

# Measuring the Structural Preservation of Semantic Hierarchy Alignments

**Cliff Joslyn, Patrick Paulson,  
Amanda White  
Sinan al-Saffar**



**Pacific Northwest**  
NATIONAL LABORATORY

**National Security  
Directorate**

Fourth International Workshop on  
Ontology Matching (OM-2009)  
October, 2009

# OUTLINE

---

- The role of taxonomic structure preservation in ontology alignment
- Order theoretical approach to semantic hierarchy alignment
- 2008 Anatomy: lower distance discrepancy only
- Participation in evaluation for OAEI 2009
  - 2009 Benchmark
  - 2009 Anatomy

# SEMANTIC HIERARCHIES

---

**Ubiquitous, universal:** Transitive, non-symmetric links

- Subsumptive, inheritance taxonomies: `is-a`
- Meronomic, compositional: `part-of`
- Implies, precedes, causes

**Claim:** Cores of ontologies, taxonomies, computational lexicons, typing systems

- Comprise the bulk of links in real-world ontological knowledgebases: 80 – 90% of links
- Becoming large:  $10^5 - 10^7$

**Challenge:** Identify/establish technology for key tasks:

- Classification, categorization, clustering
- Induction from source data
- Navigation, anomaly detection
- Visualization
- Link analysis, search, retrieval
- Merger, linkage, interoperability



# STRUCTURAL PRESERVATION IN ONTOLOGY ALIGNMENT

---

- Taxonomy alignment cast as (single- or multi-valued) mapping between two semantic hierarchies
- Relative size, structure, domains, impart different properties:
  - More partial in one direction or other
  - More concentrated in one portion or another
  - Take nodes “close together” in one hierarchy to nodes “far apart” in the other
  - Take nodes in one structural relationship (e.g. parent-child) into another (e.g. sibling)
- We can measure the *smoothness* of an alignment
- *Smooth mapping neither necessary nor sufficient to be a good mapping*
  - Strong correlation of smoothness with  $P, R, F$  promises an *automatic* predictor of alignment quality independent of gold standard
  - *Knowledge of mapping properties useful:*
    - Ontology Aligners:** Homogeneous alignments within one domain
    - Ontology Designers:** Given a determinate alignment against a reference ontology

# ORDER THEORETICAL SEMANTIC HIERARCHY ALIGNMENT

**Hierarchies:** Finite, bounded, partially ordered sets (posets)  $\mathcal{P} = \langle P, \leq \rangle$

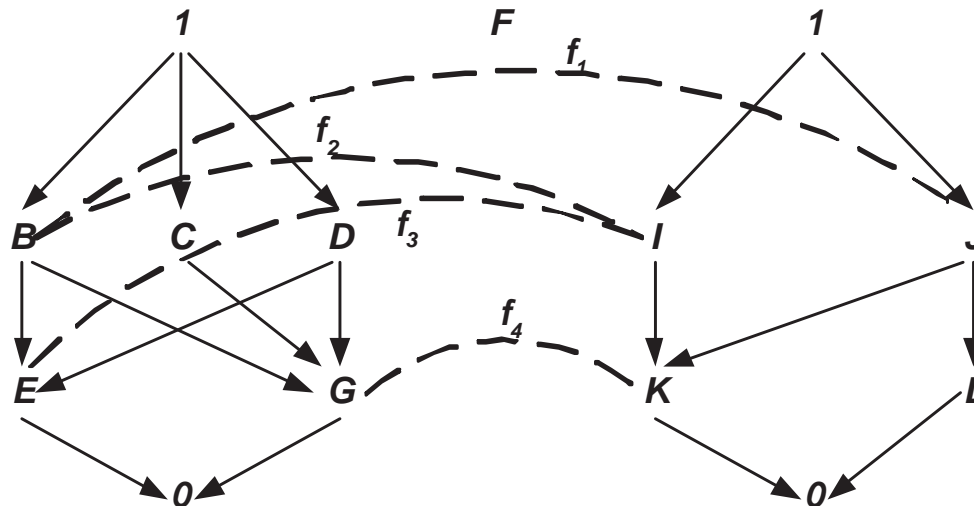
**Alignment Relation:**  $F \subseteq P \times P'$  is a collection of pairs  $\vec{f} = \langle a, a' \rangle \in F$ , for  $\mathcal{P} := \langle P, \leq \rangle, \mathcal{P}' := \langle P', \leq' \rangle$

**Links, Anchors:**  $\vec{f} \in F$

$$Q := \{a \in P, \exists a' \in P', \langle a, a' \rangle \in F\} \subseteq P$$

$$Q' := \{a' \in P', \exists a \in P, \langle a, a' \rangle \in F\} \subseteq P',$$

