Relational data and knowledge

Semantic similarity

- We want to use *background knowledge* to
  - determine similarity between classes,
  - instances,
  - and entities associated with classes

How to measure similarity?

- semantic similarity measures similarity between classes
- semantic similarity measures similarity between instances of classes
- semantic similarity measures similarity between entities *associated* 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 $sim : D \times D$ is a similarity on D if, for all $x, y \in D$, the function 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:

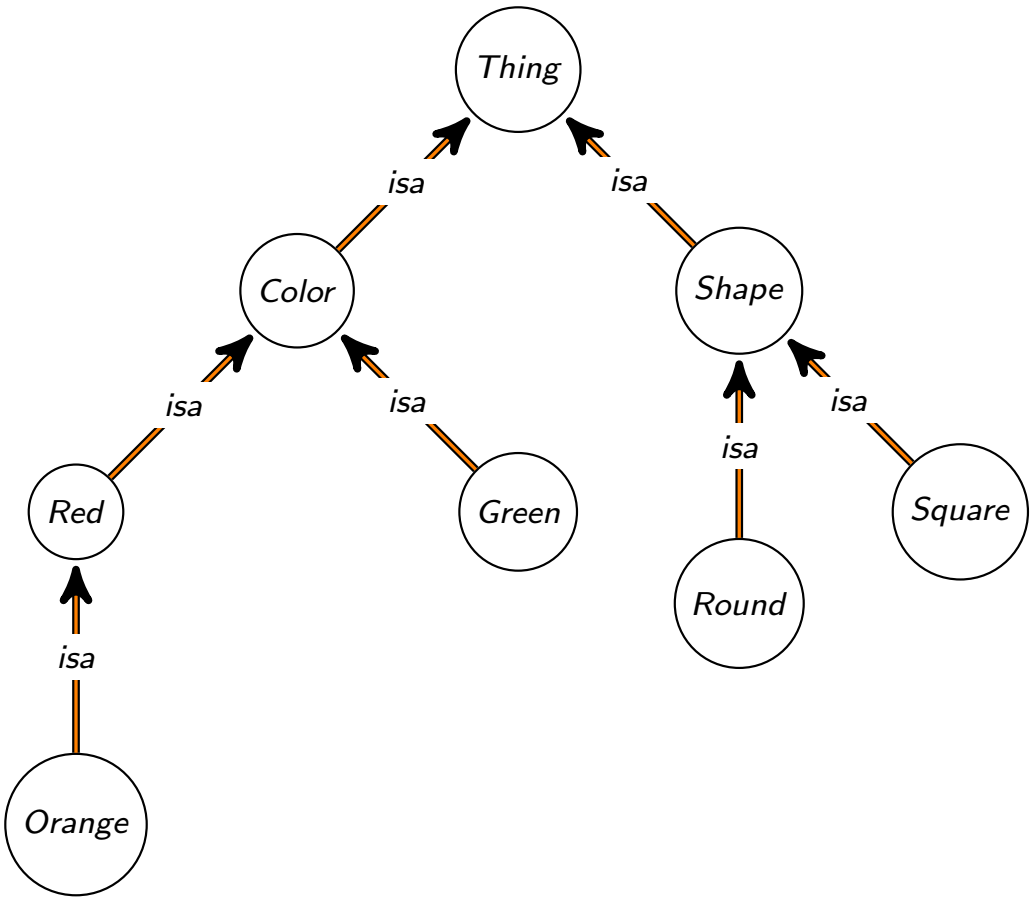
- non-negative: $sim(x, y) \geq 0$ for all x, y
- symmetric: $sim(x, y) = sim(y, x)$
- reflexive: $sim(x, x) = max_D$
  - weaker form: $sim(x, x) > sim(x, y)$ for all $x \neq y$
- $sim(x, x) > sim(x, y)$ for $x \neq y$

How to measure similarity?

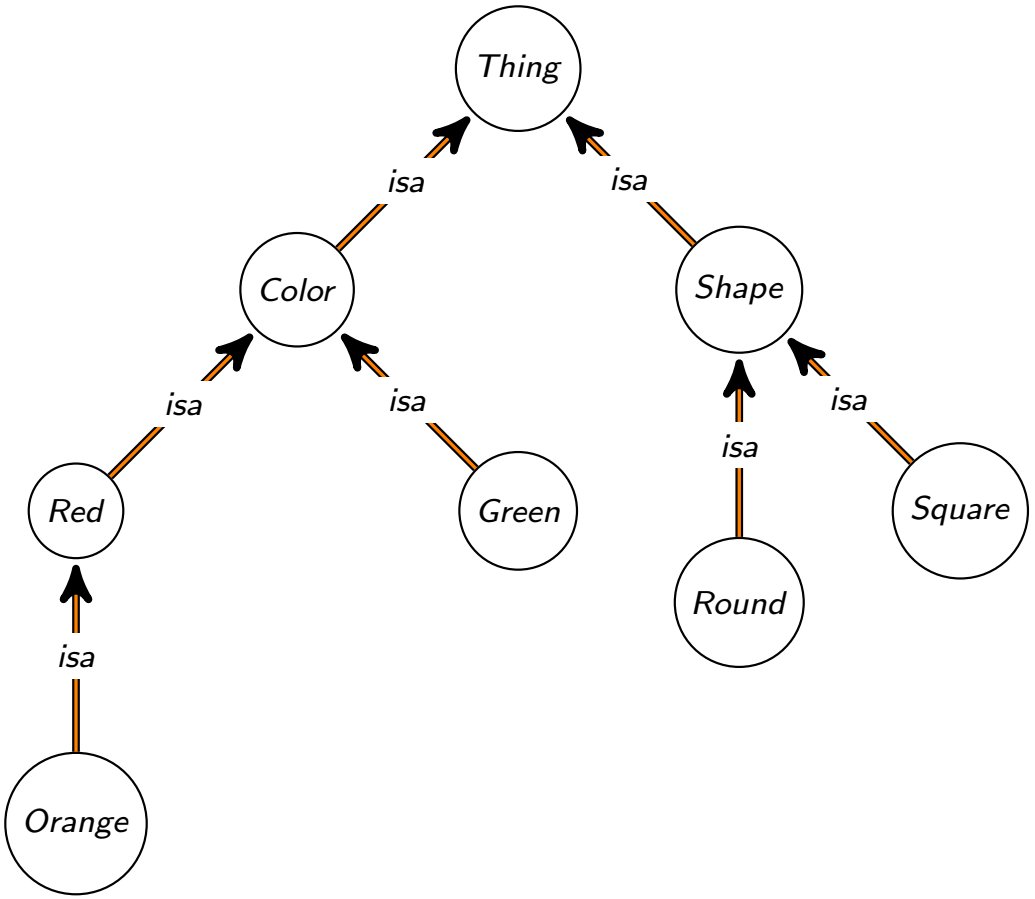
What properties do we want in a similarity measure?
A function $sim : D \times D$ is a similarity on D if, for all $x, y \in D$, the function sim is:

- 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?

