Aprendizaje de Máquina y Minería de Datos
Clase II

Curso Exploratorio de Computación – IIC 1005
Vicente Domínguez
2018-2

Google Speech Recognition

- https://www.youtube.com/watch?v=D5VN56jQMWM

Inteligencia Artificial

- **Objetivo**: Construir máquinas/software que exhiban comportamiento inteligente.

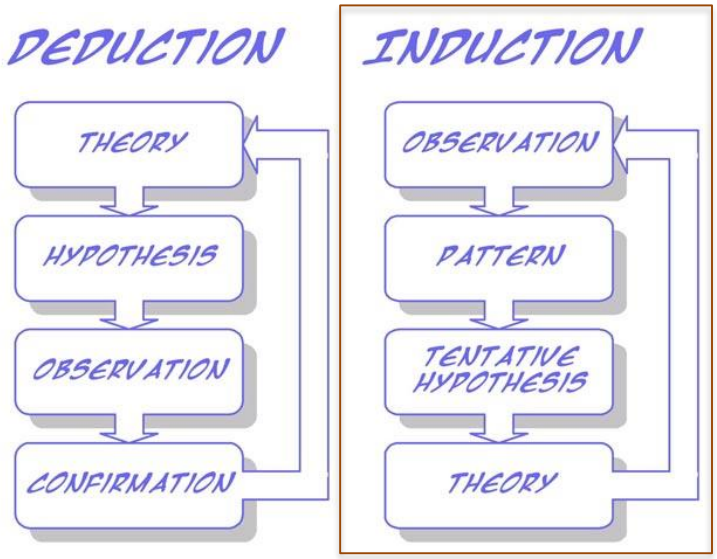
- Algunas aplicaciones comunes: percepción visual, reconocimiento del habla, toma de decisiones, traducción entre lenguajes.

Inteligencia Artificial

- El dominio de problemas en el área de Inteligencia Artificial incluye:
  - Representar conocimiento
  - Razonamiento
  - Planning
- En esta clase me enfocaré en métodos estadísticos usados para aprendizaje, lo que se conoce como **Machine Learning**, y en el procedimiento de uso de estos algoritmos para encontrar patrones en colecciones de datos (**Data Mining**)

Aprendizaje de Máquina (Machine Learning)

- Estudio de algoritmos computacionales que aprenden y mejoran automáticamente a través de la experiencia.
- Algunos investigadores lo llaman también Aprendizaje Estadístico
- Tareas típicas de Machine Learning
  - Descubrimiento de Patrones
  - Clasificación
  - Clustering
  - Regresión
  - Detección de Anomalías/Outliers
  - Reducción de Dimensionalidad



Clasificación Tradicional de algoritmos de ML

- **Aprendizaje Supervisado**: Los algoritmos reciben ejemplos (datos etiquetados) a partir de los cuales aprenden

- **Aprendizaje No Supervisado**: Los algoritmos no reciben ejemplos etiquetados.