**Counts:**  $m := |Q|, m' := |Q'|$ , and  $N := |F| \leq mm'$ .



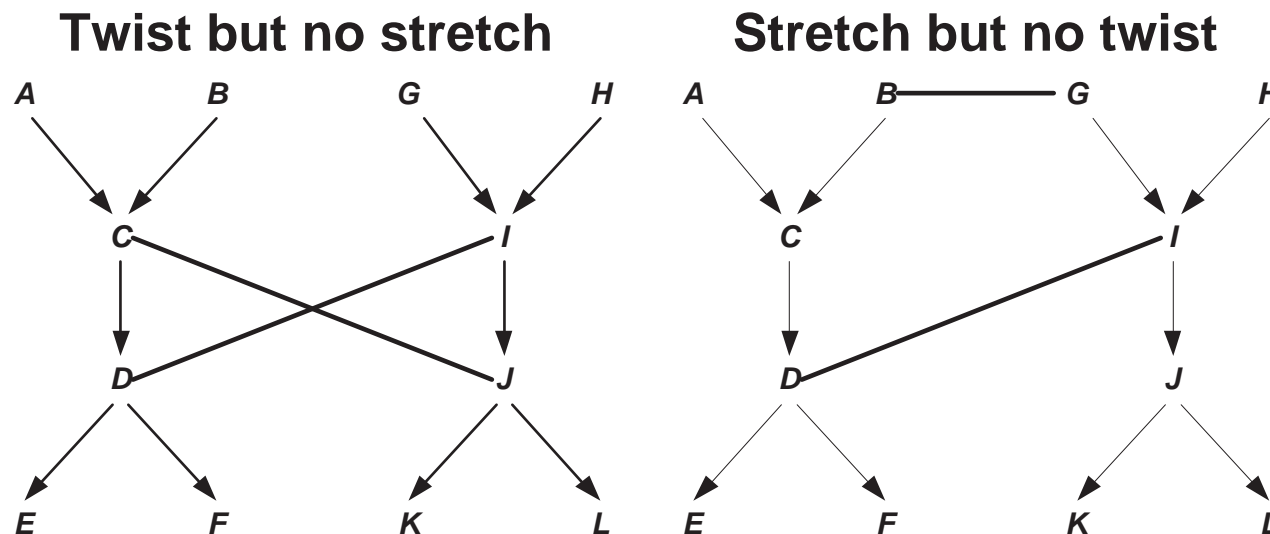
# STRUCTURAL PRESERVATION CRITERIA

**Approach:** Not *relative* measure of alignment to gold standard  
Rather an *inherent* or *independent* measure of the alignment

**Twist, Order Discrepancy:**  $a, b$  should have the same structural relations in  $\mathcal{P}$  as  $a', b'$  in  $\mathcal{P}'$

- **Comparable:** Ancestor-successor, parent-child
- **Non-comparable:** Sibling, cousin, etc.

**Stretch, Distance Discrepancy:** Relative distance between  $a, b \in P$  should be the same as  $a', b' \in P'$



# SIMPLE ORDER DISTANCE

Not path-counting

Upset (Ancestors), Downset (Descendants):

$$\uparrow a := \{b \geq a\}, \quad \downarrow a := \{b \leq a\}$$

Upper and Lower Bounds: For a set of nodes  $Q \subseteq P$

$$\text{Min}(Q) := \{a \in Q : \nexists b \in Q, b < a\} \subseteq P,$$

$$\text{Max}(Q) := \{a \in Q : \nexists b \in Q, b > a\} \subseteq P.$$

Generalized Join and Meet:

$$a \vee b := \text{Min}(\uparrow a \cap \uparrow b) \subseteq P$$

$$a \wedge b := \text{Max}(\downarrow a \cap \downarrow b) \subseteq P,$$

Upper Distance:

$$d_u(a, b) = |\uparrow a| + |\uparrow b| - 2 \max_{c \in a \vee b} |\uparrow c|$$

Lower Distance:

$$d_l(a, b) = |\downarrow a| + |\downarrow b| - 2 \max_{c \in a \wedge b} |\downarrow c|$$

Example:

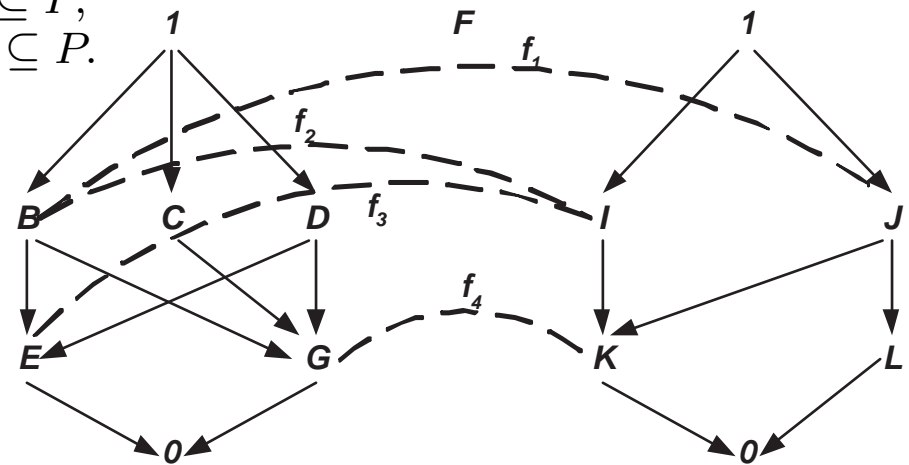
$$\uparrow E = \{1, B, D, E\},$$

$$\uparrow E \cap \uparrow G = \{1, B, D\}$$

$$E \vee G = \{B, D\}$$

$$d_u(E, G) = 4 + 5 - 2 \text{Max}(|\uparrow B|, |\uparrow D|) = 5$$

$$d_l(E, G) = 2 + 2 - 2 \cdot 1 = 2$$



# DISTANCE DISCREPANCY

---

## Normalized Distance:

$$\bar{d}_u(a, b) := \frac{d_u(a, b)}{|P| - 1} \in [0, 1], \quad \bar{d}_l(a, b) := \frac{d_l(a, b)}{|P| - 1} \in [0, 1].$$