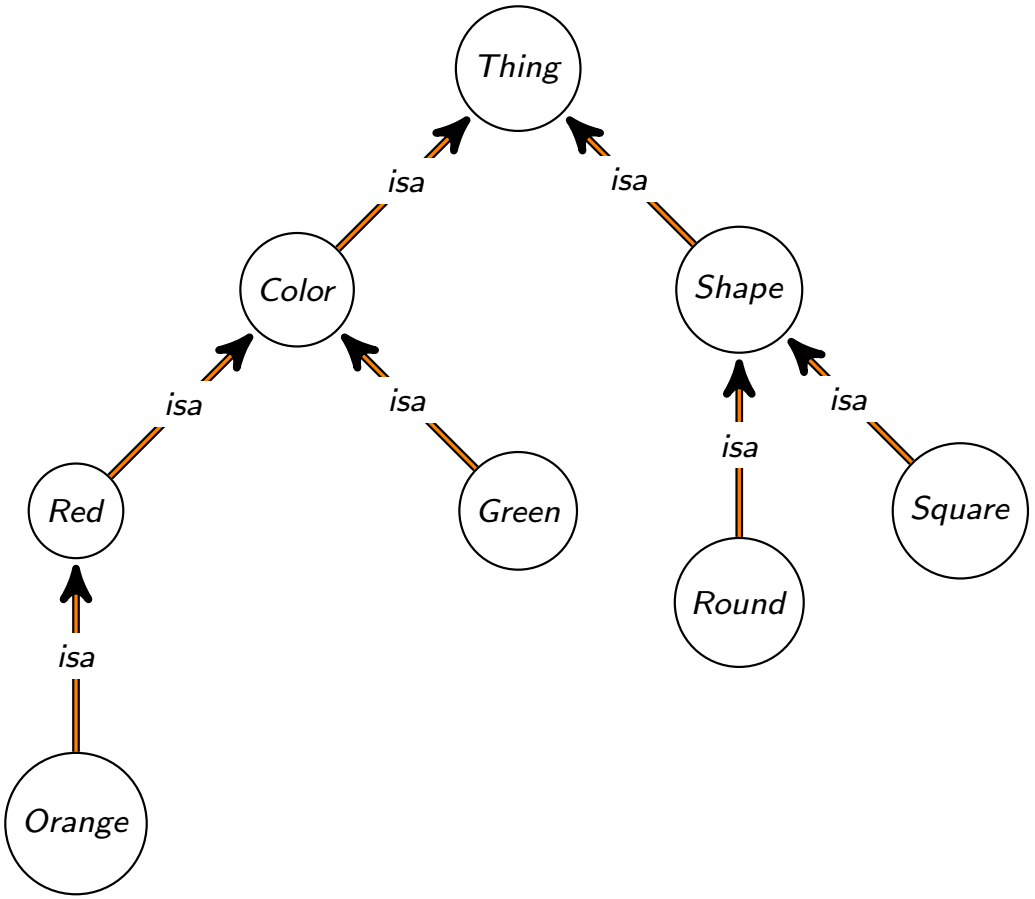


How to measure similarity?



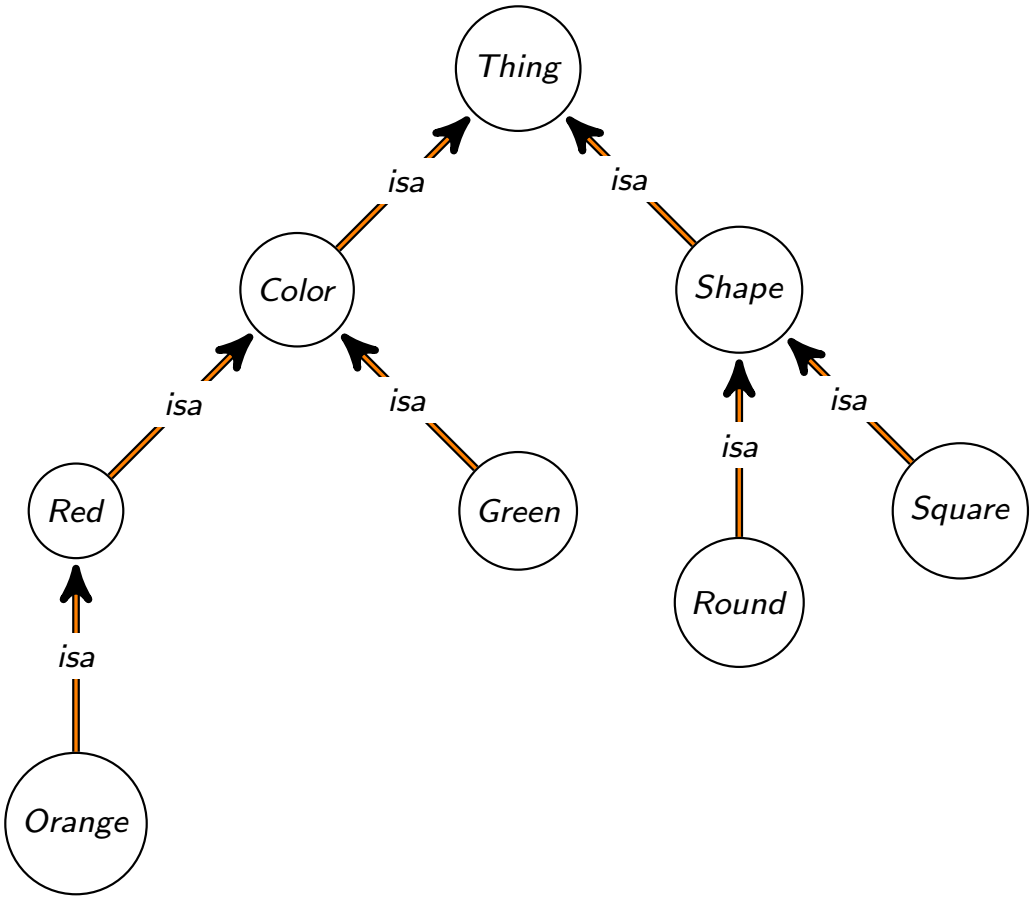
- distance on shortest path (Rada *et al.*, 1989)

How to measure similarity?



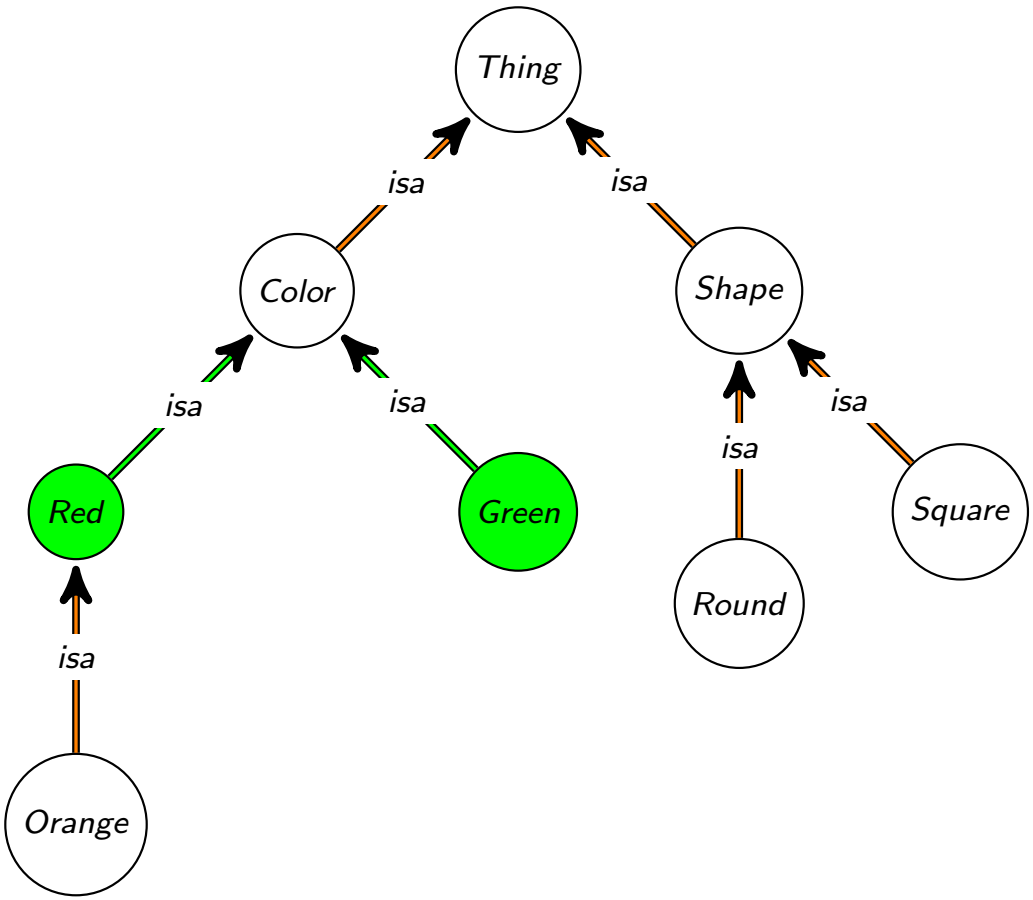
- distance on shortest path (Rada *et al.*, 1989)
- $dist_{Rada}(u, v) = sp(u, isa, v)$

How to measure similarity?



