Searching for concepts in semantic space

Vector search is not just for examples anymore

https://github.com/rmhorton/PMC_classifiers

Robert Horton, PhD
Win-Vector Labs
rhorton@win-vector.com

Bay Area Use R Group
https://www.meetup.com/r-users/events/303488652/
October 15, 2024

Technical Takeaways

- Semantic embeddings
  - Capture the meaning of text in fixed-length numeric vectors
  - Turn NLP problems into geometry problems: search & prediction
- Label Mining
  - Build upon existing captured human judgement in Pubmed Central
    - Section heading patterns
    - Key terms (MeSH)
- Concept vectors
  - Represent abstract concepts in semantic space.
  - Prediction (model scoring) can be framed as similarity search.
  - Models as data
  - Similarity search is scalable (FANN).
- Transfer Learning
  - Will models trained on PMC data work for you?

Semantic Embeddings

sentence embedding: a numeric representation of a sentence in the form of a vector of real numbers which encodes meaningful semantic information.

https://en.wikipedia.org/wiki/Sentence_embedding

All the Python

```
from sentence_transformers import SentenceTransformer

xformer = SentenceTransformer("all-mpnet-base-v2")
embeddings = xformer.encode(sentences)
```

Pubmed

Free database of biomedical and life sciences literature

https://pubmed.ncbi.nlm.nih.gov/download/

Pubmed Central (PMC)

Free full-text archive of biomedical and life sciences journal literature from the National Institutes of Health's National Library of Medicine (NIH/NLM)

ftp.ncbi.nlm.nih.gov/pub/pmc/oa_bulk/oa_comm/xml/

No librarians were harmed in the making of this demo

pmid	para	section_path	text	embedding
27146290	0	Title	Trace Detection of RDX, HMX and PETN Explosives Using a Fluorescence Spot Sensor…	[-0.01,0.01,-0.03, …]
27146290	1	Abstract	1,3,5-trinitroperhydro-1,3,5-triazine (RDX), octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazocine (HMX), and pentaerythritol tetranitrate (PETN), the majo…	[0.01,0,-0.02, …]
27146290	2	Results and Discussions || Sensor characterization in solutions	The sensor reaction of DCM was first characterized in molecular solution, as shown in Fig. 2. The three explosives used, RDX, HMX and PETN, are white …	[-0.01,-0.03,-0.02, …]
27146290	3	Results and Discussions || Sensor characterization in solutions	The similar fluorescence quenching and absorption change were also observed for the other two explosives, HMX and PETN (Fig. S2). Control experiments …	[-0.01,-0.03,-0.01, …]
27146290	4	Results and Discussions || Fluo-spot sensing in silica gel TLC plate	With the confirmed sensor reaction in solution phase, the DCM molecular system was adapted into solid matrix, to improve the practical application in …	[0.01,-0.04,-0.02, …]

Label Mining

- Labels capture human judgement about concepts.
- A lot of judgement has already been captured
  - Indexing keywords in databases
  - Section headings as metadata
- Can we extract labels from this existing metadata?

Machine Learning: use FEATURES to predict LABELS

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
47	5.1	3.8	1.6	0.2	setosa
48	4.6	3.2	1.4	0.2	setosa
49	5.3	3.7	1.5	0.2	setosa
50	5.0	3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4.0	1.3	versicolor

pmid	para	X1	X2	X3	X4	X5	...	TITLE	RESULTS	STATISTICS
15354220	0	0.036	0.052	-0.029	0.035	-0.025	...	1	0	0
15354220	1	0.030	0.027	-0.021	0.021	-0.018	...	0	0	0
15354220	2	0.024	0.057	0.002	0.008	-0.046	...	0	0	0
15354220	3	-0.015	-0.045	0.001	0.038	-0.021	...	0	0	0
15354220	4	0.049	0.017	0.010	0.028	-0.049	...	0	0	0
15354220	5	0.008	-0.069	0.021	0.014	-0.027	...	0	0	0
15354220	6	-0.007	-0.039	0.020	-0.023	0.015	...	0	0	1
15354220	7	0.043	0.050	-0.014	0.039	-0.032	...	0	1	0

