AstraZeneca

Data Science Project Report

Directory

1.1

NER

Model to recognize variables, quotations, formulas and functions from spec

Name entity recognition

General example:

On the 15th of September DATE , Tim Cook PERSON announced that Apple ORG wants to acquire ABC Group ORG from New York GPE for 1 billion dollars MONEY

Domain-specific:

If SMOKSTAT VAR in ('CURRENT', 'FORMER') FUNC then 'SMOKER' QUOT ; else if SMOKSTAT VAR eq 'NEVER' QUOT then 'NON-SMOKER' QUOT .

The challenges

Ambiguity

Something looks like variable, formula or function, but they are not:

- no-missing
- FOLLOW-UP
- date-variable
- date/time
- at a visit (LT)
- …

Length change

Long entities are hard to be recognized precisely

Example:

Assign CEGRPID with ifc(upcase(ADJAE.ADJID)='FINAL', 'FINAL', ifc(ADJAE.ADJID ne '', 'NON-FINAL', ''))

The way of writting

- Punctuation confusion between Chinese and English
- No Spaces after commas
- Variable description

Calculation of AESTDY = (Numeric version of date part of AESTDTC - Numeric version of date part of DM.RFSTDTC).

The challenges

Nested entities

The outermost entity needs to be translated

Example:

Set to "Persons Aged >=65 Years" for RAW.SCOV2RP.RFAGE55

Tokenization

Tokenization is very important because it determines whether the entity can be recognized or not

Example:

If VAR=0.Then set VAR1=1

If VAR=0.Then 'set' VAR1=1

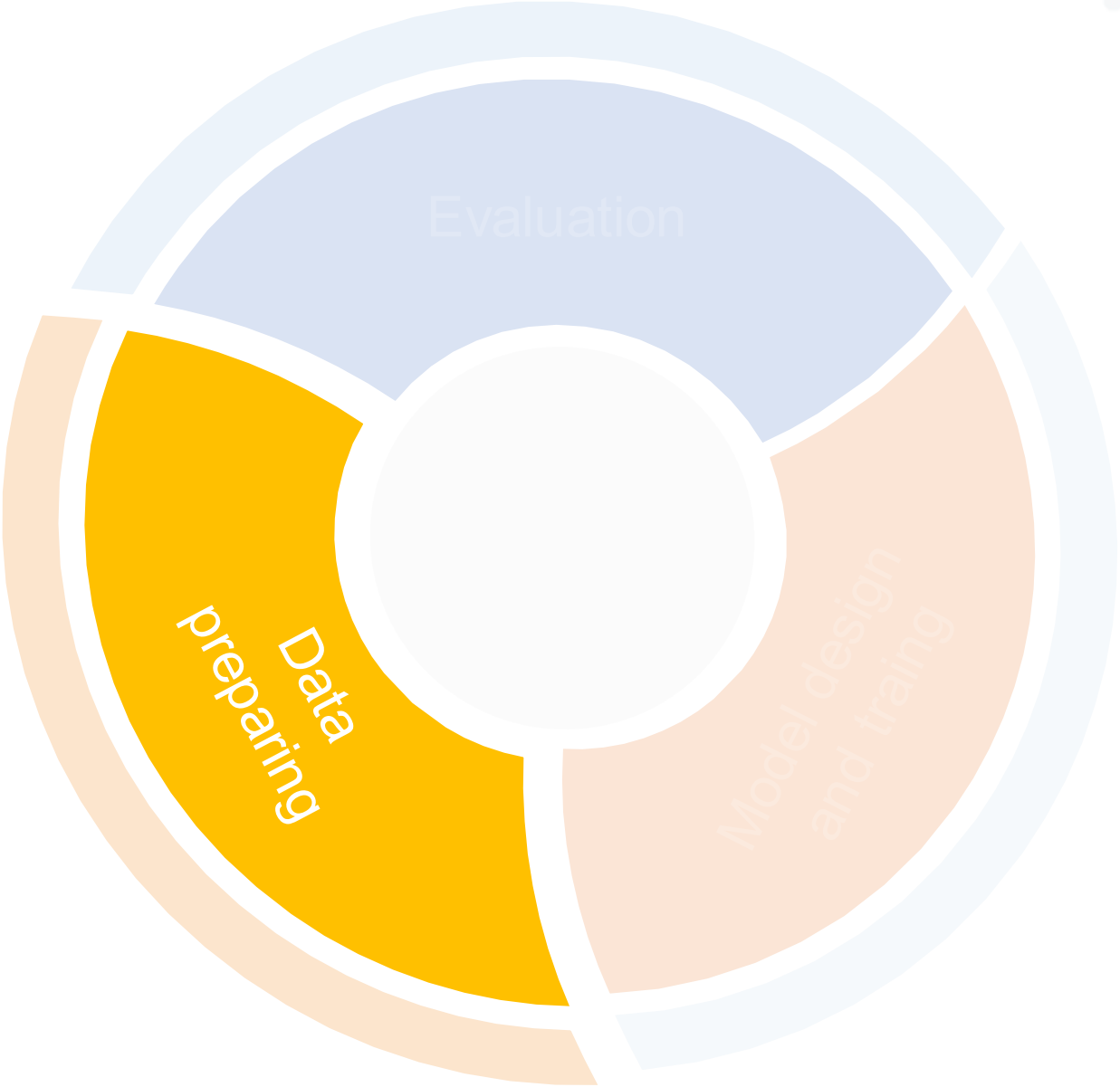
Quotes

Quotes can be confused sometimes

- it's
- don't
- patient's
- """

Steps to build an NER model

Step 1



Break down

combine

Use hundreds of specs

Rule + human

SPEC	FORMULA	FUNCTION	VARIABLE	QUOTATION	FORMAT
Set to substr (vvalue (RAW.AE.AEDIS), 1,1)		substr (vvalue (RAW.AE.AEDIS), 1,1)	RAW.AE.AEDIS		
"Considering only non-missing values:Set to 'Y' If (ADSL.DTHDTN - TRTEDT + 1) <= 30; Set to 'N' otherwise."	(ADSL.DTHDTN - TRTEDT + 1) <= 30		ADSL.DTHDTN TRTEDT	'Y' 'N'	
...	...	...	...	...	...

