BBMRI.nl
Biobanking and
BioMolecular resources
Research Infrastructure
The Netherlands

BiobankConnect: a semi-automatic 'ETL' system for biobank data integration using MOLGENIS

Chao Pang[1,2], Jonathan Jetten[1], Dennis Hendriksen[1], Mark de Haan[1], Fleur Kelpin[1], Bart Charbon[1], Tommy de Boer[1], MOLGENIS team[1], BioSHaRE and BioMedBridges collaborators, Hans Hillege[2] and Morris A. Swertz[1]

1 University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Department of Genetics, Groningen, The Netherlands
2 University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, The Netherlands
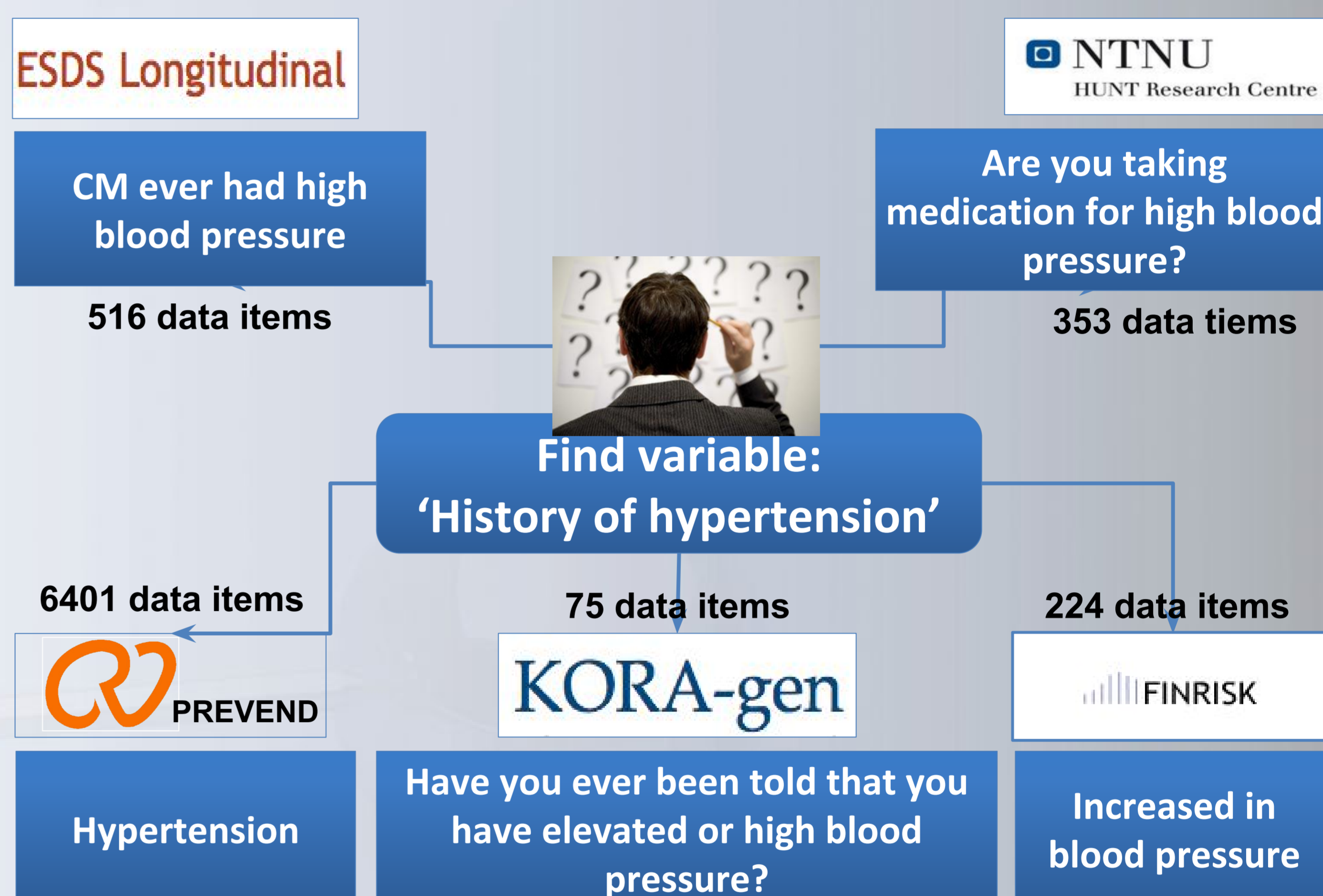
Overview

BiobankConnect aims to promote FAIR principles for data (Find, Access, Integrate and Reuse) for biobank data by enabling researchers to find, compare, and choose existing data models, formats, and guidelines. The toolbox comprises:

- **Meta-data model** – to uniformly capture the structural elements of standards models, formats, templates, data dictionaries, and guidelines so they can be compared.
- **Query interfaces** – to enable human users and scripts to search across all collected meta-data, drill down to the details of individual elements; and print the models.
- **Mapping system** – to generate, view and curate between-model mappings, to find a suitable standard for a ad-hoc uploaded data sheet, and to automatically integrate data using these models.

