Swiss Institute of Bioinformatics

SINGLE-CELL TRANSCRIPTOMICS WITH R

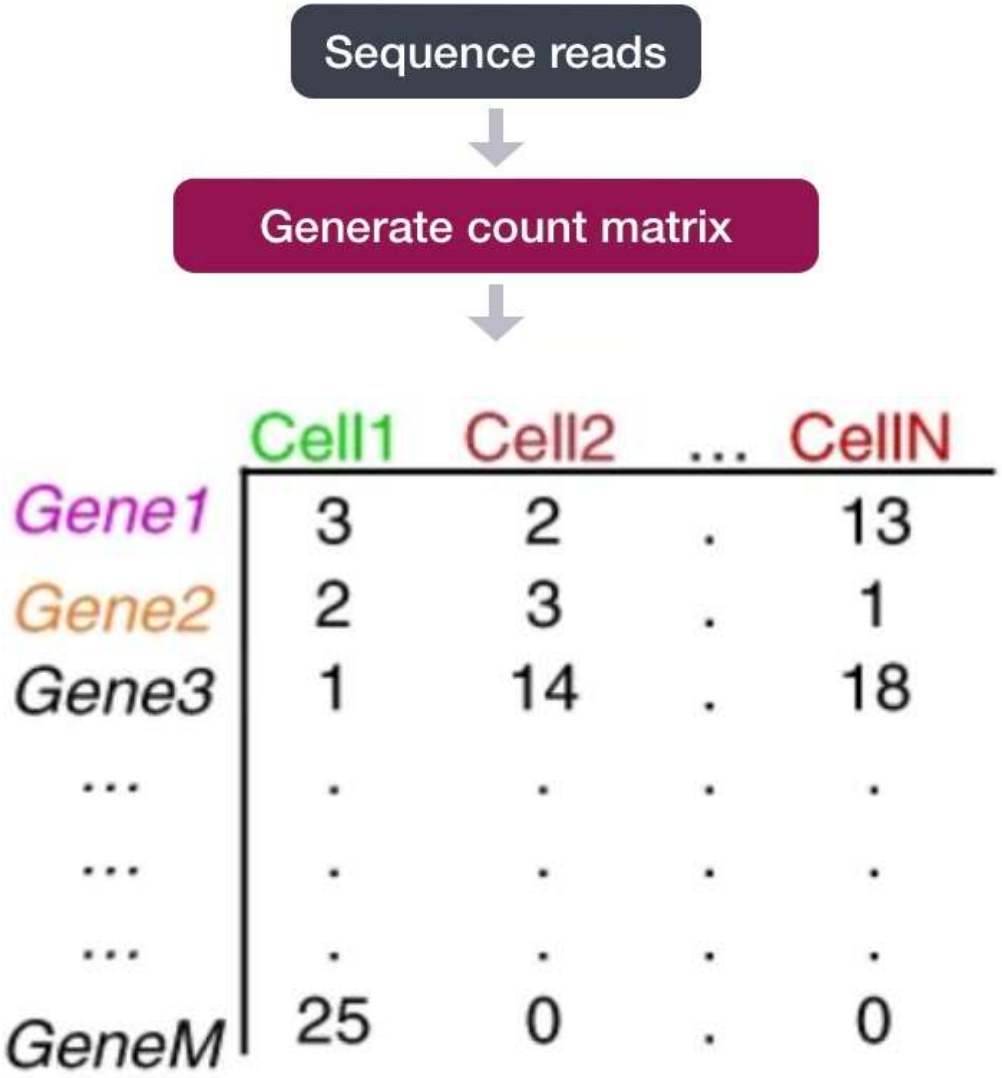
Analysis tools and QC

Deepak Tanwar

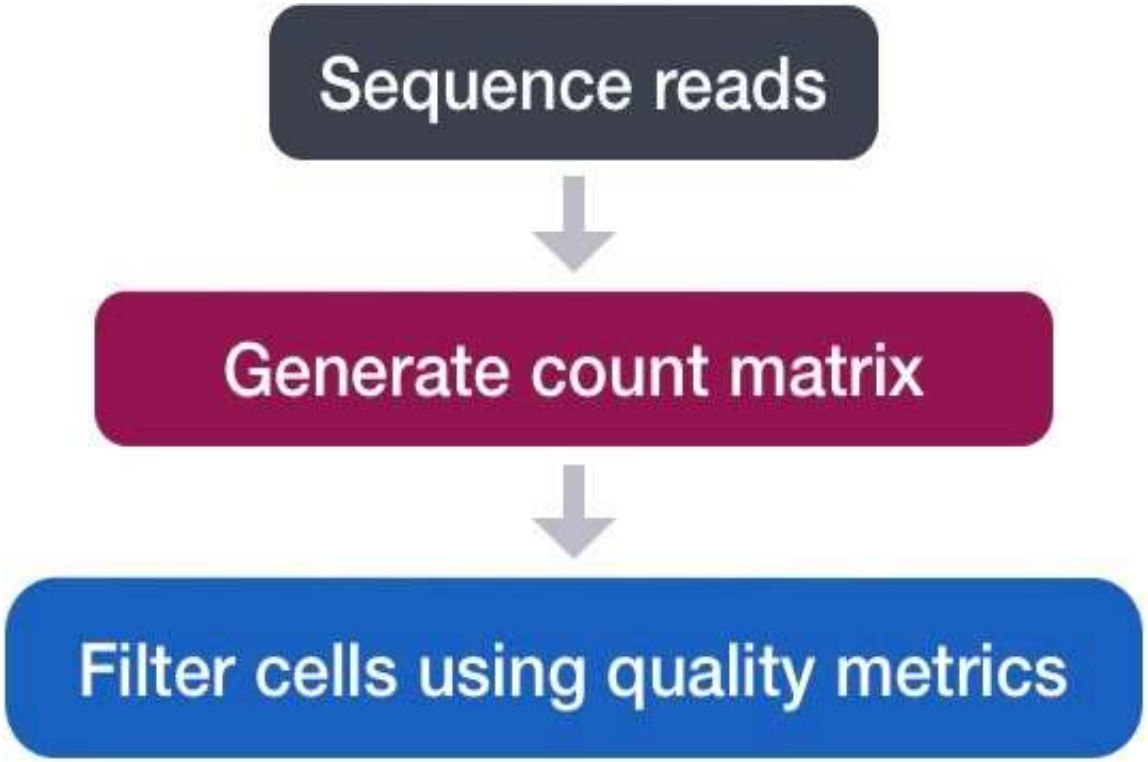
November 12-14, 2025

Adapted from previous year courses

From FASTQ to counts



Quality control of count matrix



Frequently used analysis tools

Major tools perform (at least) the following:

- QC
- normalization & scaling
- dimensionality reduction

https://nbisweden.github.io/workshop-scRNAseq/exercises

Frequently used analysis tools

Major tools perform (at least) the following:

- QC
- normalization & scaling
- dimensionality reduction

scanpy (python)

scater + scran (R, Bioconductor)

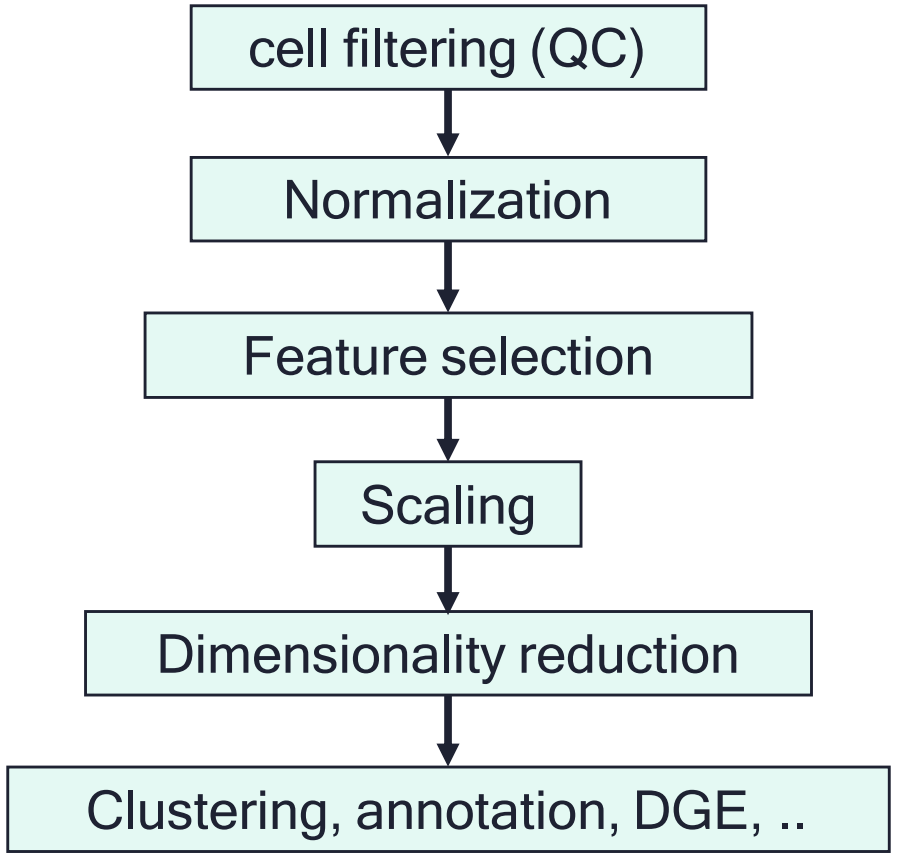
monocle3 (R, beta on github)

Seurat (R, CRAN)

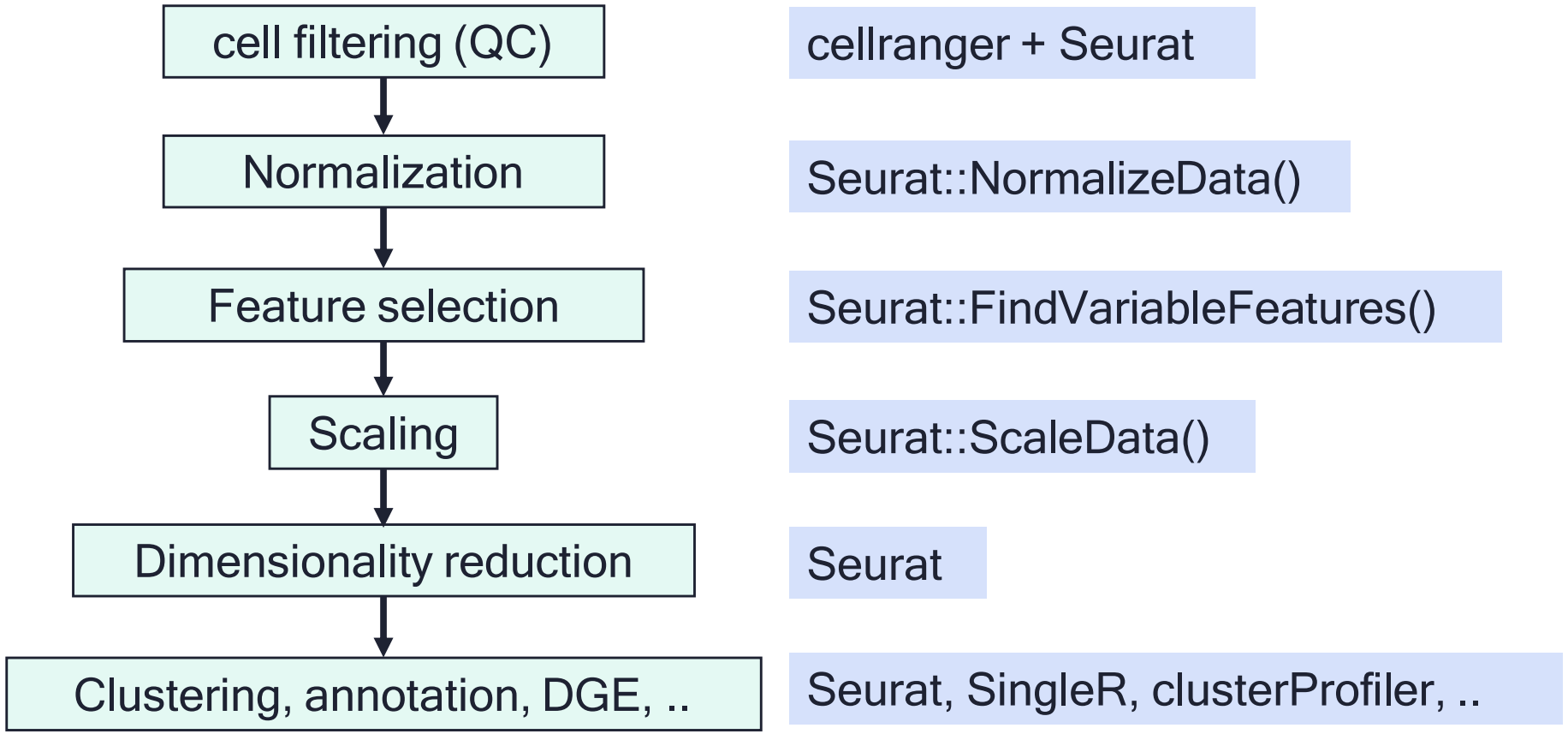
https://nbisweden.github.io/workshop-scRNAseq/exercises

Seurat

Data analysis pipeline overview



Data analysis pipeline overview



Quality control of count matrix

Goals:

- To **filter the data to only include true cells that are of high quality**, so that when we cluster our cells it is easier to identify distinct cell type populations
- To **identify any failed samples** and either try to salvage the data or remove from analysis, in addition to, trying to understand why the sample failed

Challenges:

Delineating cells that are poor quality from less complex cells

Choosing appropriate thresholds for filtering, so as to keep high quality cells without removing biologically relevant cell types

Cell filtering

Cellranger:

- cell calling (filter against low #UMI)

Manually (e.g. with Seurat):

- **#UMI**: high
- **#detected genes**
- **% mitochondrial UMI**
- **% ribosomal UMI**
- **% globin UMI**
- **Relationships between variables**

#UMI: high → possible doublet

High UMI counts per barcode often indicate **doublets/multiplets** (two or more cells in one droplet).

