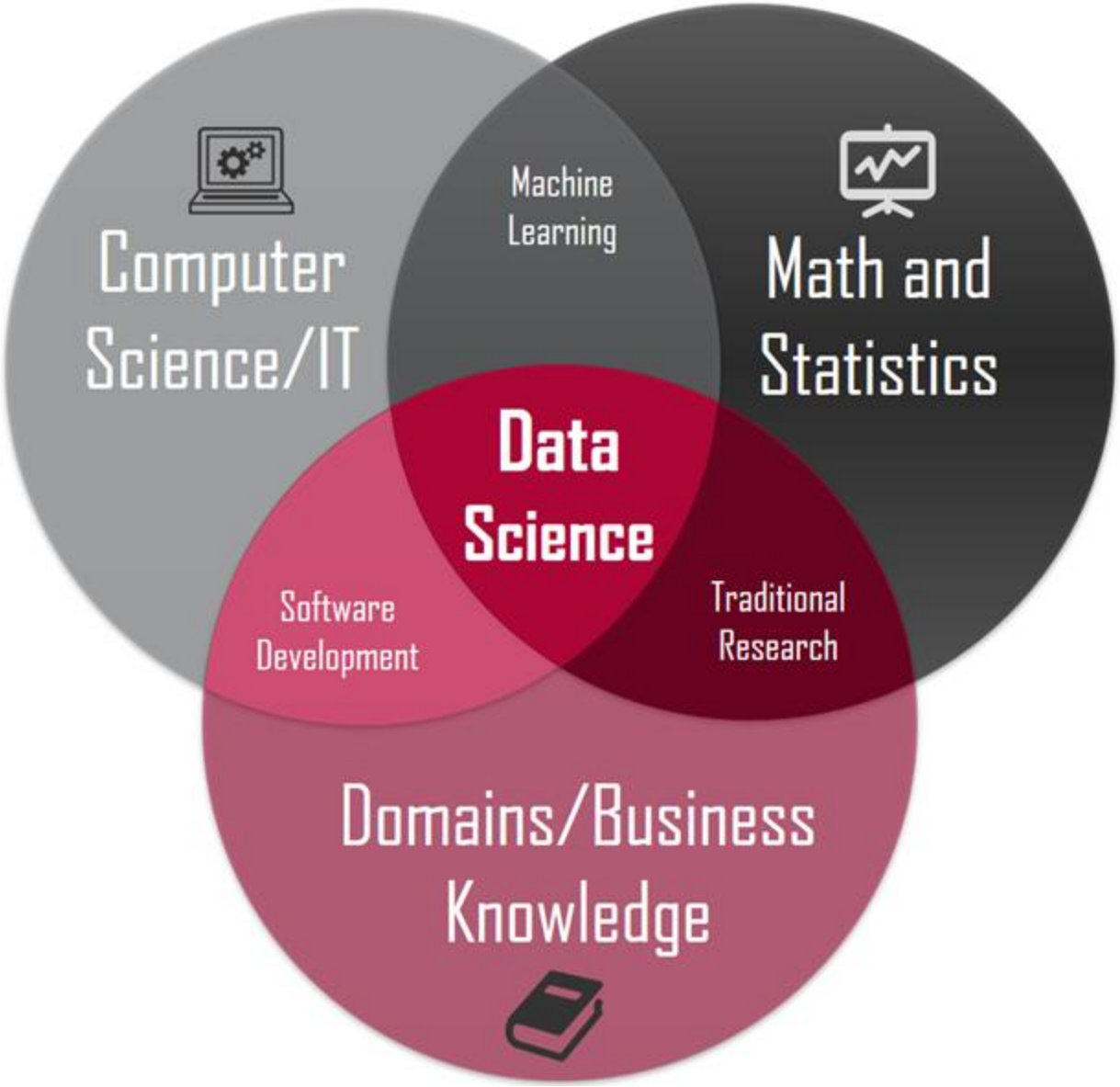
