CluePoints

How Machine Learning Can Automate Data Anomaly Detection?

Streamlining Query Management to Accelerate Database Lock and Regulatory Submission

John Hall, Senior Vice President

4th March 2025

CluePoints

Agenda

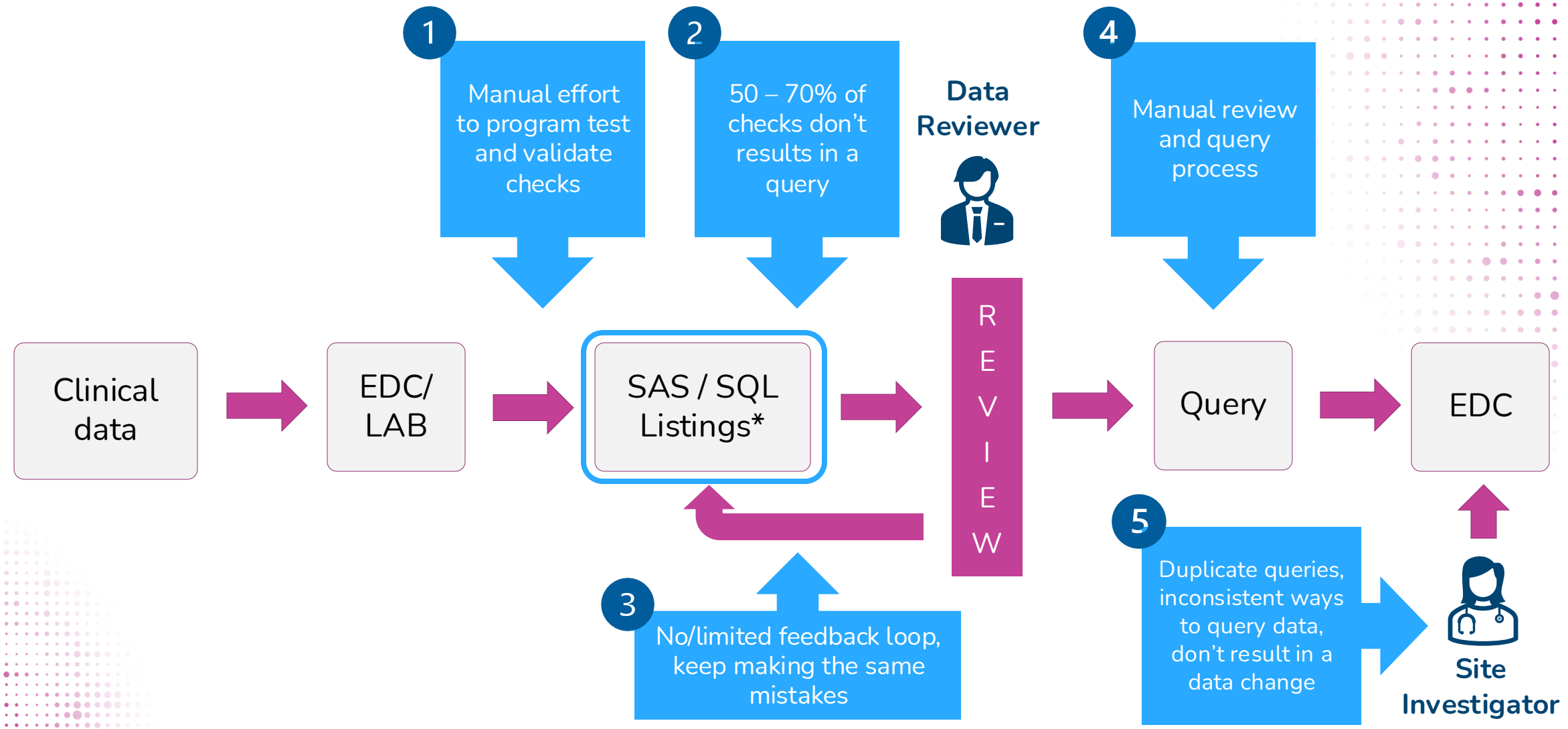
1. Data management review today – challenges and opportunities.
2. AI/ML to the rescue – or is it? Ensuring performance, scalability and regulatory compliance.
3. What could the future be like?

