

# Part 1: Ontologies and semantic similarity

Robert Hoehndorf

# Before the tutorial

See <https://github.com/bio-ontology-research-group/ontology-tutorial>:

- ▶ install Docker (e.g.: `apt-get install docker`)
- ▶ `docker run -i -t -p 8888:8888  
leechuck/ai-biomed-summer-school`
- ▶ in your browser, connect to `localhost:8888`

# Learning goals

- ▶ brief overview of ontologies in biomedicine
- ▶ semantic similarity with ontologies
- ▶ machine learning with ontologies as *features* (or background knowledge)
- ▶ unsupervised or supervised:
  - ▶ here: mostly unsupervised *feature* learning
  - ▶ “deep” learning
- ▶ focus on existing tools and methods
  - ▶ Jupyter Notebooks and code examples
- ▶ not covered:
  - ▶ learning ontologies (axioms, definitions) from data
  - ▶ (most) natural language processing
  - ▶ reasoning with ontologies
  - ▶ learning on “knowledge graphs”
  - ▶ machine learning theory

# Learning goals

## Biomedical questions:

- ▶ diagnosing rare disease
- ▶ finding functionally similar proteins
- ▶ relying on heterogeneous data integration
  - ▶ from different databases
  - ▶ model organisms

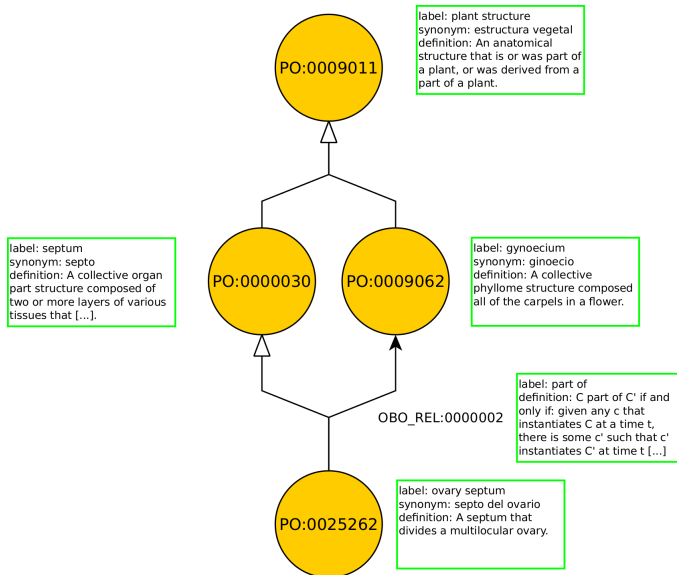
# Agenda

- ▶ Introduction: ontologies and graphs
- ▶ Semantic similarity
- ▶ Machine learning:
  - ▶ syntactic
  - ▶ graph-based
  - ▶ model-theoretic

# Ontologies, machine learning, and AI

- ▶ ontologies are ubiquitous in biomedical research
- ▶ rich formal characterization (axioms)
- ▶ how can they be used for (predictive) data analysis?
  - ▶ “fuzzy”, similarity-based search
  - ▶ background knowledge in machine learning

# Ontologies



# Ontologies for data integration

- ▶ standard identifiers (IRIs)
- ▶ labels of classes and relations
- ▶ human-readable descriptions
- ▶ axioms

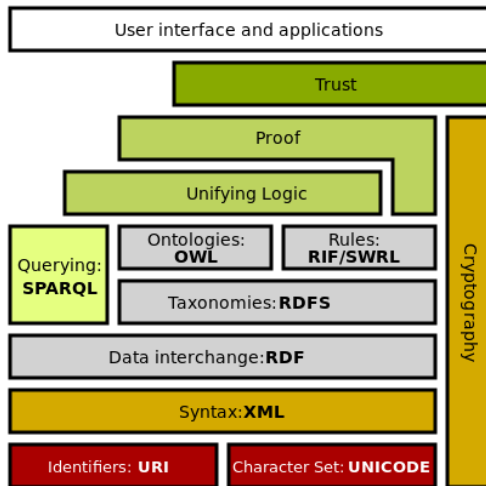


# Ontologies provide domain knowledge

[Overview](#)
[Browse](#)
[DLQuery](#)
[Download](#)

Annotation	Value
label	B cell apoptotic process
definition	Any apoptotic process in a B cell, a lymphocyte of B lineage with the phenotype CD19-positive and capable of B cell mediated immunity.
class	<a href="http://purl.obolibrary.org/obo/GO_0001783">http://purl.obolibrary.org/obo/GO_0001783</a>
ontology	GO-PLUS
Equivalent	<a href="#">apoptotic process</a> and ( <a href="#">occurs in some B cell</a> )
SubClassOf	<a href="#">occurs in some B cell</a> , <a href="#">lymphocyte apoptotic process</a>
id	GO:0001783
has_obo_namespace	biological_process

# The Semantic Web



# Manchester OWL Syntax

DL Syntax	Manchester Syntax	Example
$C \sqcap D$	C and D	Human and Male
$C \sqcup D$	C or D	Male or Female
$\neg C$	not C	not Male
$\exists R.C$	R some C	hasChild some Human
$\forall R.C$	R only C	hasChild only Human
$(\geq nR.C)$	R min n C	hasChild min 1 Human
$(\leq nR.C)$	R max n C	hasChild max 1 Human
$(= nR.C)$	R exactly n C	hasChild exactly 1 Human
$\{a\} \sqcup \{b\} \sqcup \dots$	{a b ...}	{John Robert Mary}

# Reasoning with ontologies

- ▶ 'B cell apoptosis' EquivalentTo: apoptosis and 'occurs in' some 'B cell'
- ▶ 'lymphocyte apoptosis' EquivalentTo: apoptosis and 'occurs in' some lymphocyte
- ▶ 'B cell' SubClassOf: lymphocyte

