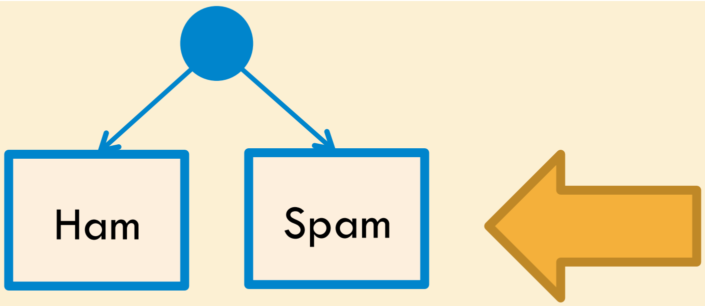
COARSE-TO-FINE, COST-SENSITIVE CLASSIFICATION OF E-MAIL

Jay Pujara jay@cs.umd.edu
Lise Getoor getoor@cs.umd.edu 12/10/2010

Parallel Coarse-to-Fine Problems

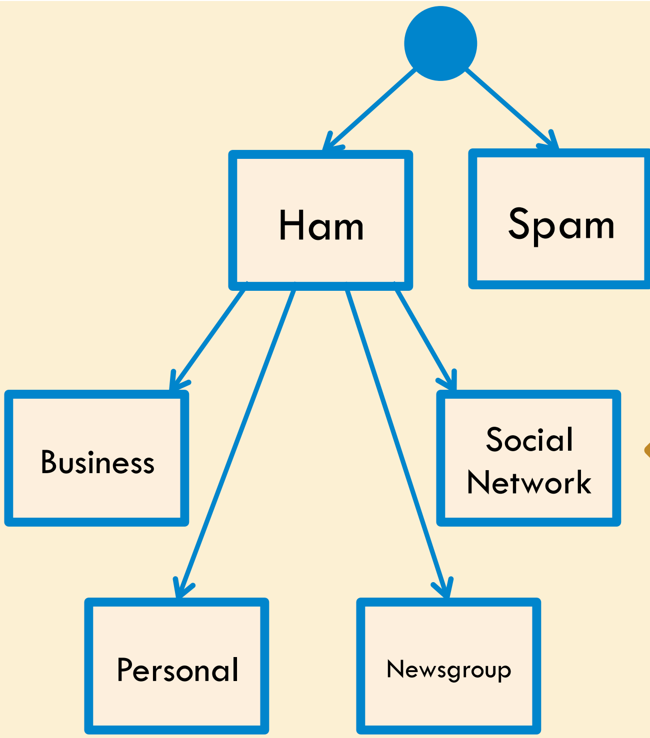
- Structure in output
  - Labels naturally have a hierarchy from coarse-to-fine
- Structure in input
  - Features may have an order or systemic dependency
  - Acquisition costs vary: cheap or expensive features
- Exploit structure during classification
- Minimize costs