Aprendizaje No Supervisado

- K-Means
- Mean Shift
- DBScan
- PCA
- t-SNE

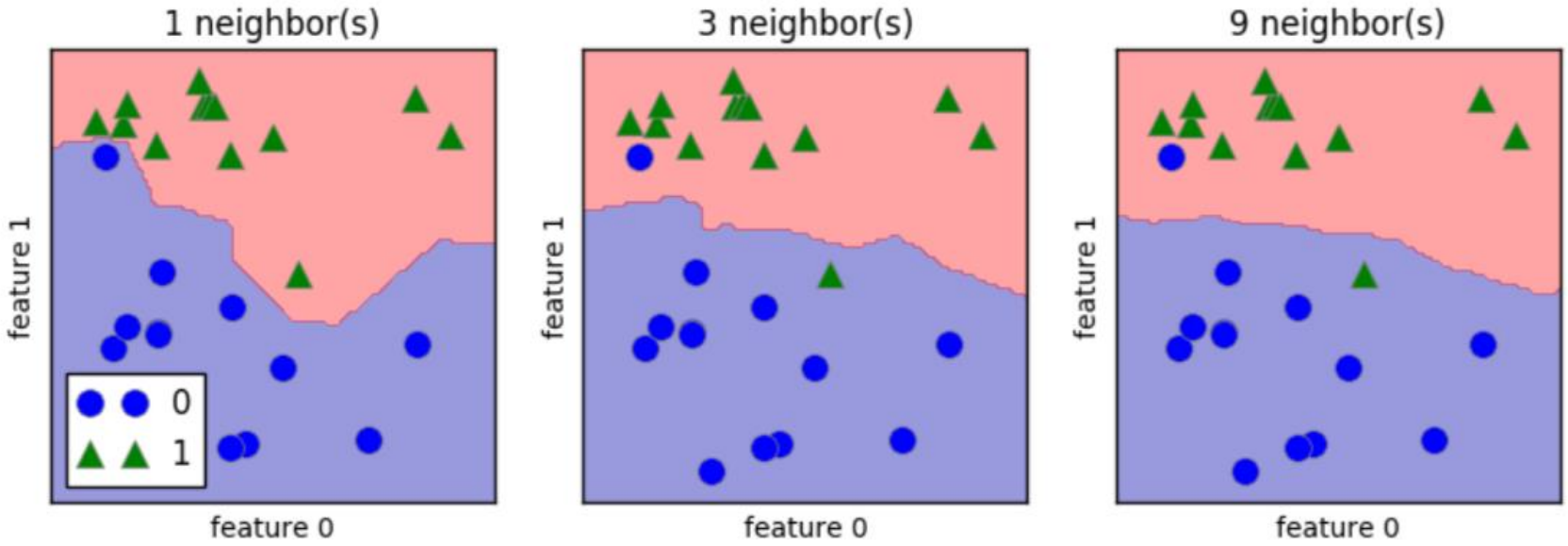
Aprendizaje No Supervisado

- **K-Means**
- **Mean Shift**
- **DBScan**
- PCA
- t-SNE

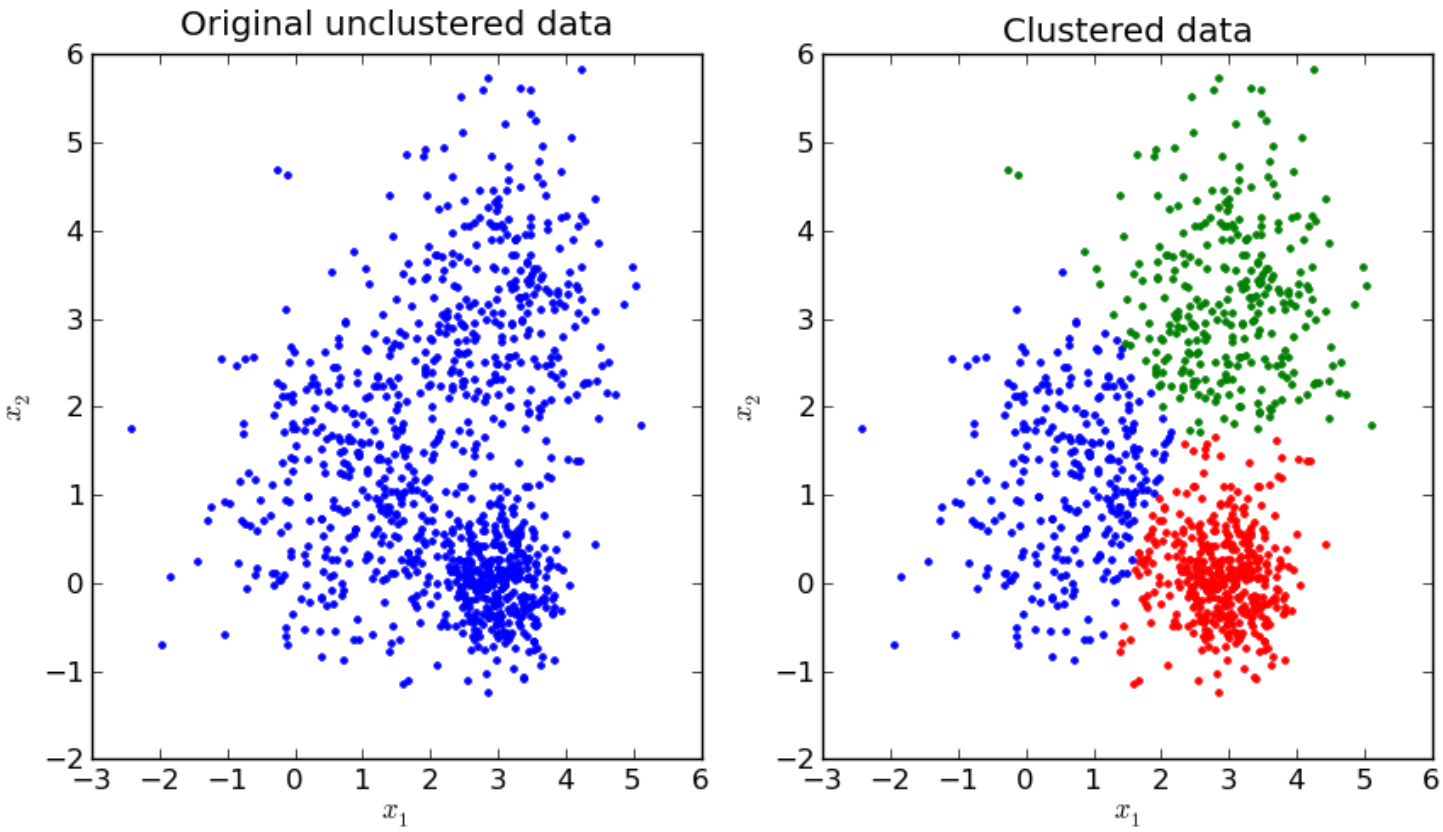
Clustering

- Técnica utilizada para análisis y visualización de datos.
- No necesita labels o clases.
- Permite identificar grupos en los datos, también posibles outliers.

