Machine Learning, Predictive Text, and Topic Models

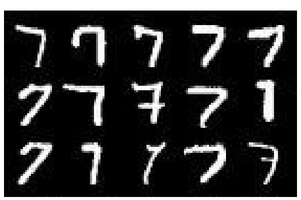
Hanna Wallach

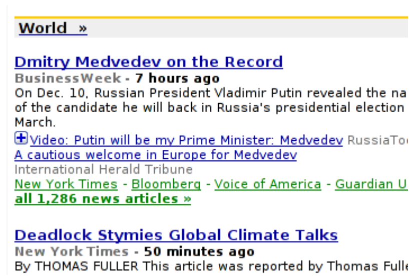
University of Massachusetts Amherst

wallach@cs.umass.edu

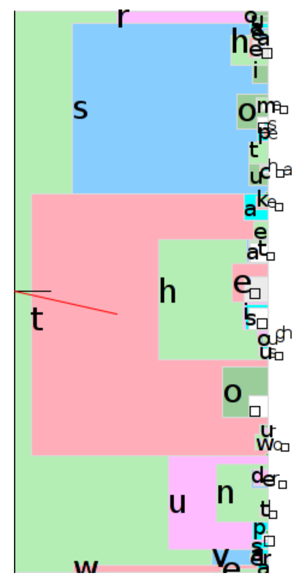
Outline

- What is machine learning?
- Examples of machine learning in practice:





$30: Dinner, Cambridge MA
$50: Bus ticket, Cambridge MA
$5000: Hotel suite, Hong Kong
$20: Beer, Amherst MA
$10: Lunch, Amherst MA



BALL	JOB
GAME	WORK
TEAM	JOBS
FOOTBALL	CAREER
BASEBALL	EXPERIENCE
PLAYERS	EMPLOYMENT
PLAY	OPPORTUNITIES
FIELD	WORKING
PLAYER	TRAINING
BASKETBALL	SKILLS
COACH	CAREERS
PLAYED	POSITIONS
PLAYING	FIND
HIT	POSITION
TENNIS	**FIELD**

Machine Learning (ML)

- There are increasingly large amounts of digital data available:

- Machine learning uses computers to find the most salient features in data to further knowledge and make life easier

  ... with as little human input as possible

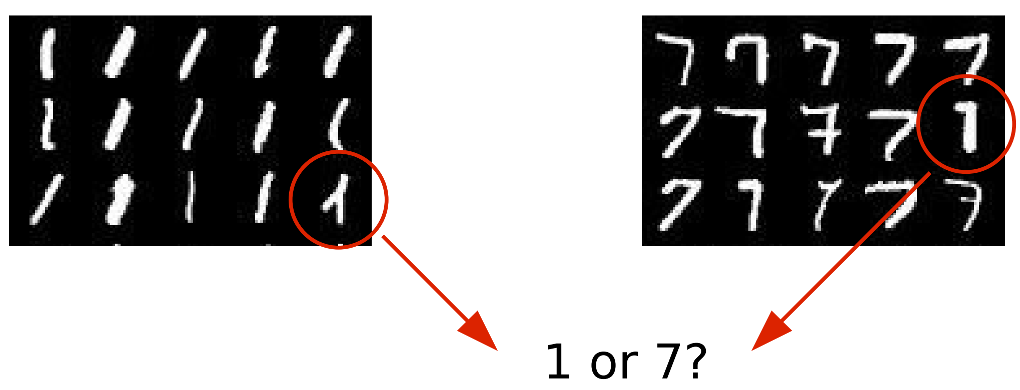
Uncertainty

- There is uncertainty in almost all real world situations:

- ML explicitly represents uncertainty using probability:
  - Pr (lemon) = how certain I am that this is a lemon
- Probability provides a framework for reasoning under uncertainty

USPS Digit Recognition

- Problem:
  - USPS needs to sort letters by zip code
- Solution:
  - Teach a computer to recognize hand-written digits
  - Only ask human when computer is uncertain:

