

# CUSTOMIZABLE CATEGORIZATION OF DOCUMENTS IN EVIDENCE-BASED RESEARCH FOR BIOPHARMA

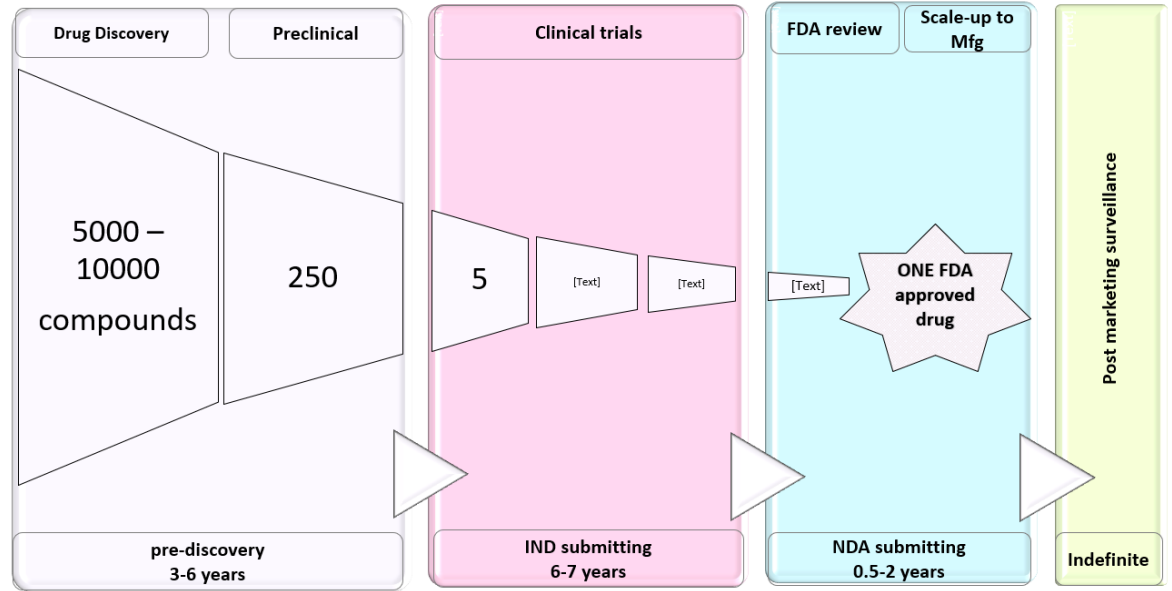
Reihaneh Manteghi (PhD candidate) - Eduardo Veas (supervisor)  
Graz University of Technology, Know-Center GmbH, Austria  
UCB Biopharmaceuticals, Brussels, Belgium



# What is the problem?



## Drug development



**Antimuscarinic Anticholinergic Medications in Parkinson Disease: To Prescribe or Deprescribe?**  
 Matthew J. Barrett MD, MSc, Lana Sargent PhD, RN, CRNP, Huma Nawaz MBBS, Daniel W. Ekin T, Price PharmD, PhD, Allison W. Willis MD, MSc  
 First published: 17 September 2021  
 https://doi.org/10.1002/mdc3.13347  
 Citations: 4  
 Relevant disclosures and conflicts of interest are listed at the end of this article.

**Patterns of Dementia Treatment and Frank Prescribing Errors in Older Adults With Parkinson Disease**  
 Original Investigation  
 January 2019  
 Sneha Mantri, MD, MS<sup>1,2</sup>, Michelle Fullard, MD, MS<sup>2</sup>, Shelly L. Gray, PharmD<sup>1,2</sup>, et al.  
 JAMA Neurol. 2019;76(1):41-49. doi:10.1001/jamaneurol.2018.2820

**Key Points**  
**Question** What are the prevalence of antedementia treatment among older adults with Parkinson disease



**HOT**  
 Health Optimization Tool  
 documents  
 Find a Category or topic  
 Add New Topic Add Edit Delete Copy Paste Search

