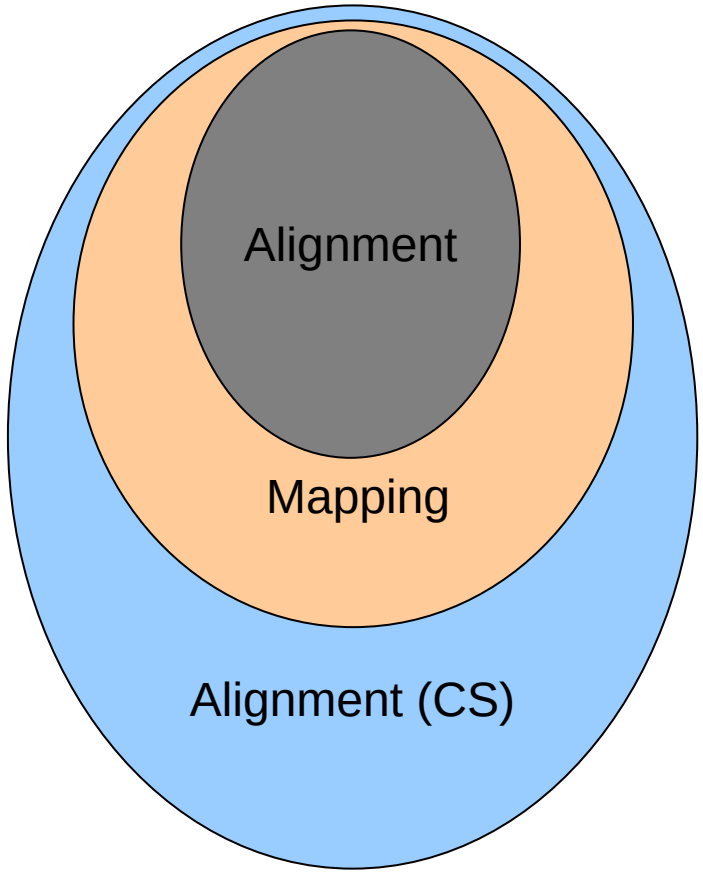
Data driven Ontology Alignment

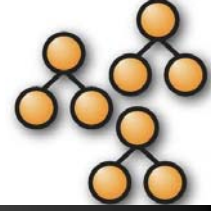
Nigam Shah
nigam@stanford.edu

What is Ontology Alignment?

- Alignment = the identification of **near synonymy relationship** b/w terms from different ontologies.
- Mapping = the identification of ***some* relationship** b/w terms from different ontologies.
- Alignment (CS) = the process of detecting potential mappings