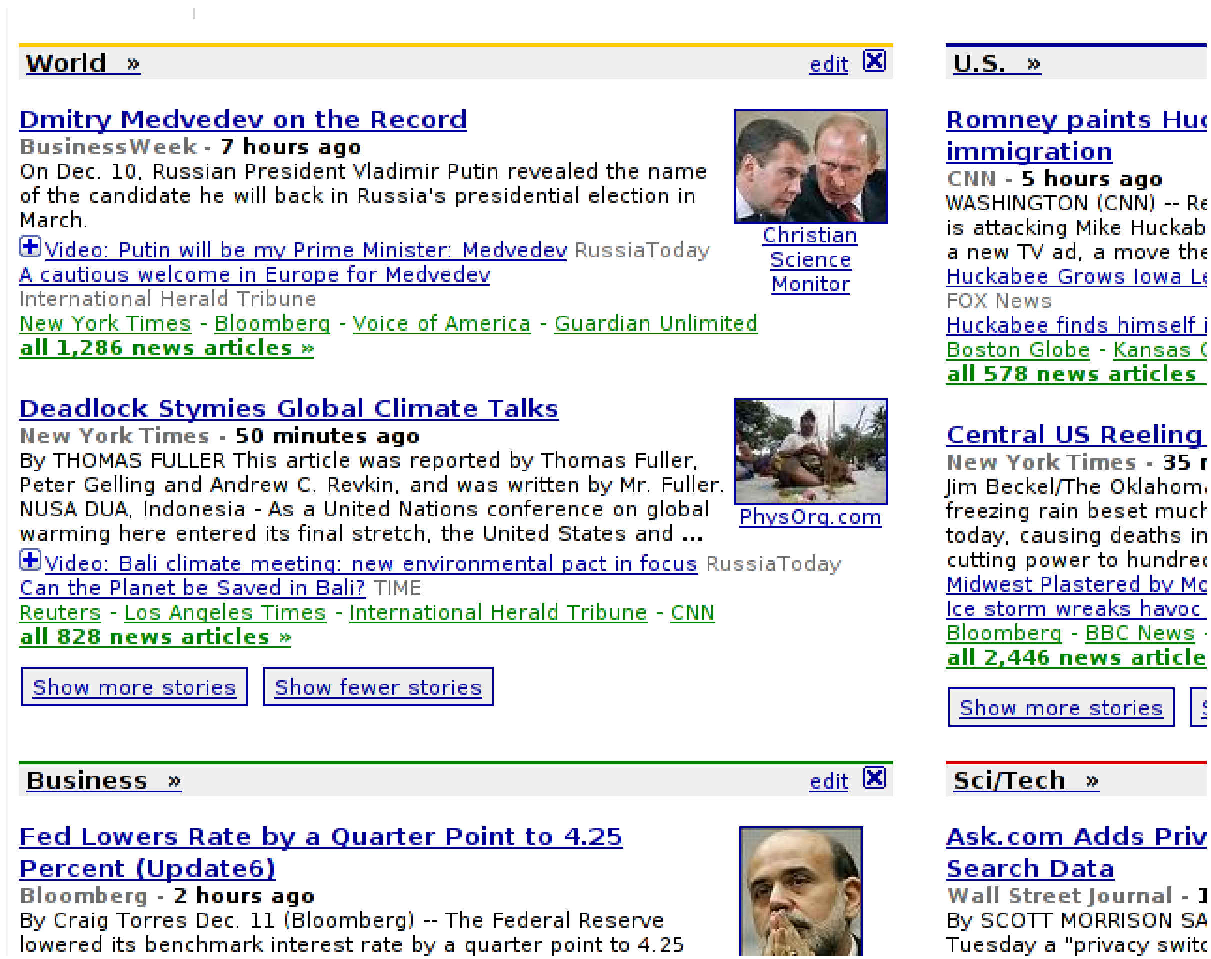


Credit Card Fraud

- Problem:
  - Want to detect credit card fraud
- Solution:
  - Train a computer to recognise normal and abnormal usages
  - Alert card-holder if abnormal pattern is detected

$30: Dinner, Cambridge MA	$30: Dinner, Cambridge MA
$50: Bus ticket, Cambridge MA	$50: Bus ticket, Cambridge MA
$10: Lunch, Amherst MA	$5000: Hotel suite, Hong Kong
$20: Beer, Amherst MA	$20: Beer, Amherst MA
$10: Lunch, Amherst MA	$10: Lunch, Amherst MA

Sorting News Stories by Genre



Predictive Text Entry

- e.g., T9 or iTAP
- Used on cell phones
- Enables use of reduced keyboard
- Enter as much text as possible with as few gestures as possible

