What's the future?

- **Wide mapping table and relationship groups**
  - For complex expressions to make ETL easy
- **CDE (common data environment)**
  - For effectiveness and consistent mappings
- **Metadata**
  - For better quality and precision
- **ML/AI**
  - For automation to optimize of costs and time

Wide mapping table and relationship groups

For “other” types of data - **entity-attribute-value (EAV)** records:

- entity is either a question or a variable
- attribute is the link
- value or answer is the value

Type	Variable / Question	Value / Answer
Lab tests with the qualitative result	SARS-CoV-2 (COVID-19) IgA+IgM [Presence] in Serum or Plasma by Immunoassay	Equivocal / Negative / Positive
Historic facts	Family history of clinical finding	Myocardial infarction
Cancer stages and assessment measures	FIGO Stage (2018 FIGO Cancer Report)	I: Tumor confined to ovaries or fallopian tube(s)
	Circumferential Resection Margin (CRM)	100 mm or greater
Survey instruments created for specific projects (UK Biobank, All Of US PPI)	Has a doctor told you that you have any of the following problems with your eyes?	Macular degeneration
	How often did you use cannabis?	1-5 times per week
Surveys by itself (PhenX, PROMIS)	Because of your problem, do you feel frustrated	No / Sometimes / Yes
	Smoking helps me concentrate	Not at all / Somewhat / Very much

Limitations of current approach

Use case	Example	Issue
One-to-many "splitting"	"Maps to" and "Maps to value" pairs: "History of" + value of "COVID-19 vaccine" together with "SARS-COV2 PCR test" + value of "POS"	It is ambiguous which "Maps to" belongs to which "Maps to value", and the standard ETL process will inflate the records
Many-to-one "merging"	HHV-6B seropositivity for Human Herpesvirus-6: False	Only a single code can be an input for a map. As a result, the ETL needs to apply a workaround and first merge the entity/value codes to map them to the target concept
	EuroQol five dimension three level self-care score: 3 (I am unable to wash or dress myself)	
Separate mapping for entities and values	Generic "Yes", "No" answers to questions; drugs, conditions and other self-sufficient concepts	Now this is managed by splitting the source codes into separate synthetic source vocabularies
Mapping to numeric content	CS Tumor Size of 32 mm	Currently, ETL needs to extract the numeric values and units from the text
Mapping of a range	Blood alcohol level of 100-119 mg/100 ml	Ranges are currently not supported
Mapping to a string	White sliced bread eaten	Currently, ETL needs to extract the values from the text
Mapping to a date	Birthdate of a relative: "1988-Sep-17"	Currently, ETL needs to extract the dates from the source

Wide mapping table

Source				Standard Concept
Source Concept	Question/Variable	Answer/Value	Range	
Ambulatory procedures - lithotripsy				Lithotripsy
	CS Tumor Size	\d	001 - 988 millimeters (mm) |(Code exact size in mm)	Estimated Tumor Size
Documentation of patients with primary headache diagnosis and imaging other than ct or mri obtained				Headache Imaging
Evidence of alcohol involvement determined by blood alcohol level of 100-119 mg/100 ml				Ethanol [Mass/volume] in Blood
Home visit, phototherapy services (e.g., bili-lite), including equipment rental, nursing services, blood draw, supplies, and other services, per diem				Home visit, phototherapy services (e.g., bili-lite), including equipment rental, nursing services, blood draw, supplies, and other services, per diem
	Wears glasses or contact lenses	Yes		Abnormal vision Uses visual aid
	Age started wearing glasses or contact lenses	**\d (e.g. 15)**		History of event **longer than 10 years ago**
	Type of sliced bread eaten	white		Food eaten

...

