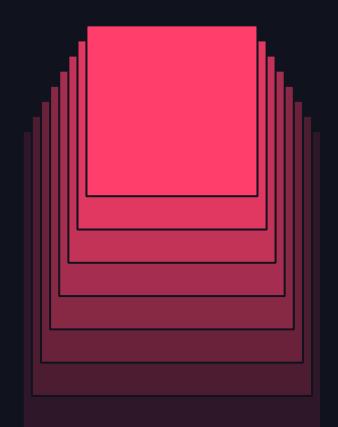


Texas Rangers' World Series Success with a Data Intelligence Platform

Texas Rangers Baseball Operations 6/12/2024



Who We Are

Texas Rangers Baseball Club

Alexander Booth

Assistant Director, R&D Texas Rangers Baseball Club Joined the club in 2018 abooth@texasrangers.com



Michael Topol

Sr. Data Scientist Texas Rangers Baseball Club Joined the club in 2023 mtopol@texasrangers.com



Oliver Dykstra

Data Engineer, R&D Texas Rangers Baseball Club Joined the club in 2022 odykstra@texasrangers.com



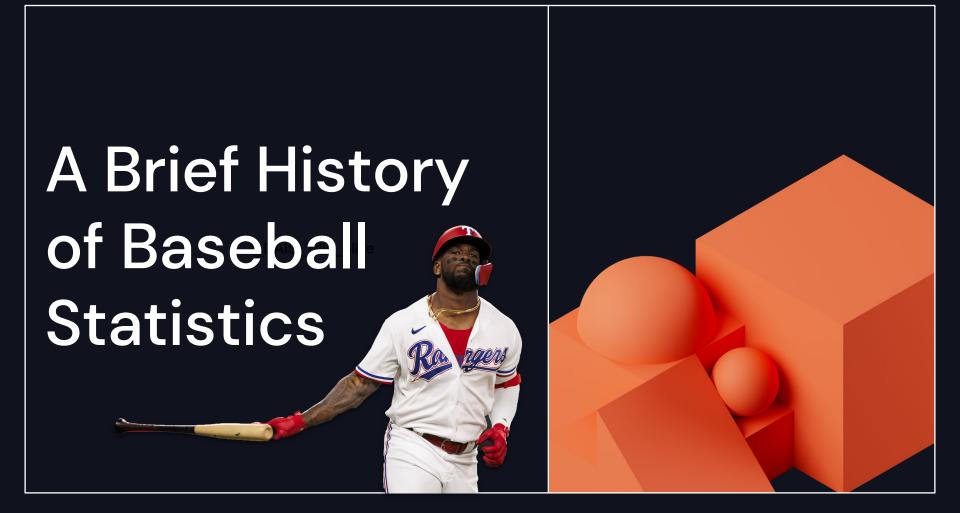
What We're Talking About

Texas Rangers Baseball Club

- History of Baseball Statistics
- Big Data Baseball
- The Winning Formula:
 - Cutting-Edge Data Engineering
 - Mastering Data Governance
 - Unleashing Al-Driven Analytics
- Putting it all together:

World Series Victory





That's all baseball is, is numbers; it's run by numbers, averages, percentages and odds...

Rollie Fingers(Hall of Fame Pitcher)

A Brief History of Baseball Statistics

1859 - Henry Chadwick publishes the first Box Score, a set of statistics compiling the runs, hits, outs, assists and errors.

Chadwick also coined the term "strike-out"! Further, to this day the reason a strikeout is recorded as a K is because Chadwick used the phrase "the batter was struck."



A Brief History of Baseball Statistics

1941 - Former major leaguer Ethan Allen invents All-Star Baseball, a tabletop game that allows kids to imitate major league games with a spinner on top of circular discs.

1952 - Topps adds full statistics lines on the back of their annual baseball cards.

Baseball statistics start to proliferate through popular culture.

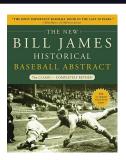


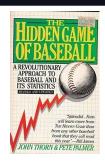
A Brief History of Baseball Statistics

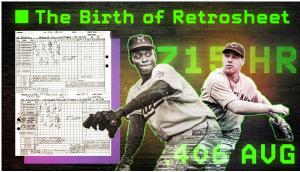
1977 - Bill James publishes his first "Baseball Abstract" which becomes a national bestseller in the early 80s.

