+

Tutorial 6: Database Integration and Data Linkage

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

+ Question 1

- Discuss different roles database views play in the following systems:
  - Relational database
  - Distributed database
  - Data warehouse
  - Data integration

+ Q1: Database View

- View feature

```
CREATE      VIEW    SYSAN(ENO, ENAME)
AS          SELECT  ENO, ENAME
            FROM    EMP
            WHERE   TITLE = “Prof.”
```

  - “Virtual tables”
    - Doesn’t store data
    - View definition: view name and retrieval query

+ Q1: Database View

- Relational DBMS
  - Easy to retrieve data
    - Present a subset of the data contained in a table
    - Act as aggregated tables
  - Access control
    - Limit the degree of exposure
    - External view
- Distributed database system
  - Hide the complexity of the underlining distributed system
  - Provide different view for different users on the same distributed system
  - As data integration/fragmentation method
  - External view

+ Q1: Database View

- Data warehousing system
  - Precomputed cubes in data warehouse
  - Act as aggregated tables
  - Easy to retrieve, query fast
- Data integration
  - Defining global schema *(front end)*
    - Integrate local schemas and support user queries

+ Question 2

- We consider the following three parties related to the Olympics information system involving swimming events:
  - Local Organisation Committee (**LOC**) for an Olympiad (e.g., the London Game in 2012)
    - **Result**(EventID, CompID, Position, Time)
  - International Olympic Committee (**IOC**)
    - **Competitor**(ID, Country, Name)
    - **OlympicRecord**(EventID, CompID, Olympiad, Time)
  - **FINA**, international swimming federation
    - **Athlete**(ID, Country, Name)
    - **WorldRecord**(EventID, AthID, Year, Time)

+ Q2: Olympic Information System

- Local Organisation Committee (LOC)
- International Olympic Committee (IOC)
- International Swimming Federation (FINA)

LOC

Result
EventID
CompID
Position
Time

FINA

Athlete
ID
Country
Name

World Record
EventID
AthID
Year
Time

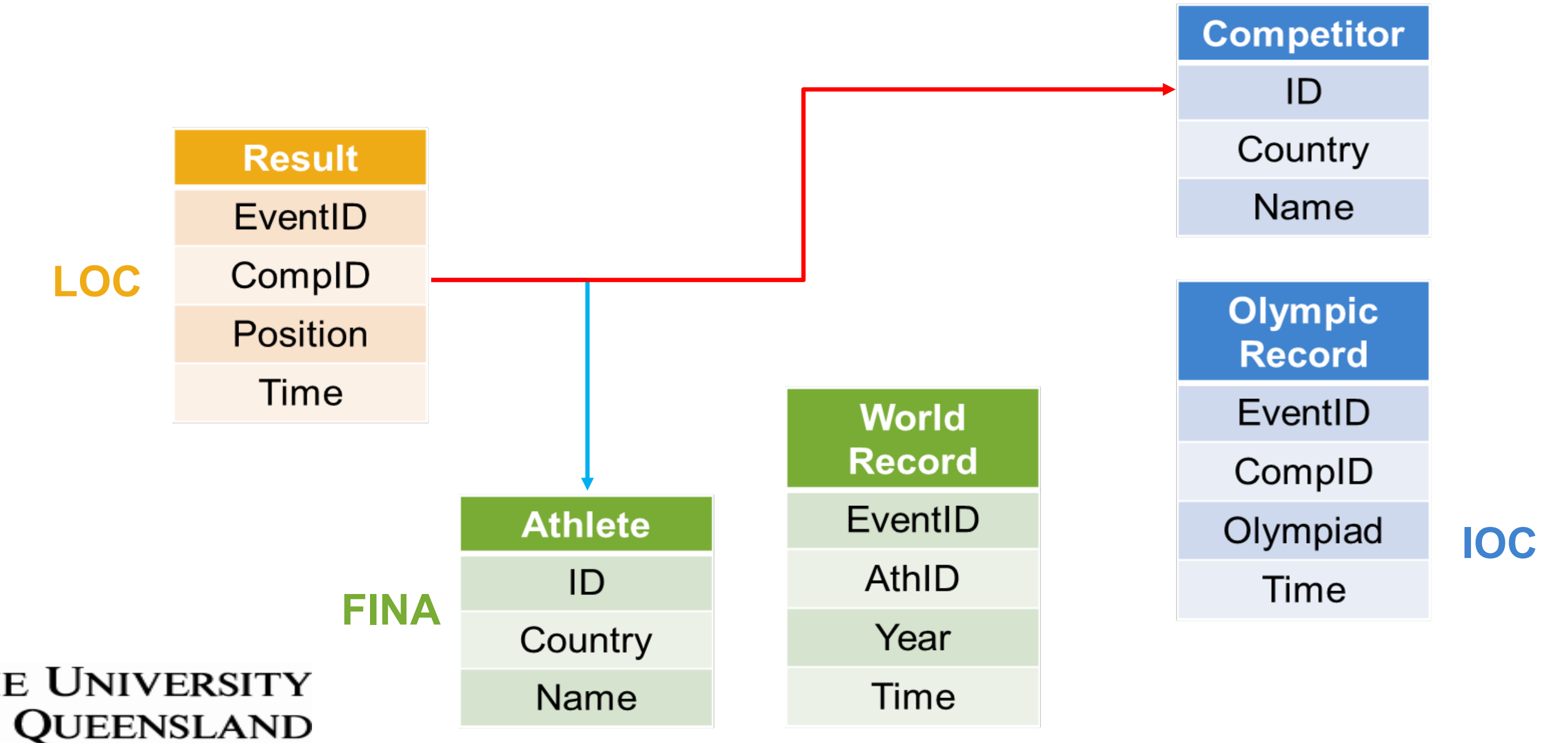
IOC

Competitor
ID
Country
Name

Olympic Record
EventID
CompID
Olympiad
Time

Q2: Olympic Information System

- All competitors participating the Game organised by the **LOC** are registered with the **IOC**, and also **FINA**.



