DATA+AI
SUMMIT 2022

Big Data in the Age of Moneyball

Analyzing baseball's modern data revolution

Alexander Booth
Senior Analyst, Texas Rangers

Ryan Stoll
Data Engineer, Texas Rangers

ORGANIZED BY databricks

Agenda

Texas Rangers Baseball Club

1) Who We Are
2) The Age of Moneyball
3) Statcast (R)Evolution
4) Big Data Discovery
   a) How the Texas Rangers use Databricks
5) Case Study: The New Science of Hitting

Who We Are

Texas Rangers Baseball Club

Alexander Booth

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

Ryan Stoll

Data Engineer, R&D
Texas Rangers Baseball Club
Joined the club in 2021
rstoll@texasrangers.com

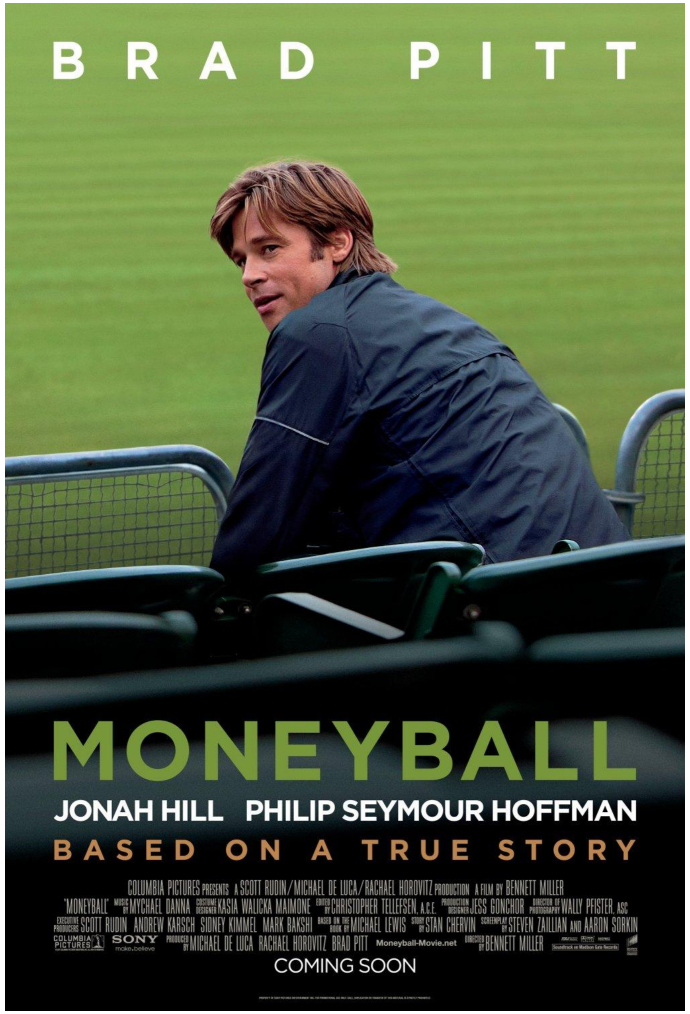
The Age of Moneyball

The Age of Moneyball

The start of a revolution

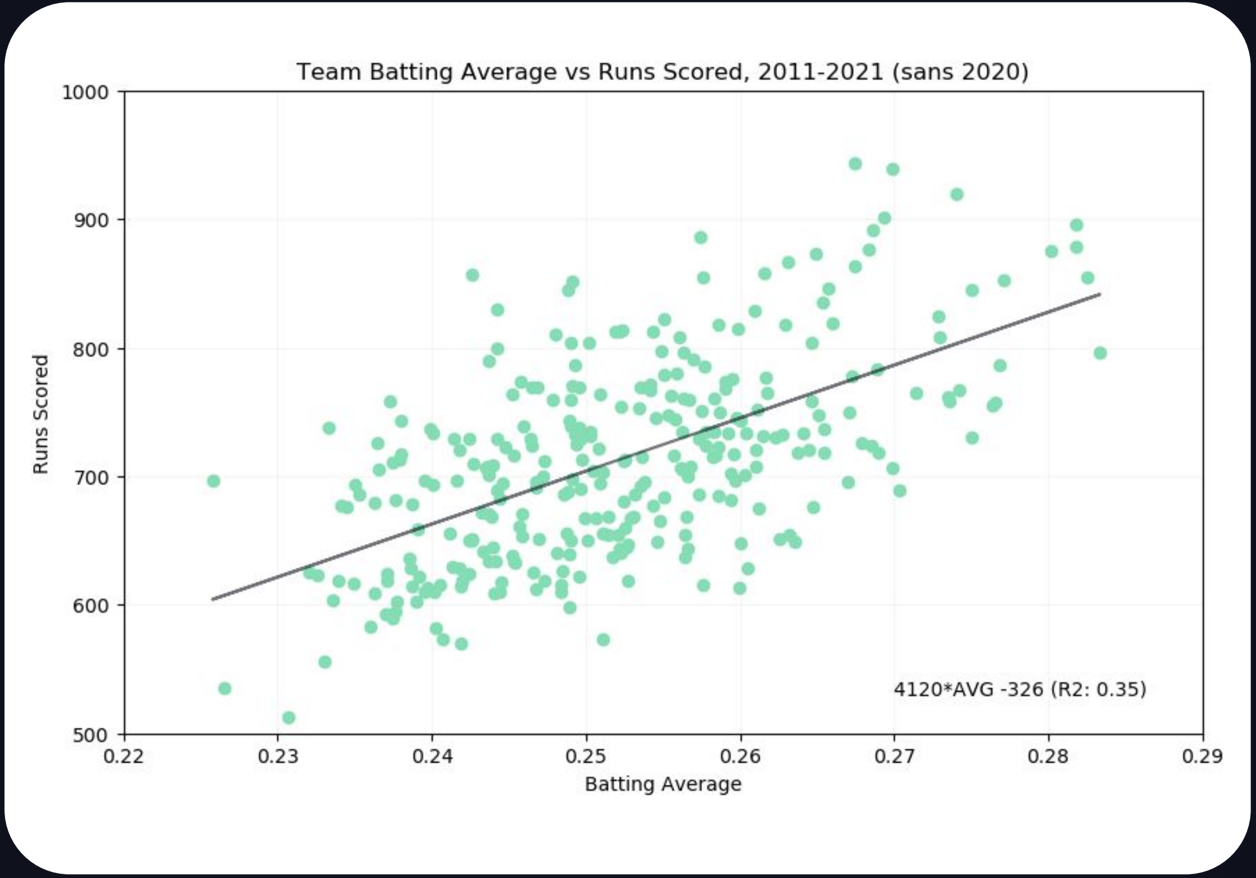
"If you challenge conventional wisdom, you will find ways to do things much better than they are currently done."

