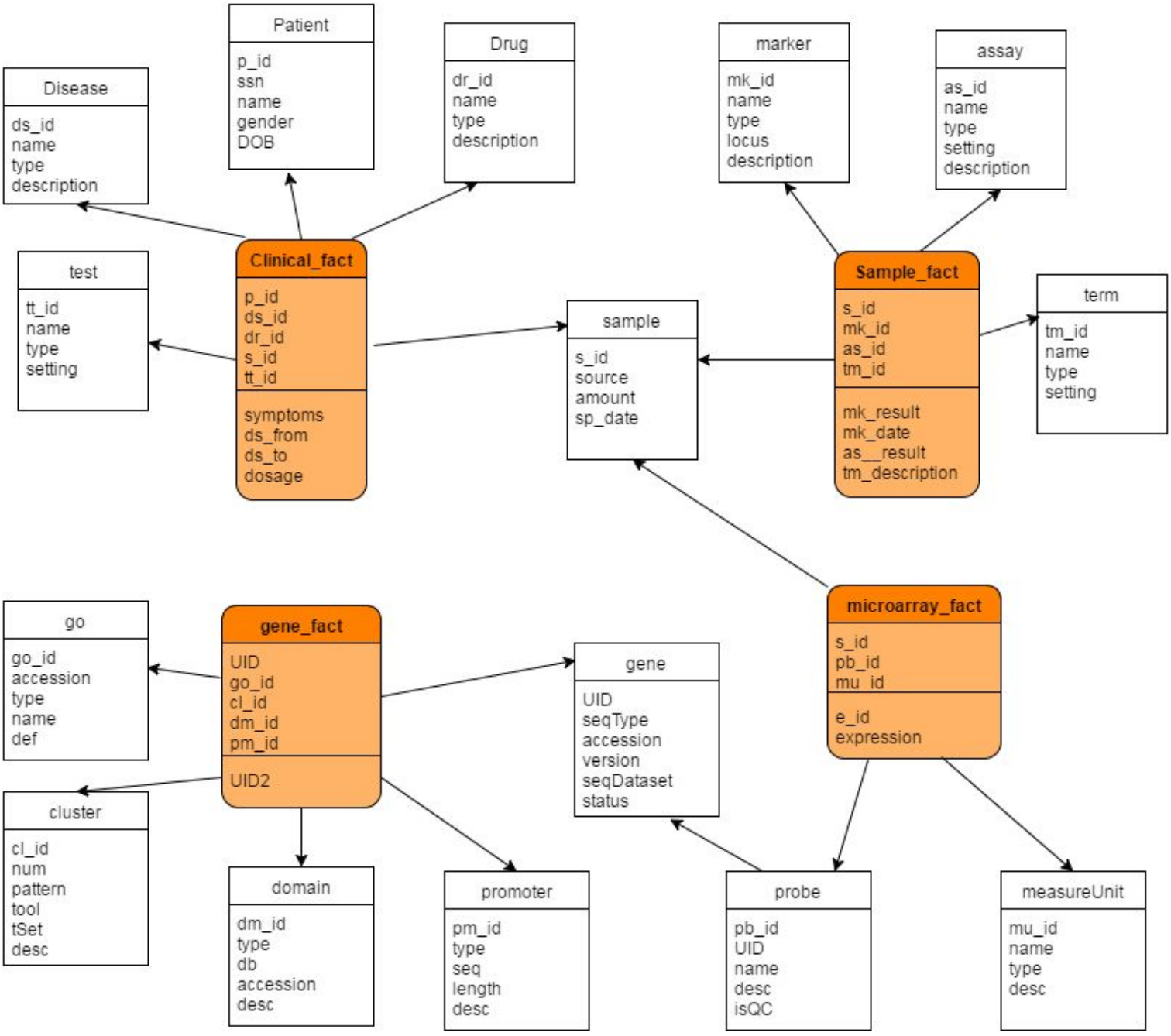
University at Buffalo
The State University of New York

# BIOLAP

By:
Akshay Kumar (50169103, akumar34)
Priyanka Singh (50169994, psingh28)
Sahil Dureja (50168872, sahildur)

# Part I: Implementation

The implementation of our schema is as follows:



- We chose this schema instead of star schema as dividing the fact table into several fact tables area wise not only save space it makes the query a little efficient as there would be less rows to process.
- Further dividing the fact tables would lead into more number of joins thereby increasing the query time.
- Also, further extending the dimension tables like in snowflake schema would have led to more number of joins and therefore making the queries less efficient.
- Further we have tried to increase the efficiency by creating a temp table which contained p_id, UID and probe_id. This helped when we tried to find the pearson correlation. As the query was running within a loop instead of processing the whole query every time, the non-changing part of the query was stored in the temp table and only a part of the query was running in the loop.

