

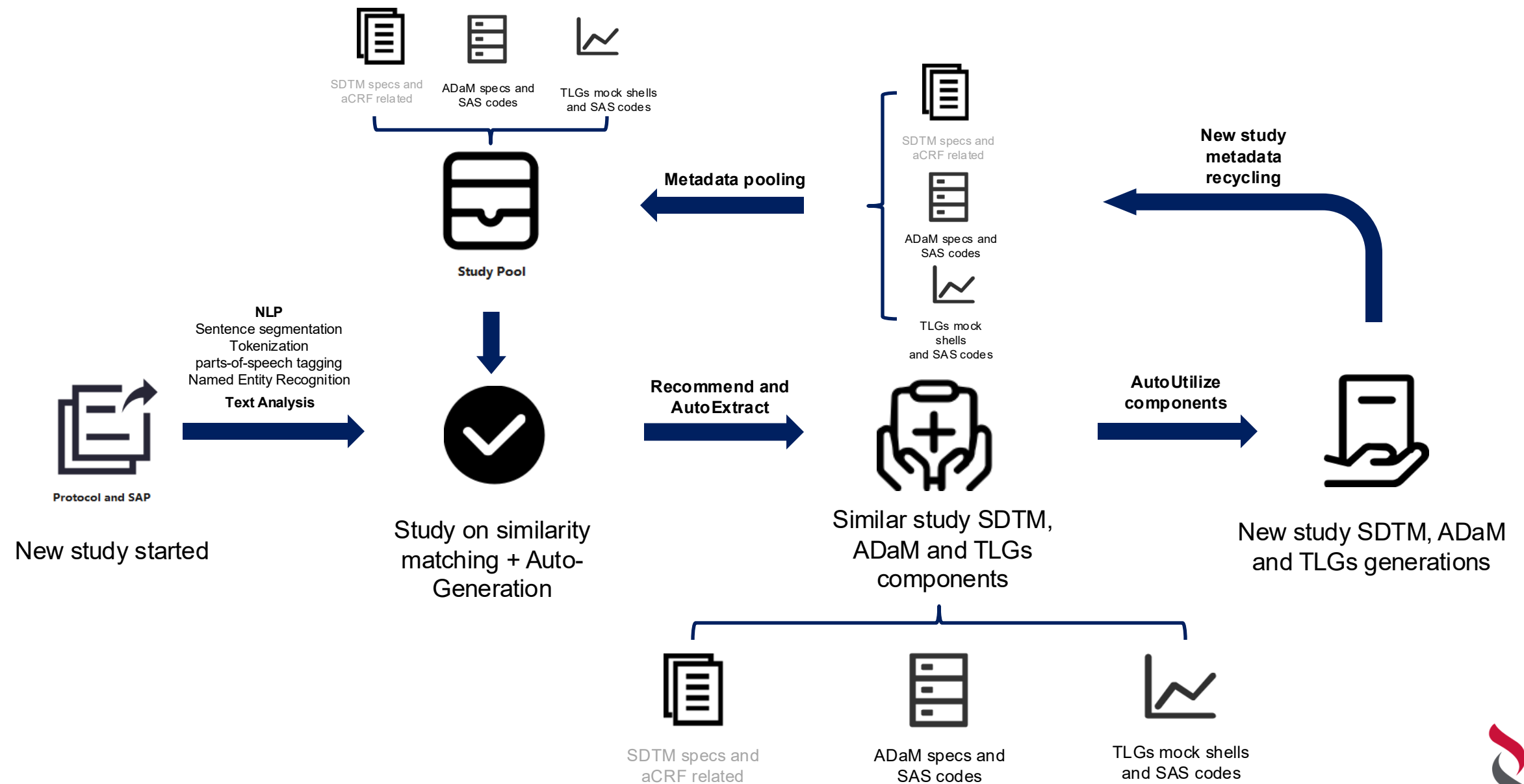
Approaches to Develop an Agent for Statistical Programming: Insights from Dizal Statistical Programming's Investigation