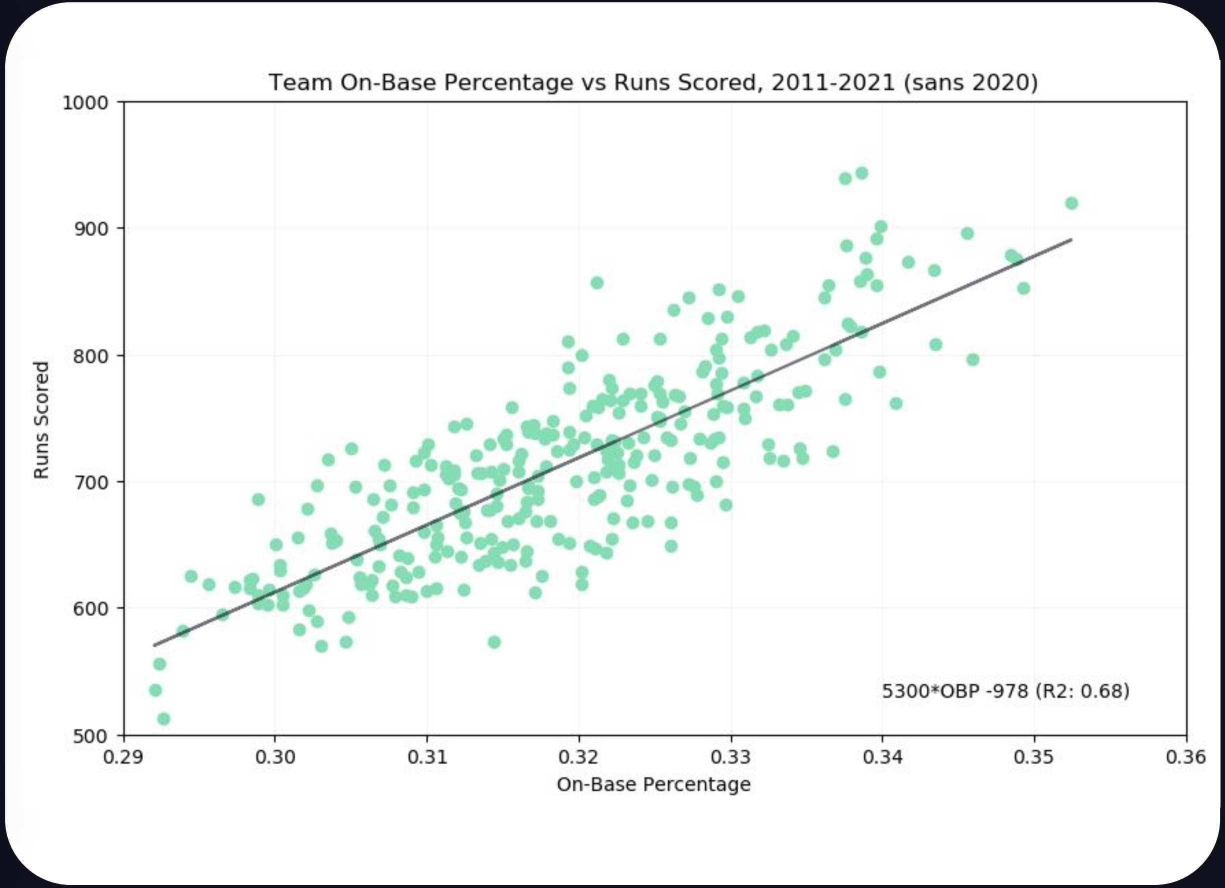
Bill James



The Age of Moneyball

"You get on base, we win. You don't, we lose. And I hate losing." - Brad Pitt/Billy Beane





The Age of Moneyball

Data Disruption

Billy Beane identified a **market inefficiency**.

The market historically 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.

Statcast (R)Evolution

Statcast (R)Evolution

Data, Data Everywhere

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

2006: Pitch F/X Ball Tracking Debut. Spin rates, velocity, and movement all tracked.

2015: Statcast Debut. Radar + HD Video measures all action on the field, per pitch.

2017: Statcast switches from Pitch F/X to TrackMan as its technology provider.