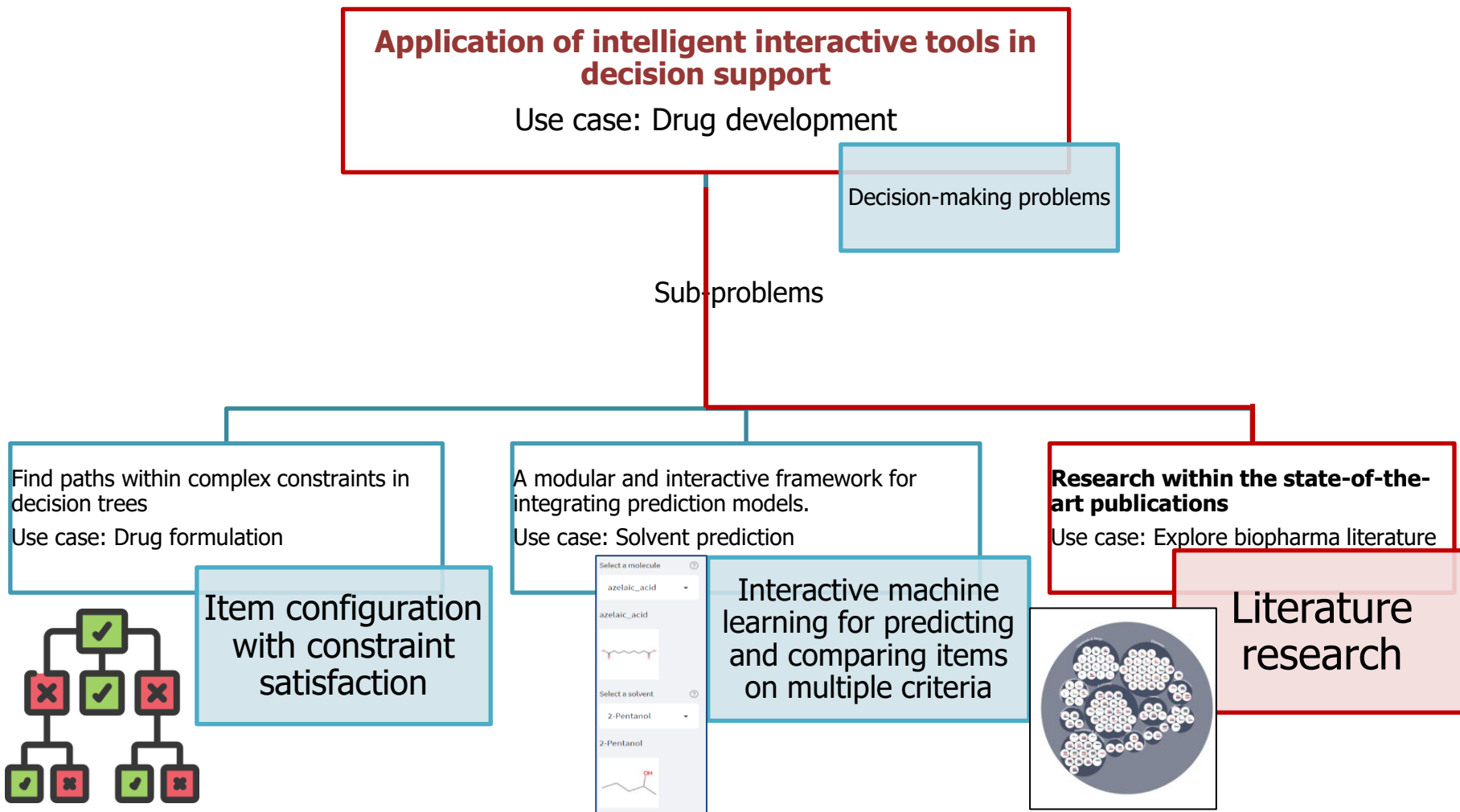
Page 1 < >

We aimed determine implementing standard care public primary care among children years acute respiratory infection ARI presenting primary care visits rural district total population Ching Rai province Thailand more seasonal effects economic public education older patients ARI performed analysis health system perspective Division levels constructed three patient categories children years children years adults time factors Model parameters based data patterns included review acute infection management primary care walk published studies procurement price list Parameters varied determinable variability markers Costs reported US dollars threshold per DAAY avoided annual direct cost public primary associated staff training monitoring If reduced deaths cover lower respiratory tract infections children years per patient annually In population children years equivalent DALYs avoided per year When downstream cost savings hospitalization inappropriate antibiotic prescription considered public primary care standard care saving annually 1.1th intervention yielded modest mortality gains older patients avoided cost savings primarily reducing inappropriate antibiotic prescription use groups The downstream intervention also demonstrated usability estimate adjusted ARI primary care management rural northern Thailand This finding their generalizable weighting countries similar disease epidemiology health system

Next infections diseases particularly correct fungal pathogens poor considerable risk global biodiversity The amphibian chytrid fungus Batrachochytrium dendrobatidis BtD demonstrated scale threat caused greatest revealed low vertebrate biodiversity attributable pathogen Despite relatively decline several countries cases affected species experienced population recovery evidence However potential impact

# Overall research question

Can interactive search tools with intelligent user interfaces help domain experts in **decision-making** processes?



**HOT**  
Health Optimization Tool

documents

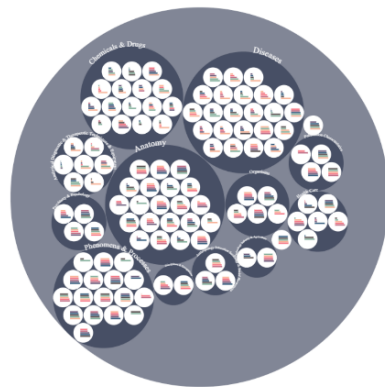
Find a Category or topic

Page 1 < >

We aimed determine supplementing standard care pulse oximetry among children years acute respiratory infection ARI preventing primary care units rural district total population Chiang Rai province Thailand assess economic effects estimating pulse oximetry older patients ARI performed analysis health systems perspective Decision trees constructed three patient categories children years children years adults time horizon Model parameters based data patients included review acute infection management primary care units published studies procurement price list Parameters varied deterministic sensitivity analyses Costs expressed US dollars threshold per DALY averted annual direct cost pulse oximetry associated staff training monitoring if reduced deaths severe acute respiratory tract infections children years per patient annually in population children years equi-cost DALYs averted per year When downstream costs related hospitalizations inappropriate antibiotic prescriptions considered pulse oximetry dominated standard care saving annually This intervention yielded smaller mortality gains older patients resulted cost savings primarily reducing inappropriate antibiotic prescriptions age groups The dominance intervention also demonstrated sensitivity oximetry address ARI primary care management rural northern Thailand This finding likely generalizable neighboring countries similar disease epidemiology health systems

link

Novel infectious diseases particularly caused fungal pathogens pose considerable risks global biodiversity The amphibian chytrid fungus Batrachochytrium dendrobatidis Bd demonstrated scale threat caused greatest recorded loss vertebrate biodiversity identifiable pathogen Despite catastrophic declines several continents many affected species experienced population recovery, rebounds, However, natural selection



## Literature research assistant

### Goal:

Customized classification and data exploration through scientific documents to improve the decision-making process

### Used in:

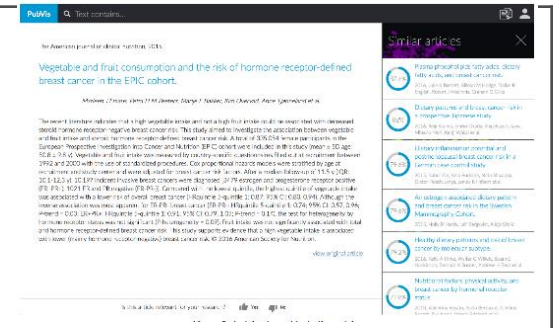
Different domains

**Use case:** Explore biopharma literature

# Similar tools

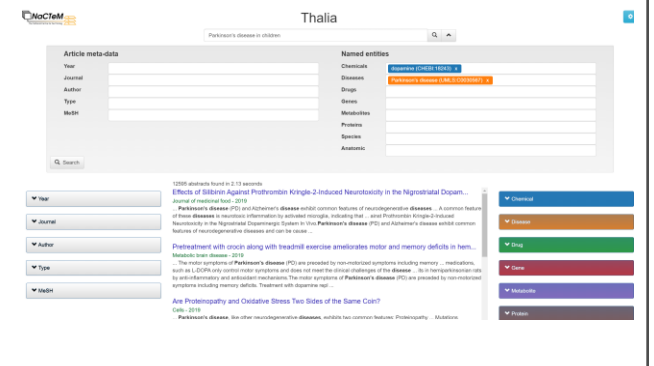
## PubVis<sup>1</sup>

Uses **interactive machine learning** and combines essential features to help researchers find the content most relevant to them—articles can be tailored to the user's interests



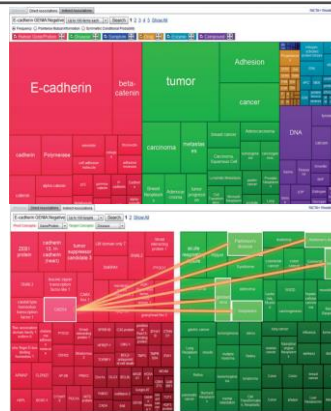
## Thalia<sup>2</sup>

A semantic search engine for exploring **biomedical abstracts**



## FACTA<sup>+</sup><sup>3</sup>

- Discovering useful **associations** between **biomedical** concepts
- Visualization of **directly** and **indirectly** associated concepts



