DIANA:
Data Intensive ANAlyzer

Seung-won Hwang
CSE, POSTECH

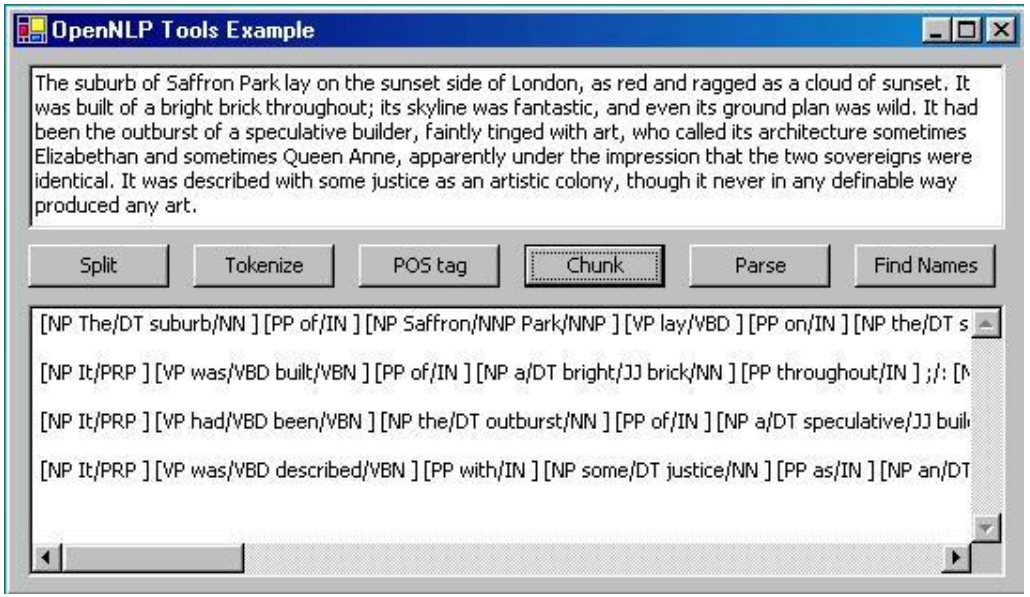
Outline

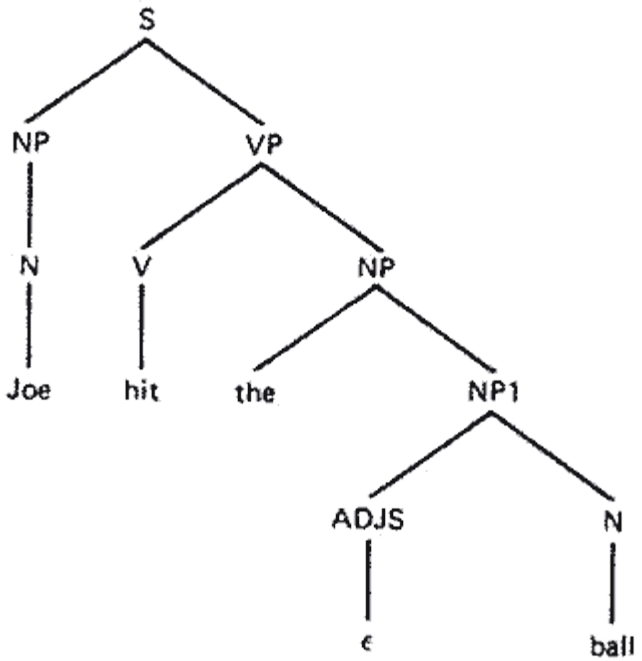
- "Data Intensive"?
- My research context
- Call for (synergetic) collaborations!

- Mantra of the talk
  - Ignorance is bliss
  - Systematic creativity: Undiscovered public knowledge

    Swanson suggested ... that novel information might be unearthed by systematically studying seemingly unrelated and non-interactive research literatures, which he called "complementary but disjoint"

Motivation: NLP meets IR

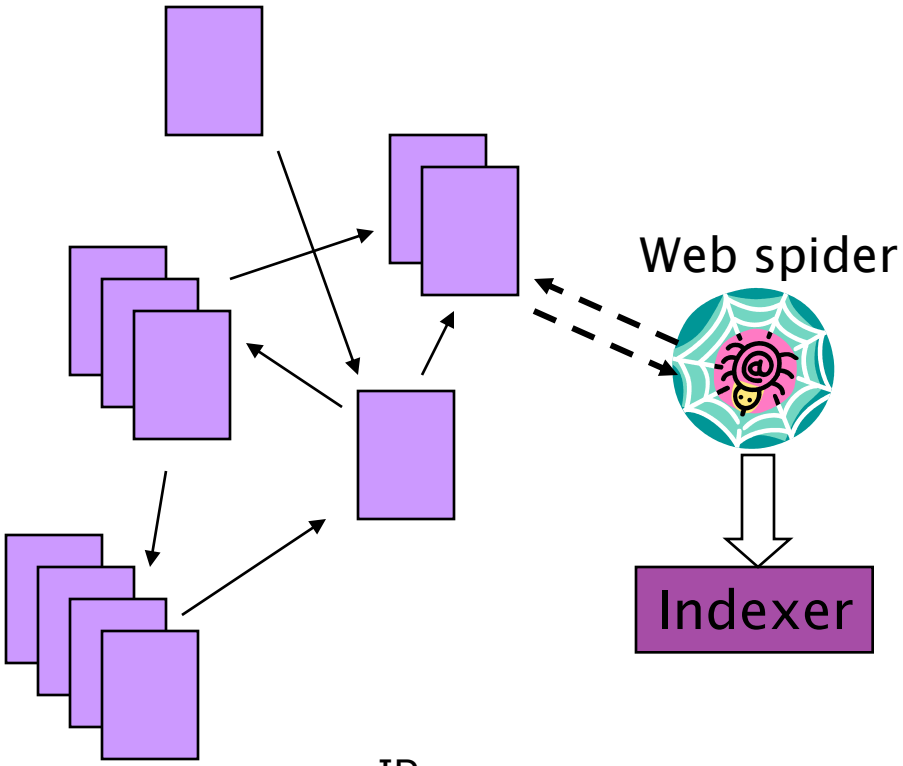




Classic NLP
- generative / model-driven
- ambitious vision: AI

NLP vision today?