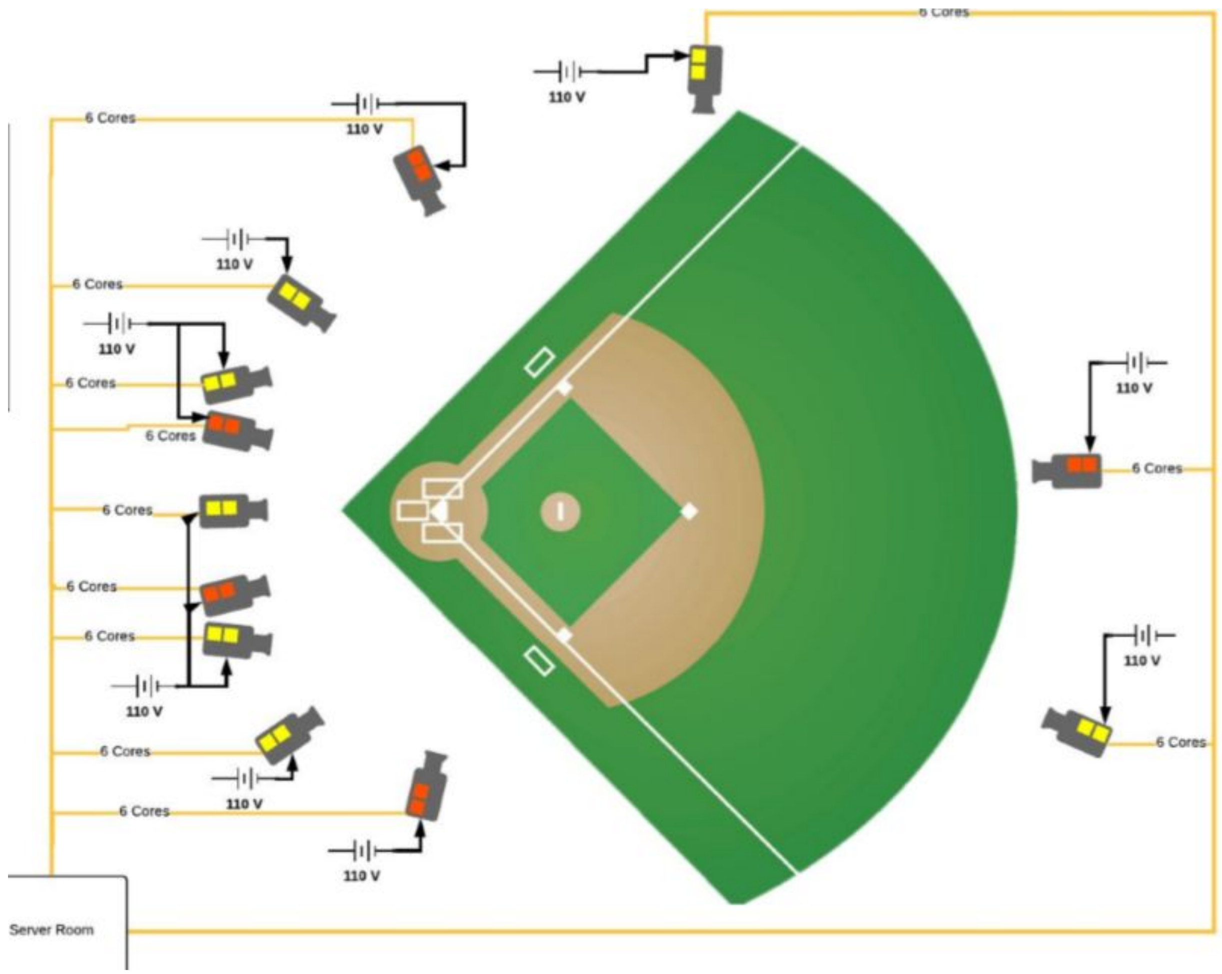
2020: Statcast switches from TrackMan to Hawk-Eye as its technology provider.