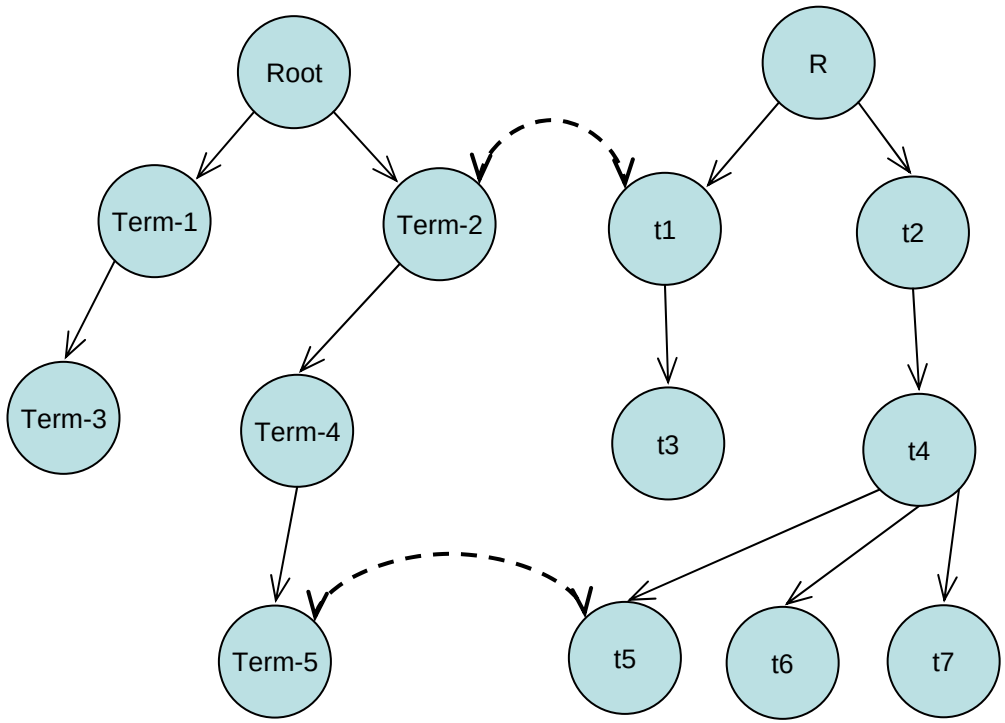
Approaches to alignment



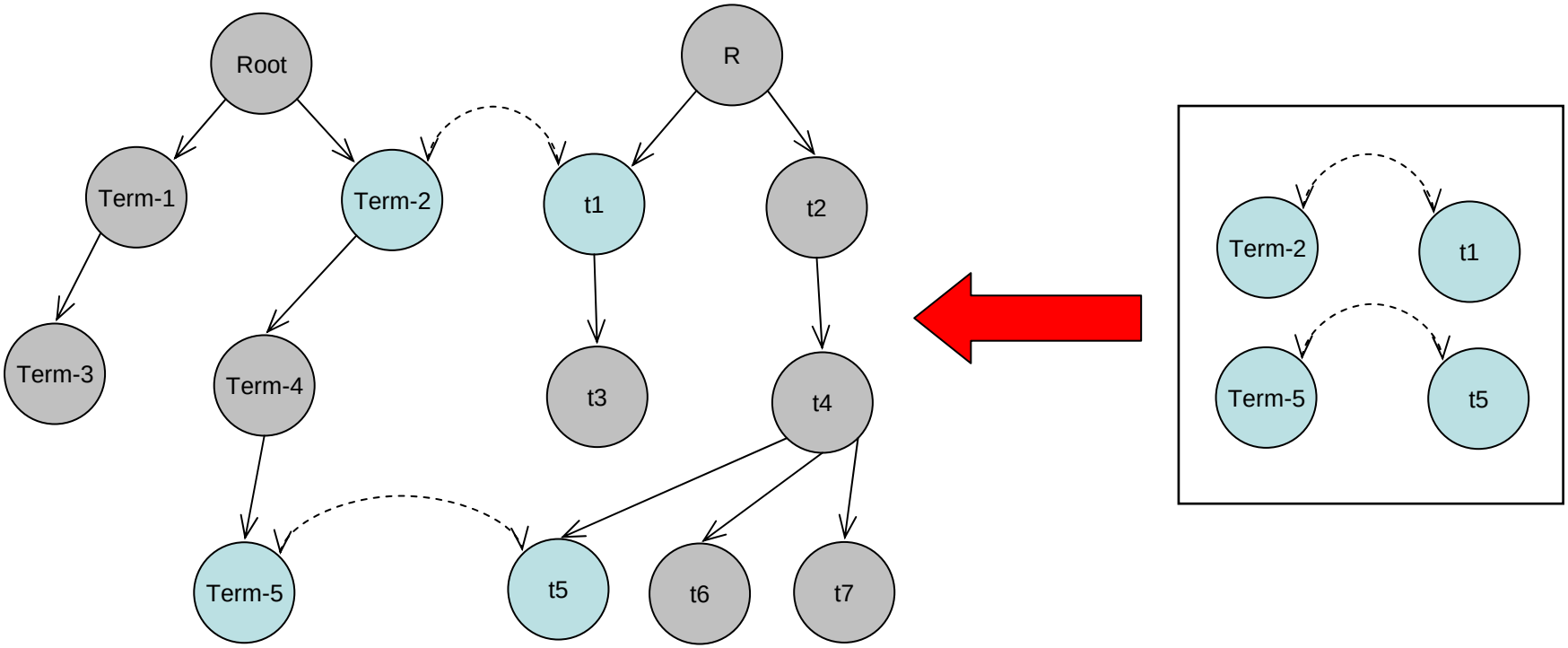
- Pre-defined, during the process of creation of the ontology…
  - The OBO Foundry paradigm (http://obofoundry.org)
  - Authors discuss, argue, vote and reach a consensus
  - Takes a long time!
- Post-hoc, after the relevant ontologies have been in use for some time
  - Human curated → does not scale
  - Algorithm driven (PROMPT, FOAM …)
  - Data driven (which we discuss today)

Steps in Alignment (CS)

- Anchor identification
  - Identify similar class labels in the ontologies to be aligned
  - Usually done by string matching
- Ontology structure
  - Use the “similar” classes as anchors and examine the local [graph] structure around them to inform the “similarity” metric



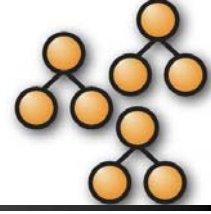
How can the annotated data help?



Ontology [graph] structure based step

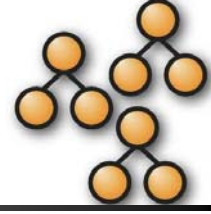
Provide Anchors from annotated data

Annotated data (biomedical)



- Annotation = A statement declaring a relationship b/w a biomedical *thing* and a term [class name] (or an instance of a class) from an ontology.
  - e.g. p53 <associated_with> cell death
- Annotations tell us what the biologists *believe* to be true (in particular or in general)
  - Most annotations are created after particular observations and then are generalized during *interpretation* by a biologist.
- Annotations of clinical / medical data are usually NOT generalized but remain at the particular (or instance) level.

