Advanced Cyberinfrastructure for Large-Scale Health Data Analysis (Tutorial)

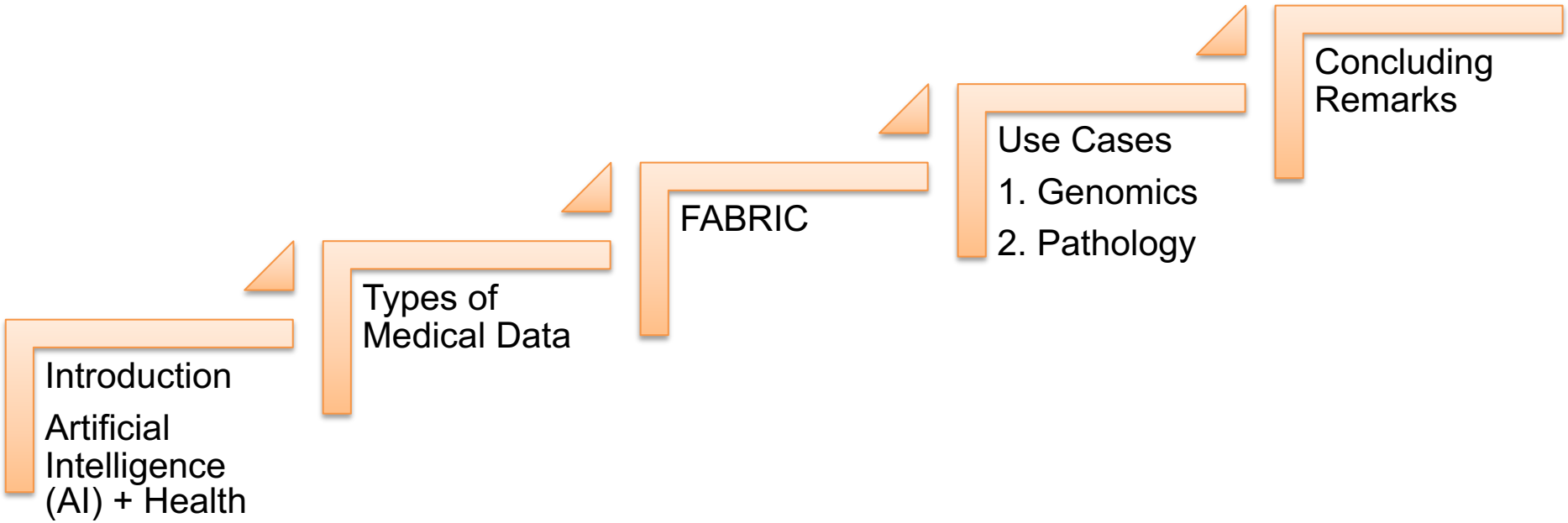
Praveen Rao, Ph.D.
Associate Professor
Dept. of Electrical Engineering & Computer Science
The University of Missouri
Columbia, Missouri, USA

EDBT ICDT PAESTUM 2024

6th International Workshop on Health Data Management in the Era of AI (HeDAI 2024)

Grant No. 2201583

Roadmap



The term “Artificial Intelligence” was coined by Prof. John McCarthy in 1956

AI in Healthcare: In the 70s

Medical Informatics

The INTERNIST-1/QUICK MEDICAL REFERENCE Project—Status Report

RANDOLPH A. MILLER, MD; MELISSA A. McNEIL, MD; SUE M. CHALLINOR, MD;
FRED E. MASARIE, Jr, MD, *and* JACK D. MYERS, MD, *Pittsburgh*

INTERNIST-1 and its successor, QUICK MEDICAL REFERENCE (QMR), are computer programs designed to provide health care professionals with diagnostic assistance in general internal medicine. Both programs rely on the INTERNIST-1 computerized knowledge base, which comprehensively describes 570 diseases in internal medicine. The philosophies behind the development of each program differ. Whereas INTERNIST-1 functions solely as a high-powered diagnostic consultant program, the QMR program acts more as an information tool, providing users with multiple ways of reviewing and manipulating the diagnostic information in the program's knowledge base. At the lowest level, the program can be viewed as an electronic textbook of medicine. In addition, the QMR program has the ability to assist users with generating hypotheses in complex patient cases. The QMR program has not been evaluated formally as an information tool for practicing physicians. A preliminary study indicates that QMR's case-analysis capabilities are of potential benefit in most patients in internal medicine admitted for diagnostic evaluation.

(Miller RA, McNeil MA, Challinor SM, et al: The INTERNIST-1/QUICK MEDICAL REFERENCE project—Status report, *In* Medical informatics [Special Issue]. West J Med 1986 Dec; 145:816-822)

AI in Healthcare: In the 70s

MYCIN: A KNOWLEDGE-BASED COMPUTER PROGRAM
APPLIED TO INFECTIOUS DISEASES*

Edward H. Shortliffe
Department of Medicine
Stanford University School of Medicine
Stanford, California 94305

A rule-based expert system is described which uses artificial intelligence techniques, and a model of the interaction between physicians and human consultants, to attempt to satisfy the demands of a user community that is often reluctant to experiment with computer technology. Experience to date has demonstrated that the program is efficient, relatively easy to use, and reliable in the domain of bacteremia therapy selection. Future work will involve broadening and evaluating the program's expertise in other areas of infectious disease therapy. To that end rules regarding diagnosis and treatment of meningitis have been written and are currently under evaluation.

Introduction

Few potential user populations are as demanding of computer technology as are practicing physicians. This is due to a variety of factors which include the physician's independence as a lone decision maker, the seriousness with which he views actions that may often have life-and-death significance, and the overwhelming time demands which tend to make him impatient with any innovation that breaks up the finely-tuned flow of his daily routine. Yet as medical science has expanded, the individual practitioner has become increasingly less able to manage all the expertise he needs if he is

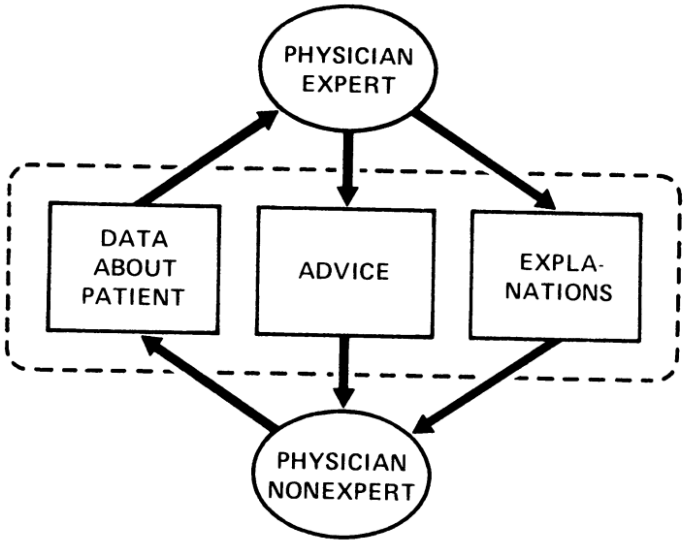


Figure 1 - Diagram summarizing the flow of information between physician and expert in the human consultation process. (Figure reproduced from reference 10).

