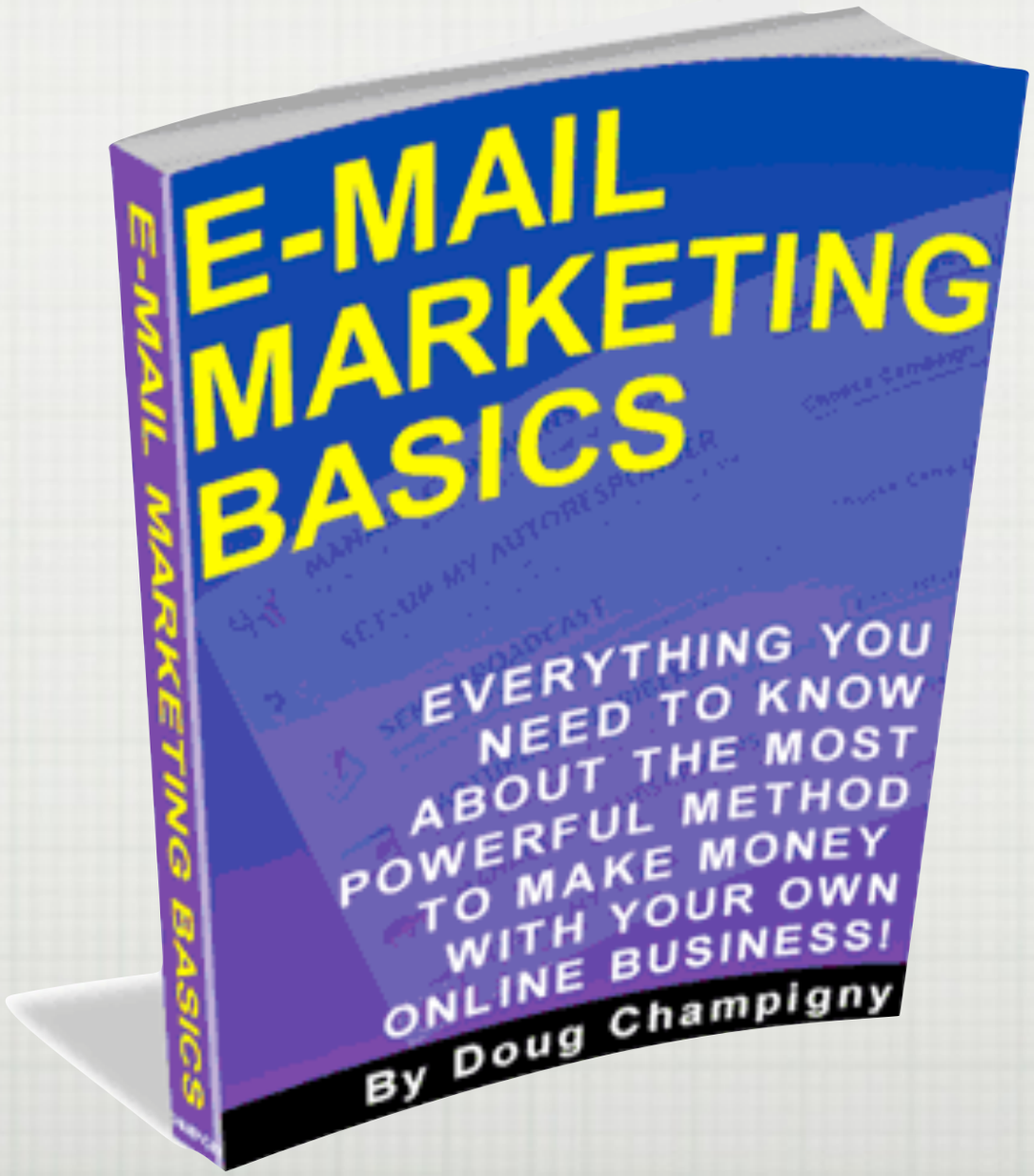
JOHDATUS TEKOÄLYYN

TEEMU ROOS

HELSINGIN YLIOPISTO

KUINKA RIKASTUA



NAIVI BAYES

FROM: "MARGARETTA NITA" <MARGUERITESEBRINA@WMLE.COM>
SUBJECT: SPECIAL OFFER : VIAGRA ON SALE AT $1.38 !!!
X-BOGOSITY: YES, TESTS=BOGOFILTER, SPAMICITY=0.99993752,
VERSION=2011-08-29
DATE: MON, 26 SEP 2011 21:52:26 +0300
X-CLASSIFICATION: JUNK - AD HOC SPAM DETECTED (CODE = 73)

SPECIAL OFFER : VIAGRA ON SALE AT $1.38 !!!

COMPARE THE BEST ONLINE PHARMACIES TO BUY VIAGRA. ORDER
VIAGRA ONLINE WITH HUGE DISCOUNT.
MULTIPLE BENEFITS INCLUDE FREE SHIPPING, REORDER DISCOUNTS,
BONUS PILLS
HTTP://RXPHARMACYCVS.RU

NAIVI BAYES

FROM: "MARGARETTA NITA" <MARGUERITESEBRINA@WMLE.COM>
SUBJECT: SPECIAL OFFER : VIAGRA ON SALE AT $1.38 !!!
X-BOGOSITY: YES, TESTS=BOGOFILTER, SPAMICITY=0.99993752,
VERSION=2011-08-29
DATE: MON, 26 SEP 2011 21:52:26 +0300
X-CLASSIFICATION: JUNK - AD HOC SPAM DETECTED (CODE = 73)

SPECIAL OFFER : VIAGRA ON SALE AT $1.38 !!!

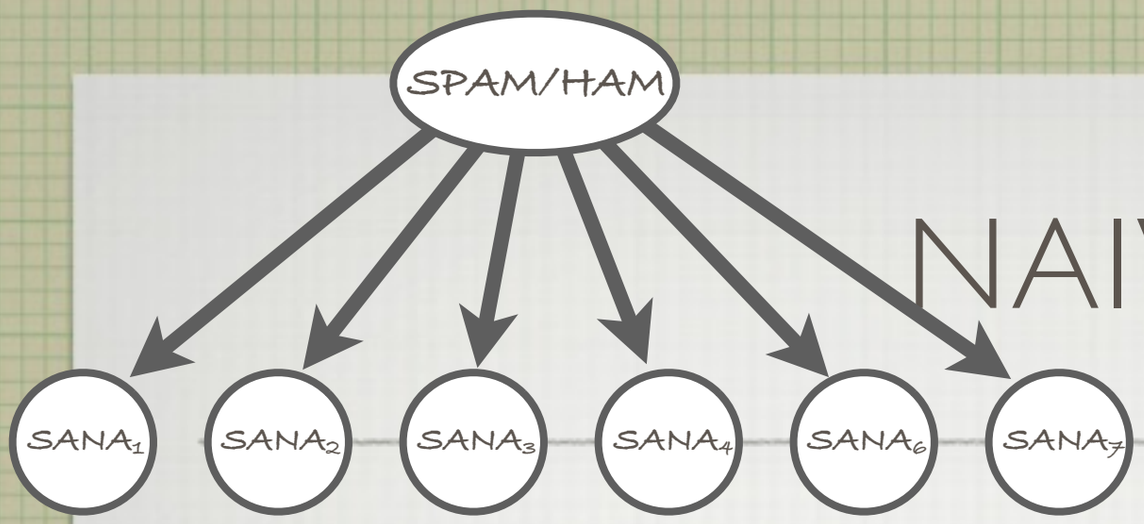
COMPARE THE BEST ONLINE PHARMACIES TO BUY VIAGRA. ORDER
VIAGRA ONLINE WITH HUGE DISCOUNT.
MULTIPLE BENEFITS INCLUDE FREE SHIPPING, REORDER DISCOUNTS,
BONUS PILLS
HTTP://RXPHARMACYCVS.RU

NAIVI BAYES

```
FROM: "MARGARETTA NITA" <MARGUERITESEBRINA@WMLE.COM>
SUBJECT: SPECIAL OFFER : VIAGRA ON SALE AT $1.38 !!!
X-BOGOSITY: YES, TESTS=BOGOFILTER, SPAMICITY=0.99993752,
VERSION=2011-08-29
DATE: MON, 26 SEP 2011 21:52:26 +0300
X-CLASSIFICATION: JUNK - AD HOC SPAM DETECTED (CODE = 73)


SPECIAL OFFER : VIAGRA ON SALE AT $1.38 !!!

COMPARE THE BEST ONLINE PHARMACIES TO BUY VIAGRA. ORDER
VIAGRA ONLINE WITH HUGE DISCOUNT.
MULTIPLE BENEFITS INCLUDE FREE SHIPPING, REORDER DISCOUNTS,
BONUS PILLS
HTTP://RXPHARMACYCVS.RU
```



