Leibniz AILab

Harnessing Big Data for Precision Medicine and Healthcare

Artificial Intelligence

- error-free and versatile
- safe and robust
- transparent, explainable and fair

Personalized Medicine

- more precise diagnoses
- individual therapies
- individual medication

Mining the electronic health record Linguistic and ontological challenges

Stefan Schulz, Medical University of Graz, Austria

June 11, 2021

Conflict of Interest Disclosure

- Professor for Medical Informatics,

  Medical University of Graz, Austria

- Project co-leader

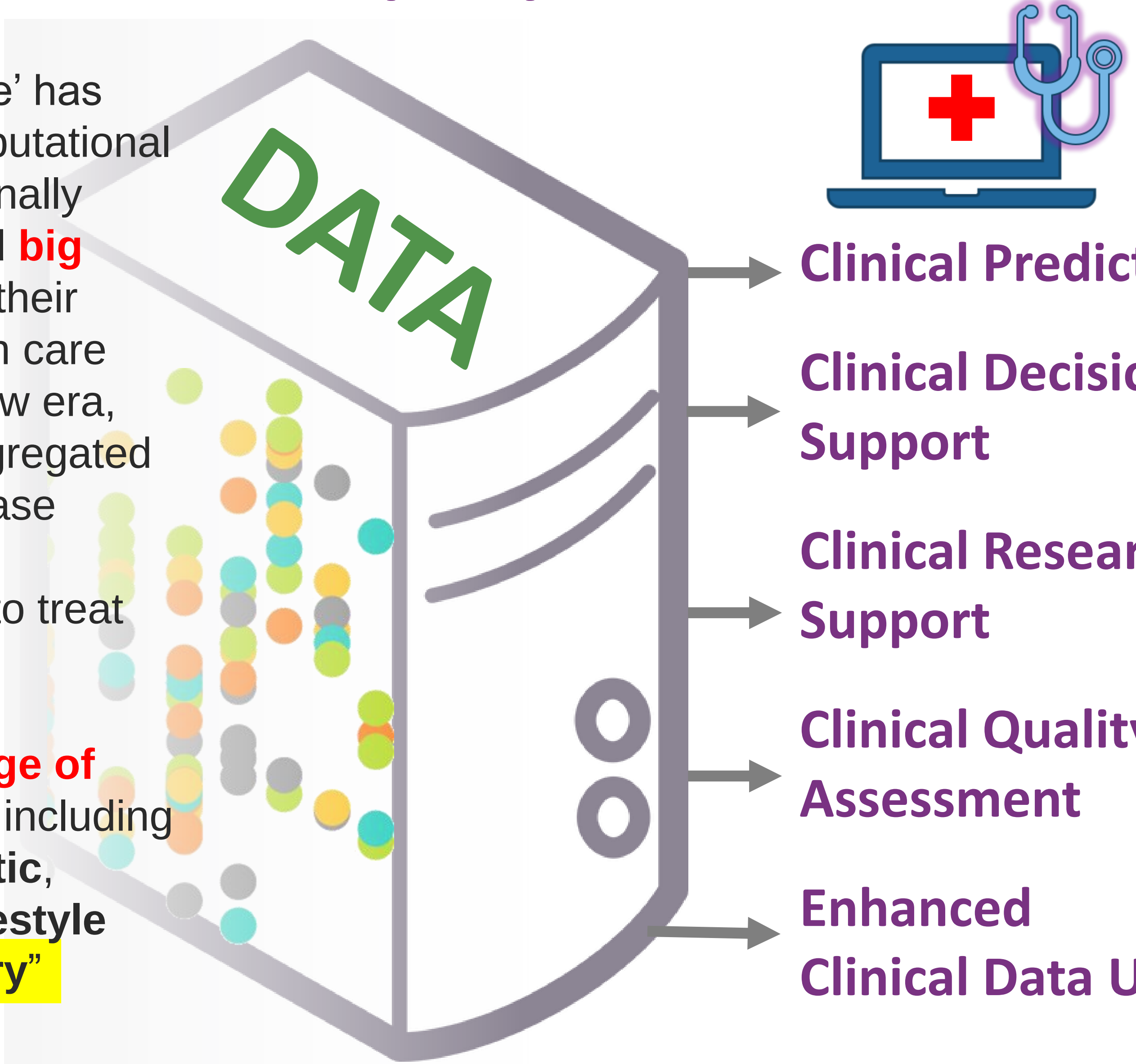
  CBmed Biomarker Research GmbH, Graz Austria

- Head of Medical Research

  Averbis GmbH, Freiburg, Germany

Precision Medicine (PM) is data-centred

"'Precision medicine' has emerged as a computational approach to functionally interpret **omics** and **big data**, and facilitate their application to health care provision. In this new era, patients are not segregated by disease, or disease subtype.
Instead, the aim is to treat every patient as an individual case, incorporating a **range of personalized data** including **genomic**, **epigenetic**, **environmental**, **lifestyle** and **medical history**"



Clinical Prediction

Clinical Decision Support

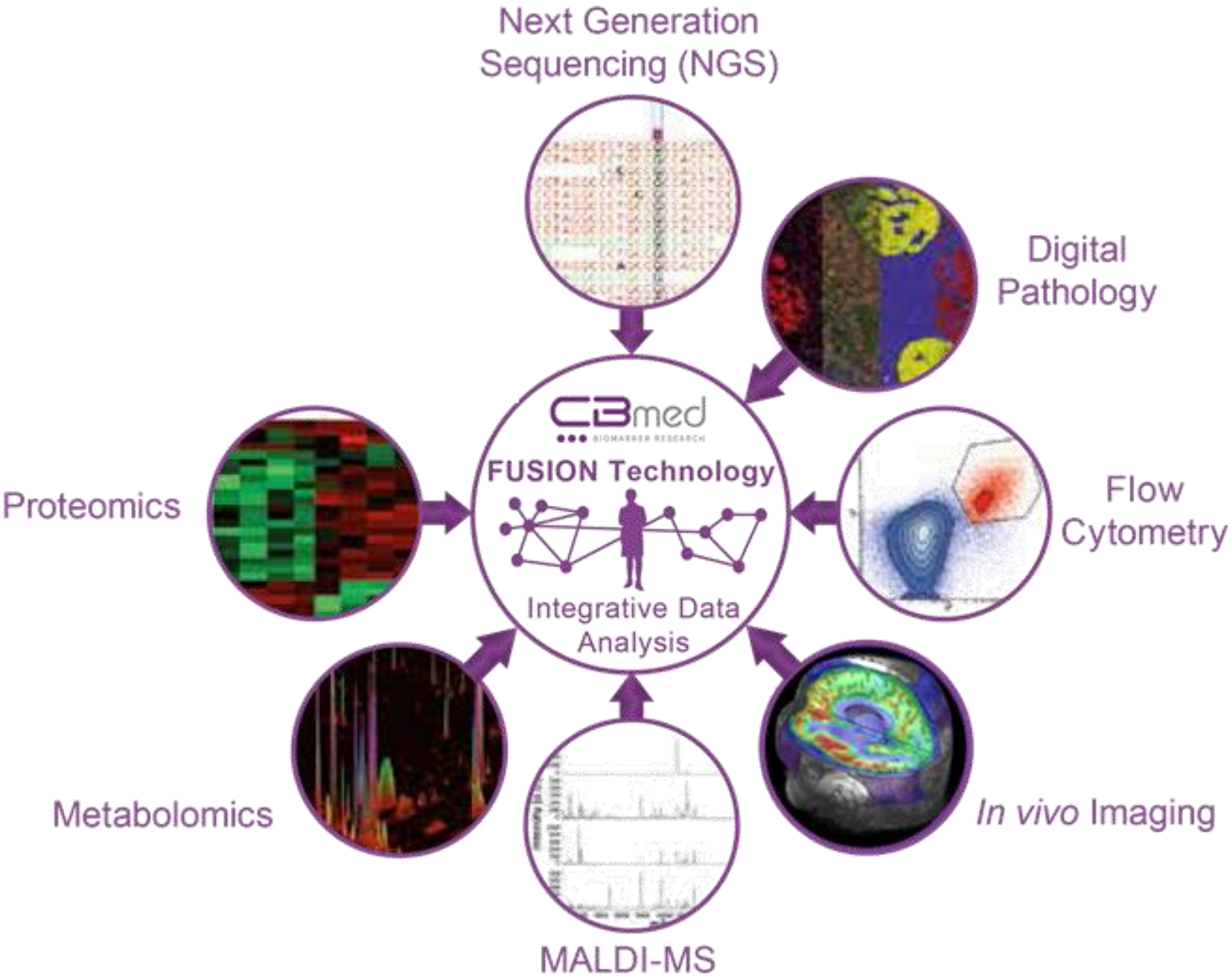
Clinical Research Support

Clinical Quality Assessment

Enhanced Clinical Data Use

David J. Duffy. Problems, challenges and promises: perspectives on precision medicine. Briefings in Bioinformatics, Volume 17, Issue 3, May 2016, Pages 494–504, https://doi.org/10.1093/bib/bbv060