Entity pattern set

Variables

- VSDY
- SUPPLB.SPCANDTC
- RAW.LB.LBORRES
- RAW.SCOV2SD.MODULE_O
- Usubjid
- …

Formulas

- [V1] = [quot1]
- [V1] <= [V2] <= [V3]
- [V1]+ [V2]
- [V1]/[V2] > 5
- [V1] + [V2]+[V3]- 1 >= 0
- [V1] = .
- …

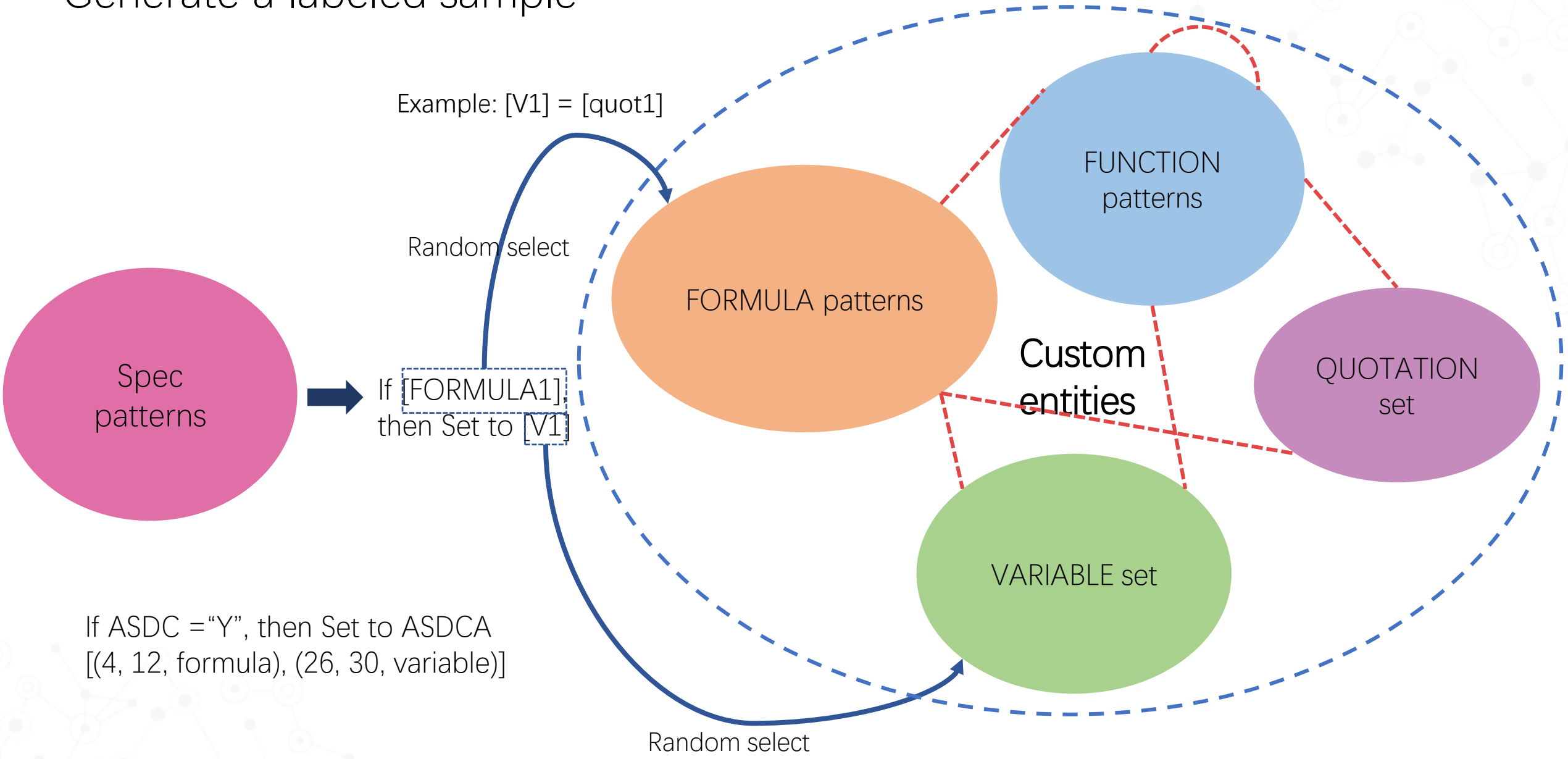
Functions

- Concatenate([V1] and [V2])
- ifc([V2] [fc6], [fc5], [fc2])
- in([quot1] ,[quot2] ,[quot3])
- input ([V1] ,[V2].)
- missing([V1])
- upcase ([V0])
- Upcase ([fc0])
- …

