Hands-on Machine Learning Workshop

R-Ladies Philly

Workshop Goals

We hope workshop participants come away with the following:

- An understanding why it's useful to study transcriptomic data from tumors and how we can use unsupervised machine learning to reach our analysis goals
- An intuition for some of the challenges for identifying groups of tumors that have similar molecular processes occurring at the time of collection
- The beginnings of an intuition for different dimensionality reduction approaches uses for visualization
- An idea of why we may want to use domain-specific approaches to learn low-dimensional representations of our data

In general, this is meant to be a crash course or jumping off point that is sparse on the details underlying more complex methodologies.

Workshop Assumptions

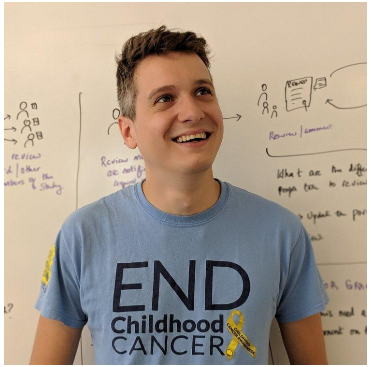
- Some familiarity with the following
  - R
  - RStudio
  - R Markdown
  - The Tidyverse
  - Navigating directory structure with file.path()
  - The concept of clustering

- RStudio Cloud Free account: https://rstudio.cloud/plans/free

The CCDL Team

Casey Greene, PhD (Director)

Deepashree V. Prasad (UX Designer)



Kurt Wheeler (Data Engineer)

David S. Mejia (Full Stack Engineer)

CHILDHOOD
CANCER
DATA LAB
Powered by: Alex's Lemonade Stand Foundation

Jaclyn Taroni, PhD (Principal Data Scientist)

Joshua Shapiro, PhD (Data Scientist)

Candace Savonen (Biological Data Analyst)

Chante Bethell (Biological Data Analyst)

Steven Foltz, PhD (Postdoc)

We create products for researchers with different expertise. We would love to hear from you about your experiences with data and collaboration and get some of your input!

Deepashree V. Prasad
(UX Designer)