IR story

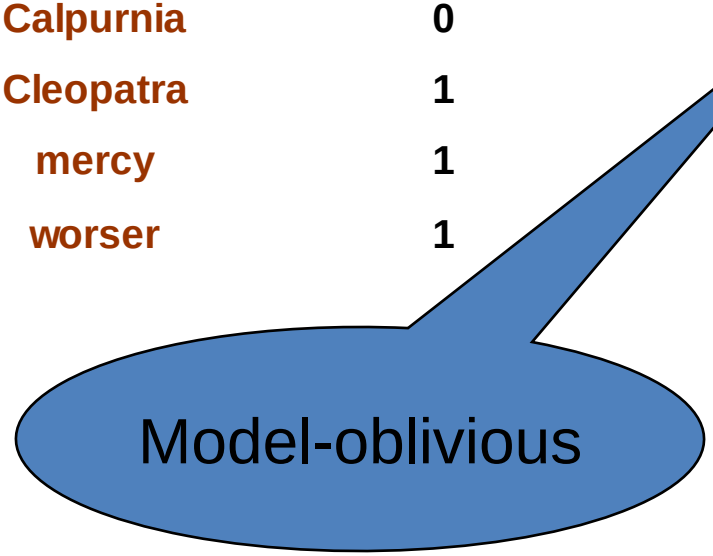


User

IR
- data-intensive / statistical
- down-to-earth vision: caching Web info to local disk + provide efficient access

IR="Ignorant" Retrieval?

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0



Source: http://www.stanford.edu/class/cs276/cs276-2006-syllabus.html

But “ignorance is bliss”

- Brutus: 1 1 0 1 0 0
- Caesar: 1 1 0 1 1 1
- Brutus AND Caesar: 1 1 0 1 0 0

- 3 out of 7? Ranking?
- Immediate answer? Indexing?

→ My research interests

NLP+IR Synergy Today

- Web-scale NLP
  - Augmenting NLP engine to exploit Web-scale corpus knowledge
  - Enriching IR view with NLP knowledge
  - Synergy showcases