Performance of pattern models on a test set
hand-labelled for adverse events

name	pattern	auc
TITLE	^title$	0.51648
AE1	adverse events	0.82069
AE2	adverse event	0.78856
AE3	adverse (event|effect)	0.81252
AE4	adverse.*(event|effect)	0.80106
AE5	results.*adverse.*(event|effect)	0.84942
AE6	results.*(adverse.*(event|effect)|tolerability)	0.84968
AE7	results.*(adverse (event|effect)|tolerability|safety)	0.86996
AE8	results.*(adverse.*(event|effect)|tolerability|safety)	$\boxed{0.87028}$
AE9	results.*(adverse.*(event|effect)|tolerability|safety|toxicit)	0.82714
TOL	tolerability	0.85133
SAFETY1	safety	0.82901
SAFETY2	results.*safety	0.84293
TOX1	toxic	0.82947
TOX2	toxicit	0.82988
TOX3	results.*toxic	0.81525
TOX4	results.*toxicit	0.82604

https://www.nlm.nih.gov/mesh/meshhome.html

Welcome to Medical Subject Headings

The Medical Subject Headings (MeSH) thesaurus is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine. It is used for indexing, cataloging, and searching of biomedical and health-related information. MeSH includes the subject headings appearing in MEDLINE/PubMed, the NLM Catalog, and other NLM databases.

Recent MeSH Updates

Visit our What's New page to see all recent MeSH developments including the most recent ones listed below

Learn About MeSH

- Tutorials and Webinars
- MeSH Vocabulary
    - Introduction to MeSH

Concept Vector

A representation of a category of items in a semantic embedding space. This represents a concept to the extent that the items in the category represent the concept.

These vectors can be constructed from the coefficients of a logistic regression classifier.

All the math

cosine similarity

dot product

$$S_C(\mathbf{A},\mathbf{B}) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \boxed{\frac{\mathbf{A}}{\|\mathbf{A}\|}}\cdot\boxed{\frac{\mathbf{B}}{\|\mathbf{B}\|}}$$

vector lengths

unit vectors

logistic regression

$$P(y \mid \mathbf{x}) = \sigma(\beta_0 + \boldsymbol{\beta}\cdot\mathbf{x})$$

logistic "squashing" function

intercept

coefficient vector

feature vector

coefficient vector

feature vector

$$P(y \mid \mathbf{x}) = \sigma(\beta_0 + \|\boldsymbol{\beta}\|\boldsymbol{b}\cdot\mathbf{x})$$

squash it!

shift it!

scale it!

dot product

where $\boldsymbol{b} = \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}$ is the coefficient unit vector

Extract model data

```r
get_parameter_row <- function(clf){
  vector_length <- function(v) sqrt(sum(v*v))

  par <- clf %>% coef(s='lambda.1se') %>% as.matrix %>% '['(,1)
  intercept <- par[[1]]
  beta <- par[-1]
  beta_length <- vector_length(beta)
  beta_unit_str <- (beta/beta_length) %>% pgvector.serialize
  list(intercept=intercept, beta_scaling_factor=beta_length, beta_unit_vector=beta_unit_str)
}

model_data <- model_list %>% lapply(get_parameter_row) %>%
  bind_rows %>% bind_cols(target=names(model_list), .)
```

```
> model_data %>% tibble
# A tibble: 19 × 4
   target  intercept beta_scaling_factor beta_unit_vector
   <chr>       <dbl>               <dbl> <chr>
 1 TITLE       0.909                72.8 [-0.00754755248026402,0.0341607196583276,0.0105754463213793,-0.0…
 2 AE1        -4.96                 30.1 [-0.0311827986111597,-0.0701493689588177,0.00360991330727554,0.0…
 3 AE2        -4.84                 24.0 [-0.0104774311105458,-0.0688404554711353,0.0637831429635,0.038…
 4 AE3        -4.45                 26.3 [-0.0881265092249542,-0.00670485889907199,0.000293081319691433,0…
 5 AE4        -4.00                 29.6 [-0.0201045147846336,-0.0251302531827192,0.0342076459567 7,0.005…
```

