Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

Columbia University Student Project Orientation

Sarangan Ravichandran
BIDS, FNLCR

July 14, 2020

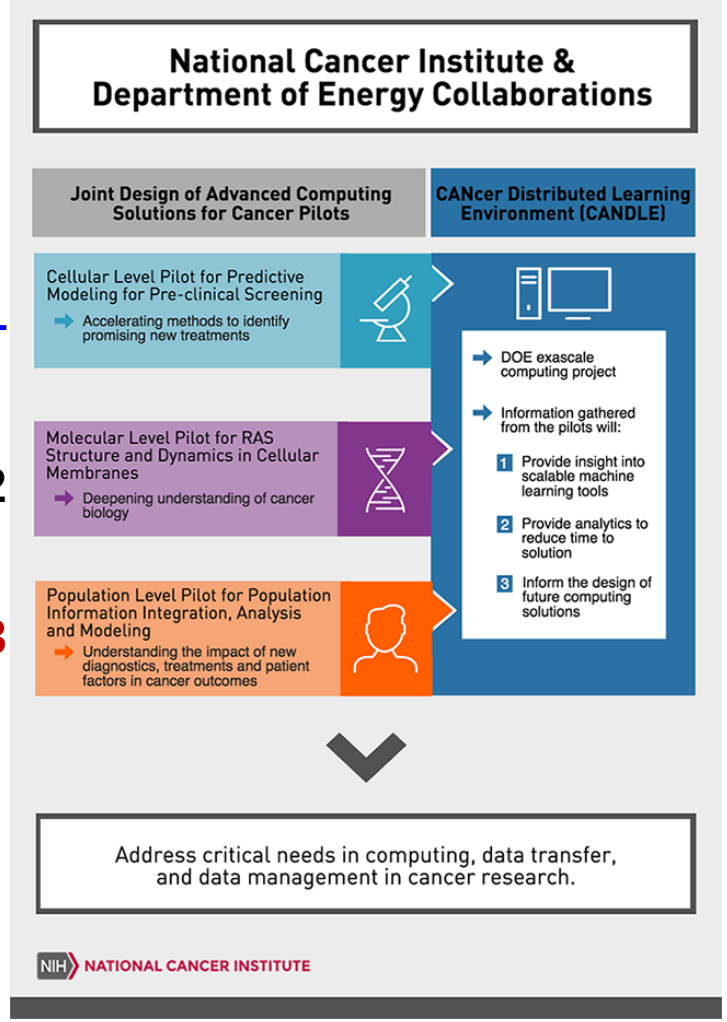
Introduction

- Eric Sthalberg, Ph.D.
  - Director, Biomedical Informatics and Data Science
- Naomi Ohashi,
  - Technical Project Manager
- Prof. Michael Robbins,
  - Columbia University
- Ravichandran Sarangan, Ph.D.,
  - Data scientist & 18 years of computational biology background

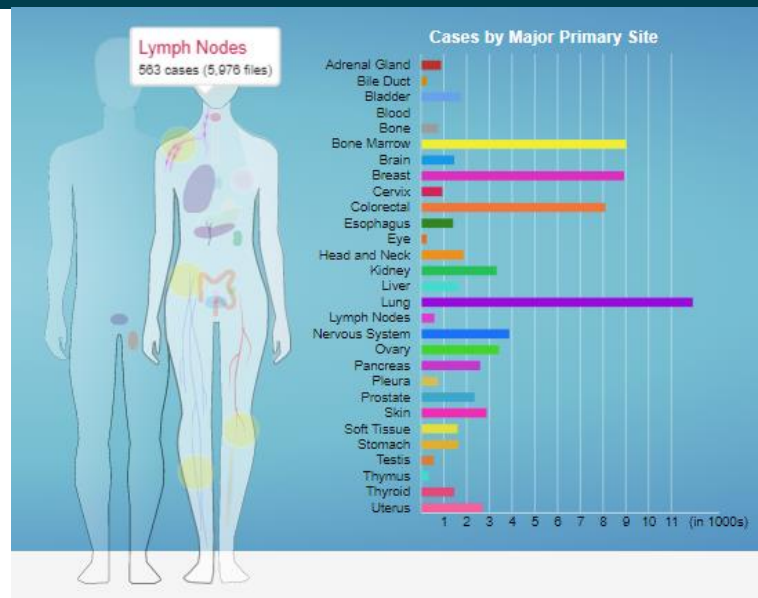
https://datascience.cancer.gov/collaborations/joint-design-advanced-computing

The Joint Design of Advanced Computing Solutions for Cancer (JDACS4C)

- JDACS4C program was created in 2016 to accelerate cancer research using emerging exascale computing capabilities.
- Part of the Cancer Moonshot
- Cross-agency collaboration between NCI and the DOE
- Pilot1:
  - *Focuses on developing predictive models, both computational and experimental, to improve pre-clinical therapeutic drug screening.*
  - https://datascience.cancer.gov/collaborations/joint-design-advanced-computing/cellular-pilot



Project-1 Overview: What Human Cancer Datasets (Biomolecular/Drug/Phenotype) are Available for Machine-Learning?



- Assignments
  - Each student 2 cancer areas
  - Report outcome & datasets
- Goal
  - Carry out literature search to identify ML related publications/datasets
- Deliverable(s)
  - Spreadsheet (summary, publication reference(s), source link, software …)
  - GitHub repository
  - Weekly meeting & final presentation/docs

Project-2 Overview: Survey to Identify Emerging Infectious Disease(s) Datasets for Machine-Learning

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

- Assignments
  - Each student infectious disease
  - Report outcome & datasets
- Goal
  - Literature search to identify ML related publications/datasets
- Deliverable(s)
  - Spreadsheet (summary, publication reference(s), source link, software …)
  - GitHub repository
  - Weekly meeting & final presentation/docs

Project-5 Overview: Cloud Deployment, Optimization Strategies for Teaching, Training and Collaborative Reproducible Research

- Assignments
  1. Identify top-5 software technologies (**ideal for one student**)
  2. After step 1, each student pick one tech. and find out product features and functions
  3. Compare technologies and report outcome (details on project document) (**Teamwork**)
- Goal
  - To identify cloud-sharable computing environments (free) and compare them to categorize (based on the progress of development, ease of use, support of GitHub/programming-languages etc.) the top five software
- Deliverable(s)
  - Create report on software comparison
  - GitHub repository
  - Weekly meeting & final presentation/docs

