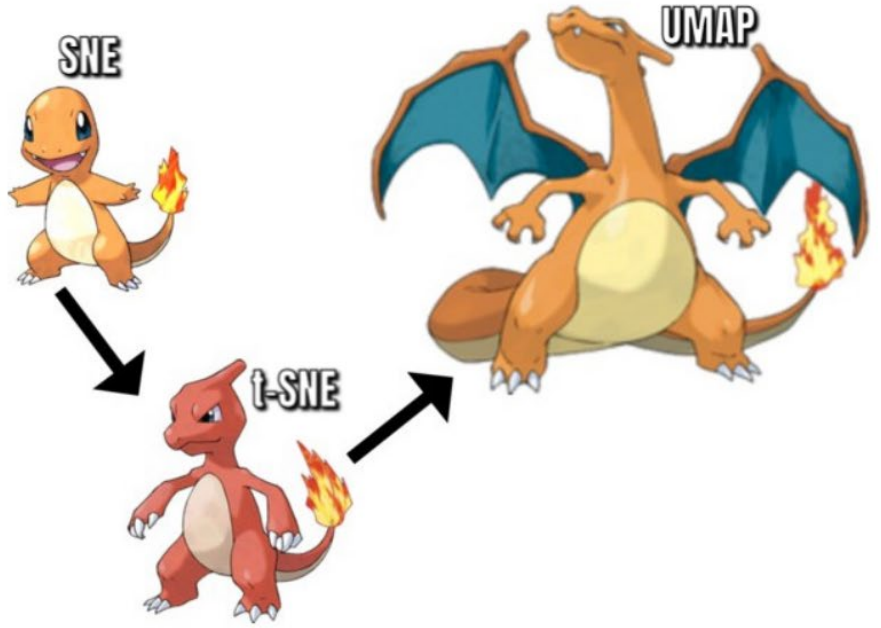
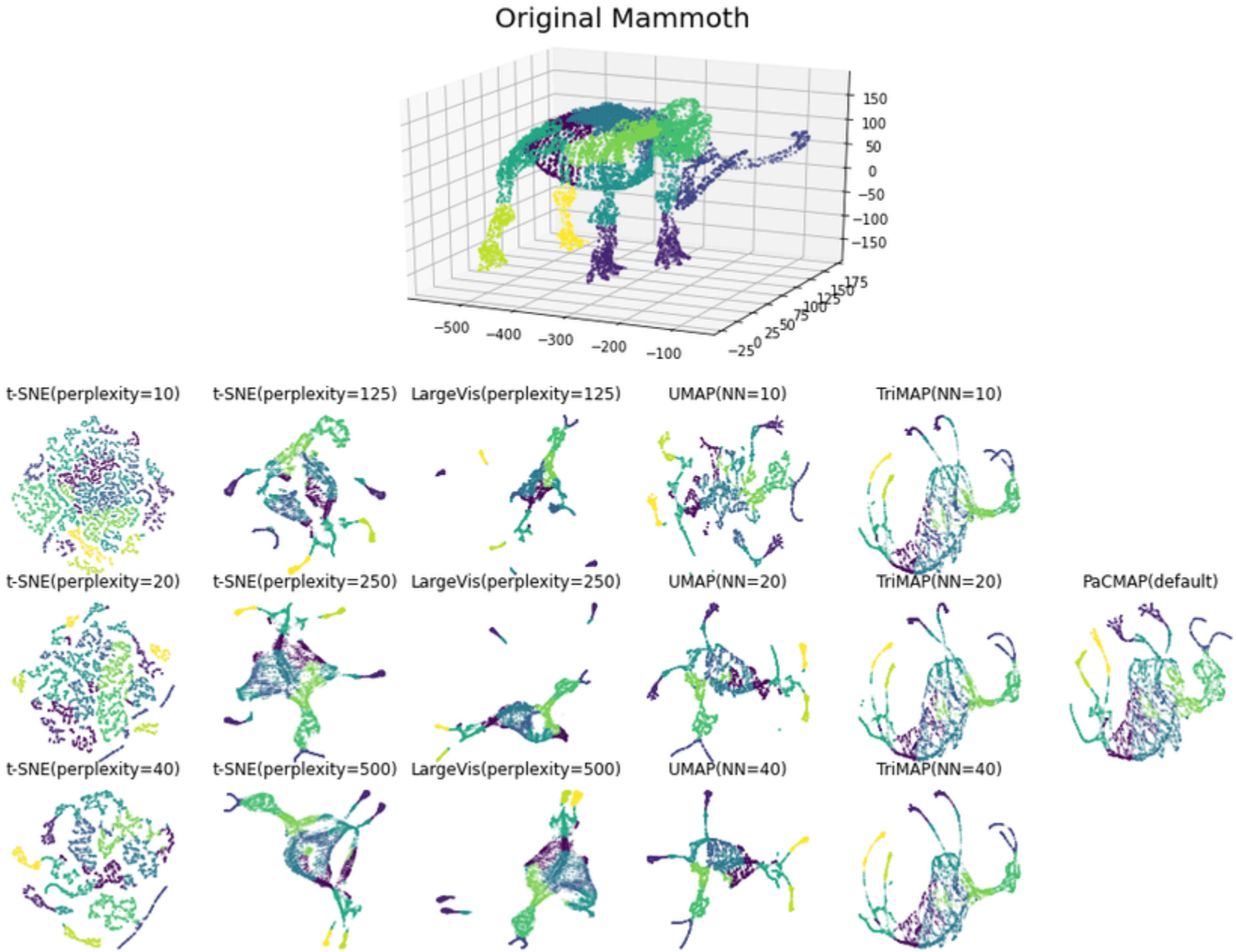


Dimension Reduction: PCA, tSNE, UMAP

PhD(e). Jonnatan Arias Garcia – jonnatan.arias@utp.edu.co – jariasg@uniquindio.edu.co

PhD. David Cardenas peña - dcardenasp@utp.edu.co

PhD. Hernán Felipe Garcia - hernanf.garcia@udea.edu.co


Original Mammoth
150
100
50
0
−50
−100
−150
−500
−400
−300
−200
−100
−25
0
25
50
75
100
125
150
175
t-SNE(perplexity=10)
t-SNE(perplexity=125)
LargeVis(perplexity=125)
UMAP(NN=10)
TriMAP(NN=10)
t-SNE(perplexity=20)
t-SNE(perplexity=250)
LargeVis(perplexity=250)
UMAP(NN=20)
TriMAP(NN=20)
PaCMAP(default)
t-SNE(perplexity=40)
t-SNE(perplexity=500)
LargeVis(perplexity=500)
UMAP(NN=40)
TriMAP(NN=40)

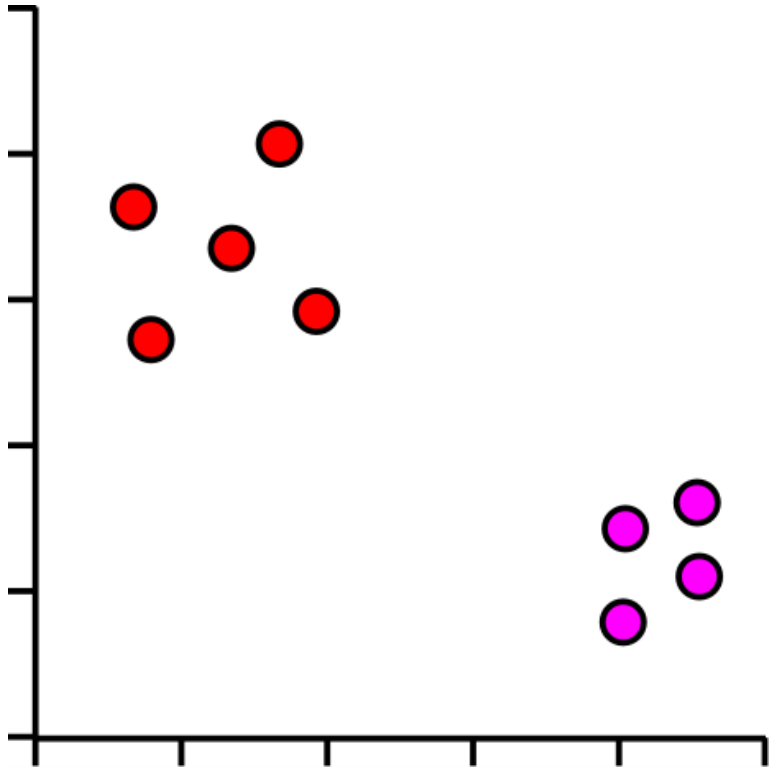
¿Pa dónde vamos?

Gene	Description	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5
Inpp5d	inositol polyphosphate-5-phosphatase D	7.00	5.45	5.89	6.03	5.75
Aim2	absent in melanoma 2	3.01	4.37	4.59	4.38	4.18
Gldn	gliomedin	3.48	3.63	4.61	4.70	4.74
Frem2	Fras1 related extracellular matrix protein 2	4.75	4.66	3.46	3.74	3.45
Rps3a1	ribosomal protein S3A1	6.10	7.23	7.44	7.36	7.34
Slc38a3	solute carrier family 38, member 3	1.90	3.16	3.52	3.61	3.19
Mt1	metallothionein 1	5.07	6.49	6.46	6.04	6.05
C1s1	complement component 1, s subcomponent 1	2.74	3.02	3.86	4.10	4.10
Cds1	CDP-diacylglycerol synthase 1	4.55	4.22	3.80	3.16	3.12
Ifi44	interferon-induced protein 44	4.82	4.52	3.87	3.42	3.59
Lefty2	left-right determination factor 2	6.95	6.28	5.88	5.60	5.61
Fmr1nb	fragile X mental retardation 1 neighbor	4.28	2.78	3.10	3.25	2.57
Tagln	transgelin	7.93	7.91	7.20	7.02	6.68

Cada punto es una celda

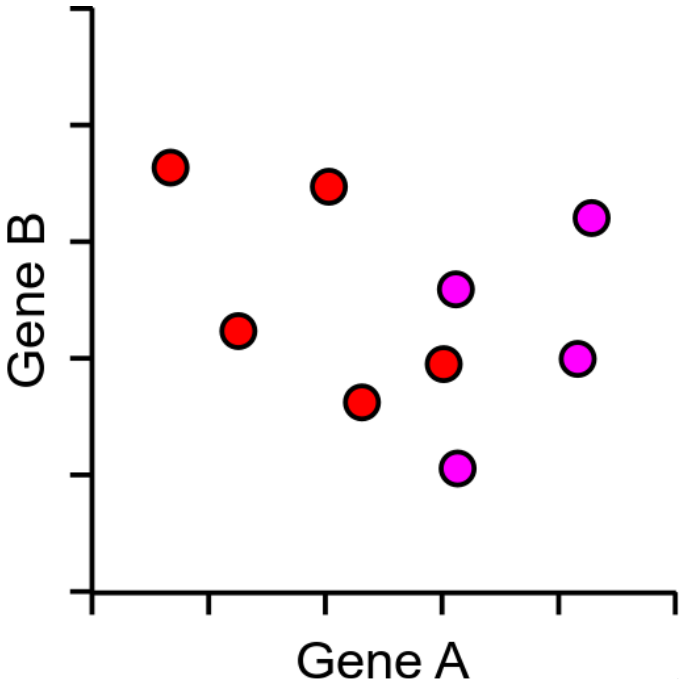
Los grupos de puntos son celdas similares

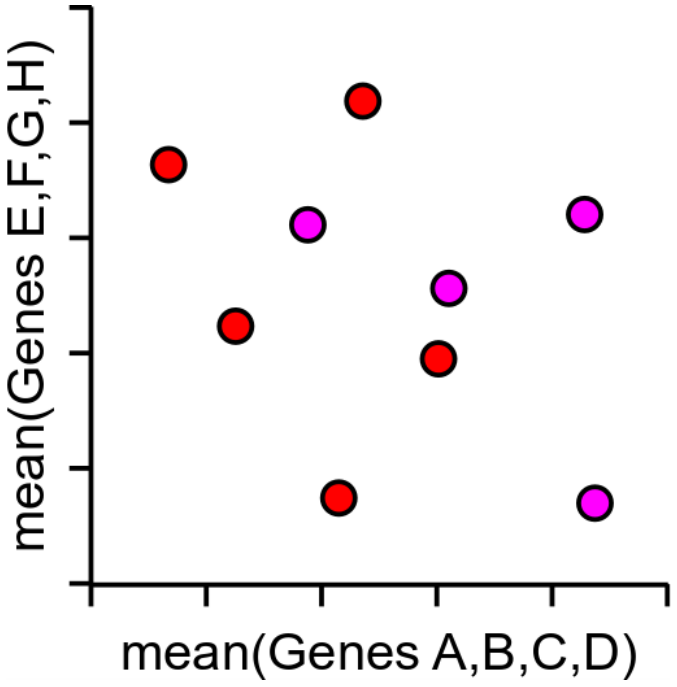
La separación de grupos podría ser interesante desde el punto de vista biológico

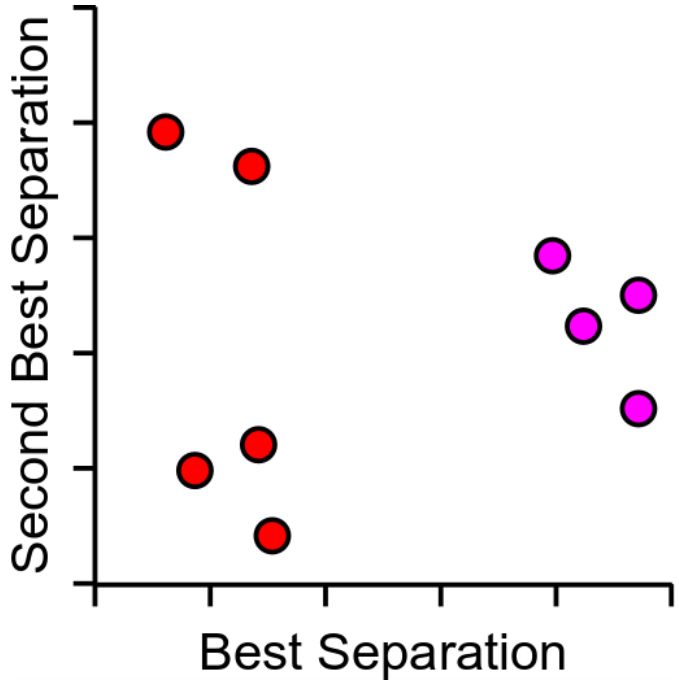


¡Demasiados datos!

