Use case & Requirement

- Model that classifies email messages as:
  - SPAM (Junk e-mail)
  - HAM (bonafide e-mail)

SPAM origin

The problem:

googleteam GOOGLE LOTTERY WINNER! CONTACT

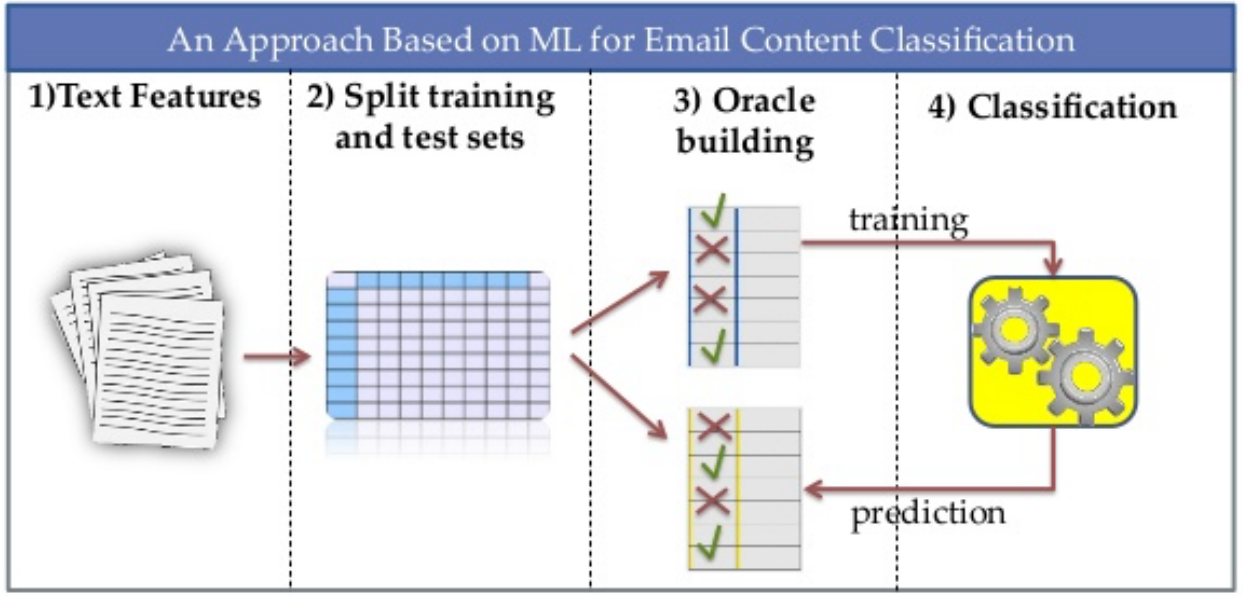
From: googleteam **To:**
Subject: GOOGLE LOTTERY WINNER! CONTACT YOUR AGENT TO CLAIM YOUR PRIZE.

GOOGLE LOTTERY INTERNATIONAL
INTERNATIONAL PROMOTION / PRIZE AWARD .
(WE ENCOURAGE GLOBALIZATION)
FROM: THE LOTTERY COORDINATOR,
GOOGLE B.V. 44 9459 PE.
RESULTS FOR CATEGORY "A" DRAWS
Congratulations to you as we bring to your notice, the results of the First Ca
inform you that your email address have emerged a winner of One Million (1,(
money of Two Million (2,000,000.00) Euro shared among the 2 winners in this
email addresses of individuals and companies from Africa, America, Asia, Au
CONGRATULATIONS!
Your fund is now deposited with the paying Bank. In your best interest to avo
award strictly from public notice until the process of transferring your claims
NOTE: to file for your claim, please contact the claim department below on e
****