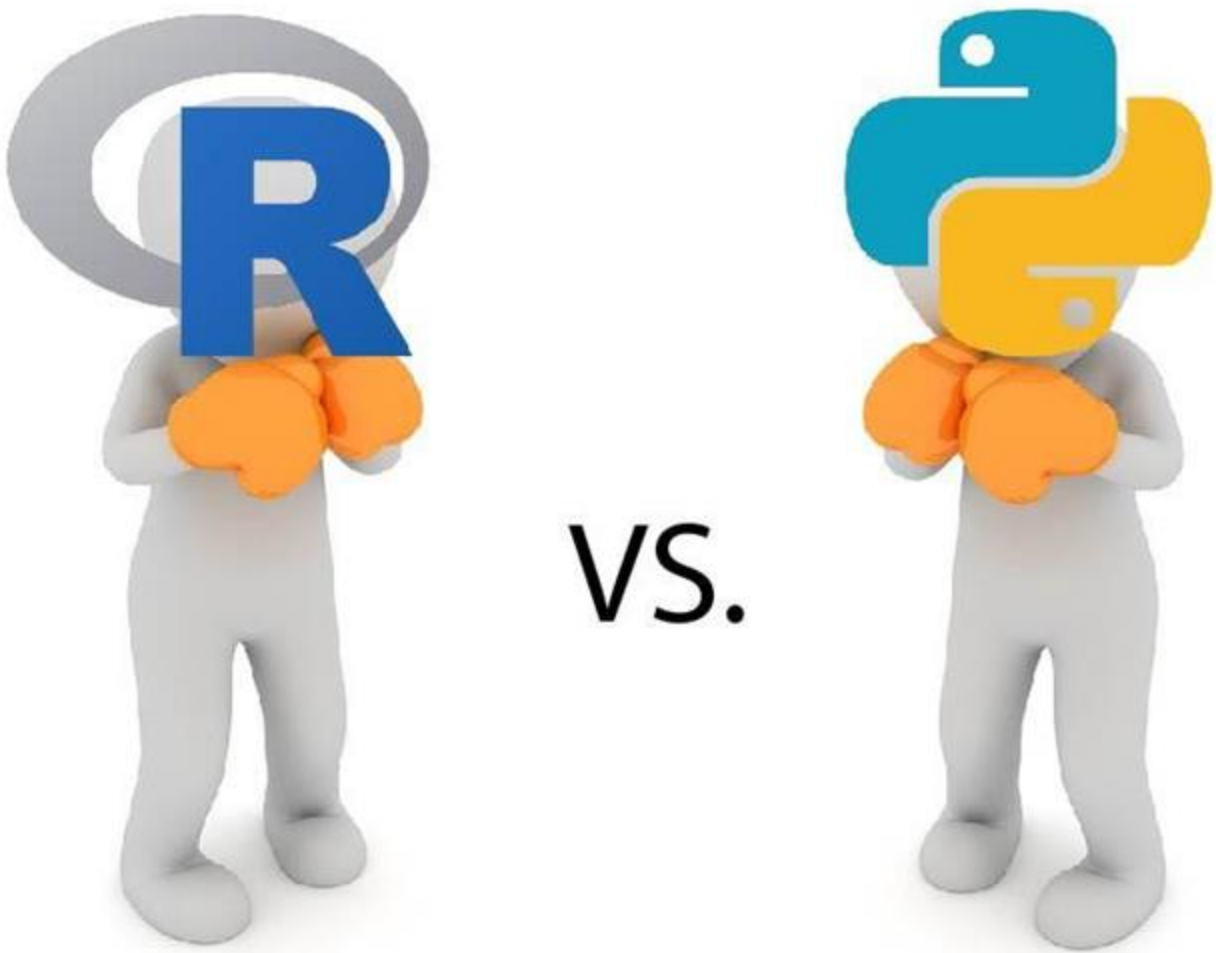
Introducción a la Ciencia de Datos