Cancer

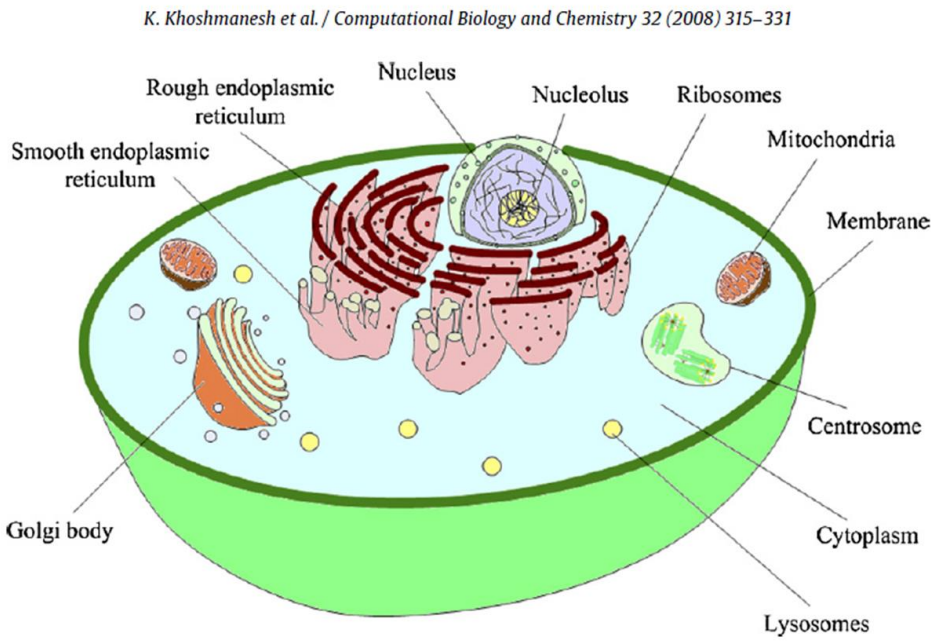
- https://www.cancer.gov/about-cancer/understanding/what-is-cancer

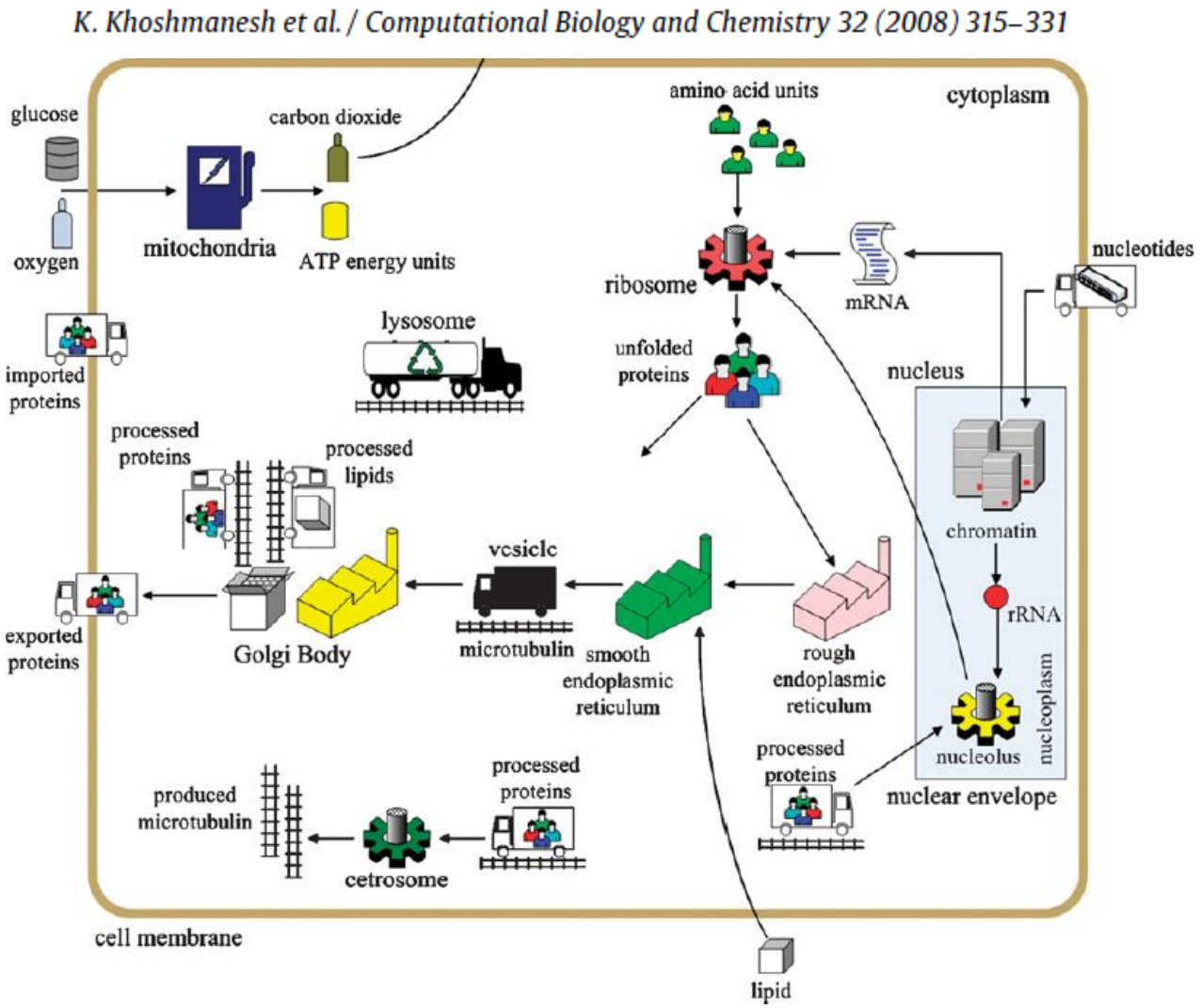
- "Cancer is the name given to a collection of related diseases. In all types of cancer, some of the body's cells begin to divide without stopping and spread into surrounding tissues." (quote from NCI website)
- Cancer
  - 100 types of cancer
  - Cells gain immortality
  - Spreading

Brief Biology Background (projects 1 & 2)

