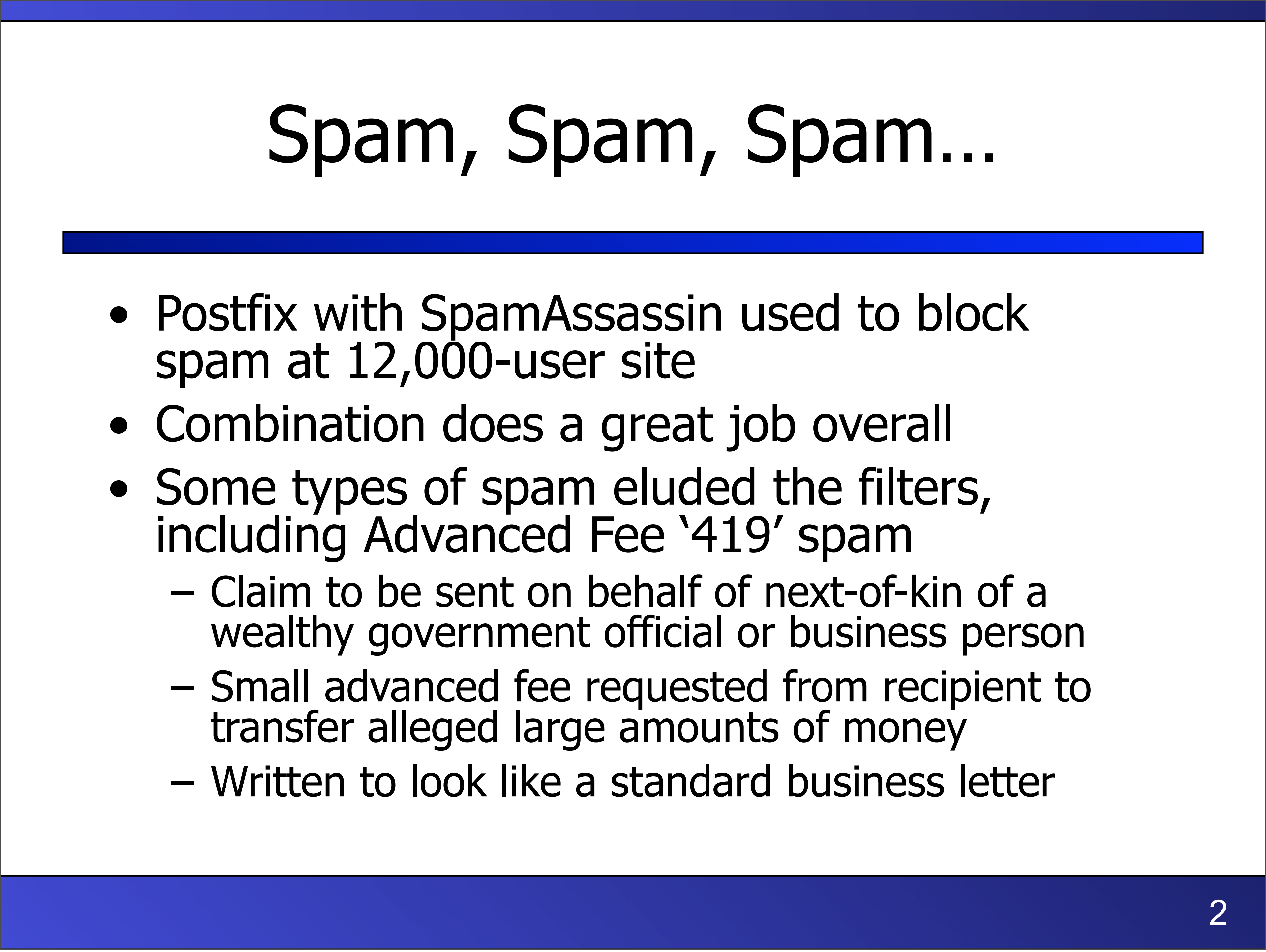
Genetic Regular Expressions: a New Way to Detect and Block Spam

Eric Conrad

April 2008



Spam, Spam, Spam…

- Postfix with SpamAssassin used to block spam at 12,000-user site
- Combination does a great job overall
- Some types of spam eluded the filters, including Advanced Fee ‘419’ spam
  - Claim to be sent on behalf of next-of-kin of a wealthy government official or business person
  - Small advanced fee requested from recipient to transfer alleged large amounts of money
  - Written to look like a standard business letter

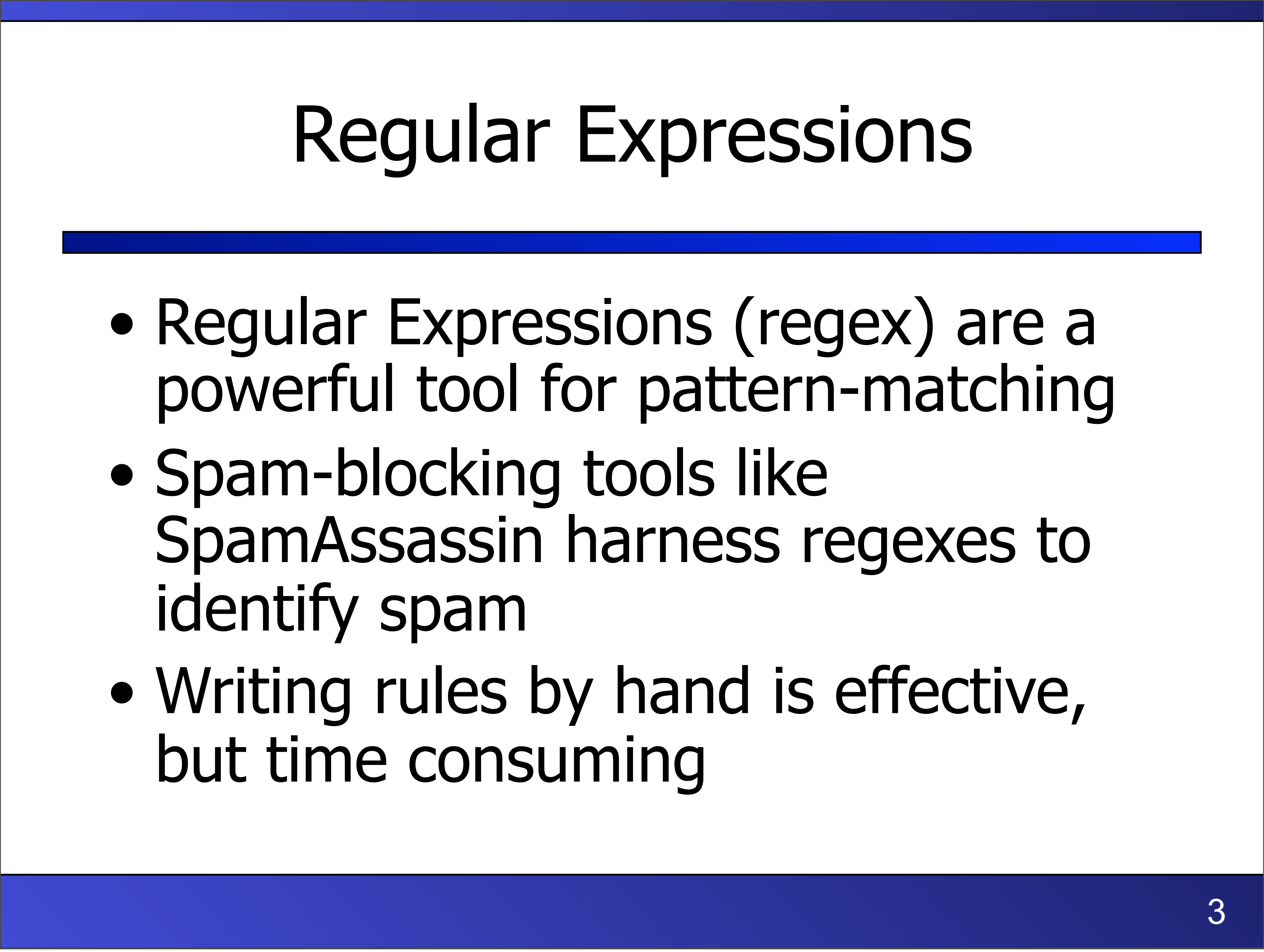
The company has 4 inbound mail relays, each running FreeBSD with the Postfix email server. Additional software includes SpamAssassin, an open source spam filter, and ClamAV, and open-source antivirus solution. Amavisd-new is used to interface Postifx, ClamAV,and SpamAssassin.

