



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

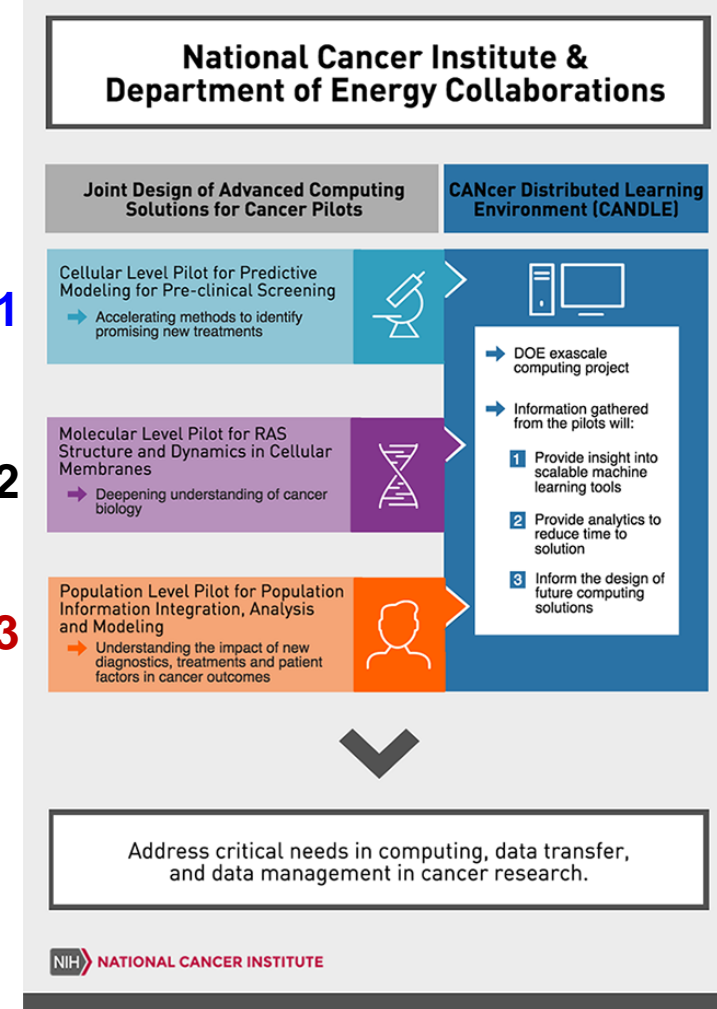
The Joint Design of Advanced Computing Solutions for Cancer (JDACS4C)