Example annotated data set

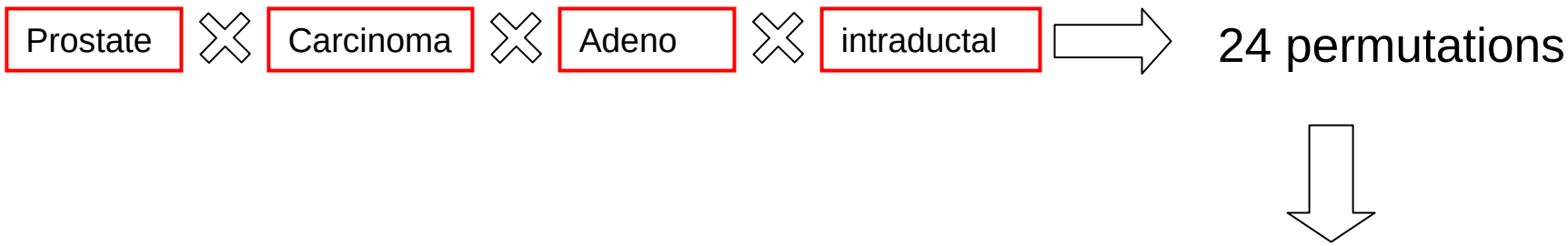


- Each donor block in the TMA has semi-structured text associated with it.

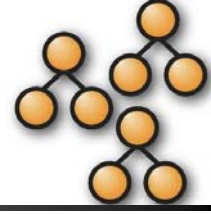
ID	Organ	Diagnosis	Subclass 1	Subclass 2	Subclass 3	Subclass 4
2334	Ovary	MMMT				
3335	Prostate	Carcinoma	Adeno	intraductal		
7022	Bladder	Carcinoma	Transitional cell	In situ		
7288	Testis	teratoma	immature	Embryonal carcinoma		
8060	Liver	Carcinoma	hepatocellular	No vascular invasion	HepC cirrhosis	
6662	Soft tissue	Sarcoma	Leiomyo	epithelioid		
6663	lung	Sarcoma	Leiomyo	epithelioid		
4713	stomach	carcinoma	unknown			

Map text to ontology terms

- Make all possible permutations
  - Rules to weed out bad permutations
- Check for an exact match with NCI and SNOMED-CT terms (and/or synonyms)
  - Rules to weed out bad matches

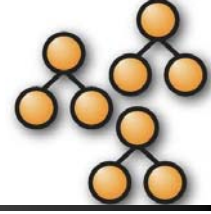


Sample matches



	Organ	Diagnosis	Subclass 1	Subclass 2	Subclass 3	Ontology Terms
2334	Ovary	MMMT				Malignant_Mixed_Mesodermal_Mullerian_Tumor
3335	Prostate	Carcinoma	Adeno	intraductal		Prostate_Ductal_Adenocarcinoma
7022	Bladder	Carcinoma	Transitional cell	In situ		Stage_0_Transitional_Cell_Carcinoma Transitional_Cell_Carcinoma Bladder_Carcinoma Carcinoma_in_situ
7288	Testis	teratoma	immature	Embryonal carcinoma		Immature|Teratoma Testicular_Embryonal_Carcinoma Immature_Teratoma
8060	Liver	Carcinoma	hepatocellular	No vascular invasion	HepC cirrhosis	Hepatocellular_Carcinoma
6662	Soft tissue	Sarcoma	Leiomyo	epithelioid		Soft_Tissue_Sarcoma Leiomyosarcoma Epithelioid_Sarcoma
6663	lung	Sarcoma	Leiomyo	epithelioid		Lung_Sarcoma Leiomyosarcoma Epithelioid_Sarcoma
4713	stomach	carcinoma	unknown			Gastric_carcinoma

Some boring results (and validation)…



- Mapped the term-sets for 8495 records, which correspond to 783 distinct term-sets.
  - 577 term-sets (6614 records) matched to the NCI thesaurus
  - 365 term-sets (3465 records) matched to SNOMED-CT
- In total mapped 6871 records (80%) of annotated records in TMAD (641 distinct term-sets) to one or more ontology terms.

Validation	NCI		SNOMED-CT	
	Appropriate	Inappropriate	Appropriate	Inappropriate
Set-1	41	9	41	9
Set-2	42	8	43	7
Set-3	46	4	38	12
Total	129	21	122	28
Average (%)	**43.0 (86%)**	**7.0 (14%)**	**40.66 (81%)**	**9.33 (19%)**

Context for the project

Data Analysis

By Array Block
By Diagnosis
By Marker (IHC and/or ISH
(Updated - Dec 2005) Explore Tissue Data

Search Tables

Database Users
Donor Blocks
Array Blocks
Cores
Stains
Antibodies
Spot Images
By NCI Thesaurus