# Reasoning with ontologies

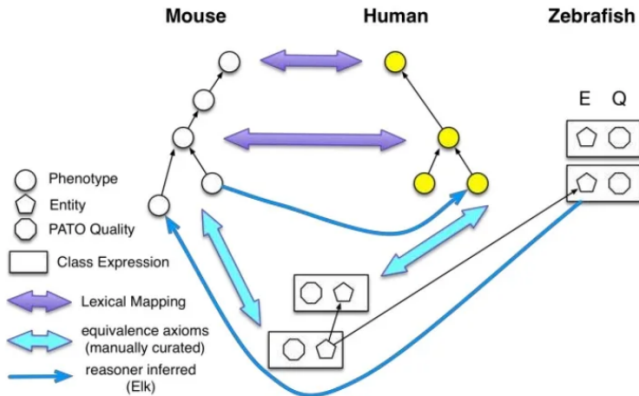
- ▶ 'B cell apoptosis' EquivalentTo: apoptosis and 'occurs in' some 'B cell'
- ▶ 'lymphocyte apoptosis' EquivalentTo: apoptosis and 'occurs in' some lymphocyte
- ▶ 'B cell' SubClassOf: lymphocyte
- ▶ 'B cell apoptosis' SubClassOf: 'lymphocyte apoptosis'

# Reasoning with phenotype ontologies

## Phenotype ontology:

- ▶ A phenotype is a quality that inheres in its bearer:
  - ▶ Enlarged heart:  $Enlarged(x) \wedge \exists y (inheresIn(x, y) \wedge Heart(y))$
  - ▶ in DL:  $Enlarged \sqcap \exists inheresIn. Heart$
  - ▶ Enlarged: reuse an ontology of qualities (PATO)
  - ▶ Heart: reuse ontology of (human, mammalian) anatomy

# UberPheno, PhenomeNET



# BLAST-like search over phenotypes

Using ontologies and reasoning, we can

1. describe phenotypes computationally
  - ▶ morphology
  - ▶ function
  - ▶  $\Rightarrow$  ontologies
2. integrate/compare phenotypes *within and between species* (to overcome limitations in phenotype data)
  - ▶ homologous organ structures
  - ▶ related/identical function
  - ▶  $\Rightarrow$  ontologies and automated reasoning
3. measure phenotypic similarity
  - ▶ use morphological or functional similarity
  - ▶ similarity in attribute values
  - ▶  $\Rightarrow$  semantic similarity, machine learning



## Other questions we want to answer

- ▶ Are cyclin dependent kinases *functionally* more similar to lipid kinases or to riboflavin kinases? How about *phenotypically*?

## Other questions we want to answer

- ▶ 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?

## Other questions we want to answer

- ▶ 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*?

## Other questions we want to answer

- ▶ 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?

## Other questions we want to answer

- ▶ 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?

## Other questions we want to answer

- ▶ 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?

# Ontologies and graphs

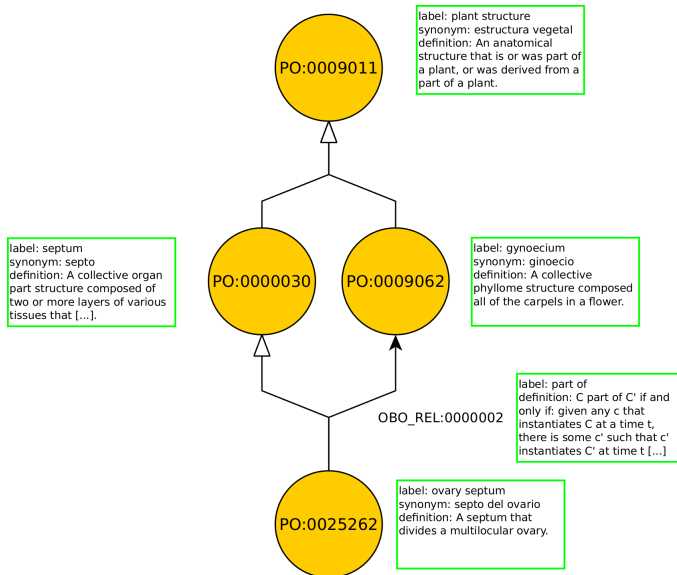
- ▶ semantic similarity measures can be graph-based, feature-based, or model-based
- ▶ we may need to generate graphs from ontologies
  - ▶ remember: ontologies are sets of axioms
  - ▶ *subclass* axioms are easy
  - ▶ how about more complex axioms?
- ▶ solution: define relational patterns

# Relations as patterns

- ▶  $X \text{ SubClassOf } Y: X \xrightarrow{\text{is-a}} Y$
- ▶  $X \text{ SubClassOf } \text{part-of some } Y: X \xrightarrow{\text{part-of}} Y$
- ▶  $X \text{ SubClassOf } \text{regulates some } Y: X \xrightarrow{\text{regulates}} Y$
- ▶  $X \text{ DisjointWith } Y: X \xleftrightarrow{\text{disjoint}} Y$
- ▶  $X \text{ EquivalentTo } Y: X \xleftrightarrow{\equiv} Y, \{X, Y\}$



# Relations as patterns



## Semantic similarity

- ▶ We want to use *background knowledge* in ontologies to
  - ▶ determine similarity between classes,
  - ▶ instances,
  - ▶ and entities with ontology annotations

# How to measure similarity?

- ▶ semantic similarity measures similarity between classes
- ▶ semantic similarity measures similarity between instances of classes
- ▶ semantic similarity measures similarity between entities *annotated* with classes
- ▶  $\Rightarrow$  reduce all of this to similarity between classes

# How to measure similarity?

What properties do we want in a similarity measure?

A function  $\text{sim} : D \times D$  is a similarity on  $D$  if, for all  $x, y \in D$ , the function  $\text{sim}$  is:

# How to measure similarity?

What properties do we want in a similarity measure?

