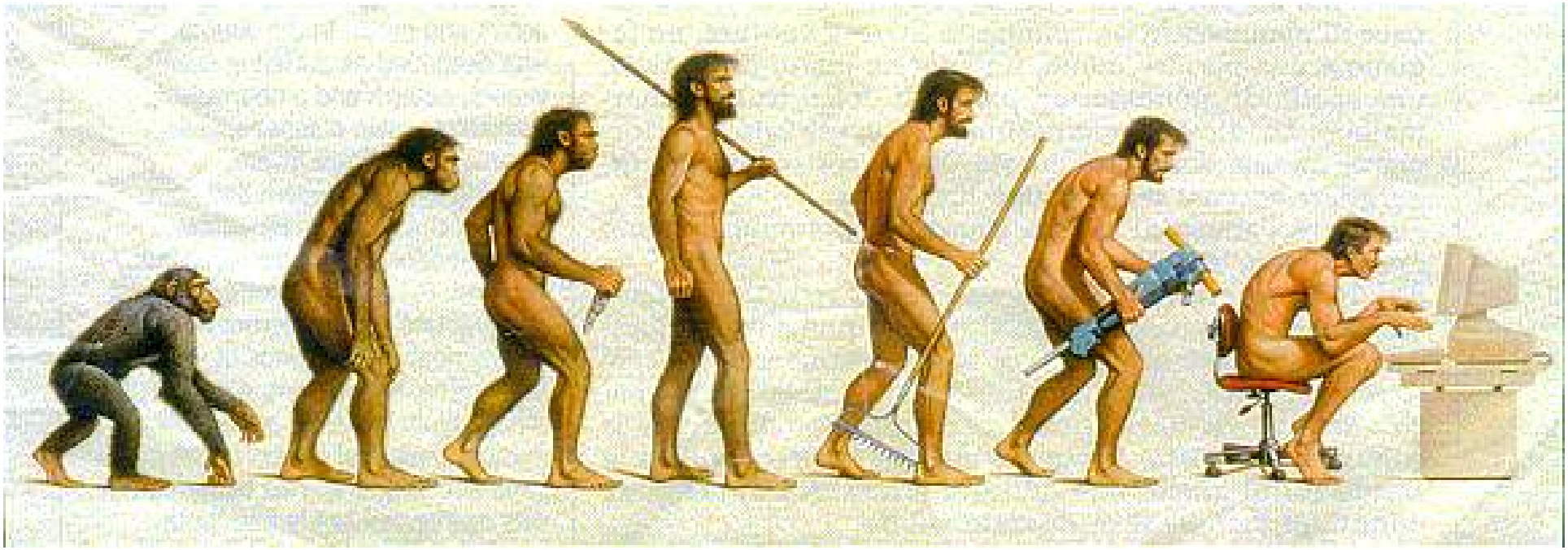
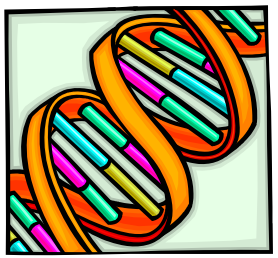
Evolutionary Computation

Theory & Application

Presented by

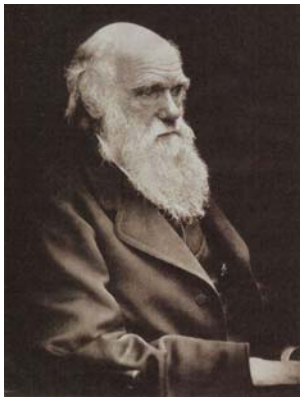
Robert M. Patton, Ph.D.

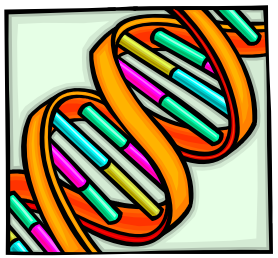




What is EC?

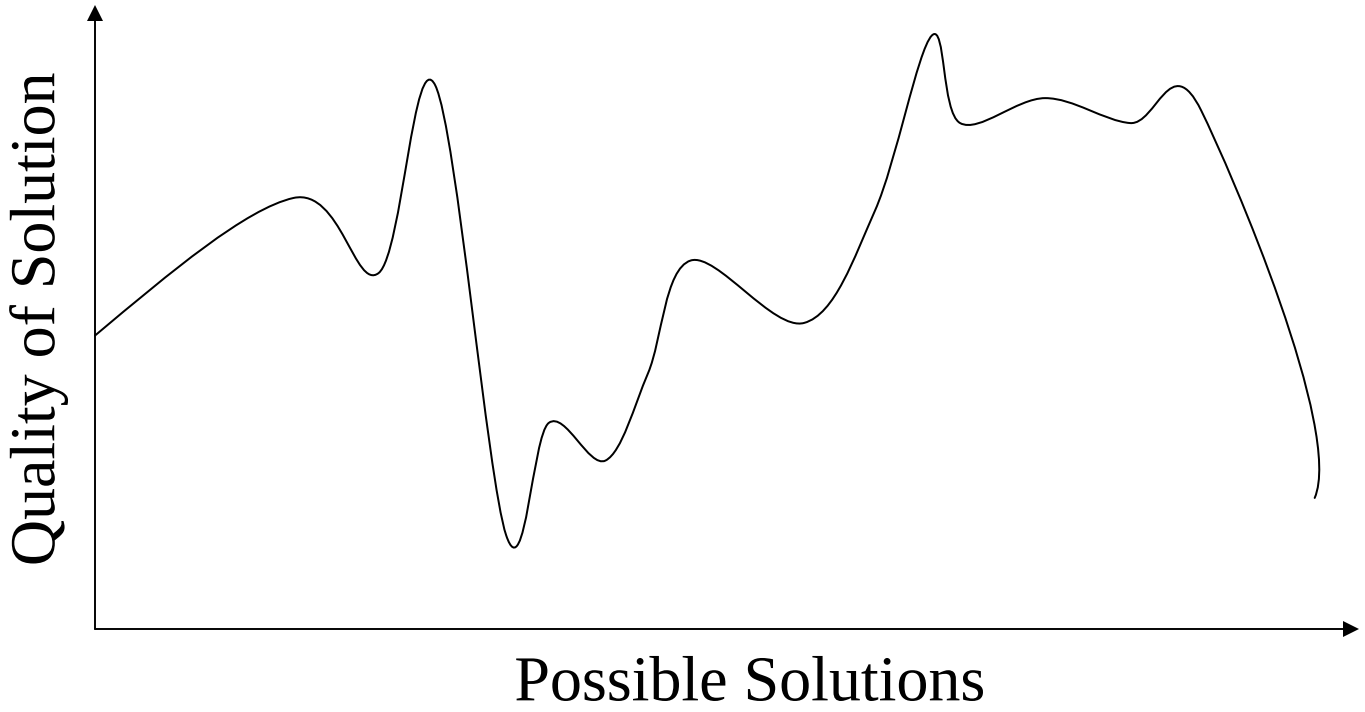
- Search algorithms based on principles of natural selection and genetic reproduction (J. Holland)
- “Only the strong survive”
- Breeding the Best of the Best

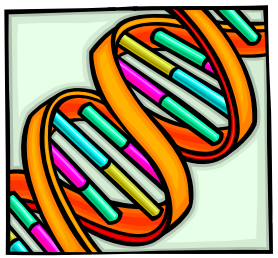




What is EC good for?

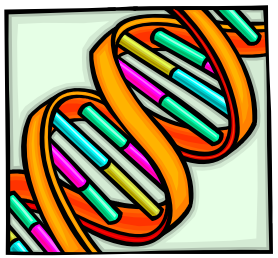
- Traversing extremely large areas that may have multiple local maximums / minimums





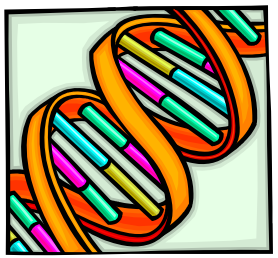
Types of Problems

- Searching
- Combinatorial & Permutation problems
- Optimization
- Learning / Evolving (artificial intelligence)
- Adaptive Sampling
- Clustering



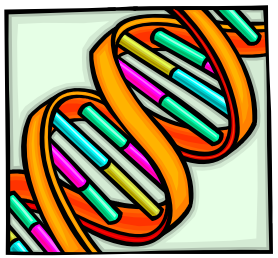
Types of EC

- Types:
  - Genetic Algorithms
  - Genetic Programming
  - Evolutionary Strategy
  - Learning Classifier Systems
  - Quantum-Inspired Evoluationary Algorithms
- Differences are generally in representation and fitness function



