Do You Know SQL?

About Semantic Errors in SQL Queries

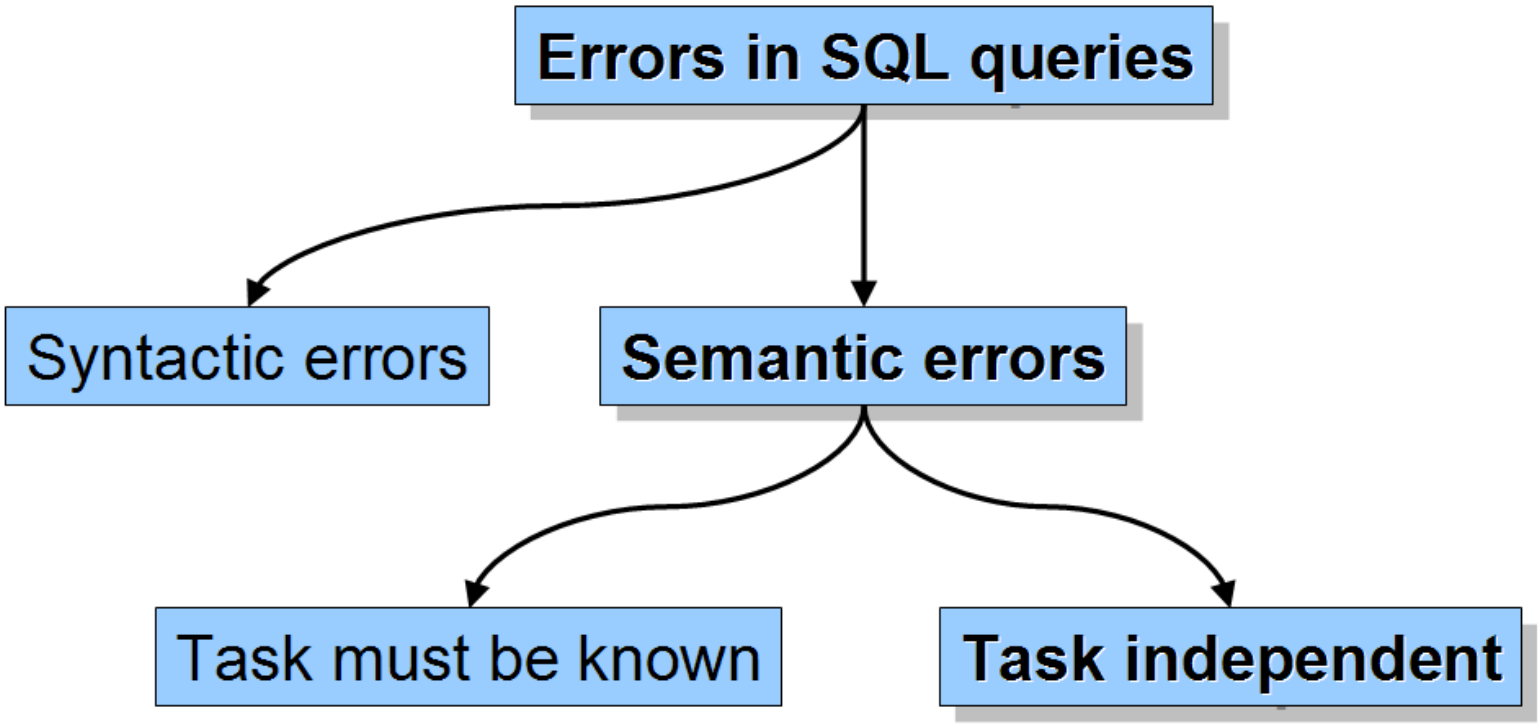
Christian Goldberg
goldberg@informatik.uni-halle.de

