# Ontology Alignment

# Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Ontology alignment challenges

# Ontologies in biomedical research

- many biomedical ontologies

  e.g. GO, OBO, SNOMED-CT

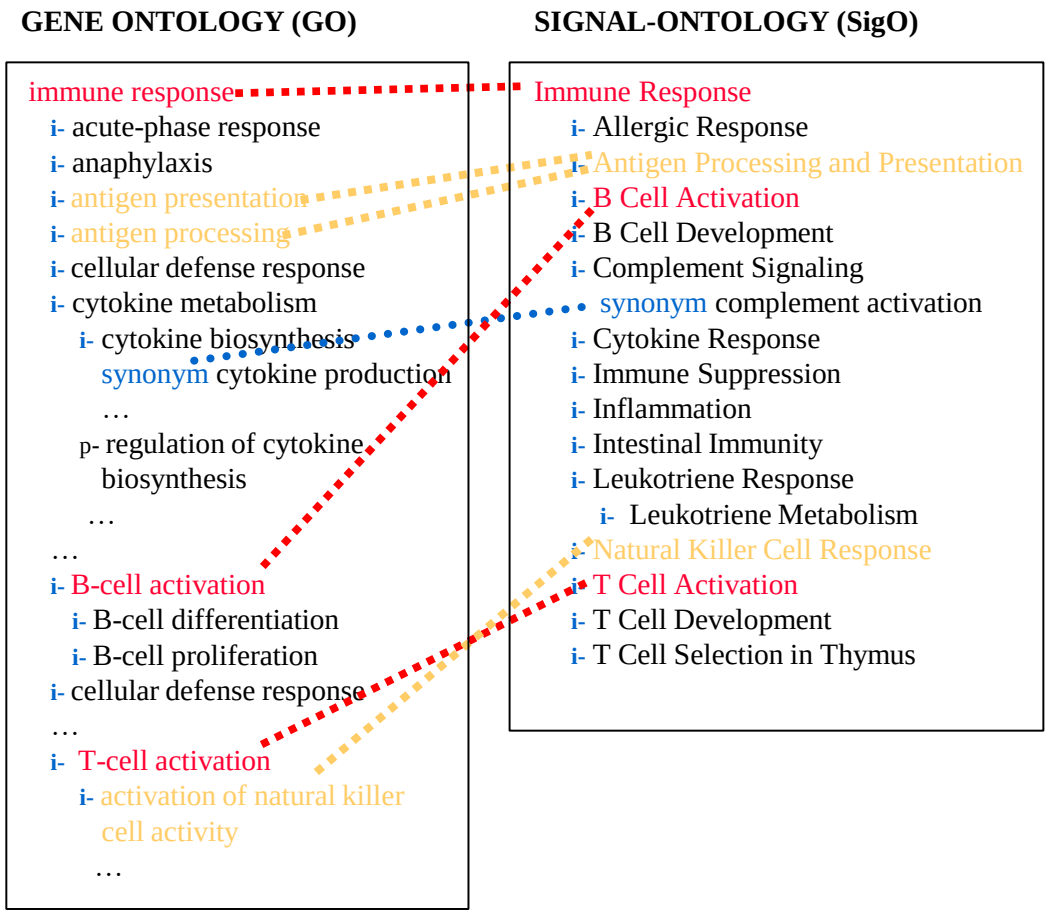
- practical use of biomedical ontologies

  e.g. databases annotated with GO

**GENE ONTOLOGY (GO)**

immune response
- **i-** acute-phase response
- **i-** anaphylaxis
- **i-** antigen presentation
- **i-** antigen processing
- **i-** cellular defense response
- **i-** cytokine metabolism
  - **i-** cytokine biosynthesis
    synonym cytokine production
    …
  - p- regulation of cytokine biosynthesis
  - …
- …
- **i-** B-cell activation
  - **i-** B-cell differentiation
  - **i-** B-cell proliferation
- **i-** cellular defense response
- …
- **i-** T-cell activation
  - **i-** activation of natural killer cell activity
  - …

# Ontologies with overlapping information



# Ontologies with overlapping information

- Use of multiple ontologies
  - custom-specific ontology + standard ontology
  - different views over same domain
  - overlapping domains
- Bottom-up creation of ontologies

  experts can focus on their domain of expertise

→ important to know the inter-ontology relationships

**GENE ONTOLOGY (GO)**

immune response
- i- acute-phase response
- i- anaphylaxis
- i- antigen presentation
- i- antigen processing
- i- cellular defense response
- i- cytokine metabolism
  - i- cytokine biosynthesis
    synonym cytokine production
    …
  - p- regulation of cytokine biosynthesis
  …

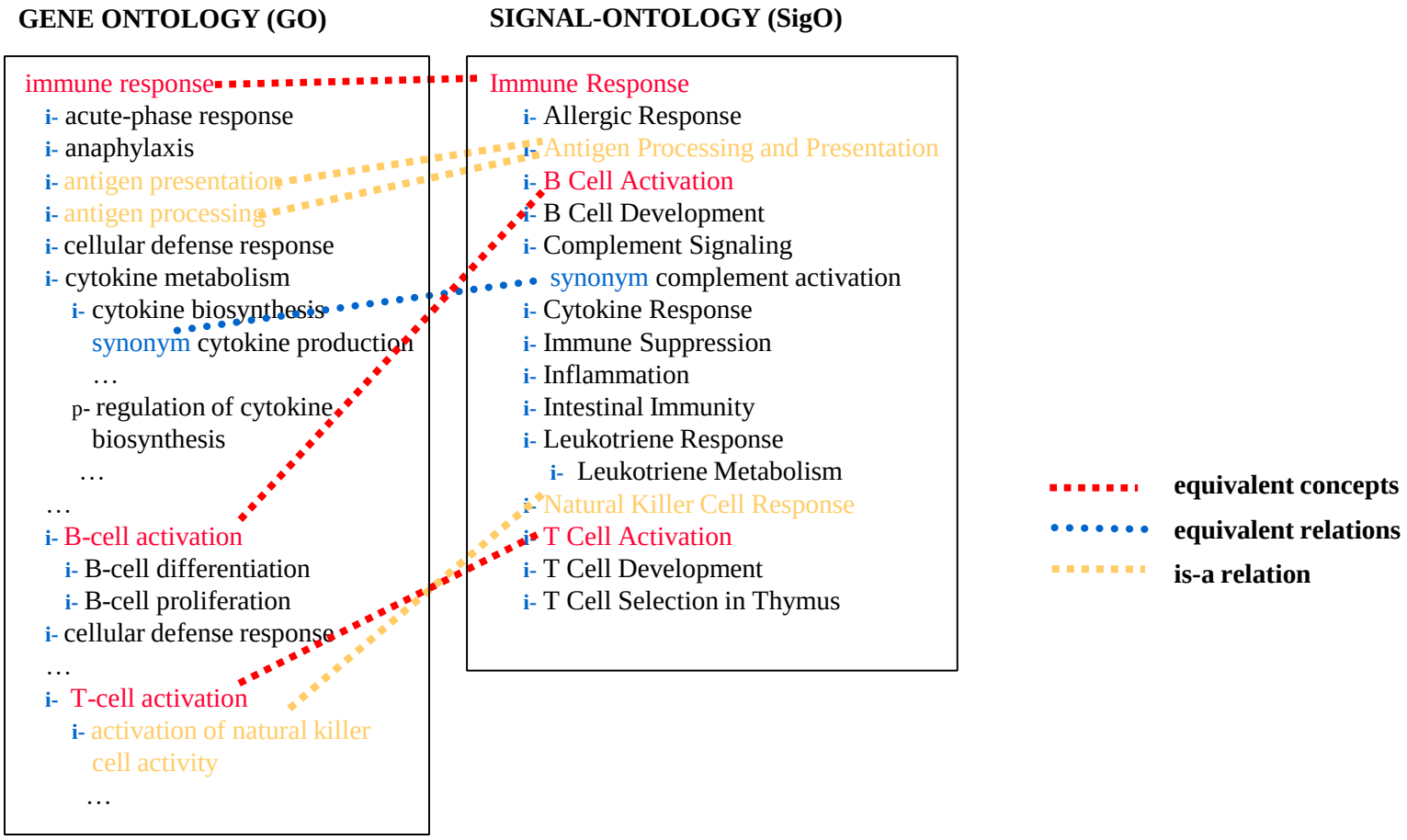
…
- i- B-cell activation
  - i- B-cell differentiation
  - i- B-cell proliferation
- i- cellular defense response

…
- i- T-cell activation
  - i- activation of natural killer cell activity
  …

**SIGNAL-ONTOLOGY (SigO)**

Immune Response
- i- Allergic Response
- i- Antigen Processing and Presentation
- i- B Cell Activation
- i- B Cell Development
- i- Complement Signaling
  synonym complement activation
- i- Cytokine Response
- i- Immune Suppression
- i- Inflammation
- i- Intestinal Immunity
- i- Leukotriene Response
  - i- Leukotriene Metabolism
- i- Natural Killer Cell Response
- i- T Cell Activation
- i- T Cell Development
- i- T Cell Selection in Thymus

# Ontology Alignment

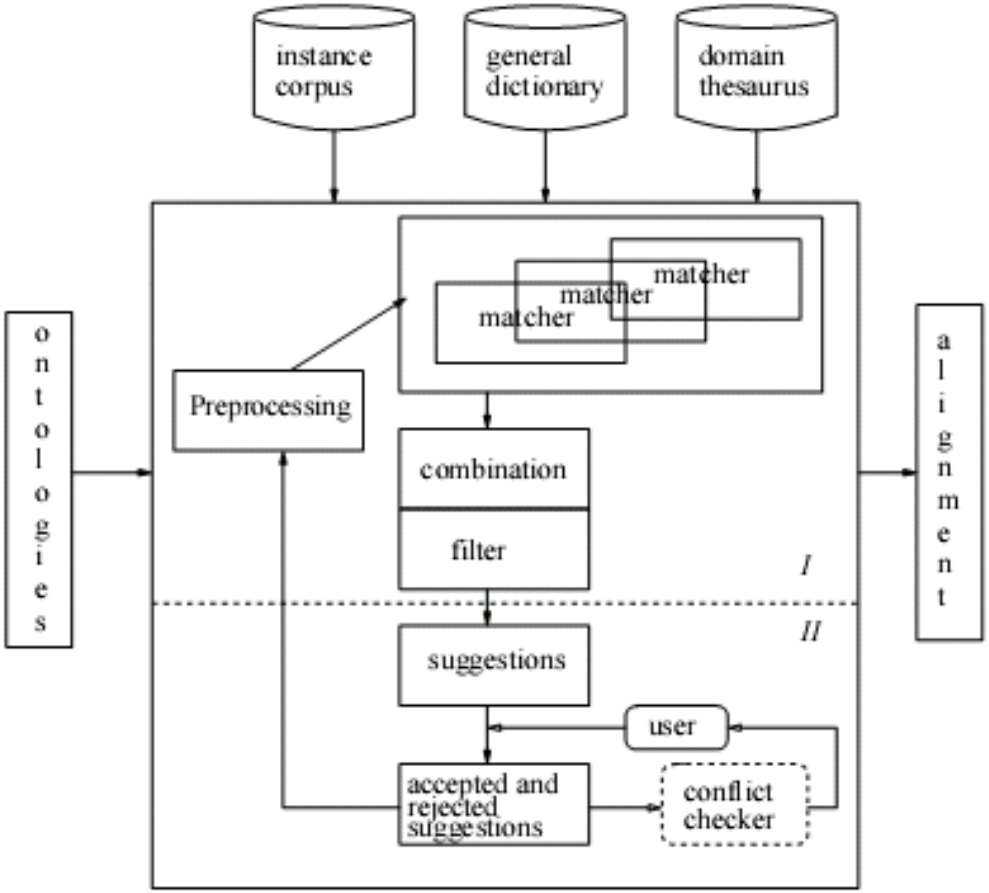



Defining the relations between the terms in different ontologies

# Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Ontology alignment challenges

# An Alignment Framework



# Preprocessing

# Preprocessing

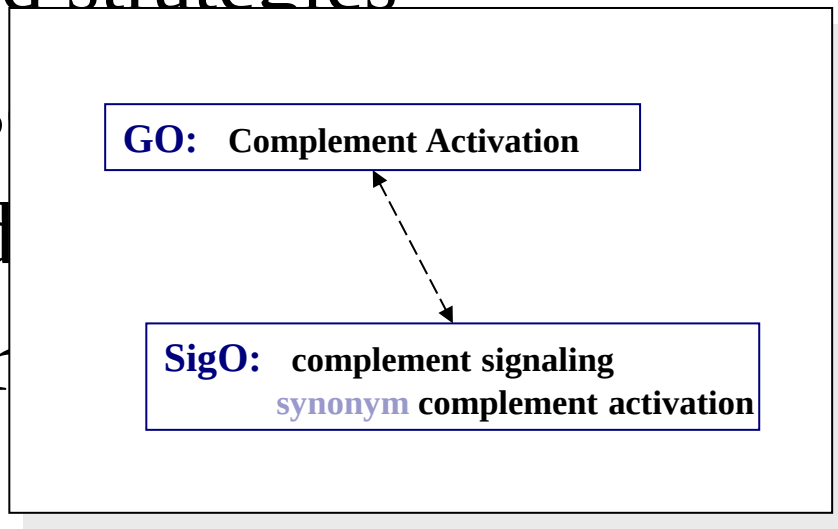
For example,

- Selection of features
- Selection of search space

# Matchers

# Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-bas
- Instance-based
- Use of auxiliar

