
# Robust Optimization of Database Queries

Jayant Haritsa
Database Systems Lab
Indian Institute of Science

# Database Management Systems (DBMS)

- Efficient and convenient mechanisms for storing, maintenance and querying of enterprise data
  - banking, inventory, insurance, travel, …
- Cornerstone of computer industry
  - Uses more than 80 percent of computers worldwide
  - Employs more than 70 percent of IT professionals
  - Largest monetary sector of computer business

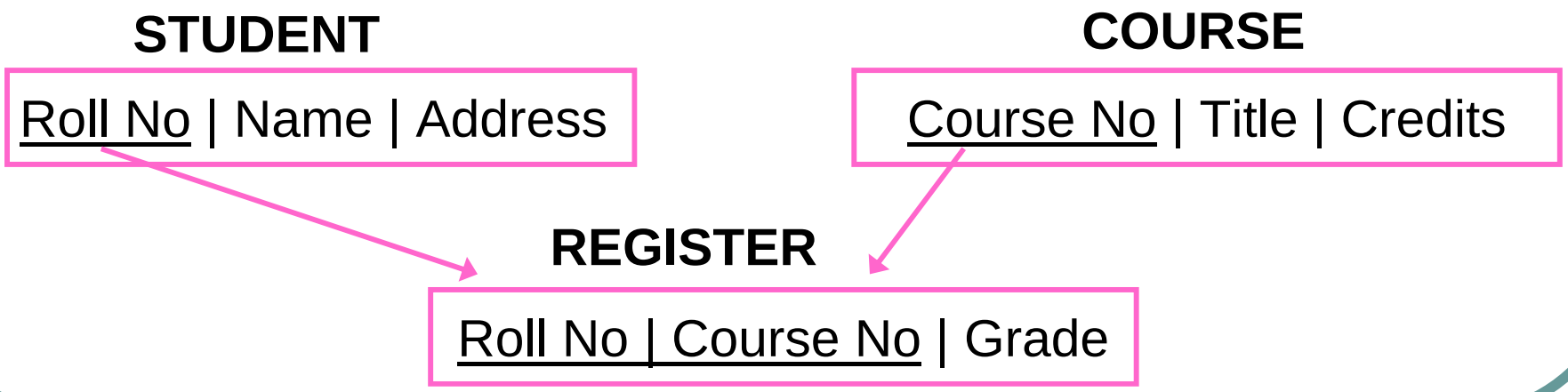
# Current Database Systems

- Commercial
  - IBM DB2 / Oracle / Microsoft SQL Server / Sybase ASE / CA Ingres
- Public-Domain
  - PostgreSQL / MySQL

# Relational Database Systems
## [RDBMS]

- Based on first-order logic
  - Edgar Codd of IBM Research, Turing Award (1981)
  - "We believe in Codd, not God"

- Data is stored in a set of relations (i.e. tables) with attributes, relationships, constraints

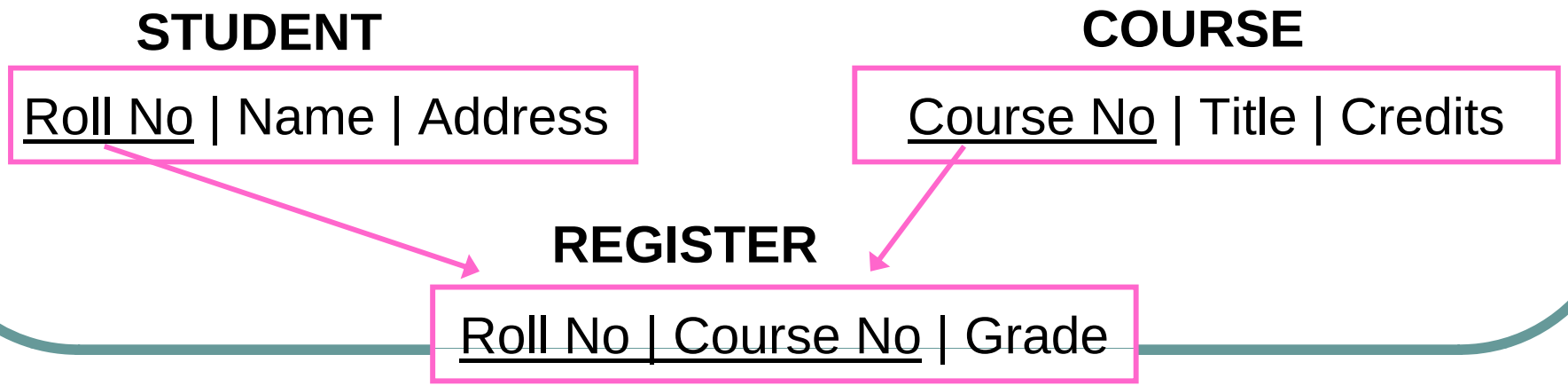


# QUERY INTERFACE

## Structured Query Language (SQL)

- Invented by IBM, 1970s
- Example: *List names of students and their course titles*

```
select  STUDENT.Name, COURSE.Title
 from    STUDENT,  COURSE,  REGISTER
where  STUDENT.RollNo = REGISTER.RollNo   and
          REGISTER.CourseNo = COURSE.CourseNo
```



# Query Execution Plans

- SQL is a declarative language
  - Specifies only what is wanted, but not how the query should be evaluated (i.e. ends, not means)

```
select   STUDENT.Name, COURSE.Title
from     STUDENT,  COURSE,  REGISTER
where    STUDENT.RollNo = REGISTER.RollNo   and
           REGISTER.CourseNo = COURSE.CourseNo
```

