

# KNN- Vecinos más cercanos

Pontificia Universidad Javeriana  
Francisco Carlos Calderón Ph.D

# Objetivos:

Aplicar la técnica de los vecinos más cercanos, o los KNN en problema de clasificación y regresión.

Identificar las ventajas y desventajas del método de clasificación no paramétrico.

# Historia: Georgy Voronoy

Nació	Georgy Feodosevich Voronoy (Георгий Феодосьевич Вороной)  28 Abril de 1868 Zhuravka, Imperio Ruso
Murió	20 Noviembre de 1908 40 años ☺ Varsovia, Imperio Ruso



Tutor doctoral de Boris Delaunay, su hijo realizó el primer trasplante de riñón.

[https://en.wikipedia.org/wiki/Georgy\\_Voronoy](https://en.wikipedia.org/wiki/Georgy_Voronoy)

# Historia: Boris Delaunay

Nació	<b>Boris Nikolaevich Delaunay or Delone</b> , Борис Николаевич Делонé 15 marzo, 1890 San petersburgo, Imperio Ruso
Died	July 17, 1980 (aged 90) Moscu, Unión Sovietica
Conocido por	Triangulación con su apellido, Ser escalador
<b>Carrera Científica:</b>	
Estudiante doctoral de:	Dmitry Grave, Georgy Voronoy

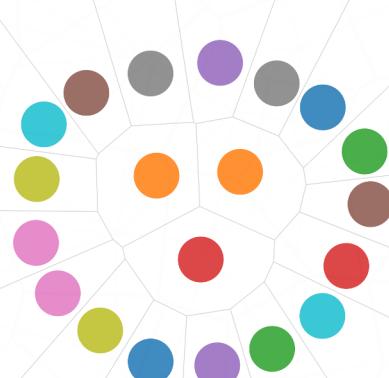


<http://www.mountain.ru/mkk/pers/delone.shtml>

# Teselaciones P2.

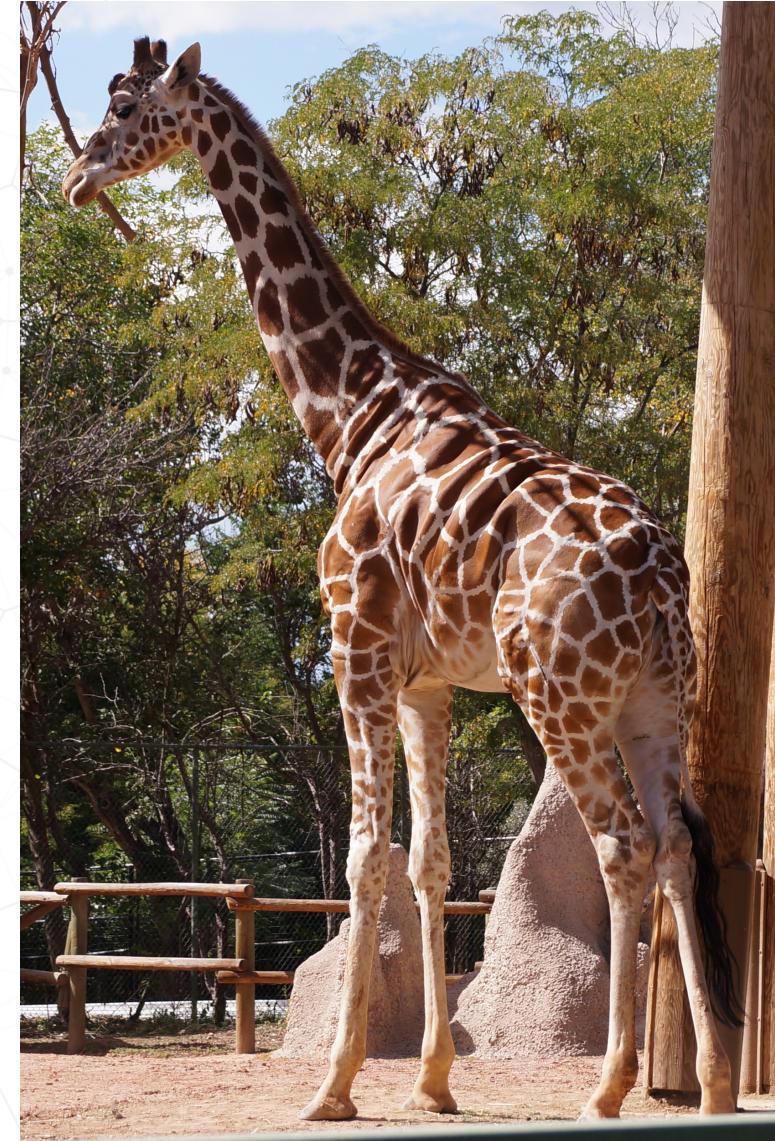
Hemos usado en nuestro curso K-Medias.

