# CSCI 443 LECTURE 6: SPARK CATALYZER & CORRELATION

Professor David Harrison

# TODAY

- Spark DAGs
- Spark Catalyzer
- Correlation, covariance, correlation matrices

# DATES OF INTEREST

| | |
|---|---|
| 2/13 TH | HW2 due |
| 2/20 TH | HW3 due |
| 2/25 T | Midterm review |
| 2/27 TH | Midterm |
| 3/3 M | Midterm grades due |
| 3/7 F | Last date to drop (no refund) |

# HOMEWORK 2

Due next Thursday, Feb 13.

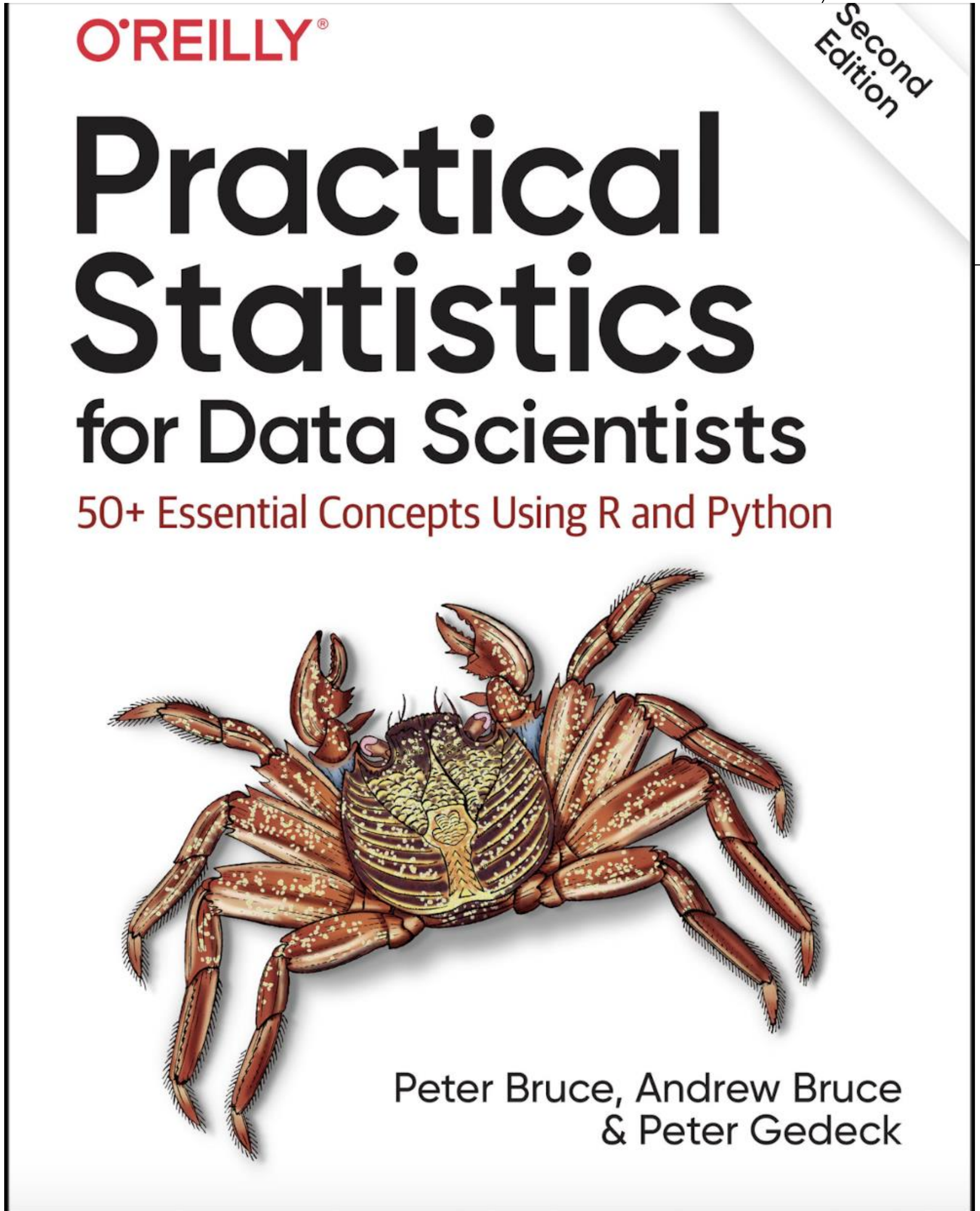
11 pm.

Submission:

- Submit archived Databricks Notebook to Blackboard.
- You MAY submit a scanned handwritten page alongside the notebook for the problems that only involve math.
- NOTE: Submission only needs to be the notebook (and optionally the scanned page).  No README is necessary.

# READING!

- Read Chapter 2: Data and Sampling Distributions.
  - We already discussed forms of bias. This is covered in the first part of chapter 2.
  - Read up through the section titled "Regression to the Mean"



# OFFICE HOURS

Due to scheduling conflict, office hours updated

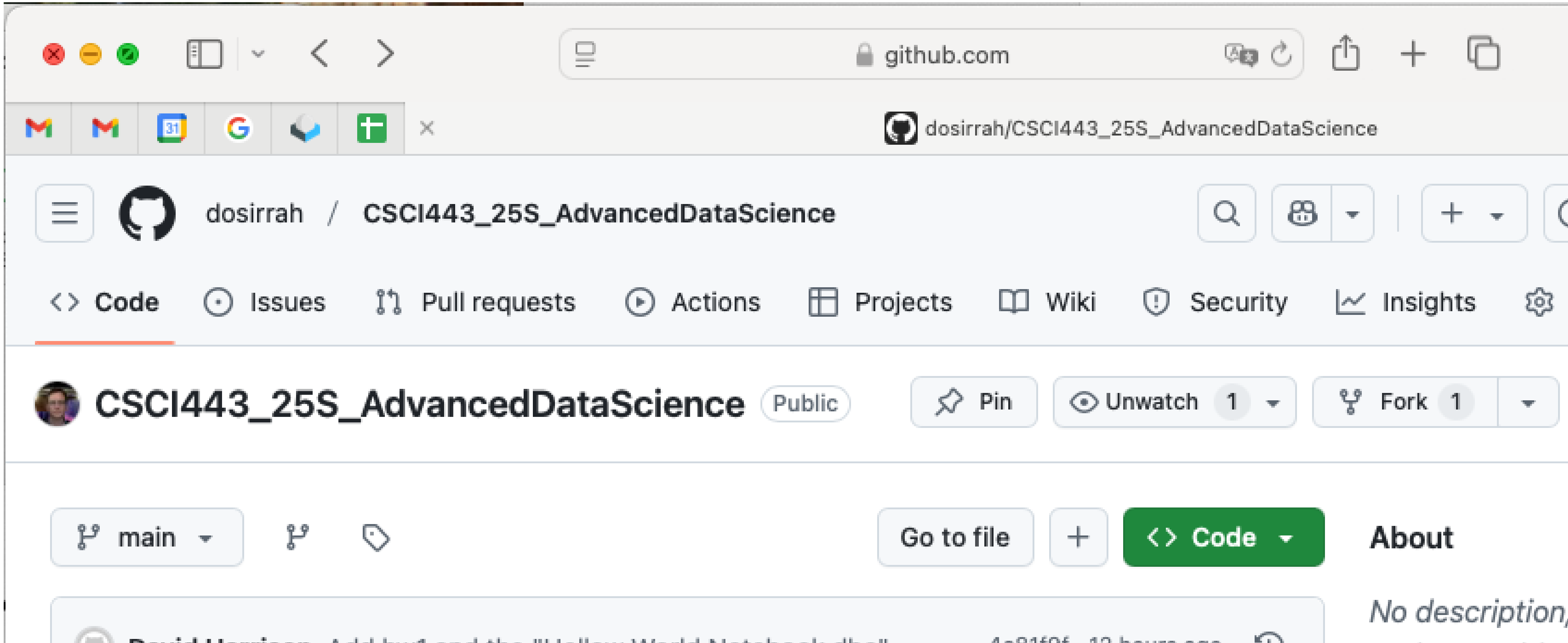
| | |
|---|---|
| Tuesday | 4-5 PM |
| Wednesday | 12:30-1:30 PM |

# GITHUB

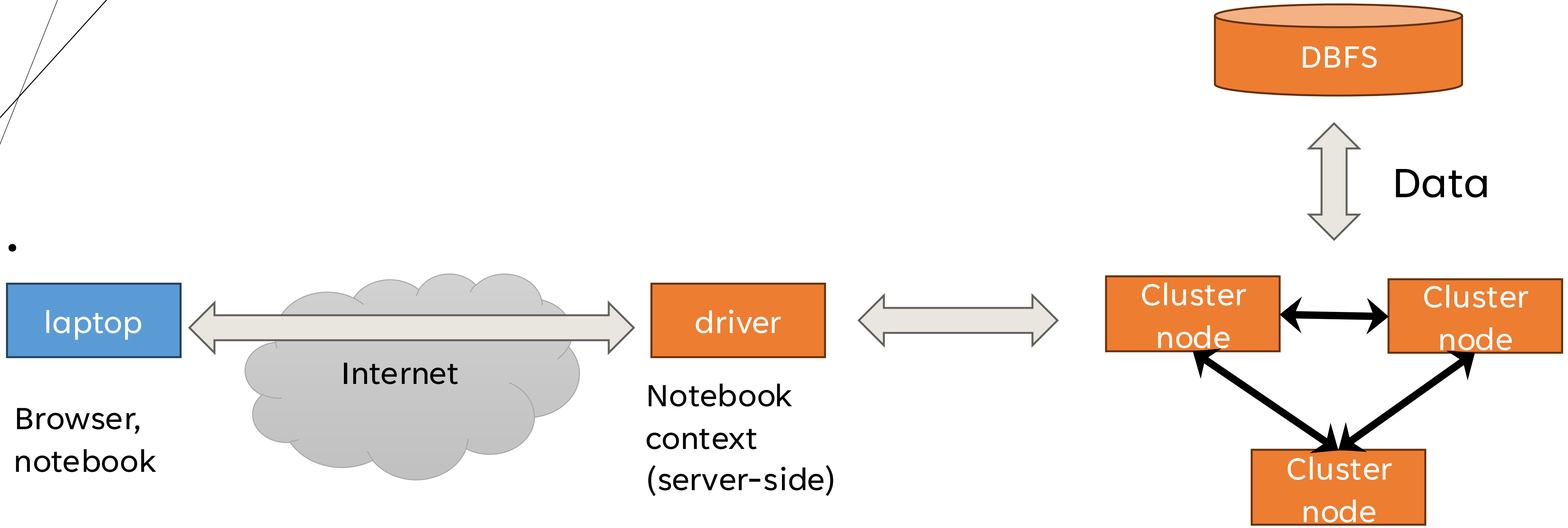
Lecture slides and examples have been committed to GitHub for lectures 1 through 4.

The project is at

https://github.com/dosirrah/CSCI443_25S_AdvancedDataScience



# SPARK ARCHITECTURE (1000 FT)



# TRANSFORMATIONS

Added to the plan.

| Transformation | Description | Example |
|---|---|---|
| `select()` | Select specific columns | `df.select("Name", "Age")` |
| `filter()` / `where()` | Filter rows based on a condition | `df.filter(df.Age > 30)` |
| `withColumn()` | Add or modify a column | `df.withColumn("AgePlusOne", df.Age + 1)` |
| `drop()` | Remove a column | `df.drop("Ticket")` |
| `orderBy()` | Sort rows by a column | `df.orderBy("Fare", ascending=False)` |
| `limit()` | Take the first N rows (still lazy) | `df.limit(10)` |
| `distinct()` | Remove duplicate rows | `df.select("Pclass").distinct()` |

# ACTIONS

Execute the plan to generate output.

| Action | Description | Example |
|---|---|---|
| `show()` | Displays results | `df.show(5)` |
| `collect()` | Brings all data to the driver (⚠️ not recommended for large data) | `df.collect()` |
| `count()` | Returns the total number of rows | `df.count()` |
| `first()` | Returns the first row | `df.first()` |
| `take(n)` | Returns the first n rows | `df.take(5)` |
| `describe()` | Computes summary statistics | `df.describe().show()` |
| `summary()` | More detailed statistics than `describe()` | `df.summary().show()` |

# ACTIONS

Execute the plan to generate output.

| Action | Description | |
|---|---|---|
| `show()` | Displays results | |
| `collect()` | Brings all data to the driver (⚠️ not recommended for large data) | |
| `count()` | Returns the total number of rows | `df.count()` |
| `first()` | Returns the first row | `df.first()` |
| `take(n)` | Returns the first n rows | `df.take(5)` |
| `describe()` | Computes summary statistics | `df.describe().show()` |
| `summary()` | More detailed statistics than `describe()` | `df.summary().show()` |

Is count() an action or a transformation?

# ACTIONS

Execute the plan to generate output.

| Action | Description | |
|---|---|---|
| `show()` | Displays results | |
| `collect()` | Brings all data to the driver (⚠️ not recommended for large data) | |
| `count()` | Returns the total number of rows | `df.count()` |
| `first()` | Returns the first row | `df.first()` |
| `take(n)` | Returns the first n rows | `df.take(5)` |
| `describe()` | Computes summary statistics | `df.describe().show()` |
| `summary()` | More detailed statistics than `describe()` | `df.summary().show()` |

Is count() an action or a transformation?
BOTH

# COUNT AS AN ACTION

df.count() immediately executes on the DataFrame.

It counts the number of rows.

1 minute ago (1s)

```
# count the number of rows.

df.count()
```

(2) Spark Jobs

```
Out[31]: 891
```

# COUNT AS A TRANSFORMATION

The count specifies a transformation to perform a count, but it doesn't actually perform the count until we associate the plan with data (in this case df), and then execute the action show().

01:56 PM (1s)

```
from pyspark.sql.functions import count

df.select(count(when(col("Age").isNull(), 1))).show()
```

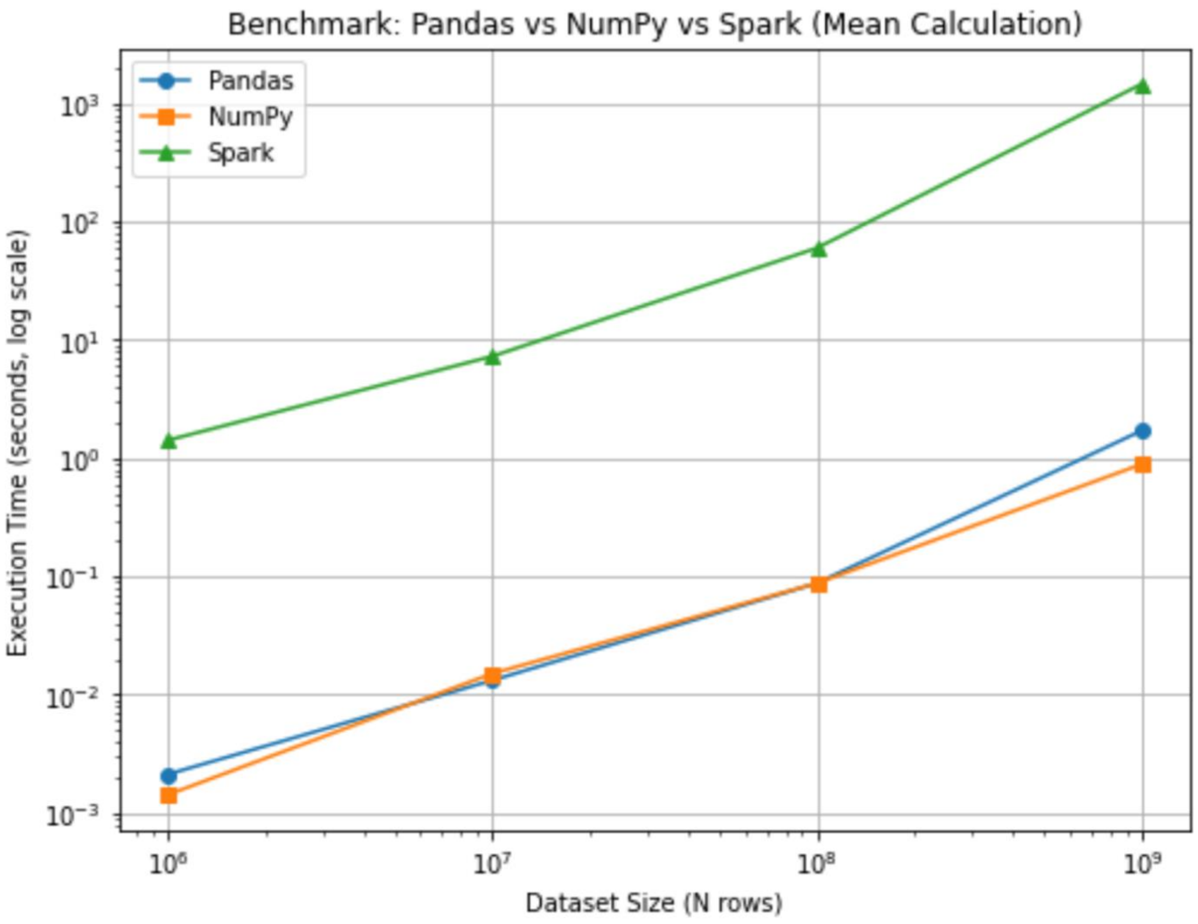
(2) Spark Jobs

```
+-------------------------------------+
|count(CASE WHEN (Age IS NULL) THEN 1 END)|
+-------------------------------------+
|                                  177|
+-------------------------------------+
```

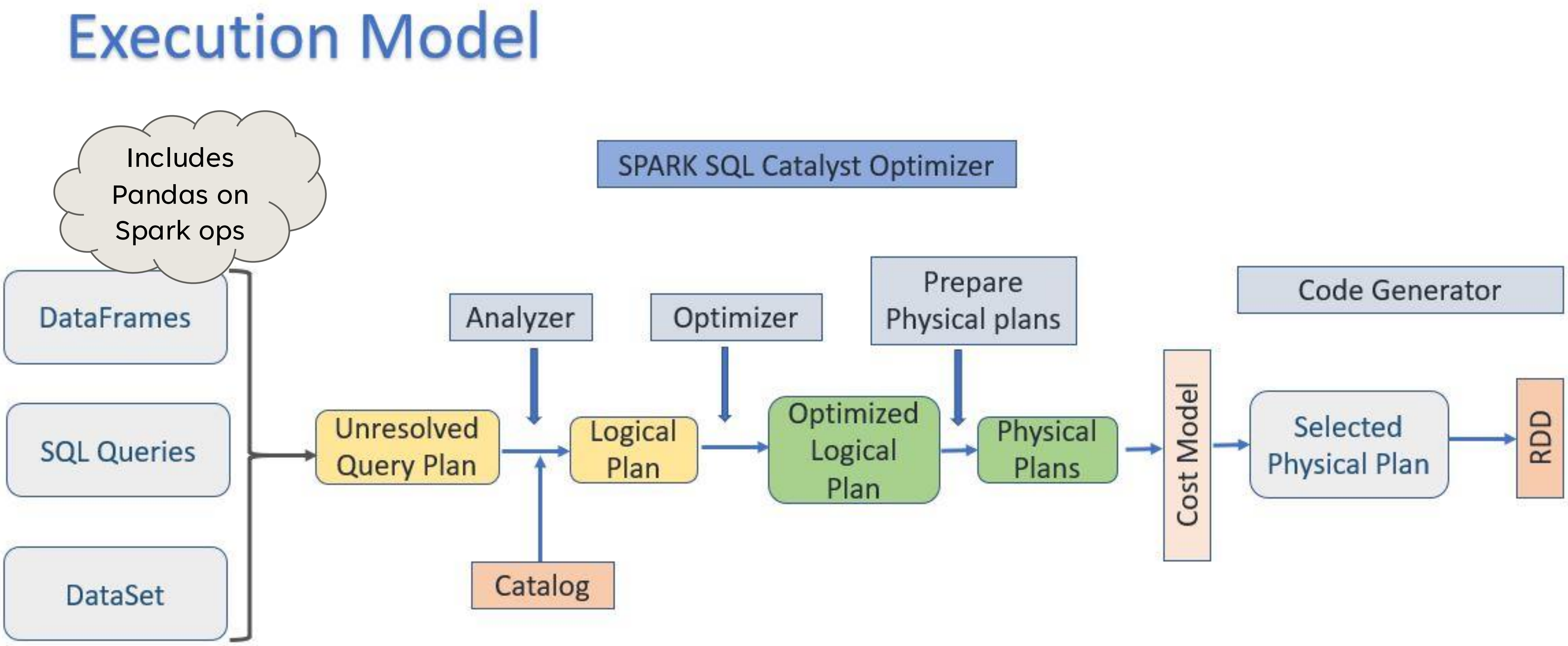
# AN EXERCISE IN BENCHMARKING

Community Edition only allocates one node for both driver and cluster.