Clasificación

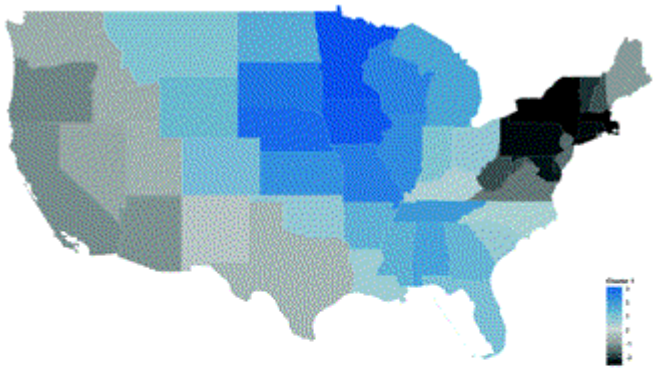


Clustering

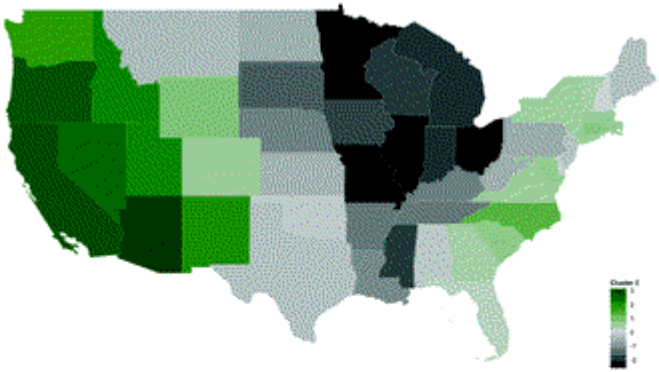


Clustering de Mapas

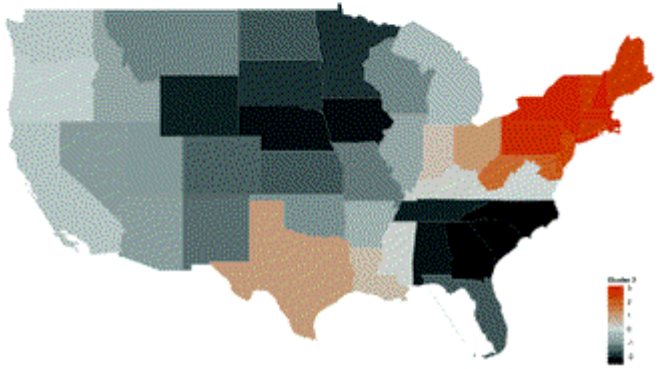
A. Cluster 1: Friendly & Conventional Region