Enter New Records

Enter New Donor Block
Enter New Array Block
Enter New Antibody

Administrative Functions

- Change Your Password

Upload Existing Excel Data

- Batch Upload Tissue Array List
- Batch Upload Master workbooks
- Batch Upload TMADonors spreadsheet
- Load images from loader.stanford.edu
- Statistics
- Cross reference of Arrays, Sectors, Images and Slides
- (Experimental) grid Fluorescence tissue array

New Terms added:

Enter (type or paste) NCI term/s to start:
[peritoneal neoplasm is filled in as an example]

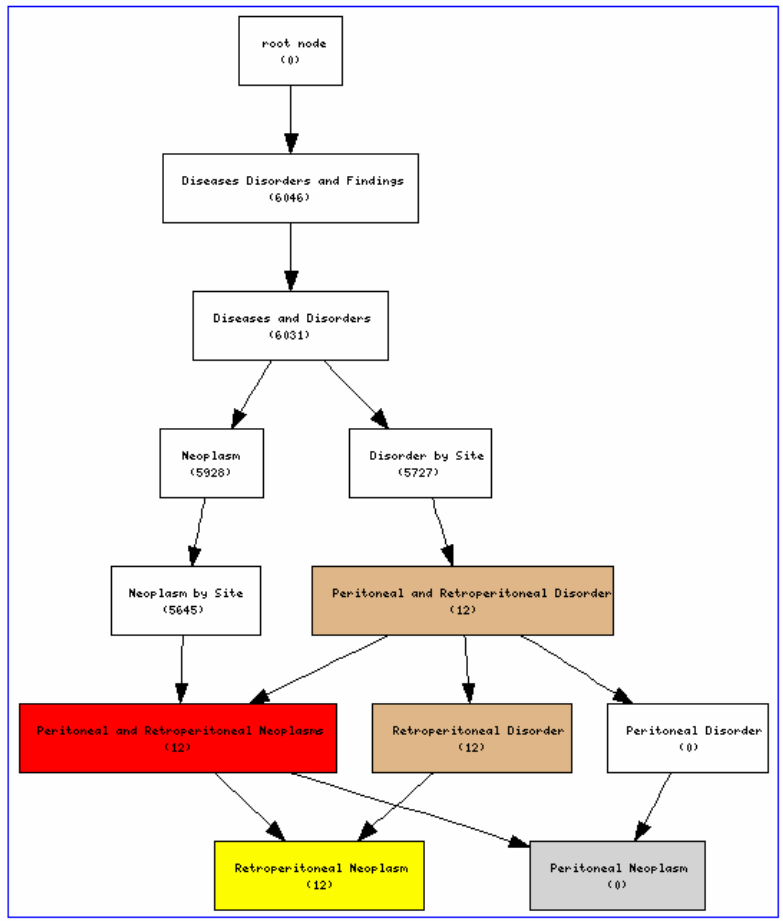
peritoneal neoplasm

Submit Terms

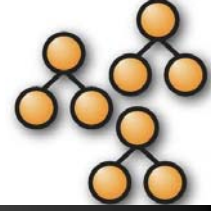
Query Ontology [Start over]
connected to nigam-tma.stanford.edu at Wed Apr 5 11:36:27 2006, running as NIGAM

Past Terms:
New Terms added:

Anchor Term: C7337
Children of Anchor: Peritoneal_Neoplasm Retroperitoneal_Neoplasm



Click on the “Red Node” link to get data



Query TMA

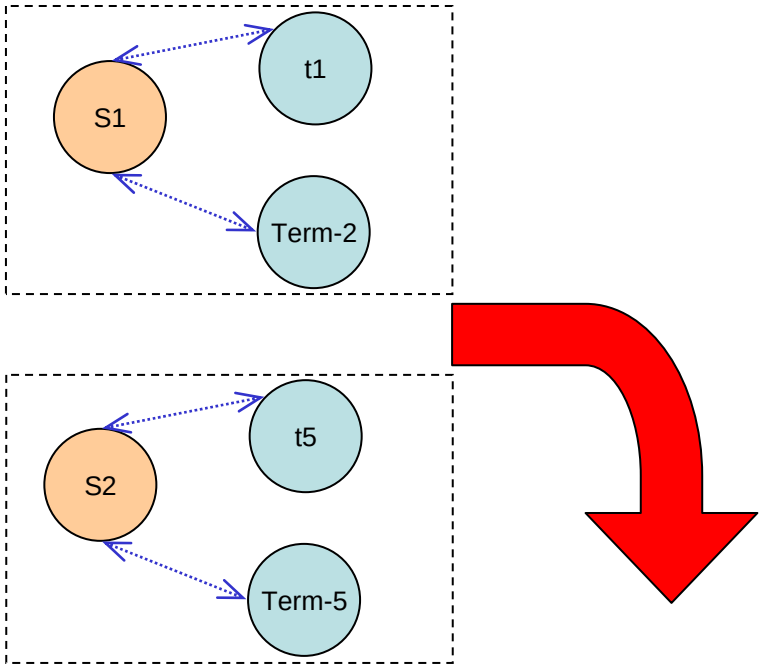
connected to DBI:mysql:tma2 on smi-protege.stanford.edu at Tue Apr 4 19:40:54 2006

