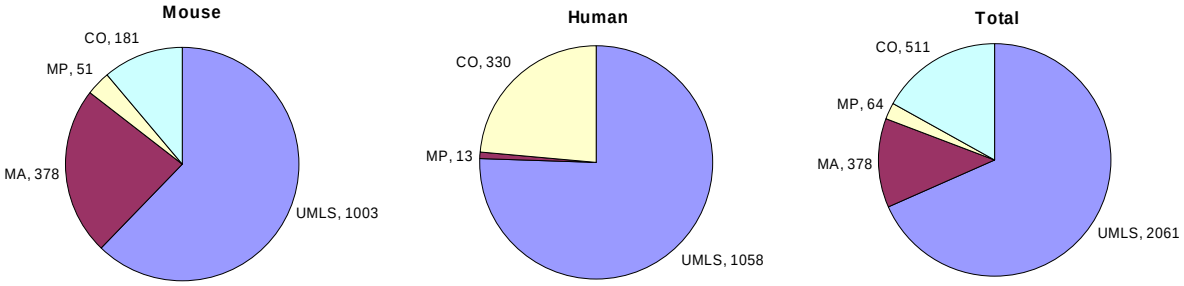
Detailed PhenoGO Statistics and Evaluation per type of mappings

[Natural Language Processing (NLP)]
[Medical Subject Headings (MeSH)]

Natural Language Processing (NLP) Statistics and Evaluation

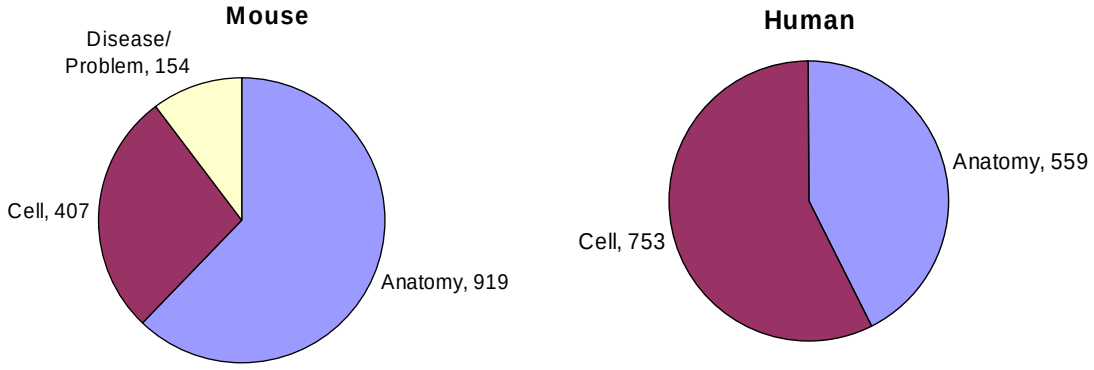
1. Distribution of **phenotypic context codes** per terminology
   a. Unified Medical Language System (UMLS)
   b. Adult Mouse Anatomy (MA)
   c. Mammalian Phenotypes (MP)
   d. Cell Ontology (CO)



Unique Codes			
Terminology	**Mouse**	**Human**	**Total**
UMLS	224	233	371
MA	57	2	57
MP	12	0	12
CO	38	52	66
Total	331	287	506

2. **Total number of annotations** (PubMed ID – gene – GO code – context code)
   a. Mouse: 1,480
   b. Human: 1,312

3. Distribution of **annotations per class** (Anatomy, Cell Type, Disease/Problem)



4. **Precision** (per class – Anatomy, Cell Type, Disease/Problem):
   a. Generate list for each class of those PubMed IDs with at least one code for the class (e.g., list of all PubMed IDs having at least one associated Anatomy code)
   b. Choose 50 random PubMed IDs from this list
   c. For each PubMed ID, randomly choose one gene-GO pair from the gold standard (GOA Database) and one context code for the class
   d. Evaluate accuracy of context code with relation to gene-GO pair and the article