EC Components

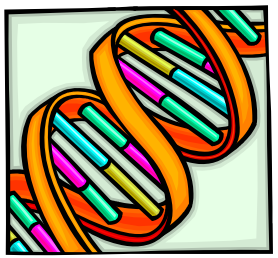
- Population of Individuals
  - Population represents solutions to a problem
- Fitness function (how to evaluate)
  - Determines how well the solutions solve the problem
- Selection function (how to select parents)
  - Only the “best” solutions survive to create new offspring
- Genetic operators (how to create children)
  - Crossover
  - Mutation



Algorithm

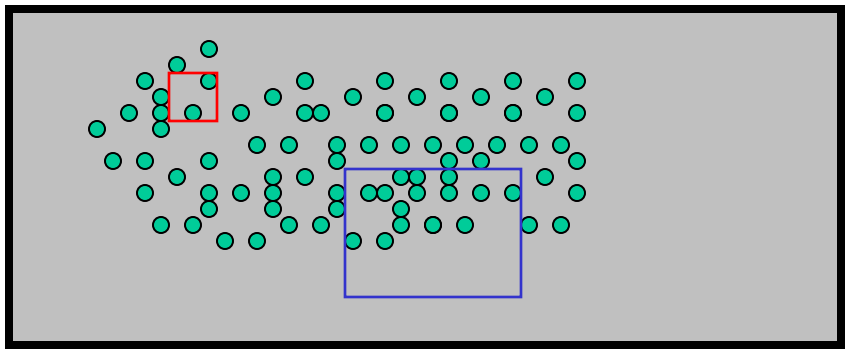
Initialize population;
Evaluate current population;
While (stopping condition* not satisfied) {
    Select parents from population;
    Apply Crossover/Mutation to parents;
    Evaluate current population;
}

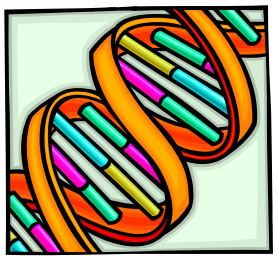
*Stopping condition could be: number of generations, fitness threshold, etc.



Population

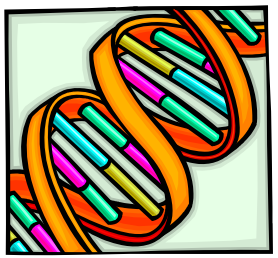
- A collection of possible solutions to the problem (i.e., a sampling of the solution space)
- Population size can be fixed or variable
- Determine a “good” size of population can be tricky
  - Small population
    - Converge fast
    - Little variety
  - Large population
    - Converge slow
    - Much variety



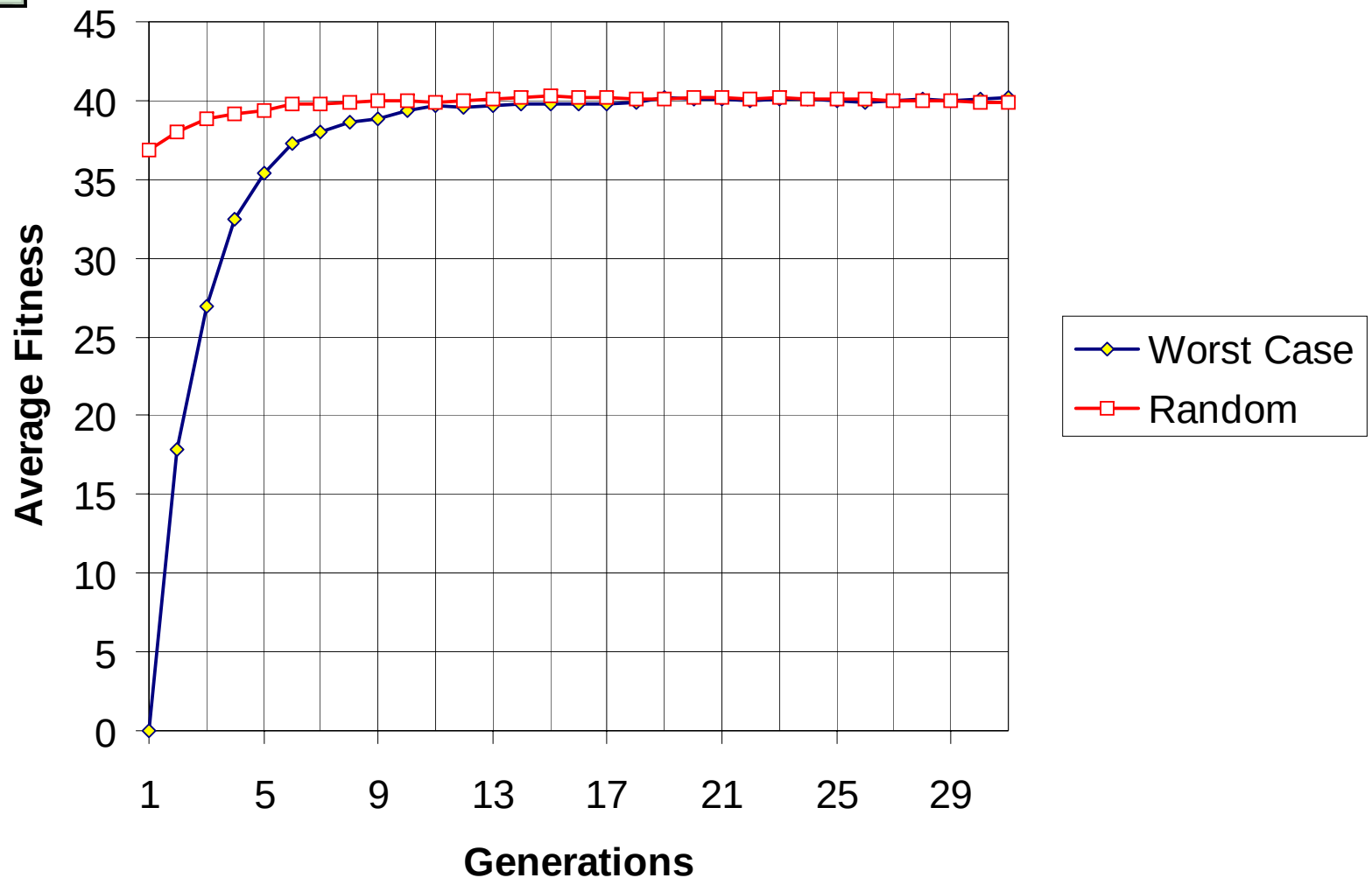


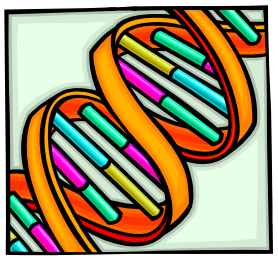
Initialize Population

- Two approaches:
  - Random (looking for a diamond in the mountain)
    - "I don't have a clue what the solution is"
  - Sub-optimal (looking for a diamond in a rock)
    - "I have a 'good enough' solution but I want to improve it"
- Starting with a mountain takes longer than if you started with a rock



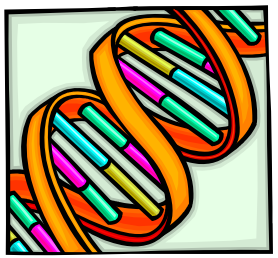
Initialization Comparison





Encoding

- Solution domain must be encoded into “DNA”
- Binary: only 1’s and 0’s
  - 0101000011100101
- Real: any number
  - 912837472910193 8
- Real numbers usually work better for real-world problems

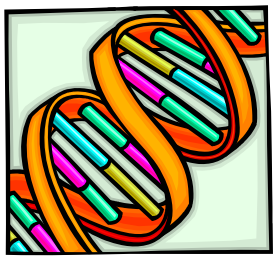


Encoding Example

DNA

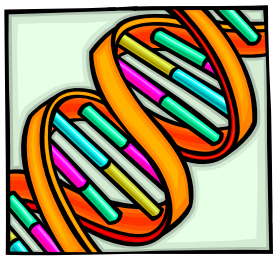
Appetite	Gene 1	Herbivore (1), Carnivore (2), Insectivore (3)
Movement	Gene 2	Climb (1), Fly (2), Run (3), Swim (3)
Feet	Gene 3	Claws (1), Flippers (2), Hooves (3), Wings (4)
Skin	Gene 4	Fur (1), Scales (2), Feathers (3)

