Who uses Python?

On-line games

Web services

Applications

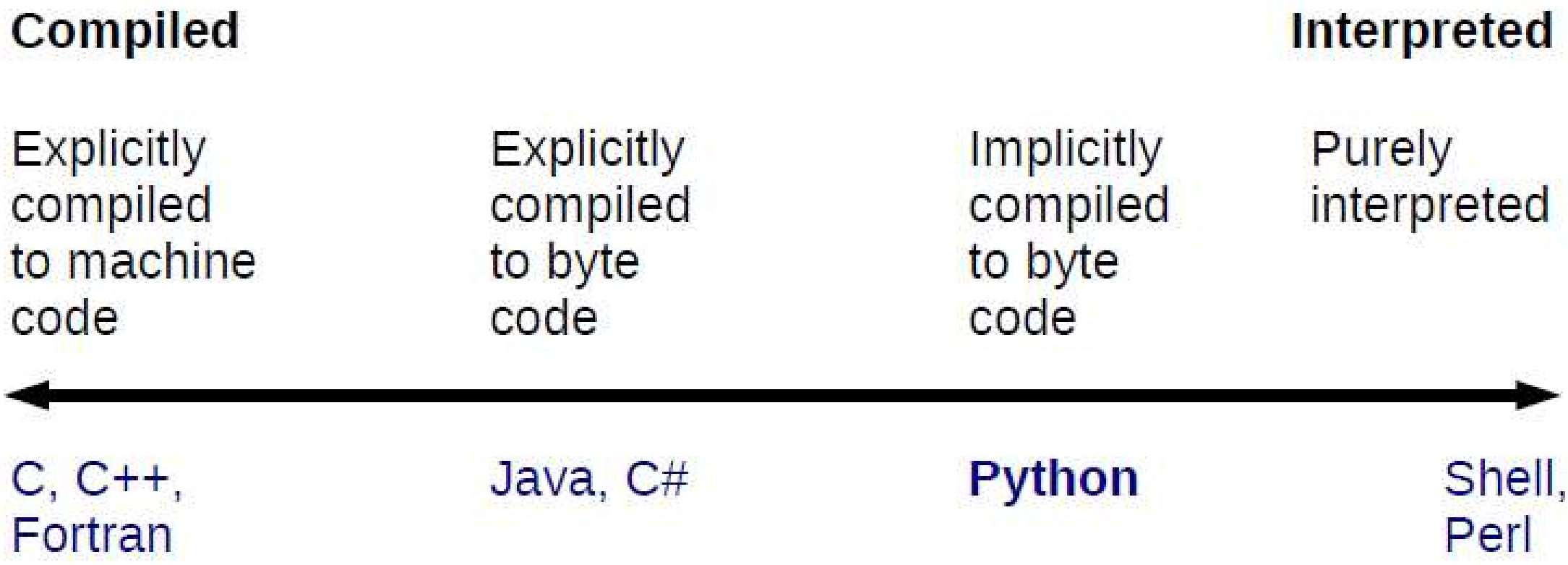
Science

Instrument control

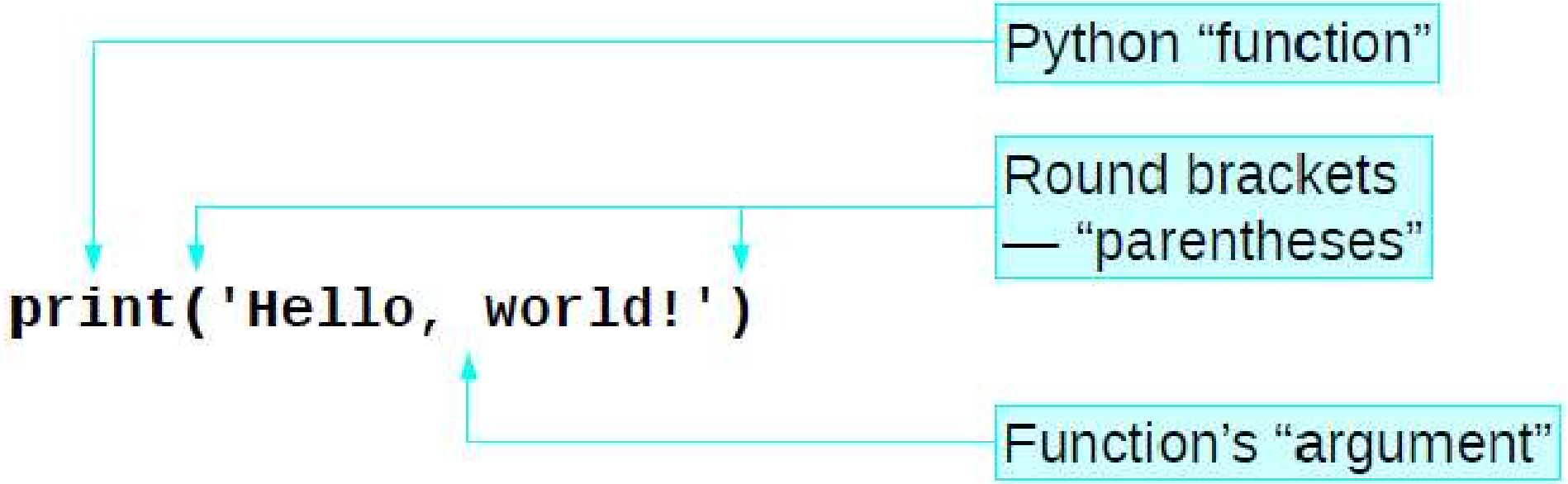
Embedded systems

en.wikipedia.org/wiki/List_of_Python_software

What sort of language is Python?



