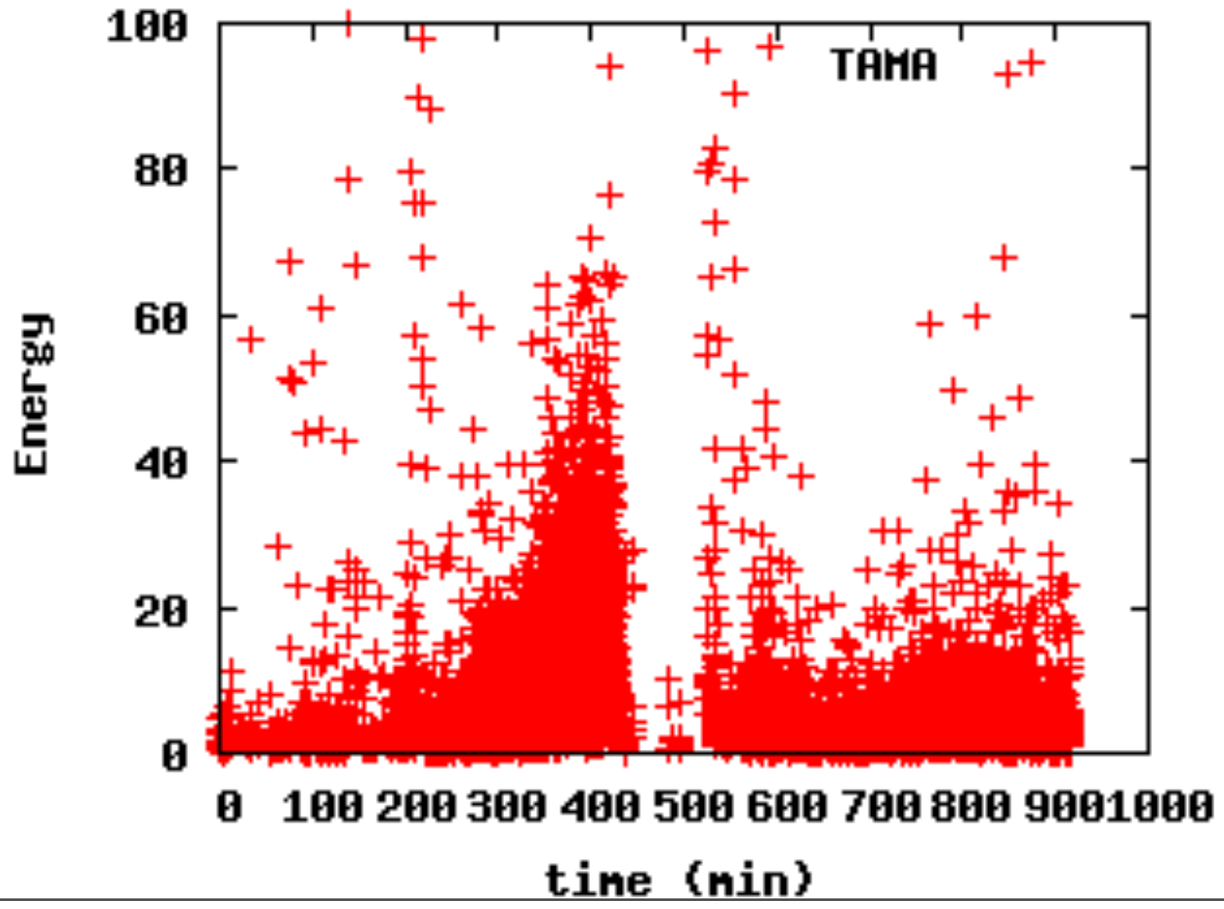
Spam Message from a GW detector

Kazuhiro Hayama

INDEX

How glitches appear

- For example, in some cases, right before lock-loss, the glitch frequency increases significantly. In some cases, no.
- We may infer the reason why the lock-loss takes place by seeing the glitch population around it.
- This imply a viewpoint, a glitch in a glitch population.
- A way of characterizing a detector by seeing the glitch population.



Bayesian Statistics for Spam Filter

Our goal is to get $p(y_S|x)$. Here, x is set of token (in this case, a token means a word), y_S is an event of spam mail. From Bayes law, combined probability of x, y_j, $(j = S, N)$ is written

$$p(x, y_j) = p(x \mid y_j)p(y_j) = p(y_j \mid x)p(x),$$

$$p(y_j \mid x) = \frac{p(x \mid y_j)p(y_j)}{p(x)}.$$

If x can represent combination of tokens x_i, $i=1,...,m$,

$$p(x \mid y_j) = \prod_{i=1}^{m} p(x^i \mid y_j),$$

$$p(x) = p(x, y_S) + p(x, y_N)$$

$$= p(y_S)\prod_{i=1}^{m} p(x^i \mid y_S) + p(y_N)\prod_{i=1}^{m} p(x^i \mid y_N)$$

Bayesian Statistics for Spam Filter

Then,

$$p(y_S \mid x) = \frac{p(y_S)\prod_{i=1}^{m} p(x^i \mid y_S)}{p(y_S)\prod_{i=1}^{m} p(x^i \mid y_S) + p(y_N)\prod_{i=1}^{m} p(x^i \mid y_N)}$$

Now,

$$p(x^i) = p(x^i \mid y_S)p(y_S) + p(x^i \mid y_N)p(y_N),$$

$$\pi(x^i) \equiv \frac{p(x^i \mid y_S)p(y_S)}{p(x^i)} = \frac{p(x^i \mid y_S)p(y_S)}{p(x^i \mid y_S)p(y_S) + p(x^i \mid y_N)p(y_N)}$$

Therefore

$$p(y_S \mid x) = \frac{(p(y_S))^{1-m}\prod_{i=1}^{m} \pi(x^i)}{(p(y_S))^{1-m}\prod_{i=1}^{m} \pi(x^i) + (1-p(y_S))^{1-m}\prod_{i=1}^{m} (1-\pi(x^i))}$$

Bayesian Statistics for Spam Filter

Assuming

$$p(y_S) = p(y_N) = 0.5$$

We obtain

$$p(y_S \mid x) = \frac{\prod_{i=1}^{m} \pi(x^i)}{\prod_{i=1}^{m} \pi(x^i) + \prod_{i=1}^{m} (1 - \pi(x^i))}$$