CluePoints

Data Cleaning Today

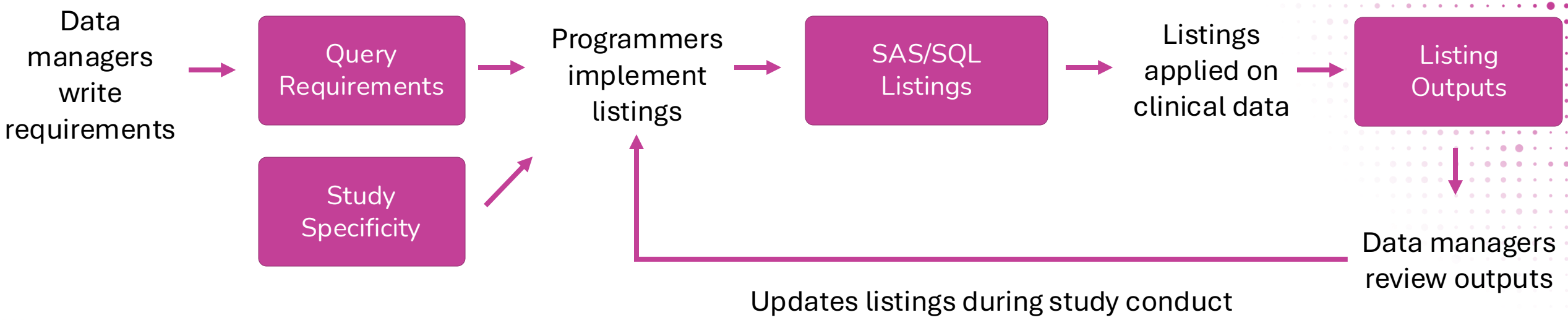
CluePoints

Data Queries: Current Challenges



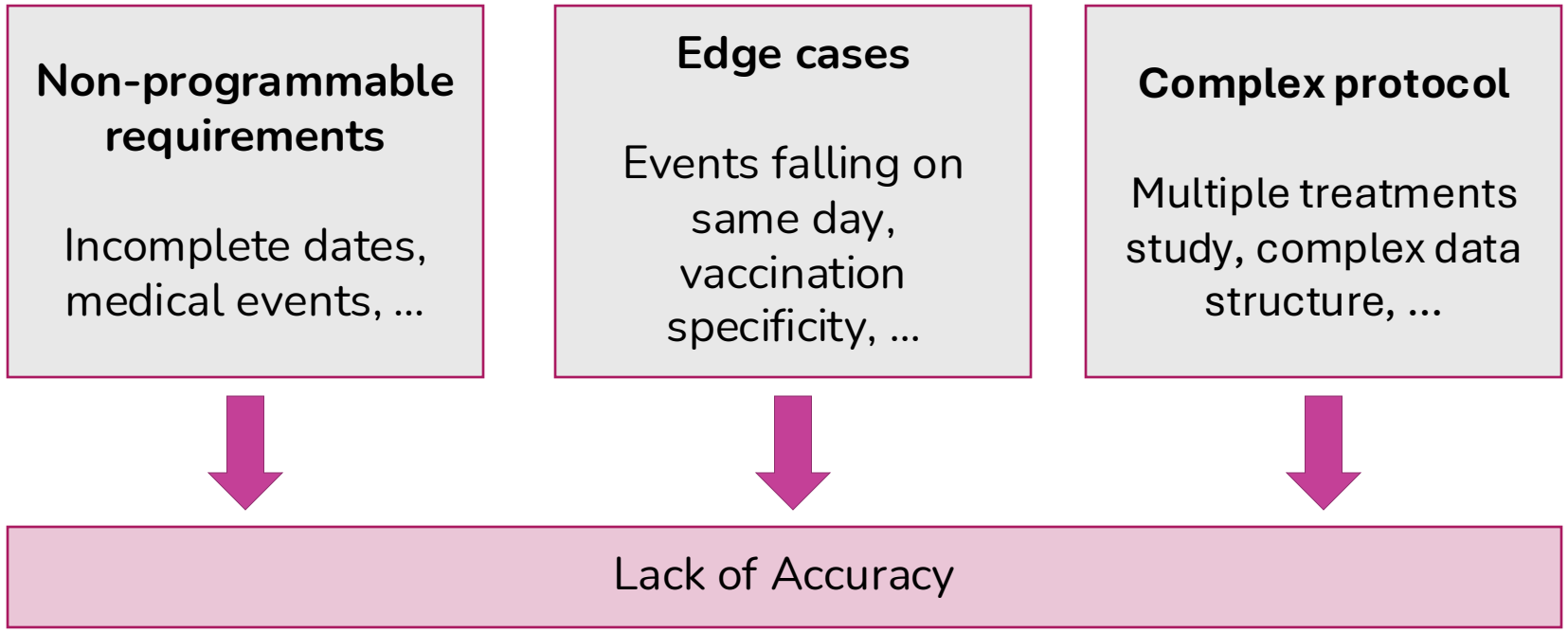
CluePoints

SAS/SQL Listings: Current Process



CluePoints

SAS/SQL Listings: Current Challenges



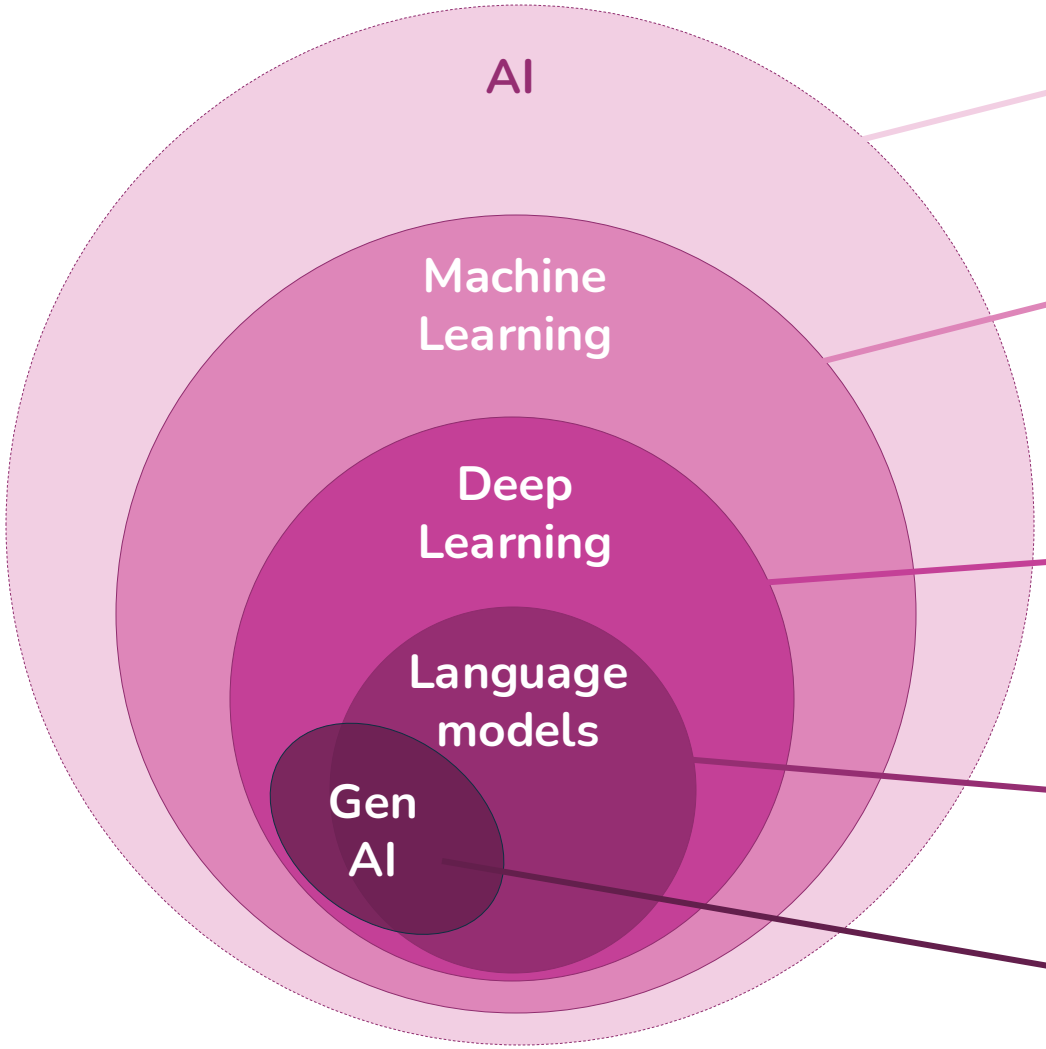
CluePoints

AI/ML to the rescue – or is it?

