10-709: Read The Web

Tom Mitchell

January 2006
Learn to Read / Read to Learn

Thesis: We can achieve a breakthrough in NLP by building a continuously learning, continuously reading system, targeted toward understanding and extracting 80% of the factual content on the internet.

Why now?
1. Recent progress in NLP
2. Recent progress in statistical machine learning
   • Especially bootstrapping methods that leverage redundancy
3. The web provides huge corpus of highly redundant text
The Idea

• Build on existing components
  – Named entity extractors, question answerers, parsers, coreference resolvers, ...
  – Self-supervised learning algorithms
  – Knowledge representations, ontologies, KBs, ...

• Create agent that formulates and pursues an infinite stream of learning/reading/fact acquisition subgoals

• Learn to read / Read to learn

• Primarily unsupervised (self-supervised)
Design goals for ReadTheWeb system

• Nonstop 24x7 operation, pursuing two goals:
  – Learning to read
  – Reading the web

• Begin with state-of-the-art methods (NLP, learning, representation)

• Architecture for improving continuously
  – A growing knowledge base (with pointers back to text sources)
  – A growing ability to understand complex text (and non-text)

• <1 day barrier to entry for researchers
Design of the course

• Become experts in state of the art of semi-supervised learning for NLP

• Design, implement, experiment with, and write up a first ReadTheWeb system

• First 4 weeks: each team implements working semi-supervised learner, for some aspect of NLP
• Next 8 weeks: we design and implement integrated system
• All 13 weeks: cover state-of-art research papers
What we’ll build on

• State of the art semi-supervised learning and NLP algorithms

• Existing software
  – Knowledge repository (SCONE)
  – Text learning package (Minor Third)
  – Text annotation framework (UIMA)
  – Web crawl / web query engine

• Your expertise, creativity and hard work
Course Logistics/Details

- This is a research project disguised as a course
- This will be hard work, and fun
- Some guest lectures (e.g., Oren Etzioni, Feb 9)
- No exams
- Grading based on projects and course participation
- Course web site will appear by tomorrow, off http://www.cs.cmu.edu/~tom
Redundantly Sufficient Features

Christos Faloutsos

U.S. mail address:
Department of Computer Science
University of Maryland
College Park, MD 20742
(97-99: on leave at CMU)
Office: 3227 A.V. Williams Bldg
Phone: (301) 405-2695
Fax: (301) 405-6707
Email: christos@cs.umd.edu

Current Position: Assoc. Professor of Computer Science, (97-98: on leave at CMU)
Join Appointment: Institute for Systems Research (ISR).
Academic Degrees: Ph.D. and M.Sc. (University of Toronto); B.Sc. (Nat. Tech. U. Athens)

Research Interests:
- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining;
Redundantly Sufficient Features

Professor Faloutsos

my advisor
Redundantly Sufficient Features

Christos Faloutsos

Current Position: Assoc. Professor of Computer Science. (97-98: on leave at CMU)
Join Appointment: Institute for Systems Research (ISR).
Academic Degrees: Ph.D and M.Sc. (University of Toronto); B.Sc. (Nat. Tech. U. Athens)

Research Interests:

- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining;

U.S. mail address:  
Department of Computer Science  
University of Maryland  
College Park, MD 20742  
(97-99: on leave at CMU)  
Office: 3227 A.V. Williams Bldg  
Phone: (301) 405-2695  
Fax: (301) 405-6707  
Email: christos@cs.umd.edu
Redundantly Sufficient Features

Professor Faloutsos

Christos Faloutsos

**Current Position:** Assoc. Professor of Computer Science, (97-98: on leave at CMU)
**Join Appointment:** Institute for Systems Research (ISR).
**Academic Degrees:** Ph.D. and M.Sc. (University of Toronto); B.Sc. (Nat. Tech. U. Athens)

**Research Interests:**

- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining;

---

**U.S. mail address:**
Department of Computer Science
University of Maryland
College Park, MD 20742
*(97-99: on leave at CMU)*
**Office:** 3227 A.V. Williams Bldg
**Phone:** (301) 405-2695
**Fax:** (301) 405-6707
**Email:** christos@cs.umd.edu
CoTraining Algorithm #1
[Blum&Mitchell, 1998]