Vector search

```r
scored_paragraphs_sql <- sprintf("with concept_vectors(name, vector) as (
  values
  ('%s', '%s'),
  ('%s', '%s')
),
cv as (
  select cast(name as text) as name, cast(vector as vector(768)) as vector from concept_vectors
),
scored_examples as (
  select pmid, paragraph_number
      , vector <#> (select vector from cv where name='AE8') as AE8_score
      , vector <#> (select vector from cv where name='TOX4') as TOX4_score
    from embedding
    limit 2000
)
select se.*, p.section_path, p.text
  from scored_examples se
  join paragraph p on se.pmid=p.pmid and se.paragraph_number=p.paragraph_number
",
"AE8", model_data[model_data$target=="AE8",][["beta_unit_vector"]],
"TOX4", model_data[model_data$target=="TOX4",][["beta_unit_vector"]])

scored_paragraphs <- dbGetQuery(con, scored_paragraphs_sql)
```

I foolishly named my vector columns 'vector'

There is also a data type named 'vector'

Inner product of two vectors

Find the top MeSH terms for a paragraph

```
pmid = '25215334'
para = 1
paragraph_sql <- sprintf("select * from paragraph where pmid='%s' and paragraph_number=%d", pmid, para)
embedding_sql <- sprintf("select vector from embedding where pmid='%s' and paragraph_number=%d", pmid, para)

paragraph_text <- dbGetQuery(con, paragraph_sql)[['text']]
query_vector <- dbGetQuery(con, embedding_sql)[['vector']][[1]]

# get top mesh terms for embedding
mesh_sql <- sprintf("
      select dmd.target, dd.name, dmd.beta_unit_vector <#> '%s' score
        from descriptor_model_data dmd
        join descriptor_detail dd on dmd.target = dd.id
        order by score limit 5", query_vector)
top_mesh_terms <- dbGetQuery(con, mesh_sql)
```

Find the top MeSH terms for a paragraph

"Most coastal structures have been built in surf zones to protect coastal areas. In general, the transformation of waves in the surf zone is quite complicated and numerous hazards to coastal communities may be associated with such phenomena. Therefore, the behavior of waves in the surf zone should be carefully analyzed and predicted. Furthermore, an accurate analysis of deformed waves around coastal structures is directly related to the construction of economically sound and safe coastal structures because wave height plays an important role in determining the weight and shape of a levee body or armoring material. In this study, a numerical model using a large eddy simulation is employed to predict the runup heights of nonlinear waves that passed a submerged structure in the surf zone. Reduced runup heights are also predicted, and their characteristics in terms of wave reflection, transmission, and dissipation coefficients are investigated."

target <chr>	name <chr>	score <dbl>
D013314	Stress, Mechanical	−0.2229931
D003247	Conservation of Natural Resources	−0.1961243
D003198	Computer Simulation	−0.1960209
D045483	Rivers	−0.1910898
D014874	Water Pollutants, Chemical	−0.1879359

