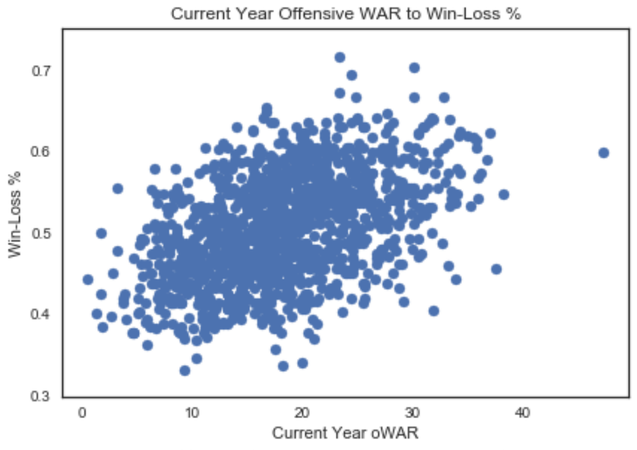
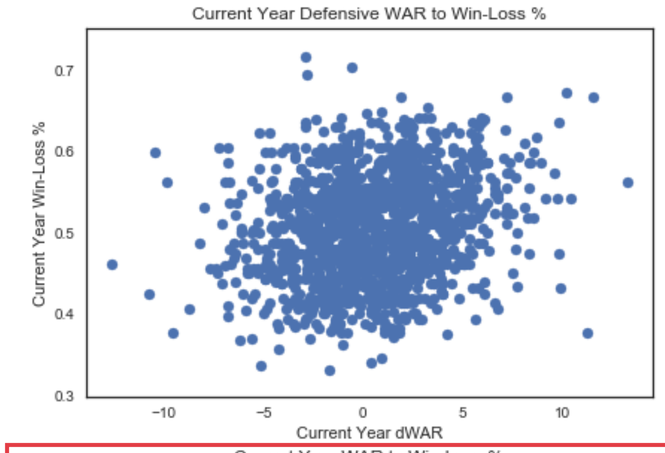
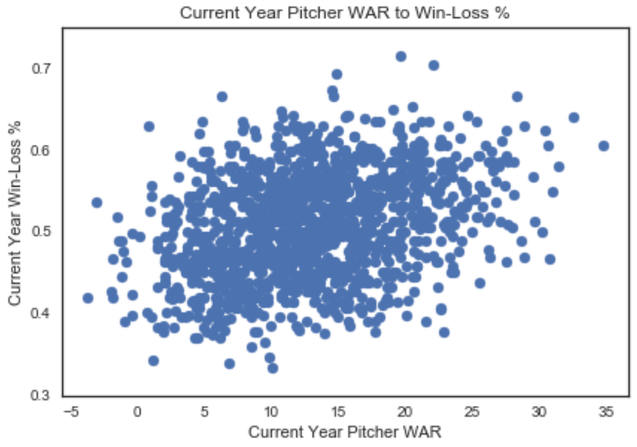
Baseball and Power of Luck

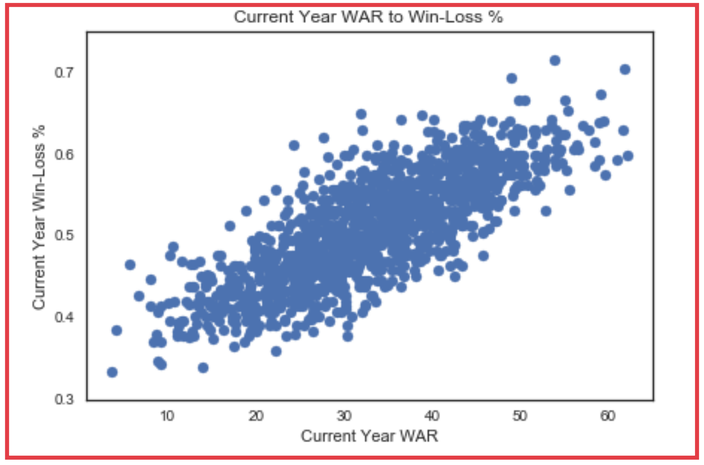
Project Luther - Web Scraping
Will Stokvis
January 26, 2018