The solution works very well, blocking over 150,000 spam per business day for a company with 12,000 employees. A small percentage of spam does get through. The most common spam complaints received by the team were in regard to Advanced Fee (‘419’) spam.,

•Postfix: http://www.postfix.org/
•ClamAV: http://www.clamav.net/
•SpamAssassin: http://spamassassin.apache.org/
•Amavisd-New: http://www.ijs.si/software/amavisd/

Here’s an example 419 scam email received ‘in the wild’:

```
To:
Subject: Hello
From: "villaran nenita" <villaran1976_n@citromail.hu>
Date: Thu, 13 Sep 2007 12:16:56 CEST
Hello Dear,
My name is nenita villaran,The wife of Mr.Panfilo Nenita a senator during president joseph Estrada regime in Philippine.who is receently killed
in phi=
lippine.  During my husband's regime as a senator,I realized some reasonable amount of money from various deals that I successfully executed and
my buisness in the united state of which the Government has block most of my account in the bank of america,trying to leave me with nothing.
well before my late husband was killed,I secretly put in a box the sum of $30,000,000 million USD (Thirty million United states dollars) and
deposit i=
t in a security company abroad.I am contacting you because I want you to help me in securing the money for the future of my children since the
government now monitor all my movement=
 .
I hope to trust you as who will not sit on this money when you claim it.i will give you 15% of the total money for your assistance.if you are
willing to help me as soon as possible
Best regards=20
Villaran Nenita
```



Regular Expressions

- Regular Expressions (regex) are a powerful tool for pattern-matching
- Spam-blocking tools like SpamAssassin harness regexes to identify spam
- Writing rules by hand is effective, but time consuming

Regular expressions are a powerful mechanism for matching text. A simple regular expression matches literal text, like ‘/This is a literal match/’.

‘

Regexes may also contain advanced features such as metacharacters, character classes, and grouping (among others).[1]

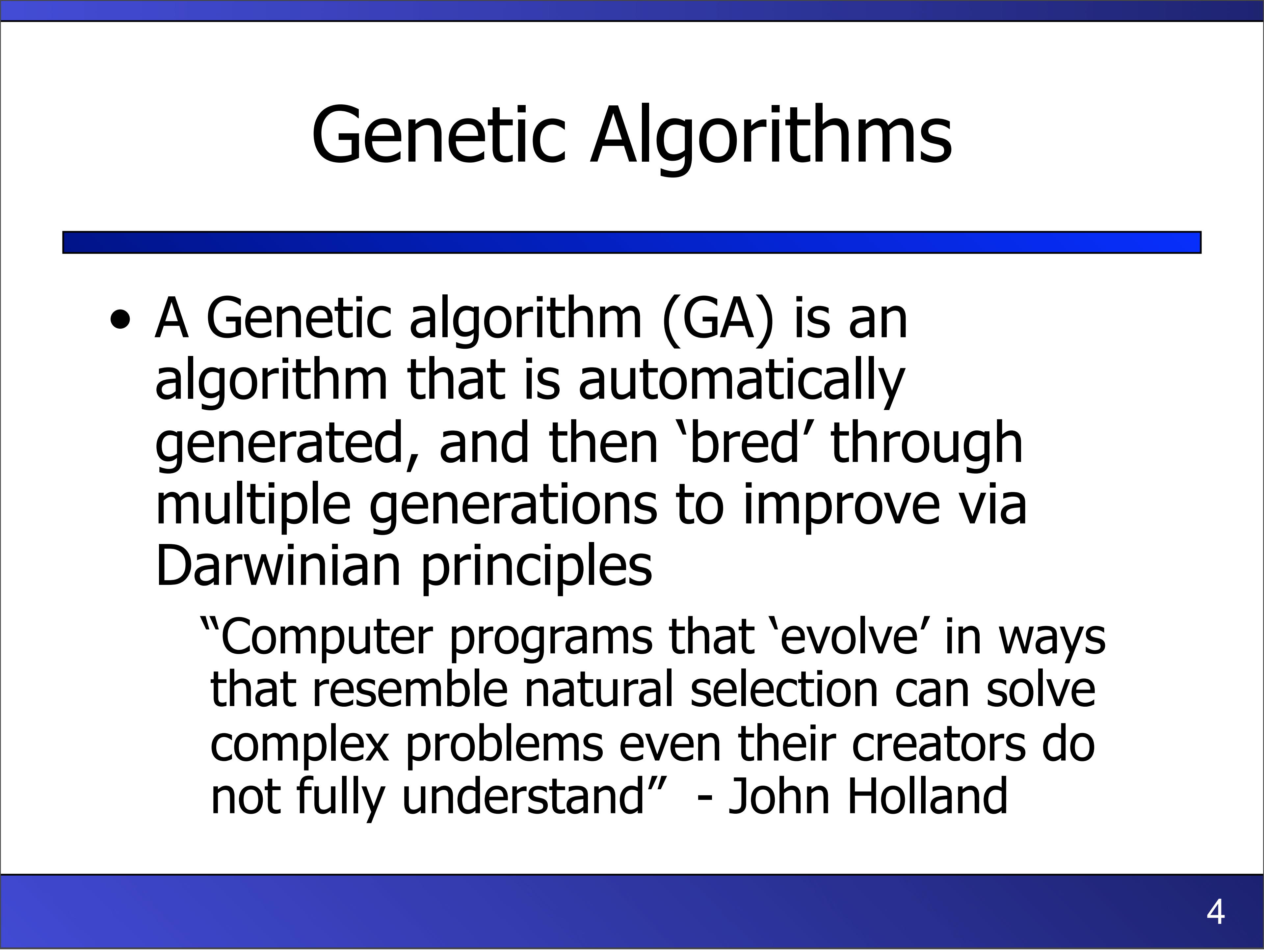
Metacharacters are concise commands which perform a function, such as ‘match the start of line’ (“^”). Character classes allow a range of characters, such as [a-z] for lowercase characters. Grouping allows a choice of words, like “(FEE|FIE|FOE|FUM)”.

This example uses all three types:

```
/^You are in a maze of twist(y|ing) little passages[, .]/
```

This regex matches lines beginning with “You are in a maze of …”, allows the choice of ‘twisty’ or ‘twisting’, and matches a space, period, or comma after ‘passages’.

[1] A good resource for learning about regular expressions is *Mastering Regular Expressions by* Jeffrey Friedl.. O'Reilly & Associates, January 1997. ISBN 0-596-00289-0.



