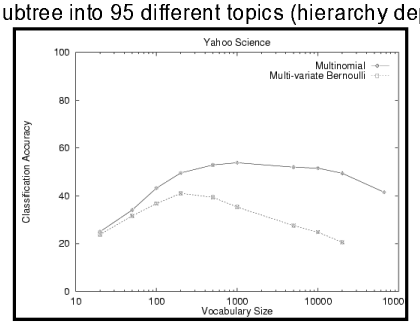
Conditional Independence

- Conditional independence assumption is typically false
  - Sinus condition not independent of runny nose, even given flu
- Nevertheless, it works surprisingly well
  - Reason 1: small number of parameters
    - if we try to fit too many parameters with sparse data, can get really strange models
  - Reason 2: Don't need probabilities to be correct, only argmax

$$\arg\max_c \hat{P}(c)\prod_i \hat{P}(x_i \mid c) = \arg\max_c P(c)\prod_i P(x_i \mid c)$$

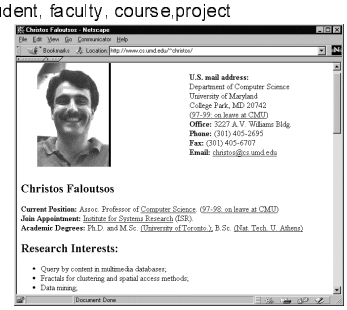
Example: AutoYahoo!

- Classify 13589 Yahoo! webpages in "Science" subtree into 95 different topics (hierarchy depth 2)



Example: WebKB (CMU)

- Classify webpages from CS departments into:
  - student, faculty, course,project

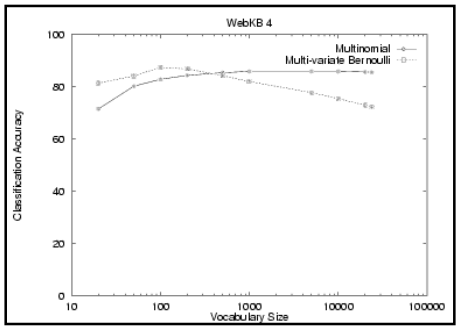


WebKB Experiment

- Train on ~5,000 hand-labeled web pages
  - Cornell, Washington, U.Texas, Wisconsin
- Crawl and classify a new site (CMU)
- Results:

	Student	Faculty	Person	Project	Course	Departmt
Extracted	180	66	246	99	28	1
Correct	130	28	194	72	25	1
Accuracy:	72%	42%	79%	73%	89%	100%

NB Model Comparison



Faculty	
associate	0.00417
chair	0.00303
member	0.00288
ph	0.00287
director	0.00282
fax	0.00279
journal	0.00271
recent	0.00260
received	0.00258
award	0.00250

Students	
resume	0.00516
advisor	0.00456
student	0.00387
working	0.00361
stuff	0.00359
links	0.00355
homepage	0.00345
interests	0.00332
personal	0.00332
favorite	0.00310

Courses	
homework	0.00413
syllabus	0.00399
assignments	0.00388
exam	0.00385
grading	0.00381
midterm	0.00374
pm	0.00371
instructor	0.00370
due	0.00364
final	0.00355

Departments	
departmental	0.01246
colloquia	0.01076
epartment	0.01045
seminars	0.00997
schedules	0.00879
webmaster	0.00879
events	0.00826
facilities	0.00807
eople	0.00772
postgraduate	0.00764

Research Projects	
investigators	0.00256
group	0.00250
members	0.00242
researchers	0.00241
laboratory	0.00238
develop	0.00201
related	0.00200
arpa	0.00187
affiliated	0.00184
project	0.00183

Others	
type	0.00164
jan	0.00148
enter	0.00145
random	0.00142
program	0.00136
net	0.00128
time	0.00128
format	0.00124
access	0.00117
begin	0.00116