Synergy Showcase #1 : Learning Language

- Generative

- Data intensive

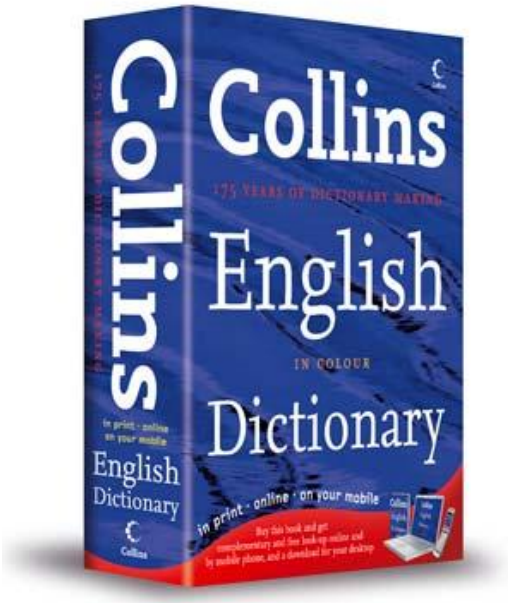
Showcase #2: Translation

- Generative

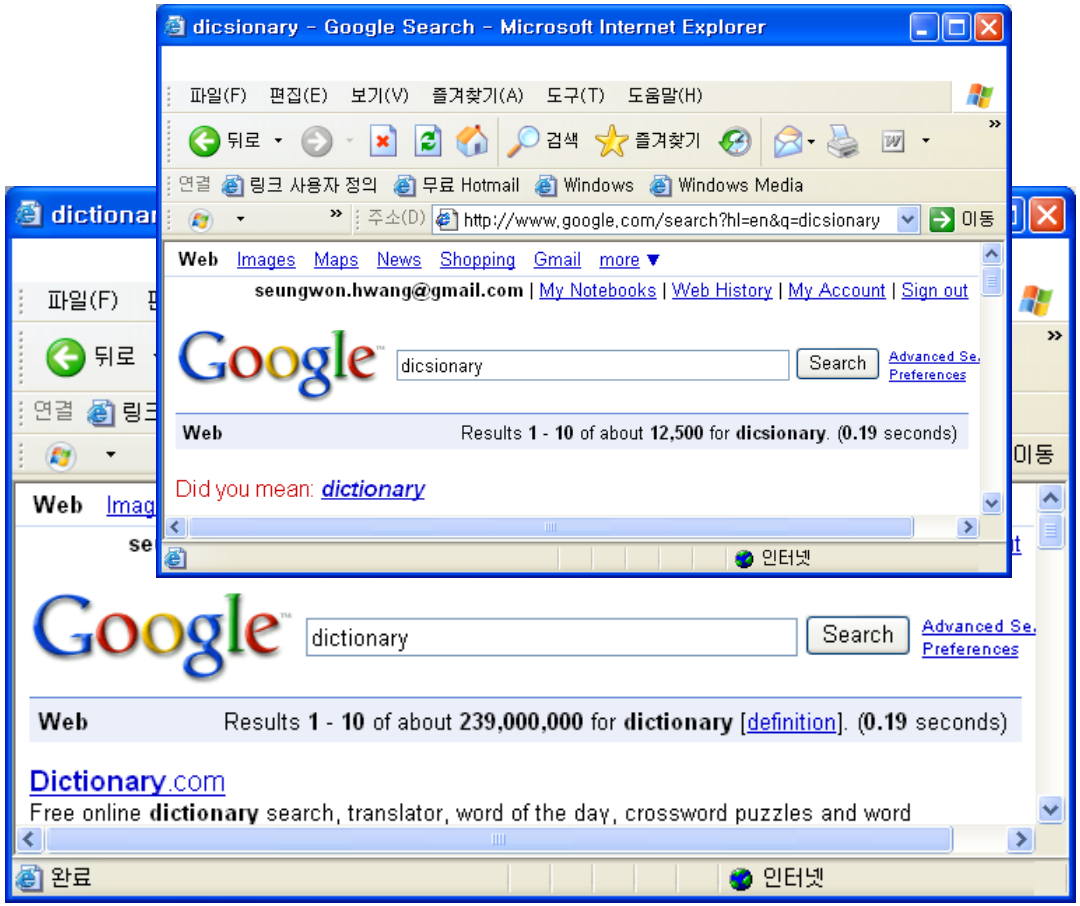
- Data intensive

Showcase #3: Writing Help

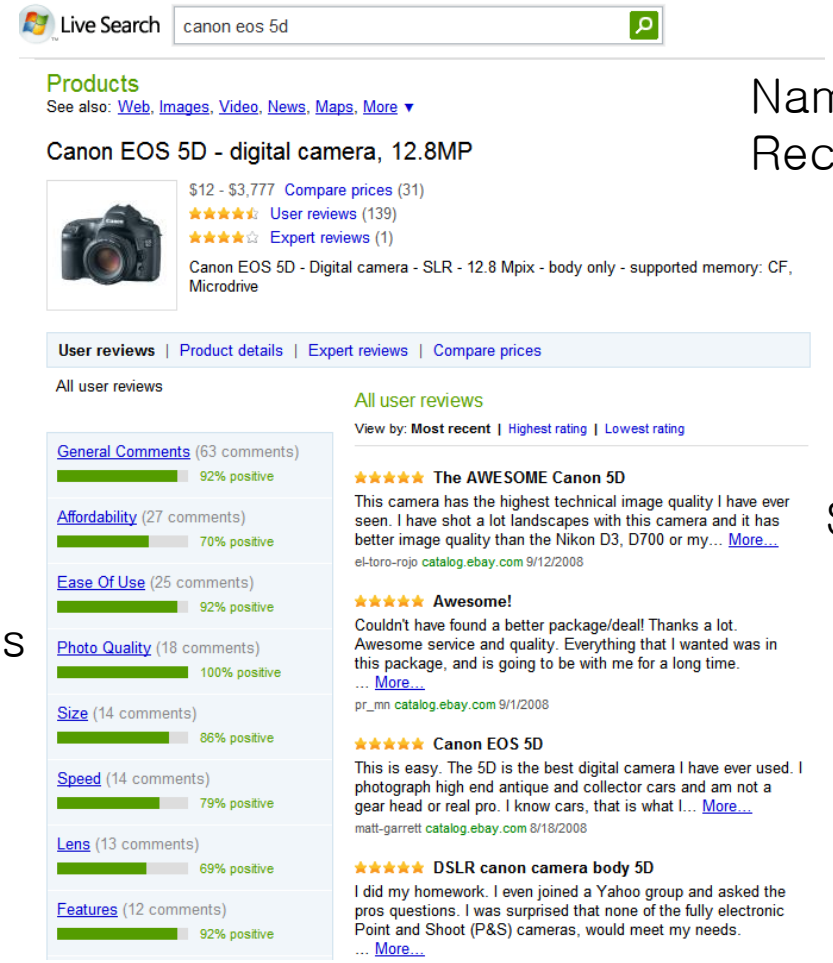
- Generative



- Data intensive



Showcase #4: Search Intent?