NAIVI BAYES

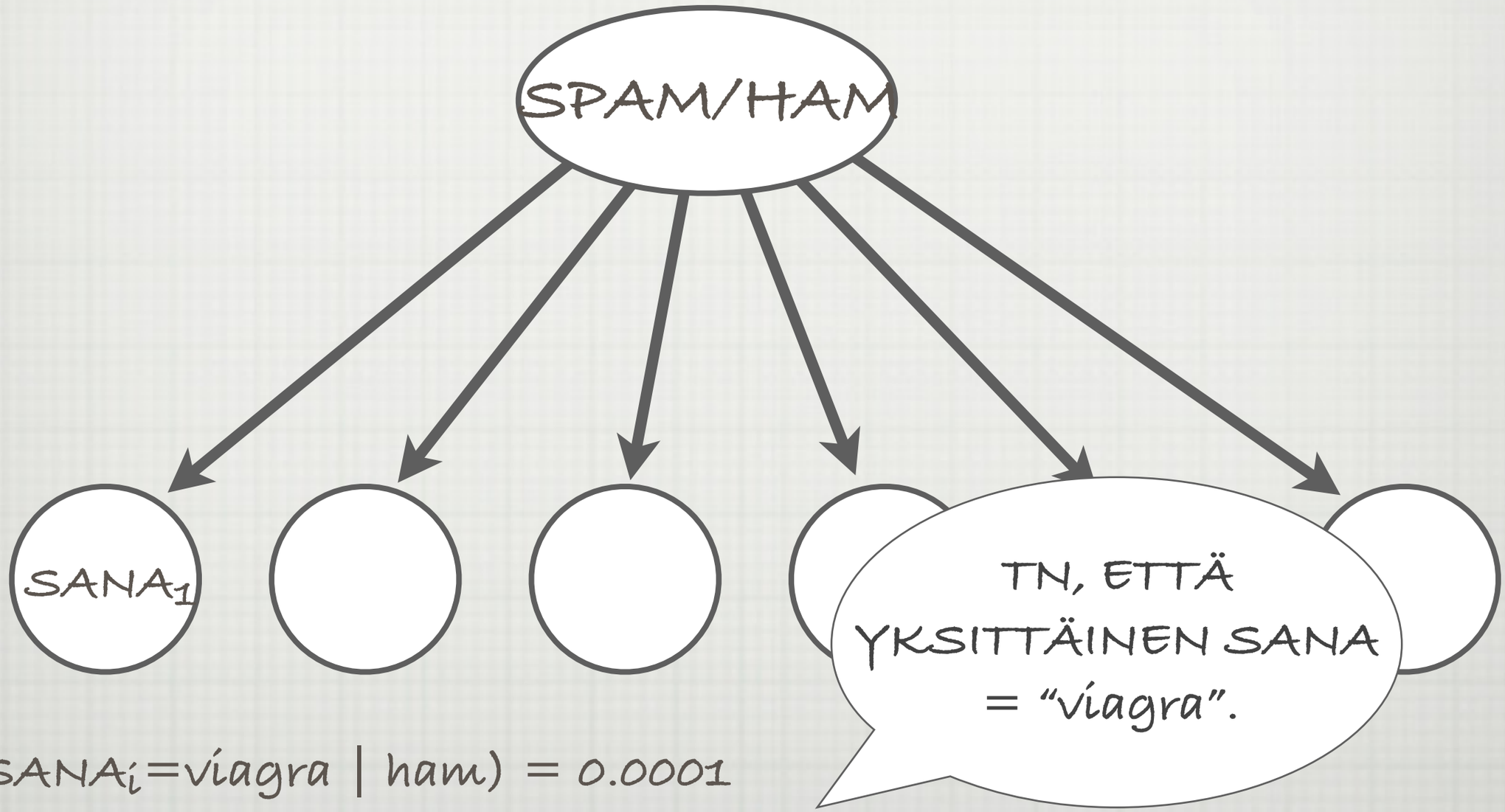
ROSKAPOSTISUODATIN:

SPAMICITY(Viesti, P):

Odds = P.Spam / P.noSpam
for each Sana in Viesti
Odds = Odds * P.Sana_Spam(Sana) /P.Sana_noSpam(Sana)
return(Odds)

