AI-powered unstructured data analysis to accelerate Biometrics activities

Hao Chen
Yiwen Wang
Kevin Huang

phuse

Agenda

- Background
- Business Problem
- New Perspective
- Case Sharing
- Summary

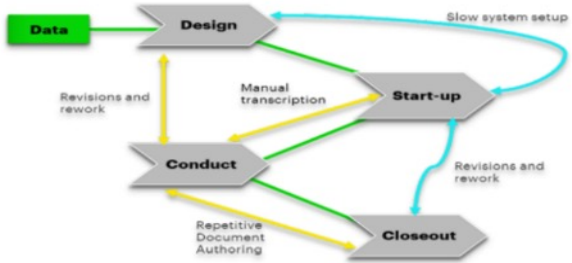
Background

TRANSITIONING TO FUTURE OF CLINICAL TRIALS

Study published by the Massachusetts Institute of Technology (MIT) has found that only 13.8% of drugs successfully pass clinical trials company can expect to pay between $161 million to $2 billion for any drug to complete the entire clinical trials. This Artificial Intelligence has the potential to Significantly reduce clinical cycle time and cost while helping to improve the outcome of clinical development.

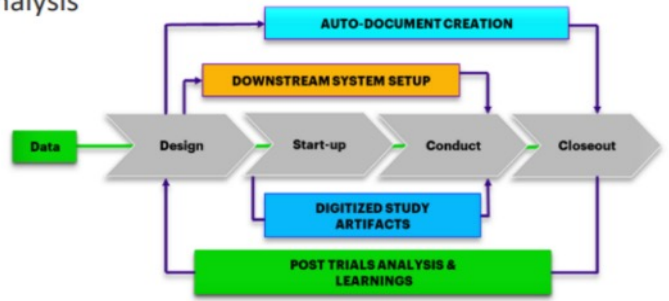
Traditional clinical trail data flow

A Non-linear maze of manual effort, rework Inefficiency



The future state vision

Streamline interoperability of trial artifacts enabled through the standardization of data/vocabulary learnings from AI & ML analysis



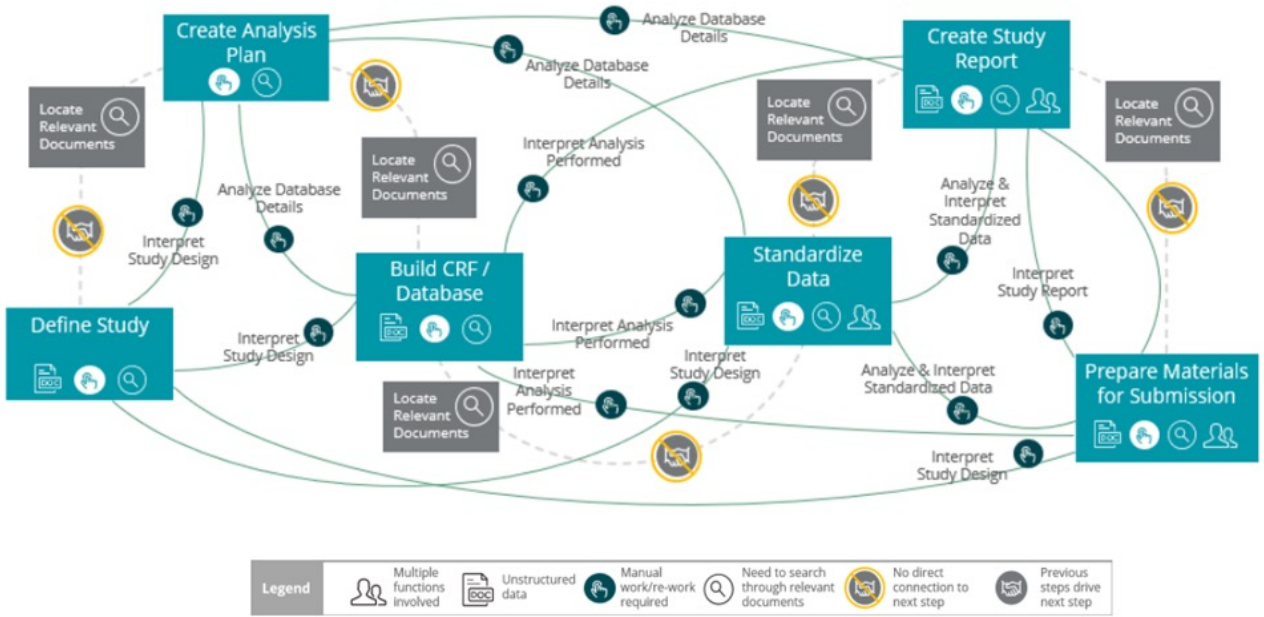
FUTURE STATE BENEFITS

- **Reduced manual** Effort & rework.
- **Increased interoperability** Across Multiple systems
- **Improved Trial** Cycle time.
- **Optimized Use of Insights** To Improve the Next Trial
- **Reduced Errors in** Trial artifacts.

phuse + accenture

Business Problem

Clinical Data Flow : Current State



Statistician

Programming

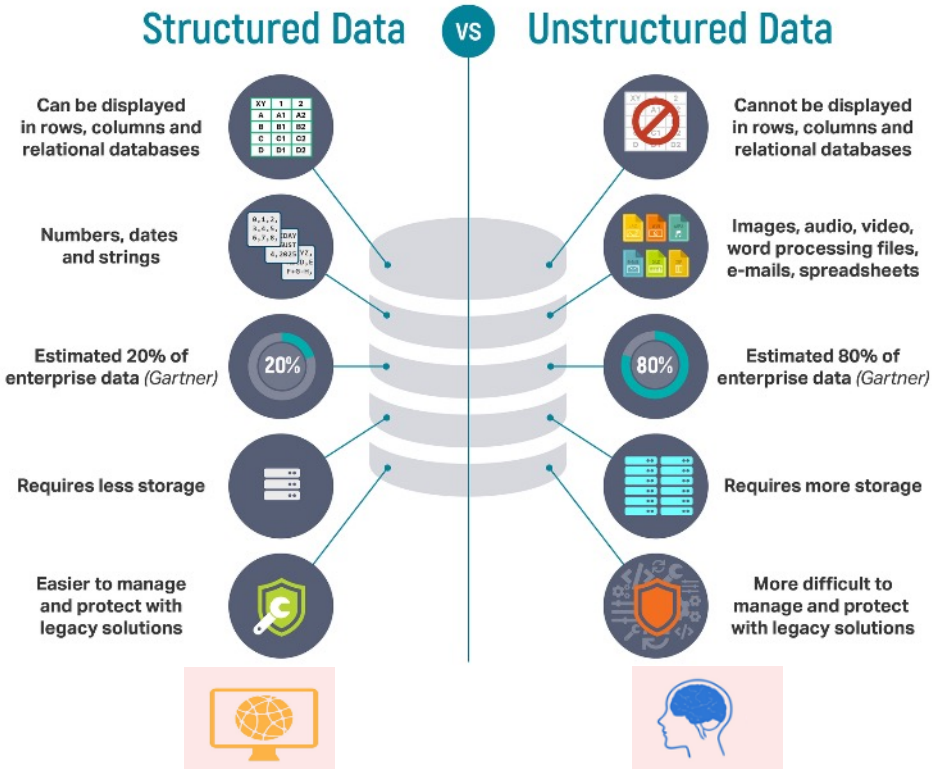
Data Management

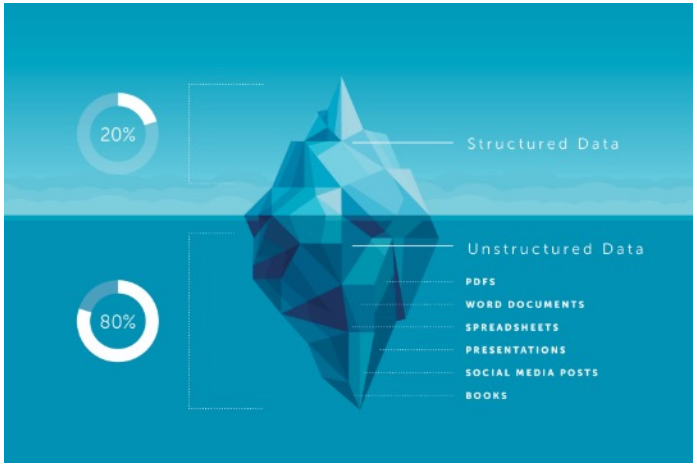
Medical Communication

[ConvergeHEALTH | Using AI to automate clinical data, from protocol to submission]

