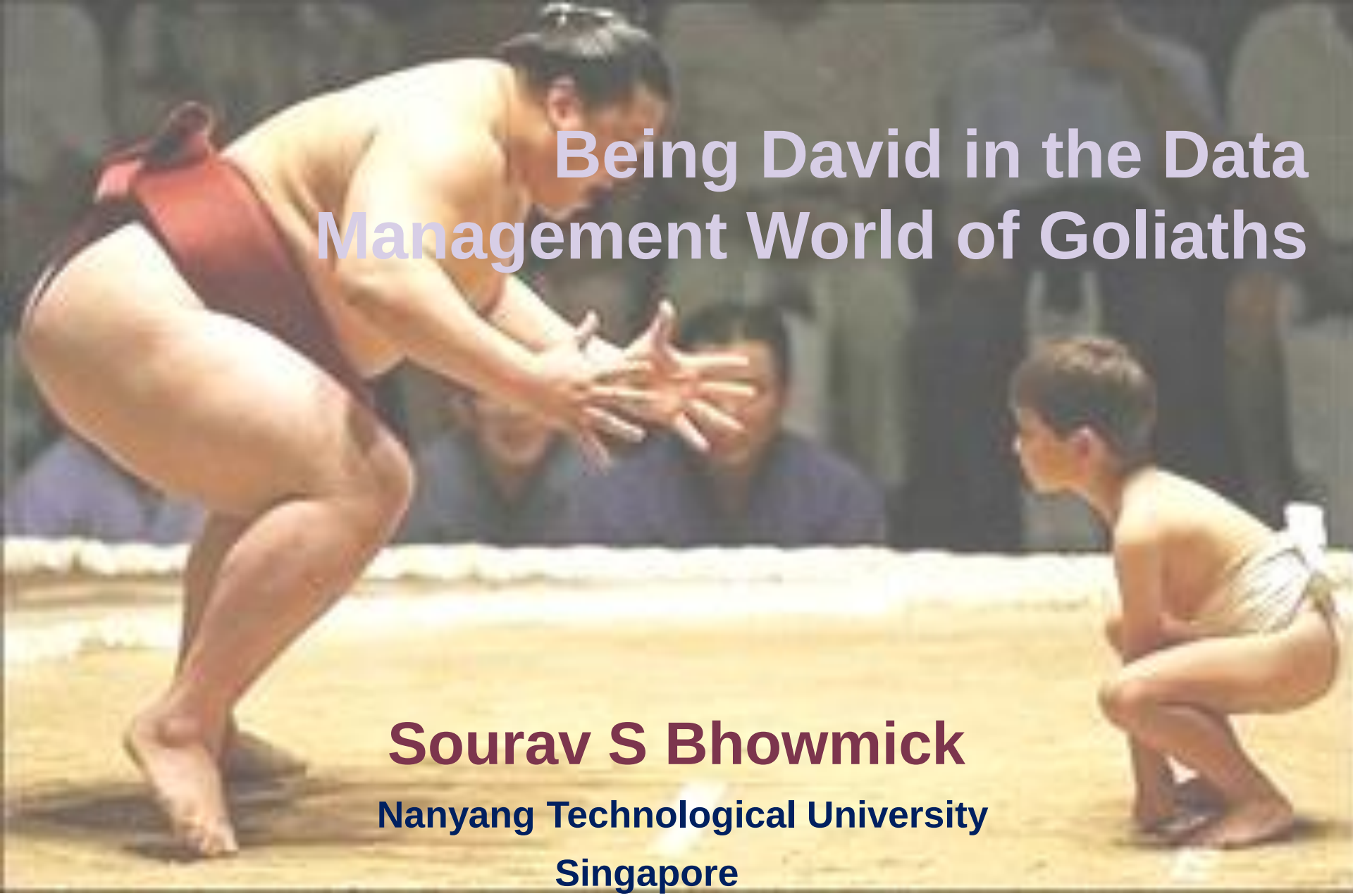


Being David in the Data Management World of Goliaths

Sourav S Bhowmick

Nanyang Technological University

Singapore

iDB 2011, Kyoto, Japan

NANYANG TECHNOLOGICAL UNIVERSITY

Disclaimer

The opinions expressed in this talk are solely of the author and does not necessarily reflect the opinions or beliefs of the community.

