PICI INTERNSHIP 2025

CATERINA PONTI

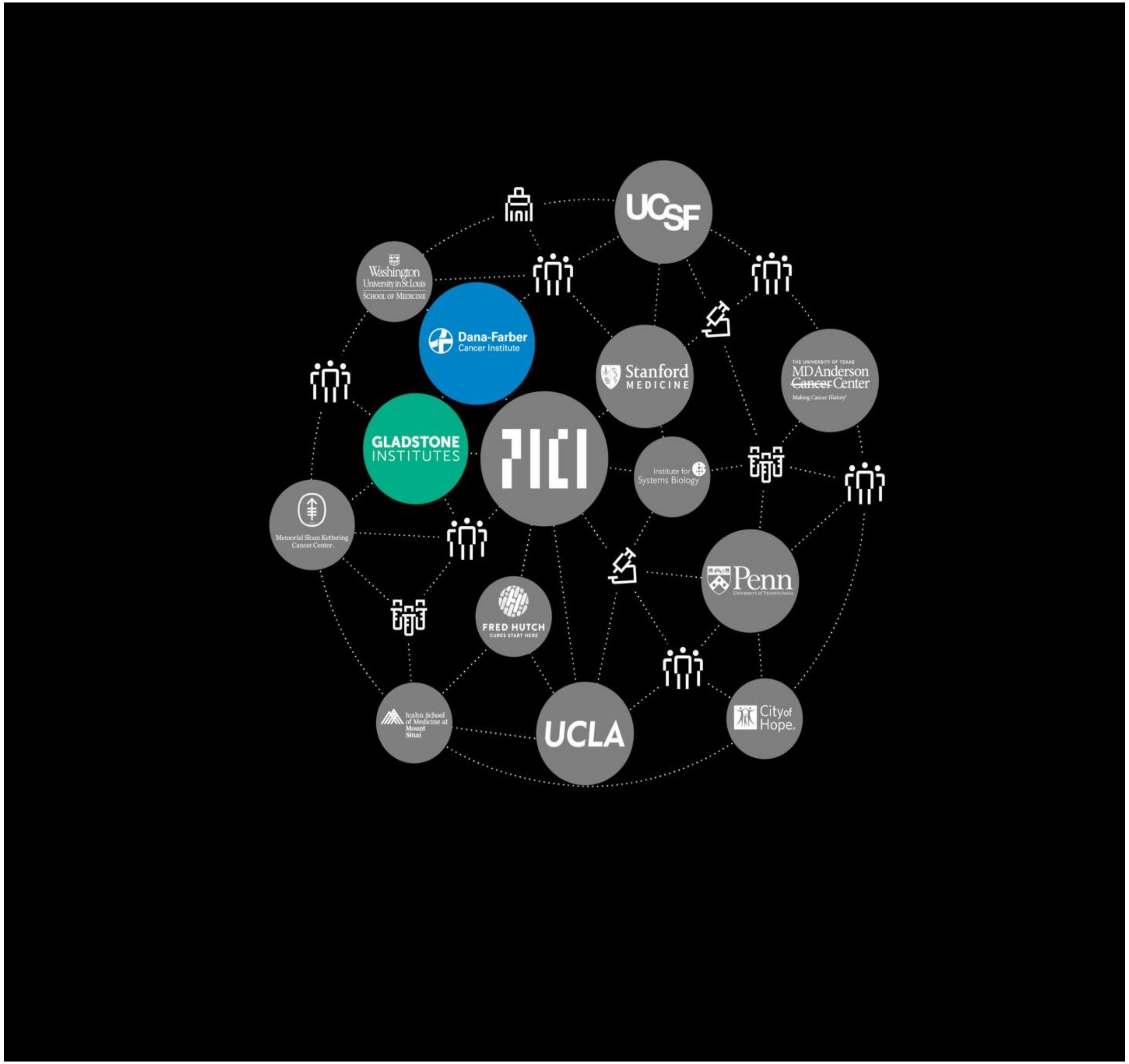
Data Science and Bioinformatics Intern

PARKER INSTITUTE FOR CANCER IMMUNO THERAPY (PICI)

Our mission is to accelerate the development of breakthrough immune therapies to turn all cancers into curable diseases.

WHAT IS PICI?

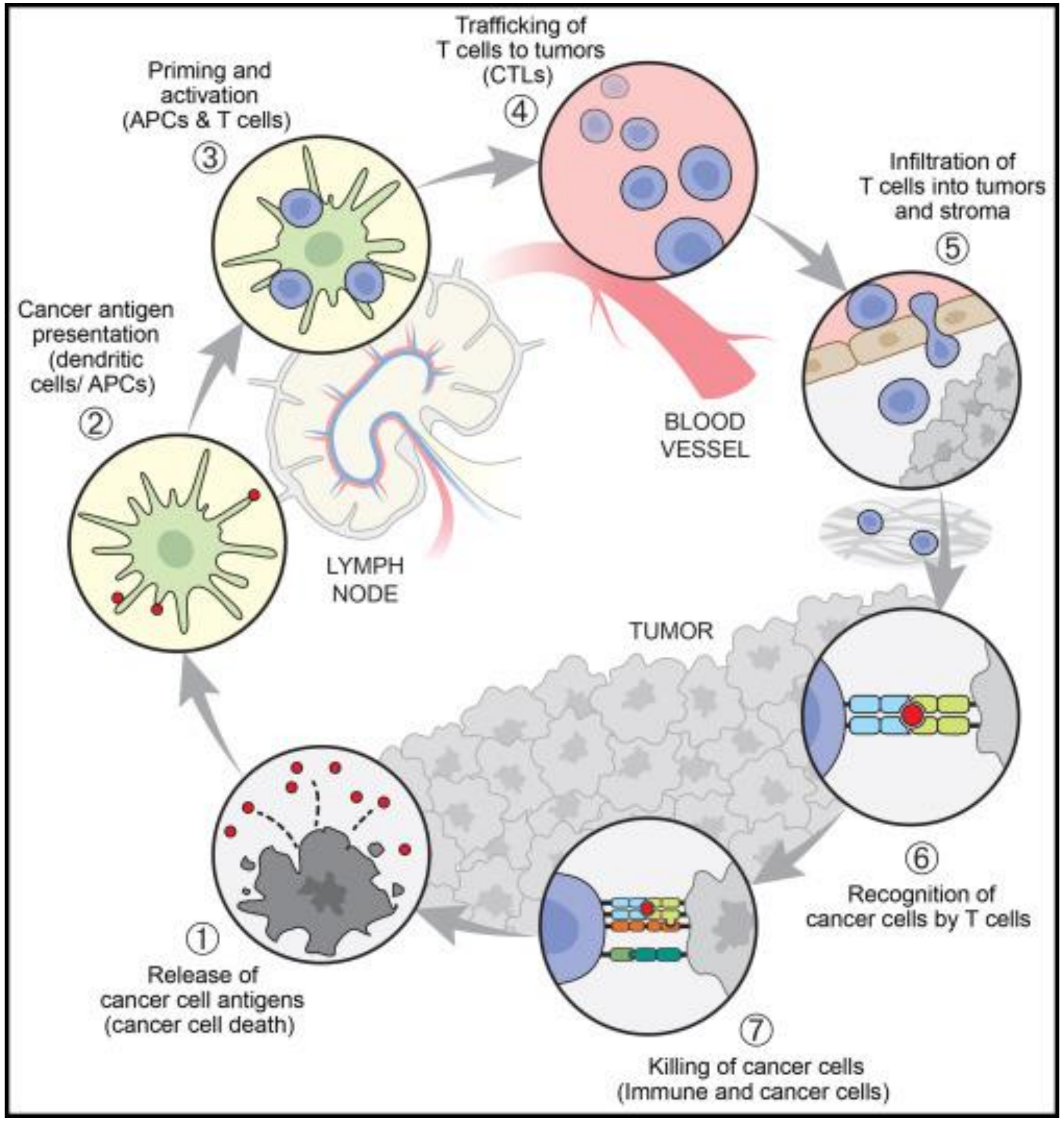
- Patient-centric, grant-giving non-profit organization
- Fund & accelerate high-impact research
- Drive collaboration (700+ Cross-Institutional PICI Investigators)
- Enable bold Innovations


UCSF
Washington University in St. Louis School of Medicine
Dana-Farber Cancer Institute
Stanford Medicine
The University of Texas MD Anderson ~~Cancer~~ Center Making Cancer History
Gladstone Institutes
PICI
Institute for Systems Biology
Memorial Sloan Kettering Cancer Center
Penn University of Pennsylvania
Fred Hutch Cures Start Here
Icahn School of Medicine at Mount Sinai
UCLA
City of Hope

THE CANCER-IMMUNITY CYCLE

1. **Release:** Dying cancer cells release "markers" (Antigens).
2. **Presentation:** Specialized scout cells (Dendritic cells) capture these markers.
3. **Priming:** Scouts travel to lymph nodes to "teach" T-cells what the cancer looks like.
4. **Trafficking:** Newly trained T-cells enter the bloodstream to find the tumor.
5. **Infiltration:** T-cells exit the blood and "invade" the tumor tissue.
6. **Recognition:** T-cells identify the specific cancer cells using the markers from Step 1.
7. **Killing:** T-cells destroy the cancer cells, which releases *more* markers, restarting the cycle at Step 1.

Where the System Breaks Down: Cancer cells can **turn off** T cells by using “brakes” on the immune system called immune checkpoints -> this stops T cells from attacking



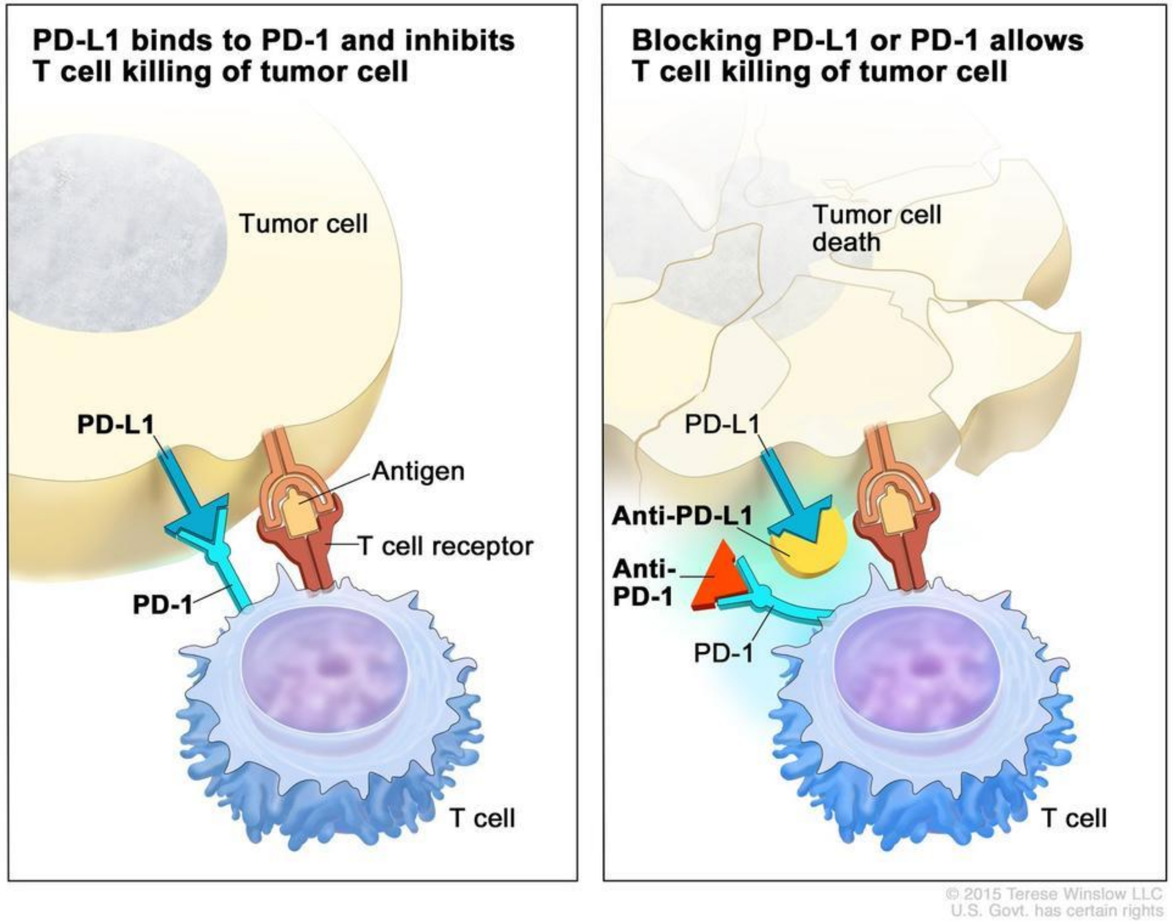
IMMUNE CHECKPOINT INHIBITOR (ICI) THERAPY

Goal: To restart the Cancer-Immunity Cycle by removing the biochemical "brakes" (checkpoints) that cancer uses to survive.