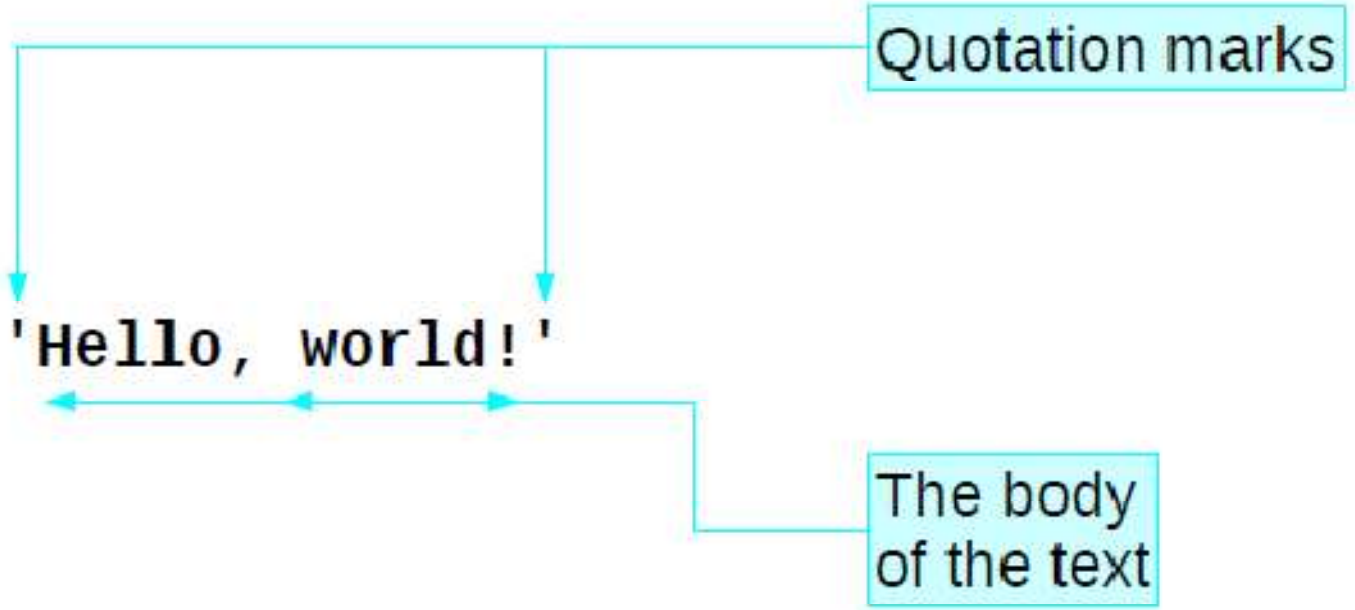
Python commands




```
print  ≠  PRINT
```

“Case sensitive”

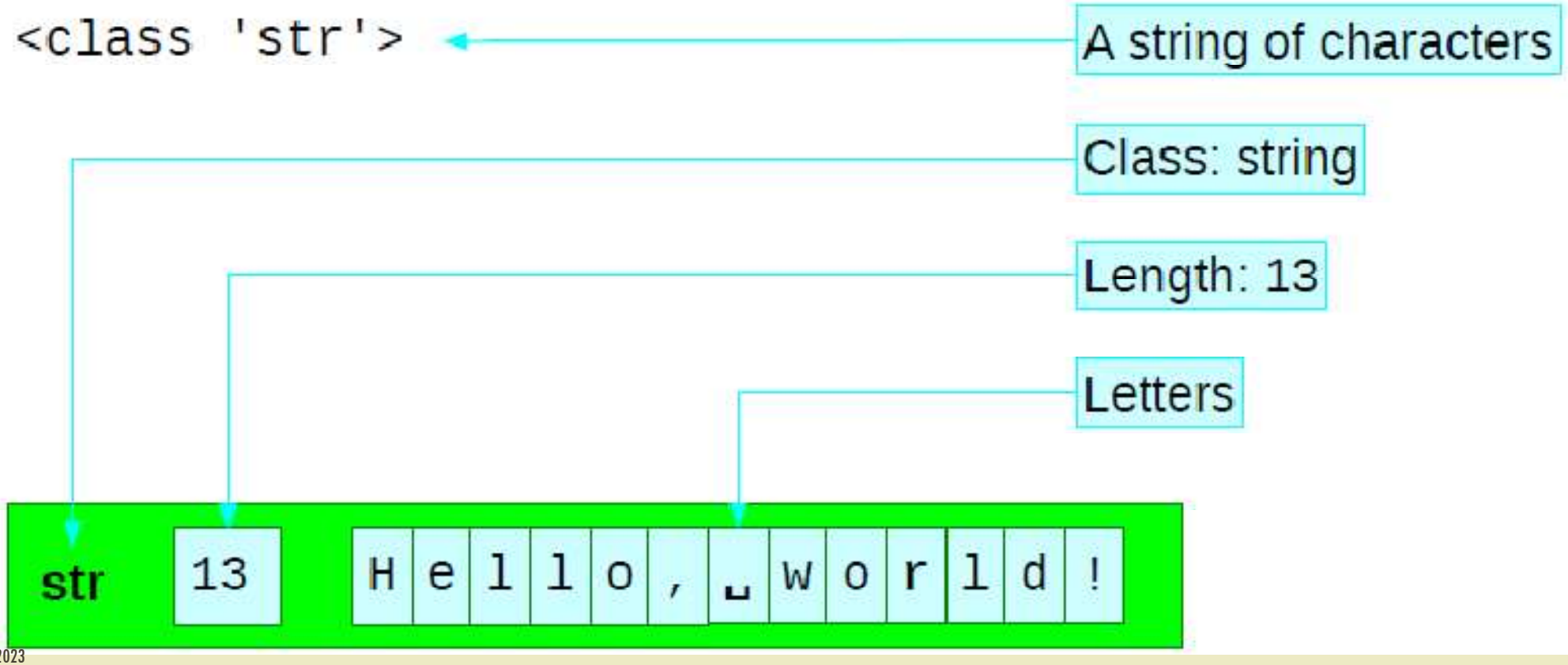
Python text



The quotes are not
part of the text itself.

Text: a “string” of characters

```
>>> type('Hello, world!')

<class 'str'>
```



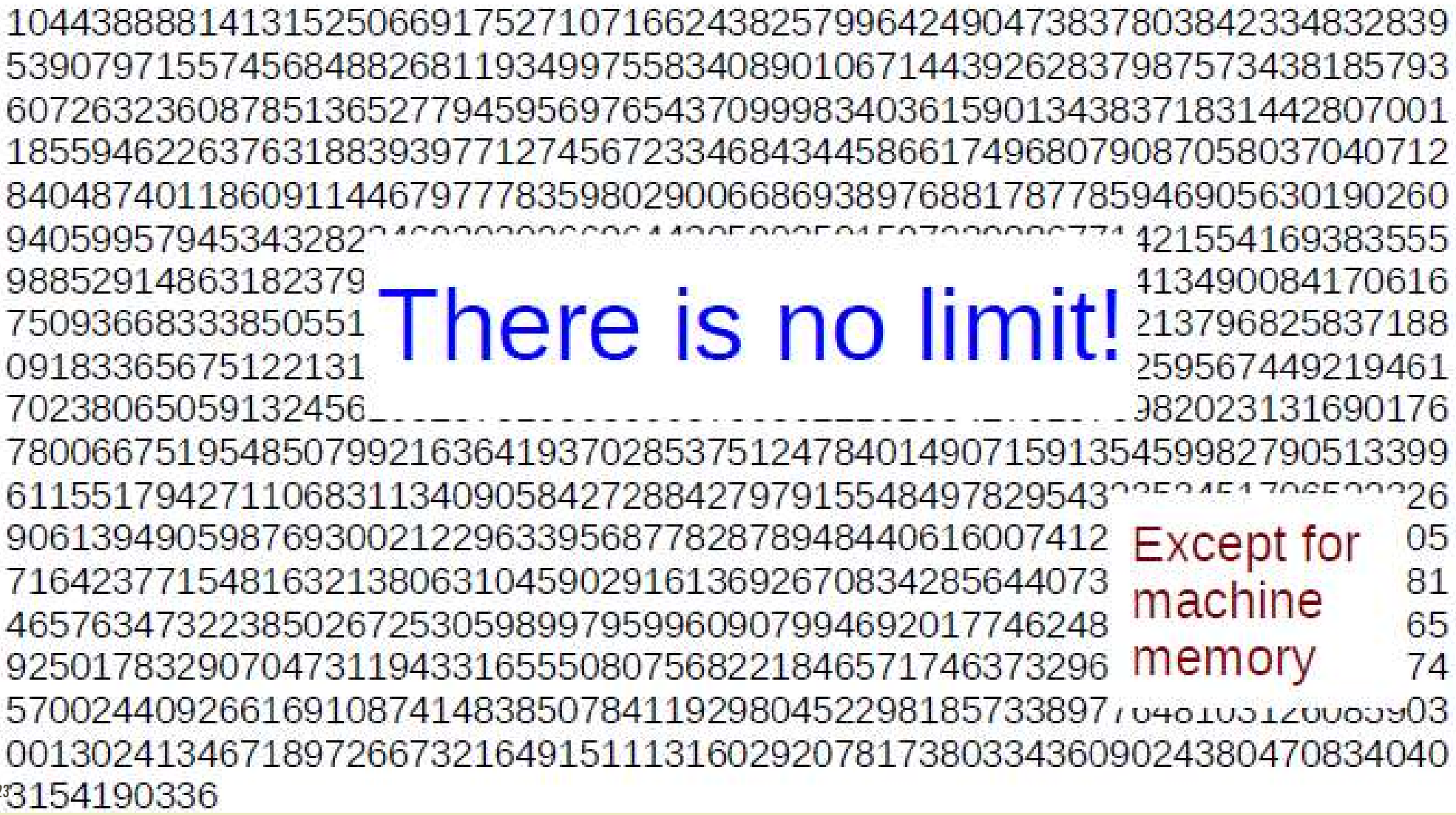
How big can a Python integer be?

```
>>> 2**2
4

>>> 4**2
16

>>> 16**2
256

>>> 256**2
65536

>>> 65536**2
4294967296
```

How big can a Python integer be?

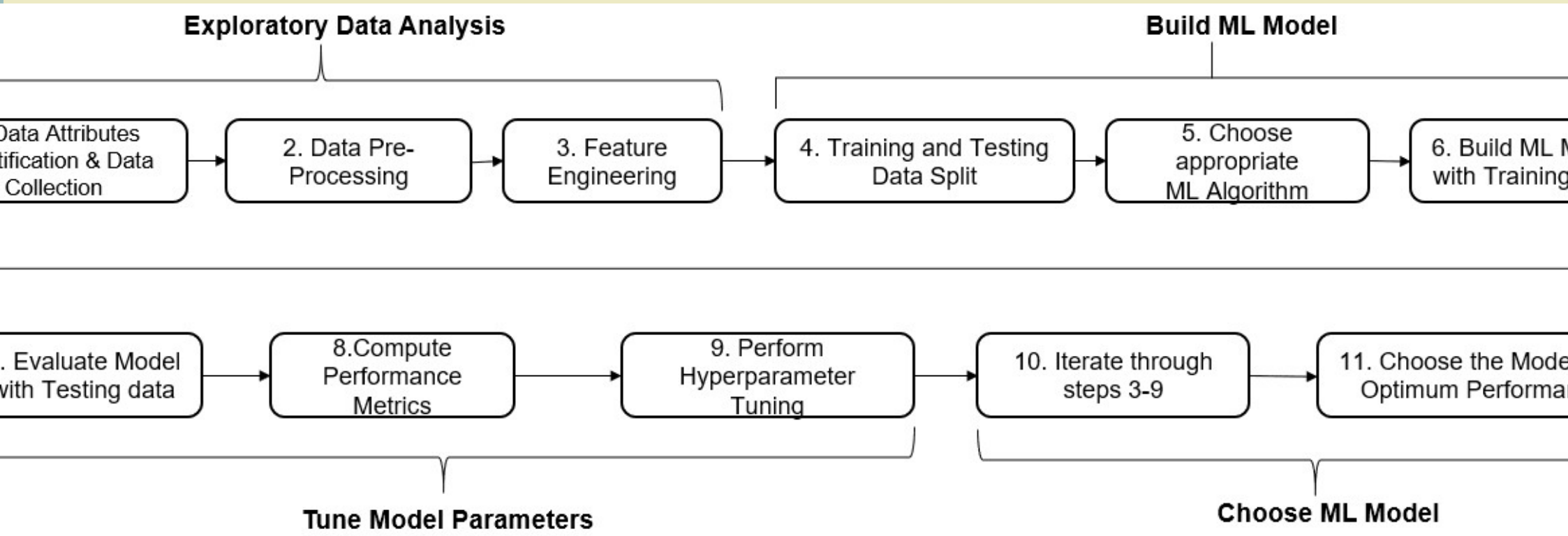
```
>>> 4294967296**2
18446744073709551616

>>> 18446744073709551616**2
340282366920938463463374607431768211456

>>> 340282366920938463463374607431768211456**2
115792089237316195423570985008687907853269984
6656405640394575840079131296399360

>>> 115792089237316195423570985008687907853269
984665640564039457584007913129639936**2
134078079299425970995740249982058461274793658
205923933777235614437217640300735469768018742
981669034276900318581864860508537538828119465
6994643364900608409 6
```

How big can a Python integer be?

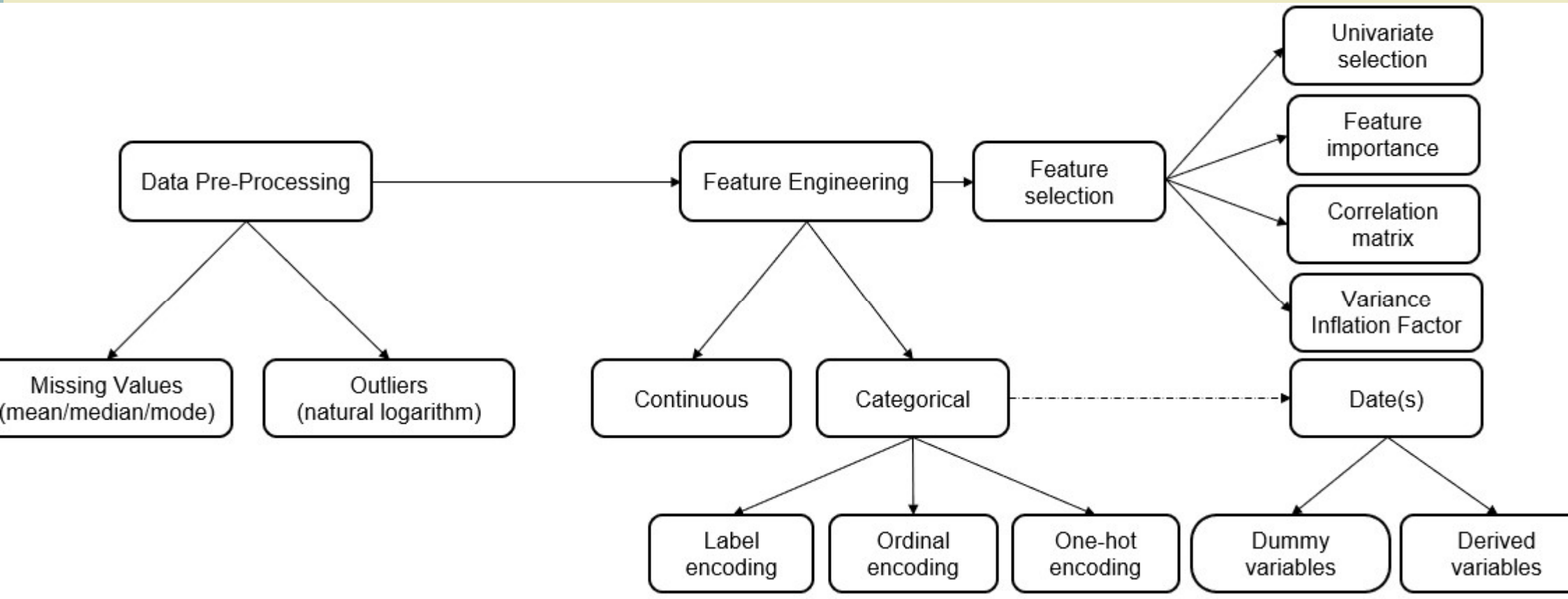


And	exec	Not
Assert	finally	or
Break	for	pass
Class	from	print
Continue	global	raise
def	if	return
del	import	try
elif	in	while
else	is	with
except	lambda	yield

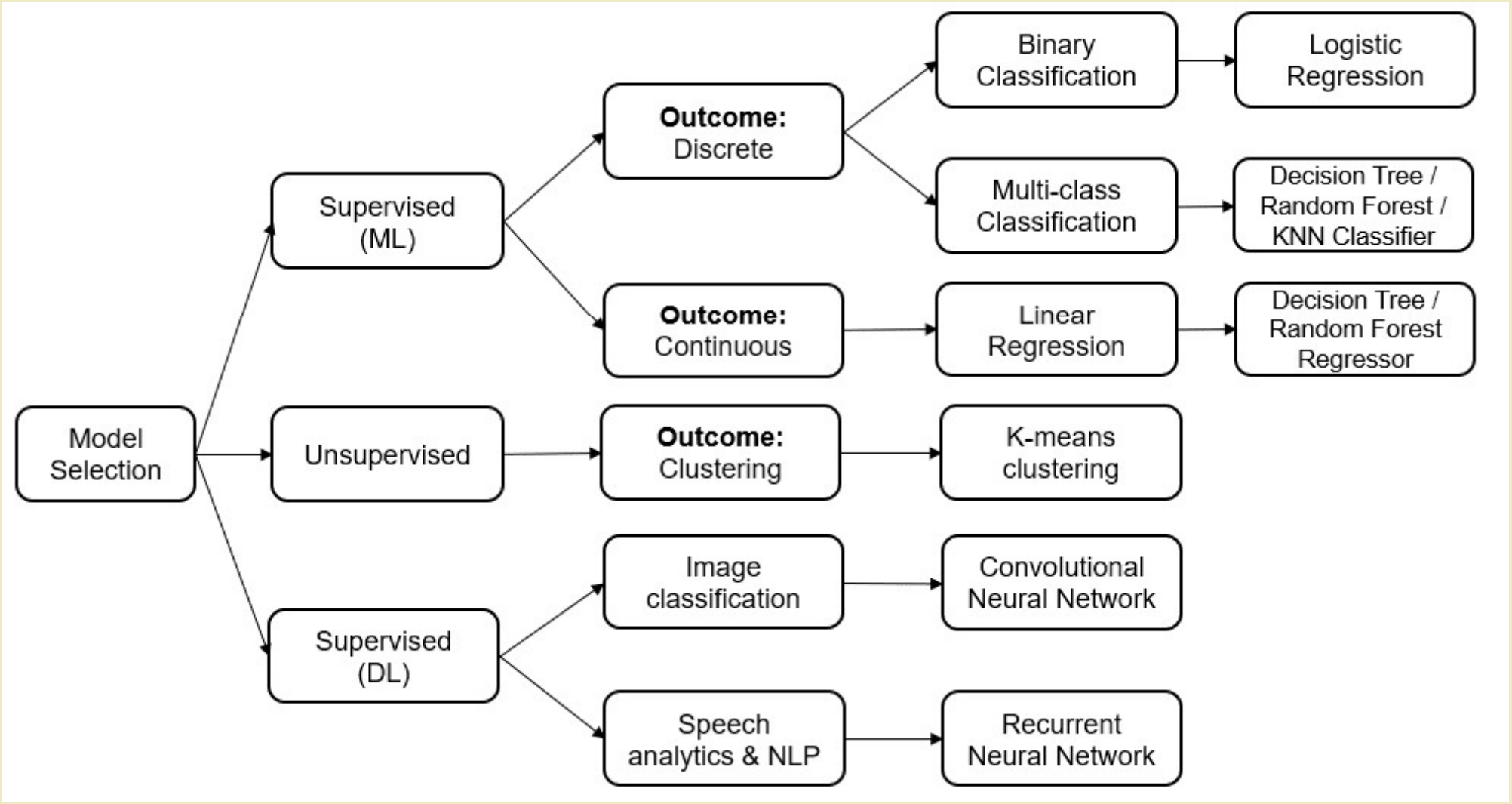
MODEL DEVELOPMENT LIFE CYCLE



STEPS INVOLVED IN EDA



MODEL SELECTION



Missing value Analysis

Why missing values

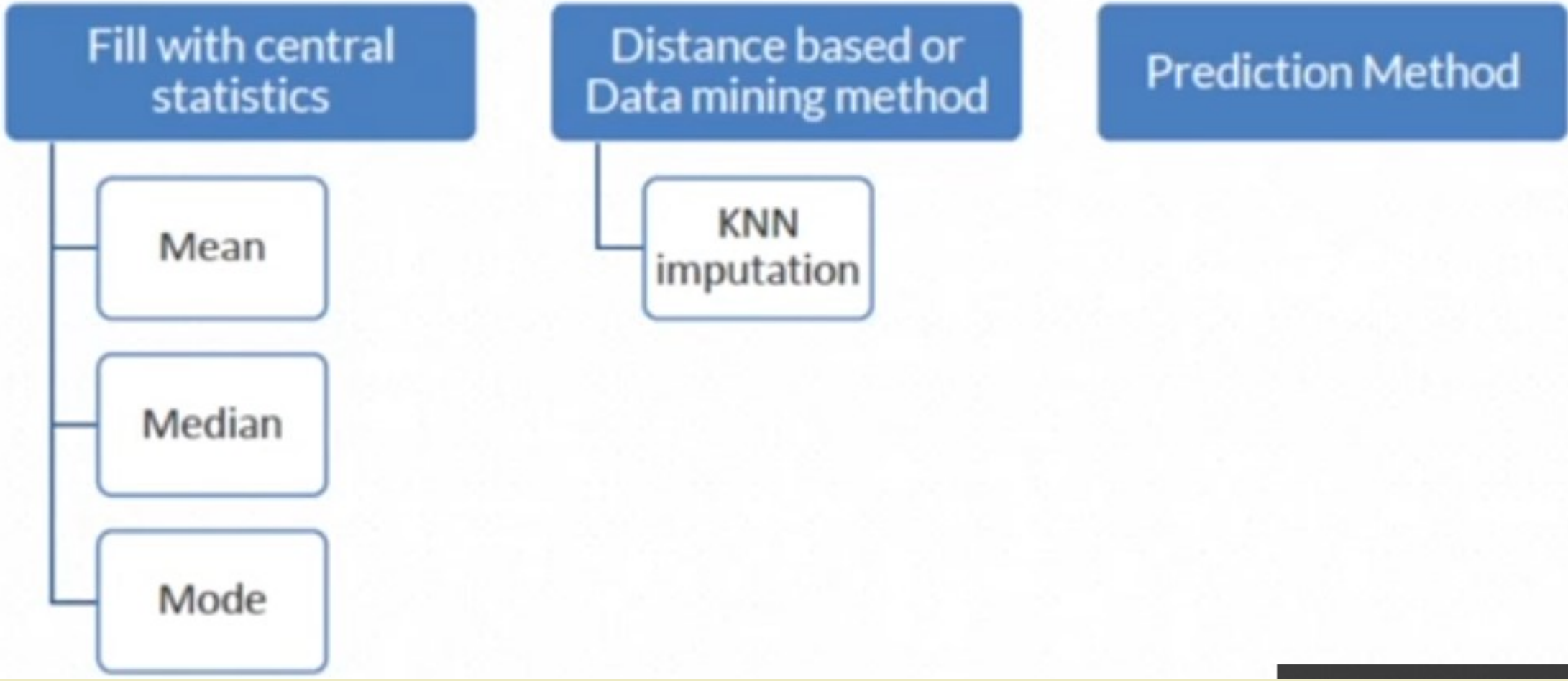
- Human Error
- Refuse to answer while surveying
- Optional box in questionnaire

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Ignore or impute missing value???

- Understand why each value is missing
- Plot bar graph
- Delete observations or variables where you do not intend to impute a value
  - Drop variable
  - Drop observation
  - Consider the variables to impute whose missing values is less than 30%

Impute missing values



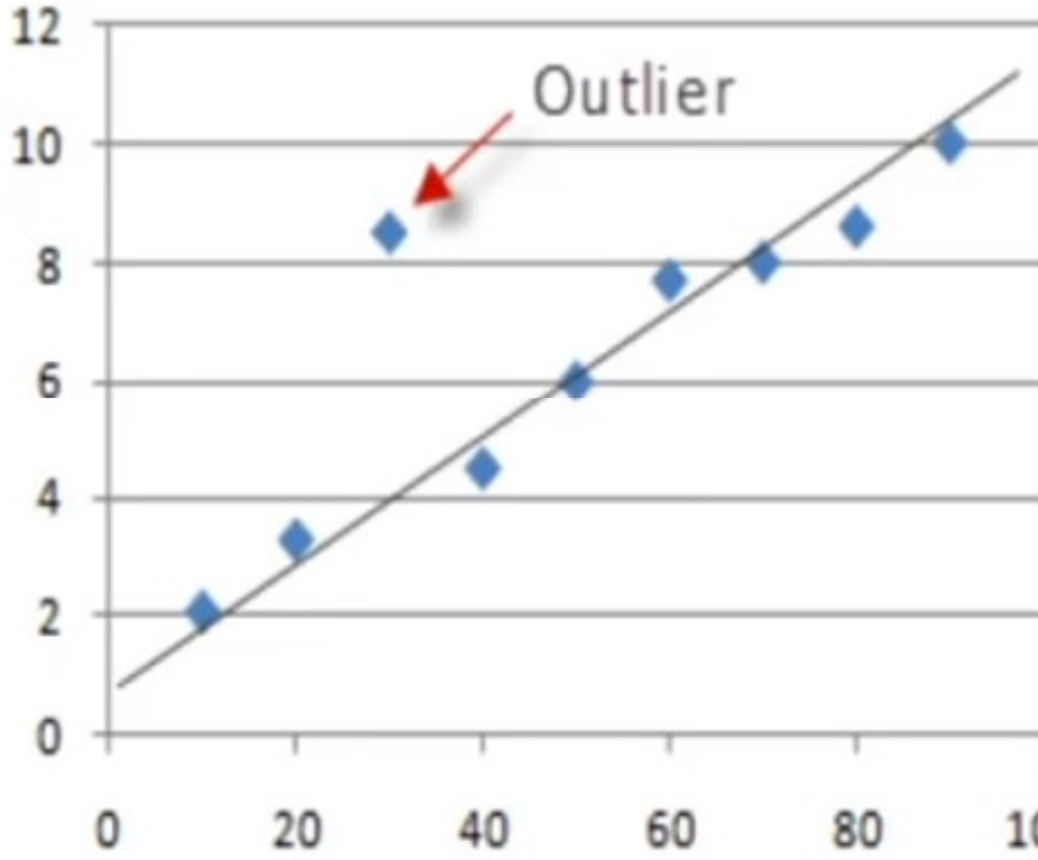
Outlier Analysis

What is an outlier?

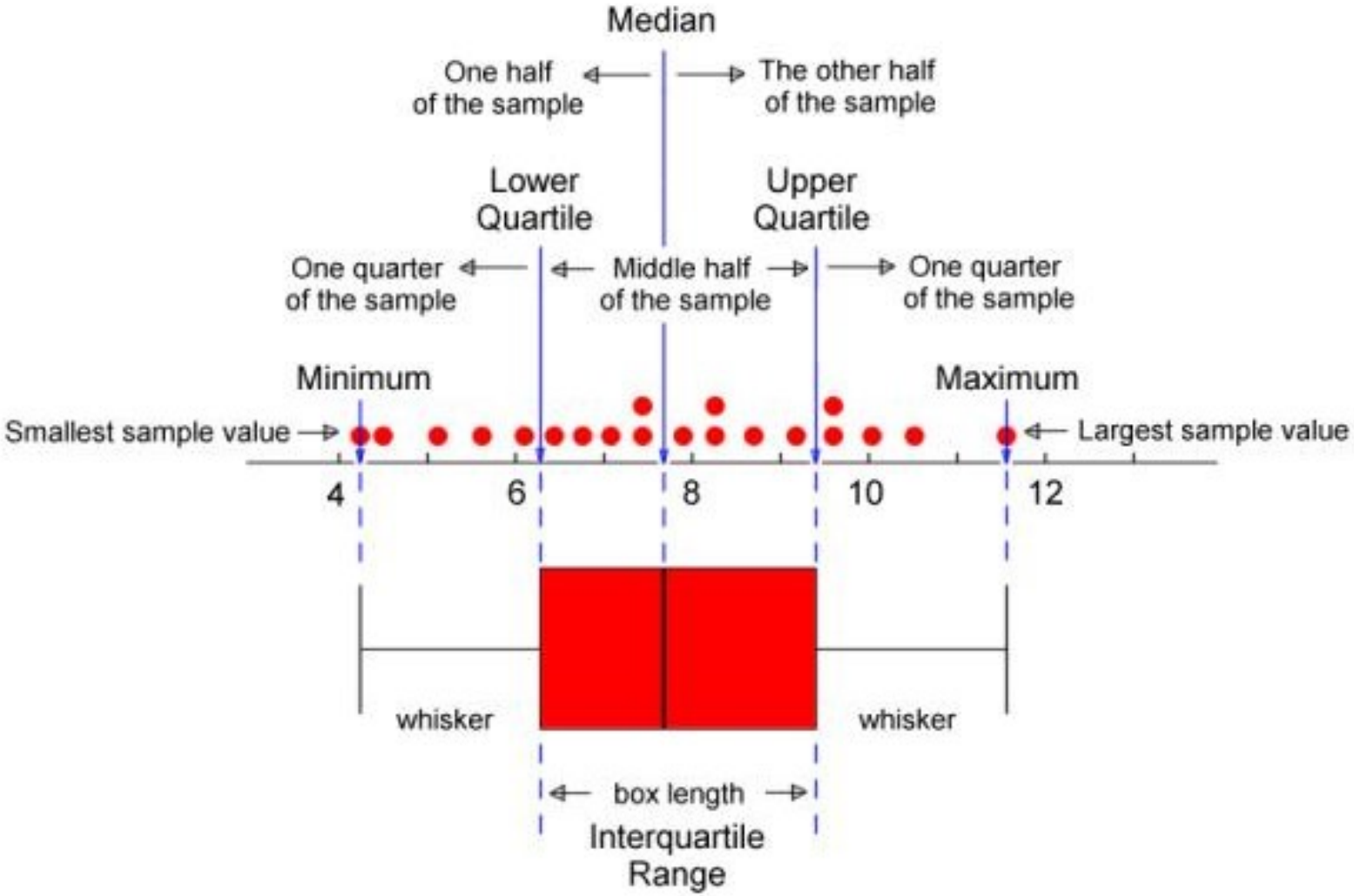
- Observations inconsistent with rest of the dataset Global Outlier

Causes of Outliers

- Poor data quality / contamination
- Low quality measurements, malfunctioning equipment, manual error
- Correct but exceptional data



BOX PLOT


Median
One half of the sample
The other half of the sample
Lower Quartile
Upper Quartile
One quarter of the sample
Middle half of the sample
One quarter of the sample
Minimum
Maximum
Smallest sample value
Largest sample value
4
6
8
10
12
whisker
whisker
box length
Interquartile Range

$$r = \frac{Cov(X, Y)}{\sigma_x \, \sigma_y}$$

$$r = \frac{\sum XY}{n \, \sigma_x \, \sigma_y}$$

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

where

$$X = x - \bar{x}$$

$$Y = y - \bar{y}$$

r ⟶ Correlation Coefficient

σ_x ⟶ standard deviation of dataset X

σ_y ⟶ standard deviation of dataset Y

Chi-Square test of Independence

- Compares two variables in a contingency table to see if they are related
- Hypothesis Testing

  Null Hypo: Two variables are independent

  Alternate Hypo: Two variables are not independent
- Uses contingency table for better representation

- Chi-square test can be calculated as

$$c^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

Contd..

- Degrees of Freedom

$$(\text{number of rows} - 1)(\text{number of columns} - 1)$$

- Calculate critical value using table
- If Chi-square statistic is greater than Critical value then reject null hypothesi

Contigency Table					
	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

$$c^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

DF = **(number of rows − 1)(number of columns − 1)**

Expected Value					
	High School	Bachelors	Masters	Ph.d.	Total
Female	50.886	49.868	50.377	49.868	201
Male	49.114	48.132	48.623	48.132	194
Total	100	98	99	98	395

50.88608

Chi-square test =	8.005907121
Degrees of Freedom =	3
Critical Value with 95% CI =	7.815

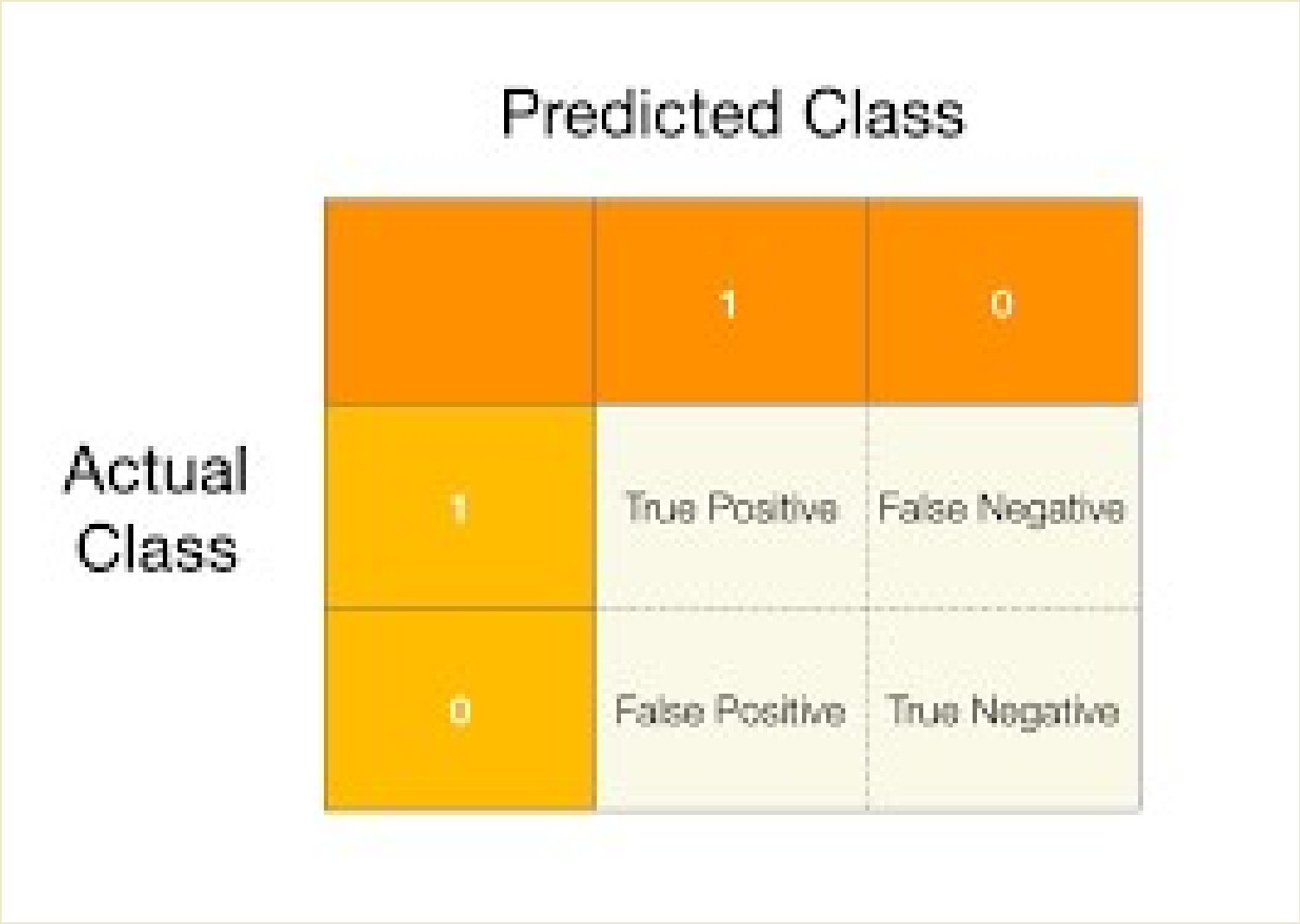
Result	Reject Null Hypothesis as ch-square test is greater than critical value

Expected Value					
	High School	Bachelors	Masters	Ph.d.	Total
Female	1.632374248	0.342372343	0.380295155	1.576992	201
Male	1.69126921	0.354720851	0.394013718	1.63387	194
Total	100	98	99	98	395

12:04

Upper-tail critical values of chi-square distribution with ν degrees of freedom

	Probability less than the critical value				
ν	0.90	0.95	0.975	0.99	0.999
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588
11	17.275	19.675	21.920	24.725	31.264
12	18.549	21.026	23.337	26.217	32.910
13	19.812	22.362	24.736	27.688	34.528
14	21.064	23.685	26.119	29.141	36.123
15	22.307	24.996	27.488	30.578	37.697
16	23.542	26.296	28.845	32.000	39.252
17	24.769	27.587	30.191	33.409	40.790
18	25.989	28.869	31.526	34.805	42.312
19	27.204	30.144	32.852	36.191	43.820
20	28.412	31.410	34.170	37.566	45.315
21	29.615	32.671	35.479	38.932	46.797


Predicted Class
Actual Class
1
0
1
True Positive
False Negative
0
False Positive
True Negative