Break down

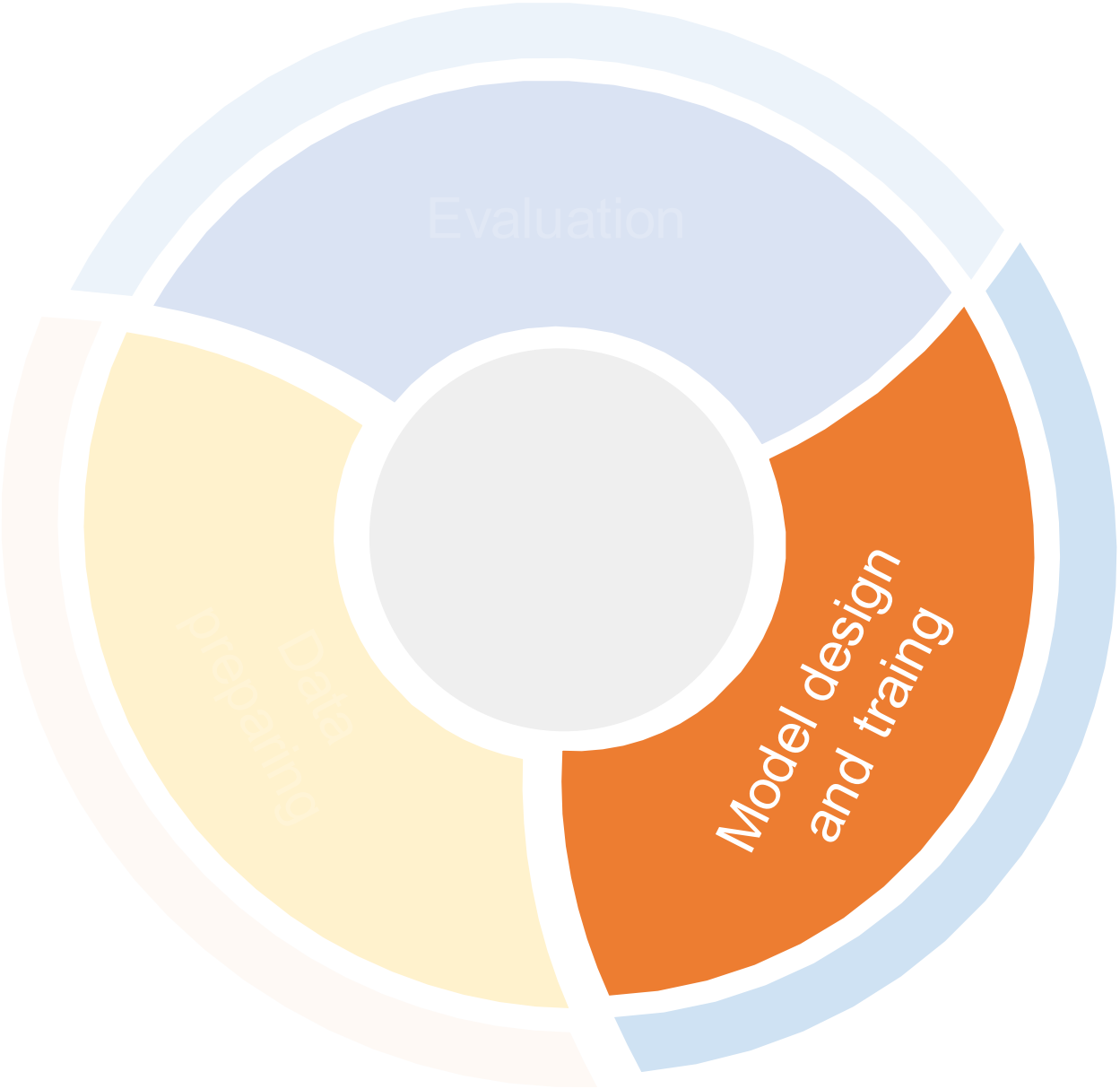
combine

Generate a labeled sample



If ASDC ="Y", then Set to ASDCA
[(4, 12, formula), (26, 30, variable)]

Step 2



Model design and training

Set to substr (vvalue (RAW.AE.AEDIS), 1,1)

tokenizer

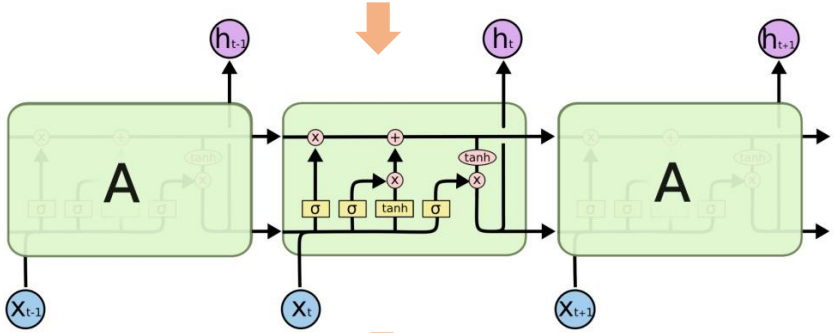
Set | to | substr | (| vvalue | (| RAW.AE.AEDIS |) | , | 1 | , | 1 |)

token to vector model

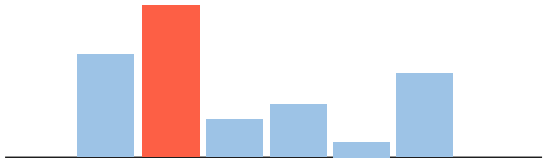
- Word2vec
- Glove
- Bert

0.1	0.1	0.5	0.7	0.4		
0.0	0.4	0.6	0.9	0.5		
0.2	...	0.0	0.1	0.5		
...	0.9	...	...	...	...	...
0.6	0.8	0.4	0.2	0.3		
0.1	0.3	0.2	0.8	0.0		
0.7	0.2	0.5	0.3	0.2		

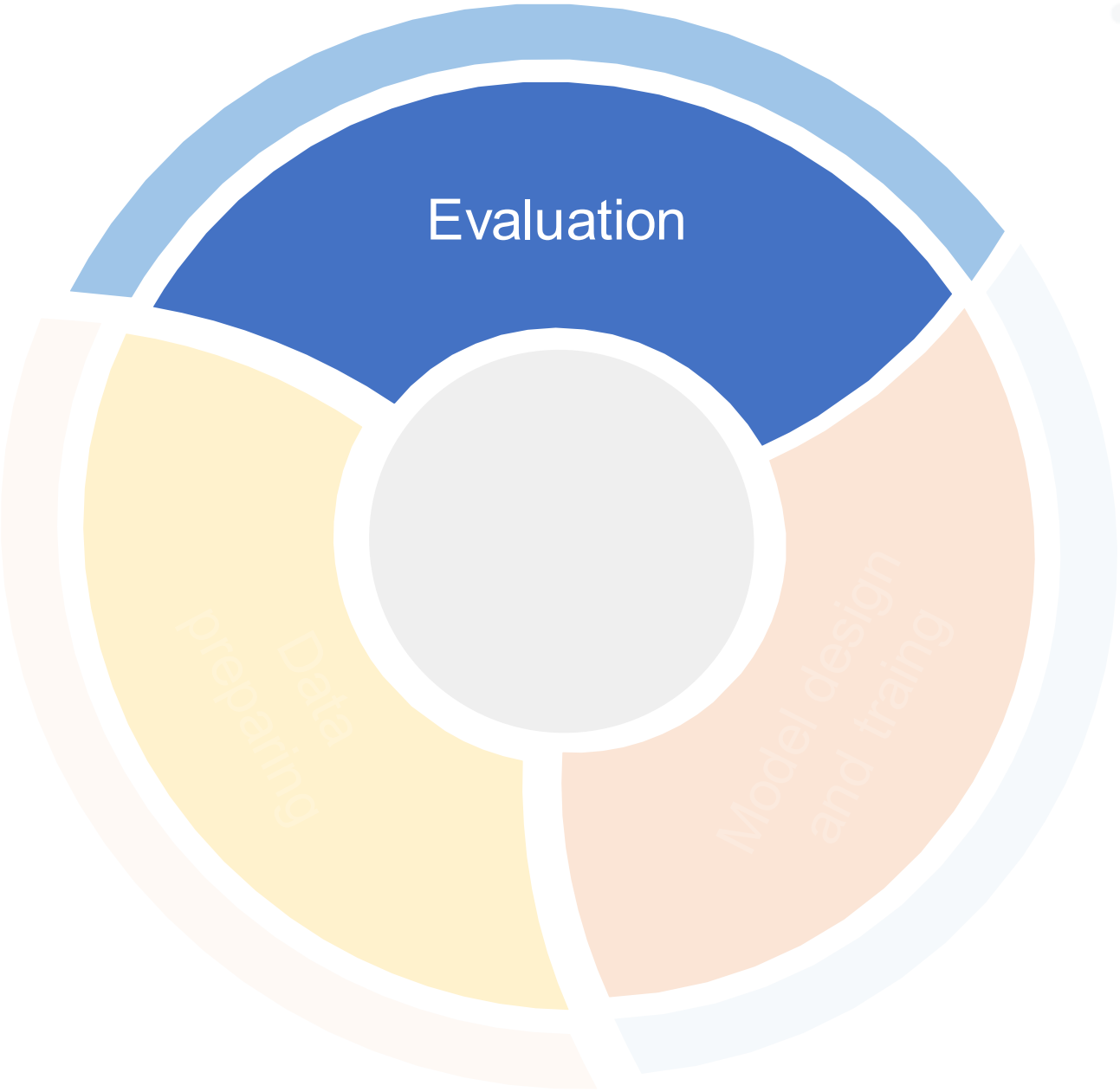
NER model
(stack-lstm)