Roughly 37.2 trillion cells in our body

Typical cell (across length) 10 x 10^{-6}m

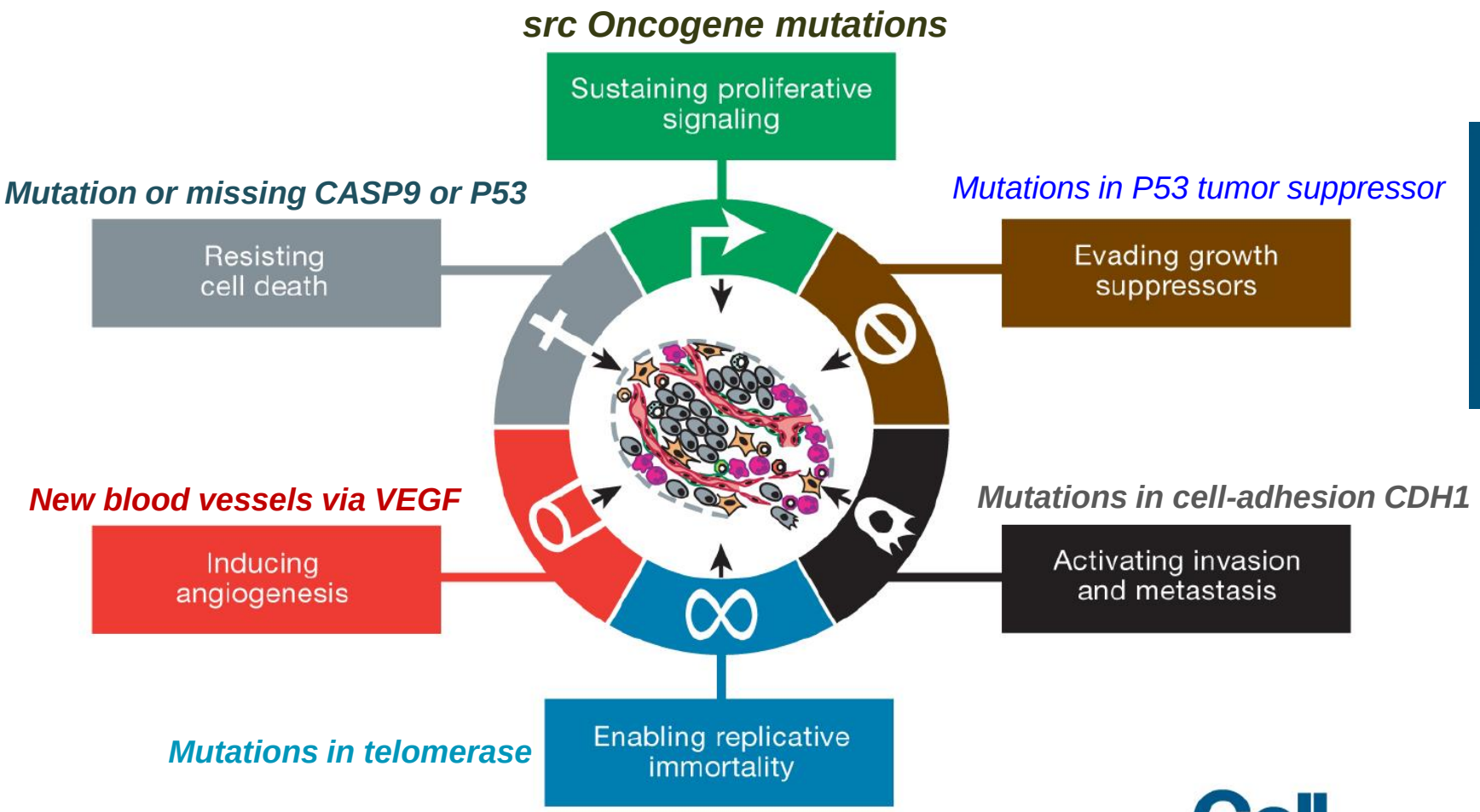




Brief Cancer Background (appropriate for projects 1 and 2)

Hallmarks of cancer: Integral Components of Most Forms of Cancer (Acquired Capabilities)

Hallmarks of Cancer: The Next Generation



Hanahan and Weinberg, 2011


REVIEW | VOLUME 100, ISSUE 1, P57-70, JANUARY 07, 2000

The Hallmarks of Cancer

Douglas Hanahan • Robert A Weinberg

Open Archive • DOI: https://doi.org/10.1016/S0092-8674(00)81683-9

Cell PRESS

Brief Reproducible Research & Documentation Background

Appropriate for project-5

Reproducibility

"More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments"

Is there a reproducibility crisis? M.Baker, Nature, 533, 452, 2016

Reproduce another scientist's experiments (failed to reproduce their own experiment)
Chemistry: 90% (60%)
Biology: 80% (60%)
Physics & Engineering: 70% (50%)
Medicine: 70% (60%)
Earth and Env. Science: 60% (40%)

Reproducibility/Teaching in Research

- Script availability via GitHub Notebook
  - Supplemental pages is a good place
  - Useful for checking the results
  - Useful for learning/teaching
  - Useful for reviewers
  - Etc.
- Converting static notebooks into dynamic and interactive

PubMed

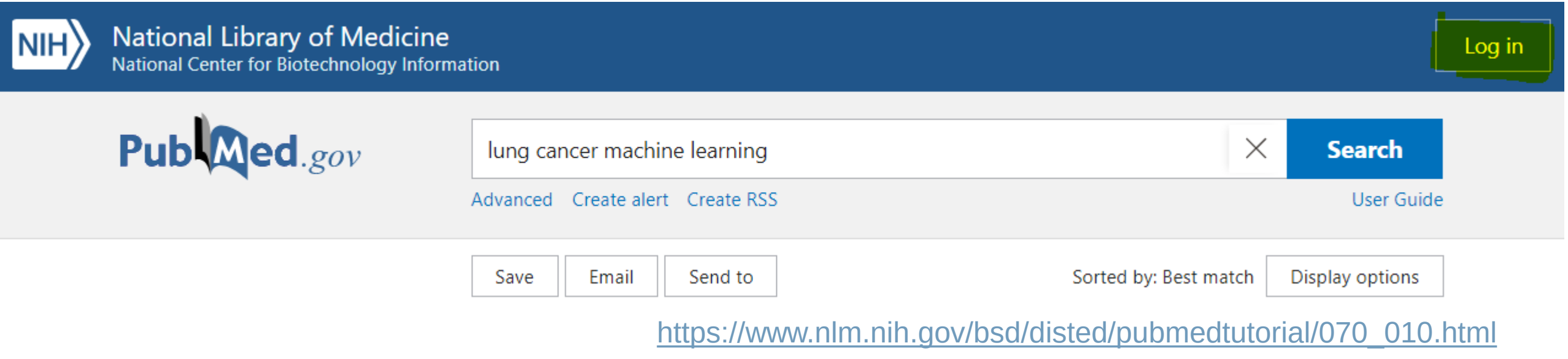
- Free resource

Medical Subject Headings (MeSH)