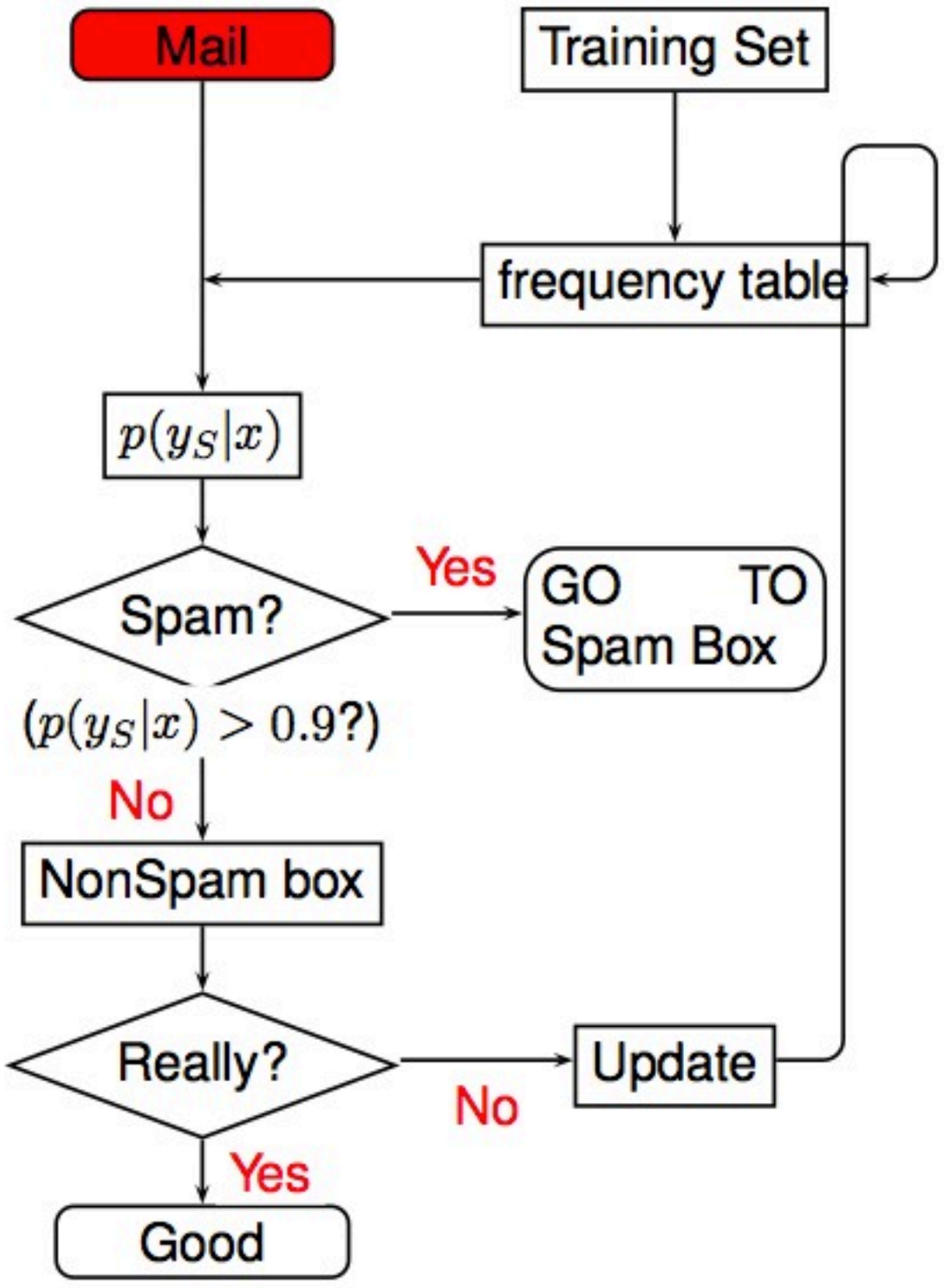
Graham uses following approximation,

$$p(x^i \mid y_S) p(y_S) = \alpha_S \frac{\#Spam[x^i]}{\#Spam}$$

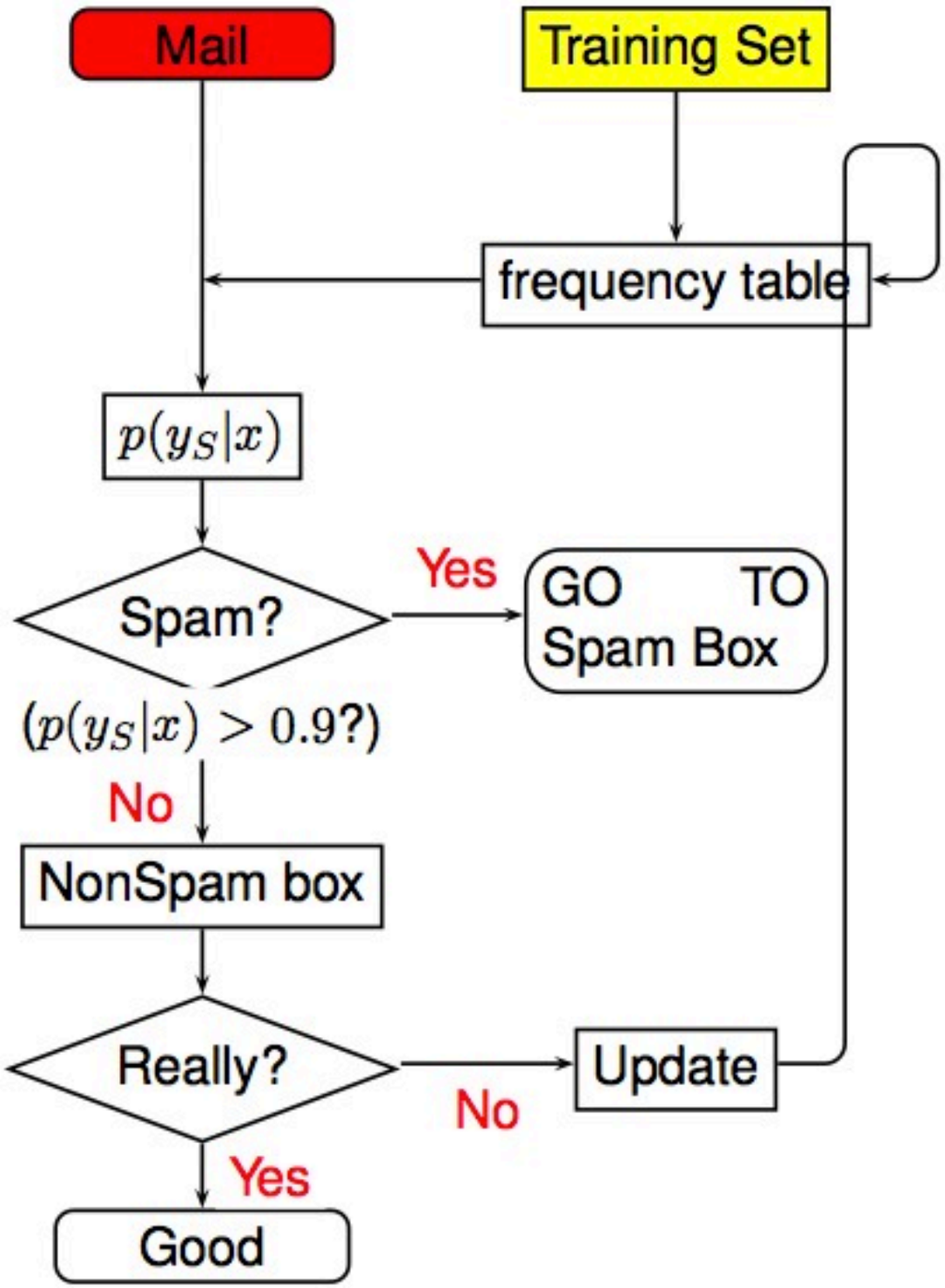
$$p(x^i \mid y_N) p(y_N) = \alpha_S \frac{\#NonSpam[x^i]}{\#NonSpam}$$