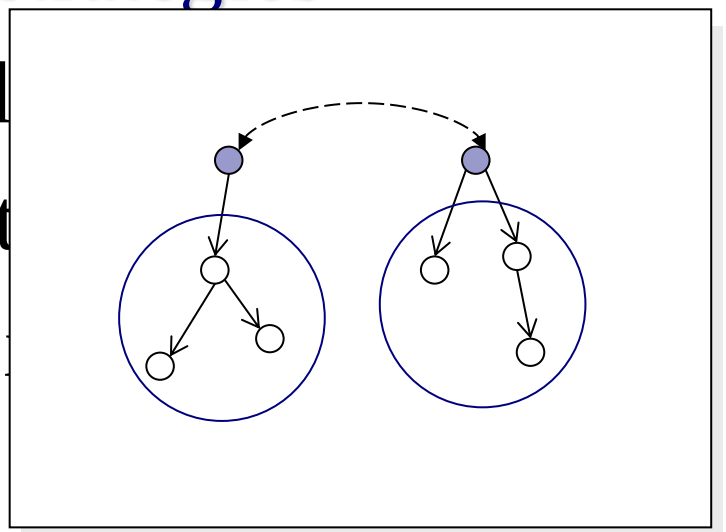


# Example matchers

- Edit distance
  - Number of deletions, insertions, substitutions required to transform one string into another
  - aaaa → baab: edit distance 2
- N-gram
  - N-gram : N consecutive characters in a string
  - Similarity based on set comparison of n-grams
  - aaaa : {aa, aa, aa};  baab : {ba, aa, ab}

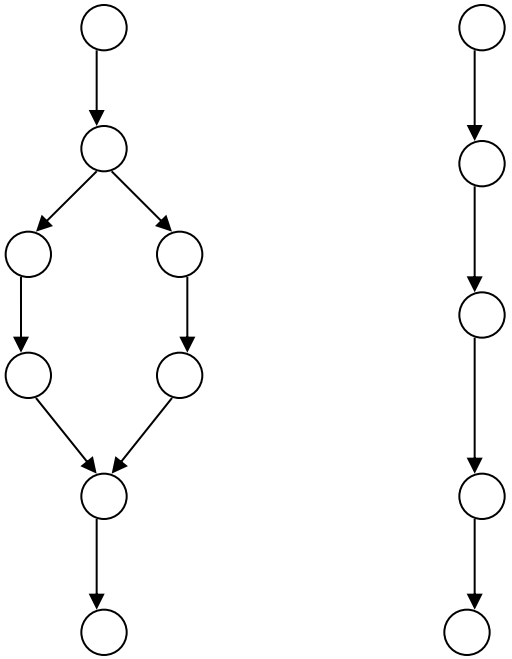
# Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based
- Instance-based st
- Use of auxiliary



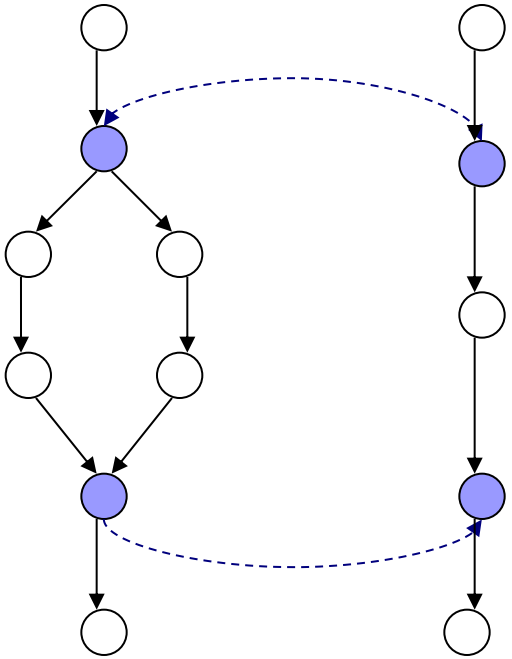
# Example matchers

- Propagation of similarity values
- Anchored matching



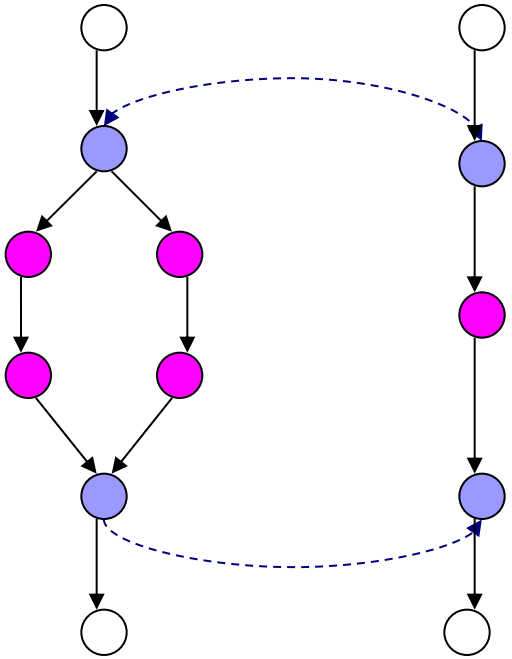
# Example matchers

- Propagation of similarity values
- Anchored matching



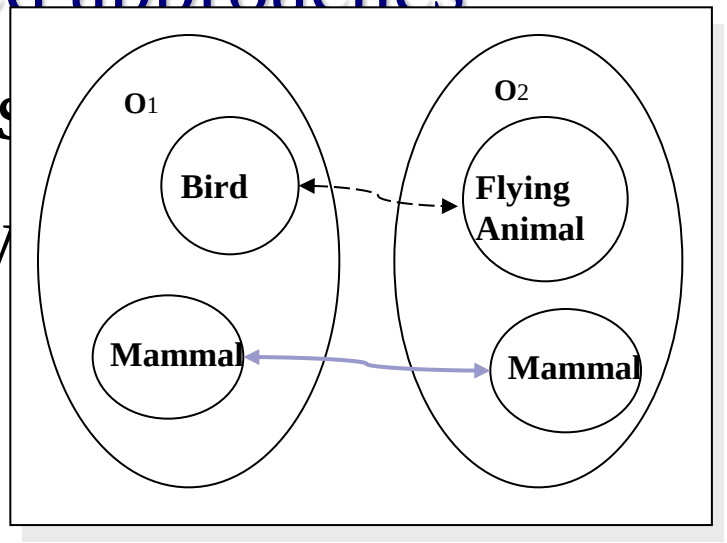
# Example matchers

- Propagation of similarity values
- Anchored matching



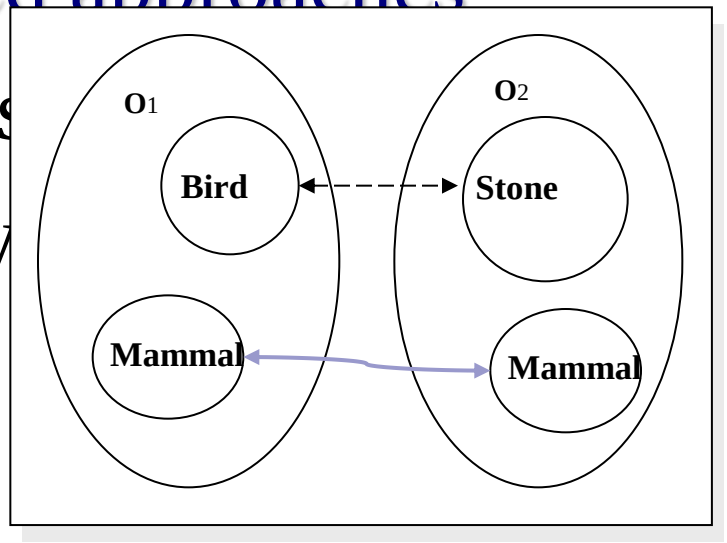
# Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based approaches
- Instance-based s
- Use of auxiliary



# Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based approaches
- Instance-based s
- Use of auxiliary

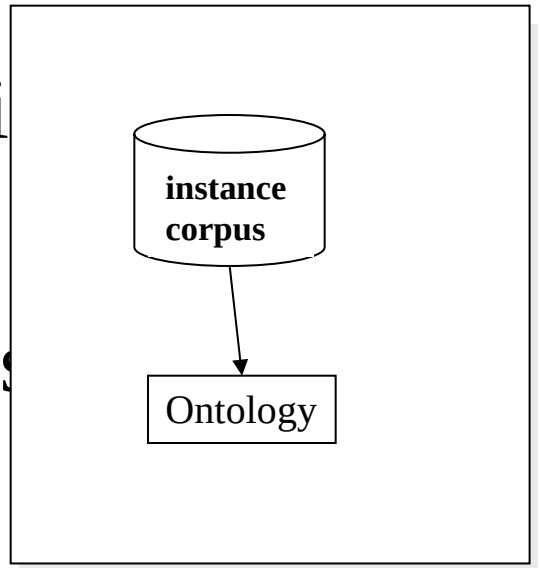


# Example matchers

- Similarities between data types
- Similarities based on cardinalities

# Matcher Strategies

- Strategies based on linguisti
- Structure-based strategies
- Constraint-based approache
- Instance-based strategies
- Use of auxiliary information



# Example matchers

- Instance-based
- Use life science literature as instances

# Learning matchers – instance-based strategies

- Basic intuition

  A similarity measure between concepts can be computed based on the probability that documents about one concept are also about the other concept and vice versa.

# Learning matchers - steps

- Generate corpora
  - Use concept as query term in PubMed
  - Retrieve most recent PubMed abstracts
- Generate text classifiers
  - One classifier per ontology / One classifier per concept
- Classification
  - Abstracts related to one ontology are classified by the other ontology's classifier(s) and vice versa
- Calculate similarities

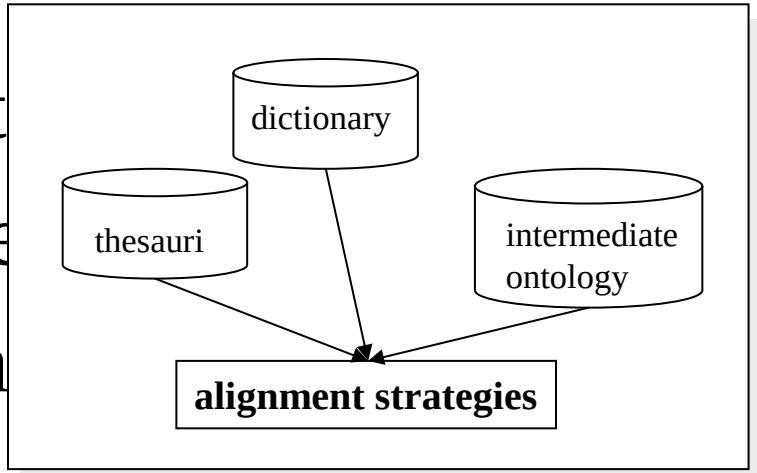
# Basic Naïve Bayes matcher

- Generate corpora
- Generate classifiers
  - Naive Bayes classifiers, one per ontology
- Classification
  - Abstracts related to one ontology are classified to the concept in the other ontology with highest posterior probability P(C|d)
- Calculate similarities

$$sim(C_1, C_2) = \frac{n_{NBC2}(C_1, C_2) + n_{NBC1}(C_2, C_1)}{n_D(C_1) + n_D(C_2)}$$

# Matcher Strategies

- Strategies based linguist
- Structure-based strategie
- Constraint-based approa
- Instance-based strategies
- Use of auxiliary information



# Example matchers

- Use of WordNet
  - Use WordNet to find synonyms
  - Use WordNet to find ancestors and descendants in the is-a hierarchy
- Use of Unified Medical Language System (UMLS)
  - Includes many ontologies
  - Includes many alignments (not complete)
  - Use UMLS alignments in the computation of the similarity values

**Table 7** Matching Strategies in the participating systems - 1

| System | String-based strategies | Structure-based strategies | Constraint-based strategies | Instance-based strategies |
|---|---|---|---|---|
| AgreementMaker | SubString, Edit-Distance, TF-IDF | ✓ | ✓ | ✓ |
| ALIN | SimMetrics APP , WS4J APP[b] | ✓ | - | - |
| AML | Jaccard, I-Sub | ✓ | ✓ | ✓ |
| Anchor-Flood | Jaro-Winkler | ✓ | - | ✓ |
| AOAS | Jaro-Winkler | ✓ | - | - |
| AOT, AOTL | Edit-Distance, Block-Distance, SLIM-Winkler, Jaro-Winkler, Smith-Winkler, Needleman-Wunsch | - | - | - |
| AROMA | Jaro-Winkler | ✓ | ✓ | - |
| ASMOV | Edit-Distance | ✓ | ✓ | ✓ |
| BLOOMS | Jaccard, Exact Match, Lin, Jaro-Winkler | - | - | - |
| CIDER-CL | Soft TF-IDF, Jaro-Winkler | ✓ | - | - |
| CODI | Edit-Distance, Jaro-Winkler, Cosine, Smith-Waterman, Jaccard, Overlap coefficient | ✓ | ✓ | ✓ |
| COMMAND | UMBC similarity Model | ✓ | - | - |
| CroMatcher | N-Gram, TF-IDF | ✓ | ✓ | ✓ |
| CSA | Edit-Distance, Wu-Palmer, TF-IDF | ✓ | - | ✓ |
| DKP-AOM, DKP-AOM-Lite | SimMetrics APP[b] | ✓ | ✓ | - |
| DSSim | Jaccard, Jaro-Winkler | ✓ | - | - |
| Eff2Match | Exact Match, TF-IDF | ✓ | - | - |
| Falcon-AO | I-Sub, TF-IDF | ✓ | - | - |
| FCA-Map | Exact Match | ✓ | - | - |
| GeRoMeSuite+SMB | Edit-Distance, Jaro-Winkler, I-Sub, Soft TF-IDF, SecondString Library[c] | ✓ | - | ✓ |
| GMap | Edit-Distance, TF-IDF | ✓ | - | - |
| GOMMA, GOMMA-bk | Exact Match, N-gram | ✓ | - | ✓ |
| Hertuda | Damerau-Levenshtein[d] | - | - | - |
| HotMatch | Damerau-Levenshtein[d] | ✓ | ✓ | ✓ |
| IAMA | Edit-Distance | - | - | ✓ |

