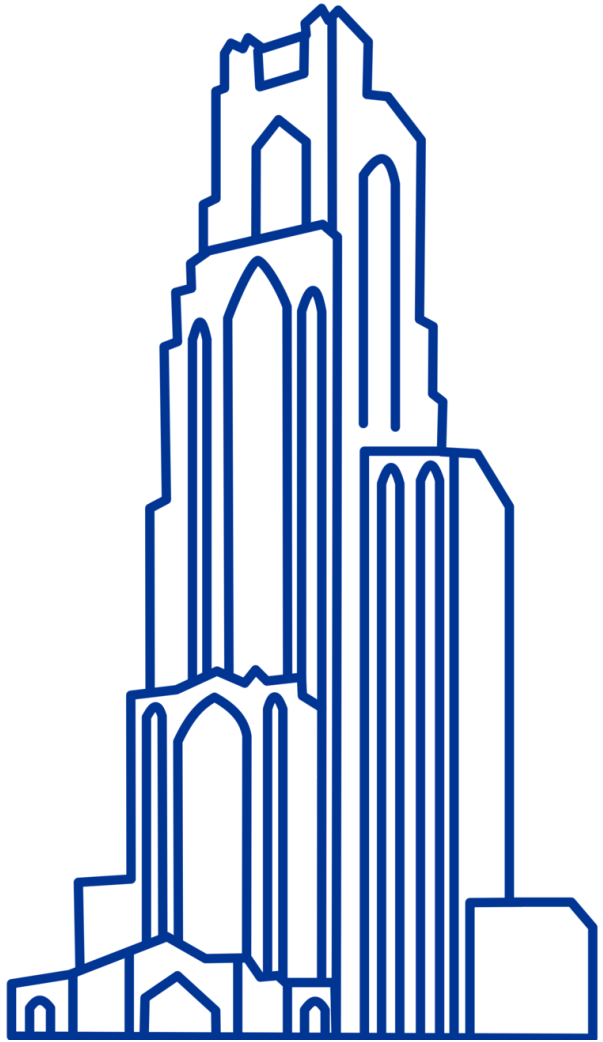


Computational Biology

(BIOSC 1540)

Lecture 07A

RNA Quantification

Foundations

Feb 18, 2025

Announcements

Assignments

- P02A will be released today or tomorrow

Quizzes

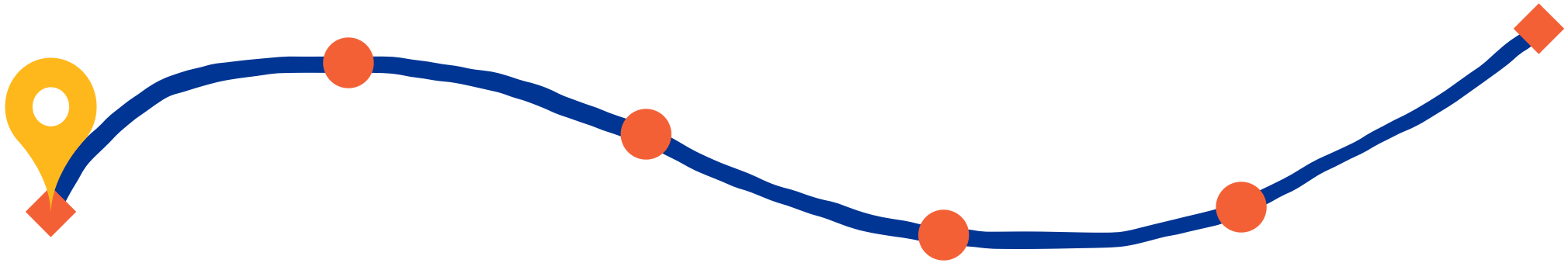
- Quiz 02 is today
- Quiz 03 is on Mar 18 and will cover L06B to L08B

~~**CBytes**~~

CBits

- César will provide optional Python recitations on Fridays from 1 - 2 pm. (Location TBD)
- Please fill out the Canvas discussion for CBit 06

After today, you should have a better understanding of



Quiz 02

Please put away all materials as we distribute the quiz

Sit with an empty seat between you and your neighbors for the quiz

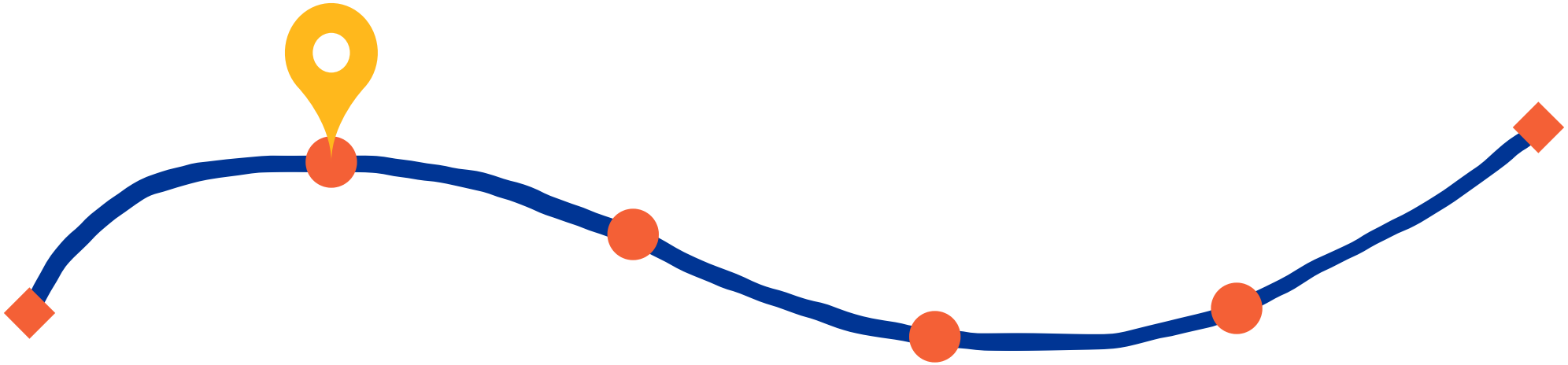
Fill out the cover page, and do not start yet