AI in Healthcare: In the 70s

Representation of Expert Knowledge for Consultation: The CASNET and EXPERT Projects[(1)]

Casimir A. Kulikowski, Sholom M. Weiss

Kulikowski, C. A. and Weiss, S. M. "Representation of Expert Knowledge for Consultation: The CASNET and EXPERT Projects." Chapter 2 in Szolovits, P. (Ed.) *Artificial Intelligence in Medicine*. Westview Press, Boulder, Colorado. 1982.

Abstract

A major focus of research within the Rutgers Research Resource on Computers in Biomedicine is to investigate representations of expert knowledge and develop strategies of reasoning for consultation systems, with particular emphasis on medical consultation problems. Past work includes the development of a causal-associational network (CASNET) model for describing disease processes, and its application in an expert-level consultation program in glaucoma (CASNET/Glaucoma), that incorporates the knowledge of a national network of clinical experts in the disease. Collaborative testing of the program has been extended to Japan, and a comparison of Japanese and American decision making rules is currently underway.

Present work involves the formulation of a new representational scheme, called EXPERT, and its application in building models for reasoning in rheumatology and endocrinology. Experience to date with consultation programs based on the EXPERT models has shown that both the acquisition and the structuring of medical knowledge are significantly facilitated by this method.

In 2019...

Royal College of Physicians | Future Healthcare Journal

Future Healthc J. 2019 Jun; 6(2): 94–98.
doi: 10.7861/futurehosp.6-2-94

PMCID: PMC6616181
PMID: 31363513

The potential for artificial intelligence in healthcare

Thomas Davenport, president's distinguished professor of information technology and management[A] and
Ravi Kalakota, managing director[B]

ABSTRACT Go to:

The complexity and rise of data in healthcare means that artificial intelligence (AI) will increasingly be applied within the field. Several types of AI are already being employed by payers and providers of care, and life sciences companies. The key categories of applications involve diagnosis and treatment recommendations, patient engagement and adherence, and administrative activities. Although there are many instances in which AI can perform healthcare tasks as well or better than humans, implementation factors will prevent large-scale automation of healthcare professional jobs for a considerable period. Ethical issues in the application of AI to healthcare are also discussed.

In 2022...

Artificial intelligence in healthcare

Applications, risks, and ethical and societal impacts

https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2022)729512

In 2023...

DECEMBER 14, 2023

Delivering on the Promise of AI to Improve Health Outcomes

BRIEFING ROOM › BLOG

Lael Brainard, National Economic Advisor
Neera Tanden, Domestic Policy Advisor
Arati Prabhakar, Director of the Office of Science and Technology Policy

As President Biden has said, artificial intelligence (AI) holds tremendous promise and potential peril. In few domains is this truer than healthcare. The President has made clear, including by signing a landmark Executive Order on October 30, that the entire Biden-Harris Administration is committed to placing the highest urgency on governing the development and use of AI safely and responsibly to drive improved health outcomes for Americans while safeguarding their security and privacy.

National Institutes of Health (NIH)

National Institutes of Health
Office of Data Science Strategy

ODSS Intranet (NIH Staff)

Home | Strategic Plan | Resources | Research Funding | News & Events | About

Artificial Intelligence at NIH

Home / Artificial Intelligence At NIH

Artificial Intelligence at the NIH

The National Institutes of Health (NIH) makes a wealth of biomedical data available to research communities and aims to make these data findable, accessible, interoperable, and reusable—or FAIR. Additionally, the NIH seeks to make these data usable with artificial intelligence and machine learning (AI/ML) applications.

NIH has unique needs that can drive the development of novel approaches and application of existing tools in AI/ML. From electronic health record data, omics data, imaging data, disease-specific data, and beyond, NIH is poised to create and implement large and far-reaching applications using AI and its components.

Learn more about artificial intelligence activities at the NIH below.

https://datascience.nih.gov/artificial-intelligence

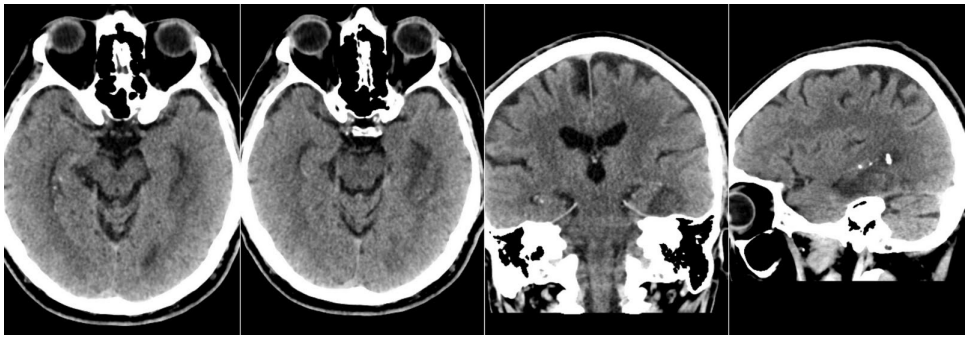
Structured and Unstructured Data

- Electronic Health Records (EHR) data of patients
  - Demographics, lab results, past medical history, procedures, clinical notes, medications, etc.

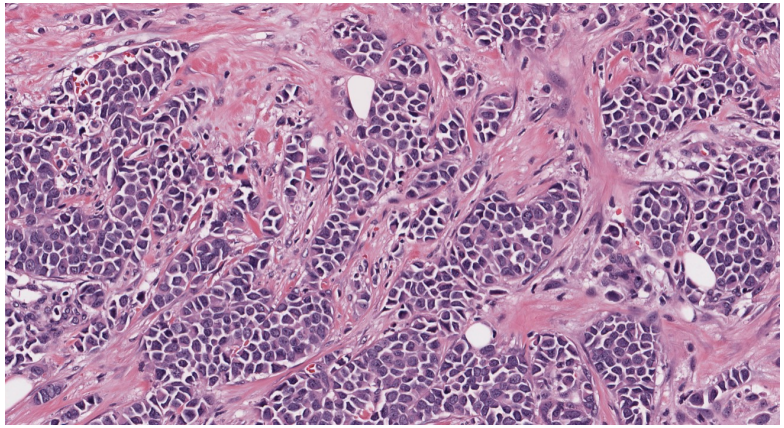
Source: Clipart Library

Medical Imaging

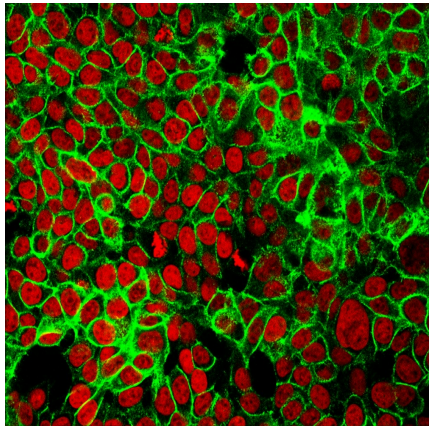
- Medical imaging data
  - Whole slide imaging, brain imaging, radiology imaging, MRI, CT-scans, X-rays, etc.