**Distance Discrepancy:** Relative distance between  $a, b \in P$  should be the same as  $a', b' \in P'$ :

$$D_{F,d} := \frac{\sum_{\vec{f}, \vec{g} \in F} |\bar{d}(a, b) - \bar{d}'(a', b')|}{\binom{|F|}{2}}$$

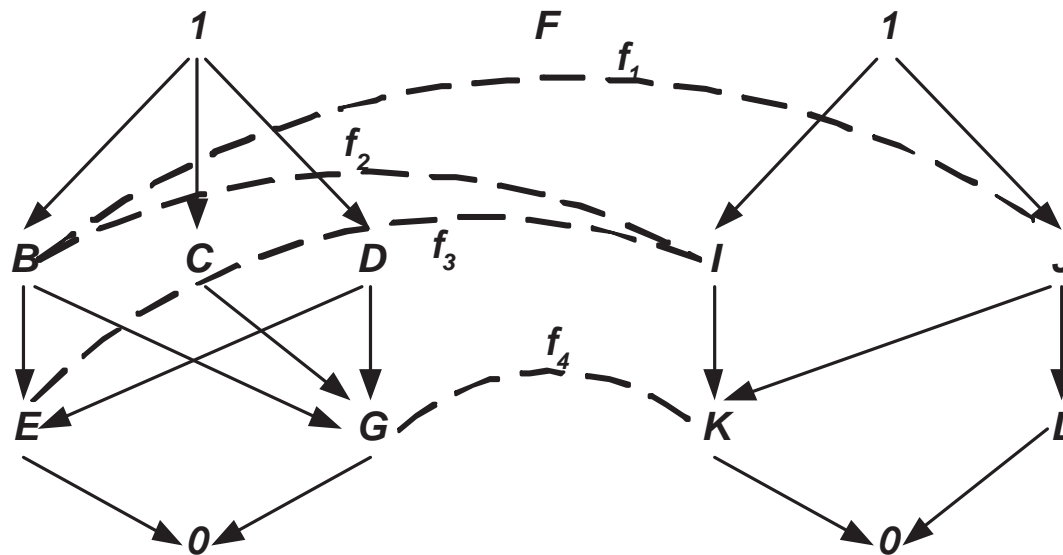
**Upper Vs. Lower Distance:** Lower distance preferable for top-bounded, downward-branching structures, especially if lacking multiple inheritance



# EXAMPLE

	$\vec{f}_1 = \langle B, J \rangle$	$\vec{f}_2 = \langle B, I \rangle$	$\vec{f}_3 = \langle E, I \rangle$	$\vec{f}_4 = \langle G, K \rangle$
$\vec{f}_1 = \langle B, J \rangle$	0.00	0.60	0.27	0.07
$\vec{f}_2 = \langle B, I \rangle$	0.60	0.00	0.33	0.13
$\vec{f}_3 = \langle E, I \rangle$	0.27	0.33	0.00	0.13
$\vec{f}_4 = \langle G, K \rangle$	0.07	0.13	0.13	0.00

$$D(F) = \frac{0.6 + 0.27 + 0.07 + 0.33 + 0.13 + 0.13 + 1.53}{6} = .255$$



# RELATION TO SEMANTIC SIMILARITIES

Compare to  $d_l(a, b) = |\downarrow a| + |\downarrow b| - 2 \max_{c \in a \wedge b} |\downarrow c|$

Poset  $\mathcal{P} = \langle P, \leq \rangle$ , probability distribution

$p: P \mapsto [0, 1], \sum_{a \in P} p(a) = 1$ , “cumulative”  $\beta(a) := \sum_{b \leq a} p(b)$

**Resnik:**  $\delta(a, b) = \max_{c \in a \vee b} [-\log_2(\beta(c))]$

**Lin:**

$$\delta(a, b) = \frac{2 \max_{c \in a \vee b} [\log_2(\beta(c))]}{\log_2(\beta(a)) + \log_2(\beta(b))}$$

**Jiang and Conrath:**

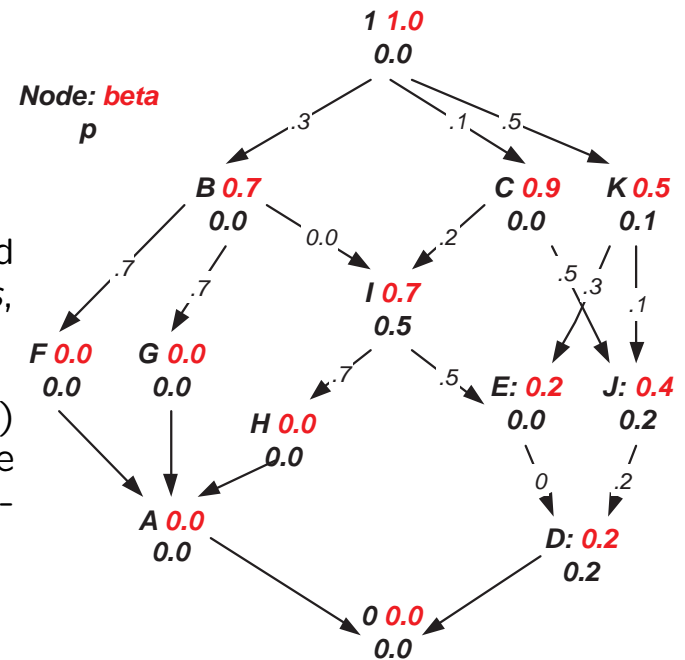
$$\delta(a, b) = 2 \max_{c \in a \vee b} [\log_2(\beta(c))] - \log_2(\beta(a)) - \log_2(\beta(b))$$