- 5000 células y 2500 genes medidos
- Siendo realistas, solo podemos trazar 2 dimensiones (x,y)







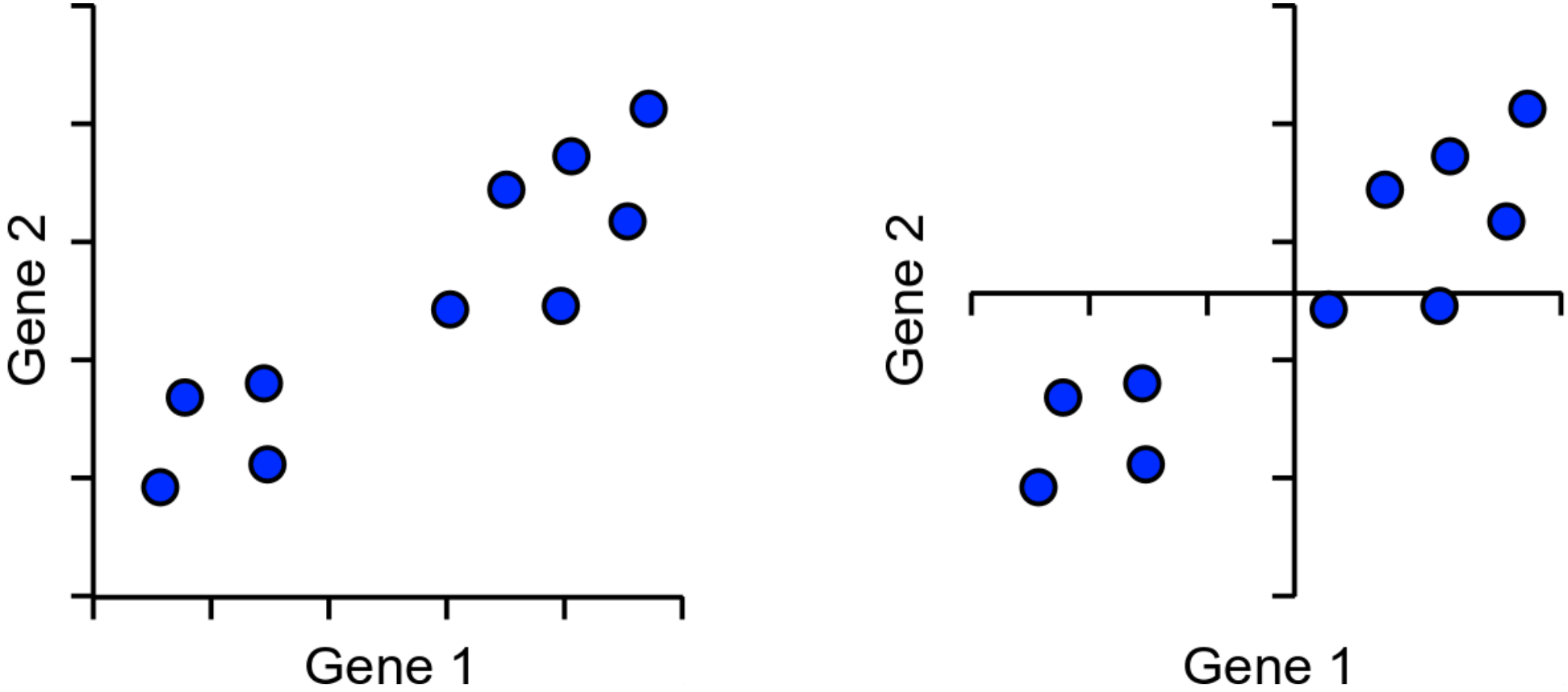
Análisis de los componentes principales (PCA)

- Método para resumir de forma óptima grandes conjuntos de datos multidimensionales
- Puede encontrar un número menor de dimensiones (idealmente 2) que retengan la mayor parte de la información útil en los datos.
- Crea una receta para convertir grandes cantidades de datos en un solo valor, llamado Componente Principal (PC), por ejemplo:

PC = (GeneA*10)+(GeneB*3)+(GeneC*-4)+(GeneD*-20)…

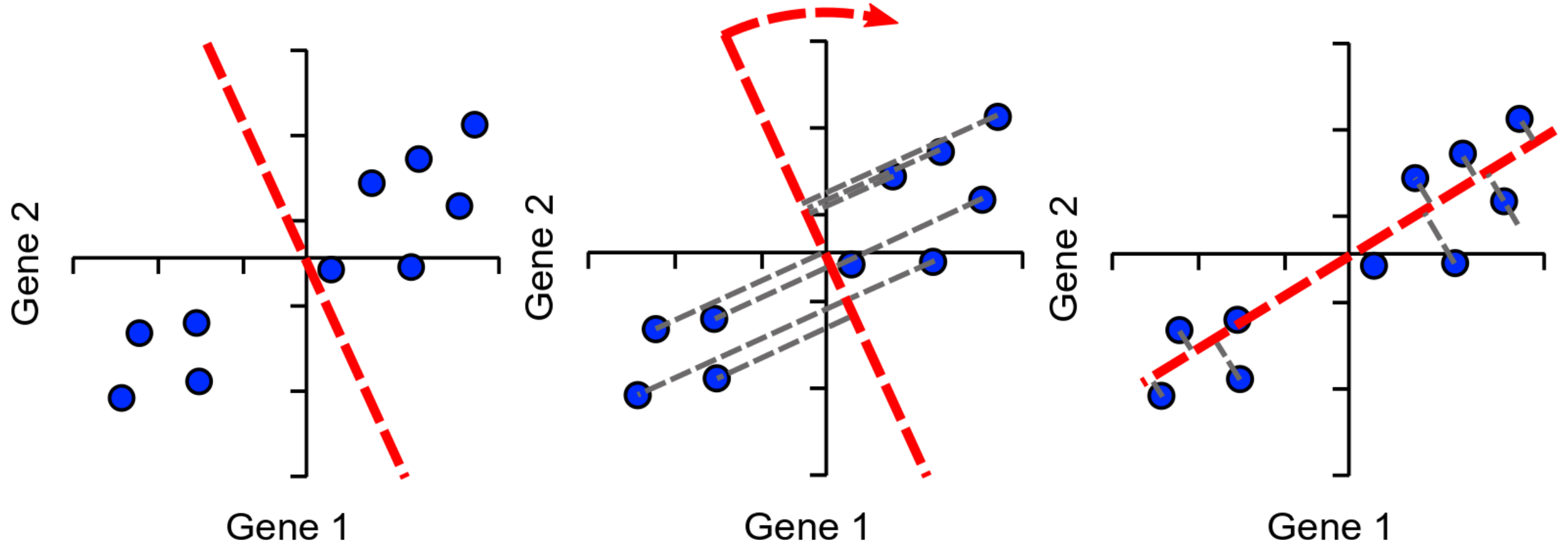
¿Cómo funciona la PCA?

Ejemplo simple usando 2 genes y 10 células

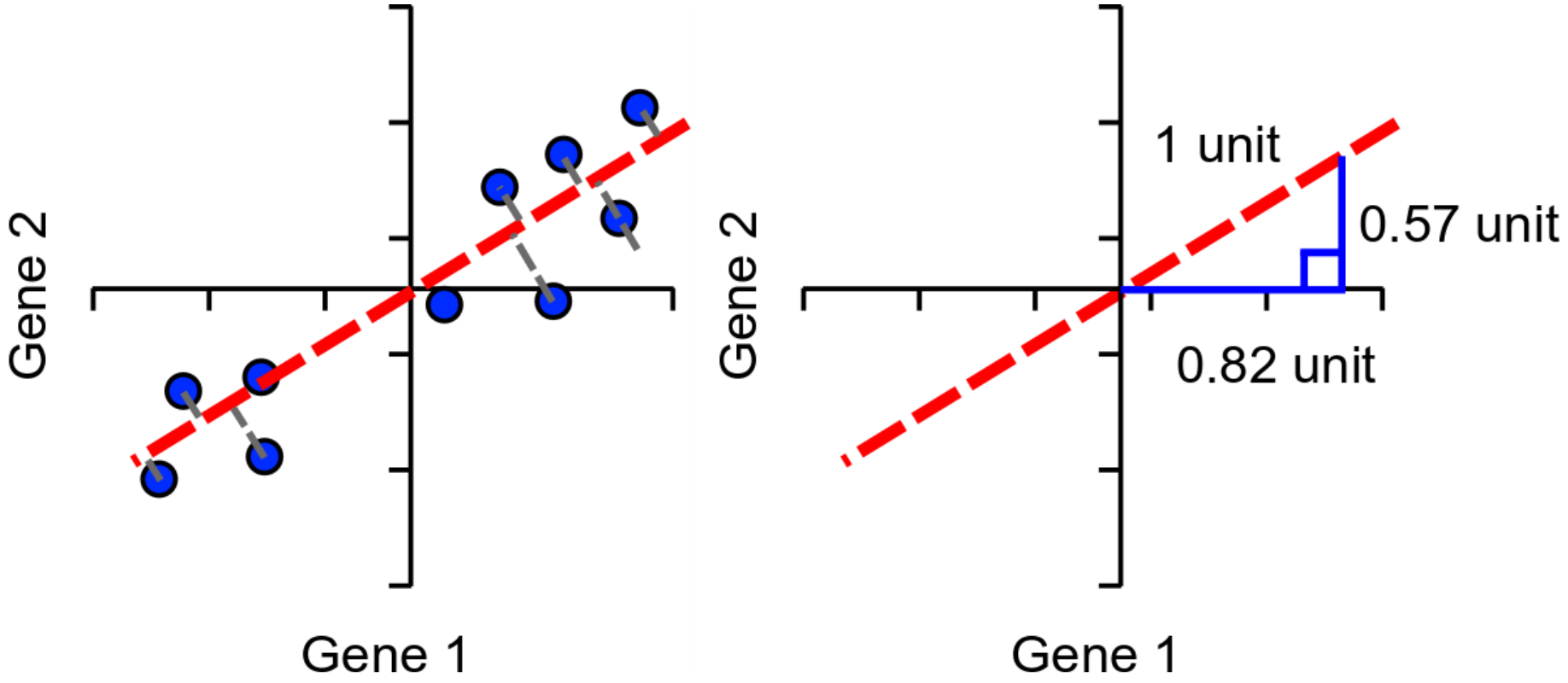


¿Cómo funciona la PCA?

Encuentre la línea de mejor ajuste, pasando por el origen



Asignación de cargas a genes



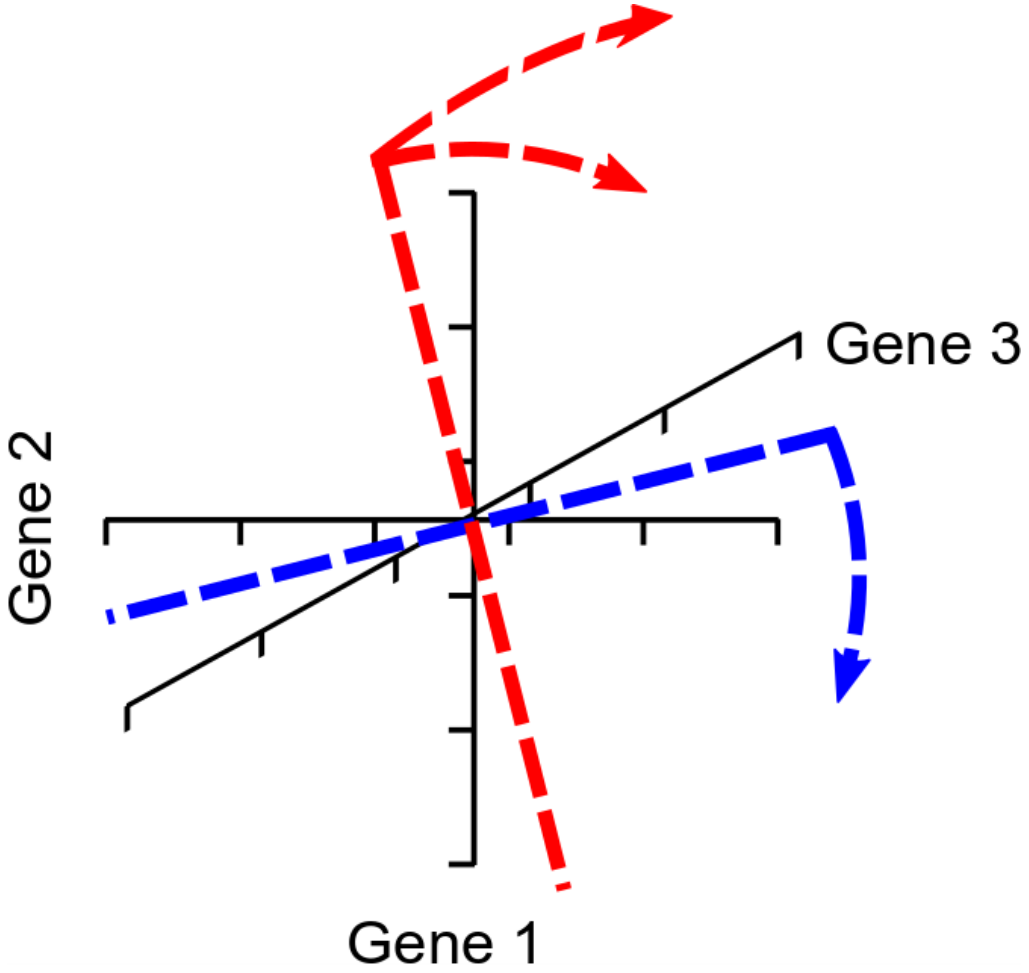
Vector único o '**eigenvector**'

Cargas:

- Gene1 = 0.82
- Gene2 = 0.57

Una mayor carga equivale a una mayor influencia en el PC

Más dimensiones

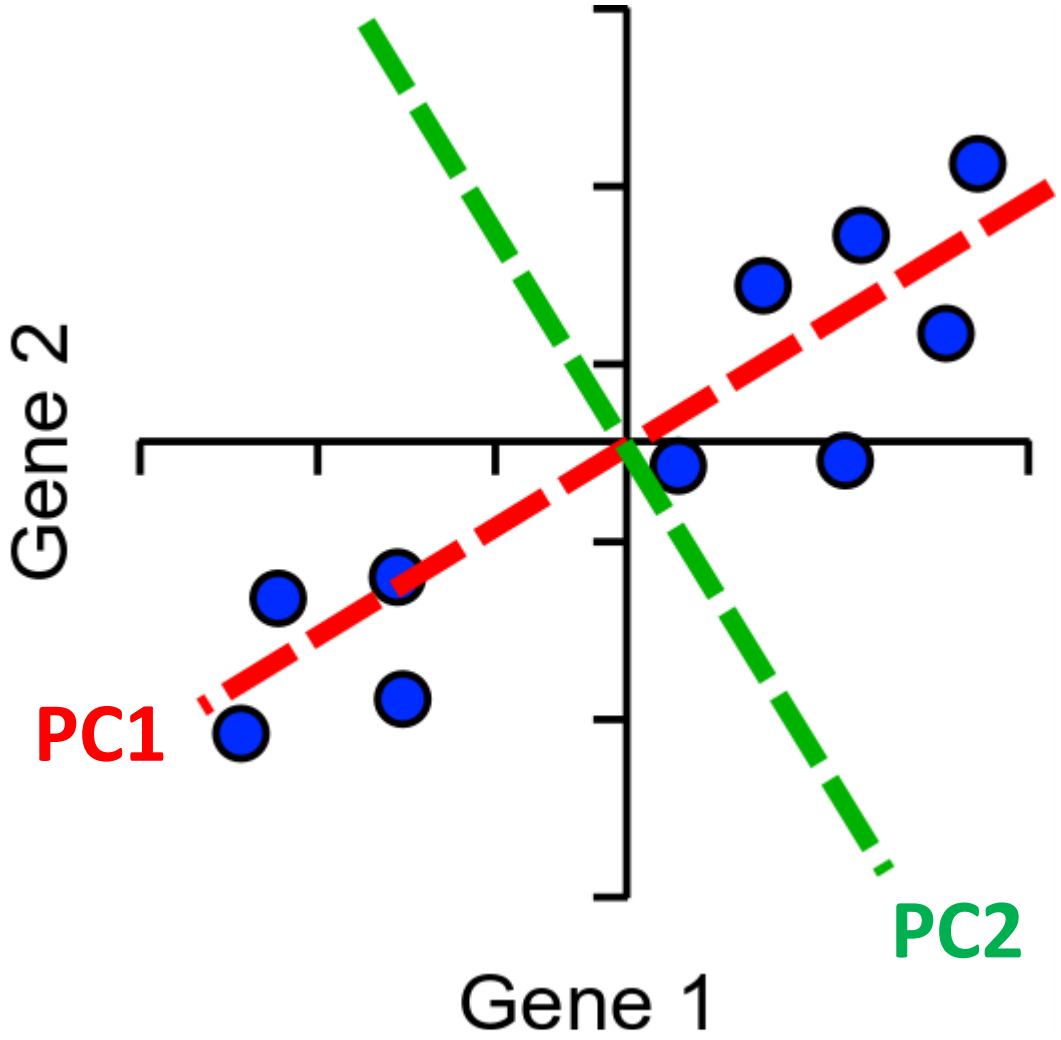


- La misma idea se extiende a un mayor número de dimensiones (n)
- El primer PC gira en dimensiones (n-1)
  - El siguiente PC es perpendicular al PC2, pero girado de manera similar (n-2)
  - El último PC permanece perpendicular (no hay opción)
  - El mismo número de PC que de genes

Explicación de la varianza

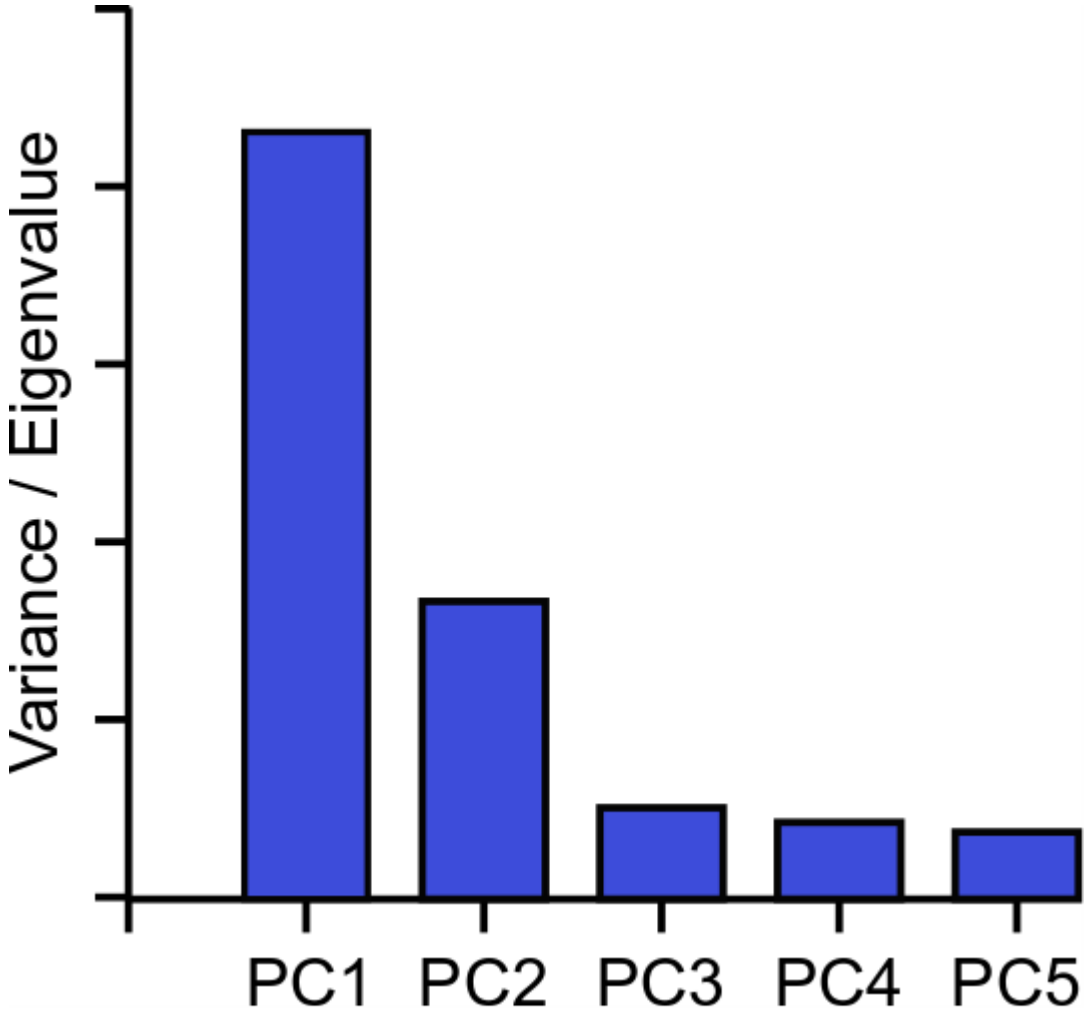
- Cada PC siempre explica una proporción de la varianza total de los datos. Entre los dos se explican todo
  - PC1 siempre explica más
  - PC2 es el siguiente más alto, etc., etc.
- Dado que solo trazamos 2 dimensiones, nos gustaría saber que estas son una buena explicación
- ¿Cómo calculamos esto?

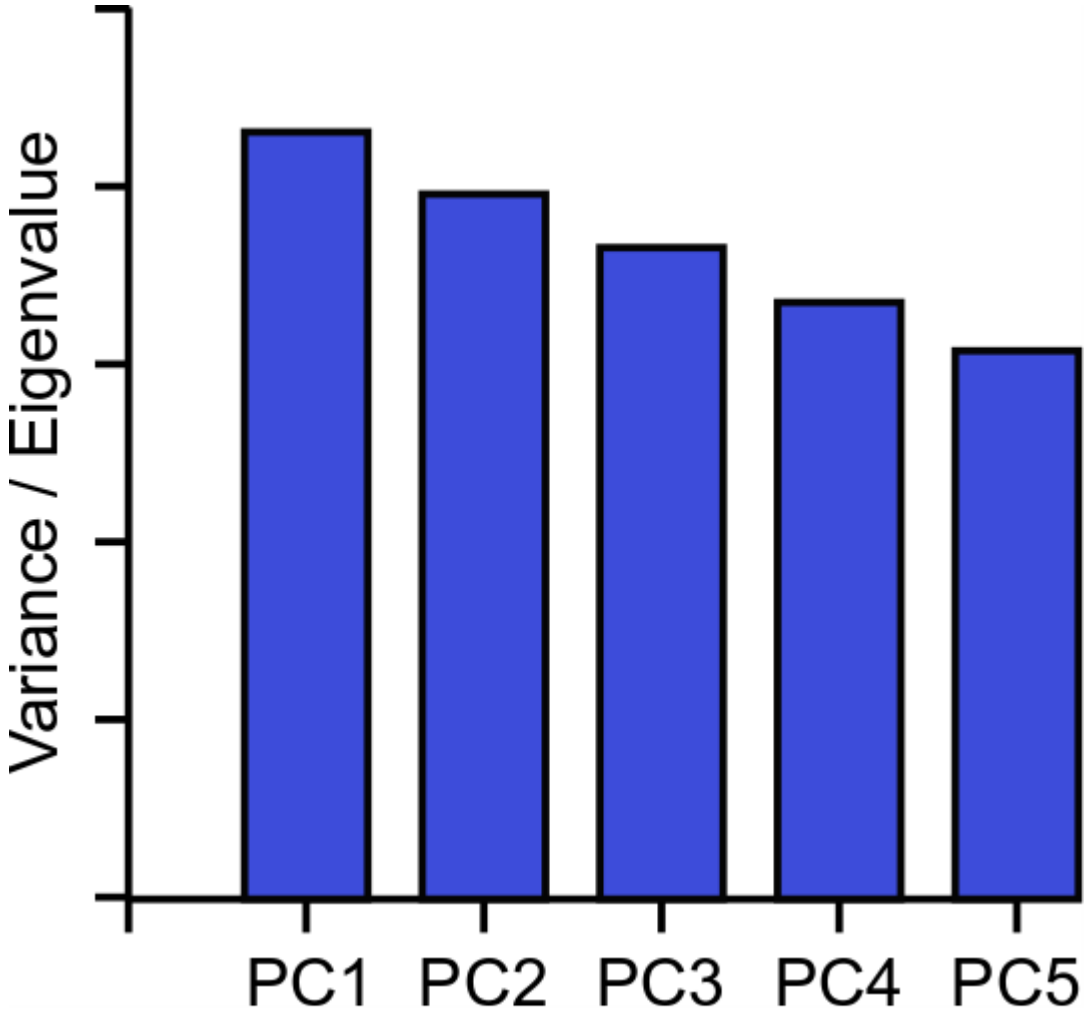
Explicación de la varianza



- Proyectar en PC
- Calcular la distancia al origen
- Calcular la suma de las diferencias al cuadrado(SSD)
  - Esta es una medida de varianza llamada‘eigenvalue’
  - Divida por (puntos-1) para obtener la varianza real

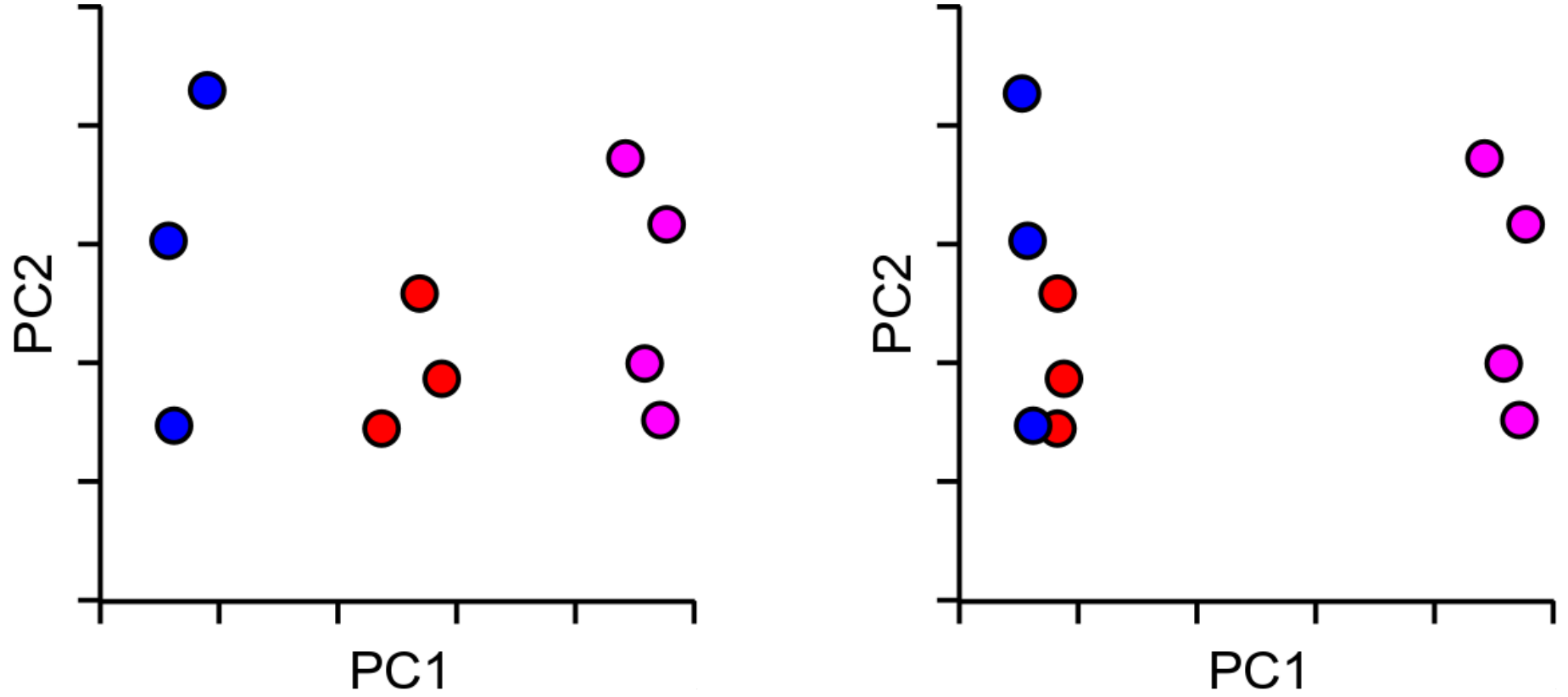
Explicación de la varianza– Scree Plots





Entonces, ¿PCA es genial?