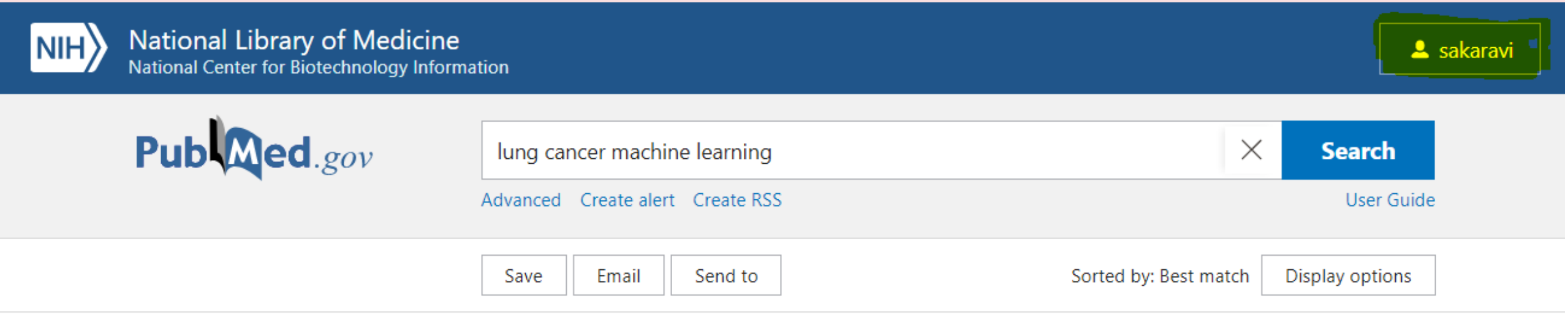
https://pubmed.ncbi.nlm.nih.gov/about/

- Made up of three components
  - **MEDLINE:** provides citations and indexed with MeSH terms
    - Access since 1996
  - **PubMed Central (PMC):** Full-article archive
  - **Bookshelf:** Full-text archive of book-chapters, reports and DBs related to biomedical sciences

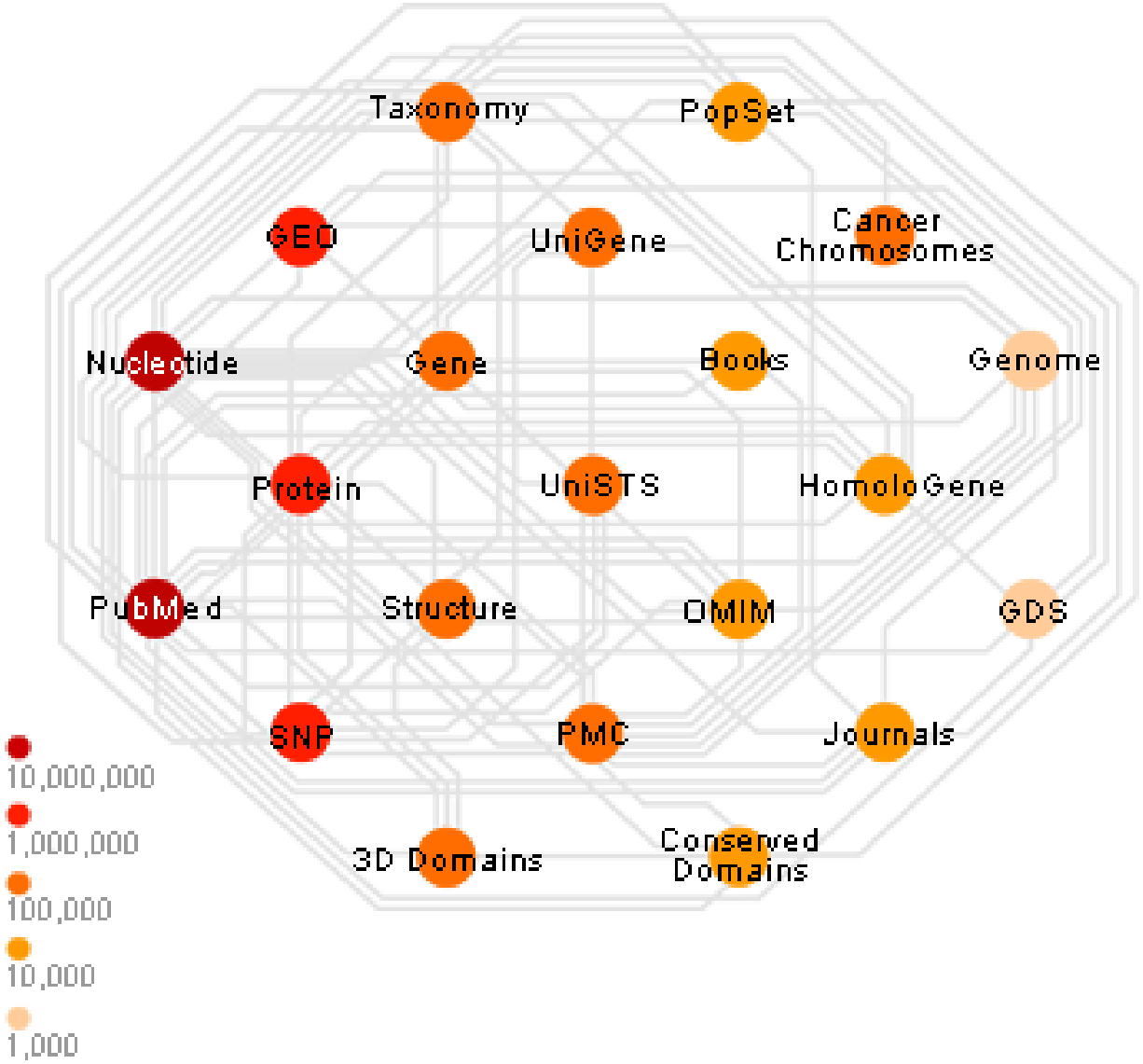
Create an account in My NCBI



https://www.nlm.nih.gov/bsd/disted/pubmedtutorial/070_010.html



PubMed is part of NCBI's vast retrieval system, known as Entrez.



