



**Stefan  
Schulz**

Medical  
University  
of Graz  
(Austria)

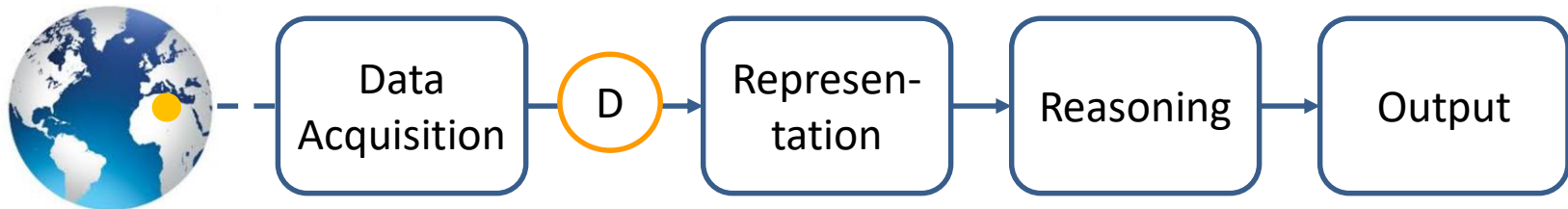
[purl.org/steschu](http://purl.org/steschu)



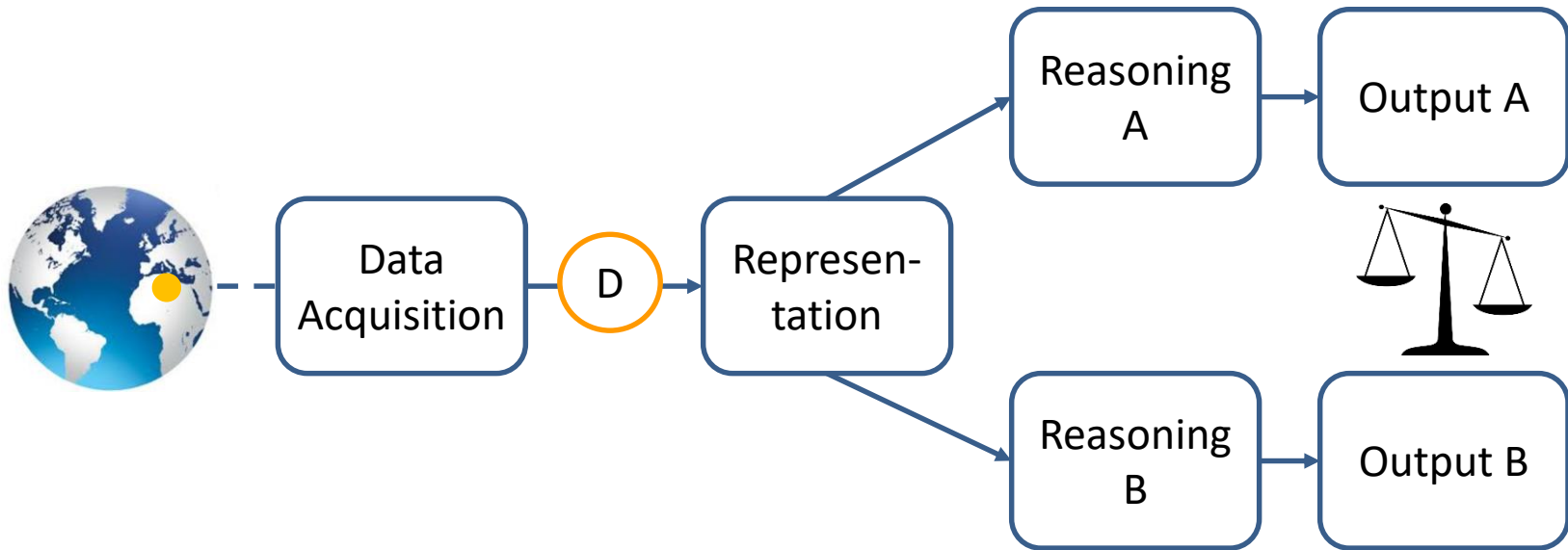
Keynote address:

**Annotating clinical narratives with  
SNOMED CT:  
The thorny way towards  
interoperability of  
clinical routine data**