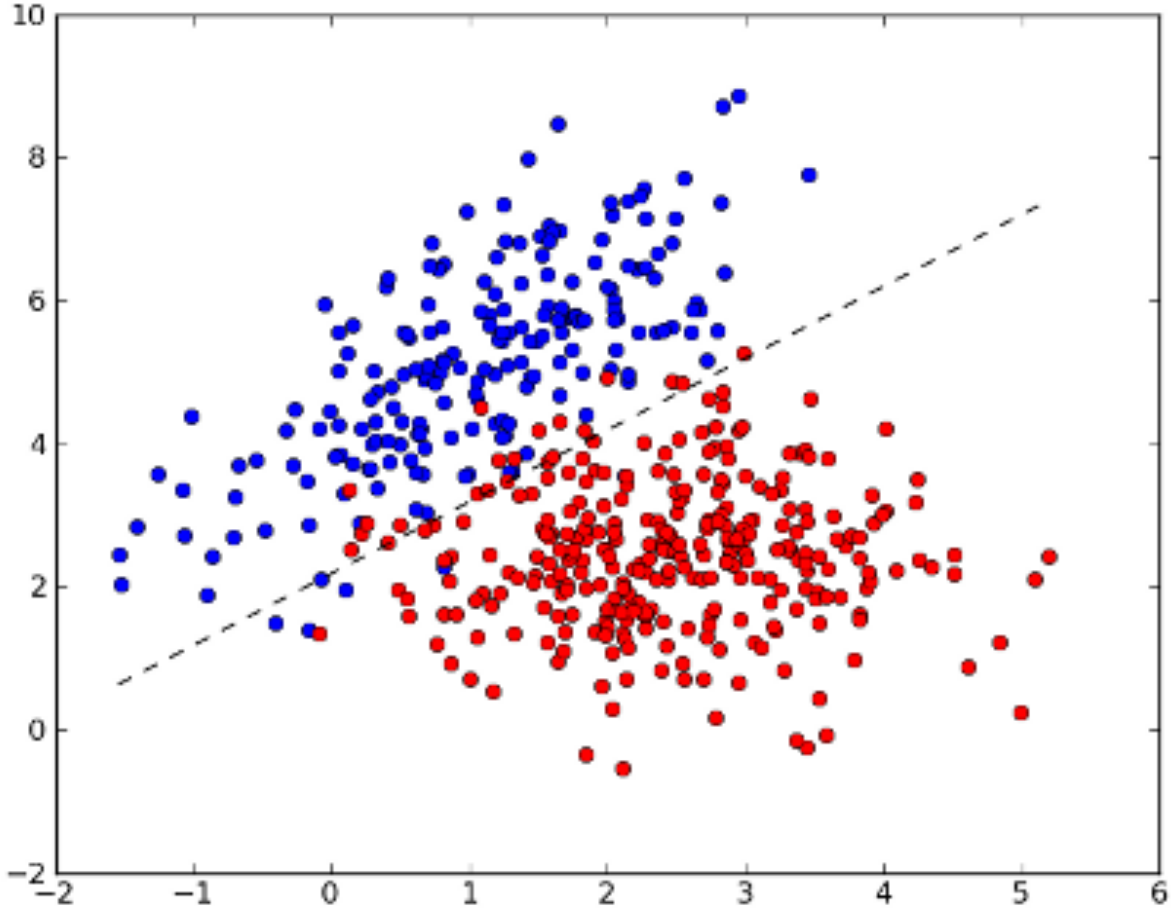
Strategy

ML for Email Classification

→ Antoniol et. al., CASCON 2008
→ Zhou et al. , ICSME 2014



What classifiers try to do:



The data

- HAM (bonafide e-mail)
  - E-mails from six Enron employees
- SPAM (junk / unsollicited e-mail)
  - the SpamAssassin corpus
  - the Honeypot project
  - Spam collection of Bruce Guenter
  - Spam collection of Georgios Paliouras

Spambase

1. Title: SPAM E-mail Database

2. Sources:
   (a) Creators: Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt
       Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304
   (b) Donor: George Forman (gforman at nospam hpl.hp.com) 650-857-7835
   (c) Generated: June-July 1999

3. Past Usage:
   (a) Hewlett-Packard Internal-only Technical Report. External forthcoming.
   (b) Determine whether a given email is spam or not.
   (c) ~7% misclassification error.
       False positives (marking good mail as spam) are very undesirable.
       If we insist on zero false positives in the training/testing set,
       20-25% of the spam passed through the filter.

4. Relevant Information:
       The "spam" concept is diverse: advertisements for products/web
       sites, make money fast schemes, chain letters, pornography...
       Our collection of spam e-mails came from our postmaster and
       individuals who had filed spam. Our collection of non-spam
       e-mails came from filed work and personal e-mails, and hence
       the word 'george' and the area code '650' are indicators of
       non-spam. These are useful when constructing a personalized
       spam filter. One would either have to blind such non-spam
       indicators or get a very wide collection of non-spam to
       generate a general purpose spam filter.

       For background on spam:
       Cranor, Lorrie F., LaMacchia, Brian A. Spam!
       Communications of the ACM, 41(8):74-83, 1998.

5. Number of Instances: 4601 (1813 Spam = 39.4%)

6. Number of Attributes: 58 (57 continuous, 1 nominal class label)

Spambase word frequency

```
1, 0.    | spam, non-spam classes

word_freq_make:          continuous.
word_freq_address:       continuous.
word_freq_all:           continuous.
word_freq_3d:            continuous.
word_freq_our:           continuous.
word_freq_over:          continuous.
word_freq_remove:        continuous.
word_freq_internet:      continuous.
word_freq_order:         continuous.
word_freq_mail:          continuous.
word_freq_receive:       continuous.
word_freq_will:          continuous.
word_freq_people:        continuous.
word_freq_report:        continuous.
word_freq_addresses:     continuous.
word_freq_free:          continuous.
word_freq_business:      continuous.
```

SPAM:

80 free spins and a €300 welcome package are waiting for you!

PLAY NOW

Follow these steps to receive your free spins:

- Visit MegaCasino.com
- Receive 10 free spins on Ninja Master, upon registration
- Deposit using the bonus code: **Mega80**
- Receive an additional **70 free spins** on your following 3 deposits **+ €300 welcome package**

These are some truly excellent opportunities to WIN BIG! GO FOR IT!

PLAY NOW

Free Occurs 4 times in message body

Good Luck!
Kevin Cairns, Games Development
Mega Casino

PyCharm

Spambase CAPITAL_run_length

```
capital_run_length_average: continuous.
capital_run_length_longest: continuous.
capital_run_length_total:   continuous.
```

SPAM:

KA-CHING!

That the sound of your first affiliate commission if you take *IMMEDIATE* action right now.

If you're completely new to making money on the internet...

Or if you're struggling to make your first dime online...

Get ready. Everything is about to change...

Click here NOW ...Go...Go...Go!

Enjoy it :)

- Steven A.
CEO, Tesler Investments

Model performance – existing models

NB version	Enr1	Enr2	Enr3	Enr4	Enr5	Enr6	Avg.
FB	90.50	93.63	96.94	95.78	99.56	99.55	95.99
MV Gauss	93.08	95.80	97.55	80.14	95.42	91.95	92.32
MN TF	95.66	96.81	95.04	97.79	99.42	98.08	97.13
MV Bern.	97.08	91.05	97.42	97.70	97.95	97.92	96.52
MN Bool.	96.00	96.68	96.94	97.79	99.69	98.10	**97.53**

Table 4: Spam recall (%) for 3000 attributes, $T = 0.5$.