PubMed Search

National Library of Medicine
National Center for Biotechnology Information

Log in

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

Search NCBI | Cancer | Search

Results by database
Results found in 33 databases

Literature	
Bookshelf	120,076
MeSH	395
NLM Catalog	50,857
PubMed	4,114,465
PubMed Central	1,805,061

Genes	
Gene	49,568
GEO DataSets	905,177
GEO Profiles	15,791,203
HomoloGene	227
PopSet	1,311

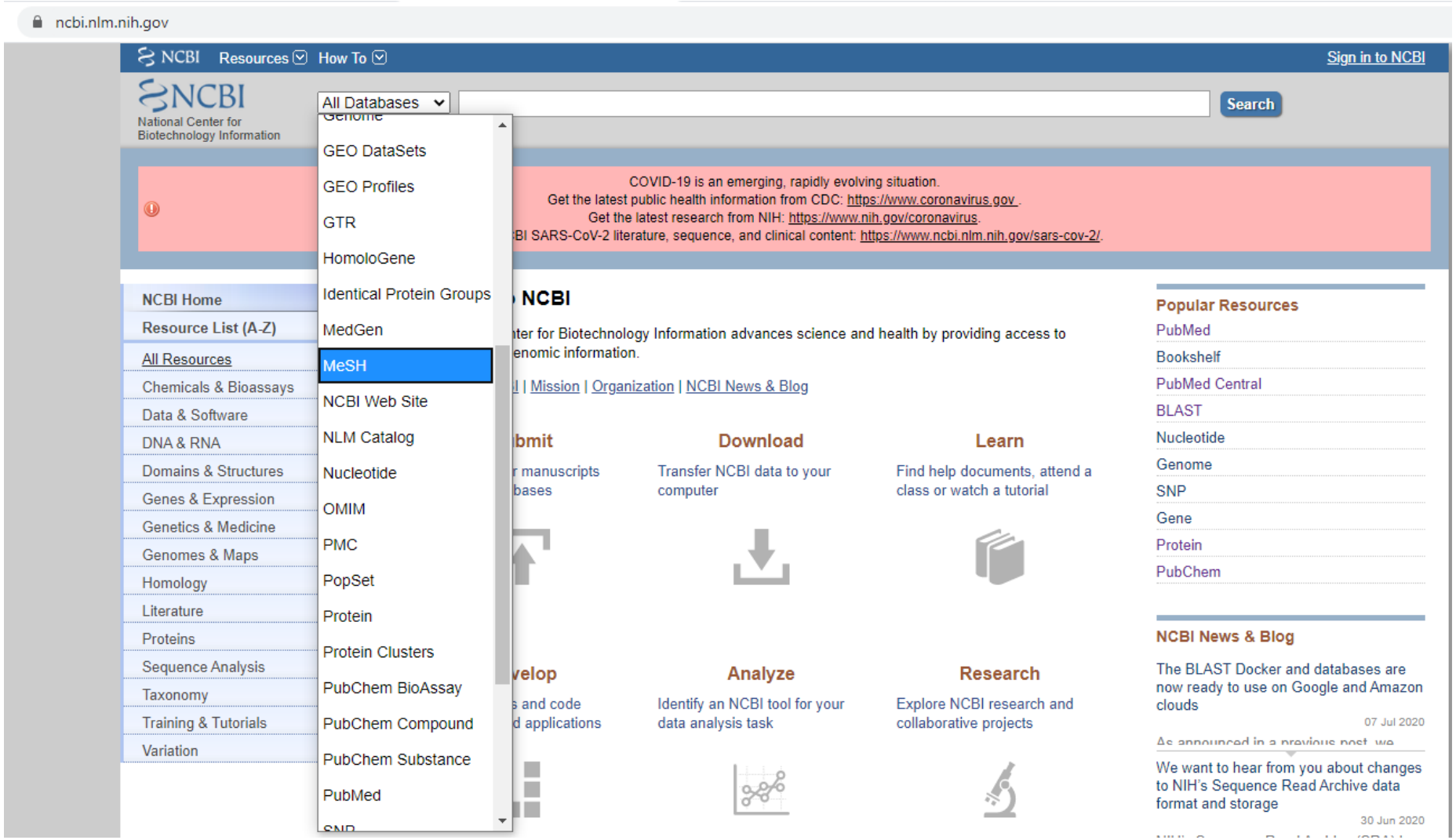
Proteins	
Conserved Domains	1,188
Identical Protein Groups	21,900
Protein	2,900,288
Protein Clusters	19
Sparcle	4,922
Structure	13,687

Genomes	
Assembly	2,513
BioCollections	0
BioProject	27,985
BioSample	1,015,742
Genome	22
Nucleotide	10,408,628
SRA	382,550
Taxonomy	1

Genetics	
ClinVar	129,791
dbGaP	632
dbSNP	0
dbVar	7,564
GTR	3,188
MedGen	7,273
OMIM	3,283

PubChem	
BioAssays	213,457
Compounds	6,238
Pathways	81
Substances	63,455

Accessing MeSH via NCBI



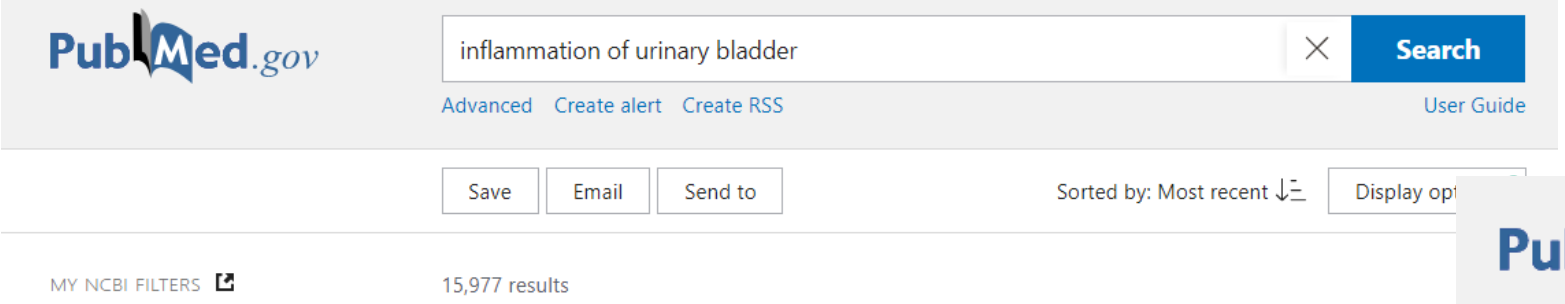
Medical Subject Headings (MeSH) in MEDLINE/PubMed