LOC

Result
EventID
CompID
Position
Time

FINA

Athlete
ID
Country
Name

World Record
EventID
AthID
Year
Time

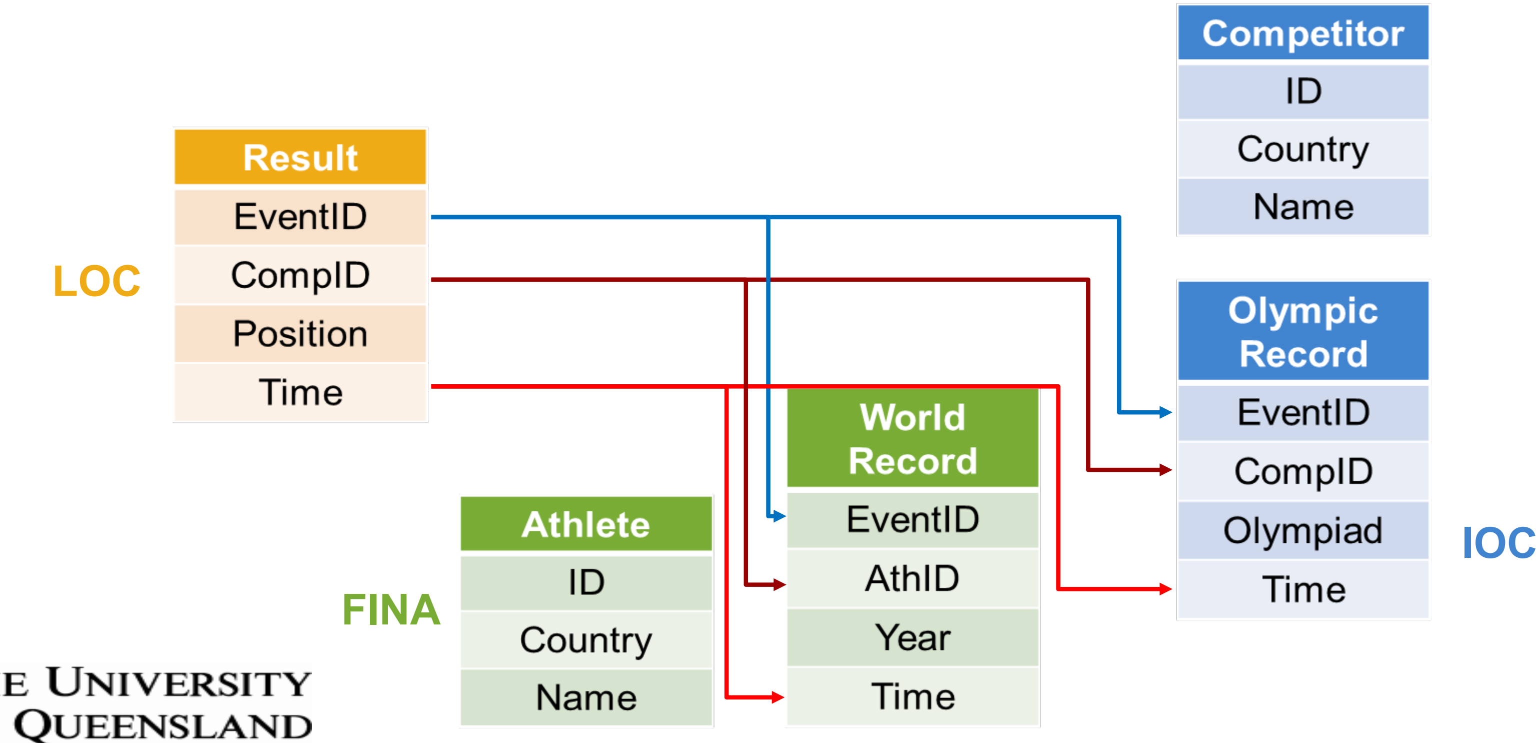
IOC

Competitor
ID
Country
Name

Olympic Record
EventID
CompID
Olympiad
Time

+ Q2: Olympic Information System

- Record-breaking game results are saved in both LOC, IOC and FINA.



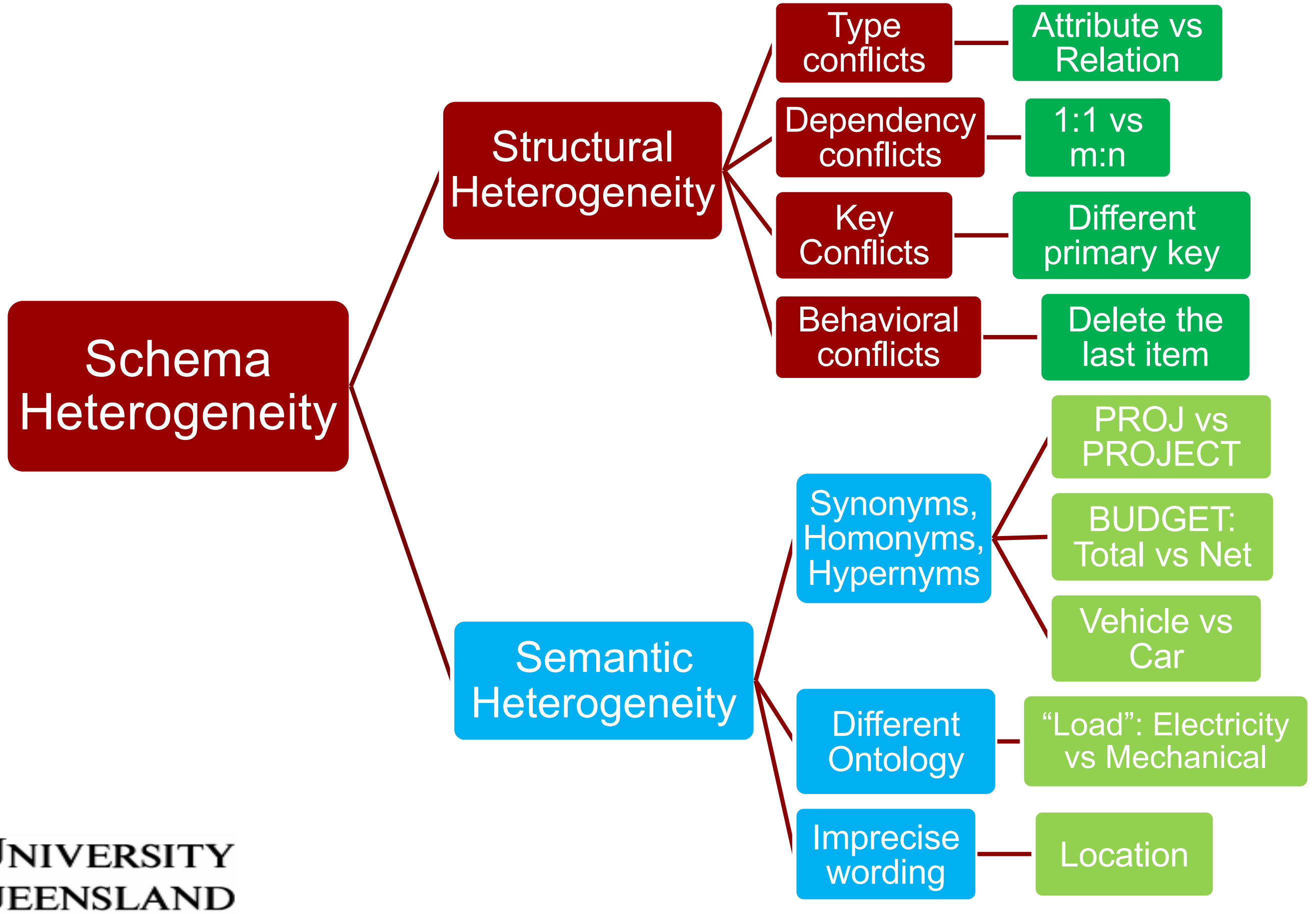
The University of Queensland
AUSTRALIA

+ Q2: Olympic Information System

- (a) Identify possible **semantic heterogeneity** when integrating these three independently developed databases into GoldMedalist, and discuss possible solutions.