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

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

Statcast (R)Evolution

Hawk-Eye 12 Camera System

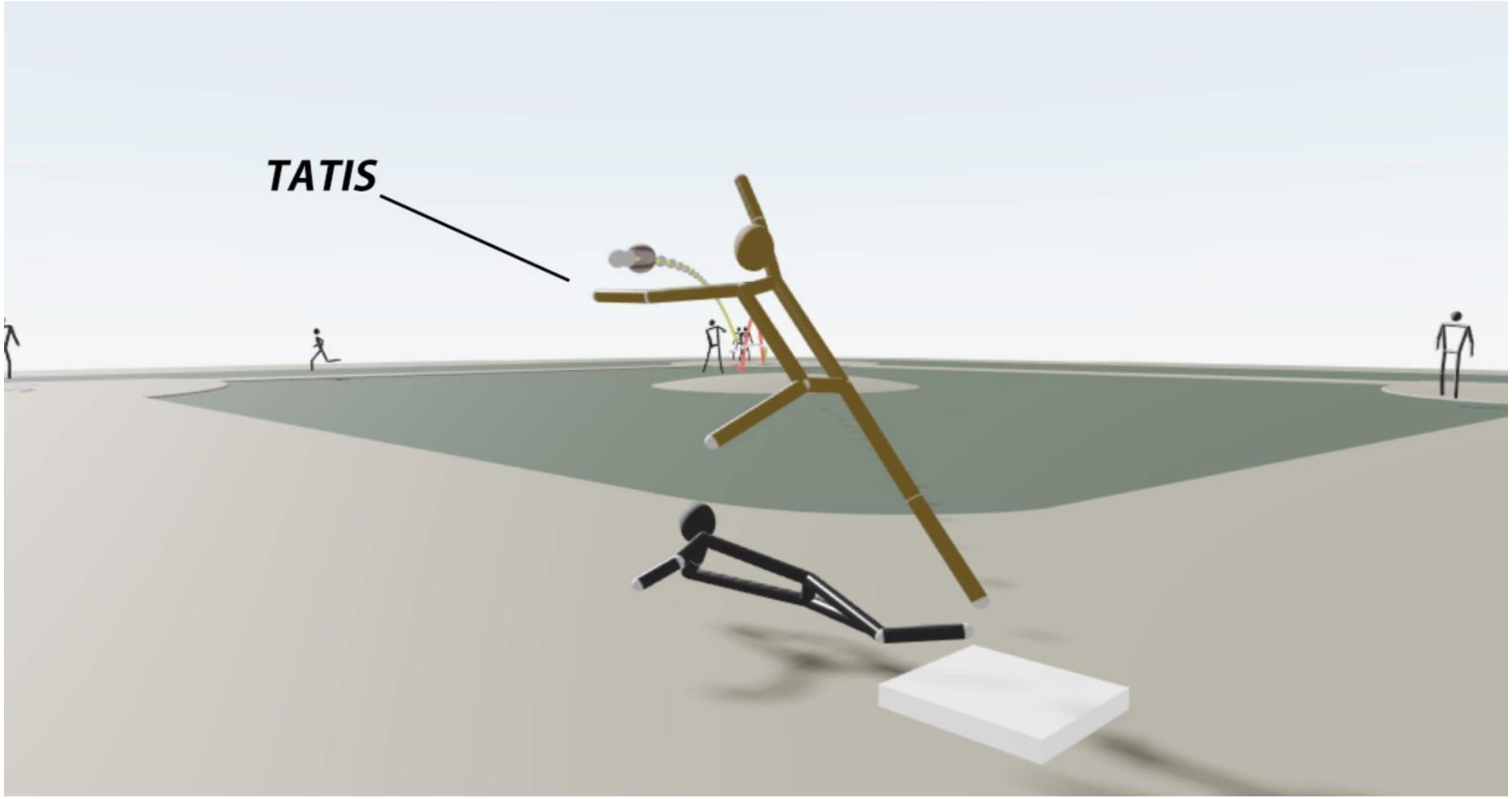


Statcast (R)Evolution

Skeleton Pose Tracking