**Issues:**

- General mathematical grounding
- Not *rely* on probabilistic weighting

A Butanitsky and G Hirst: (2006) “Evaluating WordNet-based measures of semantic distance.” *Computational Linguistics*, 32(1), 13–47.

Lord, PW; Stevens, Robert; Brass, A; and Goble, C: (2003) “Investigating Semantic Similarity Measures Across the Gene Ontology: the Relationship Between Sequence and Annotation”, *Bioinformatics*, v. 10, pp. 1275-1283



# GROUNDING IN ORDER METRICS

- Semimodular functions on finite, bounded orders  $\langle P, \leq \rangle$
- E.g. generalized Kolmogorov probabilities, information measures in graphical models (Studeny 2005)

**Generalized Join, Meet:**  $a \nabla b := \uparrow a \cap \uparrow b$ ,  $a \Delta b := \downarrow a \cap \downarrow b$

**Isotone Map:**  $v: P \mapsto \mathbb{R}, a \leq b \rightarrow v(a) \leq v(b)$

$$v^\nabla(a, b) := \min_{c \in a \nabla b} v(c), \quad v_\Delta(a, b) := \max_{c \in a \Delta b} v(c)$$

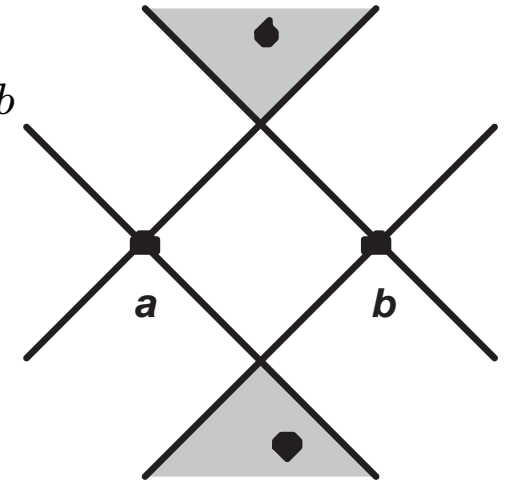
**Antitone Map:**  $v: P \mapsto \mathbb{R}, a \leq b \rightarrow v(a) \geq v(b)$

$$v^\nabla(a, b) := \max_{c \in a \nabla b} v(c), \quad v_\Delta(a, b) := \min_{c \in a \Delta b} v(c)$$

**Valuations:** Super- and sub-modularity:

**Lower:**  $v(a) + v(b) \leq v^\nabla(a, b) + v_\Delta(a, b)$

**Upper:**  $v(a) + v(b) \geq v^\nabla(a, b) + v_\Delta(a, b)$



Metric $d(a, b)$	Lower Valuation $v(a) + v(b) \leq v^\nabla(a, b) + v_\Delta(a, b)$	Upper Valuation $v(a) + v(b) \geq v^\nabla(a, b) + v_\Delta(a, b)$
Isotone	$v(a) + v(b) - 2v_\Delta(a, b)$	$2v^\nabla(a, b) - v(a) - v(b)$
Antitone	$v(a) + v(b) - 2v^\nabla(a, b)$	$2v_\Delta(a, b) - v(a) - v(b)$

Monjardet, B: (1981) "Metrics on Partially Ordered Sets - A Survey", *Discrete Mathematics*, v. 35, pp. 173-184

Orum, Chris and Joslyn, Cliff A: (2009) "Valuations and Metrics on Partially Ordered Sets", <http://arxiv.org/abs/0903.2679v1>, submitted

## 2008 ANATOMY TRACK

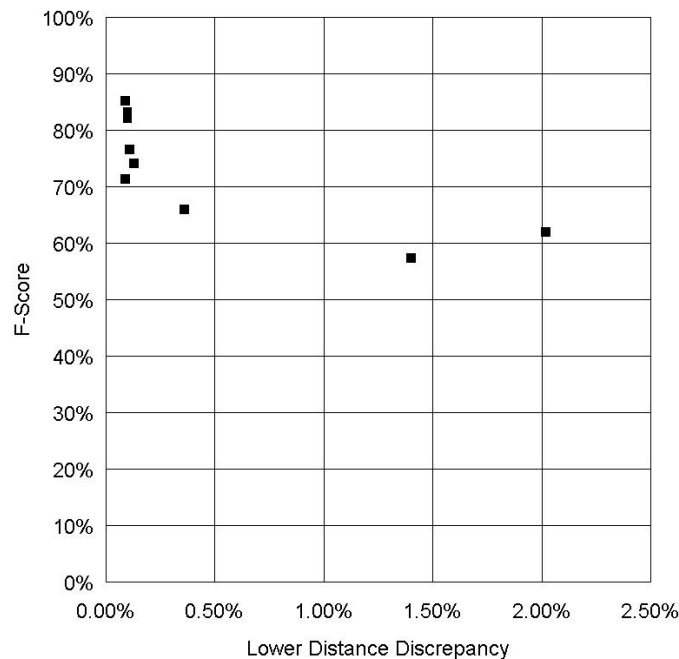
---

- Chosen for sufficient size, multiple inheritance, moderate amount of complexity, availability of reference alignment