Quiz ends around **9:50 am**

https://www.clockfaceonline.co.uk/clocks/digital/

When you are finished, please hold on to your quiz and feel free to doodle, write anything, tell me a joke, etc. on the last page

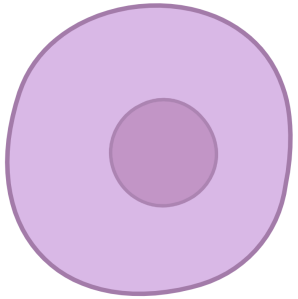
After today, you should have a better understanding of



The importance of quantification in RNA-seq data analysis

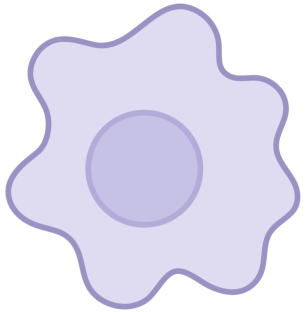
Let's remember the big picture: We want to quantify gene expression differences

Suppose we have isolated a **normal and cancerous cell**



Normal

We want to identify possible drug targets based on **overexpressed genes**

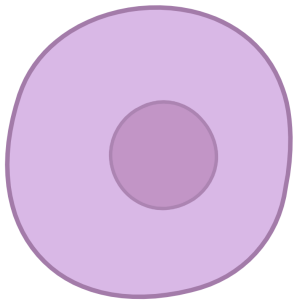


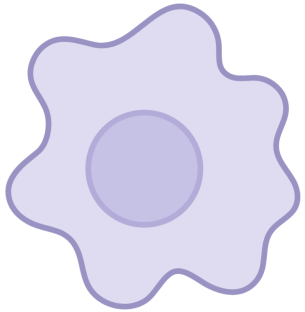
Cancerous

We have to quantify the amount of transcripts in our cell(s)

Our transcriptome is the set of mRNA transcripts our cell could have

We first have to define what possible transcripts our cells have



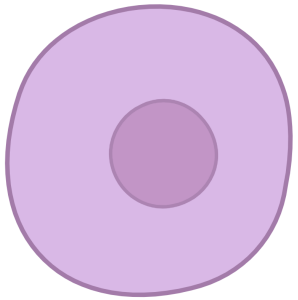


Let's consider only **three transcripts**

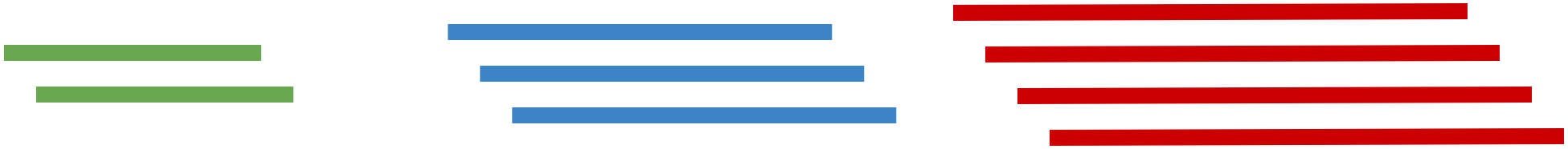
t_1 t_2 t_3

They have short, medium, and long **lengths** $l_3 > l_2 > l_1$

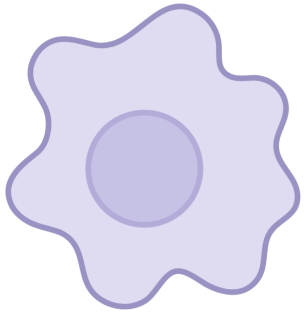
RNA quantification aims to determine the number of transcripts in my original sample