NB version	Enr1	Enr2	Enr3	Enr4	Enr5	Enr6	Avg.
FB	97.64	98.83	95.36	96.61	90.76	89.97	94.86
MV Gauss	94.83	96.97	88.81	99.39	97.28	95.87	95.53
MN TF	94.00	96.78	98.83	98.30	95.65	95.12	96.45
MV Bern.	93.19	97.22	75.41	95.86	90.08	82.52	89.05
MN Bool.	95.25	97.83	98.88	99.05	95.65	96.88	**97.26**

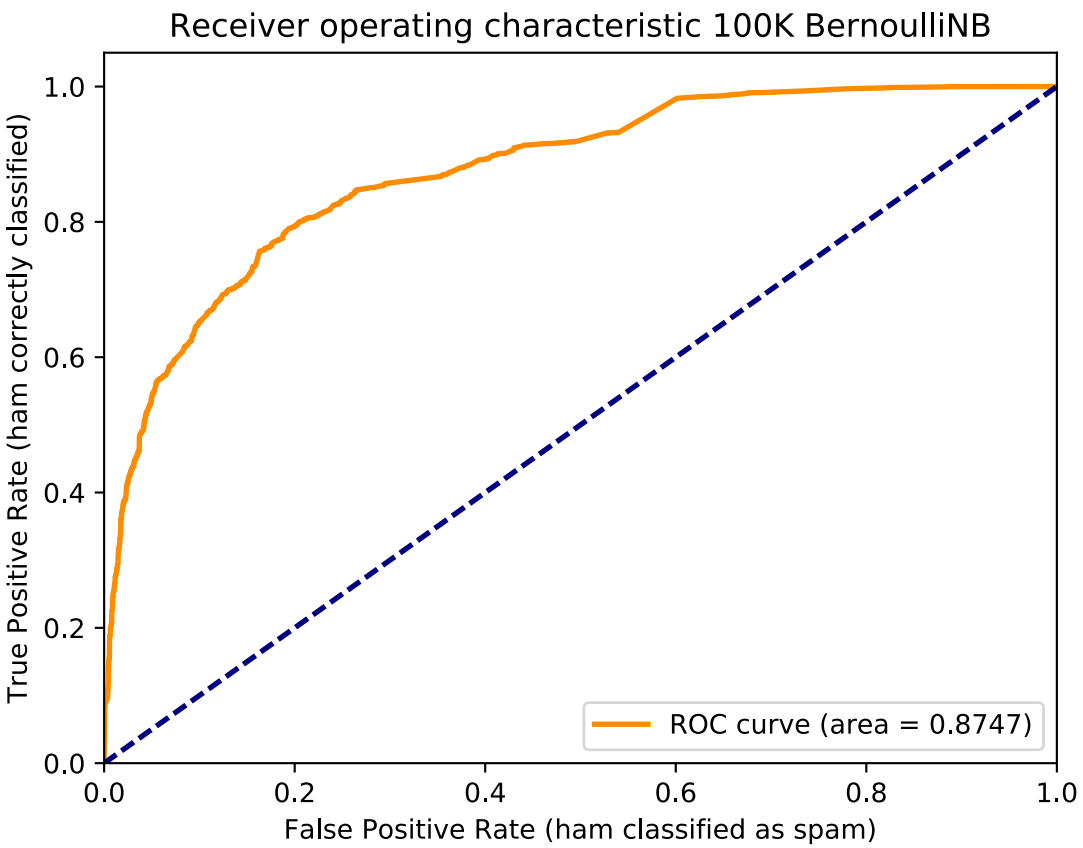
Table 5: Ham recall (%) for 3000 attributes, $T = 0.5$.

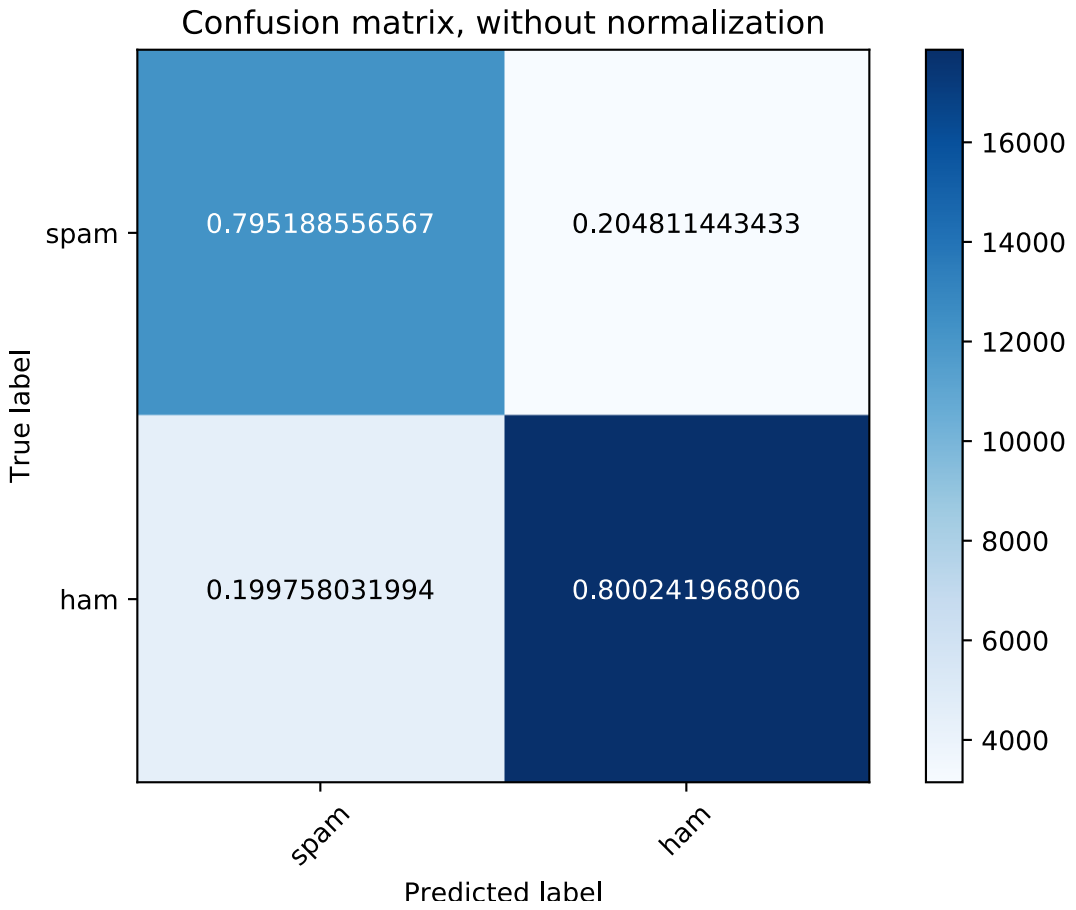
CEAS 2006 - Third Conference on Email and Anti-Spam, July 27-28, 2006, Mountain View, California USA

