Regular Expression Fundamentals

Juliette Reinders Folmer

@jrf_nl | regexcheatsheets.com

Wildcards on Steroids

Some people, when confronted with a problem, think

"*I know, I'll use regular expressions.*"

Now they have two problems.

— Jamie Zawinski, August 1997

alt.religion.emacs

What Are Regular Expressions ?

And Why Are They Useful to You ?

Regular Expression:

A sequence of characters that define a search pattern.

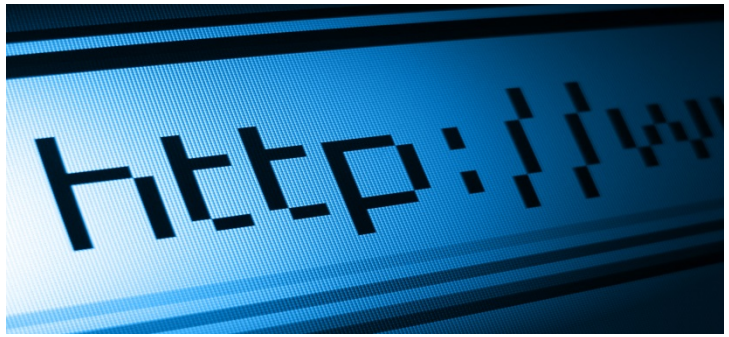
Pattern Recognition

Patterns


PIN
1 2 3
4 5 6
7 8 9
* 0 #


JPG
PNG
PSD
TIFF
BMP


SLAVE


http://


775-AB
SV 03-081
KPY-07
LJ 149-26
B43-76
3423
90-32 LJ
SXK-875
12612


EURO


RCA
AM
5:41

Patterns


User name
User
Password
Cancel
Next


VISA
MasterCard
AMERICAN EXPRESS


TSA PRE
INTL
BEIJING TO WASHINGTON
UA898
PEK-IAD
E29
5:40P
39H
BOARDING GROUP
5
CONFIRMATION: GT4PJ2

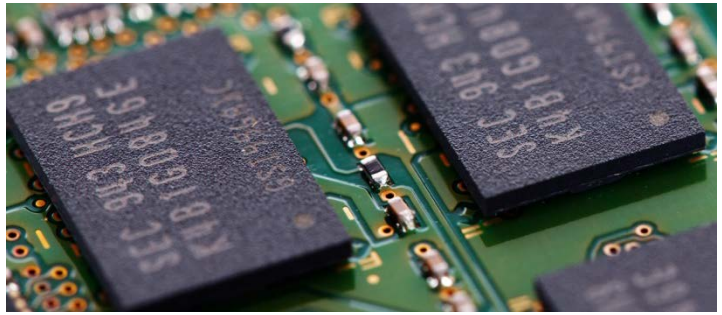

De Slegte
De Slegte Leiden
De last van de wereld
Totaal 1 artikel
17,50
EUR
Exclusief
17,50
BTW
0,00


01 GEN. 1999
trodat
printy-dater 4810
14 FEB. 2005
13 FEB. 2005
12 FEB. 2005
11 FEB. 2005
10 FEB. 2005
09 FEB. 2005
08 FEB. 2005

Patterns

:) :) (H) :(

ISBN 1 86197 271 7



Typical Uses for Regular Expressions

- Input validation
- Search (and replace)
- String parsing
- Data scraping
- Syntax highlighting
- Data mapping

