Medical University of Graz

IMPERFECTION IN STANDARDS FOR CLINICAL DATA REPRESENTATION

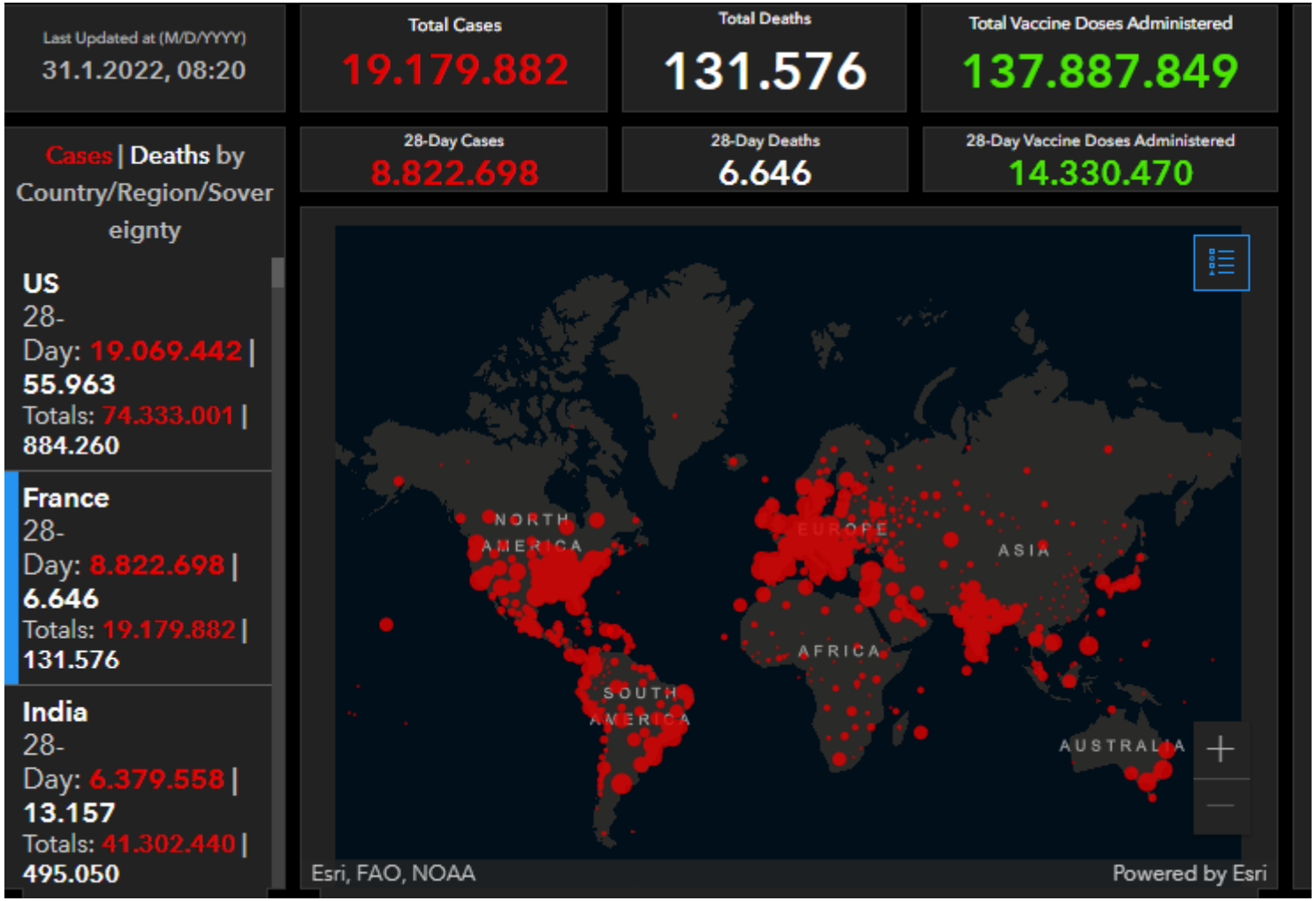
INTENDED seminar series

Stefan Schulz
Institute of Medical Informatics, Statistics and Documentation

February 1, 2022

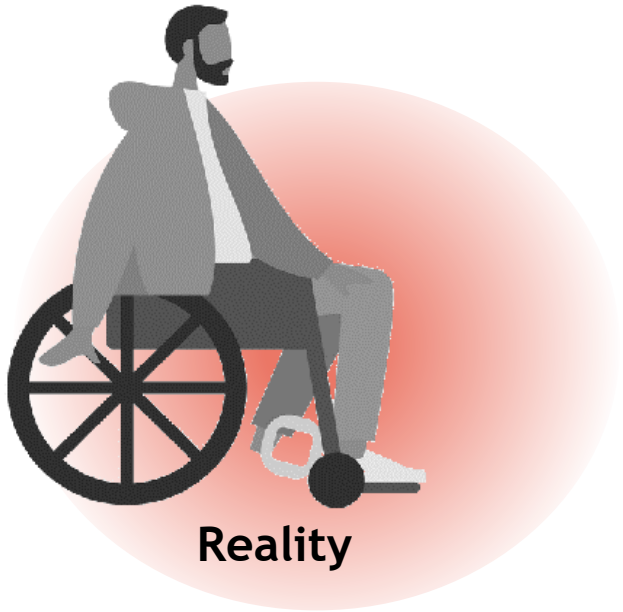