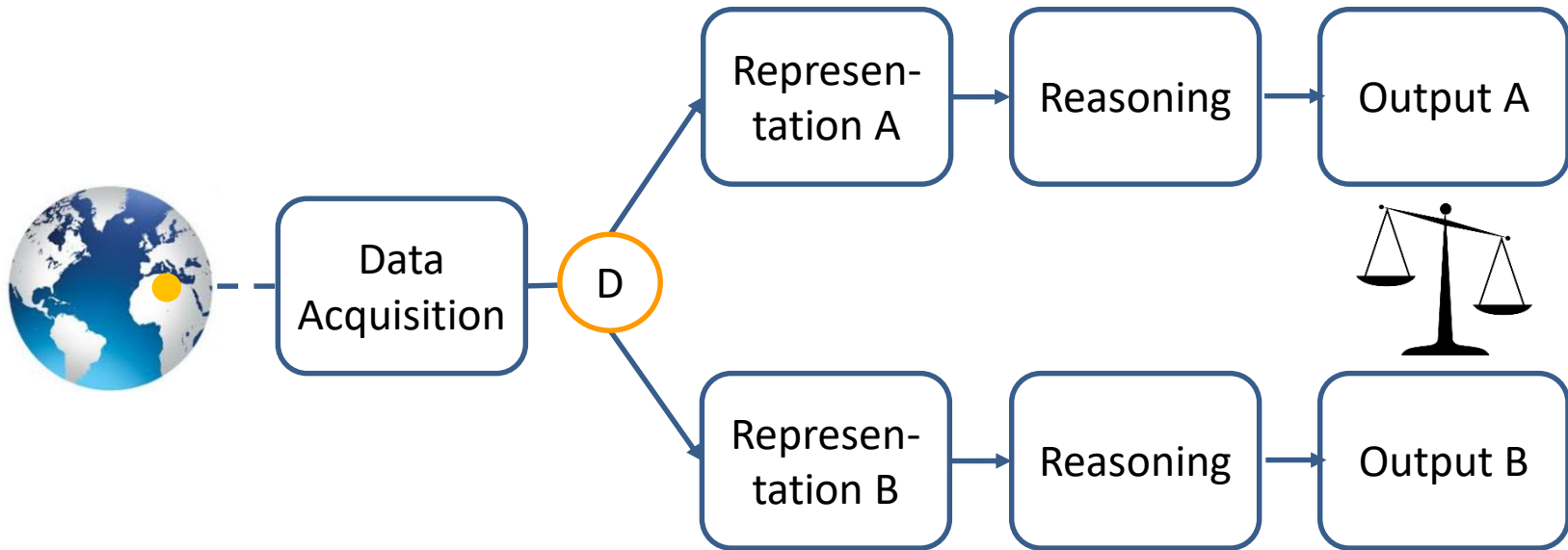
# "Classical" AI workflow



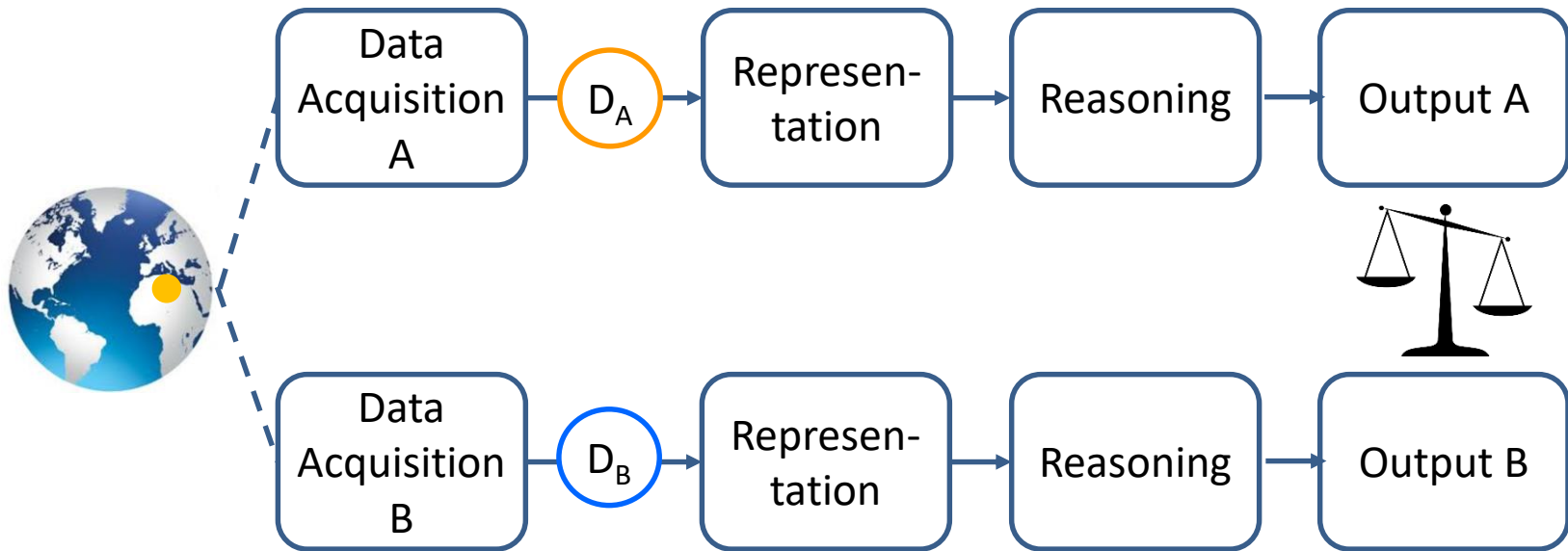
# "Classical" AI workflow



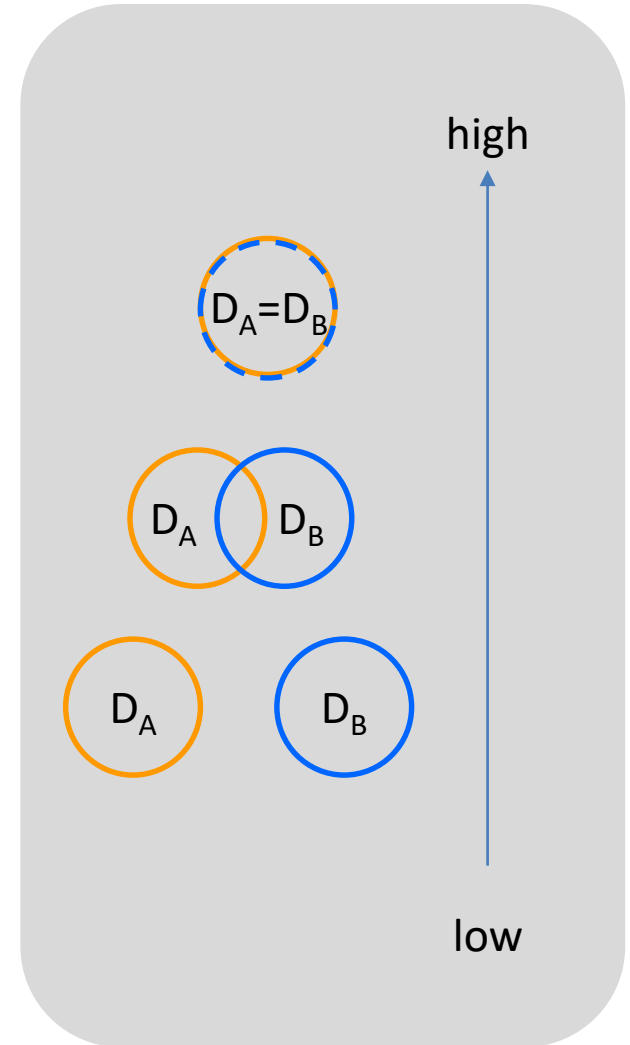
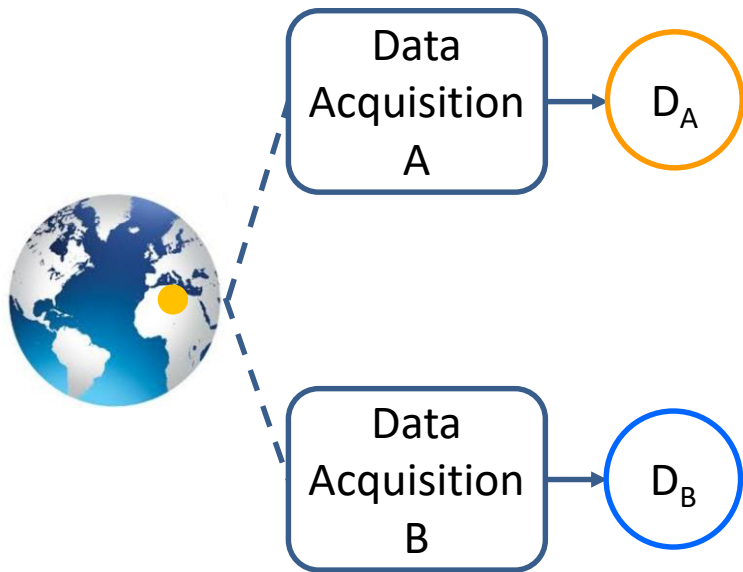
# "Classical" AI workflow