Given: labeled data L, unlabeled data U

Loop:
- Train $g_1$ (hyperlink classifier) using L
- Train $g_2$ (page classifier) using L
- Allow $g_1$ to label $p$ positive, $n$ negative examps from U
- Allow $g_2$ to label $p$ positive, $n$ negative examps from U
- Add these self-labeled examples to L
CoTraining: Experimental Results

- begin with 12 labeled web pages (academic course)
- provide 1,000 additional unlabeled web pages
- average error: learning from labeled data 11.1%;
- average error: cotraining 5.0%

Typical run:
CoTraining setting:

• wish to learn \( f: X \rightarrow Y \), given \( L \) and \( U \) drawn from \( P(X,Y) \)

• features describing \( X \) can be partitioned \( (X = X_1 \times X_2) \)

  such that \( f \) can be computed from either \( X_1 \) or \( X_2 \)

\[
(\exists g_1, g_2) (\forall x \in X) \quad g_1(x_1) = f(x) = g_2(x_2)
\]

One result [Blum&Mitchell 1998]:

• If

  – \( X_1 \) and \( X_2 \) are conditionally independent given \( Y \)
  
  – \( f \) is PAC learnable from noisy labeled data

• Then

  – \( f \) is PAC learnable from weak initial classifier plus unlabeled data
Co-Training Rote Learner
Co-Training Rote Learner

My advisor

hyperlinks

pages
Co-Training Rote Learner
Co-Training Rote Learner

My advisor

hyperlinks

pages
Expected Rote CoTraining error given $m$ labeled examples, rote learning, perfectly redundantly sufficient

**CoTraining setting:**

- **learn** $f : X \rightarrow Y$
- **where** $X = X_1 \times X_2$
- **where** $x$ drawn from unknown distribution
- **and** $\exists g_1, g_2 \ (\forall x) g_1(x_1) = g_2(x_2) = f(x)$

$$E[\text{error}] = \sum_j P(x \in g_j)(1 - P(x \in g_j))^m$$

Where $g_j$ is the $j$th connected component of graph of $L+U$, $m$ is number of labeled examples
How many *unlabeled* examples suffice?

Want to assure that connected components in the underlying
distribution, $G_D$, are connected components in the observed
sample, $G_S$

$$O(\log(N)/\alpha)$$ examples assure that with high probability, $G_S$ has same
connected components as $G_D$ [Karger, 94]

$N$ is size of $G_D$, $\alpha$ is min cut over all connected components of $G_D$
Co Training

- What’s the best-case graph? (most benefit from unlabeled data)
- What the worst case?
- What does conditional-independence imply about graph?
PAC Generalization Bounds on CoTraining

[Dasgupta et al., NIPS 2001]

This theorem assumes $X_1$ and $X_2$ are conditionally independent given $Y$.

**Theorem 1** With probability at least $1 - \delta$ over the choice of the sample $S$, we have that for all $h_1$ and $h_2$, if $\gamma_i(h_1, h_2, \delta) > 0$ for $1 \leq i \leq k$ then (a) $f$ is a permutation and (b) for all $1 \leq i \leq k$,

$$P(h_1 \neq i \mid f(y) = i, h_1 \neq \bot) \leq \frac{\hat{P}(h_1 \neq i \mid h_2 = i, h_1 \neq \bot) + \epsilon_i(h_1, h_2, \delta)}{\gamma_i(h_1, h_2, \delta)}.$$

The theorem states, in essence, that if the sample size is large, and $h_1$ and $h_2$ largely agree on the unlabeled data, then $\hat{P}(h_1 \neq i \mid h_2 = i, h_1 \neq \bot)$ is a good estimate of the error rate $P(h_1 \neq i \mid f(y) = i, h_1 \neq \bot)$. 
What if CoTraining Assumption Not Perfectly Satisfied?