Considerations

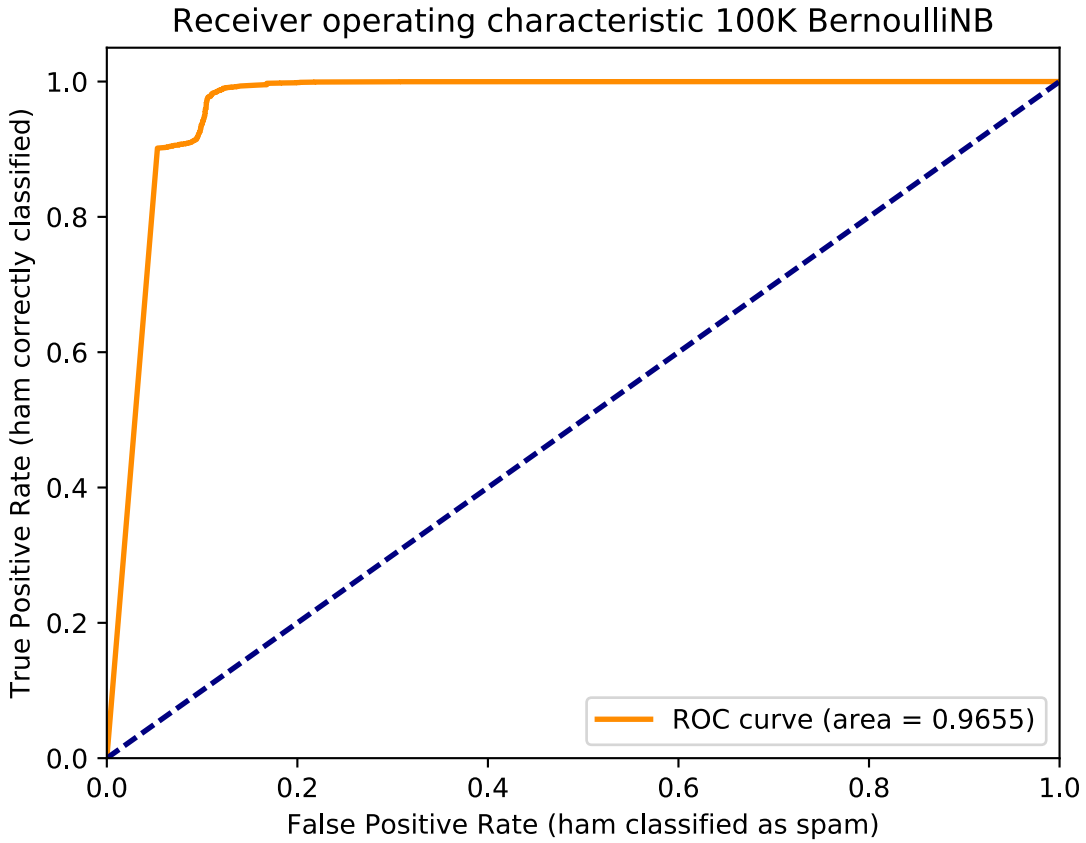
- Personalized SPAM filters
- False positives (marking good mail as spam) are very undesirable.
- ML algorithms are interesting because they can change the way they classify based on their input. So, if a classifier stops working well after a period of time (because the form of spam has changed), one merely needs to rebuild the classifier using more recent emails and the ML will output a new classifier that's much more effective. In this way, the filter can never be outdated, and no matter how hard they try, spammers won't be able to get their wares past our dutiful filter.