Ensuring performance, scalability and regulatory compliance.

CluePoints

The terminology is changing....

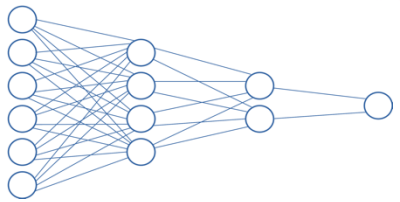


AI: a set of techniques? A solution? A goal?

Machine learning models learn patterns from data

Data →
Answers →
Machine Learning → Rules

Deep learning models are deep neural networks able to process complex data (images, text, etc.)



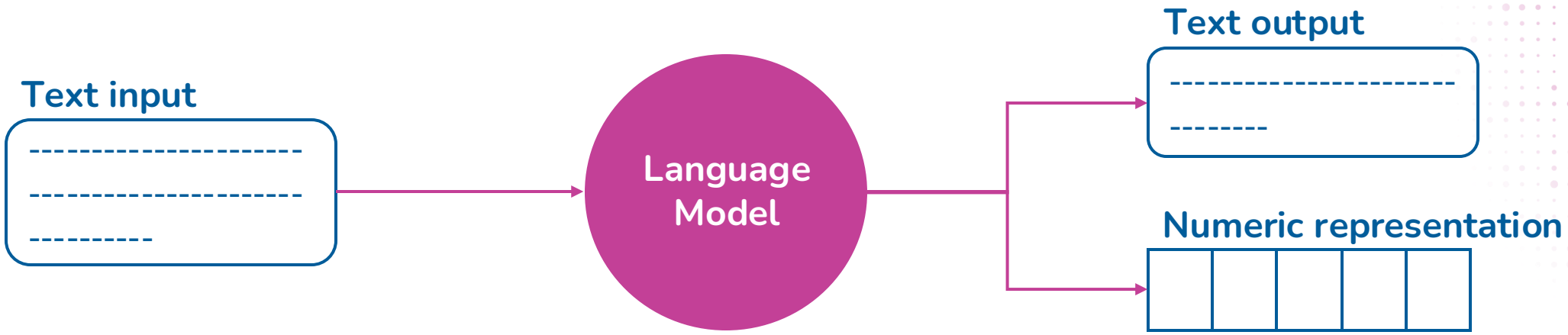
Language models learn probabilistic relationships in text

p("The dog barks") > p("The cat barks")

Gen AI refers to deep learning models (language models or others) able to generate data (text, images, etc.)

CluePoints

Language models have disrupted the field



Why using language models for clinical data review?

