Does Whiskey Cure Diabetes?

CHIDATA

CHIDATA

Hypothetical Study

A study once found that patrons at a liquor store who exclusively purchased pricier whiskey (> $40 a bottle) were less likely to have Type II diabetes.

What can we conclude?

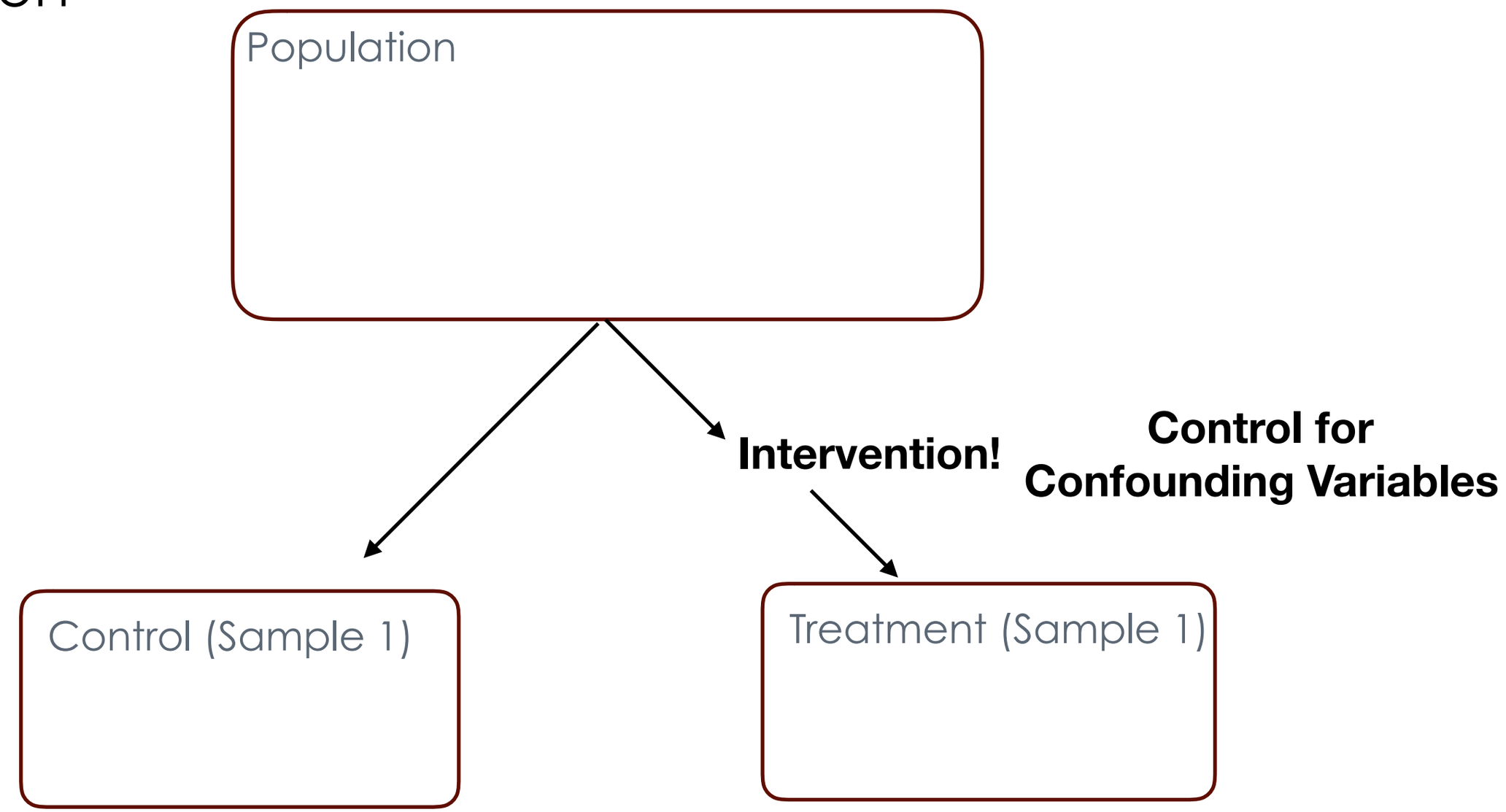
(a) Expensive whiskey mitigates the effects of diabetes

(b) Cheap whiskey exacerbates the effects of diabetes

(c) Expensive whiskey drinkers are probably wealthier and have better overall diet and exercise.

Recap: Correlation v.s. Causation

(Last Time) Impossible to guess a causal link without more information



Suggests Causation…

What if we don't care about causality?

Patrons who exclusively purchased pricier whiskey (> $40 a bottle).

What can we recommend to a customer?

(a) Other expensive/luxury products (because they can afford it)

(b) Whiskey glasses (so they can drink the whiskey)

(c) Chocolate (because they are less likely to have diabetes)

Degrees of Predictive Value

Patrons who exclusively purchased pricier whiskey (> $40 a bottle).

Likelihood of following through with a purchase

More accurate

(a) Other expensive/luxury products (because they can afford it)

(b) Whiskey glasses (so they can drink the whiskey)

(c) Chocolate (because they are less likely to have diabetes)

CHIDATA

Degrees of Predictive Value

Two variables: X and Y

How well does X predict Y?

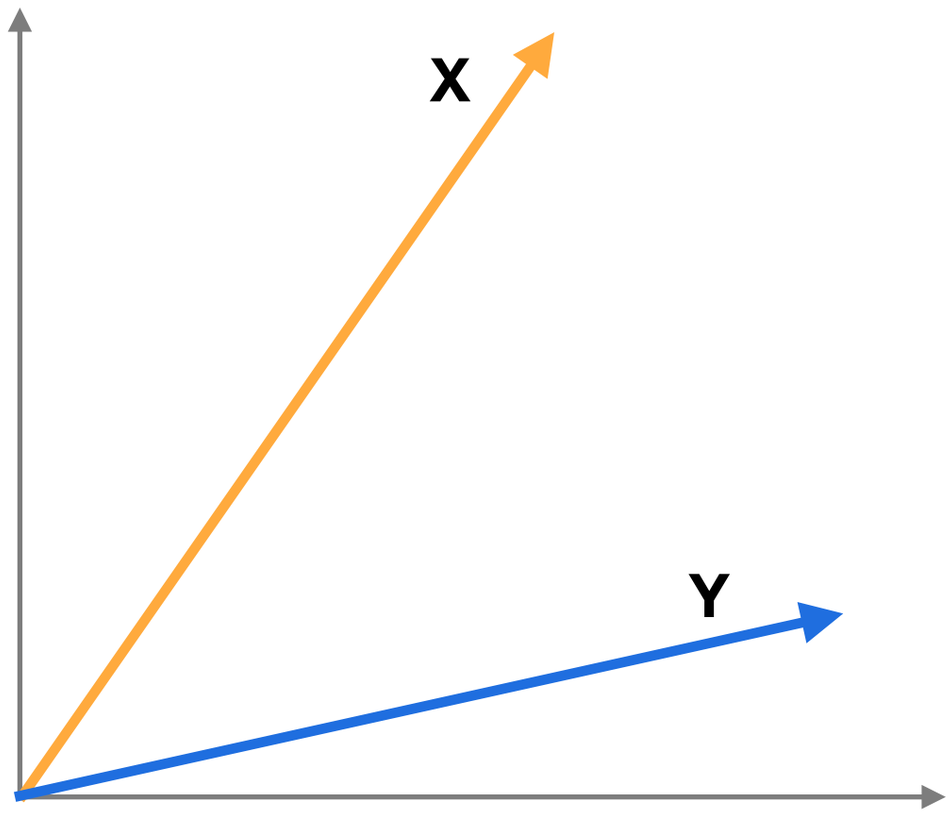
What does this mean and how do we measure predictive value with uncertain quantities?

