

CoClean

Collaborative Data Cleaning

Mashaal Musleh University of Minnesota*

Mourad Ouzzani, Nan Tang HBKU, Qatar Computing Research Institute

AnHai Doan, University of Wisconsin

* work done while doing an internship at QCRI



UNIVERSITY OF MINNESOTA



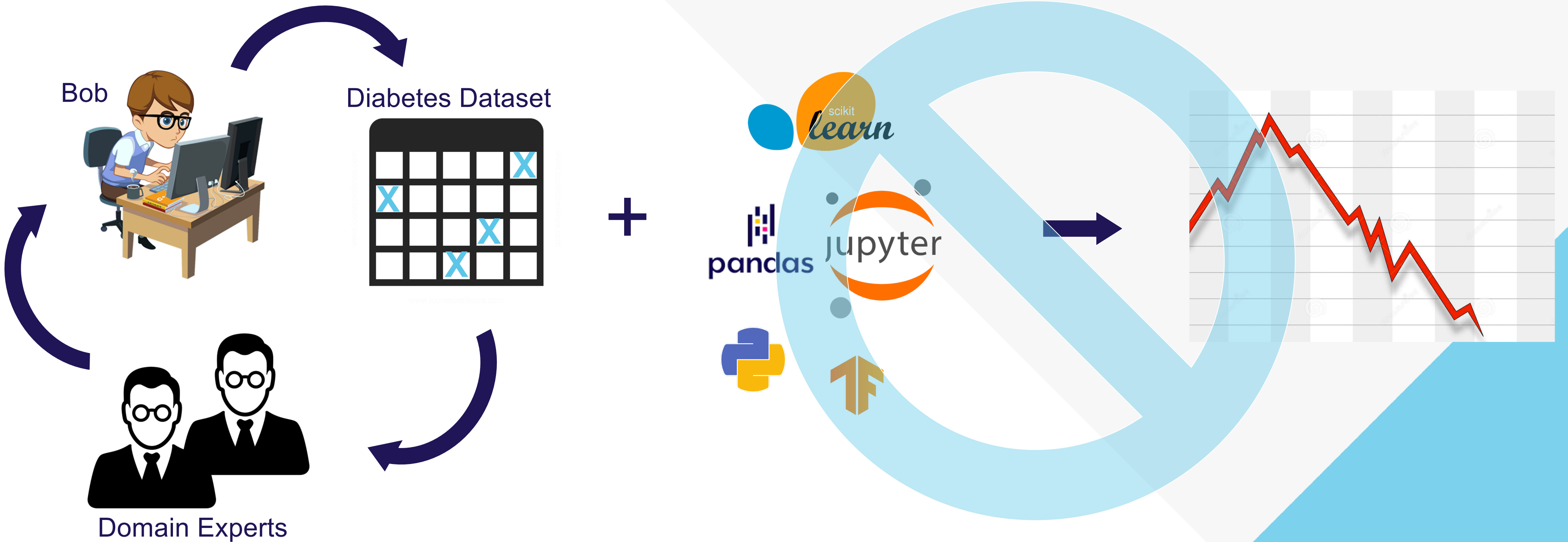
WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY



The Problem



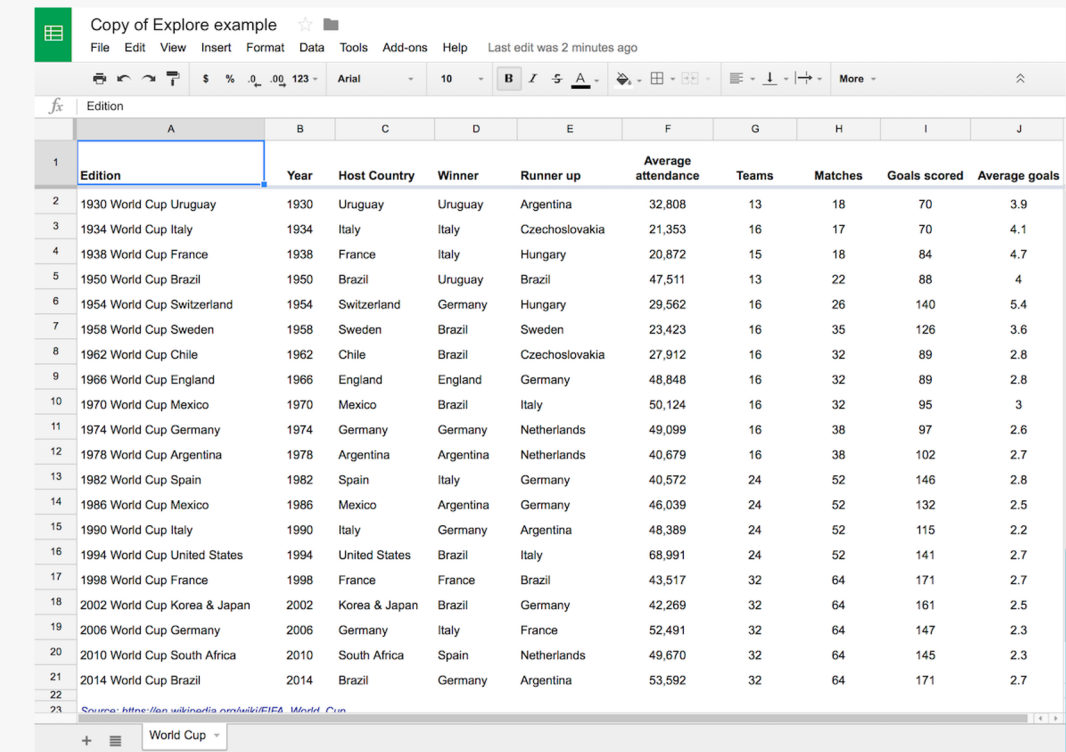
This is common and many data scientists struggle to collaborate with domain experts who are familiar with the data.

Existing Solutions

Issues:

- High Human Cost
 - Multiple Upload/Download
 - Stop/Wait
- Single User (OpenRefine)
- Users overwrite each other (google sheets)
- Not integrated with data science environments used in practice

Google Sheets Style



Copy of Explore example

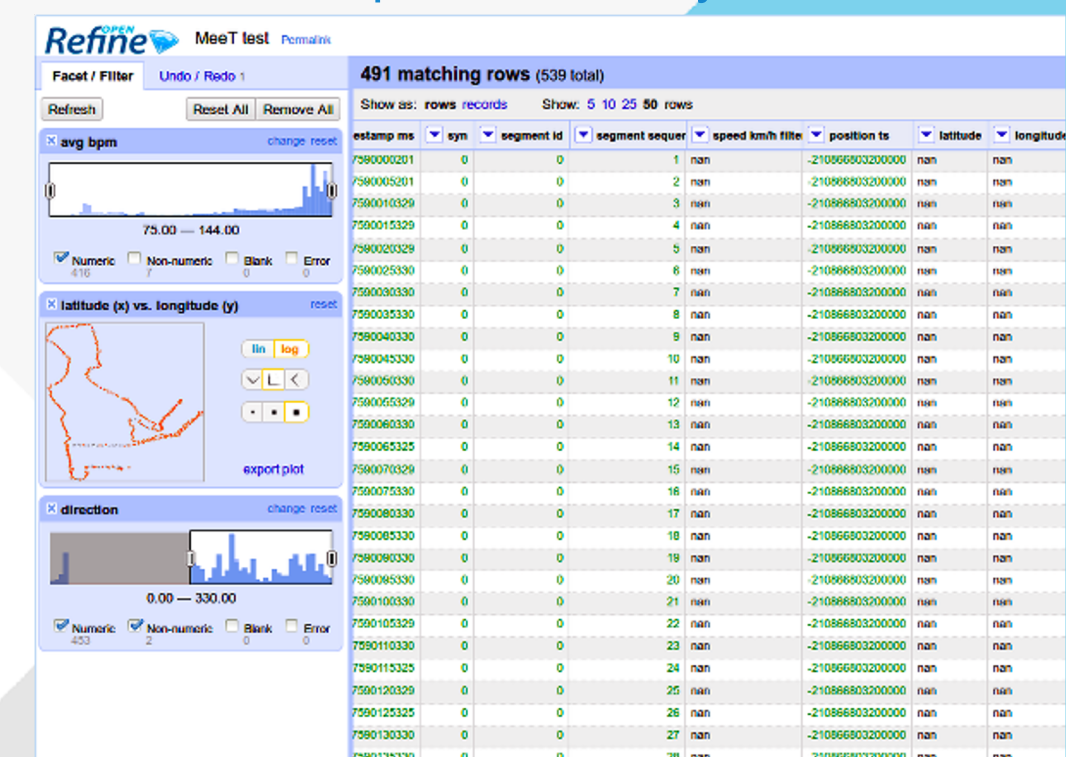
File Edit View Insert Format Data Tools Add-ons Help Last edit was 2 minutes ago