Dragisic Z, Ivanova V, Li H, Lambrix P,
Experiences from the Anatomy track in the Ontology Alignment Evaluation Initiative,
*Journal of Biomedical Semantics* 8:56, 2017

**Table 8** Matching strategies in the participating systems - 2

| System | String-based strategies | Structure-based strategies | Constraint-based strategies | Instance-based strategies |
|---|---|---|---|---|
| JarvisOM | Cosine, WuPalmer, Lin, N-gram | - | - | - |
| KOSIMap | SimMetrics API[a], Degree of commonality coefficient | ✓ | ✓ | - |
| Lily | Edit-Distance | ✓ | ✓ | ✓ |
| LogMap | I-Sub | ✓ | - | ✓ |
| LPHOM | I-Sub, Monge-Elkan, 3-Gram, Jaccard, Lin | - | - | - |
| LYAM++ | SOFT TF-IDF, Jaccard | ✓ | - | - |
| MaasMatch | Cosine, Edit-Distance, Jaccard, 3-Gram, Longest Common Substring | ✓ | - | ✓ |
| MapSSS | Edit-Distance, Choice based on [10] | ✓ | ✓ | - |
| NBJLM | Set of words-level | ✓ | - | - |
| ODGOMS | Longest Common Subsequence, SMOA, TF-IDF | ✓ | - | - |
| Optima+ | Lin, Smith-Waterman, Needleman-Wunsch Inverse Edit-Distance | ✓ | - | - |
| Prior+ | Edit-Distance | ✓ | - | - |
| RiMOM | Edit-Distance, Cosine | ✓ | - | ✓ |
| RSDLWB | Jaccard, Substring | ✓ | ✓ | - |
| SAMBO, SAMBOdtf | Edit-Distance, 3-Gram | ✓ | - | ✓ |
| ServOMap | Edit-Distance, I-Sub, Q-Gram, TF-IDF, Monge-Elkan, Jaccard | ✓ | - | - |
| SOBOM | I-Sub | ✓ | - | - |
| StringsAuto | Choice based on [10] | - | - | - |
| TaxoMap | Lin, 3-gram Degree of commonality coefficient | ✓ | ✓ | - |
| TOAST | ✓[b] | ✓ | - | - |
| WeSeE | Edit-Distance, TF-IDF | - | - | - |
| WikiMatch | Jaccard | - | - | - |
| X-SOM | Edit-Distance, Jaro | ✓ | - | ✓ |
| XMap | Edit distance, Jaro-Winkler, N-gram, Jaccard, Cosine | ✓ | ✓ | - |
| YAM++ | Tversky[c], TF-IDF | ✓ | - | ✓ |

**Table 9** Use of auxiliary information by the participating systems

| System | Background knowledge | | | | | | |
|---|---|---|---|---|---|---|---|
| | UMLS | Uberon | BioPortal | MeSH | FMA | WordNet | Other |
| AgreementMaker | ✓ | ✓ | - | - | - | ✓ | - |
| ALIN | - | - | - | - | - | ✓ | - |
| AML | ✓ | ✓ | | ✓ | | ✓ | - |
| Anchor-Flood | - | - | - | - | - | ✓ | - |
| AOAS | ✓ | - | - | - | ✓ | - | - |
| AOT, AOTL | - | - | - | - | - | ✓ | - |
| ASMOV | ✓ | - | - | - | - | ✓ | - |
| COMMAND | ✓ | - | - | - | - | ✓ | - |
| CroMatcher | - | ✓ | - | - | - | ✓ | - |
| CSA | - | - | - | - | - | ✓ | - |
| DKP-AOM | - | - | - | - | - | ✓ | - |
| DSSim | - | - | - | - | - | ✓ | - |
| Eff2Match | - | - | - | - | - | ✓ | - |
| GOMMA | ✓ | ✓ | - | - | ✓ | - | - |
| GeRoMeSuite+SMB | - | - | - | - | - | ✓ | - |
| Hotmatch | - | - | - | - | - | - | API tanes[a], WikiPedia, Big Huge Thesaurus[b] |
| JarvisOM | - | - | - | - | - | ✓ | Apache Lucene[c] |
| IAMA | - | - | - | - | - | - | Apache Lucene[c] |
| Lily | - | - | - | - | - | - | Web search (Google) |
| LogMapBio | - | - | ✓ | - | - | - | - |
| LYAM++ | - | ✓ | - | - | - | - | BabelNet[d] |
| MaasMatch | - | - | - | - | - | ✓ | - |
| MapSSS | - | - | - | - | - | - | Google |
| NBJLM | - | - | - | - | - | ✓ | - |
| Optima+ | - | - | - | - | - | ✓ | - |
| RIMOM | ✓ | - | - | - | - | ✓ | Wiki Pages |
| RSDLWB | - | - | - | - | - | ✓ | DBpedia[e] |
| SAMBO | ✓ | - | - | - | - | ✓ | - |
| ServOMap | - | - | - | - | - | ✓ | Apache Lucene[c] |
| TaxoMap | - | - | - | - | - | ✓ | - |
| TOAST | - | - | - | - | - | ✓ | - |
| WeSeE | - | - | - | - | - | - | Microsoft Bing Search JFreeWebSearch[f] |
| WikiMatch | - | - | - | - | - | - | WikiPedia |
| XMap | ✓ | | - | - | - | ✓ | - |
| X-SOM | - | - | - | - | - | ✓ | Google |
| YAM++ | - | - | - | - | - | - | Apache Lucene[c] |

# Combinations

# Combination Strategies

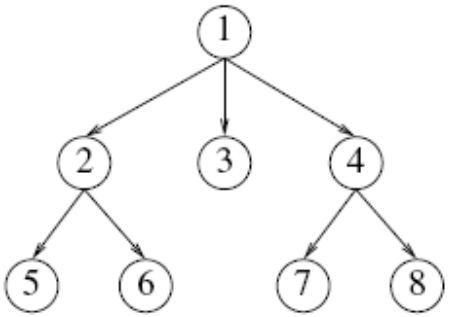
- Usually weighted sum of similarity values of different matchers
- Maximum of similarity values of different matchers

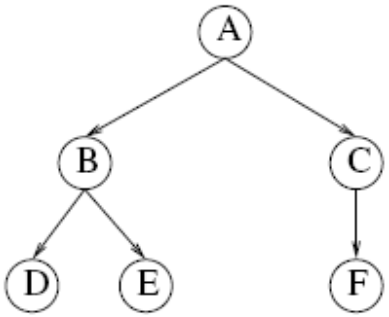
# Filtering

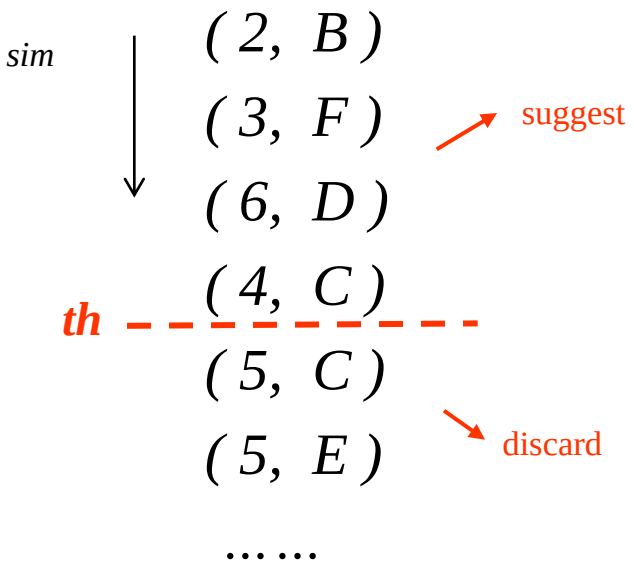
# Filtering techniques

- Threshold filtering

  Pairs of concepts with similarity higher or equal than threshold are alignment suggestions



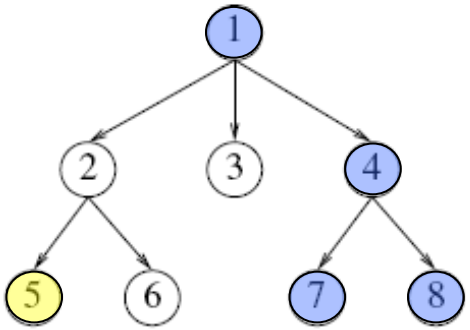



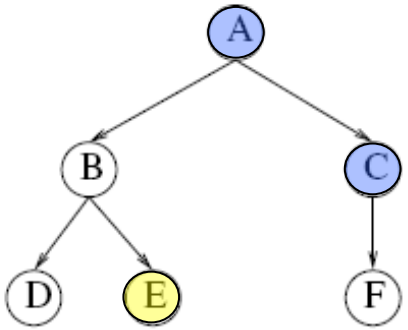
# Filtering techniques

- Double threshold filtering

  (1) Pairs of concepts with similarity higher than or equal to **upper** threshold are alignment suggestions

  (2) Pairs of concepts with similarity between **lower** and **upper** thresholds are alignment suggestions if they make sense with respect to the structure of the ontologies and the suggestions according to (1)




( 2, B )
( 3, F )
( 6, D )
( 4, C )
*upper-th*
( 5, C )
( 5, E )
*lower-th*
……

# Example alignment system

**SAMBO** – matchers, combination, filter

# Example alignment system **SAMBO** – suggestion mode

**Table 6** Analysis of the components of the participating systems

| Systems | Basic processes | | | | | |
|---|---|---|---|---|---|---|
| | Preprocessing[D/R] | Matching | Combination | Filtering | Debugging | User interaction* |
| AgreementMaker | - | ✓ | ✓ | ✓ | - | ✓[1] |
| ALIN | - | ✓ | ✓ | ✓ | - | ✓ |
| AML, AML_bk | D | ✓ | ✓ | ✓ | ✓ | ✓[1] |
| Anchor-Flood | D | ✓ | ✓ | ✓ | - | - |
| AOAS | - | ✓ | ✓ | ✓ | - | - |
| AOT, AOTL | - | ✓ | ✓ | ✓ | - | - |
| AROMA | D | ✓ | ✓ | ✓ | - | - |
| ASMOV | - | ✓ | ✓ | ✓ | ✓ | ✓ |
| BLOOMS | D | ✓ | ✓ | ✓ | - | - |
| CIDER-CL | D | ✓ | ✓ | ✓ | - | - |
| CODI | D | ✓ | ✓ | ✓ | ✓ | - |
| COMMAND | - | ✓ | ✓ | ✓ | - | - |
| CroMatcher | D | ✓ | ✓ | ✓ | - | - |
| CSA | D | ✓ | ✓ | ✓ | - | - |
| DKP-AOM, DKP-AOM-Lite | D | ✓ | ✓ | ✓ | ✓ | - |
| DSSim | R | ✓ | ✓ | ✓ | - | - |
| Eff2Match | D | ✓ | ✓ | ✓ | - | - |
| Falcon-AO | R | ✓ | ✓ | ✓ | - | ✓[1] |
| FCA-Map | D | ✓ | - | - | ✓ | - |
| GeRoMeSuite+SMB | - | ✓ | ✓ | ✓ | ✓ | ✓[1] |
| GMap | - | ✓ | ✓ | ✓ | - | - |
| GOMMA, GOMMAbk | R | ✓ | ✓ | ✓ | ✓ | ✓(✓)[1] |
| Hertuda | D | ✓ | - | ✓ | - | ✓ |
| HotMatch | D | ✓ | ✓ | ✓ | - | - |
| IAMA | D | ✓ | ✓ | ✓ | - | - |

Dragisic Z, Ivanova V, Li H, Lambrix P,
Experiences from the Anatomy track in the Ontology Alignment Evaluation Initiative,
*Journal of Biomedical Semantics* 8:56, 2017

| | | | | | | |
|---|---|---|---|---|---|---|
| JarvisOM | D | ✓ | ✓ | ✓ | - | ✓ |
| KOSIMap | D | ✓ | ✓ | ✓ | ✓ | - |
| Lily | D | ✓ | ✓ | ✓ | ✓ | ✓[1] |
| LogMap, LogMapBio, LogMapC, LogMapLite | D,R | ✓ | ✓ | ✓ | ✓ | ✓[1] |
| LPHOM | D | ✓ | ✓ | ✓ | - | - |
| LYAM++ | D | ✓ | - | ✓ | - | - |
| MaasMatch | D | ✓ | ✓ | ✓ | - | - |
| MapSSS | - | ✓ | ✓ | ✓ | - | - |
| NBJLM | - | ✓ | ✓ | ✓ | - | - |
| ODGOMS | D | ✓ | ✓ | ✓ | - | - |
| Optima+ | - | ✓ | ✓ | ✓ | - | - |
| Prior+ | D | ✓ | ✓ | ✓ | - | - |
| RiMOM | D | ✓ | ✓ | ✓ | - | - |
| RSDLWB | D | ✓ | ✓ | - | - | ✓[1] |
| SAMBO, SAMBOdtf | - | ✓ | ✓ | ✓ | ✓ | ✓[1] |
| ServOMap(Lt), ServOMBI | D | ✓ | ✓ | ✓ | ✓ | ✓ |
| SOBOM | - | ✓ | ✓ | ✓ | - | - |
| StringsAuto | - | ✓ | ✓ | ✓ | - | - |
| TaxoMap | D,R | ✓ | ✓ | ✓ | - | - |
| TOAST | - | ✓ | - | - | - | - |
| WeSeE | D | ✓ | - | ✓ | - | ✓ |
| WikiMatch | D | ✓ | - | ✓ | - | - |
| X-SOM | - | ✓ | ✓ | ✓ | ✓ | - |
| XMap, XMAPGen, XMAPSig | - | ✓ | ✓ | ✓ | - | ✓ |
| YAM++ | D | ✓ | ✓ | ✓ | ✓ | - |

# Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Ontology alignment challenges

# Evaluation measures

- Precision:

$$\frac{\text{\# correct mapping suggestions}}{\text{\# mapping suggestions}}$$

- Recall:

$$\frac{\text{\# correct mapping suggestions}}{\text{\# correct mappings}}$$

- F-measure: combination of precision and recall

# Ontology Alignment Evaluation Initiative

http://oaei.ontologymatching.org/

# OAEI

- Since 2004, Evaluation of *systems*
- Different tracks (2023)
  - Ontologies
    - Anatomy, conference, ,Bio-ML, biodiversity and ecology, food nutrition , materials science, …
    - Multilingual: multifarm (9 languages)
    - Complex
    - Interactive
  - Instance matching and link discovery
  - Tabular data to Knowledge Graphs
  - Knowledge graphs

# OAEI

- Evaluation measures
  - Precision/recall/f-measure
  - recall of non-trivial mappings
  - full / partial golden standard

# OAEI 2023

- Anatomy:
  - 9 systems
  - best system for f:  f=0.941, p=0.951, r=0.931, r+=0.818, 54 seconds
  - 2 systems produce coherent mappings

# OAEI Anatomy Track 2007-2016*

- Components
  - Almost all systems implement preprocessing, matchers, combination, filtering components
  - Debugging component and GUI rarely implemented
- Matching strategies
  - Variety of string-based strategies
  - Most often string and structured-based strategies
- Use of background knowledge
  - Almost all systems use sources of background knowledge

* Dragisic Z, Ivanova V, Li H, Lambrix P, Experiences from the Anatomy track in the Ontology Alignment Evaluation Initiative, *Journal of Biomedical Semantics* 8:56, 2017.