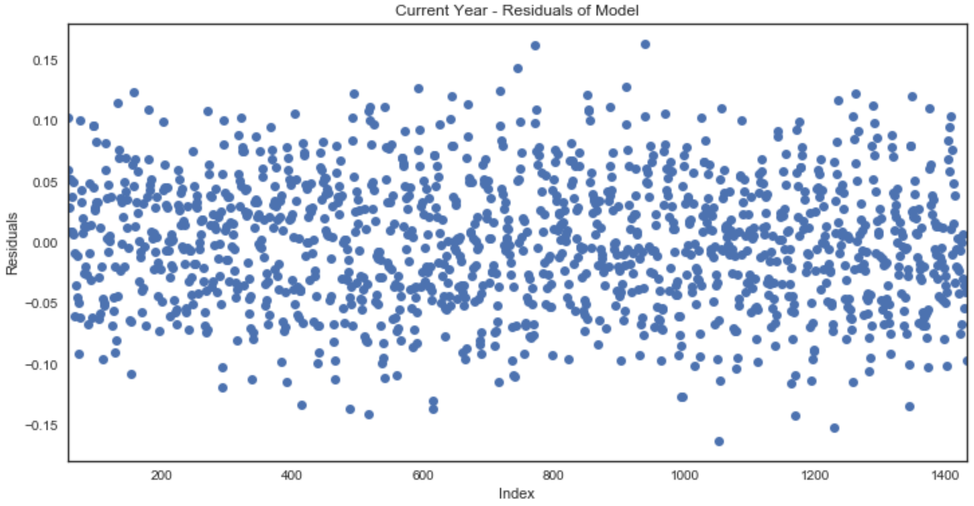
Mapping Current Year Stats against Team Win-Loss %











Objective

Can you predict a team’s Win-Loss % based on the player stats and team performance from the previous year?

