CWS ML Analyst Questionnaire, Joe DiGiovanni

1. <u>For RHHs:</u>

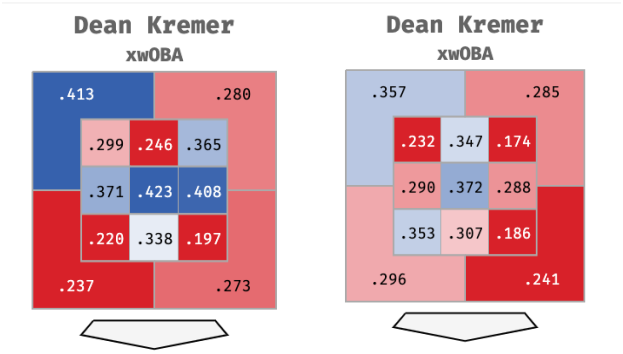
- **Vasil:** Everything will drop more. Sit sinker early. Ignore sweepers away.
- **Eisert:** Low velo. Own inner half, punish middle misses, don't chase the changeup down. Hunt heater up if he falls behind.
- **Schultz:** Force him in zone, avoid backfoot sweeper, hunt firm stuff middle, drive mistakes. Beware changeup away.

<u>For LHPs:</u>

- **Colson:** avoid anything on the inner third, get breaking stuff low and away (lower the better), hard stuff up and away. Do not miss middle away
- **Meidroth:** patient, high contact, low power. Pound the zone. Keep it inside and low, don't miss up, putout low and inside.
- **Braden:** switch hitter with all field power, high strikeouts. Live low and keep the ball out of the air.

2. Lineup (with OPS vs RHP)

<u>Starters</u>	
R Meidroth 2B	.630
L Montgomery SS	.887
L Teel C	.835
L Yastrzemski CF	.809
R Vargas 3B	.707
L Tauchman RF	.752
L Benintendi LF	.709
R Sosa 1B	.723
S Baldwin DH	.731
<u>Bench</u>	
R Mead UT	.679
S Quero C	.602
R Robert CF	.601



vs LHH vs RHH

Anderson was a below average fly ball pitcher last season. I am building a lefty-heavy lineup for Kremer with as much defensive utility as I can get.

Though he has some of the weakest results vs RHP by OPS, Meidroth leads off because he specifically does well avoiding bait up and away the zone, as well as low and away, This is where Kremer likes to target for strikes vs RHH, so he'll work long at-bats and

force Kremer into the zone. I want Montgomery second to ambush Kremer for early damage. Teel is a more complete hitter ands sets up for more damage from my free agent addition, Mike Yastrzemski. I think the Sox would benefit from an everyday lefty CF/RF with pop and a proven glove. He should only cost around $11-12m AAV.

Vargas slots in above Tauchman because he is less susceptible to Kremer's aforementioned putaway pitches. Tauchman moves down from his leadoff position to keep some high OBP in the back half of the lineup. Benintendi has reverse splits but has a better OBP vs righties than both Sosa and Baldwin, who struggle to cover Kremer's main zones.

Mead, Quero, and Robert would sit for if a lefty comes in later in the game. All three players struggle against pitcher's with Kremer's profile.

3. When teams should use an opener

Teams should consider using an opener when their internal models indicate their win probability is increased compared to using a traditional starter. This will typical happen when there is any combination of the following:

- starter matches up poorly against the top half of the opponents' lineup (platoon split, historical performance)
- starter has a notable drop off in performance after the third time through the order
- bullpen has multiple relievers projected perform better against the top half of the order
- high-stakes game (postseason game or playoff hunt)
- team is facing a planned bullpen day with bulk relief options available

Since the beginning of the Statcast era, pitchers have typically seen a jump in opposing OPS of 60+ points between the first and third time through the lineup, partly due to the best hitters in the lineup becoming familiar with their stuff/approach. Openers intend to limit damage by preventing the most dangerous hitters from seeing the starter until he settles in.

The obvious drawback is burning a quality reliever early on, limiting bullpen flexibility in higher leverage later on. However, this assumes the first inning is low-leverage in a seasonal context; the championship leverage of the first three at bats in first inning of Game 6 of the ALCS significantly exceeds the same leverage of the situation on Opening Day. Therefore, you should consider an opener if winning the game is considered more important than allowing a less experienced pitcher opportunities to learn how to pitch through and find outs in unfavorable matchups. Teams contending for a playoff spot or who are already playing are more likely to start with a reliever with a platoon advantage, simply to prevent the bulk starter from allowing the top of the lineup to see him until later.

4. Baserunning value

Relative to offense and defense, baserunning is the least important component of position player value. Elite baserunners typically contribute around 10-12 RAA (roughly 1 WAR) while top hitters can produce anywhere from 50-80+ runs above average. However, baserunning value compounds across a roster, and good baserunning teams can gain advantages of 30+ runs over a full season, roughly equal to three wins.