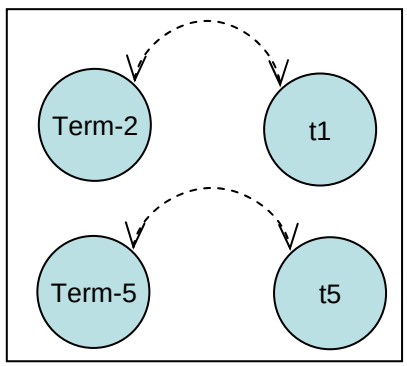
Anchor Term: C7337

1306,	Endocrine,	adrenal,	unknown,	pheochromocytoma,	Pheochromocytoma,
1276,	Endocrine,	adrenal,	unknown,	pheochromocytoma,	Pheochromocytoma,
1264,	Endocrine,	adrenal,	unknown,	carcinoma adrenocortical,	Adrenal_Cortex_Carcinoma,
1252,	Endocrine,	adrenal,	unknown,	pheochromocytoma,	Pheochromocytoma,
1253,	Endocrine,	adrenal,	unknown,	pheochromocytoma,	Pheochromocytoma,
1232,	Unknown,	unknown,	unknown,	carcinoma adrenocortical,	Adrenal_Cortex_Carcinoma,
1161,	Endocrine,	adrenal,	unknown,	pheochromocytoma signet ring,	Pheochromocytoma,
1021,	Endocrine,	adrenal,	unknown,	carcinoma adrenocortical,	Adrenal_Cortex_Carcinoma,
1008,	Endocrine,	adrenal,	unknown,	carcinoma adrenocortical,	Adrenal_Cortex_Carcinoma,
982,	Endocrine,	adrenal,	unknown,	pheochromocytoma,	Pheochromocytoma,
985,	Endocrine,	adrenal,	unknown,	pheochromocytoma,	Pheochromocytoma,
694,	Urinary,	bladder,	unknown,	pheochromocytoma,	Pheochromocytoma,

[Back to TMA Browser]

Annotations performed using multiple ontologies are the key…

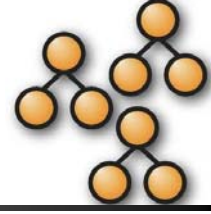




- The *relationship* [blue arrows] embodied in this annotation is fuzzy… but that's life.

- However, (depending on the data) this gives a way to say:
  - Term-2 <is synonymous to> t1
  - Term-5 <is synonymous to> t5

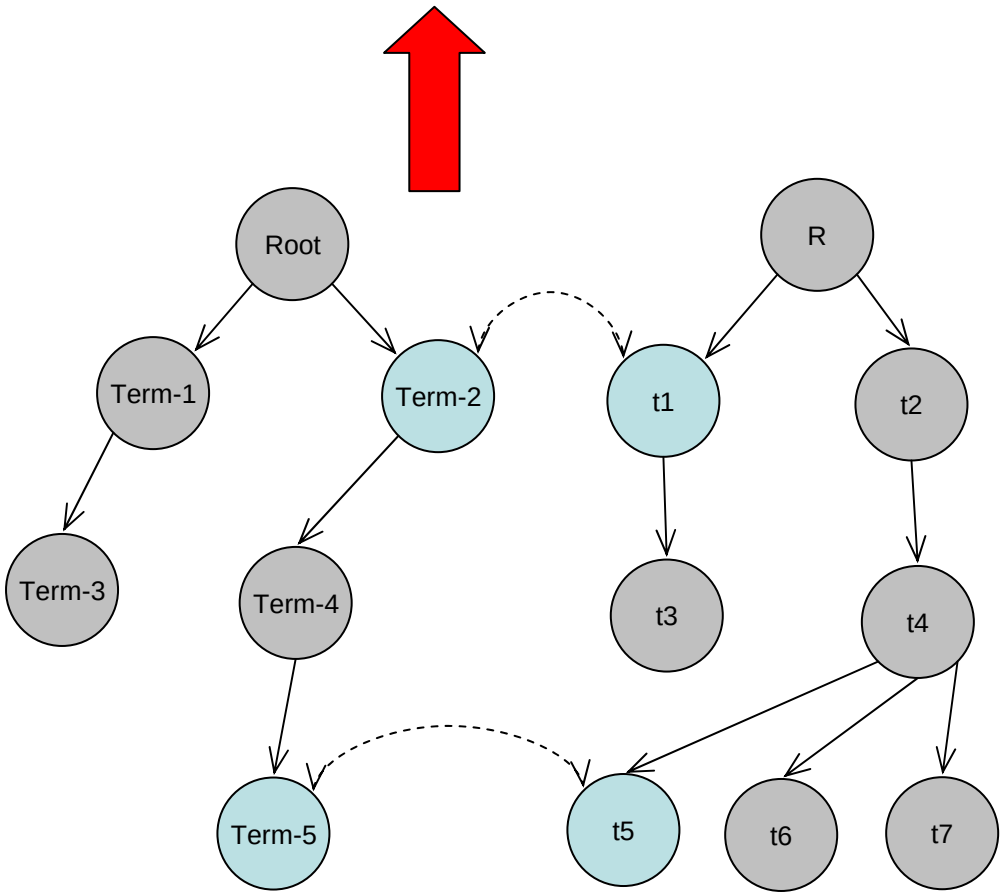
How good are the anchors?

