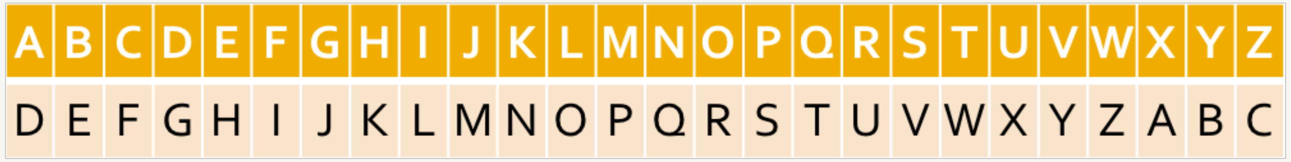
Classic cipher: substitution cipher

- Each letter is uniquely replaced by another
- Caesar cipher is an example of a substitution cipher utilised by Julius Caesar:
    - Here, each letter in the plaintext is shifted three letters on right
    - When it reaches the end, it is wrapped back at the beginning
    - The decryption would require a three left shift
- An example of a Caesar shift is given below



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C

Classic cipher: substitution cipher

- A substitution cipher can be generalised in the following way:
  - $E_k(x) = (x + k)\ mod\ 26$
  - $D_k(x) = (x - k)\ mod\ 26$
  - For Caesar cipher, $k = 3$
- k can be any value, however it is very easy to attack the system by brute forcing different values of k until a meaningful message is found

Classic cipher: substitution cipher

- We can generalise this cipher so that each letter can have an arbitrary substitution, so long as all the substitutions are unique
- This approach greatly increases the key space; hence, increasing the security of the cryptosystem
- For example, with English plaintexts, there are 26! possible substitution ciphers
  - $26! \approx 4.03 \times 10^{26}$ such ciphers!

Substitution cipher attack

- Even with this huge key space, a substitution cipher can be easily broken
- This is because letters in a natural language, like English, are not uniformly distributed
- Knowledge of letter frequencies, including pairs and triples can be used in cryptologic attacks against substitution ciphers
- For example, in English text, the letter “E” occurring just over 12% of the time, and “T” occurring less than 10% of the time
- The most frequently occurring character in a ciphertext created from English substitution cipher probably corresponds to the letter E and so on

Substitution cipher attack

a:	8.05%	b:	1.67%	c:	2.23%	d:	5.10%
e:	12.22%	f:	2.14%	g:	2.30%	h:	6.62%
i:	6.28%	j:	0.19%	k:	0.95%	l:	4.08%
m:	2.33%	n:	6.95%	o:	7.63%	p:	1.66%
q:	0.06%	r:	5.29%	s:	6.02%	t:	9.67%
u:	2.92%	v:	0.82%	w:	2.60%	x:	0.11%
y:	2.04%	z:	0.06%				

Table 1: Letter frequencies in the book *The Adventures of Tom Sawyer*, by Mark Twain.