- Anti–PD-1 / Anti–PD-L1 (e.g., pembrolizumab, atezolizumab)
- Anti–CTLA-4 (e.g., ipilimumab)

Clinical Challenges

- Immune-related Adverse Events (iRAEs) Because we are "releasing the brakes," the immune system can become overactive..
- It may mistakenly attack healthy organs (e.g., lungs, skin, or colon), leading to inflammation.


PD-L1 binds to PD-1 and inhibits T cell killing of tumor cell
Tumor cell
PD-L1
Antigen
T cell receptor
PD-1
T cell
Blocking PD-L1 or PD-1 allows T cell killing of tumor cell
Tumor cell death
PD-L1
Anti-PD-L1
Anti-PD-1
PD-1
T cell
© 2015 Terese Winslow LLC U.S. Govt. has certain rights

RADIOHEAD Study Planning

Study Support and Goals:

- Funded by the Helmsley Charitable Trust, JDRF, and Bristol Myers Squibb.
- **Goal:** Identify **biomarkers** to prevent or intervene in:
  - Cancer immunotherapy-induced **Type 1 Diabetes**.
  - Severe **Immune-Related Adverse Events** (irAEs).

Academic Guidance and Study Direction:

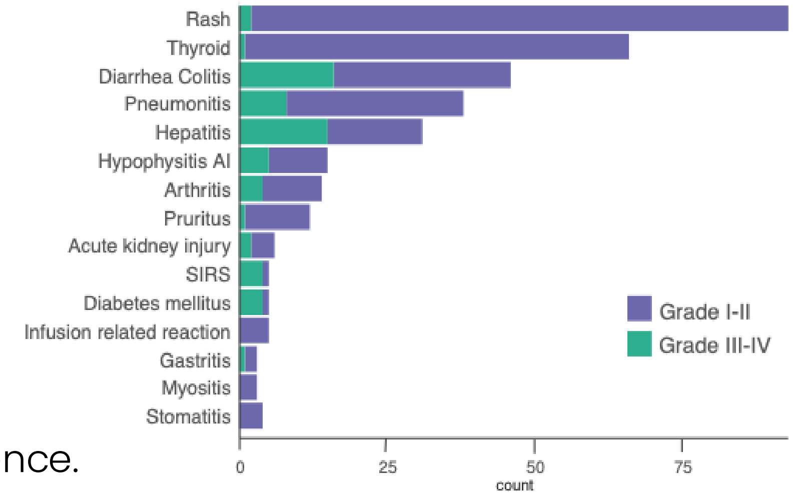
- UCSF: data collection and trial objectives.
- Huntsman Cancer Institute: data cleaning and curation work.

Key Challenges:

- Recruitment for a Rare Event was difficult: 12 Expected, 3 Recruited.

Root Cause Analysis:

- Patient Behavior
- COVID Impact
- Technical Issues
  - Difficulty in diagnosing T1D in community centers lacking specialized experience.



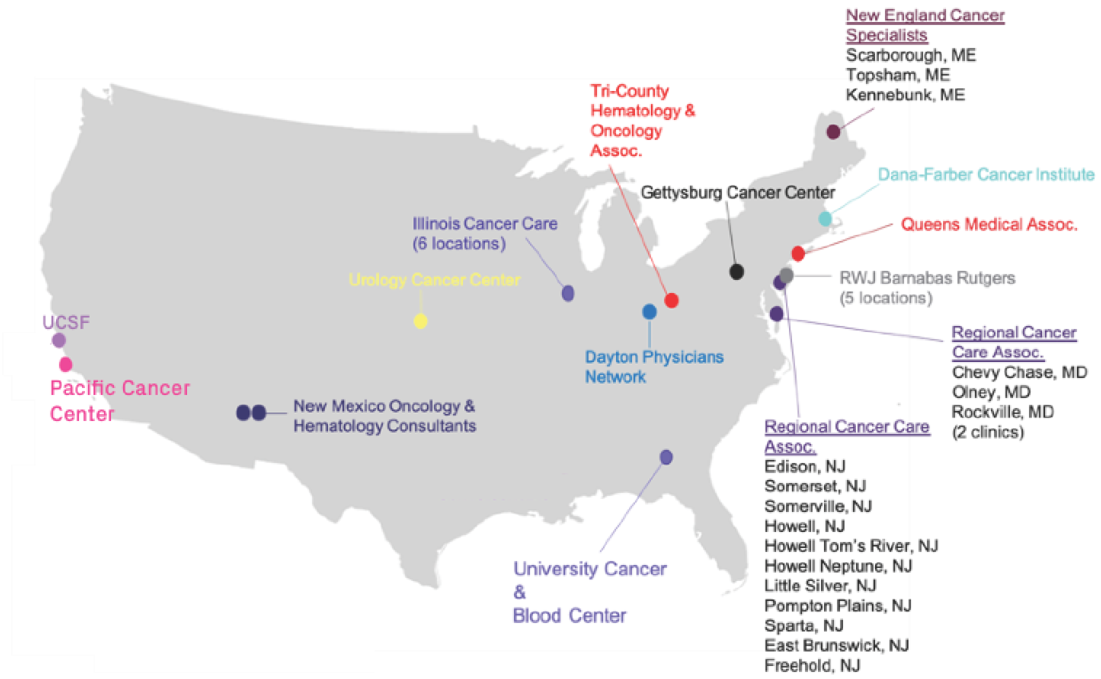
New Focus: Association between all Immune-Related Adverse Events (irAEs) and Cancer Treatment Response.

RADIOHEAD Sample Collection

- Standardized Processes: All blood samples shipped overnight to ensure consistency across sites.
- Central Laboratory: Usage of a single central facility to eliminate processing variables.
- Site Selection Strategy
- Batch Effect Analysis

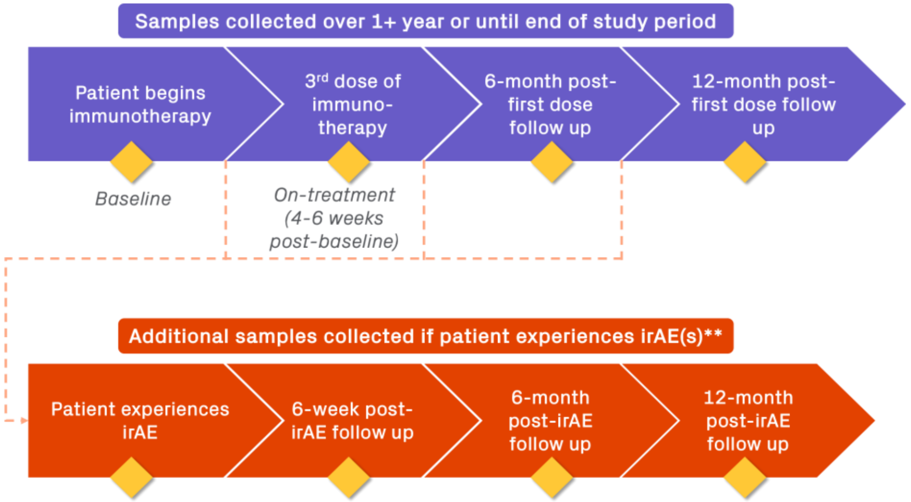
Systemic Data Biases

- Geographic Bias: study privileges expensive, resource-rich urban hubs
- Demographic Gap: underrepresentation of minority
- Economic Barriers: Cost of Access
- IRAE Potentially underreported
- Lack of Longitudinal Data (patient follow up is inconsistent because of pandemic)
- Clinical Limitations: additional disease information was often a free textbox



What Makes RADIOHEAD *Unique* ?

- Scale and Data:
  - Data collected from 1,070 patients (~2.5 draws per patients).
  - Generated over 70,000 samples across 3,500+ combinations of timepoints.
  - Samples were collected over one year.
- Event-Driven Sampling
  - Additional samples collected in the event of an adverse event (irAE)


Samples collected over 1+ year or until end of study period

Patient begins immunotherapy → 3rd dose of immuno-therapy → 6-month post-first dose follow up → 12-month post-first dose follow up

Baseline

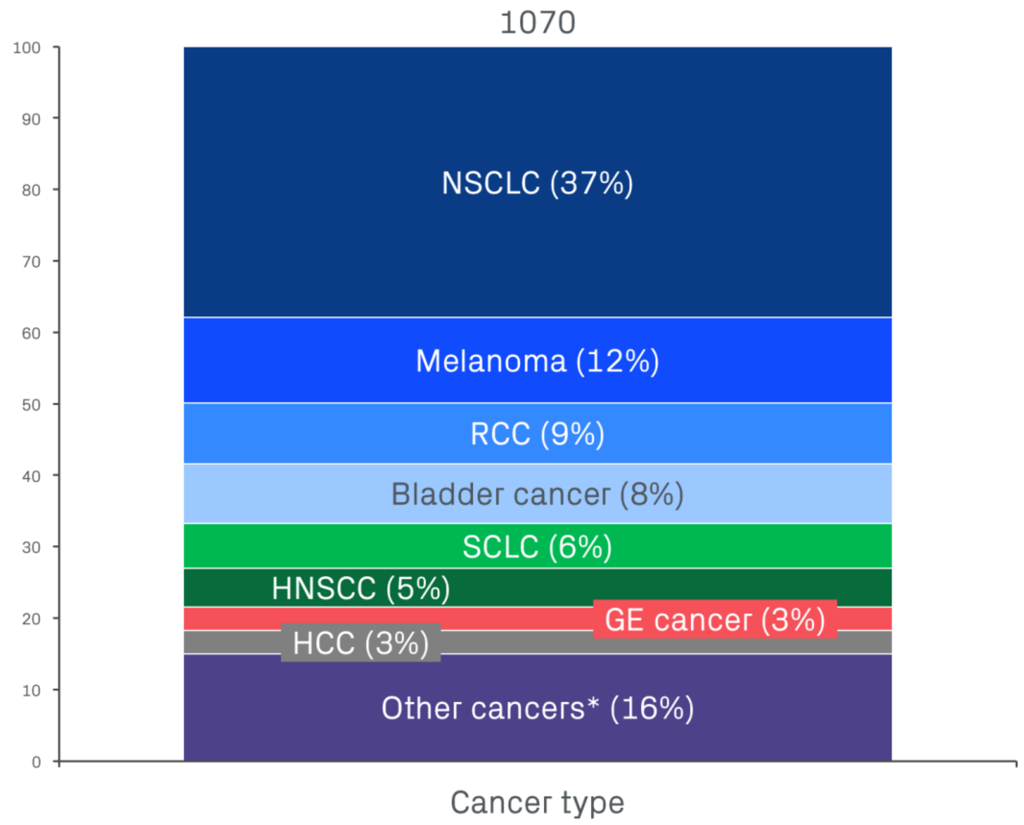
On-treatment (4-6 weeks post-baseline)

Additional samples collected if patient experiences irAE(s)**

Patient experiences irAE → 6-week post-irAE follow up → 6-month post-irAE follow up → 12-month post-irAE follow up

1070 Patients 3500+ Timepoints	70K+ Sample Aliquots Banked
Patient Information	Whole Blood
Treatment Detail	Plasma
Outcomes Data	Serum
Additional Metadata	PBMC

What Makes RADIOHEAD *Unique* ?