- 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>

Pilot1

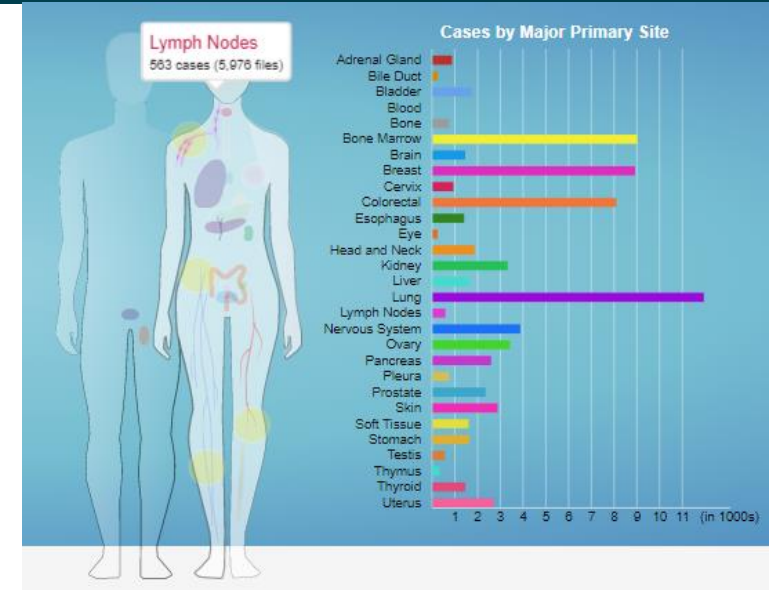
Pilot2

Pilot3



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

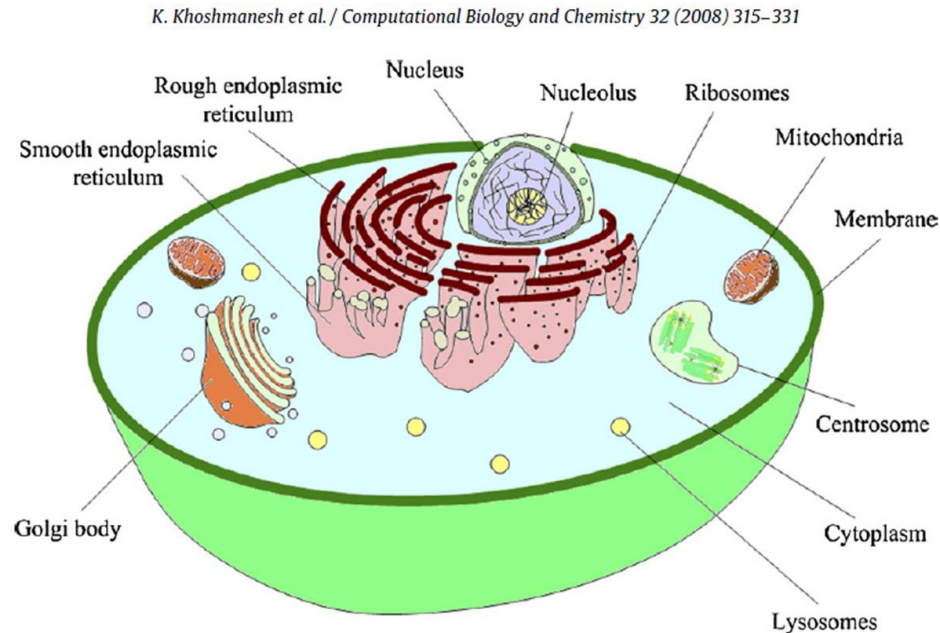
Cancer

- <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

