CompSci 590.6

Understanding Data:

Theory and Applications

Lecture 9

Explanation for Database Queries ("Detour" lecture)

Instructor: Sudeepa Roy

Email: *sudeepa@cs.duke.edu*

- Classroom changed

  North 306

Today's Paper(s)

Detour lecture:

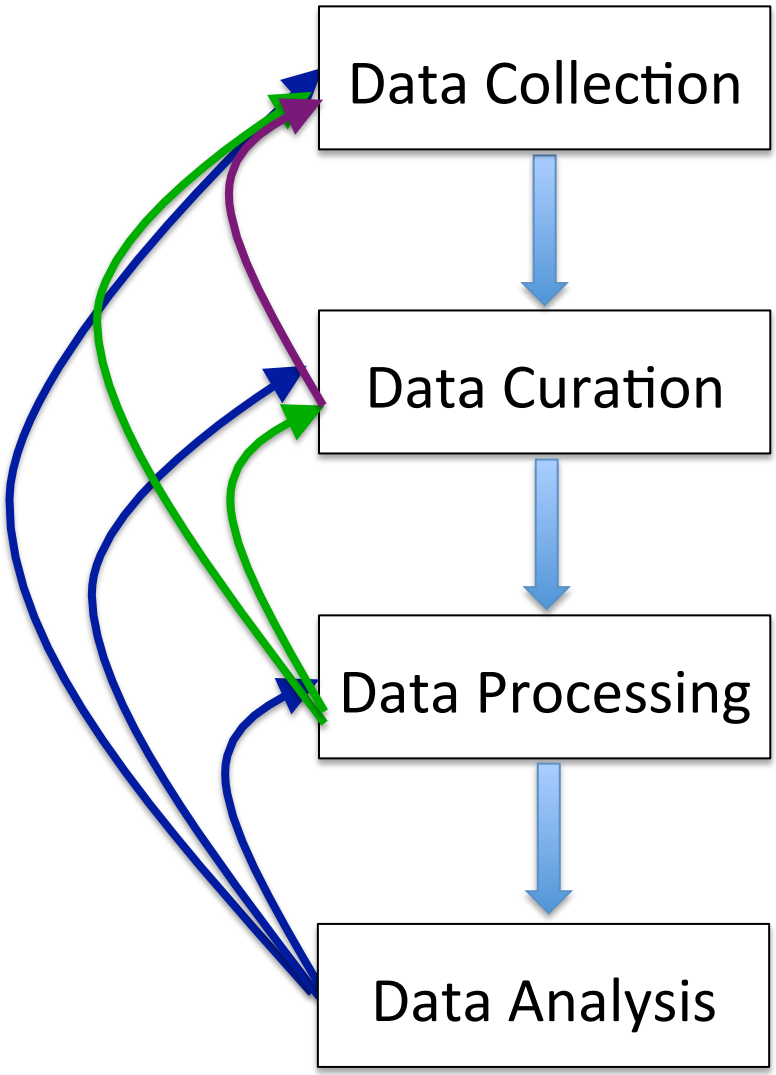
A Formal Approach to Finding Explanations for Database Queries

Roy-Suciu

SIGMOD'14

- An intro to systematic data analysis
- For your course projects!

Data Analysis Pipeline



- Acquire data from different sources

- Clean/Format
- Integrate with other datasets
- Store in database

- Run queries (aggregate)
- Plot graphs

- Examine trends and anomalies
- Understand results

Step 1: Collect datasets

- Several public datasets are available
  - Data.gov
  - CDC/NCHS
  - NSF
  - DBLP
  - Arnetminer
  - Yelp academic data
  - Duke library

Step 2: Curate datasets

- Store from XML/TEXT to DBMS
- Figure out schema
- Clean
- Extract new features
  - www.cs.duke.edu/~sudeepa
  - Duke.edu
  - cs / stat / math
  - Edu
- Integrate
  - Store multiple datasets in the same database
  - DBLP, Arnetminer, NSF