Zhiping Yan

2025-05-09



Intelligent Agent Process



Database Technologies Increasing Data Exchange

Database Design



Behavior Data



SDTM Metadata



ADaM Metadata



TLG Metadata

Study Metadata



Behavior Data



SDTM Metadata



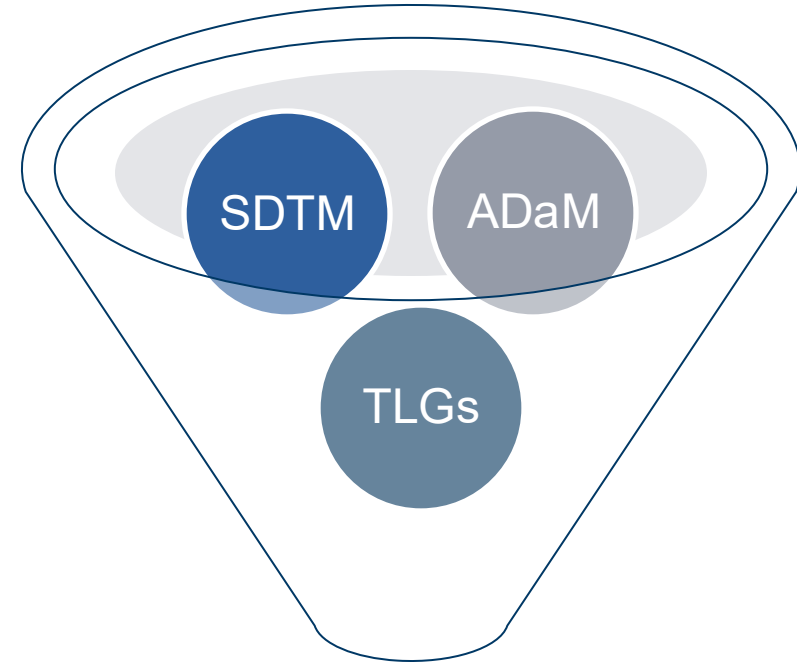
ADaM Metadata



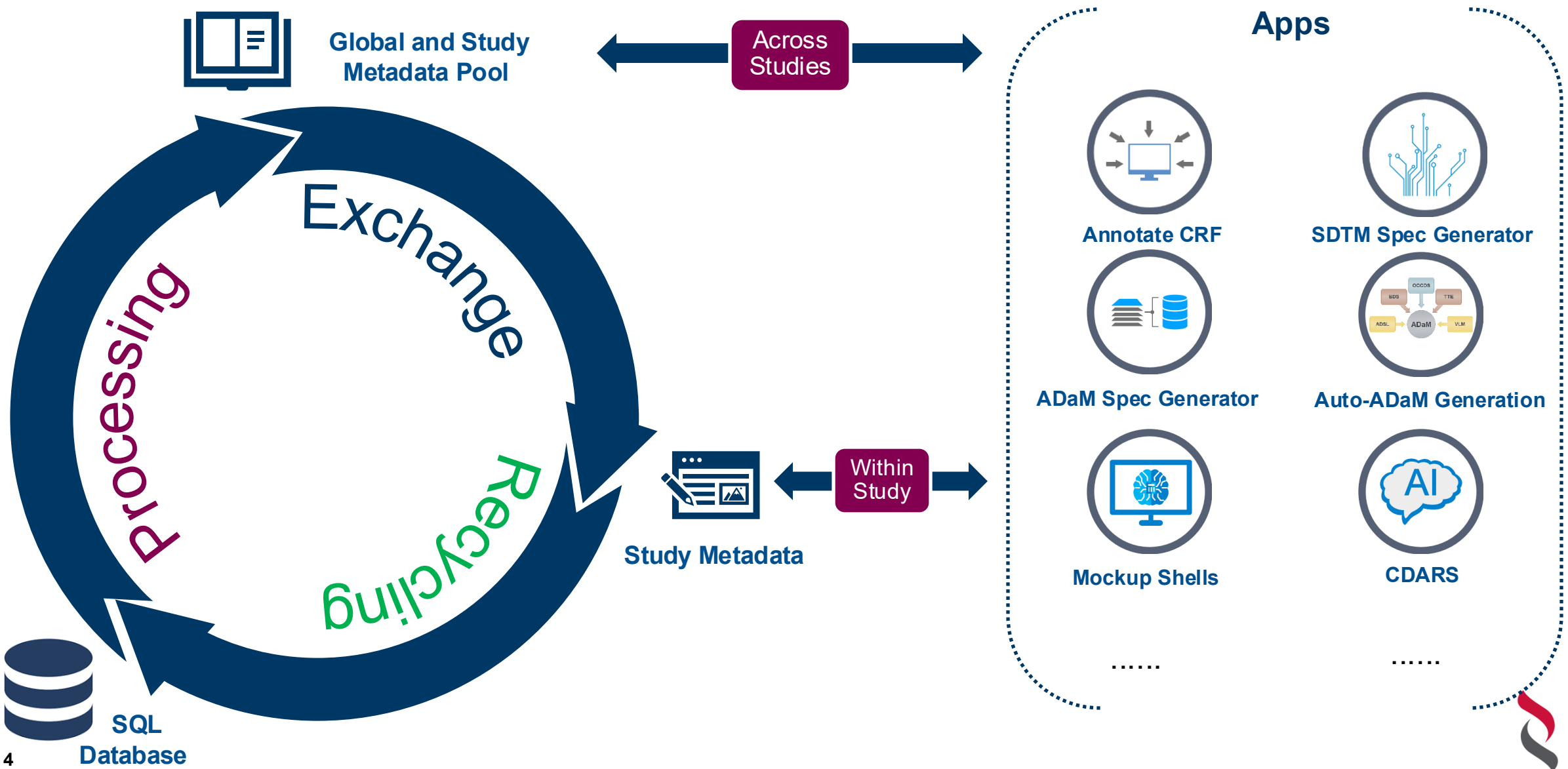
TLG Metadata



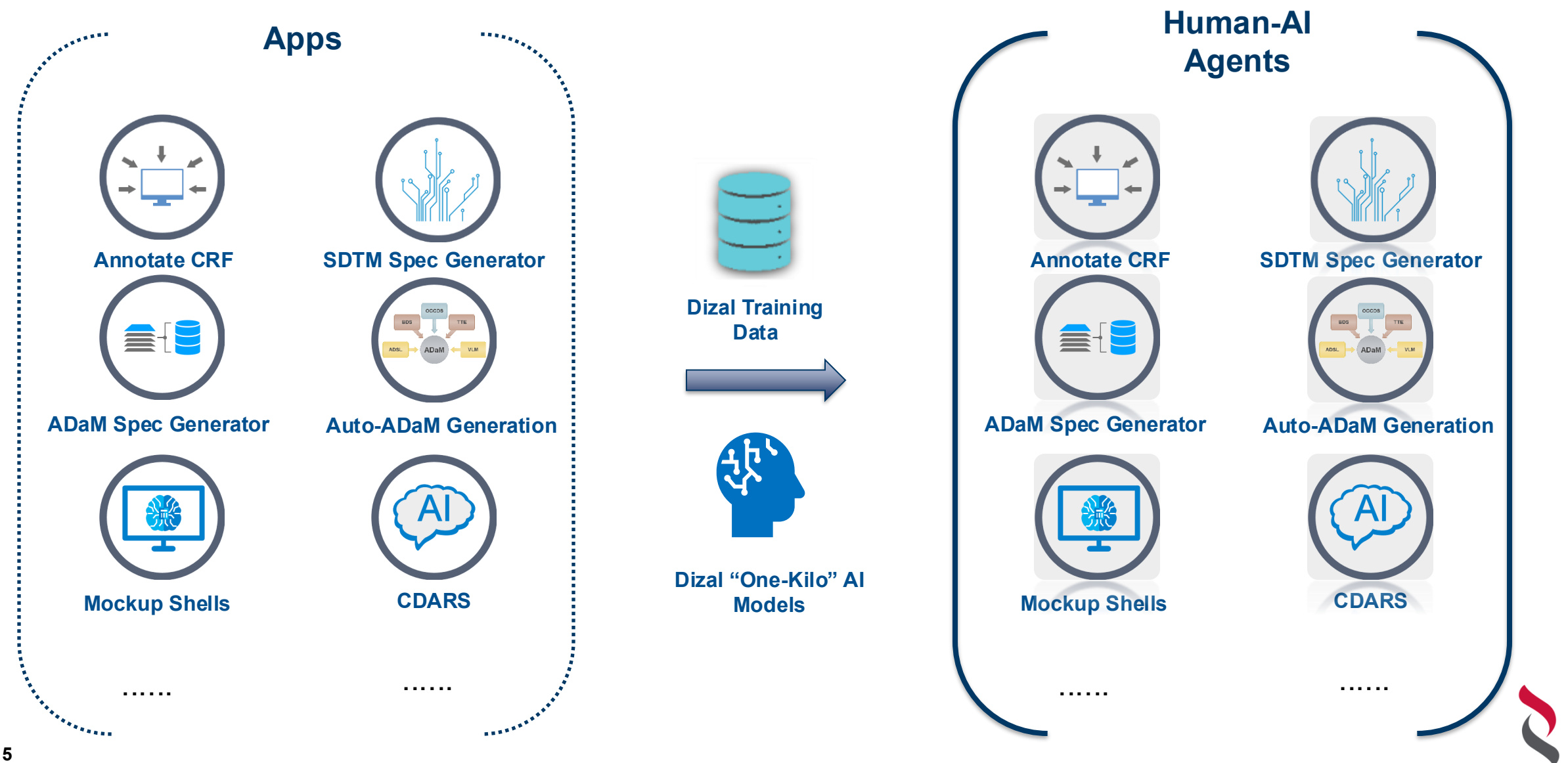
Across Studies Metadata Pool



Automated Metadata Exchange Process Within IoT Framework



Interventive AI: A Collaborative Human-AI Paradigm



Case 1: ADaM Datasets Generation

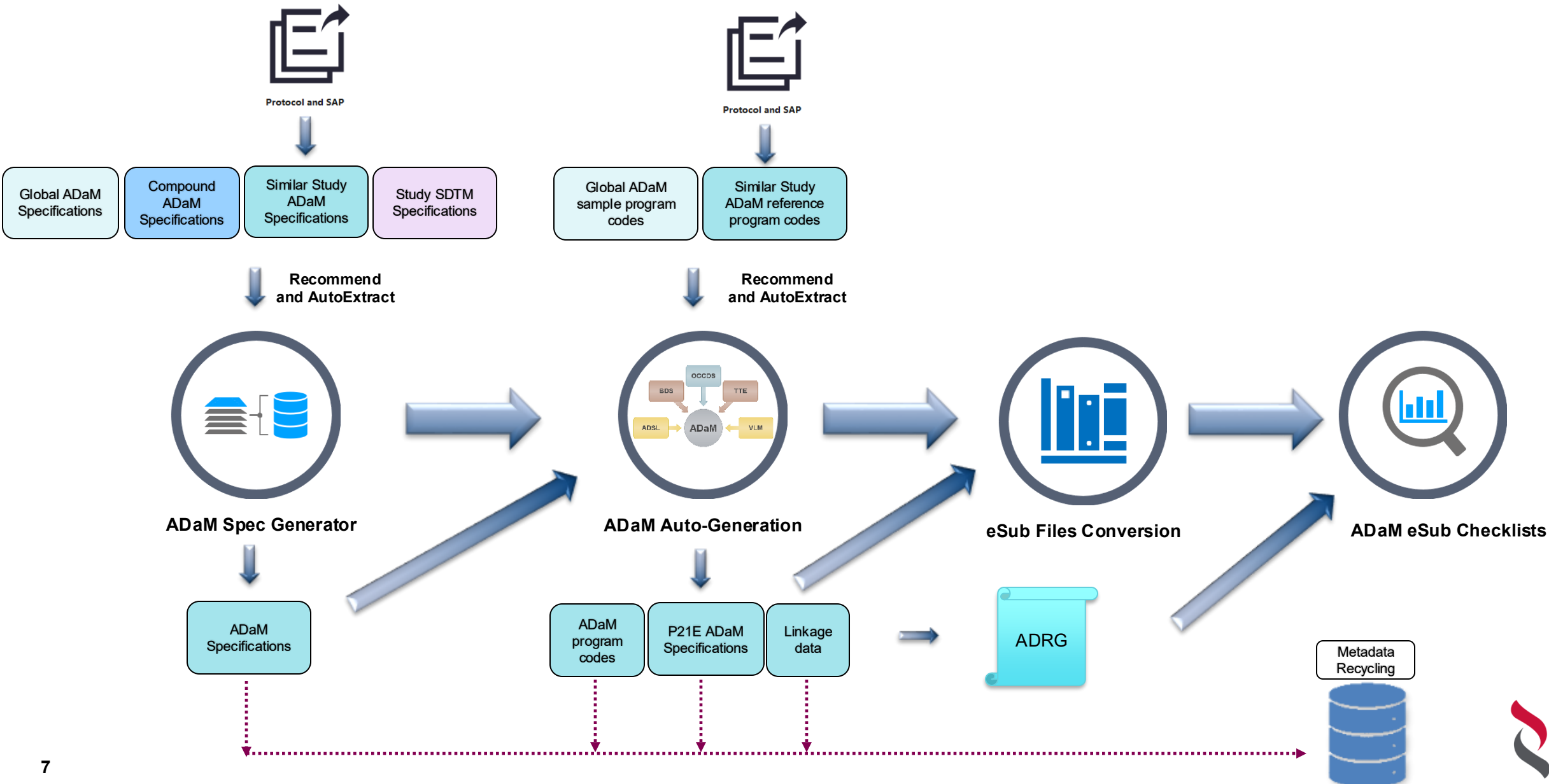
- Across studies and within study metadata exchange
- Search database and retrieve derivation rules for new variable/PARAMCD for ADaM specifications
- Recommend applicable ADaM reference code for the new study

