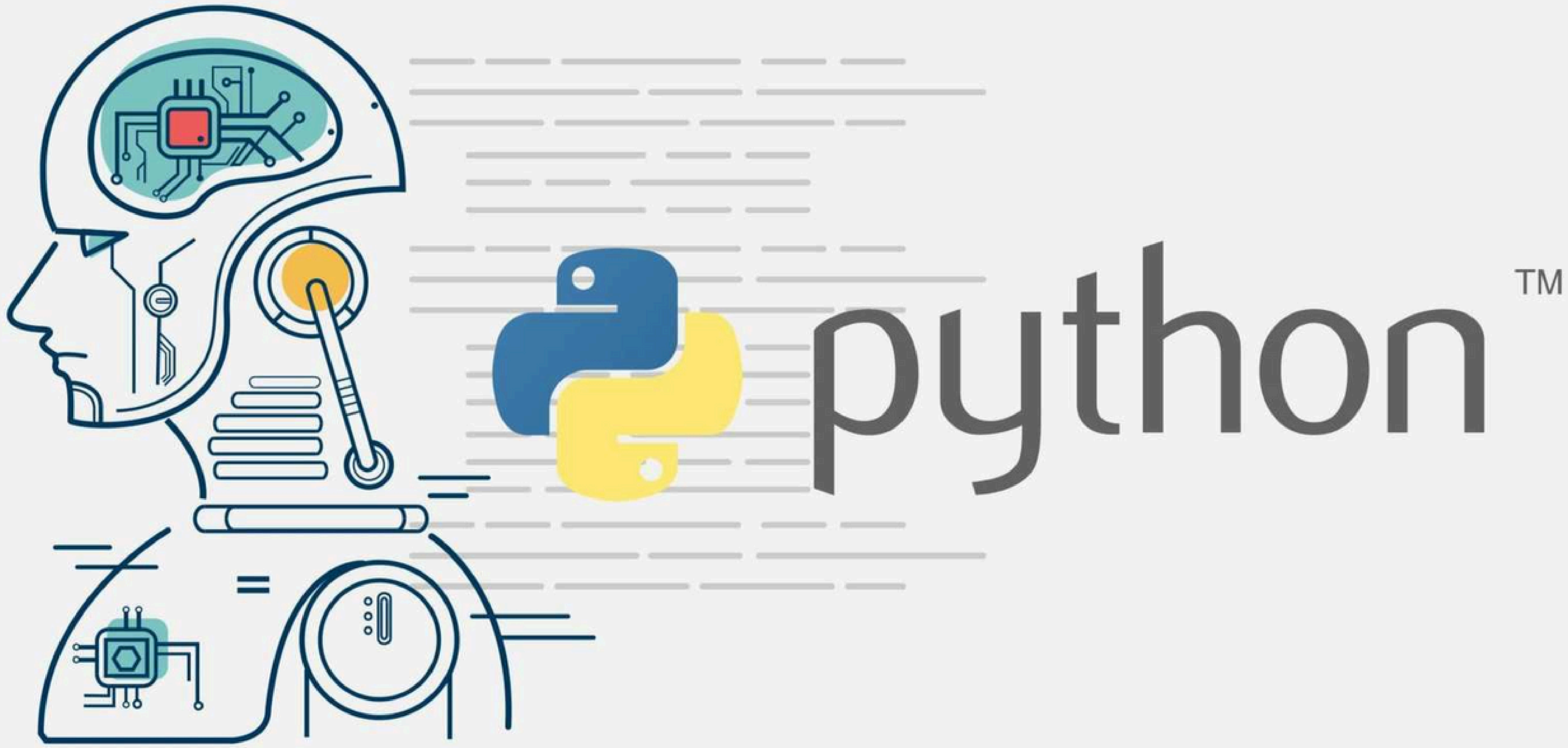
# Coding - Python

## Regular Expressions



Jowita Drozdowicz

# Why text data matters?

- 80%-90% of all data is unstructured
- Text drives insights from people, not just numbers
- Regex & NLP make this data usable
- Clean text = smarter models and decisions

# Messy examples

Hey there! My name’s Anna, I’m from the UK 🇬🇧 and I’ve just bought 3 iPhones for 2,499.99 USD!!!

Can u believe it?? 😂 Email me at anna_92@example.com or contact@tech-review.co.uk.

BTW, check out my blog @ https://techstuff.blog or follow me on Twitter #TechLife #AI #Python3.