E-mail Challenges: Spam Detection



- Most mail is spam
- Billions of classifications
- Must be incredibly fast

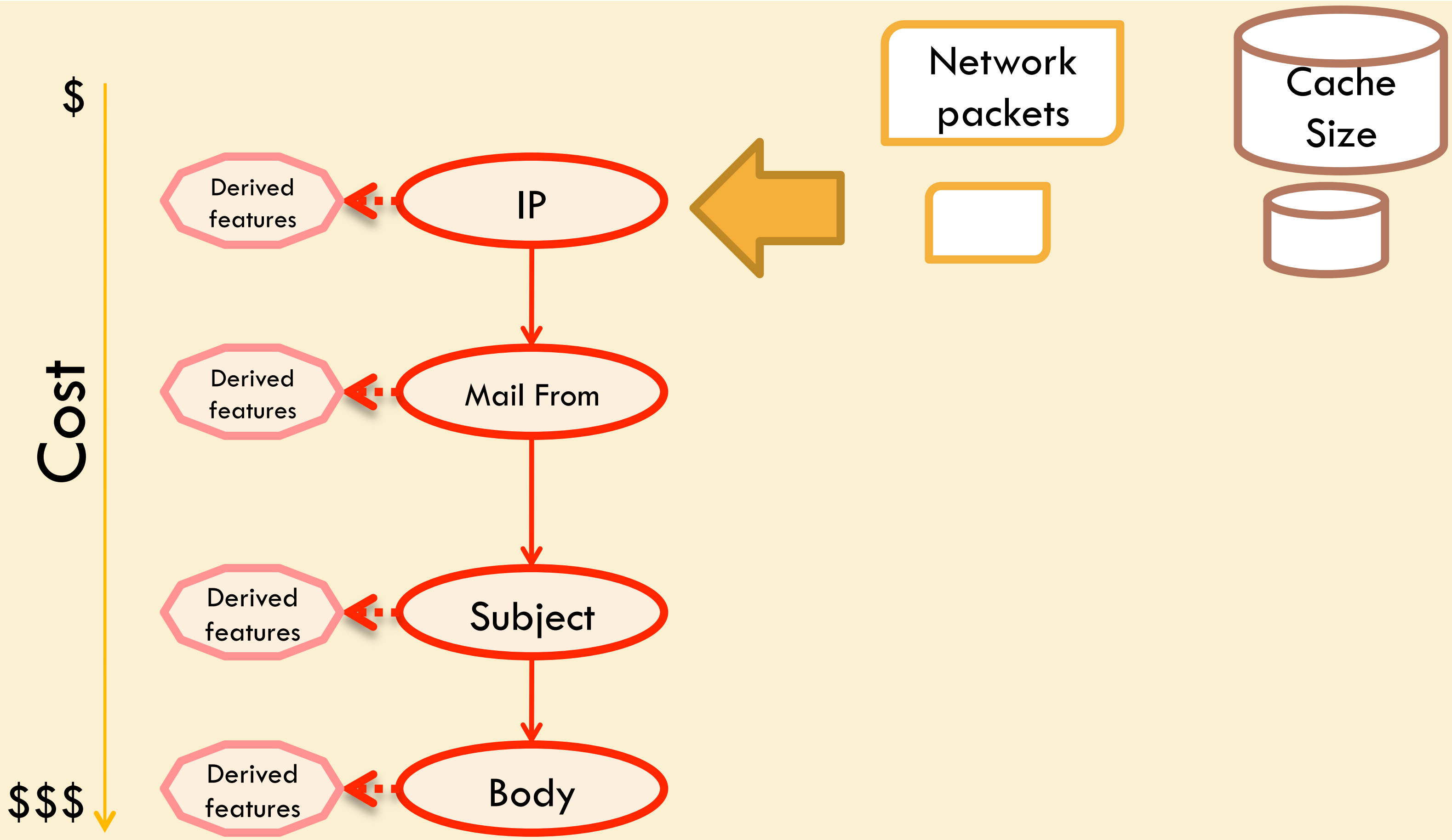
E-mail Challenges: Categorizing Mail



- E-mail does more, tasks such as:
  - Extract receipts, tracking info
  - Thread conversations
  - Filter into mailing lists
  - Inline social network response

