

PharmaSUG Osaka 2023

Automatic Generation of Python Programs for Creating SDTM Datasets

FUJITSU LIMITED



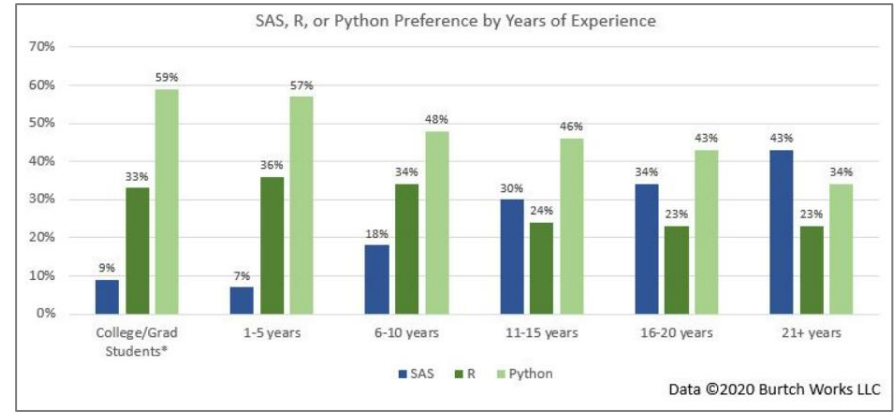
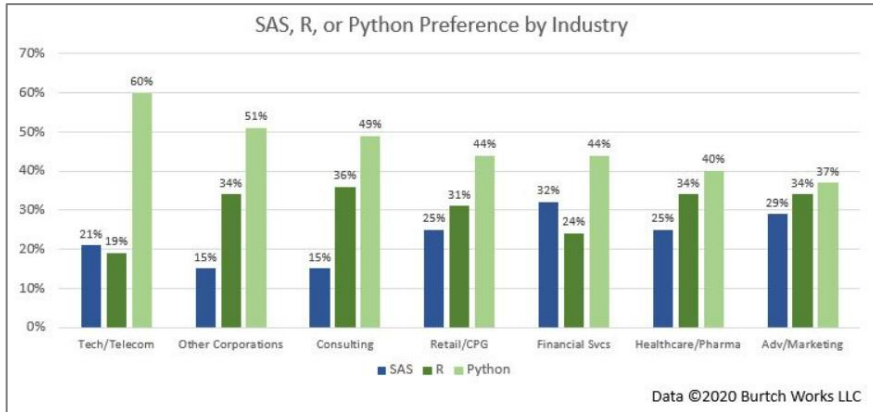
- The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of the organization where the author belongs.

- The main topic of this presentation is automatic generation of **Python** programs for creating CDISC **SDTM** datasets.
- Before diving into the main topic, this presentation will explain:
 - Why **Python** is a good option for data science in drug development
 - How programming practice is changing with **Generative AI**
 - How **Digital Data Flow** is impacting approach to data processing

Background and Trends

Popularity of SAS, R and Python

- A study conducted in 2020 for 1,000 data scientists in the United States showed that **Python** was the most popular language among SAS, R and Python, in any industry.
- It also shows that Python is more popular in younger generation (with less years of experience in data science).





Reference: Burtch Works LLC "[2020 SAS, R, or Python Survey Results: Which Tool do Data Scientists & Analytics Pros Prefer?](#)"

- Generative AI (more precisely LLM: Large Language Model) is significantly changing the way how programs are developed.
 - Code Generation
 - Code Completion
 - Test Code Generation and more

# Ranking	Programming Language	Percentage (YoY Change)	YoY Trend
1	Python	17.715% (+0.516%)	
2	Java	11.086% (-0.172%)	
3	Go	10.817% (+0.958%)	^
4	JavaScript	10.065% (+0.692%)	^
5	C++	9.712% (-0.348%)	v
34	R	0.056%	^^

