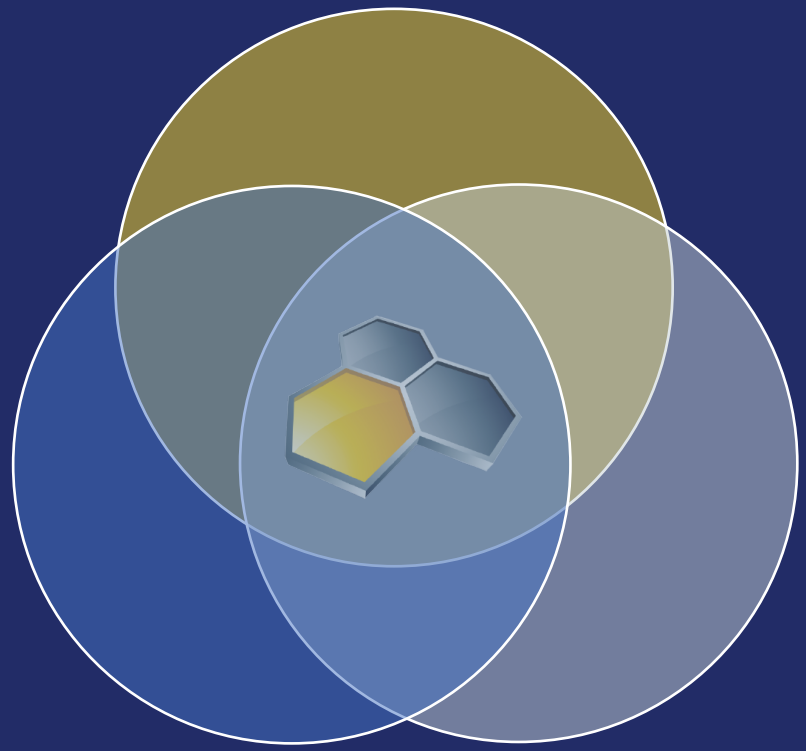
On the importance of Edge Computing: Service to the cloud and to Public Health

Luis V Santana-Quintero, PhD, Staff Fellow, Center for Biologics Evaluation & Research

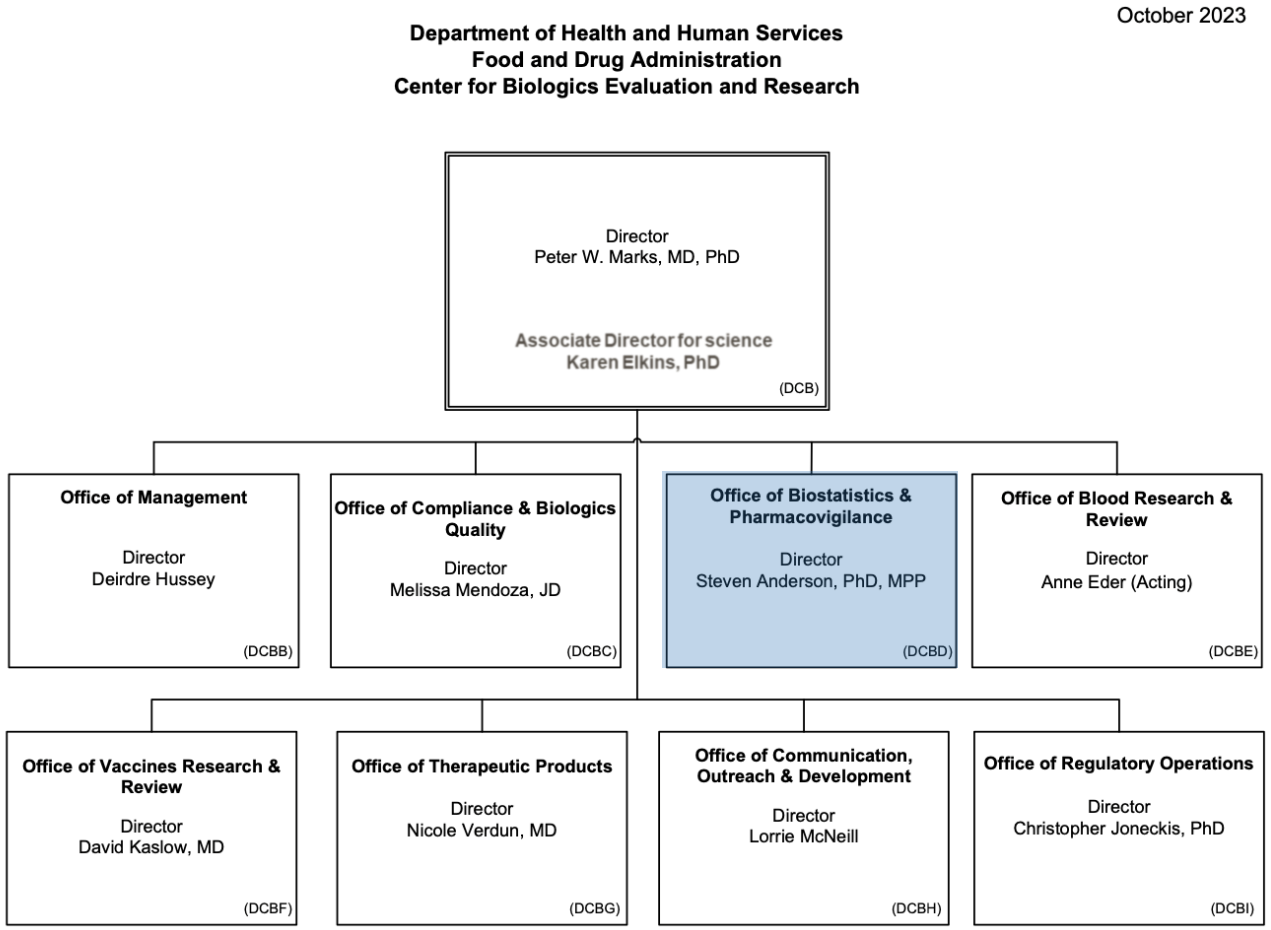
Outline

- FDA HIVE Ecosystem
  - NGS Lifecycle and Challenges
    - Data Transfer / Hardware / Pipelines / Data Analysis
- Supporting FDA's mission
  - Sharing Data
  - Regulatory Review
  - Flexible Capabilities in the cloud
- Solutions for omics and machine learning problems
  - Natural Language Processing (NLP) and deep learning

