Active Learning: Recent Advances

Charles Ling
PhD (Upenn), Professor
University of Western Ontario, Canada
(加拿大 西安大略大学)

cling@csd.uwo.ca; cling.csd.uwo.ca

Joint work with Jun Du et al

KDD08, ICDM09, TKDE10, ECML10, ICDM10, ????11

Outline

- Introduction
- Active Learning with Generalized Queries (AGQ)
- Experimental Comparison
- Discussions
- Recent Advances
- Conclusions and Future Works

Introduction: Why

- Traditional ML: large training set; passive
- Human learning: small training set…
  - But we learn actively, in many ways…
  - ML: active: acquire missing values, new examples, ask for labels of unlabeled examples, …
  - Reduce # of labelled examples significantly
- Which unlabeled examples to ask??
  - Correct labels provided by users or "Oracle"
  - More powerful than semi-supervised learning

Introduction: Previous Works

- Pool-based Active Learning (Tong & Koller 2002, Roy & Mccallum 2001; Baram, El-Yaniv, & Luz 2004)
  - Choose one with most uncertain predicted label
- Direct Query Construction (Ling & Du, KDD 2008)
  - Construct one with most uncertain predicted label
- Exponential speed-up proved for threshold concept
- On real data, speed-up is often small

Introduction – Limitations

- Previous works always ask specific queries
- Example: Predicting heart disease based on a patient dataset with 30 attributes
  - Query: For name="Jane", age="35", gender="female", weight="85 kg", blood pressure="160/90", temperature="98F", chest pain="no", and ……
  - Answer: Yes/No (for this specific patient).

Limitations of Specific Queries

- Many of the attributes may not be relevant;
  - Example: name, temperature, … etc. are not relevant
  - Too many attribute confuse human experts;
- Labels returned are also specific (only for the specific queries).

Outline

- Introduction
- Active Learning with Generalized Queries (AGQ)
- Experimental Comparison
- Discussions
- Recent Advances
- Conclusions and Future Works

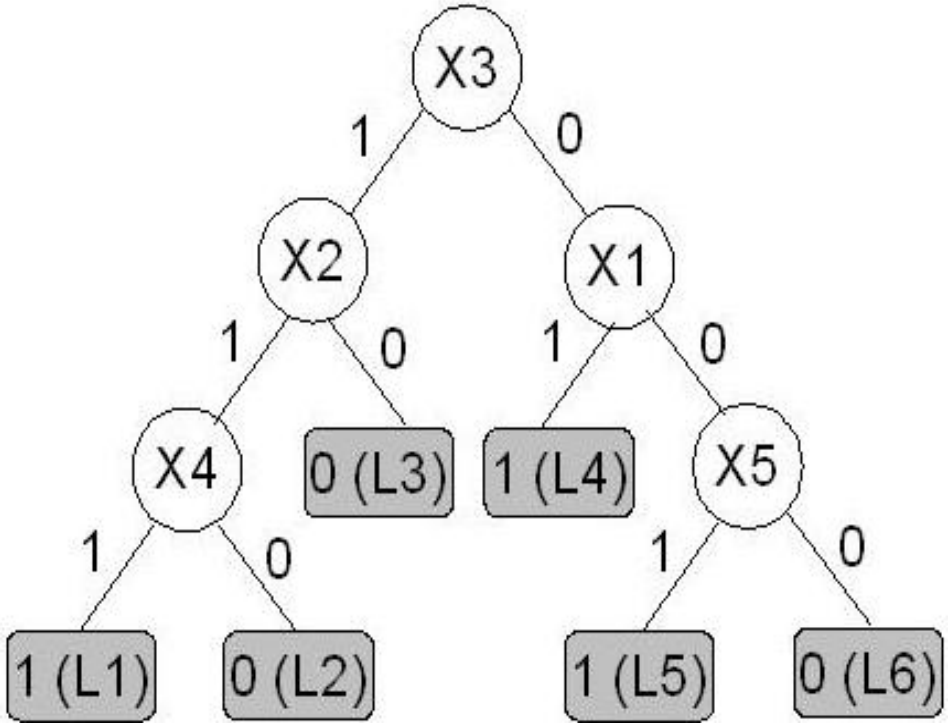
An Extension: Generalized Queries

- Active learners can ask generalized queries.
- Example: Query: “are people over 50 with chest pain likely to have heart disease?” (only 2 relevant attributes, age and chest pain)
- Advantages
  - More natural and relevant
  - One generalized query = many specific queries
  - Answer apply to all these examples
- Need to identify irrelevant features

Can Feature Selection Do It?

Feature Selection + Pool_AL < AGQ

Suppose the target is:



All attributes are relevant, thus FS could do nothing.

Generalized query for L6:
[x1=0, x3=0, x5=0] ->?, should obtain certain answer from oracle.

Difficulties of Generalized Queries

- Answer from the oracle can often be uncertain.
  - Query : “are people over 50 with chest pain likely to have heart disease?” Yes with a 90% probability
  - Query: “are people over 50 likely to have heart disease?” Yes with a 60% probability
  - More general, more powerful, but more uncertain
- Answer could be inaccurate (90% or 92%?)
- How to ask good generalized queries based on limited info (small training set)?

