# Introduction to Scientific Computation 113E

Variables - data1

corr_coef_y | allmeansdata | p_value_o | data1

27130x34 table

| | 1<br>Gene | 2<br>ARNA | 3<br>TRNA | 4<br>ARNA1 | 5<br>TRNA1 | 6<br>ARNA2 | 7<br>TRNA2 | 8<br>ARNA3 | 9<br>TRNA3 | 10<br>TRNA4 | 11<br>TRNA5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | "LOC1024... | 0 | 1.0045 | 4.0185 | 0.9162 | 2.9799 | 3.3377 | 1.3212 | 2.1511 | 1.0805 | |
| 2 | "ZBTB42" | 27.8394 | 37.1676 | 55.2547 | 30.2348 | 42.7112 | 32.2643 | 54.1705 | 48.3991 | 35.6578 | 50.26 |
| 3 | "FCAMR" | 1.1136 | 0 | 1.0046 | 0 | 0.9933 | 0 | 0 | 0 | 0 | |
| 4 | "ZNF503-... | 41.2024 | 35.1586 | 40.1853 | 16.4917 | 35.7582 | 32.2643 | 60.7767 | 23.6618 | 47.5438 | 40.84 |
| 5 | "NFU1" | 111.3578 | 123.5573 | 118.5465 | 133.7663 | 91.3821 | 86.7800 | 76.6315 | 120.4600 | 100.4903 | 68.07 |
| 6 | "ELSPBP1" | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 7 | "ZRANB3" | 190.4218 | 183.8291 | 141.6531 | 169.4984 | 155.9455 | 92.3428 | 118.9109 | 121.5355 | 152.3563 | 102.63 |
| 8 | "MECR" | 259.4637 | 291.3139 | 188.8708 | 237.2977 | 289.0455 | 189.1358 | 257.6403 | 253.8264 | 347.9341 | 271.24 |
| 9 | "LOC1057... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 10 | "LINC003... | 2.2272 | 2.0091 | 3.0139 | 4.5810 | 6.9530 | 0 | 0 | 0 | 1.0805 | 4.18 |
| 11 | "AARSD1" | 1.1136 | 0 | 0 | 0 | 0 | 2.2251 | 0 | 0 | 0 | |
| 12 | "DEXI" | 485.5200 | 435.9663 | 492.2695 | 308.7619 | 511.5410 | 493.9782 | 478.2860 | 357.0779 | 467.8742 | 569.72 |
| 13 | "DCHS1" | 1.1559e+03 | 1.0035e+03 | 1.2116e+03 | 1.3138e+03 | 1.3479e+03 | 1.4652e+03 | 1.8788e+03 | 1.4842e+03 | 1.7224e+03 | 1.2724e+ |
| 14 | "PSMD2" | 1.2550e+03 | 1.8232e+03 | 1.3914e+03 | 1.4861e+03 | 1.5545e+03 | 1.3206e+03 | 1.5630e+03 | 1.1282e+03 | 1.4382e+03 | 1.5730e+ |
| 15 | "GABRR1" | 3.3407 | 4.0181 | 2.0093 | 1.8324 | 5.9597 | 8.9005 | 6.6062 | 2.1511 | 2.1611 | 4.18 |
| 16 | "PKNOX2" | 780.6181 | 676.0491 | 640.9550 | 244.6274 | 522.4672 | 406.0857 | 486.2134 | 287.1680 | 504.6126 | 799.07 |
| 17 | "TIPARP" | 309.5747 | 294.3275 | 372.7184 | 721.0553 | 238.3881 | 354.9078 | 395.0484 | 379.6641 | 298.2293 | 273.34 |
| 18 | "ADAM20" | 113.5849 | 91.4123 | 74.3427 | 89.7883 | 100.3216 | 160.2091 | 104.3773 | 73.1364 | 63.7519 | 97.39 |
| 19 | "LOC2847... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 20 | "MIR4715" | 0 | 0 | 0 | 0 | 0 | 1.1126 | 0 | 0 | 0 | |

Assc. Prof. Halil Bayraktar

Lecture 4

The relational operators in MATLAB are:

> greater than

< less than

>= greater than or equals

<= less than or equals

== equality

~= inequality

The logical operators are:

| or for scalars

& and for scalars

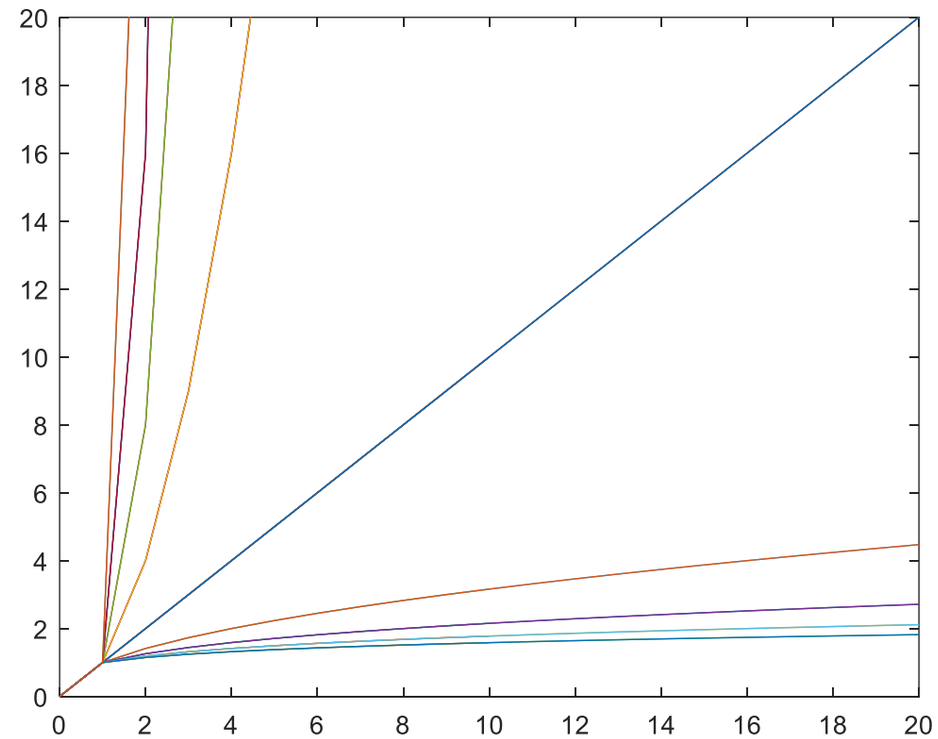
~ not  (tilde symbol)

# For loop

- Used to repeat the computation within the loop

```
% for loops
x=1:1:100
y=reshape(x,10,10)

for i=1:10
   y(i,1)=i+5

end
```

```
% repeating with for loops

for i=1:100;
   disp(i)
   fprintf('%f',i)
    disp('matlab')
    x(i,1)=i
    x(i,2)=sqrt(i)

end
```



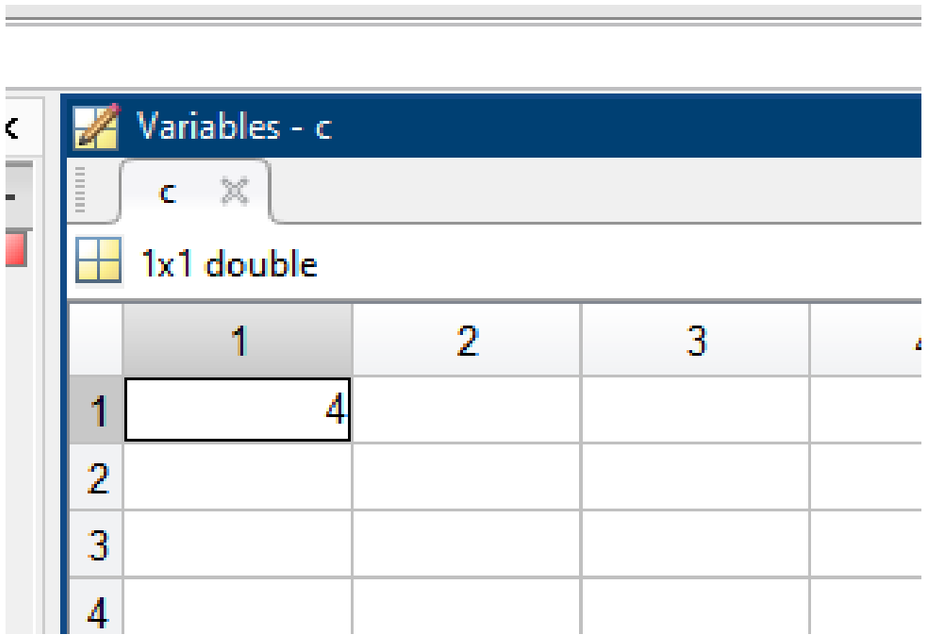

```
%%
for i=1:5
   x=0:1:20
y=x.^(1/i)
y1=x.^(i)
figure(1)
plot(x,y)
hold on
plot(x,y1)
axis([0 20 0 20])
hold on
end
```

```
geneA='attgta'
geneB='attcta'
res1=geneA==geneB
```

```
res1 =

  1×6 logical array

   1   1   1   0   1   1

fx >>
```

== egual

~= not equal




```
[r,c,u]=find(geneA~=geneB)

geneD=[geneA,geneB,geneA,'ccc']

save('geneD','geneD')
```

```
geneD=[geneA,geneB,geneA,'ccc']

save('generesults','geneD')
%%
% do not work
geneD=geneA+geneB

%%
nucleotide='atcg'
%%
chr1=''

for i=1:10000
    x=randi([1,4],1,1)
    chr1=[chr1,nucleotide(x)]

end
%%

%[r,c,u]=find(geneA=='att')
k = strfind(chr1,'ttatttt')
disp(chr1(2549:2560))
%%
```

# Example : Search a region of nucleotides

```
x='acgc'
y='atgg'
z=x==y
[r,c,l]=find(z==0)
disp(c)
%%
x='atcg'
chr1=''
arrChr={}
for j=1:10;
for i=1:10000;
    num=randi([1,4],1,1);
    chr1=[chr1,x(num)];

end
arrChr{j,1}=chr1
  chr1=''
end
disp(chr1)
%%
targetsequence='ggcgg'
k=strfind(arrChr{6,1} ,targetsequence )
disp(k)
%%
targetregion={}
targetsequence='ctgg'
for i=1:10
k=strfind(arrChr{i,1} ,targetsequence )
disp(k)
 targetregion{i,1}=k

end
```

# Applications of relational operators: Find an information in an array

returns the row and column indices of non-zero entries in a matrix.

| | | |
|---|---|---|
| 1 | 2 | 2 |
| 4 | 6 | 9 |
| 1 | 10 | 9 |

```
a=find(x>7)

ans =

     6
     8
     9
```

```
[row,col,v]=find(x>7)

row =
     3
     2
     3
col =
     2
     3
     3
v =
3×1 logical array
   1
   1
   1
```

# And, or and not operator

- And, &
- Or, |
- Not equal, ~=

```
disp(char(2049:2500))
%%

x=randi([1,100],100,1)
y=randi([1,100],100,2)
```

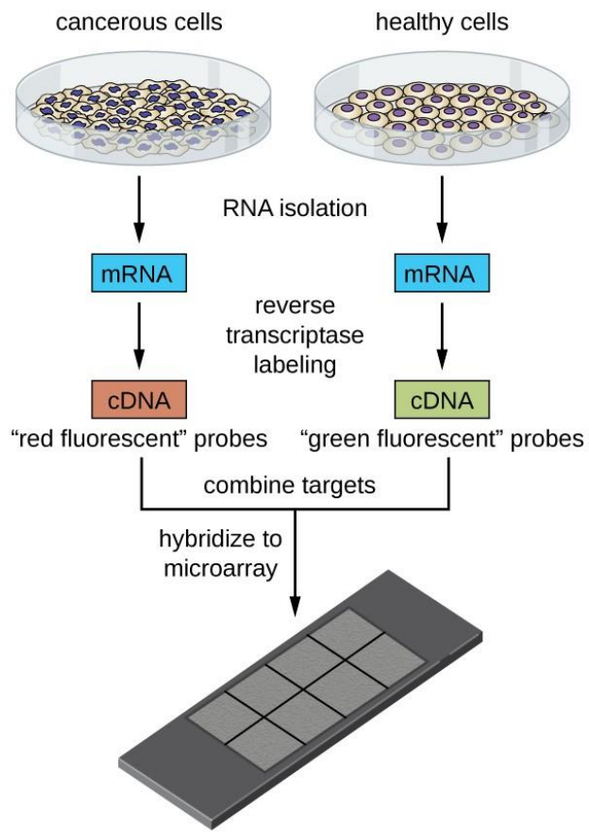
```
%%
[r,c,l]=find(x>20 & x<40)
[r,c,l]=find(y(:,1)<5 | y(:,2)<5)
```

