Rabin Fingerprinting The Hash Cannon

...

Arnaud Bellizzi
@Oodrive
Architect - Backup & Archive

Data !

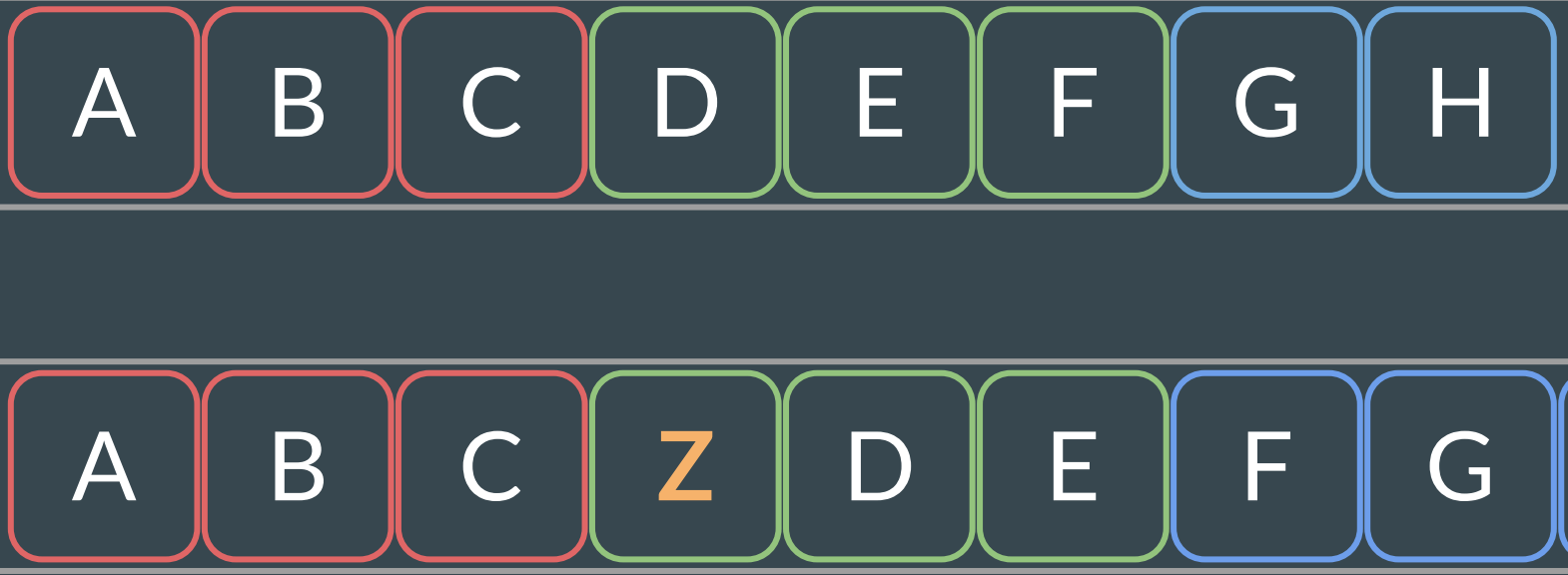
Chunks !

Chunks !

