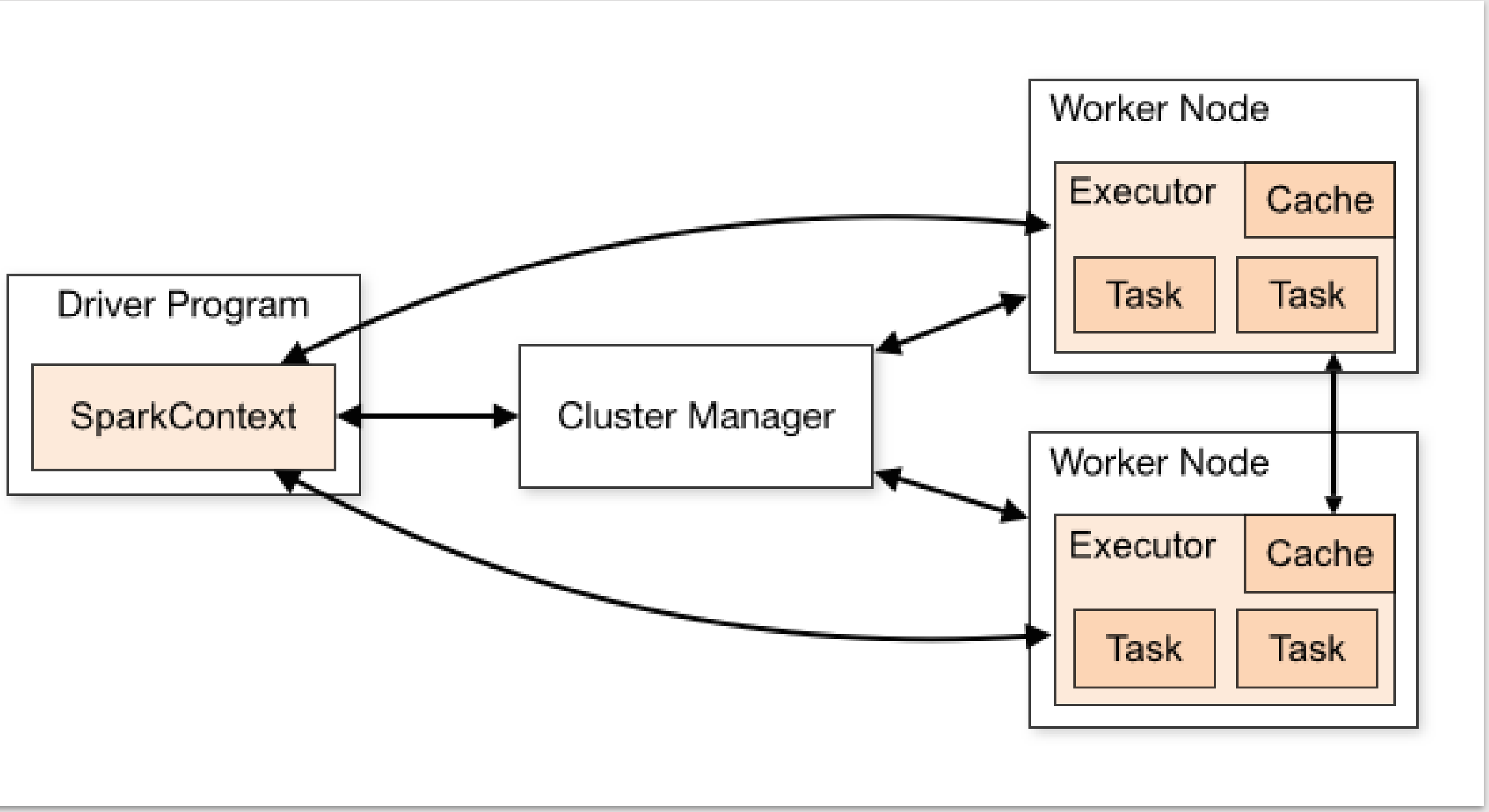
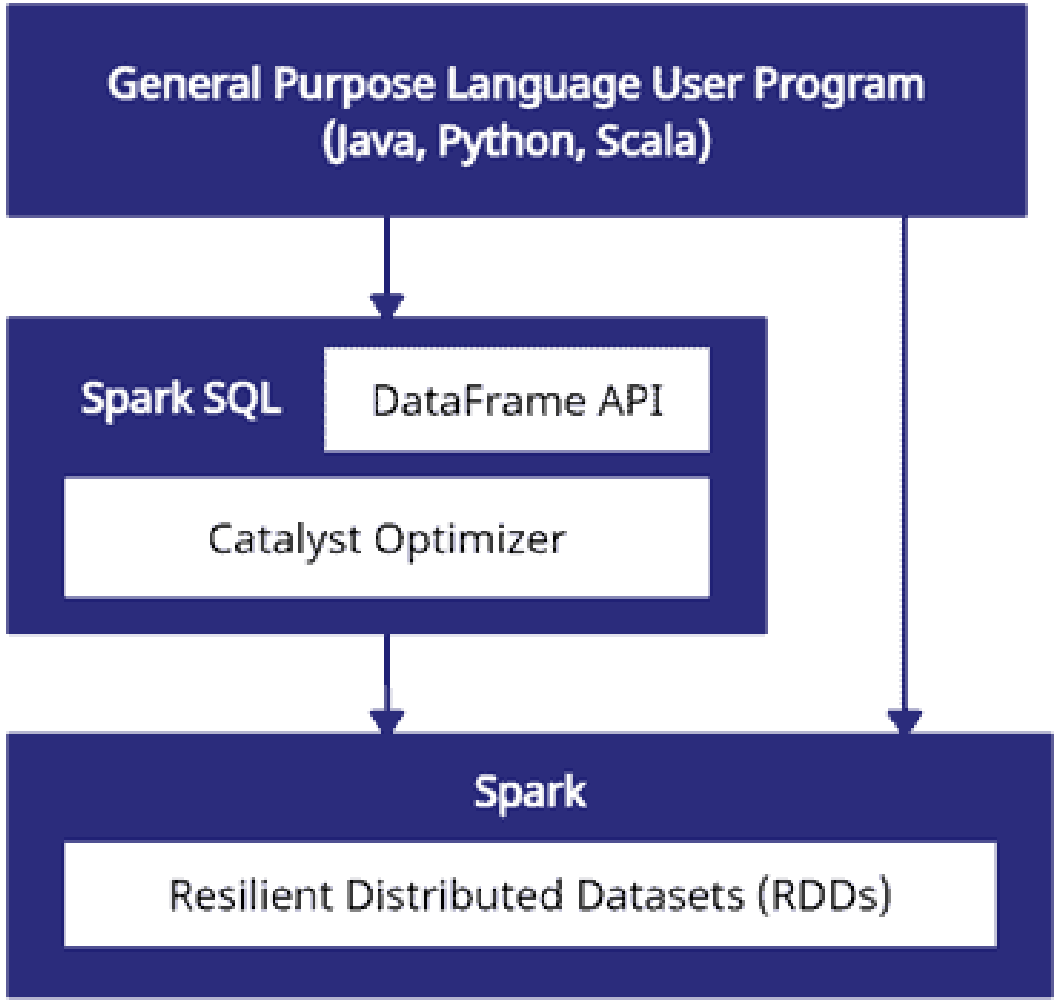
APACHE SPARK

ÍNDICE

- **Spark SQL**
- **MLlib**
  - Introducción a Machine Learning
  - Clasificaciones y Regresiones
  - Regresión: Regresión Lineal
  - Clasificación: Regresión Logística, Árboles de Decisión y Random Forest
  - Clustering: Kmeans
  - Sistemas de recomendación
- **Streaming**
- **GraphX**

RRD: Ciclo de Vida

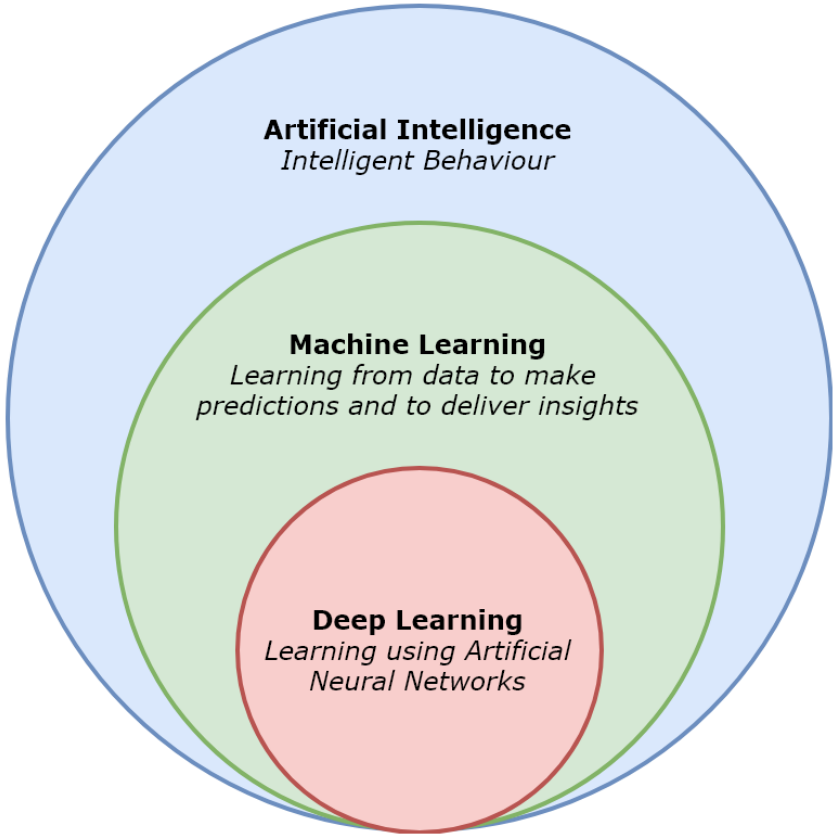

Worker Node
Executor
Cache
Task
Task
Driver Program
SparkContext
Cluster Manager
Worker Node
Executor
Cache
Task
Task


General Purpose Language User Program
(Java, Python, Scala)
Spark SQL
DataFrame API
Catalyst Optimizer
Spark
Resilient Distributed Datasets (RDDs)


'drop'
];
events.forEach(e => {
fileDropZone.addEventListener(e, (ev) => {
ev.preventDefault();
if (ev.type === 'dragenter') {
fileDropZone.classList.add('solid-border');
}
if (ev.type === 'dragleave') {
fileDropZone.classList.remove('solid-border');
}
if(ev.type === 'drop') {
fileDropZone.classList.remove('solid-border');
ev.dataTransfer.files)
'border rounded

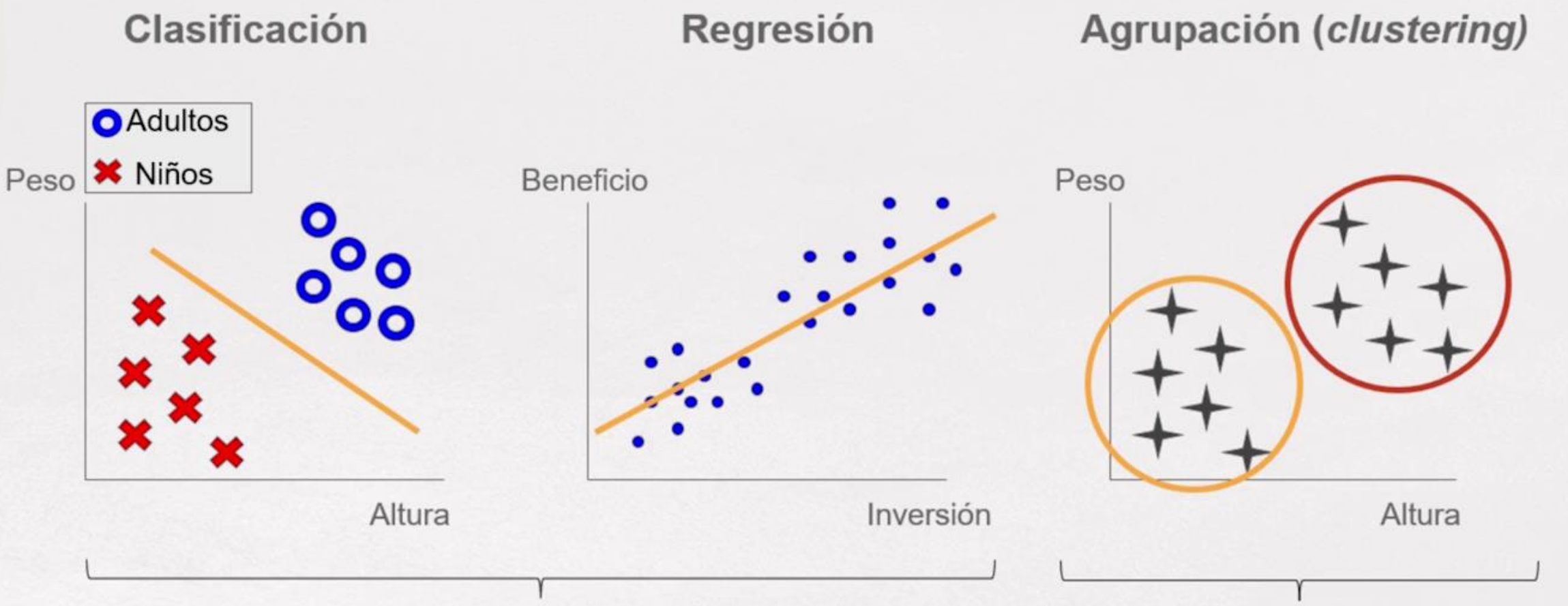
NOTEBOOK

MACHINE LEARNING

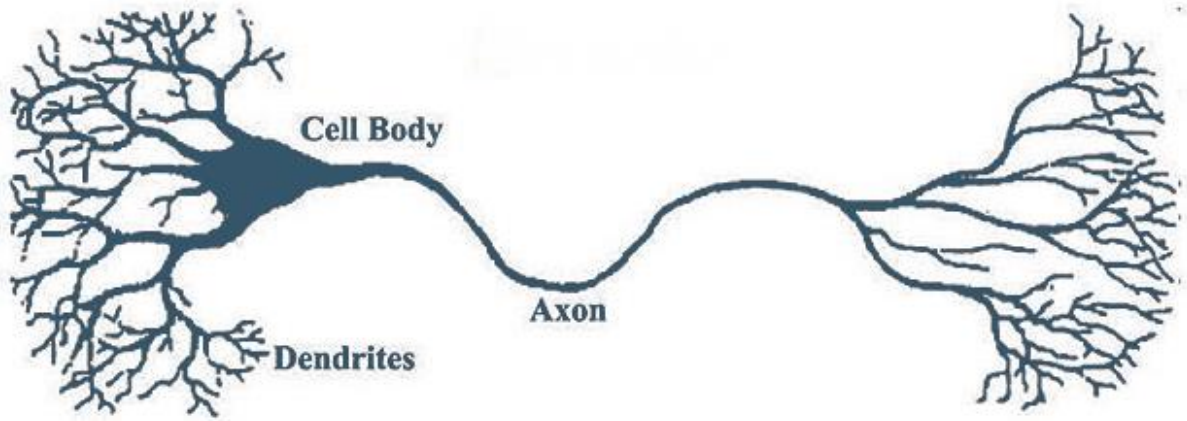


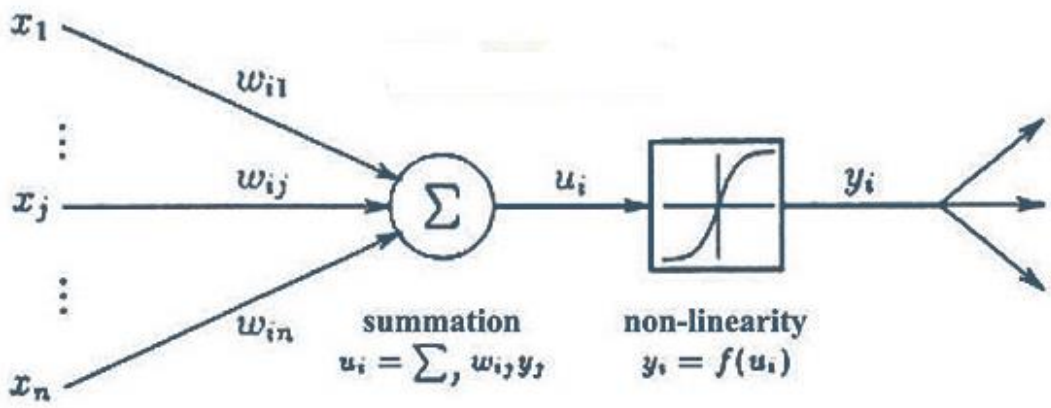
TIPOS DE MACHINE LEARNING

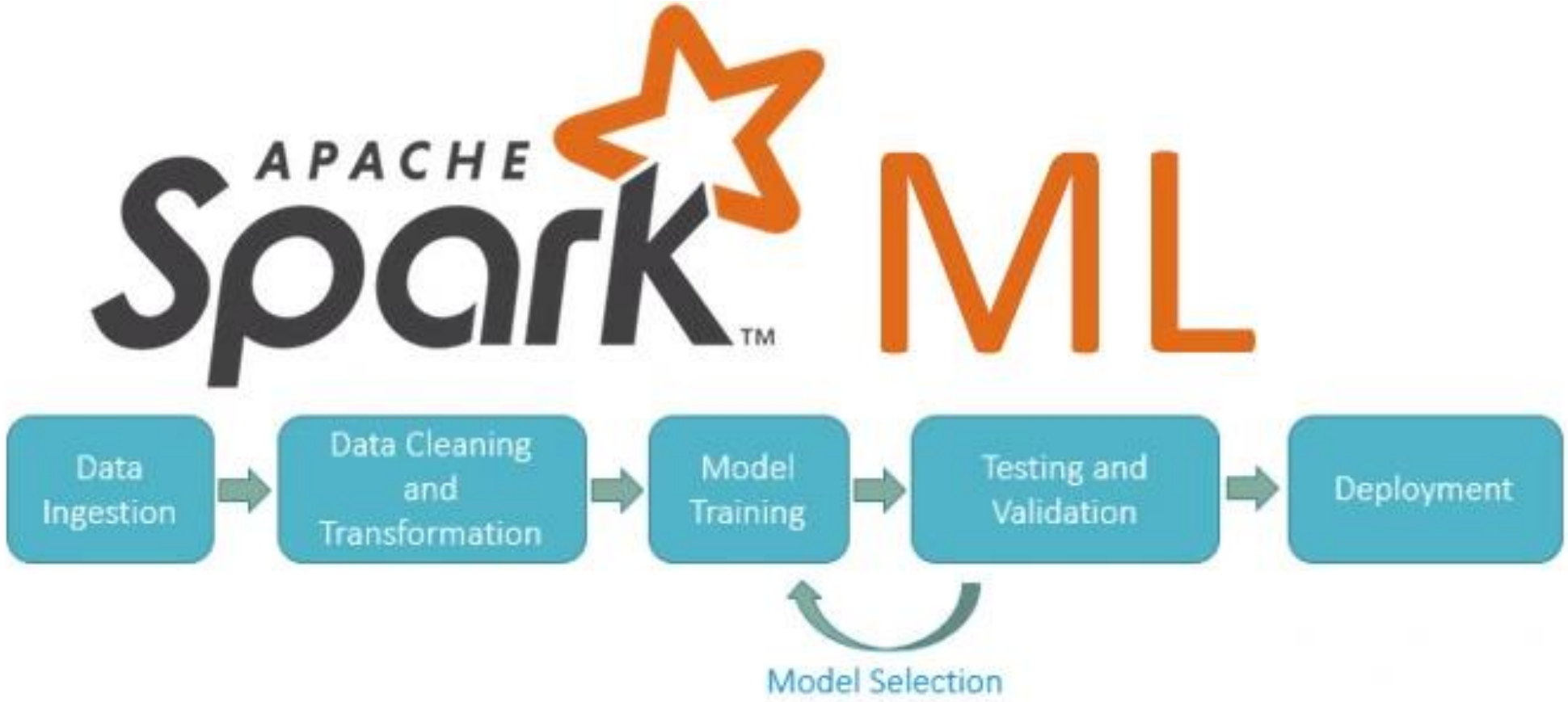
Técnicas de Machine Learning



REDES NEURONALES

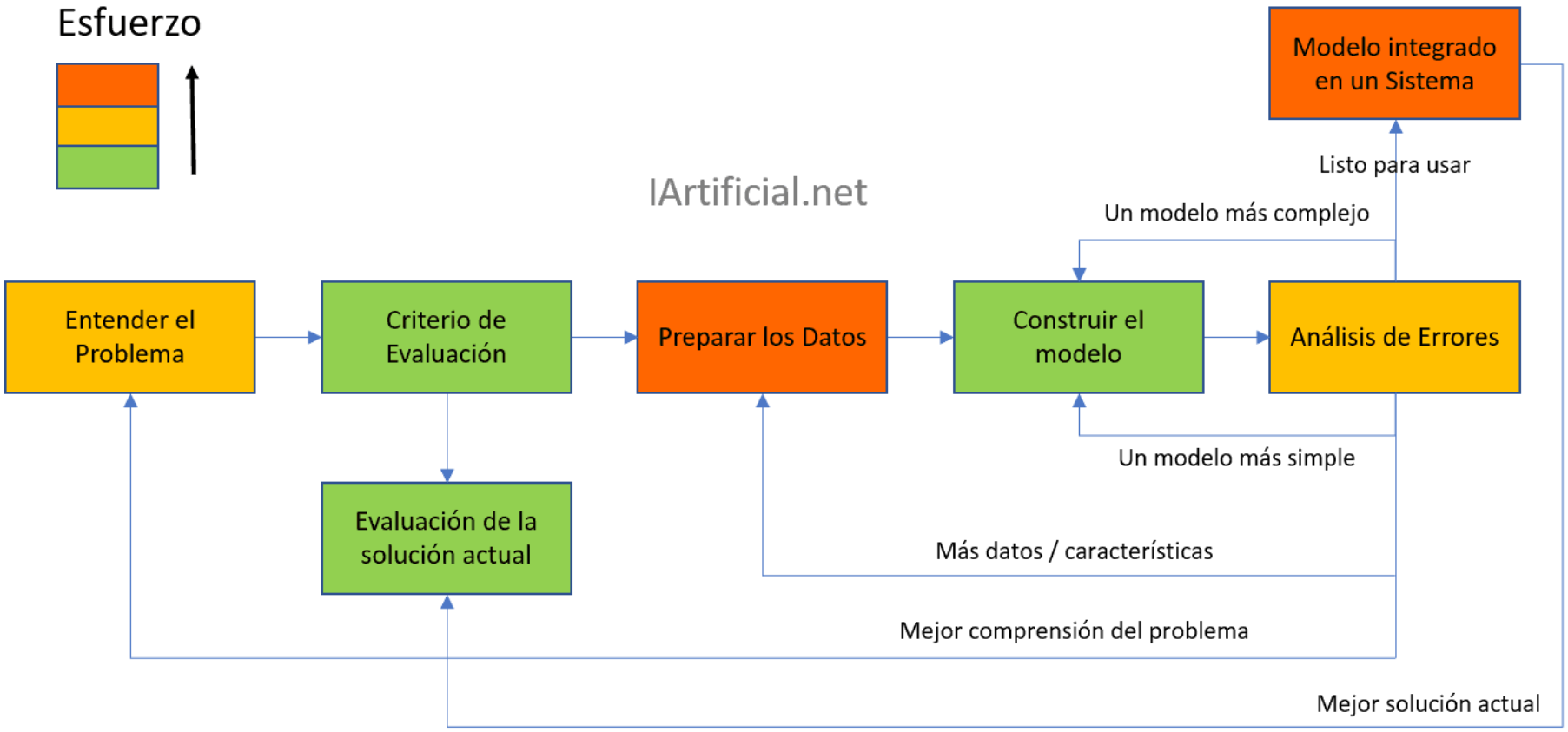


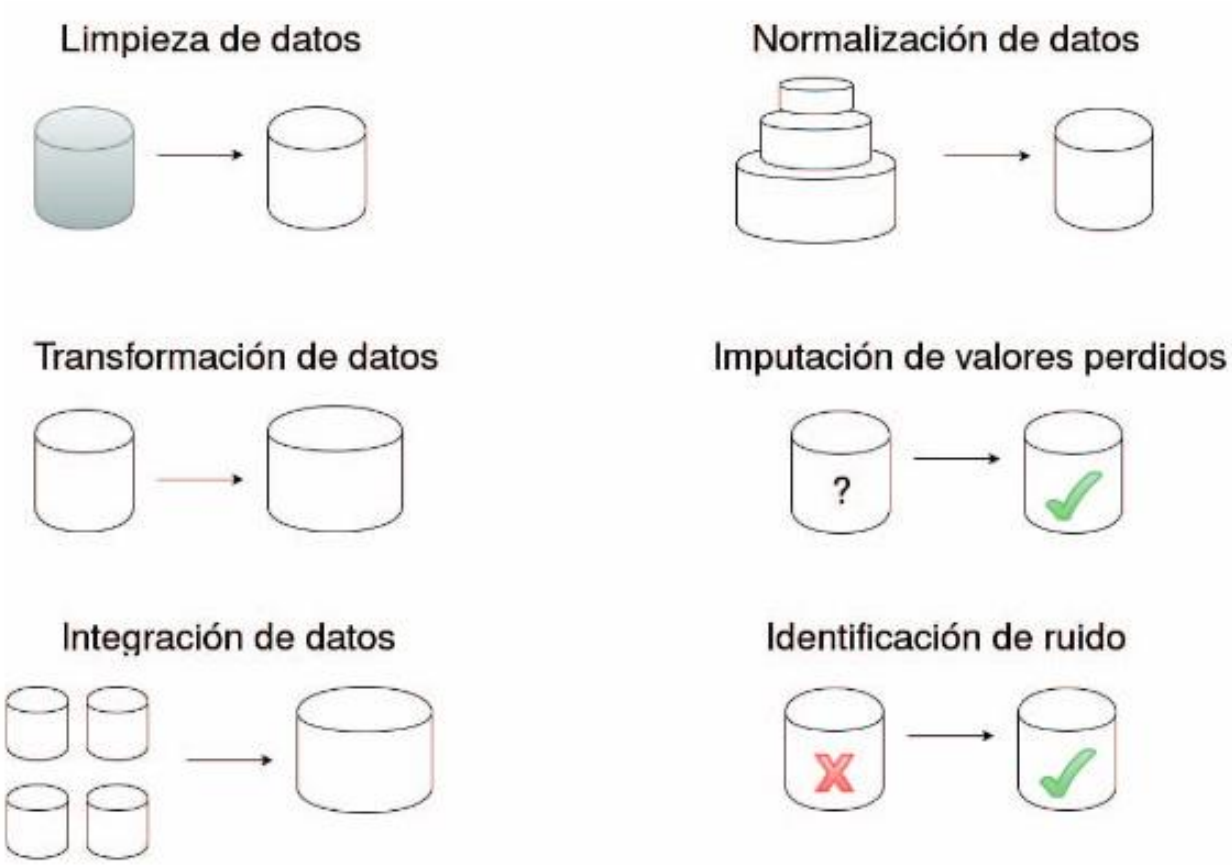



APACHE
Spark™
ML
Data Ingestion
Data Cleaning and Transformation
Model Training
Testing and Validation
Deployment
Model Selection