Transfer Learning

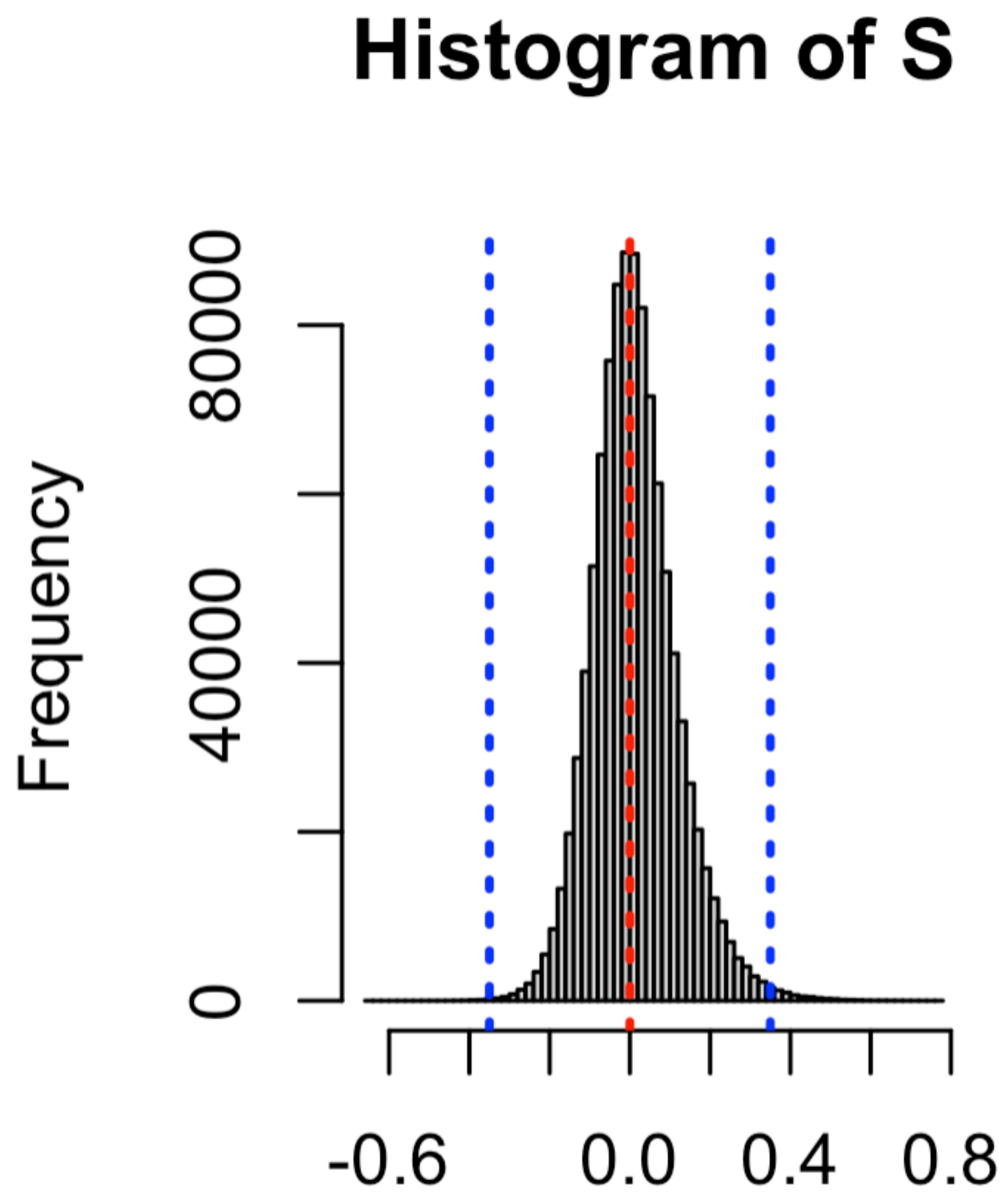
“A technique in machine learning (ML) in which knowledge learned from a task is re-used in order to boost performance on a related task.”

https://en.wikipedia.org/wiki/Transfer_learning

For our purposes it mostly means:
“Training a model to recognize a concept in PMC articles, then using it to predict that concept (or a related concept) in a different corpus.”