1989 - Retrosheet begins massive compilation and online publishing of old box scores and play-by-plays, allowing droves of historical research never before possible.

1996 - Baseball Prospectus begins publication of their annual and website - introduces statistics community to VORP, PECOTA, Pitcher Abuse points and more.





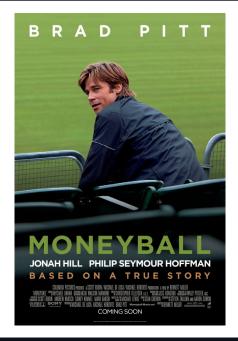


A Brief History of Baseball Statistics

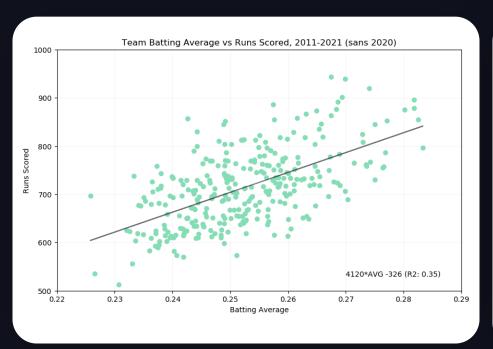
2001 - 2003 - Billy Beane's Moneyball A's use data-driven insights to identify market inefficiencies within the game of baseball.

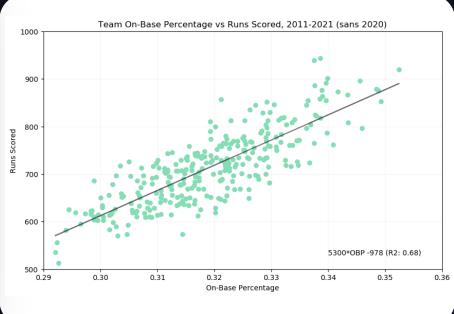
2003 - Michael Lewis publishes his book, "Moneyball: The Art of Winning an Unfair Game" about those A's.

2011 - Brad Pitt stars in the movie adaptation of the same name.



A Brief History of Baseball Statistics





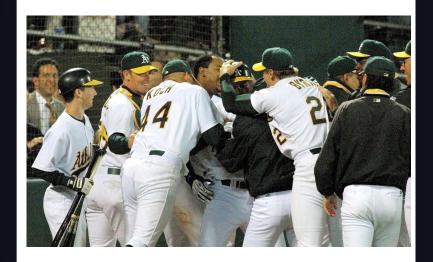
A Brief History of Baseball Statistics

Billy Beane identified a market inefficiency.

The market priced players with high batting averages higher than those with high on-base percentages. However, on-base percentage has a **higher correlation** to total runs scored.

The Oakland A's used this information to acquire players undervalued by the market that could help them compete with higher payroll teams.

This data-driven decision **disrupted the industry** and left a legacy far beyond baseball.





"If you challenge conventional wisdom, you will find ways to do things much better than they are currently done."

- Bill James

(Founder of Sabermetrics)

The Statcast Revolution

2001–2002: Moneyball, Billy Beane, Oakland A's identify data-driven market inefficiencies.

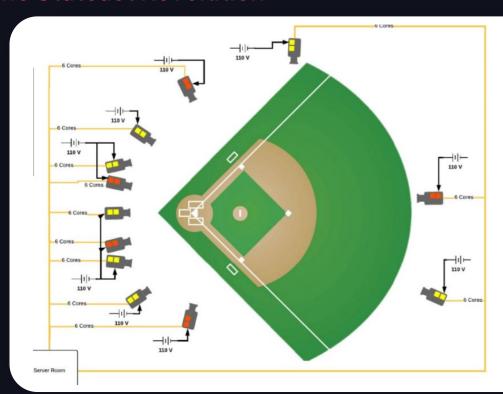
2015: Statcast Debut. Radar + HD Video measures all action on the field, per pitch. 2020: Statcast switches from TrackMan to Hawk-Eye as its technology provider.

2022: Statcast deployed to AAA. Widespread MiLB adoption planned.

2006: PITCHf/x Ball Tracking Debut. Spin rates, velocity, and movement all tracked.

2017: Statcast switches from PITCHf/x to TrackMan as its technology provider.

