

# Minería de Datos

## IIC2433

Naïve Bayes y Data Analysis

Vicente Domínguez

# ¿Qué vimos las clases pasadas?

- Random Forest

# ¿Qué veremos esta clase?

- Otra forma de clasificar: Naïve Bayes (Basado en diapositivas del prof. Denis Parra)
- Bayesian Data Analysis

# Probabilidades condicionales y conjuntas

## *Ejemplo*



Al tirar una moneda

- ¿Cuál es la probabilidad de que salga sello
- ¿Cuál es la probabilidad de que salga dos veces sello al tirarla dos veces?

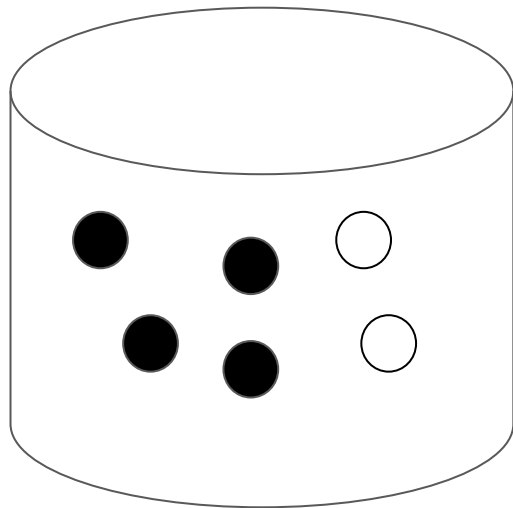
## Probabilidad conjunta con eventos **independientes**

$$P(A, B) = P(A) * P(B)$$

$$P(A, B, C) = P(A) * P(B) * P(C)$$

# Probabilidades condicionales y conjuntas

## *Ejemplo*



Tengo 4 bolitas negras y 2 blancas en una tómbola,

- Al sacar una bolita al azar, ¿Cuál es la probabilidad de que salga una blanca?
- Al sacar dos bolitas al azar, ¿Cuál es la probabilidad de que ambas salgan blancas?

## Probabilidad conjunta con eventos **dependientes**

$$P(A, B) = P(A|B)*P(B) = P(B|A)*P(A)$$

$$P(A, B, C) = P(A|B,C)*P(B|C)*P(C)$$

# Teorema de Bayes



Thomas Bayes (1701 –  
1761)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

# Teorema de Bayes

- $P(A = \text{sí})$ : Probabilidad del evento A sea “sí”
- $P(A=\text{sí} | B=\text{sí})$ : Probabilidad de que el evento A sea “sí”  
DADO QUE el evento B fue “sí”
- Por simplicidad, usamos  $P(A) = P(A=\text{“sí”})$

$$P(A | B) = \frac{P(A, B)}{P(B)} \quad P(A | B) = \frac{P(B | A) * P(A)}{P(B)}$$

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$



# Teorema de Bayes

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

**Prior:** Distribución de probabilidad a priori. El conocimiento de la probabilidad o incerteza de la clase antes de observar o condicionar los datos.

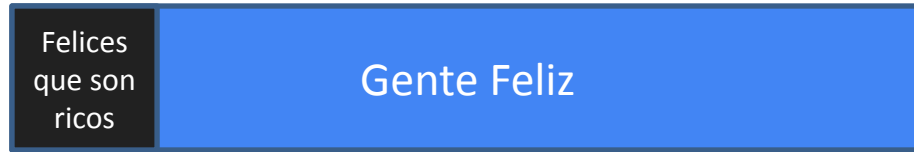
**Likelihood:** La probabilidad del evento bajo cierta clase o categoría, condicionada por los datos.

**Evidence:** Suma de las probabilidades del evento bajo todas las clases.

**Posterior:** Distribución de probabilidad condicional, que representa la probabilidad del evento condicionado luego de observar los datos.

# Noción del Teorema de Bayes

- << La riqueza hace la felicidad >>
- ¿Son felices los ricos?  $P(\text{feliz} = \text{sí} \mid \text{rico} = \text{sí})$
- ... yo sé que de la gente feliz, 20% es rica.