prediction



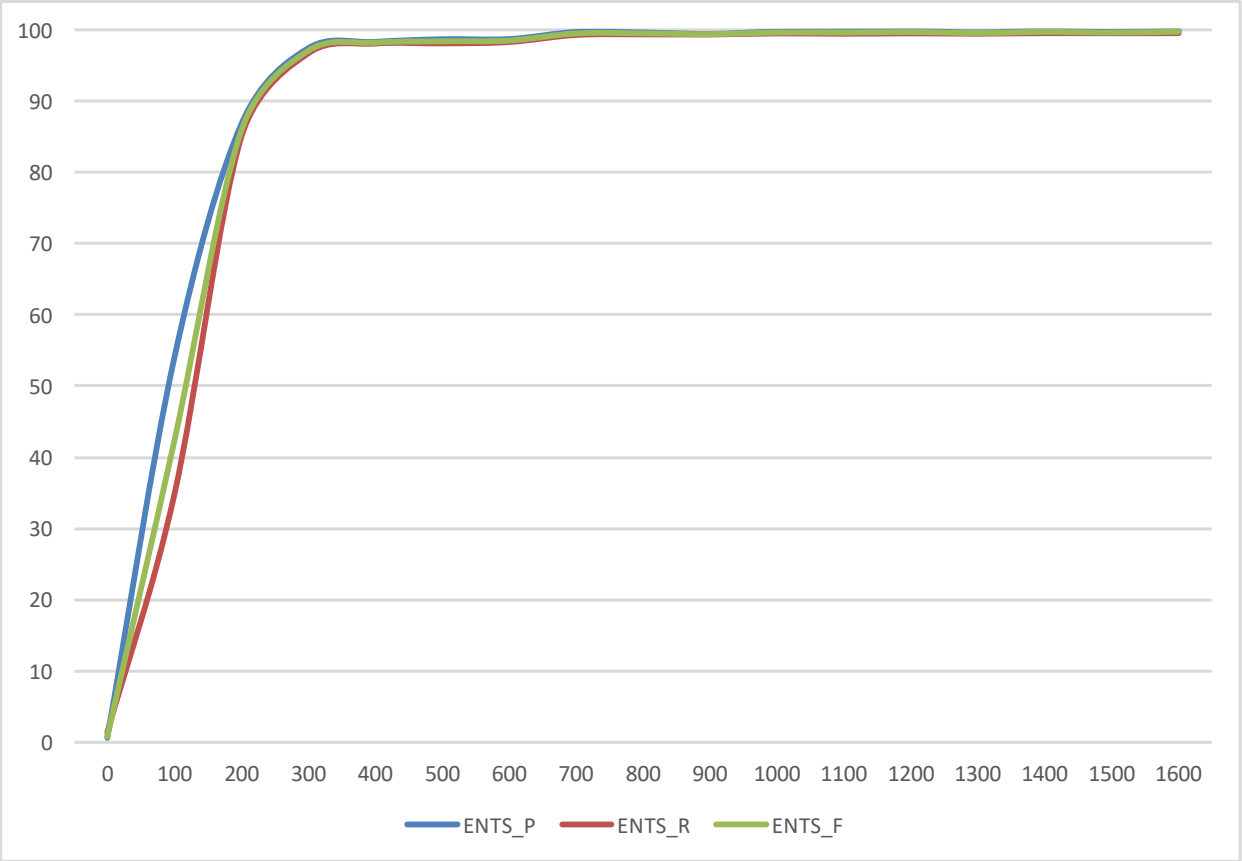
Step 3



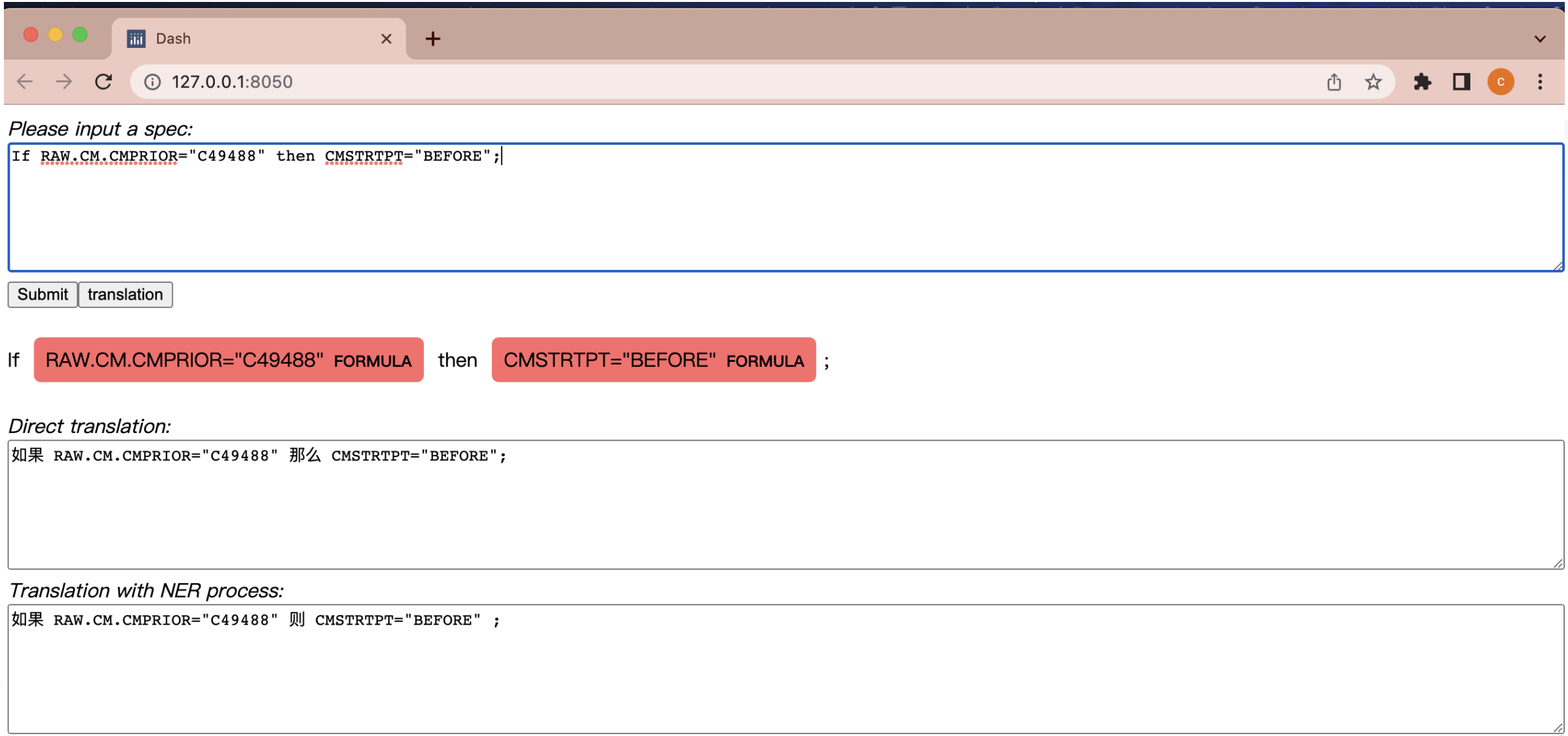
NER-train

Training samples(random generated): 100000
Test samples(random generated): 5000

BATCH	ENTS_P	ENTS_R	ENTS_F
0	0.58	1.38	0.81
100	53.84	34.75	42.24
200	86.63	85	85.8
300	97.33	96.78	97.05
400	98.22	98.16	98.19
500	98.6	98.18	98.39
600	98.62	98.37	98.49
700	99.61	99.37	99.49
800	99.54	99.45	99.5
900	99.35	99.43	99.39
1000	99.66	99.58	99.62
1100	99.7	99.53	99.61
1200	99.78	99.59	99.69
1300	99.55	99.54	99.55
1400	99.79	99.62	99.71
1500	99.62	99.6	99.61
1600	99.82	99.62	99.72



NER predict demo



1.2

NER applications

Sentence pattern analysis and AI translation

19

Sentence pattern analysis

Sentence pattern analysis

Input: spec sentences

NER predict	Embedding	cluster
Use NER model to recognize variables, formulas, functions and quotes. Then replace them with special string such as [formula]	Use NLP model to transform each sentence to a dense vector. Here we use a transformer model called MPNET	Cluster sentences based on their embedding vectors. Here k-means is enough.

Sentence pattern analysis

Analysis 23270 sentences and get their patterns. Filter patterns numbers>100

A	B	C	D
spec pattern	数量	占比	cluster
Subset the data using the logic: <formula> and not <function>	194	0.83%	0
Subset the data using the logic: <formula> ne 0 and not <function>	137	0.59%	0
Assign <variable> with <quotation>	5398	23.20%	1
Subset the data using the logic: not <function>	7651	32.88%	2
Subset the data using the logic: not <function> and not <function>	145	0.62%	2
Assign <variable> with <function>	3452	14.83%	3
Assign <variable> from <variable>	2138	9.19%	4
Assign <variable>	350	1.50%	4
Creates a date variable (<variable>) from (<variable>) in <variable>	367	1.58%	5
Subset the data using the logic: <formula>	510	2.19%	6
Convert type for attribute <variable> from <variable>	467	2.01%	7
Convert type for attribute <variable>	139	0.60%	7
Concatenate parts <variable>, <quotation>, <variable>	323	1.39%	8
Combines a date-variable (<variable>) and a time variable (<variable>) to create a date/time variable (<variable>) in <variable>	187	0.80%	10
Assign <variable> from <formula>	151	0.65%	12
汇总	22687	92.86%	

AI translation

NER to help translation

Input: spec

NER predict	Translation	restore
Use NER model to recognize variables, formulas, functions and quotes. Then replace them with special string such as AAAA0A	Use translation model to translate English to Chinese	Restore entity to original

Example: ADT is numeric date part of VS.VSDTC