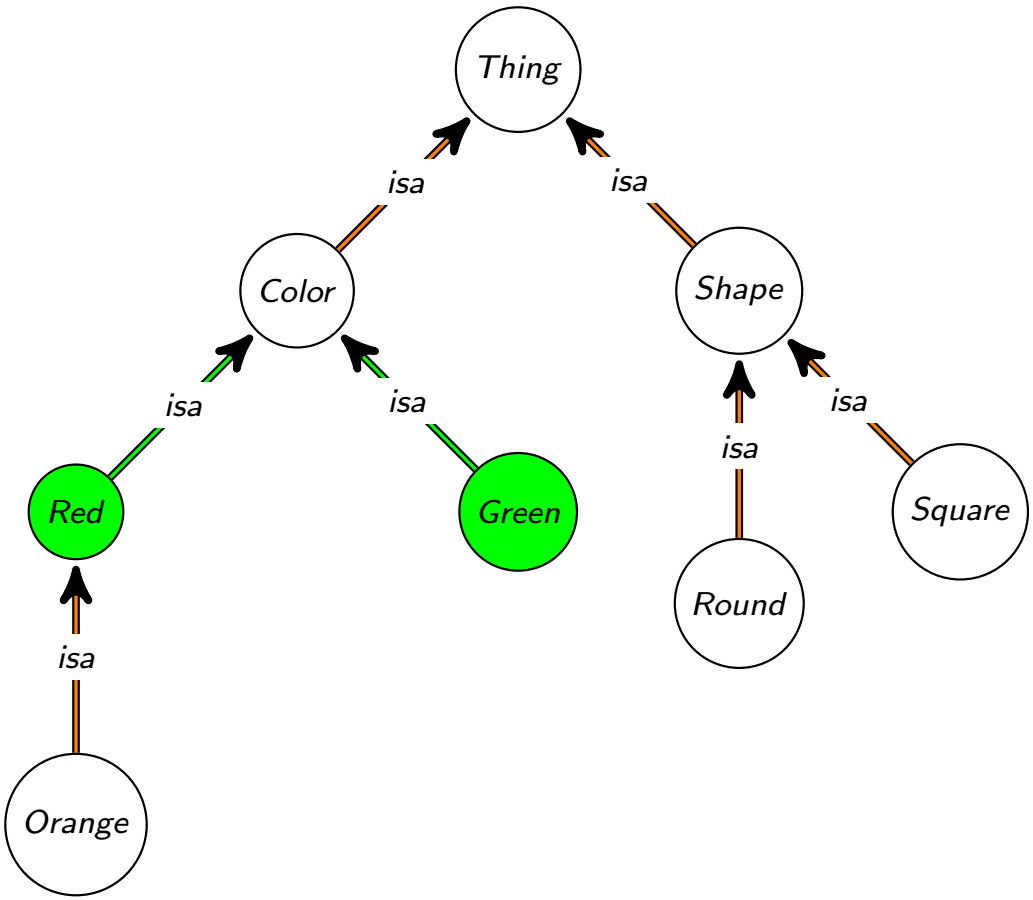
- distance on shortest path (Rada *et al.*, 1989)
- $dist_{Rada}(u, v) = sp(u, isa, v)$
- $sim_{Rada}(u, v) = \frac{1}{dist_{Rada}(u,v)+1}$

How to measure similarity?



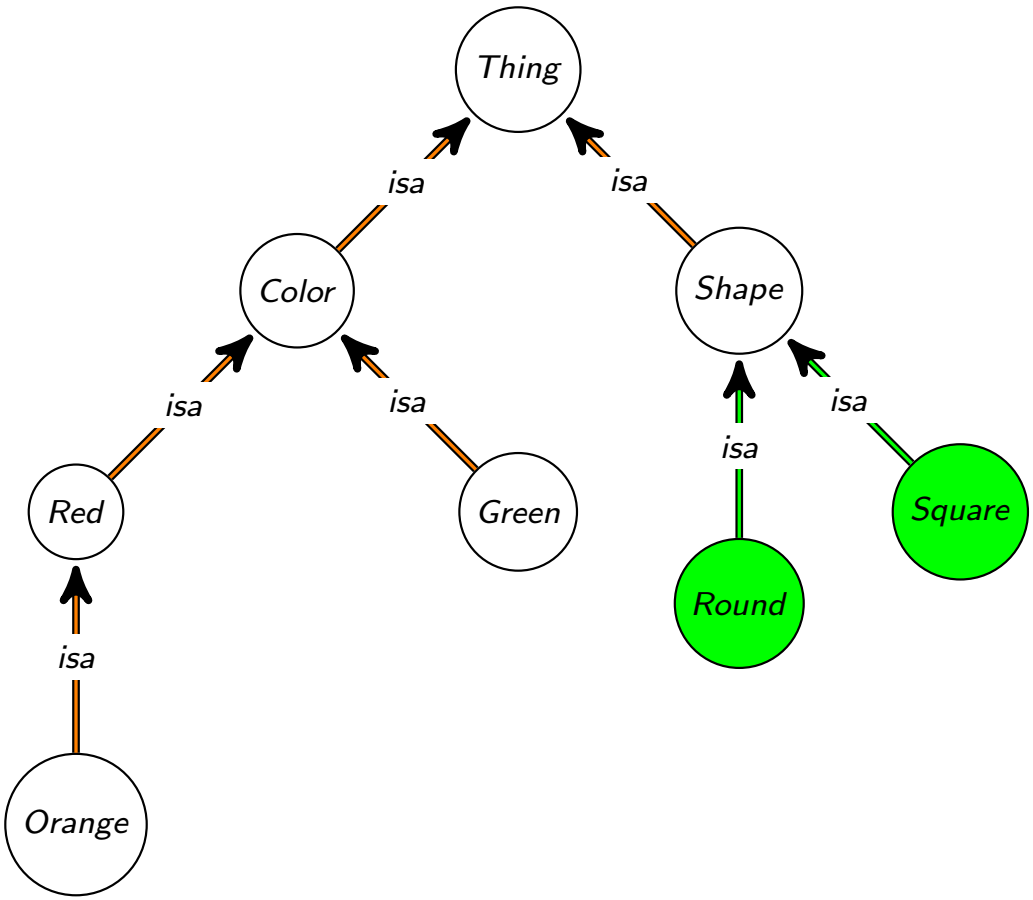
- distance on shortest path

How to measure similarity?



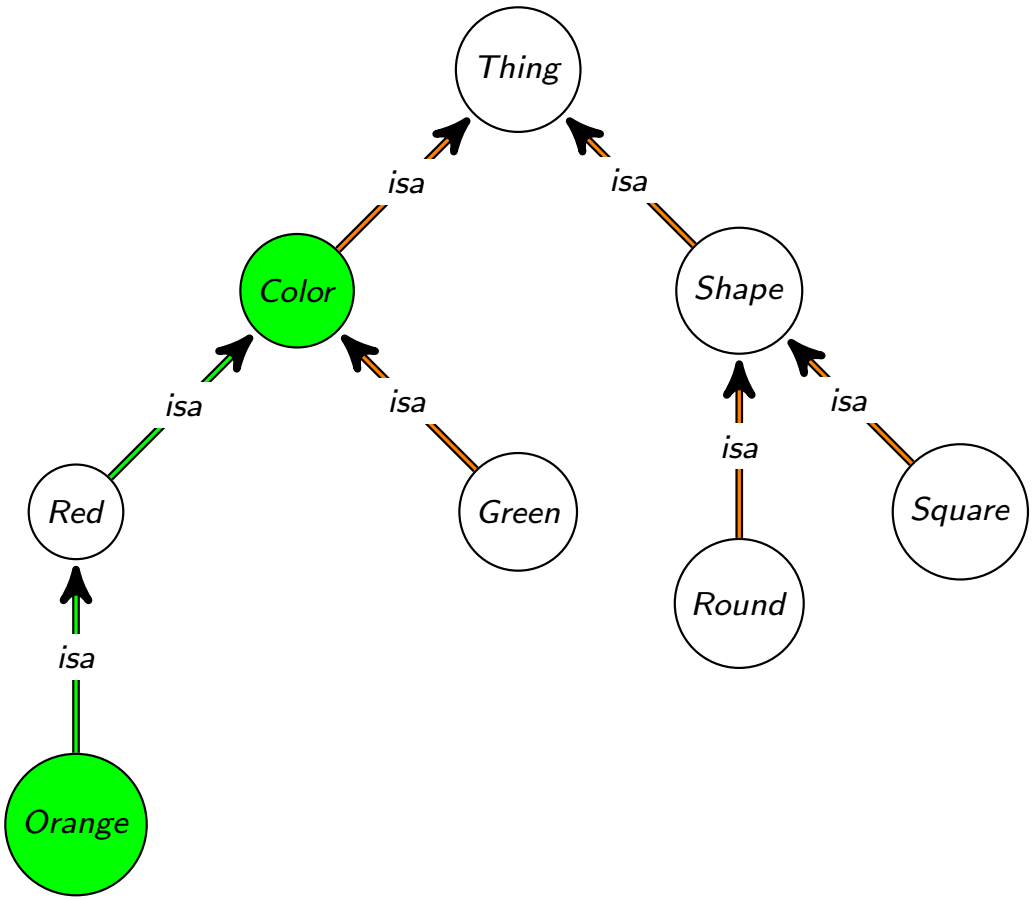
- distance on shortest path
- distance(green, red) $= 2$
- $sim_{Rada}(green, red) = \frac{1}{3}$

How to measure similarity?