+

Semantic Heterogeneity



THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

+ Q2-a

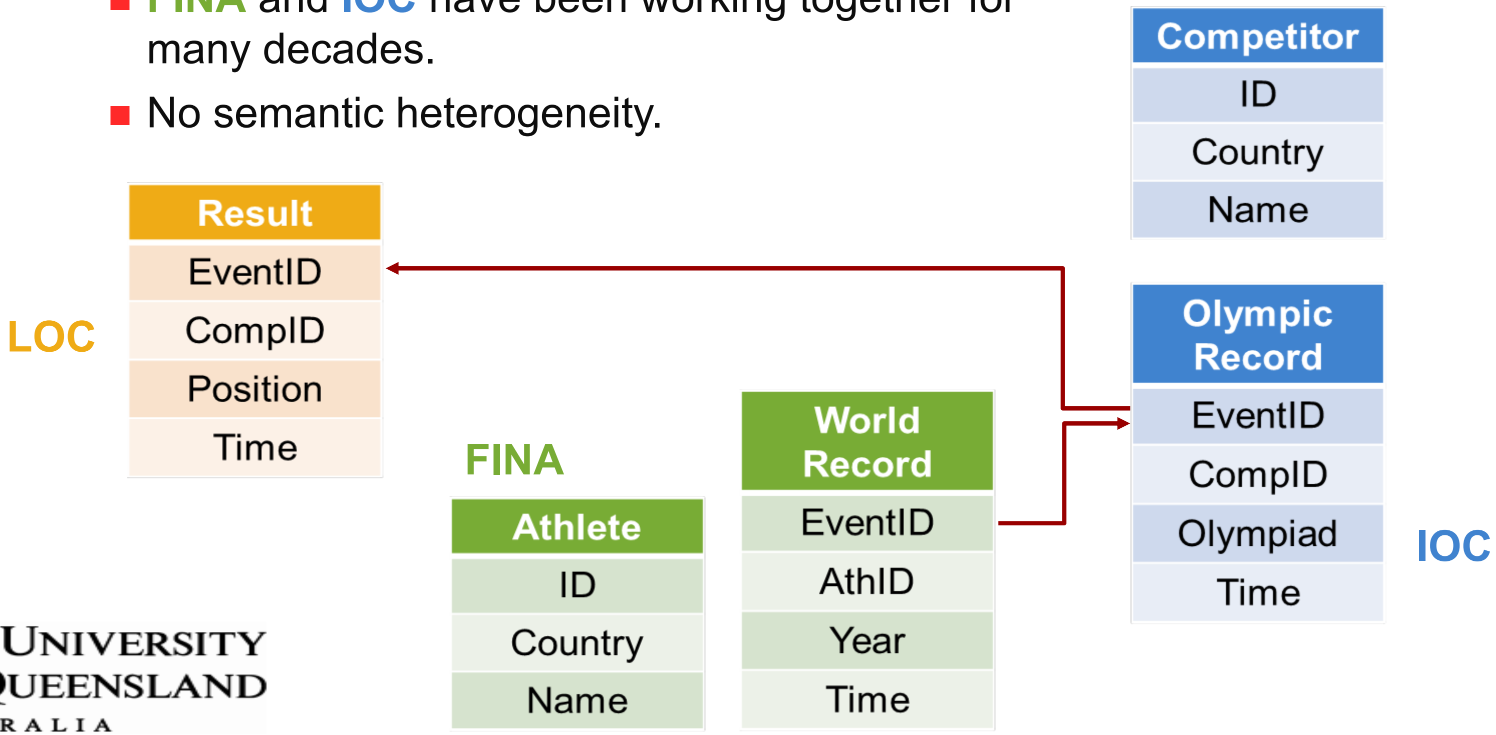
- Example (Integration)
  - WORKER(wNumber, name, title, salary)
    PROJECT(pNumber, pName, budget)
    CLIENT(cName, address)
    WORKS IN(wNumber, pNumber, responsibility, duration)
    CONTRACTED BY(pNumber, cName, contractNo)
  - EMP(eNo, eName, title)
    PROJ(pNo, pName, budget, loc, cName)
    ASG(eNo, pNo, resp, dur)
    PAY(title, sal)

+ Q2-a

GoldMedalist(CompID, EventID, Time)

at least 3 attributes

- EventID: Standardised by **FINA** and adopted by **IOC**.
  - **FINA** and **IOC** have been working together for many decades.
  - No semantic heterogeneity.



LOC

Result
EventID
CompID
Position
Time

FINA

Athlete
ID
Country
Name

World Record
EventID
AthID
Year
Time

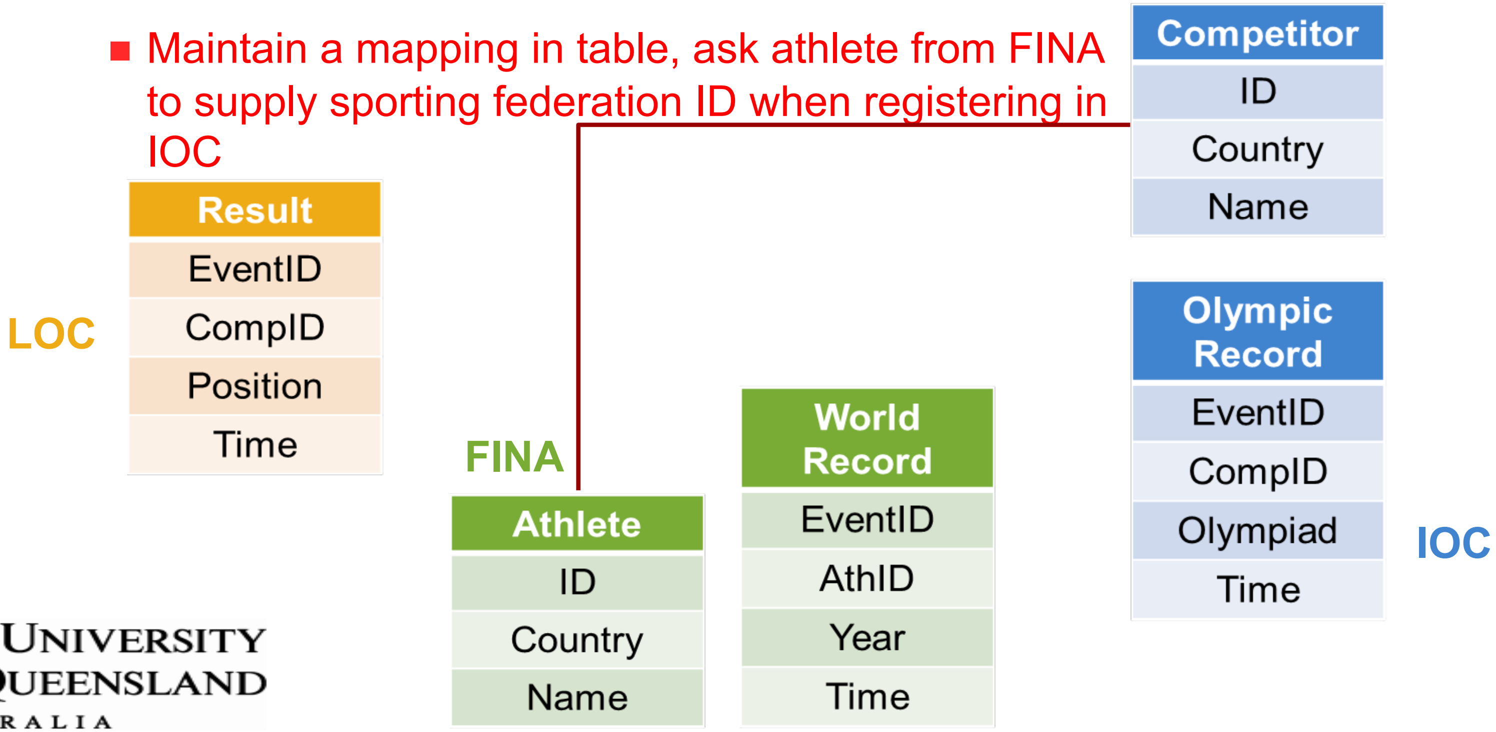
IOC

Competitor
ID
Country
Name

Olympic Record
EventID
CompID
Olympiad
Time

+ Q2-a

- FINA.Athlete.ID ≠ IOC.Competitor.ID
  - They organise ID in different ways before they start collaboration
  - Maintain a mapping in table, ask athlete from FINA to supply sporting federation ID when registering in IOC