FDA/HIVE Ecosystem



Where we are: Office of Biostatistics and Pharmacovigilance



Deputy Office Director
Richard Forshee, PhD

HIVE Lead
Luis Quintero-Santana, PhD

Anton Golikov

Sean Smith, PhD

Charles Chung, PhD

Arya Eskandarian

Tigran Ghazanchyan

Sergey Ivanovsky

Justin Philip

Alexander Lukyanov

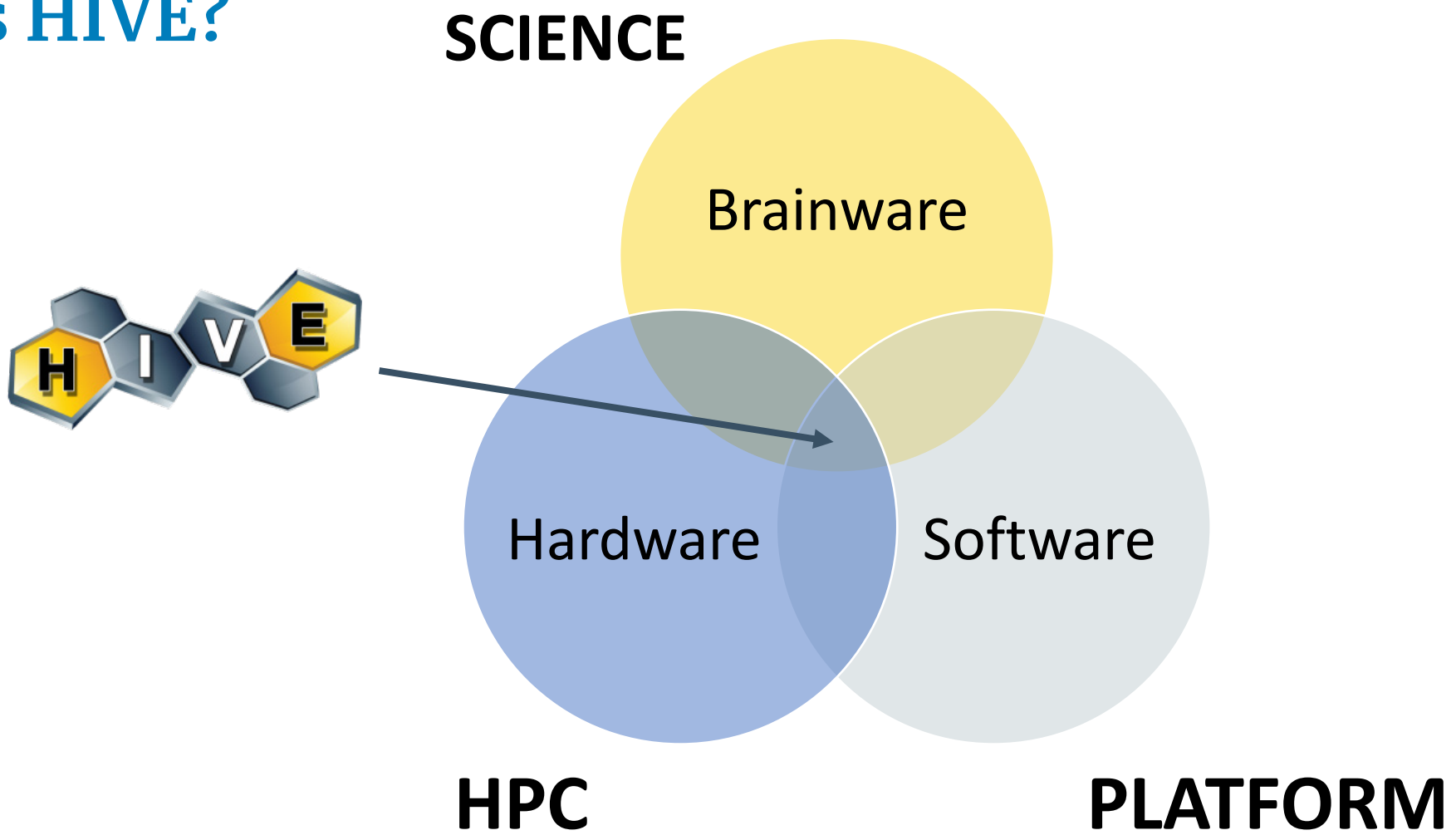
Ilya Mazo, PhD

Paul Rahul, PhD

Alexander Murray

Spyros Karaiskos, PhD

What is HIVE?



What is HIVE ?

- A lot of CPUs
- Big storage space
- Runs large computations
- Multiple applications

SCIENCE

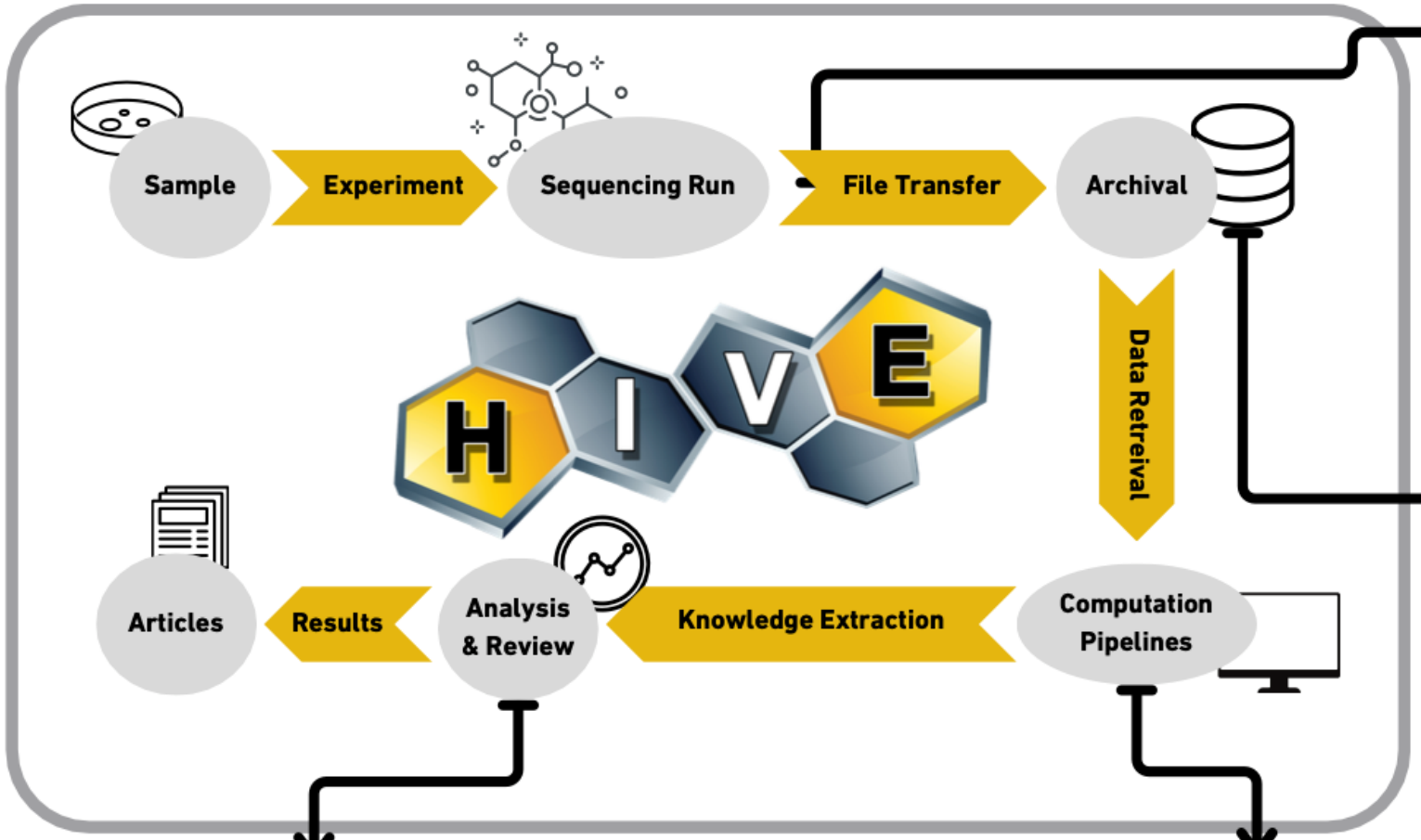
- Bioinformatics scientists
- Data Science
- Novel algorithms
- Many publications

HPC

PLATFORM

- Web-based interface
- Point and click
- Storage management
- Sharing and tracking

NGS Lifecycle – Challenges



Data Transfer

- Slow and unreliable

Hardware

- Adequate storage and computing power as datasets keep getting bigger.

Data Analysis – Difficult to visualize and interpretate data in a timely manner

Pipelines – Generate custom solutions for biological questions

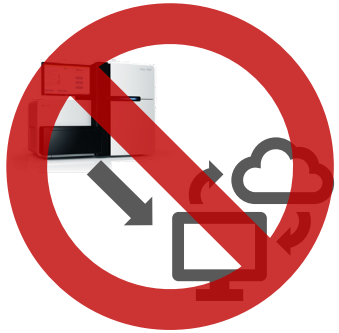
Data Transfer

Data Transfer
- Slow and unreliable

- Enhanced integration with the **NGS Core Sequencing Facilities in FDA - White Oak Campus**, building high-speed connections from sequencing machine directly to HIVE.

1. Established direct connectivity to sequencers
2. Make data available through a user-friendly platform
3. Data is secured and doesn't need to be sent around.

Reduce the transfer time from days to hours



Hardware

- High-Performance Computing
  - Cores: **6,000**
  - RAM: **30 TB**
  - Disk space: **10 PB**
  - GPU: **50 NVIDIA**
  - NGS optimized hardware

- Hosting
  - HIVE platform production
  - HIVE development
  - HIVE test

Hardware

- Adequate storage and computing power as datasets keep getting bigger.
- Specialized hardware

FDA

ATO

Pipelines and Tools

Pipelines – Generate custom solutions for biological questions

NGS Technology allows for rapid, high-resolution sequencing:

Whole Genome Sequencing

- Rapidly sequence genomes.

Targeted Sequencing

- To identify cancer-related mutations.

RNA Sequencing and single cell RNA-seq

- Gene Expression Analysis and to discover RNA variants and splice sites.

Metagenomic and microbiome sequencing

- Study microbial diversity in humans or environment

Pathway enrichment and Visualization of Omics

- Gene set enrichment analysis, gene ontologies, gene regulatory networks

VDJ Tools

Annotations
General (2)
Extract sequences from Annotations
Panel analysis portals (3)
SnpEff

Antimicrobial Resistance (AMR)
AMR Short Reads Pipeline
AMR Long Reads Pipeline
AMR Hybrid Reads Pipeline

Classifications
AlgoRLDA
Sequence Hierarchical Clustering
Genotypic Variation Clustering
Alignment Comparator
Viral Mutation Comparator
Mothur MiSeq Pipeline
DESeq2

De Novo Sequence Assemblers
Canu
Contig Extension
IDBA-HYBRID
IDBA-UD
PRICE
SPAdes
Velvet (Paired-End)

Downloaders
HTTP Downloader
FTP Downloader
HIVE AWS S3 Downloader
HIVE Network Downloader
NCBI GenBank Downloader
NCBI SRA Fastq Downloader
NCBI SRA SAM Downloader

DRAGEN
DRAGEN (full)
DRAGEN Reference Index
DRAGEN Pipeline (16)

Epigenetics
MACS2 Callpeak
MACS2 BDGCMP Function

General DNA Filters and Tools
Picard Tools (4)
Adapters Filter
Agilent Molecular Barcode Removal
Agilent SureCall Trimmer
Agilent SureCall Trimmer
ANARCI
Bcl to Fastq converter
Complexity Filter
FastP
FastQC
HIVE UMI-Tools