Source: Clipart Library



Source: Genomics Data Commons

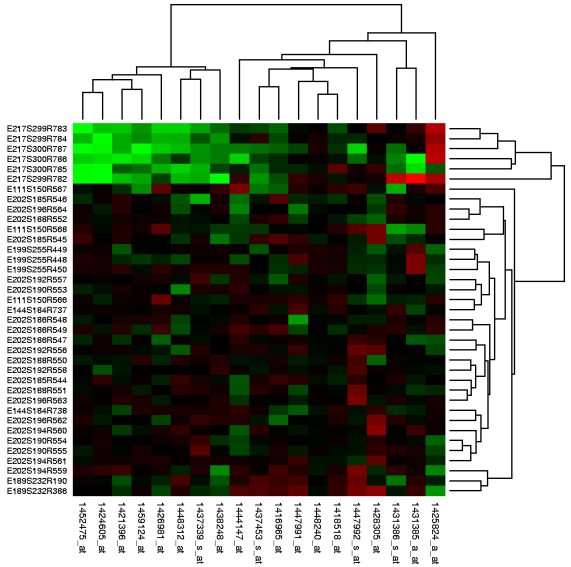


Source: Unsplash

Genomic Data

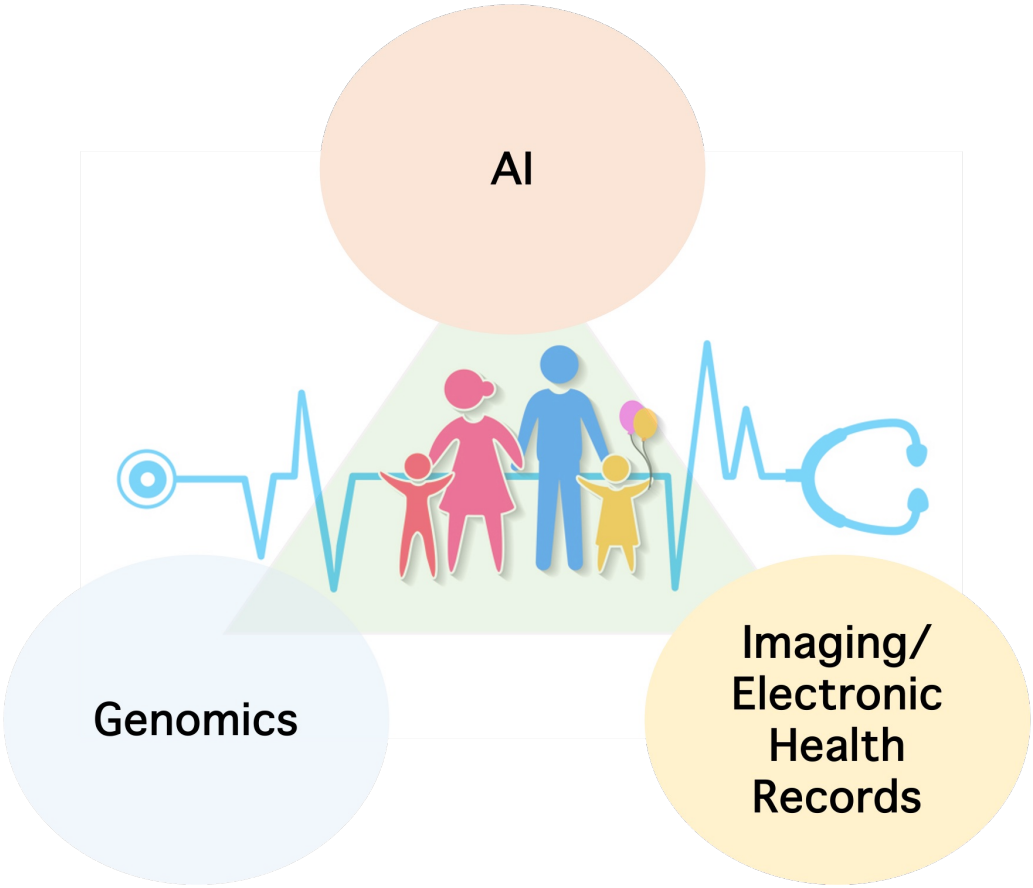
- Whole genome sequences, whole exome sequences
- Single-cell RNA sequencing; spatial transcriptomics data