NANYANG TECHNOLOGICAL UNIVERSITY

David & Goliath – The Story

Hebrew Bible

- Saul and Israelites vs Philistines
- Goliath of Gath
  - A giant Philistine warrior who challenges the Israelites to send out a champion of their own to decide the outcome in single combat
- David, younger brother of Saul, accepts the challenge
- David went to the battle only with a sling and five stones!
- Goliath in armors and shield.
- David hits Goliath's forehead->Goliath collapses -> David cut off his head with Goliath's sword

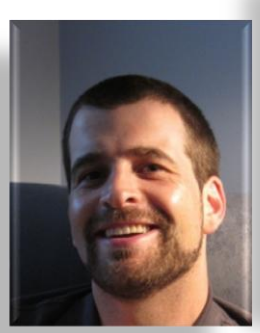
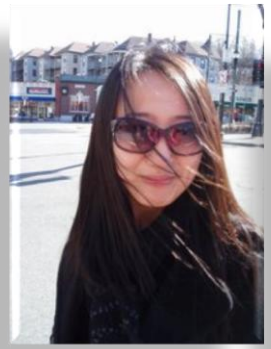
Davids & Goliaths in DB World

Davids

- Not a descendant of "star" faculty
- Not affiliated to DB-strong institutes
- Not well-connected to frequent top-tier community
- Only has a sling (laptop) and stones (ideas)

Goliaths

- Frequently publishes in top-tier conferences
- Descendant of "Goliaths"
- Affiliated to DB-strong institutes
- Strong social network

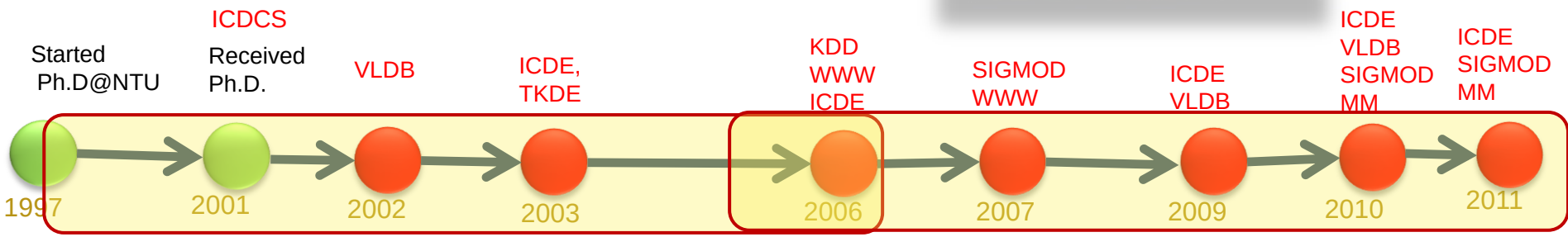





NANYANG TECHNOLOGICAL UNIVERSITY

Being David – Sharing My Experience

- NTU wasn't DB strong
- DB in Singapore = NUS DB
- My advisor was not a descendant of "star"
- Disconnected from visible network



Sanjay Madria (MUST)
Mukesh Mohania (IBM)

Curtis Dyreson (Utah)
Aixin Sun (NTU)
C F Dewey (MIT)

NANYANG TECHNOLOGICAL UNIVERSITY

Problem Statement

Top-Tier Acceptance Maximization Problem

Let there be k slots in a conference C where k << N . Let R be the set of reviewers in C. Let p be the paper of David D. The top-tier acceptance maximization problem is to find an algorithm that maximizes $f_C(p_k) = Pr(p_k | R)$.