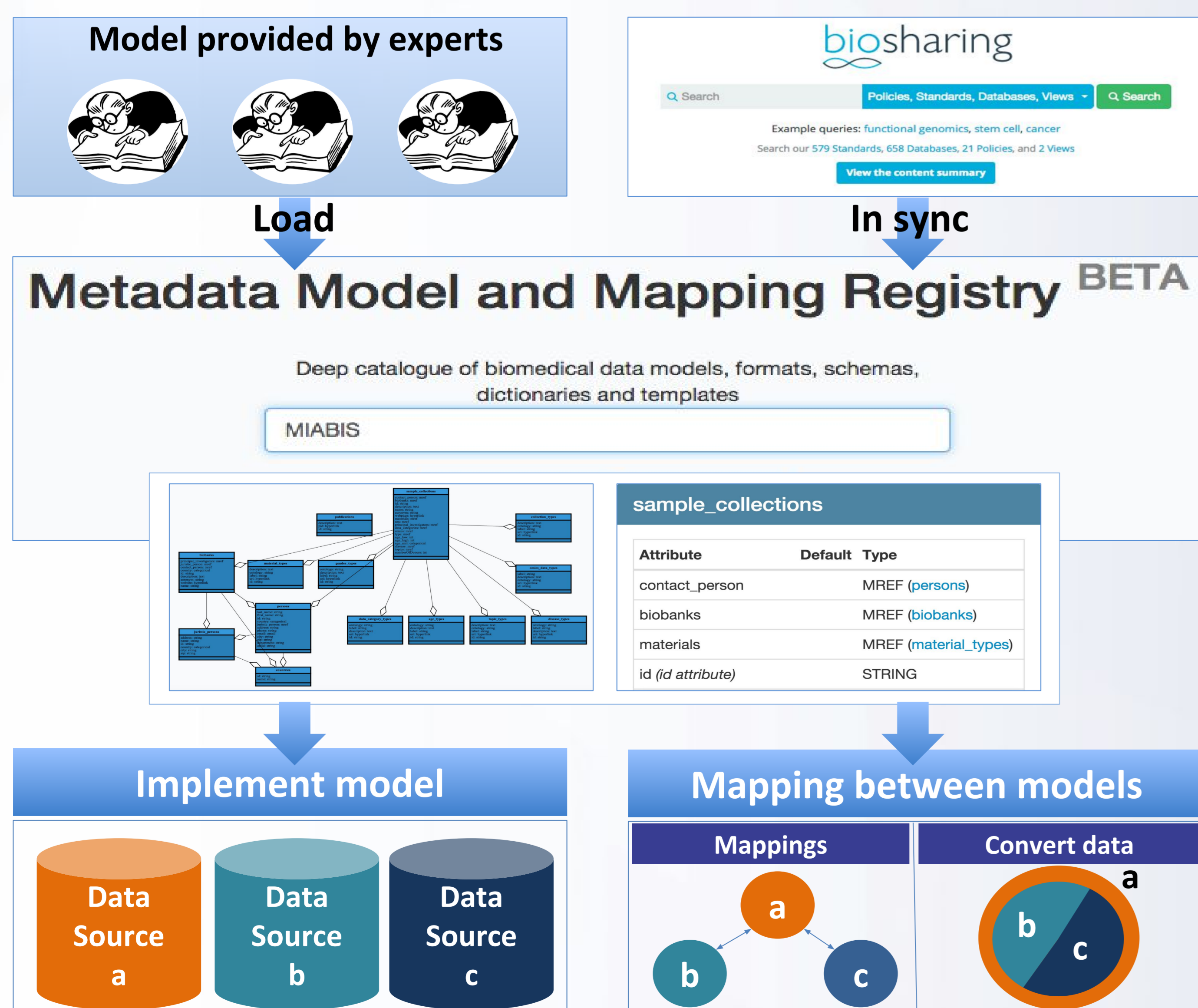BiobankConnect has been developed in collaboration with BioSHaRE, BioMedBridges, P3G and BBMRI.

Challenge

Manually searching for data items that can be pooled across data files or databases is very time-consuming because data items to be considered, are described using different structures, definitions and terminologies, e. g. in biobanking:

ESDS Longitudinal — CM ever had high blood pressure — 516 data items

NTNU HUNT Research Centre — Are you taking medication for high blood pressure? — 353 data tiems

Find variable: 'History of hypertension'

PREVEND — 6401 data items — Hypertension

KORA-gen — 75 data items — Have you ever been told that you have elevated or high blood pressure?