Glycomics
MS Spectra Analysis
Glymps for glyco-proteomics
MSGF+ for glyco-proteomics
GlycoPeptideSearch

HISAT2 Index

Lofreq
LoFreq alnqual
LoFreq call
LoFreq checkref
LoFreq indelqual
LoFreq somatic
LoFreq viterbi

Metagenomics
HUMAnN2
MetaPhlAn2

Other
AlphaFold 2.1.2
ELISA
Renamer
Differential P
Adver
CAR-T
HIV Diversity
Resistance Gene Identifier (RGI)

PyIR

RNA-SeQC

Salmon
Salmon Alevin
Salmon Index
Salmon Quant Align Mode
Salmon Quant Map Mode
Salmon Quantmerge

SARS-CoV-2
Spike Conservation Tool

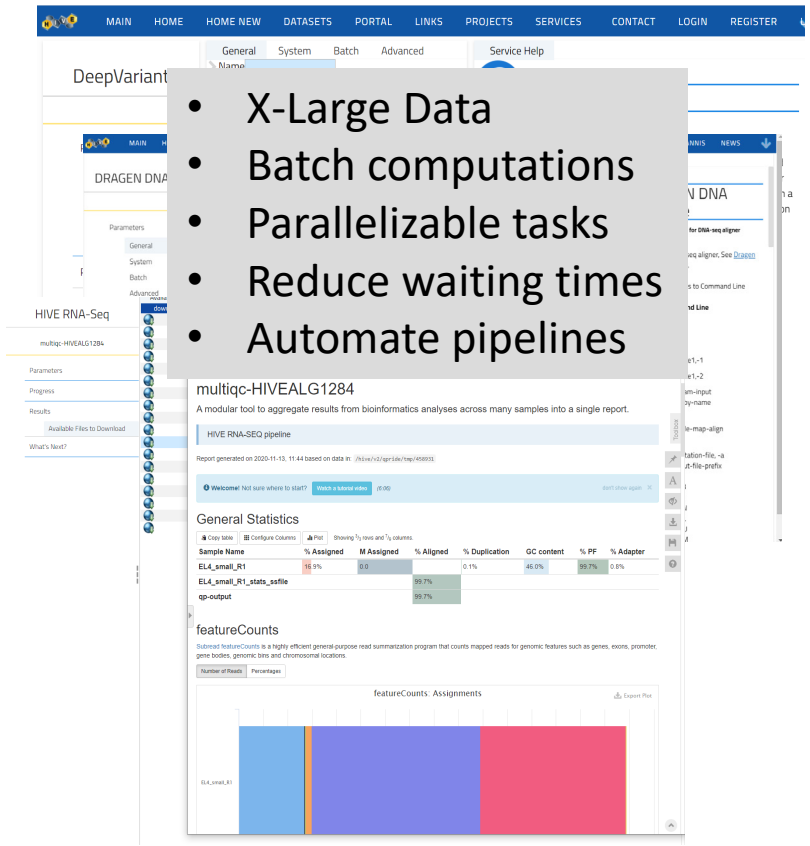
Sequence Alignment on Genome
DNA-seq (9)
Long Reads (2)
Metagenomic-seq (2)
Multiple Alignment (2)
NCBI Magic-BLAST
RNA-seq (6)

STAR
STAR Ali
Deep

Data Analysis

Data Analysis – Computing data analysis where data resides

Compute FAST



- X-Large Data
- Batch computations
- Parallelizable tasks
- Reduce waiting times
- Automate pipelines

HIVE API

Compute SLOW

MATLAB

- Smaller Datasets
- Customized scripts
- Rigorous analytics
- Visualization

jupyter

RStudio

tableau

git

ATO

HIVE is supporting FDA's mission

Regulatory Review

To create applications and services integrated with existing regulatory review services

Flexible Capabilities

Provide a venue for the exploration of rapidly evolving new technologies

Sharing Data

Efficiently share datasets across FDA

Unified HPC at FDA

To create a shared ecosystem through FDA

Regulatory Big Data

Sharing Data

FDA

Regulatory HIVE (regHIVE)

- HIVE instance Authorized to Operate (ATO) in a regulatory environment
- Supports review of regulatory submissions with NGS protocols

HIVE ATO

- **Too big for current Electronic Submission Gateway**
- **Mailing hard disk**

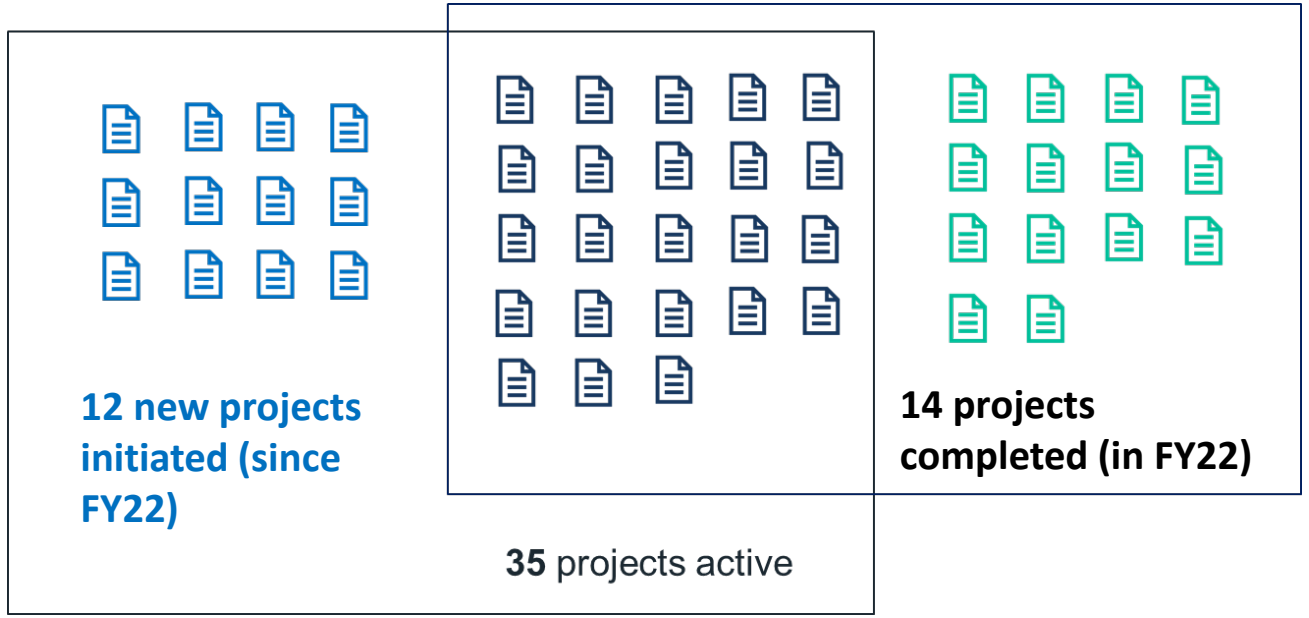
Available to Reviewers

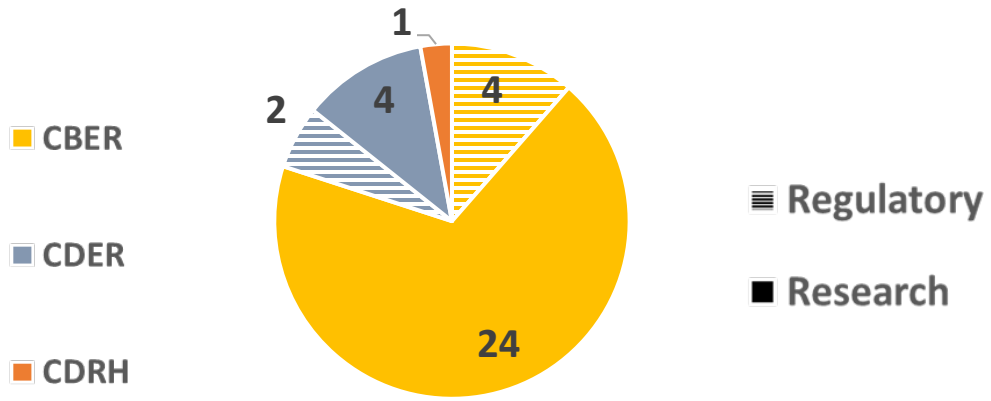
CBER Total >150TB

- **24 Investigational New Drug Applications**
- **3 Master Files**
- **2 Biological License Applications**