LOC

Result
EventID
CompID
Position
Time

FINA

Athlete
ID
Country
Name

World Record
EventID
AthID
Year
Time

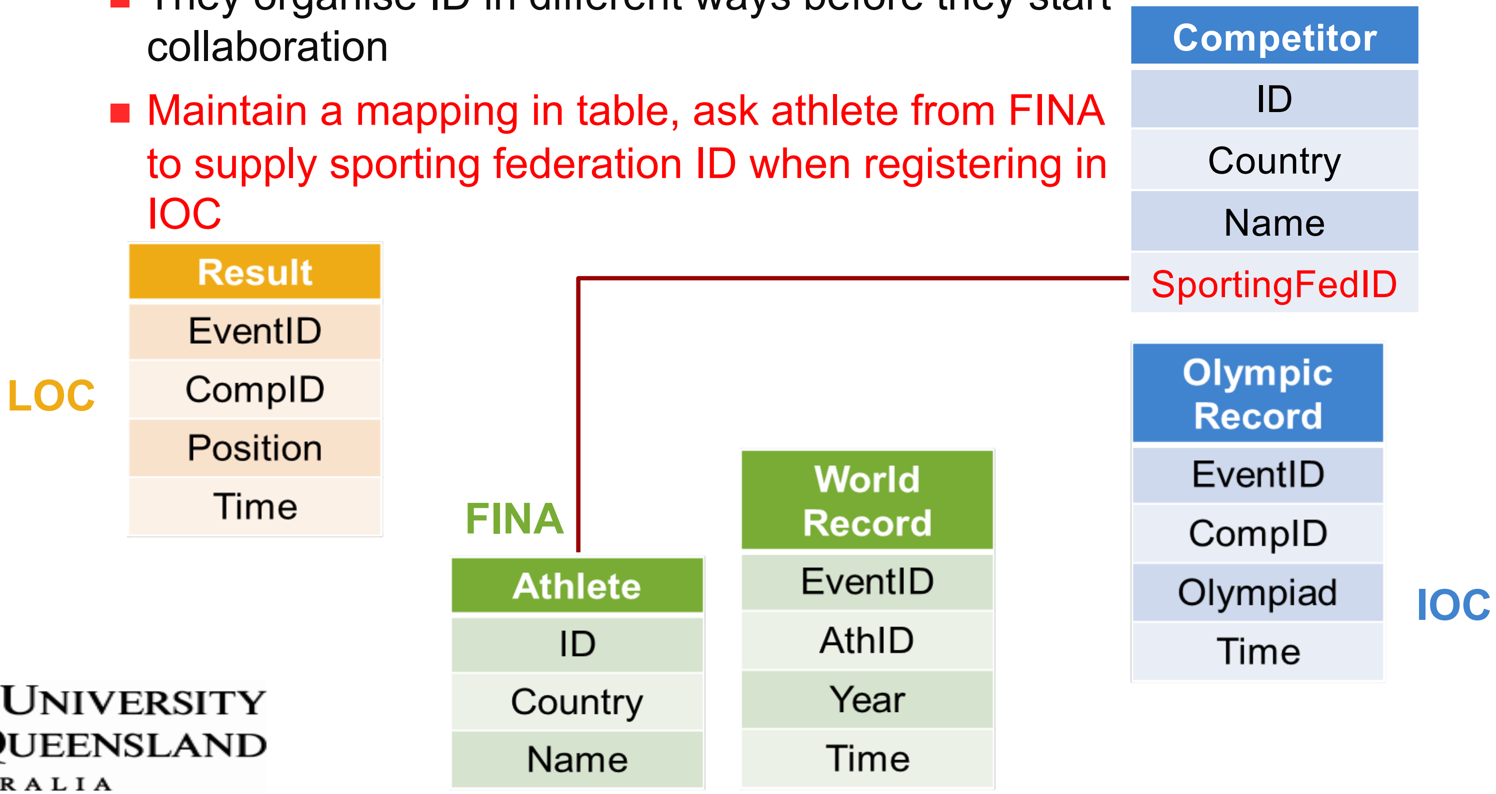
IOC

Competitor
ID
Country
Name

Olympic Record
EventID
CompID
Olympiad
Time

+ # Q2-a

- FINA.Athlete.ID ≠ IOC.Competitor.ID
  - They organise ID in different ways before they start collaboration
  - Maintain a mapping in table, ask athlete from FINA to supply sporting federation ID when registering in IOC



LOC

Result
EventID
CompID
Position
Time

FINA

Athlete
ID
Country
Name

World Record
EventID
AthID
Year
Time

IOC

Competitor
ID
Country
Name
SportingFedID

Olympic Record
EventID
CompID
Olympiad
Time

+ Q2-b

- (b) Now the LOC wants to integrate these three databases into the following table that shows all swimming event Gold medalists in the Game:
  - GoldMedalist(CompID, EventID, Time, OlympicRecord, WorldRecord)

LOC

Result
EventID
CompID
Position
Time

FINA

Athlete
ID
Country
Name

World Record
EventID
AthID
Year
Time

IOC

Olympic Record
EventID
CompID
Olympiad
Time

Competitor
ID
Country
Name

+ Q2-b

- Now the LOC wants to integrate these three databases into the following table that shows all swimming event Gold medalists in the Game:
  - GoldMedalist(CompID, EventID, Time, OlympicRecord, WorldRecord)

- Use SQL to construct GoldMedallist.
  - **Result**(EventID, CompID, Position, Time)
  - **OlympicRecord**(EventID, CompID, Olympiad, Time)
  - **WorldRecord**(EventID, AthID, Year, Time)

+ Q2-b

```
CREATE VIEW GoldMedalist
(CompID, EventID, Time, OlympicRecord, WorldRecord) AS
          SELECT B.CompID, B.EventID, B.Time, IR.Time, WR.Time
          FROM Result B, OlympicRecord IR, WorldRecord WR
          WHERE
                    IR.EventID = B.EventID AND
                    WR.EventID = B.EventID AND
                    B.Position = 1;
```

+ Q2-c

- (c) Assume the ABC Television wants to create the following table to show all the swimming records set at the Game organised by the LOC:
  - NewRecord(EventID, CompID, Record, Time)

  where Record is either "World" or "Olympic". Show an SQL query computing NewRecord.

- Solution
  - WorldRecord <= OlympicRecord
    - Break the WorldRecord
      - Time < WorldRecord
    - Break the OlympicRecord
      - Time >= WorldRecord and Time < OlympicRecord

+ Q2-c

```
CREATE VIEW NewRecord(EventID, CompID, Record, Time) AS
        SELECT EventID, CompID, "World", Time
        FROM GoldMedallist G
        WHERE Time < G.WorldRecord
 UNION
        SELECT EventID, CompID, "Olympic", Time
        FROM GoldMedallist G
        WHERE Time >= G.WorldRecord AND Time <    G.OlympicRecord
```