	<b>MA</b>	<b>NCIT</b>
# Nodes	2744	3304
# Edges	2863	3761
Edge density	1.04	1.14
Height	8	14
# Leaves	2258	2631
Leaf density	82.3%	79.6%
Average leaf top rank	2.58	5.92
Multiple inheritance	4.0%	13.2%

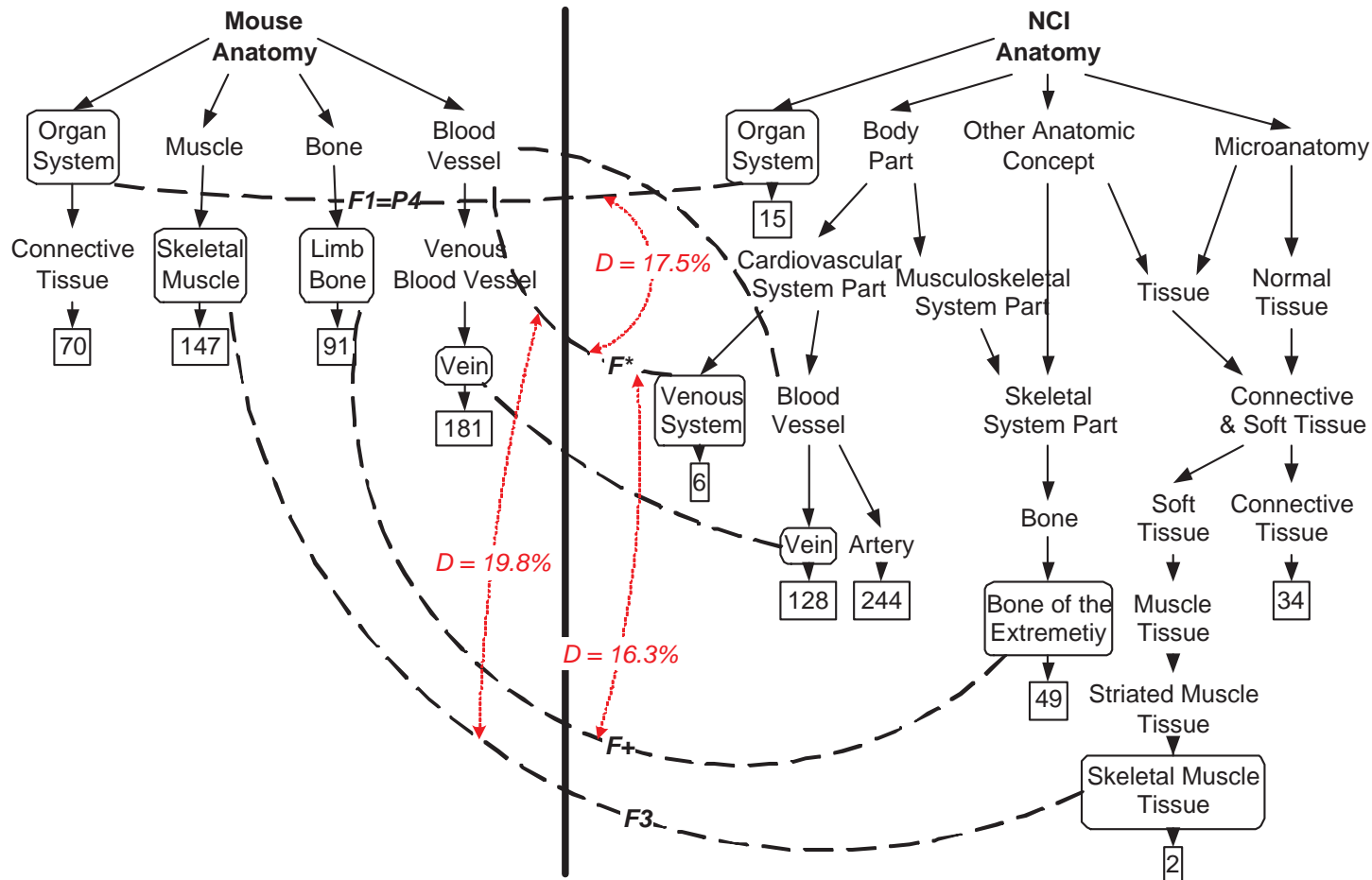
# 2008 ANATOMY TRACK RESULTS

Alignment	$m$	$m'$	$N$	$D(F)$	$P$	$R$	$F$ -score
reference_partial	986	984	988	0.08%			
reference_full	1501	1504	1523	0.11%			
aflood	1186	1186	1186	0.11%	87.4%	68.2%	76.6%
AROMA	1062	1062	1062	0.36%	80.3%	56.0%	66.0%
ASMOV	1261	1261	1261	0.09%	78.7%	65.2%	71.3%
DSSim	1170	1086	1545	2.02%	61.6%	62.4%	62.0%
Lily	1324	1324	1324	0.13%	79.6%	69.3%	74.1%
RiMOM	1205	1205	1205	0.10%	92.9%	73.5%	82.1%
SAMBO	1465	1465	1465	0.09%	86.9%	83.6%	85.2%
SAMBOdtf	1527	1527	1527	0.10%	83.1%	83.3%	83.2%
TaxoMap	2533	1279	2533	1.40%	46.0%	76.4%	57.4%