Case 2: Automation of TLG

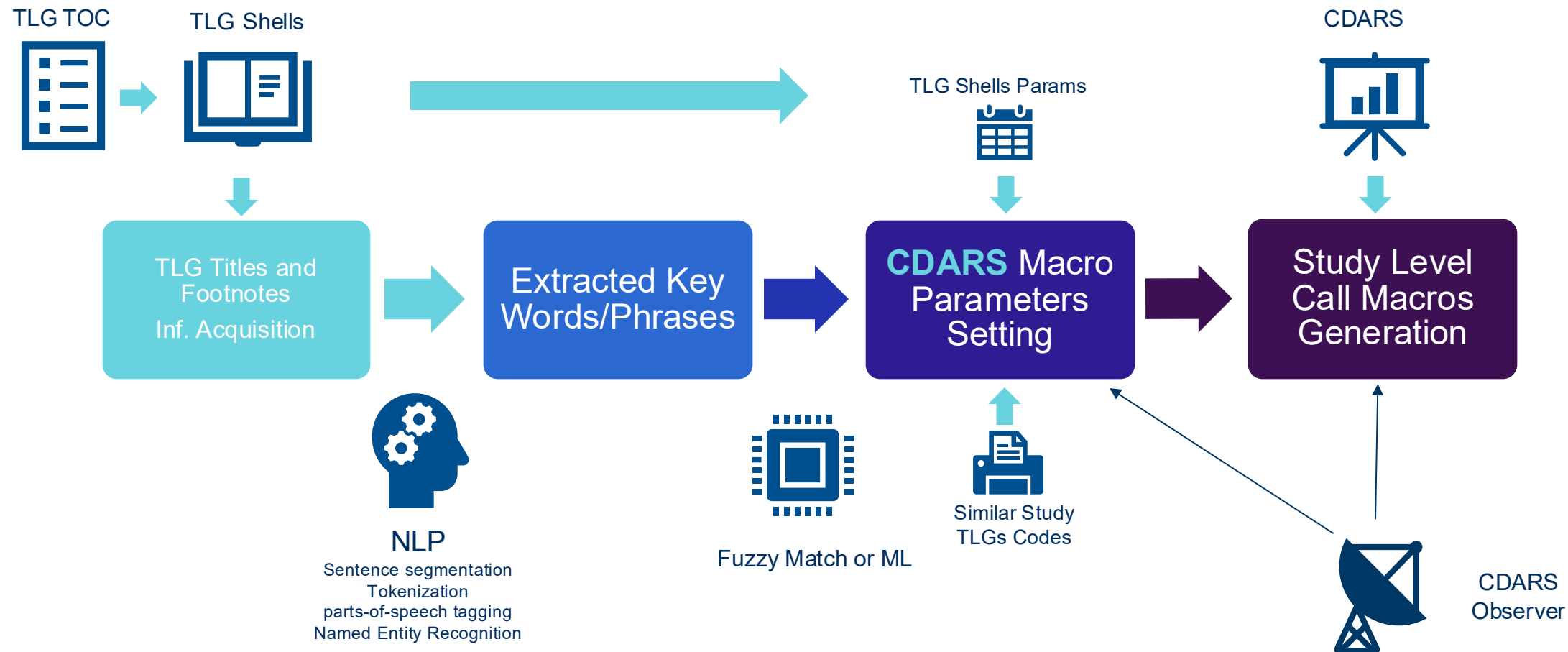
- Generate program based on NLP text analysis of title and TLG programs of similar study
- Accuracy checks for AI generated TLGs programs
- Across studies and within study metadata exchange



