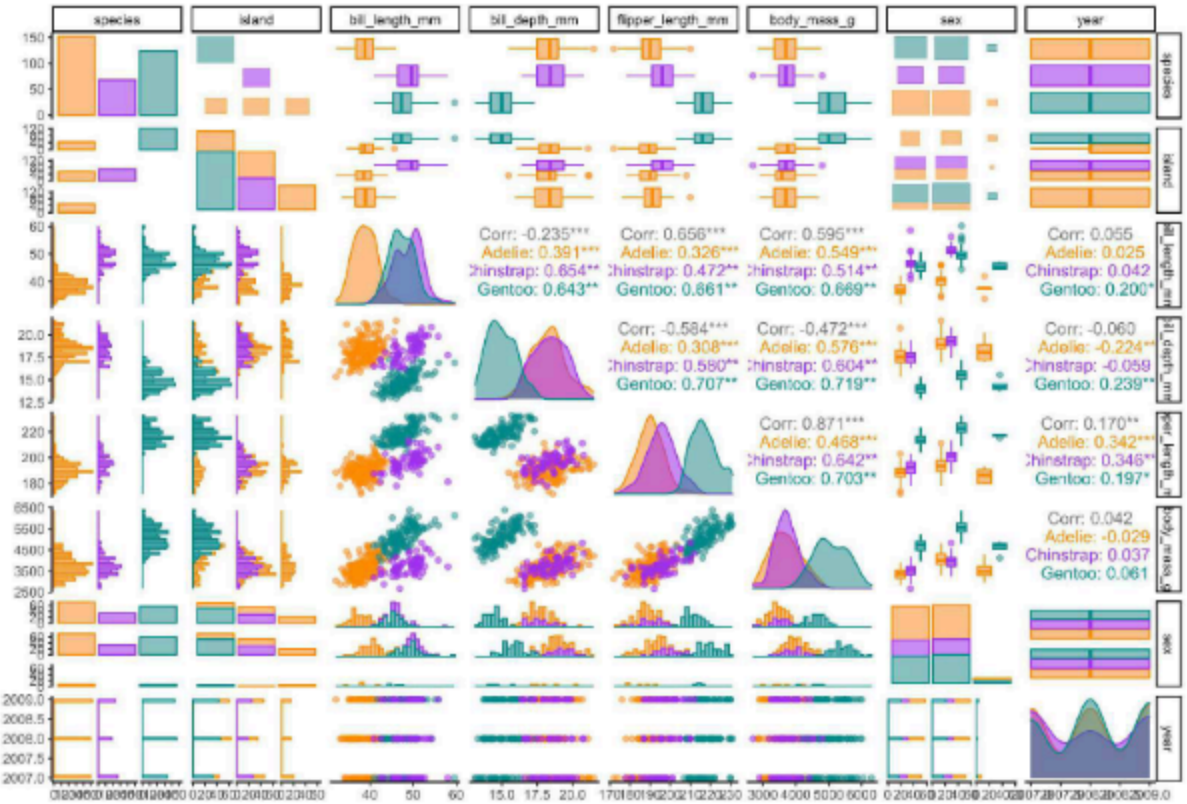
¿Qué hace cuando tengo muchas variables?

En la clase anterior vimos nos dimos cuenta de como una variable podría causar la diferencia entre saber si mi trabajo valió o no valió la pena.

Entonces ¿que pasa cuando tengo muchas variables?, lo que deberíamos hacer es hacer el análisis exploratorio considerando esta variable categórica o esta nueva variable numérica y ver como se están relacionando con otras variables.

En el ejemplo anterior hicimos una exploración con la función `pairplot()` diciéndole que queríamos graficar cada par de variables considerando un parámetro de segmentación. así de esta manera fuimos capaces de encontrar patrones interesantes, por ejemplo que en algunas ocasiones las especies coincidían en cuanto a su comportamiento gráfico y en otras tenían similitud en cuanto a una linealidad pero desplazada.

Así que recordemos cuando hicimos las regresiones tanto lineal como logística, el hecho de no contemplar la variable `specie` hacía que la información no fuera clara y que a la vez existiera un ruido que provocaba comportamientos extraños en los datos.

