

# Bayesian Filter fighting Spam



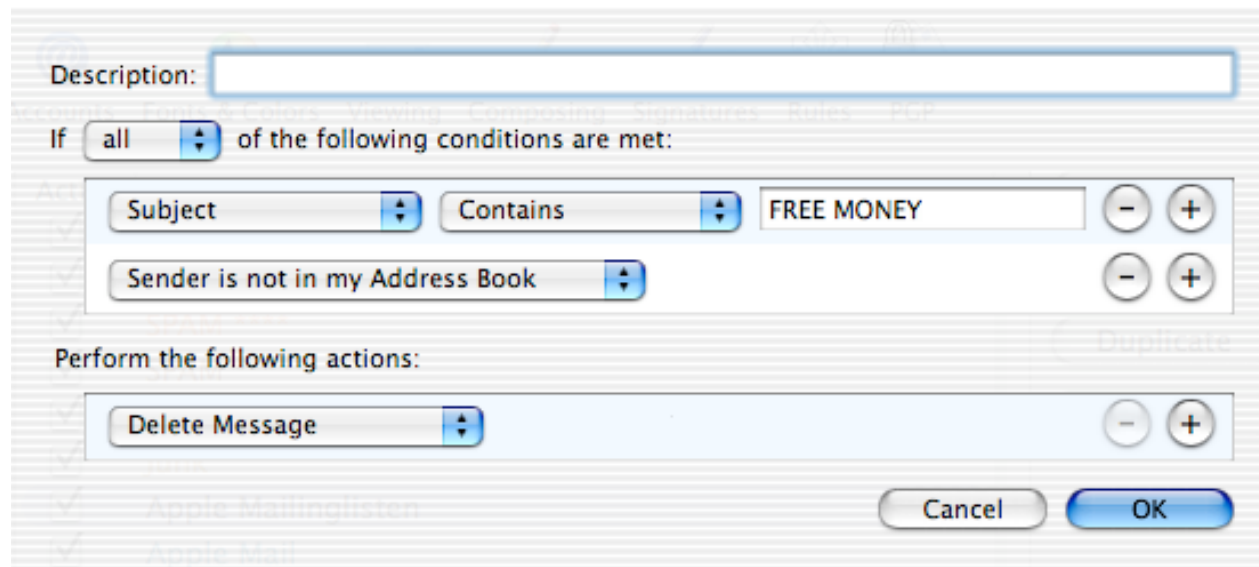
Hauptseminar Machine Learning, October 2003

# What is Spam?

- It is the act of blindly mass mailing a message that makes it spam, not the actual content.
- However, it seems that the language of spam constitutes a distinctive genre.
- Spam messages are often about topics rarely mentioned in legitimate messages.

# Motivation

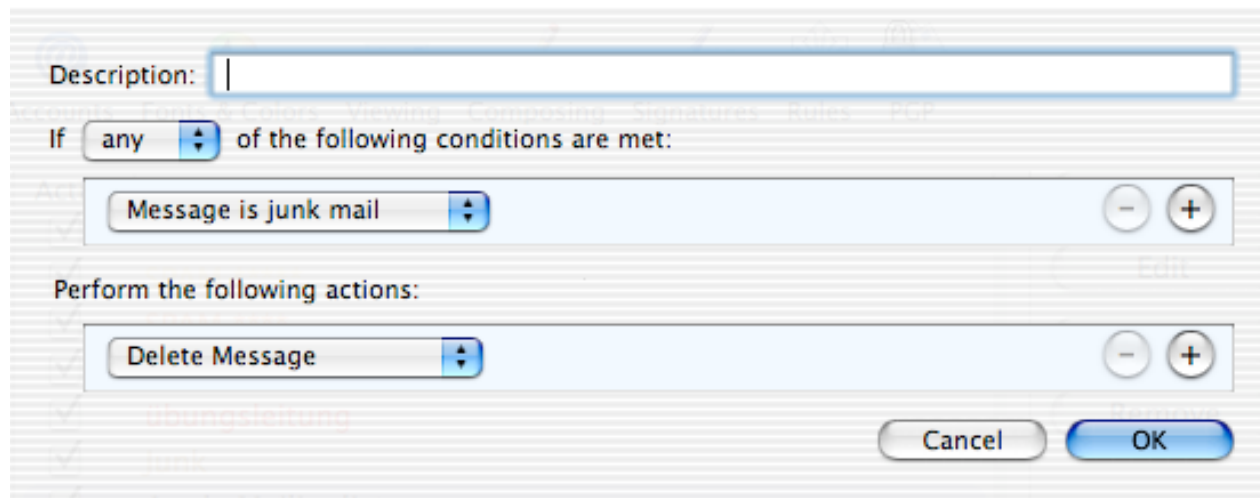
- Idea: Define rules that trigger some action
  - ECA Rules: “Event-Condition-Action”  
— but who defines those rules?



