**Slide 1 Notes:**

Over the last two weeks, I’ve been working with the TOIL pan-cancer dataset to explore CES1 expression in colorectal cancer — specifically colon adenocarcinoma (COAD).  
I focused only on primary tumour samples and used clinical metadata from TCGA.  
The core aim was to perform survival analysis using gene expression — looking at overall survival, disease-specific survival, and progression-free interval.  
After verifying CES1 individually, I built multi-gene signatures by combining CES1 with other genes like CKS1B, BIRC5, and CDCA5.  
I also set up automated R pipelines to calculate z-score signatures, apply thresholding, and generate Kaplan–Meier plots.  
The diagram here shows the basic flow — starting from expression + clinical CSVs, scaling to signature scores, and then stratifying for survival.

**Slide 2 Notes:**

The goal was to investigate CES1 expression across COAD samples using survival endpoints like OS, DSS, and PFI.  
I also built multi-gene signature scores and automated the entire process in R for future scalability.  
I also did some background research on CES1 — it's involved in fatty acid metabolism and has been implicated in various cancers, which I found fascinating.

**Slide 3 Notes:**

In Week 1, I focused on verifying and cleaning up the CES1 survival analysis.  
I fixed some issues with stage-specific subsetting and made sure I was always using the correct filtered dataset — not the raw file.  
I then added patient counts to the legends, adjusted fonts and layouts for clarity, and checked that I was using the correct survival columns — like OS\_months, DSS\_months, and PFI\_months.  
I saved all results in an organized folder: KM\_CES1\_plots.  
The plot on the right is an example — this is KM\_OS\_0.5\_v1, showing CES1 stratified by median expression.

**Slide 4 Notes:**

This task expanded the CES1 survival analysis by looping across quantile-based thresholds.  
I tested quartiles, central exclusions, and extreme groupings to define HIGH vs LOW expression.  
All plots were generated automatically across OS, DSS, and PFI, skipping imbalanced groupings.  
Outputs were saved to the KM\_task2\_looped\_thresholds folder.  
The plots shown are representative of well-balanced comparisons after fixing group size filters.

**Slide 5 Notes:**

Next, I extended the analysis beyond CES1 alone.  
First, I normalised the expression of CES1, CKS1B, BIRC5, and CDCA5 using z-scores.  
Then I created multi-gene signature scores by combining 2–4 genes — for example, CES1 + CPT1A + MGLL.  
These composite scores were stratified into quantile-based groups to define LOW vs HIGH.  
Survival plots were generated for every signature–threshold combination and saved to the KM\_signature\_plots folder.  
As shown here, not all combinations produced strong effects, especially where group sizes were very uneven.

**Slide 6 Notes:**

In the final stage, I streamlined the workflow by turning repeated code into reusable R functions — including signature creation, stratification, and survival plotting.  
This made it much easier to test multiple genes and thresholds in one go.  
All output plots were saved with descriptive filenames and organized into folders.

**Slide 7 Notes:**

I ran into a few challenges:

* Matching sample IDs between datasets — especially where Ensembl version suffixes differed
* Accidentally using the wrong dataset (unfiltered instead of filtered) in some plots
* Plotting scripts were hard to scale manually  
  To solve this, I:
* Stripped suffixes, harmonized IDs
* Subsetted to primary tumours + colon adenocarcinoma only
* Automated all repetitive code  
  By the end of Week 3, the pipeline runs reliably and outputs are saved cleanly.