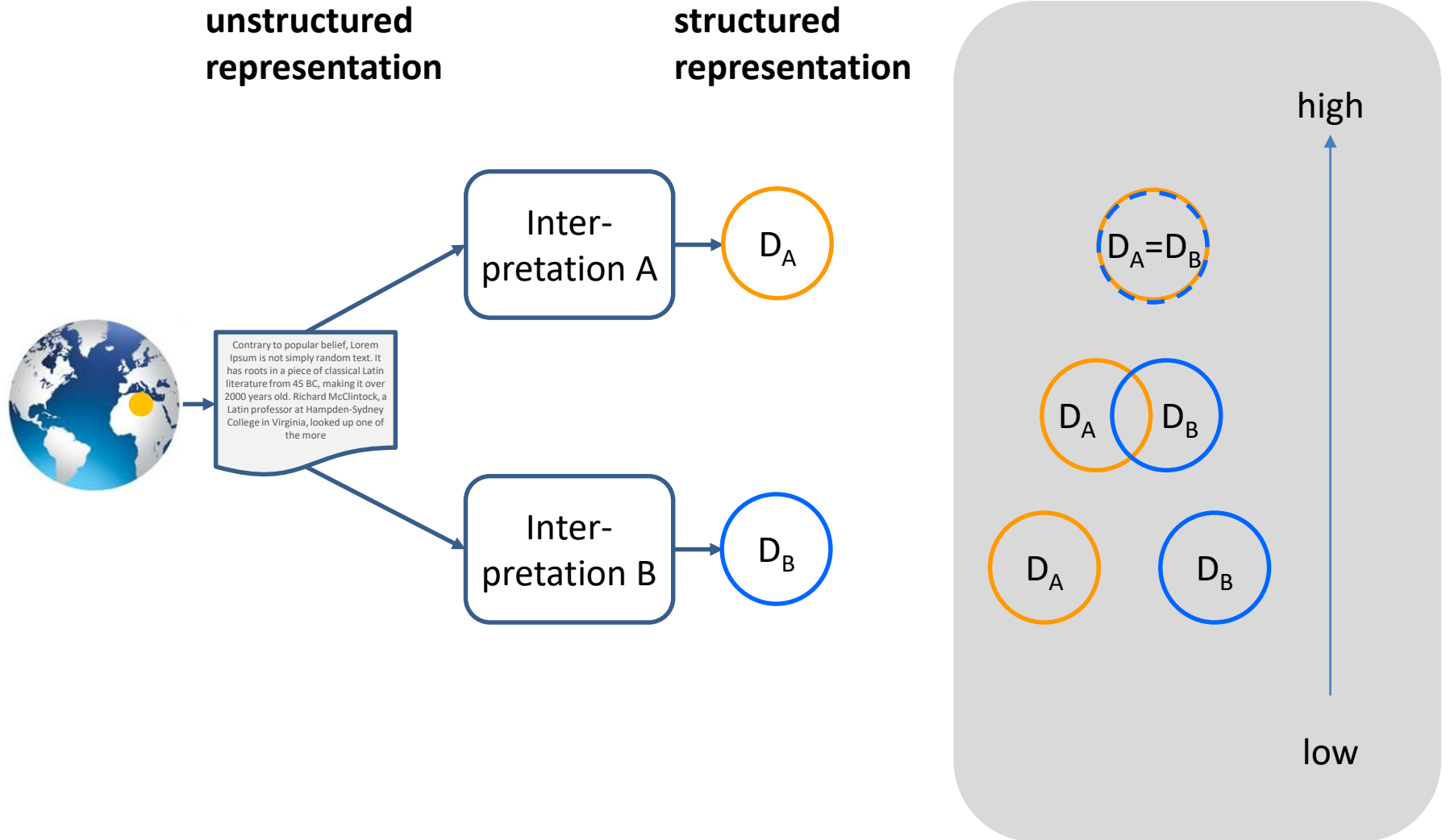
# "Classical" AI workflow



# Data reliability $\rightarrow$ Data interoperability



# Data reliability → Data interoperability



# Focus of the talk

- Structured extracts from unstructured clinical data: reliability and interoperability
- Empirical study on inter-annotator agreement
- Analysis of examples for inter-annotator disagreement
- Mechanisms to improve agreement