Why?

Each cell has a biologically plausible UMI range. Outliers above this suggest merged transcriptomes.

Common threshold: Top 1-5% of UMI distribution or > 2-3× median UMI count

#Detected genes

Low gene count: Empty droplets, damaged cells, or failed library prep.

High gene count: Doublets (two cell types → more genes detected).

Typical thresholds:

- Minimum: 200–500 genes/cell
- Maximum: Often tied to UMI count (e.g., cells with >6,000 genes may be doublets in 10x data).

% Mitochondrial UMI: dying cells

High mitochondrial content = cell stress, apoptosis, or membrane damage.

Cytoplasmic mRNA degrades faster than mitochondrial RNA → %MT increases.

Threshold: Usually <5–20%, depending on tissue (e.g., muscle has naturally high MT).

% Ribosomal UMI

High ribosomal RNA: Technical artifacts (incomplete poly-A selection).

Stress response or rapidly dividing cells.

% Globin UMI (blood samples)

True, especially in PBMC or whole blood data.

Globin mRNA (HBA, HBB) dominates in red blood cells/erythroid precursors.

Reduces sensitivity for immune cell genes.

Common practice: Remove cells with >10–20% globin.

Relationships between variables

Metrics are not independent:

- High UMI ↔ High gene count
- High %MT ↔ Low UMI/gene count (dying cells lose RNA)

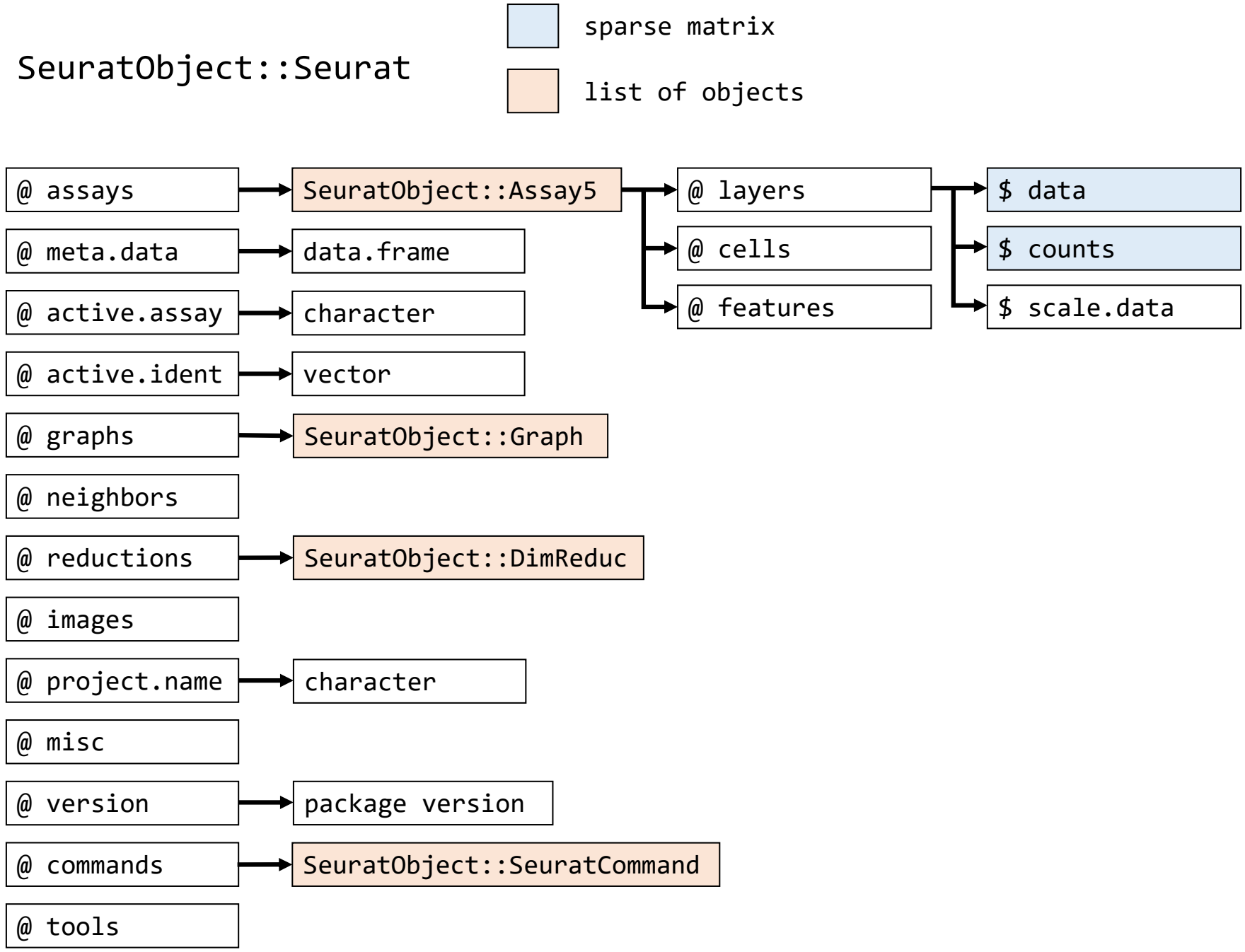
Best practice: Use scatter plots to set adaptive thresholds

Plot	What to Look For	What It Means
#UMI vs #Genes	Linear trend + outliers	Normal cells follow a line. Outliers = doublets
#Genes vs %MT	Low-gene, high-%MT cloud	Dying or empty droplets