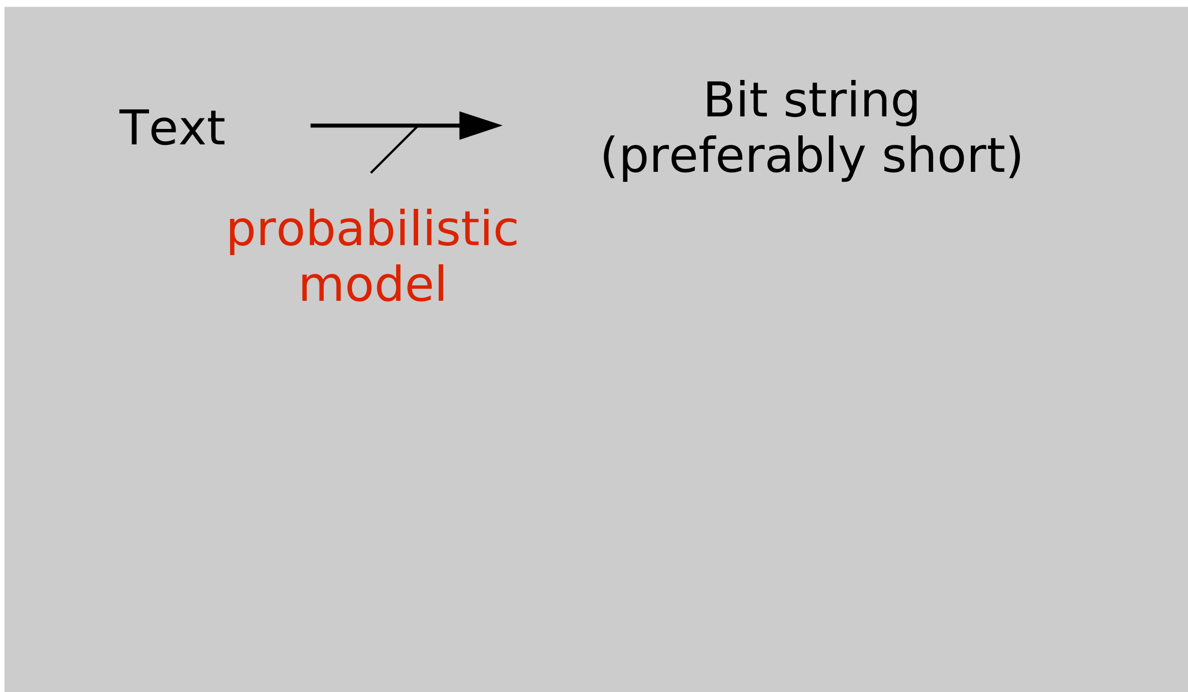
Text ← Gestures (as few as possible)

Predictive Text Entry

- This is like the reverse of text compression
- Text compression: want to go from as much text as possible to as small a representation as possible

