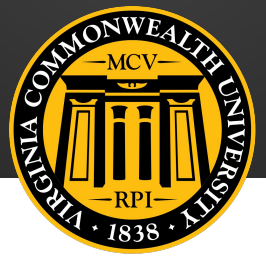
Understanding Mortality and Aging

Utilizing data-mining techniques to find conserved patterns in the different global causes of mortality



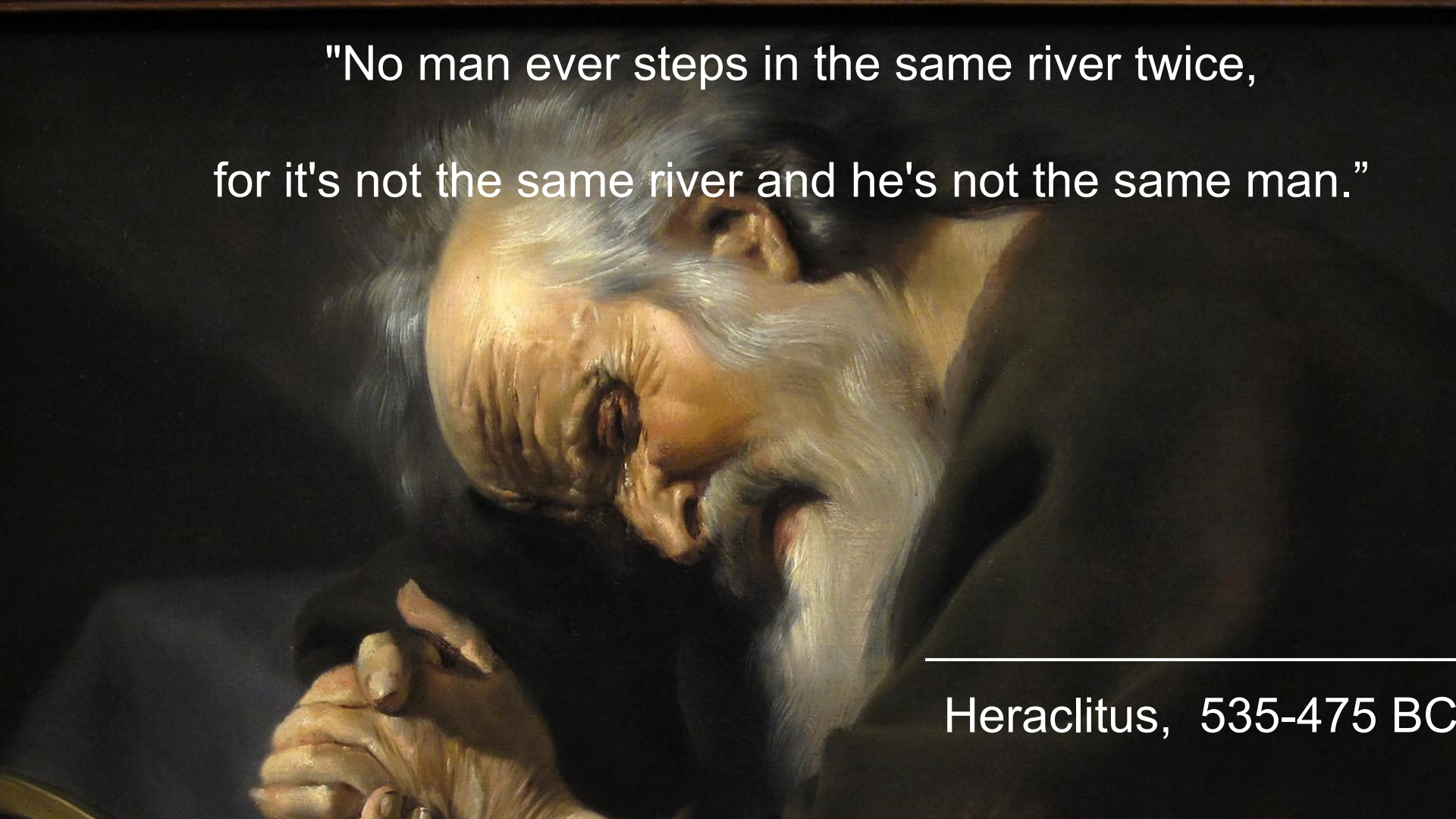
Center for the Study of
BIOLOGICAL
COMPLEXITY