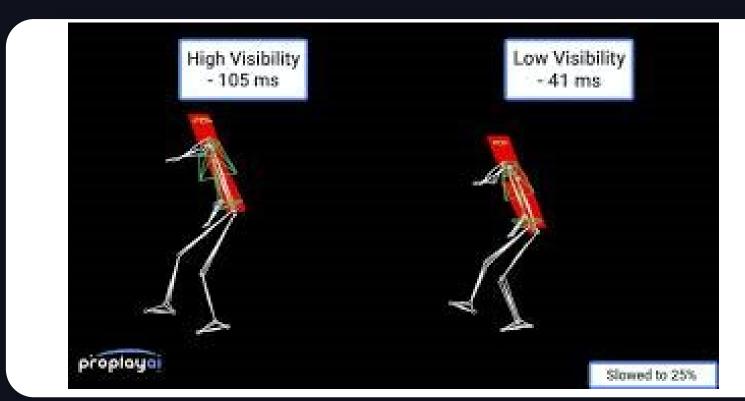
2021: Pose Tracking and FieldVision debut.
Skeleton and body movements tracked.













The Statcast Revolution

CATAPULT















TRACKMAN











P PERFECT GAME







©2024

©2024 Databricks Inc. — All rights reserved

SPARTASCIENCE



The Winning Formula:

Cutting-Edge Data Engineering



"It's math-y, but there's still the whole arts-and-science debate [with big data]. I'd argue there is an art to that sort of stuff."

Mike Fitzgerald
 (VP of R&D for the Diamondbacks)

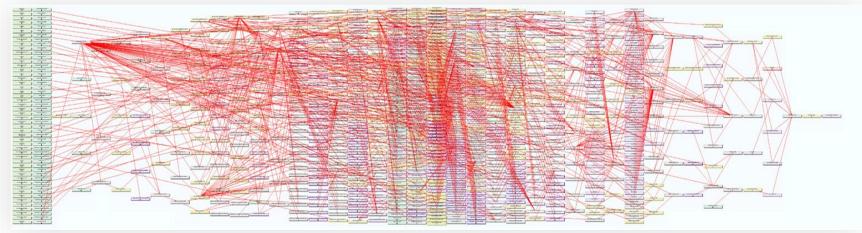
The Obstacles

Original Complexity

Our first data solution failed to adapt as we started ingesting big data.

Problems:

- Difficult to make changes
- Hard to troubleshoot
- Tough to explain



The Obstacles

Limitations of Two-Tier Architecture

Technical limitations:

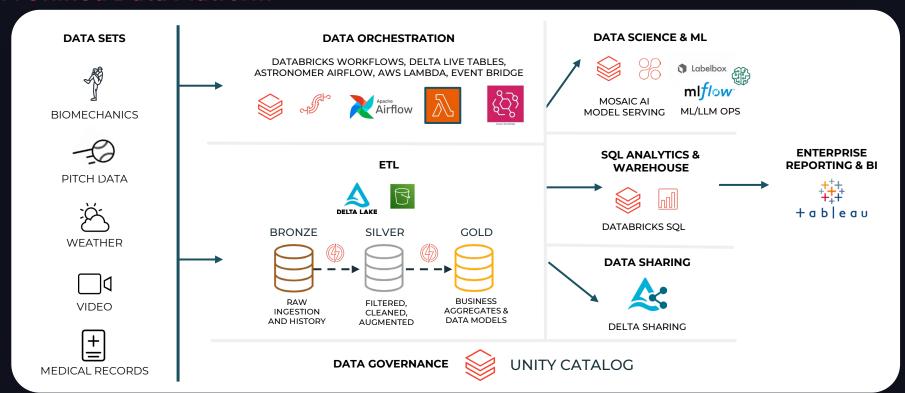
- Time-consuming and expensive maintenance efforts
- Cost prohibitive proprietary formats and redundant data replications
- Integrated storage and compute lead to inefficient and brittle transformation layers
- Not flexible enough for a variety of data formats or semi-structured data

When you're asked how the cloud migration is going:



The New Potential

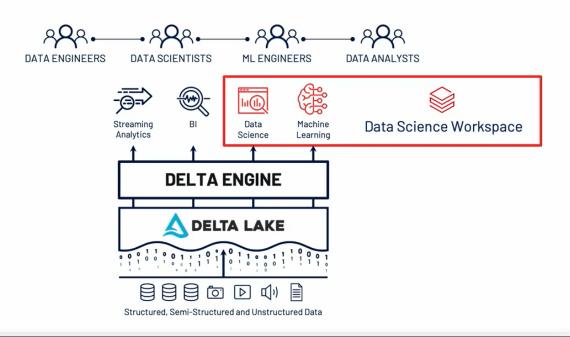
A Unified Data Platform



The New Potential

A Unified Data Platform

Solving the world's toughest problems is a team sport



27

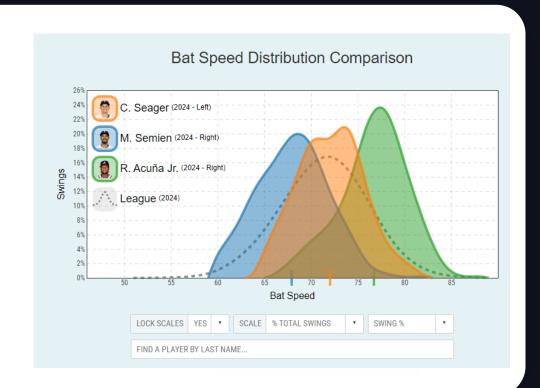
The New Potential

New Data Sources

Statcast **bat tracking** is available beginning with the 2024 season.

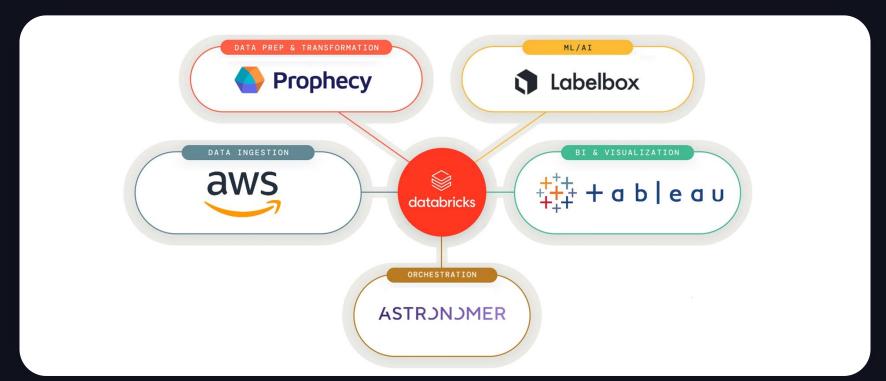
Since different parts of a bat can move at different speeds, an individual swing's speed is measured at the point six inches from the head of the bat, what is popularly called "the sweet-spot."

Swing length tracks the sum distance traveled by the head of the bat in XYZ space from the start of data until contact point.



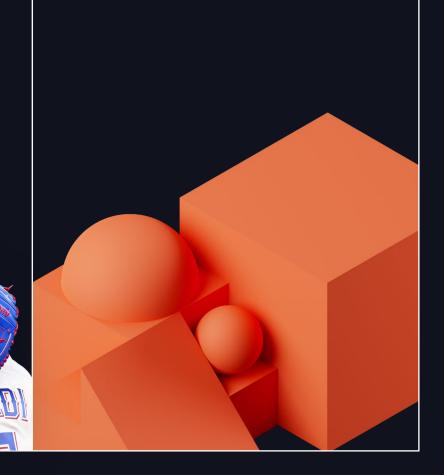
Open Source and Partner Solutions

A Pantheon of Partner Solution



The Winning Formula:

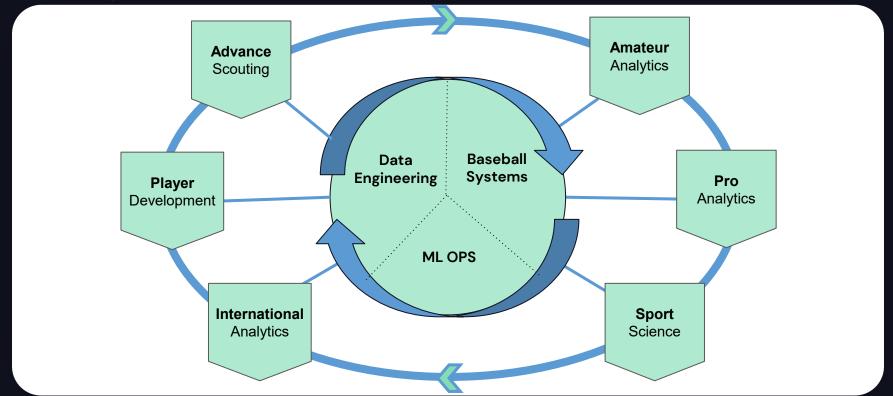
Mastering Data Governance



"There were too many things that were taken for granted that shouldn't have been."