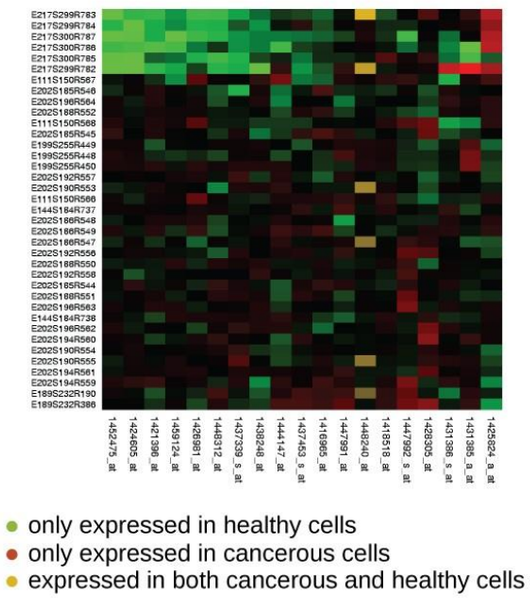
```
%%
[r,c,l]=find(x~=13)
[r1,c,l]=find(x~=[13,20,89])
[r1,c,l]=find(x~=13 & x~=20)
```

# Exampe 2: Compare genes and find unmatched nucleotides

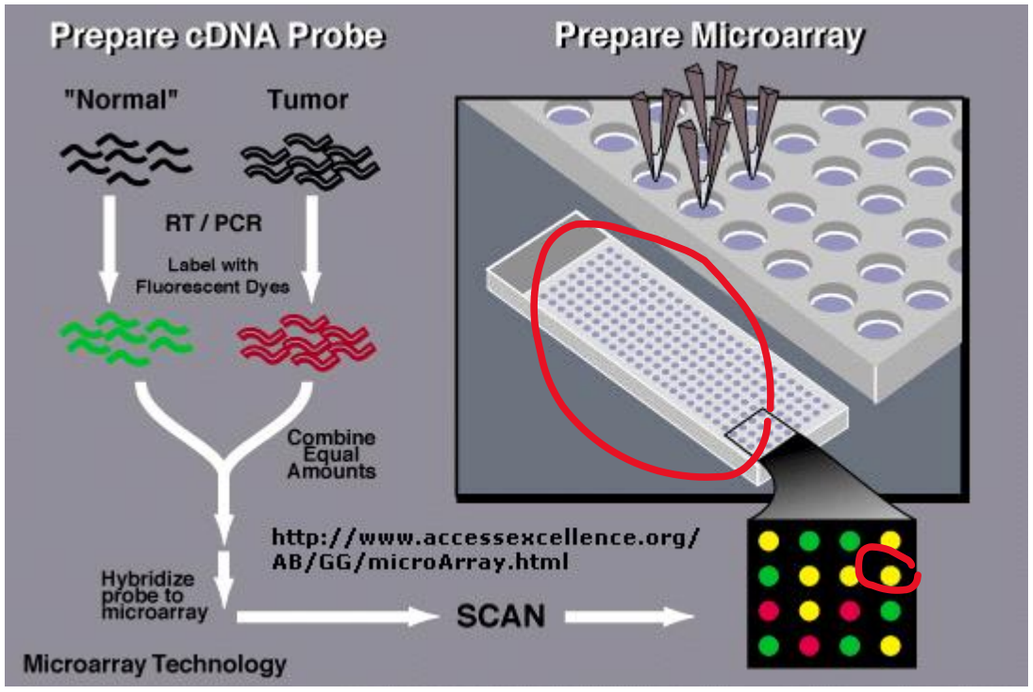
```
geneA='AAAATAGTAGATGATGATGATGATGTCCATATAT'
geneB='AAAATATGTAATTGTATGGATGTCCATATAT'
[row,col,v]=find(geneA~=geneB)
```

# Compare data with logic operators



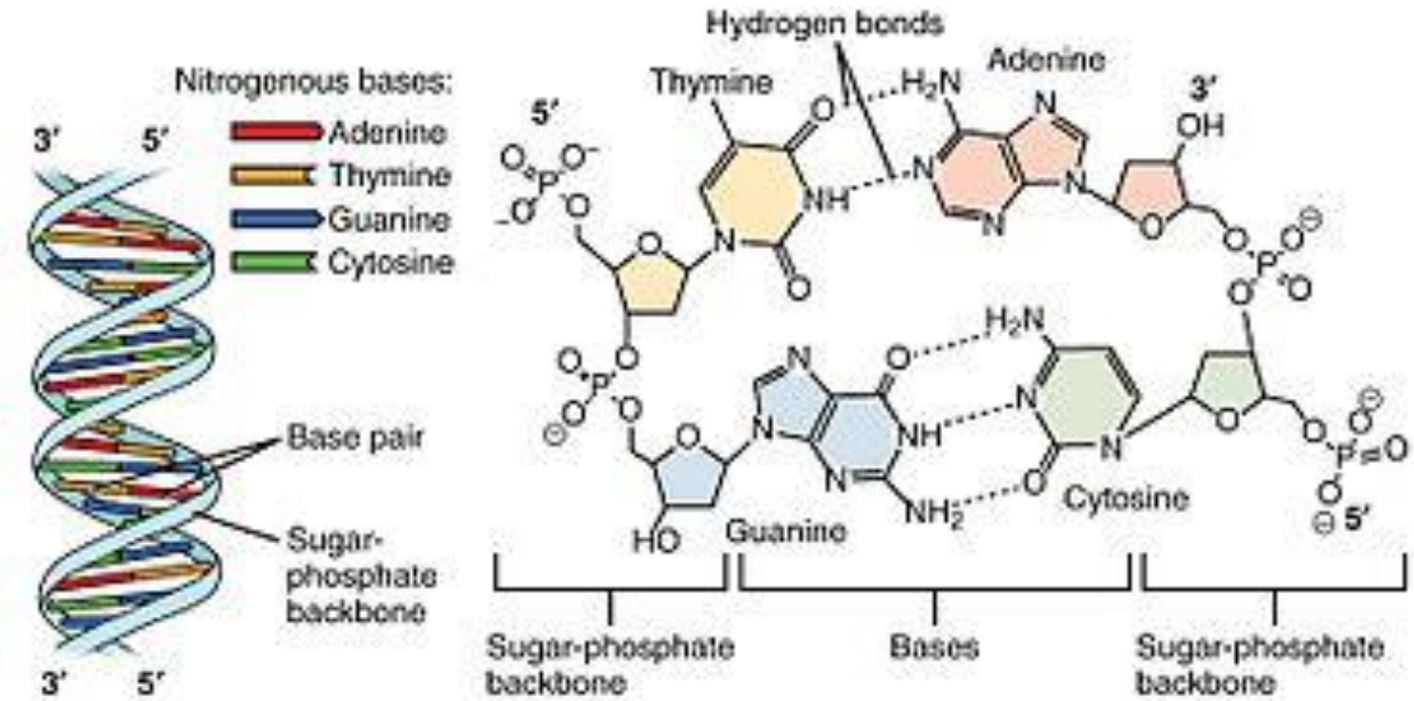

```
geneA=randi(100,1000,1)
geneB=randi(100,1000,1)
geneC=randi(100,1000,1)

%

x=find(geneA>90 & geneB>90
& geneC<90)
geneA(x,1)
geneB(x,1)
geneC(x,1)
```

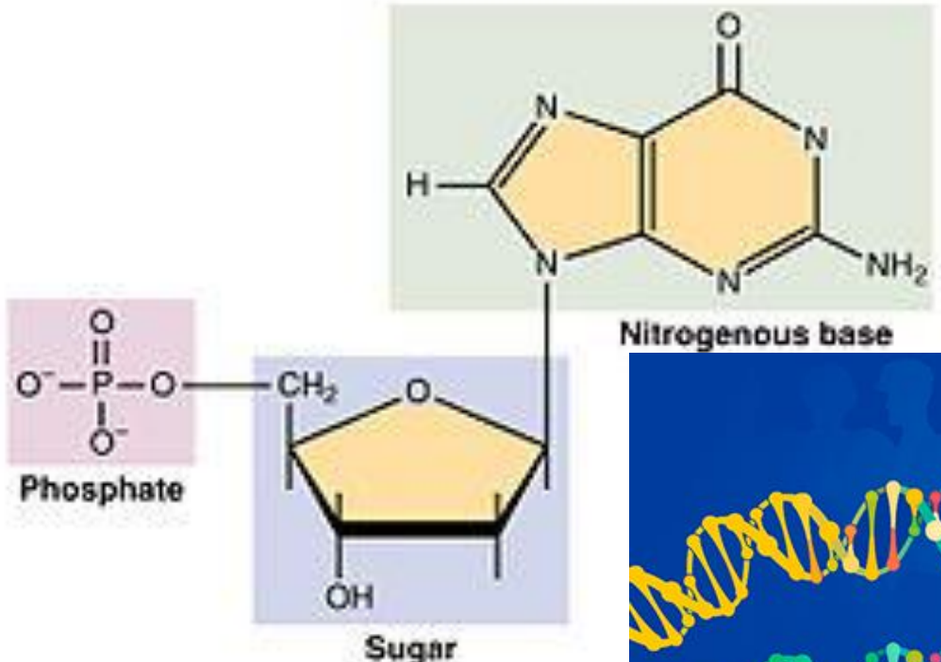


# A genome sequence with for loop

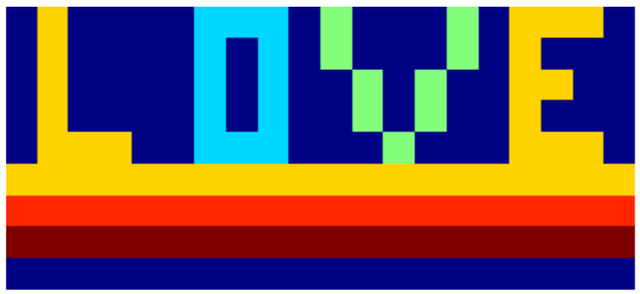
We need 4 letters

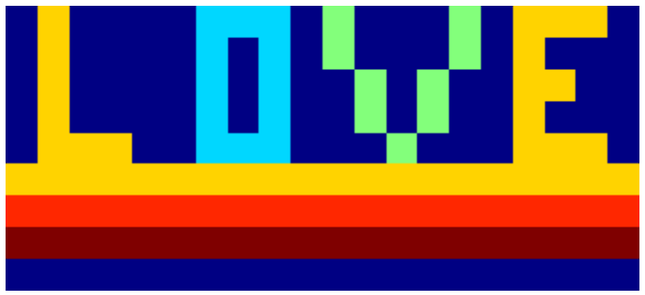
```
%%
letters=['a','g','c','t']
genex=''
for i=1:100
    a=randi([1,4],1,1)
     genex(i)=letters(a)

end
```