Sesión 1

¿Qué herramientas utiliza un Científico de Datos?



¿Qué lenguaje de programación utilizar?



¿Dónde comenzar a aprender rápido y gratis?

kaggle

https://www.kaggle.com/learn/overview

https://courses.analyticsvidhya.com/

https://towardsdatascience.com/

METODOLOGÍA
CRISP - DM

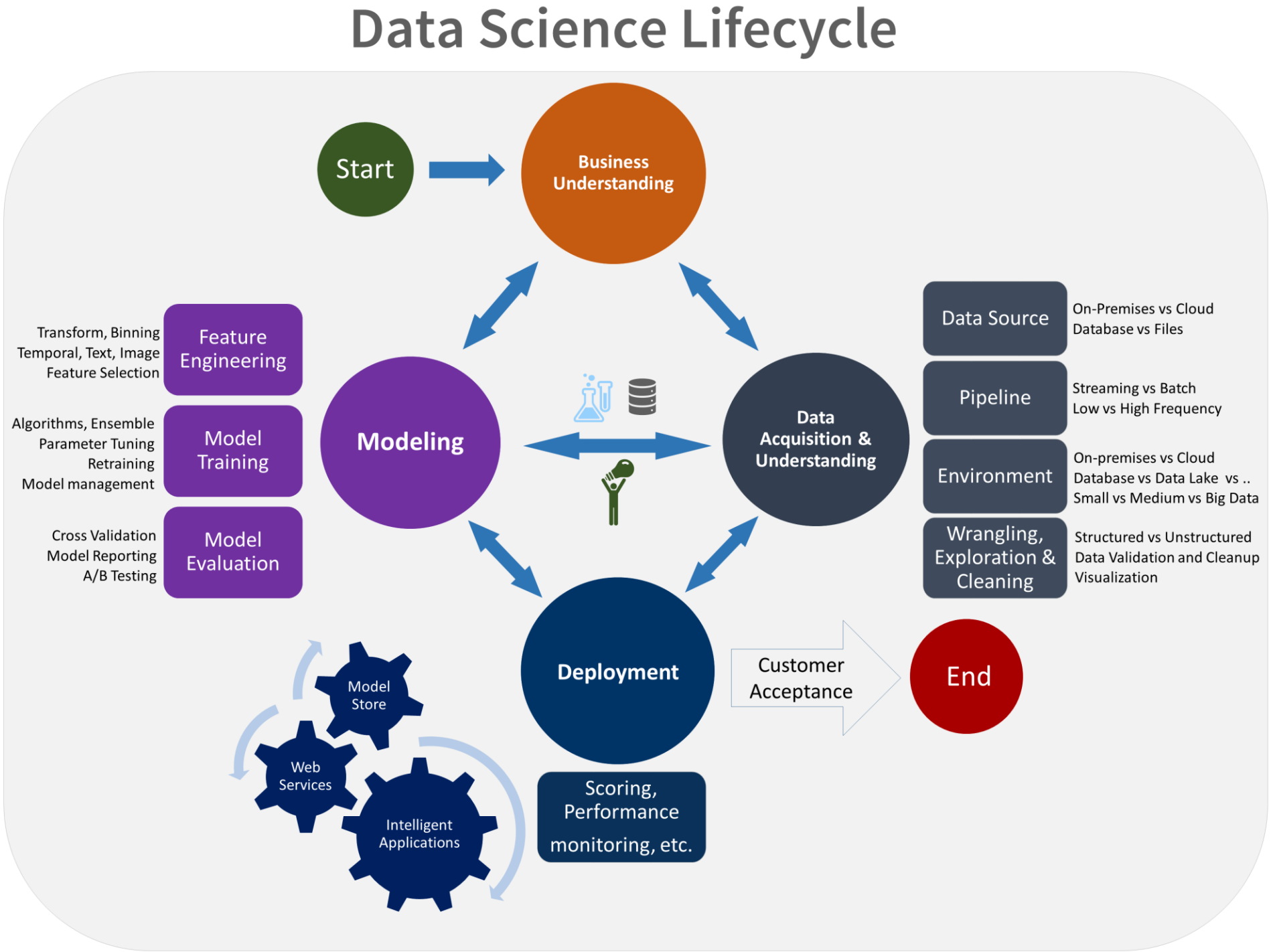
Cross-Industry Standard Process for Data Mining

(finales 90’s)

METODOLOGÍA TDSP

Team Data Science Process

(Microsoft, 2016)



Identifica una necesidad del mercado

1

¿Algún problema que puedas resolver aplicando análisis de datos?

Captura los datos necesarios y limpia la información

2

¿Tienes todos los datos que necesitas? ¿Los datos están legibles?