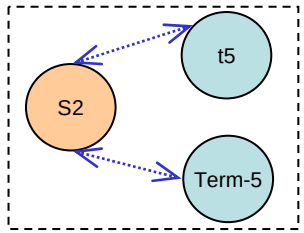


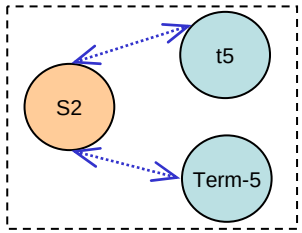
- **Strategy: Evaluate against a manually defined gold standard [UMLS]**
  - Find the CUI of the NCI-term (Nt) from the UMLS.
  - Find the CUI of the SNOMED-CT term (St) from the UMLS
  - Examine if the CUIs are the same or within two links of each other
- **Results: The CUIs were**
  - identical for 2335 records
  - at one link from each other for 403 records
  - at two links from each other for 189 records.
- Overall, Nt – St pairs from **2927 records (= 259 distinct terms) were appropriately aligned**. [259 = 88%]
- The CUIs for the Nt – St pairs for 281 records (corresponding to 36 distinct terms), were separated by more than two links.

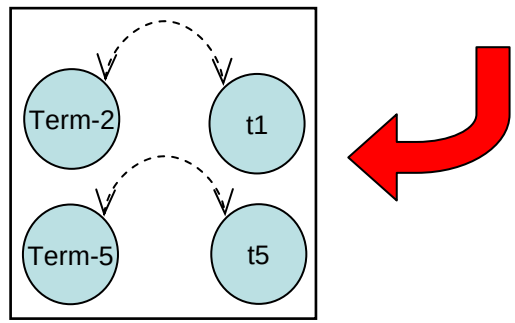
We might improve alignment …

Ontology [graph] structure based step





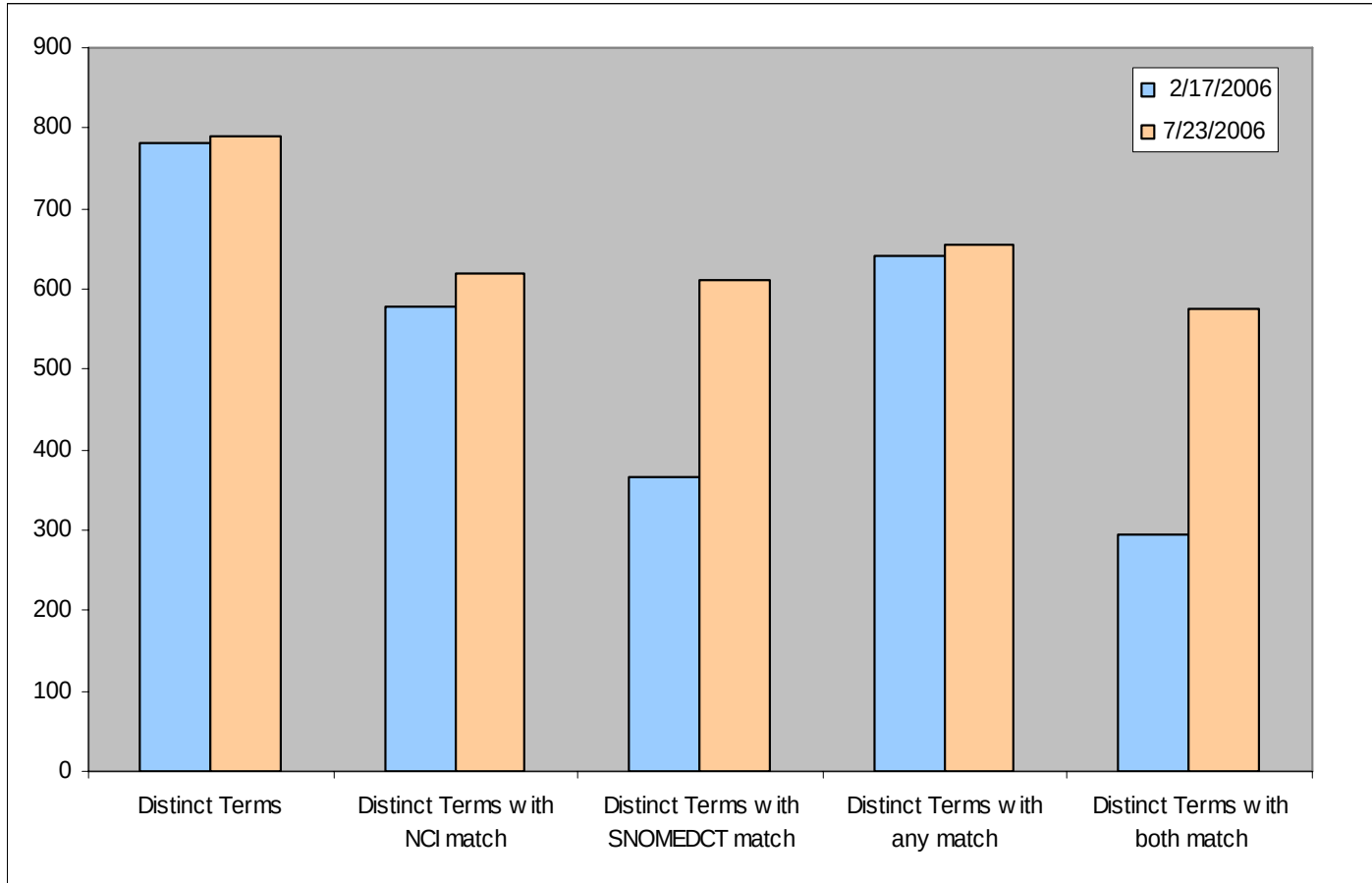




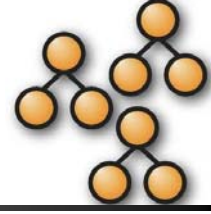
Provide Anchors from annotated data

Better Text-mapping → Better Alignment

2/17	7/23	
783	791	Distinct Terms
577	620	Terms with NCI match
365	610	Terms with SNOMEDCT match
641	654	Terms with any match
295	**576**	**Terms with both match**

