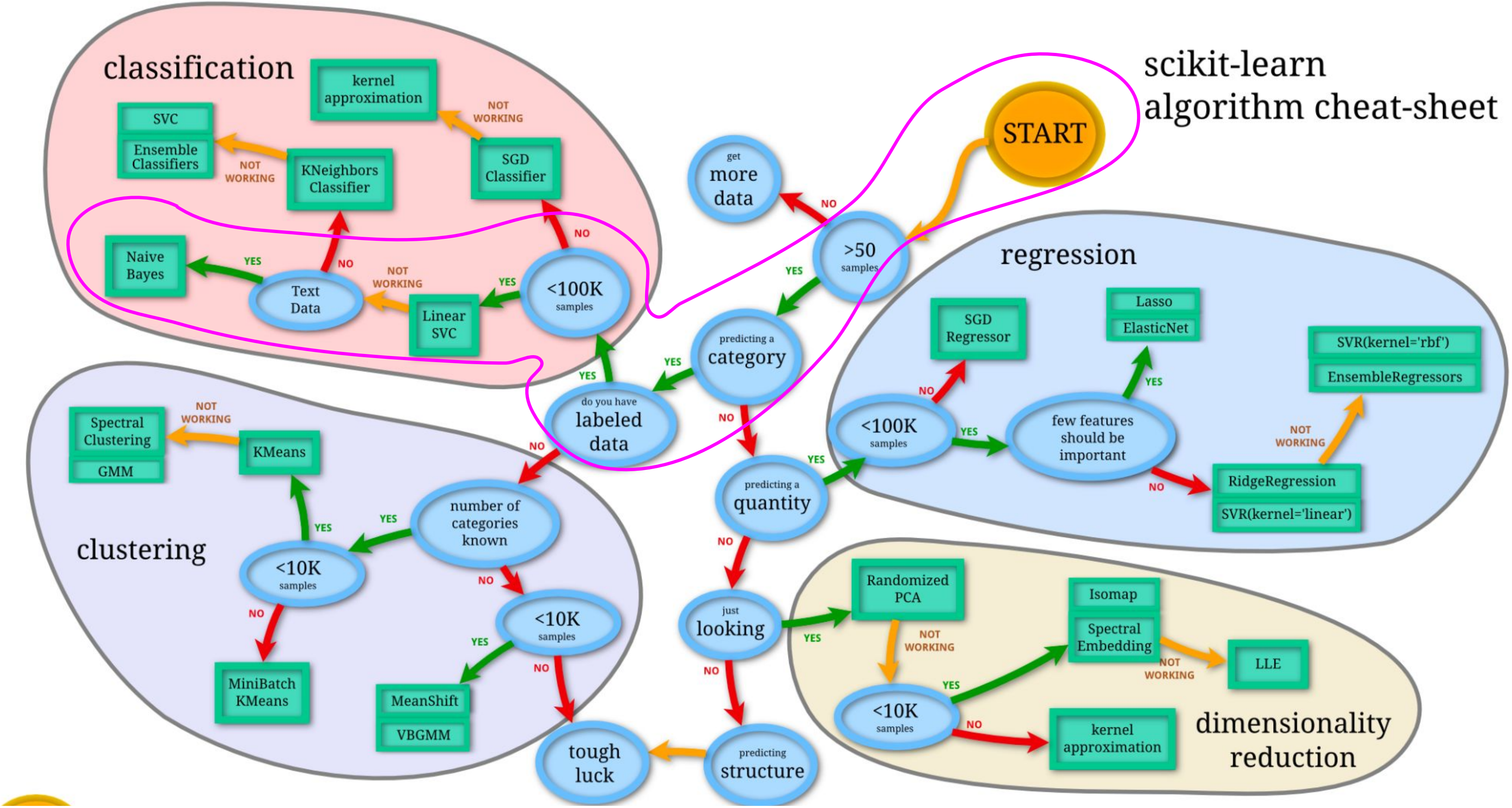
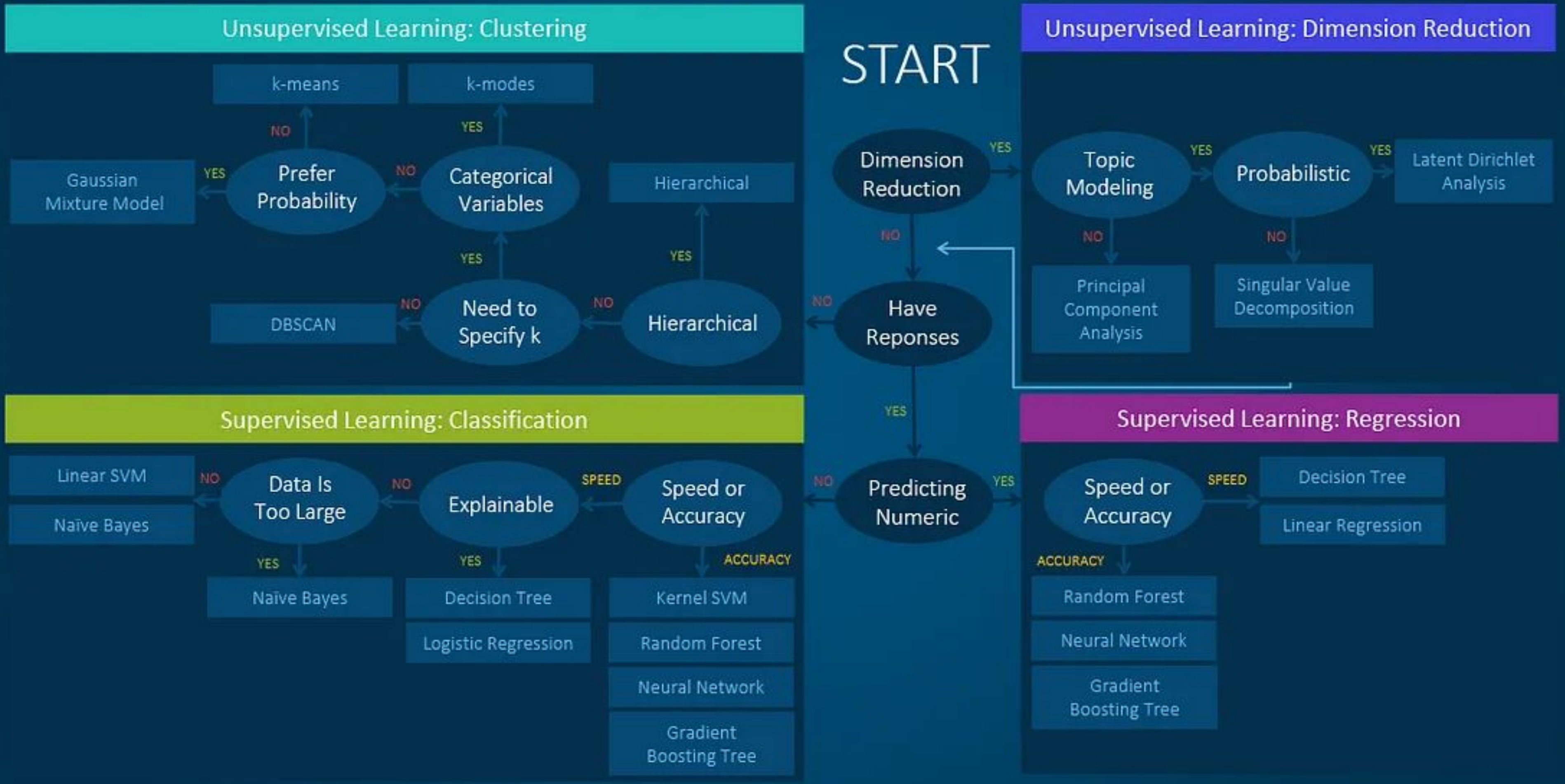
Unidad 4:
Estadística bayesiana para ML
Naive Bayes

By Ruth Chirinos



scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
classification
regression
clustering
dimensionality reduction
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
<100K samples
SGD Classifier
kernel approximation
Linear SVC
Text Data
Naive Bayes
KNeighbors Classifier
SVC
Ensemble Classifiers
SGD Regressor
few features should be important
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
RidgeRegression
SVR(kernel='linear')
number of categories known
<10K samples
KMeans
Spectral Clustering
GMM
MiniBatch KMeans
MeanShift
VBGMM
Randomized PCA
Isomap
Spectral Embedding
LLE
YES
NO
NOT WORKING

Machine Learning Algorithms Cheat Sheet



Probabilidad Condicional

Probabilidad Condicional

Ocurre cuando dos eventos o sucesos son **dependientes** entre sí, y la ocurrencia de uno **condiciona** la ocurrencia del otro.

Expresiones

Suceso → A

Suceso → B

Probabilidad de que A ocurra → $P(A)$

Probabilidad de que B ocurra → $P(B)$

Probabilidad de que A ocurra habiendo ocurrido ya B → $P(A/B)$

Probabilidad de que B ocurra habiendo ocurrido ya A → $P(B/A)$

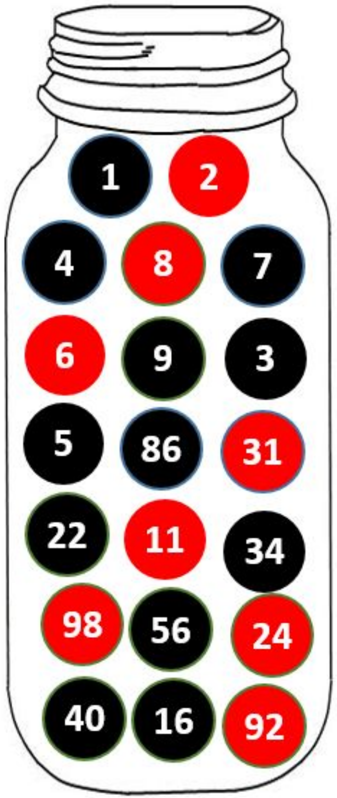
$$\textbf{Probabilidad} = \frac{\#\ \text{de casos favorables}}{\#\ \text{total de resultados posibles}}$$

$A \rightarrow$ Rojas

$\bar{A} \rightarrow$ Negras

$B \rightarrow$ Par

$\bar{B} \rightarrow$ Impar



Sucesos Favorables al sacar una bola al azar

$$\left.\begin{array}{l} A \rightarrow \text{Rojas} \rightarrow 8 \\ \bar{A} \rightarrow \text{Negras} \rightarrow 12 \end{array}\right] 20$$

$$\left.\begin{array}{l} B \rightarrow \text{Par} \rightarrow 13 \\ \bar{B} \rightarrow \text{Impar} \rightarrow 7 \end{array}\right] 20$$

Que sea **roja**

$$P(A) = \frac{8}{20}$$

Que sea **negra**

$$P(\bar{A}) = \frac{12}{20}$$

Que sea **par**

$$P(B) = \frac{13}{20}$$

Que sea **impar**

$$P(\bar{B}) = \frac{7}{20}$$

Probabilidad Condicional

Ocurre cuando dos eventos o sucesos son **dependientes** entre sí, y la ocurrencia de uno **condiciona** la ocurrencia del otro.

Expresiones

Suceso → A

Suceso → B

Probabilidad de que A ocurra → $P(A)$

Probabilidad de que B ocurra → $P(B)$

Probabilidad de que A ocurra habiendo ocurrido ya B → $P(A/B)$

Probabilidad de que B ocurra habiendo ocurrido ya A → $P(B/A)$

Fórmulas

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

Que sea **roja**

$$P(A) = \frac{8}{20}$$

Que sea **negra**

$$P(\bar{A}) = \frac{12}{20}$$

Que sea **par**

$$P(B) = \frac{13}{20}$$

Que sea **impar**

$$P(\bar{B}) = \frac{7}{20}$$

Tabla de doble entrada o de contingencia

	A: rojas	$\bar{A}$: negras	
B: pares	6	7	13
$\bar{B}$: impares	2	5	7
	8	12	20

Probabilidad de que B ocurra habiendo ocurrido ya A
Probabilidad de que sea **par** habiendo salido **roja** → $P(B/A) = \frac{6}{8}$

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{6}{20}}{\frac{8}{20}} = \frac{6 \cdot \cancel{20}}{\cancel{20} \cdot 8} = \frac{6}{8}$$

Clasificador Naive Bayes

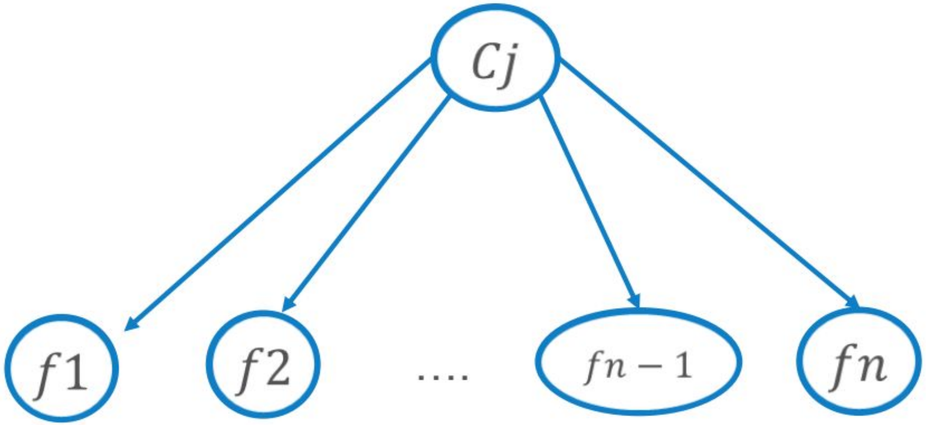
Puntos importantes

- Aprendizaje Supervisado
  - Naive Bayes
    - Teorema de Bayes
    - Tablas de Probabilidad Condicional
    - Independencia Condicional
    - Naive Bayes con variables numéricas

Naive Bayes

Cuando usamos Naive Bayes?