Example: [3, 2, 4, 3]→Bird

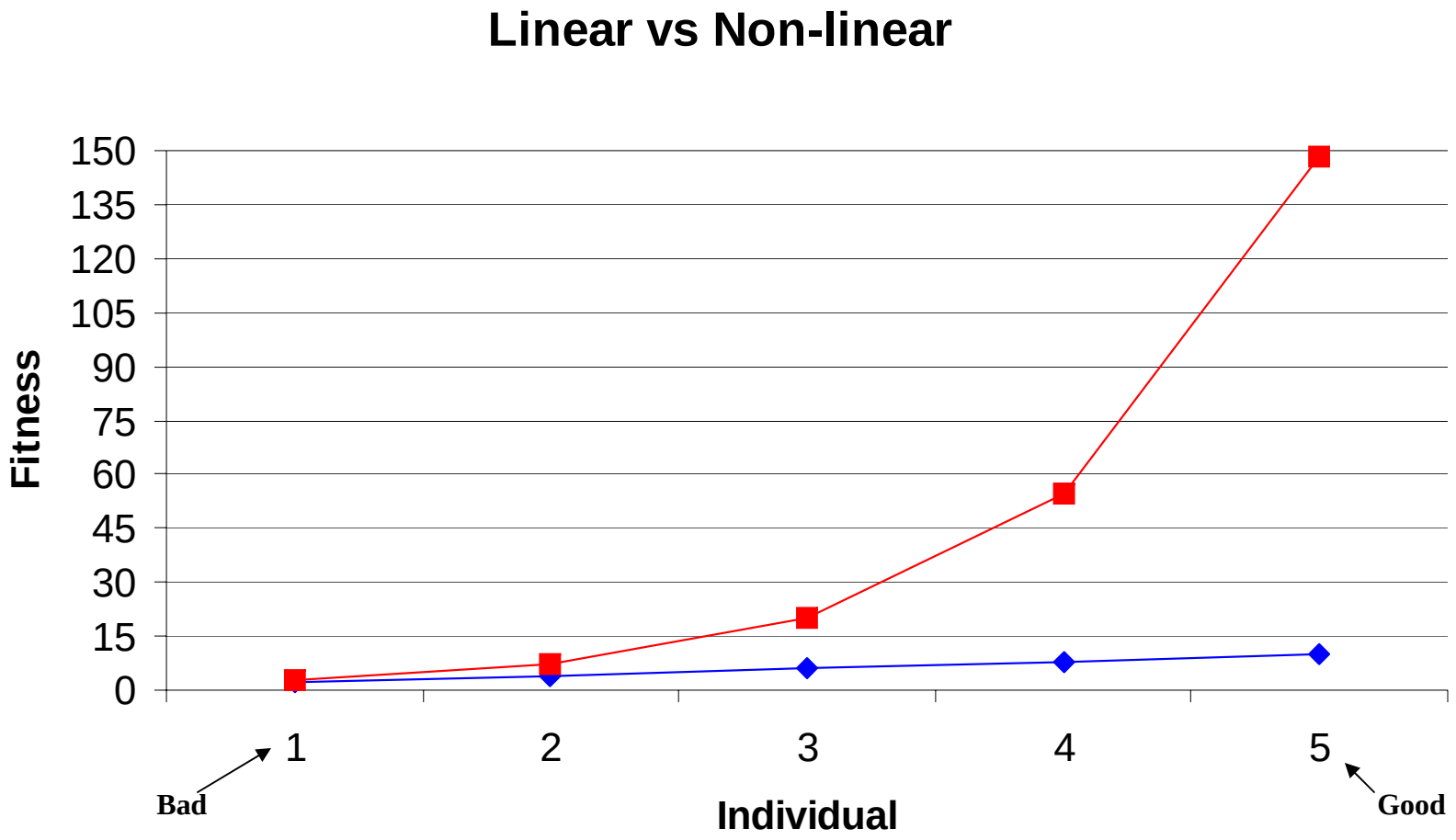


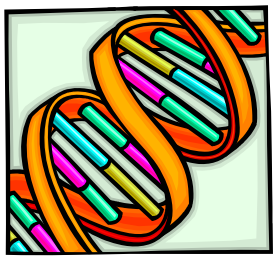
Fitness Function

- Encodes the problem space or environment
- Measures the “quality” or “strength” of an individual
  - Is this individual a good solution for this problem?
- Usually some mathematical expression
  - Linear (small variation between good / bad)
  - Non-linear (big variation between good / bad)



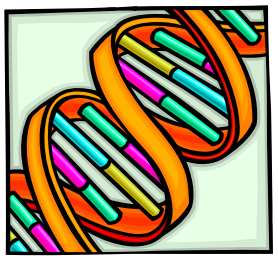
Fitness Comparison





Linear Fitness Example

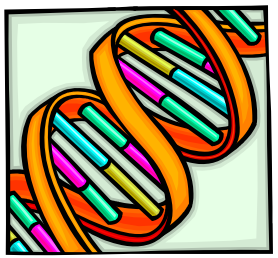
$$\text{Fitness(i)} = \sum_{j=0}^{9} \sum_{k=j+1}^{9} \Delta(gene(i, j), gene(i, k))$$



Non-linear Fitness Example

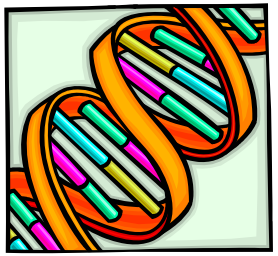
$$\text{Fitness}(i) = \frac{(\text{Likelihood}(i) \times \text{Intensity}(i))^{y}}{\text{Niche Size}(p, i)}$$

i = individual (a test case)
p = population (a set of test cases)
y = user-defined scaling factor



Selection Function

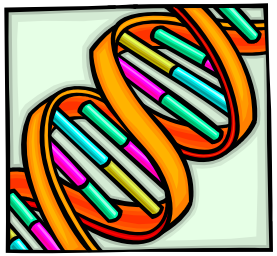
- Selects parents from the current population
- Selection is based on fitness
- Individuals must be above some criteria (i.e., fitness thresold) to survive the selection process
- Used to converge the population



Example

- Fitness Proportional Selection
  - Select parents based on proportion of the sum of the fitness values

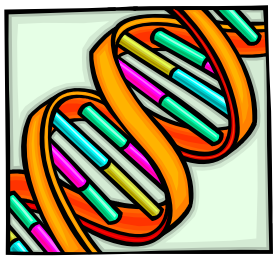
Individual	Linear	Proportion	Non-linear	Proportion
1	2	0.0667	2.718281828	0.0116
2	4	0.133	7.389056099	0.0317
3	6	0.2	20.08553692	0.0861
4	8	0.267	54.59815003	0.2341
5	10	0.33	148.4131591	0.6364
Fitness Sum	30	1	233.204184	1



Example

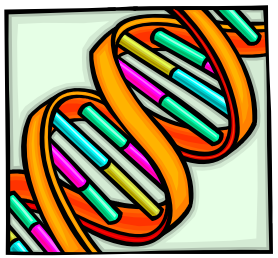
- Above Average Selection
  - Select parents who are greater than or equal to the population fitness average
  - Less computation / more aggressive convergence

Individual	Linear	Non-linear
1	2	2.72
2	4	7.39
3	6	20.09
4	8	54.60
5	10	148.41
6	12	403.43
7	14	1096.63
8	16	2980.96
9	18	8103.08
10	20	22026.47
Fitness Average	11	3484.38



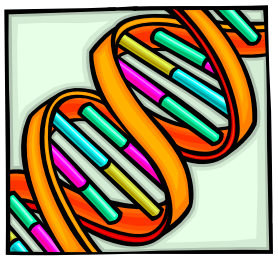
Genetic Operators

- Genetic operators are used to:
  - Introduce new “DNA” into the population
  - Control convergence / divergence of the population
- Two primary operators:
  - Crossover
  - Mutation
- There are trade-offs to consider in setting the type of operators and their frequency of use
  - Examples: too much mutation → population diverges

