Relational and semantic data

(Background) knowledge, machine learning, and AI

- knowledge bases containing background knowledge are everywhere
- rich formal characterization (axioms)
- how can they be used for (predictive) data analysis?
  - "fuzzy", similarity-based search
  - predictive analysis and machine learning

Learning goals

- machine learning with structured background knowledge
- unsupervised or supervised:
  - here: mostly unsupervised *feature* learning
  - "deep" learning
- focus on existing tools and methods
  - Jupyter Notebooks and code examples
  - mOWL library
- not covered:
  - extracting knowledge (axioms, definitions) from data
  - (most) natural language processing
  - reasoning and deduction
  - machine learning theory

Agenda

- Introduction: knowledge bases and graphs
- Semantic similarity
- Machine learning with background knowledge
- applications
- (most hands-on components based on the mOWL library)

Classifying diseases

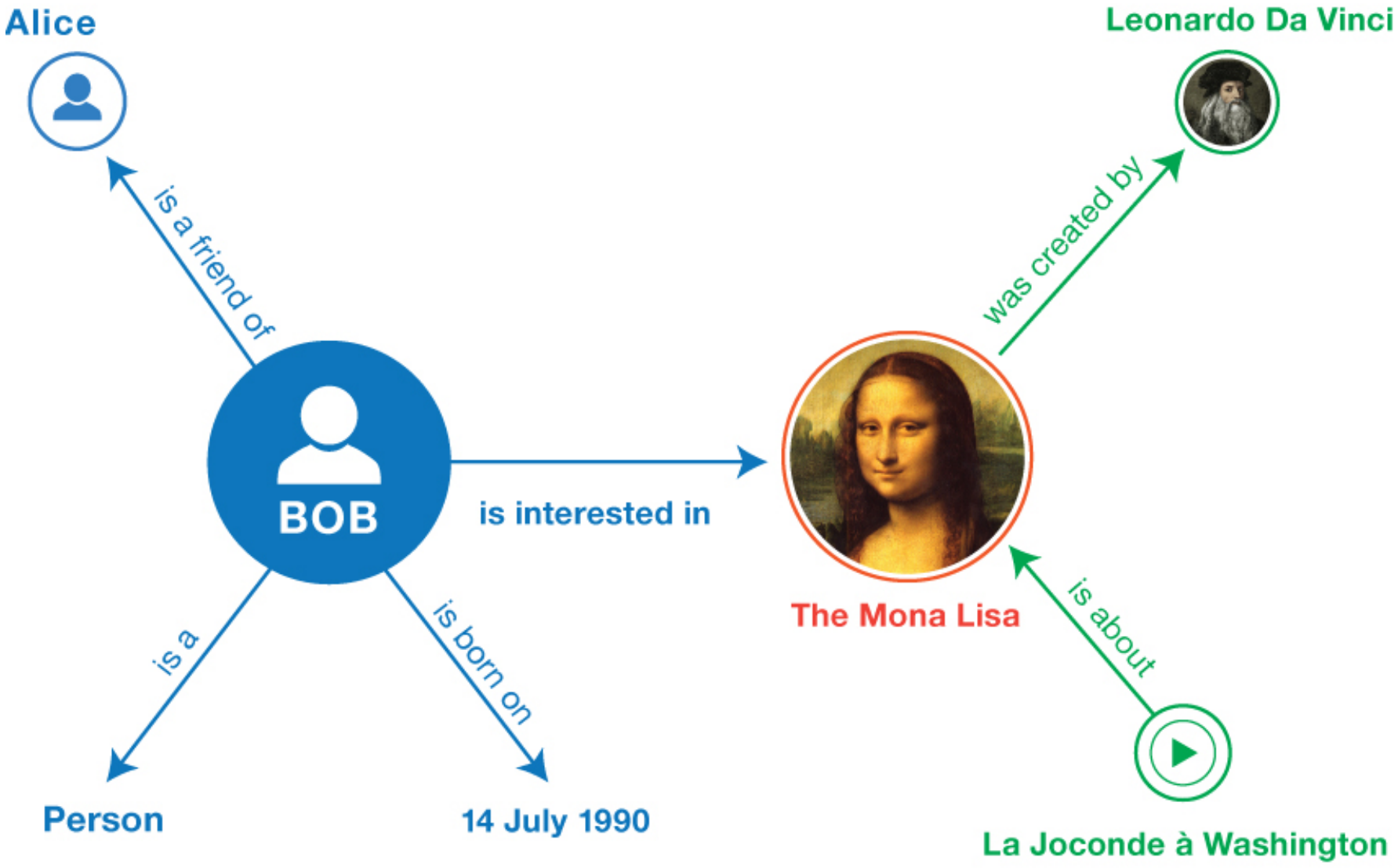
- Francois Bossier de Lacroix (18th century): Nosologia Methodica
- William Cullen of Edinburgh (1785): Synopsis nosologiae methodica
- William Farr (1837): improved classification, standardization, international collaboration
- Jaques Bertillon (1893): International (Bertillon) Classification of Causes of Death
- 1900: International Classification of Diseases (ICD) version 1
- 2022: International Classification of Diseases (ICD) version 11