- Pocas clases (2,3,4, ..)
- Muchos atributos (cientos, miles, …)
- Una RB por clase, que será padre de todos los atributos.



Naive Bayes

- **Objetivo:** realizar clasificación mediante la probabilidad, pero simplificando los métodos existentes.
- **Requisitos:** independencia condicional total.
- **Base:** teorema de Bayes.

$$P(Ai \mid B) = \frac{P(B \mid Ai)\, P(Ai)}{P(B)}$$

Donde:
$P(Ai)$ son las probabilidades a priori
$P(B \mid Ai)$ es la probabilidad de B en la hipótesis A_i
$P(Ai \mid B)$ son las probabilidades a posteriori

Naive Bayes

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

Es decir, la probabilidad de A dado que ocurre B, es igual a la probabilidad de que ocurra B dado que ha ocurrido A, multiplicado por la probabilidad de que ocurra A, dividido por la probabilidad de que ocurra B.

GAUSSIAN NAIVE BAYES CLASSIFIER

"Gaussian" because this is a normal distribution

This is our prior belief

$$p(\text{class} \mid \text{data}) = \frac{p(\text{data} \mid \text{class}) \times p(\text{class})}{p(\text{data})}$$

We don't calculate this in naive bayes classifiers

ChrisAlbon

Naive Bayes

- La relación entre los eventos dependientes se puede describir utilizando el teorema de Bayes: $P(A|B)=\{P(B|A)P(A)\} / P(B)$
- Supongamos que queremos estimar la probabilidad de que un mensaje sea spam. Sin tener evidencias adicionales es un 0.2 (probabilidad apriori).
- Si tenemos la evidencia que el mensaje contiene la palabra viagra. La probabilidad de que viagra se haya usado en mensajes de spam previos se llama verosimilitud (likelihood) y la probabilidad de que aparezca en cualquier mensaje se llama verosimilitud marginal (marginal likelihood).

Naive Bayes

- Para obtener cada uno de los componentes hay que construir una tabla de frecuencias, que indica el número de veces que la palabra viagra ha aparecido en los mensajes de spam. Esta tabla de frecuencias se puede utilizar para calcular una tabla de verosimilitud.

- Aplicando el teorema de Bayes, se puede calcular la probabilidad a posteriori (*posterior probability*). Si es mayor que 0.5, es más probable que sea *spam* que *ham,* por lo que se debería filtrar.

probabilidad previa

$$P\,(spam \mid Viagra) = \frac{P\,(Viagra \mid spam)\,P(spam)}{P(Viagra)}$$

probabilidad posterior

probabilidad marginal

Naive Bayes

	Viagra		
Frecuencia	Sí	No	Total
spam	4	16	20
ham	1	79	80
Total	5	95	100

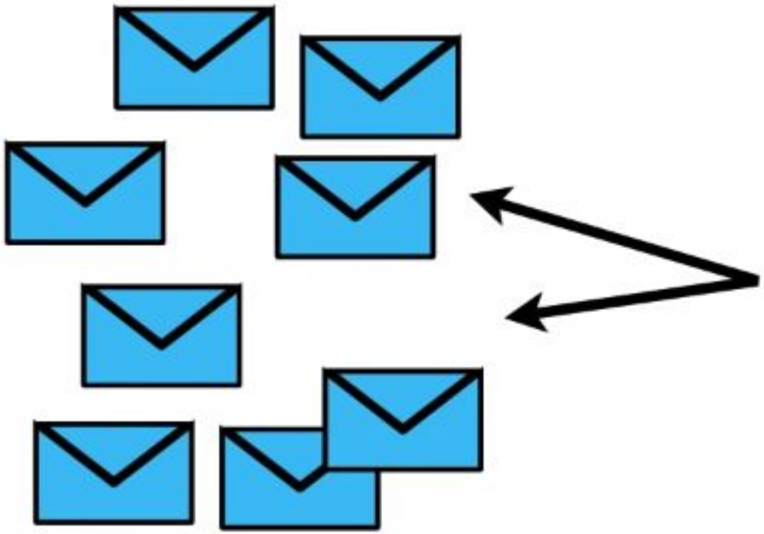
	Viagra		
Probabilidad	Sí	No	Total
spam	4/20	16/20	20
ham	1/80	79/80	80
Total	5/100	95/100	100

- Para calcular la probabilidad posteriori

  P(Spam|Viagra) = (4/20) *(20/100)/ (5/100) = 0.8

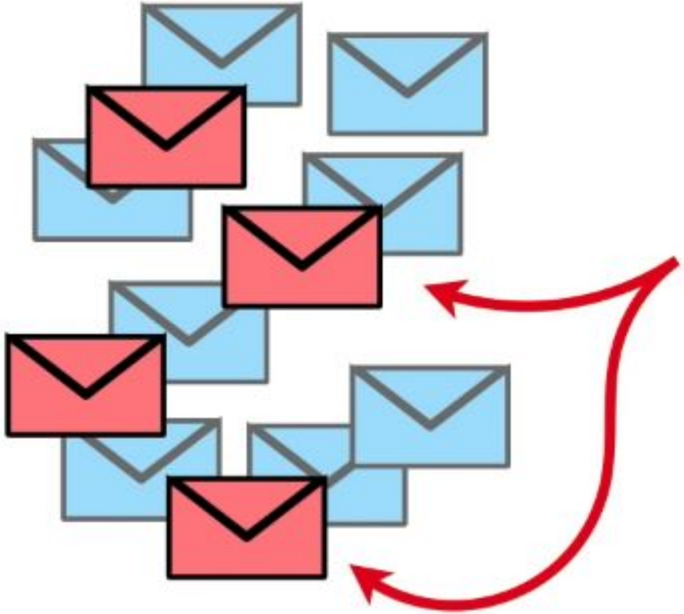
- La probabilidad de que un mail que contenga la palabra viagra sea spam es del 0.8.

Spam o Normal Emails



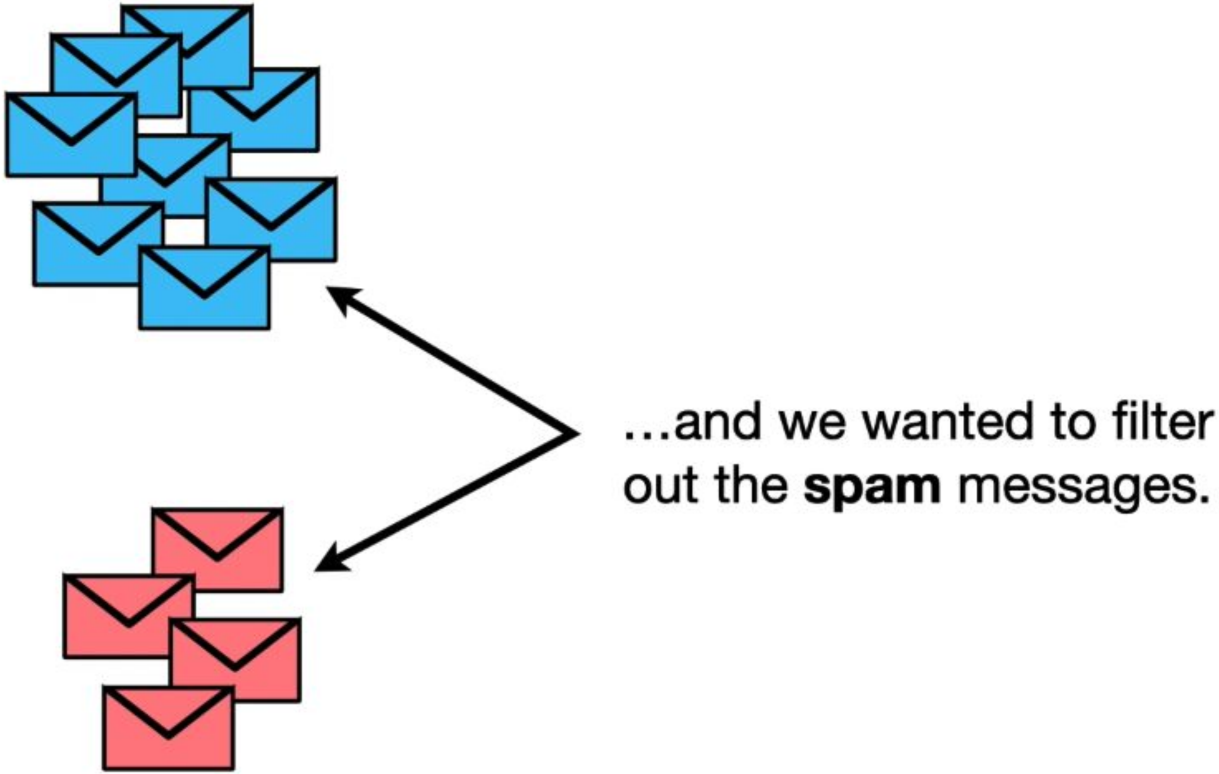
Imagine we received **normal messages** from friends and family…

Spam o Normal Emails

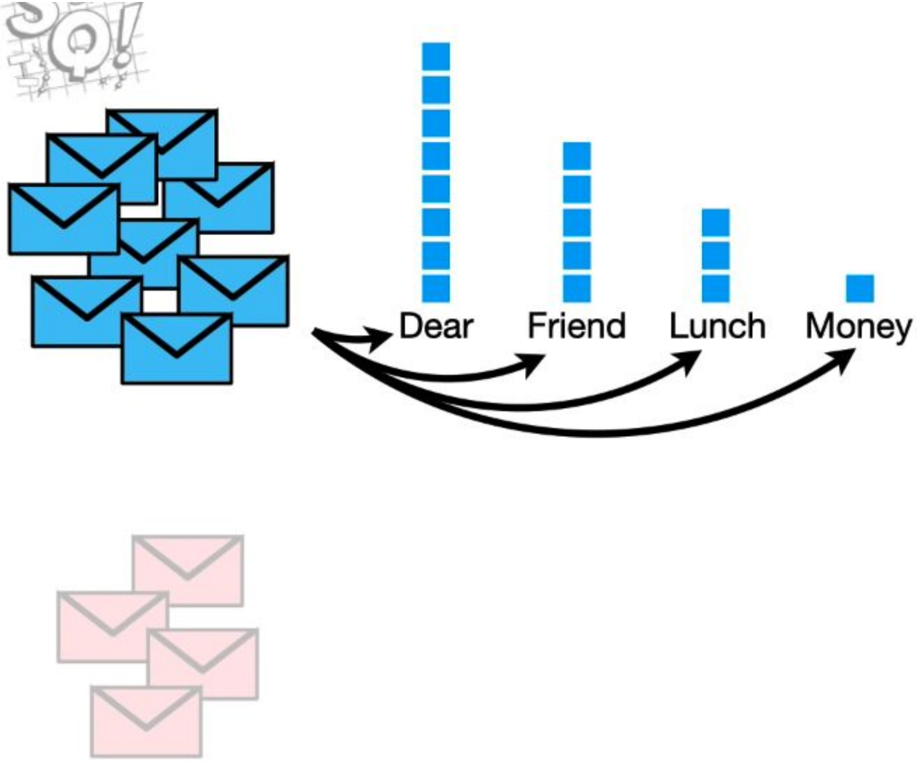


…and we also received **spam** (unwanted messages that are usually scams or unsolicited advertisements)…

Spam o Normal Emails

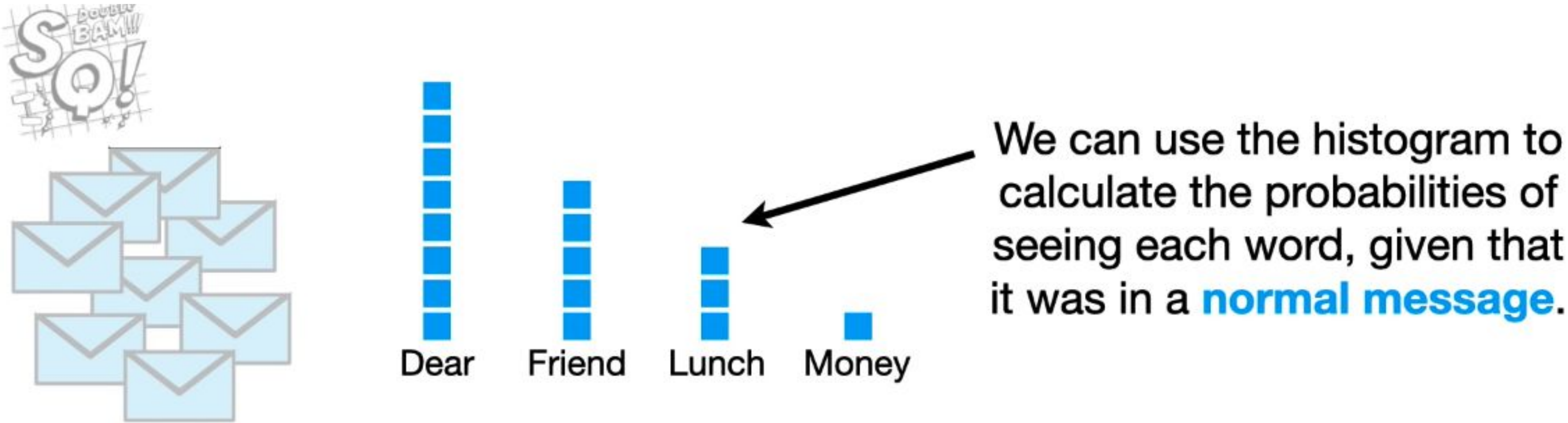


Spam o Normal Emails

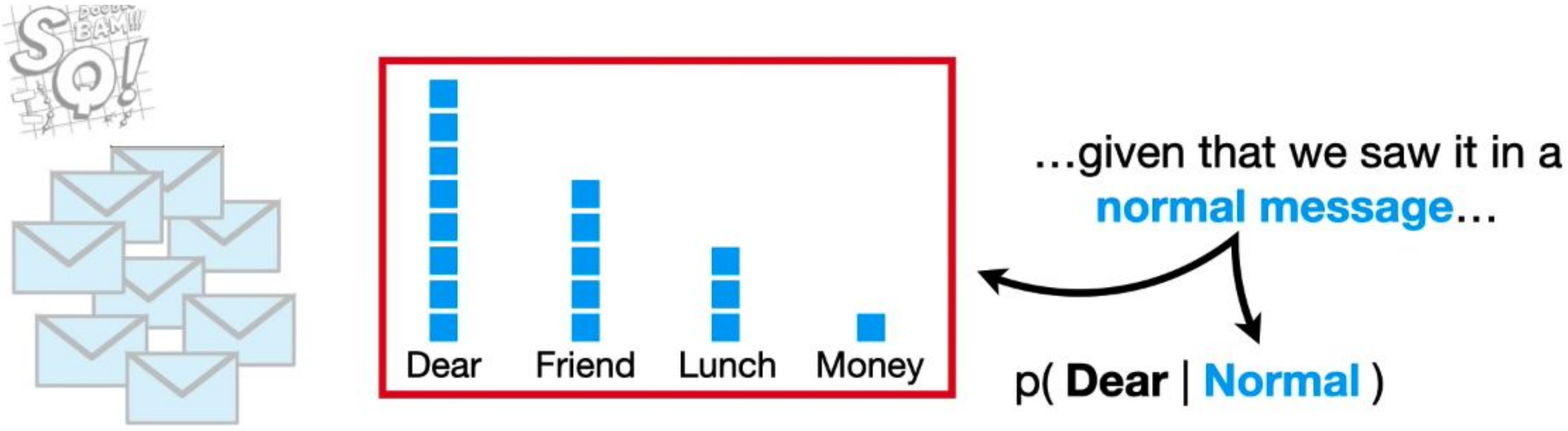


So, the first thing we do is make a **histogram** of all the words that occur in the normal messages from friends and family.