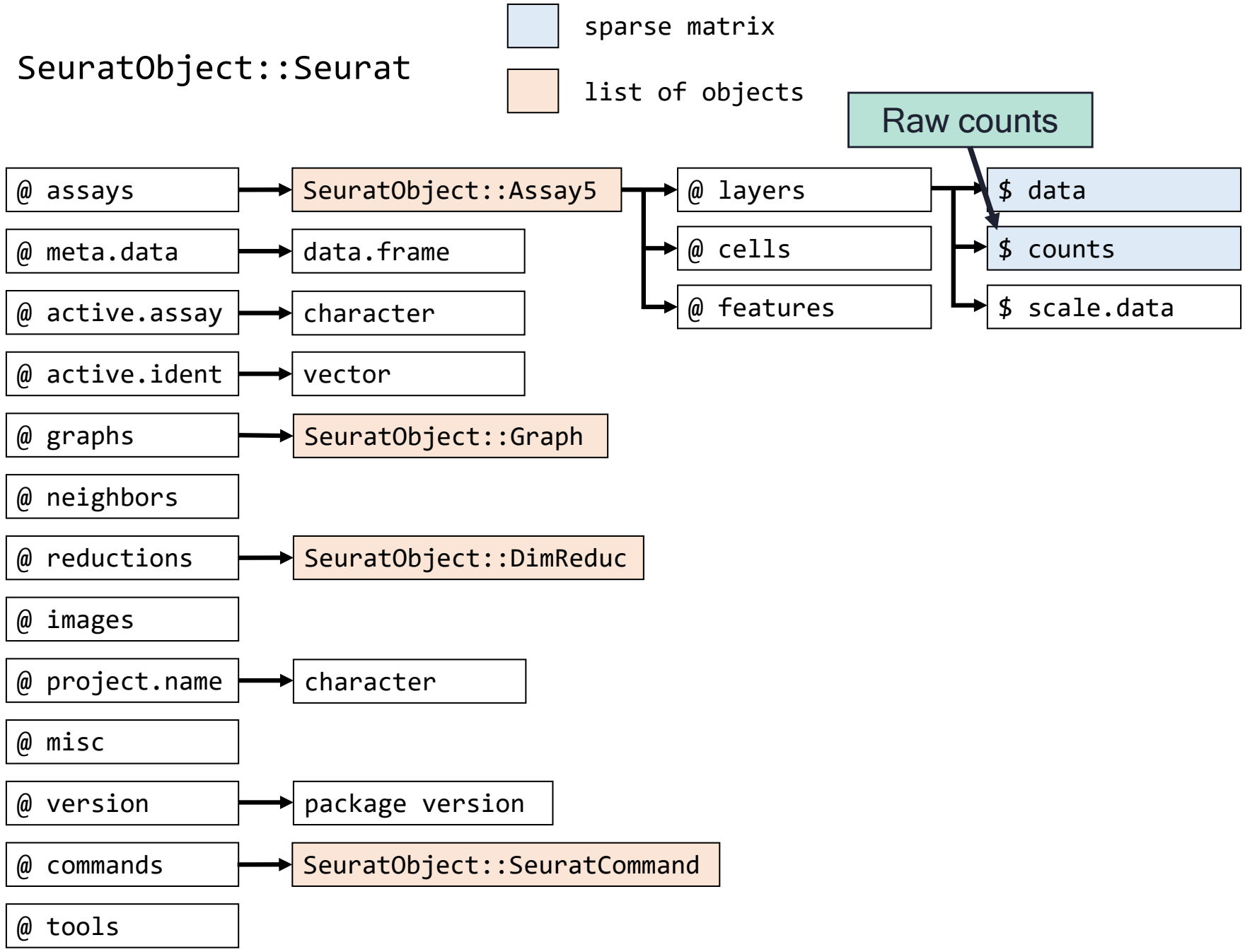
Summary of QC filtering

Metric	Low Value Indicates	High Value Indicates	Typical Action
nCount_RNA (UMI)	Empty droplet	Doublet	Filter extremes
nFeature_RNA (genes)	Low-quality cell	Doublet	Filter low & high
% Mitochondrial	—	Dying/stressed cell	Filter high (>10%)
% Ribosomal	—	Technical bias or stress	Flag or filter
% Globin	—	RBC contamination	Filter high (>10–20%)

