Curation of Large Scale EHR Data for Use with Biobank Samples

Global Biobank Week 14.9.2017
Session 6B: Biobanks and Electronic Health Records

Henrik Edgren, CSO

Conflicts of interest

- Employee of MediSapiens Ltd.
- Visit us at booth 27A.

Outline

- Benefits and challenges of EHR data.
- The challenges of scale.
- Different types of data and the data curation process.
- Discrete data and narrative text.

Benefits of EHR data

- The data exists and more is recorded continuously.
- Covers many areas of disease and health.
- The data covers a long period of time.
- Collected by healthcare professionals.

- Particularly useful for population-based biobanks with a more general research purpose.

Challenges of EHR data

- The data exists and more is recorded continuously.
- Covers many areas of disease and health.
- The data covers a long period of time.
- Collected by healthcare professionals.

- Access is often not straightforward.

The challenges of scale

- Discrete data, narrative text, laboratory measurements, images etc.

- So much diverse data that it is almost unusable.
- More is generated continuously.

- How do you turn this into clean, standardized and searchable data?

The data curation process

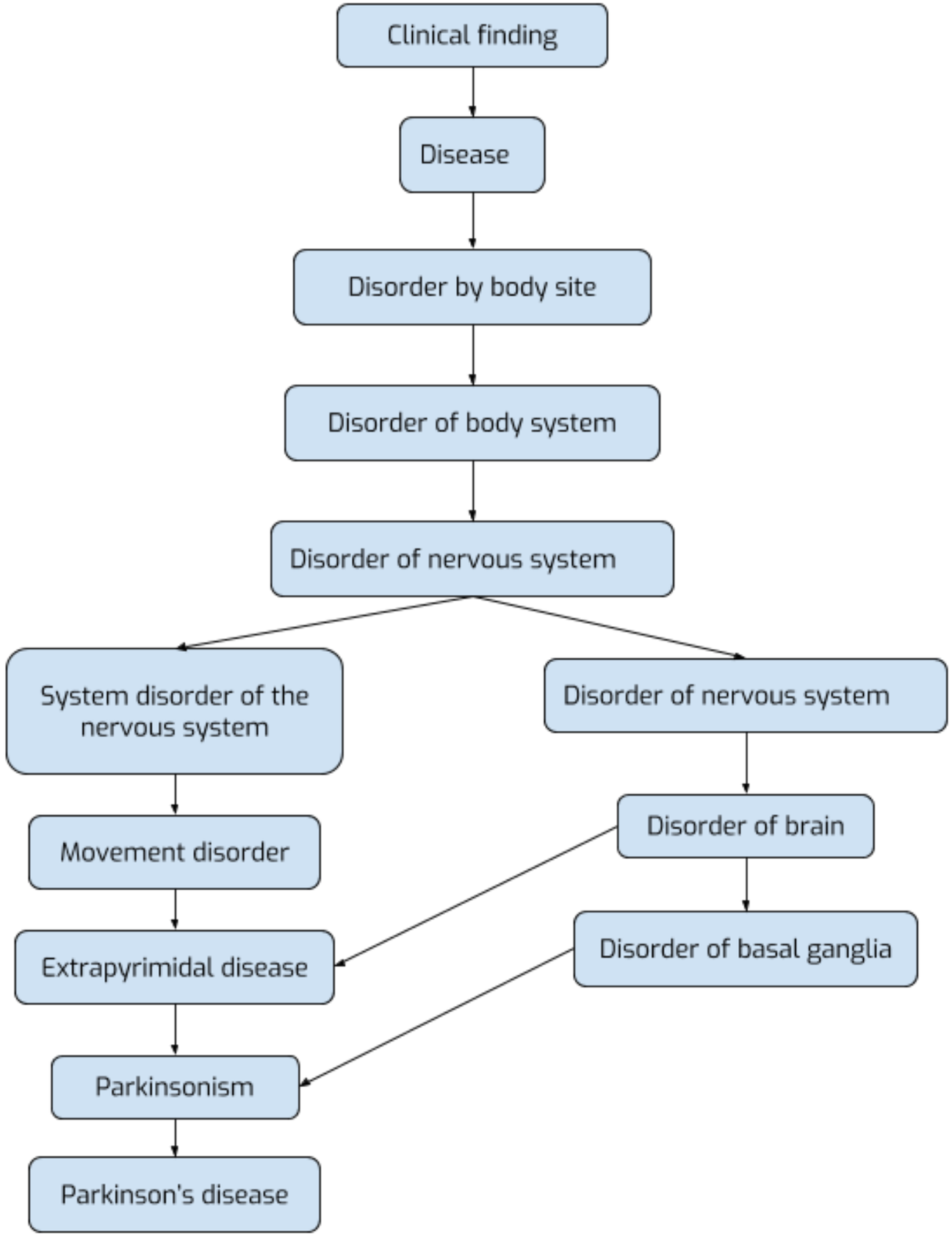
- Goals:
  - Structure
  - Query
  - Analyze
  - For all of the data