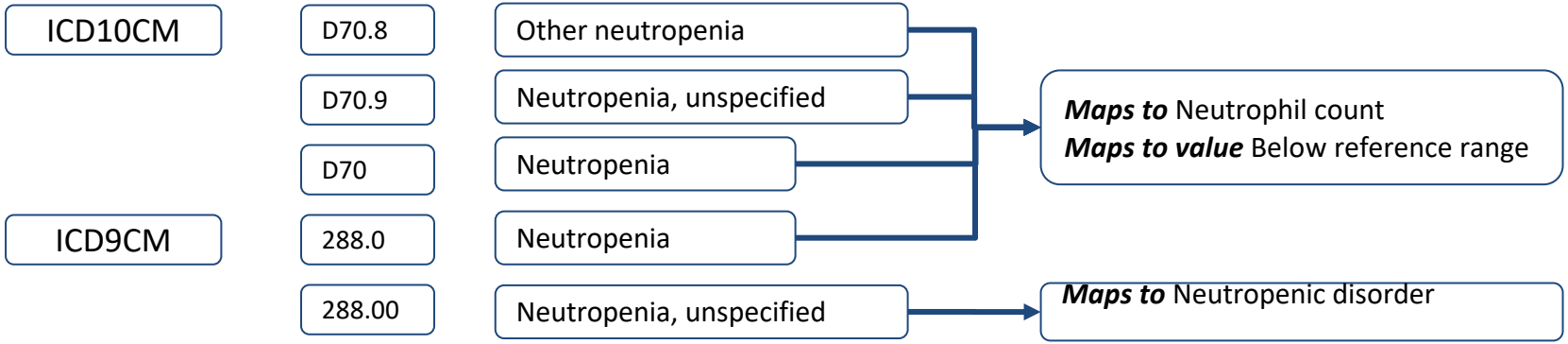
Target							
Numeric	Operator	Error	Unit Concept	Value Concept	String	Condition Status Concept	Visit Concept
							Ambulatory Surgical Center
\d			millimeter				
						Primary diagnosis	
110		10	milligram per deciliter				
							Home Visit
				Uses visual aid			
					“white sliced bread”		

...

CDE (common data environment)

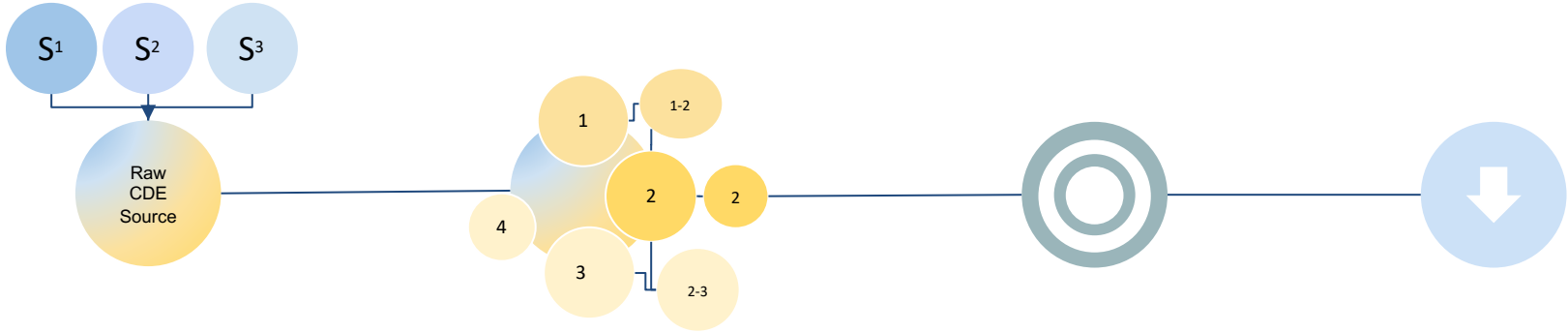
Addressing issues:

- Mapping discrepancies across vocabularies containing same or close semantic entities
- Suboptimal mappings reuse



Goal: create a structure for grouping of different source data, storage of mapping candidates of different origin and decision making on preferable mappings