123 Arial 10 B I U A

	A	B	C	D	E	F	G	H	I	J
1	Edition	Year	Host Country	Winner	Runner up	Average attendance	Teams	Matches	Goals scored	Average goals
2	1930 World Cup Uruguay	1930	Uruguay	Uruguay	Argentina	32,808	13	18	70	3.9
3	1934 World Cup Italy	1934	Italy	Italy	Czechoslovakia	21,353	16	17	70	4.1
4	1938 World Cup France	1938	France	Italy	Hungary	20,872	15	18	84	4.7
5	1950 World Cup Brazil	1950	Brazil	Uruguay	Brazil	47,511	13	22	88	4
6	1954 World Cup Switzerland	1954	Switzerland	Germany	Hungary	29,562	16	26	140	5.4
7	1958 World Cup Sweden	1958	Sweden	Brazil	Sweden	23,423	16	35	126	3.6
8	1962 World Cup Chile	1962	Chile	Brazil	Czechoslovakia	27,912	16	32	89	2.8
9	1966 World Cup England	1966	England	England	Germany	48,848	16	32	89	2.8
10	1970 World Cup Mexico	1970	Mexico	Brazil	Italy	50,124	16	32	95	3
11	1974 World Cup Germany	1974	Germany	Germany	Netherlands	49,099	16	38	97	2.6
12	1978 World Cup Argentina	1978	Argentina	Argentina	Netherlands	40,679	16	38	102	2.7
13	1982 World Cup Spain	1982	Spain	Italy	Germany	40,572	24	52	146	2.8
14	1986 World Cup Mexico	1986	Mexico	Argentina	Germany	46,039	24	52	132	2.5
15	1990 World Cup Italy	1990	Italy	Germany	Argentina	48,389	24	52	115	2.2
16	1994 World Cup United States	1994	United States	Brazil	Italy	68,991	24	52	141	2.7
17	1998 World Cup France	1998	France	Brazil	Brazil	43,517	32	64	171	2.7
18	2002 World Cup Korea & Japan	2002	Korea & Japan	Brazil	Germany	42,269	32	64	161	2.5
19	2006 World Cup Germany	2006	Germany	Italy	France	52,491	32	64	147	2.3
20	2010 World Cup South Africa	2010	South Africa	Spain	Netherlands	49,670	32	64	145	2.3
21	2014 World Cup Brazil	2014	Brazil	Germany	Argentina	53,592	32	64	171	2.7
22										
23										

World Cup

Open Refine Style



Refine MeeT test

Facet / Filter Undo / Redo 1

Refresh Reset All Remove All

491 matching rows (539 total)

Show as: rows records Show: 5 10 25 50 rows

estamp ms syn segment id segment sequer speed km/h title position ts latitude longitude

7590000201	0	0	1	nan	-210866803200000	nan	nan
7590000501	0	0	2	nan	-210866803200000	nan	nan
7590010329	0	0	3	nan	-210866803200000	nan	nan
7590015329	0	0	4	nan	-210866803200000	nan	nan
7590020329	0	0	5	nan	-210866803200000	nan	nan
7590025330	0	0	6	nan	-210866803200000	nan	nan
7590030330	0	0	7	nan	-210866803200000	nan	nan
7590035330	0	0	8	nan	-210866803200000	nan	nan
7590040330	0	0	9	nan	-210866803200000	nan	nan
7590045330	0	0	10	nan	-210866803200000	nan	nan
7590050330	0	0	11	nan	-210866803200000	nan	nan
7590055329	0	0	12	nan	-210866803200000	nan	nan
7590060330	0	0	13	nan	-210866803200000	nan	nan
7590065325	0	0	14	nan	-210866803200000	nan	nan
7590070329	0	0	15	nan	-210866803200000	nan	nan
7590075330	0	0	16	nan	-210866803200000	nan	nan
7590080330	0	0	17	nan	-210866803200000	nan	nan
7590085330	0	0	18	nan	-210866803200000	nan	nan
7590090330	0	0	19	nan	-210866803200000	nan	nan
7590095330	0	0	20	nan	-210866803200000	nan	nan
7590100330	0	0	21	nan	-210866803200000	nan	nan
7590105329	0	0	22	nan	-210866803200000	nan	nan
7590110330	0	0	23	nan	-210866803200000	nan	nan
7590115325	0	0	24	nan	-210866803200000	nan	nan
7590120329	0	0	25	nan	-210866803200000	nan	nan
7590125325	0	0	26	nan	-210866803200000	nan	nan
7590130330	0	0	27	nan	-210866803200000	nan	nan
7590135330	0	0	28	nan	-210866803200000	nan	nan