Selecciona y evalúa el modelo

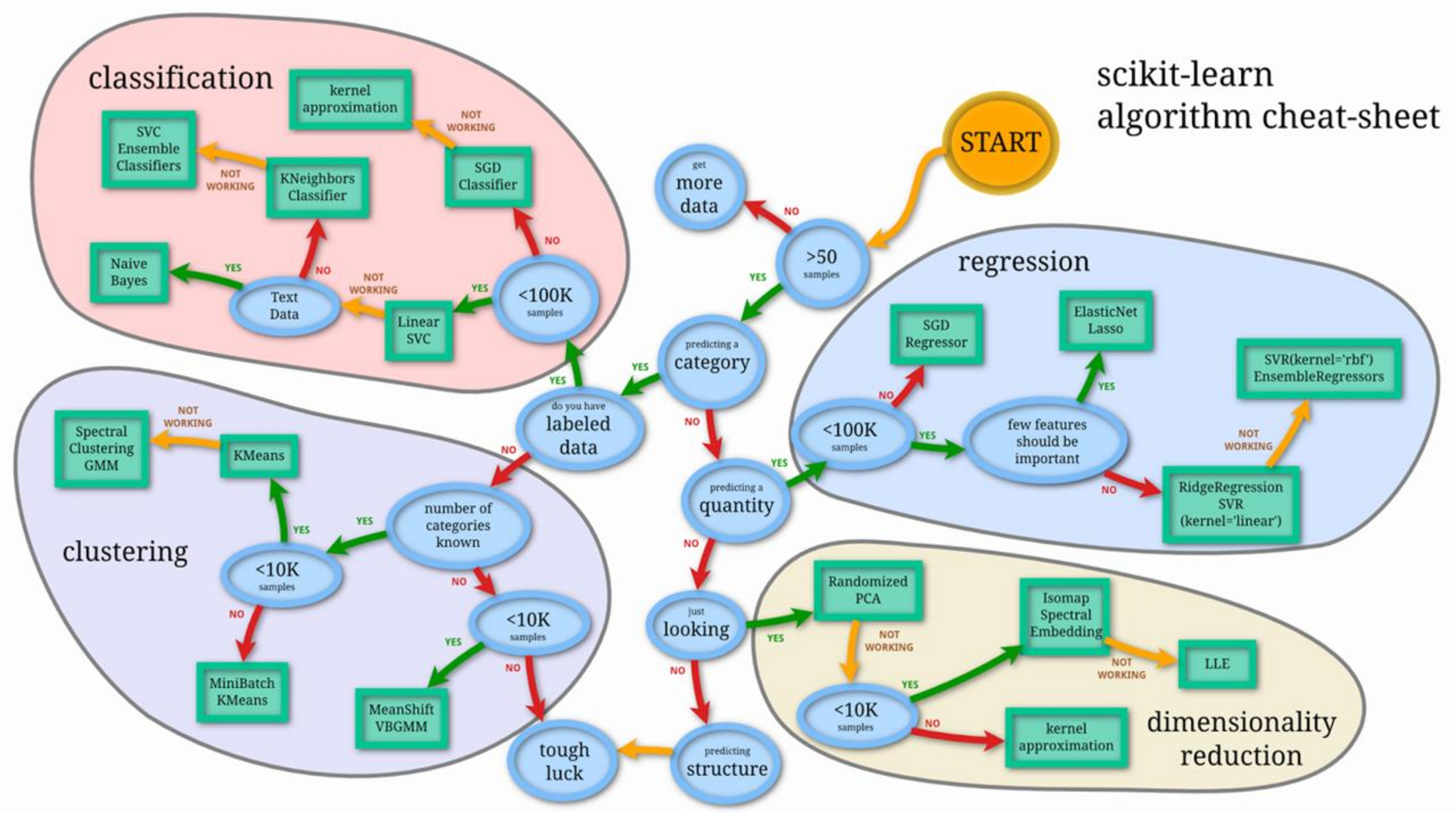
3

¿El modelo es el mejor para los datos que tienes? ¿El modelo es preciso?

Implementa el modelo para uso de tus clientes

4

¿Tu modelo puede salir a producción? ¿Tienes la infraestructura?