Ben Lindbergh(Author and Analyst)

Hub and Spoke Federation



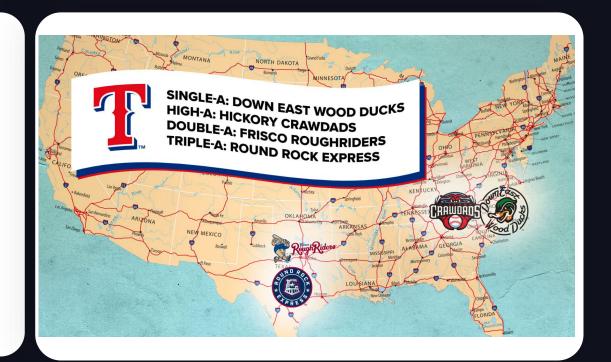
32

A World Wide Presence

Affiliates

DSL Rangers (DR, Rk)
ACL Rangers (Arizona, Rk)
Down East Wood Ducks (NC, A)
Hickory Crawdads (NC, A+)
Frisco RoughRiders (Texas, AA)
Round Rock Express (Texas, AAA)

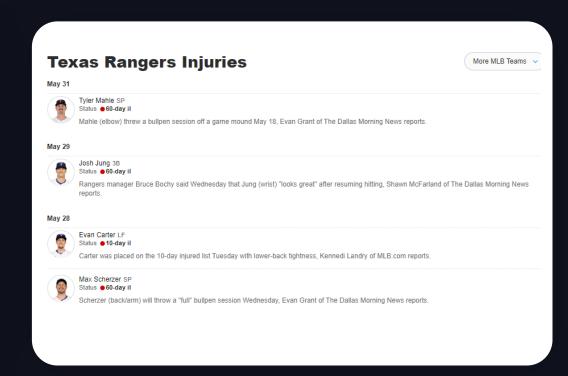
Texas Rangers (Texas, MLB)



Sensitive Information

Secure Data Sources

- Medical Information
- Addresses, PII
- Contract Details
- Financials
- Education History Amateur Draft
- FERPA, HIPAA

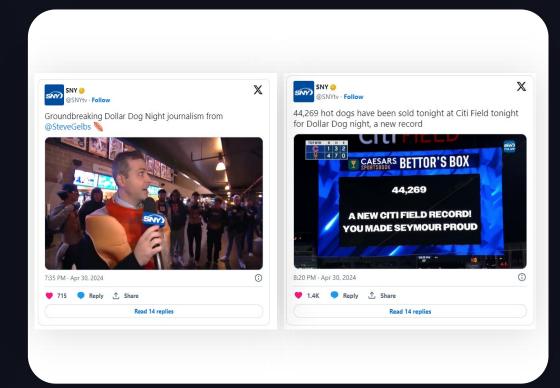


Glizzies 🏉

Case Study: NY Mets 4/30/2024

"SNY sideline reporter Steve Gelbs, dressed as a hot dog, did a hard-hitting report in front of the "\$1 Wiener Boys" where he explained that the Mets normally sell 4,100 hot dogs per game. For \$1 hot dog night, the team ordered 70,000 frankfurters...

When all was said and done, the attendance for the Cubs-Mets game was 22,880 and 44,269 hot dogs were sold."



Glizzies 🏉

Machine Learning & Al can help predict how many \$1 hot dogs will be consumed - leading to more *optimized inventory management*.

However, knowing the starting pitcher or lineup, the weather, or expected score would likely affect attendance and consequent hot dog sales.