- Llegábamos a una Teselación a partir de los centroides.
- Las fronteras de decisión están dadas por el diagrama de Voronoi.
- 
- Vamos a usar un enfoque similar para
- Hacer regresión y clasificación.

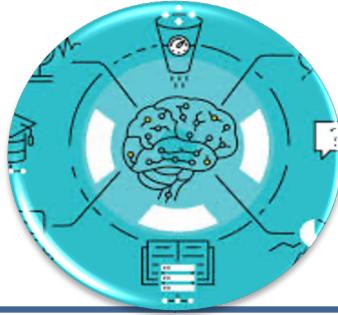


Fondo tomado de: (Ejercicio, jugar con el applet)  
<https://observablehq.com/@d3-hover-voronoi>

Al final de este curso debe estar tan encantados con las teselaciones que se las pueden tatuar, como la Jirafa!



Fuente propia, Zoo de Denver, USA



Aprendizaje



Restringidos  
“Constraint”



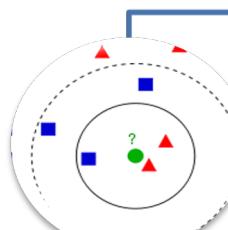
One Shot



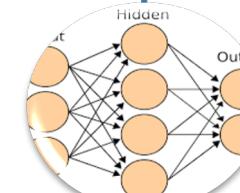
Basado en  
explicación



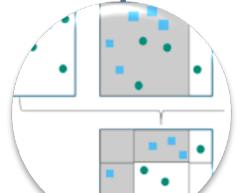
Regularizados  
“Regularity”



KNN



ANN



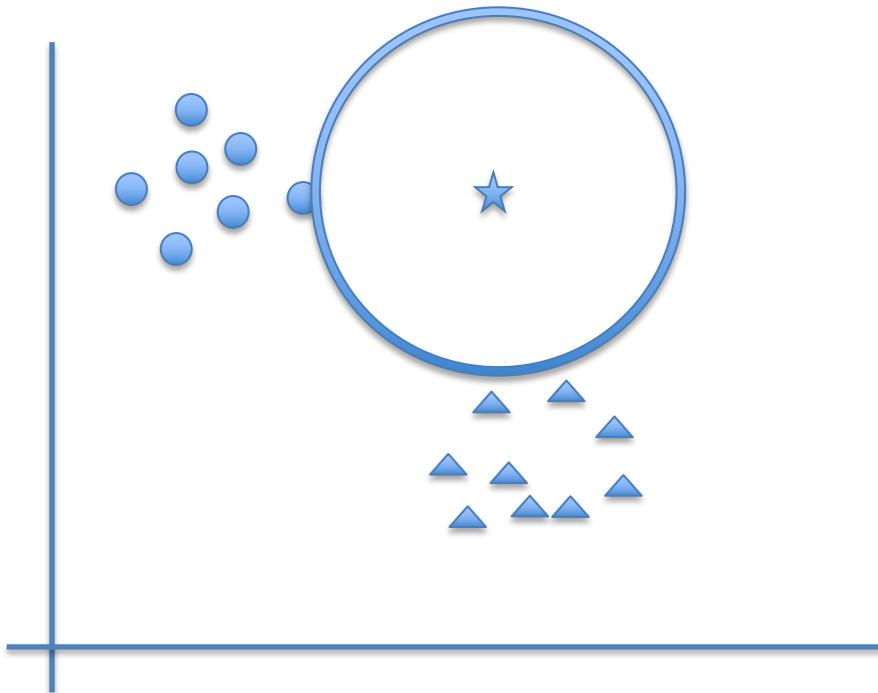
Boosting

**KNN**

<https://animals.sandiegozoo.org/animals/secretary-bird>

Facultad de ingeniería  
Deptº de electrónica

# K - Vecinos Más Cercanos, Clasificación



Para clasificar una nueva muestra  $z$  a partir del conjunto de entrenamiento  $x_i$ .

