

A decorative phylogenetic tree graphic is positioned on the left side of the slide. It consists of two main vertical columns of nodes, each represented by a small blue circle. The top column is colored cyan and branches out towards the right. The bottom column is colored blue and branches out towards the right. The two columns meet at a central point near the bottom, where several horizontal lines converge.

# **Modulo “Analisis de Datos Cientificos y Geograficos”**

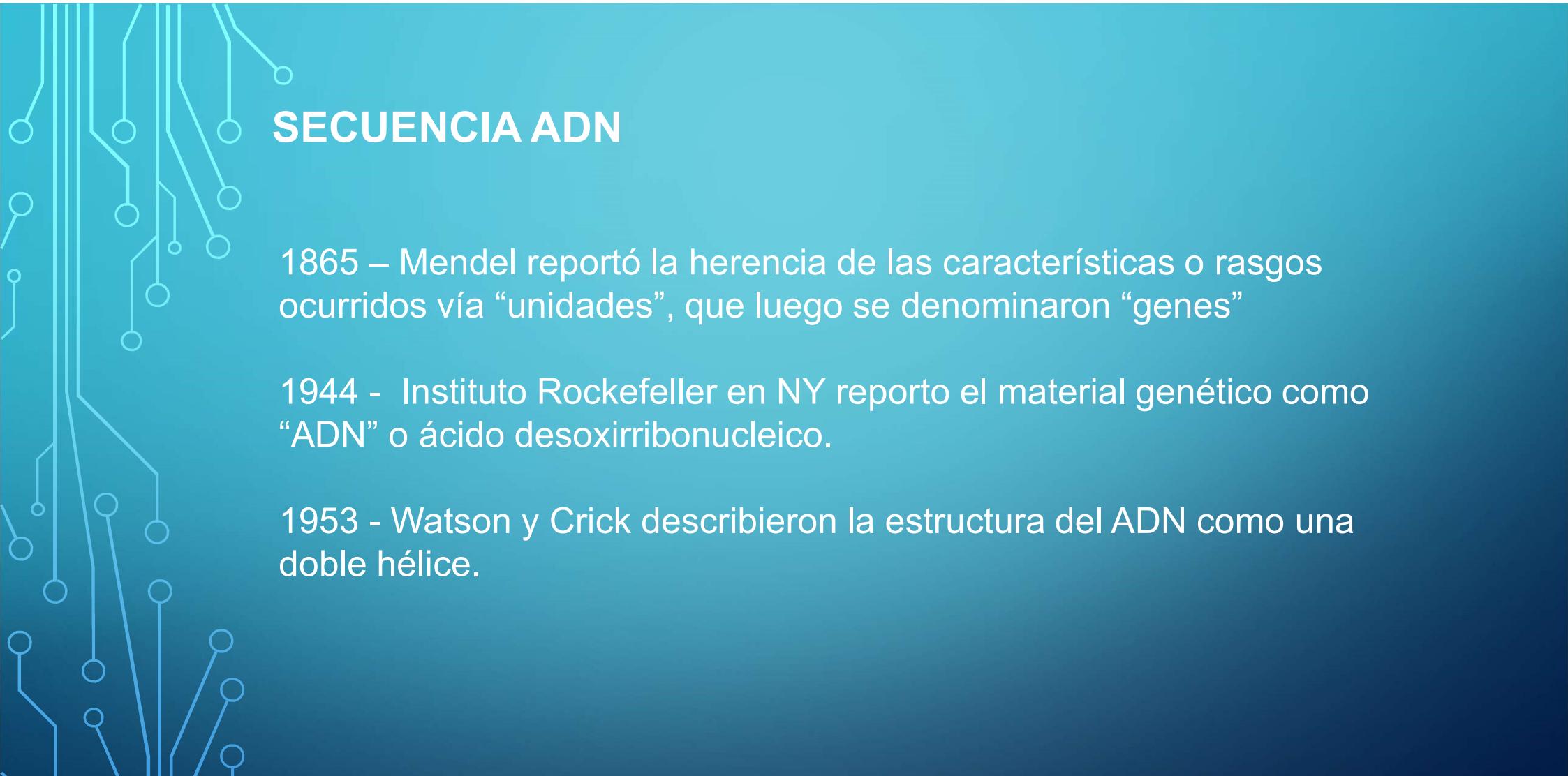
## **GENOMICA**



## GENÉTICA vs. GENÓMICA

La genética es una rama de biología enfocada en la herencia y la variación de los organismos. En términos más simples, la genética se centra en las características o rasgos que se transmiten de padres a hijos, de una generación a otra.

La genómica es una disciplina científica enfocada en el mapeo genético, la secuenciación de ADN, y el análisis del genoma completo de un organismo, incluyendo organizar los resultados en bases de datos. El genoma se refiere al material genético de un organismo.



## SECUENCIA ADN

1865 – Mendel reportó la herencia de las características o rasgos ocurridos vía “unidades”, que luego se denominaron “genes”

1944 - Instituto Rockefeller en NY reportó el material genético como “ADN” o ácido desoxirribonucleico.

1953 - Watson y Crick describieron la estructura del ADN como una doble hélice.



## SECUENCIA ADN

1975 – Se desarrollan métodos de secuenciación de ADN: determinar la secuencia de los bloques de ADN, denominados nucleótidos, que constan de:

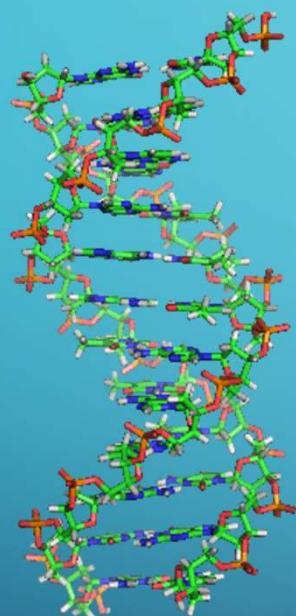
- una molécula de azúcar (desoxirribosa en el ADN)
- un grupo fosfato
- una base con nitrógeno:
  - citosina (C)
  - guanina (G)
  - adenina (A)
  - timina (T)



Aquí nos  
enfocaremos !



## SECUENCIA ADN

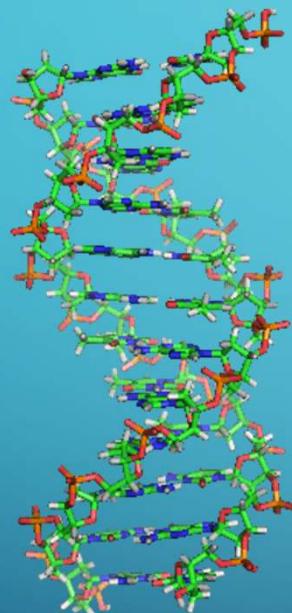


La doble hélice, similar a una escalera de caracol, permite al ADN la capacidad de almacenar y transmitir información.

Las bases se conectan a través de las dos hebras de la doble hélice:

- la citosina (C) se aparea con la guanina (G)
- la adenina (A) se aparea con la timina (T).

## SECUENCIA ADN

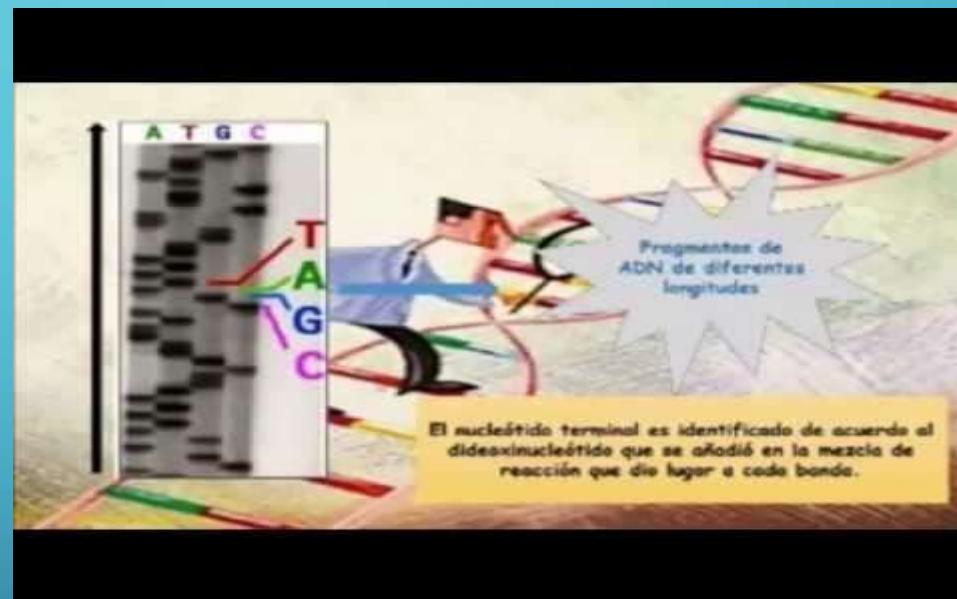


El **ADN** contiene las instrucciones genéticas utilizadas en el desarrollo y funcionamiento de todos los organismos vivos y es responsable de su transmisión hereditaria: contiene las instrucciones necesarias para construir otros componentes de las células.

Los segmentos de ADN que llevan esta información genética se llaman **GENES**, y los otros segmentos tienen propósitos estructurales o regulan el uso de esta información genética.

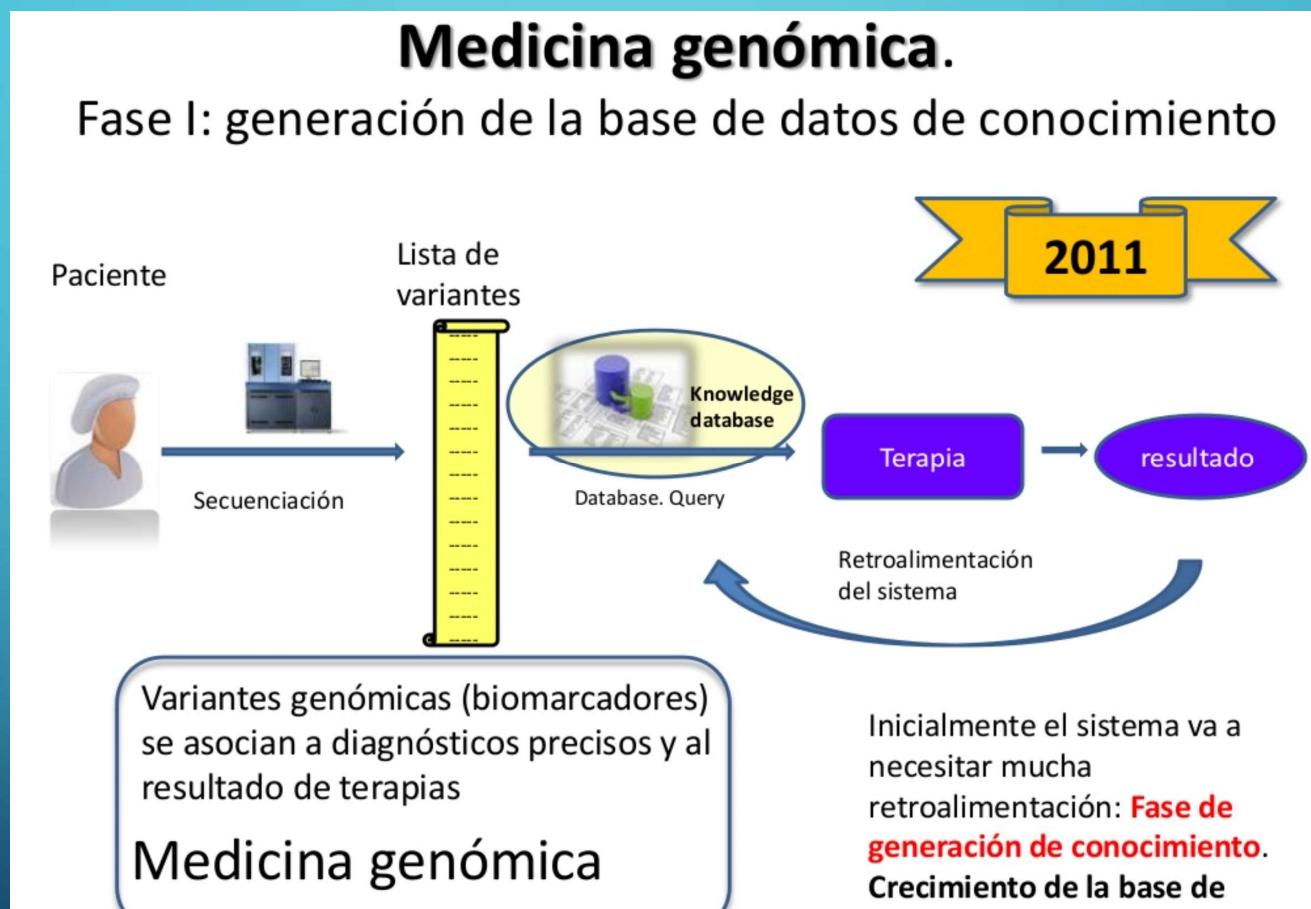
**Alfabeto: ACGT**

# SECUENCIA ADN



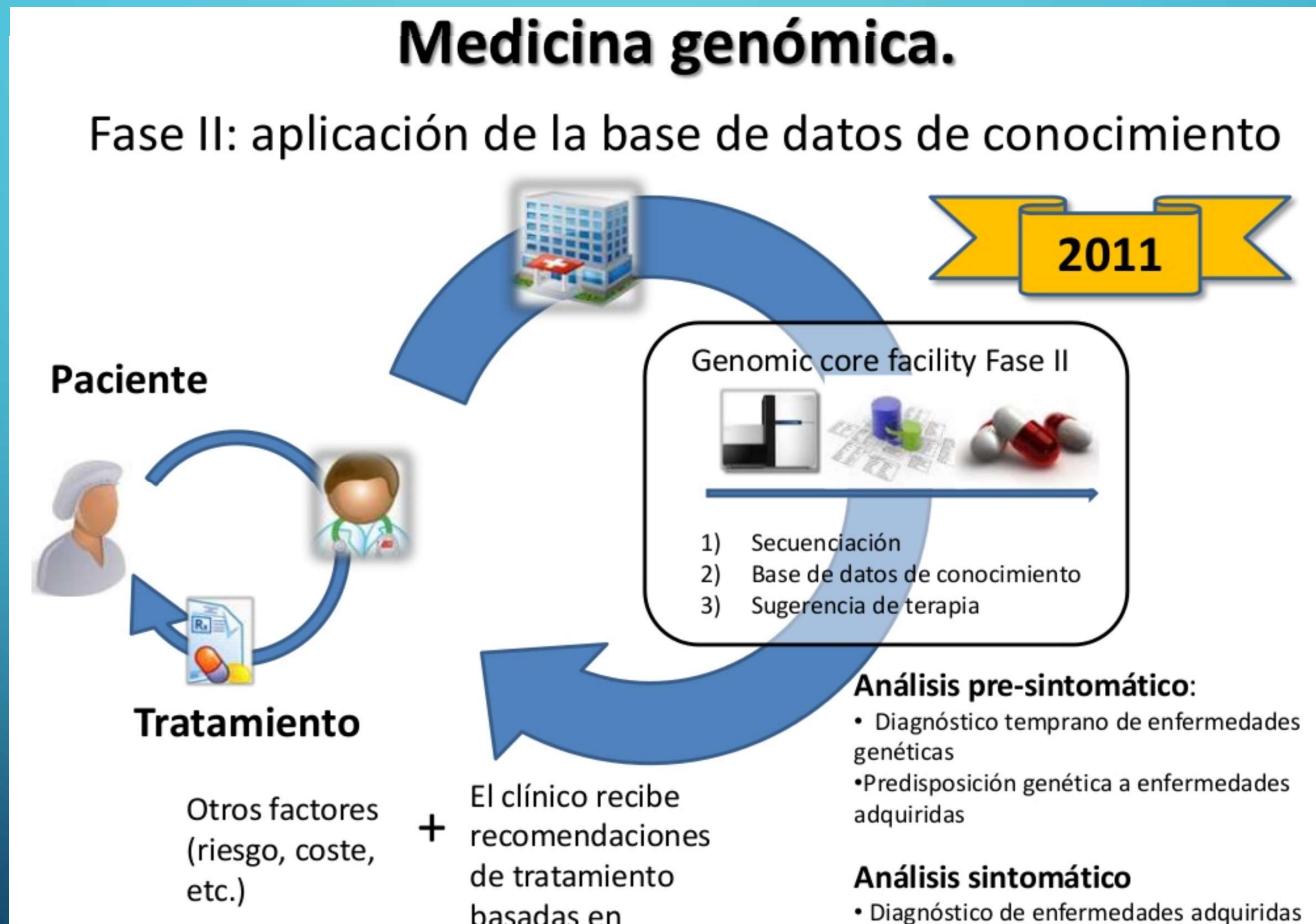
<https://www.youtube.com/watch?v=1JfHNedZUxU>

# HACIA DONDE VAMOS?



Fuente: Computational Genomics  
Departamento

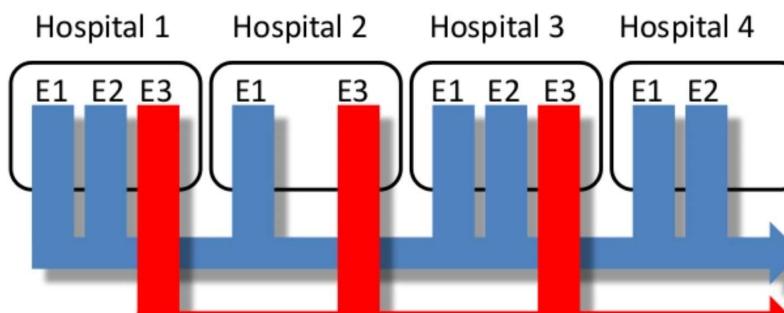
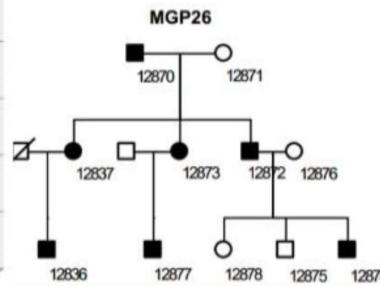
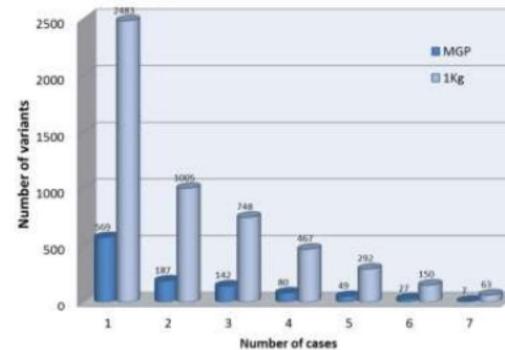
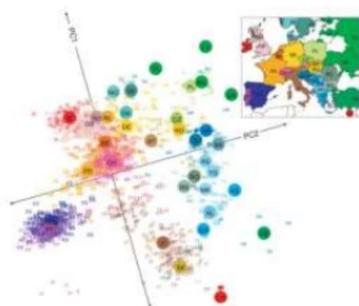
# HACIA DONDE VAMOS?