 [Using AI to drive automation for clinical data, from protocol to submission]

Structured Data vs Unstructured Data





https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

Problem examples

1) Output Table analysis

File Format/Table Layout varies across different TAs/Vendors, how to extract the data?

(.rtf/.doc/.docx)

Table 3. Patient Screening and Enrollment, Trials A and B

Disposition	Trial A	Trial B
Patients screened	XX	XX
Screening failures	XXX (XX.X%)	XXX (XX.X%)
Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
Other	XXX (XX.X%)	XXX (XX.X%)
Patients enrolled	XXX (XX.X%)	XXX (XX.X%)
Patients randomized	XXX (XX.X%)	XXX (XX.X%)

Table 3. Patient Screening and Enrollme... A and B

Disposition	Trial A	Trial B
Patients screened	XX	XX
Screening failures	XXX (XX.X%)	XXX (XX.X%)
Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
Other	XXX (XX.X%)	XXX (XX.X%)
Patients enrolled	XXX (XX.X%)	XXX (XX.X%)
Patients randomized	XXX (XX.X%)	XXX (XX.X%)

Table 3. Patient Screening and Enrollment, Trial...

Disposition	Trial A	Trial B
Patients screened	XX	XX
Screening failures	XXX (XX.X%)	XXX (XX.X%)
Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
Other	XXX (XX.X%)	XXX (XX.X%)
Patients enrolled	XXX (XX.X%)	XXX (XX.X%)
Patients randomized	XXX (XX.X%)	XXX (XX.X%)

?

	A	B	C	D	E
1	rowNum	Lev0	Lev1	V2	V3
2	1	Disposition		Trial A	Trial B
3	2	Patients screened		XX	XX
4	3	Screening failures		XXX (XX.X%)	XXX (XX.X%)
5	4	Screening failures	Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
6	5	Screening failures	Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
7	6	Screening failures	Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
8	7	Screening failures	Other	XXX (XX.X%)	XXX (XX.X%)
9	8	Patients enrolled		XXX (XX.X%)	XXX (XX.X%)
10	9	Patients randomized		XXX (XX.X%)	XXX (XX.X%)

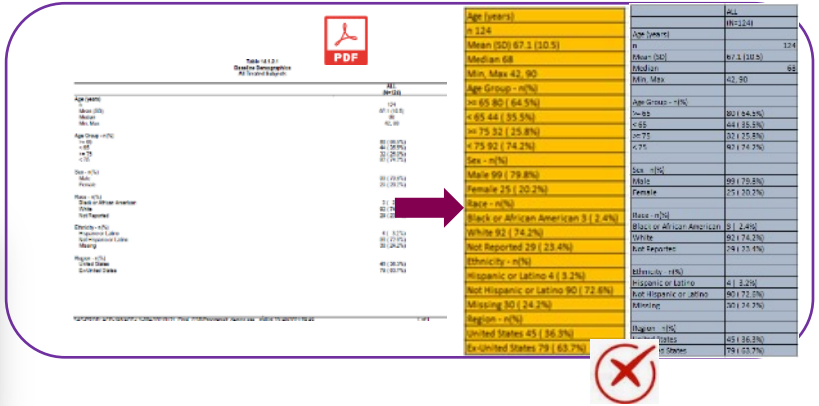
Sheet 1

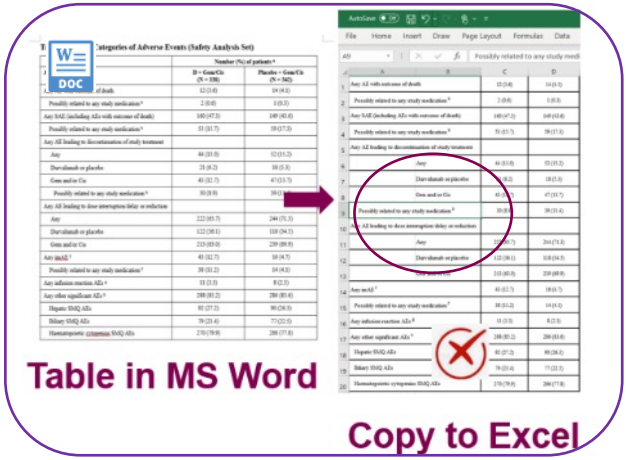
https://www.regulations.gov/document/FDA-2022-N-1961-0002

Problem examples

1) Output Table analysis

Option A: Manual Copy & Paste





Option B: Ad-hoc programming - rtf/xml/txt parser

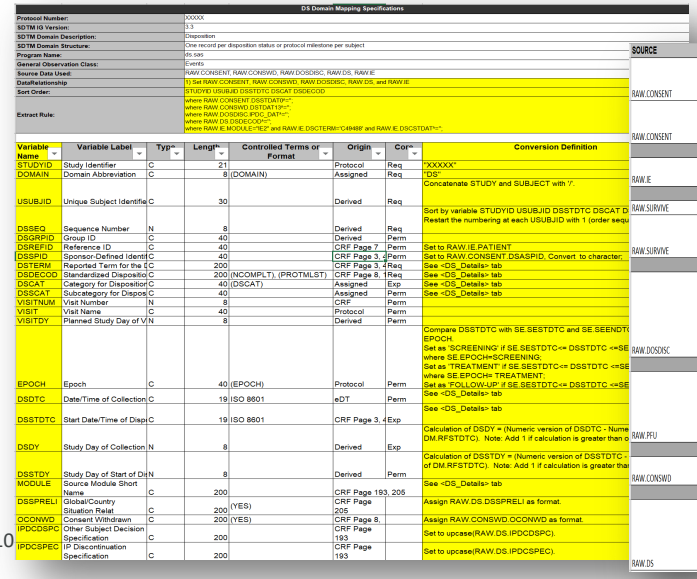
Option C: Reproduce the output via SAS programming