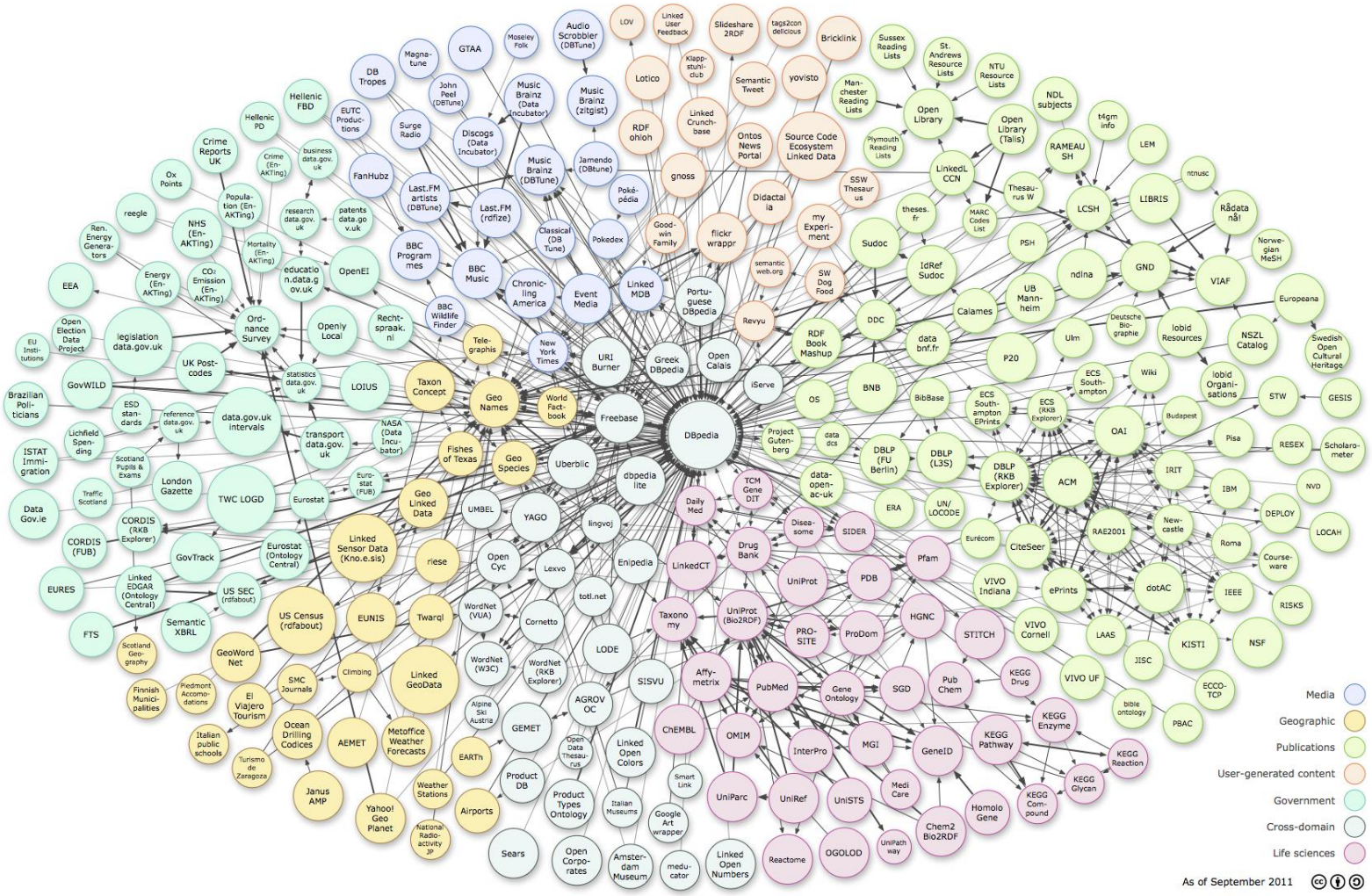
B. Cluster 2: Relaxed & Creative Region



C. Cluster 3: Temperamental & Uninhibited Region



Clustering de Grafos



Clustering de Galaxias

Clustering de Imágenes




Ejemplo Clustering de Imágenes

- https://clippingmagic.com/

K-Means

- Técnica utilizada por décadas al igual que KNN
- Busca K-Medias para lograr generar conjuntos disjuntos en los datos
- Se espera un conocimiento o una intuición previa de cuantos clusters se van a encontrar

K-Means Algoritmo

- Se eligen K centroides iniciales, pueden ser elementos del dataset