Compare concept vectors to each other

```
S <- M %*% t(M)
diag(S) <- 0 # diagonal
threshold <- 0.35
hist(S, breaks=100)
abline(
  v=c(-threshold, 0, threshold),
  col=c('blue', 'red', 'blue'),
  lty=3, lwd=3)
```

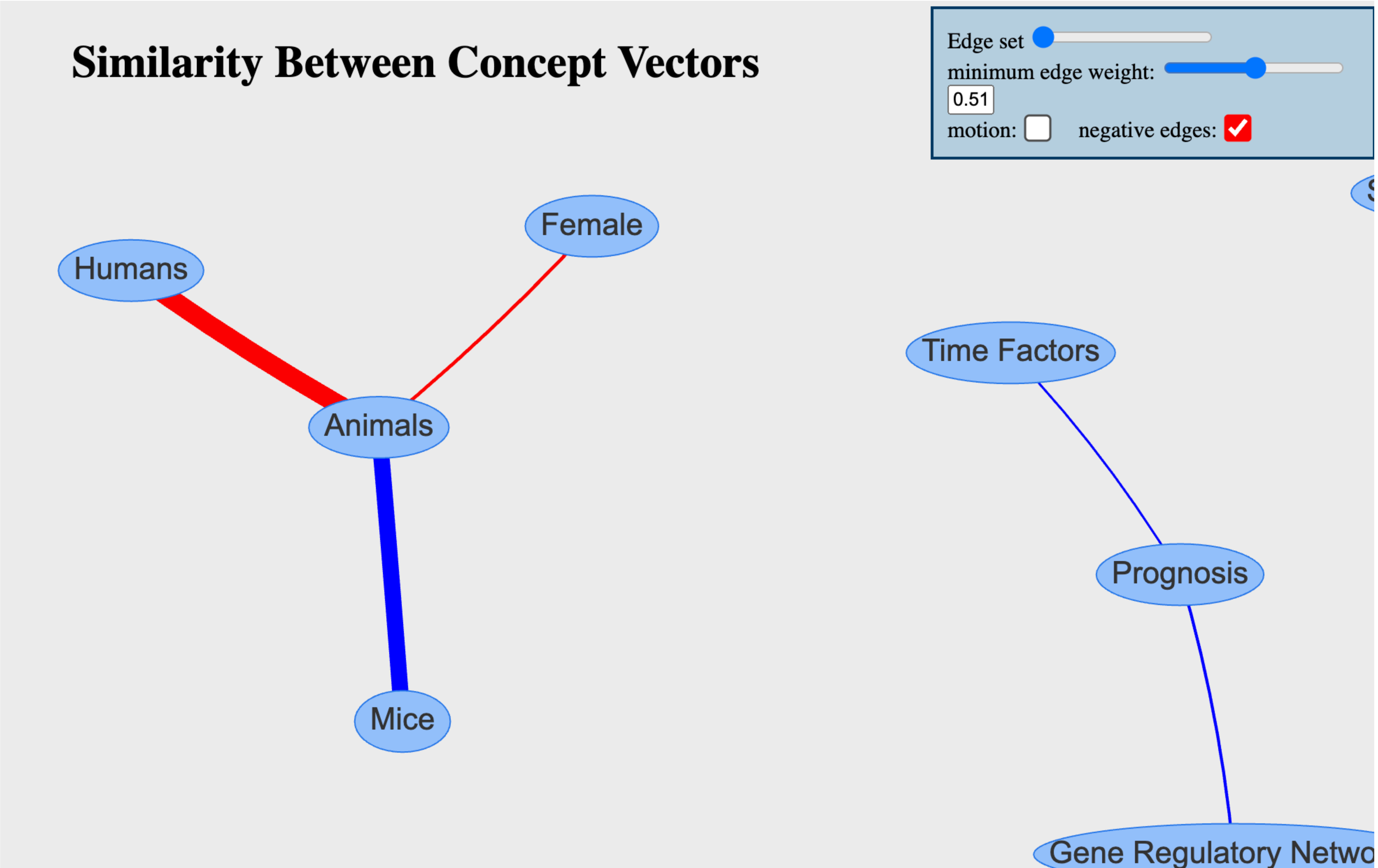


Are Humans Animals?