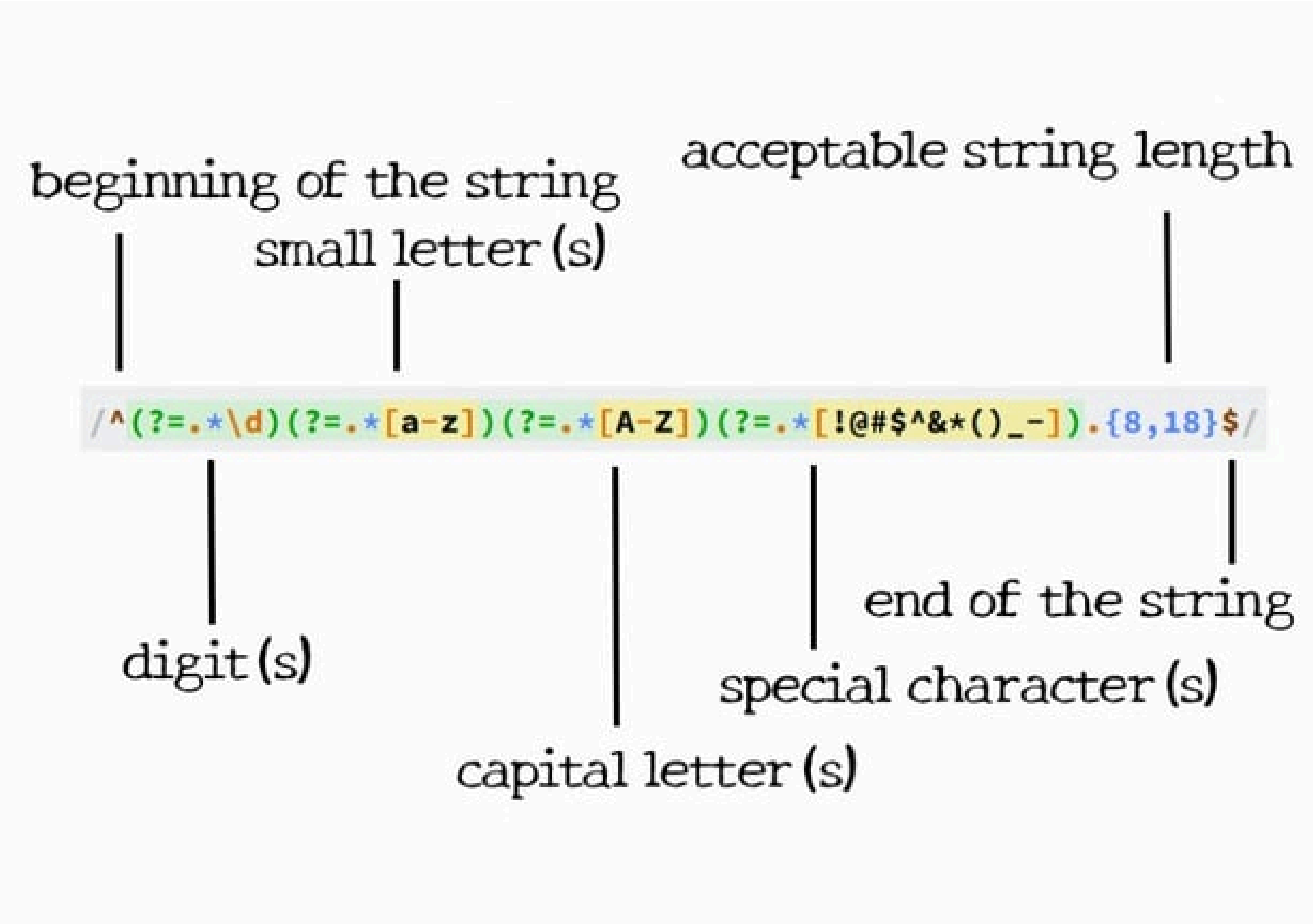
Order ID: #2025-00458 | Call me maybe? +44-20-7946-0958 📞

P.S. See you on 08/10/2025 😎

# Regular Expressions

.[RegEx]*

# What is a regex?



```
/^(?=.*\d)(?=.*[a-z])(?=.*[A-Z])(?=.*[!@#$^&*()_-]).{8,18}$/
```

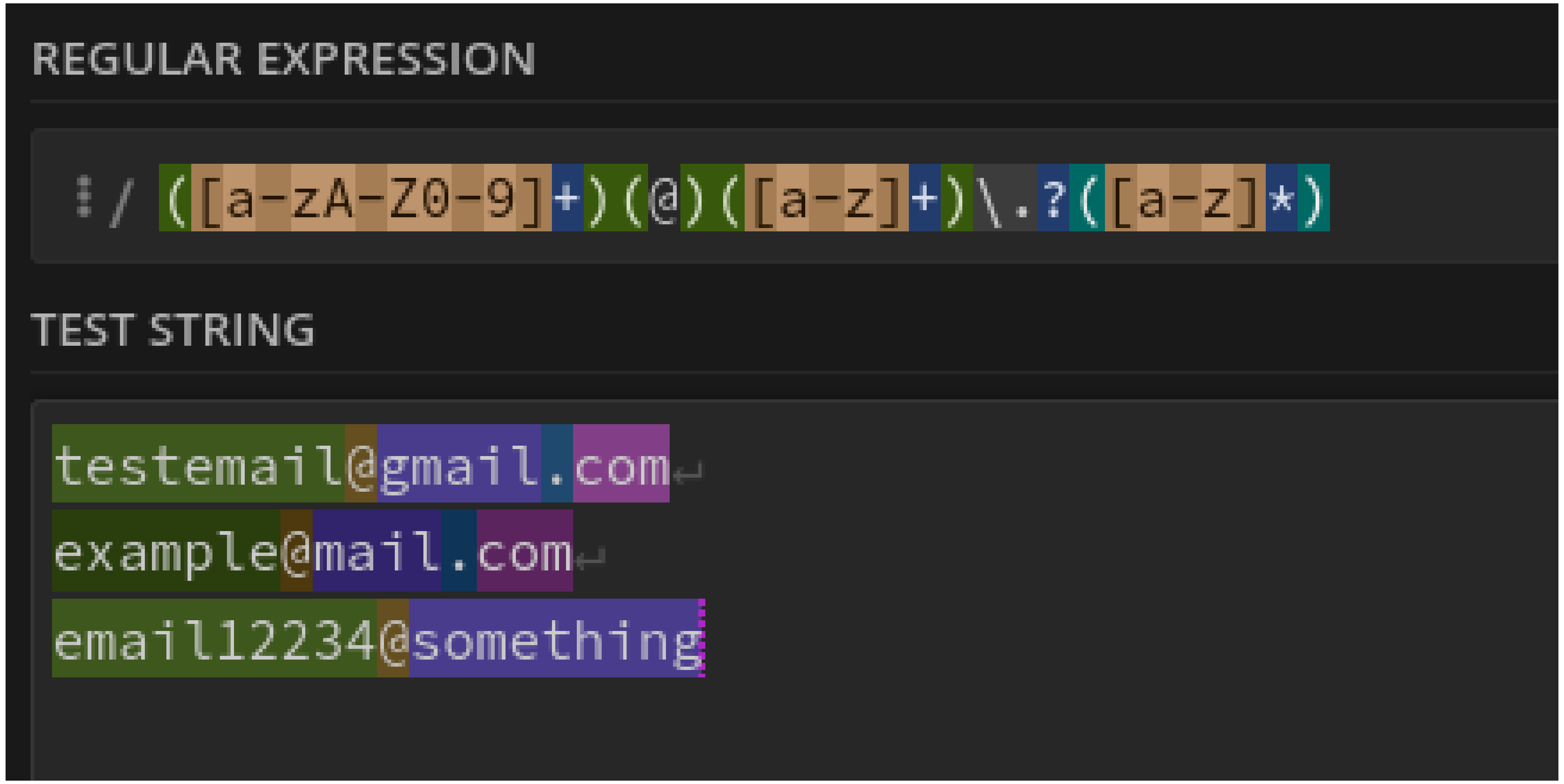
# Key syntax

- Character classes: [a-z], \d, \w
- Quantifiers: *, +, {m,n}
- Anchors: ^, $, \b
- Groups and alternation: ( ), |

Let’s try: regex101.com

# Use cases

- Find all emails, phone numbers, dates
- Replace multiple spaces
- Extract hashtags or mentions from tweets



```
/ ([a-zA-Z0-9]+)(@)([a-z]+)\.?([a-z]*)
```

```
testemail@gmail.com
example@mail.com
email12234@something
```

# Short quiz

Question: What does this regex match?

```
\+?\d{1,3}[-\s]?\d{2,4}[-\s]?\d{3}[-\s]?\d{3,4}
```

A. Email address

B. Hashtag

C. Telephone number with country code

D. Date

# Short quiz

Question: What does this regex extract from text?

```
#\w+
```

A. All capitalized words
B. Hashtags (e.g., #Python)
C. Words ending with punctuation
D. URLs

# Short quiz

Question: Which of the following strings would match this pattern?

```
\b\d{1,2}[/-]\d{1,2}[/-]\d{2,4}\b
```

A. 2025-10-08
B. 08/10/2025
C. +44-20-7946-0958
D. anna_92@example.com

# Short quiz

Question: What does this regex match?

```
(cat|dog)s?
```

A. Only the word cat
B. Only the word dog
C. Both cat and dog, with optional plural s
D. Any animal name

# Short quiz

Question: What does this regex match?

```
^[A-Z][a-z]+\s[A-Z][a-z]+$
```

A. A single word

B. A full name starting with capital letters

C. Any uppercase word

D. A sentence ending with a period

# NLP Preprocessing

# What is NLP Preprocessing?

- Raw text ≠ ready data!
- Example:
- “Cats are running faster than dogs!!! 🐱🐶 #speed”
- Has emojis, punctuation, casing, duplicates, etc.
- Preprocessing = turning messy text into analyzable form.