A function  $sim : D \times D$  is a similarity on  $D$  if, for all  $x, y \in D$ , the function  $sim$  is:

- ▶ non-negative:  $sim(x, y) \geq 0$  for all  $x, y$

# How to measure similarity?

What properties do we want in a similarity measure?

A function  $sim : D \times D$  is a similarity on  $D$  if, for all  $x, y \in D$ , the function  $sim$  is:

- ▶ non-negative:  $sim(x, y) \geq 0$  for all  $x, y$
- ▶ symmetric:  $sim(x, y) = sim(y, x)$

# How to measure similarity?

What properties do we want in a similarity measure?

A function  $sim : D \times D$  is a similarity on  $D$  if, for all  $x, y \in D$ , the function  $sim$  is:

- ▶ non-negative:  $sim(x, y) \geq 0$  for all  $x, y$
- ▶ symmetric:  $sim(x, y) = sim(y, x)$
- ▶ reflexive:  $sim(x, x) = \max_D$

# How to measure similarity?

What properties do we want in a similarity measure?

A function  $sim : D \times D$  is a similarity on  $D$  if, for all  $x, y \in D$ , the function  $sim$  is:

- ▶ non-negative:  $sim(x, y) \geq 0$  for all  $x, y$
- ▶ symmetric:  $sim(x, y) = sim(y, x)$
- ▶ reflexive:  $sim(x, x) = \max_D$ 
  - ▶ weaker form:  $sim(x, x) > sim(x, y)$  for all  $x \neq y$



# How to measure similarity?

What properties do we want in a similarity measure?

A function  $sim : D \times D$  is a similarity on  $D$  if, for all  $x, y \in D$ , the function  $sim$  is:

- ▶ non-negative:  $sim(x, y) \geq 0$  for all  $x, y$
- ▶ symmetric:  $sim(x, y) = sim(y, x)$
- ▶ reflexive:  $sim(x, x) = \max_D$ 
  - ▶ weaker form:  $sim(x, x) > sim(x, y)$  for all  $x \neq y$
- ▶  $sim(x, x) > sim(x, y)$  for  $x \neq y$

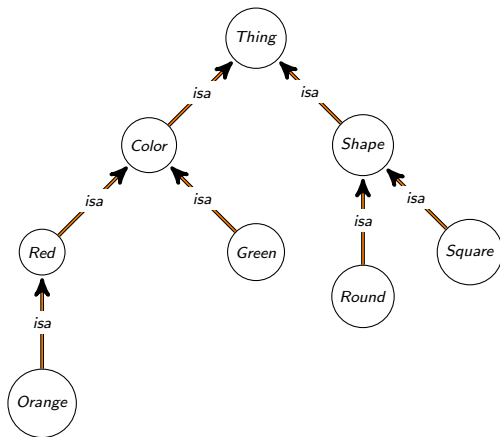
# How to measure similarity?

What properties do we want in a similarity measure?

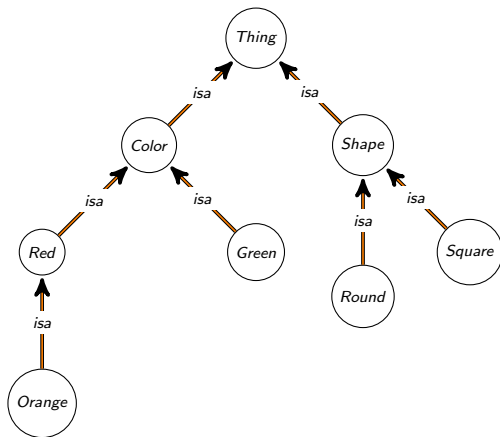
A function  $sim : D \times D$  is a similarity on  $D$  if, for all  $x, y \in D$ , the function  $sim$  is:

- ▶ non-negative:  $sim(x, y) \geq 0$  for all  $x, y$
- ▶ symmetric:  $sim(x, y) = sim(y, x)$
- ▶ reflexive:  $sim(x, x) = \max_D$ 
  - ▶ weaker form:  $sim(x, x) > sim(x, y)$  for all  $x \neq y$
- ▶  $sim(x, x) > sim(x, y)$  for  $x \neq y$
- ▶  $sim$  is a *normalized* similarity measure if it has values in  $[0, 1]$

# How to measure similarity?

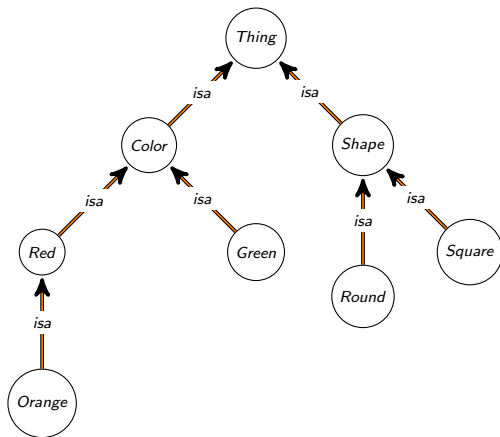


# How to measure similarity?



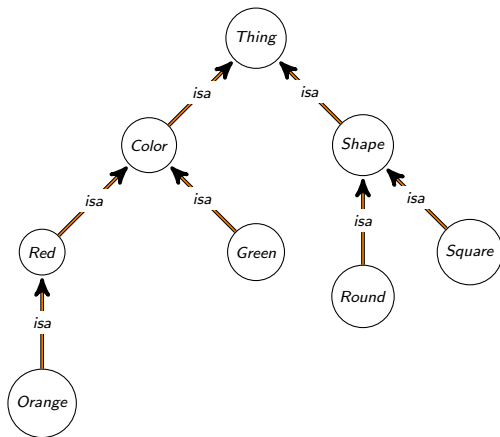
- distance on shortest path (Rada *et al.*, 1989)