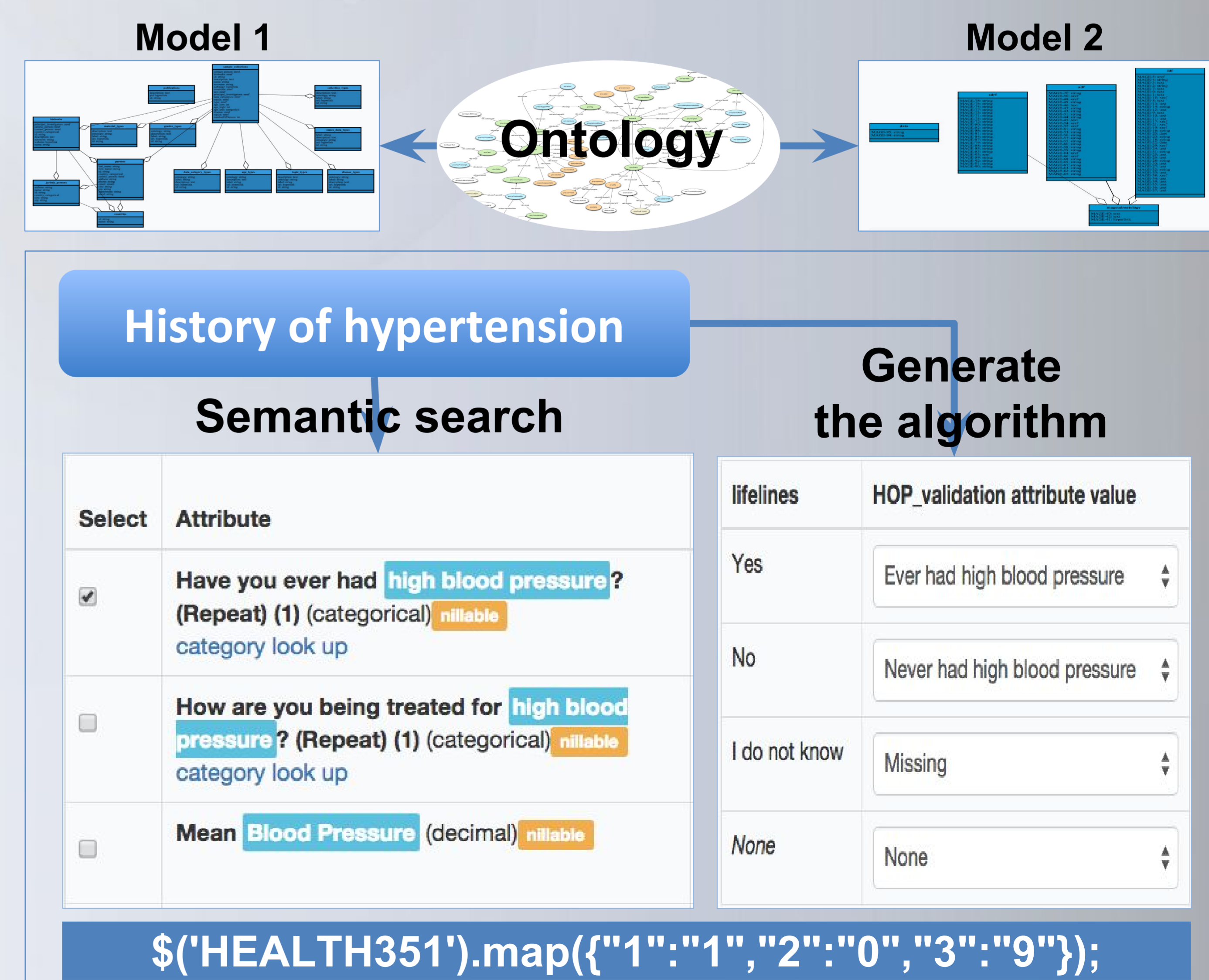FINRISK — 224 data items — Increased in blood pressure

Model registry

Metadata models are provided by domain experts and in sync with biosharing.org. Users can add models and find mappings between them.

Model provided by experts — Load
biosharing — In sync

Metadata Model and Mapping Registry BETA
Deep catalogue of biomedical data models, formats, schemas, dictionaries and templates
MIABIS

sample_collections

Attribute	Default	Type
contact_person		MREF (persons)
biobanks		MREF (biobanks)
materials		MREF (material_types)
id (id attribute)		STRING

Implement model — Data Source a, Data Source b, Data Source c

Mapping between models — Mappings — Convert data

Mapping tool

The data mapper semi-automatically generates convertors between model attributes using the information from ontologies via a semantic search process, e.g. in biobanking:

Model 1 — Ontology — Model 2

History of hypertension

Semantic search

Select	Attribute
☑	Have you ever had high blood pressure? (Repeat) (1) (categorical) nillable category look up
☐	How are you being treated for high blood pressure? (Repeat) (1) (categorical) nillable category look up
☐	Mean Blood Pressure (decimal) nillable

Generate the algorithm

lifelines	HOP_validation attribute value
Yes	Ever had high blood pressure
No	Never had high blood pressure
I do not know	Missing
None	None

$('HEALTH351').map({"1":"1","2":"0","3":"9"});