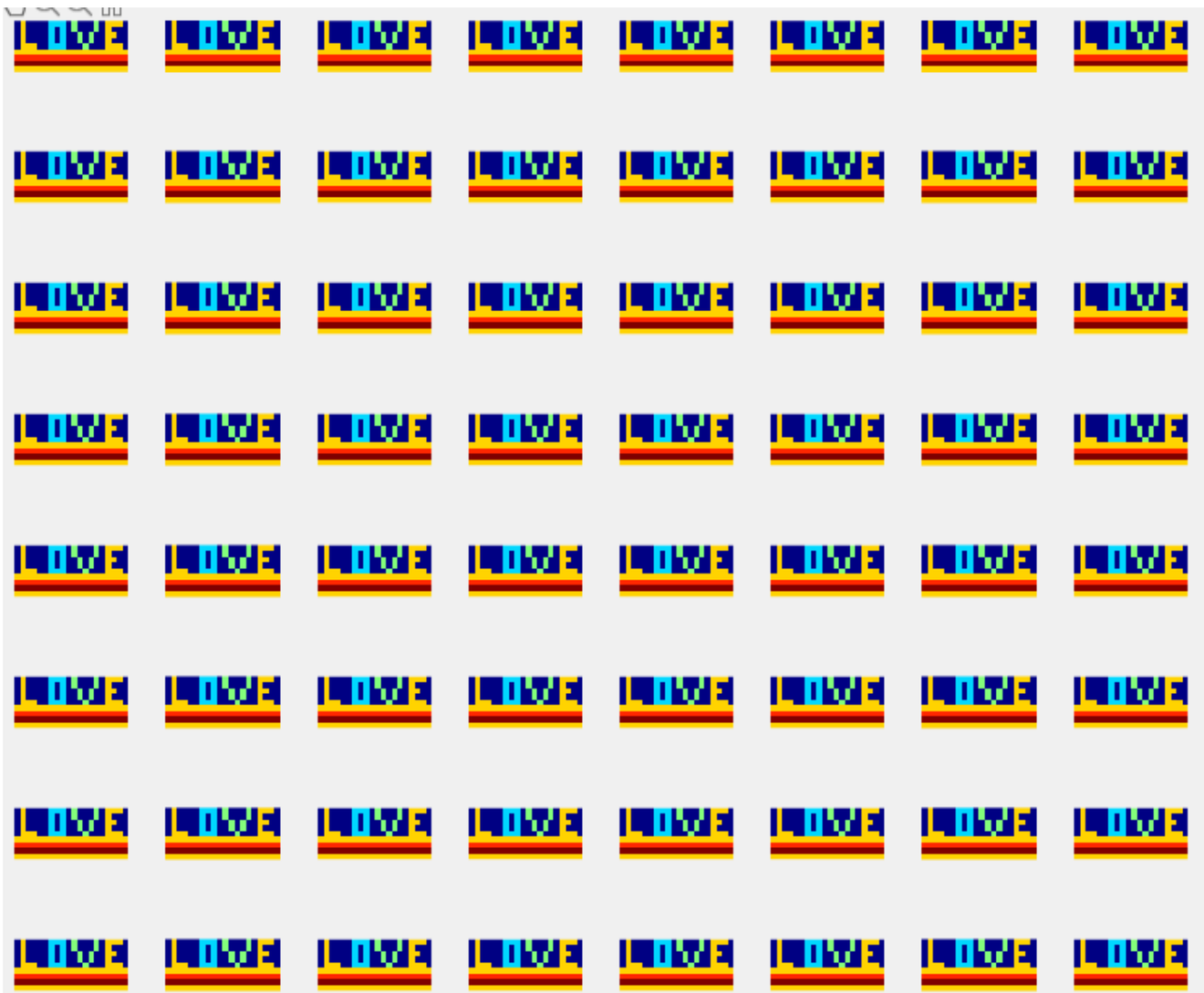




# How can we imagine Arrays in 2D? Can we print Love with many colors? Can we print mant of them?



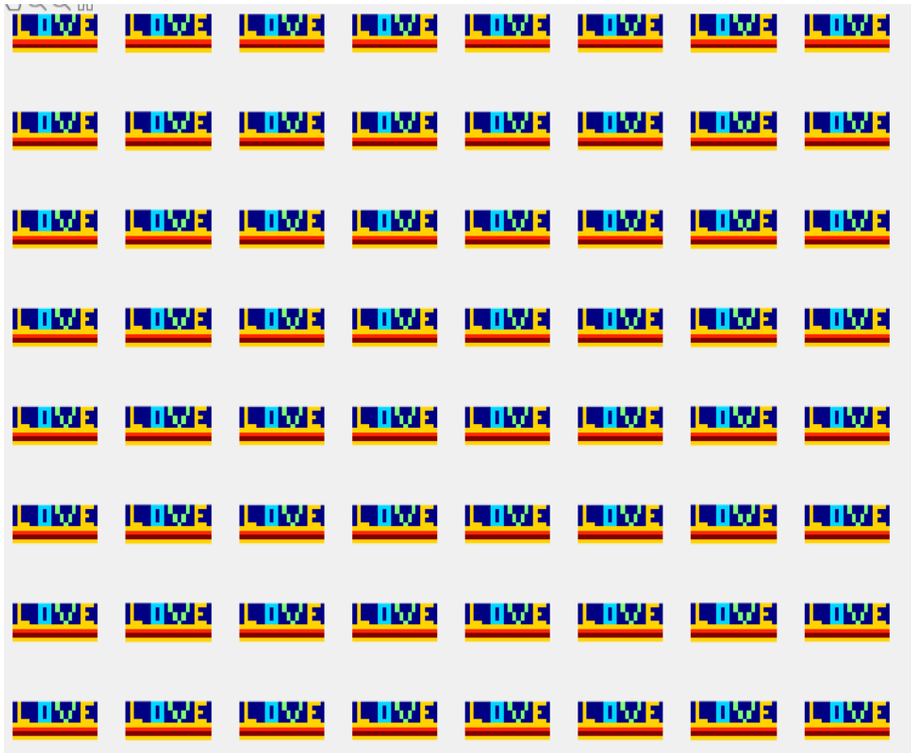


```
%%
% love with colors
k=zeros(9,20)
k(6,:)=4
k(7,:)=5
k(8,:)=6
k(9,:)=0
%L
k(1:5,2)=4
k(5,3:4)=4
%0
k(1:5,7)=2
k(1:5,9)=2
k(1,8)=2
k(5,8)=2
%v
k(1:2,11)=3
k(3:4,12)=3
k(5,13)=3
k(1:2,15)=3
k(3:4,14)=3
%E
k(1:5,17)=4
k(1,17:19)=4
k(5,17:19)=4
k(3,18)=4
figure(1)
subplot(1,2,1)
imshow(k,[],'initialmagnification',1200)
subplot(1,2,2)
imshow(k,[],'initialmagnification',1200)
colormap jet
%%
```

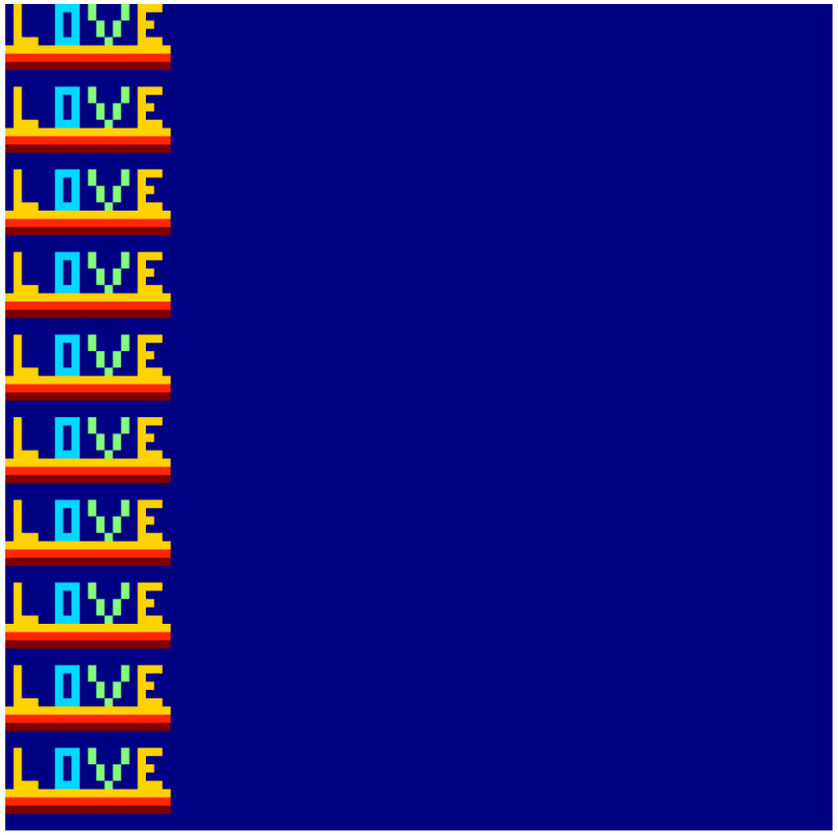


```
for i=1:64;
  figure(1)
  subplot(8,8,i)
  imshow(k,[],'initialmagnification',1200)
  hold on
  colormap jet

end
```

# Can we organize arrays with different ways?

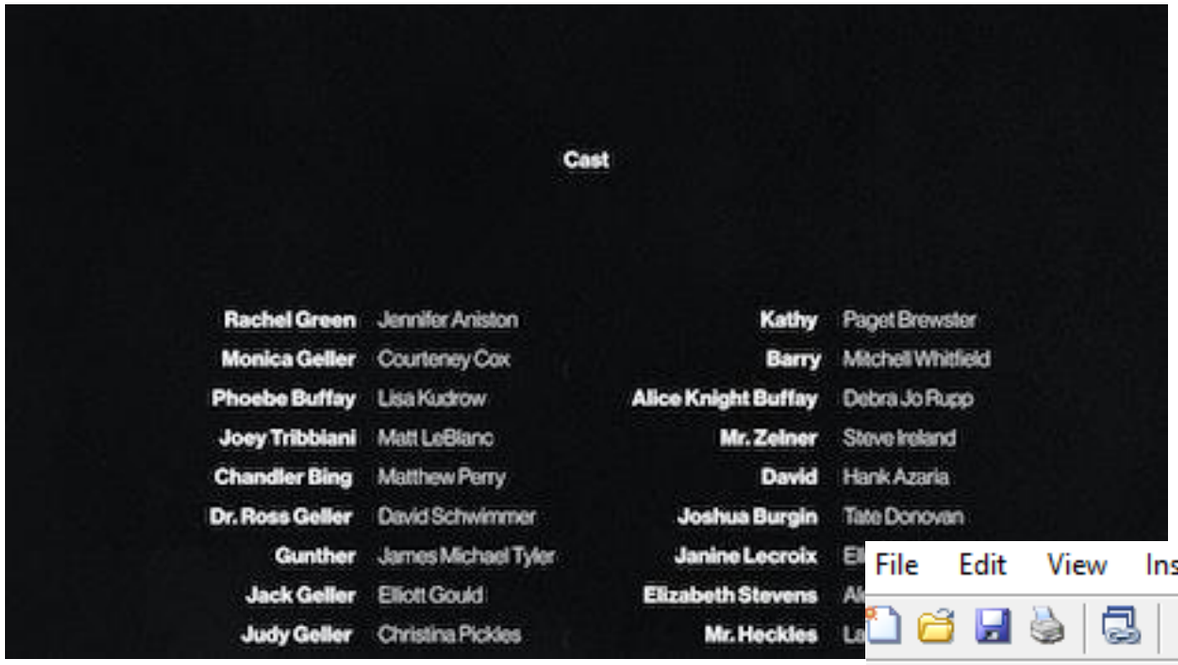


```
for i=1:64;
    figure(1)
    subplot(8,8,i)
    imshow(k,[],'initialmagnification',1200)
    hold on
    colormap jet

end
```