JOS SPAMICITY(Viesti, P) >1, LUOKITTELE VIESTI SPAMIKSI
JOS SPAMICITY(Viesti, P) <1, LUOKITTELE VIESTI HAMIKSI

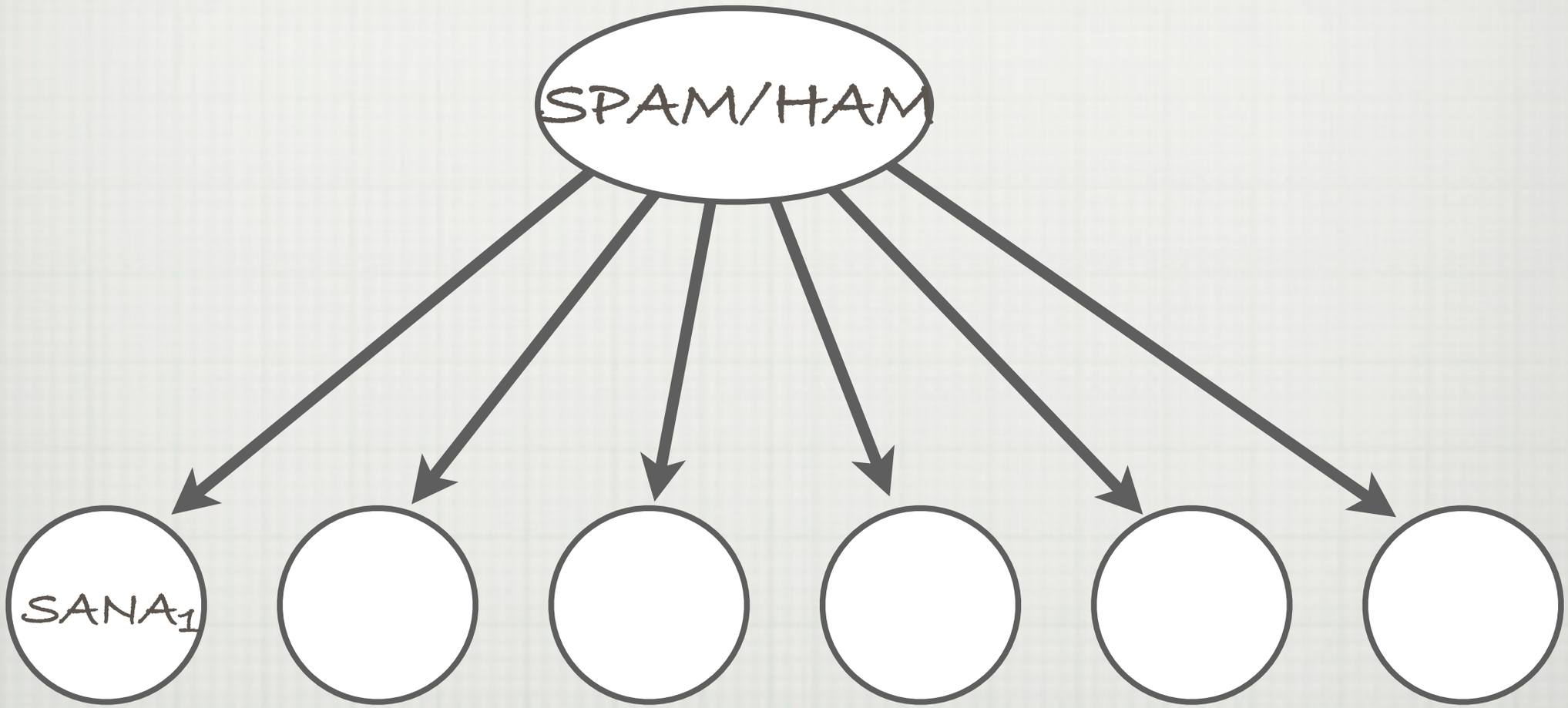
NAIVI BAYES



P(SANA$_i$=viagra | ham) = 0.0001

P(SANA$_i$=viagra | spam) = 0.002

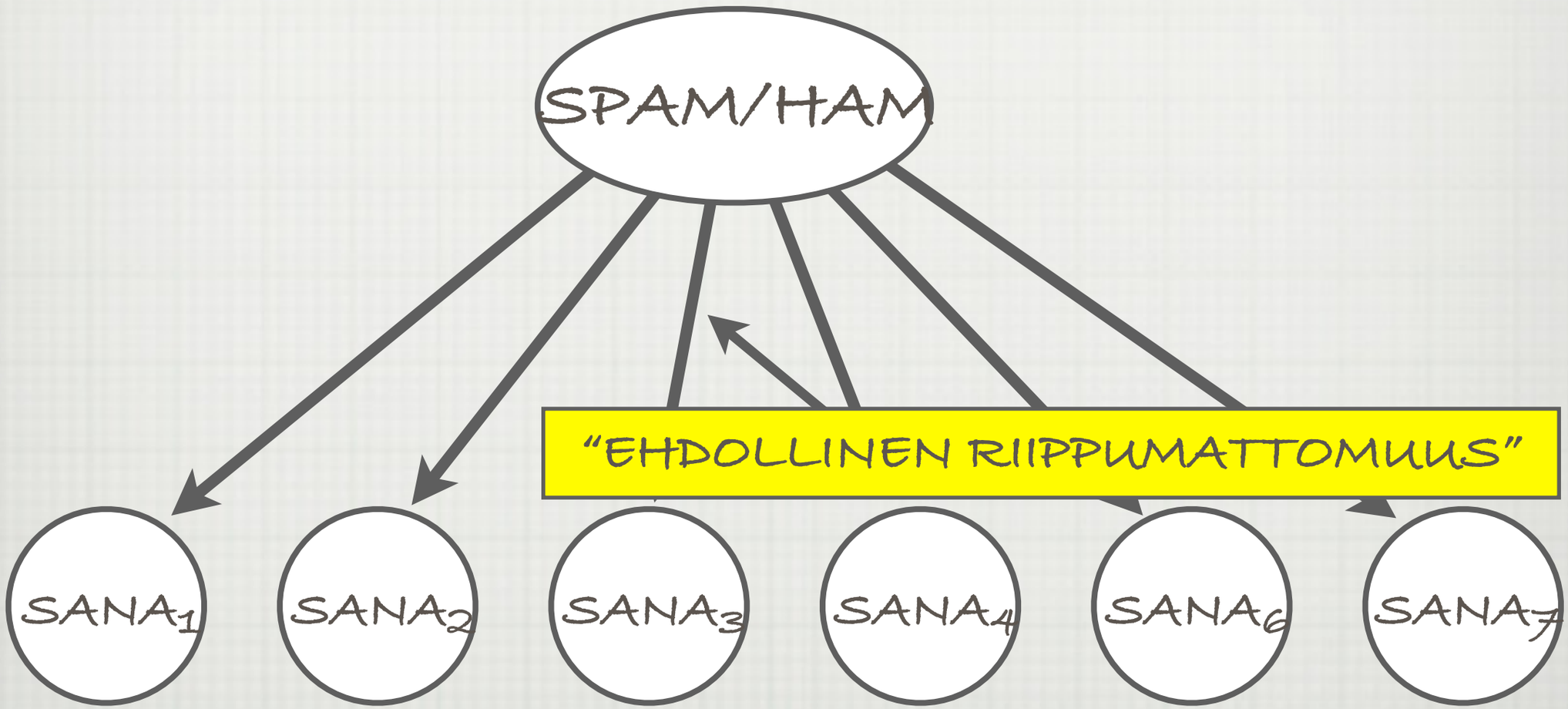
NAIVI BAYES



$P(SANA_i = \$ \mid ham) = 0.0002$

$P(SANA_i = \$ \mid spam) = 0.005$

NAIVI BAYES


SPAM/HAM
"EHDOLLINEN RIIPPUMATTOMUUS"
SANA1
SANA2
SANA3
SANA4
SANA6
SANA7

NAIVI BAYES

MUUTTUJAT:

1. LUOKKA: spam/ham
2. $SANA_1$
3. $SANA_2$
4. ...

JAKAUMAT:

$P(LUOKKA=spam) = 0.5$

$P(SANA_i=viagra \mid spam)=0.002$
$P(SANA_i=viagra \mid ham)=0.0001$

$P(SANA_i=\$ \mid spam)=0.005$
$P(SANA_i=\$ \mid ham)=0.0002$

$P(SANA_i=is \mid spam)=0.002$
$P(SANA_i=is \mid ham)=0.002$

$P(SANA_i=algorithm \mid spam)=0.0001$
$P(SANA_i=algorithm \mid ham)=0.002$

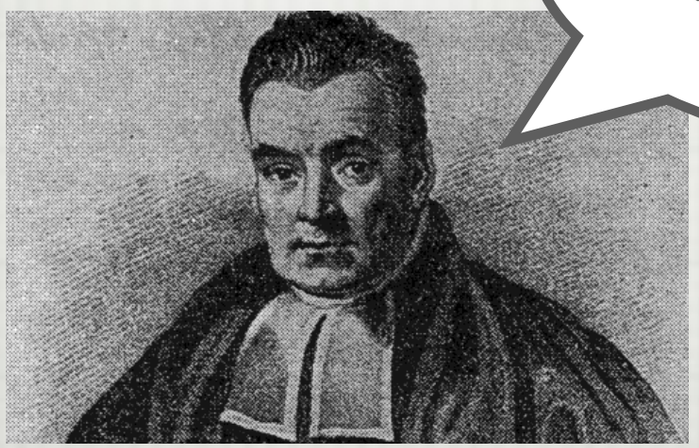
JNE...

NAIVI BAYES

PÄÄTTELY:

1. $P(spam) = 0.5$
2. $P(spam \mid SANA_1 = viagra) = \frac{P(spam)\, P(SANA_1 = viagra \mid spam)}{P(SANA_1 = viagra)}$

BAYESIN KAAVA!



NAIVI BAYES

PÄÄTTELY:

1. $P(\text{spam}) = 0.5$
2. $P(\text{spam} \mid \text{SANA}_1=\text{viagra}) = \dfrac{P(\text{spam})\, P(\text{SANA}_1=\text{viagra} \mid \text{spam})}{P(\text{SANA}_1=\text{viagra})}$

$$P(\text{SANA}_1=\text{viagra}) = P(\text{spam})\, P(\text{SANA}_1=\text{viagra} \mid \text{spam}) + P(\text{ham})\, P(\text{SANA}_1=\text{viagra} \mid \text{ham})$$

NAIVI BAYES

PÄÄTTELY:

1. $P(spam) = 0.5$
2. $P(spam \mid SANA_1 = viagra) = \frac{P(spam)\, P(SANA_1 = viagra \mid spam)}{P(SANA_1 = viagra)}$
3. $P(spam \mid SANA_1 = viagra, SANA_2 = is) = \frac{P(spam)\, P(SANA_1 = viagra, SANA_2 = is \mid spam)}{P(SANA_1 = viagra, SANA_2 = is)}$
4. $P(spam \mid SANA_1 = viagra, SANA_2 = is, SANA_3 = algorithm) = \frac{P(spam)\, P(SANA_1 = viagra, SANA_2 = is, SANA_3 = algorithm \mid spam)}{P(SANA_1 = viagra, SANA_2 = is, SANA_3 = algorithm)}$

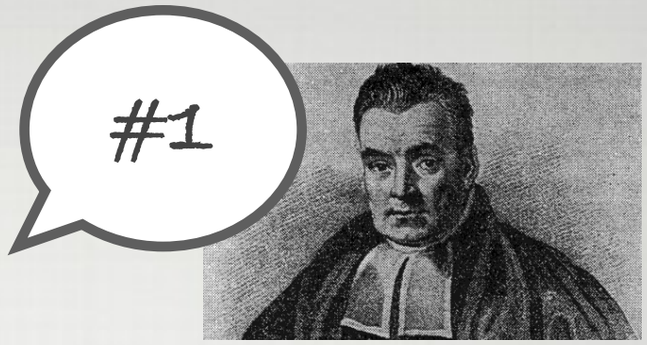
NAIVI BAYES

PÄÄTTELY:

1. $P(spam) = 0.5$
2. $P(spam \mid SANA_1 = viagra) = \frac{P(spam)\, P(SANA_1 = viagra \mid spam)}{P(SANA_1 = viagra)}$
3. $P(spam \mid SANA_1 = viagra, SANA_2 = is) = \frac{P(spam)\, P(SANA_1 = viagra, SANA_2 = is \mid spam)}{P(SA$
4. $P(spam \mid SANA_1 = $
$P(spam)\, P(SANA_1 = viagra, SANA_2 = is, SA$
$= \frac{}{P(SANA_1 = viagra, SANA_2 = is, SA$

KAKSI NIKSIÄ!

NAIVI BAYES



PÄÄTTELY:

4. $P(spam \mid SANA_1=viagra, SANA_2=is, SANA_3=algorithm)$

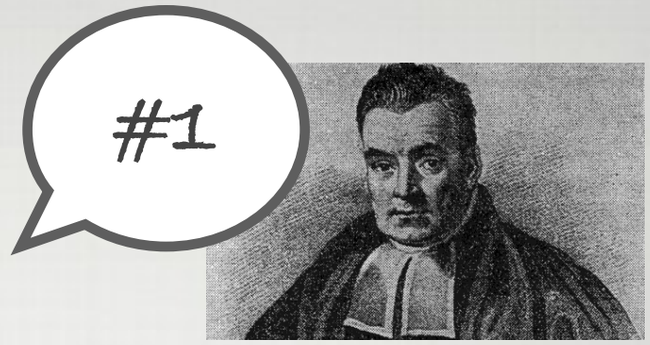
$$= \frac{P(spam)\, P(SANA_1=viagra, SANA_2=is, SANA_3=algorithm \mid spam)}{P(SANA_1=viagra, SANA_2=is, SANA_3=algorithm)}$$

$P(SANA_1=viagra, SANA_2=is, SANA_3=algorithm)$

$= P(spam, SANA_1=viagra, SANA_2=is, SANA_3=algorithm)$

$+ P(ham, SANA_1=viagra, SANA_2=is, SANA_3=algorithm)$

NAIVI BAYES



PÄÄTTELY:

4. $P(spam \mid SANA_1 = viagra, SANA_2 = is, SANA_3 = algorithm)$

$$= \frac{P(spam)\, P(SANA_1 = viagra, SANA_2 = is, SANA_3 = algorithm \mid spam)}{P(SANA_1 = viagra, SANA_2 = is, SANA_3 = algorithm)}$$

$P(evidenssi)$

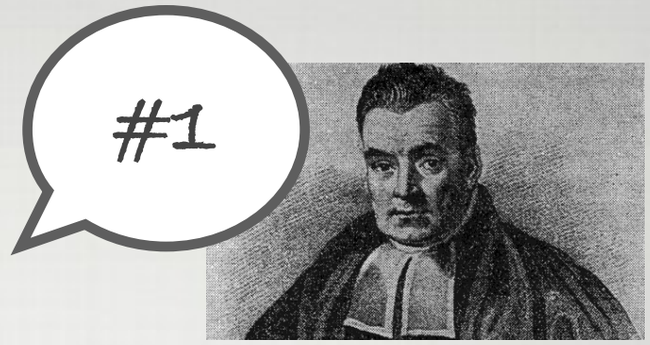
$= P(spam, evidenssi)$ // MARGINALISOINTI