Case 1: ADaM Datasets Generation



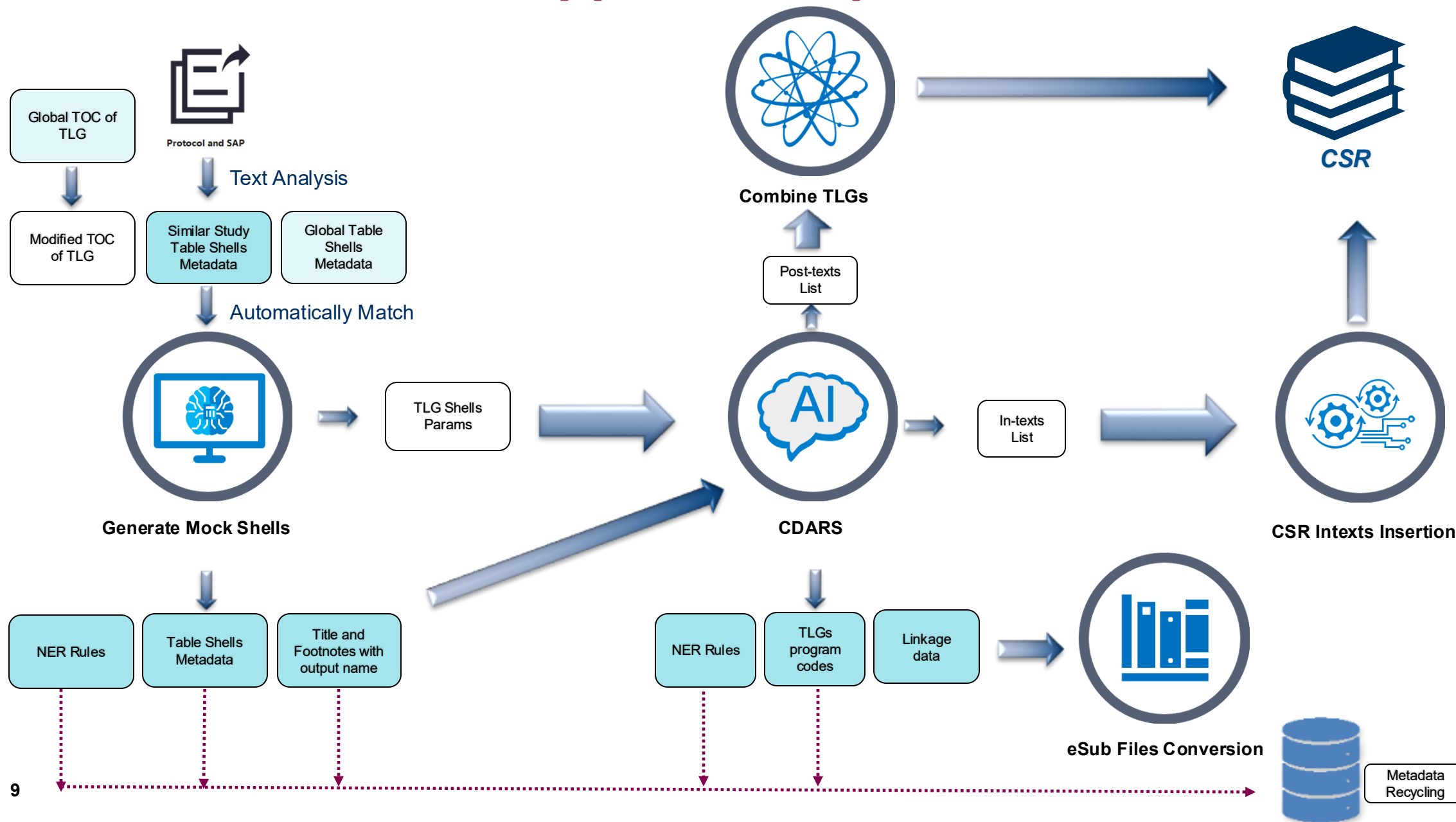
Case 2: Automation of TLG



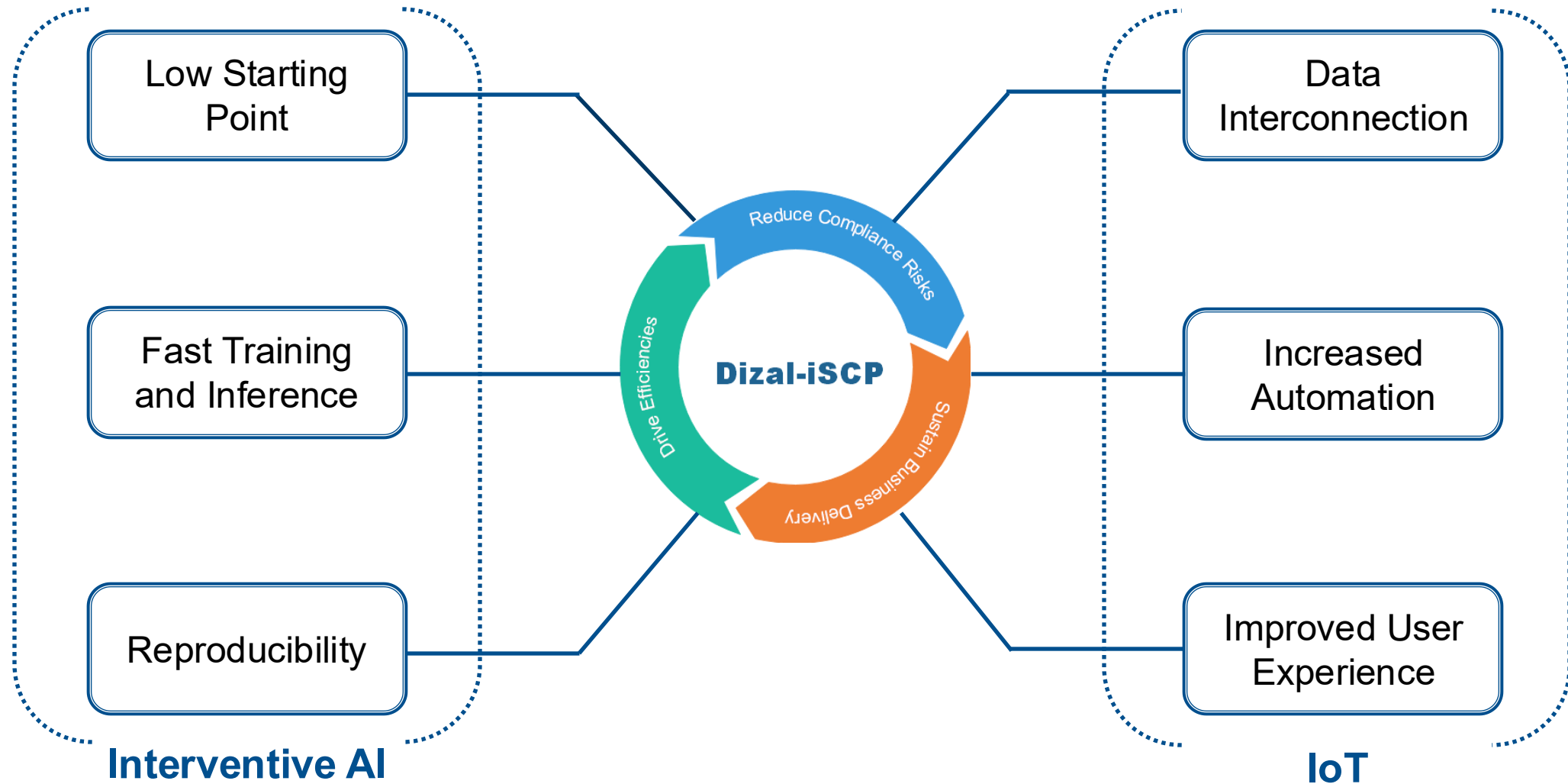
Clinical **D**ata **A**nalysis and **R**eporting **S**ystem (**CDARS**) is a One-click SAS® Application that produces standard statistical tables and listings based on Dizal's Non-efficacy (or efficacy) standard Tables, Graphs, and Listings.