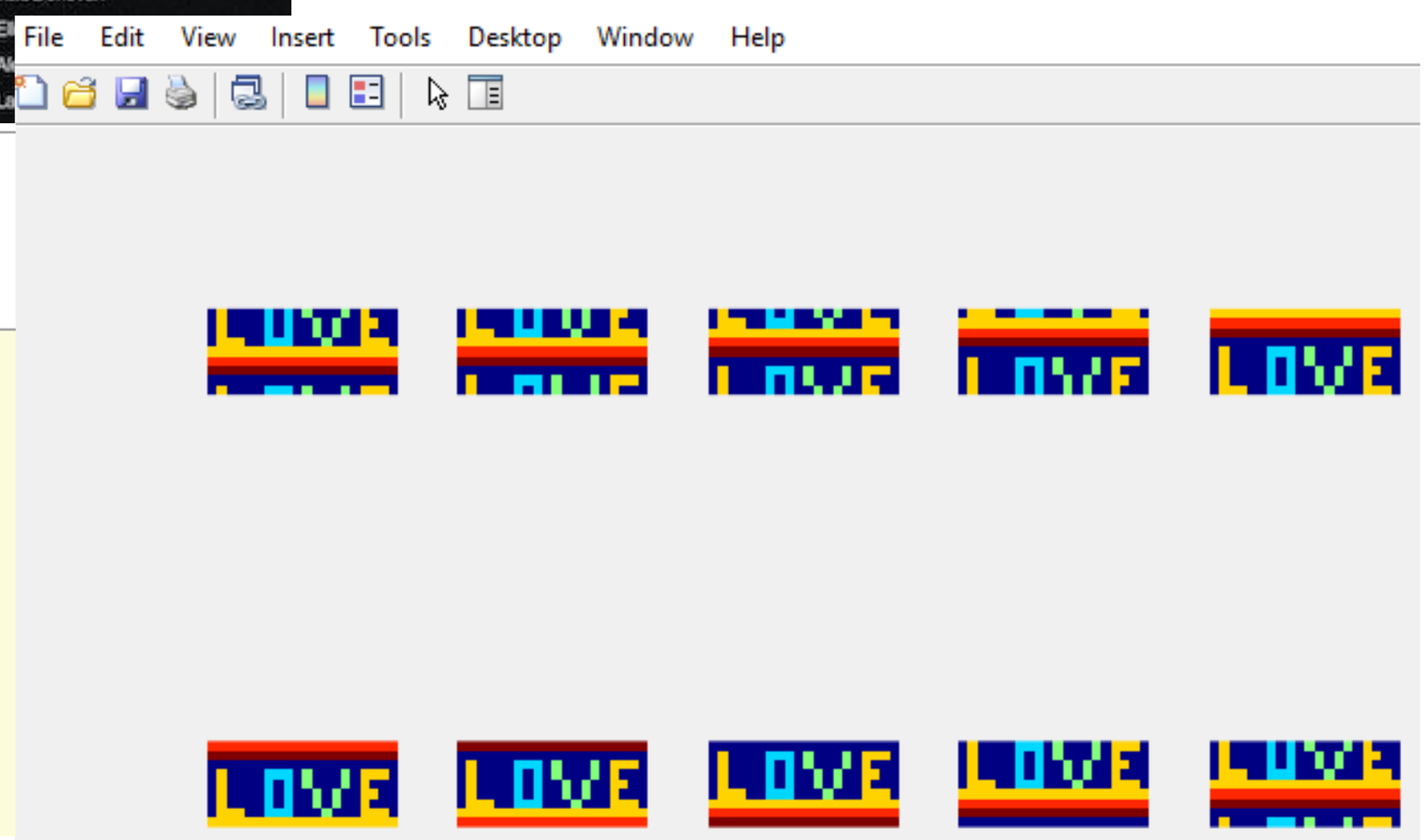


```
arr=zeros(100,100)
for i=1:10:100
    arr(i:i+8,1:20)=k
    %arr(i:i+4,30:32)=k(1:5,7:9)

end
figure(4)
%%
imshow(arr,[],'initialmagnification',1200)
hold on
colormap jet
```

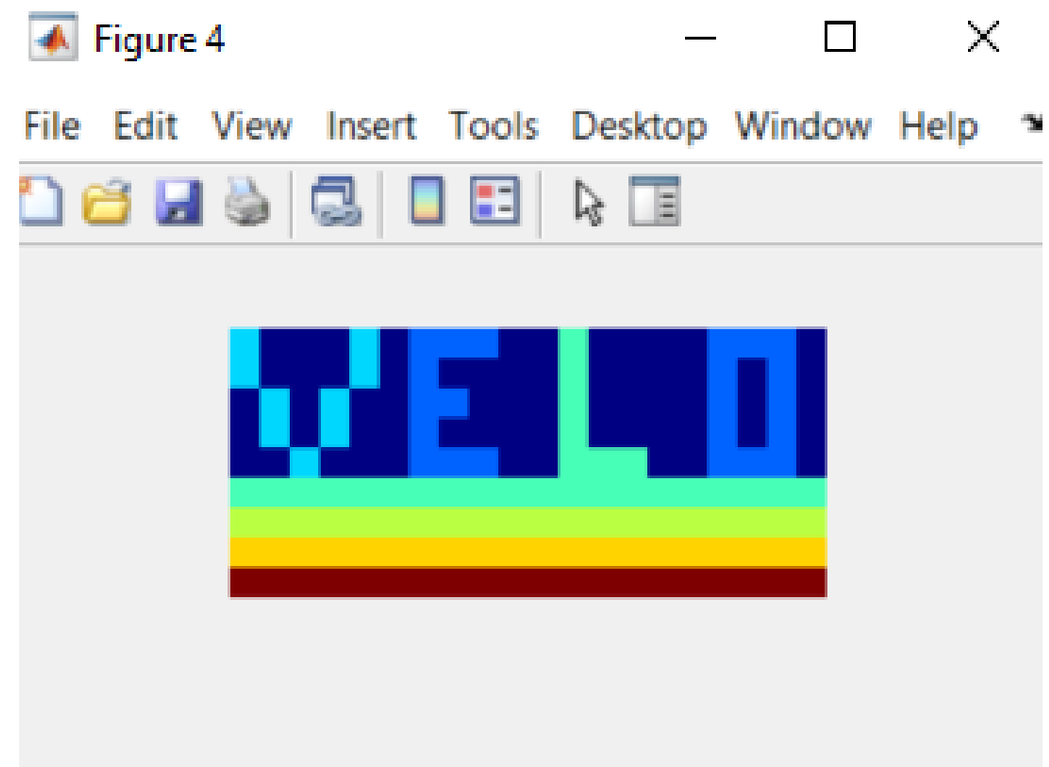
# Lets do fun with circshift, rolling the writings: Design an animation for film credits




```
%%
 figure(4)
imshow(k,[], 'initialmagnification',1200)
 colormap jet
%%
k2=k
for i=1:20
    k2=circshift(k2,-1)
    figure(4)
    imshow(k2,[],'initialmagnification',1200)
    colormap jet
    pause(0.2)
    disp(i)
end
```



# Circle data in rows and columns

# Sorting rows

examscores =

| | |
|---|---|
| 94 | 60 |
| 65 | 88 |
| 80 | 82 |
| 100 | 77 |
| 67 | 81 |
| 95 | 70 |
| 62 | 97 |
| 65 | 88 |
| 76 | 74 |
| 60 | 65 |

sortrows(examscores)

ans =

| | |
|---|---|
| 60 | 65 |
| 62 | 97 |
| 65 | 88 |
| 65 | 88 |
| 67 | 81 |
| 76 | 74 |
| 80 | 82 |
| 94 | 60 |
| 95 | 70 |
| 100 | 77 |

sortrows(examscores,2)

ans =

| | |
|---|---|
| 94 | 60 |
| 60 | 65 |
| 95 | 70 |
| 76 | 74 |
| 100 | 77 |
| 67 | 81 |
| 80 | 82 |
| 65 | 88 |
| 65 | 88 |
| 62 | 97 |

# Data sorting

sort the elements of each column in a particular order.

examscores =

98 76 71 83 70 85 89 83 71 63

sort(x,'ascend')

ans =

63 70 71 71 76 83 83 85 89 98

sort(x,'descend')

ans =

98 89 85 83 83 76 71 71 70 63

**Reshaping a Matrix**

The number of rows and columns in a matrix can be changed provided the total number of elements remains the same.

```
a=randi([1,10],3,3)
 b=reshape(a,9,1)
```

| | | |
|---|---|---|
| 2 | 2 | 8 |
| 7 | 3 | 2 |
| 1 | 8 | 3 |

| |
|---|
| 2 |
| 7 |
| 1 |
| 2 |
| 3 |
| 8 |
| 8 |
| 2 |
| 3 |

```
b=reshape(a,1,9)
```

| | | |
|---|---|---|
| 8 | 5 | 7 |
| 10 | 8 | 10 |
| 9 | 9 | 10 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 8 | 10 | 9 | 5 | 8 | 9 | 7 | 10 | 10 |

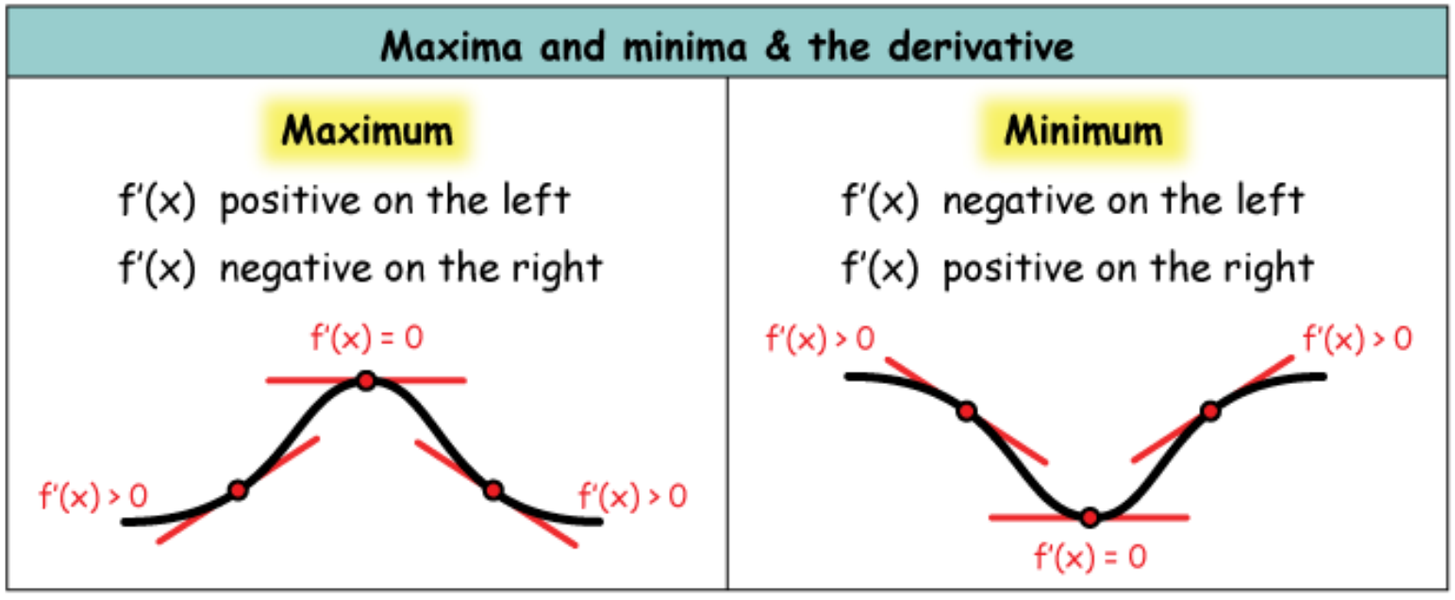
# Finding anomaly in the data. This is an harder problem for teaching the computer to find the outliers.

a=[5,8,3,6,7,200, 10, 12, 295, 34, 250]