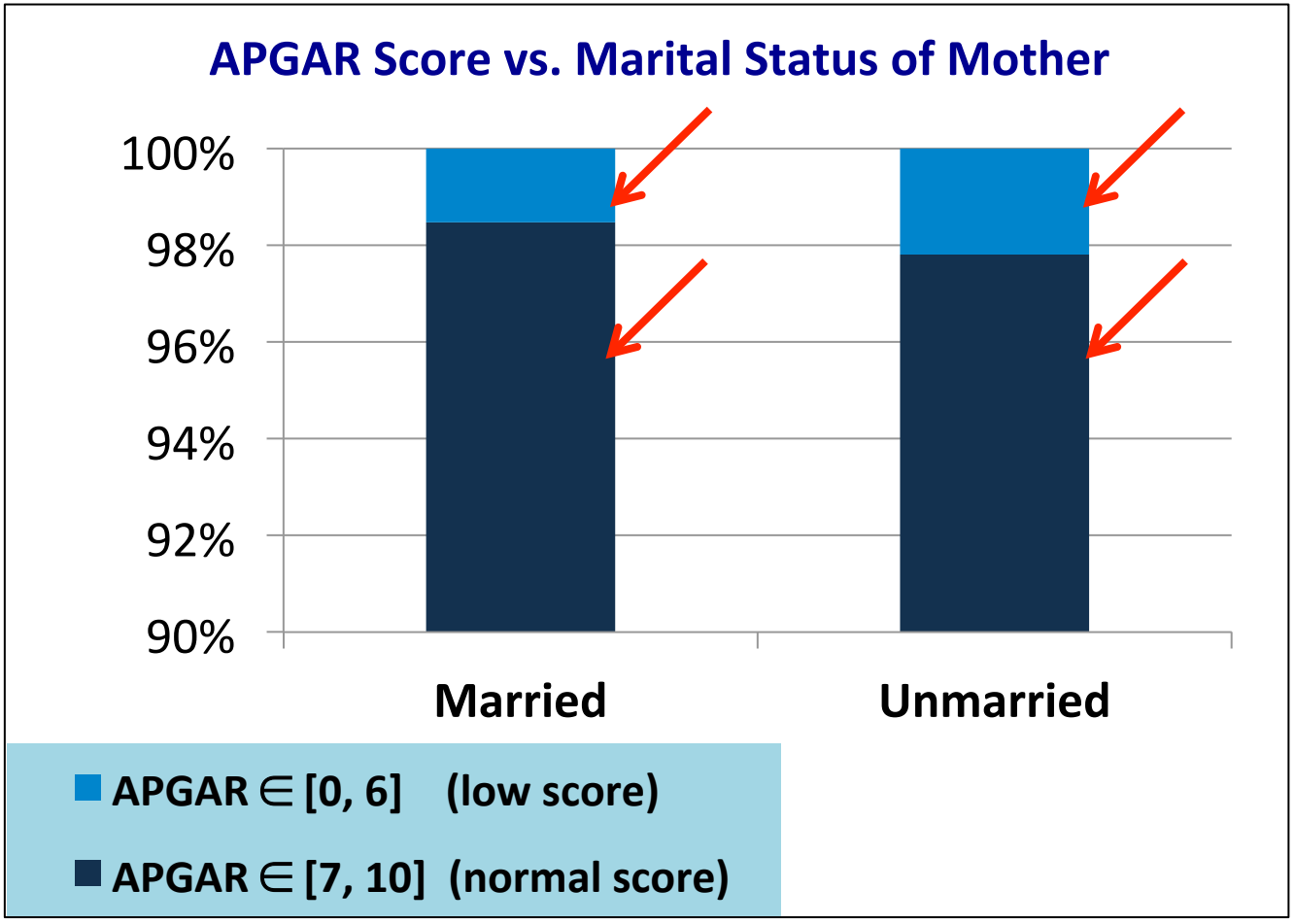
Step 3: Process and ask questions

- Fun step!

- Run several queries
- Plot graphs
- Ask questions

Natality data 2010
(CDC/NCHS)

Example 1: Health of a newborn vs. marital status of the mother



Married mothers have healthier babies. **Explain why.**

Example 2: Top-5 schools in CS with highest total NSF grant $ from 1990

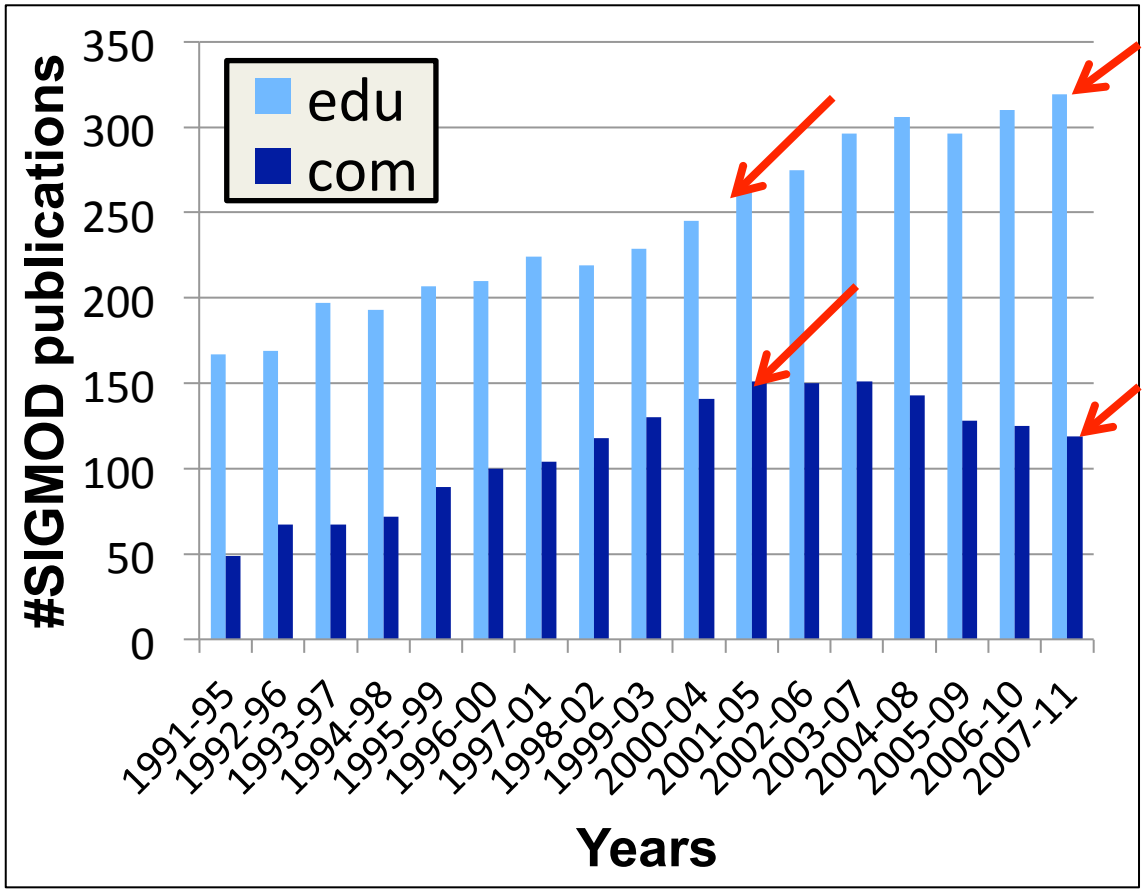
Rank	School	Total Award $ from 1990
1	UIUC	**1169.7** Million
2	UCSD	723.3 Million
3	CMU	**472.9** Million
4	UT Austin	319.4 Million
5	MIT	292.7 Million

Rank (as grad school) on US News: UIUC - 5, CMU – 1
both about 60 primary faculty in CS

UIUC received much larger amount of awards than CMU. Explain why.