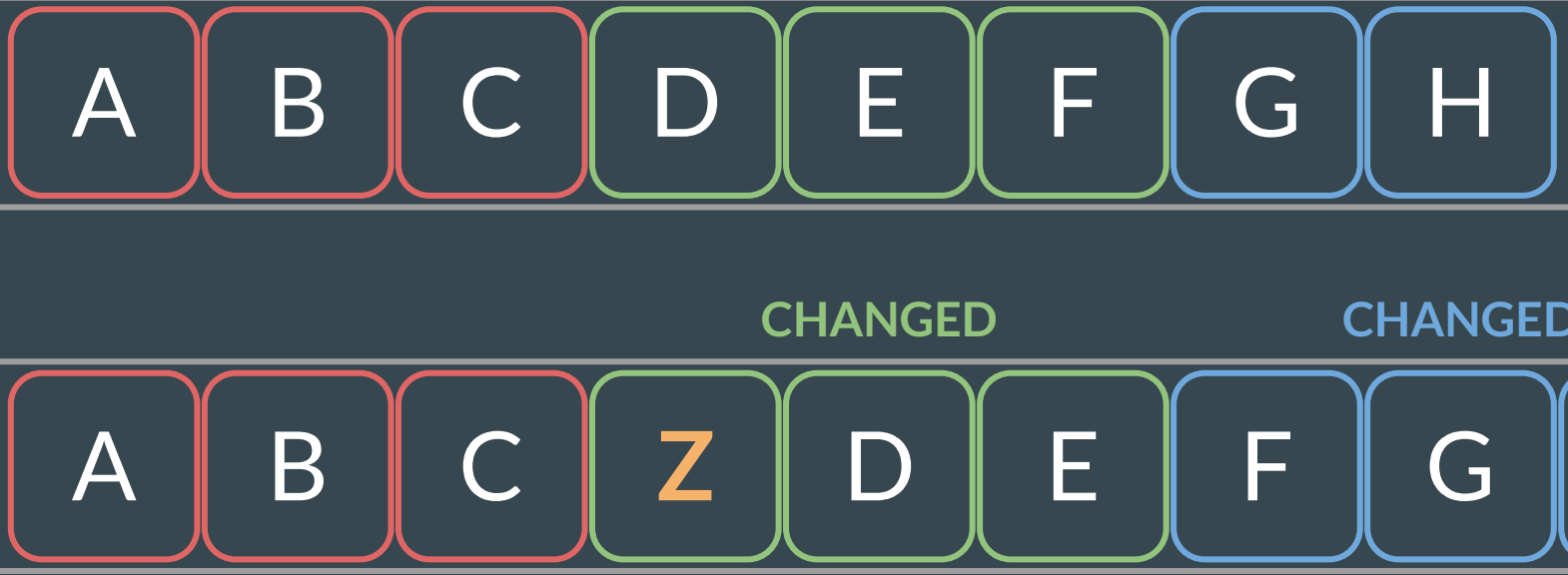
- Differential transfer
- Deduplication

Byte Shift

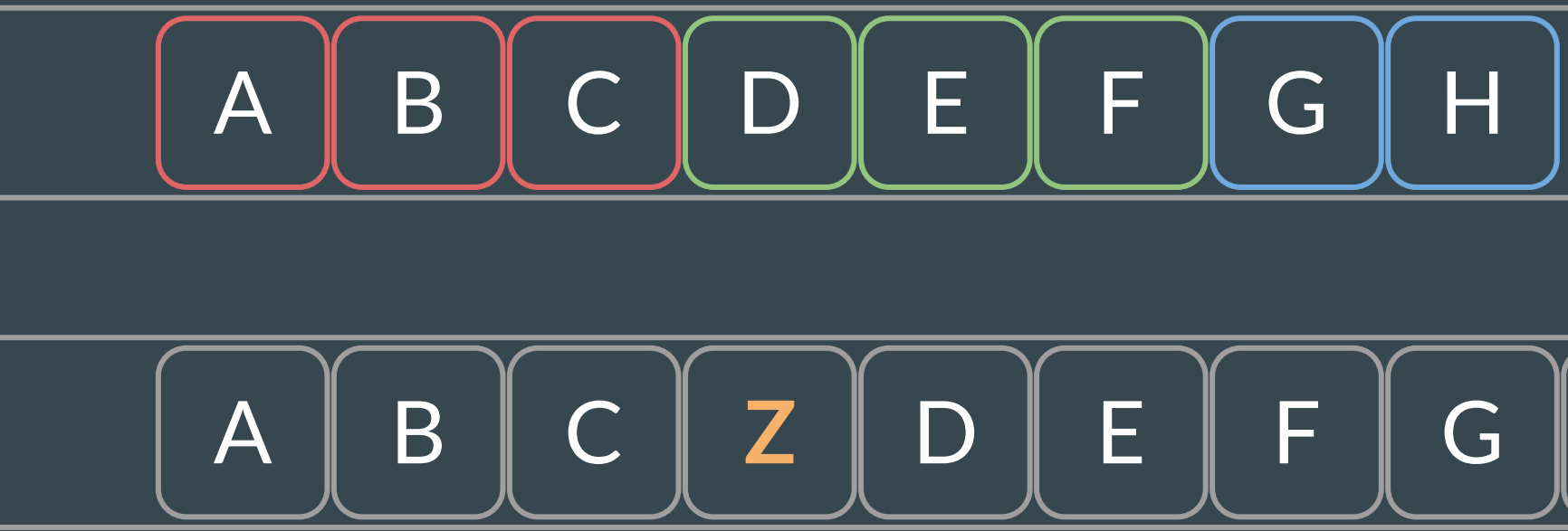
Byte Shift



Byte Shift

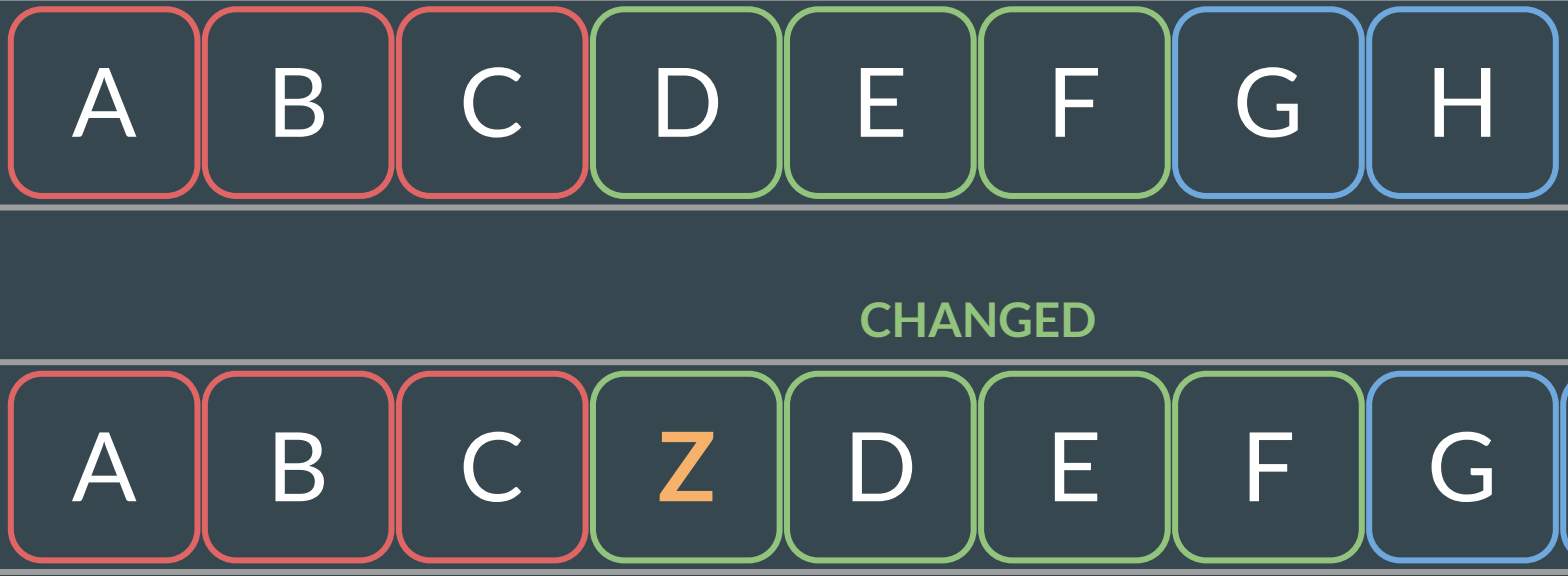


What would we like ?



What would we like ?

What would we like ?



What would we like ?

- Variable chunk size
- Content based chunking

Enter Rabin !