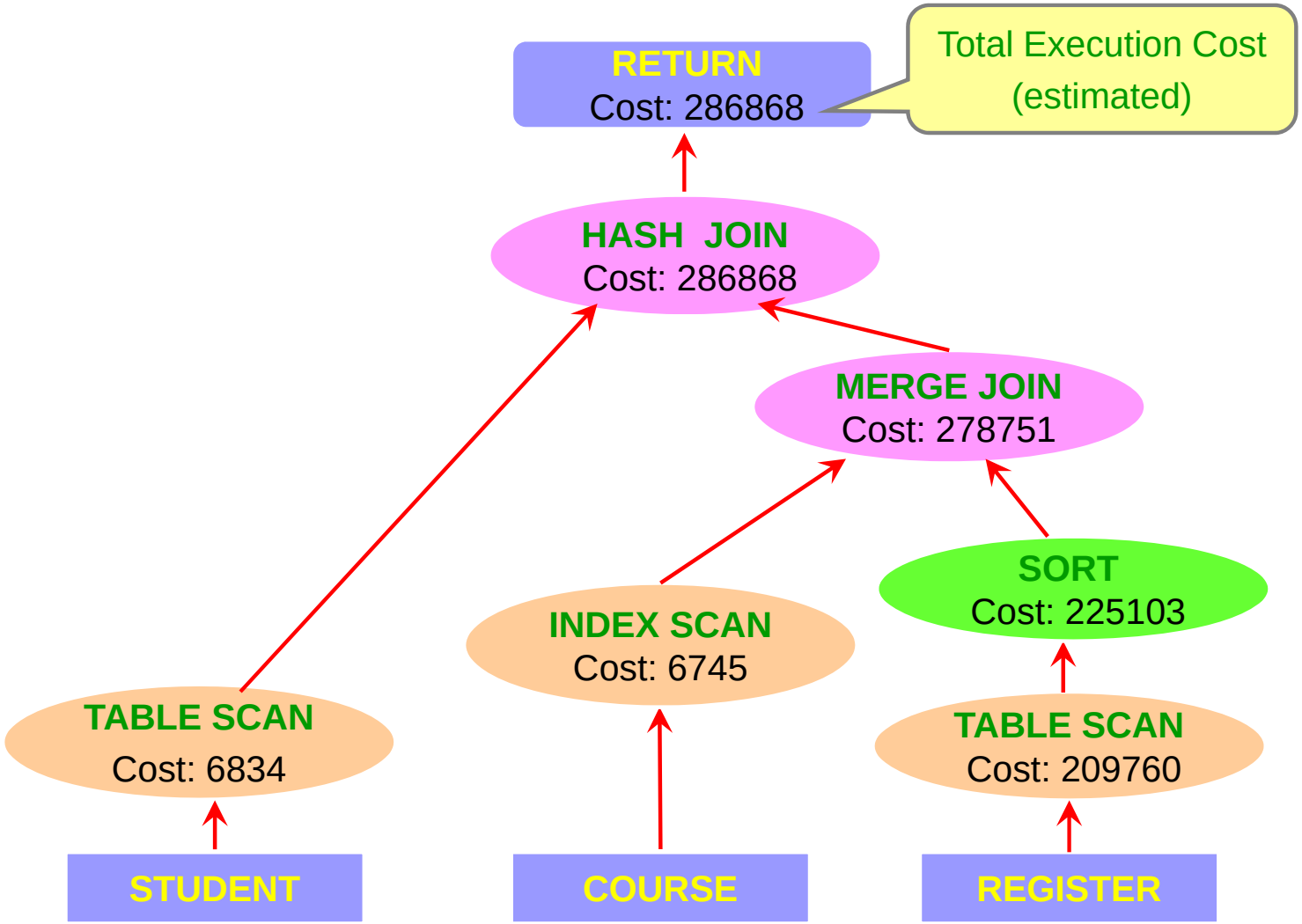
Unspecified:

**join order [ ((S ⋈ R) ⋈ C) or ((R ⋈ C) ⋈ S) ? ]**

**join techniques [ Nested-Loops or Sort-Merge or Hash ? ]**

- DBMS query optimizer identifies the optimal evaluation strategy: "query execution plan"

# Sample Execution Plan




# CONCEPT:
# PLAN DIAGRAMS

# Relational Selectivities

- Cost-based Query Optimizer's choice of

  execution plan = f (query, database, system, …)

- For a given database and system setup,

  execution plan chosen for a query =
  f (selectivities of query's base relations)

  – selectivity is the estimated percentage of rows of a relation used in producing the query result

