NIF

Neuroscience
Information
Framework

February 9, 2012

The Representation of Gender in NIF’s Data Holdings

The Neuroscience Information framework indexes 150 individual databases deeply, meaning that it exposes data from those databases and data sets. These include large aggregators of data such as the model organism databases (mouse genome informatics, MGI) and small data sets such as the 1,000 functional connectomes. These holdings break down to approximately 350 million individual data records, most of which are tagged and aligned to some extent to a structured vocabulary.

In the first pass, the term male was searched exclusively. Of the 350 million total records in NIF’s holdings, the search for male reveals that there are 159 million records that mention the term male as well as 5.9 million articles: http://neuinfo.org/nif/nifgwt.html?query=male

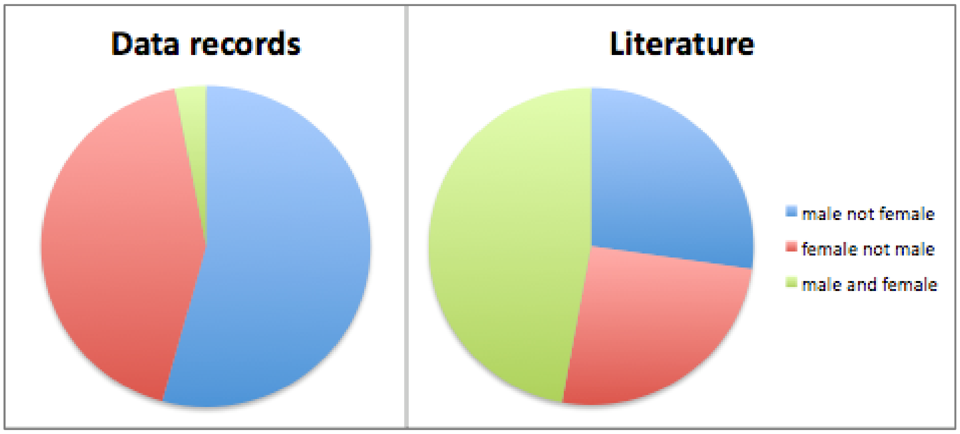
Searching female exclusively reveals that 127 million records and 5.9 million articles mention the term female: http://neuinfo.org/nif/nifgwt.html?query=female

In order to break down some of these findings, we can examine both the literature and the data results and compare the prevalence of male versus female. All raw data and numbers are available in the attached appendix.

In most cases, the data labeled with either male or female are indeed data gathered from that particular gender of animal. However, it should be noted that there are gene names or phenotypic descriptions in this set that include the term male such as “gene function required for the development of male germ cells” or the male-specific lethal gene. Currently we cannot easily exclude gene names from the search results, so an interpretation of the following data should be treated with some caution as not all of the results are specific to an organism that is male or an organism that is female.

Below are pie charts that visually represent the data collected for queries performed on NIF data records and literature (Figure 1). The results of these queries are separated into records and papers that returned male (only male with no mention of female in the paper or data record, blue color), female (only female with no mention of male, red color), and both male and female (green color). These charts suggest that data records in which a gender is recorded deal exclusively with males approximately 55% of the time while the literature deals with males and females together approximately 45% of the time, favoring males alone slightly over females. For a full set of numbers corresponding to these charts please see the figure 8 or the attached excel spreadsheet.

Figure 1:



In figures two and three, we have extracted only the records and publications that deal with humans or animals and have found that a somewhat different picture emerges. Human data records follow the trend described above with a slight male bias in the data records and literature. In the animal data, the bias toward studying males is quite strong. It appears that about half of the papers in the animal literature study males exclusively and the other half study either females or both males and females. Individual data records show that about one-sixth pertain to male, one-sixteenth pertain exclusively to females and more than three-fourths of the individual data records pertain to both males and females. This implies that, in most cases, data records in animals cannot be reliably traced back to a particular gender.