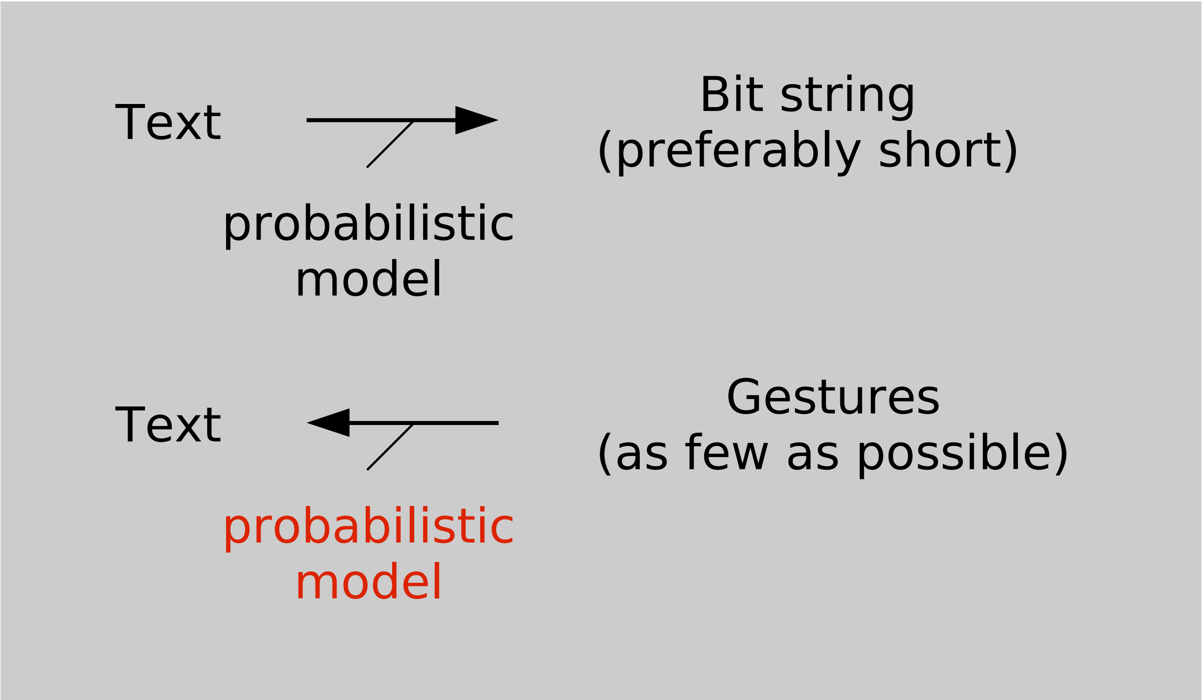
Writing and Text Compression

- Optimal text compression



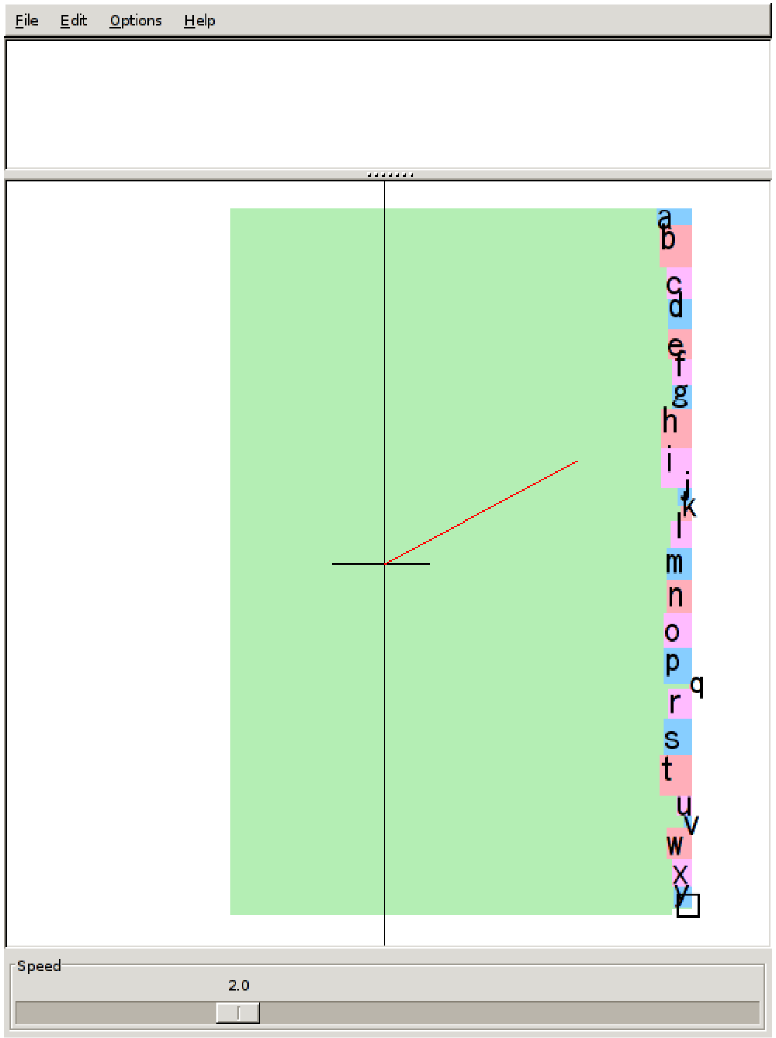
Writing and Text Compression

- Optimal text compression and writing with predictive text entry



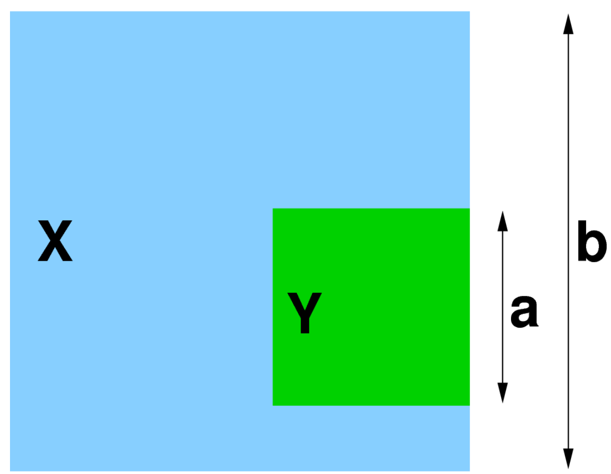
Dasher [http://www.dasher.org.uk]

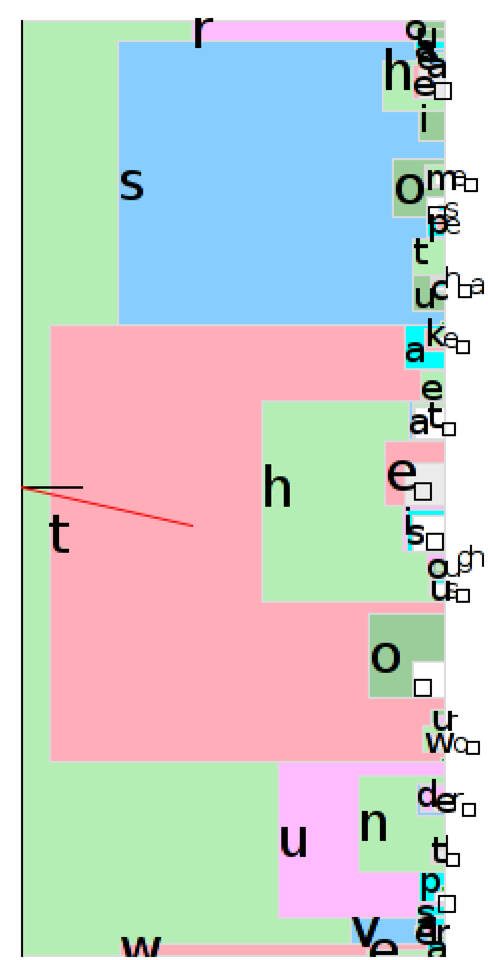
- Driven by 2D continuous gestures
- Uses a model of language
- Available for
  - Windows
  - Linux
  - Mac OS X
  - Pocket PC
  - etc.



