INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

ANALYSIS OF TOP 2 OF 3 CHOICES: GREATEST ENGINEERING ACHIEVEMENTS OF THE CENTURY DATASET USING INTELLIGENT DATA MINING TECHNIQUE

Ali Tariq Bhatti

North Carolina A&T State University, Greensboro NC USA
atbhatti@aggies.ncat.edu, ali_tariq302@hotmail.com

Abstract: -There are 20 Greatest Engineering Achievements of this 20th century. The class is to decide on what they believe the top three achievements are. Each group in the class is assigned to one of three tiers whose jobs are to determine the top singlet, pair, and triplet combinations reducing the number as it passes from tier to tier. Resolution group are to get the outcomes from top (8) singlets top (4) pairs, and top (3) triplets. As arbiter group, the goal is to determine the final (singlet, pair and triplet) achievement codes from the data passed from Resolution group. Arbiters corrected the results of Resolution group and then, they apply their methods using data mining tools to get the top achievements to see, which one's are the winners in terms of singlet, pairs, and triplets. In this research paper, leading two of three choices in our case such as pair and triplet or you can choose other combination, so how arbiters use their data mining tools for their methodology to get the top three achievements as a winner.

Keywords: - Data mining, Data sets, GAC, Weka, Weight proposed method, Bernoulli process, Breaker method

1. DATA MINING

Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. These data mining tools predict future trends and behaviors by reading through databases for hidden patterns; they allow organizations to make proactive, knowledge-driven decisions. Here, a notable success was achieved by SKICAT, a system used by astronomers to perform image analysis, classification, and cataloging of sky objects from sky-survey images [1].

2. INTRODUCTION

There were five groups assigned to work on the given GAC (Greatest Achievement of Century) database. First group worked on the top eight (8) GAC entries, second group, I worked on the top four (4) GAC pair-combinations, and third group worked on the top three (3) GAC triplets. Other two groups were Resolution and Arbiter, however, the role of the Resolution will design and administer a process by which the class selects top GAC singlets (2), pairs(2), and triplets(2). The last tier group called as Arbiter will take decisions from the

Resolution group and design and administer a process by which a singlet, pair, and triplet is chosen for the class.

3. THE TOP EIGHT GAC ENTRIES

This was the first group, who worked on the Top Eight GAC Entries. They used the data mining tool such as Microsoft Excel to get the outcome of top categories being related to electronics, software and computers. I observed from their group that there was an issue with the code assignment. Code B7T was assigned to two categories: Airplane and Radio/TV. There were 5 counts for code B7T, but no way of telling for which category. They computed the code as a singlet for the frequency of 20 greatest engineering achievements of this century as

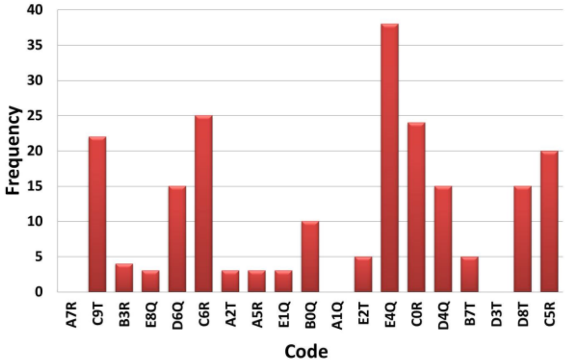
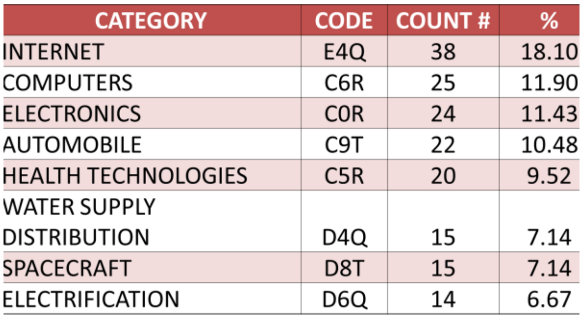


Figure 1: Frequency for singlet

The TOP 8 GAC computed using data mining tool (Microsoft Excel) were as.