Therefore, governing shared data between the business, baseball, and concessions departments is pecessary



Source:

https://www.si.com/mlb/2024/05/01/mets-dollar-hot-dog-night-madness

Mastering Data Governance

Unity Catalog

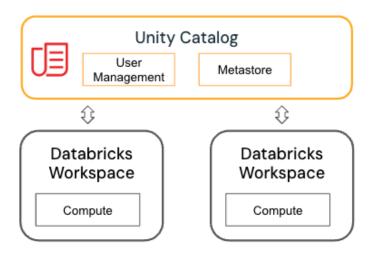


Without Unity Catalog





With Unity Catalog



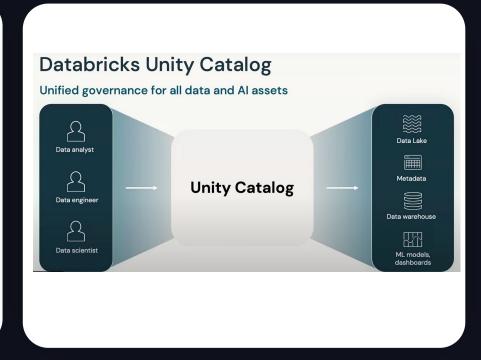
37

Mastering Data Governance

Unity Catalog

Unity Catalog enables centralized, **finegrained access control** and governance across our diverse data teams, such as player development and amateur scouting, while **ensuring compliance** with FERPA and HIPAA.

It provides **secure**, **unified access** to data from our minor league teams nationwide, while **safeguarding sensitive information** like contract financials and medical records. UC allows us to maintain transparency, ensure data integrity, and support compliance, all within a **single**, **integrated platform**.

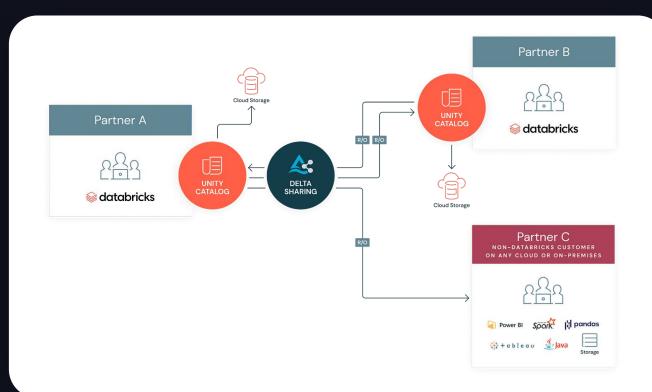


Mastering Data Governance

Delta Sharing

Unity Catalog with **Delta Sharing** allows us to connect seamlessly with data partners both on Databricks as well as on other platforms.

UC facilitates secure and efficient data sharing across our baseball and business ecosystem.





"Half of it's art... it's creativity, and then half of it is just knowing the data you're working with and being able to manipulate it in the direction that will benefit the player."

> Brian Bannister
> (Director of Pitching, Chicago White Sox)