$$\pi(x^i) = \frac{\alpha_S \dfrac{\#Spam[x^i]}{\#Spam}}{\alpha_S \dfrac{\#Spam[x^i]}{\#Spam} + \alpha_N \dfrac{\#NonSpam[x^i]}{\#NonSpam}}$$

Flow Chart

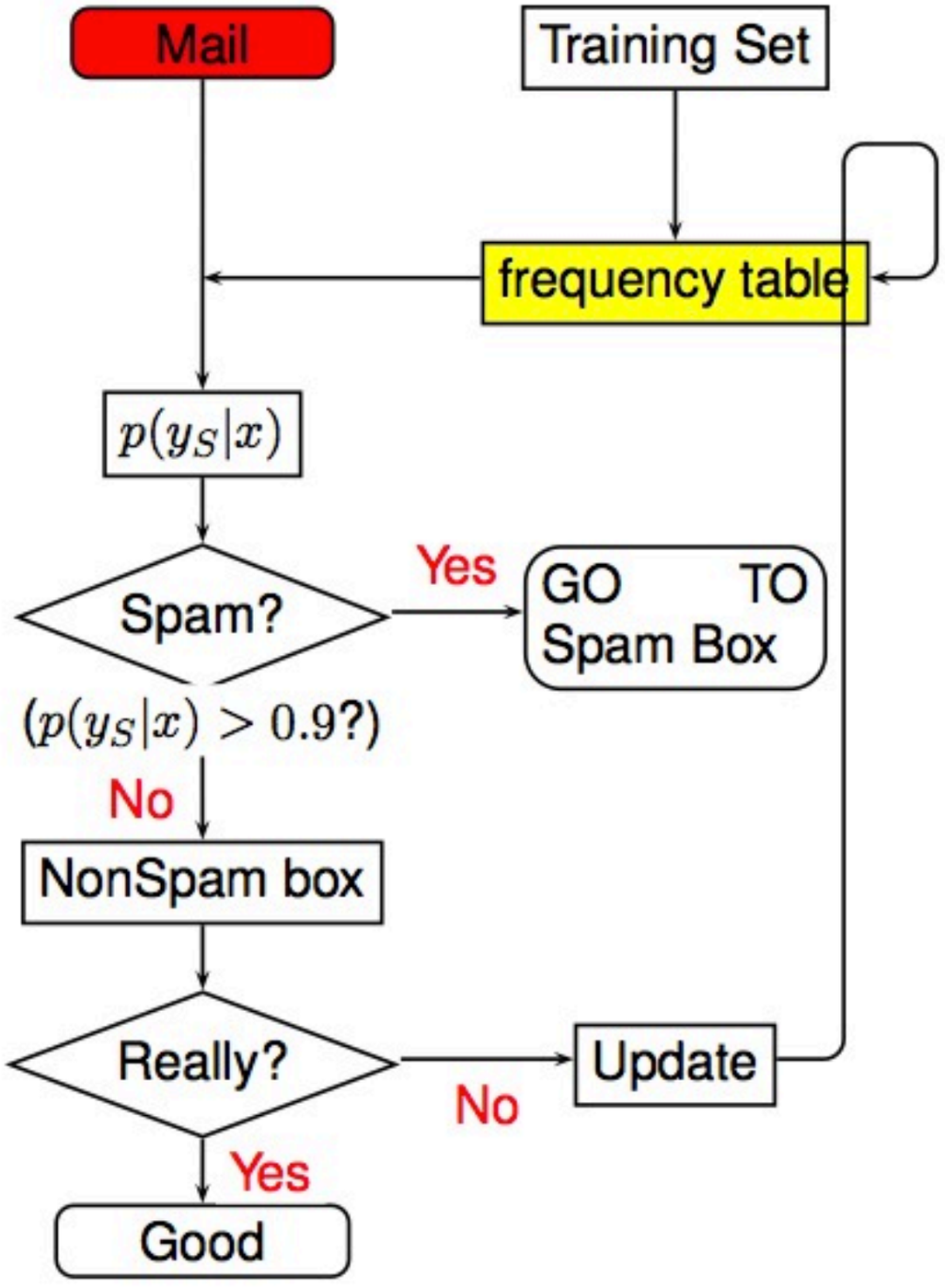

Mail
Training Set
frequency table
$p(y_S|x)$
Spam?
Yes
GO TO
Spam Box
$(p(y_S|x) > 0.9?)$
No
NonSpam box
Really?
No
Update
Yes
Good

Flow Chart



- prepare 4000 spam and non-spam mails, respectively.
- scan the entire text, including headers and embedded html and javascript, of each message in each corpus