Exp. Info. Info.	Target Concepts	Print Concepts	Query Concept
9.050   17.09	0.567 Parkinsonism	0.567 Parkinsonism	0.564 Parkinson
	0.500 (proliferation) disorder	0.500 (proliferation) disorder	0.500 Parkinson
	0.500 (proliferation) disorder	0.500 (proliferation) disorder	0.500 Parkinson
8.210   17.09	0.500 (proliferation) disorder	0.500 (proliferation) disorder	0.500 Parkinson
	0.500 (proliferation) disorder	0.500 (proliferation) disorder	0.500 Parkinson
7.116   17.09	0.567 Alzheimer's disease	0.567 Alzheimer's disease	0.572 Parkinson
	0.567 Alzheimer's disease	0.567 Alzheimer's disease	0.585 Parkinson
6.937   17.09	0.500 dementia	0.500 dementia	0.564 Parkinson
	0.500 dementia	0.500 dementia	0.589 Parkinson

1. Horn F. Interactive Exploration and Discovery of Scientific Publications with PubVis. arXiv e-prints. 2017 June. doi: 10.48550/arXiv.1706.08094.
2. Soto AJ, Przybyła P, Ananiadou S. Thalia: semantic search engine for biomedical abstracts. Bioinformatics. 2019 May 15;35(10):1799-1801. doi: 10.1093/bioinformatics/bty871. PMID: 30329013; PMCID: PMC6513154.
3. Tsuruoka Y, Miwa M, Hamamoto K, Tsujii J, Ananiadou S. Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics. 2011 Jul 1;27(13):i111-9. doi: 10.1093/bioinformatics/btr214. PMID: 21685059; PMCID: PMC3117364.

# Literature research assistant

## How it works:

In-house data exploration tool with search and customized classification option which can be used on different literature data

**HOT**  
Health Optimization Tool

documents

Find a Category or topic

Combined search query

disease and cognitive and dementia

dementia

Search

Page 1 < > 126

In central nervous system neurodegenerative disorders stem therapies considered promising therapeutic approach The safe use human Neural Stem Cells hNSCs treatment several neurological diseases currently evaluation phase clinical trials Clinical application hNSCs require development GMP standardized protocols capable generating high quantities reproducible well characterized stem cells bearing stable functional genetic aim study evaluate possible instabilities modifications microsatellite loci different culture passages high culture passages represent vitro replicative stress leading senescence Experimental method The hNSCs characterized different culture time points passage passage genetic typing ten microsatellite showed genetic stability microsatellite loci maintained cells even high passages adding demonstration safety hNSCs GMP culture method

link

brain ad cognitive alzheimer disease neuronal neurological neurodegenerative neurons dementia

disease AD common form dementia industrialized countries severely targets hippocampal formation humans mouse models condition The adult hippocampus hosts continuous addition new dentate granule cells DGCs numerous mammalian species including humans Although morphology positioning DGCs within granule cell layer GCL match developmental origin rodents similar correlation reported humans date Our data reveal DGCs located inner portions human GCL show shorter less complex dendrites found outer portions layer presumably generated developmentally Moreover AD patients DGCs show early morphological alterations aggravated disease progresses An aberrantly increased number DGCs several primary apical dendrites first morphological change detected patients stages This alteration persists throughout AD progression leads generalized dendritic atrophy late stages disease Our data reveal distinct vulnerability several morphological characteristics DGCs located inner outer portions GCL AD support notion malfunction hippocampus related cognitive impairments patients AD

link

brain ad cognitive alzheimer disease neuronal neurological neurodegenerative neurons dementia

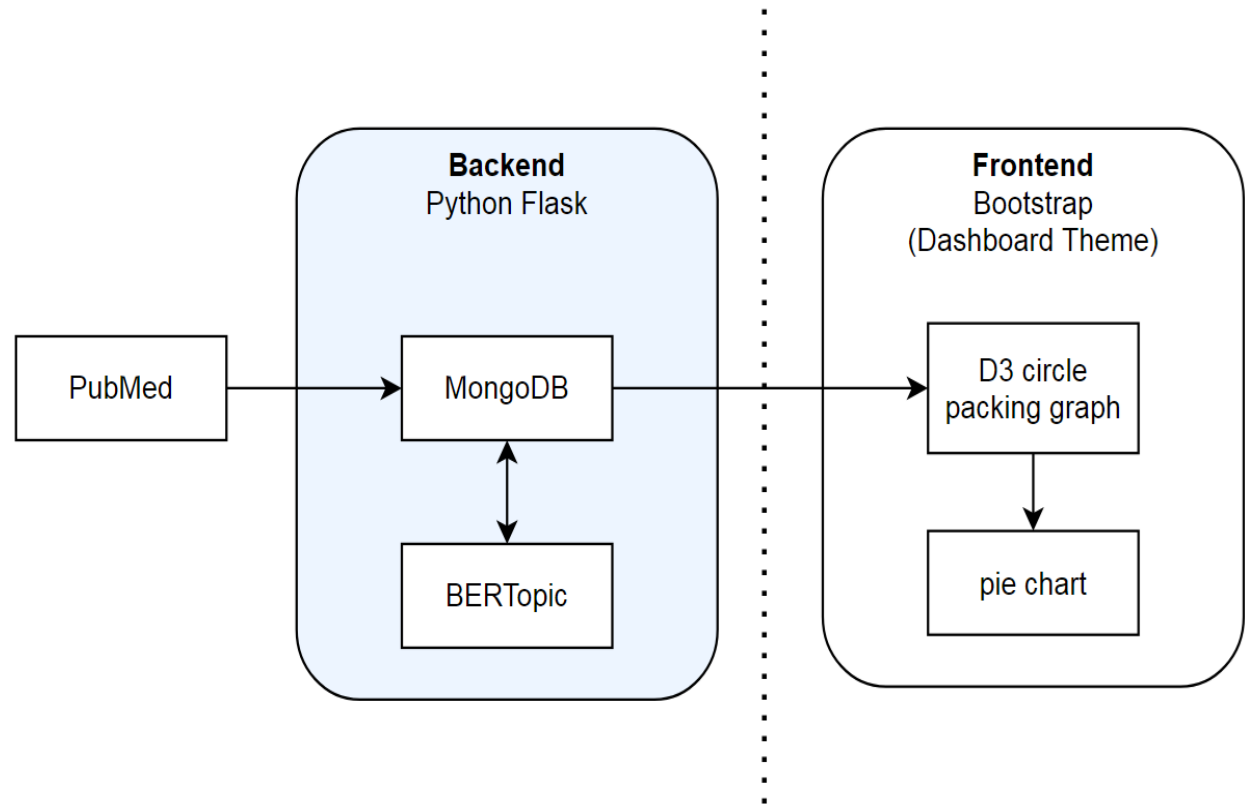
Alzheimer disease affects profoundly quality human behavior cognition The broad distribution severity across various human subjects requires quantitative diagnose Alzheimer disease beyond conventional tripartite classification cohorts cognitively normal CN mild cognitive impairment MCI Alzheimer disease AD The unfolding broad distributions quantitative continuous degree AD severity necessary precise diagnose study different stages conducted massive reanalysis MRI images human brains based accumulated bigdata Alzheimer Disease Neuroimaging Initiative The score matrix cortical thickness profile cortex points subjects constructed statistically learning cortical thickness data human brains We also developed new simple algebraic predictor provides quantitative continuous degree AD severity subjects along scale fully CN fully AD state The mathematical measure new predictor degree AD severity presented based covariance correlation matrix cortical thickness profile human subjects One remove uncertainty determination different stages