Data as “Fuel” for precision medicine



Source: CBmed – Center for Biomarker Research in Medicine, Graz, Austria

Clinical data

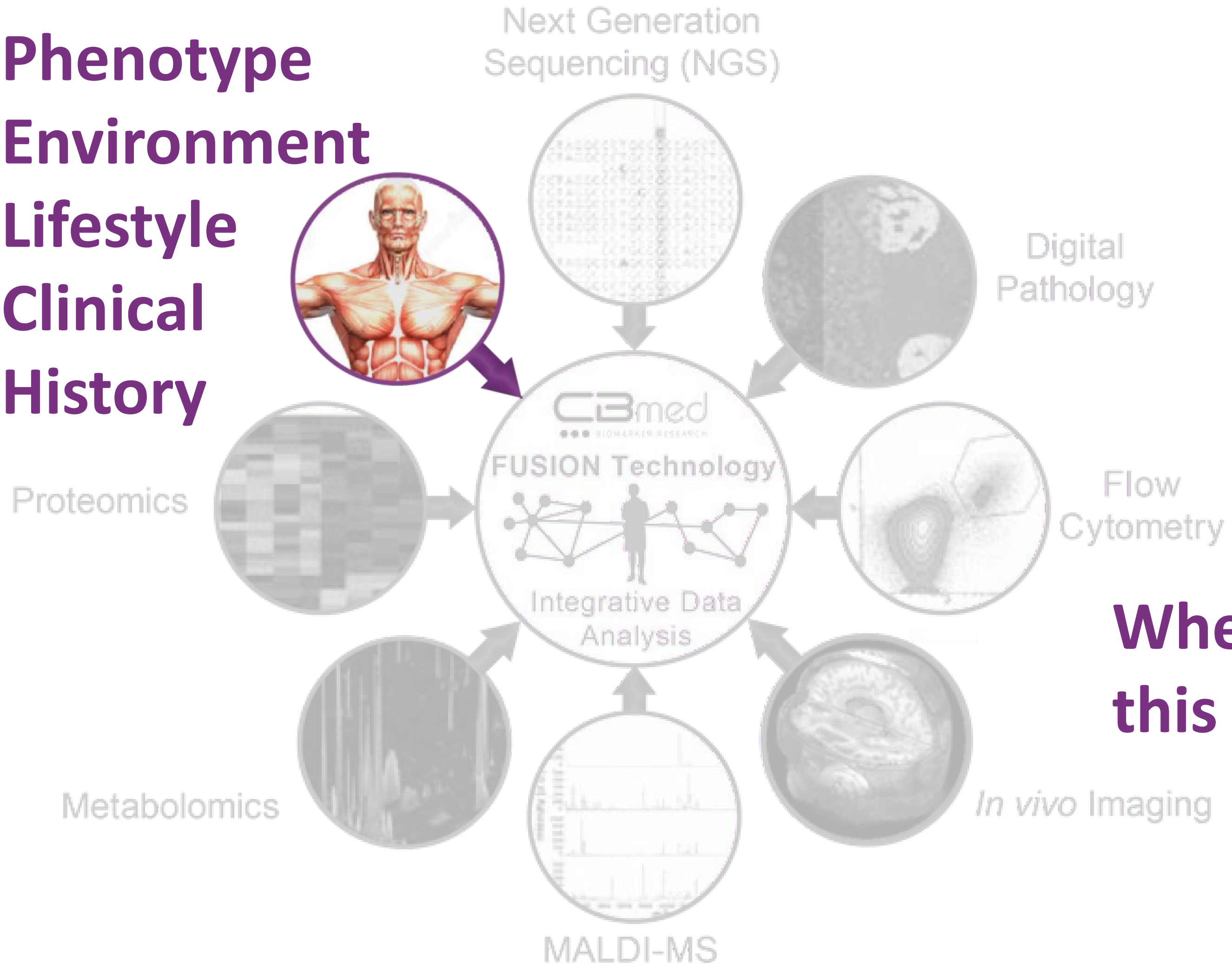
Phenotype

Environment

Lifestyle

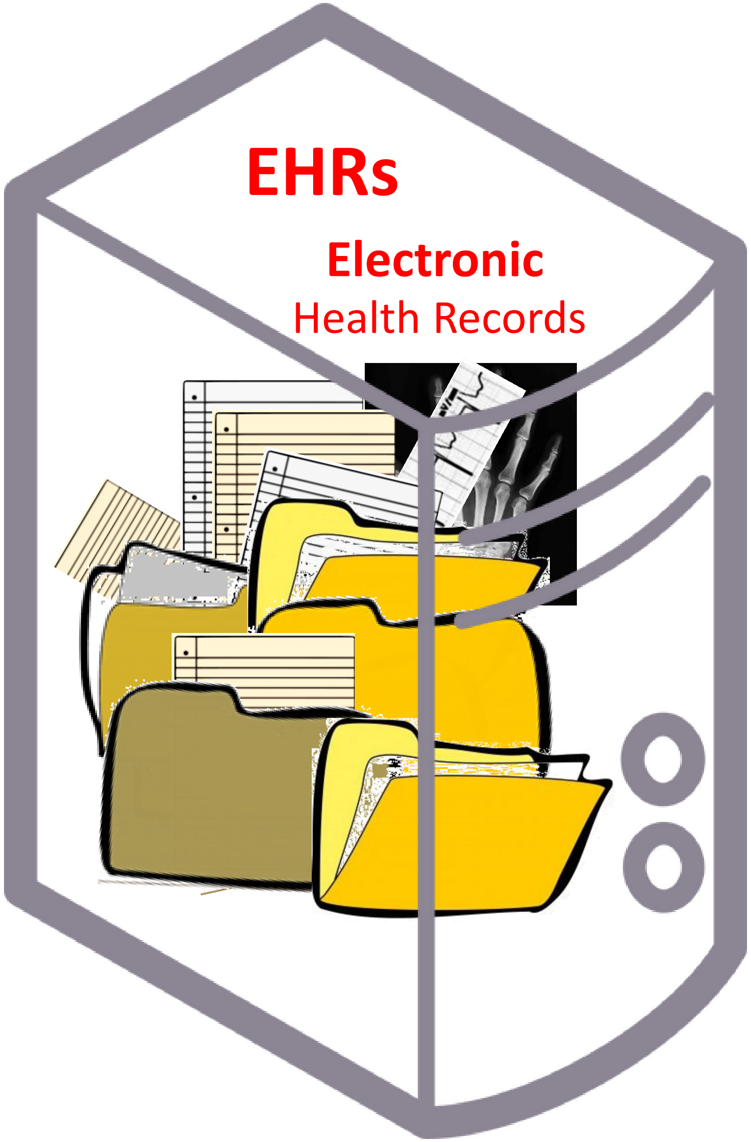
Clinical

History

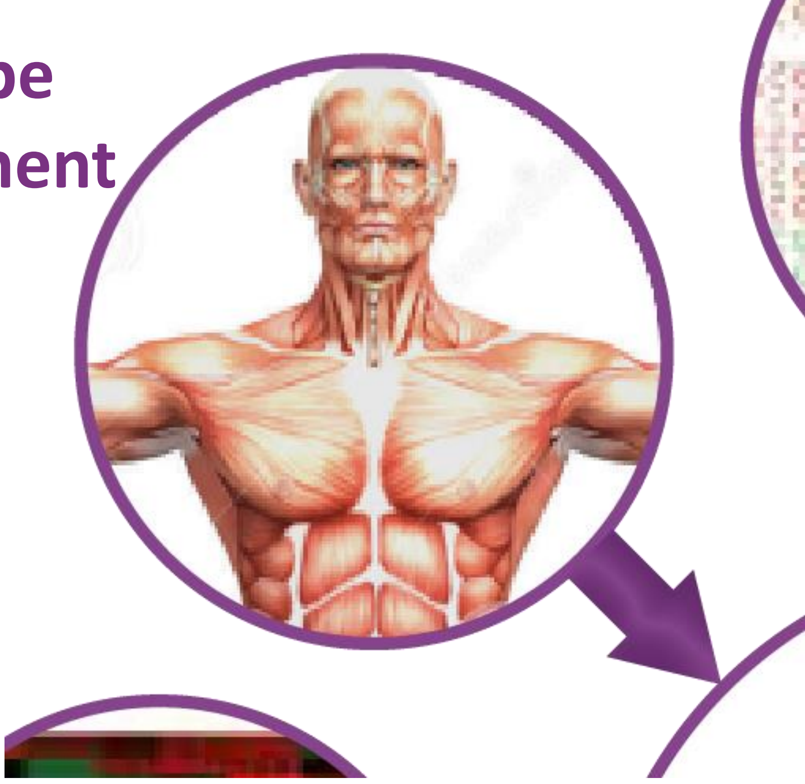


Where is this data?

Source: CBmed – Center for Biomarker Research in Medicine, Graz, Austria

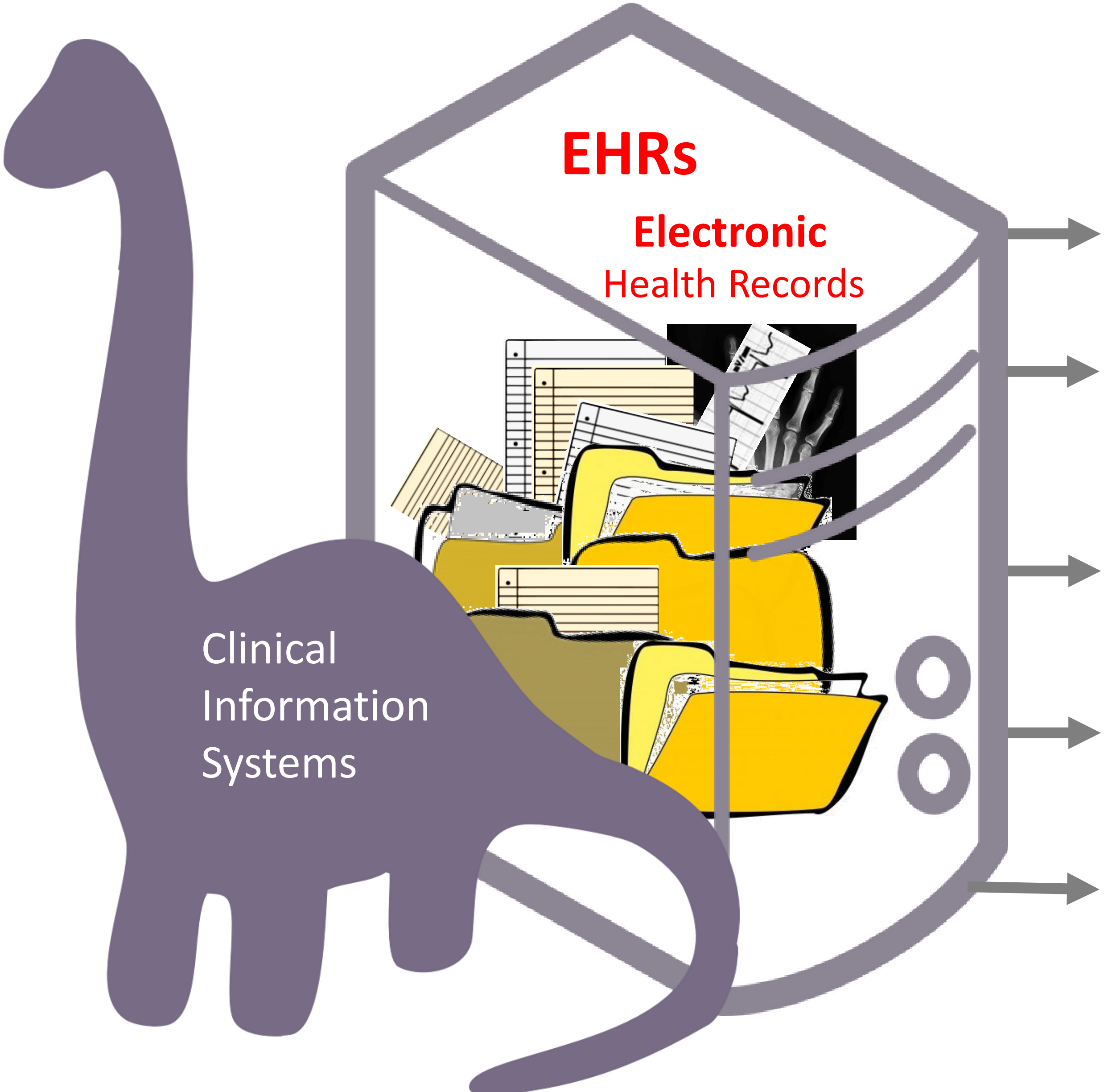


Phenotype
Environment
Lifestyle
Clinical
History



What is in EHRs?