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
YES
Linear SVC
NOT WORKING
Text Data
NO
Naive Bayes
YES
NO
KNeighbors Classifier
NOT WORKING
SVC Ensemble Classifiers
NOT WORKING
kernel approximation
NOT WORKING
SGD Classifier
NO
classification
NO
predicting a quantity
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
YES
ElasticNet Lasso
NO
RidgeRegression SVR (kernel='linear')
NOT WORKING
SVR(kernel='rbf') EnsembleRegressors
regression
NO
just looking
YES
Randomized PCA
NOT WORKING
<10K samples
YES
Isomap Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
NO
number of categories known
YES
<10K samples
YES
KMeans
NOT WORKING
Spectral Clustering GMM
NO
<10K samples
YES
MeanShift VBGMM
NO
MiniBatch KMeans
clustering

Aprendizaje No Supervisado

Clusterización con Algoritmo Kmeans

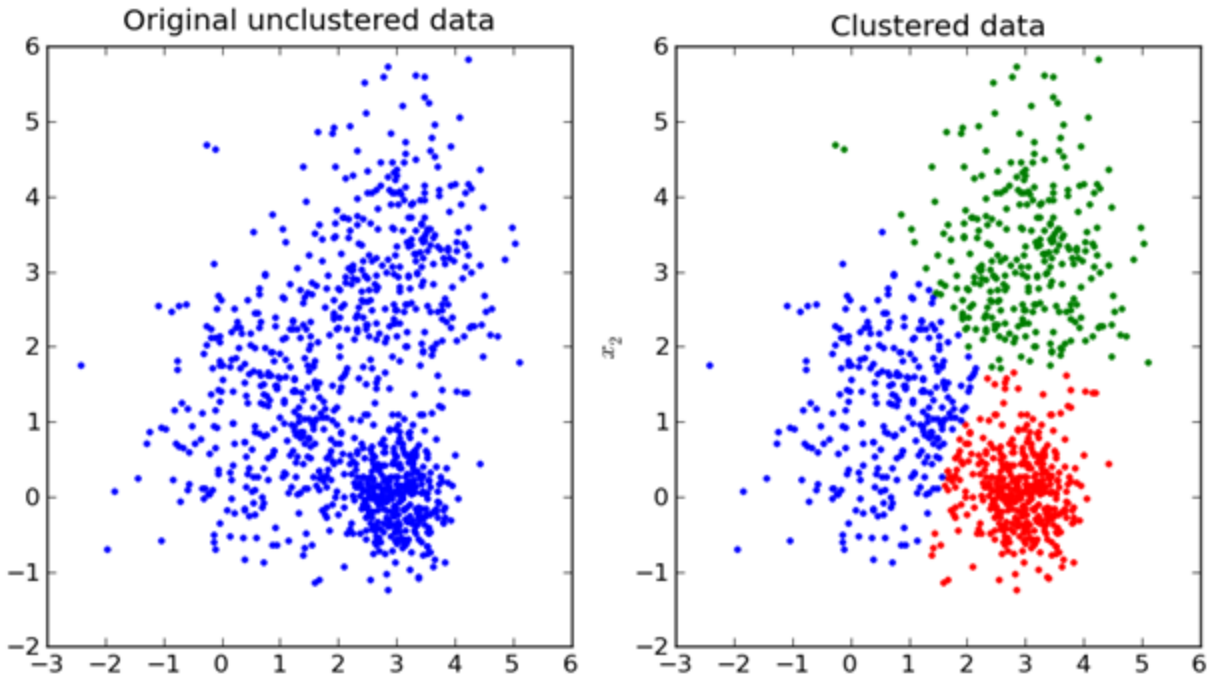
AGRUPAR ELEMENTOS

- ✓ Recopilar un conjunto
- ✓ Escoger un criterio aleatorio
- ✓ Ejecutar criterio sobre el conjunto

ORDENAR Y CLASIFICAR

- ✓ Identificar cada elemento
- ✓ Descubrir una característica en común
- ✓ Dar escala a esa característica
- ✓ Asignar cada elemento a un grupo según escala