Named Entity Recognition

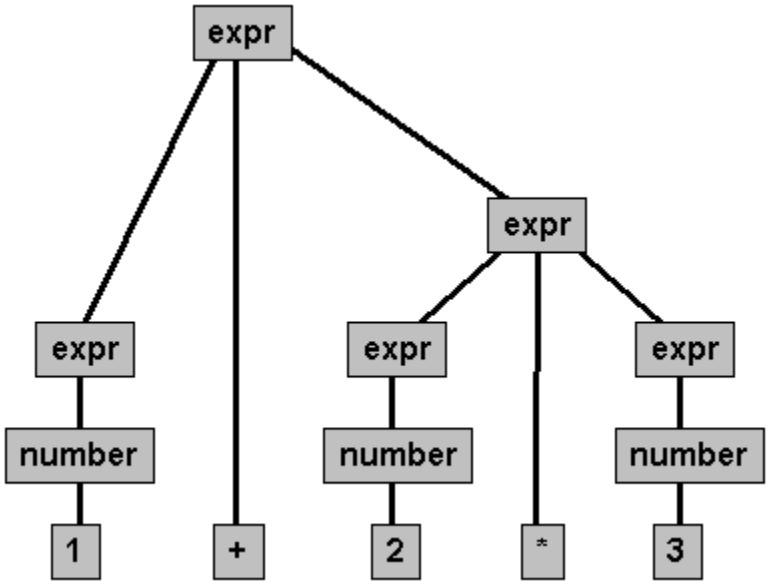
Summarization

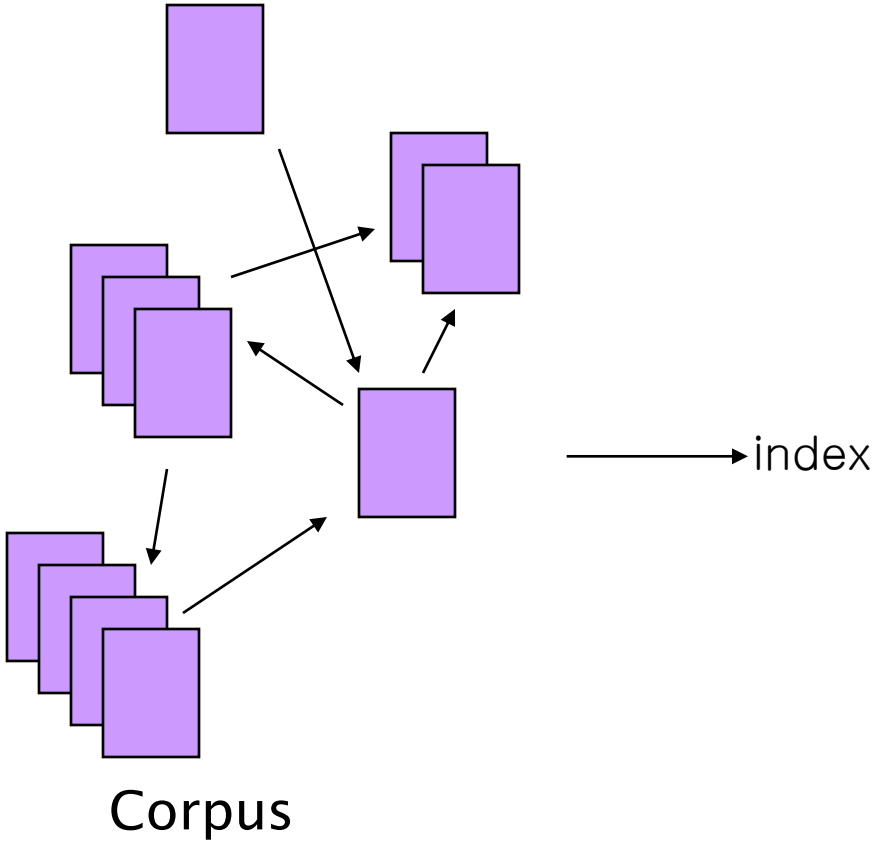
Sentiment analysis

Claim

Data intensive approaches can provide complementary and scalable solutions

How is this claim relevant to US?