- distance on shortest path
- distance(square, round) $= 2$
- $sim_{Rada}(square, round) = \frac{1}{3}$

How to measure similarity?



- distance on shortest path
- distance(orange, color) $= 2$
- $sim_{Rada}(orange, color) = \frac{1}{3}$

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 $\sigma : C \mapsto \mathbb{R}$:
  - $x \sqsubseteq y \rightarrow \sigma(x) \geq \sigma(y)$

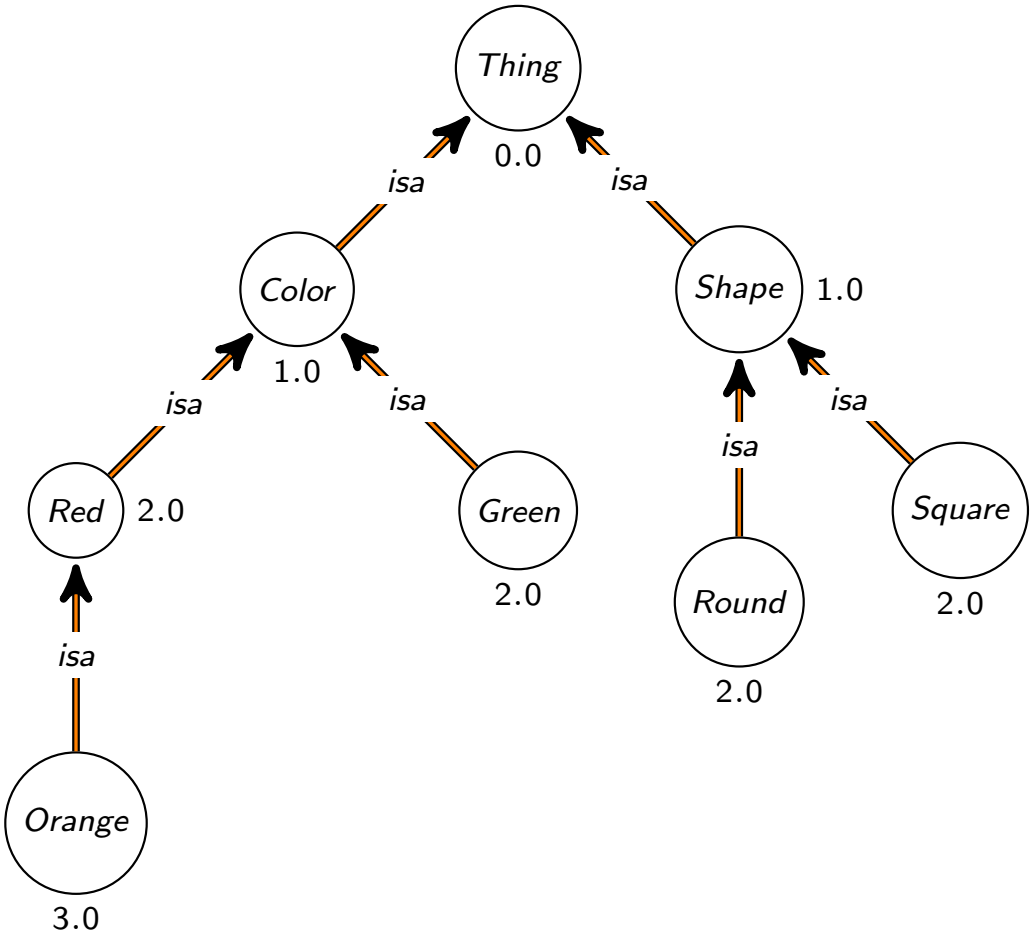
How to measure similarity

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

How to measure similarity

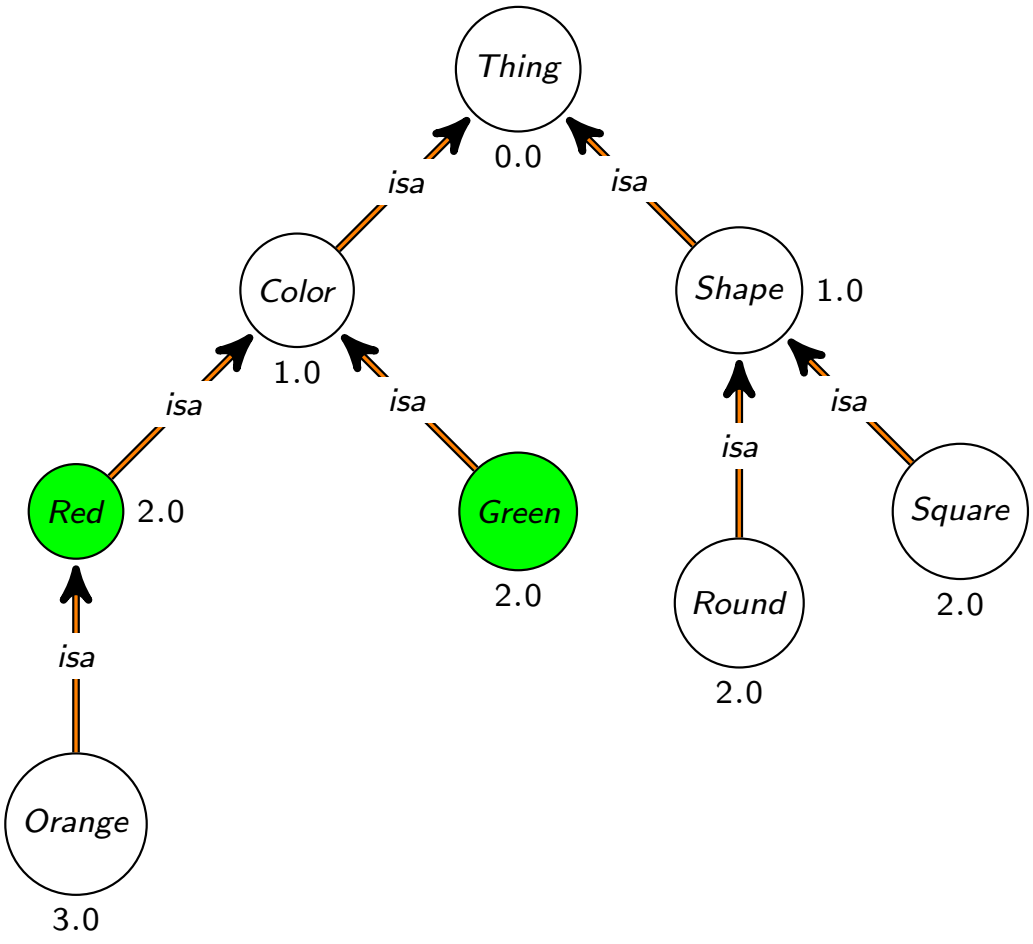
- term specificity measure $\sigma : C \mapsto \mathbb{R}$:
  - $x \sqsubseteq y \rightarrow \sigma(x) \geq \sigma(y)$
- intrinsic:
  - $\sigma(x) = f(depth(x))$
  - $\sigma(x) = f(A(x))$ (for ancestors $A(x)$)
  - $\sigma(x) = f(D(x))$ (for descendants $D(x)$)
  - many more, e.g., Zhou et al.:
    $\sigma(x) = k \cdot \left(1 - \frac{\log |D(x)|}{\log |C|}\right) + (1 - k)\frac{\log depth(x)}{\log depth(G_T)}$
- 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: $eIC_{Resnik}(x) = -\log p(x)$ (with $p(x) = \frac{|I(x)|}{|I|}$)
    - in biology, one of the most popular specificity measure when annotations are present

How to measure similarity?