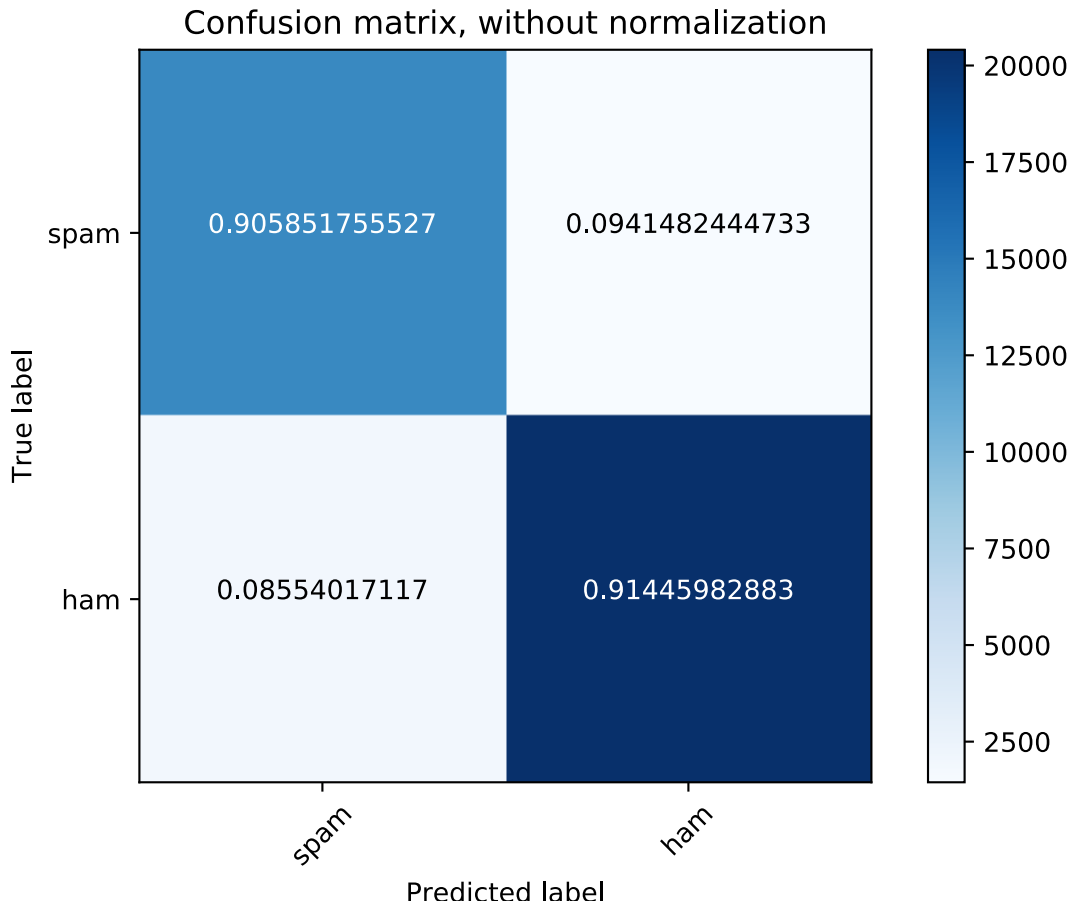
Bernoulli Naive Bayes Performance
48 Spambase features on 100K Enron emails