Michael O. Rabin
1931

- Turing Award Winner 1976
- Great smile
- Loves Math, Cryptography, Hashes

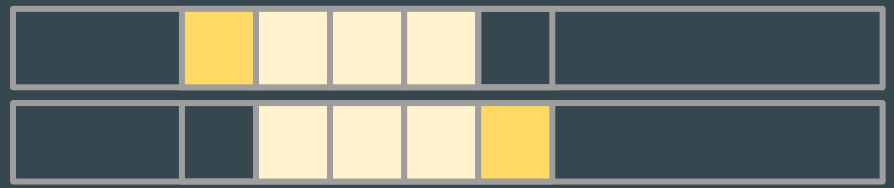
Fingerprinting

- **Hashing** function (arbitrary lengths, statistically uniform)

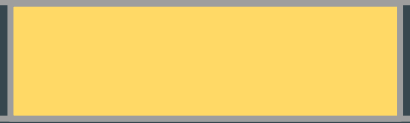
Fingerprinting

- **Hashing** function (arbitrary lengths, statistically uniform)

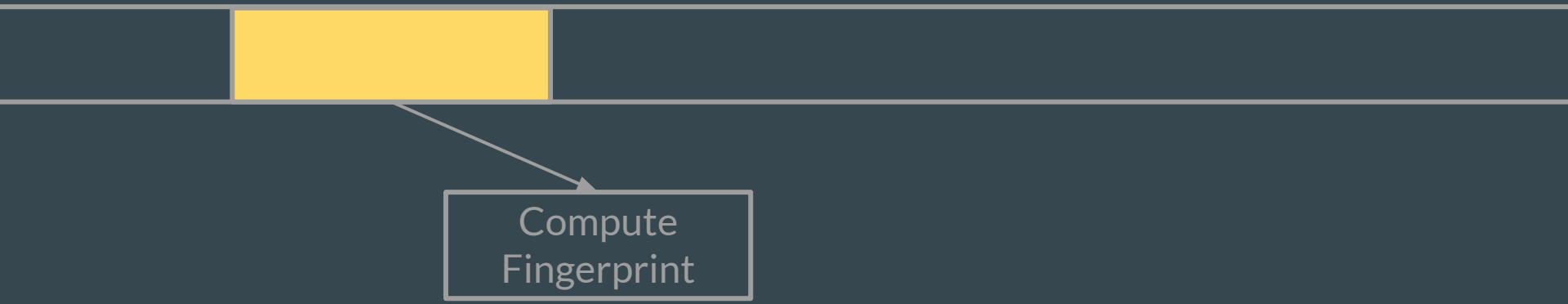
- **Rolling** hash



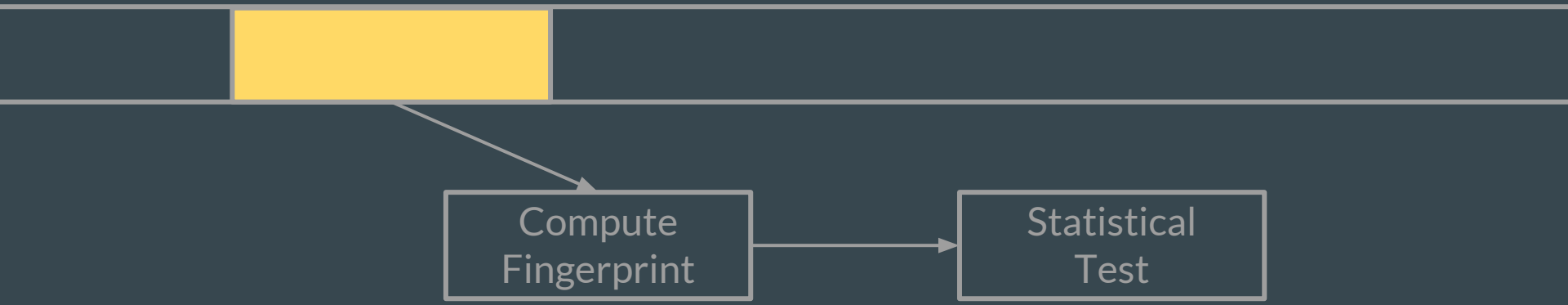
The Hash Cannon !



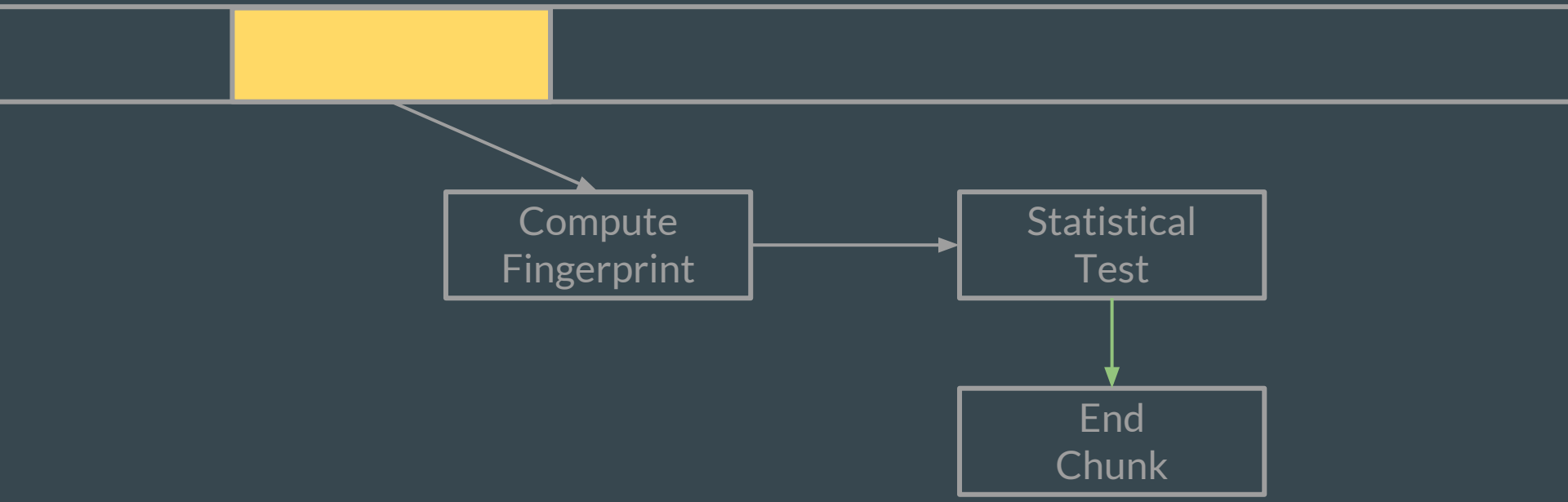
The Hash Cannon !