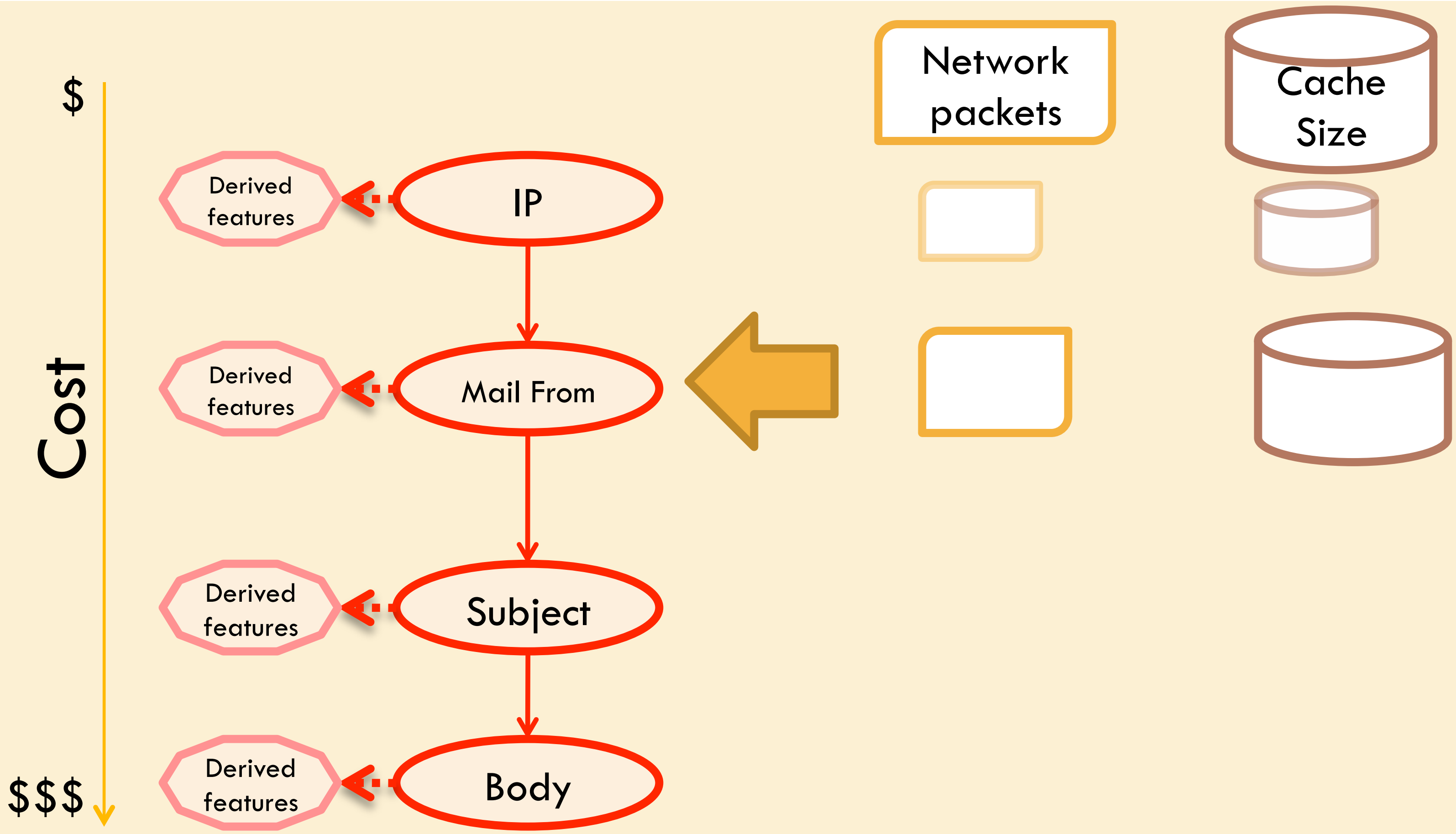
- Computationally intensive processing
- Each task applies to one class