The image shows a screenshot of an email rule configuration dialog box. At the top, there is a text field labeled "Description:". Below it, the "If" section is set to "all" and contains two conditions: "Subject" contains "FREE MONEY" and "Sender is not in my Address Book". The "Perform the following actions:" section contains one action: "Delete Message". The dialog box has "Cancel" and "OK" buttons at the bottom right. The background shows a blurred view of the "Rules" tab in an email client.

# What we want

- If it's Spam, throw it away.



- but who decides what is Spam?

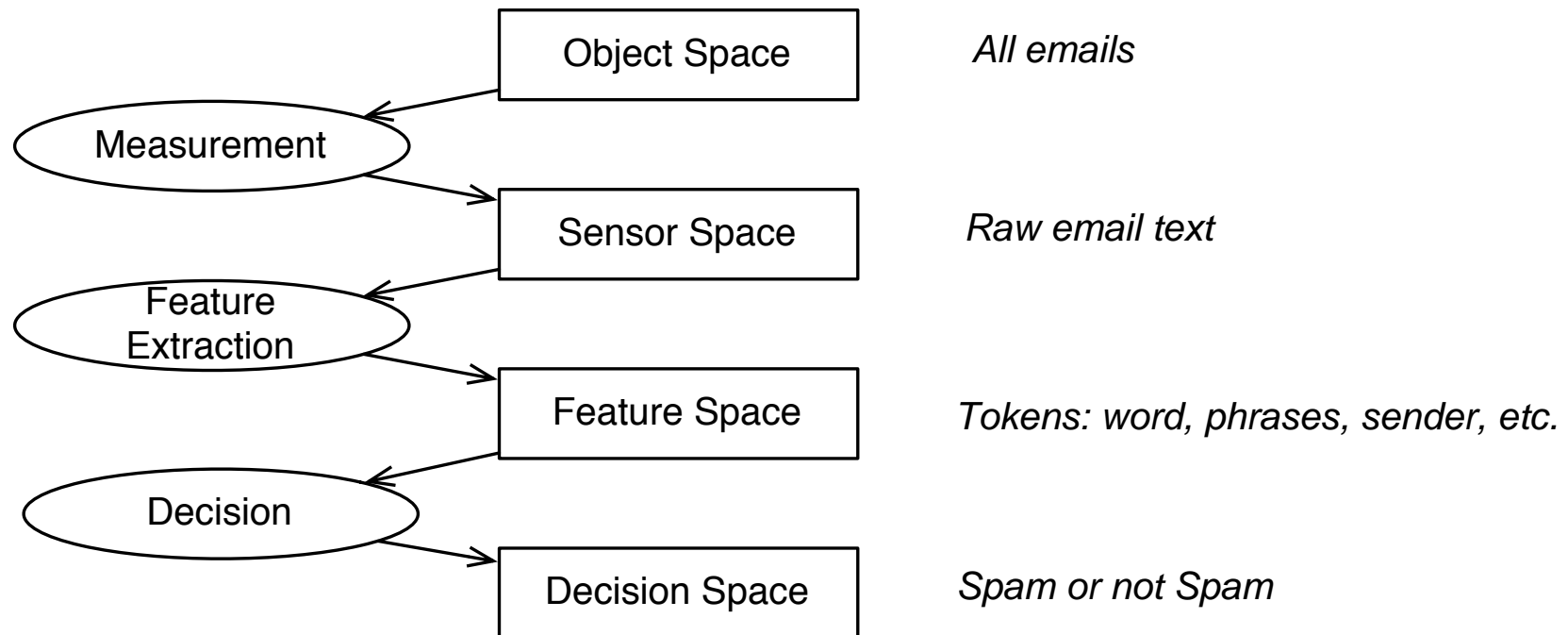
# Let the machines decide!