Skyler A. Kuhn

Dr. Tarynn Witten, Dr. Alberto Cano, Dr. Allison Johnson

Center for the Study of Biological Complexity & Department of Computer Science

Virginia Commonwealth University


"No man ever steps in the same river twice,
for it's not the same river and he's not the same man."
Heraclitus, 535-475 BC

What is Biodemography?

- Biodemography is an amalgamation of the biological sciences and demography
- Biodemographers seeks to understand how the biological determinates of death and birth vary across populations[1]

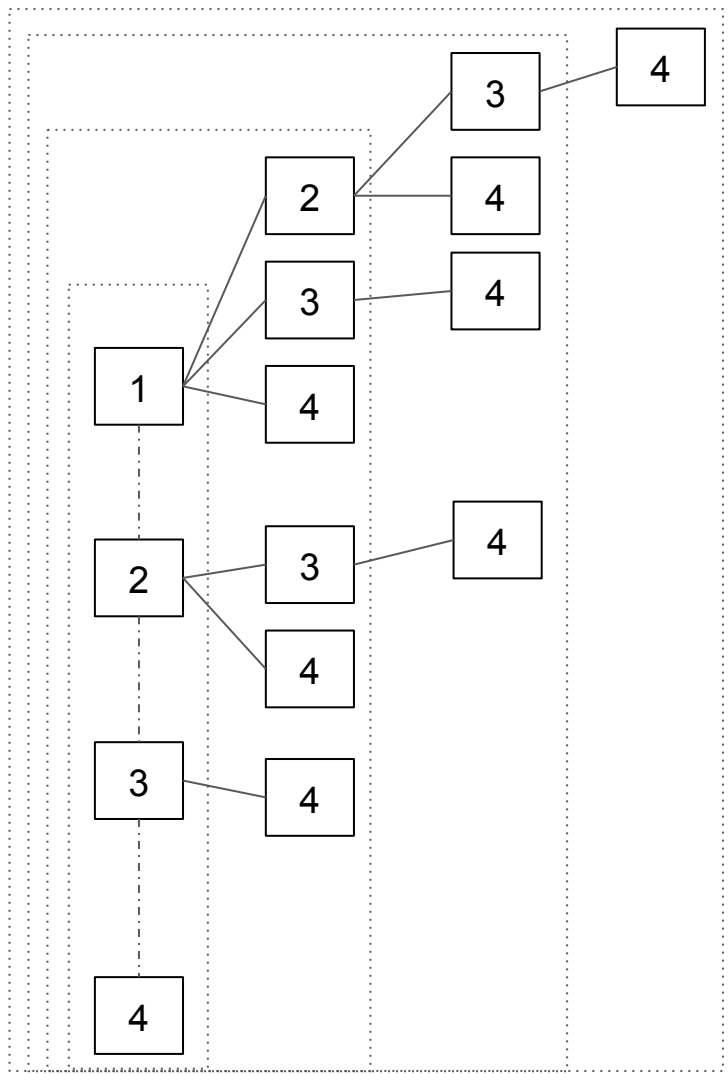
1. Crimmins, Eileen, Jung Ki Kim, and Sarinnapha Vasunilashorn. "Biodemography: New Approaches to Understanding Trends and Differences in Population Health and Mortality." *Demography* 47.S (2010). Print.

Exhaustive Searches

- A class of iterative methods typically used to solve discrete problems where no efficient solution method is known[2]
  - Also known as Brute Force Searches
- This method enumerates through all possible combinations of a given set to test each candidate (or each generated combination) against a predefined problem statement
  - Easy to implement, easy to program

---

2. "Brute-force Search." *Wikipedia*. Wikimedia Foundation, 04 Aug. 2017. Web. 07 Aug. 2017.