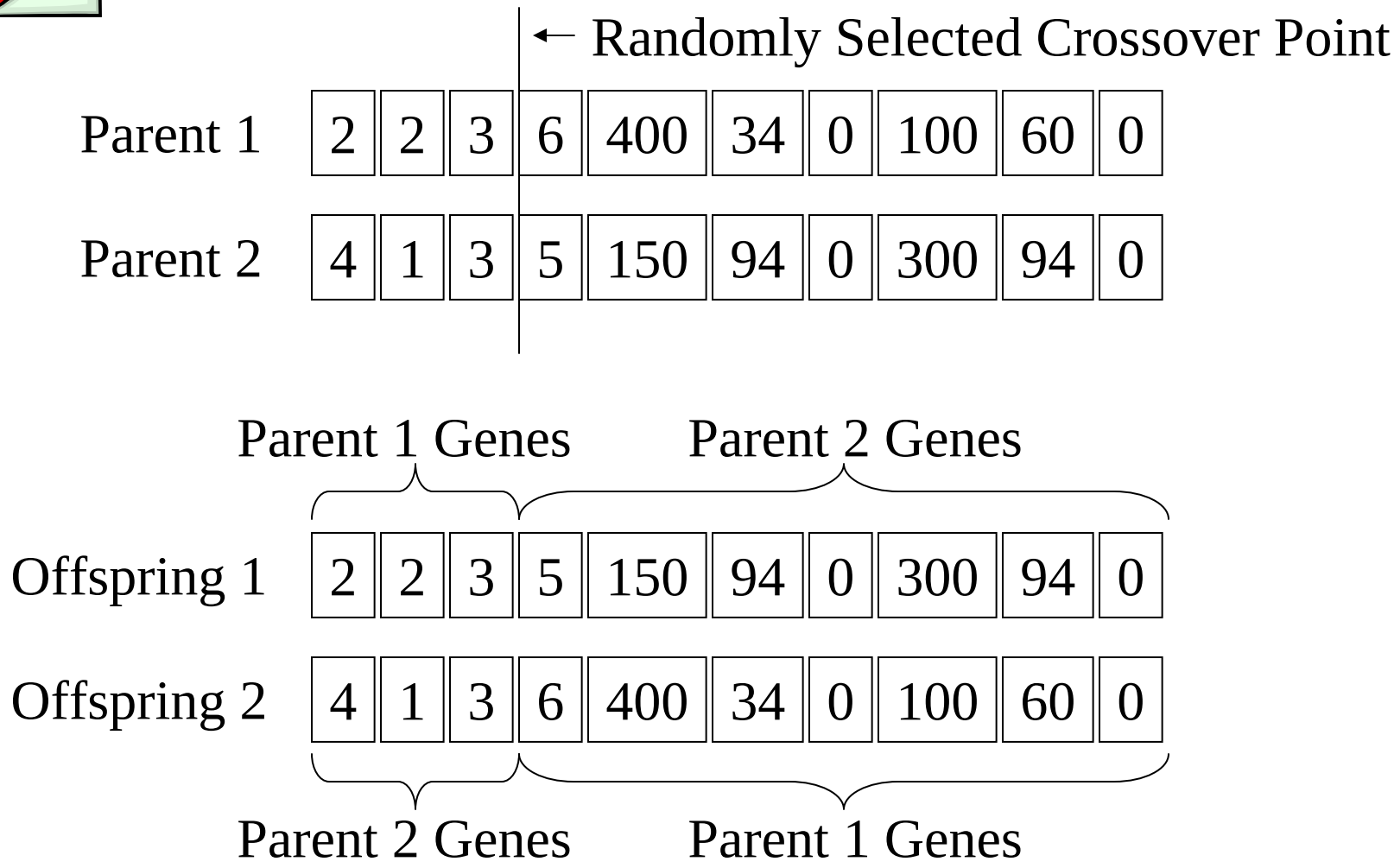


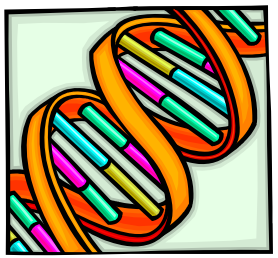
Crossover

- Takes 2 parents and creates 2 children
- No new “DNA” material, just a different combination of existing “DNA”
- Different approaches:
  - 1 point / 2 point
  - Permutation specific
- Crossover rate can be pre-defined or variable



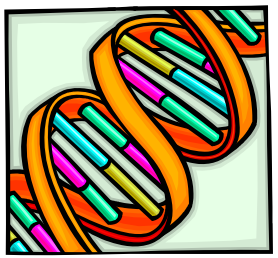
1 Point Crossover



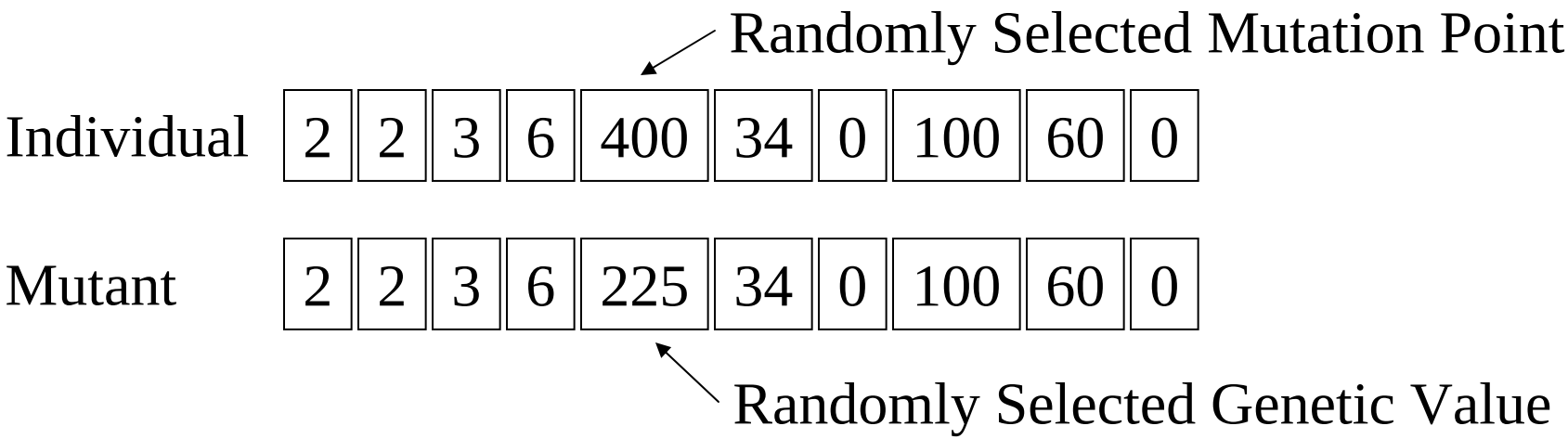


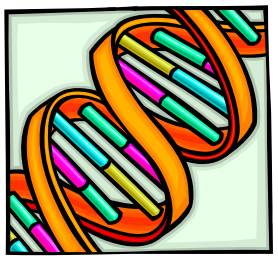
Mutation

- Takes 1 individual and creates 1 mutant
- New “DNA” material is introduced into the population as well as a different combination
- Different approaches:
  - 1-point / Permutation specific
- Mutation rate can be pre-defined or variable
- Causes more divergence than crossover



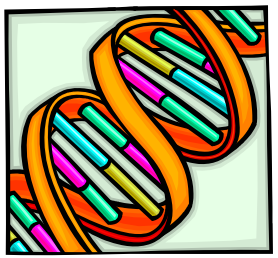
1 Point Mutation





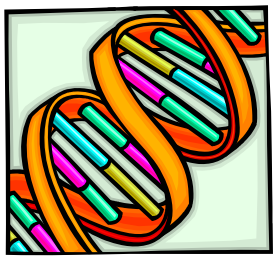
Critical Components

- Two very critical components:
  - Encoding (how to represent the solution)
  - Fitness function (how to evaluate the solution)
- Dependencies
  - Fitness function sometimes depends on encoding
  - Selection depends on the fitness function
  - Crossover / Mutation depend on the encoding



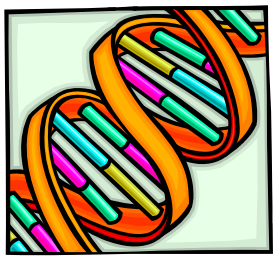
Applications

- Combination Lock
- Adaptive Sampling
- Gryffin



Combination Lock Example

- Combination lock consisting of 12 numbers
  - Only 1 set of 12 numbers will “open” the lock
- Each number of the combination ranges from 0 – 9 (inclusive)
- 10^{12} (1 trillion) possible solutions
- 1 global maximum, many local maximums