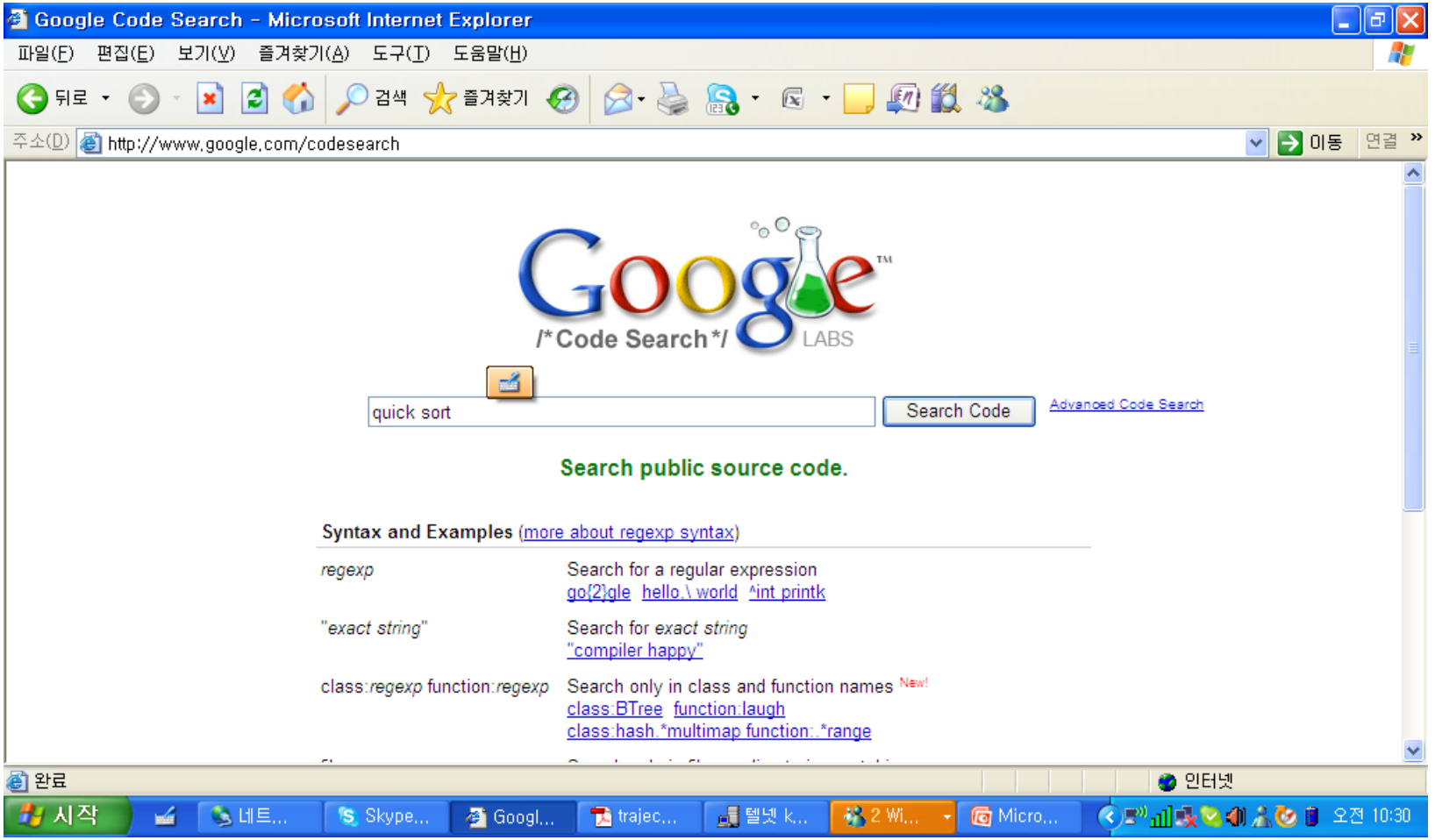




Opportunities

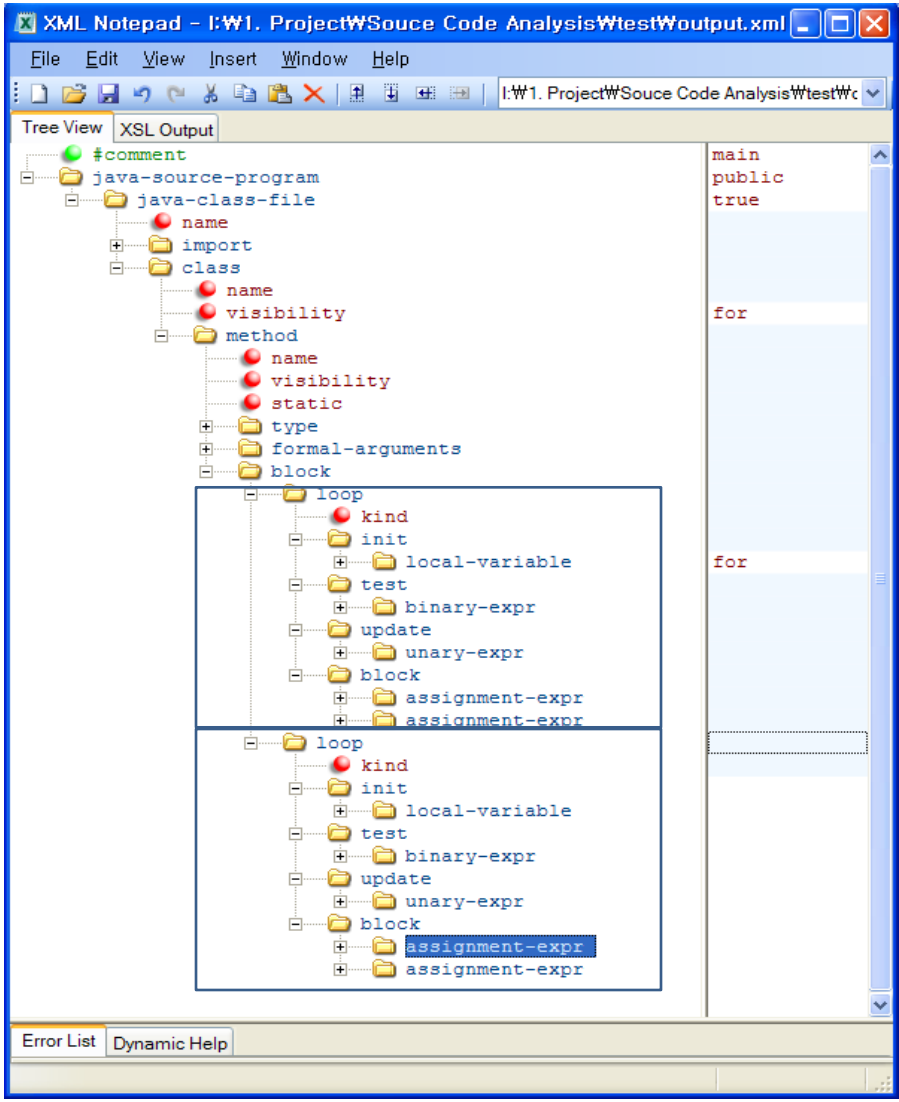
- **Corpus** as a "byproduct" of software products
  - Code corpus (search engine)
  - Bug fix corpus (development platforms)
  - Crash / bug reports (various products)
- Code search, debugging guidance, teaching coding by examples, …

Huge room to “advance the state-of-the-art”



