



IDD'AI:

Automating the Tedious So You Can Focus on the Genius

Lionel Debecq, Senior Stat Programmer
Samantha Cambier, Senior Biostatistician

Date: 17Nov2025

challenging the status quo since 1991

Agenda



**Context and
IDD'AI idea**



Prerequisites



**Workflow
overview**



IDD'AI

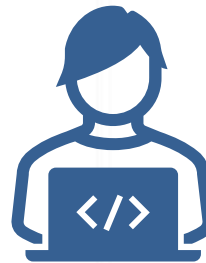
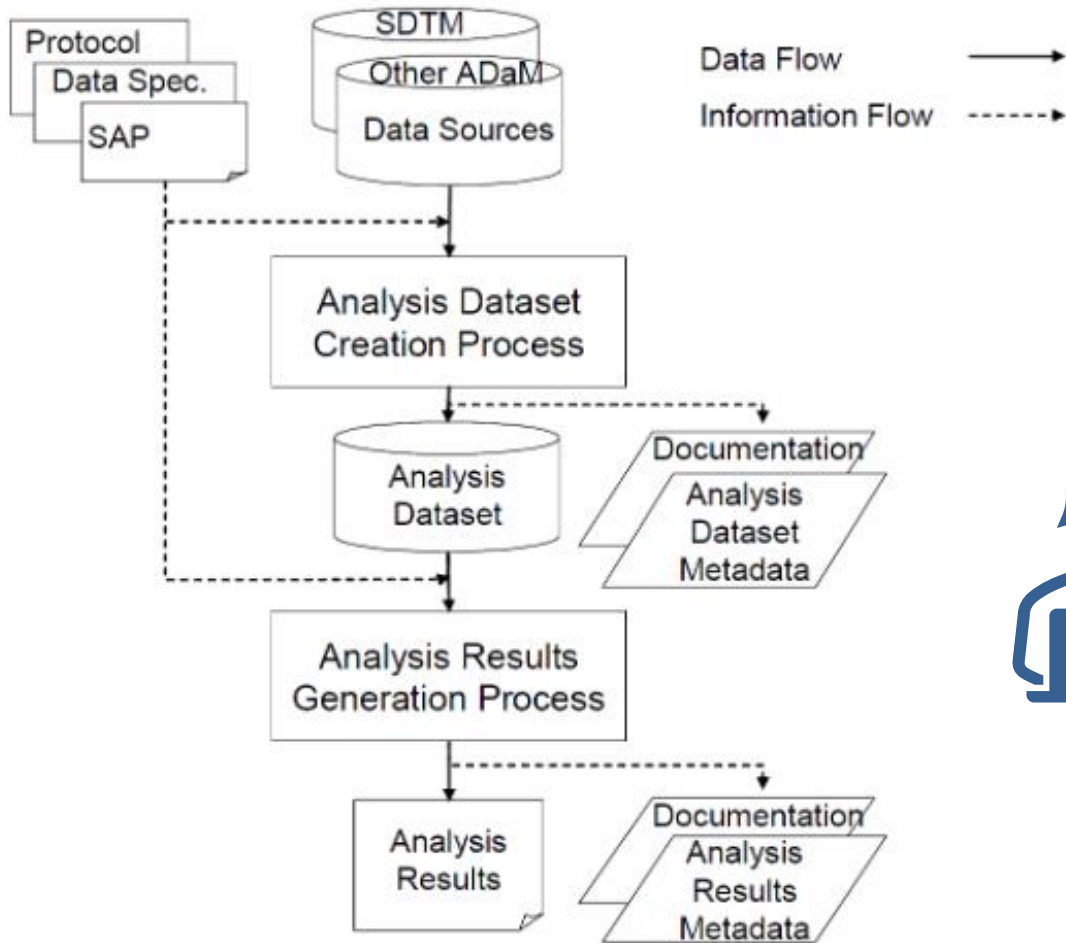