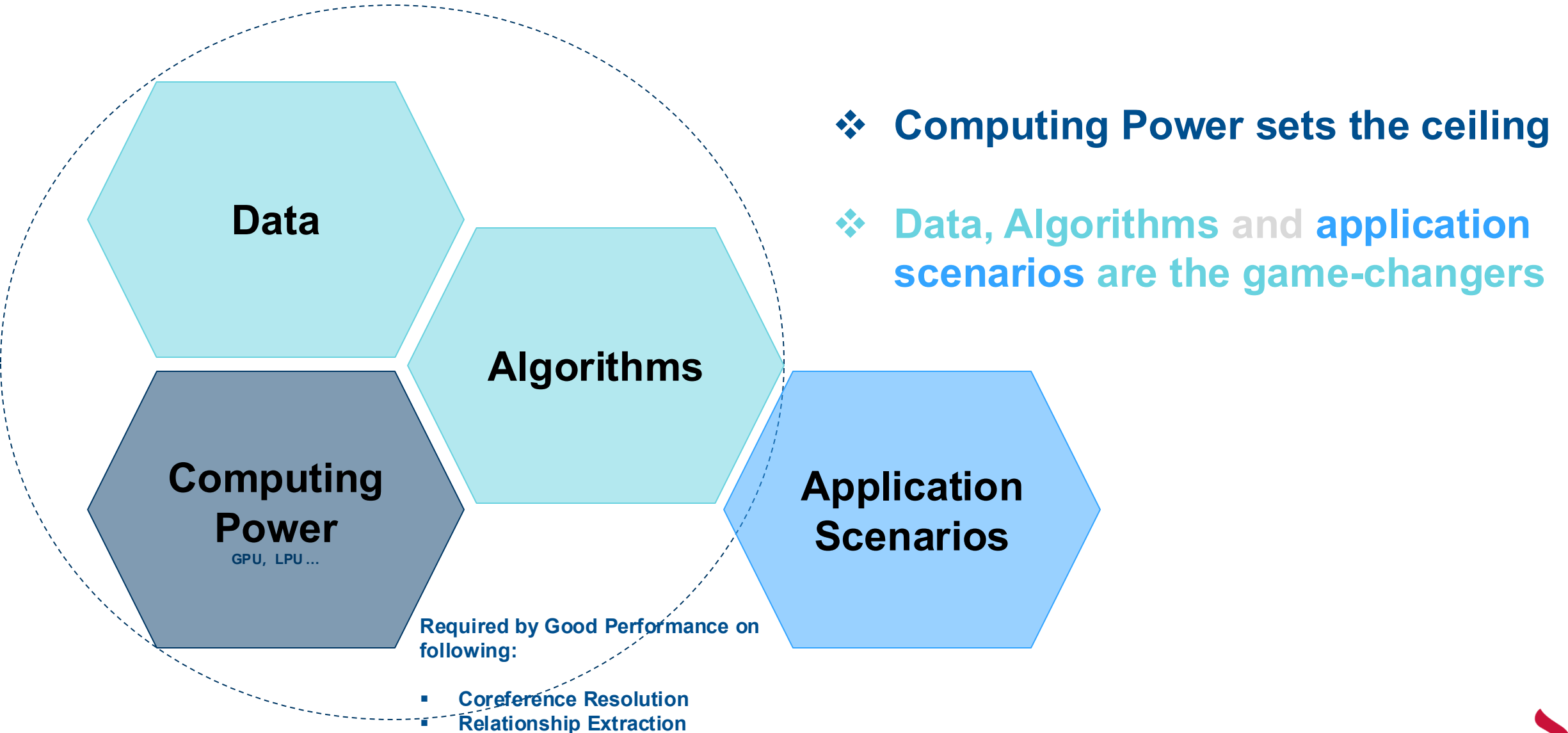
Automation of TLG Apps Development



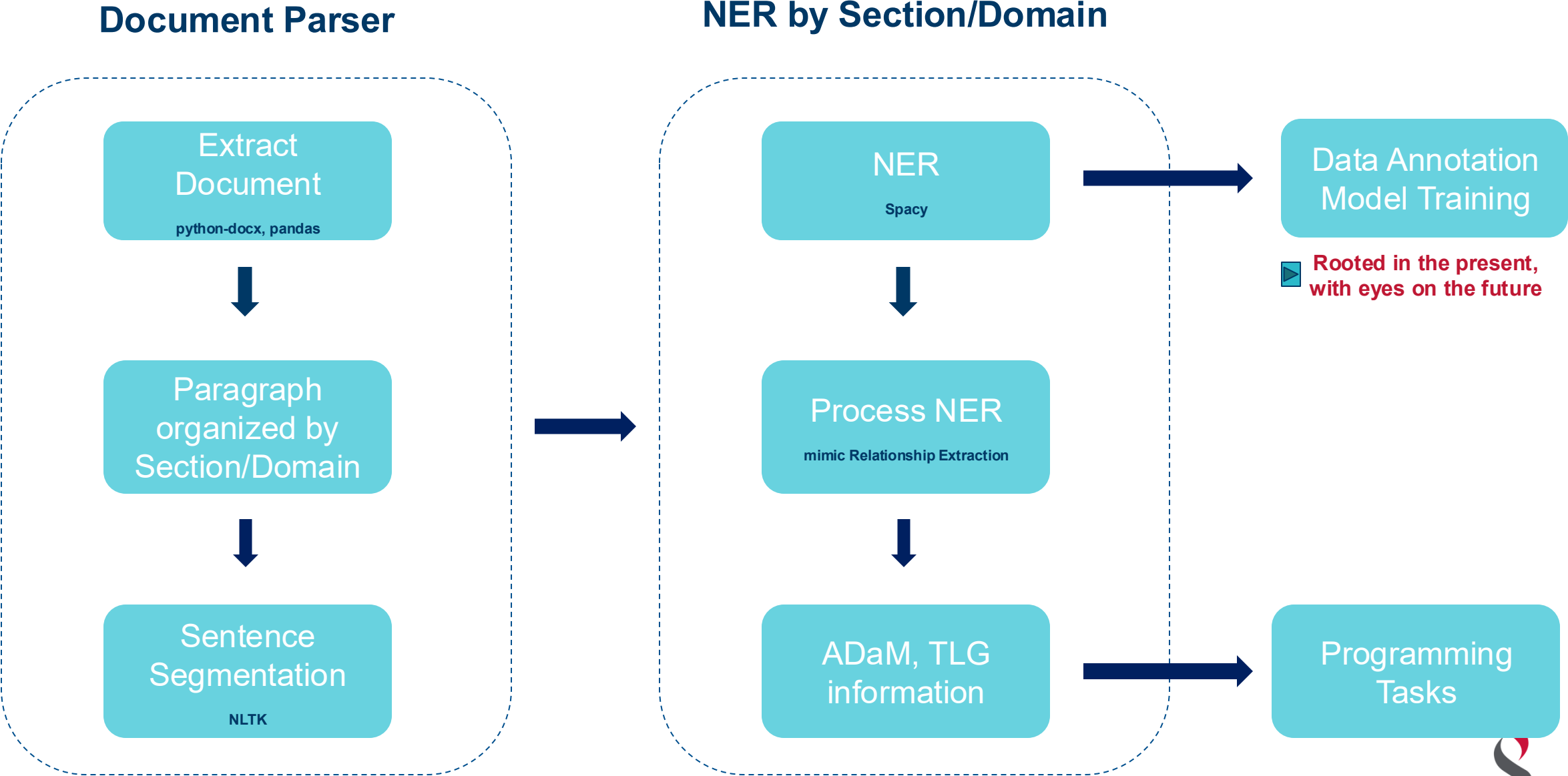
Minimum Viable AI for Biotech - Interventive AI and IoT Framework



Core Elements of AI



Dizal One-Kilo AI Models – SAP Extractor Example



SAP Extractor Example – NER

Source Text Data

Demographic and baseline characteristics will be listed and summarized in the safety analysis set and evaluable for CT-based response set, respectively. Standard descriptive statistics will be presented for the continuous variables below:

- Age (years)
- Height (cm) at baseline
- Weight (kg) at baseline
- Body mass index (BMI) in kg/m², calculated as (weight/height²) where weight is in kg and height is in m

The number and percentage of participants will be presented for the categorical variables below:

- Age group (<65 years, ≥65 years)
- Sex
- Race
- Ethnicity
- Country
- Geographical region (Asia: China and S. Korea, Non-Asia: United States and Australia)
- ECOG performance status at baseline (0, 1, 2)
- Bone marrow involvement at baseline by histology (yes, no)
- FDG-avid disease (yes, no) (only for participants with baseline PET scan)
- LDH elevation at baseline (yes, no)

No formal tests of statistical significance will be performed on the demographic and baseline data.