CDER Total >5TB

Supporting FDA Regulatory and Research





Regulatory Review

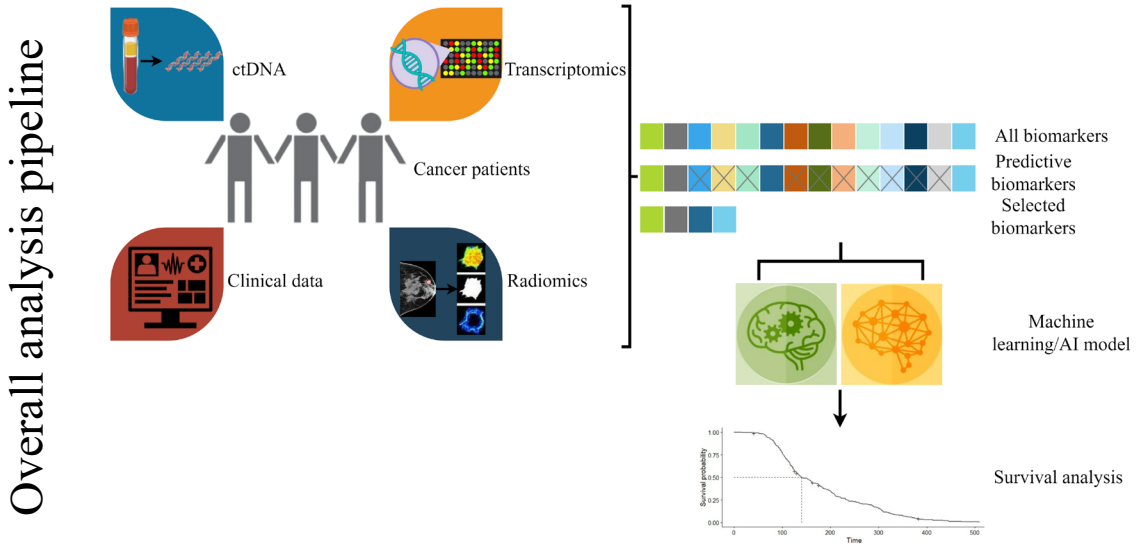
Finding biomarkers for breast cancer survival estimation using multimodal data

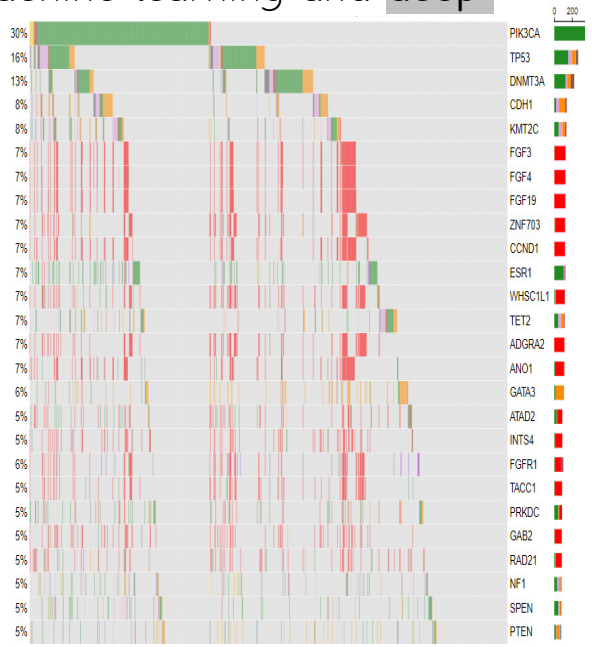
Multicenter Research Collaboration Agreement with a sponsor and multiple FDA centers (CBER/CDER/CDRH)

- Goal:
  - To find novel radiogenomics-based prognostic and predictive biomarkers for HR+/HER2- metastatic breast cancer. Use Phase III trial – clinical data from Advanced Breast Cancer.
- Methods:
  - Combining different data types to create individual and multi-omics models using machine learning and deep learning approaches. Including multiple predictive endpoints

Data Sharing

Use of HIVE to transfer large datasets with sponsor and AWS.





Regulatory QC NGS Pipeline

Regulatory Review

FDA

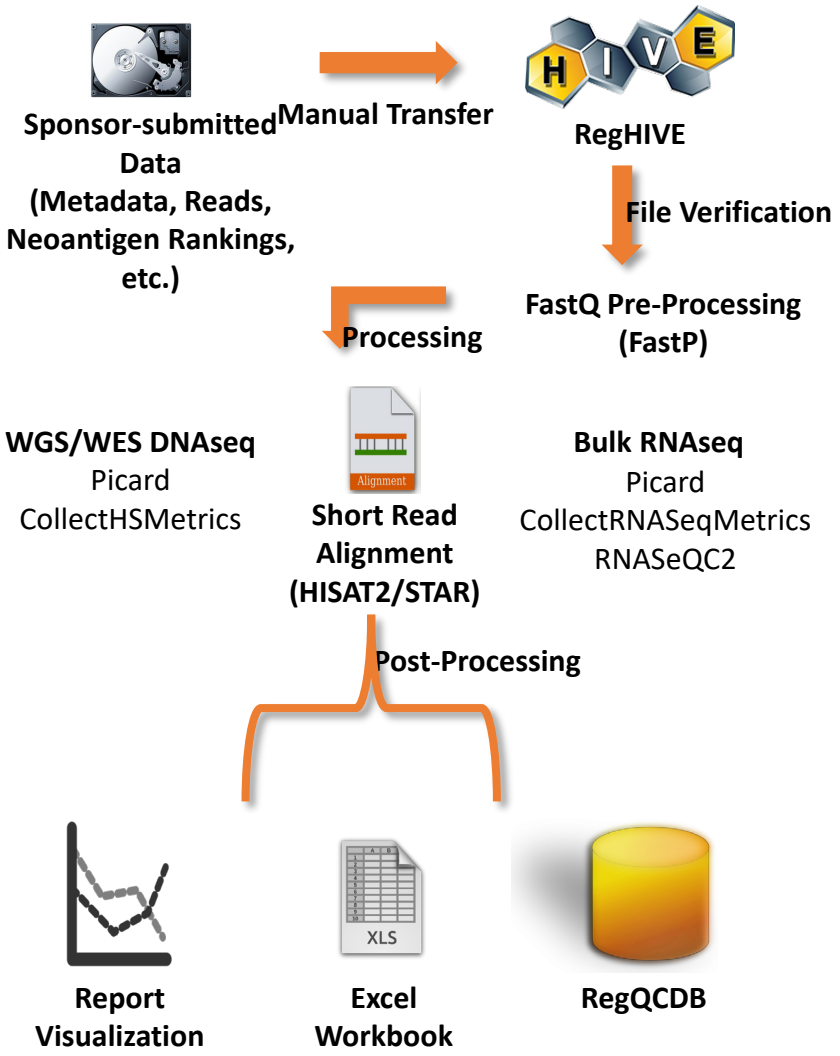
- Purpose:

Run Quality Checks on NGS datasets for:

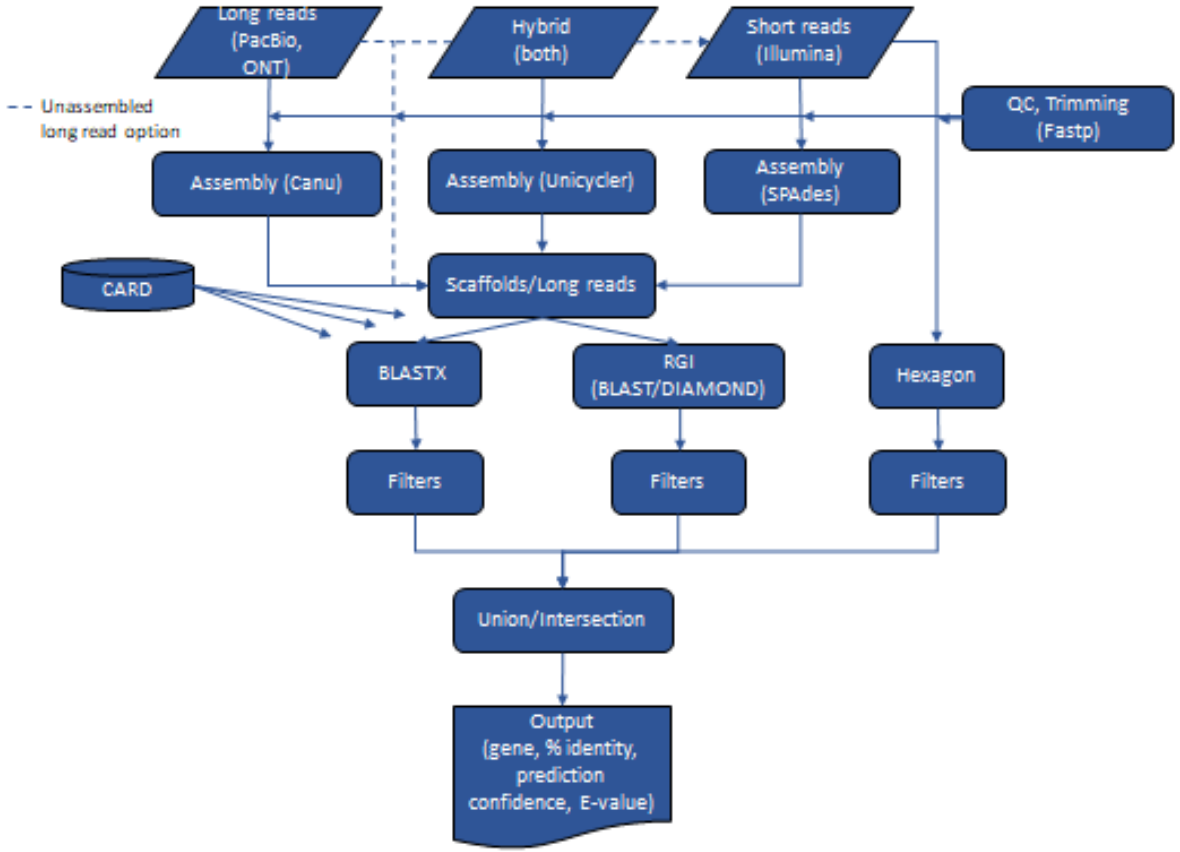
  - DNA seq
  - RNA seq

- Confirm NGS data integrity for downstream analysis:
  - Large volume of datasets
  - High quality data
  - Detect contamination
  - Identification of outliers
  - Sequencing biases (gc bias, homopolymers, duplicates, etc)