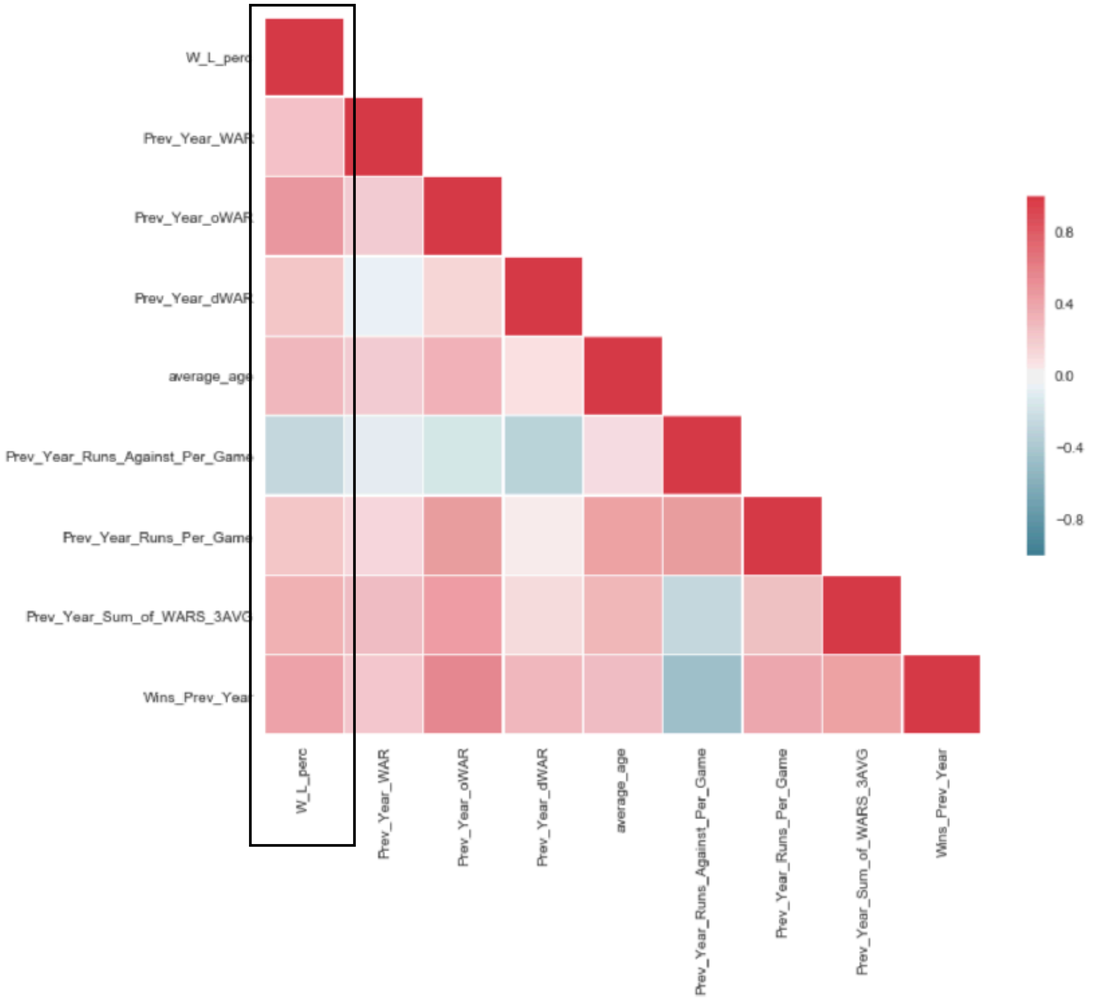
Dataset

- **Time Period** - 1962 onward

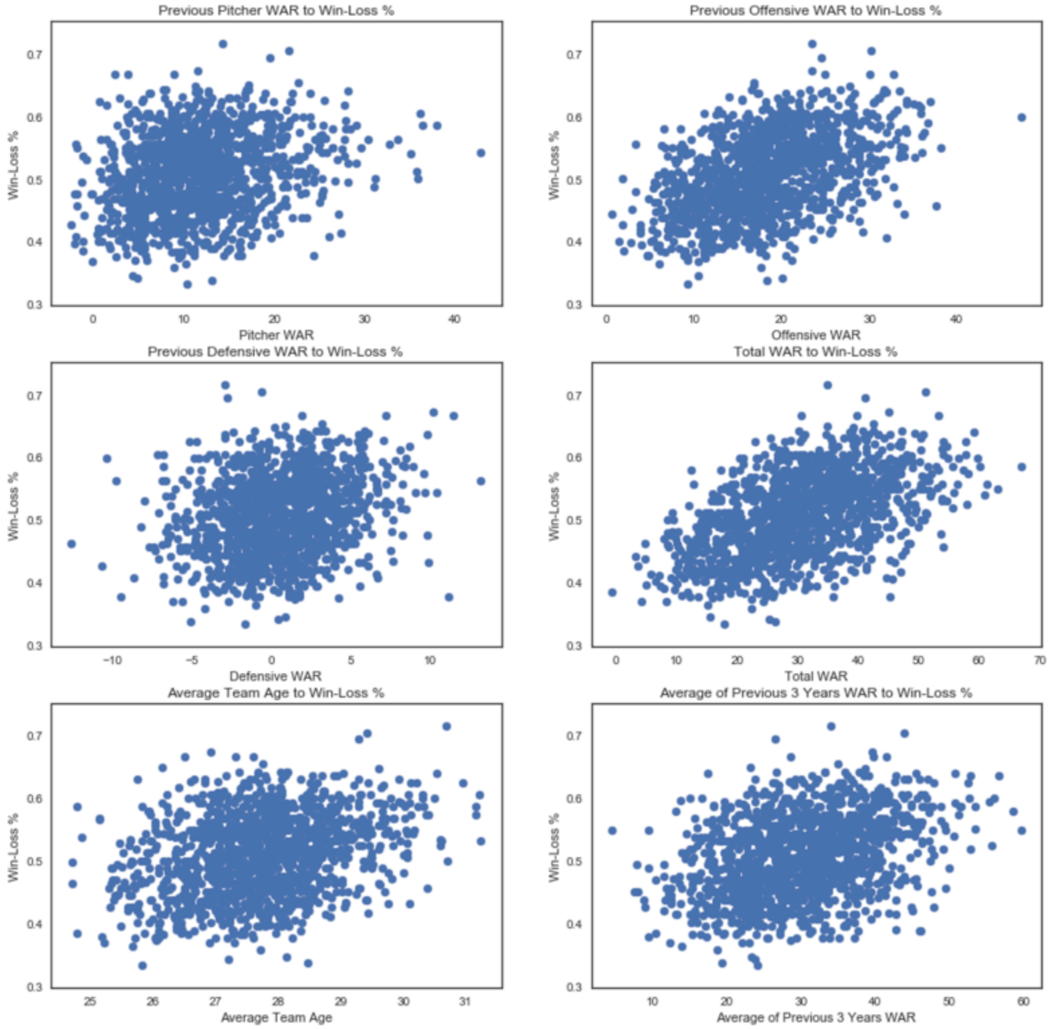
- **Inclusions** - Only players who pitched **or** played in the field
  - Pitchers - 5,240
  - Players - 4,983