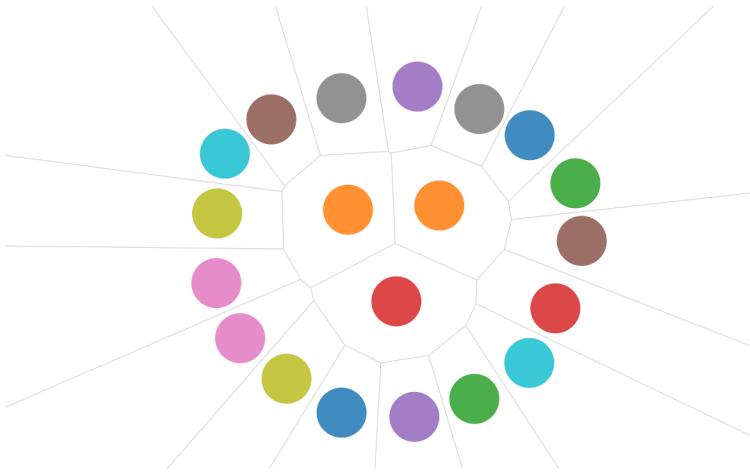
1. Se hallan las distancias\* de  $z$  a todos los  $x_i$
2. Se toman las menores  $k$  distancias.
3. Se selecciona la clase a partir del  $y_i$  asociado a los  $x_i$  de menor distancia con  $z$ .

Si  $k=1$  se crea una partición de Voronoi.

La selección puede hacerse a partir de la moda “el de mayor votación”

\*Por distancia se hace referencia a cualquier métrica, por ahora usemos la distancia euclidiana

# Consideraciones ☺



Para 2 clases se debe tomar un k impar.

- Solo es necesaria 1 muestra por clase para lograr un clasificador por KNN.
- Se debe seleccionar cuidadosamente el  $k$ .
- $K$  no debe ser un múltiplo del número de clases.
  - e.g. si hay 6 clases se puede tomar  $K=6$ .
- KNN es considerada una técnica No paramétrica en sentido estadístico.
  - KNN es clasificado como un “lazy learner” ya que no estima una función de frontera de decisión.

# Problemas ☹

- Si se toman muchas muestras  $x_i$  la clasificación puede llegar a ser lenta.
- No existe un método estándar para determinar un valor óptimo para k.
  - Valores pequeños de k son susceptibles a afectaciones por valores fuera de tendencia “Ruido”
  - Valores Grandes de k son más inmunes a ruido pero si k es muy grande las categorías con pocas muestras pueden llegar a no ser seleccionadas nunca.
  - Si se selecciona mal el k se puede llegar fácilmente a “grandes” regiones incongruentes empatadas en votación.

# Thumb rules! Selección del K



Tomado de [Quora](#)

Puede usarse k como el número impar más cercano a la raíz cuadrada de n donde n es:

- El número total de datos  $x_i$
- El número de datos en la clase más pequeña.

# Thumb rules! Cuando usar KNN

- Cuando los datos están bien etiquetados.
  - No hay ruido\* aparente en las etiquetas.
- Cuando hay pocos datos.
  - Lo contradictorio es que KNN funciona mejor si hay muchos datos!
- Cuando no les están pagando bien por el trabajo y quieren terminar rápido y entregar algo que funcione.



\*Por ejemplo bases de datos hechas a partir de encuestas pueden tener ruido

# Regresión por KNN

KNN puede ser usado para hacer regresión.

Se puede usar el promedio\* de  $y_i$  para los K-vecinos más cercanos al punto z.

\*Puede usarse en realidad cualquier estimador del valor esperado.

# Mejoras a KNN

Puede implementarse una mejora tanto para regresión como para clasificación por KNN.

El valor estimado de clasificación o regresión a partir de  $y_i$  se hace pesando el estimador a partir de las distancias a  $z$

# Ejercicio en clase

Realizar un ejemplo usando python del método de KNN para clasificación.

Experimentar con diferentes configuraciones.