Figure 2:

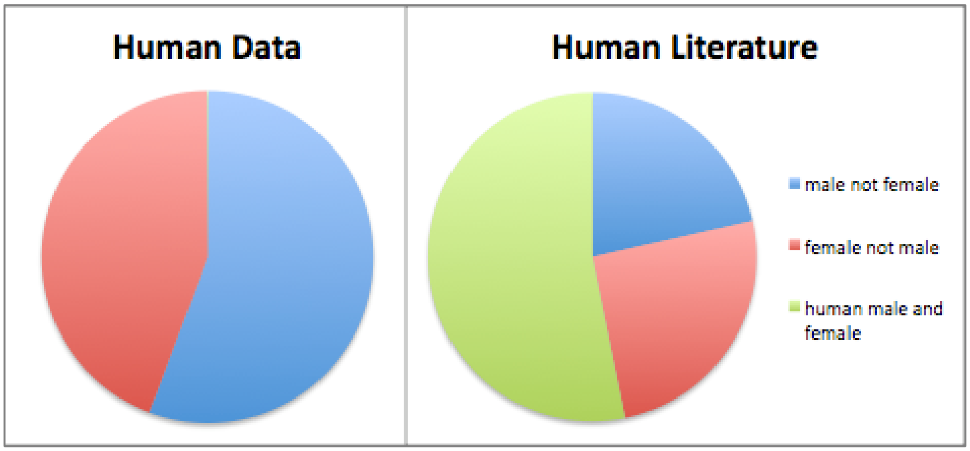
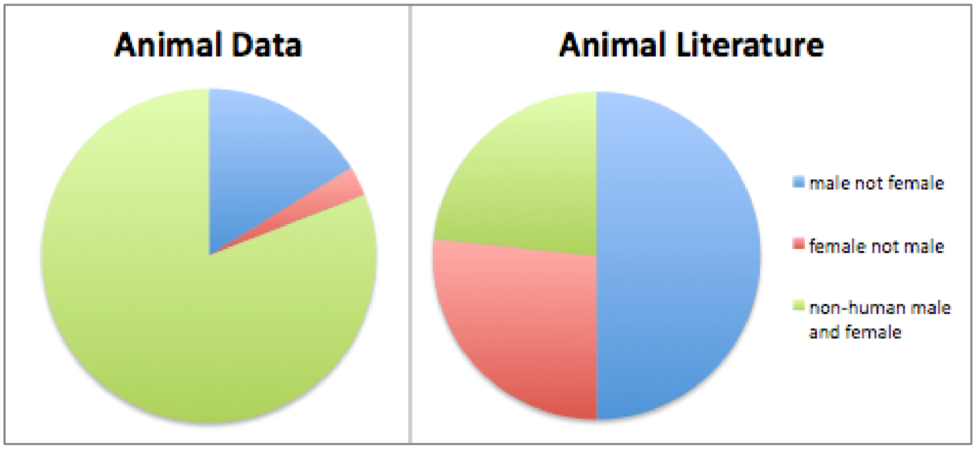


Figure 3:



Further analysis of the animal literature and data records as demonstrated in figures four and five reveals that mouse researchers generally do not keep track of gender and rat researchers largely study males.

Figure 4:

