A Human Factors Approach to Spam Filtering

Robert Beverly

MIT CSAIL
rbeverly@csail.mit.edu
July 27, 2009

Conference on Email and Anti-Spam 2009

No spam classifier is perfect

Okay in other ML fields, e.g.

- Handwriting recognition, search engines, music recommendation, etc.

But with spam:

- Adaptable, adversarial inputs
- Complexion of dataset severely unbalanced
- High cost of false positives
- Getting from 99.9% to 99.999%

Fighting a losing battle?

No spam classifier is perfect

Okay in other ML fields, e.g.

- Handwriting recognition, search engines, music recommendation, etc.

But with spam:

- Adaptable, adversarial inputs
- Complexion of dataset severely unbalanced
- High cost of false positives
- Getting from 99.9% to 99.999%

Fighting a losing battle?

No spam classifier is perfect

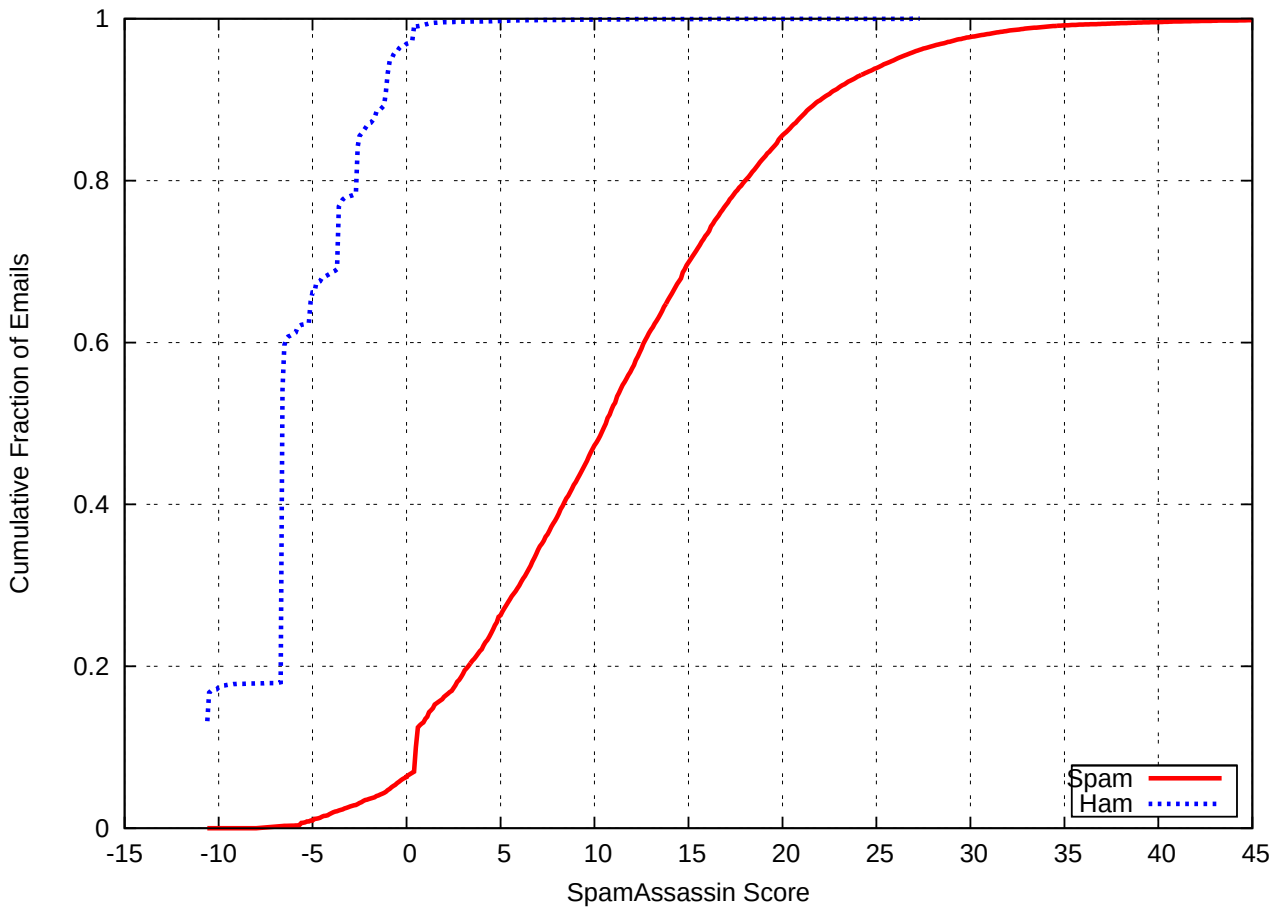
Okay in other ML fields, e.g.

- Handwriting recognition, search engines, music recommendation, etc.