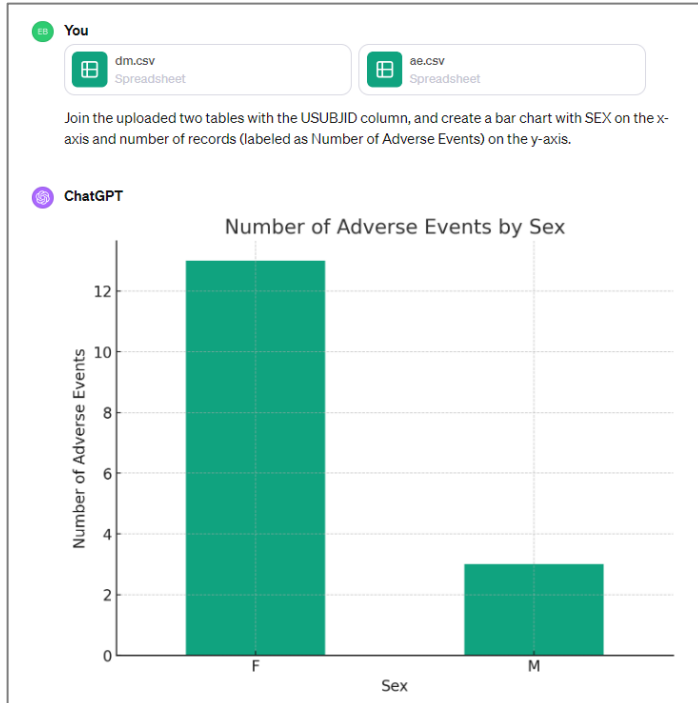
- ChatGPT, the most famous LLM-based generative AI, is built with Python and best at generating Python codes among other languages.

 Can ChatGPT create programs automatically? Please explain the language you are best at along with its reason.

 Yes, ChatGPT can automatically create programs. Python is the language that I'm most proficient in, and there are several reasons for this:

- Another famous LLM-based generative AI, Copilot, is trained on all languages that appear in public repositories (see the top 50 pull requests in Q3 of 2023 on the left).

- Advanced data analysis is another powerful feature of ChatGPT that allows uploading files, write and execute Python codes, and create graphs with instructions in natural language.



You

Please show the python codes you created to analyze files and depicted the bar chart.
< 2/2 >