CATEGORY	CODE	COUNT #	%
INTERNET	E4Q	38	18.10
COMPUTERS	C6R	25	11.90
ELECTRONICS	C0R	24	11.43
AUTOMOBILE	C9T	22	10.48
HEALTH TECHNOLOGIES	C5R	20	9.52
WATER SUPPLY DISTRIBUTION	D4Q	15	7.14
SPACECRAFT	D8T	15	7.14
ELECTRIFICATION	D6Q	14	6.67

Figure 2: TOP 8 GAC Singlets

I also observed in the group, how they show their Top 8 GAC with Pie chart as below.

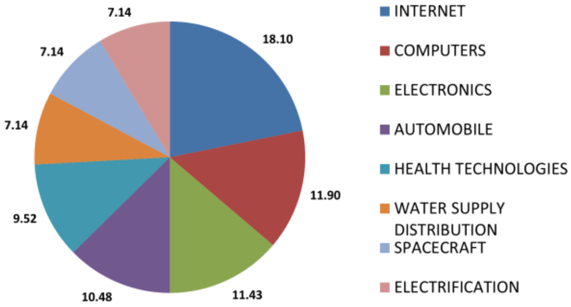


Figure 3: Pie chart

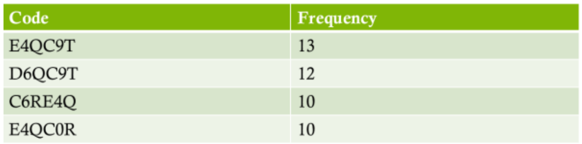
4. TOP 4 GAC PAIR COMBINATIONS

A. Weka Software:

First, we use the data mining software called as Weka to analyze our Top 4 GAC Pair Combinations. Weka is a collection of machine learning algorithms for solving real world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from your own Java code. The software enables a computer program to automatically analyze a large body of data and decide what information is most relevant. This crystallized information can then be used to automatically make predictions or to help people make decisions faster and more accurately [2].

B. Microsoft Excel Software

We use Microsoft Excel software to get the correct Top (4) Pair combinations



Code	Frequency
E4QC9T	13
D6QC9T	12
C6RE4Q	10
E4QC0R	10

Figure 4: Top 4 pair combinations

Code Definition for the top 4 GAC pair combinations were as:

1) E4QC9T - Internet, Automobile
2) D6QC9T- Electrification, Automobile
3) C6RE4Q - Computers, Internet
4) E4QC0R - Internet, Electronics

Bar Graph of Top 4 GAC pair Combinations were as:

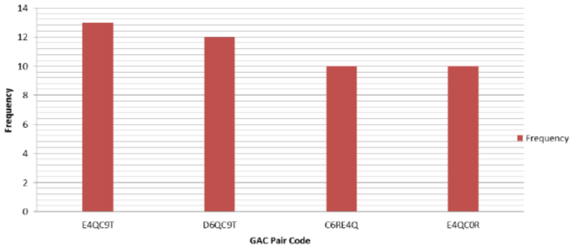


Figure 5: Bar graph

As a whole, able to produce the output results for the top 4 GAC pair combinations successfully. We are waiting for the input on the resolution and arbitrary.

5. TOP 3 TRIPLETS

This group decodes the data from the provided GAC database and provides the class with the top three (3) GAC triplets and they also specify that order doesn't matter in the triplet.