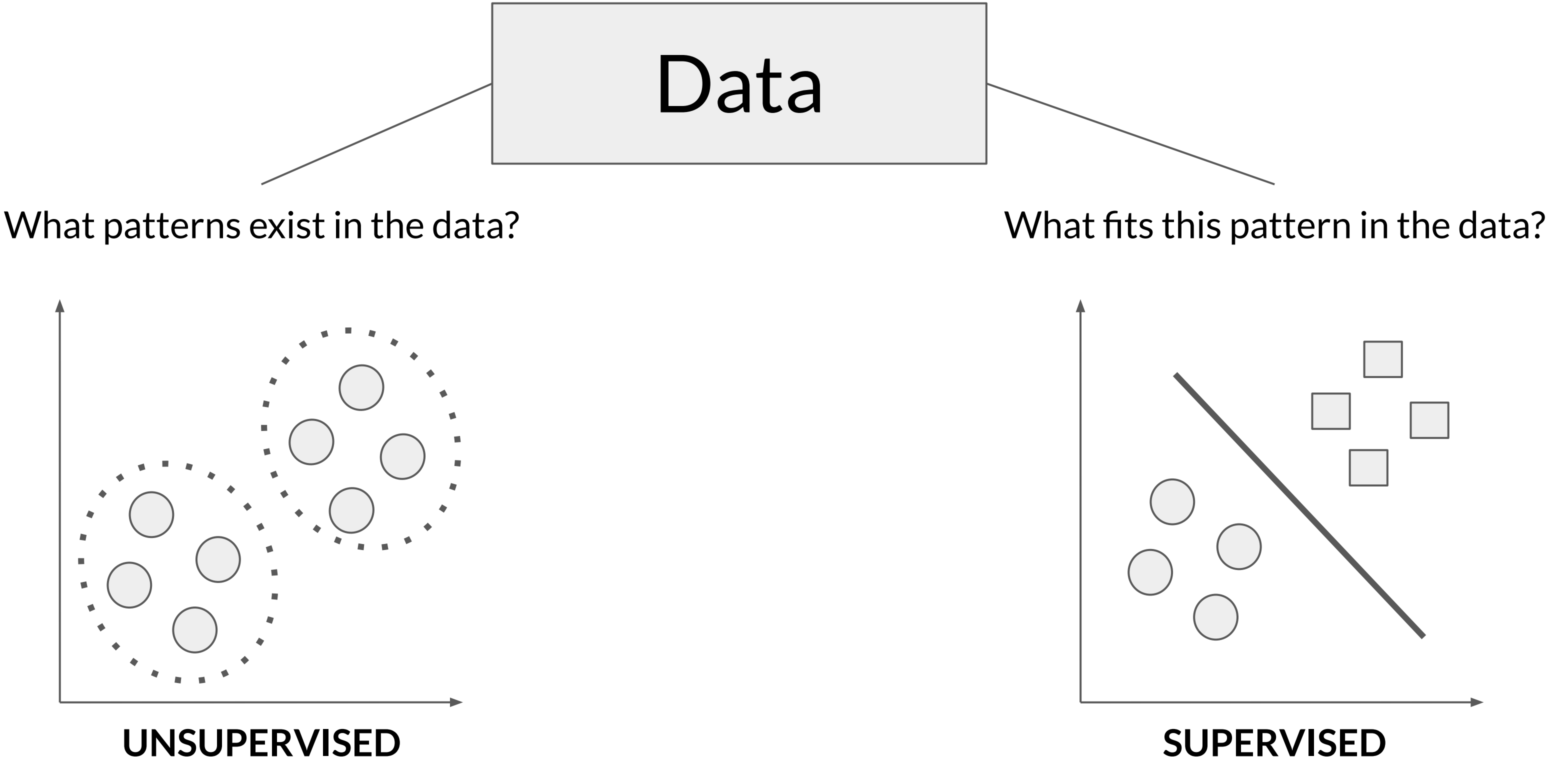
Deepa would be delighted to speak with you about this. If you are interested please fill out this form https://bit.ly/3gRNKxI or reach out to her on twitter @deepa_vprasad!

Now for a bit of background...

Machine Learning, what is it?

Having a computer program learn to perform a task (like predicting an outcome) from data, rather than programming explicit instructions

Classes of machine learning algorithms



Data

What patterns exist in the data?

What fits this pattern in the data?