Features have costs & dependencies



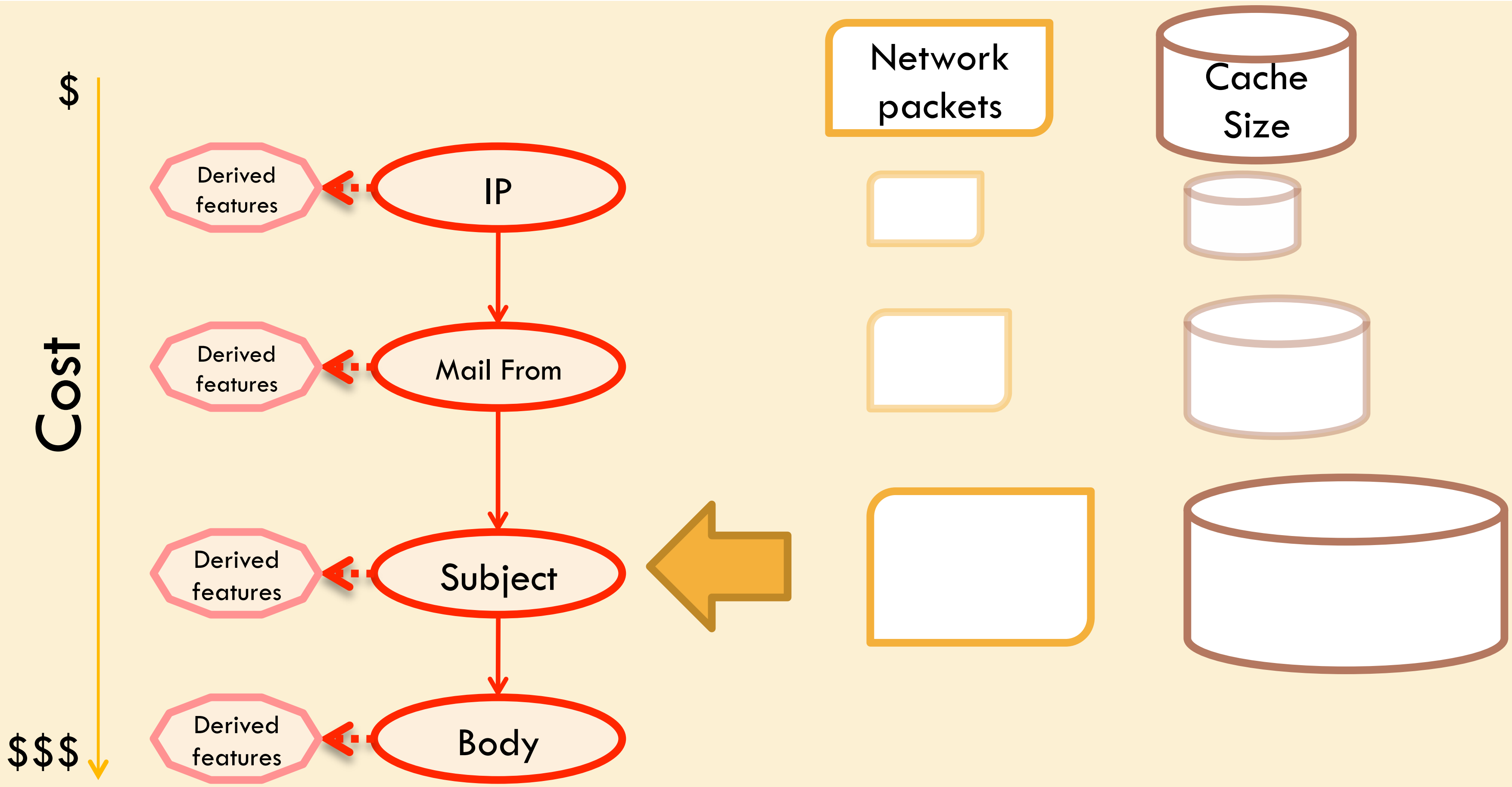
IP is known at socket connect time, is 4 bytes in size