source: CS webpages, NSF, US News

Example 3: #SIGMOD papers from academia and industry

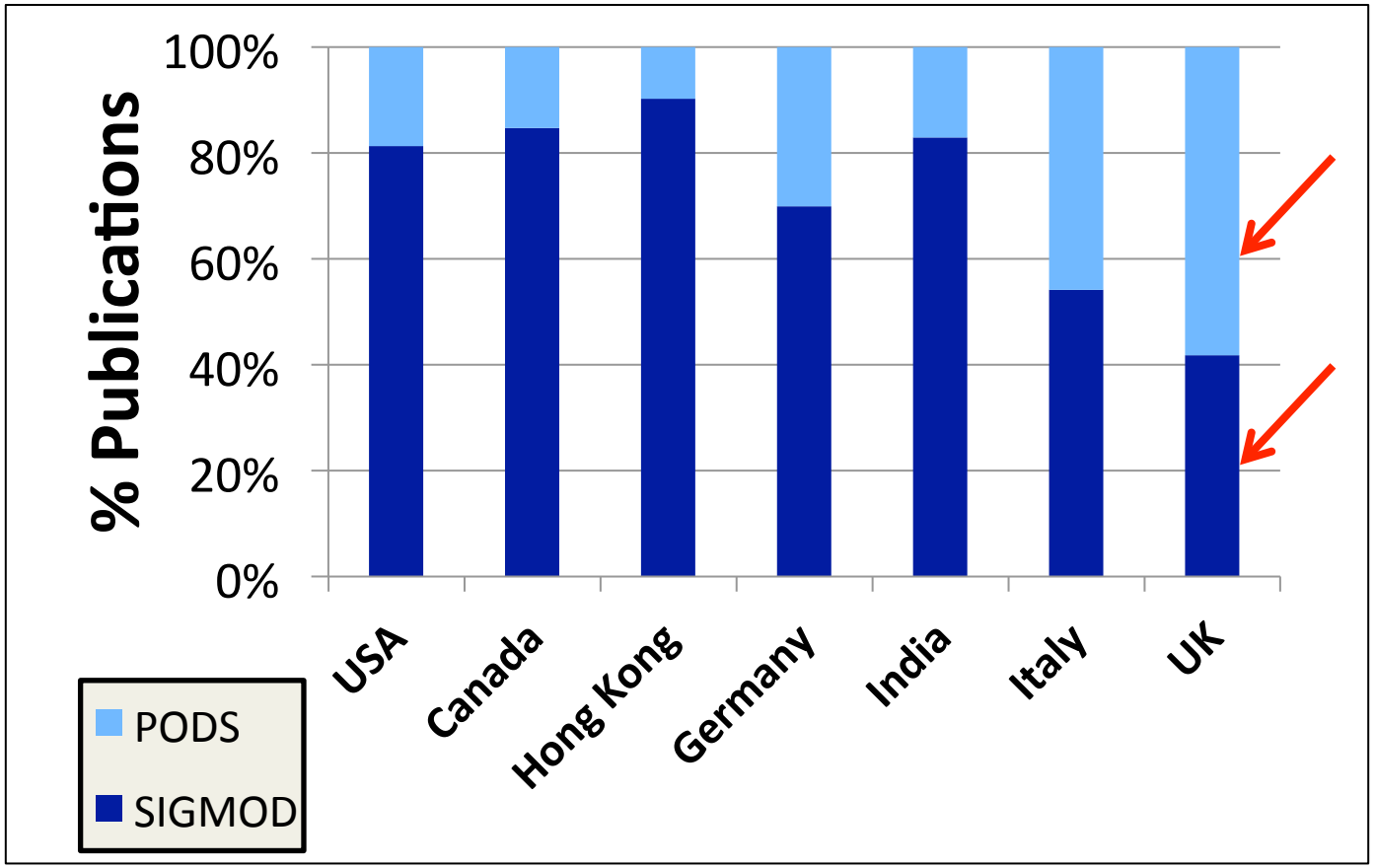


A peak for industry around 2000.

An increasing trend for academia. **Explain why.**

DBLP data

Example 4: #SIGMOD (systems) vs. #PODS (theory) papers from different countries



#SIGMOD papers ≤ #PODS papers in UK. **Explain why.**

Step 4: Data Analysis

- This is the challenging step
- How to answer these question
- Start with a clean formulation

Primary keys

Foreign keys

Database 101

aid	name	inst	dom
A1	LL	E.uk	uk
A2	DS	W.edu	edu
A3	MB	O.uk	uk

Author (A)

aid	pubid
A1	P1
A2	P1
A1	P2
A3	P2
A2	P3
A3	P3

Authored (AD)

pubid	year	venue
P1	2001	PODS
P2	2011	PODS
P3	2001	SIGMOD

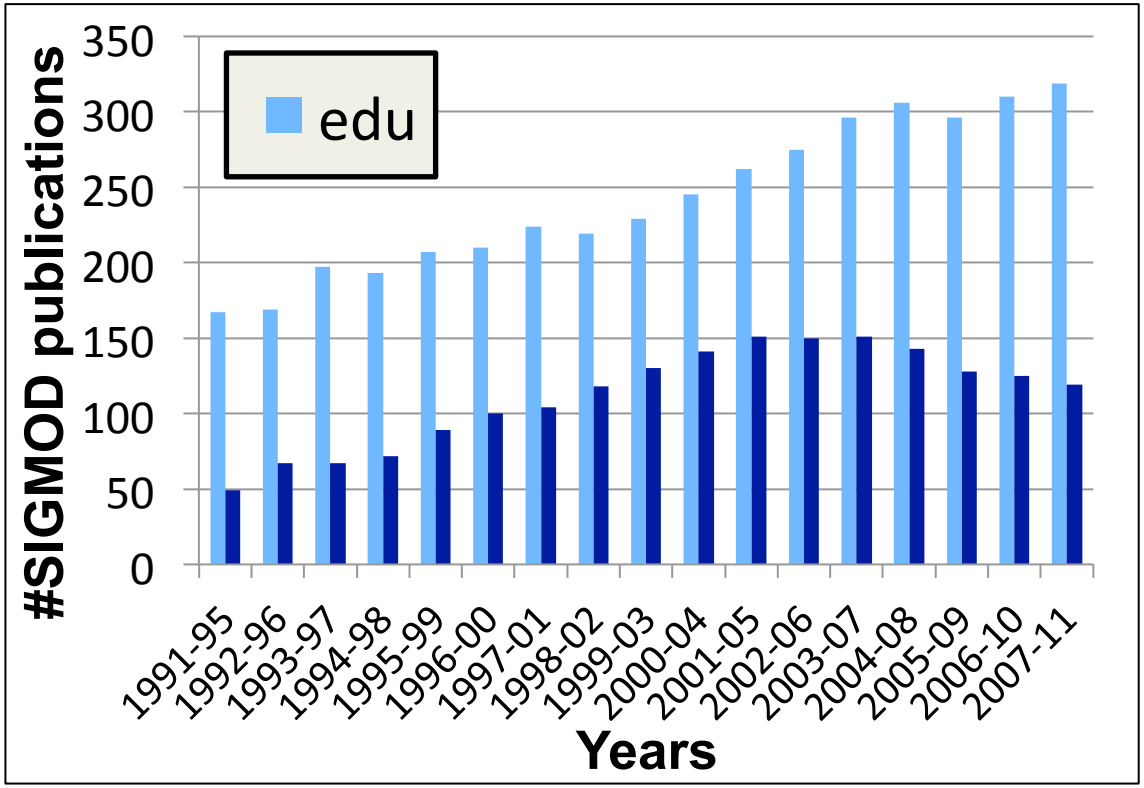
Publications (P)