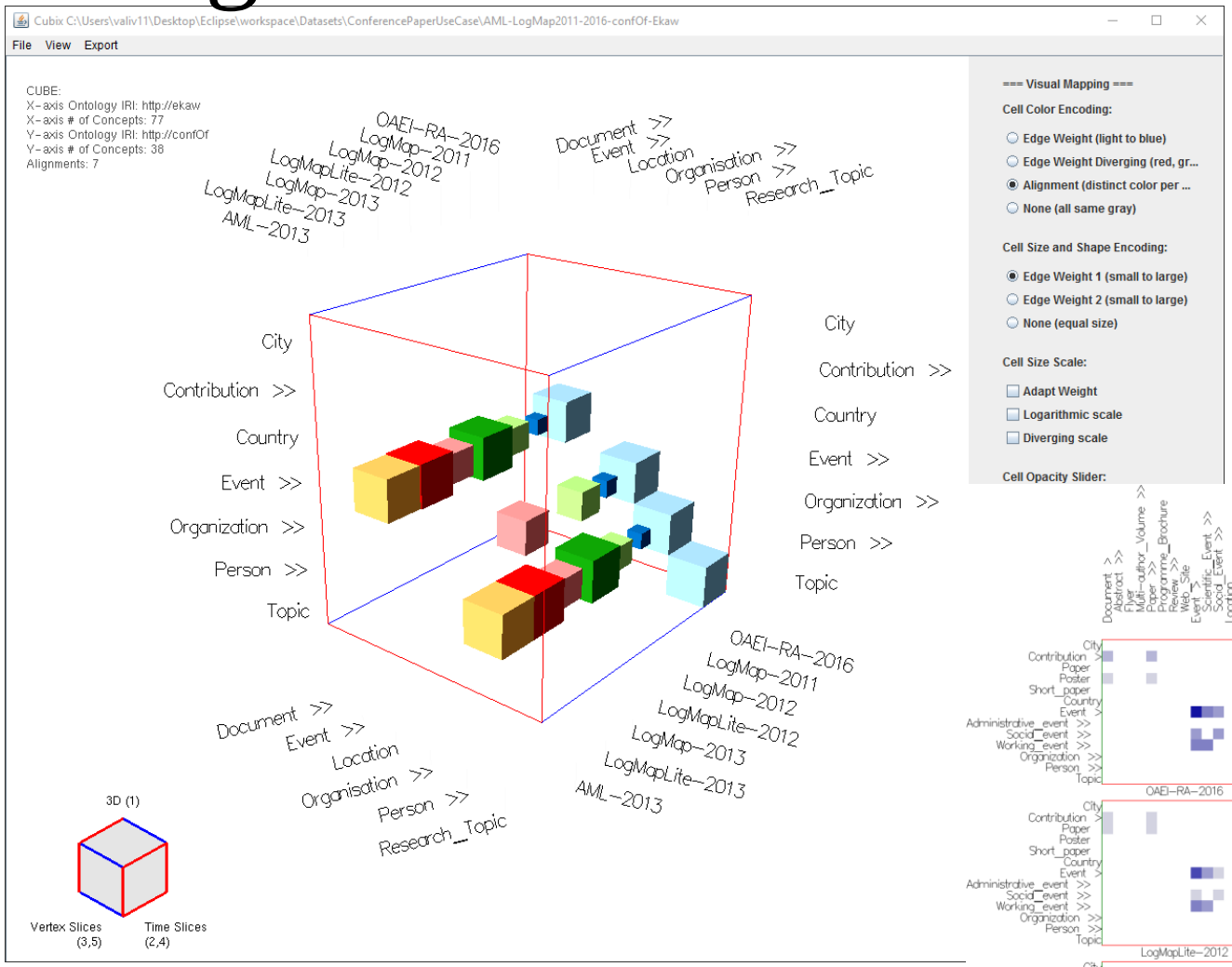
# Complementary evaluation

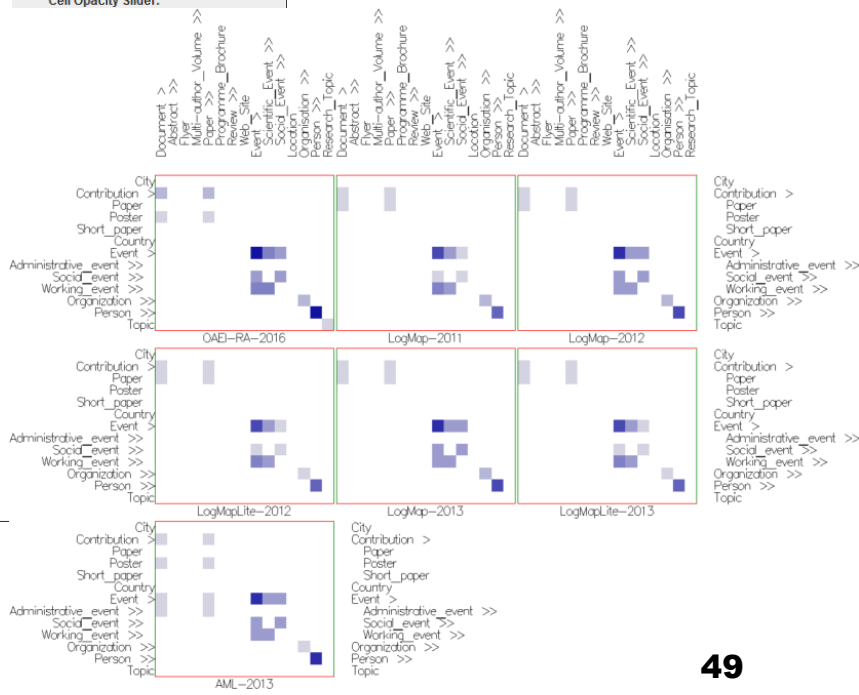
Alignment cubes

- Interactive visualization of alignments
- Region-level, mapping level
- Missing mappings
- Often found mappings

- http://www.ida.liu.se/~patla00/research/AlignmentCubes/

# Alignment cubes



# Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Ontology alignment challenges

# Challenges

- Large-scale matching evaluation
- Efficiency of matching techniques
  - parallellization
  - distribution of computation
  - approximation of matching results (not complete)
  - modularization of ontologies
  - optimization of matching methods

# Challenges

- Matching with background knowledge
  - partial alignments
  - reuse of previous matches
  - use of domain-specific corpora
  - use of domain-specific ontologies
- Matcher selection, combination and tuning
  - recommendation of algorithms and settings

# Challenges

- User involvement
  - visualization
  - user feedback
- Explanation of matching results
- Social and collaborative matching
- Alignment management: infrastructure and support

# Further reading

Starting points for further studies

# Further reading ontology alignment

- http://www.ontologymatching.org

(plenty of references to articles and systems)

- Ontology alignment evaluation initiative: http://oaei.ontologymatching.org

(home page of the initiative)

- Euzenat, Shvaiko, *Ontology Matching*, Springer, 2007.

- Shvaiko, Euzenat, Ontology Matching: state of the art and future challenges, *IEEE Transactions on Knowledge and Data Engineering* 25(1):158-176, 2013.

- Dragisic Z, Ivanova V, Li H, Lambrix P, Experiences from the Anatomy track in the Ontology Alignment Evaluation Initiative, *Journal of Biomedical Semantics* 8:56, 2017.

# Further reading ontology alignment

Systems at LiU / IDA / ADIT

- Lambrix, Tan, SAMBO – a system for aligning and merging biomedical ontologies, *Journal of Web Semantics*, 4(3):196-206, 2006.

(description of the SAMBO tool and overview of evaluations of different matchers)

- Lambrix, Tan, A tool for evaluating ontology alignment strategies, *Journal on Data Semantics*, VIII:182-202, 2007.

(description of the KitAMO tool for evaluating matchers)

- Lambrix P, Kaliyaperumal R, A Session-based Ontology Alignment Approach enabling User Involvement, *Semantic Web Journal* 8(2):225-251, 2017.

- Ivanova V, Bach B, Pietriga E, Lambrix P, Alignment Cubes: Towards Interactive Visual Exploration and Evaluation of Multiple Ontology Alignments, 16th International Semantic Web Conference, 400-417, 2017.

# Further reading ontology alignment

- Chen, Tan, Lambrix, Structure-based filtering for ontology alignment,*IEEE WETICE workshop on semantic technologies in collaborative applications*, 364-369, 2006.

(double threshold filtering technique)

- Tan, Lambrix, A method for recommending ontology alignment strategies, *International Semantic Web Conference*, 494-507, 2007.

  Ehrig, Staab, Sure, Bootstrapping ontology alignment methods with APFEL, *International Semantic Web Conference*, 186-200, 2005.

  Mochol, Jentzsch, Euzenat, Applying an analytic method for matching approach selection, *International Workshop on Ontology Matching*, 2006.

(recommendation of alignment strategies)

- Lambrix, Liu, Using partial reference alignments to align ontologies, *European Semantic Web Conference*, 188-202, 2009.

(use of partial alignments in ontology alignment)

# Further reading ontology alignment

User Involvement

- Li H, Dragisic Z, Faria D, Ivanova V, Jimenez-Ruiz E, Lambrix P, Pesquita C, User validation in ontology alignment: functional assessment and impact, *The Knowledge Engineering Review*, 2019.

- Ivanova V, Lambrix P, Åberg J, Requirements for and Evaluation of User Support for Large-Scale Ontology Alignment, *12th Extended Semantic Web Conference - ESWC 2015*, LNCS 9088, 3-20, 2015.

# Ontology Completion and Debugging

# Defects in ontologies

- Syntactic defects
  - E.g. wrong tags or incorrect format
- Semantic defects
  - E.g. unsatisfiable concepts, incoherent and inconsistent ontologies
- Modeling defects
  - E.g. wrong or missing relations

# Example - incoherent ontology

- Example: DICE ontology
  - Brain $\sqsubseteq$ CentralNervousSystem $\sqcap$ BodyPart $\sqcap$ $\exists$systempart.NervousSystem $\sqcap$ $\exists$ region.HeadAndNeck $\sqcap$ $\forall$region.HeadAndNeck

*A brain is a central nervous system and a body part which has a system part that is a nervous system and that is in the head and neck region.*

  - CentralNervousSystem $\sqsubseteq$ NervousSystem

*A central nervous system is a nervous system.*

  - BodyPart $\sqsubseteq\neg$NervousSystem

*Nothing can be at the same time a body part and a nervous system.*

Slide from G. Qi

# Example - inconsistent ontology

- **Example from Foaf:**
  - **Person(timbl)**
  - **Homepage(timbl, http://w3.org/)**
  - **Homepage(w3c, http://w3.org/)**
  - **Organization(w3c)**
  - **InverseFunctionalProperty(Homepage)**
  - **DisjointWith(Organization, Person)**
- **Example from OpenCyc:**
  - **ArtifactualFeatureType(PopulatedPlace)**
  - **ExistingStuffType(PopulatedPlace)**
  - **DisjointWith(ExistingObjectType,ExistingStuffType)**
  - **ArtifactualFeatureType ⊑ ExistingObjectType**