Benefits



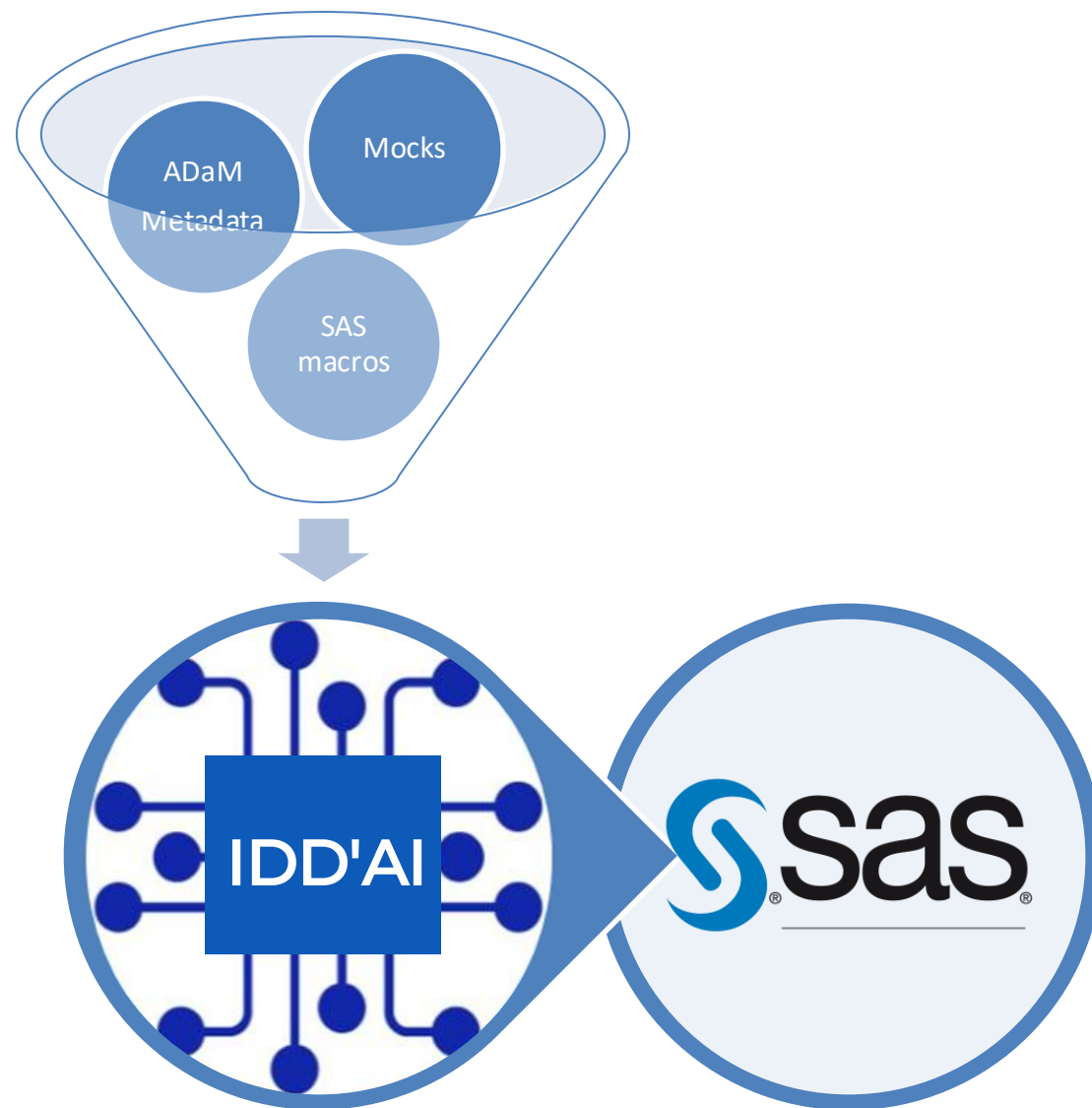
**Conclusion
and Next
steps**

Context



- Time-consuming
- Error-prone
- Manual and Slow
- Potential for Automation

Idea



Prerequisites: Mocks - Unique

•Structure:

Title → Protocol → Population → Table → Footnotes

•Table layout:

- First column = Statistic description
- Next columns = Analysis groups
- Results printed in the body

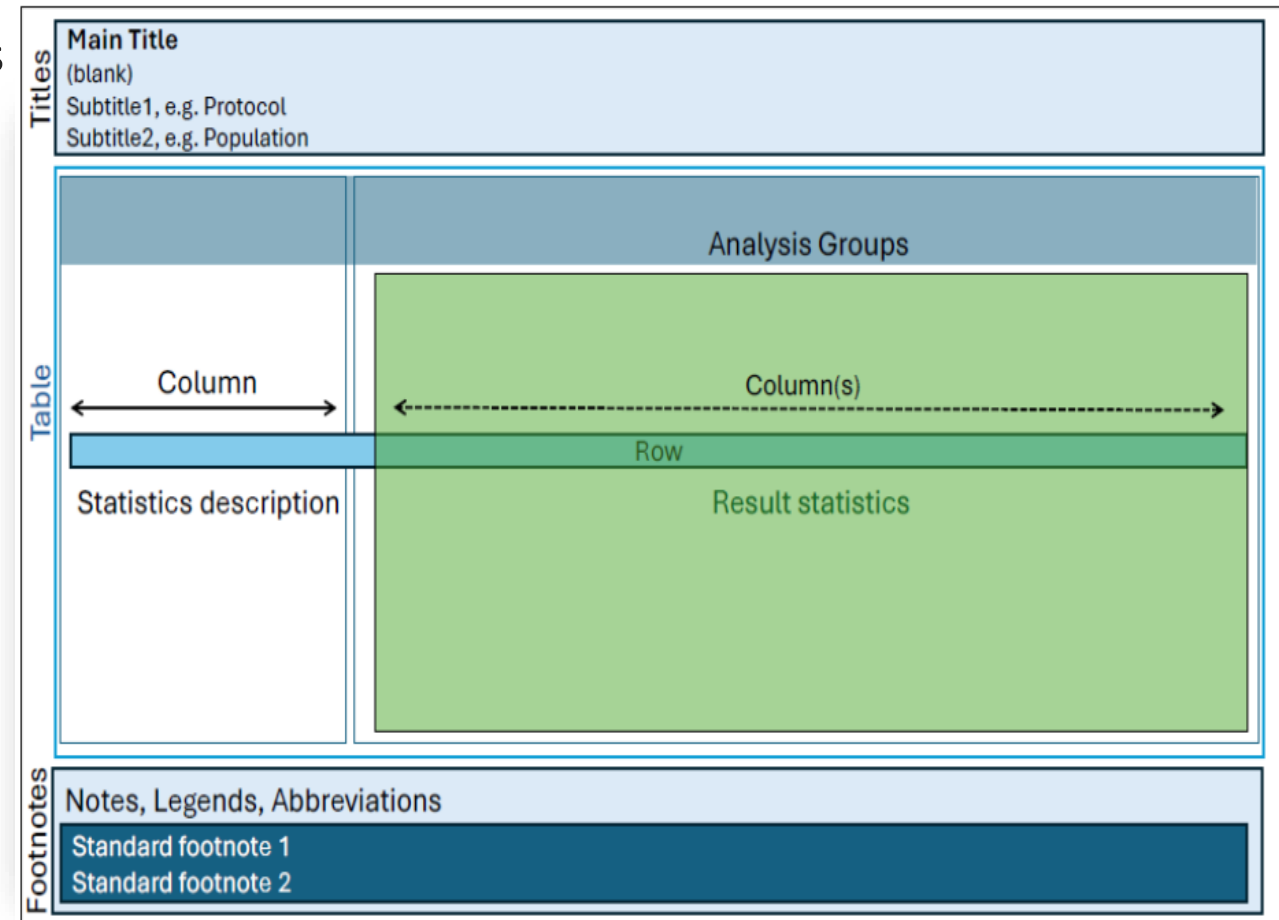
•Footnotes:

- ≤7 custom
- + 2 standard lines (date/program name)

•Indentation: Must be consistent (spaces OR tabs)

Gender	
Male	
Female	
Child-bearing potential?	
Yes	
No	

•Formatting: White background, black text



RTF file ODS RTF SAS

Prerequisites: ADaM – Unique

- ✓ Generated automatically during ADaM programming (via IDDI ADaM macro)
- ✓ Already used in ID-Spectackle to maintain study specifications
- ✓ Converted into JSON files

```
"Metadata": {  
  "Study",  
  {  
    "STUDYNAME": "SAMPLE-STUDY-001"  
  },  
  "Datasets",  
  {  
    "DATASET": "ADSL",  
    "DESCRIPTION": "Subject-Level Analysis Dataset"  
  },  
  "Variables",  
  {  
    "ORDER": 1,  
    "DATASET": "ADSL",  
    "VARIABLE": "STUDYID",  
    "LABEL": "Study Identifier",  
    "DATA_TYPE": "text",  
    "CODELIST": ""  
  },  
  {  
    "ORDER": 2,  
    "DATASET": "ADSL",  
    "VARIABLE": "USUBJID",  
    "LABEL": "Unique Subject Identifier",  
    "DATA_TYPE": "text",  
    "CODELIST": ""  
  },  
  {  
    "ORDER": 3,  
    "DATASET": "ADSL",  
    "VARIABLE": "SUBJID",  
    "LABEL": "Subject Identifier for the Study",  
    "DATA_TYPE": "text",  
    "CODELIST": ""  
  }  
}
```

```
"ADDITIONAL": {  
  {  
    "ID": "ACN",  
    "NAME": "Action Taken with Study Treatment",  
    "DATA_TYPE": "text",  
    "ORDER": 5,  
    "TERM": "DOSE NOT CHANGED",  
    "DECODED_VALUE": ""  
  },  
  {  
    "ID": "ACN",  
    "NAME": "Action Taken with Study Treatment",  
    "DATA_TYPE": "text",  
    "ORDER": 6,  
    "TERM": "DOSE REDUCED",  
    "DECODED_VALUE": ""  
  },  
  {  
    "ID": "ACN",  
    "NAME": "Action Taken with Study Treatment",  
    "DATA_TYPE": "text",  
    "ORDER": 3,  
    "TERM": "DRUG INTERRUPTED",  
    "DECODED_VALUE": ""  
  }  
}
```

Reusable across studies?

Prerequisites: IDDI SAS Macros - Standard

- Descriptive stats: univ-tab.sas
- Categorical stats: freq-tab.sas
- Categorical (occurrence): event-freqs.sas

TABLE 11: OVERVIEW OF ADVERSE EVENTS

Study: XXX
Population: Safety Analysis Population

	A N=xxx	B N=xxx	All N=xxx
Number of patients who experienced at least one			
Adverse Event	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
AE suspected to be caused by Atezolizumab/Placebo	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
AE suspected to be caused by Trastuzumab Emtansine	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
AE suspected to be caused by Trastuzumab	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
AE Leading to discontinuation of Atezolizumab/Placebo	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
AE Leading to discontinuation of Trastuzumab Emtansine	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
AE Leading to discontinuation of Trastuzumab	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
AE Leading to interruption of Atezolizumab/Placebo	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
AE Leading to interruption of Trastuzumab Emtansine	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
AE Leading to interruption of Trastuzumab	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
AE Leading to dose reduction of Trastuzumab Emtansine	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
NCI CTC Grade >=3 AE	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
Serious AE	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
Life-threatening AE	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
AE Leading to hospitalization	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
AE leading to a disability/incapacity	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
Congenital anomaly/birth defect in offspring of study subject	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
Fatal AE	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)