Features have costs & dependencies



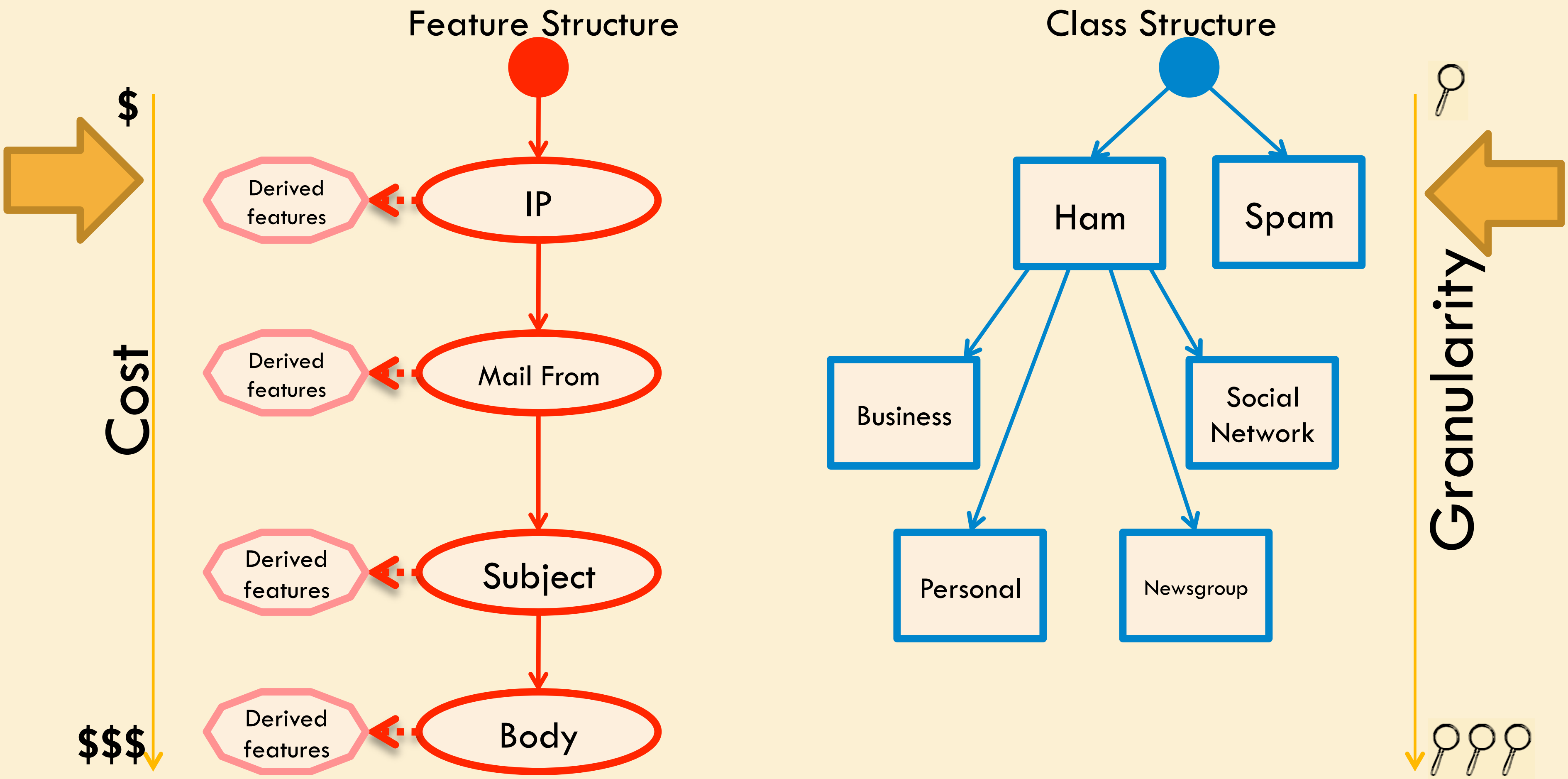
The Mail From is one of the first commands of an SMTP conversation
From addresses have a known format, but higher diversity