- 
- 20% no es tanto ... por lo cual podemos concluir que la riqueza no hace la felicidad. ¿o no?

$$P(\text{feliz} \mid \text{rico}) = \frac{P(\text{rico} \mid \text{feliz}) * P(\text{feliz})}{P(\text{rico})}$$

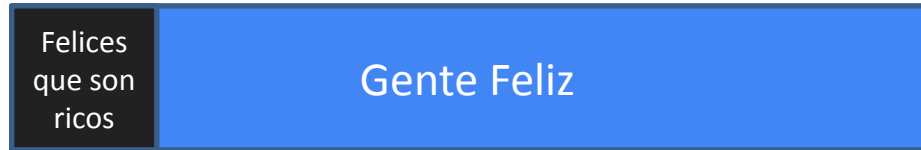
# Noción del Teorema de Bayes

- “La riqueza hace la felicidad”
- ¿Son felices los ricos?  $P(\text{feliz} = \text{sí} \mid \text{rico} = \text{sí})$
- ... yo sé que de la gente feliz, 20% es rica.

## Supongamos:

A: gente feliz = 40% de la población

B: gente rica = 10% de la población



$$P(\text{feliz} \mid \text{rico}) = \frac{P(\text{rico} \mid \text{feliz}) * P(\text{feliz})}{P(\text{rico})}$$

$$\textit{Posterior} = \frac{\textit{Likelihood} \times \textit{Prior}}{\textit{Evidence}}$$

# Noción del Teorema de Bayes

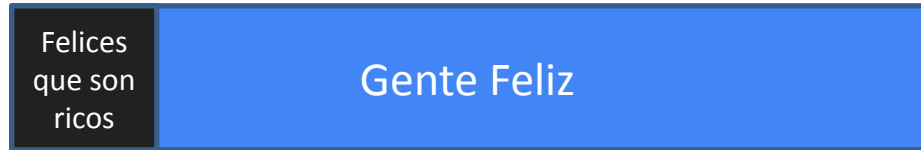
- “La riqueza hace la felicidad”
- ¿Son felices los ricos?  $P(\text{feliz} = \text{sí} \mid \text{rico} = \text{sí})$
- ... yo sé que de la gente feliz, 20% es rica.

## Supongamos:

A: gente feliz = 40% de la población

B: gente rica = 10% de la población

C:  $P(\text{rico} \mid \text{feliz}) = 20\%$



$$P(\text{feliz} \mid \text{rico}) = \frac{P(\text{rico} \mid \text{feliz}) * P(\text{feliz})}{P(\text{rico})}$$

$$\textit{Posterior} = \frac{\textit{Likelihood} \times \textit{Prior}}{\textit{Evidence}}$$

# ¿Y por qué se llama “Naïve” ?

- Naïve significa “**ingenuo**”
- Es “ingenuo” por que asume independencia de los eventos\*

- \* en realidad, asume independencia condicional

Manzana	Carne	Pastel	¿Alergia?
No	Sí	No	Sí
No	Sí	Sí	Sí
No	Sí	No	Sí
Sí	Sí	Sí	Sí
Sí	Sí	No	No
No	No	Sí	No
Sí	No	No	No
No	No	No	No

- ¿Cuál es la probabilidad de haber consumido el alimento Manzana, dado que hubo alergia, es decir  $P(\text{Manzana}=\text{Si}|\text{Alergia}=\text{Si})$ ? ¿Cuál es la probabilidad de haber consumido el alimento pastel, dado que no hubo alergia, es decir,  $P(\text{Pastel}=\text{Si}|\text{Alergia}=\text{No})$  ?

$$P(M|A) = (P(A|M) * P(M)) / P(A) = (1/3 * 3/8) / 1/2 = 1/4$$

$$P(P|\text{No } A) = (P(\text{No } A|P) * P(P)) / P(\text{No } A) = (1/3 * 3/8) / 1/2 = 1/4$$

# Volviendo: Ejemplo de Clasificación

- Consideremos un auto SUV, color rojo, doméstico. ¿La probabilidad de que la roben es mayor o menor de que no la roben?