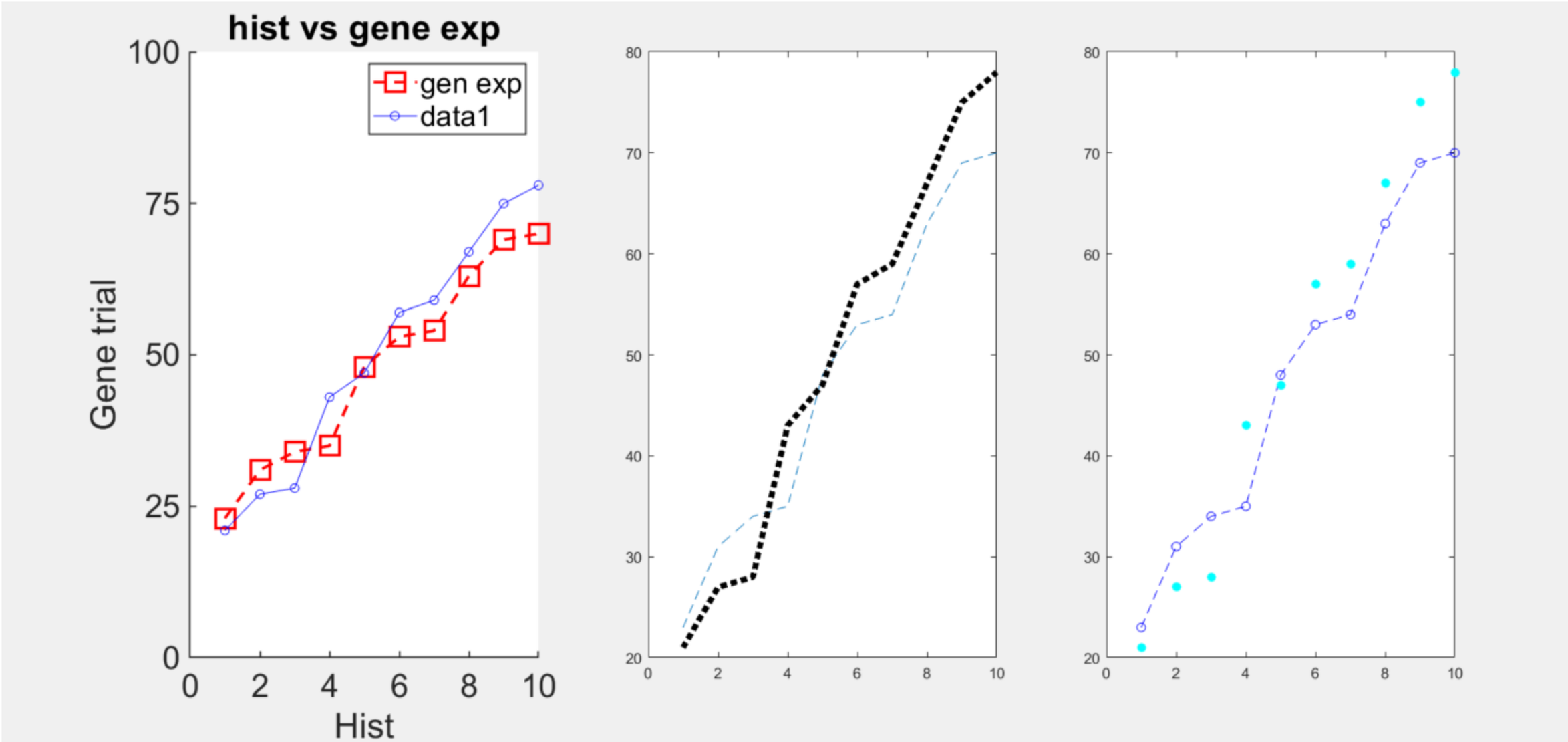
b = 3 5 6 7 8 10 12 34 200 250 295

You should take the first derivative of the function. How can you take the first derivative with matlab (circshift)?

Protocol: 1. sort the data
2. take the first derivative
3. Find the max and its index number
4. Use the index number and find the subdata?

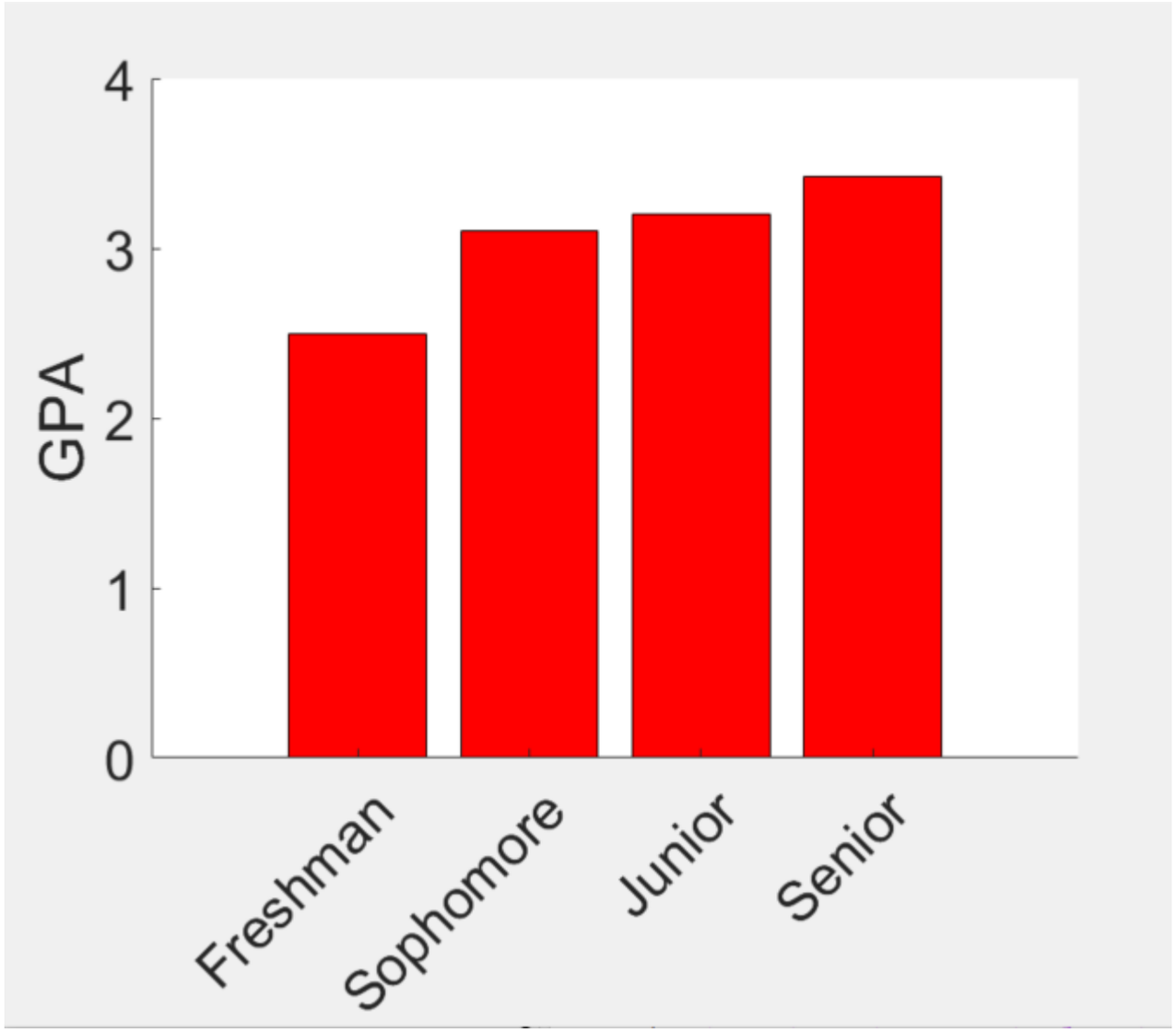



```
% finding the outliers numbers in the data sets
%
a=[5,8,3,6,7,200, 10, 12, 295, 34, 250]
b=sort(a)
c=circshift(b,-1)
% derivative
d=c-b
k=max(d)
%%
k1=find(d==k)
asub=b(1,1:k1)
```



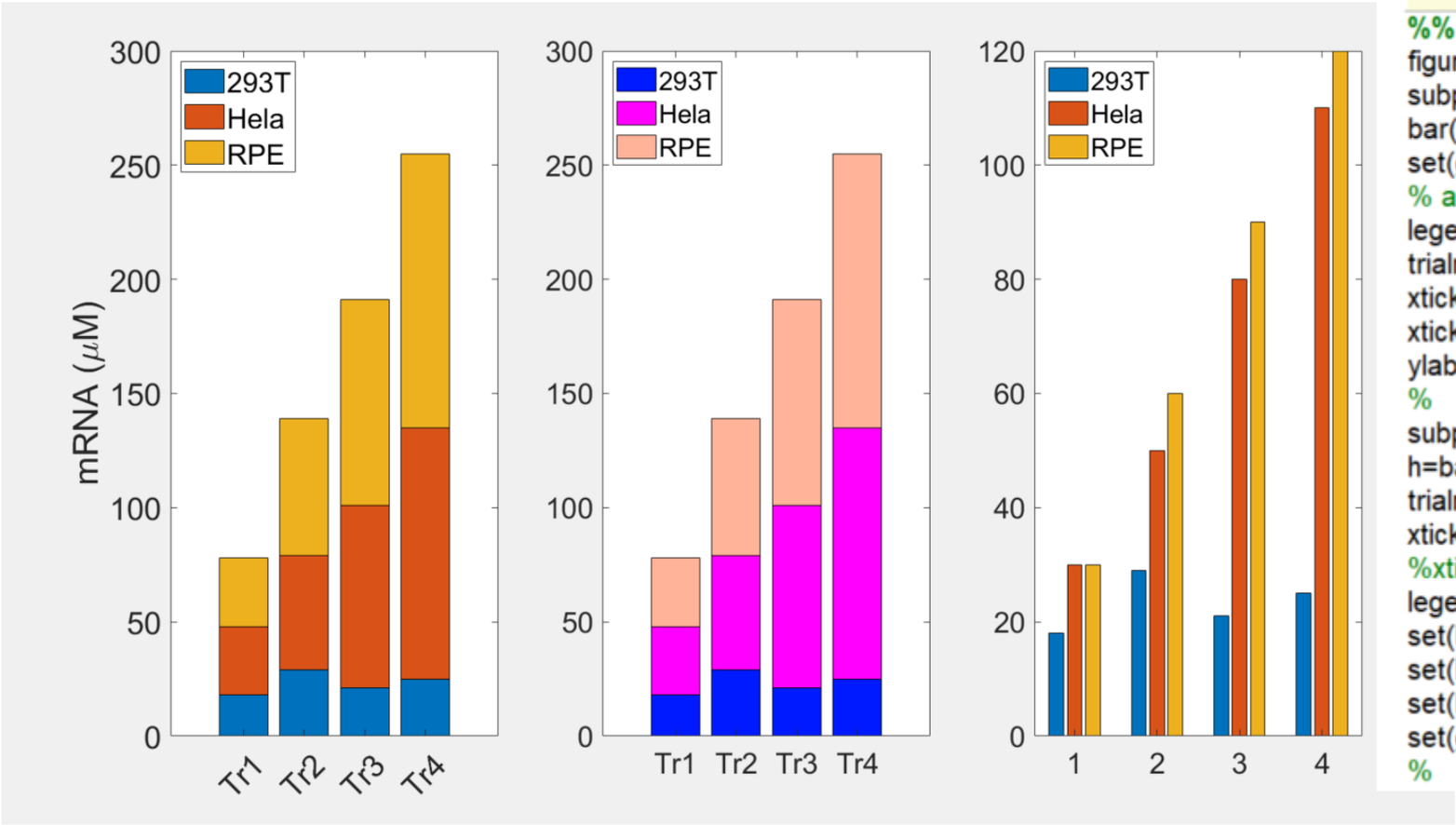

```
saveas(gcf,'firstfigure.png')
saveas(figure(1),'firstfigure.jpg')
saveas(figure(1),'firstfigure.tif')
```

# Bar plots




```
GPA = [2.5,3.1,3.2,3.42]
year1={'Freshman','Sophomore','Junior','Senior'}
figure(1)
bar(GPA,'r')
xticklabels(year1)
xtickangle(45)
ylabel('GPA')
set(gca,'Fontsize',24)
box off

xticklabels(year1)

ylabel('GPA')

saveas(figure(1),'studentsgradesaverage.pdf')
```