- Avoided playoff data as it is essentially a toss up

Correlations



How’s the model perform?



OLS Regression Results

Dep. Variable:	W_L_perc	**R-squared:**	0.330
Model:	OLS	**Adj. R-squared:**	0.326
Method:	Least Squares	**F-statistic:**	92.55
Date:	Thu, 25 Jan 2018	**Prob (F-statistic):**	7.90e-110
Time:	15:31:16	**Log-Likelihood:**	2021.3
No. Observations:	1326	**AIC:**	-4027.
Df Residuals:	1318	**BIC:**	-3985.
Df Model:	7		
Covariance Type:	nonrobust		

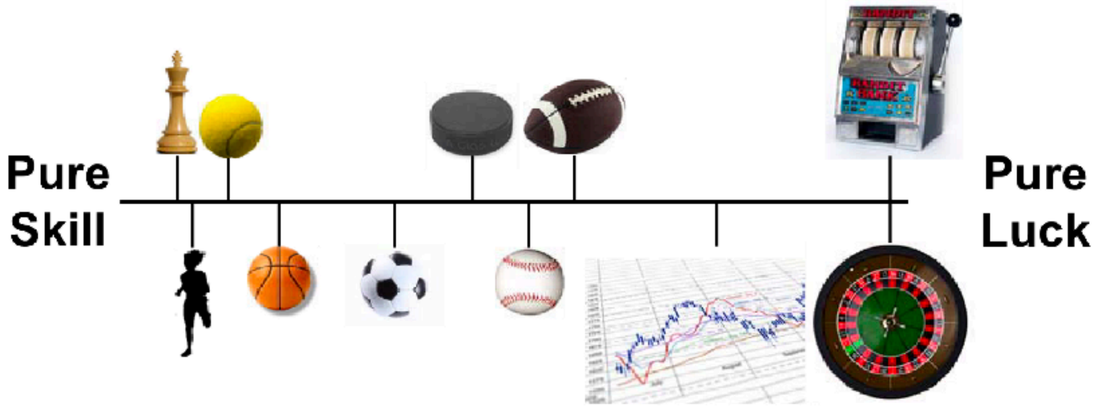
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.2455	0.039	6.240	0.000	0.168	0.323
Prev_Year_WAR	0.0017	0.000	6.306	0.000	0.001	0.002
Prev_Year_oWAR	0.0027	0.000	9.530	0.000	0.002	0.003
Prev_Year_dWAR	0.0024	0.000	4.999	0.000	0.001	0.003
average_age	0.0077	0.002	4.962	0.000	0.005	0.011
Prev_Year_Runs_Against_Per_Game	-0.0178	0.004	-4.428	0.000	-0.026	-0.010
Prev_Year_Runs_Per_Game	0.0090	0.004	2.176	0.030	0.001	0.017
Prev_Year_Sum_of_WARS_3AVG	0.0004	0.000	2.238	0.025	5.43e-05	0.001

Omnibus:	7.275	**Durbin-Watson:**	1.030
Prob(Omnibus):	0.026	**Jarque-Bera (JB):**	5.478
Skew:	0.017	**Prob(JB):**	0.0646
Kurtosis:	2.687	**Cond. No.**	1.32e+03

Conclusion

- Average player age has an unexpectedly high correlation with Win - Loss %

- In terms of correlation to subsequent year performance,
  1. Offense
  2. Pitching
  3. Defense



Next Steps

- Analyze game logs and search for patterns in day to day performance
- Utilize more advanced metrics (e.g. BABIP, HR/FB) that account for the effects of luck