- Task 1
- $r(D(F), F\text{-score}) = -0.780$

# MOST DISCREPANT LINKS, REFERENCE ALIGNMENT



$\delta$	$\vec{f}$				$\vec{g}$			
		MA	NCIT		MA	NCIT		
19.8%	F*	blood vessel	Venous System	F3	skeletal muscle	Skeletal_Muscle_Tissue		
17.6%	F*	blood vessel	Venous System	F1=P4	organ system	Organ_System		
16.3%	F*	blood vessel	Venous System	F+	limb bone	Bone_of_the_Extremity		

# MOST DISCREPANT SUBMISSIONS

DSSIM			Taxomap		
<i>D</i>	MA	NCIT	<i>D</i>	MA	NCIT
62.82%	joint	Body_Part	11.72%	tail blood vessel	Blood_Vessel
15.66%	cardiovascular system	Cardiovascular_System_Part	11.72%	foot blood vessel	Blood_Vessel
13.02%	capillary	Blood_Vessel	11.72%	neck blood vessel	Blood_Vessel
11.04%	bone	Loose_Connective_Tissue	11.72%	head blood vessel	Blood_Vessel
9.84%	perineal artery	Perineal_Artery	11.72%	lung blood vessel	Blood_Vessel
9.84%	ethmoidal artery	Artery	11.72%	upper leg blood vessel	Blood_Vessel
8.87%	brachial artery	Brachial_Artery_Branch	11.72%	lower leg blood vessel	Blood_Vessel
8.84%	celiac artery	Artery	11.72%	pelvis blood vessel	Blood_Vessel
8.82%	radial artery	Artery	11.72%	abdomen blood vessel	Blood_Vessel

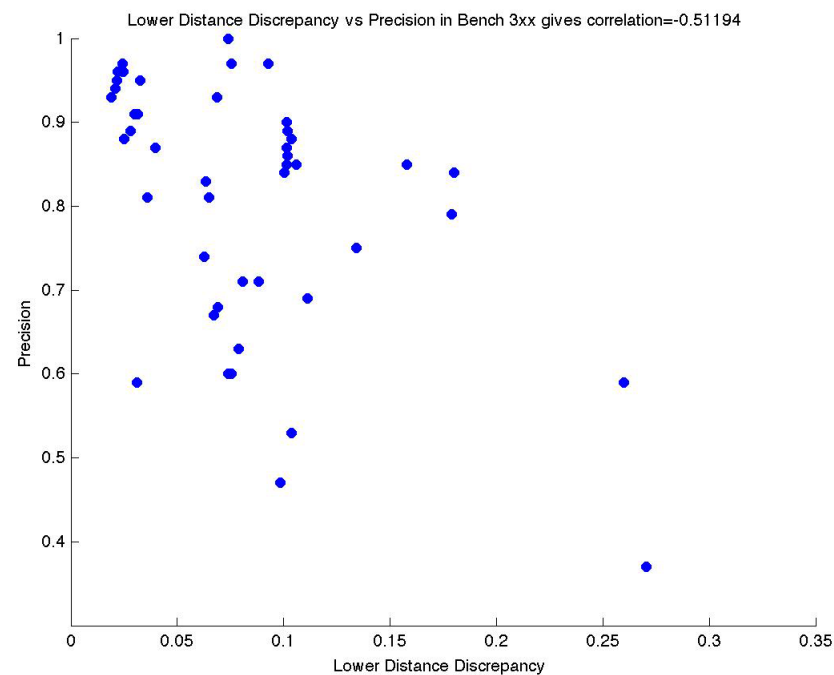
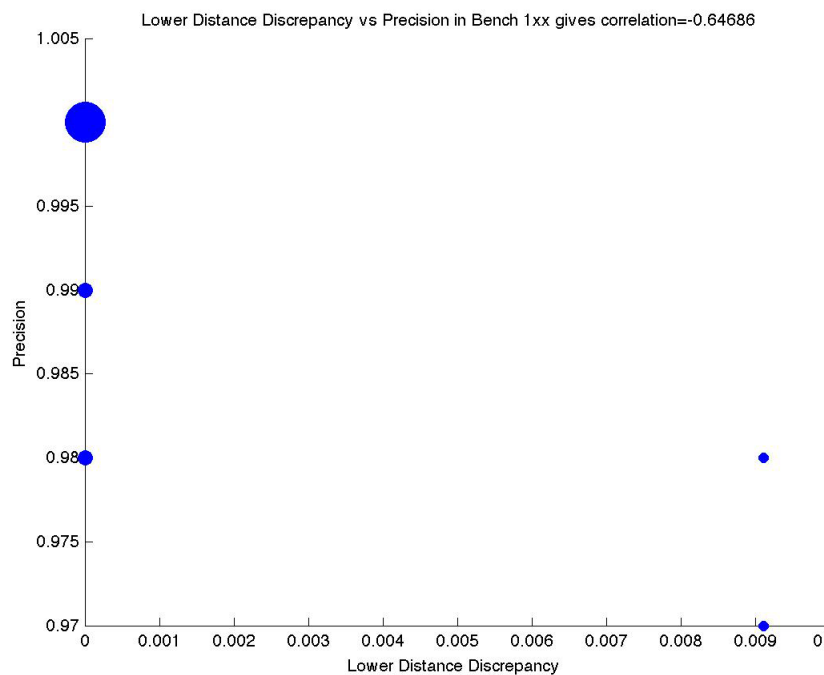
Aggregate link discrepancies point directly to anomolous mappings of high significance:

**DSSIM:** ⟨joint, Body\_Part⟩ most discrepant: Joint covers 21 nodes, Body\_Part covers 2137

**Taxomap:** Many links to NCIT Blood\_Vessel

# 2009 BENCHMARK

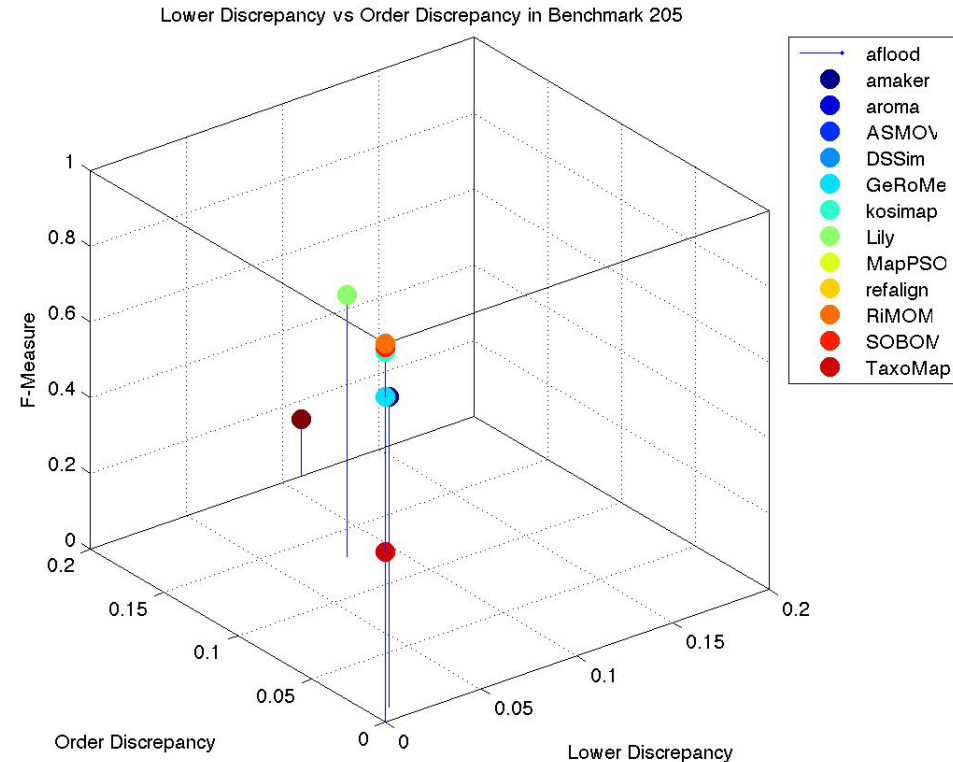
$r$		P	R	F-score
Bench [1-3]*	$D(F)$	-0.39	-0.02	-0.07
	OD			-0.20
Bench 1*	$D(F)$	-0.65	0.03	0.02
	OD	-0.65	0.03	0.02
Bench 2*	$D(F)$	-0.41	-0.13	-0.18
	OD	-0.48	-0.25	-0.29
Bench 3*	$D(F)$	-0.51	-0.48	-0.53
	OD	-0.54	-0.39	-0.46





# BENCHMARK 205

Submitter	$D(F)$	$P$	$R$	$F$
	$r(D(F), \cdot)$	-0.92	-0.73	-0.76
refalign	0.0	1.0	1.0	1.0
MapPSO	0.0	1.0	0.99	0.99
aroma	0.0	1.0	0.99	0.99
ASMOV	0.0	1.0	0.99	0.99
Lily	0.0	1.0	0.99	0.99
RiMOM	0.0	1.0	0.99	0.99
GeRoMe	0.0	1.0	0.97	0.98
amaker	0.0	1.0	0.97	0.98
DSSim	0.0	0.91	0.81	0.86
aflood	0.00801	0.91	0.75	0.82
kosimap	0.08342	0.83	0.59	0.69
SOBOM	0.0	1.0	0.29	0.45
TaxoMap	0.10811	0.53	0.09	0.15