TABLE 06: DEMOGRAPHICS AND BASELINE CHARACTERISTICS

Study: XXX
Population: ITT Population

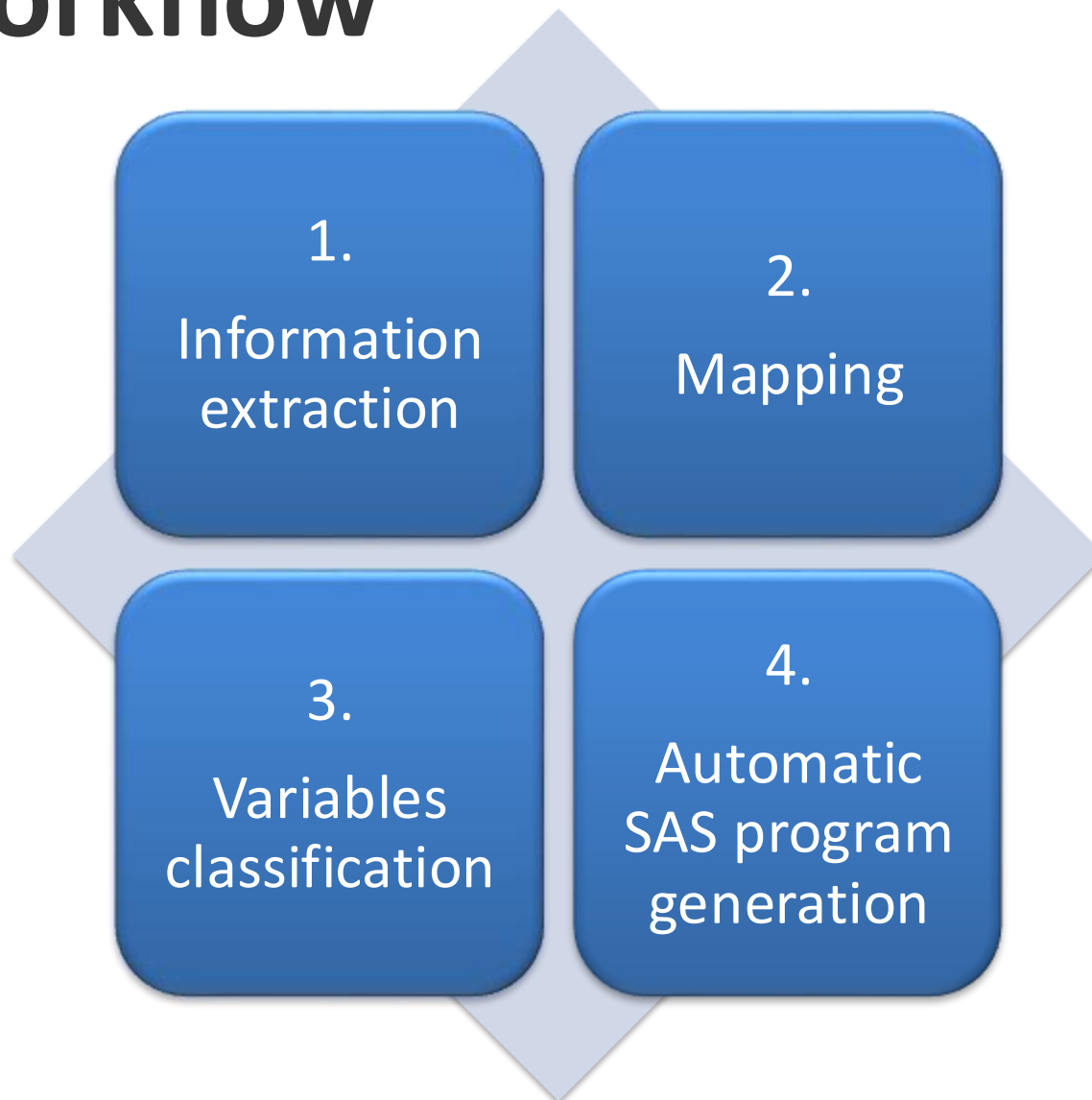
	A N=xxx	B N=xxx	All N=xxx
Age [years]			
n	xxx	xxx	xxx
Mean	xx.x	xx.x	xx.x
SD	xx.xx	xx.xx	xx.xx
Median	xx.x	xx.x	xx.x
Min-Max	xx.x - xx.x	xx.x - xx.x	xx.x - xx.x
Age [years]			
n	xxx	xxx	xxx
< 65	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
65-74	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
>=75	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
Gender			
n	xxx	xxx	xxx
Male	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
Female	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
Race			
n	xxx	xxx	xxx
American Indian or Alaska Native	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
Asian	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
Black or African American	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
Native Hawaiian or Other Pacific Islander	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
White	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
Unknown	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
Not reported	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)
Multiple	xxx (xx.x%)	xxx (xx.x%)	xxx (xx.x%)

Percentages are based on N in the column heading.

