

Learning on the Internet

Mike Perkowitz and Oren Etzioni
University of Washington

<http://www.cs.washington.edu/homes/map/ILA.html>

Finding Information on the Net

The number and diversity of information sources on the net is growing rapidly.

What kinds of software tools can help people search for information?

Standard approaches:

- Gopher, WAIS, the Webcrawler, Lycos
 - unable to interpret the results of searches
 - unable to use multiple information sources in concert

Sample AI Systems

- SIMS (Knoblock, Arens, & Hsu 1994)
- The Information Manifold (Levy, Srivastava, & Kirk 1994)
- The Internet Softbot (Etzioni & Weld 1994)

Softbots (Software Robots)

Softbot: an AI program that interacts with a real-world software environment.
(Etzioni & Segal '92)

Sensors: `netfind`, `INSPEC`, `finger`.

Effectors: `ftp`, `mail`, `lpr`.

UW Internet softbot:

- Human requests translated into planner goals.
- *Active:* plans, **executes**, and recovers from errors
(unlike Wilensky's UC, Microsoft Wizards).
- *General purpose:* a softbot is worth a thousand shell scripts.
(unlike knowbots or taskbots).

AI systems

Limitations

- rely on sophisticated **models** of information sources
- models must be hand coded
- sources unknown to the **programmers** are unknown to the AI system

How to keep up with the growth of the Web?

How to keep up with the growth of the Web?

Automatically learn models of information sources, services (S).

Four subproblems:

- **Discovery:** how does a software agent find new and unknown S ?
- **Protocol:** what are the mechanics of accessing S and parsing the response?
- **Semantics:** how does the agent understand the information available?
- **Quality:** what is the accuracy, reliability, and scope of S ?

Internet Learning Agent (*ILA*) focuses on **semantics**.

An Example Interaction

Facts in *ILA*'s model:

(lastname person37 Etzioni)

(userid person37 Etzioni)

(firstname person37 Oren)

(department person37 CS)

(mail-stop CS FR-35)

Trick: query with a familiar individual.

query> Etzioni

output> Oren Etzioni 209 FR-35

From this query and knowledge, *ILA* can hypothesize:

- The information is about people.
- The first field in the output is `firstname`.
- The second field is `lastname` or `userid`.
- The third field is *unknown*.

Key assumption: output fields have a consistent interpretation!

The Semantics Learning Problem

Given:

- *ILA's* model of the world:
 - objects: `person35, person37, ...`
 - relations: `(lastname person35 Perkowitz), ...`
- Query/response pairs from \mathcal{S}

Determine: model which explains the observed query/response pairs.

$\mathcal{S}_2 = \text{lastname}(\text{person})$

$\mathcal{S}_4 = \text{mail-stop}(\text{department}(\text{person}))$

Query = `lastname(person)`

Goal: learn accurate model of \mathcal{S} , using as few queries as you can.

ILA's Learning Method

Sketch: Query S with familiar objects and *explain* the responses.

- 1. Which object should *ILA* query the S with?**
2. What is the appropriate mapping from object to query string?
- 3. What are possible explanations for each token in the response?**
4. How to evaluate competing explanations?
5. Termination criterion?

See Perkowitz & Etzioni (IJCAI '95) for the details.

The Correspondence Heuristic

Saint Augustine's Method: learning by example.

Me: "George Washington's *uxor* was Martha."