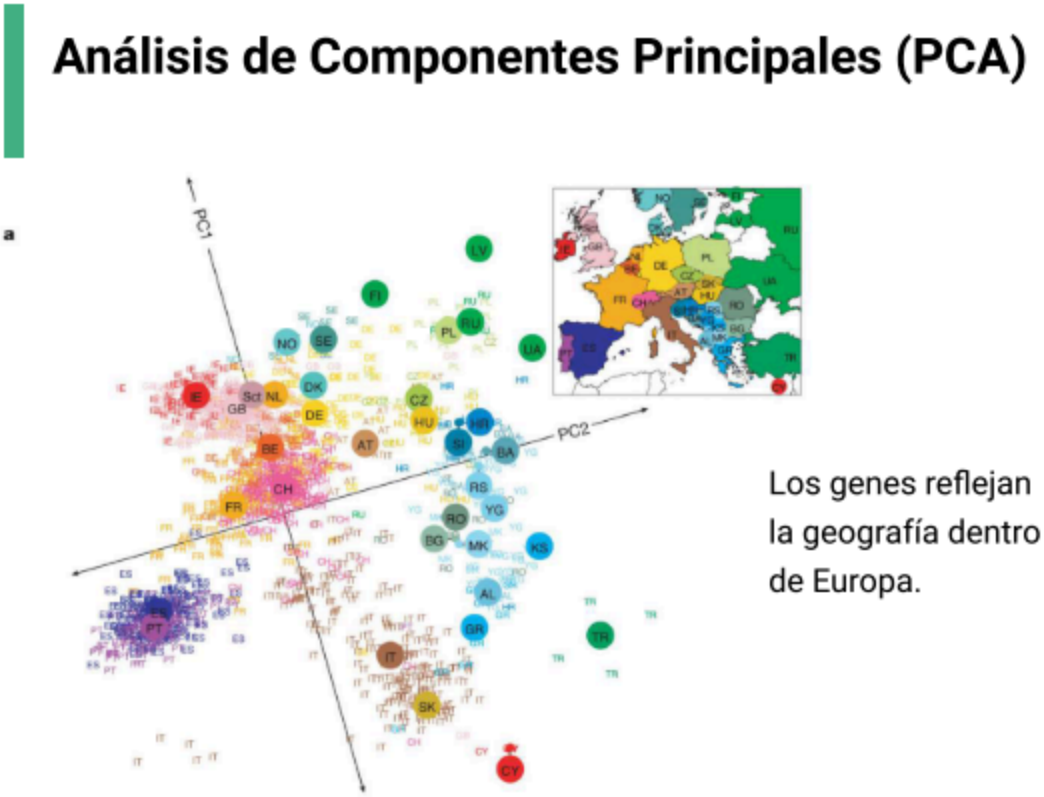


También es cierto que cuando tenemos miles de variables, este proceso se vuelve más complejo y complicado que a veces pudiera tornarse confuso.

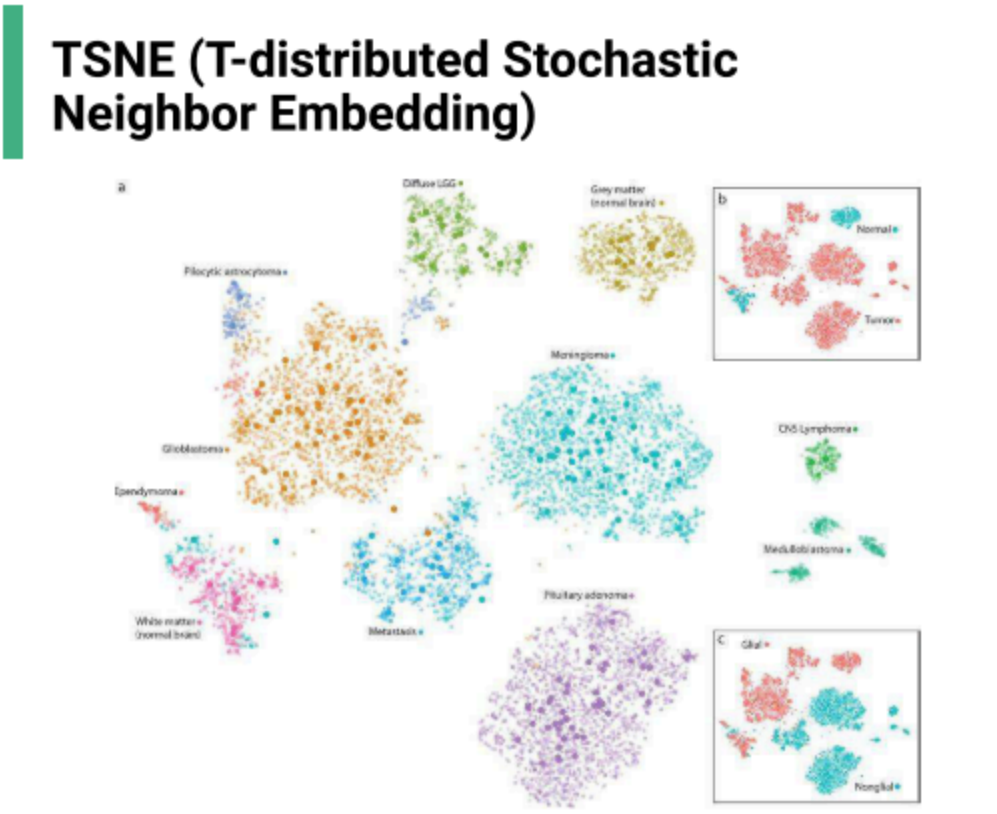
Existen técnicas especializadas y algoritmos que me van a ayudar a entender la variación de los datos de manera sencilla, reduciendo las dimensiones y colapsando el espacio a un numero menor de variables. Por ejemplo un espacio de 10 dimensiones podría reducirse a uno de 2 dimensiones, de manera sencilla.