Output: Txx_xxx.txt (Production date: DDMMYYYY)
SAS Program Name: Txx_xxx.sas

Page x of x
Data Extraction Date: DDMMYYYY

Project workflow



1. Information Extraction from Mock Tables



- **Rule-based Python module**

- **Linear pipeline**

- Scans document from start to finish

- **State machine**

- to classify elements:
 - Titles & footnotes
 - Table metadata (variables, labels...)
 - Table contents (statistics)

- **AI support**

- for linguistic & syntax corrections

Document
traversal and
metadata
extraction

Table structure
parsing

Text cleaning
and
normalization

AI-assisted
post-processing

JSON Output
generation

Table Processing & Hierarchical Parsing

- Regex detection: Identify tables titles

- Save metadata:

- Titles
- Subtitles
- Notes

- Indentation-based hierarchy:

- Non-indented row → New parent variable
- Indented row → Sub-category

- Statistic heuristic:

- “%” → categorical
- Mean/Median → continuous
- Default = categorical

⚠ Parsing depends on consistent formatting & indentation

1. Mock Outputs

1.1 Tables

Table 1.01: Summary of Treatment Emergent Adverse Events (TEAEs)

Protocol:

Analysis Population: Safety Population

Number of Subjects	Dummy A (N=xxx)	Dummy B (N=xxx)	Dummy C (N=xxx)	Total (N=xxx)
Subjects with at least one TEAE	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
All causalities	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Non-ocular	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Ocular - Study eye	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Ocular - Fellow eye	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
TEAE related to study medication	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Non-ocular	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Ocular - Study eye	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Ocular - Fellow eye	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
TEAE related to IVT procedure	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Non-ocular	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Ocular - Study eye	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Ocular - Fellow eye	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Serious TEAE	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Non-ocular	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Ocular - Study eye	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)
Ocular - Fellow eye	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)	xx (xx.x%)

Note: For the dummy treatment groups, the assignments are random and have no bearing on actual treatment assignment. Percentages are based N.
Treatment-emergent Adverse Events (TEAEs) are defined as AEs occurring on or after the first dose of study treatment.

Produced: ddmmyyyy
Program: <programname>.sas

Date of data Extraction: ddmmyyyy

Generative AI & Post-Processing

Why?	How?	What it does?
The parser gave a good base , but needed contextual refinement	- <i>Google Gemini</i> used to correct and enhance JSON output - Uses strict prompts + locked JSON schema for deterministic results - Combines zero-/few-shot prompting depending on complexity	- Refines structure and text - Detects repeated tables or logic (e.g., “Similar to”, “Repeat for each”) - Ensures consistent, machine-readable output

Output

Table 02.02 Baseline Disease Characteristics

Protocol:
Population: Full analysis set

	Trt A N=xxx	Trt B N=xxx	All N=xxx
Baseline hs-CRP (mg/L)	n=xxx	n=xxx	n=xxx
Mean	xx.x	xx.x	xx.x
SD	xx.xx	xx.xx	xx.xx
Median	xx.x	xx.x	xx.x
Q1-Q3	xx - xx	xx - xx	xx - xx
Min-Max	xx - xx	xx - xx	xx - xx
Baseline IL-6 (pg/mL)	n=xxx	n=xxx	n=xxx
Mean	xx.x	xx.x	xx.x
SD	xx.xx	xx.xx	xx.xx
Median	xx.x	xx.x	xx.x
Q1-Q3	xx - xx	xx - xx	xx - xx
Min-Max	xx - xx	xx - xx	xx - xx

hs-CRP = high-sensitivity C-reactive protein; IL-6 = interleukin 6.

Baseline value is defined as the eligible measurement from randomization visit if before or at the date of first dose, or the measurement from screening if randomization value is not available.

For UACR, eGFR and hs-CRP, baseline is defined as the mean of the two assessments from the randomisation and the screening visits.

Produced: ddmmmyyyy
Program: <programname>.sas

Data Extraction Date: ddmmmyyyy

Page xx of xx

```
{
  "table_id": "TABLE 02.02",
  "table_title": "Baseline Disease Characteristics",
  "header": [{"", "", "", ""}, [{"", "Trt A\nN=xxx", "Trt B\nN=xxx", "All\nN=xxx"}]],
  "variable": [
    {
      "name": "Baseline hs-CRP (mg/L)",
      "content": [{"name": "Mean"}, {"name": "SD"}, {"name": "Median"},
        {"name": "Q1-Q3"}, {"name": "Min-Max"}],
      "categorical": false
    },
    {
      "name": "Baseline IL-6 (pg/mL)",
      "content": [{"name": "Mean"}, {"name": "SD"}, {"name": "Median"},
        {"name": "Q1-Q3"}, {"name": "Min-Max"}],
      "categorical": false
    }
  ],
  "notes": "hs-CRP = high-sensitivity C-reactive protein; IL-6 = interleukin 6.
  Baseline value is defined as the eligible measurement from randomization
  visit if before or at the date of first dose, or the measurement from screening if
  randomization value is not available. For UACR, eGFR and hs-CRP, baseline is defined
  as the mean of the two assessments from the randomisation and the screening visits.

  Produced: ddmmmyyyy Data Extraction Date: ddmmmyyyy
  Page xx of xx
  Program: <programname>.sas
}
```