Source: Unsplash



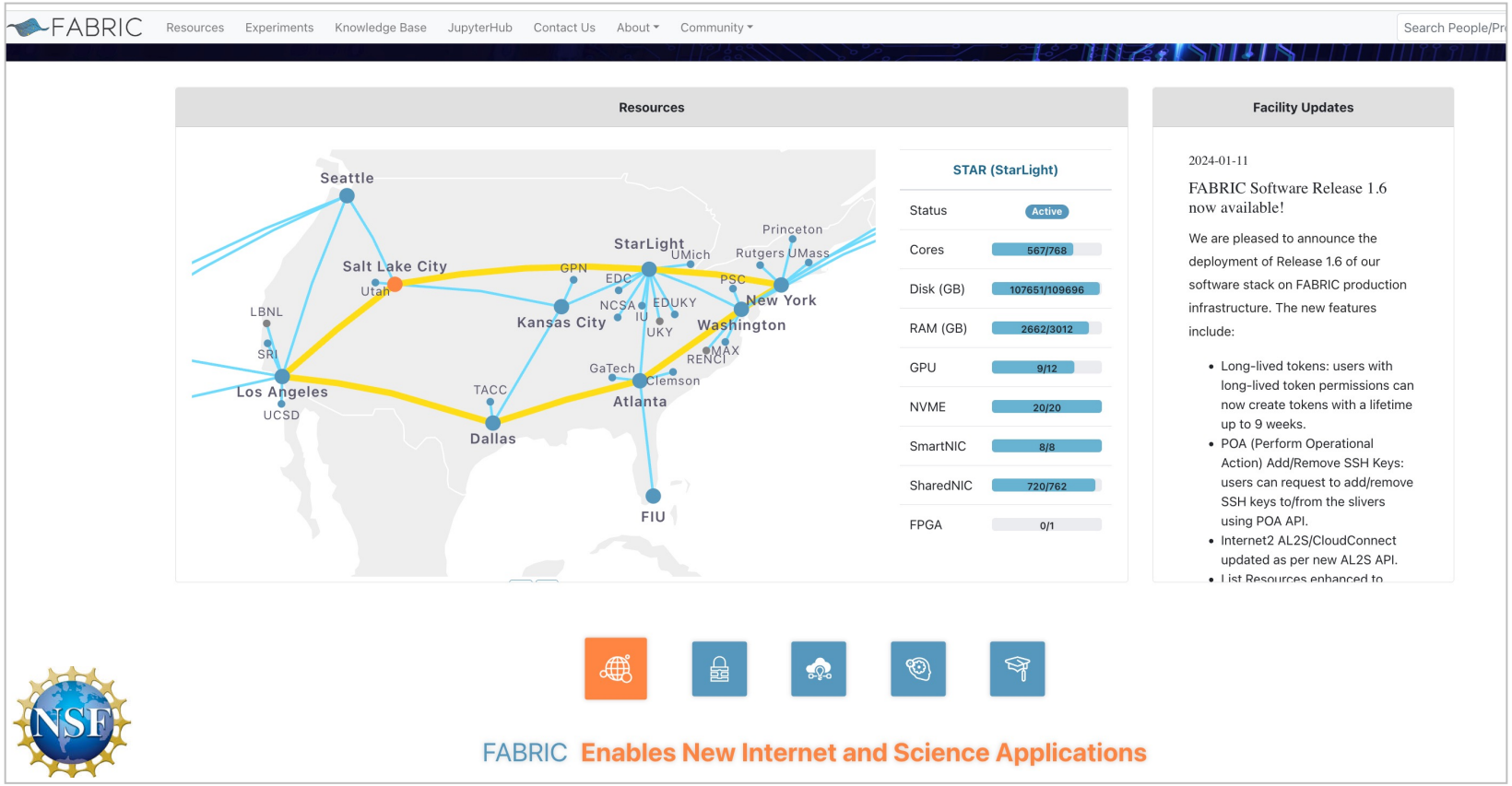
Source: Wikipedia

The Future of Medicine



FABRIC: A Programmable Research Testbed

https://portal.fabric-testbed.net/



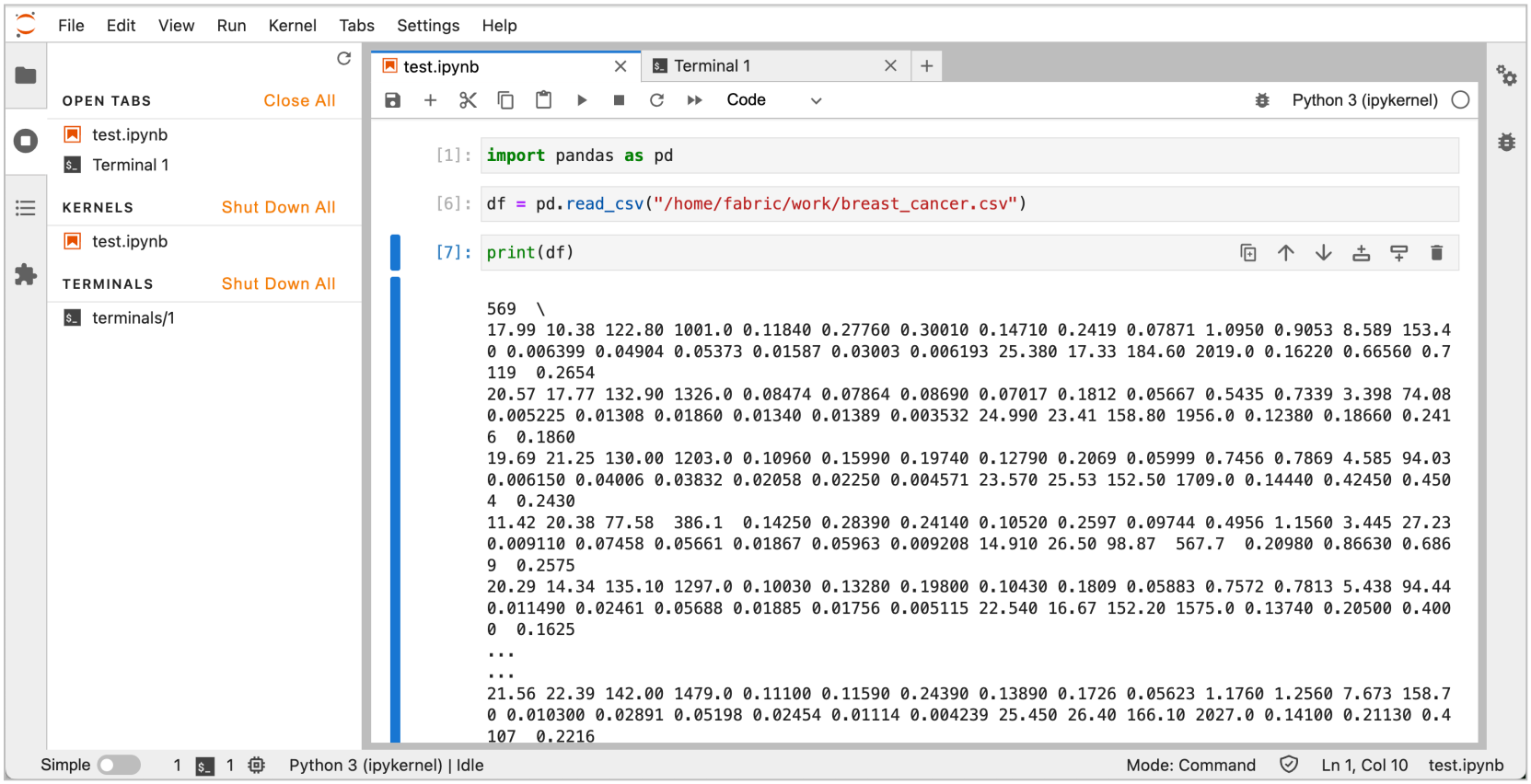
Over 30 sites

FABRIC Features

- Programmable experiments
  - Single virtual machine (VM); a cluster of VMs
  - A cluster across different sites
- Availability of hardware accelerators
  - Graphics processing units (GPUs)
  - Smart network interface cards (NICs)
  - Field programmable gate arrays (FPGAs)
- High speed optical links
  - Terabit core
  - 100G links
- Free to use!

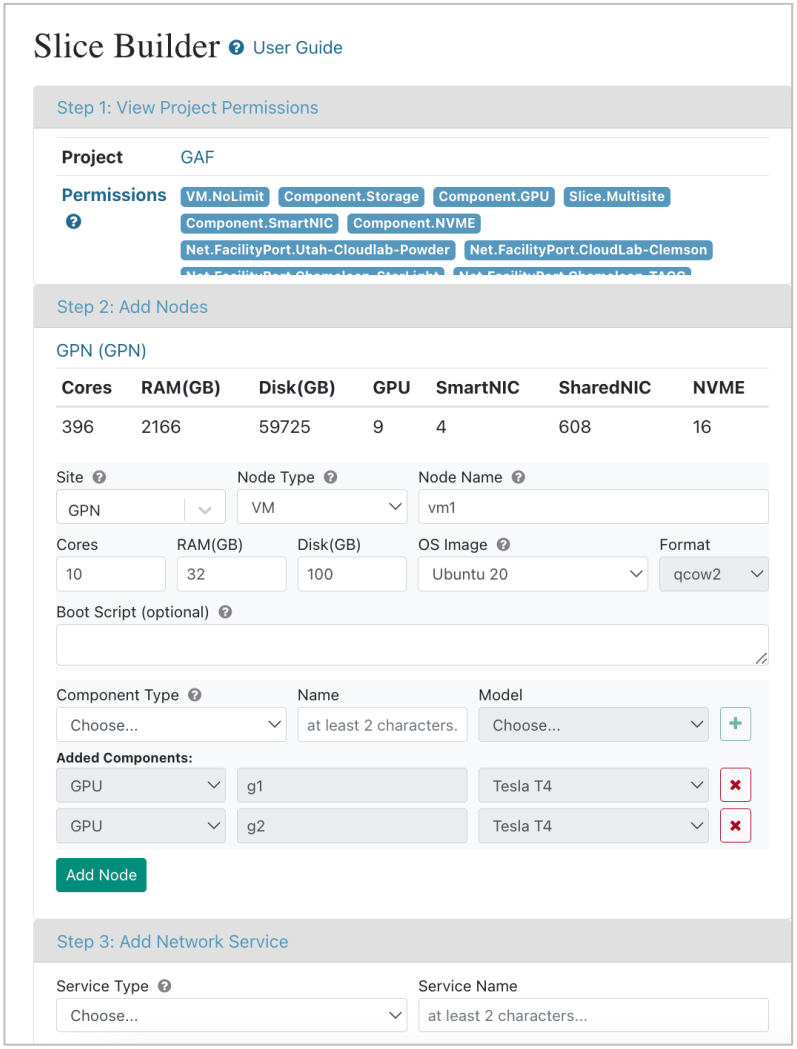
JupyterHub

For those who like Notebooks!