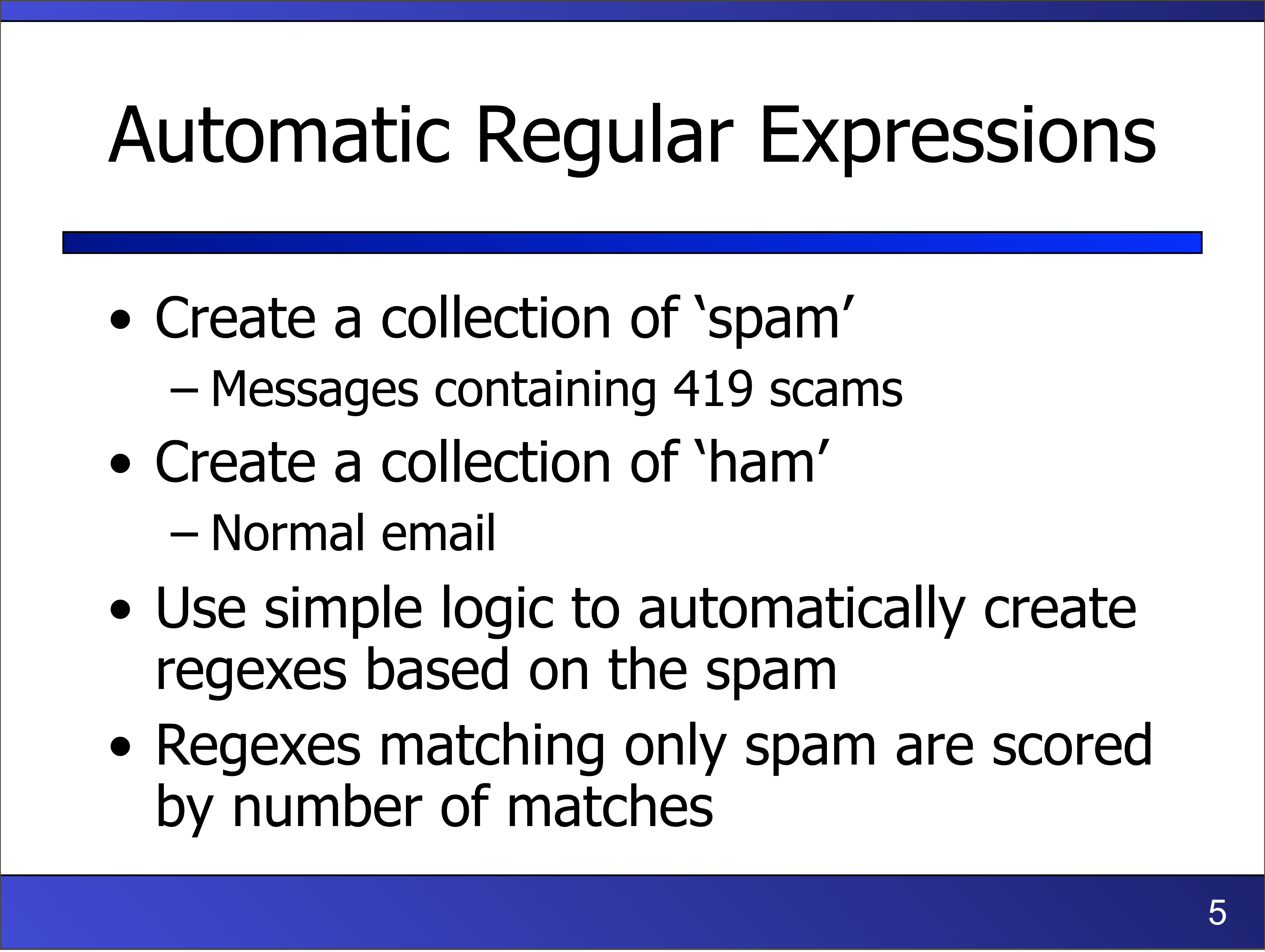
Genetic Algorithms

- A Genetic algorithm (GA) is an algorithm that is automatically generated, and then ‘bred’ through multiple generations to improve via Darwinian principles

  “Computer programs that ‘evolve’ in ways that resemble natural selection can solve complex problems even their creators do not fully understand” - John Holland

As genetic algorithm pioneer John Holland said “Computer programs that ‘evolve’ in ways that resemble natural selection can solve complex problems even their creators do not fully understand.”[1]

[1] Holland, John H. "Genetic Algorithms," *Scientific American,* July 1992, URL: http://www.econ.iastate.edu/tesfatsi/holland.GAIntro.htm



Automatic Regular Expressions

- Create a collection of ‘spam’
  - Messages containing 419 scams
- Create a collection of ‘ham’
  - Normal email
- Use simple logic to automatically create regexes based on the spam
- Regexes matching only spam are scored by number of matches

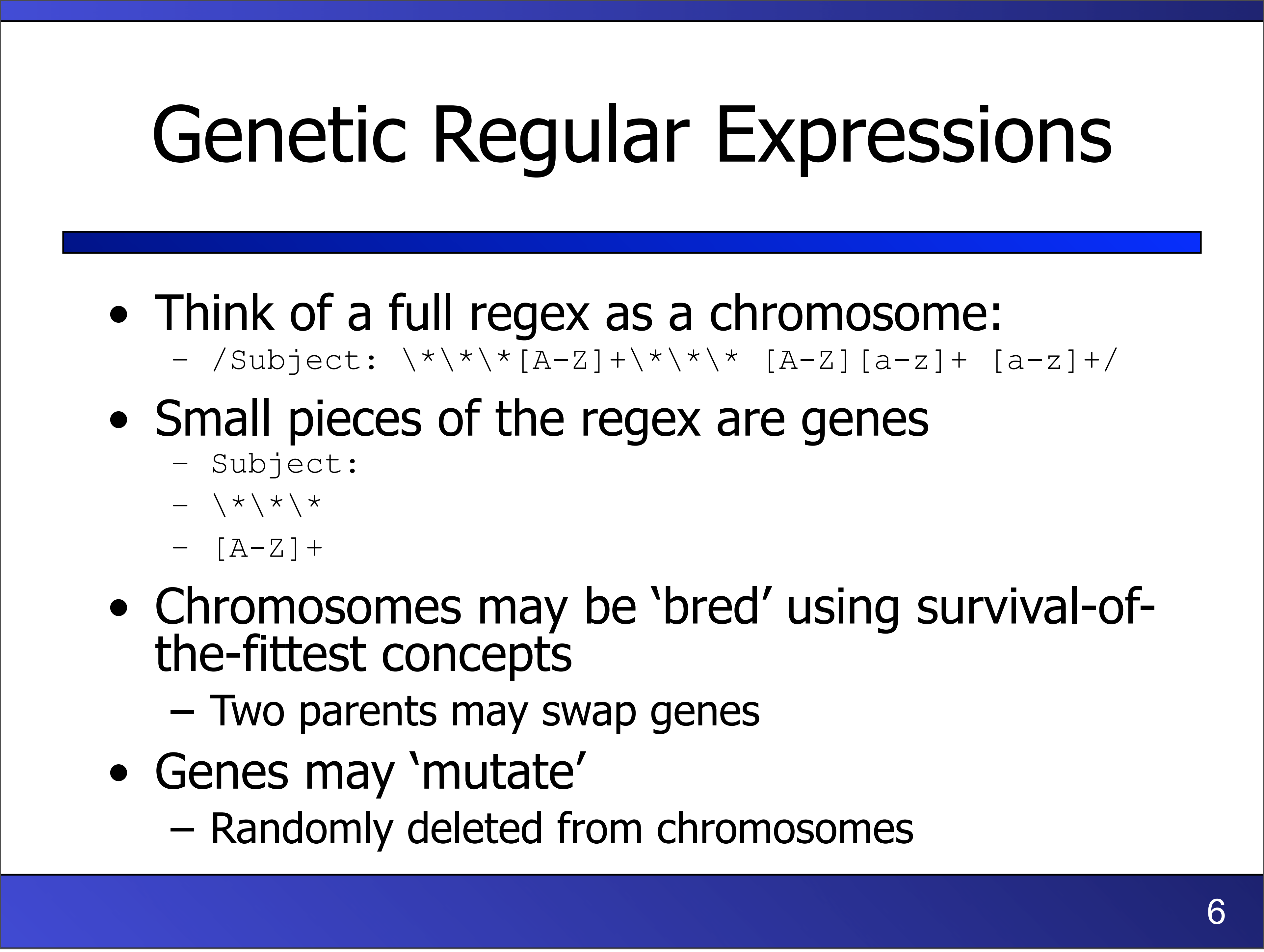
Paul Gram described Bayesian filtering to identify spam in his paper ‘A Plan for Spam.’[1] He described using a ‘corpus’ of ‘spam’ and ‘ham’, human-selected groups of spam and non-spam., respectively He then used Bayesian filtering techniques to automatically assign a mathematical probability that certain ‘tokens’ (words in the email) were indications of spam.

We will use the spam and ham corpus approach to identify the fitness of genetic regular expressions.

The regular expressions are generated using the following algorithm, applied in order (first match wins):

If it’s a number, replace with \d+ (1 or more digits)
If it’s capital hexadecimal, replace with [A-F0-9]+ (1 or more capital hex digits)
If it’s lowercase hexadecimal, replace with [a-f0-9]+ (1 or more lowercase hex digits)
If it’s a common TLD, replace with (com|net|org|edu|biz|info|us)
If it’s a lowercase word, replace with [a-z]+ (one or more lowercase letters)
If it’s an uppercase word, replace with [A-Z]+ (one or more uppercase letters)
If it’s an abbreviated day of the week, replace with (Mon|Tue|Wed|Thu|Fri|Sat|Sun)
If it’s an abbreviated month, replace with (Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)
If it’s a word with the first letter capitalized, replace with [A-Z][a-z]+ (uppercase letter, followed by one or more lowercase letters).
Else: keep the token as a literal regex.