Assumptions

- There are G Goliaths in the community
- M slots are taken by G' $\subset$ G

NANYANG TECHNOLOGICAL UNIVERSITY

Visual Representation

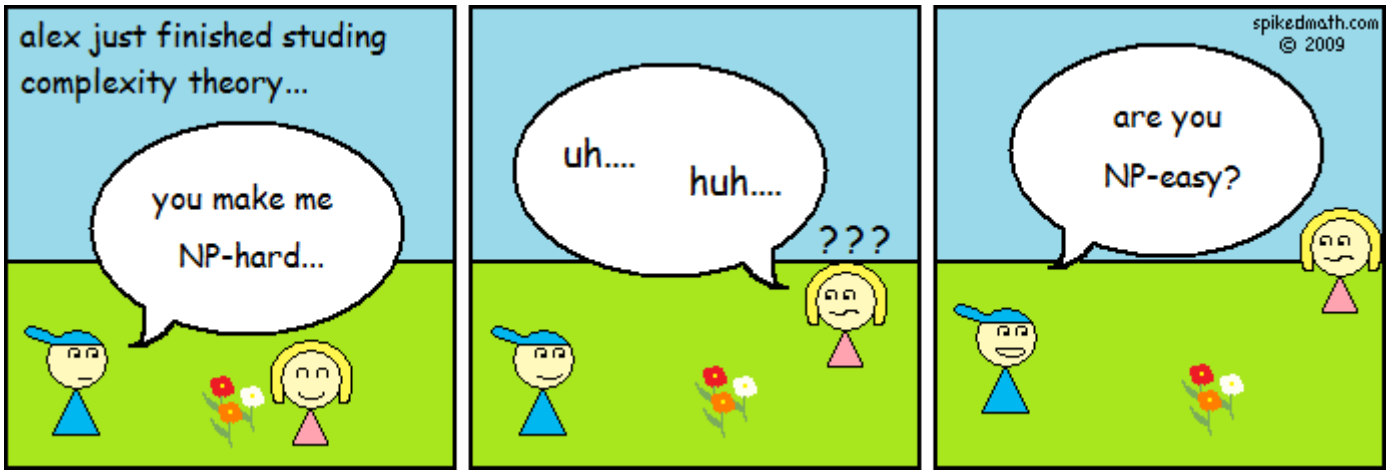
Characteristics

Theorem 1

The top-tier acceptance maximization problem is NP-Hard.

Solution

- Heuristic-based algorithm



NANYANG TECHNOLOGICAL UNIVERSITY

Challenges

Hard Issues

- Real-world reviewers
- No Goliath as co-author
- Affiliation of the paper is not DB-strong

Ideal vs Real Reviewers

Real-world Reviewers

•May not be aware of all works in the field
• May not read every sentence carefully
• May be biased to authors weight and social network
• May have a biased view of a solution

Ideal Reviewers

• Expert in the field
•Reads every line of the paper carefully
• Is not influenced by names of authors/affiliations
• Have solid vision and open mind
• Can see through the proposed solution
• Not biased to any social network

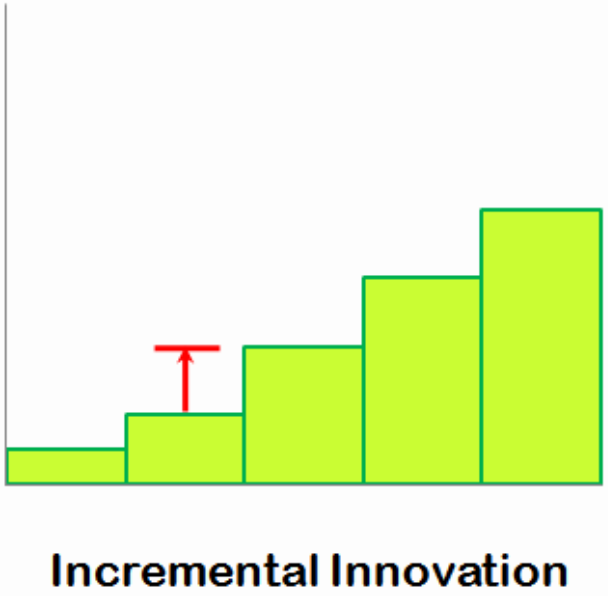
NANYANG TECHNOLOGICAL UNIVERSITY

Curse of Delta

"The work seems to be delta research…

What is delta?

- Different reviewers have different measure of delta
- Most papers are based on prior work
- Twig join algorithm lead to many enhancements



Solution

- Not easy as you do not know who will review
- Safe solution: propose a new line of research

Lack of Significant Impact

"The impact of the work is low…..

How do you know?

- Crystal ball comment
- DB community has been wrong in prediction many times
  - Relational model, B trees, Lorel, Page Rank, ARM….
  - Best Paper ≠ 10 Years Best Paper

Solution

- Intractable problem!