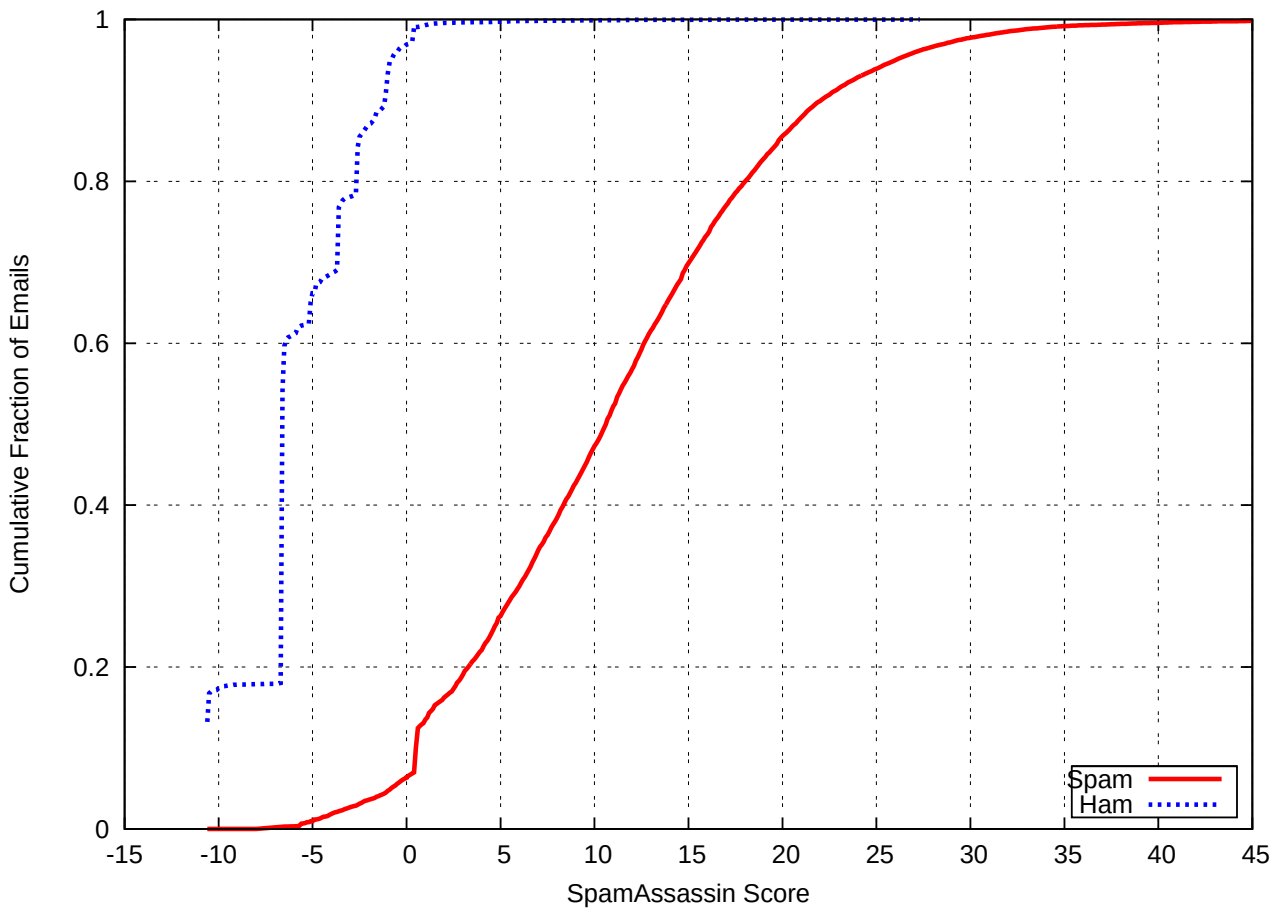
But with spam:

- Adaptable, adversarial inputs
- Complexion of dataset severely unbalanced
- High cost of false positives
- Getting from 99.9% to 99.999%

