PyCon PL '24

Fundamentals of Pandas and Polars in scientific applications

•••

Igor Zubrycki, Lidia Lipińska-Zubrycka

Gliwice, 30.08.2024

Project partially funded by The National Centre for Research and Development, grant number LIDER/50/0203/L-11/19/NCBR/2020.

About us

Lidia Lipińska-Zubrycka

- Scientist / bioinformatician
- University of Warsaw

Igor Zubrycki

- Scientist / Machine learning specialist
- Lodz University of Technology
- R&D Machine Learning team manager

Scripts and link to Google Colab

Let's visit: https://github.com/AdoHaha/PyConPL24_Pandas_Polars

Dataset 1: Breast cancer data

The results show data obtained from a breast tumor biopsy along with the diagnosis: malignant or benign form of cancer.

Features:

1. ID number
2. Diagnosis (M = malignant (pl. złośliwy), B = benign (pl. łagodny))

Ten real-valued features are computed for each cell nucleus:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

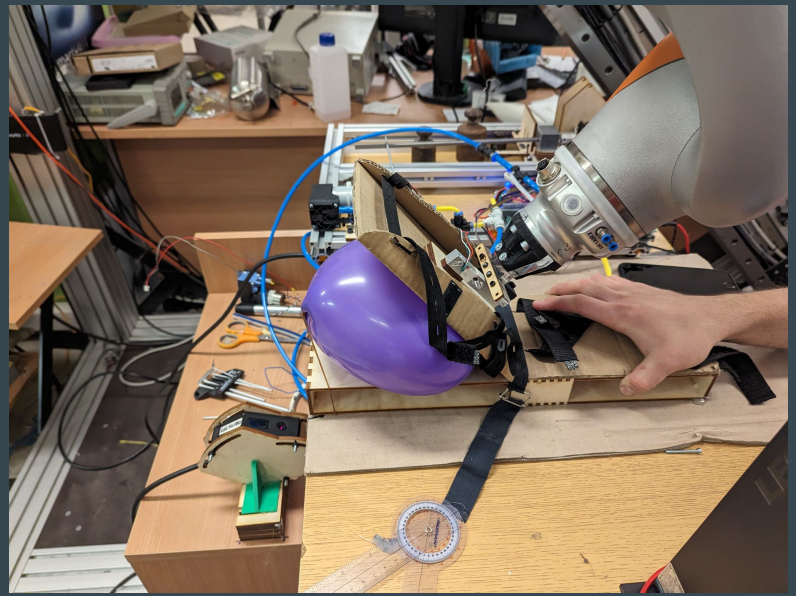
Dataset 1: Breast cancer dataset (Pandas)

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concav points_mea
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.1471
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.0701
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.1279
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.1052
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.1043
...	...	...	...	...	...	...	...	...	...	
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.1389
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.0979
566	926954	M	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.0530
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.1520
568	92751	B	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.0000

569 rows × 33 columns

1. Do we see differences between patients with benign and malignant forms of cancer based on the experimental data obtained?
2. Which of the determined parameters best describes the form of the cancer?

Dataset 2: Pneumatic robot dataset (Polars)



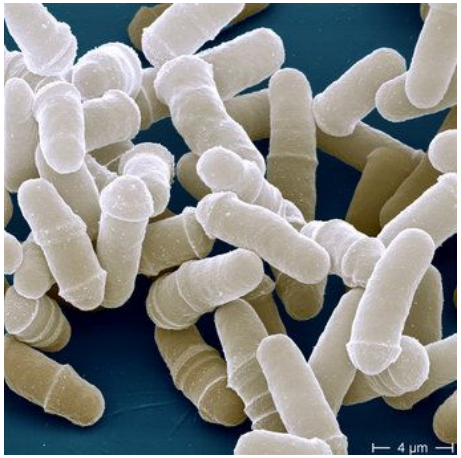
time	weight	gas_flow	pressure	laser	angle	angle_vel
f64	f64	f64	f64	f64	f64	f64
0.0	-0.253109	0.0	-0.968333	67.863029	0.036505	0.0
0.005	-0.253109	0.0	-0.968333	67.863029	0.036505	0.0
0.01	-0.253109	0.0	-1.4229	67.863029	0.036505	0.0
0.015	-0.253109	0.0	-1.486328	67.863029	0.036505	0.0
0.02	-0.253109	0.0	-1.528613	67.863029	0.036505	0.0

Bonus Dataset for Pandas: RNA tails

Real dataset from my work.