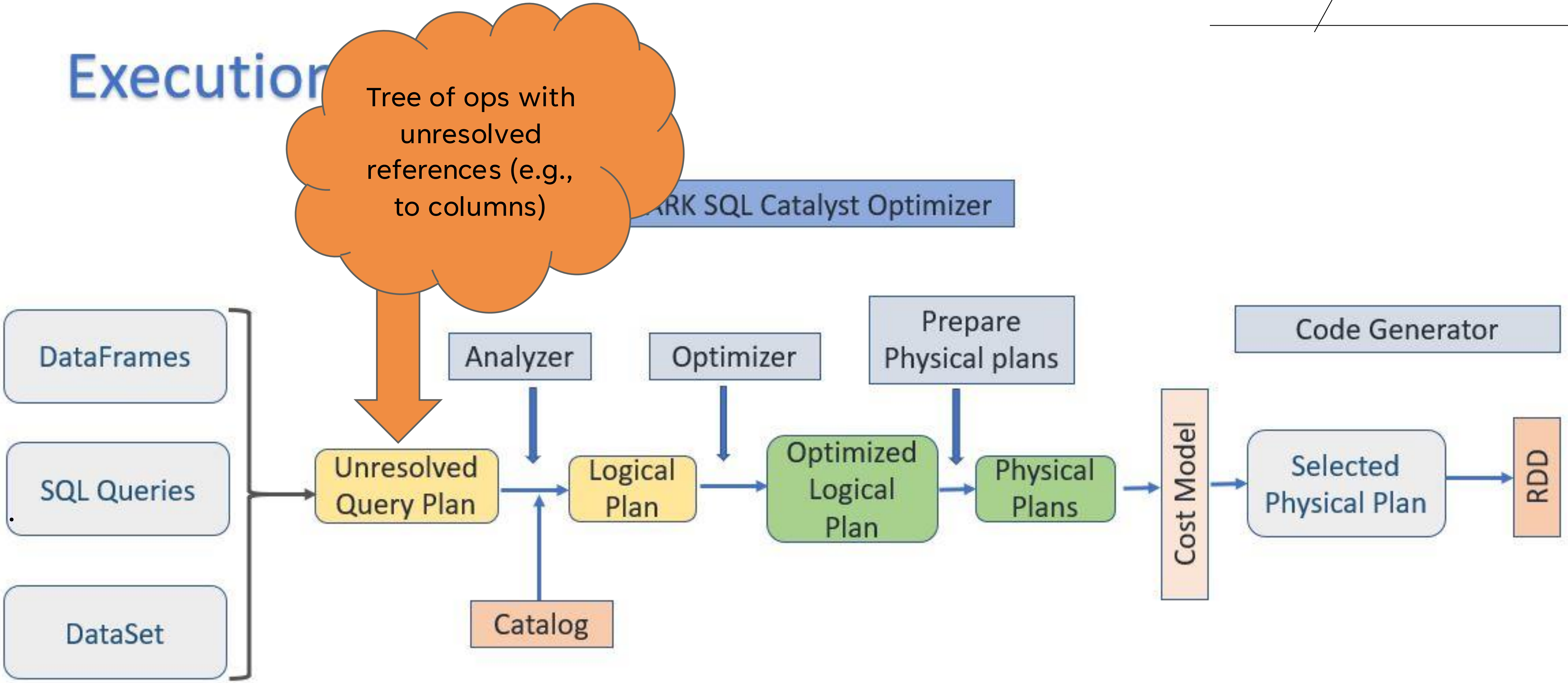
- Spark and driver share the same resources.
- Cannot exploit the scaling properties of a cluster.
- Spark overhead results in worse performance at all scales compared to Numpy and Pandas



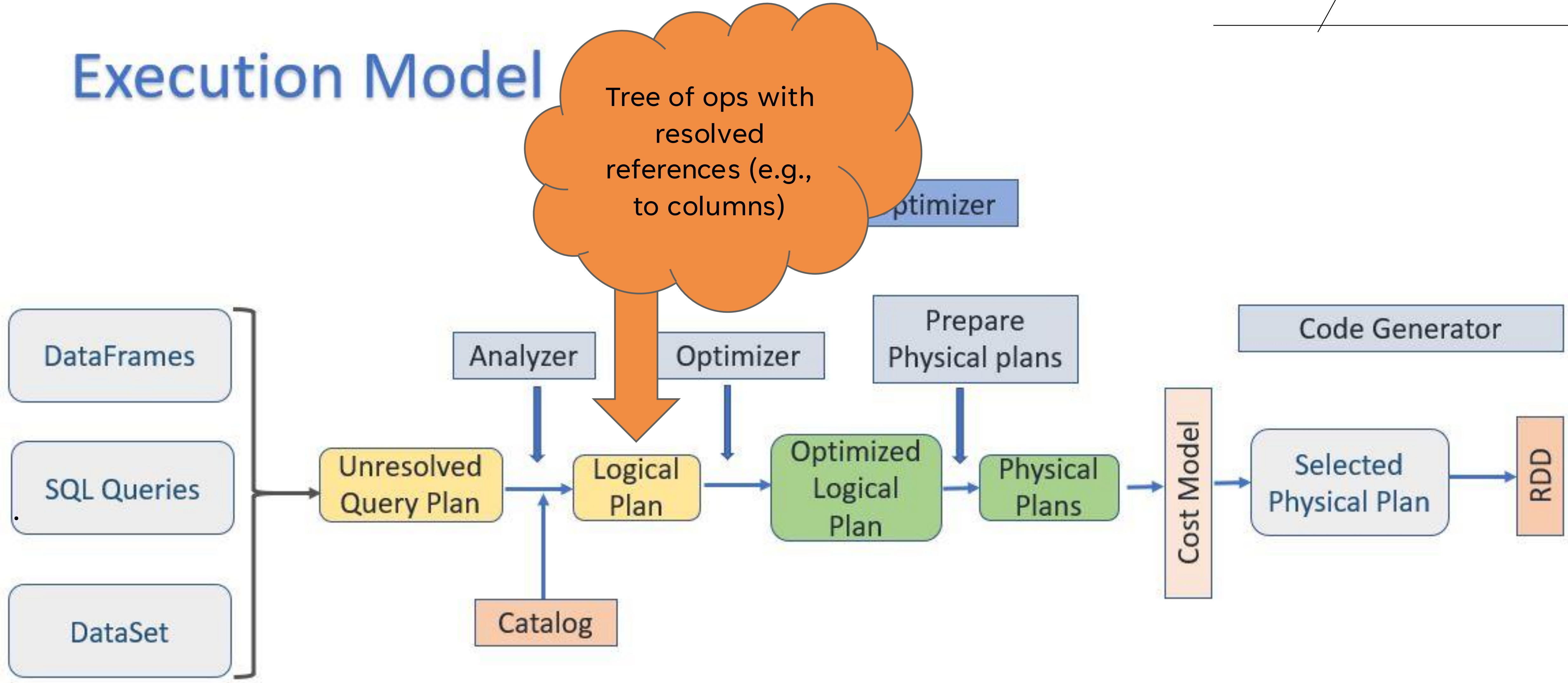
# CATALYST OPTIMIZER



# CATALYST OPTIMIZER

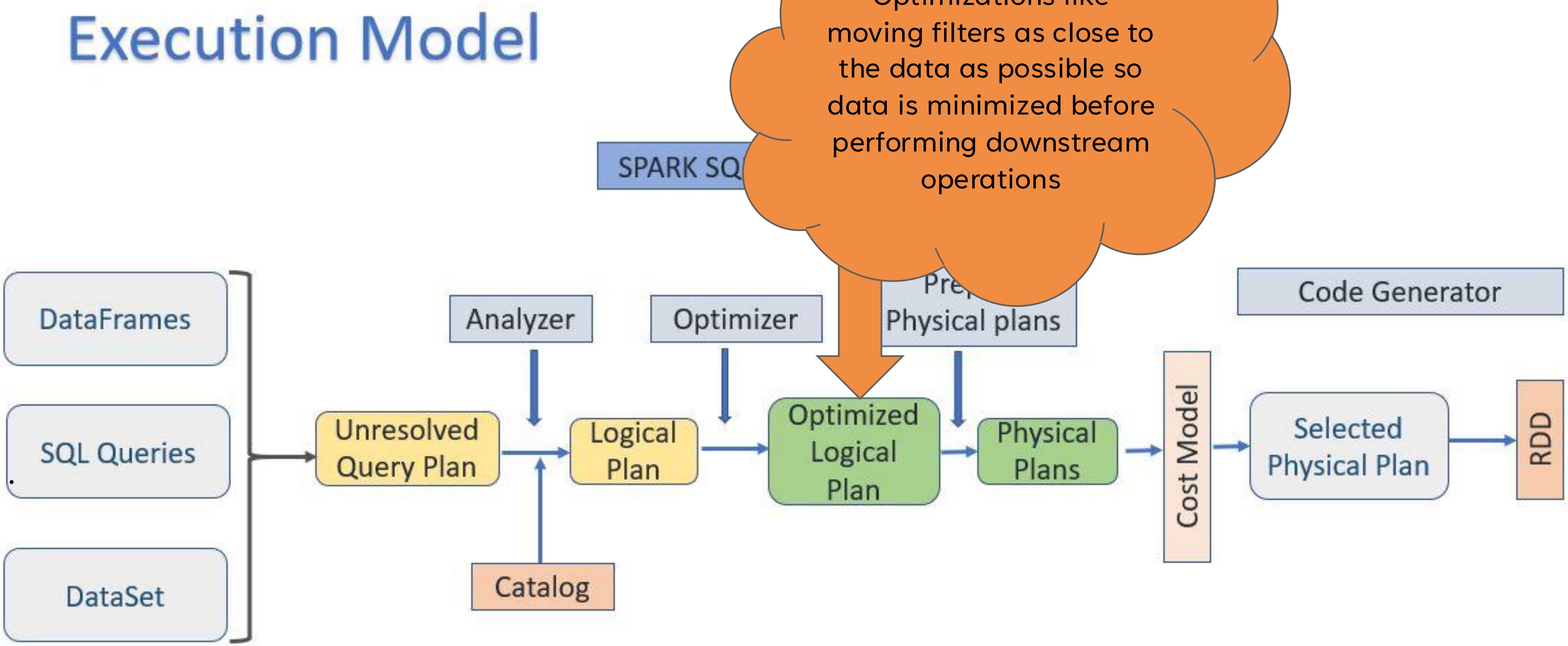


# CATALYST OPTIMIZER

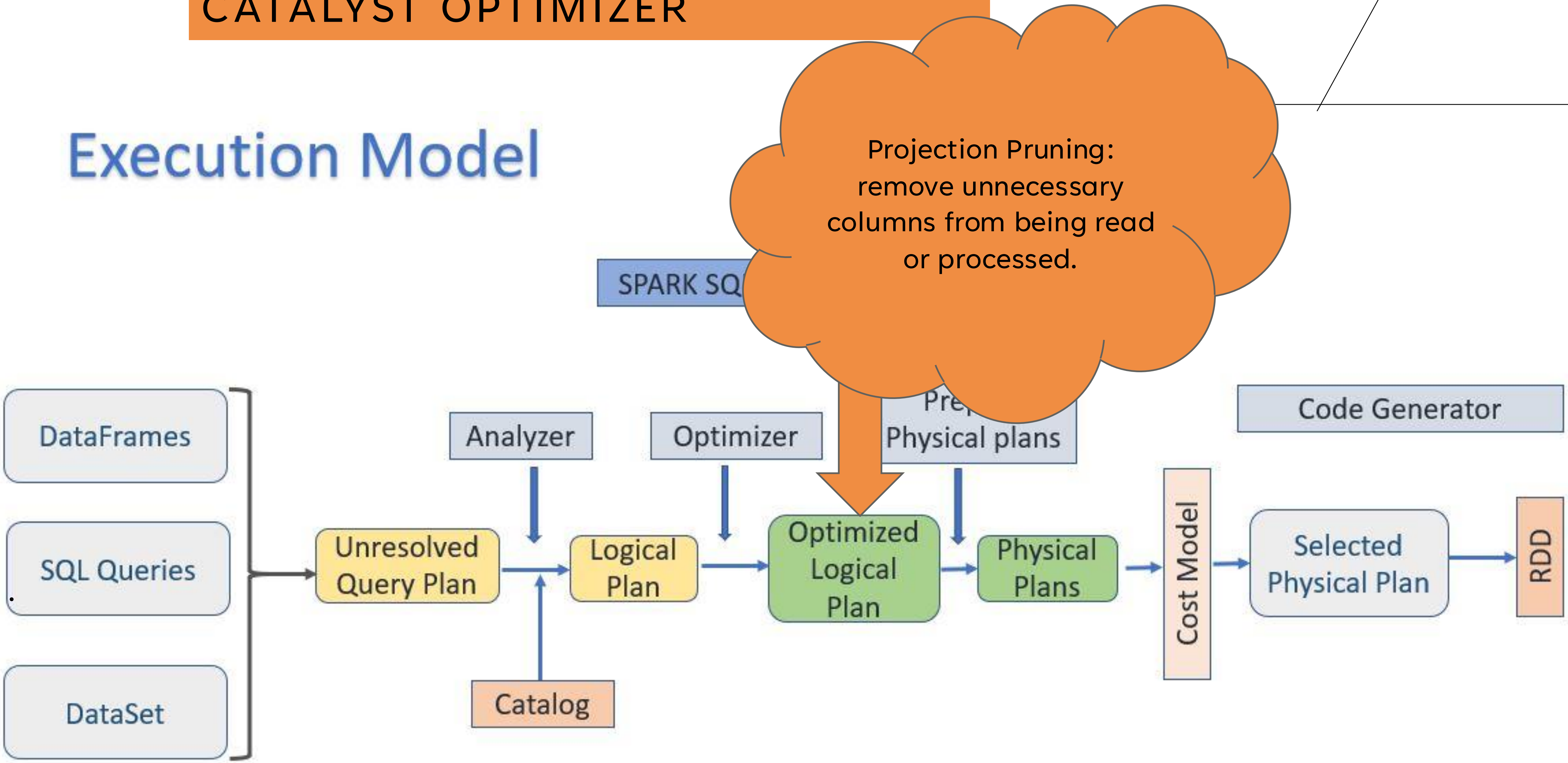


# CATALYST OPTIMIZER

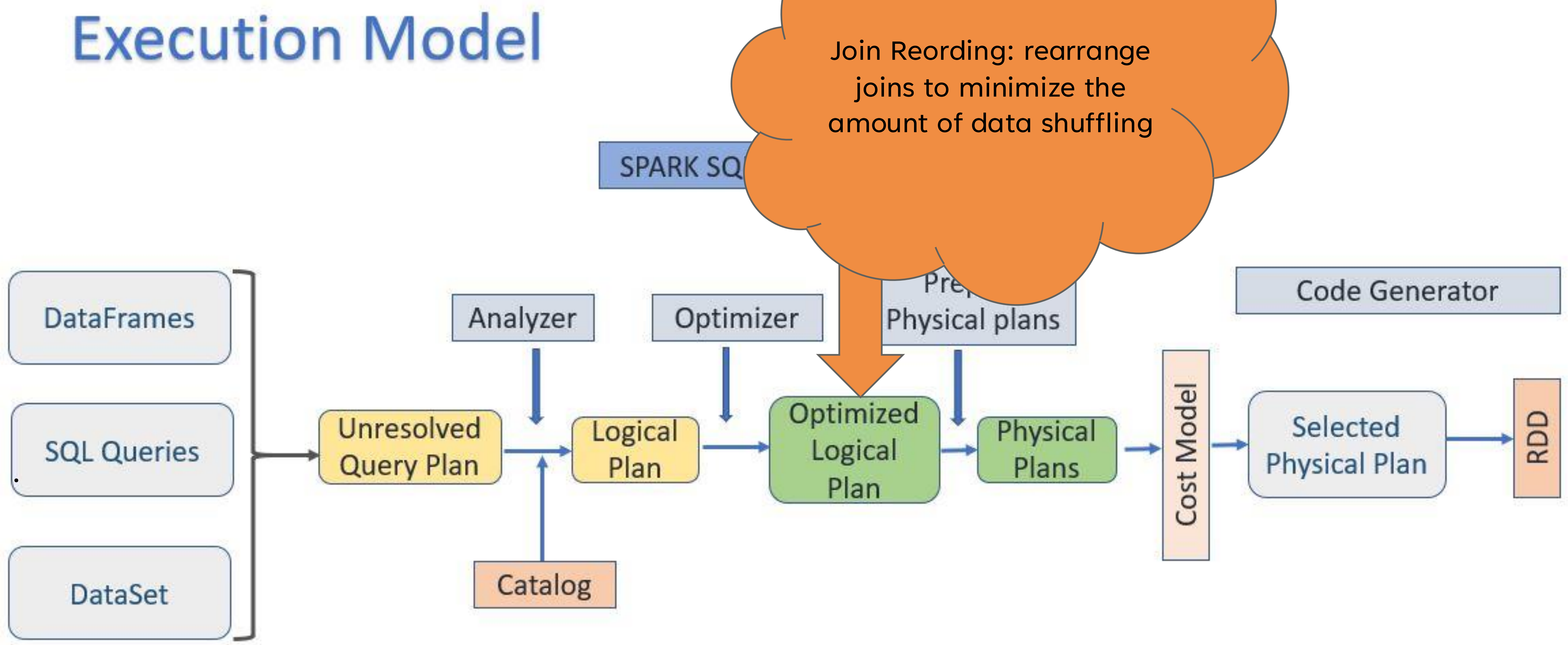
Optimizations like moving filters as close to the data as possible so data is minimized before performing downstream operations