# How to measure similarity?



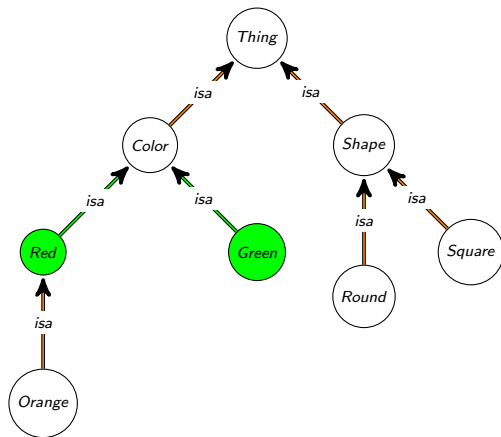
- ▶ distance on shortest path (Rada *et al.*, 1989)
- ▶  $dist_{Rada}(u, v) = sp(u, isa, v)$

# How to measure similarity?



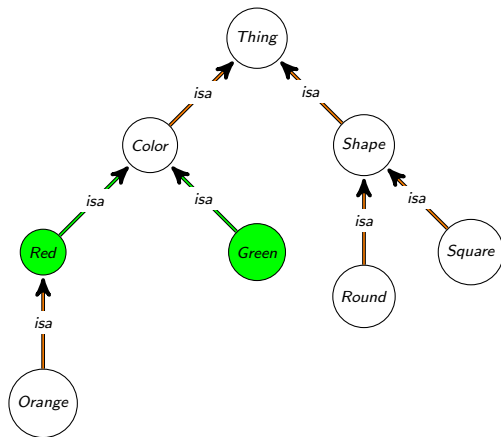
- ▶ distance on shortest path (Rada *et al.*, 1989)
- ▶  $dist_{Rada}(u, v) = sp(u, isa, v)$
- ▶  $sim_{Rada}(u, v) = \frac{1}{dist_{Rada}(u, v) + 1}$

# How to measure similarity?



- distance on shortest path

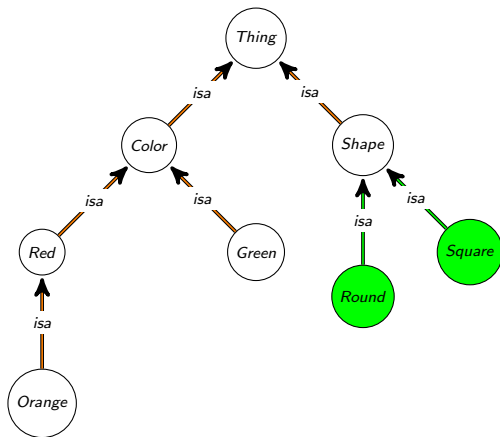
# How to measure similarity?



- ▶ distance on shortest path
- ▶  $\text{distance}(\text{green}, \text{red}) = 2$
- ▶  $\text{sim}_{\text{Rada}}(\text{green}, \text{red}) = \frac{1}{3}$



# How to measure similarity?



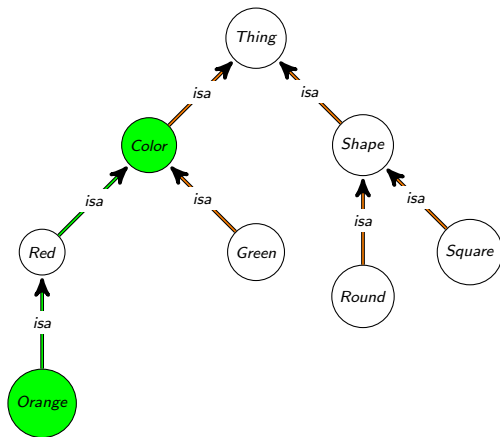
► distance on shortest path

►  $\text{distance}(\text{square}, \text{round}) = 2$

►

$$\text{sim}_{\text{Rada}}(\text{square}, \text{round}) = \frac{1}{3}$$

# How to measure similarity?



► distance on shortest path

►  $\text{distance}(\text{orange}, \text{color}) = 2$

►

$$\text{sim}_{\text{Rada}}(\text{orange}, \text{color}) = \frac{1}{3}$$

# How to measure similarity?

- ▶ shortest path is not always intuitive

# How to measure similarity?

- ▶ shortest path is not always intuitive
- ▶ we need a way to determine *specificity* of a class
  - ▶ number of ancestors
  - ▶ number of children
  - ▶ information content

# How to measure similarity?

- ▶ shortest path is not always intuitive
- ▶ we need a way to determine *specificity* of a class
  - ▶ number of ancestors
  - ▶ number of children
  - ▶ information content
- ▶ *density* of a branch in the ontology
  - ▶ number of siblings
  - ▶ information content

# How to measure similarity?

- ▶ shortest path is not always intuitive
- ▶ we need a way to determine *specificity* of a class
  - ▶ number of ancestors
  - ▶ number of children
  - ▶ information content
- ▶ *density* of a branch in the ontology
  - ▶ number of siblings
  - ▶ information content
- ▶ account for different edge types
  - ▶ non-uniform edge weighting

# How to measure similarity

- ▶ term specificity measure  $\sigma : \mathcal{C} \mapsto \mathbb{R}$ :
  - ▶  $x \sqsubseteq y \rightarrow \sigma(x) \geq \sigma(y)$

# How to measure similarity

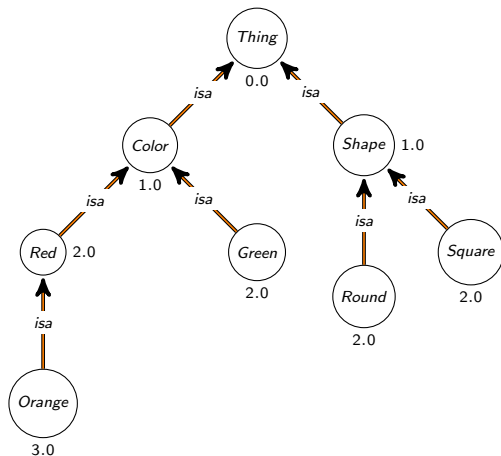
- ▶ term specificity measure  $\sigma : C \mapsto \mathbb{R}$ :
  - ▶  $x \sqsubseteq y \rightarrow \sigma(x) \geq \sigma(y)$
- ▶ intrinsic:
  - ▶  $\sigma(x) = f(\text{depth}(x))$
  - ▶  $\sigma(x) = f(A(x))$  (for ancestors  $A(x)$ )
  - ▶  $\sigma(x) = f(D(x))$  (for descendants  $D(x)$ )
  - ▶ many more, e.g., Zhou et al.:
$$\sigma(x) = k \cdot \left(1 - \frac{\log |D(x)|}{\log |C|}\right) + (1 - k) \frac{\log \text{depth}(x)}{\log \text{depth}(G_T)}$$