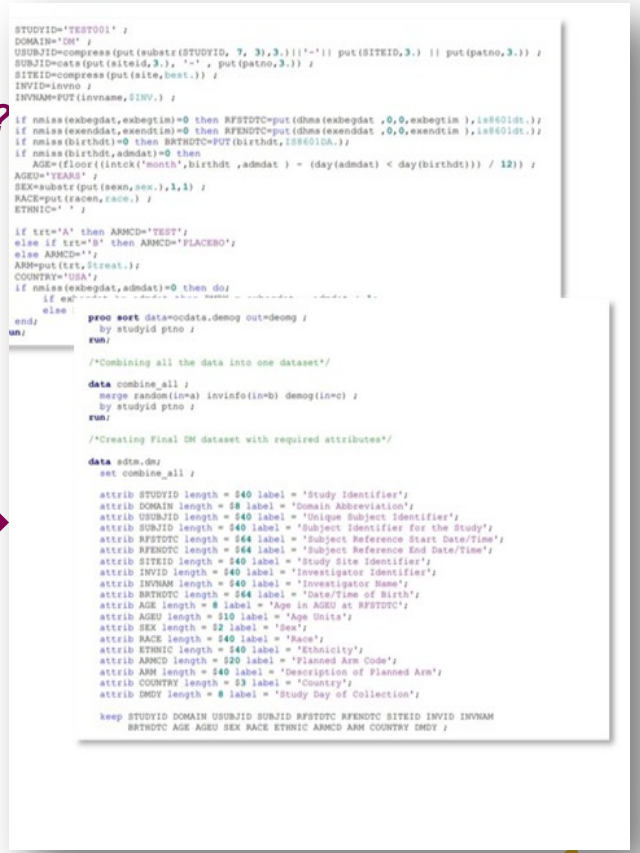
Problem examples

2) Text Spec to Code Translation

How to translate the Text Spec into codes for SDTM Code Generation?

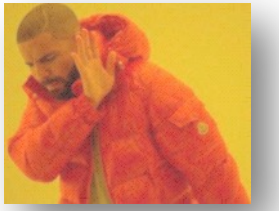
LBORRES = For RAW.LB records:
If RAW.LB.MODULE='LB3' and RAW.LB.LBVDIP not missing then set to upcase(vvalue(RAW.LB.LBVDIP)), else if RAW.LB.MODULE='LB3' and RAW.LB.LBORRES not missing then set to RAW.LB.LBORRES, else if RAW.LB.MODULE='LB4' and RAW.LB.LBVDIP not missing then set to upcase(vvalue(RAW.LB.LBVDIP)), else if RAW.LB.MODULE='LB4' and RAW.LB.LBORRES not missing then set to RAW.LB.LBORRES, else if RAW.LB.MODULE='LB6' and RAW.LB.LBVDIP not missing then set to upcase(vvalue(RAW.LB.LBVDIP)), else if RAW.LB.MODULE='LB6' and RAW.LB.LBORRES not missing then set to RAW.LB.LBORRES, else set to RAW.LB.LBORRES.





Problem examples

Manual coding?



2) Text Spec to Code Translation

```
/*************************************************************************;
*  Program name      : AE.sas
*  Type              : SDTM*
*  Purpose           : Creation of sdtm.AE and sdtm.suppAE*
*  Author            : xxxx                 *
*  Datetime created  : 07/13/2022, 12:53:15 *
*  Input datasets    : [NOTE: Complete as needed from the mapping specification]
*  Extract Rule      : [where RAW.AE.AEANY="1"]
*  Macros used       : %setup, %localsetup*
*  Output files      : sdtm.AE, sdtm.suppAE*
*  Revision history  :
*
*  Date        Author      Ref    Revision
*************************************************************************/
```

Program header

```
*************************************************************************;
**  Setup calls                                                        **;
*************************************************************************;
%autoexec;

%let sdtm_domain = AE;

%let AESort_Order = STUDYID USUBJID AESTDTC AEDECOD;

%let kname = STUDYID DOMAIN USUBJID AESEQ AESPID AETERM AELLT AELLTCD AEDECOD AEPTCD AEHLT AEHLTC
    AESDTH AESHOSP AESLIFE AESMIE AECONTRT EPOCH AESTDTC AEENDTC AESTDY AEENDY;

%preprocess(domain=AE);

/*-------------------RAW.AE-------------------*/
data RAW_AE;

 attrib
    STUDYID    label='Study Identifier'                          length=$21
 ...
    AEREL      label='Causality'                                 length=$200
    AEOUT      label='Outcome of Adverse Event'                  length=$200
 ;
 set
    RAW.AE(rename=(AESER = _AESER ) where=(AEANY="1"));

    /* Set to "AE-"||left(trim(put(RAW.AE.AENO,best.))) if RAW.AE.AENO is not missing. */
    if ^missing(AENO) then do;AESPID      = "AE-"||left(trim(put(AENO,best.)));end;

    /* Set to RAW.AE.SOC_CODE  */
    AESOCCD = SOC_CODE;

    /* Assign RAW.AE.AESEVMAX as format.  */
    AESEV=
            ifc(AESEVMAX = "1",>"MILD",
            ifc(AESEVMAX = "2",>"MODERATE",
            ifc(AESEVMAX = "3",>"SEVERE",
                AESEVMAX)))      ;
    /* Assign RAW.AE.AESER as format.  */
```

System macro variable

Attributes, good format

Rename, Where

Codelist

```
/*==================temp==================*/
Data temp;
 merge  RAW_AE(in=a)  RAW_SERAE;
 by SUBJECT AENO;
 if a
 ;
 run;
 ;

data temp;

 attrib
    STUDYID    label='Study Identifier'                          length=$21
    DOMAIN     label='Domain Abbreviation'                       length=$2
    .....
 /* Compare VSDTC with SE.SESTDTC and SE.SEENDTC per subject per EPOCH. Set as 'SCREENING' if SE.SESTDTC<= QSDTC <=SE
 <=SE.SEENDTC where SE.EPOCH= TREATMENT; Set as 'FOLLOW-UP' if SE.SESTDTC<= QSDTC <=SE.SEENDTC where SE.EPOCH= FOLLOW
 if SESTDTC and SEENDTC include time part, then we need to compare the time, if no, only to compare date part. */
 /* Variable EPOCH tod - CANNOT PARSE
 EPOCH=Compare VSDTC with SE.SESTDTC and SE.SEENDTC per subject per EPOCH.
 Set as 'SCREENING' if SE.SESTDTC<= QSDTC <=SE.SEENDTC where SE.EPOCH=SCREENING; Set as 'TREATMENT' if SE.SESTDTC<= Q
 /* Variable EPOCH tod */;
 ;
    /*'      00008".*/
    STUDYID  =  "      :00008";

    /*."AE".*/
    DOMAIN   =  "AE";

    /*Concatenate STUDY and SUBJECT separating by '/'.                    */
    USUBJID  =  cat(strip(STUDY),'/',strip(SUBJECT));

    /*Calculation of AESTDY = (Numeric version of date part of AESTDTC - Numeric version of date part of DM.RFSTDTC).  Note1: Add 1 if cal
    then AESTDY is equal to missing. */
    %sdtm_dy(INVAR_DTC = AEstdtc, OUTVAR_DY = AEstdy);

    /*Calculation of AEENDY = (Numeric version of date part of AEENDTC - Numeric version of date part of DM.RFSTDTC).  Note1: Add 1 if cal
    then AEENDY is equal to missing. */
    %sdtm_dy(INVAR_DTC = AEendtc, OUTVAR_DY = AEendy);
 ;
 run;

    /*Compare AESTDTC with SE.SESTDTC and SE.SEENDTC per subject per EPOCH. Set as 'SCREENING' if SE.SESTDTC<= AESTDTC <=SE.SEENDTC where
    <=SE.SEENDTC where SE.EPOCH= TREATMENT; Set as 'FOLLOW-UP' if SE.SESTDTC<=AESTDTC <=SE.SEENDTC where SE.EPOCH= FOLLOW-UP; If date part
    Note: if SESTDTC and SEENDTC include time part, then we need to compare the time, if no, only to compare date part. */
    %sdtm_epoch (indata=final1, outdata=final2, domain=AE, debug=Y);
```