Available to Reviewers



Antimicrobial Resistance Pipeline – Hybrid (Short/Long Read) Pipeline



- Designed in collaboration with Paul Carlson (OVRR)

FDA Review

Biologic therapeutics (bacteria, phages) reviewed at FDA
Screen for safety issues, antibiotic resistance genes

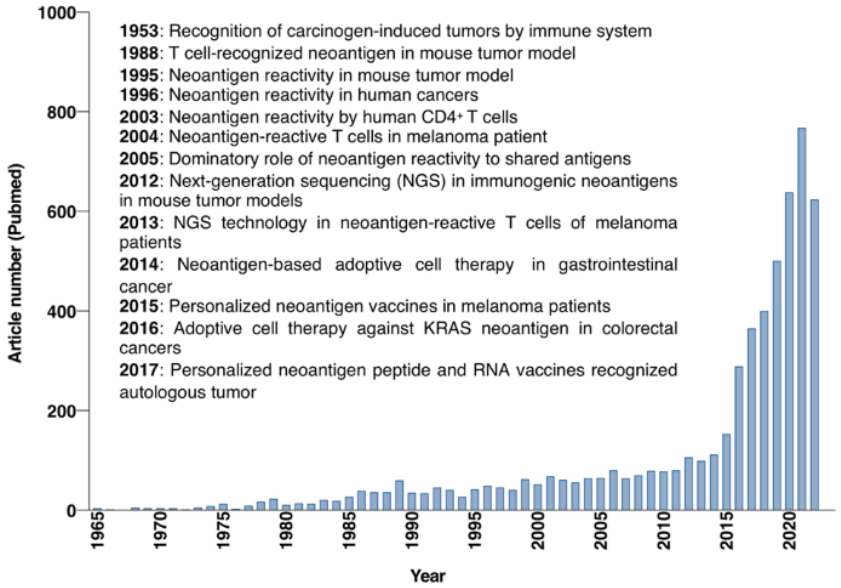
Bioinformatics Objectives

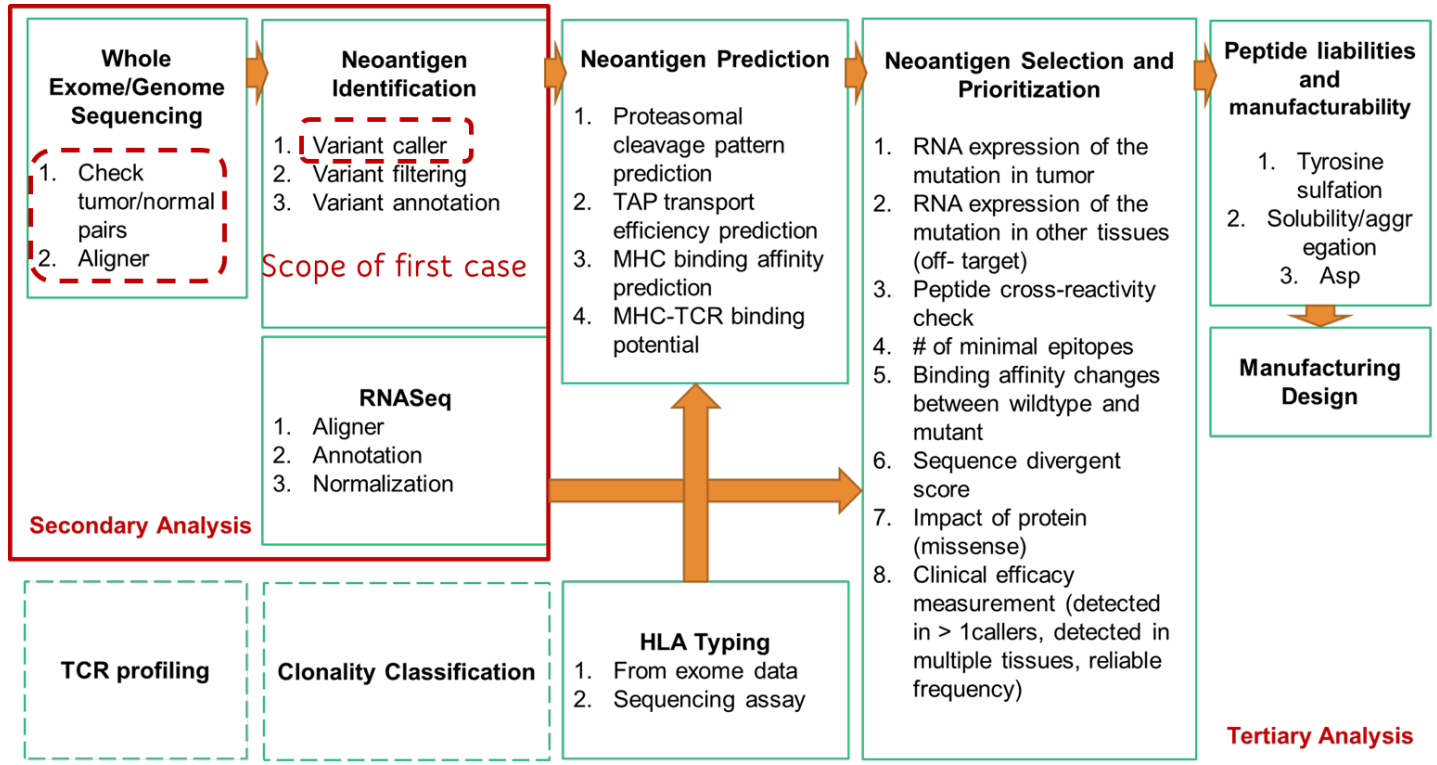
Accept high-throughput sequencing (HTS) reads as input and report potential antimicrobial resistant (AMR) genes present in the biologic product

New Bioinformatics Software packages for neoantigen detection pipeline

Regulatory Review

- Neoantigen discovery for personalized cancer vaccines
  - Neoantigens
    - Self-antigens generated by tumor cells because of genomic mutations
  - Use of neoantigen targets in treatment
    - Cancer vaccine therapy
    - T cell therapy