# How to measure similarity

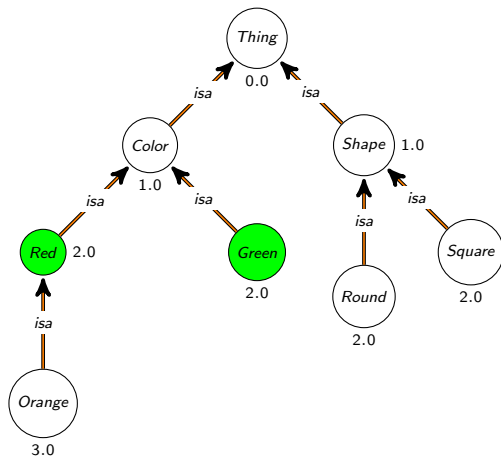
- ▶ term specificity measure  $\sigma : \mathcal{C} \mapsto \mathbb{R}$ :
  - ▶  $x \sqsubseteq y \rightarrow \sigma(x) \geq \sigma(y)$
- ▶ intrinsic:
  - ▶  $\sigma(x) = f(\text{depth}(x))$
  - ▶  $\sigma(x) = f(A(x))$  (for ancestors  $A(x)$ )
  - ▶  $\sigma(x) = f(D(x))$  (for descendants  $D(x)$ )
  - ▶ many more, e.g., Zhou et al.:
$$\sigma(x) = k \cdot \left(1 - \frac{\log |D(x)|}{\log |\mathcal{C}|}\right) + (1 - k) \frac{\log \text{depth}(x)}{\log \text{depth}(G_T)}$$
- ▶ extrinsic:
  - ▶  $\sigma(x)$  defined as a function of instances (or annotations)  $I$ 
    - ▶ note: the number of instances monotonically decreases with increasing depth in taxonomies
  - ▶ Resnik 1995:  $eIC_{\text{Resnik}}(x) = -\log p(x)$  (with  $p(x) = \frac{|I(x)|}{|I|}$ )
    - ▶ in biology, one of the most popular specificity measure when annotations are present

# How to measure similarity?



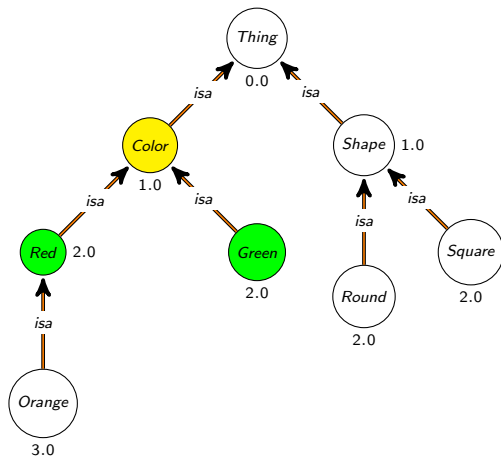
- Resnik 1995:  
similarity between  $x$  and  $y$  is the information content of the *most informative common ancestor*

# How to measure similarity?



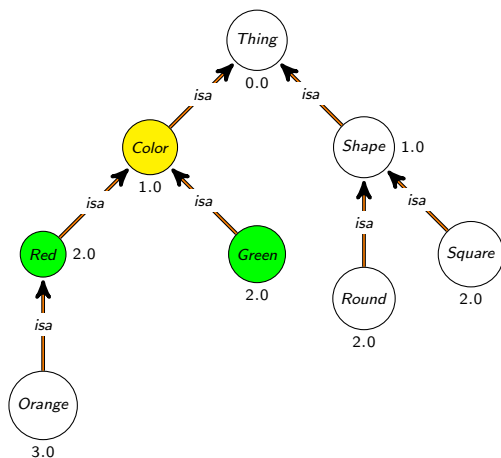
- Resnik 1995:  
similarity between  $x$  and  $y$  is the  
information content  
of the *most  
informative common  
ancestor*

# How to measure similarity?



- Resnik 1995:  
similarity between  $x$  and  $y$  is the information content of the *most informative common ancestor*

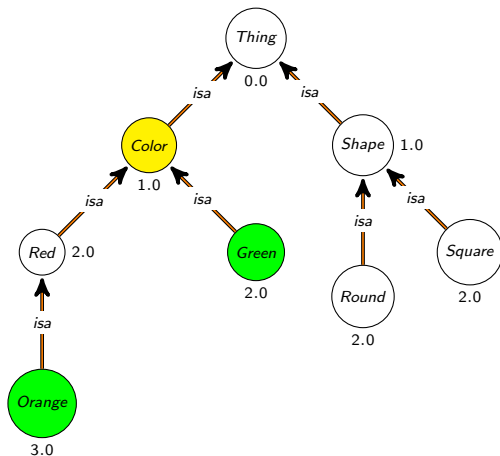
# How to measure similarity?