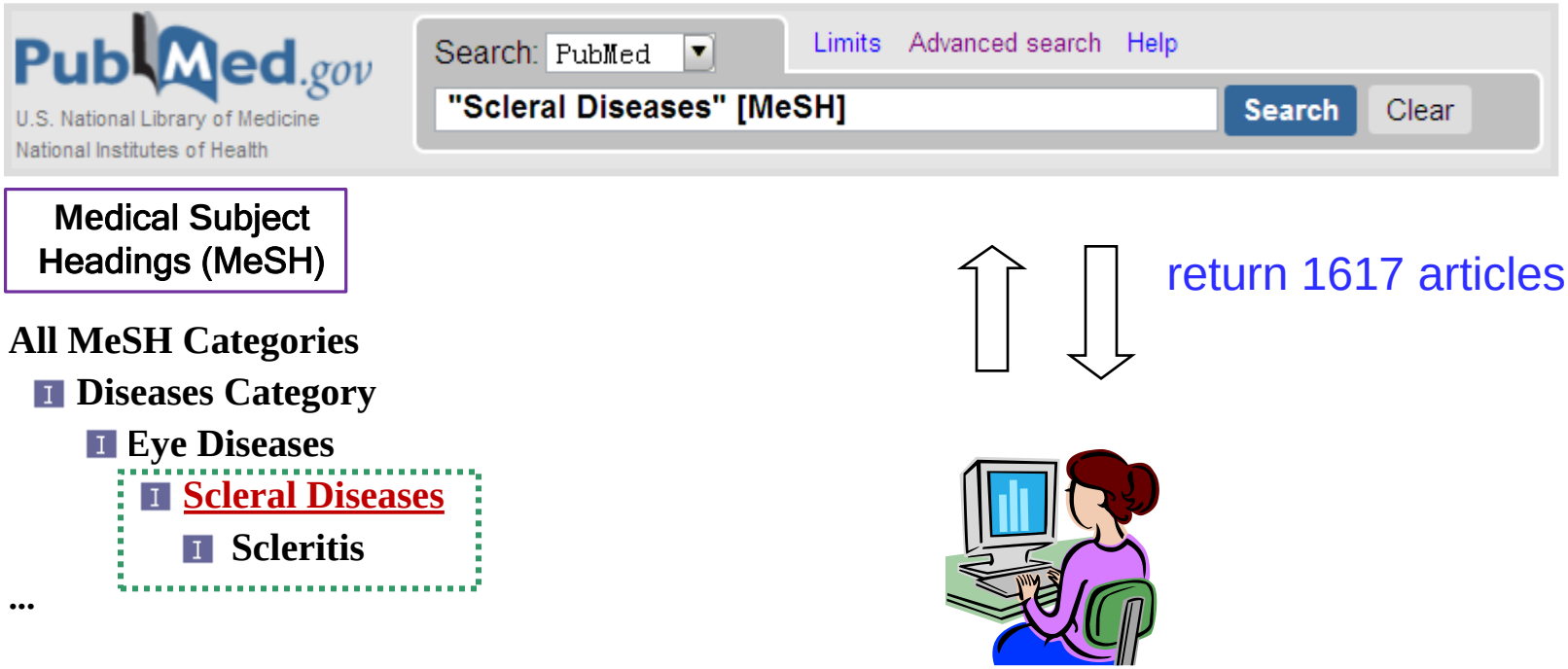
Slide from G. Qi

# Example - missing is-a relations

- In 2008 Ontology Alignment Evaluation Initiative (OAEI) Anatomy track, task 4
  - Ontology MA : Adult Mouse Anatomy Dictionary (2744 concepts)
  - Ontology NCI-A : NCI Thesaurus - anatomy (3304 concepts)
  - 988 mappings between MA and NCI-A
    - 121 missing is-a relations in MA
    - 83 missing is-a relations in NCI-A

# Influence of missing structure

- Ontology-based querying.



# Influence of missing structure

- Incomplete results from ontology-based queries


Medical Subject Headings (MeSH)

**All MeSH Categories**
- **Diseases Category**
  - **Eye Diseases**
    - **Scleral Diseases**
      - ~~**Scleritis**~~

...

return 1617 articles

return 695 articles

57% results are missed !

# Defects in ontologies and ontology networks

- Ontologies and ontology networks with defects, although often useful, also lead to problems when used in semantically-enabled applications.

→ Wrong conclusions may be derived or valid conclusions may be missed.

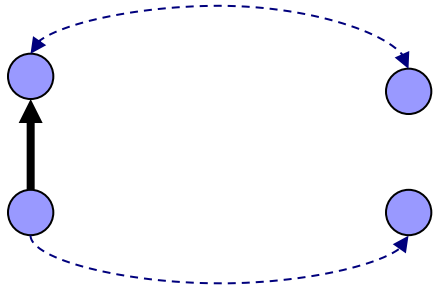
# Completion and debugging process

- Detection  (find candidate defects)
- Validation (real defects)
- Repair (remove wrong, add correct)

# Detection

Many approaches

- inspection
- ontology learning or evolution
- using linguistic and logical patterns
  - animals *such as* dogs and cats
- by using knowledge intrinsic to an ontology network



- by using machine learning and statistical methods

# Repairing

**Definition 1.** (Repair)[15]

Let $\mathcal{T}$ be a TBox. Let $M$ and $W$ be finite sets of TBox axioms. Let $Or$ be an oracle that, given a TBox axiom, returns true or false. A repair for Complete-Debug-Problem CDP$(\mathcal{T}, Or, M, W)$ is any pair of finite sets of TBox axioms $(A, D)$ such that

(i) $\forall\ \psi_a \in A$: $Or(\psi_a) =$ true;

(ii) $\forall\ \psi_d \in D$: $Or(\psi_d) =$ false;

(iii) $(\mathcal{T} \cup A) \setminus D$ is consistent;

(iv) $\forall\ \psi_m \in M$: $(\mathcal{T} \cup A) \setminus D \models \psi_m$;

(v) $\forall\ \psi_w \in W$: $(\mathcal{T} \cup A) \setminus D \not\models \psi_w$.

Current work usually focuses on debugging or completion, but not both.

Most work on debugging.

# Ontology Debugging

# Example : an Incoherent Ontology

Consider the following TBox $\mathcal{T}^*$, where $A$, $B$ and $C$ are primitive and $A_1, \ldots, A_7$ defined concept names:

| | |
|---|---|
| $ax_1 : A_1 \dot{\sqsubseteq} \neg A \sqcap A_2 \sqcap A_3$ | $ax_2 : A_2 \dot{\sqsubseteq} A \sqcap A_4$ |
| $ax_3 : A_3 \dot{\sqsubseteq} A_4 \sqcap A_5$ | $ax_4 : A_4 \dot{\sqsubseteq} \forall s.B \sqcap C$ |
| $ax_5 : A_5 \dot{\sqsubseteq} \exists s.\neg B$ | $ax_6 : A_6 \dot{\sqsubseteq} A_1 \sqcup \exists r.(A_3 \sqcap \neg C \sqcap A_4)$ |
| $ax_7 : A_7 \dot{\sqsubseteq} A_4 \sqcap \exists s.\neg B$ | |

DL Reasoner

The ontology is incoherent!
The set of unsatisfiable concepts are : $\{A_1, A_3, A_6, A_7\}$.

**What are the root causes of these defects?**

# Explain the Semantic Defects

- We need to identify the sets of axioms which are necessary for causing the logic contradictions.

| | |
|---|---|
| $ax_1{:}A_1 \dot{\sqsubseteq} \neg A \sqcap A_2 \sqcap A_3$ | $ax_2{:}A_2 \dot{\sqsubseteq} A \sqcap A_4$ |
| $ax_3{:}A_3 \dot{\sqsubseteq} A_4 \sqcap A_5$ | $ax_4{:}A_4 \dot{\sqsubseteq} \forall s.B \sqcap C$ |
| $ax_5{:}A_5 \dot{\sqsubseteq} \exists s.\neg B$ | $ax_6{:}A_6 \dot{\sqsubseteq} A_1 \sqcup \exists r.(A_3 \sqcap \neg C \sqcap A_4)$ |
| $ax_7{:}A_7 \dot{\sqsubseteq} A_4 \sqcap \exists s.\neg B$ | |

- For example, for the unsatisfiable concept "$A_1$", there are two sets of axioms.

$ax_1{:}A_1 \dot{\sqsubseteq} \neg A \sqcap A_2 \sqcap A_3$
$ax_2{:}A_2 \dot{\sqsubseteq} A \sqcap A_4$

$ax_1{:}A_1 \dot{\sqsubseteq} \neg A \sqcap A_2 \sqcap A_3$
$ax_3{:}A_3 \dot{\sqsubseteq} A_4 \sqcap A_5$
$ax_4{:}A_4 \dot{\sqsubseteq} \forall s.B \sqcap C$
$ax_5{:}A_5 \dot{\sqsubseteq} \exists s.\neg B$

# Minimal Unsatisfiability Preserving Sub-TBoxes (MUPS)

**Definition 1** Let $A$ be a concept which is unsatisfiable in a TBox $\mathcal{T}$. A set $\mathcal{T}' \subseteq \mathcal{T}$ is a *minimal unsatisfiability-preserving sub-TBox (MUPS)* of $\mathcal{T}$ if

- $A$ is unsatisfiable in $\mathcal{T}'$, and
- $A$ is satisfiable in every sub-TBox $\mathcal{T}'' \subset \mathcal{T}'$.

We will abbreviate the set of MUPS of $\mathcal{T}$ and $A$ by $mups(\mathcal{T}, A)$.

$$mups(\mathcal{T}^*, A_1) = \{\{ax_1, ax_2\}, \{ax_1, ax_3, ax_4, ax_5\}\}$$

- The MUPS of an unsatisfiable concept imply the solutions for repairing.

  → Remove at least one axiom from each axiom set in the MUPS

# Example

$$
\begin{aligned}
mups(\mathcal{T}^*, A_1) &= \{\{\cancel{ax_1}, ax_2\}, \{\cancel{ax_1}, \cancel{ax_3}, \cancel{ax_4}, ax_5\}\} \\
mups(\mathcal{T}^*, A_3) &= \{\{\cancel{ax_3}, \cancel{ax_4}, ax_5\}\} \\
mups(\mathcal{T}^*, A_6) &= \{\{\cancel{ax_1}, ax_2, \cancel{ax_4}, ax_6\}, \\
&\quad \{\cancel{ax_1}, \cancel{ax_3}, \cancel{ax_4}, ax_5, ax_6\}\} \\
mups(\mathcal{T}^*, A_7) &= \{\{\cancel{ax_4}, ax_7\}\}
\end{aligned}
$$

- Possible ways of repairing all the unsatisfiable concepts in the ontology:

$$\{ax_1, ax_3, ax_4\}$$

**How to represent all these possibilities?**

# Minimal Incoherence Preserving Sub-TBox (MIPS)

**Definition 2** Let $\mathcal{T}$ be an incoherent TBox. A TBox $\mathcal{T}' \subseteq \mathcal{T}$ is a *minimal incoherence-preserving sub-TBox (MIPS) of* $\mathcal{T}$ if

- $\mathcal{T}'$ is incoherent, and
- every sub-TBox $\mathcal{T}'' \subset \mathcal{T}'$ is coherent.

$$
\begin{aligned}
mups(\mathcal{T}^*, A_1) &= \{\{ax_1, ax_2\}, \{ax_1, ax_3, ax_4, ax_5\}\} \\
mups(\mathcal{T}^*, A_3) &= \{\{ax_3, ax_4, ax_5\}\} \\
mups(\mathcal{T}^*, A_6) &= \{\{ax_1, ax_2, ax_4, ax_6\}, \\
&\qquad \{ax_1, ax_3, ax_4, ax_5, ax_6\}\} \\
mups(\mathcal{T}^*, A_7) &= \{\{ax_4, ax_7\}\}
\end{aligned}
$$

We will abbreviate the set of MIPS of $\mathcal{T}$ by $mips(\mathcal{T})$. For $\mathcal{T}^*$ we get three MIPS:

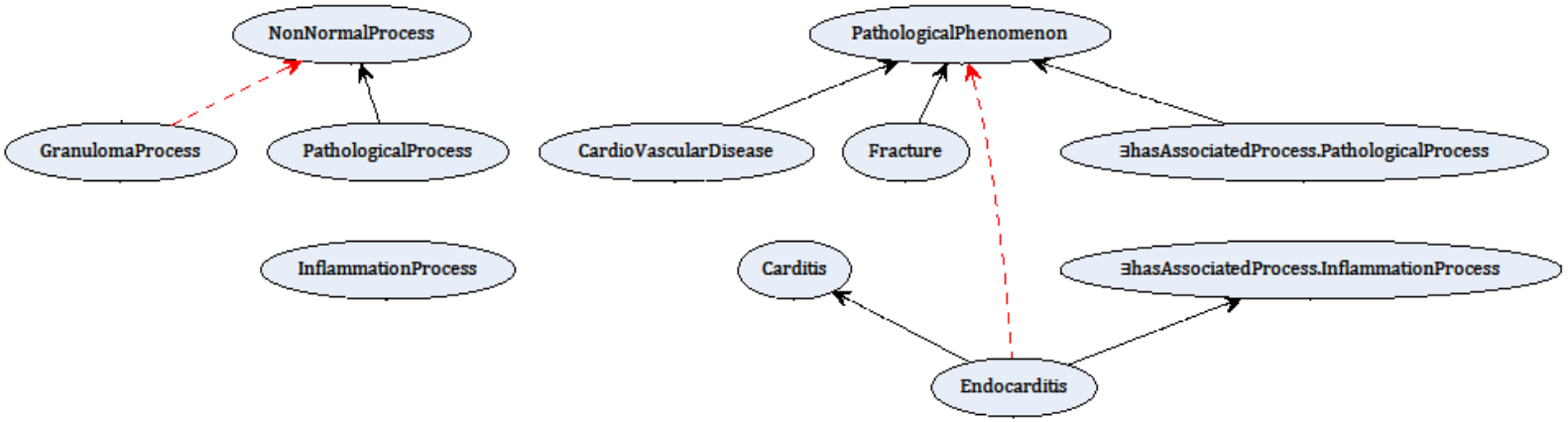
$$mips(\mathcal{T}^*) = \{\{ax_1, ax_2\}, \{ax_3, ax_4, ax_5\}, \{ax_4, ax_7\}\}$$

A possible repairing is $\{ax_i\} \cup \{ax_j\} \cup \{ax_k\}$, where

- $ax_i \in \{ax_1, ax_2\}$
- $ax_j \in \{ax_3, ax_4, ax_5\}$
- $ax_k \in \{ax_4, ax_7\}$