$+ P(ham, evidenssi)$

$$P(spam \mid evidenssi) = \frac{P(spam)P(evidenssi \mid spam)}{P(spam, evidenssi) + P(ham, evidenssi)}$$

$$P(ham \mid evidenssi) = \frac{P(ham)P(evidenssi \mid ham)}{P(spam, evidenssi) + P(ham, evidenssi)}$$

NAIVI BAYES



PÄÄTTELY:

4. $P(spam \mid SANA_1=viagra, SANA_2=is, SANA_3=algorithm)$

$$= \frac{P(spam)\,P(SANA_1=viagra, SANA_2=is, SANA_3=algorithm \mid spam)}{P(SANA_1=viagra, SANA_2=is, SANA_3=algorithm)}$$

$P(evidenssi)$

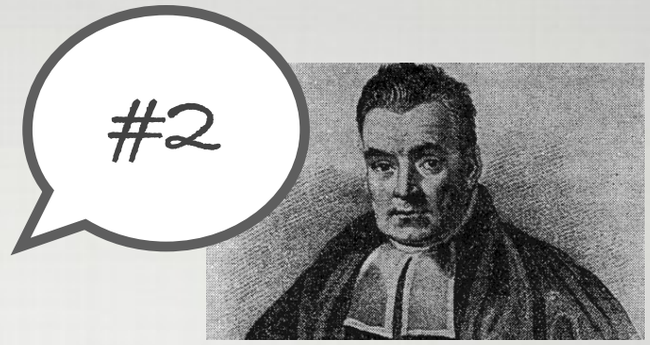
$= P(spam, evidenssi)$ // MAR

$+ P(ham, evidenssi)$

$$\frac{P(S \mid E)}{P(\neg S \mid E)} = \frac{P(S)P(E \mid S)}{P(\neg S)P(E \mid \neg S)}$$

$$\frac{P(spam \mid evidenssi)}{P(ham \mid evidenssi)} = \frac{P(spam)P(evidenssi \mid spam)}{P(ham)P(evidenssi \mid ham)}$$

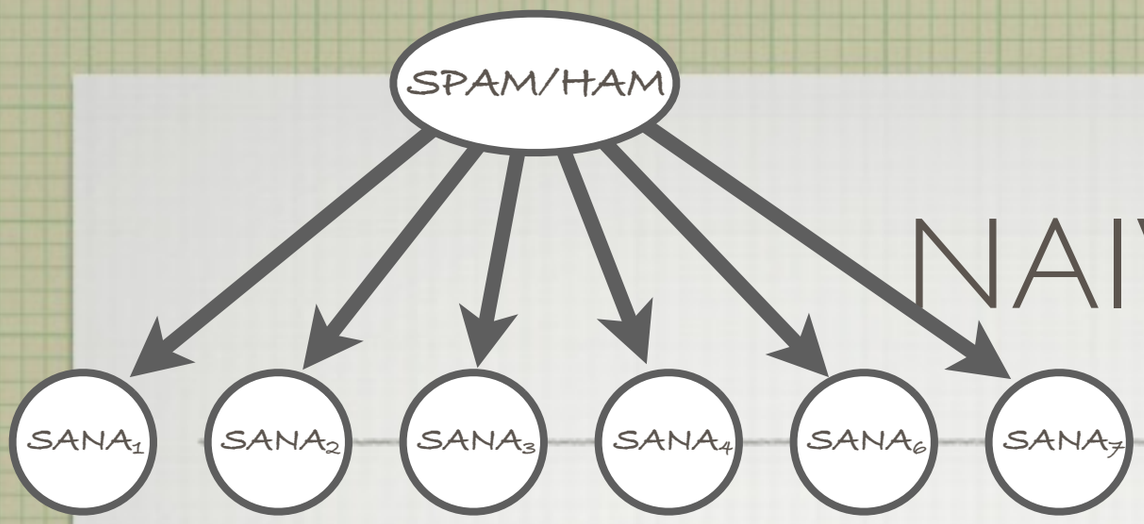
NAIVI BAYES



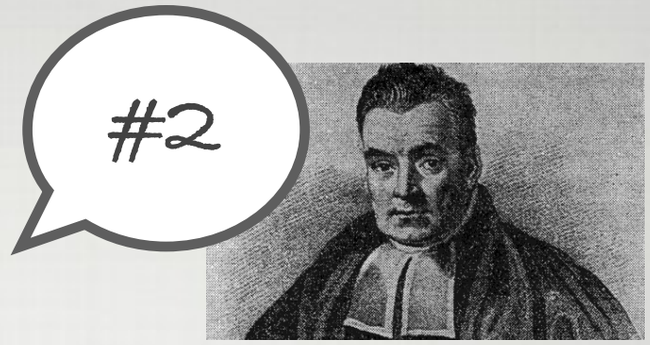
PÄÄTTELY:

$P(SANA_1 = viagra, SANA_2 = is, SANA_3 = algorithm \mid spam)$

$= P(SANA_1 = viagra \mid spam)$ // KETJUSÄÄNTÖ
$\times P(SANA_2 = is \mid SANA_1 = viagra, spam)$
$\times P(SANA_3 = algorithm \mid SANA_1 = viagra, SANA_2 = is, spam)$



NAIVI BAYES



PÄÄTTELY:

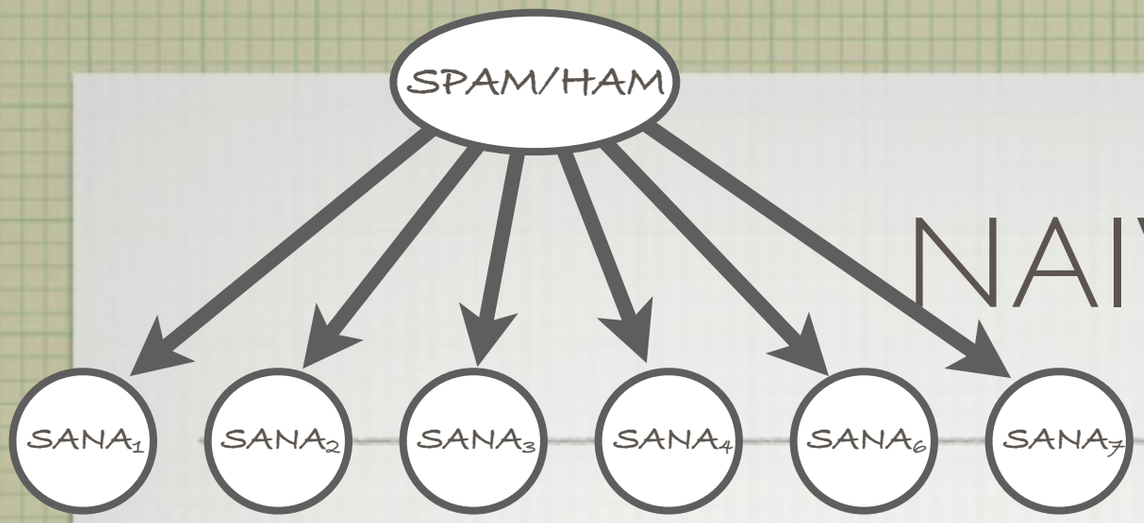
$P(SANA_1 = viagra, SANA_2 = is, SANA_3 = algorithm \mid spam)$

$= P(SANA_1 = viagra \mid spam)$ // KETJUSÄÄNTÖ

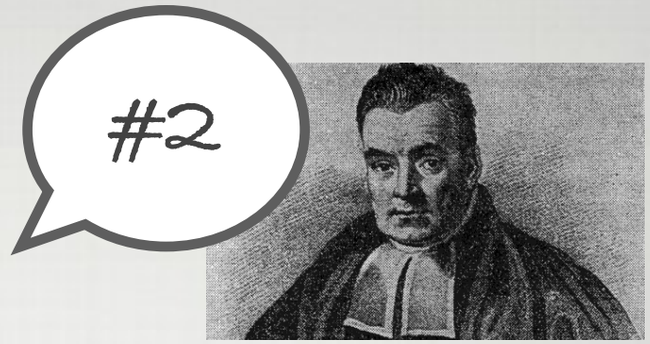
$\times P(SANA_2 = is \mid \cancel{SANA_1 = viagra}, spam)$

$\times P(SANA_3 = algorithm \mid \cancel{SANA_1 = viagra, SANA_2 = is}, spam)$

// RIIPPUMATTOMUUS



NAIVI BAYES



PÄÄTTELY:

$P(SANA_1 = viagra, SANA_2 = is, SANA_3 = algorithm \mid spam)$

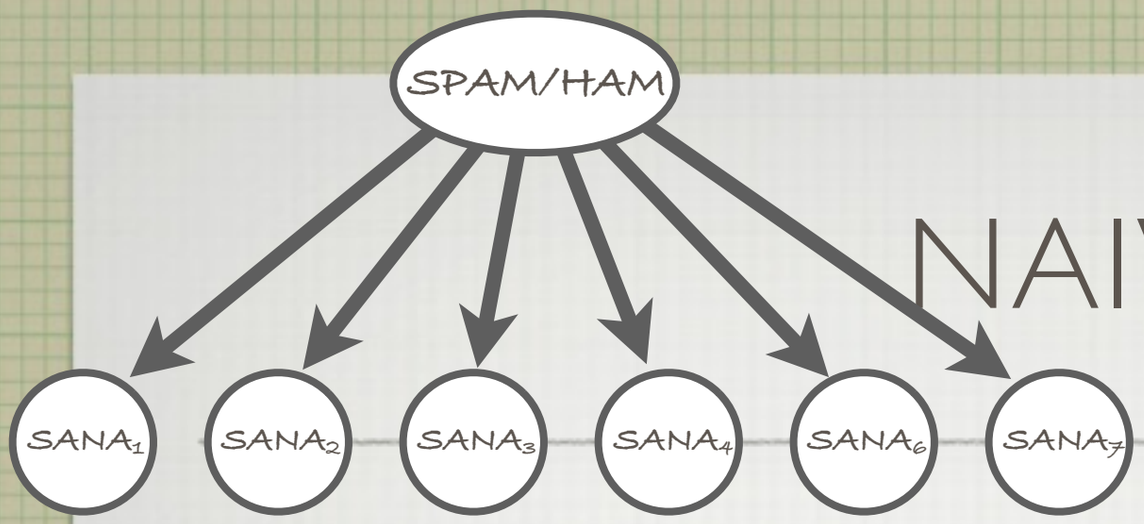
$= P(SANA_1 = viagra \mid spam)$
$\times P(SANA_2 = is \mid spam)$
$\times P(SANA_3 = algorithm \mid spam)$

$$\frac{P(S \mid E)}{P(\neg S \mid E)} = \frac{P(S)P(E \mid S)}{P(\neg S)P(E \mid \neg S)}$$

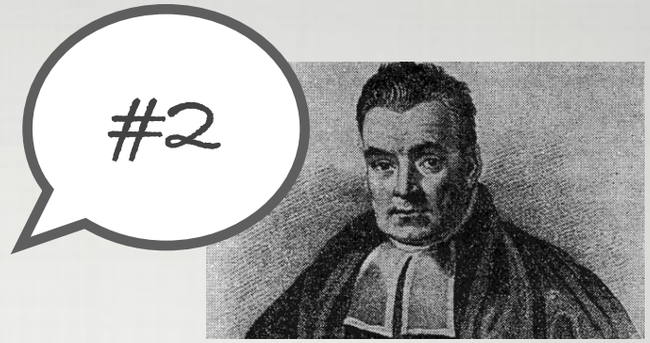
$P(SANA_1 = viagra, SANA_2 = is, SANA_3 = algorithm \mid ham)$

$= P(SANA_1 = viagra \mid ham)$
$\times P(SANA_2 = is \mid ham)$
$\times P(SANA_3 = algorithm \mid ham)$

// KETJUSÄÄNTÖ



NAIVI BAYES



PÄÄTTELY:

$P(SANA_1 = viagra, SANA_2 = is, SANA_3 = algorithm \mid spam)$

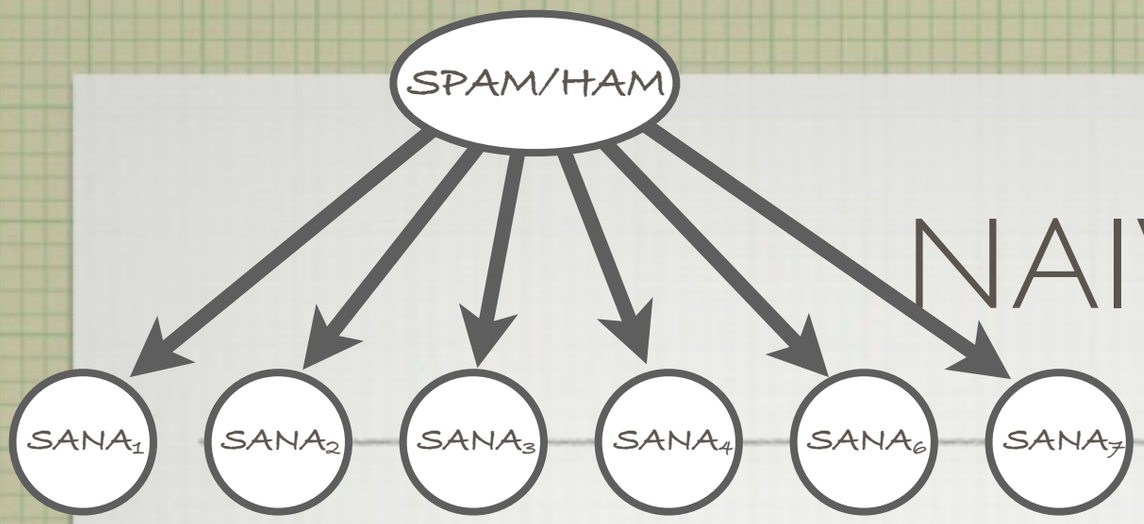
$= P(SANA_1 = viagra \mid spam)$
$\times P(SANA_2 = is \mid spam)$
$\times P(SANA_3 = algorithm \mid spam)$

$$\frac{P(S \mid E)}{P(\neg S \mid E)} = \frac{P(S)P(E \mid S)}{P(\neg S)P(E \mid \neg S)}$$

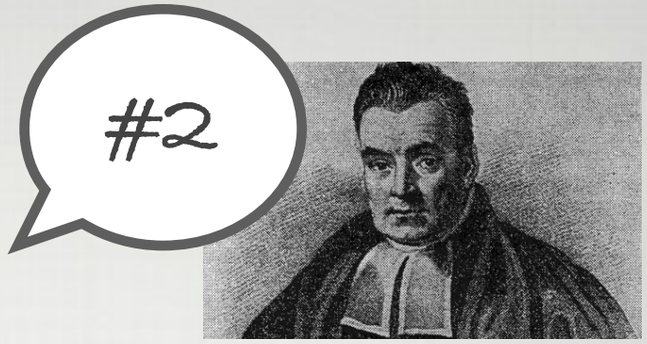
$P(SANA_1 = viagra, SANA_2 = is, SANA_3 = algorithm \mid ham)$

$= P(SANA_1 = viagra \mid ham)$
$\times P(SANA_2 = is \mid ham)$
$\times P(SANA_3 = algorithm \mid ham)$

// KETJUSÄÄNTÖ



NAIVI BAYES



PÄÄTTELY:

$P(spam|evidenssi) / P(ham|evidenssi)$

$= P(spam) / P(ham)$

$\times P(SANA_1=viagra|spam) / P(SANA_1=viagra|ham)$ >1

$\times P(SANA_2=is|spam) / P(SANA_2=is|ham)$ $=1$

$\times P(SANA_3=algorithm|spam) / P(SANA_3=algorithm|ham)$ <1

...

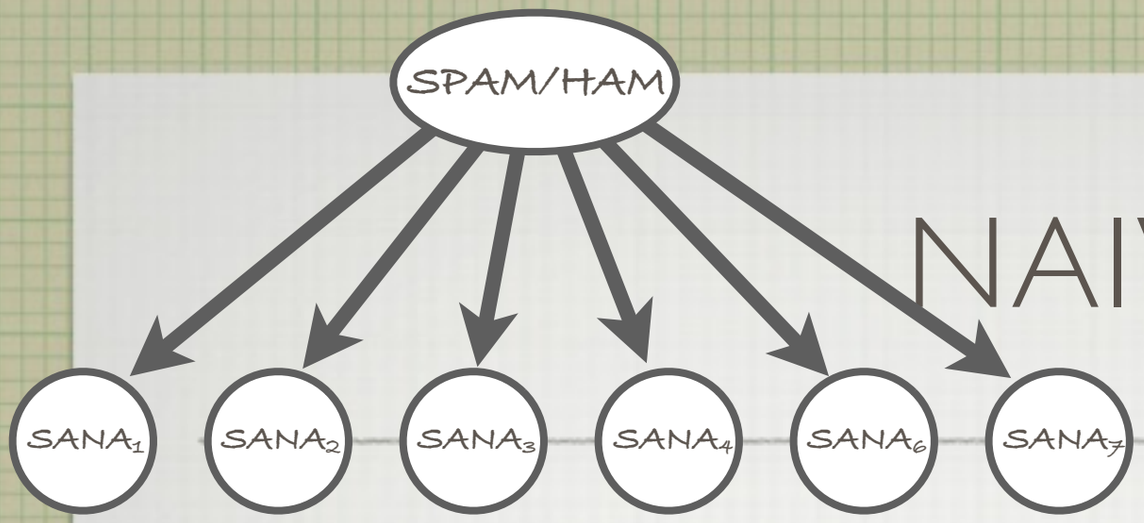
...

...

OSAMÄÄRÄ!