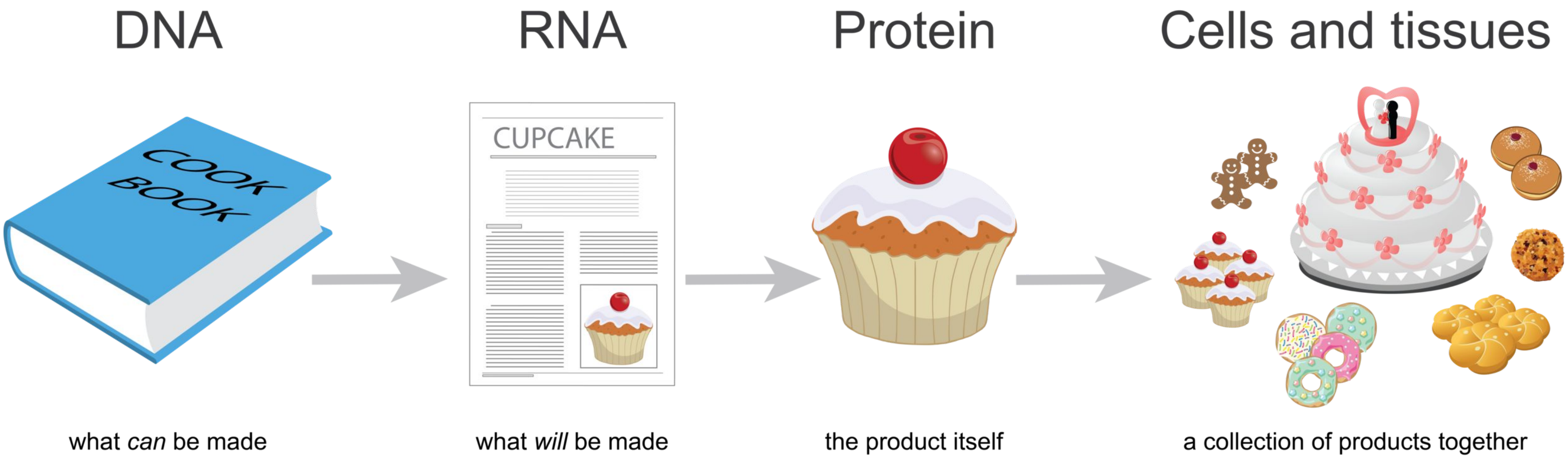
UNSUPERVISED

SUPERVISED

We’ll focus on **unsupervised** learning

- Which samples are most similar to one another?
- How many groups of samples exist in my data?
- What patterns of gene expression exist in my data? How do the genes vary together?