- ▶ Resnik 1995:  
similarity between  $x$  and  $y$  is the information content of the *most informative common ancestor*

- ▶ 
$$\text{sim}_{\text{Resnik}}(\text{Green}, \text{Red}) = 1.0$$

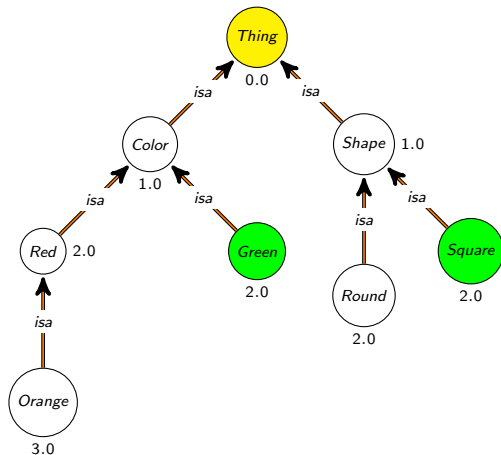
# How to measure similarity?



- ▶ Resnik 1995:  
similarity between  $x$  and  $y$  is the  
information content  
of the *most informative common ancestor*

- ▶ 
$$\text{sim}_{\text{Resnik}}(\text{Green}, \text{Orange}) = 1.0$$

# How to measure similarity?



- ▶ Resnik 1995:  
similarity between  $x$   
and  $y$  is the  
information content  
of the *most  
informative common  
ancestor*

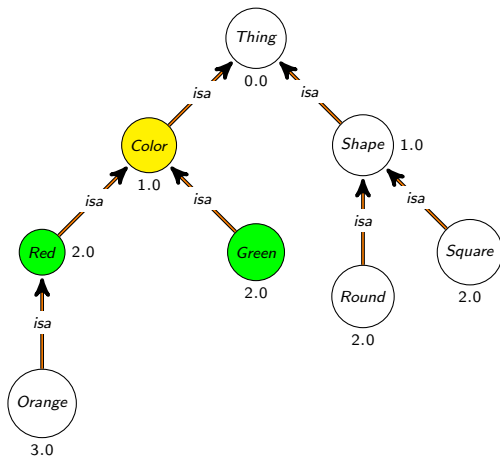
- ▶  $sim_{Resnik}(Square, Orange)$   
0.0

# How to measure similarity?

- ▶ (Red, Green) and (Orange, Green) have the same similarity
- ▶ need to incorporate the specificity of the compared classes



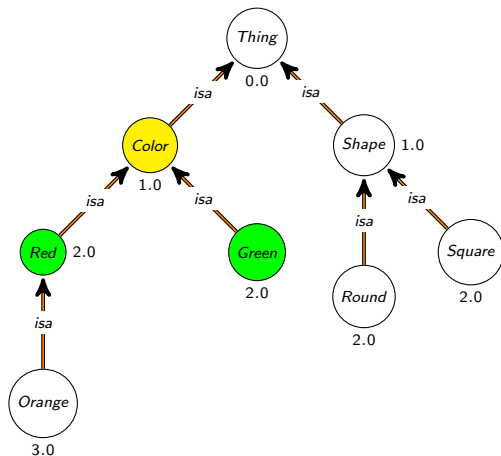
# How to measure similarity?



► Lin 1998:

$$sim_{Lin}(x, y) = \frac{2 \cdot IC(MICA(x, y))}{IC(x) + IC(y)}$$

# How to measure similarity?

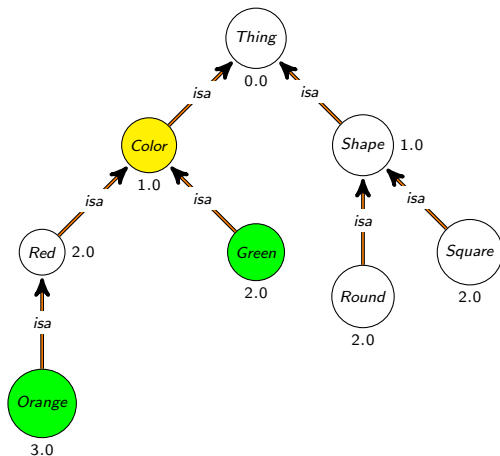


► Lin 1998:

$$sim_{Lin}(x, y) = \frac{2 \cdot IC(MICA(x, y))}{IC(x) + IC(y)}$$

►  $sim_{Lin}(Green, Red) = 0.5$

# How to measure similarity?



► Lin 1998:

$$sim_{Lin}(x, y) = \frac{2 \cdot IC(MICA(x, y))}{IC(x) + IC(y)}$$

►

$$sim_{Lin}(Green, Orange) = 0.4$$

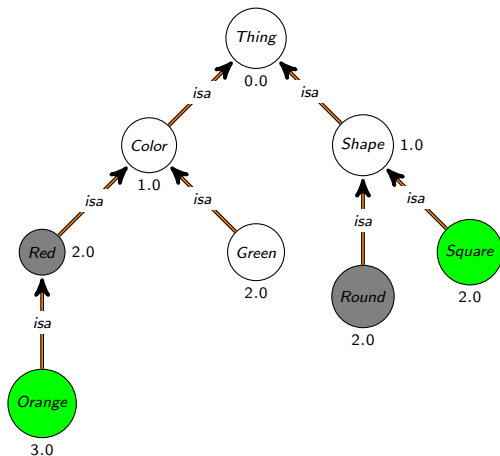
# How to measure similarity?

- ▶ many(!) others:
  - ▶ Jiang & Conrath 1997
  - ▶ Mazandu & Mulder 2013
  - ▶ Schlicker et al. 2009
  - ▶ ...