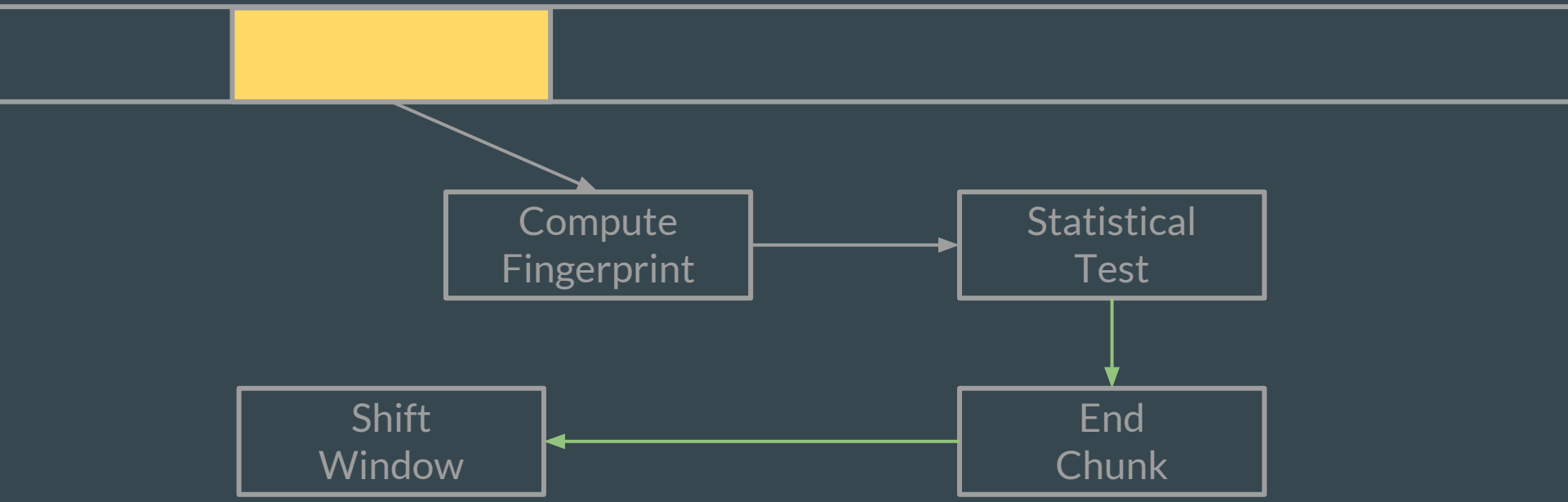
The Hash Cannon !



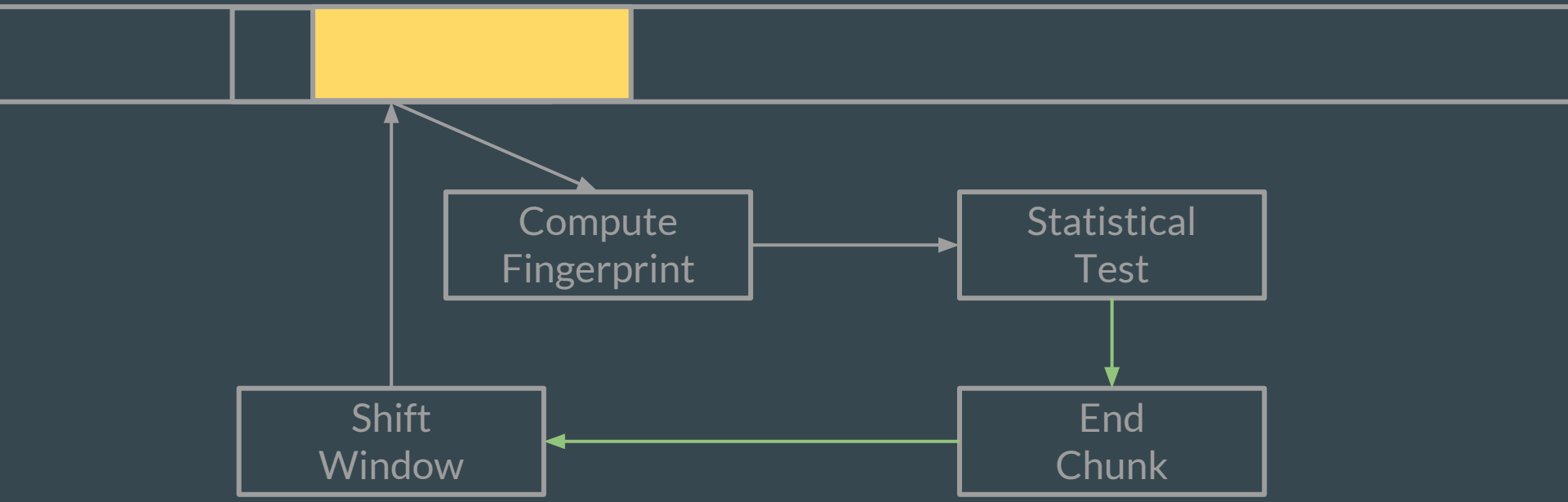
The Hash Cannon !