Health care *data tend to be imperfect*

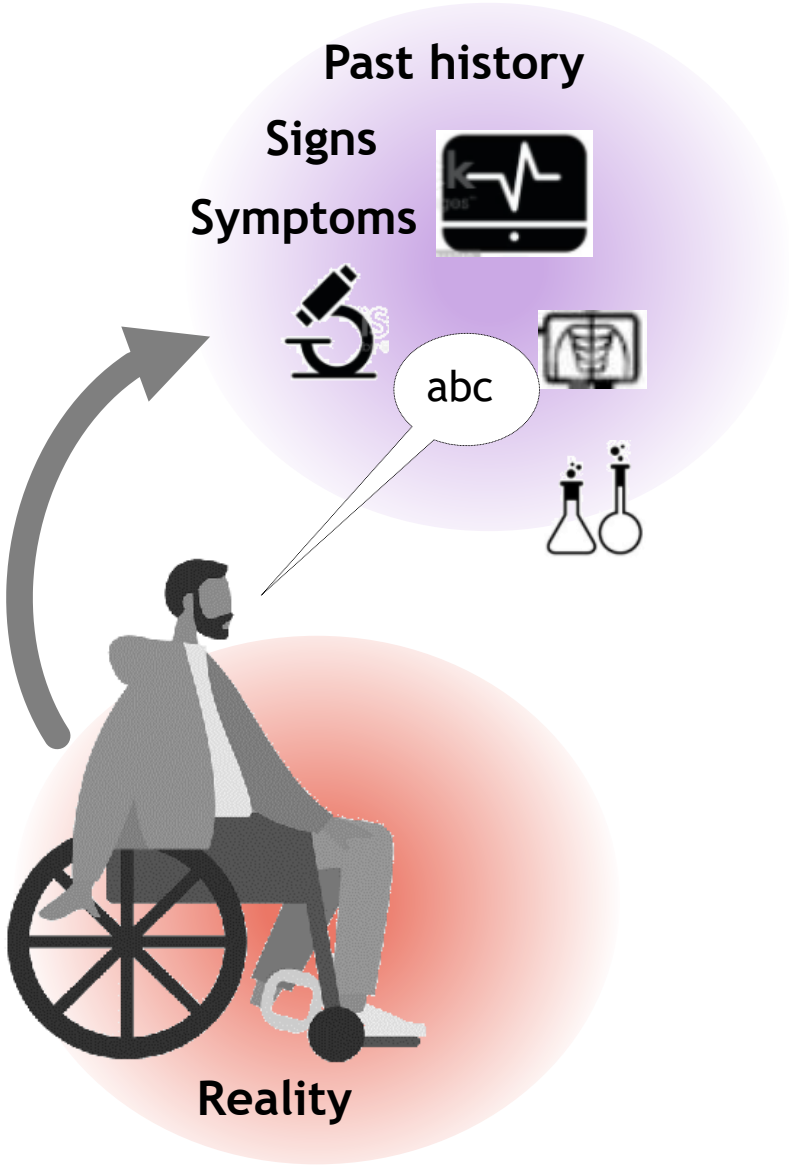
Health care *data tend to be imperfect*

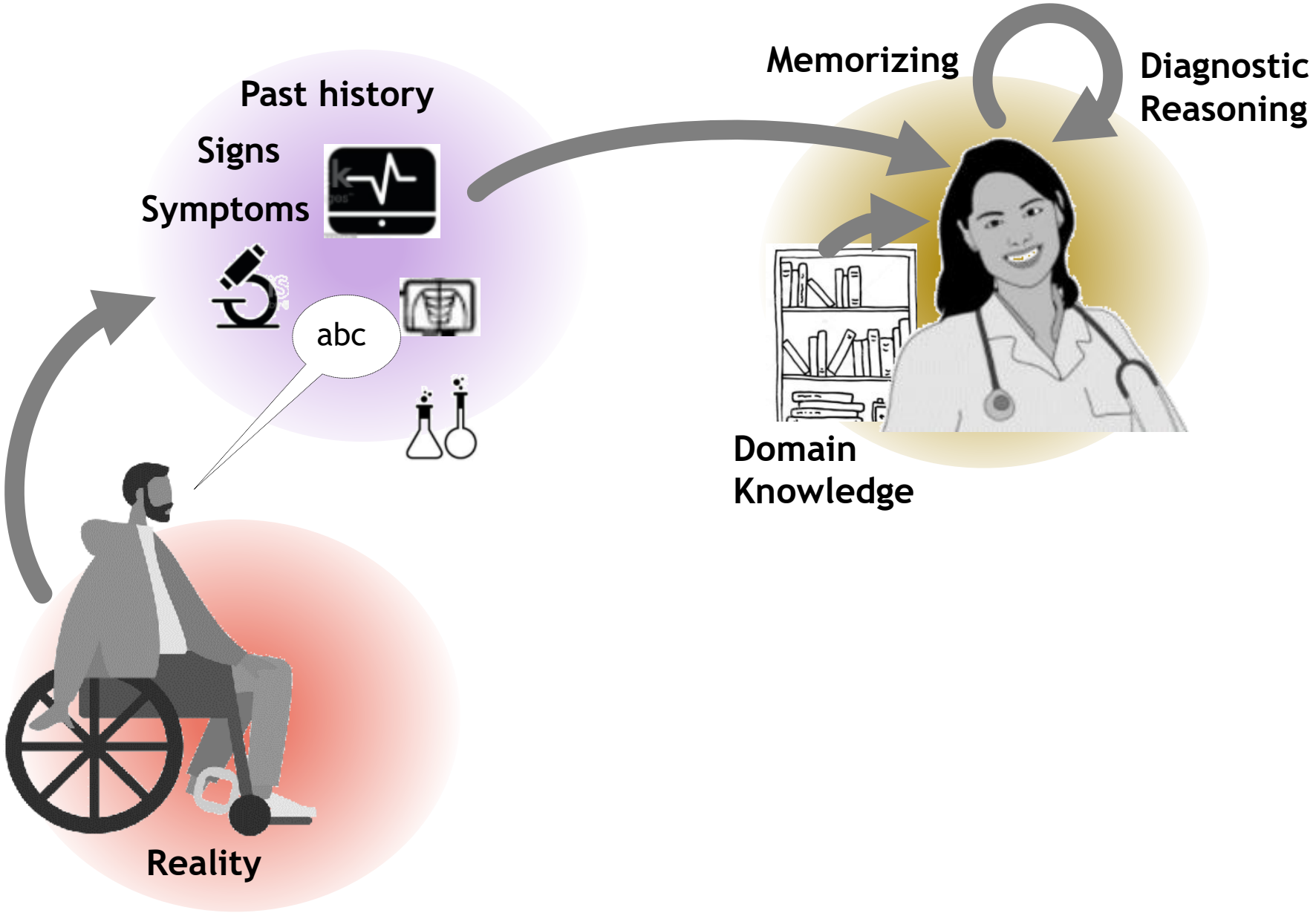


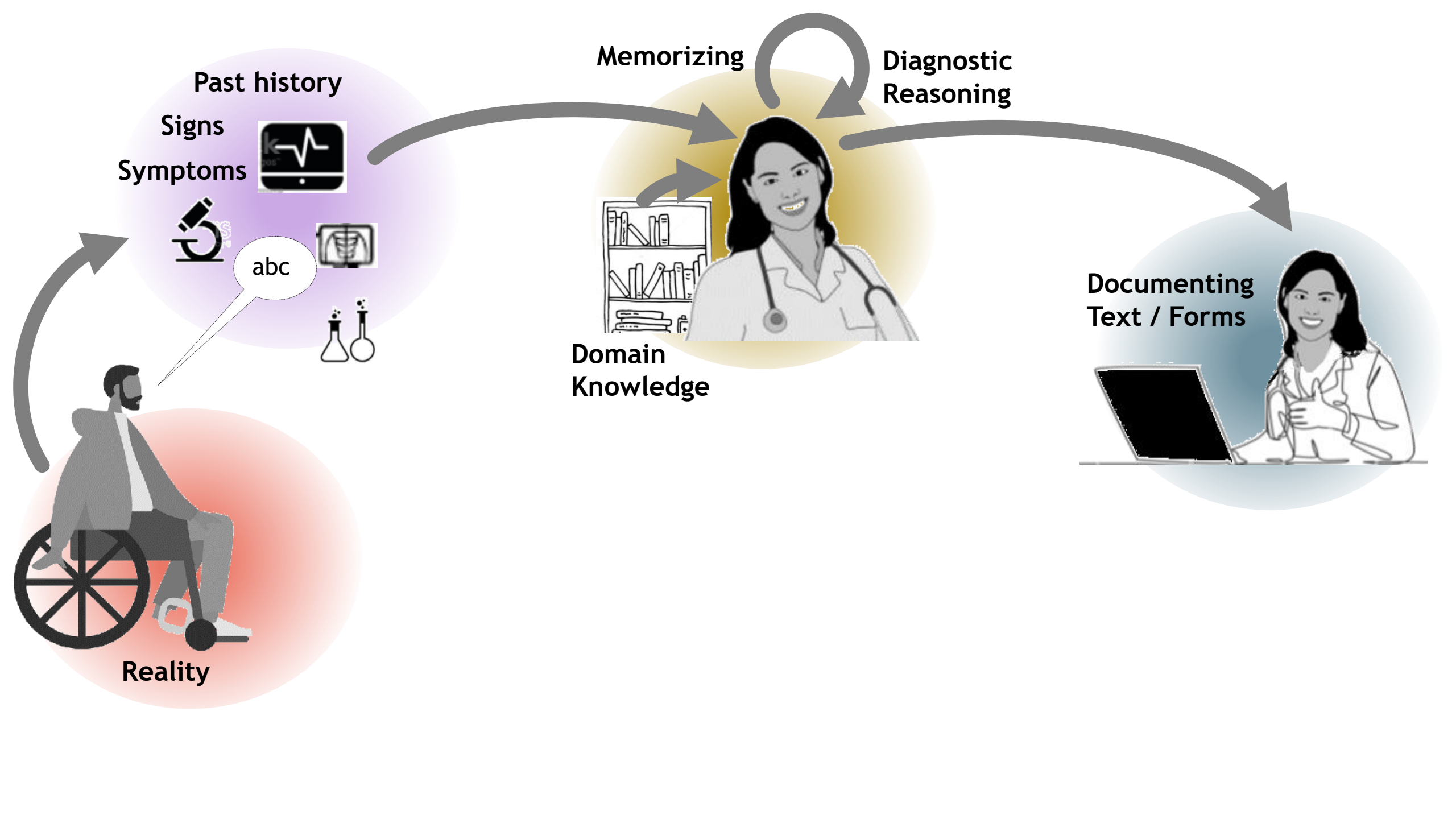
- Who was tested?
- Which test methods were used?
- How is mortality counted