CHIDATA

High School Geometry

Two **vectors**: X and Y

How well does X **align with** Y?



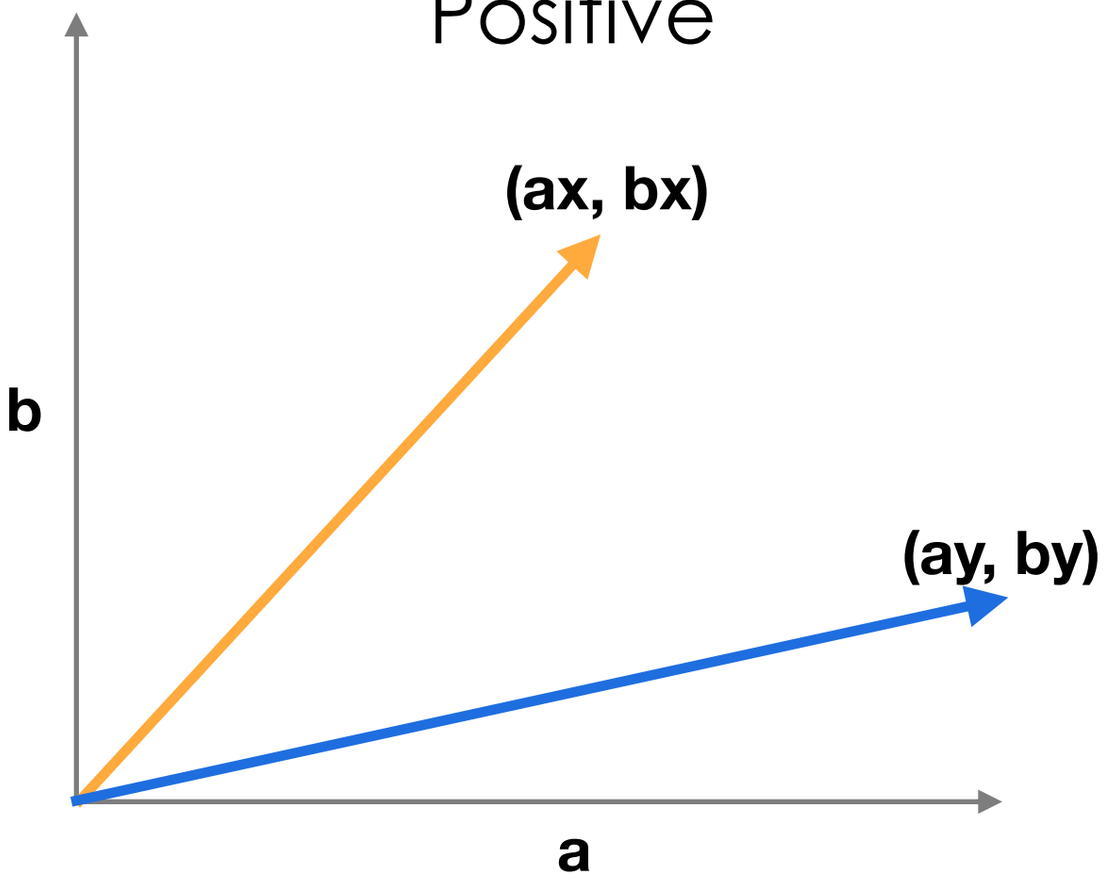
Are they going in the same direction? How closely?