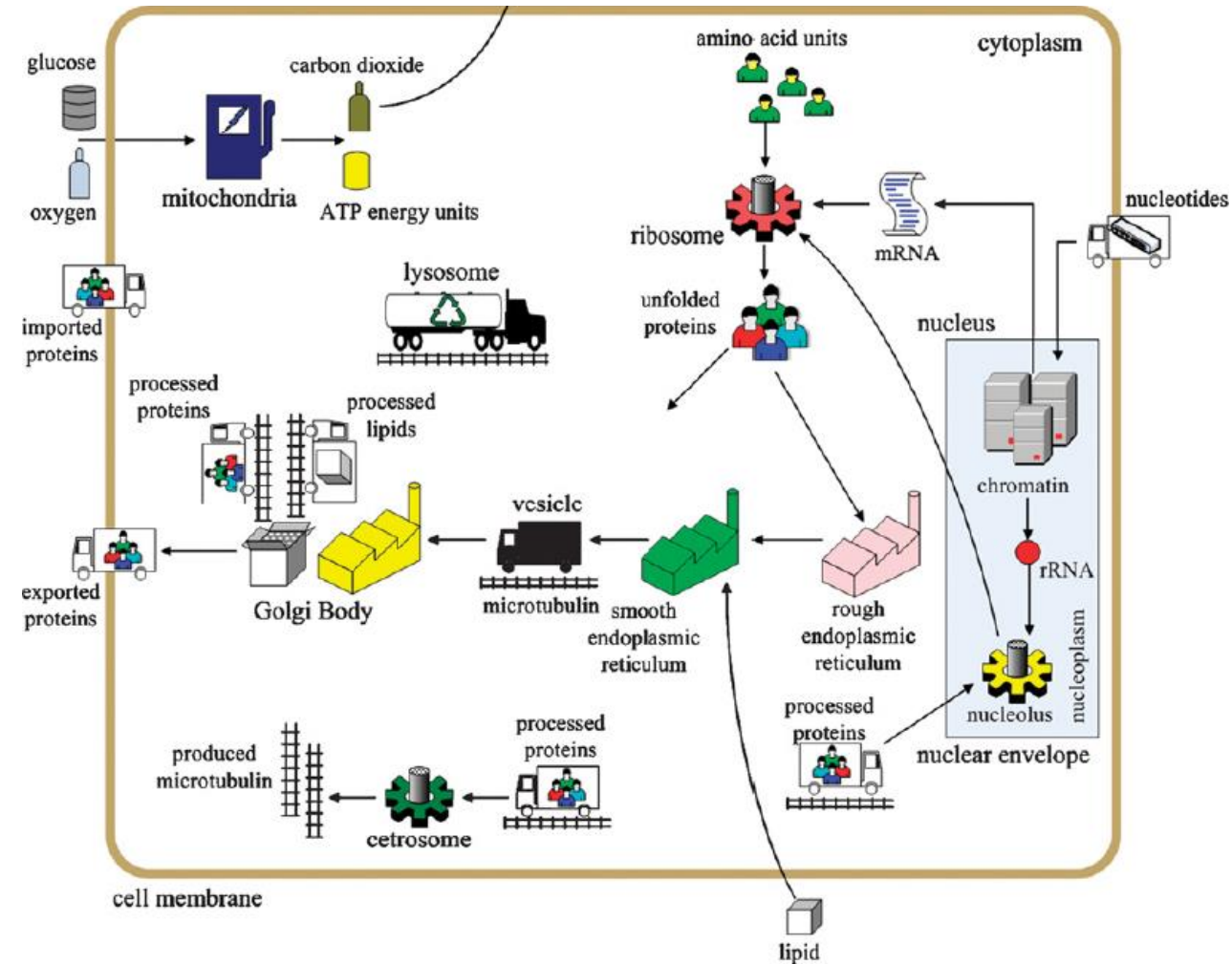
Brief Biology Background (projects 1 & 2)

Roughly 37.2 trillion cells in our body

Typical cell (across length) $10 \times 10^{-6}\text{m}$



K. Khoshmanesh et al. / Computational Biology and Chemistry 32 (2008) 315–331



Brief Cancer Background (appropriate for projects 1 and 2)