- Todos los datos son asignados al centroide más cercano
- Se calcula el nuevo centroide de cada conjunto de datos y se vuelve al paso anterior
- Se itera hasta que ningún punto cambie de cluster de una iteración a la siguiente.

Video K-Means

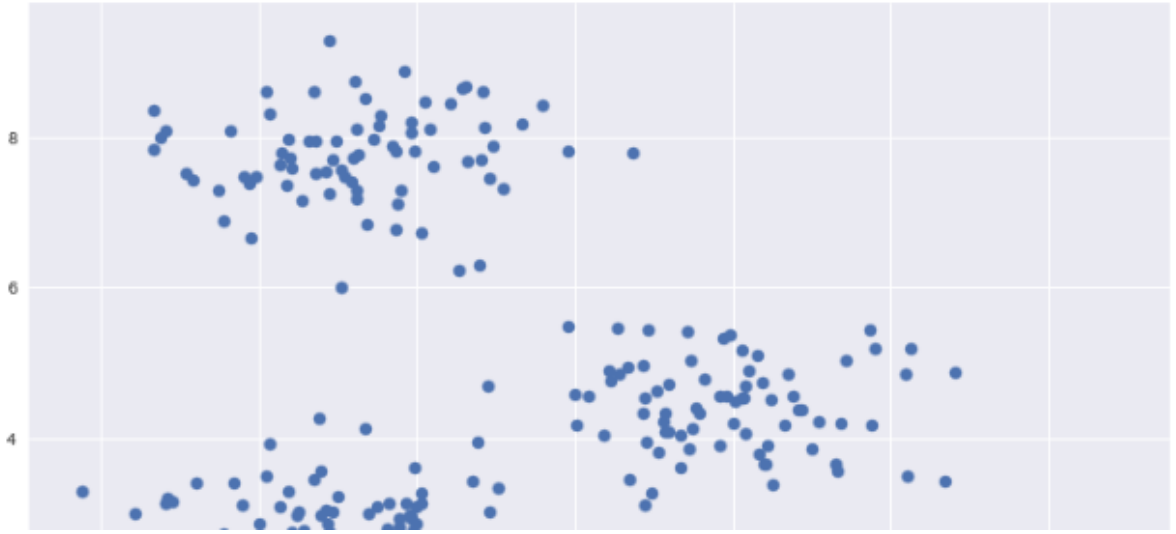
- https://www.youtube.com/watch?v=5I3Ei69I40s

Ejemplo en Vivo

In [8]:
```
%matplotlib inline # Opción para que se vea el output inline

import matplotlib.pyplot as plt
import seaborn as sns; sns.set()  # Estilo del plot
import numpy as np
from pylab import rcParams
rcParams['figure.figsize'] = 12, 9 #Tamaño del plot
```