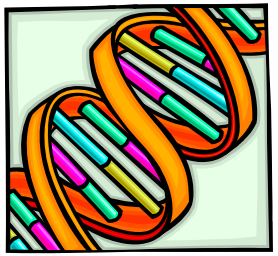


ComboGA

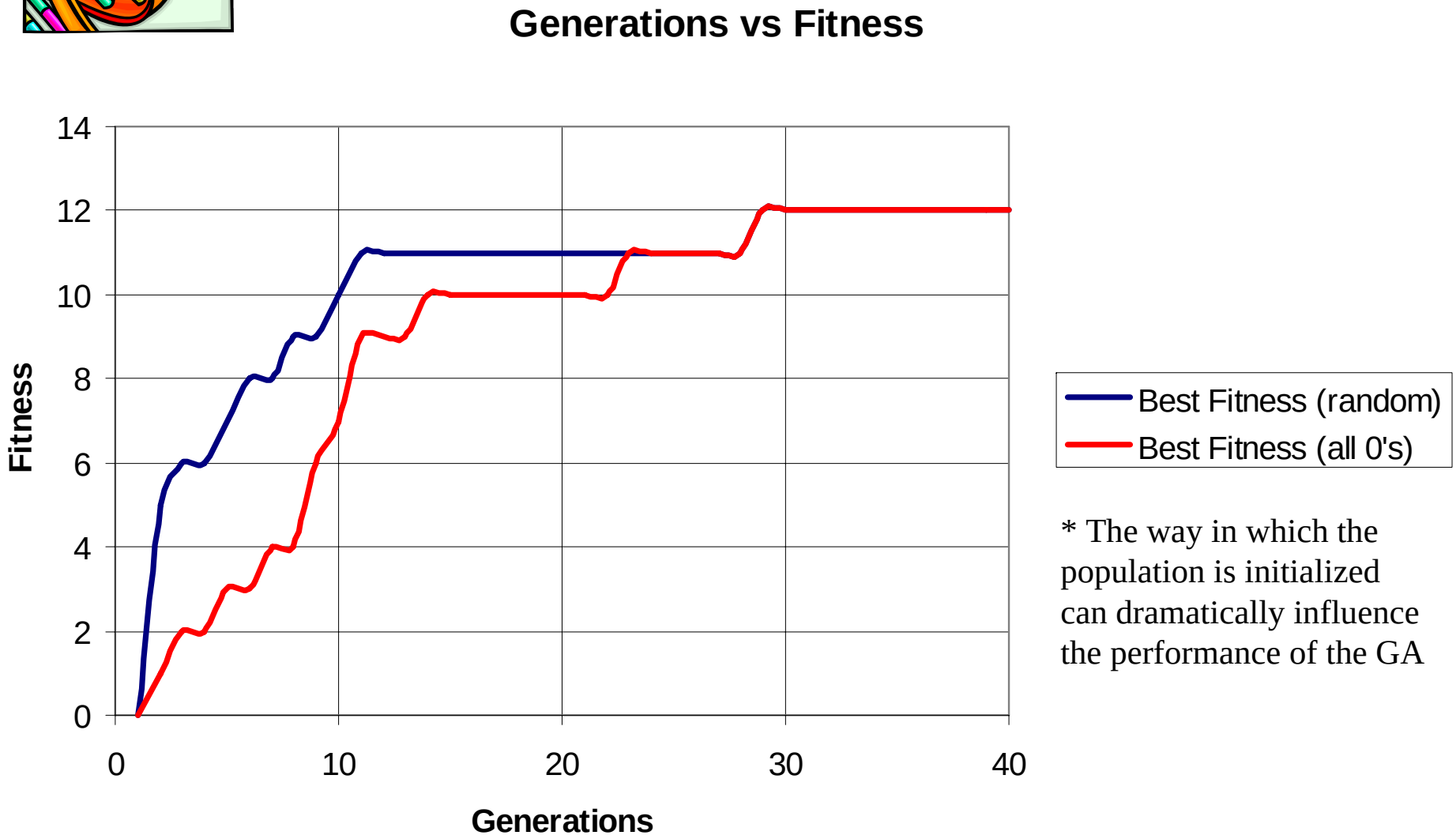
- GA population: 100 (initialized randomly)
- Encoding: 12 genes in the DNA with each gene representing a single digit in the combination
- Linear fitness function: Count the number of genes that match the combination lock (doesn't say which ones, just how many)
- Maximum fitness is 12 (this "opens" the lock)
- Above Average selection function
- 1-point Crossover and Mutation

Example Individual

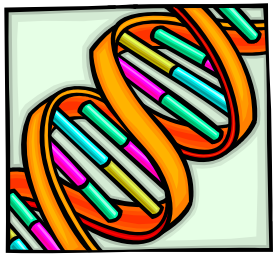
2	2	3	6	9	4	0	0	6	0	1	5



Result

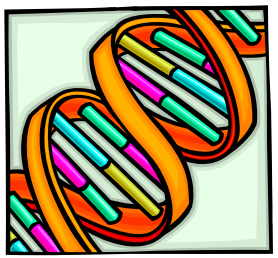


* The way in which the population is initialized can dramatically influence the performance of the GA



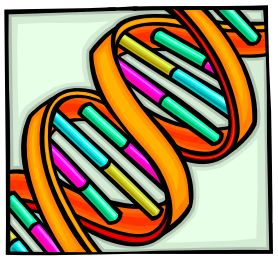
ComboGA Output

Generation	Best Fitness	Best Individual
1	0	000,000,000,000
2	1	000,000,000,000
3	2	000,000,000,010
4	2	000,000,000,010
5	3	100,000,000,700
6	3	100,000,000,700
7	4	100,001,000,700
8	4	100,001,000,700
9	6	100,000,036,710
10	7	102,001,400,710
11	9	102,001,436,710
12	9	102,001,436,710
13	9	102,001,436,710
14	10	102,901,436,710
…	…	…
23	11	102,921,436,710
…	…	…
29	12	102,921,436,714



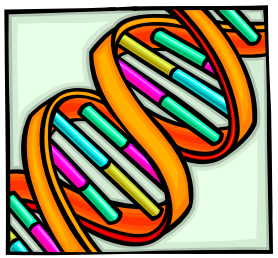
ComboGA Summary

- Correct solution was found very quickly
  - Not a brute force approach
  - Did not require expert rules
- Take the Best of the Best, and breed them to improve the population
- In the worst case, only 3,000 candidate solutions out of 1 trillion were tested to find the correct solution

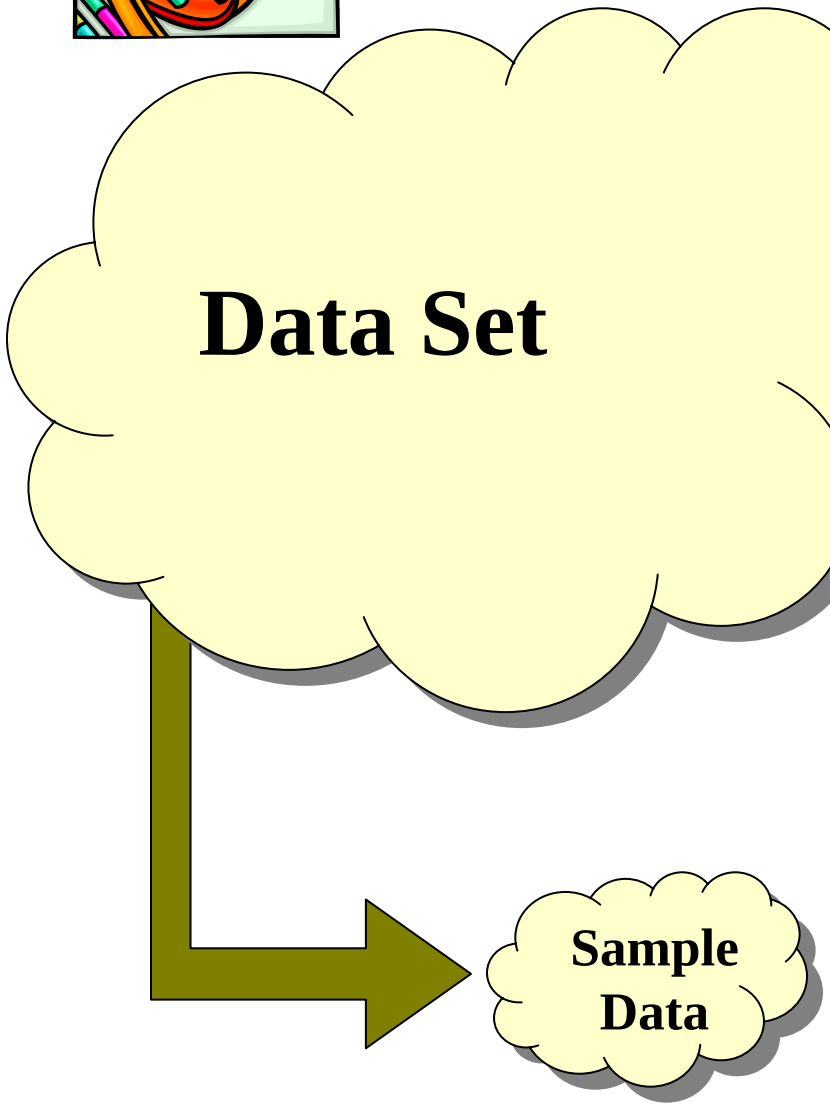


Adaptive Sampling

- NEED: Identify diverse values in a data set
- PROBLEM: Don’t have time to analyze every single data point
- GOAL: Find an ideal sample that represents the diversity without
  - Applying clustering techniques
  - Prior knowledge of what the categories of the population are