ChatGPT

Here is the Python code I used for analyzing the uploaded files and depicting the bar chart:

```
python
Copy code

import pandas as pd
import matplotlib.pyplot as plt

# Load the datasets
dm_path = '/mnt/data/dm.csv'
ae_path = '/mnt/data/ae.csv'

dm_data = pd.read_csv(dm_path)
ae_data = pd.read_csv(ae_path)

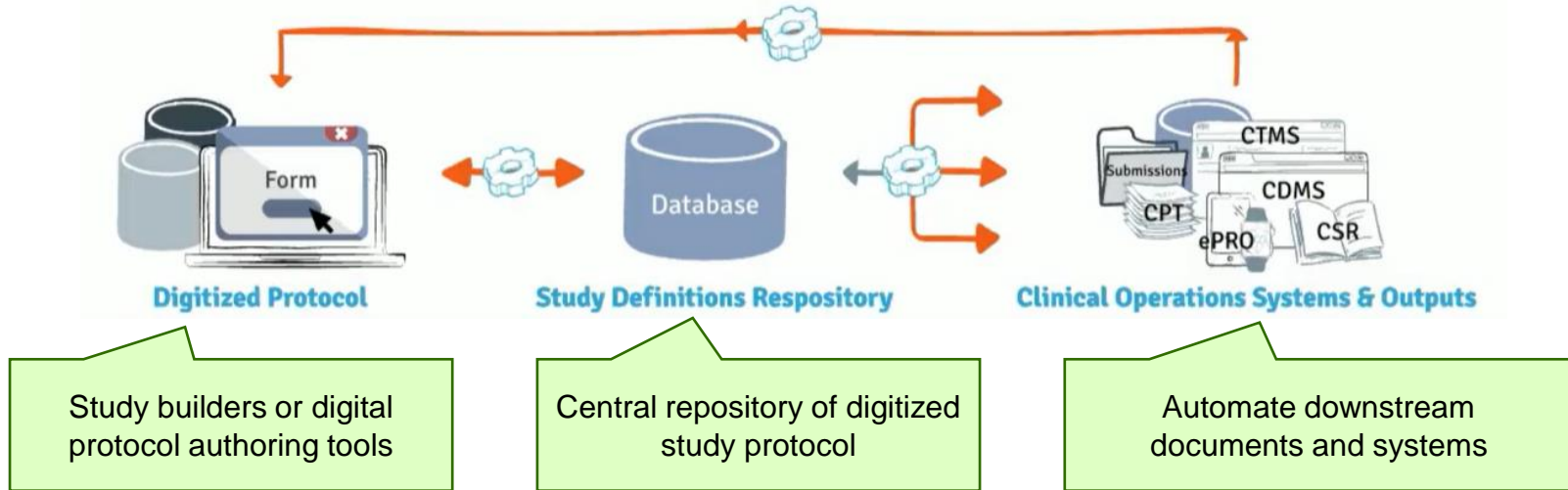
# Joining the two tables on 'USUBJID' column
joined_data = pd.merge(ae_data, dm_data[['USUBJID', 'SEX']], on='USUBJID')

# Counting the number of records (adverse events) for each sex
sex_counts = joined_data['SEX'].value_counts()

# Plotting the bar chart
plt.figure(figsize=(8, 6))
sex_counts.plot(kind='bar')
plt.title('Number of Adverse Events by Sex')
plt.xlabel('Sex')
plt.ylabel('Number of Adverse Events')
plt.xticks(rotation=0) # Keeping the sex labels horizontal
plt.show()
```

Digital Data Flow (DDF) Overview

- Digital Data Flow (DDF) is an initiative by TransCelerate that aims to create digitized study protocol and automate creation of study assets.



- CDISC is collaborating with TransCelerate to develop a standard model for Study Definitions Repository as part of their journey to achieve end-to-end automation.

Note: The content in this slide has been created by summarizing contents on the TransCelerate web site at the author's own discretion.

DDF Achieves More SDTM Automation

(1) Activities/BCs are populated via Study Builder or read from Study Protocol.

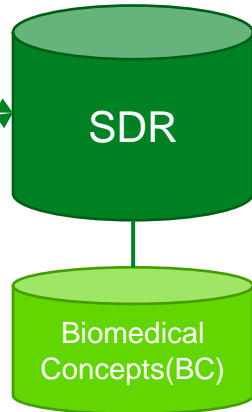
(2) Activities with BCs are mapped to forms when a draft eCRF Spec is created from SDR.

(3) Activities with BCs are mapped to SDTM domains when a draft SDTM Spec is created from SDR.

Study Builder

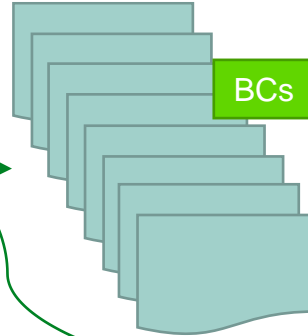


Protocol



Biomedical Concepts(BC)