- The computer makes the decision
- The user can help in the decision by training the machine in advance.  
“Offline Learning”
- The user can help in the decision by correction wrong decisions.  
“Online Learning”

# More Classification Tasks

- Speech Recognition
- OCR (Optical Character Recognition)
- Biometric Sensors for authentication
- Quality control in production
- many more...

# Abstract Model



# Measurements vs. Features

- Measurements come directly from sensors, like CCD cameras, microphones, etc.
  - usually lots of data
  - contains a lot of unimportant data
- Features are extracted from raw data to reduce complexity



# a priory Probabilities

- assume we know nothing about an email, but we know that 20% of the mail we receive are spam.
- Then for a new email, we know with:

$$P(\omega_1) = 0.2$$

$\omega_1$  : Mail is spam

$$P(\omega_2) = 0.8$$

$\omega_2$  : Mail is not spam

# a priori Probabilities (2)

- we decide:

$$\omega_1 \text{ if } P(\omega_1) > P(\omega_2)$$

$$\omega_2 \text{ if } P(\omega_1) \leq P(\omega_2)$$

- but this means that we classify every email not to be spam.  
Obviously, this is not what we want.

# Feature Vector

- the Feature Vector contains all our extracted features.
- for example, count the occurrences of words in the email

$$X = (x_1, x_2, \dots, x_n)$$

- more on the choice of appropriate features later...

# a posteriori Probability

- The a posteriori probability is the a conditional probability after a measurement:

$$P(x_1, x_2, \dots, x_n | \omega_i)$$

i.e. the probability of the occurrence of certain words in spam (and not spam)

- but what we want is:

$$P(\omega_i | x_1, x_2, \dots, x_n)$$

# Conditional Probabilities

- from highschool we know:

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

which leads to

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

- This is called the “Bayes Formula”

# Bayes Formula for Spam

$$P(\textit{Spam} | x_1, \dots, x_n) = \frac{P(x_1, \dots, x_n | \textit{Spam}) \cdot P(\textit{Spam})}{P(x_1, \dots, x_n)}$$

- A problem remains: How can we calculate

$$P(x_1, \dots, x_n | \textit{Spam})$$

- trivial solution: assume independence of the individual features

# Naive Bayes

- Assuming independence, we can compute

$$\begin{aligned} P(x_1, \dots, x_n | Spam) &= \frac{P(x_1, \dots, x_n \wedge Spam)}{P(Spam)} \\ &= \frac{P(x_1 | x_2 \dots, x_n \wedge Spam) \cdot P(x_2 \dots, x_n \wedge Spam)}{P(Spam)} \\ &= \frac{P(x_1 | x_2 \dots, x_n \wedge Spam) \cdot P(x_2 \dots, x_n | Spam) \cdot P(Spam)}{P(Spam)} \\ &= P(x_1 | x_2 \dots, x_n \wedge Spam) \cdot P(x_2 \dots, x_n | Spam) \\ &= \dots \\ &= \prod_{i=1}^n P(x_i | x_{i+1} \dots, x_n \wedge Spam) \\ &= \prod_{i=1}^n P(x_i | Spam) \end{aligned}$$

# Naive Bayes (2)

- Now we can build our classifier:  
We classify an email as spam, if

$$\frac{P(\textit{Spam}|x_1, \dots, x_n)}{P(\textit{Ham}|x_1, \dots, x_n)} \geq \lambda$$

- The choice of  $\lambda$  depends on the “cost” we imply on missclassification.



# Loss Function

- Sometimes the cost of missclassification is different for different classes:
  - mistakenly deleting an important email is much worse than letting a spam mail slip through
  - selling a defect climbing rope is much worse than rejecting a good one in quality assurance.