Fuente: Computational Genomics  
Departamento

## HACIA DONDE VAMOS?

### El efecto de la variación local

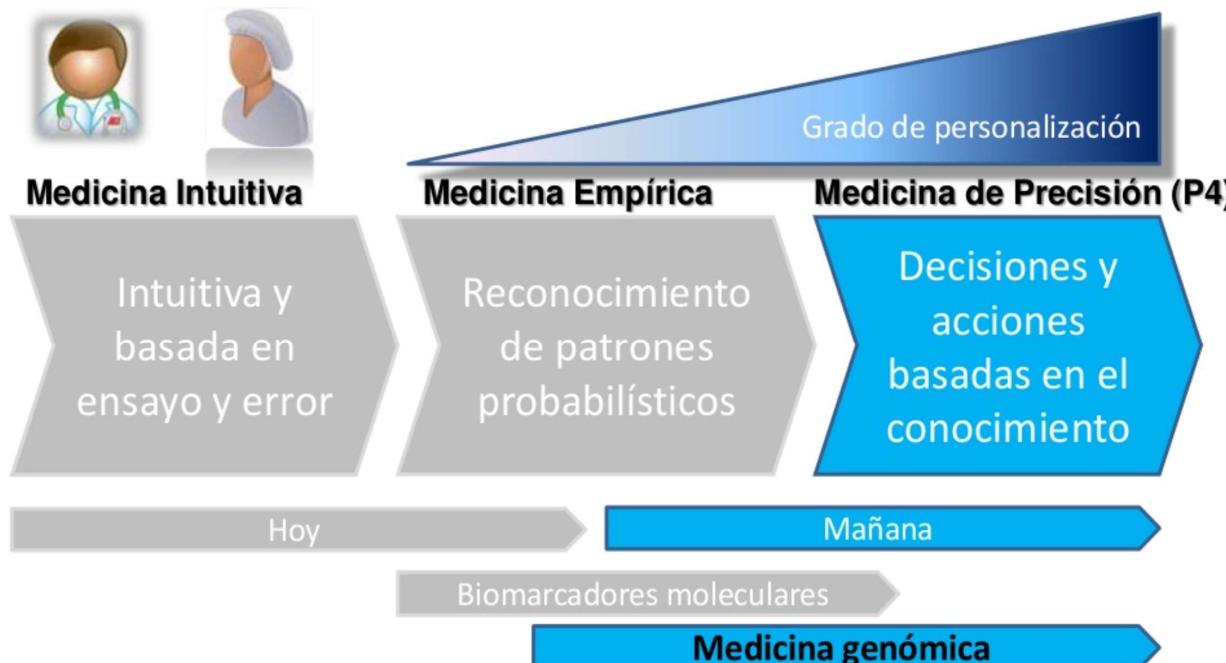


Y su uso sin comprometer la confidencialidad de los datos genómicos

Fuente: Computational Genomics Departamnate

# HACIA DONDE VAMOS?

## La transición a la medicina de precisión



El uso de nuevos algoritmos que permitan hacer modelos que **relacionen** funcionalmente el **genotipo** con la **enfermedad** o con los mecanismos de acción de los **fármacos** permitirá una verdadera **transición a la medicina de**

Fuente: Computational Genomics  
Departamento



## DESAFIOS DE LA GENOMICA

La mayor parte de los datos biológicos serán secuencias del genoma humano, junto a información médica relacionada

En la actualidad, las tecnologías de secuenciación generan hasta 6 Gb de lecturas de cadenas de ADN en cada ejecución...

**Procesar dicho volumen lleva varios días**

Al ritmo actual, la cantidad de datos de genómica producidos diariamente se duplicará cada 7 meses....

En 2025, esa cifra oscilará entre 2 y 40 exabytes por año !  
**(1 exabyte = 1.000 petabytes = 1.000.000 terabytes)**



## DESAFIOS DE LA GENOMICA

El gran desafío es encontrar la manera de capturar, almacenar, procesar e interpretar toda esa información biológica codificada en el genoma.

Ademas, habrá que homogeneizarlos, ya que se van almacenando en distintos formatos.



## DESAFIOS DE LA GENOMICA

El gran desafío es encontrar la manera de capturar, **almacenar**, procesar e interpretar toda esa información biológica codificada en el genoma.

Ademas, habrá que homogeneizarlos, ya que se van almacenando en distintos formatos.



Compactacion?



## DESAFIOS DE LA GENOMICA

El gran desafío es encontrar la manera de capturar, almacenar, **procesar e interpretar** toda esa información biológica codificada en el genoma.

Ademas, habrá que homogeneizarlos, ya que se van almacenando en distintos formatos.

Alineación eficiente?



## DESAFIOS DE LA GENOMICA

**Podremos aplicar técnicas de Big Data?**

Universidad de Missouri: RNAMiner permite analizar hasta cinco variables en los genomas completos de cinco especies: humano, ratón, un tipo de mosca, una pequeña planta de flores y un tipo de bacteria.

Tambien permite aportar datos genómicos de cualquier especie para acrecentar la base de datos.

2 Gb de datos requieren aproximadamente 10 horas...  
(pero 1 exabyte= 1.000.000.000 Gigabytes)



## DESAFIOS DE LA GENOMICA

Podremos aplicar técnicas de Big Data?

El centro CITIUS crea BigBWA, una nueva herramienta que permite aprovechar las ventajas de las tecnologías *Big Data* para incrementar el rendimiento de las operaciones de alineado.

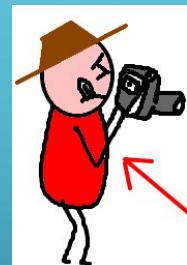
Esta basado en Hadoop: con un cluster de 6 servidores, reduce un procesamiento típico de 4 días a 8 horas !



# ALINEAMIENTO DE SECUENCIAS ADN



# ALINEAMIENTO DE SECUENCIAS ADN



# ALINEAMIENTO DE SECUENCIAS ADN



# ALINEAMIENTO DE SECUENCIAS ADN



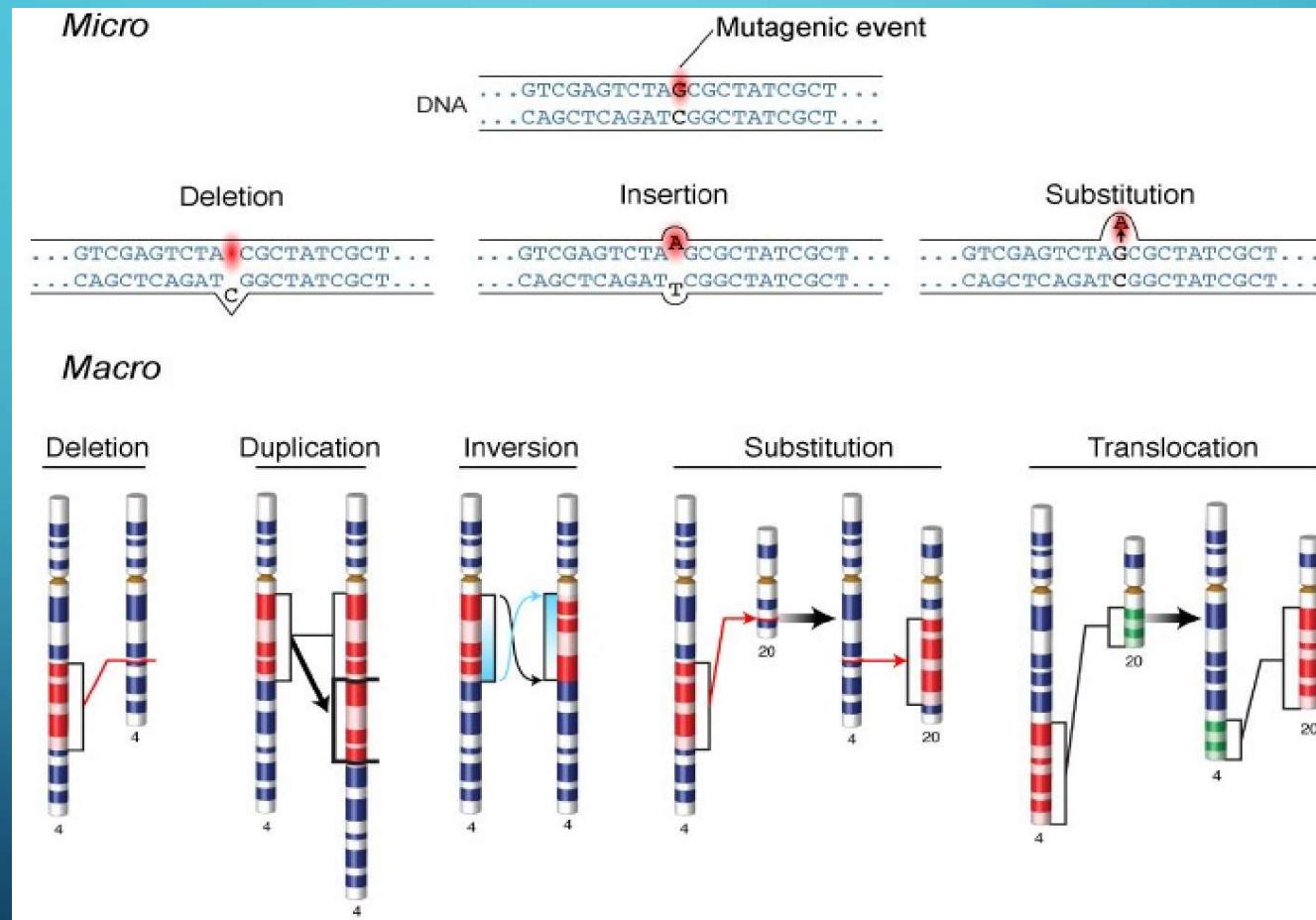
# ALINEAMIENTO DE SECUENCIAS ADN



ALINEACION



# ALINEAMIENTO DE SECUENCIAS ADN





# ALINEAMIENTO DE SECUENCIAS ADN

Ediciones posibles de una secuencia

**Mutations**

AGGCCTC

**Insertions**

AGGACTC

**Deletions**

AGGGCCTC

AGG.CTC



## ALINEAMIENTO DE SECUENCIAS ADN

Dada dos cadenas  $X = x_1 x_2 \dots x_M$ ,  $Y = y_1 y_2 \dots y_N$ ,

una alineación es una asignación de huecos a las posiciones  $0, \dots, N$  en  $X$ , y  $0, \dots, N$  en  $Y$ , para alinear cada letra en una secuencia con **una letra**, o un **gap** en la otra secuencia



## ALINEAMIENTO DE SECUENCIAS ADN

AGGCTATCACCTGACCTCCAGGCCGATGCC  
TAGCTATCACGACCGCGGTGATTGCCCGAC





## ALINEAMIENTO DE SECUENCIAS ADN

AGGCTATCACCTGACCTCCAGGCCGATGCC  
TAGCTATCACGACCGCGGTGATTGCCCGAC

-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---  
TAG-CTATCAC--GACCGC--GGTCGATTGCCCGAC



## QUE SERIA UN BUEN ALINEAMIENTO?

AGGCTAGTT , AGCGAAGTTT

AGGCTAGTT-  
AGCGAAGTTT

6 matches,  
3 mismatches  
1 gap

AGGCTA-GTT-  
AG-CGAAGTTT

7 matches,  
1 mismatch  
3 gaps

AGGC-TA-GTT-  
AG-CG-AAGTTT

7 matches,  
0 mismatch  
5 gaps



## QUE SERIA UN BUEN ALINEAMIENTO?

AGGCTAGTT , AGCGAAGTTT

AGGCTAGTT-  
AGCGAAGTTT

AGGCTA-GTT-  
AG-CGAAGTTT

AGGC-TA-GTT-  
AG-CG-AAGTTT

Match: +m  
Mismatch: -s  
Gap: -d

Score F = (# matches) × m - (# mismatches) × s - (#gaps) × d ???



# METODOS AVANZADOS DE ALINEACION

Para dos secuencias:

- Alineamiento Global
- Alineamiento Local



## Algoritmos con Programacion Dinamica sobre Matrices





## ALINEACION GLOBAL: Algoritmo NEEDLEMAN-WUNSCH

Paso 1: Inicializar

$$F(0, 0) = 0$$

$$F(0, i) = -i * d$$

$$F(j, 0) = -j * d$$

Paso 2: Iterar por las celdas

For each  $i=1 \dots M, j=1 \dots N$

$$F(i,j) = \max \begin{cases} F(i-1, j-1 + s(x_i, y_j)) & \text{caso 1} \\ F(i-1, j) - d & \text{caso 2} \\ F(i, j-1) - d & \text{caso 3} \end{cases}$$

$$P(i,j) = \begin{cases} \text{DIAG} & \text{if caso 1} \\ \text{IZQ} & \text{if caso 2} \\ \text{DER} & \text{if caso 3} \end{cases}$$

**$x_i$  alinea con  $y_i$**   
 **$x_i$  alinea con gap**  
 **$y_i$  alinea con gap**



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$		A	G	T	A
1	A				
2	T				
3	A				



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	-1	-2	-3	-4
1	A				
2	T				
3	A				



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	-1	-2	-3	-4
1	A	-1			
2	T	-2			
3	A	-3			

$s=1$   
 $d=1$

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	-1	-2	-3	-4
$1$	A	-1	?		
$2$	T	-2			
$3$	A	-3			



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

		$i=0$	$1$	$2$	$3$	$4$	
		$j=0$	0	-1	-2	-3	-4
$i=0$	$j=1$	A	G	T	A		
$i=1$	$j=0$	-1	?				
$i=2$	$j=1$	T	-2				
$i=3$	$j=0$	A	-3				

$s=1$   
 $d=1$

$$F(1, 1) = \max \begin{cases} F(0, 0) + 1 & (\text{match}) \\ F(0, 1) - 1 \\ F(1, 0) - 1 \end{cases}$$



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$s=1$   
 $d=1$

$F(i,j)$

		$i=0$	$1$	$2$	$3$	$4$	
		$j=0$	0	-1	-2	-3	-4
$i=0$	$j=1$	A	G	T	A		
$i=1$	$j=0$	-1	?				
$i=1$	$j=1$	A	-1				
$i=2$	$j=0$	T	-2				
$i=2$	$j=1$	A	-3				

$$F(1, 1) = \text{MAX} \begin{cases} 0 + 1 \text{ (match)} \\ -1 - 1 \\ -1 - 1 \end{cases}$$

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

		$i=0$	$1$	$2$	$3$	$4$	
		$j=0$	0	-1	-2	-3	-4
		$j=1$	A	-1	1 (D)		
		2	T	-2			
		3	A	-3			

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	-1	-2	-3	-4
$j=1$	A	-1	1	?	
$j=2$	T	-2			
$j=3$	A	-3			

$$F(2, 1) = \max \begin{cases} F(1, 0) - 1 & (\text{no match}) \\ F(1, 1) - 1 \\ F(2, 0) - 1 \end{cases}$$

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	-1	-2	-3	-4
1	A	-1	1	?	
2	T	-2			
3	A	-3			

$$F(2, 1) = \text{MAX} \begin{bmatrix} -1 & -1 \\ 1 & -1 \\ -2 & 1 \end{bmatrix}$$

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	-1	-2	-3	-4
1	A	-1	1	0 (L)	
2	T	-2			
3	A	-3			

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	-1	-2	-3	-4
1	A	-1	1	0	
2	T	-2			
3	A	-3			



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	-1	-2	-3	-4
1	A	-1	1	0	-1 (L)
2	T	-2			
3	A	-3			



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	-1	-2	-3	-4
1	A	-1	1	0	-1
2	T	-2			
3	A	-3			



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	-1	-2	-3	-4
1	A	-1	1	0	-1
2	T	-2			
3	A	-3			

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2				
3	A	-3				

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0 (U)			
3	A	-3				

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0			
3	A	-3				

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0 <small>(D)</small>		
3	A	-3				

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0		
3	A	-3				

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1 (D)	
3	A	-3				

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	
3	A	-3				

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0 (L)
3	A	-3				

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0
3	A	-3				

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0
3	A	-3	-1 (D)			

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0
3	A	-3	-1			

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0
3	A	-3	-1	-1 (D)		

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0
3	A	-3	-1	-1		

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0
3	A	-3	-1	-1	0 (U)	

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0
3	A	-3	-1	-1	0	

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0
3	A	-3	-1	-1	0	2 (D)



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0
3	A	-3	-1	-1	0	2 (D)

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1 (D)	0
3	A	-3	-1	-1	0	2

A  
A

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	-1	-2	-3	-4
$1$	A	-1	1	0	-1
$2$	T	-2	0	0	1
$3$	A	-3	-1	-1	0

TA  
TA

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	-1	-2	-3	-4
$A$	-1	1 (D)	0	-1	-2
$T$	-2	0	0	1	0
$A$	-3	-1	-1	0	2

GTA  
-TA

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	1	2	3	4	
$j=0$	0	-1	-2	-3	-4	
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0
3	A	-3	-1	-1	0	2