# SQL Query Template
## [Q7 of TPC-H Benchmark]

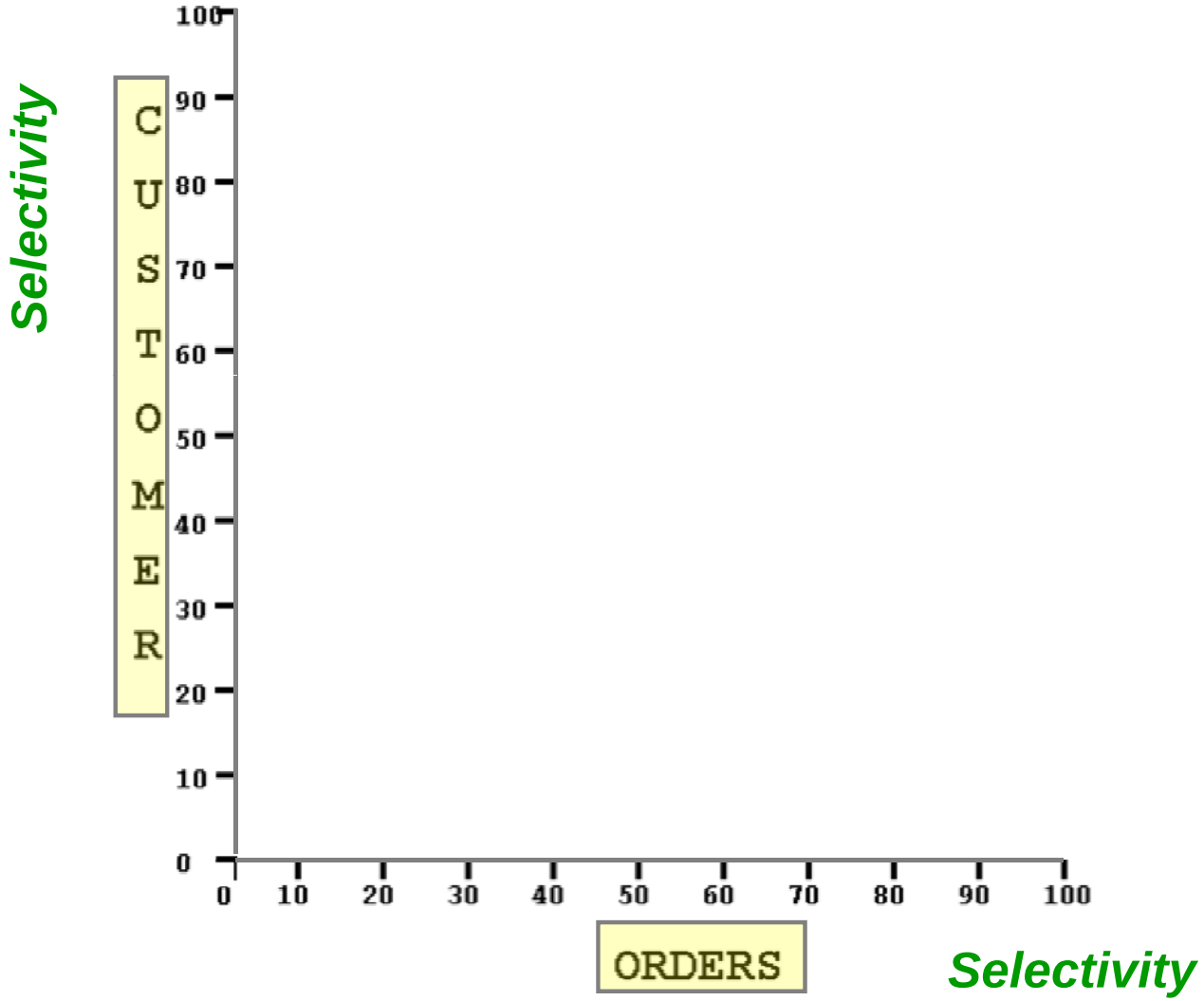
***Determines the values of goods shipped between nations in a time period***

```
select
    supp_nation, cust_nation, l_year, sum(volume) as revenue
from
(select n1.n_name as supp_nation,  n2.n_name as cust_nation,
        extract(year from l_shipdate) as l_year,
        l_extendedprice * (1 - l_discount) as volume
 from   supplier, lineitem, orders, customer, nation n1, nation n2
 wher                    okey and o_orderkey = l
                         key and s_nationkey = n
                         n2.n_nationkey and
        ((n1.n_name =  RANCE' and n2.n_name = 'GE     NY') or
         (n1.n_name = 'G RMANY' and n2.n_name = '     ANCE')) and
         l_shipdate between date '1995-01-01' and da  '1996-12-31'
         and  o_totalprice ≤ C1  and  c_acctbal ≤ C2   )  as shipping
 group by supp_nation, cust_nation, l_year
 order by supp_nation, cust_nation, l_year
```

Value determines selectivity of ORDERS relation

Value determines selectivity of CUSTOMER relation

# Relational Selectivity Space



# Plan and Cost Diagrams

- A plan diagram is a pictorial enumeration of the plan choices of the query optimizer over the relational selectivity space
- A cost diagram is a visualization of the (estimated) plan execution costs over the same relational selectivity space

# PICASSO

Picasso is a Java tool developed in our lab that automatically generates plan and cost diagrams

- ~50000 lines of code (2004-11) with ~100 classes
- Operational on DB2/Oracle/SQLServer/Sybase/PostgreSQL
- Copyrighted by IISc in May 2006
- Released as free software in Nov 2006 by Prof. N Balakrishnan, Associate Director of IISc
- Release of v1.0 in May 2007, v2.0 in Feb 2009, v2.1 in Feb 2011
- In use at academic and industrial labs worldwide
  - CMU, Purdue, Duke, TU Munich, NU Singapore, IIT-B, …
  - IBM, Microsoft, Oracle, Sybase, HP, …
- Received Best Software award in Very Large Data Base (VLDB) conference, 2010