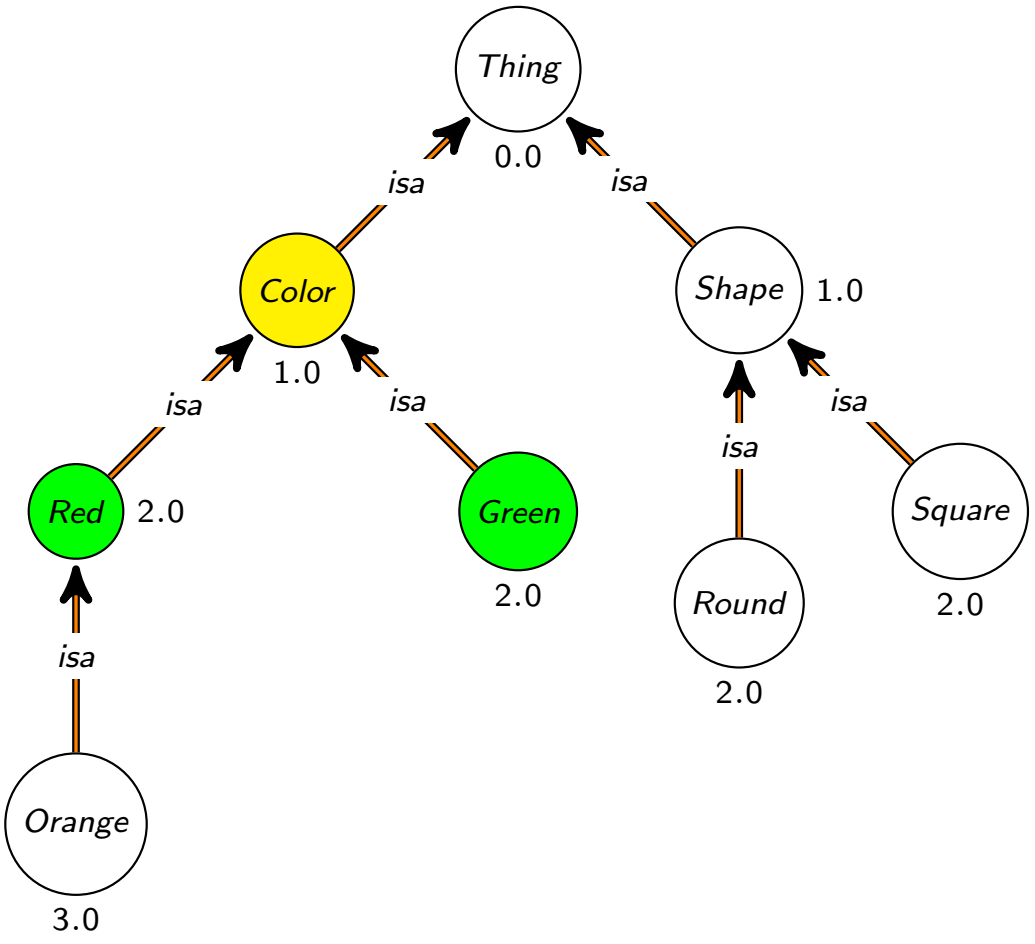
- Resnik 1995: similarity between x and y is the information content of the *most informative common ancestor*

How to measure similarity?



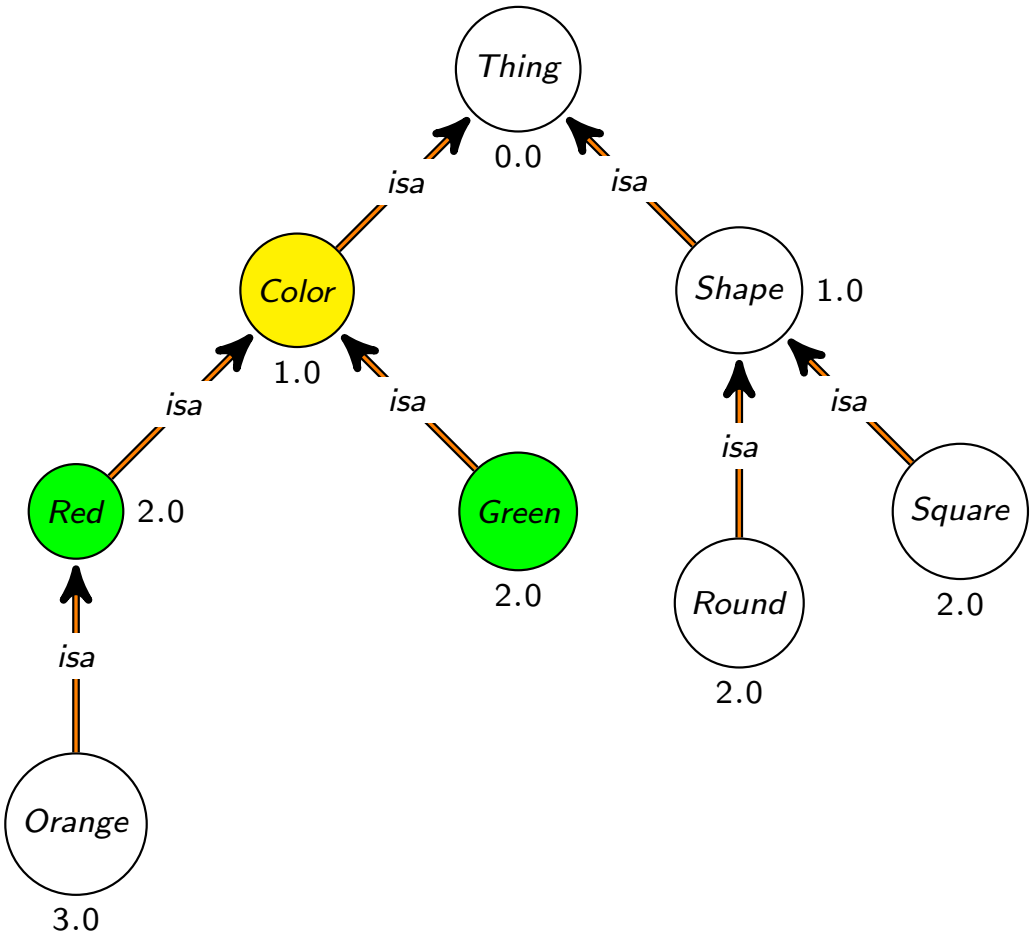
- Resnik 1995: similarity between x and y is the information content of the *most informative common ancestor*

How to measure similarity?



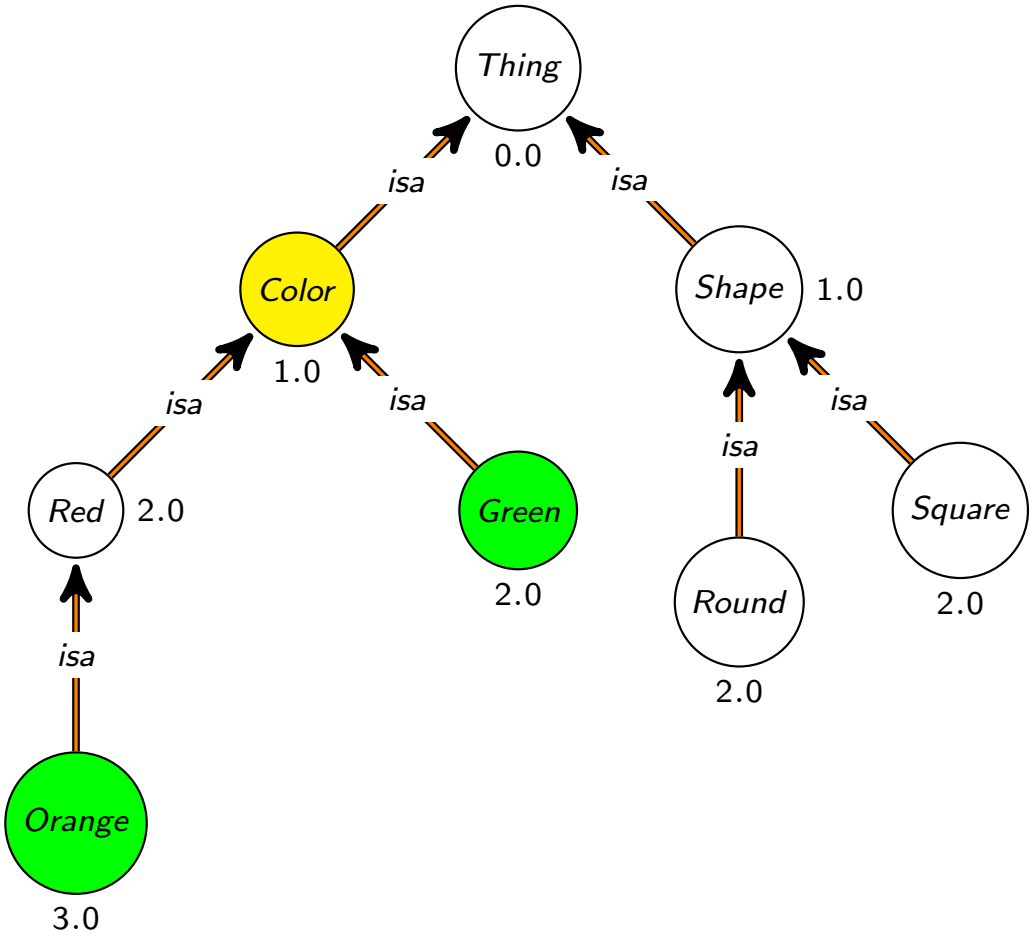
- Resnik 1995: similarity between x and y is the information content of the *most informative common ancestor*

How to measure similarity?