+ Q2-d

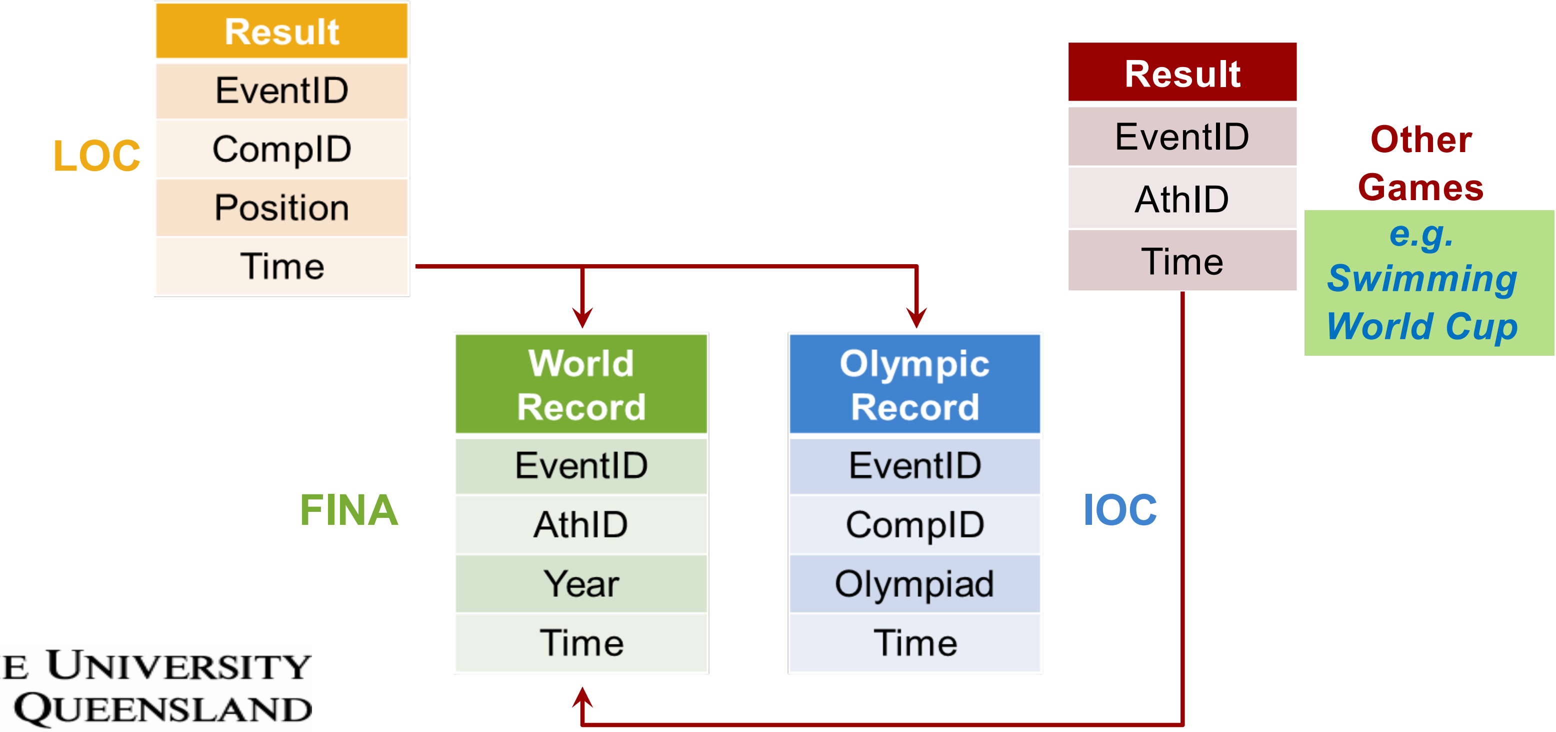
- (d) Assume that GoldMedalist is maintained by the LOC, with any new records updated to the OlympicRecord and WorldRecord tables are done by IOC and FINA respectively. It is a requirement that the Olympic records and World records in GoldMedalist must be accurate all the time. What "quality of service" guarantees do the IOC and FINA need to make to ensure such accuracy in GoldMedalist?

+ Q2-d

- GoldMedalist(CompID, EventID, Time, OlympicRecord, WorldRecord)

This is a VIEW !

- Both IOC and FINA need to guarantee the OlympicRecord and WorldRecord are the current.



LOC

Result
EventID
CompID
Position
Time

Other Games

e.g. Swimming World Cup

Result
EventID
AthID
Time

FINA

World Record
EventID
AthID
Year
Time

IOC

Olympic Record
EventID
CompID
Olympiad
Time

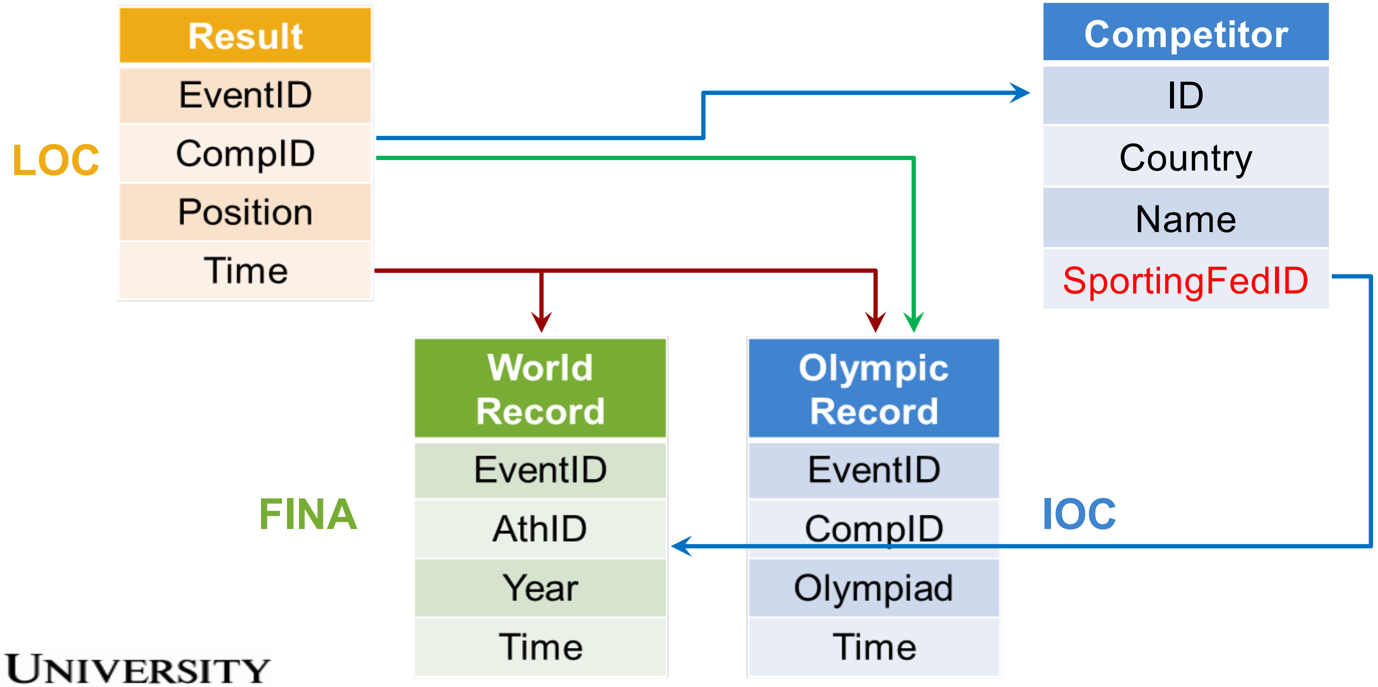
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

+ Q2-d

- GoldMedalist(CompID, EventID, Time, OlympicRecord, WorldRecord)

This is a VIEW !