Structured Excel Data

Sentence	Token_ID	Start	End	String	ADaM
Demographic and baseline characteristics will be listed and summarized in the safety analysis set and evaluable for CT-based response set, respectively. Standard descriptive statistics will be presented for the continuous variables below:	DM Data	0	4	Demographic and baseline characteristics	
Demographic and baseline characteristics will be listed and summarized in the safety analysis set and evaluable for CT-based response set, respectively. Standard descriptive statistics will be presented for the continuous variables below:	AnalysisType	6	7	listed	
Demographic and baseline characteristics will be listed and summarized in the safety analysis set and evaluable for CT-based response set, respectively. Standard descriptive statistics will be presented for the continuous variables below:	AnalysisType	8	9	summarized	
Demographic and baseline characteristics will be listed and summarized in the safety analysis set and evaluable for CT-based response set, respectively. Standard descriptive statistics will be presented for the continuous variables below:	SAFFL	11	14	safety analysis set	Y
Demographic and baseline characteristics will be listed and summarized in the safety analysis set and evaluable for CT-based response set, respectively. Standard descriptive statistics will be presented for the continuous variables below:	EVACTFL	15	22	evaluable for CT	Y
Age (years)	AGE	0	3	Age	Y
Age (years)	AGEU	5	10	years	Y
Height (cm) at baseline	HEIGHTBL	0	6	Height	Y
Height (cm) at baseline	UNIT	8	10	cm	Y
Weight (kg) at baseline	WEIGHTBL	0	6	Weight	Y
Weight (kg) at baseline	UNIT	8	10	kg	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) where weight is in kg and height is in m	BMI	17	20	BMI	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) where weight is in kg and height is in m	UNIT	25	27	kg	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) where weight is in kg and height is in m	UNIT	25	30	kg/m2	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) where weight is in kg and height is in m	WEIGHTBL	47	53	weight	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) where weight is in kg and height is in m	WEIGHTBL	69	75	weight	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) where weight is in kg and height is in m	UNIT	82	84	kg	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) where weight is in kg and height is in m	HEIGHTBL	89	95	height	Y
Age group (<65 years, ≥65 years)	AGEGR1	0	9	Age group	
Age group (<65 years, ≥65 years)	Categories	11	14	<65	
Age group (<65 years, ≥65 years)	Categories	22	25	≥65	
Sex	SEX	0	3	Sex	Y
Race	RACEGR1	0	4	Race	
Ethnicity	ETHNIC	0	9	Ethnicity	Y
Country	COUNTRY	0	7	Country	Y
Geographical region (Asia: China and S. Korea, Non-Asia: United States and Australia)	REGION1	0	19	Geographical region	
Geographical region (Asia: China and S. Korea, Non-Asia: United States and Australia)	Categories	21	25	Asia	
Geographical region (Asia: China and S. Korea, Non-Asia: United States and Australia)	Subcategories	27	32	China	
Geographical region (Asia: China and S. Korea, Non-Asia: United States and Australia)	Subcategories	37	45	S. Korea	
Geographical region (Asia: China and S. Korea, Non-Asia: United States and Australia)	Categories	47	55	Non-Asia	
Geographical region (Asia: China and S. Korea, Non-Asia: United States and Australia)	Categories	51	55	Asia	
Geographical region (Asia: China and S. Korea, Non-Asia: United States and Australia)	Subcategories	57	70	United States	
Geographical region (Asia: China and S. Korea, Non-Asia: United States and Australia)	Subcategories	75	84	Australia	
ECOG performance status at baseline (0, 1, 2)	Other	0	4	ECOG	
ECOG performance status at baseline (0, 1, 2)	ECOGBL	0	4	ECOG	Y
ECOG performance status at baseline (0, 1, 2)	Categories	37	38	0	
ECOG performance status at baseline (0, 1, 2)	Categories	40	41	1	
ECOG performance status at baseline (0, 1, 2)	Categories	43	44	2	

SAP Extractor Example – TLG Information

The laboratory tests as specified in Section 4.2.2 will be summarized in the safety analysis set.

For the laboratory test with applicable CTCAE grade criteria, shift tables will be provided to show the shift from baseline CTCAE grade to the worst grade post-baseline. For the tests (like ALT, AST, ALP, and total bilirubin) where CTCAE grade is not applicable for the baseline value, summary of worst grade post-baseline by baseline status (normal/abnormal) will be provided instead of shift table. In addition, summary tables will be provided to present the maximum grade increase from baseline for each abnormality applicable.

To present the status change in clinical significance from baseline to the worst case (abnormal clinically significant > abnormal not clinically significant > normal) post-baseline, shift tables will be provided for the laboratory tests not covered by CTCAE.

In the shift tables, only the patients with non-missing data at both baseline and the corresponding post-baseline visit(s) will be included.

Number and percentage of patients meeting the criteria in Appendix D during the analysis period (from the first dose date of drug until 28 days after the last dose) will be reported. Shift table will also be provided to tabulate the AST or ALT elevation status at baseline against the worst case during the analysis period (only list the number and percentage of patients worsening against baseline).

Graphic displays may be provided for some laboratory tests of interest (such as tests related to common AEs) to show the trend of change relative to the baseline.

All abnormal laboratory data observed in the study will be included in the individual data listing.

Source Text Data

Structured Excel Data



Sentence	Shell	Output	Title	Analysis Set
The laboratory tests as specified in Section 4.2.2	LB0101a	14.3.2	Summary of Chemistry with Change from Baseline by Visit	Safety Analysis Set
The laboratory tests as specified in Section 4.2.2	LB0101a	14.3.2	Summary of Hematology with Change from Baseline by Visit	Safety Analysis Set
The laboratory tests as specified in Section 4.2.2	LB0101a	14.3.2	Summary of Coagulation with Change from Baseline by Visit	Safety Analysis Set
For the laboratory test with applicable CTCAE grade criteria	LB0302	14.3.2	Shift Table of Chemistry Based on CTCAE Grade by Visit	Safety Analysis Set
For the laboratory test with applicable CTCAE grade criteria	LB0302	14.3.2	Shift Table of Hematology Based on CTCAE Grade by Visit	Safety Analysis Set
For the laboratory test with applicable CTCAE grade criteria	LB0302	14.3.2	Shift Table of Coagulation Based on CTCAE Grade by Visit	Safety Analysis Set
For the laboratory test with applicable CTCAE grade criteria	LB0302	14.3.2	Shift Table of Urinalysis Based on CTCAE Grade by Visit	Safety Analysis Set
For the tests (like ALT, AST, ALP, and total bilirubin)	LB0302	14.3.2	Shift Table of Chemistry Based on CTCAE Grade by Visit	Safety Analysis Set
For the tests (like ALT, AST, ALP, and total bilirubin)	LB0302	14.3.2	Shift Table of Hematology Based on CTCAE Grade by Visit	Safety Analysis Set
For the tests (like ALT, AST, ALP, and total bilirubin)	LB0302	14.3.2	Shift Table of Coagulation Based on CTCAE Grade by Visit	Safety Analysis Set
For the tests (like ALT, AST, ALP, and total bilirubin)	LB0302	14.3.2	Shift Table of Urinalysis Based on CTCAE Grade by Visit	Safety Analysis Set
In addition, summary tables will be provided to present the maximum grade increase from baseline for each abnormality applicable	LB0305a	14.3.2	Summary of Maximum CTCAE Grade Increase from Baseline in Chemistry	Safety Analysis Set
In addition, summary tables will be provided to present the maximum grade increase from baseline for each abnormality applicable	LB0305a	14.3.2	Summary of Maximum CTCAE Grade Increase from Baseline in Hematology	Safety Analysis Set
In addition, summary tables will be provided to present the maximum grade increase from baseline for each abnormality applicable	LB0305a	14.3.2	Summary of Maximum CTCAE Grade Increase from Baseline in Coagulation	Safety Analysis Set
In addition, summary tables will be provided to present the maximum grade increase from baseline for each abnormality applicable	LB0305a	14.3.2	Summary of Maximum CTCAE Grade Increase from Baseline in Urinalysis	Safety Analysis Set
To present the status change in clinical significance from baseline to the worst case	LB0202a	14.3.2	Shift Table of Chemistry Regarding Clinical Significance by Visit	Safety Analysis Set
To present the status change in clinical significance from baseline to the worst case	LB0202a	14.3.2	Shift Table of Hematology Regarding Clinical Significance by Visit	Safety Analysis Set
To present the status change in clinical significance from baseline to the worst case	LB0202a	14.3.2	Shift Table of Coagulation Regarding Clinical Significance by Visit	Safety Analysis Set
To present the status change in clinical significance from baseline to the worst case	LB0202a	14.3.2	Shift Table of Urinalysis Regarding Clinical Significance by Visit	Safety Analysis Set
To present the status change in clinical significance from baseline to the worst case	LB0201a	14.3.2	Shift Table of Chemistry Regarding the Worst Clinical Significance	Safety Analysis Set
To present the status change in clinical significance from baseline to the worst case	LB0201a	14.3.2	Shift Table of Hematology Regarding the Worst Clinical Significance	Safety Analysis Set
To present the status change in clinical significance from baseline to the worst case	LB0201a	14.3.2	Shift Table of Coagulation Regarding the Worst Clinical Significance	Safety Analysis Set
To present the status change in clinical significance from baseline to the worst case	LB0201a	14.3.2	Shift Table of Urinalysis Regarding the Worst Clinical Significance	Safety Analysis Set
In the shift tables, only the patients with non-missing data at both baseline and the corresponding post-baseline visit(s) will be included	LB0202a	14.3.2	Shift Table of Chemistry Regarding Clinical Significance by Visit	Safety Analysis Set
In the shift tables, only the patients with non-missing data at both baseline and the corresponding post-baseline visit(s) will be included	LB0202a	14.3.2	Shift Table of Hematology Regarding Clinical Significance by Visit	Safety Analysis Set
In the shift tables, only the patients with non-missing data at both baseline and the corresponding post-baseline visit(s) will be included	LB0202a	14.3.2	Shift Table of Coagulation Regarding Clinical Significance by Visit	Safety Analysis Set
In the shift tables, only the patients with non-missing data at both baseline and the corresponding post-baseline visit(s) will be included	LB0202a	14.3.2	Shift Table of Urinalysis Regarding Clinical Significance by Visit	Safety Analysis Set
Shift table will also be provided to tabulate the AST or ALT elevation status at baseline against the worst case during the analysis period	LB0401b	14.3.2	Summary of Laboratory Abnormalities Associated with Liver Function Tests	Safety Analysis Set
Graphic displays may be provided for some laboratory tests of interest	F-LB0101	14.3.2	Line Graph for Mean (±SD) for Change from Baseline in Chemistry	Safety Analysis Set
Graphic displays may be provided for some laboratory tests of interest	F-LB0101	14.3.2	Line Graph for Mean (±SD) for Change from Baseline in Hematology	Safety Analysis Set
Graphic displays may be provided for some laboratory tests of interest	F-LB0101	14.3.2	Line Graph for Mean (±SD) for Change from Baseline in Coagulation	Safety Analysis Set
All abnormal laboratory data observed in the study will be included in the individual data listing	LB9901a	16.2.1	Abnormal Laboratory Value	Safety Analysis Set