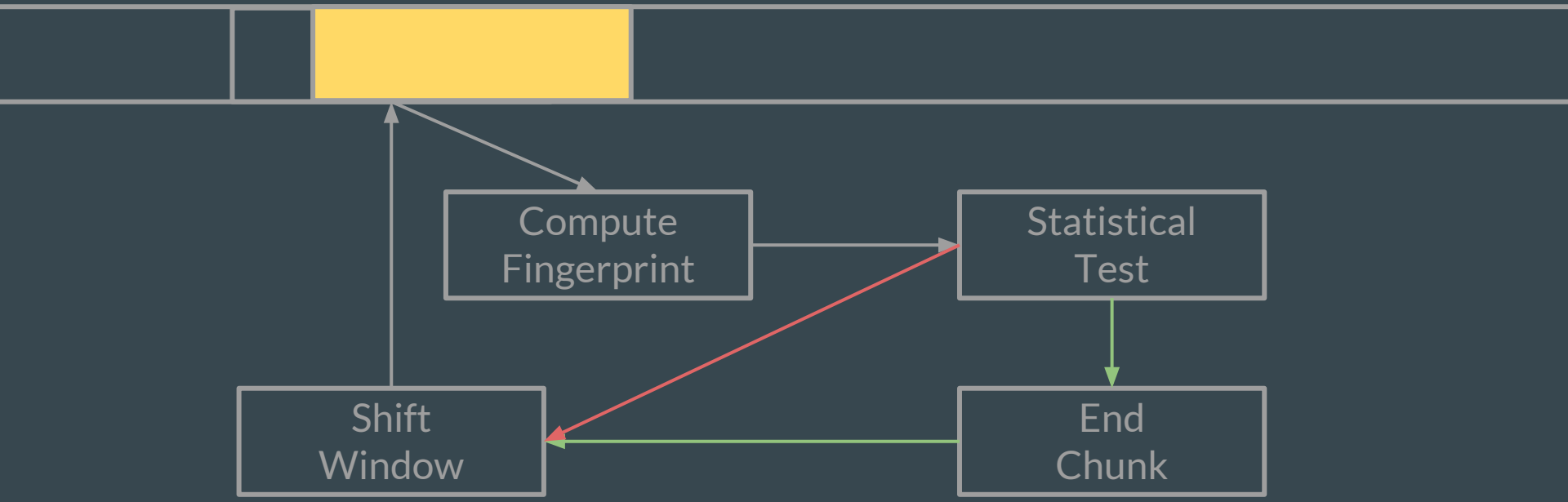
The Hash Cannon !



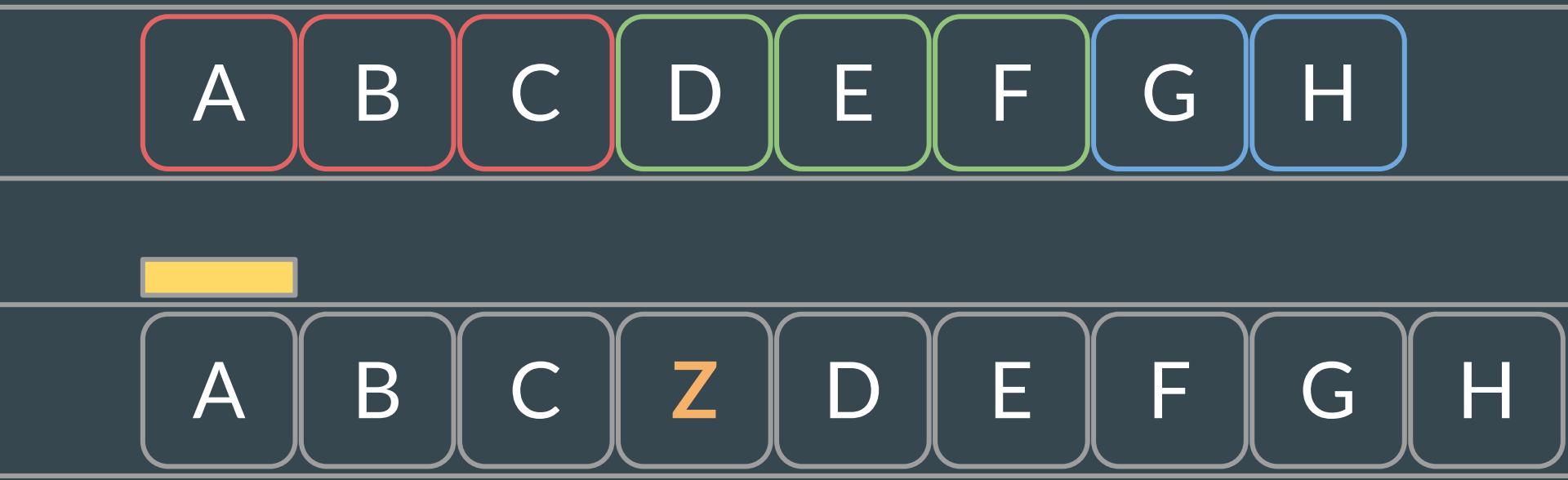
The Hash Cannon !



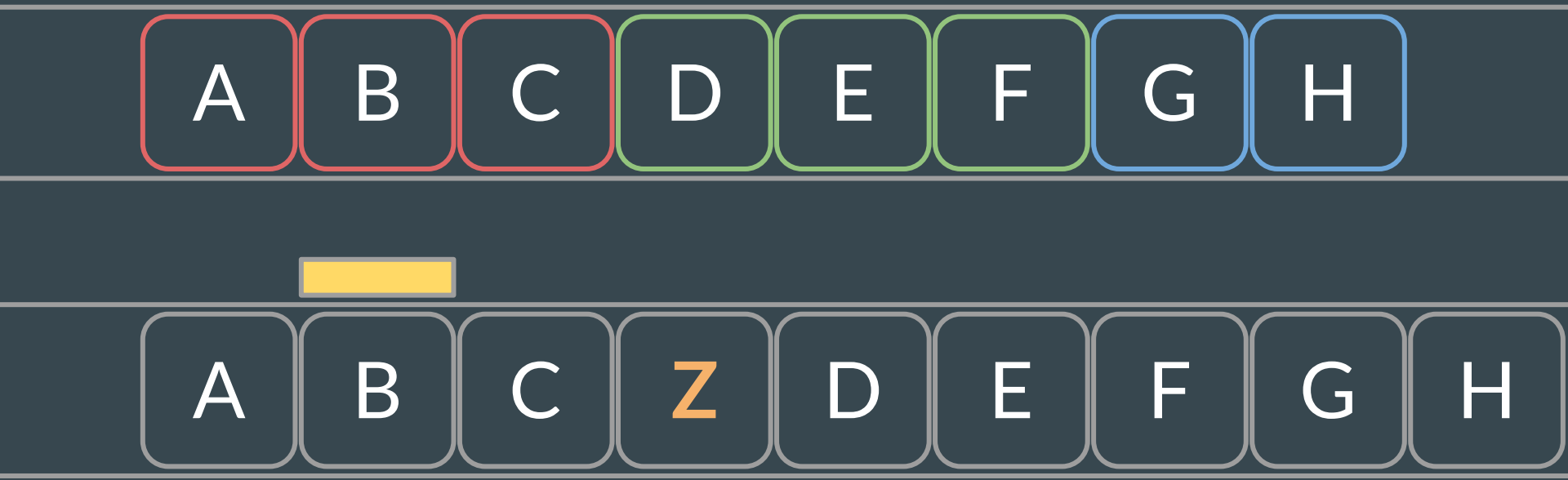
The Hash Cannon !



