Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Data Management for Digital Health
Revision of Exercise 2

Borchert, Rasheed, Dr. Bayat, Dr. Schapranow

Data Management for Digital Health
Winter 2022

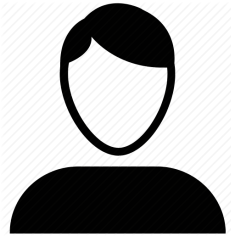
Exercise I
Topics

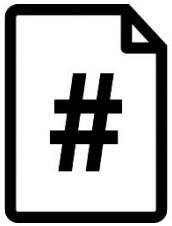
- Medical Use Case Oncology
- Bio Recap
- Genome Data Acquisition and Processing
- NLP & Text Data
- Text Data and Machine Learning

Exercise 2
Key Stats

Hasso Plattner Institut

25 Questions 50 Points	46 Students 44 Passed	Average score 40 / 80%	Average time 1h 26 min





<< 3h

Q12: Concerning global and local sequence alignment, which answers are correct

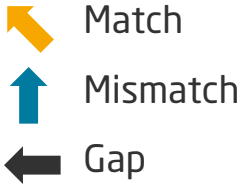
- ✓ Sequence alignment aims to identify the best matches for two given sequences.
- ✗ Global alignment produces more accurate results than local alignment.
- ✗ Global alignment always results in exactly one best alignment having the lowest score.
- ✓ There might be multiple local sequence alignments having the same alignment score.

Frequently missed

Frequent incorrect answer

Needleman-Wunsch Algorithm
Questions?

- Reference: ACTGC
- Alignment: -CTG-
- The score of the alignment is: 1 which is the **best** global score



	-	A	C	T	G	C
-	0	← -1	-2	-3	-4	-5
C	-1	-1 ↖	0	-1	-2	-3
T	-2	-2	-1 ↖	1	0	-1
G	-3	-3	-2	0 ↖	2 ←	1

Q15: Please apply the inverse BWT function BWT-1 for the given string. Which of the following statements are correct for the result?

Hasso Plattner Institut

- ✓ The original string contains three non-consecutive letters ‘a’.
- ✗ The original string contains three consecutive letters ‘a’.
- ✓ The original string contains ‘n’ and ‘m’, whereas ‘n’ occurs first.
- ✗ The original string contains ‘n’ and ‘m’, whereas ‘m’ occurs first.

Frequently missed

Frequent incorrect answer

nr*a^{3}g#m → *anagram#

Read	Sorted	-	BWT
3.	A_1	<	N
6.	A_2	<	R
1.	A_3	<	*
4.	G	<	A_1
7.	M	<	A_2
2.	N	<	A_3
5.	R	<	G
9.	*	<	#
8.	#	<	M

Q16: Which statements are true about ontologies?

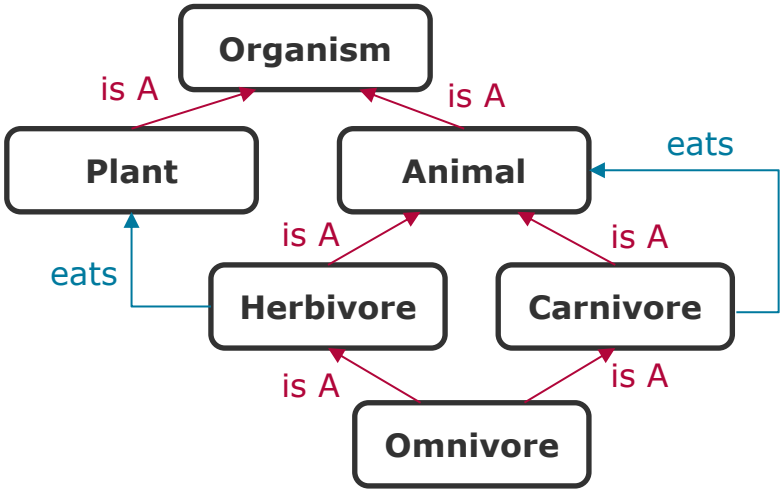
- ✓ Ontologies and controlled vocabularies can be used to derive dictionaries for Named Entity Recognition.
- ✗ Ontologies are statistical models of a particular domain.
- ✗ Ontologies always have a tree structure, i.e., each concept only has a single parent concept.
- ✗ The graph structure of ontologies makes machine-processing infeasible.

Frequently missed

Frequent incorrect answer

Representing Meaning: Ontologies

- Strings have no “meaning” per se
- Ontologies = representation and formal naming of concepts in a domain and relations among them
- Process of creating an ontology is known as knowledge engineering
- Have a long history in philosophy and are used in Artificial Intelligence since 1970s



Q18: Extraction of accurate phenotypes [...] Why is it not sufficient to use ICD codes in the EHR for this purpose?