# CATALYST OPTIMIZER



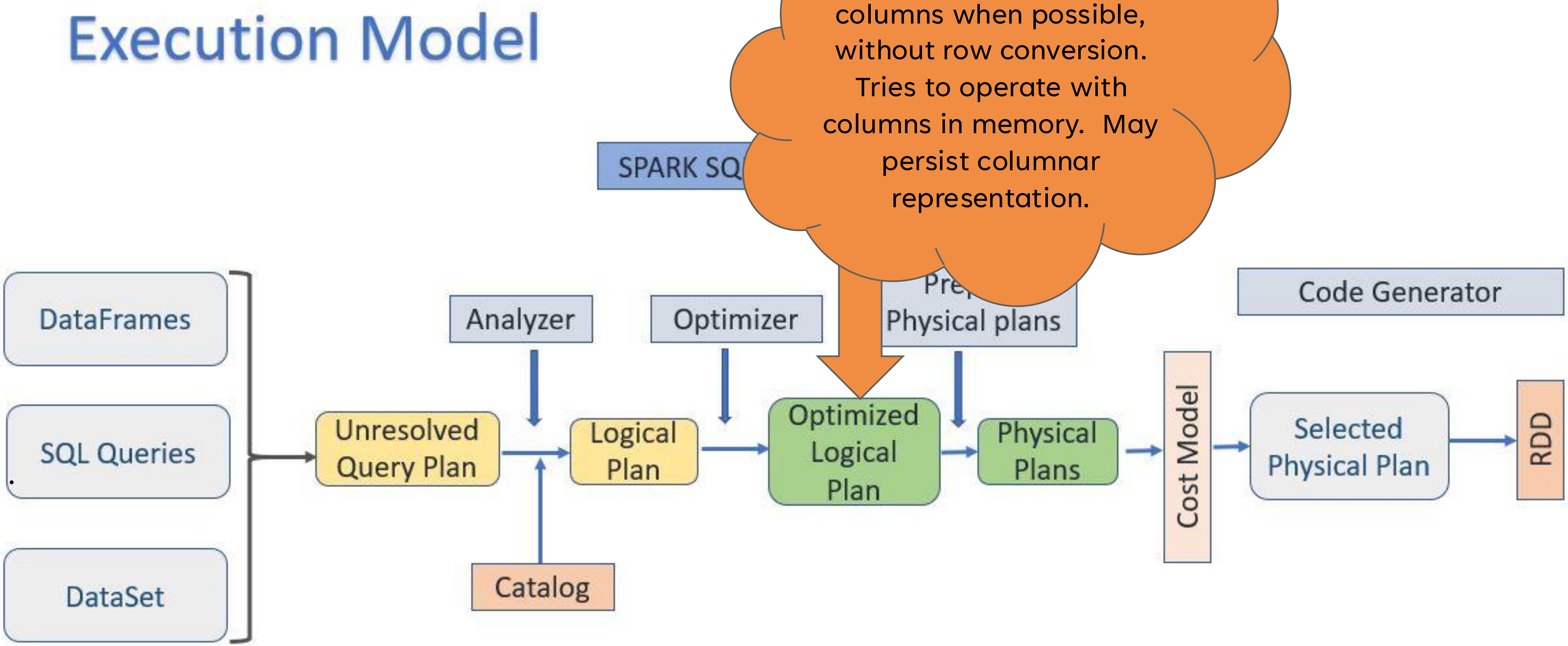
# CATALYST OPTIMIZER



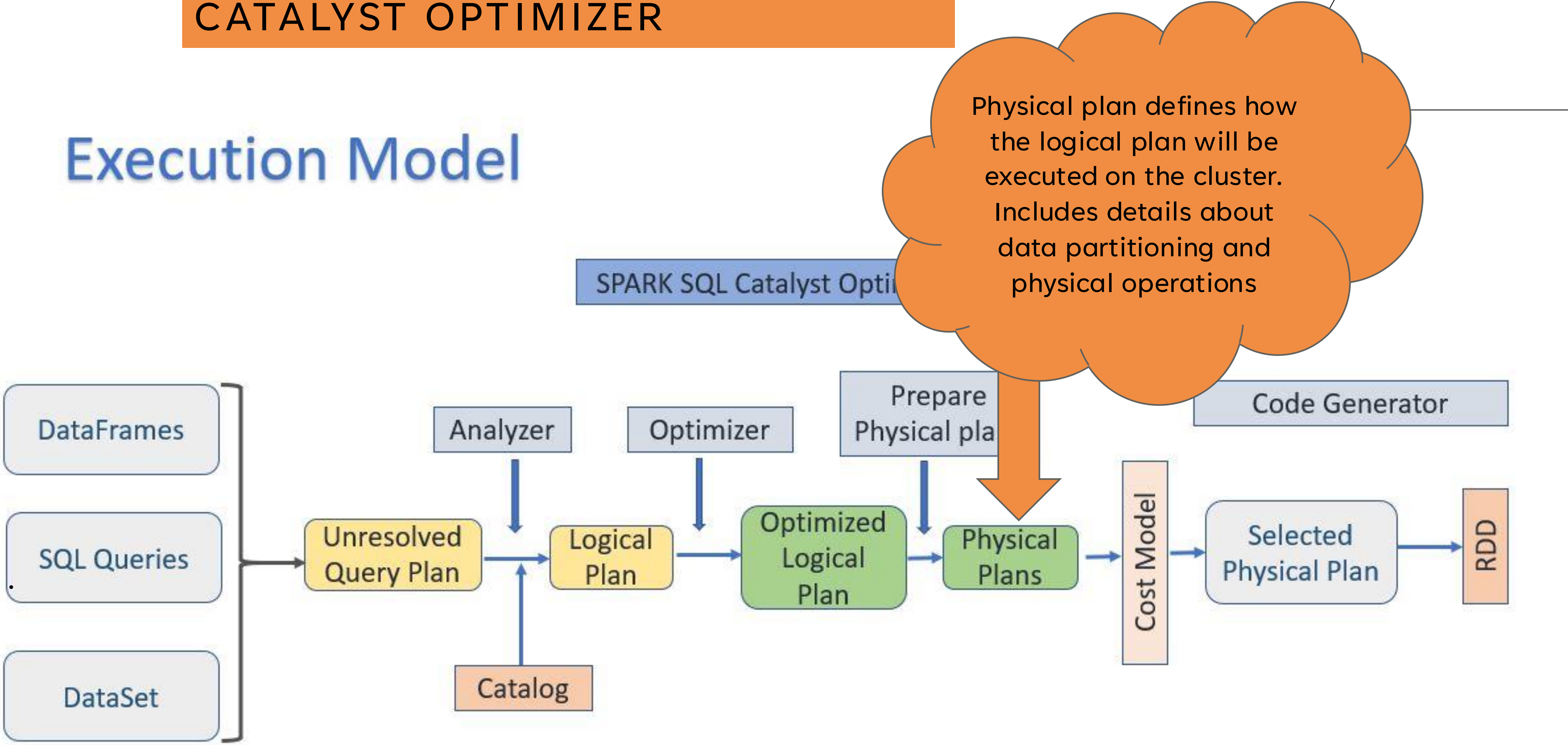

Join Reording: rearrange joins to minimize the amount of data shuffling

# CATALYST OPTIMIZER

Columnar processing: use columns when possible, without row conversion. Tries to operate with columns in memory.  May persist columnar representation.

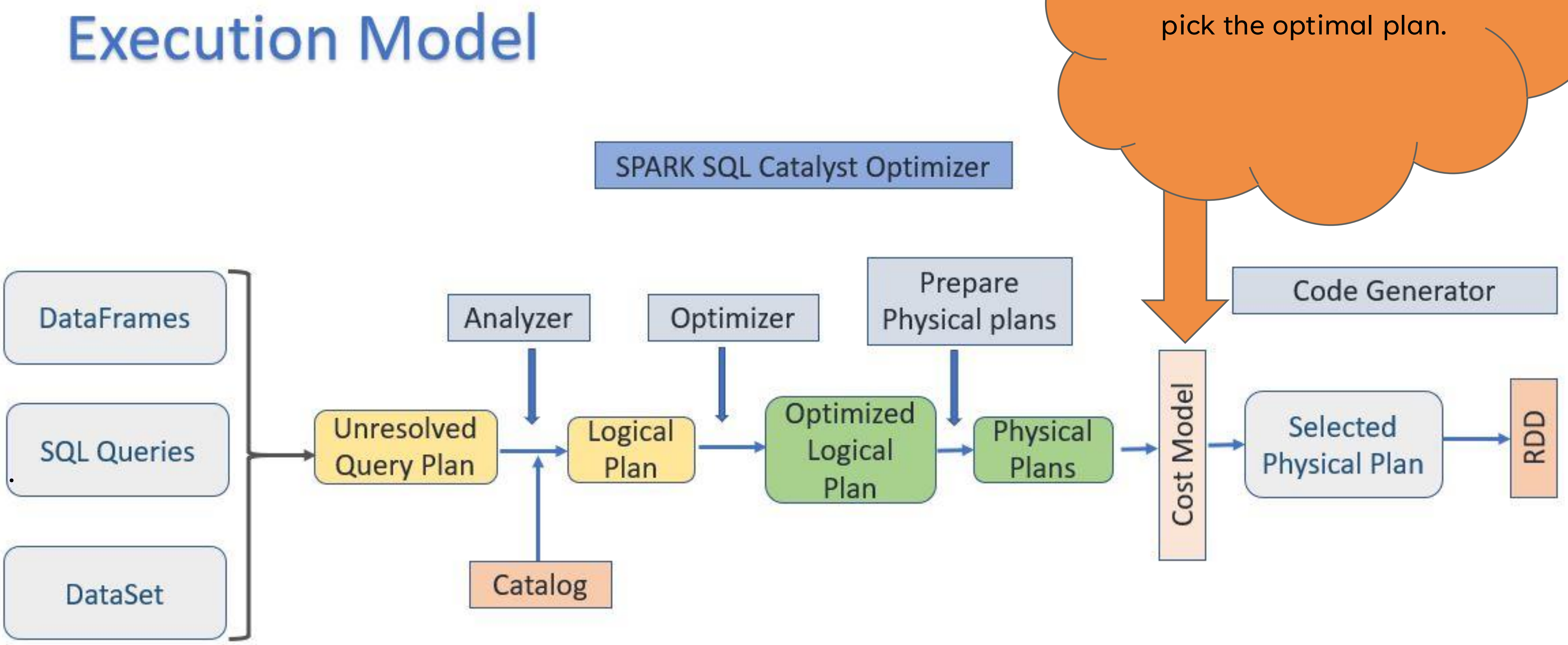


# CATALYST OPTIMIZER



# CATALYST OPTIMIZER

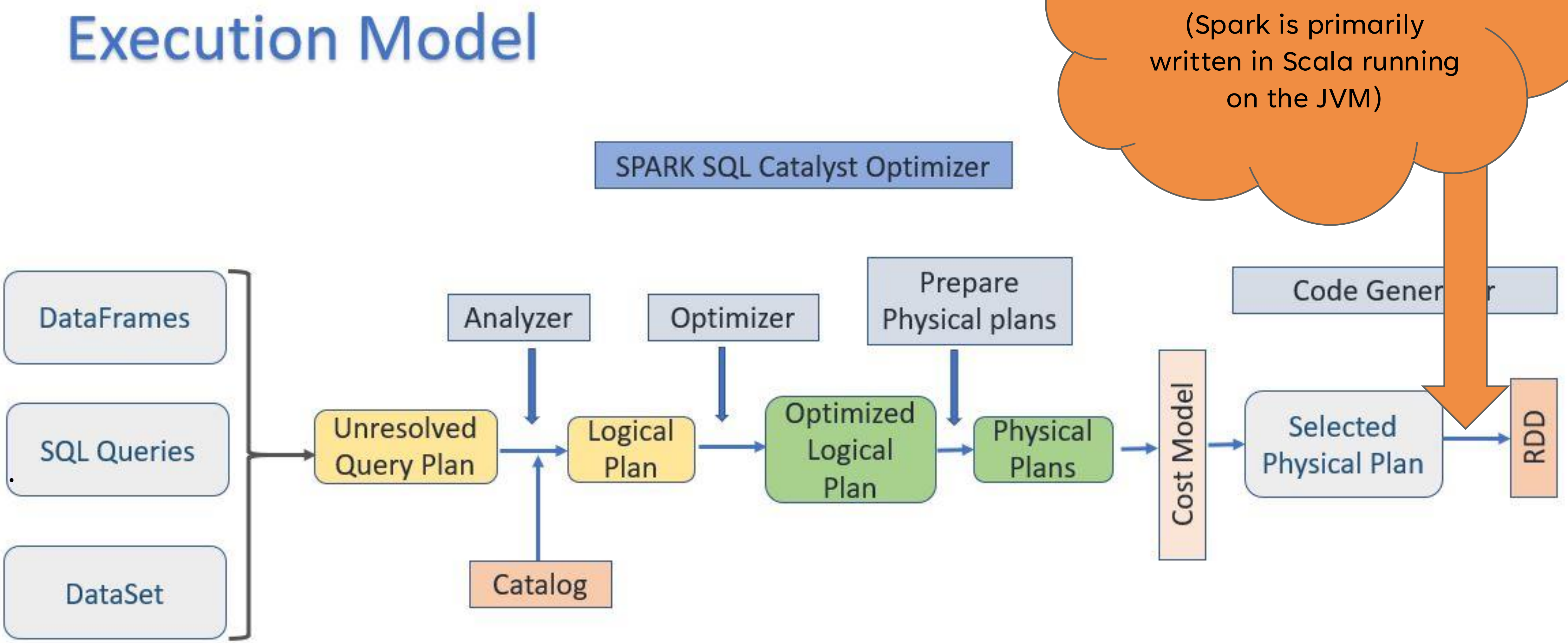
Cost model is used to pick the optimal plan.



# CATALYST OPTIMIZER

Code generator outputs optimized Java bytecode

(Spark is primarily written in Scala running on the JVM)



# CORRELATION

## KEY TERMS FOR CORRELATION

***Correlation coefficient***

A metric that measures the extent to which numeric variables are associated with one another (ranges from –1 to +1).

***Correlation matrix***

A table where the variables are shown on both rows and columns, and the cell values are the correlations between the variables.

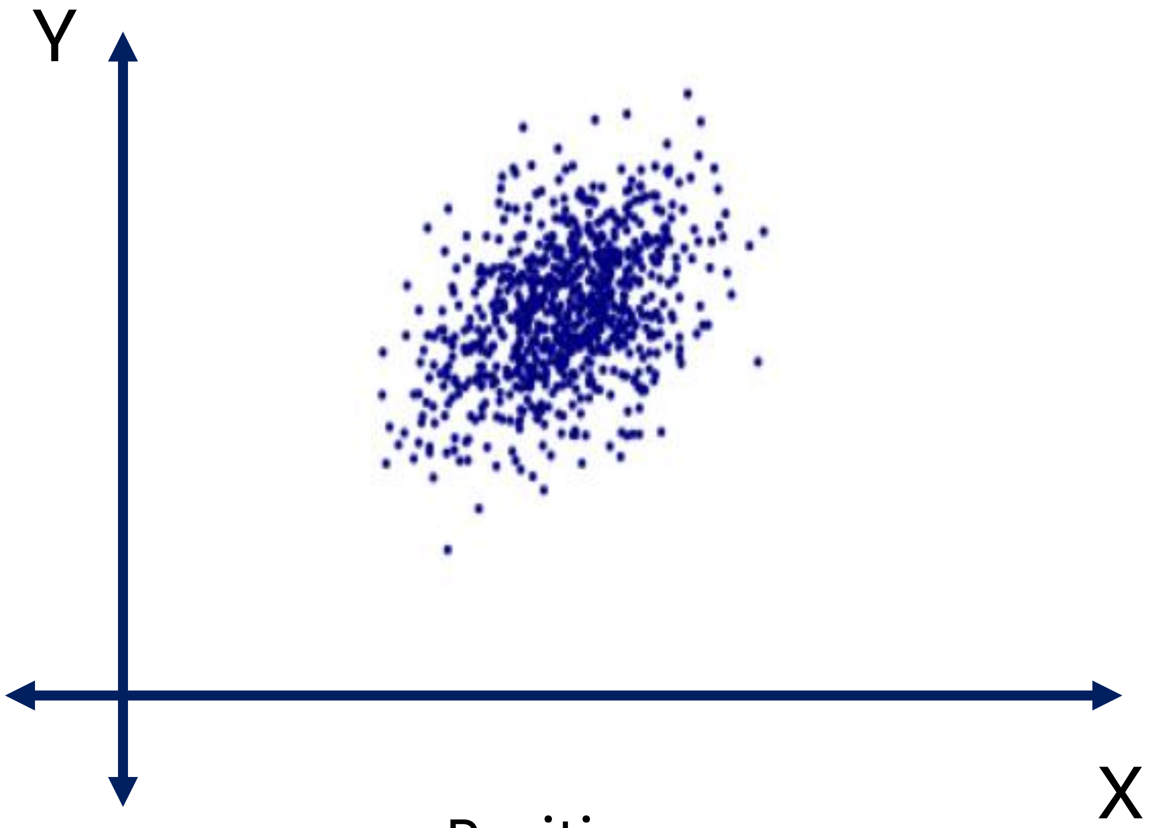
***Scatterplot***

A plot in which the x-axis is the value of one variable, and the y-axis the value of another.

# CORRELATION

Correlation between two random variables means they tend to move together.

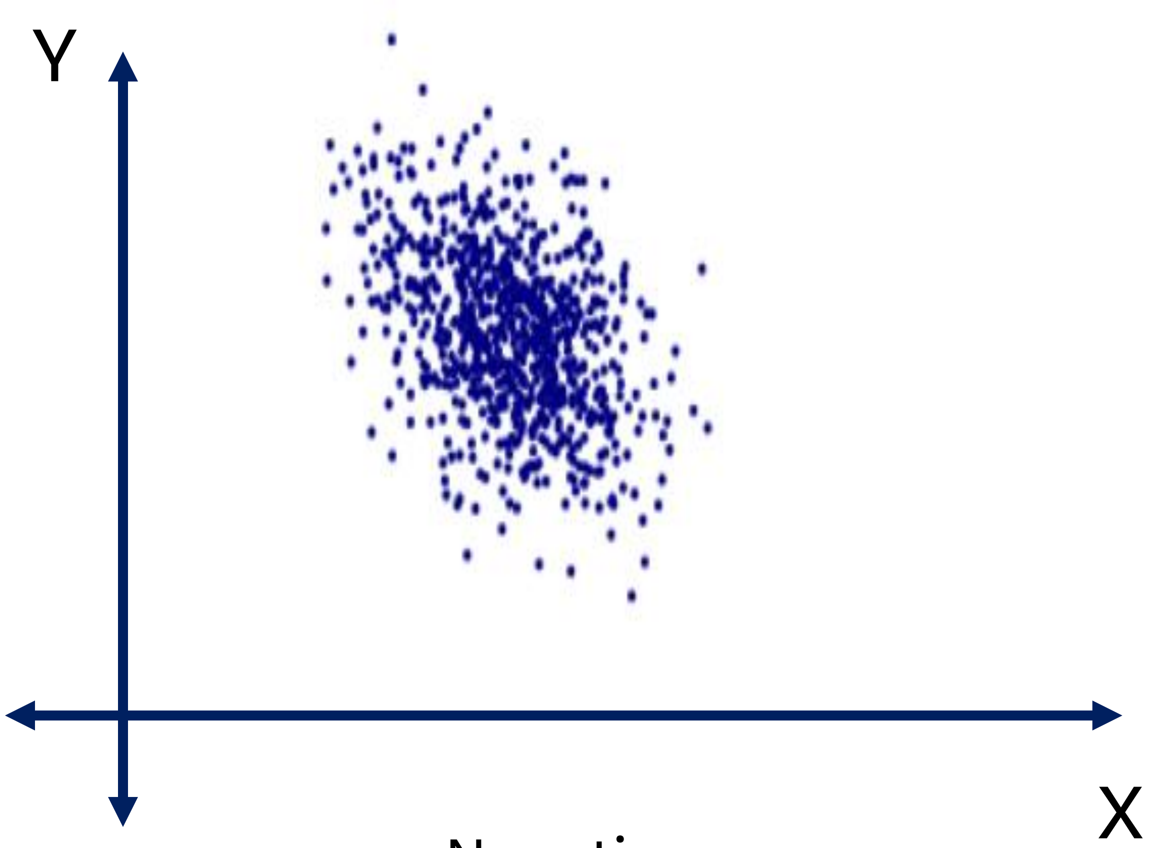
When one increases the other does, and vice versa.




Positive correlation

# CORRELATION

Negative correlation means they tend to move opposite to one another.