  - Caused by Covid-19
  - Covid-19 present
- How are data collected?
- How are data processed?
- How are data aggregated?
- How are data analysed

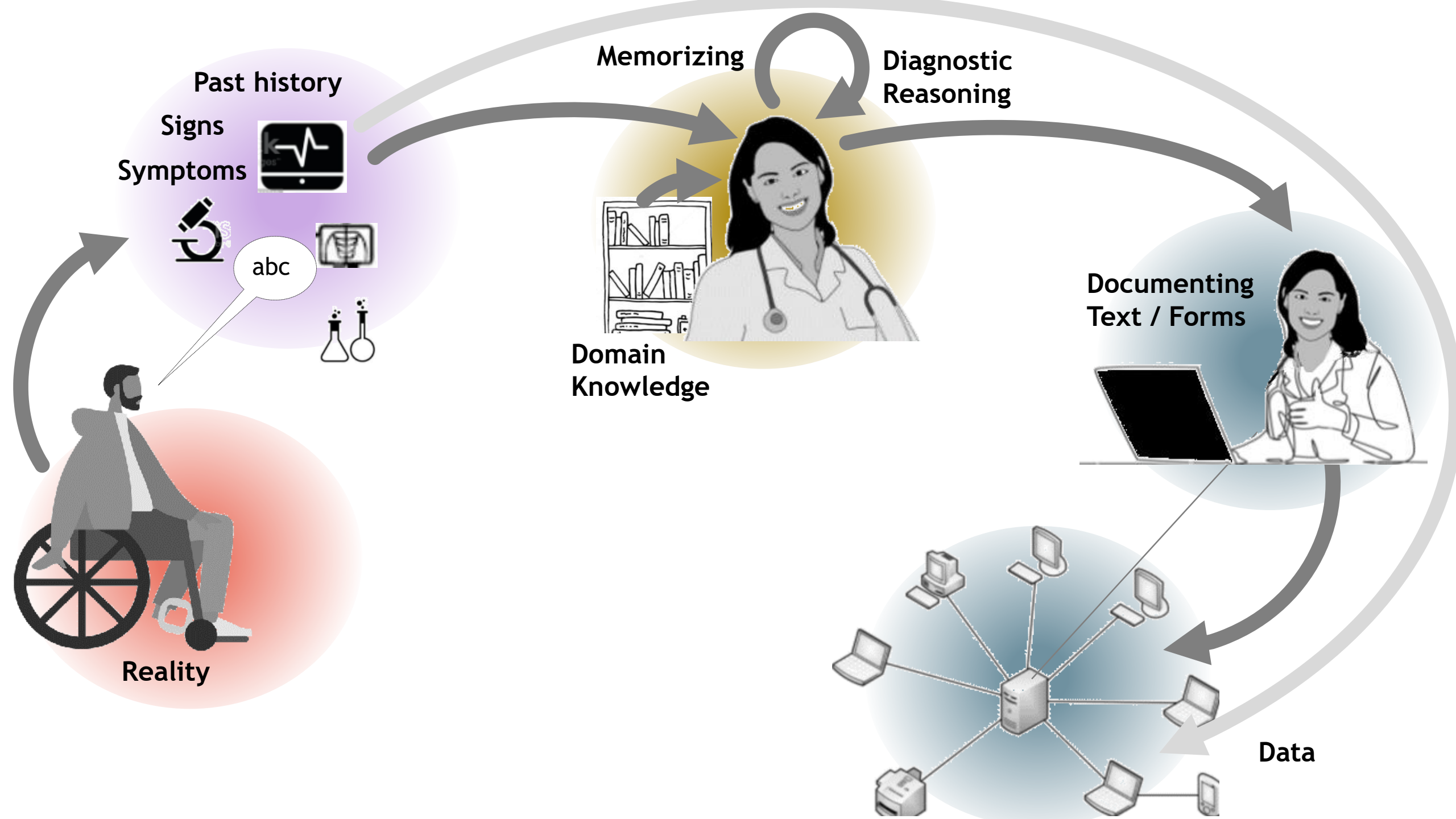
The reality of biology and medicine is complex and hidden
Health care decisions often rely on incomplete data