Data table including existing publications and the links

Existing documents within the categories related to the search query

Zoomable graph to find publications by categories

# Architecture



# Architecture

## Story behind the categories

<https://meshb.nlm.nih.gov/treeView>

### Medical Subject Headings 2022

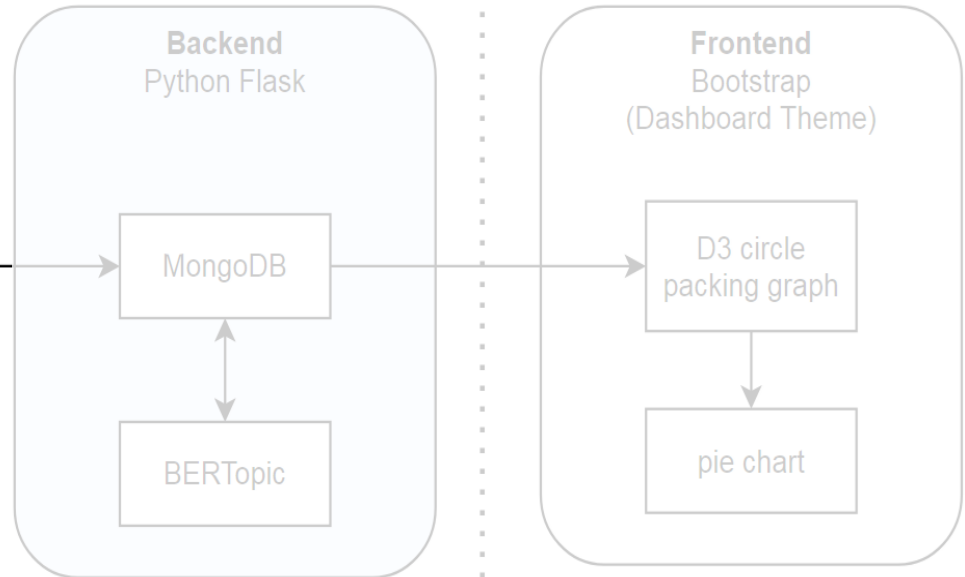
The files are updated each week day Monday-Friday by 8AM EST

Search MeSH... FullWord ▾

- All Terms
  - Main Heading (Descriptor) Terms
  - Qualifier Terms
  - Supplementary Concept Record Terms
- MeSH Unique ID
- Search in all Supplementary Concept Record Fields
  - Heading Mapped To
  - Indexing Information
- Pharmacological Action
- Search Related Registry and CAS Registry/EC Number/UNII Code/NCBI Taxonomy ID Number (RN)
  - Related Registry Search
  - CAS Registry/EC Number/UNII Code/NCBI Taxonomy ID Number (RN)
- Search in all Free Text Fields
  - Annotation
  - ScopeNote
  - SCR Note

- Anatomy [A] +
- Organisms [B] +
- Diseases [C] +
- Chemicals and Drugs [D] +
- Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E] +
- Psychiatry and Psychology [F] +
- Phenomena and Processes [G] +
- Disciplines and Occupations [H] +
- Anthropology, Education, Sociology, and Social Phenomena [I] +
- Technology, Industry, and Agriculture [J] +
- Humanities [K] +
- Information Science [L] +
- Named Groups [M] +
- Health Care [N] +
- Publication Characteristics [V] +
- Geographicals [Z] +