- Steps
  - Exploration
  - Cleaning
  - Enrichment
  - Standardization

Discrete data

- Goal: all discrete data is encoded in a standardized way.
- What should these standards be?
- Some data may already be coded (ad hoc, ICD-9/10, SNOMED-CT), most is not.
- Ontologies provide existing standards.

What is an ontology?



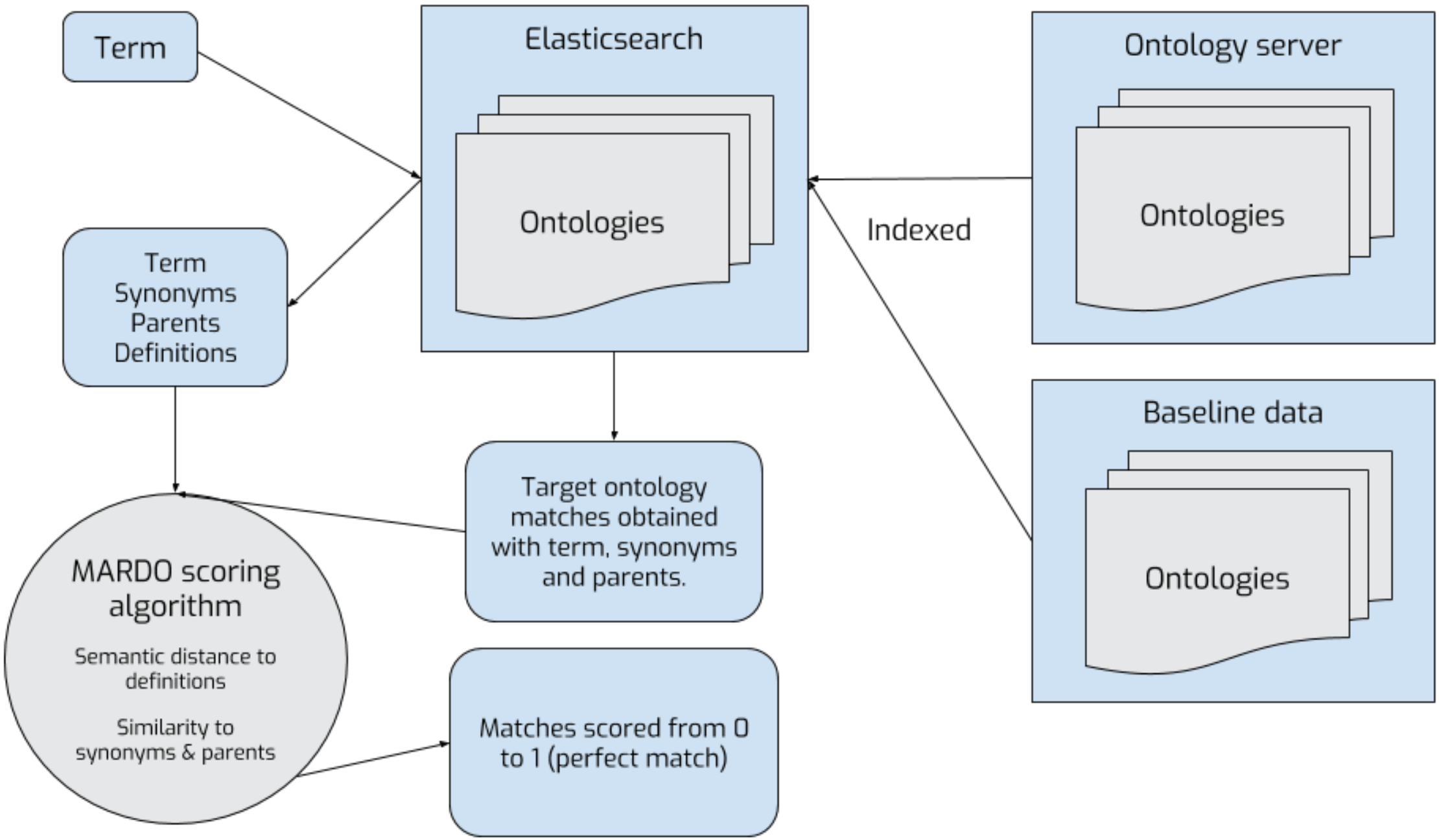
Discrete data

- Goal: all discrete data is encoded in a standardized way.
- What should these standards be?
- Some data may already be coded (e.g. ICD-10, SNOMED-CT), most is not.
- Ontologies provide existing standards.

- Manual mapping to them is very time consuming -> automation.