# Completing the is-a structure of ontologies

# Example




Repairing actions:

$$\{\text{Endocarditis} \dot{\sqsubseteq} \text{PathologicalPhenomenon}, \text{GranulomaProcess} \dot{\sqsubseteq} \text{NonNormalProcess}\}$$

$$\{\text{Carditis} \dot{\sqsubseteq} \text{CardioVascularDisease}, \text{GranulomaProcess} \dot{\sqsubseteq} \text{PathologicalProcess}\}$$

$$\{\text{Carditis} \dot{\sqsubseteq} \text{Fracture}, \text{GranulomaProcess} \dot{\sqsubseteq} \text{NonNormalProcess}\}$$

# Description logic EL

- Concepts

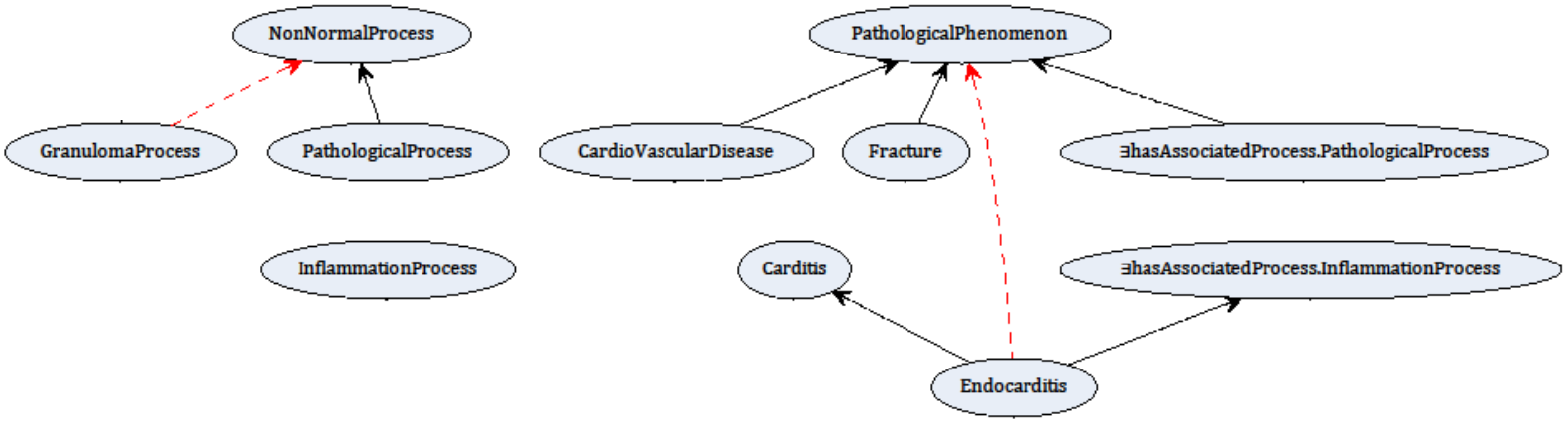
| | |
|---|---|
| Atomic concept | $A$ |
| Universal concept | $\top$ |
| Intersection of concepts | $C \sqcap D$ |
| Existential restriction | $\exists r.C$ |

- Terminological axioms: equivalence and subsumption

# Generalized Tbox Abduction Problem – GTAP(**T**,**C**,Or,M)

- Given
  - **T**- a Tbox in EL
  - **C**- a set of atomic concepts in **T**
  - $M = \{A_i \subseteq B_i\}_{i=1..n}$ and $\forall$ i:1..n: $A_i, B_i \in \mathbf{C}$
  - Or: $\{C_i \subseteq D_i \mid C_i, D_i \in \mathbf{C}\} \rightarrow \{true, false\}$
- Find
  - $S = \{E_i \subseteq F_i\}_{i=1..k}$ such that
    $\forall$ i:1..k: $E_i, F_i \in \mathbf{C}$ and $Or(E_i \subseteq F_i) = true$
    and T U S is consistent and T U S $\models$ M

# GTAP - example



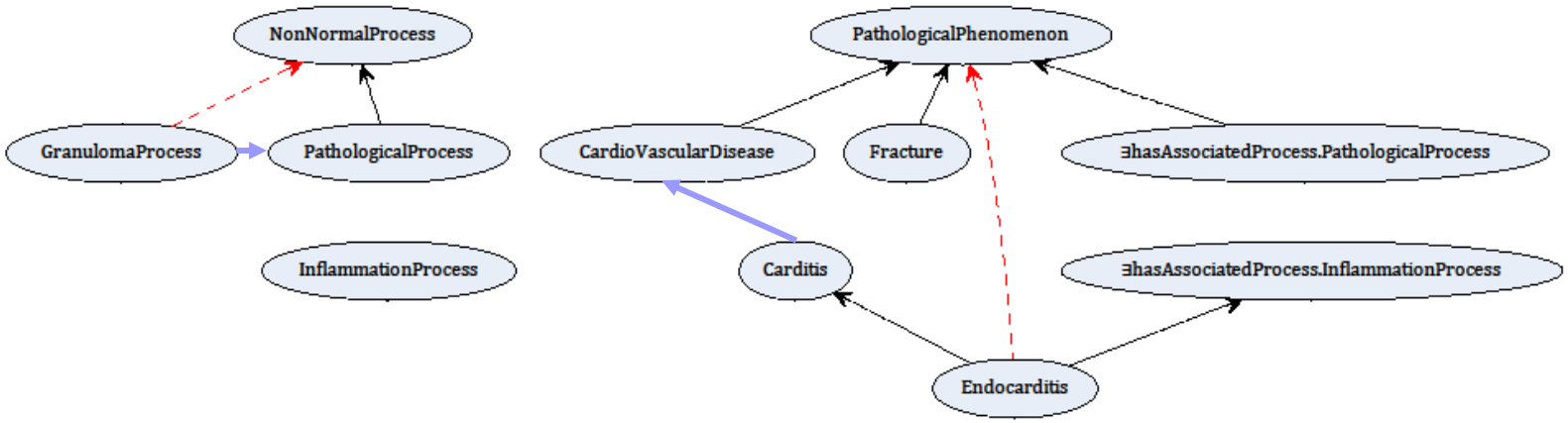

$C$ = { GranulomaProcess, CardioVascularDisease, PathologicalPhenomenon, Fracture, Endocarditis, Carditis, InflammationProcess, PathologicalProcess, NonNormalProcess}

$T$ = { GranulomaProcess $\dot{\sqsubseteq} \top$, hasAssociatedProcess $\dot{\sqsubseteq} \top \times \top$,
CardioVascularDisease $\dot{\sqsubseteq}$ PathologicalPhenomenon, Fracture $\dot{\sqsubseteq}$ PathologicalPhenomenon,
$\exists$hasAssociatedProcess.PathologicalProcess $\dot{\sqsubseteq}$ PathologicalPhenomenon,
Endocarditis $\dot{\sqsubseteq}$ Carditis, Endocarditis $\dot{\sqsubseteq}$ $\exists$hasAssociatedProcess.InflammationProcess,
PathologicalProcess $\dot{\sqsubseteq}$ NonNormalProcess }

$M$ = { Endocarditis $\dot{\sqsubseteq}$ PathologicalPhenomenon, GranulomaProcess $\dot{\sqsubseteq}$ NonNormalProcess }

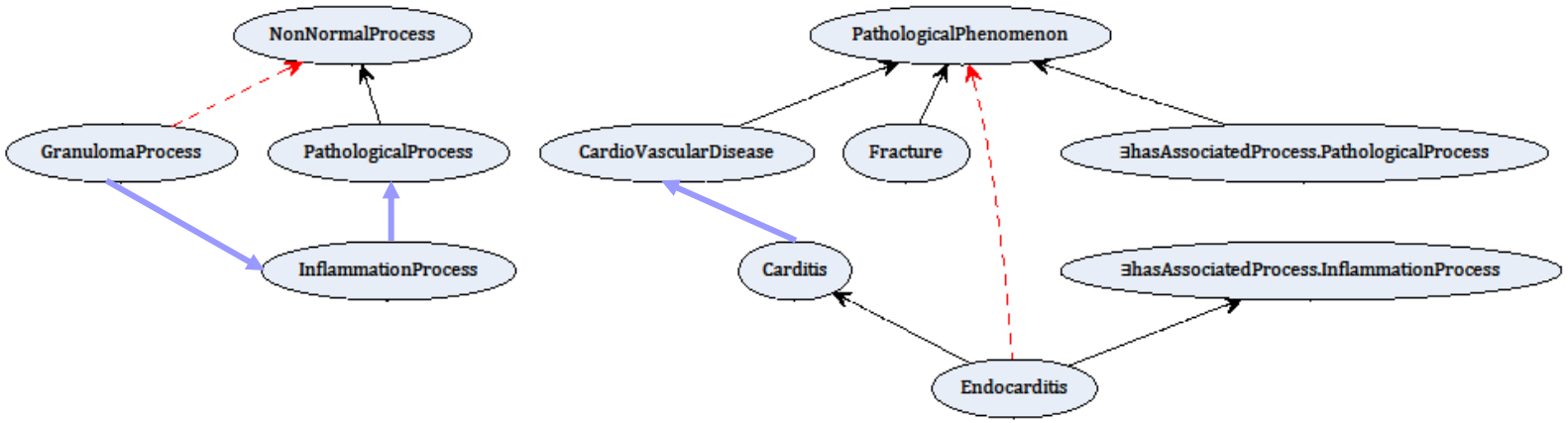
# Preference criteria

- There can be many solutions for GTAP



# Preference criteria

- There can be many solutions for GTAP




Not all are equally interesting.

# More complete

*Definition 4 (More Complete - Ontologies).* Let $\mathcal{O}_1$ and $\mathcal{O}_2$ be two ontologies represented by TBoxes $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively. Then, we say that $\mathcal{O}_1$ is *more complete* than $\mathcal{O}_2$ (or $\mathcal{O}_2$ is *less complete* than $\mathcal{O}_1$) iff $(\forall \psi : (\mathcal{T}_2 \models \psi \wedge Or(\psi) = true) \rightarrow \mathcal{T}_1 \models \psi)) \wedge (\exists \psi : Or(\psi) = true \wedge \mathcal{T}_1 \models \psi \wedge \mathcal{T}_2 \not\models \psi)$. $\mathcal{O}_1$ and $\mathcal{O}_2$ are *equally complete* iff $\forall \psi : Or(\psi) = true \rightarrow (\mathcal{T}_1 \models \psi \leftrightarrow \mathcal{T}_2 \models \psi)$.

*Definition 5 (More Complete - Repairs).* Let $\mathcal{O}$ be an ontology represented by TBox $\mathcal{T}$ and let $(A_1, D_1)$ and $(A_2, D_2)$ be two repairs for $\mathrm{CDP}(\mathcal{T}, Or, M, W)$. Let $\mathcal{O}_1$ be the ontology represented by $((\mathcal{T} \cup A_1) \setminus D_1)$ and $\mathcal{O}_2$ the ontology represented by $((\mathcal{T} \cup A_2) \setminus D_2)$. Then, repair $(A_1, D_1)$ is *more complete* than repair $(A_2, D_2)$ (or $(A_1, D_1)$ is preferred to $(A_2, D_2)$ w.r.t. "more complete") iff $\mathcal{O}_1$ is *more complete* than $\mathcal{O}_2$.

Note: 'more complete' was called 'more informative' in older papers.

# More informative

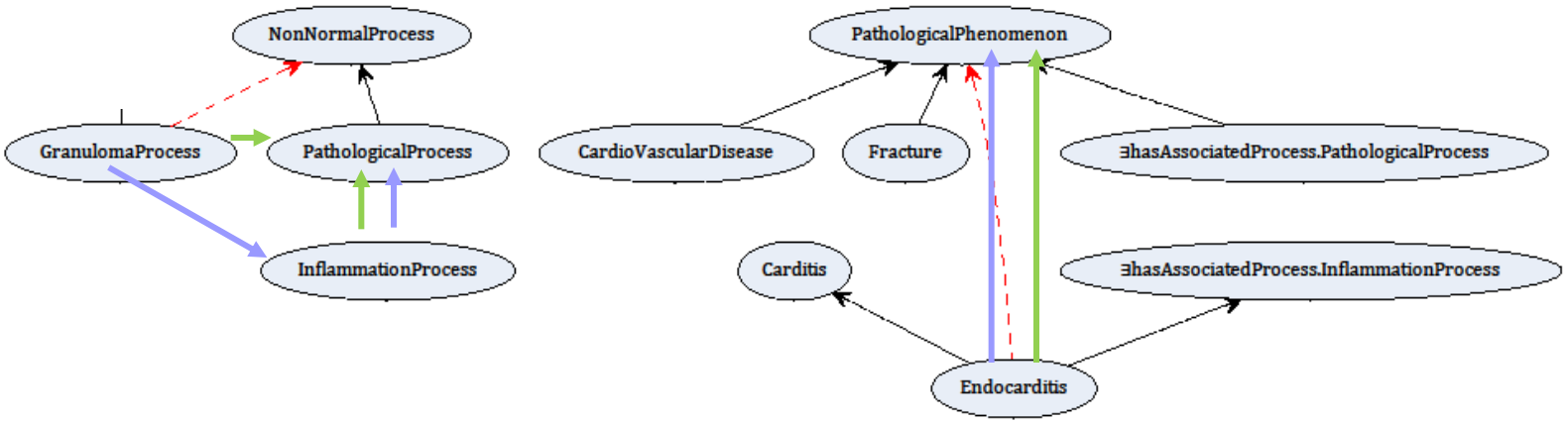
- Let *S* and *S'* be two solutions to GTAP(**T**,**C**,Or,M). Then,