Appendix

Tools Used


Scrapy


BeautifulSoup

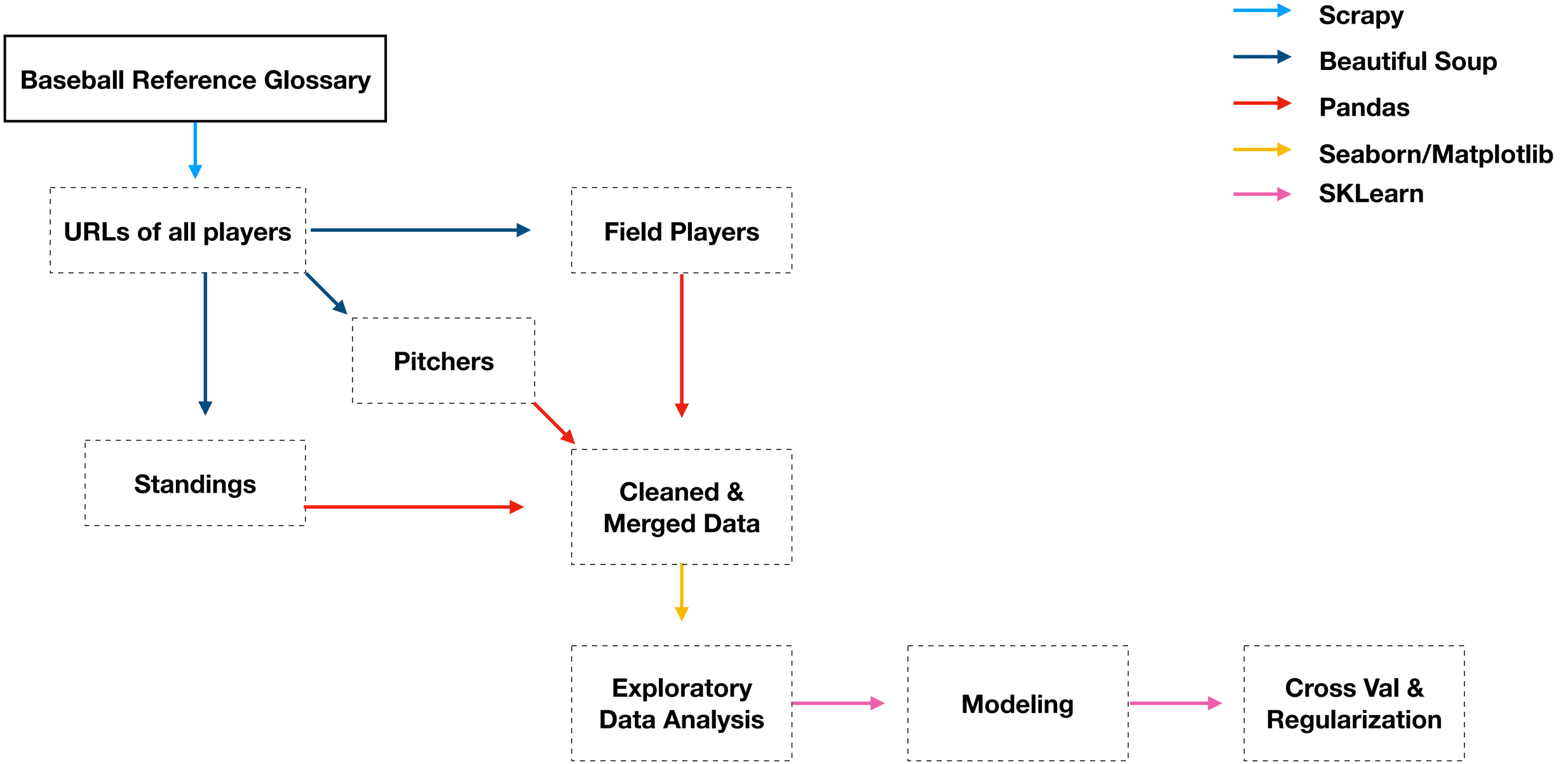

Pandas


NumPy


matplotlib


scikit
learn


SM
StatsModels
Statistics in Python


Baseball Reference Glossary
URLs of all players
Field Players
Pitchers
Standings
Cleaned &
Merged Data
Exploratory
Data Analysis
Modeling
Cross Val &
Regularization
Scrapy
Beautiful Soup
Pandas
Seaborn/Matplotlib
SKLearn