Using data intelligence to optimize performance delivers competitive advantage



Talent Acquisition

MLB DRAFT ANALYSIS



Player Performance

BIOMECHANICS

SWING ADJUSTMENT

PITCH OPTIMIZATION



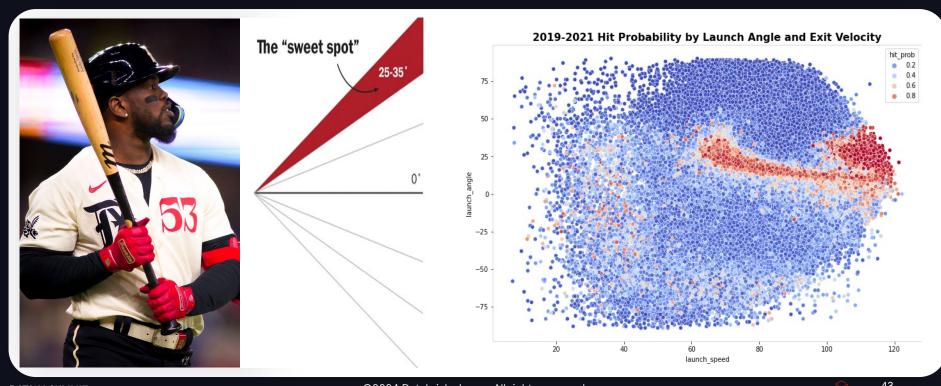
In-Game Strategy

DEFENSIVE POSITIONING

LINE-UP OPTIMIZATION

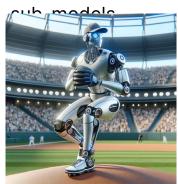
BULLPEN MATCH-UPS

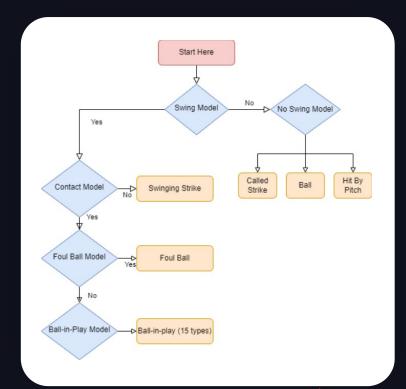
The Launch Angle Revolution



Grading Pitch Quality

- Model pitch-tracking data to assess the quality of a pitch given its expected results
 - O PitchingBot & Stuff+ are public models from Fangraphs
 - O Often a combination of many



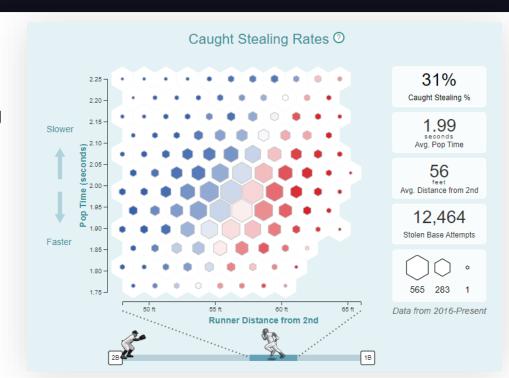


Base Stealing Probability



What is this? A <u>Statcast metric</u> designed to express the skill of catchers at throwing out runners on steal attempts, given the specifics of the opportunities they are presented with.

How this works: Each steal attempt is assigned a *modeled probability* of being successful or not based on several features at the time the pitch crosses the plate, most notably: runner distance from 2B, runner speed, catcher pop time, pitch location, and pitcher/batter handedness.



Defensive Positioning

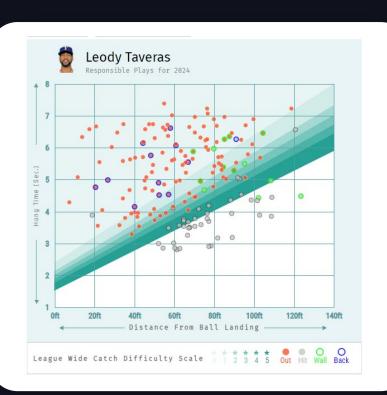


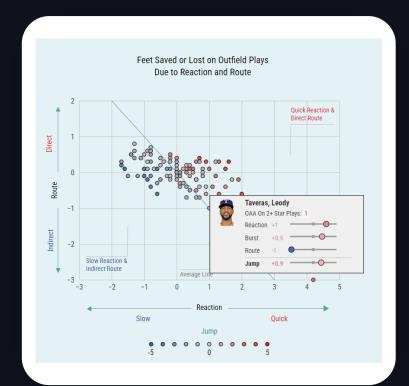
Right Handed Hitters



Various
modeling
frameworks
can
recommend
positioning
for each
player and
pitcher
combo

Analyzing Defense





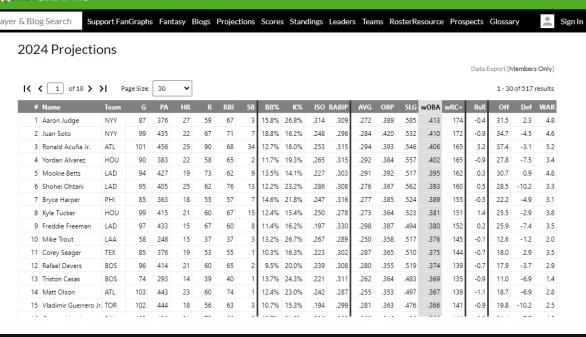
.XIFANGRAPHS

Player Projections

Statistical and machine learning models predict player projections.

Projections give us confidence into future player performance and expected returns from trades and free agents.

ZiPS on Fangraphs is one example of a projections system in the public sphere.