eCRF Spec



BCs

Automated Mapping

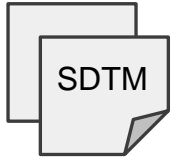
SDTM Spec



BCs

Generate

SDTM Datasets

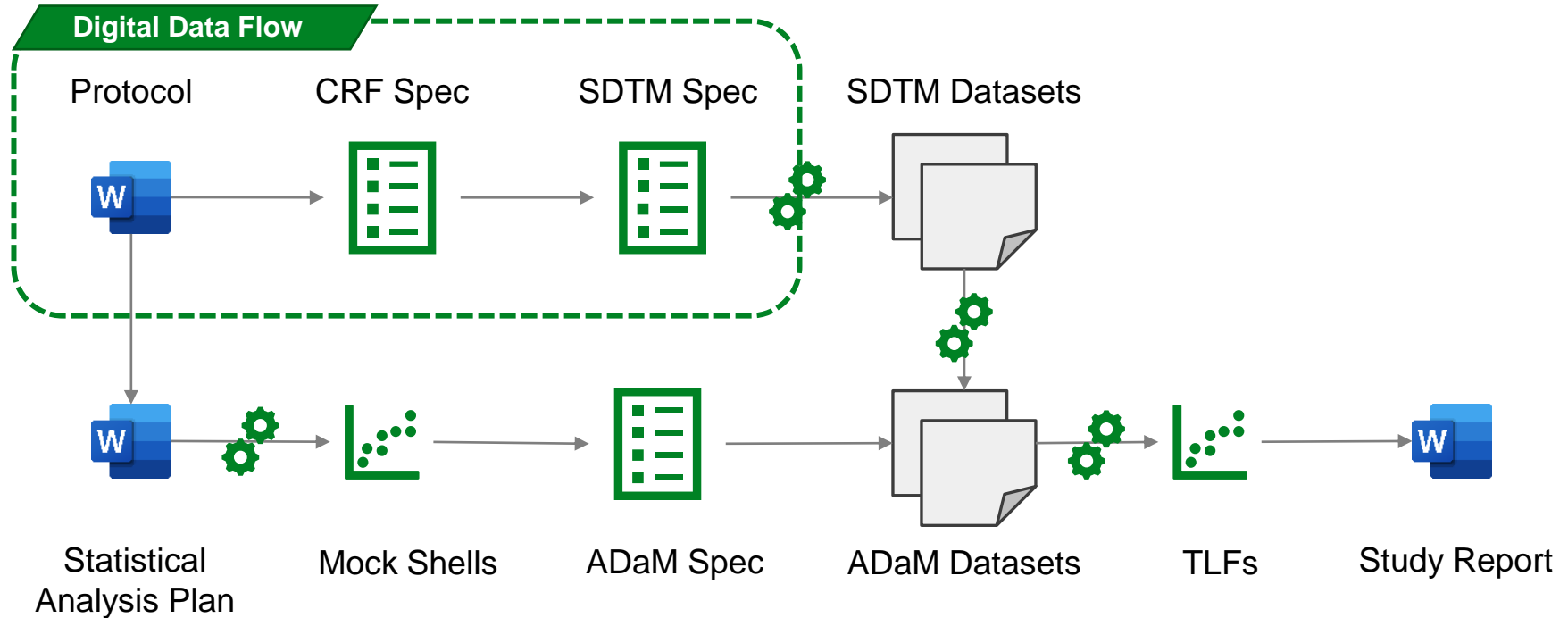


(4) SDR knows which BCs are referenced from which forms/domains

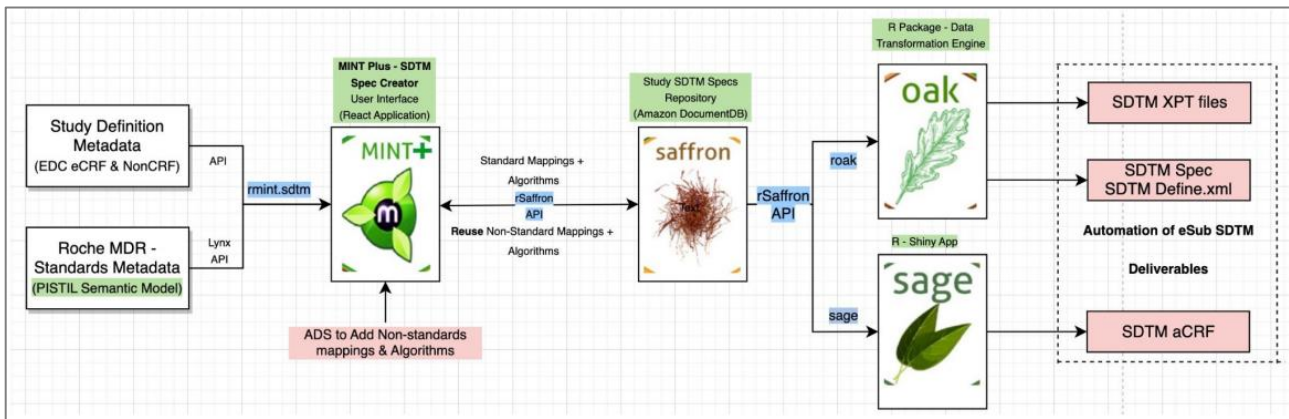
Digital Data Flow

Metadata Repository

- Do they conflict or complement to each other?



- A well-known example of R-based SDTM automation

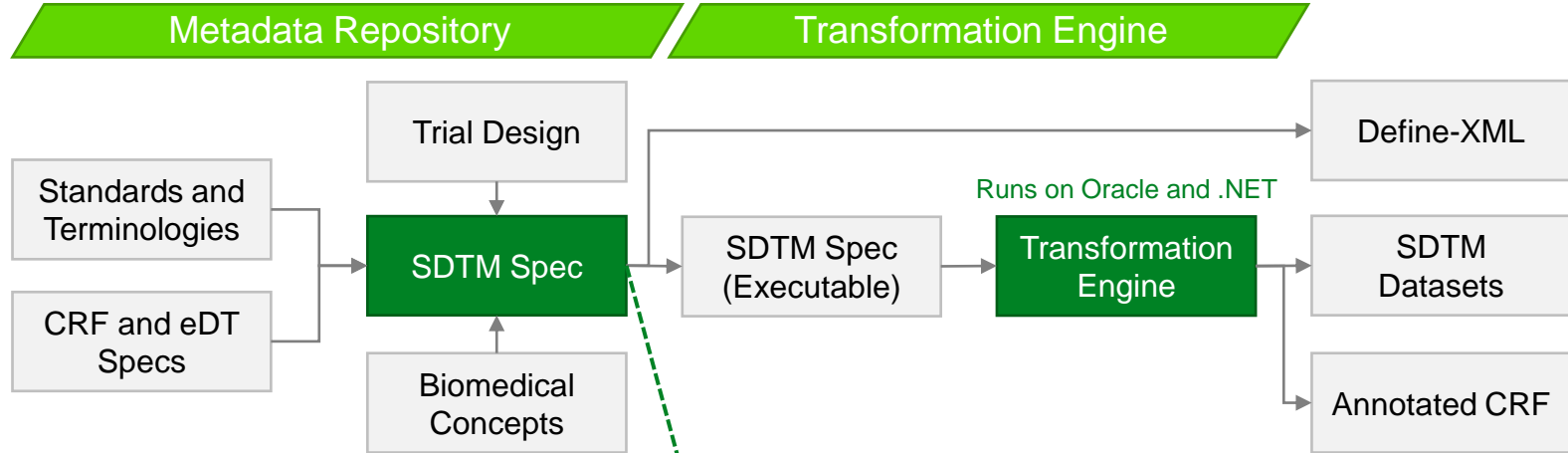


Up to 22 Reusable Algorithms

Only as Algorithms	Only as Sub-Algorithms	Algorithm & Sub-Algorithms
03_AE_AEREL	11_MERGE	01_ASSIGN_NO_CT
07_DATASET_LEVEL	18_REMOVE_DUP	02_ASSIGN_CT
09_IF_THEN_ELSE	19_GROUP_BY	05_HARDCODE_CT
17_WHODRUG_FA	20_NEED_USER_INPUT	06_HARDCODE_NO_CT
13_RELREC	08_NOTSUBMITTED	
14_RELREC_CONDITION	15_MULTIPLE_RESPONSES	
21_NONCRF_LAB		
22_NONCRF_PKC		
23_PAISED_VARS		

Reference: F. Hoffmann-La Roche AG "[OAK Garden - SDTM Automation The flourishing Data Transformation Engine](#)"

- Similar components and processes to OAK Garden, but more variety of methods.