Baserunning value comes from a player developing a high baserunning IQ and then using it with their physical ability to maximize their value. Physical abilities include sprint speed, acceleration from standstill, and raw athleticism.

Although the top 5 players by BsR in 2025 all had a 90th sprint speed or greater (right), speed only has a moderate positive correlation with baserunning value. Wyatt Langford and Josh Naylor had an 85 percentile difference in sprint speed and both finished with 0 BsR. Being faster doesn't make you a better baserunner, it provides you with a wider range of opportunities to take advantage of the defense.

Rk.	Player	Team	Year	Baserunning Runs
1	Carroll, Corbin		2025	10
2	De La Cruz, Elly		2025	9
3	Buxton, Byron		2025	8
4	Turner, Trea		2025	7
5	Witt Jr., Bobby		2025	7

Baserunning IQ is a blanket term for the decision-making skills converting physical tools into runs. This can be measured in several ways (many of these are not independent of other factors):

- stolen base attempt volume, efficiency, and jump timing
- reading balls off the bat relative to fielder positioning
- taking bases on batted balls
- sliding technique, tag avoidance, and swim moves
- avoiding non-double-play outs

Because many of these depend of a player's game sense, these skills are generally interconnected. It would be odd to have a player excel at one while completely failing at others. The risk-reward dynamic matters significantly. Aggressive baserunning can produce negative value if executed poorly, making sound decision-making crucial. Context also plays a role as game situation, score, and inning influence optimal baserunning strategy.

5. Intentional high choppers

I believe the league has forgotten the situational value of intentional high choppers. With modern bat-tracking data measuring attack angles, we can now quantify and teach a swing path with a negative attack angle designed to drive the ball into the dirt or the plate with high exit velocity. If done correctly, you are legally left with a ground ball that has the hang time of a popup, eliminating runner tag-up requirements. Executed with a runner on third and fewer than two outs, this can increase run expectancy while limiting defensive options.

6. Pitch-type prediction model

I'd use a Statcast pitch-level data table (2021-2025), where each row represents a pitch including outcome, location, velocity, and categorical game context. This table then joins to pre-aggregated pitcher and hitter tendency tables: pitcher metrics (pitch usage % by handedness/count for both season and in-game, zone distribution, xwOBA/whiff rates by pitch type), batter metrics (xwOBA and whiff rates vs each pitch type from LHP/RHP—generalized, not pitcher-specific), and sample size features (PA/BF counts), along with every aggregated stat so the model learns to appropriately weight reliability.

Splitting would use time-based rolling windows:

Train	Validation	Test
2021-2023	2024, 1st Half	2024, 2nd Half
2022-2024	2025, 1st Half	2025, 2nd Half

I would use a LightGBM multi-class classifier to predict the next pitch type in an at-bat. It handles categorical features natively, captures non-linear interactions, and easily scales to the to millions of pitches in the data. If there are signs of overfitting, I would switch to an XGBoost.

The LightGBM model would ingest the following features:

- **Context:** balls, strikes, outs, runners for each base, pitching/batting team runs, leverage, inning, pitcher/hitter/catcher ID, pitcher/hitter sides, encoded platoon advantage (pitcher, hitter, or none). times faced (career and in-game), former teammates indicator
- **Pitcher tendencies:** pitch count, days of rest, usage % by batter handedness/count (season + in-game), zone tendencies by pitch % (per pitch type, handedness), xwOBA/whiff rates by pitch type, batters faced count.
- **Batter tendencies:** xwOBA and whiff% vs each pitch type from LHP/RHP, chase rate, swing rates by zone, plate appearance counts.
- **Sequential:** previous 3 pitches in current at-bat (type, location, result, velocity), back-to-back indicator

Sample size features allow the model handle uncertainty naturally instead of filtering out non-qualified players. When PA vs sliders from RHP = 15, model leans on general chase rate, but if PA = 200, it trusts specific performance.

To ensure the model is interpretable for coaches, I would run a SHAP analysis after training to confirm the sample-size features behave logically (low-sample data receives low SHAP magnitude), and that count, platoon, and pitcher usage are in the top predictors.

Validation

Primary metric: log loss (penalizes overconfident predictions)
Secondary: accuracy, top-3 accuracy (correct pitch in top-3 probabilities), per-pitch-type precision/recall
Baseline to beat: ~35% accuracy (always predicting four seamer, most common pitch type in 2025).

Sanity checks: count and sequential features should rank top-10 in feature importance; sample size features in top-30 (confirms model uses uncertainty signal); temporal stability across windows (no degradation); comparable performance on rookies vs veterans (proper generalization).

	Training	Validation	Test
Window 1	2021, 2022, 2023	2024, 1st half	2024, 2nd half
Window 2	2022, 2023, 2024	2025, 1st half	2025, 2nd half