Mapping

Challenge: No large labeled datasets → unsupervised/ zero-shot methods

LLM (Gemini)

- Direct prompting
- Hallucinations

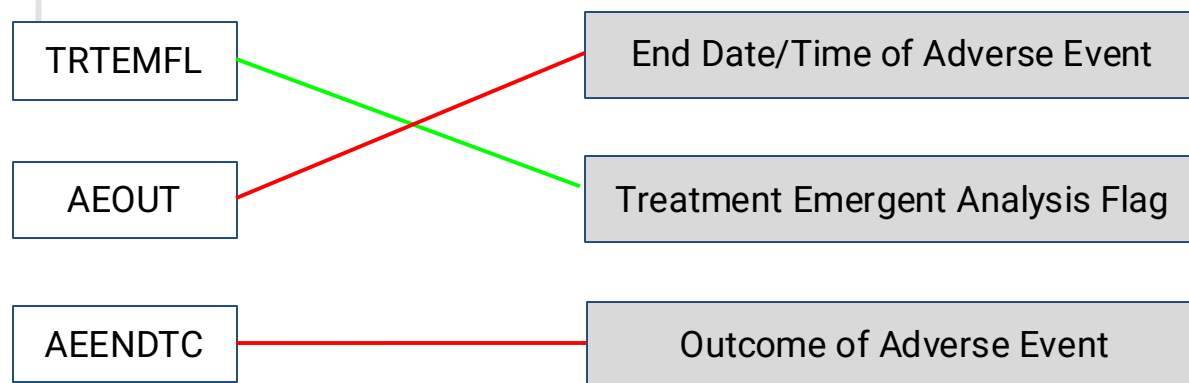
Fuzzy string matching

- Lexical only
- Fragile (synonyms, similar phrasing)

Sentence embeddings (transformer)

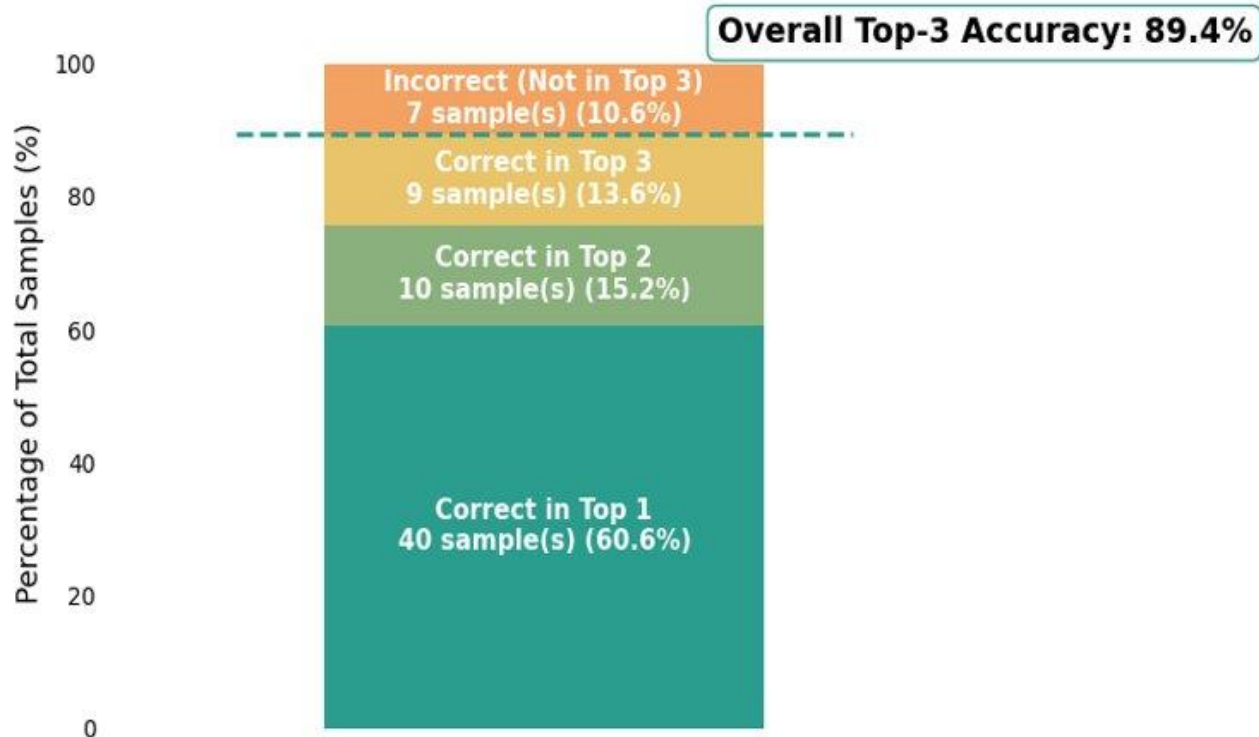
- Semantic similarity
- Best results
- But sensitive to long text

Mapping: Sentence embeddings, results



$$\text{Accuracy} = \frac{\text{number of correct connections}}{\text{number of connections}}$$