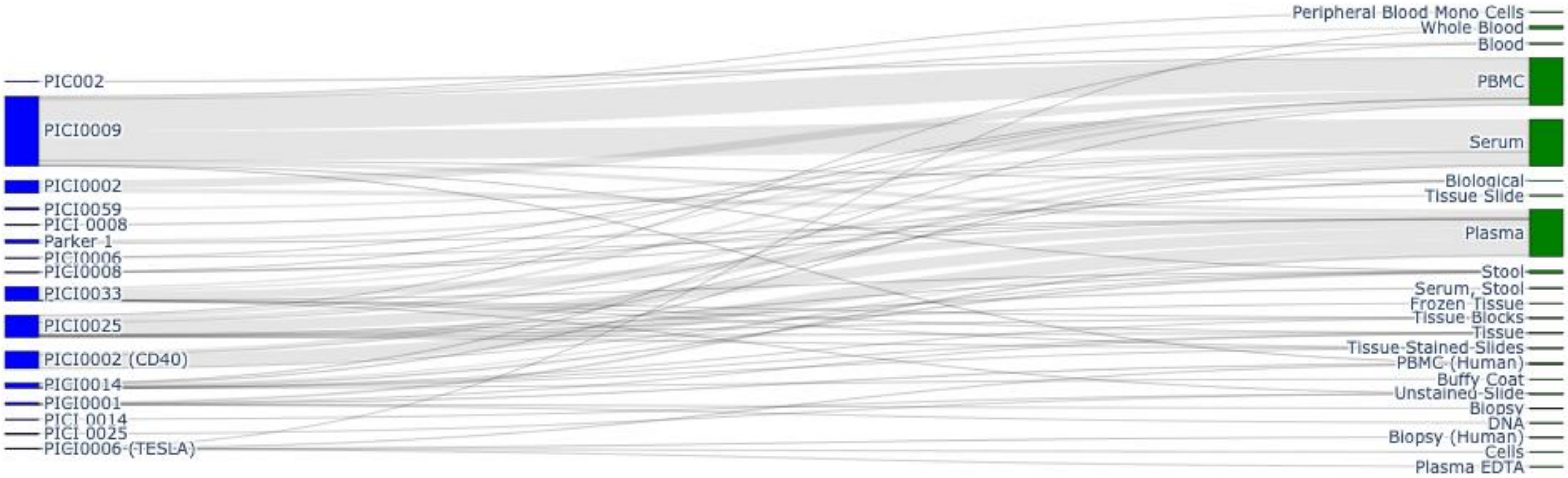
Distribution of patient tumor types in RADIOHEAD dataset
Number of patients

Comprehensive Multiomic Analyses

ctDNA	~750 genes, completed	GUARDANT
Serum Proteomics	600+ proteins, completed	nomic
WES / SNP-panel	completed	Genentech A Member of the Roche Group Teiko.bio
HD Flow Cytometry	80 markers, in progress (complete by mid-2025)	BostonGene
Bulk-RNAseq	In progress (complete by mid-2025)	
Single cell RNA-seq	Projected to start in April 2025	immunai
Metabolomics	Evaluating partners	

BIOREPOSITORY

BIOREPOSITORY 2024 PULL



- The biorepository includes samples from 10 additional PICI-owned clinical trials.
- Limitations: inconsistent naming conventions and duplicate entries.
  - I developed a **data clean-up and standardization script** to support future data pulls.
- RADIOHEAD (PICI-009) is a high-priority study due to its external partnerships and multi-omics integration capabilities.

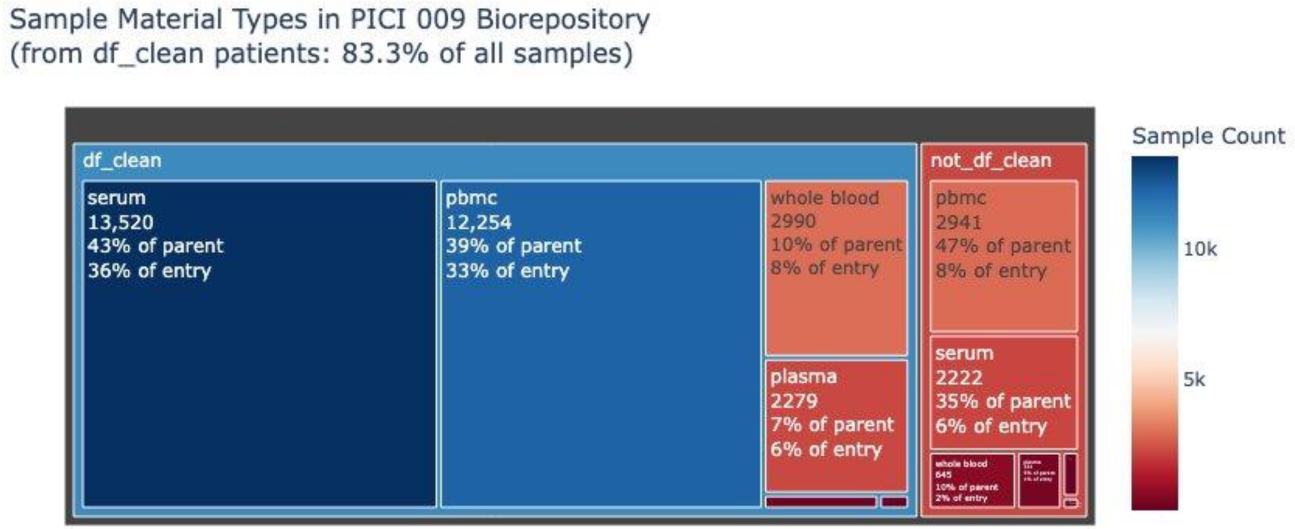
RADIOHEAD DATASETS

Biorepository

Multiple pulls: keeping track of sample shipments

REDCap: clinical annotations + samples' barcodes of patients (1,300) collected during clinical trial

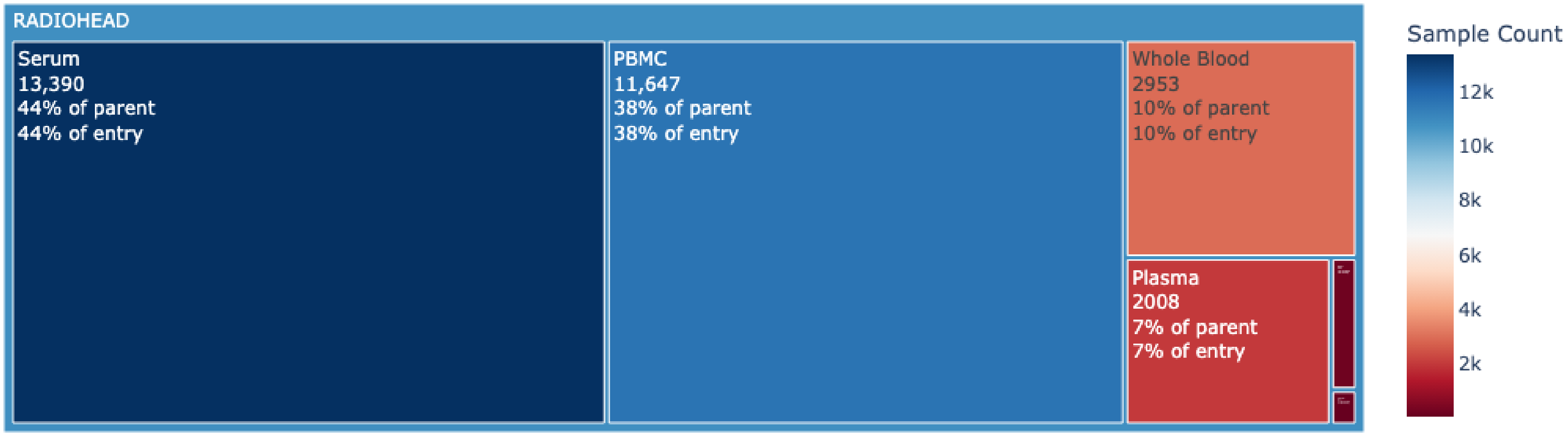
df_clean: cleaned REDCap with 1,070 patients of all patients with a pretreatment sample



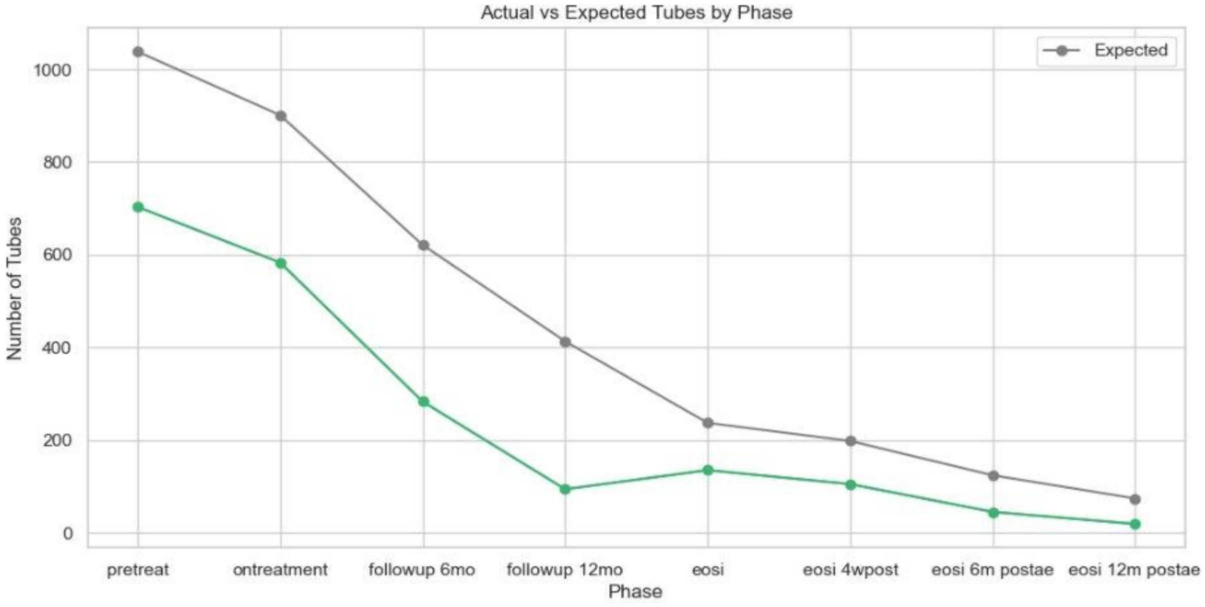
Sample Material Types in PICI 009 Biorepository
(from df_clean patients: 83.3% of all samples)

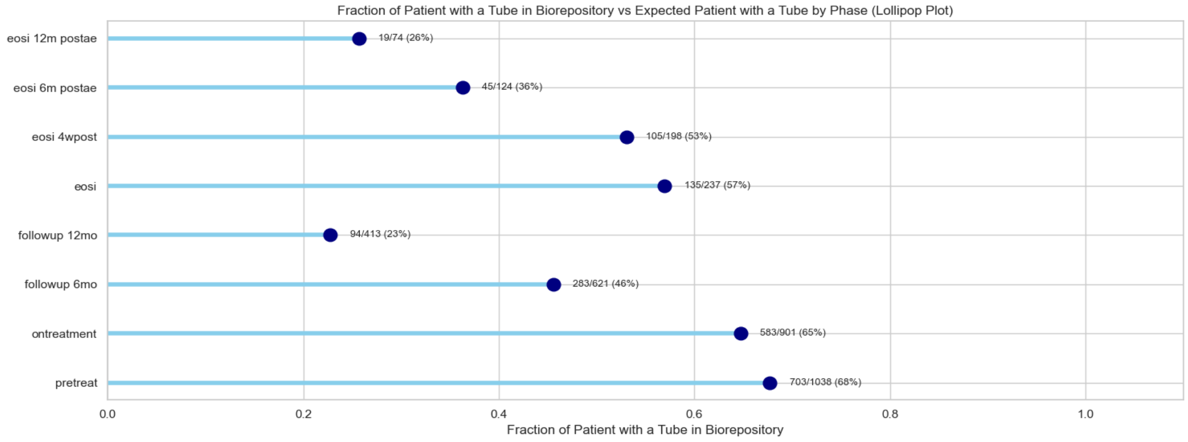
BARCODES MATCHING AND OUTLIER DETECTION (RADIOHEAD)

Sample Material Types in Biorepository from the 1070 RADIOHEAD patients selected in df_clean.