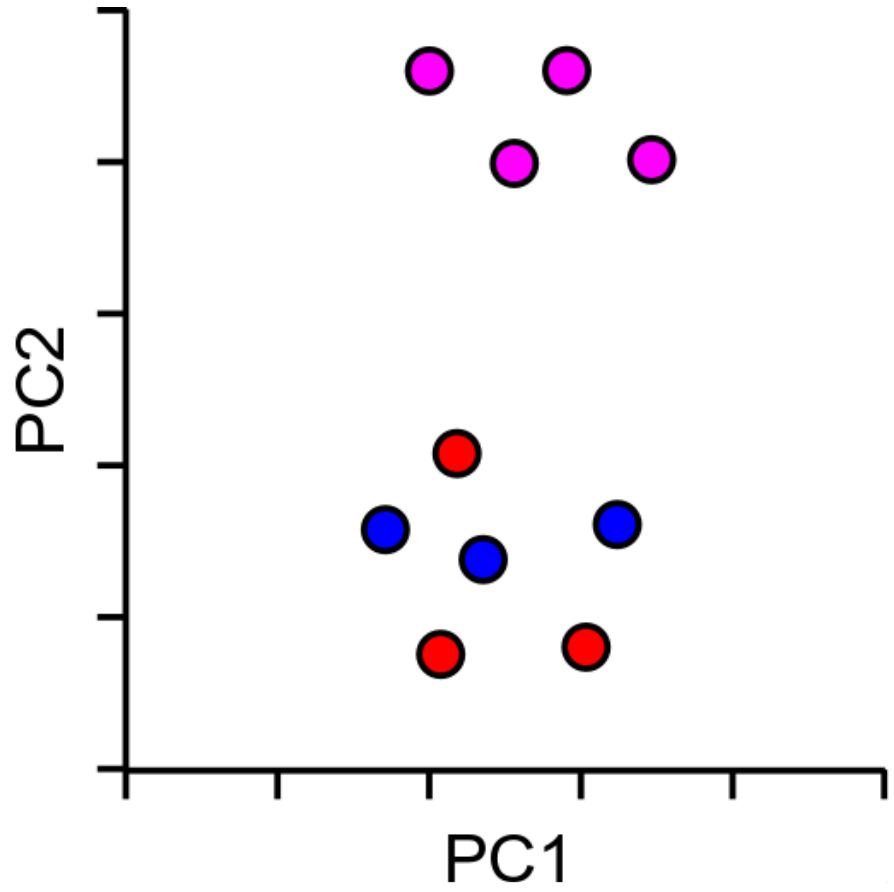
Algo así…

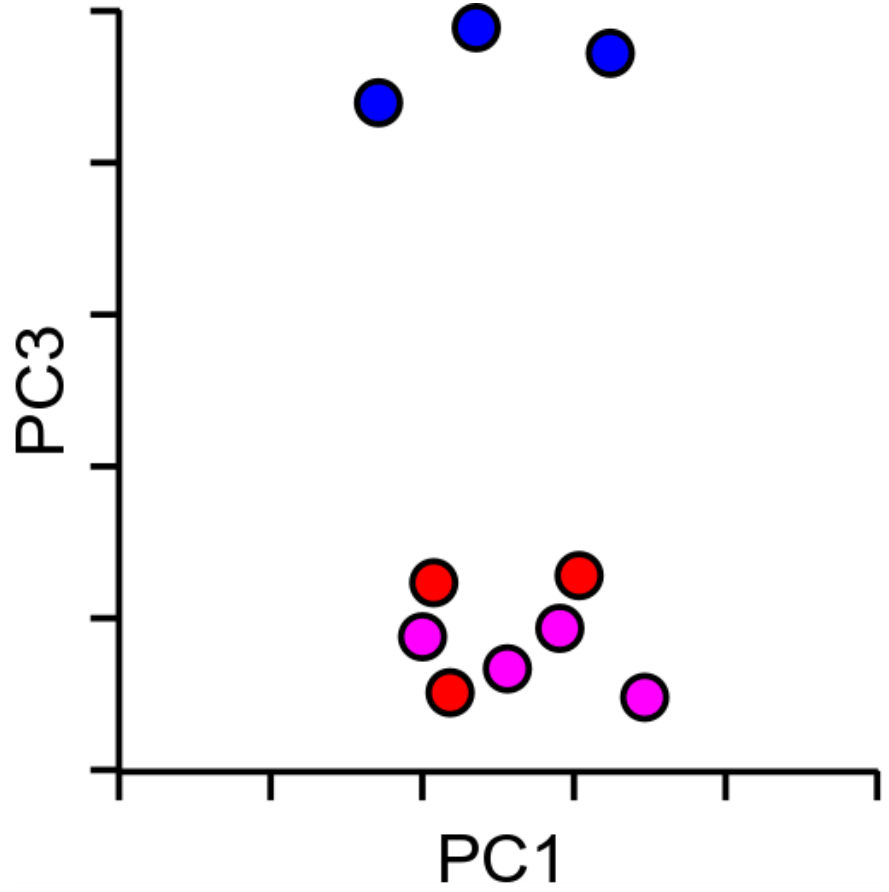


Separación no lineal de valores

Entonces, ¿PCA es genial?

Algo así…





No optimizado para 2 dimensiones

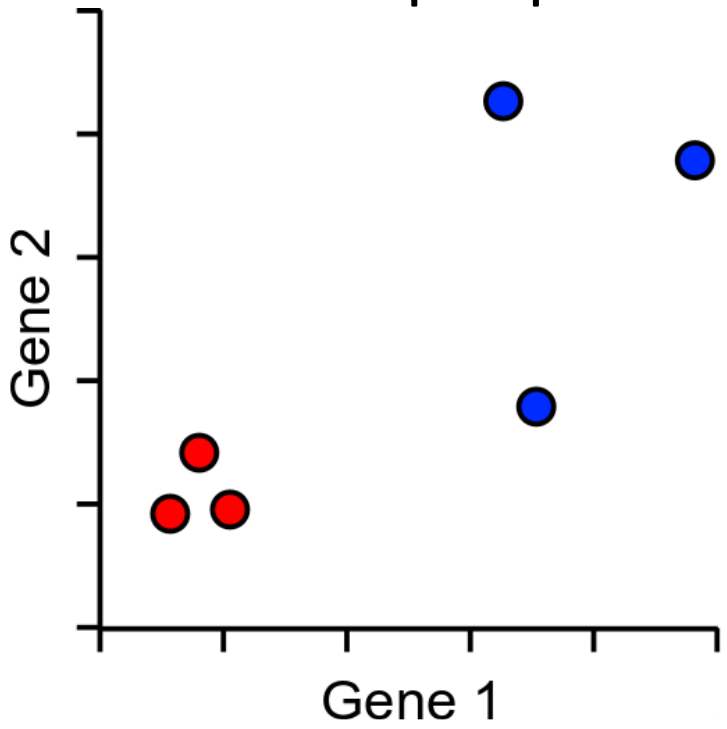
tSNE al rescate...

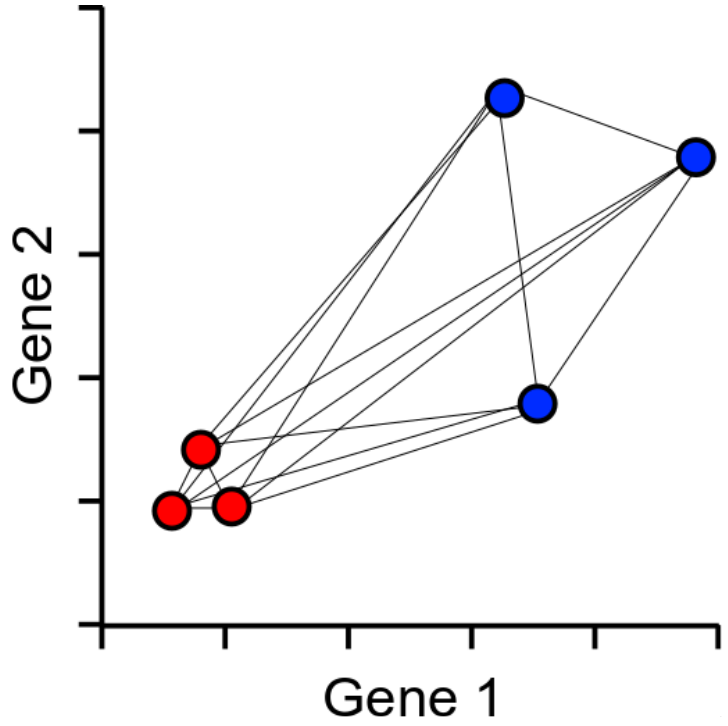
- Embedding de vecino estocástico distribuido en T

- Tiene como objetivo resolver los problemas de PCA
  - Escalado no lineal para representar cambios en diferentes niveles
  - Separación óptima en 2 dimensiones

¿Cómo funciona tSNE?

- Basado en la tabla de todos contra todos de distancias de celda a celda por pares

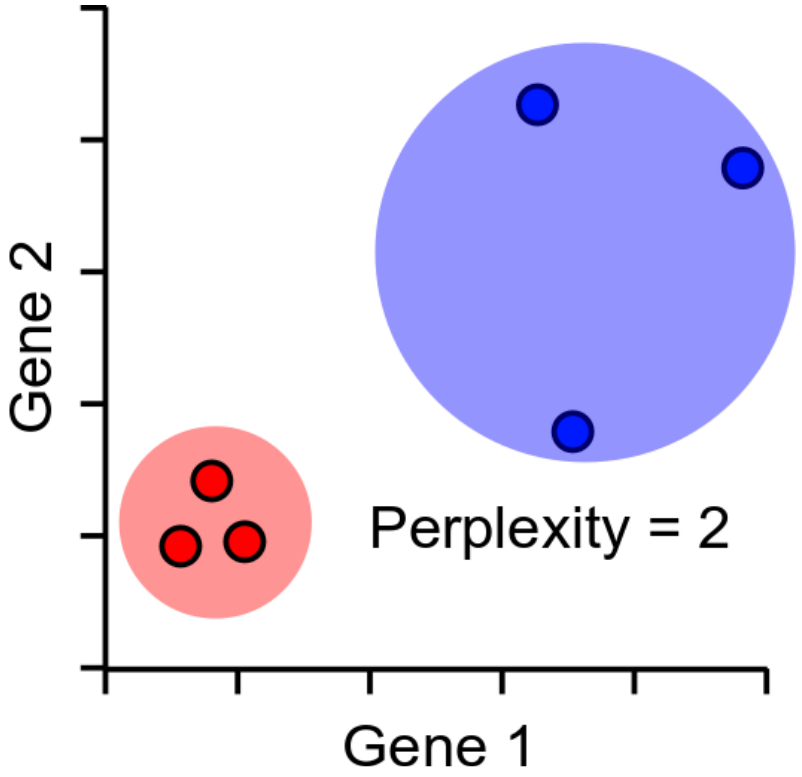




	0	10	10	295	158	153
	9	0	1	217	227	213
	1	8	0	154	225	238
	205	189	260	0	23	45
	248	227	246	44	0	54
	233	176	184	41	36	0

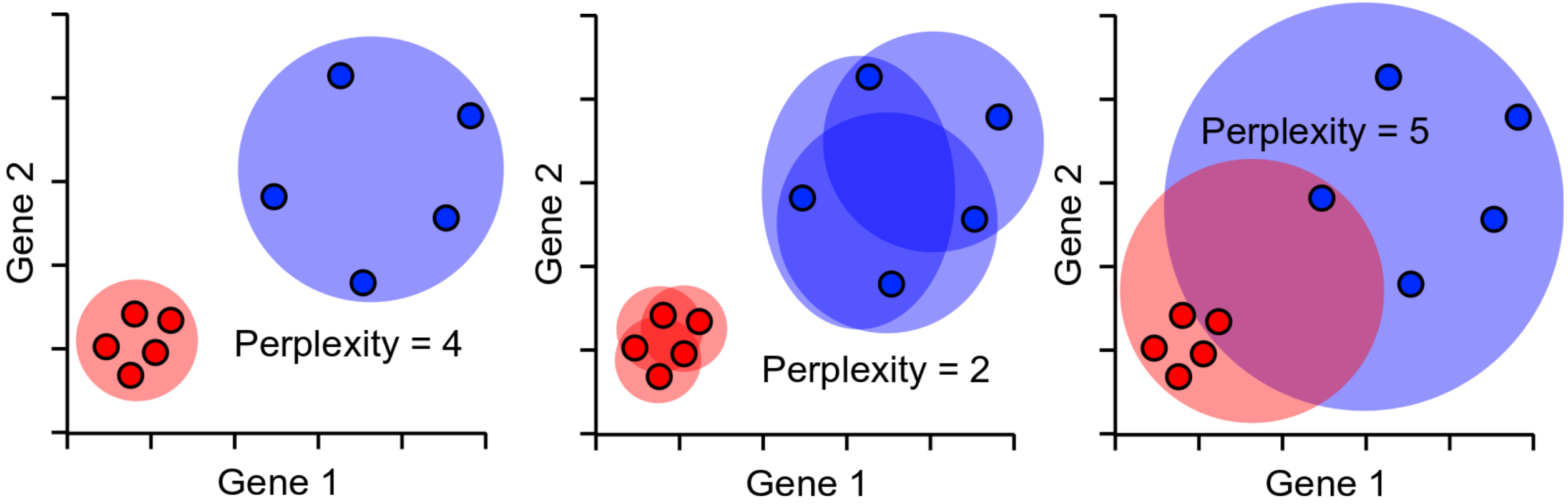
Escalado de distancia y perplejidad

- Perplejidad = número esperado de vecinos dentro de un clúster
- Distancias escaladas en relación con los vecinos de perplejidad



	0	4	6	586	657	836
	4	0	4	815	527	776
	9	3	0	752	656	732
	31	28	29	0	4	7
	31	24	25	4	0	7
	40	37	32	8	8	0