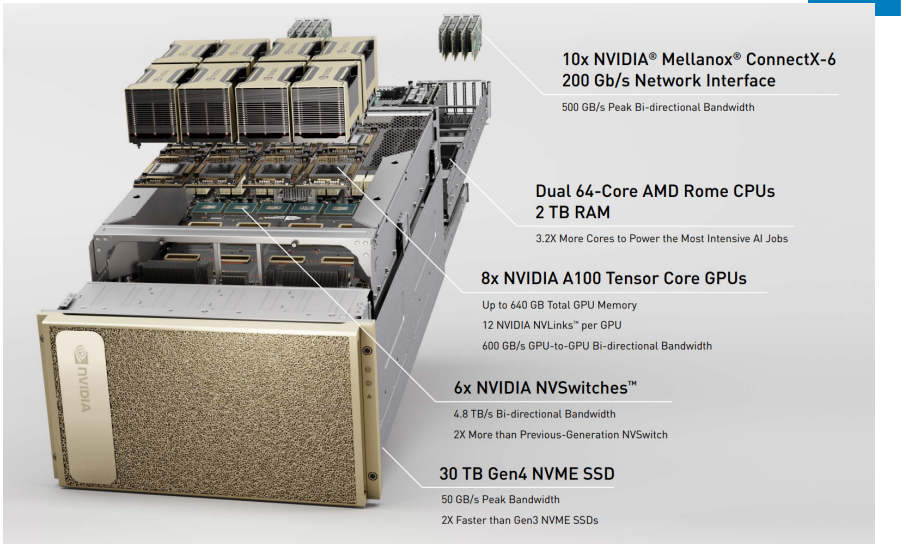


Xie, 2023, Signal Transduction and Targeted Therapy

Flexible Capabilities

Applications using deep learning

Using cutting edge Hardware that supports AI/DL



2x NVIDIA DGX-A100

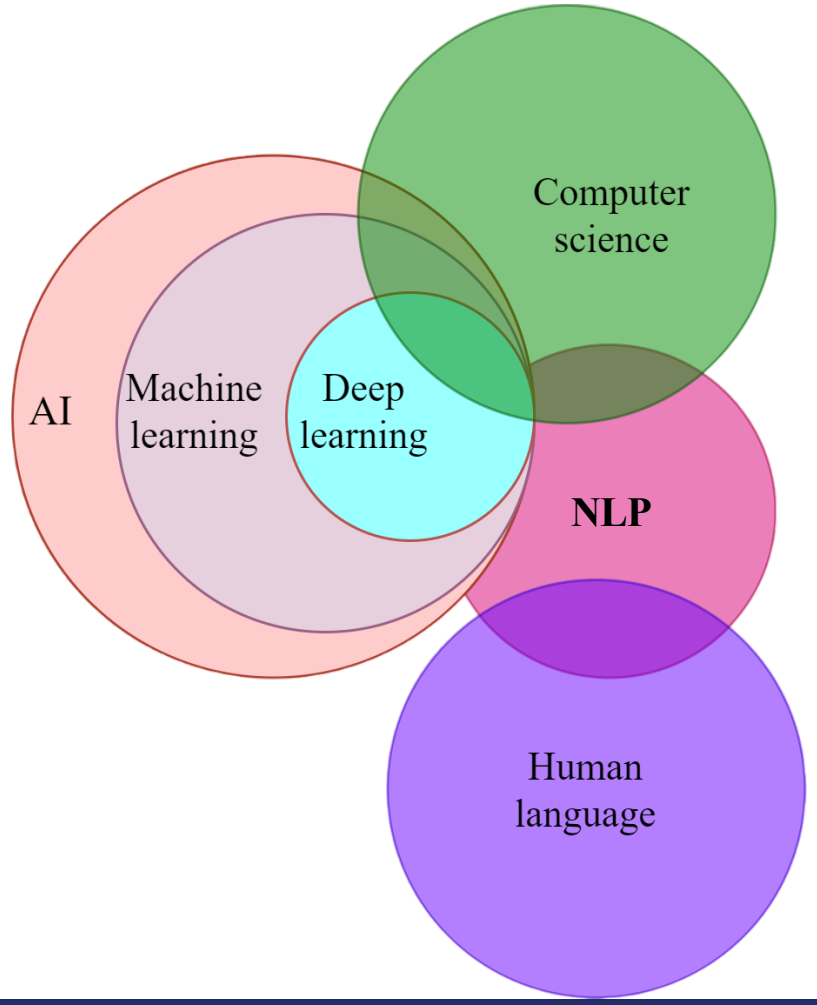
Total A100 GPUs across all HIVE servers: **48**

Supporting:

- Use of Large Language Models
- VAERS report using natural language processing (NLP)
- Docket comments – automatic processing

Natural Language Processing

- Natural language: the language that human beings use to communicate.
- Natural language processing: the idea for computers to derive meaning from natural language.
- NLP is an interdisciplinary research area.

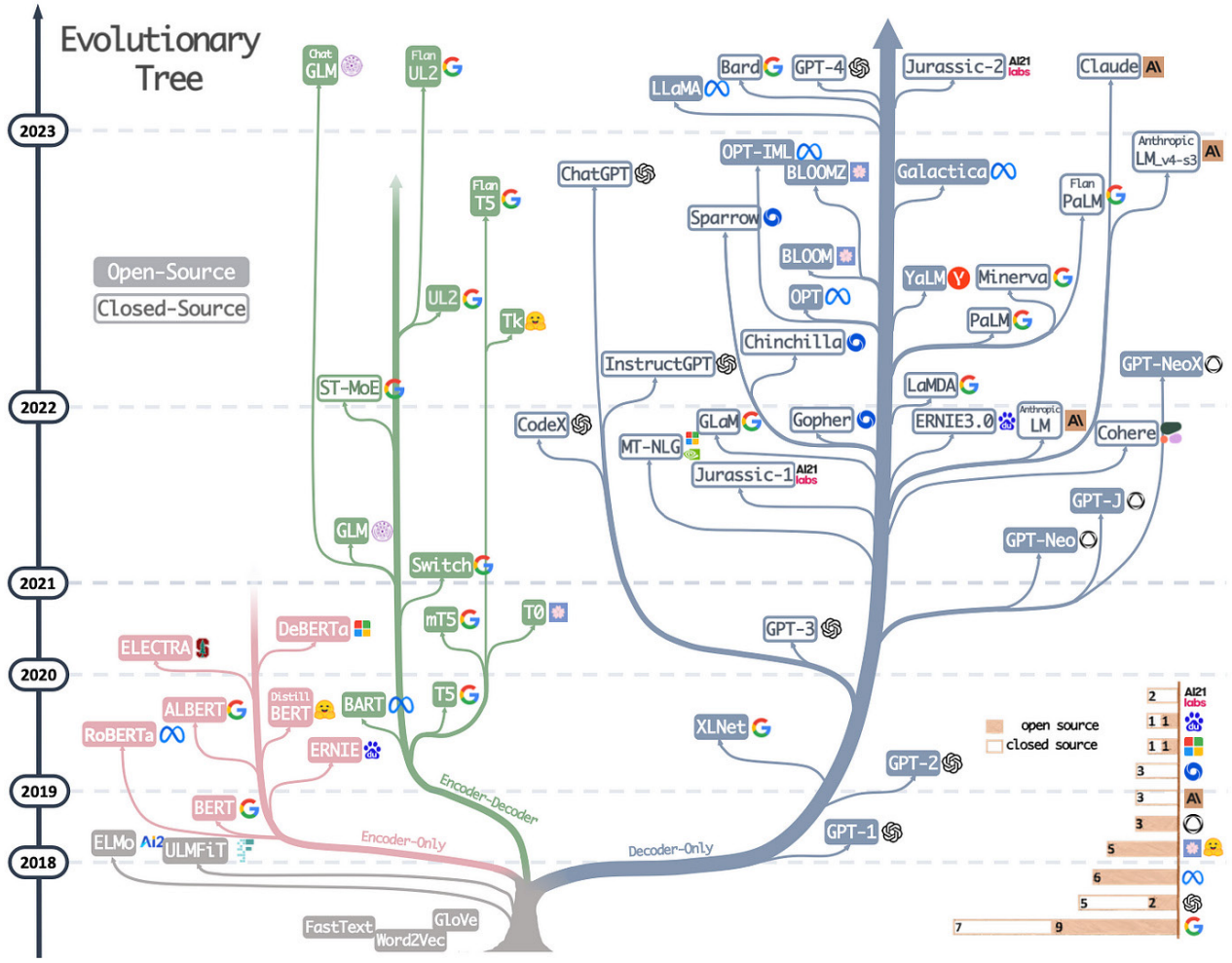


Flexible Capabilities

Large language models

- A deep learning algorithm with the ability to **recognize, summarize, translate, predict, and generate text** and other content based on knowledge gained from massive training datasets.

- The most powerful LLMs are **based on transformers** (computational efficiency when processing sequences in parallel).



Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond - https://doi.org/10.48550/arXiv.2304.13712

LLMs are rapidly advancing

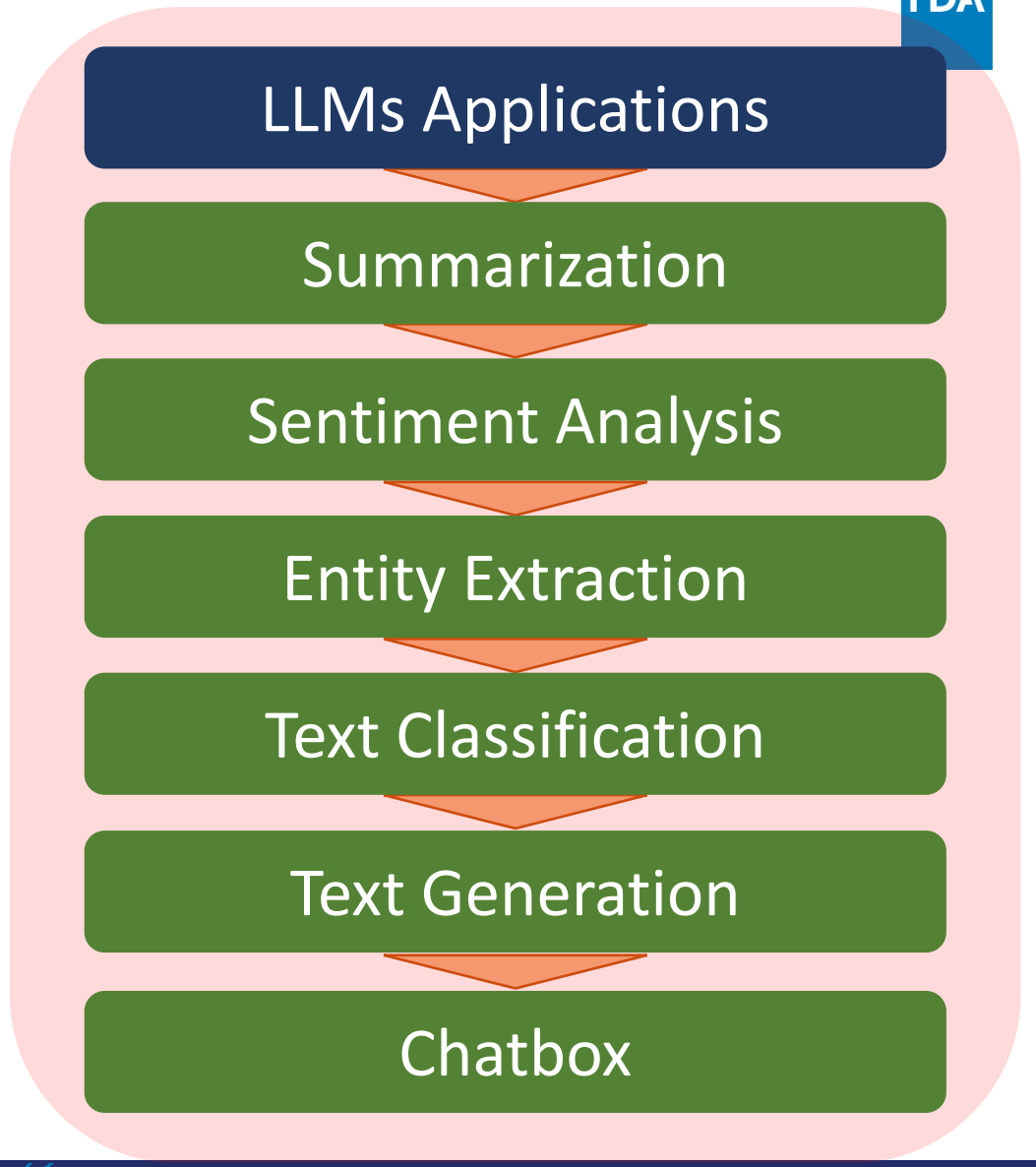
- No single model is the best
- Identify use-cases and apply best model for each case