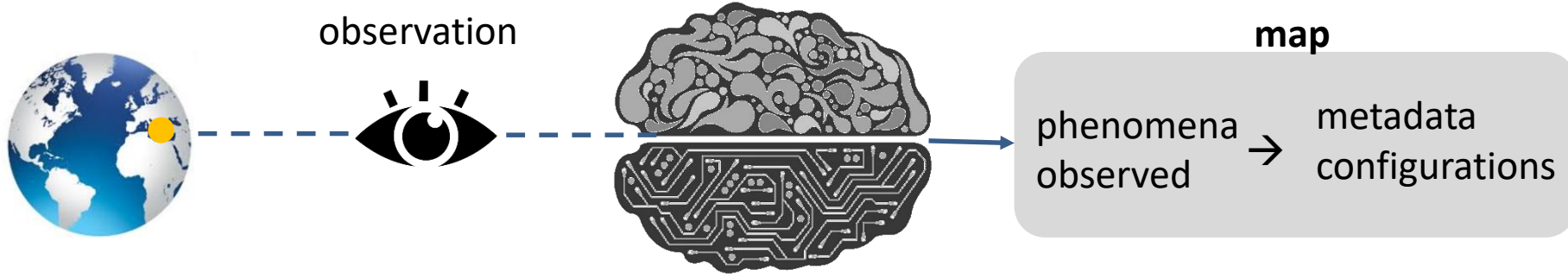
- better data reliability
- better interoperability
- better training data
- better gold standards



# Annotating clinical narratives with SNOMED CT

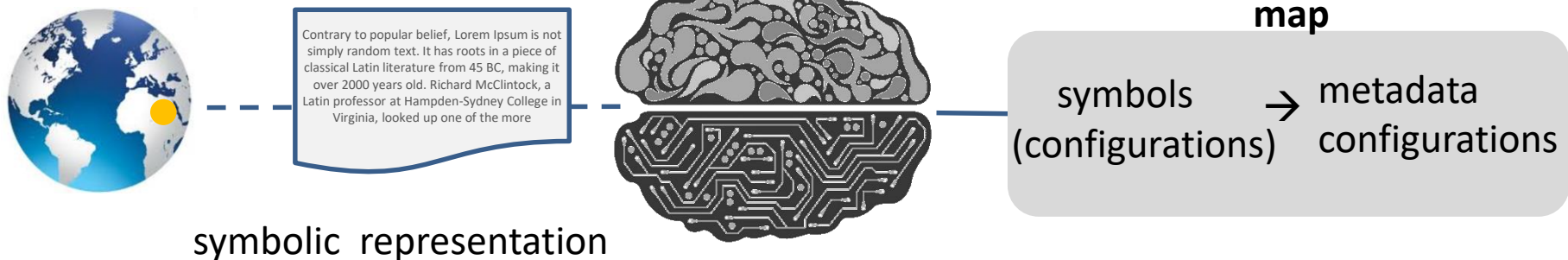
# Annotating clinical narratives with SNOMED CT

## Coding



Vocabulary

## Annotation



# Annotating clinical narratives with SNOMED CT

Huge clinical  
reference  
terminology

eHealth standard,  
maintained by  
transnational SDO

~300,000  
"concepts"

representable as  
OWL EL

SNOMED CT

preferred terms  
and synonyms in  
several languages

(quasi-)  
ontological  
definitional and  
qualifying axioms

multiple  
hierarchies

covers disorders,  
procedures, body parts,  
substances, devices,  
organisms, qualities...

# Annotation: Sources of complexity

Contrary to popular belief, Lorem Ipsum is not simply random text. It has roots in a piece of classical Latin literature from 45 BC, making it over 2000 years old. Richard McClintock, a Latin professor at Hampden-Sydney College in Virginia, looked up one of the more

Map

SNOMED CT

## Clinical narrative

- sequence of Tokens
- syntactic structures
- relations at various levels

## Ontology

- entities, codes
- relations
- logical constructors
- axioms

## Terminology

- preferred terms
- synonyms
- definitions

- Compactness
- Agrammaticality
- Short forms
- Implicit contexts

best text span to annotate?  
Naïve or analytic annotation?