A. Methodology

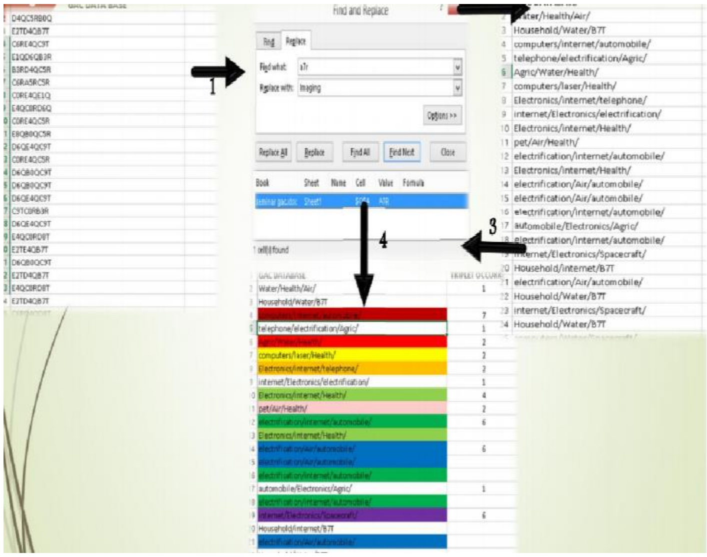
We analyzed that they use the Microsoft Excel software for the Top 3 Triplets

1) Import Data into Microsoft Excel
2) Arrange data
3) Use to find and replace function to decode data
4) Use of keyword search to highlight and count triplets
5) Collate result

6) Perform tie breaker
7) Finalize Result

B. PICTORIAL VIEW OF THE PROCESS

By using the data mining tool such as Microsoft Excel, they follow the steps which are discussed in the methodology part as:



GAC DATA BASE

D4QC5RB0Q
E2TD4QB7T
C6RE4QC9T
E1QD6QB3R
B3RD4QC5R
C6RA5RC5R
C0RE4QE1Q
E4QC0RD6Q
C0RE4QC5R
E8QB0QC5R
D6QE4QC9T
C0RE4QC5R
D6QB0QC9T
D6QB0QC9T
D6QE4QC9T
C9TC0RB3R
D6QE4QC9T
E4QC0RD8T
E2TE4QB7T
D6QB0QC9T
E2TD4QB7T
E4QC0RD8T
E2TD4QB7T

Find and Replace

Find | Replace

Find what: a7r
Replace with: Imaging

Options >>

Replace All | Replace | Find All | Find Next | Close

Book	Sheet	Name	Cell	Value	Formula
seminar gac.xlsx	Sheet1		D4	A7R	

1 cell(s) found

2 Water/Health/Air/
3 Household/Water/B7T
4 computers/internet/automobile/
5 telephone/electrification/Agric/
6 Agric/Water/Health/
7 computers/laser/Health/
8 Electronics/internet/telephone/
9 internet/Electronics/electrification/
10 Electronics/internet/Health/
11 pet/Air/Health/
12 electrification/internet/automobile/
13 Electronics/internet/Health/
14 electrification/Air/automobile/
15 electrification/Air/automobile/
16 electrification/internet/automobile/
17 automobile/Electronics/Agric/
18 electrification/internet/automobile/
19 internet/Electronics/Spacecraft/
20 Household/internet/B7T
21 electrification/Air/automobile/
22 Household/Water/B7T
23 internet/Electronics/Spacecraft/
24 Household/Water/B7T

	GAC DATABASE	TRIPLET OCCURR
2	Water/Health/Air/	1
3	Household/Water/B7T	
4	computers/internet/automobile/	7
5	telephone/electrification/Agric/	1
6	Agric/Water/Health/	2
7	computers/laser/Health/	2
8	Electronics/internet/telephone/	2
9	internet/Electronics/electrification/	1
10	Electronics/internet/Health/	4
11	pet/Air/Health/	2
12	electrification/internet/automobile/	6
13	Electronics/internet/Health/	
14	electrification/Air/automobile/	6
15	electrification/Air/automobile/	
16	electrification/internet/automobile/	
17	automobile/Electronics/Agric/	1
18	electrification/internet/automobile/	
19	internet/Electronics/Spacecraft/	6
20	Household/internet/B7T	
21	electrification/Air/automobile/	

Figure 6: Triplet

The bar chart representation for triplets is as:

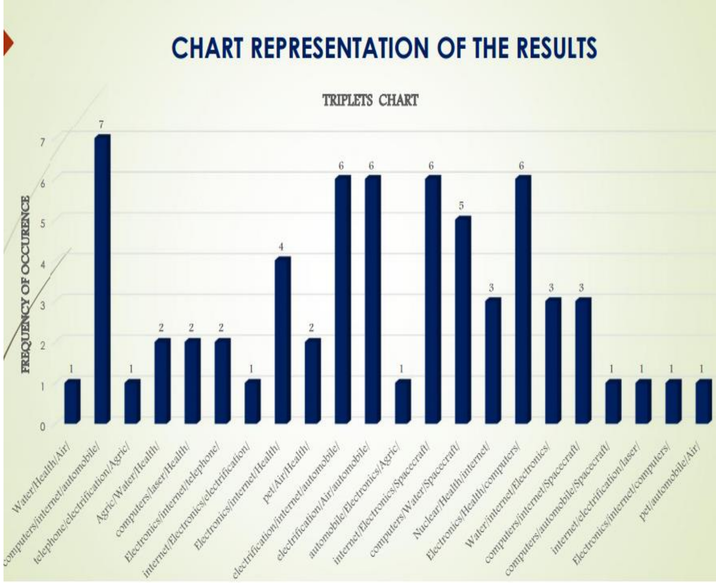
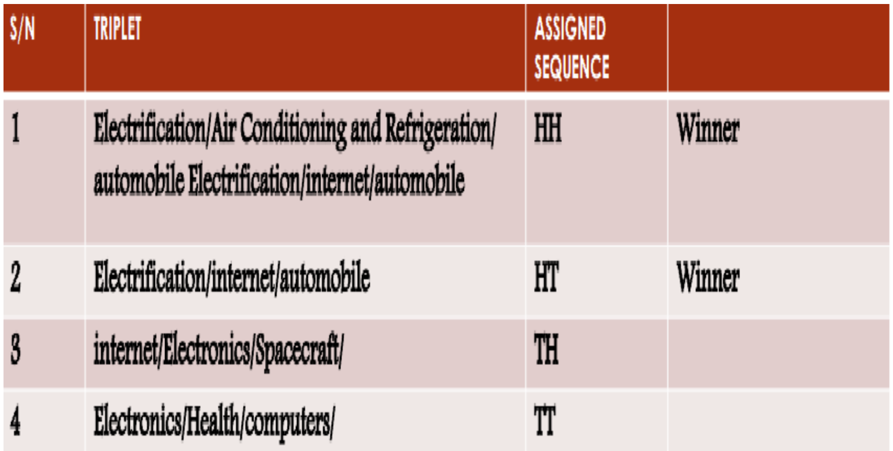


Figure 7: Bar chart for Triplet

C. Tie Breaker

1) Once they used their Microsoft Excel software, so the tie breaker was settled using two coins.

2) Firstly, we assigned sequences to each of the triplets, then we flipped the coins a certain amount of times, the first two sequences to appear twice, were declared winners.



S/N	TRIPLET	ASSIGNED SEQUENCE	
1	Electrification/Air Conditioning and Refrigeration/ automobile Electrification/internet/automobile	HH	Winner
2	Electrification/internet/automobile	HT	Winner
3	internet/Electronics/Spacecraft/	TH	
4	Electronics/Health/computers/	TT	

Figure 8: Winner of Tie Breaker using Microsoft Excel

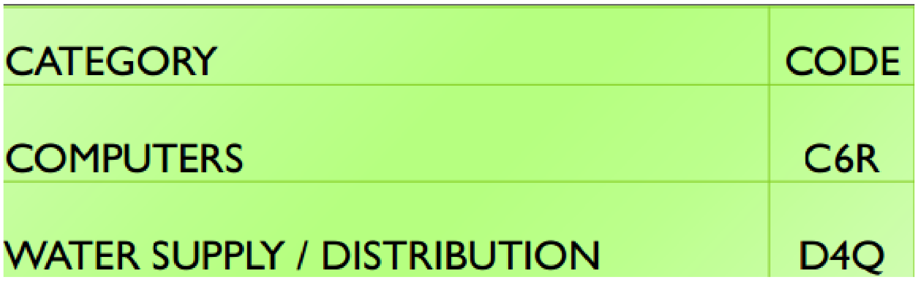
The results of Top 3 Triplets along with their frequency using data mining software are as:
1) Computers/internet/automobile -7
2) Electrification/Air Conditioning and Refrigeration/ automobile- 6
3) Electrification/internet/automobile- 6