- Any updates initiated from GoldMedalist to WorldRecord need to use the revised Competitor table



THE UNIVERSITY OF QUEENSLAND AUSTRALIA

+ Q3. Edit Distance

- The edit distance between two strings is the minimum number of operations to transform one string to another
  - Operations:

    delete, insert or substitution/replace one character
  - Normalization to $[0, 1]$
    - Divided by $max(|A|,\ |B|)$

- What's the edit distance?
  - 'John', 'Jon' $\mathbf{1/4}$
  - 'John', 'Josh' $\mathbf{2/4}$
  - 'Smith', 'Sitch' $\mathbf{2/5}$

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

+ Q3. Jaccard Distance

- Jaccard Distance
  - Based on string tokenization
    - Breaking a stream of ***text*** up into words, phrases, symbols, or other forms of elements called tokens
    - An n-gram is a contiguous sequence of n items from a given sequence of text or speech

It was the best of times

1-*gram* sequence

I, t, #, w, a, s, #, t, h, e, #, b, e, s, t, #, o, f, #, t, i, m, e, s

2-*gram* sequence

It, t#, #w, wa, as, s#, #t, th, he, e#, #b, be, es, st, t#, #o, of, f#, #t, ti, im, me, es

3-*gram* sequence

It#, t#w, #wa, was, as#, s#t, #th, the, he#, e#b, #be, bes, est, st#, t#o, #of, of#, f#t, #ti, tim, ime, mes

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

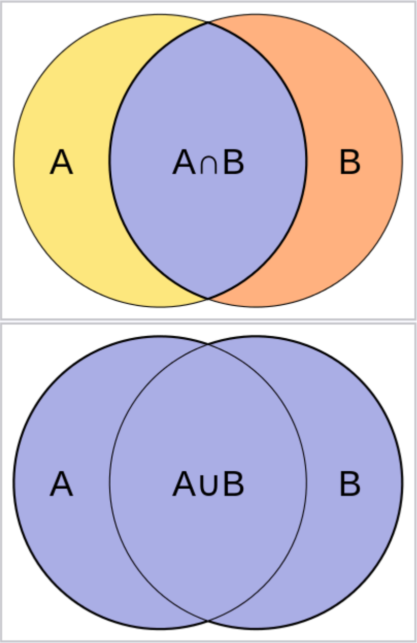
+ Q3. Jaccard Distance

- Jaccard coefficient

$$J(A,B) = \frac{|intersect\ (A,B)|}{|union\ (A,B)|} = \frac{|A \cap B|}{|A \cup B|}$$

$$= \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

$$0 \le J(A,B) \le 1$$



- Jaccard distance

$$d_J(A,B) = 1 - J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

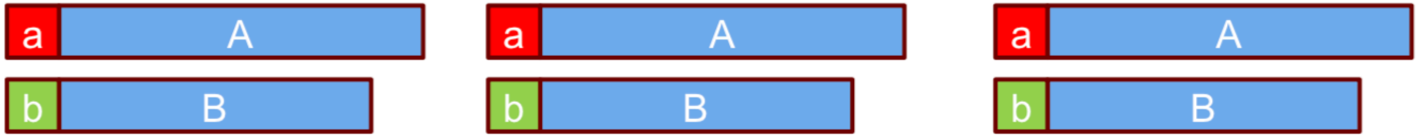
$$= \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

dissimilarity

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

+ Q3. Edit Distance Computation

- Intuition



d(A, B)+1

d(A,“b”+B)+1

d(“a”+A, B)+1

replace

$$A_0 = a_1 a_2 a_3 a_4 \cdots a_{m-1} a_m$$
$$B_0 = b_1 b_2 b_3 b_4 \cdots b_{n-1} b_n$$

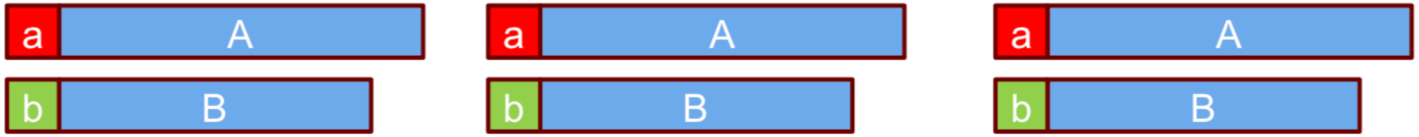
$$A_0 = \mathbf{a_1} a_2 a_3 a_4 \cdots a_{m-1} a_m$$
$$B_0 = \mathbf{b_1} b_2 b_3 b_4 \cdots b_{n-1} b_n$$

$$A_1 = \mathbf{a_1} a_2 a_3 a_4 \cdots a_{m-1} a_m$$
$$B_1 = \mathbf{a_1} b_2 b_3 b_4 \cdots b_{n-1} b_n$$

$$d(A_0, B_0) = d(A_1, B_1) + Cost_{\mathbf{replace}}$$

+ Q3. Edit Distance Computation

- Intuition



d(A, B)+1

d(A,“b”+B)+1

d(“a”+A, B)+1

delete

$$A_0 = a_1 a_2 a_3 a_4 \cdots a_{m-1} a_m$$
$$B_0 = b_1 b_2 b_3 b_4 \cdots b_{n-1} b_n$$

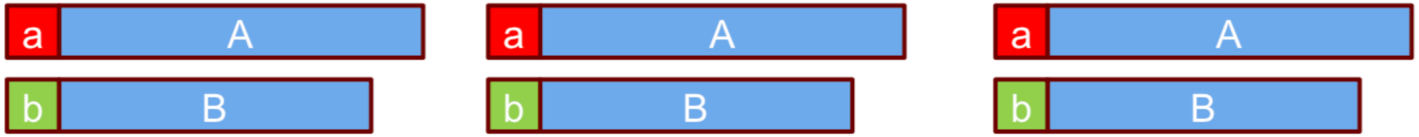
$$A_0 = \boldsymbol{a_1} a_2 a_3 a_4 \cdots a_{m-1} a_m$$
$$B_0 = b_1 b_2 b_3 b_4 \cdots b_{n-1} b_n$$

$$A_1 = \quad a_2 a_3 a_4 \cdots a_{m-1} a_m$$
$$B_0 = b_1 b_2 b_3 b_4 \cdots b_{n-1} b_n$$

$$\begin{aligned} & d(A_0, B_0) \\ &= d(A_1, B_0) + Cost_{\boldsymbol{delete}} \\ &= d(A_1, b_1 + B_1) + Cost_{\boldsymbol{delete}} \end{aligned}$$

+ Q3. Edit Distance Computation

- Intuition



d(A, B)+1

d(A,“b”+B)+1

d(“a”+A, B)+1

insert

$$A_0 = a_1 a_2 a_3 a_4 \cdots a_{m-1} a_m$$
$$B_0 = b_1 b_2 b_3 b_4 \cdots b_{n-1} b_n$$

$$A_1 = \boldsymbol{a_0}\, a_1 a_2 a_3 a_4 \cdots a_{m-1} a_m$$
$$B_0 = \quad b_1 b_2 b_3 b_4 \cdots b_{n-1} b_n$$

$$\begin{aligned} d(A_0, B_0) &= d(A_1, B_0) + Cost_{insert} \\ &= d(a_0 + A_0, B_0) + Cost_{insert} \end{aligned}$$

THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

+ Q3. Edit Distance Computation

$$d_{ij} = \begin{cases} d_{i-1,j-1} & \text{for } a_j = b_i \\ \min \begin{cases} d_{i-1,j} + w_{\text{del}}(b_i) \\ d_{i,j-1} + w_{\text{ins}}(a_j) \\ d_{i-1,j-1} + w_{\text{sub}}(a_j, b_i) \end{cases} & \text{for } a_j \neq b_i \end{cases}$$

University Queensland

Queensland University

sub	ins
del	

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

		U	N	I	V	E	R	S	I	T	Y	#	Q	U	E	E	N	S	L	A	N	D
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Q	1																					
U	2																					
E	3																					
E	4																					
N	5																					
S	6																					
L	7																					
A	8																					
N	9																					
D	10																					
#	11																					
U	12																					
N	13																					
I	14																					
V	15																					
E	16																					
R	17																					
S	18																					
I	19																					
T	20																					
Y	21																					

$$d_{ij} = \begin{cases} d_{i-1,\,j-1} & (a_j = b_i) \\ \min\begin{pmatrix} d_{i-1,\,j} + 1 \\ d_{i,\,j-1} + 1 \\ d_{i-1,\,j-1} + 1 \end{pmatrix} & (a_j \neq b_i) \end{cases}$$