Toy DBLP database

1. Relational databases
   - multiple tables
   - row = tuple, col = attribute
   - primary/foreign keys
2. Aggregate queries

A simpler SQL query

```
SELECT P.year, count(distinct P.pubid)
FROM  A, AD, P
WHERE A.aid = AD.aid
   AND AD.pubid = P.pubid
   AND P.venue = 'SIGMOD'
    .....
GROUP BY P.year
ORDER BY P.year
```

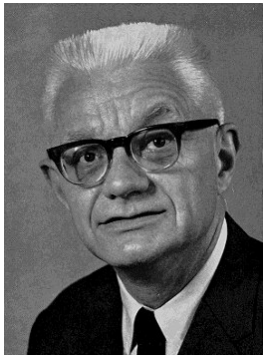


Causality and Explanations: A Brief History

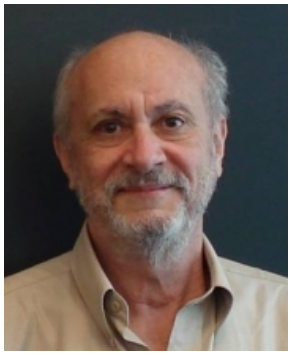
Aristotle (384-322 BC)

David Hume (1711-76)

Karl Pearson (1857-1936)



Carl Gustav Hempel (1905-97)



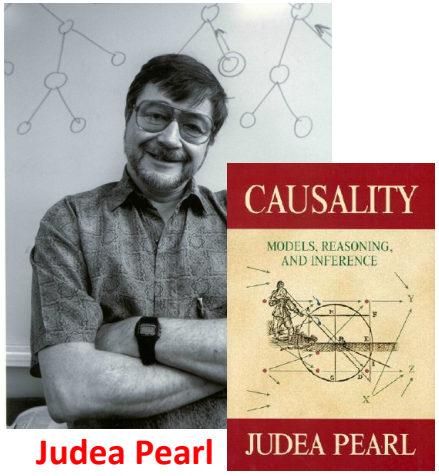
Donald Rubin (1943-)

Philosophy

Statistics

AI

Economics



Judea Pearl (1936-)

Turing awardee: 2011

"It took another 25 years … to formulate the **randomized experiment – the only scientifically proven method of testing causal relations from data**, and to this day, the one and only causal concept permitted in mainstream statistics. And that is roughly where things stand today."

Judea Pearl, 2000
(Causality: Models, Reasoning, and Inference)

Not possible with available data

Explanation by Intervention (our SIGMOD'14 paper)

Controlled Experiments ≡

Causation by Intervention (J. Pearl):

"A variable Y is a cause of Z if we can change Z by manipulating Y"

- i.e. $\Delta Y \Rightarrow \Delta Z$

Databases:

A set of input tuples is an explanation of one or more query answers if we can change the answers by "manipulating" these tuples

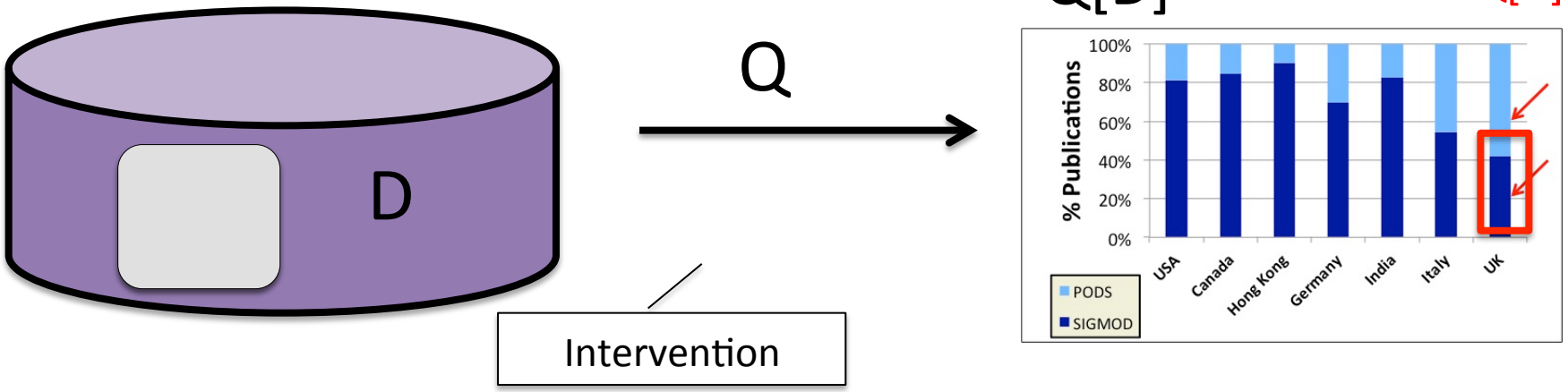
- i.e. $\Delta D \Rightarrow \Delta Q[D]$

What kinds of manipulation can we support?

$\Delta D \Rightarrow \Delta Q[D]$

Intervention in Databases

- Modification?
  - *"E.F. Codd proposed Relational Model in 1970"*
- Insertion?
  - *"S. Roy wrote a SIGMOD paper in 1970"*
- Restricted to Tuple Deletion



Explanation = a compact, high level description

Explanation Desiderata

- Define explanation ϕ
- Compute its intervention Δ_ϕ
- Find top-k explanations

- Succinctness
- Computability of intervention
- Formalize user question and scoring function
- Efficient exploration of explanation space and rank

Explanation Desiderata

- Define explanation ϕ
- Compute its intervention Δ_ϕ
- Find top-k explanations

Class of explanation

- Succinctness
- Computability of intervention
- Formalize user question and scoring function
- Efficient exploration of explanation space and rank