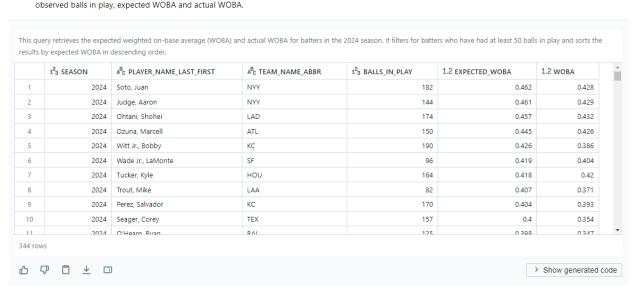


Generative Al

Beyond Copilots, Databricks Genie Rooms allow for tabular data queries using natural language.

This example finds top players this season by sorting on the KPI "Expected WOBA".

Our own Corey Seager makes the top 10!

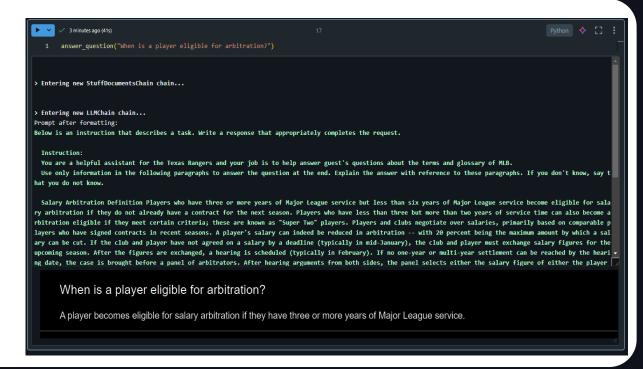


Which Batters in the 2024 season, with at least 50 balls in play, have the highest expected WOBA? Bring back the season, their name, team,

Generative Al

Using LangChain, Chroma DB, HuggingFace and DBRX, we have built RAG chat applications from entirely open-source technologies on top of our own documentation and glossaries.

These are used to quickly query and summarize our written documents given baseball specific language.



Generative AI - Fan Engagement





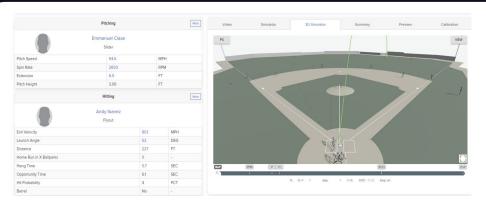
51

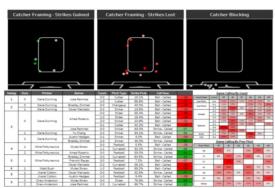


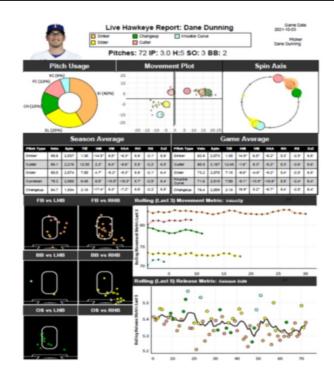
"If you can make yourself half a percent better... then that's a win. It can be the difference between making the playoffs and not making the playoffs."

> - Ryan Murray (Sr. Director of Baseball R&D, Texas Rangers)

Reports and Visualizations

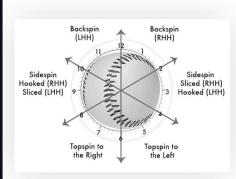






54

KPIs and Development



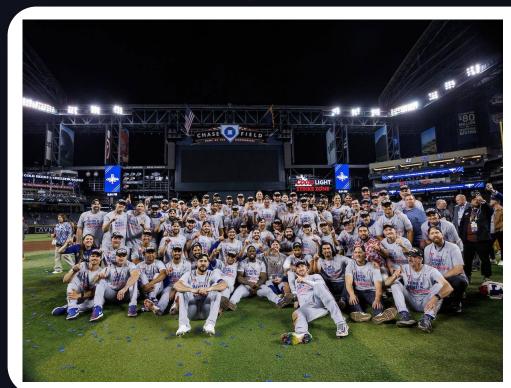




World Series Victory



World Series Victory





World Series Victory







Databricks is delivering the data intelligence that helps the Rangers to win

4X

300+

10X

ST

more data at the same cost as other multi-cloud data warehouses

users accessing
data securely
under a unified
governance model

faster democratization of insights across the organization ever world series championship