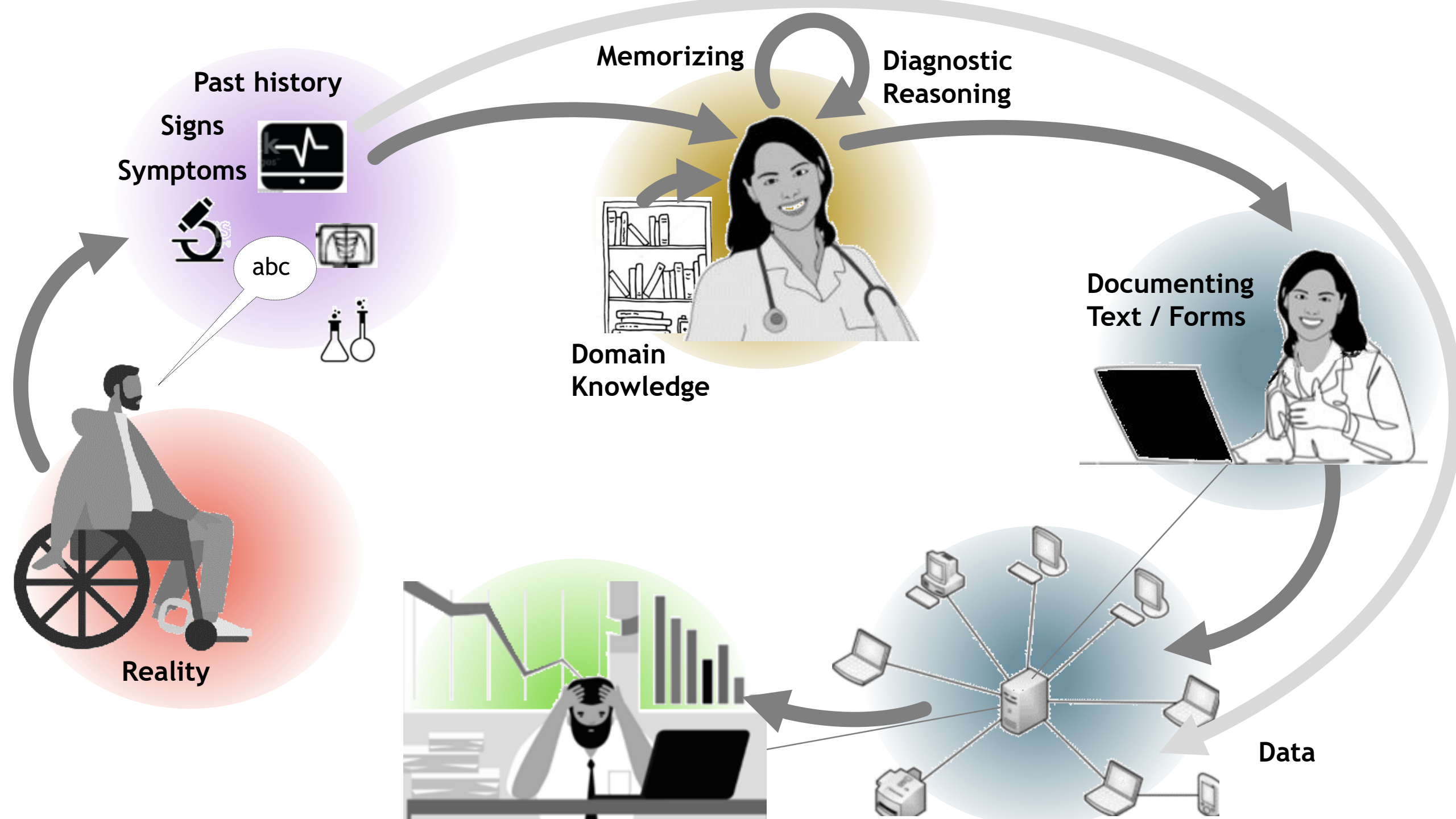



Past history
Signs
Symptoms
abc
Reality


Past history
Signs
Symptoms
abc
Reality
Memorizing
Diagnostic
Reasoning
Domain
Knowledge


Past history
Signs
Symptoms
abc
Reality
Memorizing
Diagnostic
Reasoning
Domain
Knowledge
Documenting
Text / Forms

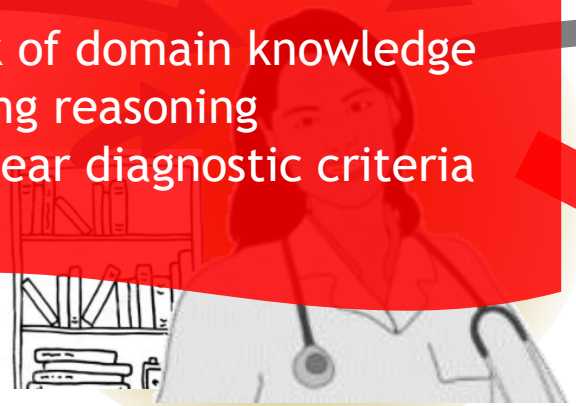

Past history
Signs
Symptoms
abc
Reality
Memorizing
Diagnostic Reasoning
Domain Knowledge
Documenting Text / Forms
Data


Past history
Signs
Symptoms
abc
Reality
Memorizing
Diagnostic Reasoning
Domain Knowledge
Documenting Text / Forms
Data

Inaccuracy and Incompleteness
- Lab results (FN / FP)
- Patient-reported data
- History

Reality

Imperfection of diagnosis
- Lack of domain knowledge
- Wrong reasoning
- Unclear diagnostic criteria



Domain Knowledge

Diagnostic Reasoning

Imperfection of documentation
- Reporting incomplete, wrong, biased
- Use of natural language (ambiguous, fuzzy terms, contexts)

Imperfection of data processing
- Information extraction
- Natural language analysis
- Coding and classification systems

Data

Imperfection of data retrieval
- Tools
- Reasoning (e.g. via taxonomies)
- Query language

Inaccuracy and Incompleteness
- Lab results (FN / FP)
- Patient-reported data
- History

Past history
Signs
Symptoms

abc

Reality

Memorizing

Diagnostic Reasoning

Imperfection of diagnosis
- Lack of domain knowledge
- Wrong reasoning
- Unclear diagnostic criteria

Domain Knowledge

Imperfection of documentation
- Reporting incomplete, wrong, biased
- Use of natural language (ambiguous, fuzzy terms, contexts)

Documenting
Text / Forms

Semantics
data standards

Imperfection of data retrieval
- Tools
- Reasoning (e.g. via taxonomies)
- Query language

Imperfection of data processing
- Information extraction
- Natural language analysis
- Coding and classification systems

Data

FAIR Principles

Findability

Resource and its metadata are easy to find by both, humans and computer systems. Basic machine readable descriptive metadata allows the discovery of interesting data sets and services.

Accessibility

Resource and metadata are stored for the long term such that they can be easily accessed and downloaded or locally used by humans and ideally also machines using standard communication protocols.

Interoperability

Metadata should be ready to be exchanged, interpreted and combined in a (semi)automated way with other data sets by humans as well as computer systems.

Reusability

Data and metadata are sufficiently well-described to allow data to be reused in future research, allowing for integration with other compatible data sources. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines and humans.

Semantics
data standards

Compliance

- **F1**. Resource is uploaded to a public repository.
- **F2**. Metadata are assigned a globally unique and persistent identifier.