6. RESOLUTION GROUP:

This group has had to take the input from singlets, pair, and triplets. Their job was to (a) Selecting Top 2 out of the top eight (8) GAC entries (b) Selecting Top 2 out of the top four (4) GAC pair-combinations (c) Selecting Top 2 out of the top three (3) GAC triplets.

A. Top Two Singlets Analysis Procedure

Resolution took the data from singlets as in figure (2), and analyzed that they did these steps such as:
1) Listed top 8 singlets.
2) Columnize the singlets into columns 1, 2 and 3.
3) Pick most frequent in each column.
4) List top two and identified with the shortest distance to the provided database codes.
I analyzed that they finally got Top 2 singlets such as

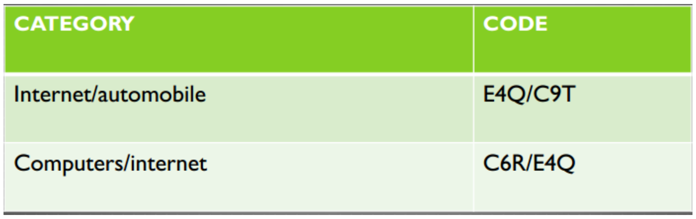


CATEGORY	CODE
COMPUTERS	C6R
WATER SUPPLY / DISTRIBUTION	D4Q

Figure 9: Top 2 singlets by Resolution

B. Top Two Pairs Analysis Procedure

Resolution took the data from pairs as in figure (5), and analyzed that they did these steps such as,
1) List top four combinations
2) Identified most frequent codes
3) Identified pairs that contained the most frequent codes
I analyzed that they finally got Top 2 pairs such as:



CATEGORY	CODE
Internet/automobile	E4Q/C9T
Computers/internet	C6R/E4Q

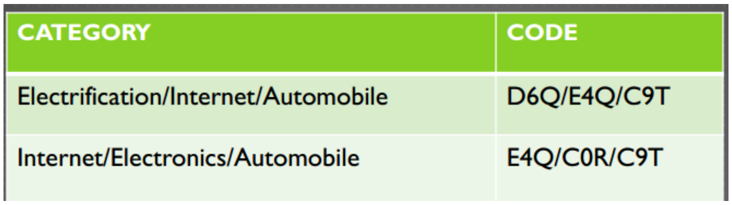
Figure 10: Top 2 pairs by Resolution

C. Top Two Triplets Analysis Procedure

Resolution took the data from triplets as mentioned in Triplets section, and analyzed that they did these steps such as:

1) List top three triplets
2) Identified most frequent codes
3) Ranked top two with the contained winners

I analyzed that they finally got Top 2 triplets such as:



CATEGORY	CODE
Electrification/Internet/Automobile	D6Q/E4Q/C9T
Internet/Electronics/Automobile	E4Q/C0R/C9T

Figure 11: Top 2 triplets by Resolution

7. ARBITERS GROUP

Arbiters group are the one to determine the final singlet, pair and triplet achievement codes from the data passed from Resolution group. I understand that Arbiters use three methods to approach the final singlet, pair a pair and triplets. Those methods they used are as: (a)Proposed Method: Weighting based on the singlet frequency, (b) Bernoulli process Tie Breaker method1, (c) Random decision Tie Breaker method2

A. Methods of Arbiter

1) Proposed Method: Weighting based on the singlet frequency: This method is a single array code based on each code type such as singlet, pair, and triplet determined to be the one with the highest weight. Each code is weighted based on the cumulative frequency of the base singlet combination. However, ties are broken by a coin flip. In fact, ties and winners are recorded in the event alternate methods are desired to break ties or determine overall winner.