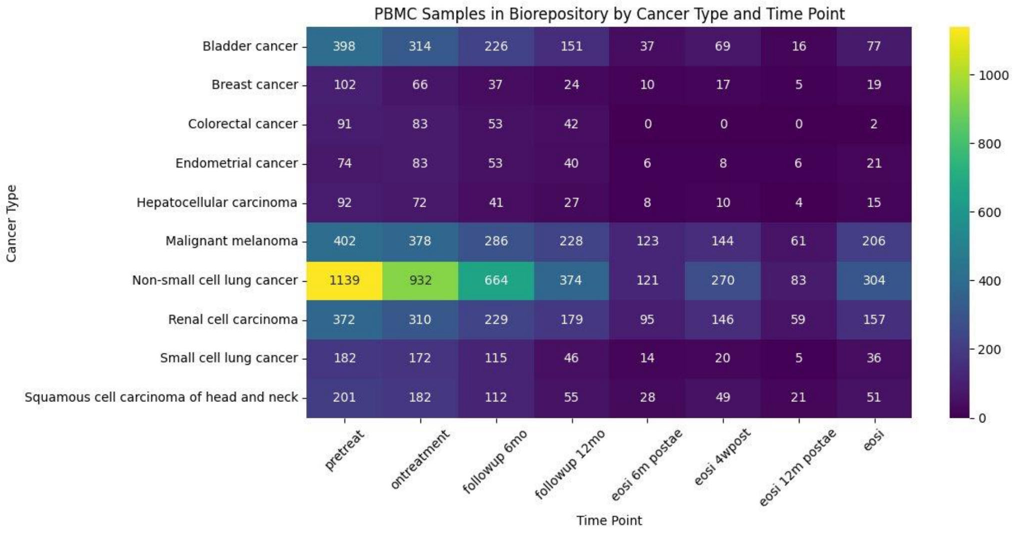
- Cleaned and standardized barcodes using custom parsing rules.
- Detected outliers (barcodes assigned to the wrong project name) using pairwise **Levenshtein distances** (via RapidFuzz) via RapidFuzz to compare alphanumeric barcodes.
- Matched barcodes between biorepository and REDCap using exact and substring matching.

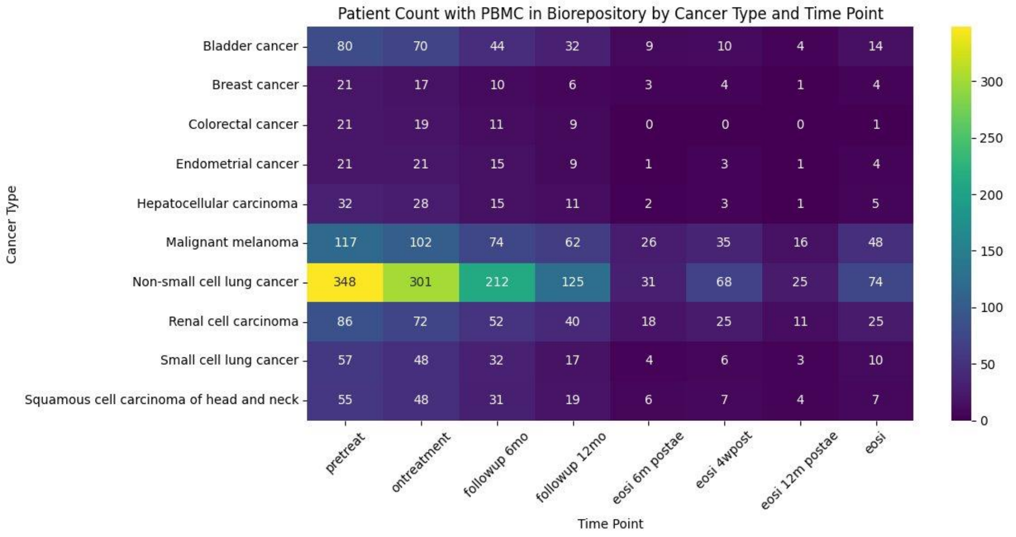




SAMPLE COUNT AND PATIENT BY TIMEPOINT ACTUAL VS. EXPECTED (FROM REDCAP)

- Gap between the expected samples tracked in REDCap and the actual samples available in the biorepository as of 2025
- Percentage of expected tubes collected per patient during clinical trial vs samples available now in the biorepository





SAMPLES AND PATIENT COUNT BY TIMEPOINT AND CANCER TYPE

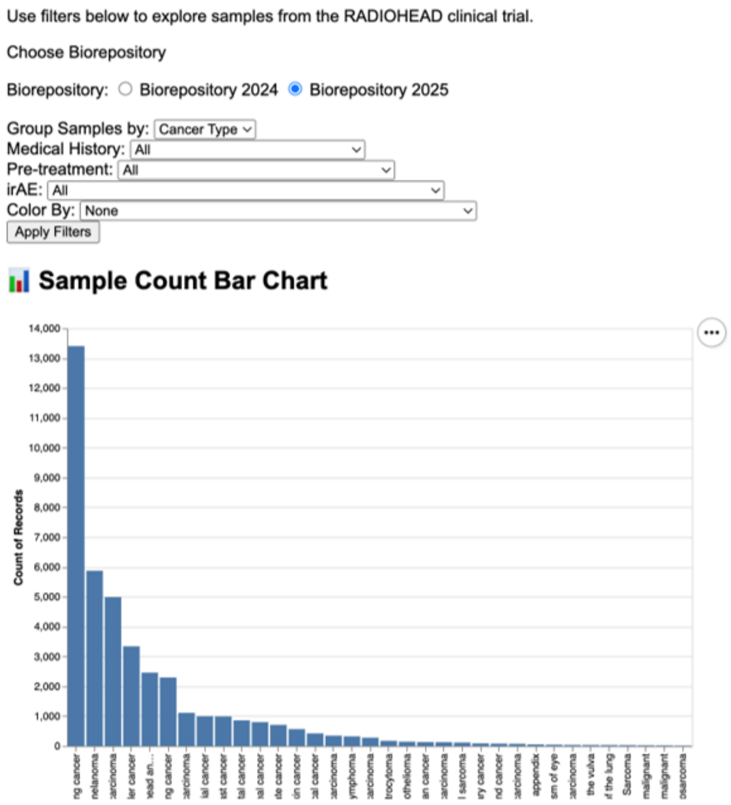
1. Counts unique samples and patients remaining in the biorepository, grouped by cancer type and timepoint.

2. Provides a clear view of sample availability across disease groups and visits.

3. Helps scientists quickly identify gaps, trends, and well-represented cohorts when planning analyses or selecting samples.

WHY MOVING FROM SCIRPTS TO A WEB PLATFORM?

- Internal Efficiency: Automated scripts to replace manual queries
- Partner Collaboration
- Real-Time Updates
- User Friendly



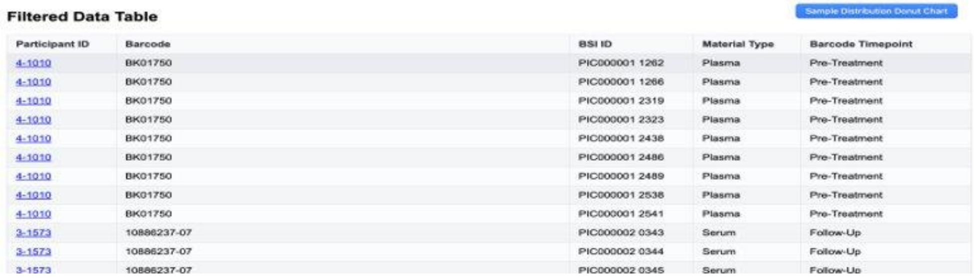
Use filters below to explore samples from the RADIOHEAD clinical trial.

Choose Biorepository

Biorepository: ○ Biorepository 2024 ● Biorepository 2025

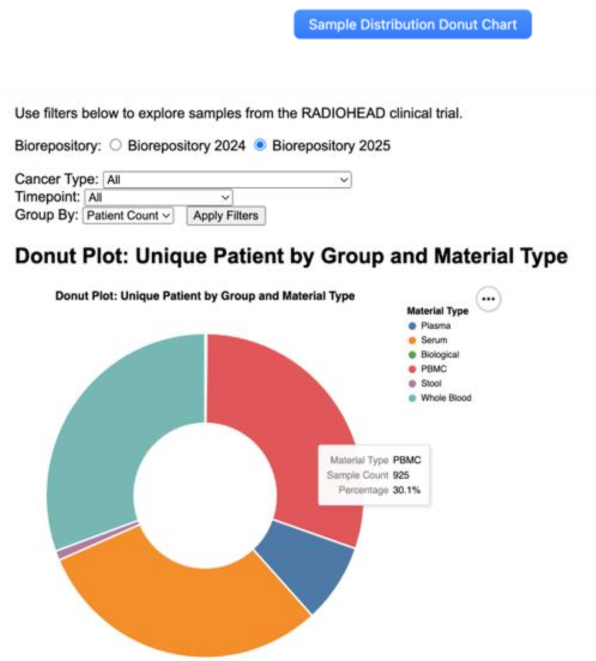
Group Samples by: Cancer Type
Medical History: All
Pre-treatment: All
irAE: All
Color By: None

📊 Sample Count Bar Chart



Filtered Data Table

Participant ID	Barcode	BSI ID	Material Type	Barcode Timepoint
4-1010	BK01750	PIC000001 1262	Plasma	Pre-Treatment
4-1010	BK01750	PIC000001 1266	Plasma	Pre-Treatment
4-1010	BK01750	PIC000001 2319	Plasma	Pre-Treatment
4-1010	BK01750	PIC000001 2323	Plasma	Pre-Treatment
4-1010	BK01750	PIC000001 2438	Plasma	Pre-Treatment
4-1010	BK01750	PIC000001 2486	Plasma	Pre-Treatment
4-1010	BK01750	PIC000001 2489	Plasma	Pre-Treatment
4-1010	BK01750	PIC000001 2538	Plasma	Pre-Treatment
4-1010	BK01750	PIC000001 2541	Plasma	Pre-Treatment
3-1573	10886237-07	PIC000002 0343	Serum	Follow-Up
3-1573	10886237-07	PIC000002 0344	Serum	Follow-Up
3-1573	10886237-07	PIC000002 0345	Serum	Follow-Up



Use filters below to explore samples from the RADIOHEAD clinical trial.