NANYANG TECHNOLOGICAL UNIVERSITY

Not “Skyline” Performance

“Performance gain is not significant enough…

Delta performance

- Can-do-better papers
- Improvement by a factor of 3 or less
- Weakens the proposed technique

Solution

- Propose technique that can bring in at least an order of magnitude performance improvement

Not Enough Experiments!

"Not enough experiments…

Performance study

- How much is "enough"?
- Controlled by page limit

Solution

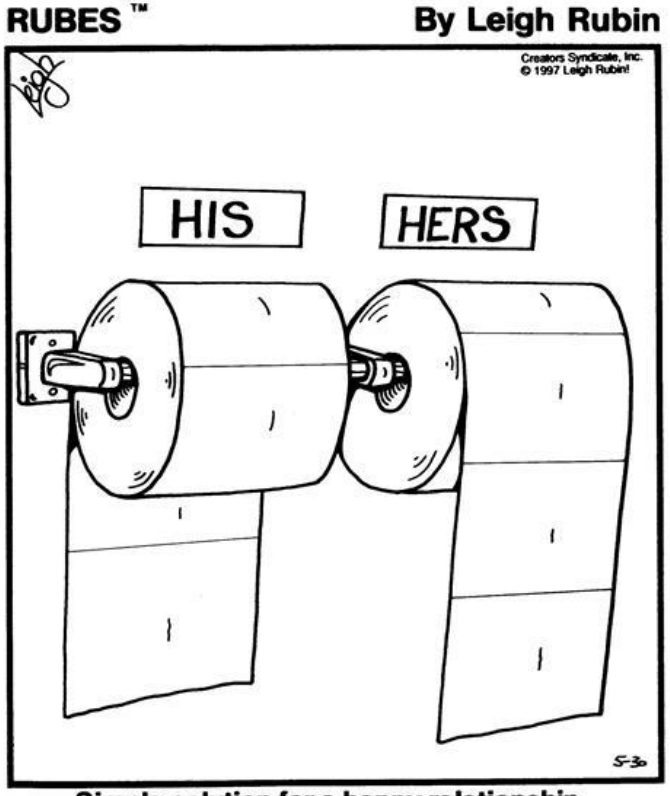
- Report exhaustive experimental study (~3 pages)
- Even report results that are obvious
- Highlight the strong results

Too Simple Solution!

"Solution is simple…

Simplicity

- Simplicity is not a crime
- Techniques that can stand test-of-time generally have simple solution
- Interpretation -> No complex formula/maths, theorems are visible ☺



Simple solution for a happy relationship

Solution

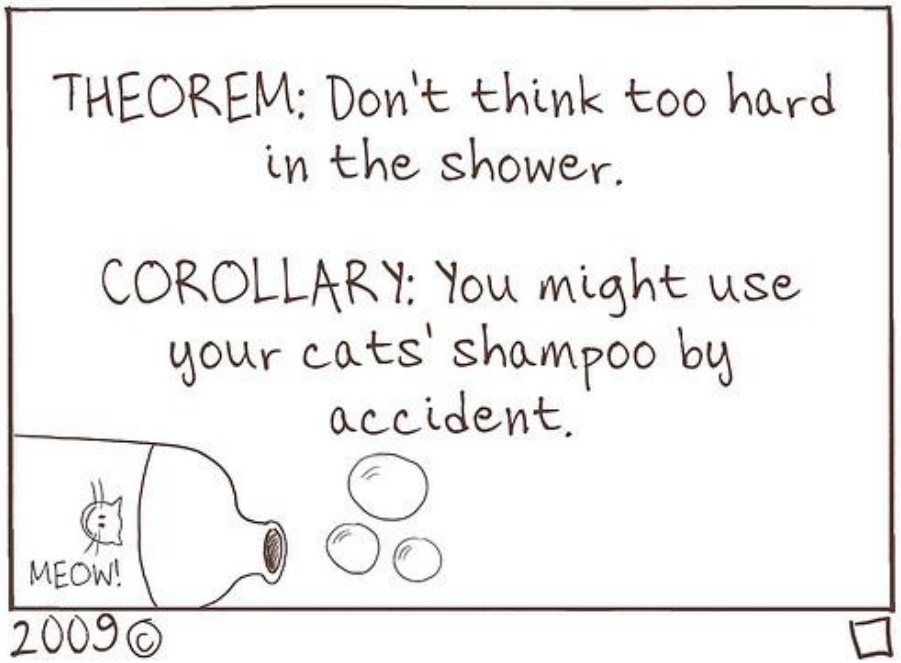
- Don't solve this problem
- Hope for a better reviewer next time ☺

Not Enough Theorems!