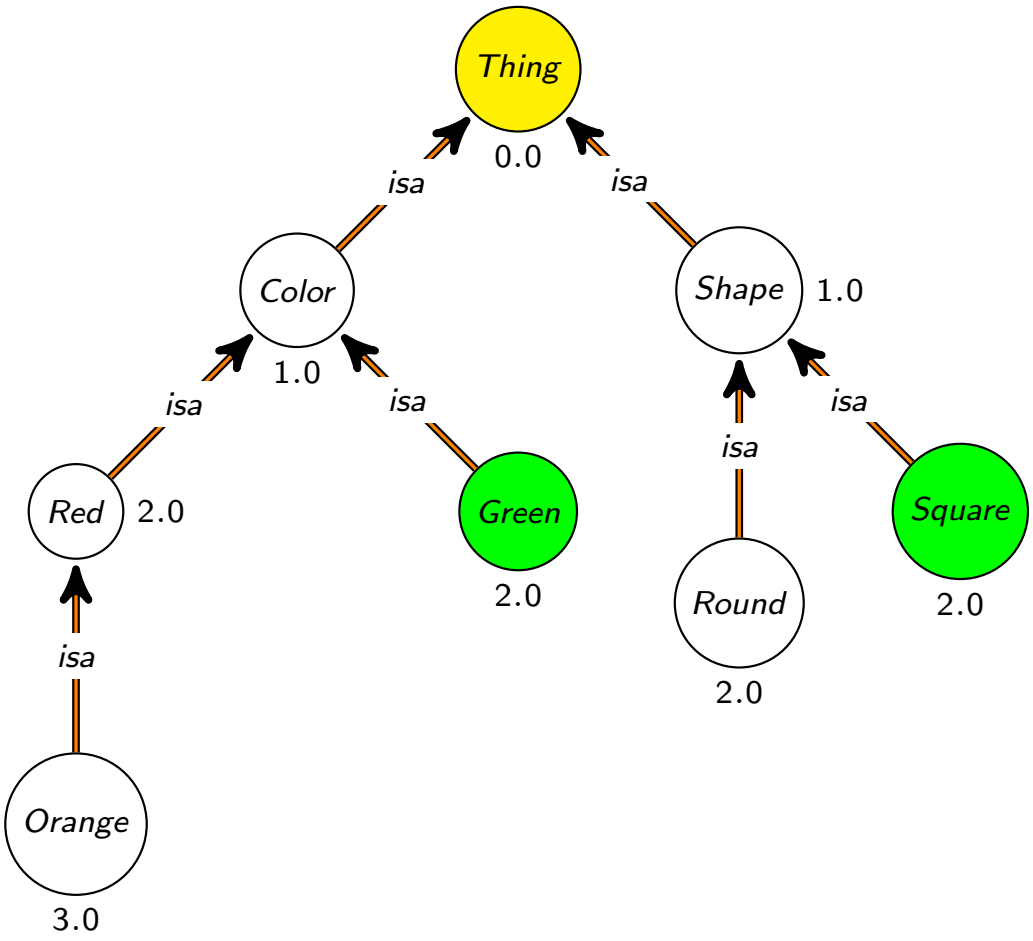
- Resnik 1995: similarity between x and y is the information content of the *most informative common ancestor*
- $sim_{Resnik}(Green, Red) = 1.0$

How to measure similarity?



- Resnik 1995: similarity between x and y is the information content of the *most informative common ancestor*
- $sim_{Resnik}(Green, Orange) = 1.0$

How to measure similarity?

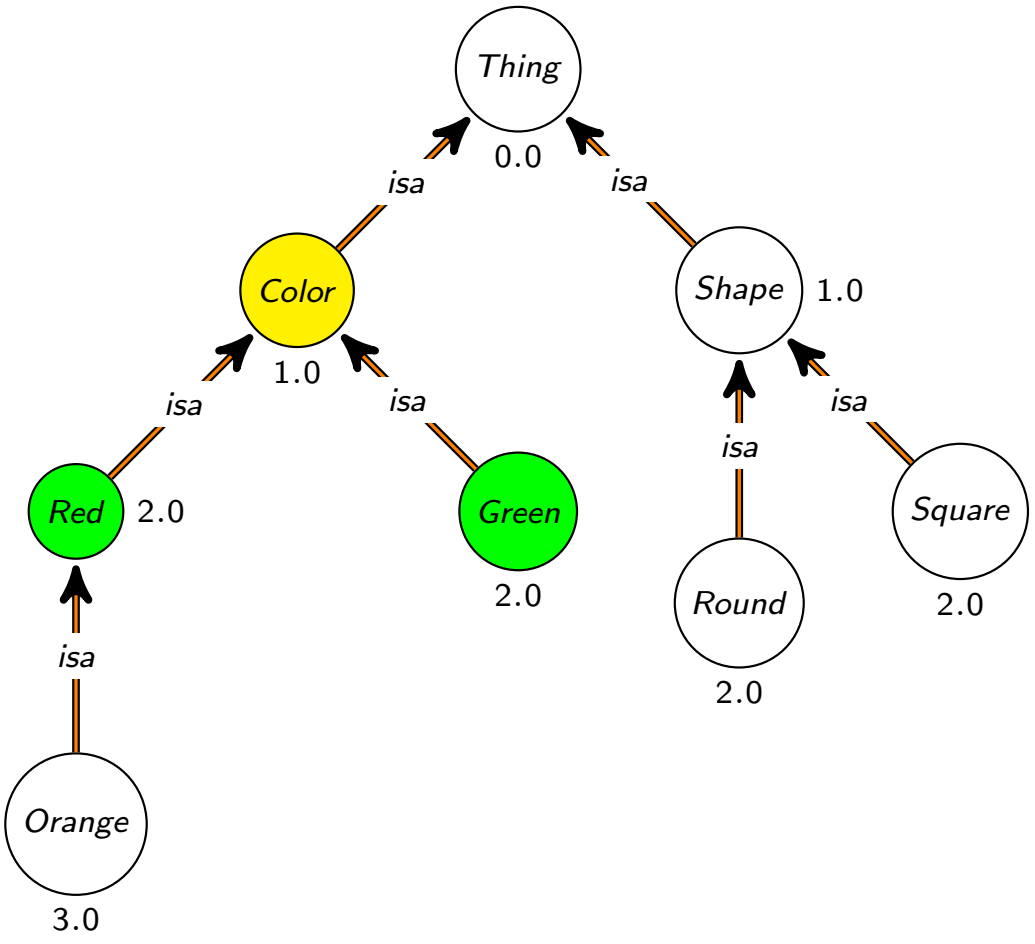


- Resnik 1995: similarity between x and y is the information content of the *most informative common ancestor*
- $sim_{Resnik}(Square, Orange)$ 0.0

How to measure similarity?