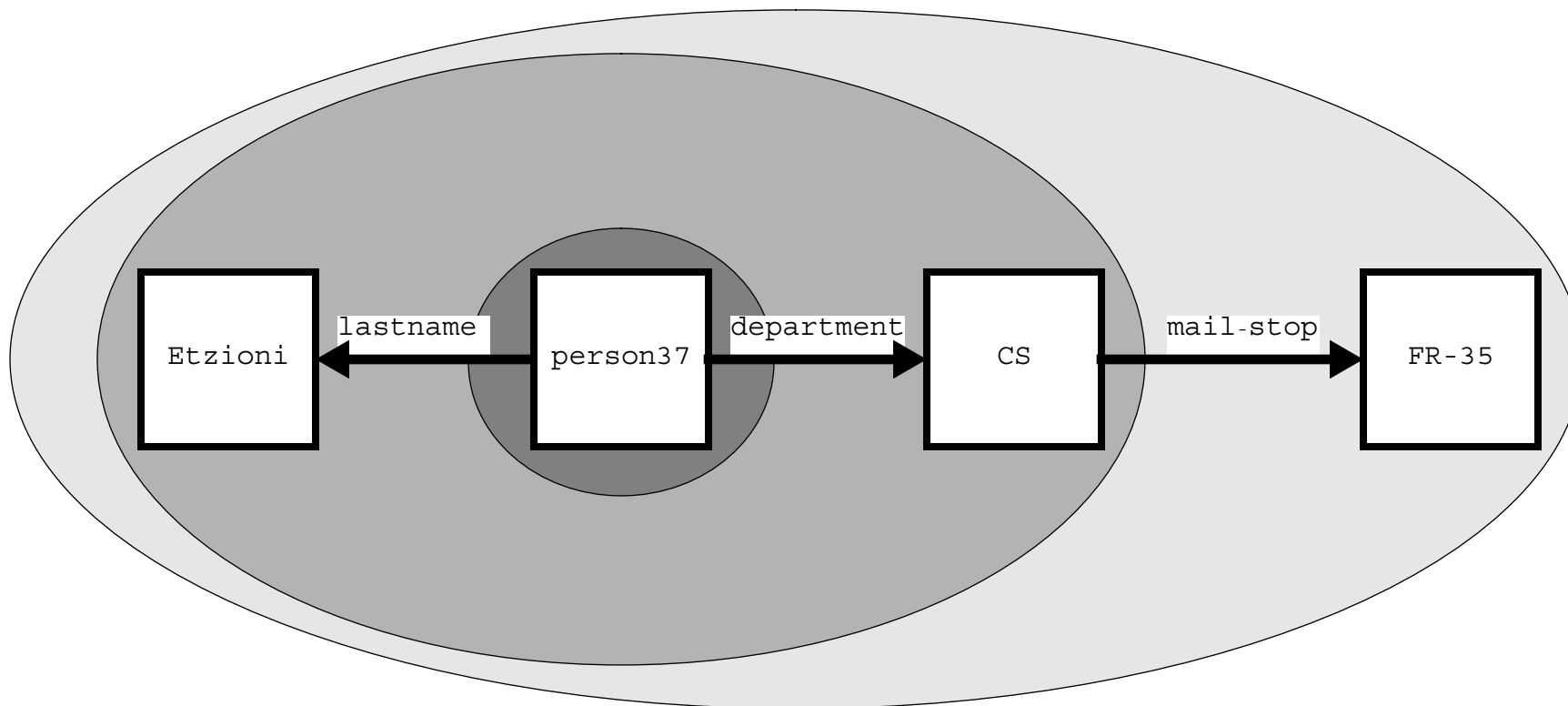
You: "Who was Jefferson's *uxor*?"

I assume you:

- are familiar with George Washington.
- have a concept corresponding to *uxor* (e.g. *wife*).
- will establish a *general correspondence* between *uxor* and *wife* based on the example.

Leap of faith formalized as a determination in Perkowski & Etzioni (IJCAI '95).

Fuzzy Relational Pathfinding (Richards & Mooney '92)



- Starting node
- Frontier after spreading out one link
- Frontier after spreading out two links

`FR-35 = mail-stop(department(person37))`

Discriminating Queries

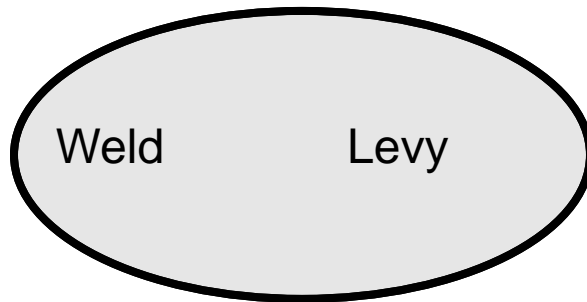
How does *ILA* decide between the hypothesis `lastname` and the hypothesis `userid`?