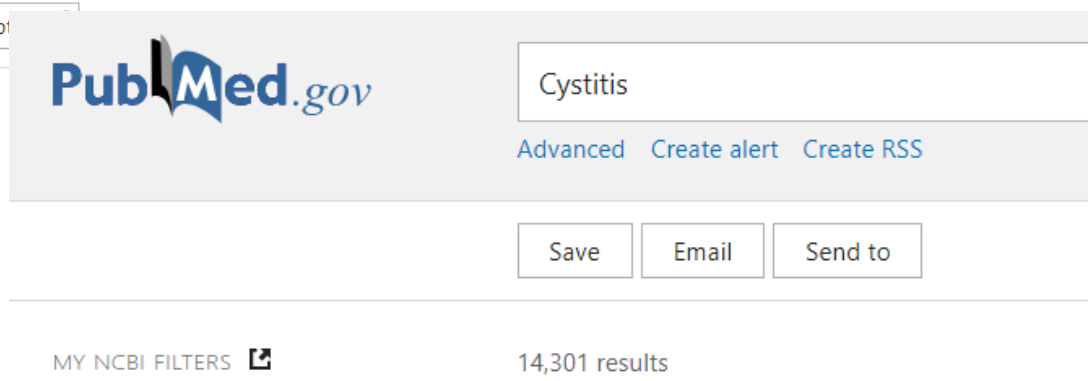
- "MEDLINE uses a controlled vocabulary, meaning that there is a specific set of terms used to describe each article."
- MeSH consists of
  - **Headings** (a concept in medical literature; Ex. eye-lashes)
    - Updated regularly
  - **Sub-headings**
    - DIAG used for diagnosis for disease associated papers
  - **Publication Types**
    - Articles or Reviews
  - **Supplementary Concept Records**
    - Topics discussed in the articles. For example, coq10, substances mentioned in articles

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

PubMed Search helps to identify associated MeSH terms

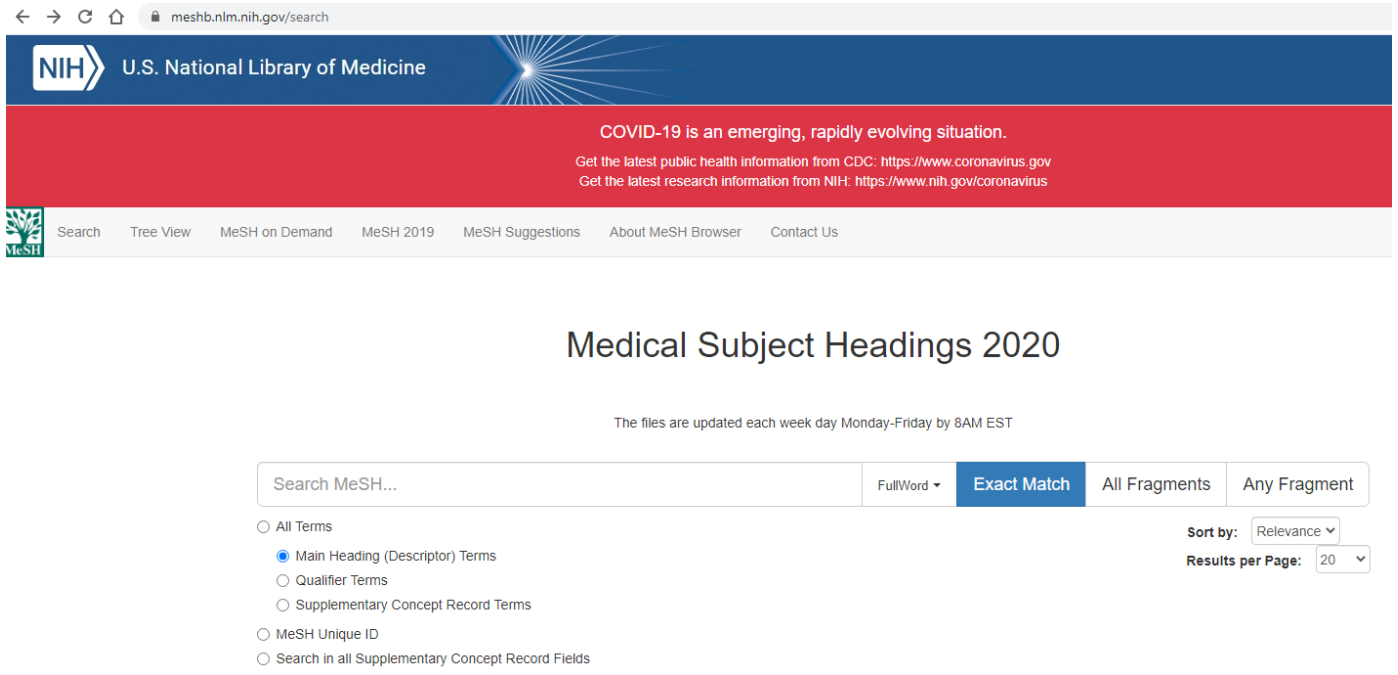
- Indexers assign MeSH terms to each article
  - This will provide specific entry points for search using PubMed
- It is useful to search for articles using MeSH terms
  - Example
    - Cystitis instead of “bladder diseases”





How to find MeSH terms?

https://meshb.nlm.nih.gov/search



https://meshb.nlm.nih.gov/search

- Search for “machine learning”