CLUSTERING



1. Recopilar datos
2. Identificar centroides
3. Asignar cada dato a un *centroide*[1]
4. Formar clusters

[1] https://youtu.be/V28U9cGeSdA

FUNDAMENTO MATEMÁTICO

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}.$$

✓ Definir una distancia **euclidiana**

$$\arg\min_{c_i \in C} dist(c_i, x)^2$$

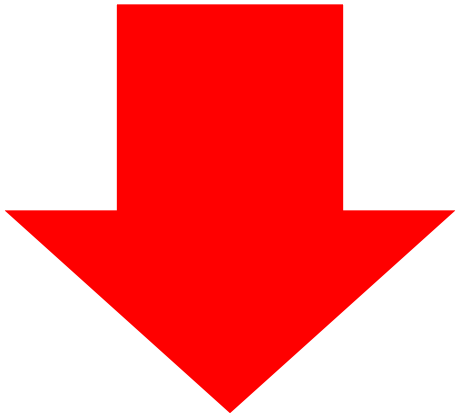
✓ Asignar cada dato al centroide (c_i) separado por la mínima distancia

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$$

✓ Hallar nuevo centroide y repetir

https://muthu.co/mathematics-behind-k-mean-clustering-algorithm/

EQUILIBRIO BUSCADO



Minimizar la **distancia** entre cada elemento **dentro** del cluster)

- ✓ Los elementos de un mismo grupo son muy similares

Maximizar la **distancia** entre cada elemento **fuera** del cluster

- ✓ Un grupo se distingue de otro fácilmente

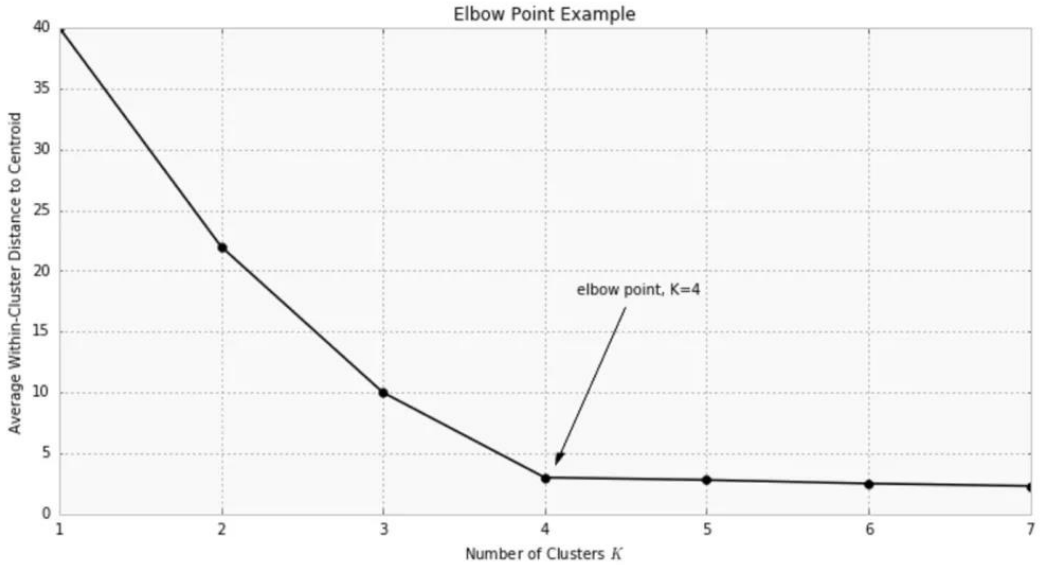
REQUISITOS DEL MODELO

- ✓ Los datos deben ser numéricos y continuos (no flags)

	usuario	op	co	ex	ag
0	3gerardpique	34.297953	28.148819	41.948819	29.370315
1	aguerosergiokun	44.986842	20.525865	37.938947	24.279098
2	albertochicote	41.733854	13.745417	38.999896	34.645521
3	AlejandroSanz	40.377154	15.377462	52.337538	31.082154
4	alfredocasero1	36.664677	19.642258	48.530806	31.138871

- ✓ Definir previamente el número de cluster, es decir el número final de segmentos deseados

¿CÓMO DETERMINAR EL NÚMERO DE CLUSTER?



- ✓ Iterar KMeans con un número **cada vez mayor** de cluster
- ✓ Identificar cuando la **mejora** del score sea **mínima**

Let's Code

https://drive.google.com/file/d/16EOFMzByzl5GapnV30Cv7BAxIeO81DRF/view?usp=sharing