We have implemented the schema given with some changes to increase the efficiency of some queries. Like for queries trying to relate patient and disease with another entity like drug direct querying does not work because of many of the values being null instead of repeated data. For this we saved the patient-disease information in a separate table. This table can be joined with other fact table. This can also be done by nested query but former approach is faster as the required lookup is already stored.

# Part II: Sample Queries

### Question1:

**Query A:**

```
SELECT COUNT(*) FROM testtable4 as tt4
        INNER JOIN disease as d
                ON tt4.ds_id = d.ds_id
                WHERE d.description = "tumor";
```

**Result:**

| COUNT(*) |
| --- |
| 53 |

**Query B:**

```
SELECT COUNT(*) FROM testtable4 as tt4
        INNER JOIN disease as d
                ON tt4.ds_id = d.ds_id
                WHERE d.type = "leukemia";
```

**Result:**

| COUNT(*) |
| --- |
| 27 |

**Query C:**

```
SELECT COUNT(*) FROM testtable4 as tt4
        INNER JOIN disease as d
                ON tt4.ds_id = d.ds_id
                WHERE d.name = "ALL";
```

**Result:**

| COUNT(*) |
| --- |
| 13 |

## Question 2:

**Query:**

```
SELECT DISTINCT(drug.type) FROM `testtable4` AS tt4
        INNER JOIN disease AS d
                ON tt4.ds_id=d.ds_id
        INNER JOIN clinical_fact AS cf
                ON tt4.p_id=cf.p_id
        INNER JOIN drug
                ON drug.dr_id=cf.dr_id
                WHERE d.description = "tumor";
```

**Result:**

20 results

| type |
| --- |
| Drug Type 017 |
| Drug Type 005 |
| Drug Type 016 |
| Drug Type 002 |
| Drug Type 010 |
| Drug Type 009 |
| Drug Type 008 |
| Drug Type 007 |
| Drug Type 019 |
| Drug Type 001 |
| Drug Type 004 |
| Drug Type 006 |
| Drug Type 014 |
| Drug Type 013 |

| |
|---|
| Drug Type 020 |
| Drug Type 015 |
| Drug Type 003 |
| Drug Type 018 |
| Drug Type 012 |
| Drug Type 011 |

### Question3:

**Query:**

```
SELECT tt3.p_id, tt3.s_id, tt4.name, mf.s_id, mf.exp, pb.UID FROM testtable3 AS tt3
    INNER JOIN testtable4 AS tt4 ON tt3.p_id = tt4.p_id
    INNER JOIN microarray_fact AS mf ON tt3.s_id = mf.s_id
    INNER JOIN probe AS pb ON mf.pb_id = pb.pb_id
    INNER JOIN gene_fact AS gf ON pb.UID = gf.UID
    WHERE tt4.name = "ALL"
    AND mf.mu_id = "001"
    AND gf.cl_id = "00002";
```

**Result:**

325 results

| p_id | s_id | name | exp | UID |
|---|---|---|---|---|
| 47880 | 973218 | ALL | 36 | 2616922 |
| 47880 | 973218 | ALL | 102 | 55191224 |
| 47880 | 973218 | ALL | 142 | 51323880 |
| 47880 | 973218 | ALL | 42 | 58732744 |
| 47880 | 973218 | ALL | 115 | 40425172 |
| 47880 | 973218 | ALL | 179 | 45666597 |
| 47880 | 973218 | ALL | 177 | 71758989 |
| 47880 | 973218 | ALL | 133 | 88305510 |

## Question 4:

**Query 1:**

SELECT AVG( mf.exp ) , VAR_SAMP( mf.exp ) , COUNT( mf.exp ) FROM `testtable3` AS tt3 INNER JOIN `testtable4` AS tt4 ON tt3.p_id = tt4.p_id INNER JOIN microarray_fact AS mf ON tt3.s_id = mf.s_id INNER JOIN probe ON probe.pb_id = mf.pb_id INNER JOIN testtable5 AS tt5 ON tt5.UID = probe.UID WHERE tt4.name = "ALL" AND tt5.go_id = "0012502"

**Query 2:**

SELECT AVG( mf.exp ) , VAR_SAMP( mf.exp ) , COUNT( mf.exp ) FROM `testtable3` AS tt3 INNER JOIN `testtable4` AS tt4 ON tt3.p_id = tt4.p_id INNER JOIN microarray_fact AS mf ON tt3.s_id = mf.s_id INNER JOIN probe ON probe.pb_id = mf.pb_id INNER JOIN testtable5 AS tt5 ON tt5.UID = probe.UID WHERE tt4.name != "ALL" AND tt5.go_id = "0012502"