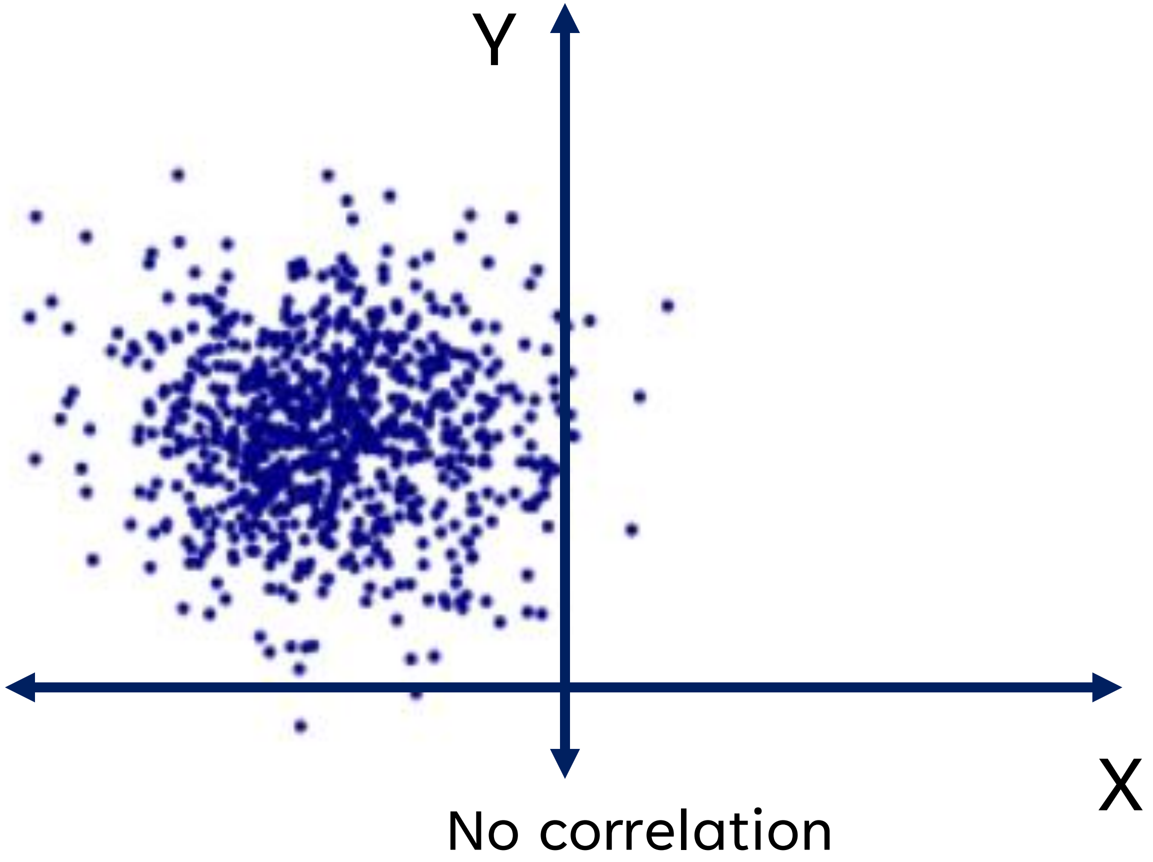



Negative correlation

# CORRELATION

There is no correlation if they do not move together.

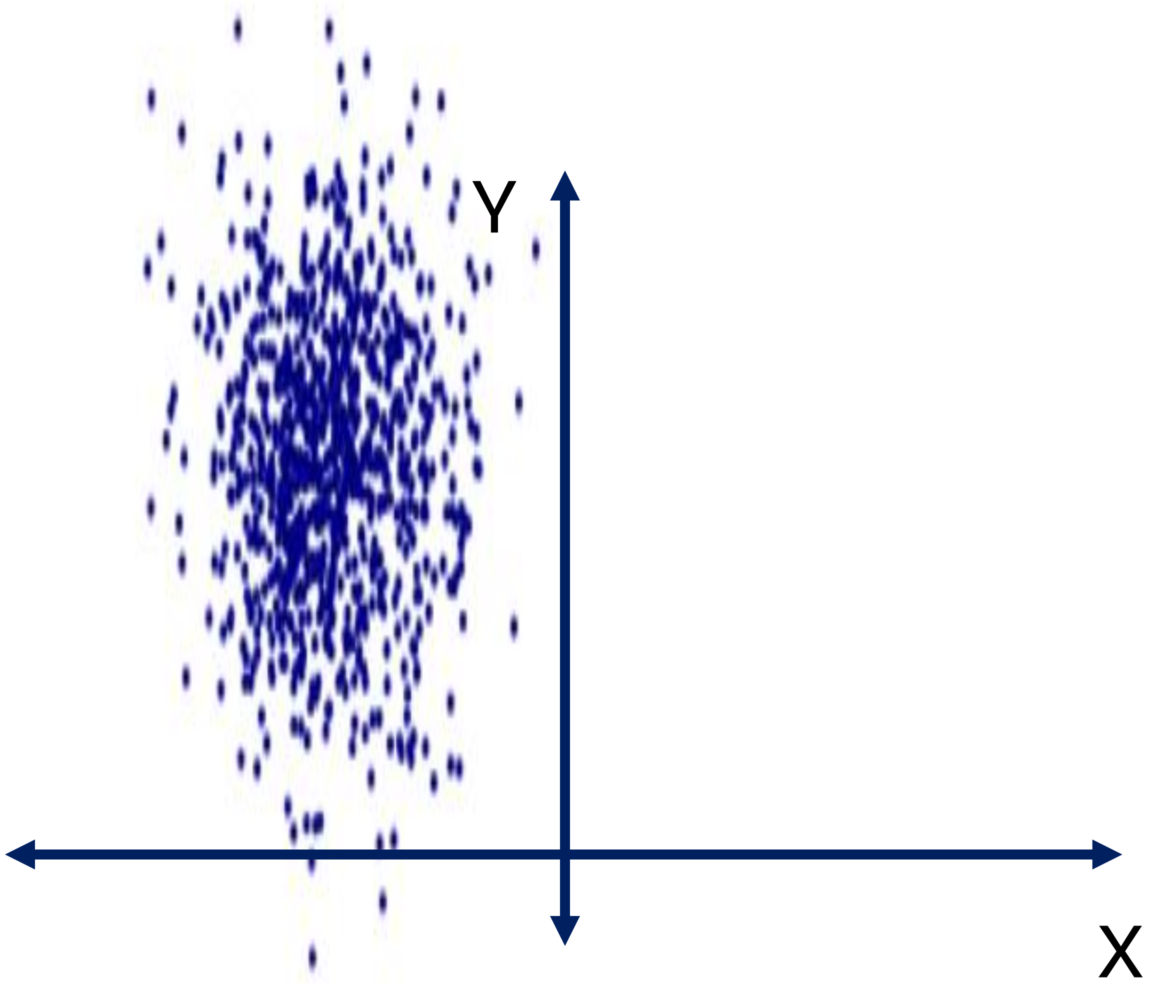
On a Cartesian plane this appears as NO tilt to the scatter of samples.



# CORRELATION

There is no correlation if they do not move together.

On a Cartesian plane this appears as NO tilt to the scatter of samples.

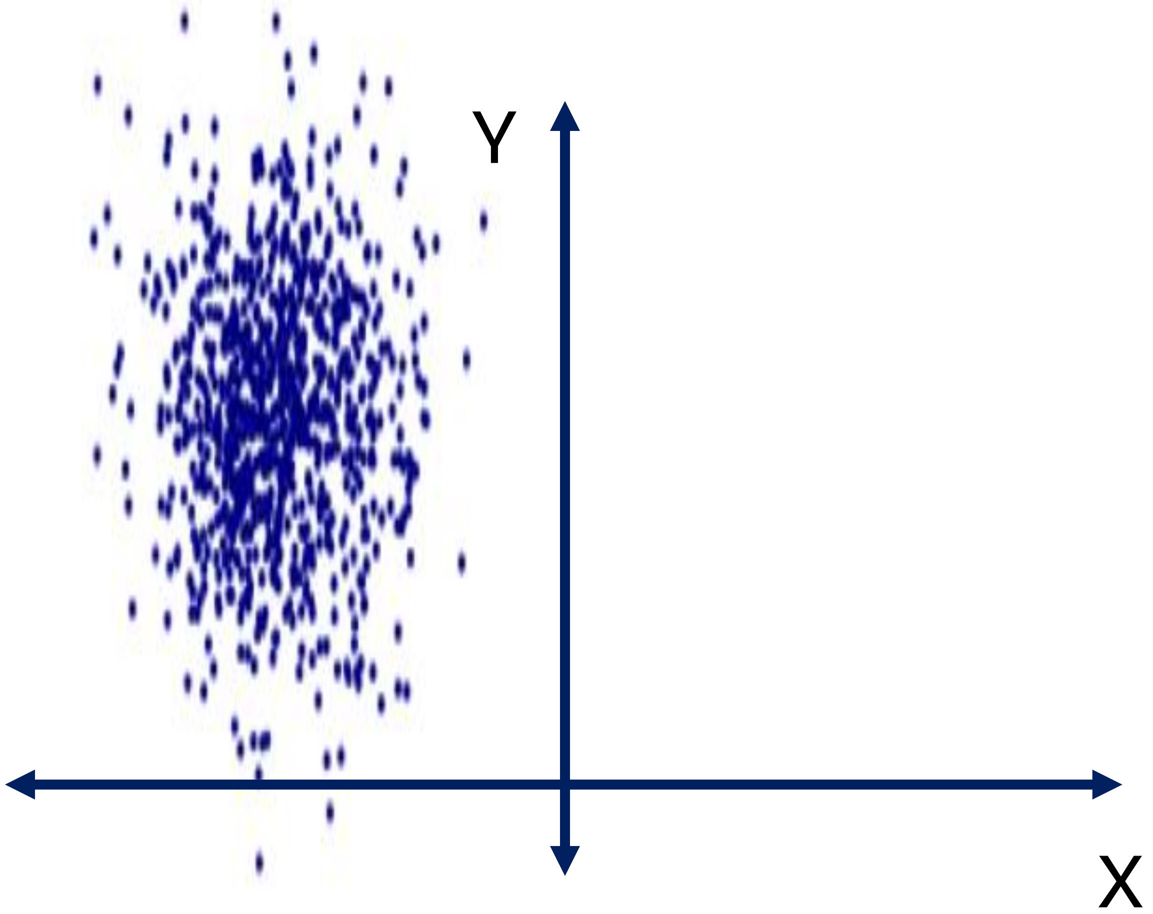



Correlated??????

# CORRELATION

There is no correlation if they do not move together.

On a Cartesian plane this appears as NO tilt to the scatter of samples.

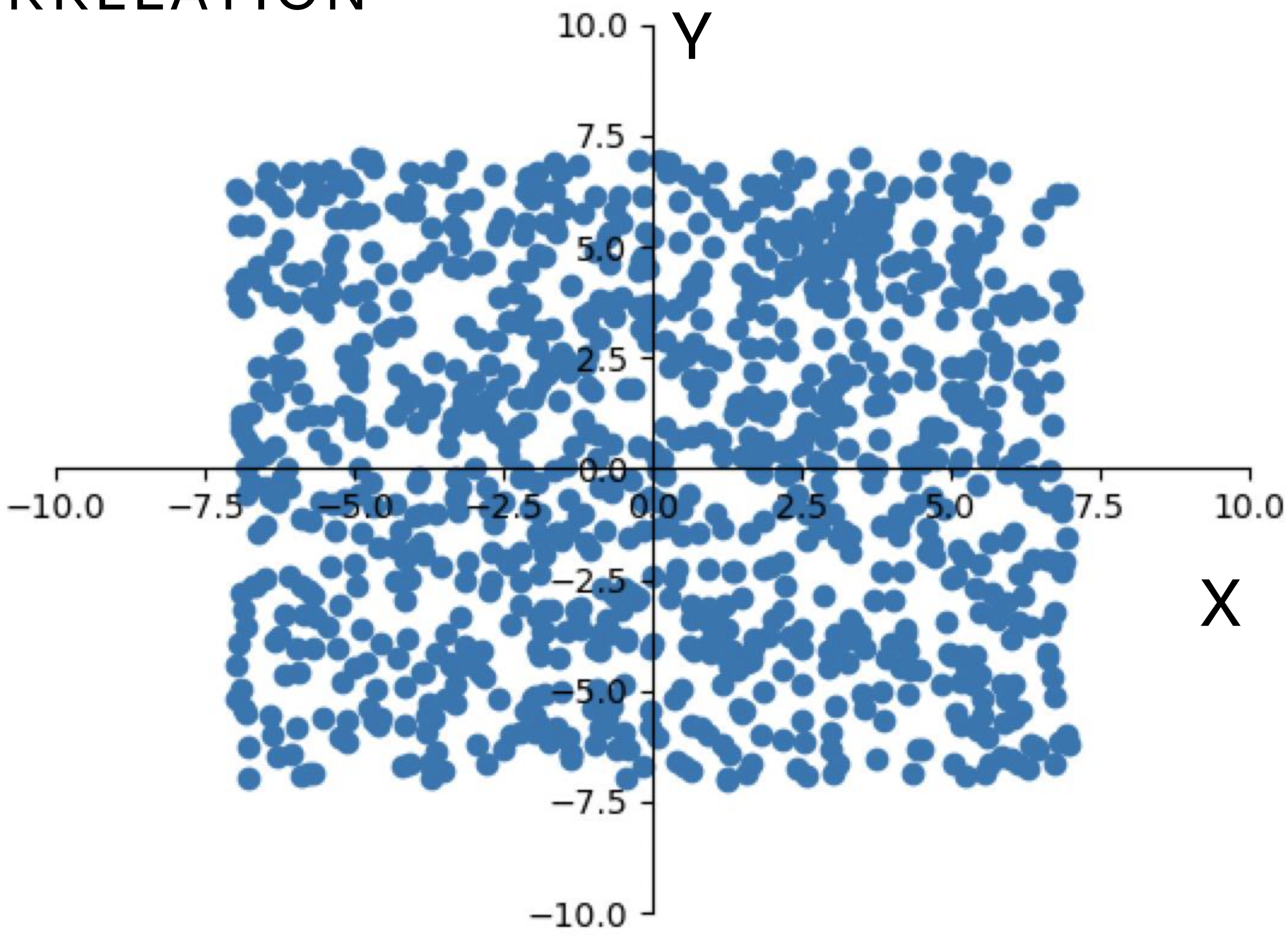



More variation in Y than X but still no correlation

# CORRELATION

There is no correlation if they do not move together.

On a Cartesian plane this appears as NO tilt to the scatter of samples.

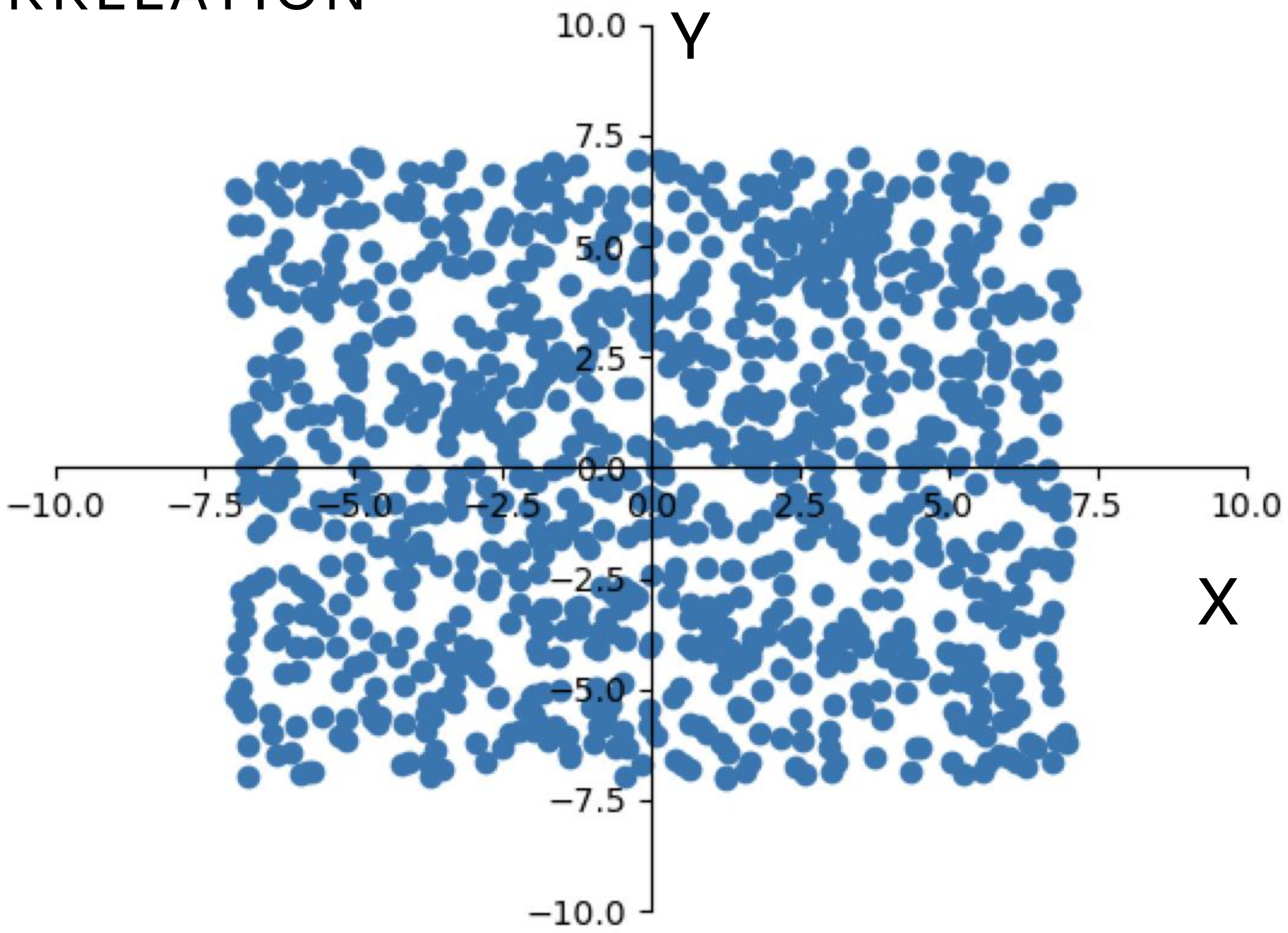



Correlated?????

# CORRELATION

There is no correlation if they do not move together.

On a Cartesian plane this appears as NO tilt to the scatter of samples.