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

$P(\text{Robo} \mid \text{Red, SUV, Domestic}) = \text{Posterior}$

$$P(\text{Robo}) = \underbrace{(p(\text{Robo}))}_{\text{Prior}} \underbrace{(p(\text{Color} \mid \text{Robo}) * p(\text{Tipo} \mid \text{Robo}) * p(\text{Origin} \mid \text{Robo}))}_{\text{Likelihood}} / N$$

Rojo
SUV
Domestic

$$N = p(\text{Robo})p(\text{Color} \mid \text{Robo}) * p(\text{Tipo} \mid \text{Robo}) * p(\text{Origin} \mid \text{Robo}) + p(\text{No Robo})p(\text{Color} \mid \text{No Robo}) * p(\text{Tipo} \mid \text{No Robo}) * p(\text{Origin} \mid \text{No Robo})$$

} Evidence



Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

$$P(\text{Robo}) = (p(\text{Robo})p(\text{Color} | \text{Robo}) * p(\text{Tipo} | \text{Robo}) * p(\text{Origen} | \text{Robo})) / N$$

$$N = p(\text{Robo})p(\text{Color} | \text{Robo}) * p(\text{Tipo} | \text{Robo}) * p(\text{Origen} | \text{Robo})$$

+

$$p(\text{No Robo})p(\text{Color} | \text{No Robo}) * p(\text{Tipo} | \text{No Robo}) * p(\text{Origen} | \text{No Robo})$$

$$P(\text{Robo}) = (5/10 * 3/5 * 1/5 * 2/5) / ((5/10 * 3/5 * 1/5 * 2/5) + (5/10 * 2/5 * 3/5 * 3/5))$$

$$P(\text{Robo}) = 0.25$$

Manzana	Carne	Pastel	¿Alergia?
No	Sí	No	Sí
No	Sí	Sí	Sí
No	Sí	No	Sí
Sí	Sí	Sí	Sí
Sí	Sí	No	No
No	No	Sí	No
Sí	No	No	No
No	No	No	No

- Basado en los datos de la Tabla 1, usando un clasificador Naive Bayes, clasifique los siguientes dos casos dados los datos de la Tabla 1.

Manzana	Carne	Pastel	¿Alergia?
Sí	No	Sí	??
Sí	Sí	No	??

# Ejemplo de Clasificación Numérico

- ¿Qué ocurrirá en el caso con datos numéricos y no categóricos?

# Ejemplo de Clasificación Numérico

- ¿Qué ocurrirá en el caso con datos numéricos y no categóricos?
- R: En general se asume que los datos distribuyen como una gaussiana (puede ser otra distribución) y se calculan sus parámetros acorde a los datos.
- Luego, se utilizan para utilizar la función de densidad para calcular un estimado de la probabilidad del dato a predecir.

# Referencias

- Material de Tom M. Mitchell, CMU:

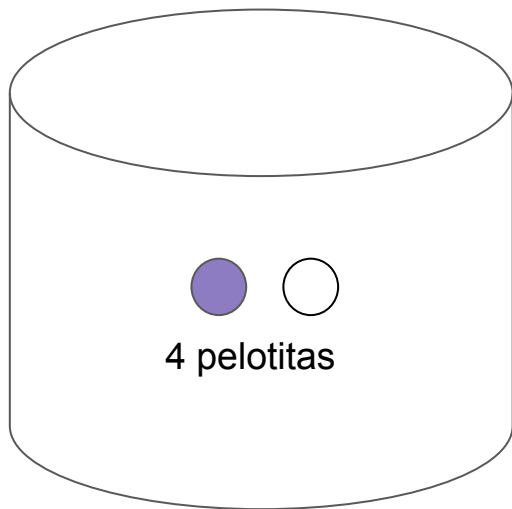
<http://www.cs.cmu.edu/~awm/15781/slides/NBayes-9-27-05.pdf>

<http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf>

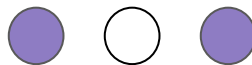
# Bayesian Data Analysis

- Supongamos que tienes un conjunto de datos, ¿cómo podrías utilizarlo para aprender del mundo?
- Las siguientes slides están basadas en el libro y curso Statistical Rethinking de Richard McElreath (recomendadísimo)

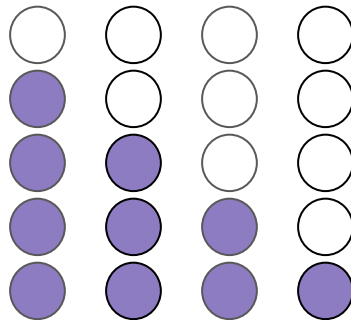
# Bayesian Data Analysis



Luego de sacar 3 veces pelotitas con reposición, obtienes lo siguiente:

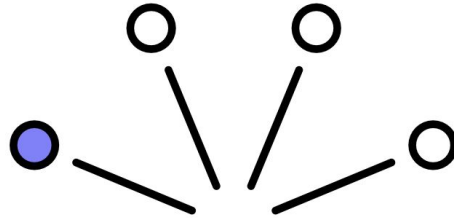


De qué color son las pelotitas que hay?  
Posibles opciones



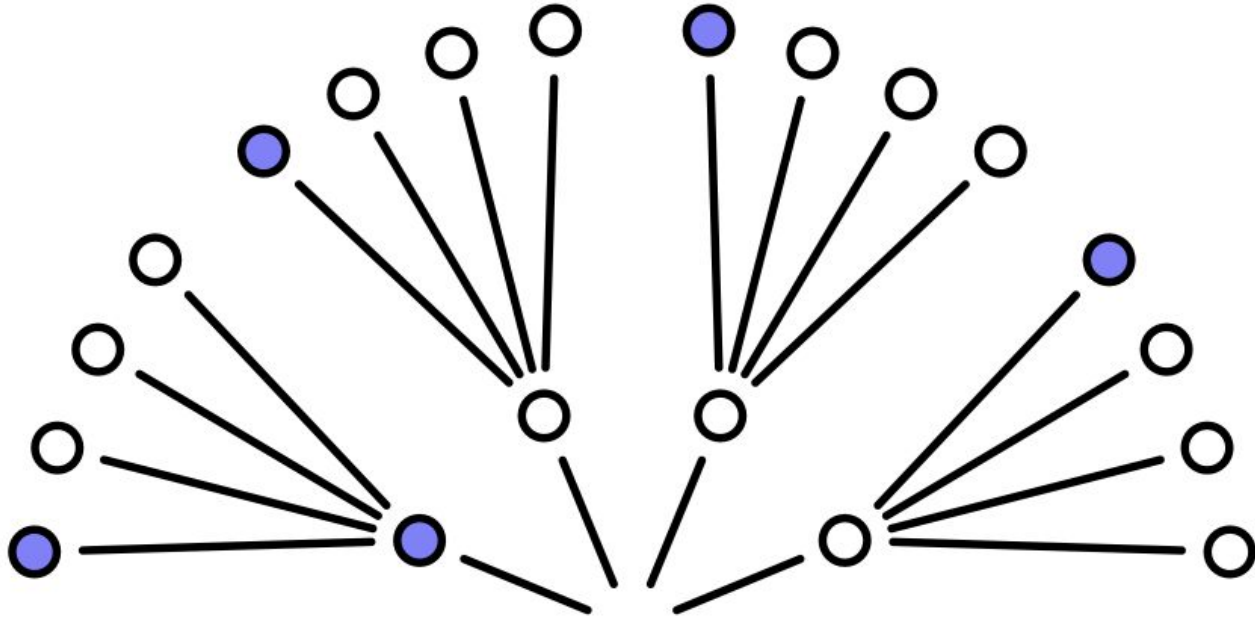
# Bayesian Data Analysis

- Supongamos la siguiente conjetura: En la bolsa hay 

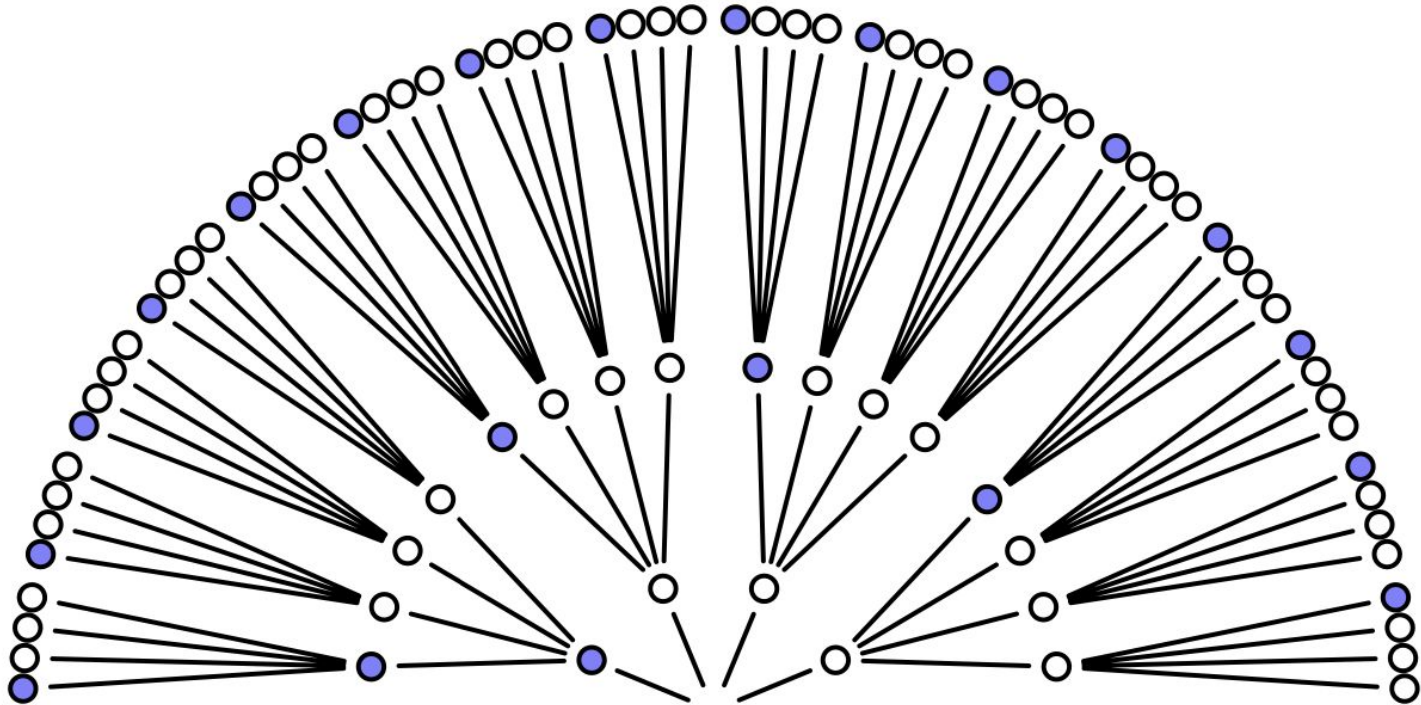




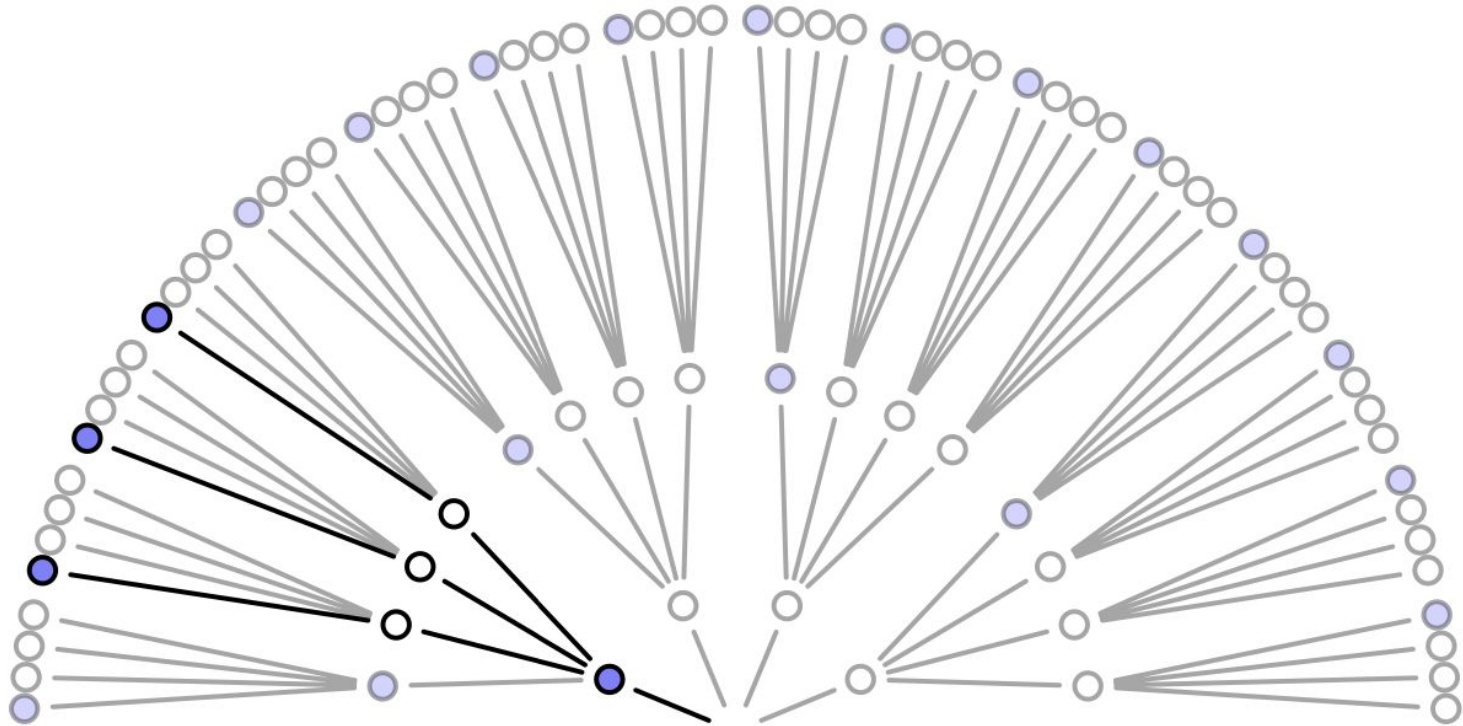
# Bayesian Data Analysis


























# Bayesian Data Analysis

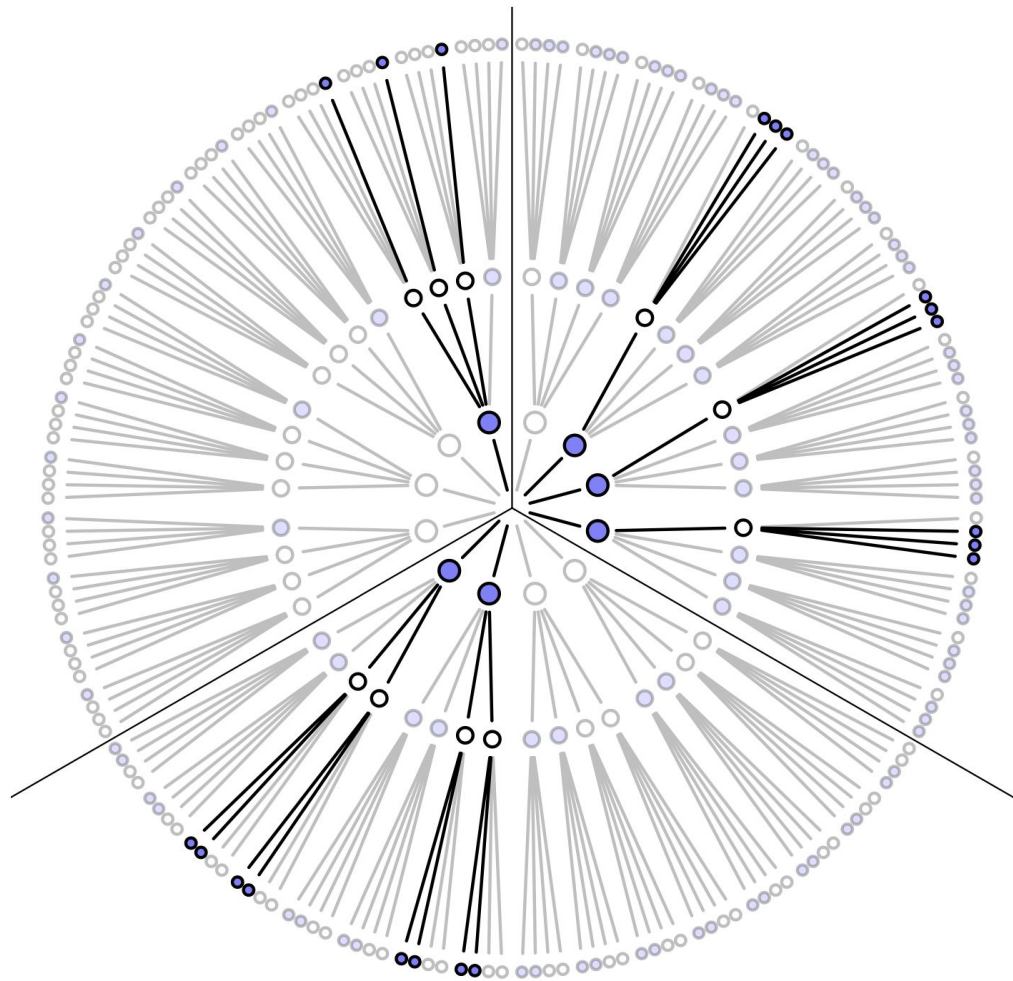


# Bayesian Data Analysis

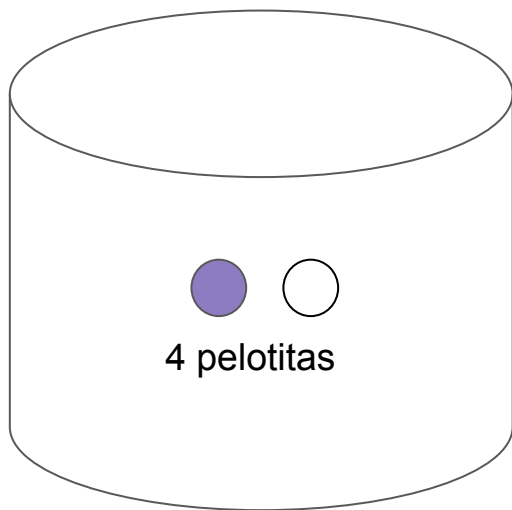



# Bayesian Data Analysis

Conjecture	Ways to produce   
[     ]	$0 \times 4 \times 0 = 0$
[     ]	$1 \times 3 \times 1 = 3$
[     ]	$2 \times 2 \times 2 = 8$
[     ]	$3 \times 1 \times 3 = 9$
[     ]	$4 \times 0 \times 4 = 0$




