Session 4: Substitution ciphers and frequency analysis

To develop knowledge and use of ciphers to include substitution ciphers and frequency analysis

Let’s learn

In this session, you will discuss how you might make a Caesar **cipher** more secure by using a mixed-up alphabet. You will look at examples of substitution ciphers and learn how to use frequency analysis and common words to crack **codes**.

Let’s discuss

The security of the Caesar cipher rests on keeping the ‘key’ (how far along the alphabet the letters of the **message** are shifted) secret. This is a weak **encryption** system as it is easy to test all possibilities.

1. How do you think it could be made more secure?

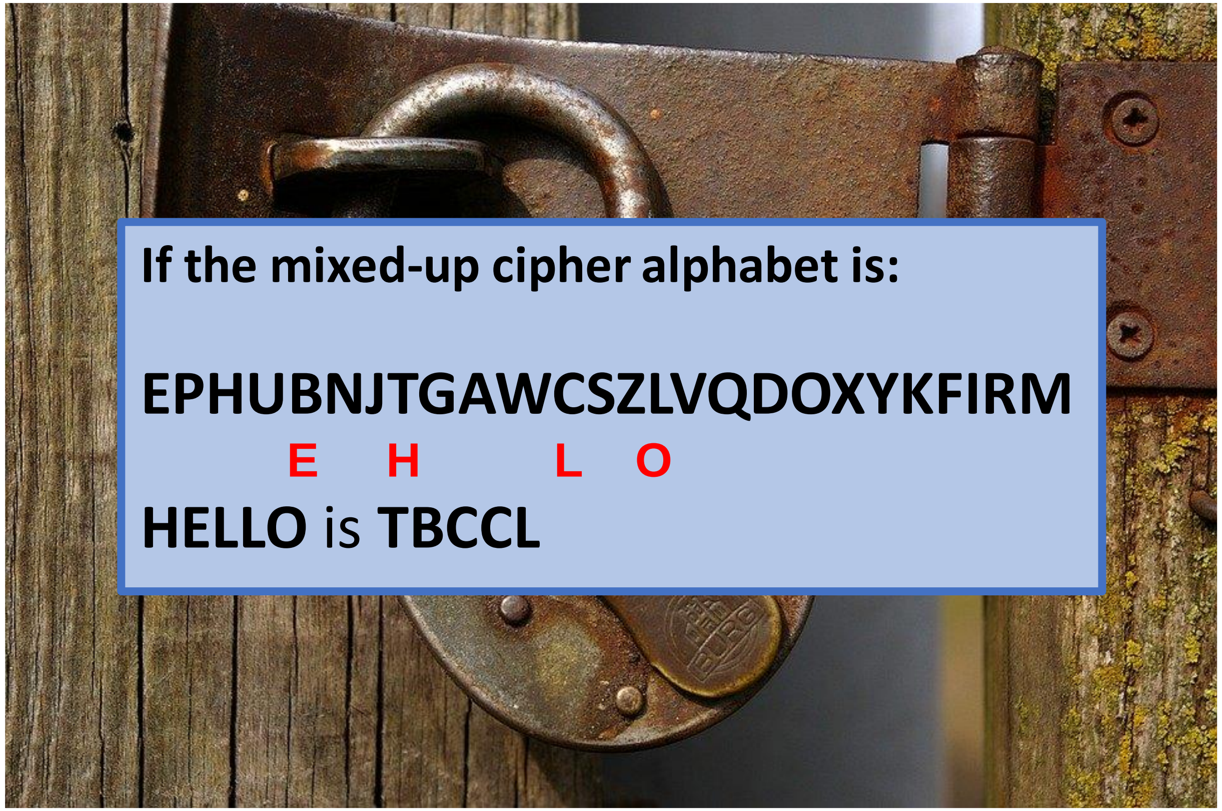
Let’s learn

One idea might be that there is no need for the cipher alphabet to be kept in alphabetical order.

You can **encode** a message by substituting letters in the plain text alphabet for the letter in the corresponding position of a mixed-up alphabet. This is called a substitution cipher.

Let’s discuss

1. What other words can you make using the cipher alphabet on the right?



If the mixed-up cipher alphabet is:

EPHUBNJTGAWCSZLVQDOXYKFIRM

E H L O

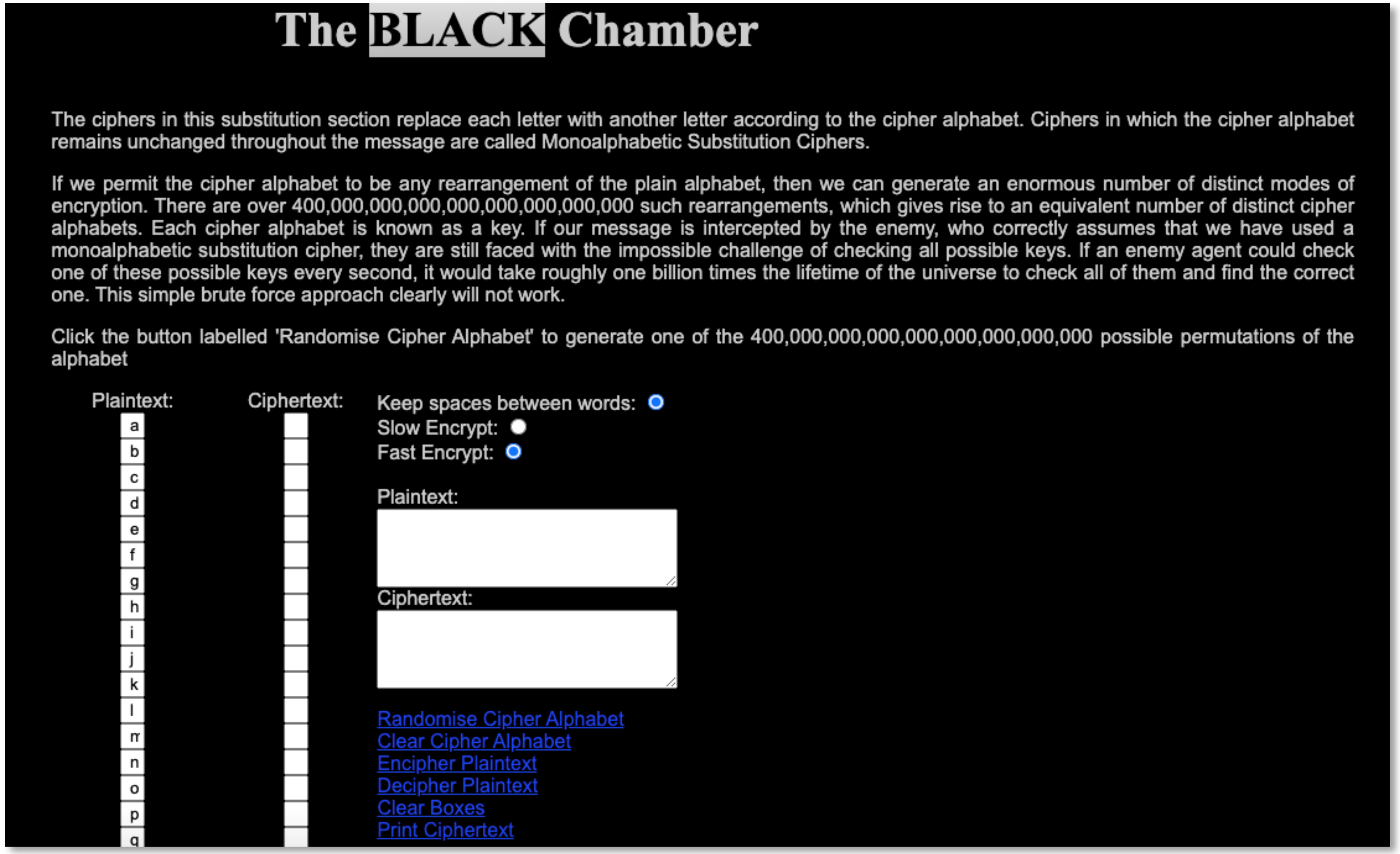
HELLO is **TBCCL**

Do you know what **encode** means?
Click on this box to see the definition.

Let’s do

There is a tool **on The Black Chamber** website. It creates substitution ciphers by mixing up the alphabet.

Have a go and explore creating and using substitution ciphers using this online tool.



The BLACK Chamber

The ciphers in this substitution section replace each letter with another letter according to the cipher alphabet. Ciphers in which the cipher alphabet remains unchanged throughout the message are called Monoalphabetic Substitution Ciphers.

If we permit the cipher alphabet to be any rearrangement of the plain alphabet, then we can generate an enormous number of distinct modes of encryption. There are over 400,000,000,000,000,000,000,000,000 such rearrangements, which gives rise to an equivalent number of distinct cipher alphabets. Each cipher alphabet is known as a key. If our message is intercepted by the enemy, who correctly assumes that we have used a monoalphabetic substitution cipher, they are still faced with the impossible challenge of checking all possible keys. If an enemy agent could check one of these possible keys every second, it would take roughly one billion times the lifetime of the universe to check all of them and find the correct one. This simple brute force approach clearly will not work.

Click the button labelled 'Randomise Cipher Alphabet' to generate one of the 400,000,000,000,000,000,000,000,000 possible permutations of the alphabet

Plaintext:	Ciphertext:
a	
b	
c	
d	
e	
f	
g	
h	
i	
j	
k	
l	
m	
n	
o	
p	
q	

Keep spaces between words: ●
Slow Encrypt: ○
Fast Encrypt: ●

Plaintext:

Ciphertext:

Tap the image to open the substitution tool on The Black Chamber website!

Let’s discuss

Think about the security of the substitution cipher method.

Would it still be easy to test all possible keys?

Click to reveal the answer!

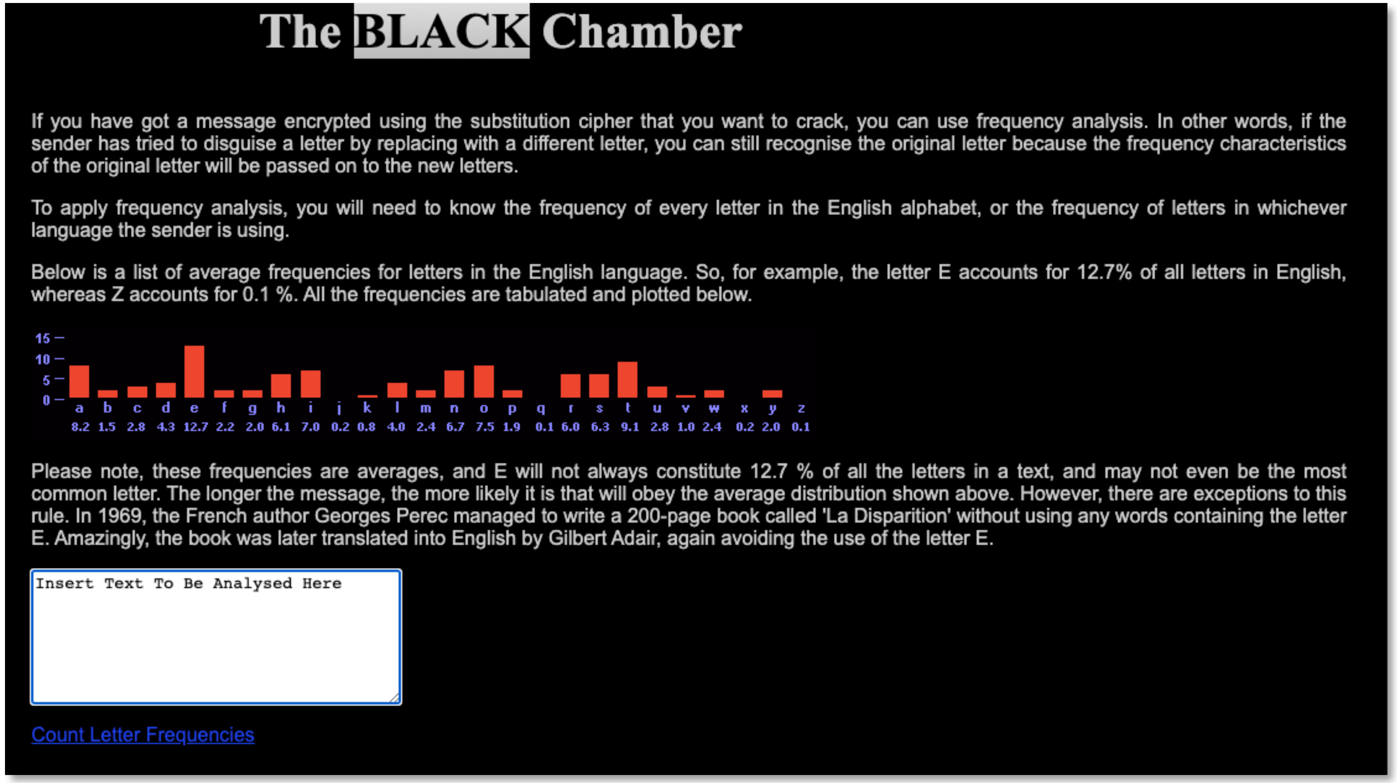
It is not easy to test all possible keys because there are endless possibilities. However, the system can still be quite easily broken because some letters and words occur in English more than others (letters e/t/a, words and/of/the).

Letter Count

There is a frequency tool on the code cracking website The Black Chamber.

This helps us to identify the possible letters as some as more popular than others.

See if you can crack the code.



The BLACK Chamber

If you have got a message encrypted using the substitution cipher that you want to crack, you can use frequency analysis. In other words, if the sender has tried to disguise a letter by replacing with a different letter, you can still recognise the original letter because the frequency characteristics of the original letter will be passed on to the new letters.

To apply frequency analysis, you will need to know the frequency of every letter in the English alphabet, or the frequency of letters in whichever language the sender is using.

Below is a list of average frequencies for letters in the English language. So, for example, the letter E accounts for 12.7% of all letters in English, whereas Z accounts for 0.1 %. All the frequencies are tabulated and plotted below.

Please note, these frequencies are averages, and E will not always constitute 12.7 % of all the letters in a text, and may not even be the most common letter. The longer the message, the more likely it is that will obey the average distribution shown above. However, there are exceptions to this rule. In 1969, the French author Georges Perec managed to write a 200-page book called 'La Disparition' without using any words containing the letter E. Amazingly, the book was later translated into English by Gilbert Adair, again avoiding the use of the letter E.

Click the image to open the frequency tool on The Black Chamber website!