		U	N	I	V	E	R	S	I	T	Y	#	Q	U	E	E	N	S	L	A	N	D
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Q	1	1	2	3	4	5	6	7	8	9	10	11	**11**	12	13	14	15	16	17	18	19	20
U	2	**1**																				
E	3	2																				
E	4	3																				
N	5	4																				
S	6	5																				
L	7	6																				
A	8	7																				
N	9	8																				
D	10	9																				
#	11	10																				
U	12	**11**																				
N	13	12																				
I	14	13																				
V	15	14																				
E	16	15																				
R	17	16																				
S	18	17																				
I	19	18																				
T	20	19																				
Y	21	20																				

$$d_{ij} = \begin{cases} d_{i-1,\,j-1} & (a_j = b_i) \\ \min\begin{pmatrix} d_{i-1,\,j} + 1 \\ d_{i,\,j-1} + 1 \\ d_{i-1,\,j-1} + 1 \end{pmatrix} & (a_j \neq b_i) \end{cases}$$

		U	N	I	V	E	R	S	I	T	Y	#	Q	U	E	E	N	S	L	A	N	D
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Q	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20
U	2	1	2	3	4	5	6	7	8	9	10	11	12	**11**	12	13	14	15	16	17	18	19
E	3	2	2																			
E	4	3	3																			
N	5	4	**3**																			
S	6	5	4																			
L	7	6	5																			
A	8	7	6																			
N	9	8	**7**																			
D	10	9	8																			
#	11	10	9																			
U	12	11	10																			
N	13	12	**11**																			
I	14	13	12																			
V	15	14	13																			
E	16	15	14																			
R	17	16	15																			
S	18	17	16																			
I	19	18	17																			
T	20	19	18																			
Y	21	20	19																			

$$d_{ij} = \begin{cases} d_{i-1,\,j-1} & (a_j = b_i) \\ \min\begin{pmatrix} d_{i-1,\,j} + 1 \\ d_{i,\,j-1} + 1 \\ d_{i-1,\,j-1} + 1 \end{pmatrix} & (a_j \neq b_i) \end{cases}$$

		U	N	I	V	E	R	S	I	T	Y	#	Q	U	E	E	N	S	L	A	N	D
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Q	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20
U	2	1	2	3	4	5	6	7	8	9	10	11	12	11	12	13	14	15	16	17	18	19
E	3	2	2	3	4	**4**	5	6	7	8	9	10	11	12	**11**	12	13	14	15	16	17	18
E	4	3	3	3																		
N	5	4	3	4																		
S	6	5	4	5																		
L	7	6	5	6																		
A	8	7	6	7																		
N	9	8	7	8																		
D	10	9	8	9																		
#	11	10	9	10																		
U	12	11	10	11																		
N	13	12	11	12																		
I	14	13	12	**11**																		
V	15	14	13	12																		
E	16	15	14	13																		
R	17	16	15	14																		
S	18	17	16	15																		
I	19	18	17	16																		
T	20	19	18	17																		
Y	21	20	19	18																		

$$d_{ij} = \begin{cases} d_{i-1,\,j-1} & (a_j = b_i) \\ \min\begin{pmatrix} d_{i-1,\,j} + 1 \\ d_{i,\,j-1} + 1 \\ d_{i-1,\,j-1} + 1 \end{pmatrix} & (a_j \neq b_i) \end{cases}$$

		U	N	I	V	E	R	S	I	T	Y	#	Q	U	E	E	N	S	L	A	N	D
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Q	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20
U	2	1	2	3	4	5	6	7	8	9	10	11	12	11	12	13	14	15	16	17	18	19
E	3	2	2	3	4	4	5	6	7	8	9	10	11	12	11	12	13	14	15	16	17	18
E	4	3	3	3	4	**4**	5	6	7	8	9	10	11	12	**12**	**11**	12	13	14	15	16	17
N	5	4	3	4	4																	
S	6	5	4	5	5																	
L	7	6	5	6	6																	
A	8	7	6	7	7																	
N	9	8	7	8	8																	
D	10	9	8	9	9																	
#	11	10	9	10	10																	
U	12	11	10	11	11																	
N	13	12	11	12	12																	
I	14	13	12	11	12																	
V	15	14	13	12	**11**																	
E	16	15	14	13	12																	
R	17	16	15	14	13																	
S	18	17	16	15	14																	
I	19	18	17	16	15																	
T	20	19	18	17	16																	
Y	21	20	19	18	17																	

$$d_{ij} = \begin{cases} d_{i-1,\,j-1} & (a_j = b_i) \\ \min\begin{pmatrix} d_{i-1,\,j} + 1 \\ d_{i,\,j-1} + 1 \\ d_{i-1,\,j-1} + 1 \end{pmatrix} & (a_j \neq b_i) \end{cases}$$

		U	N	I	V	E	R	S	I	T	Y	#	Q	U	E	E	N	S	L	A	N	D
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Q	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20
U	2	1	2	3	4	5	6	7	8	9	10	11	12	11	12	13	14	15	16	17	18	19
E	3	2	2	3	4	4	5	6	7	8	9	10	11	12	11	12	13	14	15	16	17	18
E	4	3	3	3	4	4	5	6	7	8	9	10	11	12	12	11	12	13	14	15	16	17
N	5	4	3	4	4	5	5	6	7	8	9	10	11	12	13	12	**11**	12	13	14	15	16
S	6	5	4	5	5	5																
L	7	6	5	6	6	6																
A	8	7	6	7	7	7																
N	9	8	7	8	8	8																
D	10	9	8	9	9	9																
#	11	10	9	10	10	10																
U	12	11	10	11	11	11																
N	13	12	11	12	12	12																
I	14	13	12	11	12	13																
V	15	14	13	12	11	12																
E	16	15	14	13	12	**11**																
R	17	16	15	14	13	12																
S	18	17	16	15	14	13																
I	19	18	17	16	15	14																
T	20	19	18	17	16	15																
Y	21	20	19	18	17	16																

$$d_{ij} = \begin{cases} d_{i-1,\,j-1} & (a_j = b_i) \\ \min\begin{pmatrix} d_{i-1,\,j} + 1 \\ d_{i,\,j-1} + 1 \\ d_{i-1,\,j-1} + 1 \end{pmatrix} & (a_j \neq b_i) \end{cases}$$

		U	N	I	V	E	R	S	I	T	Y	#	Q	U	E	E	N	S	L	A	N	D
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Q	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20
U	2	1	2	3	4	5	6	7	8	9	10	11	12	11	12	13	14	15	16	17	18	19
E	3	2	2	3	4	4	5	6	7	8	9	10	11	12	11	12	13	14	15	16	17	18
E	4	3	3	3	4	4	5	6	7	8	9	10	11	12	12	11	12	13	14	15	16	17
N	5	4	3	4	4	5	5	6	7	8	9	10	11	12	13	12	11	12	13	14	15	16
S	6	5	4	5	5	5	6	**5**	6	7	8	9	10	11	12	13	12	**11**	12	13	14	15
L	7	6	5	6	6	6	6															
A	8	7	6	7	7	7	7															
N	9	8	7	8	8	8	8															
D	10	9	8	9	9	9	9															
#	11	10	9	10	10	10	10															
U	12	11	10	11	11	11	11															
N	13	12	11	12	12	12	12															
I	14	13	12	11	12	13	13															
V	15	14	13	12	11	12	13															
E	16	15	14	13	12	11	12															
R	17	16	15	14	13	12	**11**															
S	18	17	16	15	14	13	12															
I	19	18	17	16	15	14	13															
T	20	19	18	17	16	15	14															
Y	21	20	19	18	17	16	15															

$$d_{ij} = \begin{cases} d_{i-1,\,j-1} & (a_j = b_i) \\ \min\begin{pmatrix} d_{i-1,\,j} + 1 \\ d_{i,\,j-1} + 1 \\ d_{i-1,\,j-1} + 1 \end{pmatrix} & (a_j \neq b_i) \end{cases}$$