"The paper don't have enough theorems …

Theorem problem

- Good research ≠ No. of theorems
- Generating large number of insignificant theorems obfuscate the main idea



Solution

- Don't solve this problem
- Hope for a better reviewer next time ☺

NANYANG TECHNOLOGICAL UNIVERSITY

Ok..But What About Y?

"The paper doesn't discuss how it can handle …

Interpretation

- Often Y = future work
- Comments made without considering page limit
- Often indicates that the reviewer cannot find any other issue to criticize

Solution

- Intractable problem!

Summary

Key Issues

- Curse of delta ✓
- Lack of significant impact ✗
- Not “skyline” performance ✓
- Not enough experiments ✓
- Too simple solution ✗
- Not enough theorems ✗
- The Missing Y ✗

David's Paper

Rank 1 strategy

New line of research!

Pros

- No competition!
- Curse of delta
- Lack of novelty
- Too simple soln
- Not "skyline" perf.

Cons

- Its hard!
- Difficult to get accepted

Rank 2 strategy

Prove a famous work wrong!

Pros

- Curse of delta
- Lack of impact
- Too simple soln
- Performance

Cons

- Its hard too!
- Can be controversial

Rank 3 strategy

Can-do-better paper

Pros

- Relatively easy
- Pro-performance (at least 1-2 orders of mag)

Cons

- Curse of delta
- Novelty
- Perf-not-enough

Secret of Great Papers

Secrets to tackle Real-world Reviewers

- Story telling (Sell your story in 3 pages)
- Understandable without reading the entire paper
- Make it readable even if certain sections are skipped
- Realistic assumptions & Real problems
- Elegant solution
- Solid experiments (even obvious ones!)
- Know your area well! (Don't be oblivious to related research)

Heartware of Top-tier Research

Heartware necessary for success

- Be driven
- Never say die
- Be creative (Don’t always believe what you read)
- Think deep vs Think fast
- Work hard
- Be a one-man army (David don’t have choice)

Post-Relational Reality

Impact of DB research

- How to measure?
  - Publications in top venues
  - Citations
  - Who uses it? How it enhances human lives?

Then

•Data is generated in companies
- Resides in companies
- Used by companies
- DB-literate users

Now

•Data is generated by everyone
- Resides everywhere
- Used by everyone
- Non-DB literate users

NANYANG TECHNOLOGICAL UNIVERSITY

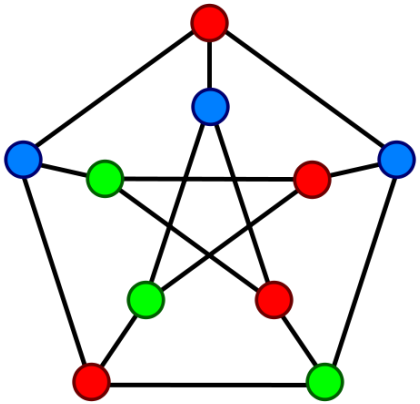
What Areas to Focus?

Data streams

Parallel & Distributed DB

Cloud data management

semistructured
data management



graph
data management

DB+ IR

mobile data management

Object-Oriented Model

Object 1: Maintenance Report

Date	
Activity Code	
Route No.	
Daily Production	
Equipment Hours	
Labor Hours	

Object 1 Instance

01-12-01
24
I-95
2.5
6.0
6.0

Object 2: Maintenance Activity

Activity Code	
Activity Name	
Production Unit	
Average Daily Production Rate	

OO DB

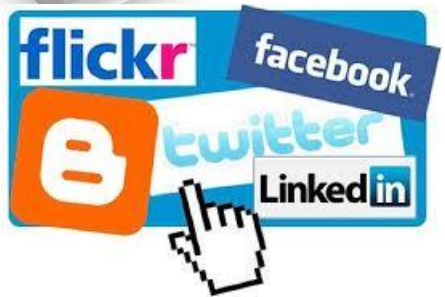
Key	Product ID	Price ($)	Prob.
a_1	a	120	0.7
a_2	a	80	0.3
b_1	b	110	0.6
b_2	b	90	0.4
c_1	c	140	0.5
c_2	c	110	0.3
c_3	c	100	0.2
d_1	d	10	1

Probabilistic DB

NANYANG TECHNOLOGICAL UNIVERSITY

Biggest Impact?