Todo placeholder

Variable level

Data level

```
    /*for source data RAW.CONSENT*/
 if not missing(DSSTDAT2) then do;
        DSSPID=strip(put(DSASPID1,best.));
        DSTERM="MAIN INFORMED CONSENT 2";
        DSDECOD ="INFORMED CONSENT 2 OBTAINED";
        DSCAT="PROTOCOL MILESTONE";
        DSDTC=VIS_DAT;
        DSSTDTC=DSSTDAT2;
        VISITDTC=VIS_DAT;
 output;
 end;
;
```

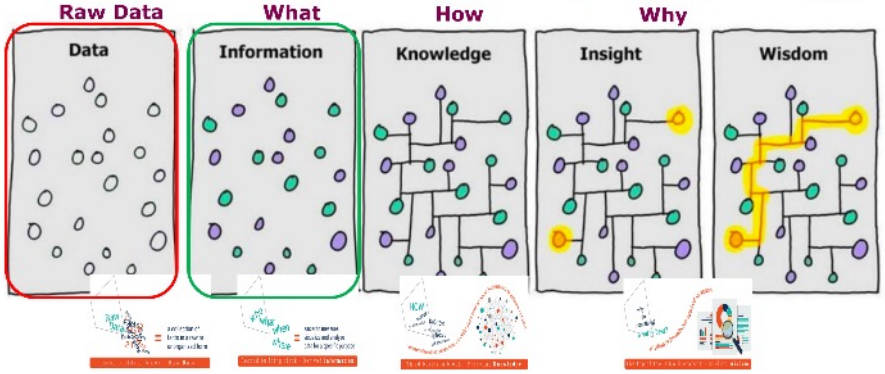
Data Join/Combine

Grouping process

New Perspective

From Unstructured Data to Structured Data – AI Solution

Human-driven to ***Software-driven***



Unstructured Data

The university has 5600 students. Shaun (ID Number: 160801), 18 years old Communication study. Linh with ID number 160802, majoring in Accounting and is 20 years old. Ahmed from Psychology study program, 19 years old, ID number 160803.

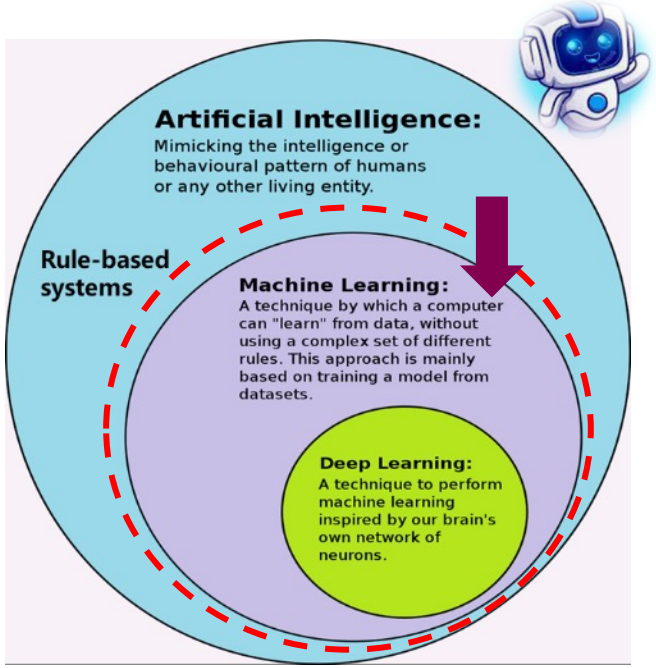
Semi-Structured Data

```
<University>
  <ID Number="160801">
    <Name="Shaun">
    <Age="18">
    <Program="Communication">
  <ID Number="160802">
    <Name="Linh">
    <Age="20">
    <Program="Accounting">
..... </University>
```

Structured Data

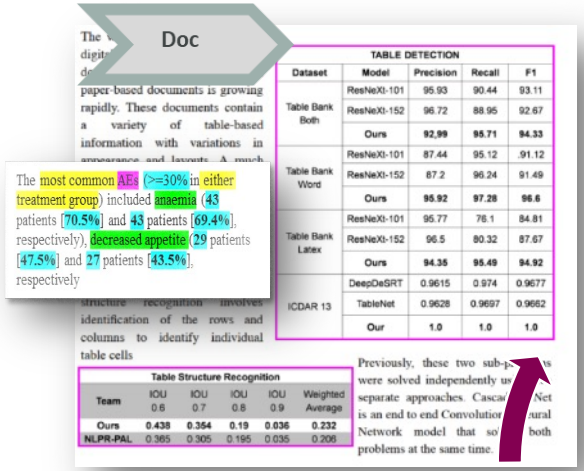
ID	Name	Age	Program
160801	Shaun	18	Communication
160802	Linh	20	Accounting
160803	Ahmed	19	Psychology

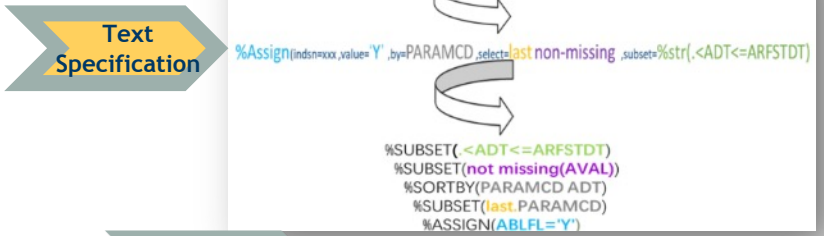
GLEEMATIC A.I.



How deep learning is a subset of machine learning and how machine learning is a subset of artificial intelligence (AI) https://en.wikipedia.org/wiki/Deep_learning

Digitalization for the Trial Artifacts





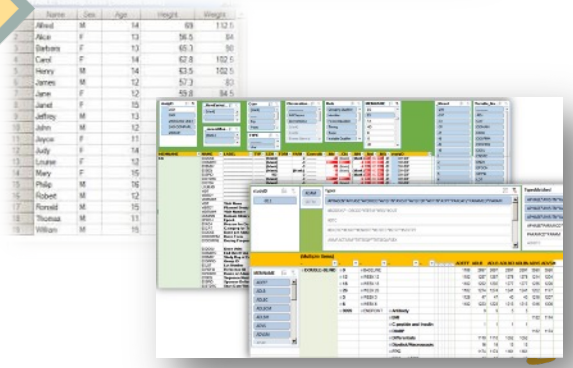
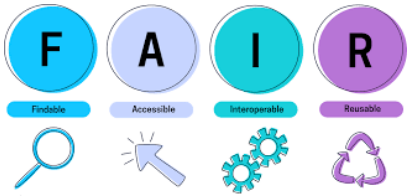