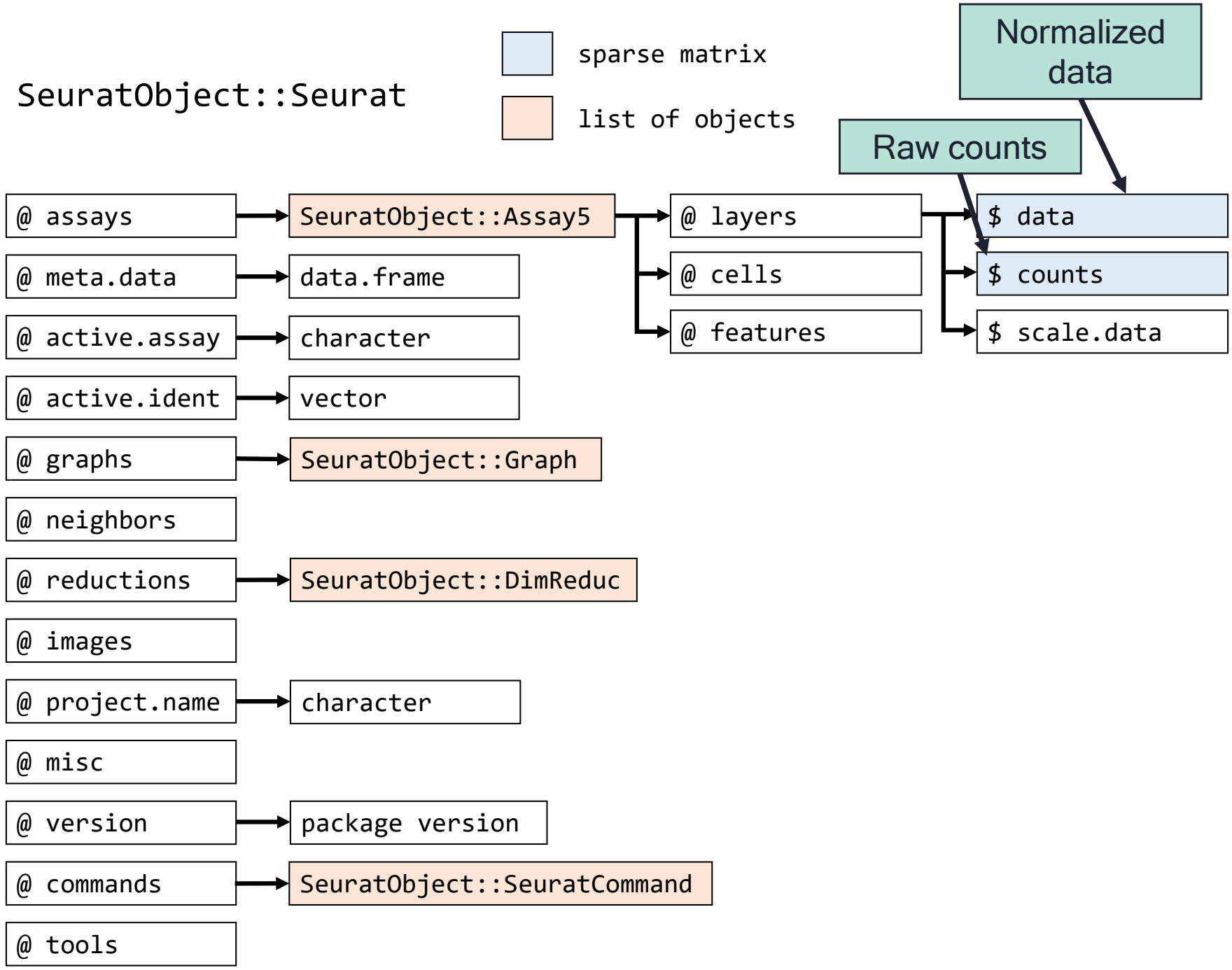
Data structure for the course



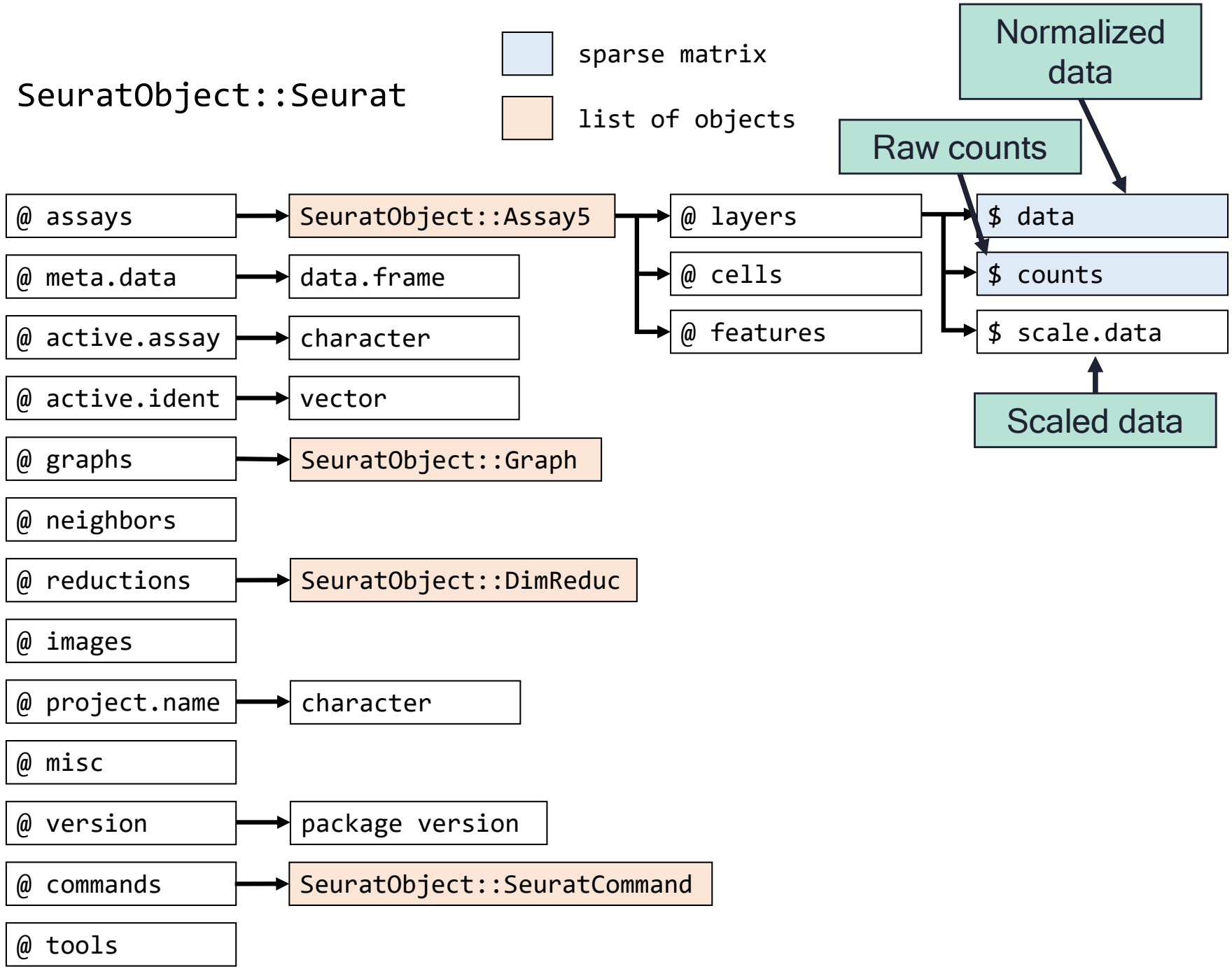
Data structure for the course



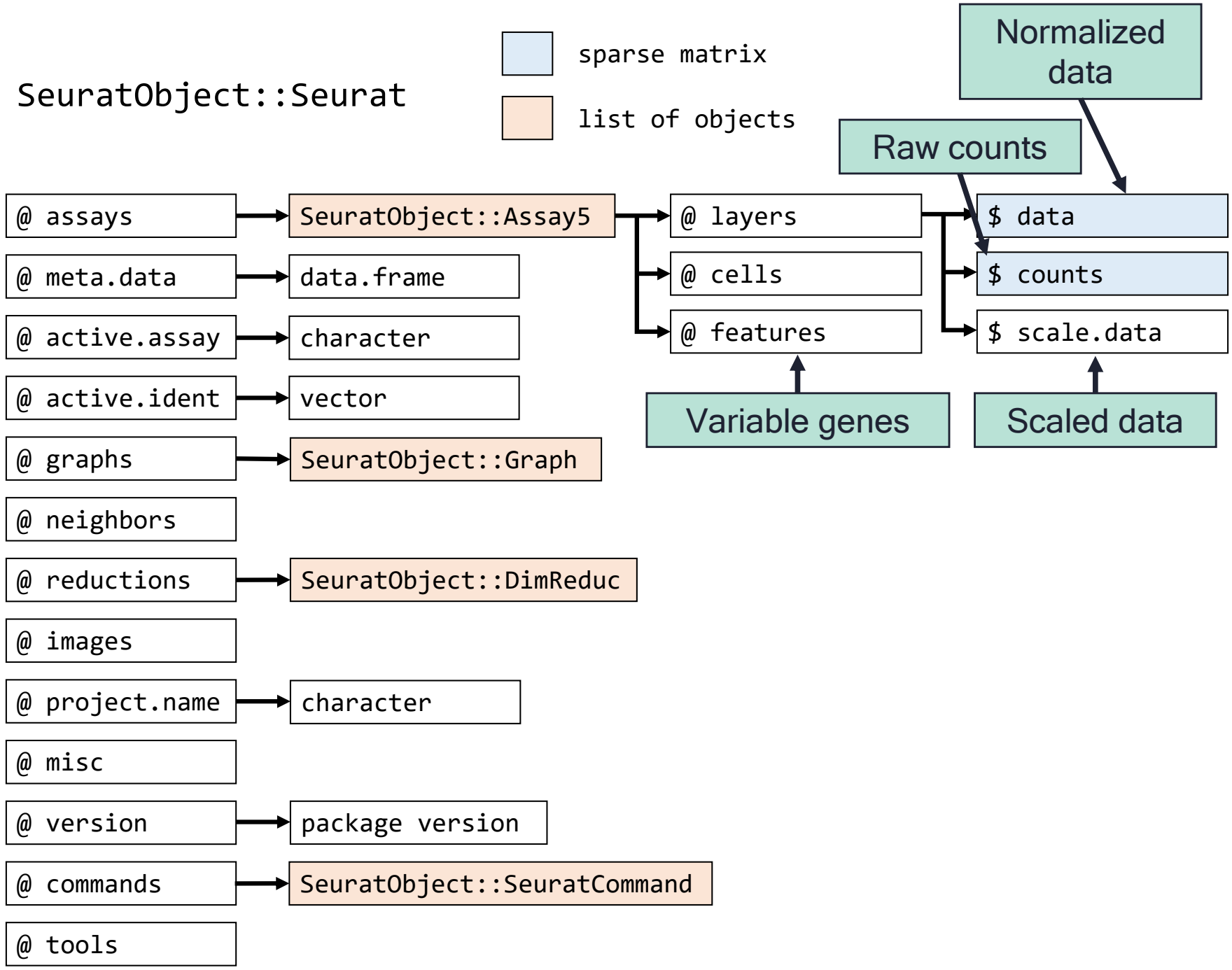