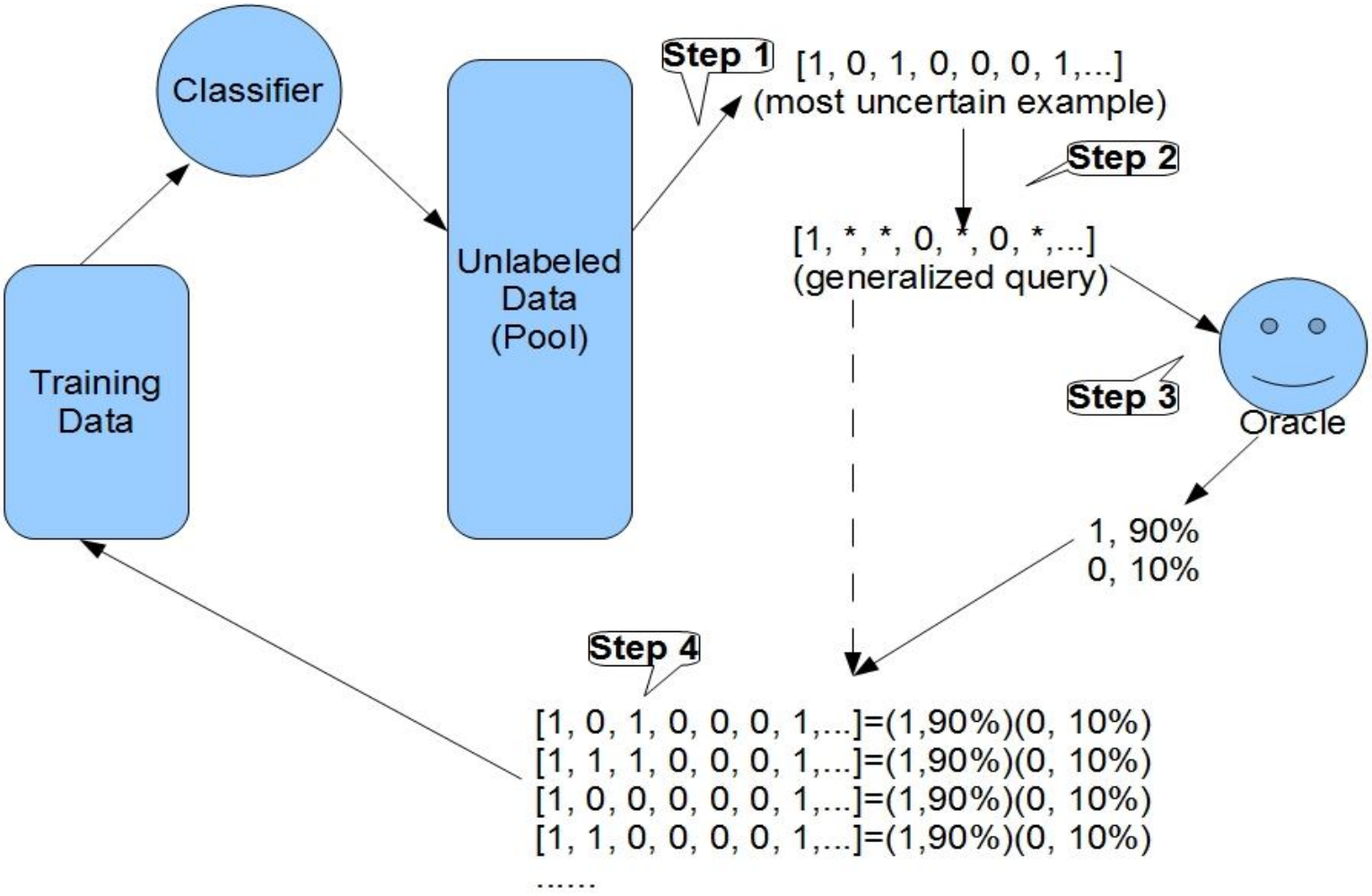
Our Task…

Assuming oracle can answer GQ with prob label, design a robust active learner that attempts to ask few generalized queries (large speed-up)

End results:

- AGQ (Active learning with Generalized Queries)
  - Query: over 50 with chest pain [ICDM 2009]
- AGQ+ [IEEE TKDE 2010]
  - Query: age 50-60, mild or severe chest pain
- Applications in medical domain, text mining…
- More recent advances in AL

AGQ Process


Classifier
Unlabeled Data (Pool)
Training Data
Step 1
[1, 0, 1, 0, 0, 0, 1,...]
(most uncertain example)
Step 2
[1, *, *, 0, *, 0, *,...]
(generalized query)
Step 3
Oracle
1, 90%
0, 10%
Step 4
[1, 0, 1, 0, 0, 0, 1,...]=(1,90%)(0, 10%)
[1, 1, 1, 0, 0, 0, 1,...]=(1,90%)(0, 10%)
[1, 0, 0, 0, 0, 0, 1,...]=(1,90%)(0, 10%)
[1, 1, 0, 0, 0, 0, 1,...]=(1,90%)(0, 10%)
.....

Key Steps

1. Finding the Most Uncertain Example;
2. Constructing the Generalized Query;
3. Asking Generalized Queries to Oracle;
4. Updating the Training Dataset;

1. Finding Most Uncertain Example

- Build model on current training examples;
- Find most uncertain example
  - From a pool of unlabeled examples (pool-based)
  - Direct query construction (KDD 2008)
- Example: [1, 0, 1, 1, 0, 1] (from the pool)
  - Prediction: 52% for class 1; 48% for class 0;
  - Most uncertain: prob closest to 50%

2. Constructing Generalized Query

- Find irrelevant attributes in the chosen example (i.e., the most uncertain example)

- Example:
  - Input: the chosen example [1, 0, 1, 1, 0, 1]
  - Output: generalize to [1, *, 1, *, 0, 1]

How to Generalize?

- Main Idea
  - For all irrelevant attributes, examples with any combination of their values have same predicted label and probability.
- Example:
  - Given: x_1=[1, 0, 1, 1, 0, 1], and $P(1|x_1) = 52\%$

    2^{nd} and 4^{th} attributes (red ones) are irrelevant
  - Then: $P(1|x_2) = 52\%$, x_2=[1, 0, 1, 0, 0, 1];

    $P(1|x_3) = 52\%$, x_3=[1, 1, 1, 0, 0, 1];

    $P(1|x_4) = 52\%$, x_4=[1, 1, 1, 1, 0, 1];

Greedy Hill-Climbing Search

Given a small threshold p:

- For each attribute, construct a fix number of examples with different attribute values;
- Estimate the probability of these examples by the current classifier;
- Choose the attribute with minimum probability variance v; if v <p, add it to the subset of irrelevant attributes
- Repeat, until v > p

P represents how “conservative” the generalization would be.

In AGQ, p = 0.0001

(More details: see paper)

3. Asking Generalized Queries

- Asking generalized queries to the oracle, and obtain an answer with probability.
- If generalized queries are “conservative”, answers should be certain
- On UCI datasets, we first construct the “target concept” on the whole datasets to simulate the oracle

- Example:
  - Query: How to classify [1, *, 1, *, 0, 1]?
  - Answer: 90% for positive

4. Updating Training Set

- Update training set, according to the generalized queries and probability answers
- Example:
  - Given: [1, *, 1, *, 0, 1]; 90% for 1 (10% for 0)
  - Add four specific examples into training set, with probability labels (90% for 1 and 10% for 0)