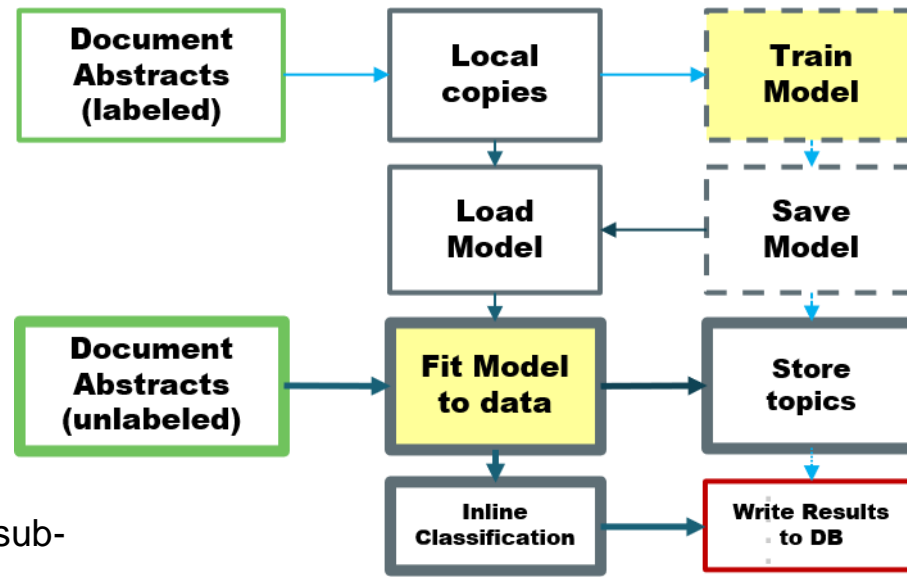
PubMed





# Architecture

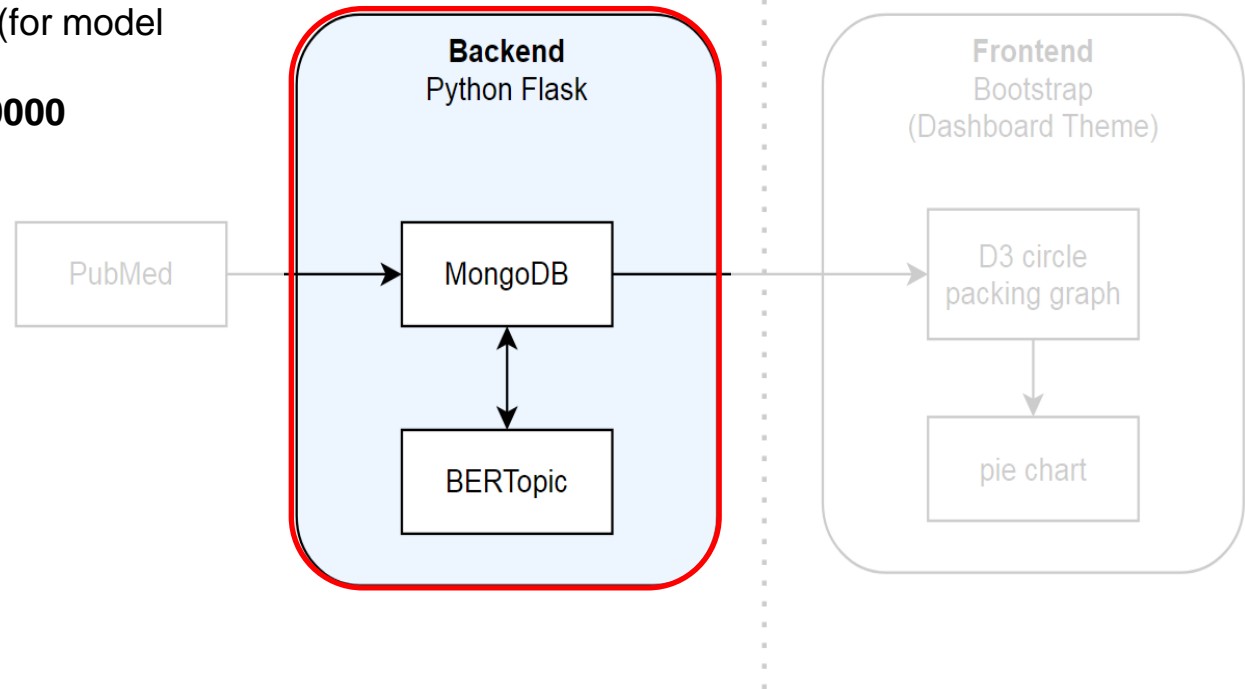
## Back-end model



Number of classes (categories and sub-categories) : **47**

Number of documents per category (for model training): **10000**

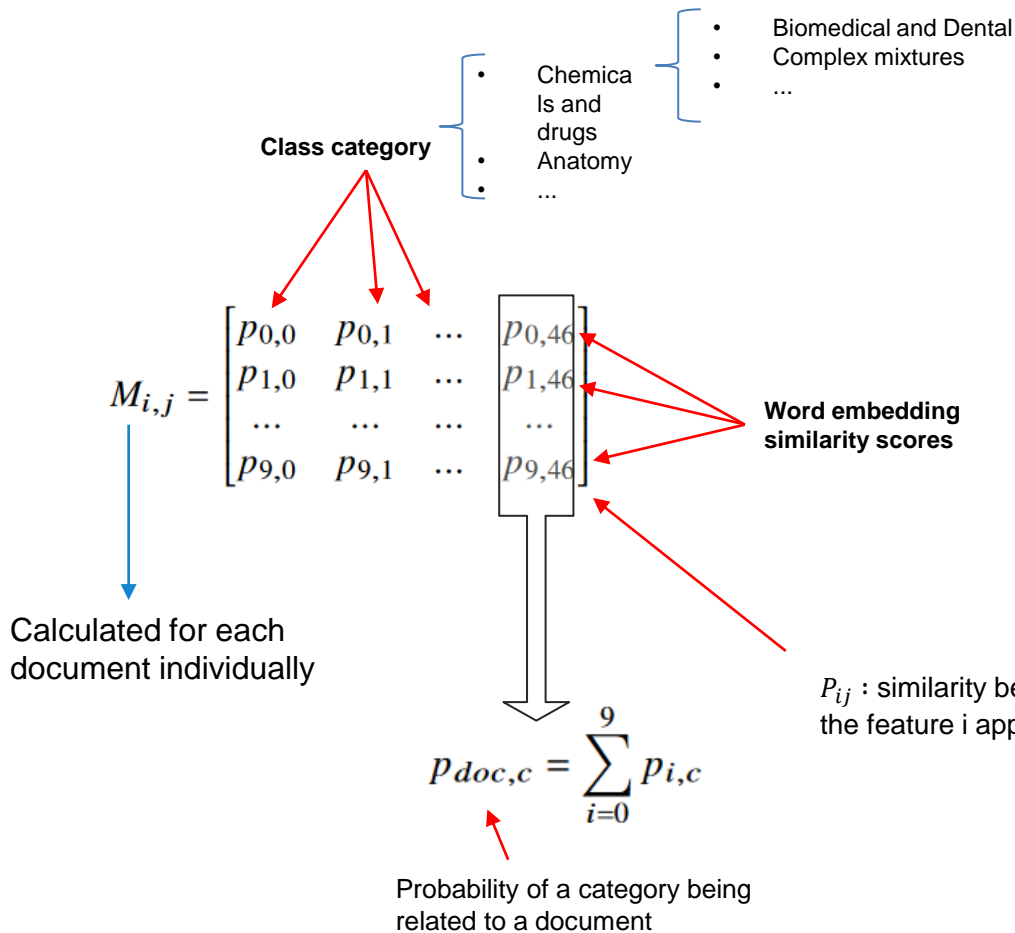
Number of unlabeled documents: **10000**