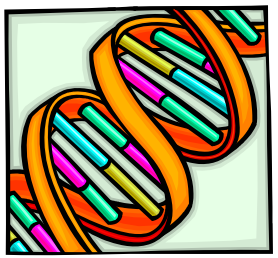
Background on Sampling



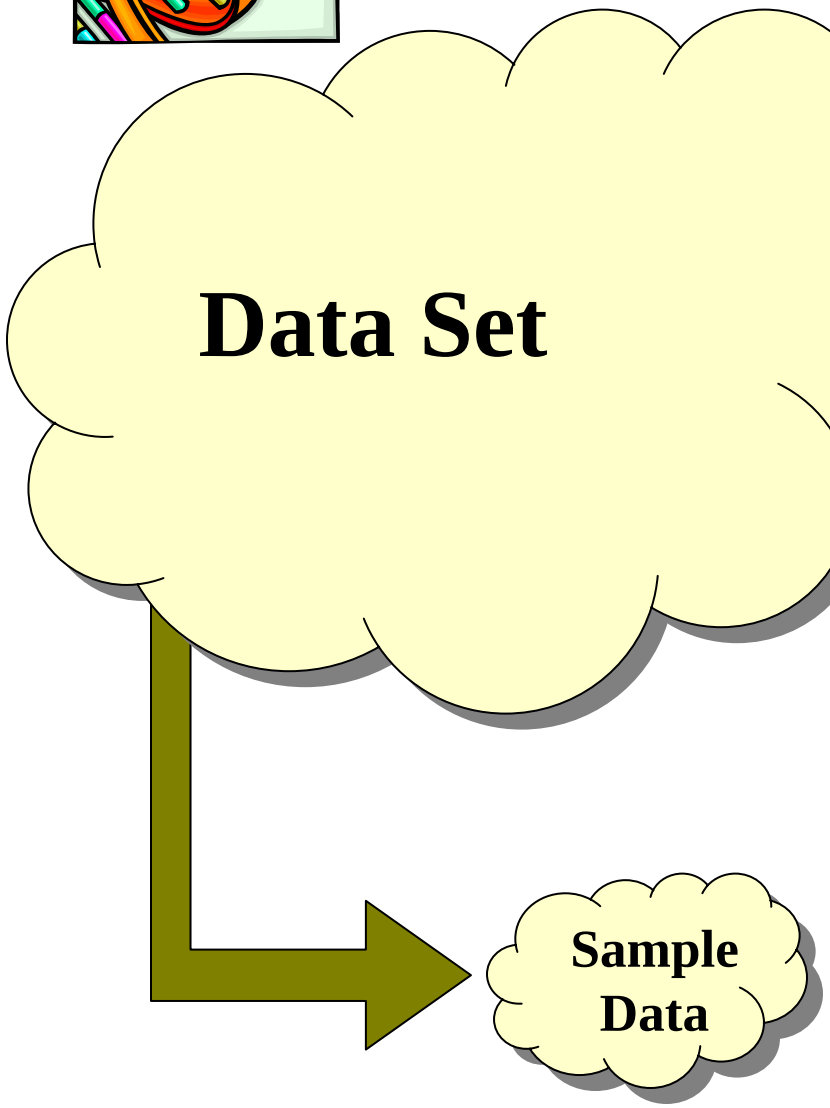
Problem: Characterizations about a large set of data must be made.

Solution: Sample the data set to obtain a smaller, more manageable data set

Types of Sampling:
-- Probabilistic
(smaller, but identical)
-- Non-probabilistic
(smaller, but NOT identical)

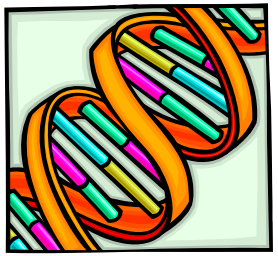


Non-probabilistic Sampling



Problem: How can we obtain an accurate, representative sample of the diversity of the entire data?

Solution: Adaptive sampling using a *Genetic Algorithm*.



Adaptive Sampling: How it works

Population

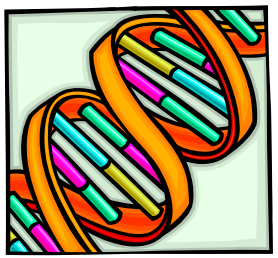
Resample based on results from evalution

Results of the evaluation are then used to alter the frequency of sampling, the method of sampling, or both

Evaluate the sample

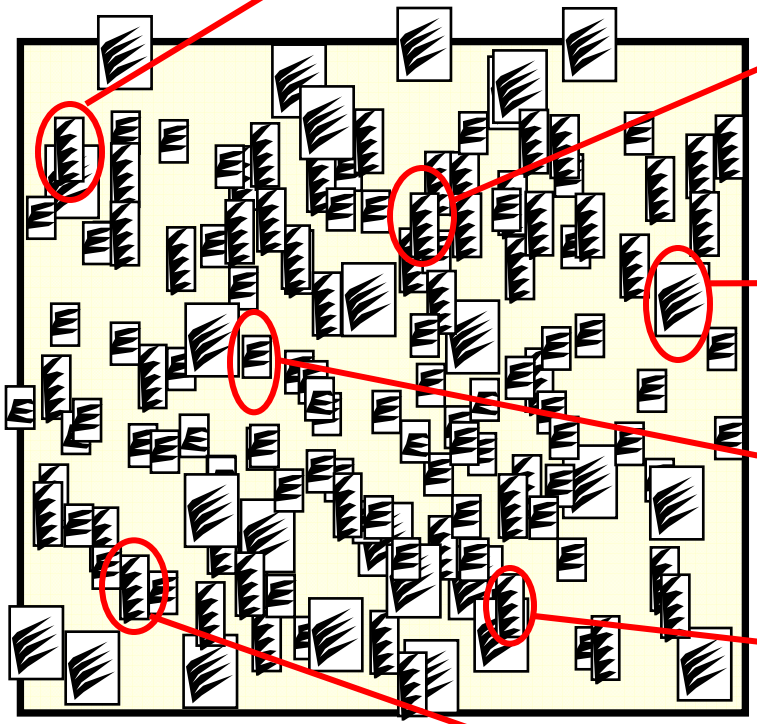
Each sample is evaluated to determine its quality

Sample

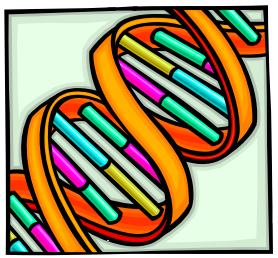


Genetic Algorithm

Sample Articles that are representative of the diversity of the entire set



Set of Newspaper Articles



Genetic Algorithm: A Simple Example

Data Set

Document ID	Title
D1:	China Intercepts U.S. Military Aircraft
D2:	Russia to Extend Life of Nuclear Military Fleet
D3:	U.S. -Based Chinese Military Expert Denied Release
D*m*:	….

1. Generate Samples

D1	D2	D3	….	D*N*
D3	D5	D6	….	D*N*
D1	D3	D6	….	D*N*
D2	D4	D7	….	D*N*

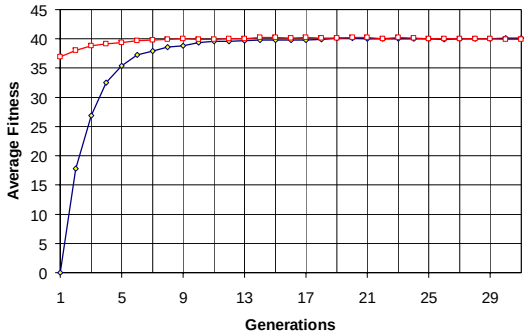
2. Evaluate Samples

$$\text{Fitness(i)} = \sum_{j=0}^{N} \sum_{k=j+1}^{N} \Delta(gene(i,j), gene(i,k))$$

3. Generate New Samples

D3	D2	D6	….	D*N*
D1	D5	D3	….	D*N*
D1	D4	D6	….	D*N*
D2	D3	D7	….	D*N*

4. Measure Performance



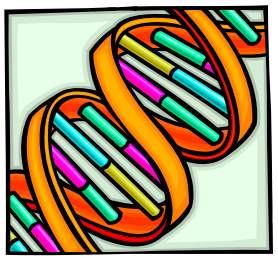
5. Final Sample

D1:

D4:

D6:

D9:

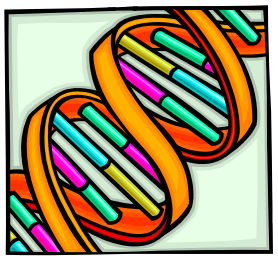


Genetic Representation

Sample size of *N*

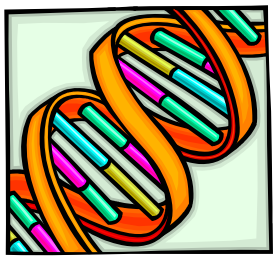
Document 1	Document 2	…	Document *N*

Gene 1 Gene 2 Gene *N*



Fitness Function Goal

- Achieve an ideal sample that represents diversity of population
- Without applying:
  - Clustering techniques
  - Prior knowledge of categories
- Goal: Find sample with the most dissimilarity between documents within the sample