How can it be used for PM?



EHRs

Electronic Health Records

Clinical Information Systems

- Clinical Prediction
- Clinical Decision Support
- Clinical Research Support
- Clinical Quality Assessment
- Enhanced Clinical Data Use

PM requires precision clinical data

Vitality
NUTRITIONALS
UltraPure
DATA
100% NATURAL MSM FROM VEGETABLE OILS
GUARANTEED PURE AND GMO-FREE
500g Powder

PM requires precision clinical data

- **FAIR data:**

  Findable, Accessible, Interoperable, Reusable

- **Barriers:**

  Technical: clinical information systems not designed for data export and secondary use

  Legal / ethical: patient consent, de-identification

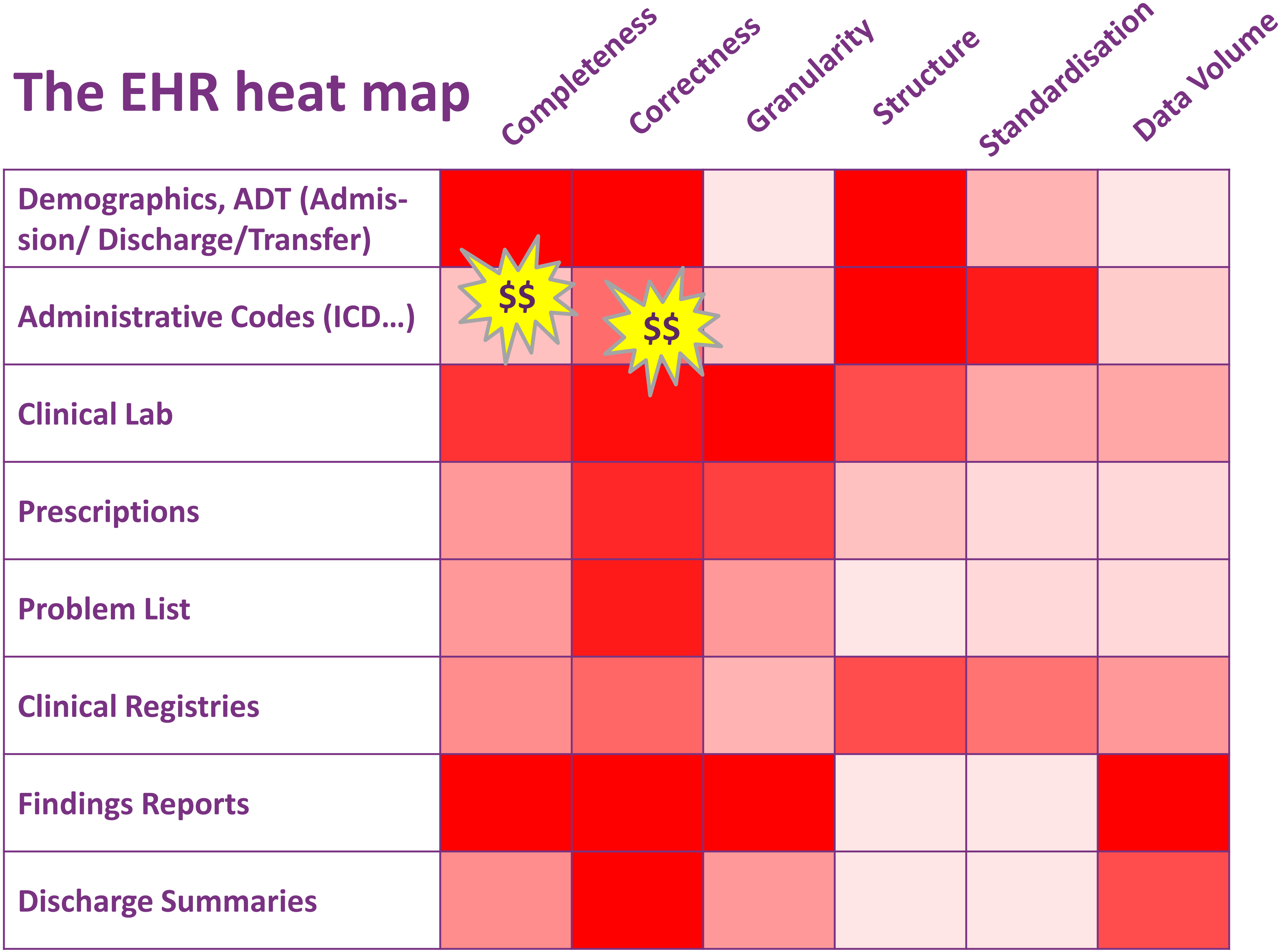
  Structure: Lack of structured data, unstructured data produced for humans, not for machines

  Contexts and provenance: data generation workflows, data creators, intent, motivation and purposes for data collection

  Standardisation: standards for meaning (ontologies), standards for information collection and exchange templates

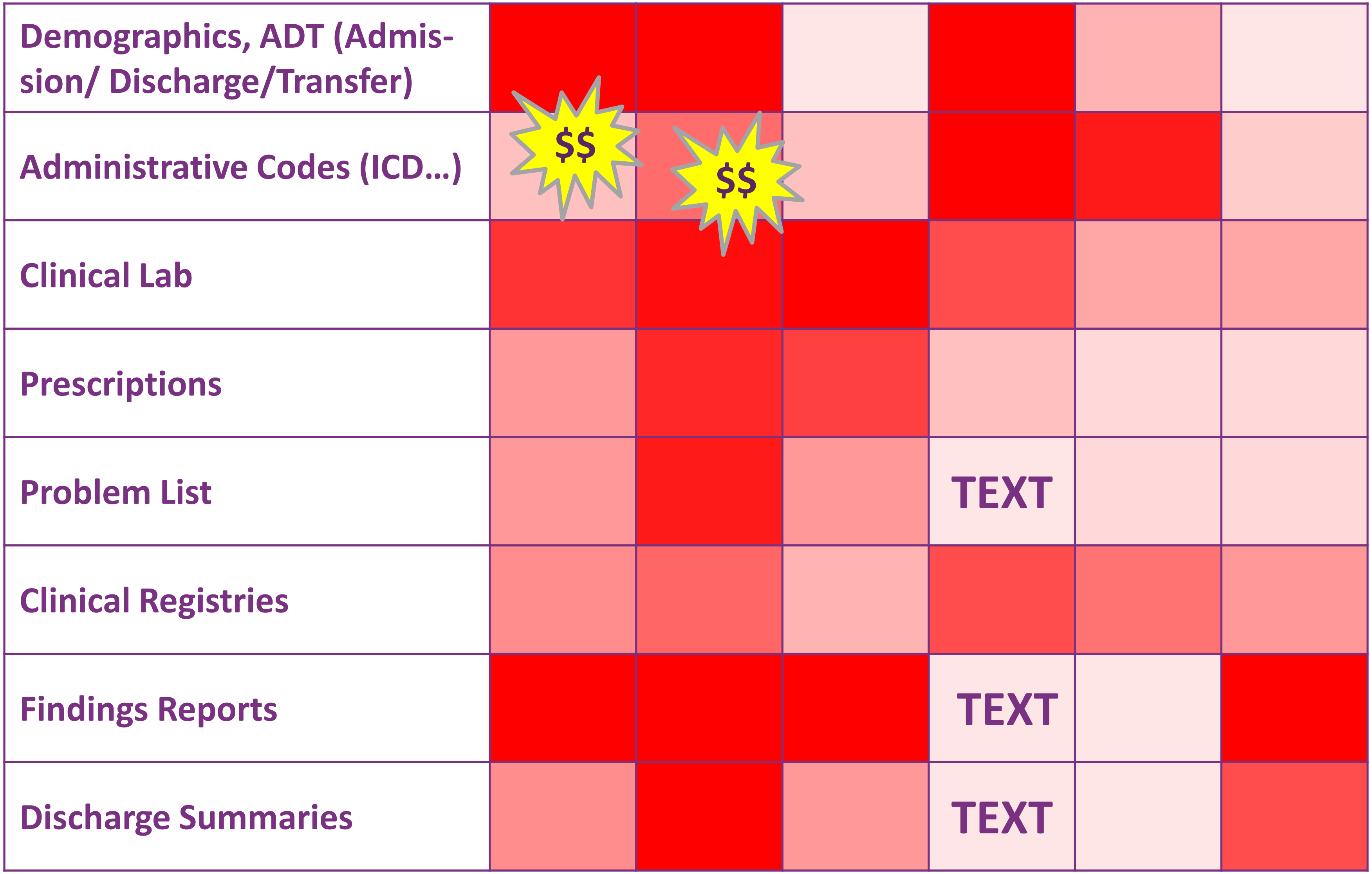
Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018

The EHR heat map



Schulz S. Clinical Informatics Challenges in Precision Medicine. Pathways to Precision Medicine. To Appear 2021

The EHR heat map



	Completeness	Correctness	Granularity	Structure	Standardisation	Data Volume
Demographics, ADT (Admission/ Discharge/Transfer)						
Administrative Codes (ICD…)	\$\$	\$\$				
Clinical Lab						
Prescriptions						
Problem List				TEXT		
Clinical Registries						
Findings Reports				TEXT		
Discharge Summaries				TEXT		

Schulz S. Clinical Informatics Challenges in Precision Medicine. Pathways to Precision Medicine. To Appear 2021