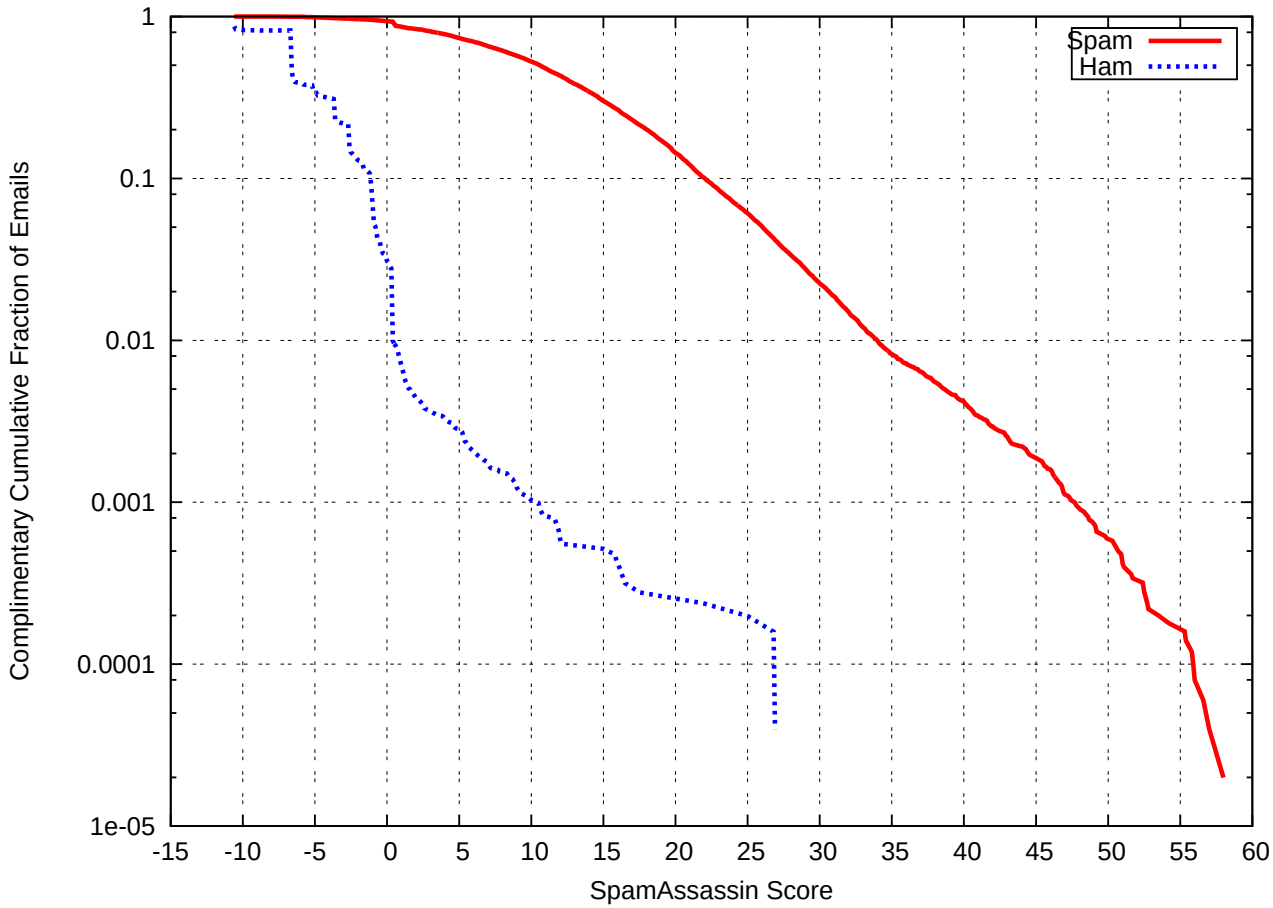
Fighting a losing battle?



- TREC 2007 dataset (∼75k messages)
- Classified with SpamAssassin
- How close are mails to the threshold (5)?



- How close are mails to the threshold (5)?
- 99.72% of ham below threshold... good?



- *No* threshold gives zero FP/FN (well-known compromise)
- Deluge of spam implies this compromise is flawed
- 0.28% above → 71 false positives

Approaching from a different direction...

The User Agent:

- Users interact with their email via a Mail User Agent (MUA), e.g. Outlook, Hotmail, etc.
- Note that besides going graphical, MUAs have changed little over past $\sim$ 30 years
- Better incorporate human factors into a MUA

Human Factors Approach – Potential:

1. Make email more useful to the user
   - How are emails presented?
2. Humans ultimate arbiter of any mail's importance
   - How to better include, scale their decision process?
3. Remove burden of perfect classification from classifier
   - "good enough" filtering
4. Eliminate false positives

Innovate in the user agent

Human Factors Approach – Potential:

1. Make email more useful to the user
   - How are emails presented?
2. Humans ultimate arbiter of any mail's importance
   - How to better include, scale their decision process?
3. Remove burden of perfect classification from classifier
   - "good enough" filtering
4. Eliminate false positives

Innovate in the user agent

Position

- Separate *classification* from *filtering*

The `inbox`:

- Rethink the inbox: use a **single** mail folder, don't attempt to filter into spam, ham "folders"
- Use color, size, shade, order, and other *human factors* to present the inbox
- Presentation of email a function of *importance*

Proof-of-concept: SpamGUI Thunderbird extension...