Dot Products

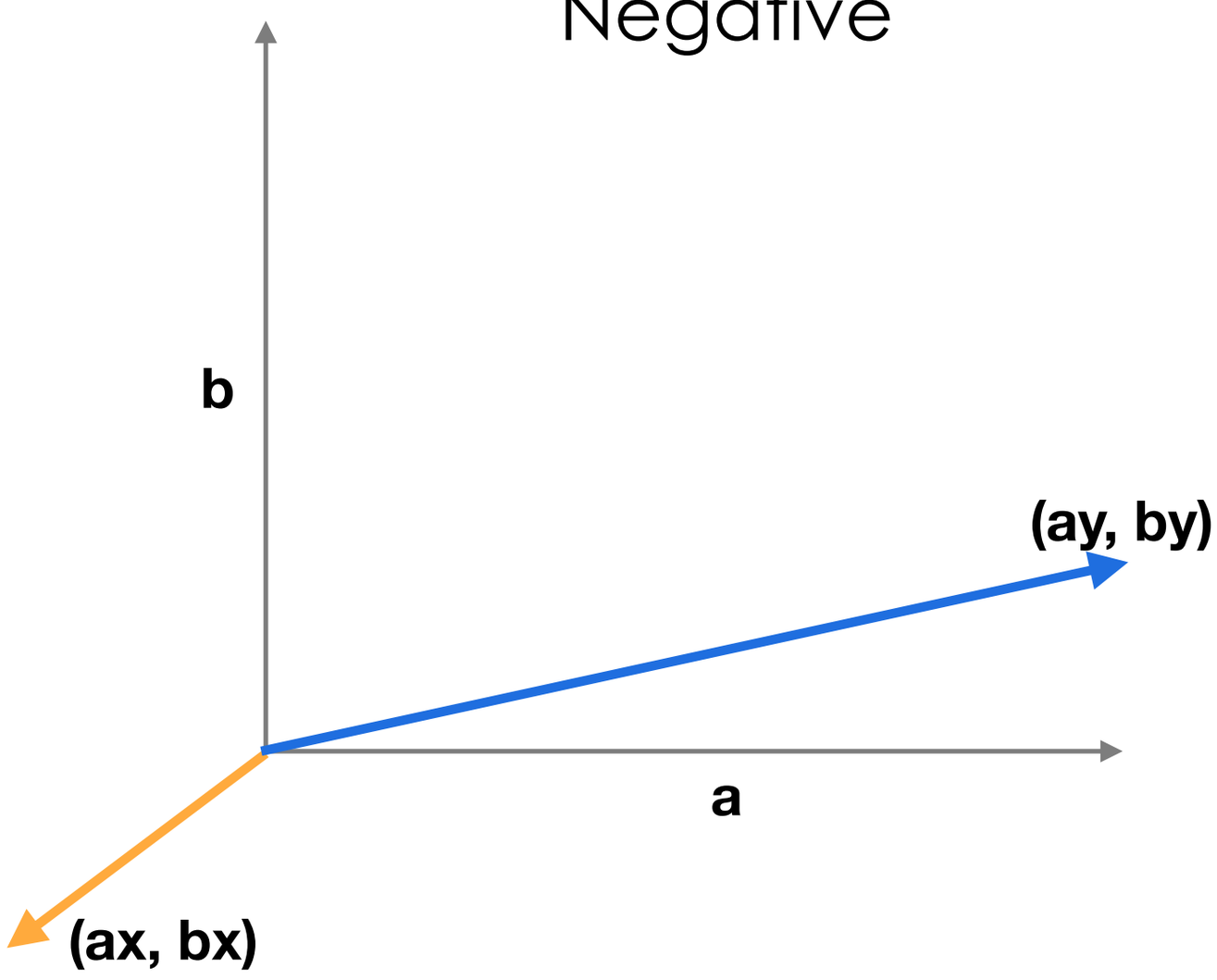
Two **vectors**: X and Y

$x \bullet y = ax*bx + ay*by$

Positive



Negative

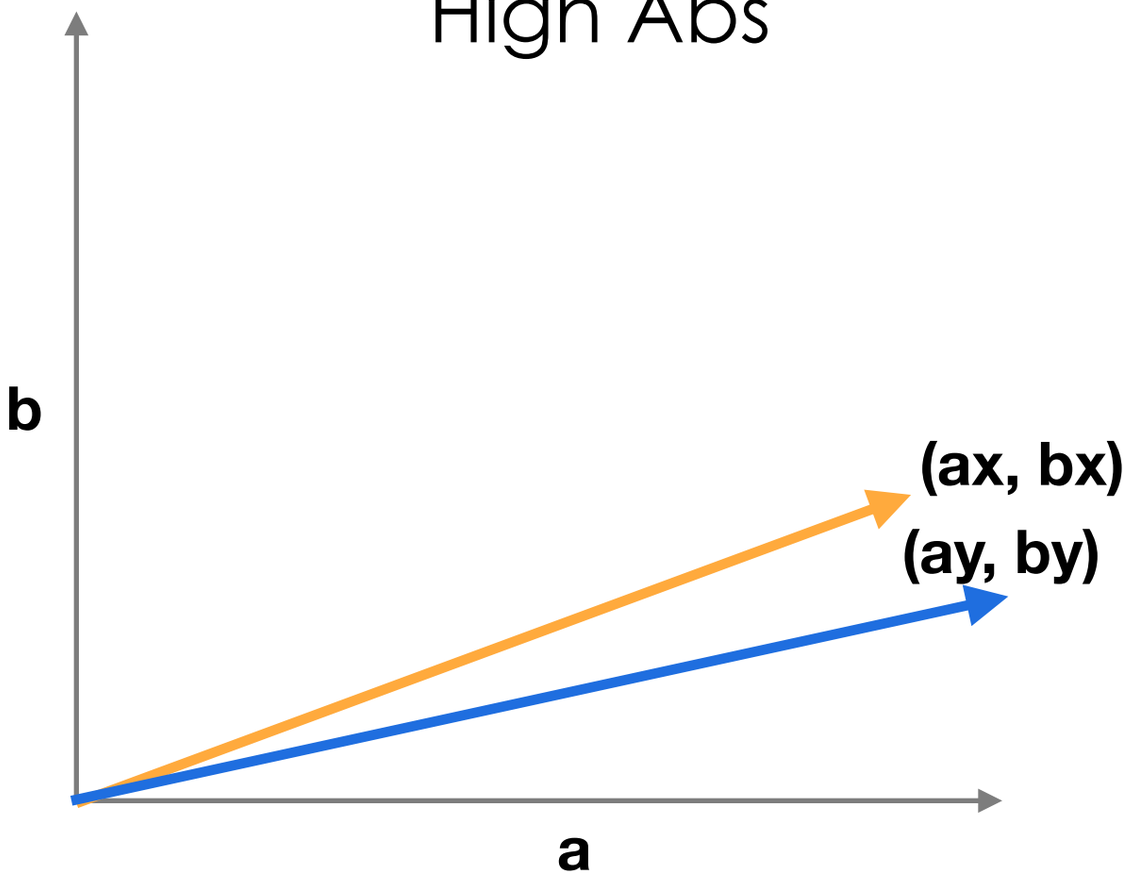


Absolute Value Depends on Angle

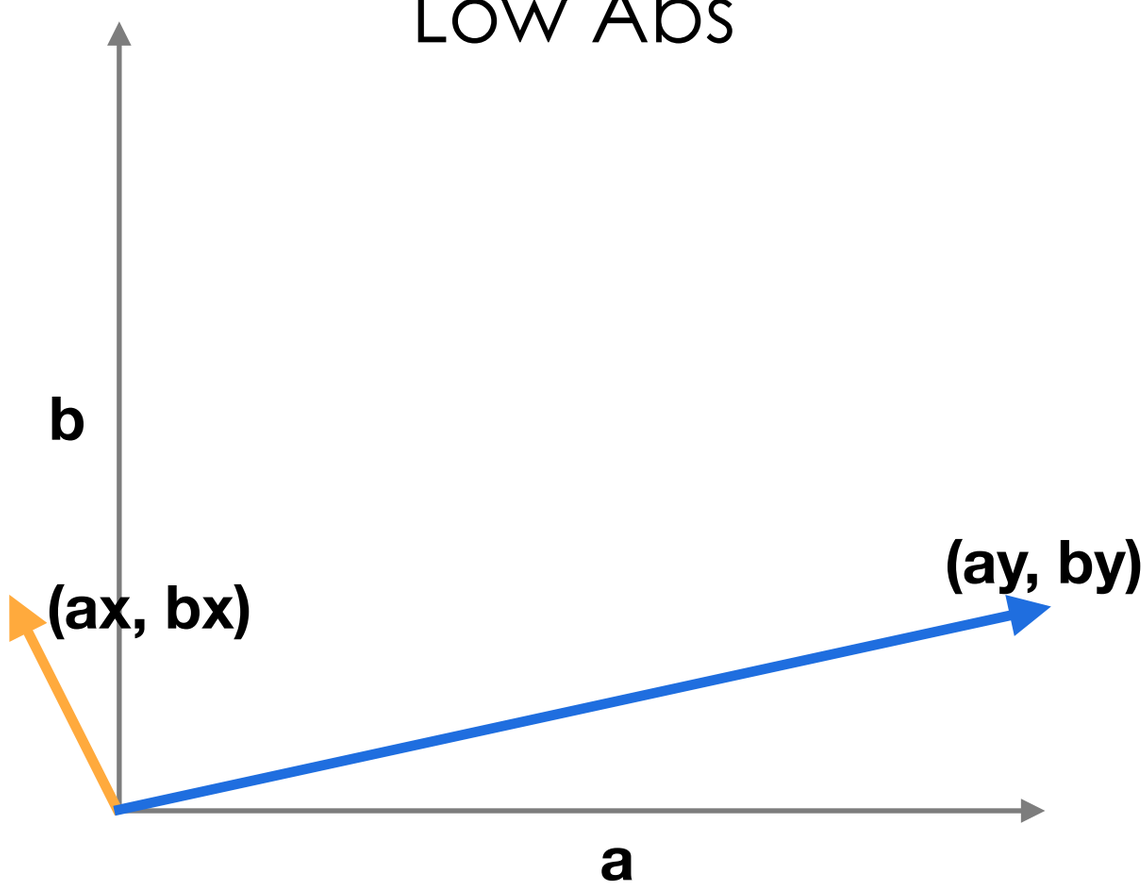
Two **vectors**: X and Y

$x \bullet y = ax*bx + ay*by$

High Abs



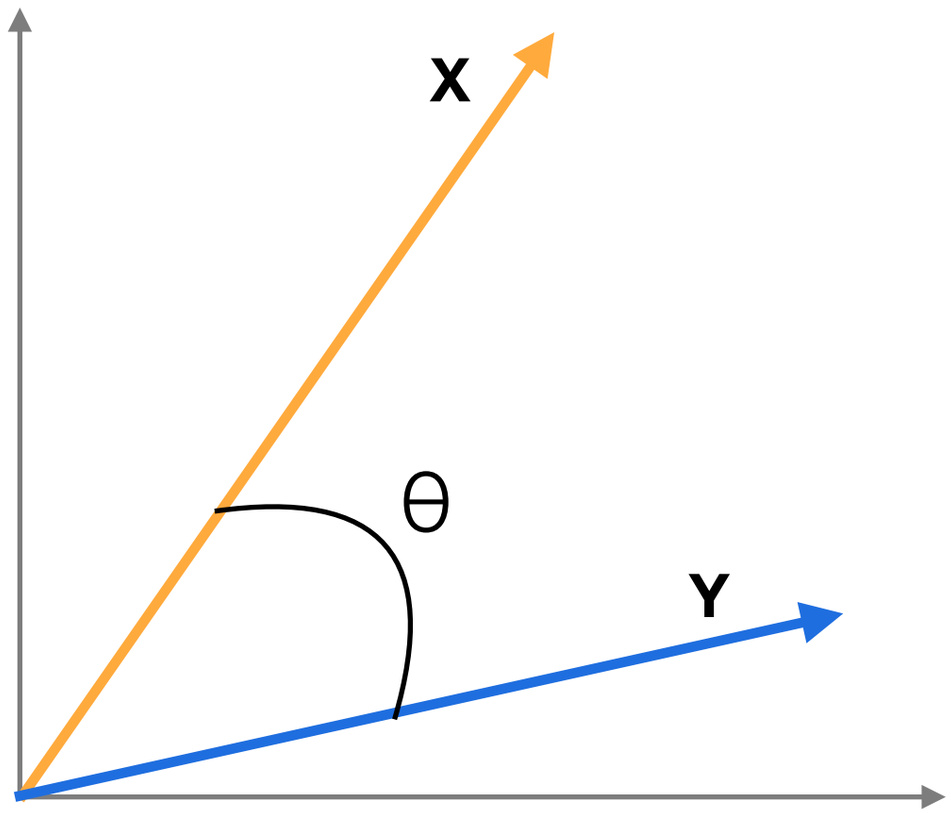
Low Abs



CHIDATA

High School Geometry

Two **vectors**: X and Y Calculate the Angle?

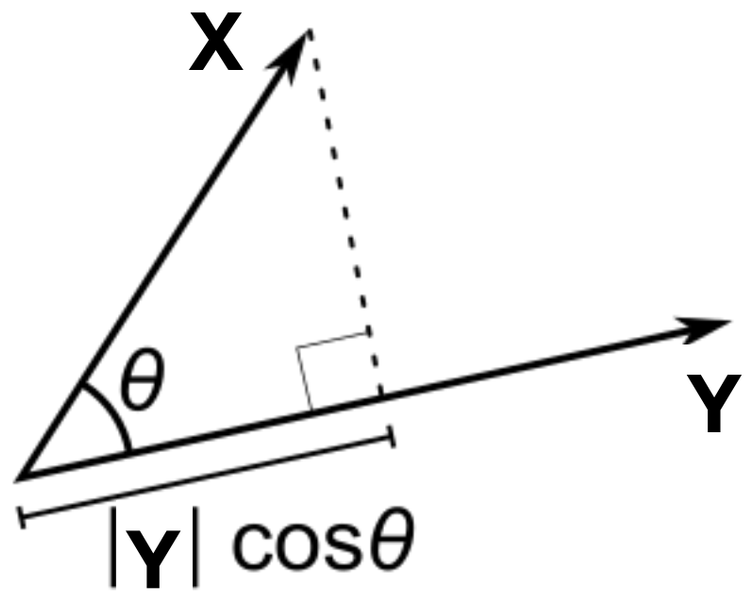


$$\frac{X \bullet Y}{|X|\,|Y|} = \cos(\theta)$$

High School Geometry

$x \bullet y$ “Distance” both vectors travel in the same direction

$$\frac{X \bullet Y}{|X|\,|Y|}$$ Angle between the two vectors



If I travel along X instead of Y, how far will I be from Y

Projection of X onto Y is linear

CHIDATA

Back to Stats

X and Y are measured quantities of individuals in the same population.

X = [1.5, 3.6, 2.3, ….., 10.7]

Y = [2.9, 7.6, 4.3, ….., 20.1]

Can think of them as big vectors

If I travel along X instead of Y, how far will I be from Y

If I use X to “predict” Y, how close will I be

Make This More Precise

X and Y are measured quantities of individuals in the same population.

X = [1.5, 3.6, 2.3, ….., 10.7]

X' = X - mean(X)

Y = [2.9, 7.6, 4.3, ….., 20.1]

Y' = Y - mean(Y)

X' • Y' Covariance

$$\frac{X' \bullet Y'}{|X'|\,|Y'|}$$ Correlation

Exactly the standard deviation!

$$S = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$$

Covariance and Correlation

The strength of "linear" dependence between two sets of observations.