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

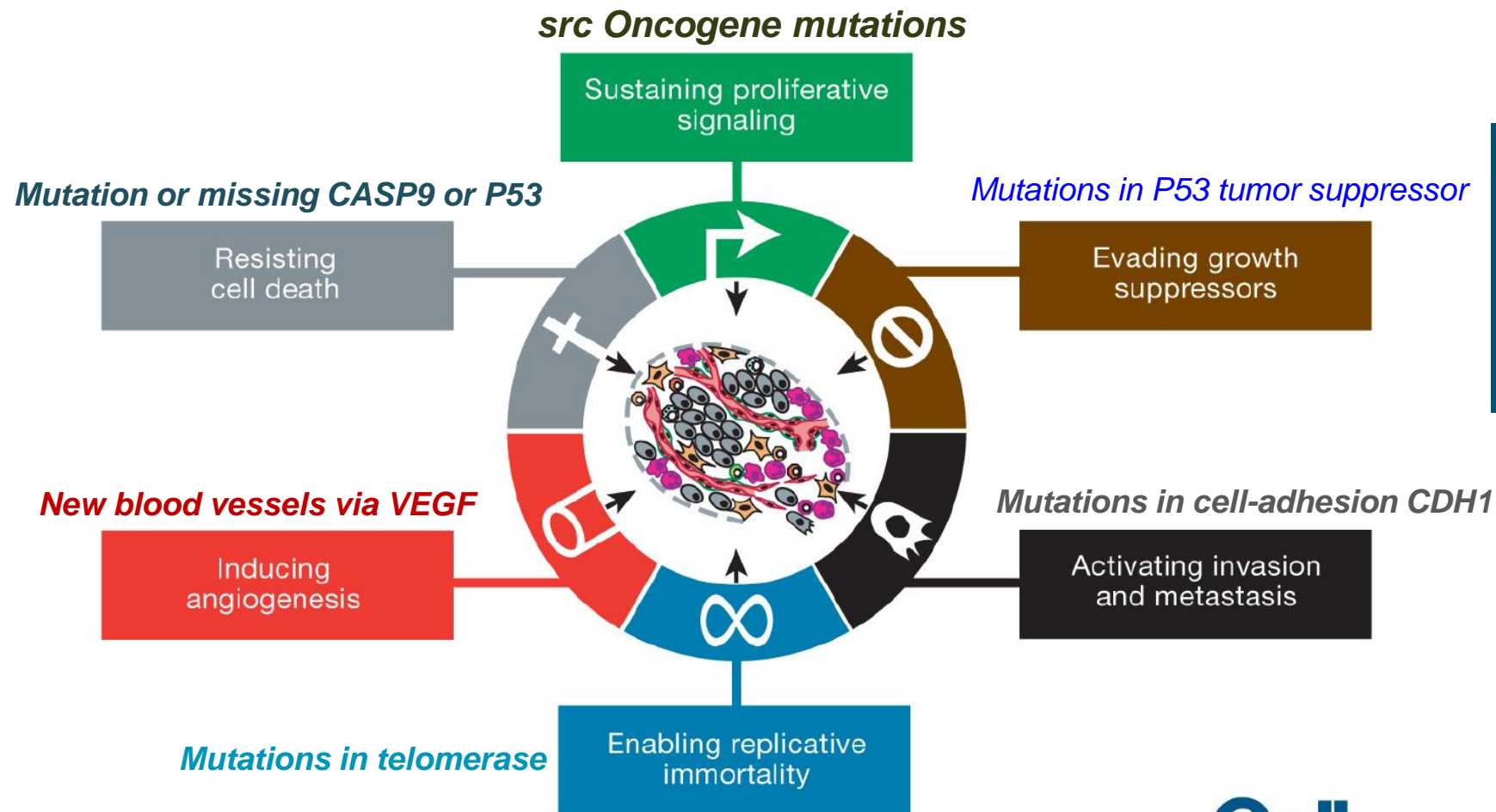
Hallmarks of Cancer: The Next Generation

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](https://doi.org/10.1016/S0092-8674(00)81683-9)



Hanahan and Weinberg, 2011

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

National Library of Medicine
National Center for Biotechnology Information



lung cancer machine learning

×

Search

[Advanced](#) [Create alert](#) [Create RSS](#) [User Guide](#)

Save


Email


Send to

Sorted by: Best match

Display options

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

National Library of Medicine
National Center for Biotechnology Information



lung cancer machine learning

×

Search

[Advanced](#) [Create alert](#) [Create RSS](#) [User Guide](#)