Biorepository: ○ Biorepository 2024 ● Biorepository 2025

Cancer Type: All
Timepoint: All
Group By: Patient Count

Donut Plot: Unique Patient by Group and Material Type

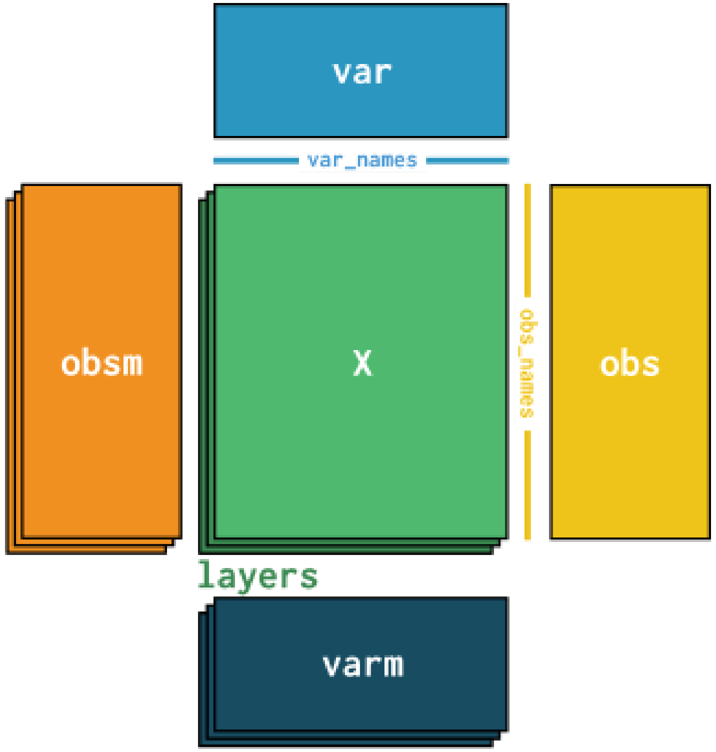
Mass Cytometry Data (CyTOF) from Teiko

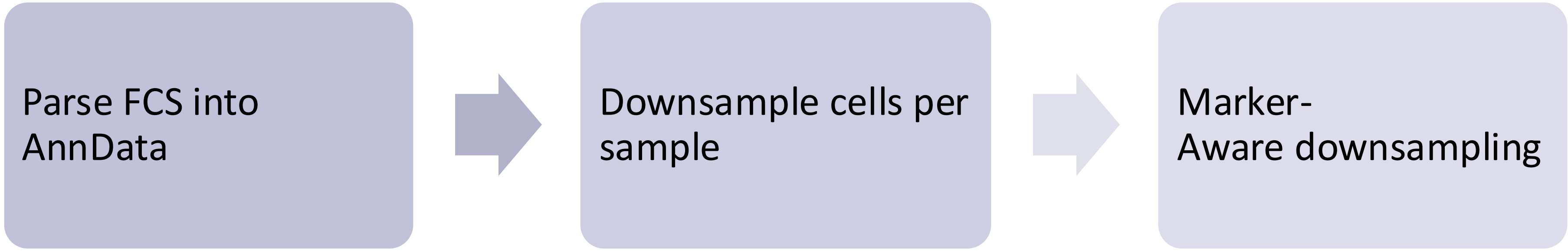
Technology and Overview:

- Uses **antibodies** to quantify up to 50 single-cell biological markers.
- Partnership with **Teiko** completed data acquisition for all melanoma patients.

Raw Data:

- Each sample ~10,000 cells
- High dimensionality: 50 markers x 10,000 cells x 500 samples
- **AnnData** (Annotated Data) object




Parse FCS into AnnData

Downsample cells per sample

Marker-Aware downsampling

The Cloud Solution

Challenges

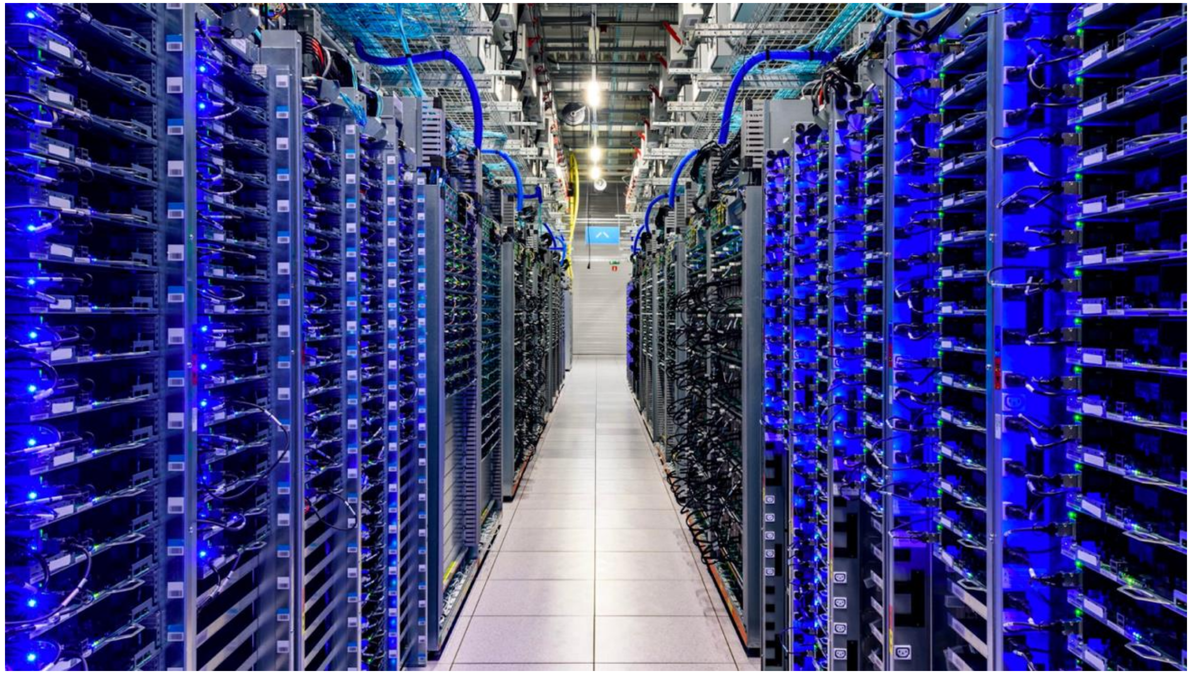
- High-dimensional data → long runtimes & heavy compute
- Running on a laptop took ~8 hrs
- Local workflow limits collaboration

The Cloud Solution

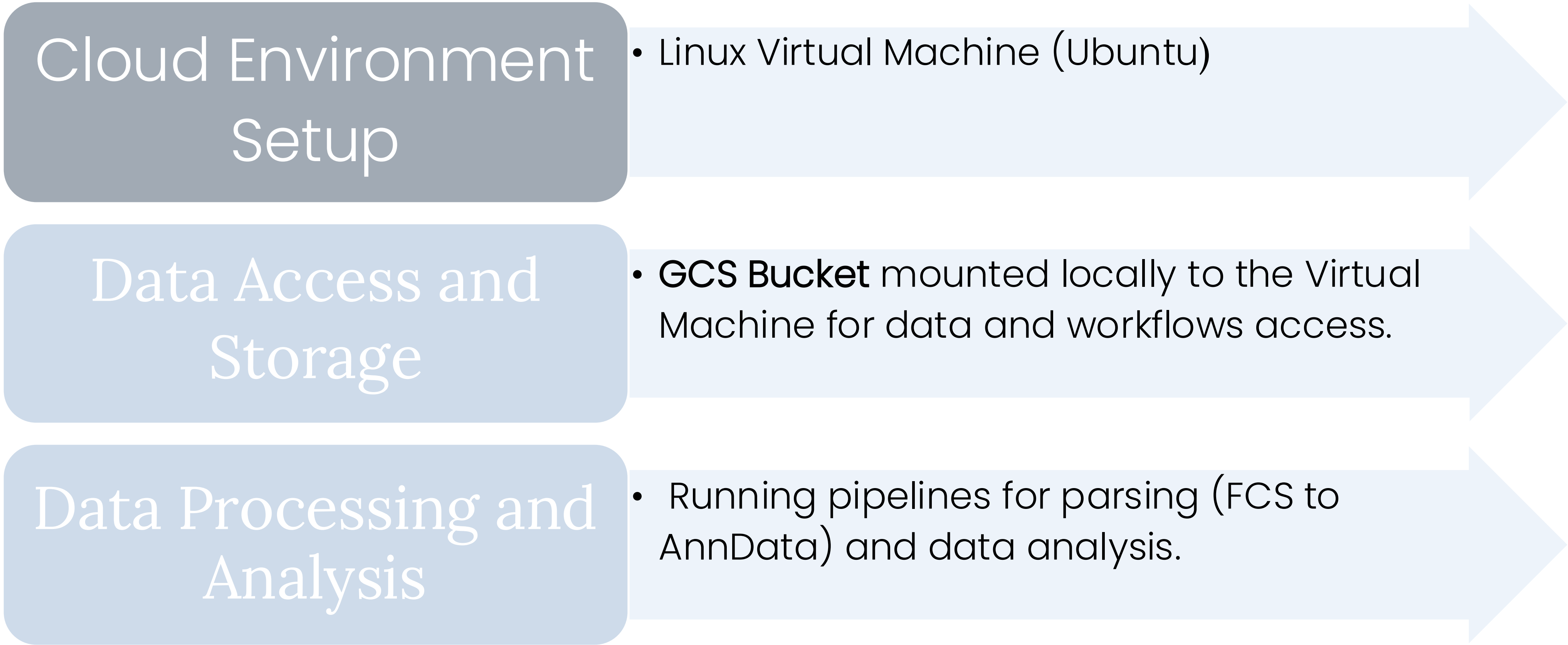
- What is the Cloud?
- Remote compute & storage instead of local machine

Why the Cloud

- Scales analysis beyond local computing limits.
- Enables fast processing of large single-cell datasets.
- Supports reproducible, shareable pipelines.



Cloud Workflow Architecture



Cloud Environment Setup

- Linux Virtual Machine (Ubuntu)

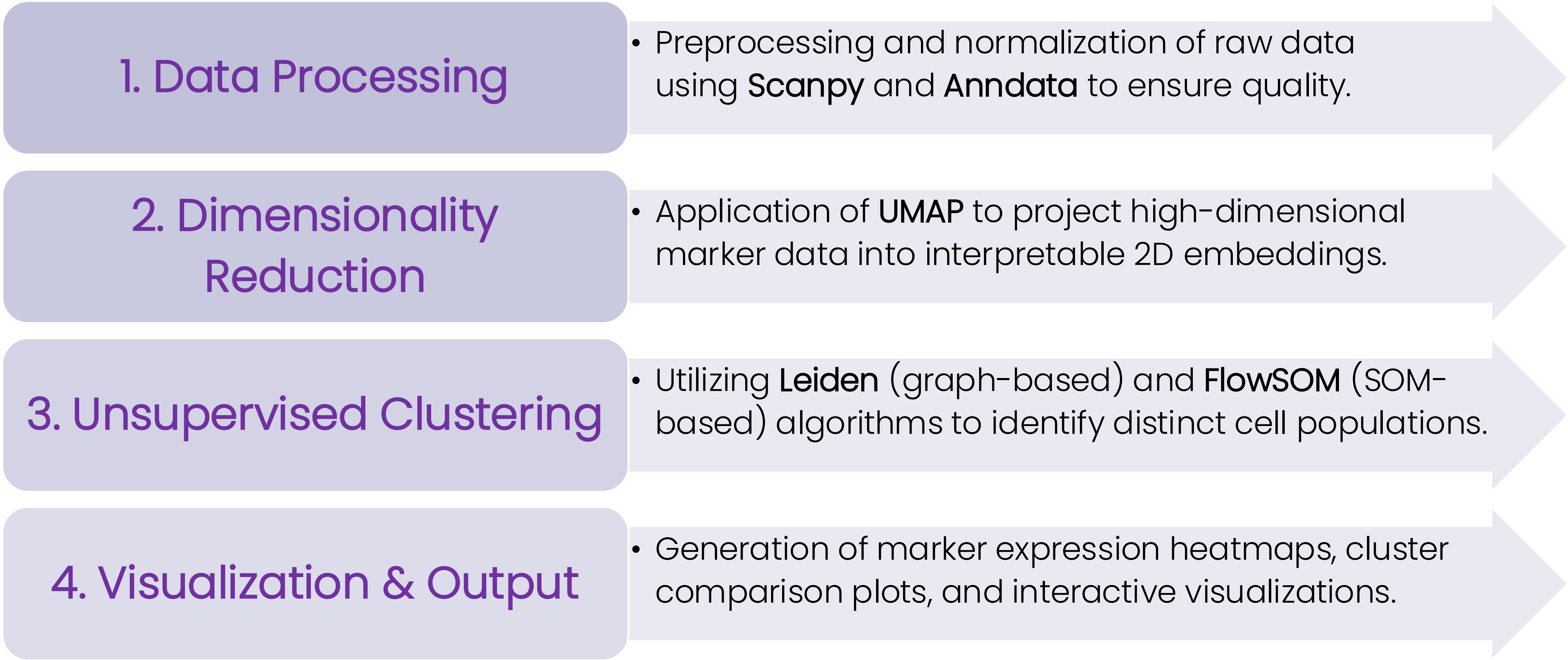
Data Access and Storage

- **GCS Bucket** mounted locally to the Virtual Machine for data and workflows access.