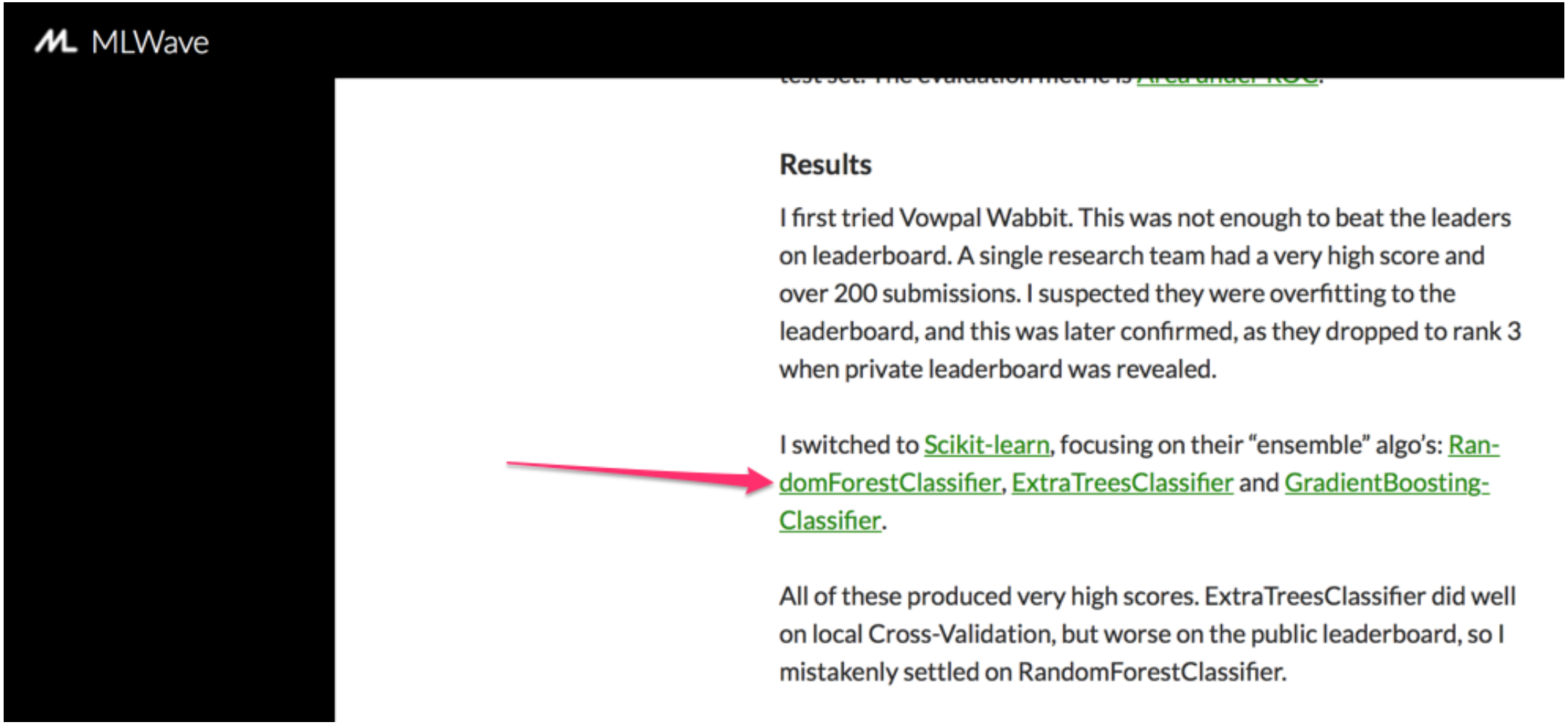


Bernoulli Naive Bayes Performance
527 New features



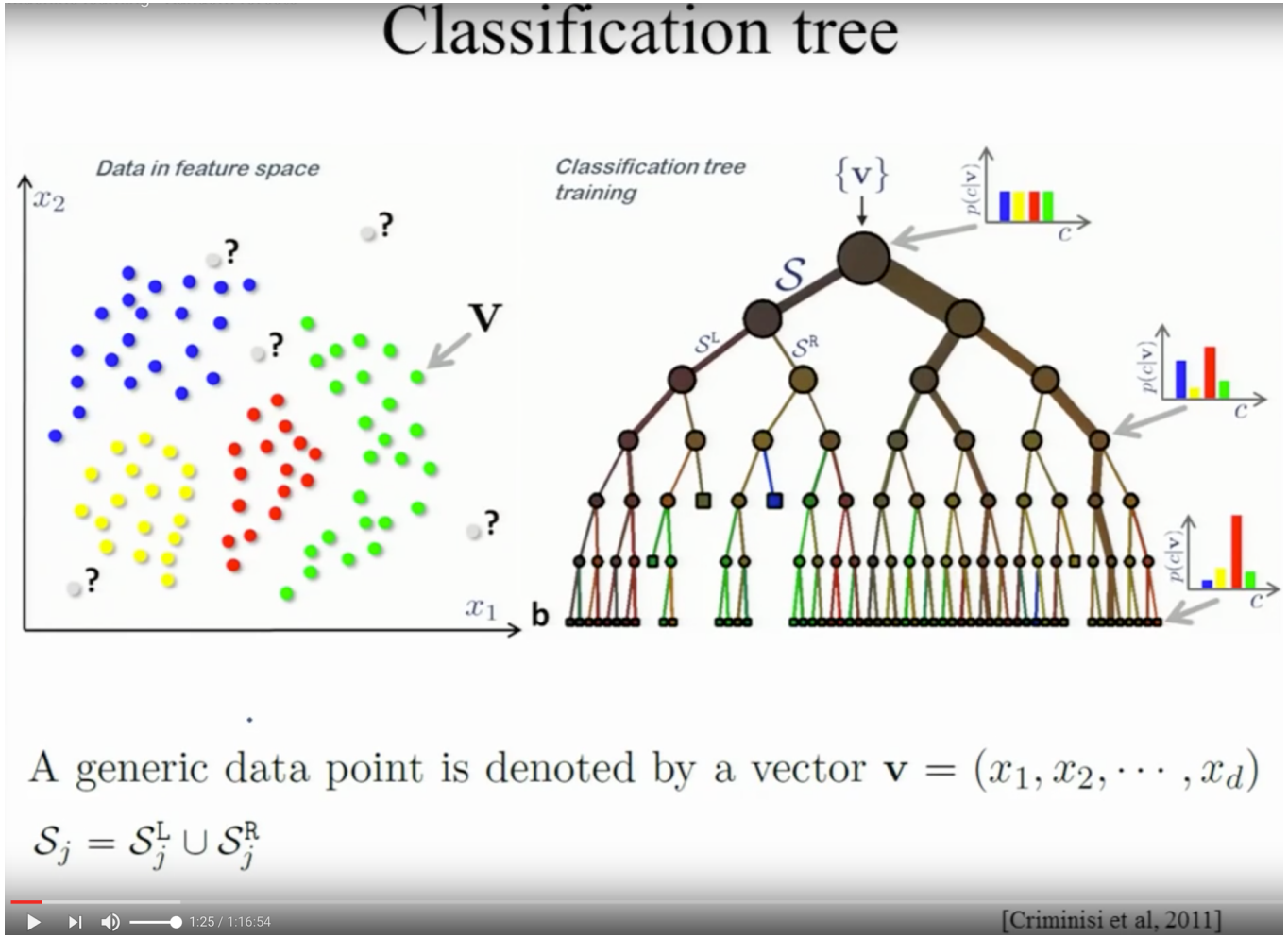


A winning spam filtering strategy

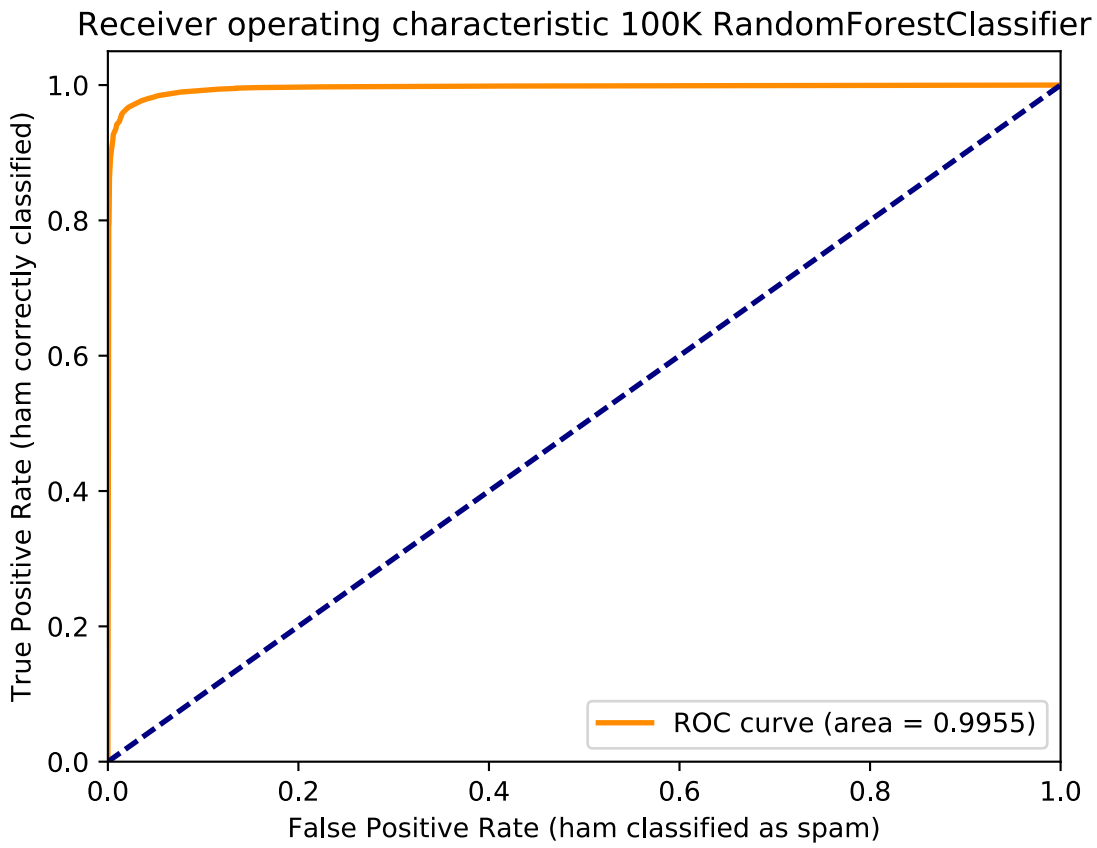


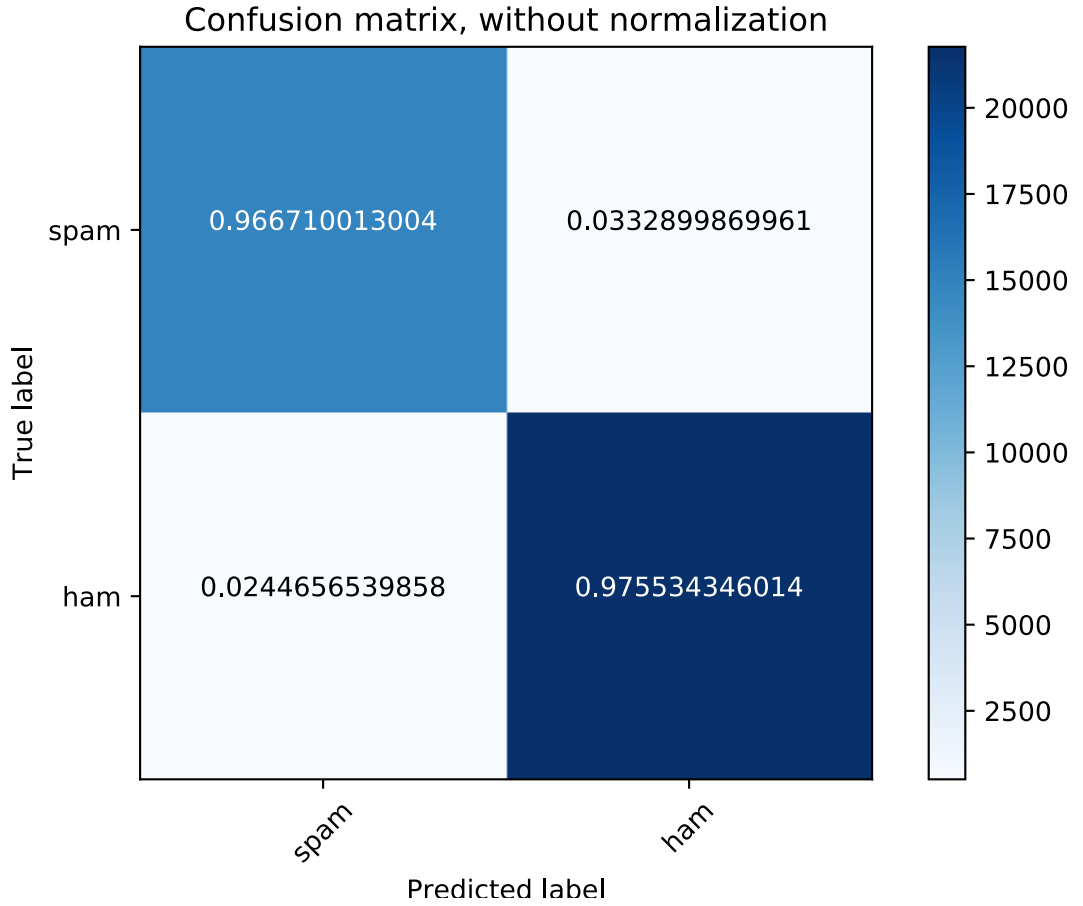