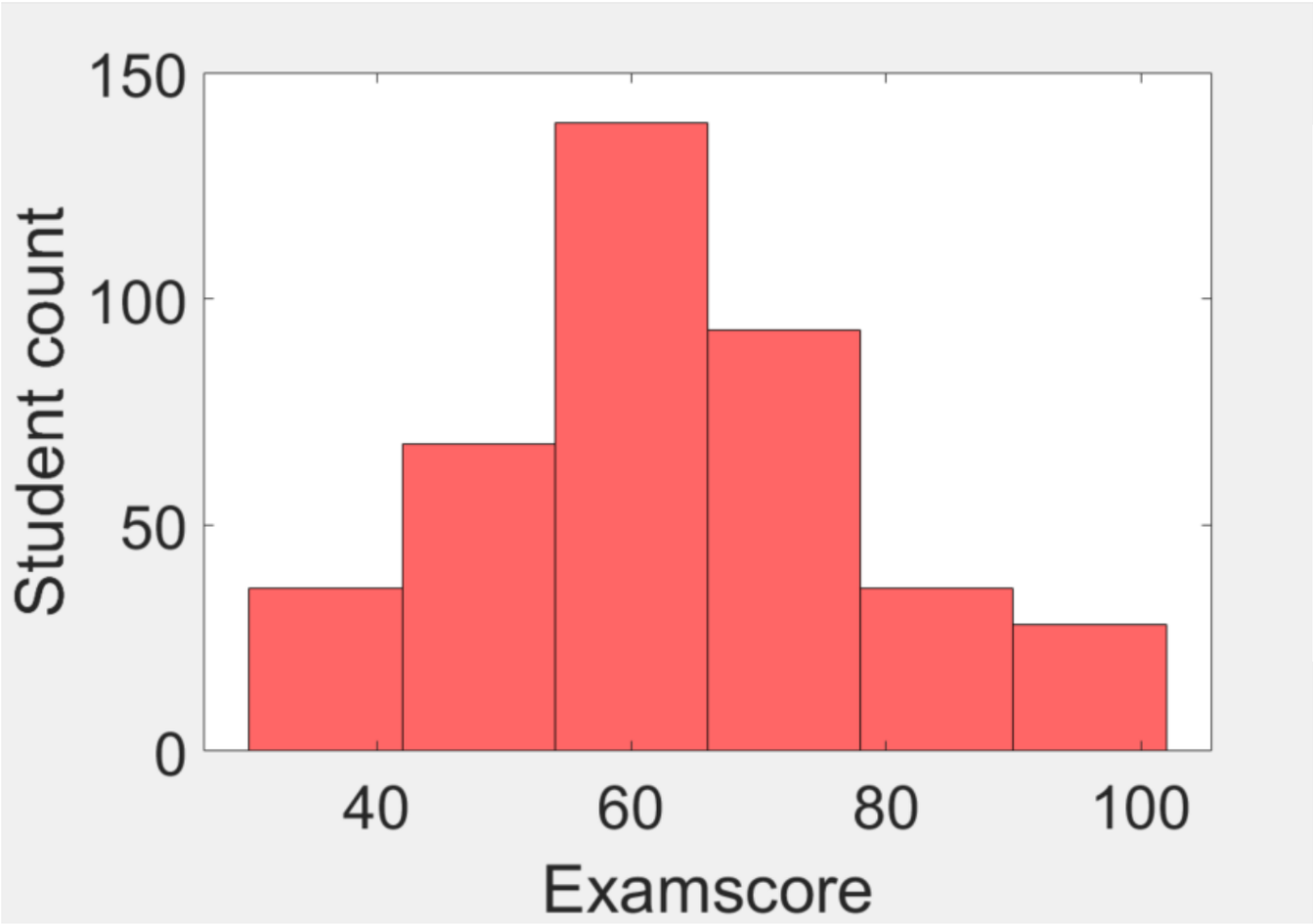
# Stacking or grouping bars




```
mRNA = [18 30 30; 29 50 60; 21 80 90; 25 110 120];

%%
figure(6)
subplot(1,3,1)
bar(mRNA,'stacked')
set(gca,'Fontsize',24)
% adding a legend
legend('293T','Hela','RPE','Location','Northwest')
trialname={'Tr1','Tr2','Tr3','Tr4'}
xticklabels(trialname)
xtickangle(45)
ylabel('mRNA (\muM)')
%
subplot(1,3,2)
h=bar(mRNA,'stacked')
trialname={'Tr1','Tr2','Tr3','Tr4'}
xticklabels(trialname)
%xtickangle(45)
legend('Location','Northwest');
set(h(1),'DisplayName','293T','Facecolor',[0 0.1 1])
set(h(2),'DisplayName','Hela','Facecolor',[1 0 1])
set(h(3),'DisplayName','RPE','Facecolor',[1 0.7 0.6])
set(gca,'Fontsize',22)
%
```

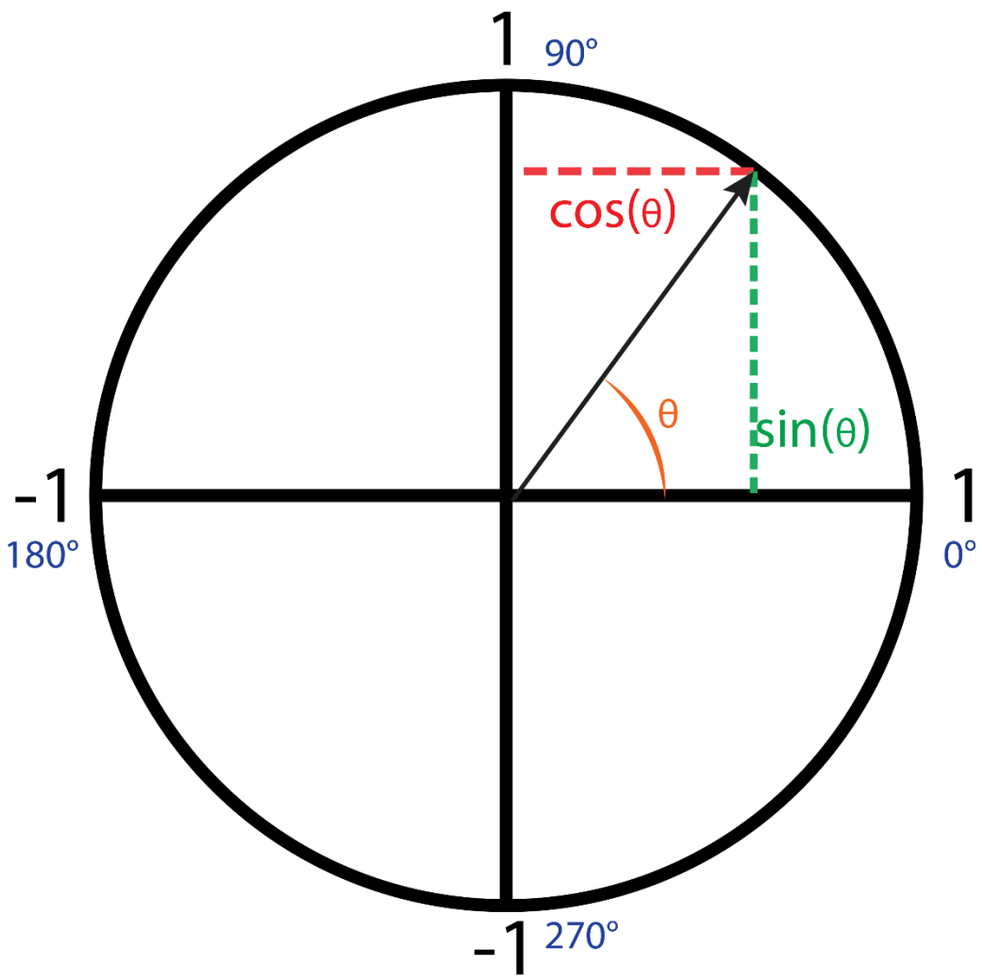
```
%
subplot(1,3,3)
bar(mRNA ,'grouped')
set(gca,'Fontsize',22)
labels = {'293T','Hela','RPE'};
legend(labels,'Location','NorthWest');
```

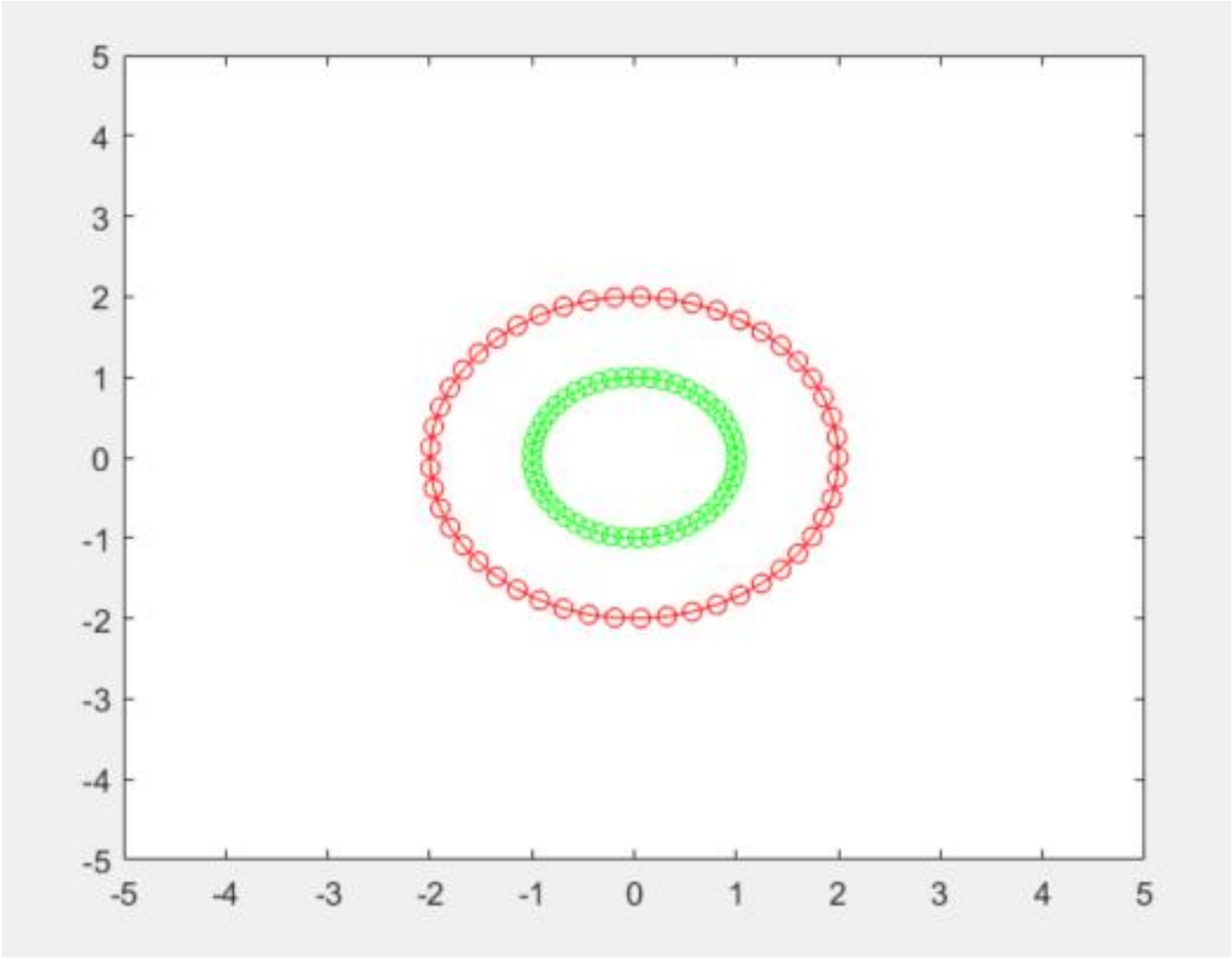
# Histogram plot



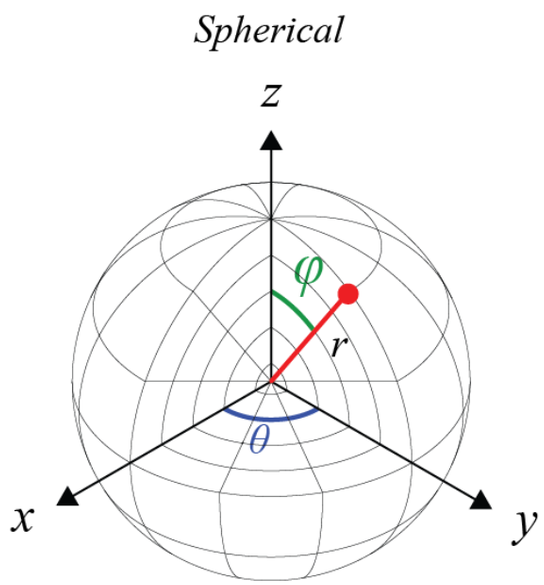

```
examscores(1:200,2)= randi([50,70],200,1)
%%
[counts,edges]=histcounts(examscores,6)
%%
figure(6)
h=histogram(examscores(:,1:2),6)
set(gca,'Fontsize',30)
xlabel('Examscore')
ylabel('Student count')
set(h,'Facecolor',[1 0 0])
```