Data Processing and Analysis

- Running pipelines for parsing (FCS to AnnData) and data analysis.

Analytical Workflow



1. Data Processing

- Preprocessing and normalization of raw data using **Scanpy** and **Anndata** to ensure quality.

2. Dimensionality Reduction

- Application of **UMAP** to project high-dimensional marker data into interpretable 2D embeddings.

3. Unsupervised Clustering

- Utilizing **Leiden** (graph-based) and **FlowSOM** (SOM-based) algorithms to identify distinct cell populations.

4. Visualization & Output

- Generation of marker expression heatmaps, cluster comparison plots, and interactive visualizations.

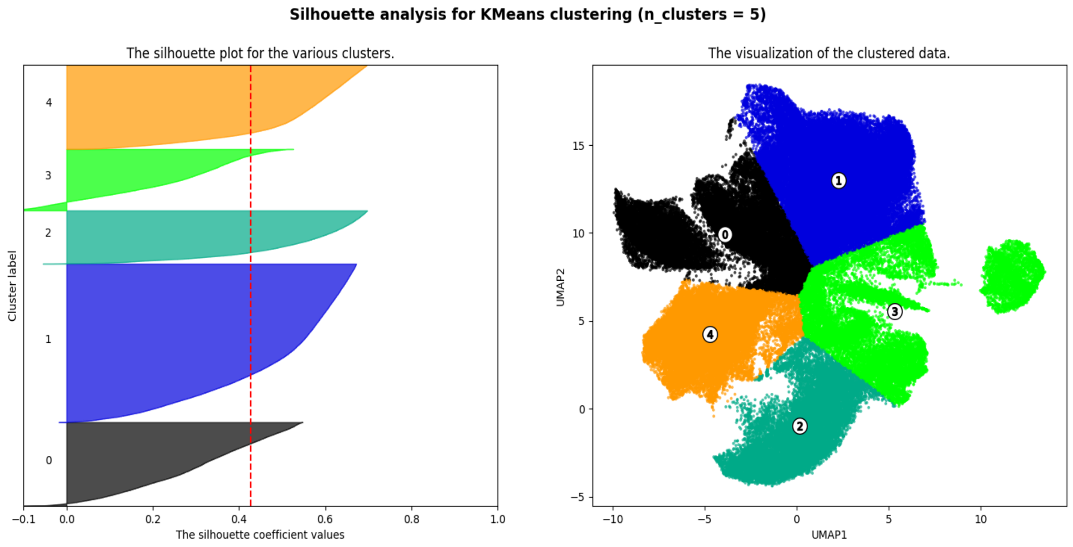
CLUSTER VALIDATION: SILHOUETTE ANALYSIS

Clustering: Cells were clustered using unsupervised methods (Leiden / FlowSOM), with K = 5 clusters selected for evaluation.

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

For every cell (data point) *i*, we calculate the Silhouette Score based on two distances:

- **Intra-cluster distance (*a*):** Distance to cells in its **own** cluster (should be small).
- **Inter-cluster distance (*b*):** Distance to cells in the **nearest neighbor** cluster (should be large).

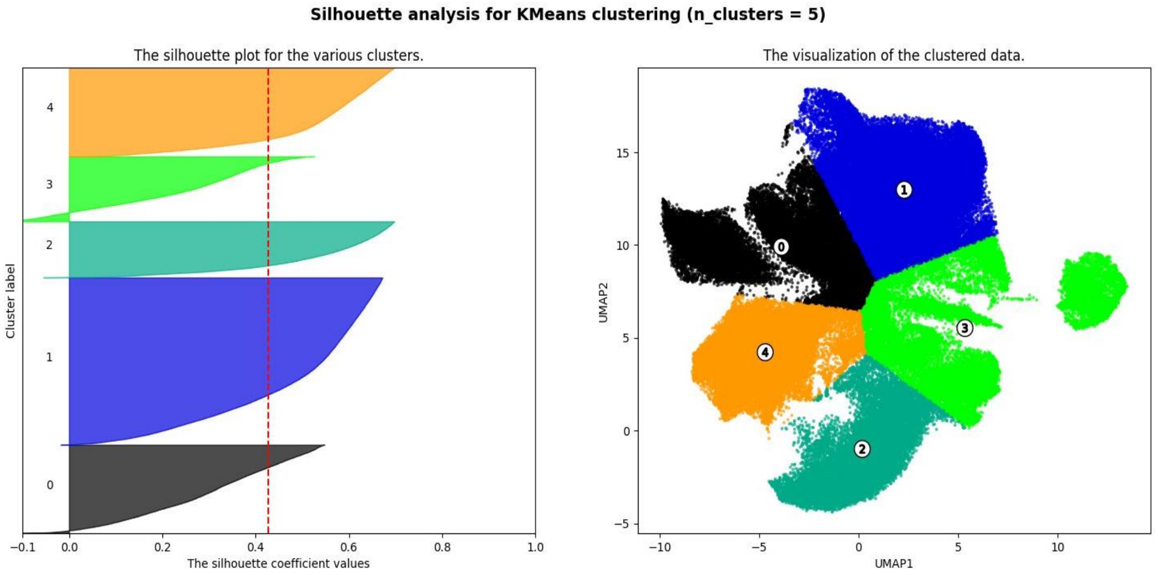


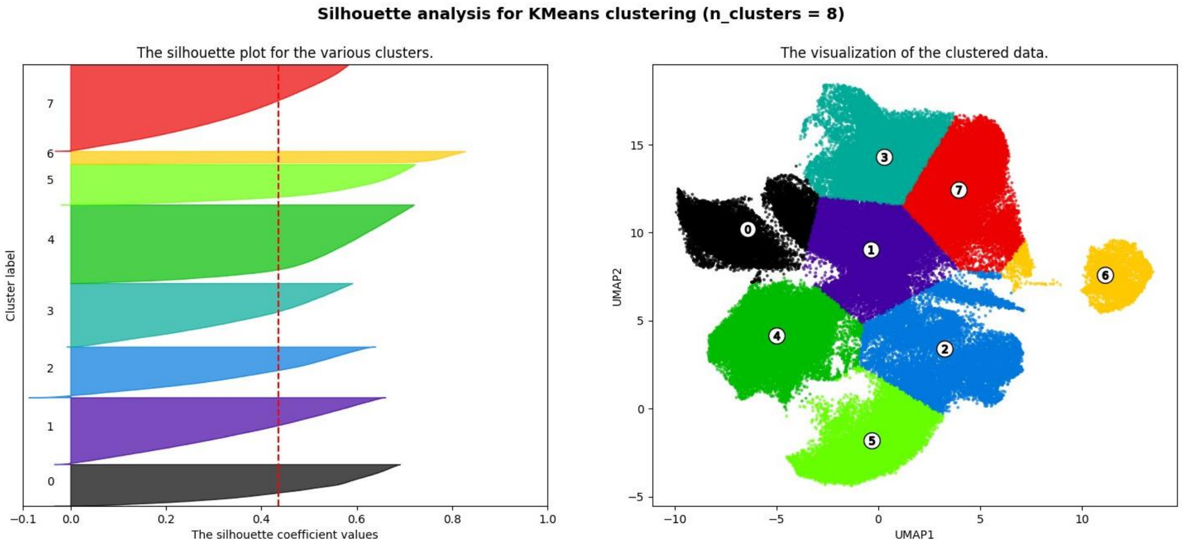
Interpreting the Silhouette Plot

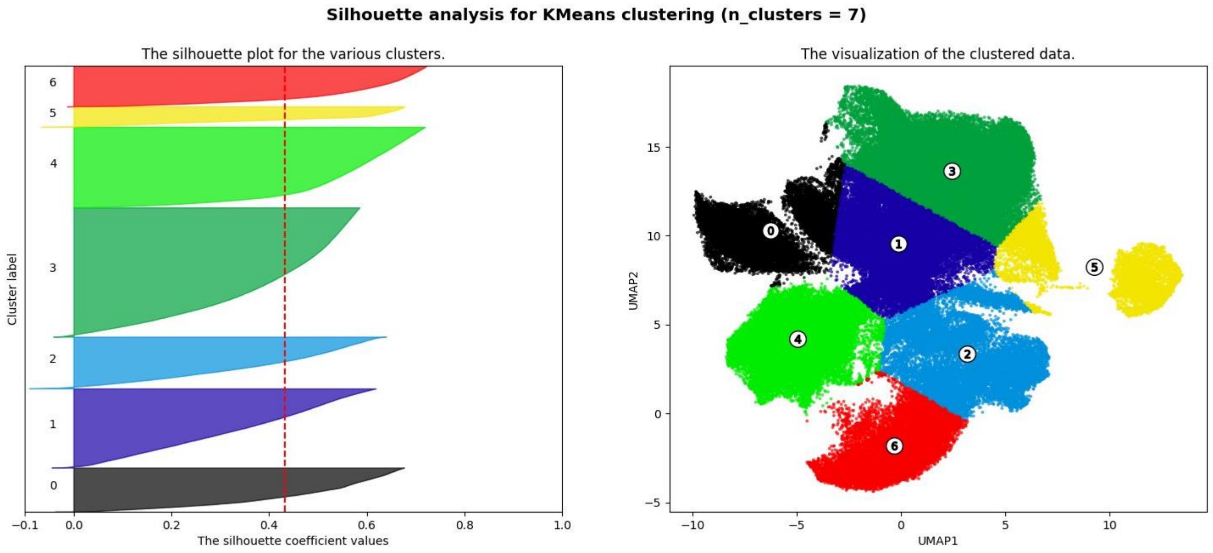
- **Each Line = One Cell:** The plot is made of horizontal bars; every single line represents an individual cell's score.
- **The "Knife" Shape:** You want thick, "knife-shaped" blocks. This indicates that most cells have high scores and are well-assigned.
- **The Threshold:** The vertical dashed line is the **average score**. You want most of your "knives" to extend past this line.