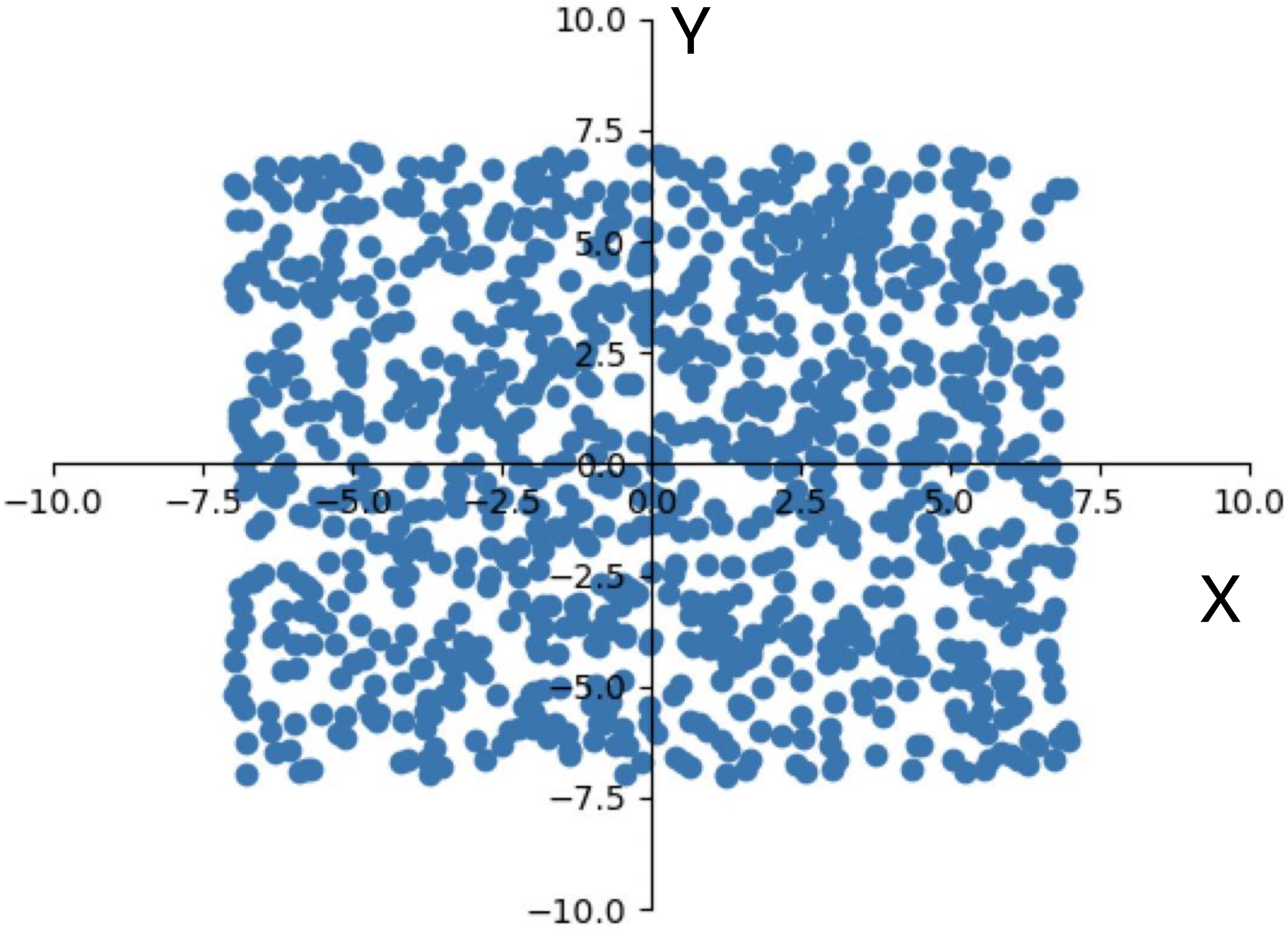



Uniformly distributed in X and Y, but they don’t move together so NO CORRELATION!

# CORRELATION

Correlation is qualitative.

We want some way to quantify correlation.




No correlation

# COVARIANCE

Let $X$ and $Y$ be two random variables, covariance is

$$cov(X,Y) = E[(X-\mu_x)(Y-\mu_Y)]$$

If $\mu_x = 0$ and $\mu_y = 0$ then this simplifies to

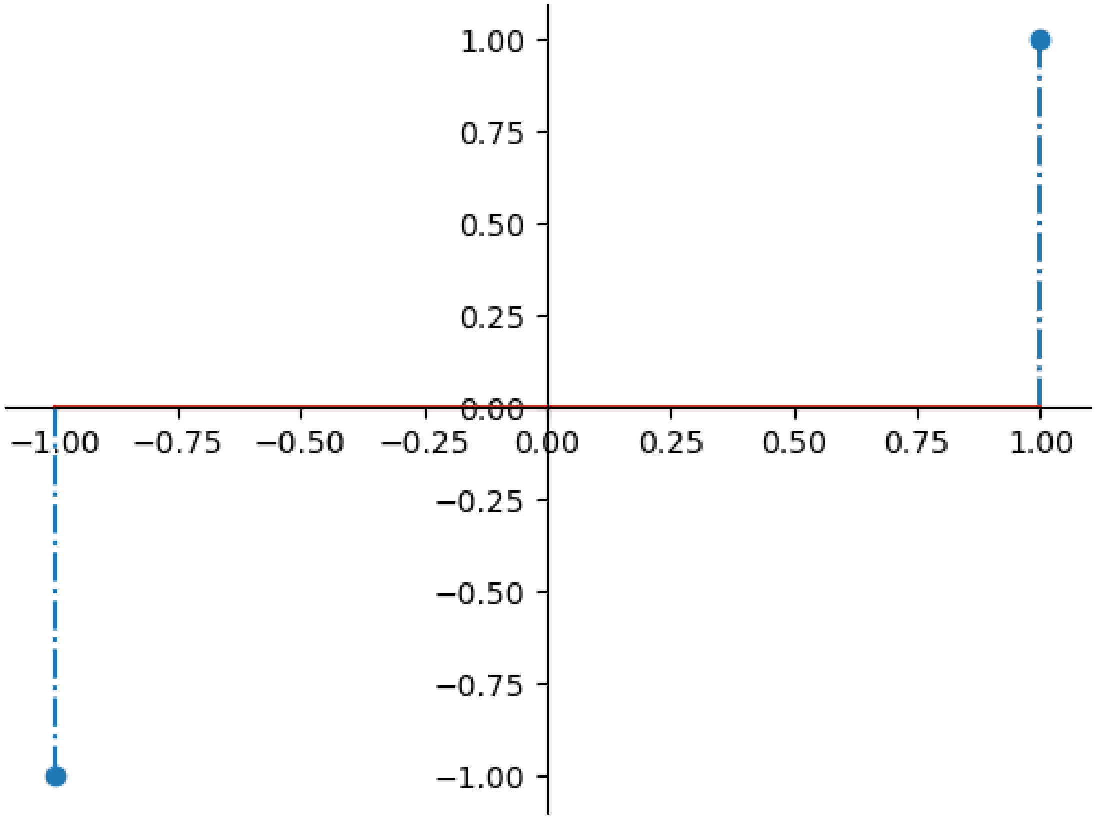
$$cov(X,Y) = E[XY]$$

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

let $S$ denote the samples drawn from $X$ and $Y$.

$$S = [(x_1,y_1),(x_2,y_2)] = [(-1,-1),(1,1)]$$

$$E[XY] = \frac{1}{2}[(1\cdot 1) + (-1\cdot -1)] = 1$$




REMINDER! Specifically chose mean = 0 to simplify equation.

# COVARIANCE

Let $X$ and $Y$ be two random variables, covariance is

$$cov(X,Y) = E[(X - \mu_x)(Y - \mu_Y)]$$

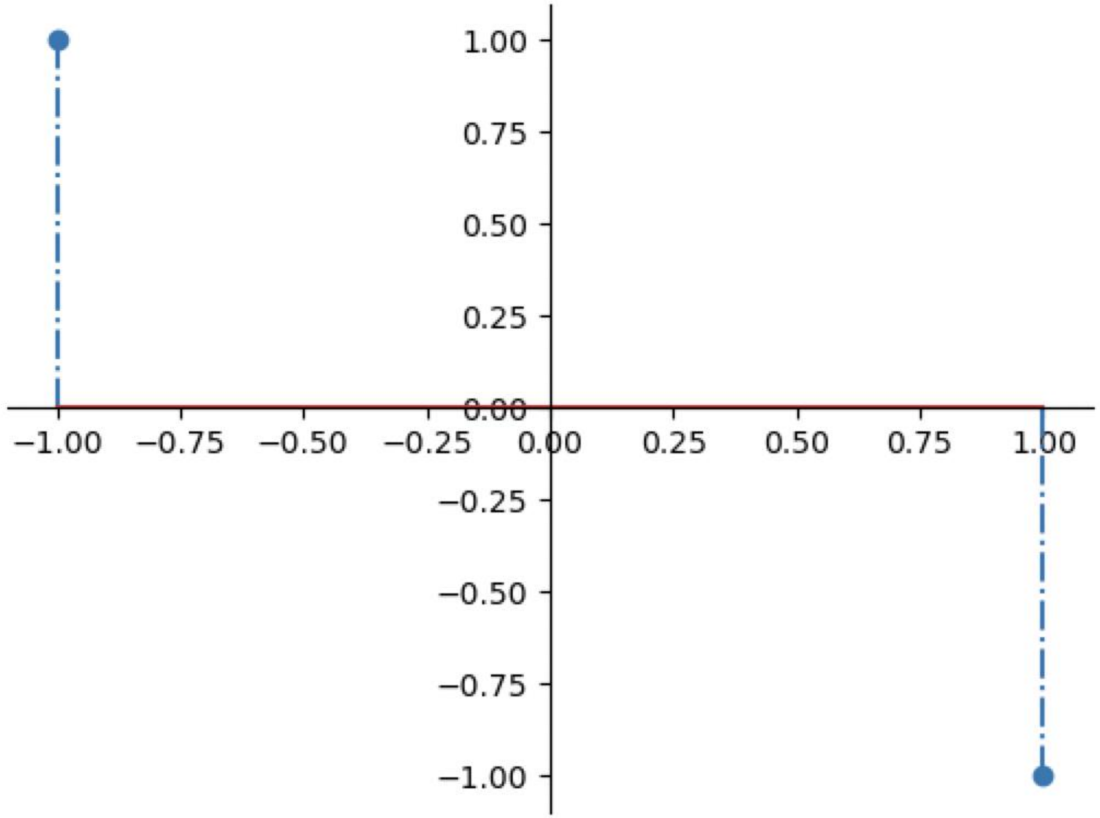
If $\mu_x = 0$ and $\mu_y = 0$ then this simplifies to

$$cov(X,Y) = E[XY]$$

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$$S = [(-1,1),(1,-1)]$$

$$E[XY] = \frac{1}{2}[(-1 \cdot 1) + (1 \cdot -1)] = -1$$




REMINDER! Specifically chose mean = 0 to simplify equation.

# COVARIANCE

Let $X$ and $Y$ be two random variables, covariance is

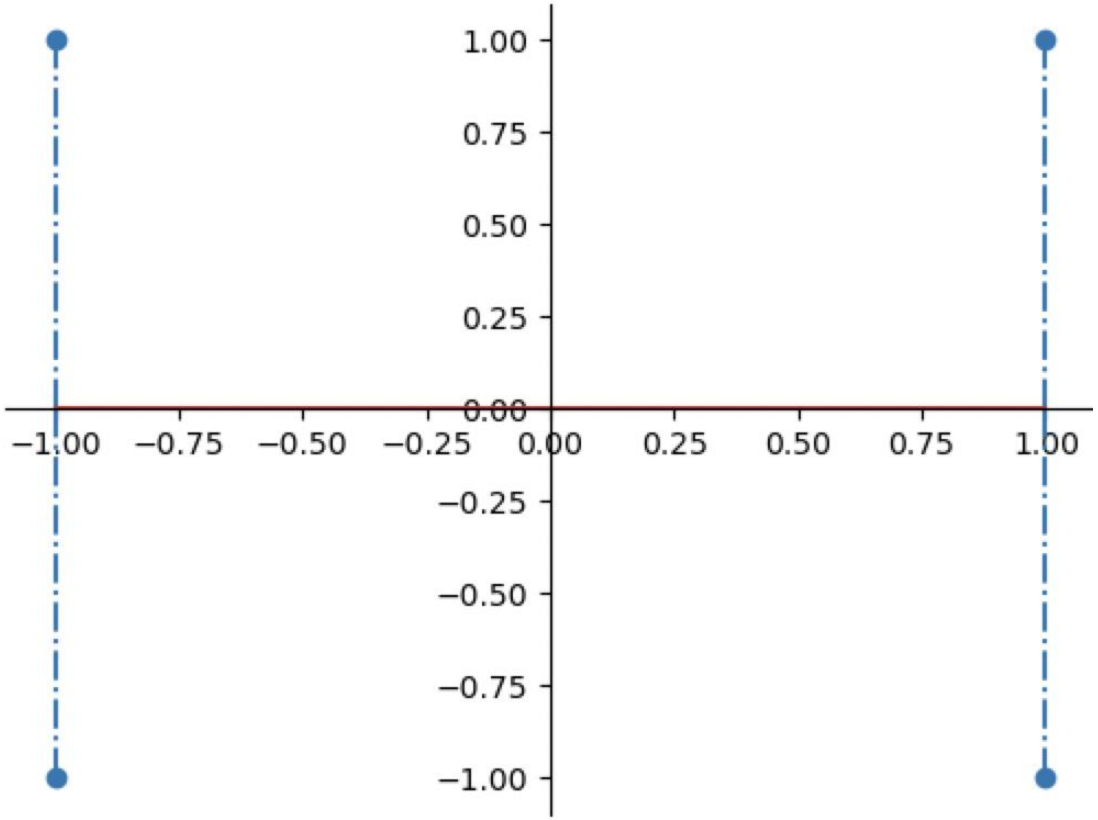
$$cov(X,Y) = E[(X - \mu_x)(Y - \mu_Y)]$$

If $\mu_x = 0$ and $\mu_y = 0$ then this simplifies to

$$cov(X,Y) = E[XY]$$

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$$S = [(-1,1),(-1,-1),(1,1),(1,-1)]$$




$$E[XY] = \frac{1}{4}[(-1\cdot 1) + (-1\cdot -1) + (1\cdot 1) + (1\cdot -1)] = 0$$

REMINDER! Specifically chose mean = 0 to simplify equation.

# COVARIANCE

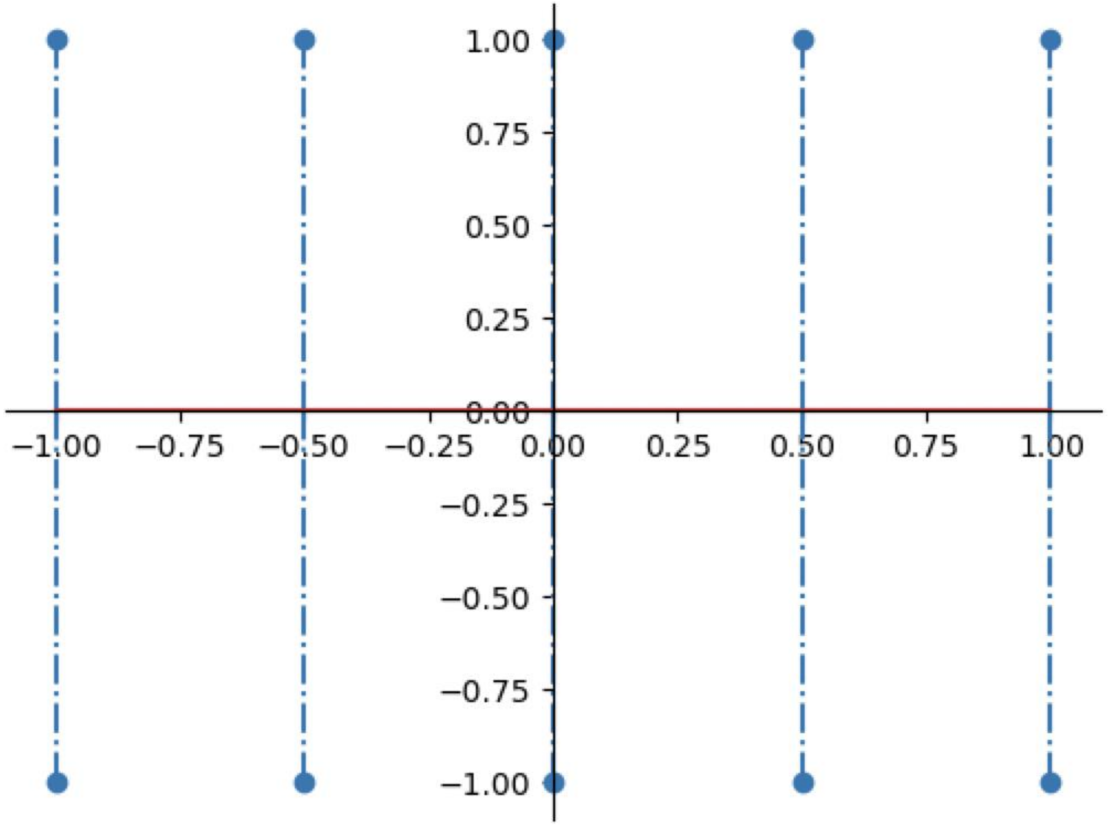
Let $X$ and $Y$ be two random variables, covariance is

$$cov(X,Y) = E[(X-\mu_x)(Y-\mu_Y)]$$

If $\mu_x = 0$ and $\mu_y = 0$ then this simplifies to

$$cov(X,Y) = E[XY]$$

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$




$$E[XY] = \frac{1}{10}[(-1\cdot-1)+(-1\cdot 1)+(-\frac{1}{2}\cdot\frac{1}{2})+(-\frac{1}{2}\cdot-\frac{1}{2})+\cdots+(1\cdot 1)+(1\cdot-1)] = 0$$

# COVARIANCE

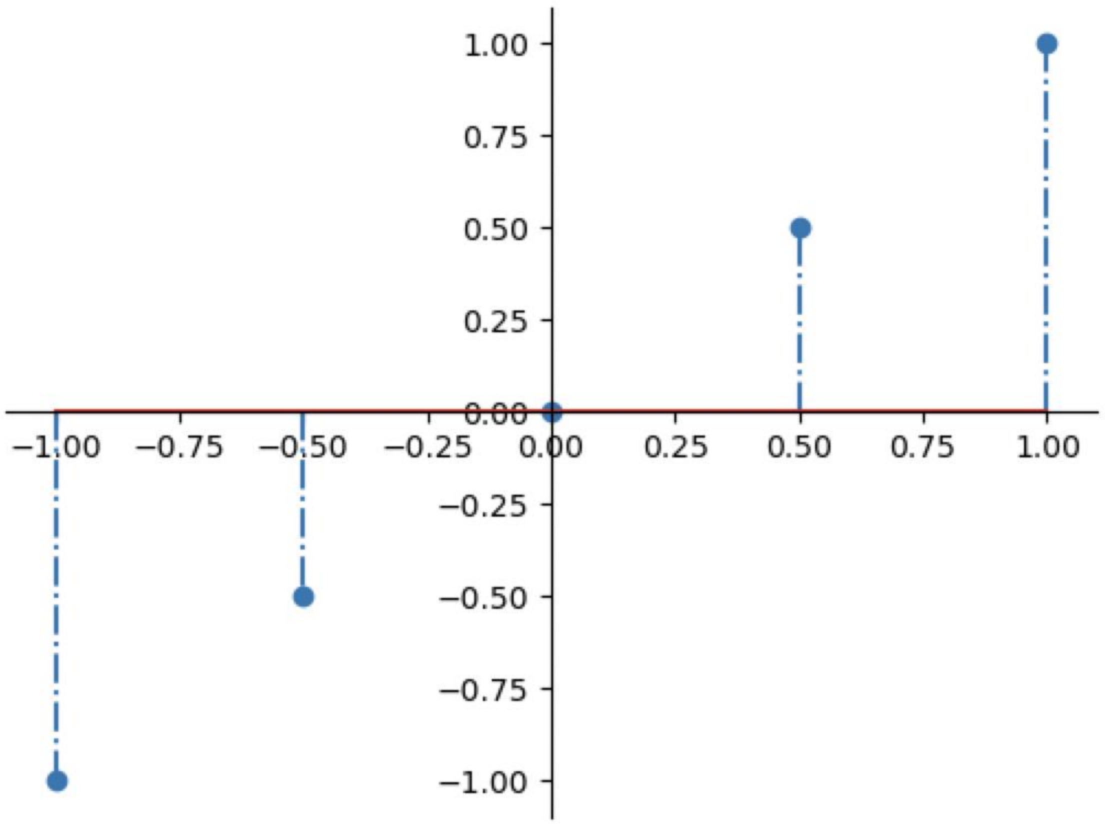
Let $X$ and $Y$ be two random variables, covariance is

$$cov(X,Y) = E[(X-\mu_x)(Y-\mu_Y)]$$

If $\mu_x = 0$ and $\mu_y = 0$ then this simplifies to

$$cov(X,Y) = E[XY]$$

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$




$$E[XY] = \frac{1}{5}[(-1\cdot-1)+(-\frac{1}{2}\cdot-\frac{1}{2})+(0\cdot0)+(\frac{1}{2}\cdot\frac{1}{2})+(1\cdot1)] = ???$$