# The Picasso Connection

*Woman with a guitar*

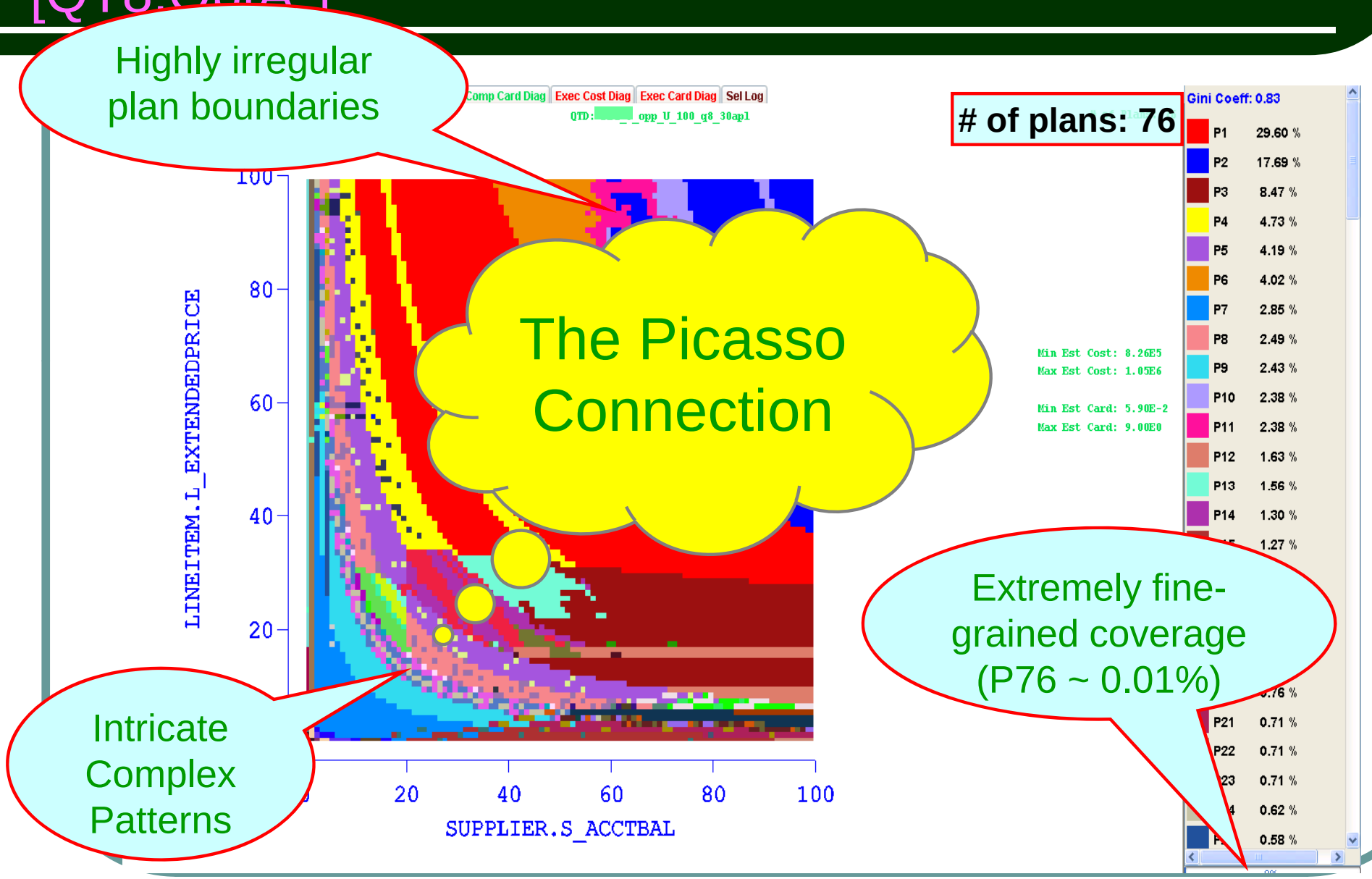
Georges Braque, 1913

Plan diagrams are often similar to cubist paintings !

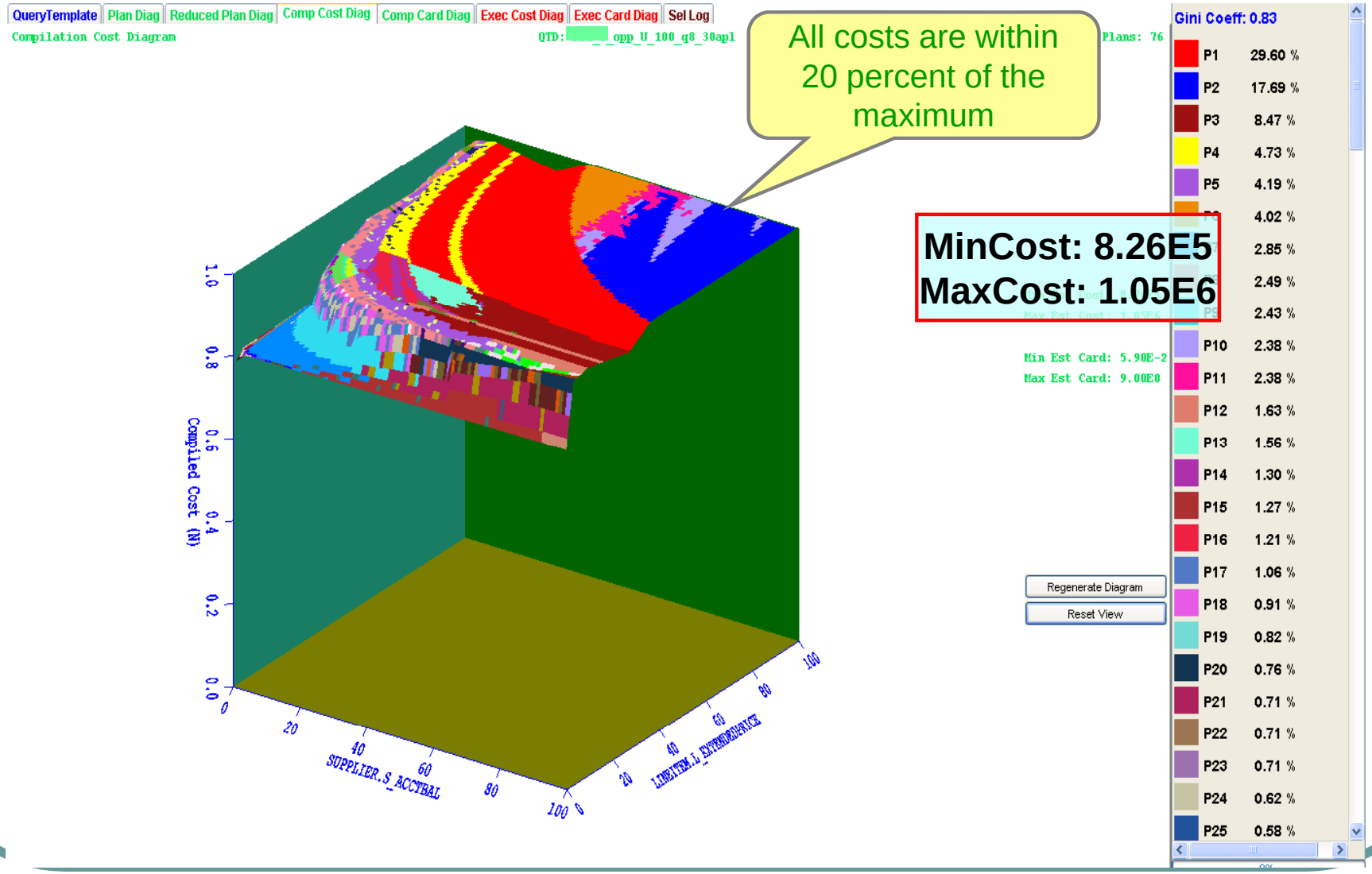
[ Pablo Picasso – founder of cubist genre ]