Users of Regular Expressions

Developers

Working with strings

Data Professionals

Query data

(Sys-)Admins

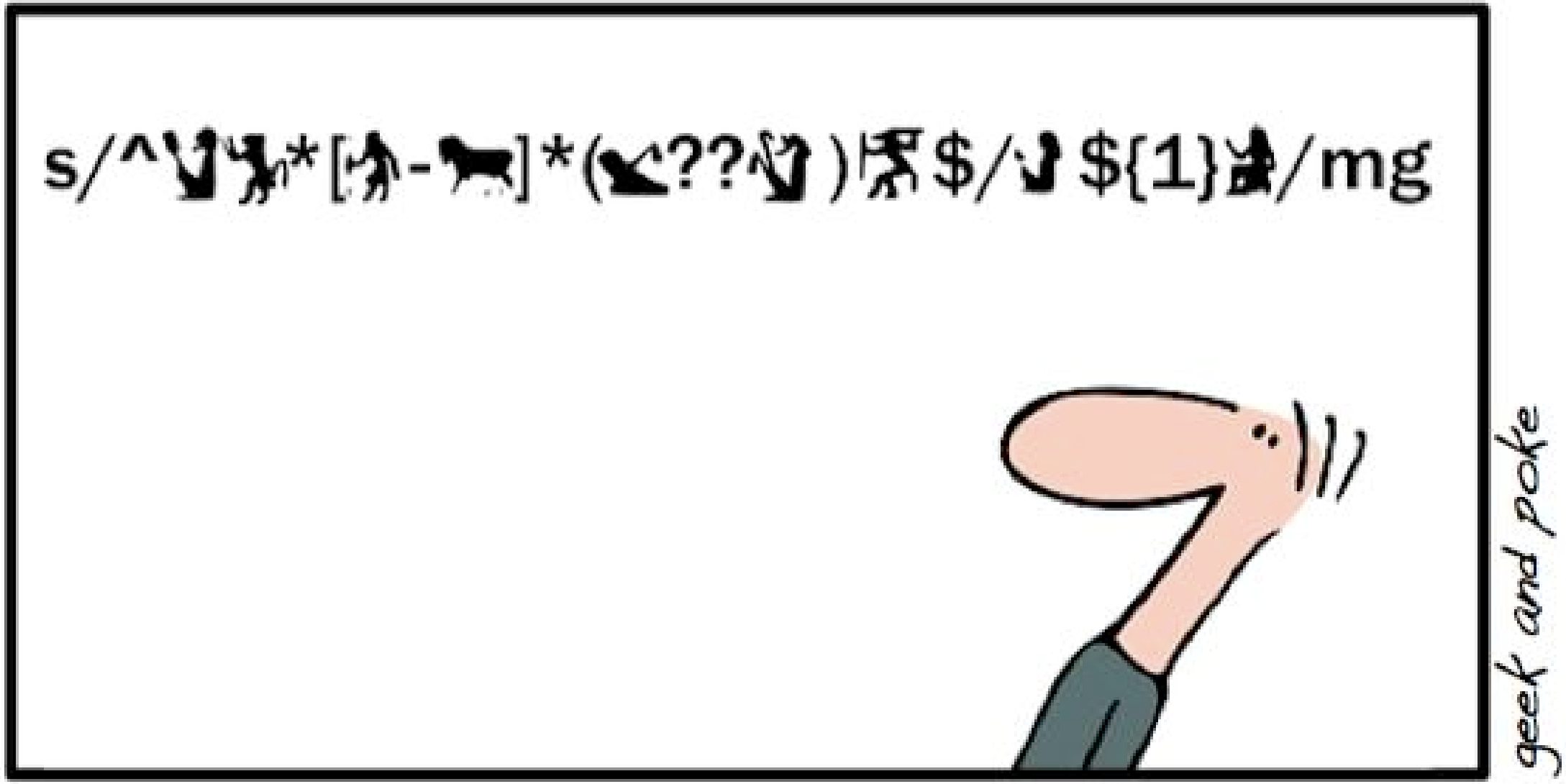
File system

Server directives

Top 15 Programming Languages

Language Rank	Types	Spectrum Ranking	
1. Java		100.0	✓
2. C		99.9	✓
3. C++		99.6	✓
4. Python		95.8	✓
5. C#		91.8	✓
6. R		84.7	✓
7. PHP		84.5	✓
8. JavaScript		83.0	✓
9. Ruby		75.3	✓
10. Matlab		72.4	✓
11. Shell		71.4	✓
12. SQL		70.9	✓
13. Assembly		67.9	✗
14. Go		67.9	✓
15. Perl		66.9	✓

Source: IEEE, July 2015



ANCIENT EGYPTIAN REGEXP

A Brief History

1940s
McCulloch and Pitts
Neural network theory

1956
Stephen Cole Kleene
Regular events/sets
=> Regular Expressions

1968
Ken Thompson
First implementation in computing

grep

g/re/p

g/<regular expression>/p

g/<regular expression>/p

[G]lobal search

[P]rint result


Regular
Expressions
Other pattern
matching
implementations

1940s
McCulloch and Pitts
Neural network theory

1956
Stephen Cole Kleene
Regular events/sets
=> Regular Expressions

1968
Ken Thompson
First implementation in computing

1973
Ken Thompson
First release of *grep*

1986
Henry Spencer
regex library

1987
Larry Wall
Integration into Perl

I define UNIX as *30 definitions of regular expressions living under one roof*.