# Loss Function (2)

- Formally, we assign each class a cost by defining a cost function

$$\lambda(\alpha_i, \omega_j)$$

- The overall risk is then:

$$R(\alpha_i | \vec{x}) = \sum_{i=1}^n \lambda(\alpha_i, \omega_j) \cdot P(\omega_i | \vec{x})$$

# Loss Function (3)

- We decide for that class that gives the minimum risk given the observation.
- In the two-category case this is the same as applying a threshold

$$\frac{P(\textit{Spam}|x_1, \dots, x_n)}{P(\textit{Ham}|x_1, \dots, x_n)} \underset{=}{\geq} \lambda$$

# Training Data

- We use our training data to compute the probabilities

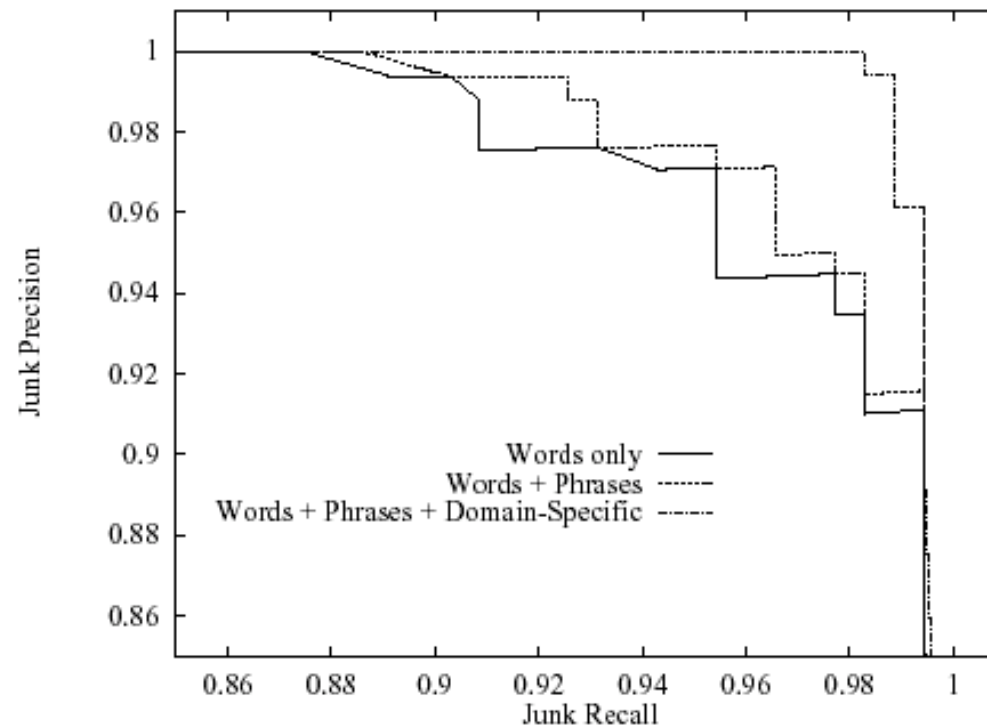
$$\begin{aligned} \frac{P(\textit{Spam}|x_1, \dots, x_n)}{P(\textit{Ham}|x_1, \dots, x_n)} &= \frac{P(\textit{Spam}) \cdot \prod_{i=1}^n p(x_i|\textit{Spam})}{P(\textit{Ham}) \cdot \prod_{i=1}^n p(x_i|\textit{Ham})} \\ &= \frac{N_{\textit{Spam}} \cdot \prod_{i=1}^n \frac{N_{\textit{Spam},x_i}}{N_{\textit{Spam}}}}{N_{\textit{Ham}} \cdot \prod_{i=1}^n \frac{N_{\textit{Ham},x_i}}{N_{\textit{Ham}}}} \\ &= \frac{\prod_{i=1}^n N_{\textit{Spam},x_i}}{\prod_{i=1}^n N_{\textit{Ham},x_i}} \end{aligned}$$

# Precision and Recall

- whenever we have a threshold value, we can write the precision and the recall as function of this parameter
- precision: the percentage of emails classified as spam that are in fact spam
- recall: the percentage of all spam emails that are correctly classified as spam