AAAA0A is numeric date part of AAAA1A → AAAA0A是AAAA1A的数值日期部分 → ADT是VS.VSDTC的数值日期部分

Advantages

- Quotes will not affect the translation

spec

A patient is randomized when a record with DS.DSDECOD="RANDOMIZATION CODE ALLOCATED" exists in DS and when a planned treatment is available

A patient is randomized when a record with DS.DSDECOD="RANDOMIZATION CODE ALLOCATED" FORMULA exists in DS VAR and when a planned treatment is available

Direct translation

当DS中存在DS.DSDECOD="分配随机化CODE"的记录和计划治疗可用（DM.ARMCD）时，对患者进行随机分组。

NER translation

当DS中存在DS.DSDECOD="RANDOMIZATION CODE ALLOCATED"记录时和计划治疗可用（DM.ARMCD）时，对患者进行随机分组。

Advantages

- Variables will not be translated

spec

Create at most one record with DVSPID='1.2' for each subject in the FAS, satisfying any of the following criteria: 1) ADSL.SEX ne 'M'. 2) ADSL.AGE < 18 and ADSL.COUNTRY not in ('JPN','TWN') . 3) ADSL.AGE < 20 and ADSL.COUNTRY in ('JPN','TWN')

Create at most one record with DVSPID='1.2' FORMULA for each subject in the FAS VAR , satisfying any of the following criteria: 1) ADSL.SEX ne 'M' FORMULA . 2) ADSL.AGE < 18 FORMULA and ADSL.COUNTRY VAR not in ('JPN','TWN') FUNC . 3) ADSL.AGE < 20 FORMULA and ADSL.COUNTRY VAR in ('JPN','TWN') FUNC

Direct translation

为FAS中的每例受试者创建最多一条记录，DVSPID="1.2"，符合以下任一标准：1）ADSL.SEX ne'M'。2）ADSL.年龄<18和ADSL.国家不在（'JPN'，'TWN'）。3）ADSL.年龄<20岁和ADSL.国家（"JPN"、"TWN"）

NER translation

为FAS中的每例受试者创建最多一条DVSPID='1.2'记录，符合以下任一标准：1）ADSL.SEX ne 'M'。2）ADSL.AGE < 18和ADSL.COUNTRY而非in ('JPN','TWN')。3）ADSL.AGE < 20和ADSL.COUNTRY in ('JPN','TWN')

Advantages

- View VAR=. VAR ne . as formula

spec

if ASTDT ne . and TRTSDT ne . And ASTDT >= TRTSDT then APHASE = "On and Off Treatment"

if | ASTDT ne . FORMULA | and | TRTSDT ne . FORMULA | And | ASTDT >= TRTSDT FORMULA | then | APHASE = "On and Off Treatment" FORMULA |

Direct translation

如果 ASTDT ne 。和 TRTSDT ne 。并且 ASTDT >= TRTSDT 然后 APHASE = "开和关治疗"

NER translation

如果 ASTDT ne . 和 TRTSDT ne . 和 ASTDT >= TRTSDT 则 APHASE = "On and Off Treatment"

Advantages

- Sentence structure can be simpler and more clear

spec

If patient took opiods at baseline (i.e. If (SDTM.CM.CMSTDTC <= ADSL.RANDDTC <= SDTM.CM.CMENDTC) or if (SDTM.CM.CMSTDTC <= ADSL.RANDDTC and SDTM.CM.CMENRTPT = "ONGOING")) then equal to "Taking opioids at baseline". Otherwise equal to "Not taking opioids at baseline".

If patient took opiods at baseline (i.e. If (SDTM.CM.CMSTDTC <= ADSL.RANDDTC <= SDTM.CM.CMENDTC FORMULA) or if (SDTM.CM.CMSTDTC <= ADSL.RANDDTC FORMULA and SDTM.CM.CMENRTPT = "ONGOING" FORMULA)) then equal to "Taking opioids at baseline" QUOT . Otherwise equal to "Not taking opioids at baseline" QUOT .