ML LIB

PASOS DE UN PROYECTO DE ML

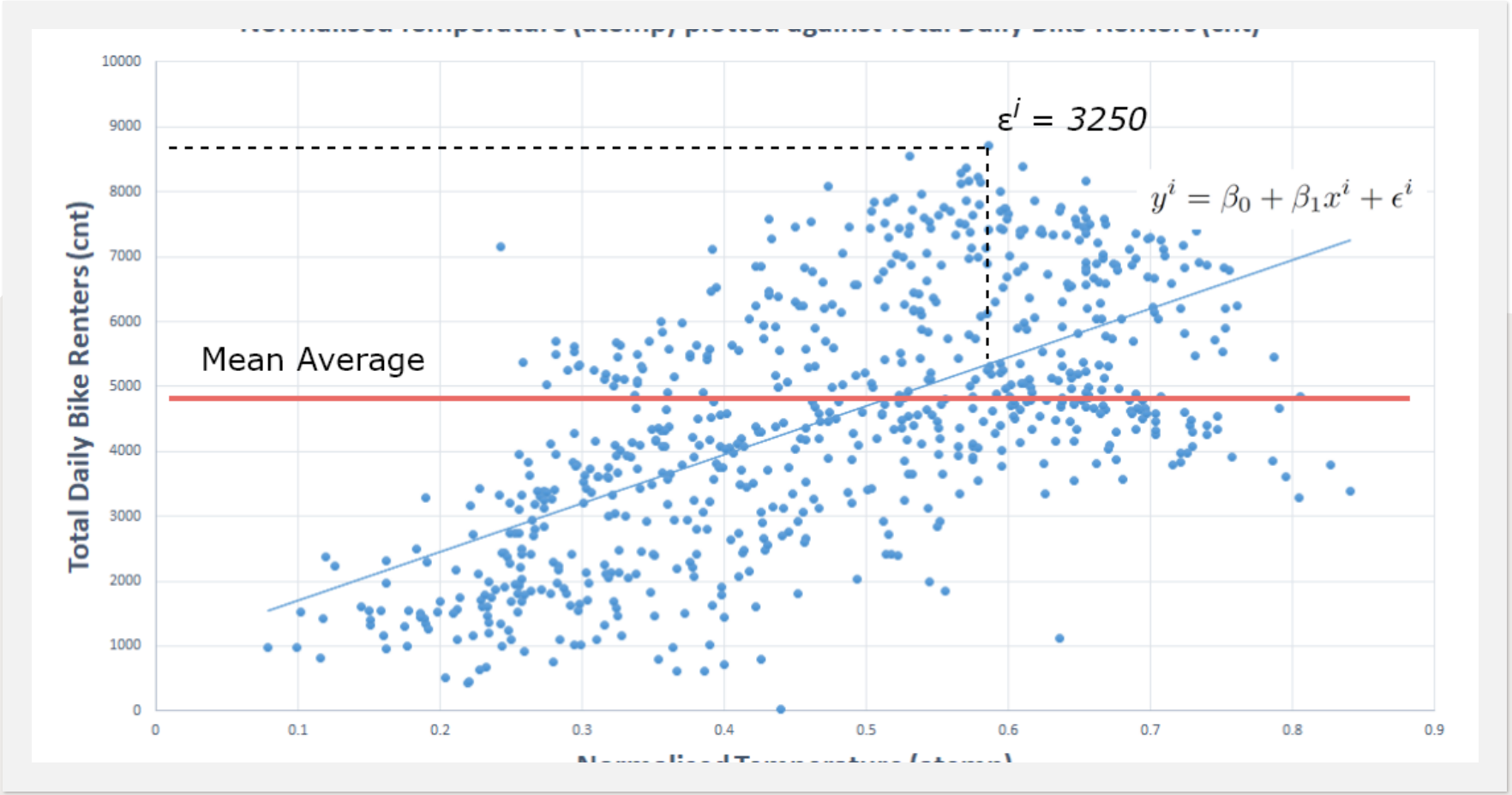




PROBLEMAS EN LOS DATOS


'drop'
];
events.forEach(e => {
fileDropZone.addEventListener(e, (ev) => {
ev.preventDefault();
if (ev.type === 'dragenter') {
fileDropZone.classList.add('solid-border');
}
if (ev.type === 'dragleave') {
fileDropZone.classList.remove('solid-border');
}
if(ev.type === 'drop') {
fileDropZone.classList.remove('solid-border');
ev.dataTransfer.files)
'border rounded

NOTEBOOK



REGRESIÓN LINEAL

$$y^i = \beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + \ldots + \beta_k x_k^i + \epsilon^i$$

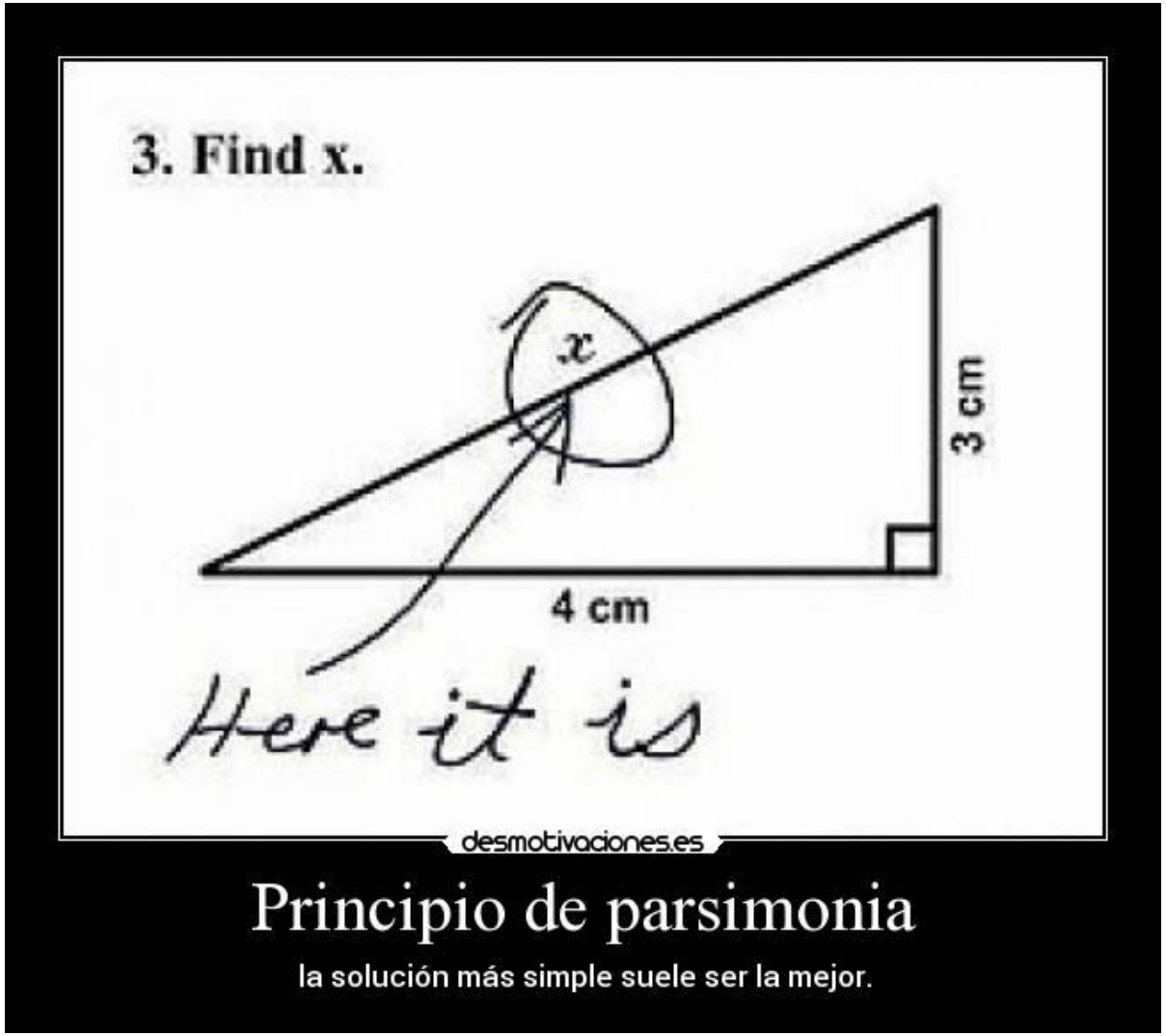
$$SSE = (\epsilon_1)^2 + (\epsilon_2)^2 + \ldots + (\epsilon_n)^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

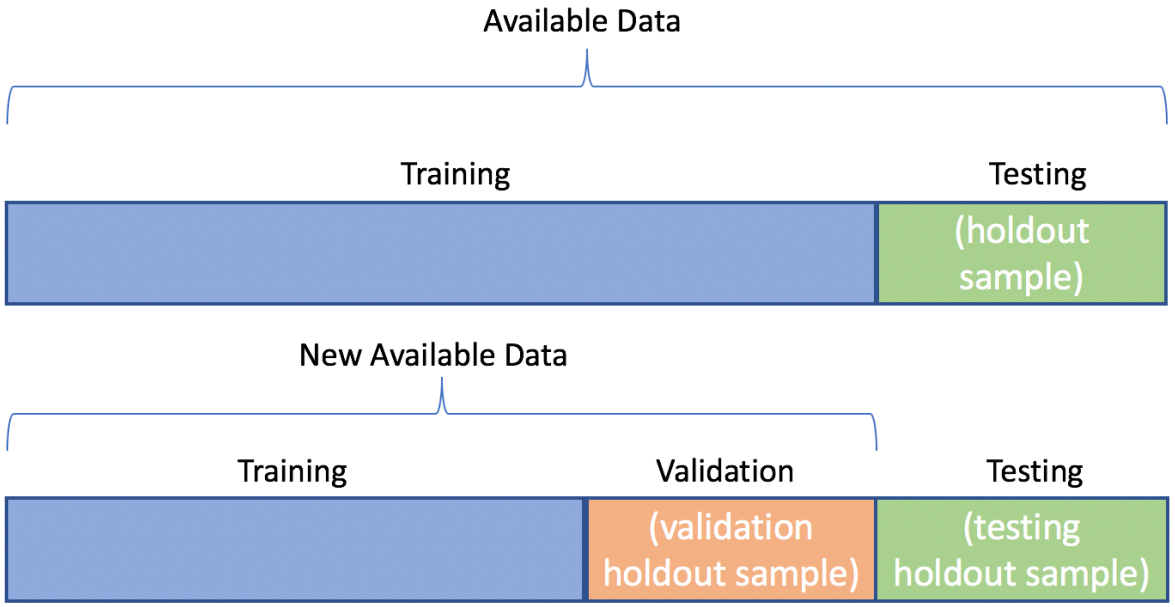
FÓRMULAS

CORRELACIÓN

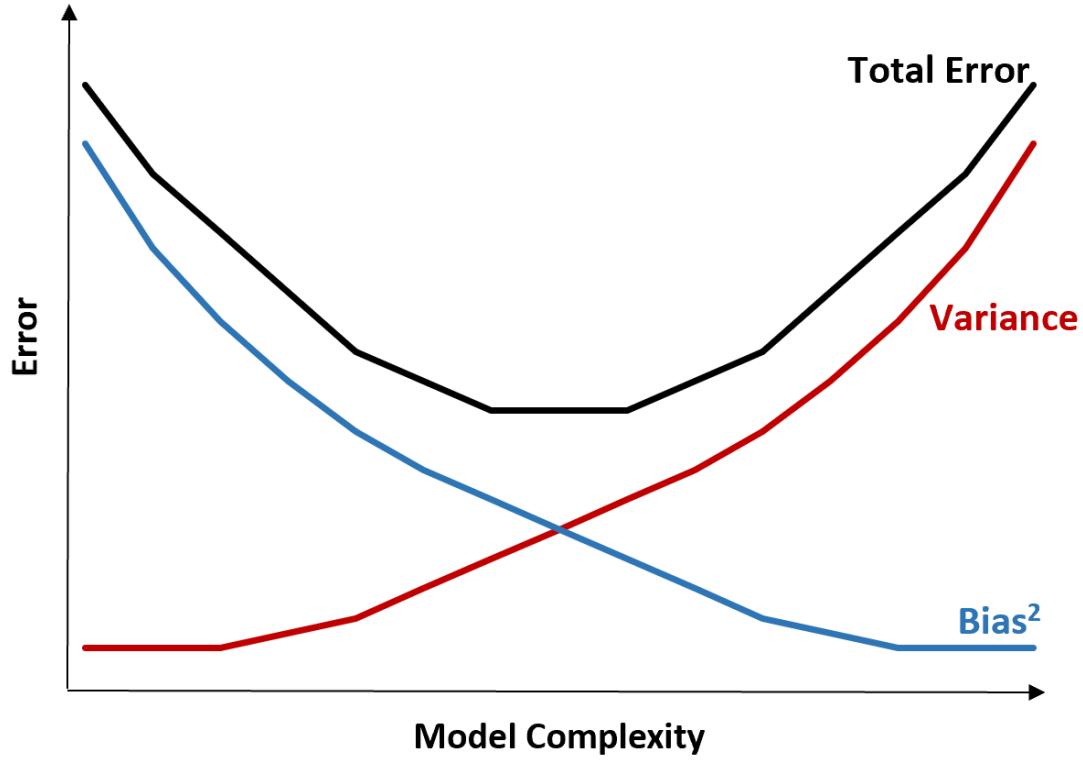
PRINCIPIO DE PARSIMONIA



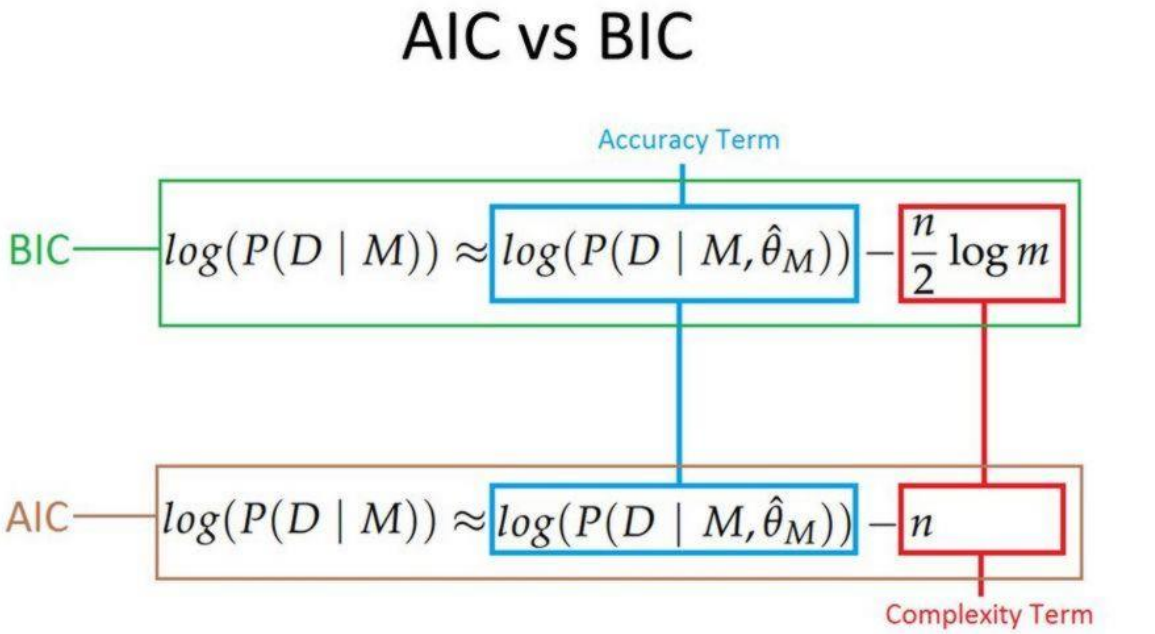
TRAIN Y TEST


Available Data
Training
Testing
(holdout sample)
New Available Data
Training
Validation
Testing
(validation holdout sample)
(testing holdout sample)

