

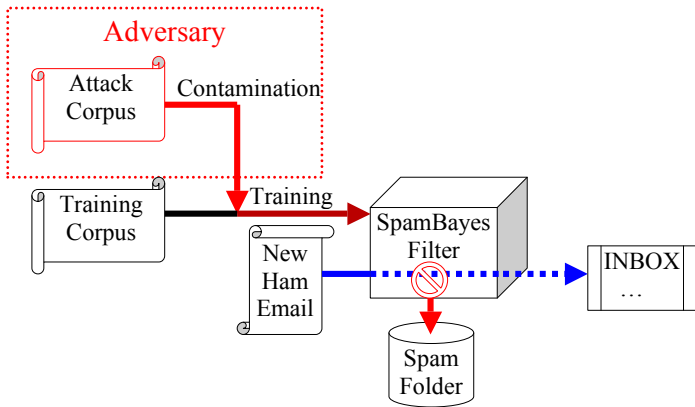
# Attacking SpamBayes: Compromising a Statistical Spam Filter

Marco Barreno   Fuching Jack Chi   Anthony D. Joseph  
Blaine Nelson   Benjamin I. P. Rubinstein   Udam Saini  
Charles Sutton   J.D. Tygar   Kai Xia

Department of Computer Science  
University of California at Berkeley

December 8, 2007

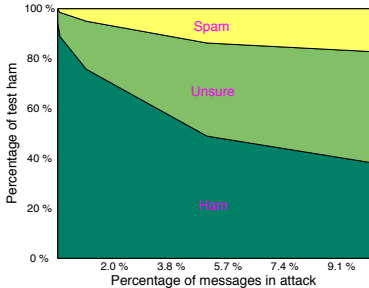
## Attacks against SpamBayes



- Attacks consist of messages that mislead SpamBayes.
- Attack messages cause filter to misclassify ham as spam.
- Result is denial-of-service making filter unusable.

# Our Attacks

**Dictionary Attack** — makes rare words look like spam tokens.



**Targeted Attack** — makes targeted message appear to be spam.