Reasoning

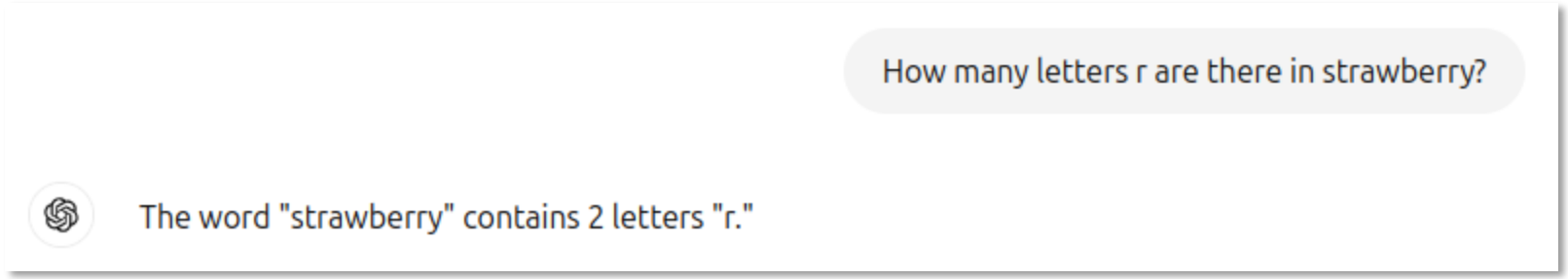
Medical knowledge

Flexibility in input data

CluePoints

But be careful….

Language models make mistakes, sometimes unexpected ones



GPT-4o (Sep 2024)

Needs process to avoid this behavior

CluePoints

Requirements for language models

Specialized in clinical data	Understand the task	Actionable answers	Reliable and consistent
• Understand the concepts related to clinical data • Understand clinical data structure	• Know what action is expected • Know how to respond	• Possible to parse the answer • Take decision from the answer	• Consistent in the process • High accuracy and no unexpected mistakes

CluePoints

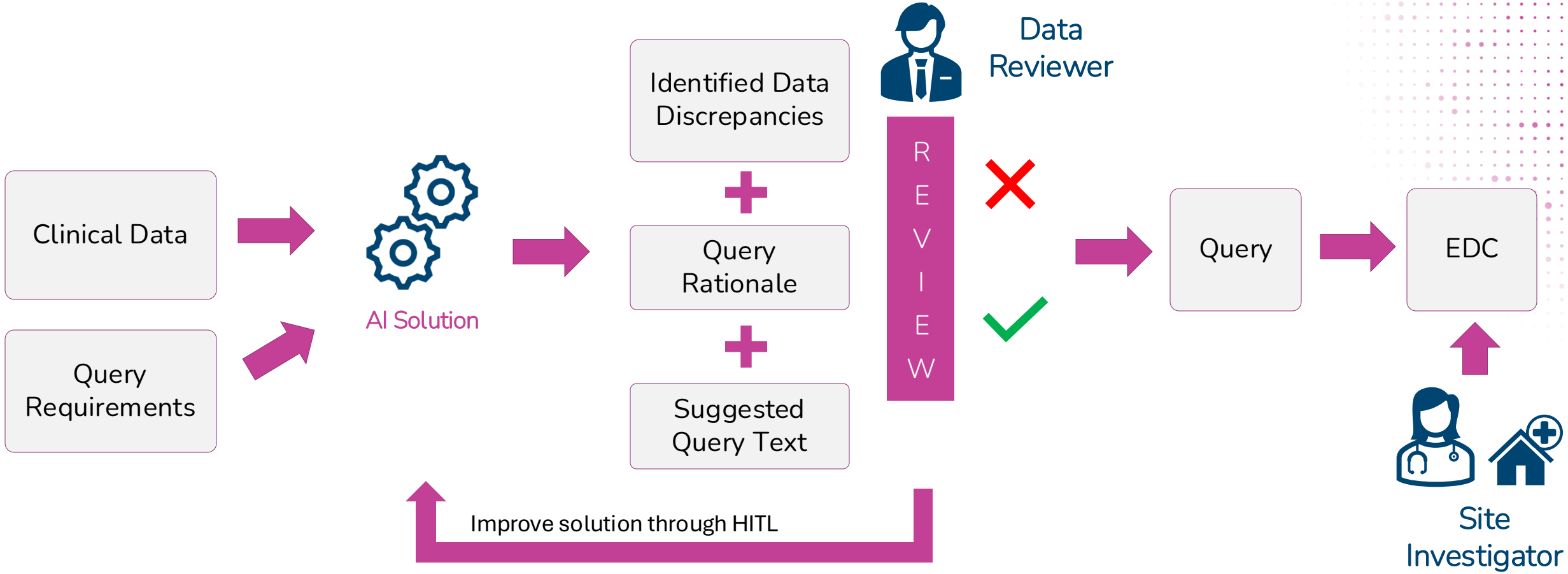
How to achieve these requirements?