Exhaustive Search of set S {1,2,3,4} yielding k-itemsets

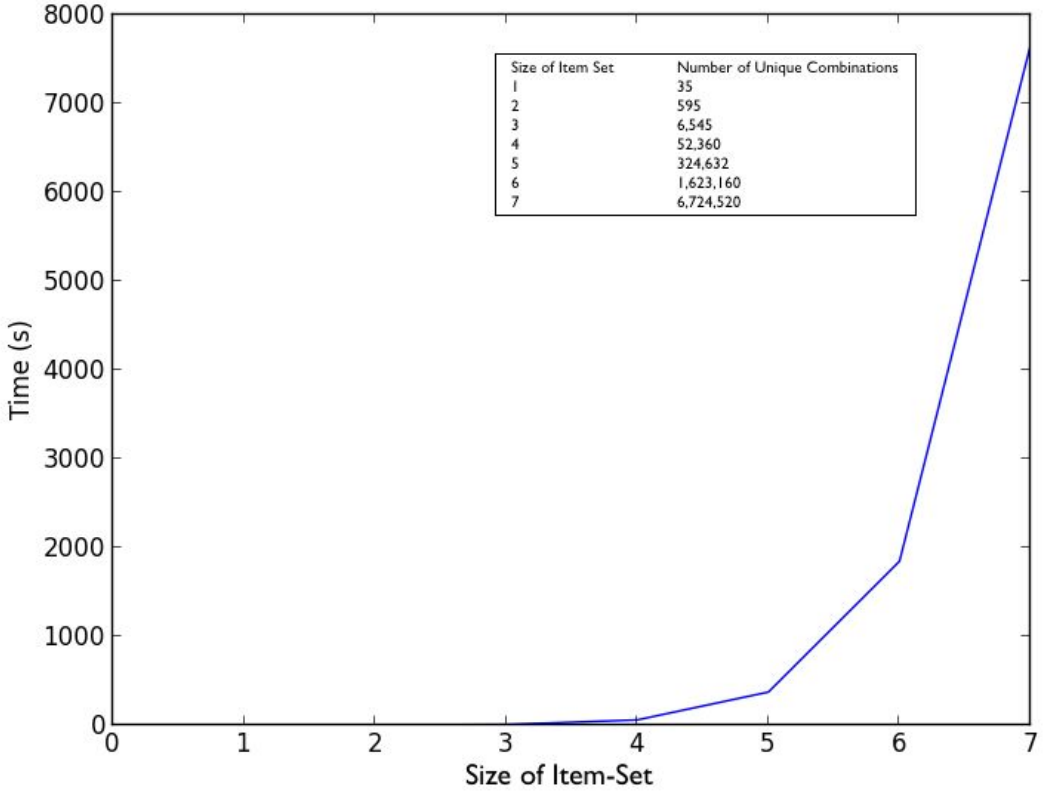
1-itemsets	2-itemsets	3-itemsets	4-itemsets
{1}	{1,2}	{1,2,3}	{1,2,3,4}
{2}	{1,3}	{1,2,4}	
{3}	{1,4}	{1,3,4}	
{4}	{2,3}	{2,3,4}	
	{2,4}		
	{3,4}		

Disadvantages of Exhaustive Searches

- Suffers from severe performance issues
- As the size of the set to be permuted grows, the time complexity increases in a polynomial manner
  - O(n^k): where n is the cardinal number of the set and k is the cardinal number of the candidate itemset[3]
  - Last resort option

3. "Computational Complexity Theory." *Wikipedia*. Wikimedia Foundation, 06 Aug. 2017. Web. 07 Aug. 2017.