# Complex Plan Diagram

## [QT8.OptA*]



# Cost Diagram
## [QT8, Opt A*]




# PHENOMENA:
# PLAN DIAGRAM REDUCTION

# Problem Statement

Can the plan diagram be recolored with a smaller set of colors (i.e. some plans are "swallowed" by others), such that

Guarantee:

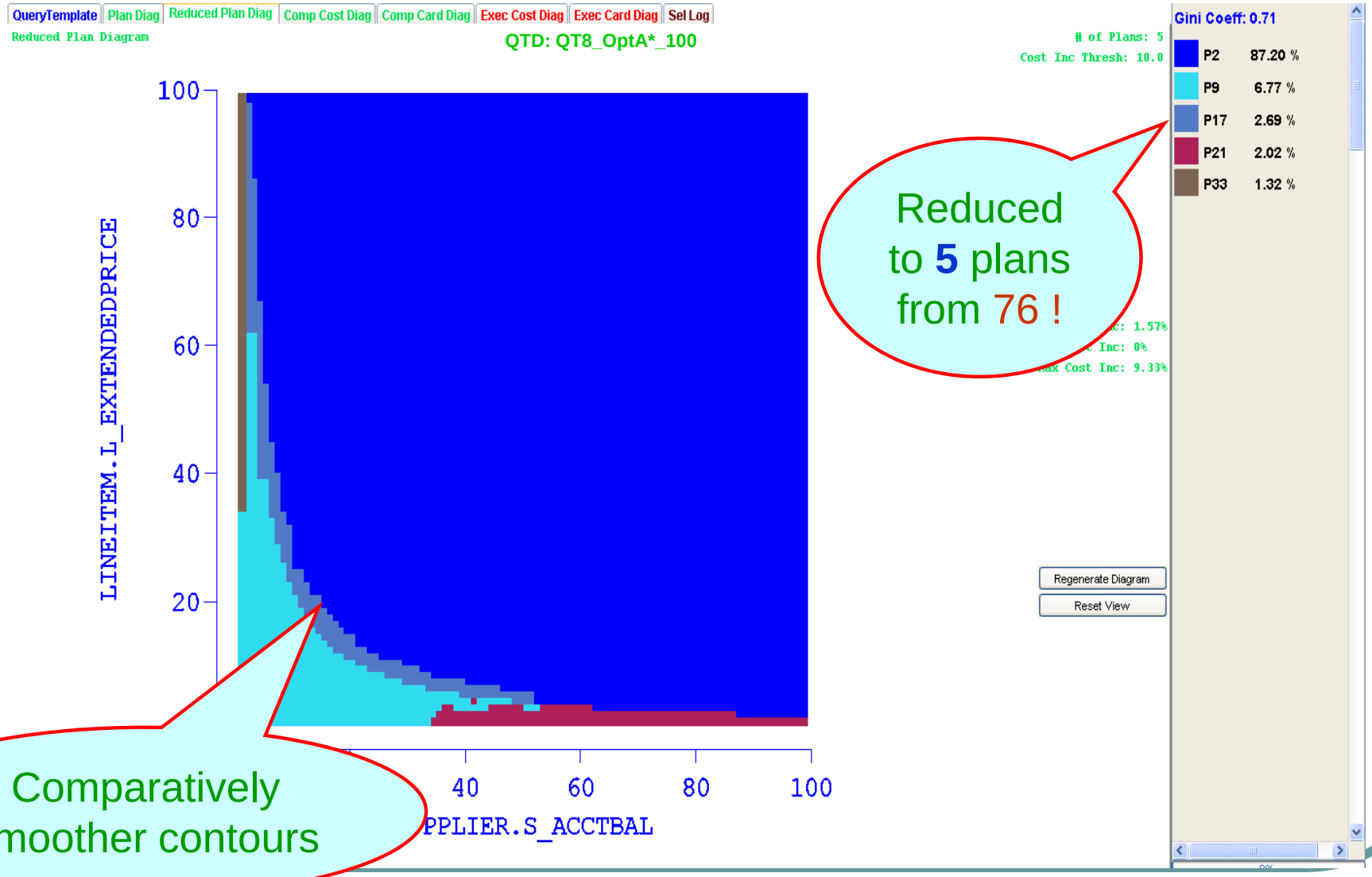
*No query point in the original diagram has its estimated cost increased, post-swallowing, by more than $\lambda$ percent* (user-defined)