- Idea: Want classifiers that produce a *maximally consistent* labeling of the data
- If learning is an optimization problem, what function should we optimize?
What Objective Function?

\[ E = E_1 + E_2 + c_3 E_3 + c_4 E_4 \]

\[ E_1 = \sum_{<x,y> \in L} (y - \hat{g}_1(x_1))^2 \]

Error on labeled examples

\[ E_2 = \sum_{<x,y> \in L} (y - \hat{g}_2(x_2))^2 \]

Disagreement over unlabeled

\[ E_3 = \sum_{x \in U} (\hat{g}_1(x_1) - \hat{g}_2(x_2))^2 \]

Misfit to estimated class priors

\[ E_4 = \left( \frac{1}{|L|} \sum_{<x,y> \in L} y \right) - \left( \frac{1}{|L| + |U|} \sum_{x \in L \cup U} \frac{\hat{g}_1(x_1) + \hat{g}_2(x_2)}{2} \right)^2 \]
What Function Approximators?

\[ \hat{g}_1(x) = \frac{1}{\sum_{j} w_{j,1} x_j} \]
\[ \hat{g}_2(x) = \frac{1}{\sum_{j} w_{j,2} x_j} \]

- Same functional form as logistic regression
- Use gradient descent to simultaneously learn \( g_1 \) and \( g_2 \), directly minimizing \( E = E_1 + E_2 + E_3 + E_4 \)
- No word independence assumption, use both labeled and unlabeled data
## Classifying Jobs for FlipDog

### X1: job title

<table>
<thead>
<tr>
<th>Job Title</th>
<th>Company</th>
<th>Location</th>
<th>Salary</th>
<th>Job Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C++/Java Consultants</td>
<td>Elite Placement Services</td>
<td>Houston, TX</td>
<td>$60K</td>
<td>Functions of this position include the consulting, development and implementation of EAI solutions supporting e-commerce and B2B initiatives for...</td>
</tr>
<tr>
<td>Chief Software Architect</td>
<td>Elite Placement Services</td>
<td>Houston, TX</td>
<td>$150K</td>
<td>Responsible for the end-to-end architecture of all n-tiered web-based applications and complementary products. Provide design direction for the...</td>
</tr>
<tr>
<td>Web Application Developers</td>
<td>MI Systems, Inc.</td>
<td>Houston, TX</td>
<td>Open Hourly</td>
<td>See job synopsis for permanent opportunities. (2) Application Developers with...</td>
</tr>
<tr>
<td>Sales Consulting Engineer</td>
<td>Visual Numerics, Inc.</td>
<td>Houston, TX</td>
<td>Open Hourly</td>
<td>Performs pre-sales technical support and product training to customers and non-customers. Technical support includes providing verbal and written response...</td>
</tr>
<tr>
<td>Peoplesoft Software Analyst (Systems Analyst III)</td>
<td>I.T. Staffing, Inc.</td>
<td>Houston, TX</td>
<td>Open Hourly</td>
<td>Some international travel required. Job Description: CLIENT/SERVER APPLICATION ADMINISTRATION. SETTING UP USERS AND SECURITY FOR DATABASE AND APPLICATION...</td>
</tr>
<tr>
<td>Peoplesoft Software Analyst (Systems Analyst III)</td>
<td>I.T. Staffing, Inc.</td>
<td>Houston, TX</td>
<td>Open Hourly</td>
<td>Some international travel required. Job Description: CLIENT/SERVER APPLICATION ADMINISTRATION. SETTING UP USERS AND SECURITY FOR DATABASE AND APPLICATION...</td>
</tr>
</tbody>
</table>

### X2: job description

<table>
<thead>
<tr>
<th>Description</th>
<th>Date Posted</th>
<th>Location</th>
<th>Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functions of this position include...</td>
<td>November 01, 2000</td>
<td>Houston, TX</td>
<td>Computing/MIS Software Development</td>
</tr>
<tr>
<td>Responsible for the end-to-end architecture...</td>
<td>November 01, 2000</td>
<td>Houston, TX</td>
<td>Computing/MIS Software Development</td>
</tr>
<tr>
<td>See job synopsis for permanent opportunities...</td>
<td>November 01, 2000</td>
<td>Houston, TX</td>
<td>Computing/MIS Software Development</td>
</tr>
<tr>
<td>Performs pre-sales technical support...</td>
<td>November 01, 2000</td>
<td>Houston, TX</td>
<td>Computing/MIS Technical Support/Help Desk</td>
</tr>
<tr>
<td>Some international travel required. Job Description:...</td>
<td>October 27, 2000</td>
<td>Houston, TX</td>
<td>Computing/MIS Software Development</td>
</tr>
<tr>
<td>Some international travel required. Job Description:...</td>
<td>October 27, 2000</td>
<td>Houston, TX</td>
<td>Computing/MIS Software Development</td>
</tr>
</tbody>
</table>
Gradient CoTraining
Classifying FlipDog job descriptions: SysAdmin vs. WebProgrammer