- Text to map -> search for matching terms in target ontology -> return the best matches along with a score indicating match quality

Automating ontology mapping



Discrete data & ontologies

- Goal: all discrete data is encoded in a standardized way.
- What should these standards be?
- Some data may already be coded (e.g. ICD-10, SNOMED-CT), most is not.
- Ontologies provide existing standards.

- Manual mapping to them is very time consuming -> automation.

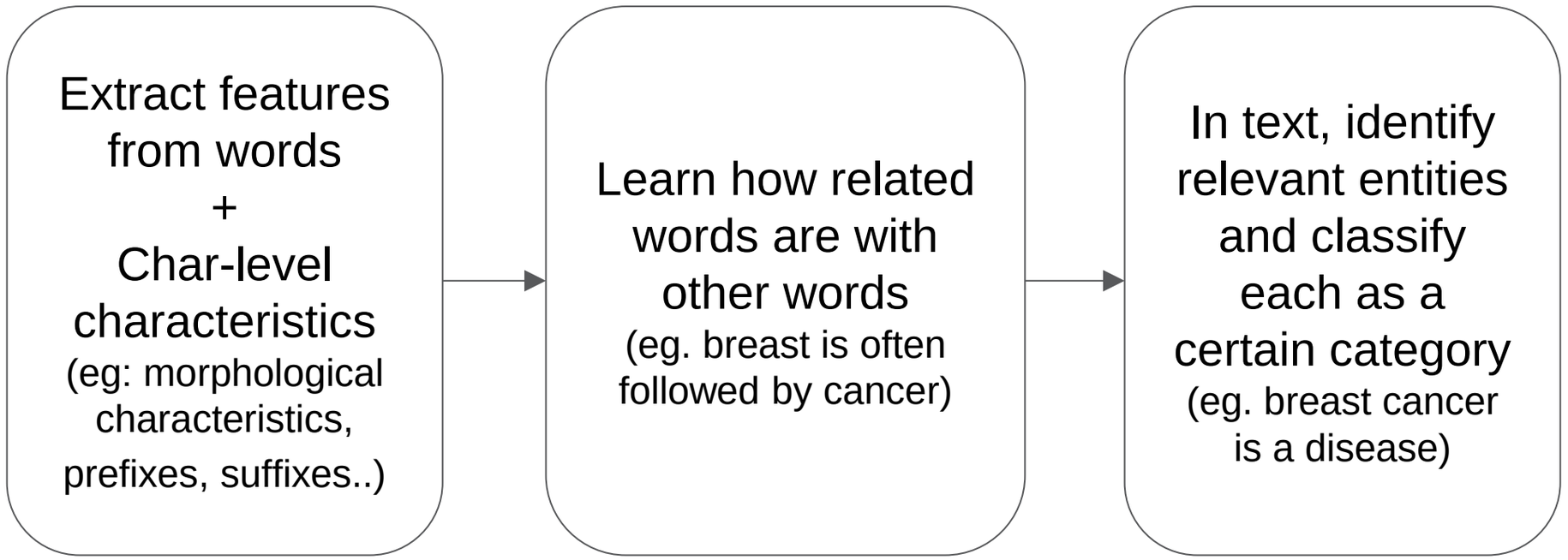
- Solutions:
  - Data cleaning and enrichment can be automated quite far.
  - Standardization through ontology mapping:
    - Automation
    - Learning