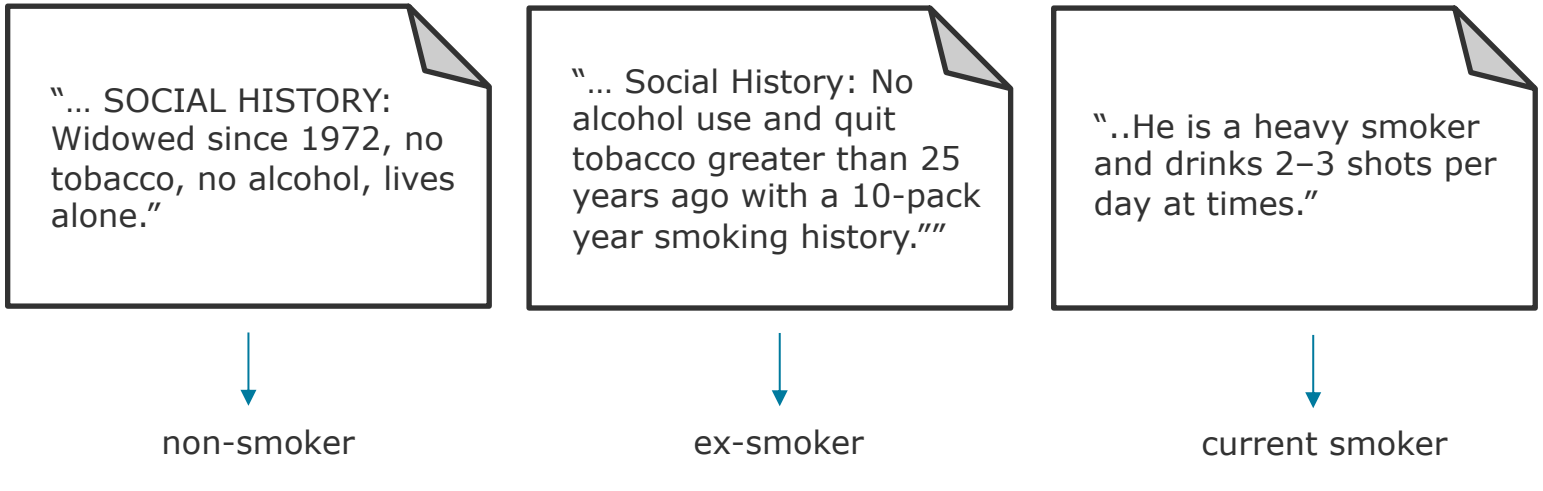
Hasso Plattner Institut

- ✓ ICD codes are used for billing purposes and may therefore not serve the purpose of our downstream application.
- ✓ The granularity of ICD codes may be too coarse for many downstream applications, e.g., for distinguishing certain disease subtypes.
- ✗ ICD codes are not machine-readable and cannot easily be turned into features for ML algorithms.
- ✗ ICD codes fall under higher levels of data protection than clinical free-text but cannot be de-identified.

Frequently missed

Frequent incorrect answer

Example: Phenotyping



Heinze et al. (2008): Medical i2b2 NLP Smoking Challenge: The A-Life System Architecture and Methodology

Q21: You are planning to use an off-the-shelf rule-based information extraction tool [...] What might prevent you from doing so?

Hasso Plattner Institut

- ✗ You do not have enough training data to train the tool.
- ✓ The tool was developed for another language as your target language.
- ✓ The tool was developed for other types of entities than the ones you care about and you do not have the resources to extend the tool.
- ✗ Using open-source tools is not allowed in a clinical context, because you can never know if they don't have security issues.

Frequently missed

Frequent incorrect answer

Rule- or ML-based Information Extraction?

	Pros	Cons
Rule-based	• Declarative • Easy to comprehend • Easy to maintain • Easy to incorporate domain knowledge • Easy to trace and fix the cause of errors	• Heuristic • Requires tedious manual labor
ML-based	• Trainable • Adaptable • Reduces manual effort	• Requires labeled data • Requires retraining for domain adaptation • Requires ML expertise to use or maintain • Opaque

CAVE:

- from 2013 (ML-based back then != today)
- by IBM researchers (IBM sells / sold rule-based IE software)

Chiticariu, Laura et. Al (2013): Rule-based Information Extraction is Dead! Long Live Rule-based Information Extraction Systems!