Convert xlsx to JSONL Data

Sentence	Token ID	Start	End	String	ADaM
Demographic and baseline characteristics will be listed and summarized	DM Data	0	4	Demographic and baseline cha	
Demographic and baseline characteristics will be listed and summarized	AnalysisType	6	7	listed	
Demographic and baseline characteristics will be listed and summarized	AnalysisType	8	9	summarized	
Demographic and baseline characteristics will be listed and summarized	SAFFL	11	14	safety analysis s Y	
Demographic and baseline characteristics will be listed and summarized	EVACTFL	15	22	evaluable for CT Y	
Age (years)	AGE	0	3	Age	Y
Age (years)	AGEU	5	10	years	Y
Height (cm) at baseline	HEIGHTBL	0	6	Height	Y
Height (cm) at baseline	UNIT	8	10	cm	Y
Weight (kg) at baseline	WEIGHTBL	0	6	Weight	Y
Weight (kg) at baseline	UNIT	8	10	kg	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) whe	BMIBL	17	20	BMI	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) whe	UNIT	25	27	kg	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) whe	UNIT	25	30	kg/m2	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) whe	WEIGHTBL	47	53	weight	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) whe	WEIGHTBL	69	75	weight	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) whe	UNIT	82	84	kg	Y
Body mass index (BMI) in kg/m2, calculated as (weight/height2) whe	HEIGHTBL	89	95	height	Y
Age group (<65 years, ≥65 years)	AGEGR1	0	9	Age group	
Age group (<65 years, ≥65 years)	Categories	11	14	<65	
Age group (<65 years, ≥65 years)	Categories	22	25	≥65	
Sex	SEX	0	3	Sex	Y
Race	RACEGR1	0	4	Race	
Ethnicity	ETHNIC	0	9	Ethnicity	Y
Country	COUNTRY	0	7	Country	Y
Geographical region (Asia: China and S. Korea, Non-Asia: United Stat	REGION1	0	19	Geographical region	
Geographical region (Asia: China and S. Korea, Non-Asia: United Stat	Categories	21	25	Asia	
Geographical region (Asia: China and S. Korea, Non-Asia: United Stat	Subcategories	27	32	China	
Geographical region (Asia: China and S. Korea, Non-Asia: United Stat	Subcategories	37	45	S. Korea	
Geographical region (Asia: China and S. Korea, Non-Asia: United Stat	Categories	47	55	Non-Asia	
Geographical region (Asia: China and S. Korea, Non-Asia: United Stat	Categories	51	55	Asia	
Geographical region (Asia: China and S. Korea, Non-Asia: United Stat	Subcategories	57	70	United States	
Geographical region (Asia: China and S. Korea, Non-Asia: United Stat	Subcategories	75	84	Australia	
ECOG performance status at baseline (0, 1, 2)	Other	0	4	ECOG	
ECOG performance status at baseline (0, 1, 2)	ECOGBL	0	4	ECOG	Y
ECOG performance status at baseline (0, 1, 2)	Categories	37	38	0	
ECOG performance status at baseline (0, 1, 2)	Categories	40	41	1	
ECOG performance status at baseline (0, 1, 2)	Categories	43	44	2	



```

{"text": "Age (years)",
 "entities": [
   {"id": 0, "start_offset": 0, "end_offset": 3, "label": "AGE", "text": "Age"},
   {"id": 1, "start_offset": 5, "end_offset": 10, "label": "AGEU", "text": "years"}
 ],
 "relations": []
}

{"text": "Age group (<65 years, ≥65 years)",
 "entities": [
   {"id": 0, "start_offset": 0, "end_offset": 9, "label": "AGEGR1", "text": "Age group"},
   {"id": 1, "start_offset": 11, "end_offset": 14, "label": "Categories", "text": "<65"},
   {"id": 2, "start_offset": 22, "end_offset": 25, "label": "Categories", "text": "≥65"}
 ],
 "relations": [
   {"id": 0, "from_id": 0, "to_id": 1, "type": "has_category"},
   {"id": 1, "from_id": 0, "to_id": 2, "type": "has_category"}
 ]
}

{"text": "ECOG performance status at baseline (0, 1, 2)",
 "entities": [
   {"id": 0, "start_offset": 0, "end_offset": 4, "label": "Other", "text": "ECOG"},
   {"id": 1, "start_offset": 0, "end_offset": 4, "label": "ECOGBL", "text": "ECOG"},
   {"id": 2, "start_offset": 37, "end_offset": 38, "label": "Categories", "text": "0"},
   {"id": 3, "start_offset": 40, "end_offset": 41, "label": "Categories", "text": "1"},
   {"id": 4, "start_offset": 43, "end_offset": 44, "label": "Categories", "text": "2"}
 ],
 "relations": [
   {"id": 0, "from_id": 1, "to_id": 2, "type": "has_category"},
   {"id": 1, "from_id": 1, "to_id": 3, "type": "has_category"},
   {"id": 2, "from_id": 1, "to_id": 4, "type": "has_category"}
 ]
}

```



Deploy Doccano and Create Sequence Labeling Project

1. `conda` create new environment

```
conda create --name doccano_env python=3.8
```

2. `Activate` new environment

```
conda activate doccano_env
```

3. `Install` pandas version = 1.3.5

```
conda install pandas=1.3.5
```

4. `Install` doccano

```
pip install doccano
```

5. `Create` database `and` account

```
doccano init  
doccano createuser --username admin --password pass
```

6. `Start` doccano

```
doccano webserver --port 9000
```

7. `open` another terminal to start doccano task

```
doccano task
```

8. `open` webpage `and` input localhost:9000 to start working

Project name

sap_dm

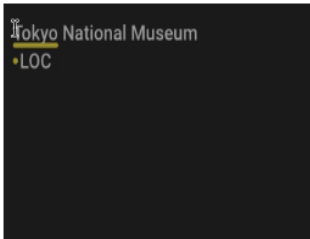
Description

This project is to label NER for demographic data section from SAP

Tags

☒ Allow project members to create label types

☒ Allow overlapping spans



Tokyo National Museum
•LOC

☒ Use relation labeling

☐ Count [grapheme clusters](#) as one character

☐ Randomize document order

☒ Share annotations across all users

Create

Import JSONL Data into Doccano for Annotation

Start Annotation

Home

Dataset

Labels

Relations

Members

Comments

Guideline

Metrics

Settings

×

⌵

⇅

📄

💬

✎

🗑

📄

has_category

has_category

Age group (<65 years, ≥65 years)