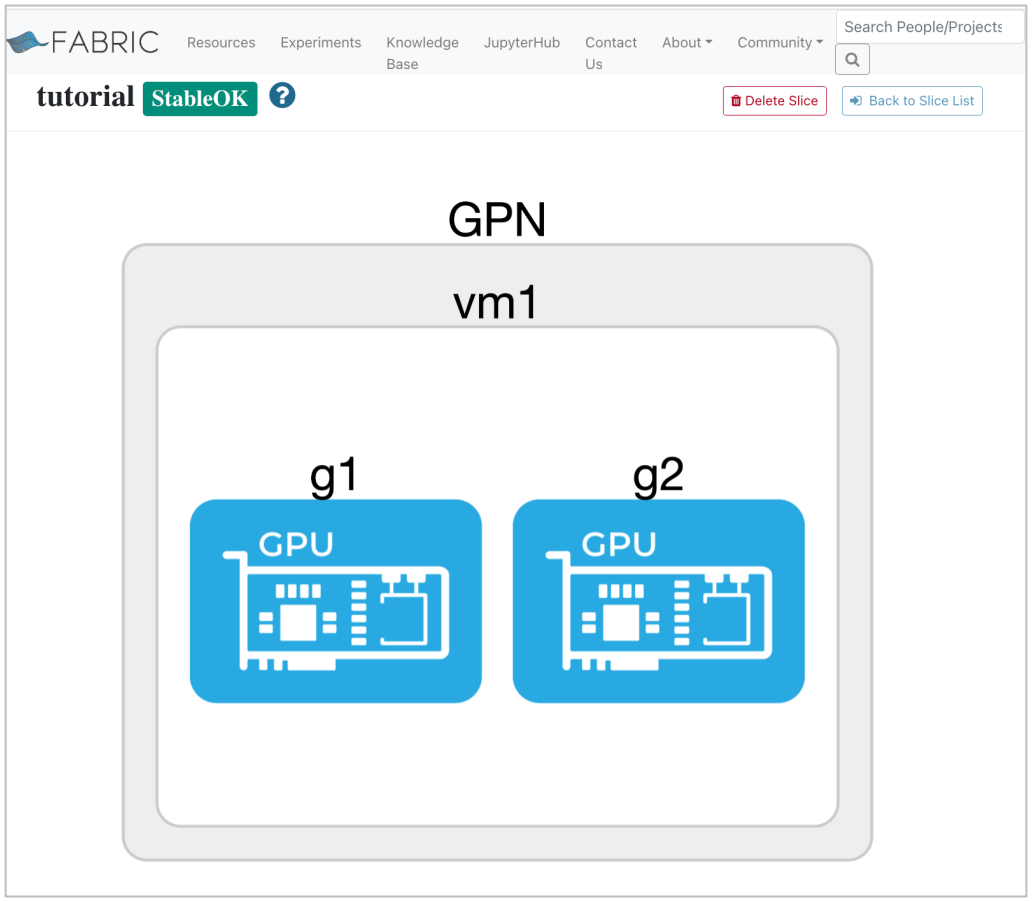
Setting Up an Experiment

Step 1

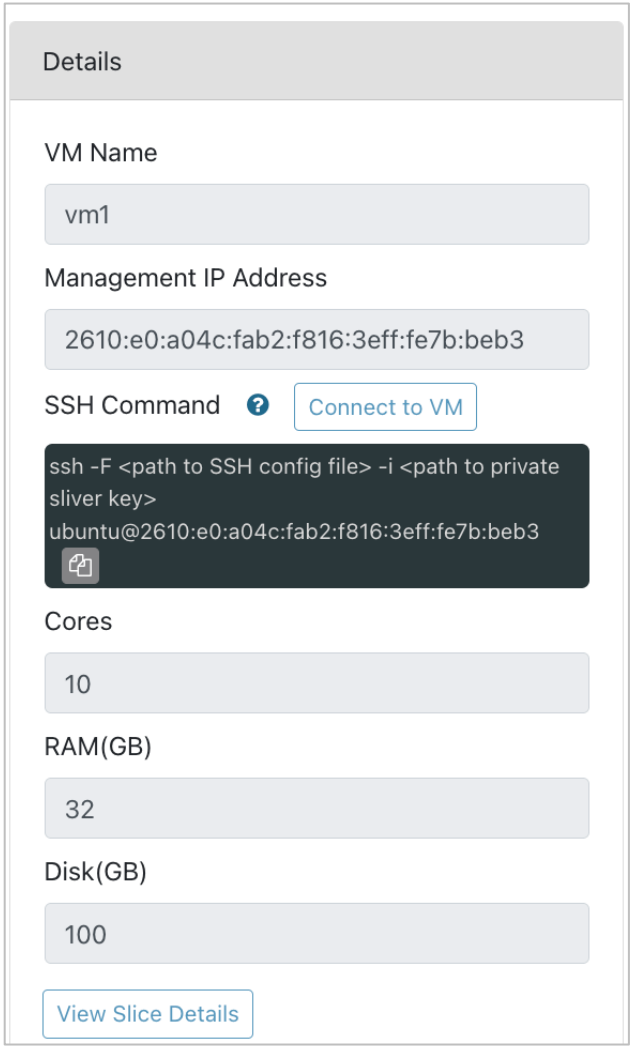


Setting Up an Experiment

Step 2



Step 3



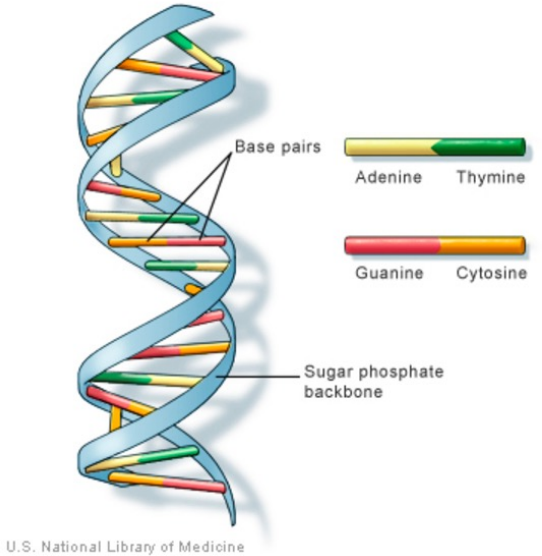
First Use Case: Genomics

- Human genome sequence analysis
  - Variant calling pipeline
    - Identify variants in an individual's DNA