PM requires precision extraction tools

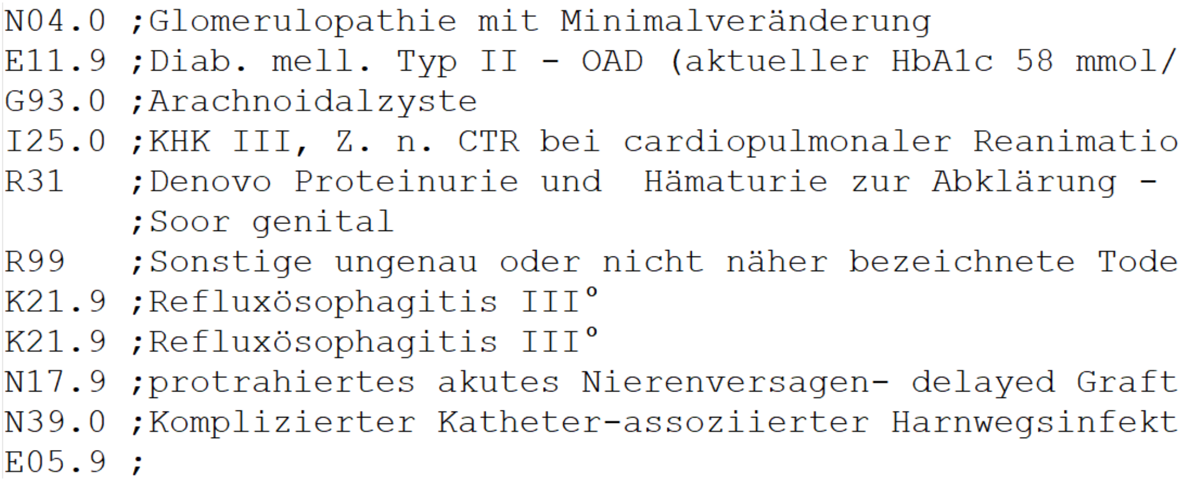
Structured data in context

PM requires precision extraction tools

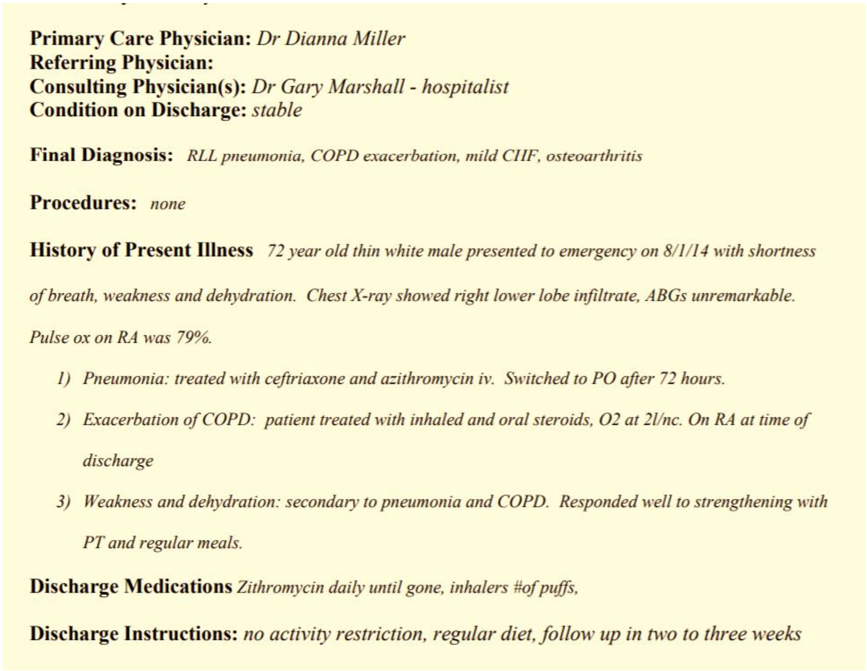
- **Automated analysis of unstructured data:**
  - Images
  - Biosignals
  - **Natural language: information extraction by natural language processing:**
    large parts of EHR content is free text:
    - Findings reports (radiology, pathology,...)
    - Progress notes
    - Nursing notes
    - Problem lists
    - Discharge summaries and letters

Large parts of information only in free text

St. p. TE eines exulc. sek.knot.SSM li US dors. 5/11 Level IV 2,4 mm Tumordurchm. Sentinnel LK ing. li. tumorfr.



```
N04.0 ;Glomerulopathie mit Minimalveränderung
E11.9 ;Diab. mell. Typ II - OAD (aktueller HbA1c 58 mmol/
G93.0 ;Arachnoidalzyste
I25.0 ;KHK III, Z. n. CTR bei cardiopulmonaler Reanimatio
R31   ;Denovo Proteinurie und  Hämaturie zur Abklärung -
      ;Soor genital
R99   ;Sonstige ungenau oder nicht näher bezeichnete Tode
K21.9 ;Refluxösophagitis III°
K21.9 ;Refluxösophagitis III°
N17.9 ;protrahiertes akutes Nierenversagen- delayed Graft
N39.0 ;Komplizierter Katheter-assoziierter Harnwegsinfekt
E05.9 ;
```



Primary Care Physician: *Dr Dianna Miller*
Referring Physician:
Consulting Physician(s): *Dr Gary Marshall - hospitalist*
Condition on Discharge: *stable*

Final Diagnosis: *RLL pneumonia, COPD exacerbation, mild CHF, osteoarthritis*

Procedures: *none*

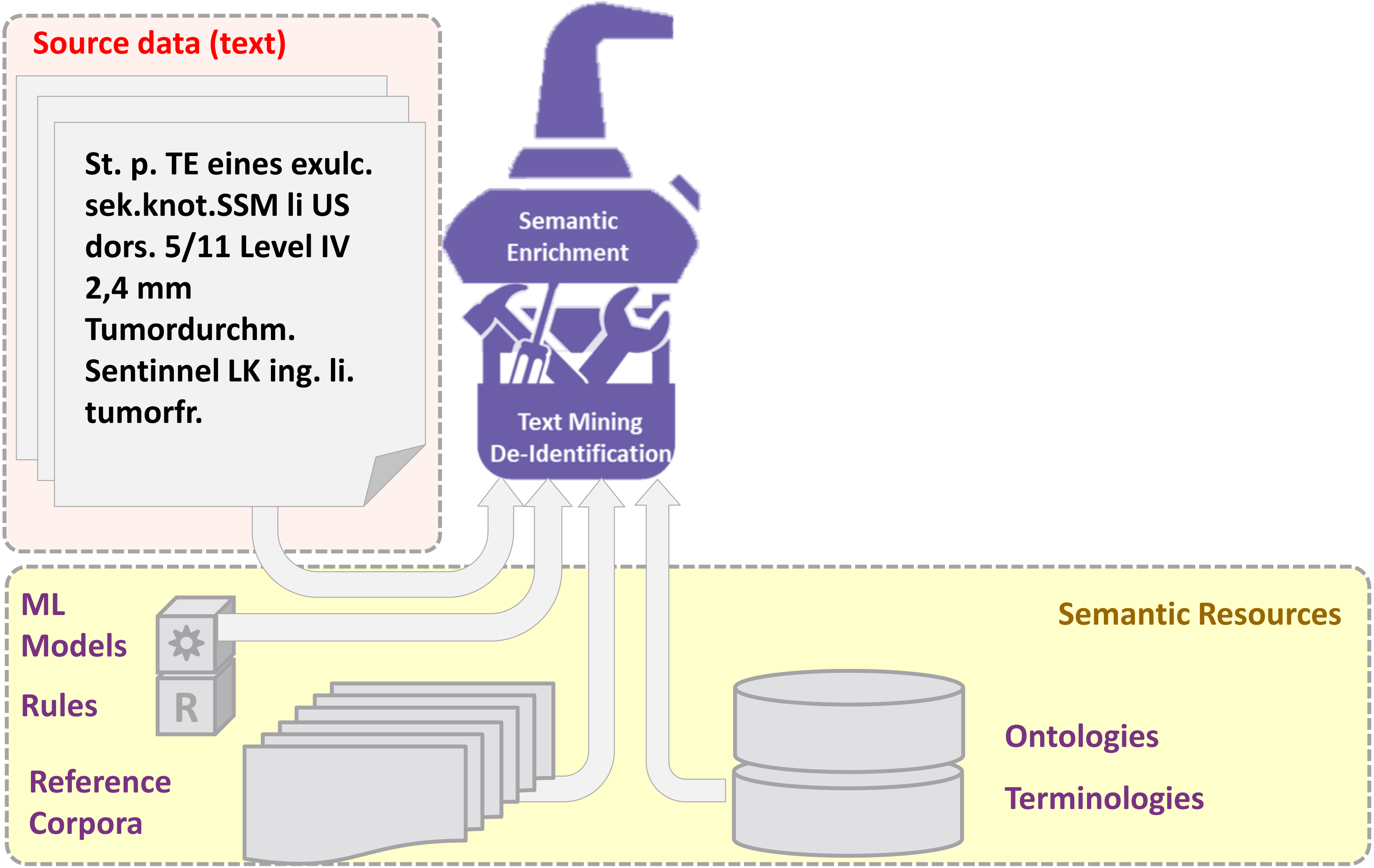
History of Present Illness *72 year old thin white male presented to emergency on 8/1/14 with shortness of breath, weakness and dehydration. Chest X-ray showed right lower lobe infiltrate, ABGs unremarkable. Pulse ox on RA was 79%.*

1) *Pneumonia: treated with ceftriaxone and azithromycin iv. Switched to PO after 72 hours.*
2) *Exacerbation of COPD: patient treated with inhaled and oral steroids, O2 at 2l/nc. On RA at time of discharge*
3) *Weakness and dehydration: secondary to pneumonia and COPD. Responded well to strengthening with PT and regular meals.*

Discharge Medications *Zithromycin daily until gone, inhalers #of puffs,*

Discharge Instructions: *no activity restriction, regular diet, follow up in two to three weeks*

Natural language processing (NLP)



Source data (text)

St. p. TE eines exulc. sek.knot.SSM li US dors. 5/11 Level IV 2,4 mm Tumordurchm. Sentinnel LK ing. li. tumorfr.