# COVARIANCE

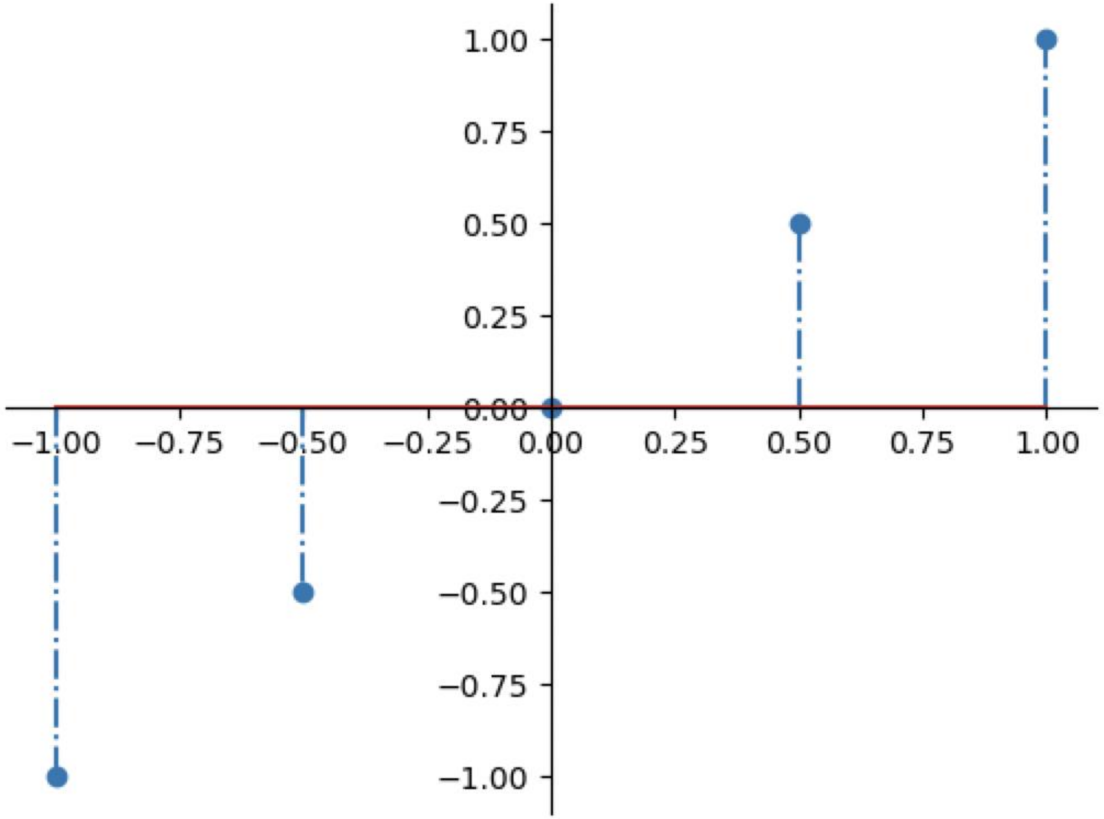
Let $X$ and $Y$ be two random variables, covariance is

$$cov(X,Y) = E[(X-\mu_x)(Y-\mu_Y)]$$

If $\mu_x = 0$ and $\mu_y = 0$ then this simplifies to

$$cov(X,Y) = E[XY]$$

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$




$$E[XY] = \frac{1}{5}[(-1\cdot-1)+(-\frac{1}{2}\cdot-\frac{1}{2})+(0\cdot0)+(\frac{1}{2}\cdot\frac{1}{2})+(1\cdot1)] = \frac{2.5}{5} = \frac{1}{2}$$

# COVARIANCE

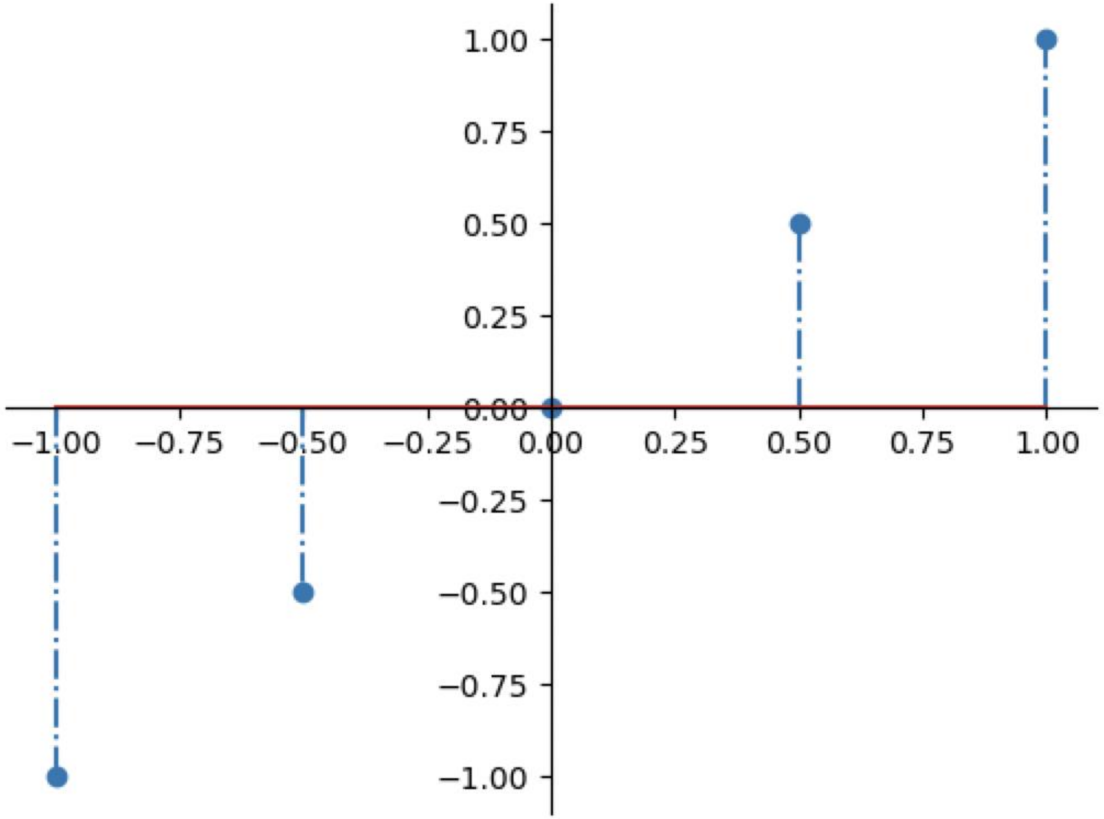
Let $X$ and $Y$ be two random variables, covariance is

$$cov(X,Y) = E[(X - \mu_x)(Y - \mu_Y)]$$

If $\mu_x = 0$ and $\mu_y = 0$ then this simplifies to

$$cov(X,Y) = E[XY]$$

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$




$$E[XY] = \frac{1}{5}[(-1 \cdot -1) + (-\frac{1}{2} \cdot -\frac{1}{2}) + (0 \cdot 0) + (\frac{1}{2} \cdot \frac{1}{2}) + (1 \cdot 1)] = \frac{2.5}{5} = \frac{1}{2}$$

Why? Why not 1?

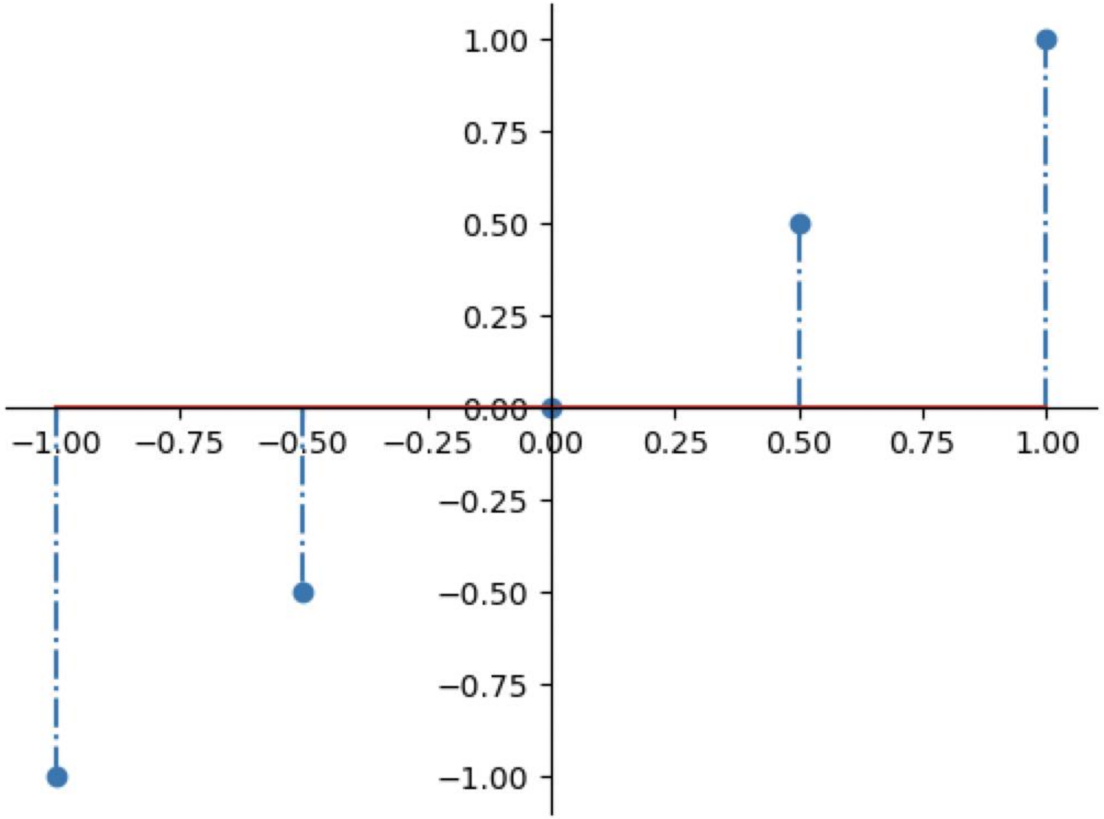
# COVARIANCE

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

When X and Y are equal, they square.

The impact of a sample grows with the square of the distance from the mean (here mean is 0).

Numbers farther out have greater impact on E[XY].




$$E[XY] = \frac{1}{5}[(-1\cdot-1)+(-\frac{1}{2}\cdot-\frac{1}{2})+(0\cdot0)+(\frac{1}{2}\cdot\frac{1}{2})+(1\cdot1)] = \frac{2.5}{5} = \frac{1}{2}$$

# COVARIANCE

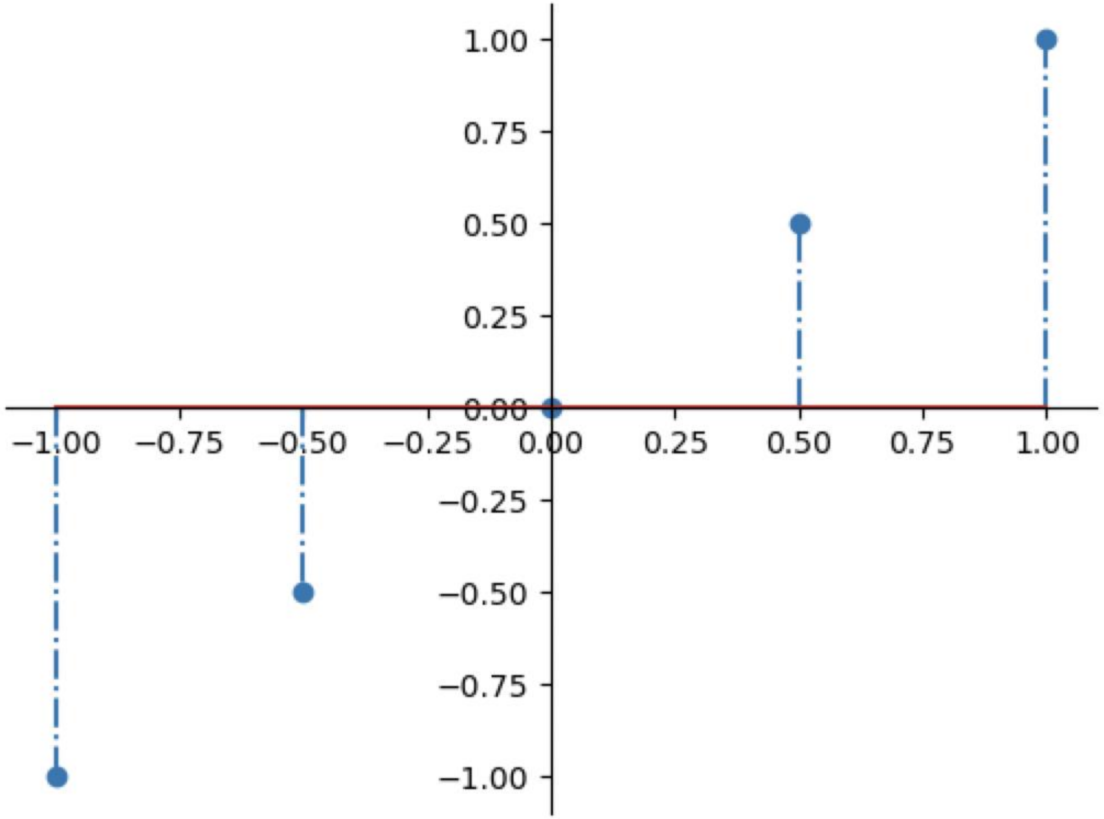
$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

Variance ALSO grows with the square.

When the mean is zero

$$Var[X] = E[X^2]$$

$$\frac{E[XY]}{Var[X]} = ???$$




$$E[XY] = \frac{\frac{1}{5}[(-1\cdot -1) + (-\frac{1}{2}\cdot -\frac{1}{2}) + (0\cdot 0) + (\frac{1}{2}\cdot\frac{1}{2}) + (1\cdot 1)]}{\frac{1}{5}[(-1\cdot -1) + (-\frac{1}{2}\cdot -\frac{1}{2}) + (0\cdot 0) + (\frac{1}{2}\cdot\frac{1}{2}) + (1\cdot 1)]} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1$$

# COVARIANCE

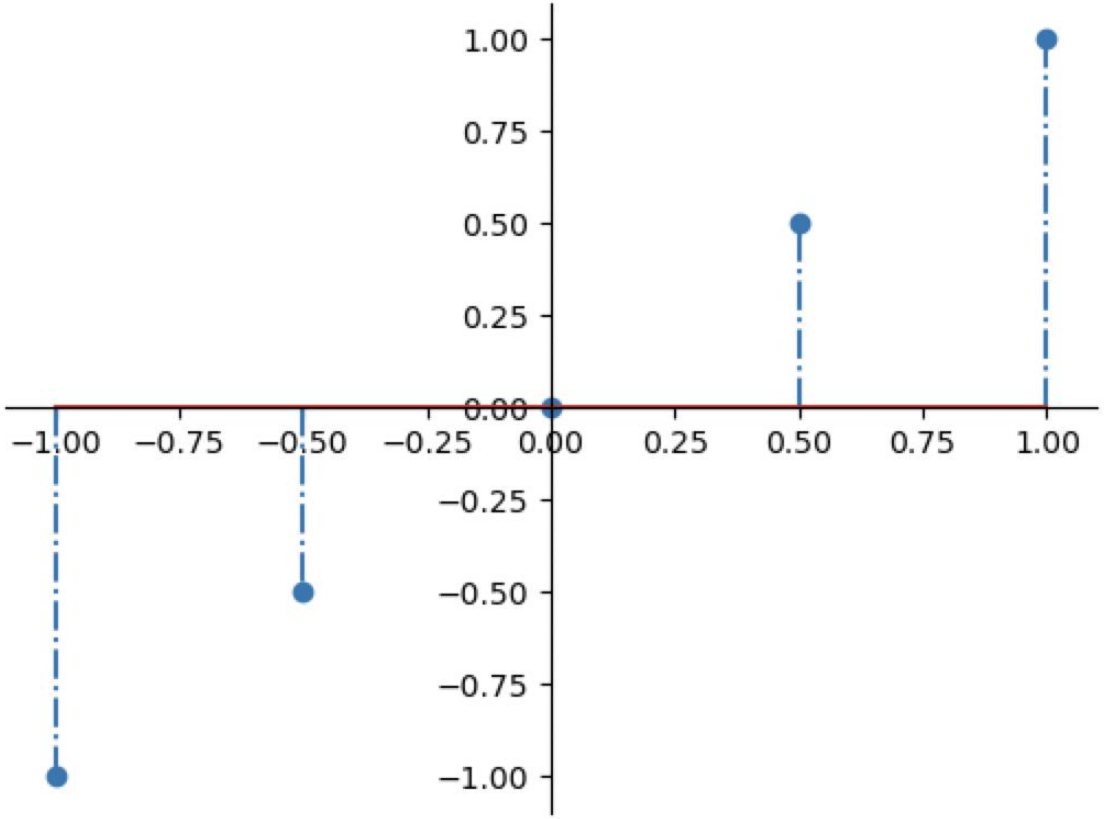
$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

Variance ALSO grows with the square.

When the mean is zero

$$Var[X] = E[X^2]$$

$$\frac{E[XY]}{Var[X]} = ???$$




But why only Var[X]? Shouldn't the variation in Y also matter?

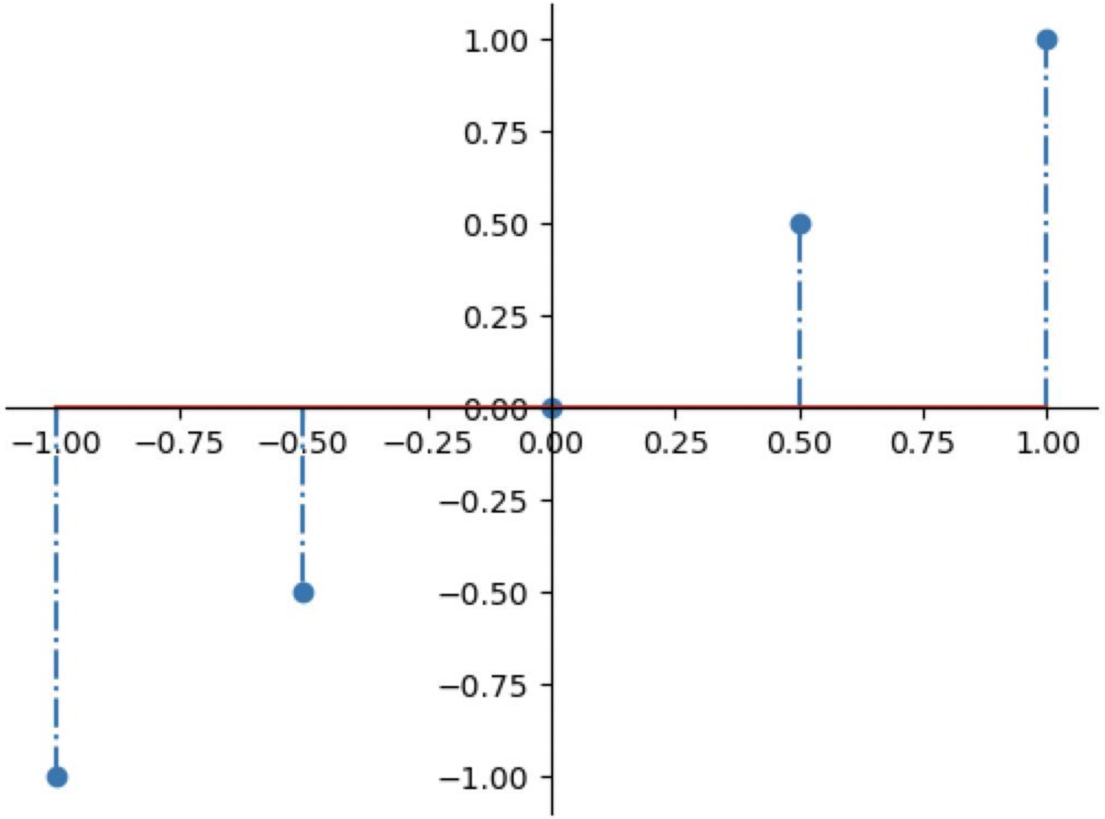
# COVARIANCE

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

Variance ALSO grows with the square.