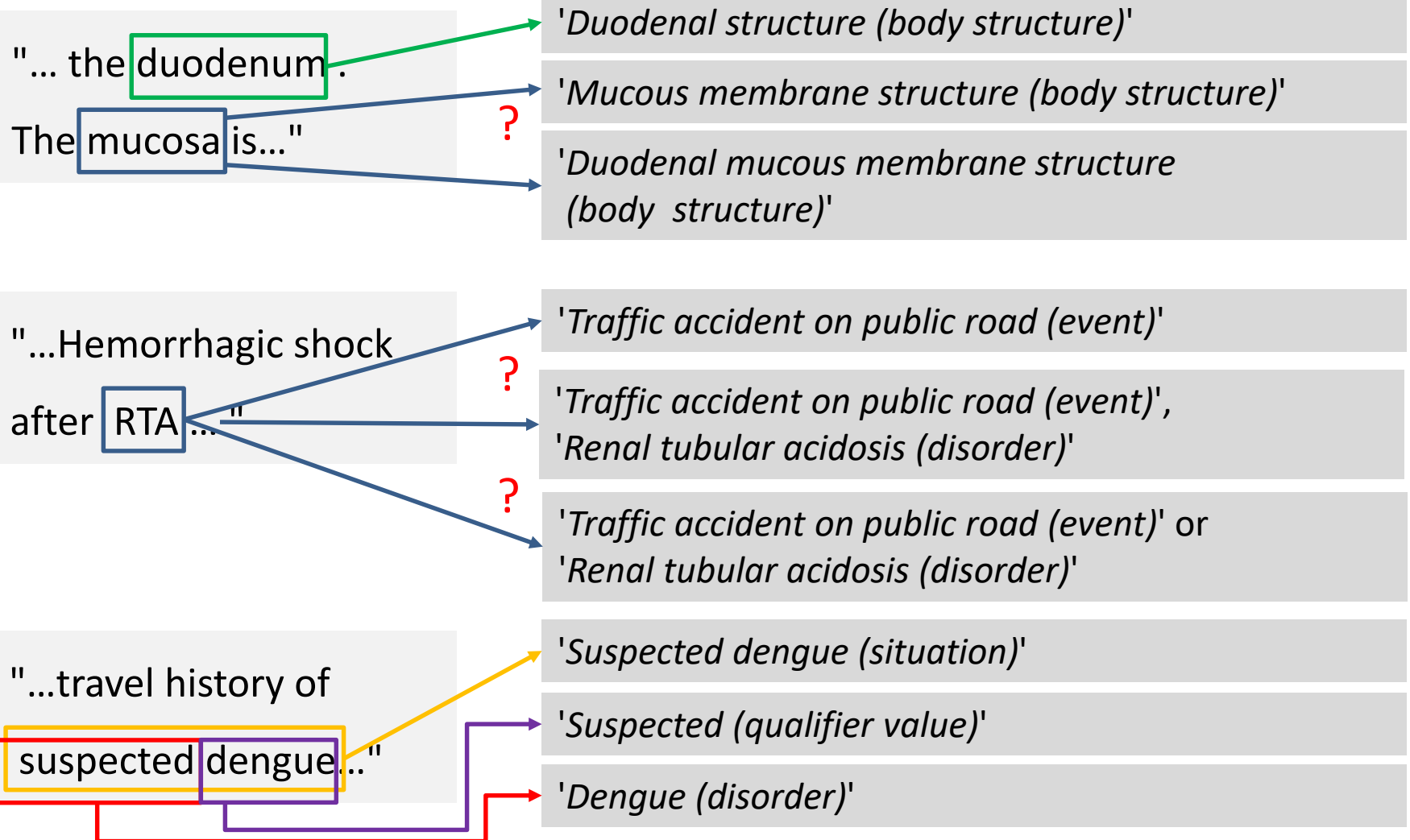
- Ill-defined concepts
- Similar concepts
- Pre-coordination vs. post-coordination

Complex annotations (> 1 concept)  
Degree of formality?

# Examples

## Clinical text

## SNOMED CT concepts (FSNs)



# Coding / Annotation guidelines

- Examples:

1. German coding guidelines for ICD and OPS, 171 pages
2. Using SNOMED CT in CDA models: 147 pages
3. CHEMDNER-patents: annotation of chemical entities in patent corpus: annotation manual 30 pages
4. CRAFT Concept Annotation guidelines: 47 pages
5. Gene Ontology Annotation conventions: 7 pages

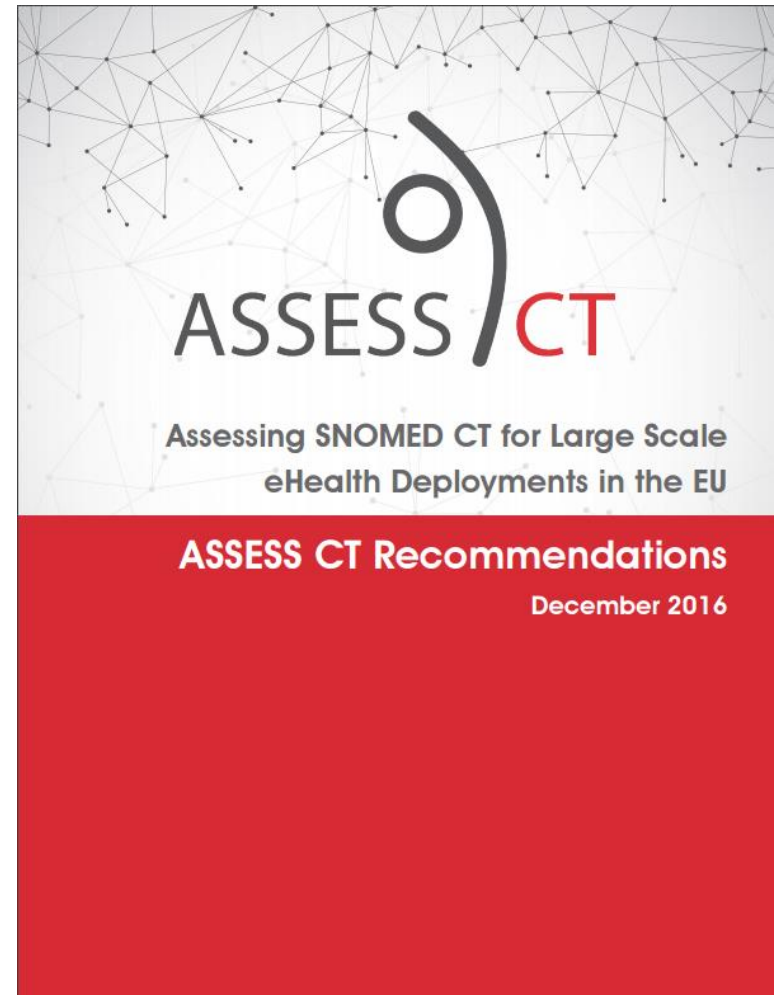
- Complex rule sets, requiring intensive training

1. [http://www.dkgev.de/media/file/21502.Deutsche\\_Kodierrichtlinien\\_Version\\_2016.pdf](http://www.dkgev.de/media/file/21502.Deutsche_Kodierrichtlinien_Version_2016.pdf)
2. <http://www.snomed.org/resource/resource/249>
3. [http://www.biocreative.org/media/store/files/2015/cemp\\_patent\\_guidelines\\_v1.pdf](http://www.biocreative.org/media/store/files/2015/cemp_patent_guidelines_v1.pdf)
4. [http://bionlp-corpora.sourceforge.net/CRAFT/guidelines/CRAFT\\_concept\\_annotation\\_guidelines.pdf](http://bionlp-corpora.sourceforge.net/CRAFT/guidelines/CRAFT_concept_annotation_guidelines.pdf)
5. <http://geneontology.org/page/go-annotation-conventions>