Succinctness

aid	name	inst	dom
A1	LL	E.uk	uk
A2	DS	W.edu	edu
A3	MB	O.uk	uk

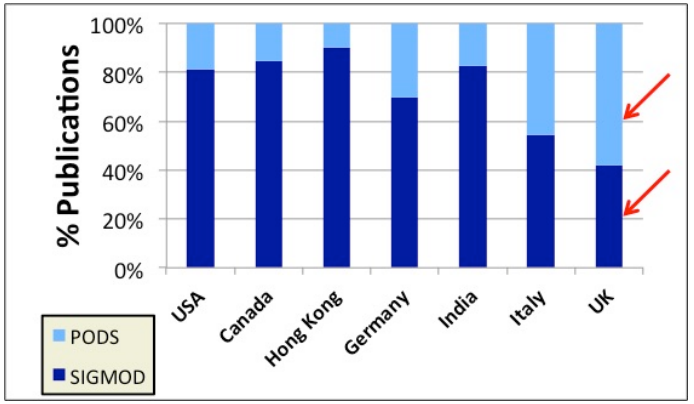
Author

aid	pubid
A1	P1
A2	P1
A1	P2
A3	P2
A2	P3
A3	P3

Authored

pubid	year	venue
P1	2001	PODS
P2	2011	PODS
P3	2001	SIGMOD

Publications



- Explanation = a subset of tuples
- Arbitrary subset?
  - Long explanations, may lack common properties
  - Hard to interpret, overfitting
  - Exponential search space in #tuples

Succinctness

aid	name	inst	dom
A1	LL	E.uk	uk
A2	DS	W.edu	edu
A3	MB	O.uk	uk

Author

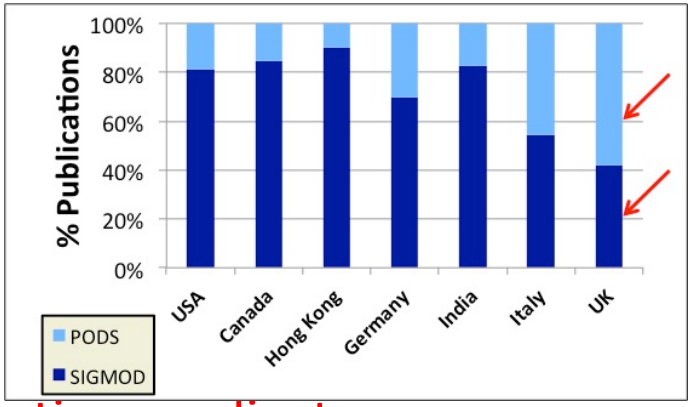
aid	pubid
A1	P1
A2	P1
A1	P2
A3	P2
A2	P3
A3	P3

Authored

pubid	year	venue
P1	2001	PODS
P2	2011	PODS
P3	2001	SIGMOD

Publications

[inst = E.uk]
[name = LL] ∧ [year = 2001]
(multiple tables)



- Allow subsets that are specified by conjunctive predicates
  - Explanations are succinct (~ #attributes)
  - Polynomial search space in #tuples
  - Exponential in #attributes (can be controlled)

Explanation Desiderata

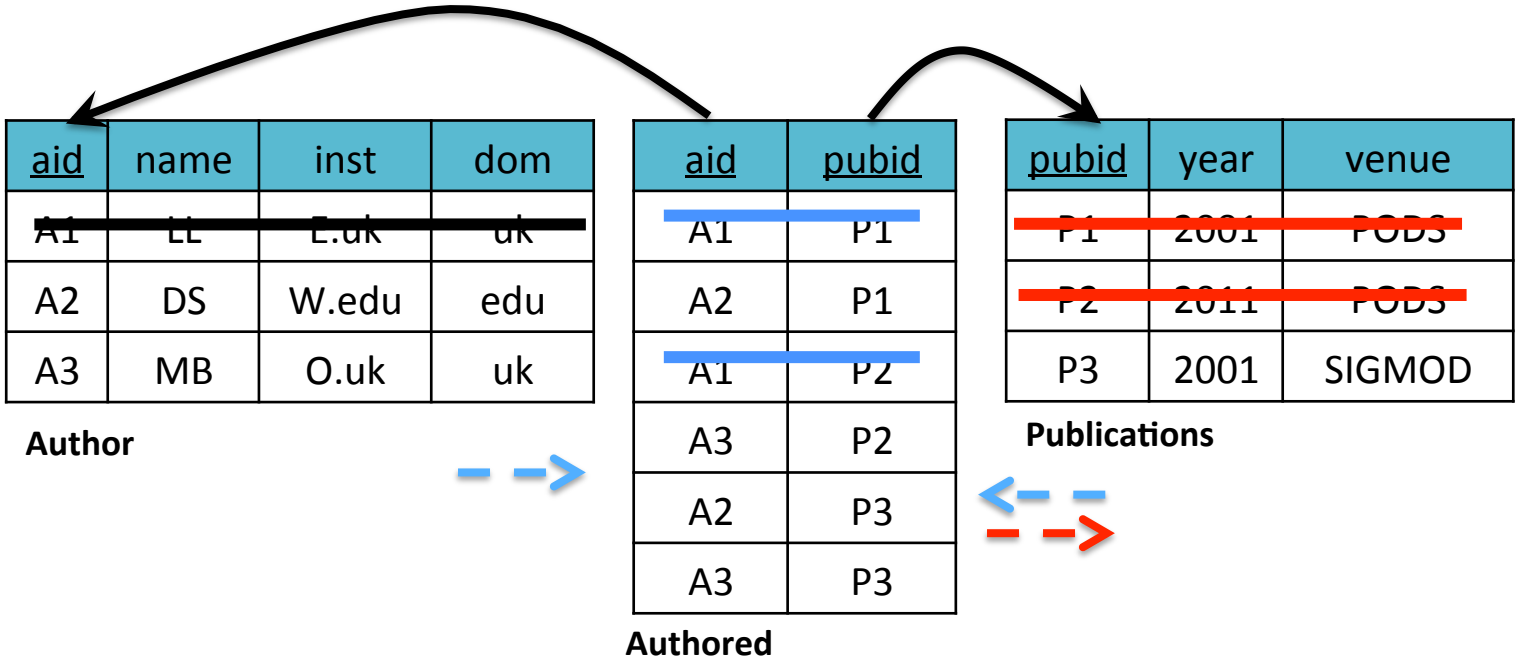
- Define explanation ϕ
- Compute its intervention Δ_ϕ
- Find top-k explanations

✔ Succinctness — Conjunctive predicates

- Computability of intervention — Only satisfying a predicate is not sufficient
- Formalize user question and scoring function
- Efficient exploration of explanation space and rank