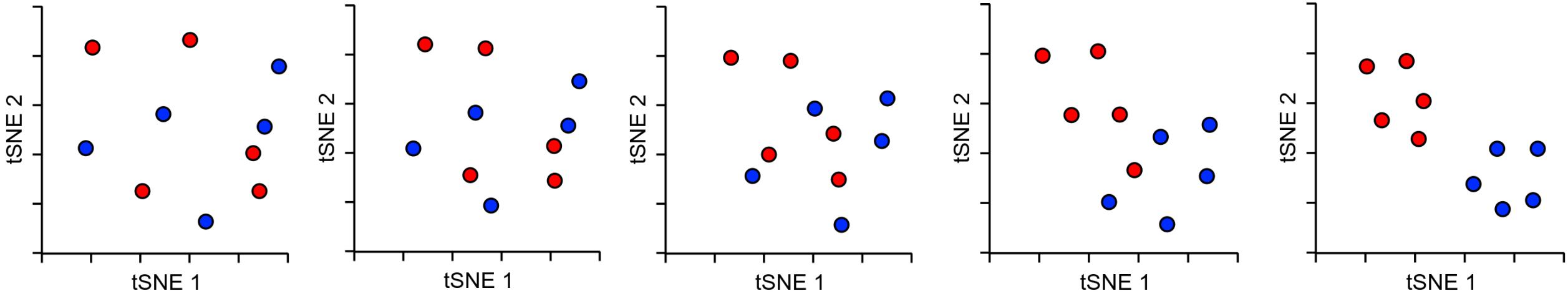
Perplejidad Robustez



Proyección tSNE

- Dispersar aleatoriamente todos los puntos dentro del espacio (normalmente 2D)
- Iniciar una simulación
  - El objetivo es hacer que las distancias de los puntos coincidan con la matriz de distancias
  - Baraja los puntos en función de lo bien que coincidan
  - Detenerse después de un número fijo de iteraciones, o
  - Detenerse después de que las distancias hayan convergido

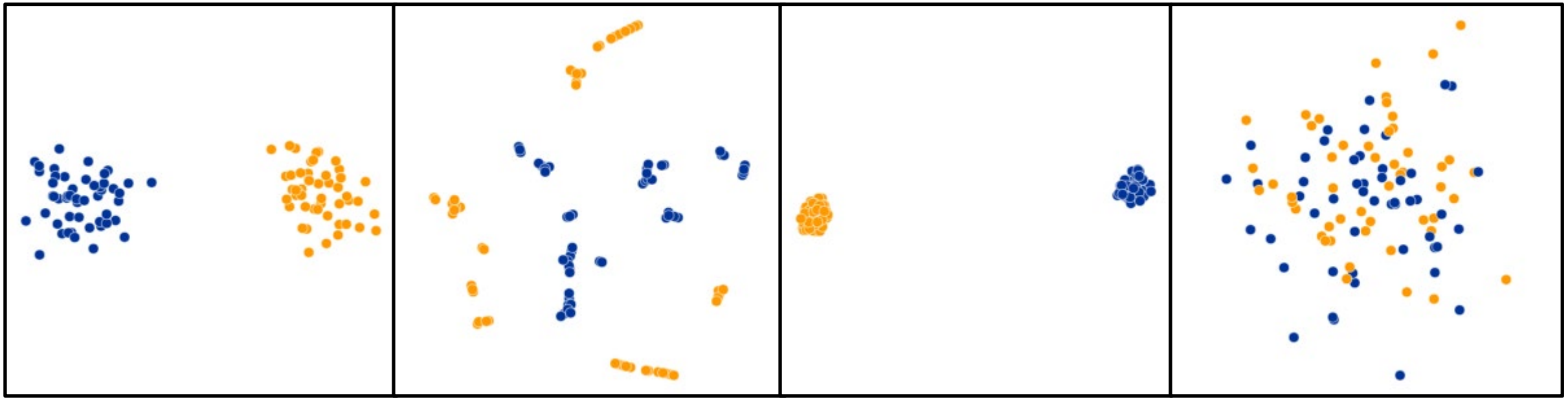
Proyección tSNE



- X e Y no significan nada (a diferencia de PCA)
- La distancia no significa nada (a diferencia de PCA)
- La proximidad es muy informativa
- La proximidad lejana no es muy interesante
- No se pueden racionalizar las distancias ni añadir más datos

Ejemplos prácticos de tSNE

Perplexity Settings Matter

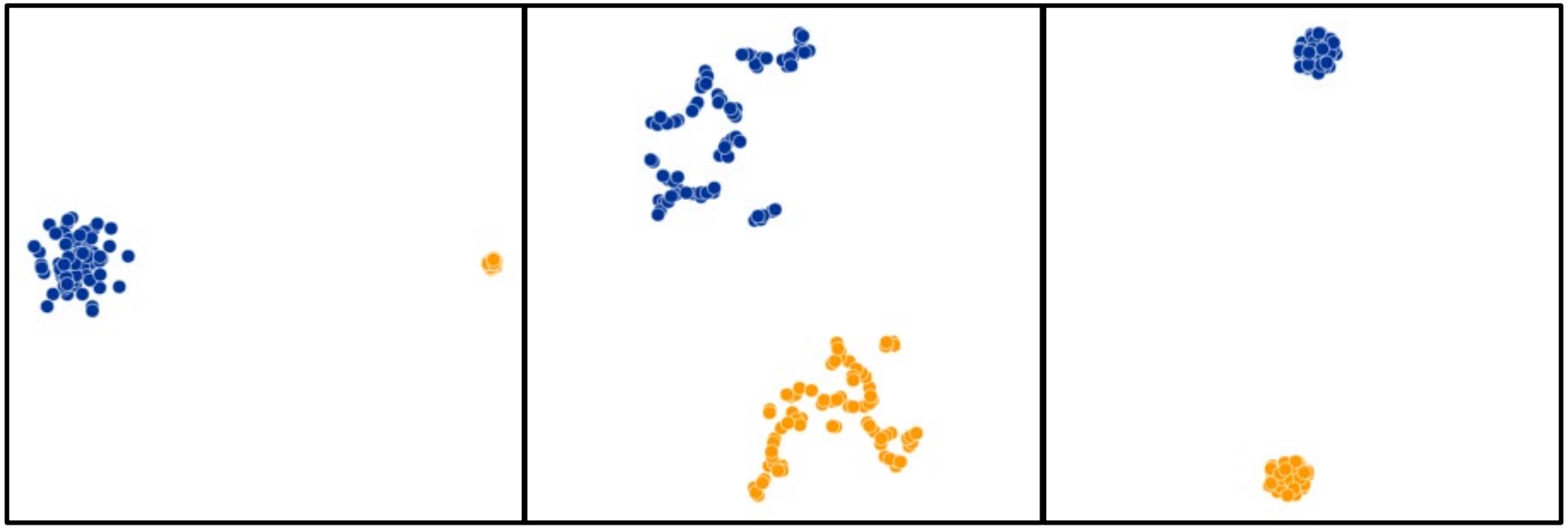


Original | Perplexity = 2 | Perplexity = 30 | Perplexity = 100

https://distill.pub/2016/misread-tsne/

Ejemplos prácticos de tSNE

Los tamaños de los clústeres no tienen sentido

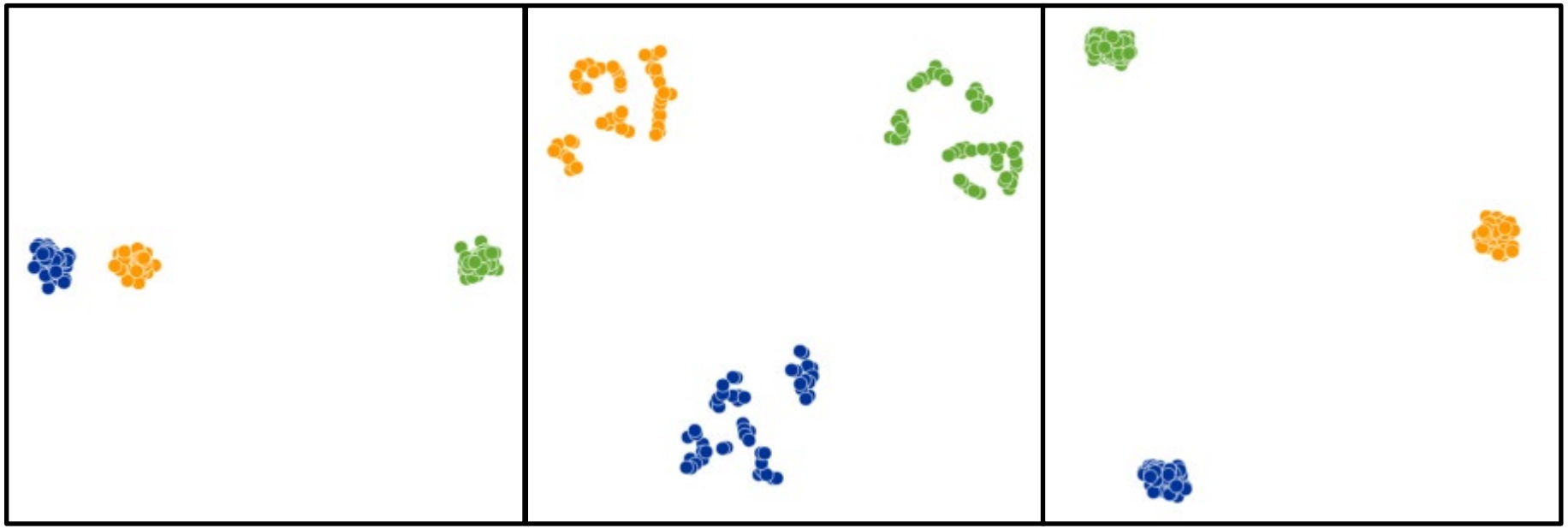


Original Perplexity = 5 Perplexity = 50

https://distill.pub/2016/misread-tsne/

Ejemplos prácticos de tSNE

No se puede confiar en las distancias entre clústeres



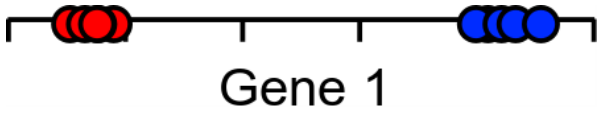
Original Perplexity = 5 Perplexity = 30

https://distill.pub/2016/misread-tsne/

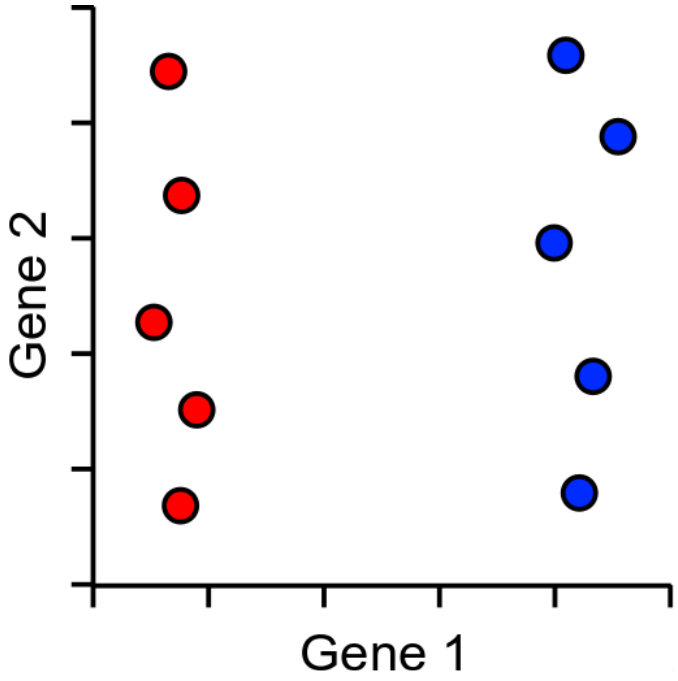
Entonces, ¿tSNE es genial?

Algo así...

Imagine un conjunto de datos con un solo gen superinformativo



Distance within cluster = low
Distance between clusters = high



Distance within cluster = higher
Distance between clusters = lower

- Now 3 genes
- Now 3,000 genes

- Todo está a la misma distancia de todo

¿Así que todo apesta?

- PCA
  - Requiere más de 2 dimensiones
  - Desconcertado por los datos cuantificados
  - Espera relaciones lineales
- tSNE
  - No puede hacer frente a los datos ruidosos
  - Pierde la capacidad de agruparse

Respuesta: Combine los dos métodos, obtenga lo mejor de ambos mundos

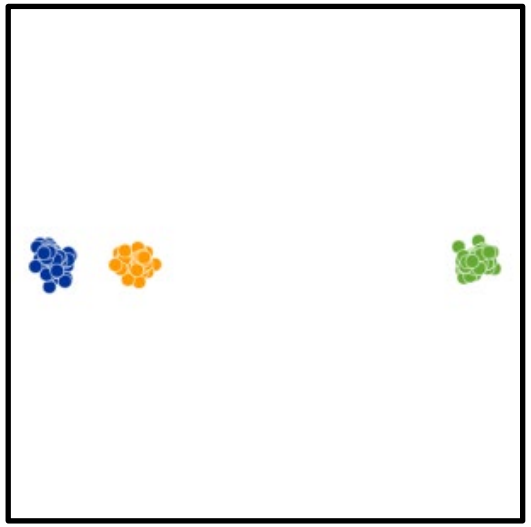
- PCA
  - Bueno para extraer la señal del ruido
  - Extrae dimensiones informativas
- tSNE
  - Puede reducir a 2D
  - Puede hacer frente al escalado no lineal

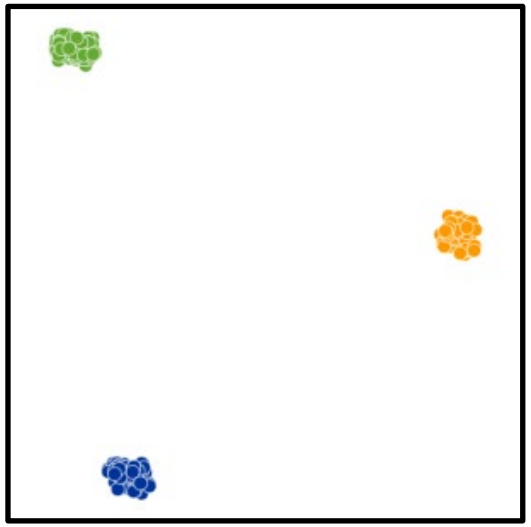
Esto es lo que hacen muchas pipelines en su análisis predeterminado

Entonces, ¿PCA + tSNE es genial?