Final Accuracy
Labeled data alone: 86%
CoTraining: 96%
Gradient CoTraining
Classifying Capitalized sequences as Person Names

Eg., “Company president Mary Smith said today…”

<table>
<thead>
<tr>
<th>Using labeled data only</th>
<th>25 labeled</th>
<th>2300 labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5000 unlabeled</td>
<td>5000 unlabeled</td>
</tr>
<tr>
<td>Cotraining</td>
<td>.24</td>
<td>.13</td>
</tr>
<tr>
<td>Cotraining without fitting class priors (E4)</td>
<td>.15 *</td>
<td>.11 *</td>
</tr>
<tr>
<td></td>
<td>.27 *</td>
<td></td>
</tr>
</tbody>
</table>

Error Rates

* Quite sensitive to weights of error terms E3 and E4
Co-EM  [Nigam & Ghani, 2000]

Idea:
• Like co-training, use one set of features to label the other
• Like EM, iterate
  – Assigning probabilistic values to unobserved class labels
  – Updating model parameters (= labels of other feature set)

\[
P(\text{class}|\text{context}_i) = \sum_j P(\text{class}|NP_j)P(NP_j|\text{context}_i)
\]
\[
P(\text{class}|NP_i) = \sum_j P(\text{class}|\text{context}_j)P(\text{context}_j|NP_i)
\]
CoEM applied to Named Entity Recognition

[Rosie Jones, 2005], [Ghani & Nigam, 2000]

Update rules:

\[
P(\text{class}|\text{context}_i) = \sum_j P(\text{class}|NP_j)P(NP_j|\text{context}_i)
\]

\[
P(\text{class}|NP_i) = \sum_j P(\text{class}|\text{context}_j)P(\text{context}_j|NP_i)
\]
CoEM applied to Named Entity Recognition

[Rosie Jones, 2005], [Ghani & Nigam, 2000]

Update rules:

\[
P(class|context_i) = \sum_j P(class|NP_j)P(NP_j|context_i)
\]

\[
P(class|NP_i) = \sum_j P(class|context_j)P(context_j|NP_i)
\]
CoEM applied to Named Entity Recognition
[Rosie Jones, 2005], [Ghani & Nigam, 2000]

Update rules:

\[
P(\text{class}|context_i) = \sum_j P(\text{class}|NP_j)P(NP_j|context_i)
\]

\[
P(\text{class}|NP_i) = \sum_j P(\text{class}|context_j)P(context_j|NP_i)
\]
Bootstrapping Results

locations

Precision

Recall
Some nodes are more important than others  [Jones, 2005]

Can use this for active learning...