Dasher: Screen Layout

- Box sizes are proportional to probabilities
- Probabilities come from a letter-based language model
- P(X) = b
  P(X, Y) = a

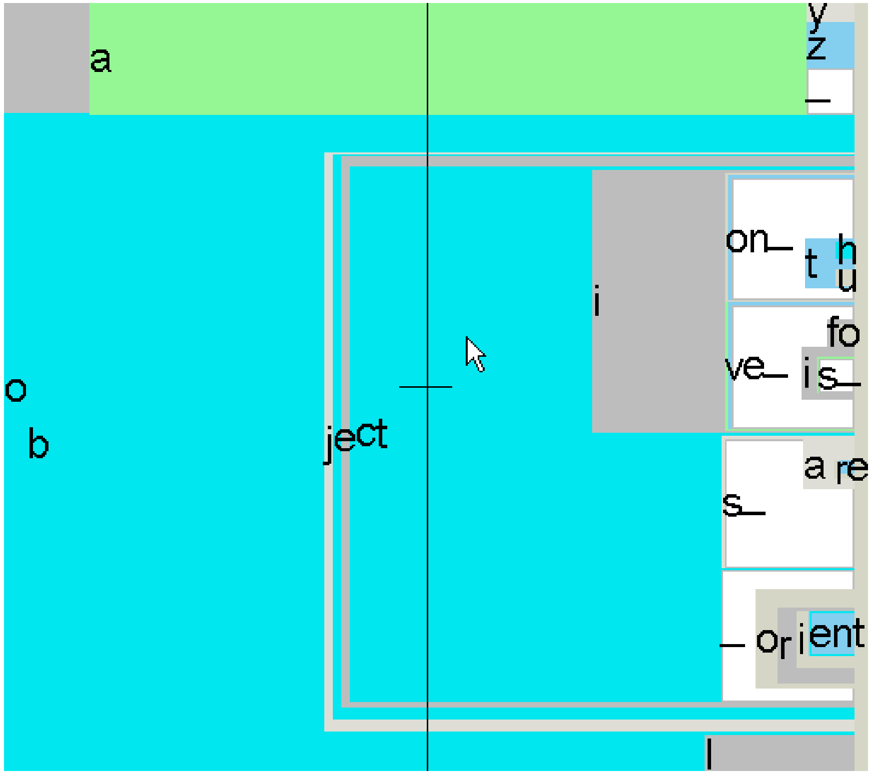




Dasher: Dynamics

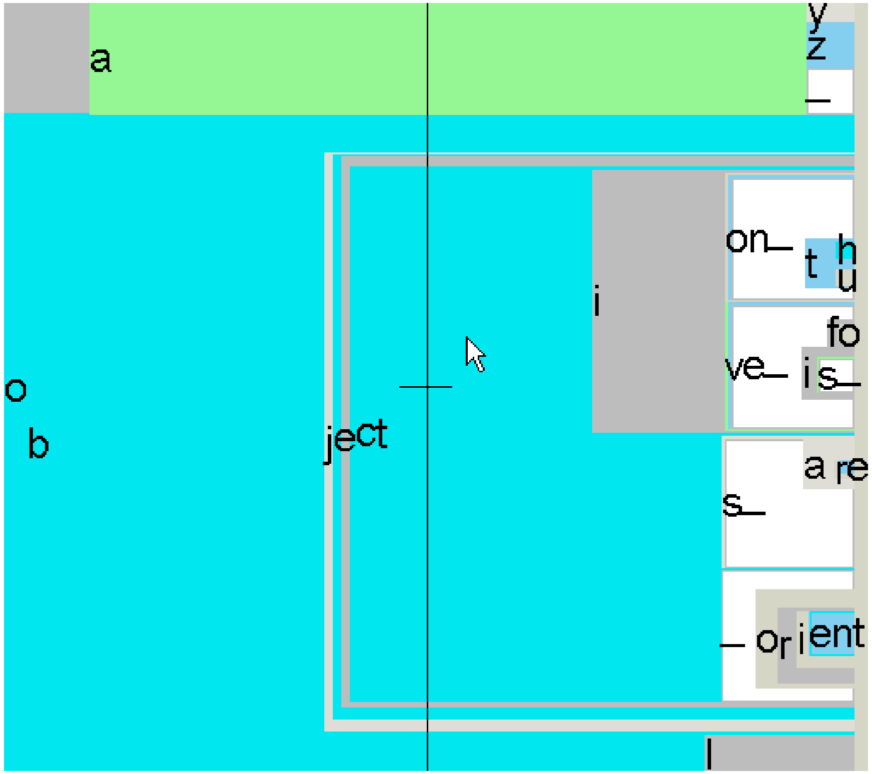
Point where you want to go

- Like driving a car
- Motion sickness?
- Not if you're driving!



Dasher: Benefits

- Keyboards: one gesture per character
- Dasher: some gestures select many characters
- Works with any language
- Inaccurate gestures can be compensated for by later gestures



Topic Models

- Humans can read a document and identify the small number of topics that best characterize that document

The Beverly Hills love nest that Jennifer Aniston and Brad Pitt called home during their marriage is on the block – for $28 million – the Los Angeles Times reported on Sunday, which happened to be the same day that the 4 1/2-year union of the actors was officially dissolved.