Save

Email

Send to

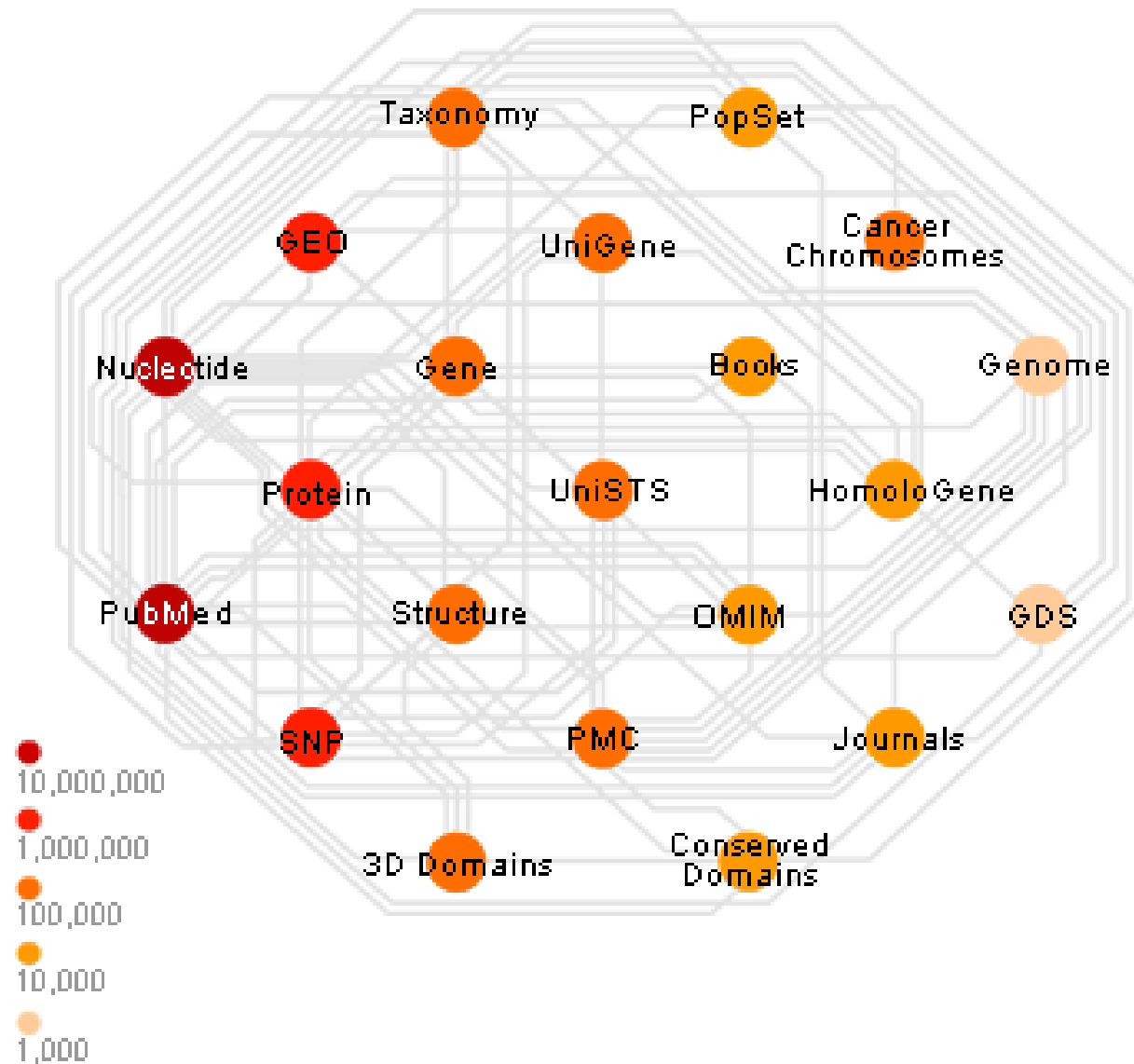
Sorted by: Best match

Display options


Log in

sakaravi

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	Genes	Proteins
Bookshelf 120,076	Gene 49,568	Conserved Domains 1,188
MeSH 395	GEO DataSets 905,177	Identical Protein Groups 21,900
NLM Catalog 50,857	GEO Profiles 15,791,203	Protein 2,900,288
PubMed 4,114,465	HomoloGene 227	Protein Clusters 19
PubMed Central 1,805,061	PopSet 1,311	Sparcle 4,922
		Structure 13,687
Genomes	Genetics	PubChem
Assembly 2,513	ClinVar 129,791	BioAssays 213,457
BioCollections 0	dbGaP 632	Compounds 6,238
BioProject 27,985	dbSNP 0	Pathways 81
BioSample 1,015,742	dbVar 7,564	Substances 63,455
Genome 22	GTR 3,188	
Nucleotide 10,408,628	MedGen 7,273	
SRA 382,550	OMIM 3,283	
Taxonomy 1		

Accessing MeSH via NCBI

The screenshot shows the NCBI website (ncbi.nlm.nih.gov) with the 'All Databases' dropdown menu open. The 'MeSH' option is highlighted in blue. The main navigation bar includes 'NCBI', 'Resources', and 'How To'. A search bar is located on the right. A red banner at the top provides COVID-19 information with links to CDC and NIH resources. The left sidebar lists various NCBI resources, including 'NCBI Home', 'Resource List (A-Z)', 'All Resources', 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', 'Domains & Structures', 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'. The main content area features a 'MeSH' link, a 'Download' button, and a 'Learn' button. The right sidebar lists 'Popular Resources' (PubMed, Bookshelf, PubMed Central, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem) and 'NCBI News & Blog' (The BLAST Docker and databases are now ready to use on Google and Amazon clouds, 07 Jul 2020).

ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases
Genome
GEO DataSets
GEO Profiles
GTR
HomoloGene
Identical Protein Groups
MedGen
MeSH
NCBI Web Site
NLM Catalog
Nucleotide
OMIM
PMC
PopSet
Protein
Protein Clusters
PubChem BioAssay
PubChem Compound
PubChem Substance
PubMed
SNP

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>.
NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

NCBI
National Center for Biotechnology Information advances science and health by providing access to genomic information.
[Home](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit
Upload manuscripts and code to public databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Build tools and code for data analysis applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources
[PubMed](#)
[Bookshelf](#)
[PubMed Central](#)
[BLAST](#)
[Nucleotide](#)
[Genome](#)
[SNP](#)
[Gene](#)
[Protein](#)
[PubChem](#)

NCBI News & Blog
The BLAST Docker and databases are now ready to use on Google and Amazon clouds
07 Jul 2020
As announced in a previous post, we
We want to hear from you about changes to NIH's Sequence Read Archive data format and storage
30 Jun 2020

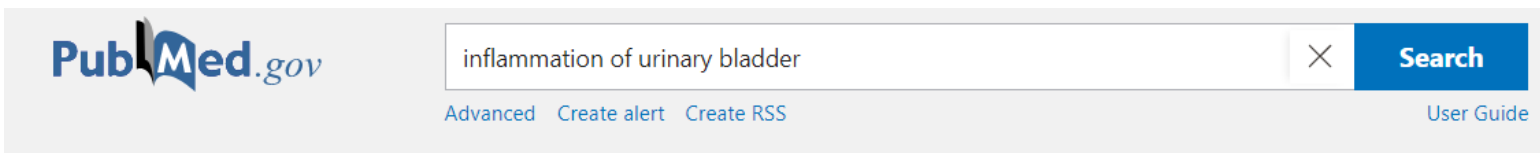
Medical Subject Headings (MeSH) in MEDLINE/PubMed

- “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”



PubMed.gov

inflammation of urinary bladder

Search

[Advanced](#) [Create alert](#) [Create RSS](#) [User Guide](#)

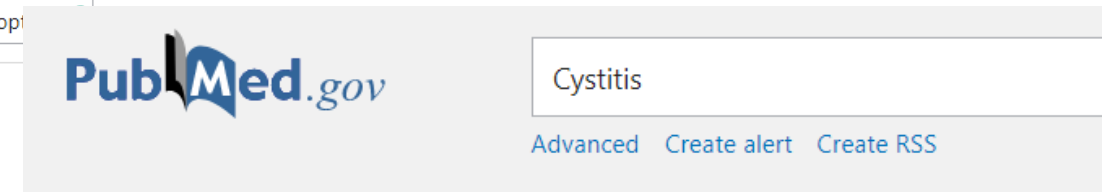
Save Email Send to

Sorted by: Most recent ↓

Display options

MY NCBI FILTERS

15,977 results



PubMed.gov

Cystitis

[Advanced](#) [Create alert](#) [Create RSS](#)