Overall Algorithm Performance Breakdown



Variable classification

Default

Mean /
Median

%

freq-tab.sas

univ-tab.sas

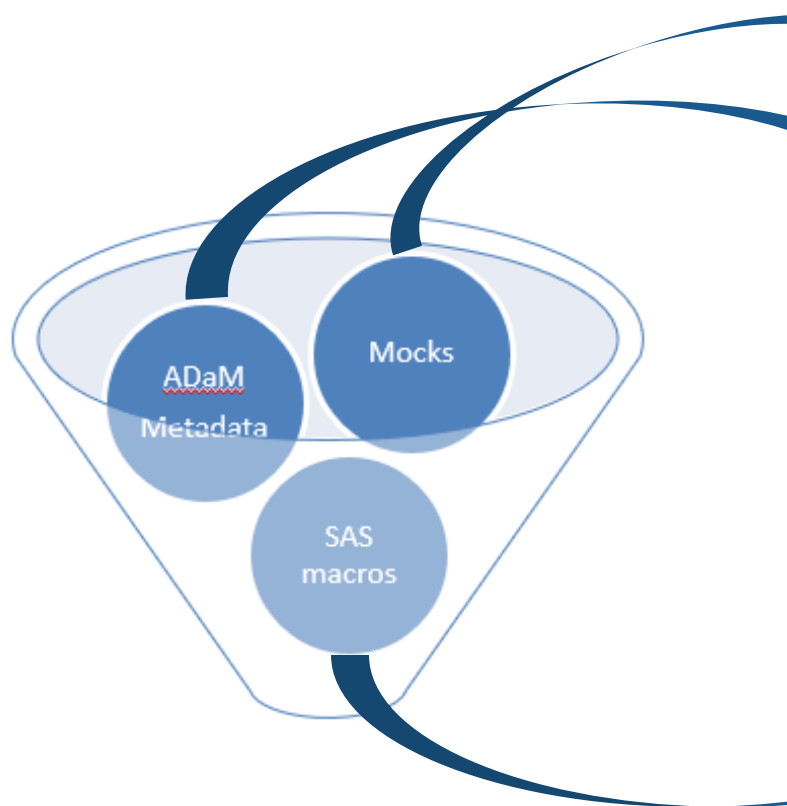
event-freqs.sas

Automatic SAS program generation

- 1 Program per mock table
- Pre-complete macro calls
- **Ready-to-edit SAS program**



IDD'AI Tool



This app will help you build the code needed for the construction of mocks

☐ Perform cleaning with Gemini ✨ ?

Upload a docx file and a json file to start the process

Upload your mock tables here as a .docx file



Drag and drop file here

Limit 200MB per file • DOCX

Browse files

Upload your metadata here as a .json file



Drag and drop file here

Limit 200MB per file • JSON

Browse files

Upload a docx file and a json file to start the process.

Upload an alternative sas code template here as a .j2 file



Drag and drop file here

Limit 200MB per file • J2

Browse files

Using default template. 📄

IDD'AI Tool

TABLE 02.01

Table Title

Demographic and Baseline Characteristics

Table Population

Full analysis set

Variables

Variable Name

Sex

Delete Variable

Variable Type

CATEGORICAL

Content

Male

Female

<Missing>

+ Add Content

Variable Name

Is the subject of childbearing potential?

Delete Variable

Variable Type

CATEGORICAL

IDD'AI Tool



Matching Results Verifier

For each row, select the correct match. If none are right, type the correct value in the 'Other' column.

Select one or more table:

Choose options

Validate Matches

Mock Parameter	Table title	Top 1 Match Display	Top 2 Match Display	Top 3 Match Display	Select Match	Other (Searchable)	PARA
Sex	Demographic and Baseline Characteristics	SEX	RACE	AGE	Top 1	None	None
Is the subject of childbearing potential?	Demographic and Baseline Characteristics	CHILDP0	SEX	HSRRISK	Top 1	None	None
Age (years)	Demographic and Baseline Characteristics	AGE	AGEU	ADT	Top 1	None	None
Subject Self-reported Race	Demographic and Baseline Characteristics	RACE	ETHNIC	ITTFL	Top 1	None	None
Subject Self-reported Ethnicity	Demographic and Baseline Characteristics	ETHNIC	RACE	ITTFL	Top 1	None	None
Smoking status	Demographic and Baseline Characteristics	SMOKER	LBFAST	CMSTAT	Top 1	None	None
BMI (kg/m2)	Demographic and Baseline Characteristics	AVAL PARAMCD_ADVS:BMI	AVAL PARAMCD_ADVS:WEIGHT	AVAL PARAMCD_ADVS:HEIGHT	Top 1	None	None
Pulse Rate (beats/min)	Demographic and Baseline Characteristics	AVAL PARAMCD_ADVS:PULSE	AVAL PARAMCD_ADVS:PULSEAVG	ARRCLTA	Top 1	None	None
Systolic Blood Pressure (mmHg)	Demographic and Baseline Characteristics	AVAL PARAMCD_ADVS:SYSBP	AVAL PARAMCD_ADVS:SYSBPAVG	AVAL PARAMCD_ADVS:DIABP	Top 1	None	None
Diastolic Blood Pressure (mmHg)	Demographic and Baseline Characteristics	AVAL PARAMCD_ADVS:DIABP	AVAL PARAMCD_ADVS:DIABPAVG	AVAL PARAMCD_ADVS:SYSBP	Top 1	None	None
<Lipids (units)>	Demographic and Baseline Characteristics	AVAL PARAMCD_ADLB:CHOL	AVAL PARAMCD_ADLB:TRIG	LBORRES	Top 1	None	None