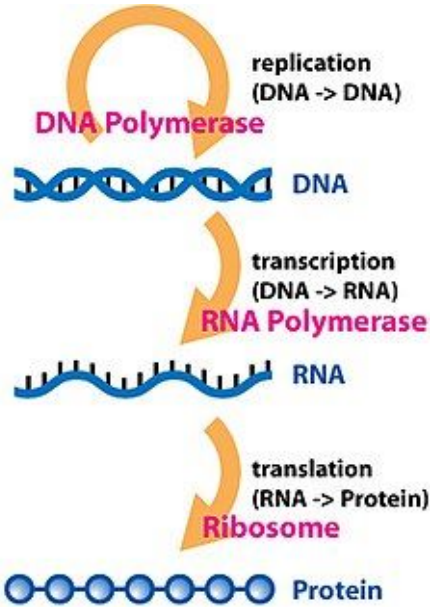
More details about ai after notebook 1.

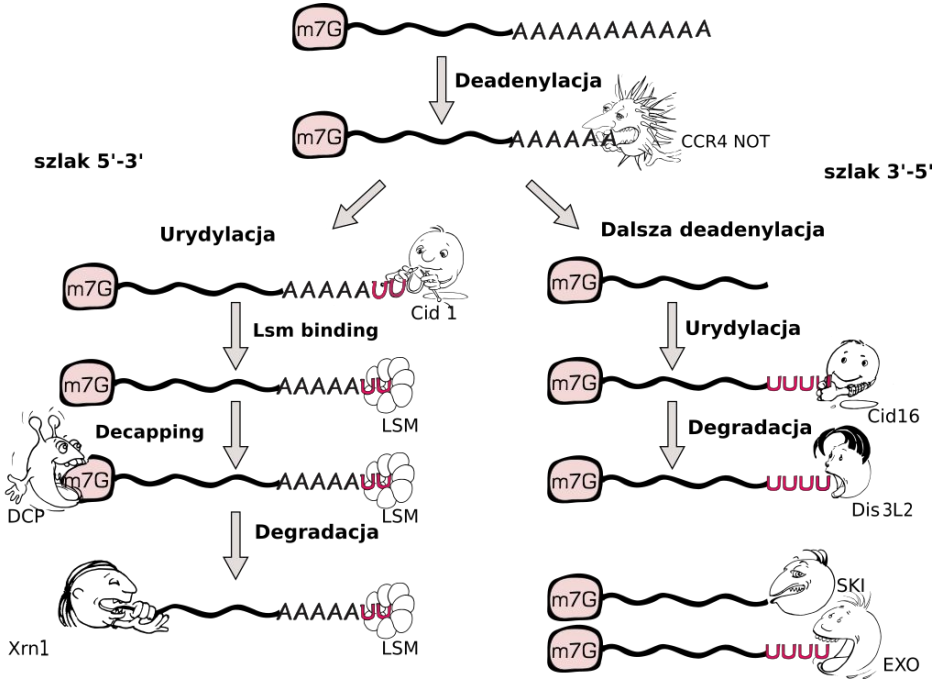
Let's start with "Intro_Pandas"