Flow Chart

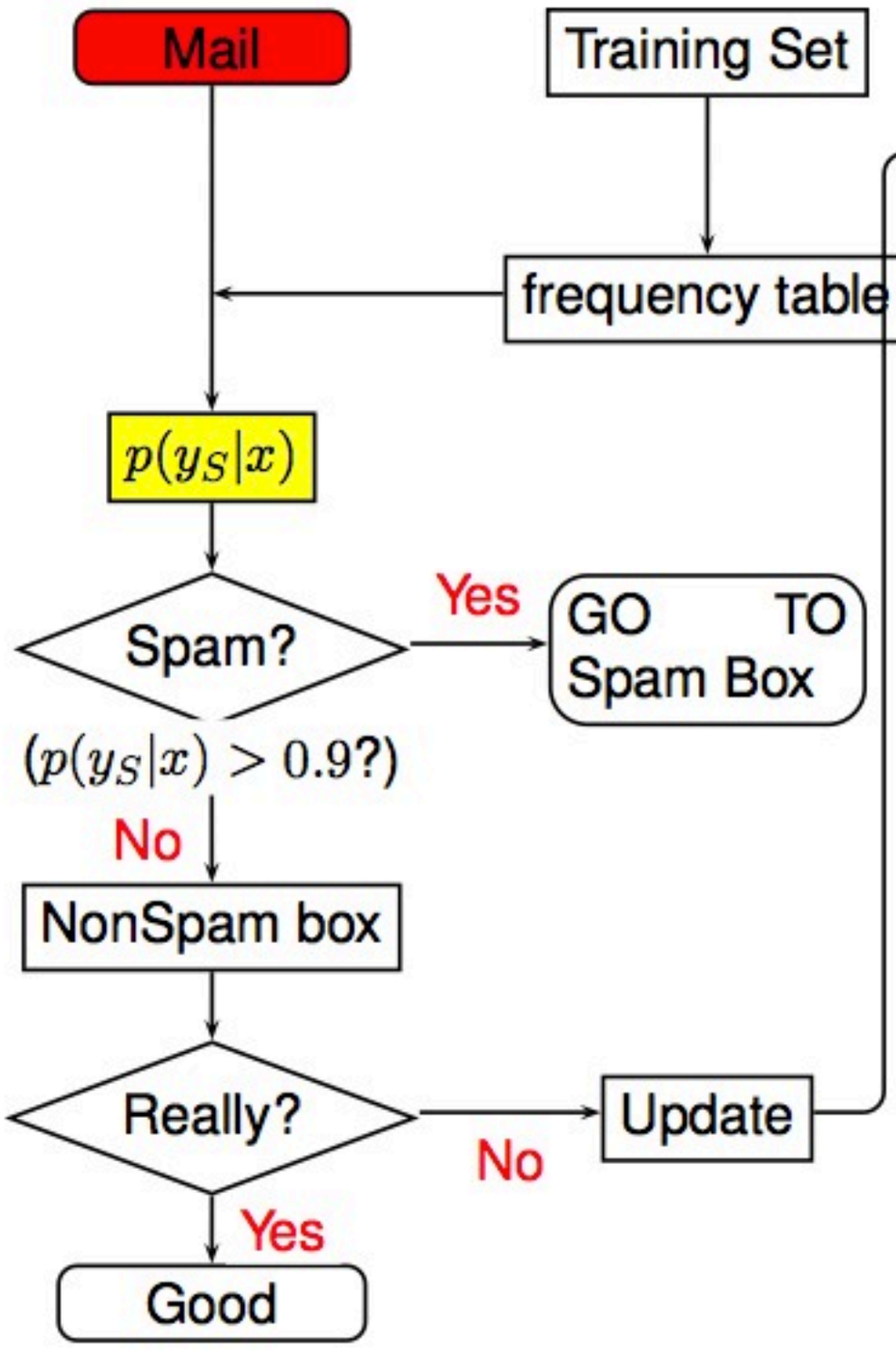


- **mapping each token to the probability that an email containing it is a spam**

$$\pi(x^i) = \frac{\alpha_S \frac{\#\,Spam[x^i]}{\#\,Spam}}{\alpha_S \frac{\#\,Spam[x^i]}{\#\,Spam} + \alpha_N \frac{\#\,NonSpam[x^i]}{\#\,NonSpam}}$$