Direct translation

如果患者在基线时服用阿片类药物（即如果 (SDTM.CM.CMSTDTC <= ADSL.RANDDTC <= SDTM.CM.CMENDTC) 或如果 (SDTM.CM.CMSTDTC <= ADSL.RANDDTC 和 SDTM.CM.CMENRTPT = "ONGOING ")) 然后等于“在基线时服用阿片类药物”。否则等于“基线时不服用阿片类药物”。

NER translation

如果患者在基线时服用阿片类药物（即如果（SDTM.CM.CMSTDTC <= ADSL.RANDDTC <= SDTM.CM.CMENDTC）或如果（SDTM.CM.CMSTDTC <= ADSL.RANDDTC 和 SDTM.CM.CMENRTPT = "ONGOING"）），则等于 "Taking opioids at baseline"。否则等于 "Not taking opioids at baseline" 。

Problems

- Noise of input

1 AENDT – CORE.TRTSDT + 1 if it is on or after TRTSDT,
else AENDT – CORE.TRTSDT

AENDT – CORE.TRTSDT + 1 FORMULA if it is on or after TRTSDT VAR , else AENDT – FORMULA CORE.TRTSDT

2 AENDT – CORE.TRTSDT + 1 if it is on or after TRTSDT,
else AENDT - CORE.TRTSDT

AENDT – CORE.TRTSDT + 1 FORMULA if it is on or after TRTSDT VAR , else AENDT – CORE.TRTSDT FORMULA

Problems

- Special formulas or functions

1 ADTR.AVAL (where PARAMCD = "TRSUMND") - sum (the diameter of lesions at this visit which correspond to the missing lesions) (ADTR.AVAL=. (where PARAMCD = "TRLONDD")) .

ADTR.AVAL VAR (where PARAMCD = "TRSUMND" FORMULA) – sum (the diameter of lesions at this visit which correspond to the missing lesions) (ADTR.AVAL=. FORMULA (where PARAMCD = "TRLONDD" FORMULA)) .

2 ADTR.AVAL where PARAMCD = "TRSUMND" - sum (the diameter of lesions at this visit which correspond to the missing lesions (ADTR.AVAL=. (where PARAMCD = "TRLONDD")) or lesions with intervention (TRINTRFL = "Y" (where PARAMCD = "TRLONDD"))

ADTR.AVAL VAR where PARAMCD = "TRSUMND" – sum (the diameter of lesions at this visit which FORMULA correspond to the missing lesions (ADTR.AVAL=. FORMULA (where PARAMCD = "TRLONDD" FORMULA)) or lesions with intervention (TRINTRFL = "Y" FORMULA (where PARAMCD = "TRLONDD" FORMULA))

2.

Graph database

Challenges

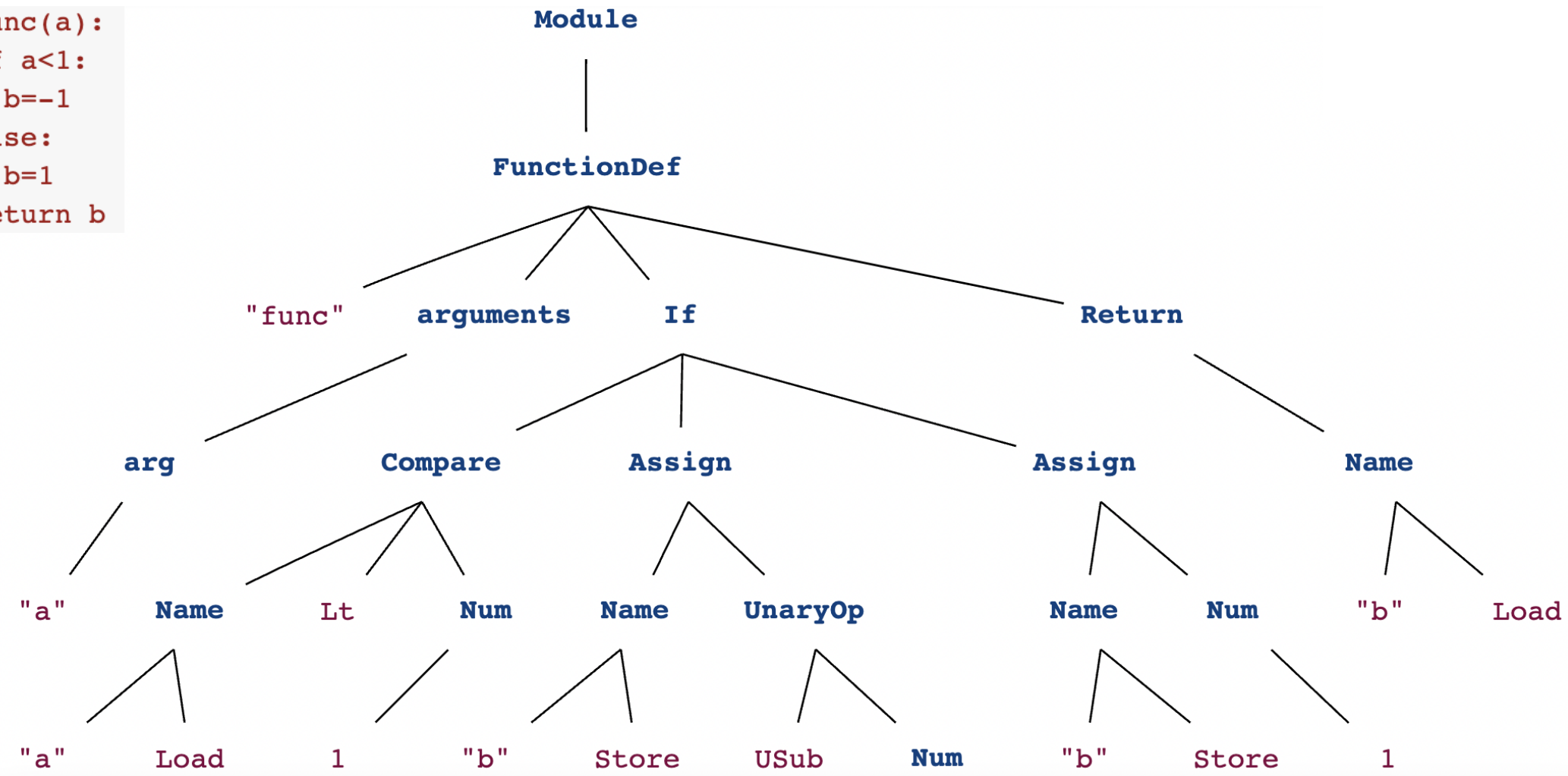
1.How to define nodes types and relation types?

2.How to extract relations?