- **A1**. Resource is accessible for download or manipulation by humans and is ideally also machine readable.
- **A2**. Publications and data repositories have contingency plans to assure that metadata remain accessible, even when the resource or the repository are no longer available.

- **I1**. Resource is uploaded to a repository that is interoperable with other platforms.
- **I2**. Repository meta- data schema maps to or implements the CG Core metadata schema.
- **I3**. Metadata use standard vocabularies and/or ontologies.

- **R1.** Metadata are released with a clear and accessible usage license.
- **R2.** Metadata about data and datasets are richly described with a plurality of accurate and relevant attributes.

Wilkinson, M., Dumontier, M., *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016).

Desiderata

- FAIR:
  - Using persistent unique IDs
  - Linked to persistent schemas
  - Rich descriptions based on standardised vocabularies and ontologies
- Precision Medicine / precision science
  - Require precision semantic standards for
    - Data acquisition
    - Cohort and risk group identification
    - Monitoring and quality assurance
    - Decision support

How precise are existing semantic standards ?

Case study: "Recreational drug abuse in adolescents"

Case study: "Recreational drug abuse in adolescents"

Name	**Azra** 	**Benjamin** 	**Chiara** 
Age	13	17	18
Substance	MDMA (Ecstasy)	Alcohol	Nicotine
Form	Pills (~ 100 mg)	Heineken beer (~5.5 Vol.%)	Cigarettes Marlboro
Dose /Frequency	2-3 tablets /Party 2 mal / month	2 * 0.5 liter / week	1 – 2 packs/day

Who fulfils these requirements?

Definition: Adolescence

SNOMED CT	undefined
NCI Thesaurus	12 – 21 years
Dictionnaire de l'Académie Nationale de Médecine	Période de la vie comprise entre la puberté et l'âge adulte, soit entre 12 et 18 ans.
Larousse	Période de la vie entre l'enfance et l'âge adulte, pendant laquelle se produit la puberté et se forme la pensée abstraite.
World Health Organization	10 – 19 years

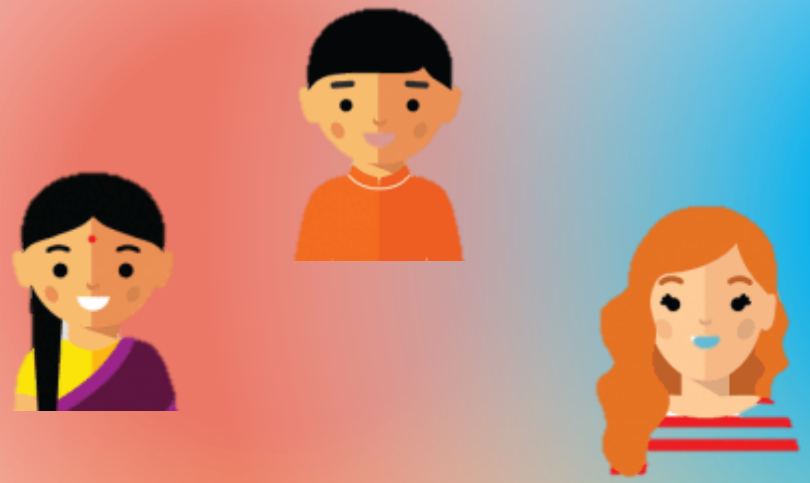
Definition: (Recreational) drug

SNOMED CT	• "Illegal drug" • "Psychoactive substance"
NCI Thesaurus	• "Recreational Drugs": A drug used for personal enjoyment rather than for medicinal purposes
Dictionnaire de l'Académie Nationale de Médecine	• (Drogue) Substance naturelle ou de synthèse dont les effets psychotropes suscitent des sensations apparentées au plaisir, incitant à un usage répétitif qui conduit à instaurer la permanence de cet effet
Larousse	• (Drogue) Substance psychotrope naturelle ou synthétique, généralement nuisible pour la santé, susceptible de provoquer une toxicomanie, et consommée en dehors d'une prescription médicale.

Definition: abuse

SNOMED CT	Substance abuse: primitive subtype of Disease
NCI Thesaurus	"Drug abuse": The use of a drug for a reason other than which it was intended (…).
Dictionnaire de l'Académie Nationale de Médecine	(abus) …utilisation excessive et volontaire, permanente ou intermittente, d'une ou de plusieurs substances psycho-actives ou non, ayant des conséquences préjudiciables à la santé physique ou psychique
Larousse	(abus) Mauvais emploi, usage excessif ou injuste de quelque chose

Recreational drug



Adolescent

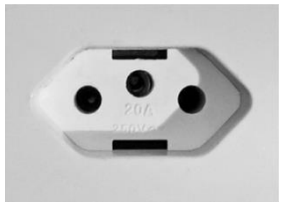
Substance abuse

Recreational drug

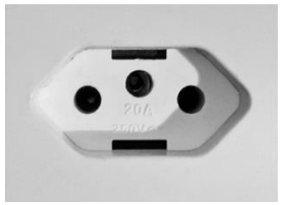
Adolescent

Substance abuse

Industry standards: definitions with crisp boundaries



Industry standards: definitions with crisp boundaries



Label	Euro flat connector	Tomada no padrão NBR 14136
Code	CEE 7/16	NBR 14136
Source	IECEE (Europ6)	NBR (Brazil)
Max voltage	230 V	127 V / 220 V
Max current	16 A	10 A
Pins / Holes: Diameter	4 mm	4.8 mm (protective contact + 3mm)
Pins / Holes: Distance	17,5 mm	17,5 mm
Pins: length	19 mm	
Pit		10 mm
Contours	Hexagon: (35,3 mm, 13.7 mm)	Hexagon: 35,5 mm, 17mm

Making medical ontologies more standards-like

Label (suggested)	**Person of minimum 14 years and younger than 18**
Code	133937008
Source	SNOMED CT
Parent class	**Person**
Definition (suggested)	>= 14.000 years and < 17.999 years

Label	**Psychoactive substance**
Code	418149003
Source	SNOMED CT
Parent class	**Substance**
Definition	Alcohol OR Nicotine OR Morphine OR Benzodiazepine OR Amphetamine OR LSD OR Cannabis OR ….