Analogy:

(with due apologies to Sri Lankans in the audience)
Sri Lanka agrees to be annexed by India if it is assured that the cost of living of each Lankan citizen is not increased by more than $\lambda$ percent

# Reduced Plan Diagram [λ=10%]

## [QT8, OptA*, Res=100]



# Anorexic Reduction

Extensive empirical evaluation with a spectrum of multi-dimensional query templates indicates that

"With a cost-increase-threshold of just 20%, virtually all complex plan diagrams
[irrespective of query templates, data distribution, query distribution, system configurations, etc.]

reduce to "anorexic levels" (~10 or less plans)!


# APPLICATION: ROBUST PLANS

# Selectivity Estimation Errors

$q_e(x_e, y_e)$ : estimated location by optimizer

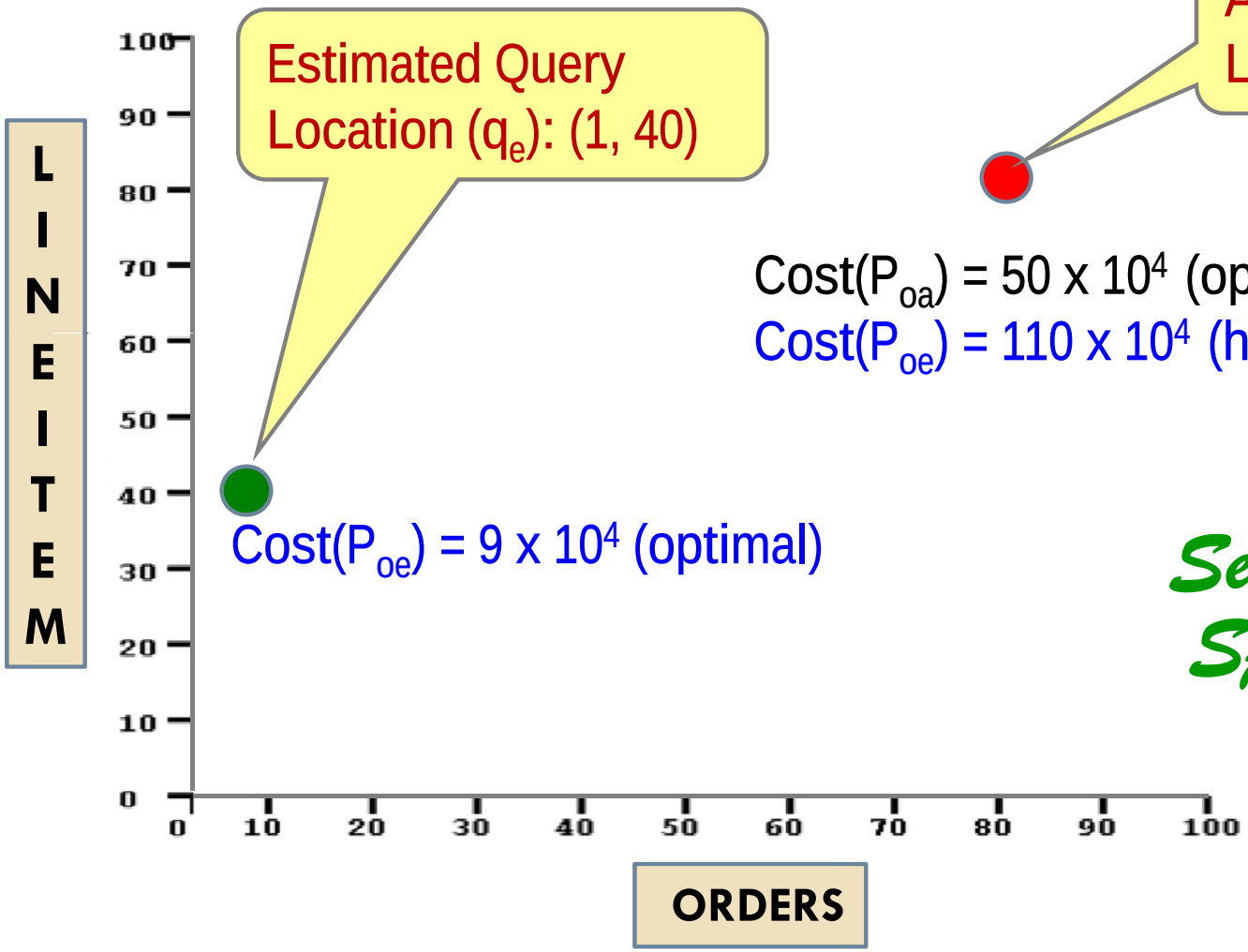
$q_a(x_a, y_a)$: actual location during execution