Table 3. Patient Screening and Enrollment, Trials A and B

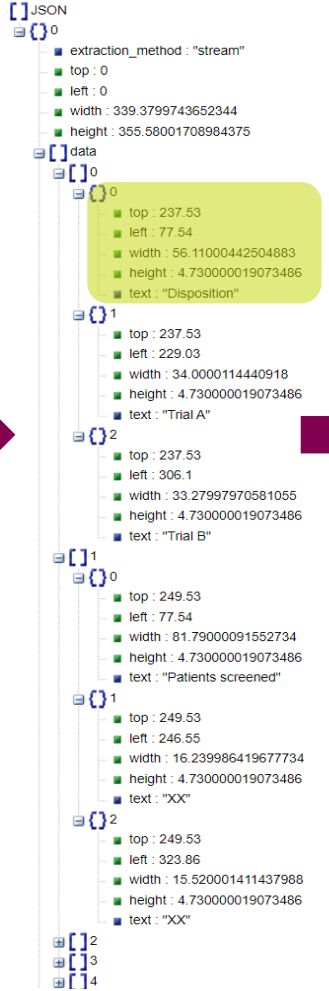
Disposition	Trial A	Trial B
Patients screened	XX	XX
Screening failures	XXX (XX.X%)	XXX (XX.X%)
Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
Other	XXX (XX.X%)	XXX (XX.X%)
Patients enrolled	XXX (XX.X%)	XXX (XX.X%)
Patients randomized	XXX (XX.X%)	XXX (XX.X%)

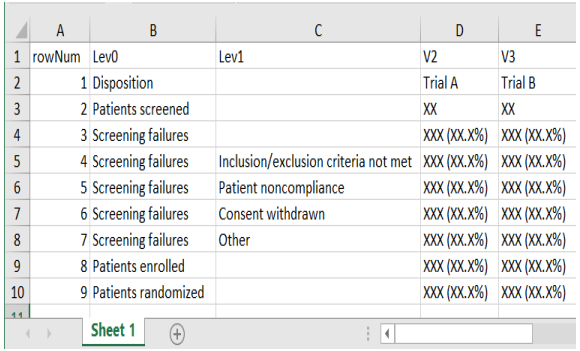


Solutions

1) Output Table analysis

Disposition	Trial A	Trial B
Patients screened	XX	XX
Screening failures	XXX (XX.X%)	XXX (XX.X%)
Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
Other	XXX (XX.X%)	XXX (XX.X%)
Patients enrolled	XXX (XX.X%)	XXX (XX.X%)
Patients randomized	XXX (XX.X%)	XXX (XX.X%)





	A	B	C	D	E
1	rowNum	Lev0	Lev1	V2	V3
2	1	Disposition		Trial A	Trial B
3	2	Patients screened		XX	XX
4	3	Screening failures		XXX (XX.X%)	XXX (XX.X%)
5	4	Screening failures	Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
6	5	Screening failures	Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
7	6	Screening failures	Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
8	7	Screening failures	Other	XXX (XX.X%)	XXX (XX.X%)
9	8	Patients enrolled		XXX (XX.X%)	XXX (XX.X%)
10	9	Patients randomized		XXX (XX.X%)	XXX (XX.X%)

Sheet 1

- Challenges
  - Rows/Columns
  - Nested Levels
  - Paragraphs/Pages
  - Dynamic table area
  - Broken lines
  - Dynamic spacing
  - …

Solutions

2) Text Spec to Code Translation

Text Spec (human language + code piece)

LBORRES = For RAW.LB records:
If RAW.LB.MODULE='LB3' and RAW.LB.LBVDIP not missing then set to upcase(vvalue(RAW.LB.LBVDIP)), else if RAW.LB.MODULE='LB3' and RAW.LB.LBORRES not missing then set to RAW.LB.LBORRES, else if RAW.LB.MODULE='LB4' and RAW.LB.LBVDIP not missing then set to upcase(vvalue(RAW.LB.LBVDIP)), else if RAW.LB.MODULE='LB4' and RAW.LB.LBORRES not missing then set to RAW.LB.LBORRES, else if RAW.LB.MODULE='LB6' and RAW.LB.LBVDIP not missing then set to upcase(vvalue(RAW.LB.LBVDIP)), else if RAW.LB.MODULE='LB6' and RAW.LB.LBORRES not missing then set to RAW.LB.LBORRES, else set to RAW.LB.LBORRES.

If RAW.LB.MODULE='LB3' FORMULA and RAW.LB.LBVDIP VAR not missing then set to upcase(vvalue(RAW.LB.LBVDIP)) FUNC , else if RAW.LB.MODULE='LB3' FORMULA and RAW.LB.LBORRES VAR not missing then set to RAW.LB.LBORRES VAR , else if RAW.LB.MODULE='LB4' FORMULA and RAW.LB.LBVDIP VAR not missing then set to upcase(vvalue(RAW.LB.LBVDIP)) FUNC , else if RAW.LB.MODULE='LB4' FORMULA and RAW.LB.LBORRES VAR not missing then set to RAW.LB.LBORRES VAR , else if RAW.LB.MODULE='LB6' FORMULA and RAW.LB.LBVDIP VAR not missing then set to upcase(vvalue(RAW.LB.LBVDIP)) FUNC , else if RAW.LB.MODULE='LB6' FORMULA and RAW.LB.LBORRES VAR not missing then set to RAW.LB.LBORRES VAR , else set to RAW.LB.LBORRES VAR .

```
LBORRES= ifc(MODULE='LB3' and ^missing(LBVDIP), upcase(vvalue(LBVDIP)),
         ifc(MODULE='LB3' and ^missing(LBORRES) ,   LBORRES,
         ifc(MODULE='LB4' and ^missing(LBVDIP),   upcase(vvalue(LBVDIP)),
         ifc(MODULE='LB4' and ^missing(LBORRES),   LBORRES,
         ifc(MODULE='LB6' and ^missing(LBVDIP) , upcase(vvalue(LBVDIP)),
         ifc(MODULE='LB6' and ^missing(LBORRES) , LBORRES,
          LBORRES )))))));
```

SAS programs (or R/Python/SQL)

- **Case Sharing**

1) Output Table Analysis

– Side by Side Table Generation

2) Text Spec to Code Translation

– SDTM Mapping Code Generation

Side by Side Table Generation

- Business needs

Needs for clinical document writing

- Side-by-side tables (Global vs Asia vs China populations) are needed for China submission

SAS programming not the best solution

- Additional resource/budget/time needed
- Start from Scratch/Format Adjustment

Manual creation not the best solution

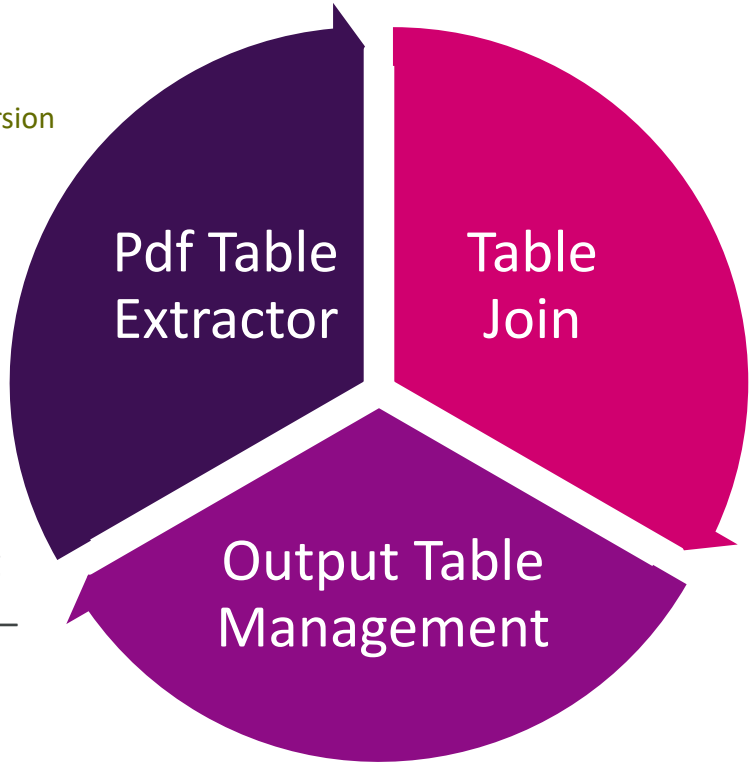
- Time consuming
- Pron to error
- Intensive QC required