# Annotation experiments in ASSESS-CT

# Annotation experiments in ASSESS-CT

- EU project on the fitness of purpose of SNOMED CT as a core reference terminology for the EU: [www.assess-ct.eu](http://www.assess-ct.eu)  
Feb 2015 – Jul 2016
- Scrutinising clinical, technical, financial, and organisational aspects of reference terminology introduction
- Summary of results: brochure published, scientific papers to appear





# Annotation of clinical narratives

- Comparing
  - **SNOMED CT** vs.
  - **UMLS** derived terminology
- Resources
  - Parallel corpus: 60 **clinical text snippets** from 6 languages, high diversity
  - For each language: 2 annotators \* 40 samples → **20 snippets annotated twice**
  - Annotators
    - trained by webinars
    - follow annotation guideline (10 pages)

Nitroglycerin pump spray as required	387404004;385074009;225761000
Amantadine bds	372763006;229799001
Allopurinol 300 ½ tablet every other day (last dose on	387135004;385055001;225760004

- e.g.
- chunking into noun phrases
- annotation of chunks by sets of codes
- give preference to maximally pre-coordinated codes
- understanding text and assign maximally specific codes

Normal mucous membranes in mouth pharynx and on the larynx.	17621005;33044003;71248005
Hyoid and thyroid cartilage are intact.	21387005;52940008;11163003
Fragmental fractures of the two upper vertebrae of the cervical spine.	13321001;207984009;207983003
Otherwise the cervical spine is intact.	122494005;11163003
Oesophagus as well as trachea are torn at the lower end of the neck.	262793000;282459005;261122009;123958008

# Principal quantitative results (English)

Concept coverage [95% CI]	SNOMED CT	Alternative
Text annotations – English	.86 [.82-.88]	.88 [.86-.91]



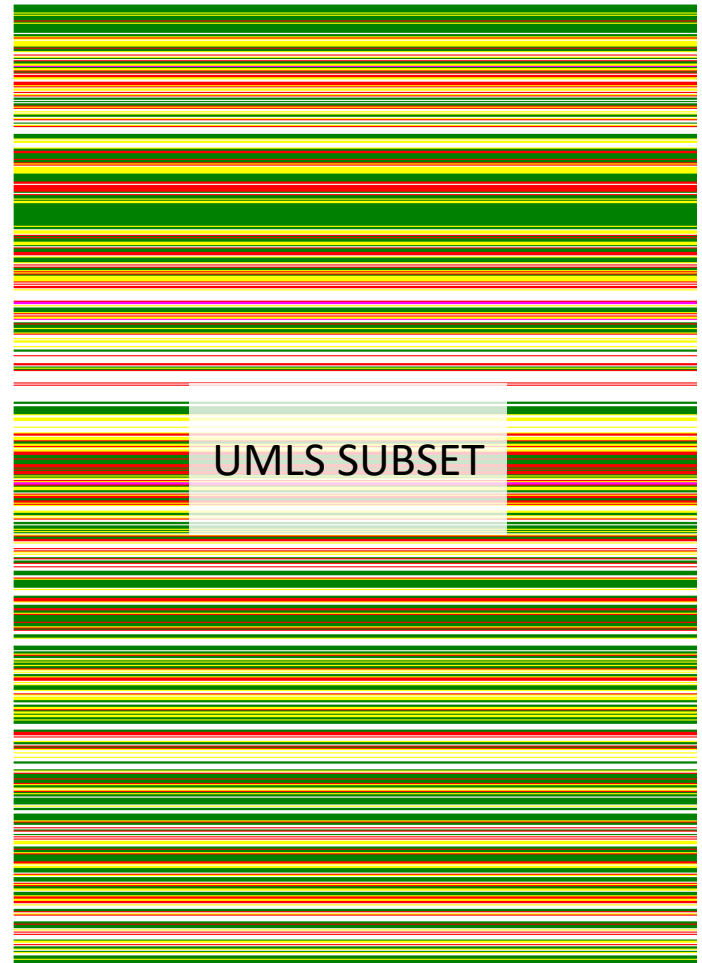
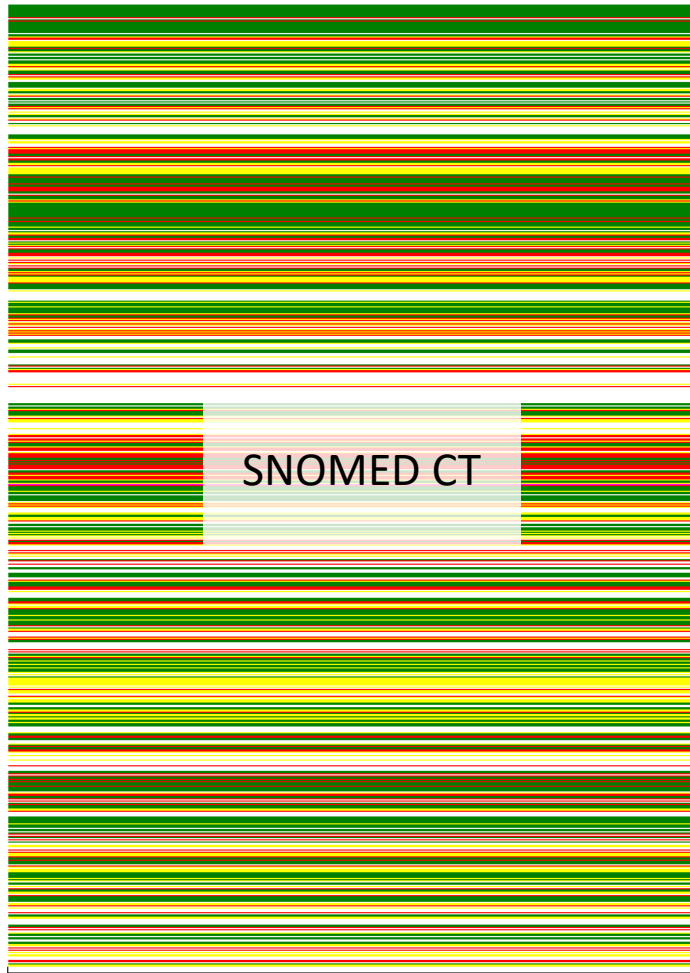
Term coverage [95% CI]	SNOMED CT	Alternative
Text annotations – English	.68 [.64; .70]	.73 [.69; .76]