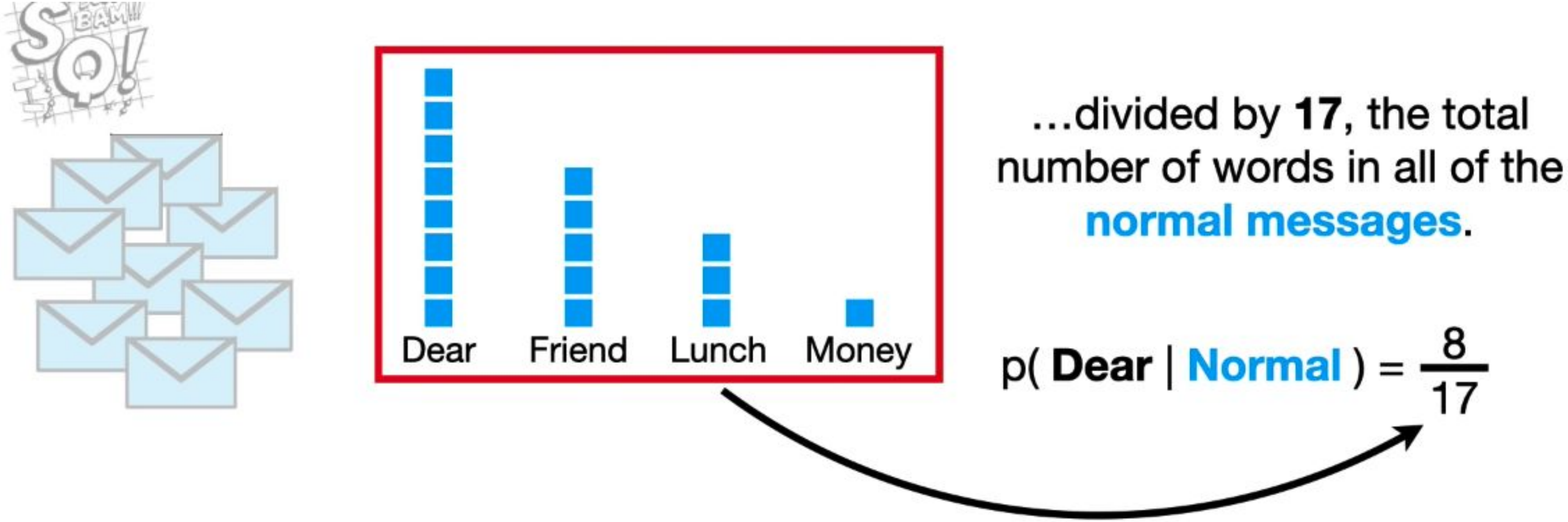
Spam o Normal Emails



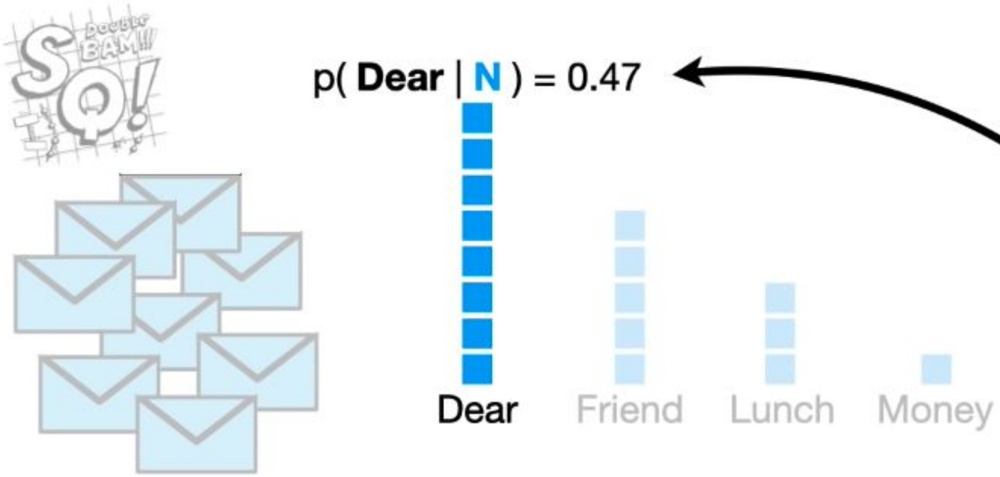
Spam o Normal Emails



Spam o Normal Emails



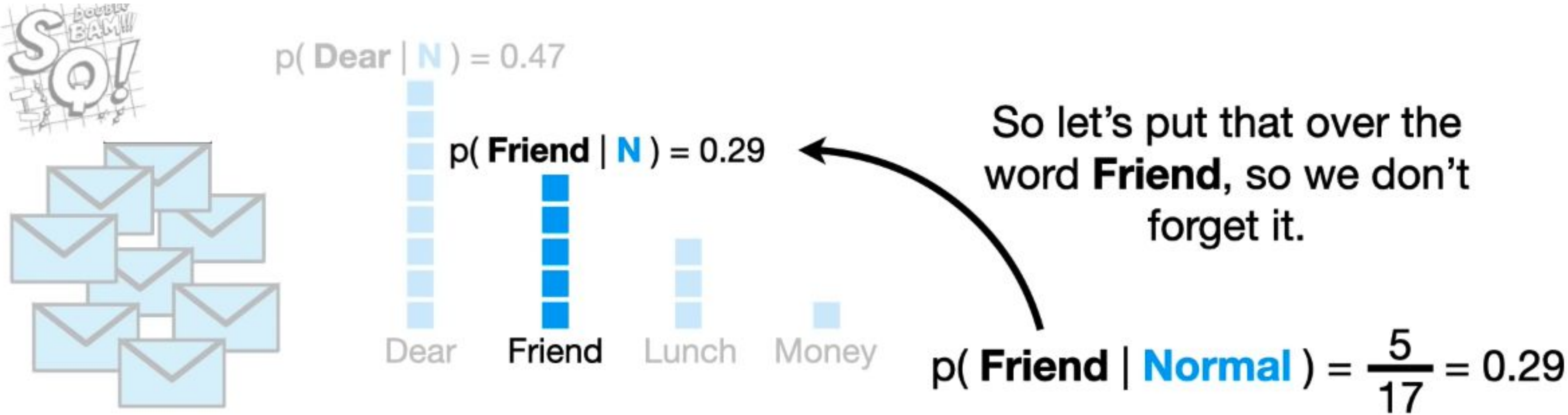
Spam o Normal Emails



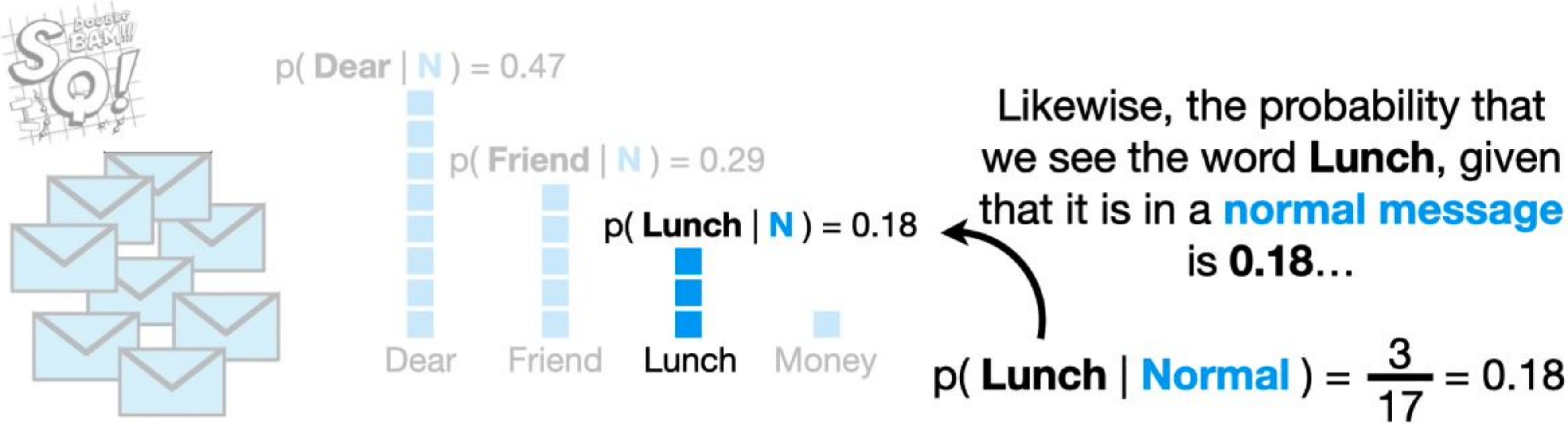
So let's put that over the word **Dear**, so we don't forget it.