In [9]:
```
from sklearn.datasets.samples_generator import make_blobs
X, y_true = make_blobs(n_samples=300, centers=4,
                       cluster_std=0.60, random_state=0)
plt.scatter(X[:, 0], X[:, 1], s=50);
```



K-Means

- Los cluster finales son muy dependientes de los puntos iniciales de los centros
- Un centroide puede no ser un punto de la base de datos (como los centroides iniciales)

K-Means

- ¿Siempre es bueno usar K-Means?
- ¿Cómo encuentro clusters que son tendencias en los datos?

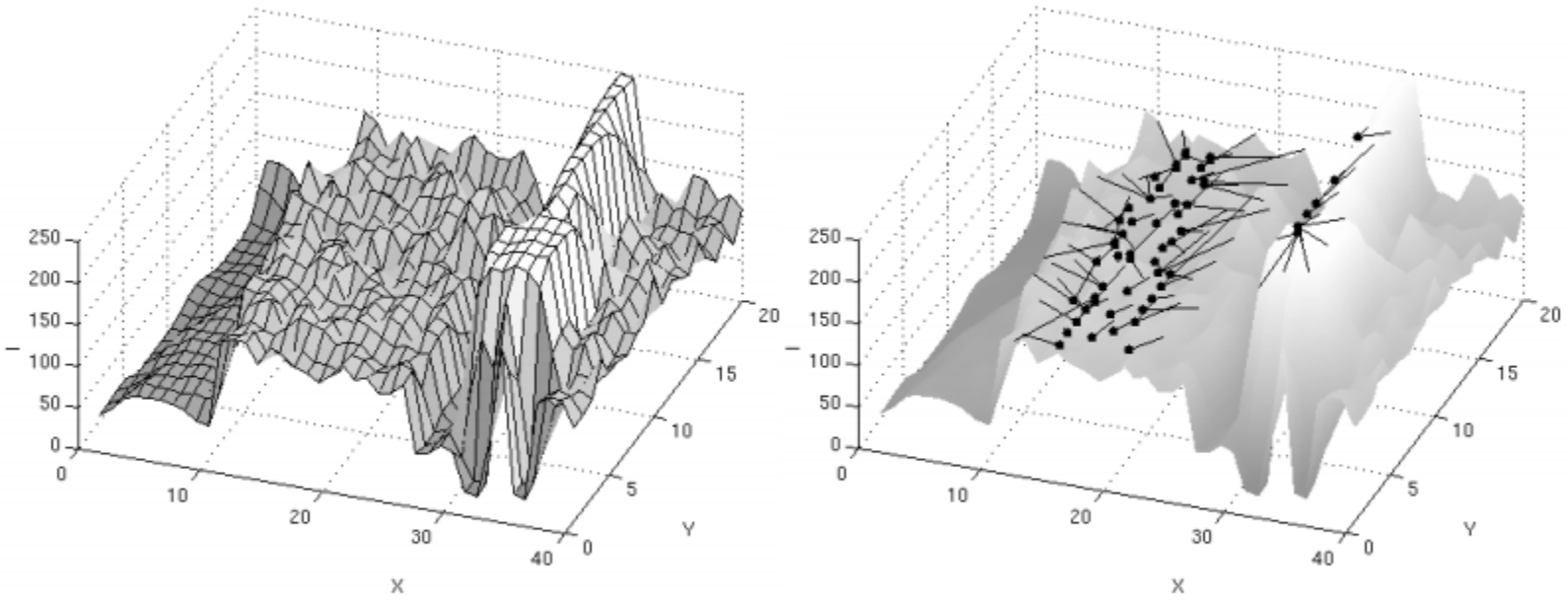
Mean Shift

- Algoritmo que busca aglomeraciones de puntos que siguen una tendencia.
- No es necesario saber a priori la cantidad de clusters a encontrar pero si conocer la distribución de los datos
- Muy sensible a sus parámetros iniciales.