Alineamiento Optimo

**AGTA**  
**A-TA**

Score:  $1+0+1+2 = 4$



## EJERCICIO

T G A C T A A G T

T G C G T A G T



## EJERCICIO

		T	G	A	C	T	A	A	G	T
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
T	-1	1	0	-1	-2	-3	-4	-5	-6	-7
G	-2	0	2	1	0	-1	-2	-3	-4	-5
C	-3	-1	1	1	2	1	0	-1	-2	-3
G	-4	-2	0	0	1	1	0	-1	0	-1
T	-5	-3	-1	-1	0	2	1	0	-1	1
A	-6	-4	-2	0	-1	1	3	2	1	0
G	-7	-5	-3	-1	-1	0	2	2	3	2
T	-8	-6	-4	-2	-2	0	1	1	2	4



## ALINEACION LOCAL: Algoritmo Smith-Waterman

Paso 1: Inicializar

$$F(0, 0) = 0$$

$$F(0, i) = 0$$

$$F(j, 0) = 0$$

Paso 2: Iterar por las celdas

For each  $i=1 \dots M, j=1 \dots N$

$$F(i,j) = \max \begin{cases} 0 \\ F(i-1,j-1+s(x_i,y_j)) & \text{caso 1} \\ F(i-1,j)-d & \text{caso 2} \\ F(i,j-1)-d & \text{caso 3} \end{cases}$$

$P(i,j)$  se construye igual al anterior,  
pero termina en el primer 0 que encuentra



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
1	A	0			
2	T	0			
3	A	0			

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
$1$	A	?			
$2$	T	0			
$3$	A	0			

$$F(1, 1) = \max \begin{cases} 0 \\ F(0, 0) + 1 \text{ (match)} \\ F(0, 1) - 1 \\ F(1, 0) - 1 \end{cases}$$

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
$1$	A	?			
$2$	T	0			
$3$	A	0			

$$F(1, 1) = \text{MAX} \begin{cases} 0 \\ 0 + 1 \text{ (match)} \\ 0 - 1 \\ 0 - 1 \end{cases}$$



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
$1$	A	0	$1_{(D)}$		
$2$	T	0			
$3$	A	0			



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
1	<b>A</b>	0	1		
2	<b>T</b>	0			
3	<b>A</b>	0			



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
1	A	0	1	0 (L)	
2	T	0			
3	A	0			



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
1	A	0	1	0 <sub>(L)</sub>	0
2	T	0			
3	A	0			



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
1	A	0	1	0 <sub>(L)</sub>	0
2	T	0			
3	A	0			



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
1	A	0	1	0 <sub>(L)</sub>	1 <sub>(D)</sub>
2	T	0			
3	A	0			

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
1	A	0	1	0 <sub>(L)</sub>	0
2	T	0			
3	A	0			

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
1	A	0	1	0 <sub>(L)</sub>	0
2	T	0	0 <sub>(U)</sub>		
3	A	0			

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
1	A	0	1	0 <sub>(L)</sub>	0
2	T	0	0		
3	A	0			

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
1	A	0	1	0 <sub>(L)</sub>	0
2	T	0	0		
3	A	0			

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	0	0	0	0	
1	A	0	1	0 <sub>(L)</sub>	0	1 <sub>(D)</sub>
2	T	0	0	0 <sub>(D)</sub>		
3	A	0				

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	0	0	0	0	
1	A	0	1	0 <sub>(L)</sub>	0	1 <sub>(D)</sub>
2	T	0	0	0 <sub>(D)</sub>		
3	A	0				

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	0	0	0	0	
1	A	0	1	0 <sub>(L)</sub>	0	1 <sub>(D)</sub>
2	T	0	0	0 <sub>(D)</sub>	1 <sub>(D)</sub>	
3	A	0				



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	0	0	0	0	
1	A	0	1	0 <sub>(L)</sub>	0	1 <sub>(D)</sub>
2	T	0	0	0 <sub>(D)</sub>	1 <sub>(D)</sub>	0 <sub>(L,U)</sub>
3	A	0				

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	0	0	0	0	
1	A	0	1	0 <sub>(L)</sub>	0	1 <sub>(D)</sub>
2	T	0	0	0 <sub>(D)</sub>	1 <sub>(D)</sub>	0 <sub>(L,U)</sub>
3	A	0				

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	0	0	0	0	
1	A	0	1	0 <sub>(L)</sub>	0	1 <sub>(D)</sub>
2	T	0	0	0 <sub>(D)</sub>	1 <sub>(D)</sub>	0 <sub>(L,U)</sub>
3	A	0	1 <sub>(D)</sub>			

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	0	0	0	0	
1	A	0	1	0 <sub>(L)</sub>	0	1 <sub>(D)</sub>
2	T	0	0	0 <sub>(D)</sub>	1 <sub>(D)</sub>	0 <sub>(L,U)</sub>
3	A	0	1 <sub>(D)</sub>			

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	0	0	0	0	
1	A	0	1	0 <sub>(L)</sub>	0	1 <sub>(D)</sub>
2	T	0	0	0 <sub>(D)</sub>	1 <sub>(D)</sub>	0 <sub>(L,U)</sub>
3	A	0	1 <sub>(D)</sub>	0 <sub>(L)</sub>		

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	0	0	0	0	
1	A	0	1	0 <sub>(L)</sub>	0	1 <sub>(D)</sub>
2	T	0	0	0 <sub>(D)</sub>	1 <sub>(D)</sub>	0 <sub>(L,U)</sub>
3	A	0	1 <sub>(D)</sub>	0 <sub>(L)</sub>		

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	0	0	0	0	
1	A	0	1	0 <sub>(L)</sub>	0	1 <sub>(D)</sub>
2	T	0	0	0 <sub>(D)</sub>	1 <sub>(D)</sub>	0 <sub>(L,U)</sub>
3	A	0	1 <sub>(D)</sub>	0 <sub>(L)</sub>	0 <sub>(U)</sub>	

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	0	0	0	0	
1	A	0	1	0 <sub>(L)</sub>	0	1 <sub>(D)</sub>
2	T	0	0	0 <sub>(D)</sub>	1 <sub>(D)</sub>	0 <sub>(L,U)</sub>
3	A	0	1 <sub>(D)</sub>	0 <sub>(L)</sub>	0 <sub>(U)</sub>	



Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

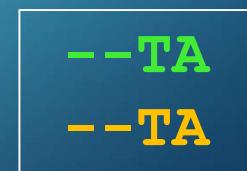
	$i=0$	$1$	$2$	$3$	$4$	
$j=0$	0	0	0	0	0	
1	A	0	1	0 <sub>(L)</sub>	0	1 <sub>(D)</sub>
2	T	0	0	0 <sub>(D)</sub>	1 <sub>(D)</sub>	0 <sub>(L,U)</sub>
3	A	0	1 <sub>(D)</sub>	0 <sub>(L)</sub>	0 <sub>(U)</sub>	2 <sub>(D)</sub>

Ejemplo:  $x = \text{AGTA}$ ,  $y = \text{ATA}$

$F(i,j)$

	$i=0$	$1$	$2$	$3$	$4$
$j=0$	0	0	0	0	0
1	A	0	1	0 (L)	0 (D)
2	T	0	0	0 (D)	1 (D, U)
3	A	0	1 (D)	0 (L)	0 (U)

Alineamiento Optimo



Score:  $0+1+2 = 3$



## EJERCICIO

T G A C T A A G T

T G C G T A G T

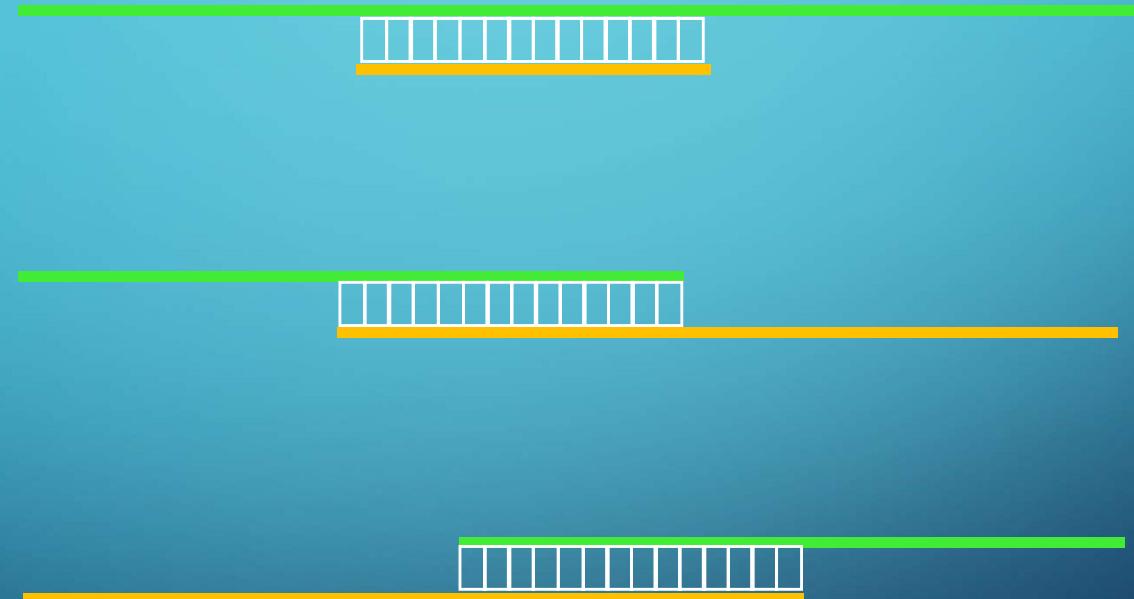


## EJERCICIO

		T	G	A	C	T	A	A	G	T
	0	0	0	0	0	0	0	0	0	0
T	0	1	0	0	0	1	0	0	0	1
G	0	0	2	1	0	0	0	0	1	0
C	0	0	1	1	2	1	0	0	0	0
G	0	0	1	0	1	1	0	0	1	0
T	0	1	0	0	0	2	1	0	0	2
A	0	0	0	1	0	1	3	2	1	1
G	0	0	1	0	0	0	2	2	3	2
T	0	1	0	0	0	1	1	1	2	4



## DIFERENTES TIPOS DE ALINEAMIENTO

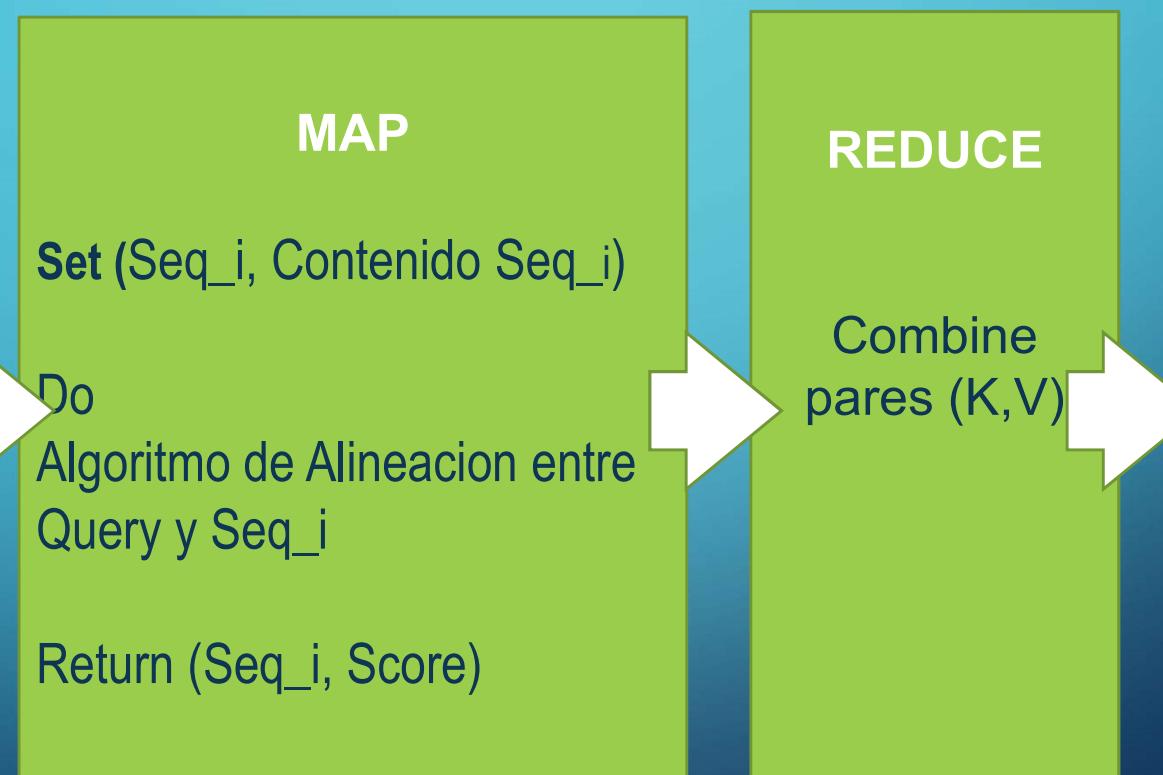


# DIFERENTES TIPOS DE ALINEAMIENTO

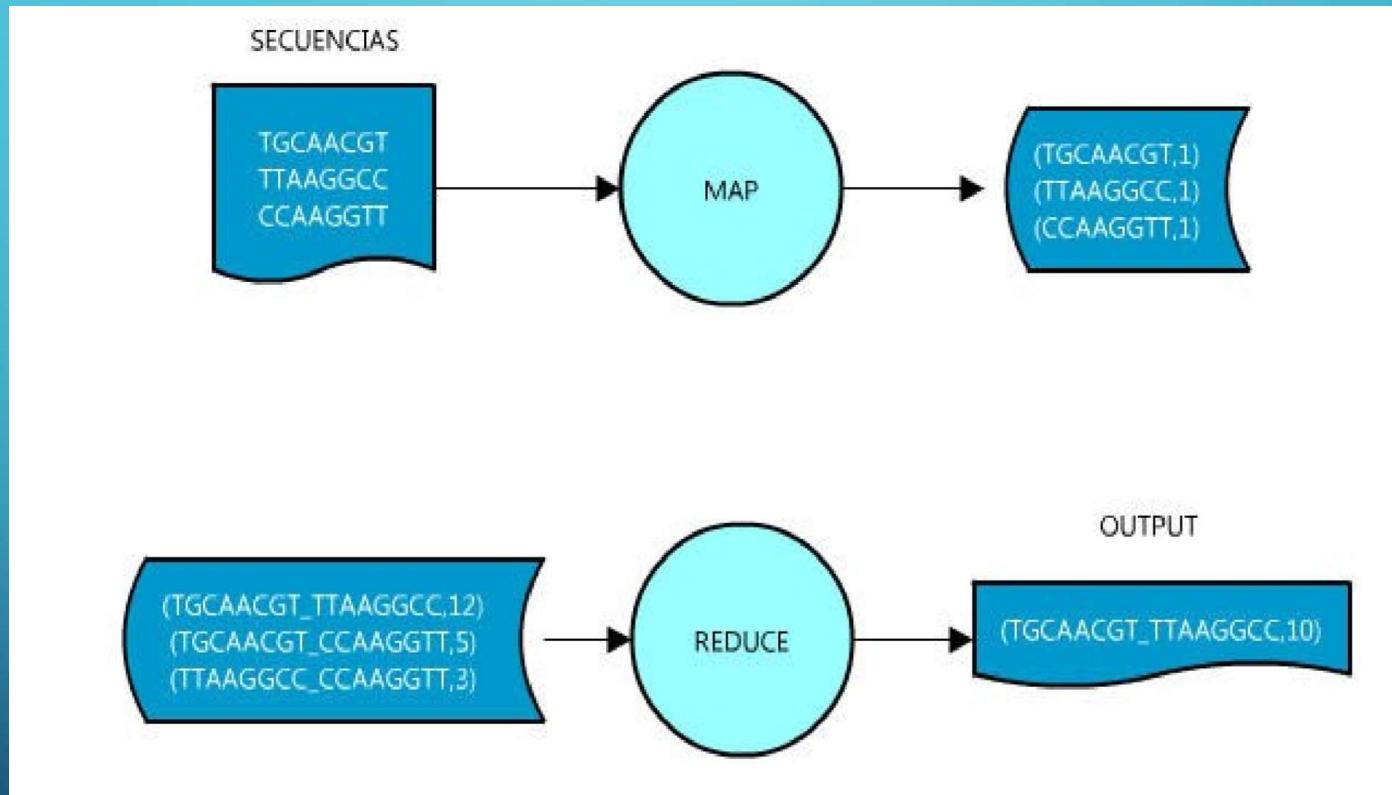
Un ejemplo seria, no penalizar los gaps de los extremos

# ALINEACION DE ADN USANDO HADOOP

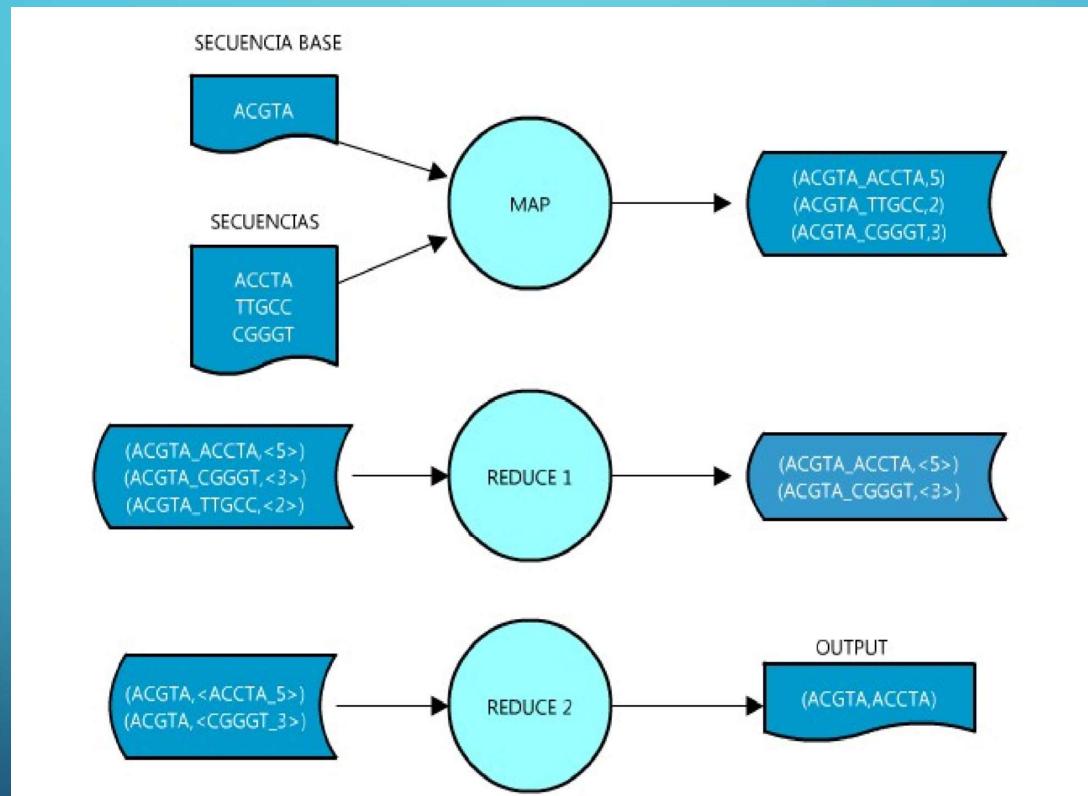
Q  
Seq\_1  
Seq\_2  
....  
Seq\_n



# ALINEACION DE ADN USANDO HADOOP



# ALINEACION DE ADN USANDO HADOOP



## OTRAS APLICACIONES



*Abraham Wald: Blindemos el resto!*



# **Modulo “Analisis de Datos Cientificos y Geograficos”**

## **ANALISIS DE MUSICA**

(Ver paper completo en Campus→Lecturas)



## BIG DATA EN MUSICA

### The Evolution of Popular Music: USA1960–2010

Matthias Mauch, Robert MacCallum, Mark Levy, Armand Leroi

El equipo analizó las propiedades musicales en segmentos de 30s de **17094** canciones del **US Billboard Hot 100** en el periodo **1960-2010**, observando los patrones de cambios de acordes y tonos.