Semantic Enrichment

Text Mining De-Identification

ML Models

Rules

Reference Corpora

Semantic Resources

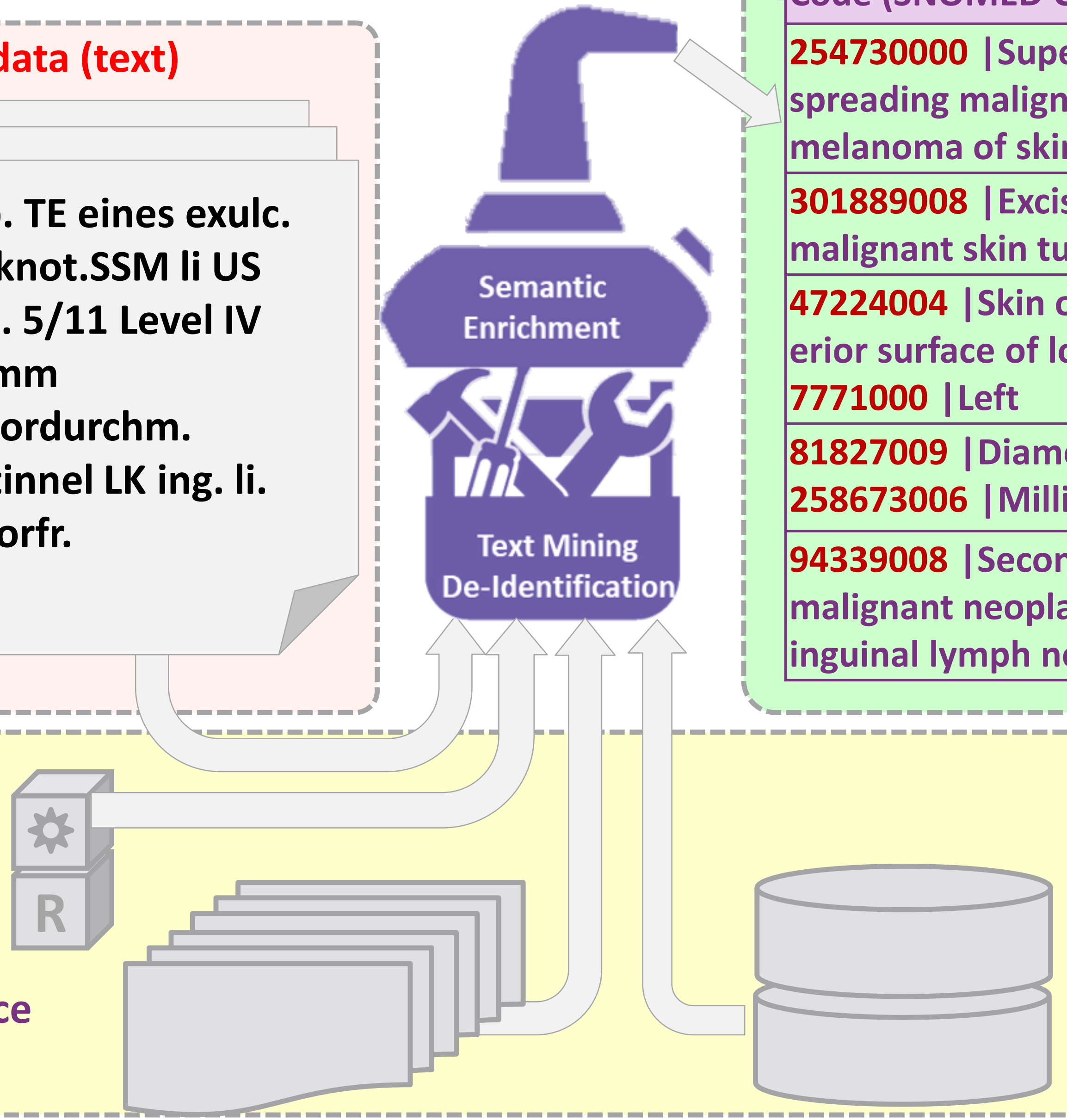
Ontologies

Terminologies

Natural language processing (NLP)

Source data (text)

St. p. TE eines exulc.
sek.knot.SSM li US
dors. 5/11 Level IV
2,4 mm
Tumordurchm.
Sentinnel LK ing. li.
tumorfr.



Standardised Target Representation

Code (SNOMED CT)	Value	Context
254730000 |Superficial spreading malignant melanoma of skin		History of
301889008 |Excision of malignant skin tumour		History of
47224004 |Skin of posterior surface of lower leg 7771000 |Left		Current
81827009 |Diameter 258673006 |Millimetre	2.4	Current
94339008 |Secondary malignant neoplasm of inguinal lymph nodes		Current Absent

Semantic Resources

ML Models

Rules

Reference Corpora

Ontologies

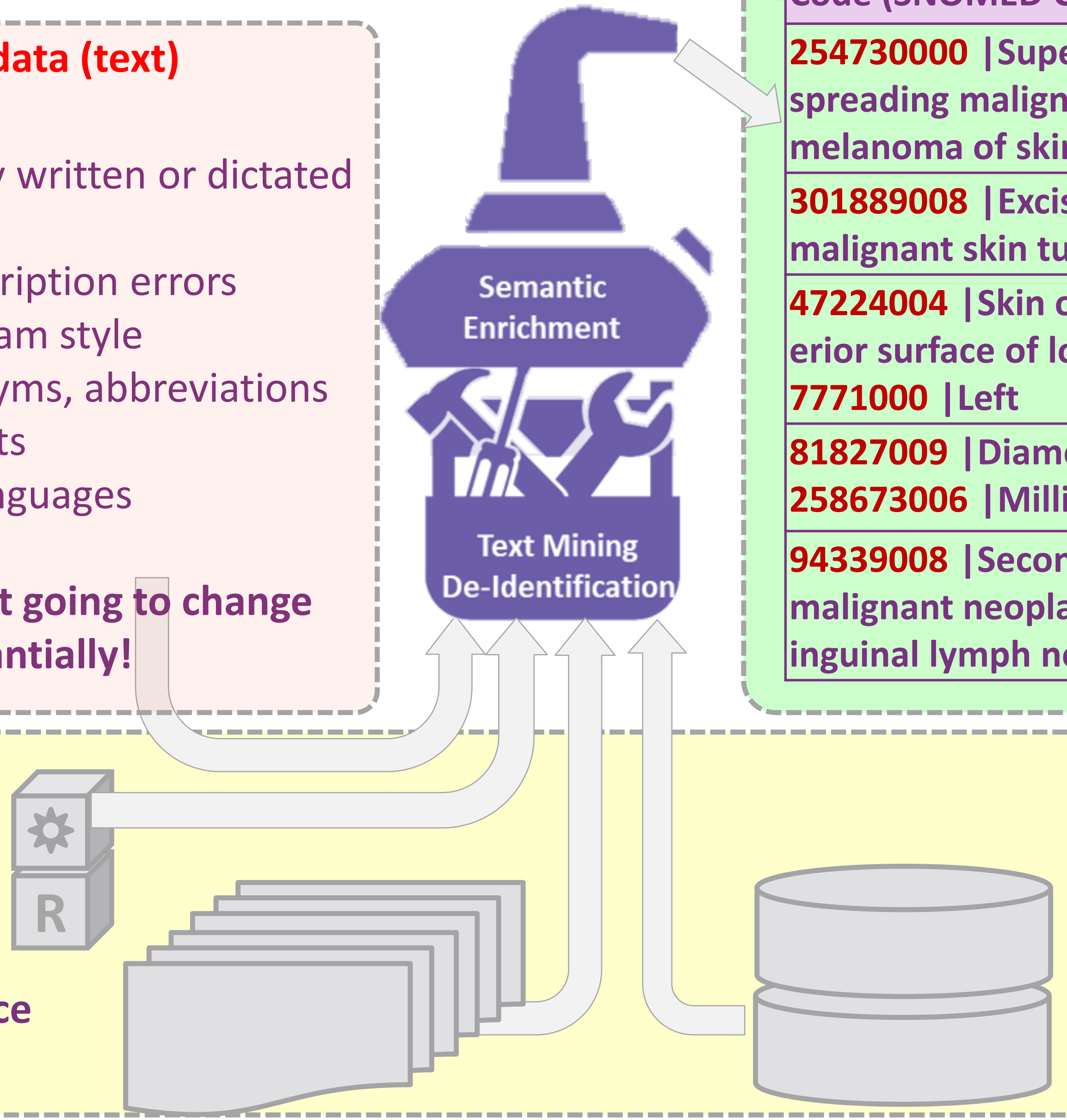
Terminologies

Natural language processing (NLP)

Source data (text)

- Hastily written or dictated
- Typos
- Transcription errors
- Telegram style
- Acronyms, abbreviations
- Dialects
- Sublanguages

- **It's not going to change substantially!**



Standardised Target Representation

Code (SNOMED CT)	Value	Context
254730000 |Superficial spreading malignant melanoma of skin		History of
301889008 |Excision of malignant skin tumour		History of
47224004 |Skin of posterior surface of lower leg 7771000 |Left		Current
81827009 |Diameter 258673006 |Millimetre	2.4	Current
94339008 |Secondary malignant neoplasm of inguinal lymph nodes		Current Absent

ML Models

Rules

Reference Corpora

Semantic Resources

Ontologies

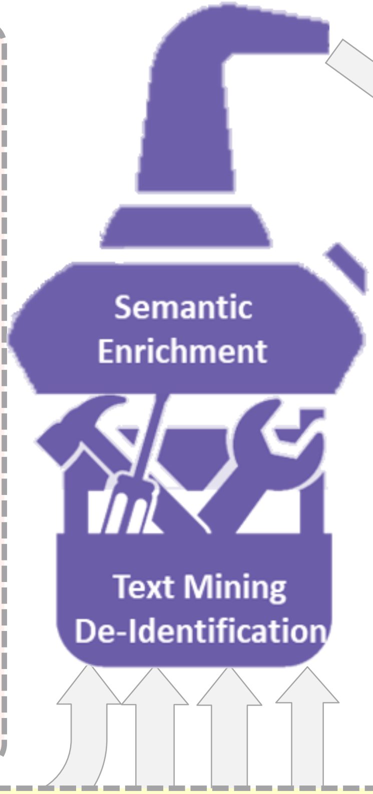
Terminologies

Natural language processing (NLP)

Source data (text)

- Hastily written or dictated
- Typos
- Transcription errors
- Telegram style
- Acronyms, abbreviations
- Dialects
- Sublanguages

- **It's not going to change substantially!**



Standardised Target Representation

Code (SNOMED CT)	Value	Context
254730000 |Superficial spreading malignant melanoma of skin		History of
301889008 |Excision of malignant skin tumour		History of
47224004 |Skin of posterior surface of lower leg 7771000 |Left		Current
81827009 |Diameter 258673006 |Millimetre	2.4	Current
94339008 |Secondary malignant neoplasm of inguinal lymph nodes		Current Absent

Semantic Resources

- Clinical NLP lagging behind
- Privacy vs. sharing of annotated corpora
- Reliability of de-identification
- Data ownership vs. sharing of models

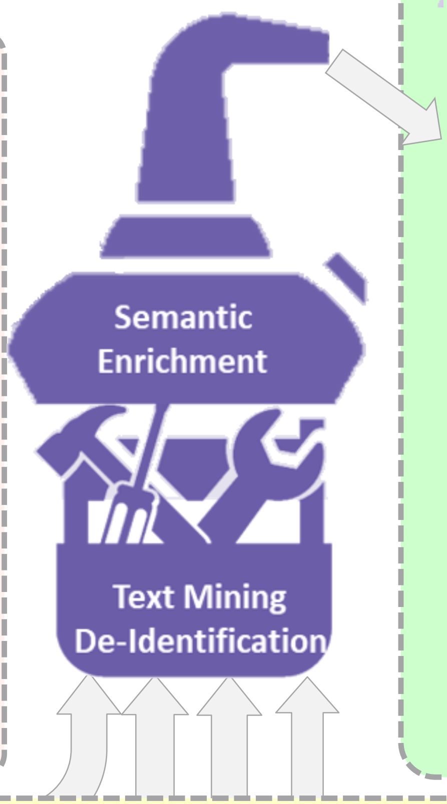
- Low adherence to standards (e.g. SNOMED CT)
- Quality issues of standards
- Coverage of clinical jargon by terminologies: Translation vs. interface terminology creation → (PMID 29295238)