**Result:**
t = 1.0071347875363
p(one tailed) = 0.15703092528161
p(two tailed) = 0.31406185056321

## Question 5:

**Query 1:**

SELECT tt4.name, AVG(mf.exp) , COUNT(*) , SUM(mf.exp) FROM testtable4 AS tt4 INNER JOIN testtable3 AS tt3 ON tt4.p_id = tt3.p_id INNER JOIN microarray_fact AS mf ON tt3.s_id = mf.s_id INNER JOIN probe ON mf.pb_id = probe.pb_id INNER JOIN testtable5 AS tt5 ON probe.UID = tt5.UID WHERE tt5.go_id = "0007154" AND ( tt4.name = "ALL"OR tt4.name = "AML"OR tt4.name = "Breast tumor"OR tt4.name = "Colon tumor" ) GROUP BY tt4.name

**Query 2 (Run for each selected disease):**

SELECT mf.exp FROM testtable4 AS tt4 INNER JOIN testtable3 AS tt3 ON tt4.p_id = tt3.p_id INNER JOIN microarray_fact AS mf ON tt3.s_id = mf.s_id INNER JOIN probe ON mf.pb_id = probe.pb_id INNER JOIN testtable5 AS tt5 ON probe.UID = tt5.UID WHERE tt5.go_id = "0007154" AND tt4.name = "Colon tumor"

**Result:**
SSd = 30415.354295497, SSe = 3165344.330037
MSd = 10138.451431832, MSe = 3229.9431939153
F = 3.1388946563926

## Question 6:

**For one same disease (ALL)**

**Query 1:**

SELECT distinct clinical_fact.p_id FROM clinical_fact join disease on clinical_fact.ds_id=disease.ds_id left join (SELECT p_id,s_id FROM clinical_fact where clinical_fact.s_id!="null") as samplejoined on samplejoined.p_id=clinical_fact.p_id left join microarray_fact on microarray_fact.s_id=samplejoined.s_id left join probe on probe.pb_id=microarray_fact.pb_id left join gene_fact on gene_fact.UID=probe.UID where disease.name="ALL" and gene_fact.go_id="0012502"

**Query 2 (Run for each patient with disease):**

SELECT exp FROM clinical_fact join disease on clinical_fact.ds_id=disease.ds_id left join (SELECT p_id,s_id FROM clinical_fact where clinical_fact.s_id!="null") as samplejoined on samplejoined.p_id=clinical_fact.p_id left join microarray_fact on microarray_fact.s_id=samplejoined.s_id left join probe on probe.pb_id=microarray_fact.pb_id left join gene_fact on gene_fact.UID=probe.UID where disease.name="ALL" and gene_fact.go_id="0012502" AND clinical_fact.p_id="77689" ORDER BY probe.pb_id ASC

**Result:**

Sum of all possible correlations = 7.689842120158
Count of N1 X (N1 - 1) /2 = 78
Average Correlation = 0.098587719489205

**For one disease with all the rest:**

**Query 1A:**

SELECT distinct clinical_fact.p_id FROM clinical_fact join disease on clinical_fact.ds_id=disease.ds_id left join (SELECT p_id,s_id FROM clinical_fact where clinical_fact.s_id!="null") as samplejoined on samplejoined.p_id=clinical_fact.p_id left join microarray_fact on microarray_fact.s_id=samplejoined.s_id left join probe on probe.pb_id=microarray_fact.pb_id left join gene_fact on gene_fact.UID=probe.UID where disease.name="ALL" and gene_fact.go_id="0012502"

**Query 1B (for each patient with disease):**

SELECT exp FROM clinical_fact join disease on clinical_fact.ds_id=disease.ds_id left join (SELECT p_id,s_id FROM clinical_fact where clinical_fact.s_id!="null") as samplejoined on samplejoined.p_id=clinical_fact.p_id left join microarray_fact on microarray_fact.s_id=samplejoined.s_id left join probe on probe.pb_id=microarray_fact.pb_id left join gene_fact on gene_fact.UID=probe.UID where disease.name="ALL" and gene_fact.go_id="0012502" AND clinical_fact.p_id="77689" ORDER BY probe.pb_id ASC