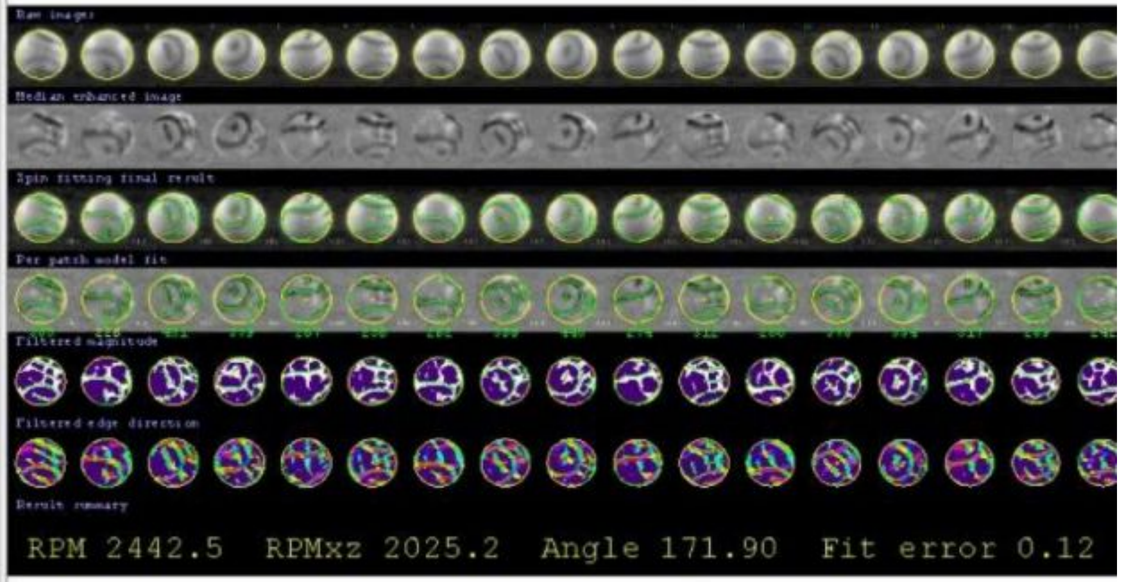
Statcast (R)Evolution

FieldVision



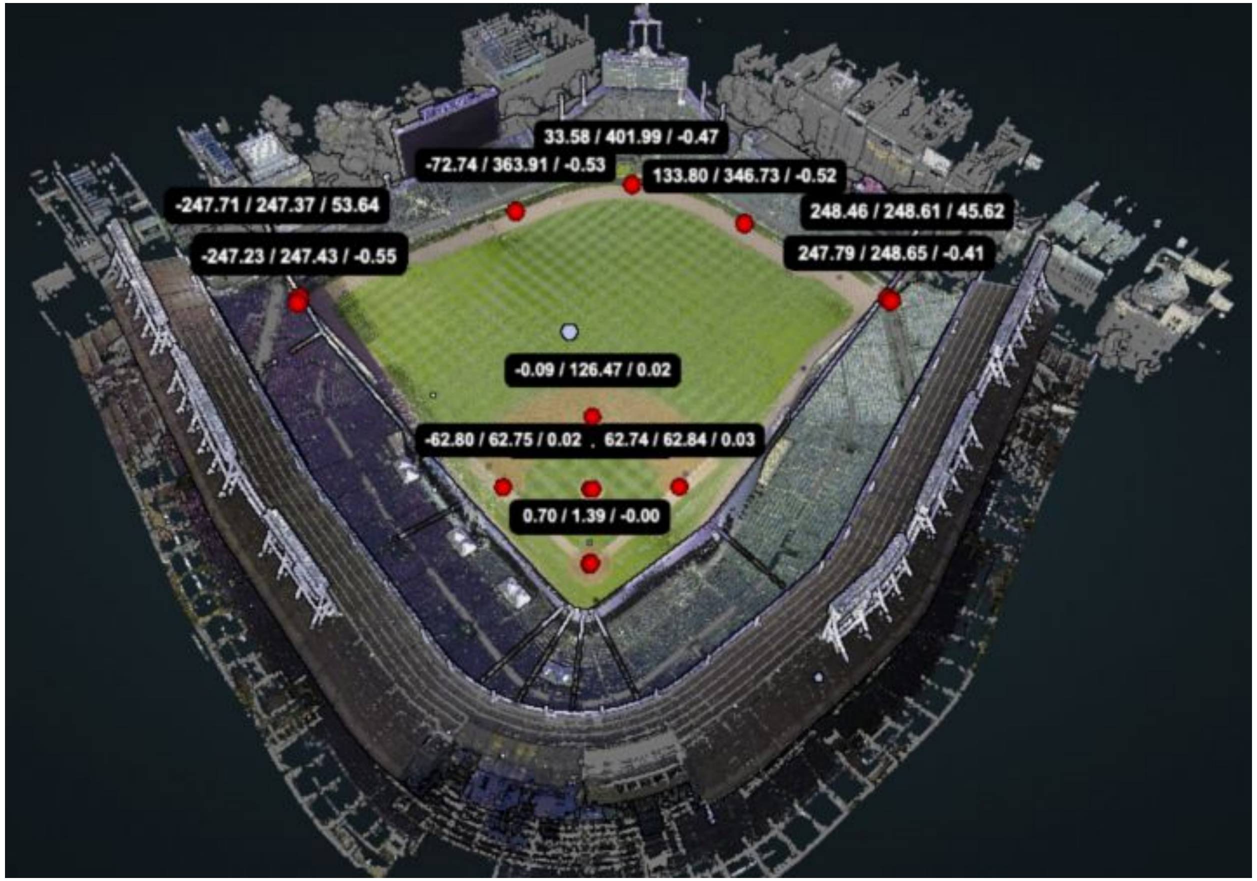
Statcast (R)Evolution

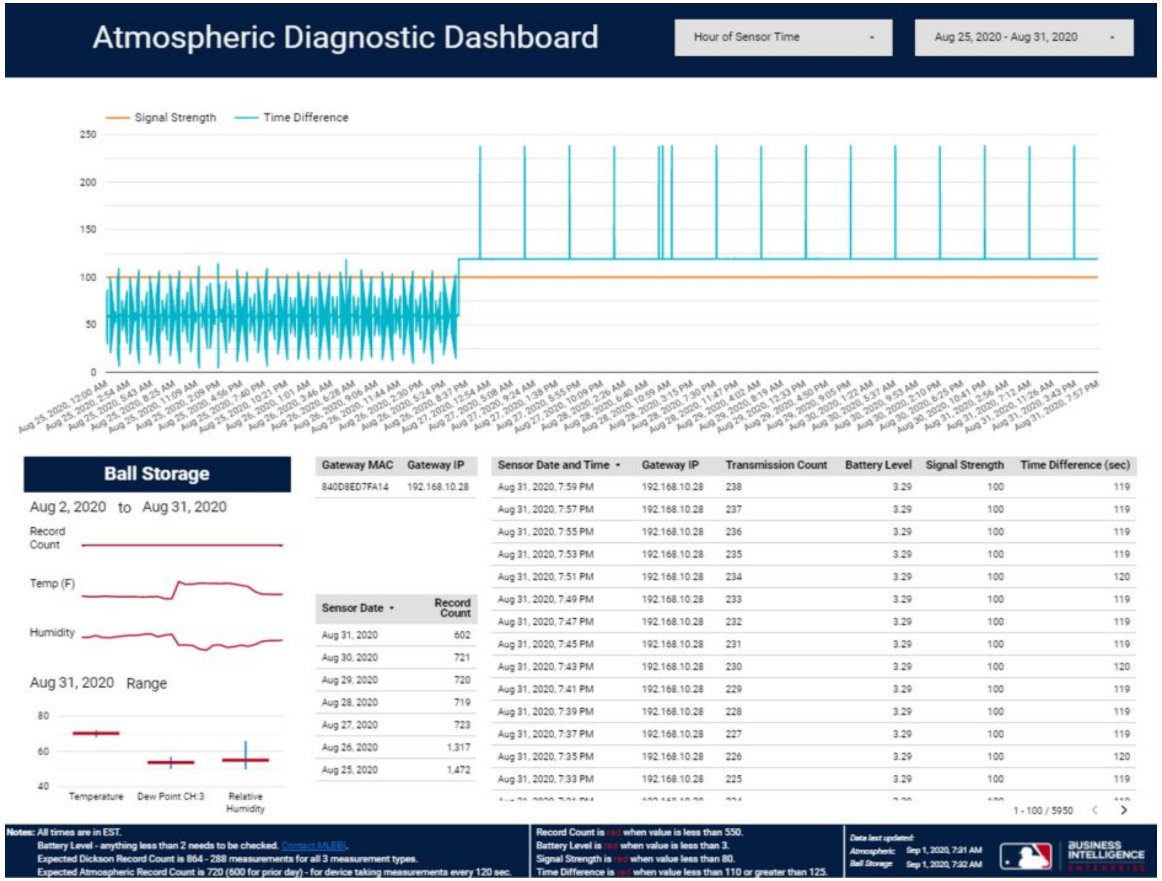
Seam Orientation and Observed Spin Tracking