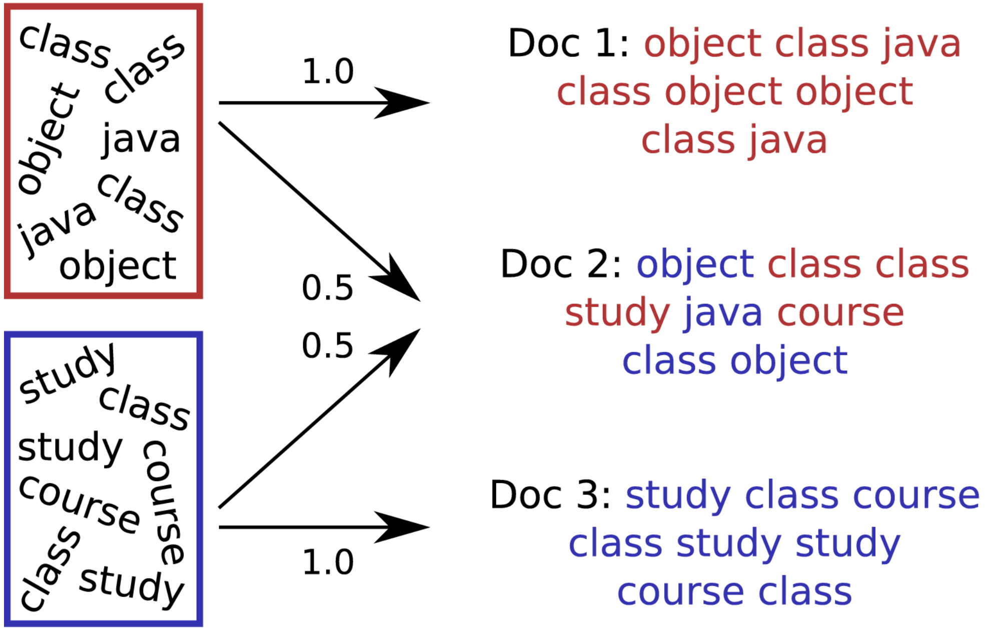
More on this story

- **Pop Quiz: Do You Know Jennifer?**

The more than 10,000-sq.ft French Normandy house, originally designed by noted architect Wallace Neff for *A Star Is Born* actor Fredric March in the 1930s, was purchased by the Pitts in 2001 for about $13.5 million. They then spent two years refurbishing it and are now selling it as part of their divorce settlement.