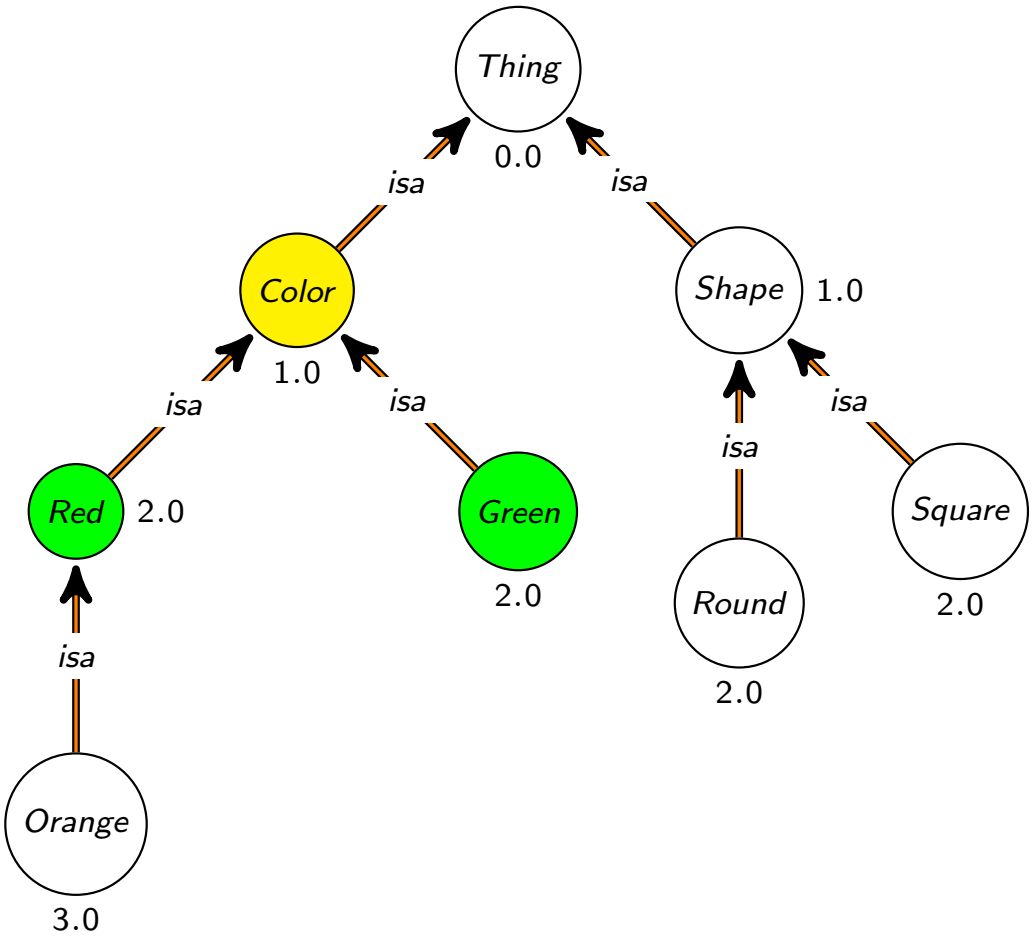
- (Red, Green) and (Orange, Green) have the same similarity
- need to incorporate the specificity of the compared classes

How to measure similarity?



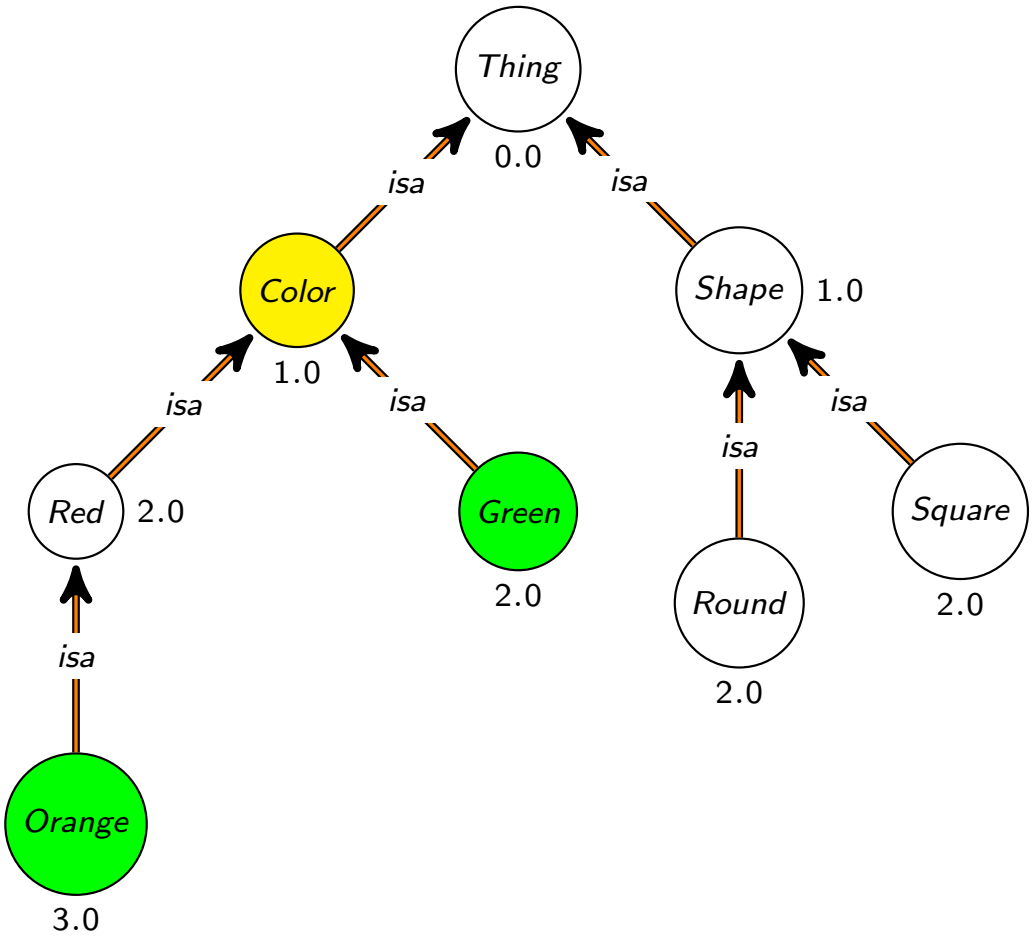
- Lin 1998:
  $sim_{Lin}(x, y) = \frac{2 \cdot IC(MICA(x,y))}{IC(x)+IC(y)}$

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:
  $sim_{Lin}(x, y) = \frac{2 \cdot IC(MICA(x,y))}{IC(x)+IC(y)}$
- $sim_{Lin}(Green, Orange) = 0.4$

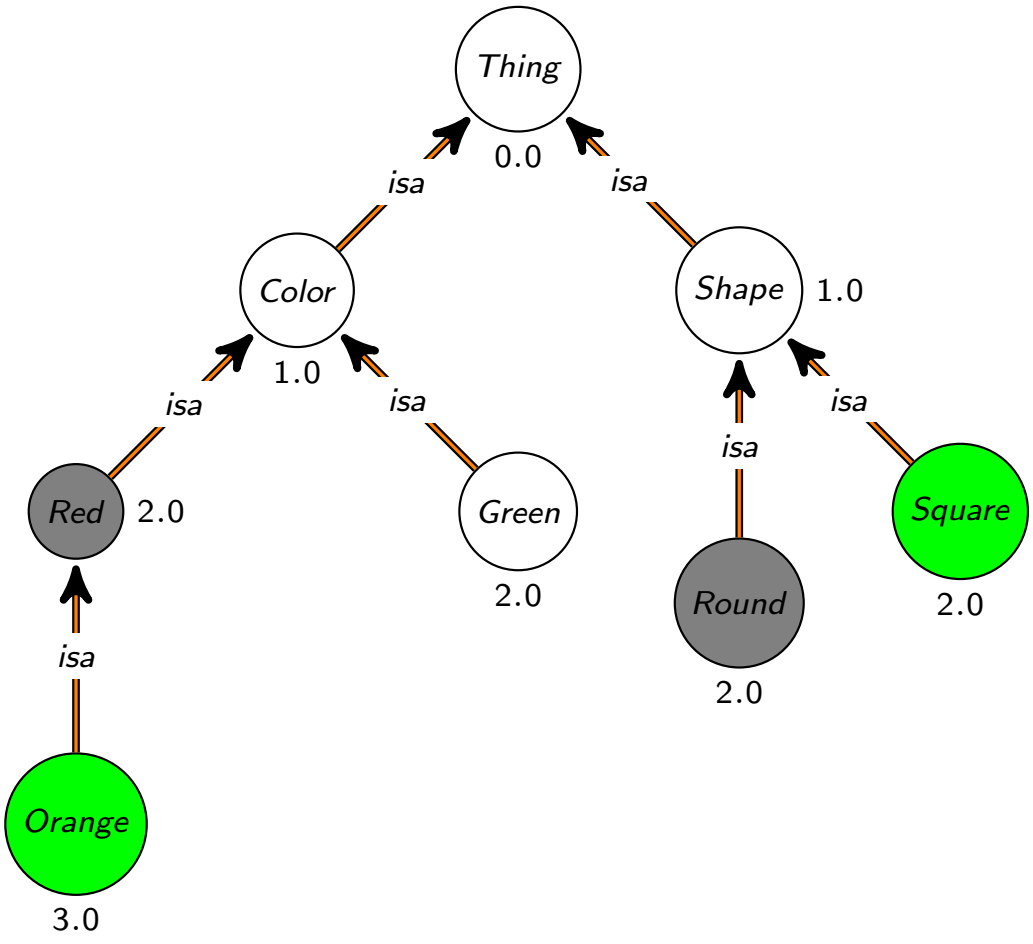
How to measure similarity?