Side by Side Table Generation

- Current key functions

  Extract data from TFL files:

  - Legacy TFLs without editable version
  - Inconsistent format of tables
  - Table across multiple pages



Join table:

- Join by auto-populated/flexible keys
- 2+ tables join

Benefits

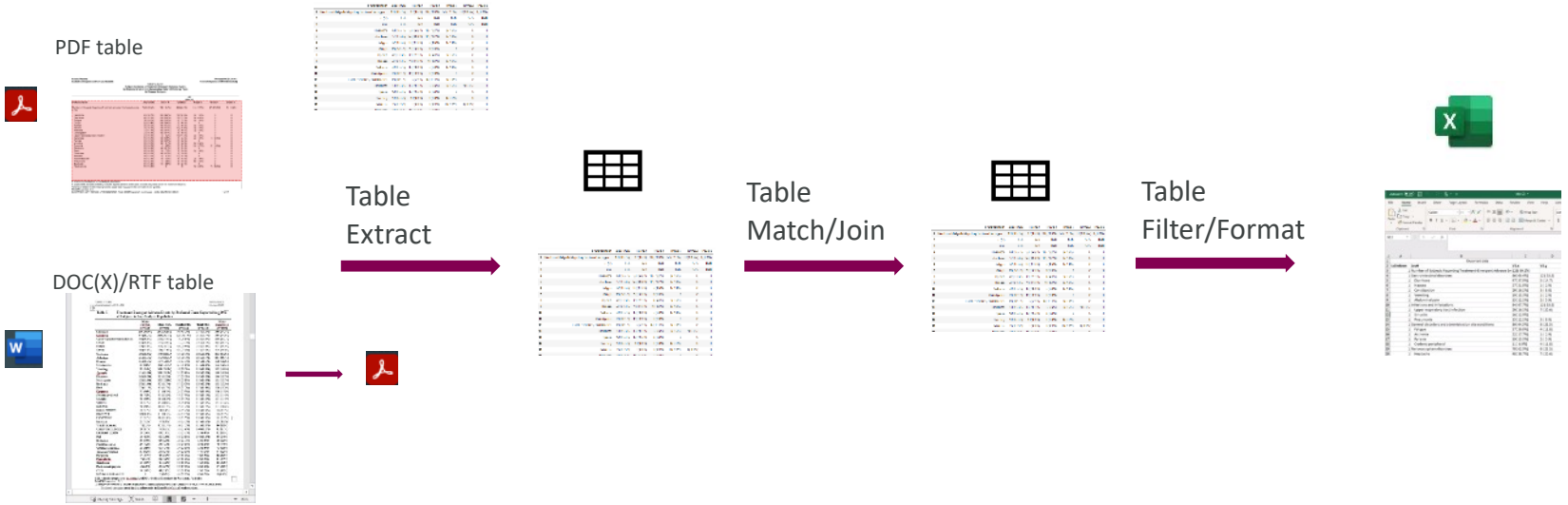
- 3-5 minutes/table
  - Spotted QC is still required
- Batch and Real-time processing
- Data extraction accuracy 100% – No OCR(Optical character recognition)

Manage the outlook of a table:

- Filter content or data
- Sort values
- Other calculations users may need

Side by Side Table Generation

- Process



PDF Table Extractor | Table Join | Output Table

Use this tool to extract structured, machine-readable tables from PDF reports in a few clicks. Load a PDF into the app and extract tables right in the browser, ready for export to Excel

Upload PDF file

	V1	V2	V3	V4	V5
1					

SDTM Mapping Code Generation

- Business needs

SDTM spec and programming codes consistency

- Updates should be taken in both files.

SDTM programming code not reusable

- Reference programming code should be checked line by line before implementation.

Manual creation not the best solution

- Time consuming
- Pron to error
- Intensive QC required for simple and standardized variable generation

SDTM Mapping Code Generation

- **Designed feature**
  - **Natural Language Spec -> Metadata -> code**
  - **Spec adaptive to different standard (across TA/vendor)**
  - **70/30% -> 80/20%**
  - **Standard algorithm accumulated**
  - **Rerun applicable with updated user-defined code**

SDTM Mapping Code Generation

- **Text Spec to Code**
  - **Minimum input from Spec**

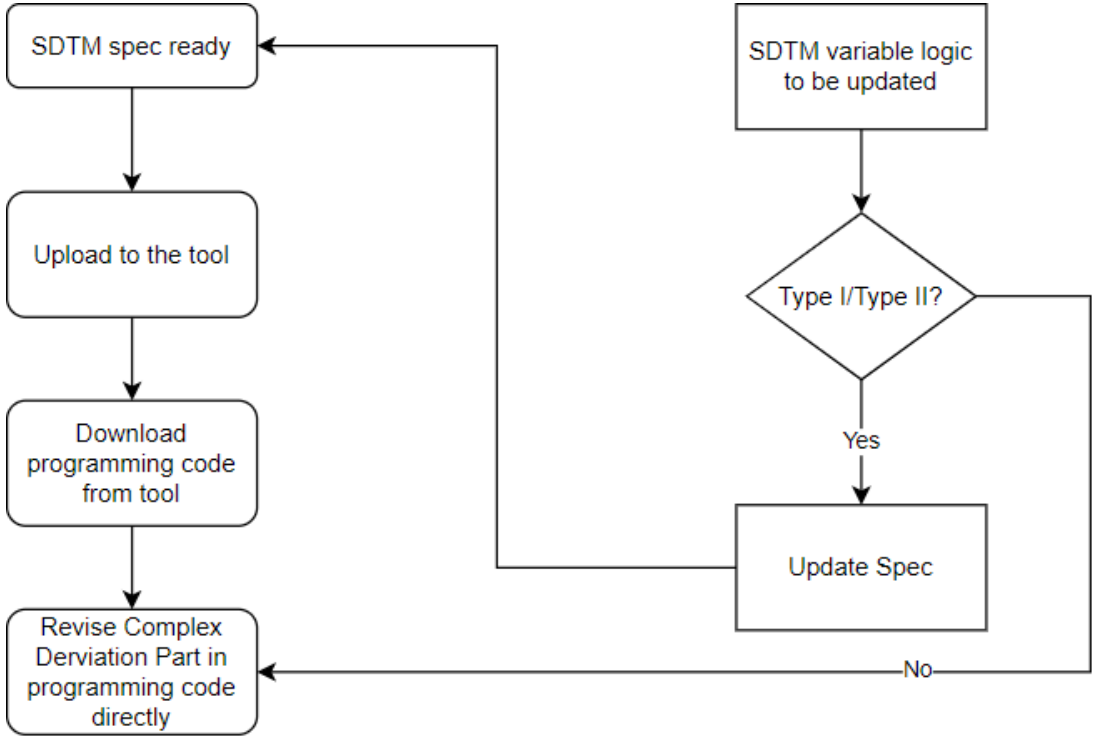
- **Type I (from Spec) – simple derivation**
  - **Copy/Drop/Rename**
  - **Assign - function call**
- **Type II (from Spec) – standard derivation**
  - **Standard Mapping macro call (top down)**
  - **User-defined Pattern macro call, some chance to reuse (bottom up)**
- **Type III (from open code) – complex derivation**
  - **User-defined complex algorithm, little chance to reuse**
  - **Do it in open code**
  - **Rerun applicable with updated Code & Spec**

```
206   _domain='AE';
207   _module='AE SERAE';
208   _aewd=put(  aedis, $NY.);
209   _aendtc=%dtc(INVAR_DAT= ae.aeendat ,INVAR_TIM= ae.aeentim ,OUTVAR_DTC= ae._aeendtc );
210   _aeout=upcase(putc(  aeout,vformat(aeout)));
211   _aerel=put(  aerel, $AERELF.);
212   _aeser=upcase(putc(  aeser,vformat(aeser)));
213   _aesev=upcase(putc(  aesevmax,vformat(aesevmax)));
214   _aestdtc=%dtc(INVAR_DAT= ae.aestdat ,INVAR_TIM= ae.aesttim ,OUTVAR_DTC= ae._aestdtc );
215   _aeterm=aeterm;
216   _aehlgtcd=hlgtcode;
217   _aehlgt=hlgtname;
218   _aehltcd=hlt_code;
219   _aehlt=hlt_name;
220   _ip=ip;
221   run;
222
223 ******************************************************************************;
224 * Execute standalone macros
225 ******************************************************************************;
226 %dy(INVAR_REFDTC=&_templib.._dm.rfstdtc ,INVAR_DTC=&_templib..ae.aeendtc ,OUTVAR_DY=&_templib..ae.aeendy ); ;
227 %dy(INVAR_REFDTC=&_templib.._dm.rfstdtc ,INVAR_DTC=&_templib..ae.aestdtc ,OUTVAR_DY=&_templib..ae.aestdy ); ;
228 %seq(DATASET_LAYOUT=STUDYID USUBJID AESTDTC AEDECOD, OUTVAR_SEQ=&_templib..ae.aeseq); ;
229
```

```
1  ***For RAW.PREG records:\n'if RAW.PREG.MODULE='PRE...***;
5  /* variable LBSPEC todo */
6      VISITNUM=visit;
7      %sdlm dtc(INVAR_DAT=samp dat,INVAR_TIM=,OUTVAR_DTC=&sdlm domain.dtc );
8      if vvalue(ASM_APP)="N" then LBSTAT="NOT DONE";
9      if upcase(module)="PREG" then
0          do;
1          LBORRES=upcase(vvalue(PREGTSTU));
2          LBSPEC="URINE";
3          end;
4      if upcase(module)="PREG1" then
5          do;
6          LBORRES=upcase(vvalue(PREGTSTS));
7          LBSPEC="SERUM";
8          end;
9      LBTESTCD="HCG";
0      LBTEST="Choriogonadotropin Beta";
1      LBCAT="PREGNANCY TEST";
2  /*     LBSPEC= */
3  /*             nan             nan             nan             nan
4  /* variable LBSPEC todo */
5  run;
```