Fine-tuning open-source language models!

Open-source model	Clinical data for training	Examples to learn from	SME support
• Existing powerful open-source models • Full control on these models	• Access to diverse clinical studies	• Examples to understand the task • Consistency in the ground truth	• Feedback from solution's output

CluePoints

AI Driven Detection of Data Queries: How?



CluePoints

Example Query Scenario

Scenario

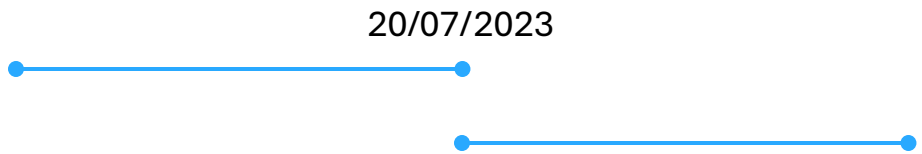
Identify Duplicate/Overlapping Concomitant Medication/Therapy records

Study specific exceptions (wouldn't raise a query)

1. Except for such overlapping Concomitant Medications where the end date of 1st CM and start date of the overlapping CM are the same day and where the frequency has changed.
2. Except where the CMs have different start dates but missing end dates, but that the type of CM is a vaccination.
3. ....

E.g. 1

Con Med Treatment	Dose	Frequency
Con Med 1	7.5mg	BID
Con Med 2	7.5 mg	TID



E.g. 2

Con Med Treatment	CM Category
Con Med 1	Vaccination
Con Med 2	Vaccination

CluePoints

Example Query Scenario

Scenario

Identify Duplicate/Overlapping Concomitant Medication/Therapy records

Study specific conditions

1. Except for such overlapping Concomitant Medications where the end date of 1st CM and start date of the overlapping CM are the same day and where the frequency has changed.
2. Except where the CM is a vaccination and the end date are missing, as long as the start dates are different.
3. ....

Model Benefit

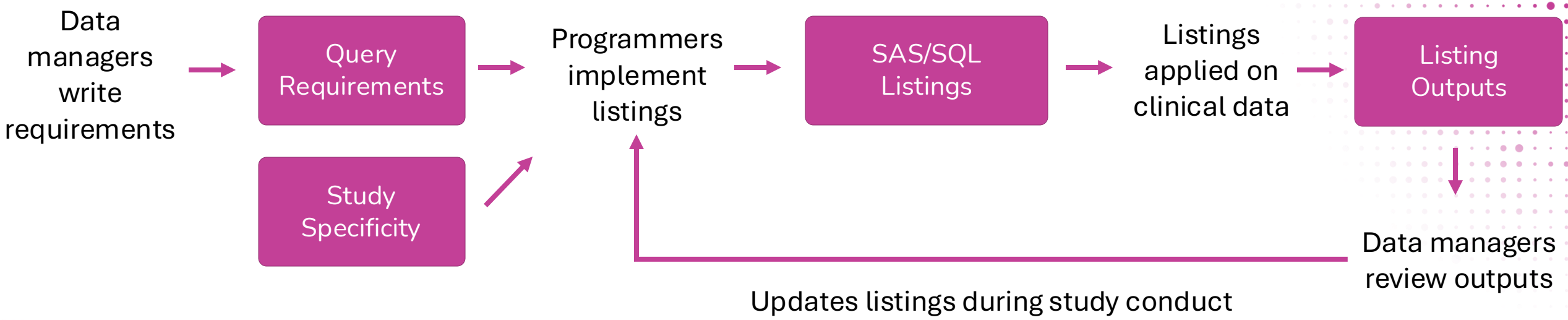
- Reduced the burden to program the scenario.
- Model is able to understand all the relevant conditions (indication, route, dosage, frequency, missing dates)
- Significantly reduce false positive rate (model achieves 95-100% true positive rate)
- Improved speed to review query scenarios

CluePoints


I'm being sued for a missed diagnosis! What do I do now?!
Sorry, buddy, comes with the job.
When handing radiology over to artificial intelligence sounds appealing.