    [1, 0, 1, 0, 0, 1], [1, 0, 1, 1, 0, 1],

    [1, 1, 1, 0, 0, 1], [1, 1, 1, 1, 0, 1].
- Limit the number of examples added

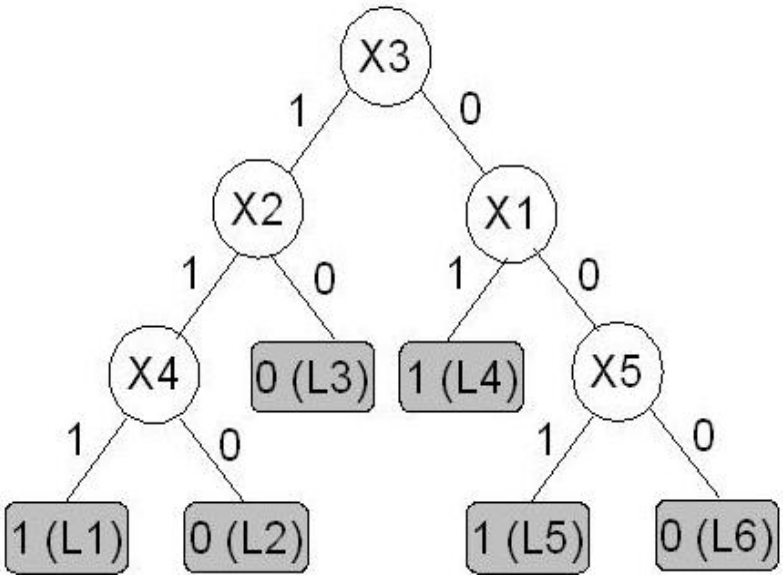
Outline

- Introduction
- Active Learning with Generalized Queries (AGQ)
- Experimental Comparison
- Discussions
- Recent Advances
- Conclusions and Future Works

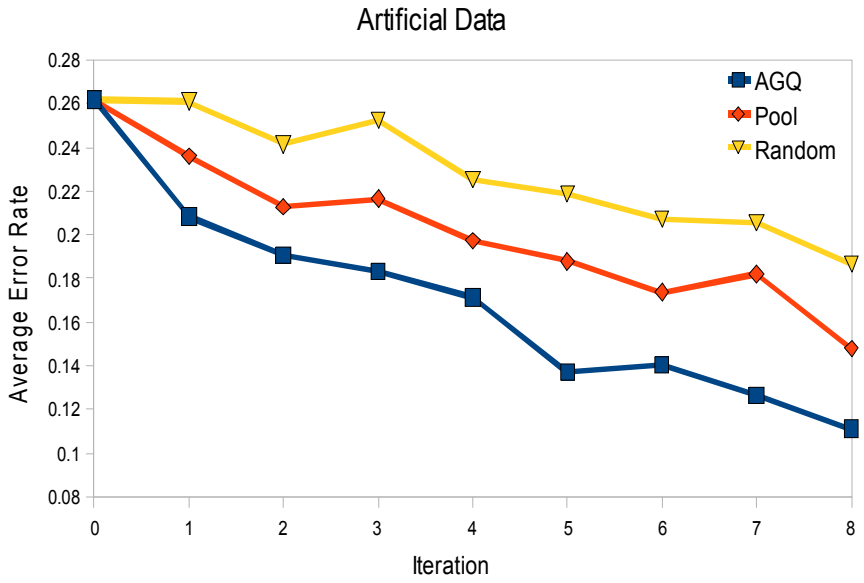
Artificial Data

Target:

5 relevant + 5 irrelevant att.



Comparison:



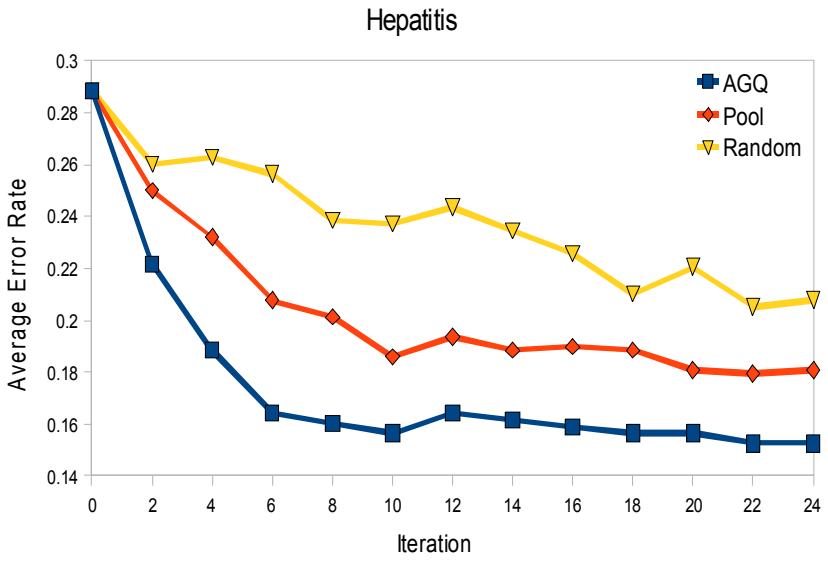
	AGQ	Pool
Query 1	[1, 1, 1, 0, *, *, *, *, *, *]	[1, 1, 1, 0, 1, 1, 1, 1, 0, 0]
Classified by Leaf(ves)	L2	L2
Ideal Query	[*, 1, 1, 0, *, *, *, *, *, *]	[*, 1, 1, 0, *, *, *, *, *, *]
Answer	0, 100%	0
No. of Examples	10	1
Error Rate	0.18	0.27
Query 2	[0, *, 0, 1, *, *, *, *, *, *]	[1, 0, 1, 1, 0, 0, 1, 0, 0, 1]
Classified by Leaf(ves)	L5, L6	L3
Ideal Query	-	[*, 0, 1, *, *, *, *, *, *, *]
Answer	0, 54%	0
No. of Examples	10	1
Error Rate	0.21	0.22
Query 3	[0, 1, 0, 1, 1, 0, 0, *, 1, *]	[1, 1, 1, 1, 0, 1, 1, 1, 0, 1]
Classified by Leaf(ves)	L5	L1
Ideal Query	[0, *, 0, *, 1, *, *, *, *, *]	[*, 1, 1, 1, *, *, *, *, *, *]
Answer	1, 100%	1
No. of Examples	8	1
Error Rate	0.16	0.26
Query 4	[0, 1, 0, 1, 0, 1, *, *, 0, *]	[1, 0, 1, 1, 0, 1, 0, 0, 1, 1]
Classified by Leaf(ves)	L6	L3
Ideal Query	[0, *, 0, *, 0, *, *, *, *, *]	[*, 0, 1, *, *, *, *, *, *, *]
Error Rate	0, 100%	0
No. of Examples	8	1
Error Rate	0.17	0.26
Query 5	[1, *, 0, *, 0, *, 1, *, *, *]	[1, 1, 1, 0, 0, 1, 0, 0, 1, 1]
Classified by Leaf(ves)	L4	L2
Ideal Query	[1, *, 0, *, *, *, *, *, *, *]	[*, 1, 1, 0, *, *, *, *, *, *]
Answer	1, 100%	0
No. of Examples	10	1
Error Rate	0.13	0.2