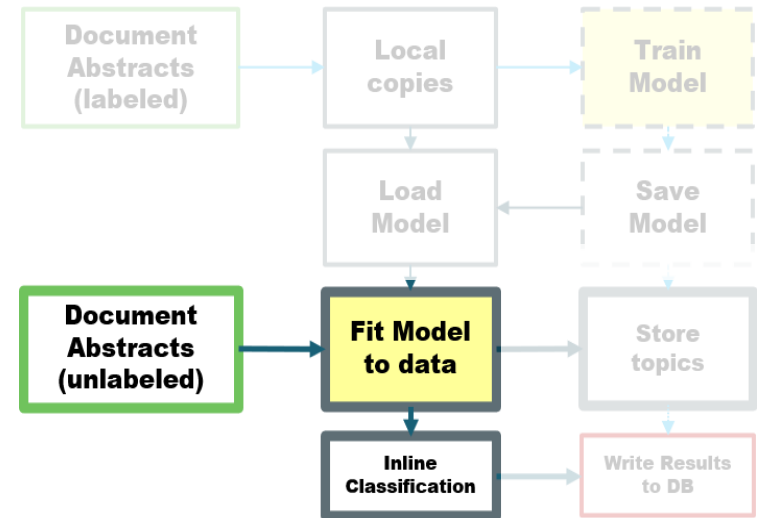
# Architecture

## Back-end model

### Similarity detection and classification



Top 10 Features are extracted from introduced categories using bertopic



# Architecture

## Front-end

### HOT

Health Optimization Tool

Skin and Connective Tissue Diseases

Find a Category or topic

Add New Topic And Or Not Clear Search

Page 1 < > 134

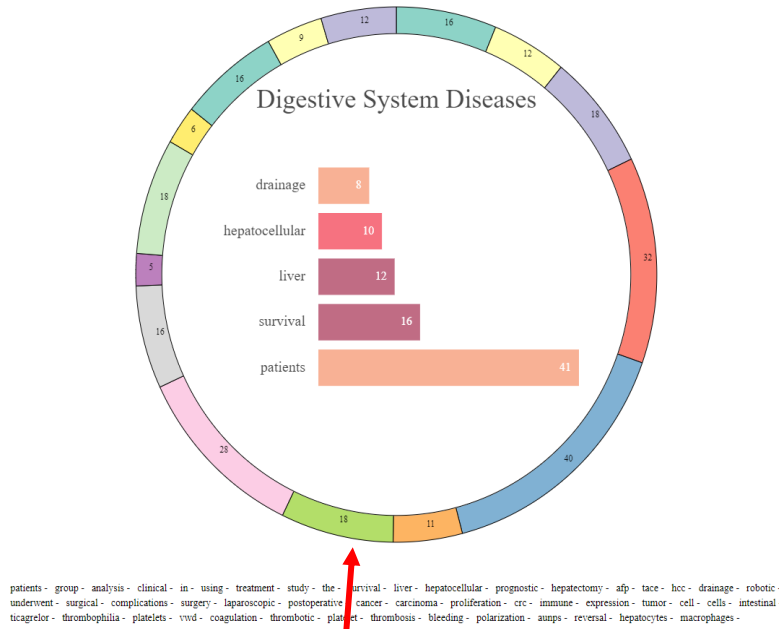
Efficacy transarterial chemoembolization TACE combined antiviral therapy AVT outcome hepatitis B HCC microvascular invasion MVI hepatic resection multicenter retrospective study conducted All patients divided four groups according postoperative adjuvant therapy control group AVT group TACE group combined group The overall survival OS survival RFS total patients enrolled study including control group TACE group AVT group combined group Multivariate Cox analysis showed postoperative adjuvant AVT TACE independent protective factors OS RFS The median OS among control group TACE group AVT group combined group months months months respectively p The median RFS among group months months months respectively p adjuvant TACE AVT independent protective factors associated mortality tumor recurrence HCC MVI resection This combined treatment strategy may provide useful clinical significance prevention tumor recurrence patients

link hcc survival tace liver afp hepatectomy prognostic hepatocellular group patients

This study aimed assess role expression Hepatocellular Carcinoma HCC study carried subjects classified three groups Group I healthy controls Group II patients liver cirrhosis Group III patients HCV related HCC The controls frequency matched based age sex groups All individuals subjected testing liver function AFP viral markers detected plasma using quantitative significantly upregulated whereas significantly downregulated HCC patients compared cirrhotic patients healthy controls There significant correlation AFP tumor size Receiver operating curve ROC analyses suggested plasma significant diagnostic value HCC sensitivity specificity A sensitivity specificity observed may play role disease progression development HCC may act potential biomarkers diagnosis HCC

link hcc survival tace liver afp hepatectomy prognostic hepatocellular group patients

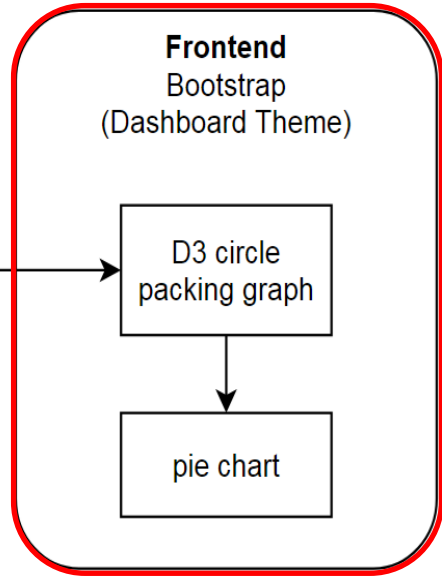
Hepatic macrophages regulate liver granuloma formation fibrosis caused infection Schistosoma japonicum manner regulation dependent macrophage activation state Interleukin IL may immunomodulatory effects macrophages However whether affect liver granuloma formation fibrosis affecting polarization macrophages japonicum infection remains unclear The aim study investigate macrophage polarization liver granuloma formation fibrosis japonicum



patients - group - analysis - clinical - in - using - treatment - study - the - survival - liver - hepatocellular - prognostic - hepatectomy - afp - tace - hcc - drainage - robotic - underwert - surgical - complications - surgery - laparoscopic - postoperative - cancer - carcinoma - proliferation - crc - immune - expression - tumor - cell - cells - intestinal - ticagrelor - thrombophilia - platelets - vwd - coagulation - thrombotic - platelet - thrombosis - bleeding - polarization - aumps - reversal - hepatocytes - macrophages -