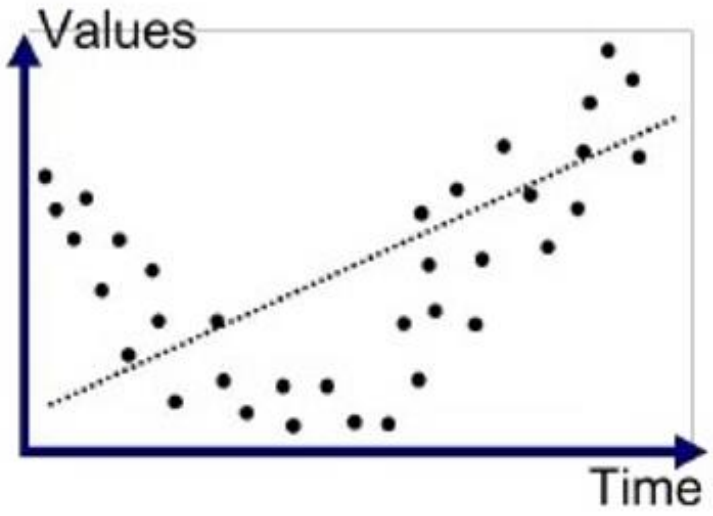
TRADE OFF SESGO Y VARIANZA



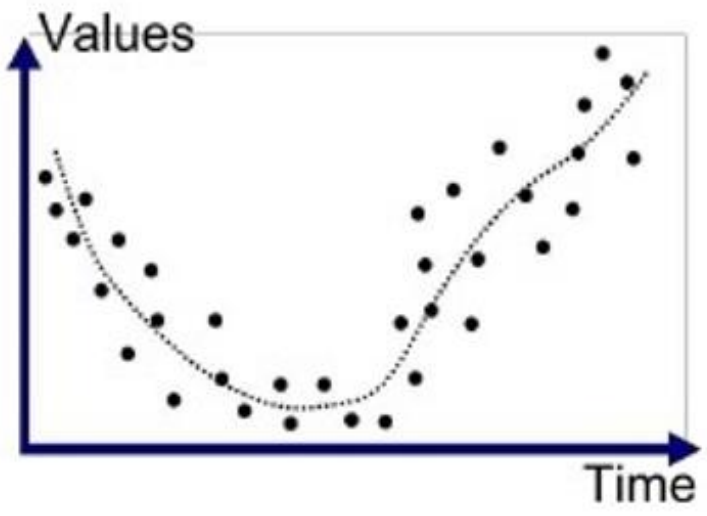
AIC VS BIC



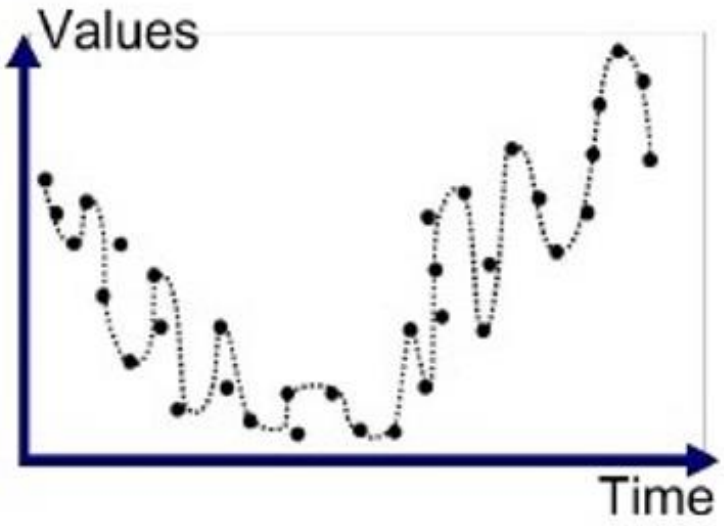
BetweenMates.com



Underfitted

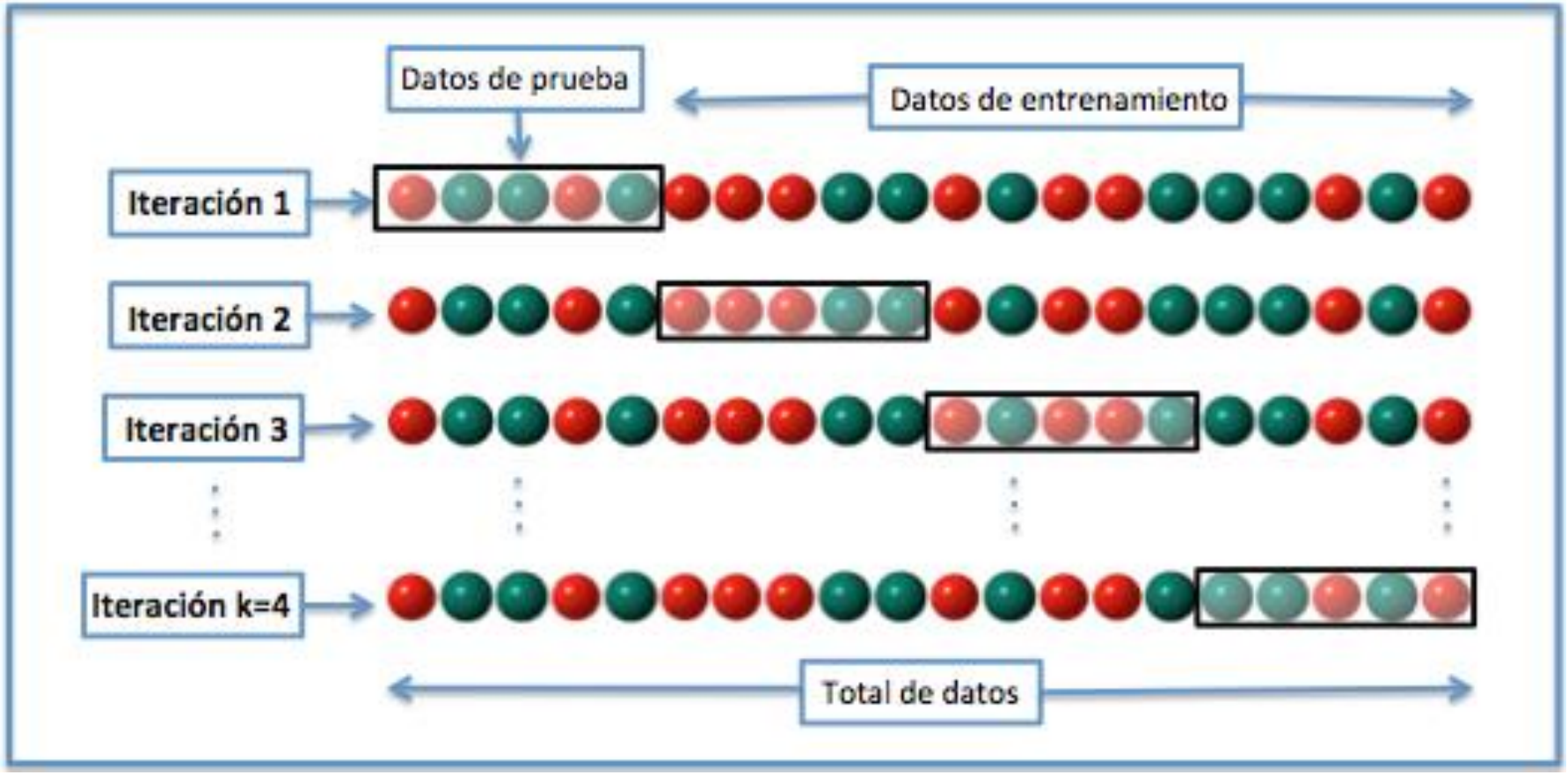


Good Fit/Robust



Overfitted

LASSO, RIDGE Y ELASTIC NET



VALIDACION CRUZADA

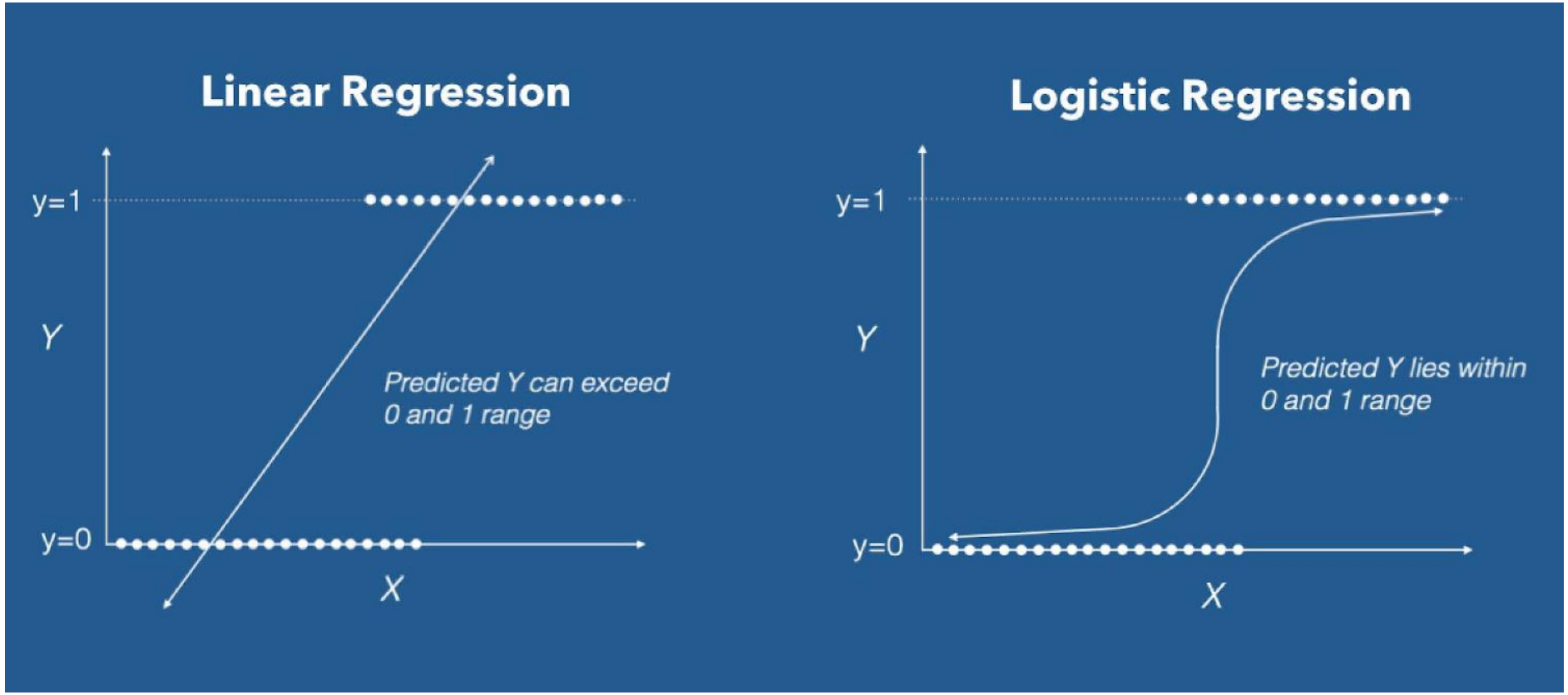

'drop'
];
events.forEach(e => {
fileDropZone.addEventListener(e, (ev) => {
ev.preventDefault();
if (ev.type === 'dragenter') {
fileDropZone.classList.add('solid-border');
}
if (ev.type === 'dragleave') {
fileDropZone.classList.remove('solid-border');
}
if(ev.type === 'drop') {
fileDropZone.classList.remove('solid-border');
ev.dataTransfer.files)
'border rounded

NOTEBOOK

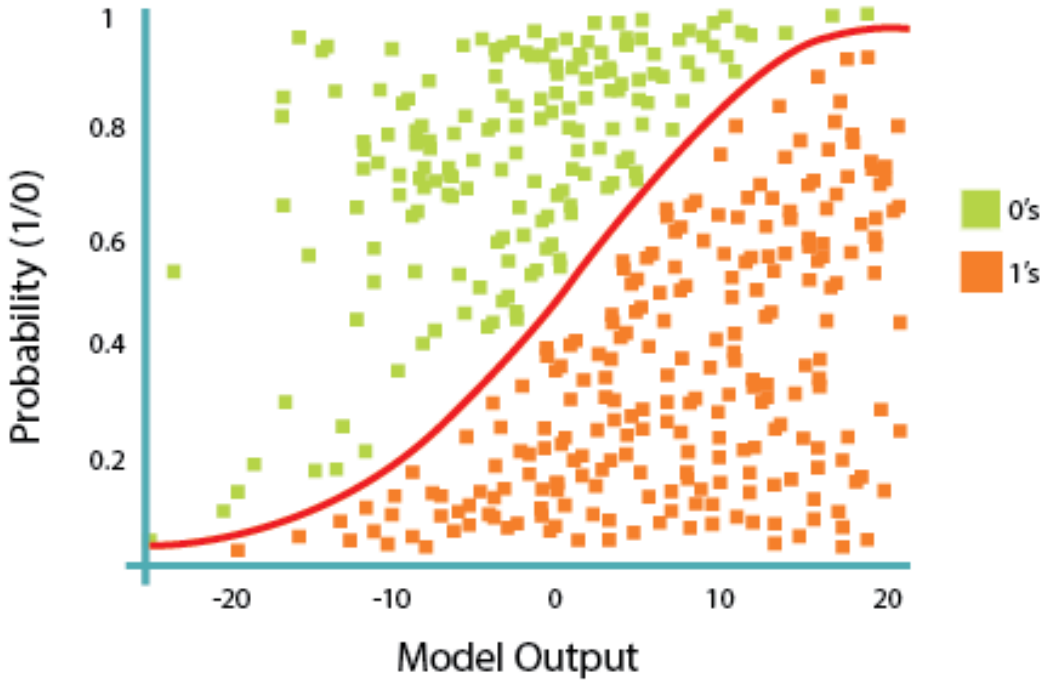

'drop'
];
events.forEach(e => {
fileDropZone.addEventListener(e, (ev) => {
ev.preventDefault();
if (ev.type === 'dragenter') {
fileDropZone.classList.add('solid-border');
}
if (ev.type === 'dragleave') {
fileDropZone.classList.remove('solid-border');
}
if(ev.type === 'drop') {
fileDropZone.classList.remove('solid-border');
ev.dataTransfer.files)
'border rounded

NOTEBOOK

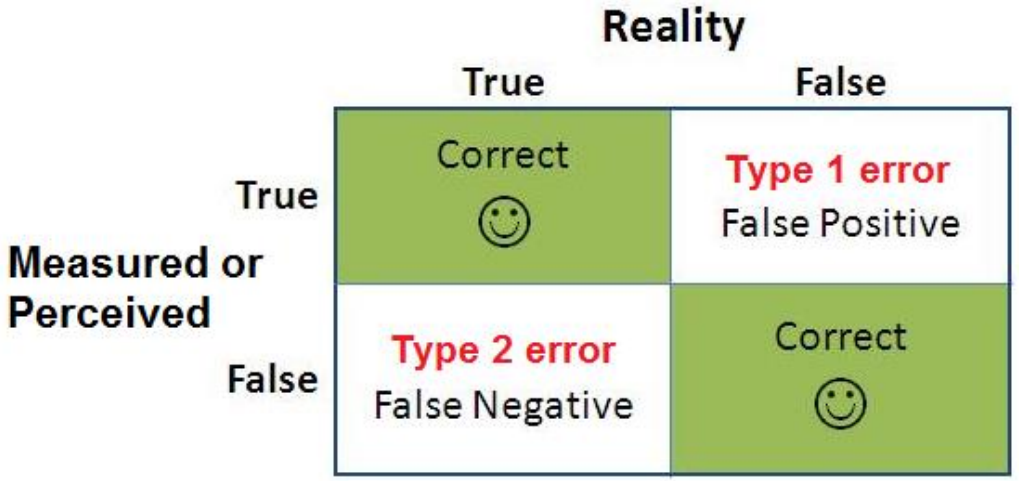
REGRESIÓN LOGÍSTICA



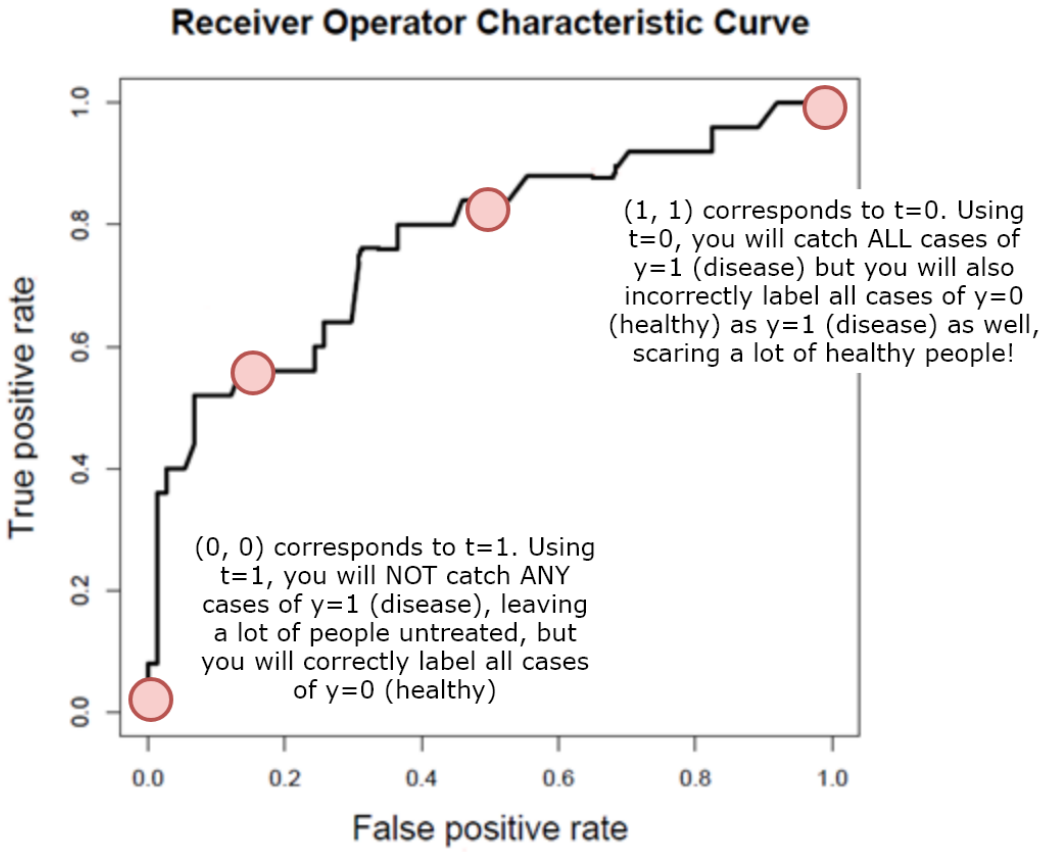
FUNCIÓN SIGMOIDE



THRESHOLD

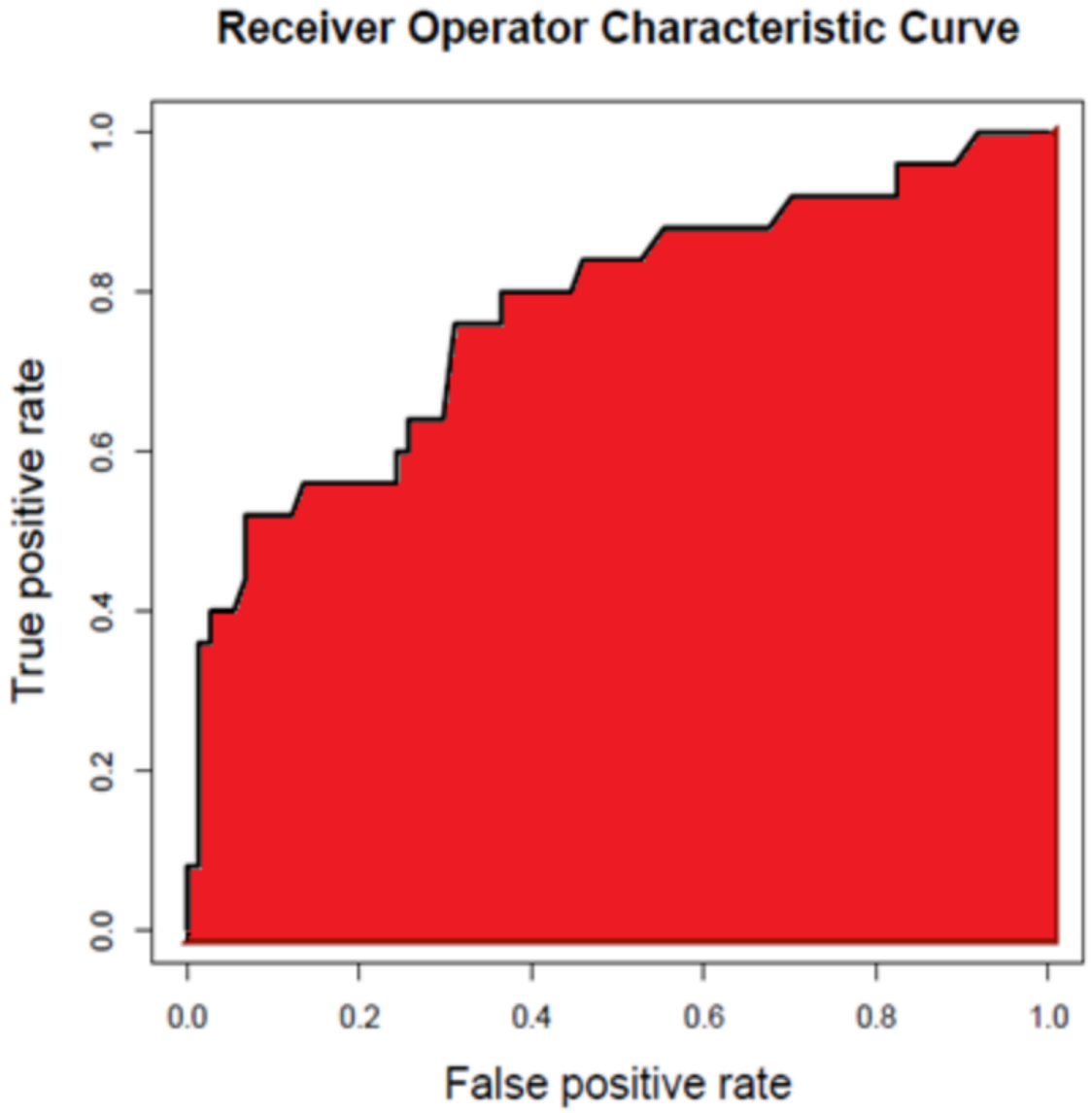

Reality
True
False
Measured or Perceived
True
False
Correct
Type 1 error
False Positive
Type 2 error
False Negative
Correct

Matriz de confusión		Estimado por el modelo			
		Negativo (N)	*Positivo (P)*		
Real	*Negativo*	a: (TN)	b: (FP)		
	Positivo	c: (FN)	d: (TP)	**Precisión *("precision")*** Porcentaje predicciones positivas correctas:	d/(b+d)
		Sensibilidad, exhaustividad *("Recall")* Porcentaje casos positivos detectados	**Especifidad (*Specifity*)** Porcentaje casos negativos detectados	**Exactitud ("accuracy")** Porcentaje de predicciones correctas *(No sirve en datasets poco equilibrados)*	
		d/(d+c)	a/(a+b)	(a+d)/(a+b+c+d)	


Receiver Operator Characteristic Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
(1, 1) corresponds to t=0. Using t=0, you will catch ALL cases of y=1 (disease) but you will also incorrectly label all cases of y=0 (healthy) as y=1 (disease) as well, scaring a lot of healthy people!
(0, 0) corresponds to t=1. Using t=1, you will NOT catch ANY cases of y=1 (disease), leaving a lot of people untreated, but you will correctly label all cases of y=0 (healthy)

CURVA ROC

AUC


Receiver Operator Characteristic Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

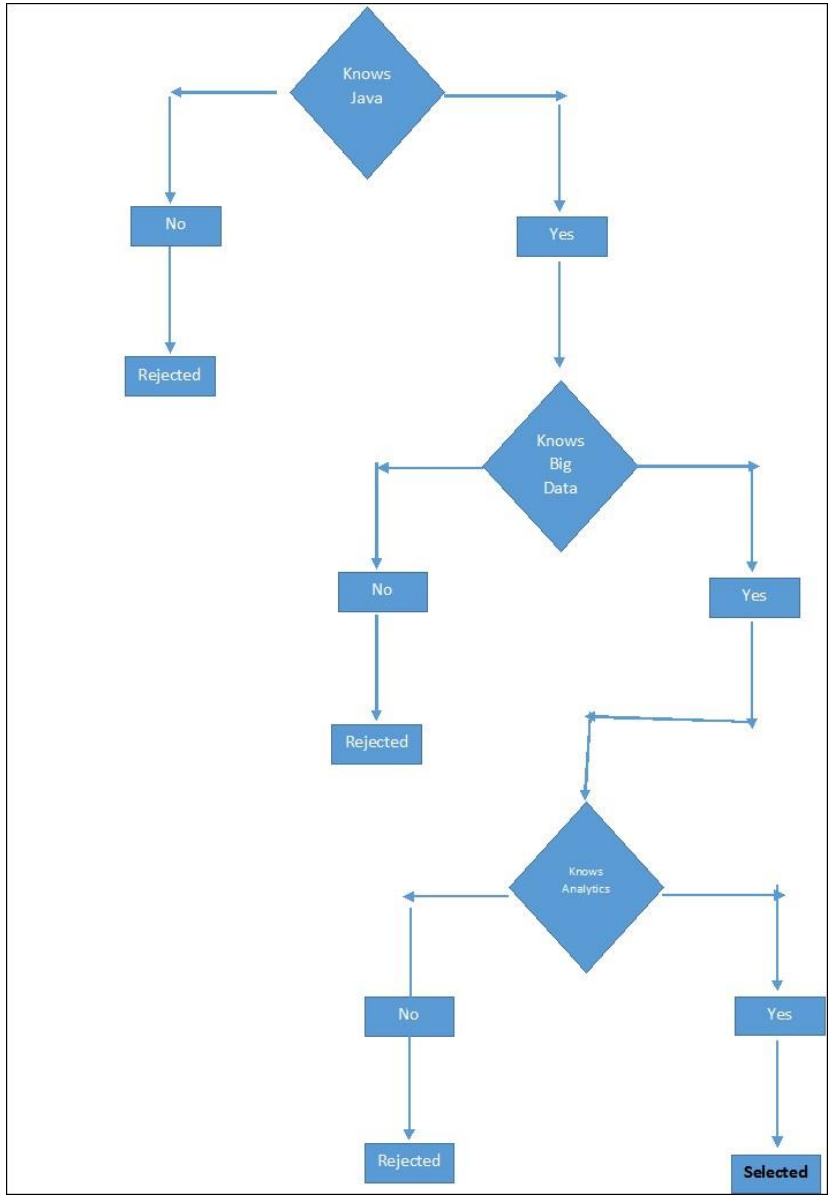

'drop'
];
events.forEach(e => {
fileDropZone.addEventListener(e, (ev) => {
ev.preventDefault();
if (ev.type === 'dragenter') {
fileDropZone.classList.add('solid-border');
}
if (ev.type === 'dragleave') {
fileDropZone.classList.remove('solid-border');
}
if(ev.type === 'drop') {
fileDropZone.classList.remove('solid-border');
ev.dataTransfer.files)
'border rounded

NOTEBOOK


'drop'
];
events.forEach(e => {
fileDropZone.addEventListener(e, (ev) => {
ev.preventDefault();
if (ev.type === 'dragenter') {
fileDropZone.classList.add('solid-border');
}
if (ev.type === 'dragleave') {
fileDropZone.classList.remove('solid-border');
}
if(ev.type === 'drop') {
fileDropZone.classList.remove('solid-border');
ev.dataTransfer.files)
'border rounded

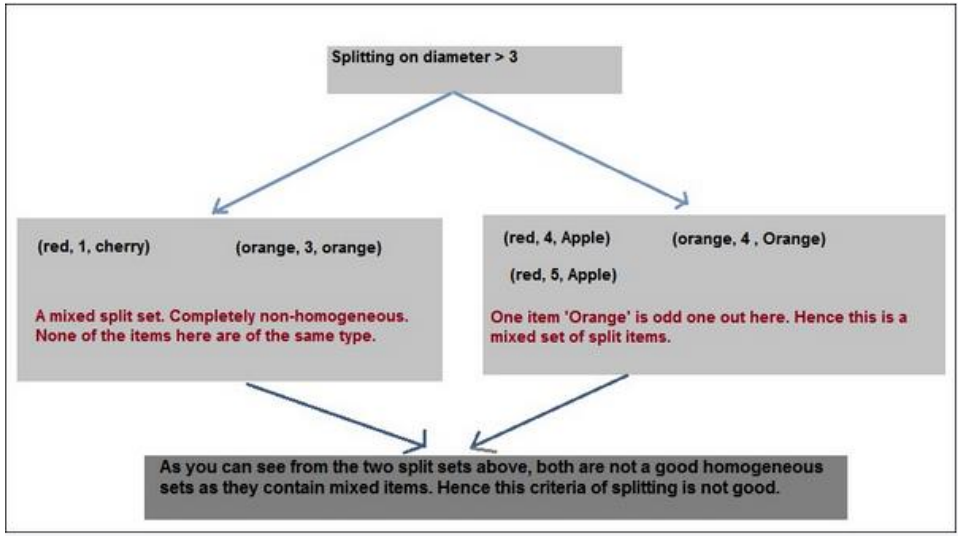
NOTEBOOK

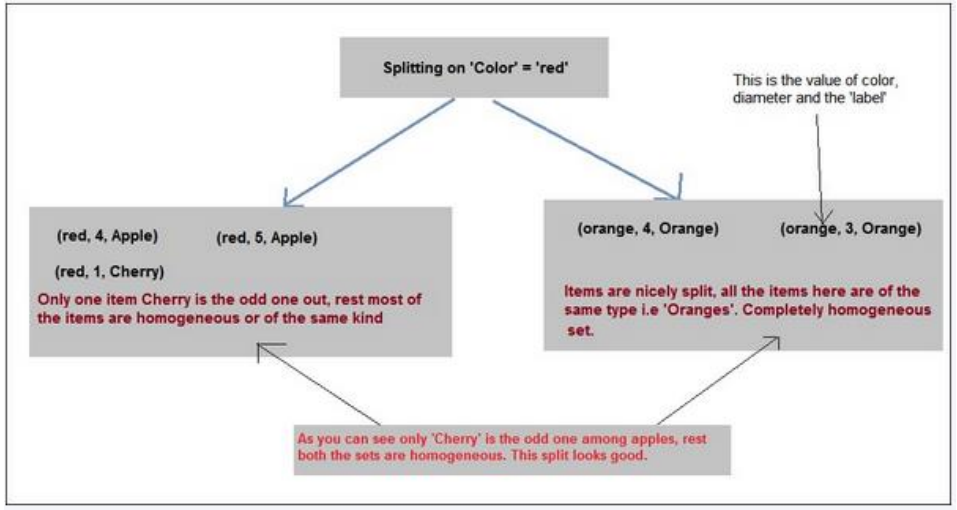
ARBOLES DE DECISIÓN



EJEMPLO DE ARBOL DE DECISIÓN

Color	Diámetro	Fruta
Rojo	4	Manzana
Naranja	4	Naranja
Rojo	1	Cereza
Rojo	5	Manzana
Naranja	3	Naranja





COMPARANDO DIVISIONES

$$Entropy(fruits) = -\left[\left(\frac{3}{8}\right)\log_2\frac{3}{8} + \left(\frac{5}{8}\right)\log_2\frac{5}{8}\right] = 0.95$$

$$Gini\ Impurity = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = 0.48$$

INDICE DE GINI Y ENTROPIA

VENTAJAS Y DESVENTAJAS

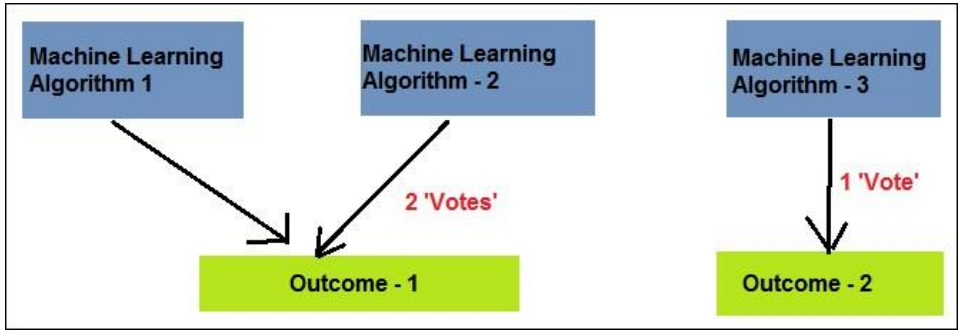
- ✓ Son muy simples de construir y usar. Son esencialmente un con-junto de if...else que conducen a un resultado concluyente.
- ✓ La entrada puede ser de cualquier tipo, ya sea numérica o string.
- ✓ Desde la perspectiva de los grandes datos y la informática distribuida, es más fácil crear un modelo de árbol de decisiones que se pueda distribuir en un grupo de máquinas. Por lo tanto, puede ejecutarse en paralelo y puede ser muy rápido.
- ❖ Los árboles de decisión sufren el problema del sobreajuste. Debido al sobreajuste, los se desempeñan extremadamente bien con los datos de entrenamiento, pero se desempeñan mal con cualquier dato nuevo que no sea parte del conjunto de entrenamiento.
- ❖ Si hay demasiadas reglas de decisión, el modelo pronto puede volverse bastante complejo. Dado que estamos tratando con grandes datos, este problema es más común.