— Donald Knuth, 1999
Digital Typography, ch. 33

1940s
McCulloch and Pitts
Neural network theory

1956
Stephen Cole Kleene
Regular events/sets
=> Regular Expressions

1968
Ken Thompson
First implementation in computing

1973
Ken Thompson
First release of *grep*

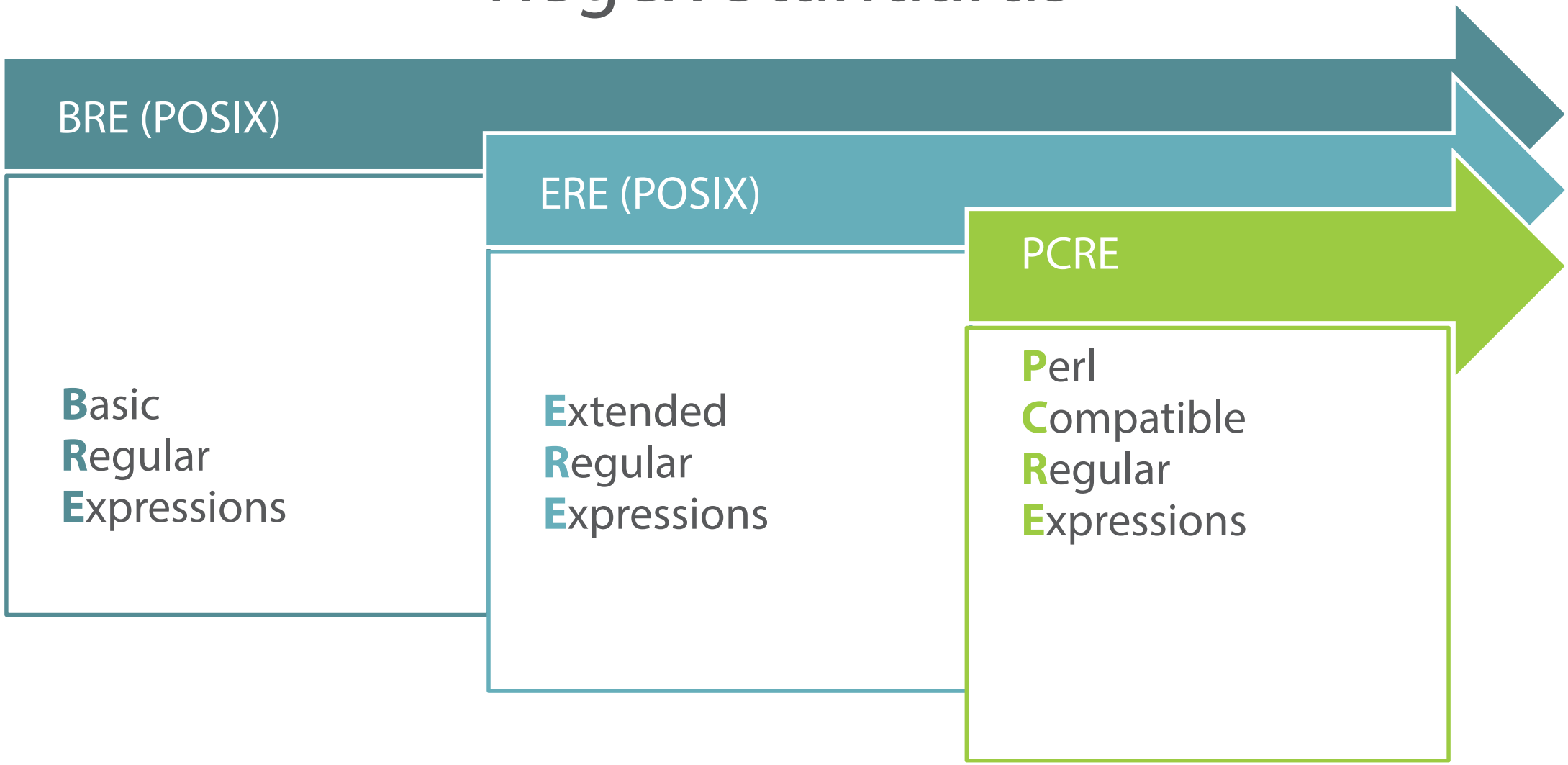
1986
Henry Spencer
regex library

1987
Larry Wall
Integration into Perl

1992
POSIX.2: standarization
into BRE and ERE

1997
Philip Hazel
PCRE

Regex Standards



* **POSIX** = **P**ortable **O**perating **S**ystem **I**nterface for uni**X**

Regular Expressions vs. Regex

How It Works

How It Works

Pattern

Determining Pattern Rules

Photo by Squaio

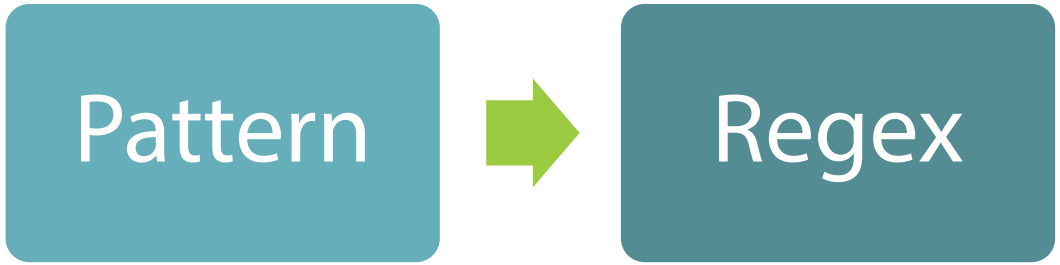
RFC 4122	A Universally Unique IDentifier (UUID) URN Namespace	July 2005	
RFC 4213	Basic Transition Mechanisms for IPv6 Hosts and Routers	October 2005	6in4
RFC 4217	Securing FTP with TLS	October 2005	SSL FTP (FTPS)
RFC 4271	Border Gateway Protocol 4	January 2006	Border Gateway Protocol
RFC 4287	The Atom Syndication Format	December 2005	Atom
RFC 4251	The Secure Shell (SSH) Protocol Architecture	January 2006	SSH-2
RFC 4291	IP Version 6 Addressing Architecture	February 2006	IPv6
RFC 4353	A Framework for Conferencing with the Session Initiation Protocol (SIP)	February 2006	Conference call
RFC 4408	Sender Policy Framework (SPF) for Authorizing Use of Domains in E-Mail, Version 1	January 2006	SPF
RFC 4422	Simple Authentication and Security Layer (SASL)	June 2006	SASL
RFC 4541	Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches	May 2006	IGMP snooping
RFC 4575	A Session Initiation Protocol (SIP) Event Package for Conference State	August 2006	Conference call
RFC 4579	Session Initiation Protocol (SIP) Call Control - Conferencing for User Agents	August 2006	Conference call
RFC 4634	US Secure Hash Algorithms (SHA and HMAC-SHA)	July 2006	SHA-1, SHA-2
RFC 4646	Tags for Identifying Languages	September 2006	language tags
RFC 4787	Network Address Translation (NAT) Behavioral Requirements for Unicast UDP	January 2007	NAT
RFC 4880	OpenPGP Message Format	November 2007	OpenPGP
RFC 4960	Stream Control Transmission Protocol	September 2007	SCTP
RFC 5023	The Atom Publishing Protocol	October 2007	Atom

Photo by Succo

Pattern Rules

- RFC 821
- RFC 822
- RFC 1035
- RFC 1123
- RFC 2142
- RFC 2821
- RFC 2822
- RFC 3696
  + associated errata
- RFC 4291
- **RFC 5321**
- **RFC 5322**
  section 3.2.3 and 3.4.1
- RFC 5952
- **RFC 6530, 6531 (i8n)**

How It Works



```
/^(?!(?:(?:\x22?\x5C[\x00-\x7E]\x22?)|(?:\x22?[^\x5C\x22]\x22?)){255,})(?!(?:
(?:\x22?\x5C[\x00-\x7E]\x22?)|(?:\x22?[^\x5C\x22]\x22?)){65,}@)(?:(?:[\x21
\x23-\x27\x2A\x2B\x2D\x2F-\x39\x3D\x3F\x5E-\x7E]+)|(?:\x22(?:[\x01-\x08\x0B
\x0C\x0E-\x1F\x21\x23-\x5B\x5D-\x7F]|(?:\x5C[\x00-\x7F]))*\x22))(?:\.(?:(?:
[\x21\x23-\x27\x2A\x2B\x2D\x2F-\x39\x3D\x3F\x5E-\x7E]+)|(?:\x22(?:[\x01-\x08
\x0B\x0C\x0E-\x1F\x21\x23-\x5B\x5D-\x7F]|(?:\x5C[\x00-\x7F]))*\x22)))*@(?:(?:
(?!.*[^.]{64,})(?:(?:(?:xn--)?[a-z0-9]+(?:-+[a-z0-9]+)*\.){1,126}){1,}(?:(?:
[a-z][a-z0-9]*)|(?:(?:xn--)[a-z0-9]+))(?:-+[a-z0-9]+)*)|(?:\[(?:(?:IPv6:(?:(?:
[a-f0-9]{1,4}(?::[a-f0-9]{1,4}){7})|(?:(?!(?:.*[a-f0-9][:\]]){7,})(?:[a-f0-9]
{1,4}(?::[a-f0-9]{1,4}){0,5})?::(?:[a-f0-9]{1,4}(?::[a-f0-9]{1,4}){0,5})?)))|
(?:(?:IPv6:(?:(?:[a-f0-9]{1,4}(?::[a-f0-9]{1,4}){5}:)|(?:(?!(?:.*[a-f0-9]:)
{5,})(?:[a-f0-9]{1,4}(?::[a-f0-9]{1,4}){0,3})?::(?:[a-f0-9]{1,4}(?::[a-f0-9]
{1,4}){0,3}:)?)))?(?:(?:25[0-5])|(?:2[0-4][0-9])|(?:1[0-9]{2})|(?:[1-9]?[0-9]
))(?:\.(?:(?:25[0-5])|(?:2[0-4][0-9])|(?:1[0-9]{2})|(?:[1-9]?[0-
9])))){3}))\]))\z/i
```

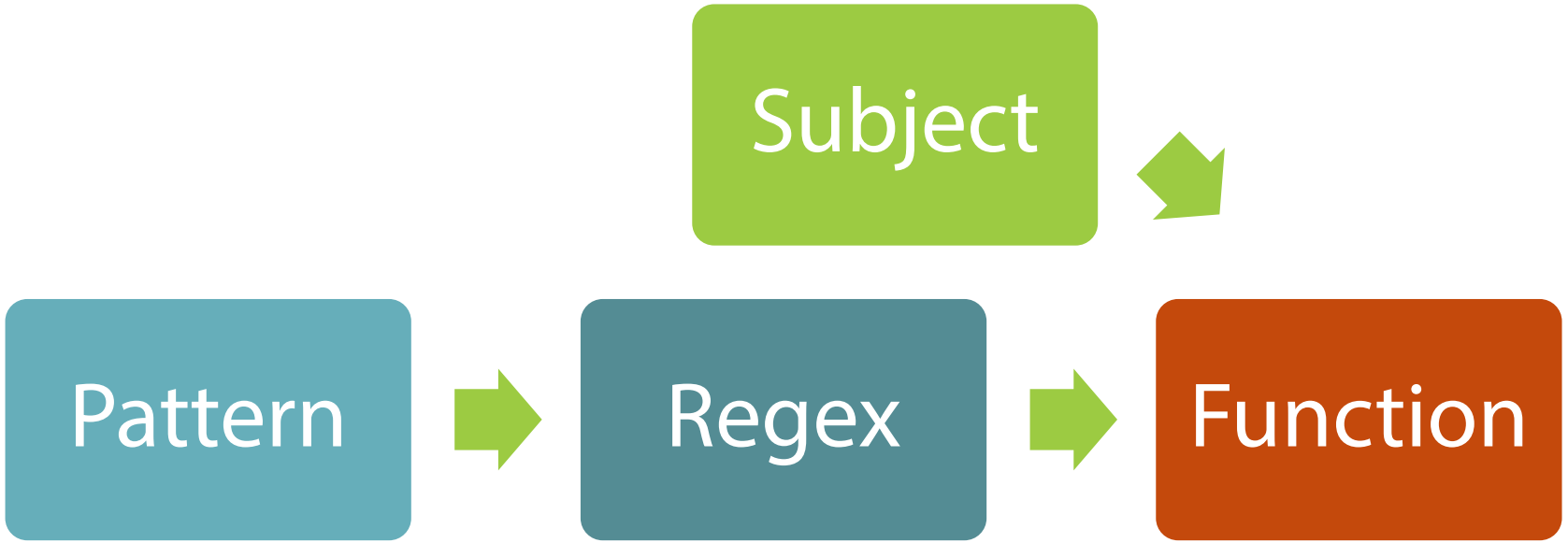

Email According to RFC 5321

```
/^[a-z0-9!#$%&'*+/=?^_`{|}~-]+(?:\.[a-z0-9!#$%&'*+/=?^_`{|}~-
]+)*@(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.)+[a-z0-9](?:[a-z0-
9-]*[a-z0-9])?\z/i
```

Basic Email Validation

- does not allow for i8n domains
- is flexible towards new top-level domain additions

How It Works



Subject String

Subject

مرحبا العالم!

Hallo Welt!

Hej Värld!

Hello World!

Ciao Modo

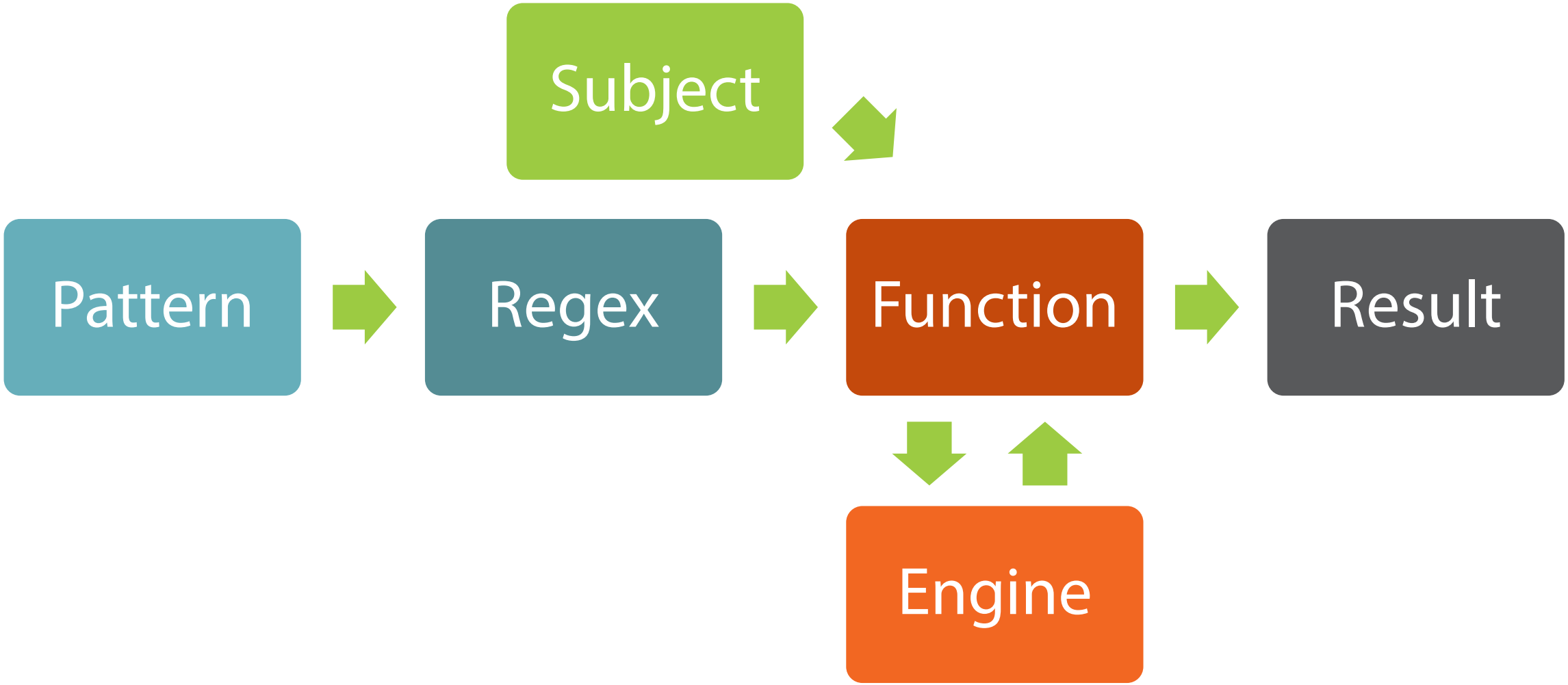
ハローワールド！

¡Holá mundo!

世界您好！

Salut le Monde!

How It Works



Result Types

Does it match ?	How many matches have been found ?	What are the matches ?
Boolean or similar	Integer	Array

Let’s Get Started