UCI Data

14 UCI datasets

Dataset	# Att	# Inst	Class Dist.	Train
breast-cancer	9	277	196/81	1/5
breast-w	9	699	458/241	1/10
colic	22	368	232/136	1/5
credit-a	15	690	307/383	1/20
credit-g	20	1000	700/300	1/100
diabetes	8	768	500/268	1/10
…				

Comparison on 1 typical data



Statistics on UCI Data

Dataset	# Don't-care Att.	#. Inst	Certainty of Oracle	Iteration of "Pool"	Iteration of AGQ	% Reduction of Iteration	AGQ (w/t/l)
breast-cancer	2.7 (30%)	14.54	95%	35	18	49%	W
breast-w	5.35 (59%)	32.31	87%	18	18	0%	T
colic	13.15 (60%)	35.68	91%	15	8	47%	W
credit-a	6.38 (43%)	16.43	88%	12	5	58%	W
credit-g	8.54 (43%)	4.97	87%	50	12	76%	W
diabetes	3.02 (38%)	27.31	89%	50	16	68%	W
heart-statlog	5.92 (46%)	12.52	89%	50	25	50%	W
…							
Avg.	12.53 (51.36%)	16.49	90.21%	35.14	24.21	36%	9/4/1

Outline

- Introduction
- Active Learning with Generalized Queries (AGQ)
- Experimental Comparison
- Discussions (skip)
- Recent Advances
- Conclusions and Future Works

Outline

- Introduction
- Active Learning with Generalized Queries (AGQ)
- Experimental Comparison
- Discussions
- Recent Advances
- Conclusions and Future Works

Recent Advances in AL (Skip)

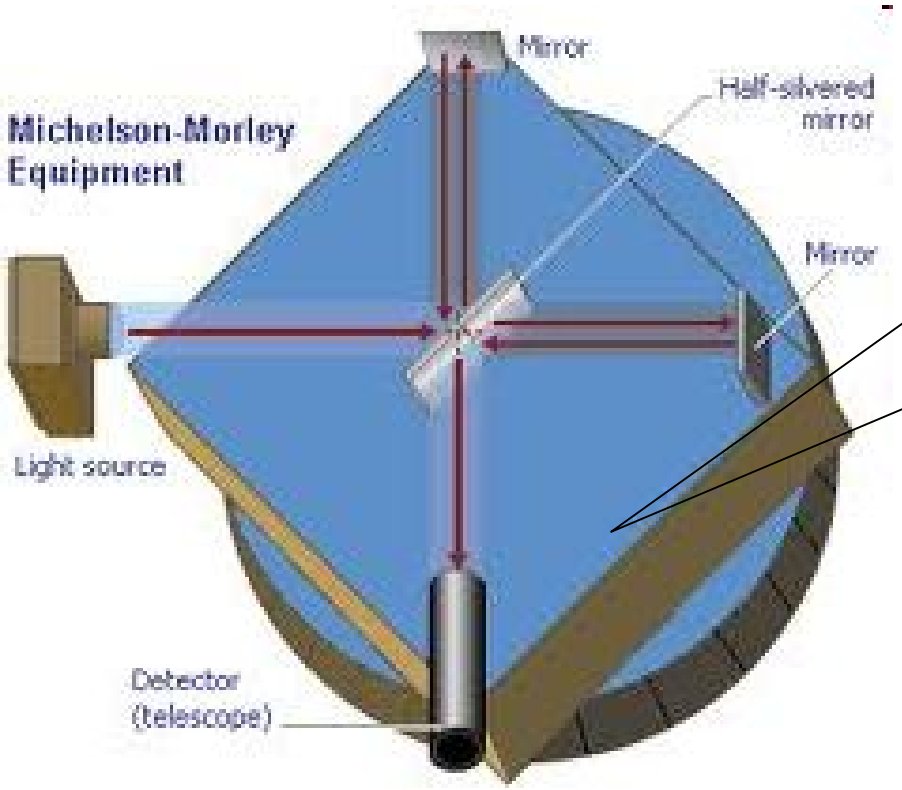
- A New Paradigm of Active Learning
- Hierarchical Classification for Next Search Engine
- Making reasonable assumptions on oracle
  - Cost-sensitive oracle
  - Ambiguous oracle
  - Oracle with explanation
  - …