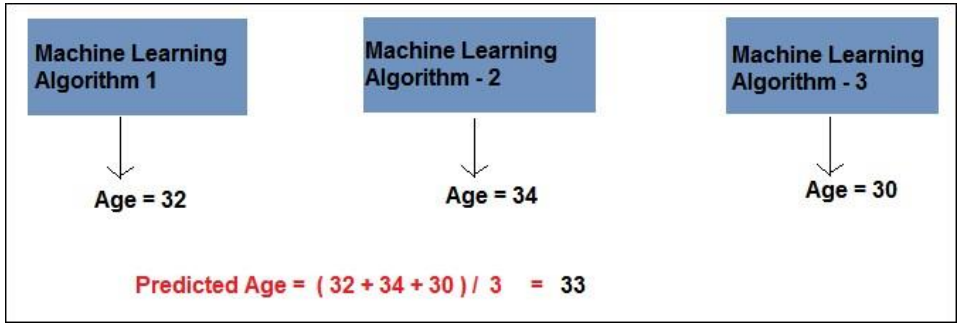
KINNECT


'drop'
];
events.forEach(e => {
fileDropZone.addEventListener(e, (ev) => {
ev.preventDefault();
if (ev.type === 'dragenter') {
fileDropZone.classList.add('solid-border');
}
if (ev.type === 'dragleave') {
fileDropZone.classList.remove('solid-border');
}
if(ev.type === 'drop') {
fileDropZone.classList.remove('solid-border');
ev.dataTransfer.files)
'border rounded

NOTEBOOK

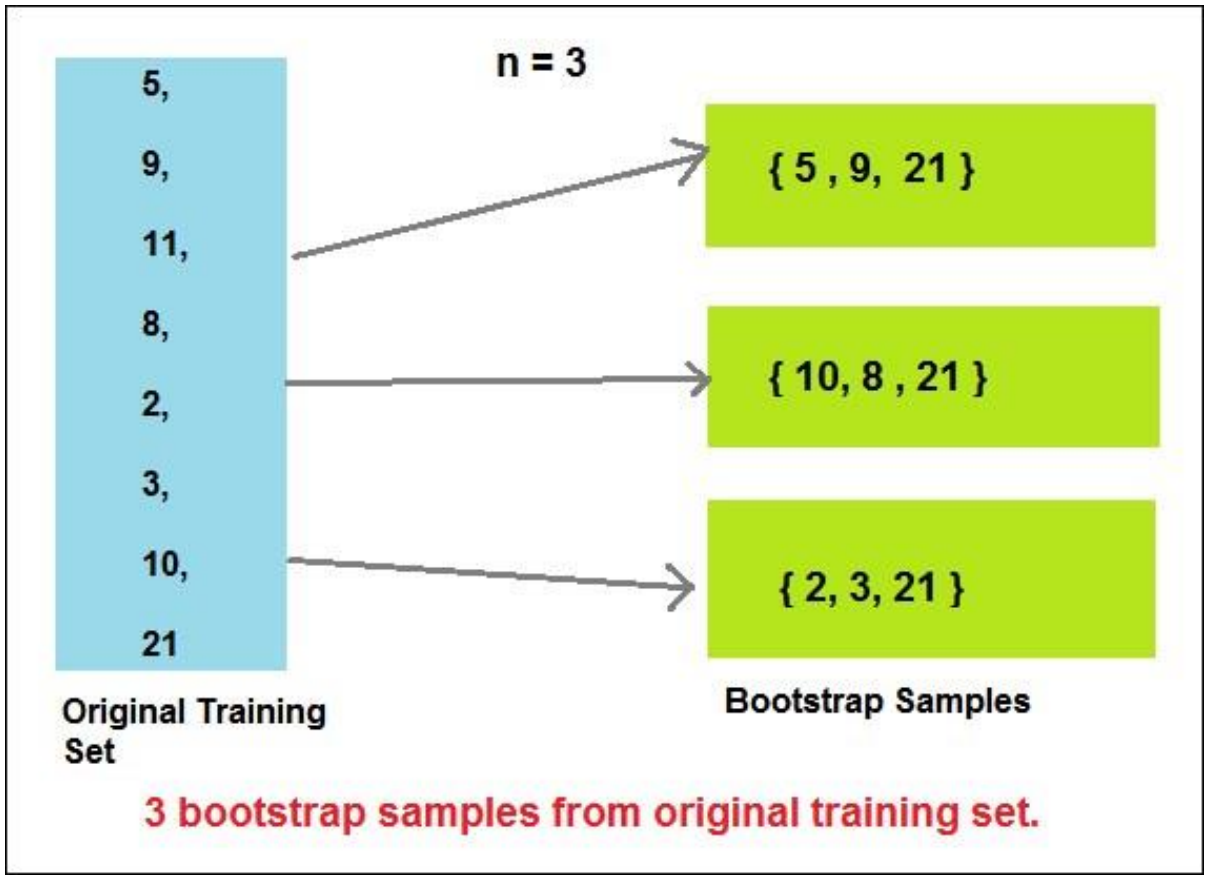
ENSEMBLING


Machine Learning Algorithm 1
Machine Learning Algorithm - 2
Machine Learning Algorithm - 3
2 'Votes'
1 'Vote'
Outcome - 1
Outcome - 2

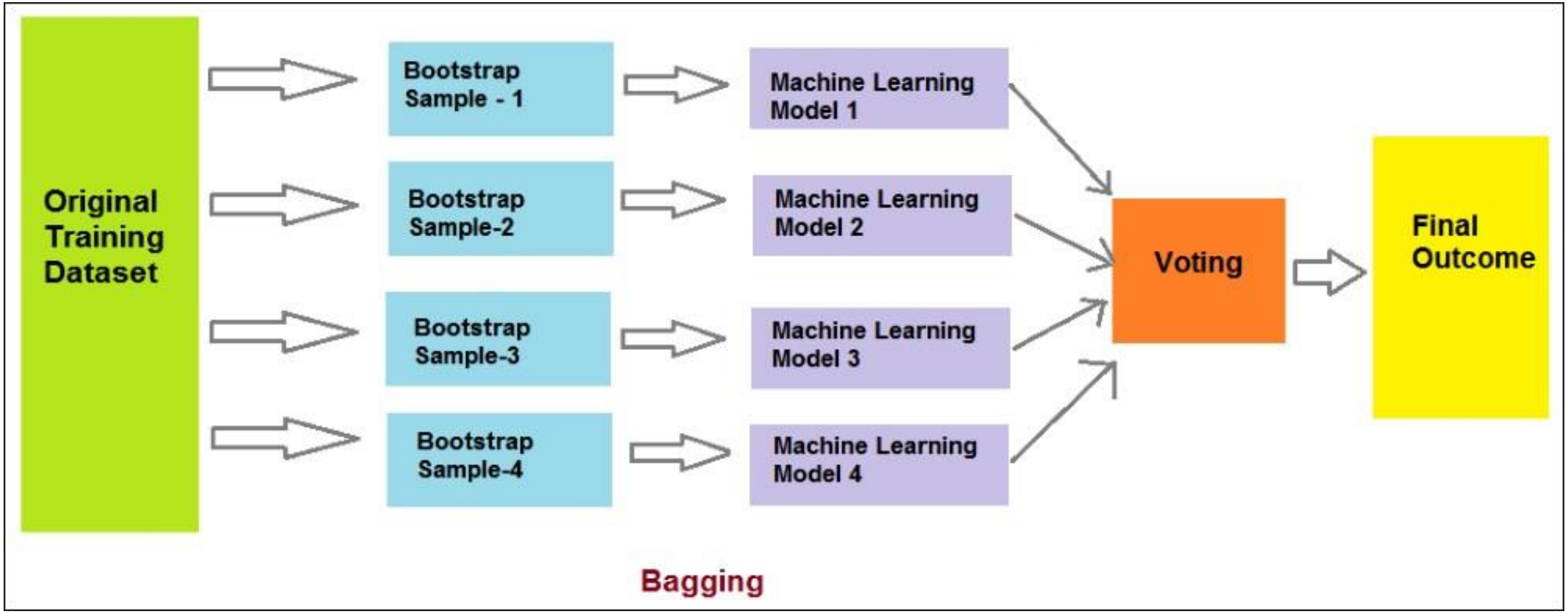

Machine Learning Algorithm 1
Machine Learning Algorithm - 2
Machine Learning Algorithm - 3
Age = 32
Age = 34
Age = 30
Predicted Age = (32 + 34 + 30) / 3 = 33

ENFOQUES

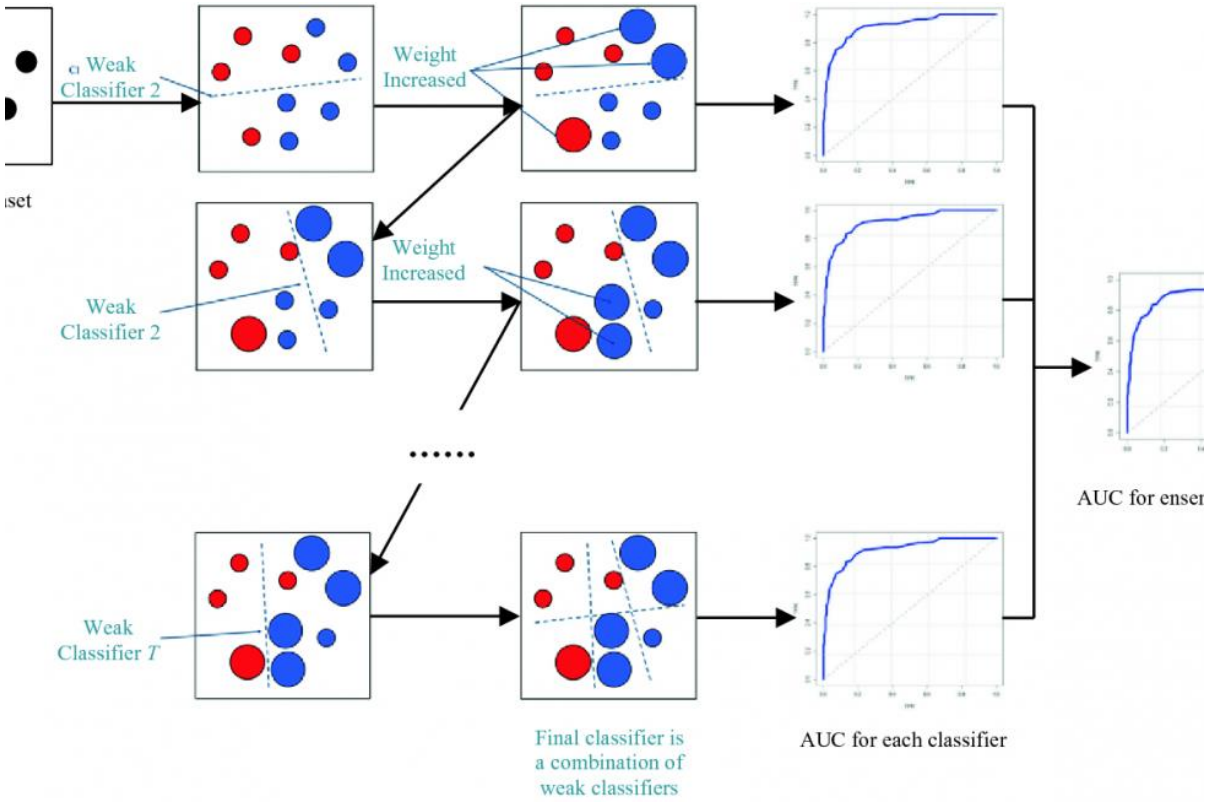
BOOTSTRAPING



BAGGING

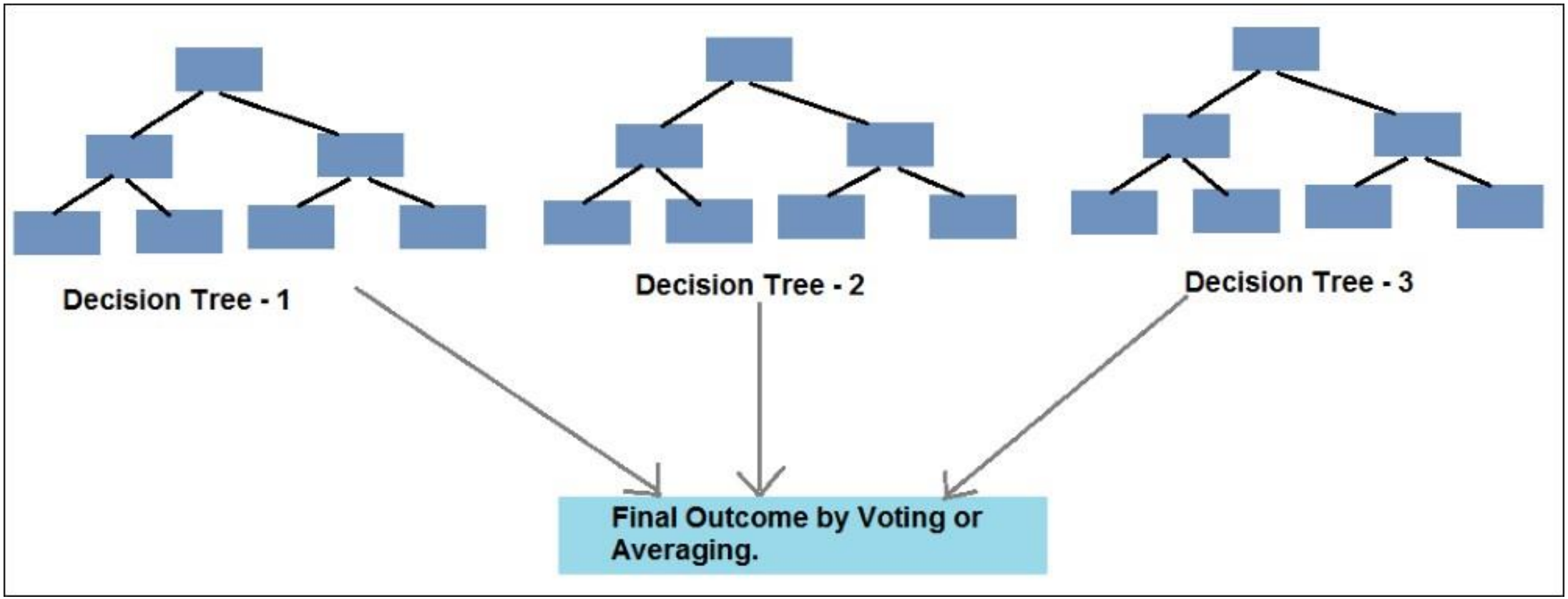

Original Training Dataset
Bootstrap Sample - 1
Bootstrap Sample-2
Bootstrap Sample-3
Bootstrap Sample-4
Machine Learning Model 1
Machine Learning Model 2
Machine Learning Model 3
Machine Learning Model 4
Voting
Final Outcome
Bagging

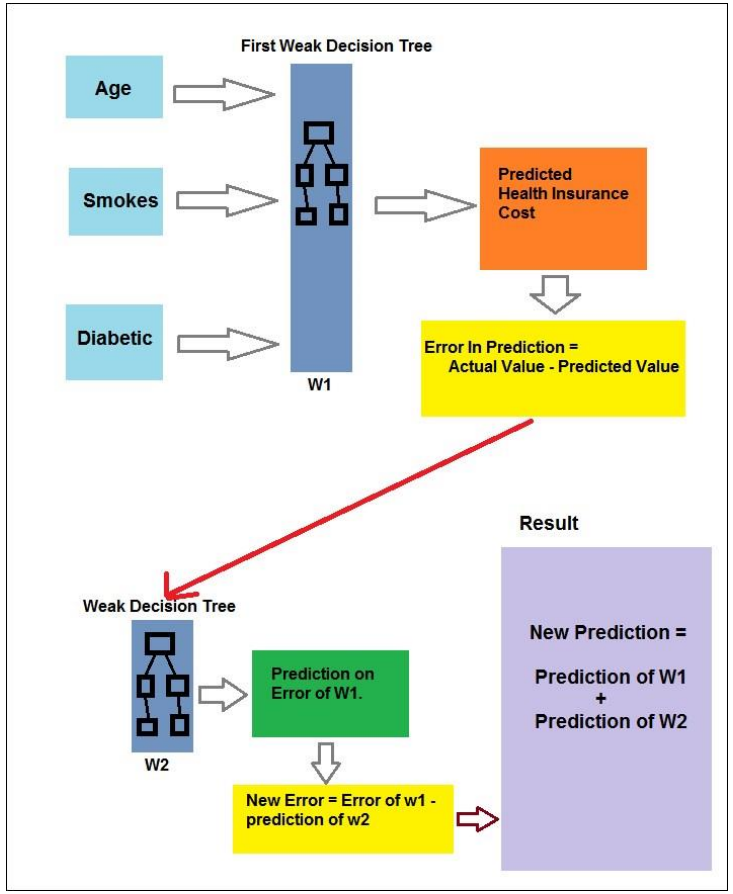
BOOSTING


Weak
Classifier 2
Weight
Increased
Weight
Increased
Weak
Classifier 2
.....
Weak
Classifier T
Final classifier is
a combination of
weak classifiers
AUC for each classifier
AUC for ense

VENTAJAS Y DESVENTAJAS

RANDOM FOREST

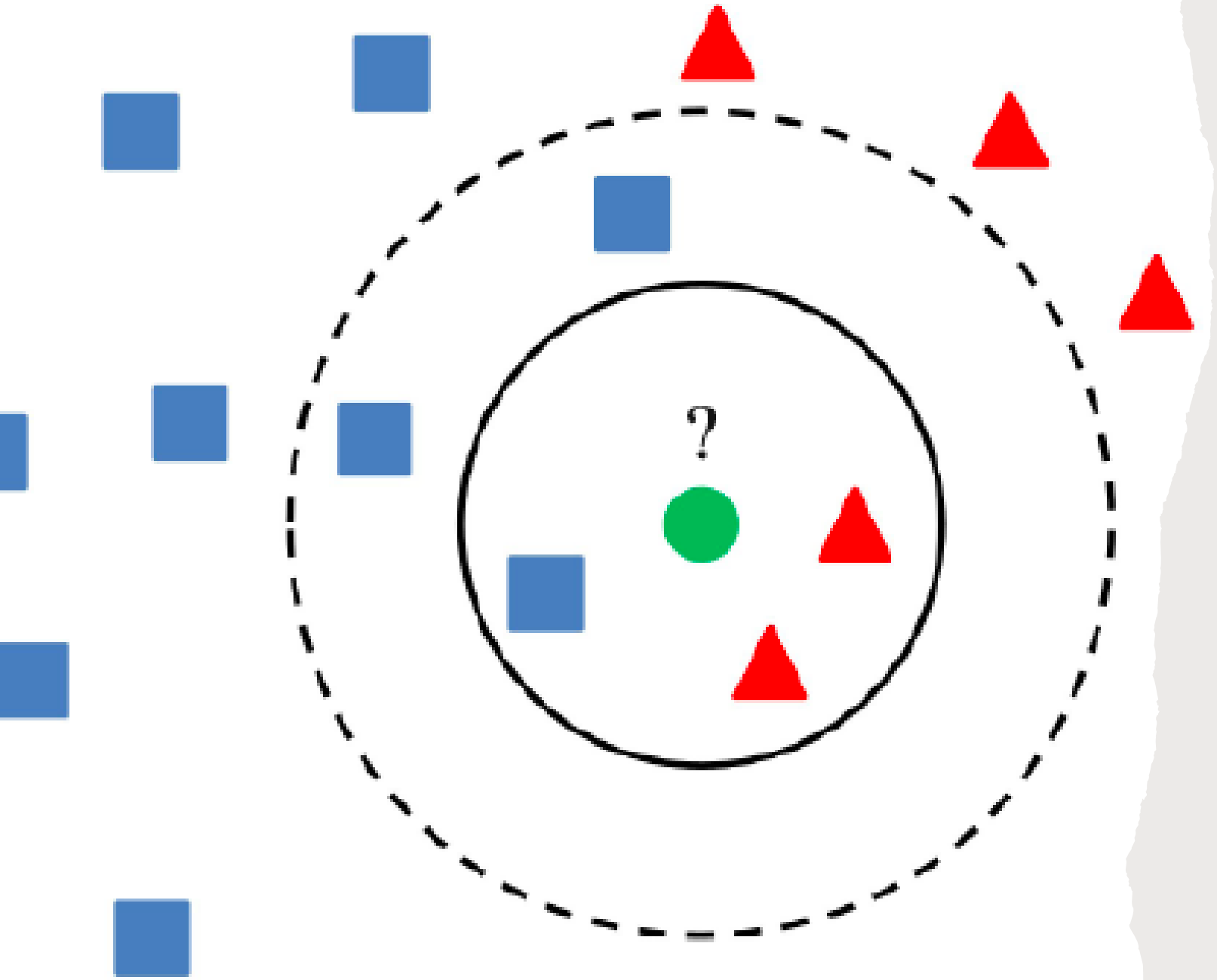



First Weak Decision Tree
Age
Smokes
Diabetic
W1
Predicted Health Insurance Cost
Error In Prediction = Actual Value - Predicted Value
Result
Weak Decision Tree
W2
Prediction on Error of W1.
New Error = Error of w1 - prediction of w2
New Prediction = Prediction of W1 + Prediction of W2