Data structure for the course



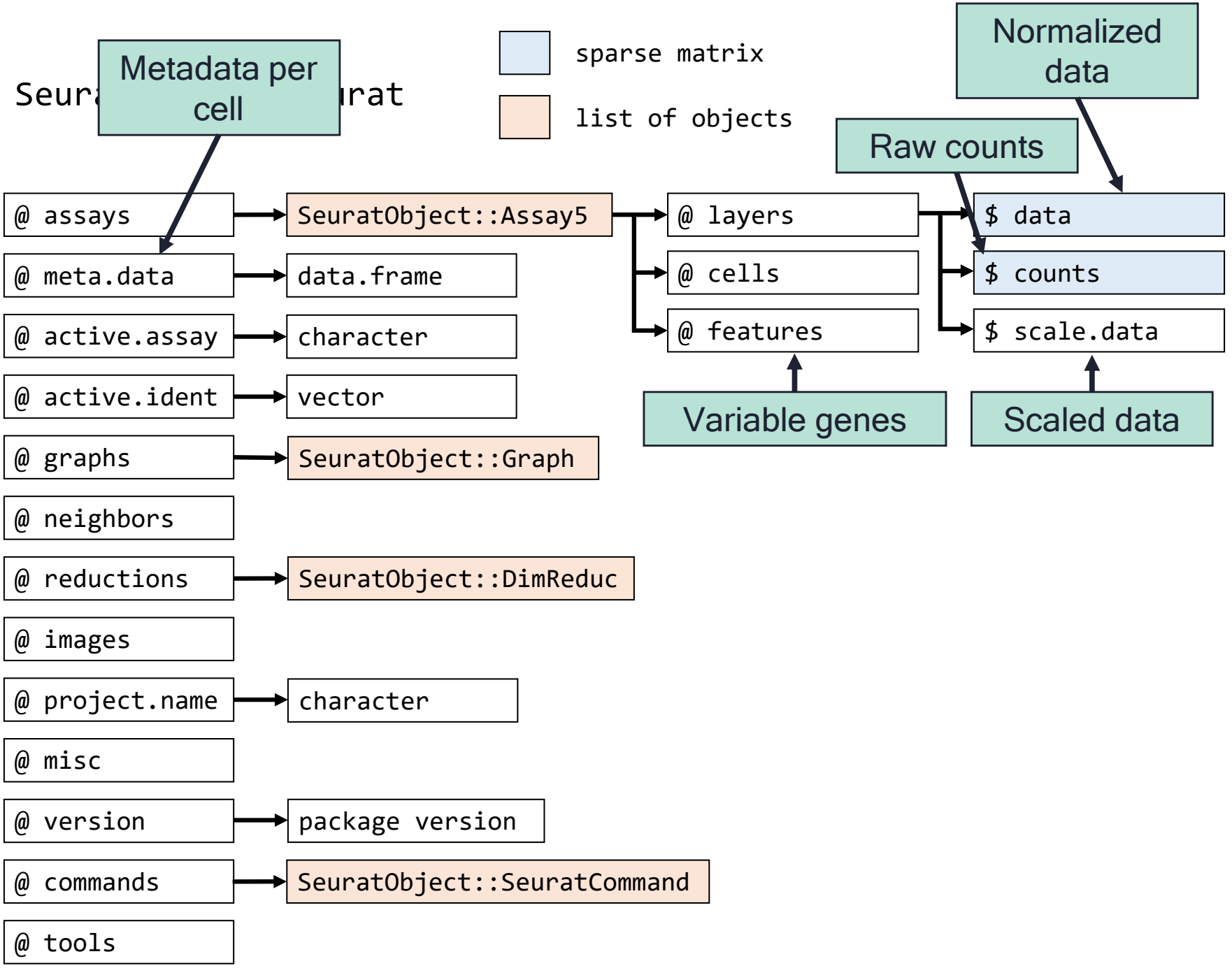
Data structure for the course



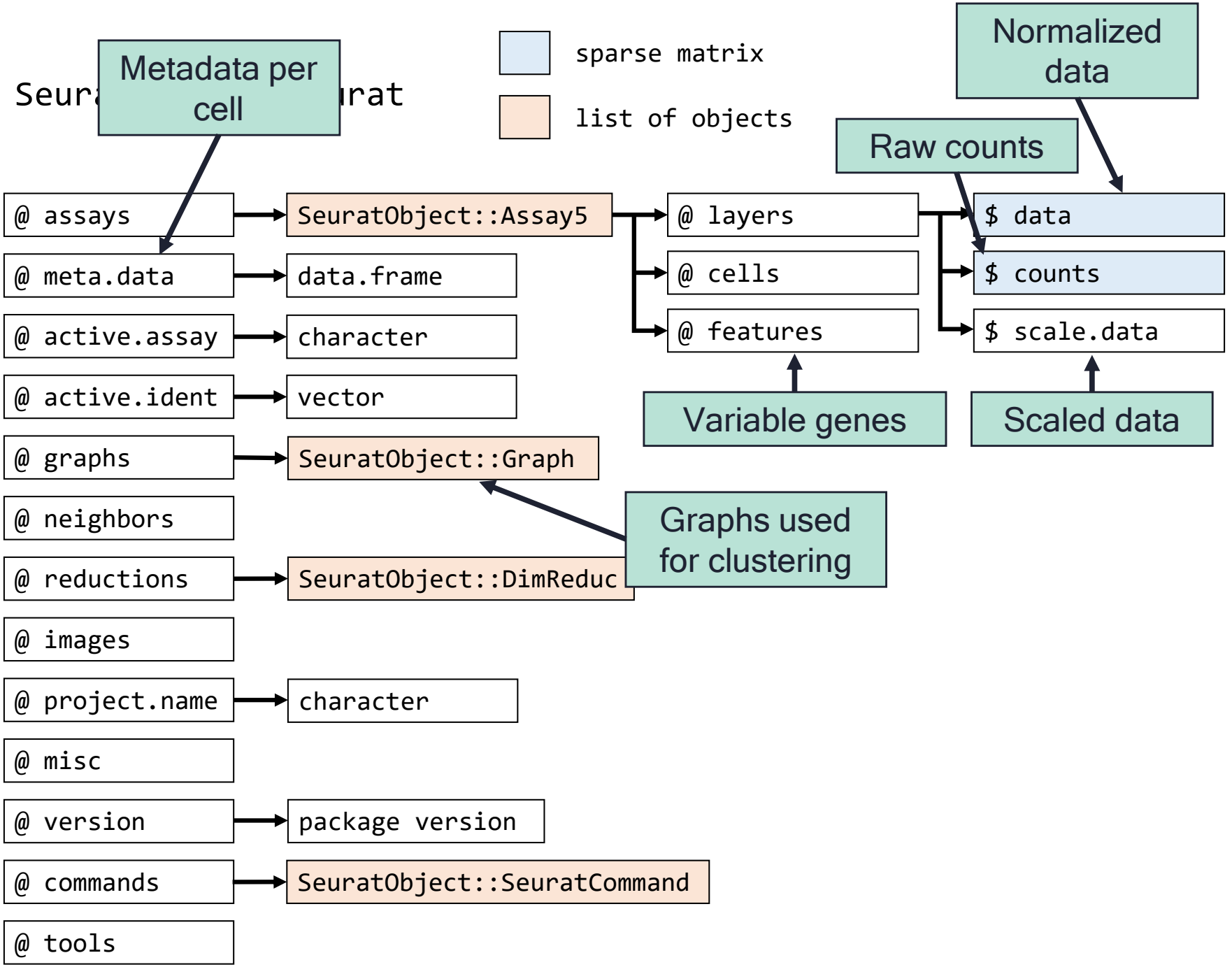