A New Paradigm of Active Learning

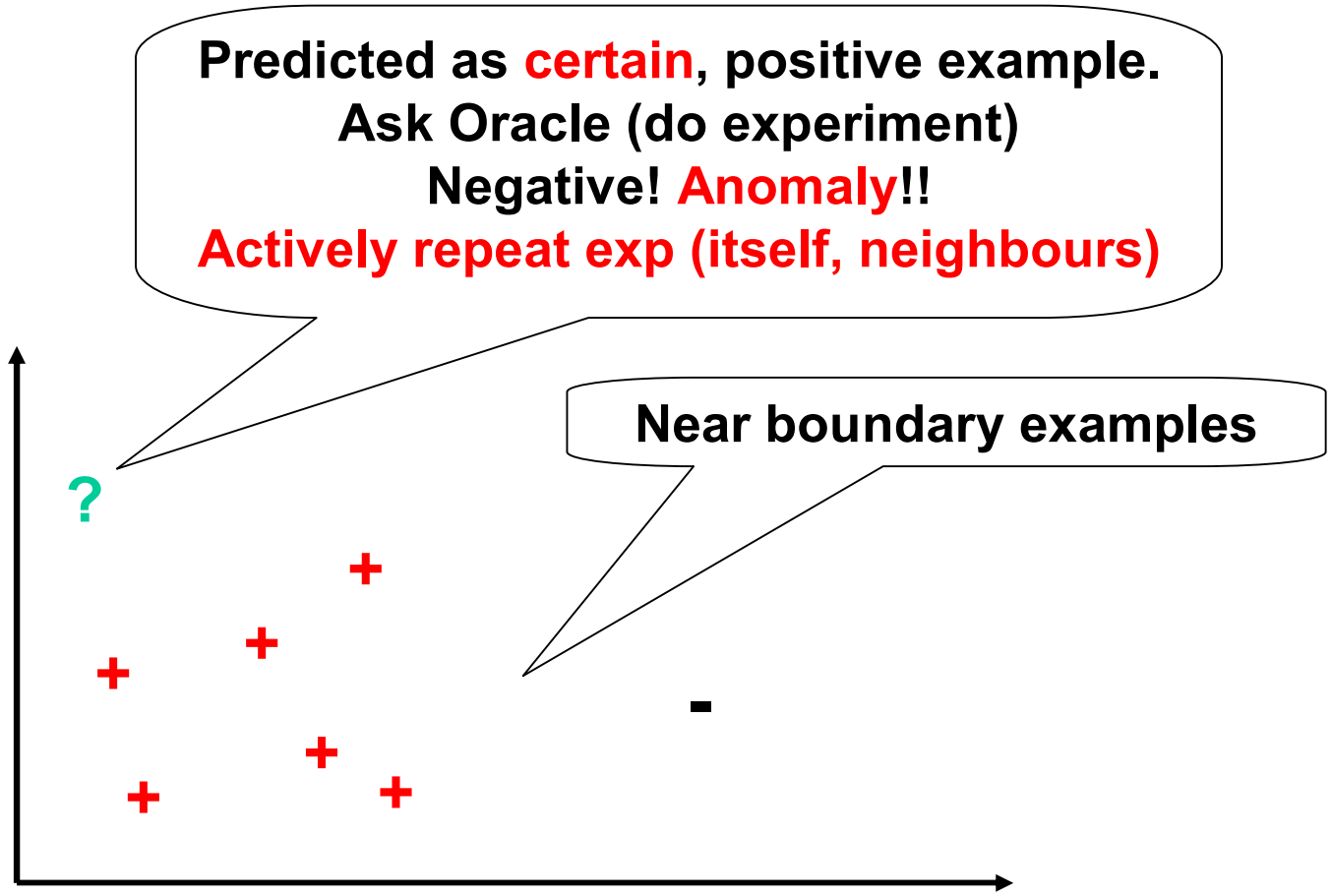
- Most previous works of AL (pool-based, AGQ, …) are based on asking near-boundary examples
- In scientific discovery, scientists are active learners... They do perform near-boundary exp
- What else?
- When anomaly (surprise) is found, they “repeat” experiments to confirm or disapprove it!

The Most Famous Failed Exp

- Newtonian physics predicted that the speed of light will change in the “aether wind”
- In 1887, Michelson–Morley experiment



Anomaly is found
Exp. repeatedly: AL
A new theory in 1905
Einstein Special Relativity