Making medical ontologies more standards-like

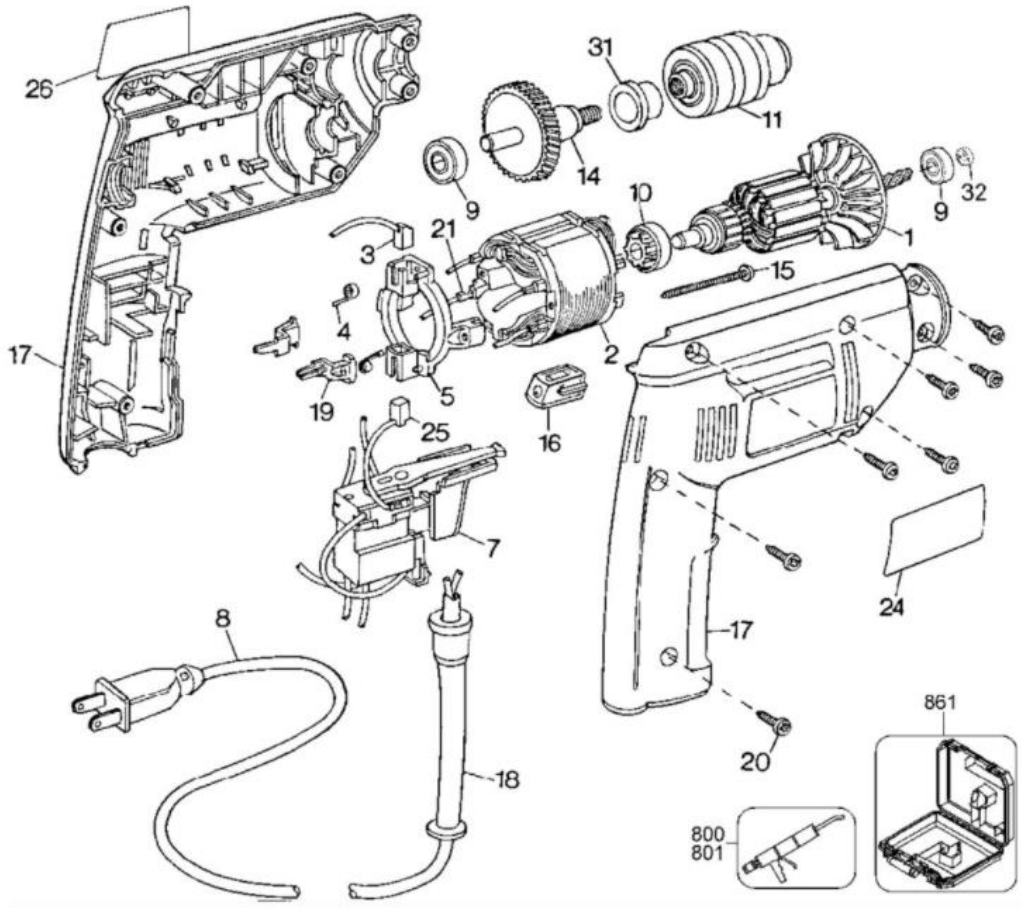
Label	**Ethanol abuse**	**Nicotine abuse**	**MDMA abuse**
Code	15167005	724697004	724703003
Source	SNOMED CT	SNOMED CT	SNOMED CT
Parent class		**Psychoactive substance abuse**	
Substance	**Ethanol**	**Nicotine**	**MDMA**
Limit	20 ♀♂ 30	?	0
UNit	g / day	g / kg body weight/d	g / kg body weight/d

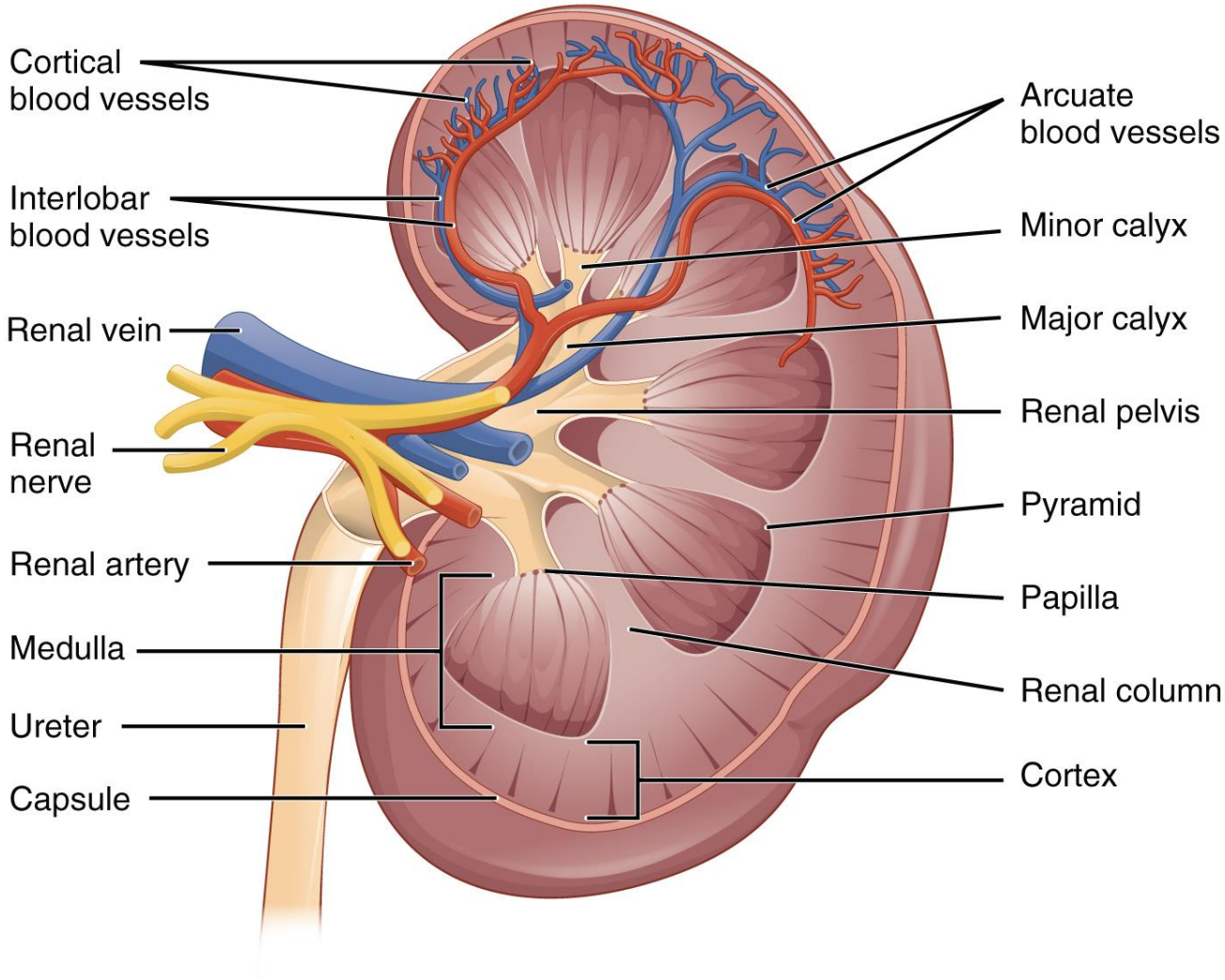
Label	**Psychoactive substance abuse**
Code	91388009
Source	**SNOMED CT**
Parent class	**Substance abuse**
Definition	**Ethanol abuse** OR **Nicotine abuse** OR **MDMA abuse** OR … OR…

Industry standards vs. biomedical ontologies

- Industry Standards:
  - Precise description of technical artefacts and assignment to categories
  - Safety, comparability, technical compatibility, interchangeability of components
- Biomedical Ontologies
  - Precise description of biological objects (objects, processes, properties) and their assignment to categories
  - Patient safety, comparability, merging of data (semantic interoperability)

Industry standards vs. biomedical ontologies





Industry standards vs. biomedical ontologies
Where is the difference?