# Precision and Recall

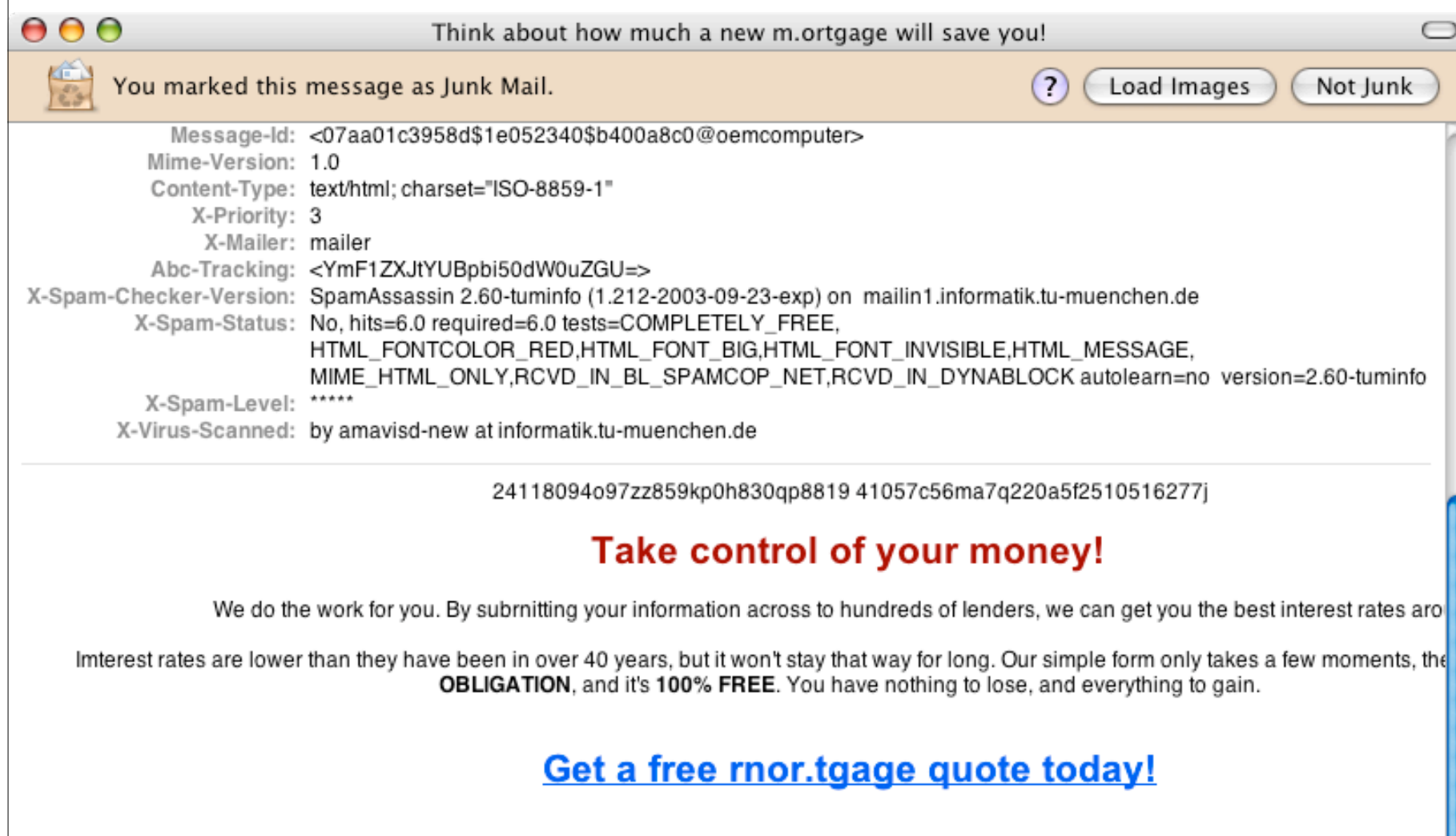
- example for precision/recall curve:



# How to select Features

- How to select Features
  - words, phrases, meta information:  
HTML messages, header fields, email address
  - removing insignificant features:  
calculate the mutual information between each feature and the class.

# Advanced Rules



Think about how much a new m.ortgage will save you!

You marked this message as Junk Mail. ? Load Images Not Junk

Message-Id: <07aa01c3958d\$1e052340\$b400a8c0@oemcomputer>  
Mime-Version: 1.0  
Content-Type: text/html; charset="ISO-8859-1"  
X-Priority: 3  
X-Mailer: mailer  
Abc-Tracking: <YmF1ZXJtYUBpbi50dW0uZGU=>  
X-Spam-Checker-Version: SpamAssassin 2.60-tuminfo (1.212-2003-09-23-exp) on mailin1.informatik.tu-muenchen.de  
X-Spam-Status: No, hits=6.0 required=6.0 tests=COMPLETELY\_FREE,  
HTML\_FONTCOLOR\_RED,HTML\_FONT\_BIG,HTML\_FONT\_INVISIBLE,HTML\_MESSAGE,  
MIME\_HTML\_ONLY,RCVD\_IN\_BL\_SPAMCOP\_NET,RCVD\_IN\_DYNABLOCK autolearn=no version=2.60-tuminfo  
X-Spam-Level: \*\*\*\*\*  
X-Virus-Scanned: by amavisd-new at informatik.tu-muenchen.de

