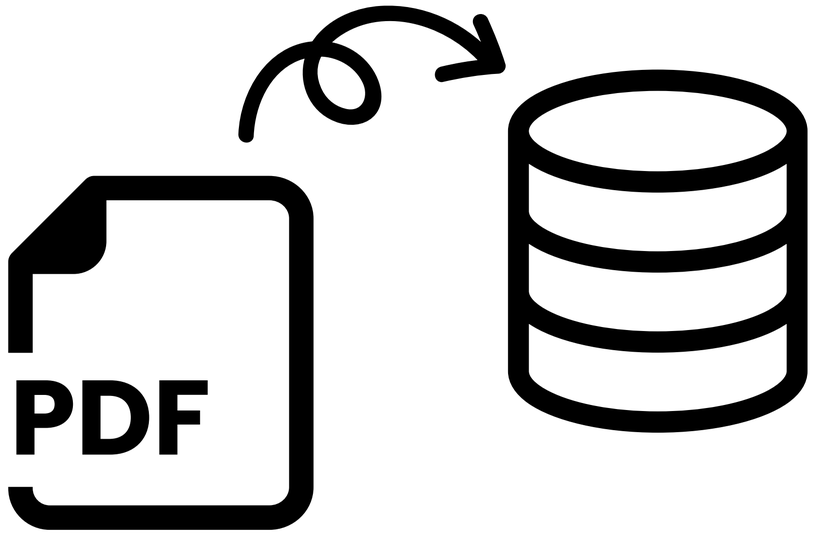
Web Scraping Data from Online PDF Files

Creation of a database of middle distance and distance performances from World Athletics Championships

GitHub repository



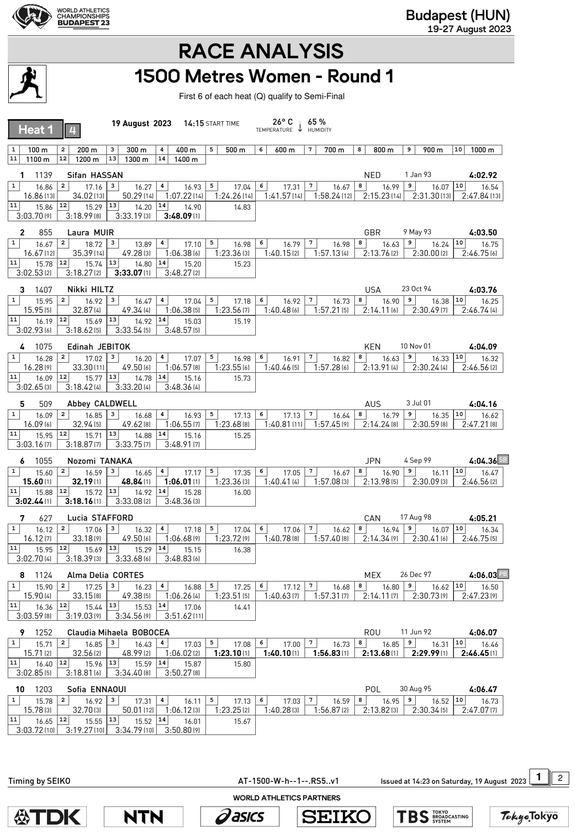
Baptiste Gorteau

May 2025

Context & Problematic

Key Points to Understand the Project Goals

- No detailed middle distance and distance data available for analysis.
- Detailed data available but in online PDFs.
- **How to extract the data from the online PDFs to create a database?**



WORLD ATHLETICS CHAMPIONSHIPS BUDAPEST 23

Budapest (HUN)
19-27 August 2023

RACE ANALYSIS

1500 Metres Women - Round 1

First 6 of each heat (Q) qualify to Semi-Final

Heat 1 | 19 August 2023 | 14:15 START TIME | 26° C TEMPERATURE | 65 % HUMIDITY

1	100 m	2	200 m	3	300 m	4	400 m	5	500 m	6	400 m	7	700 m	8	800 m	9	900 m	10	1000 m
11	1100 m	12	1200 m	13	1300 m	14	1400 m												