# How to measure similarity?

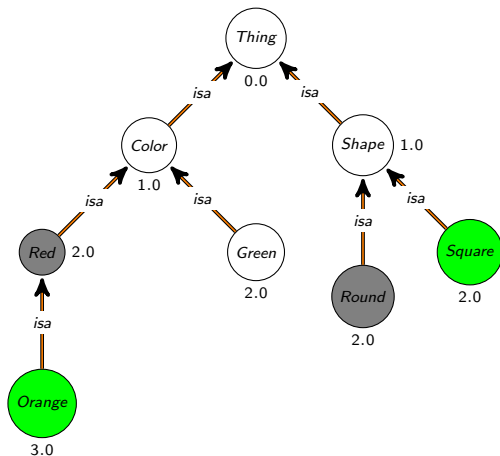
- ▶ we only looked at comparing pairs of classes
- ▶ mostly, we want to compare *sets* of classes
  - ▶ set of GO annotations
  - ▶ set of signs and symptoms
  - ▶ set of phenotypes
- ▶ two approaches:
  - ▶ compare each class individually, then merge
  - ▶ directly set-based similarity measures

# How to measure similarity?



- similarity between a square-and-orange thing and a round-and-red thing

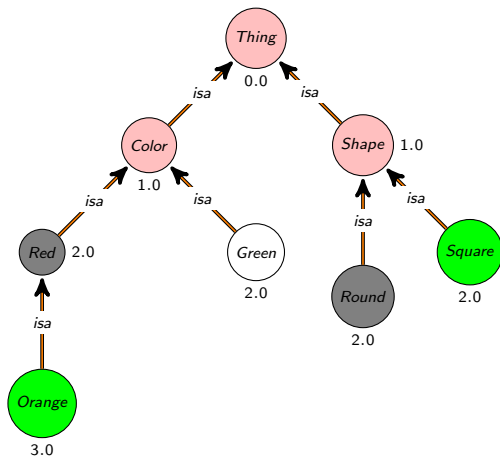
# How to measure similarity?



- ▶ similarity between a square-and-orange thing and a round-and-red thing
- ▶ Pesquita et al., 2007:

$$\text{simGIC}(X, Y) = \frac{\sum_{c \in A(X) \cap A(Y)} IC(c)}{\sum_{c \in A(X) \cup A(Y)} IC(c)}$$

# How to measure similarity?



- ▶ similarity between a square-and-orange thing and a round-and-red thing
- ▶ Pesquita et al., 2007:  
$$\text{simGIC}(X, Y) = \frac{\sum_{c \in A(X) \cap A(Y)} IC(c)}{\sum_{c \in A(X) \cup A(Y)} IC(c)}$$
- ▶  $\text{simGIC}(so, rr) = \frac{2}{11}$



# How to measure similarity?

- ▶ alternatively: use different merging strategies
- ▶ common: average, maximum, **best-matching average**
  - ▶ Average:  $sim_A(X, Y) = \frac{\sum_{x \in X} \sum_{y \in Y} sim(x, y)}{|X| \times |Y|}$
  - ▶ Max average:  $sim_{MA}(X, Y) = \frac{1}{|X|} \sum_{x \in X} \max_{y \in Y} sim(x, y)$
  - ▶ Best match average:  $sim_{BMA}(X, Y) = \frac{sim_{MA}(X, Y) + sim_{MA}(Y, X)}{2}$

# How to measure similarity?

- ▶ Semantic Measures Library:
  - ▶ comprehensive Java library
  - ▶ <http://www.semantic-measures-library.org/>
- ▶ R packages: GOSim, GOSemSim, HPOSim, LSAfun, ontologySimilarity,...
- ▶ Python: sematch, fastsemsim (GO only)

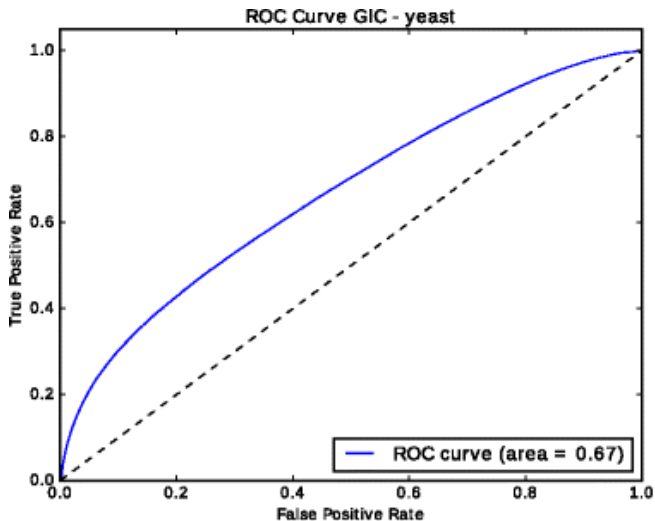
# Applications of semantic similarity

## Hypothesis

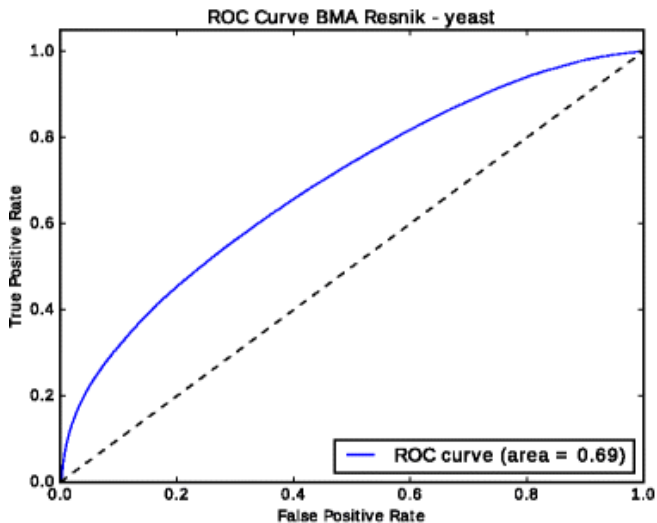
Proteins with similar functions are more likely to interact.

- ▶ relies on background knowledge about functions (encoded in GO)
- ▶ “similarity” can mean:
  - ▶ part of the same pathway
  - ▶ siblings of a common super-class
  - ▶ located in the same location
- ▶ set-based comparison of GO functions
  - ▶ single GO hierarchy or all?
  - ▶ which similarity measure?