Classifying diseases

Breast cancer:

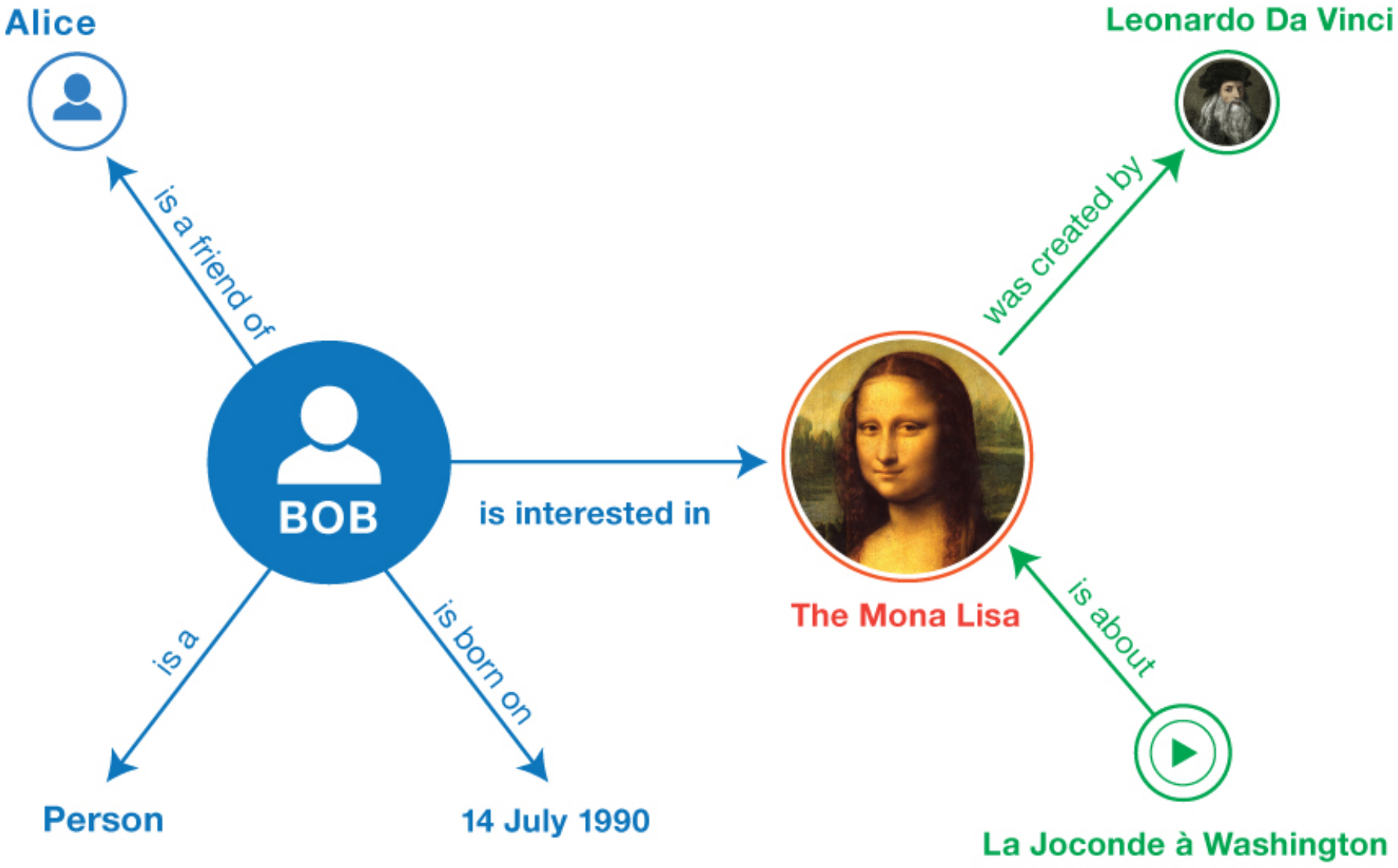
- specific anatomy
- histopathology
- grade and stage
- molecular subtype
- celltype of origin
- (etiology)