Data are generated and consumed by non-DB experts users!



DB Community's Love Affair

SPEED, SPEED, SPEED
SCALE, SCALE, SCALE

Hard Reality

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rss: <http://purl.org/rss/1.0/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>

SELECT ?title ?known_name ?link

FROM http://planetrdf.com/index.rdf
FROM NAMED <phil-foaf.rdf>
WHERE {
        GRAPH <phil-foaf.rdf> {
        ?me foaf:name "Phil McCarthy" .
        ?me foaf:knows ?known_person .
        ?known_person foaf:name ?known_name .
        } .
        ?item dc:creator ?known_name .
        ?item rss:title ?title .
        ?item rss:link ?link .
        ?item dc:date ?date.
    }
ORDER BY DESC[?date] LIMIT 10
```

NANYANG TECHNOLOGICAL UNIVERSITY

Hard Reality!

We know the problem

> Thirty years of research on query languages can be summarized by: we have moved from SQL to XQuery. At best we have moved from one declarative language to a second declarative language with roughly the same level of expressiveness. It has been well documented that end users will not learn SQL; rather SQL is notation for professional programmers.

The Lowell Database Research Self-Assessment, Communication of the ACM (May 2005)

Usability

> If the user can't use it, it doesn't work.

Susan Dray, Distinguished Engineer of ACM

NANYANG TECHNOLOGICAL UNIVERSITY

My Favorite Problems

Unifying Theme

Data management for the people, by the people

Future Healthcare

- Data-driven drug targets discovery
- Functional visualization of molecular networks
- ACM BCB 11

Social media

- Analyzing social networks
- Improving social image search
- CIKM[09,10,11], MM[10,11], VLDB [10]

Usable DBs

- DB as iPad
- Making XML & graph DBs usable
- ICDE [06,09,10], VLDB[10], SIGMOD [10,11]

NANYANG TECHNOLOGICAL UNIVERSITY

Towards Usable DBs

Querying without XQuery

- User-friendly DB-ignorant GUI (towards iPAD for DB)
- Structure and query language independence
- ICDE 2010, VLDB 2010
- XMORPH

DB meets HCI

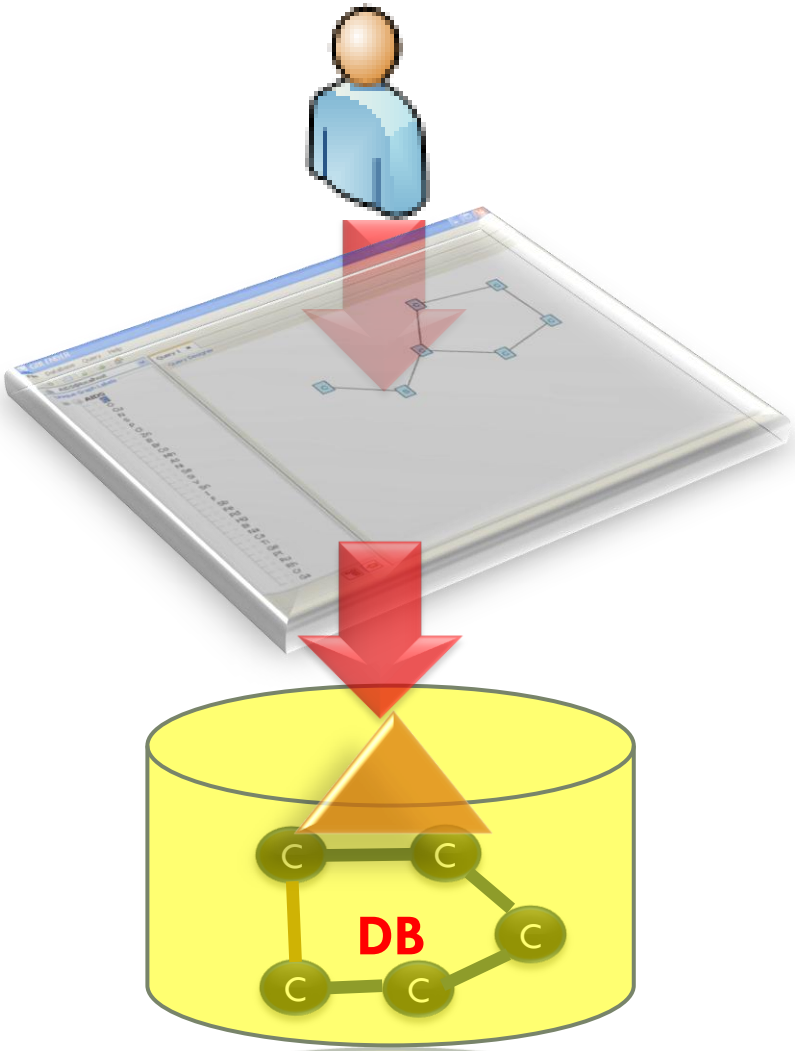
- Blending visual query formulation and query processing
- Rank 1 strategy (we are the first!)
- ICDE [06,09], SIGMOD [10,11]
- XBLEND, GBLENDER

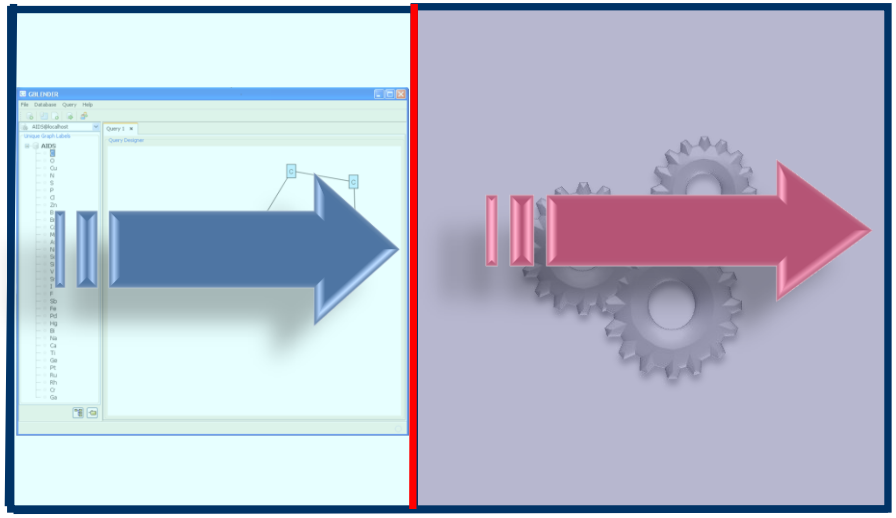
"A picture is worth a thousand words. An interface is worth a thousand pictures

Ben Shneiderman, 2003

NANYANG TECHNOLOGICAL UNIVERSITY