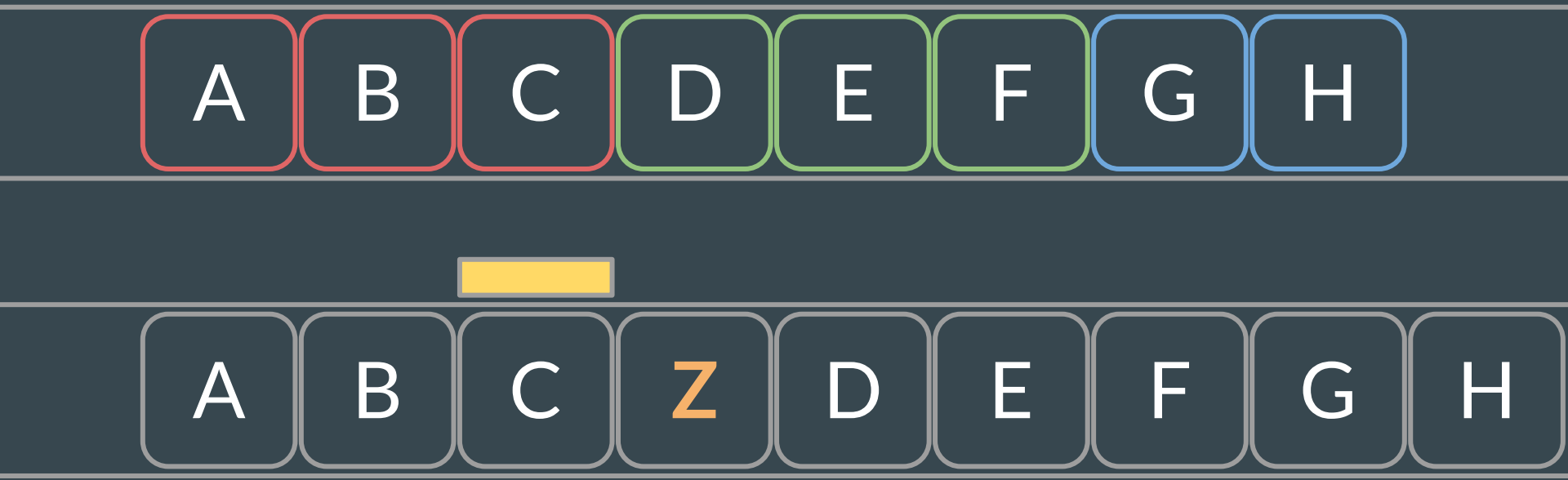
Byte Shift (again)



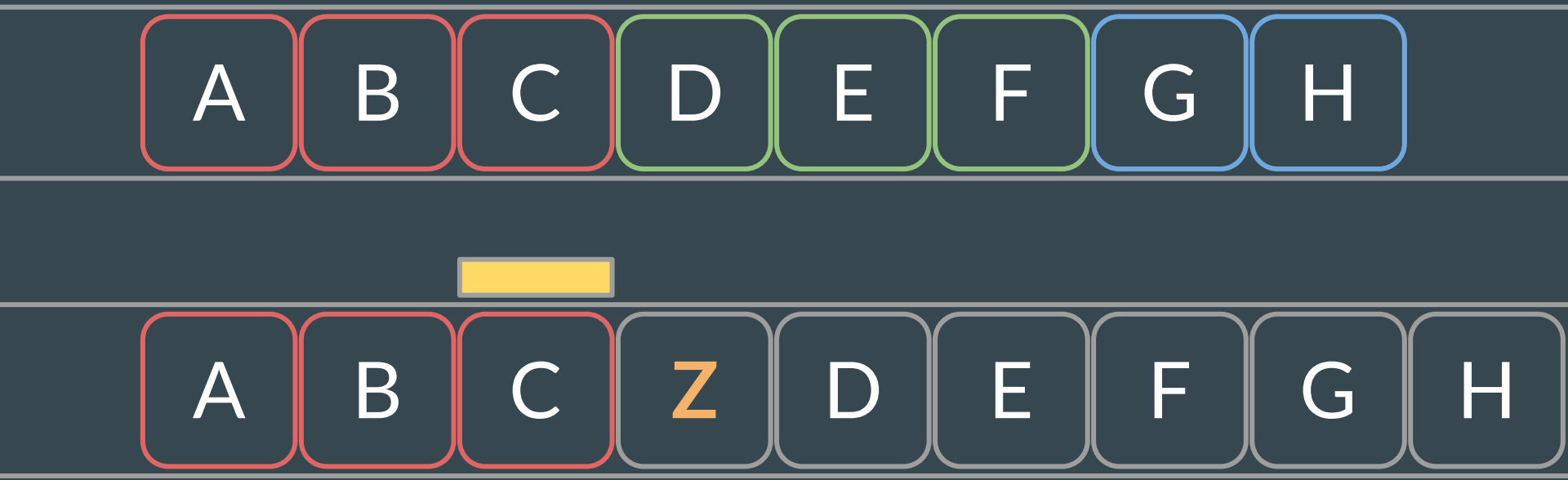
Byte Shift (again)