# BENCHMARK 205: REFERENCE VS. KOSIMAP

$D(F) = 0.083, P = .83, R = 0.59, F\text{-score} = 0.69$



Display from AgreementMaker

# 2009 ANATOMY

Test	Submitter	$D(F)$	OD	$P$	$R$	F-score
1	aflood	0.00133	0.00155	0.873	0.653	0.747
1	amaker	0.00127	0.00147	0.865	0.798	0.831
1	aroma	0.00288	0.00298	0.775	0.678	0.723
1	ASMOV	0.00314	0.00368	0.746	0.755	0.751
1	DSSim	0.00156	0.00233	0.853	0.676	0.754
1	kosimap	9.9E-4	0.00123	0.866	0.619	0.722
1	Lily	0.00259	0.00346	0.738	0.739	0.739
1	Ref_Full	7.8E-4	6.6E-4	1.0	1.0	1.0
1	SOBOM	8.8E-4	9.1E-4	0.952	0.777	0.855
1	taxomap	0.00149	0.00225	0.87	0.678	0.762
2	aflood	0.00105	9.8E-4	0.892	0.712	0.792
2	amaker	8.6E-4	8.1E-4	0.967	0.682	0.8
2	ASMOV	0.00133	0.00161	0.821	0.736	0.776
2	DSSim	0.00113	0.00123	0.973	0.62	0.757
2	kosimap	0.0023	0.00443	0.907	0.446	0.598
2	Lily	0.00236	0.00341	0.869	0.559	0.681
2	taxomap	7.5E-4	8.6E-4	0.953	0.609	0.743
3	aflood	0.00148	0.0016	0.827	0.763	0.794
3	amaker	0.00332	0.00368	0.511	0.815	0.628
3	ASMOV	0.00306	0.00386	0.725	0.767	0.745
3	kosimap	9.9E-4	0.00123	0.866	0.619	0.722
3	Lily	0.00332	0.00393	0.534	0.774	0.632
3	taxomap	0.01486	0.02115	0.458	0.716	0.559
4	aflood	0.00145	0.00155			
4	amaker	7.7E-4	6.6E-4			
4	ASMOV	0.00373	0.0041			
4	taxomap	0.00474	0.00748			



# 2009 ANATOMY

<i>r</i>		P	R	F-score
Anatomy [1-3]	$D(F)$	-0.73	0.04	-0.59
	OD	-0.69	-0.03	-0.61
Anatomy 1	$D(F)$	-0.91	-0.25	-0.55
	OD	-0.94	-0.35	-0.64
Anatomy 2	$D(F)$	-0.47	-0.71	-0.85
	OD	-0.36	-0.82	-0.94
Anatomy 3	$D(F)$	-0.68	-0.03	-0.76
	OD	-0.65	-0.07	-0.74

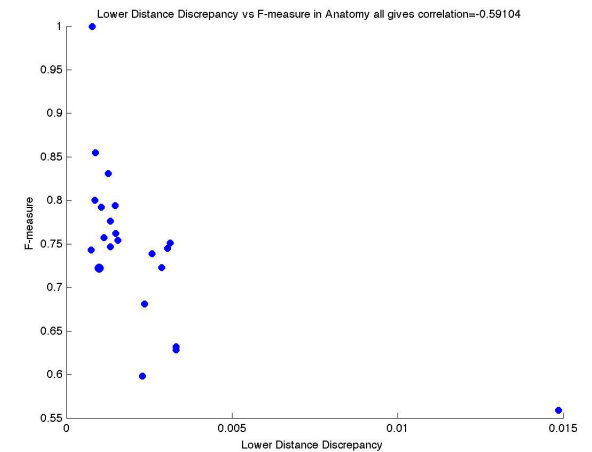
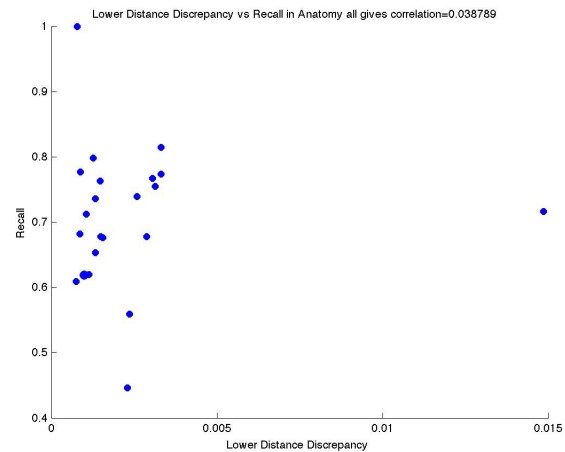
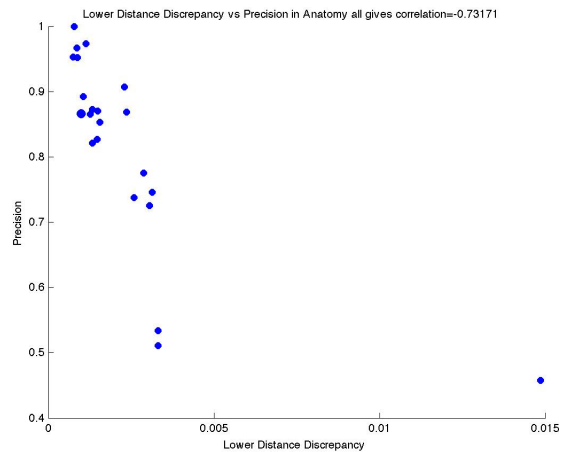
- Outlier: Taxomap, excluding:

$$r(D(F), P) = -0.84$$

$$r(D(F), R) = 0.05$$

$$r(D(F), \text{F-score}) = -0.61$$

- Scaling issues remain



# THANKS

---

## Reviewers

## Organizers

## PNNL:

- Alan Chappell
- Eric Marshall
- Joshua Short

## Universität Mannheim:

- Christian Meilicke

## The Jackson Laboratory:

- Martin Ringwald
- Terry Hayamizu

## U Illinois Chicago:

- Isabel Cruz
- Flavio Antonelli