Each individual color is assigned to a related category

BERTopic



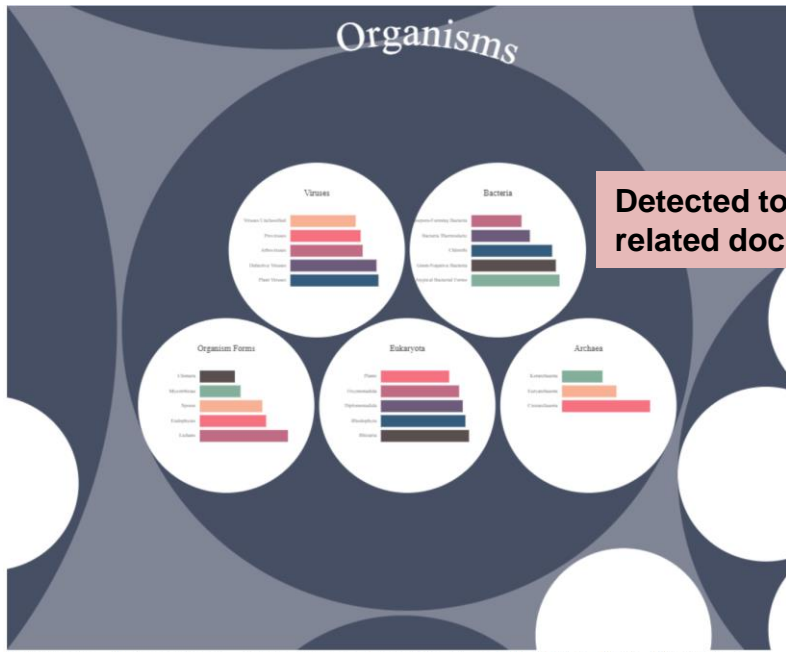
# Features in details

## Zoom-in/out to explore categories

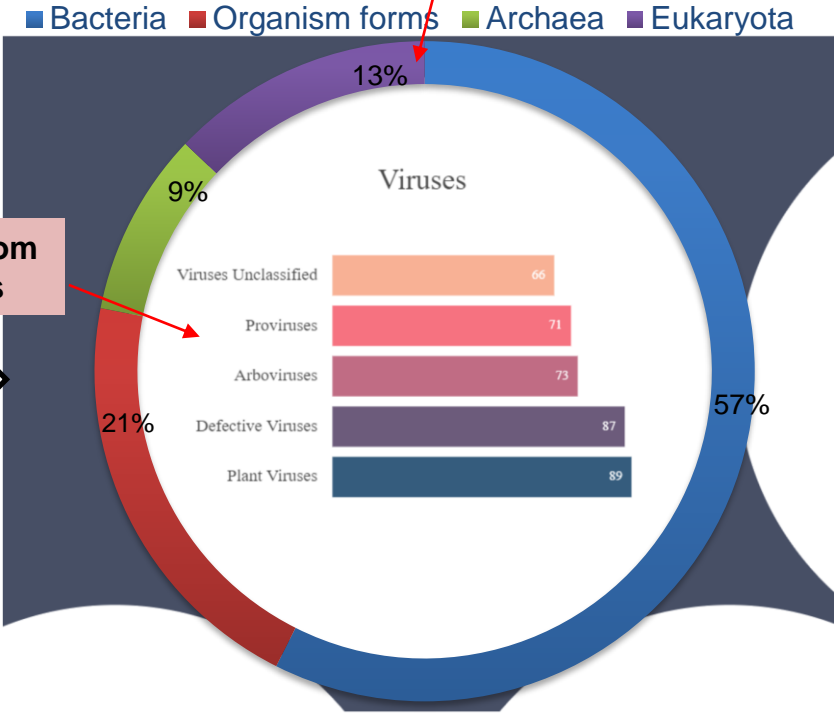
### Visualize relational information using a donut chart

- Provide additional information about the other related categories.
- Each piece of the donut represents the similarity of the selected category to another category based on mutual topics

Percentages of documents being related to other categories



Aquatic Organisms - Blood-Borne Pathogens - Chimera - Endophytes - Extremophiles - Hermaphroditic Organisms - Lichens - Life Cycle Stages - Microfilariae - Mycorrhizae - Organisms Genetically Modified - Sentinel Species - Spores -



Arboviruses - Archaeal Viruses - Bacteriophages - Defective Viruses - DNA Viruses - Fungal Viruses - Helper Viruses - Hepatitis Viruses - Insect Viruses - Oncogenic Viruses - Oncolytic Viruses - Plant Viruses - Provirus - Reassortant Viruses - RNA Viruses - Viroids - Viruses Unclassified -

Rest of the detected topics with less frequencies

# Evaluations

**Goal:** Assist pharma scientists in literature research and exploration

**Baseline:** Conventional methods : (e.g., online database search, in-house document search, ... )

**Users:** **Three** domain expert users

**Task:** Find results for given search queries functions

**Evaluation method:** Usability study using questionnaires and quantitative evaluation

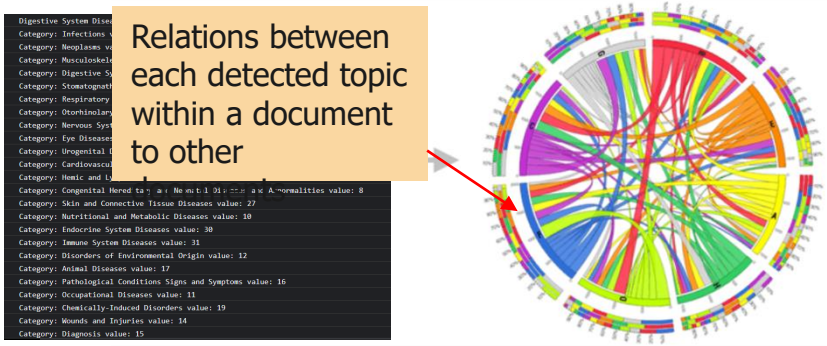
## Results:

1. All three users found existing features and the graph representation useful
2. Some users found it difficult to interact with the tool except the user who was in close collaboration with the tool development
3. Users found it highly useful to have manual guide for the tool
4. Users found it very useful to offer different solutions for navigation through the graph (e.g., mouse wheel)

# Literature research assistant future work

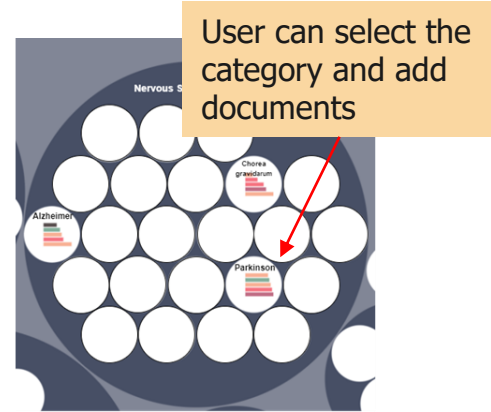
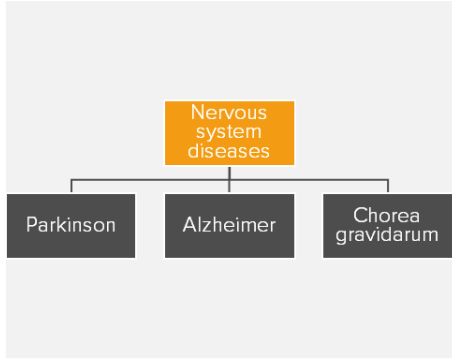
## Visualize relational information using a circus chart

- Visualize the relationship between categories
- Visualize which topic on the bar is related to other categories