Answer: *ILA* queries with a person whose `userid` \neq `lastname`.

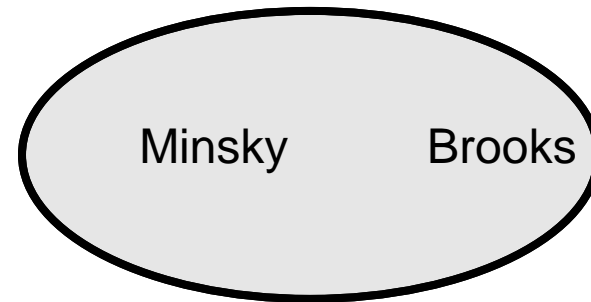
e.g., `(lastname person35 Perkowitz)`
`(userid person35 map)`

- *ILA* converges quickly on correct hypothesis
- Decreases the number of queries needed
- An *optimization*, not required to get a good solution

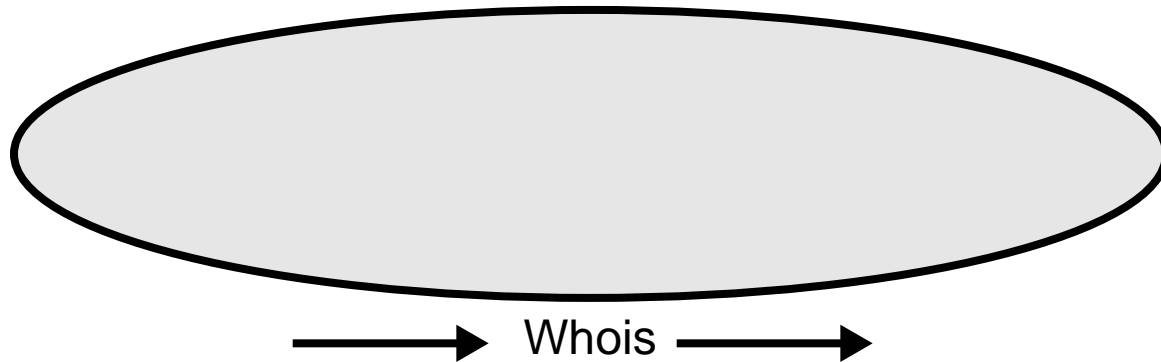
Bootstrapping Approach



UW phonebook



MIT phonebook

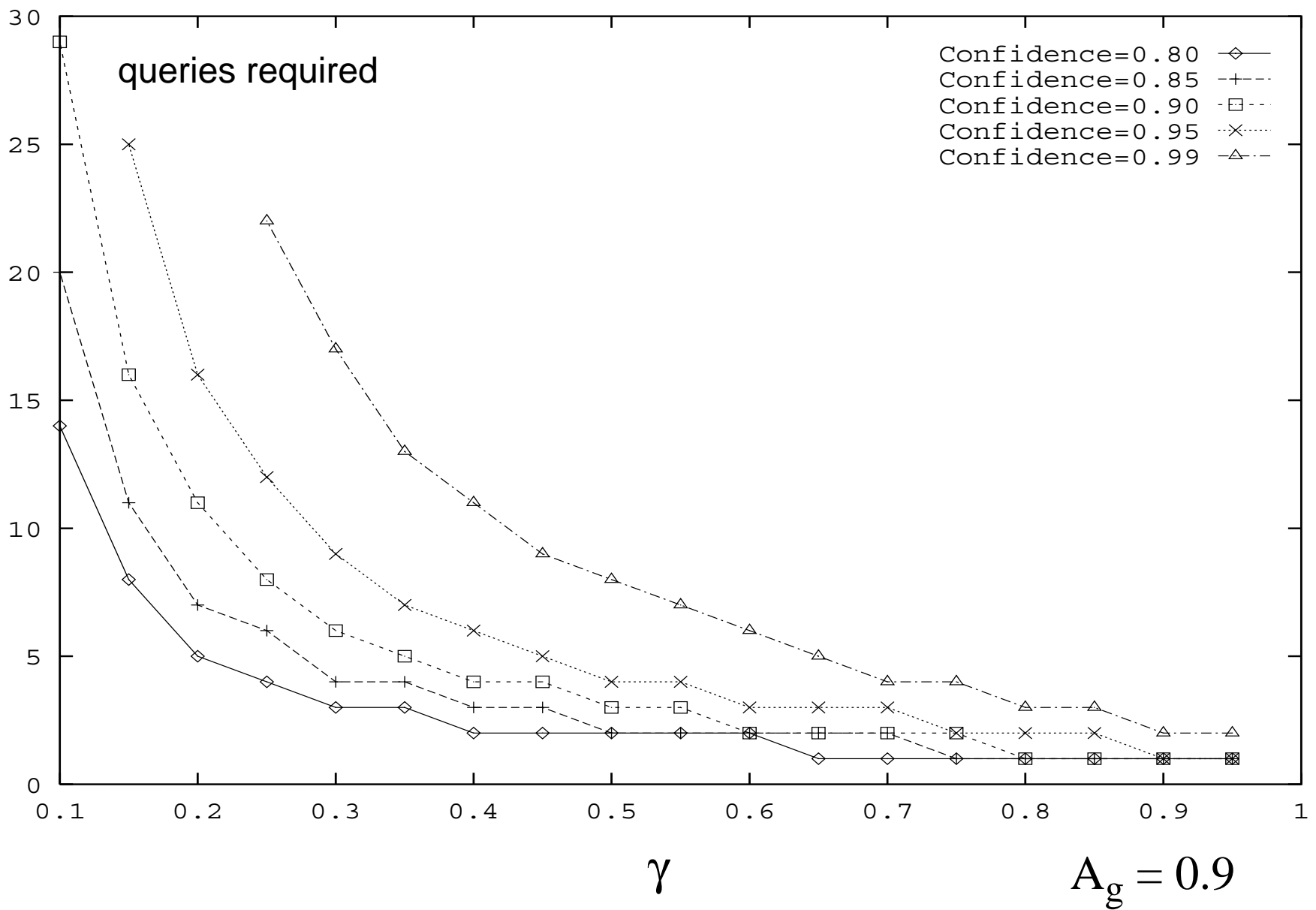


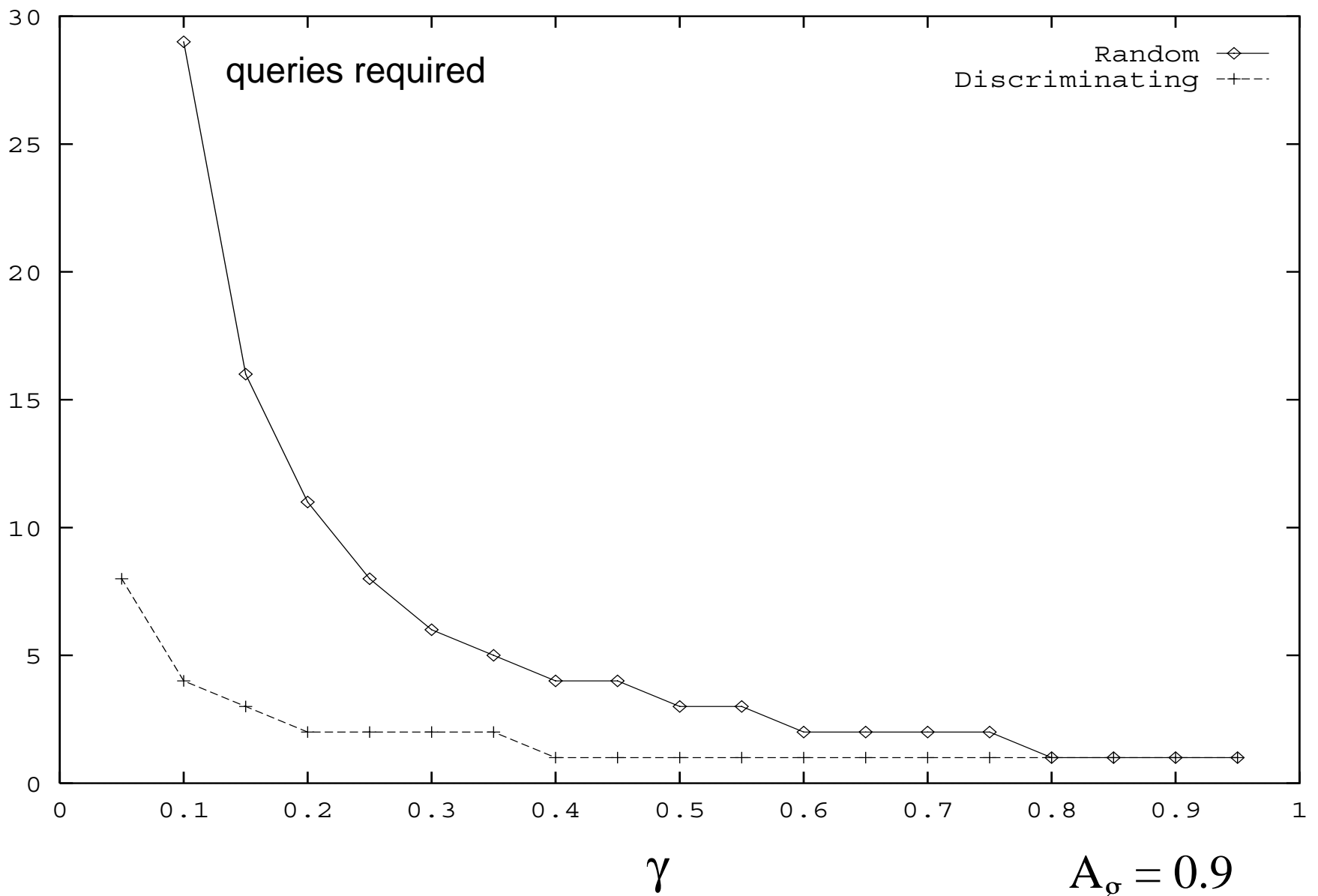