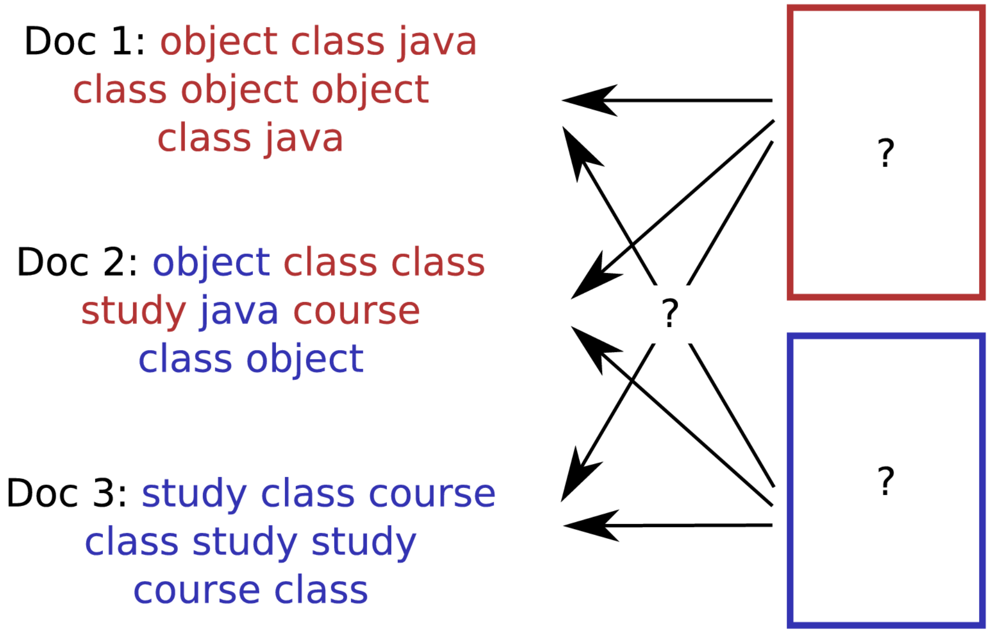
Topic Models

- Topics are mixtures of words and documents are mixtures of topics



Topic Models

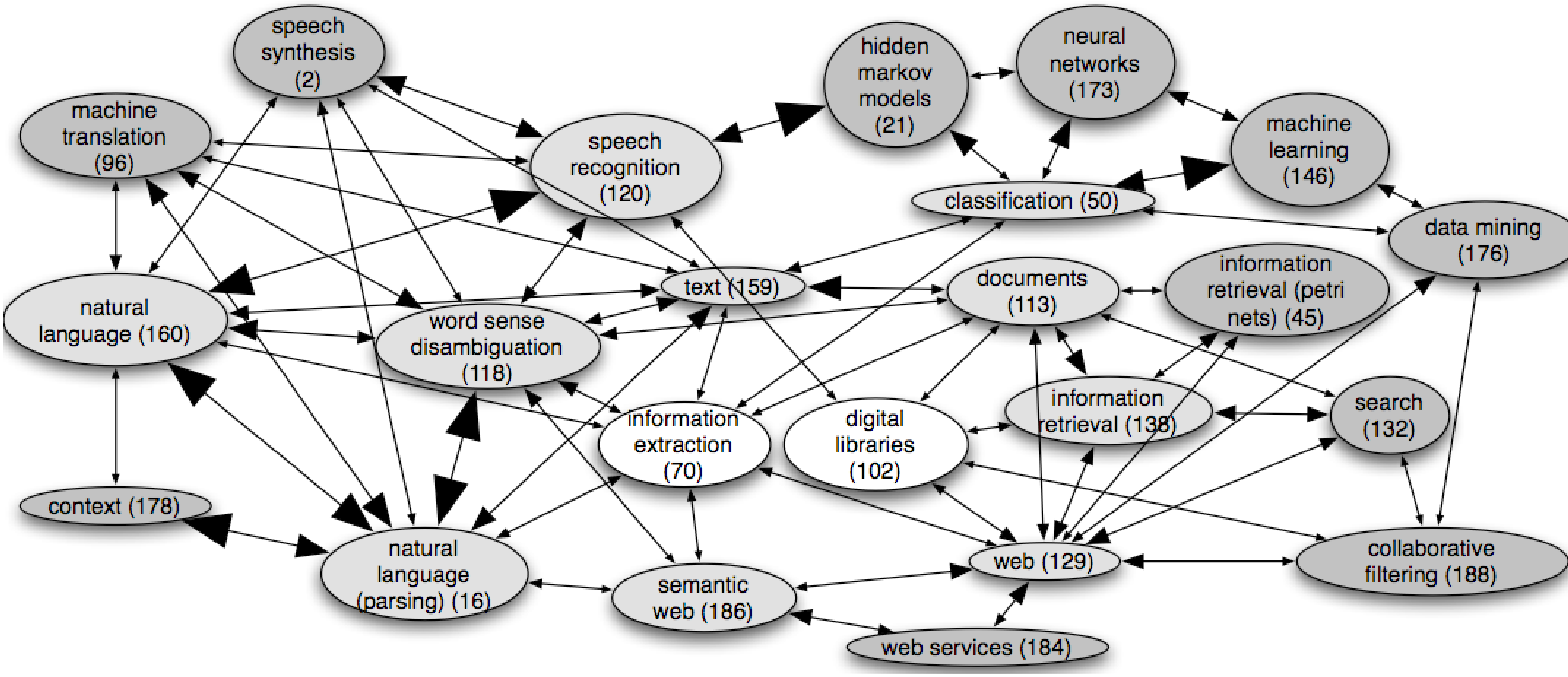
- Infer topic information from word-document co-occurrences



Example Topics [Tenenbaum et al.]

STORY	**FIELD**	SCIENCE	BALL	JOB
STORIES	MAGNETIC	STUDY	GAME	WORK
TELL	MAGNET	SCIENTISTS	TEAM	JOBS
CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
TOLD	POLES	CHEMISTRY	**FIELD**	WORKING
SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
TALES	COMPASS	MANY	BASKETBALL	SKILLS
PLOT	LINES	MATHEMATICS	COACH	CAREERS
TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
SHORT	ELECTRIC	**FIELD**	PLAYING	FIND
FICTION	DIRECTION	PHYSICS	HIT	POSITION
ACTION	FORCE	LABORATORY	TENNIS	**FIELD**

Transfer between Topics [Mimno]



Entities and Topics [Newman et al.]

Sept. 11		Fear		US Pride		Defense		Agencies	
attack	0.017	fear	0.023	american	0.062	defense	0.039	agencies	0.029
victim	0.016	public	0.019	flag	0.046	missile	0.039	department	0.019
tragedy	0.015	threat	0.011	country	0.035	system	0.032	staff	0.017
missing	0.013	concern	0.010	war	0.028	administration	0.019	mission	0.017
lost	0.012	anger	0.008	nation	0.022	arms	0.019	agency	0.017
families	0.012	crisis	0.008	history	0.012	weapon	0.019	policy	0.016
lives	0.010	support	0.007	feel	0.010	nuclear	0.015	problem	0.012
memorial	0.010	sense	0.007	symbol	0.009	test	0.014	resources	0.011
happened	0.009	seen	0.007			missiles	0.013	program	0.009
dead	0.009	changed	0.006			treaty	0.012	security	0.009
E130	0.980	**E55**	0.720	**E55**	1.000	**E6**	0.900	**E145**	0.780
		E130	0.110			**E145**	0.100	**E161**	0.220
		E161	0.060						