Natural language processing (NLP)

Source data (text)

- Hastily written or dictated
- Typos
- Transcription errors
- Telegram style
- Acronyms, abbreviations
- Dialects
- Sublanguages

- **It’s not going to change substantially!**



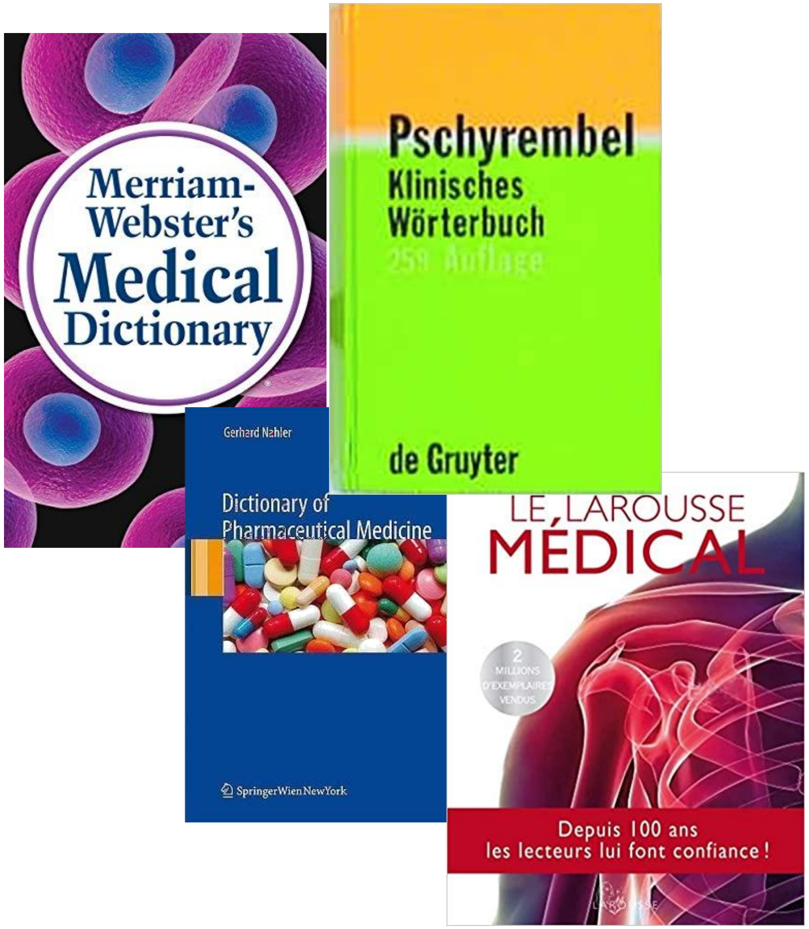
Standardised Target Representation

- Competing representations of same content
  - Low inter-coder agreement → ASSESS CT (PMID: 30654902)
- Meaning vs. context:
  - Negation
  - Plan
  - Uncertainty
  - Other subjects (family history)
- Ontologies (e.g. SNOMED CT) vs. information models (e.g. FHIR)

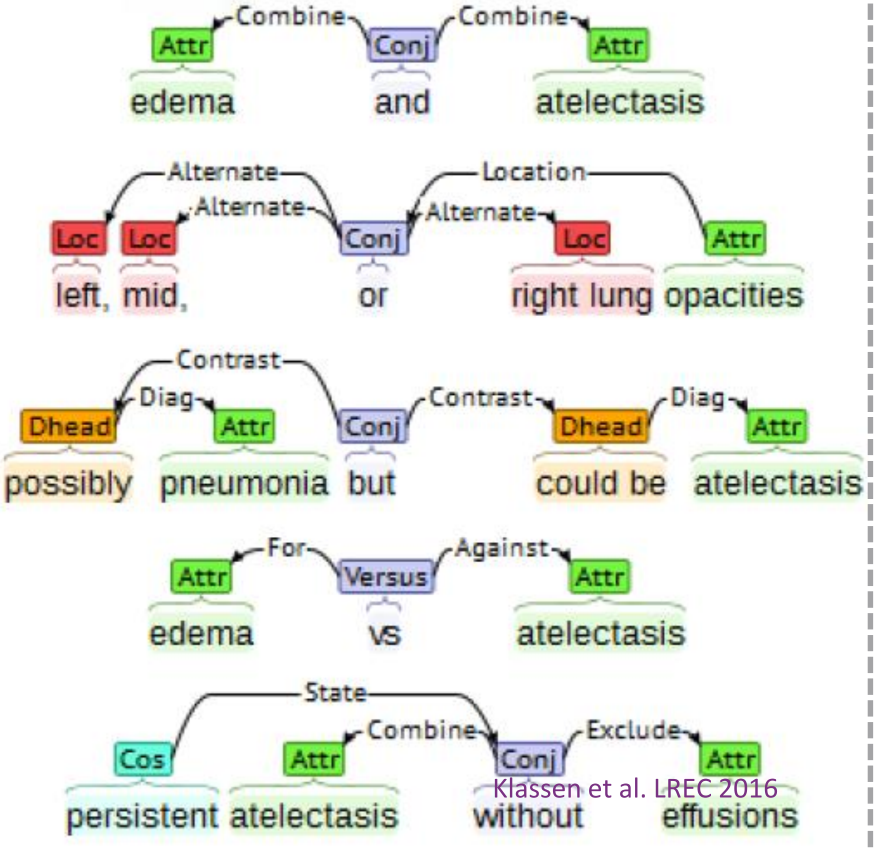
Semantic Resources

- Clinical NLP lagging behind
- Privacy vs. sharing of annotated corpora
- Reliability of de-identification
- Data ownership vs. sharing of models
- Low adherence to standards (e.g. SNOMED CT)
- Quality issues of standards
- Coverage of clinical jargon by terminologies: Translation vs. interface terminology creation → (PMID 29295238)

Precision medicine requires precision representations of clinical language


Merriam-Webster's Medical Dictionary
Pschyrembel Klinisches Wörterbuch 259. Auflage
de Gruyter
Gerhard Nahler
Dictionary of Pharmaceutical Medicine
SpringerWienNewYork
LE LAROUSSE MÉDICAL
2 MILLIONS D'EXEMPLAIRES VENDUS
Depuis 100 ans les lecteurs lui font confiance !
LAROUSSE

Dictionaries



Klassen et al. LREC 2016

Annotated Corpora

Precision medicine requires precision representations of clinical language

- **Subtle difference in spelling – large difference in meaning**
  - "Sodium chloride", "Sodium chlorite", "Sodium chlorate"
  - "AIDS", "ARDS", "STEMI", "NSTEMI"
  - "Hepatitis A", "Hepatitis B", "Hepatitis C"
- **Synonyms**
  - "2019-nCoV", "SARS-CoV-2", "Wuhan Coronavirus", "2019 novel coronavirus"
- **Homonyms**
  - "RTA": "road traffic accident" vs. "renal tubular acidosis"
- **Neologisms**
  - Single-word compounds,
    e.g. in German: "Mediainfarktverdacht", "Botulismustoxinvergiftung"

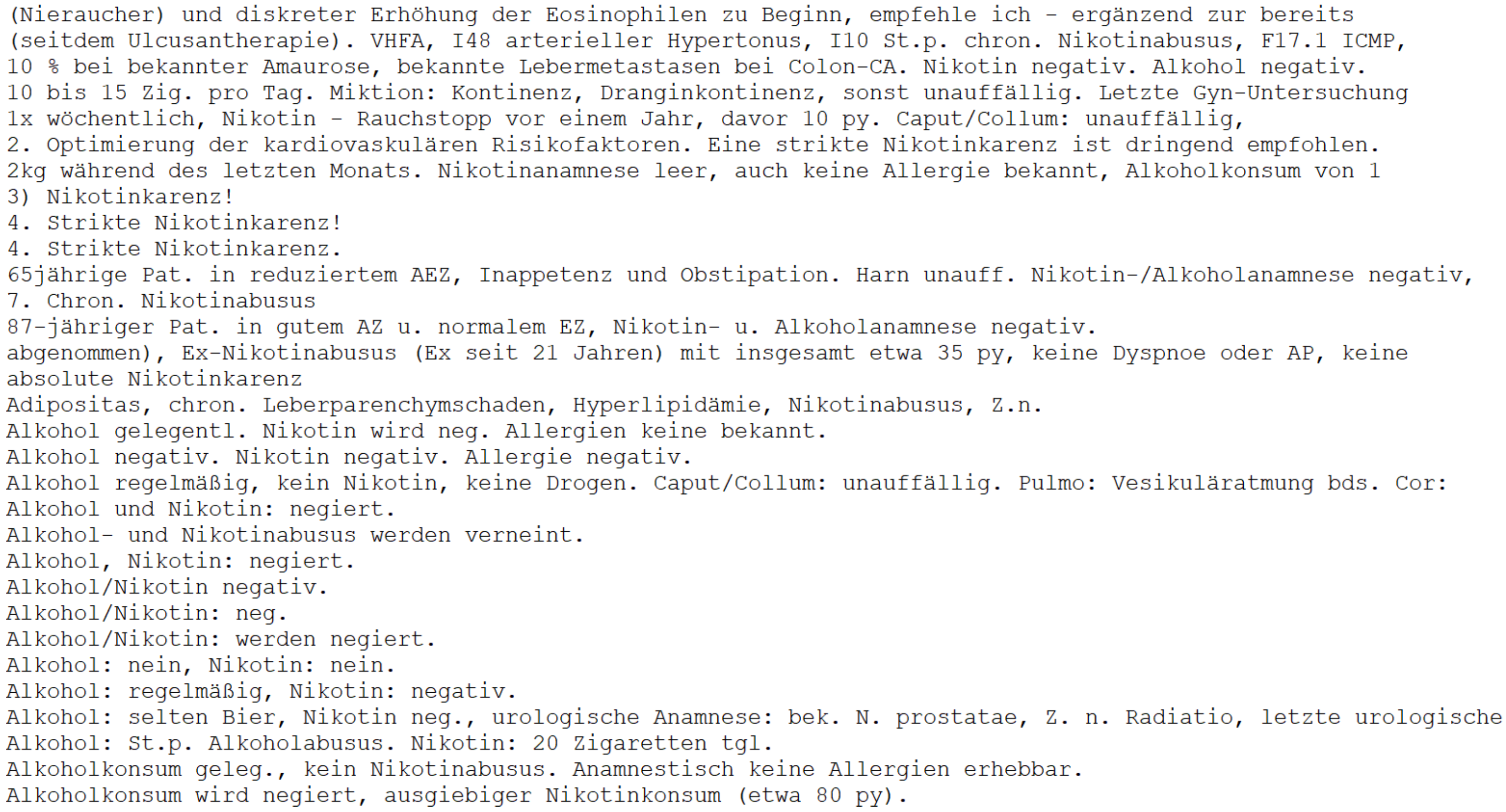
Example: SNOMED CT Interface Terminology for German