Source: Clipart Library

NVIDIA PARABRICKS

Background



Source: https://medlineplus.gov/genetics/understanding/basics/dna/

	Reference Genome	A Person's Genome
What is it?	+ Mitochondrial DNA	+ Mitochondrial DNA
How many chromosomes?	24 (22 + X + Y)	46 (23 PAIRS)
How many letters?	~3.2 bn	~6.4 bn
How to think about it?	• The Human Genome Project and its goal of a first draft of "the human genome" • Serves as a standard for comparison • A "consensus" genome sequence	• The genome of a person • The genome within a person's cells • The whole genome sequence of an individual

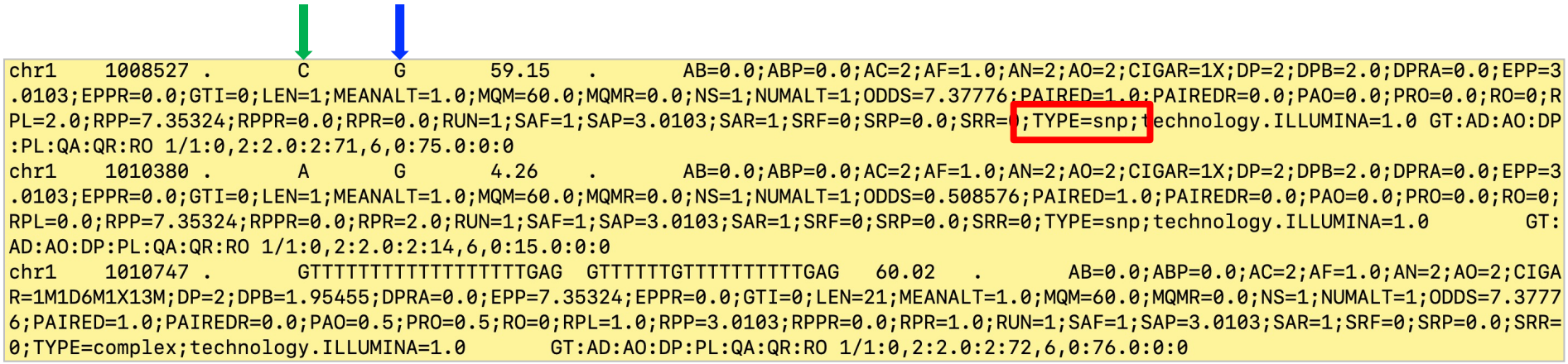
Source: https://www.veritasgenetics.com/our-thinking/whole-story/

By 2025, there will be 2-40 exabytes of human genome data [Stephens et.al., PLOS Biology, 2015]

Human genomes are very large in size!

Variant Calling

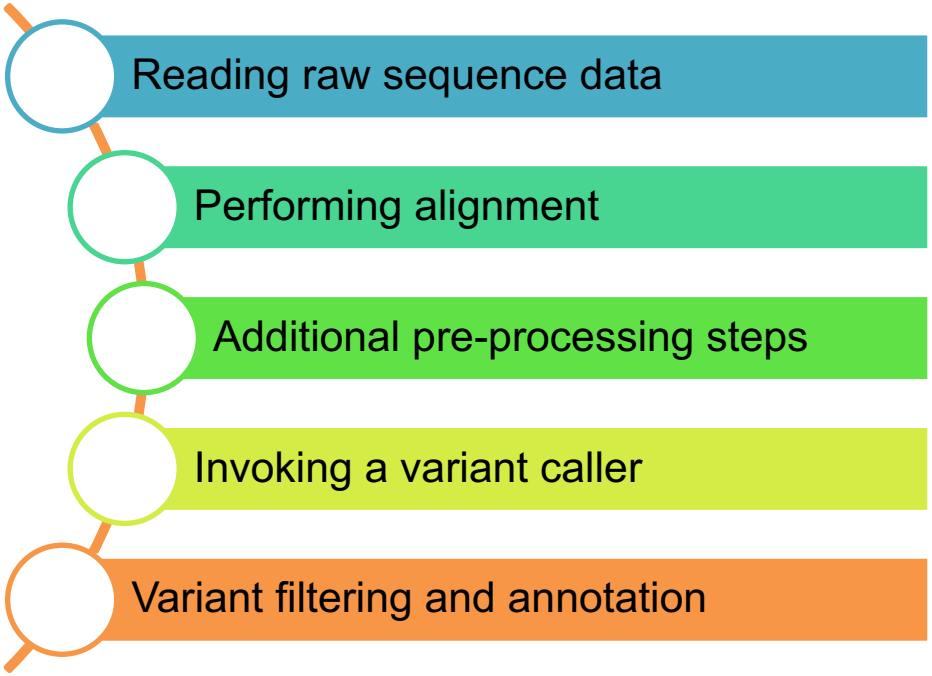
- Variant calling is a fundamental task
  - To identify variants in an individual's genome compared to a reference genome
- Types of variants[1]
  - Single nucleotide polymorphism (SNP)
  - Short insertions/deletions (indels)
  - Copy number variation, and other structural variants



[1] https://m.ensembl.org/info/genome/variation/prediction/classification.html

Variant Calling

- Identifying variants can enable
  - Better understanding of an individual's risk to diseases
  - New innovations in precision medicine and drug discovery
- Variant calling pipeline (DNA sample)[1]



[1] https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

Setup

- Tools needed
  - CUDA, Docker, nvidia-docker2
  - NVIDIA's Parabricks

```
$ git clone https://github.com/MU-Data-Science/GAF.git
$ ${HOME}/GAF/Tutorial/scripts/setup-GPUs-genomics.sh
```