- *S* is more informative than *S'*
iff **T** U *S* |= *S'* but not **T** U *S'* |= *S*

- *S* is equally informative as *S'*
iff **T** U *S* |= *S'* and **T** U *S'* |= *S*

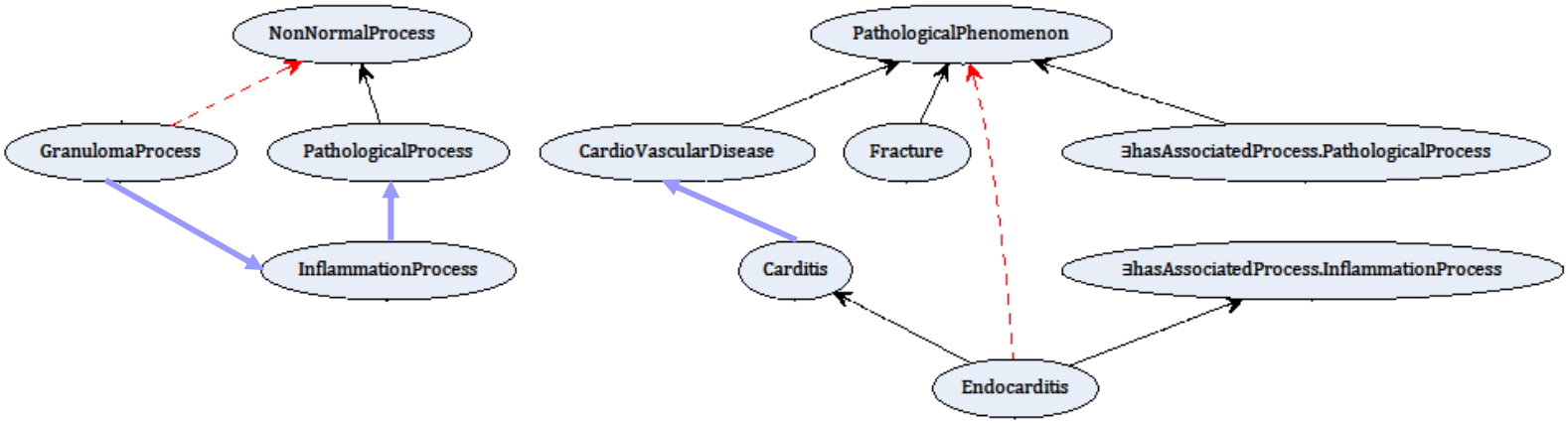
# More complete

- ’Blue’ solution is more complete than ’green’ solution



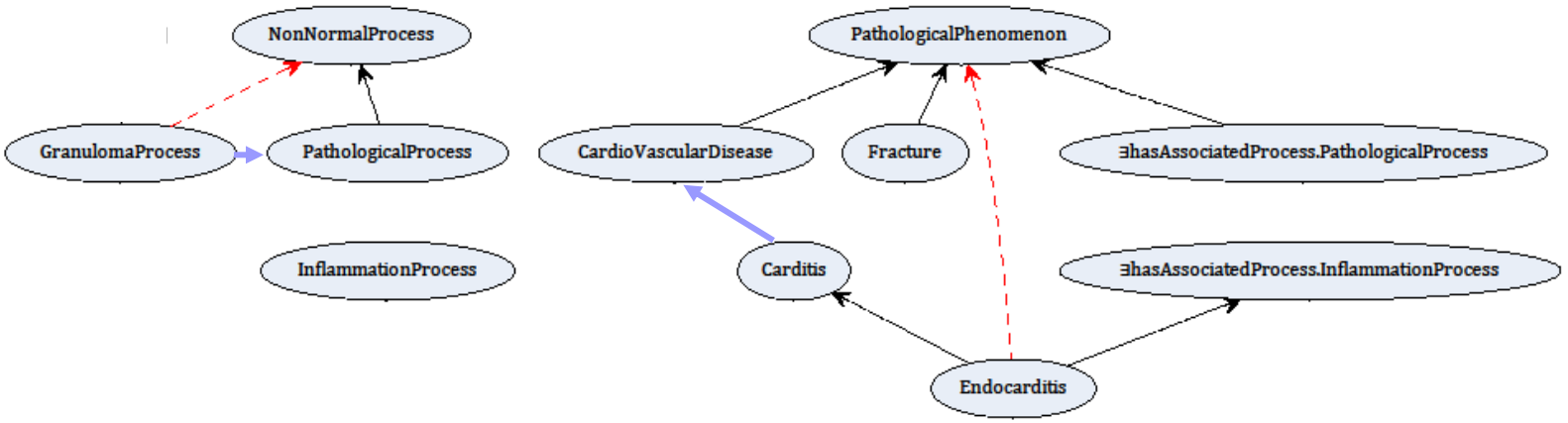
# Maximally complete

- A solution *S* to GTAP(**T**,**C**,Or,M) is maximally complete (or semantically maximal) iff there is no solution *S'* which is more complete than *S*.



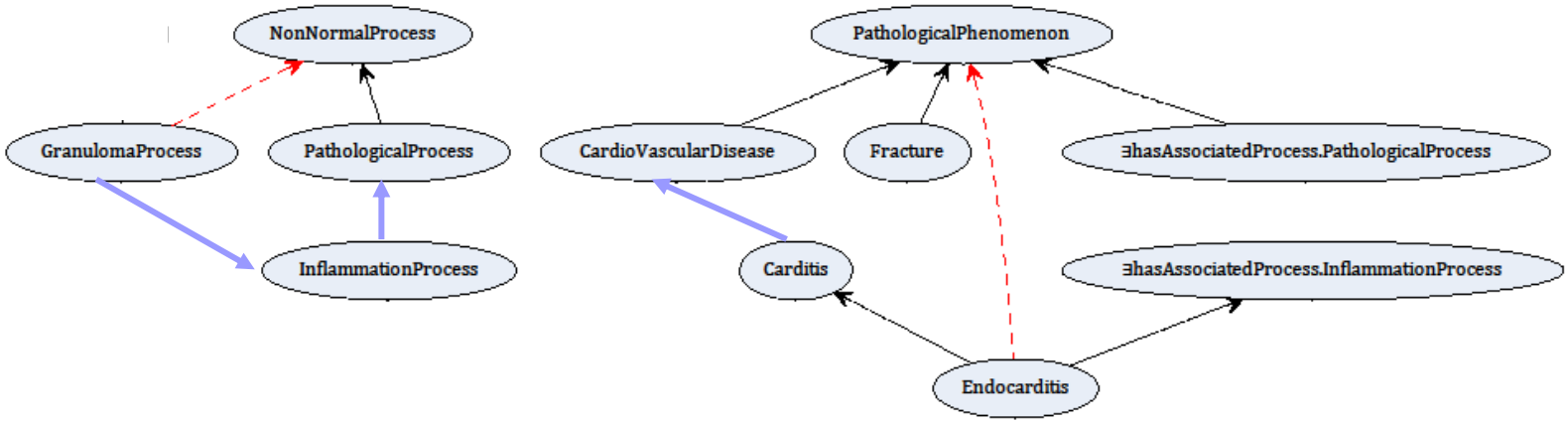
# Subset minimality

- A solution *S* to GTAP(**T**,**C**,Or,M) is subset minimal iff there is no proper subset *S'* of *S* that is a solution.



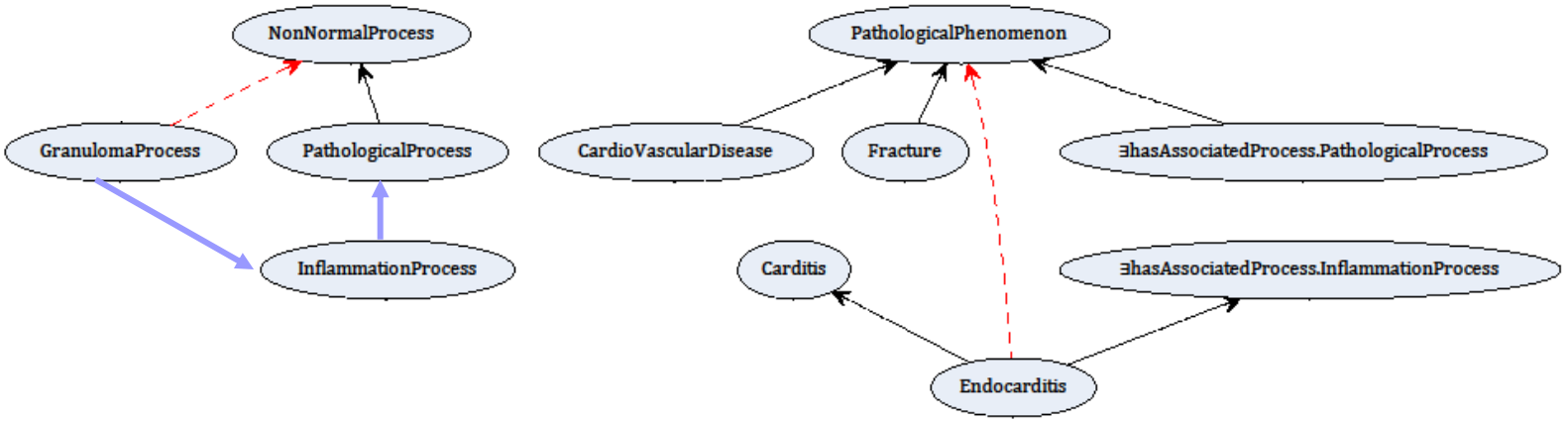
# Combining with priority for semantic maximality

- A solution *S* to GTAP(**T**,**C**,Or,M) is maxmin optimal iff *S* is maximally complete and there is no other maximally complete solution that is a proper subset of *S*.



# Combining with priority for subset minimality

- A solution *S* to GTAP(**T**,**C**,Or,M) is minmax optimal iff *S* is subset minimal and there is no other subset minimal solution that is more complete than *S*.



## Combining with equal preferences

- A solution *S* to GTAP(**T**,**C**,Or,M) is skyline optimal iff there is no other solution that is a proper subset of *S* and that is equally complete than *S*.
  - All subset minimal, minmax optimal and maxmin optimal solutions are also skyline optimal solutions.
  - Maximally complete solutions may or may not be skyline optimal.

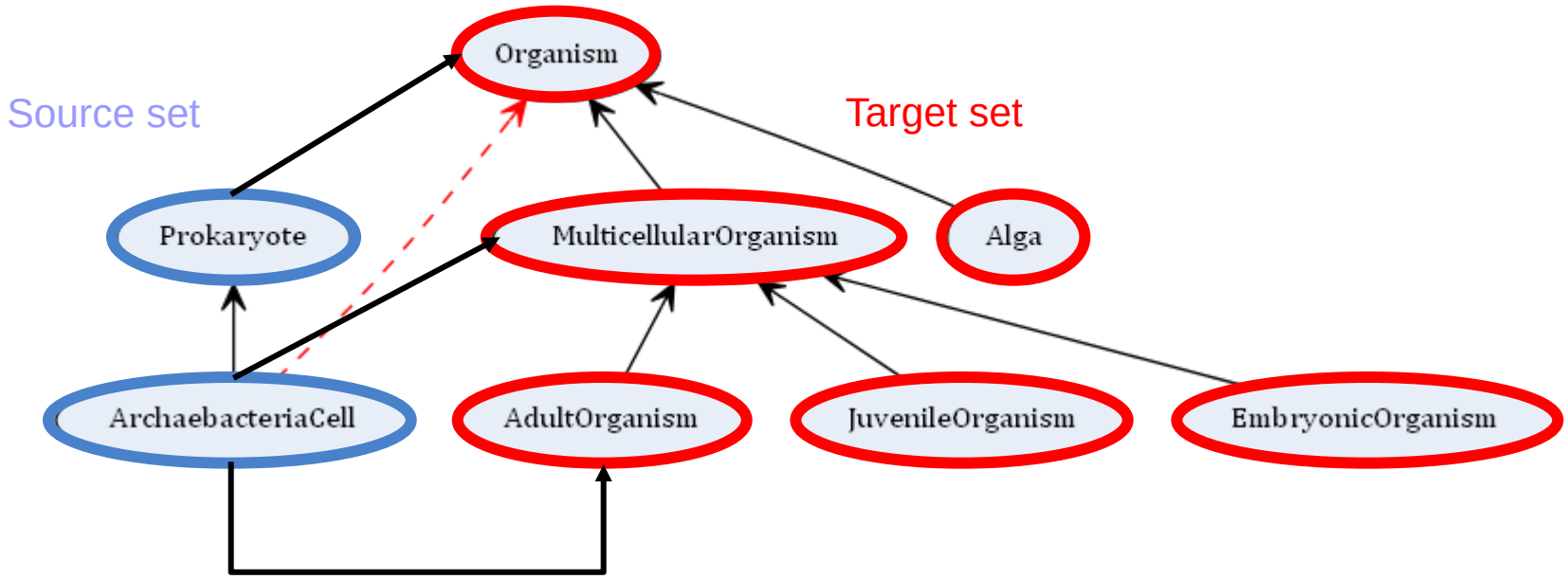
# Preference criteria - conclusions

- In practice it is not clear how to generate maxmin or maximally complete solutions (the preferred solutions)
- Skyline optimal solutions are the next best thing and are easy to generate