SNOMED ID	Score	Fully Specified Name (English)	German Interface Term
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Hirninfarkt verursacht durch Stenose der A. carotis
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Hirninfarkt verursacht durch Stenose der A. karotis
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Schlaganfall wegen Stenose der Halsschlagader
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Insult wegen Stenose der Halsschlagader
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Schlaganfall wegen Karotisstenose
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Insult wegen Karotisstenose
99451000119105	0.800	Cerebral infarction due to stenosis of carotid artery (disorder)	Gehirninfarkt verursacht durch Verengung der Halsschlagader

Hashemian Nik D, Kasáč Z, Goda Z, Semlitsch A, Schulz S. Building an Experimental German User Interface Terminology Linked to SNOMED CT. Stud Health Technol Inform. 2019 Aug 21;264:153-157

Example: annotation for smoking status

- Text snippets from discharge summaries.
  Annotations: {current smoker, past smoker, never smoked}

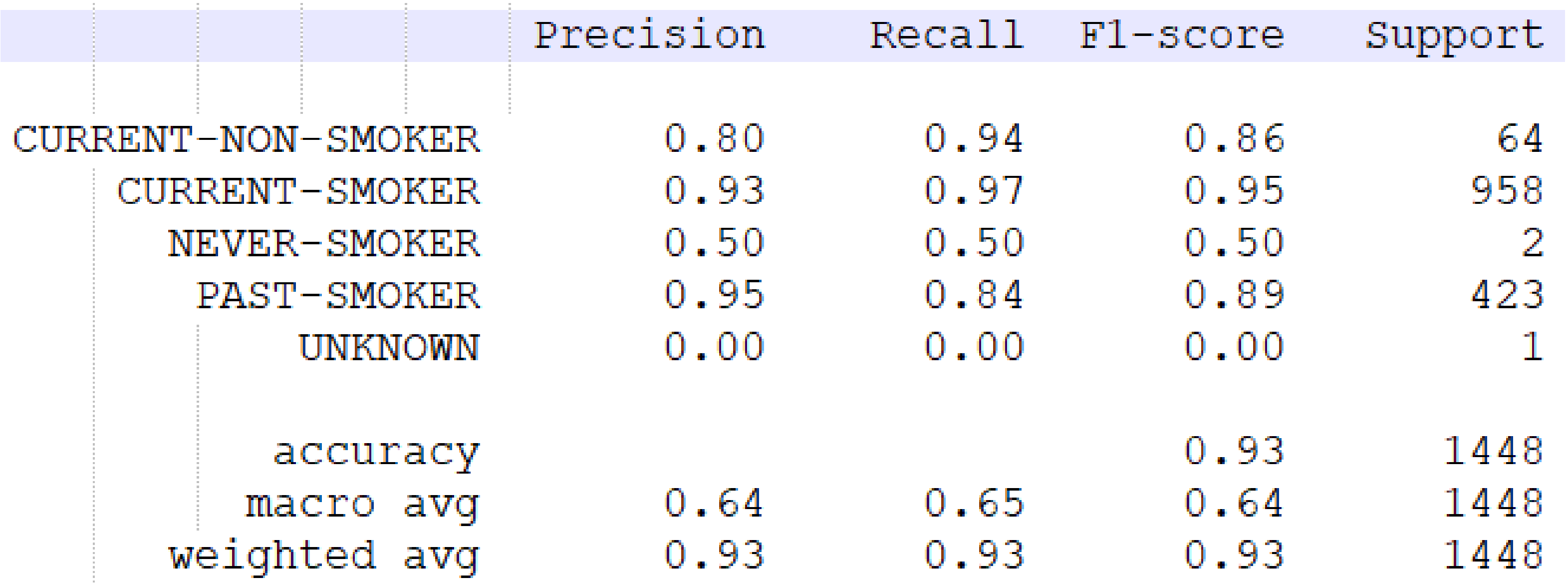


```
(Nieraucher) und diskreter Erhöhung der Eosinophilen zu Beginn, empfehle ich - ergänzend zur bereits
(seitdem Ulcusantherapie). VHFA, I48 arterieller Hypertonus, I10 St.p. chron. Nikotinabusus, F17.1 ICMP,
10 % bei bekannter Amaurose, bekannte Lebermetastasen bei Colon-CA. Nikotin negativ. Alkohol negativ.
10 bis 15 Zig. pro Tag. Miktion: Kontinenz, Dranginkontinenz, sonst unauffällig. Letzte Gyn-Untersuchung
1x wöchentlich, Nikotin - Rauchstopp vor einem Jahr, davor 10 py. Caput/Collum: unauffällig,
2. Optimierung der kardiovaskulären Risikofaktoren. Eine strikte Nikotinkarenz ist dringend empfohlen.
2kg während des letzten Monats. Nikotinanamnese leer, auch keine Allergie bekannt, Alkoholkonsum von 1
3) Nikotinkarenz!
4. Strikte Nikotinkarenz!
4. Strikte Nikotinkarenz.
65jährige Pat. in reduziertem AEZ, Inappetenz und Obstipation. Harn unauff. Nikotin-/Alkoholanamnese negativ,
7. Chron. Nikotinabusus
87-jähriger Pat. in gutem AZ u. normalem EZ, Nikotin- u. Alkoholanamnese negativ.
abgenommen), Ex-Nikotinabusus (Ex seit 21 Jahren) mit insgesamt etwa 35 py, keine Dyspnoe oder AP, keine
absolute Nikotinkarenz
Adipositas, chron. Leberparenchymschaden, Hyperlipidämie, Nikotinabusus, Z.n.
Alkohol gelegentl. Nikotin wird neg. Allergien keine bekannt.
Alkohol negativ. Nikotin negativ. Allergie negativ.
Alkohol regelmäßig, kein Nikotin, keine Drogen. Caput/Collum: unauffällig. Pulmo: Vesikuläratmung bds. Cor:
Alkohol und Nikotin: negiert.
Alkohol- und Nikotinabusus werden verneint.
Alkohol, Nikotin: negiert.
Alkohol/Nikotin negativ.
Alkohol/Nikotin: neg.
Alkohol/Nikotin: werden negiert.
Alkohol: nein, Nikotin: nein.
Alkohol: regelmäßig, Nikotin: negativ.
Alkohol: selten Bier, Nikotin neg., urologische Anamnese: bek. N. prostatae, Z. n. Radiatio, letzte urologische
Alkohol: St.p. Alkoholabusus. Nikotin: 20 Zigaretten tgl.
Alkoholkonsum geleg., kein Nikotinabusus. Anamnestisch keine Allergien erhebbar.
Alkoholkonsum wird negiert, ausgiebiger Nikotinkonsum (etwa 80 py).
```

Smoking Status Custom annotators

7242 manually annotated context lines

fastText



```
                        Precision    Recall  F1-score   Support

CURRENT-NON-SMOKER         0.80       0.94      0.86        64
    CURRENT-SMOKER         0.93       0.97      0.95       958
      NEVER-SMOKER         0.50       0.50      0.50         2
       PAST-SMOKER         0.95       0.84      0.89       423
           UNKNOWN         0.00       0.00      0.00         1

          accuracy                              0.93      1448
         macro avg         0.64       0.65      0.64      1448
      weighted avg         0.93       0.93      0.93      1448
```

Parameter optimized **shallow neural network**

Annotator integrated into Averbis health discovery platform

https://fasttext.cc/

https://averbis.com/de/health-discovery/

Precision medicine requires precision standards

Precision medicine requires precision standards

- **Two kinds of semantic standards for interoperable representation of EHR content**
  - Information models (models of use):
    Standardised templates for recurring clinical documentation needs, e.g.
    - condition, observation, procedure, medication administration
  - Ontologies (models of meaning)
    Standardised formal and informal descriptions for types of entities that are referred to by the EHR
    - diseases, procedures, substances,
      body parts, organisms, lab observables
    - linked to technical terms in several languages
- **Ontology IDs provide standardised meaning for the patient-specific instantiations of FHIR resources**

Standards require precise definitions

- **Problem: ill-defined primitives**

Condition	
Element Id	Condition
Definition	A clinical condition, problem, diagnosis, or other event, situation, issue, or clinical concept that has risen to a level of concern.

FHIR

SNOMED CT
The global language of healthcare

|Clinical finding| represents the result of a clinical observation, assessment or judgment and includes normal and abnormal clinical states e.g. |asthma|, |headache|, |normal breath sounds|). The |clinical finding| hierarchy includes concepts used to represent diagnoses.

$$\begin{aligned} \text{Appendicitis} \equiv \text{Disease} \sqcap {} & \\ & \exists\, \text{Role\_Group}.(\exists\, \text{Finding\_site}.\text{Appendix\_structure} \sqcap \\ & \qquad \exists\, \text{Associated\_morphology}.\text{Inflammatory\_morphology}) \end{aligned}$$