Others....

What is the ideal solution?

Ideal Solution

- **Multiple Users**
 - Multiple feedback/ability to consolidate
 - Power/Lay users
- **Integrated with Existing data science workflows**
 - Python, Pandas, etc.
 - Jupyter Notebooks.
- **Minimal Human Cost**
 - Quick & easy use.
 - No upload/download, wait, etc.

Proposed Solution

CoClean: Collaborative Data Cleaning

Crowd in the loop

- Collaborative Data Frames (CDF)

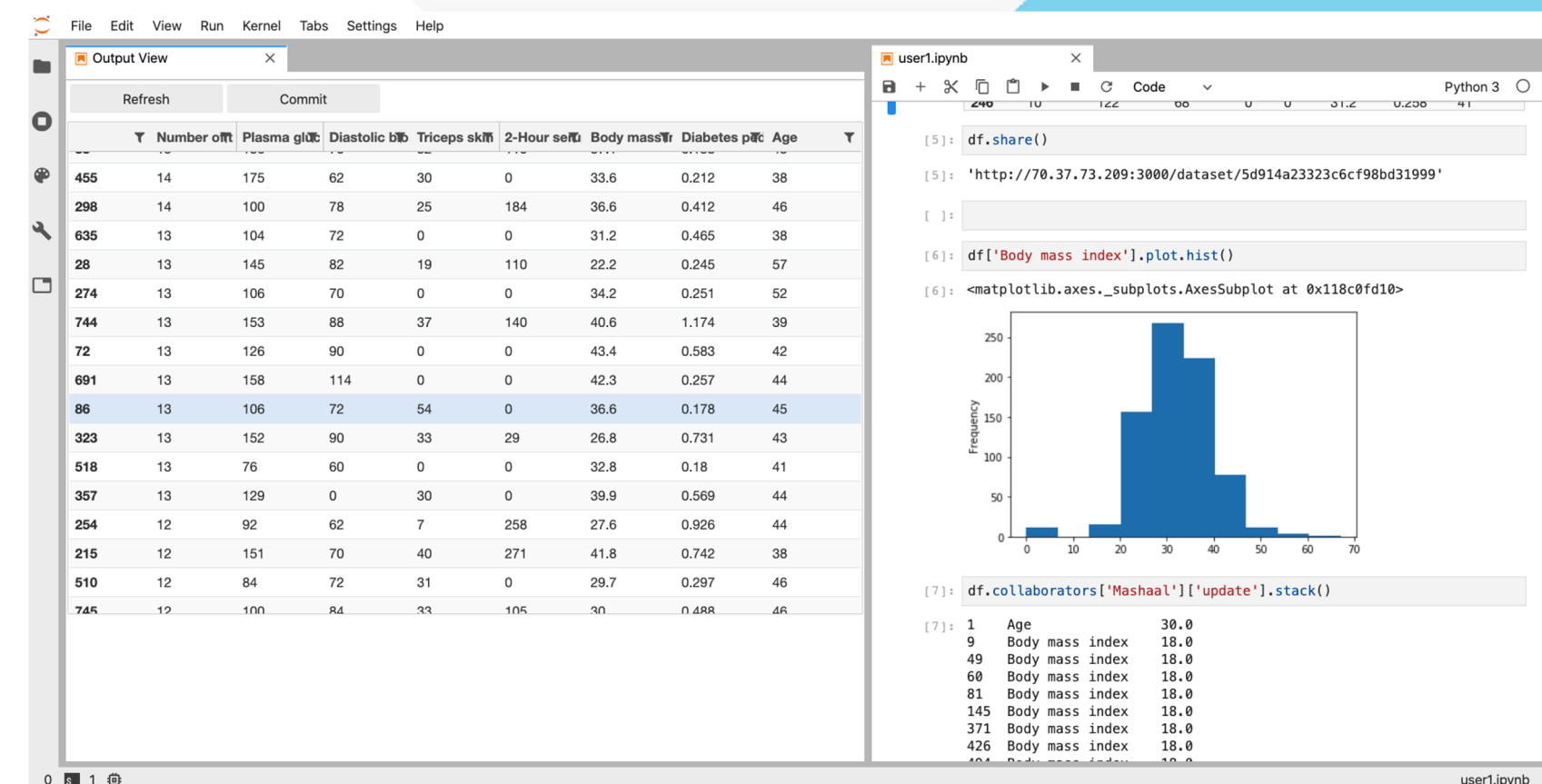