Classical Visual Querying Paradigm





Query formulation Query processing

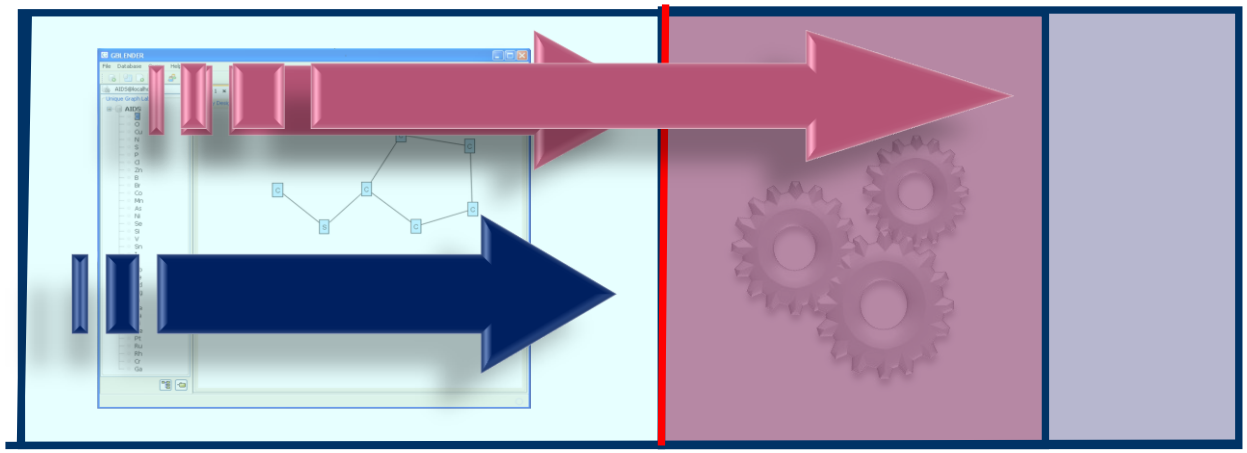
time

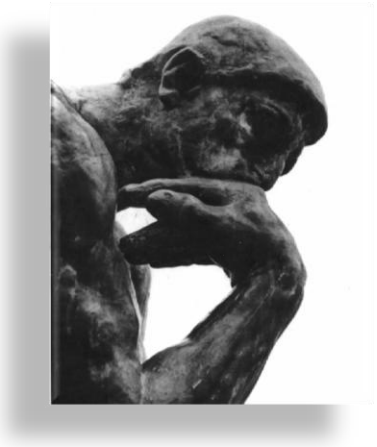
NANYANG TECHNOLOGICAL UNIVERSITY

Visual Graph Query Formulation Meets Query Processing

A novel paradigm

- Why wait for the complete visual query to be constructed before initiating query evaluation? How can we blend these two steps?
- By initiating query processing "early", can we significantly reduce the system response time?





Query formulation

Query processing

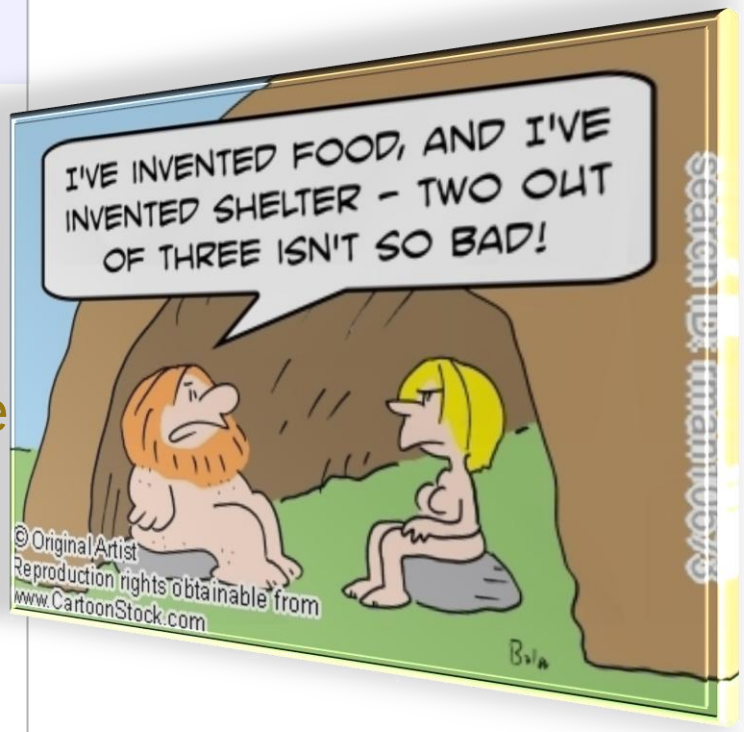
time

NANYANG TECHNOLOGICAL UNIVERSITY

Final Thoughts

VLDB/SIGMOD/ICDE

- Not be-all-end-all
  - PageRank, B Trees, Lorel, etc
- Solve problem that enhances human lives
- Solve problems that you are passionate about
- Don't just follow Goliaths
- Publish and go beyond by building prototypes
  - Always nice to see your ideas working in real rather than on just a piece of paper ☺
- Someday you will be recognized ☺



NANYANG TECHNOLOGICAL UNIVERSITY


Merci beaucoup
شكرا على الواجب
ありがとうございます
Thank you
Спасибо
Mahalo
ARRIGATO
Tak
Vielen Dank
muchas gracias!
Salamat
धन्यवाद
GRACIAS
ευχαριστία