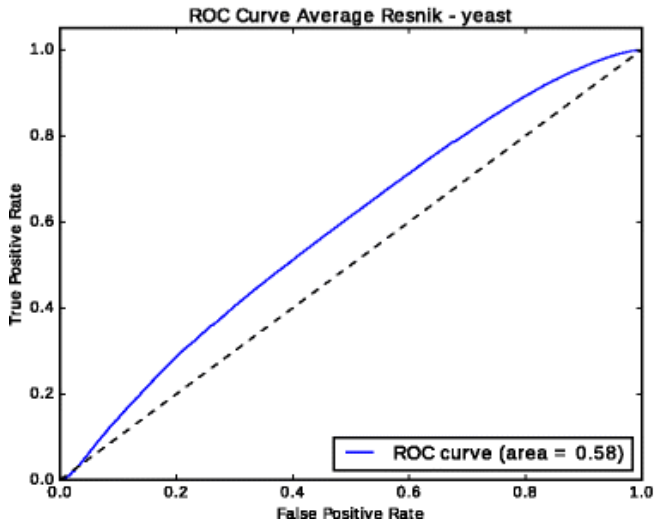
# Applications of semantic similarity



# Applications of semantic similarity



# Applications of semantic similarity



# Applications of semantic similarity

- ▶ no obvious choice of similarity measure
- ▶ depends on application
  - ▶ predicting PPIs in different organisms may benefit from a different similarity measure!
- ▶ different similarity measures may react differently to biases in data
- ▶ needs some testing and experience

# Applications of semantic similarity

Recommendations for using semantic similarity::

- ▶ use Resnik's information content measure
- ▶ use Resnik's similarity
- ▶ use Best Match Average
- ▶ use the full ontology
- ▶ classify your ontology using a reasoner before applying semantic similarity
  - ▶ although many ontologies come pre-classified
- ▶  $\Rightarrow$  but there are many exceptions
  - ▶ similar location  $\Rightarrow$  use location subset of GO
  - ▶ developmental phenotypes  $\Rightarrow$  use developmental branch of phenotype ontology



# Applications of semantic similarity

- ▶ choice of ontology determines the kind of similarity
- ▶ functional similarity: Gene Ontology
- ▶ anatomical, structural similarity: anatomy ontologies (Uberon, MA, FMA, etc.)
- ▶ phenotypic similarity: phenotype ontology (HPO, MP, etc.)
- ▶ chemical structural similarity: ChEBI

# Applications of semantic similarity

- ▶ phenotypic similarity used to:
  - ▶ diagnosis: similarity between patient phenotypes and disease phenotypes
    - ▶ also between patient phenotypes, e.g., Phenomizer:  
<http://compbio.charite.de/phenomizer/>
  - ▶ disease modules: similarity between disease and disease
  - ▶ clustering/stratification: similarity between patient and patient
  - ▶ disease gene discovery: similarity between patient/disease phenotypes and gene–phenotype associations
    - ▶ in humans
    - ▶ in model organisms
  - ▶ drug repurposing: side-effect similarity; similarity between side effect profile and gene–disease associations

# Applications of semantic similarity

- ▶ comparing entities annotated with *different* ontologies/vocabularies of the *same* (or related) domains
  - ▶ medical: UMLS, HPO, DO, ORDO, NCIT, ICD, SNOMED CT, MeSH, ...
  - ▶ phenotype: HPO, MP, CPO, WBPhenotype, FBCV, MeSH, ...
  - ▶ chemical: ChEBI, MeSH, DrOn, RXNorm, DrugBank, ...

# Applications of semantic similarity

- ▶ needs mapping, alignment, or integration
  - ▶ mapping: given a term  $t$ , find corresponding class in ontology  $O$ 
    - ▶ can be 1:1, 1:n, n:1, n:m
    - ▶  $t$  can be from ontology, vocabulary, database, or text
    - ▶ use  $O$  for analysis
  - ▶ alignment: given two ontologies or vocabularies  $O_1$  and  $O_2$ , find all mappings between classes/terms in  $O_1$  and  $O_2$ 
    - ▶ applicable to ontologies and vocabularies
    - ▶ use  $O_1$  or  $O_2$  for analysis
  - ▶ integration: given two ontologies  $O_1$  and  $O_2$ , combine both ontologies into a single ontology  $O$ 
    - ▶ maintain meaning of classes
    - ▶ use  $O$  for analysis

# Applications of semantic similarity

- ▶ lexical mappings: use class labels (and synonyms) to find matches
  - ▶ hypertension (HP:0000822) and hypertension (MP:0000231)
- ▶ semantic mappings: use class axioms to find matches
  - ▶ pulmonary valve stenosis (MP:0006182) and Pulmonic stenosis (HP:0001642)
  - ▶ both definitions based on constricted (PATO:0001847) and pulmonary valve (UBERON:0002146)
- ▶ hybrid: combine lexical and semantic mappings

# Applications of semantic similarity

tools for ontology mapping, matching, integration:

- ▶ AgreementMaker Light:  
<https://github.com/AgreementMakerLight/AML-Jar>
  - ▶ structural (semantic) and lexical matches
  - ▶ can use domain-specific background knowledge
- ▶ LogMap: <https://github.com/ernestojimenezruiz/logmap-matcher>
  - ▶ structural (semantic) and lexical matches
  - ▶ biology-themed versions
- ▶ NCBO Annotator:  
<https://bioportal.bioontology.org/annotator>
  - ▶ lexical matches only
  - ▶ can annotate full text
- ▶ recent tools and comprehensive ongoing evaluation:
  - ▶ OAEI: <http://oei.ontologymatching.org/>

# Hands-on part: diagnosing rare disease using mouse phenotypes

- ▶ run the “Semantic Similarity” notebook
  - ▶ then find the mouse genotype with the most similar set of phenotypes to “Tetralogy of Fallot” (OMIM:187500)
  - ▶ or: use the data from <https://hpo.jax.org/app/download/annotation> to add more diseases and query by disease (hint: a disease is really just a set of phenotypes)