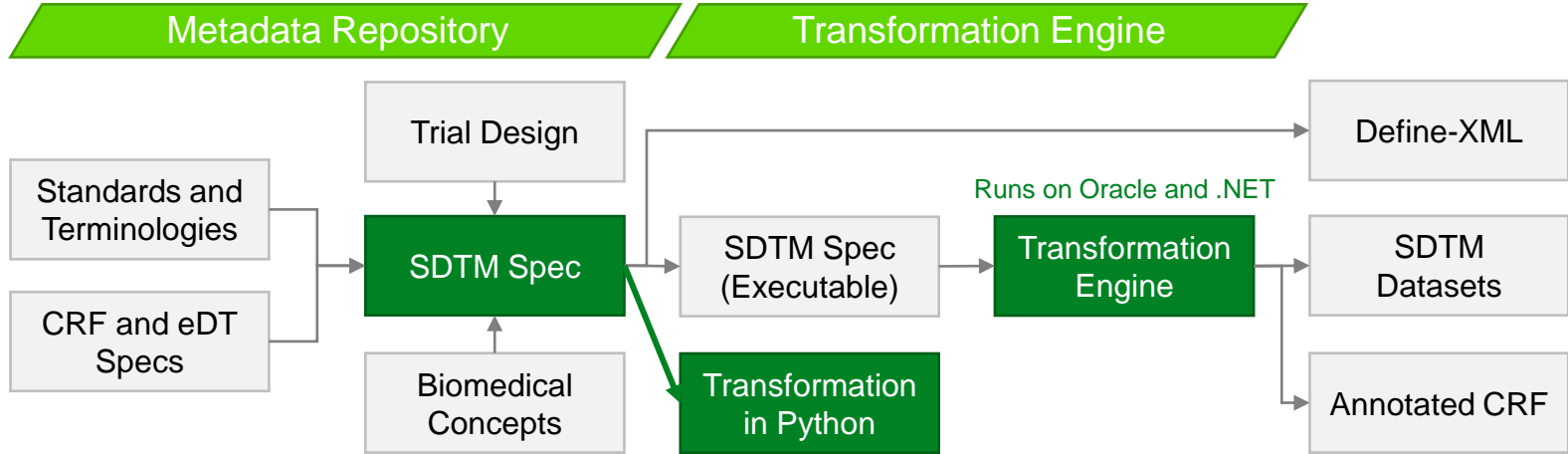
55 Methods and 286 Use Cases**

**: Pre-configured method calls

Topic Methods	Non-Topic Methods
assignTopic copy copyDesign copyUnique extrt joinUnique lookupConcept lookupConceptNormalized lookupDomainREGEX n1mapTopic se sv	bift calc coeval comment concat concatDate convertToStd copy copyDesign copyIf copyToSUPP ...and more

Adding Python Implementation to tsClinical

- The 55 methods are being implemented in Python.



Python currently implemented for real-time mapping review as SDTM spec is edited.

EDC.AE (Created from CRF Spec)