Data structure for the course



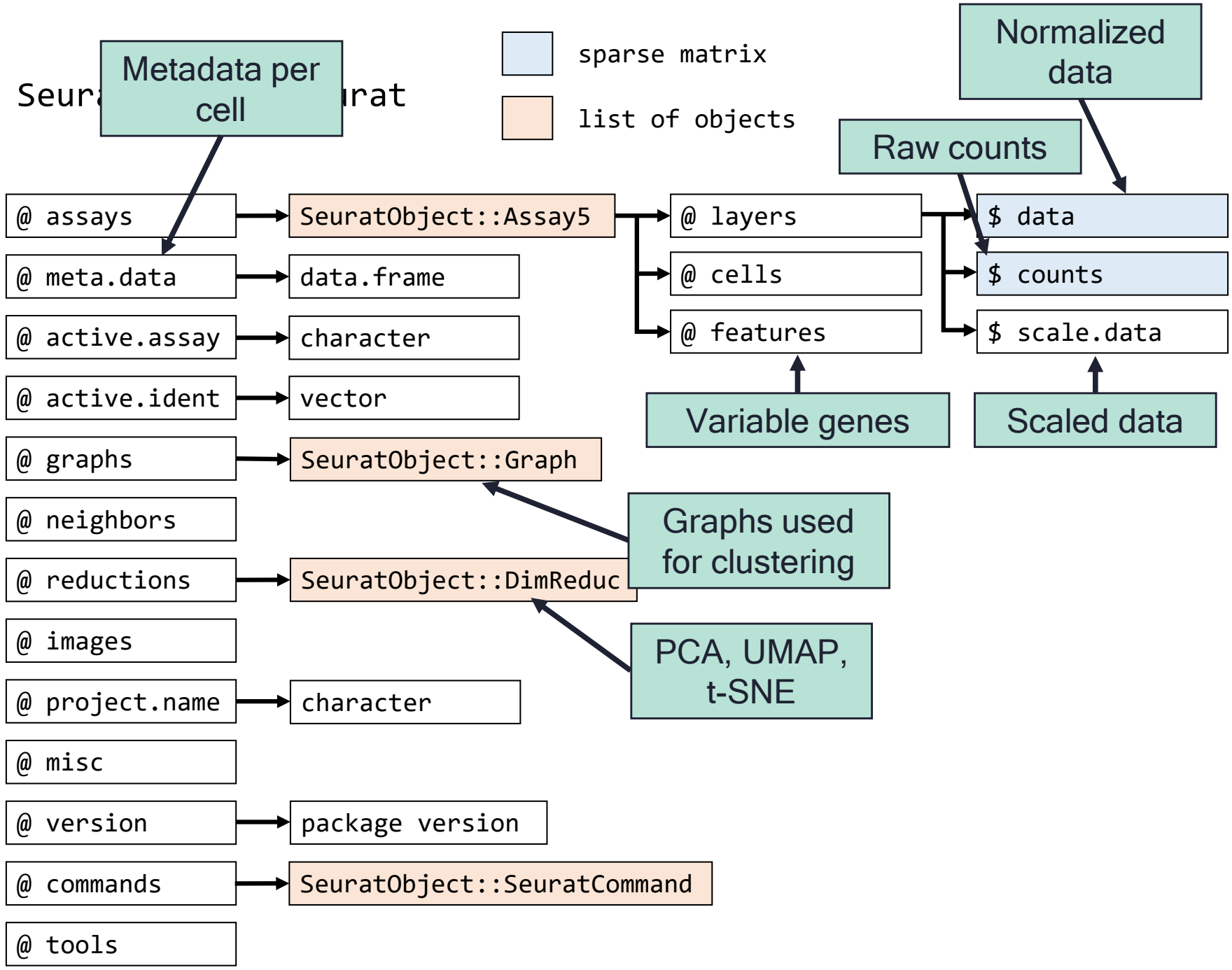
Data structure for the course



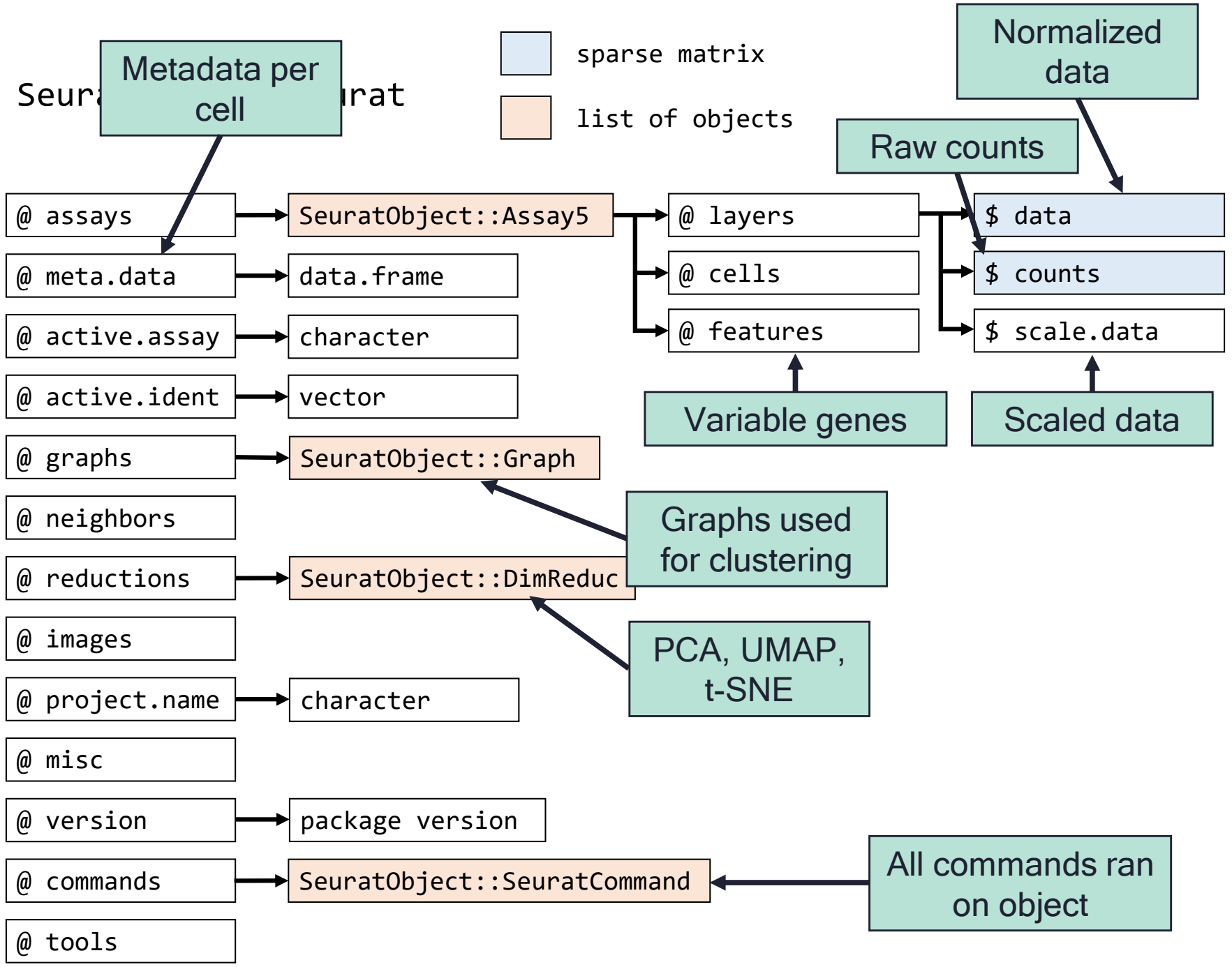
Data structure for the course