Statcast (R)Evolution

LIDAR Scans and Weather Tracking





Statcast (R)Evolution

Current Technology Landscape

Diamond KINETICS

SiMi
reality motion systems

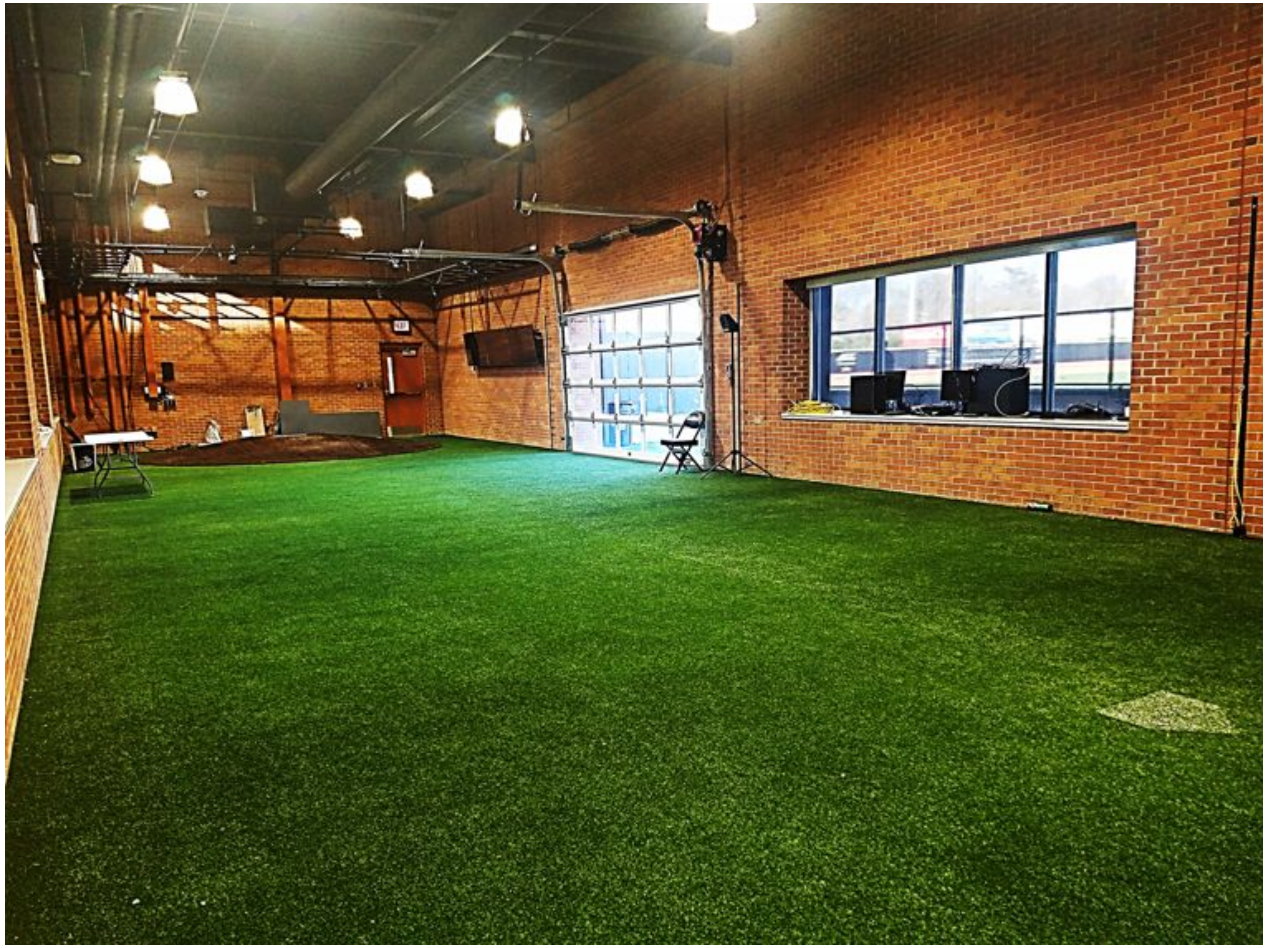
BLAST.

QUALISYS

KINATRAX

Statcast (R)Evolution

Wake Forest Pitching Lab



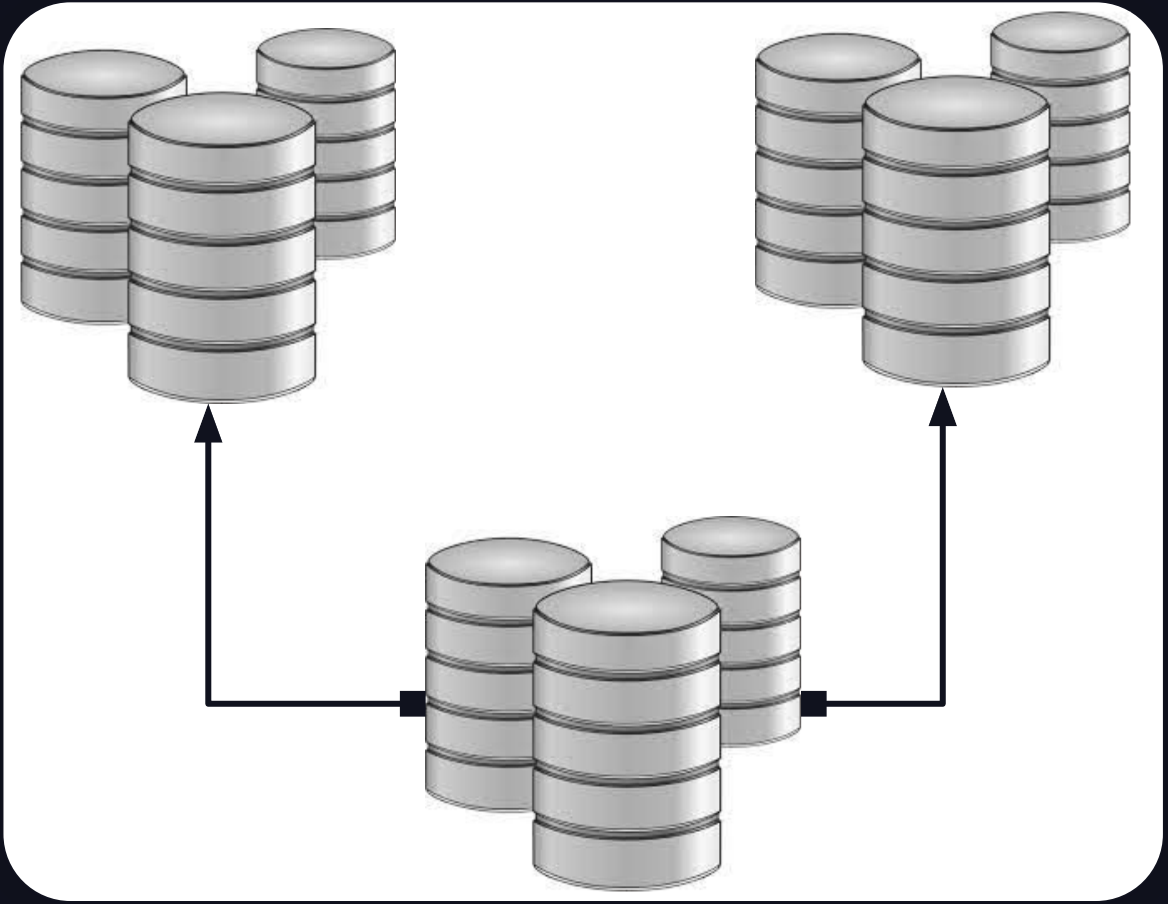
Big Data Discovery

Big Data Discovery

Siloed teams, divided data

Baseball Analytics Departments

- Pro Scouting
- Amateur Scouting
- International Scouting
- Player Development
- Advance Game Preparation
- Player Contract Negotiations
- Internal Player Evaluation