$$p(\textbf{Dear} \mid \textbf{Normal}) = \frac{8}{17} = 0.47$$

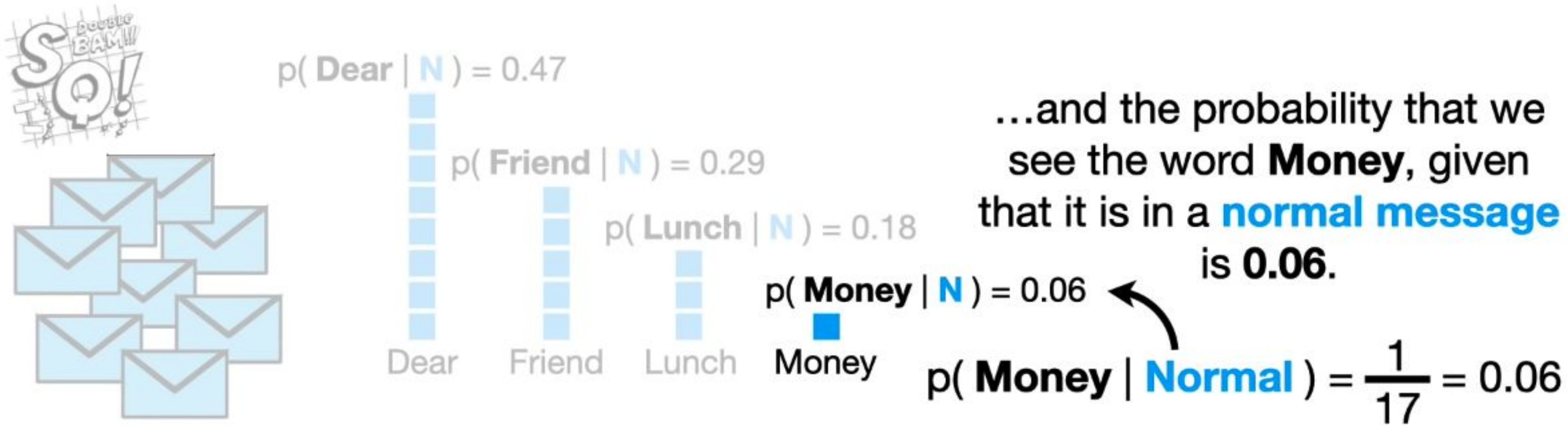
Spam o Normal Emails



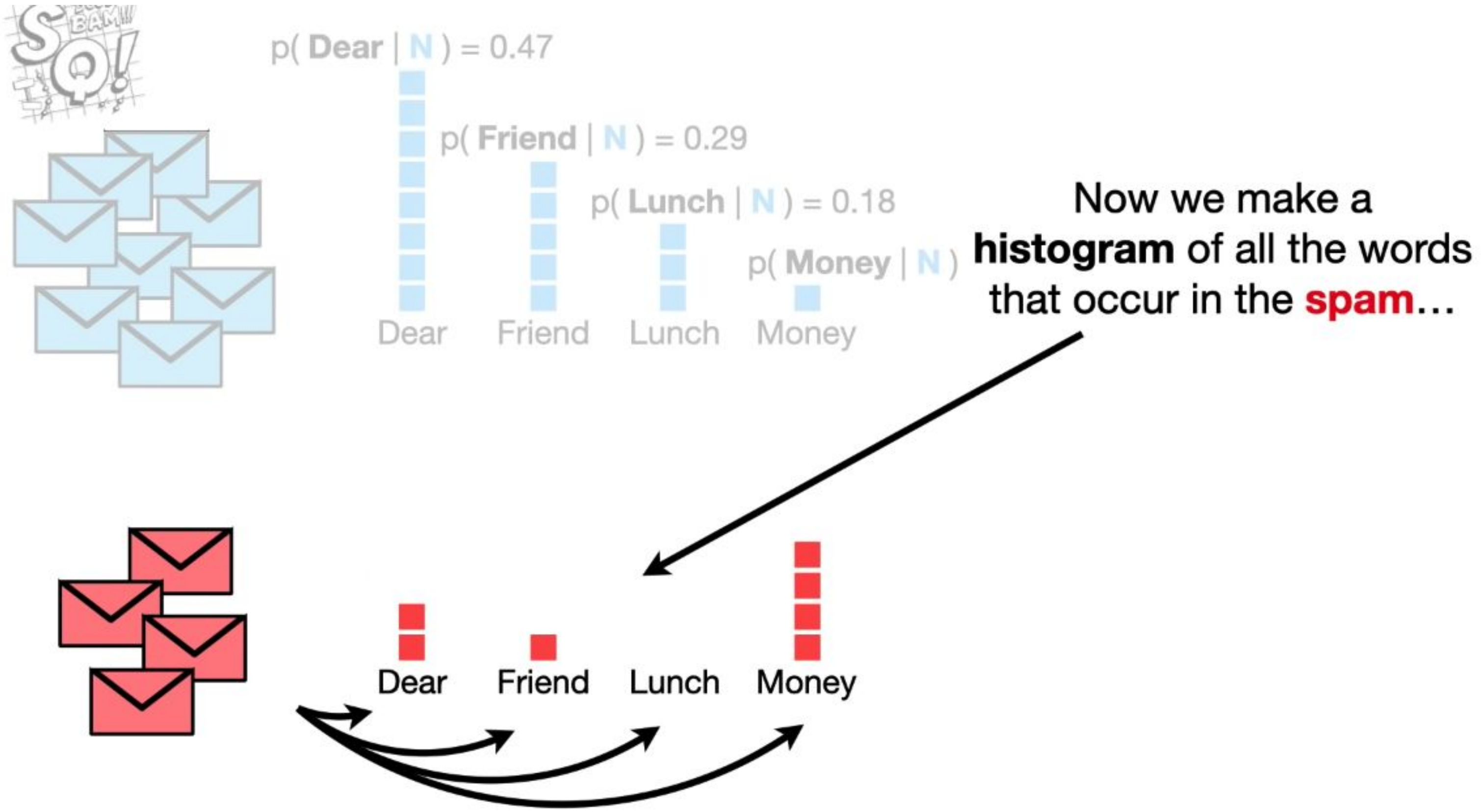
Spam o Normal Emails



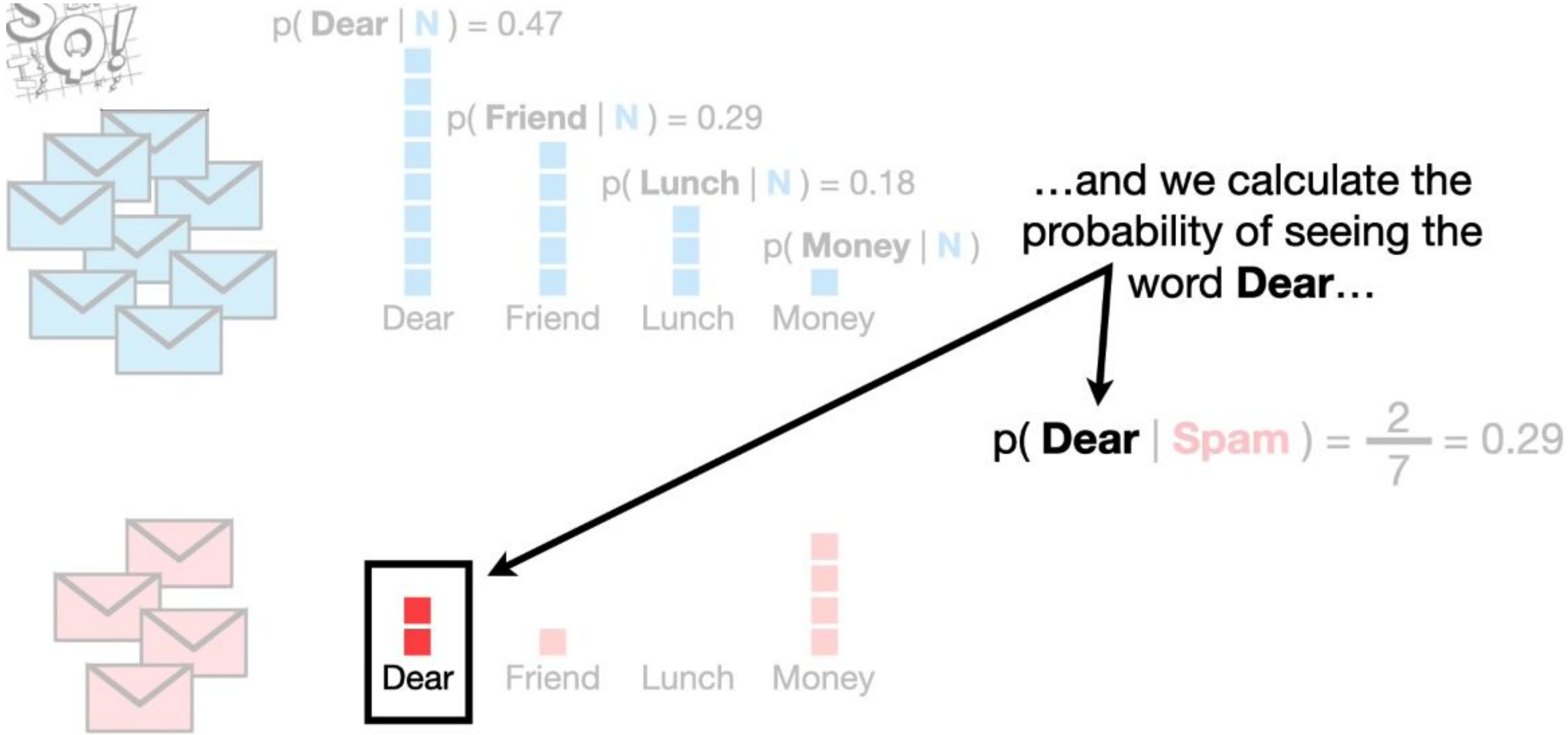
Spam o Normal Emails



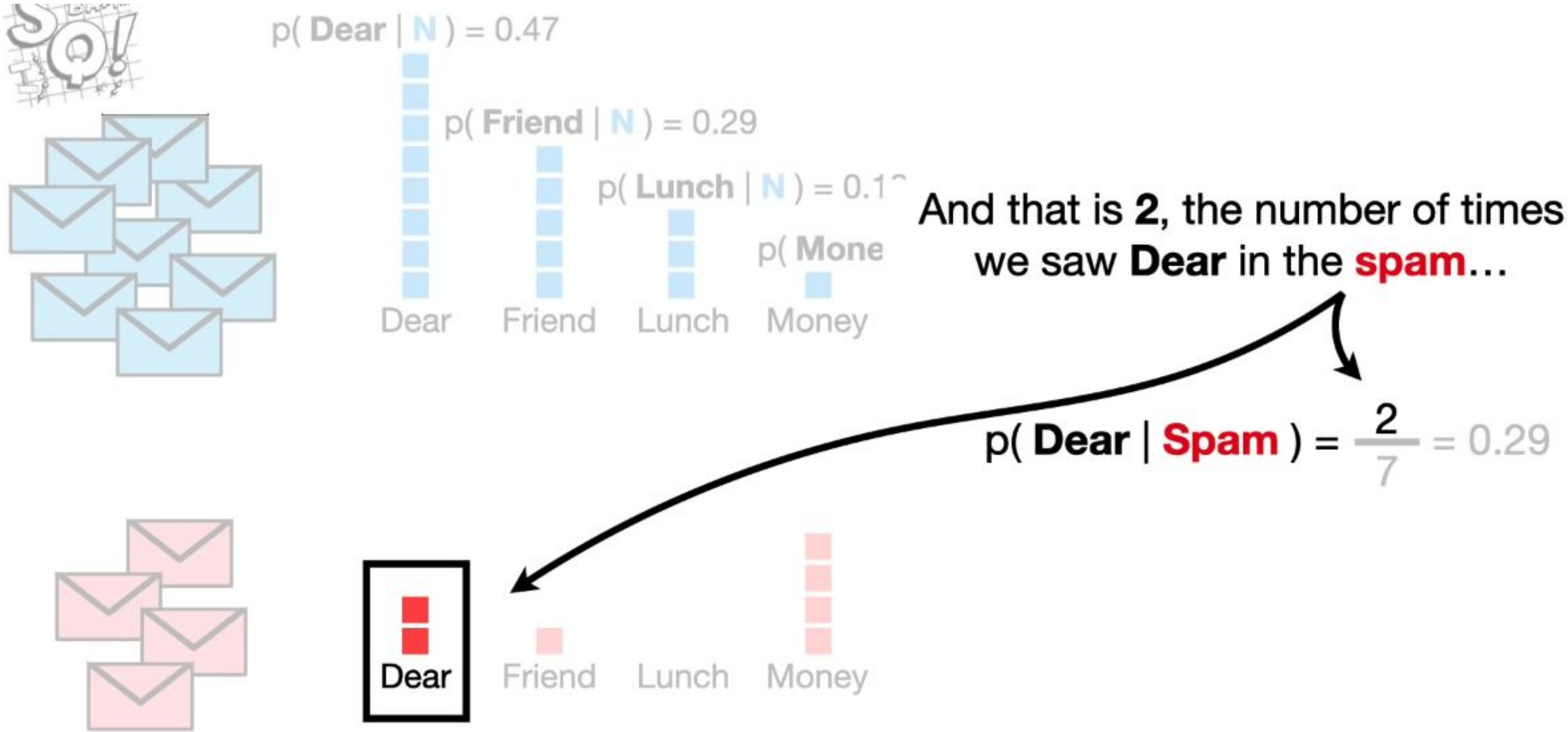
Spam o Normal Emails



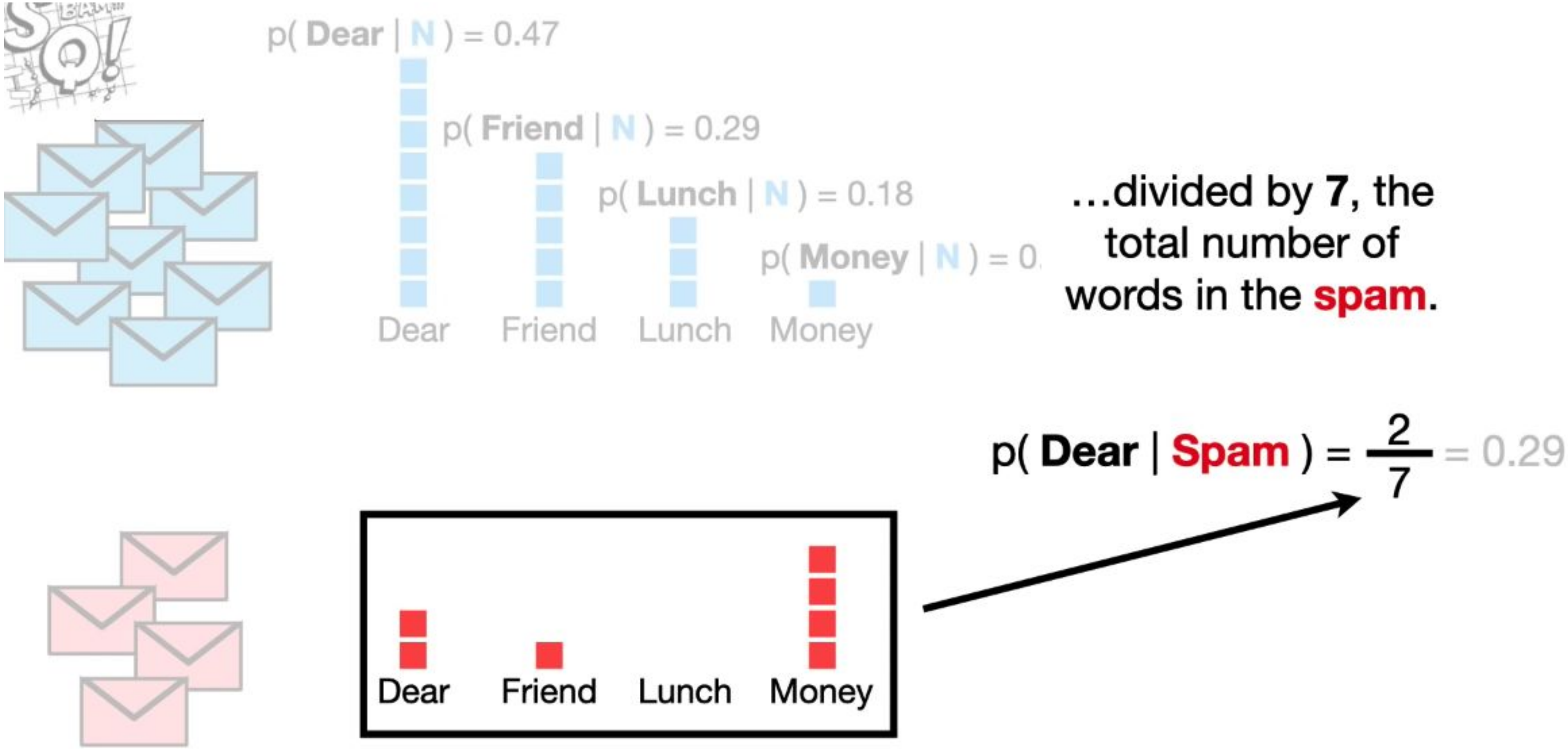
Spam o Normal Emails



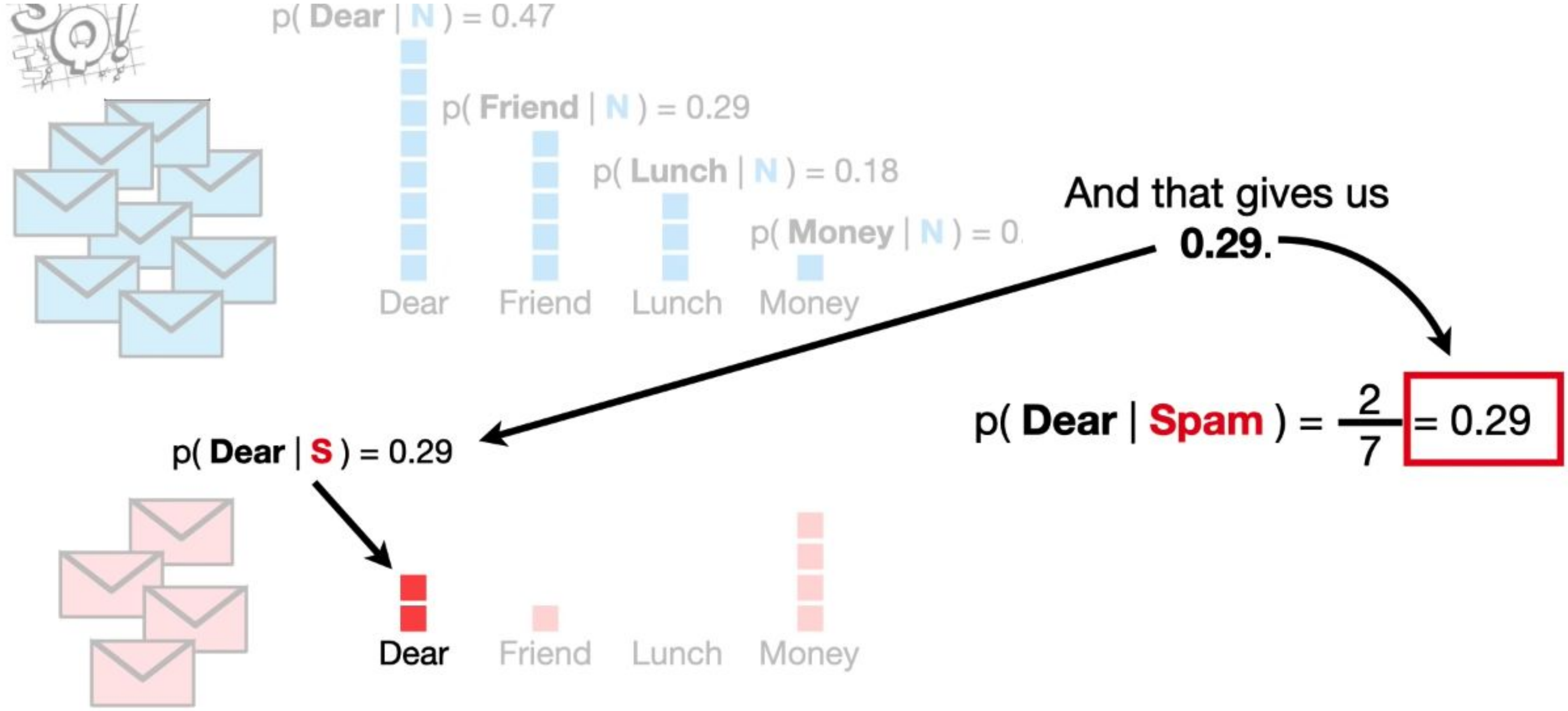
Spam o Normal Emails



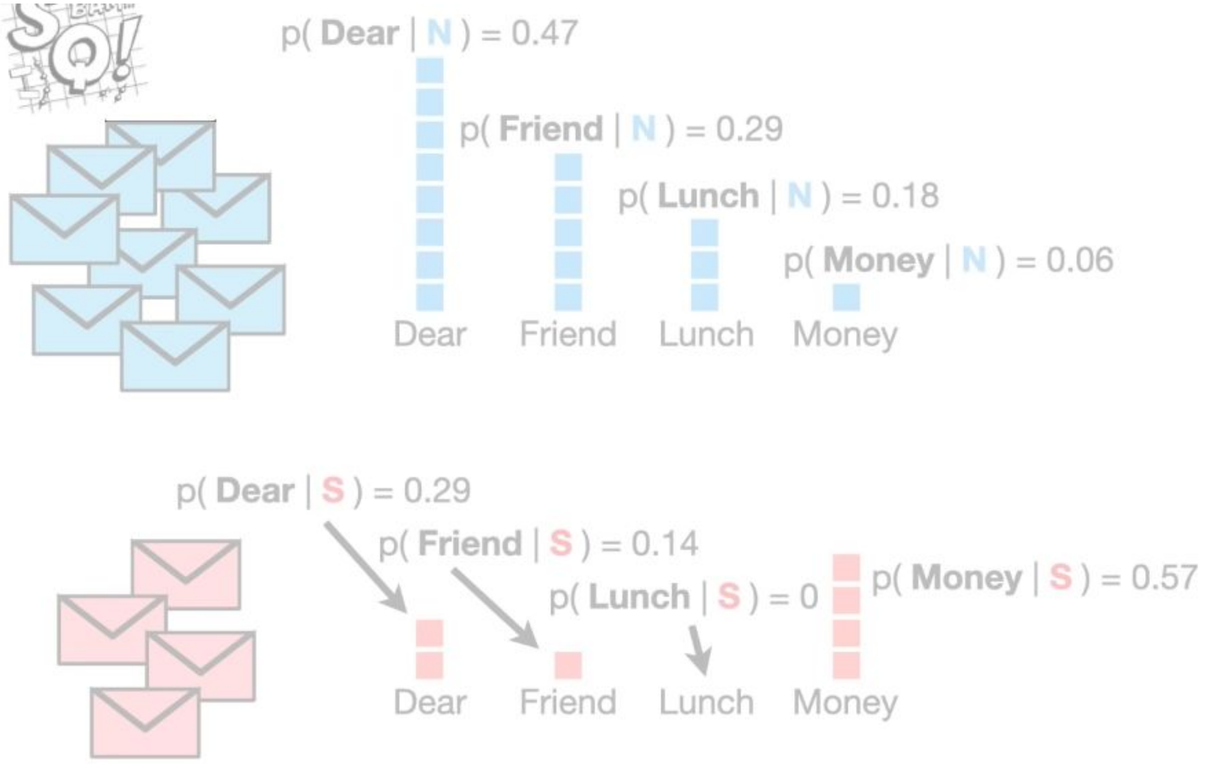
Spam o Normal Emails



Spam o Normal Emails



Spam o Normal Emails



Bam!

Spam o Normal Emails

p(**Dear** | **N**) = 0.47
p(**Friend** | **N**) = 0.29
p(**Lunch** | **N**) = 0.18
p(**Money** | **N**) = 0.06

Now, because these histograms are taking up a lot of space, let's get rid of them, but keep the probabilities.

p(**Dear** | **S**) = 0.29
p(**Friend** | **S**) = 0.14
p(**Lunch** | **S**) = 0.00
p(**Money** | **S**) = 0.57

Spam o Normal Emails

p(**Dear** | **N**) = 0.47
p(**Friend** | **N**) = 0.29
p(**Lunch** | **N**) = 0.18
p(**Money** | **N**) = 0.06

p(**Dear** | **S**) = 0.29
p(**Friend** | **S**) = 0.14
p(**Lunch** | **S**) = 0.00
p(**Money** | **S**) = 0.57

Terminology Alert!!!

Because we have calculated the probabilities of discrete, individual words, and not the probability of something continuous, like weight or height, these **Probabilities** are also called **Likelihoods.**

Spam o Normal Emails

p(**Dear** | **N**) = 0.47
p(**Friend** | **N**) = 0.29