# Other graphics: draw a circle with matlab



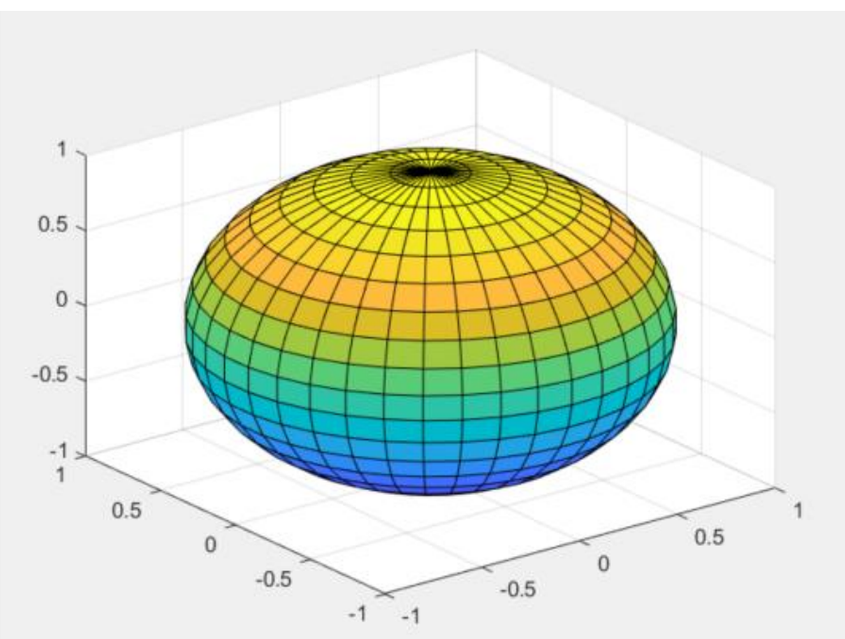



# Drawing a sphere

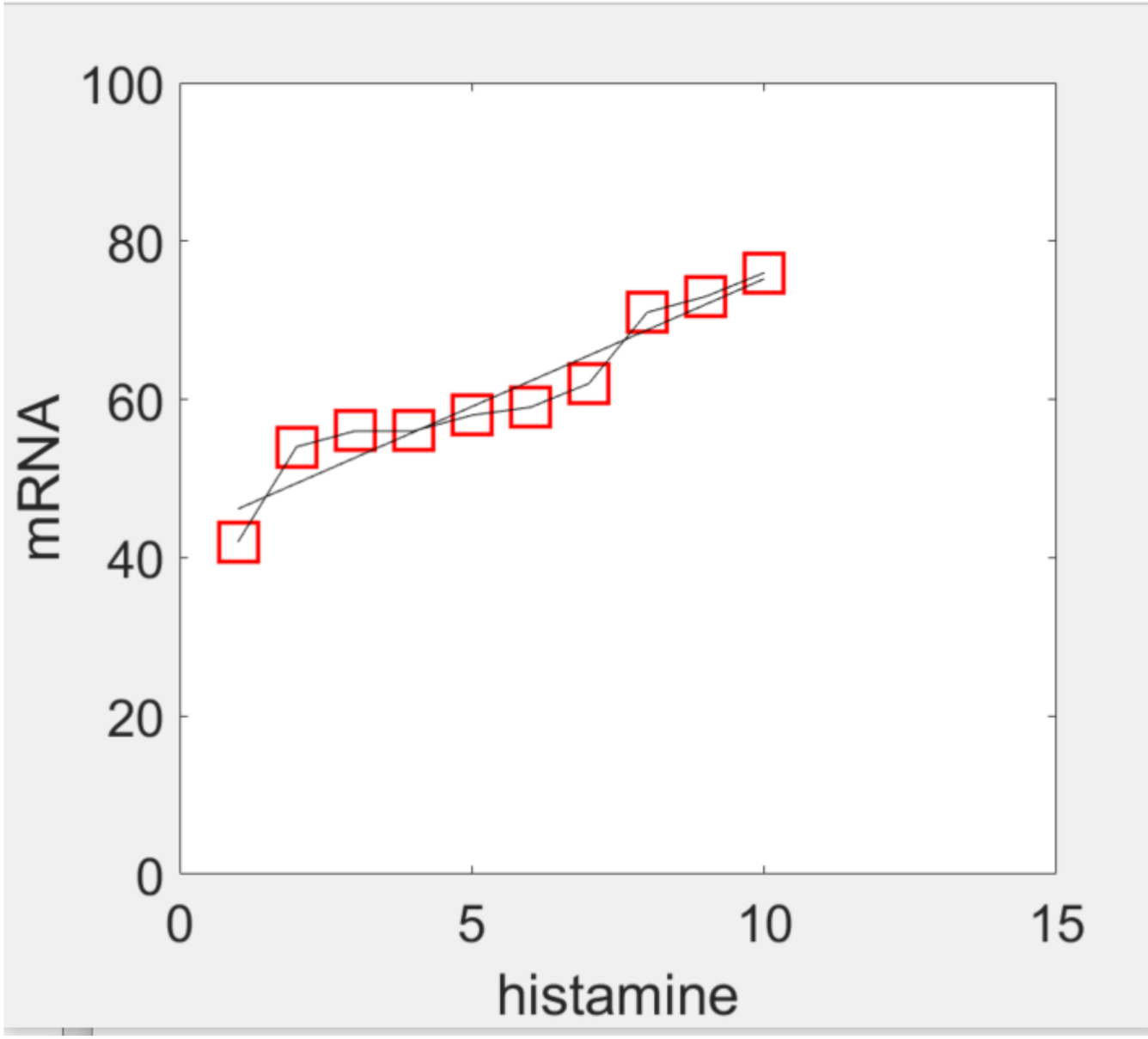



$$(r,\theta,\varphi) \longrightarrow \begin{aligned} x &= r \sin\varphi \cos\theta \\ y &= r \sin\varphi \sin\theta \\ z &= r \cos\varphi \end{aligned}$$

```
%%
N = 20;
thetavec = linspace(0,pi,N);
phivec = linspace(0,2*pi,2*N);
[th, ph] = meshgrid(thetavec,phivec);
R = ones(size(th));
x = R.*sin(th).*cos(ph);
y = R.*sin(th).*sin(ph);
z = R/1.*cos(th);

figure(3)
surf(x,y,z);
```