2) Bernoulli process Tie Breaker method1: The data sets used to demonstrate the performance of the aspect Bernoulli model are quite distinct in their nature [3]. It is a coin flip method consisting of 2 achievement codes whether they are singlet, pairs, and triplets. However, number of iteration is passed through the function that signifying the number of tosses or flip. Therefore, whom ever achieve the number of occurrences be the winner and if there is a tie, the coin is flipped once more to determine the winner.

3) Random decision Tie Breaker method2: This method is different than Bernoulli process Tie Breaker method1 consisting of achievement codes whether they are singlet, pairs, and triplets. A prime number plays an important role for the number of iterations to be passed through the function. In this method, we are not looking for number of occurrences as the quantity is chosen for continuous randomization before deciding a winner. Moreover, winner receives "1" in their place and summation frequency shows output as a result.

B. Input data:

Once Arbiters get the input data from Resolution group such as:

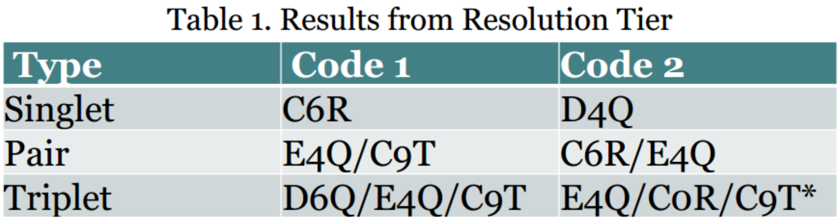
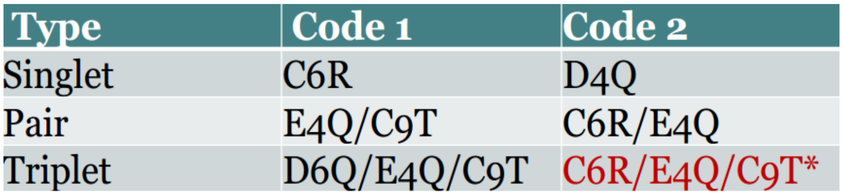


Table 1. Results from Resolution Tier

Type	Code 1	Code 2
Singlet	C6R	D4Q
Pair	E4Q/C9T	C6R/E4Q
Triplet	D6Q/E4Q/C9T	E4Q/C0R/C9T*

Figure 12: Results from Resolution tier

After identifying error in achievement codes, we see for triplet change such as C0R from C6R.

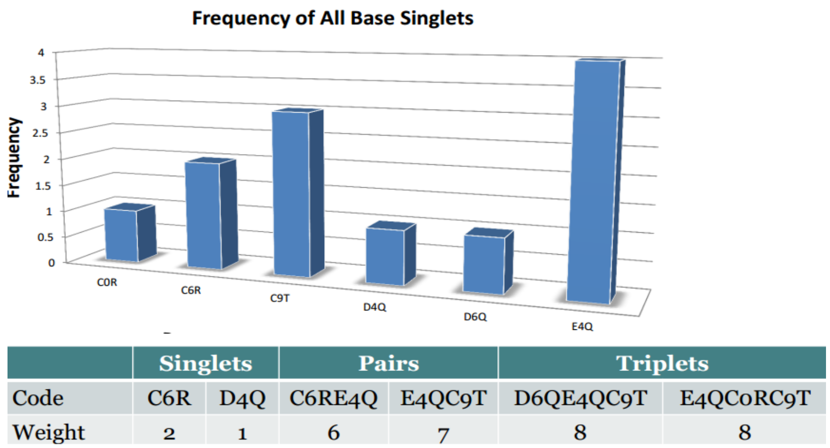


Type	Code 1	Code 2
Singlet	C6R	D4Q
Pair	E4Q/C9T	C6R/E4Q
Triplet	D6Q/E4Q/C9T	C6R/E4Q/C9T*