Intervention
= tuple deletion

Induced Tuple Deletion by Foreign Keys



aid	name	inst	dom
~~A1~~	~~LL~~	~~L.uk~~	~~uk~~
A2	DS	W.edu	edu
A3	MB	O.uk	uk

Author

aid	pubid
~~A1~~	~~P1~~
A2	P1
~~A1~~	~~P2~~
A3	P2
A2	P3
A3	P3

Authored

pubid	year	venue
~~P1~~	~~2001~~	~~PODS~~
~~P2~~	~~2011~~	~~PODS~~
P3	2001	SIGMOD

Publications

- Standard Foreign Keys
  - Forward cascade delete
- Back-and-forth Foreign Keys
  - Forward cascade delete
  - Reverse cascade delete

Additional causal paths (semantic):

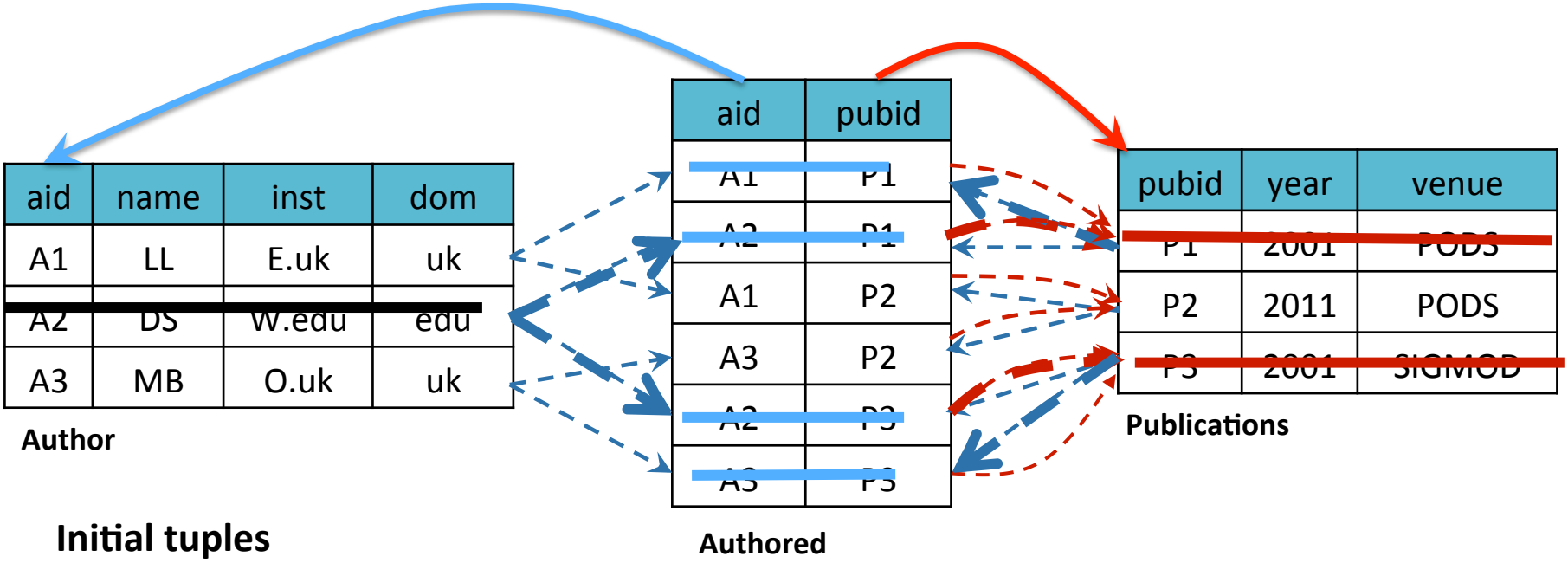
- Author ⟶ Publication
- Publication ⟶̸ Author

"Causal dependency" between tuples

Intervention Δ_ϕ for a given ϕ

- Intervention Δ_ϕ contains 7 tuples
- Recursion even for acyclic schema
- Not reachability
  - No explicit edges
  - Predicates can span multiple tables

Candidate explanation
ϕ : [name = 'DS']



Input:

Relations $R_1, \ldots, R_k$ (+ foreign keys)

Attributes $A_1, \ldots, A_k$

Fixed predicate ϕ

Output:

Interventions $\Delta_1, \ldots, \Delta_k$

(Theorem-proving step) Recursive Query to Compute Δ_ϕ

No need to look at the details here

(Initial tuples) Rule 1: $\Delta_i = R_i - \Pi_{A_i} \sigma_{\neg\phi} [R_1 \bowtie \ldots \bowtie R_k]$

(Forward delete) Rule 2: $\Delta_i = R_i - \Pi_{A_i} [(R_1 - \Delta_1) \bowtie \ldots \bowtie (R_k - \Delta_k)]$

(Reverse delete) Rule 3: $\Delta_i = R_i \ltimes_{pk=fk} \Delta_j$ $\quad R_j \longrightarrow R_i$

- Query is not monotone in database
  - i.e., if $D \subseteq D'$, not necessarily $\Delta(D) \subseteq \Delta(D')$
  - Standard techniques (Datalog) do not directly work
- Query has a unique least fixpoint, poly-time convergence
- #Steps to converge depend on the schema (characterization)

Explanation Desiderata

- Define explanation ϕ
- Compute its intervention Δ_ϕ
- Find top-k explanations

✔ Succinctness — Conjunctive predicates

✔ Computability of intervention — Recursion for a given predicate

- Formalize user question and scoring function — Aggregate queries
- Efficient exploration of explanation space and rank