Relational data



- entities
- relations

Relational data

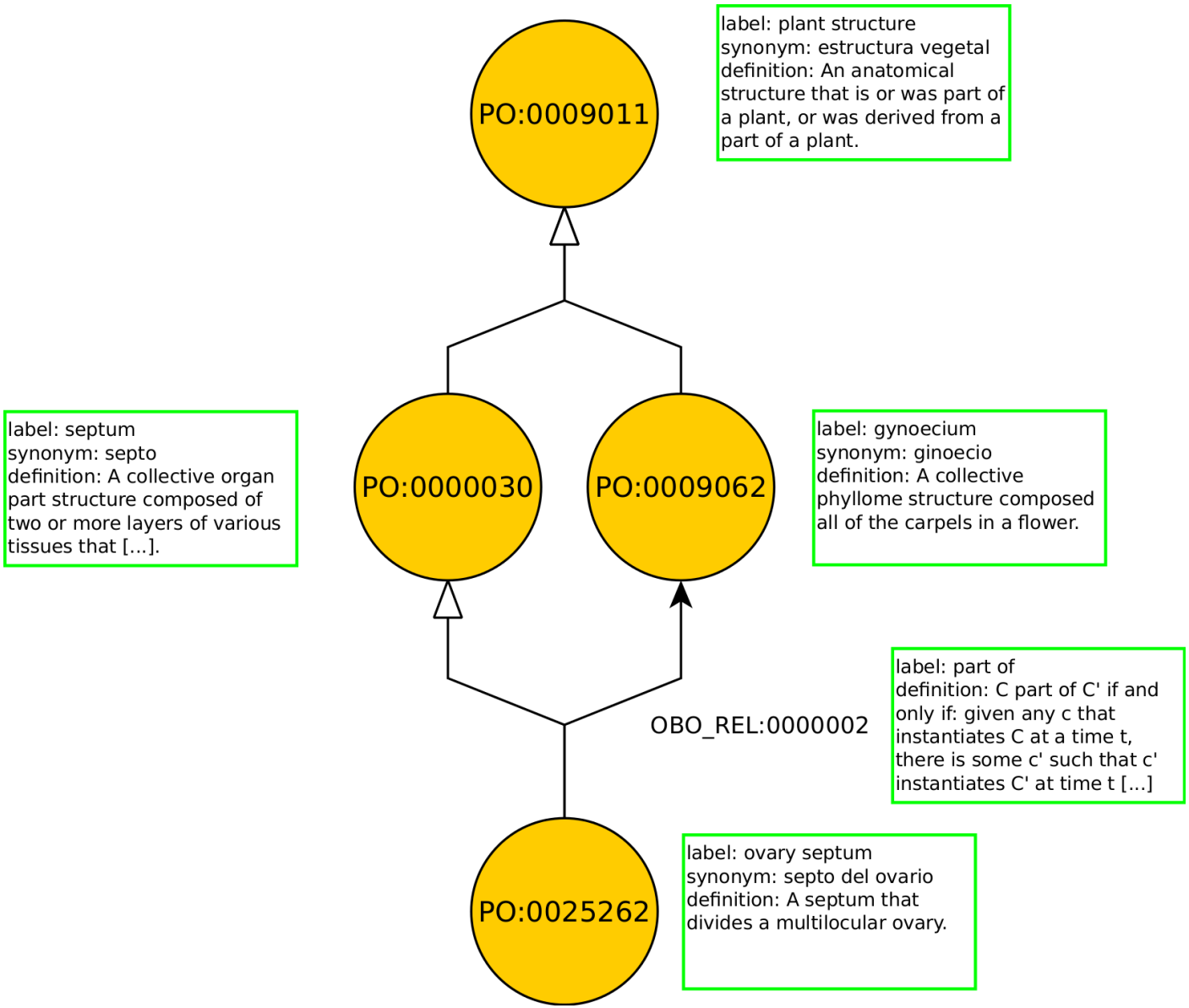


- every "breast cancer" is a type of cancer
- every person has exactly two parents

Axioms

- *classes* represent kinds of things in the world
  - *Arm*, *Apoptosis*, *Influenza*, *Homo sapiens*, *Drinking behavior*, *Membrane*
- *instances* of classes are individuals satisfying the classes' intension
  - my arm, the influenza I had last year, one ethanol molecule, etc.
- *relations* between instances arise from interactions, configurations, etc., of individuals
  - my arm is **part of** me, the **duration of** my influenza was 10 days
- *axioms* specify the conditions that instances of a class must satisfy
  - every instance of *Hand* is a **part of** an instance of *Arm*

Ontologies


PO:0009011
label: plant structure
synonym: estructura vegetal
definition: An anatomical structure that is or was part of a plant, or was derived from a part of a plant.
PO:0000030
label: septum
synonym: septo
definition: A collective organ part structure composed of two or more layers of various tissues that [...].
PO:0009062
label: gynoecium
synonym: ginoecio
definition: A collective phyllome structure composed all of the carpels in a flower.
OBO_REL:0000002
label: part of
definition: C part of C' if and only if: given any c that instantiates C at a time t, there is some c' such that c' instantiates C' at time t [...]
PO:0025262
label: ovary septum
synonym: septo del ovario
definition: A septum that divides a multilocular ovary.

Semantic similarity: some examples

- Are cyclin dependent kinases *functionally* more similar to lipid kinases or to riboflavin kinases? How about *phenotypically*?