---

24118094o97zz859kp0h830qp8819 41057c56ma7q220a5f2510516277j

**Take control of your money!**

We do the work for you. By submitting your information across to hundreds of lenders, we can get you the best interest rates around. Interest rates are lower than they have been in over 40 years, but it won't stay that way for long. Our simple form only takes a few moments, the **OBLIGATION**, and it's **100% FREE**. You have nothing to lose, and everything to gain.

[Get a free rnor.tgage quote today!](#)



# Advanced Features

header	subject: domain registration spam subject	DOMAIN_SUBJECT	1
header	Domain in From header has no MX or A DNS records	NO_DNS_FOR_FROM	0 1.105 0 1.650
header	From and To are the same, but not exactly	FROM_AND_TO_SAME	0.718 1.443 2.097 0.522
header	Received via buggy SMTP server (MDaemon 2.7.4SP4R)	MDAEMON_2_7_4	2.900 2.800 2.800 2.700
header	Received: contains a forged HELO	FORGED_RCVD_HELO	1
header	Received: contains a numeric HELO	RCVD_NUMERIC_HELO	1.271 0.326 1.526 1.502
header	Received: contains a name with a faked IP-address	FAKED_IP_IN_RCVD	2.900 2.800 2.800 2.700
header	Received via SMTPD32 server (SMTPD32-n.n)	SMTPD_IN_RCVD	1
header	Lots and lots of Cc: headers	LOTS_OF_CC_LINES	2.900 2.800 0 0
header	Received forged, contains fake AOL relays	FORGED_AOL_RCVD	4.300 4.300 4.100 4.100
header	Contains forged hostname for a DSL IP in Brazil	FORGED_TELESP_RCVD	2.900 2.800 2.800 2.700
header	Forged hotmail.com 'Received:' header found	FORGED_HOTMAIL_RCVD	0.470 0.001 0.500 0.500
header	hotmail.com 'From' address, but no 'Received:'	FORGED_HOTMAIL_RCVD2	0.051 0 1.884 2.499
header	Forged eudoramail.com 'Received:' header found	FORGED_EUDORAMAIL_RCVD	2.799 2.796 2.696 2.700
header	'From' yahoo.com does not match 'Received' headers	FORGED_YAHOO_RCVD	0.375 0.477 1.181 0.901
header	'From' junos.com does not match 'Received' headers	FORGED_JUNO_RCVD	1.538 2.796 2.696 2.058
header	Forged 'by gw05' 'Received:' header found	FORGED_GW05_RCVD	2.900 2.800 2.800 2.700
header	Forged hotmail.com Received 'from mx' header	FORGED_MX_HOTMAIL	2.900 2.800 2.800 2.700
header	Sent by a known spamhaus (qvcs)	RCVD_BY_QVES_COM	2.900 0 0 0
header	Character set doesn't exist	NONEXISTENT_CHARSET	2.900 2.800 2.800 2.700
header	A foreign language charset used in headers	CHARSET_FARAWAY_HEADER	3.200
header	'X-Mailer' line contains gibberish	X_MAILER_GIBBERISH	1
header	Sent with 'X-Priority' set to high	X_PRIORITY_HIGH	1.495 0.516 1.486 1.305
header	Sent with 'X-Msmail-Priority' set to high	X_MSMAIL_PRIORITY_HIGH	0.500 0.501 0.501 0.500

# Advanced Features

Content analysis details: (23.6 points, 6.0 required)