<table>
<thead>
<tr>
<th>Noun-phrase</th>
<th>Outdegree</th>
</tr>
</thead>
<tbody>
<tr>
<td>you</td>
<td>1656</td>
</tr>
<tr>
<td>we</td>
<td>1479</td>
</tr>
<tr>
<td>it</td>
<td>1173</td>
</tr>
<tr>
<td>company</td>
<td>1043</td>
</tr>
<tr>
<td>this</td>
<td>635</td>
</tr>
<tr>
<td>all</td>
<td>520</td>
</tr>
<tr>
<td>they</td>
<td>500</td>
</tr>
<tr>
<td>information</td>
<td>448</td>
</tr>
<tr>
<td>us</td>
<td>367</td>
</tr>
<tr>
<td>any</td>
<td>339</td>
</tr>
<tr>
<td>products</td>
<td>332</td>
</tr>
<tr>
<td>i</td>
<td>319</td>
</tr>
<tr>
<td>site</td>
<td>314</td>
</tr>
<tr>
<td>one</td>
<td>311</td>
</tr>
<tr>
<td>1996</td>
<td>282</td>
</tr>
<tr>
<td>he</td>
<td>269</td>
</tr>
<tr>
<td>customers</td>
<td>269</td>
</tr>
<tr>
<td>these</td>
<td>263</td>
</tr>
<tr>
<td>them</td>
<td>263</td>
</tr>
<tr>
<td>time</td>
<td>234</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context</th>
<th>Outdegree</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;x&gt; including</td>
<td>683</td>
</tr>
<tr>
<td>including &lt;x&gt;</td>
<td>612</td>
</tr>
<tr>
<td>&lt;x&gt; provides</td>
<td>565</td>
</tr>
<tr>
<td>provides &lt;x&gt;</td>
<td>565</td>
</tr>
<tr>
<td>provide &lt;x&gt;</td>
<td>390</td>
</tr>
<tr>
<td>&lt;x&gt; include</td>
<td>389</td>
</tr>
<tr>
<td>include &lt;x&gt;</td>
<td>375</td>
</tr>
<tr>
<td>&lt;x&gt; provide</td>
<td>364</td>
</tr>
<tr>
<td>one of &lt;x&gt;</td>
<td>354</td>
</tr>
<tr>
<td>&lt;x&gt; made</td>
<td>345</td>
</tr>
<tr>
<td>&lt;x&gt; offers</td>
<td>338</td>
</tr>
<tr>
<td>offers &lt;x&gt;</td>
<td>320</td>
</tr>
<tr>
<td>&lt;x&gt; said</td>
<td>287</td>
</tr>
<tr>
<td>&lt;x&gt; used</td>
<td>283</td>
</tr>
<tr>
<td>includes &lt;x&gt;</td>
<td>279</td>
</tr>
<tr>
<td>to provide &lt;x&gt;</td>
<td>266</td>
</tr>
<tr>
<td>use &lt;x&gt;</td>
<td>263</td>
</tr>
<tr>
<td>like &lt;x&gt;</td>
<td>260</td>
</tr>
<tr>
<td>variety of &lt;x&gt;</td>
<td>252</td>
</tr>
<tr>
<td>&lt;x&gt; includes</td>
<td>250</td>
</tr>
</tbody>
</table>
Component Size is Power-Law Distributed

Node Degree is Power-Law Distributed

[Jones, 2005]
CoTraining Summary

• Unlabeled data improves supervised learning when example features are redundantly sufficient
  – Family of algorithms that train multiple classifiers

• Theoretical results

• Many real-world problems of this type
  – Semantic lexicon generation [Riloff, Jones 99], [Collins, Singer 99]
  – Web page classification [Blum, Mitchell 98]
  – Word sense disambiguation [Yarowsky 95]
  – Speech recognition [de Sa, Ballard 98]
  – Visual classification of cars [Levin, Viola, Freund 03]
Bootstrapping: Learning to extract named entities

I arrived in Beijing on Saturday.

$x_1$: I arrived in ______ on Saturday.

$x_2$: Beijing
Example 3: Word sense disambiguation  [Yarowsky]

• “bank” = river bank, or financial bank??

• Assumes a single word sense per document
  – X1: the document containing the word
  – X2: the immediate context of the word (‘swim near the ___’)

Successfully learns “context $\rightarrow$ word sense” rules when word occurs multiples times in document.
Example 4: Bootstrap learning for IE from HTML structure
[Muslea, et al. 2001]

X₁: HTML preceding the target
X₂: HTML following the target

R₁: \texttt{SkipTo "Phone : <i>"}

Name: <i> Gino’s </i><p> Phone: <i> (800) 111-1717 </i><p> Fax: (616) 111-...
Example Bootstrap learning algorithms:

- Classifying web pages [Blum&Mitchell 98; Slattery 99]
- Classifying email [Kiritchenko&Matwin 01; Chan et al. 04]
- Named entity extraction [Collins&Singer 99; Jones&Riloff 99]
- Wrapper induction [Muslea et al., 01; Mohapatra et al. 04]
- Word sense disambiguation [Yarowsky 96]
- Discovering new word senses [Pantel&Lin 02]
- Synonym discovery [Lin et al., 03]
- Relation extraction [Brin et al.; Yangarber et al. 00]
- Statistical parsing [Sarkar 01]
Many Exploitable Redundancies

- Hyperlink words, web page words
  - (page classification, hyperlink word sense)
- Email subject line, email body
  - (email classification)
- Statements of same fact on many different websites
  - EventDates(ElvisBirthday, January 28)
- Assertions in both text, and tables
  - Semi-structured HTML
  - Excel spreadsheets
- Directory names, directory contents
- Activity clusters from email text, or social network
- Calendar events, email before and after meeting
- Deductive inference, when knowledge available
Easily obtained lists for some entities...