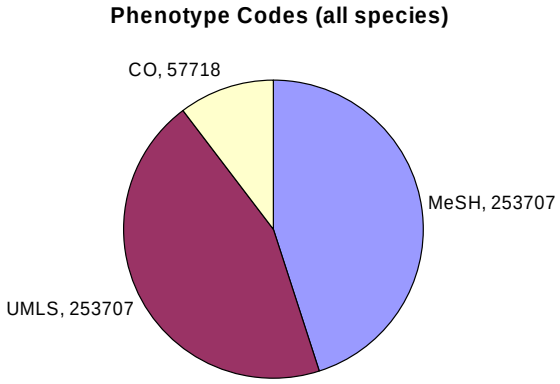
Class	**Precision**			
	Mouse		**Human**	
Anatomy	90.74	[86, 94]	76	[67, 83]
Cell	93.73	[89, 96]	80	[74, 86]
Disease	79.17	[73, 85]	-	-
Average	87.88	-	78	-

5. **Recall** (per class – Anatomy, Cell Type, Disease/Problem):
   a. Choose 50 random PubMed IDs
   b. Expert analyzed each title/abstract extracting terms associated with each class
   c. Evaluate accuracy of terms extracted by NLP system as compared to expert

Class	**Recall** | [Confidence Interval]			
	Mouse		**Human**	
Anatomy	94	[91, 96]	91	[88, 94]
Cell	82	[77, 87]	72	[66, 78]
Disease	79	[73, 84]	-	-
Average	85	-	81	-

Medical Subject Headings (MeSH) Statistics and Evaluation

1. Distribution of **phenotypic context codes** per terminology
   a. Medical Subject Headings (MeSH)
   b. Unified Medical Language System (UMLS) – mapped from MeSH
   c. Cell Ontology (CO) – mapped from MeSH

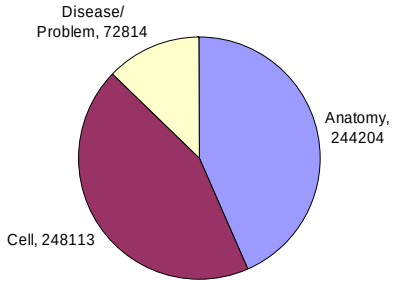


Terminology	Unique Codes
MeSH	1,945
UMLS	1,943
CO	118

2. **Total number of annotations** (database - PubMed ID – gene – GO code – context)
   a. Total (all species): 565131

Species	Annotations
Baker's yeast	5,264
Bovine	931
C. elegans	13,510
Chicken	394
D. melanogaster	103,830
Drosphila sp.	272
Fission Yeast	382
Human	109,704
Mouse	310,663
Rat	16,764
Zebrafish	3,417
Total	**565,131**

3. Distribution of **annotations per class** (Anatomy, Cell Type, Disease/Problem)



Species	Anatomy	Cell	Disease/Problem
Baker's yeast	484	4,074	706
Bovine	178	519	234
C. elegans	8,128	4,410	972
Chicken	178	204	12
D. melanogaster	58,110	40,852	4,868
Drosphila sp.	148	114	10
Fission Yeast	58	208	116
Human	32484	54,838	22,382
Mouse	133,522	135,757	41,384
Rat	8,636	6,208	1,920
Zebrafish	2,278	929	210
Total	244,204	248,113	72,814

4. **Precision** (per class – Anatomy, Cell Type, Disease/Problem):
   a. Generate list for each class of those PubMed IDs with at least one associated MeSH code for the class
   b. Choose 50 random PubMed IDs from this list
   c. For each PubMed ID, randomly choose one gene-GO pair from the gold standard (GOA Database) and one context code for the class
   d. Evaluate accuracy of context code with relation to gene-GO pair and the article

5. **Recall** (per class – Anatomy, Cell Type, Disease/Problem):
   a. Choose 50 random PubMed IDs
   b. Expert analyzed each title/abstract extracting terms associated with each class
   c. Evaluate accuracy of terms coded by MeSH as compared to expert

Class	**Precision** [Confidence Interval]		**Recall** [Confidence Interval]	
Anatomy	88	[83, 93]	75	[70, 80]
Cell	88	[84, 91]	79	[72, 86]
Disease	80	[74, 86]	74	[66, 83]
Average	85	-	76	-