Features have costs & dependencies

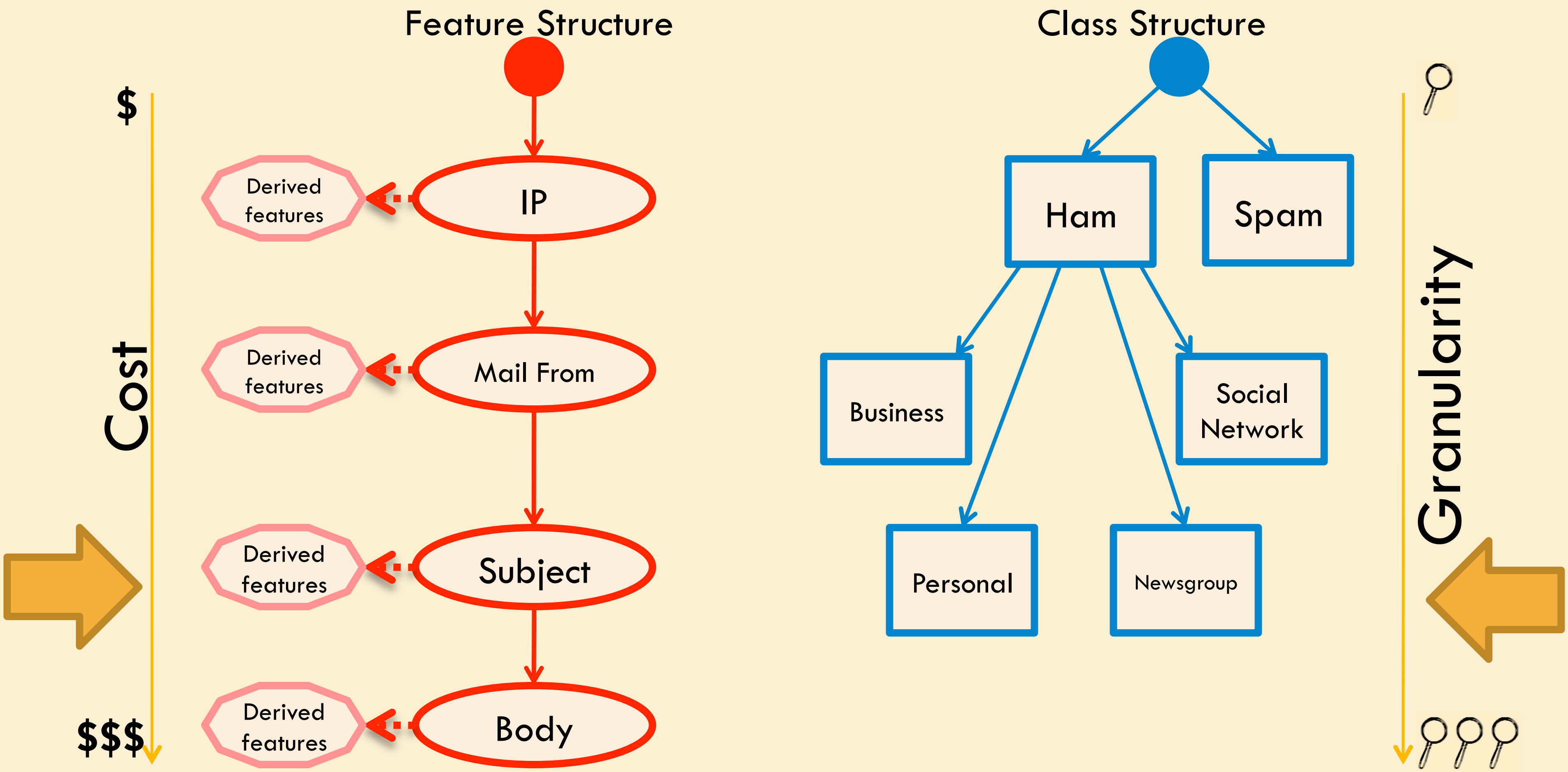


The subject, one of the mail headers, occurs after a number of network exchanges. Since the subject is user-generated, it is very diverse and often lacks a defined format

Coarse task is constrained by feature cost



Fine task is constrained by misclassification cost



Approach: Granular Cost Sensitive Classifier

Training:

- Loss functions of form: $L=\alpha$ FC + $(1-\alpha)$ MC
- Choose α_c and α_f for coarse and fine tasks
- Calculate margin threshold where feature acquisition decreases loss across training data

Test:

- Compute decision margin with available features
- Acquire features until margin above threshold
- Classify instance

Experimental Setup

Class	Messages
Spam	531
Business	187
Social Network	223
Newsletter	174
Personal/Other	102