Algo así…

- tSNE es lento. Este es probablemente su mayor crimen
  - tSNE no escala bien a un gran número de células (10k+)
- tSNE solo proporciona información fiable sobre los vecinos más cercanos, la información a gran distancia es casi irrelevante





¡UMAP al rescate!

- UMAP es un sustituto de tSNE para cumplir la misma función

- Conceptualmente muy similar a tSNE, pero con un par de cambios relevantes (y algo técnicos)
- El resultado práctico es el siguiente:
  - UMAP es bastante más rápido que tSNE
  - UMAP puede conservar una estructura más global que tSNE*
  - UMAP puede ejecutarse en datos sin procesar sin preprocesamiento de PCA*
  - UMAP puede permitir que se agreguen nuevos datos a una proyección existente

* En teoría, pero posiblemente no en la práctica

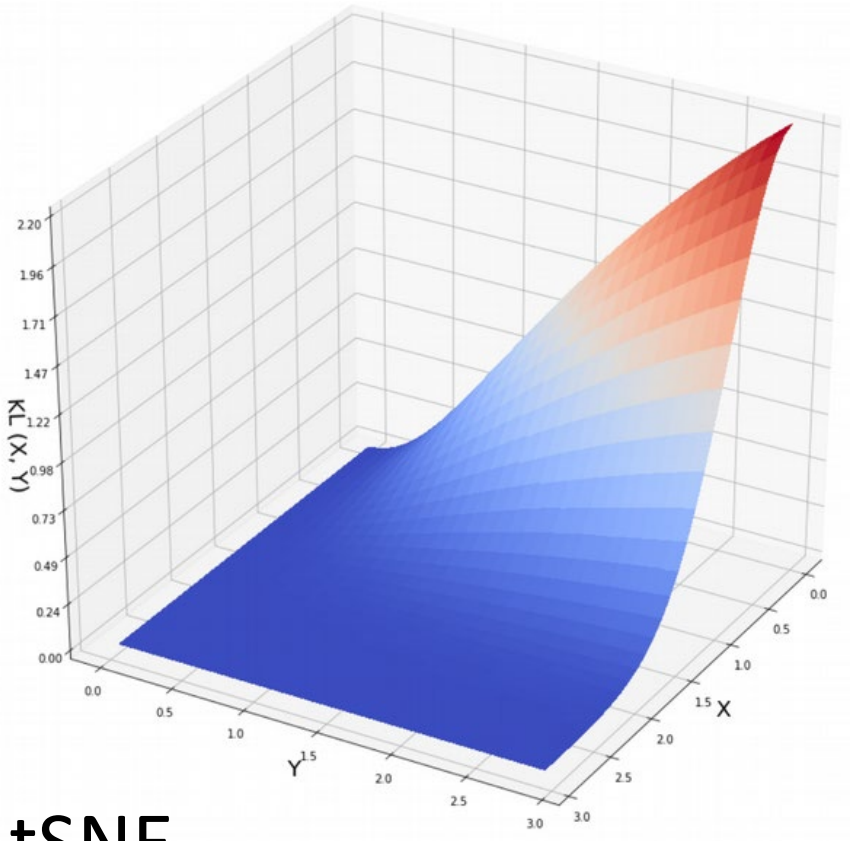
Diferencias UMAP

- En lugar del valor de perplejidad único en tSNE, UMAP define
  - **Nearest neighbours**: el número de vecinos más cercanos esperados, básicamente el mismo concepto que perplejidad
  - **Minimum distance**: la fuerza con la que UMAP empaqueta los puntos que están muy juntos

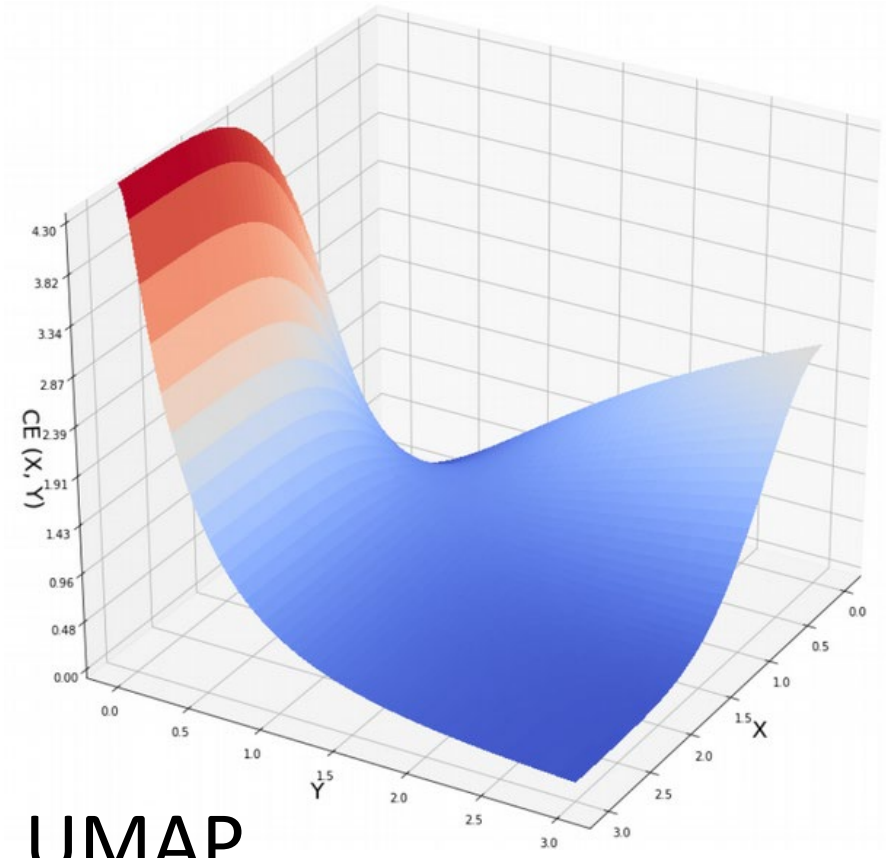
- Nearest neighbours afectará a la influencia que se le da a la información global frente a la local. La distancia mínima afectará a la compactación de las partes locales de la parcela.

Diferencias UMAP

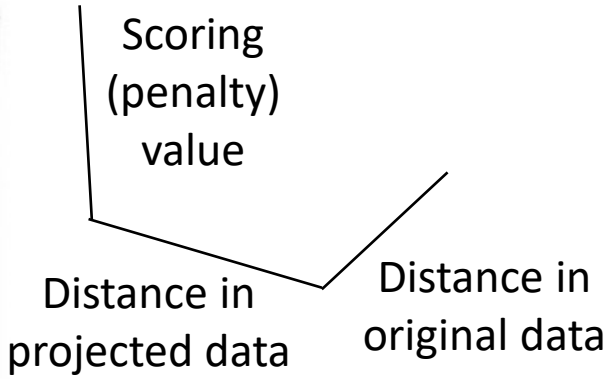
- Preservación de la estructura, principalmente en la puntuación de la proyección 2D



tSNE



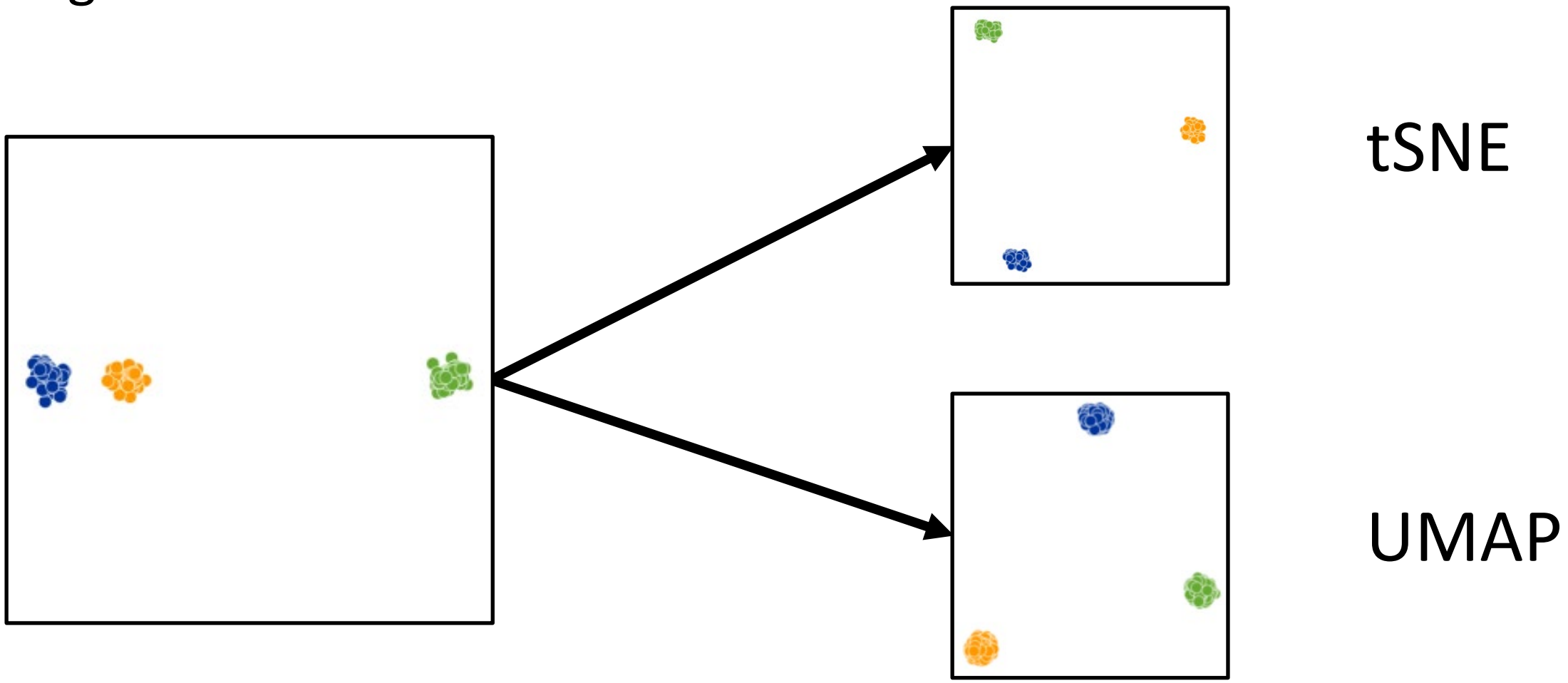
UMAP



https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668

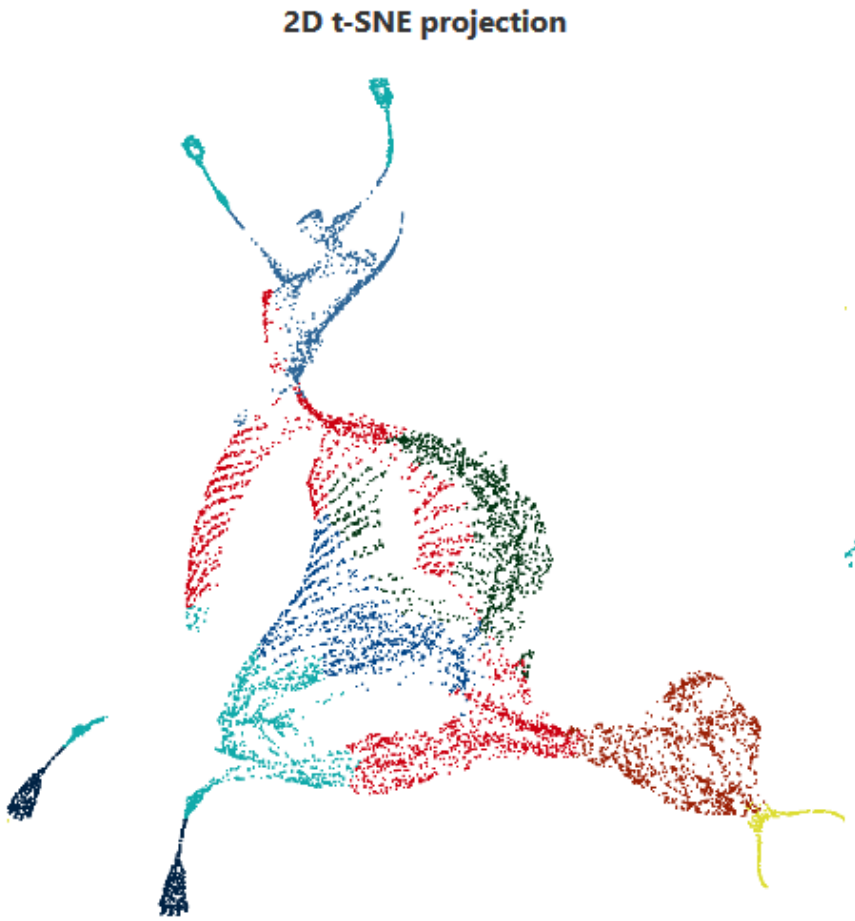
Entonces, ¿UMAP es genial?