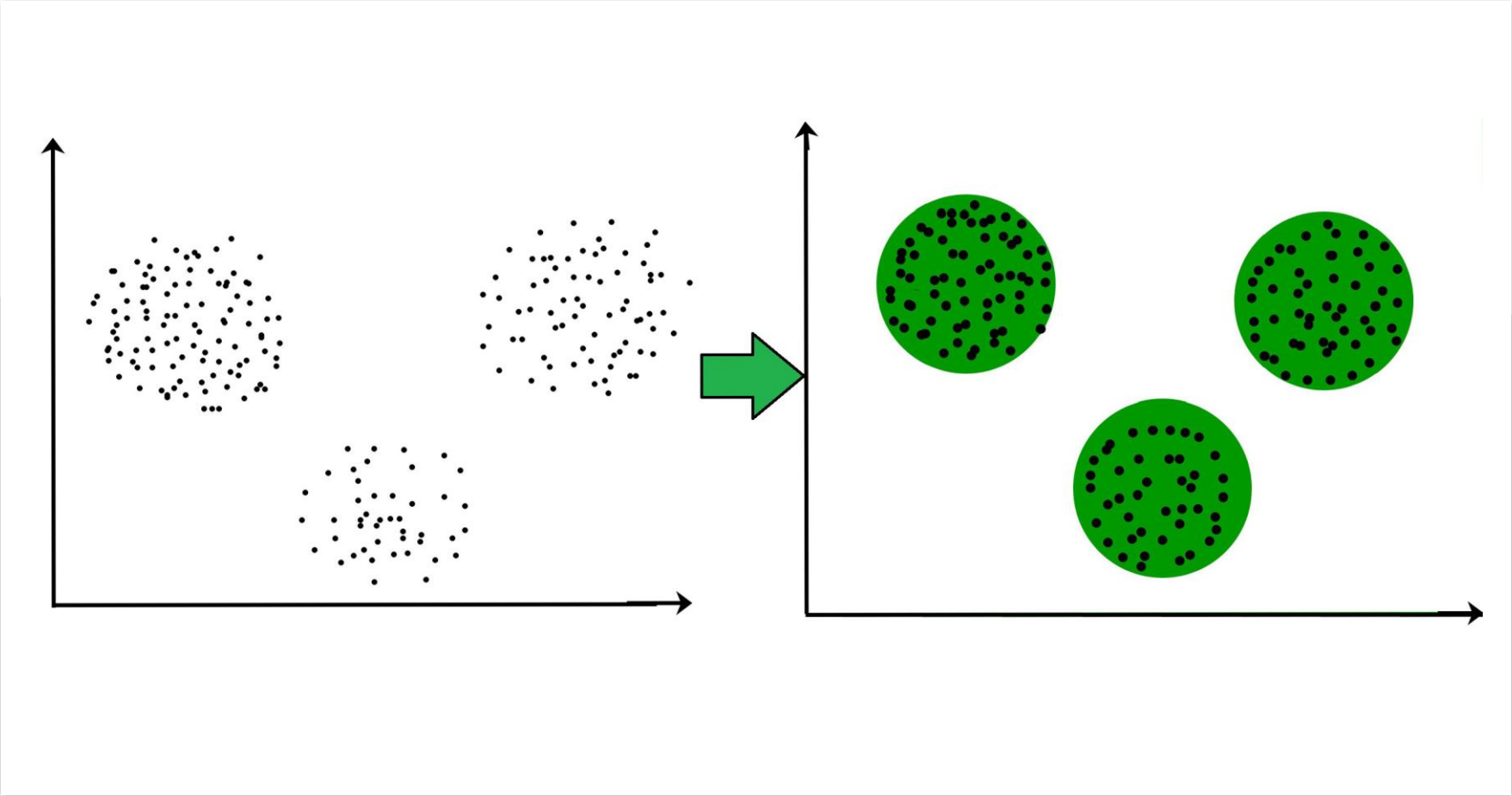
GBT


'drop'
];
events.forEach(e => {
fileDropZone.addEventListener(e, (ev) => {
ev.preventDefault();
if (ev.type === 'dragenter') {
fileDropZone.classList.add('solid-border');
}
if (ev.type === 'dragleave') {
fileDropZone.classList.remove('solid-border');
}
if(ev.type === 'drop') {
fileDropZone.classList.remove('solid-border');
ev.dataTransfer.files)
'border rounded

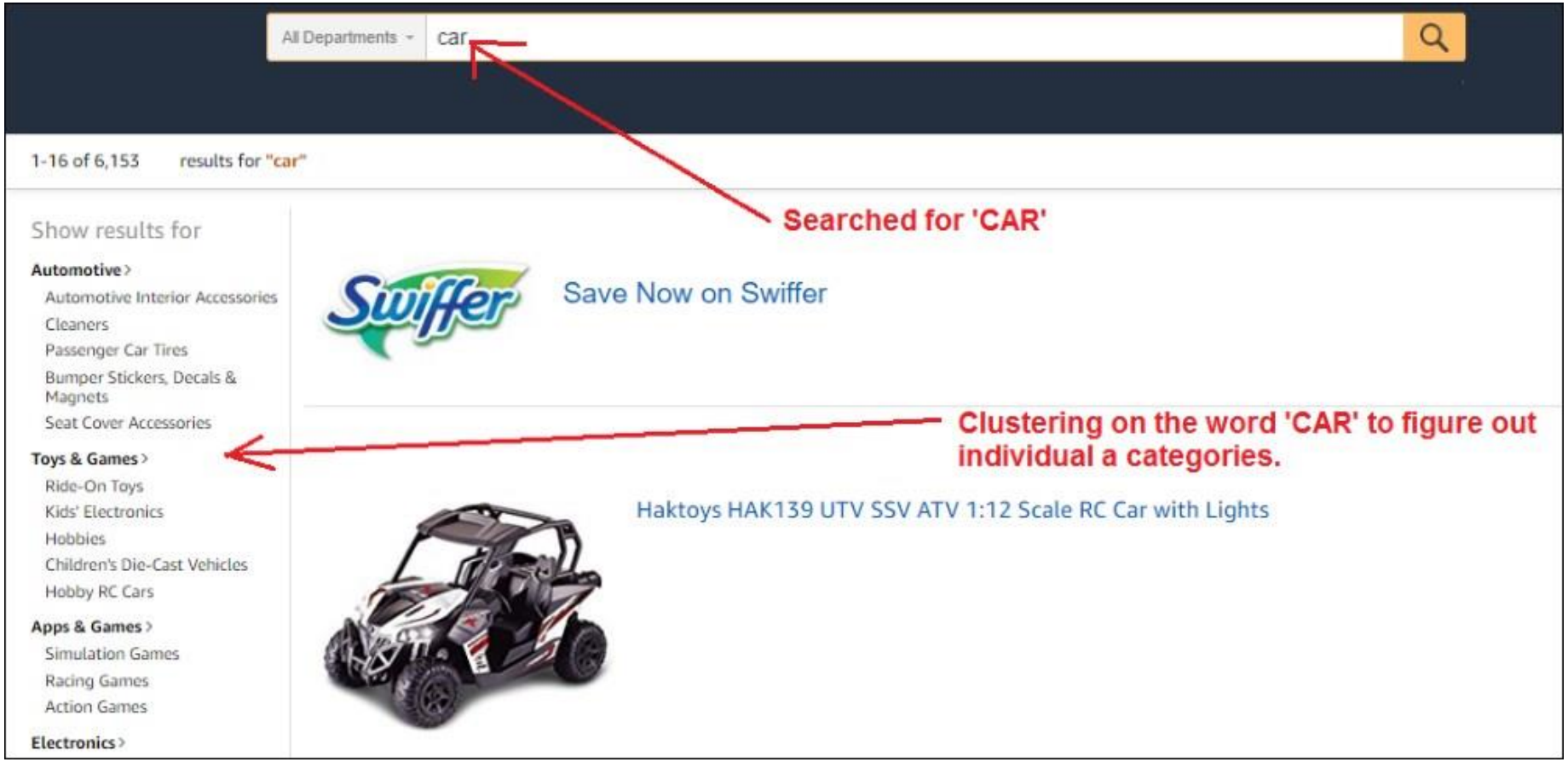
NOTEBOOK



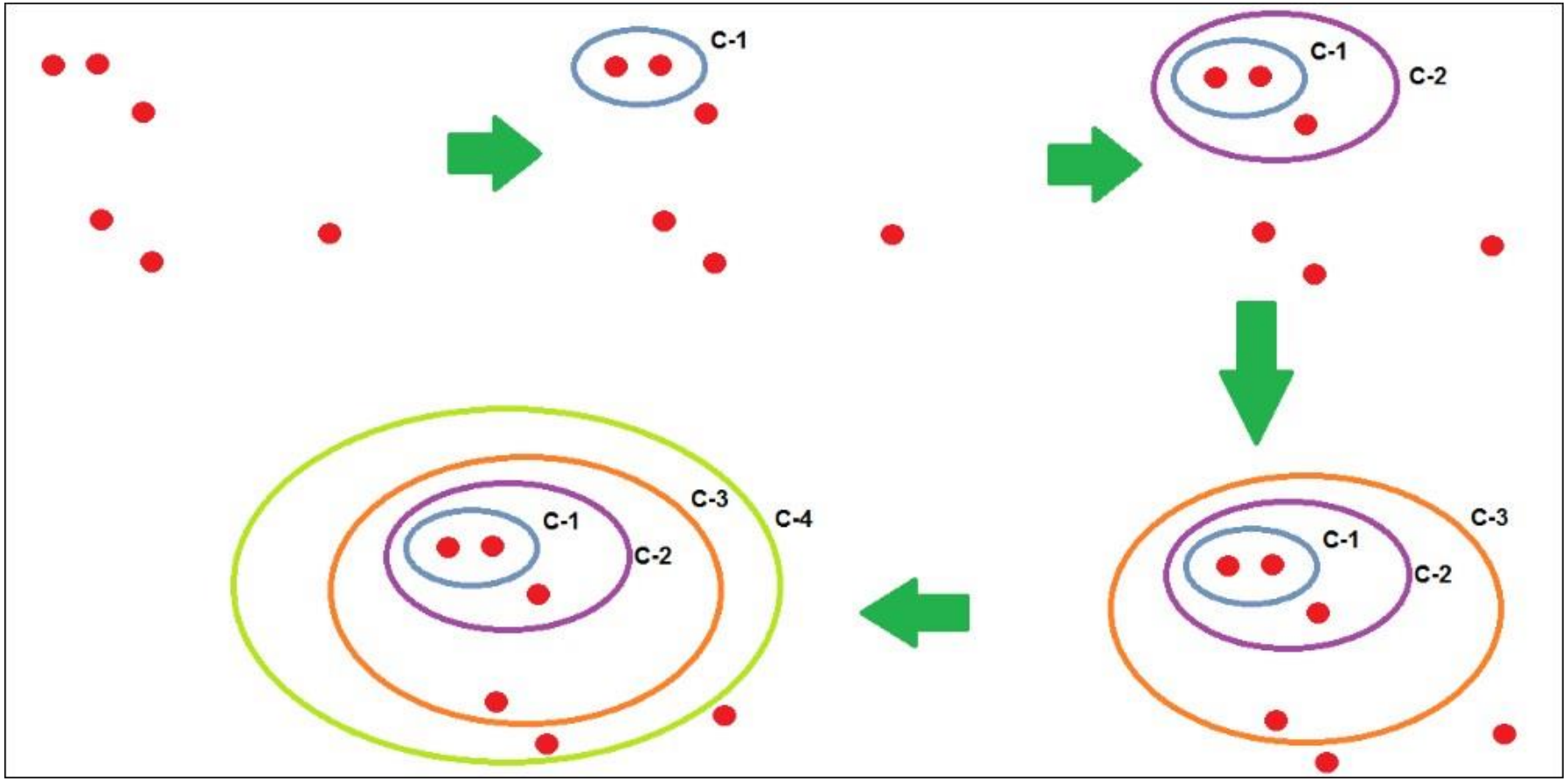
KNN



CLUSTERING

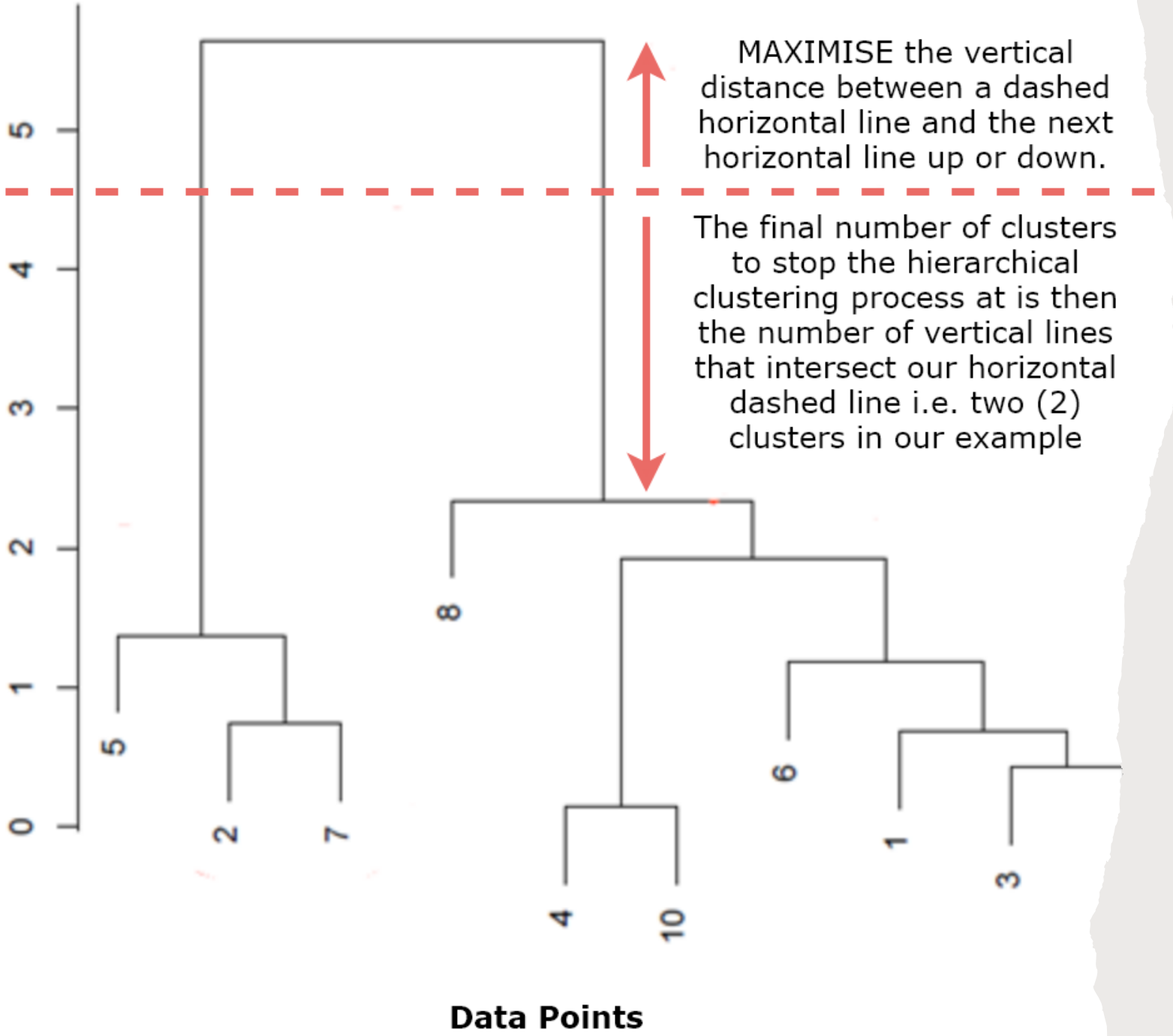

All Departments
car
1-16 of 6,153 results for "car"
Searched for 'CAR'
Show results for
Automotive
Automotive Interior Accessories
Cleaners
Passenger Car Tires
Bumper Stickers, Decals & Magnets
Seat Cover Accessories
Toys & Games
Ride-On Toys
Kids' Electronics
Hobbies
Children's Die-Cast Vehicles
Hobby RC Cars
Apps & Games
Simulation Games
Racing Games
Action Games
Electronics
Swiffer
Save Now on Swiffer
Clustering on the word 'CAR' to figure out individual a categories.
Haktoys HAK139 UTV SSV ATV 1:12 Scale RC Car with Lights

CASOS DE USO

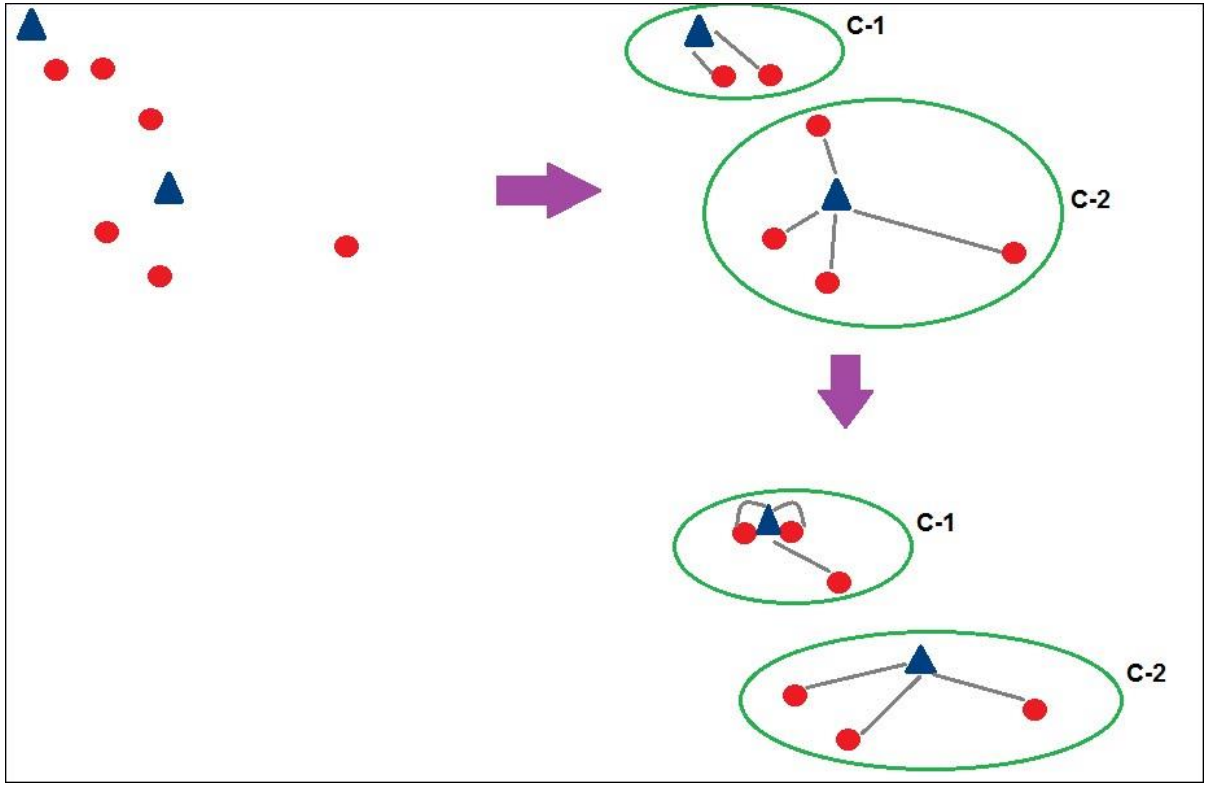


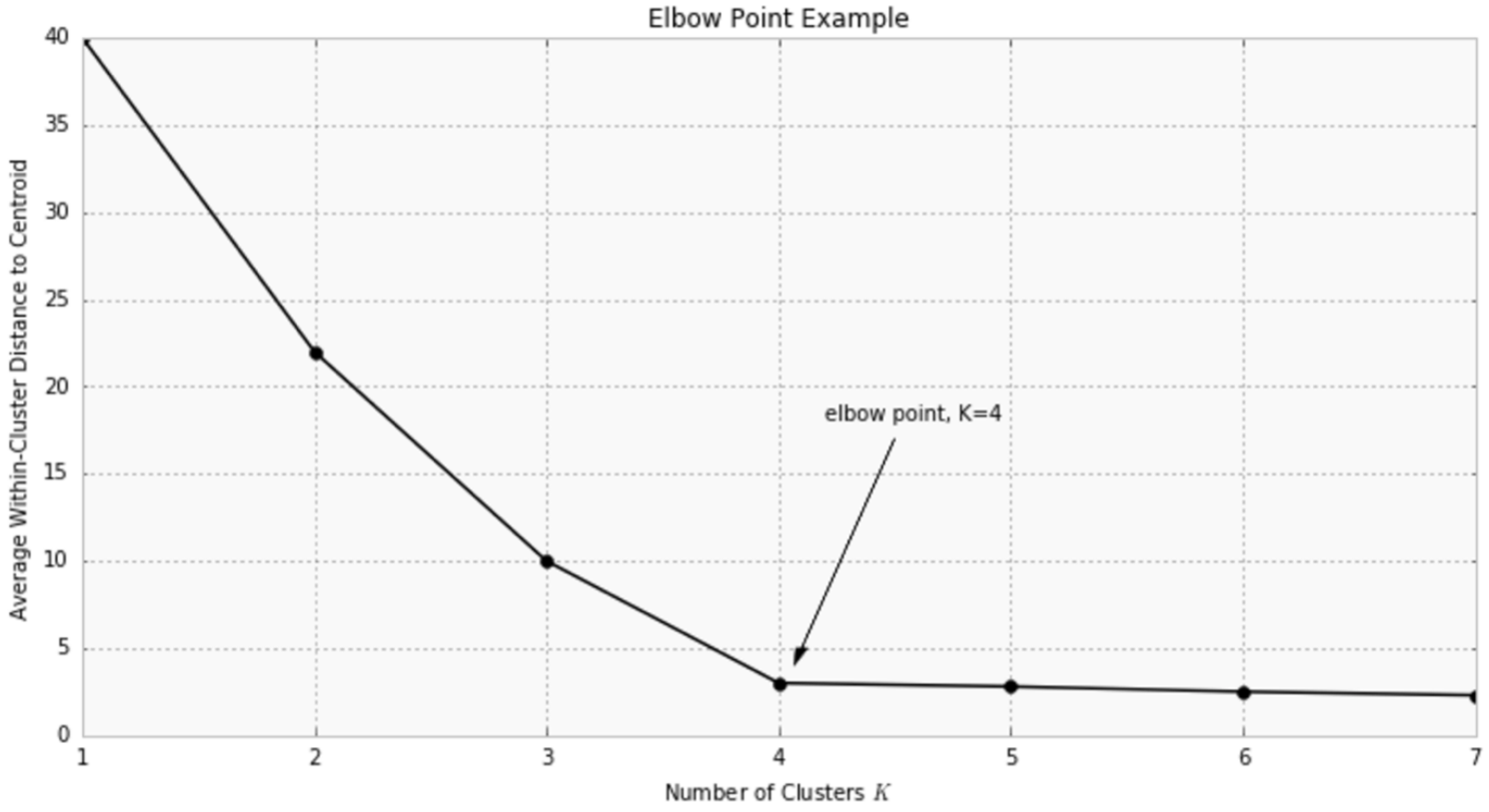
AGRUPAMIENTO JERARQUICO

DENDOGRAMA


MAXIMISE the vertical distance between a dashed horizontal line and the next horizontal line up or down.
The final number of clusters to stop the hierarchical clustering process at is then the number of vertical lines that intersect our horizontal dashed line i.e. two (2) clusters in our example
0
1
2
3
4
5
5
2
7
8
4
10
6
1
3
Data Points