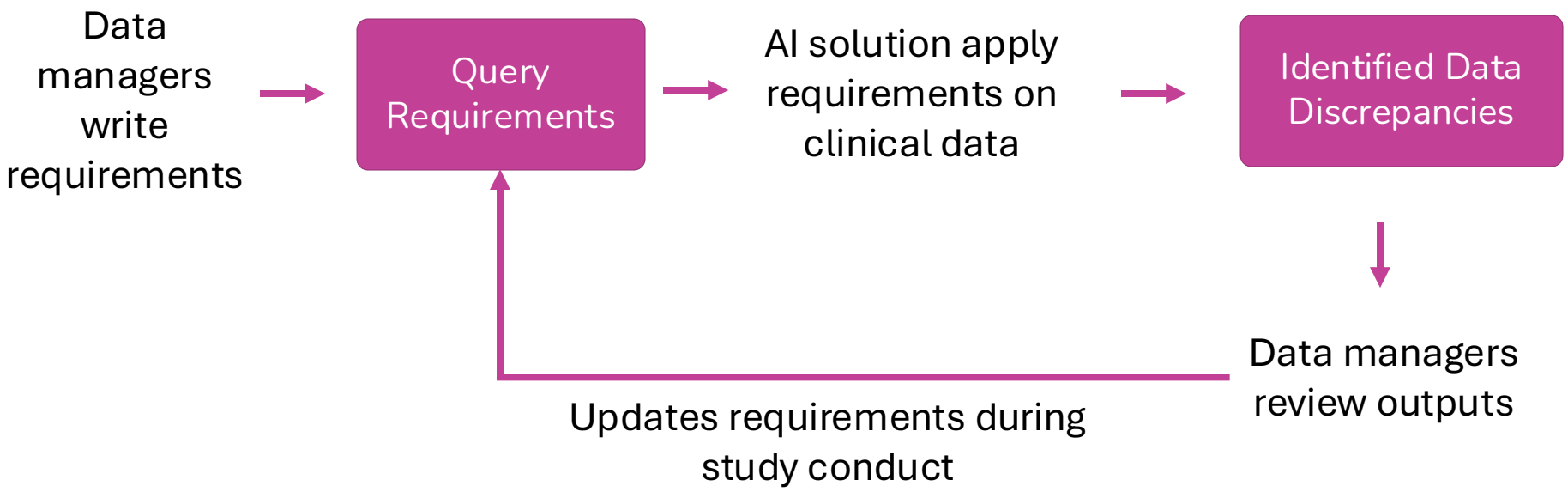
CluePoints

SAS/SQL Listings: Current Process



CluePoints

AI Driven Detection of Data Queries: How?



- Query requirements are part of the implementation and are an input to the AI solution
- AI solution automatically adapts to study specificity (data structure, study protocol, …)
- Link between requirements and quality of queries created
- Improve consistency in how requirements are applied and queries are raised

CluePoints

AI Driven Detection of Data Queries: Improve regulatory compliance

Traceability of Query Requirements

Keep track of what exact logic has been applied to clean data

Keep track on how requirements led to database change

Traceability of Model Decisions and Reasoning

Plain English text description of the models reasoning for each query scenario that can be accepted or rejected by a DM.

Used to for re-training.

Consistency on Raising Queries

Thinking is done upfront instead of real time

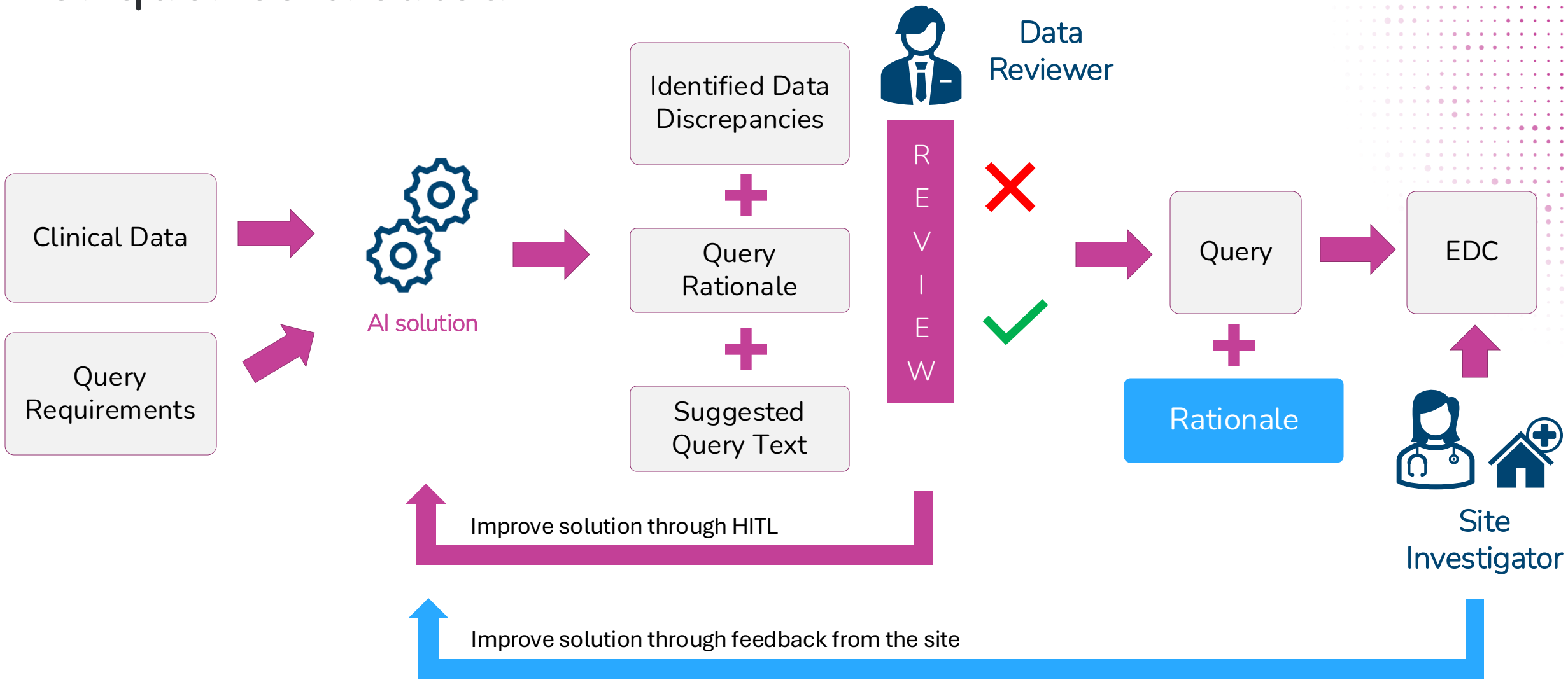
Logic to raise queries is applied globally and comprehensively and not dependent on experience of DM

CluePoints

What could the future be like?

CluePoints

Collect feedback from the site: Improve impact of queries created



CluePoints

Vision for the future

- Data managers can define any query requirements “using plain English” without programming expertise
- Raise queries in a uniform way across datasouces and studies (EDC, ECOA, EPRO, Lab)
- Improves consistency of query text for site staff, improve % of queries that result in a data change, reducing site burden
- Solution suggests relevant query requirements tailored to a given study

CluePoints

Thank you

Any Questions?