here, $\alpha_N/\alpha_S=2$

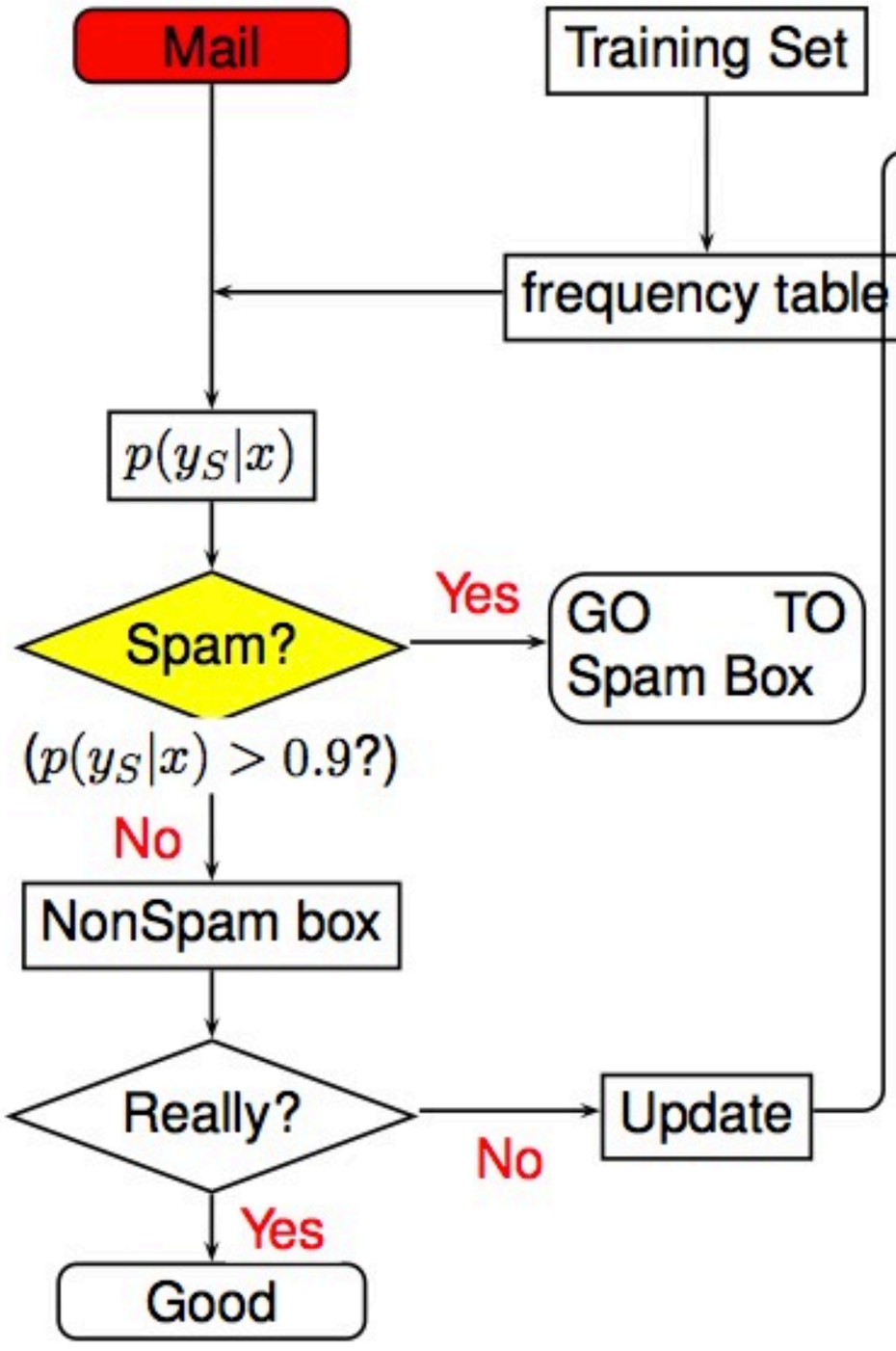
Flow Chart



- **calculate combined probability**

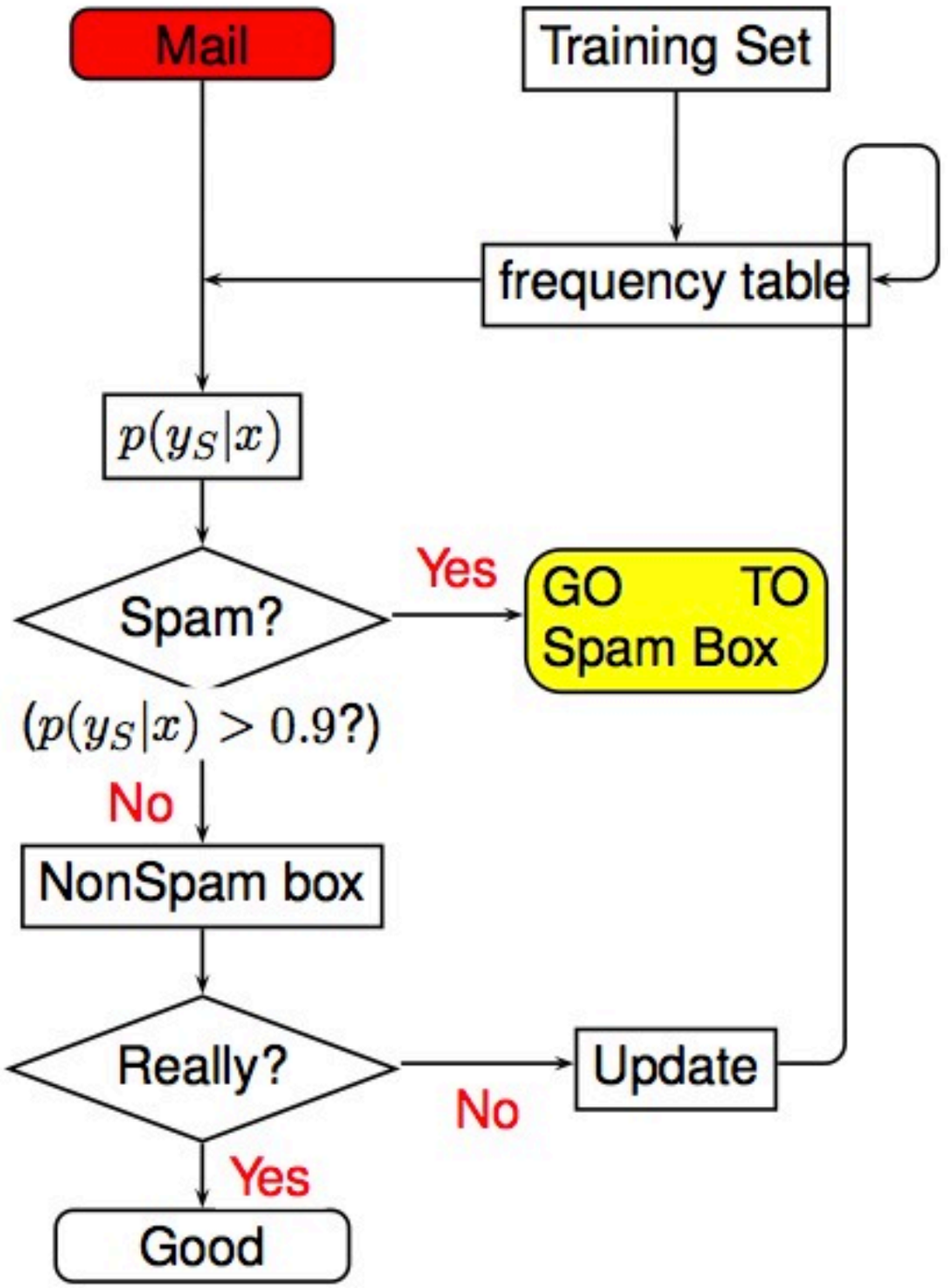
$$y(y_S \mid x) = \frac{\prod_{i=1}^{m} \pi(x^i)}{\prod_{i=1}^{m} \pi(x^i) + \prod_{i=1}^{m} (1 - \pi(x^i))}$$