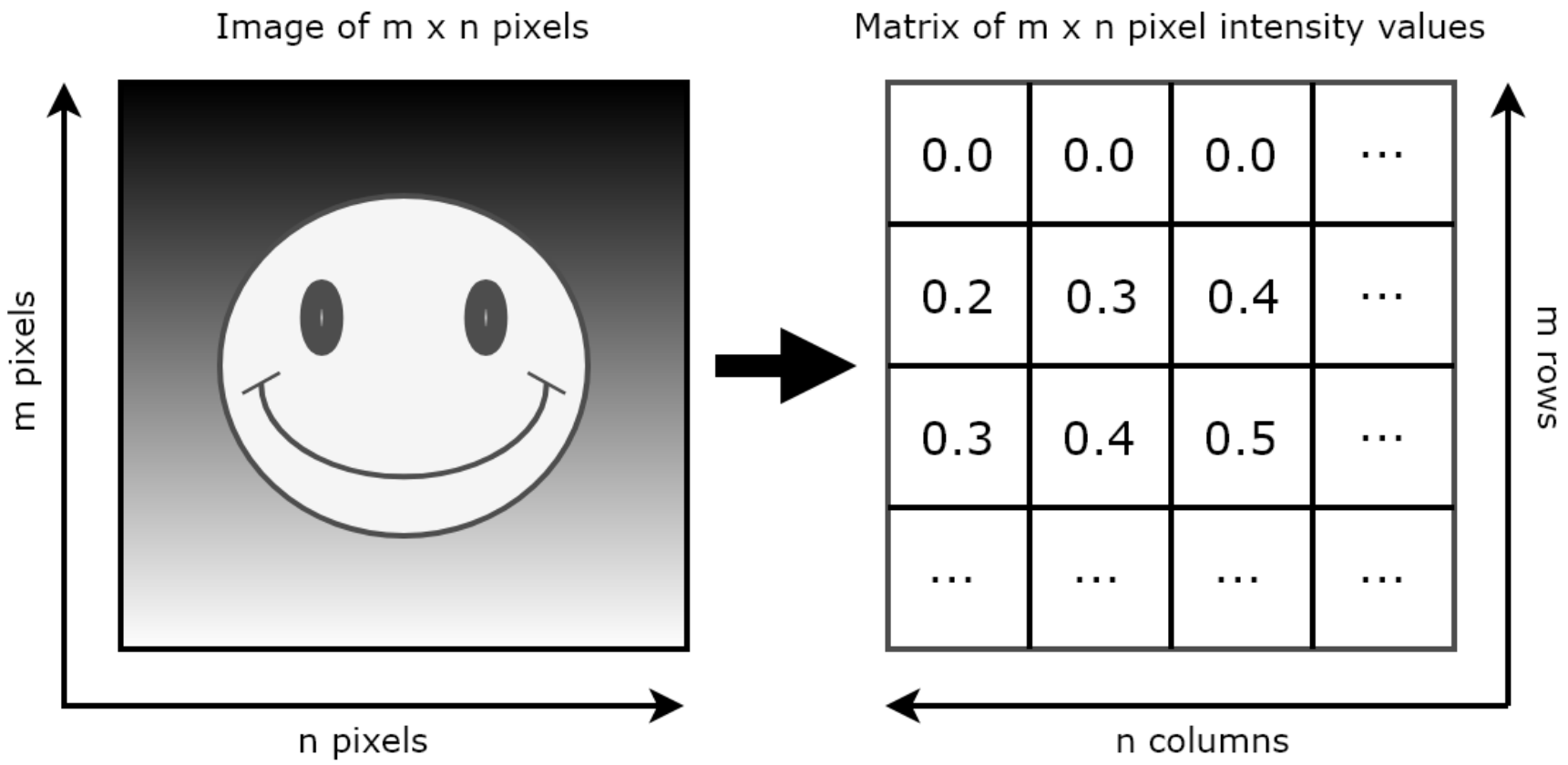
K-MEDIAS





ELBOW METHOD

IMAGEN = MATRIZ

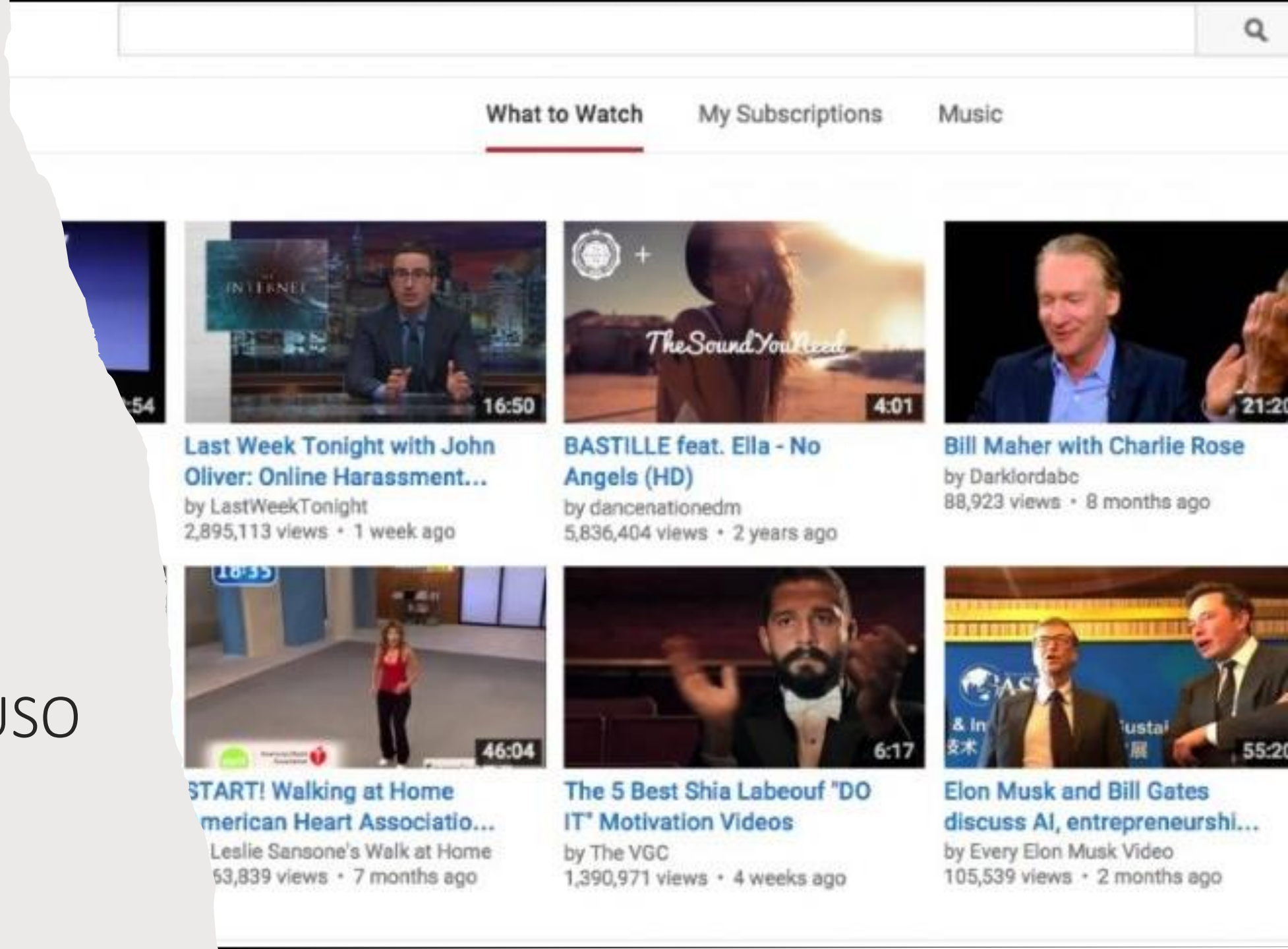



'drop'
];
events.forEach(e => {
fileDropZone.addEventListener(e, (ev) => {
ev.preventDefault();
if (ev.type === 'dragenter') {
fileDropZone.classList.add('solid-border');
}
if (ev.type === 'dragleave') {
fileDropZone.classList.remove('solid-border');
}
if(ev.type === 'drop') {
fileDropZone.classList.remove('solid-border');
ev.dataTransfer.files)
'border rounded

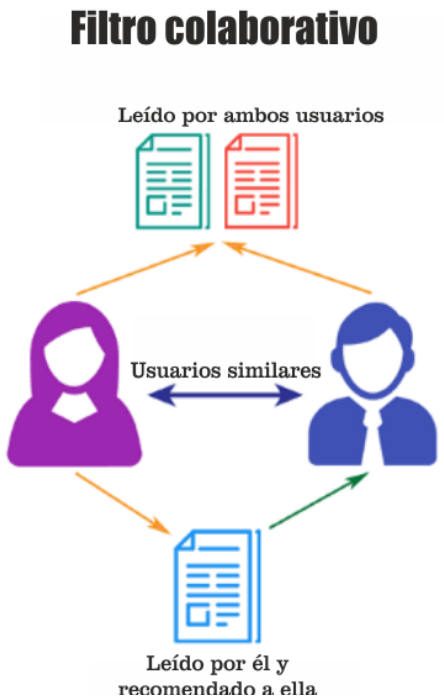
NOTEBOOK

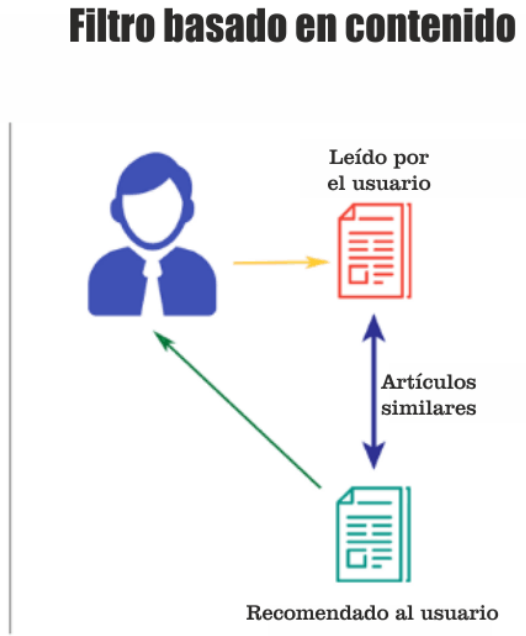
SISTEMAS DE RECOMENDACIÓN

CASOS DE USO

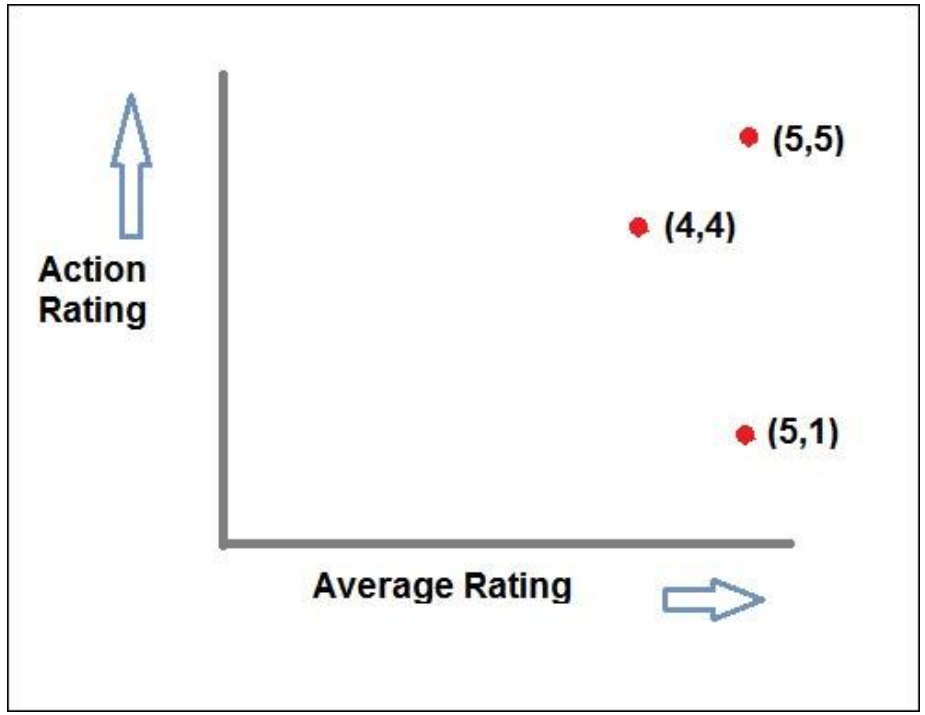

What to Watch
My Subscriptions
Music
16:50
Last Week Tonight with John Oliver: Online Harassment...
by LastWeekTonight
2,895,113 views • 1 week ago
4:01
BASTILLE feat. Ella - No Angels (HD)
by dancenationedm
5,836,404 views • 2 years ago
Bill Maher with Charlie Rose
by Darklordabc
88,923 views • 8 months ago
46:04
START! Walking at Home
merican Heart Associatio...
Leslie Sansone's Walk at Home
53,839 views • 7 months ago
6:17
The 5 Best Shia Labeouf "DO IT" Motivation Videos
by The VGC
1,390,971 views • 4 weeks ago
Elon Musk and Bill Gates discuss AI, entrepreneurshi...
by Every Elon Musk Video
105,539 views • 2 months ago

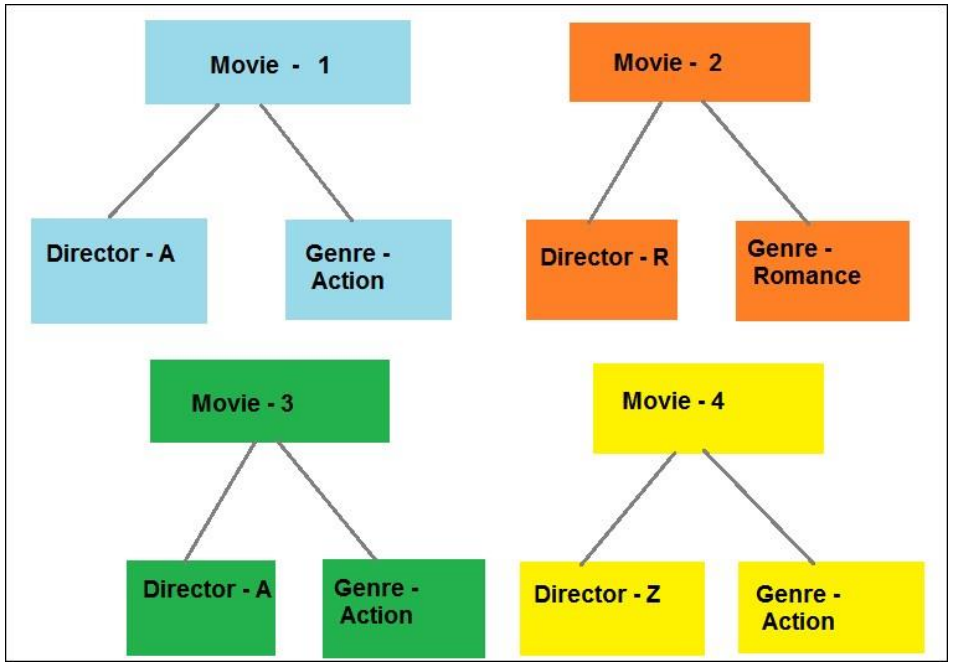
TIPOS DE SISTEMAS DE RECOMENDACIÓN

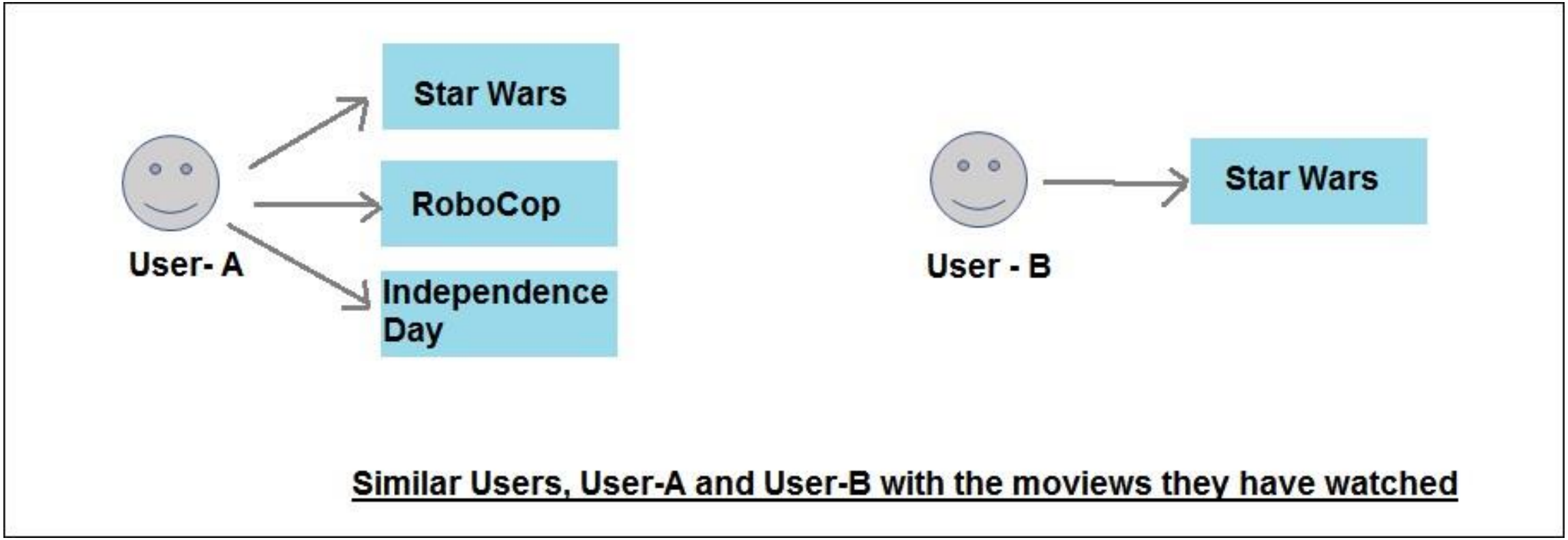




BASADOS EN CONTENIDO







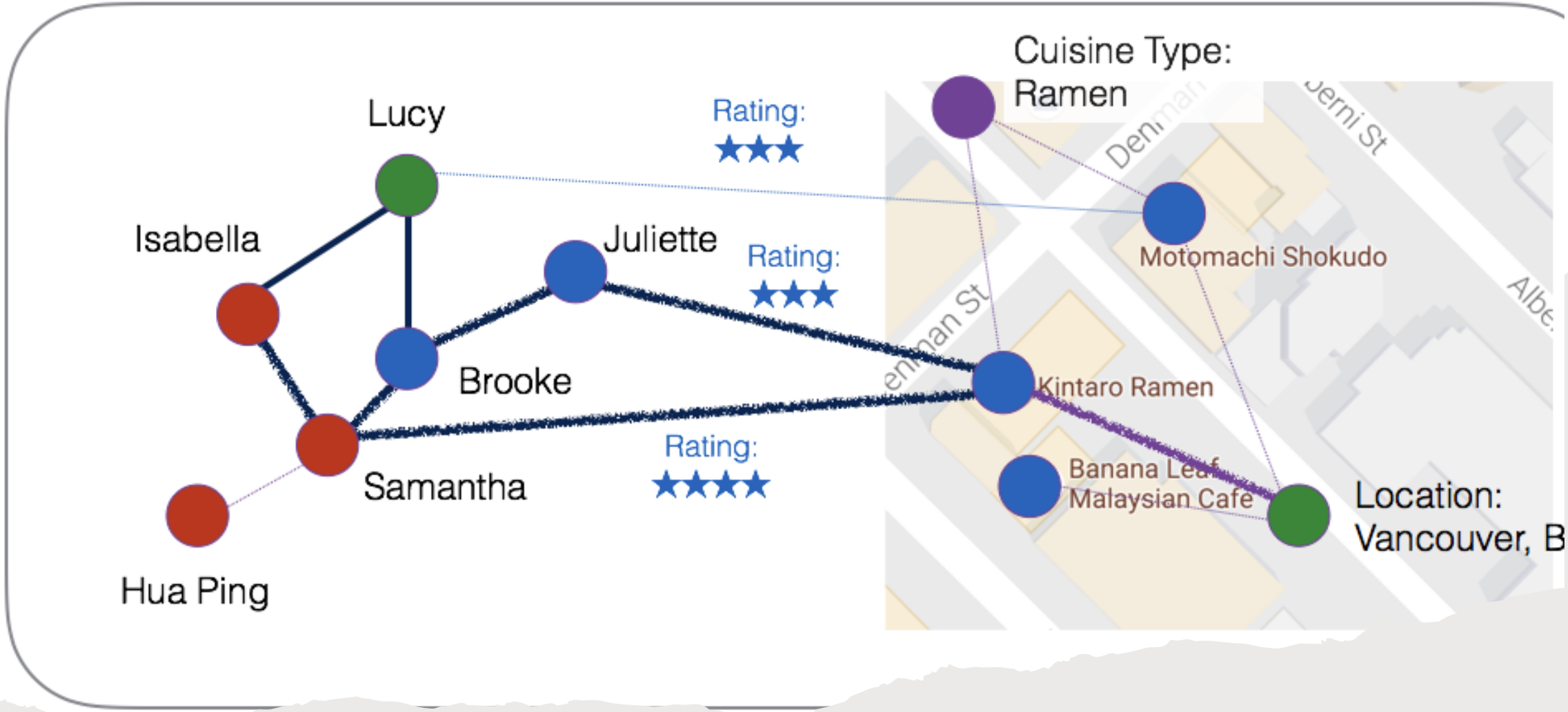
FILTRO COLABORATIVO


'drop'
];
events.forEach(e => {
fileDropZone.addEventListener(e, (ev) => {
ev.preventDefault();
if (ev.type === 'dragenter') {
fileDropZone.classList.add('solid-border');
}
if (ev.type === 'dragleave') {
fileDropZone.classList.remove('solid-border');
}
if(ev.type === 'drop') {
fileDropZone.classList.remove('solid-border');
ev.dataTransfer.files)
'border rounded