p(**Lunch** | **N**) = 0.18
p(**Money** | **N**) = 0.06

p(**Dear** | **S**) = 0.29
p(**Friend** | **S**) = 0.14
p(**Lunch** | **S**) = 0.00
p(**Money** | **S**) = 0.57

I mention this because some tutorials say these are **Probabilities**, and others say they are **Likelihoods**.

In this case, the terms are interchangeable. So don't sweat it.

Spam o Normal Emails

$p(\textbf{Dear} \mid \textbf{N}) = 0.47$

$p(\textbf{Friend} \mid \textbf{N}) = 0.29$

$p(\textbf{Lunch} \mid \textbf{N}) = 0.18$

$p(\textbf{Money} \mid \textbf{N}) = 0.06$

$p(\textbf{Dear} \mid \textbf{S}) = 0.29$

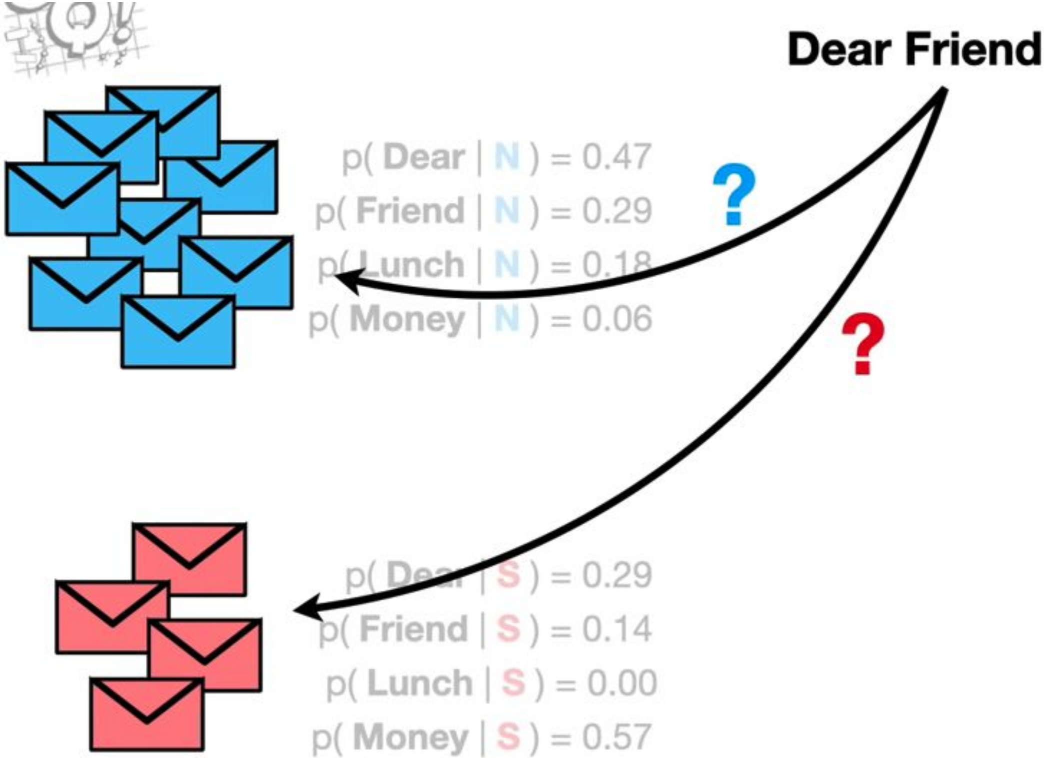
$p(\textbf{Friend} \mid \textbf{S}) = 0.14$

$p(\textbf{Lunch} \mid \textbf{S}) = 0.00$

$p(\textbf{Money} \mid \textbf{S}) = 0.57$

We'll talk more about **Probabilities** vs **Likelihoods** when we talk about **Gaussian Naive Bayes** in the follow up **'Quest.**

Spam o Normal Emails



And we want to decide if is a **normal message** or **spam**.

Spam o Normal Emails

$p(\textbf{Dear} \mid N) = 0.47$

$p(\textbf{Friend} \mid N) = 0.29$

$p(\textbf{Lunch} \mid N) = 0.18$

$p(\textbf{Money} \mid N) = 0.06$

$p(\textbf{Dear} \mid S) = 0.29$

$p(\textbf{Friend} \mid S) = 0.14$

$p(\textbf{Lunch} \mid S) = 0.00$

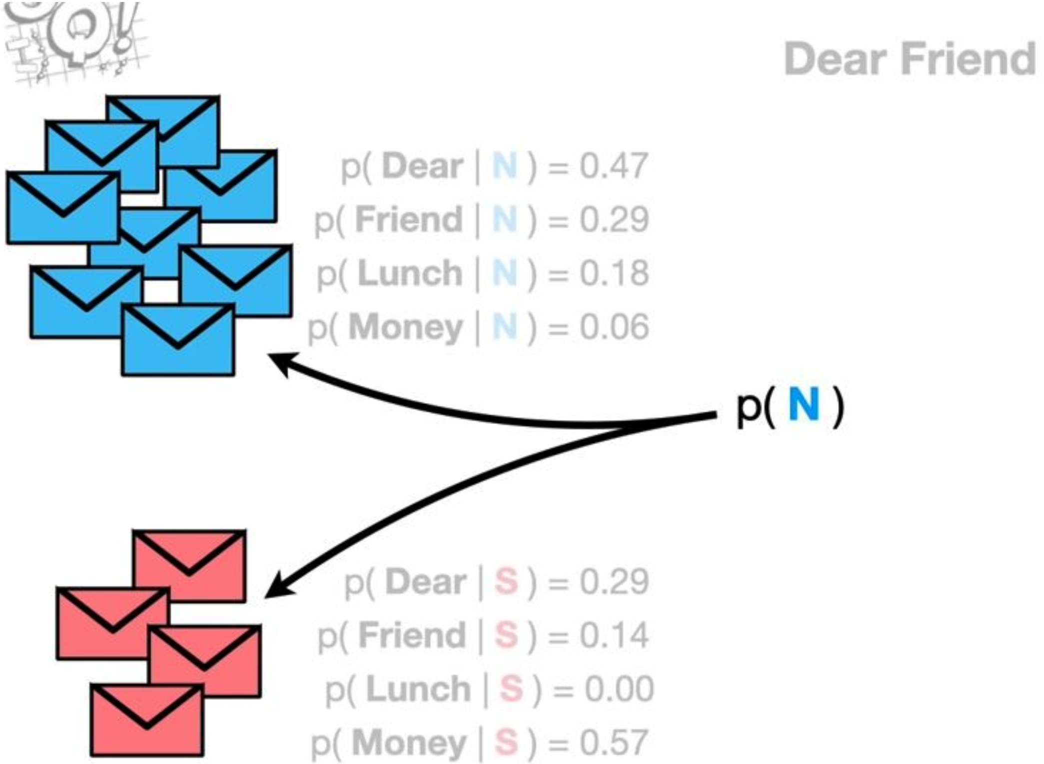
$p(\textbf{Money} \mid S) = 0.57$

Dear Friend

We start with an initial guess about the probability that any message, regardless of what it says, is a **normal message**.