**Query 2A:**

SELECT distinct clinical_fact.p_id FROM clinical_fact join disease on clinical_fact.ds_id=disease.ds_id left join (SELECT p_id,s_id FROM clinical_fact where clinical_fact.s_id!="null") as samplejoined on samplejoined.p_id=clinical_fact.p_id left join microarray_fact on microarray_fact.s_id=samplejoined.s_id left join probe on probe.pb_id=microarray_fact.pb_id left join gene_fact on gene_fact.UID=probe.UID where disease.name="AML" and gene_fact.go_id="0012502"

**Query 2B (for each patient with disease):**

SELECT exp FROM clinical_fact join disease on clinical_fact.ds_id=disease.ds_id left join (SELECT p_id,s_id FROM clinical_fact where clinical_fact.s_id!="null") as samplejoined on samplejoined.p_id=clinical_fact.p_id left join microarray_fact on microarray_fact.s_id=samplejoined.s_id left join probe on probe.pb_id=microarray_fact.pb_id left join gene_fact on gene_fact.UID=probe.UID where disease.name="AML" and gene_fact.go_id="0012502" AND clinical_fact.p_id="48802" ORDER BY probe.pb_id ASC

**Result:**

Sum of all possible correlations = -7.1455367404653
Count of N1 X N2 = 182
Average Correlation = -0.039261190881678

# Part III: Knowledge Discovery

### Finding Informative Genes (Patients with Disease):

**Query 1:**
SELECT probe.UID,AVG( microarray_fact.exp ) , VAR_SAMP( microarray_fact.exp ) , COUNT( microarray_fact.exp ) FROM clinical_fact join disease on clinical_fact.ds_id=disease.ds_id left join (SELECT p_id,s_id FROM clinical_fact where clinical_fact.s_id!="null") as samplejoined on samplejoined.p_id=clinical_fact.p_id left join microarray_fact on microarray_fact.s_id=samplejoined.s_id left join probe on probe.pb_id=microarray_fact.pb_id where disease.name="ALL" and samplejoined.s_id!="NULL" group by probe.UID

**Query 2 (Run for each patient with disease):**

SELECT probe.UID,AVG( microarray_fact.exp ) , VAR_SAMP( microarray_fact.exp ) , COUNT( microarray_fact.exp ) FROM clinical_fact join disease on clinical_fact.ds_id=disease.ds_id left join (SELECT p_id,s_id FROM clinical_fact where clinical_fact.s_id!="null") as samplejoined on samplejoined.p_id=clinical_fact.p_id left join microarray_fact on microarray_fact.s_id=samplejoined.s_id left join probe on probe.pb_id=microarray_fact.pb_id where disease.name!="ALL" and samplejoined.s_id!="NULL" group by probe.UID

**Result (After taking t-test and for p-value less than 0.01):**
No. of Informative genes: 38

### Classifying new patients:

**Finding rA values for Group A (with disease):**

| TEST1 | TEST2 | TEST3 | TEST4 | TEST5 |
|---|---|---|---|---|
| 0.802834887399 56 | 0.173629121316 53 | -0.165998214559 95 | 0.810212920574 14 | -0.139551971085 8 |
| 0.739950428140 6 | 0.240137154772 62 | -0.048497294838 268 | 0.721366240284 72 | -0.162658852504 96 |
| 0.836240203354 78 | 0.114763332103 83 | -0.016327282280 226 | 0.791851309645 46 | -0.119096728411 |
| 0.815521014430 19 | 0.127334607418 93 | 0.061845636291436 | 0.863751207132 35 | -0.112543782179 64 |

Similar 38 rows.

**Finding rA values for Group B (without disease):**

| TEST1 | TEST2 | TEST3 | TEST4 | TEST5 |
|---|---|---|---|---|
| -0.140802103473 03 | -0.254668248005 19 | -0.103502043019 14 | -0.081836200265 547 | -0.111421832653 07 |
| 0.018930905589 693 | -0.137224824819 22 | 0.053641992422 049 | -0.059533015721 666 | -0.195035087003 17 |
| -0.031511429717 815 | 0.249646214640 29 | -0.187848661882 6 | -0.026082963228 017 | 0.202563208692 4 |
| 0.196469636646 43 | -0.218579070416 43 | -0.083727826464 568 | 0.016484361608 417 | -0.135230200224 35 |

Similar 38 rows.

**Final Classification:**

| USER ID | p VALUE | CLASSIFCATION |
|---|---|---|
| test1 | 4.4408920985006E-16 | classfied as ALL |
| test2 | 3.2547485151468E-8 | classfied as ALL |
| test3 | 0.77357051847198 | NOT classfied as ALL |
| test4 | 4.4408920985006E-16 | classfied as ALL |
| test5 | 0.0038238124509271 | classfied as ALL |