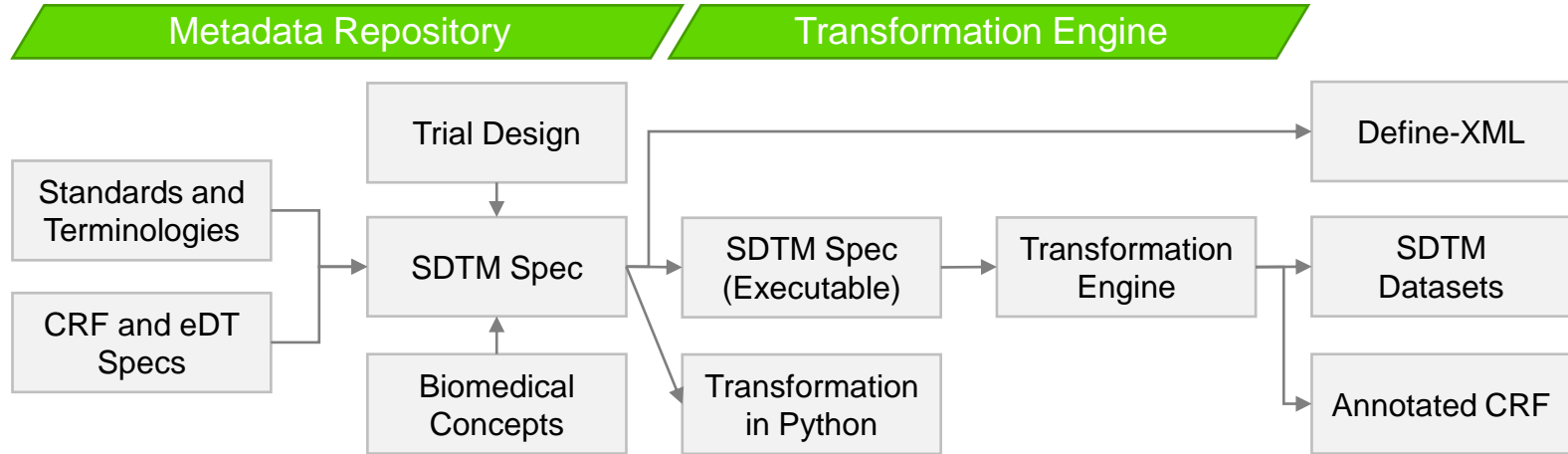
Subject	Folder	InstanceRepeatNumber	PageRepeatNumber	RecordPosition	AE.AESPID	AE.AETERM	AE.AESTDAT_RAW	AE.AEENDAT_RAW
02-02	V_FOLLOWUP_00000	1	0	0	AE003	疼痛	05 May 2020	24 Jan 2020
02-01	V_DISPOSITION_00000	0	0	2	AE001	高血压	22 Jan 2020	24 Apr 2020
03-01	V_DISCONTINUED_00000	3	0	1	AE002	心臟麻痺	22 Apr 2020	26 Mar 2020
01-03	V_AE_00000	2	0	3	AE003	胃痛	25 Mar 2020	30 May 2020
01-04	V_TREATMENT4W_00000	1	0	0	AE001	頭痛	18 Jan 2020	01 Apr 2020
01-01	V_TREATMENT1W_00000	0	0	2	AE002	腹痛	15 Apr 2020	20 Jan 2020

SDTM.AE (N/A indicates the review is not supported)