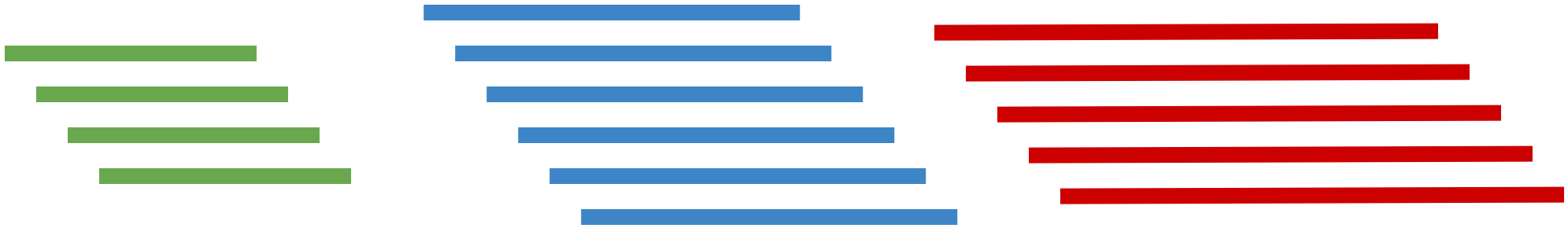
Normal



Example distribution of mRNA under different conditions

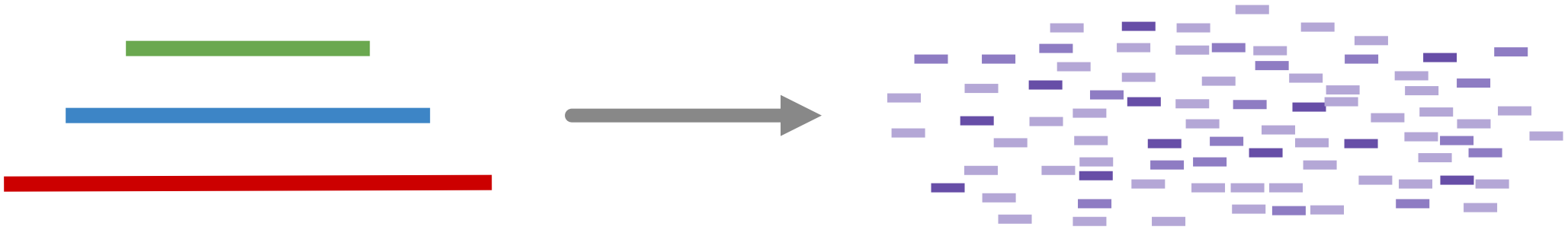


Cancerous



The RNA quantification problem statement

Given the sequencing reads that were sampled from these transcripts



Transcriptome

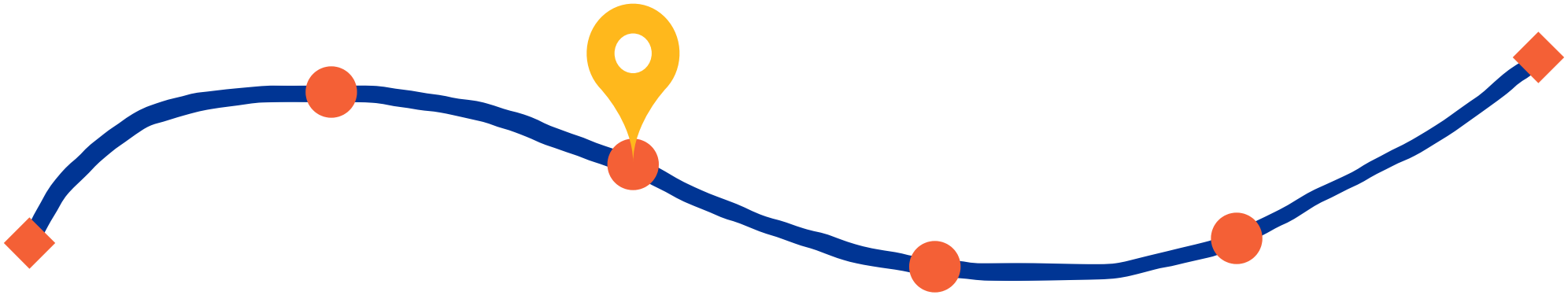
Unknown quantity

Reads/Fragments

Experimental
biases and errors

How many copies of each transcript were in my original sample?

After today, you should have a better understanding of



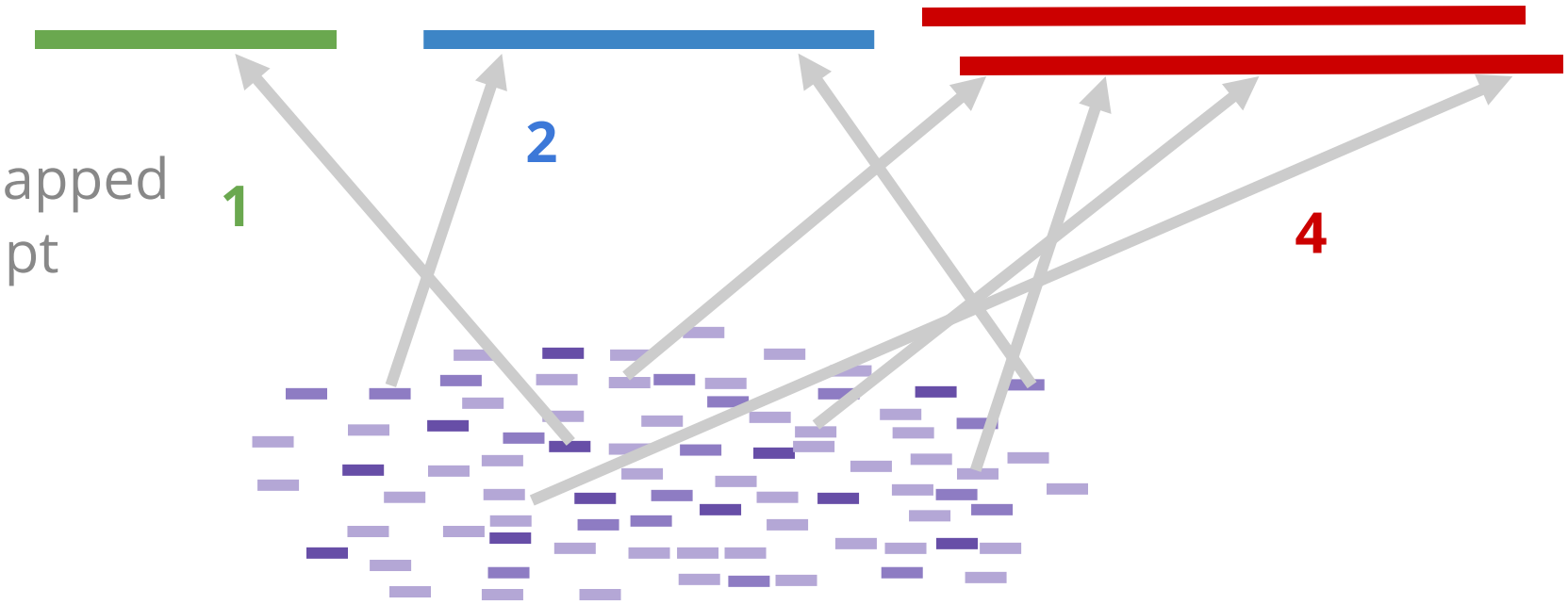
The relevance of pseudoalignment

One way we *could* quantify transcripts is to use read mapping (i.e., read alignment)

Transcriptome

Number of reads mapped to each transcript

Reads/Fragments



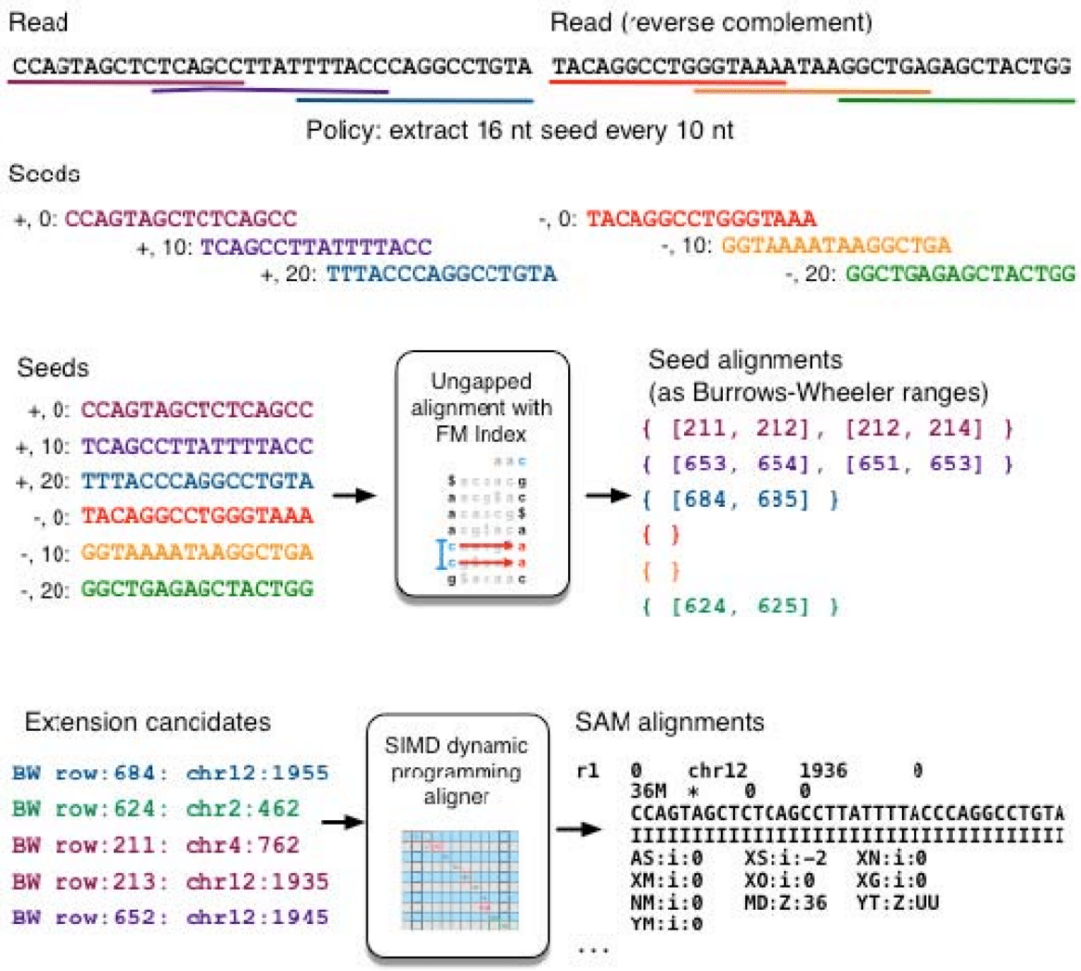
We align each read to single transcript using our read mapping algorithms

Example: Bowtie 2 uses a FM-index for read mapping

1. Extract *k*-mer seeds from all reads in our sample

2. Use **FM-index of transcriptome** to search for *k*-mer matches

3. Extend alignments starting from *k*-mer seeds



These approaches are extremely slow
(we have 30 to 60 million reads)

DOI: 10.1038/nmeth.1923

Alignment-based methods are computationally expensive

Why?

Suppose someone took library books (**transcripts**) and then shredded them (**reads**)

We want to tape these books back together by using another library's copy as a reference

We would need to find **which book** each piece came from, but also **which page**

This would take a **long time**

Relating it back: Alignment-based methods need to determine the read's exact position in the transcript

Pseudoalignment finds which transcript reads came from but not the exact position

Instead of having to find the exact page each piece is from, suppose we just need to **figure out which book**

This is all we need to know for RNA quantification:
Where did this read come from?

Strategy: Identify which transcripts are compatible with the read, skipping the precise location (i.e., alignment) step

Comparison

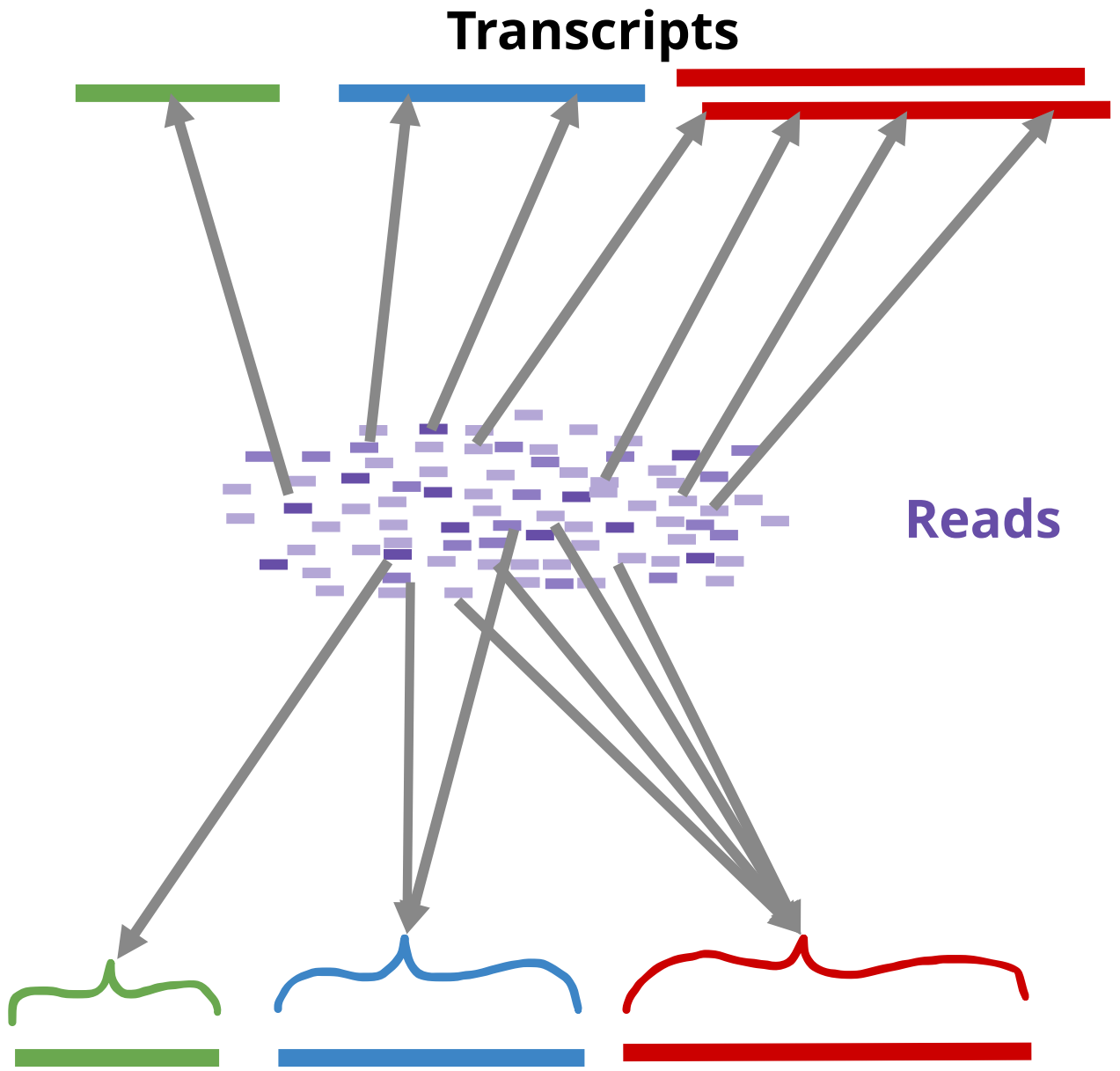
Alignment

Specifies where exactly in the transcript this read came from (e.g., at position 478)

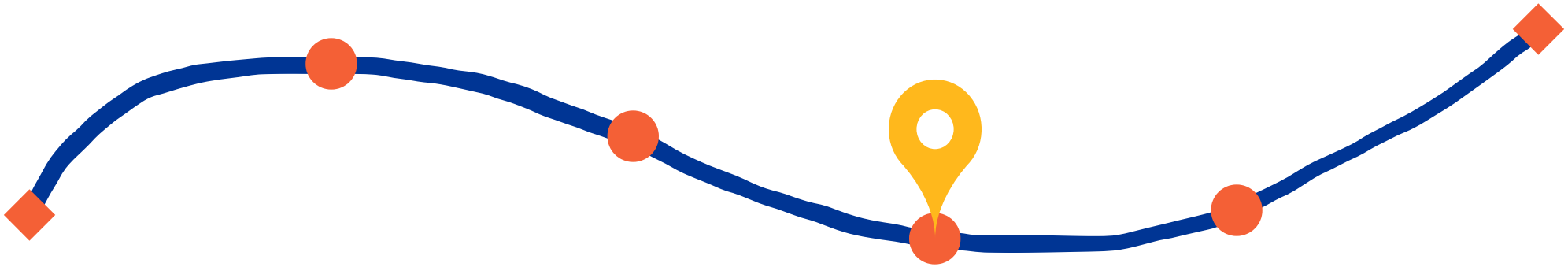
Pseudoalignment

Specifies that it came somewhere from this transcript (i.e., compatible)

Bypassing alignment accelerates quantification, but how can we do this?



After today, you should have a better understanding of



Understand the use of generative models

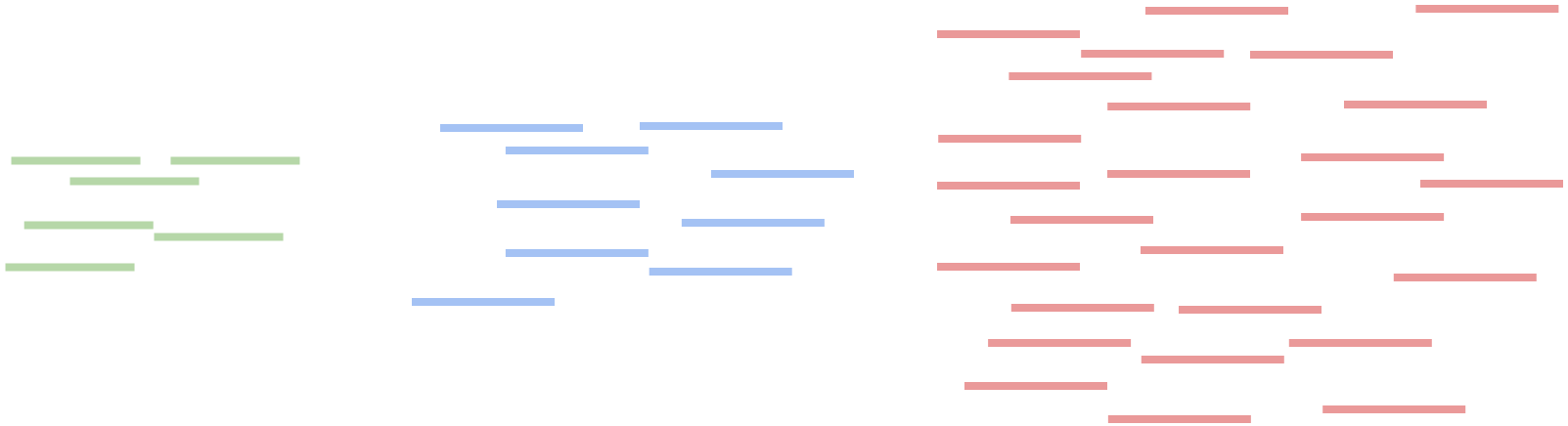
Let's understand our problem

Initial sample

Has some number of transcripts

Fragments

After PCR amplification and fragmentation



Reads

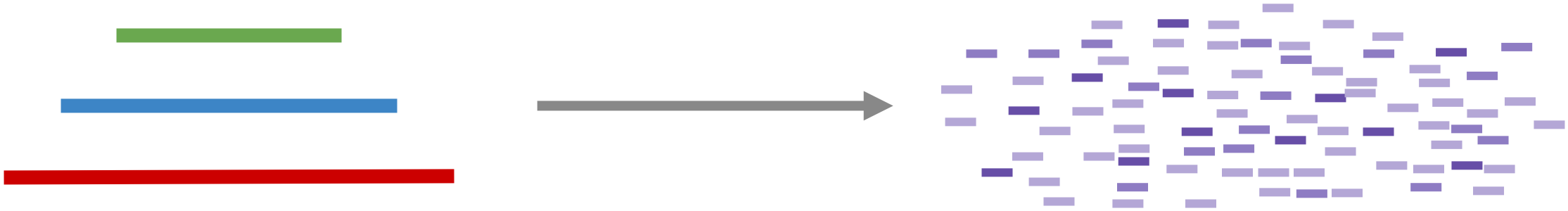
Sequencing with errors

We can only use reads to quantity our initial mRNA sample

We can use a generative model to back-calculate transcript quantities from our reads

Generative model: A statistical model that explains how the observed data are generated from the underlying system

Defines a computational framework that produces sequencing reads from a transcriptome



Let's walk through a conceptual example

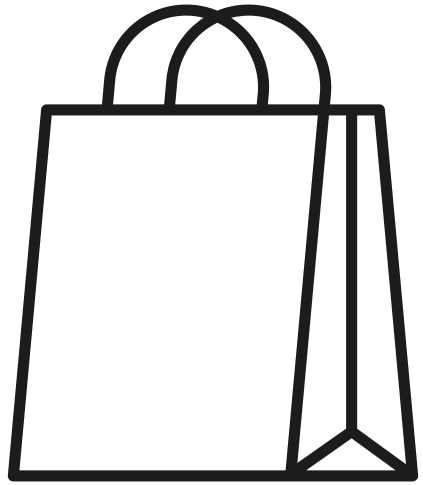
Exploring generative modeling with a bag of marbles