Big Data Discovery

All departments want to consume data

Big Data Discovery

Siloed teams, divided technology

Disparate technologies

- On-prem Databases
- Cloud Databases
- Cloud Data warehouses
- Python
- R
- Tableau/PowerBI
- Multiple cloud providers

Big Data Discovery

Databricks Unified Analytics Platform

databricks

Big Data Discovery

Unified Data Engineering

How do you ingest dozens of disparate data sources at scale?

Before, we had different ingestion scripts, running on different on-prem and cloud based servers, saving to different databases.

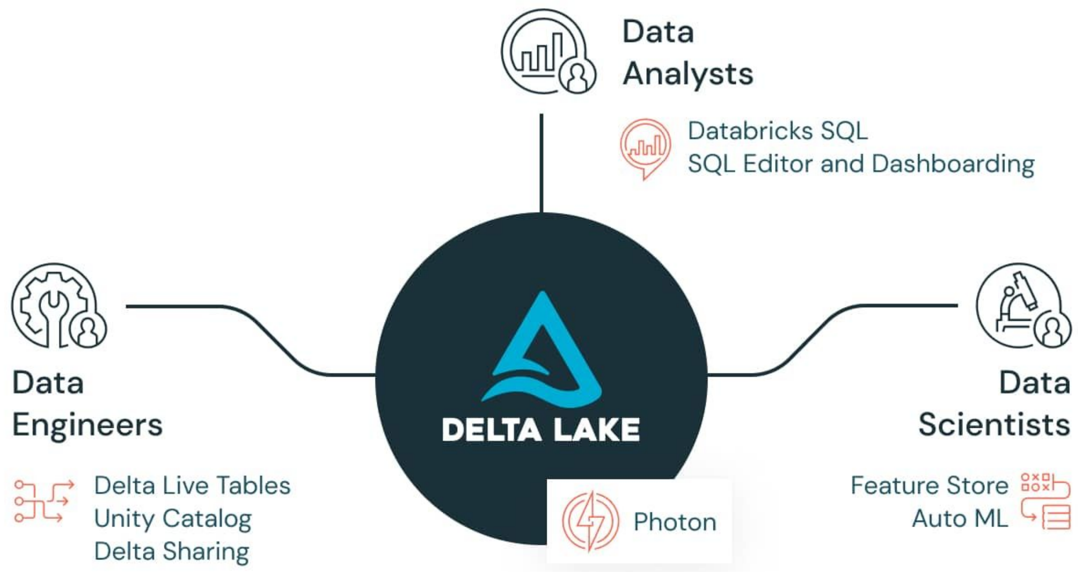
Big Data Discovery

Unified Data Engineering

Extract: APIs, FTPs, CSVs, other databases

Transform: Flatten, combine, clean

Load: Into staging Delta Lake table as needed, before loading into a single, cloud-hosted, production data warehouse.



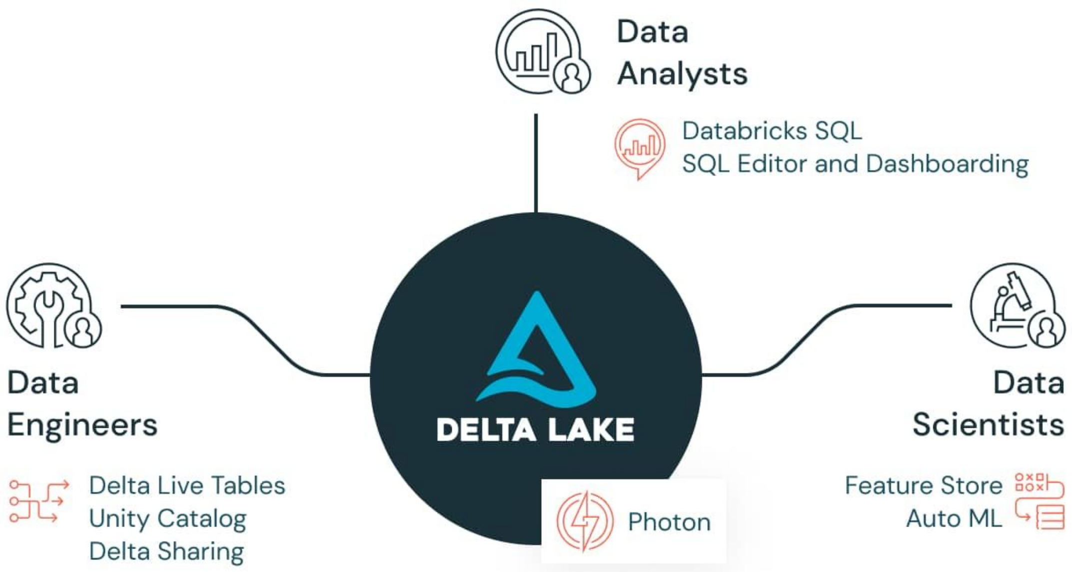
Big Data Discovery

Unified Data Engineering

By using Spark, Koalas, and the new integration of Koalas into PySpark, we can perform distributed extraction requests.

We can transform millions of pitches with as much compute as required.

We can load at the speed of Spark.



Big Data Discovery

Unified Data Engineering and MLOps

For the first time, since our engineering scripts and ML models are hosted on a **unified analytics platform**, we are also able to score and generate predictions as the data is extracted and transformed.

This allows us to **communicate insights** at a more rapid pace to our players and coaches to create fast decisions.