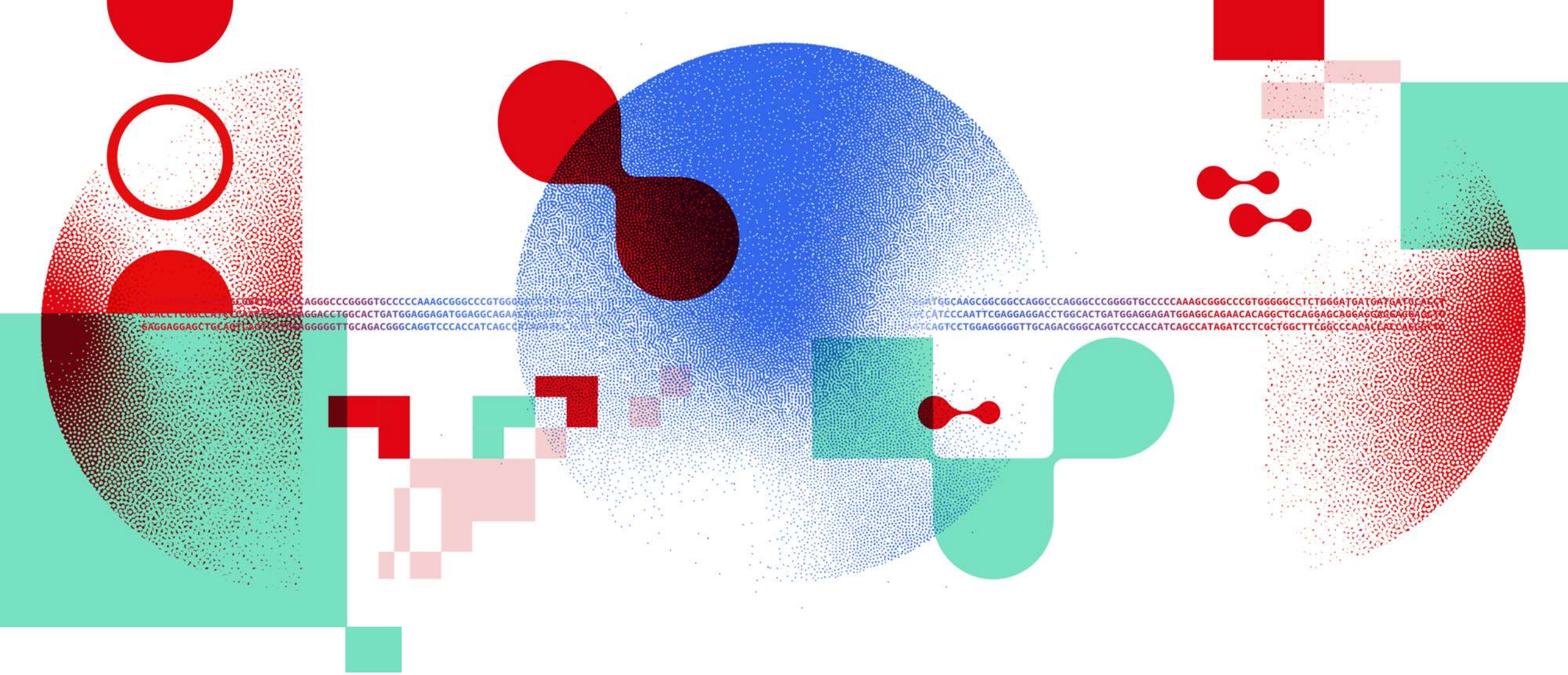


Data structure for the course



Data structure for the course





Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss