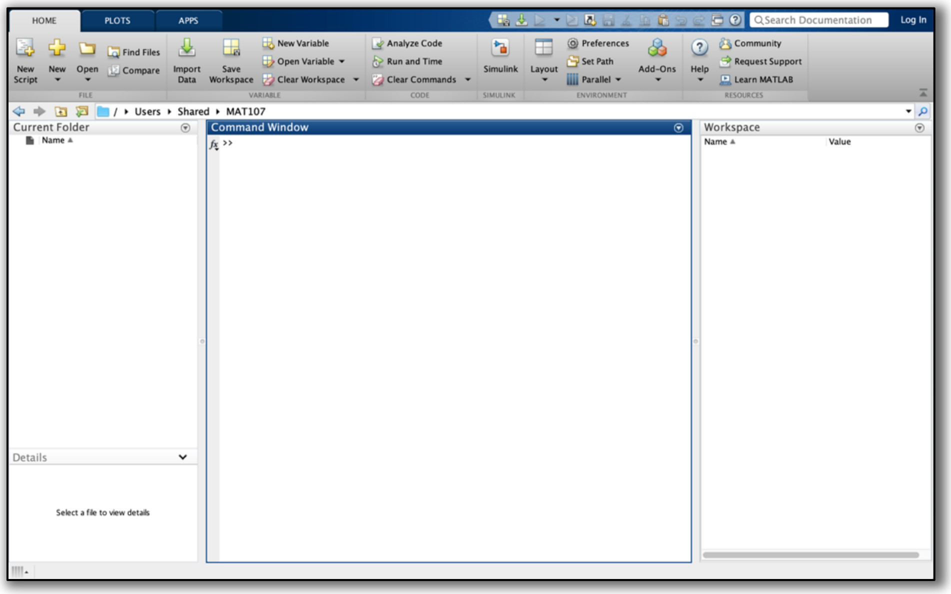
BIM 105
Fall 2024

Lab 1: Counting Problems

Instructions: Use this lab to get started working with the MATLAB interface. Once you get comfortable with basic calculations, work through the problems in Part 1. If you're already familiar with MATLAB, you can skip Part 0 and go right to Part 1. Submit your solution as a single script which outputs all answers to the Command Window. A template script is available on Canvas.

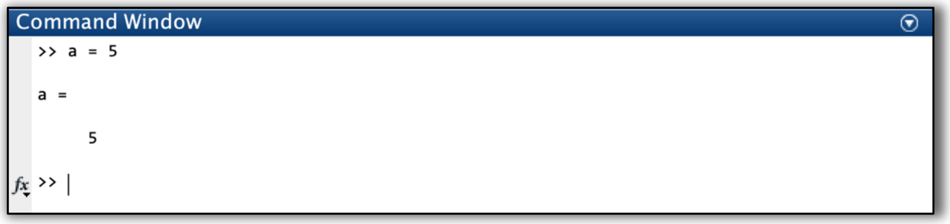
Part 0. Getting Started with MATLAB

MATLAB is a technical computing framework with an accompanying desktop application interface. When you open MATLAB, you will be presented with the default layout, which consists primarily of three panels: **Current Folder** (left), **Command Window** (center), and **Workspace** (right).



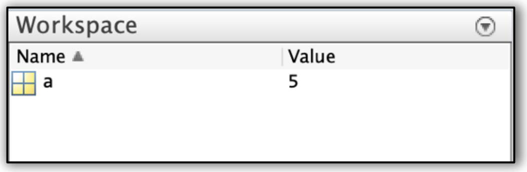
```
>>
```

The Current Folder shows the contents of the current working directory. The address bar above the panels shows the full path of this directory (in this image we are in a folder called "BIM 105"). The Command Window is where you can enter commands/calculations. For example, you can type **`a = 5`** and hit enter to declare that a variable named "a" should be assigned a value of 5.



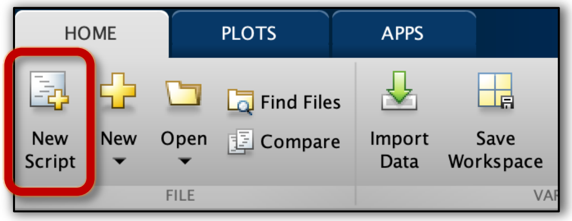
```
>> a = 5

a =

     5

>> |
```

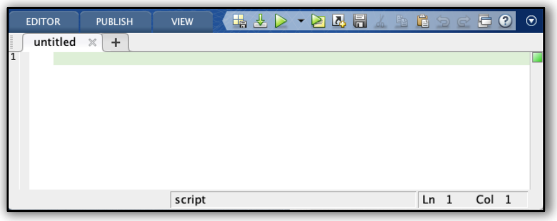
If you now inspect the Workspace panel, you'll see that the variable "a" is listed. The Workspace will show any variables that are currently defined. Generally speaking, variables in MATLAB will persist across commands unless a command which clears the workspace is called.



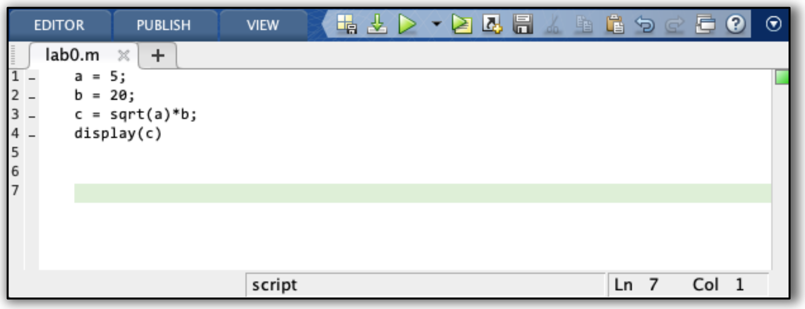
Name ▲	Value
a	5

Typing individual commands in the Command Window is a good way to test commands quickly, but you may find it easier to compile larger pieces of code in a file called a "script". To create a new script, click the "New Script" button above the Current Folder panel. This action will open a new window called the Editor to write and edit your script in (depending on your configuration, the Editor may appear as a fourth panel in the initial desktop window – you can click and drag the panel out of that window if you wish to keep this part of the interface separated).



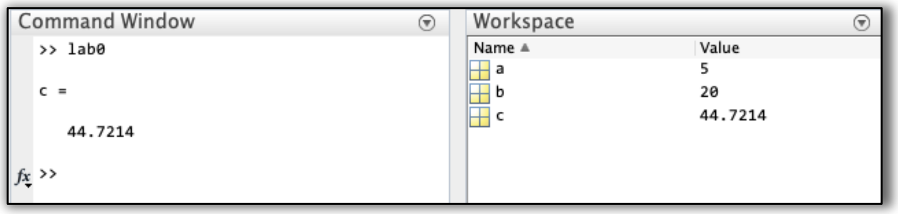


We're now ready to write some code. Let's write a simple example script to make sure everything works.



```
a = 5;
b = 20;
c = sqrt(a)*b;
display(c)
```

After writing the script, it can be run by pressing the green arrow at the top center of the window. You'll have to give the script a name and save it before it can be executed. When the script is run, its variables and output will appear in the Command Window.



```
>> lab0

c =

   44.7214

>>
```

Name ▲	Value
a	5
b	20
c	44.7214

We see that the final value of "c" is displayed in the Command Window, and all variables declared in our script are now in the Workspace.

Here, semicolons ("`;`") make an important difference in your script – a semicolon at the end of a line *suppresses the output* of that line. Any calculations and variable assignments will still be performed, but the final result of that line will not be explicitly printed in the Command Window. It's MATLAB convention to always suppress the output of code within scripts, so if you forget a semicolon on a line, the interface will inform you via a warning. To print the value of "c" at the end of our example script, we subsequently had to use the **`display`** command to print out its value.

There is a plethora of resources to help you learn the basics of MATLAB. The "Help" menu of the application can help you learn basic variable types, understand MATLAB syntax, and even search the full documentation of built-in functions. The official Mathworks website also contains tutorials, videos, and example code to assist in the learning process (https://www.mathworks.com/help/). **It's particularly important to understand arrays/matrices/vectors and the ways MATLAB allows you to index them** (https://www.mathworks.com/help/matlab/learn_matlab/array-indexing.html). If you have questions about how to use a certain function, type **`help %function name%`** to get information on how that function is used (for example, type **`help sum`** to learn about the **`sum`** function).

Part 1. Counting and Combinatorics

DNA has four types of base nucleotides which are denoted with the letters G, C, T, and A. The "sequence" of a DNA molecule is simply the sequence of nucleotide bases that comprise it, for example TAGGCCG or GTTCAAA.

Let's consider how many unique DNA sequences can be constructed from this four-character alphabet. For a sequence with a length of three nucleotides, for example, there are **`4*4*4 = 64`** possibilities.

Question 1: How many unique DNA sequences can be constructed with a length of 12?

One of DNA's primary roles in biology is to encode the genetic information needed to produce proteins. Whereas DNA is constructed by nucleotides, proteins are instead constructed by *amino acids*. When the information from DNA is read to make proteins, its sequence is read in nucleotide *triplets* which are called "codons." For instance, the DNA sequence **ATG** encodes the amino acid methionine; **TGG** encodes the amino acid tryptophan; and **GGG** encodes the amino acid glycine. The sequence **ATGTGGGGG** would subsequently encode a protein with the amino acid sequence methionine-tryptophan-glycine.

There are 20 different amino acids that are used to make proteins. There are 64 possible combinations of 3 nucleotides, however. As such, some amino acids can be encoded by more than one triplet (for example, glycine can be encoded by 4 different triplets: **GGG**, **GGA**, **GGT**, or **GGC**).

Assume we know the amino acid sequence of a protein is methionine-alanine-glycine-leucine. Methionine can be represented by only one triplet (**ATG**); alanine can be represented by four different triplets (**GCT**, **GCC**, **GCA**, **GCG**); glycine can be represented by four different triplets (**GGT**, **GGC**, **GGC**, **GGG**); and leucine can be represented by six different triplets (**TTA**, **TTG**, **CTT**, **CTC**, **CTA**, **CTG**). Because there are multiple ways to encode these amino acids in a DNA sequence, there are many different DNA sequences that would produce this protein. Specifically, there are **`1*4*4*6 = 96`** distinct DNA sequences that encode this protein product.

Question 2: Suppose that an experiment informs you that a protein is comprised by one methionine, one alanine, one glycine, and one leucine. You don't know the order of amino acids in the protein, however. How many ways can these four amino acids be arranged into a protein sequence?

Question 3: In reality, proteins always start with methionine, as **ATG** indicates the beginning of a protein-coding sequence in DNA. In other words, this amino acid must be at the first position of our sequence. How many ways can the protein sequence be formed, given that methionine is at the beginning?

<u>Question 4</u>: Given your answer to Question 3, how many DNA sequences exist that could produce the protein?

<u>Question 5</u>: In Question 1, we computed the total number of unique DNA sequences with length 12. Our protein with four amino acids also corresponds to a DNA sequence with length 12, because each codon is a triplet. What percentage of DNA sequences with length 12 would produce our protein?

Suppose that you have two DNA sequences: $x = (x_1, x_2, x_3, x_4, x_5)$ and $y = (y_1, y_2, y_3, y_4)$. You believe that the sequence y evolved from x, and are interested in investigating this hypothesis. For simplicity, we model the evolution by asserting that differences in the two sequences must be due to either **substitution**, **insertion**, or **deletion**. We characterize the relationship between the two sequences as an *alignment*, and note that for any such pair of sequences there are many different, possible alignments. For example, we may suppose that base x_4 was deleted, and that any other differences are due to substitution. This would produce the following alignment:

$$\begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \\ y_1 & y_2 & y_3 & * & y_4 \end{array}$$

You could also suppose that y_3 was inserted, x_2 and x_4 deleted, and any remaining differences are due to substitution:

$$\begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & * & x_5 \\ y_1 & * & y_2 & * & y_3 & y_4 \end{array}$$

Let's try to compute how many possible alignments exist for these two sequences.

To start, we can simplify the problem to consider only how many alignments are possible <u>under the assumption that exactly 2 differences are due to a substitution</u>. To account for two substitutions, two bases from each x and y must be chosen. This leaves three bases from x and two bases from y remaining to be aligned. Because there are no remaining substitutions possible under our assumption, the three remaining bases from x must have been deleted and the two remaining bases from y must have been inserted. The number of alignments under this assumption, then, is the number of choices from two (unordered) bases from each of x and y. We can use the **`nchoosek`** function in MATLAB to compute this number as:

```
n_alignments_2_subs = nchoosek(5,2)*nchoosek(4,2)
```

which works out to 60 possible alignments. In this computation, we are calculating the number of ways to assign two substitutions between our two sequences. This number is the product of the number of ways to choose two bases from x (which has five elements) and the number of ways to choose two bases from y (which has four elements). We are also making an assumption that the ordering of the remaining changes (deletions and insertions) are irrelevant. In other words, the two alignments below would be considered the same:

$$\begin{array}{cccc} x_1 & x_2 & x_3 & * \\ y_1 & * & * & y_2 \end{array} \qquad \begin{array}{cccc} x_1 & x_2 & * & x_3 \\ y_1 & * & y_2 & * \end{array}$$

Question 6: How many possible alignments exist for the two sequences x and y? (Hint: Compute the number of alignments at each possibility of the number of substitutions in the alignment, then sum the total number of alignments from this)

Question 7: How many possible alignments exist for two sequences if both sequences have length 15?