Privacy

Quality

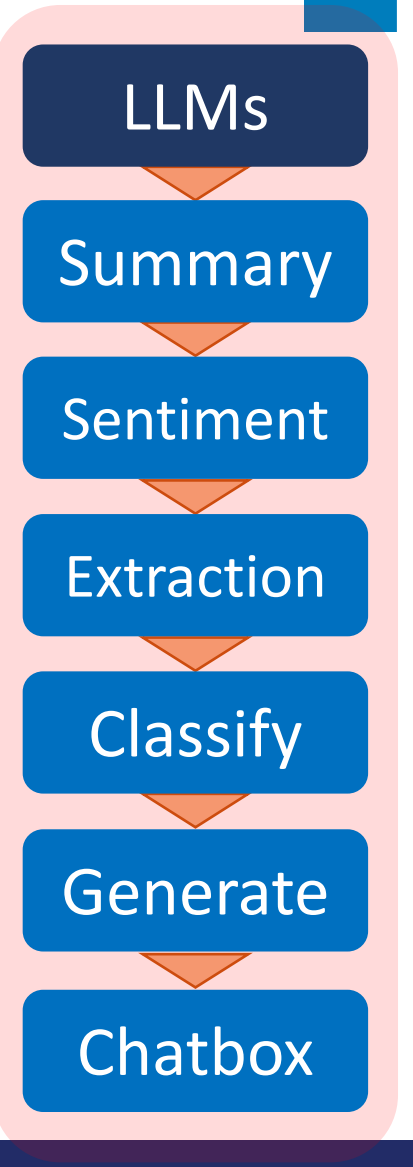
Speed

Cost



Experimental Results using LLMs

- Summarization of Large Documents (> 5,000 words)
- Video-to-Text transcription
- Questions and Answers from documents
- Identification of AI-generated comments
- Use a chatbot



LlaMA2

- Llama 2 is an open-source large language AI model capable of generating text and code in response to prompts.
- Llama 2 was pretrained on publicly available online data sources.
- The fine-tuned model, Llama Chat, leverages publicly available instruction datasets and over 1 million human annotations.
- Llama2 model –
  - Parameters: 13B
  - pre-training tokens: 2 trillion
  - context length: 4096

Update:
Released in July 2023

LLaMA2 analysis on VAERS

Flexible Capabilities

FDA/VAERS (Vaccine Adverse Event Reporting System) is a national vaccine safety surveillance program that helps to detect adverse events for vaccines.

Public VAERS dataset (1000 clinical reports)

Information has been received from a number of health professionals concerning a 12 month old male pt who on 3/19/92 was vaccinated with MMRII and within 7 days it was reported that the pt developed a rash, fever, frequent vomiting and diarrhea. Also, within a few weeks, post vax, the pt developed various infections of the ear, throat, and chest. Finally, the pt was dx'd with autism. Upon internal review, autism was considered to be "an important medical event." His exams clearly came within the dx of autism. We felt he demonstrated repetitive and absorbing interests, used his visual skills better than auditory ones, was fascinated by words, displayed little self-help behavior or pretend play and has a deviant pattern for feeding. His language is disordered with minimal understanding of speech, and a need for contextural and situational cues or gesture in order to make appropriate response. We indicated that he will likely always have communication difficulties within the autistic spectrum and that appropriate educational input was absolutely vital. His developmental progress first concerned parents around 18 to 22 months of age. His progress before this was normal with independent walking by 13 months and 2 words by fifteen months. He was a happy and affectionate baby. F/U 1/19/98 the pt was seen by a pediatrician.She noted that the parents became concerned about their son's progress round about 18-22 months. at the time of the visit it was indicated that the pt still required some help with dressing himself. He did not like being weighed and indicated the need to use toilet. During the visit he asked for a tissue after using the toilet. He was reported a fussy eater. It was noted that there was no family history of behavioral problems or learning difficulties. The pt's mother suffered from kidney problems. On exam his weight of 39.7kg eas on the 97t percentile. He did not cooperate for having his height measured. His head circumference was 55.5cms which ws on the 90th percentile.

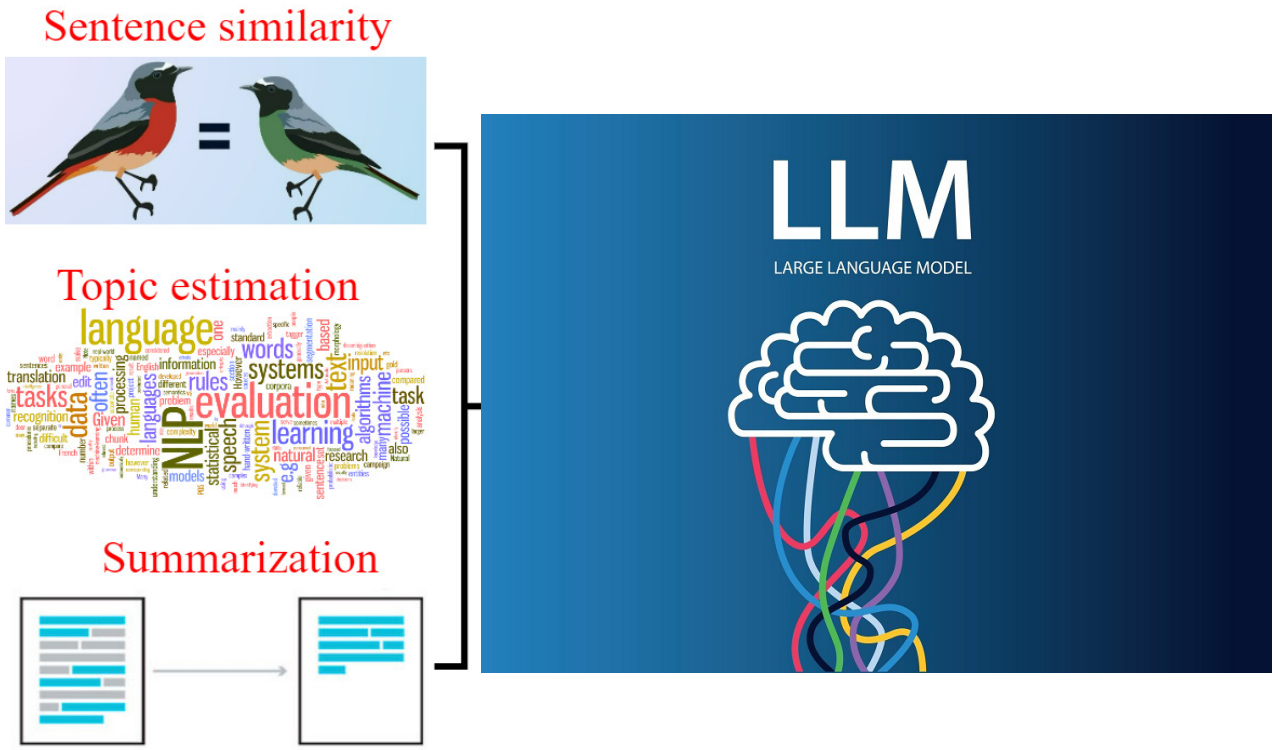
```
['age: 12 months,',
 'gender: male,',
 'vaccination_history: [',
 'MMRII on 3/19/92,',
 'developed symptoms within 7 days',
 '],',
 'symptoms: [',
 'rash,',
 'fever,',
 'frequent vomiting,',
 'diarrhea,',
 'various infections of the ear, throat, and chest',
 '],',
 'dx: autism,',
 'developmental_progress: [',
 'normal until 18-22 months,',
 'concerns raised by parents at 18-22 months,',
 'requires help with dressing,',
 'does not like being weighed,',
 'uses toilet but requires assistance',
 '],',
 'family_history: [',
 'no history of behavioral problems or learning difficulties',
 '],',
 'medical_history: [',
 'mother suffers from kidney problems',
 '],',
 'physical_exam: [',
 'weight of 39.7kg (97th percentile),',
 'head circumference of 55.5cms (90th percentile)',
 ']']
```