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

How to measure similarity?

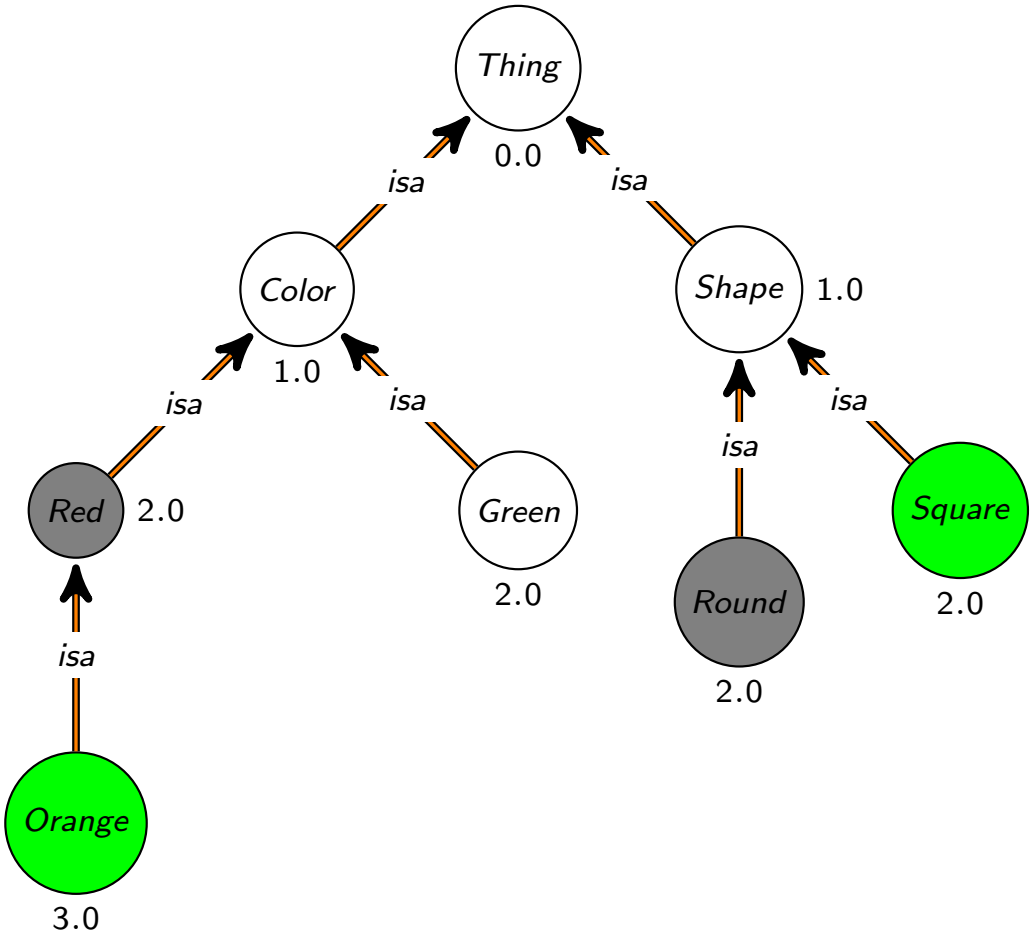
- we only looked at comparing pairs of classes
- mostly, we want to compare *sets* of classes
  - set of GO annotations
  - set of signs and symptoms
  - set of phenotypes
- two approaches:
  - compare each class individually, then merge
  - directly set-based similarity measures

How to measure similarity?



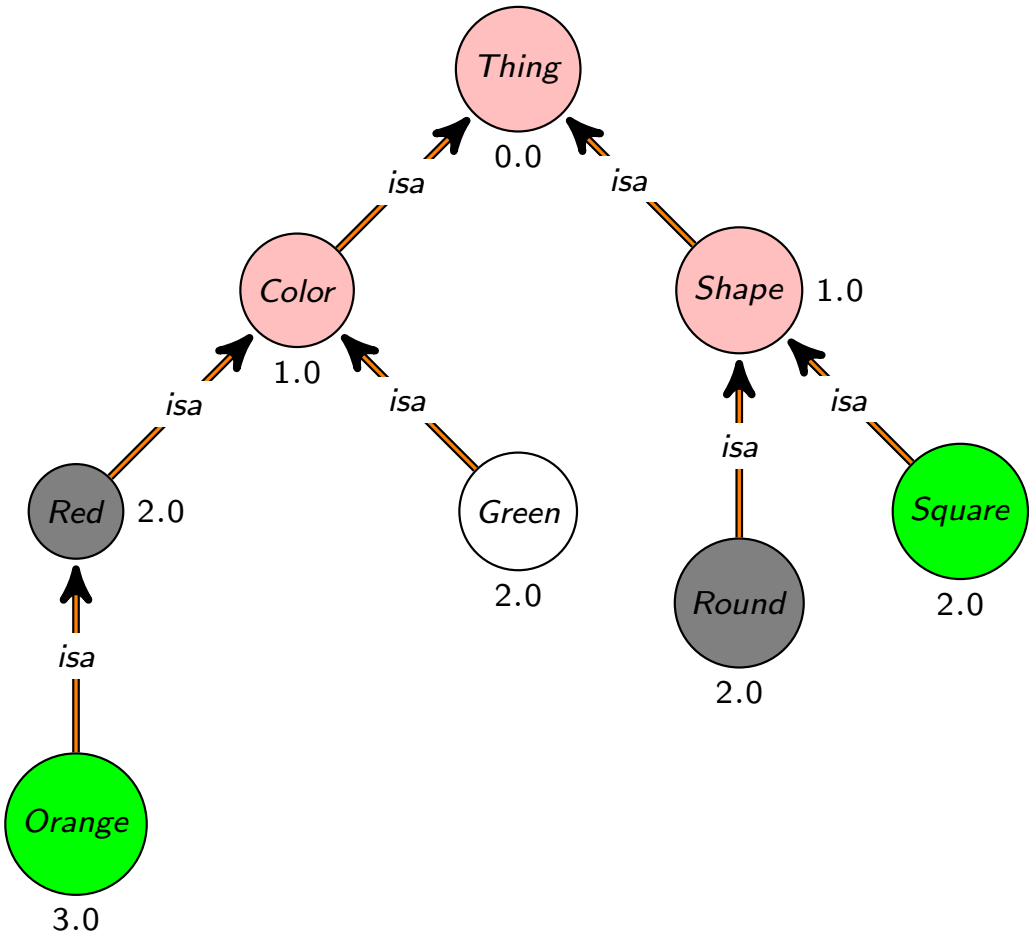
- similarity between a square-and-orange thing and a round-and-red thing

How to measure similarity?



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

How to measure similarity?



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

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

- no obvious choice of similarity measure
- depends on application
  - e.g., predicting PPIs in different organisms through similarity 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:

- use Resnik's information content measure
- use Resnik's similarity
- use Best Match Average
- use all background knowledge
- classify knowledge using an automated reasoner before applying semantic similarity

Quiz: Semantic Similarity

What defines a semantic similarity measure?

1. Number of classes, OWL profile
2. Graph structure, term specificity, pairwise similarity, aggregation operation
3. Graph structure, term specificity, groupwise similarity

Temporary page!

LaTeX was unable to guess the total number of pages correctly. A
there was some unprocessed data that should have been added
the final page this extra page has been added to receive it.
If you rerun the document (without altering it) this surplus page
will go away, because LaTeX now knows how many pages to expe
for this document.