$X' \bullet Y'$ Covariance

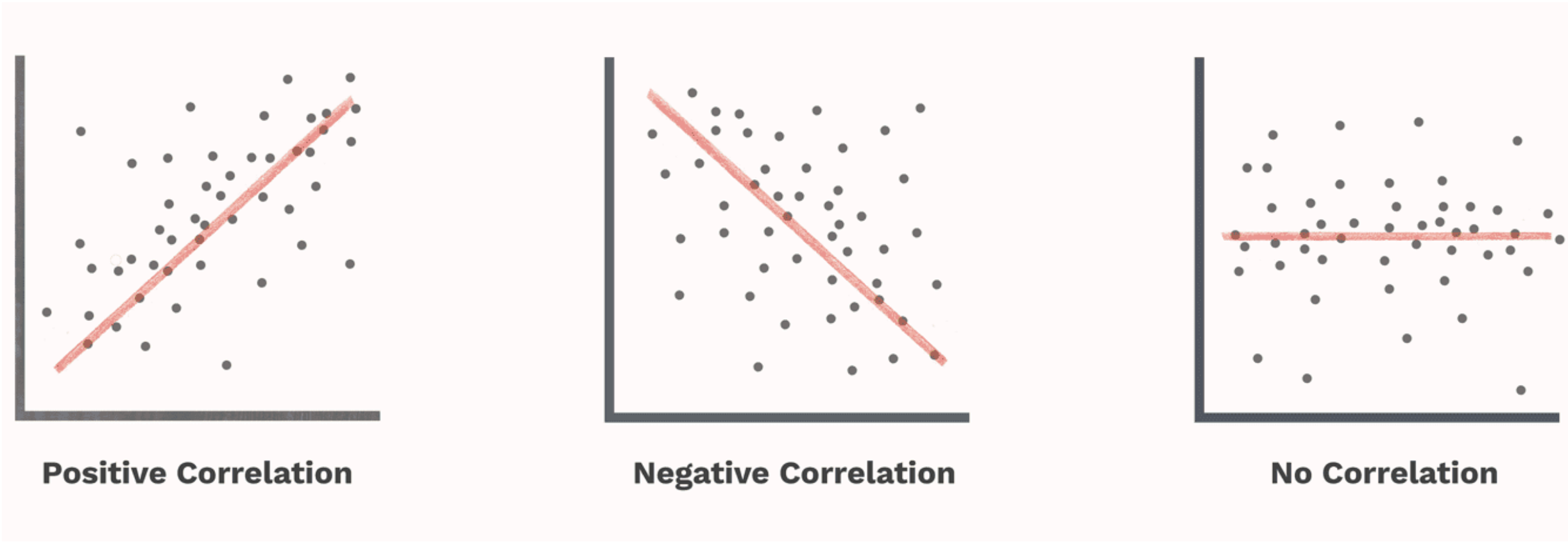
$$\mathrm{Cov}\,(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

$$\frac{X' \bullet Y'}{|X'|\,|Y'|}$$ Correlation "Pearson"

$$\frac{Cov\,(x, y)}{\sigma x * \sigma y}$$

CHIDATA

Covariance and Correlation



Recap: Significance

Control

Treatment

Measure how strongly correlated treatment is to a positive outcome.

CHIDATA

Example

Hypothesis A proposed relationship between two variables.

X = {treatment, control}
Y = Weight Loss

X = [0,1,0,…,1,1,0]
Y = [1.7,-2.3,7.6,…,4.6]

← **All of the data collected in the trial**

$$\frac{X' \bullet Y'}{|X'|\,|Y'|} \sim 1$$ Strong association

$$\frac{X' \bullet Y'}{|X'|\,|Y'|} \sim 0$$ Weak association

$$\frac{X' \bullet Y'}{|X'|\,|Y'|} < 0$$ Counterproductive

CHIDATA

Extend to Random Variables

The strength of "linear" dependence between two random variables X and Y.

$$cov(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

$$= \mathrm{E}\big[XY - X\mathrm{E}[Y] - Y\mathrm{E}[X] - \mathrm{E}[X][Y]\big]$$

$$= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] - \mathrm{E}[Y]\mathrm{E}[X] - \mathrm{E}[X][Y]$$

$$= \mathrm{E}[XY] - \mathrm{E}[X][Y]$$

CHIDATA

Properties

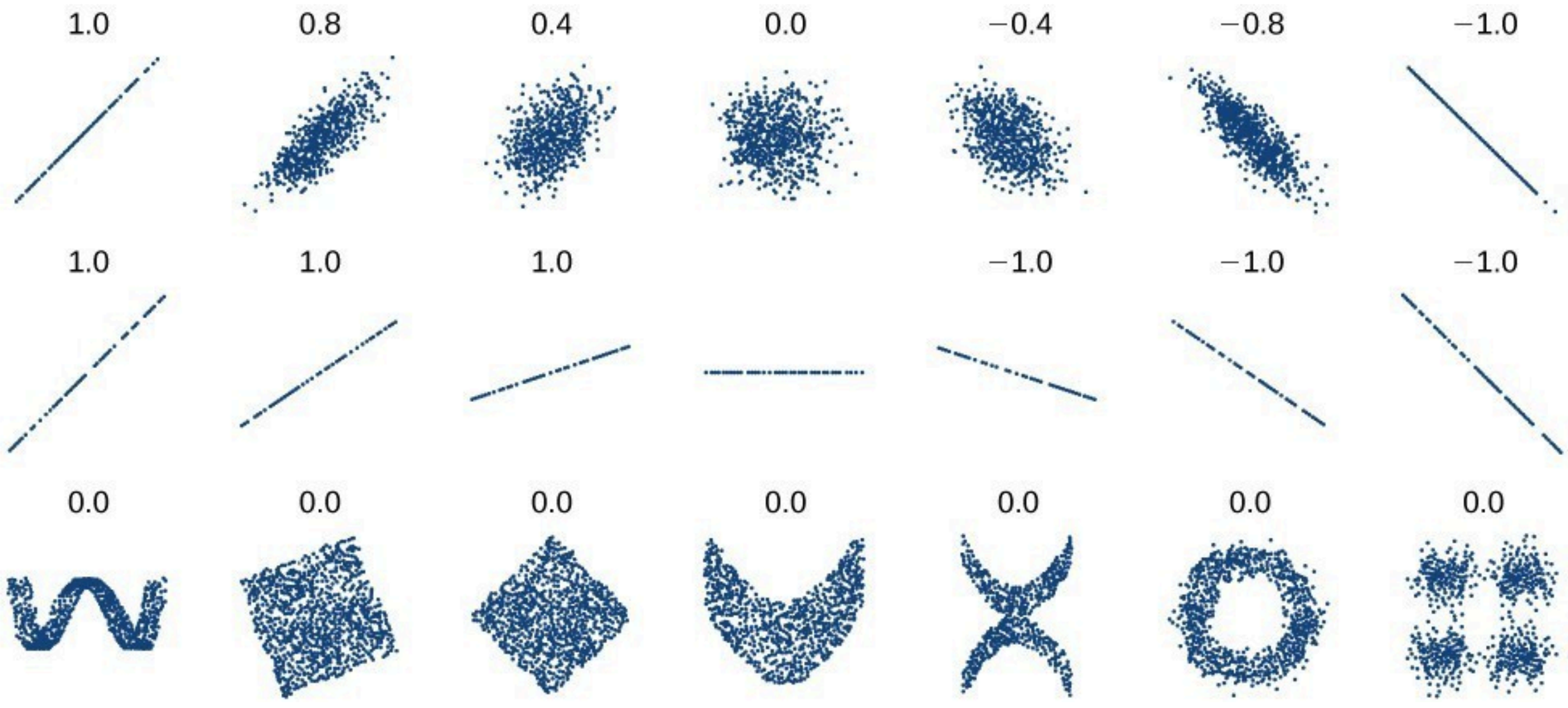
Cov(X,X) = Var(X)

Cov(X, constant) = 0

X, Y independent => Cov(X,Y) = 0 !!!Converse not true!!!

Reasons to Be Careful!