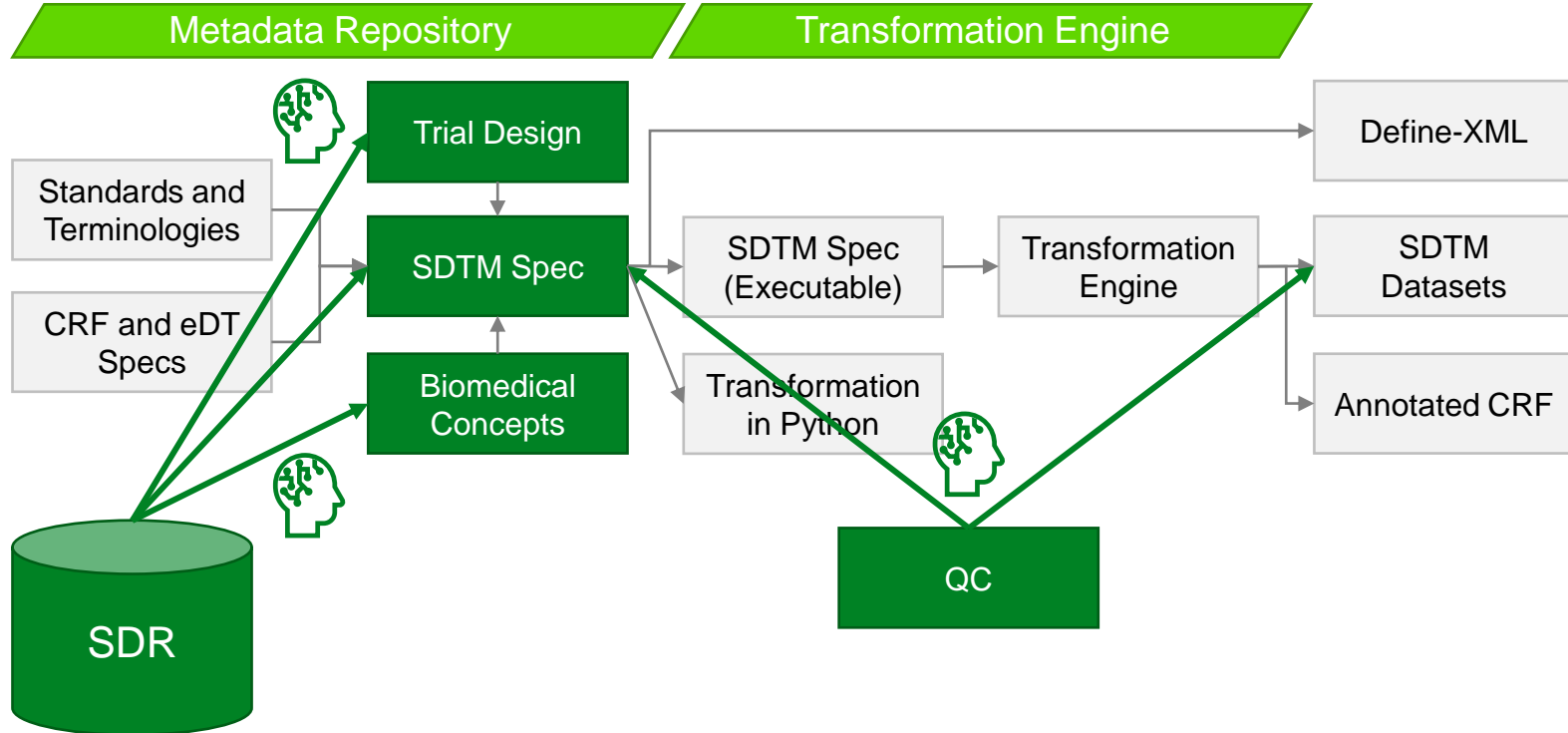
STUDYID	DOMAIN	USUBJID	AESQ	AESPID	AETERM	AELLT	AELLTCD	AEECOD	AEPTCD	AEHLT	AEHLTCD	AEHLGT
TRAINING25	AE	(N/A)	(N/A)	(N/A)	疼痛	(N/A)	(N/A)	(N/A)	(N/A)	(N/A)	(N/A)	(N/A)
TRAINING25	AE	(N/A)	(N/A)	(N/A)	高血压	(N/A)	(N/A)	(N/A)	(N/A)	(N/A)	(N/A)	(N/A)
TRAINING25	AE	(N/A)	(N/A)	(N/A)	心臟麻痺	(N/A)	(N/A)	(N/A)	(N/A)	(N/A)	(N/A)	(N/A)

- Generative AI is multipurpose, and thus your ideas matter to make it valuable.
- Python sits in a unique position in the AI era.
- Digital Data Flow is built on CDISC standards and has good familiarity with data processing.
- After all, **there are many chances of automation in your job with the rise of LLM-based generative AI and Digital Data Flow.**

- Which part of the process below can be further automated?



- Which part of the process below can be further automated?



Thank you

Please enjoy chatting in our booth
to find more information.