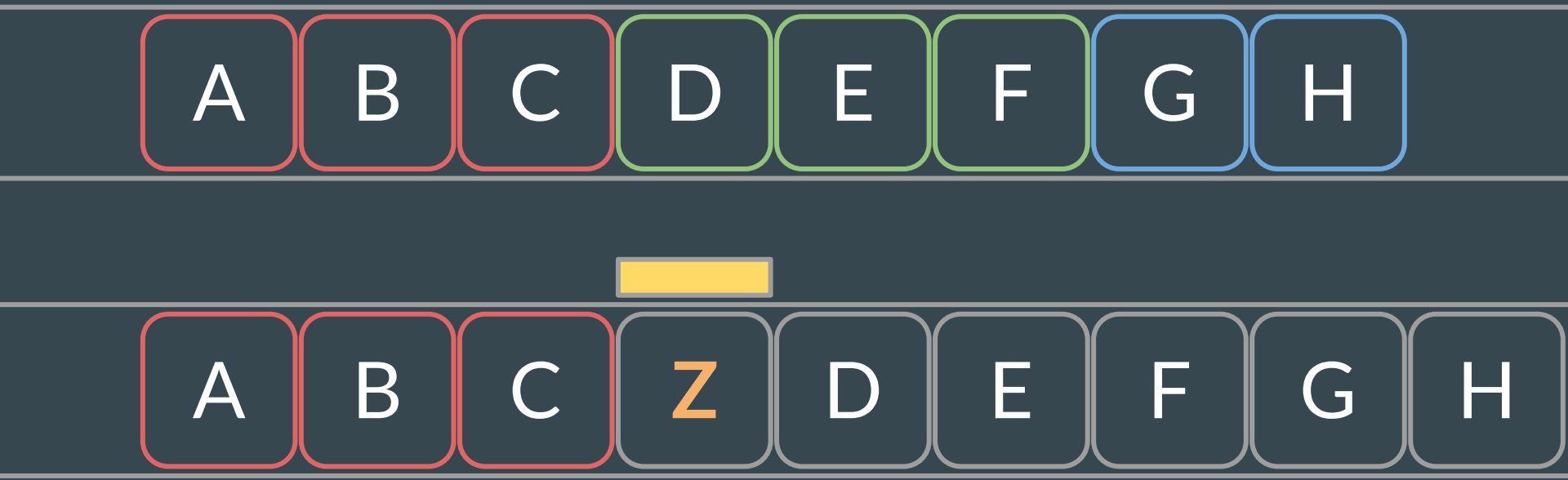
Byte Shift (again)



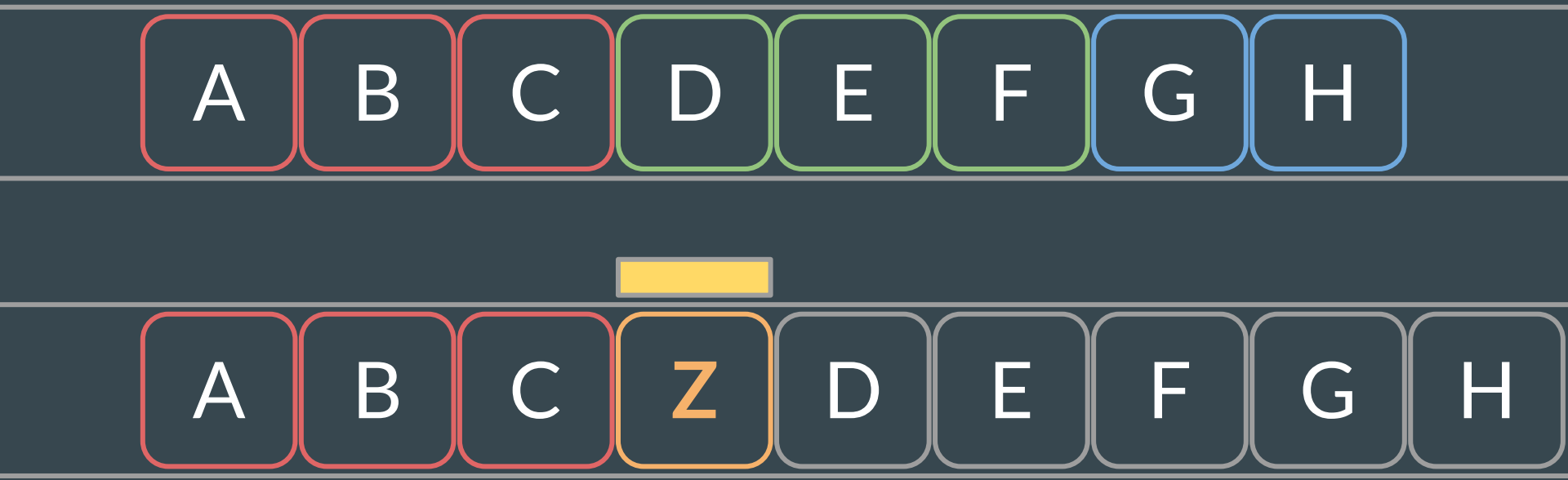
Byte Shift (again)



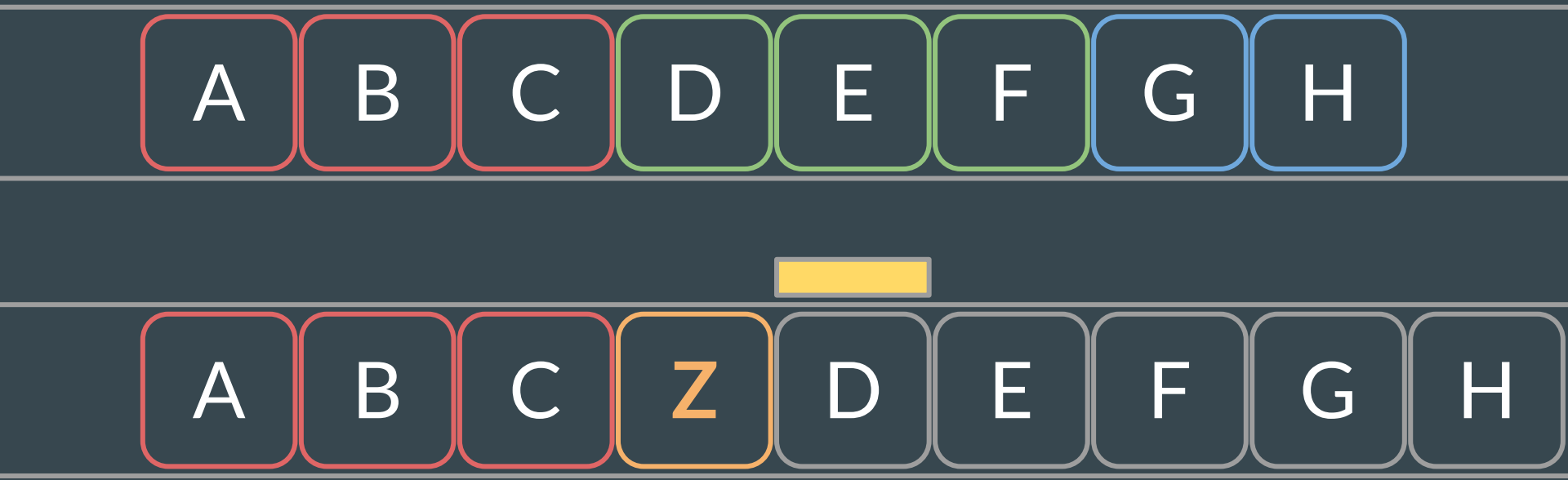
Byte Shift (worst case)