Flow Chart



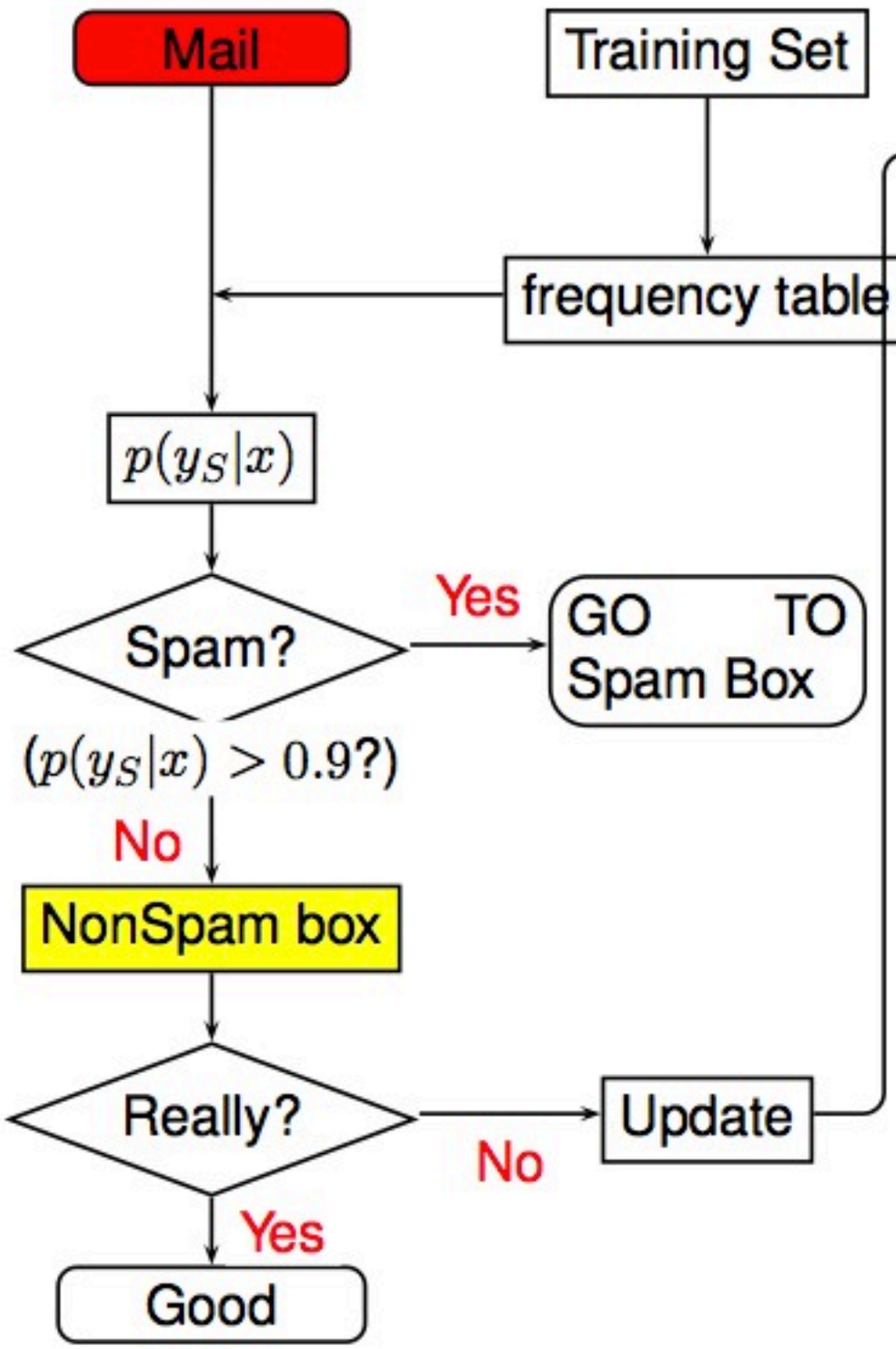
- If $p(y_S|x) > 0.9$, go to the spam box
- If other, go to the non-spam box