Institut für Informatik
Martin-Luther-Universität Halle-Wittenberg

TLAD 2009
Birmingham, 6th July 2009

Contents

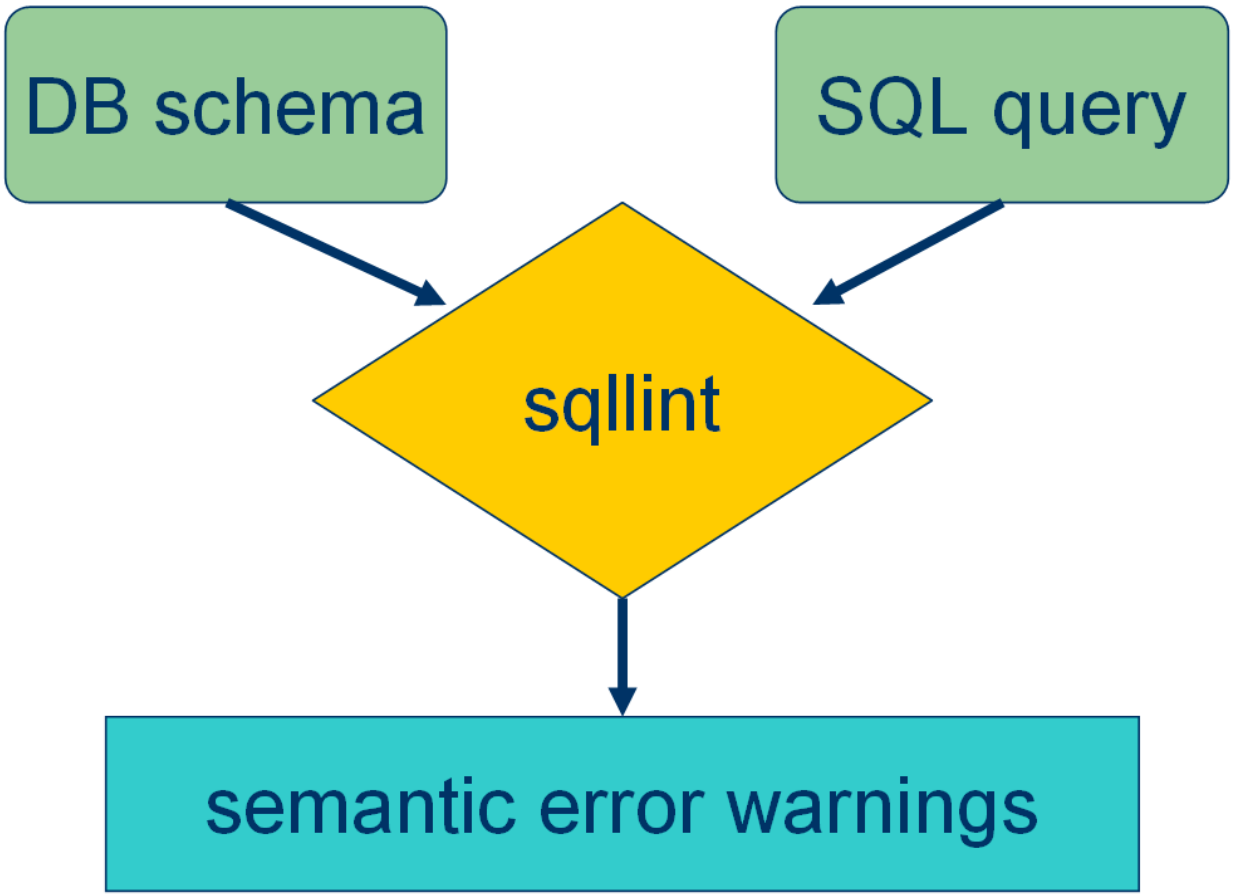
Classification of Errors



Example (Inconsistent Condition)

```
SELECT ENAME
FROM   EMP
WHERE  JOB = 'CLERK' AND JOB = 'MANAGER'
```

- Empty result in all database states
  (certainly not intended)
- Inconsistent condition is a frequent student error
- In general not decidable


DB schema
SQL query
sqllint
semantic error warnings

Unnecessary Complications

- Unnecessarily complicated query
  $\longrightarrow$ "probably not intended"
- Situation:
  1. User wrote query A.
  2. B exists equivalent to A.
  3. B is significantly simpler than A:
     B results from A by deleting parts of the query.

- Complications possible in all query parts:

SELECT	Constant / duplicate output columns
FROM	Unused tuple vars, Unnecessary joins
WHERE	Implied, tautological or inconsistent sub-conditions, Unnecessary general comparison operator
GROUP BY	singleton groups, only one group

...

Entire query unnecessary (Inconsistent Condition)

Example (singleton groups)

```
SELECT   EMPNO, MAX(SAL)
FROM     EMP
WHERE    JOB = 'MANAGER'
GROUP BY EMPNO
```

Example (comparison operator)

```
SELECT ENAME, SAL
FROM   EMP
WHERE  SAL >= (SELECT MAX(SAL) FROM EMP)
```

Violation of Standard Patterns

- Missing join conditions
- Uncorrelated EXISTS-subqueries
- SELECT clause of subquery uses no tuple variable from the subquery
- Conditions in subquery that can be moved up
- Comparison between different domains
- HAVING without GROUP BY
- DISTINCT in SUM and AVG
- Wildcards without LIKE

Further Classes of Semantic Errors

- Duplicates (Unnecessary `DISTINCT`, Many duplicates)
- Inefficient Formulations (Inefficient `HAVING`/`UNION`)
- Possible Runtime Errors (`SELECT INTO` that might return more than one tuple)
- "Bad Style" (Inconsistent use of defaults)