## BIG DATA EN MUSICA

Características cuantitativas de audio:

- 12 descriptores de contenido tonal
- 14 descriptores de timbre



**Discretizados**



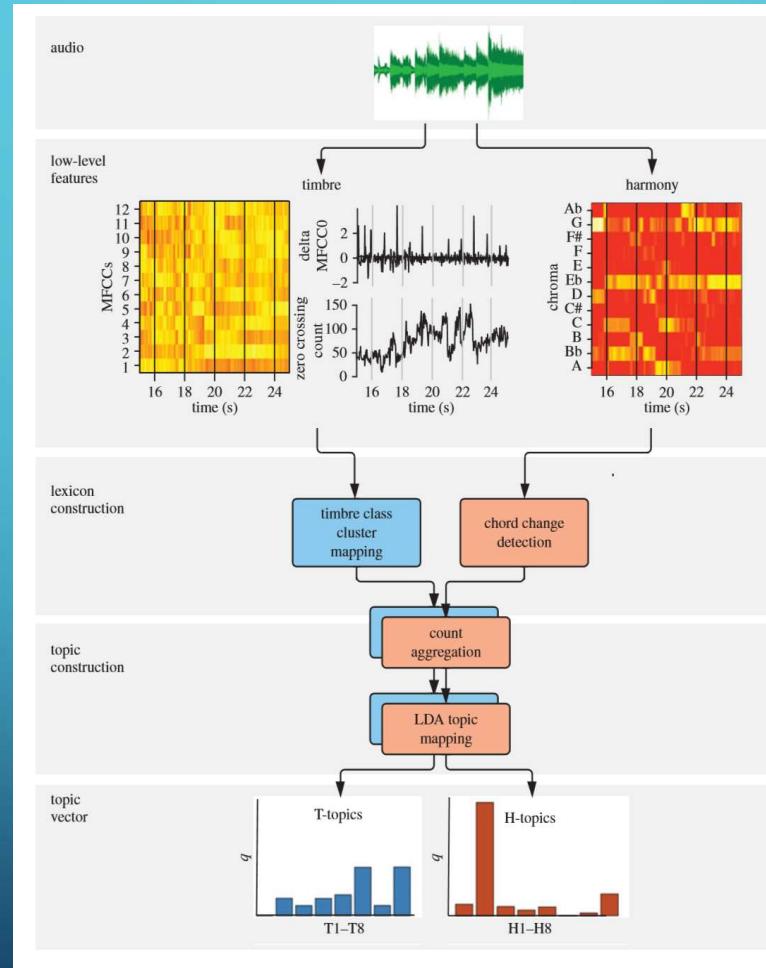
**H: léxico armónico de los cambios de acordes**

**T: léxico timbral de los grupos de timbre**

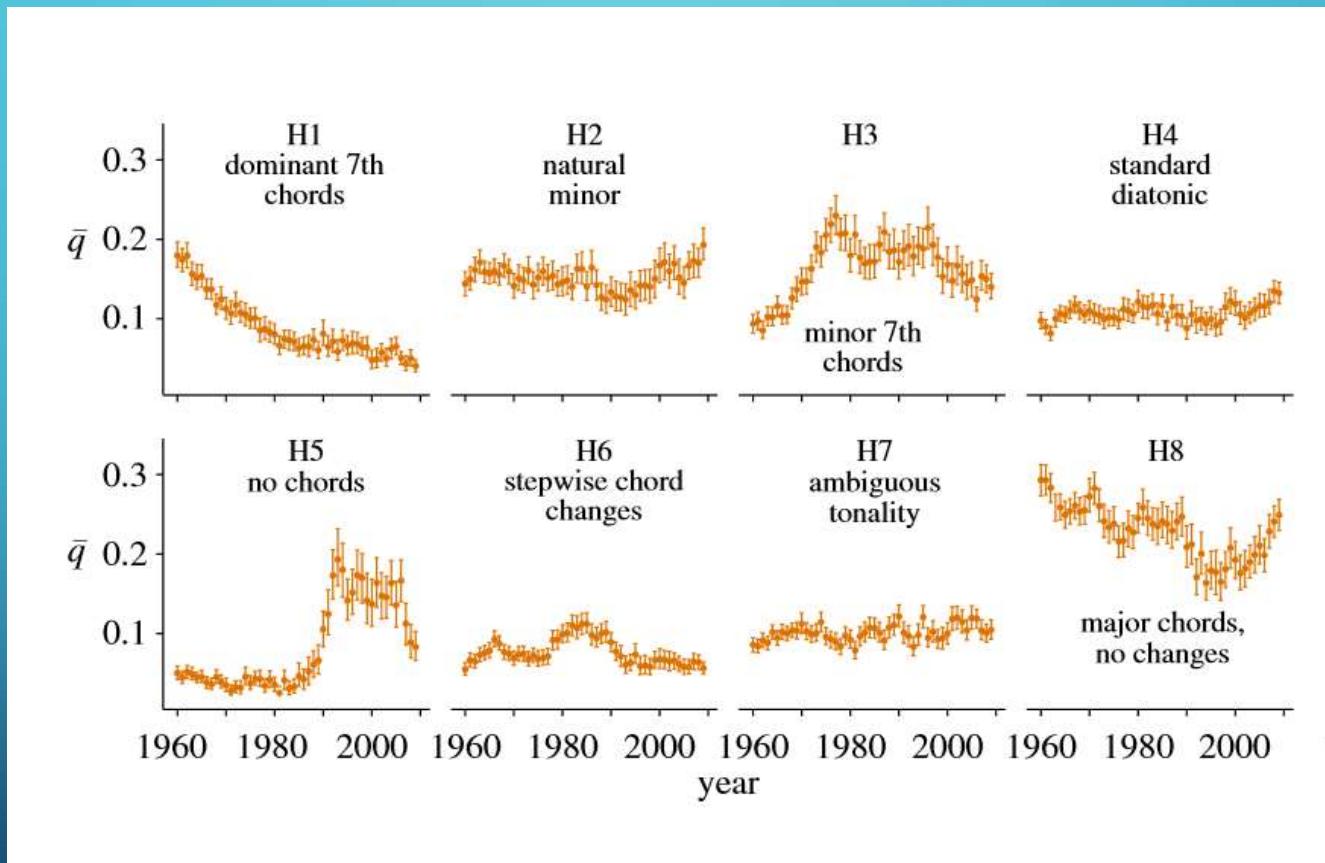


*etiquetas semánticas de género  
(asignadas por los oyentes de Last.fm  
(web service con 50 millones de usuarios !)*

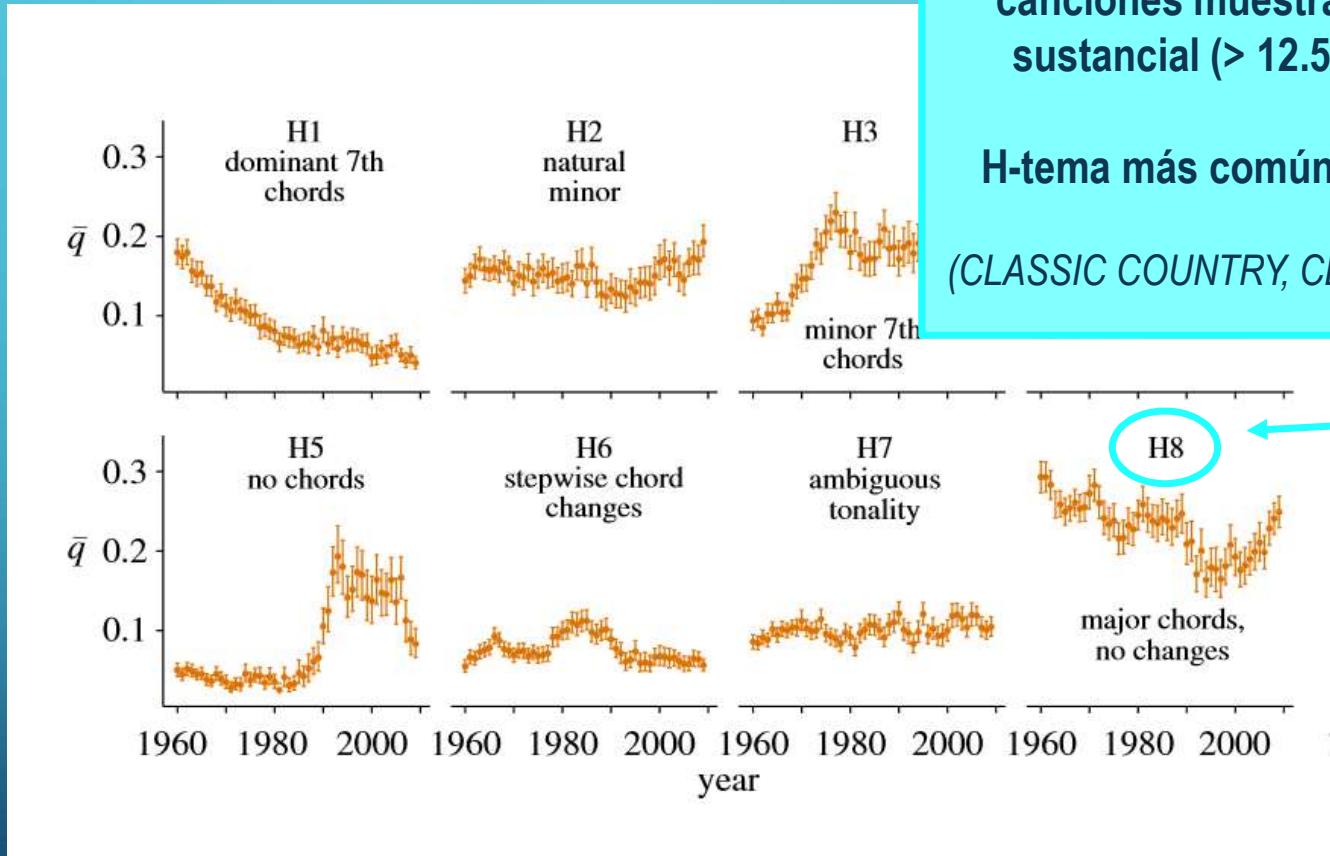
# BIG DATA EN MUSICA



# BIG DATA EN MUSICA



# BIG DATA EN MUSICA

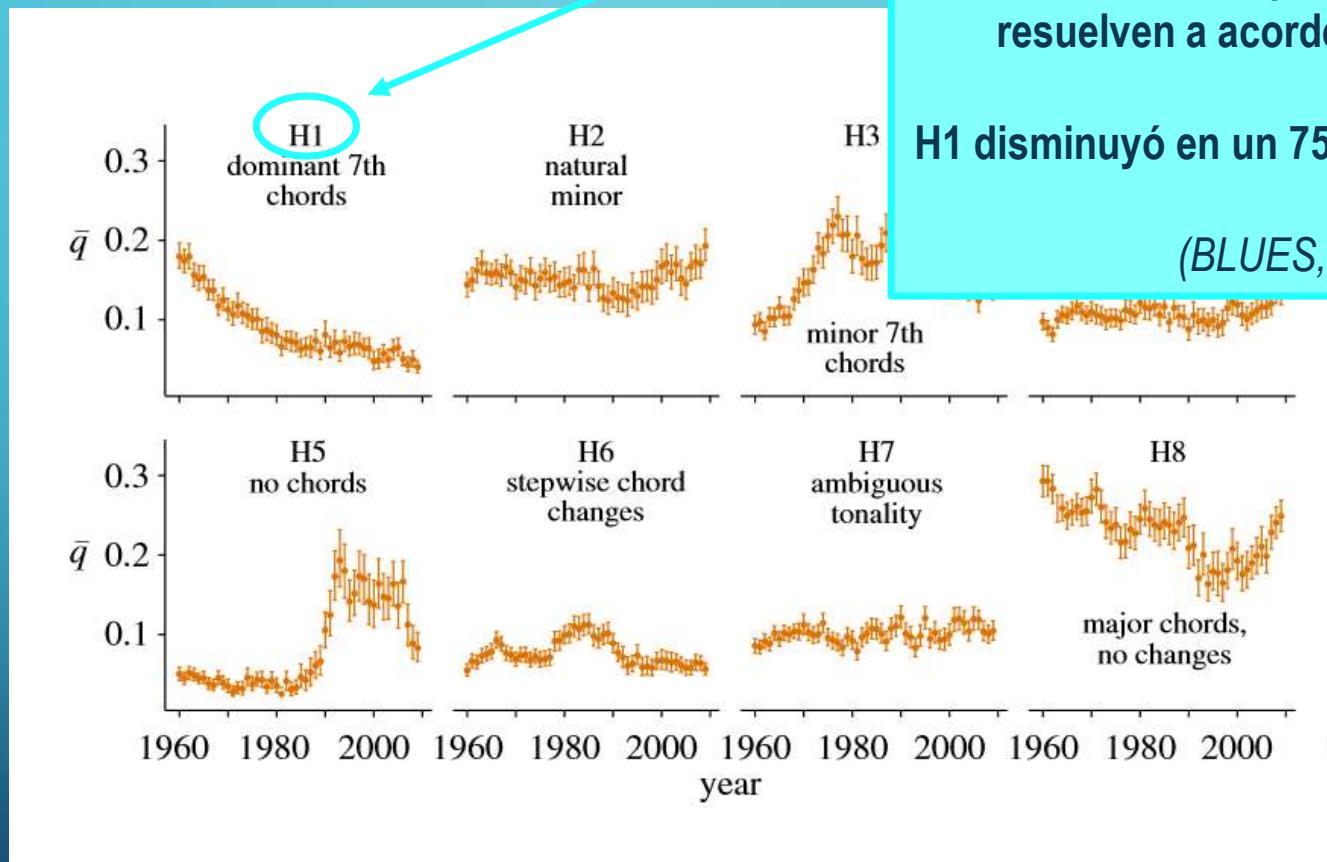


H-tema más frecuente: el 66% de las canciones muestran una frecuencia sustancial ( $> 12.5\%$ ) de este tema.

H-tema más común en 43 de 50 años.

(CLASSIC COUNTRY, CLASSIC ROCK y LOVE)

# BIG DATA EN MUSICA



H-tema inherentemente disonante, para crear tensiones que eventualmente se resuelven a acordes consonantes.

H1 disminuyó en un 75% entre 1960 y 2009.