		U	N	I	V	E	R	S	I	T	Y	#	Q	U	E	E	N	S	L	A	N	D
	0	1	2	3	4	5	6	7	8	9	10											
Q	1	1	2	3	4	5	6	7	8	9	10											
U	2	1	2	3	4	5	6	7	8	9	10											
E	3	2	2	3	4	4	5	6	7	8	9											
E	4	3	3	3	4	4	5	6	7	8	9											
N	5	4	3	4	4	5	5	6	7	8	9	10	11	12	13	12	11	12	13	14	15	16
S	6	5	4	5	5	5	6	5	6	7	8	9	10	11	12	13	12	11	12	13	14	15
L	7	6	5	6	6	6	6	6	6	7	8	9	10	11	12	13	13	12	**11**	12	13	14
A	8	7	6	7	7	7	7	7	7	7	8	9	10	11	12	13	14	13	12	**11**	12	13
N	9	8	7	8	8	8	8	8	8	8	8	9	10	11	12	13	**13**	14	13	12	**11**	12
D	10	9	8	9	9	9	9	9	9	9	9	9	10	11	12	13	14	14	14	13	12	**11**
#	11	10	9	10	10	10	10	10	10	10	10	**9**	10	11	12	13	14	15	15	14	13	12
U	12	11	10	11	11	11	11	11	11	11	11	10										
N	13	12	11	12	12	12	12	12	12	12	12	11										
I	14	13	12	11	12	13	13	13	**12**	13	13	12										
V	15	14	13	12	11	12	13	14	13	13	14	13										
E	16	15	14	13	12	11	12	13	14	14	14	14										
R	17	16	15	14	13	12	11	12	13	14	15	15										
S	18	17	16	15	14	13	12	**11**	12	13	14	15										
I	19	18	17	16	15	14	13	12	**11**	12	13	14										
T	20	19	18	17	16	15	14	13	12	**11**	12	13										
Y	21	20	19	18	17	16	15	14	13	12	**11**	12										

$$d_{ij} = \begin{cases} d_{i-1,\,j-1} & (a_j = b_i) \\ \min\begin{pmatrix} d_{i-1,\,j} + 1 \\ d_{i,\,j-1} + 1 \\ d_{i-1,\,j-1} + 1 \end{pmatrix} & (a_j \neq b_i) \end{cases}$$

		U	N	I	V	E	R	S	I	T	Y	#	Q	U	E	E	N	S	L	A	N	D
	0	1	2	3	4	5	6	7	8	9	10											
Q	1	1	2	3	4	5	6	7	8	9	10											
U	2	1	2	3	4	5	6	7	8	9	10											
E	3	2	2	3	4	4	5	6	7	8	9											
E	4	3	3	3	4	4	5	6	7	8	9											
N	5	4	3	4	4	5	5	6	7	8	9	10	11	12	13	12	11	12	13	14	15	16
S	6	5	4	5	5	5	6	5	6	7	8	9	10	11	12	13	12	11	12	13	14	15
L	7	6	5	6	6	6	6	6	6	7	8	9	10	11	12	13	13	12	11	12	13	14
A	8	7	6	7	7	7	7	7	7	7	8	9	10	11	12	13	14	13	12	11	12	13
N	9	8	7	8	8	8	8	8	8	8	8	9	10	11	12	13	13	14	13	12	11	12
D	10	9	8	9	9	9	9	9	9	9	9	9	10	11	12	13	14	14	14	13	12	11
#	11	10	9	10	10	10	10	10	10	10	10	9	10	11	12	13	14	15	15	14	13	12
U	12	11	10	11	11	11	11	11	11	11	11	10	10	**10**	11	12	13	14	15	15	14	13
N	13	12	11	12	12	12	12	12	12	12	12	11	11	11	11	12	**12**	13	14	15	15	14
I	14	13	12	11	12	13	13	13	12	13	13	12	12	12	12	12	13	13	14	15	16	15
V	15	14	13	12	11	12	13	14	13	13	14	13	13	13	13	13	13	14	14	15	16	16
E	16	15	14	13	12	11	12	13	14	14	14	14	14	14	**13**	**13**	14	14	15	15	16	17
R	17	16	15	14	13	12	11	12	13	14	15	15	15	15	14	14	14	15	15	16	16	17
S	18	17	16	15	14	13	12	11	12	13	14	15	16	16	15	15	15	**14**	15	16	17	17
I	19	18	17	16	15	14	13	12	11	12	13	14	15	16	16	16	16	15	15	16	17	18
T	20	19	18	17	16	15	14	13	12	11	12	13	14	15	16	17	17	16	16	17	17	18
Y	21	20	19	18	17	16	15	14	13	12	11	12	13	14	15	16	17	17	17	17	18	**18**

$$d_{ij} = \begin{cases} d_{i-1,\,j-1} & (a_j = b_i) \\ \min\begin{pmatrix} d_{i-1,\,j} + 1 \\ d_{i,\,j-1} + 1 \\ d_{i-1,\,j-1} + 1 \end{pmatrix} & (a_j \neq b_i) \end{cases}$$

Q3. Edit Distance Computation

UNIVERSITY#QUEENSLAND

Length = 21

UNIVERSITY#QUEENSLAND

UNIVERSITY#~~Q~~UNERSITY~~D~~

UNIVERSITY# UNIVERSITY

QUNIVERSITYD#UNIVERSITY

QUEIVENSLAND#UNIVERSITY

QUE~~IV~~ENSLAND#UNIVERSITY

QUEENSLAND#UNIVERSITY

THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

+ Q3. Edit Distance Computation

- distance = 18

  at least 18 operations should be performed to transform string A to B

- similarity

$$sim(a,b) = 1 - \frac{ED(a,b)}{\max(|a|,|b|)}$$

$$= 1 - \frac{18}{21} = 0.143$$

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Q3. Jaccard Coefficient Comp.

- 3-grams

A = University Queensland

**Uni, niv, ive, ver, ers, rsi, sit, ity,
ty#, y#Q, #Qu, Que, uee, een, ens, nsl,
sla, lan, and**

B = Queensland University

**Que, uee, een, ens, nsl, sla, lan, and,
nd#, d#U, #Un, Uni, niv, ive, ver, ers,
rsi, sit, ity**

of n-gram: $21 - n + 1 = 19$

+ Q3. Jaccard Coefficient Comp.

A

Uni, niv, ive, ver, ers, rsi, sit, ity,
ty#, y#Q, #Qu, Que, uee, een, ens, nsl,
sla, lan, and

B

Que, uee, een, ens, nsl, sla, lan, and,
nd#, d#U, #Un, Uni, niv, ive, ver, ers,
rsi, sit, ity

19 - 3

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{16}{16+3+3} = 0.727$$

$$d_J(A,B) = 1 - J(A,B) = 0.273$$

+ Q3–(d) Conclusion

- similarity
  - Edit Distance: 0.143
  - Jaccard Coefficient: 0.727
- dissimilarity
  - Edit Distance: 1 – 0.143 = 0.857
  - Jaccard Distance: 1 – 0.727 = 0.273
- Jaccard distance is less sensitive to word orders

University Queensland

Queensland University