- 3.2 FROM\_HAS\_MIXED\_NUMS3 From: contains numbers mixed in with letters
- 0.3 FROM\_HAS\_MIXED\_NUMS From: contains numbers mixed in with letters
- 1.0 ACCEPT\_CREDIT\_CARDS BODY: Accept Credit Cards
- 0.5 CLICK\_BELOW\_CAPS BODY: Asks you to click below (in capital letters)
- 0.6 FOR\_FREE BODY: No such thing as a free lunch (1)
- 5.4 BAYES\_99 BODY: Bayesian spam probability is 99 to 100%  
[score: 1.0000]
- 0.3 MIME\_HTML\_ONLY BODY: Message only has text/html MIME parts
- 0.1 HTML\_MESSAGE BODY: HTML included in message
- 0.5 HTML\_LINK\_CLICK\_CAPS BODY: HTML link text says "CLICK"
- 0.1 HTML\_LINK\_CLICK\_HERE BODY: HTML link text says "click here"
- 0.6 MIME\_HTML\_NO\_CHARSET RAW: Message text in HTML without charset
- 0.5 REMOVE\_PAGE URI: URL of page called "remove"
- 2.6 SUSPICIOUS\_RECIPS Similar addresses in recipient list
- 0.7 RCVD\_IN\_DSBL RBL: Received via a relay in list.dsbl.org  
[<<http://dsbl.org/listing?ip=212.214.158.101>>]
- 1.5 RCVD\_IN\_BL\_SPAMCOP\_NET RBL: Received via a relay in bl.spamcop.net  
[Blocked - see <<http://www.spamcop.net/bl.shtml?212.214.158.101>>]
- 2.6 FORGED\_MUA\_OUTLOOK Forged mail pretending to be from MS Outlook
- 1.0 FORGED\_OUTLOOK\_TAGS Outlook can't send HTML in this format
- 0.0 UPPERCASE\_25\_50 message body is 25-50% uppercase
- 1.0 FORGED\_OUTLOOK\_HTML Outlook can't send HTML message only
- 1.1 MIME\_HTML\_ONLY\_MULTI Multipart message only has text/html MIME parts

# Spammers are fighting back

- HTML tricks:
  - Make mo<foo>ney f<bar>ast
  - used background color to hide words
    - include non-spam words
- using their own Bayes classifier to test the spam, trying to make it indistinguishable from legitimate email

# Simple Example

```
<html><center>24118094o97zz859kp0h830qp8819 41057c56ma7q220a5f251051627  
7j<br><font color="#ffffff">The demise of my hamster made me cry!</font><br>  
<font color="#990000" face="arial" size="6"><b>Take control of your money!</b></  
font>  
<br><br>We do the work for you. By submitting your information across to hundreds of  
lenders, we can get you the best interest rates around.<br><font color="#ffffff">All your  
efforts to be me have been futile! I rule!</font><br>Interest rates are lower than they  
have been in over 40 years, but it won't stay that way for long. Our simple form only  
takes a few moments, there is absolutly <b>NO OBLIGATION</b>, and it's <b>100%  
FREE</b>. You have nothing to lose, and everything to gain.<br><br><br><a  
href="http://www.greatmdz2s.com/cgi-bin/affiliates/clickthru.cgi?id=mail01"><b><font  
face="arial" size="6">Get a free rnor.tgage quote today!</font></b></  
a><br><br><br><br>ibb14kw23ug6l425sc3g4v90o84y6 135t3nqrw8d66596e0vunxr7x  
9e77<br><font  
size="1">tax4d886ys4924118094o97zz859kp0h830qp881941057c56ma7q220a5<br>  
<br>  
To get off our list, <a href="http://www.greatmdz2s.com/gone/">un s ubscr11be</a>.</  
font>
```

# One more Idea...

- Spam email can be roughly divided in two subgroups:
  - pornographic
  - other spam
- what about classifying into three classes instead of two classes?

# ... that did not work out.

- Tests showed that the combined classifier using porn-spam, other spam and legitimate emails had an overall worse performance
- The reasons for this are:
  - a model with more degrees of freedom must fit many more parameters from the data, and additionally
  - less data for each class is available

# Further Reading

- *Dr. Andreas Linke*  
**Spam oder nicht Spam? E-Mail sortieren mit Bayes-Filtern;**  
c't 17/03, Seite 150
- *Mehran Sahami, Susan Dumais, David Heckerman, Eric Horvitz ,*  
Microsoft Research  
**A Bayesian Approach to Filtering Junk E-Mail** (1998)
- *Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinou,*  
*George Paliouras, Constantine D. Spyropoulos*  
**An Evaluation of Naive Bayesian Anti-Spam Filtering** (2000)
- *Jefferson Provost*  
**Naive-Bayes vs. Rule-Learning in Classification of Email**