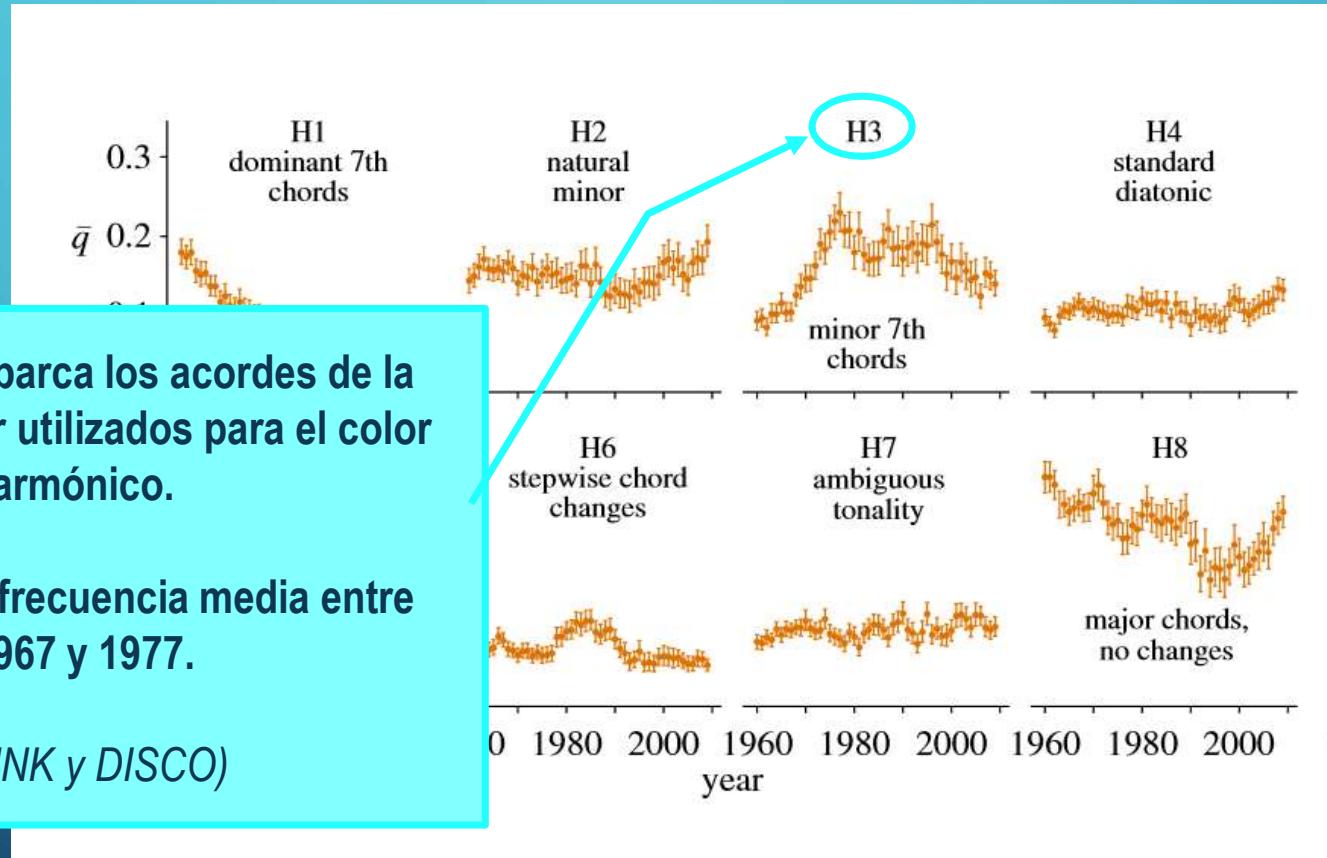
(BLUES, JAZZ)

# BIG DATA EN MUSICA

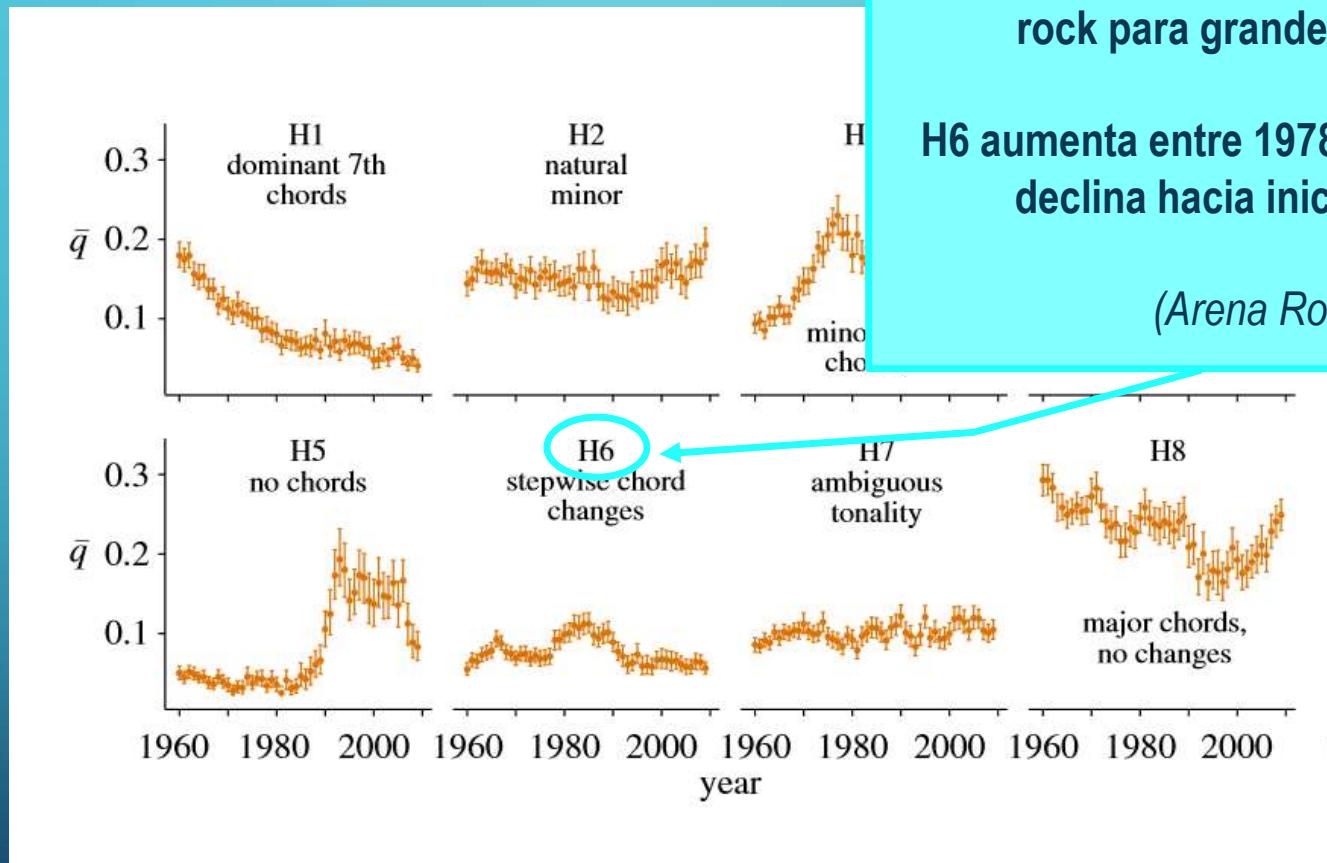
H-tema que abarca los acordes de la séptima menor utilizados para el color armónico.

H3 duplica la frecuencia media entre 1967 y 1977.

(FUNK y DISCO)



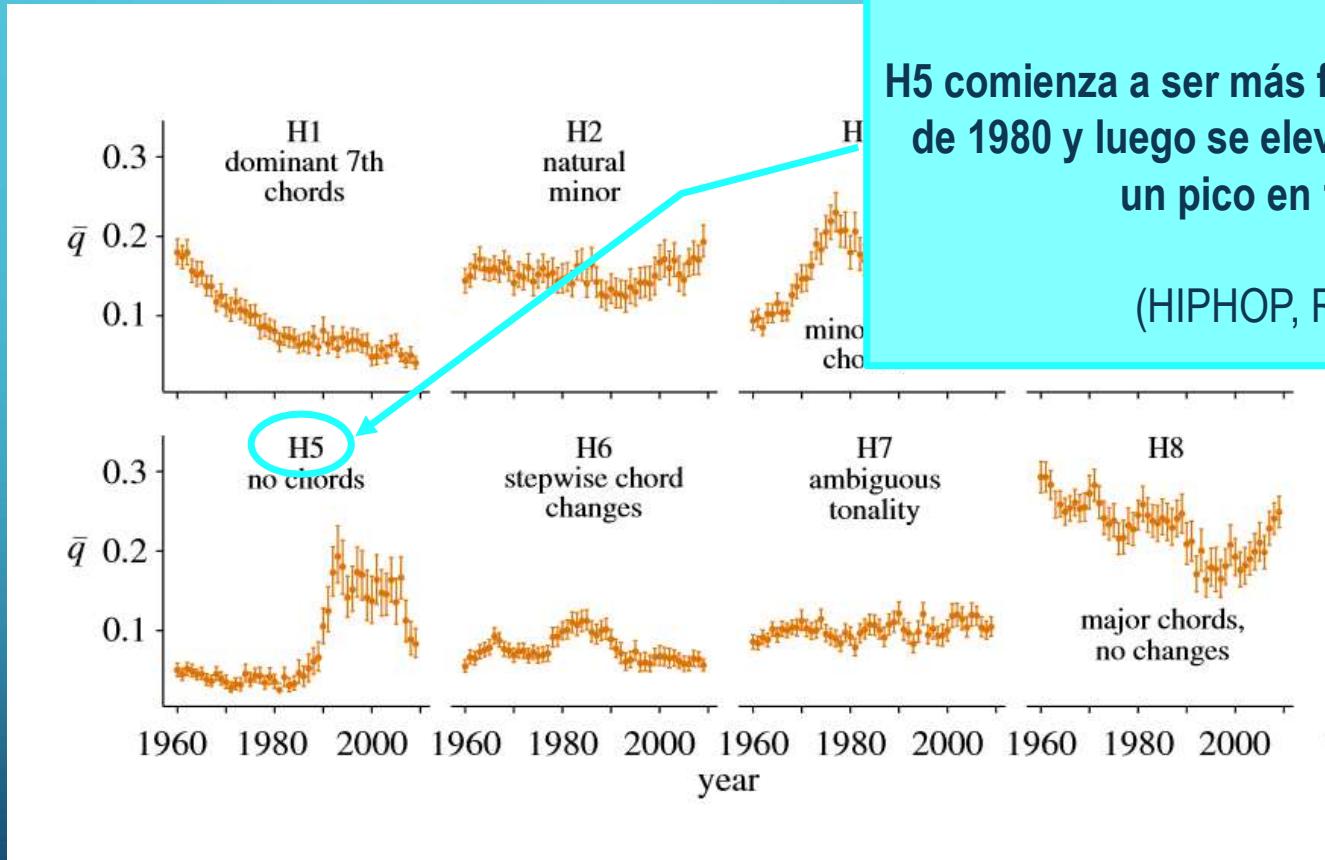
# BIG DATA EN MUSICA



H-tema que combina varios cambios de acordes que son un pilar en melodías de rock para grandes estadios.

H6 aumenta entre 1978 y 1985, y luego declina hacia inicio de 1990.

# BIG DATA EN MUSICA

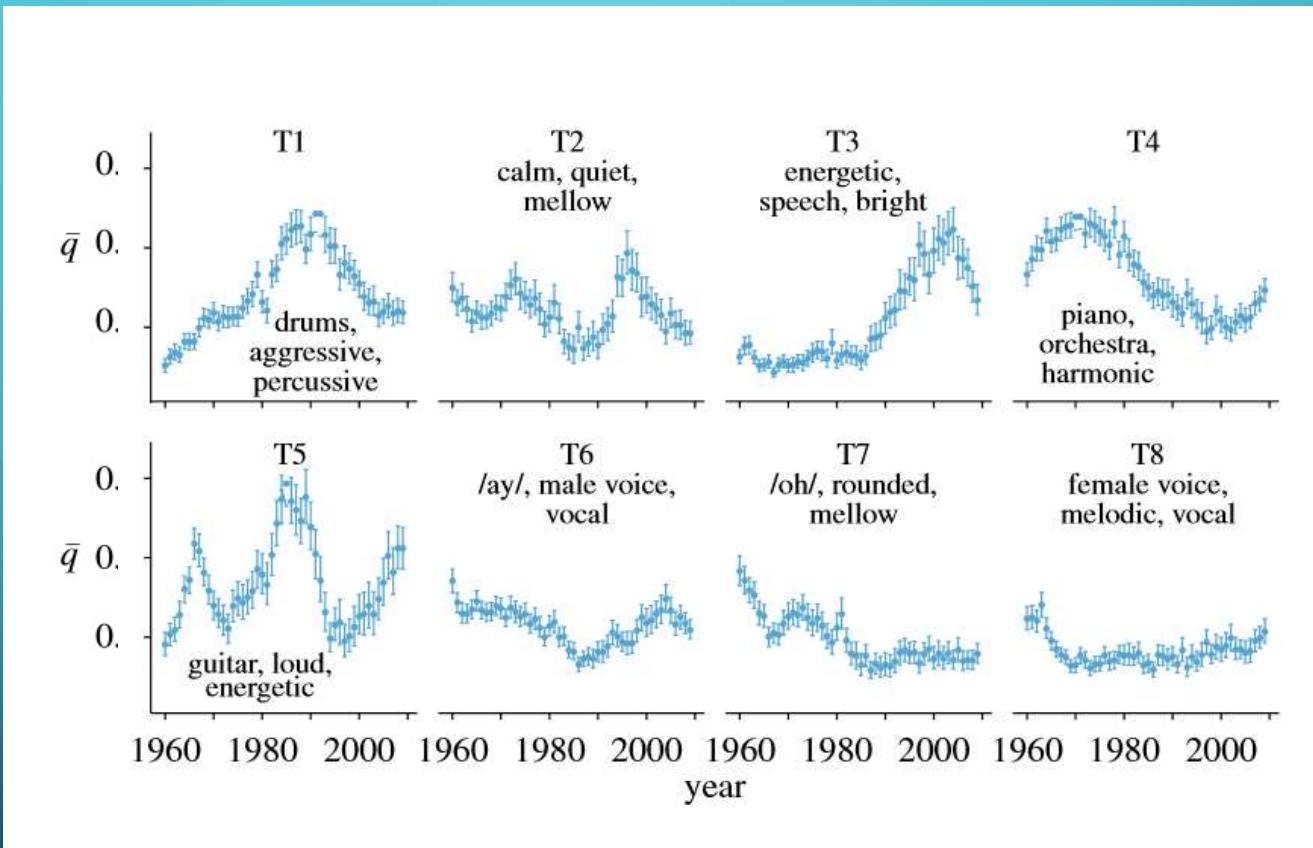


H-tema que representa poca melodía, tonos monótonos conversados.

H5 comienza a ser más frecuente a finales de 1980 y luego se eleva rápidamente a un pico en 1993.

(HIPHOP, RAP)

# BIG DATA EN MUSICA

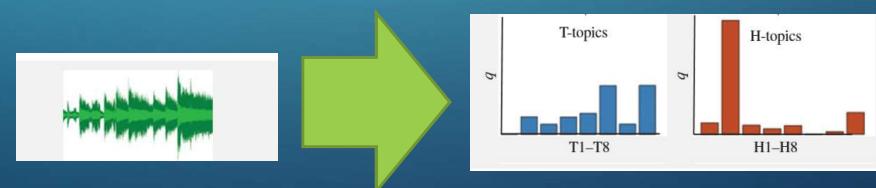


## BIG DATA EN MUSICA

**8 H + 8 T = 16 tópicos** con LDA (Latent Dirichlet Allocation)

Cada canción se representa como una distribución sobre una serie de s, y cada tema se representa como una distribución sobre todas las palabras posibles (H-lexicon, T-lexicon) y se obtiene el modelo más probable mediante inferencia probabilística.

En consecuencia, cada canción se representa como una distribución sobre ocho temas armónicos (H-temas) que captura las clases de cambios de acordes (ej: "tonos violentos, percusivos", "voz femenina, melódica, vocal", derivada de las anotaciones de expertos), con una proporción  $q$  sobre los tópicos.

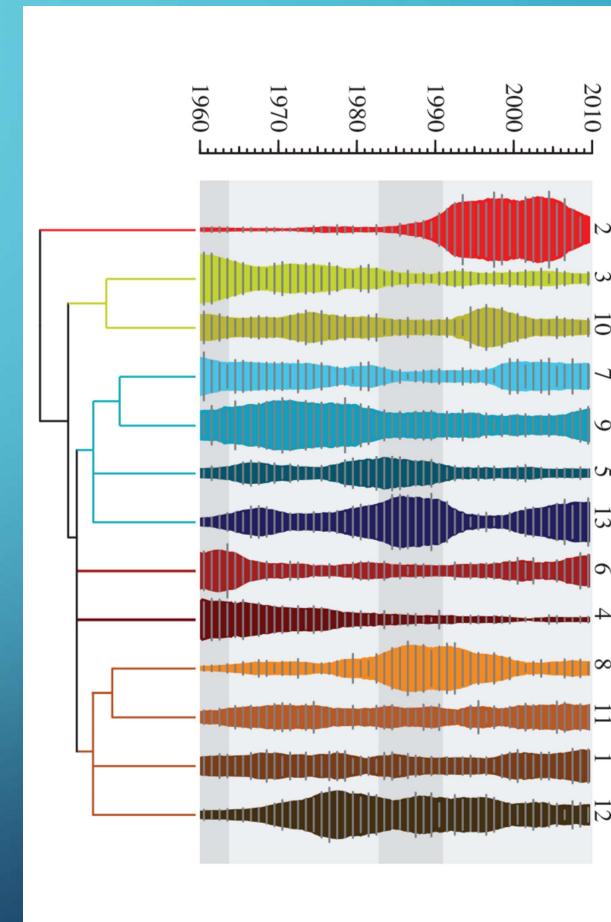


## BIG DATA EN MUSICA

Aplicaron K-means sobre componentes principales de frecuencias.

Investigaron los  $k < 25$  y la mejor solución de agrupamiento fue  $k = 13$ .

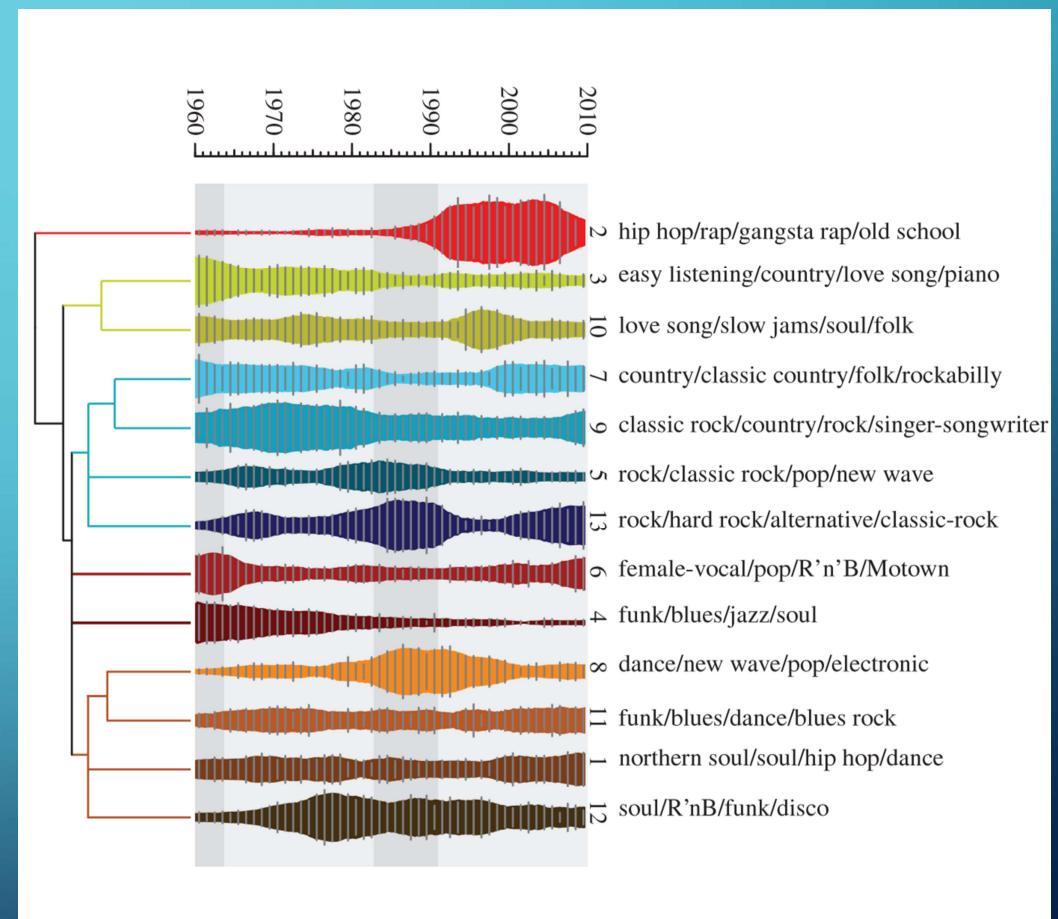
El ancho de cada eje es proporcional a la frecuencia de ese estilo, normalizado a cada año.