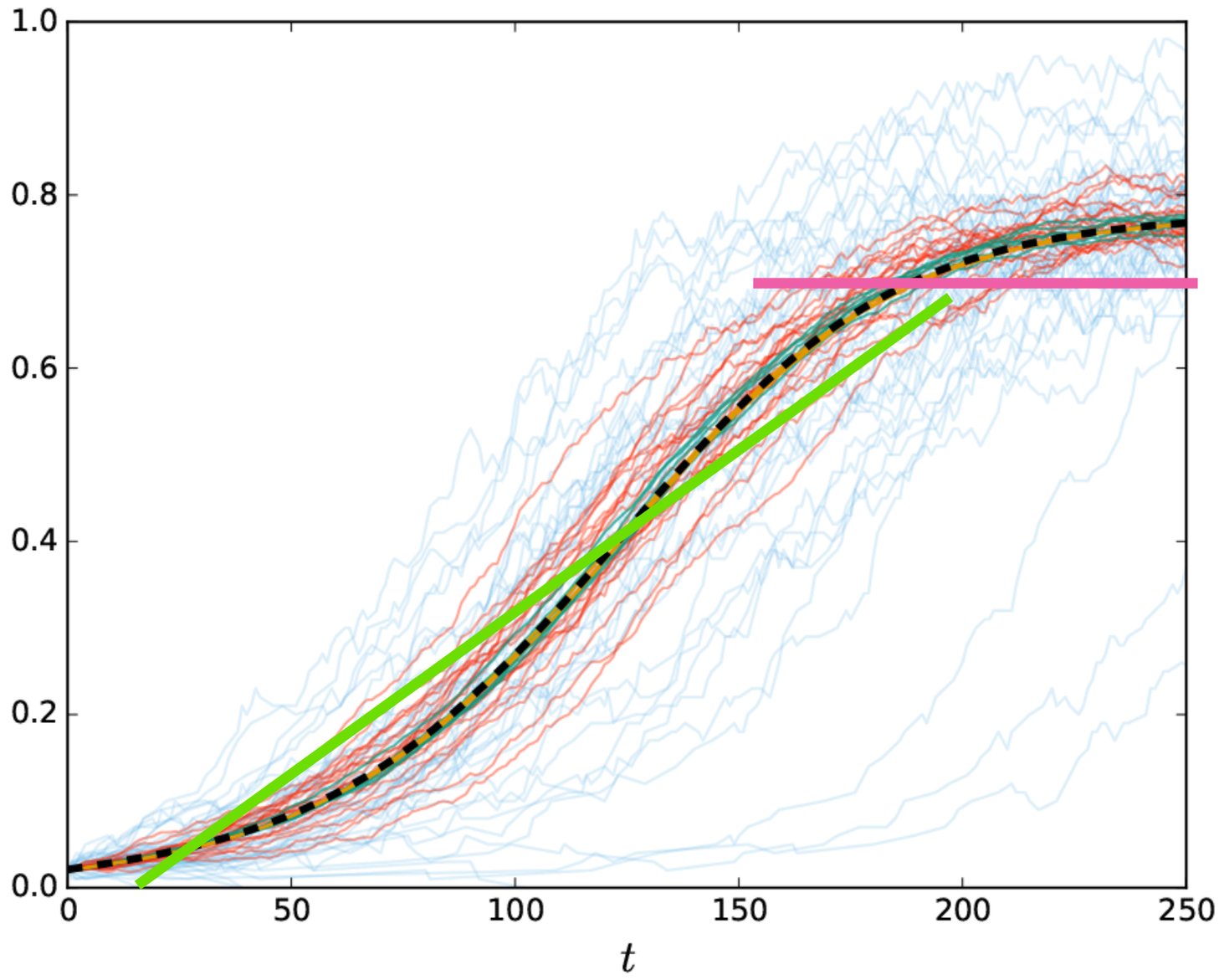
Covariance only considers “linear” predictive value



CHIDATA

Data Representation Matters

Saturation Effects



CHIDATA

Data Representation Matters

Cities with > 10 Fire Stations	Cities with <= 10 Fire Stations
86 fires	3 fires
101 fires	2 fires
54 fires	4 fires