$p(N)$

Spam o Normal Emails



The guess can be any probability that we want, but a common guess is estimated from the training data.

Spam o Normal Emails

Dear Friend

$p(\textbf{Dear} \mid \textbf{N}) = 0.47$

$p(\textbf{Friend} \mid \textbf{N}) = 0.29$

$p(\textbf{Lunch} \mid \textbf{N}) = 0.18$

$p(\textbf{Money} \mid \textbf{N}) = 0.06$

For example, since **8** of the **12** messages are **normal messages**, our initial guess will be **0.67**.

$p(\textbf{N}) =$

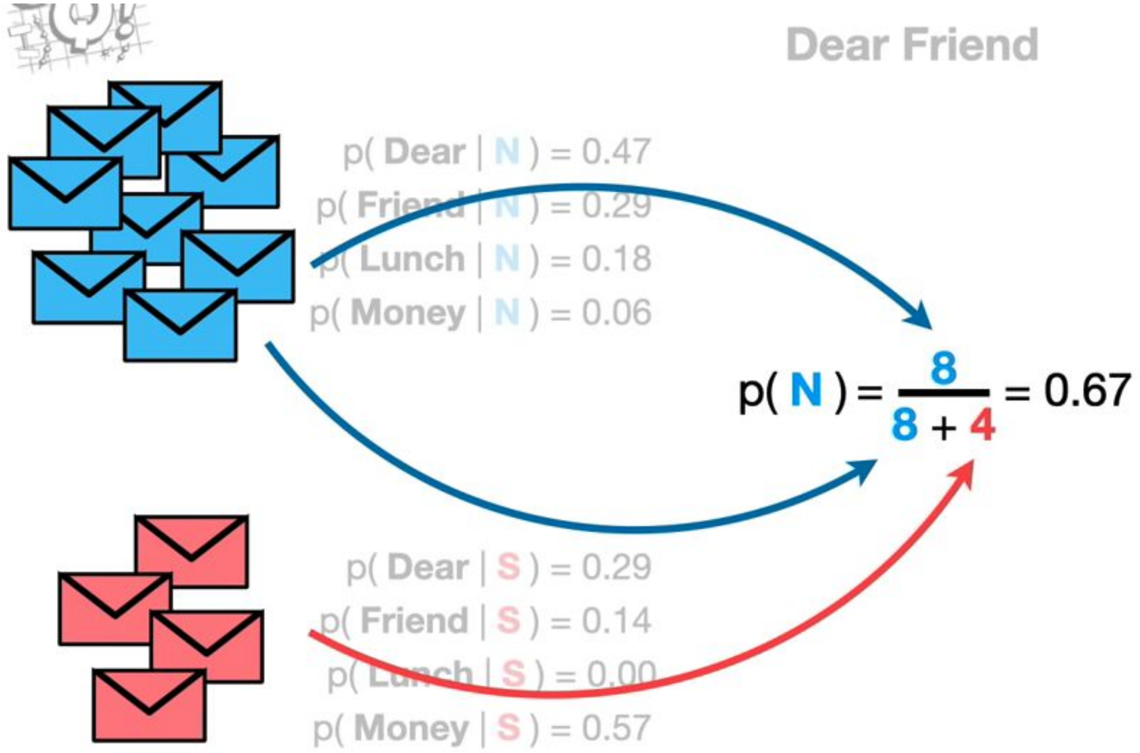
$p(\textbf{Dear} \mid \textbf{S}) = 0.29$

$p(\textbf{Friend} \mid \textbf{S}) = 0.14$

$p(\textbf{Lunch} \mid \textbf{S}) = 0.00$

$p(\textbf{Money} \mid \textbf{S}) = 0.57$

Spam o Normal Emails



For example, since **8** of the **12** messages are **normal messages**, our initial guess will be **0.67**.

Spam o Normal Emails

$p(\text{Dear} \mid N) = 0.47$

$p(\text{Friend} \mid N) = 0.29$

$p(\text{Lunch} \mid N) = 0.18$

$p(\text{Money} \mid N) = 0.06$

$p(N) = 0.67$

$p(\text{Dear} \mid S) = 0.29$

$p(\text{Friend} \mid S) = 0.14$

$p(\text{Lunch} \mid S) = 0.00$

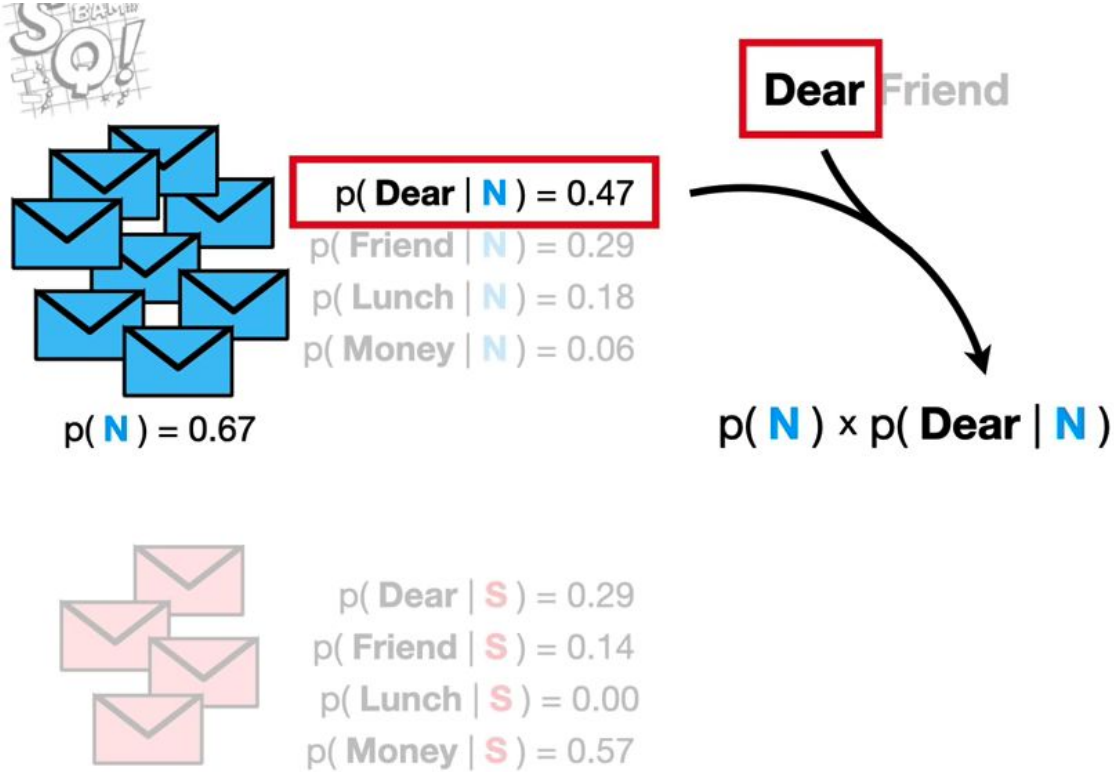
$p(\text{Money} \mid S) = 0.57$

Dear Friend

TERMINOLOGY ALERT!!!!

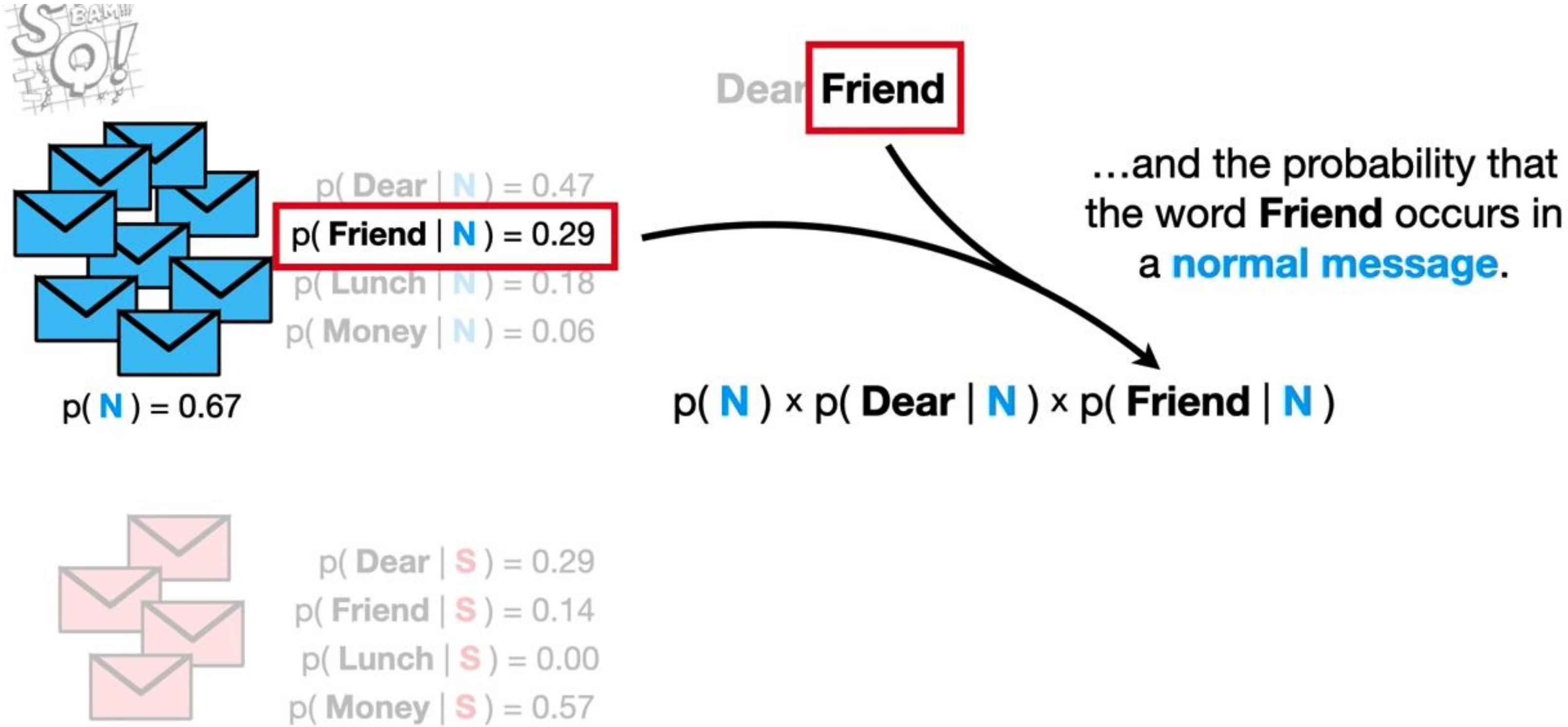
$p(N)$ ← The initial guess that we observe a **Normal** messages is called a **Prior Probability.**

Spam o Normal Emails



Now we multiply that initial guess by the probability that the word **Dear** occurs in a **normal message**…

Spam o Normal Emails



Spam o Normal Emails

Dear Friend

p(Dear | N) = 0.47
p(Friend | N) = 0.29
p(Lunch | N) = 0.18
p(Money | N) = 0.06

$\boxed{p(N) = 0.67}$

Now we just plug in the values that we worked out earlier and do the math…

$$p(N) \times p(\textbf{Dear} \mid N) \times p(\textbf{Friend} \mid N)$$

p(Dear | S) = 0.29
p(Friend | S) = 0.14
p(Lunch | S) = 0.00
p(Money | S) = 0.57

Spam o Normal Emails

p(**Dear** | **N**) = 0.47
p(**Friend** | **N**) = 0.29
p(**Lunch** | **N**) = 0.18
p(**Money** | **N**) = 0.06

p(**N**) = 0.67

p(**Dear** | **S**) = 0.29
p(**Friend** | **S**) = 0.14
p(**Lunch** | **S**) = 0.00
p(**Money** | **S**) = 0.57

Dear Friend

We can think of **0.09** as the score that **Dear Friend** gets if it is a **Normal Message**.

$$0.67 \times 0.47 \times 0.29 \boxed{= 0.09}$$

Spam o Normal Emails

p(**Dear** | **N**) = 0.47
p(**Friend** | **N**) = 0.29
p(**Lunch** | **N**) = 0.18
p(**Money** | **N**) = 0.06

p(**N**) = 0.67

Dear Friend

However, technically, it is *proportional* to the probability that the message is **normal**, given that it says **Dear Friend**.

$0.67 \times 0.47 \times 0.29 = 0.09 \propto p(\mathbf{N} \mid \textbf{Dear Friend})$

p(**Dear** | **S**) = 0.29
p(**Friend** | **S**) = 0.14
p(**Lunch** | **S**) = 0.00
p(**Money** | **S**) = 0.57

Spam o Normal Emails

$p(\text{ N } | \text{ Dear Friend }) \propto 0.09$

Dear Friend

Now, just like we did before, We start with an initial guess about the probability that any message, regardless of what it says, is **spam**.