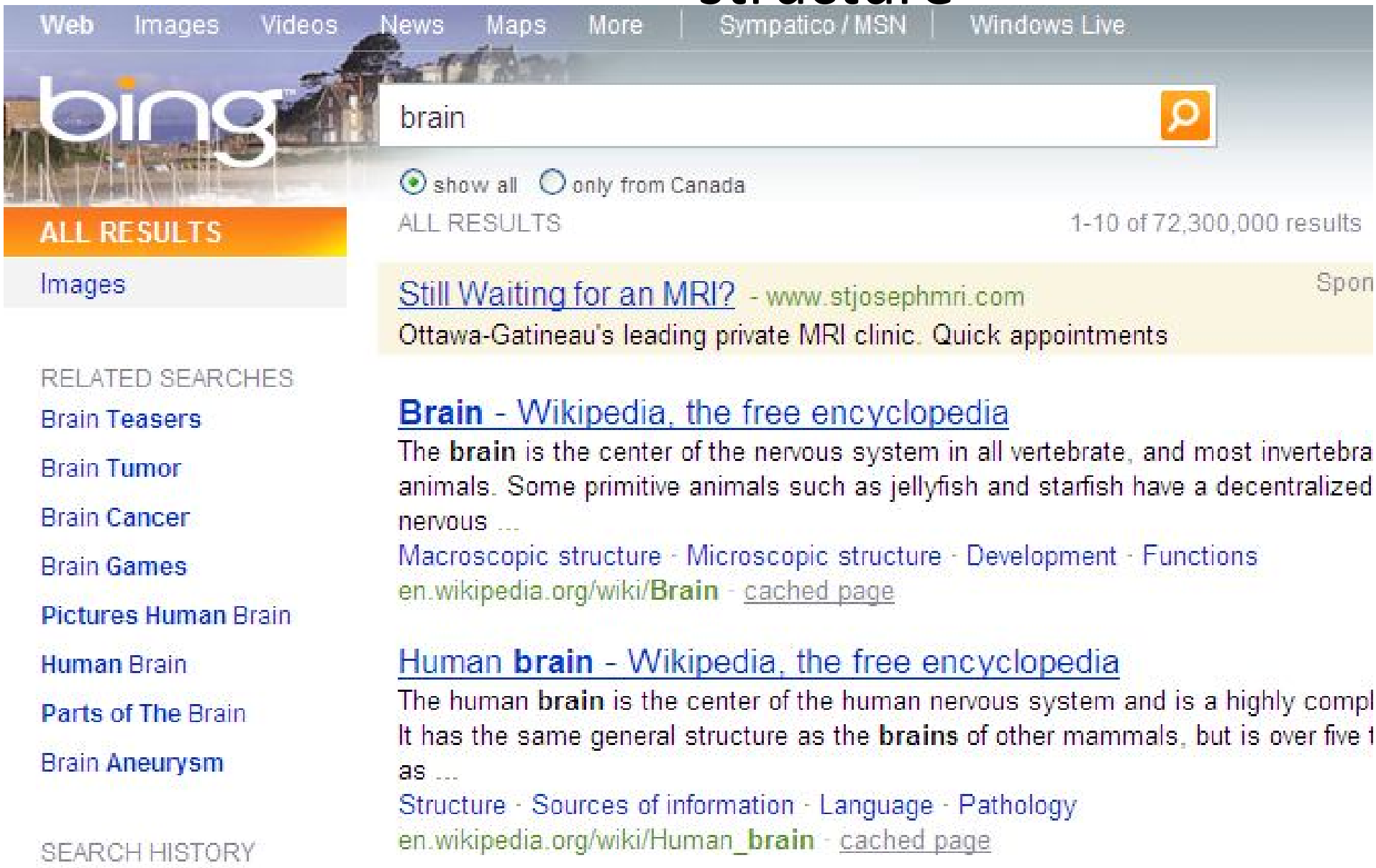


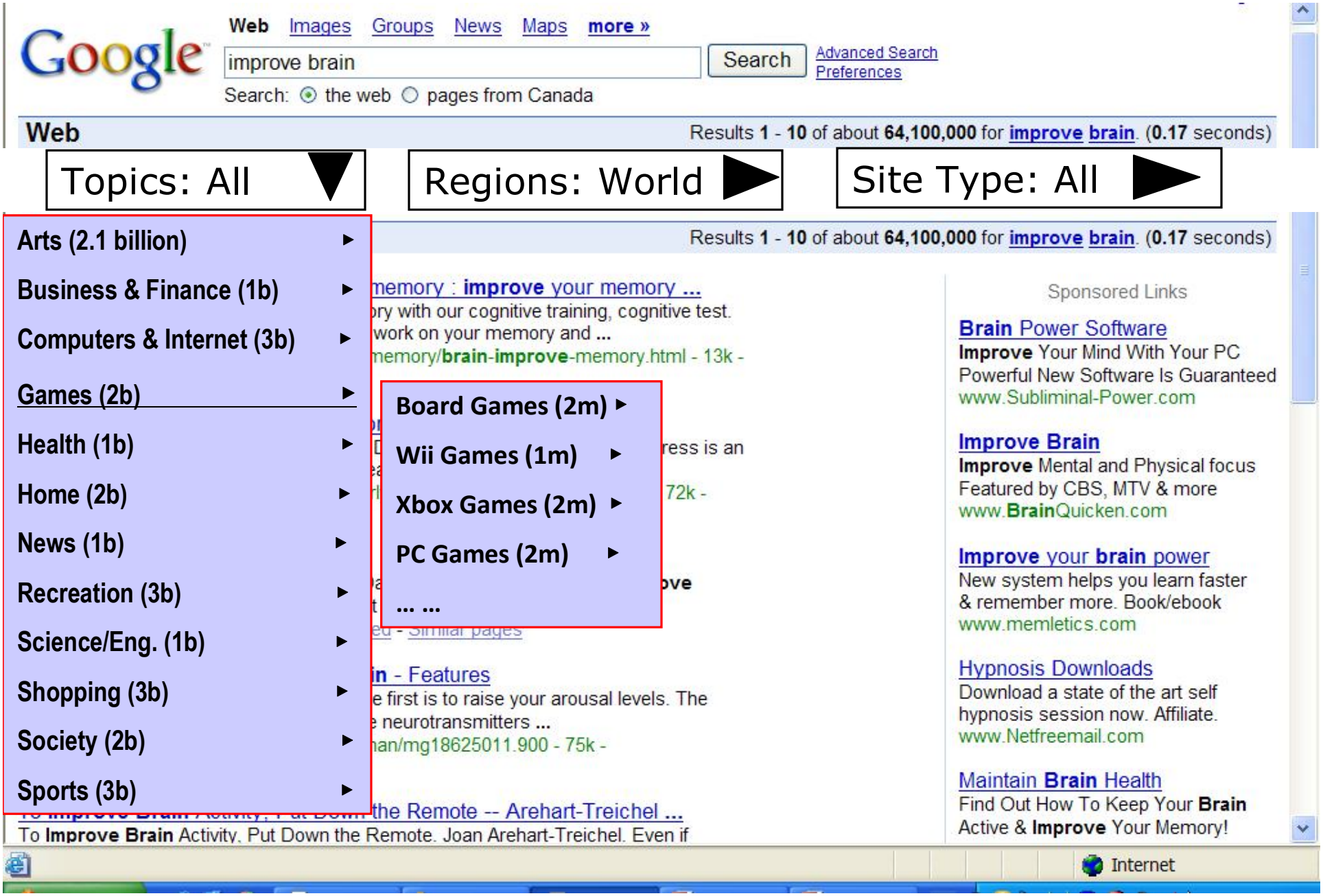
For noisy, probabilistic concepts, and unexplored space in complex concepts

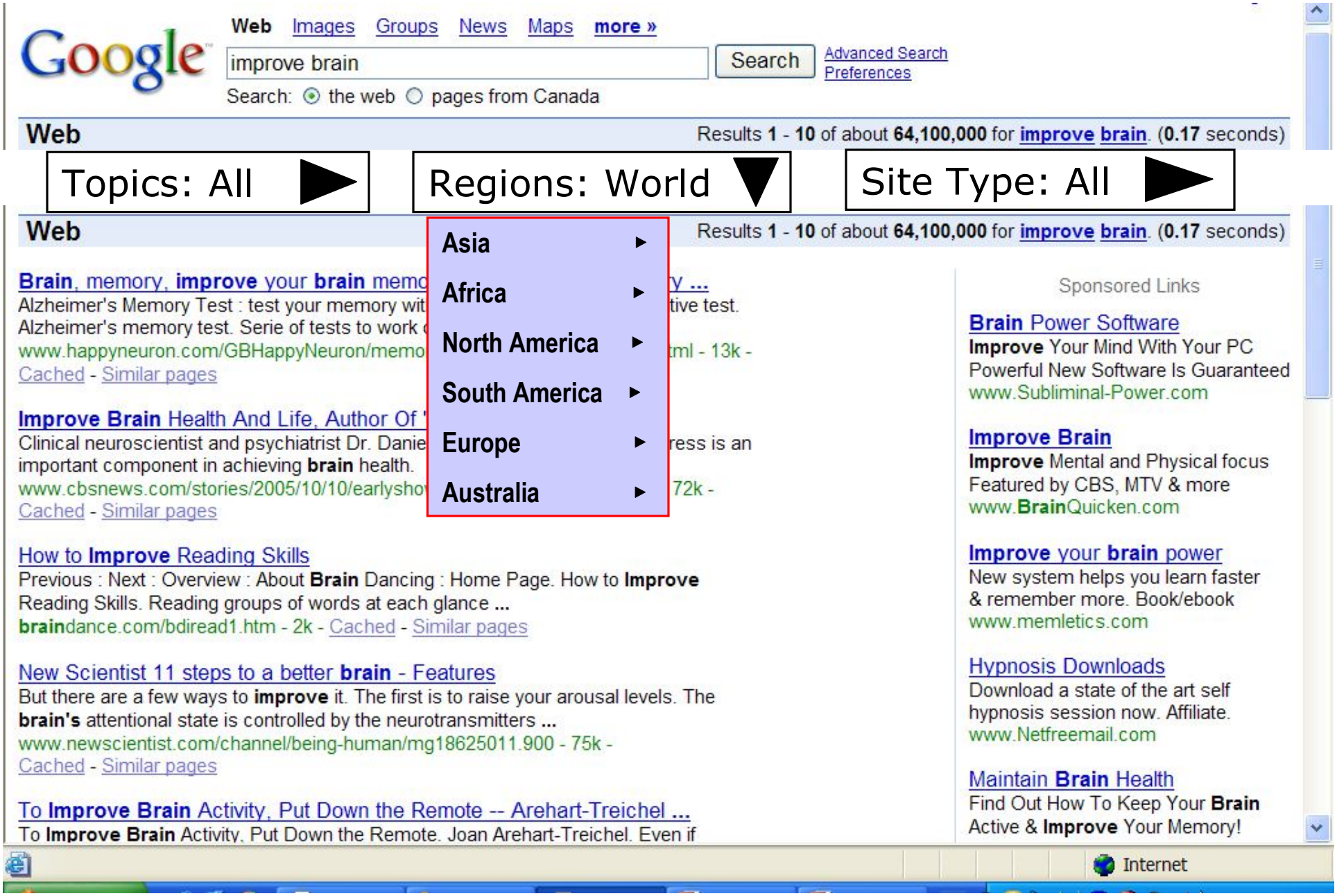
AL in Hierarchical Classification For the Next Search Engines