1 1139 **Sifan HASSAN** NED 1 Jan 93 **4:02.92**

1	16.86	2	17.16	3	16.27	4	16.93	5	17.04	6	17.31	7	16.67	8	16.99	9	16.07	10	16.54
	16.86 (13)		34.02 (13)		50.29 (14)		1:07.22 (14)		1:24.26 (14)		1:41.57 (14)		1:58.24 (12)		2:15.23 (14)		2:31.30 (13)		2:47.84 (13)
11	15.86	12	15.29	13	14.20	14	14.90		14.83										
	3:03.70 (9)		3:18.99 (8)		3:33.19 (3)		**3:48.09** (1)												

2 855 **Laura MUIR** GBR 9 May 93 **4:03.50**

1	16.67	2	18.72	3	13.89	4	17.10	5	16.98	6	16.79	7	16.98	8	16.63	9	16.24	10	16.75
	16.67 (12)		35.39 (14)		49.28 (3)		1:06.38 (6)		1:23.36 (3)		1:40.15 (2)		1:57.13 (4)		2:13.76 (2)		2:30.00 (2)		2:46.75 (6)
11	15.78	12	15.74	13	14.80	14	15.20		15.23										
	3:02.53 (2)		3:18.27 (2)		**3:33.07** (1)		3:48.27 (2)												

3 1407 **Nikki HILTZ** USA 23 Oct 94 **4:03.76**

1	15.95	2	16.92	3	16.47	4	17.04	5	17.18	6	16.92	7	16.73	8	16.90	9	16.38	10	16.25
	15.95 (5)		32.87 (4)		49.34 (4)		1:06.38 (6)		1:23.56 (7)		1:40.48 (6)		1:57.21 (5)		2:14.11 (6)		2:30.49 (7)		2:46.74 (4)
11	16.19	12	15.49	13	14.92	14	15.03		15.19										
	3:02.93 (6)		3:18.42 (6)		3:33.54 (5)		3:48.57 (5)												

4 1075 **Edinah JEBITOK** KEN 10 Nov 01 **4:04.09**

1	16.28	2	17.02	3	16.20	4	17.07	5	16.98	6	16.91	7	16.82	8	16.63	9	16.33	10	16.32
	16.28 (9)		33.30 (11)		49.50 (6)		1:06.57 (8)		1:23.55 (6)		1:40.46 (5)		1:57.28 (6)		2:13.91 (6)		2:30.24 (6)		2:46.56 (2)
11	16.09	12	15.77	13	14.78	14	15.16		15.73										
	3:02.65 (3)		3:18.42 (4)		3:33.20 (4)		3:48.36 (4)												

5 509 **Abbey CALDWELL** AUS 3 Jul 01 **4:04.16**

1	16.09	2	16.85	3	16.68	4	16.93	5	17.13	6	17.13	7	16.64	8	16.79	9	16.35	10	16.62
	16.09 (6)		32.94 (5)		49.62 (8)		1:06.55 (7)		1:23.68 (8)		1:40.81 (11)		1:57.45 (9)		2:14.24 (8)		2:30.59 (8)		2:47.21 (8)
11	15.95	12	15.71	13	14.88	14	15.16		15.25										
	3:03.16 (7)		3:18.87 (7)		3:33.75 (7)		3:48.91 (7)												

6 1055 **Nozomi TANAKA** JPN 4 Sep 99 **4:04.36**

1	15.60	2	16.59	3	16.65	4	17.17	5	17.35	6	17.05	7	16.67	8	16.90	9	16.11	10	16.47
	15.60 (1)		**32.19** (1)		**48.84** (1)		**1:06.01** (1)		1:23.36 (3)		1:40.41 (4)		1:57.08 (3)		2:13.98 (5)		2:30.09 (3)		2:46.56 (2)
11	15.88	12	15.72	13	14.92	14	15.28		16.00										
	3:02.44 (1)		**3:18.16** (1)		3:33.08 (2)		3:48.36 (3)												