Feature	Cost
IP	.168
MailFrom	.322
Subject	.510

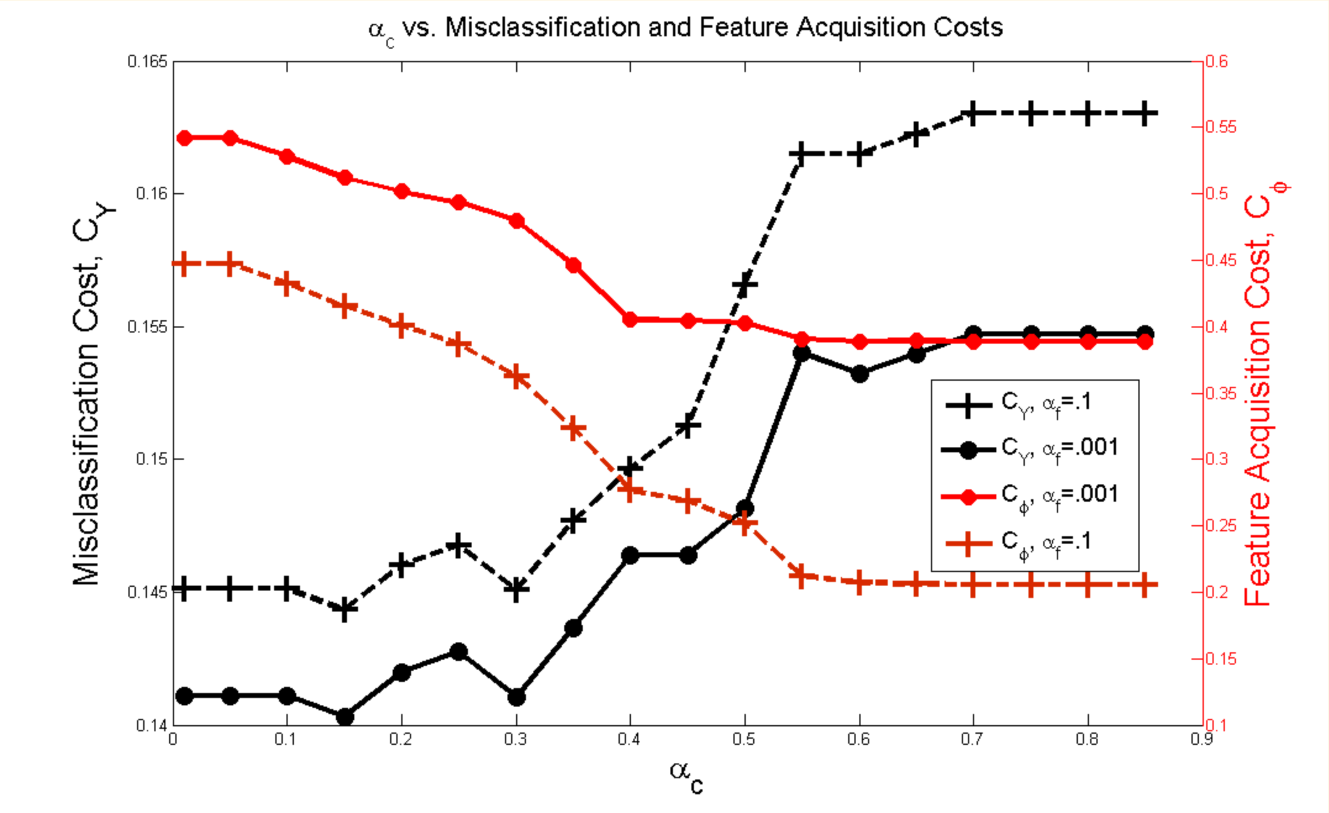
- Data from 1227 Yahoo! Mail messages from 8/2010
- Feature costs calculated from network + storage cost

Results

Feature Set	Feature Cost	Misclass Cost		
		Coarse	Fine	Overall
Fixed: IP+MailFrom	.490	.098	.214	.164
GCSC: α_c=.3, α_f=.05	.479	.091	.174	.141
Fixed: IP+MailFrom+Subject	1.00	.090	.176	.144
GCSC: α_c=.15, α_f=.01	.511	.088	.175	.140

- Evaluated NB & SVM base classifiers, NB results shown
- Compare fixed features vs. GCSC with 10-fold L1O CV
- Same feature cost, decrease misclassification cost
- Decrease feature cost, same misclassification cost

Dynamics of choosing α_c and α_f



As α_c increases, disparity in costs for different values of α_f widens

Conclusion

- Examine a problem setting with coarse-to-fine structure in both input and output
- Propose a classifier, mapping input to output
  - at different granularities
  - sensitive to feature and misclassification costs
- Demonstrate results superior to baseline
- Details at http://bit.ly/jay_c2f_2010

Questions?

Research funded by Yahoo! Faculty Research Engagement Program