> 1 => LUOKITTELE: “SPAM”

< 1 => LUOKITTELE: “HAM”



NAIVI BAYES

YHTEENVETO TOISTAISEKSI:

- TARVITAAN:
  - "PRIORIJAKAUMA" $P(spam) = 0.\_\_\_$
  - "LUOKKAEHDOLLISET" JAKAUMAT

$P(SANA_i=viagra \mid spam)=0.\_\_\_$ $P(SANA_i=viagra \mid ham)=0.\_\_\_$

$P(SANA_i=is \mid spam) = 0.\_\_\_\_$ $P(SANA_i=is \mid ham)=0.\_\_\_$

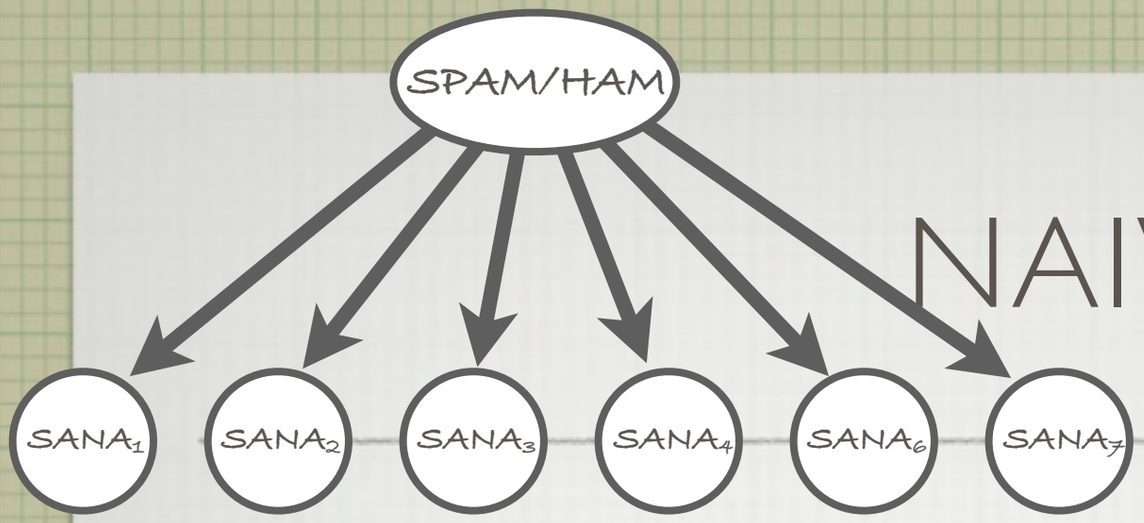
$P(SANA_i=algorithm \mid spam) = 0.\_\_\_$ $P(SANA_i=algorithm \mid ham)=0.\_\_\_$

- OLETETAAN ETTÄ $P(SANA_i \mid SANA_j, spam) = P(SANA_i \mid spam)$
  (EHDOLLINEN RIIPPUMATTOMUUS)

- OLENNAISTA ON OSAMÄÄRÄ

$$\frac{P(SANA_i=viagra \mid spam)}{P(SANA_i=viagra \mid ham)}$$

(OTETAAN NÄIDEN TULO)



NAIVI BAYES

PSEUDOKOODINA:

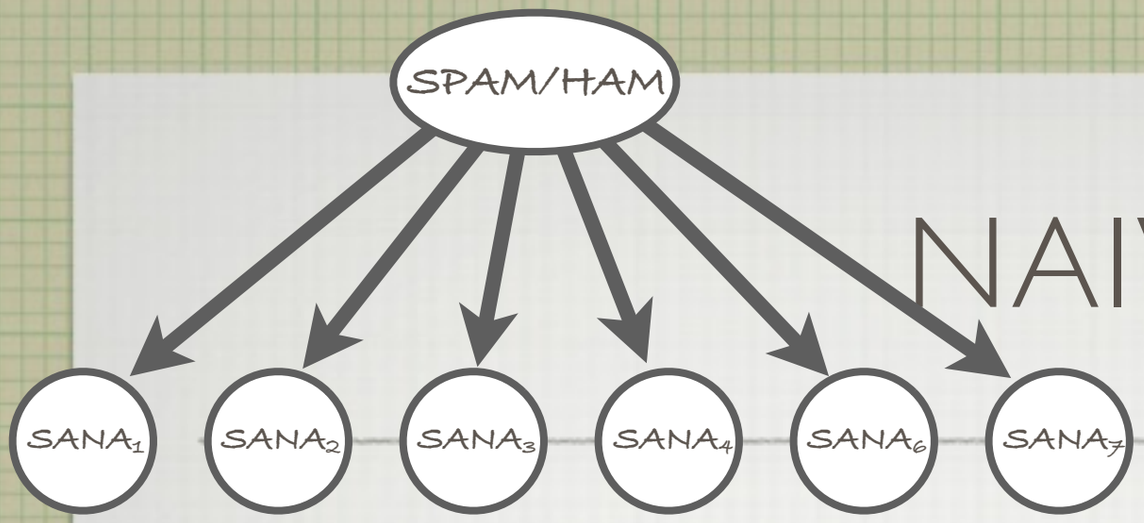
SPAMICITY(Viesti, P):

Odds = P.Spam / P.noSpam
for each Sana in Viesti
Odds = Odds * P.Sana_Spam(Sana) /P.Sana_noSpam(Sana)
return(Odds)

PÄÄTTELY:

P(spam|EVIDENSSI)/P(ham|EVIDENSSI)

= P(spam) / P(ham)
P(SANA$_1$=viagra| spam) / P(SANA$_1$=viagra| ha
P(SANA$_2$=is| spam) / P(SANA$_2$=is| ham)
P(SANA$_3$=algorithm| spam) / P(SANA$_3$=algorit



NAIVI BAYES

PSEUDOKOODINA:

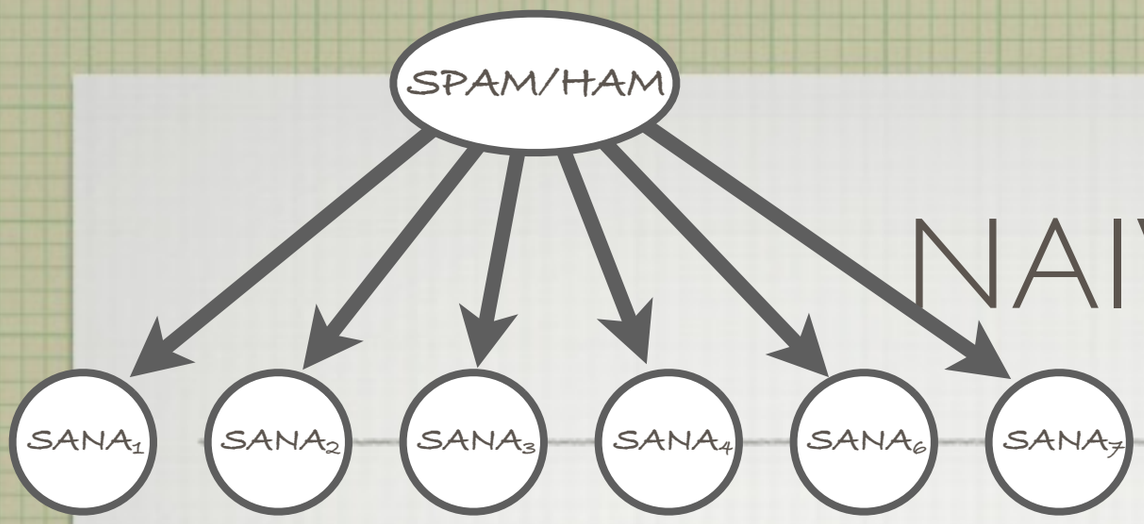
SPAMICITY(Viesti, P):

Odds = P.Spam / P.noSpam
for each Sana in Viesti
  Odds = Odds * P.Sana_Spam(Sana) /P.Sana_noSpam(Sana)
return(Odds)

PÄÄTTELY:

P(spam|EVIDENSSI)/P(ham|EVIDENSSI)

= P(spam) / P(ham)
P(SANA1=viagra| spam) / P(SANA1=viagra| ha
P(SANA2=is| spam) / P(SANA2=is| ham)
P(SANA3=algorithm| spam) / P(SANA3=algorit



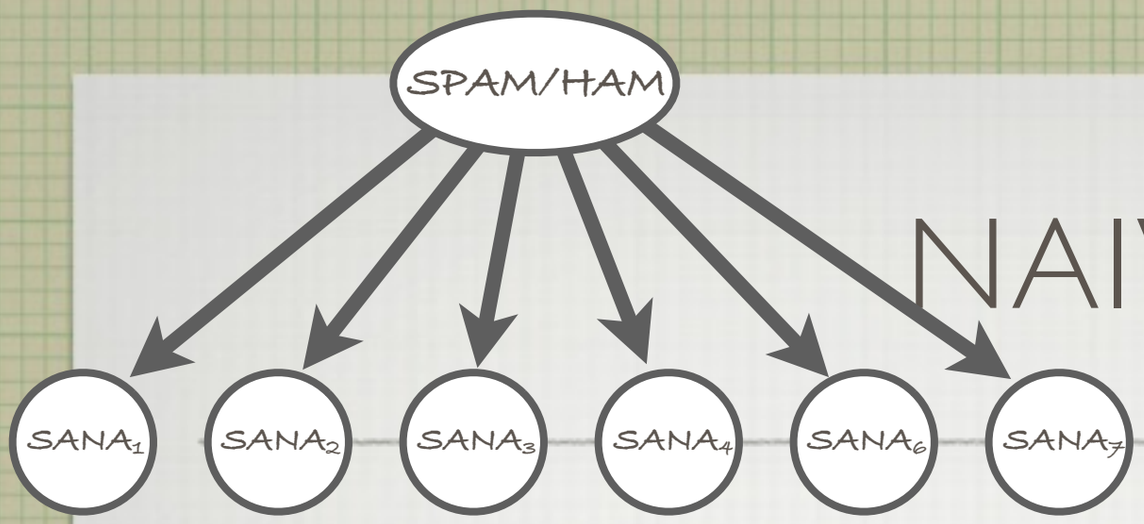
NAIVI BAYES

PSEUDOKOODINA:

SPAMICITY(Viesti, P):

```
Odds = P.Spam / P.noSpam
for each Sana in Viesti
   Odds = Odds * P.Sana_Spam(Sana) /P.Sana_noSpam(Sana)
return(Odds)
```

JOS SPAMICITY(Viesti, P) >1, LUOKITTELE VIESTI SPAMIKSI
JOS SPAMICITY(Viesti, P) <1, LUOKITTELE VIESTI HAMIKSI
JOS SPAMICITY(Viesti, P) =1, EN TIEDÄ



NAIVI BAYES

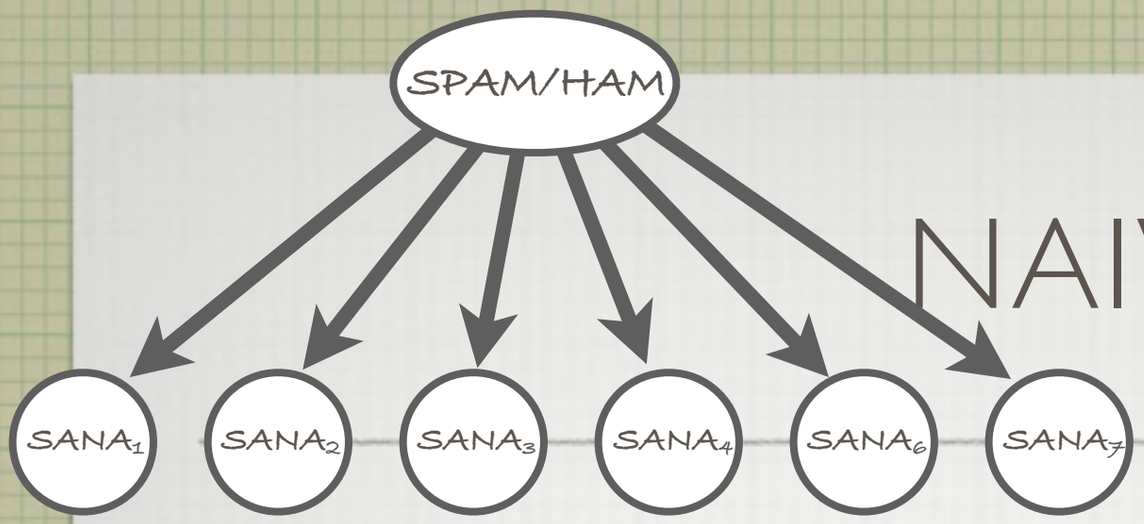
PSEUDOKOODINA:

SPAMICITY(Viesti, P):

Odds = P.Spam / P.noSpam
for each Sana in Viesti
Odds = Odds * P.Sana_Spam(Sana) /P.Sana_noSpam(Sana)
return(Odds)

JOS SPAMICITY(Viesti, P) >1+α, LUOKITTELE VIESTI SPAMIKSI
JOS SPAMICITY(Viesti, P) <1-β, LUOKITTELE VIESTI HAMIKSI
MUUTEN, EN TIEDÄ

! Epäsymmetrinen kustannusfunktio:
Asiallisen viestin luokittelu spamiksi pahempi virhe kuin toisin päin



NAIVI BAYES

PSEUDOKOODINA:

LOG(A*B) = LOG(A) + LOG(B)

SPAMICITY(Viesti, P):

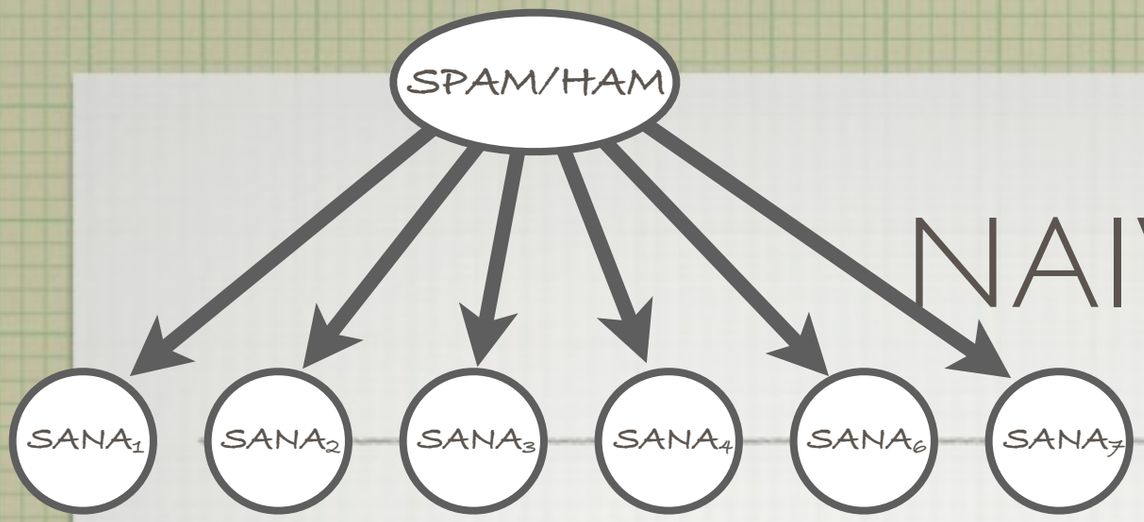
logOdds = log(P.Spam / P.noSpam)
for each Sana in Viesti
logOdds = logOdds + log(P.Sana_Spam(Sana) / P.Sana_noSpam(Sana))
return(exp(logOdds))

KÄYTÄNNÖN ONGELMA: ALI- JA YLIVUODOT

Odds ARVOSTA TULEE HELPOSTI LIIAN PIENI (LÄHELLÄ NOLLAA) TAI LIIAN SUURI.

RATKAISU:

KÄYTÄ log(Odds)



NAIVI BAYES

PARAMETRIEN OPPIMISESTA:

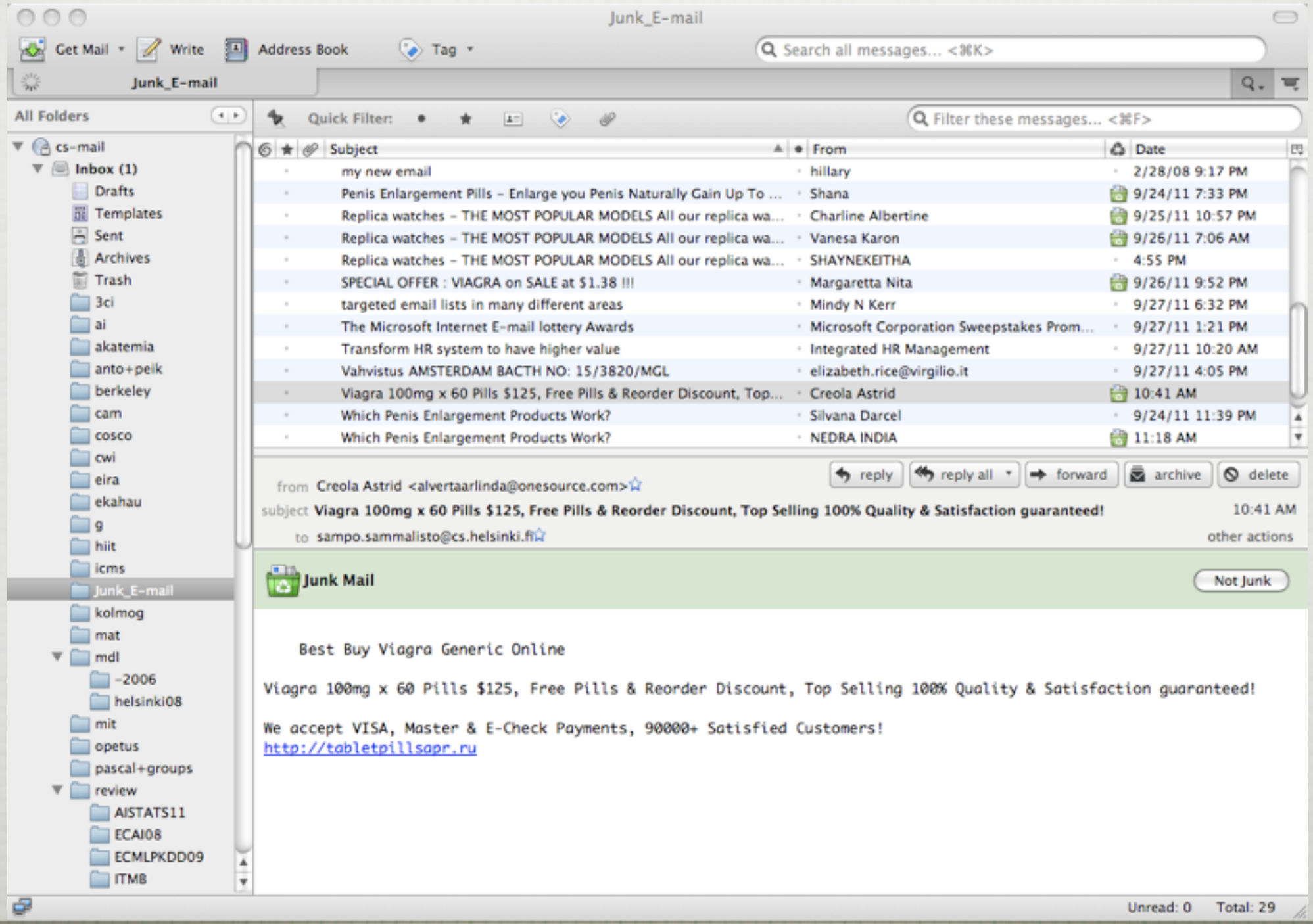
- VAIKEA KEKSIÄ PÄÄSTÄ EHDOLLISIA TN:IÄ.