Inter annotator agreement Krippendorff's Alpha [95% CI]	SNOMED CT	Alternative
Text annotations	.37 [.33-.41]	.36 [.32-.40]



# Agreement map: text annotations (English)



green: agreement – yellow: only annotated by one coder – red: disagreement

# Systematic error analysis

- Creation of gold standard for SNOMED CT
  - 20 English text samples annotated twice → 208 NPs
  - Analysis of English SNOMED CT annotations by two additional terminology experts
  - Consensus finding, according to pre-established annotation guidelines
- Inspection, analysis and classification of text annotation disagreements
- Presentation of some disagreement cases for SNOMED CT

# Reasons for disagreement

# Human issues

- Lack of domain knowledge / carelessness

Tokens	Annotator #1	Annotator #2	Gold standard
"IV"	'Structure of abductor hallucis muscle (body structure)'	'Abducens nerve structure (body structure)'	'Abducens nerve structure (body structure)'

- Retrieval error (synonym not recognised)




Tokens	Annotator #1	Annotator #2	Gold standard
"Glibenclamide"	'Glyburide (substance)'	–	'Glyburide (substance)'

- Non-compliance with annotation rules

# Ontology issues (I)

## ■ Polysemy ("dot categories")\*

Tokens	Annotator #1	Annotator #2	Gold standard
'Lymphoma'	'Malignant lymphoma (disorder)'	'Malignant lymphoma - category (morphologic abnormality)'	'Malignant lymphoma (disorder)'

 Malignant lymphoma (disorder)  

SCTID: 118600007

118600007 | Malignant lymphoma (disorder) |

- Lymphoma (clinical)
- Malignant lymphoma (clinical)
- Malignant lymphoma (disorder)
- Malignant lymphoma
- Lymphoma

Associated morphology →  
Malignant lymphoma - category

\*Alexandra Arapinis, Laure Vieu: A plea for complex categories in ontologies. Applied Ontology 10(3-4): 285-296 (2015)

# Ontology issues (I)

- Polysemy ("dot categories")\*

<b>Tokens</b>	<b>Annotator #1</b>	<b>Annotator #2</b>	<b>Gold standard</b>
'Lymphoma'	'Malignant lymphoma (disorder)'	'Malignant lymphoma - category (morphologic abnormality)'	'Malignant lymphoma (disorder)'

- "Pseudo-polysemy"

- Incomplete definitions

<b>Tokens</b>	<b>Annotator #1</b>	<b>Annotator #2</b>	<b>Gold standard</b>
"Former	'In the past (qualifier value)'	'History of (contextual qualifier) (qualifier value)'	'Ex-smoker (finding)'
Smoker"	'Smoker (finding)'	'Smoker (finding)'	




\*Alexandra Arapinis, Laure Vieu: A plea for complex categories in ontologies. Applied Ontology 10(3-4): 285-296 (2015)



# Ontological issues (II)

## ■ Incomplete definitions

Tokens	Annotator #1	Annotator #2	Gold standard
"Motor:	<i>'Skeletal muscle structure (body structure)'</i>	<i>'Muscle finding (finding)'</i>	<i>'Skeletal muscle normal (finding)'</i>
normal bulk and tone"	<i>'Normal (qualifier value)'</i>	<i>'Normal (qualifier value)'</i>	

 **Skeletal muscle normal (finding)**  

SCTID: 298300008

298300008 | Skeletal muscle normal (finding) |

Skeletal muscle normal  
Skeletal muscle normal (finding)

Finding site → Skeletal muscle structure

# Ontological issues (II)

## ■ Normal findings, incomplete definitions




Tokens	Annotator #1	Annotator #2	Gold standard
"Motor: normal bulk and tone"	'Skeletal muscle structure (body structure)'	'Muscle finding (finding)'	'Skeletal muscle normal (finding)'
	'Normal (qualifier value)'	'Normal (qualifier value)'	

## ■ Fuzziness of qualifiers

Tokens	Annotator #1	Annotator #2	Gold standard
"Significant bleeding"	'Significant (qualifier value)'	'Severe (severity modifier) (qualifier value)'	'Moderate (severity modifier) (qualifier value)'
	'Bleeding (finding)'	'Bleeding (finding)'	'Bleeding (finding)'

# Interface term (synonym) issues

Tokens	Annotator #1	Annotator #2	Gold standard
"Blood	'Blood (substance)'	'Hemorrhage (morphologic abnormality)'	'Hemorrhage (morphologic abnormality)'
extravasation"	'Extravasation (morphologic abnormality)'		
		"extravasation of blood"	

 Hemorrhage (morphologic abnormality)  

SCTID: 50960005

50960005 | Hemorrhage (morphologic abnormality) |

- Hemorrhage
- Extravasation of blood
- Blood loss
- Bleeding
- Haemorrhage
- Hemorrhage (morphologic abnormality)