7 627 **Lucia STAFFORD** CAN 17 Aug 98 **4:05.21**

1	16.12	2	17.06	3	16.32	4	17.18	5	17.04	6	17.06	7	16.62	8	16.94	9	16.07	10	16.34
	16.12 (7)		33.18 (9)		49.50 (6)		1:06.68 (9)		1:23.72 (9)		1:40.78 (8)		1:57.40 (8)		2:14.34 (9)		2:30.41 (6)		2:46.75 (5)
11	15.95	12	15.69	13	15.29	14	15.15		16.38										
	3:02.70 (4)		3:18.39 (3)		3:33.68 (6)		3:48.83 (6)												

8 1124 **Alma Delia CORTES** MEX 26 Dec 97 **4:06.03**

1	15.90	2	17.25	3	16.23	4	16.88	5	17.25	6	17.12	7	16.68	8	16.80	9	16.62	10	16.50
	15.90 (4)		33.15 (8)		49.38 (5)		1:06.26 (4)		1:23.51 (5)		1:40.63 (7)		1:57.31 (7)		2:14.11 (7)		2:30.73 (9)		2:47.23 (9)
11	16.36	12	15.44	13	15.53	14	17.06		14.41										
	3:03.59 (8)		3:19.03 (9)		3:34.56 (9)		3:51.62 (11)												

9 1252 **Claudia Mihaela BOBOCEA** ROU 11 Jun 92 **4:06.07**

1	15.71	2	16.85	3	16.43	4	17.03	5	17.08	6	17.00	7	16.73	8	16.85	9	16.31	10	16.44
	15.71 (2)		32.56 (2)		48.99 (2)		1:06.02 (2)		**1:23.10** (1)		**1:40.10** (1)		**1:56.83** (1)		**2:13.68** (1)		**2:29.99** (1)		**2:46.45** (1)
11	16.40	12	15.96	13	15.59	14	15.87		15.80										
	3:02.85 (5)		3:18.81 (6)		3:34.40 (8)		3:50.27 (8)												

10 1203 **Sofia ENNAOUI** POL 30 Aug 95 **4:06.47**

1	15.78	2	16.92	3	17.31	4	16.11	5	17.13	6	17.03	7	16.59	8	16.95	9	16.52	10	16.73
	15.78 (3)		32.70 (3)		50.01 (12)		1:06.12 (3)		1:23.25 (2)		1:40.28 (3)		1:56.87 (2)		2:13.82 (3)		2:30.34 (5)		2:47.07 (7)
11	16.65	12	15.55	13	15.52	14	16.01		15.67										
	3:03.72 (10)		3:19.27 (10)		3:34.79 (10)		3:50.80 (9)												