When the mean is zero

$$Var[X] = E[X^2]$$




$$\frac{E[XY]}{\sqrt{Var[X]}\sqrt{Var[Y]}} = \frac{E[XY]}{\sigma_X \sigma_Y} = \frac{\frac{1}{2}}{\sqrt{\frac{1}{2}}\sqrt{\frac{1}{2}}} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1$$

# COVARIANCE

We can adjust the equations to take into account non-zero mean.

$$cov(X,Y) = E[(X - \mu_x)(Y - \mu_Y)]$$

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \quad \Rightarrow \quad \frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

$$\frac{E[XY]}{\sigma_X \sigma_Y} \quad \Rightarrow \quad \frac{\frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_X \sigma_Y}$$

# COVARIANCE

We can adjust the equations to take into account non-zero mean.

$$cov(X,Y) = E[(X-\mu_x)(Y-\mu_Y)]$$

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \quad \Rightarrow \quad \frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

$$\frac{E[XY]}{\sigma_X \sigma_Y} \quad \Rightarrow \quad \frac{\frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_X \sigma_Y}$$

Pearson Correlation

# PEARSON CORRELATION

*Correlation Coefficient*

a.k.a.,

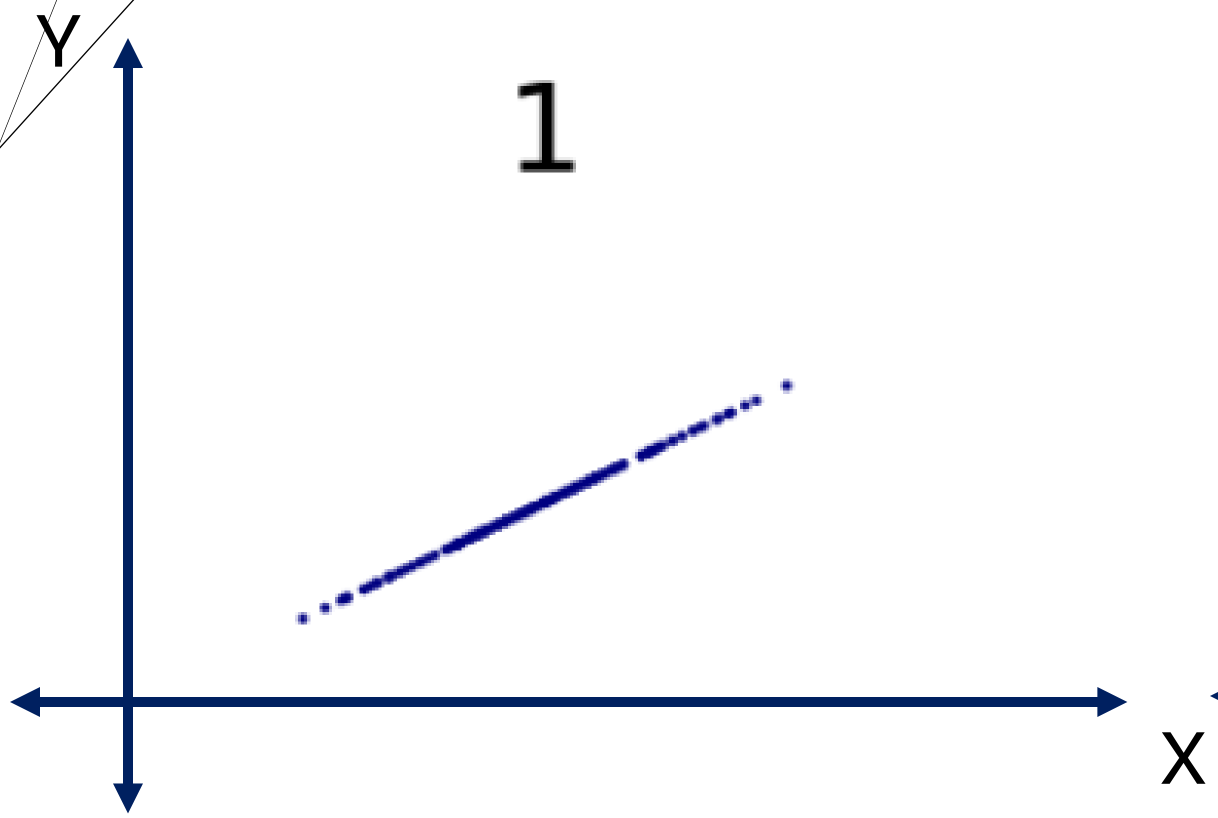
*Linear Correlation Coefficient*.

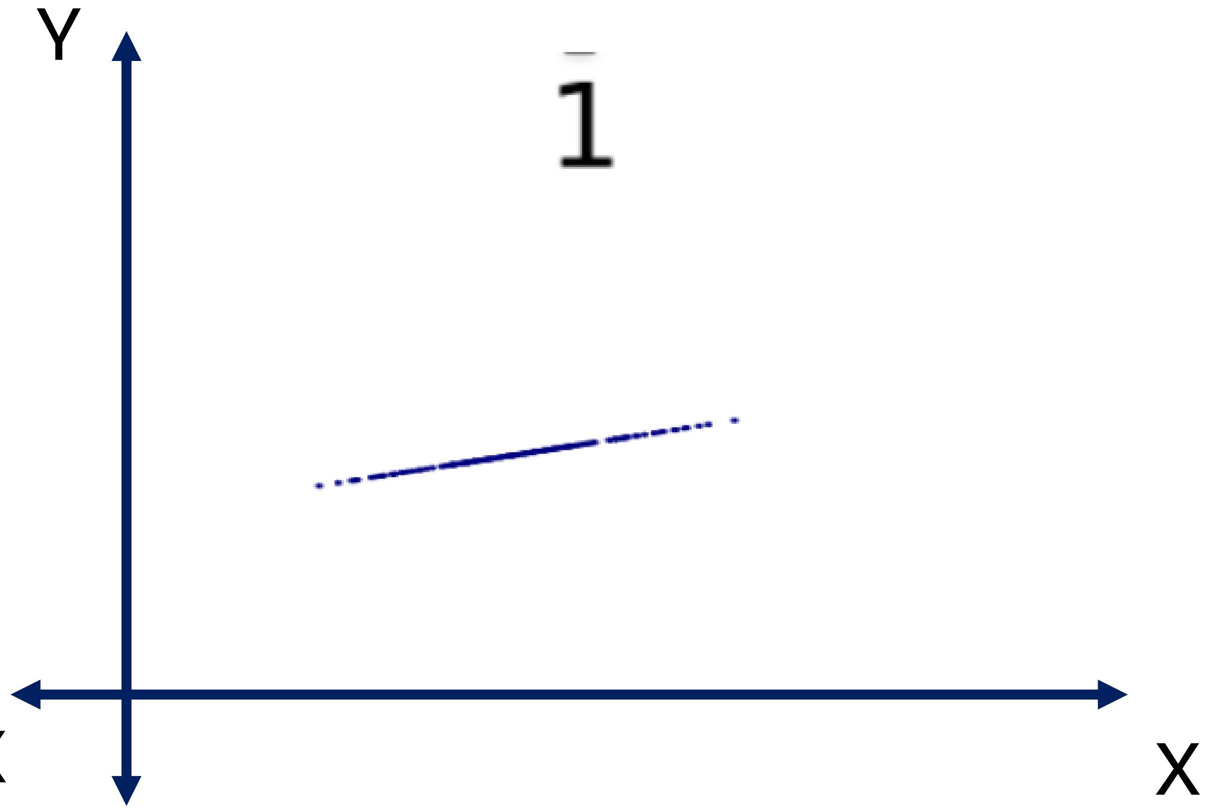
a.k.a.,

the *Pearson Correlation Coefficient*.

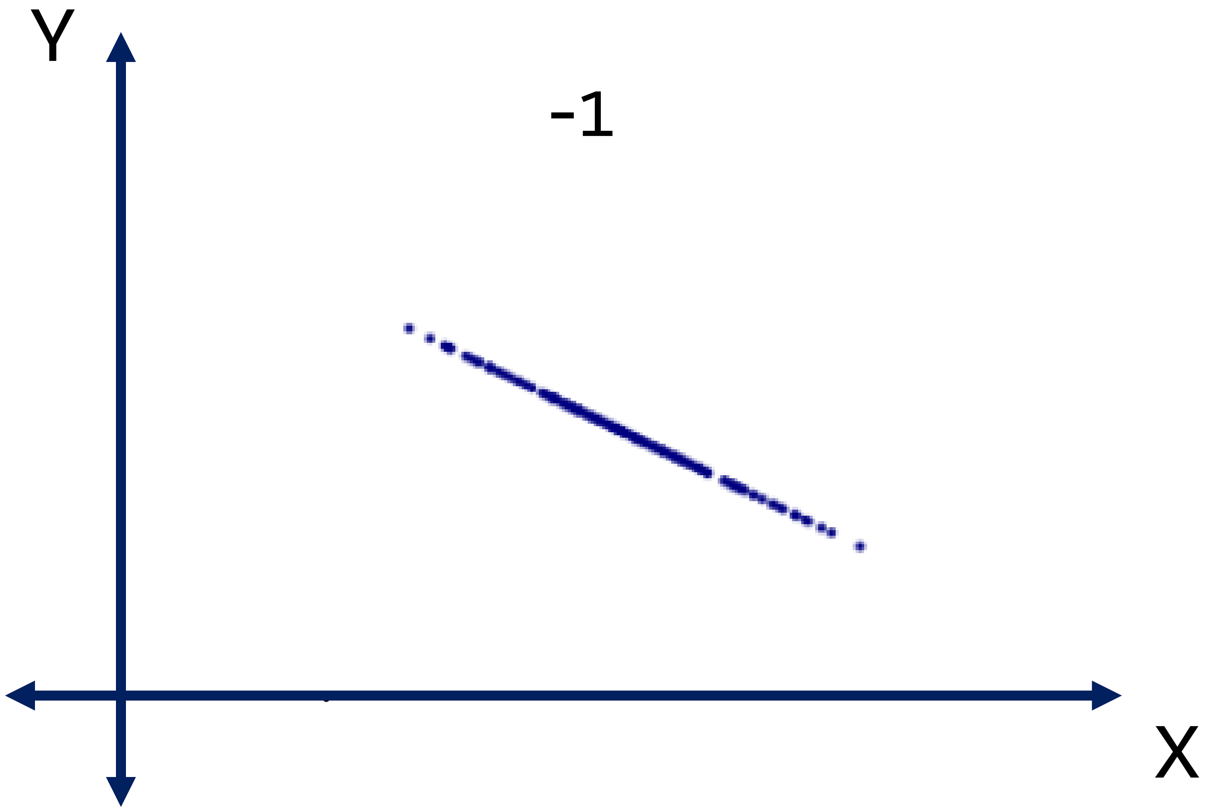
$$r = \frac{\frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_X \sigma_Y}$$

# PEARSON CORRELATION



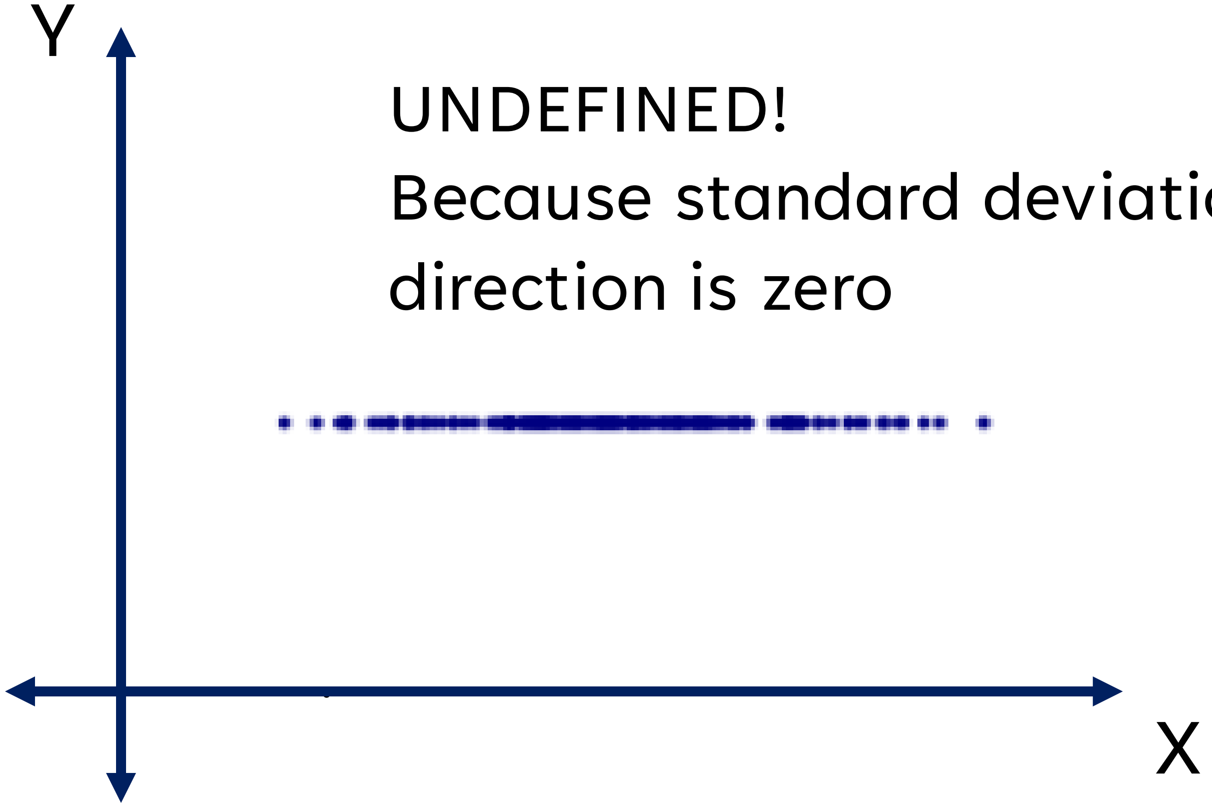
# PEARSON CORRELATION



# PEARSON CORRELATION

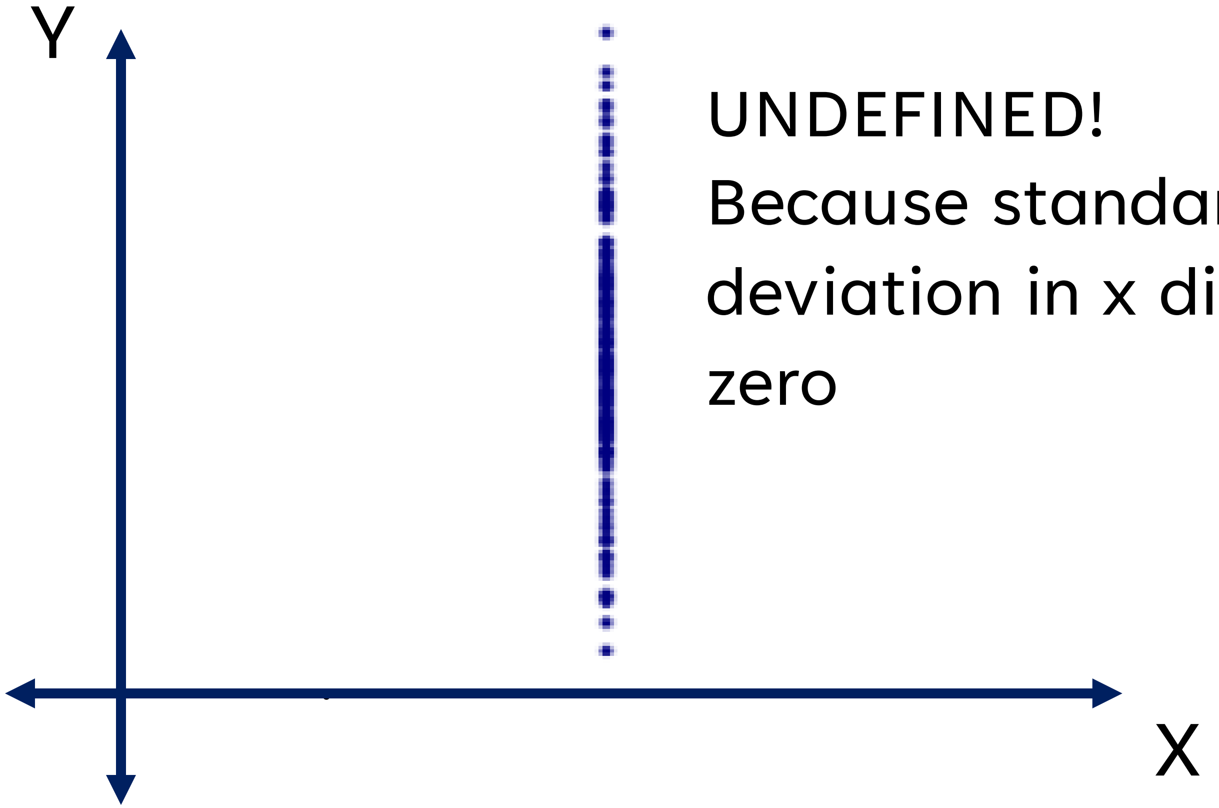
UNDEFINED!