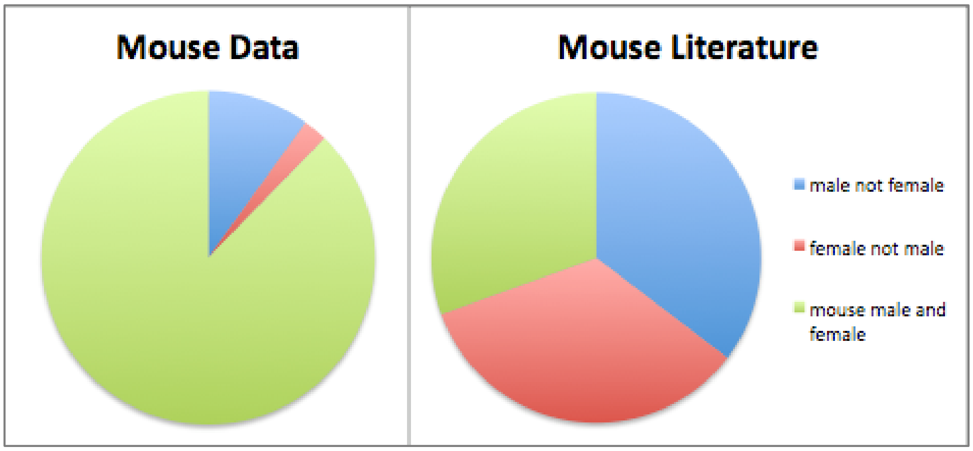
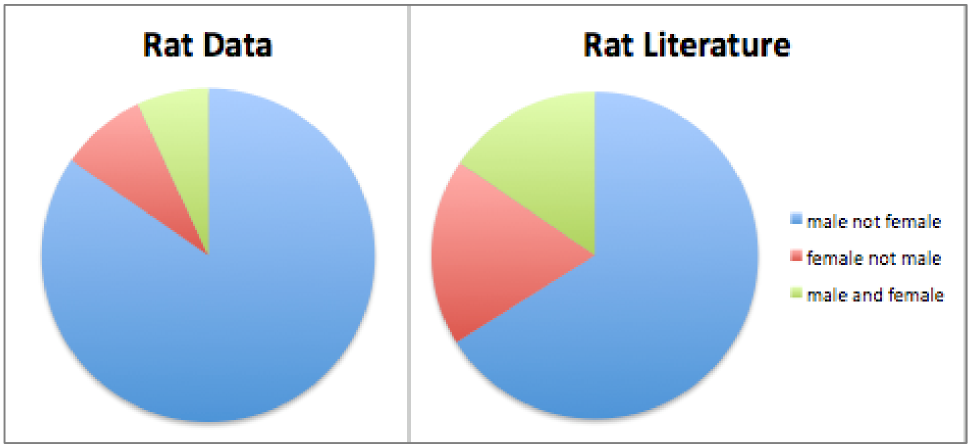


Figure 5:



Breaking down the literature more granularly allows us to generate the following graphs (Figures 6 and 7). These graphs were generated by searching gopubmed.org. It is worth noting that, before open access literature, we had almost no data about the gender of subjects being studied.