No attributes

# Interface term (synonym) issues

Tokens	Annotator #1	Annotator #2	Gold standard
"Blood	'Blood (substance)'	'Hemorrhage (morphologic abnormality)'	'Hemorrhage (morphologic abnormality)'
extravasation"	'Extravasation (morphologic abnormality)'		
		"extravasation of blood"	

Tokens	Annotator #1	Annotator #2	Gold standard
"anxious"	'Anxiety (finding)'	'Worried (finding)'	'Anxiety (finding)'
		"anxious cognitions"	

# Language issues

- Ellipsis / anaphora
  - "Cold and wind are provoking factors." (provoking factors for angina)
  - "These ailments have substantially increased since October 2013" (weakness)
  - "No surface irregularities" (breast)
  - "Significant bleeding" (intestinal bleeding)
- Ambiguity of short forms
  - "IV" (intravenous? Fourth intracranial nerve?)
- Co-ordination:
  - "normal factors 5, 9, 10, and 11"
- Scope of negation
  - "no tremor, rigidity or bradykinesia"

- Addressed by annotation guideline
- Manageable by human annotators
- Known challenges for NLP systems

# Prevention and remediation of annotation disagreements

# Prevention: annotation processes

- Training with continuous feedback
  - Early detection of inter annotator disagreement triggers guideline enforcement / guideline revision
- Tooling
  - Optimised concept retrieval (fuzzy, substring, synonyms)
  - Guideline enforcement by appropriate tools
  - Postcoordination support (complex syntactic expressions instead of grouping of concepts)
  - Anti-patterns, e.g. avoid unrelated primitive concepts (?)

# Prevention: improve terminology structure

- Fill gaps
  - equivalence axioms (reasoning)
  - Self-explaining labels (FSNs), especially for qualifiers
  - Scope notes / text definitions where necessary
- Manage polysemy
- Flag navigational and modifier concepts
- Strengthen ontological foundations
  - Upper-level ontology alignment
  - Clear division between domain entities and information entities
  - Overhaul problematic subhierarchies, especially qualifiers



# Prevention: improve content maintenance

- Analysis of real data to support terminology maintenance process
  - Harvest notorious disagreements between text passages and annotations from clinical datasets
  - Compare concept frequency and concept co-occurrence between comparable institutions and users to detect imbalances
- Stimulate community processes for ontology-guided content evolution:
  - Crowdsourcing of interface terms by languages, dialects specialties, user groups (separation of interface terminologies from reference terminologies is one of the ASSESS-CT recommendations)

# Remediation of annotation disagreements

# Remediation of annotation disagreements

- Exploit ontological dependencies / implications

Concept A	Concept B	Dependency
'Mast cell neoplasm (disorder)'	'Mast cell neoplasm (morphologic abnormality)'	A subclassOf <b>AssociatedMorphology</b> some B
'Isosorbide dinitrate (product)'	'Isosorbide dinitrate (substance)'	A subclassOf <b>HasActiveIngredient</b> some B
'Palpation (procedure)'	'Palpation - action (qualifier value)'	A subclassOf <b>Method</b> some B
'Blood pressure taking (procedure)'	'Blood pressure (observable entity)'	A subclassOf <b>hasOutcome</b> some B
'Increased size (finding)'	'Increased (qualifier value)'	A subclassOf <b>isBearerOf</b> some B
'Finding of heart rate (finding)'	'Heart rate (observable entity)'	A subclassOf <b>Interprets</b> some B

# Experiment

- Gold standard expansion:
  - Step 1: include concepts linked by attributive relations:
    - *A subclassOf Rel some B*
  - Step 2: include additional first-level taxonomic relations:
    - *A subclassOf B*

Language of text sample	Gold standard expansion	F measure
English	no expansion	0.28
	expansion step 1	0.28
	expansion step 2	0.29

- only insignificant improvement
- possibly due to missing relations in SNOMED CT, e.g. haemorrhage - blood

# Conclusion (I)

- Low inter-annotator agreement limits successful use of clinical terminologies / ontologies
  - for manual annotation scenarios
  - for benchmarking of NLP-based annotations
  - for optimised training data for ML
- Structured data essential for many intelligent systems, but unreliable information extracted from clinical narratives raises patient safety issues when used for decision support

# Conclusion (II)

- Prevention of disagreements
  - Education, tooling, guideline support
  - Terminology content improvement: labelling, scope notes, ontological clarity, full definitions, community processes
  - High coverage interface terminologies
- Remediation of disagreements
  - So far no clear evidence of ontology-based resolution of agreement issues
  - Big data approaches ?

# Conclusion (III)

- R & D required:
  - "Learning systems" for improvement terminology content / structure / tooling. Clinical "big data" underused resource
  - Harmonization of annotation guideline creation and validation efforts
  - Formulate and enforce good quality criteria for clinical terminologies used as annotation vocabularies
  - Better ontological underpinning of clinical terminologies
  - Ontologically founded patterns for recurring clinical documentation tasks: Information extraction rather than concept mapping\*

# Thanks for your attention

- Slides will be accessible via at [purl.org/steschu](https://purl.org/steschu)
- Acknowledgements: ASSESS CT team:  
Jose Antonio Miñarro-Giménez, Catalina Martínez-Costa, Daniel Karlsson, Kirstine Rosenbeck Gøeg, Kornél Markó, Benny Van Bruwaene, Ronald Cornet, Marie-Christine Jaulent, Päivi Hämäläinen, Heike Dewenter, Reza Fathollah Nejad, Sylvia Thun, Veli Stroetmann, Dipak Kalra
- Contact: [stefan.schulz@medunigraz.at](mailto:stefan.schulz@medunigraz.at)



- Vibhu Agarwal, Tanya Podchiyska, Juan M. Banda, Veena Goel, Tiffany I. Leung, Evan P. Minty, Timothy E. Sweeney, Elsie Gyang, Nigam H. Shah:  
Learning statistical models of phenotypes using noisy labeled training data. JAMIA 23(6): 1166-1173 (2016)