Readability **Expansibility** **Efficiency**

AST structure

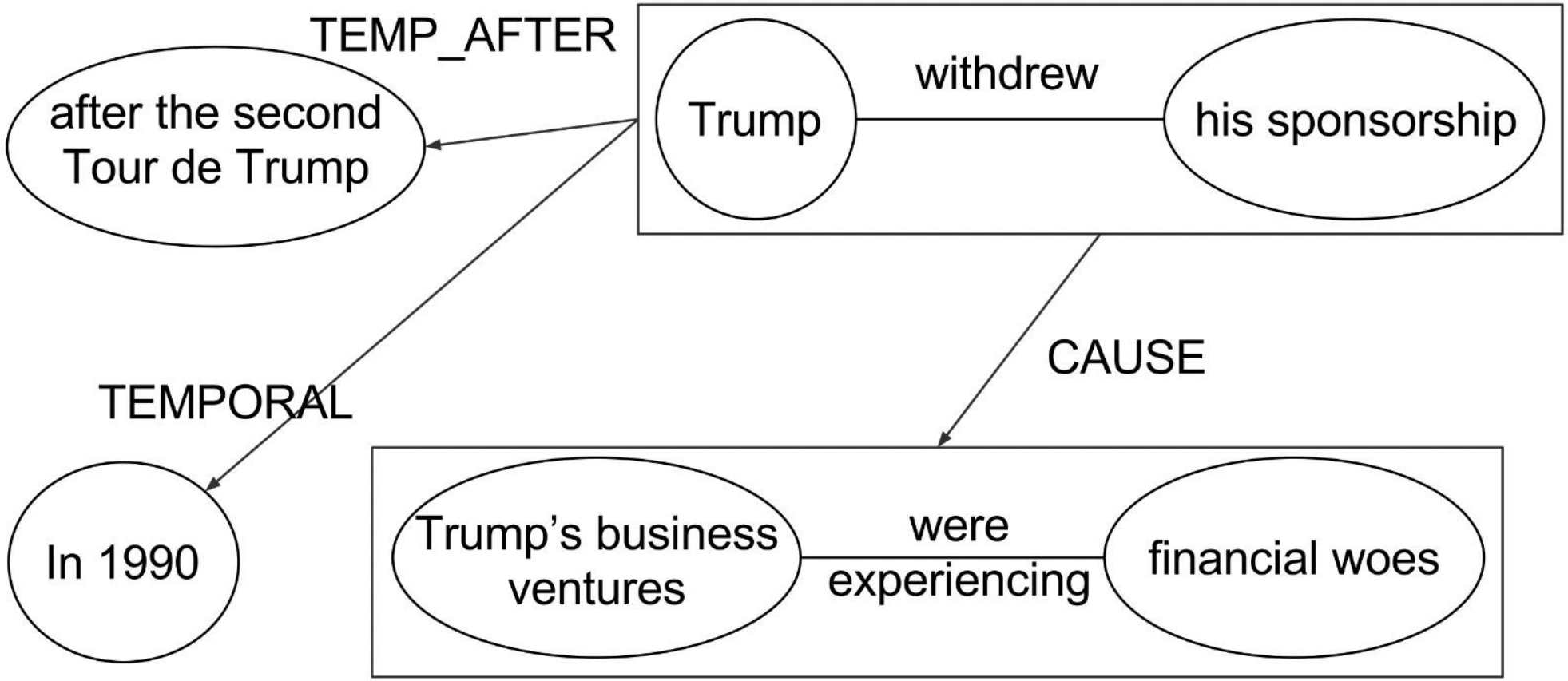
```
def func(a):
    if a<1:
      b=-1
    else:
      b=1
    return b
```



General methods to extract relations

Graphene

```
Trump withdrew his sponsorship after the second Tour de Trump in 1990 because his business
ventures were experiencing financial woes.
```



Key words to extract relations

level1	if, else, otherwise, else if, where, when, then
level2	and, or
level3	convert to, set to, assign to, equal to, contain

For example:

- If [condition] then [operation]
- If [condition], [operation]
- [operation] if [condition]

[condition] and [operation] can be divided by level2 or level3 key words!

nodes

Spec entry

DEF

<id>	7273
domain	unknow
nodeId	71
nodes_raw	DEF
nodes_rep	DEF15
spec_domain	DM
stage	sdtm
study	d3250c00034
target_var	AGEU

Key words relations

LOGIC

<id>	4
domain	unknow
nodeId	17
nodes_raw	WHERE
nodes_rep	WHERE0
spec_domain	DM
stage	sdtm
study	d3250c00034
target_var	RFSTDTC

From NER

VAR

<id>	5956
description	Subject Death Flag
domain	DM
nodeId	57
nodes_raw	DTHFL
nodes_rep	DTHFL
spec_domain	DM
stage	sdtm
study	d3250c00034
target_var	DTHFL

VALUE

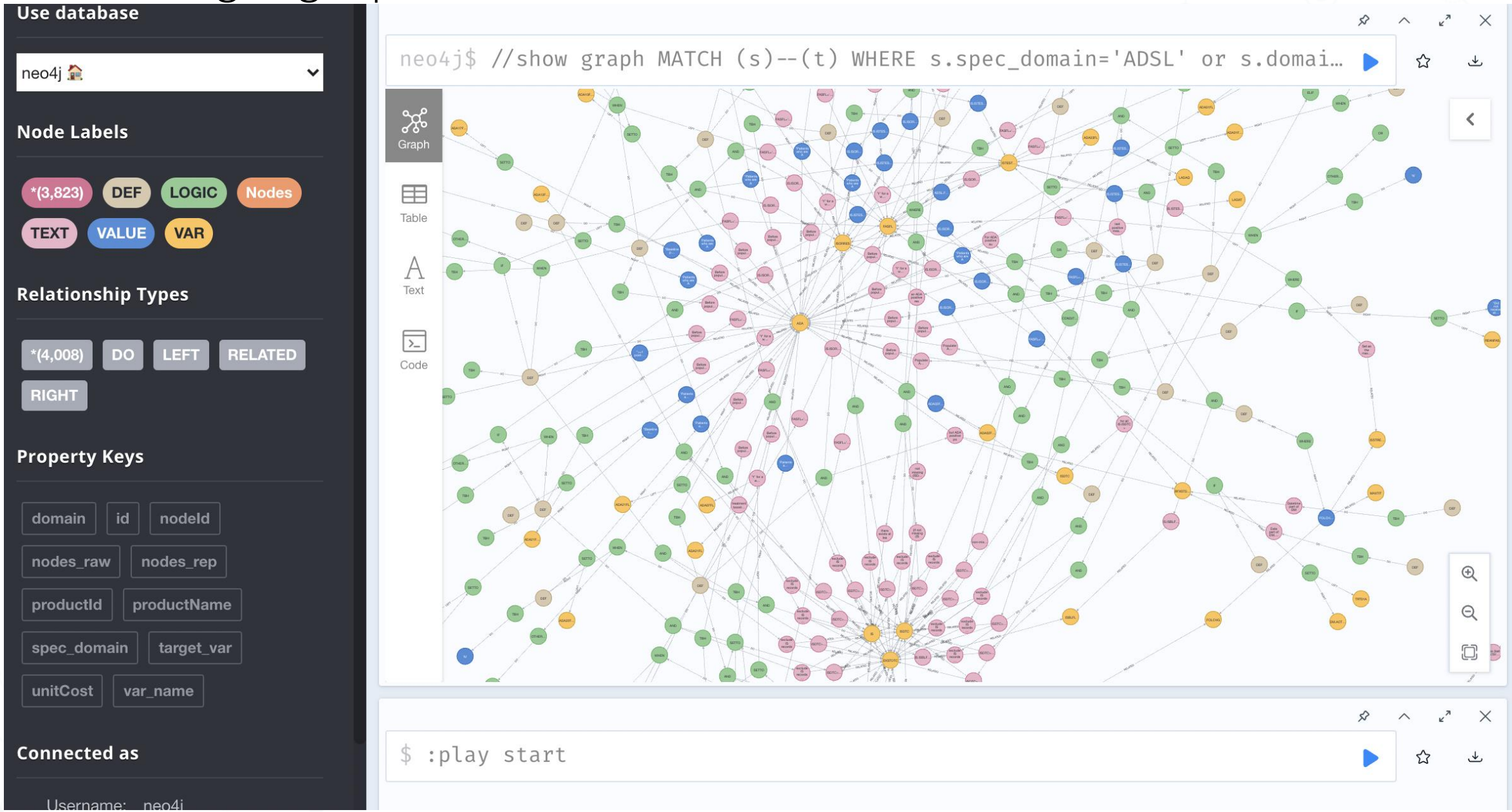
<id>	8744
domain	unknow
nodeId	84
nodes_raw	RAW.DM.SEX='C20197'
nodes_rep	FORMULA5
spec_domain	DM
stage	sdtm
study	d3250c00034
target_var	SEX