- HUONOT ARVOT HUONONTAVAT FILTTERIN TOIMINTAA

- PAREMPI RATKAISU:
  1. KERÄÄ ISO KASA SPAM-VIESTEJÄ
  2. KERÄÄ ISO KASA HAM-VIESTEJÄ
  3. ARVIOI $P(SANA_i = \_\_\_ \mid SPAM) = 0.\_\_\_$ JA
     $P(SANA_i = \_\_\_ \mid HAM) = 0.\_\_\_$ LASKEMALLA DATASTA
     (VRT. LASKUHARJ.)

- VAROTTAVA NOLLATODENNÄKÖISYYKSIÄ!
  (JOS KASA EI TARPEEKSI ISO, JOTKUT SANAT EIVÄT VAIN SATU ESIINTYMÄÄN SIINÄ.)

ESIM.


Junk_E-mail
Get Mail
Write
Address Book
Tag
Search all messages... <⌘K>
Junk_E-mail
All Folders
Quick Filter:
Filter these messages... <⌘F>
cs-mail
Inbox (1)
Drafts
Templates
Sent
Archives
Trash
3ci
ai
akatemia
anto+peik
berkeley
cam
cosco
cwi
eira
ekahau
g
hiit
icms
Junk_E-mail
kolmog
mat
mdl
-2006
helsinki08
mit
opetus
pascal+groups
review
AISTATS11
ECAI08
ECMLPKDD09
ITMB
Subject
From
Date
my new email
hillary
2/28/08 9:17 PM
Penis Enlargement Pills – Enlarge you Penis Naturally Gain Up To ...
Shana
9/24/11 7:33 PM
Replica watches – THE MOST POPULAR MODELS All our replica wa...
Charline Albertine
9/25/11 10:57 PM
Replica watches – THE MOST POPULAR MODELS All our replica wa...
Vanesa Karon
9/26/11 7:06 AM
Replica watches – THE MOST POPULAR MODELS All our replica wa...
SHAYNEKEITHA
4:55 PM
SPECIAL OFFER : VIAGRA on SALE at $1.38 !!!
Margaretta Nita
9/26/11 9:52 PM
targeted email lists in many different areas
Mindy N Kerr
9/27/11 6:32 PM
The Microsoft Internet E-mail lottery Awards
Microsoft Corporation Sweepstakes Prom...
9/27/11 1:21 PM
Transform HR system to have higher value
Integrated HR Management
9/27/11 10:20 AM
Vahvistus AMSTERDAM BACTH NO: 15/3820/MGL
elizabeth.rice@virgilio.it
9/27/11 4:05 PM
Viagra 100mg x 60 Pills $125, Free Pills & Reorder Discount, Top...
Creola Astrid
10:41 AM
Which Penis Enlargement Products Work?
Silvana Darcel
9/24/11 11:39 PM
Which Penis Enlargement Products Work?
NEDRA INDIA
11:18 AM
reply
reply all
forward
archive
delete
from Creola Astrid <alvertaarlinda@onesource.com>
subject Viagra 100mg x 60 Pills $125, Free Pills & Reorder Discount, Top Selling 100% Quality & Satisfaction guaranteed!
10:41 AM
to sampo.sammalisto@cs.helsinki.fi
other actions
Junk Mail
Not Junk
Best Buy Viagra Generic Online
Viagra 100mg x 60 Pills $125, Free Pills & Reorder Discount, Top Selling 100% Quality & Satisfaction guaranteed!
We accept VISA, Master & E-Check Payments, 90000+ Satisfied Customers!
http://tabletpillsapr.ru
Unread: 0
Total: 29

ESIM.

spam		ham	
1 MONEY	0.04 %	21 ALGORITHM	0.01 %
...		...	
5 VIAGRA	0.21 %	62 MONEY	0.02 %
...		...	
10 IS	0.42 %	2199 FOR	0.78 %
...		2492 THAT	0.88 %
19 REPLICA	0.80 %	2990 YOU	1.05 %
20 EMAIL	0.84 %	3141 IN	1.11 %
20 YOU	0.84 %	3160 I	1.11 %
21 DATABASE	0.88 %	3218 AND	1.13 %
25 EMAILS	1.05 %	3283 IS	1.16 %
26 OF	1.09 %	3472 OF	1.22 %
31 TO	1.30 %	3874 A	1.37 %
43 AND	1.80 %	5442 TO	1.92 %
48 THE	2.01 %	9196 THE	3.24 %
TOTAL 2386		TOTAL 283736	

ESIM.

spam		ham	
1 MONEY	0.04 %	21 ALGORITHM	0.01 %
...		...	
5 VIAGRA	0.21 %	62 MONEY	0.02 %
...		...	
10 IS	0.42 %	2199 FOR	0.78 %
...		2492 THAT	0.88 %
19 REPLICA	0.80 %	2990 YOU	1.05 %
20 EMAIL	0.84 %	3141 IN	1.11 %
20 YOU	0.84 %	3160 I	1.11 %
21 DATABASE	0.88 %	3218 AND	1.13 %
25 EMAILS	1.05 %	3283 IS	1.16 %
26 OF	1.09 %	3472 OF	1.22 %
31 TO	1.30 %	3874 A	1.37 %
43 AND	1.80 %	5442 TO	1.92 %
48 THE	2.01 %	9196 THE	3.24 %
TOTAL 2386		TOTAL 283736	

AKTIVATOR 4 & 5

ESIM.

spam		ham	
1 MONEY	0.04 %	21 ALGORITHM	0.01 %
...		...	
5 VIAGRA	0.21 %	62 MONEY	0.02 %
...			
10 I			
...			
19 R			
20 E			
20 Y			
21 D		3218 AND	1.13 %
25 E			
26 O			
31 T			
43 A			
48 T			
TOTA			

$$\frac{P(SANA_i=MONEY|SPAM)}{P(SANA_i=MONEY|\neg SPAM)} = \frac{0.0004}{0.0002} = 1.918 > 1$$

$$\frac{P(SANA_i=IS|SPAM)}{P(SANA_i=IS|\neg SPAM)} = \frac{0.0042}{0.0116} = 0.3622 < 1$$

ESIM.

$P(\text{SANA}_i = \text{'ALGORITHM'} \mid \text{SPAM}) = 0$?
VOI JOHTAA TILANTEESEEN, JOSSA OSAMÄÄRÄ 0/0

spam

```
...
5 VIAGRA        0.21 %
...
10 IS           0.42 %
...
19 REPLICA      0.80 %
20 EMAIL        0.84 %
20 YOU          0.84 %
21 DATABASE     0.88 %
25 EMAILS       1.05 %
26 OF           1.09 %
31 TO           1.30 %
43 AND          1.80 %
48 THE          2.01 %
TOTAL 2386
```

ham

```
21 ALGORITHM    0.01 %
...
62 MONEY        0.02 %
...
2199 FOR        0.78 %
2492 THAT       0.88 %
2990 YOU        1.05 %
3141 IN         1.11 %
3160 I          1.11 %
3218 AND        1.13 %
3283 IS         1.16 %
3472 OF         1.22 %
3874 A          1.37 %
5442 TO         1.92 %
9196 THE        3.24 %
TOTAL 283736
```

YHTEENVETO

YHTEENVETO NAIVI BAYES-SPAMFILTTERISTÄ:

- TARVITAAN:
  - "PRIORIJAKAUMA" $P(SPAM) = 0.\_\_\_$
  - "LUOKKAEHDOLLISET" JAKAUMAT

$P(SANA_i=VIAGRA|SPAM)=0.\_\_\_$ $P(SANA_i=VIAGRA|\neg SPAM)=0.\_\_\_$

$P(SANA_i=IS|SPAM) = 0.\_\_\_\_$ $P(SANA_i=IS|\neg SPAM)=0.\_\_\_$

$P(SANA_i=ALGORITHM|SPAM) = 0.\_\_\_$ $P(SANA_i=ALG.|\neg SPAM)=0.\_\_\_$

- HUOM! EI OLE KYSEESSÄ P("SANA VIAGRA ESIINTYY" | SPAM), JNE.
  (SEKIN OLISI MAHDOLLISTA, MUTTA VIESTIEN PITUUDET VAIKUTTAISIVAT TULOKSEEN.)

- OLETETAAN ETTÄ $P(SANA_i|SANA_j,SPAM) = P(SANA_i|SPAM)$
  (EHDOLLINEN RIIPPUMATTOMUUS)

YHTEENVETO

(JATKOA...):

- OLENNAISTA ON OSAMÄÄRÄ

$$\frac{P(SANA_i = VIAGRA \mid SPAM)}{P(SANA_i = VIAGRA \mid \neg SPAM)}$$

- KUSTANNUSFUNKTIO EPÄSYMMETRINEN: PAREMPI SÄÄSTÄÄ MUUTAMA SPAM KUIN HUKATA OIKEA VIESTI (HAM)
- JAKAUMAT PARAS ESTIMOIDA DATASTA
- NOLLATODENNÄKÖISYYKSILLE TEHTÄVÄ JOTAIN