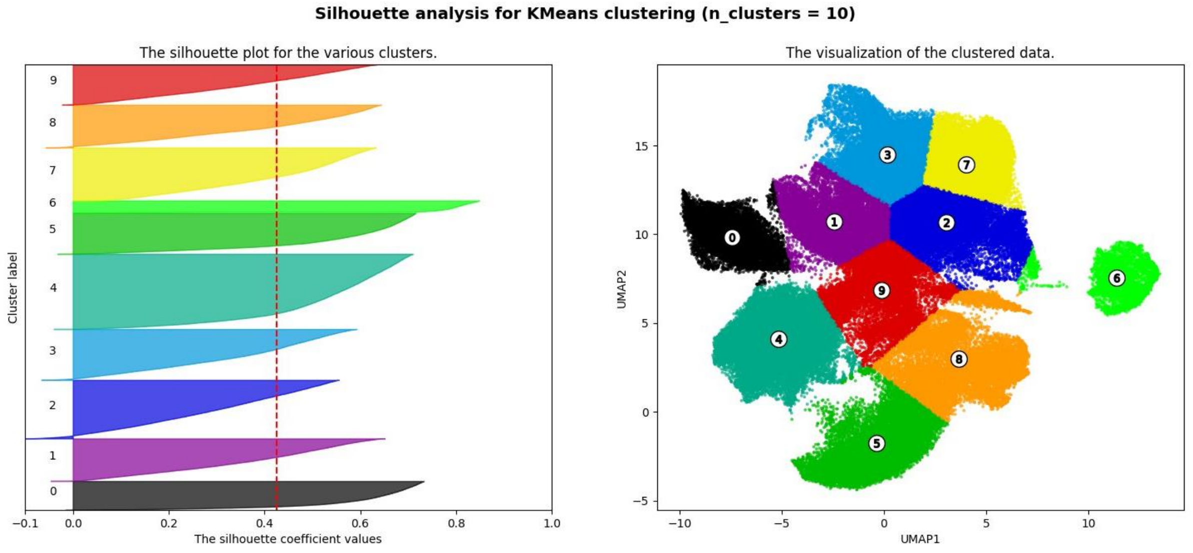
HOW TO DETERMINE THE OPTIMAL NUMBER OF CLUSTERS (K)?

SILHOUETTE PLOTS FOR K MEANS CLUSTERING ANALYSIS


Silhouette analysis for KMeans clustering (n_clusters = 5)


Silhouette analysis for KMeans clustering (n_clusters = 8)


Silhouette analysis for KMeans clustering (n_clusters = 7)


Silhouette analysis for KMeans clustering (n_clusters = 10)

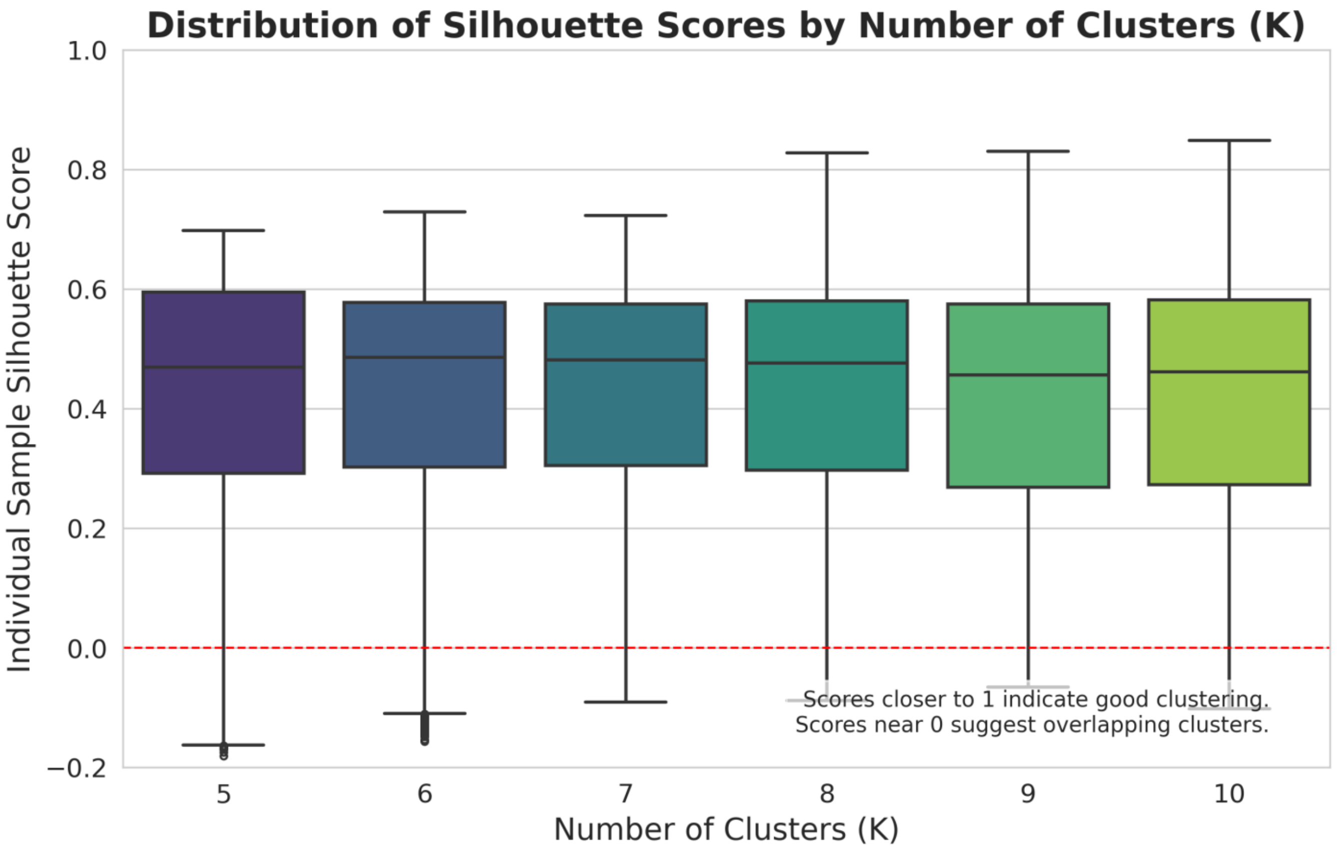
HOW TO DETERMINE THE OPTIMAL NUMBER OF CLUSTERS (K)?

What is a **Silhouette Scores** Boxplot?

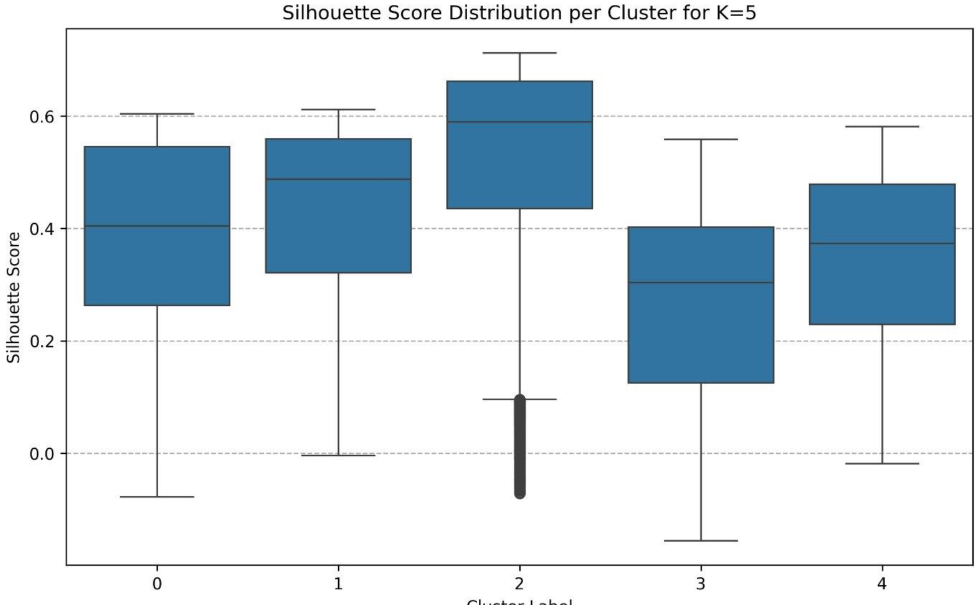
Represents the distribution of the silhouette scores for every data point across a range of cluster counts (K)

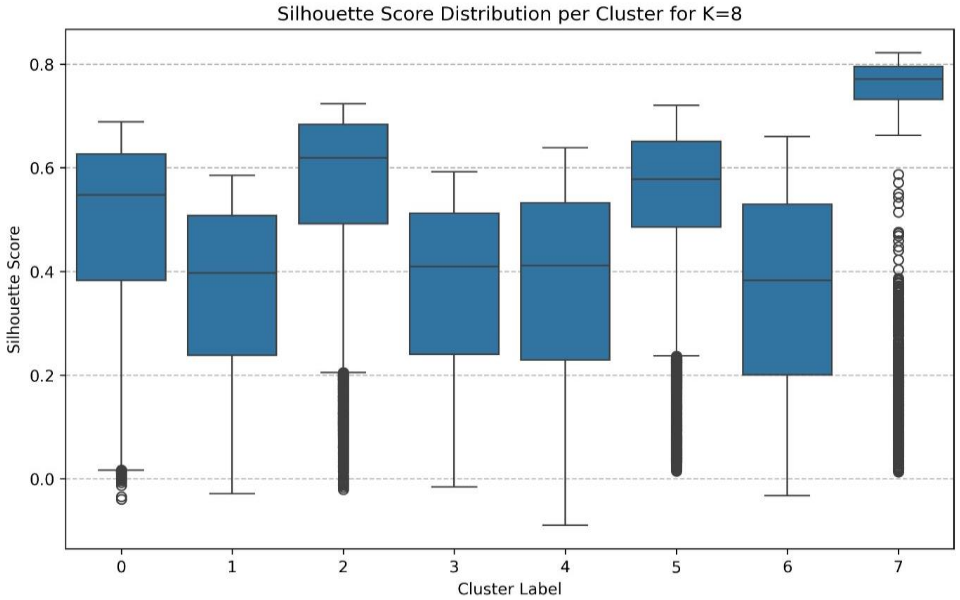
Optimal K:

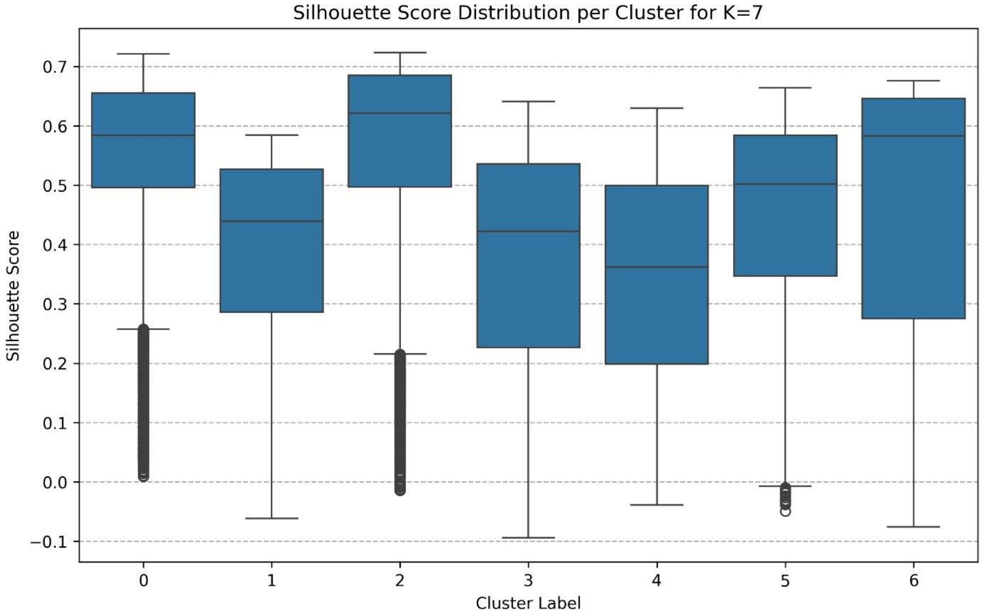
- **Median of Each Silhouette Score**: higher indicates better separation on average
- **Interpreting Cluster Cohesion** (Variability/IQR): less spread suggests more consistent cluster assignments across samples

