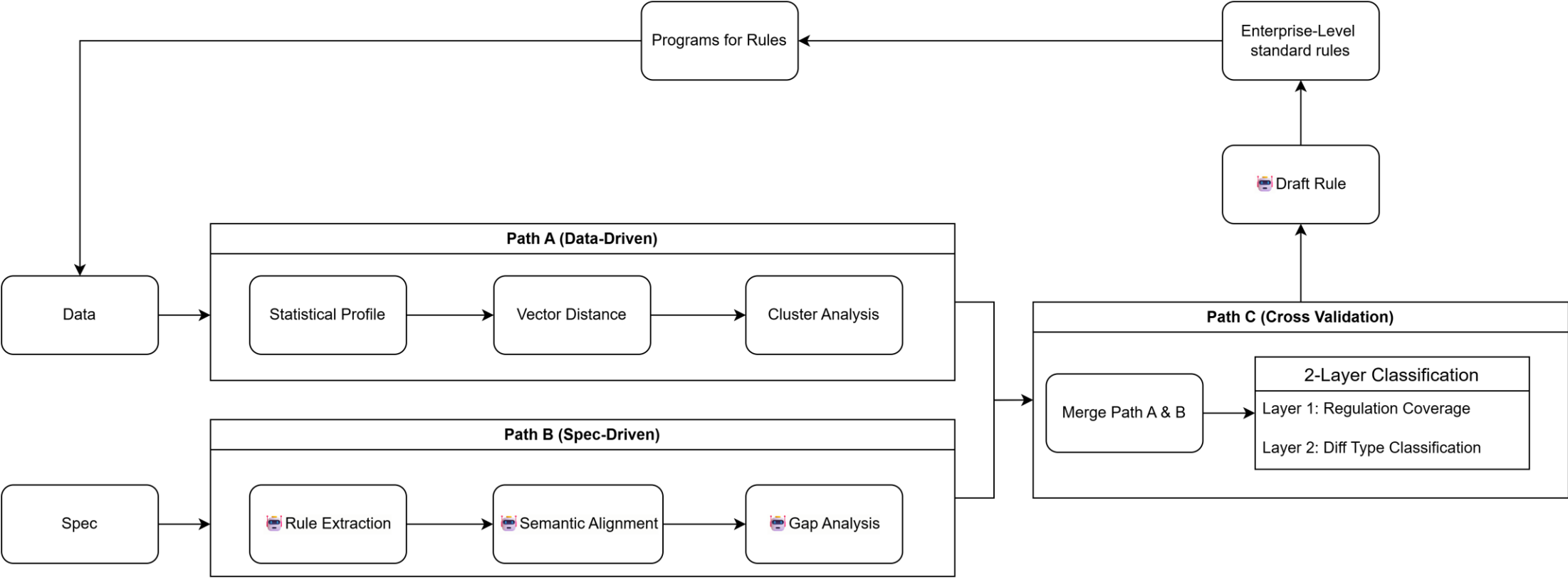
How To Reconcile Data and Document?

	Difference in Data			
Difference in Document		No	Yes, same group	Yes, diff group
	No	/	Hidden Divergence	
	Yes, same group	Spec Drift	Cross-Validated	Full Divergence
	Yes, diff group		Full Divergence	Full Divergence



- Same Group: The number of groups in the data equals the number of groups in the document.
- Diff Group: The number of groups in the data does not equal the number of groups in the document.

Workflow



Demo Video

Company Rule Discovery

127.0.0.1:5000

Summarize

Company Rule Discovery

Company Rules: 1

STUDIES

Select All (8)

CARDIO (3)

Phase 1 (1)
S08

Phase 2 (1)
S04

Phase 3 (1)
S02

NEURO (2)

Phase 2 (1)
S07

Phase 3 (1)
S05

ONCOLOGY (3)

Phase 2 (2)
S01
S03

Phase 3 (1)
S06

1 Data-Driven Ready

2 Spec-Driven Ready

3 Cross Validation Ready

DISCOVERED RULES

RULE : DOMAIN : VARIABLE : IS THERE A RULE? : WHAT'S THE EVIDENCE? : LEVEL :

Filter... Filter... Filter... All All All

Select studies and run analysis to discover cross-study rule variations

HOW TO READ THIS TABLE

Run Analysis



What's Next





What's the Value

- Issues can be systematically identified across all projects, making the process controllable and independent of individual expertise.
- Issues can be identified in advance, rather than only at the time of the pooled analysis. This means that the time required for the pooled analysis can be shortened.
- A faster pooled analysis means that the decision-making timeline can be moved forward, the time to market can also be advanced, and ultimately patients can benefit sooner.





What's the Future

- More source document (e.g. SAS/R code)
- Cross-variable and cross-domain consistency checks
- Stronger data discrepancy exploration
- More precise semantic comparison





Thank You!





Questions?