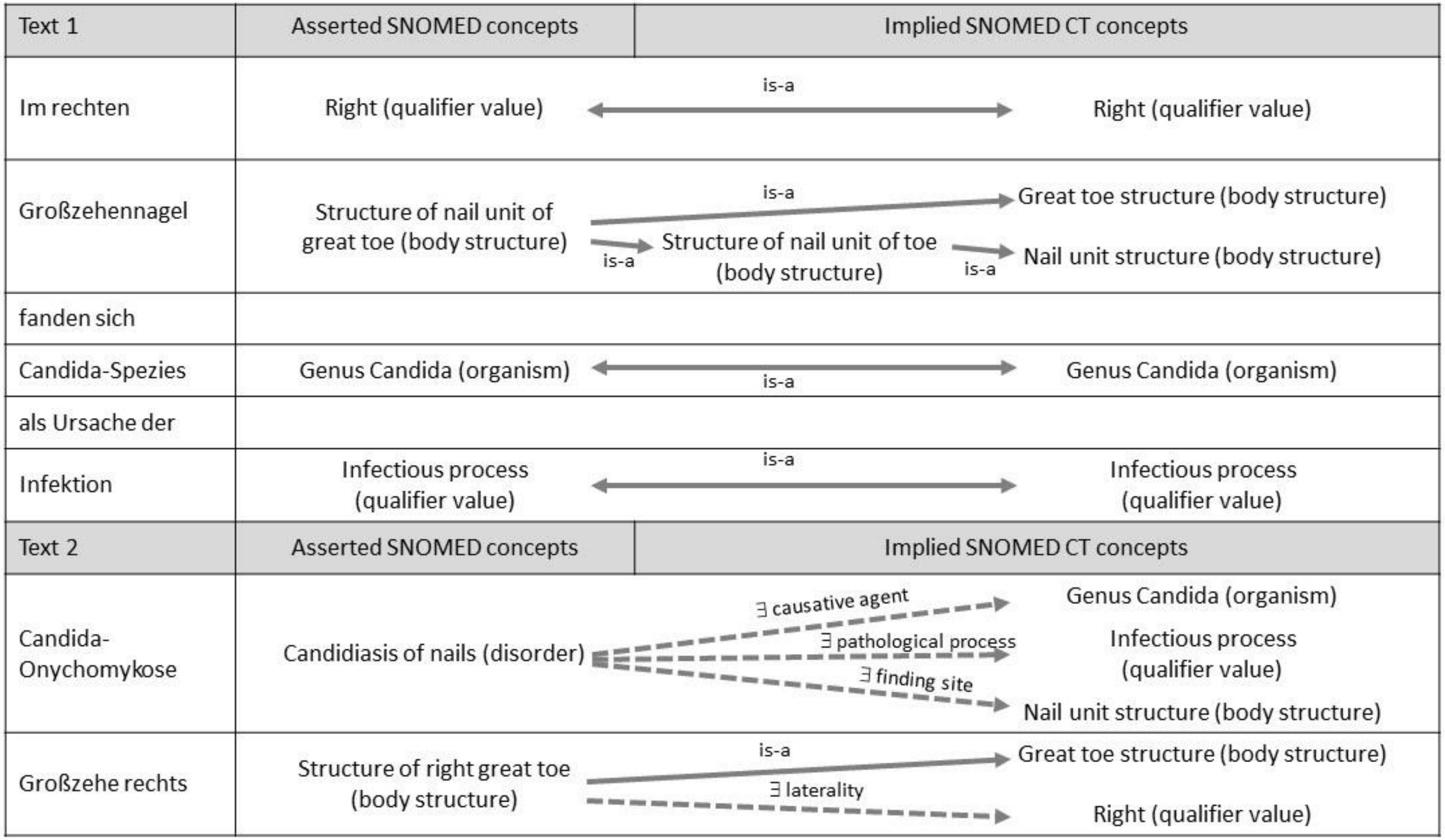
$\text{Adolescent} \sqsubseteq \text{Minor}$

$\text{Infant} \sqsubseteq \text{Minor}$

(no text definition, no formal definition)

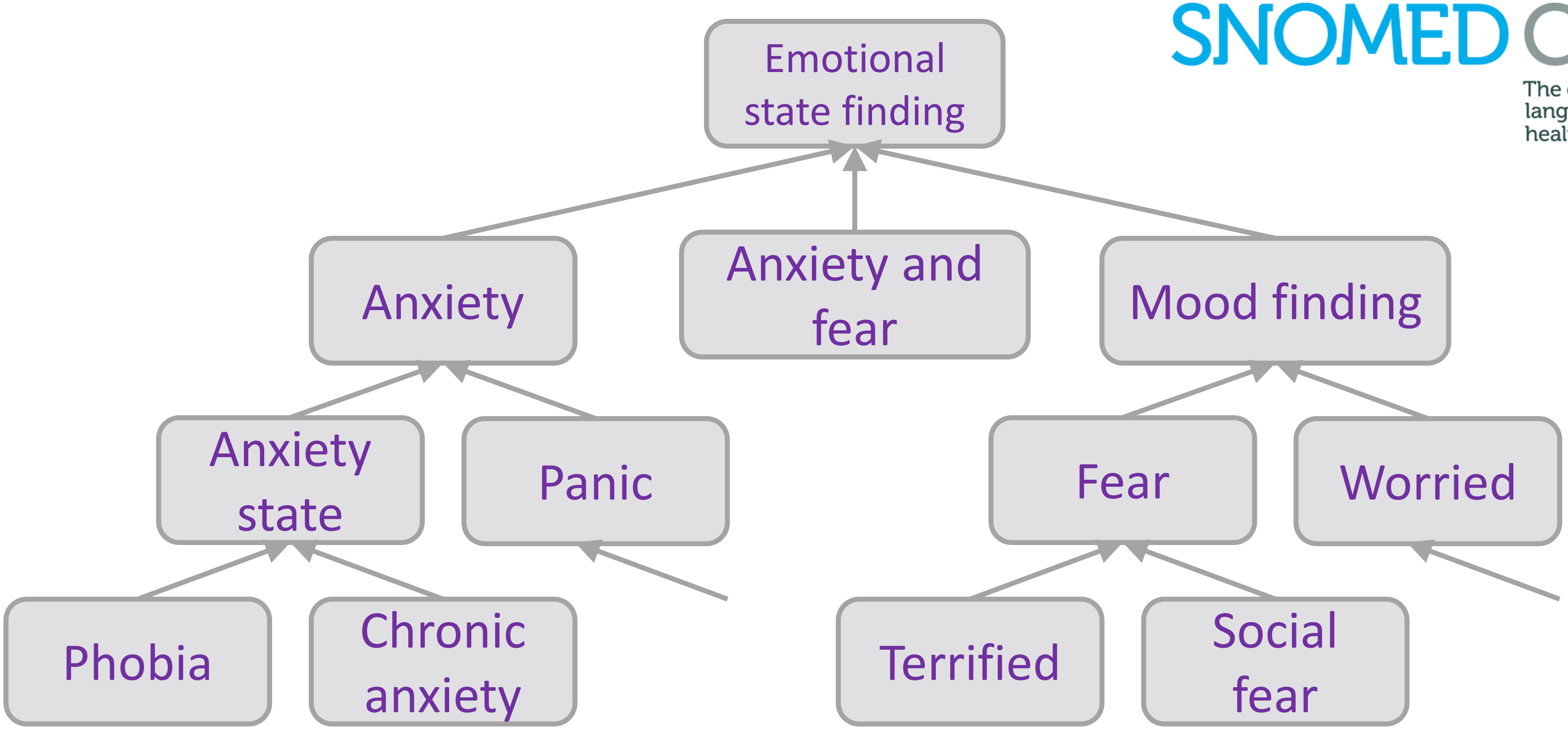
Standards should support the detection of "isosemantic" expressions

Text 1: "in the nail of the right great toe, candida species were found as cause of infection"
Text 2: "candida onychomycosis, right great toe"



Text 1	Asserted SNOMED concepts	Implied SNOMED CT concepts	
Im rechten	Right (qualifier value)	is-a	Right (qualifier value)
Großzehennagel	Structure of nail unit of great toe (body structure)	is-a is-a	Great toe structure (body structure) Structure of nail unit of toe (body structure) is-a Nail unit structure (body structure)
fanden sich			
Candida-Spezies	Genus Candida (organism)	is-a	Genus Candida (organism)
als Ursache der			
Infektion	Infectious process (qualifier value)	is-a	Infectious process (qualifier value)
Text 2	**Asserted SNOMED concepts**	**Implied SNOMED CT concepts**	
Candida-Onychomykose	Candidiasis of nails (disorder)	∃ causative agent ∃ pathological process ∃ finding site	Genus Candida (organism) Infectious process (qualifier value) Nail unit structure (body structure)
Großzehe rechts	Structure of right great toe (body structure)	is-a ∃ laterality	Great toe structure (body structure) Right (qualifier value)

Standards + data should allow detecting semantically close expressions



SNOMED CT
The global language of healthcare

Problem of large ontologies and terminologies: semantically close, undefined classes

Take-home messages

- Clinical data are overly heterogeneous
- Much information needs to be extracted from free text
- NLP-based information extraction requires costly resources
- Lack of openly-accessible clinical text
- Precision medicine needs
  - Precision information extraction tools
  - Precision language resources
  - Precision semantic standards

Thank you!

Stefan Schulz
stefan.schulz@medunigraz.at

References:

- **Schulz S**. Clinical Informatics Challenges in Precision Medicine. Pathways to Precision Medicine. To Appear 2021
- Jauk S, Kramer D, Großauer B, Rienmüller S, Avian A, Berghold A, Leodolter W, Schulz S. Risk prediction of delirium in hospitalized patients using machine learning: An implementation and prospective evaluation study. J Am Med Inform Assoc. 2020 Jul 1;27(9):1383-1392.
- Kreuzthaler M, Pfeifer B, Vera Ramos JA, Kramer D, Grogger V, Bredenfeldt S, Pedevilla M, Krisper P, Schulz S. EHR Text Categorization for Enhanced Patient-Based Document Navigation.
- Stud Health Technol Inform. 2018;248:100-107.Miñarro-Giménez, JA; Cornet, R; Jaulent, MC; Dewenter, H; Thun, S; Gøeg, KR; Karlsson, D; **Schulz, S**. Quantitative analysis of manual annotation of clinical text samples. Int J Med Inform. 2019; 123:37-48
- **Schulz, S**; Kreuzthaler, M; Huppertz, B; Sargsyan K; Kaiser, P; Fasching, R; Pieber, T. Secondary Use of Clinical Routine Data for Enhanced Phenotyping of Biobank Sample Data. Proceedings of the 1st Global Biobank Week. 2017; 1(1):53-53.-Global Biobank Week; SEP 13-15, 2017; Stockholm, SWEDEN.
- Kreuzthaler M, Martínez-Costa C, Kaiser P, Schulz S. Semantic Technologies for Re-Use of Clinical Routine Data.
- Stud Health Technol Inform. 2017;236:24-31.
- **Schulz S**, Rodrigues JM, Rector A, Chute CG. Interface Terminologies, Reference Terminologies and Aggregation Terminologies: A Strategy for Better Integration.
- Oleynik M, Kreuzthaler M, **Schulz S**. Unsupervised Abbreviation Expansion in Clinical Narratives. Stud Health Technol Inform. 2017;245:539-543.

Acknowledgements:

This work has been carried out with the K1 COMET Competence Center CBmed, which is funded by the Federal Ministry of Transport, Innovation and Technology (BMVIT); the Federal Ministry of Science, Research and Economy (BMWFW); Land Steiermark (Department 12, Business and Innovation); the Styrian Business Promotion Agency (SFG); and the Vienna Business Agency. The COMET program is executed by the FFG. KAGes and SAP provided significant resources, manpower and data as basis for research and innovation.