Time for an analogy that involves cupcakes…



DNA

what *can* be made

RNA

what *will* be made

Protein

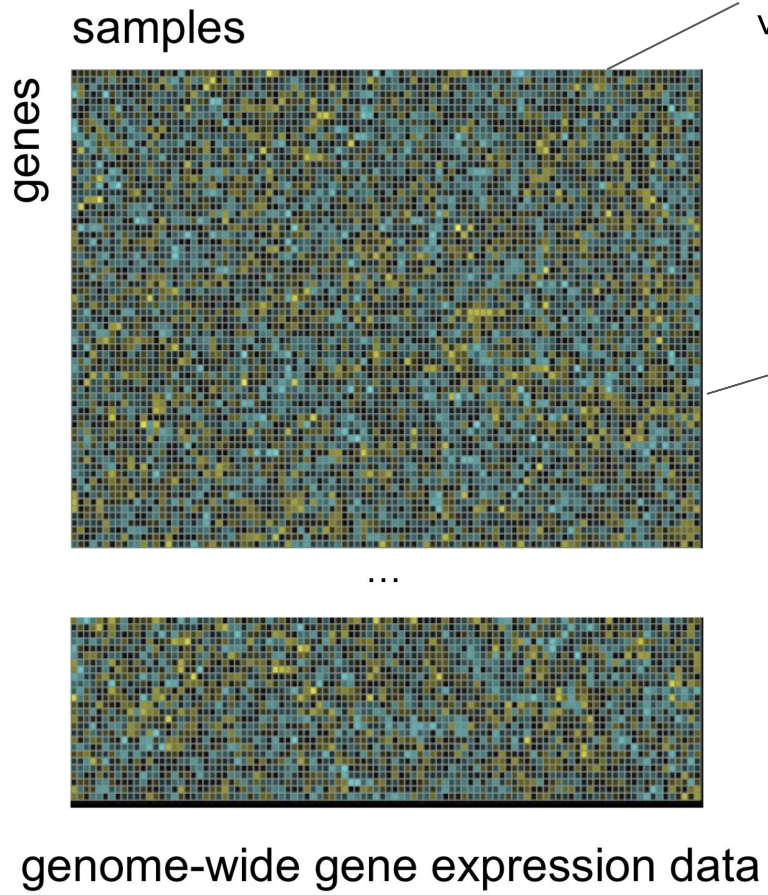
the product itself

Cells and tissues

a collection of products together

The Central Dogma of Molecular Biology

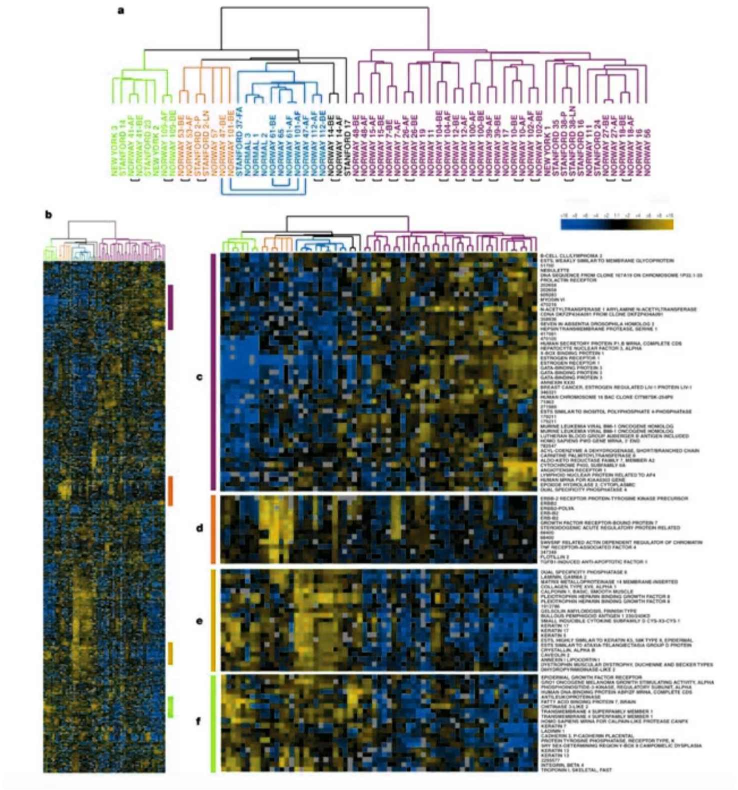
When we measure genome-wide gene expression data, it’s like taking a peek at all the baked goods being made in the bakery by measuring what recipes were around at the time.



A **column** is the collection of expression values for a given sample or observation.

A **row** is measure of the expression of the RNA (recipe) for all samples in our study. These are our features.

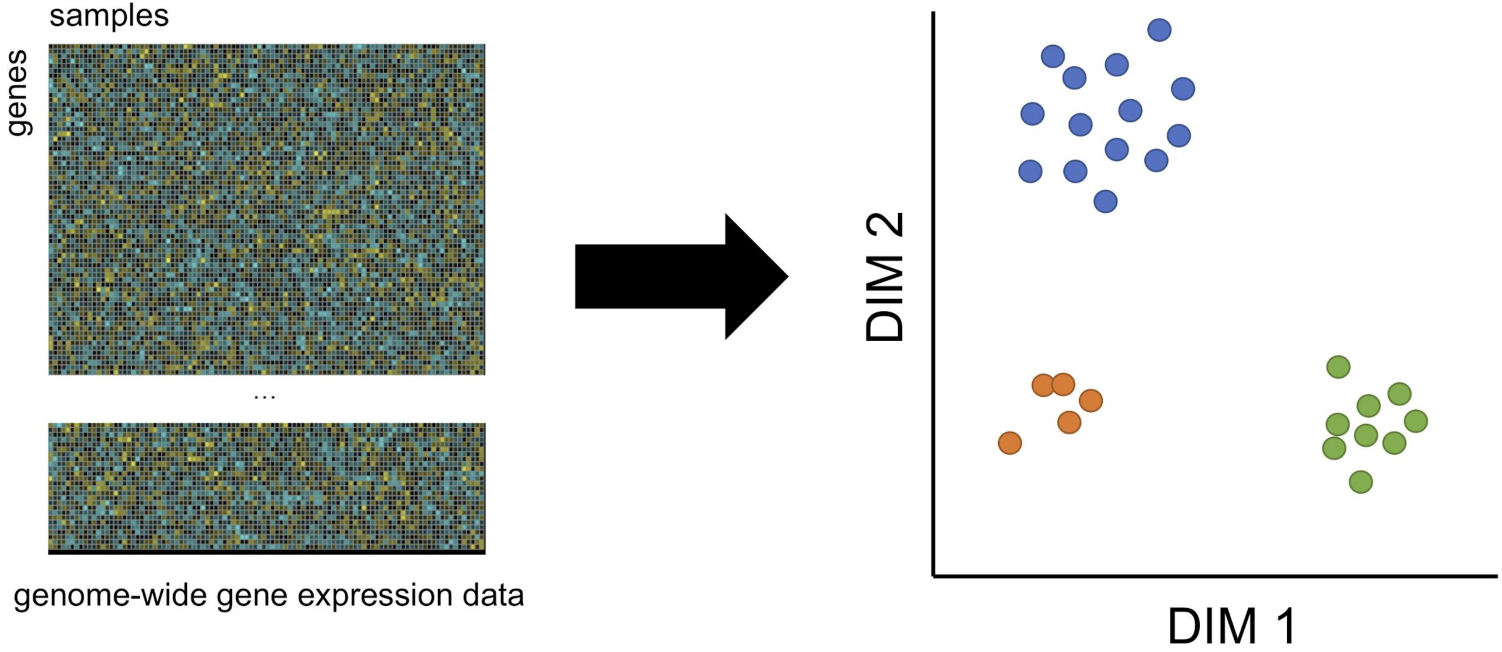
Why do we care about sample-sample relationships?