- **Animals:** Unicellular or multicellular, heterotrophic organisms, that have sensation and the power of voluntary movement. Under the older five kingdom paradigm, Animalia was one of the kingdoms. Under the modern three domain model, Animalia represents one of the many groups in the domain EUKARYOTA.

- **Humans:** Members of the species Homo sapiens.

Graph Visualization



Compare concept vectors to definition embeddings

```
definition_embeddings <- read_parquet(DEFINITION_EMBEDDINGS_FILE)

D <- definition_embeddings$vector %>% do.call('rbind',.)

definition_meshterms <- D %*% t(M)
dimnames(definition_meshterms) <- list(
  definition_embeddings$term,
  model_data$name
)

dim(definition_meshterms) # [1] 30605  1014
```

```
# definition is row, MeSH term is column
definition_meshterms['Animals', 'Mice']    #  0.028
definition_meshterms['Mice', 'Animals']    #  0.314
definition_meshterms['Animals', 'Humans'] # -0.106
definition_meshterms['Humans', 'Animals'] # -0.024
```

Models vs. Definitions

model_term	0_x	1_x	2_x	3_x	4_x	5_x	6_x	7_x	8_x	9_x	10_x	11_x	12_x	13_x	14_x	15_x	16_x	17_x	18_x	19_x
3' Untranslated Regions	RNA 3' Pol	Polymorph	Genes, cd	Amplified	Transcripti	Transcripti	Transcripti	Internal Ri	Chromoso	Cleavage S	Ribonuclea	MicroRNA	Fusion Pro	ELAV-Like	RNA Meth	Polynucle	Valine-tRN	Polyadeny	ELAV-Like	NEDD8 Pr
A549 Cells	A549 Cells	Virus Phys	AKR murin	Pulmonary	Murine pn	Cytostatic	Pipobroma	Oncolytic	RNA Virus	DNA Virus	Mitosis M	Anti-Asthn	Anticarcin	FANFT	Sestrins	Pulmonary	E-Cigarette	Fanconi Ar	Fanconi Ar	Antineopla
Actins	Tropomod	Tensins	Myosin VII	Myosins	Actins	Activin Re	Lim Kinase	Myristoyla	alpha Cate	Calponins	CapZ Actir	AlkB Homo	Gelsolin	L Cells	Actin Capp	Contactin	Cortactin	Microfilam	Actin Cyto	Myosin-Li
Action Potentials	Small-Con	Large-Con	Purkinje Fi	Calcium Cl	Autonomi	Shaw Pota	Autonomi	Kv1.6 Pota	Calcium Cl	Potassium	Adrenergi	Cholinergi	Delayed R	Large-Con	Calcium Cl	Kv1.5 Pota	Intermedia	Calcium Cl	Calcium Cl	Large-Con
Activities of Daily Living	Geriatric A	Motor Dis	Accelerom	Homebou	Activities	Presentee	Frail Elderl	Centenaria	Motor Skil	Mobility Li	Housing fo	Fatigue Sy	Human Ac	Actigraphy	Mental Sta	Absenteei	Octogenar	Caregiver	Homes for	Hemiplegi
Acute Disease	Acute Che	Acute Aor	Transfusio	Pancreatit	Middle Lo	Acute Dise	Hyphema	Case Repo	Pituitary A	Medical A	Out-of-Ho	Case Repo	Karoshi De	Advanced	Exsanguin	Air Ambul	Pancreatit	Diagnostic	Pancreatit	Acute Care
Acute Kidney Injury	Kidney Co	Kidney Tu	Acute Kidr	Kidney Dis	Uremic To	Benzolami	Nephritis,	Kallikrein-l	Fanconi Sy	NADPH Ox	Halofenate	Cilastatin	Carbonic A	Hypertens	Perinephri	Azotemia	Madin Dar	Acecainide	Glomerulo	Hemolytic
Adaptation, Physiological	Extremoph	Bacterial P	Salt Tolera	Adaptatio	Biophysica	Plant Phys	Nonlinear	Adaptatio	Crassulace	Musculosk	Biomecha	Physiologi	Baroreflex	Thermoto	Urinary Tr	Adaptatio	Musculosk	Heat-Shoc	Microbiolo	Freshwate
Adaptation, Psychological	Psychologi	Coping Ski	Survivorsh	Psychosoc	Mental Sta	Resilience,	Work-Life	Psychosoc	Life Course	Subjective	Adjustmer	Psychome	Play Thera	Transtheo	Social Sup	Orientatio	Caregiver	Schizophre	Counselor	Narrative
Adaptor Proteins, Signal Tr	Basal Cell	Methyl-Cp	Bone Mor	Silver-Russ	Focal Facia	Hajdu-Che	B-Cell Lym	Bone Mor	Bone Mor	RNA-Bindi	Nasophary	Tuberous S	Sp7 Transc	T-Cell Intra	Inhibitor o	Proto-Onc	Costello Sy	Retinoblas	Osteochor	Genes, bcl
Adenocarcinoma	Adenocarc	Colorectal	Colonic Ne	Lung Neop	Digestive S	Esophagea	Adenocarc	Gastrointe	Bronchial	Gallbladde	Colitis-Ass	Endoscopy	Respirator	Proctosco	Endoscopy	Barrett Eso	Anal Gland	Anus Neop	Transanal	Retroperit
Adenosine Triphosphate	ATPase Inh	Mitochon	ATP Synth	Mitochon	AAA Prote	Rhodamin	P-type ATF	Mitochon	ATPases A	DNA Ligas	Adenylyl C	Sodium-Po	AAA Doma	Sarcoplasr	Membran	Oxidative	Pyruvate K	Ryanodine	Excitation	Adenylyl I
Adipose Tissue	Adipogene	Adipose Ti	Receptors,	Adipocyte	Adipocyte	Adipose Ti	Adipose Ti	Adipokine	Anti-Obesi	Adipose Ti	3T3-L1 Cel	Adiponect	Lipid Mobi	Lipogenes	Epicardial	Fat Necros	Obesity Hy	Intra-Abdo	Subcutane	Ketone Bo
Administration, Oral	Clinical Tri	Administra	Dosage Fo	Pharmaco	Clinical Tri	Drug Thera	Medicatio	Vaccinia	Controlled	Administra	Alprostadi	Clinical Tri	Drug Presc	Valgancicl	Smallpox V	Medicatio	Injections,	Dispensato	Clinical Tri	Injections,
Adolescent	Adolescen	National L	Child Heal	Adolescen	Myoclonic	Adolescen	Adolescen	Homeless	Child Guid	Adolescen	Adverse Cl	Adolescen	Exposure t	Personality	Adolescen	Adolescen	Minors	Puberty, D	Neisseria r	Epilepsy, A
Adult	Cornell Me	Premarital	Attitude o	Case Repo	Practice Pa	Health Co	Health Sur	Practice Pa	Attitude to	Diagnostic	Direct-To-C	Head-Dow	Practice G	Medical H	Patients	Case Repo	Presentee	Symptom	Consumer	Clinical Me
Aedes	Aedes	Mosquito	Densovirin	Anopheles	West Nile	Insect Vec	Insect Pro	Culex	Entomobir	Encephalit	Insecticide	Insect Rep	Mosquito-	La Crosse	Encephalit	Mosquito	Zika Virus	Genes, Ins	Insect Viru	Encephaliti
Africa	HIV Serop	HIV Seroso	Tropical M	Global Hea	Western V	Anthropol	Sub-Sahar	Leishmani	Rift Valley	Neglected	Culturally	South Ame	Civilizatior	Cross-Cult	Communic	Pandemics	Epidemics	Naja haje	Indians, Ce	Hepatitis E
Age Distribution	Carcinoma	Osteosarc	Mortality	Opioid Epi	Influenza I	Vaccinatio	Morbidity	Epidemiol	Carcinoma	Incidence	Keratosis,	Centenaria	Prevalence	SEER Prog	Choroid N	Cause of D	Child Mort	Myopia, D	Age Distrib	Mortality,
Age Factors	Geriatric A	Adolescen	Adolescen	Centenaria	Health Ser	National L	Adolescen	Child Heal	Health Tra	Age Distrib	Geriatric A	Elder Nutr	Transition	Adult	Ageism	Age Deter	Adult Chil	Young Adu	Child Nutr	Puberty, P
Aged	Centenaria	Geriatric A	Aged, 80 a	SEER Prog	Practice G	Mixed Der	Vascular D	Octogenar	Geriatric A	Hospitals,	Elder Nutr	Health Ser	Nonagena	Geriatricia	Dementia,	Middle Ag	Therapeut	Dental Car	Practice G	Geriatric D
Aged, 80 and over	Octogenar	Centenaria	Geriatric A	Nurses Im	Geriatric A	Aged, 80 a	Nonagena	Geriatricia	Elder Nutr	Health Ser	Aged	Mixed Der	Homes for	Medicare	Middle Ag	Aftercare	Housing fo	Hospitals,	Medicare	Senior Cer
Aging	Aging	Aging, Pre	Cognitive	Immunose	Healthy Ag	National I	Geriatricia	Elder Nutr	Centenaria	Age Deter	Skin Aging	Geriatric A	Age Factor	Senescenc	Geriatrics	Ageism	Alzheimer	Octogenar	Chronobio	Housing fo
Agriculture	Technolog	Farmers	Forests	Grassland	Crops, Agr	Agrochem	Plant Disp	Horticultu	Agriculture	Agricultura	Farms	Agricultura	Soil Pollut	Organic Ag	Rainforest	Crop Prod	Ecological	Gardening	Environme	Lot Quality
Air Pollutants	Traffic-Rel	Air Polluti	Air Filters	Air Polluta	Air Polluti	Air Polluta	Smog	Climatic P	Air Polluta	Tobacco S	Vehicle En	Air Polluti	Anthropog	Capnograp	Weather	Non-Point	Models, Sp	Greenhous	Environme	Petroleum
Air Pollution	Traffic-Rel	Smog	Air Polluti	Air Polluta	Air Polluta	Vehicle En	Air Polluta	Greenhou	Air Polluti	Non-Point	Air Filters	Light Pollu	Tobacco S	Petroleum	Capnograp	Smoke	Carcinoge	Nitrogen D	Weather	Automobi
Alcohol Drinking	Alcohol-In	Alcohol Dr	Alcohol Dr	Alcoholism	Alcohol-Re	Alcohol-In	Alcohol Ab	Drinking	Alcohol Ar	Alcoholic I	Binge Drin	Alcoholics	Cardiomyo	Alcoholic E	Pancreatit	Drinking B	Underage	Substance	Alcoholic N	Alcoholics
Algorithms	Mathemat	Unsupervi	Radiomics	Deep Lear	Serial Lear	Supervise	Cellular Au	Neural Ne	Electronic	Computer	Medical In	Dimensior	Soft Comp	Machine L	Automate	Signal-To-I	Computer	Models, N	Models, Cl	Autosugge
Alleles	Genetic Ca	Human Ge	Genetic Va	Quantitati	Genetic He	Consangui	Hemoglob	Inbreeding	HapMap P	Genome-V	Transplant	Amplified	HLA-C Ant	HLA-B Ant	Haplotype	Immunogl	Congenita	Gene-Envi	Polymorph	Homozygo

top_20_definition_terms.xlsx

Future Goals

- Assess and document biases in MeSH term models
- Train models on big datasets using GPU
- How many MeSH terms can we predict reasonably?
- Interpretable representation in 'MeSH term space'