Machine Learning MeSH Descriptor Data 2020

Details | Qualifiers | MeSH Tree Structures | Concepts

MeSH Heading	Machine Learning
Tree Number(s)	G17.035.250.500
	L01.224.050.375.530
Unique ID	D000069550
RDF Unique Identifier	http://id.nlm.nih.gov/mesh/D000069550
Scope Note	A type of ARTIFICIAL INTELLIGENCE that enable COMPUTERS to independently initiate and execute LEARNING
Entry Term(s)	Transfer Learning
Public MeSH Note	2016; see ARTIFICIAL INTELLIGENCE 1997-2015
History Note	2016; use ARTIFICIAL INTELLIGENCE 1997-2015
Date Established	2016/01/01
Date of Entry	2015/07/10
Revision Date	2019/04/29

Machine Learning MeSH Descriptor Data 2020

Details | Qualifiers | MeSH Tree Structures | Concepts

- Mathematical Concepts [G17]
  - Algorithms [G17.035]
    - Artificial Intelligence [G17.035.250]
      - **Machine Learning [G17.035.250.500]**
        - Deep Learning [G17.035.250.500.250]
        - Supervised Machine Learning [G17.035.250.500.500]
        - Unsupervised Machine Learning [G17.035.250.500.750]

- Information Science [L01]
  - Computing Methodologies [L01.224]
    - Algorithms [L01.224.050]
      - Artificial Intelligence [L01.224.050.375]
        - Computer Heuristics [L01.224.050.375.095]
        - Expert Systems [L01.224.050.375.190]
        - Fuzzy Logic [L01.224.050.375.250]
        - Knowledge Bases [L01.224.050.375.480]
        - **Machine Learning [L01.224.050.375.530]**
          - Deep Learning [L01.224.050.375.530.250]
          - Supervised Machine Learning [L01.224.050.375.530.500]
          - Unsupervised Machine Learning [L01.224.050.375.530.750]
        - Natural Language Processing [L01.224.050.375.580]
        - Neural Networks, Computer [L01.224.050.375.605]
        - Robotics [L01.224.050.375.630]

Search Details

https://www.ncbi.nlm.nih.gov/mesh/2010029

Machine Learning

A type of ARTIFICIAL INTELLIGENCE that enable COMPUTERS to independently initiate and execute LEARNING when exposed to new data.

Year introduced: 2016

Introduced in 2016

PubMed search builder options

Subheadings:

- [ ] classification
- [ ] economics
- [ ] ethics
- [ ] history
- [ ] legislation and jurisprudence
- [ ] organization and administration
- [ ] standards
- [ ] statistics and numerical data
- [ ] trends

Terms commonly found with ML in publications

- [ ] Restrict to MeSH Major Topic.
- [ ] Do not include MeSH terms found below this term in the MeSH hierarchy.

Tree Number(s): G17.035.250.500, L01.224.050.375.530

MeSH Unique ID: D000069550

Entry Terms:

- Learning, Machine
- Transfer Learning
- Learning, Transfer

Synonyms; if you search using this term; appropriate MeSH will be included in your search

All MeSH Categories
- Phenomena and Processes Category
  - Mathematical Concepts
    - Algorithms
      - Artificial Intelligence
        - **Machine Learning**
          - Deep Learning
          - Supervised Machine Learning
            - Support Vector Machine
          - Unsupervised Machine Learning

All MeSH Categories
- Information Science Category
  - Information Science
    - Computing Methodologies
      - Algorithms
        - Artificial Intelligence
          - **Machine Learning**
            - Deep Learning
            - Supervised Machine Learning
              - Support Vector Machine
            - Unsupervised Machine Learning

Placed in two branches of the MeSH tree

Full

Send to:

Machine Learning

A type of ARTIFICIAL INTELLIGENCE that enable COMPUTERS to independently initiate and execute LEARNING when exposed to new data.
Year introduced: 2016

PubMed search builder options

Subheadings:

- [x] classification
- [ ] economics
- [ ] ethics
- [ ] history
- [ ] legislation and jurisprudence
- [ ] organization and administration
- [ ] standards
- [ ] statistics and numerical data
- [ ] trends

PubMed Search Builder

```
"Machine Learning/classification"
[Mesh]
```

Add to search builder | AND

Search PubMed

YouTube Tutorial

Related information

PubMed.gov

"Machine Learning/classification"[Mesh]

Search

Advanced Create alert Create RSS User Guide

Save Email Send to

Sorted by: Most recent

Display options

MY NCBI FILTERS

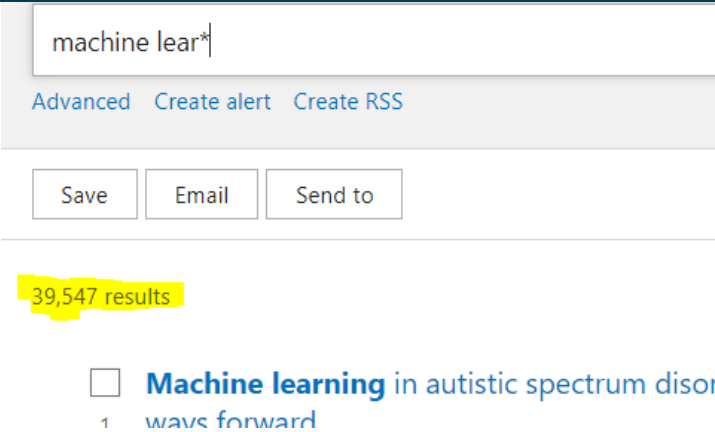
31 results

RESULTS BY YEAR

Classification of Current Procedural Terminology Codes from Electronic Health

How do I search PubMed?