Figure 6: Published papers from 1998-2012 that contain the keyword male

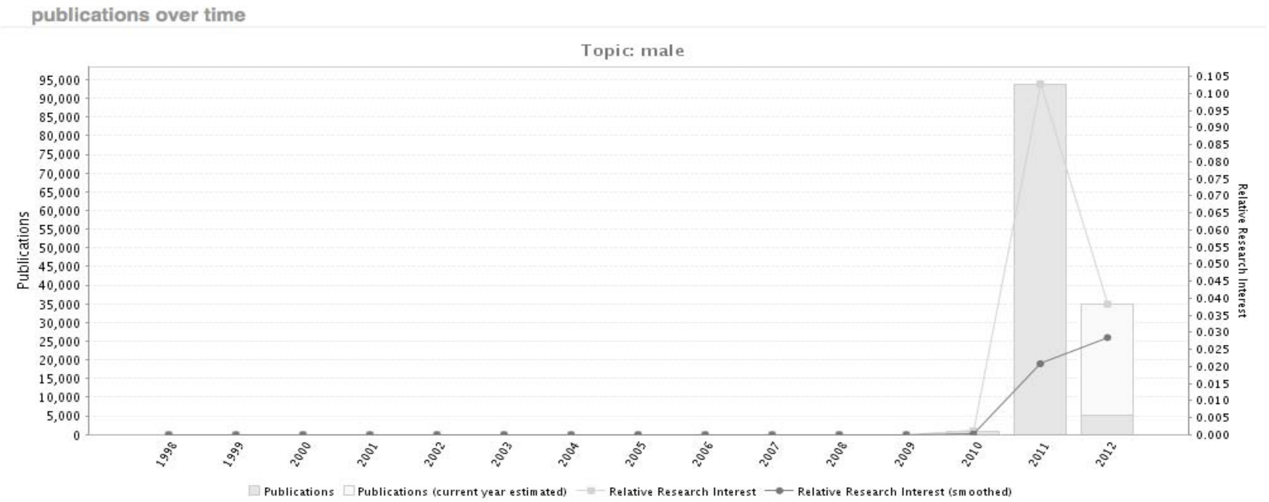


Figure 7: Published papers from 1998-2012 that contain the keyword female

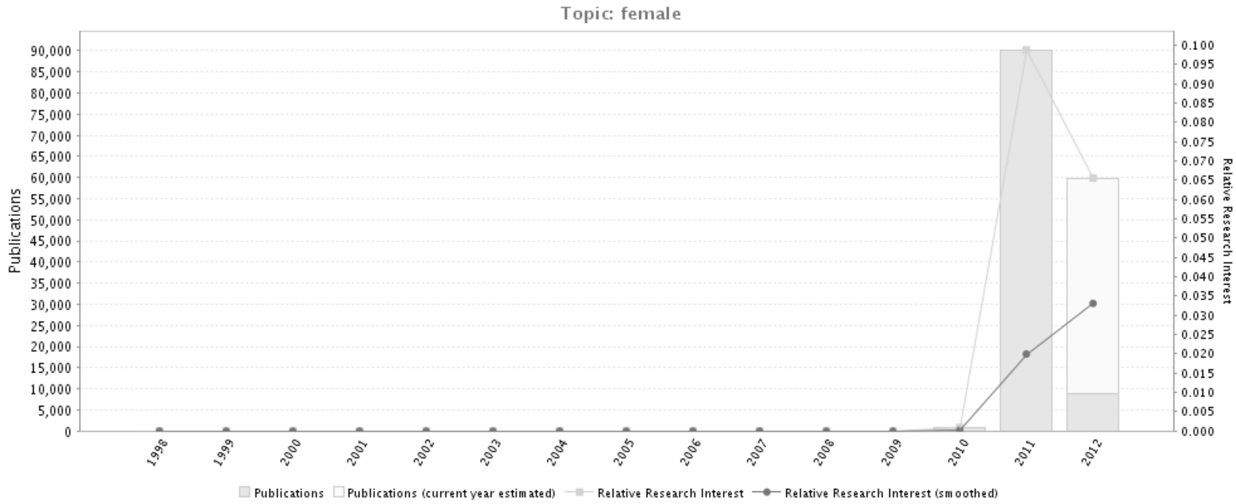


Figure 8: Heat map of directed search queries into NIF's data holdings broken down by source and query. Each box contains the number of results for each search executed in NIF; these numbers were used to construct the figures above. For a fully interactive heat map please see the attached excel sheet. Clicking on the column names should lead to a general search of the NIF data; each row header is labeled with a database name and links to a descriptive page about that database. Clicking on all green cells should execute a search against a specific database.

	Total Records	male	female	male and female	human male	human female	human male and female	animal male	animal female	animal male and female	mouse male	mouse female	mouse male and female	rat male	rat female	rat male and female
NIF Registry	0	##	9	15	##	4	4	5	11	11	4	1	2	1	0	0
Literature	##	##	##	##	##	##	##	##	##	##	##	##	##	##	##	##
Data Federation	##	##	##	##	##	##	##	##	##	##	##	##	##	##	##	##
AddGene	##	7			7											
AgingGenesDB Genes	##	7	8	14	1		1	8	13	13	1	4	5	3		5
AgingGenesDB Interventions	##	17	4	9	2		1	4	8	8	1		3			
AllenInstitute	##	##	17					17			85	17				
AmiGO	##	##	##	##	##	##	##	##	##	##		##	3	15	44	5
AntibodyRegistry	##	##	4	5	##	4	3		2	2	21	1	3	11	1	
AutDB	##	57	21	4				21	4	4	3	1				
BAMS	##	6		1					1	1						1
BioGRID	##	##			12											
BMI	##	48	2	2	38	1	1	1	1	1		1		1		1
BrainInfo	##	1														
BrainMaps	##	29			29											
BrainSpan1	##	##	##	##	##	##	##									
BrainSpan2	##	##	##	##	##	##	##									
BrainSpan3	##	##	##	##	##	##	##									