Suppose we take some number of green, blue, and red marbles (i.e., transcripts)

n_1 

n_2 

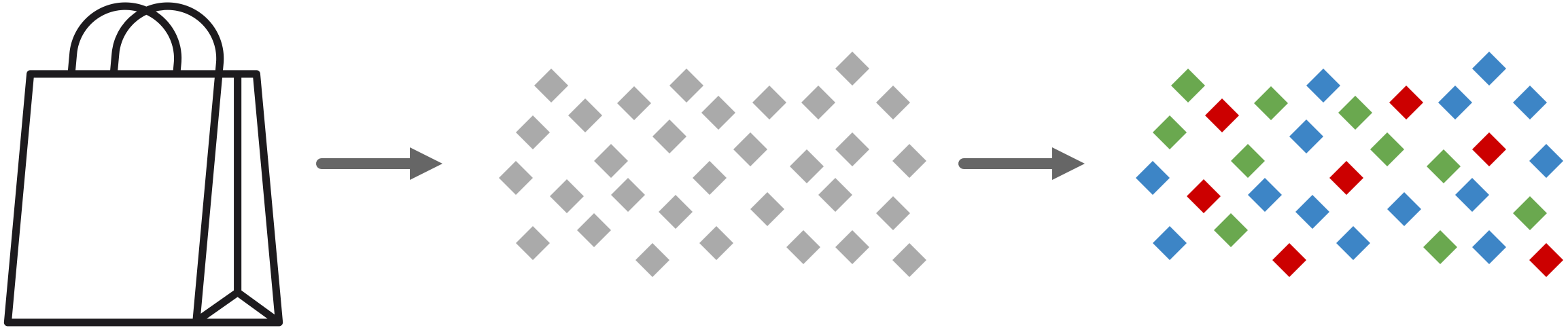
n_3 



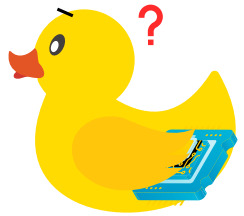
We put the marbles into a bag and crush them (i.e., PCR amplification and fragmentation)

Exploring generative modeling with a bag of marbles

We then take a handful of marble fragments (i.e., reads)
and then determine their color (i.e., map)



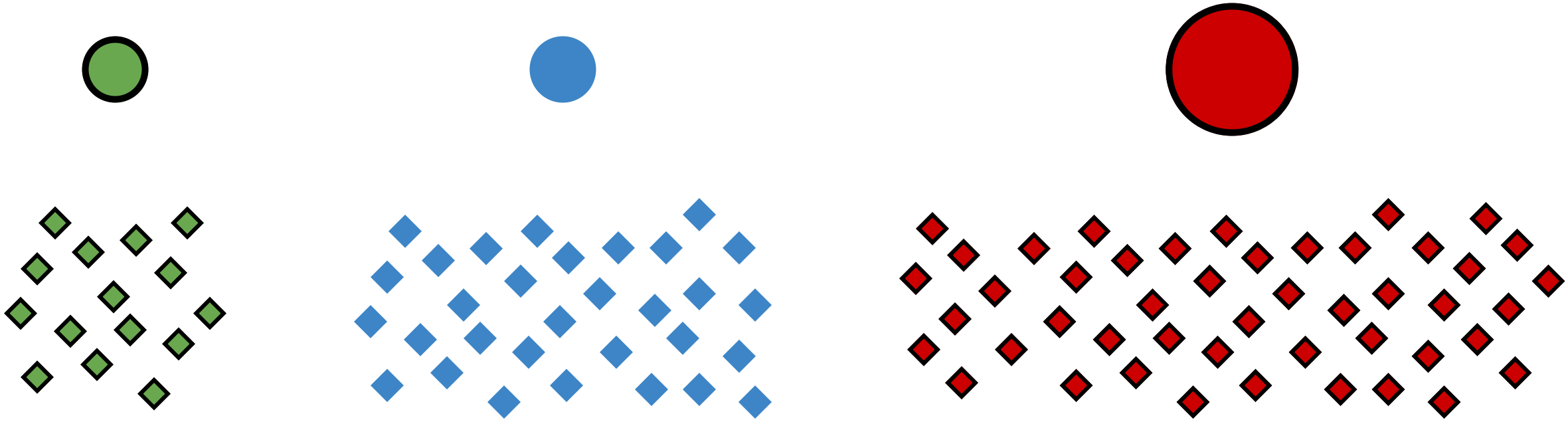
How would you estimate the distribution
of marbles in the bag?



Suppose our marbles have different characteristics

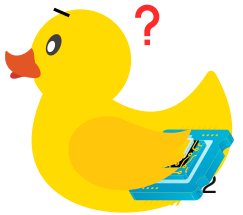
Blue marbles break easily

Red marbles are twice the size



These differences are called **biases**

How would you adjust your approach?