X = {>10 FSs, <= 10 FSs}
Y = #Fires

CHIDATA

Data Representation Matters

Cities

86 fires,15 FSs

101 fires,32 FSs

54 fires, 16FSs

3 fires, 1FS

2 fires, 4FSs

4 fires, 3FSs

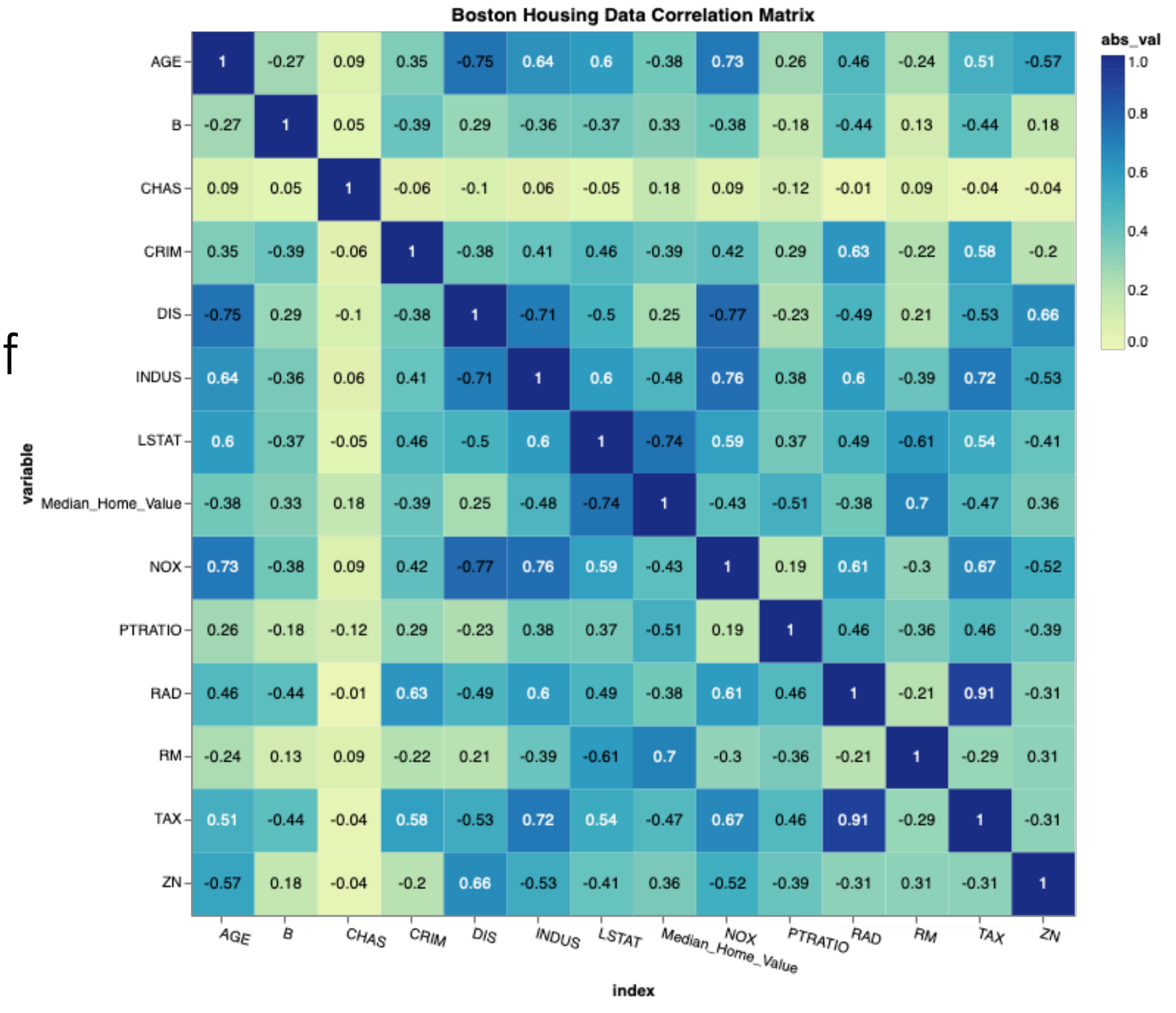
X = #FSs

Y = #Fires

CHIDATA

Multi-Variable Correlation

Compare all pairs of variables



CHIDATA

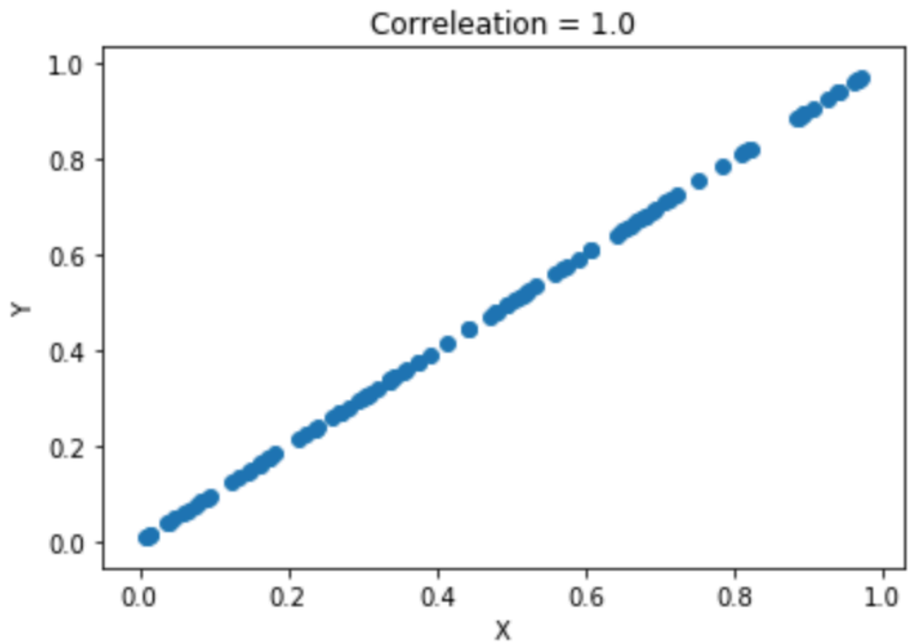
Examples

In [18]:
```
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats

x = np.random.rand(100)
y = x

corr = scipy.stats.pearsonr(x,y)[0]

plt.scatter(x,y)
plt.xlabel('X')
plt.ylabel('Y')
plt.title('Correleation = '+ str(corr))
plt.show()
```



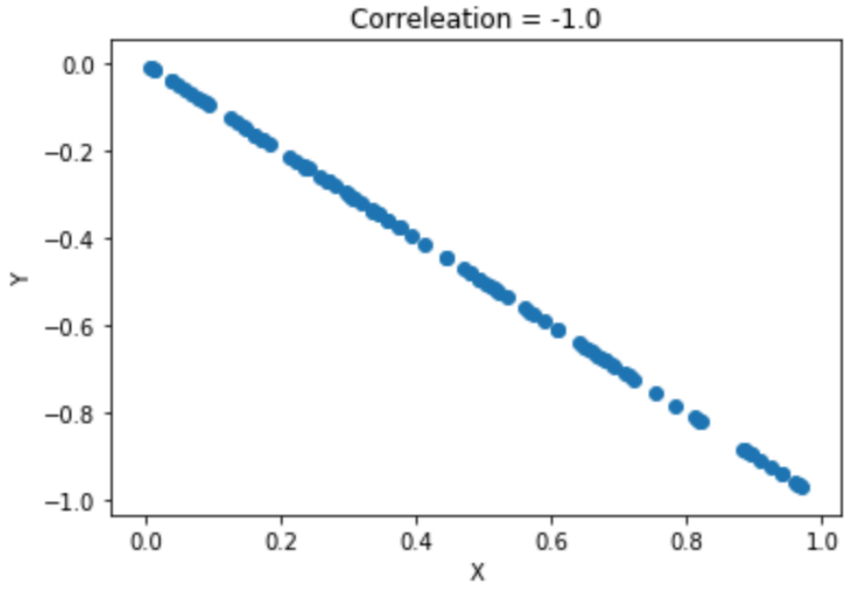
CHIDATA

Examples

In [19]:
```
y = -x

corr = scipy.stats.pearsonr(x,y)[0]

plt.scatter(x,y)
plt.xlabel('X')
plt.ylabel('Y')
plt.title('Correleation = '+ str(corr))
plt.show()
```



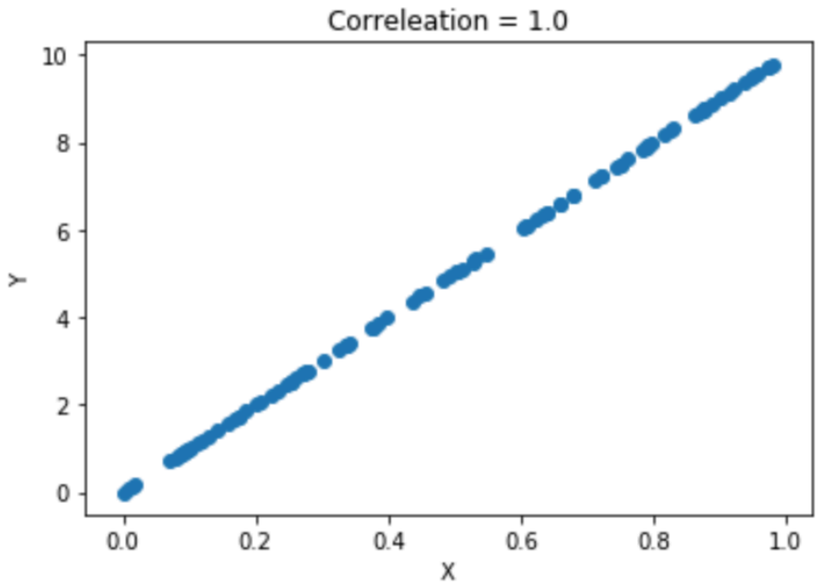
CHIDATA

Examples

In [20]:
```
x = np.random.rand(100)
y = 10*x

corr = scipy.stats.pearsonr(x,y)[0]

plt.scatter(x,y)
plt.xlabel('X')
plt.ylabel('Y')
plt.title('Correleation = '+ str(corr))
plt.show()
```