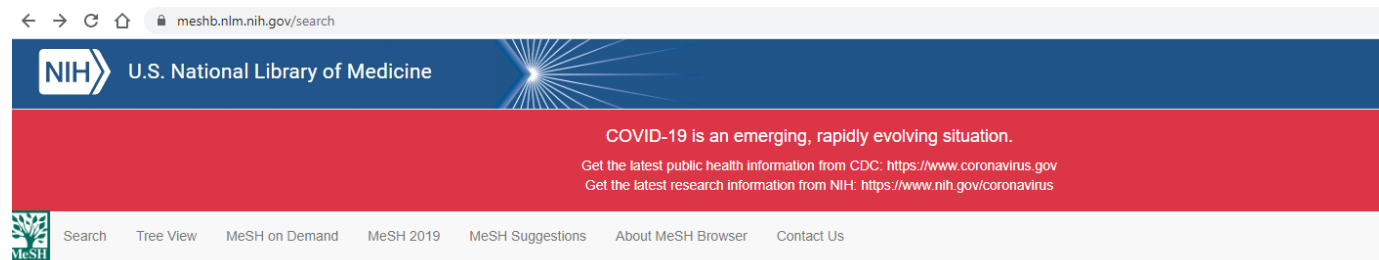
Save Email Send to

MY NCBI FILTERS

14,301 results

How to find MeSH terms?

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



Medical Subject Headings 2020

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

Search MeSH...

FullWord ▾ **Exact Match** All Fragments Any Fragment

☐ All Terms

☒ Main Heading (Descriptor) Terms

☐ Qualifier Terms

☐ Supplementary Concept Record Terms

☐ MeSH Unique ID

☐ Search in all Supplementary Concept Record Fields

Sort by: Relevance ▾

Results per Page: 20 ▾

<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

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

<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

PubMed search builder options

Subheadings:

☒ 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



"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

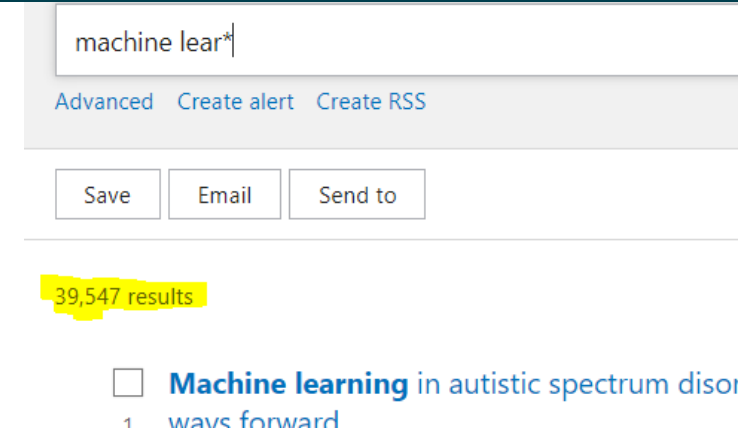
RESULTS BY YEAR



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 find "Machine learning" but also find "machine learn")



Searching PubMed



lung cancer machine learning

[Advanced](#) [Create alert](#) [Create RSS](#)

Save

Email

Send to

MY NCBI FILTERS

722 results

"lung cancer" "machine learning"

[Advanced](#) [Create alert](#) [Create RSS](#)

Save

Email

Send to

333 results

lung cancer AND machine learning

[Advanced](#) [Create alert](#) [Create RSS](#)

Save

Email

Send to

720 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

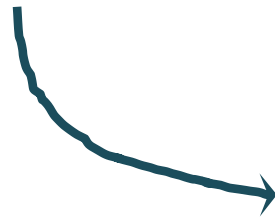
Searching PubMed

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/>



Thanks to Google Cloud, OVH, GESIS Notebooks and the Turing Institute for supporting us! 🐛



Turn a Git repo into a collection of interactive notebooks

Build and launch a repository

GitHub repository name or URL

GitHub

Git branch, tag, or commit

Git branch, tag, or commit

Path to a notebook file (optional)

Path to a notebook file (optional) File

Copy the URL below and share your Binder with others:

Fill in the fields to see a URL for sharing your Binder.

Copy the text below, then paste into your README to show a binder badge:

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!

[This Photo](#) by Unknown Author is licensed under [CC BY-SA-NC](#)

Questions/Comments

S. Ravichandran
ravichandrans@mail.nih.gov

Naomi Ohashi
naomi.ohashi@nih.gov