The difference could be substantial due to

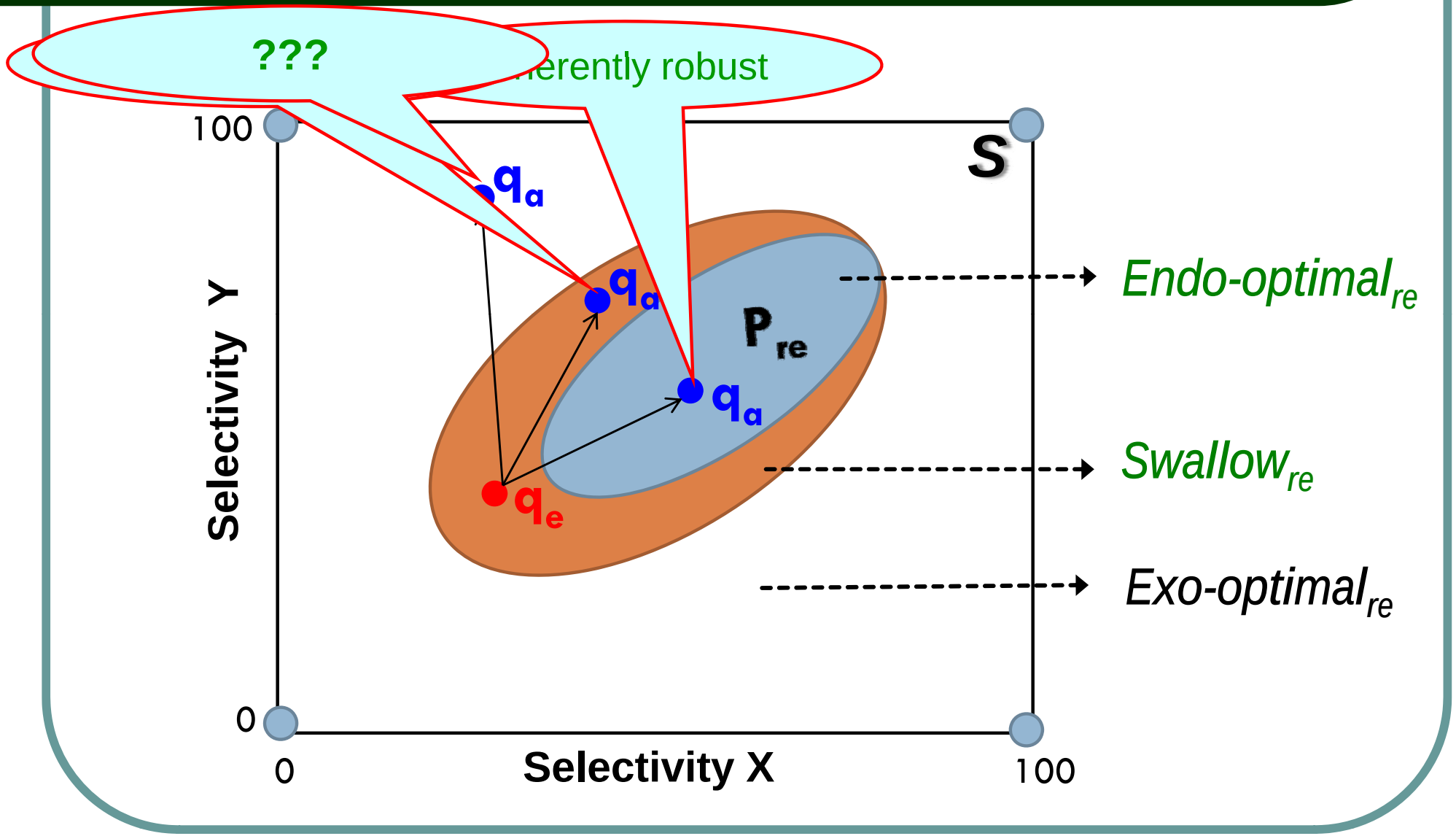
- Outdated Statistics (expensive to maintain)
- Coarse Summaries (histograms)
- Attribute Value Independence (AVI) assumptions