User Question and Scoring Function

aid	name	inst	dom
A1	LL	E.uk	uk
A2	DS	W.edu	edu
A3	MB	O.uk	uk

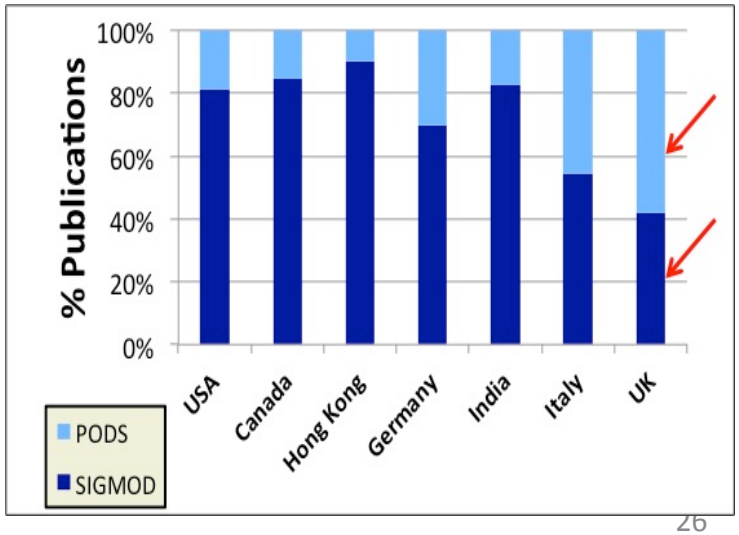
Author

aid	pubid
A1	P1
A2	P1
A1	P2
A3	P2
A2	P3
A3	P3

Authored

pubid	year	venue
P1	2001	PODS
P2	2011	PODS
P3	2001	SIGMOD

Publications



User Question and Scoring Function

aid	name	inst	dom
A1	LL	E.uk	uk
A2	DS	W.edu	edu
A3	MB	O.uk	uk

Author

aid	pubid
A1	P1
A2	P1
A1	P2
A3	P2
A2	P3
A3	P3

Authored

pubid	year	venue
P1	2001	PODS
P2	2011	PODS
P3	2001	SIGMOD

Publications

F

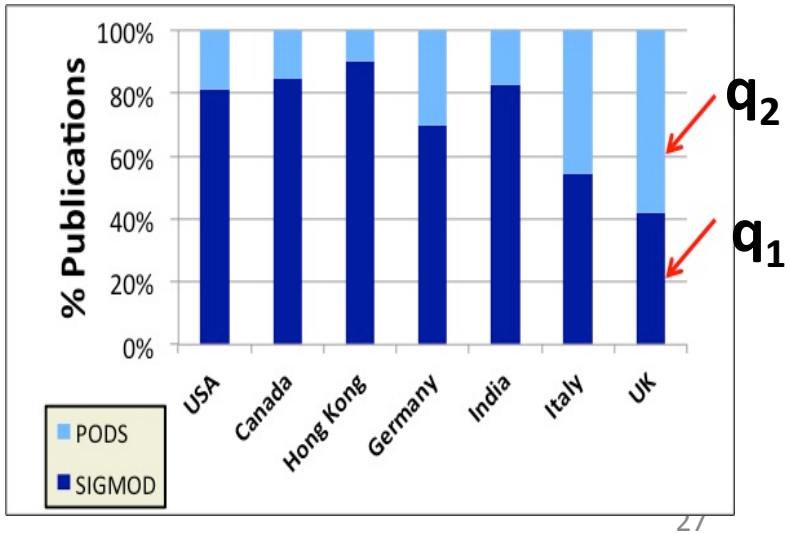
Explain why is q_1/q_2 low

q_1: count distinct 'SIGMOD' papers from 'uk'
q_2: count distinct 'PODS' papers from 'uk'

User question: A numeric function **F** of simple aggregate queries

Score of ϕ = $F(D - \Delta_\phi)$
$= q_1(D - \Delta_\phi) / q_2(D - \Delta_\phi)$

Δ_ϕ = Intervention of ϕ



Explanation Desiderata

- Define explanation ϕ
- Compute its intervention Δ_ϕ
- Find top-k explanations

- Single explanation: recursion
- Large #explanations
- Don't run a FOR LOOP
  – use DATA CUBE

✔ Succinctness — Conjunctive predicates

✔ Computability of intervention — Recursion for a given predicate

✔ Formalize user question and scoring function — Numeric function $F(q_1, q_2, \ldots)$

- Efficient exploration of explanation space and rank

Data cube: Gray et al.'97

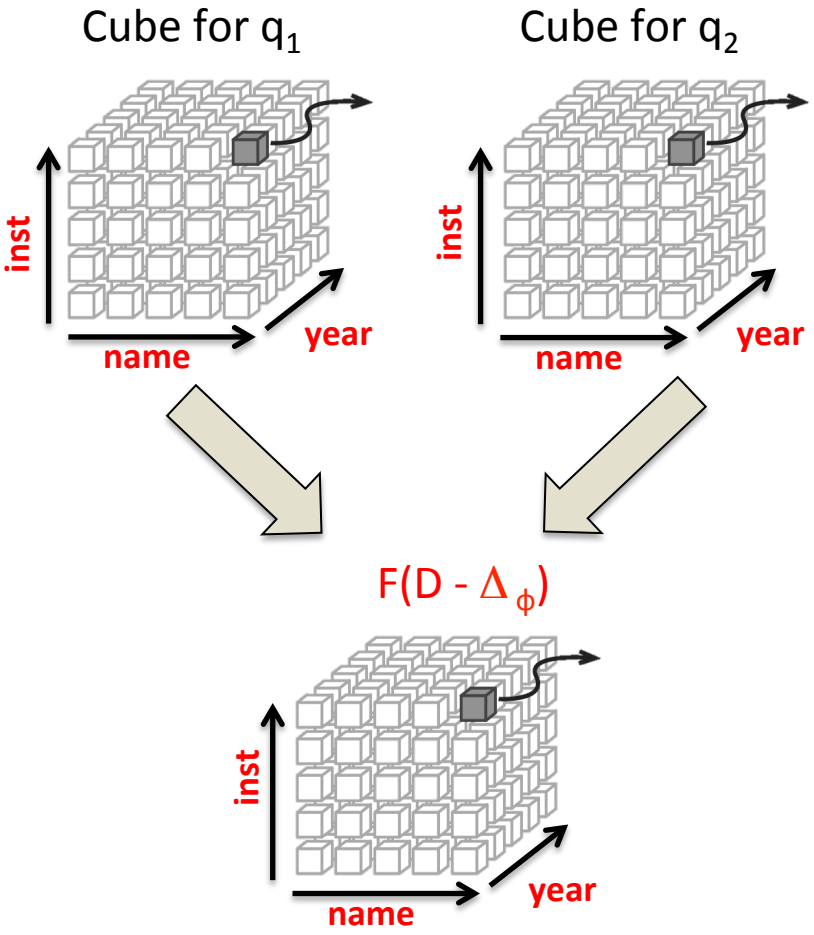
Current topic in 590.6

(Algorithm and optimization step) Optimization with OLAP Data Cube

q_1: count distinct 'SIGMOD' papers from 'uk'
q_2: count distinct 'PODS' papers from 'uk'

Goal:
Compute for all ϕ
$F(D - \Delta_\phi) = q_1(D - \Delta_\phi)/q_2(D - \Delta_\phi)$