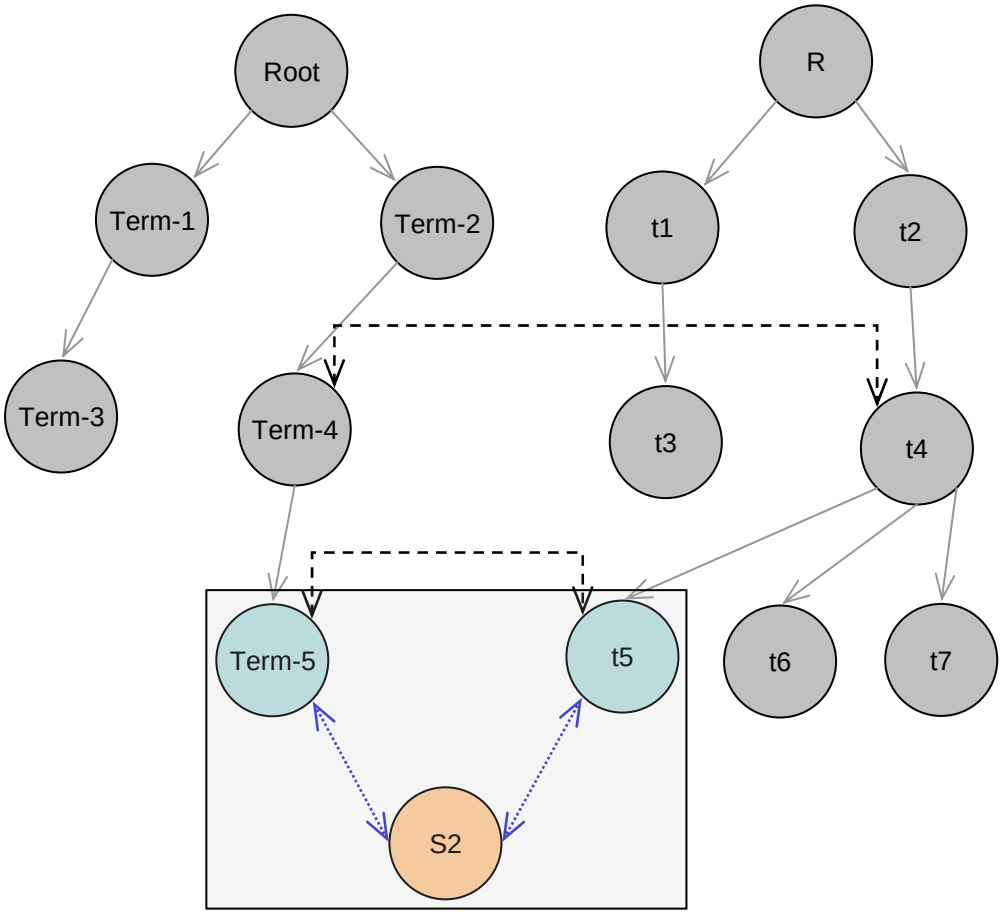


Validation of the [new] alignment



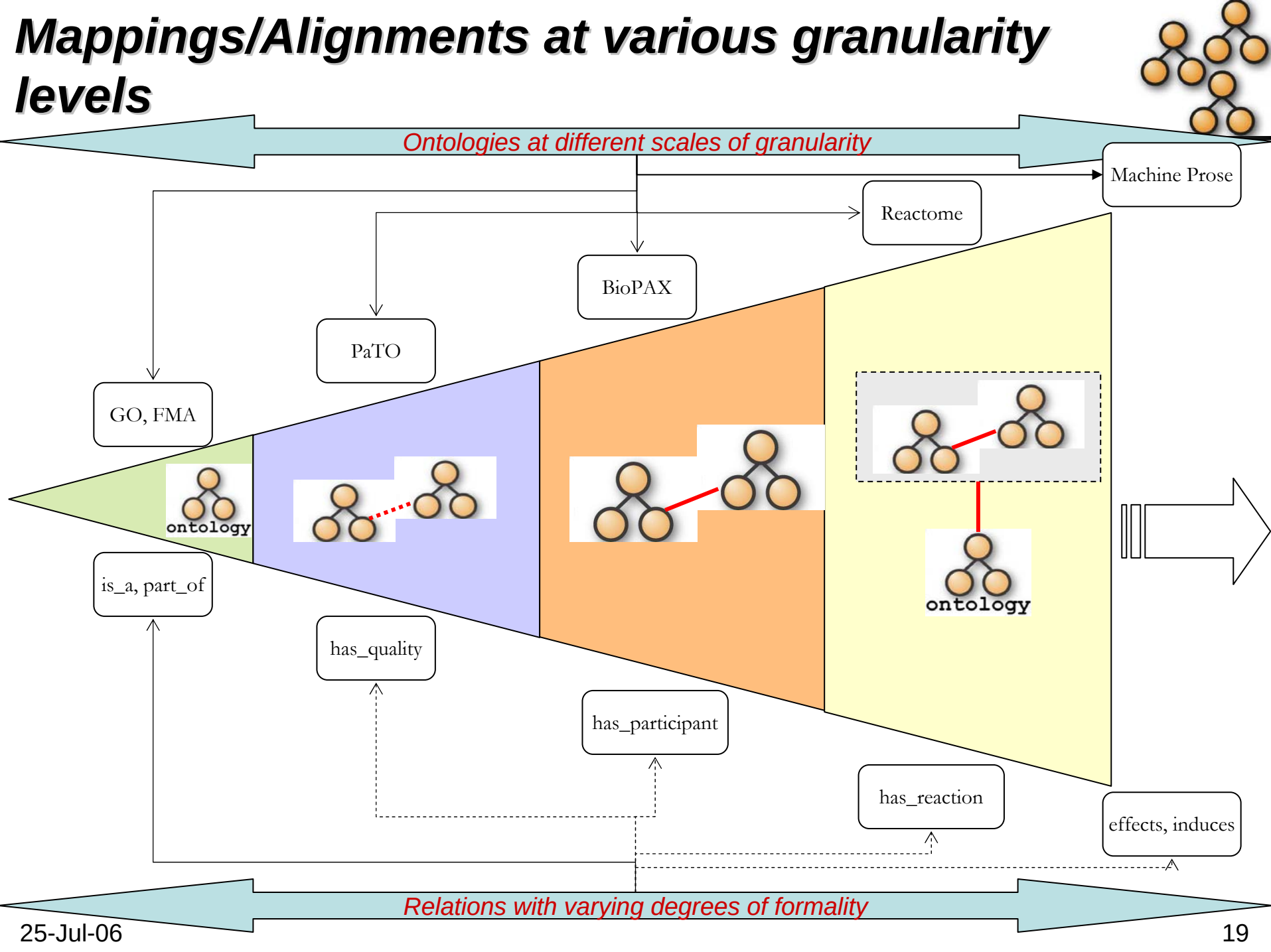
- Identify anchors using [standard] methods for the set of terms aligned using annotated data
  - Run the structural step of the alignment
- Use anchors identified using annotated data
  - Run the structural step using the annotation derived anchors
  - Also looking at indexing for text-mapping [instead of permutation generation] – With Sean Falconer
- Compare the two alignments
  - Either using an expert created gold standard (UMLS)
  - Or by direct review by experts
- We will have results at the next Protégé conference ;)

Use of “more structured” annotations



- If the *relationship* embodied in this annotation is well defined (the blue arrows)
- We might be able to say:
  - Term-5 <*has this relationship with*> t5
- If S2 is an instance of Term-5 and/or t5, we might be able to *propagate* the relationship to the parents of Term-5 and t5 (until we “see” a counter example)

Mappings/Alignments at various granularity levels



Acknowledgements

- Natasha Noy
- Kaustubh Supekar
- Daniel Rubin
- Mark Musen

- National Center for Biomedical Ontology
  www.bioontology.org

- York Sure
- (Tricia d'Entremont)
  - Pictorial Ontology Navigation