A quite complete list of over 40 semantic errors can be found in:
[S.Brass and C.Goldberg, Journal of Systems and Software 79(5), 2006]

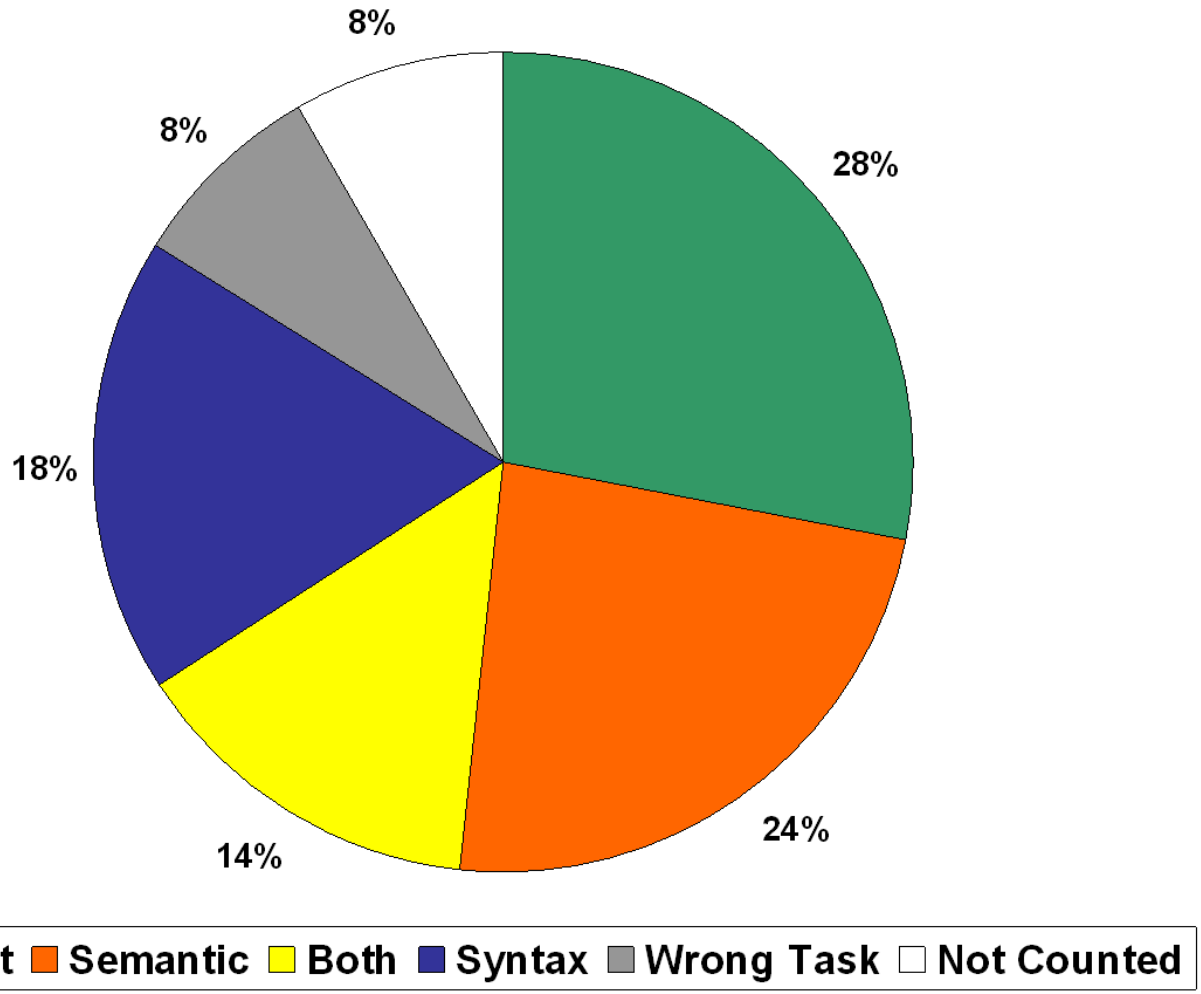
Base Data

- Five "Databases I" exams analyzed with 22 SQL exercises

Exam	Part.	Exercises SQL	Points SQL	Points Total
Final 02/03	67	4	10	50
Midterm 03/04	153	3	9	23
Final 03/04	148	3	9	20
Final 05/06	40	6	18	37
Final 08/09	53	6	15	35