## BIG DATA EN MUSICA

Para relacionar las etiquetas de Last.fm con los clústeres de estilos, usaron la técnica **GeneMerge** de Bioinformatica !

Encontraon que para cada estilo algunas etiquetas Last.fm están significativamente sobre representadas...

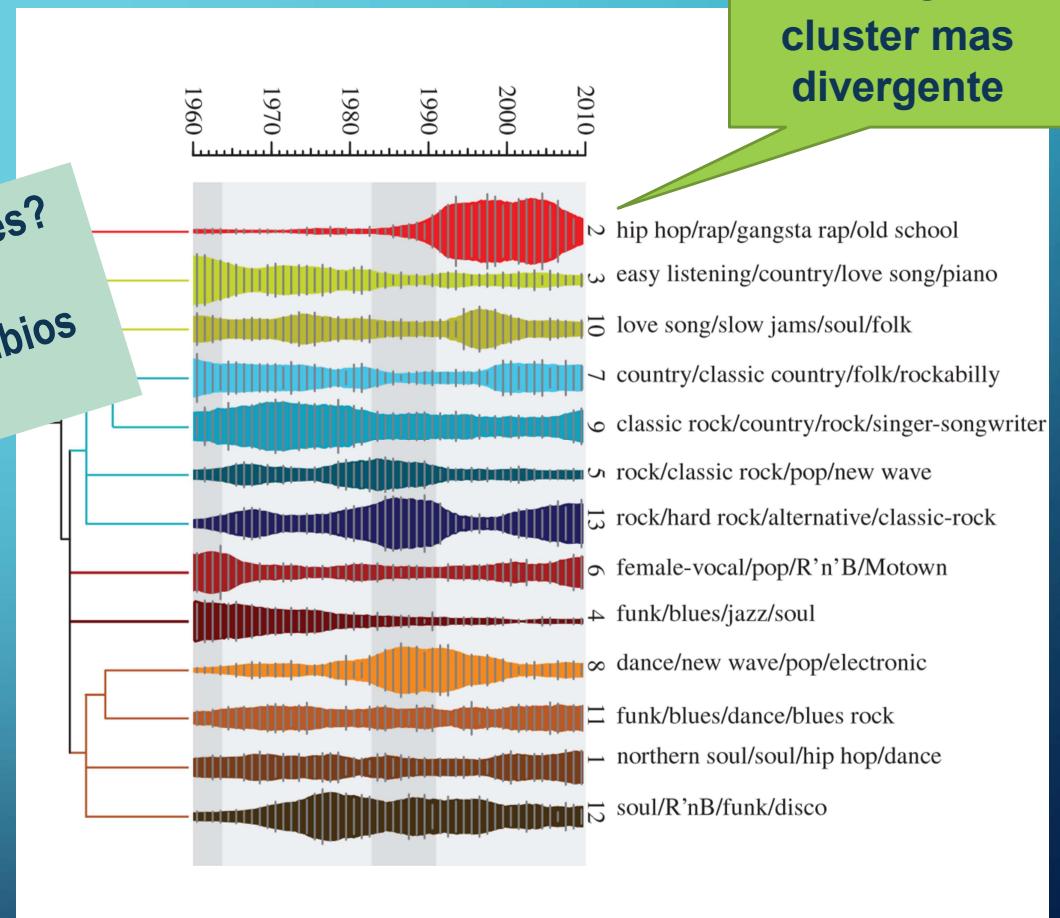


# BIG DATA EN MUSICA

Para relacionar las etiquetas de Last.fm con los clústeres de usuarios...

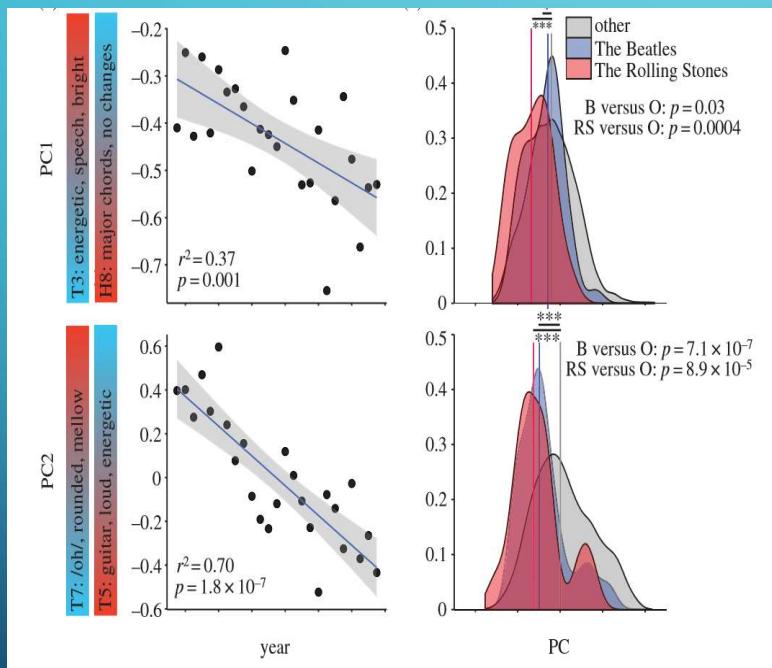
Revela los orígenes de los estilos musicales?  
NO, pero muestra tendencias en los cambios  
en la frecuencia del estilo

...se entraon que para cada estilo algunas etiquetas Last.fm están significativamente sobre representadas...



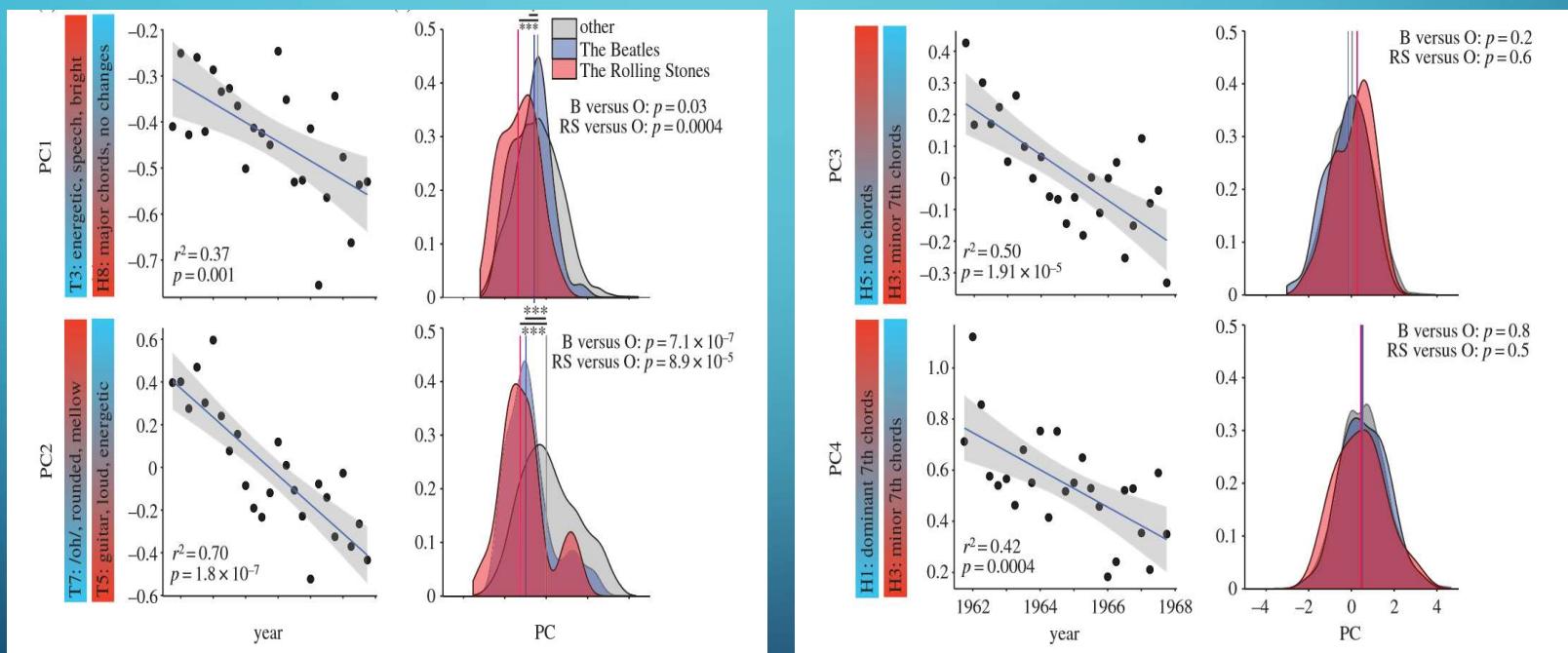
# BIG DATA EN MUSICA

La revolución musical en USA por la Invasion Britanica 1964.  
Analisis de componentes principales T y H para “Beatles” y “Rolling Stones”.



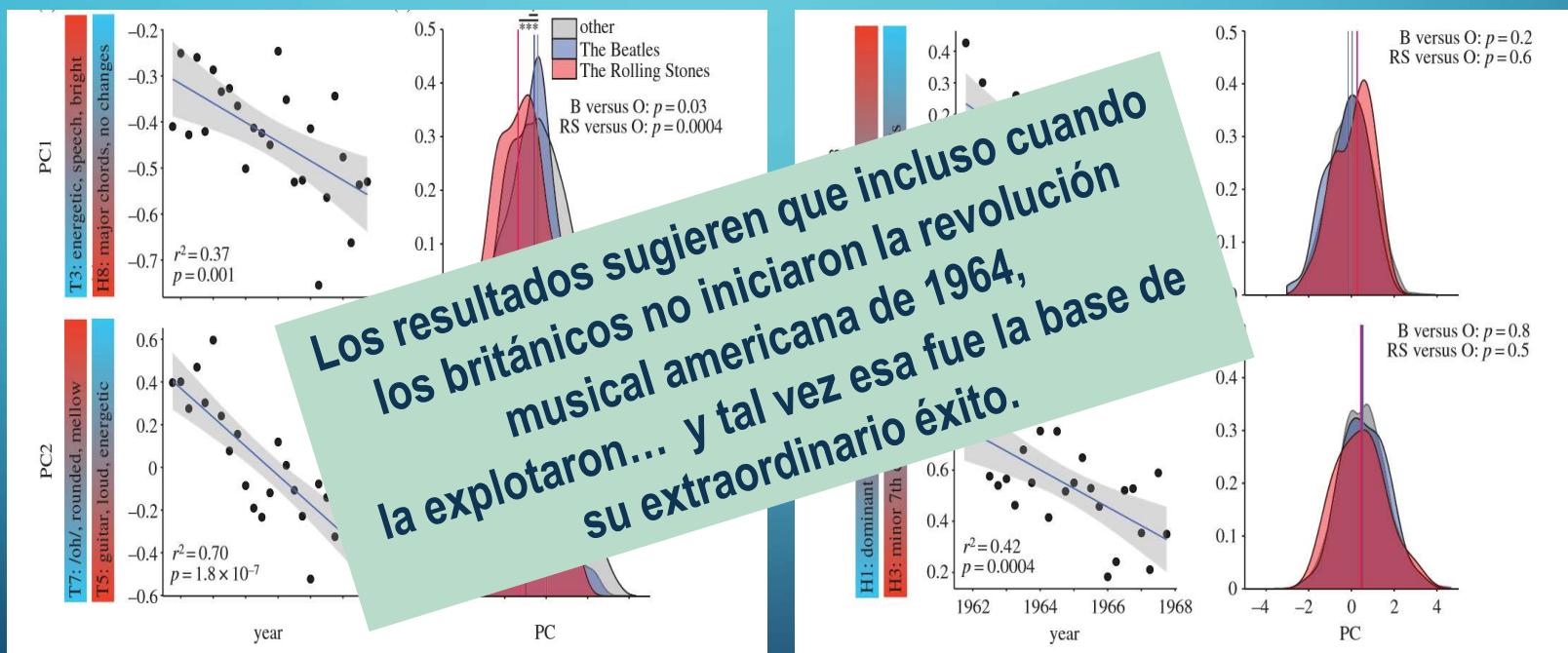
# BIG DATA EN MUSICA

La revolución musical en USA por la Invasion Britanica 1964.  
Analisis de componentes principales T y H para “Beatles” y “Rolling Stones”.



# BIG DATA EN MUSICA

La revolución musical en USA por la Invasion Britanica1964.  
Analisis de componentes principales T y H para “Beatles” y “Rolling Stones”.





# **Modulo “Analisis de Datos Cientificos y Geograficos”**

## **PREDICCION ELECTORAL**

(Ver paper completo en Campus→Lecturas)



## BIG DATA POLITICA

Puede la audiencia televisiva predecir los resultados de las elecciones presidenciales?

Arash Barfar, Balaji Padmanabhan

*En este trabajo ponen de relieve la posibilidad de pronosticar los resultados de las elecciones presidenciales de USA a nivel estatal y de condado basados únicamente en los datos sobre la audiencia de programas de televisión.*



## BIG DATA POLITICA

**Clave:** Dada la naturaleza poco frecuente de las elecciones, tales modelos son útiles sólo si pueden ser entrenados usando datos recientes.

**Problema:** La variable objetivo (resultado de la elección) generalmente no se conoce hasta que se decide la elección.



**Solución:** Los modelos pueden ser entrenados con los datos de los telespectadores en los estados "seguros" para predecir potencialmente los resultados en los estados oscilantes.

## BIG DATA POLITICA

**Clave:** Dada la naturaleza poco frecuente de las elecciones, tales modelos son útiles sólo si pueden ser aplicados a elecciones recientes.

**Problema:** Los datos históricos no se conocen bien.

Extra: estos modelos también podrían ayudar a las campañas a orientar los programas de publicidad (en 2012 se gastaron en publicidad televisiva de la campaña presidencial u\$s 2.000.000)

**Solución:** Los algoritmos de aprendizaje de máquina pueden ser entrenados con los datos de los telespectadores en los estados "seguros" para predecir potencialmente los resultados en los estados oscilantes.



## BIG DATA POLITICA

En este trabajo ponemos de relieve la posibilidad de pronosticar los resultados de las elecciones presidenciales de USA a nivel estatal y de condado basados únicamente en los datos sobre la audiencia de programas de televisión.

Objetivo: utilizar los datos televisivos para predecir los resultados de las elecciones presidenciales en dos niveles diferentes de divisiones políticas:

- 49 estados
- 165 condados



# BIG DATA POLITICA

## Primer descriptor: MPV

Minutos por votante: el tiempo que un votante del padrón de un estado o condado pasó mirando el programa durante toda la fase “caliente” T.

$$MPV(P,C) = \left( \sum_{v \in V_C} \sum_{D \in T} time(v, P, D) \right) / |V_C| ; time(v, P, D) \geq 2$$

- $V_C$  : conjunto de electorado dentro del grupo C
- $time(v, P, D)$  : minutos que el votante v ha pasado mirando el programa P el día D



# BIG DATA POLITICA

## Segundo descriptor: POF

Porcentaje de " fans " del programa dentro de un condado (o estado).

$$\text{POF } (P, C) = (|F_{C,P}| / |V_c|) * 100$$

- $F_{C,P}$  : conjunto fans del programa P en el condado C



**Como detectar el “fan”  
de un programa P ?**



## BIG DATA POLITICA

En la fase caliente, un votante del padrón se define aquí como **fan de un programa** si al menos una de las siguientes condiciones se cumple:

- 1) Ha visto el programa al menos cuatro veces, cada vez durante al menos 10 minutos
- 2) Ha visto el programa por lo menos tres veces, cada vez durante al menos 20 minutos
- 3) Ha visto el programa al menos dos veces, cada vez durante al menos 40 minutos
- 4) Ha pasado al menos 90 minutos en total viendo el programa

# BIG DATA POLITICA

Experimento real:

T = 36 días

49 estados

165 condados

547 programas

138.000 transmisiones

500.000 minutos de audiencias

547 Programs with two watch measures						
Political division (state/county)	2012 Election results	Show 1, MPV	Show 1, POF	...	Show 547, MPV	Show 547, POF



## BIG DATA POLITICA

Se utilizó validación cruzada **leave-one-out** como método de estimación del error:

A nivel estatal se usan 48 puntos de datos, se construye un modelo y luego se prueba en el punto restante y se registra el error.

Se repite 49 veces (dejando cada estado fuera una vez) y los errores se promedian para reportar el error total.