- Ontological:
  - Artefacts consist of discrete parts, biological entities are continuous
  - Artefacts are realizations of plans. Planning requires defining and ordering efforts. Biological entities resulted from evolutionary processes
  - Artefacts often instantiate new entity types. Diversification / fusion are more obvious than in biology. Legal definitions are often mandatory
- Linguistic
  - New artefacts come with new names – names of biological kinds reflect legacy. Old names for new discoveries
  - Translation problems, change of meaning ("drug", "drogue", "Droge"), different meanings according to different views (e.g. legal vs. biological in "adolescent"

Stefan Schulz and Ingvar Johansson. Continua in Biological Systems - The Monist, 2007, 90: 4

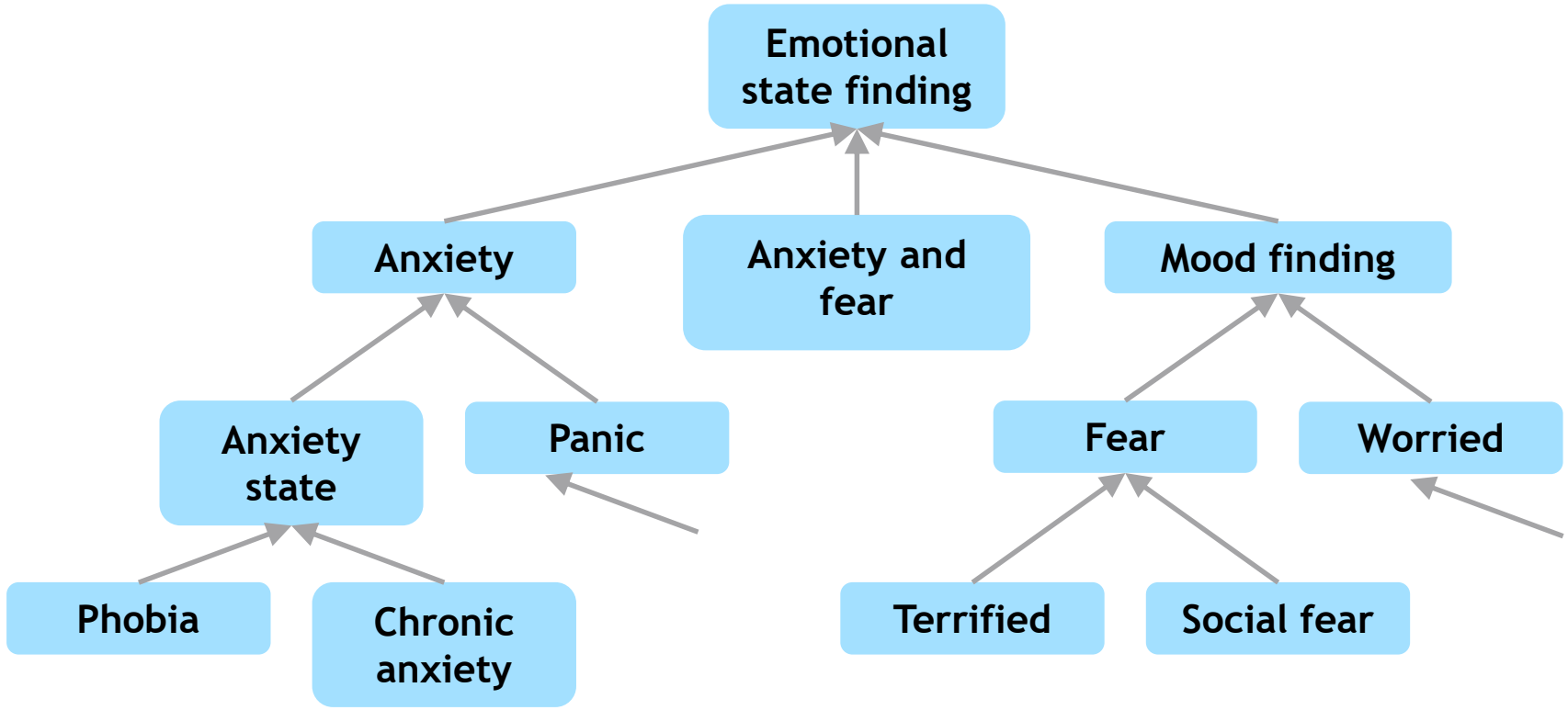
Imperfection of ontological standards

- Example: SNOMED CT, one of the largest ontologies with > 1M axioms, providing codes and standardised meaning for health care over all medical specialties
  - Started as a nomenclature: collection and standardisation of meaning, rather than standardisation of the (classes of) referents
  - Increasingly adopted ontological principles: description-logics based concept definitions, self-explaining labels.
  - Very few textual definitions / scope notes
- Two examples

SNOMED CT
The global language of healthcare

www.snomed.org

Taxonomies without defining or describing characteristics



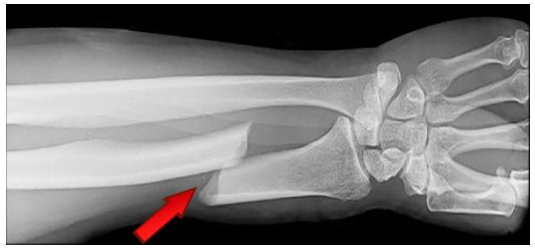
Consequence of imprecise ontologies

- Medical text annotation experiments:
- Inter-annotator agreement between medical terminology experts:

SNOMED CT (EN)		UMLS without SNOMED (EN)	
Strict	Non-strict	Strict	Non-strict
0,37	0,64	0,36	0,64

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 Mar;123:37-48

Unclear categorization – what is a disorder?



Fracture of radius (disorder)

SCTID: 12676007

12676007 | Fracture of radius (disorder) |

en Fracture of radius
en Fracture of radius (disorder)

Fracture of ulna (disorder)

SCTID: 54556006

54556006 | Fracture of ulna (disorder) |

en Fracture of ulna (disorder)
en Fracture of ulna

owl:subclassOf

owl:subclassOf

Fracture of radius AND ulna (disorder)