FDA/VAERS and potential Use Cases with AI:

- Try the use of LLMs as an alternative to extract text entities from the reports.
- Summarization in specific format (requested by Medical Officers)
- Incorporation of MeDRA terms (vaccine names and synonyms)
- Highlighting adverse events not on the product label.
- Query model with conversational questions and answers about adverse events

Docket comment processing using NLP

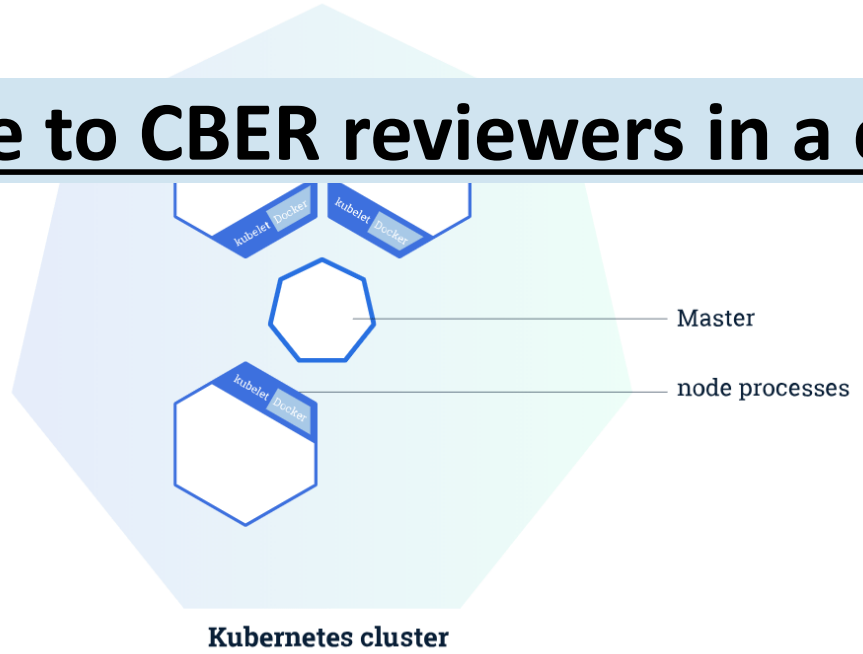
- Problem:
  - FDA receives information in DOCKETS
  - Many DOCKETS have thousands of comments
  - Manual inspection is time consuming.
- Goal: To use NLP and AI to automatically:
  - Identify and remove **duplicated** comments.
  - Identify smaller sub-groups of **related** comments.
  - Understand **sentiments/agreements** of docket comments.
  - Create **summary** of long docket comments.

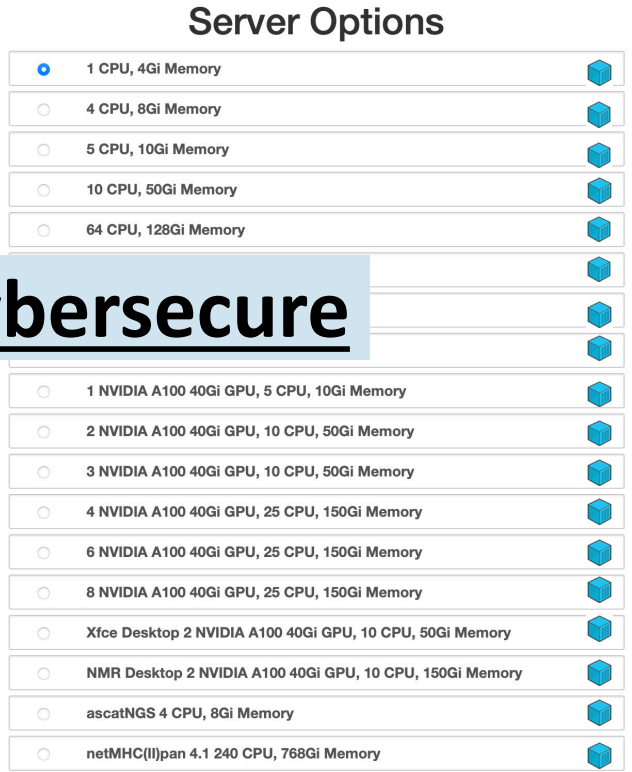


Containers

A container is a package with the tool(s) and its dependencies, so the application runs and is portable

- **Receive Containers from Sponsors**
- **Made available to CBER reviewers in a cybersecure environment**





Summary

- HIVE Ecosystem – Supports overall FDA IT infrastructure of cloud and edge computing

Regulatory Review

Provide a venue for the exploration of rapidly evolving technologies like LLMs

- Docket comments
- VAERS

Sharing Data

Sharing data across FDA as well as outside FDA with the HIVE platform

Unified HPC at FDA

Use of containers to create packages and distribute applications. Platforms like Docker and Kubernetes make this task easier to manage

Acknowledgments

David Menschik, MD, MPH

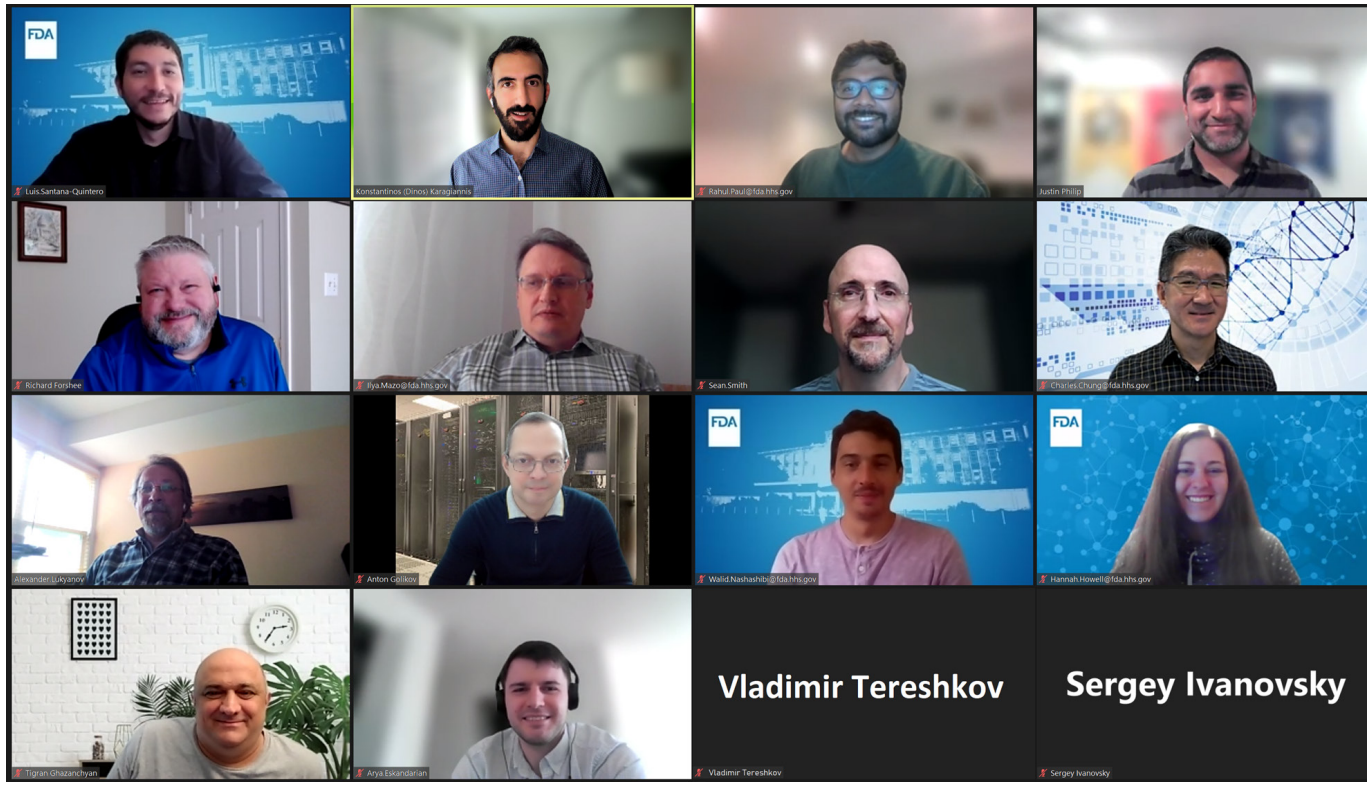
Mark Walderhaug, PhD

Rama Rayavarapu, PhD

Harinder Chahal, PharmD, MSc

OBPV Director
Steven Anderson, PhD

Deputy Office Director
Richard Forshee, PhD



HIVE TEAM

Rahul Paul, PhD

Arya Eskandarian

Anton Golikov

Sean Smith, PhD

Charles Chung, PhD

Alexander Murray

Tigran Ghazanchyan

Justin Philip

Sergey Ivanovsky

Spyros Karaiskos, PhD

Alexander Lukyanov

Konstantinos Karagiannis*

Hyok Song

Hannah Howell *

Ilya Mazo, PhD