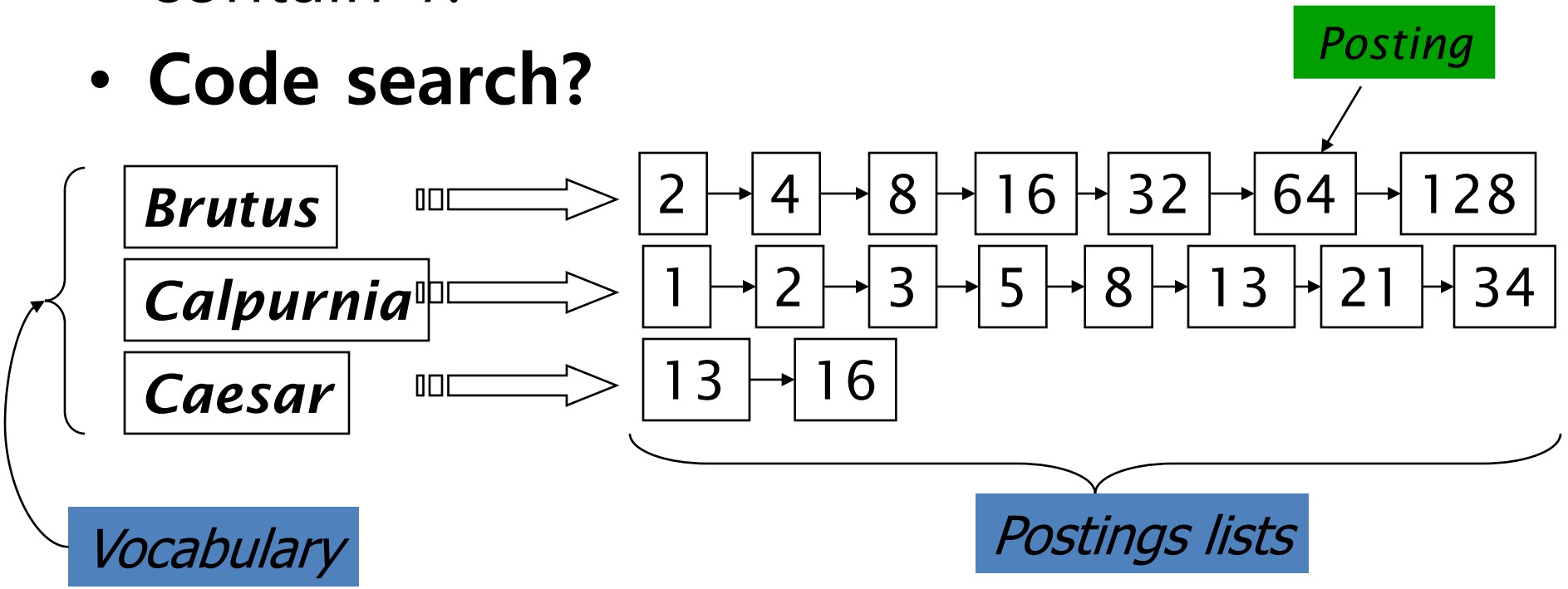
Instant Structure Match

- Sample code search
- Code review
- Education



Indexing?

- **Search engine:** For each term *T*, we must store a list of all documents that contain *T*.
- **Code search?**

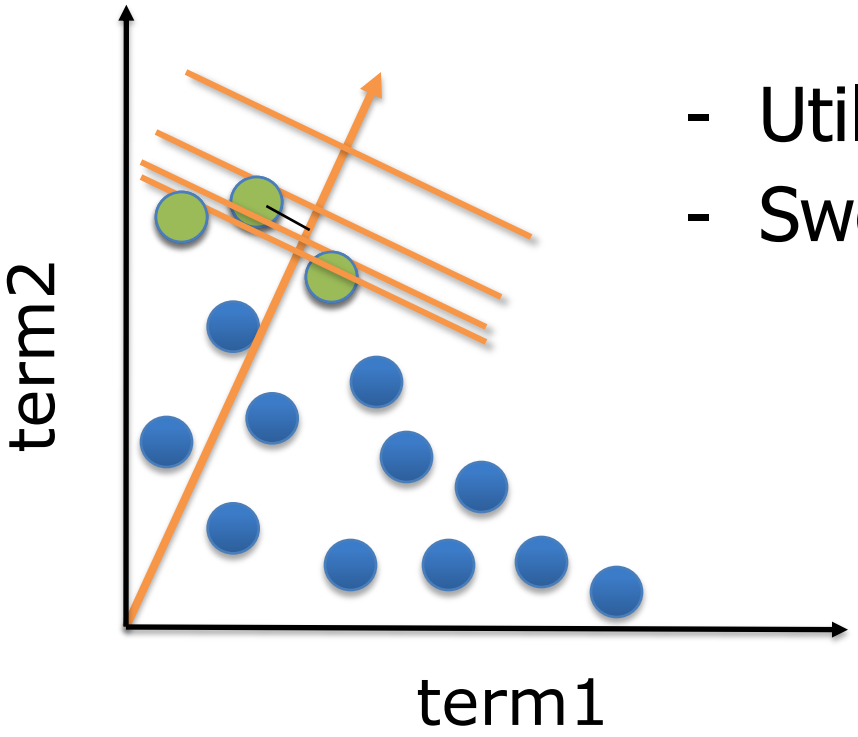


We need to define

- Ranking function
- Ranking algorithm
- Structural vocabulary

My Research Context: Ranking in DB

- Solution for intuitive exploration
  - Return best matches → No empty results
  - Deliver only k results → No flooding



- Utility relies on scoring function
- Sweepline collects best results

Pioneering Algorithm

- **Fagin's Algorithm [Fag96] (PODS)**

Sorted list on Mileage

#	Mileage
1	0.9
2	0.8
3	0.7
	.....
4	0.6

Sorted list on Age

#	Age
4	0.9
1	0.8
2	0.7
	.....
3	0.2

K = 2

F(id)=2*mileage+age (monotonic)

#	Mileage	Age	F(id)
1	0.9	0.8	2.6
4	0.6	0.9	2.1
2	0.8	0.7	2.3
3	0.7	0.2	1.6

My Research Context: Ranking in DB

Sorted Access	Random Access		
	r =1 (*cheap*)	*r = h* (*expensive*)	*r = ∞* (*impossible*)
s =1 (*cheap*)			
s = h (*expensive*)			
s = ∞ (*impossible*)			

Unified Top-k Optimization [ICDE05,ITKDE07]

"Expensive" Ranking Conditions [TODS07]

- Goal: Perform only necessary random accesses (or, "probes")
- Necessary probes
  - A probe is necessary if top-k answers cannot be determined by any algorithm without it, regardless of the outcomes of other probes.
- Optimal algorithm
  - An algorithm is probe-optimal if it performs only the necessary probes.