•AGEGR1 •Categories

•Categories

2 of 16 |< < > >|

Progress

Total16

Complete0

0%

Label Types

Span

AGE AGEU

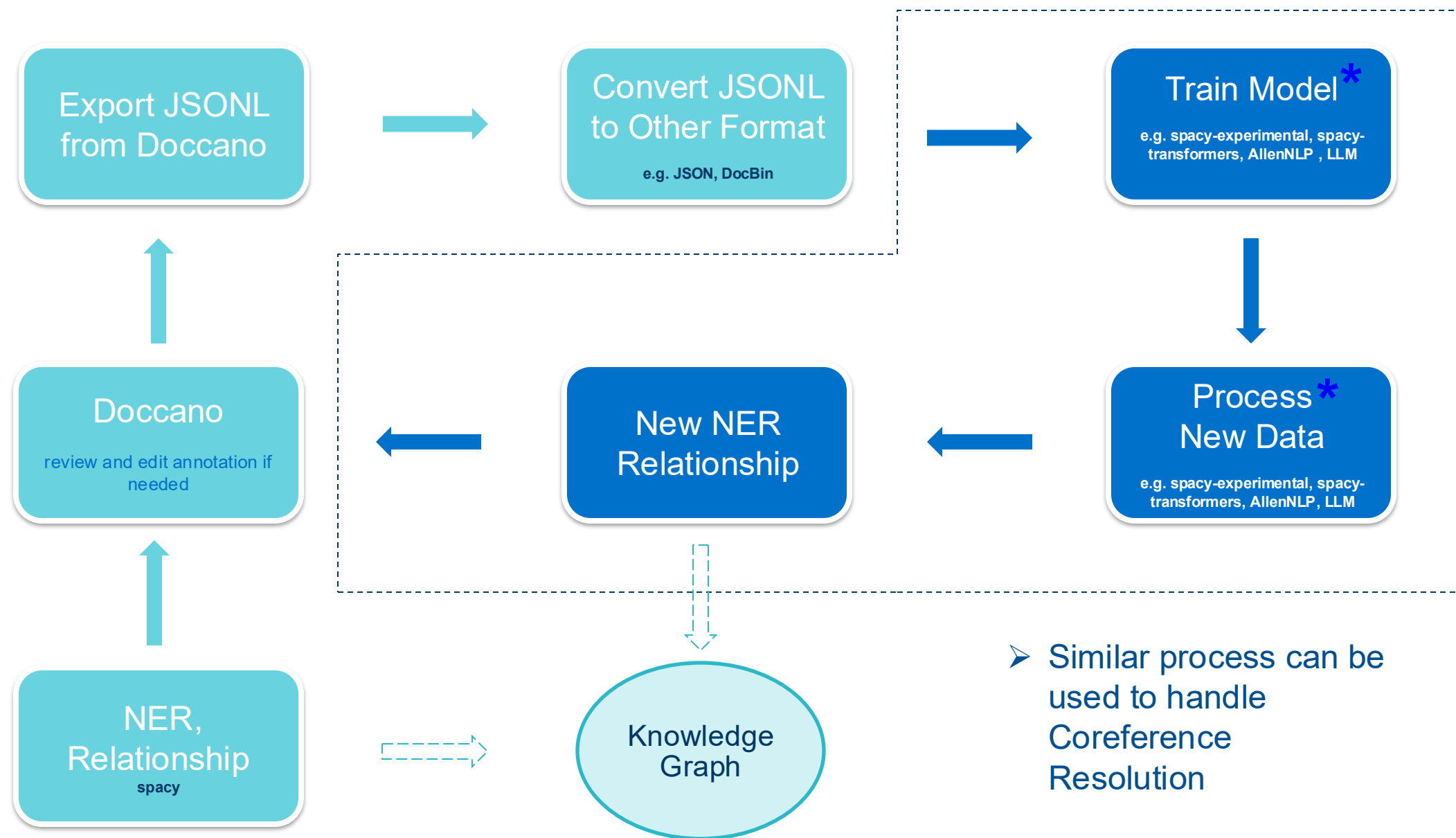
AGEGR1 Categories

BMIBL UNIT

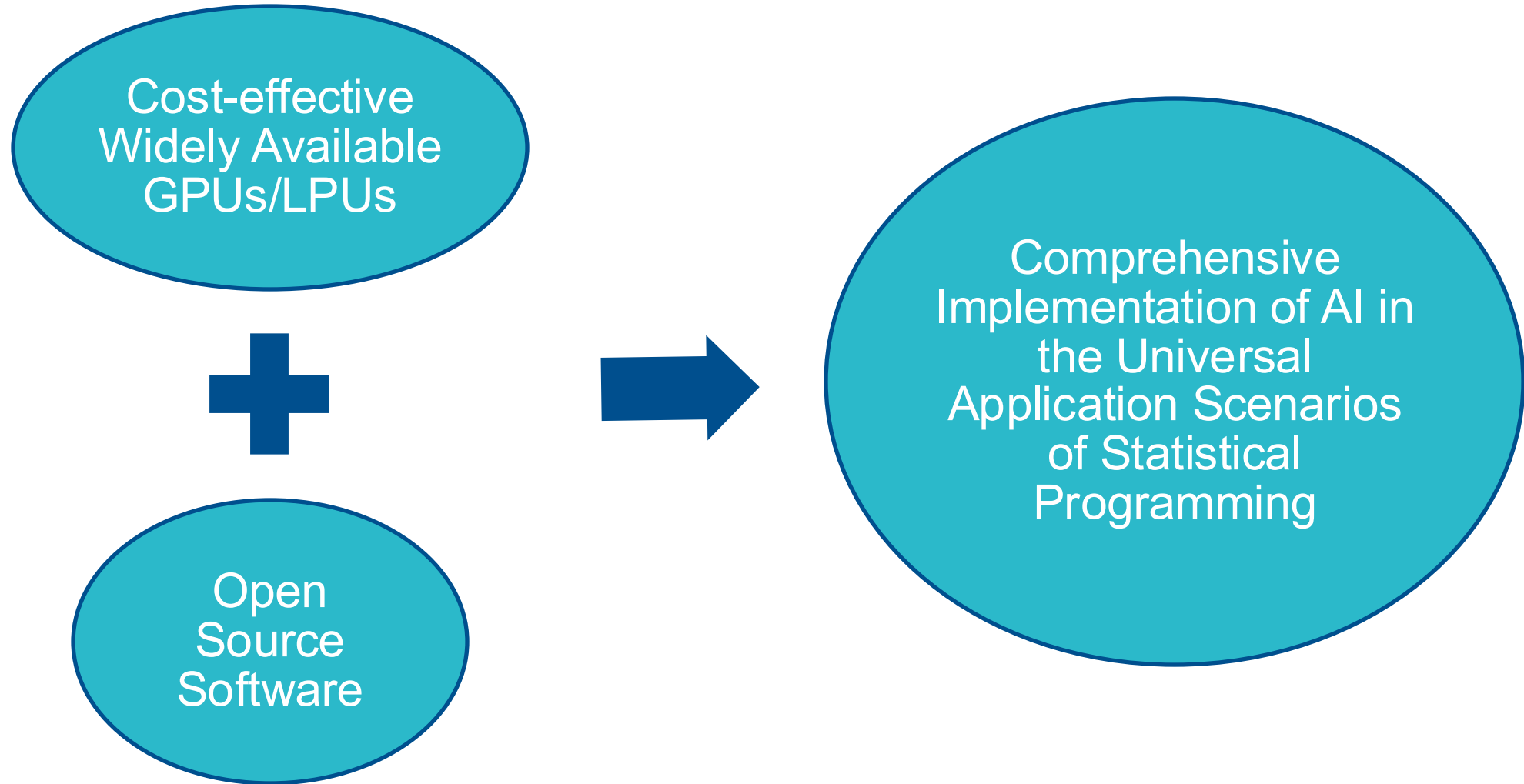
WEIGHTBL

17

Dizal NER & Relationship Extraction Models Process



The Road Ahead?



Acknowledgements

Thanks Huadan Li, Zongming Guo, Wei Zhang and all statistical programming team members for the great supports!