Flow Chart



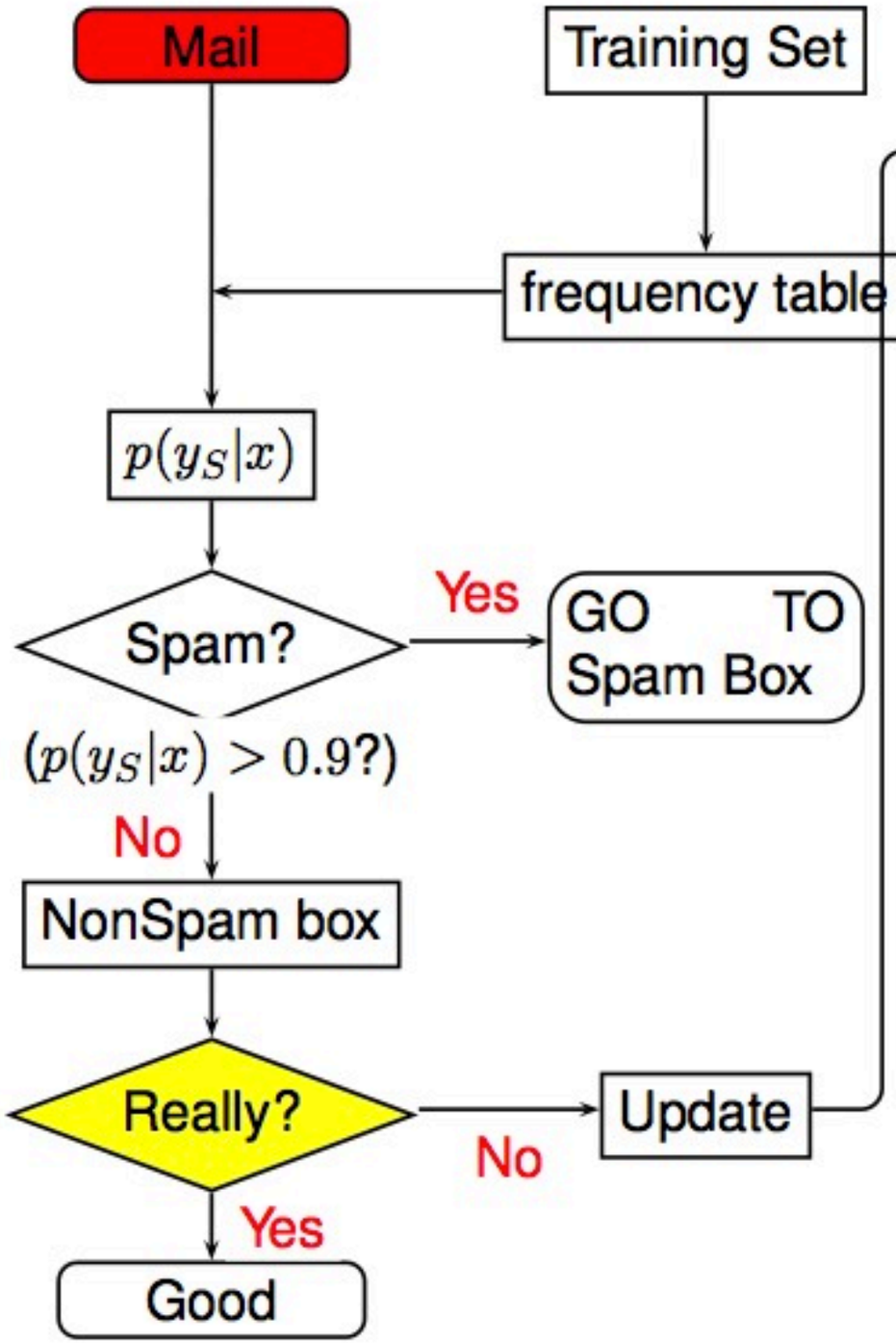
- $p(y_S|x) > 0.9$

Flow Chart



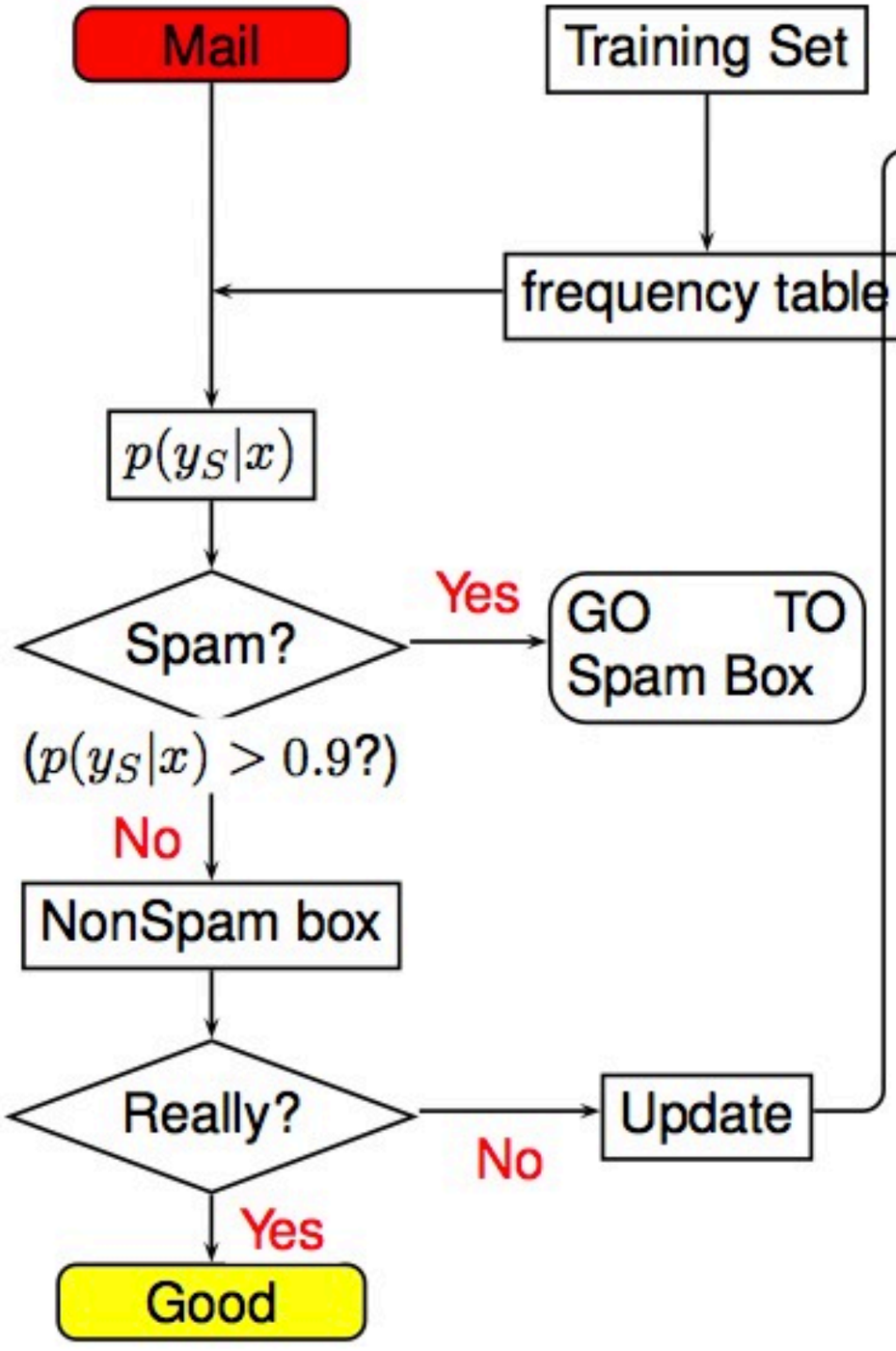
- $p(y_S|x) < 0.9$

Flow Chart



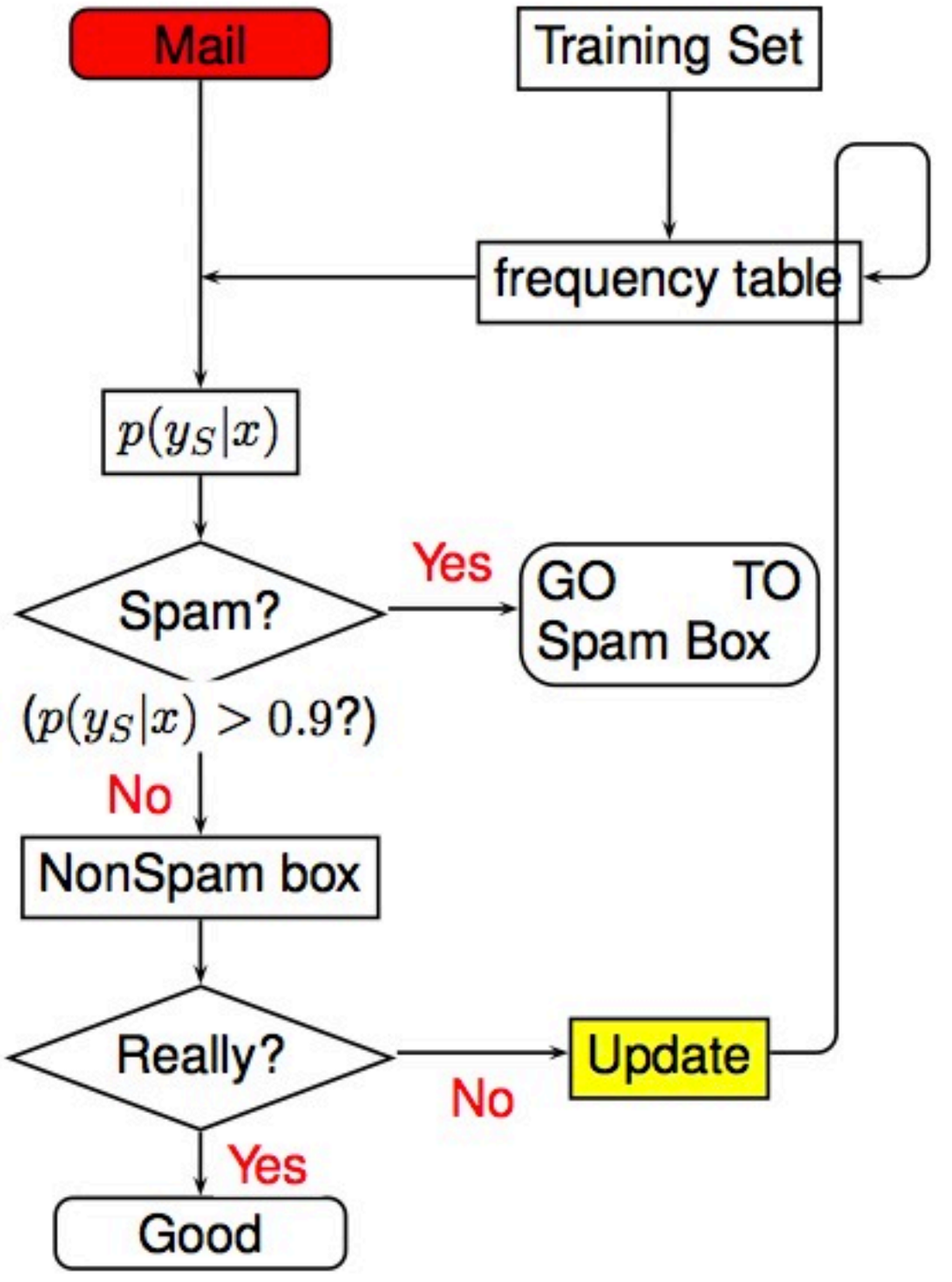
- **The Bayesian spam filter works correctly?**