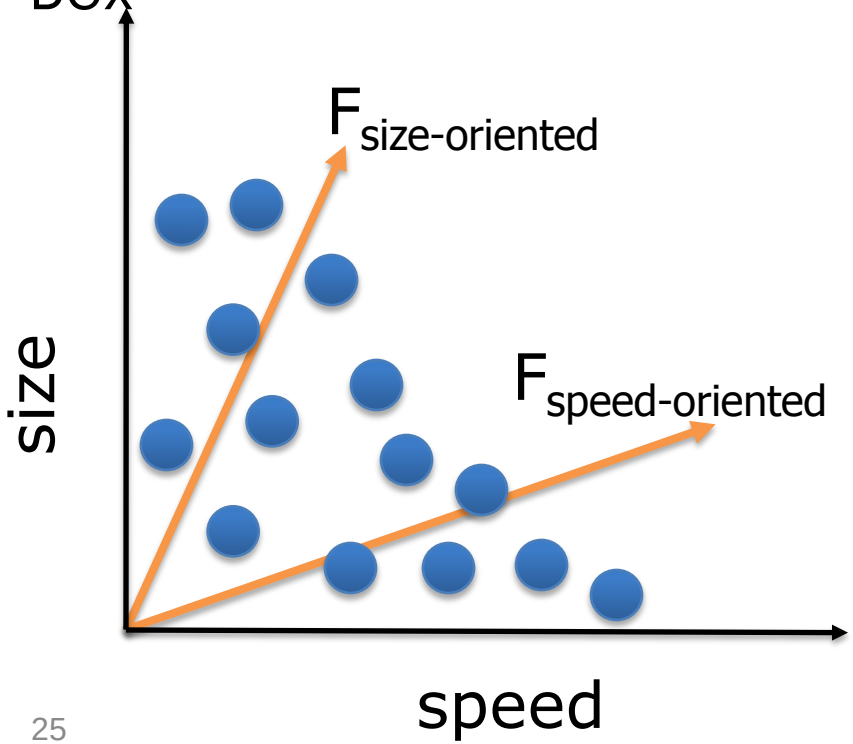
Pioneering Algorithms

- **k=1, F=min(x,p1,p2);** suppose H=(p1,p2)

OID	x	p1	p2	F=min(x,p1,p2)	
a	0.9	?	1	0.9	← top 1
b	0.8	Maybe Not!		≤ 0.8	
c	0.7	?	1	0.7	
d	0.6	?	1	0.6	
e	0.5	?	1	0.5	

Skylines: Needs for "Portfolio" Approach

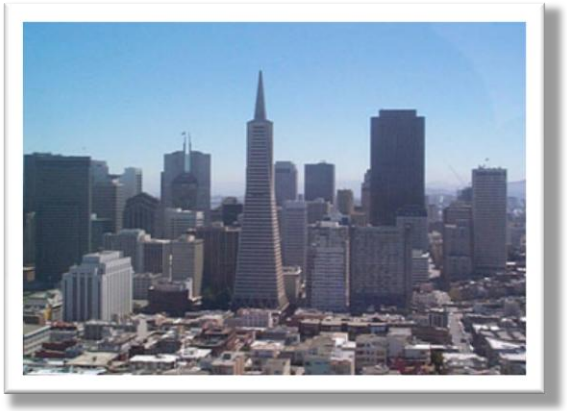
- Top-k retrieval requires accuracy of scoring function F
→ **risk** for inaccuracy
- All models are wrong; only some are useful – George Box



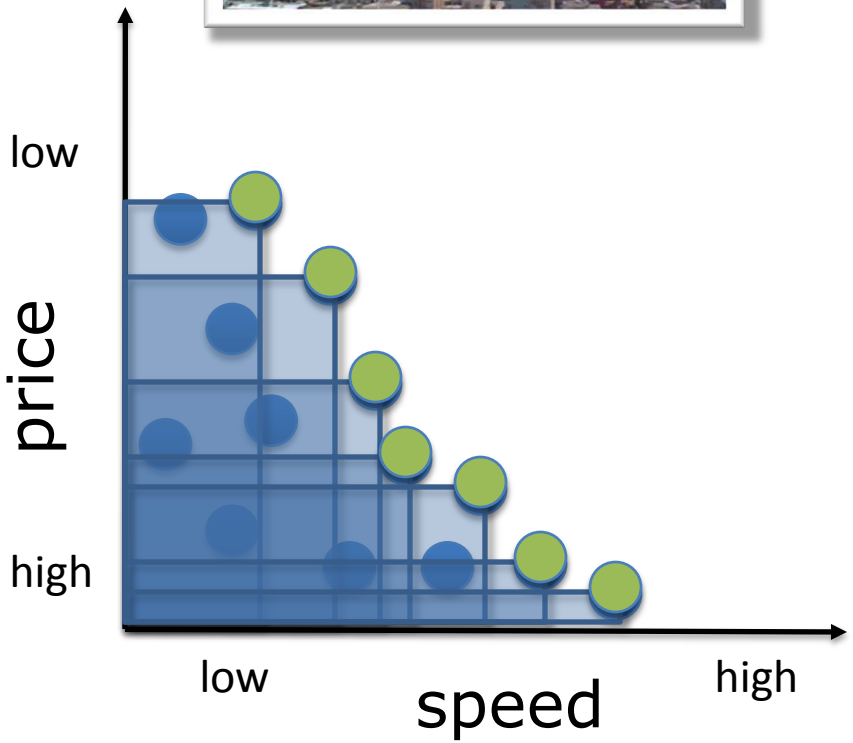
"Score = 0.6 * size
+ 0.4 * speed" ?!