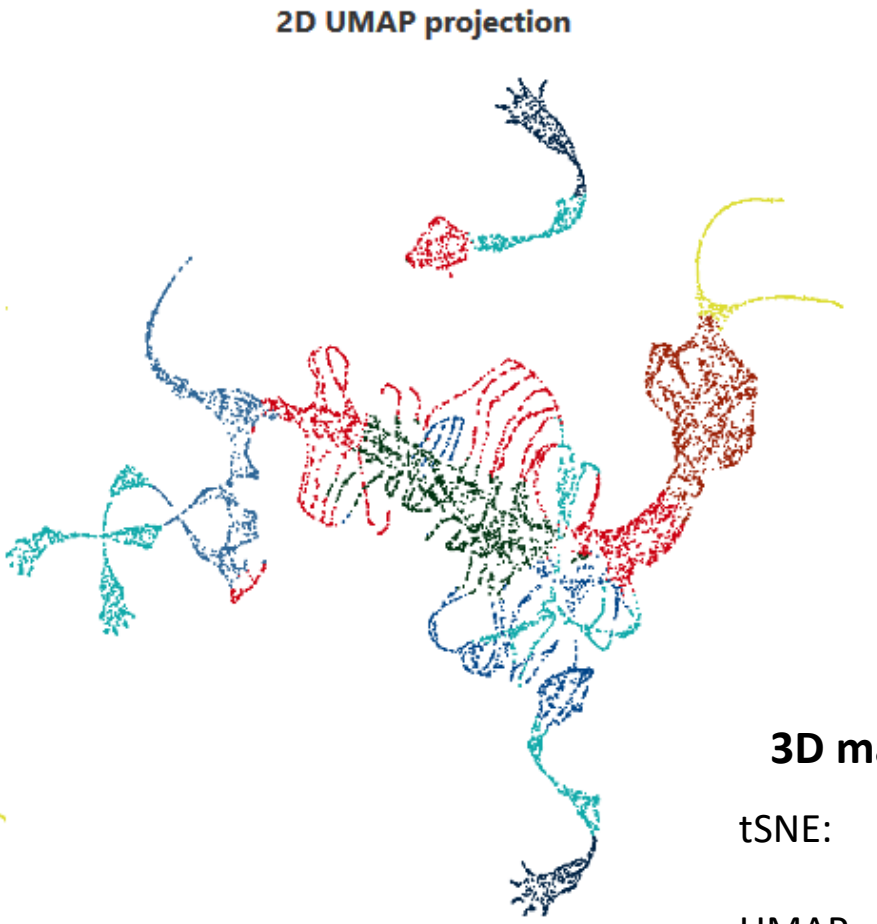
Algo así...



Entonces, ¿UMAP es todo exageración?

No, realmente funciona mejor para algunos conjuntos de datos...





3D mammoth skeleton projected into 2D

tSNE:	Perplexity 2000	2h 5min
UMAP:	Nneigh 200, mindist 0.25,	3min

https://pair-code.github.io/understanding-umap/

Enfoque práctico PCA + tSNE/UMAP

- Filtre mucho antes de comenzar
  - Células que se comportan bien
  - Genes expresados
  - Genes variables

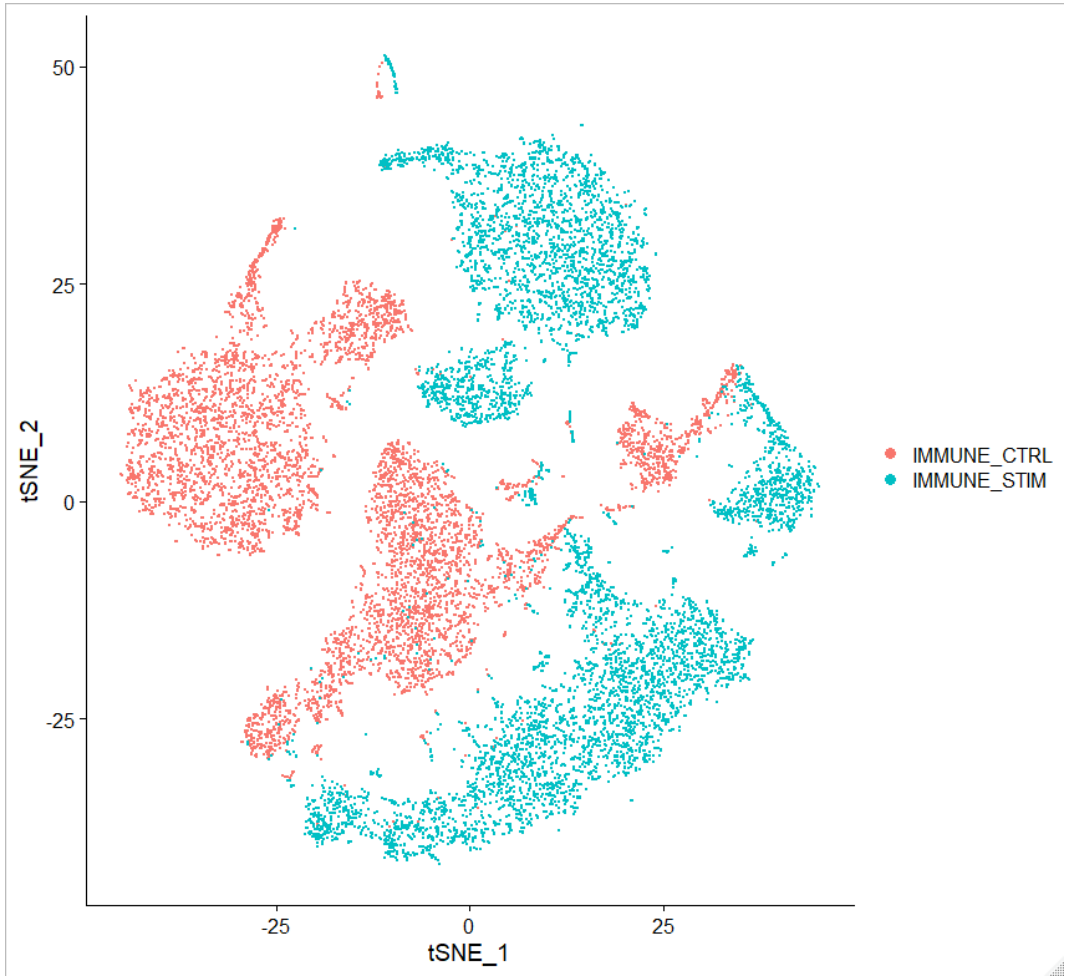
- Do PCA
  - Extraer la señal más interesante
  - Por ejemplo, los mejores PC. Reducir la dimensionalidad (pero no a 2)

- Do tSNE/UMAP
  - Calcular distancias a partir de proyecciones de PCA
  - Escala distancias y proyecta en 2 dimensiones

Entonces, ¿PCA + UMAP es genial?

Algo así... siempre y cuando solo tenga un conjunto de datos

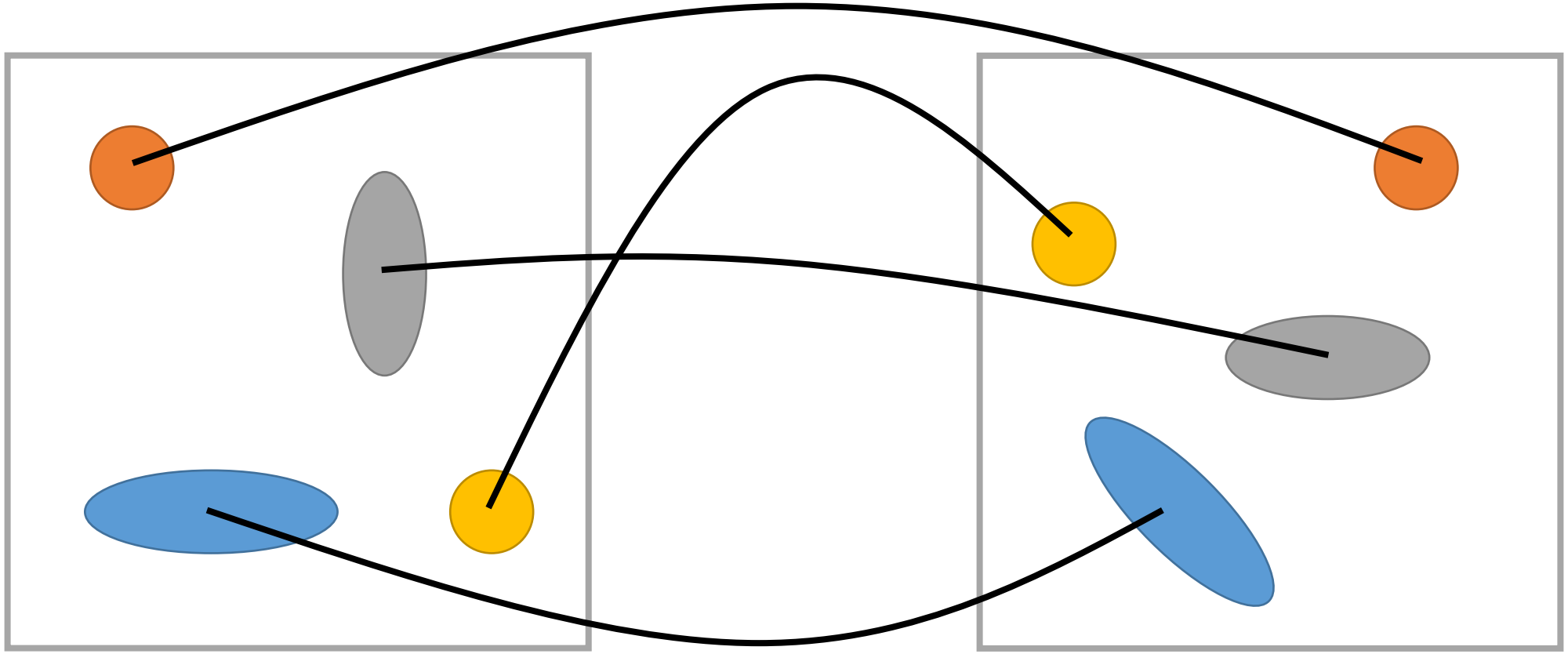
- En 10X, cada biblioteca es un "lote"
- Más sesgos a lo largo del tiempo/distancia
- Los sesgos impiden las comparaciones
- Necesidad de alinear los conjuntos de datos



Integración de datos

- Funciona sobre la base de que hay colecciones "equivalentes" de celdas en dos (o más) conjuntos de datos
- Encuentre puntos de "anclaje" que sean celdas equivalentes que deben estar alineadas
- Sesgar cuantitativamente los datos para alinear de forma óptima los anclajes

Integración UMAP/tSNE



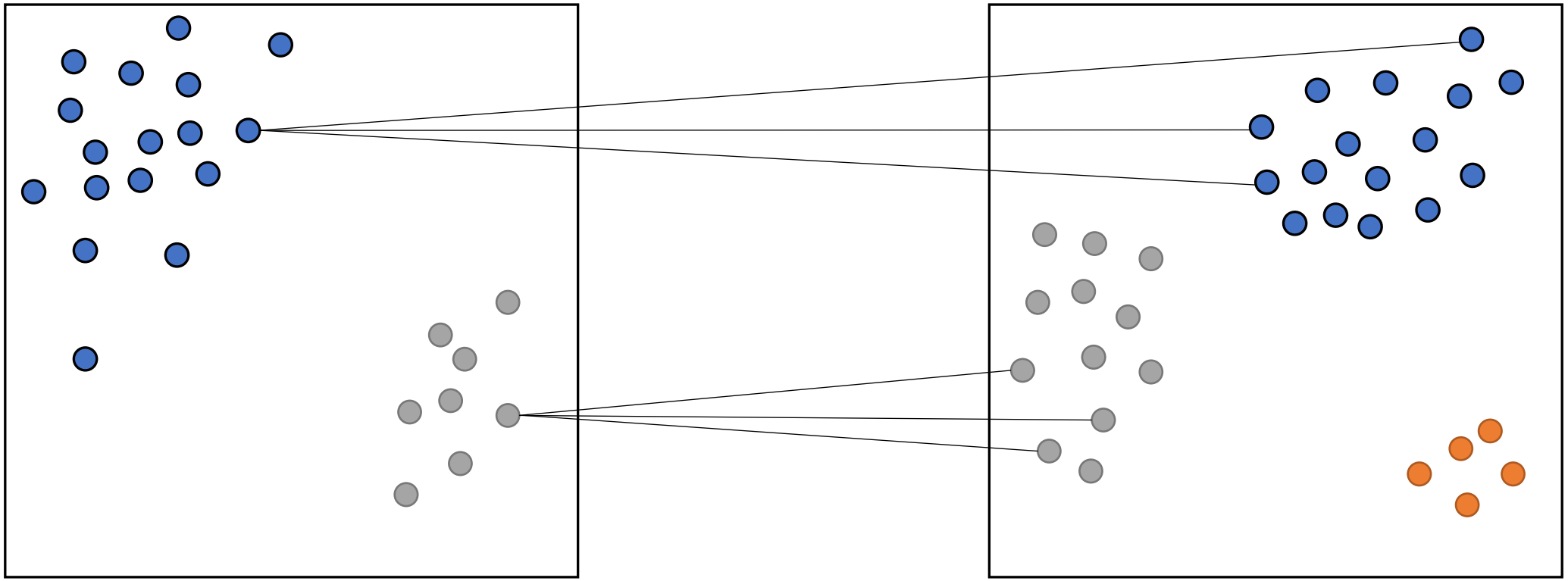
Definir puntos de "anclaje" clave entre celdas equivalentes

Integración UMAP/tSNE

Sesgar los datos para alinear los anclajes

Definición de anclajes de integración

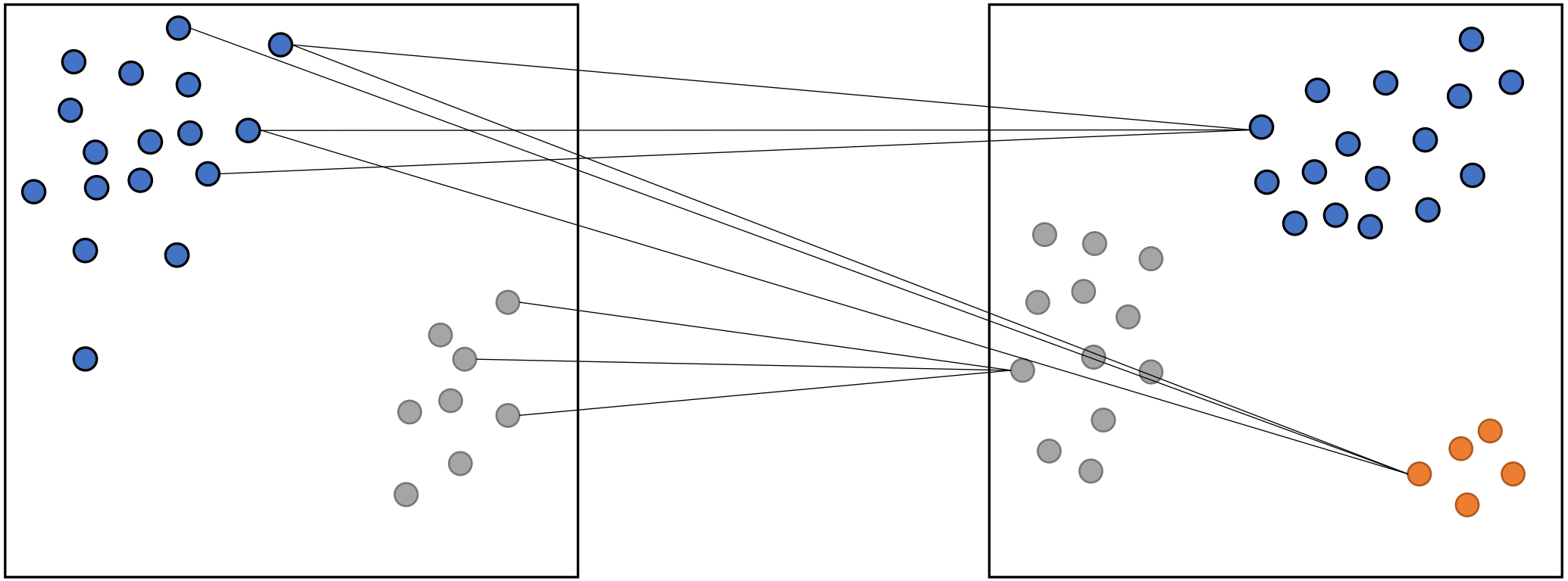
Mutual Nearest Neighbours (MNN)