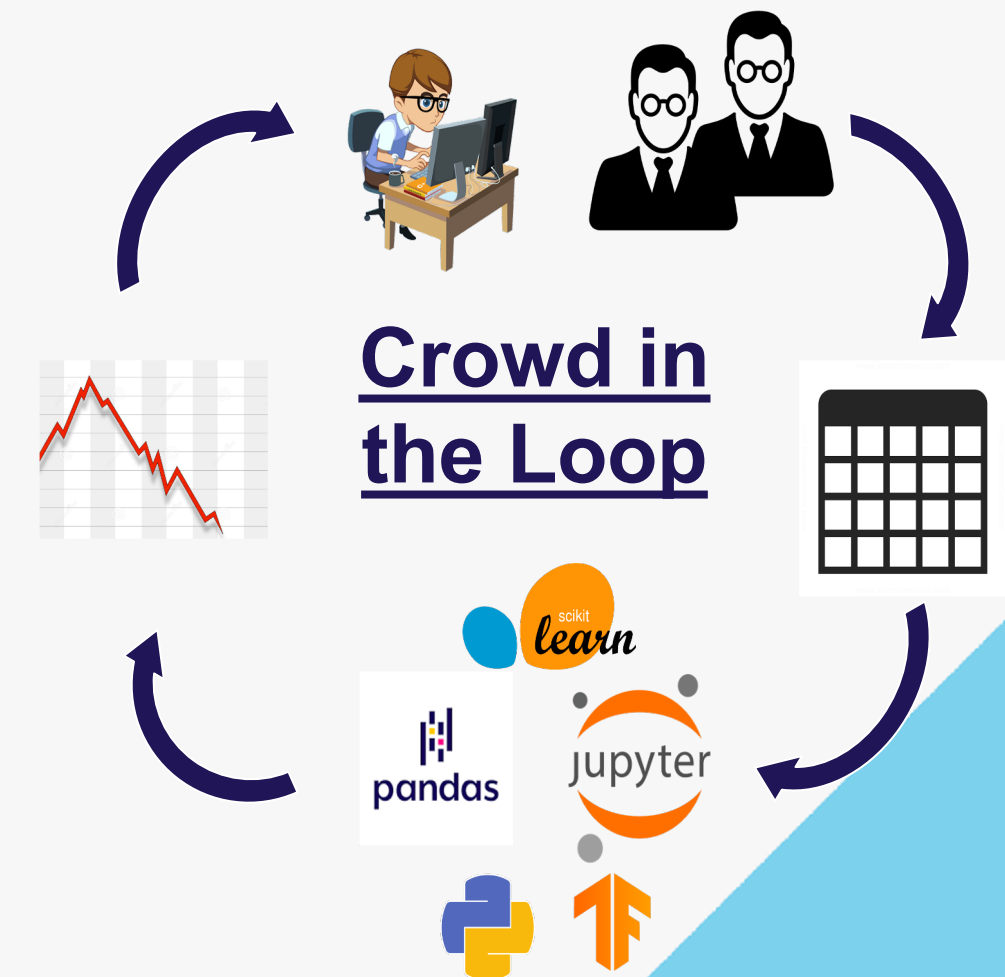
- Compatible with Pandas

- Jupyter Integrated + GUI

- No upload/download, wait, etc.
- Reduce human cost
- Support both power and lay users.

- Side by Side workflow

- More productive



DEMO

- Share data or a subset of data across multiple users, all without stopping the workflow.
- Support both lay and power users.
- Integrate the output of machine algorithms as hints.
- Allow different modes of interactions.

THANK YOU