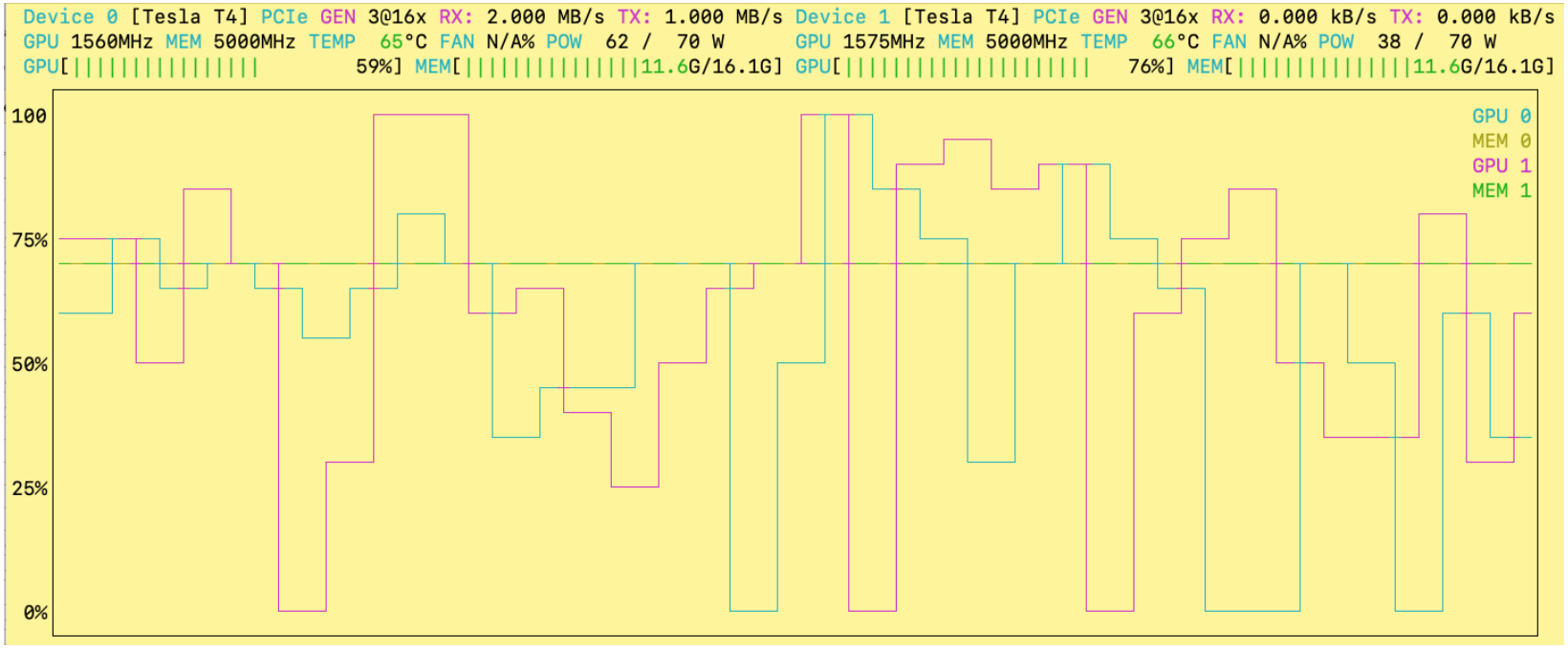
```
ubuntu@1a02b0b2-5777-4ea2-a309-f96ea5afe09c-vm1:~$ sudo docker run --rm --gpus all nvidia/cuda:12.0.0-base-u
buntu20.04 nvidia-smi
Thu Mar 14 13:13:35 2024
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 525.60.13    Driver Version: 525.60.13    CUDA Version: 12.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla T4            On   | 00000000:06:00.0 Off |                    0 |
| N/A   32C    P8    11W /  70W |      6MiB / 15360MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
+-----------------------------------------------------------------------------+
```

Variant Calling

- Run the variant calling pipeline on a publicly available genome (ERR062934)

```
ubuntu@13ac12dc-fa3d-4da5-ada1-ad63e88223dc-vm1:~$ sudo docker run --gpus 1,2 --rm --volume $(pwd):/workdir --volume $(pwd):/outputdir nvcr.io/nvidia/clara/clara-parabric
ks:4.2.1-1 pbrun germline --ref /workdir/hs38.fa --in-fq /workdir/ERR062934_1.fastq.gz /workdir/ERR062934_2.fastq.gz  --out-bam ERR062934.bam --out-variants /outputdir/ER
R062934.vcf --low-memory
Please visit https://docs.nvidia.com/clara/#parabricks for detailed documentation


[Parabricks Options Mesg]: Automatically generating ID prefix
[Parabricks Options Mesg]: Read group created for /workdir/ERR062934_1.fastq.gz and /workdir/ERR062934_2.fastq.gz
[Parabricks Options Mesg]: @RG\tID:ERR062934.1.1\tLB:lib1\tPL:bar\tSM:sample\tPU:ERR062934.1.1

[Parabricks Options Mesg]: Checking argument compatibility
[Parabricks Options Mesg]: Read group created for /workdir/ERR062934_1.fastq.gz and /workdir/ERR062934_2.fastq.gz
[Parabricks Options Mesg]: @RG\tID:ERR062934.1.1\tLB:lib1\tPL:bar\tSM:sample\tPU:ERR062934.1.1
[PB Info 2024-Mar-15 13:14:11] ------------------------------------------------------------------------------
[PB Info 2024-Mar-15 13:14:11] ||                 Parabricks accelerated Genomics Pipeline                ||
[PB Info 2024-Mar-15 13:14:11] ||                           Version 4.2.1-1.beta4                         ||
[PB Info 2024-Mar-15 13:14:11] ||                       GPU-BWA mem, Sorting Phase-I                      ||
[PB Info 2024-Mar-15 13:14:11] ------------------------------------------------------------------------------
[M::bwa_idx_load_from_disk] read 0 ALT contigs
[PB Warning 2024-Mar-15 13:14:13][ParaBricks/src/pbOpts.cu:254]

WARNING
The system has 20 threads, however recommended number of threads with 2 GPU is 24.
The run might not finish or might have less than expected performance.

[PB Info 2024-Mar-15 13:14:13] GPU-BWA mem
[PB Info 2024-Mar-15 13:14:13] ProgressMeter    Reads           Base Pairs Aligned
```