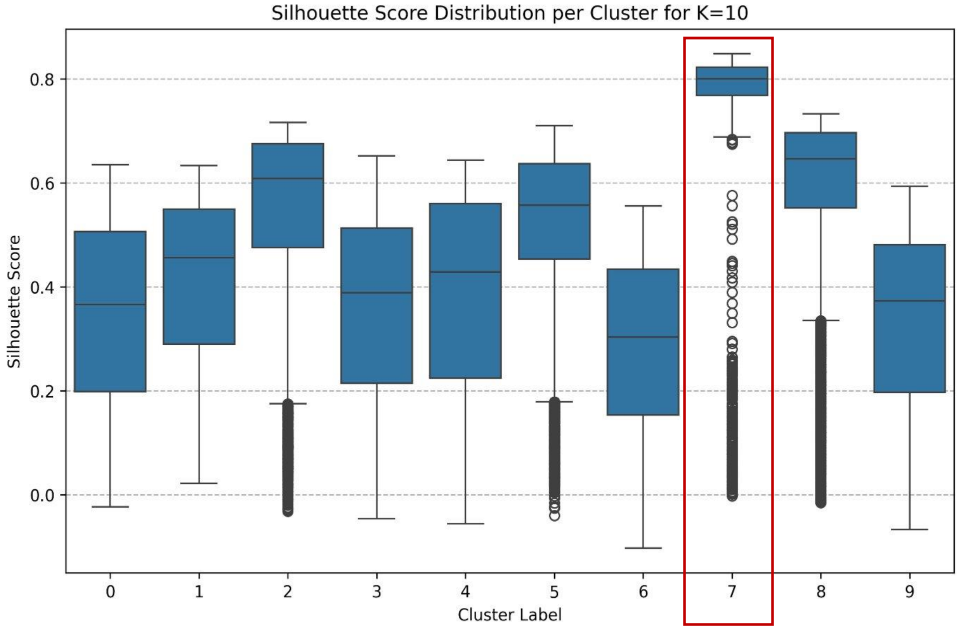


HOW GOOD IN AVERAGE ARE MY CLUSTERS?

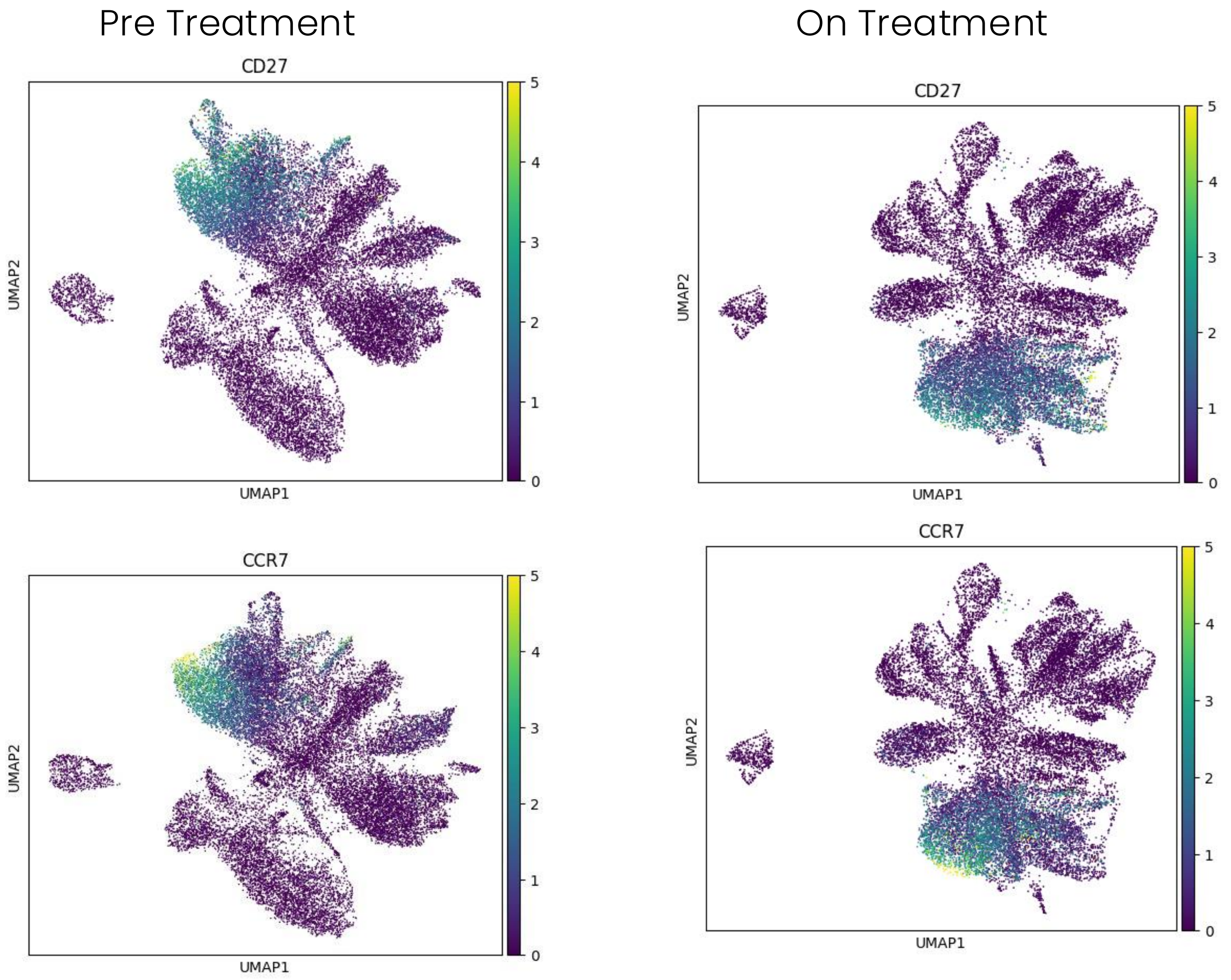

Silhouette Score Distribution per Cluster for K=5


Silhouette Score Distribution per Cluster for K=8


Silhouette Score Distribution per Cluster for K=7

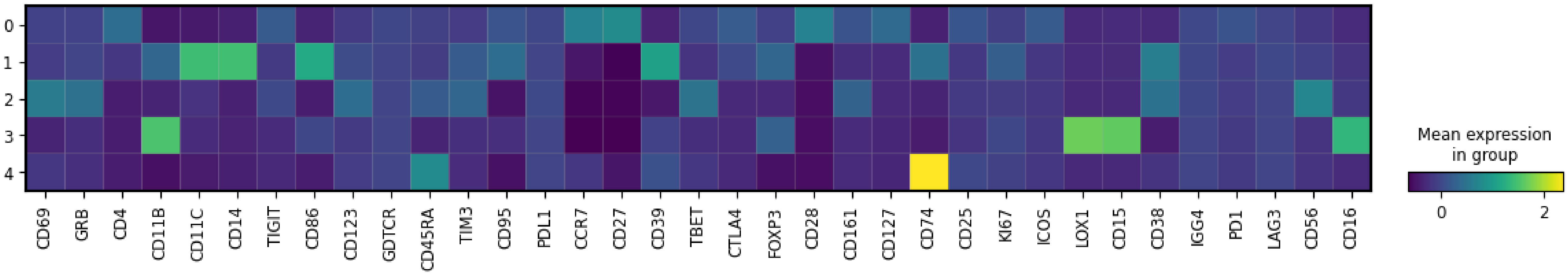

Silhouette Score Distribution per Cluster for K=10

Marker Expressions Across Treatment Stages



- **CD27:** co-stimulatory receptor on CD8+ T cells, essential for their activation, clonal expansion, survival, and long-term memory formation in anti-tumor and anti-viral immunity

- **CCR7**: Chemokine receptor marking naive and central-memory CD8^{+} T cells.



HOW CAN A COMPUTATIONAL APPROACH SUPPORT CELL MARKER CLUSTERING? HOW DATA ANALYSIS CAN SUPPORT BIOLOGY RESEARCH?

- Unbiased gating and clustering: can reveal new cell populations and relationships
- Validate biological gating
- Handling High Dimensionality and Automated Clustering

Tools and Technologies

Analysis & Parsing

- Pandas
- Anndata
- Scanpy
- Fcsparser
- pytometry

DevOps & Environment

- Linux (Ubuntu)
- Miniconda
- Pyenv
- Git/GitHub
- Bash Scripting

Web & Visualization

- Matplotlib
- Altair
- Flask
- SQLite3

Next Steps

Apply mixed linear models to appropriately account for longitudinal data.

Integrate Circulating Tumor DNA (ctDNA) and mass cytometry together in analysis

Perform clustering analyses on additional cell types

Explore determinants of immune checkpoint responses and iRAEs

Run analyses in parallel across samples, donors, or batches.

THANK YOU!

CHECK OUT MY CODE:
https://github.com/caterinaponti/radiohead-pici-internship.git

CONTACT ME

Caterina personal – cponti@dons.usfca.edu

LinkedIn – https://www.linkedin.com/in/caterina-ponti

SOURCES

- *Associations between Immune Checkpoint Inhibitor Response, Immune-Related Adverse Events, and Steroid Use in Radiohead: A Prospective Pan-Tumor Cohort Study | Journal for Immunotherapy of Cancer*, jitc.bmj.com/content/13/5/e011545.
- *The Cancer-Immunity Cycle: Indication, Genotype, and Immunotype: Immunity*, www.cell.com/immunity/fulltext/S1074-7613(23)00416-8.
- Dyikanov D;Zaitsev A;Vasileva T;Wang I;Sokolov AA;Bolshakov ES;Frank A;Turova P;Golubeva O;Gantseva A;Kamysheva A;Shpudeiko P;Krauz I;Abdou M;Chasse M;Conroy T;Merriam NR;Alesse JE;English N;Shpak B;Shchetsova A;Tikhonov E;Filatov I;Radko A;Bolshakova A;K. “Comprehensive Peripheral Blood Immunoprofiling Reveals Five Immunotypes with Immunotherapy Response Characteristics in Patients with Cancer.” *Cancer Cell*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/38744245/.
- Liang, Samantha I, et al. “Methylation-Based Ctdna Tumor Fraction Changes Predict Long-Term Clinical Benefit from Immune Checkpoint Inhibitors in Radiohead, a Real-World Pan-Cancer Study.” *Cancer Research Communications*, U.S. National Library of Medicine, 1 Aug. 2025, pmc.ncbi.nlm.nih.gov/articles/PMC12365632/%E2%80%8B.
- Mellman, Ira, et al. “The cancer-immunity cycle: Indication, genotype, and immunotype.” *Immunity*, vol. 56, no. 10, 2023, pp. 2188–2205, doi:10.1016/j.immuni.2023.09.011.
- Figure. Immune-related adverse events and impact on survival outcomes. Image from Zoe Quandt et al., Associations between immune checkpoint inhibitor response, immune-related adverse events, and steroid use in RADIOHEAD: a prospective pan-tumor cohort study, Journal for ImmunoTherapy of Cancer, vol. 13, no. 5, 12 May 2025, e011545, doi:10.1136/jitc-2025-011545. PubMed, https://pubmed.ncbi.nlm.nih.gov/40355283/
- Figure. Data center security layers, from One percent of Googlers get to visit a data center, but I did, Google Blog, 30 June 2020, blog.google/inside-google/infrastructure/how-data-center-security-works/
- Figure. Mellman, Ira, et al. “The cancer-immunity cycle: Indication, genotype, and immunotype.” *Immunity*, vol. 56, no. 10, 2023, pp. 2188–2205, doi:10.1016/j.immuni.2023.09.011. Figure of the cancer-immunity cycle.
- Figure: Immune checkpoint inhibitors illustrating PD-1/PD-L1 interaction and blockade (Credit: © Terese Winslow, National Cancer Institute, cancer.gov). Originally published on Cancer.gov in “Immune Checkpoint Inhibitors.” https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/checkpoint-inhibitors.