GA Fitness Function

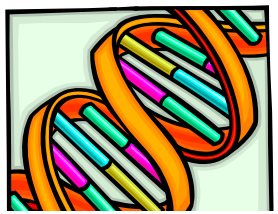
Calculate distance between VSM of gene j and k of the individual i

$$\text{Fitness(i)} = \sum_{j=0}^{N} \sum_{k=j+1}^{N} \text{Dissimilarity(Gene(i,j), Gene(i,k))}$$

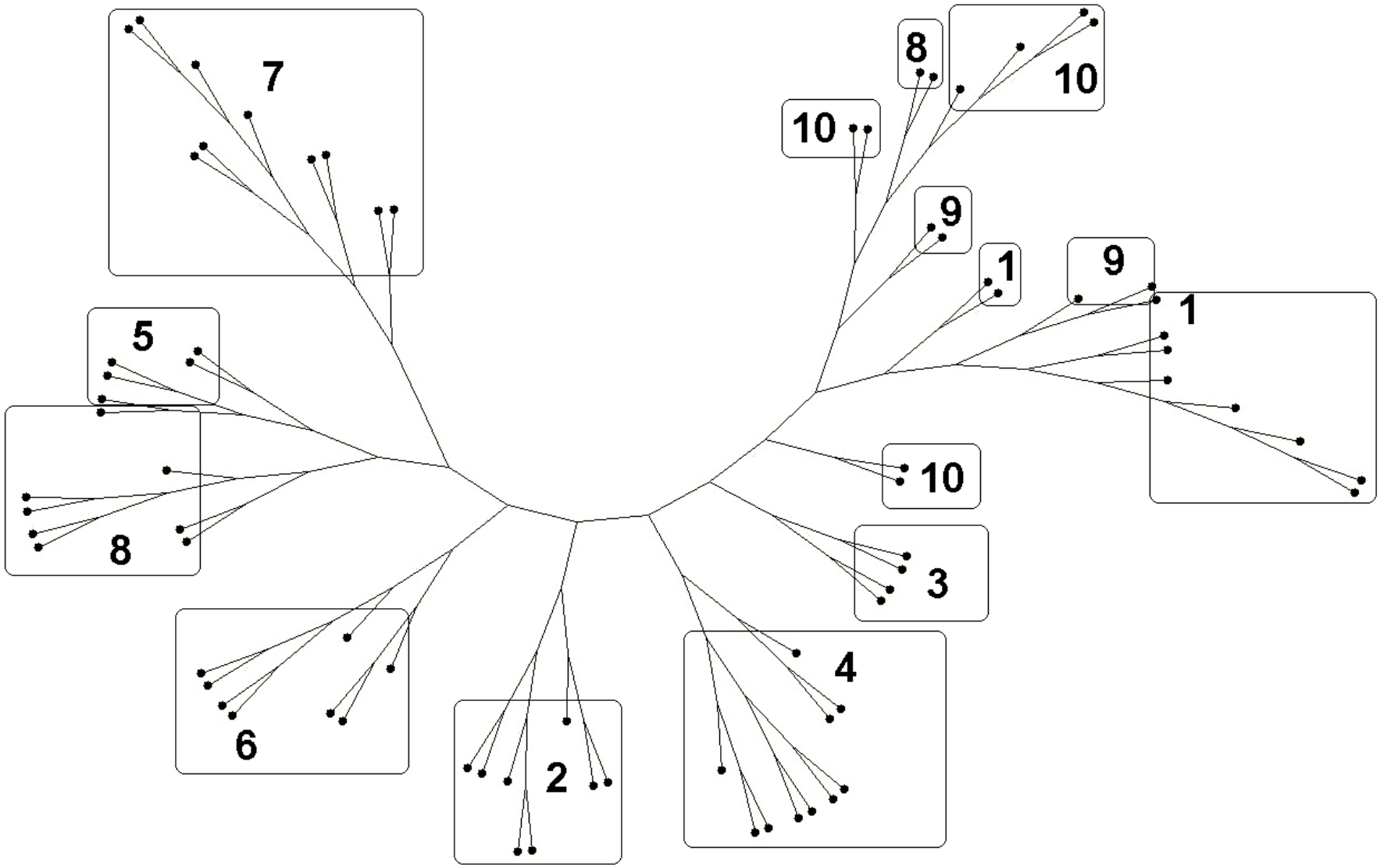
$$\frac{(N^2 - N)}{2} \text{ calculations}$$

Test Data Categories

	Category	Number of Articles
1	Airline Safety	10
2	Amphetamine	8
3	China Spy Plane Captives	4
4	Hoof and Mouth Disease	10
5	Korean Nuclear Capability	5
6	Mortgage Rates	8
7	Ocean Pollution	10
8	Saddam Hussein	10
9	Satanic Cult	4
10	Volcanoes	8

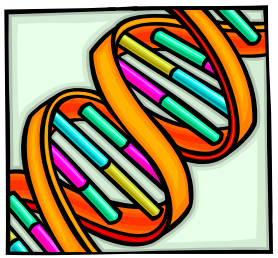


Cluster Diagram of Test Data



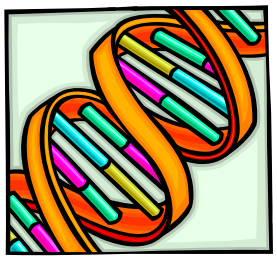
Comparison of Sample Distribution

Category	Initial Sample	Average Final Sample
Airline Safety	0	1.0
Amphetamine	0	1.0
China Spy Plane Captives	0	0.7
Hoof and Mouth Disease	0	0.9
Korean Nuclear Capability	0	0.8
Mortgage Rates	0	1.0
Ocean Pollution	0	1.1
Saddam Hussein	0	1.2
Satanic Cult	10	1.0
Volcanoes	0	1.3

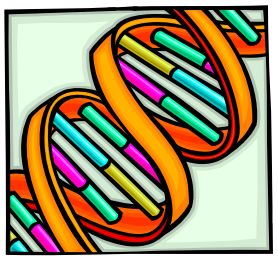


Adaptive Sampling Summary

- Impossible to analyze every member in a set to understand the whole document set
- Need to quickly scan document set and extract a representative sample
- Adaptive sampling with GA provides a solution
- GA is easily scalable to larger data, distributed sets

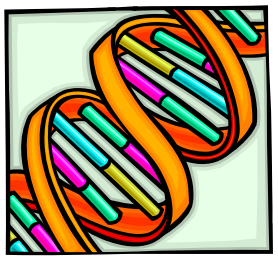


Gryffin



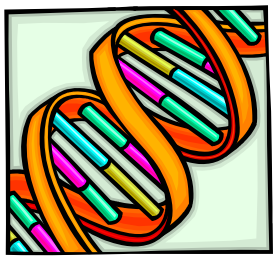
Problem

- Analysts must sift through large amounts of data to find evidence of events that:
  - have occurred (past)
  - are occurring (present)
  - may occur (future)
- Current news headlines help identify the present, but not the past or future
- Current technology can not automatically extract event information from raw text



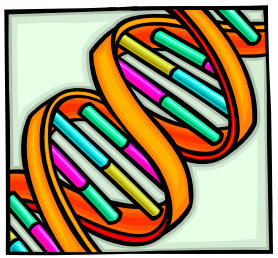
Questions

- If we can't "extract" events, can we accurately "detect" events?
- Analogous to detecting planets around distant stars
  - Distant planets are not detected directly but indirectly based on gravitational pull on the star
- Can events be detected based on occurrence of specific words or word phrases?