List of Companies

(Corrected)

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

3Com Corp
3M Company
A.G. Edwards Inc.
Abbott Laboratories
Abercrombie & Fitch Co.
ABM Industries Incorporated
Acc Hardware Corporation
ACT Manufacturing Inc.
Acterna Corp
Adams Resources & Energy, Inc.
ADC Telecommunications, Inc.
Adelphia Communications Corporation
Admistrav, Inc.
Adobe Systems Incorporated
Adolph Coors Company
Advance Auto Parts, Inc.
Advanced Micro Devices, Inc.
AdvancePCS, Inc.
Advantica Restaurant Group, Inc.
The ABS Corporation
Aetna Inc.
Affiliated Computer Services, Inc.

---

Erik Demaine's List of Events

Disclaimer

Calendar for: October 2004
Search for keywords: in any field sorted by

September 2004 -- October 2004 -- November 2004

I CDCCG 2004
10
11
I CDCCG 2004
final

POCS 2004
17
POCS 2004
SODA 2005
final

20
21
22
23
What is relation between “Elvis” and “January 8”?

Elvis Presley: The Early Years
View. Elvis Aaron Presley January 8, 1935 - August 16, 1977. Elvis charted more songs on Billboard’s Hot 100 than any other artist. ...
www.titlesweb.com/elvis.htm - 65k - 19 Sep 2004 - Cached - Similar pages

Listmanial Celebrate the king every January 8 and August 16
... Celebrate the king every January 8 and August 16 by Stephen Verhaeren, Movie Watcher. ...
4 Elvis - His Best Friend Remembers DVD (DVD) Average Customer Review: ...
www.amazon.com/exec/obidos/fg/listmanial/list-browse/J2SR569MZ2WZ - 72k - Cached - Similar pages

Elvis Aaron Presley was born on January 8
Back, Elvis Aaron Presley was born on January 8, 1935. Elvis is known as the “King of Rock and Roll”. Elvis music was influenced ... www.rock%20website/Elvis.htm - 6k - Cached - Similar pages

metaMaze - Born on January 8
... 18 people born on January 8 1937 - Shirley Bassey (singing) 1992 ... would take off."
1935 - Elvis Presley (singing/entertainment icon) “I knew ... www.metascape.com/days/0108.html - 13k - Cached - Similar pages

www.On-This-Day.com - January 8
January 8: 1705 - Georg Friedrich Handel’s opera “Almira” was produced in Hamburg. ...
1957 - Elvis took the US Army pre-induction exam on his 22nd birthday. ...
www.on-this-day.com/onthisday/thedays/music/jan08.htm - 4k - Cached - Similar pages

Elvis Presley
... Gladys, Elvis and Vernon Presley 1937. Born January 8, 1935, in East Tupelo, Mississippi, Presley was the son of Gladys and Vernon Presley, a sewing machine ... www.history-of-rock.com/elvis_presley.htm - 8k - Cached - Similar pages

Elvis Presley January 8, 1935 - August 16, 1977
Some agent strategies for generating tasks

- Collect more data from web
  - To learn about specific entities (e.g., “Rolling Stones”)
  - To learn meaning of particular language (e.g., “will attend”)
  - To locate easy-to extract facts (e.g., web pages with lists)

- Learn regularities from the populated KB
  - “Most LTI office names are of the form “NSH dddd”

- Explore specializations of ontological categories
  - What distinguishes personal home pages that contain publications from those that don’t? Can this be predicted from other (extractable) features of the home page?

- Explore specializations of language structures
  - Which ‘location’ entities share surrounding language? e.g., “the city of ?x,” Do they share other properties?
Some Types of Knowledge to Learn

• Linguistic regularities
  – \{“spoon”,”fork”,”chopsticks”\} occur often in “eat with my ___”

• HTML layout regularities
  – HTML lists often contain items of the same type

• Web site regularities
  – University departments often have a page listing all faculty

• Regularities over extracted facts
  – ‘Professors typically have more publications than their advisees’

• Temporal stability
  – Birthdays don’t change. Stock prices do.