CHIDATA

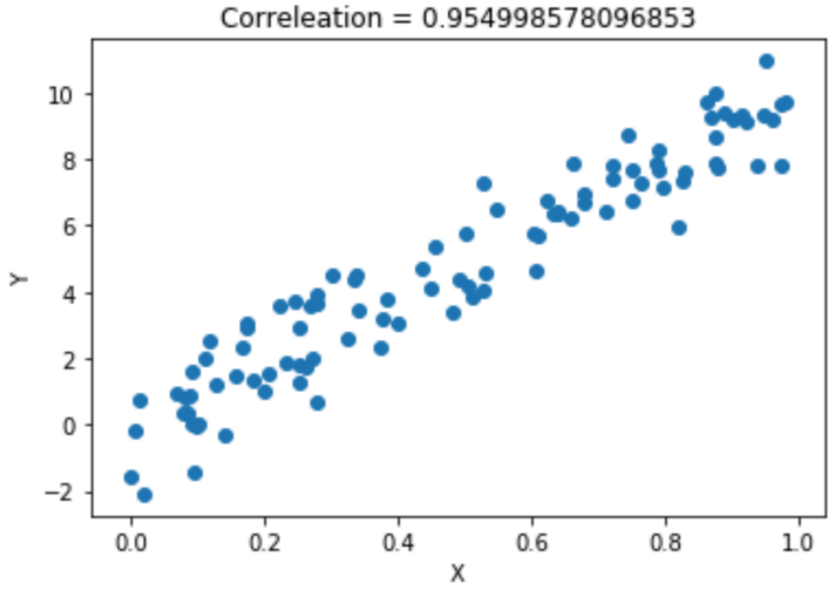
Examples

$$Y = Ax + noise$$

In [22]:
```
y = 10*x + np.random.randn(100)

corr = scipy.stats.pearsonr(x,y)[0]

plt.scatter(x,y)
plt.xlabel('X')
plt.ylabel('Y')
plt.title('Correleation = '+ str(corr))
plt.show()
```



CHIDATA

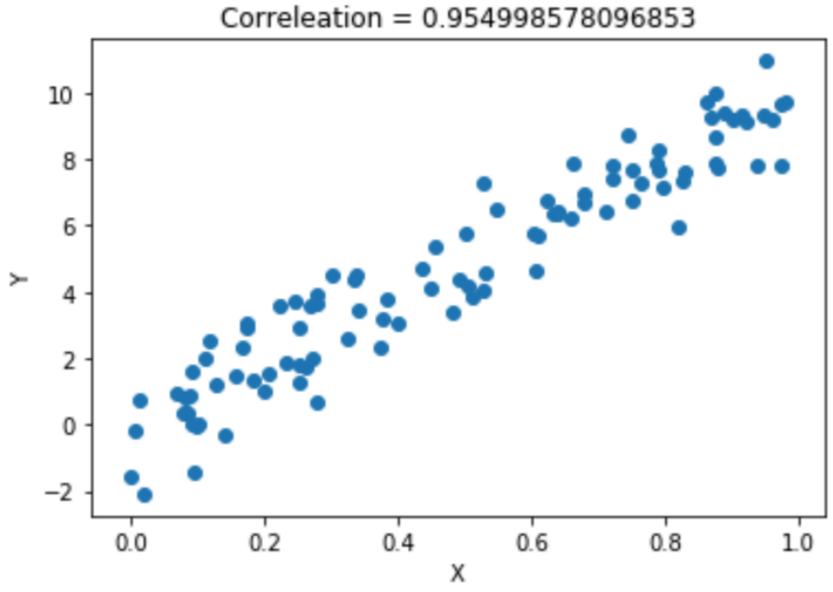
Examples

$$Y = Ax + noise$$

In [22]:
```
y = 10*x + np.random.randn(100)

corr = scipy.stats.pearsonr(x,y)[0]

plt.scatter(x,y)
plt.xlabel('X')
plt.ylabel('Y')
plt.title('Correleation = '+ str(corr))
plt.show()
```



CHIDATA

Examples

$$Y = A \cdot X + \epsilon \quad \epsilon \sim \text{i.i.d}$$

$$Cov(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$$

$$Cov(X,Y) = \mathbf{E}[X(AX+\epsilon)] - \mathbf{E}[X]\mathbf{E}[(AX+\epsilon)]$$

$$Cov(X,Y) = \mathbf{E}[AX^2] + \mathbf{E}[X\epsilon] - \mathbf{E}[X]\mathbf{E}[AX] - \mathbf{E}[X]\mathbf{E}[\epsilon]$$

$$Cov(X,Y) = A \cdot Var[X]$$

CHIDATA

Examples

$$Y = A \cdot X + \epsilon \quad \epsilon \sim \text{i.i.d}$$

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

Exercise for you guys….

$$Corr(X,Y) = \frac{1}{\sqrt{1 + \frac{Var(\epsilon)}{A^2 \cdot Var(X)}}}$$

$$Corr(X,Y) = \frac{1}{\sqrt{1 + SNR^{-1}}}$$

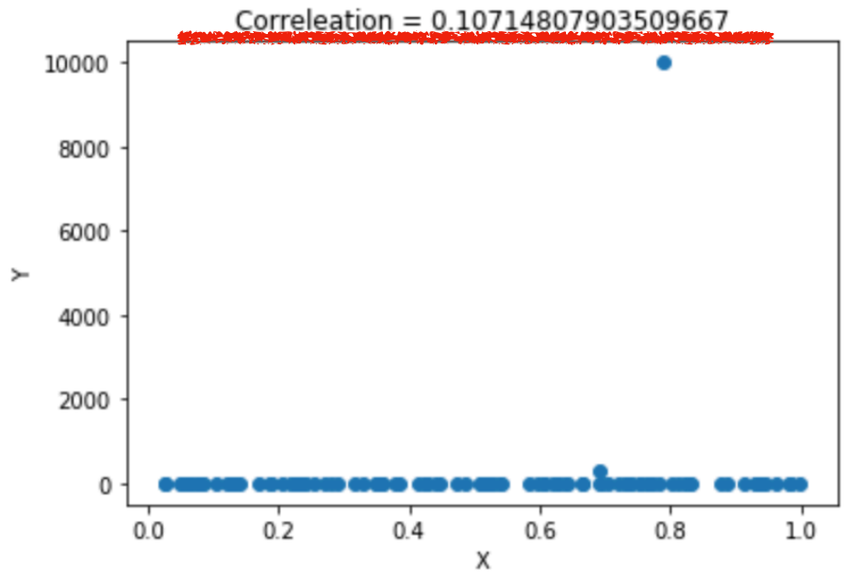
CHIDATA

Examples

In [16]:
```
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats

x = np.random.rand(100)
y = 10*x + np.random.randn(100)
y[0] = 300
y[1] = 10000
corr = scipy.stats.pearsonr(x,y)[0]

plt.scatter(x,y)
plt.xlabel('X')
plt.ylabel('Y')
plt.title('Correleation = '+ str(corr))
plt.show()
```



Summary: Covariance and Correlation

The strength of "linear" dependence between two sets of observations.

$X' \bullet Y'$ Covariance

$$\mathrm{Cov}\,(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

$$\frac{X' \bullet Y'}{|X'|\,|Y'|}$$ Correlation "Pearson"

$$\frac{Cov\,(x, y)}{\sigma x * \sigma y}$$