- Fix a set of explanation attributes
  - A = {inst, name, year}
- Compute data cube on A for q_1, q_2
- Combine to compute final score and rank
- Computation mostly by DBMS
- Matches the semantic or a heuristic



Explanation Desiderata

- Define explanation ϕ
- Compute its intervention Δ_ϕ
- Find top-k explanations

✔ Succinctness — Conjunctive predicates

✔ Computability of intervention — Recursion for a given predicate

✔ Formalize user question and scoring function — Numeric function $F(q_1, q_2, \ldots)$

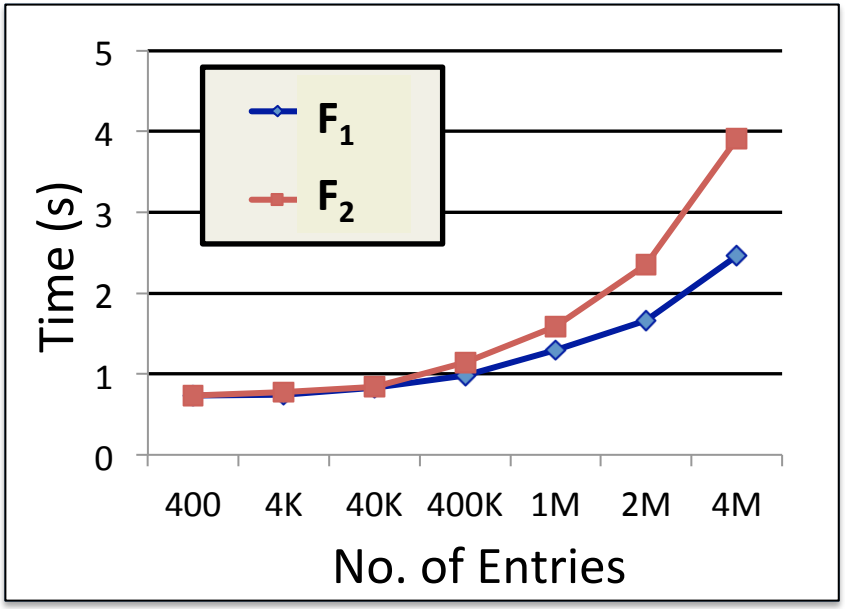
✔ Efficient exploration of explanation space and rank — Data cube

F_1: Explain why: (q_1/q_2) low

F_2: Explain why: $(q_1*q_2) / (q_3*q_4)$ low

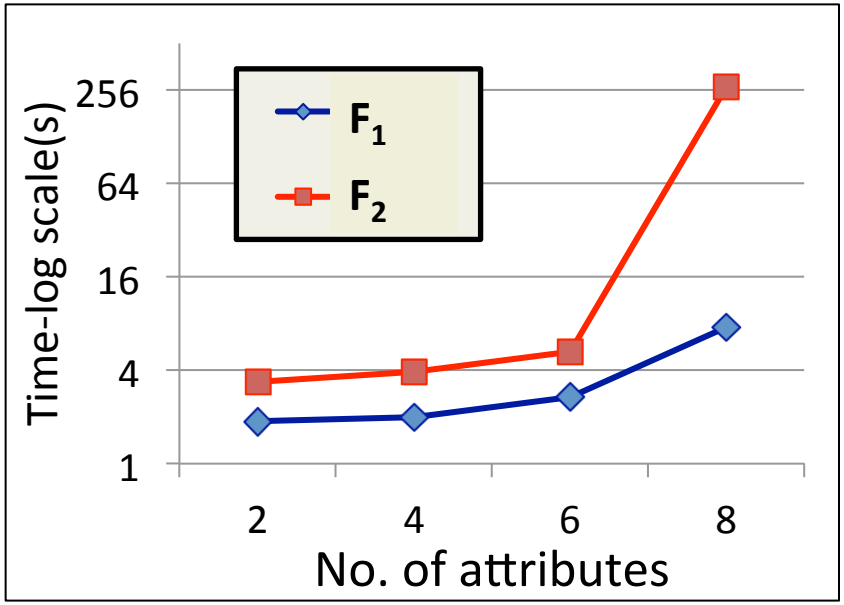
Scalability of Data Cube

Data size vs. time



4 attributes

#Explanation attributes vs. time

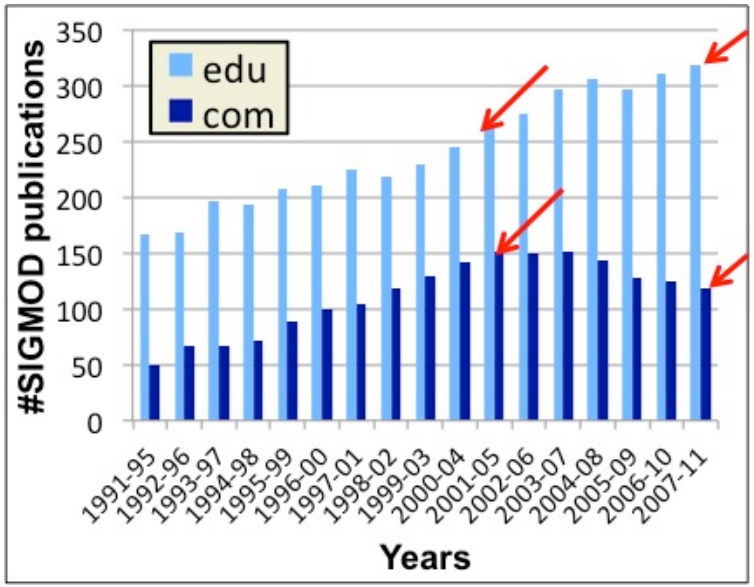


4M entries

Interactive speed (well..)
Slows down with #attr, #tuples, and query complexity

Natality Dataset 2010: (CDC/NCHS)
Single table with 233 attributes, ~4M entries, 2.89GB size

Qualitative Evaluation



	Explanations
1	inst = ibm.com
2	inst = bell-labs.com
3	name = Rajeev Rastogi
4	inst = ucla.edu
5	name = Hamid Pirahesh
6	inst = asu.edu
7	name = Rakesh Agrawal

A peak for #sigmod papers from industry, while academia papers kept increasing. Explain why.

1. Leading industrial labs and their senior researchers
2. New highly active academic database groups

For NSF data, some large "ACI" grants, and PIs who got > 500 M awards (in total from 1990)

Research Directions

- Several interesting directions
  - more complex explanations
  - uncertainty
  - performance
  - interactive exploration
  - ……