# Bayesian Data Analysis



Supón que ahora sacas una nueva pelotita y es 

Podrías hacer todos los cálculos de nuevo, o usar la información previa

Conjecture	Ways to produce 	Prior counts	New count
[○○○○]	0	0	$0 \times 0 = 0$
[●○○○]	1	3	$3 \times 1 = 3$
[●●○○]	2	8	$8 \times 2 = 16$
[●●●○]	3	9	$9 \times 3 = 27$
[●●●●]	4	0	$0 \times 4 = 0$

# Bayesian Data Analysis

Imagina que ahora te dan información adicional, los de la fábrica te cuentan que por cada bolsa con  $[\bullet\bullet\bullet\circ]$  hicieron **2** bolsas con  $[\bullet\bullet\circ\circ]$  y **3** bolsas con  $[\bullet\circ\circ\circ]$

Además, te comentan que se aseguraron de que cada bolsa contendría al menos **1** pelotita blanca y **1** azul. Con esto podemos actualizar la información

Conjecture	Prior count	Factory	
		count	New count
$[\circ\circ\circ\circ]$	0	0	$0 \times 0 = 0$
$[\bullet\circ\circ\circ]$	3	3	$3 \times 3 = 9$
$[\bullet\bullet\circ\circ]$	16	2	$16 \times 2 = 32$
$[\bullet\bullet\bullet\circ]$	27	1	$27 \times 1 = 27$
$[\bullet\bullet\bullet\bullet]$	0	0	$0 \times 0 = 0$

# Bayesian Data Analysis

- ¿Es posible ver esto como probabilidades?
- Se puede definir el proceso realizado como lo siguiente

$$\begin{aligned} &\text{plausibility of } [\bullet \circ \circ \circ] \text{ after seeing } \bullet \circ \bullet \\ &\propto \\ &\text{ways } [\bullet \circ \circ \circ] \text{ can produce } \bullet \circ \bullet \\ &\times \\ &\text{prior plausibility } [\bullet \circ \circ \circ] \end{aligned}$$



# Bayesian Data Analysis

- Definiendo esto con fórmulas

plausibility of  $p$  after  $D_{\text{new}} \propto$  ways  $p$  can produce  $D_{\text{new}} \times$  prior plausibility of  $p$

- Normalizando

plausibility of  $p$  after  $D_{\text{new}} = \frac{\text{ways } p \text{ can produce } D_{\text{new}} \times \text{prior plausibility } p}{\text{sum of products}}$

# Bayesian Data Analysis

- Finalmente reemplazando la tabla anterior obtenemos

Possible composition	$p$	Ways to produce data	Plausibility
[○○○○]	0	0	0
[●○○○]	0.25	3	0.15
[●●○○]	0.5	8	0.40
[●●●○]	0.75	9	0.45
[●●●●]	1	0	0

# Inferencia Bayesiana

- Hay muchas más herramientas basadas en estos principios
  - Bayesian data analysis
  - Model comparison
  - Multilevel models
  - Graphical causal models