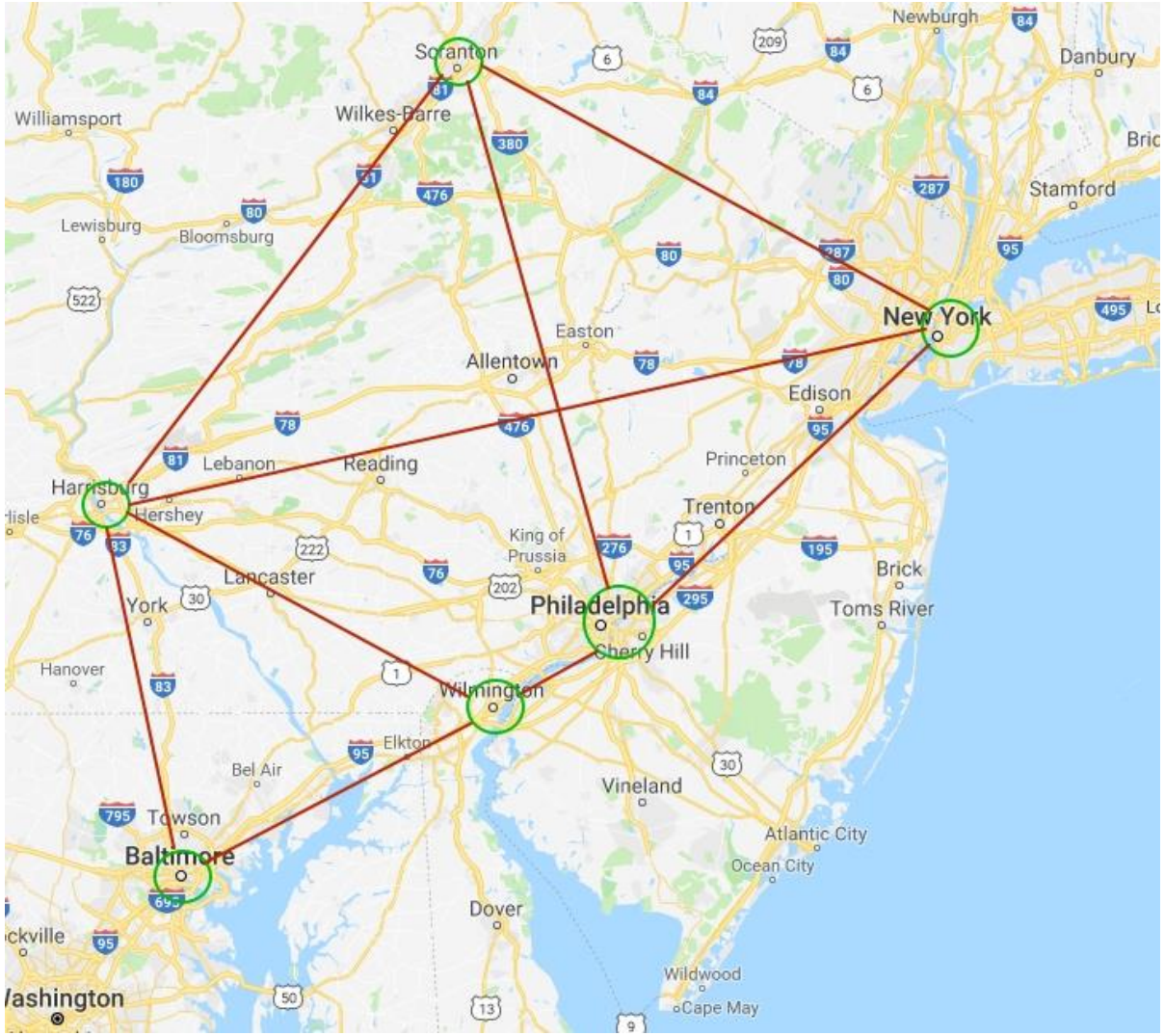
NOTEBOOK

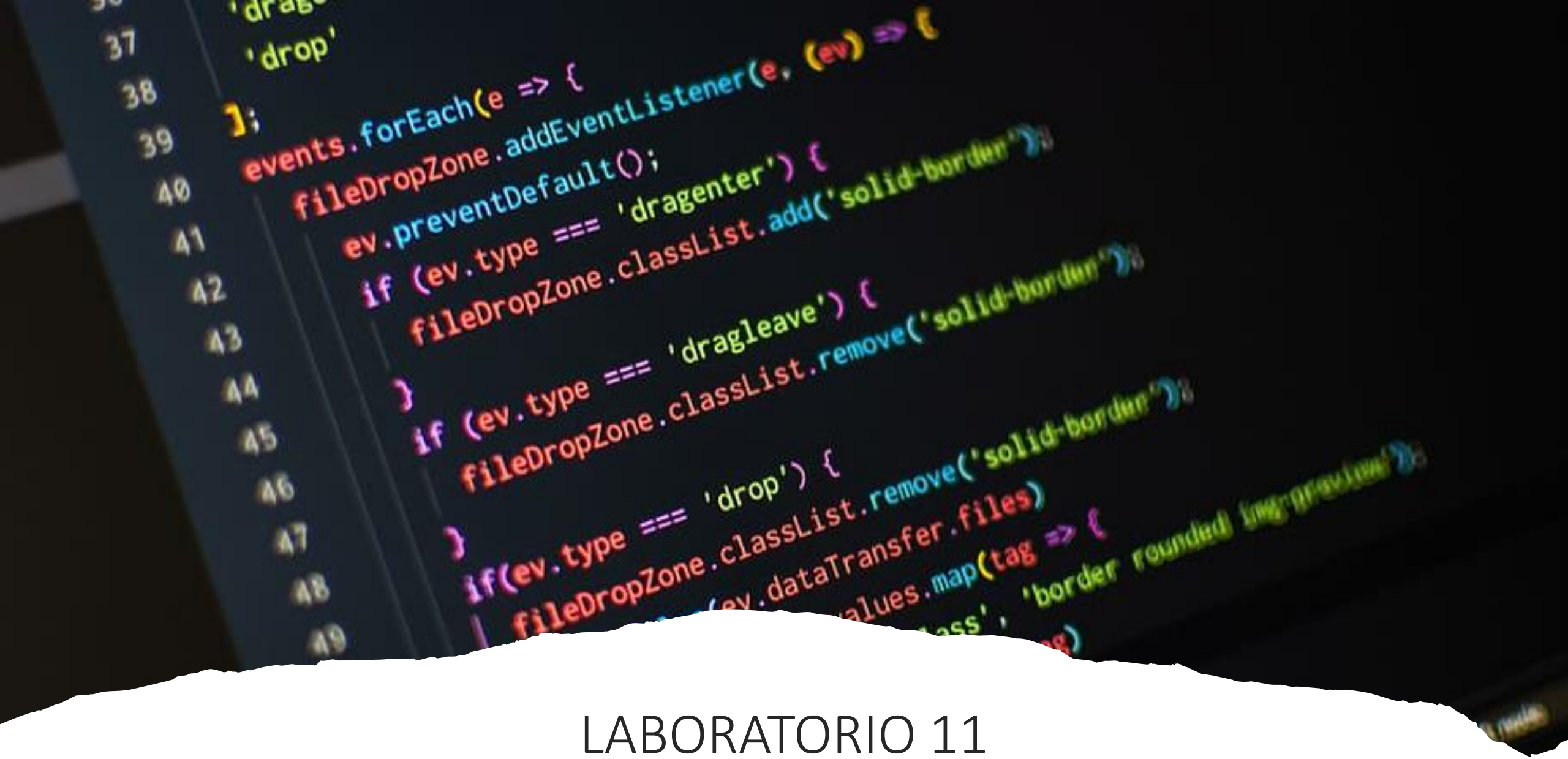
CASOS DE USO DE GRAFOS

social network + restaurant recommendations

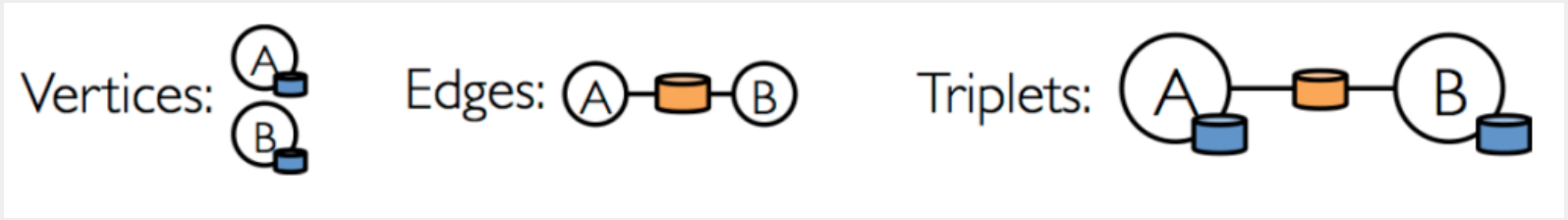


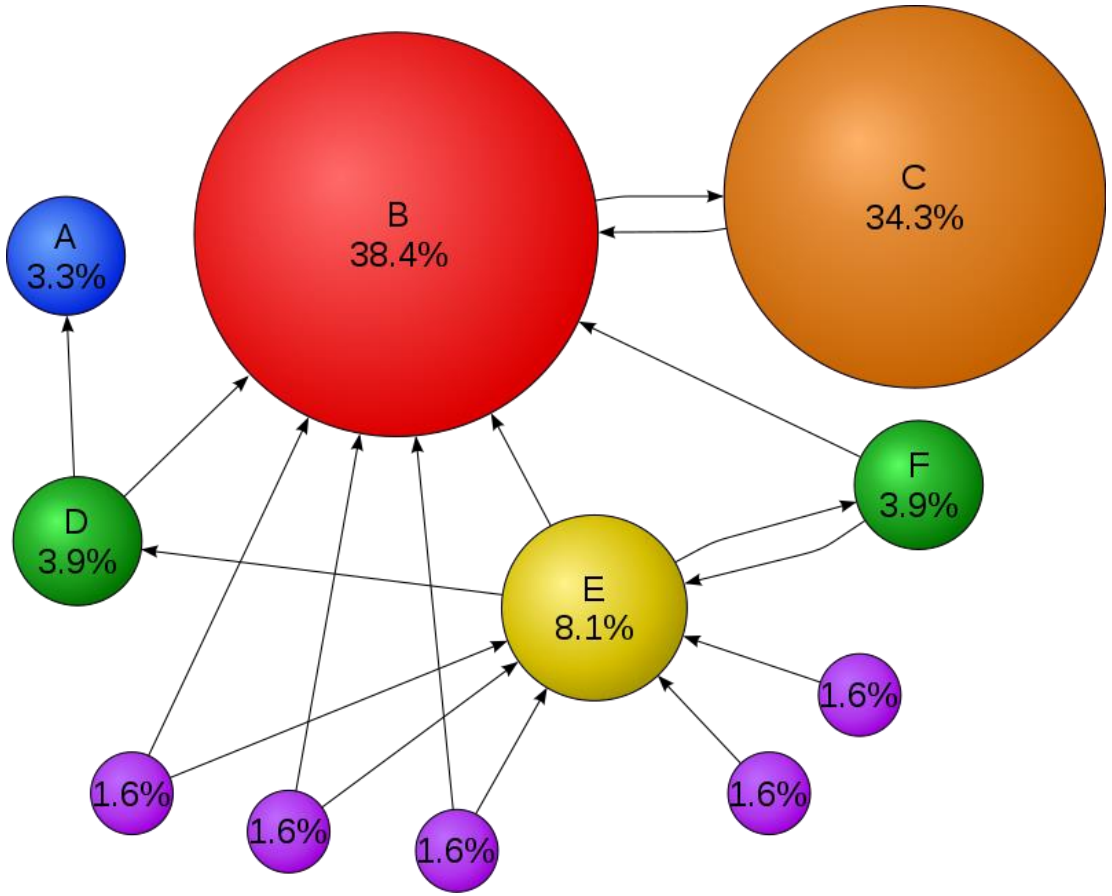
GRAFO DEL EJEMPLO



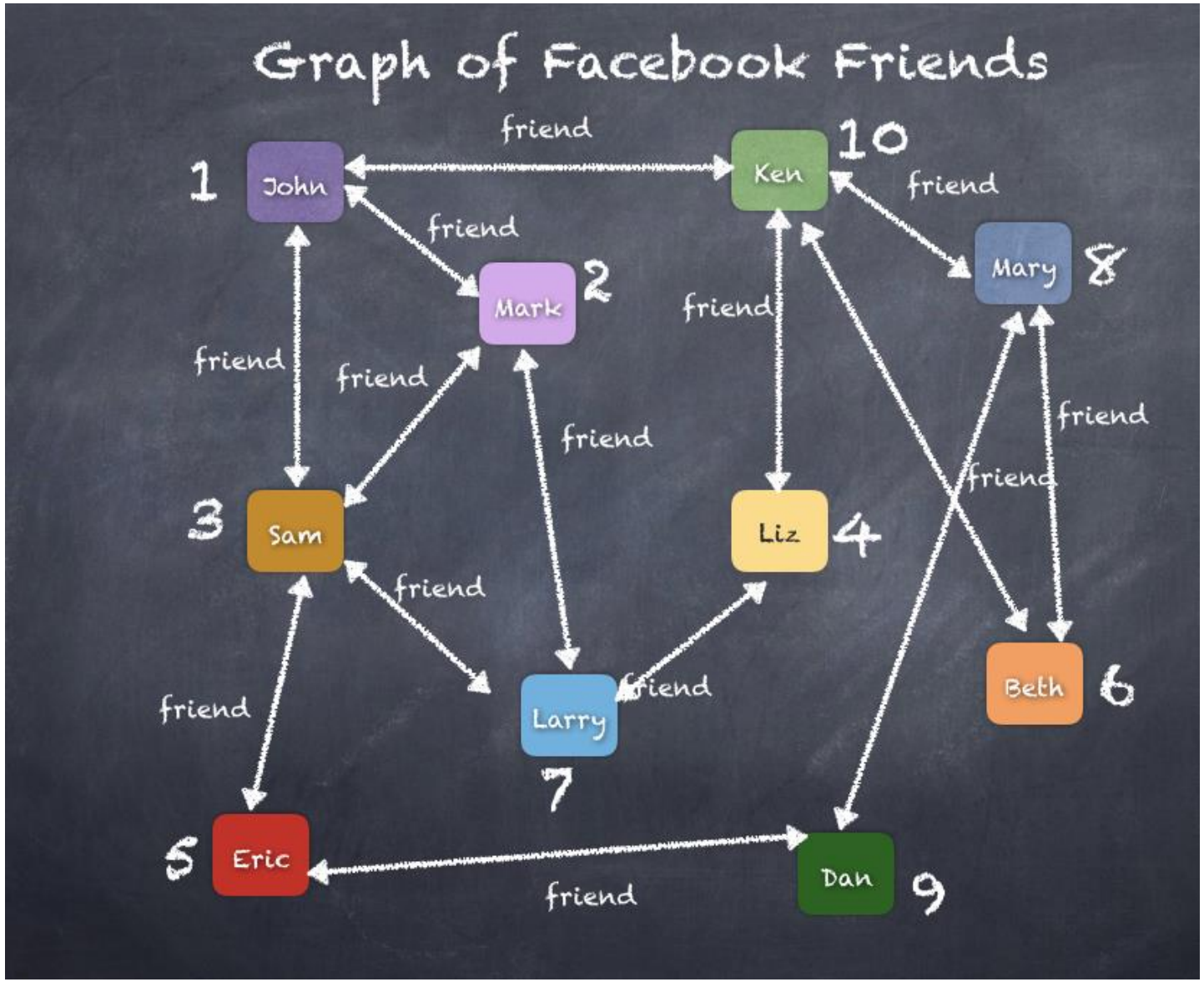


LABORATORIO 11

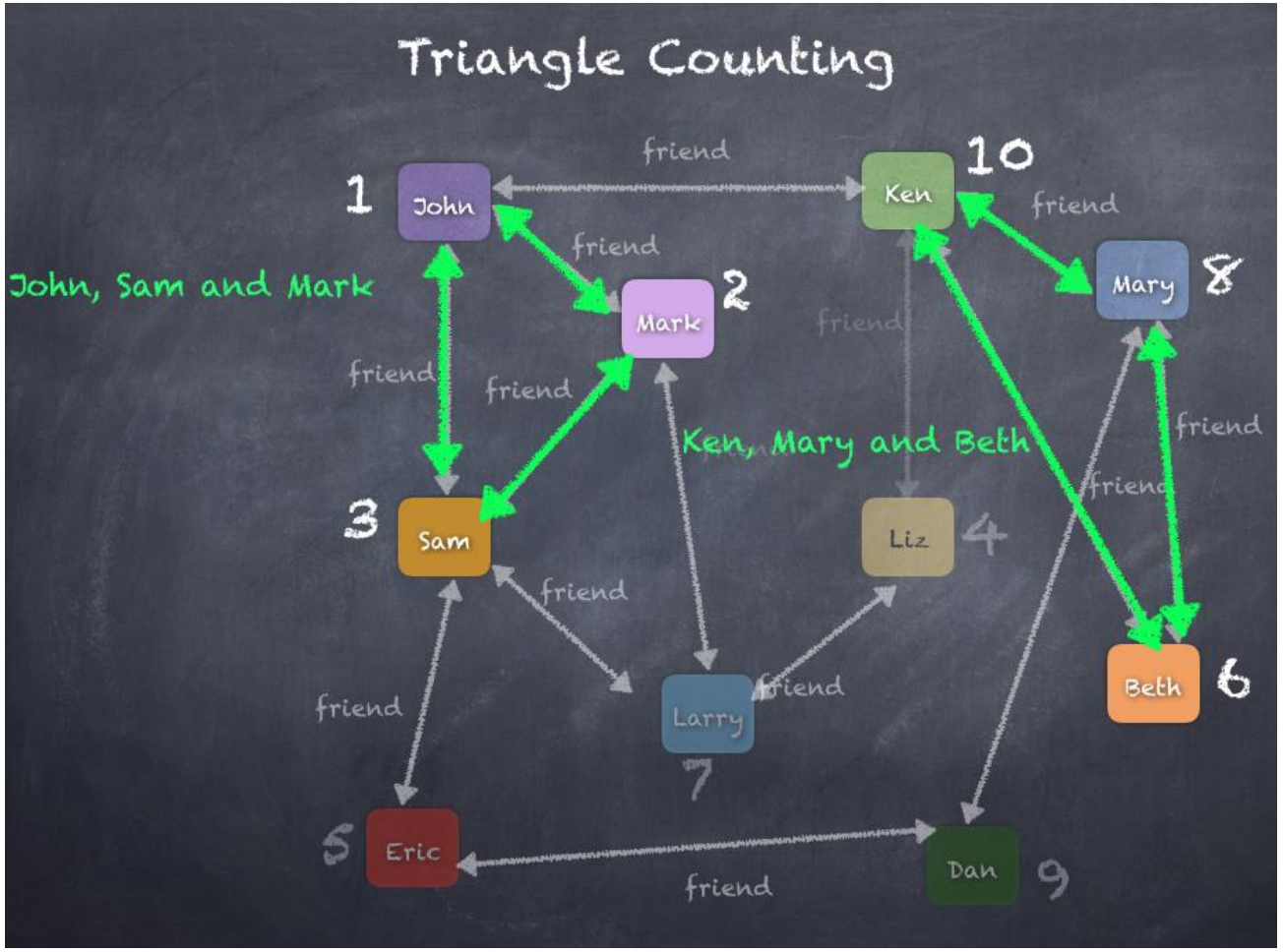

Vertices:
A
B
Edges:
A
B
Triplets:
A
B



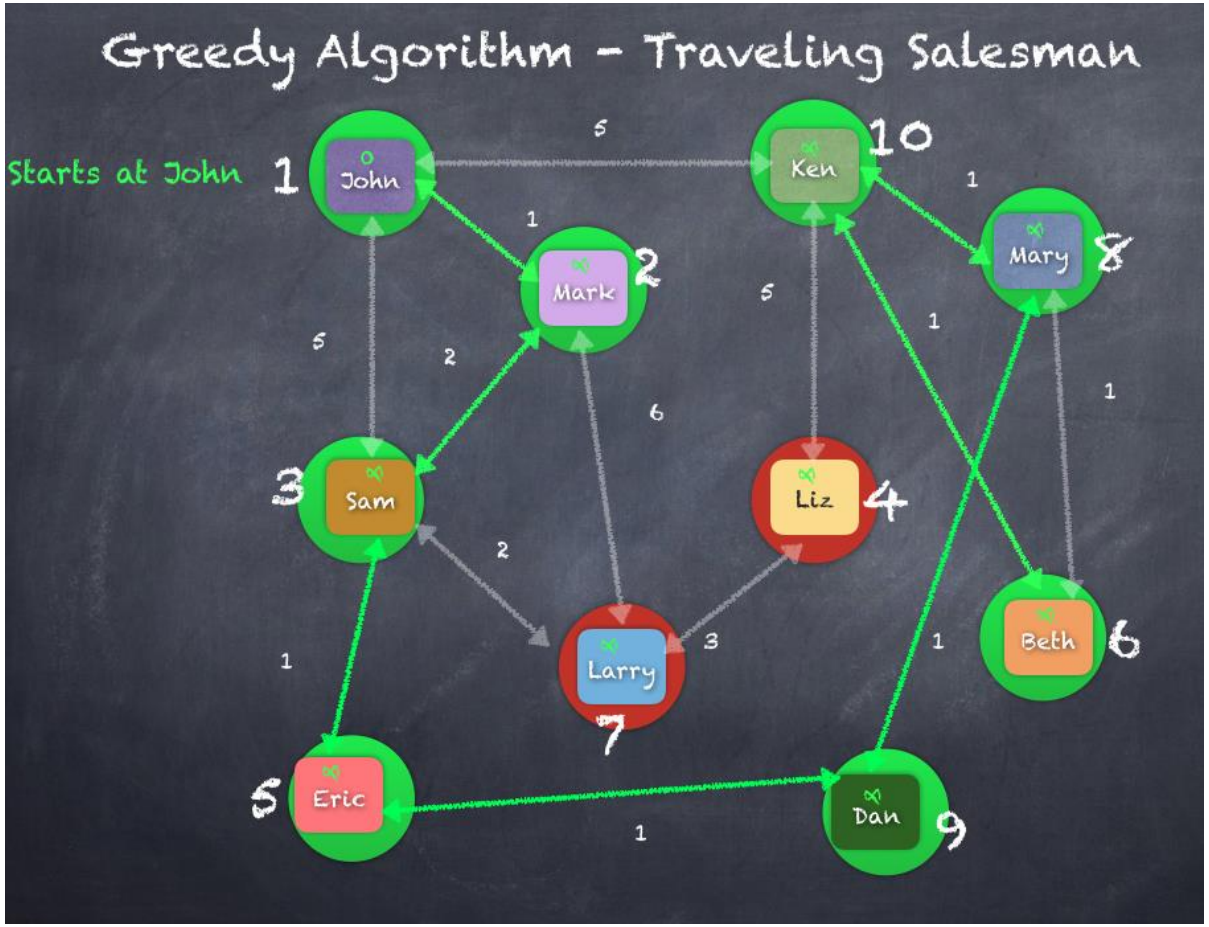
ALGORITMOS


Graph of Facebook Friends
1
John
2
Mark
3
Sam
4
Liz
5
Eric
6
Beth
7
Larry
8
Mary
9
Dan
10
Ken
friend

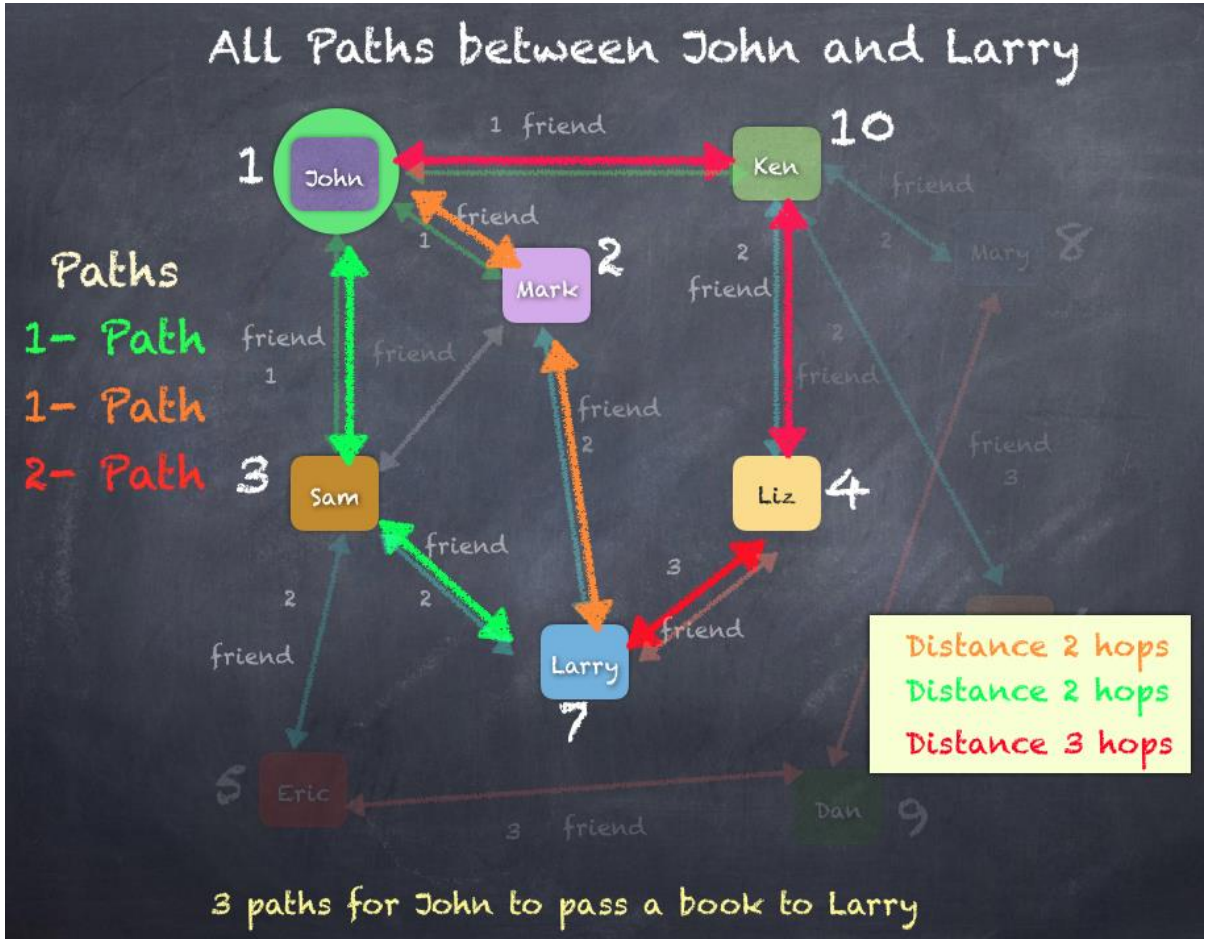
TRIANGULOS


Triangle Counting
1
John
friend
10
Ken
friend
friend
Mary
8
John, Sam and Mark
Mark
2
friend
friend
friend
friend
Ken, Mary and Beth
friend
3
Sam
Liz
4
friend
friend
Larry
friend
Beth
6
7
5
Eric
friend
Dan
9