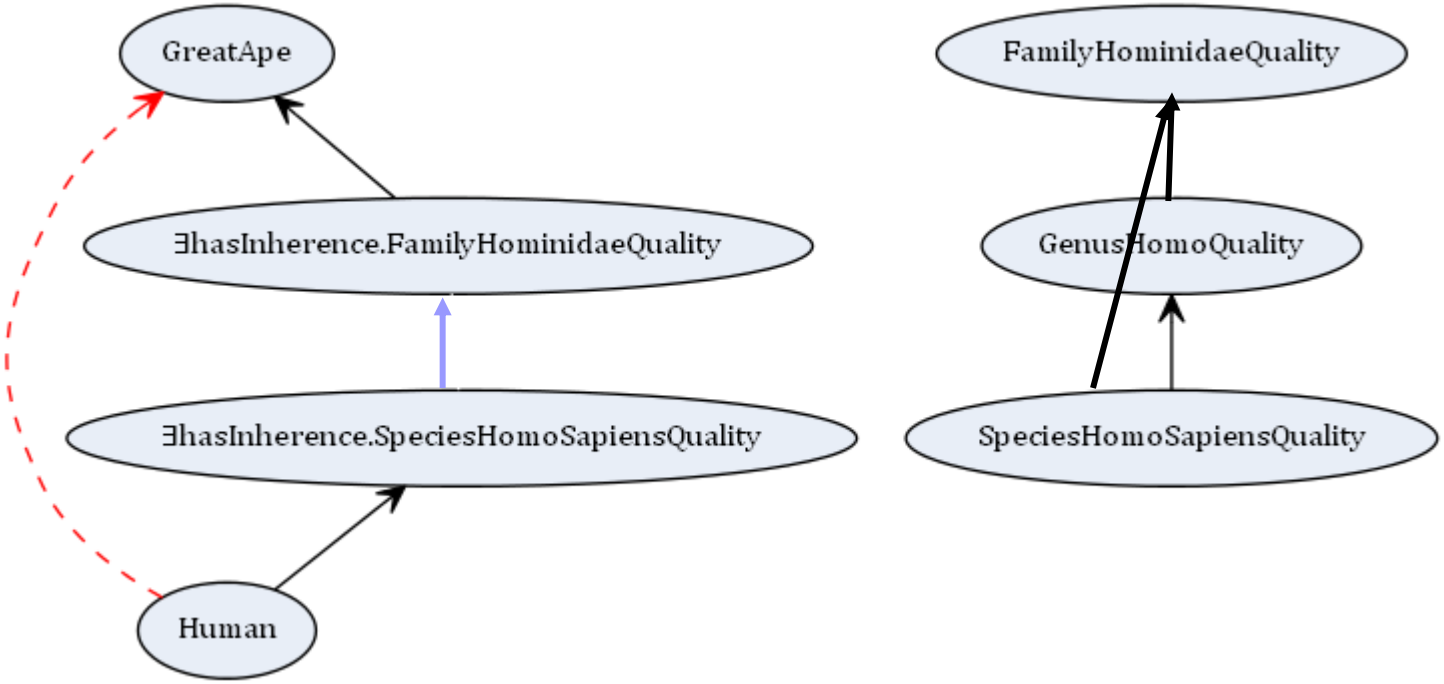
# Approach

- Input
  - Normalized EL - TBox
  - Set of missing is-a relations (correct according to the domain)
- Output – a skyline-optimal solution to GTAP
- Iteration of three main steps:
  - Creating solutions for individual missing is-a relations
  - Combining individual solutions
  - Trying to improve the result by finding a solution which introduces additional new knowledge (more complete)

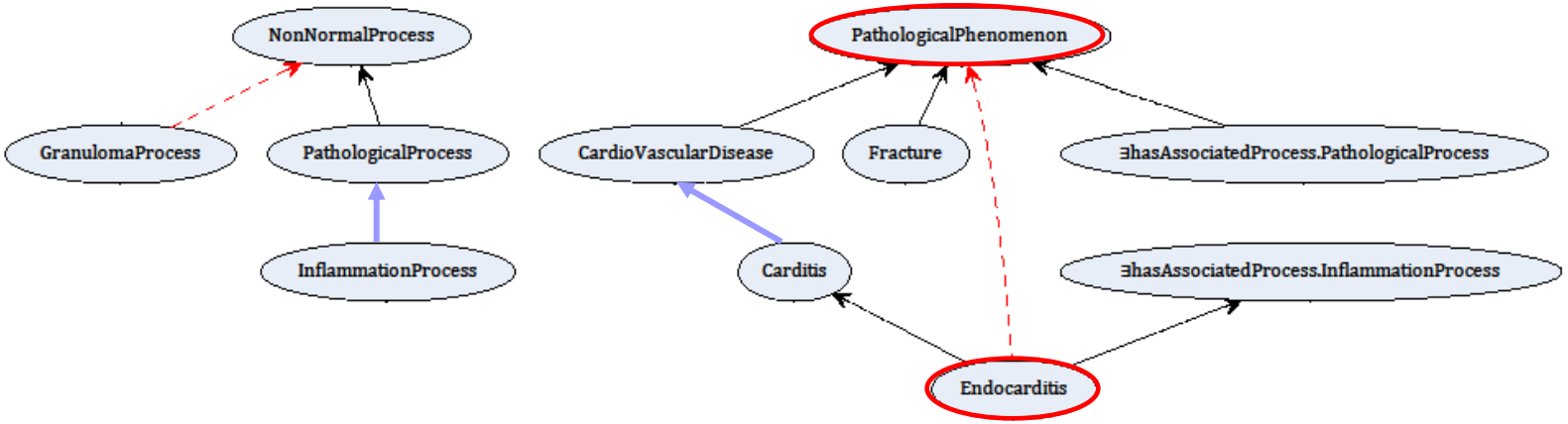
# Intuition 1


Source set
Target set
Organism
Prokaryote
MulticellularOrganism
Alga
ArchaebacteriaCell
AdultOrganism
JuvenileOrganism
EmbryonicOrganism

# Intuitions 2/3

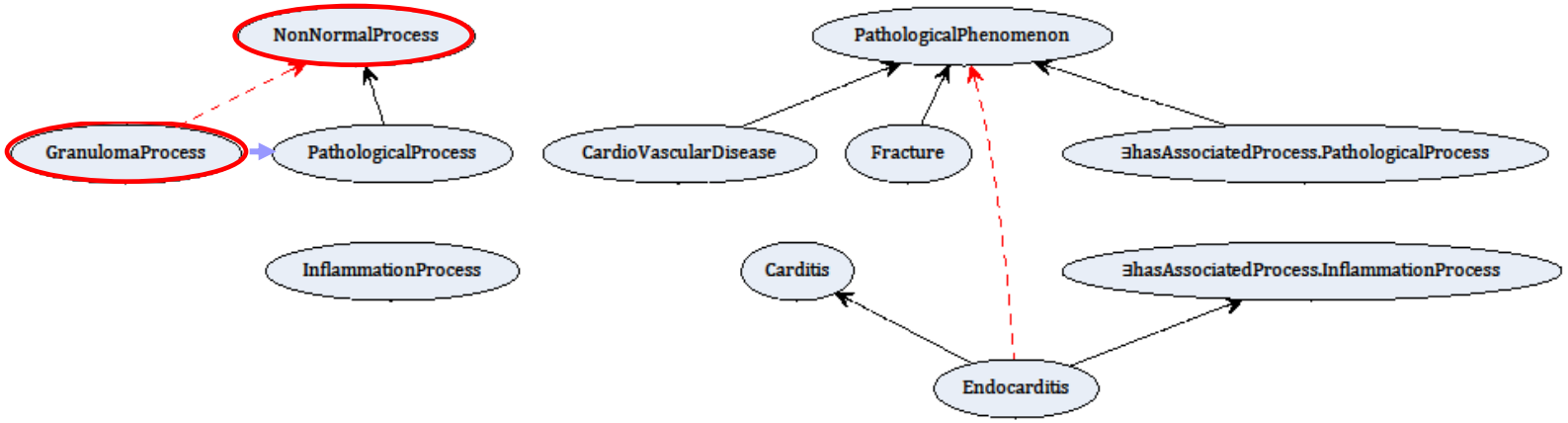


# Example – repairing single is–a relation




~~Endocarditis $\dot{\sqsubseteq}$ PathologicalPhenomenon~~
~~Endocarditis $\dot{\sqsubseteq}$ Fracture~~ false
~~Endocarditis $\dot{\sqsubseteq}$ CardioVascularDisease~~
~~Carditis $\dot{\sqsubseteq}$ PathologicalPhenomenon~~
~~Carditis $\dot{\sqsubseteq}$ Fracture~~ false
Carditis $\dot{\sqsubseteq}$ CardioVascularDisease
InflammationProcess $\dot{\sqsubseteq}$ PathologicalProcess

# Example – repairing single is–a relation




~~GranulomaProcess $\dot{\sqsubseteq}$ NonNormalProcess~~

GranulomaProcess $\dot{\sqsubseteq}$ PathologicalProcess

# Algorithm - Repairing multiple is-a relations

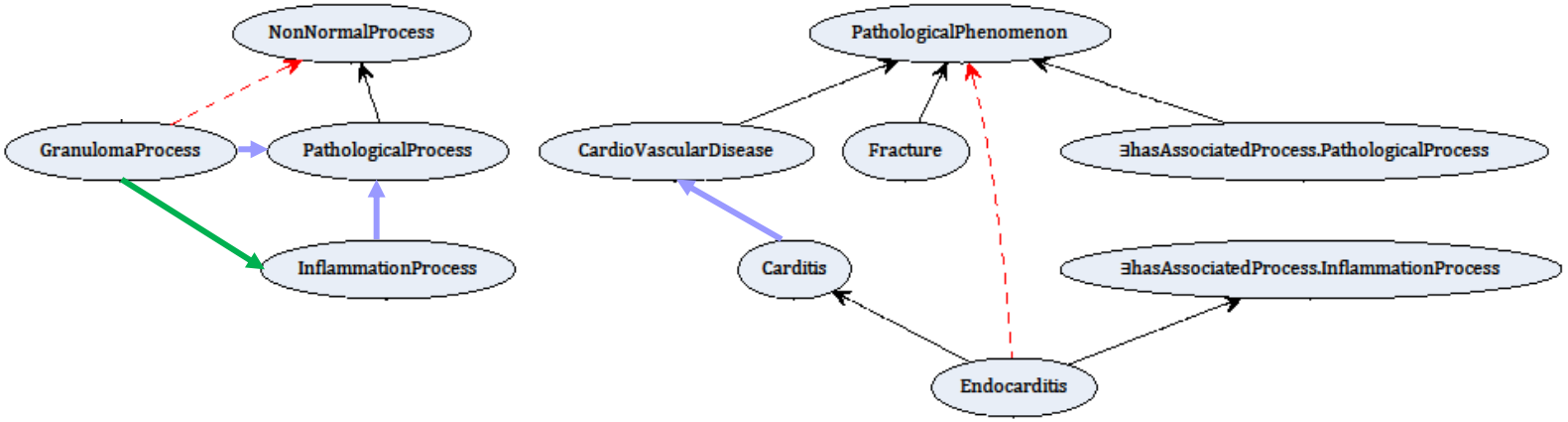
- Combine solutions for individual missing is-a relations
- Remove redundant relations while keeping the same level of completeness
- Resulting solution is a skyline optimal solution

$$\{\text{InflammationProcess} \dot{\sqsubseteq} \text{PathologicalProcess},$$
$$\text{Carditis} \dot{\sqsubseteq} \text{CardioVascularDisease},$$
$$\text{GranulomaProcess} \dot{\sqsubseteq} \text{PathologicalProcess}\}$$

# Algorithm – improving solution

- Solution S from previous step may contain relations which are not derivable from the ontology.
- These can be seen as new missing is-a relations.
- We can solve a new GTAP problem: GTAP(**T** U *S*, **C**, Or, *S*)

# Example – improving solutions




$$\text{GranulomaProcess} \dot{\sqsubseteq} \text{InflammationProcess}$$

$$\{\text{InflammationProcess} \dot{\sqsubseteq} \text{PathologicalProcess},$$
$$\text{Carditis} \dot{\sqsubseteq} \text{CardioVascularDisease},$$
$$\text{GranulomaProcess} \dot{\sqsubseteq} \text{InflammationProcess}\}$$

# Algorithm properties

- Sound
- Skyline optimal solutions

# Experiments

Two use-cases

- Case 1: given missing is-a relations
  AMA and a fragment of NCI-A ontology – OAEI 2013
  - AMA (2744 concepts) – 94 missing is-a relations
    → 3 iterations, 101 in repairing (47 additional new knowledge)
  - NCI-A (3304 concepts) – 58 missing is-a relations
    → 3 iterations, 54 in repairing (10 additional new knowledge)

- Case 2: no given missing is-a relations
  Modified BioTop ontology
  - Biotop (280 concepts, 42 object properties)
    randomly choose is-a relations and remove them: 47 ‘missing’
    → 4 iterations, 41 in repairing (40 additional new knowledge)

# Further reading

Starting points for further studies

# Further reading ontology debugging

Debugging and Completing Ontologies

- Lambrix P, Completing and Debugging Ontologies: state of the art and challenges in repairing ontologies. *Journal of Data and Information Quality*, 15(4):41, 2023.

Debugging Ontologies

- Schlobach S, Cornet R. Non-Standard Reasoning Services for the Debugging of Description Logic Terminologies. *18th International Joint Conference on Artificial Intelligence - IJCAI03*, 355-362, 2003.
- Schlobach S. Debugging and Semantic Clarification by Pinpointing. *2nd European Semantic Web Conference - ESWC05*, LNCS 3532, 226-240, 2005.

# Further reading ontology debugging

Completing ontologies

- Fang Wei-Kleiner, Zlatan Dragisic, Patrick Lambrix. Abduction Framework for Repairing Incomplete EL Ontologies: Complexity Results and Algorithms. 28th AAAI Conference on Artificial Intelligence - AAAI 2014, 1120-1127, 2014.

- Lambrix P, Ivanova V, A unified approach for debugging is-a structure and mappings in networked taxonomies, *Journal of Biomedical Semantics* 4:10, 2013.