If we want to find groups of samples with different underlying molecular processes, we will want to know about how stable those groupings are going forward.

Perou et al. “Molecular portraits of human breast tumors.” *Nature.* 2000.

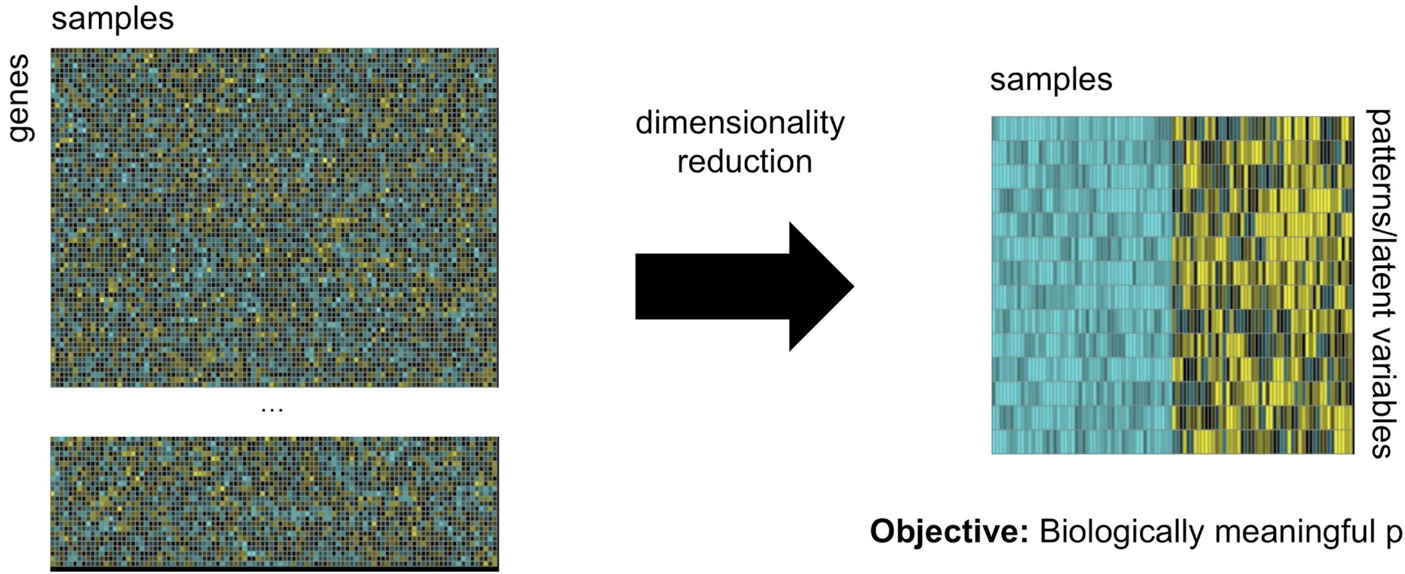
We often want to use dimensionality reduction to explore the structure in our data



We’ll focus on **unsupervised** learning

- Which samples are most similar to one another?
- How many groups of samples exist in my data?
- What patterns of gene expression exist in my data? How do the genes vary together?

Individual genes are coordinated in their expression



genome-wide gene expression data

dimensionality reduction

Objective: Biologically meaningful patterns