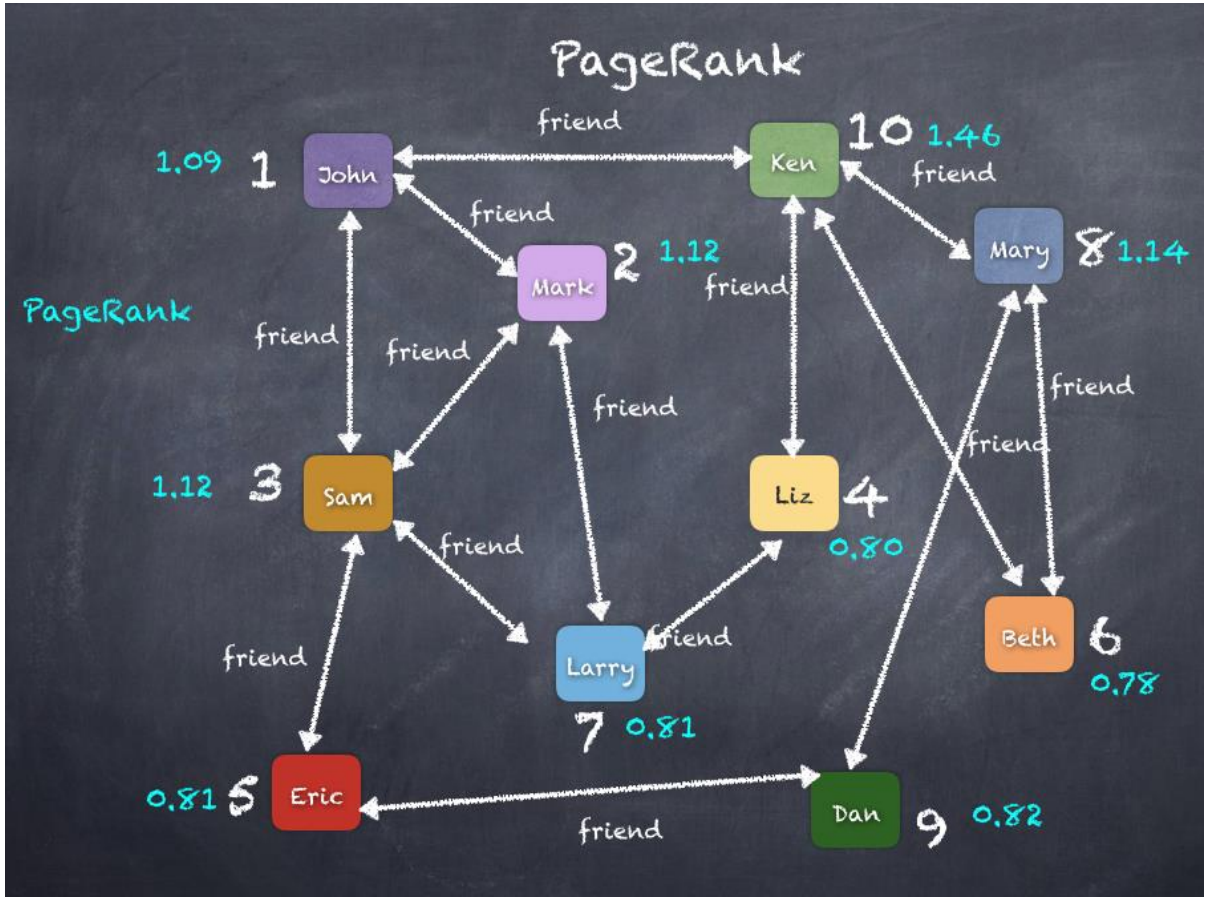
PROBLEMA DEL VENDEDOR AMBULANTE



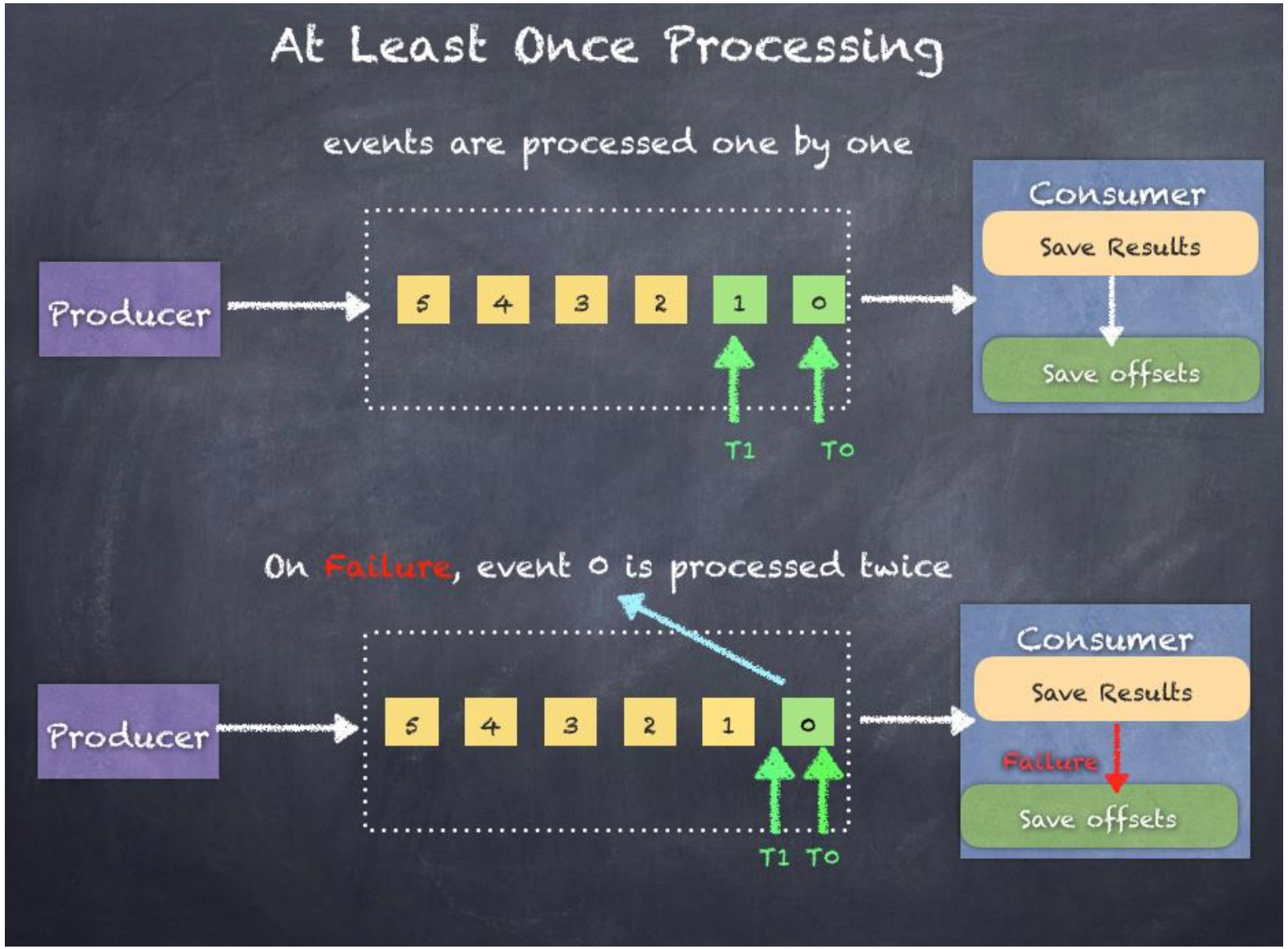
RUTA MAS CORTA

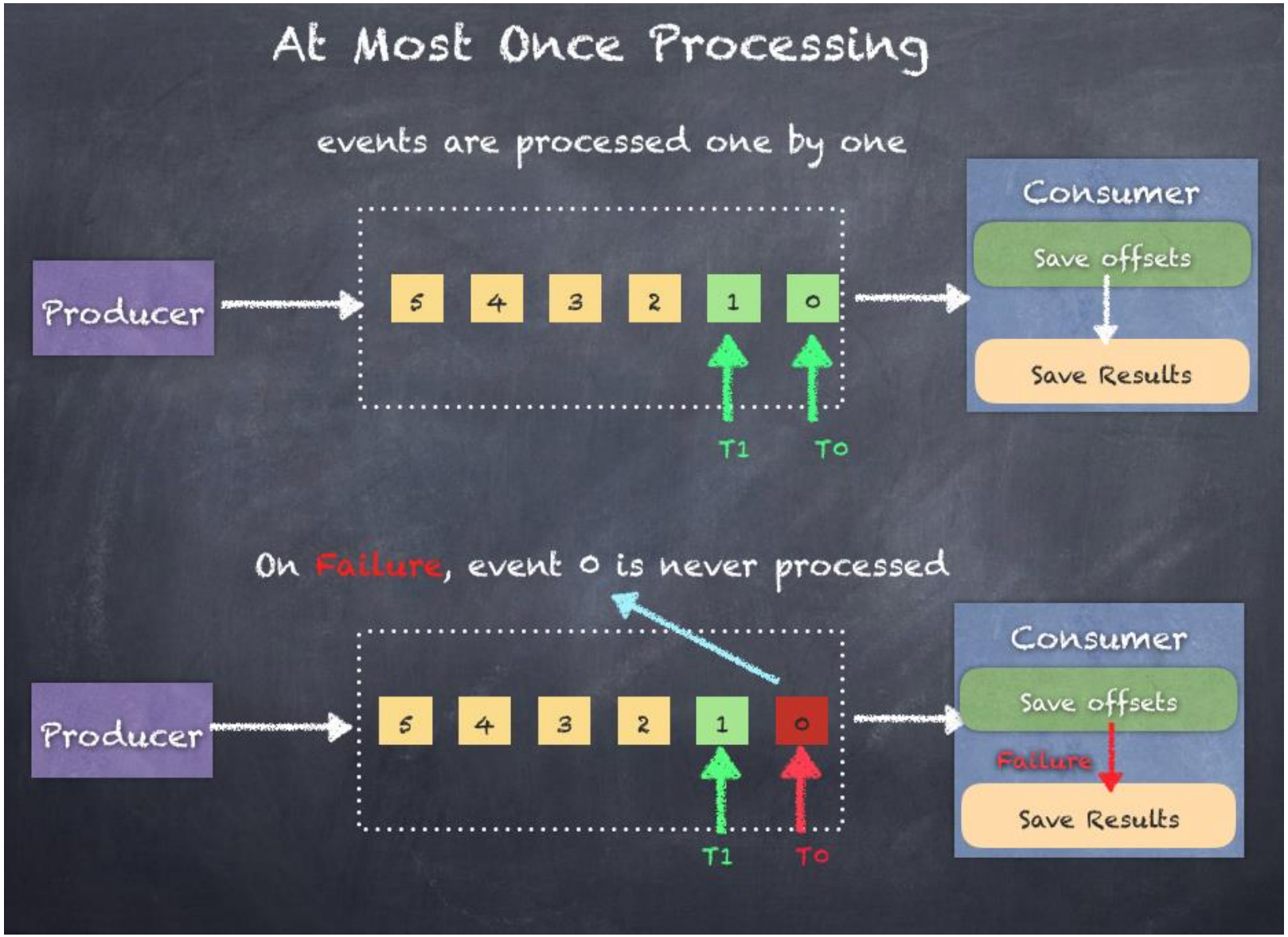


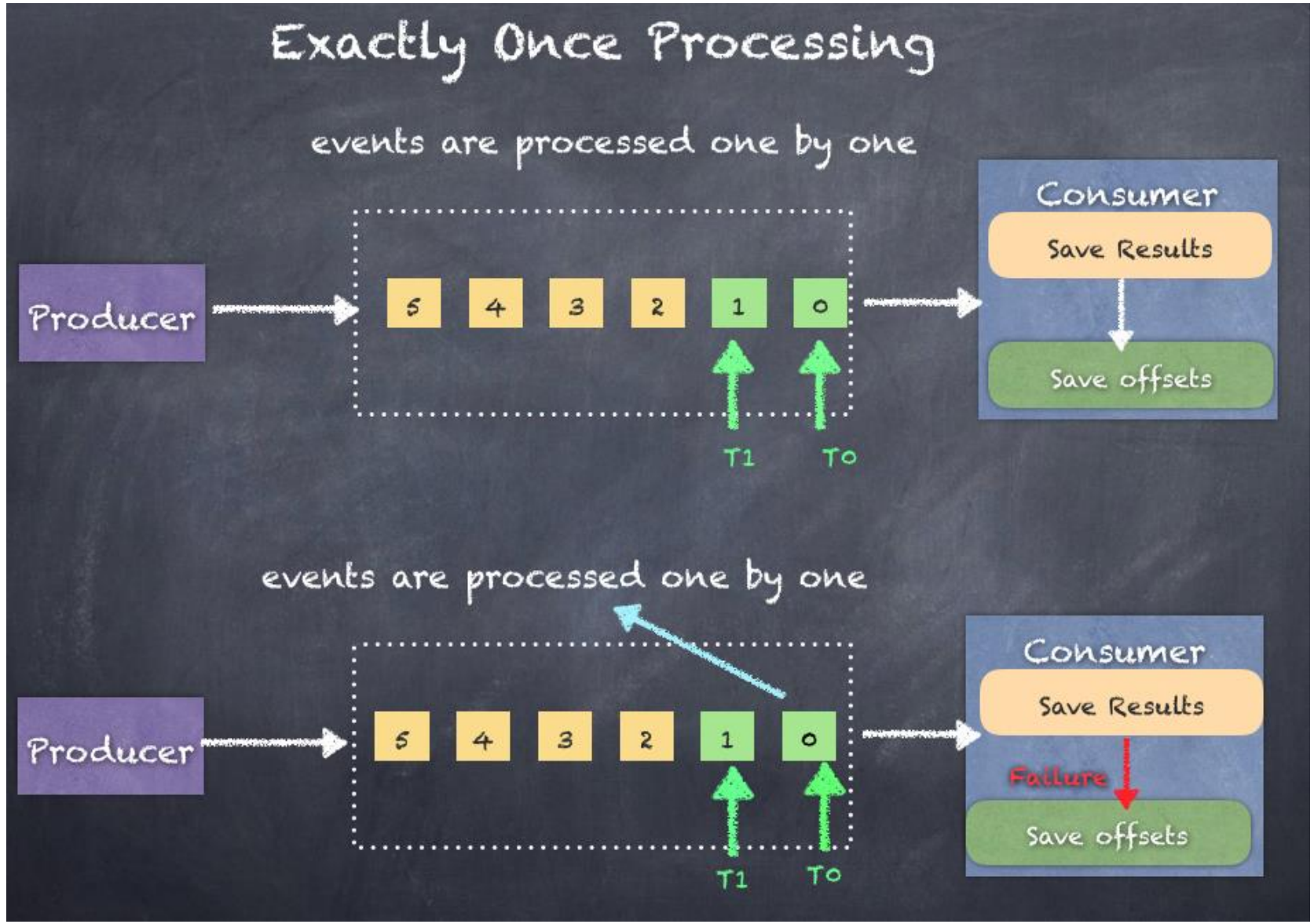
PAGE RANK


PageRank
PageRank
1.09 1 John
Ken 10 1.46
Mark 2 1.12
Mary 8 1.14
1.12 3 Sam
Liz 4 0.80
Larry 7 0.81
Beth 6 0.78
0.81 5 Eric
Dan 9 0.82
friend

STREAMING


At Least Once Processing
events are processed one by one
Producer
5
4
3
2
1
0
T1
T0
Consumer
Save Results
Save offsets
On Failure, event 0 is processed twice
Producer
5
4
3
2
1
0
T1
T0
Consumer
Save Results
Failure
Save offsets

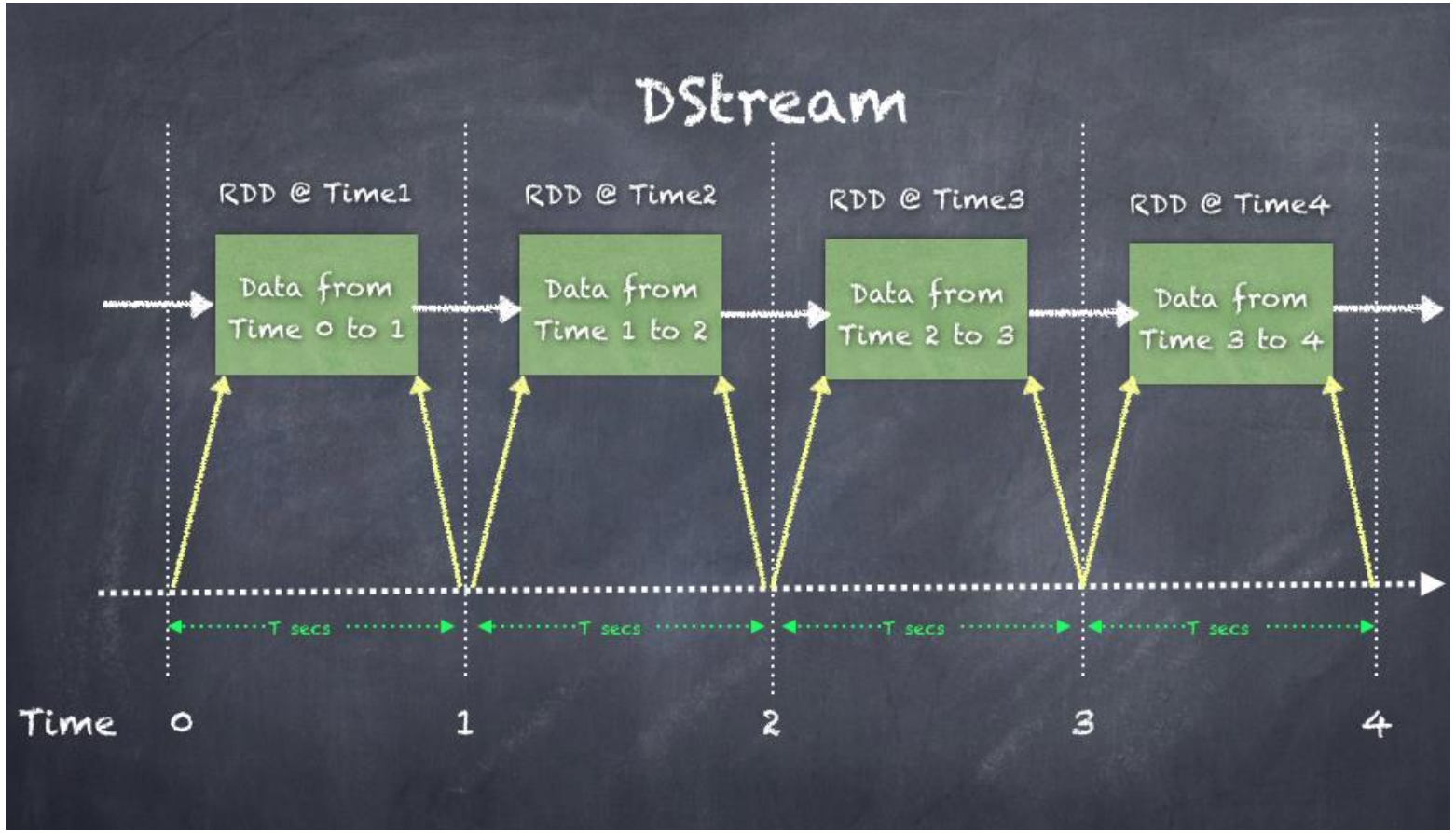

At Most Once Processing
events are processed one by one
Producer
5
4
3
2
1
0
T1
T0
Consumer
Save offsets
Save Results
On Failure, event 0 is never processed
Producer
5
4
3
2
1
0
T1
T0
Consumer
Save offsets
Failure
Save Results


Exactly Once Processing
events are processed one by one
Producer
5
4
3
2
1
0
T1
T0
Consumer
Save Results
Save offsets
events are processed one by one
Producer
5
4
3
2
1
0
T1
T0
Consumer
Save Results
Failure
Save offsets

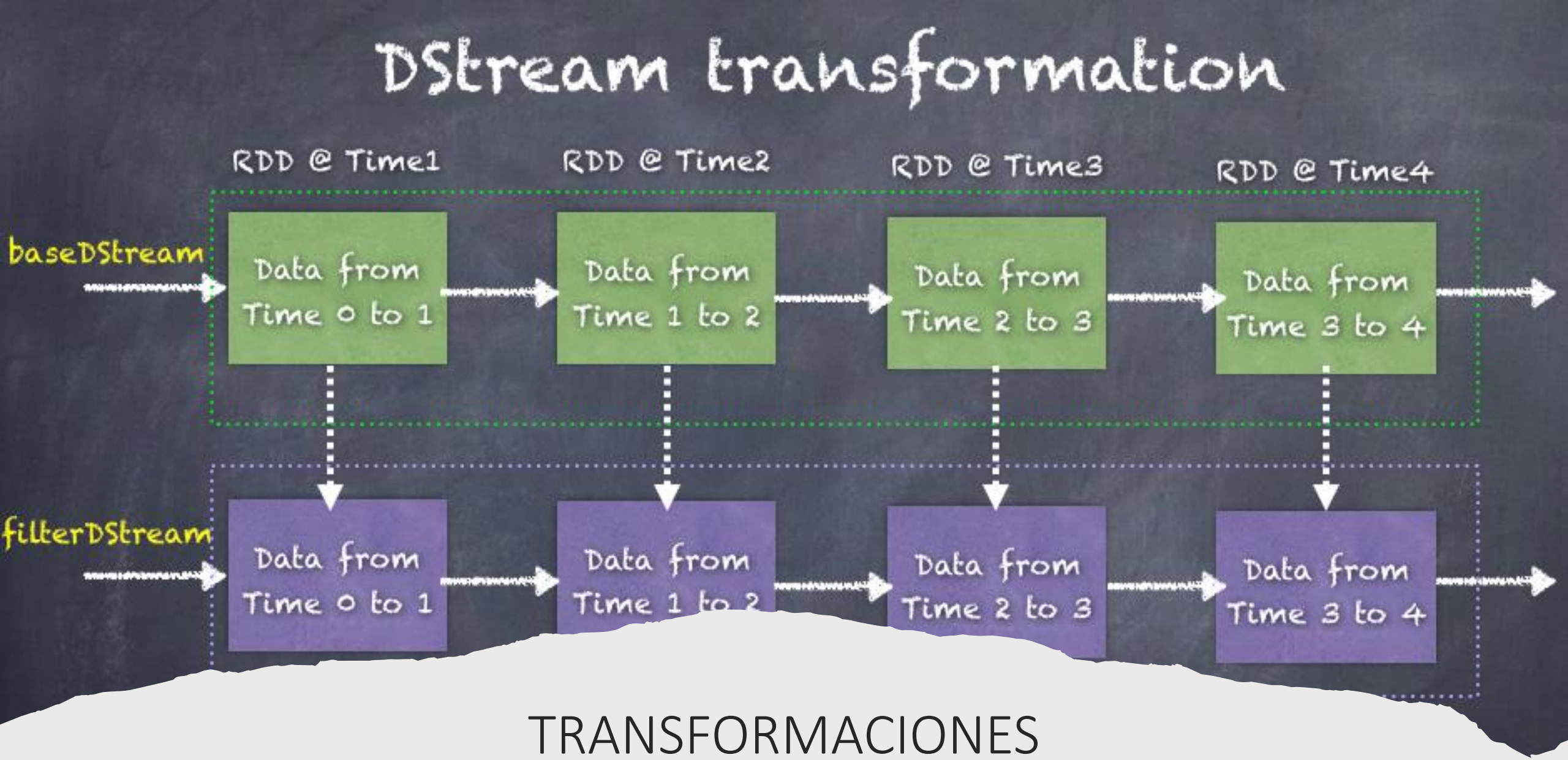
SPARK STREAMING


input data stream
Spark Streaming
batches of input data
Spark Engine
batches of processed data

MICROLOTES


DStream
RDD @ Time1
RDD @ Time2
RDD @ Time3
RDD @ Time4
Data from Time 0 to 1
Data from Time 1 to 2
Data from Time 2 to 3
Data from Time 3 to 4
T secs
T secs
T secs
T secs
Time
0
1
2
3
4

DSTREAMS



TRANSFORMACIONES


'drop'
];
events.forEach(e => {
fileDropZone.addEventListener(e, (ev) => {
ev.preventDefault();
if (ev.type === 'dragenter') {
fileDropZone.classList.add('solid-border');
}
if (ev.type === 'dragleave') {
fileDropZone.classList.remove('solid-border');
}
if(ev.type === 'drop') {
fileDropZone.classList.remove('solid-border');
ev.dataTransfer.files)
'border rounded

NOTEBOOK