cdisc

2022
US
INTERCHANGE
26-27 OCTOBER | AUSTIN

Automating Efficiency
with **Machine Learning**
for **Data Classification**

Presented by: Robert Musterer, VP Product Management
eClinical Solutions

Meet the Speaker

Robert Musterer

**Title:** VP Product Management

Organization: eClinical Solutions

Industry veteran whose career has been based on using technology to make processing clinical trial data easier and more efficient.

Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*

- *eClinical Solutions provides services and technology to the Life Sciences industry*
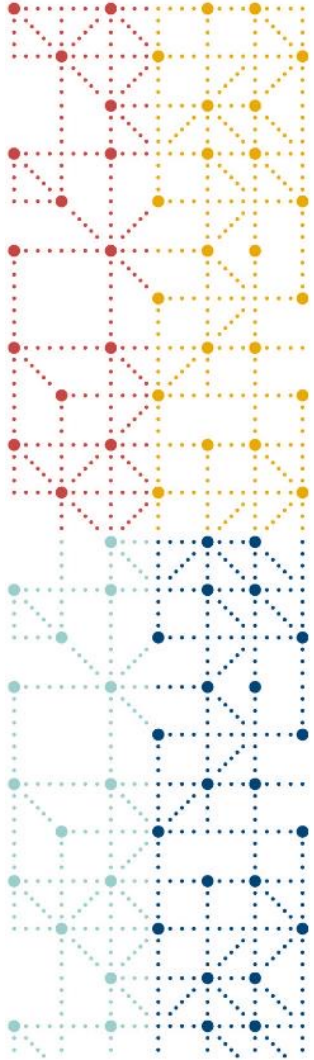
Automating Efficiency
with Machine Learning
for Data Classification

Agenda

1. Why Automate
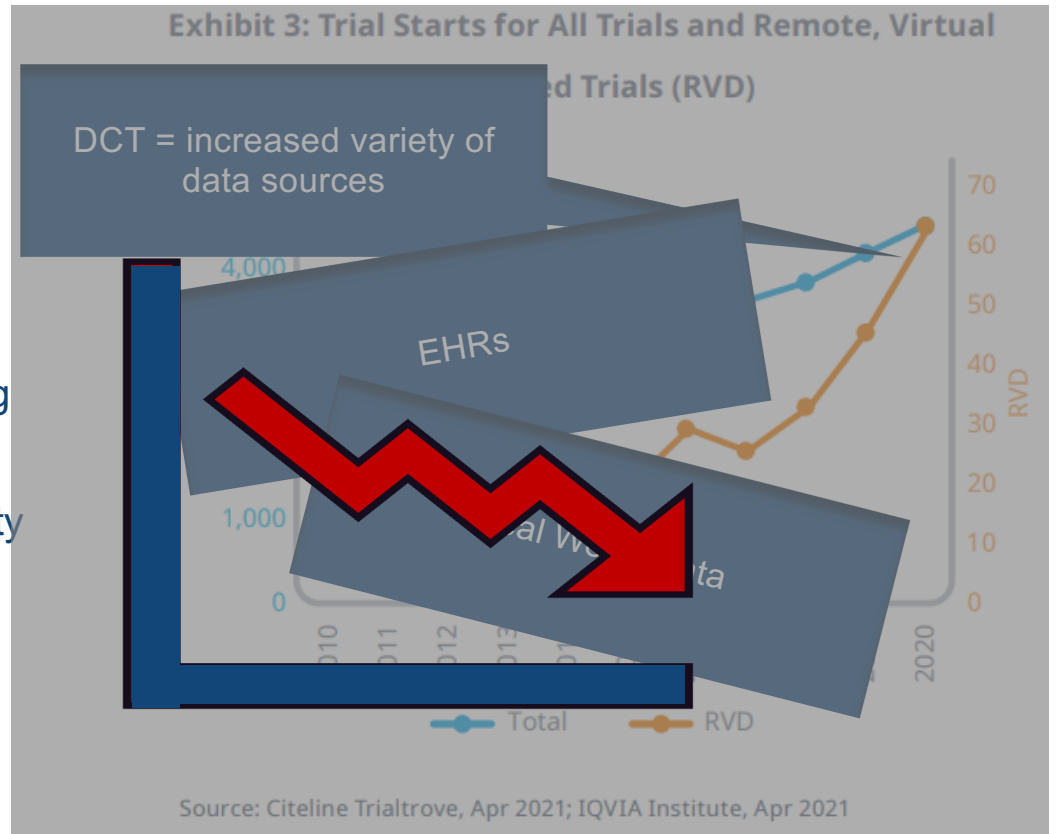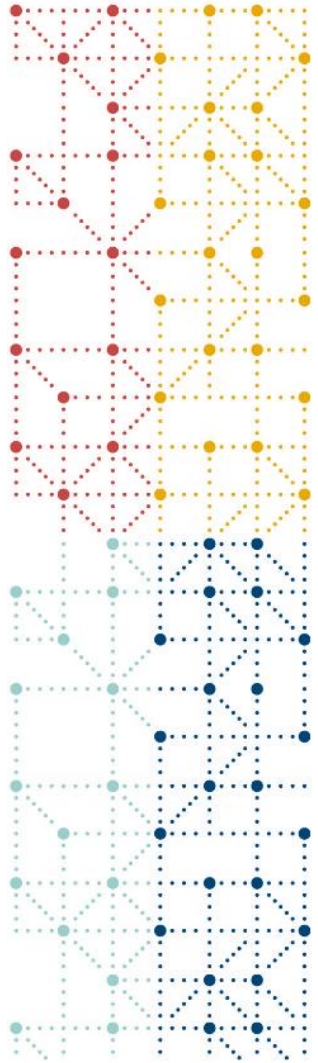2. using Machine Learning
3. for Data Classification
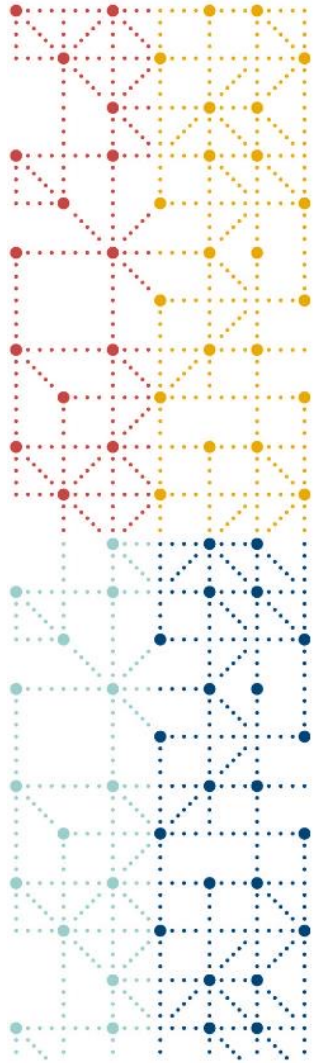
Why Automate

Automating Efficiency

WHY?

- Trial complexity increasing
- Trial size increasing
- Number of trials increasing
- Number of data sources increasing

- Downward pressures on profitability

- Economic Conditions

Exhibit 3: Trial Starts for All Trials and Remote, Virtual ...ed Trials (RVD)