Submit

Download as ZIP

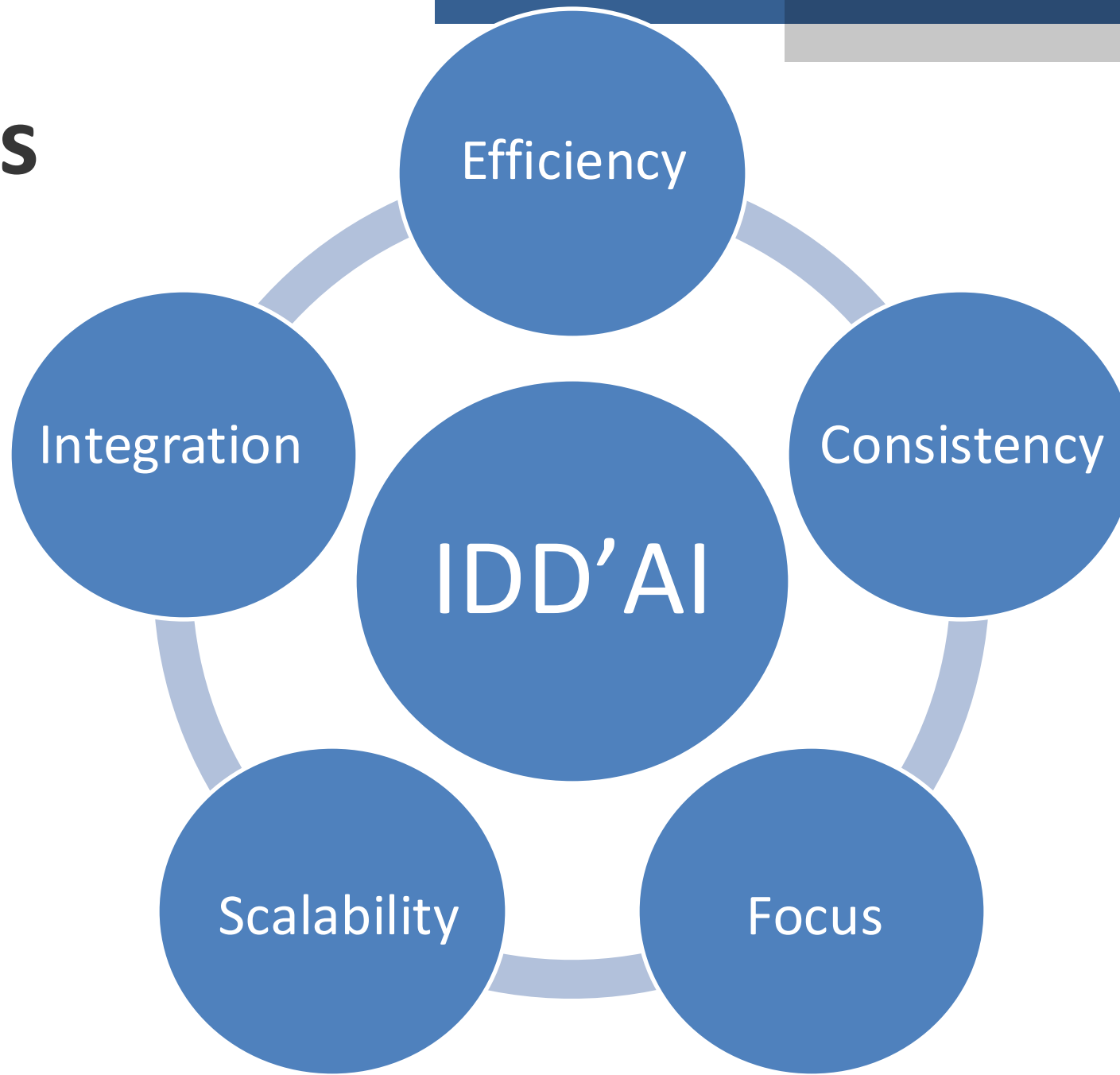
IDD'AI Tool



“Generated Code”

1 Table = 1 SAS program

Benefits



Conclusion and next steps

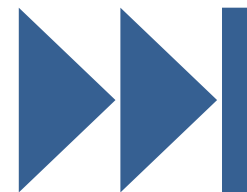


Limitations:

Manual intervention still needed (flags, data transformations)

Basic programs, not fully optimized

Standard mocks library needed



Future work:

Build standardized mock library aligned with macros

Leverage CDISC ARM/ARS standards (SAP & ToC parsing)

Develop more robust, scalable framework

Efficacy, Figures and listings



Thank you

USA



info@iddi.com



7751 Brier Creek Parkway, Suite 204
Raleigh, NC 27617



+1 984 227 8599

BELGIUM



info@iddi.com



Avenue Provinciale, 30
1341 Louvain-la-Neuve



+32 10 61 44 44



www.IDDI.com