- *Bootstrapping approach*: use known people to learn **spanning** knowledge sources, and then use those knowledge sources to learn about new sites.
- **Area queries** allow ILA to learn about people at a new site
 - e.g. ILA will query Whois with “MIT”

Learning by Bootstrapping

	Fields							Queries		Time	
	1	2	3	4	5	6	7	Total	Hits	Internet	CPU
UW	✓	✓	✓	✓	✓	✓	✓	16	16	0:19	0:24
Whois	✓	✓			✓	✓	✓	50	22	26:09	3:04
Berkeley	✓	✓			---	---	✓	24	5	6:07	0:54
Brown	✓	✓			✓	---	---	69	6	11:06	2:58
Caltech	✓	✓				✓	✓	22	11	4:02	0:17
Rice	✓	✓			---	✓	X	36	2	6:53	1:15
Rutgers	✓	✓						36	8	5:29	0:57
UCI	✓	✓			---	---	✓	34	13	12:02	6:60

Fields: 1=firstname, 2=lastname, 3=title, 4=dept, 5=phone, 6=email, 7=userid
 ✓=learned, ---=unlearned, Space=unavailable, X=wrong





$A_g = 0.9$
Confidence = 0.9

Conclusion

Problem: information explosion.

Approach: software agents that learn about the net (*ILA*).

Results:

- **very few queries necessary for learning**
- ILA uses bootstrapping to learn models of information sources
- probabilistic analysis of sample complexity (see paper for details)

Future work:

- category mismatch (Wiederhold 1992)
- more experiments
- more domains
- discovery & quality problems