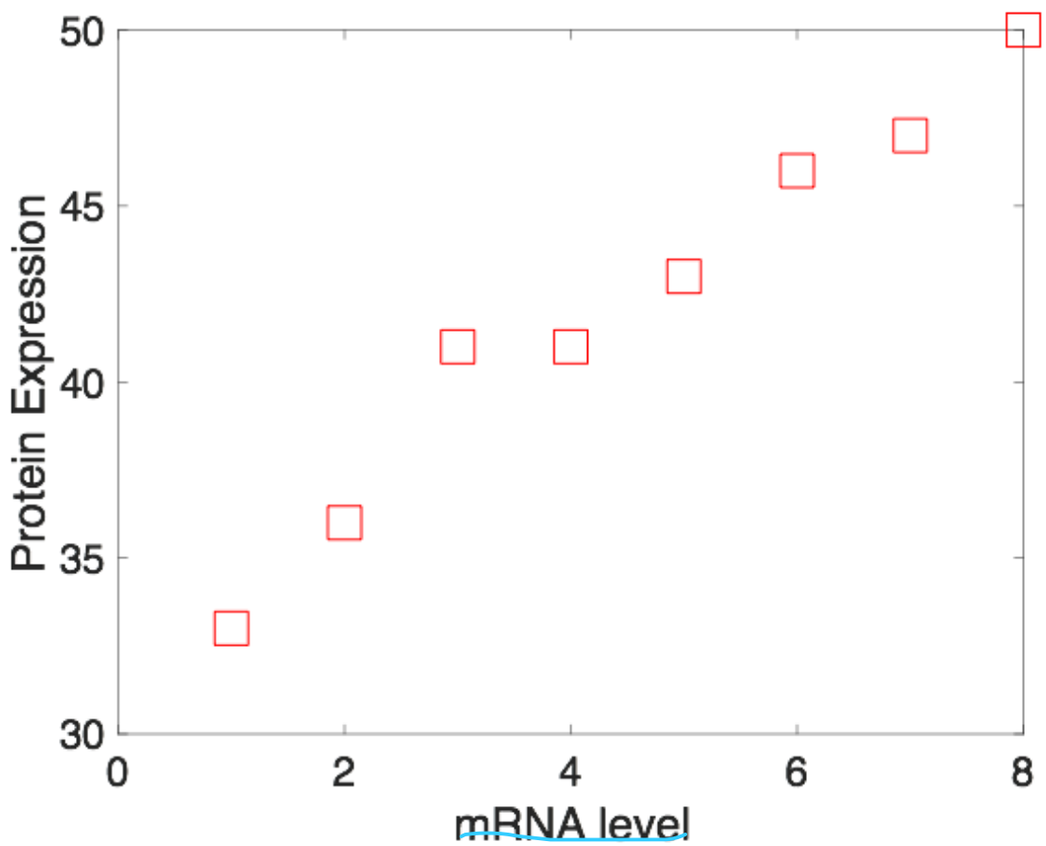


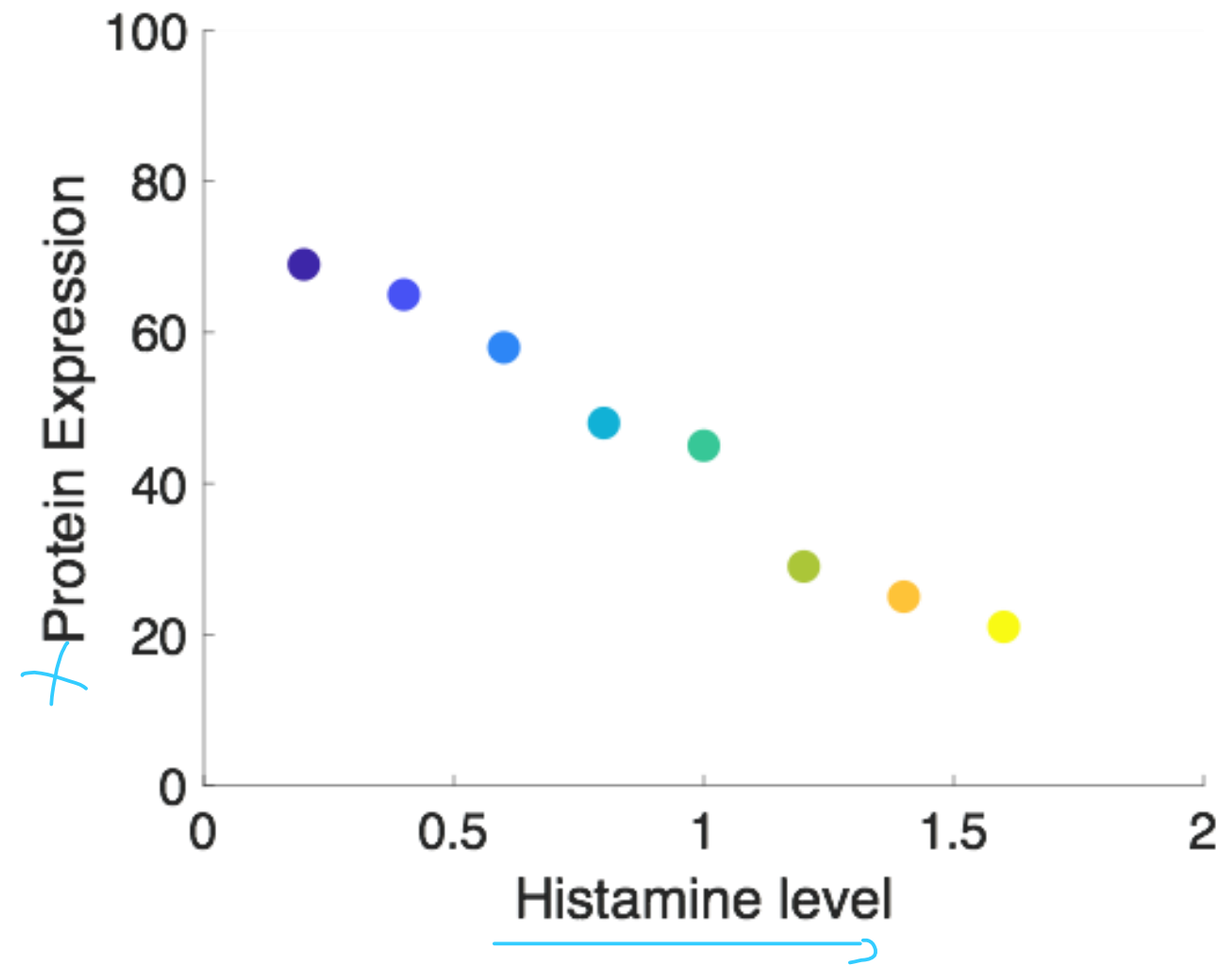
# Linear Regression



# Scatter plot

Shows the relation between two variables

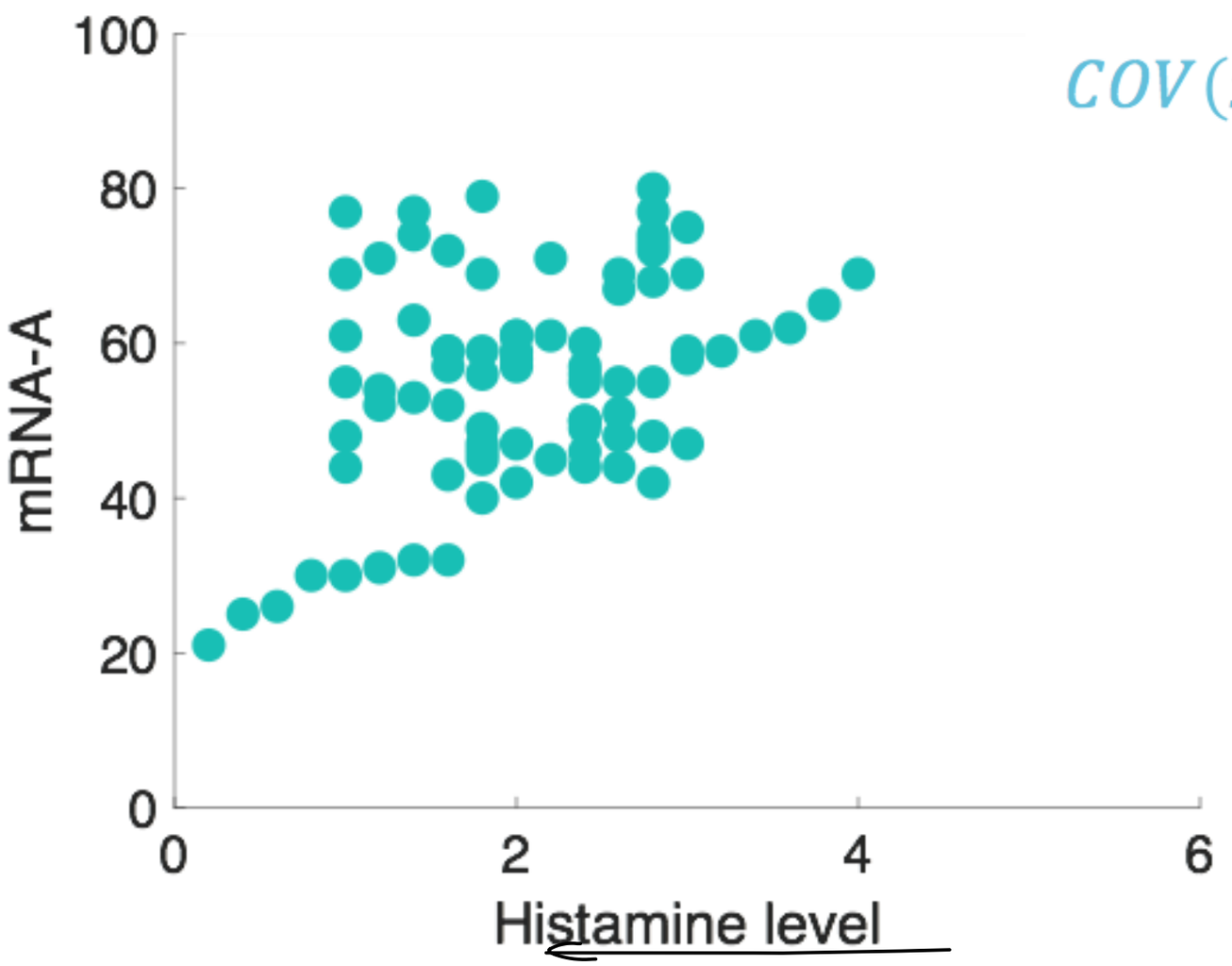



Can we quantitatively measure the strength of relationship between variables?

# Covariance

Does Y get larges (smalleR) as Y increase?




$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

mean of x= 2.0525
mean of y = 55.4125
Sx = 0.7916
Sy= 13.6537

n=80

Covariance > 0 if X and Y variables gets larger

Covariance < 0 if X and Y variables moves opposite direction

# Correlation (r)

- measures the direction and strength of relationship between two quantitative variable.
- The correlation *r* measures the direction and strength of the linear (straight line) association between two quantitative variables *x* and *y*.
- Although you can calculate a correlation for any scatterplot, *r* measures only linear relationships.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i-\bar{x})^2)(\sum_{i=1}^{n}(y_i-\bar{y})^2)}} \\ &= \frac{1}{n-1}\cdot\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} \end{aligned}$$

close to n-1 if x and y have

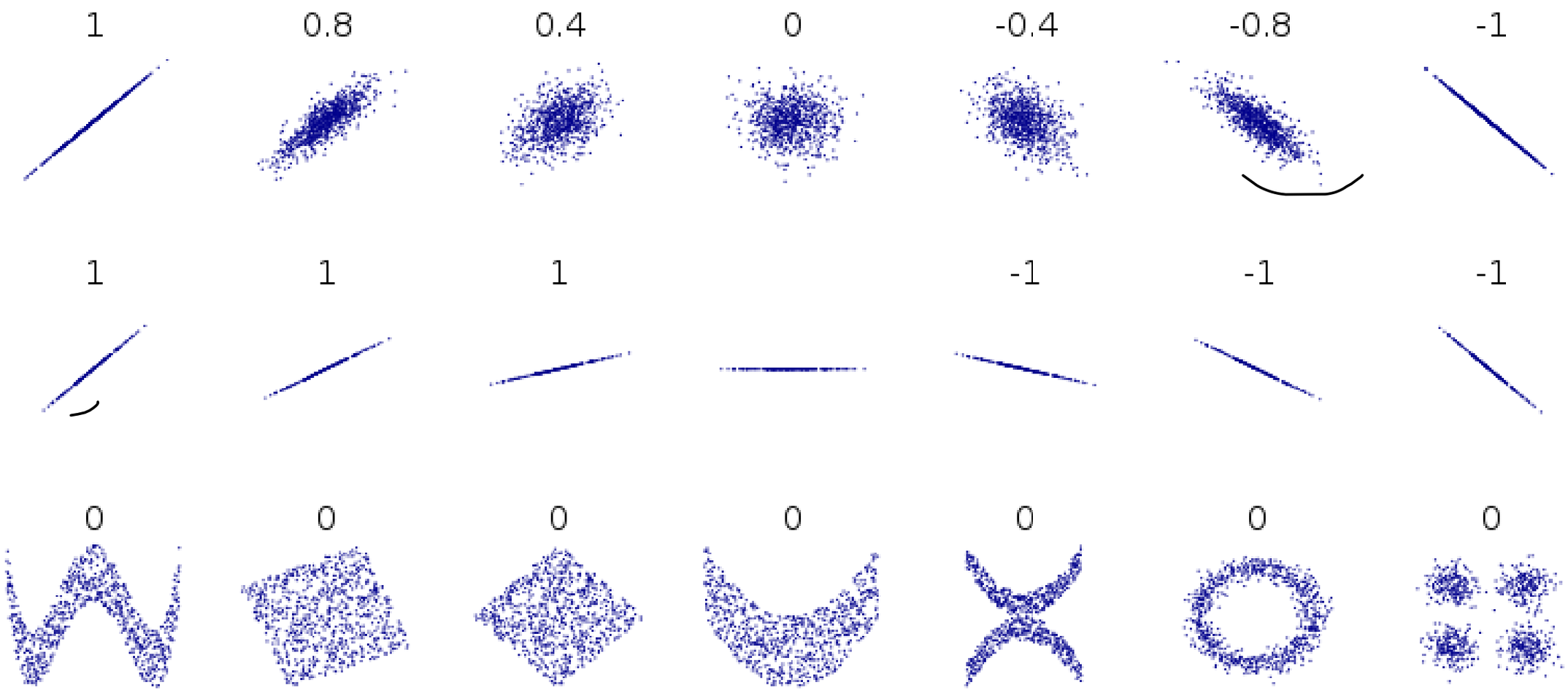
$\bar{x}$ = the sample mean of $x_1,\ldots,x_n$,

$\bar{y}$ = the sample mean of $y_1,\ldots,y_n$,

$s_x$ = the standard deviation of $x_1, \ldots, x_n$,

$s_y$ =thestandarddeviationof$y_1,\ldots,y_n$.

# Correlation sets




Remember that correlation coefficient is an indicator of the strength of a *linear* relationship between two variables, but its value generally does not completely characterize their relationship

# Summary of Correlation between two variables

- $-1 \leq r \leq 1$ always
- $r = 1$ when all the points $(x_i, y_i)$ lie on a line with positive slope
- $r = -1$ when all the points $(x_i, y_i)$ lie on a line with negative slope
- When $r = 0$, then there is no positive or negative linear association between the two variables (though the two variables may have a non-linear relationship).

# Fitlm and polyfit functions

b = fitlm(hist',genetrial')

```
New to MATLAB? See resources for Getting Started.
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    _______    ______    __________

    (Intercept)     42.933      2.1767    19.724     4.544e-08
    x1              3.2303     0.35081    9.2082    1.5659e-05


Number of observations: 10, Error degrees of freedom: 8
Root Mean Squared Error: 3.19
R-squared: 0.914,  Adjusted R-Squared: 0.903
F-statistic vs. constant model: 84.8, p-value = 1.57e-05
fx >>
```

[co,S]=polyfit(hist,genetrial,1)

```
co =

    3.2303   42.9333


S =

  struct with fields:

        R: [2×2 double]
       df: 8
    normr: 9.0124
```