Combinatorial explosion of k-itemset generation in an exhaustive search of set S {1,2,3,4,5,6,7,..., 35}



Observed

Itemset Cardinality	Generative Timing (secs)
0	0.0
1	0.04
2	0.68
3	7.39
4	59.46
5	373.39
6	1845.85
7	7660.54

Extrapolated

Itemset Cardinality	Generative Timing (secs)
8	26934.39
10	210088.26
12	954946.64
14	2654961.54
16	4646182.69
18	5192792.42
20	3716946.15
35	0.0011444

Decades you say, no problem!

… 44 years later

```
>>> primt("Done!")

Traceback (most recent call last):

File "<stdin>", line 789, in <module>

NameError: name 'primt' is not defined


>>> print("Hello Darkness, my old friend!")
```

Metaheuristics

- Alternative to exhaustive search methods that can be employed to find a sufficient solution in a set which is too large to be completely sampled[4]
  - Uses a heuristic to cut off subsets of the search, effectively reducing search space and runtime[5]
  - Does not always yield a global optimal solution (a solution to all possible combinations)[4, 6]

4. Blum, Christian, and Andrea Roli. "Metaheuristics in Combinatorial Optimization." *ACM Computing Surveys* 35.3 (2003): 268-308. Print..
5. Yang, Xin-She. "Metaheuristic Optimization." *Scholarpedia* 6.8 (2011): 11472. Print.
6. "Computational Complexity Theory." *Wikipedia*. Wikimedia Foundation, 06 Aug. 2017. Web. 07 Aug. 2017.

Computational Complexity Theory

- Field that identifies the inherent difficulty that some problems yield and classifies them accordingly[6]

- Biggest unsolved problem is whether P = NP
  - If true, it would mean that all very hard problems have seemingly easy solutions
    - Implications of this proof would be drastic and widespread

---

6. "Computational Complexity Theory." *Wikipedia*. Wikimedia Foundation, 06 Aug. 2017. Web. 07 Aug. 2017.

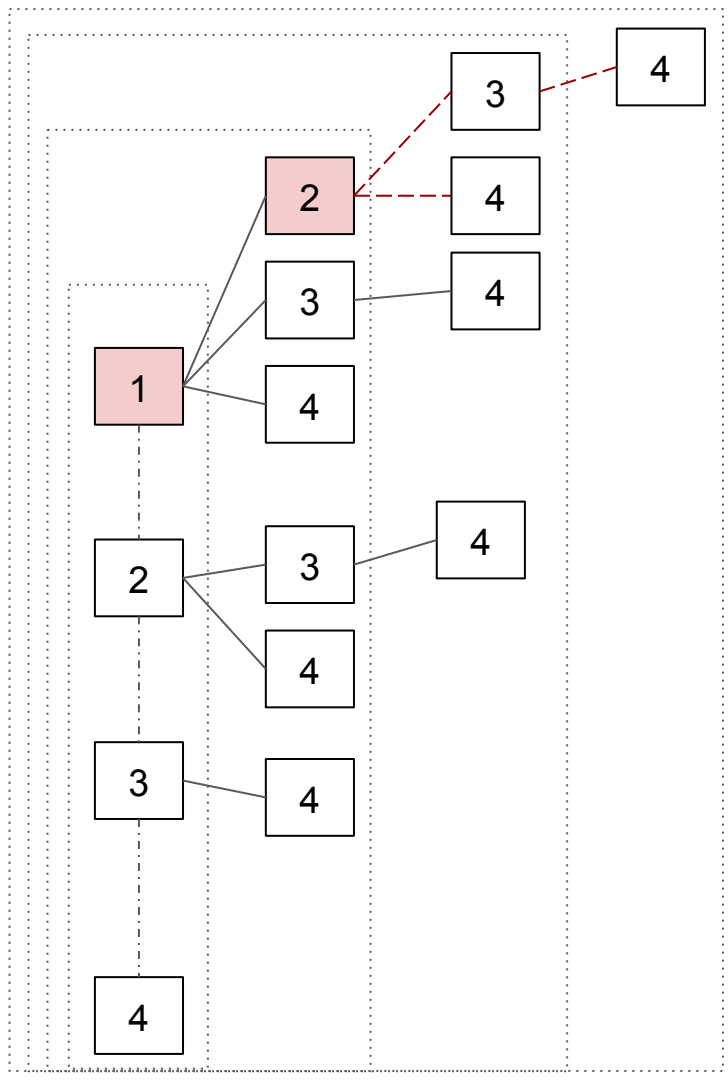
.