E130: Sept. 11		E161: US Admin		E55: US/War		E6: Foreign		E145: US Security	
NY	0.188	BUSH	0.290	AMERICA	0.164	RUSSIA	0.113	US	0.196
WTC	0.091	CLINTON	0.133	US	0.102	PENTAGON	0.073	STATE DEPT	0.052
AMERICA	0.071	WHITE HSE	0.094	WASH. DC	0.064	CHINA	0.057	GOVT.	0.041
GOD	0.036	WASH DC	0.075	BUSH	0.037	CLINTON	0.055	NSC	0.027
WASH. DC	0.035	CONGRESS	0.062	WW2	0.024	BUSH	0.052	CONGRESS	0.024
NYC	0.027	POWELL	0.032	CIVIL WAR	0.021	PUTIN	0.046	CIA	0.022
GIULIANI	0.023	UN	0.014	WEST	0.012	N. KOREA	0.033	PENTAGON	0.018
		PRESIDENT	0.014	RIGHT	0.012	IRAQ	0.029		

Topics and Email

- Enron email corpus:
  - 250k email messages, 23k people

Sally -
Attached are the hypertiles from the final report out at yesterday’s ASE Studio Workshop. The CD is finished and on its way to Houston. The files are organized by team:
Hammer - Sales and Marketing, Vision Stmt, Mission Stmt, Target Market, How to Approach, Pricing, SLA
Pliers - Producst and Services - Consulting Based
Saw - Infrastructure Transition Plan
Wrench - Producst and Services - Basic Outsourcing
I hope these help with your meeting tomorrow. Let me know if there is anything else I can do to help.
Lisa P

Selecting Email Keywords [Dredze et al.]

Sally -
Attached are the hypertiles from the final report out at yesterday's ASE Studio Workshop. The CD is finished and on its way to Houston. The files are organized by team:
Hammer - Sales and Marketing, Vision Stmt, Mission Stmt, Target Market, How to Approach, Pricing, SLA
Pliers - Producst and Services - Consulting Based
Saw - Infrastructure Transition Plan
Wrench - Producst and Services - Basic Outsourcing
I hope these help with your meeting tomorrow. Let me know if there is anything else I can do to help.
Lisa P

- Without topics: producst pliers stmt hammer wrench
- With topics: team meeting services lisa ase

Senders, Recipients, Topics [McCallum et al.]

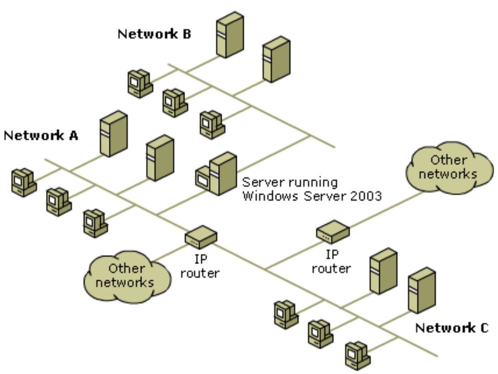
Topic 34 "Operations"		Topic 37 "Power Market"		Topic 41 "Government Relations"		Topic 42 "Wireless"	
operations	0.0321	market	0.0567	state	0.0404	blackberry	0.0726
team	0.0234	power	0.0563	california	0.0367	net	0.0557
office	0.0173	price	0.0280	power	0.0337	www	0.0409
list	0.0144	system	0.0206	energy	0.0239	website	0.0375
bob	0.0129	prices	0.0182	electricity	0.0203	report	0.0373
open	0.0126	high	0.0124	davis	0.0183	wireless	0.0364
meeting	0.0107	based	0.0120	utilities	0.0158	handheld	0.0362
gas	0.0107	buy	0.0117	commission	0.0136	stan	0.0282
business	0.0106	customers	0.0110	governor	0.0132	fyi	0.0271
houston	0.0099	costs	0.0106	prices	0.0089	named	0.0260
S.Beck	0.2158	J.Dasovich	0.1231	J.Dasovich	0.3338	R.Haylett	0.1432
L.Kitchen		J.Steffes		R.Shapiro		T.Geaccone	
S.Beck	0.0826	J.Dasovich	0.1133	J.Dasovich	0.2440	T.Geaccone	0.0737
J.Lavorato		R.Shapiro		J.Steffes		R.Haylett	
S.Beck	0.0530	M.Taylor	0.0218	J.Dasovich	0.1394	R.Haylett	0.0420
S.White		E.Sager		R.Sanders		D.Fossum	

"Chief Operations Officer"

"Government Relations Executive"

Summary

- Machines can learn a lot from unstructured digital data
- We can use machine learning to build useful applications, some of which you are already using!



Questions?