Figure 13: Identifying error in triplet

C. Original Results of Resolution

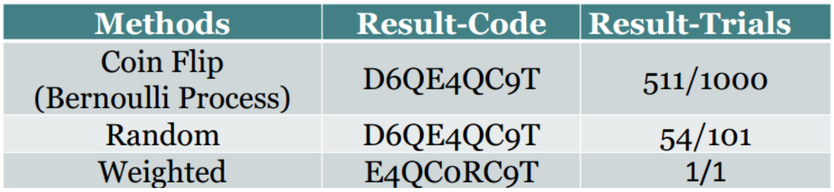
However, results originally from Resolution, the Arbiters use their methods to compute the frequency of all base singlets as:



	Singlets		Pairs		Triplets	
Code	C6R	D4Q	C6RE4Q	E4QC9T	D6QE4QC9T	E4QC0RC9T
Weight	2	1	6	7	8	8

Figure 14: Original result using Weight proposed method

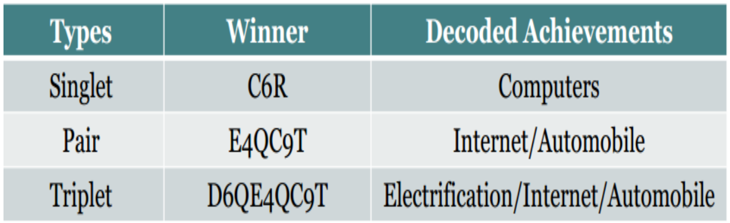
I analyzed that they use three data mining tools for triplets such as their methods to tie break for original result of Resolution.



Methods	Result-Code	Result-Trials
Coin Flip (Bernoulli Process)	D6QE4QC9T	511/1000
Random	D6QE4QC9T	54/101
Weighted	E4QC0RC9T	1/1

Figure 15: Tie breaker for Original result of Resolution

Therefore, the original results of resolution by arbiter are as:

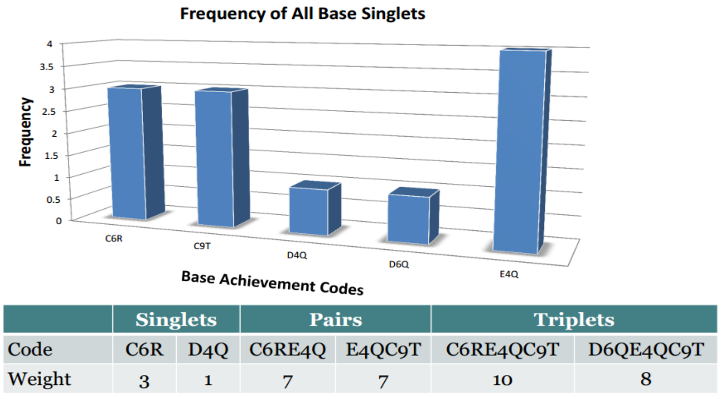


Types	Winner	Decoded Achievements
Singlet	C6R	Computers
Pair	E4QC9T	Internet/Automobile
Triplet	D6QE4QC9T	Electrification/Internet/Automobile

Figure 16: Original results of resolution by arbiter

D. Corrected Results of Resolution

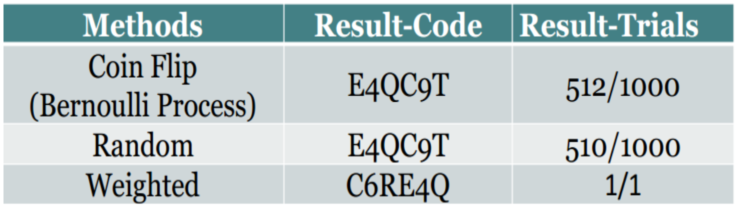
However, the Arbiters use their methods to compute the frequency of all base singlets as:



	Singlets		Pairs		Triplets	
Code	C6R	D4Q	C6RE4Q	E4QC9T	C6RE4QC9T	D6QE4QC9T
Weight	3	1	7	7	10	8