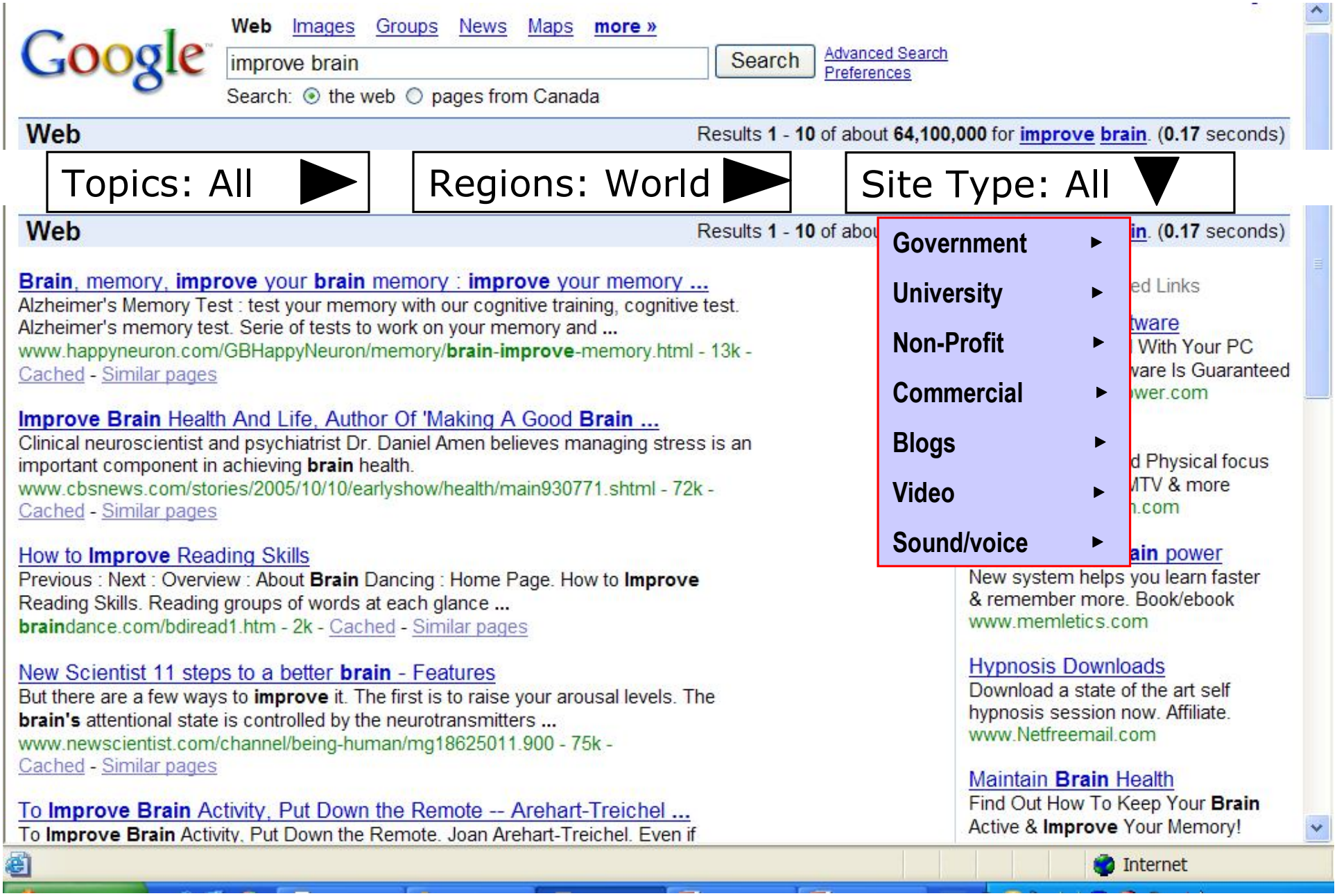
Skip

Current: Flat list, only limited structure




Google
Web Images Groups News Maps more »
improve brain
Search
Advanced Search
Preferences
Search: the web pages from Canada
Web
Results 1 - 10 of about 64,100,000 for improve brain. (0.17 seconds)
Topics: All
Regions: World
Site Type: All
Arts (2.1 billion)
Business & Finance (1b)
Computers & Internet (3b)
Games (2b)
Health (1b)
Home (2b)
News (1b)
Recreation (3b)
Science/Eng. (1b)
Shopping (3b)
Society (2b)
Sports (3b)
Board Games (2m)
Wii Games (1m)
Xbox Games (2m)
PC Games (2m)
... ...
Results 1 - 10 of about 64,100,000 for improve brain. (0.17 seconds)
memory : improve your memory ...
ory with our cognitive training, cognitive test.
work on your memory and ...
memory/brain-improve-memory.html - 13k -
ress is an
72k -
ove
in - Features
e first is to raise your arousal levels. The
e neurotransmitters ...
han/mg18625011.900 - 75k -
the Remote -- Arehart-Treichel ...
To Improve Brain Activity. Put Down the Remote. Joan Arehart-Treichel. Even if
Sponsored Links
Brain Power Software
Improve Your Mind With Your PC
Powerful New Software Is Guaranteed
www.Subliminal-Power.com
Improve Brain
Improve Mental and Physical focus
Featured by CBS, MTV & more
www.BrainQuicken.com
Improve your brain power
New system helps you learn faster
& remember more. Book/ebook
www.memletics.com
Hypnosis Downloads
Download a state of the art self
hypnosis session now. Affiliate.
www.Netfreemail.com
Maintain Brain Health
Find Out How To Keep Your Brain
Active & Improve Your Memory!
Internet