[1] URL: http://www.paulgraham.com/spam.html



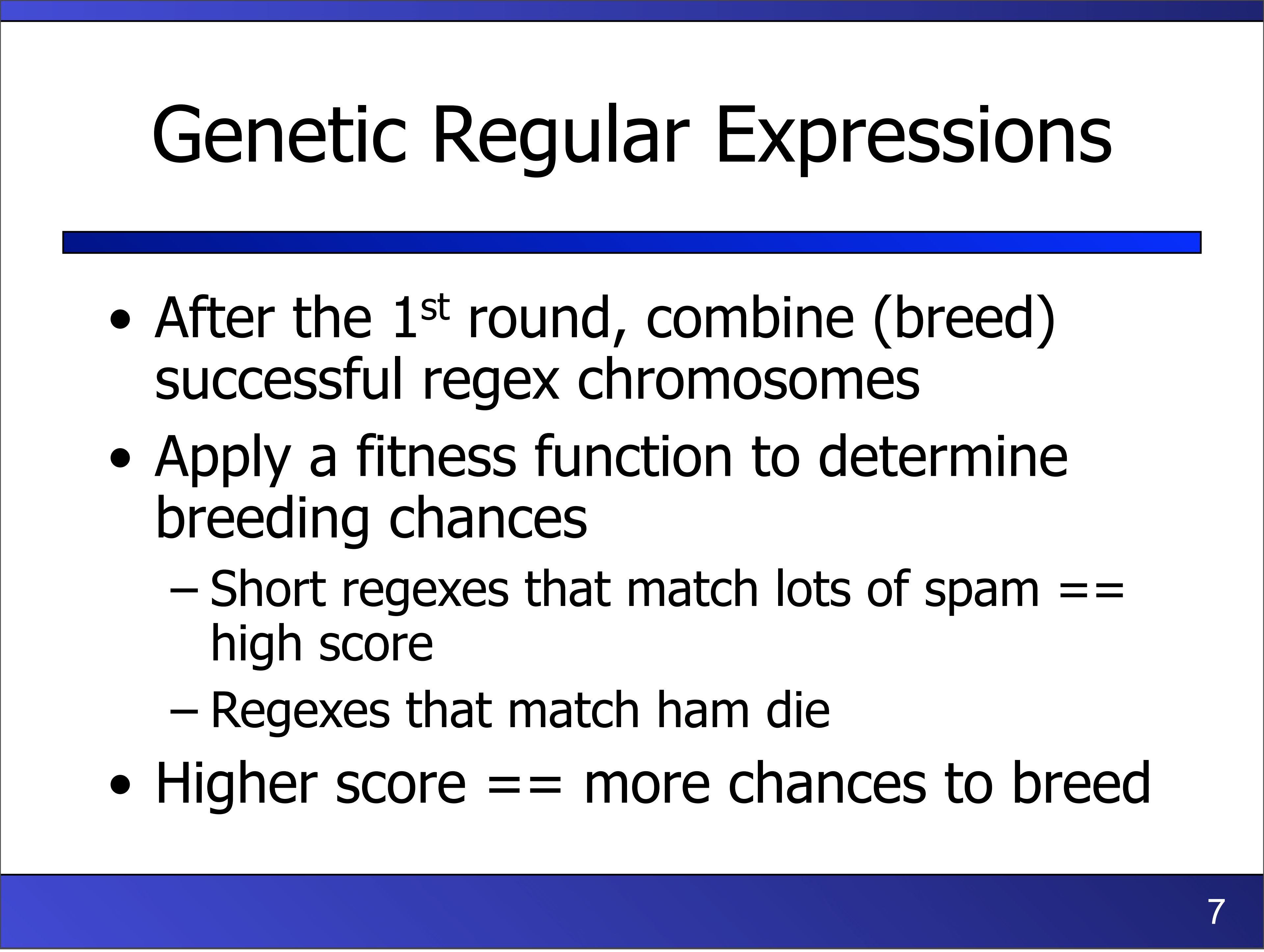
Genetic Regular Expressions

- Think of a full regex as a chromosome:
  - `/Subject: \*\*\*[A-Z]+\*\*\* [A-Z][a-z]+ [a-z]+/`
- Small pieces of the regex are genes
  - `Subject:`
  - `\*\*\*`
  - `[A-Z]+`
- Chromosomes may be ‘bred’ using survival-of-the-fittest concepts
  - Two parents may swap genes
- Genes may ‘mutate’
  - Randomly deleted from chromosomes

Genetic algorithms use chromosomes made up of individual genes. Chromosomes are designed to perform a specific function, such as sorting numbers, playing tic-tac-toe, or, in this case, matching spam text.

Chromosomes are scored according to a fitness function. Higher-scoring chromosomes are more likely to survive and are able to breed future generations. Breeding includes mating with other chromosomes and exchanging genes via a ‘crossover’ function. Stronger chromosomes are more likely to be chosen to breed.

Mutation may occur, where a random change is made to the chromosome. Many mutations may be damaging, but some could improve the ‘health’ of the chromosome.



Genetic Regular Expressions

- After the 1st round, combine (breed) successful regex chromosomes
- Apply a fitness function to determine breeding chances
  - Short regexes that match lots of spam == high score
  - Regexes that match ham die
- Higher score == more chances to breed

Chromosomes that match any ham ‘die’ and are ignored by the program. The surviving chromosomes are then scored by the fitness function.

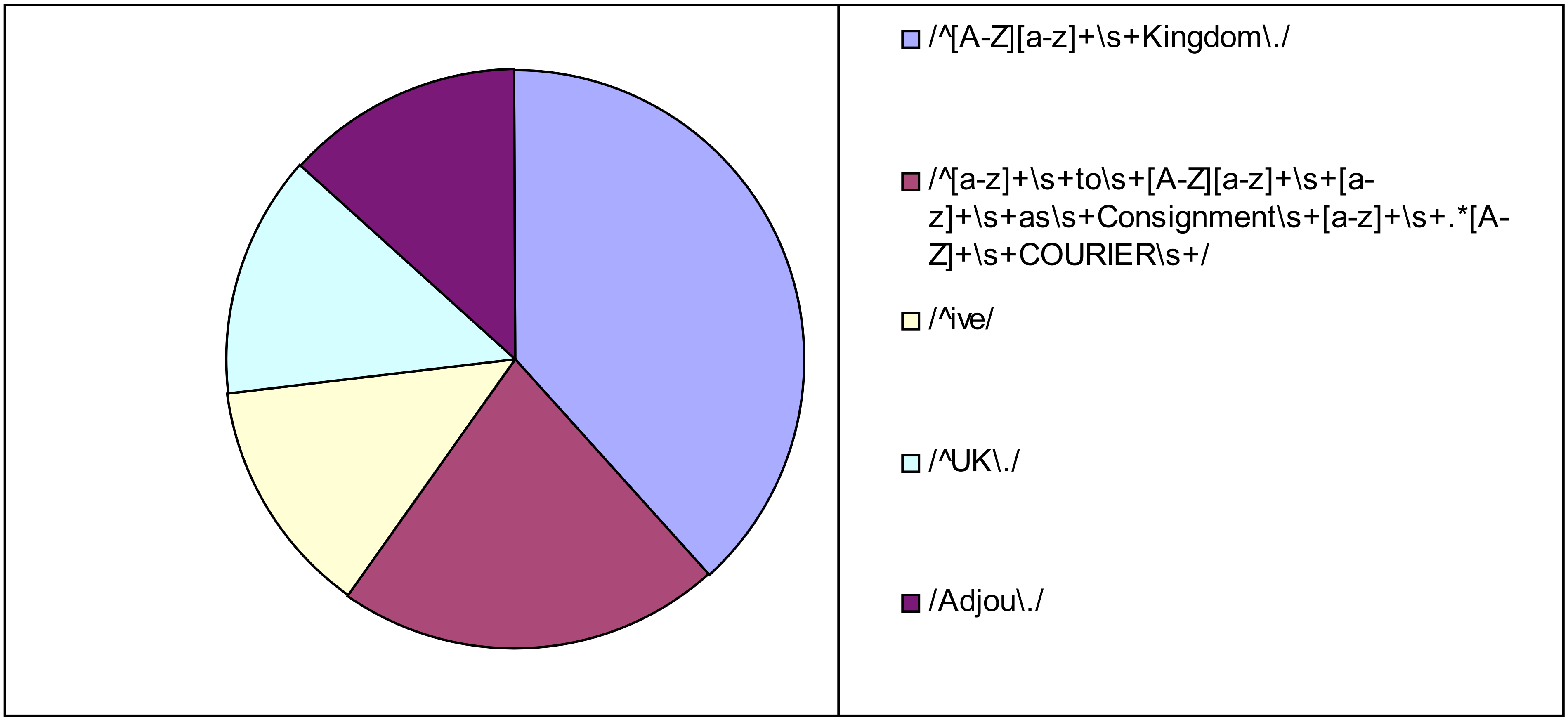
The fitness function has 2 components: number of spam lines matched, and length of the chromosome. Short chromosomes that match lots of spam receive high scores; long chromosomes that match little spam receive low scores.