Figure 17: Corrected result using Weight proposed method

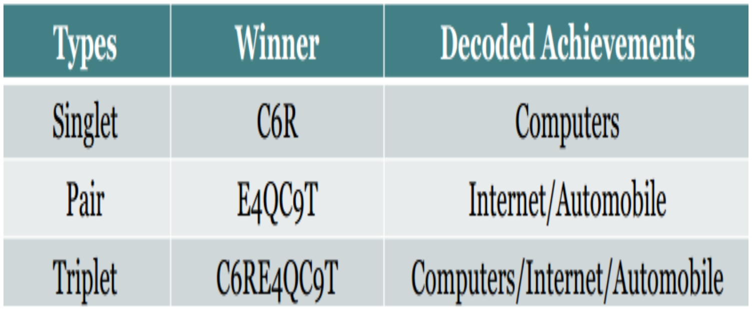
I analyzed that they use three data mining tools for pairs such as their methods to tie break for correct result of Resolution.



Methods	Result-Code	Result-Trials
Coin Flip (Bernoulli Process)	E4QC9T	512/1000
Random	E4QC9T	510/1000
Weighted	C6RE4Q	1/1

Figure 18: Tie breaker for correct result of Resolution

Therefore, the corrected results by arbiter are as:



Types	Winner	Decoded Achievements
Singlet	C6R	Computers
Pair	E4QC9T	Internet/Automobile
Triplet	C6RE4QC9T	Computers/Internet/Automobile

Figure 19: Corrected results by Arbiter

8. Conclusion

We are leading two of three choices such as pair and triplet, so how arbiters use their three methodologies by using data-mining tools to get the top three achievements. In this research paper, analysis to be determined that the two greatest achievements from the Arbiter are:

1) Pair: Internet/Automobile

2) Triplet: Computers/Internet/Automobile

These two greatest achievements play a vital role in this 20th century. I feel that Internet and Computers are useful in our professional life, and Automobile is our daily usages and requirement in aspect of our life.

The data mining methods for triplets used to get the greatest achievement are Microsoft Excel software, so the tie breaker was settled using two coins. Firstly, assigned sequences to each of the triplets, then we flipped the coins a certain amount of times, the first two sequences to appear twice, were declared winners. Other data mining methods such as Bernoulli and Random decision Tie Breaker method to achieve the number of occurrences be the winner and if there is a tie, the coin is flipped once more to determine the winner. The data mining methods for pair, we also used Microsoft Excel Software to check the number of occurrence the one with highest weight or most frequent code. Other data mining methods such as Weight proposed method, Bernoulli and Random decision Tie Breaker method also help us to get greatest achievement for pair as Internet/Automobile and triplet as Computers/Internet/Automobile as analyzed in this research paper.

REFERENCES

[1] S. Fayad, U.M; Djorgovski and N. Weir, "From digitized images to online catalogs: Data mining a sky survey." Al Magazine, vol. 17, no. 2, pp. 51–56, 1996.

[2] R. Peter and W. Ian, "The weka data mining software: An update," SIGKDD Explorations, vol. 11, 2009.

[3] E. Kaban, A; Bingham and T. Hirsimaki, "Learning to read between the lines: The aspect bernoulli model." Proceedings of the 4th SIAM international conference on data mining. pp. 462–466, 2004.

BIOGRAPHY

Ali Tariq Bhatti received his Associate degree in Information System Security (Highest Honors) from Rockingham Community College, NC USA, B.Sc. in Software engineering (Honors) from UET Taxila, Pakistan, M.Sc. in Electrical engineering (Honors) from North Carolina A&T State University, NC USA, and currently pursuing PhD in Electrical engineering from North Carolina A&T State University. Working as a researcher in campus and working off-campus too. His area of interests and current research includes Coding Algorithm, Networking Security, Mobile Telecommunication, Biosensors, Genetic Algorithm, Swarm Algorithm, Health, Bioinformatics, Systems Biology, Control system, Power, Software development, Software Quality Assurance, Communication, and Signal Processing. For more information, contact **Ali Tariq Bhatti** at ali_tariq302@hotmail.com.