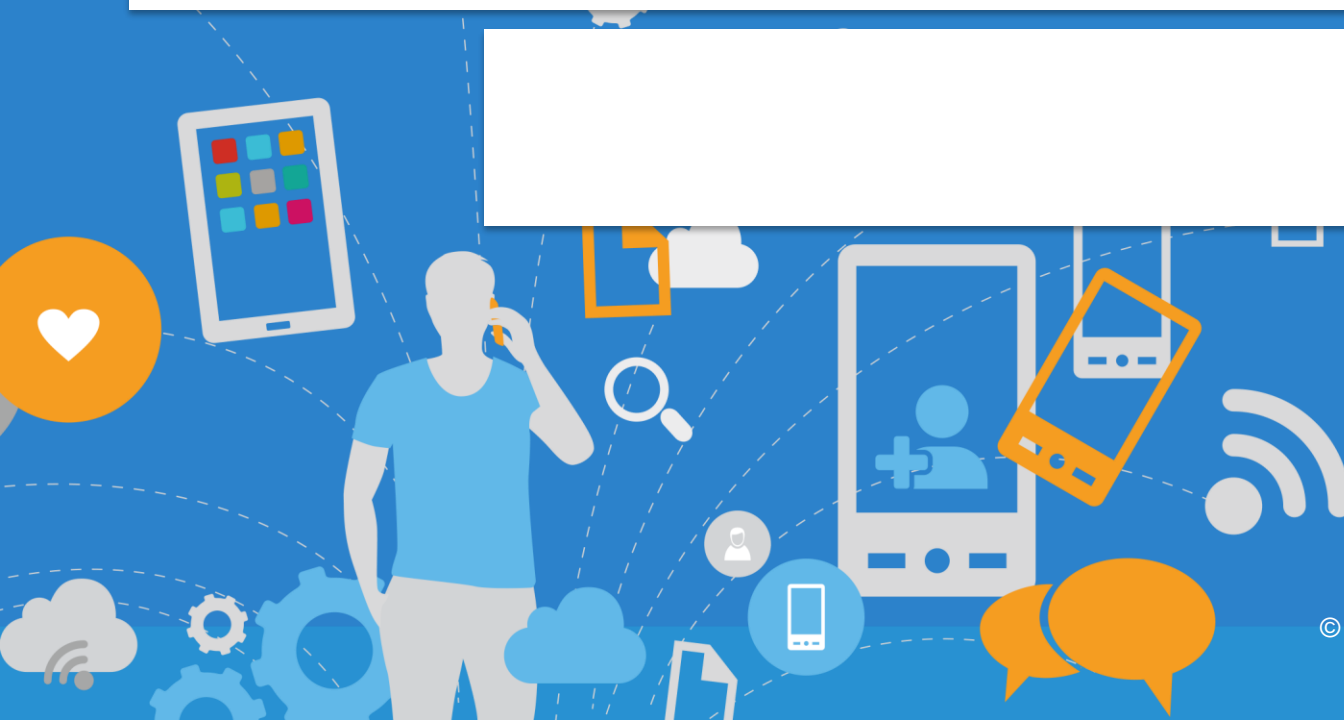


## Customize the classification by retraining the classifier

Users can upload batches of 100 classified documents to a desired class and retrain the model to update the target class definition



THANK YOU



# Features in details

## View other related categories

Health Optimization Tool

Skin and Connective Tissue Diseases

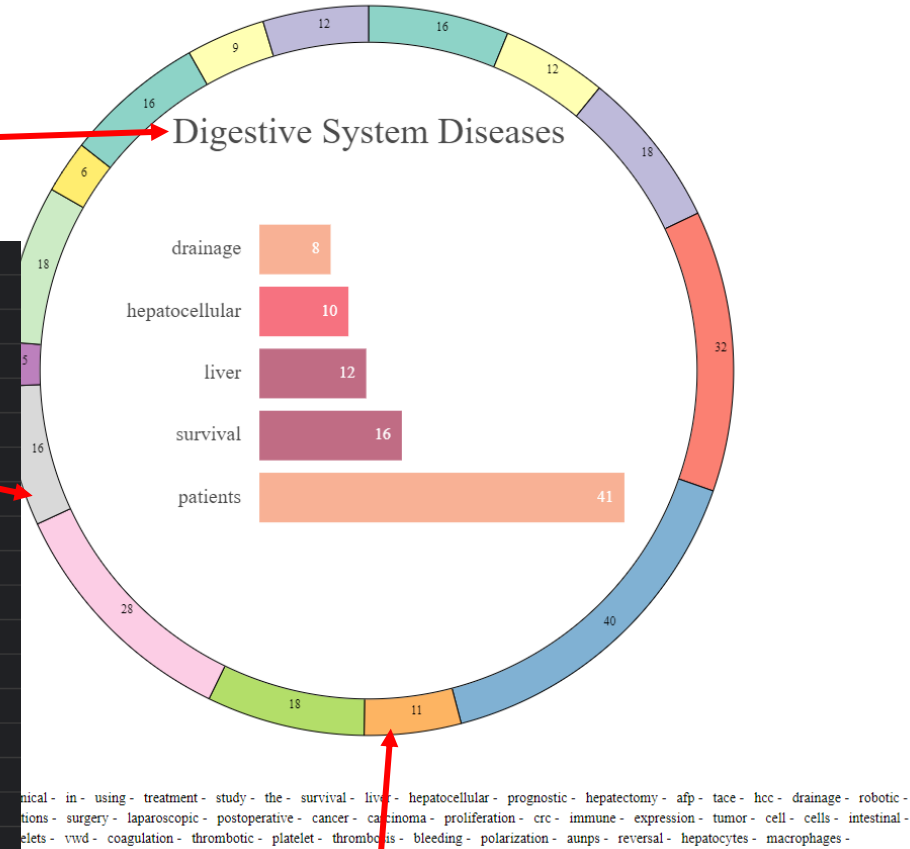
Find a Category or topic

**Selected category**

**Digestive System Diseases**

- Category: Infections value: 28
- Category: Neoplasms value: 28
- Category: Musculoskeletal Diseases value: 14
- Category: Digestive System Diseases value: 49
- Category: Stomatognathic Diseases value: 22
- Category: Respiratory Tract Diseases value: 21
- Category: Otorhinolaryngologic Diseases value: 21
- Category: Nervous System Diseases value: 7
- Category: Eye Diseases value: 11
- Category: Urogenital Diseases value: 14
- Category: Cardiovascular Diseases value: 13
- Category: Hemic and Lymphatic Diseases value: 31
- Category: Congenital Hereditary and Neonatal Diseases and Abnormalities value: 8
- Category: Skin and Connective Tissue Diseases value: 27
- Category: Nutritional and Metabolic Diseases value: 10
- Category: Endocrine System Diseases value: 30
- Category: Immune System Diseases value: 31
- Category: Disorders of Environmental Origin value: 12
- Category: Animal Diseases value: 17
- Category: Pathological Conditions Signs and Symptoms value: 16
- Category: Occupational Diseases value: 11
- Category: Chemically-Induced Disorders value: 19
- Category: Wounds and Injuries value: 14
- Category: Diagnosis value: 15

**Other related categories**



**Each individual color is assigned to a category**