- Chronic problem in database design

# Impact of Error Example



# Error Locations wrt Plan Replacement Regions

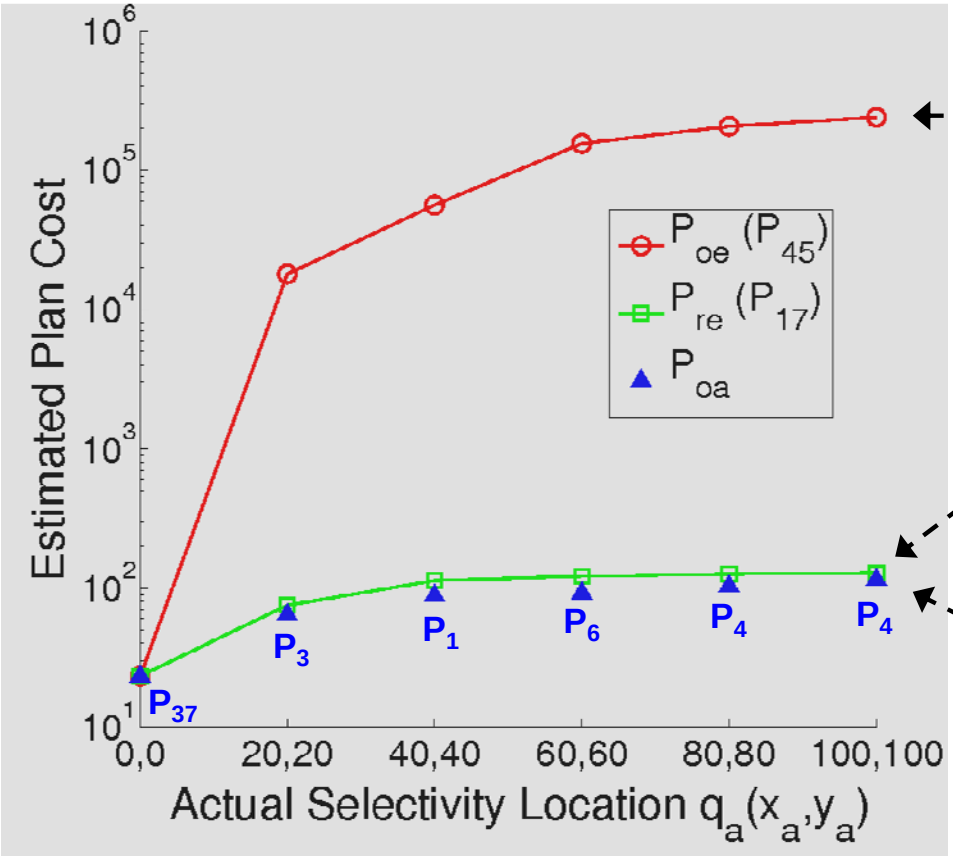


# Positive Impact of Reduction

In most cases, replacement plan provides robustness to selectivity errors even in exo-optimal region

QT5

$q_e$ = (0.36, 0.05)




Original Plan ($P_{oe}$)

Replacement Plan ($P_{re}$)

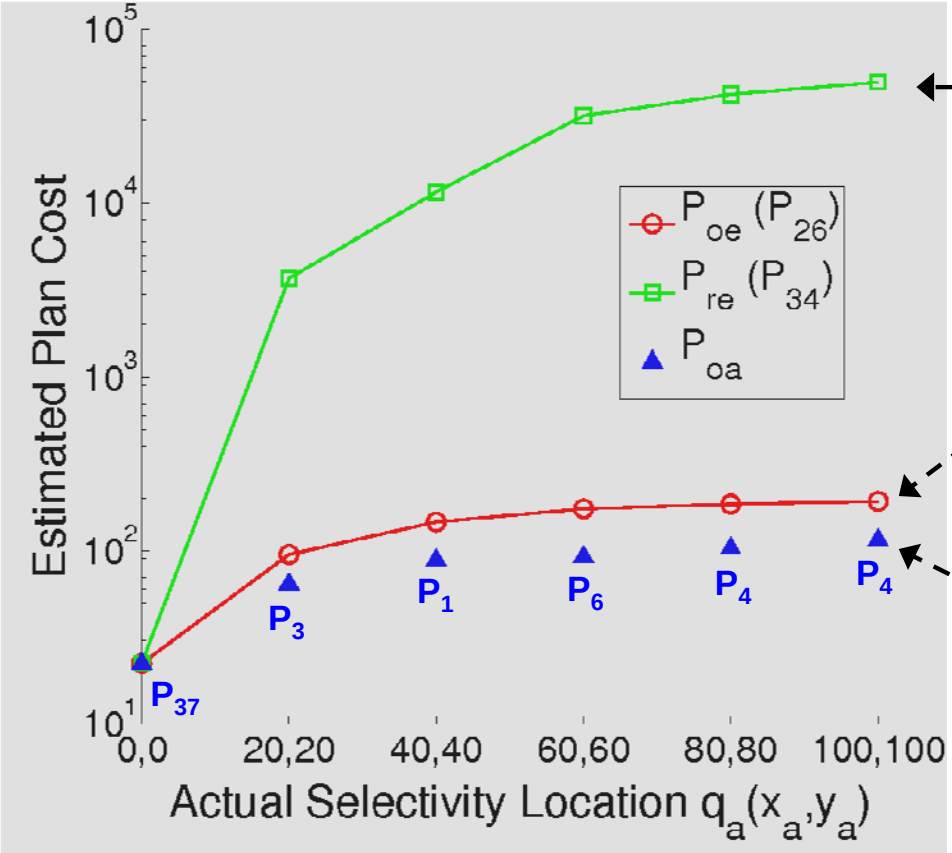
Local Optimal Plan ($P_{oa}$)

# Negative Impact of Reduction

But, occasionally, the replacement is much worse than the original plan !

QT5

$q_e$ = (0.03, 0.14)




Replacement Plan ($P_{re}$)

Original Plan ($P_{oe}$)

Local Optimal Plan ($P_{oa}$)

# Research Challenge

How do we ensure that plan replacements can only help, but never materially hurt the expected performance?

# Our Contribution

Mathematical analysis to show that only the perimeter of the selectivity space suffices to determine global safety

Border Safety ⇒ Interior Safety !

# Take Away

We can efficiently produce plan diagrams that simultaneously possess the desirable properties of being online, anorexic, safe and robust.

This result could play a meaningful role in designing the next generation of database query optimizers.

# More Details ...

http://dsl.serc.iisc.ernet.in/projects/PICASSO

Publications, Software, Sample Diagrams

# Dedication

- **Name:** Dr B K Narayana Rao (Founding Fellow)
  **Born:** 1881 **Elected:** 1934 **Section:** Medicine
  **Council Service:** 1934-61; Vice President 1938-43

- **Name:** Prof T S Subbaraya
  **Born:** 1905 **Elected:** 1935 **Section:** Physics

- **Name:** Dr B N Balakrishna Rao
  **Born:** 1910 **Elected:** 1945 **Section:** Medicine


# QUESTIONS?


# END PRESENTATION