Byte Shift (worst case)

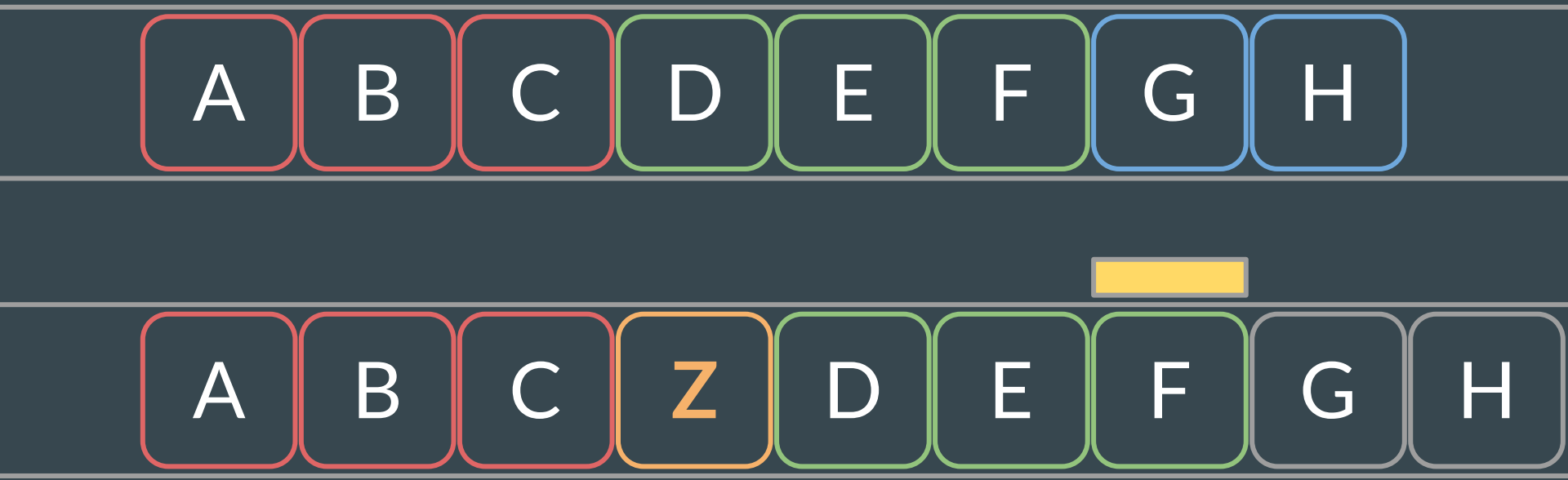


Byte Shift (worst case)



Byte Shift (worst case)

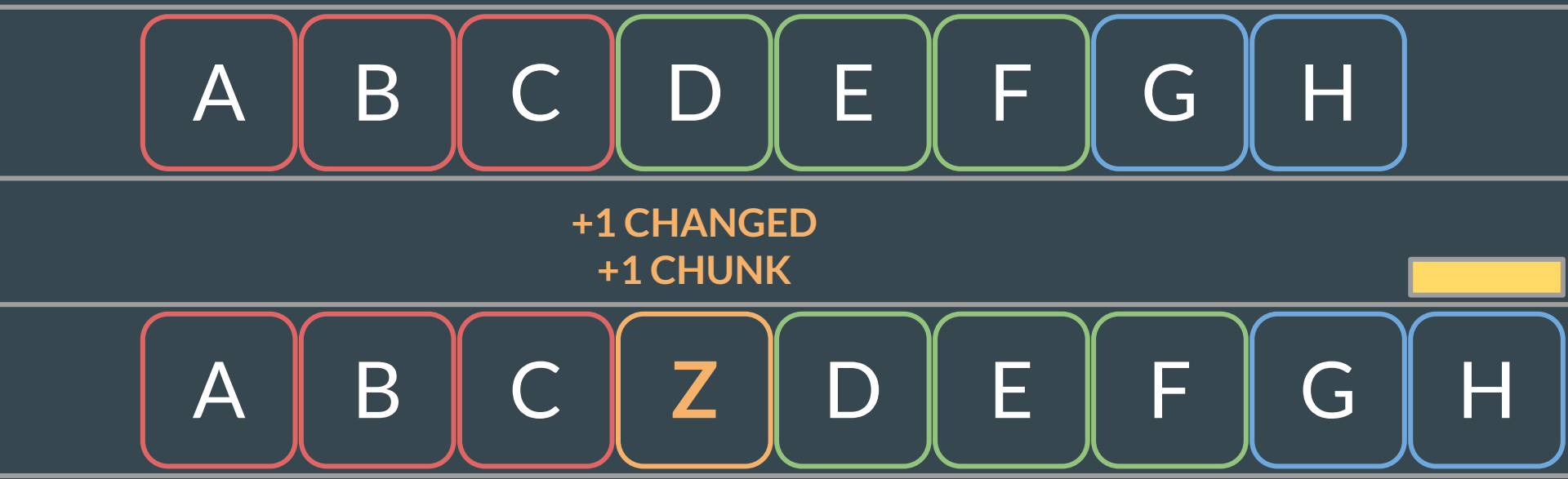
Byte Shift (worst case)



Byte Shift (worst case)

Byte Shift (worst case)

Byte Shift (worst case)



Profit !

~20% less chunks for same data

3-15% less data on the wire

Chunks are now **resistant** to Byte Shift

Thank you