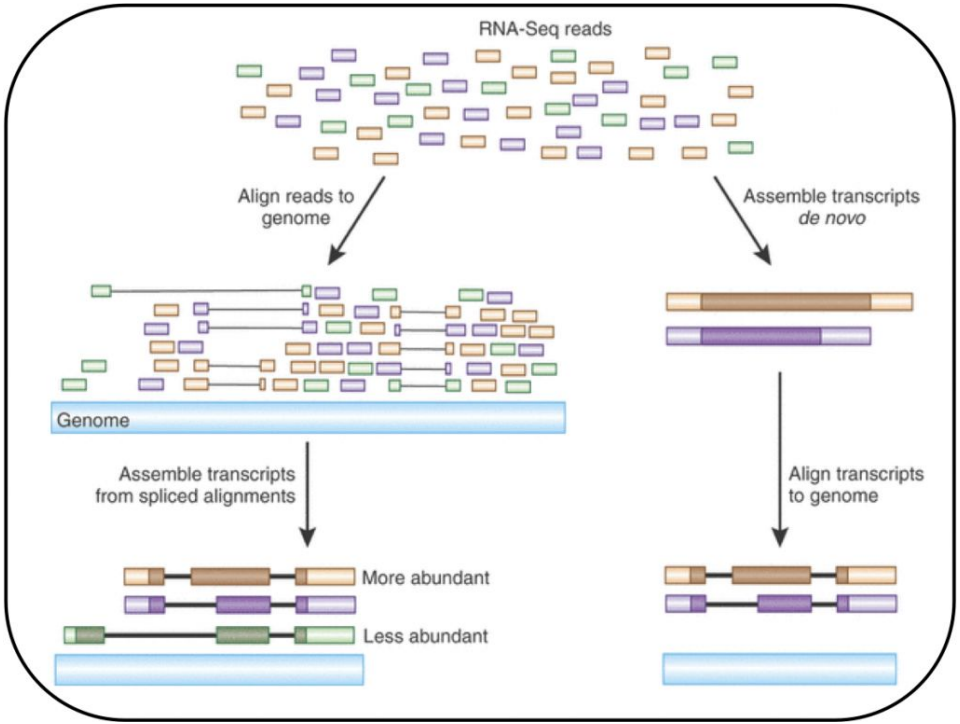
	read_ID	chr	start_R1	stop_R1	strand_R1	gene_start	gene_stop	gene	coord_R2	cigar	
0	A01330:100:HJ2G7DRX2:1:1101:10059:25551:TTGAGT	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844743	14S76M	A
1	A01330:100:HJ2G7DRX2:1:1101:11080:2268:CGTCCC	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844745	13S77M	A
2	A01330:100:HJ2G7DRX2:1:1101:11442:27007:TGGTAG	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844798	23S65M2S	
3	A01330:100:HJ2G7DRX2:1:1101:12138:28588:CGGAAA	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844743	21S69M	A
4	A01330:100:HJ2G7DRX2:1:1101:12156:29215:ATTCCC	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844803	16S74M	
5	A01330:100:HJ2G7DRX2:1:1101:12165:29293:TCAACA	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844779	15S75M	A
6	A01330:100:HJ2G7DRX2:1:1101:12319:21605:GATGAT	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844784	17S71M2S	T
7	A01330:100:HJ2G7DRX2:1:1101:12472:28604:CGGAAA	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844743	21S69M	A
8	A01330:100:HJ2G7DRX2:1:1101:1298:28917:CCTTTT	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	0	*	
9	A01330:100:HJ2G7DRX2:1:1101:13376:30827:GCTATT	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	0	*	

Dataset 2: RNA tails

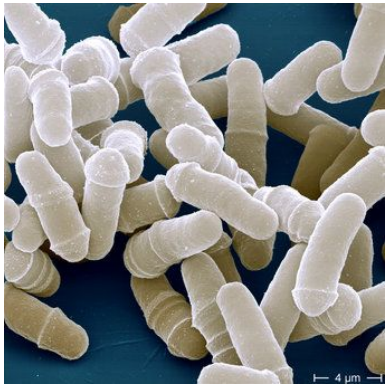




Dataset 2: RNA tails



Dataset 2: RNA tails



20 columns:

'read_ID', 'chr', 'start_R1', 'stop_R1', 'strand_R1', 'gene_start', 'gene_stop', 'gene', 'coord_R2', 'cigar', 'seq_R2', 'RNA_type', 'tail_fromcigar', 'tail_LENcigar', 'tail_fromGREP', 'tail_from', 'tail_type', 'stop_R2', 'distance_to_TES', 'tail'

	read_ID	chr	start_R1	stop_R1	strand_R1	gene_start	gene_stop	gene	coord_R2	cigar	
0	A01330:100:HJ2G7DRX2:1:1101:10059:25551:TTGAGT	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844743	14S76M	A
1	A01330:100:HJ2G7DRX2:1:1101:11080:2268:CGTCCC	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844745	13S77M	A
2	A01330:100:HJ2G7DRX2:1:1101:11442:27007:TGGTAG	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844798	23S65M2S	
3	A01330:100:HJ2G7DRX2:1:1101:12138:28588:CGGAAA	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844743	21S69M	A
4	A01330:100:HJ2G7DRX2:1:1101:12156:29215:ATTCCC	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844803	16S74M	
5	A01330:100:HJ2G7DRX2:1:1101:12165:29293:TCAACA	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844779	15S75M	A
6	A01330:100:HJ2G7DRX2:1:1101:12319:21605:GATGAT	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844784	17S71M2S	T
7	A01330:100:HJ2G7DRX2:1:1101:12472:28604:CGGAAA	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	3844743	21S69M	A
8	A01330:100:HJ2G7DRX2:1:1101:1298:28917:CCTTTT	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	0	*	
9	A01330:100:HJ2G7DRX2:1:1101:13376:30827:GCTATT	I	3845085	3845186	-	3844742	3846645	SPAC4H3.10c	0	*	