SDTM Mapping Code Generation

- **Process***



***Spec update and code update can happen simultaneously**

SDTM Mapping Code Generation

- Training Summary in first stage

Training Set	Total Variable Number	Automated Variable Number (Type I)	Automated Variable Number (Type II)	Actual Automation Rate (Type I + Type II)	Potential Automation Rate
AE	62	37(59.6%)	16(25.8%)	53(85.4%)	62(100%)
VS	26	14(53.8%)	8(30.8%)	22(84.6%)	24(100%)
EC	38	22(57.9%)	5(13.2%)	27(71.1%)	36(94.7%)
DM	31	14(45.1%)	4(12.9%)	18(58.0%)	26(83.8%)

rstudio.scp.astrazeneca.net/s/aef2cc391f1239bee297a/?view=shiny_foreground

https://rstudio.scp.astrazeneca.net/s/aef2cc391f1239bee297a/p/48ecb705/ | Open in Browser | Publish

Mapping Spec File | Code Generated | Generation Process

Upload Mapping Spec File

Browse... No file selected

Name of Sheet (Case-Sensitive)

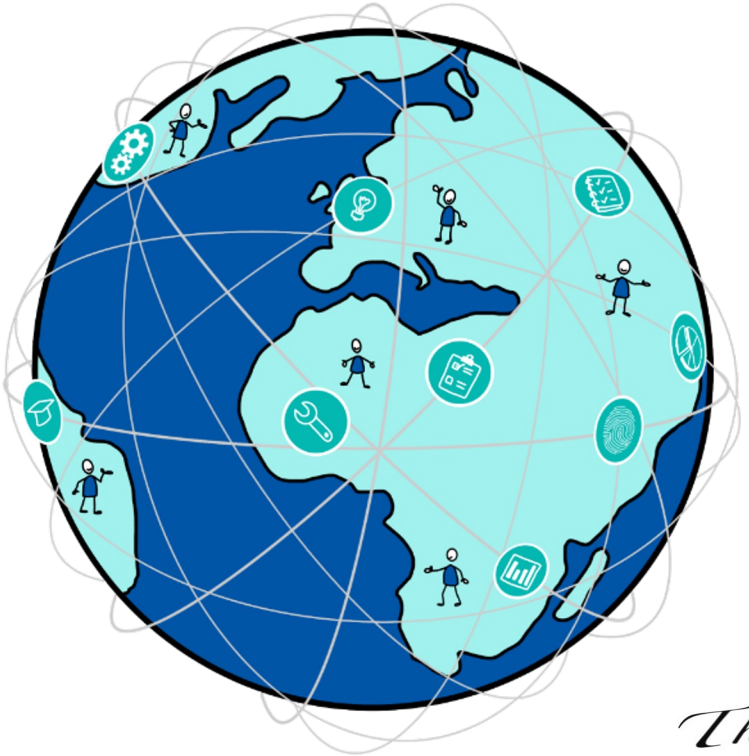
use template SAS code?

Summary

Summary

- New technology is an evitable trend in pharma industry
  - General solution for unstructured data digitization
  - Save human effort to avoid low-value work
  - Combination of Machine and Human(AI+HI)
  - From Rule-based solution to Machine Learning solution

- Business always pursuits efficiency and shorten timeline
  - Side by Side Table Generation Tool accelerates post DBL submission
  - SDTM Mapping Code Generation empowers the SDTM creation



Acknowledgements

Sponsor: Meng Chen

Programming: Xinya Wang, Zhanjian Dong, Kang Zhao, Fan Zhang, Zhilong Qi

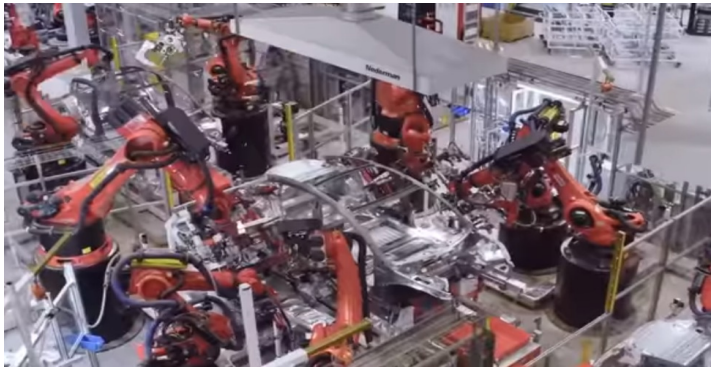
Medical Communication: Ning Zheng, Miao Tong, Simei Yu, Qingdan Du