Remaining chromosomes are scored according to this fitness function:

```
(Number of lines of spam matched) *  (1+((200-<length of chromosome>)/200));
```

In other words, any chromosome under 200 characters long will receive a fitness bonus; chromosomes over 200 characters receive a penalty.

Roulette Wheel Selection

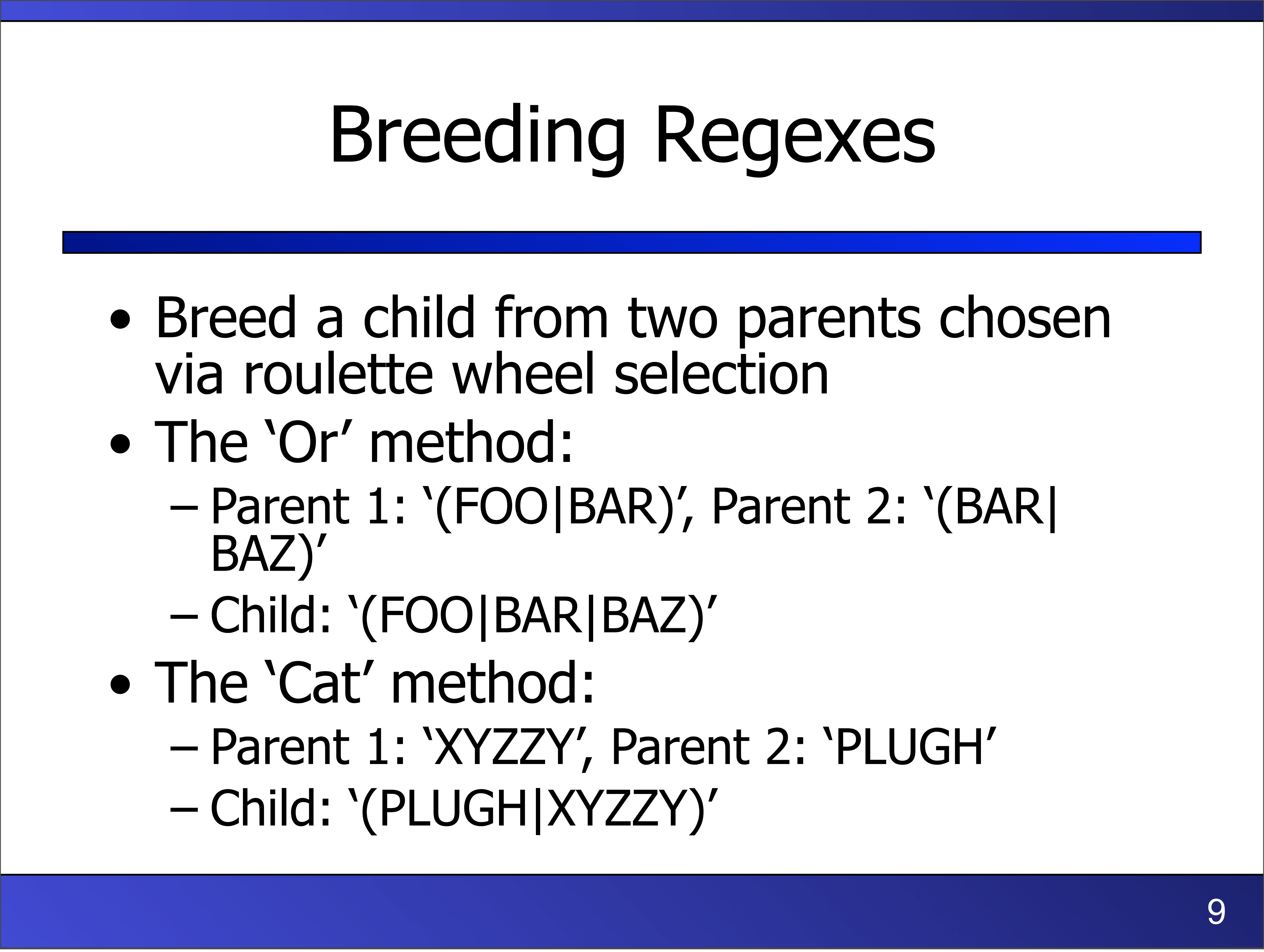


Roulette Wheel selection is a method for choosing chromosomes from a breeding pool. Higher-scoring chromosomes are more likely to be chosen than lower-scoring: a chromosome with a score of 3 is three times more likely to be chosen than a chromosome scoring 1.

An imaginary roulette wheel is constructed with a segment for each individual in the population... the size of the segment is based on the aptness of the particular individual. A fit individual will occupy a larger slice of the roulette wheel than a weaker one[1]

The above image contains the 5 highest-scoring chromosomes from the 1st generation of our results. '/^[A-Z][a-z]+\s+Kingdom\./' scored 5.64, compared with '/Adjou\./', which scored 1.96. The latter has a proportionally larger 'slot' in the roulette wheel, and is therefore more likely to be chosen for breeding.

[1] Dalton, John. *Newcastle Engineering Design Centre, January 2007: Genetic Algorithms (GAs).* URL: http://www.edc.ncl.ac.uk/highlight/rhjanuary2007.php



Breeding Regexes

- Breed a child from two parents chosen via roulette wheel selection
- The ‘Or’ method:
  - Parent 1: ‘(FOO|BAR)’, Parent 2: ‘(BAR|BAZ)’
  - Child: ‘(FOO|BAR|BAZ)’
- The ‘Cat’ method:
  - Parent 1: ‘XYZZY’, Parent 2: ‘PLUGH’
  - Child: ‘(PLUGH|XYZZY)’

Two parent chromosomes are selected via Roulette Wheel selection. A child is then bred via one of 2 methods:

The ‘Or Method’

An example of the ‘Or Method’ takes 2 parent chromosomes, such as ‘(FOO|BAR)’ and ‘(BAR|BAZ)’, breaks them down to their unique genes, and ‘ors’ them together to create a child called ‘(FOO|BAR|BAZ)’.

The ‘Cat Method’

The ‘Cat Method’ takes 2 parent chromosomes and concatenates them together, creating a longer chromosome.   Take these 2 parents:

```
        /(?:Coulibaly\.|LOTTERY\s+WINNING\s+|^deposit|the\s+finance)/
        /YOUR\s+[A-Z]+\s+[A-Z]+/
```

Simply concatenate them, taking (/Chromosome1/) and (/Chromsosme2/), creating the child (/Chromosome1|Chromsosme2/):

```
/
        (?:Coulibaly\.|LOTTERY\s+WINNING\s+|^deposit|the\s+finance|YOUR\s+[A-Z]+\s+[A-Z]+)/
```

Genetic Regex Pseudocode

- Loop until final generation:
  - Generate automatic chromosomes based on a portion of spam
  - Score chromosomes
  - Keep the fittest 3rd
  - Breed survivors
  - Mutate some of the children
  - Move to next slice of spam

This slide shows pseudocode used to breed regular expressions. In our example, we use 10 generations.

The actual code used for this presentation was written in Perl.

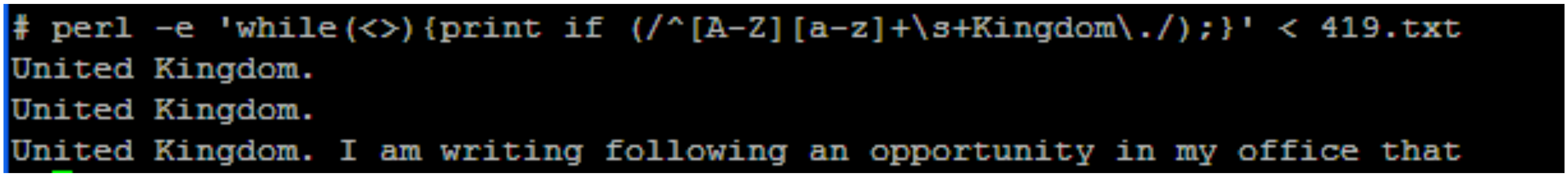
Proof-of-concept code, including the spam and ham corpora used in this presentation, may be downloaded from: http://files.ericconrad.com/genregex.tgz

The 1st Generation’s Winner

- This regex scored 5.64

  `/^[A-Z][a-z]+\s+Kingdom\./`

- Matches the string ‘United Kingdom’



```
# perl -e 'while(<>){print if (/^[A-Z][a-z]+\s+Kingdom\./);}' < 419.txt
United Kingdom.
United Kingdom.
United Kingdom. I am writing following an opportunity in my office that
```

The first generation contains automatic chromosomes, with no breeding or mutation (yet).

The fittest chromosome is ‘`/^[A-Z][a-z]+\s+Kingdom\./`’.

In plain English, that means ‘“A line of text beginning with a capital letter, followed by 1 or more lower-case letters, followed by one or more whitespace characters, followed by ‘Kingdom.’”

It matches 3 spam lines beginning with “United Kingdom”.

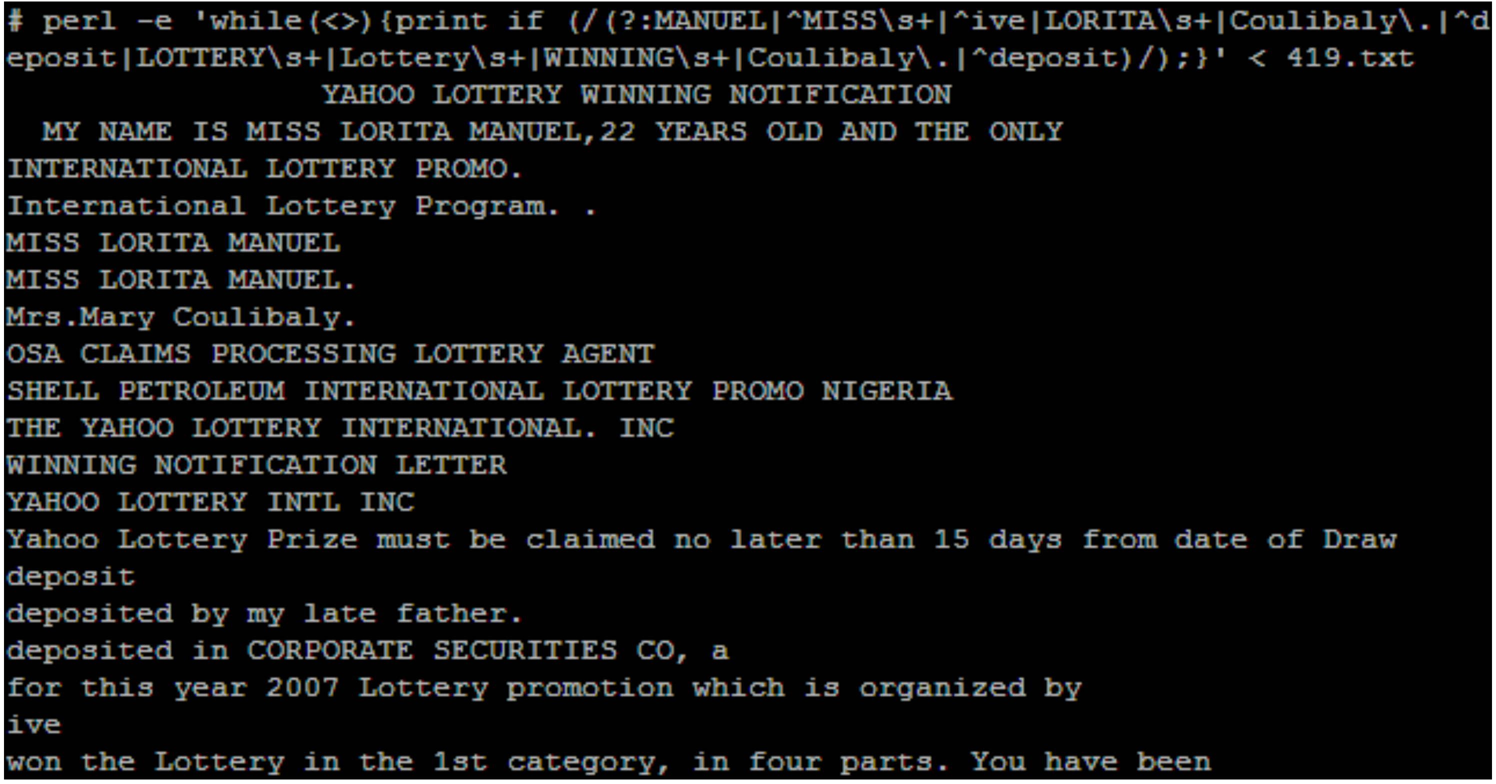
Although it only matches 3 lines, it receives a fitness bonus due to its short length, for a total fitness score of 5.64.   Note: to see the spam lines matched by each chromosome, use the following command:

```
perl -e 'while(<>){print if (/<REGEX>/);}' < 419.txt
```

This is a simple Perl script that takes standard input (the file 419.txt), checks each line against a regex, and prints the line of there is a match

The 5th generation

- This regex scored 27.65:



```
# perl -e 'while(<>){print if (/(?:MANUEL|^MISS\s+|^ive|LORITA\s+|Coulibaly\.|^d
eposit|LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.|^deposit)/);}' < 419.txt
                 YAHOO LOTTERY WINNING NOTIFICATION
  MY NAME IS MISS LORITA MANUEL,22 YEARS OLD AND THE ONLY
INTERNATIONAL LOTTERY PROMO.
International Lottery Program. .
MISS LORITA MANUEL
MISS LORITA MANUEL.
Mrs.Mary Coulibaly.
OSA CLAIMS PROCESSING LOTTERY AGENT
SHELL PETROLEUM INTERNATIONAL LOTTERY PROMO NIGERIA
THE YAHOO LOTTERY INTERNATIONAL. INC
WINNING NOTIFICATION LETTER
YAHOO LOTTERY INTL INC
Yahoo Lottery Prize must be claimed no later than 15 days from date of Draw
deposit
deposited by my late father.
deposited in CORPORATE SECURITIES CO, a
for this year 2007 Lottery promotion which is organized by
ive
won the Lottery in the 1st category, in four parts. You have been
```

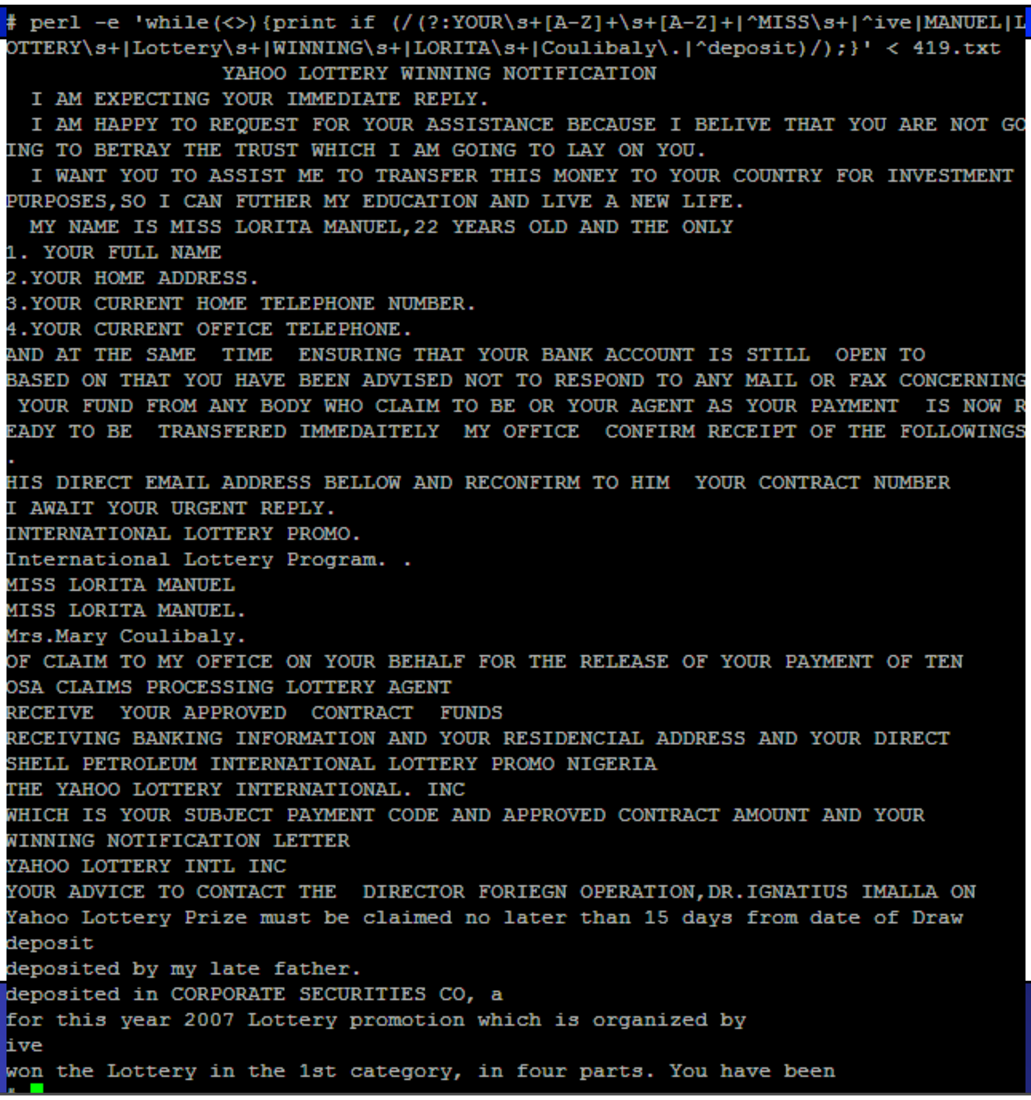
In generation 5, a new chromosome takes the lead:

```
/(?:MANUEL|^MISS\s+|^ive|LORITA\s+|Coulibaly\.|^deposit|LOTTERY\s+|Lottery\s+|WINNING\s+|Coulibaly\.|^deposit)/
```

It matches scams involving ‘lottery,’ ‘winnings,’ and ‘deposit’. In addition to the ‘lottery’ matches, it adds genes matching the names of alleged widows, such as ‘Lorita Manuel,’ and ‘Mary Coulibaly.

The 10th (final) generation

- This regex scored 50.58



```
# perl -e 'while(<>){print if (/(?:YOUR\s+[A-Z]+\s+[A-Z]+|^MISS\s+|^ive|MANUEL|L
OTTERY\s+|Lottery\s+|WINNING\s+|LORITA\s+|Coulibaly\.|^deposit)/);}' < 419.txt
                 YAHOO LOTTERY WINNING NOTIFICATION
  I AM EXPECTING YOUR IMMEDIATE REPLY.
  I AM HAPPY TO REQUEST FOR YOUR ASSISTANCE BECAUSE I BELIVE THAT YOU ARE NOT GO
ING TO BETRAY THE TRUST WHICH I AM GOING TO LAY ON YOU.
  I WANT YOU TO ASSIST ME TO TRANSFER THIS MONEY TO YOUR COUNTRY FOR INVESTMENT
PURPOSES,SO I CAN FUTHER MY EDUCATION AND LIVE A NEW LIFE.
  MY NAME IS MISS LORITA MANUEL,22 YEARS OLD AND THE ONLY
1. YOUR FULL NAME
2.YOUR HOME ADDRESS.
3.YOUR CURRENT HOME TELEPHONE NUMBER.
4.YOUR CURRENT OFFICE TELEPHONE.
AND AT THE SAME  TIME  ENSURING THAT YOUR BANK ACCOUNT IS STILL  OPEN TO
BASED ON THAT YOU HAVE BEEN ADVISED NOT TO RESPOND TO ANY MAIL OR FAX CONCERNING
 YOUR FUND FROM ANY BODY WHO CLAIM TO BE OR YOUR AGENT AS YOUR PAYMENT  IS NOW R
EADY TO BE  TRANSFERED IMMEDAITELY  MY OFFICE  CONFIRM RECEIPT OF THE FOLLOWINGS
.
HIS DIRECT EMAIL ADDRESS BELLOW AND RECONFIRM TO HIM  YOUR CONTRACT NUMBER
I AWAIT YOUR URGENT REPLY.
INTERNATIONAL LOTTERY PROMO.
International Lottery Program. .
MISS LORITA MANUEL
MISS LORITA MANUEL.
Mrs.Mary Coulibaly.
OF CLAIM TO MY OFFICE ON YOUR BEHALF FOR THE RELEASE OF YOUR PAYMENT OF TEN
OSA CLAIMS PROCESSING LOTTERY AGENT
RECEIVE  YOUR APPROVED  CONTRACT  FUNDS
RECEIVING BANKING INFORMATION AND YOUR RESIDENCIAL ADDRESS AND YOUR DIRECT
SHELL PETROLEUM INTERNATIONAL LOTTERY PROMO NIGERIA
THE YAHOO LOTTERY INTERNATIONAL. INC
WHICH IS YOUR SUBJECT PAYMENT CODE AND APPROVED CONTRACT AMOUNT AND YOUR
WINNING NOTIFICATION LETTER
YAHOO LOTTERY INTL INC
YOUR ADVICE TO CONTACT THE  DIRECTOR FORIEGN OPERATION,DR.IGNATIUS IMALLA ON
Yahoo Lottery Prize must be claimed no later than 15 days from date of Draw
deposit
deposited by my late father.
deposited in CORPORATE SECURITIES CO, a
for this year 2007 Lottery promotion which is organized by
ive
won the Lottery in the 1st category, in four parts. You have been
#
```

In the final generation, a new chromosome takes the lead, scoring over 50:

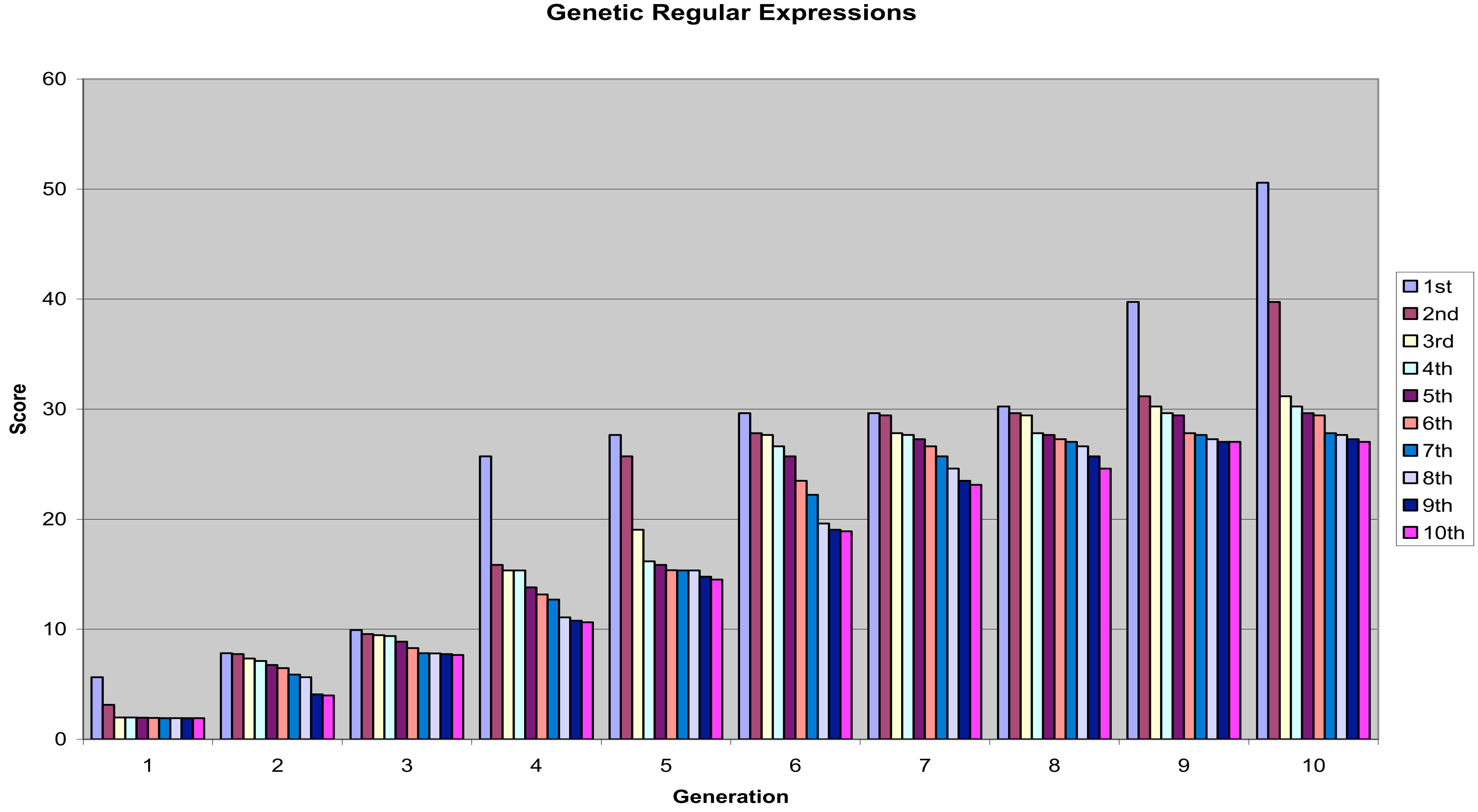
```
/(?:YOUR\s+[A-Z]+\s+[A-Z]+|^MISS\s+|^ive|MANUEL|LOTTERY\s+|Lottery\s+|WINNING\s+|LORITA\s+|Coulibaly\.|^deposit)/
```

It combines the genes from generation 9's winner (which scored 39.75):

```
/(?:Coulibaly\.|LOTTERY\s+WINNING\s+|^deposit|the\s+finance|YOUR\s+[A-Z]+\s+[A-Z]+)/
```

Generation 10's winner added the gene 'MANUEL' (matching lines containing 'MISS LORITA MANUEL', and the gene 'Lottery\s+' (matching any line containing 'Lottery,' followed by whitespace.

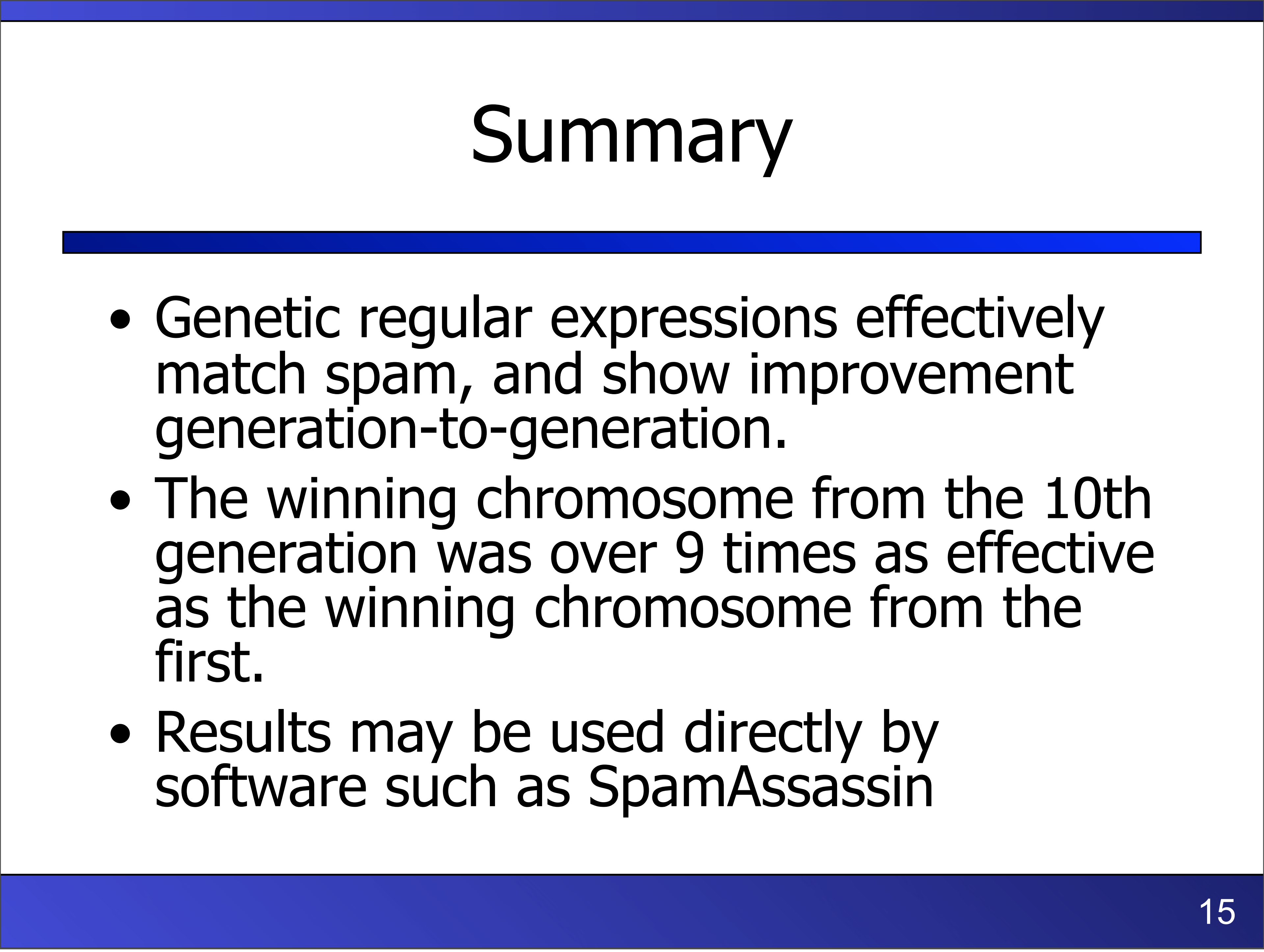
Scores Over 10 Generations



This graph shows the top 10 scores over 10 generations.

Note that overall performance increased with each generation.

During generations 6, 7, and 8 the top-scoring chromosome did not improve much, but others in the top 10 did. Then there is a jump in the top performer for rounds 9 and 10.



Summary

- Genetic regular expressions effectively match spam, and show improvement generation-to-generation.
- The winning chromosome from the 10th generation was over 9 times as effective as the winning chromosome from the first.
- Results may be used directly by software such as SpamAssassin

Genetic regular expressions effectively match spam, and show improvement generation-to-generation. The winning chromosome from the 10th generation was over 9 times as effective as the winning chromosome from the first.

Genetic Regular expressions exhibit the following characteristics, as described by Hiu Wong: ‘The basic techniques of the GAs are designed to simulate processes in natural systems necessary for evolution, specially those follow the principles first laid down by Charles Darwin of "survival of the fittest."’[1]

Genetic regular expressions leverage the Genetic Algorithm concepts of fitness, crossover, and mutation to evolve chromosomes across generations to find a superior solution.

This SpamAssassin rule is based on the winning rule:

```
body GENETIC_REGEX_1            /(?:YOUR\s+[A-Z]+\s+[A-Z]+|^MISS\s+|^ive|MANUEL|LOTTERY\s+|Lottery\s+|WINNING
\s+|LORITA\s+|Coulibaly\.|^deposit)/
score GENETIC_REGEX_1           1.0
describe GENETIC_REGEX_1        'Advanced Fee' spam rule written by genregex.pl
```

[1] Wong, Hiu. Introduction to Genetic Algorithms. Surveys and Presentations in Information Systems Engineering (SURPRISE) Journal 96 Volume 4. URL: http://www-dse.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/hmw/article1.html