Apriori Algorithm

- Robust metaheuristic used in data mining for the evaluation of frequent itemsets for boolean association rules
  - Heuristic evaluates itemsets which have met a minimum support (a count of how frequently the itemset appears within the dataset or database).
  - Based off of the apriori property which states that any subset of a frequent itemset must also be frequent[1]

---

7. Adamo, Jean-Marc. "Apriori and Other Algorithms." *Data Mining for Association Rules and Sequential Patterns* (2001): 33-48. Print.



Exhaustive Search of set S {1,2,3,4} yielding k-itemsets

1-itemsets	2-itemsets	3-itemsets	4-itemsets
{1}	{1,2}	{1,2,3}	{1,2,3,4}
{2}	{1,3}	{1,2,4}	
{3}	{1,4}	{1,3,4}	
{4}	{2,3}	{2,3,4}	
	{2,4}		
	{3,4}		

Apriori Algorithm of set S {1,2,3,4} yielding k-itemsets

1-itemsets	2-itemsets	3-itemsets	4-itemsets
{1} ✓	{1,2} ✘	~~{1,2,3}~~	~~{1,2,3,4}~~
{2} ✓	{1,3} ✓	~~{1,2,4}~~	
{3} ✓	{1,4} ✓	{1,3,4} ✓	
{4} ✓	{2,3} ✓	{2,3,4} ✓	
	{2,4} ✓		
	{3,4} ✓		

Adapted Apriori Algorithm

- Heuristic: p-value threshold
  - For each country, an itemset frequency is calculated by summating the frequencies for each item in a generated k-itemset; these k-itemset frequencies are then evaluated with a one-way chi-squared test to find whether there's a significant difference between each countries observed frequency.
    - If a candidate k-itemset does not meet this threshold then its search space is effectively pruned.

H_o: C_1k-itemset = C_2k-itemset = … = C_6k-itemset

H_A: C_ik-itemset ≠ C_jk-item

Pipeline

- Step I: Data preprocessing: “preproces.py”
  - Cleans up the datasets, before any data analysis begins, the function “pre_process_data” will analyze all columns containing icd code mortality rates, and it will replace null fields with “0”
    - If not done, problems will arise when the chi-squared analysis begins
  - The program’s second major role is to parse each of the datasets (which contain data on the top icd mortality rates for a range of ages for a given country) by biological sex into two new files-- instantiated if there is not a 1:2 ratio between datafile types

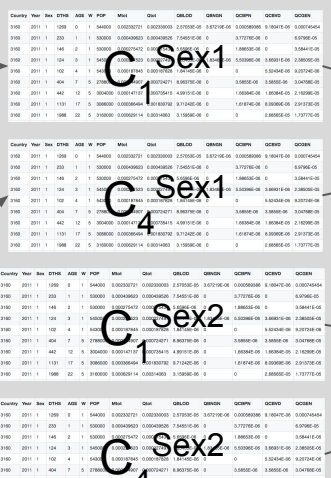
Pipeline

- Step II: Data management: “pipeline.files2dictionary(filename, country_ID, supp_dict)”
  - For the management and housing of all the project’s data, two dictionaries were created to house all the data for each biological sex (~six datasets)
  - Each age was evaluated to see if it met the minimum support (~6 or the number of countries)
    - If an age did not meet the minimum support, then one or more of the data set(s) did not contain any information on that age group, and that age was not evaluated. *The goal is to find diseases or groups of diseases conserved across all six countries.*

Pipeline

- Step III: Data analysis: “pipeline.apriori_v3(q, insig, sex_file_dict, countries_list, age)”
  - To implement this algorithm, a queue and a list of insignificant itemsets were used generate tentative candidates
    - If an itemset was found to be significant, it was added to the queue and appended to a list of significant itemsets
    - If found to be insignificant, the tentative candidate was popped from the queue
- This effectively and efficiently reduces the algorithm’s search space from years to mins.