CDE dataflow



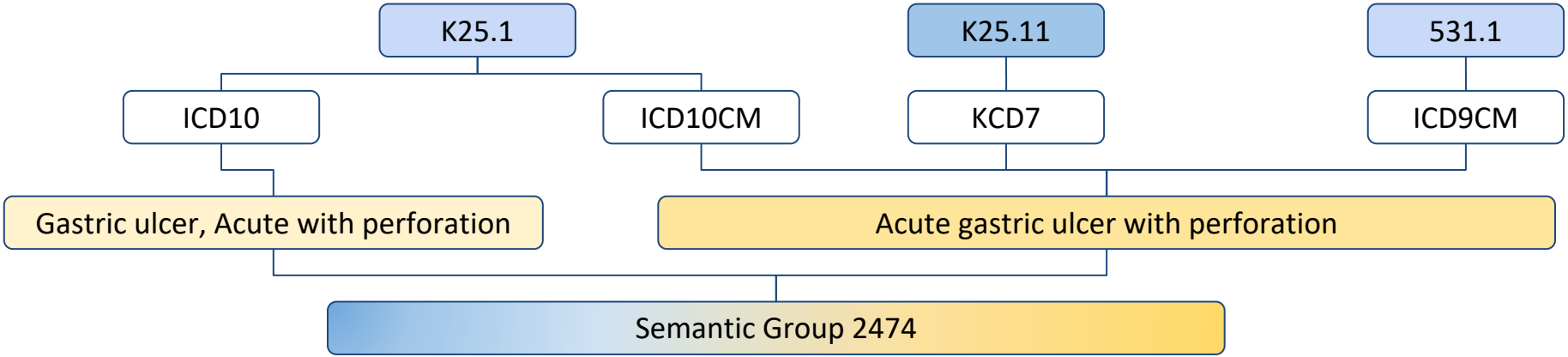
Assemble of sources
- Source agnostic
- Intentional redundancy
- Target agnostic
  - Version agnostic
  - Status agnostic

Grouping
- Multiparameter
- Multilayer

Mapping
- Get the same target

Vocabulary **integration**

CDE example: ICD family

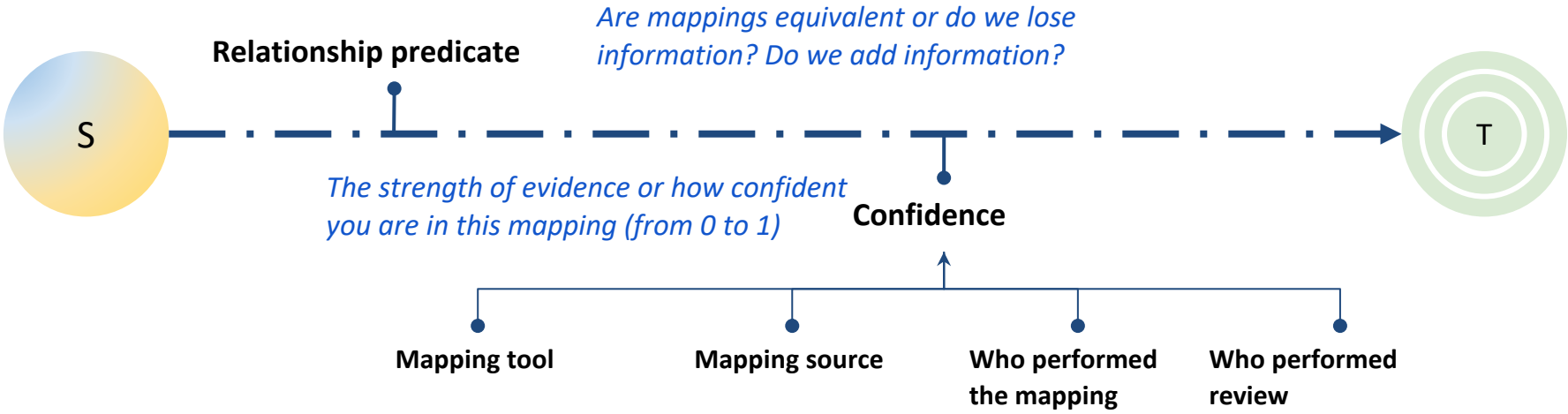


4 codes - Maps to (eq) - 1 target:

4057953	19850005	Acute gastric ulcer with perforation	Disorder	Standard	Valid	Condition	SNOMED

Metadata, inspired by sssom

SIMPLE STANDARD FOR SHARING ONTOLOGY MAPPINGS



Metadata – relationship predicate

The source concept is a **narrower** term than the target concept. Data loss happens. Typical scenario when no exact match can be found.