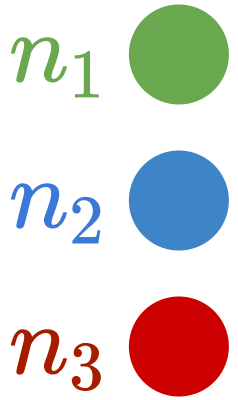


Generative models estimate parameters that explain our observed experimental results

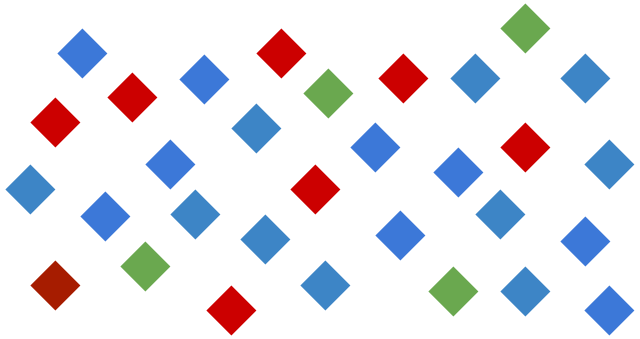
0. Determine/learn bias probabilities of sampling that specify our framework

- The probability of grabbing a blue fragment is 1.4x the green probability
- The probability of grabbing a red fragment is 2.0x the green probability

1. Guess some number of marbles

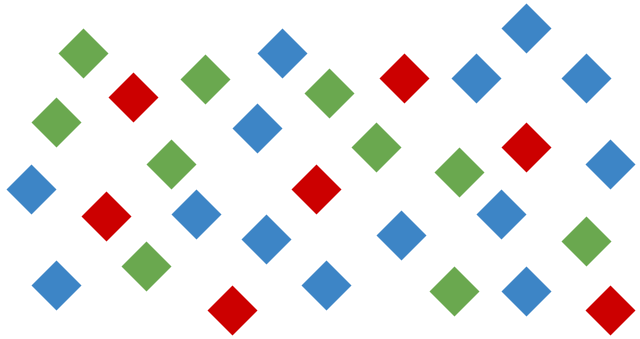


2. Randomly sample *n* fragments



Does not match

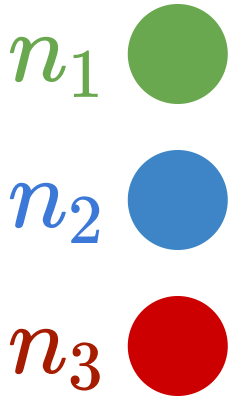
3. Compare to our original distribution



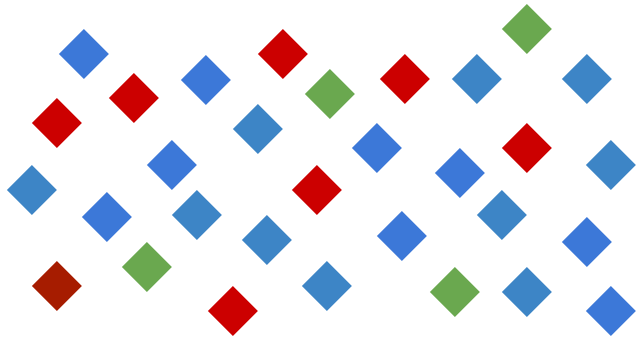
If our simulated distribution closely matches our observed distribution, we have a likely estimate of the original (hidden) distribution

We need to maximize the probability that our generative model and parameters explain our observations

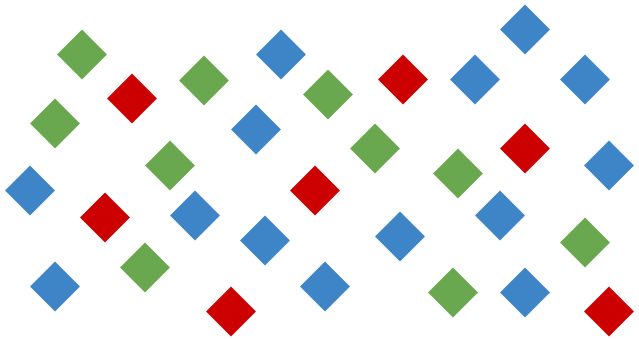
1. Guess some number of marbles



2. Randomly sample *n* fragments

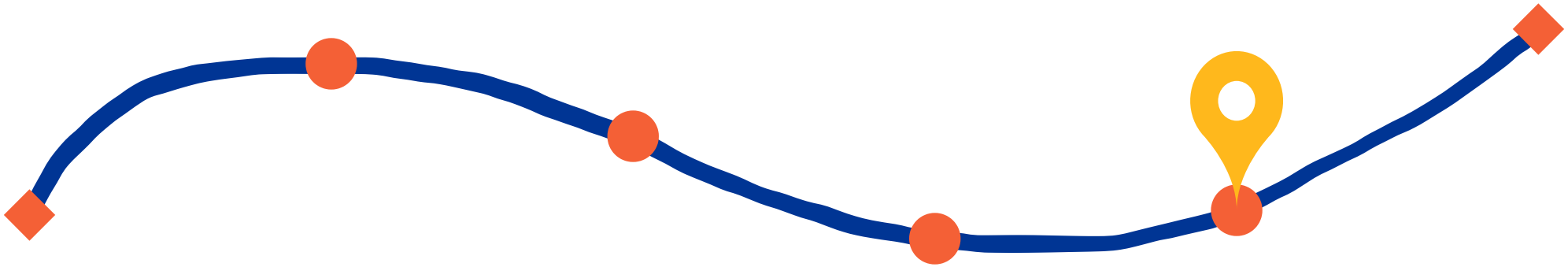


We keep trying new marble combinations until our generated fragments look very similar to our observed fragments



This is conceptually similar to how RNA quantification methods work (e.g., Salmon)

After today, you should have a better understanding of



Python practice

Before the next class, you should

Lecture 07A:

Quantification - Foundations

Lecture 07B:

Quantification - Methodology

Today

Thursday

- Start working on P02A (likely due Mar 4)