DCT = increased variety of data sources

EHRs

Source: Citeline Trialtrove, Apr 2021; IQVIA Institute, Apr 2021
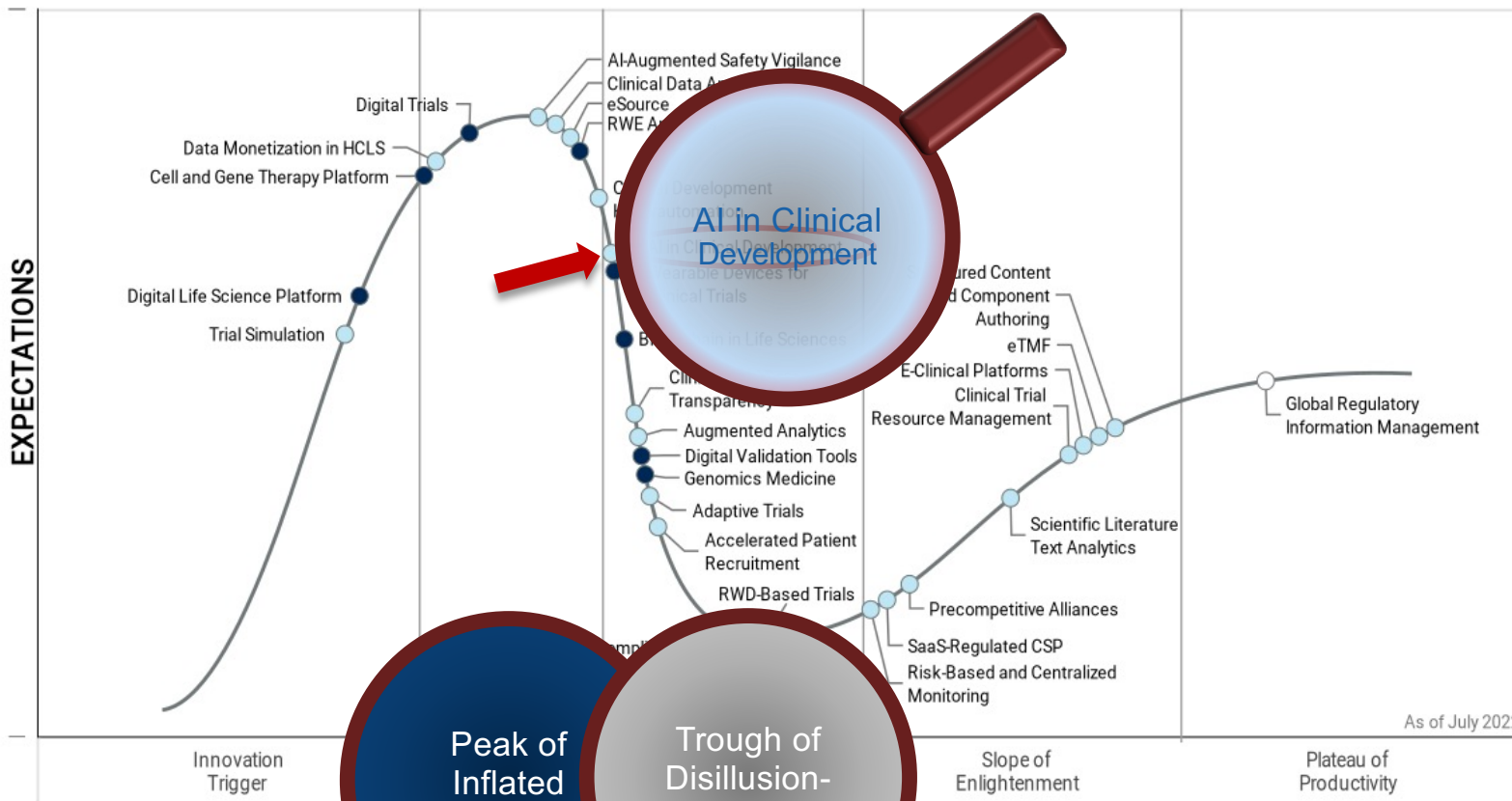
Why with Machine Learning

with **Machine Learning**

WHY?

- AI/ML techniques hold promise for significant efficiency gains

- Traditional rules based / hard coded logic is fragile and requires regular maintenance

- As each trial design is a bit different, hard coded logic needs to modified per study

- Can make very complex problems manageable

# Hype Cycle for Life Science Clinical Development, 2022

Unrealistic Expectations:

- ## Just works **immediately** out of the box
  - o Dismisses need for training data
  - o Dismisses need for feature set definition
  - o Dismisses need for ground truth data
  - o Dismisses need to train models

- ## Assume all studies are equal
  - o Ignores variance by therapeutic area
  - o Ignores variability by drug class
  - o Ignores variability by treatment arm (exposure duration and dose)

- ## Assumption of "Auto-Magic"
  - o Ignores need for Human in the Loop
  - o Ignores need to define use cases

Hype Cycle for Life Science Clinical Development, 2022

cdisc
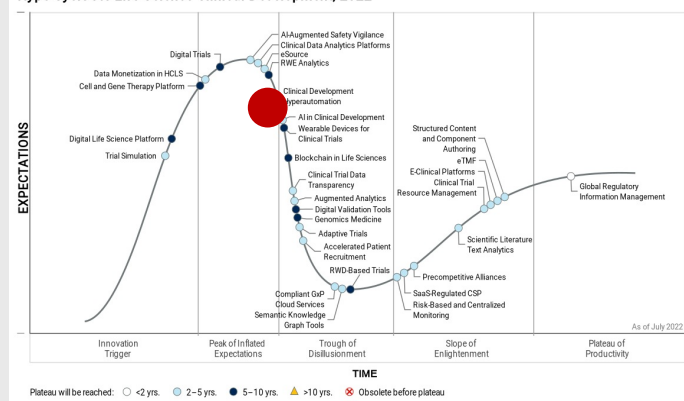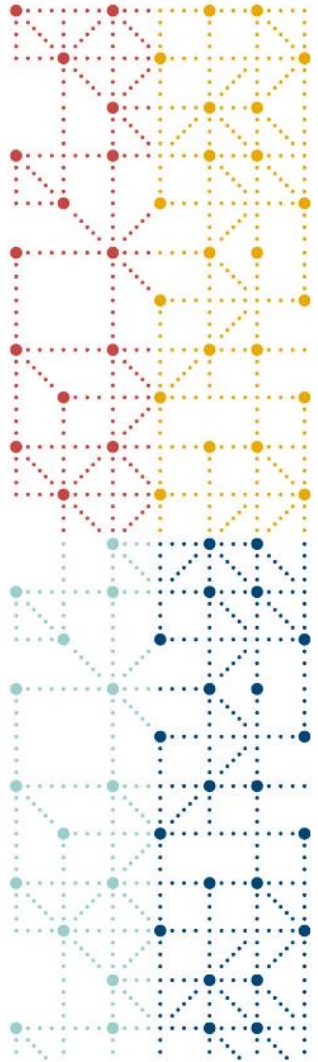
for Data Classification

for **Data Classification**

WHY?

- Pragmatic
- Rapid time to value
- Hybrid approach:
  - Metadata driven solutions
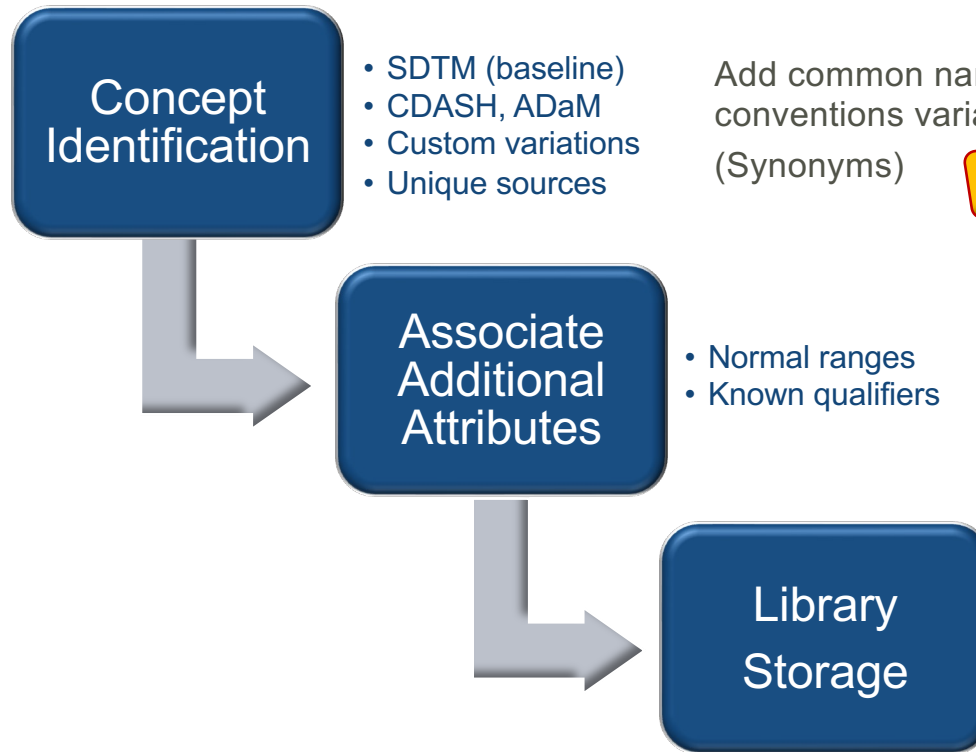 - Wide ranging use cases

cdisc

How do we go about it?

1. Preparation (laying the foundation)
2. Programmatic Comparisons
3. Review – Human in the Loop
4. Feedback (continuous improvement)

Preparation

("Expert system" / "Ground Truthing" / Laying the Foundation)

Concept Identification

- SDTM (baseline)
- CDASH, ADaM
- Custom variations
- Unique sources

Add common naming conventions variations

(Synonyms)

AEEXP = AETERM

LBTEST=LABTEST=Analyte

Associate Additional Attributes

- Normal ranges
- Known qualifiers

Library Storage

cdisc

Programmatic Comparisons

- As new sources come in:
  - ➢ **Compare:**
    - Naming conventions
      - Include common variations
    - Attributes (type, length, controlled terminology)
      - Some source provide associate metadata, for others you'll need to programmatically determine them
    - Determine if item values are stand-alone or if a qualifier constrains the range of values (most common is units).
    - Compare actual data values against commonly encountered ranges (can start with normal ranges)

Naming Conventions	Attributes	Qualified?	Actual Data Values

cdisc

ML Algorithm Determinations

For the items that fall through the pragmatic comparison

- AI/ML techniques are never 100% accurate.

- So, by handling the known knowns up front, we:
  - minimize processing demands
    - good for performance from both perspectives
      - system response time and
      - reliability of solution results
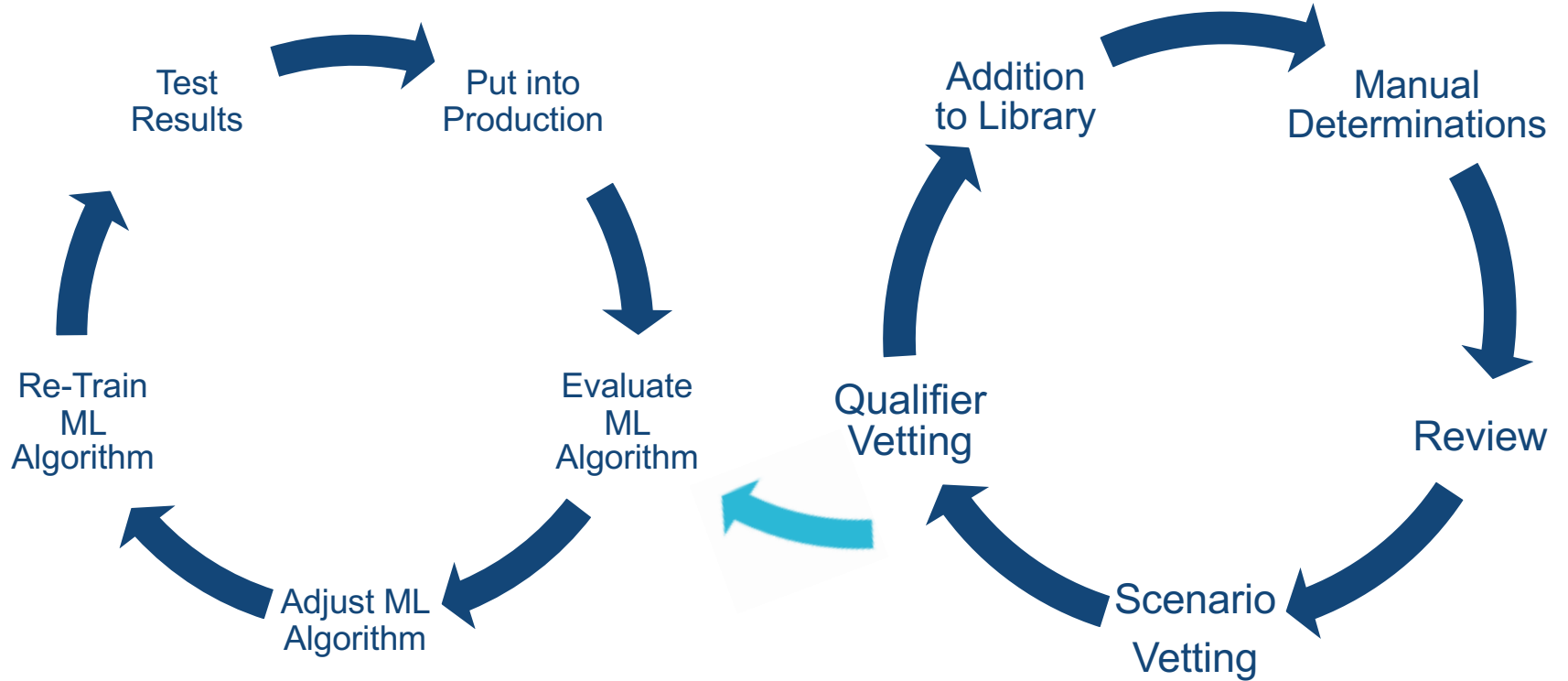 - avoid frustrating users

Human in the Loop

- Present determinations to a person for:

  - Review,
  - Acceptance,
  - Modification,
  - Manual Determination

Feedback

Test Results → Put into Production → Evaluate ML Algorithm → Adjust ML Algorithm → Re-Train ML Algorithm → Test Results

Qualifier Vetting → Addition to Library → Manual Determinations → Review → Scenario Vetting → Qualifier Vetting
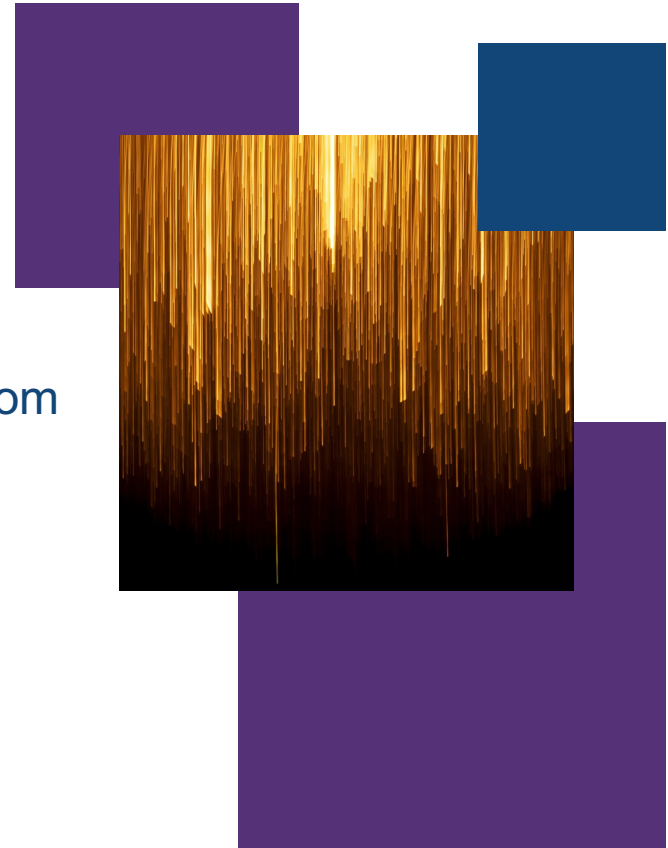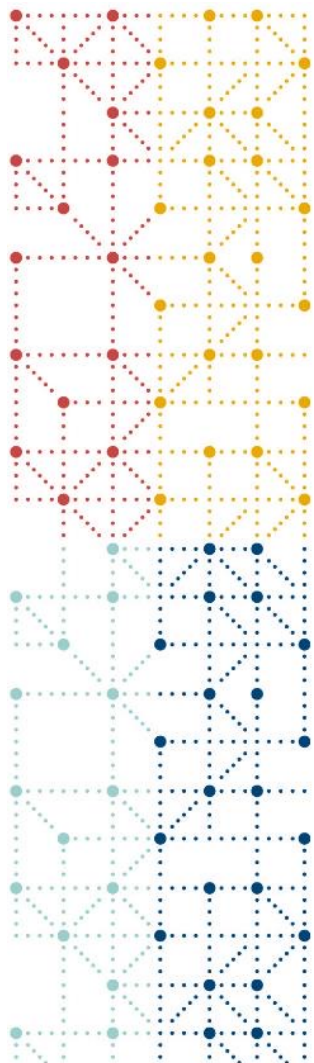
cdisc

What are the benefits?

Improved Efficiency

Foundational for the next step which is the tremendous benefits that can be obtained from using metadata driven automation!

Thank You!

cdisc