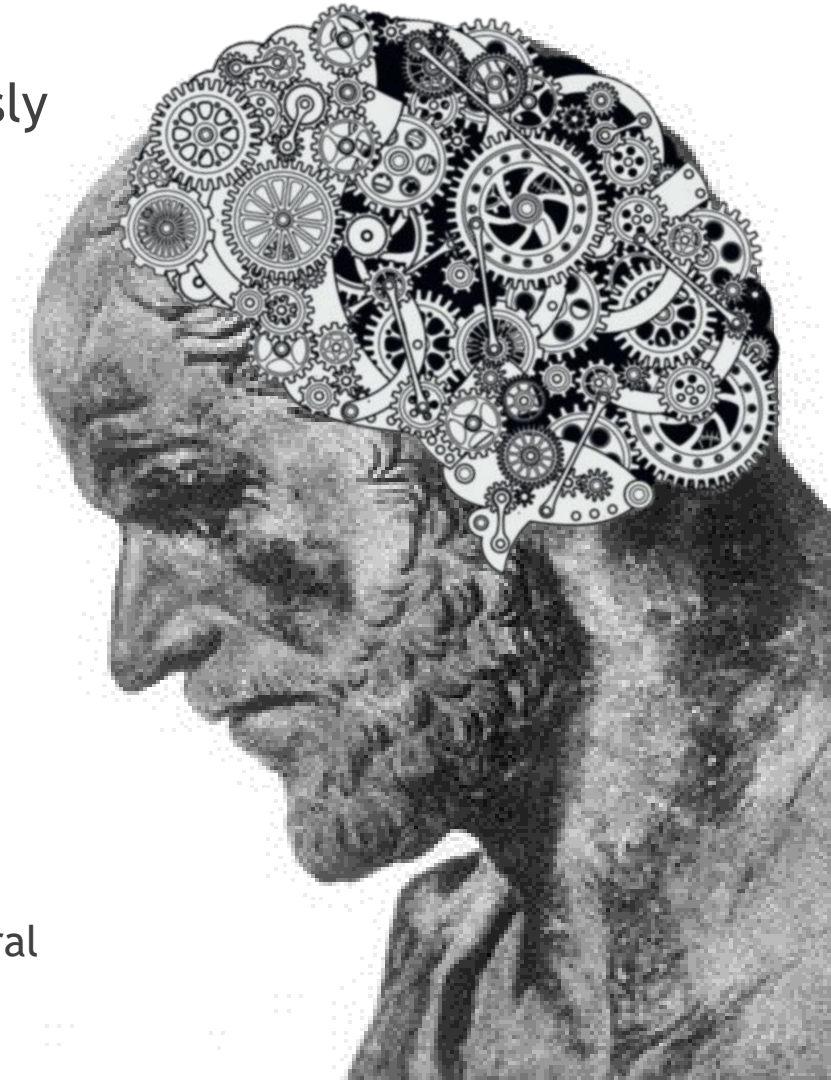
SCTID: 75857000

75857000 | Fracture of radius AND ulna (disorder) |

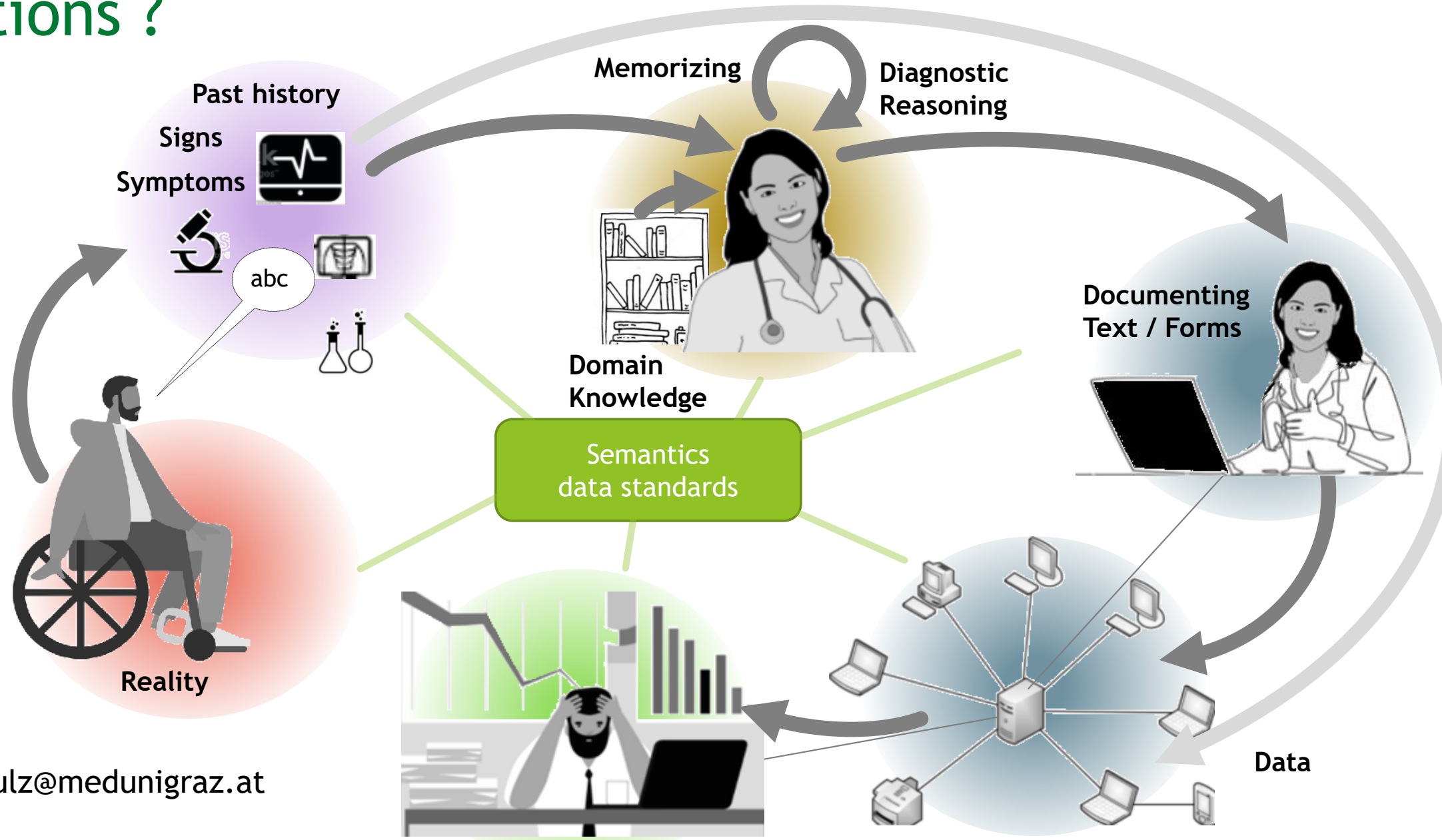
en Fracture of radius AND ulna
en Fracture of radius AND ulna (disorder)

Conclusion: reducing biomedical data standard imprecision by ontological thinking when creating semantic artefacts

- Take semantic standardisation in biology and medicine as seriously as manufacturing does
- Always start with analysing the entities of interest and their relations, then categorise, describe, define and label them
- Names / Terms do not sufficiently describe the meaning of a concept
- Make ambiguities explicit (adolescence as age group, vs. developmental stage)
- When constructing ontologies:
  - Commit to upper-level properties (object, process, quality,...)
  - Use few standardised relation types (part-of, is-about, inheres-in)
  - Commit to the description of scientific reality, as long you represent a natural science domain



Questions ?



Contact:
stefan.schulz@medunigraz.at