$p(\text{ Dear } | \text{ N }) = 0.47$

$p(\text{ Friend } | \text{ N }) = 0.29$

$p(\text{ Lunch } | \text{ N }) = 0.18$

$p(\text{ Money } | \text{ N }) = 0.06$

$p(\text{ N }) = 0.67$

$p(\text{ S })$

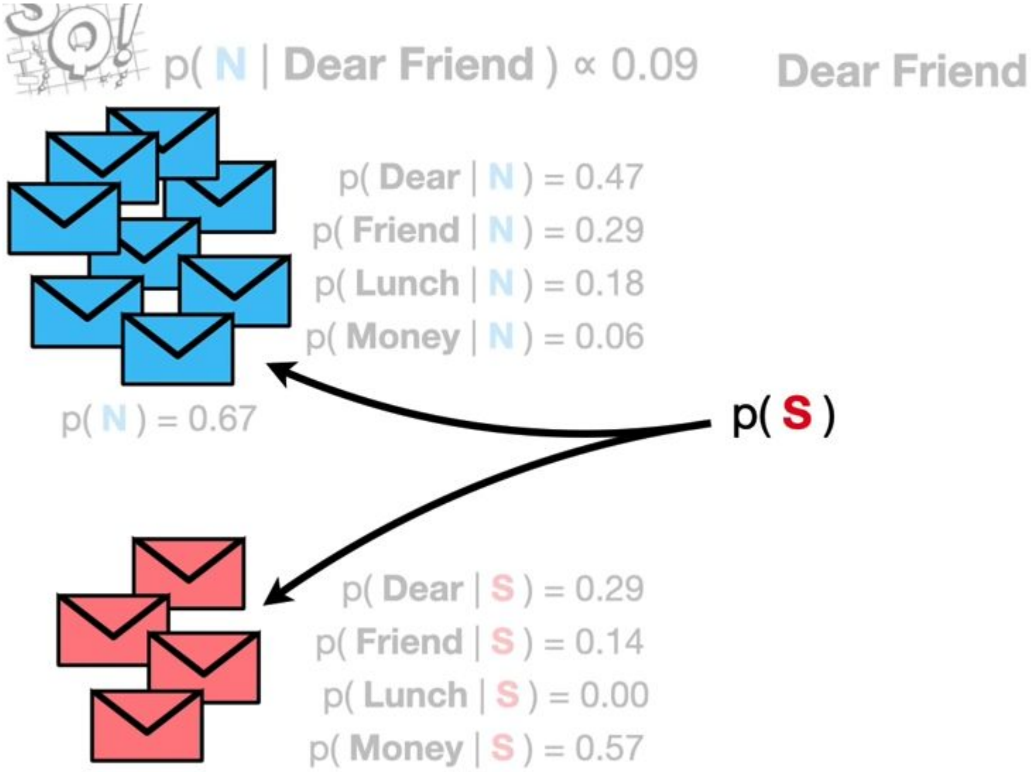
$p(\text{ Dear } | \text{ S }) = 0.29$

$p(\text{ Friend } | \text{ S }) = 0.14$

$p(\text{ Lunch } | \text{ S }) = 0.00$

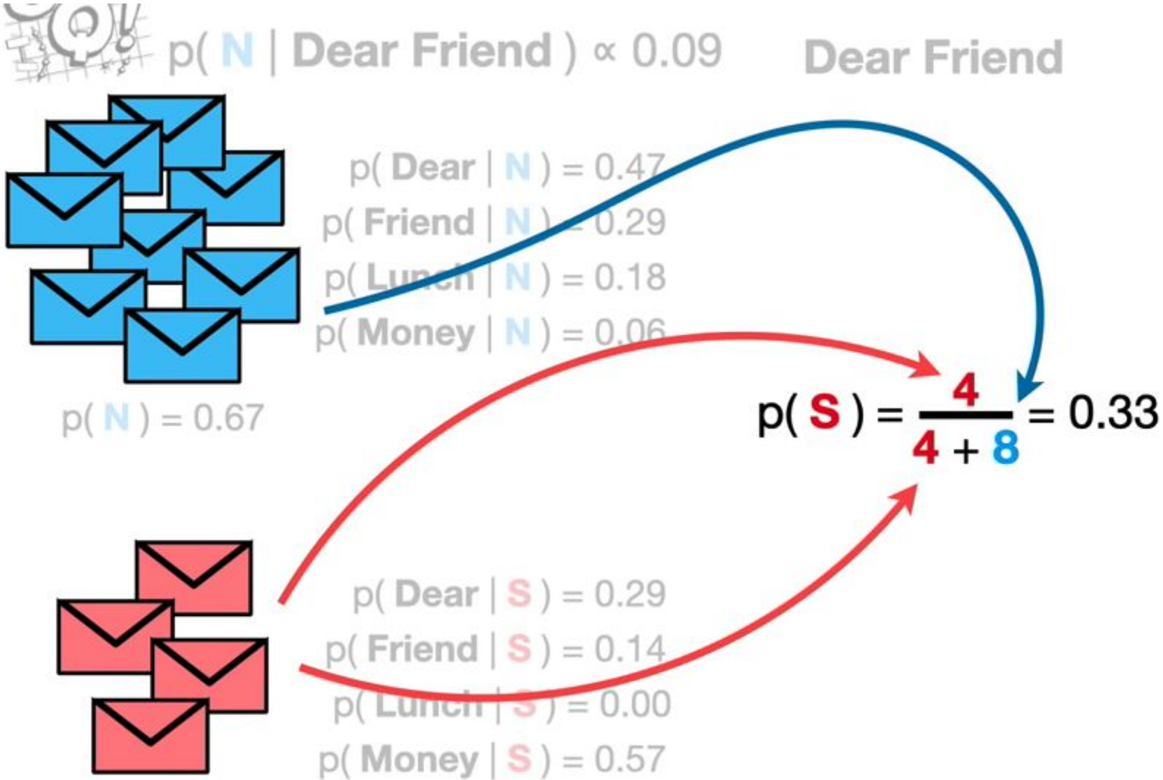
$p(\text{ Money } | \text{ S }) = 0.57$

Spam o Normal Emails



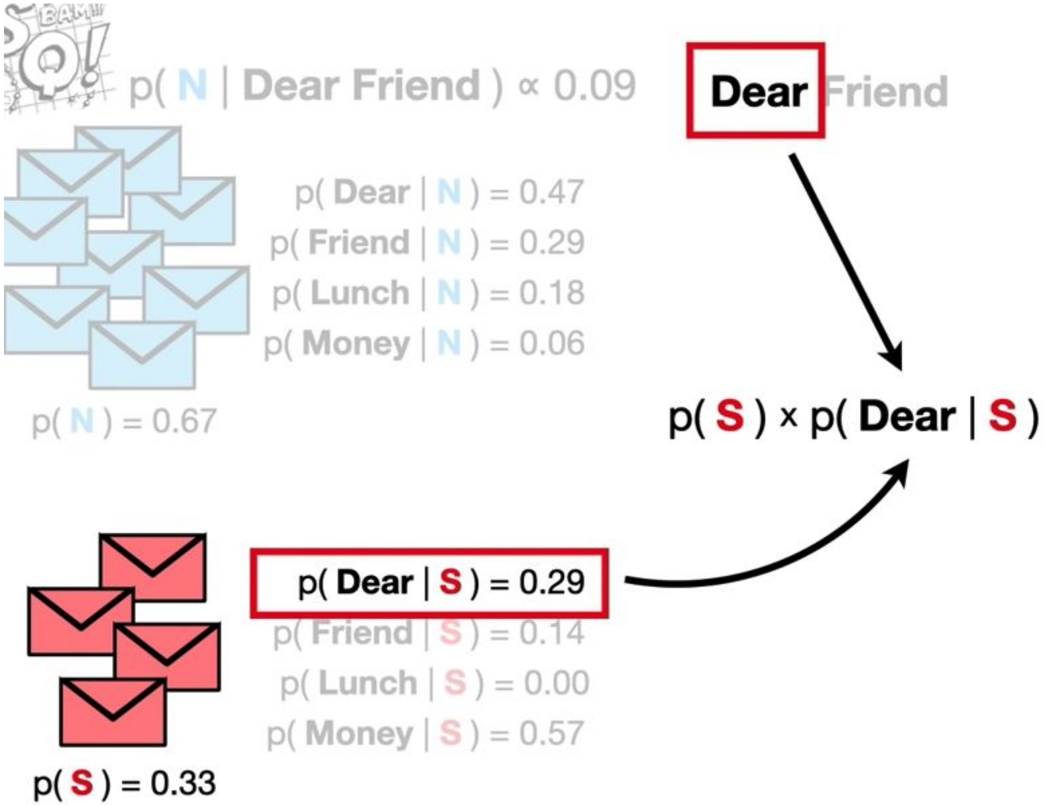
And just like before, the guess can be any probability that we want, but a common guess is estimated from the training data.

Spam o Normal Emails



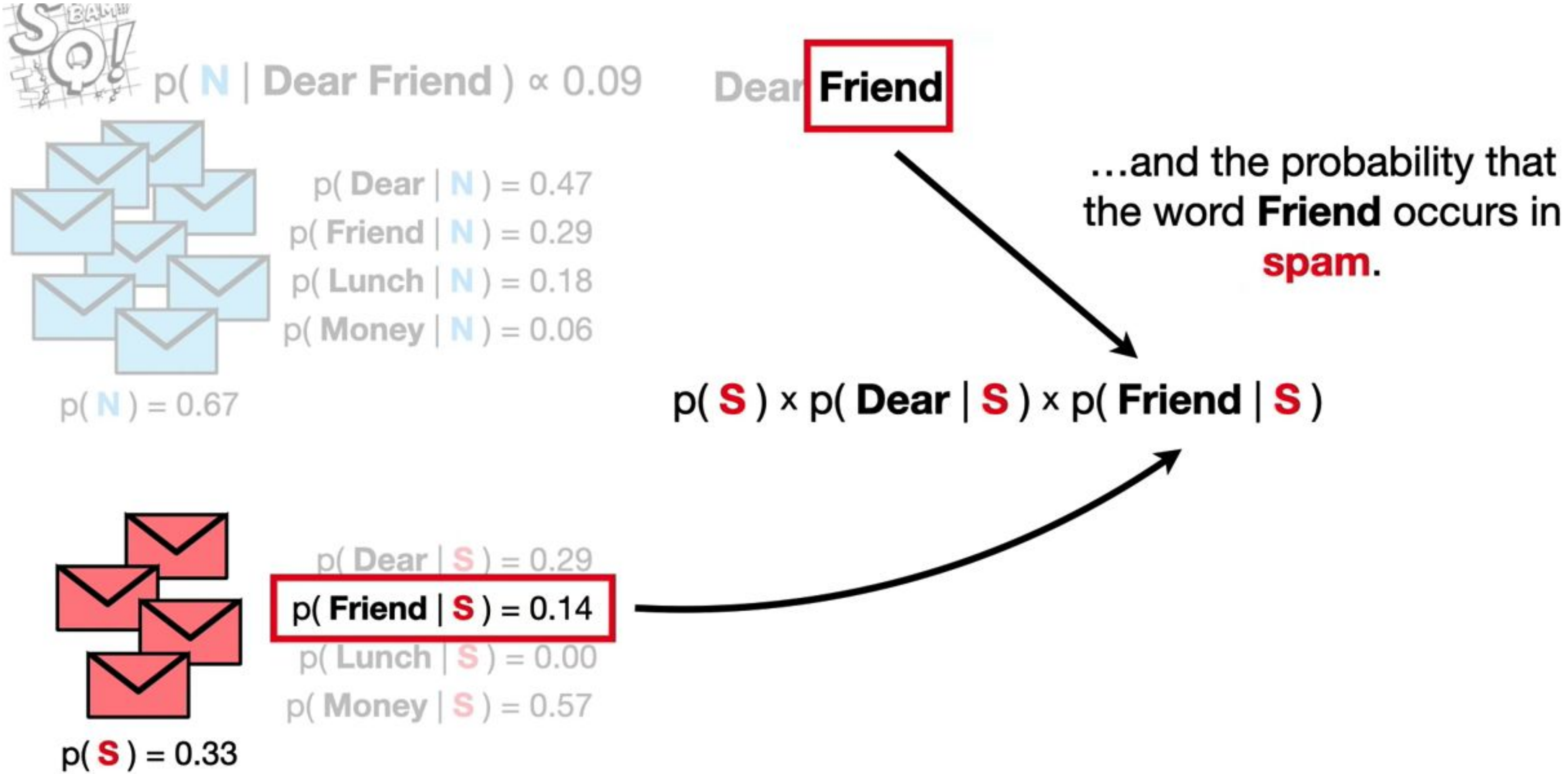
And since **4** of the **12** messages are **spam**, our initial guess will be **0.33**.