Step I: Data Preprocessing



Sex 1

[country][age][icd] = freq

Sex 2

[country][age][icd] = freq

Step II: Data Management

Set S = {1,2,3,4}

Countries: C_1 C_2 C_3 C_4

{1} = chi_square([32, 40, 30, 20]) ✓

{2} = chi_square([8, 15, 5, 6]) ✓

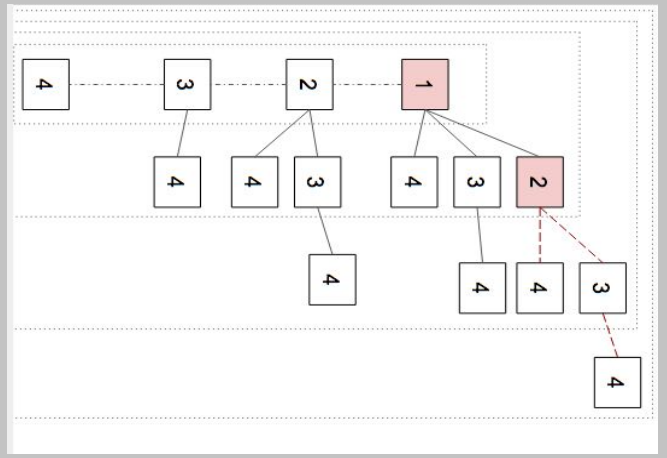
$\{1,2\}_K$ [40, 55, 35, 26]+

C_1 C_2 C_3 C_4

{1,2} chi_square([40,55,35,26]) ✘

p-value < alpha

Does not get added to the Queue!



1-itemsets

{1} ✓

{2} ✓

{3} ✓

{4} ✓

2-itemsets

{1,2} ✘ = [40,55,35,26]

{1,3} ✓

{1,4} ✓

{2,3} ✓

{2,4} ✓

{3,4} ✓

3-itemsets

~~{1,2,3}~~

~~{1,2,4}~~

{1,3,4} ✓

{2,3,4} ✓

4-itemsets

~~{1,2,3,4}~~

Step III: Data Analysis

Special Acknowledgements

It is of genuine pleasure to express my deepest sympathies to you all today. None of this would have been possible without you; thank you for your timely feedback, contagious kindness, and boundless knowledge that has enabled myself to accomplish this tumultuous task. And for that, I say thank you!

Thank you for supporting the power of mentorship. I believe it has the power to transform the world into a better place. The passing of knowledge is one of humanity's most sacred acts of veneration. I am humbled and honored to have worked with each one of you.

Sincerely,

Skyler Kuhn

References

1. Crimmins, Eileen, Jung Ki Kim, and Sarinnapha Vasunilashorn. "Biodemography: New Approaches to Understanding Trends and Differences in Population Health and Mortality." *Demography* 47.S (2010). Print.
2. "Brute-force Search." *Wikipedia*. Wikimedia Foundation, 04 Aug. 2017. Web. 07 Aug. 2017.
3. "Computational Complexity Theory." *Wikipedia*. Wikimedia Foundation, 06 Aug. 2017. Web. 07 Aug. 2017.
4. Blum, Christian, and Andrea Roli. "Metaheuristics in Combinatorial Optimization." *ACM Computing Surveys* 35.3 (2003): 268-308. Print..
5. Yang, Xin-She. "Metaheuristic Optimization." *Scholarpedia* 6.8 (2011): 11472. Print.
6. "Computational Complexity Theory." *Wikipedia*. Wikimedia Foundation, 06 Aug. 2017. Web. 07 Aug. 2017.
7. Adamo, Jean-Marc. "Apriori and Other Algorithms." *Data Mining for Association Rules and Sequential Patterns* (2001): 33-48. Print.

#primt(“Done!”)

print(“Done!”)