Métodos

- Tenemos uno llamado **PCA (Análisis de Componenetes Principales)**, aquí un ejemplo: Los genes reflejan la geografía dentro de Europa.

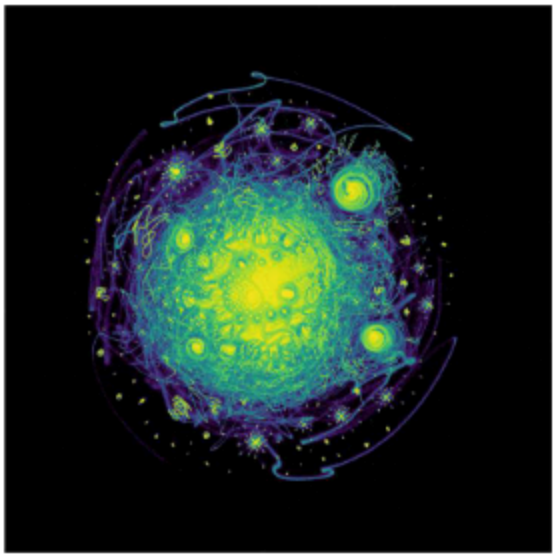


- Otro método es el **TSNE (T-distributed Stochastic Neighbor Embedding)**



- Tenemos otro llamado UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

> **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**



Visualización de 30,000,000 de números enteros representados por vectores binarios divisibilidad prima, coloreados por densidad de puntos.

UMAP

Es otro tipo de reducción de dimensiones que no es muy escuchado pero que realmente su potencial va más allá de un **PCA** y un **TSNE**.
Este algoritmo de reducción de dimensiones, trata de reducir las dimensiones, tratar de capturar la estructura global y preservar parte de la estructura local de los datos.

Por ejemplo si tenemos 10,000 variables y pudiéramos visualizarlas en un plano o pudiéramos visualizar más de 5 dimensiones a la vez, se podría hacer una proyección para visualizarlas.

Es decir como tener un mapa, nosotros sabemos que la tierra no es plana, pero la podemos representar en forma de mapa, básicamente es lo que hace **UMAP**, esta idea de que tenemos una esfera pero la podemos representar como si fuera un plano

Otra forma de comparar el algoritmo, es con los data set que utilizan en Machine Learning.

Cada una de estas herramientas nos provee la capacidad de resumir insights visuales y extraer el comportamiento o patrones que tienen nuestros datos al reducirlos a un espacio tan pequeño

Resumen

En esta clase dimos una previsualización de que existen diferentes algoritmos que nos van a ayudar a reducir las dimensiones de nuestro set de datos y extraer la información más relevante de el. Así que cuando necesitemos trabajar con un conjunto de variables grande, esto nos puede ayudar a trabajar de una forma más sencilla, para que podamos explorar sus estructuras globales o locales.

- Inteligencia Artificial contra el cancer
- How to use MAP
- 13 técnicas de reducción de dimensionalidad