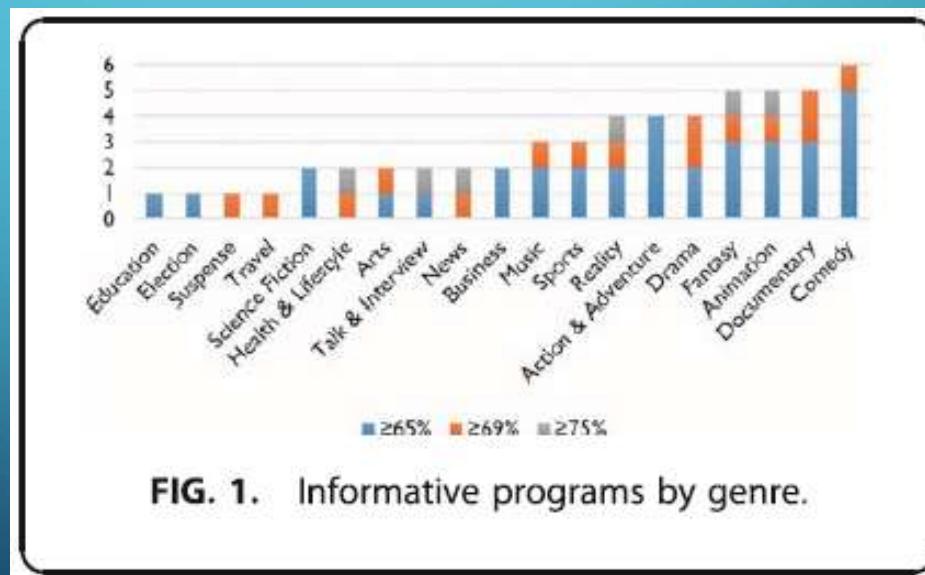
A nivel de condado, se realiza diez veces la validación cruzada.

A nivel estatal, la precisión de referencia es del 53% (26/49). Por lo tanto, un modelo que simplemente predice la clase mayoritaria será correcto 53% de las veces. Del mismo modo, a nivel de condado la línea de base se fijó en 66% (110/165).

# BIG DATA POLITICA

Se utilizaron programas presentes en todos los estados....

Se construye un árbol de decisión sobre las elecciones pasadas para cada programa: 547 árboles por estado o condado.

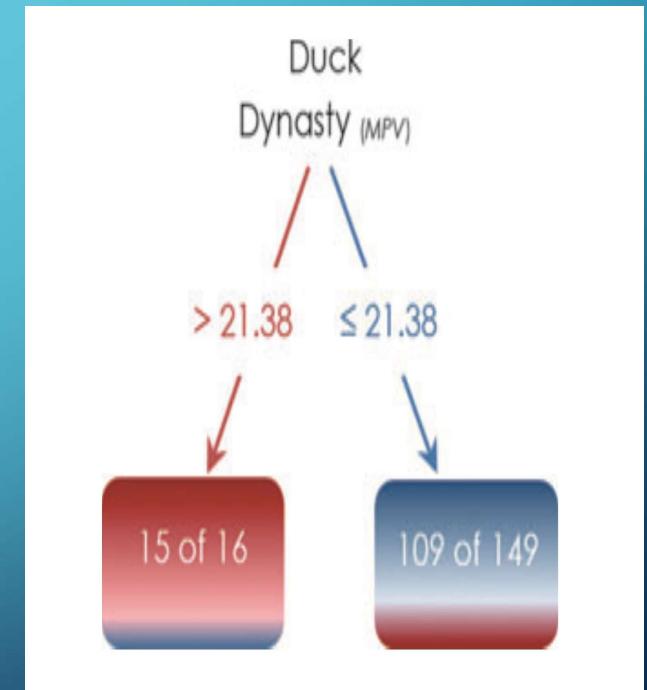


# BIG DATA POLITICA

Algunos resultados:

Duck Dynasty, que es un programa predictivo, tanto a nivel de condado como de estado.

El nivel de condado de Duck Dynasty es del 75%. Las probabilidades de votar por los republicanos aumentan a medida que el electorado dentro de la pestaña observa más ese programa.

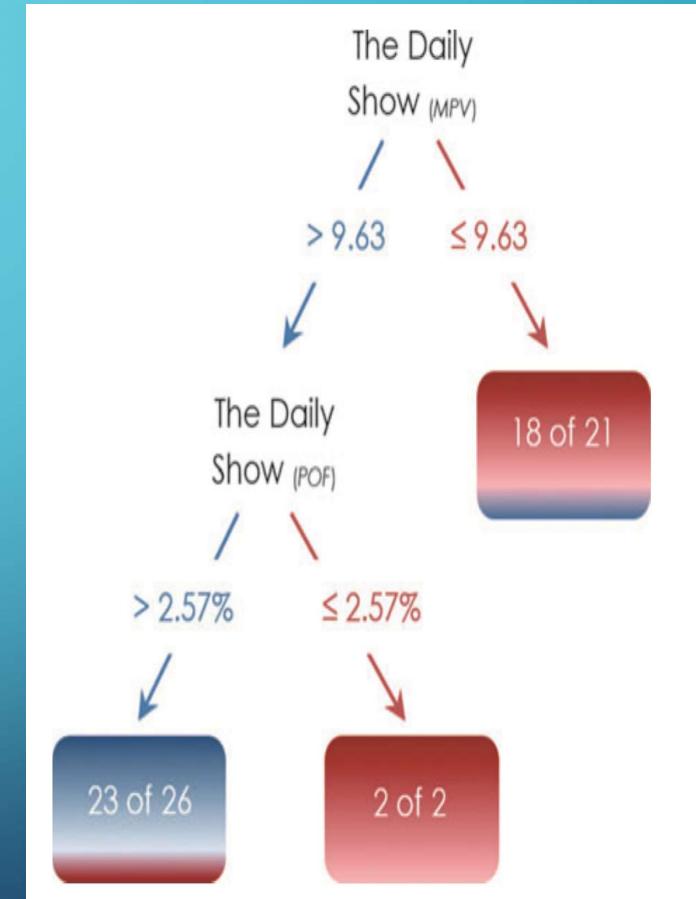


# BIG DATA POLITICA

Algunos resultados:

The Daily Show con Jon Stewart está entre los espectáculos cuyos árboles separan a los estados en Demócratas vs Republicanos.

La precisión predictiva del show es del 82% a nivel estatal.





# BIG DATA POLITICA

Algunos resultados:

Tossup state	Daily Show (MPV)	Daily Show (POF)	Daily Show prediction	2012 Election
Colorado	14.15	5.74%	D	D
Florida	11.38	3.78%	D	D
Iowa	14.89	3.27%	D	D
Nevada	6.04	1.87%	R	D
New Hampshire	11.37	3.08%	D	D
North Carolina	9.12	2.91%	R	R
Ohio	14.75	4.39%	D	D
Pennsylvania	11.93	3.99%	D	D
Virginia	14.27	6.76%	D	D
Wisconsin	9.33	3.68%	R	D



# **Modulo “Analisis de Datos Cientificos y Geograficos”**

## **PREDICCION DE DEMORA EN VUELOS**

**(Ver paper completo en Campus→Lecturas)**



# PREDICCION DE DEMORAS EN VUELOS

**Using Scalable Data Mining for Predicting Flight Delays**

Loris Belcastro, Fabrizio Marozzo, Domenico Talia, And Paolo Trunfio

Los retrasos en los vuelos son frecuentes en todo el mundo: aproximadamente el 20% de los vuelos aéreos llegan con más de 15 minutos de demora, con una consecuencia de un costo anual de miles de millones de dólares.



## PREDICCION DE DEMORAS EN VUELOS

El objetivo principal de este trabajo es implementar un predictor del retraso de llegada de un vuelo programado debido a las condiciones climáticas.

La demora de llegada prevista toma en consideración la información de vuelo (aeropuerto de origen, aeropuerto de destino, salida y llegada programadas) y las condiciones climáticas en el aeropuerto de origen y el aeropuerto de destino de acuerdo con el horario de vuelo.

# PREDICCION DE DEMORAS EN VUELOS

## Modelado

aeropuerto de origen

aeropuerto de destino

Vuelo **F** = <Ao, Ad, tsd, tad, tsa, taa>

horario de salida programado

hora de salida real

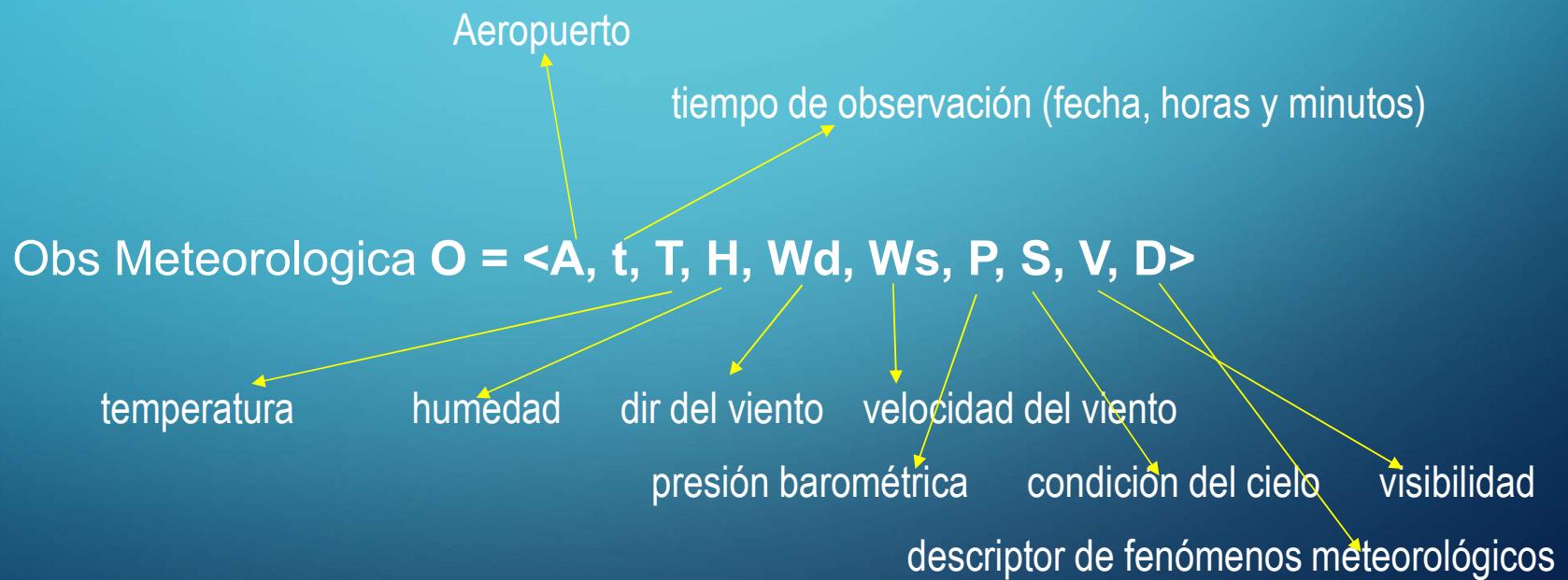
hora de llegada programada

hora de llegada real

la notación de puntos se usa para obtener los diferentes campos de una tupla

# PREDICCION DE DEMORAS EN VUELOS

## Modelado





# PREDICCION DE DEMORAS EN VUELOS

## Modelado

Retraso de llegada de un vuelo F

$$AD(F) = F.taa - F.tsa$$

tiempo de llegada real

tiempo de llegada programada



# PREDICCION DE DEMORAS EN VUELOS

## Modelado

Dado un vuelo  $F$  y un umbral  $Th$ :

$F$  es un **vuelo puntual** si  $AD(F) < Th$

$F$  es un **vuelo demorado** si  $AD(F) \geq Th$



# PREDICCION DE DEMORAS EN VUELOS

## Objetivo

Predecir el retraso de llegada de un vuelo programado debido a las condiciones climáticas.

El retraso de llegada previsto toma en consideración la información de vuelo (aeropuerto de origen y destino, hora de salida y llegada programadas) y las condiciones climáticas en el aeropuerto de origen y de destino en el horario de vuelo.



# PREDICCION DE DEMORAS EN VUELOS

## Objetivo

Si el retraso de llegada es menor que un umbral dado, se clasifica como un vuelo a tiempo; de lo contrario, es clasificado como un vuelo retrasado.

*No se toman en cuenta las condiciones climáticas en ruta de vuelo  
(habría que combinar medidas de estaciones meteorológicas para cada posición de la aeronave, teniendo en cuenta la altitud)*



# PREDICCION DE DEMORAS EN VUELOS

## Fuentes de Datos

Datos para F:

AOTP (Airline On-Time Performance) provistos por RITA (Bureau of Transportation Statistics)

Periodo= Desde Enero 2009 a Diciembre 2013

<http://transtats.bts.gov/>

# PREDICCION DE DEMORAS EN VUELOS

## Fuentes de Datos

Datos para F:

AOTP (Airline On-Time Performance  
Transportation Statistics)

Periodo= Desde Enero 2009 a Diciembre 2017

<http://transtats.bts.gov/>

The screenshot shows the homepage of the Bureau of Transportation Statistics (BTS) website. The header includes the United States Department of Transportation logo, a search bar, and links for Ask a Research Question, A-Z Index, Newsroom, Learn About BTS and Our Work, Browse Statistical Products and Data, and Explore Topics and Geography. The main content area features a sidebar with links for TranStats, Search this site, Advanced Search, Resources (Database Directory, Glossary, Upcoming Releases, Data Release History), and Data Tools (Analysis, Table Profile, Table Contents). The main content area is titled "On-Time : On-Time Performance" and includes sections for Download Instructions (Latest Available Data: September 2017), Filter Geography, Filter Year, and Filter Period. It also lists Prezipped File, % Missing, Documentation, and Terms options. Below this is a table titled "Field Name" with columns for Description and Support Table, listing Time Period (Year, Quarter, Month, DayofMonth, DayOfWeek, FlightDate), Airline (UniqueCarrier, AirlineID, Carrier), and other fields like %Missing, Documentation, and Terms. Each entry in the table has a "Get Lookup Table" link.



# PREDICCION DE DEMORAS EN VUELOS

## Fuentes de Datos

Datos para O:

QCLCD (Quality Controlled Local Climatological Data) disponible en NCDC (National Climatic Data Center)

Periodo= Desde Enero 2009 a Diciembre 2013

<https://www.ncdc.noaa.gov/climate-information>

# PREDICCION DE DEMORAS EN VUELOS

## Fuentes de Datos

Datos para O:  
QCLCD (Quality Controlled Local  
en NCDC (National Climatic Data Center)

Periodo= Desde Enero 2009 a la fecha

<https://www.ncdc.noaa.gov/climate-information>

The screenshot shows the NOAA National Centers for Environmental Information website. The header includes the NOAA logo, the text "NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION", and "NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION". Below the header, a sub-header reads "Formerly the National Climatic Data Center (NCDC)... [more about NCEI»](#)". The main navigation bar includes links for Home, Climate Information, Data Access, Customer Support, Contact, and About. A search bar is located in the top right corner. The breadcrumb navigation shows the path: Home > Data Access > Land-Based Station > Datasets > Local Climatological Data (LCD). On the left, a sidebar titled "Quick Links" lists various datasets: Land-Based Station, Datasets (with sub-links for LCD, COOP, Climate Normals, USHCN, GHCN, GSOD, USCRN, GOSIC, ASOS, AWOS, Solar Radiation, World War II Era Data, Integrated Surface Database (ISD)), Find a Station, Station Metadata, and Climate Data Online. The main content area is titled "Local Climatological Data (LCD)". It describes LCD as hourly, daily, and monthly summaries for approximately 1,600 U.S. locations. It includes sections for "Local Climatological Data (LCD)", "LCD Sample", and "QCLCD ASCII Files".