Others

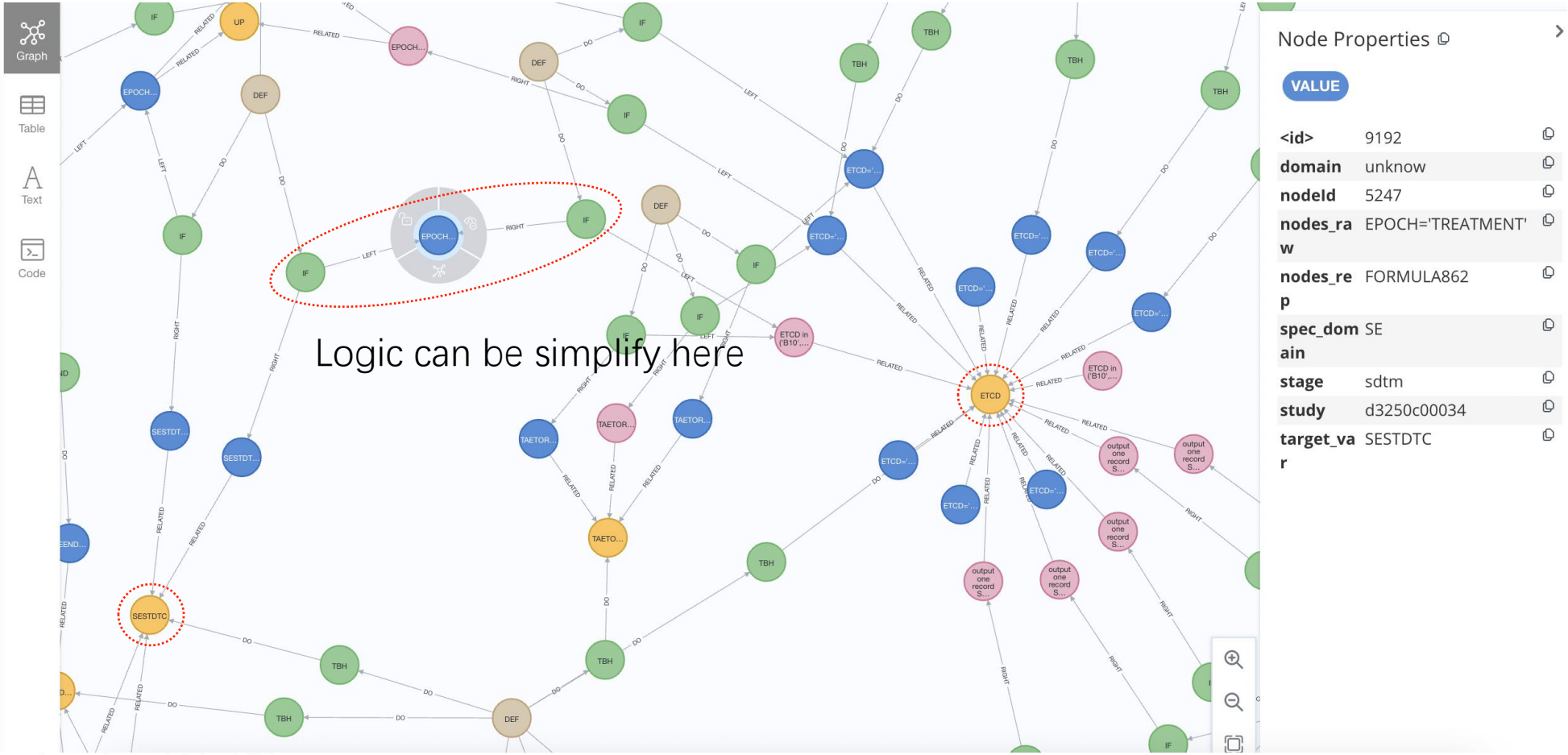
TEXT

<id>	3836
domain	unknow
nodeId	74
nodes_raw	SDTM.DM.AGE not missing
nodes_rep	FORMULA4 not missing
spec_domain	DM
stage	sdtm
study	d3250c00034
target_var	AGEU

Text to logic graph



Use case


Graph
Table
Text
Code
Logic can be simplify here
Node Properties
VALUE
<id> 9192
domain unknow
nodeId 5247
nodes_raw EPOCH='TREATMENT'
nodes_rep FORMULA862
spec_domain SE
stage sdtm
study d3250c00034
target_var SESTDTC

Use spec konwledge graph to support mutiple tasks

* Nodes classification
* Nodes importance analysis
* Logic simplify
* Cross domain analysis
* Hierarchical graph analysis
* (Sub)graph to generate description
* (Sub)graph to generate code framework.

Summary

First

Several hundreds unlabeled spec. → Lots of labeled high-quality samples → NER (High accuracy and robust) → Sentence pattern analysis

AI translation

Second

Design rules to extract relations from complex specs → Build knowledge graph in Neo4j database → Explore potential application

Thanks!