Because standard deviation in y direction is zero




$$r = \frac{\frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_X \sigma_Y}$$

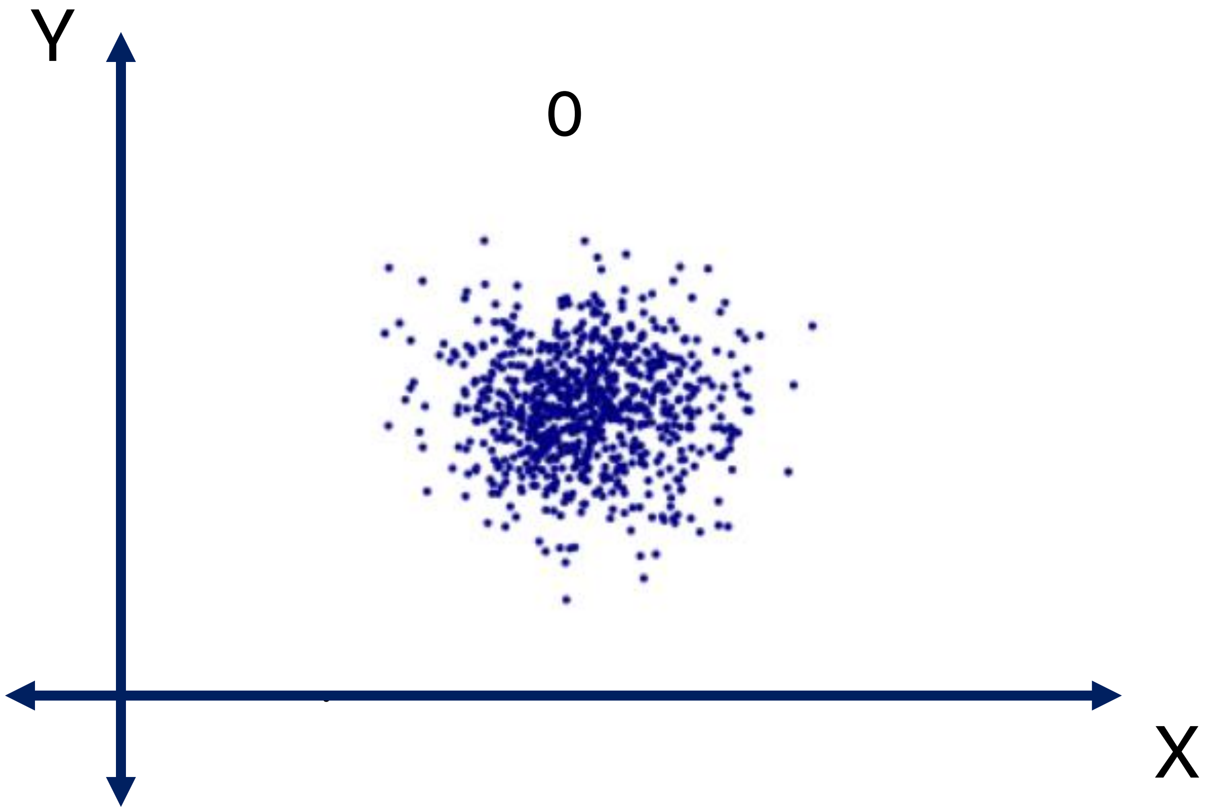
# PEARSON CORRELATION



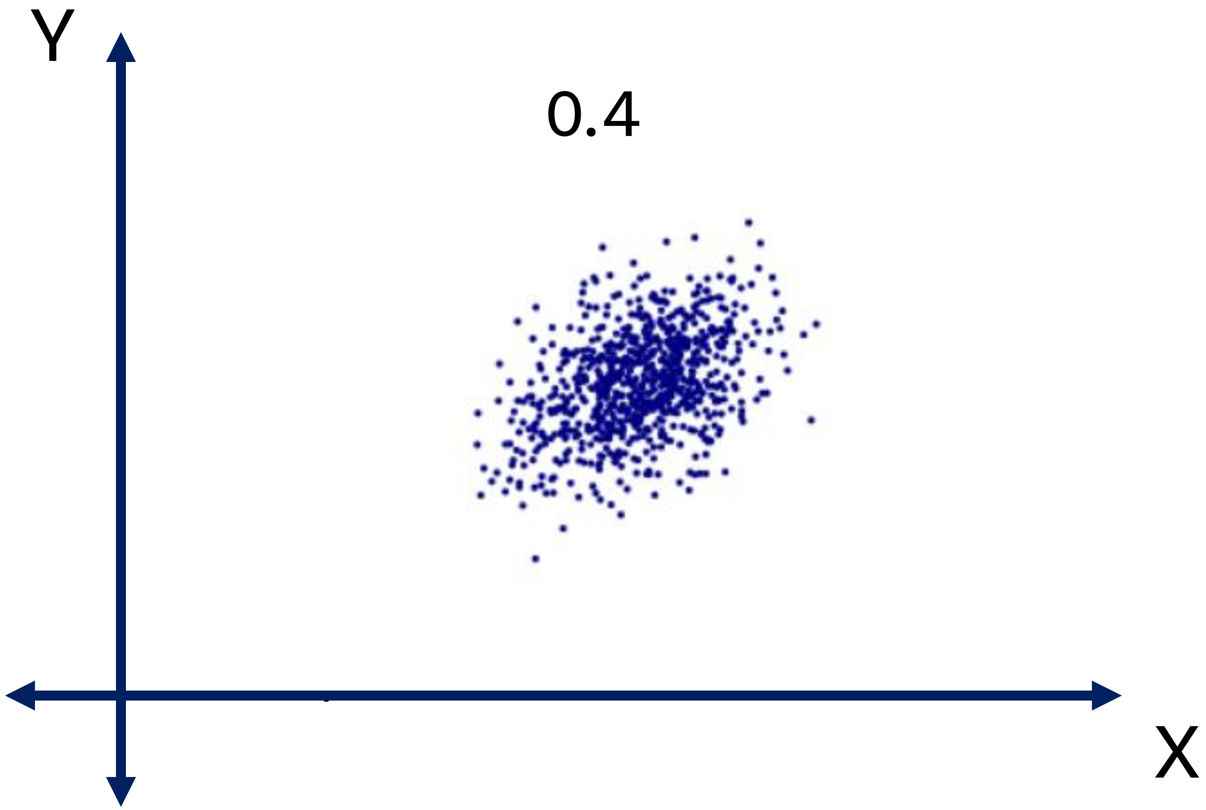

UNDEFINED!
Because standard deviation in x direction is zero

$$r = \frac{\frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_X \sigma_Y}$$

# PEARSON CORRELATION



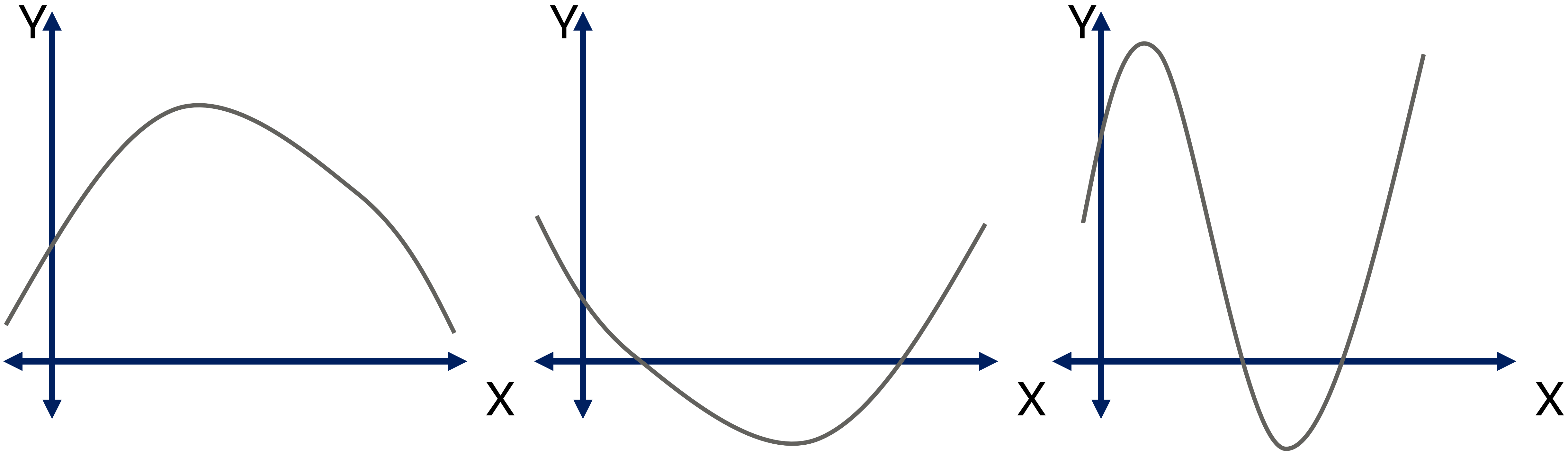
# PEARSON CORRELATION



# DEPENDENCE

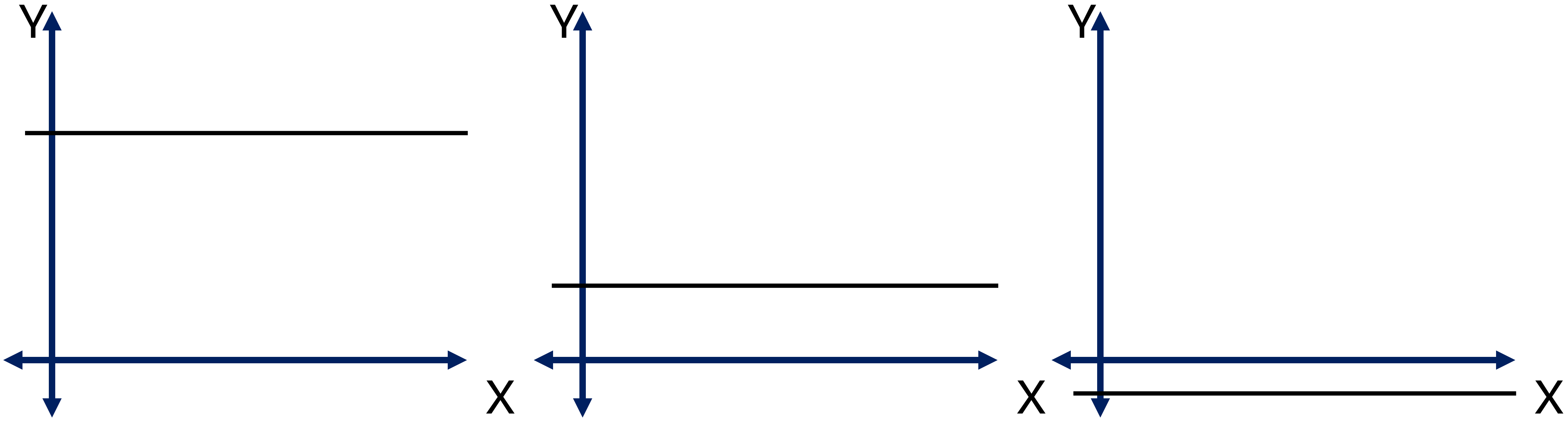
A variable y is dependent on another variable x if y=f(x).

Meaning y is a function of x.

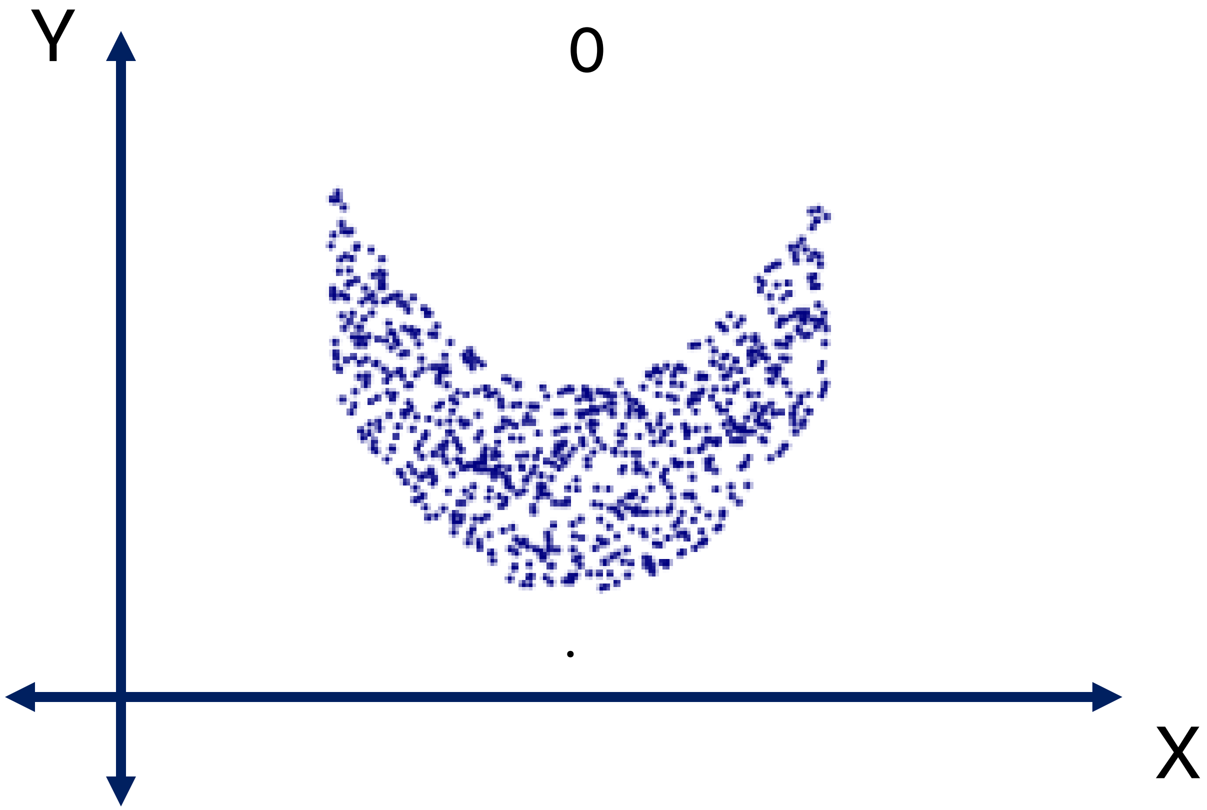


# INDEPENDENCE

A variable y is *independent* of x if y remains constant as x changes.



# CORRELATION



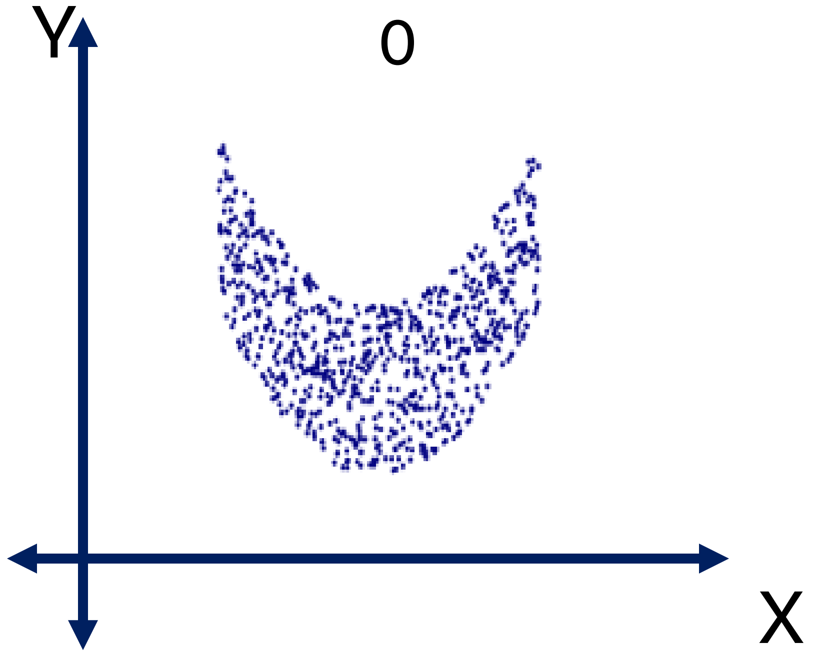

PEARSON CORRELATION COEFFICIENT ONLY CAPTURES LINEAR RELATIONSHIPS

Y is dependent on X, but has zero Pearson Correlation

# CORRELATION VS. DEPENDENCE VS. INDEPENDENCE

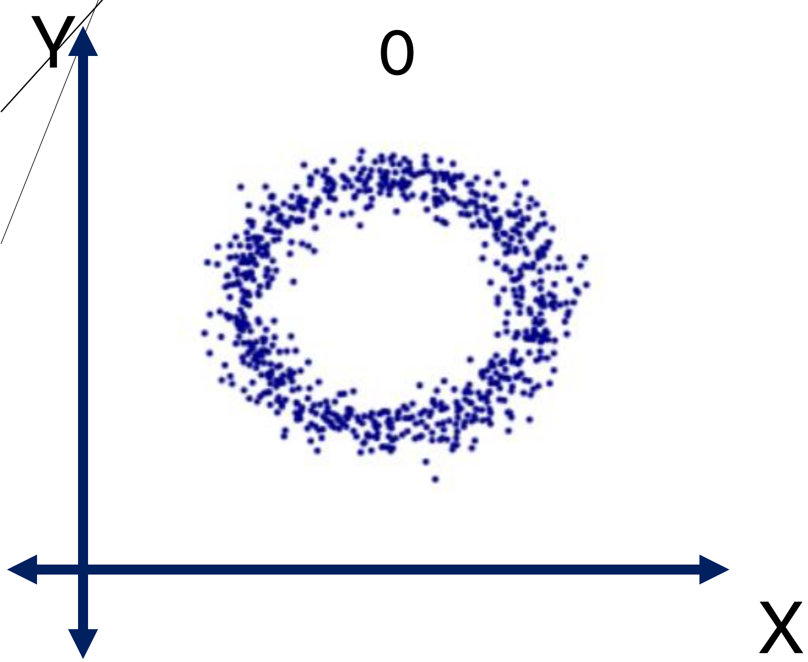
If two random variables are linearly correlated then they are dependent.

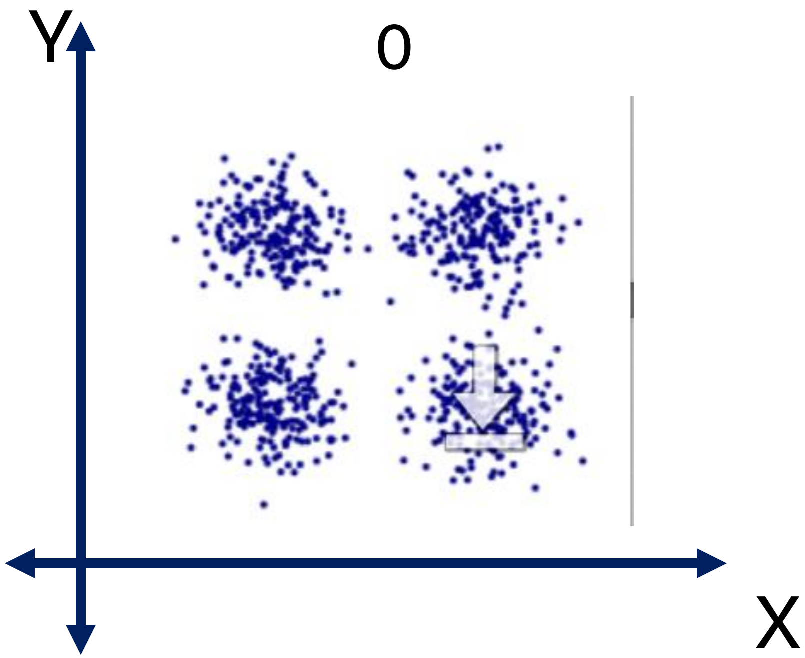
If two random variables are related in a non-linear way, they may have zero correlation and yet still be dependent!

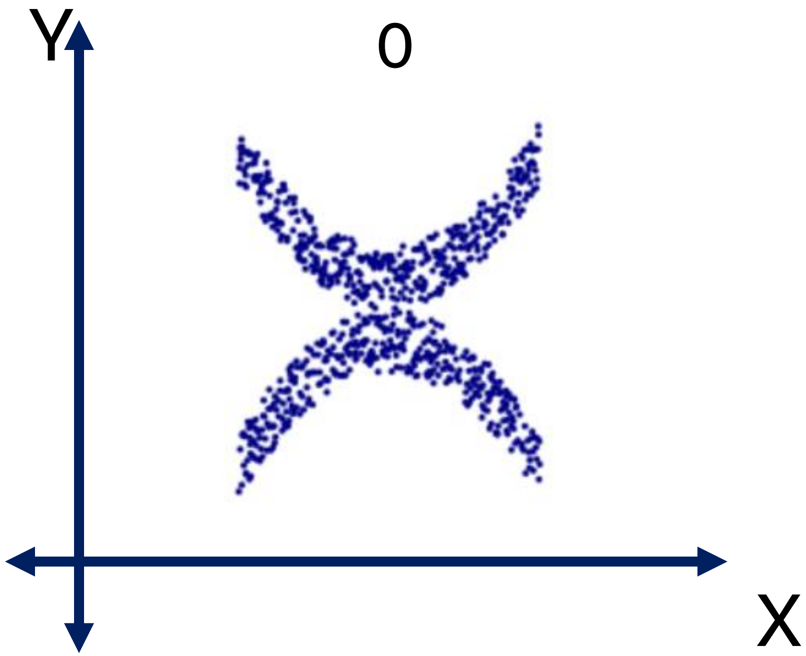


# ALL DEPENDENT BUT NOT LINEARLY CORRELATED

•



# THANK YOU

David Harrison

Harrison@cs.olemiss.edu