Mean Shift Algoritmo

- Por cada punto, computo su vecindad a una distancia dada.
- Calculo la media de la vecindad.
- Me muevo a esa nueva posición de la media y vuelvo al paso anterior.
- Repetir hasta que converger a un punto.
- Finamente, todos los puntos que llegaron al mismo punto final son un cluster.

Mean Shift



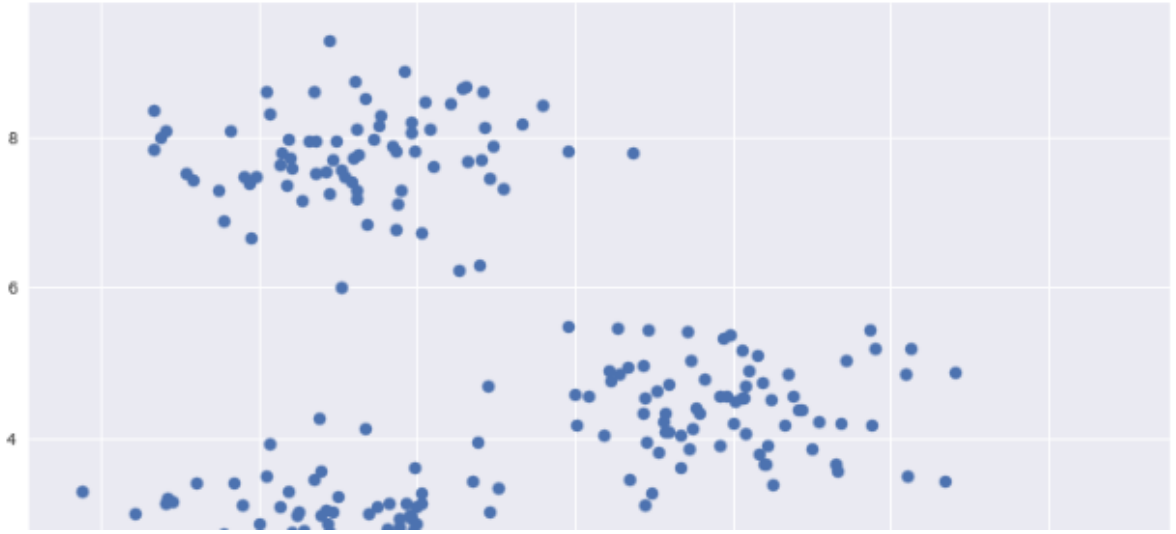
(a) (b)

Volviendo al Ejemplo

In [8]:
```
%matplotlib inline # Opción para que se vea el output inline

import matplotlib.pyplot as plt
import seaborn as sns; sns.set()  # Estilo del plot
import numpy as np
from pylab import rcParams
rcParams['figure.figsize'] = 12, 9 #Tamaño del plot
```

In [9]:
```
from sklearn.datasets.samples_generator import make_blobs
X, y_true = make_blobs(n_samples=300, centers=4,
                       cluster_std=0.60, random_state=0)
plt.scatter(X[:, 0], X[:, 1], s=50);
```



Como evaluar los clusters

- En general se hace una inspección visual de estos.
- También hay métricas como distancias intra o inter clusters, para evaluar que tan bien formados están.

Resumen

- **Aprendizaje Supervisado**: Los algoritmos reciben ejemplos (datos etiquetados) a partir de los cuales aprenden
  - KNN
  - Decision Tree

- **Aprendizaje No Supervisado**: Los algoritmos no reciben ejemplos etiquetados.
  - K-Means
  - Mean Shift

Cómo aprender más

- Hay muchos cursos en el departamento de Machine Learning
  - Inteligencia Artificial
  - Reconocimiento de Patrones
  - Minería de Datos
  - Sistemas Recomendadores
  - Deep Learning
- Hay mucho material en Internet
  - Cursos en Coursera
  - Libros
  - Tutoriales

Gracias!