Output of nvtop



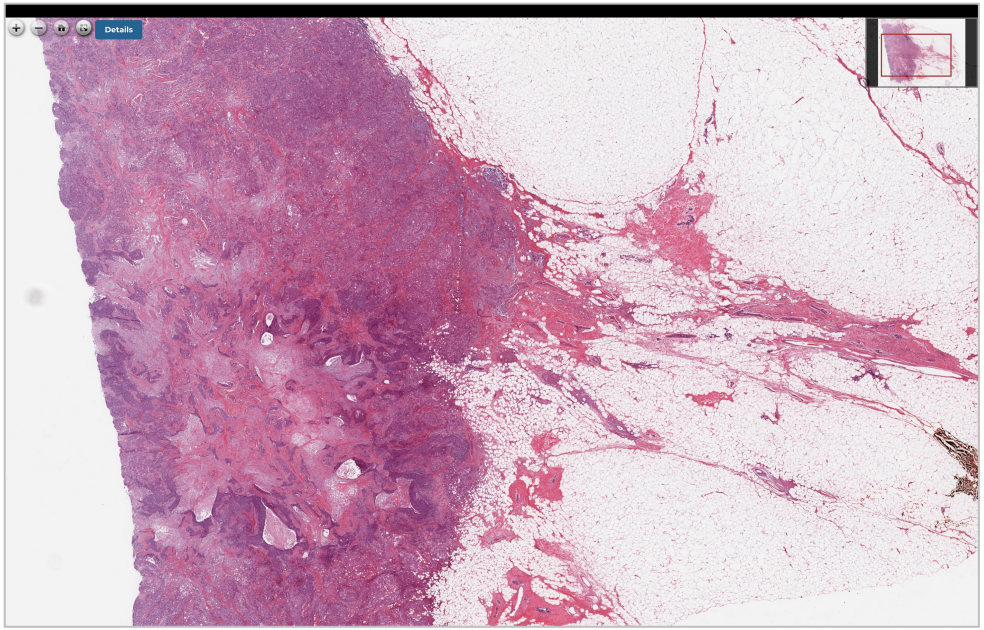
Second Use Case: Pathology

- Deep learning for medical imaging
  - Whole slide image analysis

PyTorch

TIAToolbox
Tissue Image Analytics

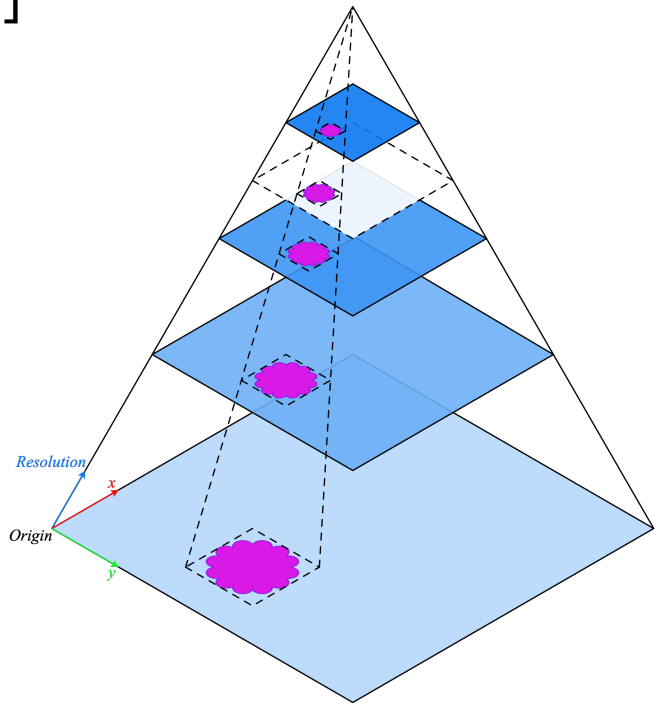
https://github.com/TissueImageAnalytics/tiatoolbox



Source: Genome Data Commons

Whole Slide Image Classification

- Pathologists use 20-40x magnification in optical microscopes for tissue diagnosis
- Whole slide imaging
  - Each digital slide can have 10+ billion pixels [Wang et. al., SPIE 2012]



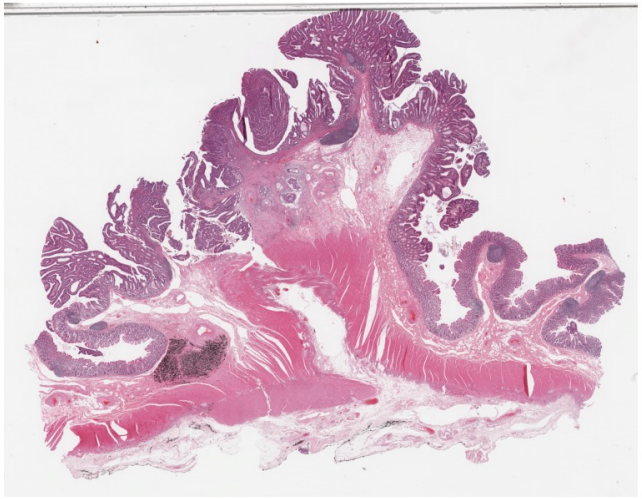
Source: TIAToolbox

PyTorch + TIAToolbox

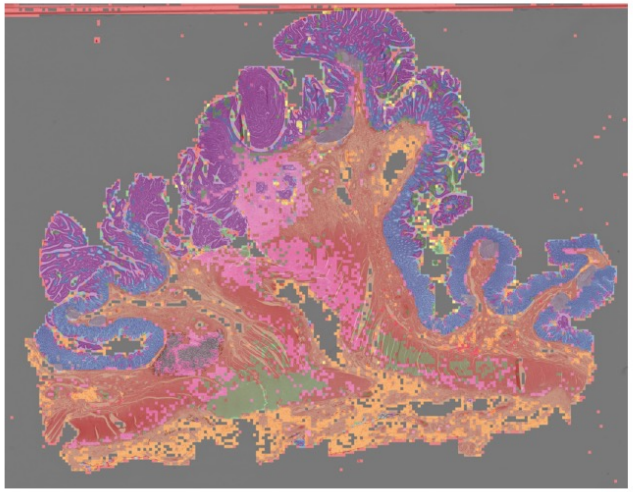
- See https://pytorch.org/tutorials/intermediate/tiatoolbox_tutorial.html

```
$ ${HOME}/GAF/Tutorial/scripts/setup-GPUs-WSI.sh
$ python3 ${HOME}/GAF/Tutorial/scripts/run_WSI_classification.py
```

Input



Output



224x224 pixels patch size

Concluding Remarks

- FABRIC provide computing resources for large-scale health data analysis – at no charge
- FABRIC experiments are programmable
  - Can be used for a number of scientific applications
- Two use cases
  - Genomics
    - Variant calling pipelines
  - Pathology
    - Whole slide image analysis
- Be creative and explore the testbed!

Recent Publications Enabled By FABRIC and CloudLab

1. Manas Das, Khawar Shehzad, and Praveen Rao - **Efficient Variant Calling on Human Genome Sequences Using a GPU-Enabled Commodity Cluster**. In *32nd ACM International Conference on Information and Knowledge Management* (CIKM 2023), 6 pages, Birmingham, UK, 2023.
2. Andrew Rommitti, Jiya Shetty, and Praveen Rao - **Evaluating the Effectiveness of Synthetic Datasets for Dementia Diagnosis Using Deep Learning**. In *52nd IEEE Applied Imagery and Pattern Recognition* (AIPR) Workshop, 5 pages, St. Louis, 2023.
3. Abdulkadir Korkmaz, Ahmad Alhonainy, and Praveen Rao - **An Evaluation of Federated Learning Techniques for Secure and Privacy-Preserving Machine Learning on Medical Datasets**. In *51st IEEE Annual Applied Imagery Pattern Recognition* (AIPR) Workshop, 7 pages, Washington D.C., 2022.
4. Praveen Rao and Arun Zachariah - **Enabling Large-Scale Human Genome Sequence Analysis on CloudLab**. In *IEEE INFOCOM Workshops: Computer and Networking Experimental Research using Testbeds* (CNERT 2022), 2 pages, 2022.
5. Praveen Rao, Arun Zachariah, Deepthi Rao, Peter Tonellato, Wesley Warren, and Eduardo Simoes - **Accelerating Variant Calling on Human Genomes Using a Commodity Cluster**. In *30th ACM International Conference on Information and Knowledge Management* (CIKM), pages 3388-3392, Australia, 2021. Nominated for Best Short Paper Award.

Questions?

- Project site
  - https://github.com/MU-Data-Science/GAF
- Team

  Faculty: Dr. Praveen Rao (**PI**), Dr. Eduardo Simoes, Dr. Deepthi Rao

  Ph.D. Students: Khawar Shehzad, Polycarp Nalela, Ajay Kumar

  B.S. Students: Matthew Schutz, Chase Webb

  Alumni: Dr. Manas Das (Southern Illinois University Edwardsville)
- Acknowledgments
  - National Science Foundation Grant No. 2201583

Thank you!