- Be specific
- For initial searches, don't use
  - punctuation/quotation marks).
    - PubMed will find phrases for you.
- Improve later searches, use operators (e.g., AND, OR; note capital letters)
    - PubMed will add logical operators between concepts.
  - Use no tags
    - PubMed will differentiate topic words, journal titles and author names.
  - You can use wildcards
    - Machine lear* (will fine "Machine learning" but also find "machine learn"



Searching PubMed

PubMed.gov

lung cancer machine learning

Advanced Create alert Create RSS

Save Email Send to

MY NCBI FILTERS

722 results

lung cancer AND machine learning

Advanced Create alert Create RSS

Save Email Send to

720 results

"lung cancer" "machine learning"

Advanced Create alert Create RSS

Save Email Send to

333 results

"Lung Neoplasms"[Mesh] "machine learning"[Mesh]

Search

Advanced Create alert Create RSS

User Guide

Save Email Send to

Sorted by: Most recent

Display options

301 results

Searching PubMed

PubMed.gov

lung cancer machine learning

Advanced Create alert Create RSS

Save Email Send to

Search: **lung cancer machine learning** Sort by: **Most Recent**

((("lung neoplasms"[MeSH Terms] OR ("lung"[All Fields] AND "neoplasms"[All Fields])) OR "lung neoplasms"[All Fields]) OR ("lung"[All Fields] AND "cancer"[All Fields])) OR "lung cancer"[All Fields]) AND (("machine learning"[MeSH Terms] OR ("machine"[All Fields] AND "learning"[All Fields])) OR "machine learning"[All Fields])

Translations

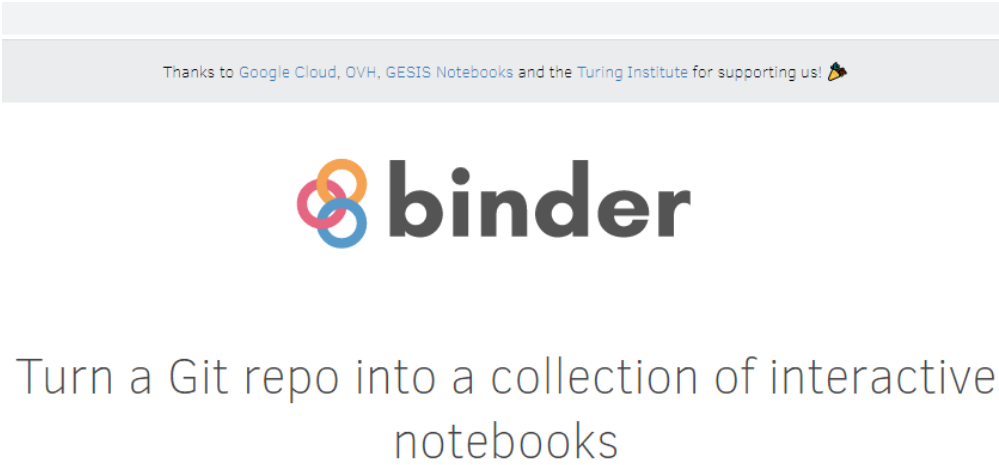
lung cancer: "lung neoplasms"[MeSH Terms] OR ("lung"[All Fields] AND "neoplasms"[All Fields]) OR "lung neoplasms"[All Fields] OR ("lung"[All Fields] AND "cancer"[All Fields]) OR "lung cancer"[All Fields]

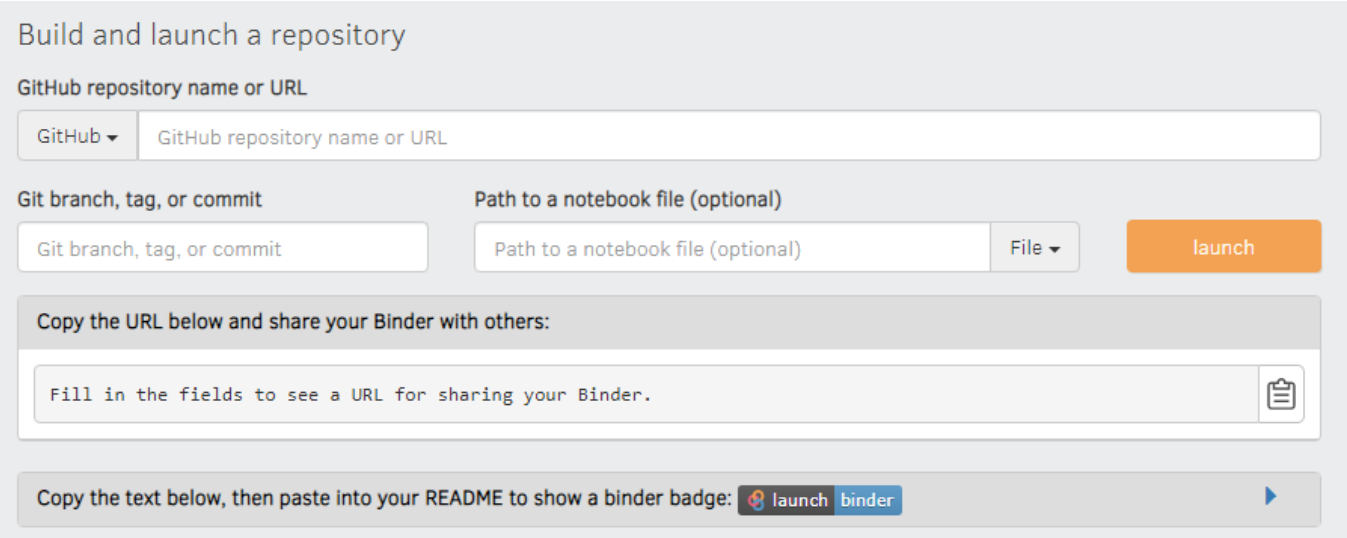
machine learning: "machine learning"[MeSH Terms] OR ("machine"[All Fields] AND "learning"[All Fields]) OR "machine learning"[All Fields]

Project-5

- Live demo using one of the software technology called BINDER
- We will use the sample GitHub page for demo
- https://github.com/ravichas/ML-predict-drugclass

https://mybinder.org/





Project-5

- ?s
- What other server/software can turn github into dynamic/interactive notebooks?
  - Notebooks: could be R, Python, Julia etc.
- How easy?
- How fast?
- How can we optimize it?
  - Binder YML file can be used to tweak this option
- What support these software provide?
- ….
- …
- Comparison in the form of report

Helpful links

- Projects-1 and 2:
  - https://www.nlm.nih.gov/bsd/disted/pubmedtutorial/015_010.html
  - https://learn.nlm.nih.gov/documentation/training-packets/T0042010P/
  - https://jamanetwork.com/journals/jama/article-abstract/369515 (MeSH)
- Project-5:
  - BINDER online software: https://mybinder.org/
  - Google's COLAB: https://colab.research.google.com/notebooks/intro.ipynb#recent=true
  - https://www.dataschool.io/cloud-services-for-jupyter-notebook/ (helpful site for our project)
  - https://github.com/jupyterhub/binderhub

Thank you!

Questions/Comments

S. Ravichandran
ravichandrans@mail.nih.gov

Naomi Ohashi
naomi.ohashi@nih.gov