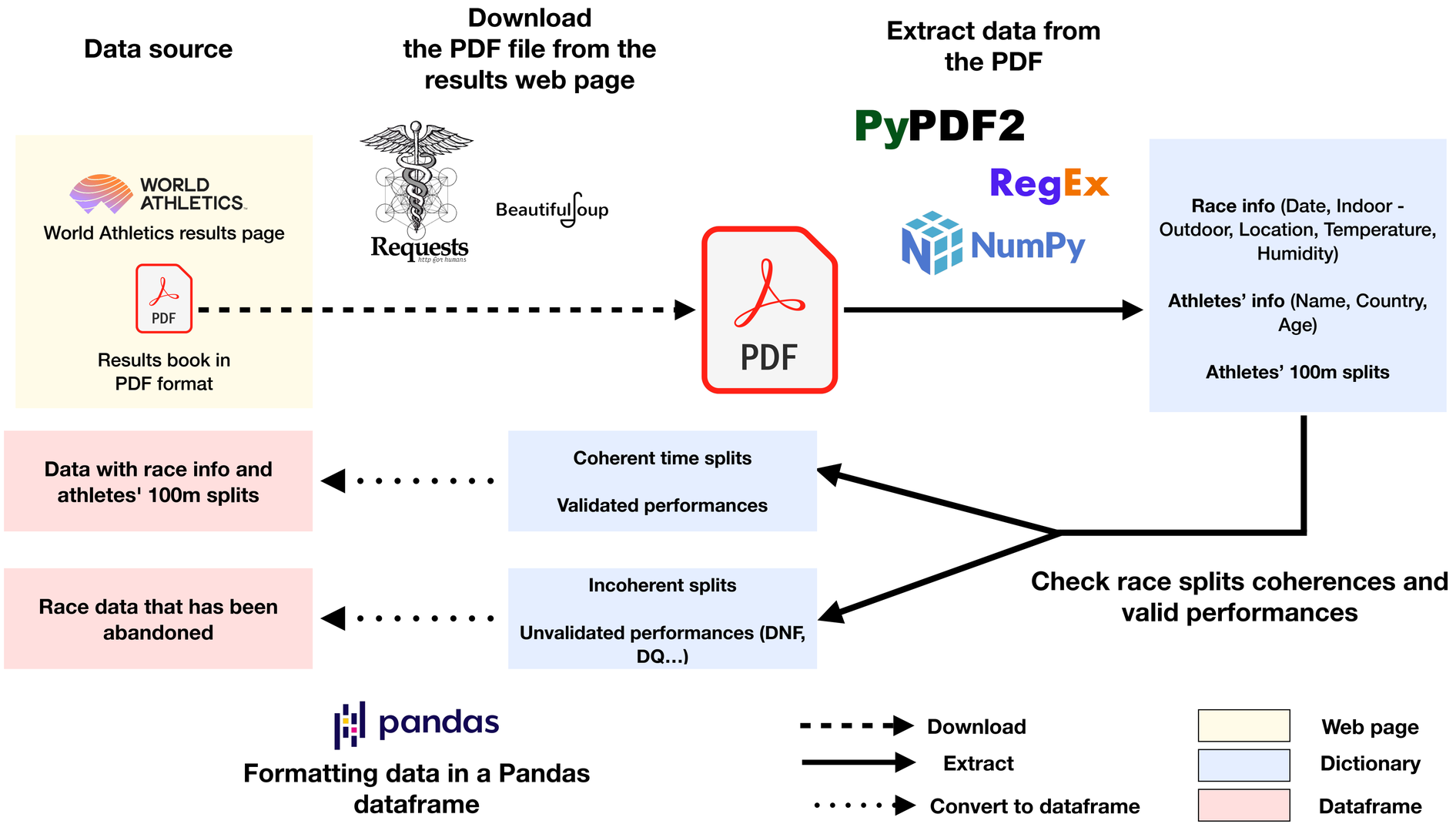
Timing by SEIKO | AT-1500-W-h--1--.RS5..v1 | Issued at 14:23 on Saturday, 19 August 2023 | 1 | 2

WORLD ATHLETICS PARTNERS

TDK NTN asics SEIKO TBS TOKYO BROADCASTING SYSTEM TokyoTokyo

Example of detailed World Athletics race results in PDF.

Project architecture



Dataframes Structure

Race Data Example

- **One dataframe per event** (800m, 1500m, 3000m, 3000m steeplechase, 5000m and 10000m)
- **Race information** (Indoor or Outdoor, Location, Date, Temperature, Humidity, Men or Women, Stage)
- **Athlete information** (Name, Country, Age)
- **Time splits at each 100m of the race**