- Which protein in the *mouse* is functionally most similar to the zebrafish *gustducin* protein?
- Which mouse knockout resembles *Bardet-Biedl Syndrome 8*?
- Are there mouse knockouts that resemble the side effects of diclofenac?
- Which genetic disease produces similar symptoms to ebola?
- Does functional similarity correlate with phenotypic similarity?

Semantic similarity

semantic similarity measures:

- for words, terms, classes
- role of background knowledge:
  - statistical/distributional semantics, large corpora
  - ontologies: (graph) topology
- similarity measures: hand-crafted or data-driven?

Semantic similarity or machine learning

- semantic similarity measures are mostly hand-crafted
  - capture certain intuition about what constitutes “similarity”
  - different measures for different kinds of similarity
  - usually interpretable (and explainable)

Semantic similarity or machine learning

- semantic similarity measures are mostly hand-crafted
  - capture certain intuition about what constitutes “similarity”
  - different measures for different kinds of similarity
  - usually interpretable (and explainable)
- machine learning methods are mostly data-driven
  - the architecture of the model is still hand-crafted
  - usually hard to interpret

Knowledge bases and graphs

- semantic similarity measures *and machine learning models* on knowledge bases (ontologies) can be graph-based, feature-based, or model-based
  - graph-based: ontology as a graph
  - feature-based: extract (or obtain) features for classes/relations
  - model-based: define similarity within (special) Σ-algebras