BrainSpan4	##	##	##	##	##	##	##									
BREDE	##	##	##	##	##	##	##	5					24		21	
CCDB	##	28	6		18	2		4								
CCDB	##	##	16		12	15		1			75	14		##		
CellImageLibrary	##		2					2							2	
Cerebellar Platform	##	50	23	23	3	1	4	22	19	19	1			22		3
chebi	##	50	2					2								
Chemoreceptors	##															
ClinicalTrials	##	##	##	##	##	##	##				14	28	##	30		##
CoriellTissue NIGMS	##	##	##	12	##	##	2									
CoriellTissue NINDS	##	##	##		##	##	12	2								
CoriellTissue	##	##	##	12	##	##	12	2								
CRCNS	8	1														
CTN	25			1					1	1						
DRG	##	##	##	##	##		14	##	##	##	##		12	##	##	##
DrugBank	##	34	30	5	6	6	1	22	4	4						
EEGbase	61	37	24					24								
EntrezGene	##	##	##	5	45	4		##	5	5	44	20	1	17		
F1000	##	82	55	58	8	5	4	45	50	50	3	1	##	15	1	
GARA	##	##	23	53												
Gemma	##	##	##	##	##	##	##	##	##	##	##	##	##	##	##	##
GeneNetwork	##	##	##	36	95	44	3	##	33	33	##	##	33	73		
GENSAT	##	##														
GENSAT Retina	##	79	##	57												
GEO	##	##	##	##	##	##	##	##	##	##	##	##	##	##	28	18
Grants.gov	##	10	9	12	3	1	7	5	5	5						
HomoloGene	##	##	78	1	54	7		62	1	1	32	5		18	2	
HumanBrainAtlas	##															
IBVD	##														2	
KawasakiDisease	##							57								
KiDatabase	##	1														
MGI	##	##	36					36			##	36				
MGITransgenes	##	22	1					1			9	1				
ModelDB	##															
Neurodatabase	14															
Neurofed	##	1														
NeuroMorpho	##	##	##	##	##	##	##	##	##	##	##		##	##	72	87
NeuronDB	60															
NeuronDB	##															
NeuronDB	##															
NIF Blogs	##	10	10	4	5	1	2	8	2	2	1		1			
Integrated Animals	##				31	6	3	17	50	50	13	3				
Integrated BGE	##	##	17					17			##	17			41	
Integrated Disease	##	8	8					6								
Integrated Connectivi	##	49	21					21							21	
Integrated Podcast	##	36	15	10	4	2		9	6	6						1
Integrated Software	##	1	1		1	1	1									
Integrated Video	##	48	39	12	7	8		26	9	9	1	1	1	1		1
NIHNeuro	##			8					8	8			2			
NIRTC	##	##	##					##								
ODE	##	##	##	##				##	##	##	12	20	2	2	1	
OdorMapDB	48	36	4	6				4	6	6	12	4	6	24		
OMIM	##	##		##	##		##		##	##	##	##	##	##	##	##

OneMind	##	4	1	1	4	1	1									
PhysioNetGaitND	##	98	68					68								
PhysioNetGaitNDD	64	28	36					36								
PubMedHealth	##	68	81	48		2		78	34	34						
RePORTER	##	##	##	##	##	##	##	##	##	##	##	##	##	##	##	##
ResearchCrossroads	##	##	##	##	##	##	##	##	##	##	##	##	##	##	##	##
RGD	##		1					1							1	
SfN	52	3		5	3		3		1	1	1					
Simtk	12	3			3											
StemCellInfo	96	1	3					3			1	3				
SumsDB	##	##	##	##	##	21	##	##	##	##						
SynapseWeb	##															
T3DB	##	28	39	7	5	5	1	34	6	6				3	1	1
Visiome	##	##	34	79	41	3	40	31	39	39	3	1		32		7
WikiPathways	##	1														
WormBase	##	##	6	11	48	1		5	11	11						
WormBase	##	##	2					2			5					
XNAT	##														2	
ZFIN	##	43	64					##	57	57						
ZFIN	##							64								