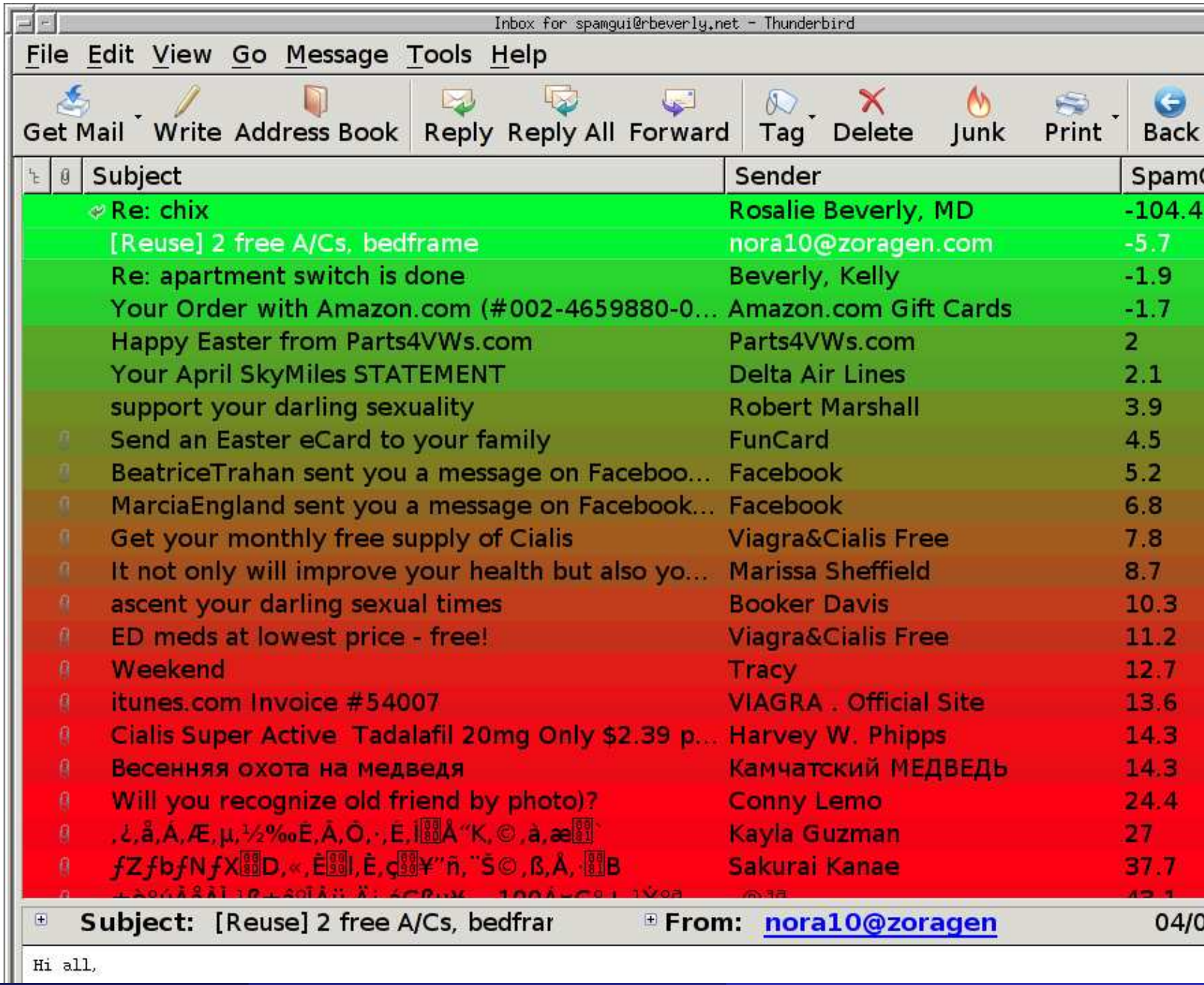
Position

- Separate *classification* from *filtering*

The `inbox`:

- Rethink the inbox: use a **single** mail folder, don't attempt to filter into spam, ham "folders"
- Use color, size, shade, order, and other *human factors* to present the inbox
- Presentation of email a function of *importance*

Proof-of-concept: SpamGUI Thunderbird extension...



A Few Observations:

- A demarcation “line” naturally emerges to the eye, above which user (or UI) can ignore messages
- User part of filtering process, but only burdened by making spam decisions on a small number of emails around line
- Easy to scan for formerly false positive emails on the threshold border

Lots of work remains:

- No user studies performed yet
- Experimenting with several approaches

More generally:

- Users inundated with information, how can UI help?
- Spam is just one class of very unimportant information
- Lots of unused input "features;" systems designers should use them
- Learn best way to present email to user

Recognize that innovation is possible in the user agent

- We're fighting a losing battle trying to make spam classifiers perfect
- Separate act of classification from filtering
- As a community, think more about how HCI / human factors methods can help

Thanks!

`http://www.rbeverly.net/spamgui/`

Questions?