Random Forest

- Decision tree based
- Insensitive to overtraining
- Relatively simple

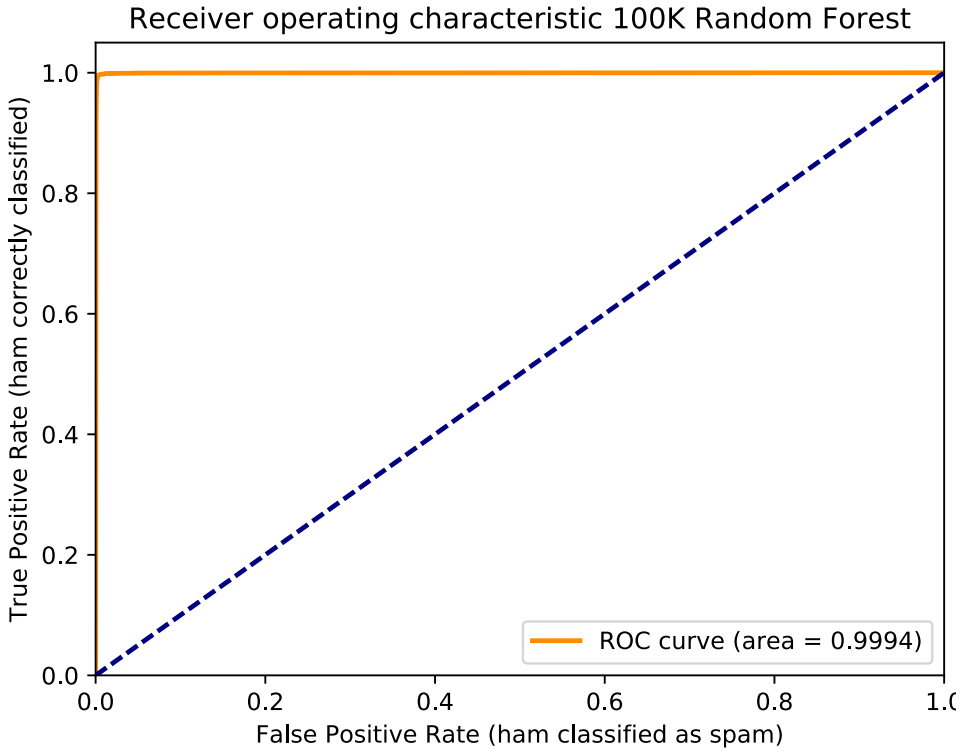


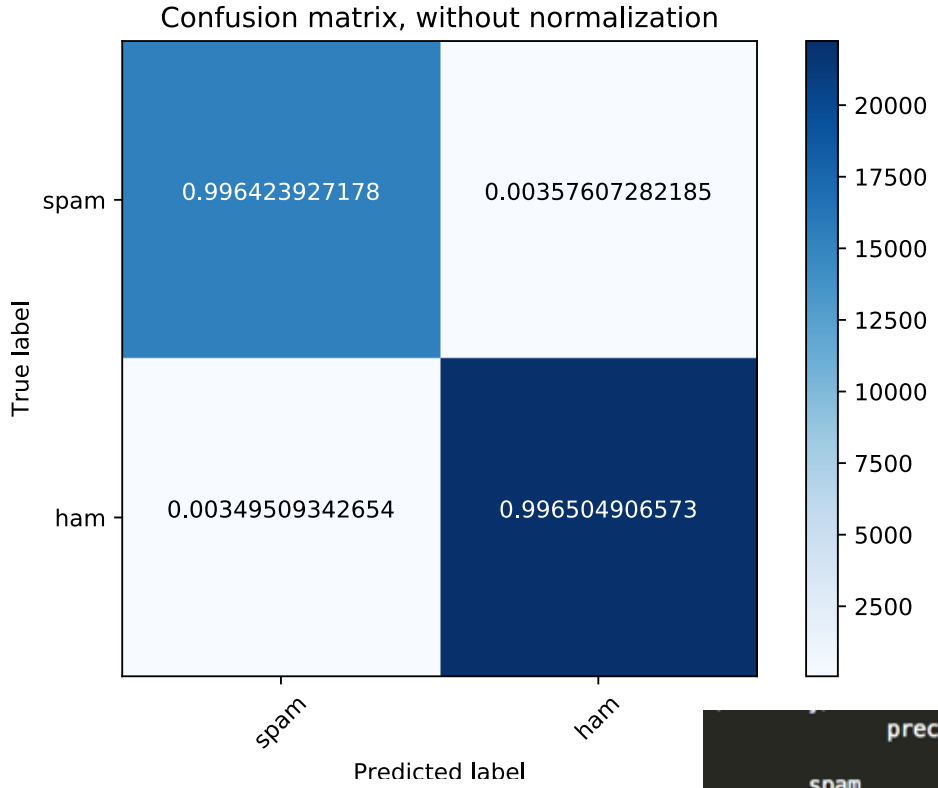
Random Forest Performance
48 Spambase features





Random Forest Performance
527 extracted features





	precision	recall	f1-score	support
spam	0.99	1.00	1.00	15380
ham	1.00	1.00	1.00	22317
avg / total	1.00	1.00	1.00	37697

False positives: 0.36%
False negatives: 0.35%

emails used for training 56545
emails used for testing 37697