Narrative text

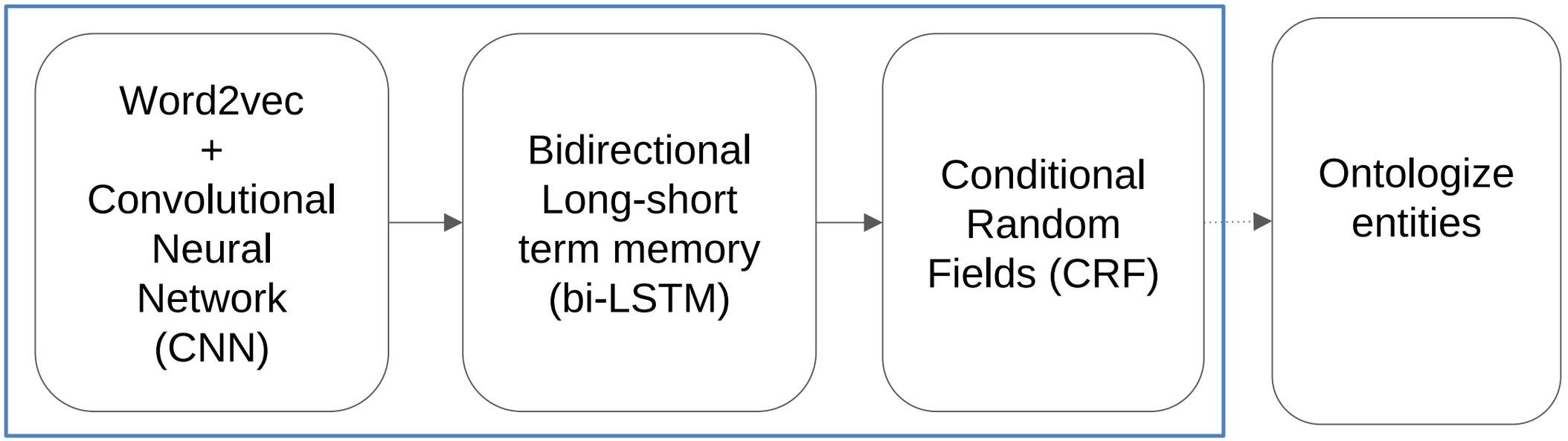
- Any free-form text, such as medical statements, pathology reports, etc.
- A wealth of information not found elsewhere.
- Nuance and context for discrete data.
- Varying goals:
  - Make searchable vs. understand

- Searchable: identify concepts in narrative text -> standardize their representation using ontology terms

Tagging of narrative text with ontology terms



NER: Algorithmic level



Summing it all up

- There are great opportunities in using EHR data for biobank samples.
- Significant challenges remain.
- Automation, learning and user experience.

Thank you!

Henrik Edgren, CSO
henrik.edgren@medisapiens.com

MediSapiens Ltd.
Mikonkatu 17C
00100 Helsinki
Finland

MediSapiens Inc.
One Broadway
Cambridge, MA 02142
USA

+358 45 8478878 (FI)
(781) 519-1801 (US)
contact@medisapiens.com
medisapiens.com