COMP_TYPE	COMP_LOC	YEAR	RACE_DATE	RACE_TEMP	RACE_HUMID	M_W	EVENT	STAGE	RACE_NB	ATHLETE_NAME	ATHLETE_COUNTRY	ATHLETE_AGE	100M	200M	300M	400M	500M	600M	700M	800M	900M	1000M	1100M	1200M	1300M	1400M	1500M
outdoor	Budapest (HUN)	2023	19 August 2023	26	65	women	1500-metres	heats	1.0	Sifan HASSAN	NED	31.0	16.86	34.02	50.29	1:07.22	1:24.26	1:41.57	1:58.24	2:15.23	2:31.30	2:47.84	3:03.70	3:18.99	3:33.19	3:48.09	4:02.92
outdoor	Budapest (HUN)	2023	19 August 2023	26	65	women	1500-metres	heats	1.0	Laura MUIR	GBR	30.0	16.67	35.39	49.28	1:06.38	1:23.36	1:40.15	1:57.13	2:13.76	2:30.00	2:46.75	3:02.53	3:18.27	3:33.07	3:48.27	4:03.50
outdoor	Budapest (HUN)	2023	19 August 2023	26	65	women	1500-metres	heats	1.0	Nikki HILTZ	USA	29.0	15.95	32.87	49.34	1:06.38	1:23.56	1:40.48	1:57.21	2:14.11	2:30.49	2:46.74	3:02.93	3:18.62	3:33.54	3:48.57	4:03.76
outdoor	Budapest (HUN)	2023	19 August 2023	26	65	women	1500-metres	heats	1.0	Edinah JEBITOK	KEN	22.0	16.28	33.30	49.50	1:06.57	1:23.55	1:40.46	1:57.28	2:13.91	2:30.24	2:46.56	3:02.65	3:18.42	3:33.20	3:48.36	4:04.09
outdoor	Budapest (HUN)	2023	19 August 2023	26	65	women	1500-metres	heats	1.0	Abbey CALDWELL	AUS	22.0	16.09	32.94	49.62	1:06.55	1:23.68	1:40.81	1:57.45	2:14.24	2:30.59	2:47.21	3:03.16	3:18.87	3:33.75	3:48.91	4:04.16
outdoor	Budapest (HUN)	2023	19 August 2023	26	65	women	1500-metres	heats	1.0	Nozomi TANAKA	JPN	24.0	15.60	32.19	48.84	1:06.01	1:23.36	1:40.41	1:57.08	2:13.98	2:30.09	2:46.56	3:02.44	3:18.16	3:33.08	3:48.36	4:04.36
outdoor	Budapest (HUN)	2023	19 August 2023	26	65	women	1500-metres	heats	1.0	Lucia STAFFORD	CAN	25.0	16.12	33.18	49.50	1:06.68	1:23.72	1:40.78	1:57.40	2:14.34	2:30.41	2:46.75	3:02.70	3:18.39	3:33.68	3:48.83	4:05.21

Dataframes Structure

Abandoned Data Example

- **One dataframe for all abandoned performances**
- **Race information** (Location, Date, Temperature, Humidity, Men or Women, Stage)
- **Athlete information** (Name, Country, Age)
- **Reason for abandoning performance** (Incoherent splits, Did not finish, Missing time splits, Disqualified)

COMP_TYPE	COMP_LOC	YEAR	RACE_DATE	RACE_TEMP	RACE_HUMID	M_W	EVENT	STAGE	RACE_NB	ATHLETE_NAME	ATHLETE_COUNTRY	ATHLETE_AGE	REASON
outdoor	Oregon (USA)	2022	15 July 2022	28	35	men	3000-metres-steeplechase	heats	3.0	Jean-Simon DESGAGNÉS	CAN	24.0	Incoherent splits
outdoor	Oregon (USA)	2022	15 July 2022	28	35	men	3000-metres-steeplechase	heats	3.0	Mohamed TINDOUFT	MAR	29.0	Did not finish
outdoor	Oregon (USA)	2022	15 July 2022	28	41	women	1500-metres	heats	1.0	Hirut MESHESHA	ETH	21.0	Missing time split

Race Data Analysis

Figures

- Comparison of indoor and outdoor 1500m and 800m races for men and women.
- Using spline bases to smooth speed curves.
- Higher speed when running outdoor than indoor.
- Peak speed 300m from the finish on a 1500m and only after 100m on an 800m.

Evolution of average speed over 800m and 1500m races during the world championships

World Athletics Championships races data from 2019 to 2025.

Source: World Athetics data (worldathletics.org/competition/calendar-results)

Baptiste Gorteau - bgorteau.github.io