Spam o Normal Emails

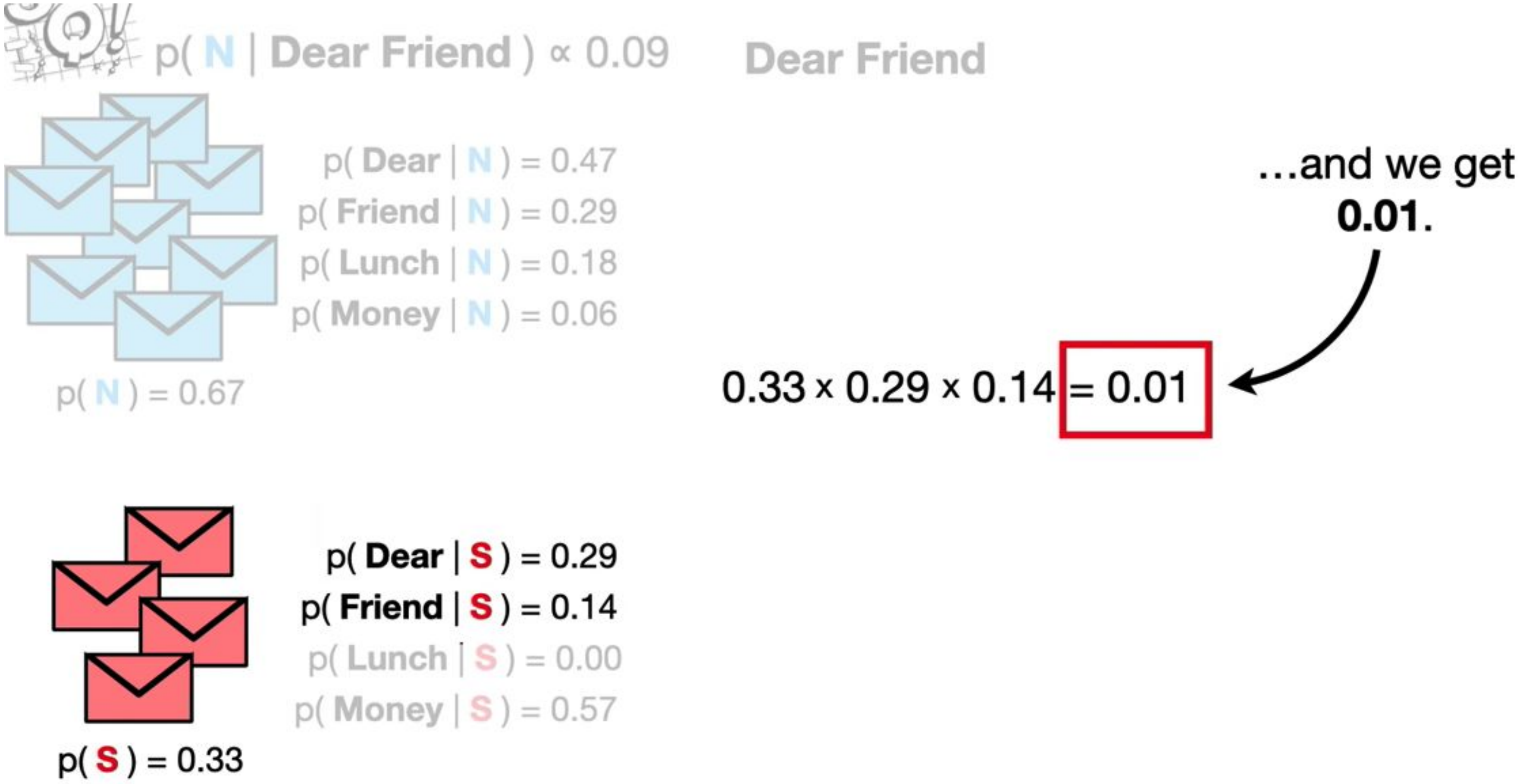


Now we multiply that initial guess by the probability that the word **Dear** occurs in **spam**…

Spam o Normal Emails



Spam o Normal Emails



Spam o Normal Emails

p(N | Dear Friend) ∝ 0.09

Dear Friend

p(**Dear** | N) = 0.47
p(**Friend** | N) = 0.29
p(**Lunch** | N) = 0.18
p(**Money** | N) = 0.06

p(N) = 0.67

Like before, we can think of **0.01** as the score that **Dear Friend** gets if it is **Spam**.

$0.33 \times 0.29 \times 0.14 \boxed{= 0.01}$

p(**Dear** | **S**) = 0.29
p(**Friend** | **S**) = 0.14
p(**Lunch** | S) = 0.00
p(**Money** | S) = 0.57

p(**S**) = 0.33

Spam o Normal Emails

p(N | **Dear Friend**) ∝ 0.09

Dear Friend

p(**Dear** | N) = 0.47
p(**Friend** | N) = 0.29
p(**Lunch** | N) = 0.18
p(**Money** | N) = 0.06

p(N) = 0.67

However, technically, it is *proportional* to the probability that the message is **spam** given that it says **Dear Friend.**

0.33 x 0.29 x 0.14 = 0.01 ∝ p(**S** | **Dear Friend**)

p(**Dear** | S) = 0.29
p(**Friend** | S) = 0.14
p(**Lunch** | S) = 0.00
p(**Money** | S) = 0.57

p(S) = 0.33

Spam o Normal Emails

$p(\,\mathbf{N} \mid \textbf{Dear Friend}\,) \propto \boxed{0.09}$

Dear Friend

$p(\,\textbf{Dear} \mid \mathbf{N}\,) = 0.47$

$p(\,\textbf{Friend} \mid \mathbf{N}\,) = 0.29$

$p(\,\textbf{Lunch} \mid \mathbf{N}\,) = 0.18$

$p(\,\textbf{Money} \mid \mathbf{N}\,) = 0.06$

$p(\,\mathbf{N}\,) = 0.67$

And because the score we got for **Normal Message**, **0.09**…

$0.33 \times 0.29 \times 0.14 = 0.01 \propto p(\,\mathbf{S} \mid \textbf{Dear Friend}\,)$

$p(\,\textbf{Dear} \mid \mathbf{S}\,) = 0.29$

$p(\,\textbf{Friend} \mid \mathbf{S}\,) = 0.14$

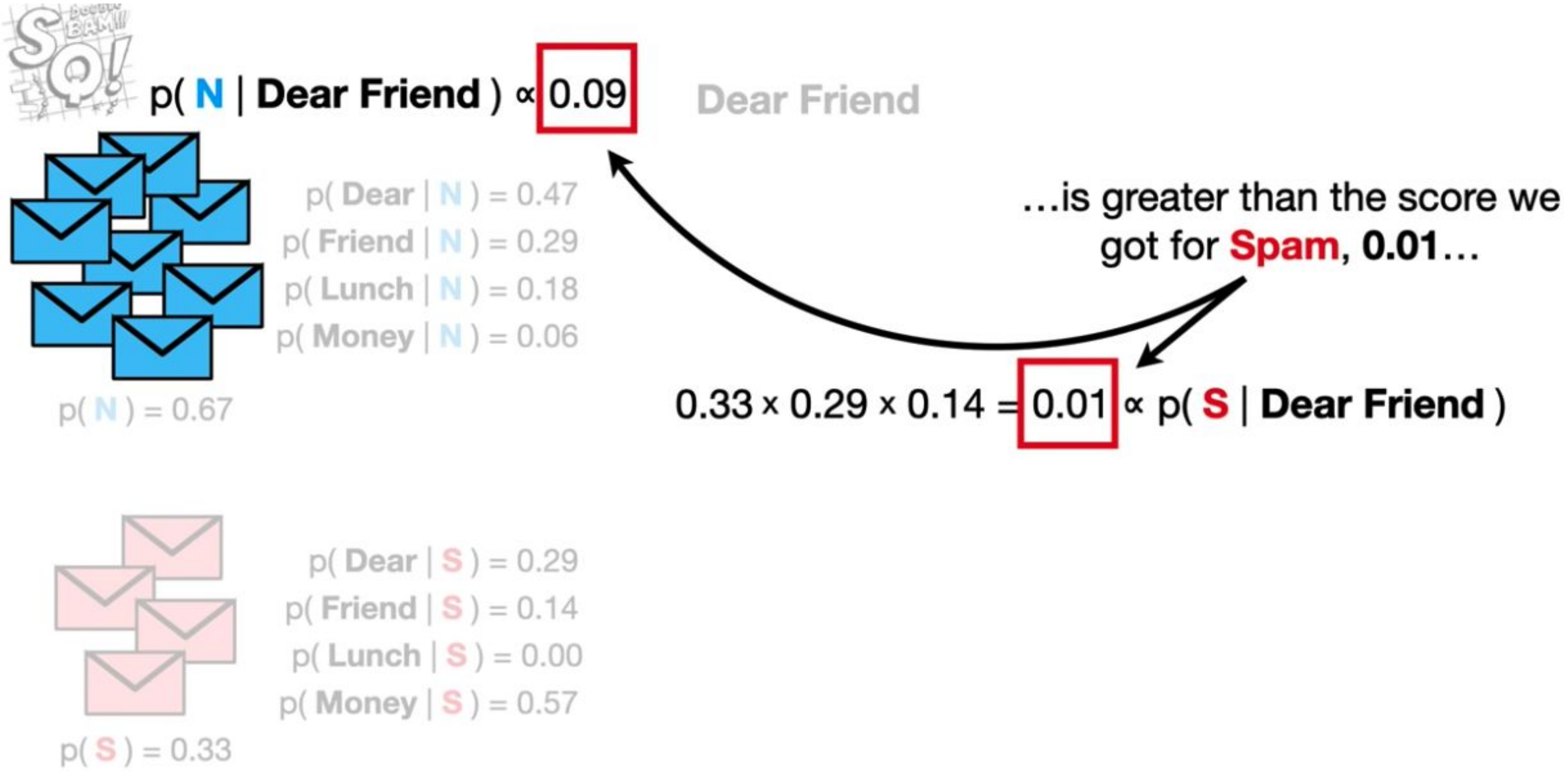
$p(\,\textbf{Lunch} \mid \mathbf{S}\,) = 0.00$

$p(\,\textbf{Money} \mid \mathbf{S}\,) = 0.57$

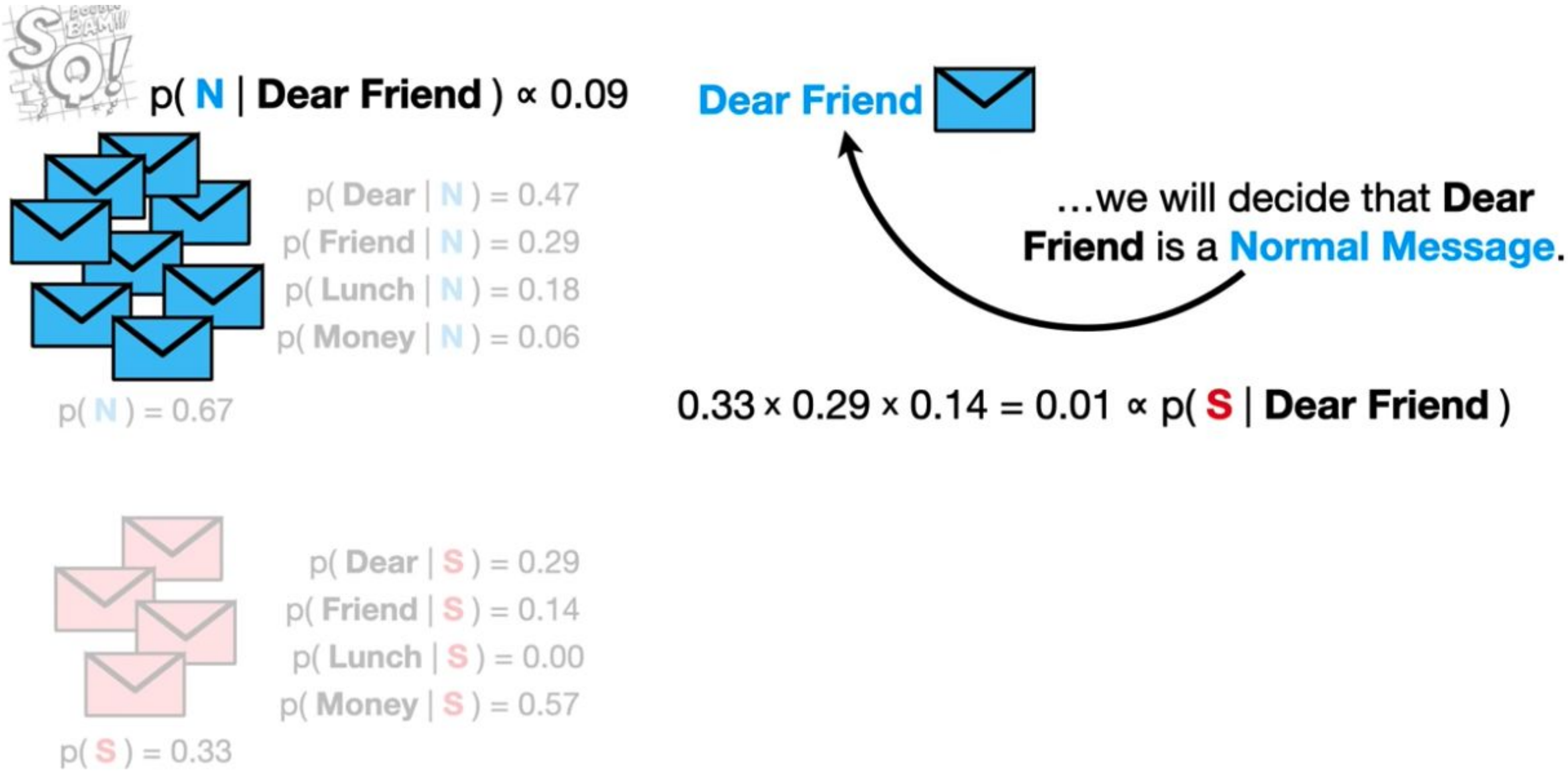
$p(\,\mathbf{S}\,) = 0.33$

Extensions

Spam o Normal Emails



Spam o Normal Emails



Métricas

Métricas de Evaluación: Accuracy, Precision, Recall

Métricas de Evaluación: Accuracy, Precision, Recall

		Predicted		
		Positive	**Negative**	
Actual	**Positive**	True positive(TP)	False Negative(FN)	Sensitivity or Recall or True Positive Rate=TP/(TP+FN)
	Negative	False Positive (FP)	True Negative(TN)	Specificity or True Negative Rate=TN/(TN+FP)
		Precision or Positive Predictive Value=TP/(TP+FP)	Negative Predictive Value=FN/(FN+TN)	Accuracy=TP+TN/TP+TN+FP+FN

Thanks!

Any questions ?

You can find me at

- Twitter: @ruthy_root
- Email: ruth.chirinos@gmail.com