Para cada celda de los datos1, encuentre las 3 celdas más cercanas de los datos2

Definición de anclajes de integración

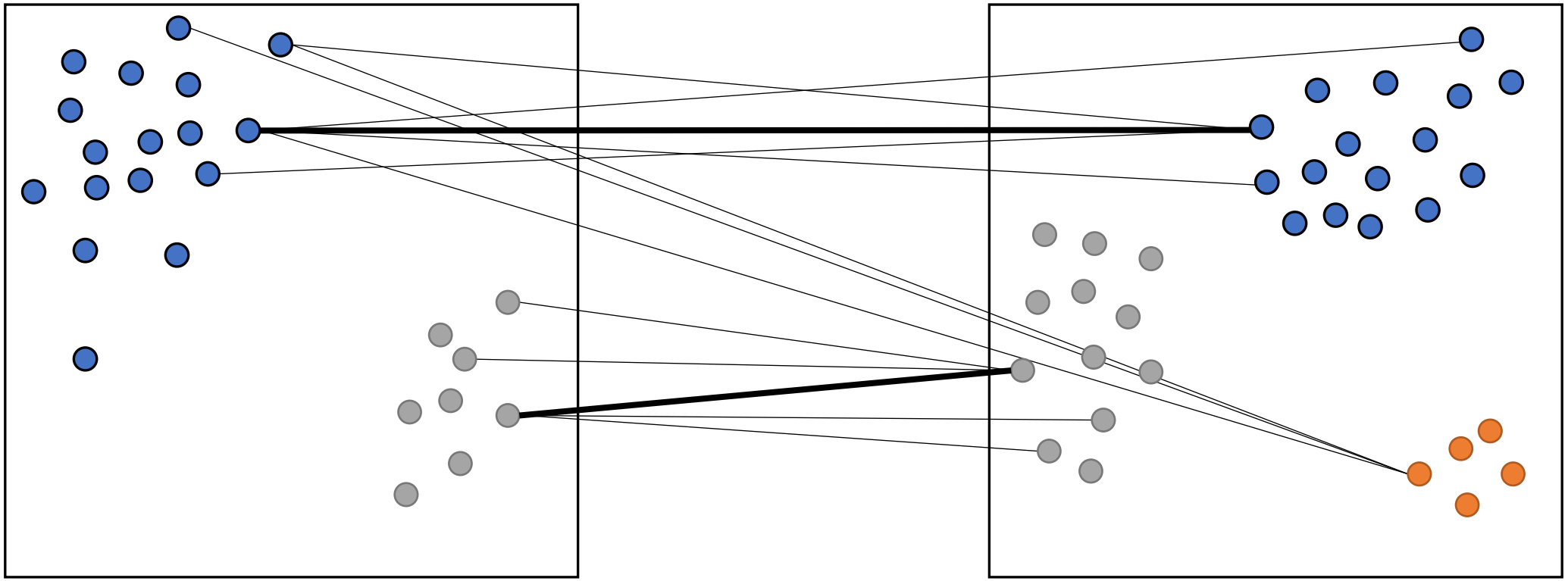
Mutual Nearest Neighbours (MNN)



Haz lo mismo al revés

Definición de anclajes de integración

Mutual Nearest Neighbours (MNN)



Seleccione pares de celdas que se encuentren en otros grupos vecinos más cercanos

Definición de los vecinos más cercanos

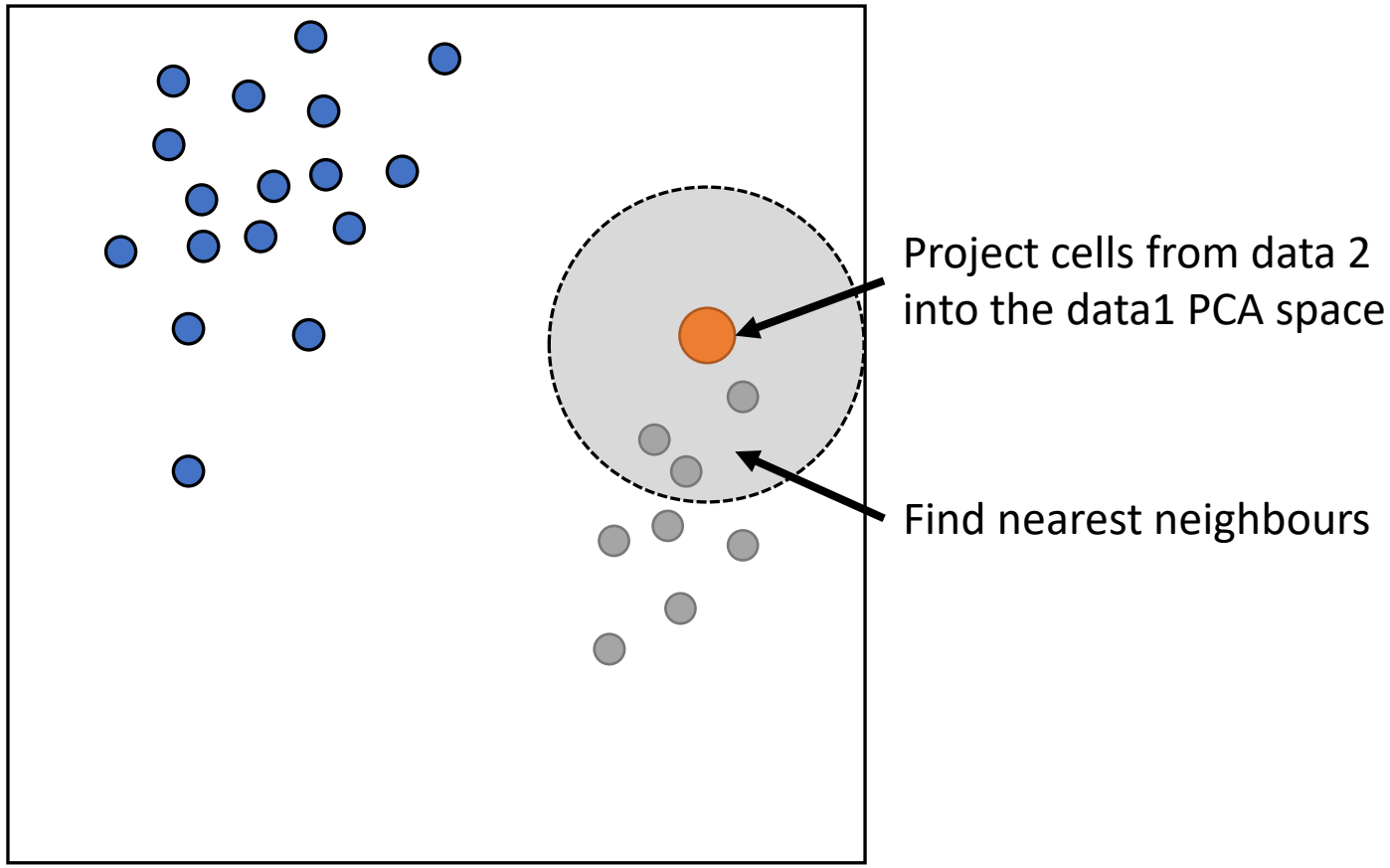
- Distancia en la cuantificación de la expresión original
  - Muy ruidoso (tecnología diferente, normalización, profundidad)
  - Lento y propenso a predicciones erróneas

- Usar una representación más limpia (menos ruidosa)
  - Principal Components (rPCA)

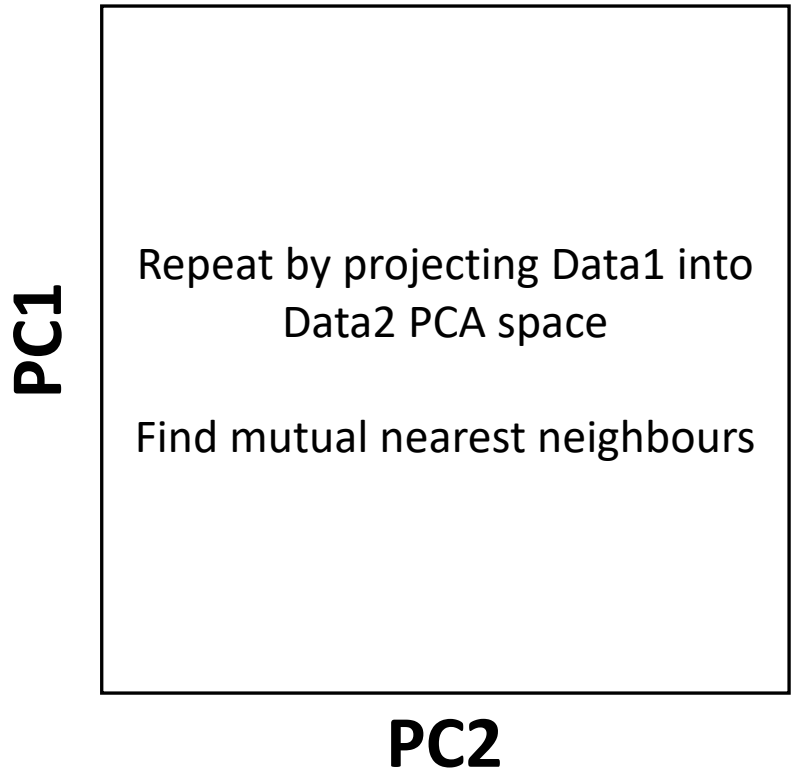
Definición de anclajes de integración

Reciprocal PCA

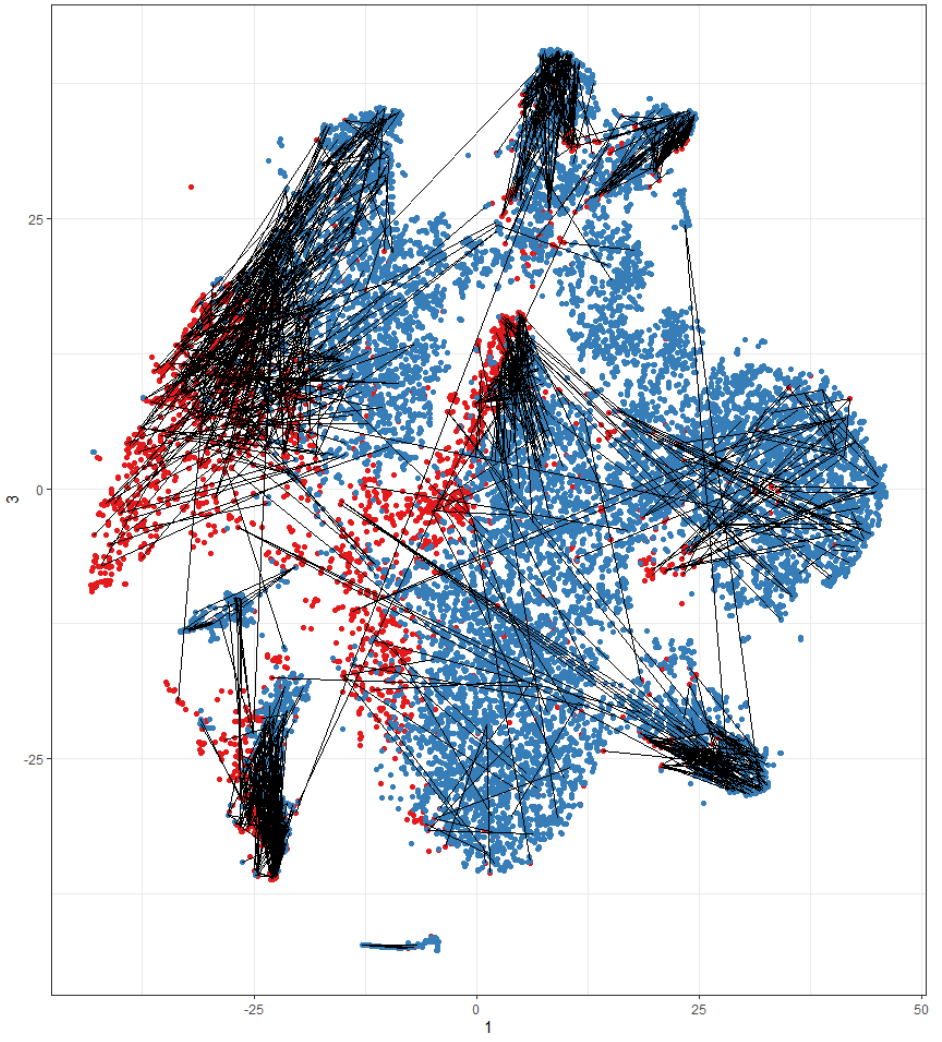
Define PCA Space for Data 1



Define PCA Space for Data 2



Anclajes de integración



Factores que afectan a la integración

- ¿Qué genes se someten a la integración?
  - Expresado en todos los conjuntos de datos
  - Variable en todos los conjuntos de datos
- Qué método se utiliza para definir los vecinos más cercanos
  - Datos normalizados, Correlación, PCA inverso
- ¿Cuántos vecinos más cercanos consideras?
  - El valor predeterminado es alrededor de 5, algunos clústeres requieren más (20)
- Otros filtros para eliminar artefactos