Data Scientist Intern: Wenkai Xiang, Jessica Chen, Hengjie Cui

backup

Our Problem

Automobile Manufacturing Process



Tesla Gigafactory 3 Shanghai
https://www.youtube.com/watch?v=-Ds1xV7M2gI

- 100% repeatable process
- Machine is the center
- Defective product destroyed

?

or

Transportation System

Pony.ai self-drive
https://www.youtube.com/watch?v=jKFV7qjWaD4

- Always change
- Human in the loop
- Corner case matters

- .rtf/.doc/.docx -> .pdf

Content / SidebySideTable

PDF Table Extractor | Table Join | Output Table

Use this tool to extract structured, machine-readable tables from PDF reports in a few clicks. Load a PDF into the app and extract tables right in the browser, ready for export to Excel

Upload PDF file | FDA-2022-N-1961-0002 | Add suffix to download file | Download

1 | Table Detect | Level Hierarchy | Scrape | Go! | Cancel

	tableID	tableNum	pageNum	rowNum	Lev0	Lev1	V2	V3
1	no tableID detected	1	1	1	Patients screened		XX	XX
2	no tableID detected	1	1	2	Screening failures		XXX (XX.X%)	XXX (XX.X%)
3	no tableID detected	1	1	3	Screening failures	Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
4	no tableID detected	1	1	5	Screening failures	Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
5	no tableID detected	1	1	7	Screening failures	Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
6	no tableID detected	1	1	8	Screening failures	Other	XXX (XX.X%)	XXX (XX.X%)
7	no tableID detected	1	1	9	Patients enrolled		XXX (XX.X%)	XXX (XX.X%)
8	no tableID detected	1	1	10	Patients randomized		XXX (XX.X%)	XXX (XX.X%)

Table 3 provides an example of how screening and enrollment data should be displayed. The populations should be defined as follows:

- **Screened population:** All patients screened for entry into the trial
- **Enrolled population:** All patients who signed a consent form for participation in the trial
- **Randomized population:** All patients randomized to a treatment arm; often referred to as the intention-to-treat population
- If screening failure data are available, provide reasons for screen failure, such as "patient noncompliance," "consent withdrawn," "inclusion/exclusion criteria not met," "other." Request breakdown of "inclusion/exclusion criteria not met" and "other," because understanding reasons for screen failure can be informative to assess the generalizability of the trial.

Table 3. Patient Screening and Enrollment, Trials A and B

Disposition	Trial A	Trial B
Patients screened	XX	XX
Screening failures	XXX (XX.X%)	XXX (XX.X%)
Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
Other	XXX (XX.X%)	XXX (XX.X%)
Patients enrolled	XXX (XX.X%)	XXX (XX.X%)
Patients randomized	XXX (XX.X%)	XXX (XX.X%)

Source: [include Applicant source, datasets and/or software tools used].

Table 4 is an example table showing all reasons reported in the study disposition dataset that should be included. Consider the following study populations:

- **Modified intention-to-treat (mITT) population:** Subset of the ITT population allowing the exclusion of some randomized patients in a justified way (e.g., patients who were deemed ineligible after randomization or certain patients who never started treatment)
- **Safety population:** All patients considered in the safety analyses who received at least one dose of the study drug in the trials submitted
- **Per-protocol population:** Only those patients who completed the treatment originally allocated and planned without major protocol violations

https://www.regulations.gov/document/FDA-2022-N-1961-0002

- .lst -> .pdf

Content / SidebySideTable

PDF Table Extractor | Table Join | Output Table

Use this tool to extract structured, machine-readable tables from PDF reports in a few clicks. Load a PDF into the app and extract tables right in the browser, ready for export to Excel

Upload PDF file | table layout test lst.pdf

Add suffix to download file

Download

1

Table Detect | Level Hierarchy | Scrape

	tableID	tableNum	pageNum	rowNum	Lev0	Lev1	V2	V3
1	No tableID detected	1	1	1	Patients screened		XX	XX
2	No tableID detected	1	1	2	Screening failures		XXX (XX.X%)	XXX (XX.X%)
3	No tableID detected	1	1	3	Screening failures	Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
4	No tableID detected	1	1	5	Screening failures	Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
5	No tableID detected	1	1	6	Screening failures	Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
6	No tableID detected	1	1	7	Screening failures	Other	XXX (XX.X%)	XXX (XX.X%)
7	No tableID detected	1	1	8	Patients enrolled		XXX (XX.X%)	XXX (XX.X%)
8	No tableID detected	1	1	9	Patients randomized		XXX (XX.X%)	XXX (XX.X%)

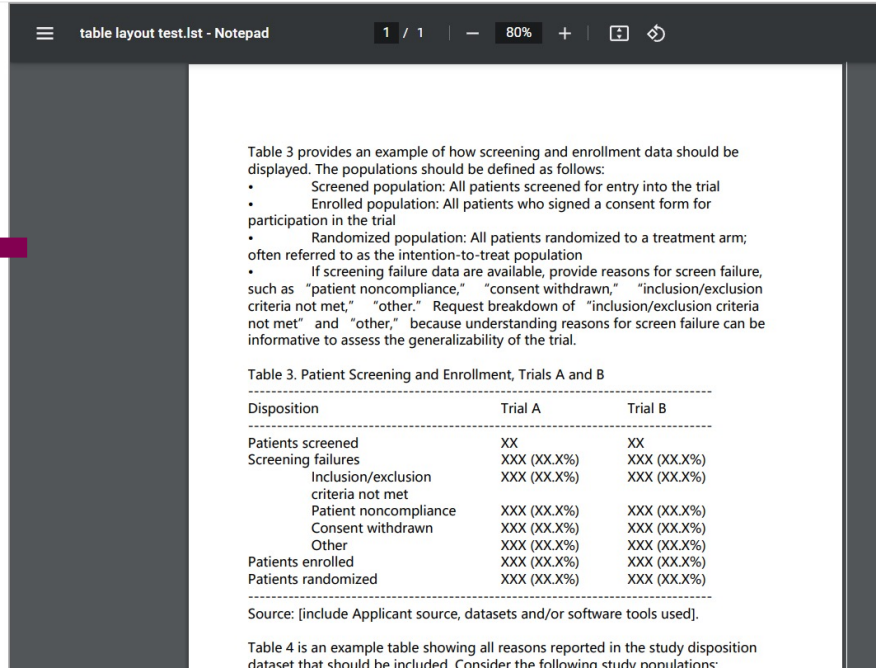


Table 3 provides an example of how screening and enrollment data should be displayed. The populations should be defined as follows:

- Screened population: All patients screened for entry into the trial
- Enrolled population: All patients who signed a consent form for participation in the trial
- Randomized population: All patients randomized to a treatment arm; often referred to as the intention-to-treat population
- If screening failure data are available, provide reasons for screen failure, such as "patient noncompliance," "consent withdrawn," "inclusion/exclusion criteria not met," "other." Request breakdown of "inclusion/exclusion criteria not met" and "other," because understanding reasons for screen failure can be informative to assess the generalizability of the trial.

Table 3. Patient Screening and Enrollment, Trials A and B

Disposition	Trial A	Trial B
Patients screened	XX	XX
Screening failures	XXX (XX.X%)	XXX (XX.X%)
Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
Other	XXX (XX.X%)	XXX (XX.X%)
Patients enrolled	XXX (XX.X%)	XXX (XX.X%)
Patients randomized	XXX (XX.X%)	XXX (XX.X%)

Source: [include Applicant source, datasets and/or software tools used].

Table 4 is an example table showing all reasons reported in the study disposition dataset that should be included. Consider the following study populations: