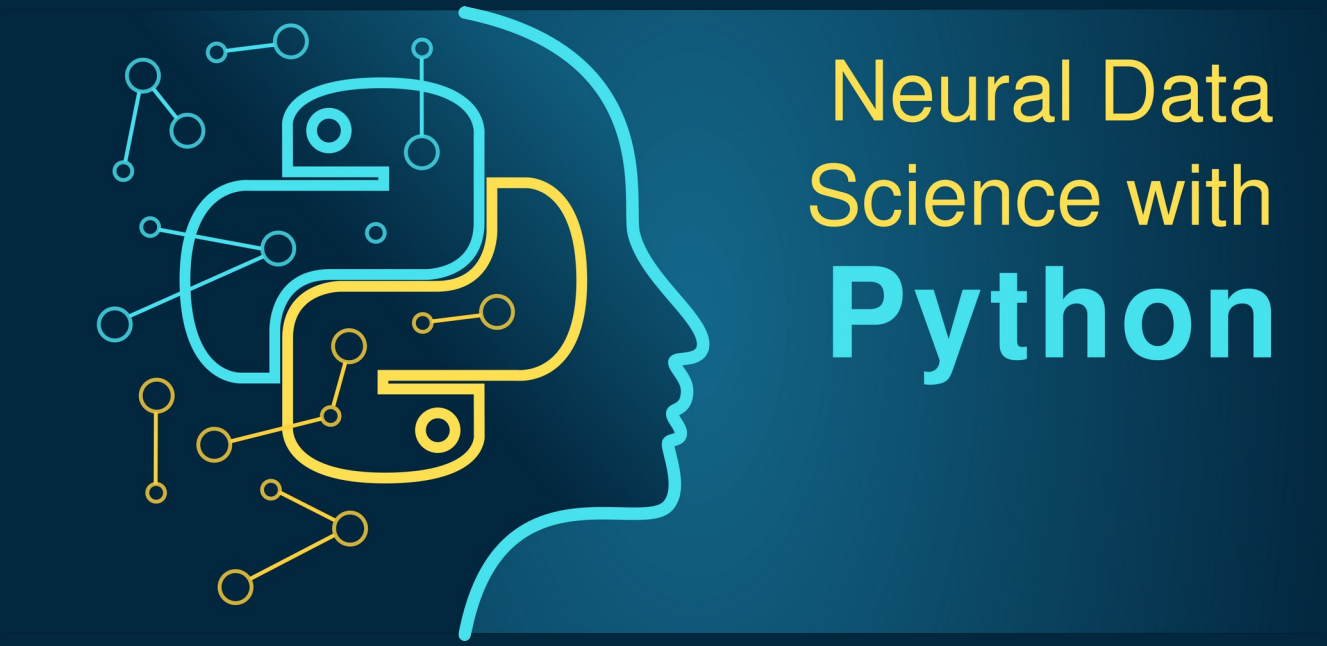


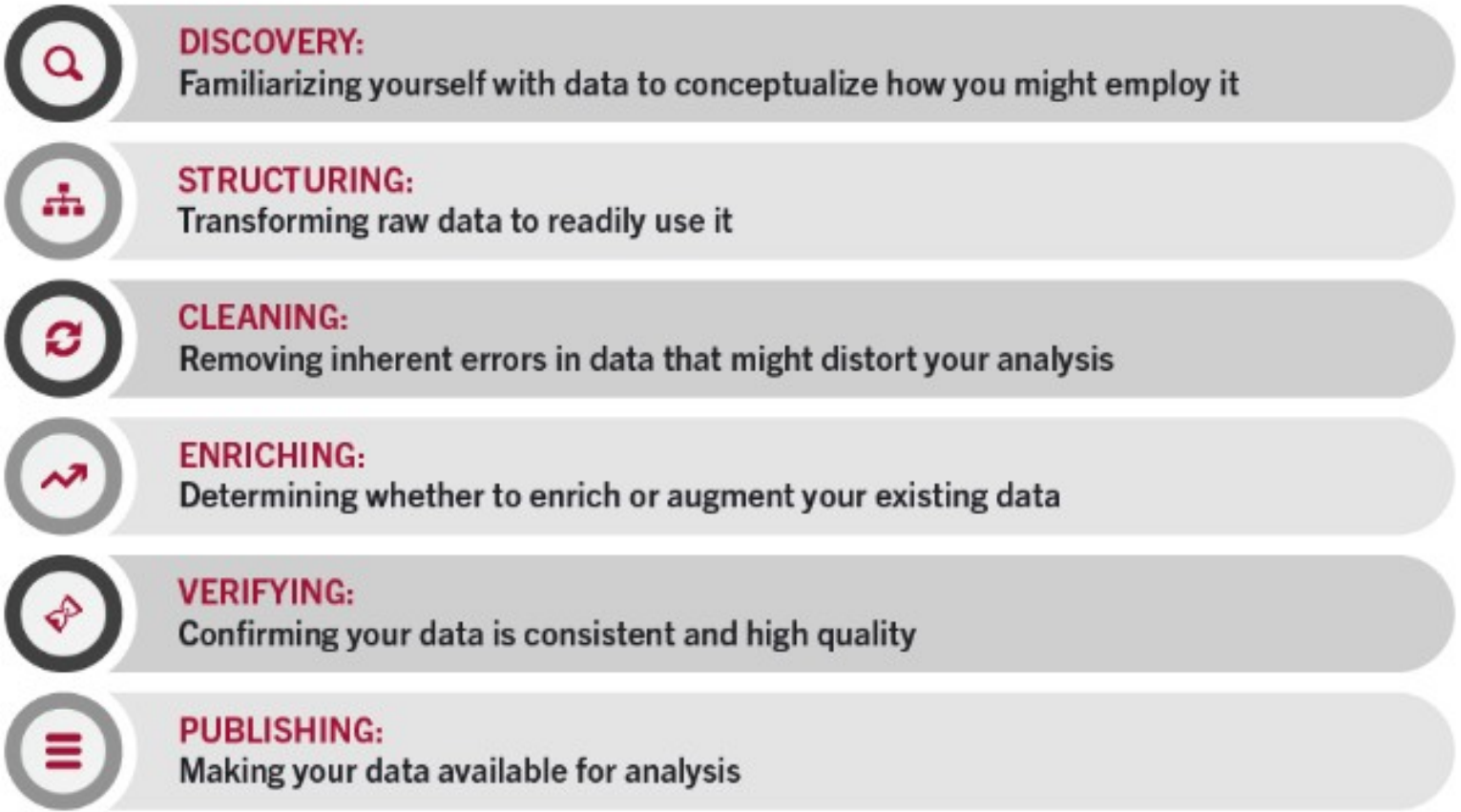
L3 : Data Wrangling with Python

Michael Graupner

SPPIN – Saint-Pères Institute for the Neurosciences

Université de Paris, CNRS

What is Data Wrangling ?



Typical workflow in the lab involves data wrangling

Data collection/gathering

Data preparation (pre-analysis)

Data analysis

Why data preparation

Data collection/gathering

Data preparation (pre-analysis)

Data analysis

For further analysis data needs to be :

- well constructed
- clean
- accurately formatted
- suitable for statistical analysis

Raw data has its challenges

Data comes in all shapes and size :

- binary files, text (csv) files, PDFs, stone tablets, images (jpg)

Different files have different formatting

- Spaces instead of NULLs, spaces instead of coma, extra rows, different column names

“Dirty” data

- unwanted anomalies
- Duplicates
- missing data

Concrete pre-processing steps

1) Cleaning data
2) Concatenating and merging data
3) Normalizing data
4) Running functions on the data
5) Data encoding

1) Cleaning data : example data-set

	0	1	2	3	4	5	6	7	8	9
0	1.0	Mickéy Mousé	56.0	70kgs	72	69	71	-	-	-
1	2.0	Donald Duck	34.0	154.89lbs	-	-	-	85	84	76
2	3.0	Mini Mouse	16.0	NaN	-	-	-	65	69	72
3	4.0	Scrooge McDuck	NaN	78kgs	78	79	72	-	-	-
4	5.0	Pink Panther	54.0	198.658lbs	-	-	-	69	NaN	75
5	6.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
6	7.0	Dewey McDuck	19.0	56kgs	-	-	-	71	78	75
7	8.0	Scööpy Doo	32.0	78kgs	78	76	75	-	-	-
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	9.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
10	10.0	Louie McDuck	12.0	45kgs	-	-	-	92	95	87

Do you notice any issues with that data-set?

Missing data

	0	1	2	3	4	5	6	7	8	9
0	1.0	Mickéy Mousé	56.0	70kgs	72	69	71	-	-	-
1	2.0	Donald Duck	34.0	154.89lbs	-	-	-	85	84	76
2	3.0	Mini Mouse	16.0	NaN	-	-	-	65	69	72
3	4.0	Scrooge McDuck	NaN	78kgs	78	79	72	-	-	-
4	5.0	Pink Panther	54.0	198.658lbs	-	-	-	69	NaN	75
5	6.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
6	7.0	Dewey McDuck	19.0	56kgs	-	-	-	71	78	75
7	8.0	Scööpy Doo	32.0	78kgs	78	76	75	-	-	-
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	9.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
10	10.0	Louie McDuck	12.0	45kgs	-	-	-	92	95	87

missing data can be dealt with by the following methods :

- *delete* : remove record
- *mean* : use mean of column
- *most frequent* : replace by most frequent value of column

Blank lines

	0	1	2	3	4	5	6	7	8	9
0	1.0	Mickéy Mousé	56.0	70kgs	72	69	71	-	-	-
1	2.0	Donald Duck	34.0	154.89lbs	-	-	-	85	84	76
2	3.0	Mini Mouse	16.0	NaN	-	-	-	65	69	72
3	4.0	Scrooge McDuck	NaN	78kgs	78	79	72	-	-	-
4	5.0	Pink Panther	54.0	198.658lbs	-	-	-	69	NaN	75
5	6.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
6	7.0	Dewey McDuck	19.0	56kgs	-	-	-	71	78	75
7	8.0	Scööpy Doo	32.0	78kgs	78	76	75	-	-	-
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	9.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
10	10.0	Louie McDuck	12.0	45kgs	-	-	-	92	95	87

Blank records need to be removed from data-set

Data format inconsistent

	0	1	2	3	4	5	6	7	8	9
0	1.0	Mickéy Mousé	56.0	70kgs	72	69	71	-	-	-
1	2.0	Donald Duck	34.0	154.89lbs	-	-	-	85	84	76
2	3.0	Mini Mouse	16.0	NaN	-	-	-	65	69	72
3	4.0	Scrooge McDuck	NaN	78kgs	78	79	72	-	-	-
4	5.0	Pink Panther	54.0	198.658lbs	-	-	-	69	NaN	75
5	6.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
6	7.0	Dewey McDuck	19.0	56kgs	-	-	-	71	78	75
7	8.0	Scööpy Doo	32.0	78kgs	78	76	75	-	-	-
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	9.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
10	10.0	Louie McDuck	12.0	45kgs	-	-	-	92	95	87

Convert data to make uniform (same unit and format of all entries)

Multiple parameters in one column

	0	1	2	3	4	5	6	7	8	9
0	1.0	Mickéy Mousé	56.0	70kgs	72	69	71	-	-	-
1	2.0	Donald Duck	34.0	154.89lbs	-	-	-	85	84	76
2	3.0	Mini Mouse	16.0	NaN	-	-	-	65	69	72
3	4.0	Scrooge McDuck	NaN	78kgs	78	79	72	-	-	-
4	5.0	Pink Panther	54.0	198.658lbs	-	-	-	69	NaN	75
5	6.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
6	7.0	Dewey McDuck	19.0	56kgs	-	-	-	71	78	75
7	8.0	Scöopy Doo	32.0	78kgs	78	76	75	-	-	-
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	9.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
10	10.0	Louie McDuck	12.0	45kgs	-	-	-	92	95	87

Separate into individual parameter entries (here one column for first and one for last name)

Duplicate entries

	0	1	2	3	4	5	6	7	8	9
0	1.0	Mickéy Mousé	56.0	70kgs	72	69	71	-	-	-
1	2.0	Donald Duck	34.0	154.89lbs	-	-	-	85	84	76
2	3.0	Mini Mouse	16.0	NaN	-	-	-	65	69	72
3	4.0	Scrooge McDuck	NaN	78kgs	78	79	72	-	-	-
4	5.0	Pink Panther	54.0	198.658lbs	-	-	-	69	NaN	75
5	6.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
6	7.0	Dewey McDuck	19.0	56kgs	-	-	-	71	78	75
7	8.0	Scööpy Doo	32.0	78kgs	78	76	75	-	-	-
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	9.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
10	10.0	Louie McDuck	12.0	45kgs	-	-	-	92	95	87

Remove duplicates/multiple entries of the same data-point

Same parameter in different columns

	0	1	2	3	4	5	6	7	8	9
0	1.0	Mickéy Mousé	56.0	70kgs	72	69	71	-	-	-
1	2.0	Donald Duck	34.0	154.89lbs	-	-	-	85	84	76
2	3.0	Mini Mouse	16.0	NaN	-	-	-	65	69	72
3	4.0	Scrooge McDuck	NaN	78kgs	78	79	72	-	-	-
4	5.0	Pink Panther	54.0	198.658lbs	-	-	-	69	NaN	75
5	6.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
6	7.0	Dewey McDuck	19.0	56kgs	-	-	-	71	78	75
7	8.0	Scööpy Doo	32.0	78kgs	78	76	75	-	-	-
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	9.0	Huey McDuck	52.0	189lbs	-	-	-	68	75	72
10	10.0	Louie McDuck	12.0	45kgs	-	-	-	92	95	87

Merge columns containing the same information

2) Concatenating and merging data

January Data

Date	Open	High	Low	Close
2-Jan-20	27.21	27.95	26.62	27.89
3-Jan-20	28.16	28.95	27.73	28.82
6-Jan-20	28.53	28.81	28	28.39
...	...	...	...	...
29-Jan-20	26.33	26.64	25.85	25.98
30-Jan-20	25.7	26.54	25.48	26
31-Jan-20	25.68	26.55	25.6	26.28

+

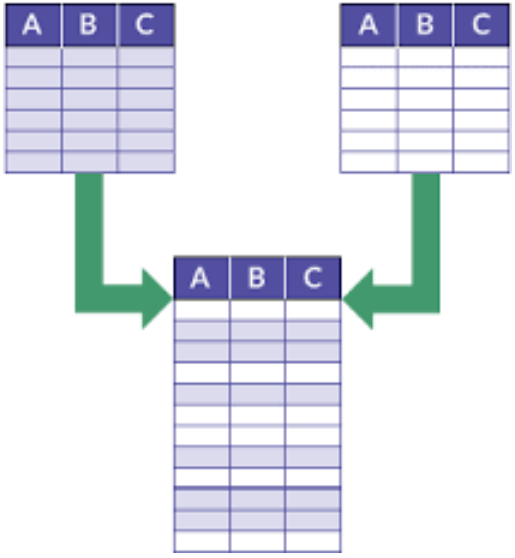
February Data

Date	Open	High	Low	Close
3-Feb-20	26.35	27.59	26.25	26.52
4-Feb-20	26.98	27.62	26.34	27.34
5-Feb-20	27.77	28.03	27.3	27.89
...	...	...	...	...
26-Feb-20	27.59	28.93	27.3	27.84
27-Feb-20	27.13	27.56	25.85	25.89
28-Feb-20	24.9	26.66	24.51	26.45

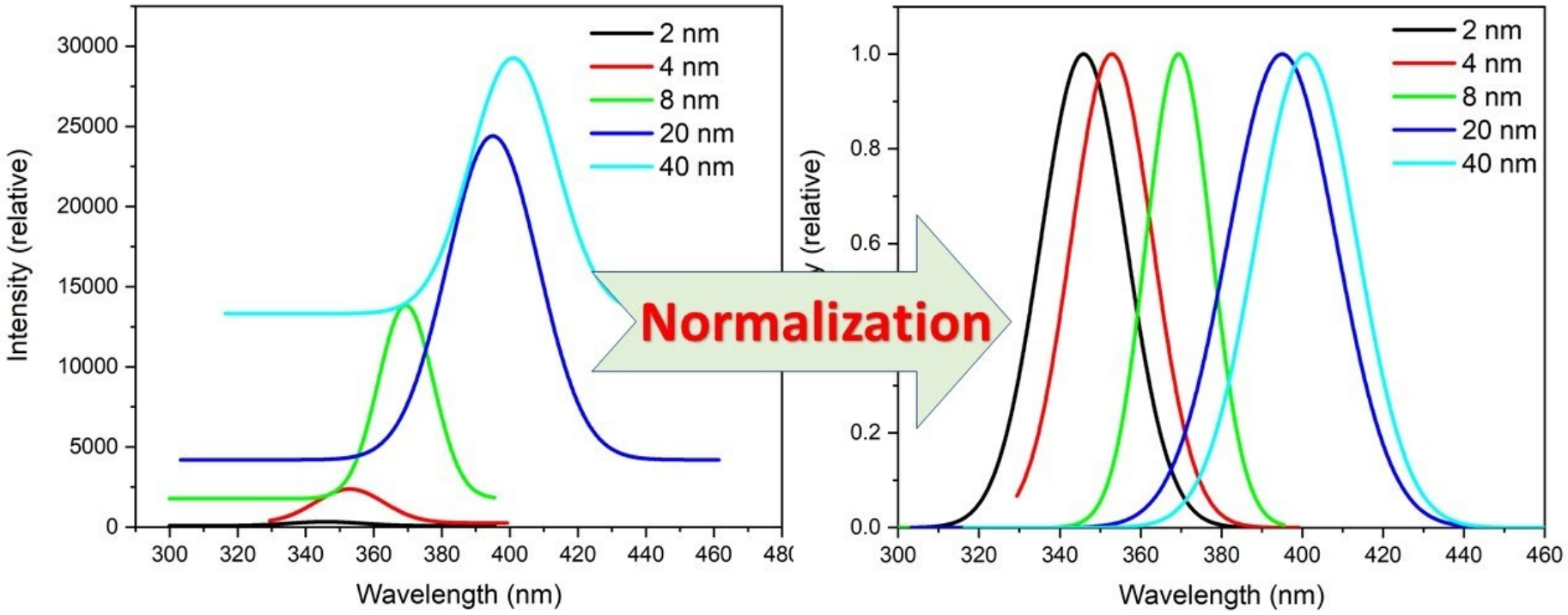
=

January Data / February Data

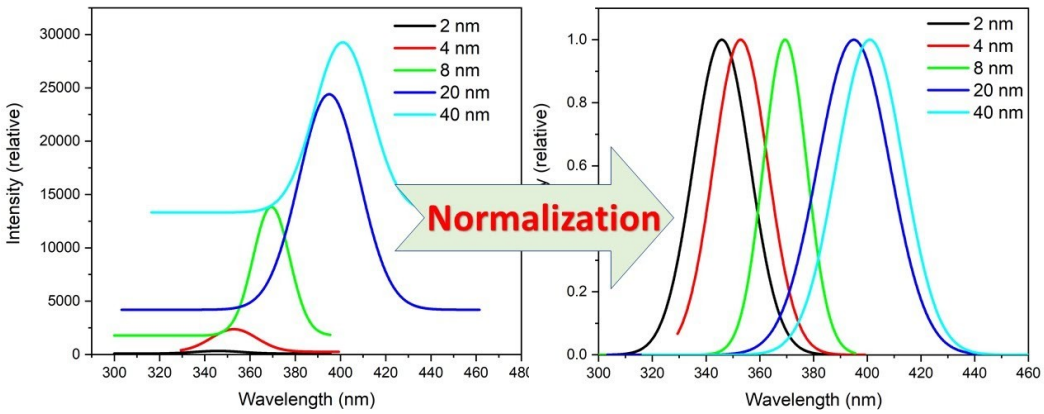
Date	Open	High	Low	Close
2-Jan-20	27.21	27.95	26.62	27.89
3-Jan-20	28.16	28.95	27.73	28.82
6-Jan-20	28.53	28.81	28	28.39
...	...	...	...	...
29-Jan-20	26.33	26.64	25.85	25.98
30-Jan-20	25.7	26.54	25.48	26
31-Jan-20	25.68	26.55	25.6	26.28
3-Feb-20	26.35	27.59	26.25	26.52
4-Feb-20	26.98	27.62	26.34	27.34
5-Feb-20	27.77	28.03	27.3	27.89
...	...	...	...	...
26-Feb-20	27.59	28.93	27.3	27.84
27-Feb-20	27.13	27.56	25.85	25.89
28-Feb-20	24.9	26.66	24.51	26.45



3) Normalize data



2) Common normalization/standardization methods



- correct for horizontal offset : baseline subtraction, mean subtraction
- Rescale : common minimum-maximum range (*e.g.* [0,1], [-1,1], [0,100])
- Standardization : subtract mean and normalized by Standard deviation (Guassian data) :

z-score

$$z = \frac{x - \mu}{\sigma}$$

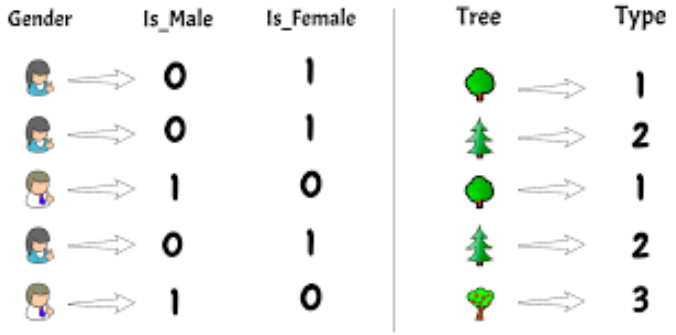
μ = Mean
σ = Standard Deviation

- Log transformation : for data following exponential distribution (large spread, reduces impact of extreme values)
- Normalization by sum : normalize each data-point by the sum of all values in the dataset, i.e., the sum of all values is 1 (e.g. probability distributions)

4) Running functions on the data

- combining data : *add, mean*
- mathematical functions : *sin(), cos(), exp()*
- custom functions : *def ...()*

4) Data encoding



- put data in format suitable for further analysis