Flow Chart



- Well done

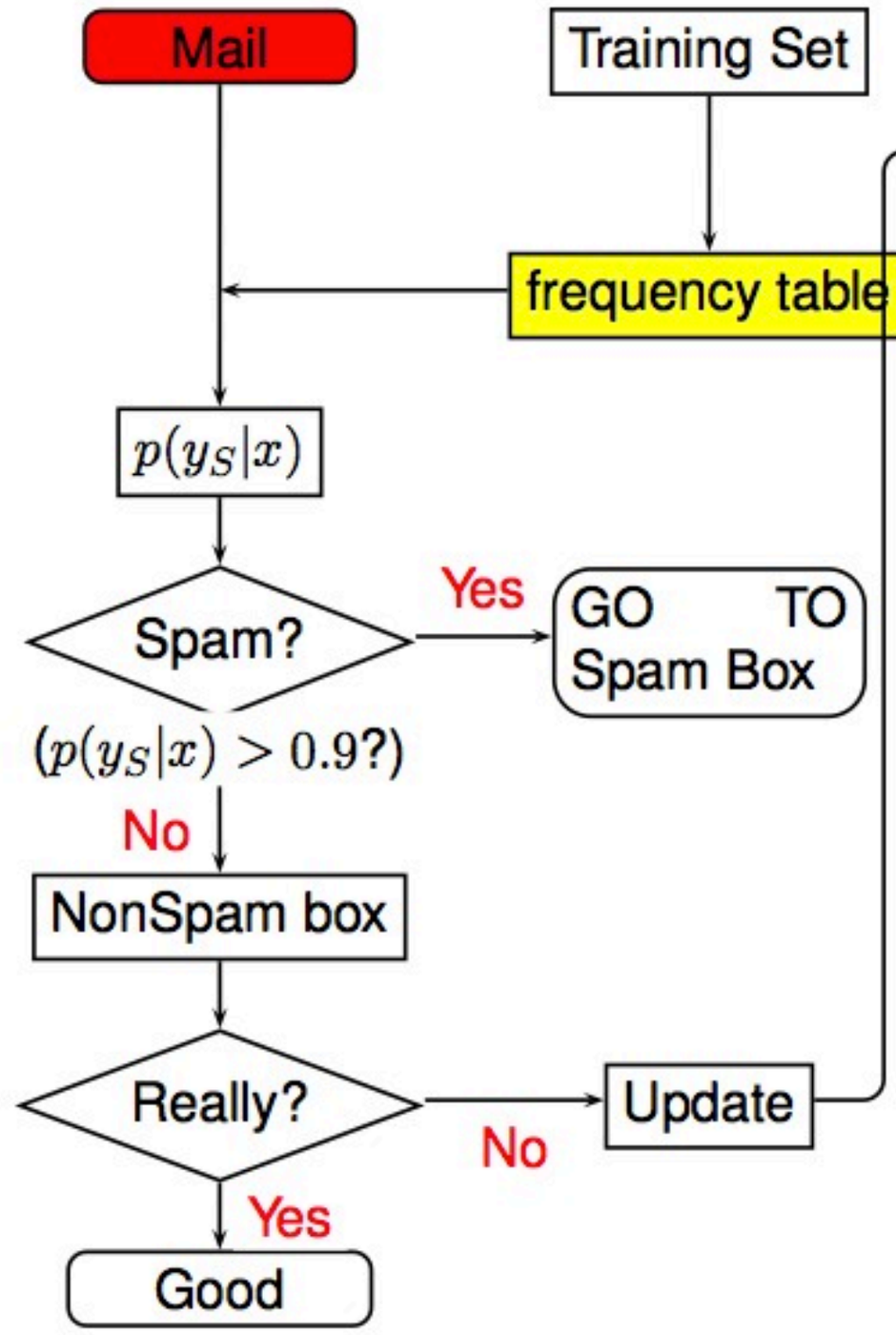
Flow Chart



- **prior probability is updated.**

$$\pi(x^i) = \frac{\alpha_S \frac{\#Spam[x^i]}{\#Spam}}{\alpha_S \frac{\#Spam[x^i]}{\#Spam} + \alpha_N \frac{\#NonSpam[x^i]}{\#NonSpam}}$$

Flow Chart



- **Database is updated and the filter evolve into more powerful one**
- **Repeat this procedure**

Spam/NonSpam and Glitch/GWB

- **Both want to avoid false dismissal**
- **Spammer -> lean to cope with spam filter -> new spam**
- **GW tel -> detector updated -> new glitch**
  (GW tel ~Spammer)
- **Glitches have many important information about the detector status**