Maps uphill

Education of patient and **household** providers: **guardian/friend's** education level

↓

Highest level of education of Personnel

Standard full **equivalent** 'Maps to' with no data loss. The two terms are intended to refer to the same thing.

Neither Up nor Down

Maps to equivalent

Abdominal aneurysm

↓

Abdominal aortic aneurysm

Rare scenario when the source concept is **broader** than the target concept. It should not happen generally.

Maps downhill

Amount of smoking

↓

Tobacco amount per day

Metadata in MedDRA

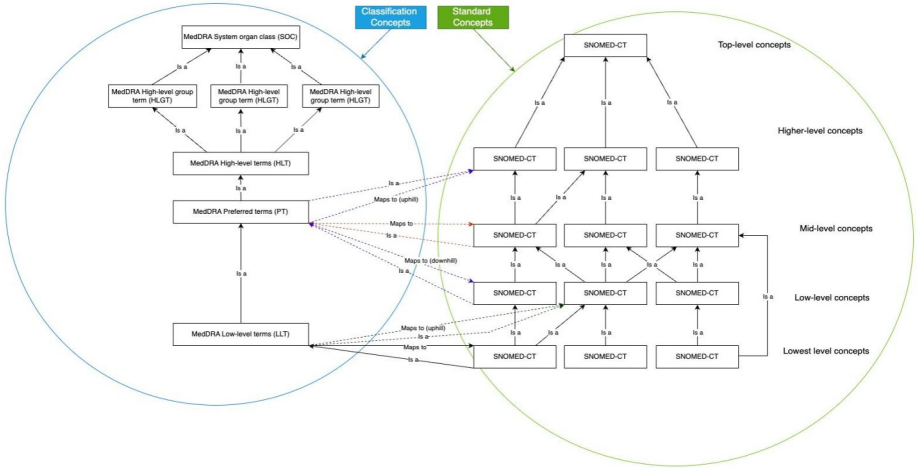
Description	Concept code	Concept name	Vocabulary	relationship_id	relationship_id _predicate	Concept code	Concept name	Vocabulary
Standard full equivalent 'Maps to' with no data loss. The two terms are intended to refer to the same thing	10001389	Adrenocortical insufficiency acute	MedDRA	Maps to	eq	766986002	Acute adrenal insufficiency	SNOMED
	10050701	Congenital pulmonary hypertension	MedDRA	Maps to	eq	1010627004	Pulmonary hypertension due to developmental abnormality	SNOMED
The source concept is a narrower term than the target concept. Data loss happens. Typical scenario when no exact match can be found	10002244	Anastomotic ulcer haemorrhage	MedDRA	Maps to	up	74474003	Gastrointestinal hemorrhage	SNOMED
	10002244	Anastomotic ulcer haemorrhage	MedDRA	Maps to	up	447408004	Ulcer of anastomosis	SNOMED
	10050821	Groin infection	MedDRA	Maps to	up	40733004	Disorder due to infection	SNOMED
	10050821	Groin infection	MedDRA	Maps to	up	118936007	Disorder of inguinal region	SNOMED
Rare scenario when the source concept is broader than the target concept. It should not happen generally if not stated otherwise	10048547	Suture rupture	MedDRA	Maps to	down	217008000	Suture failure during surgical operation	SNOMED
	10050681	Epstein-Barr virus test	MedDRA	Maps to	down	408219003	Epstein-Barr virus serology	SNOMED

out of total:

Equivalent 30634

Uphill 1344

Downhill 69



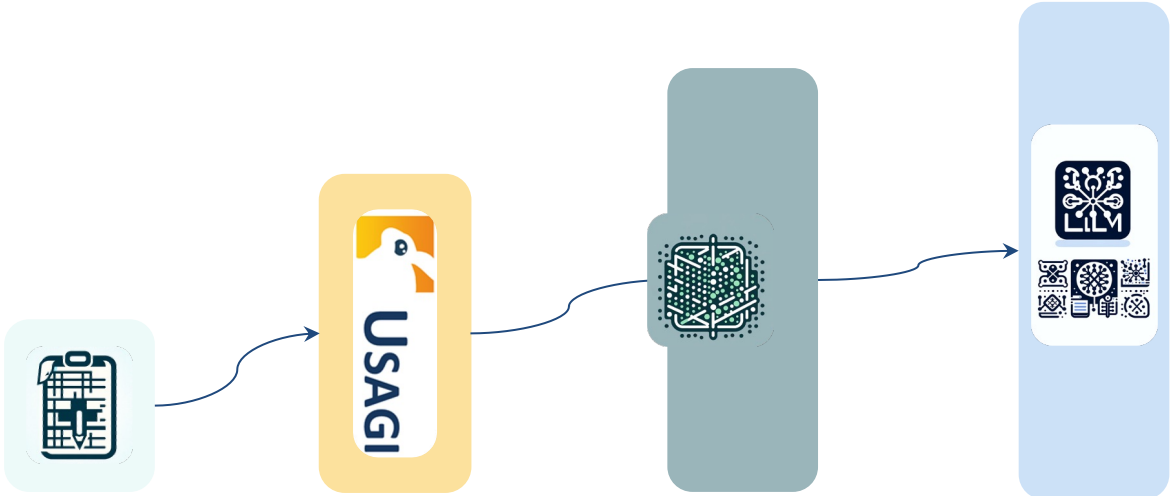
ML/AI – Problem space

- Lots of mapping work
- That requires unique knowledge
- Cost and time constraints: mapping is expensive, slow process

Why is it that **hard to solve**?

- This is the **reasoning** task, ML is still far away from it.
- Highly **specialized** data => there's no good, validated and big enough datasource to learn from.
- Data **heterogeneity**: biomedical data varies widely in terminology and representation, leading to variety of ways to represent the same clinical concept.
- Concept **evolution**: continuous medical knowledge updates, both source and standard lifecycle, changing conventions.
- Vocabulary **volume**: computationally heavy task (400k possible targets * 100 objects = 40M).

Evolution of mapping approach in OMOP



Manual mapping

TF-IDF statistics

Classic neural networks

Transformers
LLMs
Knowledge graphs

Mapping continuum

Categories of sources

Controlled vocabulary

Abdominal pain after abortion
Calyceal fistula
Cervical shortening, second trimester

Free-text

A FIB, CAD, PE
DIARRHEA, JAW TIGHTNESS, HEADACHES, CHEST PAIN
24 hour urine protein output

Complex terms

Cancer | Yes | Active: No | Remission: Yes | Origin: Other | Histology: Adenocarcinoma
DEFINITY CONTRAST DMINISTERED IV PER PROTOCOL FOR LV OPACIFICATION
FUS 2-7 T JT W INTBD FUS DEV, POST APPR P COL, OPN
Swollen Indicator|METATARSOPHALANGEAL JOINT 1|RIGHT
RIBOSOMAL P AB.SER/PLAS.QN (AI)

Different methods and accuracy

	Usagi	Embedders: BioWordVec	Embedders: BioSentVec	Enhanced approach: BioSentVec + in-home NN	Enhanced approach: (TF-IDF, Fuzzy ratio, BioWordVec) + ChatGPT
Controlled vocabulary	58%	82%	81%	70%	68%
Free-text	67%	70%	68%	64%	63%
Complex terms	5%	9%	11%	13%	31%

AI drops accuracy for trivial mappings

Auto-mapping pipeline