Intuitive "Diversification"

- Skylines or Pareto-optimality
  - Find "diversifying" frontiers



- Advantages
  - Intuitive querying
- Disadvantages
  - Curse of dimensionality



Ranking+Skylines Showcase: Web Mining Driven Visual "Portfolio" Example

Live Search canon eos 5d

Products

See also: Web, Images, Video, News, Maps, More ▾

Canon EOS 5D - digital camera, 12.8MP

$12 - $3,777 Compare prices (31)
User reviews (139)
Expert reviews (1)

Canon EOS 5D - Digital camera - SLR - 12.8 Mpix - body only - supported memory: CF, Microdrive

User reviews | Product details | Expert reviews | Compare prices

All user reviews

General Comments (63 comments) 92% positive
Affordability (27 comments) 70% positive
Ease Of Use (25 comments) 92% positive
Photo Quality (18 comments) 100% positive
Size (14 comments) 86% positive
Speed (14 comments) 79% positive
Lens (13 comments) 69% positive
Features (12 comments) 92% positive

All user reviews

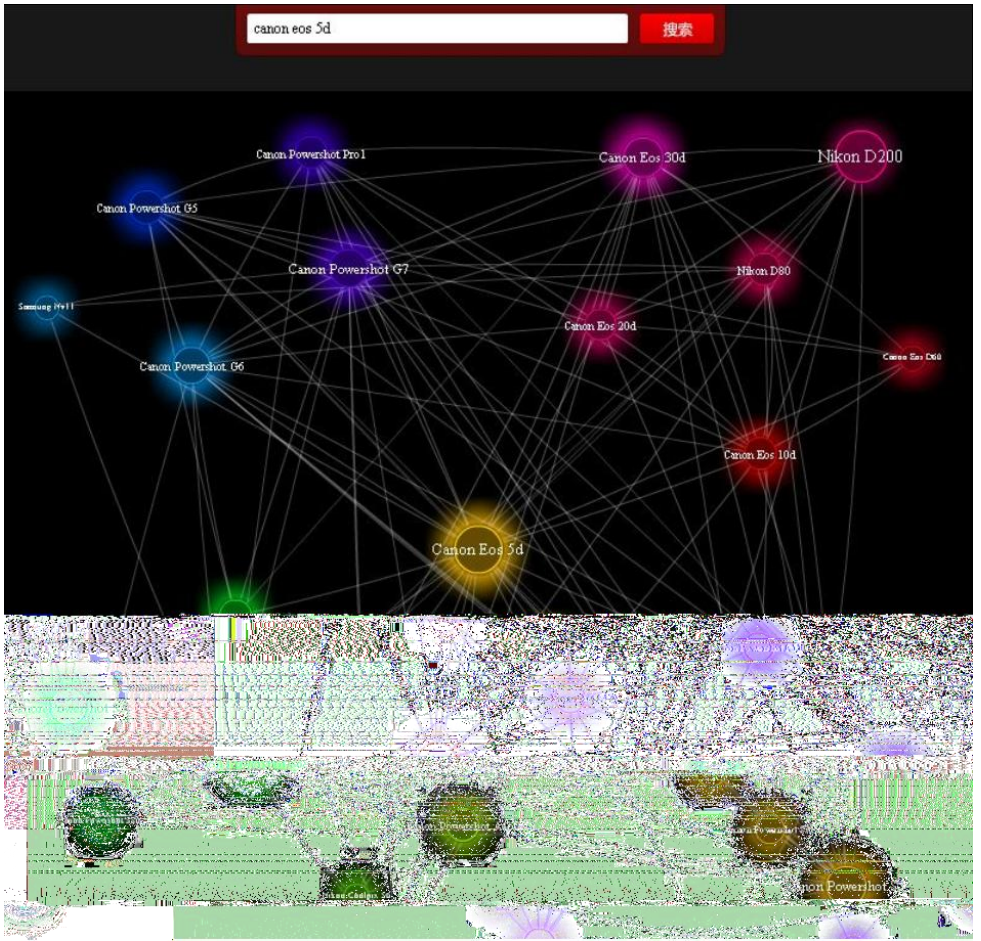
View by: **Most recent** | Highest rating | Lowest rating

★★★★★ **The AWESOME Canon 5D**
This camera has the highest technical image quality I have ever seen. I have shot a lot landscapes with this camera and it has better image quality than the Nikon D3, D700 or my… More…
el-toro-rojo catalog.ebay.com 9/12/2008

★★★★★ **Awesome!**
Couldn't have found a better package/deal! Thanks a lot. Awesome service and quality. Everything that I wanted was in this package, and is going to be with me for a long time. … More…
pr_mn catalog.ebay.com 9/1/2008

★★★★★ **Canon EOS 5D**
This is easy. The 5D is the best digital camera I have ever used. I photograph high end antique and collector cars and am not a gear head or real pro. I know cars, that is what I… More…
matt-garrett catalog.ebay.com 8/18/2008

★★★★★ **DSLR canon camera body 5D**
I did my homework. I even joined a Yahoo group and asked the pros questions. I was surprised that none of the fully electronic Point and Shoot (P&S) cameras, would meet my needs. … More…



Data Intensive Backends

- Relationship
  - Magnitude (Distance): Strong/Weak
  - Types (Color) : Sisters (D50,D70), Competitor (Canon), Accessory (Bag, Memory)..
- Mining
  - Magnitude (mass collaborative; voted by Web content creators)
    - Text co-occurrence (crawler)
  - Types (Feature-based)
    - Feature space (data extractor)

Short-term plans

- Code search feasibility study:
  - Using code-similarity metrics
- Structural indexing
  - Beta release
- Call-for-help:
  - Code is much more than a text w/ syntactic structure
    - E.g., Runtime behaviors (“expensive”. Use sparingly)
    - Enlighten me!

Thank You!

http://www.postech.edu/~swhwang