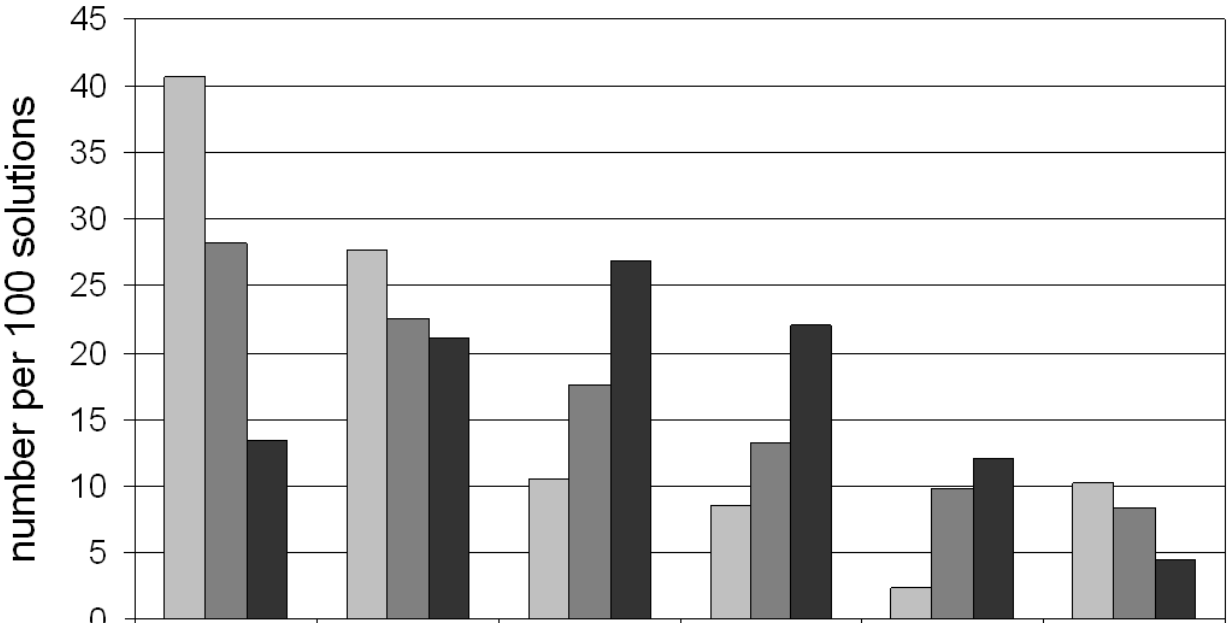
- Classification according to degree of difficulty:
  - beginner (6 exercises)
  - intermediate (9 exercises)
  - advanced (7 exercises)

Analytical Result for 1411 queries in five exams



Martin-Luther-Universität
Halle-Wittenberg

- Normalized distribution per difficulty class:



	Correct	Sem.	Syntax	Both	N.C.	W.T.
beginner	40,71	27,67	10,47	8,5	2,37	10,28
intermediate	28,22	22,63	17,56	13,26	9,88	8,45
advanced	13,44	21,15	26,87	22,03	12,11	4,41

Most frequent semantic errors

Ratio	Semantic Error
15%	Missing join condition
13%	Many duplicates
11%	Unnecessary join
8%	Inconsistent condition
6%	Unnecessary argument of `COUNT`
5%	Implied, tautological or inconsistent subcondition
5%	Unnecessary `DISTINCT`

- Percentages are relative to all detected semantic errors

Possible Causes and Solutions

"I thought it will be joined if I type it under `FROM`."

- Lack of preparation
- Absence from lectures and exercises
- Improperly reading of given tasks
- Insufficient experience in programming SQL
- But also: Insufficient comprehension of underlying DB schema

Sometimes fewer errors by making use of connection graphs and discussion

Possible Causes and Solutions

"I thought it will be joined if I type it under `FROM`."

- Lack of preparation
- Absence from lectures and exercises
- Improperly reading of given tasks
- Insufficient experience in programming SQL
- But also: Insufficient comprehension of underlying DB schema

Sometimes fewer errors by making use of connection graphs and discussion

Conclusions

- Current database systems print no warnings, only error messages if query is not executable
- We develop a semantic checker for SQL called `sqllint`
- The paper gives a survey of how often and which semantic errors appear
- Sensible error message possible in nearly a quarter of all cases
- For detailed exam descriptions and `sqllint` prototype, see:

  `http://dbs.informatik.uni-halle.de/sqllint/`

Further Literature


Stefan Brass and Christian Goldberg:
Semantic Errors in SQL Queries: A Quite Complete List.
In: *Journal of Systems and Software 79(5)*, 633–644, 2006.


Stefan Brass, Christian Goldberg:
Proving the Safety of SQL Queries.
In: *Proceedings of the Fifth International Conference on Quality Software (QSIC'05)*, 197–204, 2005