Google
Web Images Groups News Maps more »
improve brain
Search
Advanced Search
Preferences
Search: the web pages from Canada
Web
Results 1 - 10 of about 64,100,000 for improve brain. (0.17 seconds)
Topics: All
Regions: World
Site Type: All
Asia
Africa
North America
South America
Europe
Australia
Brain, memory, improve your brain memo ... y ...
Alzheimer's Memory Test : test your memory wit tive test.
Alzheimer's memory test. Serie of tests to work
www.happyneuron.com/GBHappyNeuron/memo tml - 13k -
Cached - Similar pages
Improve Brain Health And Life, Author Of
Clinical neuroscientist and psychiatrist Dr. Danie ress is an
important component in achieving brain health.
www.cbsnews.com/stories/2005/10/10/earlysho 72k -
Cached - Similar pages
How to Improve Reading Skills
Previous : Next : Overview : About Brain Dancing : Home Page. How to Improve
Reading Skills. Reading groups of words at each glance ...
braindance.com/bdiread1.htm - 2k - Cached - Similar pages
New Scientist 11 steps to a better brain - Features
But there are a few ways to improve it. The first is to raise your arousal levels. The
brain's attentional state is controlled by the neurotransmitters ...
www.newscientist.com/channel/being-human/mg18625011.900 - 75k -
Cached - Similar pages
To Improve Brain Activity, Put Down the Remote -- Arehart-Treichel ...
To Improve Brain Activity, Put Down the Remote. Joan Arehart-Treichel. Even if
Sponsored Links
Brain Power Software
Improve Your Mind With Your PC
Powerful New Software Is Guaranteed
www.Subliminal-Power.com
Improve Brain
Improve Mental and Physical focus
Featured by CBS, MTV & more
www.BrainQuicken.com
Improve your brain power
New system helps you learn faster
& remember more. Book/ebook
www.memletics.com
Hypnosis Downloads
Download a state of the art self
hypnosis session now. Affiliate.
www.Netfreemail.com
Maintain Brain Health
Find Out How To Keep Your Brain
Active & Improve Your Memory!
Internet


Google
Web Images Groups News Maps more »
improve brain
Search
Advanced Search
Preferences
Search: the web pages from Canada
Web
Results 1 - 10 of about 64,100,000 for improve brain. (0.17 seconds)
Topics: All
Regions: World
Site Type: All
Government
University
Non-Profit
Commercial
Blogs
Video
Sound/voice
Brain, memory, improve your brain memory : improve your memory ...
Alzheimer's Memory Test : test your memory with our cognitive training, cognitive test. Alzheimer's memory test. Serie of tests to work on your memory and ...
www.happyneuron.com/GBHappyNeuron/memory/brain-improve-memory.html - 13k - Cached - Similar pages
Improve Brain Health And Life, Author Of 'Making A Good Brain ...
Clinical neuroscientist and psychiatrist Dr. Daniel Amen believes managing stress is an important component in achieving brain health.
www.cbsnews.com/stories/2005/10/10/earlyshow/health/main930771.shtml - 72k - Cached - Similar pages
How to Improve Reading Skills
Previous : Next : Overview : About Brain Dancing : Home Page. How to Improve Reading Skills. Reading groups of words at each glance ...
braindance.com/bdiread1.htm - 2k - Cached - Similar pages
New Scientist 11 steps to a better brain - Features
But there are a few ways to improve it. The first is to raise your arousal levels. The brain's attentional state is controlled by the neurotransmitters ...
www.newscientist.com/channel/being-human/mg18625011.900 - 75k - Cached - Similar pages
To Improve Brain Activity, Put Down the Remote -- Arehart-Treichel ...
To Improve Brain Activity, Put Down the Remote. Joan Arehart-Treichel. Even if
Hypnosis Downloads
Download a state of the art self hypnosis session now. Affiliate.
www.Netfreemail.com
Maintain Brain Health
Find Out How To Keep Your Brain Active & Improve Your Memory!
Internet

Topics: Games > Wii games

Regions: North America > USA

Site Types: All

Amazon.com: Big **Brain** Academy: **Wii** Degree: Video **Games**
The **Wii** sequel to Big **Brain** Academy for Nintendo DS includes three multiplayer modes for up to eight players. Players also can exchange student-record books **...**
www.amazon.com › Video Games › Wii › Puzzle - Cached - Similar

Big **Brain** Academy: **Wii** Degree - Wikipedia, the free encyclopedia
Big **Brain** Academy: **Wii** Degree includes a single player mode whereby the player uses a **brain** to effectively answer questions correctly. The **game** also **...**
en.wikipedia.org/wiki/Big_**Brain**_Academy:_**Wii**_Degree - Cached

Big **Brain** Academy: **Wii** Degree for Nintendo **Wii** | EBGames
EBGames: Buy Big **Brain** Academy: **Wii** Degree, Nintendo of America, Nintendo **Wii**, Find release dates, customer reviews, previews and screenshots.
www.eb**games**.com/Catalog/ProductDetails.aspx?product_id... - Cached - Similar

Big **Brain** Academy: **Wii** Degree Video **Game** | Reviews, Trailers **...**
View Big **Brain** Academy: **Wii** Degree video **game** trailers, exclusive features, and online reviews. View exclusive interviews, actual Big **Brain** Academy: **Wii ...**
www.**game**trailers.com/**game**/big-**brain**-academy-**wii**-degree/3424 - Cached

Integrating search and browsing together

- Search alone: filters left at top level (“all pages”)
- Browsing alone: search keyword left blank
- Search within browse; browse then search

Multi-label, Multi-instance Hierarchical Classification with Active Learning

Outline

- Introduction
- Active Learning with Generalized Queries (AGQ)
- Experimental Comparison
- Discussions
- Recent Advances
- Conclusions and Future Works

Conclusions

- Active Learning can be very powerful
- Many different ways to be “active”
  - Generalized queries reduces queries significantly
  - Different assumptions on oracle
    - Cost-sensitive, ambiguous, with explanation, …
  - Anomaly-based AL (vs boundary-based AL)
  - … …
- Widespread applications (search engines, text, …)
- …

Current & Future Works

- Theoretical Guarantee
  - Some results already
- Active learning in human and machines
- Real-World Applications

Thanks for Active Listening!