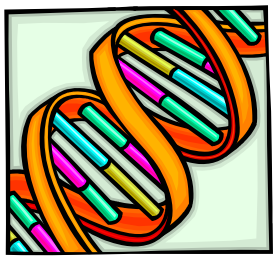
Events

- How are events characterized?
- Foundational characteristics
  - Characteristics that, when changed, dramatically alter the scenario
    - {ETA, Bombing, Trains}
    - {Al-Qaeda, Hijack, Planes}
- Descriptive characteristics
  - Characteristics whose presence enhances the detail, but does not dramatically alter the scenario (e.g., dates, times, colors)



Generic Events

- Foundational characteristics can be abstracted to general terms
  - People
  - Groups
  - Countries
  - Places
  - Infrastructure
  - Technology



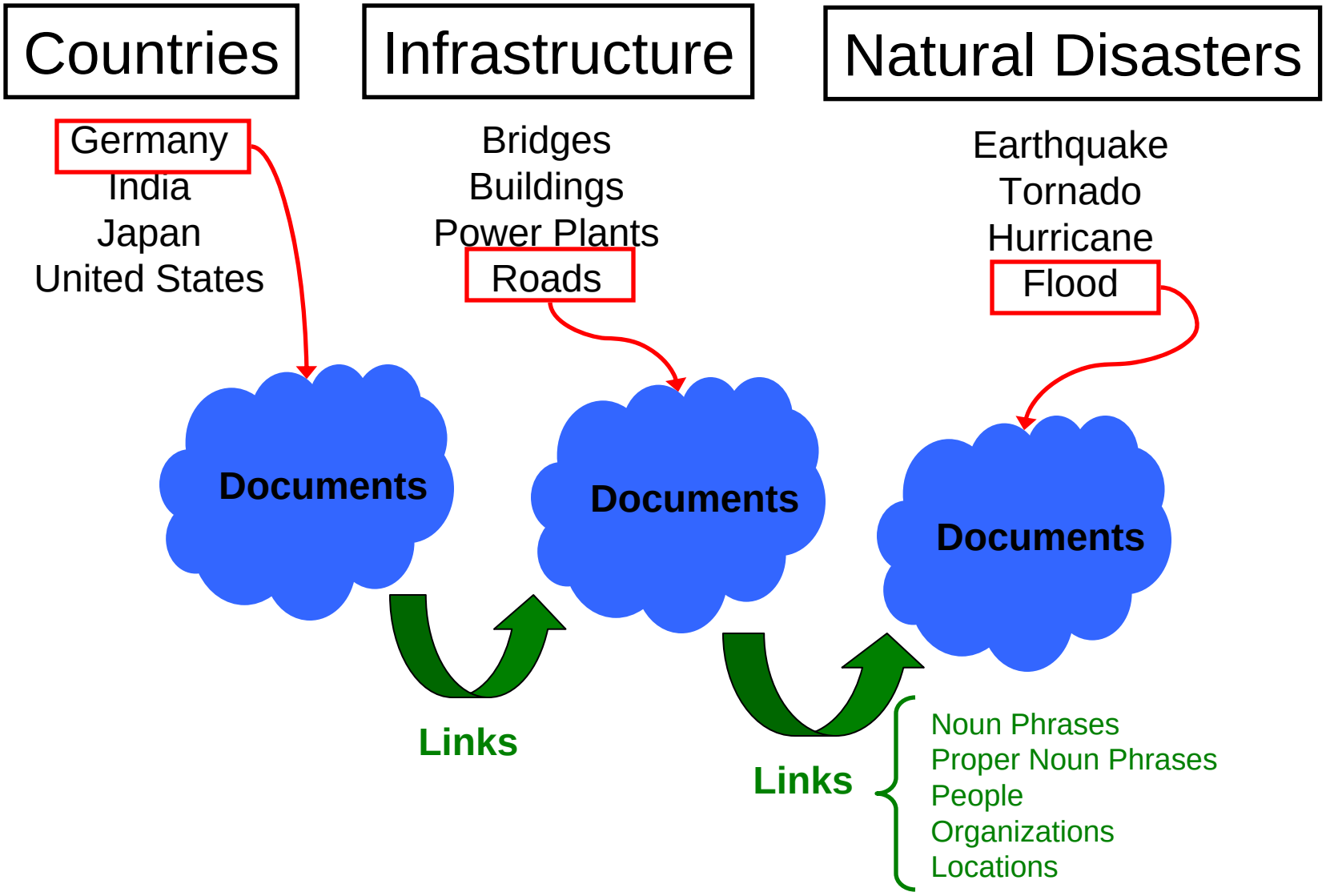
Example Event

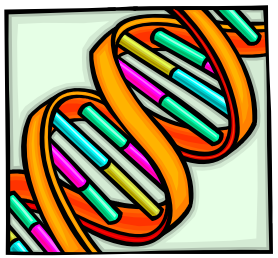
- Countries
  - **Germany, India, Japan, United States**
- Infrastructure
  - **Bridges, Buildings, Power Plants, Roads**
- Natural Disasters
  - **Earthquake, Tornado, Hurricane, Flood**

- Example Scenario: Germany – Flood – Roads

- 64 different combinations to explore !!!!

Conceptual Approach





Test Data

- FBIS data
- AP & Reuters news articles
- 1,000 total news articles
  - 366 Basketball (36.6%)
  - 240 Financial news (24.0%)
  - 162 Biological weapon (16.2%)
  - 98 Soccer (9.8%)
  - 75 Dirty bomb (7.5%)
  - 49 Gas prices (4.9%)
  - 10 Earthquake disaster (1.0%)

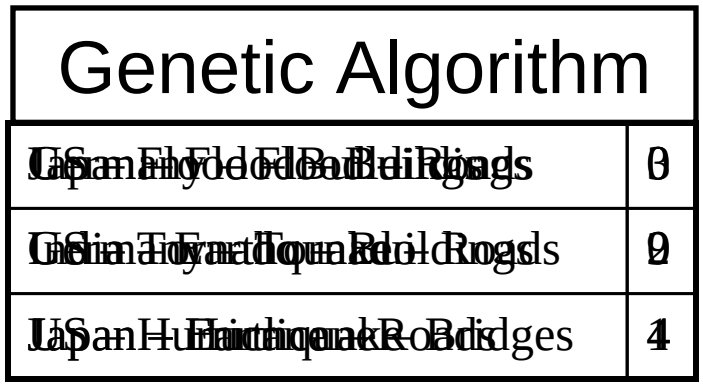
Technical Approach

1,000 documents

Reduce by Category:
- Country
- Natural Disaster
- Infrastructure

42 documents

Genetic Algorithm

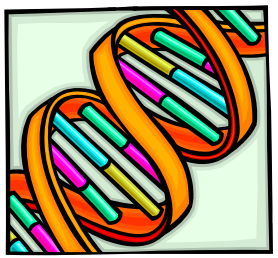


Significant reduction in effort on the part of the analyst to find the critical information

Final Scenario

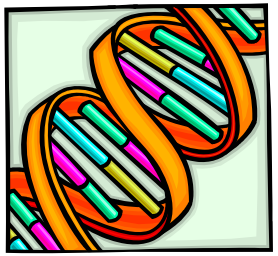
India – Earthquake
- 10 supporting documents
- 10 Proper noun phrases
- 34 Noun phrases

From 1,000 documents to 10 documents !



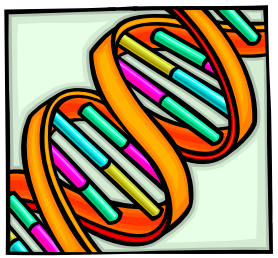
Proper Noun Phrases

- Pakistan High Commissioner in New Delhi
- Ahmedabad city
- Indian Air Force
- Richter scale
- Ahmedabad
- Bhuj



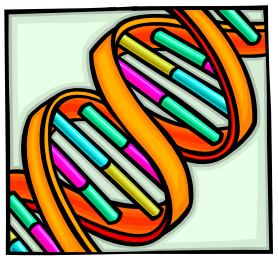
Noun Phrases

- engineering equipment
- disaster management
- devastating earthquake
- prime minister
- relief material
- earthquake-affected areas
- relief operations
- military hospitals
- MI-26 helicopters
- quake victims
- earthquake disaster
- medical experts
- death toll
- medical care
- government-run radio
- fresh tremors
- naval ships
- rescue operations
- relief supplies
- sniffer dogs



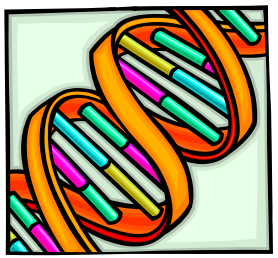
Sample Text

The death toll in the massive earthquake which rocked several parts of Gujarat yesterday has crossed 2,000. Hundreds are still feared trapped in the debris and the toll is likely to be much higher. Thousands in Ahmedabad and Bhuj have been rendered homeless. Bhuj, the epicenter of the quake measuring 6.9 on the Richter scale, bore the brunt of the devastation. Over 1,000 deaths have been reported from there.



Summary

- Large amounts of data pose significantly challenges
- Traditional techniques often break down as the data set size goes up
- Evolutionary Computation approaches can help people survive the data tsunami



References

- Illinois GA Lab (IlliGAL) run by Dave Goldberg
  - http://www-illigal.ge.uiuc.edu/
- GA Archive (Naval Research Lab)
  - http://www.aic.nrl.navy.mil/galist/
- Intro to GA
  - http://cs.felk.cvut.cz/~xobitko/ga/