# PREDICCION DE DEMORAS EN VUELOS

## Fuentes de Datos

La Oficina de Estadísticas de Transporte (BTS) ha identificado 5 causas de demora:

- **Air carrier:** la causa del retraso es debido a la aerolínea (mantenimiento, limpieza)
- **Avion Llegada tardía:** un vuelo anterior con el mismo avión llegó tarde, causando que el vuelo actual salga tarde
- **NAS:** demoras debido al Sistema Nacional de Aviación que se refieren a un gran conjunto de condiciones (climáticas no extremas, control de tráfico aéreo)
- **Clima extremo:** condiciones meteorológicas significativas (reales o pronosticadas) que retrasan o impiden el funcionamiento de un vuelo
- **Seguridad:** evacuación de una terminal, reboarding por violación de seguridad,etc

# PREDICCION DE DEMORAS EN VUELOS

## Fuentes de Datos

Table IV. Analysis of Flight Delay Causes by Year

Year	Air carrier	Late-arriving aircraft	NAS	Extreme weather	Security
2009	26.6%	32.8%	37.0%	3.4%	0.2%
2010	28.9%	35.8%	32.1%	3.1%	0.3%
2011	28.2%	37.0%	31.8%	2.8%	0.2%
2012	29.8%	37.6%	29.6%	2.8%	0.2%
2013	27.8%	38.8%	30.3%	2.9%	0.2%

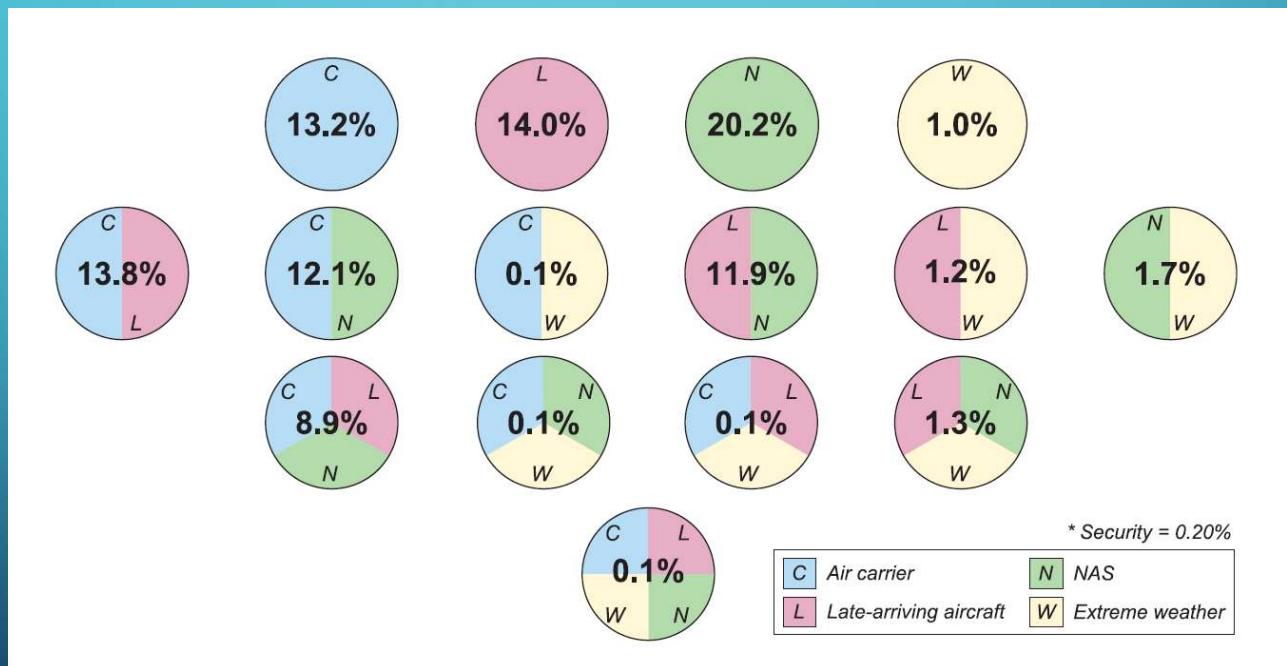
Table V. Analysis of Delayed Flights Due to Weather Conditions by Year

Year	Extreme weather	NAS related to weather	Late-arriving aircraft related to weather	Total weather
2009	3.4%	24.3%	14.5%	42.3%
2010	3.1%	20.4%	14.0%	37.4%
2011	2.8%	20.1%	14.3%	37.2%
2012	2.8%	17.4%	12.6%	32.8%
2013	2.9%	17.7%	14.1%	34.6%

# PREDICCION DE DEMORAS EN VUELOS

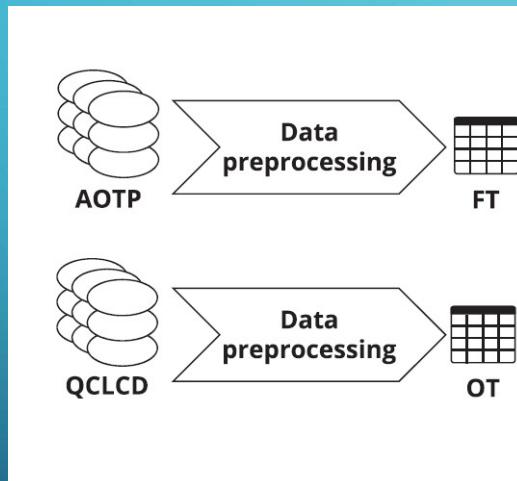
## Fuentes de Datos

Un vuelo demorado se puede deber a un único retraso o a un retraso múltiple de varias categorías



# PREDICCION DE DEMORAS EN VUELOS

## Preprocesamiento de Datos



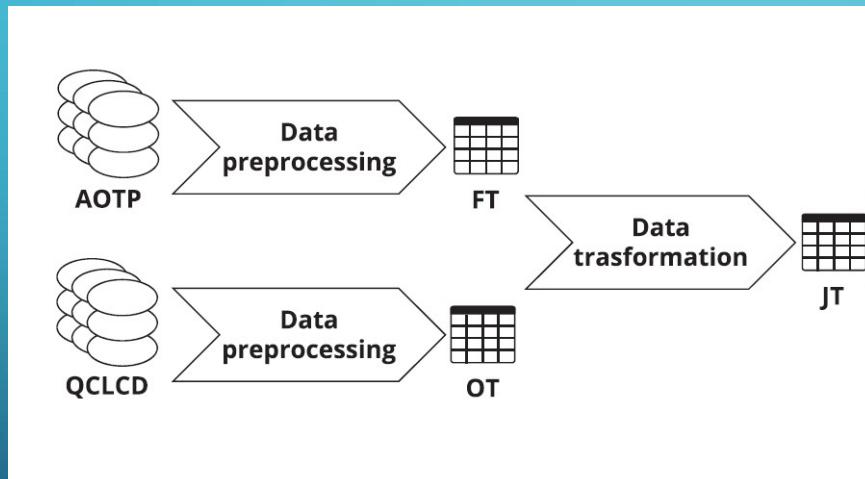
Limpieza + eliminación de vuelos cancelados

Limpieza + eliminación de observaciones no relacionadas  
con ubicaciones de aeropuertos

# PREDICCION DE DEMORAS EN VUELOS

## Preprocesamiento de Datos

Para cada vuelo  $F$  en  $FT$  se crea en  $JT$  una tupla  $\{F, Wo, Wd, C\}$



$$F = \langle Ao, Ad, tsd, tsa, \dots \rangle$$

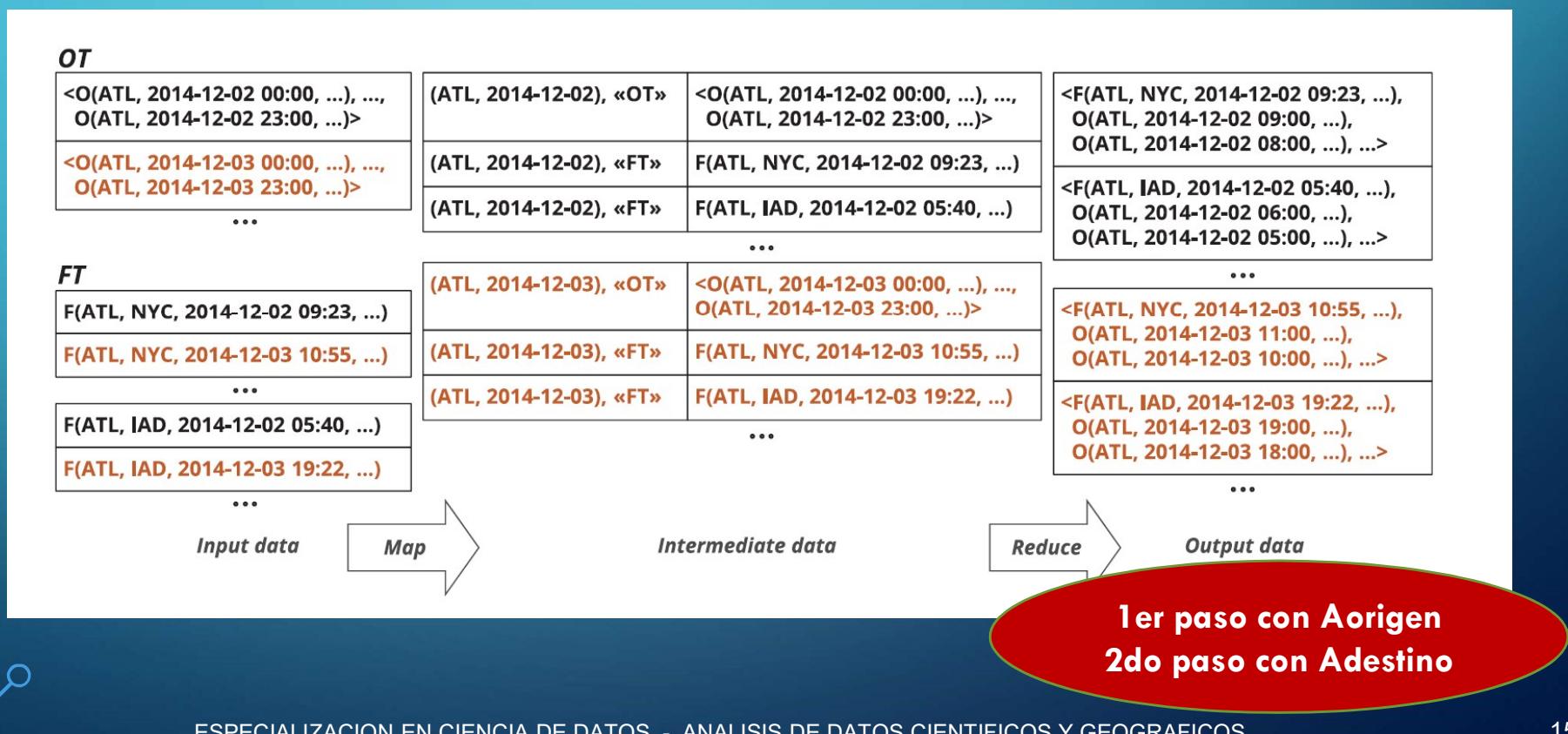
$$Wo = \langle O(Ao, tsd), O(Ao, tsd-1h), \dots, O(Ao, tsd - 12h) \rangle$$

$$Wd = \langle O(Ad, tsa), O(Ad, tsa-1h), \dots, O(Ad, tsa - 12h) \rangle$$

C es el atributo de clase que indica si F está puntual o demorado con un determinado umbral Th.

# PREDICCION DE DEMORAS EN VUELOS

## Preprocesamiento de Datos

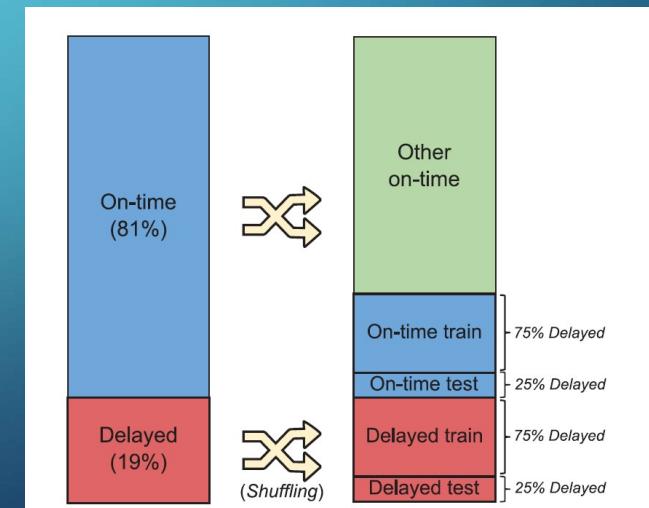


# PREDICCION DE DEMORAS EN VUELOS

## Preparación del Training and Test Set

Dataset ID	% Delayed tuples selected	% Delayed tuples ( $Th = 15\text{min}$ )	# Delayed tuples ( $Th = 15\text{min}$ )	% Delayed tuples ( $Th = 60\text{min}$ )	# Delayed tuples ( $Th = 60\text{min}$ )
D1	<i>Solo Extreme U Solo NAS U</i>	22.9%	1.3M	15.4%	257k
D2	<i>Solo (Extreme and NAS)</i>				
D3	<i>Extreme U <math>NAS \geq Th</math></i>	37.1%	2.1M	25.9%	433k
D4	<i>Extreme U NAS</i>	58.9%	3.4M	56.8%	950k
	All	100.0%	5.8M	100.0%	1.7M

Dado que los conjuntos están desequilibrados (hay mas a tiempo que demorados), realizan un algoritmo de submuestreo aleatorio que equilibra la distribución de clases en partes iguales: las tuplas con demoras se distribuyen al azar entre los grupos de las tuplas a tiempo.

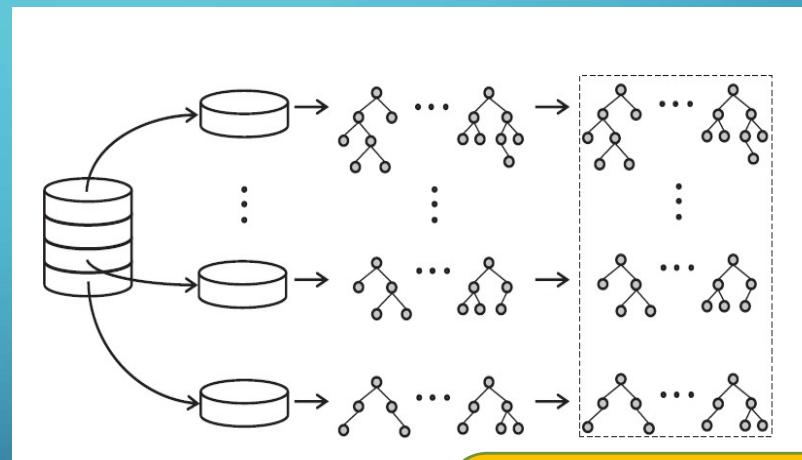


# PREDICCION DE DEMORAS EN VUELOS

## Modelado del Clasificador

Se probaron diferentes machine-learning y se termino eligiendo el de mejor performance: **Random Forest**

RF crea una colección de diferentes árboles de decisión. Cada árbol se construye a partir de un conjunto de datos de entrenamiento obtenido aplicando subconjuntos en el conjunto de entrenamiento original.



Una vez que se crean los árboles forestales, la clasificación de una tupla sin etiqueta se realiza agregando las predicciones de la diferentes árboles a través de la votación por mayoría.



# PREDICCION DE DEMORAS EN VUELOS

## Prediccion

Para predecir la clase (a tiempo o demorado) de un vuelo programado  $F_s$ , la entrada al predictor es:

- $F_s = \langle A_o, A_d, t_{sd}, t_{sa} \rangle$
- $W_o = \langle O(A_o, t_{sd}), O(A_o, t_{sd}-1h), \dots, O(A_o, t_{sd}-(m-1)h) \rangle$
- $W_d = \langle O(A_d, t_{sa}), O(A_d, t_{sa}-1h), \dots, O(A_d, t_{sa}-(n-1)h) \rangle$
- Umbral  $d \in \{Th_1, Th_2, \dots, Th_z\}$ .

# PREDICCION DE DEMORAS EN VUELOS

## Indicadores de Performance

Matriz de confusión:

		<i>Predicho</i>	
		Puntual	Demorado
<i>Real</i>	Puntual	TP	FN
	Demorado	FP	TN



# PREDICCION DE DEMORAS EN VUELOS

## Indicadores de Performance

Precisión

$$Acc = (TP + TN) / (TP + TN + FP + FN)$$

Recall Puntual

$$Rec-o = TP / (TP + FN)$$

Recall Demorado

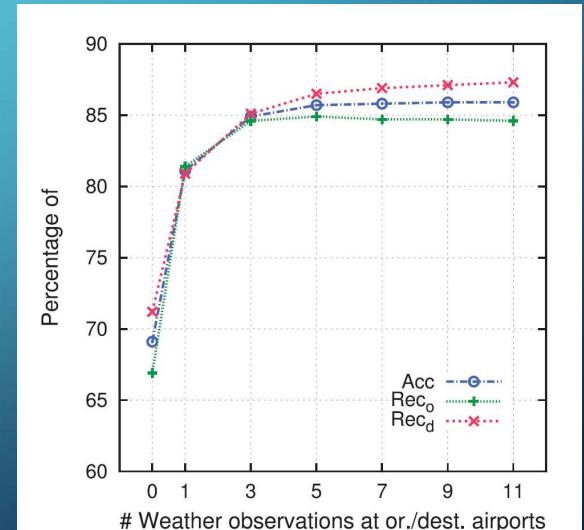
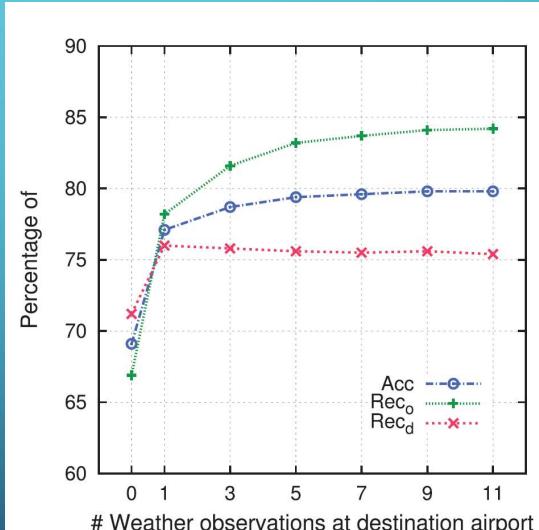
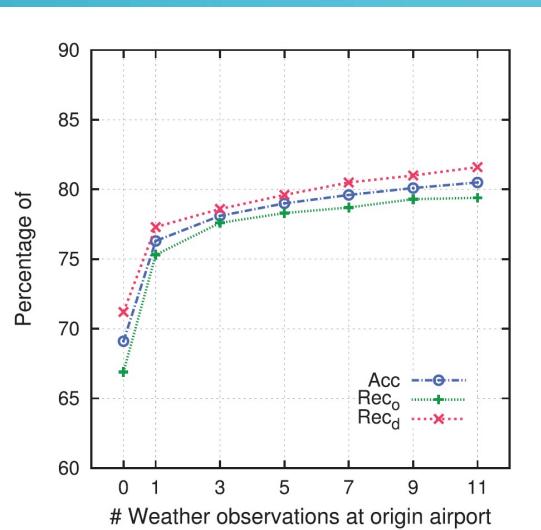
$$Rec-d = TN / (TN + FP)$$

		<i>Predicho</i>	
		Puntual	Demorado
<i>Real</i>	Puntual	TP	FN
	Demorado	FP	TN

# PREDICCION DE DEMORAS EN VUELOS

## Resultados

Para D2 y Th 60 minutos



# PREDICCION DE DEMORAS EN VUELOS

## Resultados

Table IX. Turnaround Time and Relative Speedup Values (Calculated with Respect to 2 Workers) of the Four Data-Mining Phases

Operation	1× (2 workers)		2× (4 workers)		4× (8 workers)		6× (12 workers)	
	Turn. time	Speed up	Turn. time	Speed up	Turn. time	Speed up	Turn. time	Speed up
Data preprocessing and transformation	03:08:55	–	01:40:52	1.9	00:49:16	3.8	00:34:39	5.5
Target data creation	02:14:06	–	01:06:59	2.0	00:33:19	4.0	00:23:16	5.8
Modeling	02:29:20	–	01:13:12	2.0	00:37:35	4.0	00:24:44	6.0
Evaluation	04:19:28	–	02:14:17	1.9	01:08:51	3.8	00:49:18	5.3
Total	12:11:49	–	06:15:20	1.9	03:09:01	3.9	02:11:57	5.5

# PREDICCION DE DEMORAS EN VUELOS

## Resultados