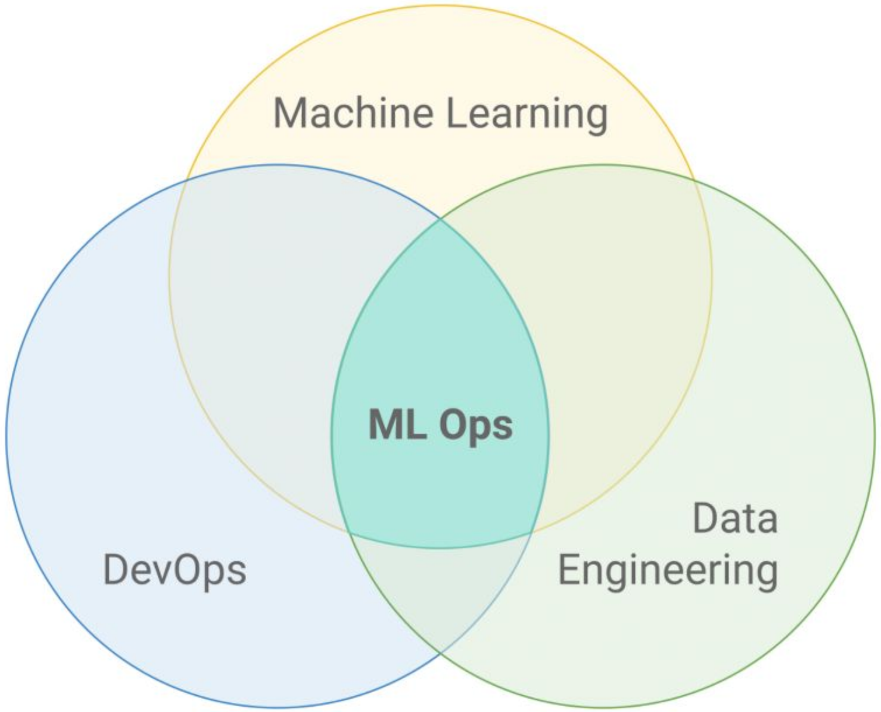
DELTA LAKE

mlflow™

Big Data Discovery

Unified Machine Learning Development

- DevOps is characterized by key principles: shared ownership, workflow automation, and rapid feedback.
- Automation is a core principle for achieving DevOps success and CI/CD is also a critical component.
- MLOps Involves building, deploying, and maintaining ML models reliably & continuously in an automated way.

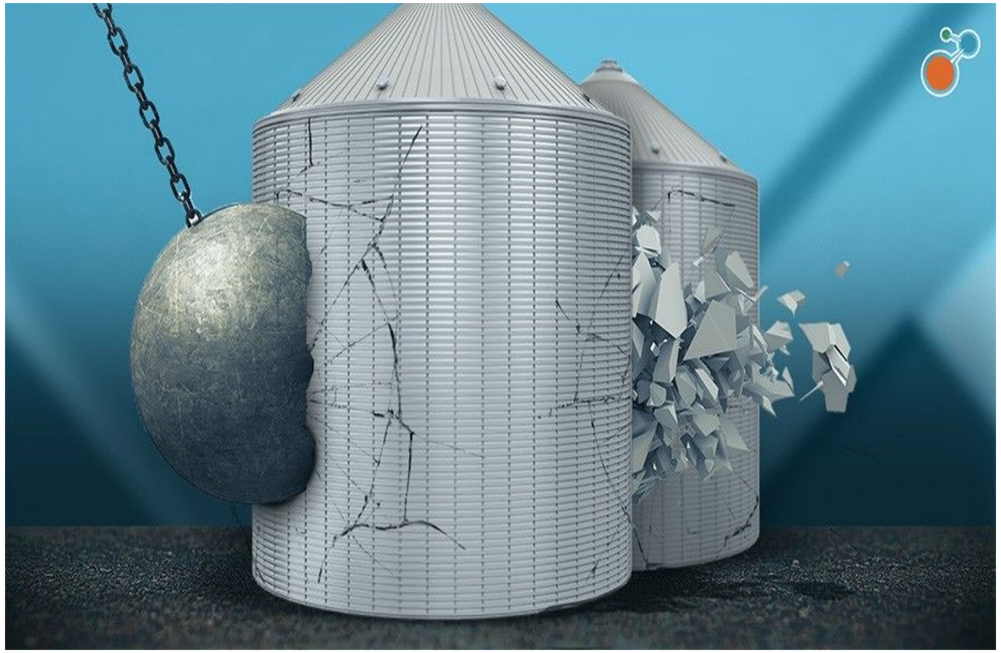


Big Data Discovery

Unified Machine Learning Development

Benefits of MLOps

- Models stored in the cloud, so everyone has access – **transparency**
- Easy peer and code reviews
- Models are retrained & promoted into production **automatically**
- Models are maintained & **monitored**
- Changes to models are tracked



Big Data Discovery

Unified Machine Learning Development

mlflow™

Integrations with:

TensorFlow, PyTorch, Keras, Spark, scikit-learn, RAPIDS, H2O.ai, python, R, Java, mleap, ONNX, GLUON, XGBoost, LightGBM, spaCy, fast.ai, statsmodels, CatBoost, PyCaret, ALGORITHMIA, OPTUNA, RAY, CONDA, kubernetes, docker, Amazon SageMaker, Azure Machine Learning, Google Cloud, databricks

Organizations using and contributing to MLflow:

databricks, Microsoft, RStudio, facebook, University of Washington, Virgin Hyperloop One, Booking.com, Toyota, Zillow, SK telecom, accenture, criteo, Outreach, splice machine, Telia Company, TEKION, Bering, SCRY Analytics, WiX.com, ABRAXAS, Tiger Shark, virtusa, Health 2 Sync, mined.ai, dL, SYSTEMATIC, despegar.com, ModelWorks, Brandfolder, ASML, fractal, Acerta, ENABLE AI, Linde, TENPOINT7, plain concepts, xpandit, Kobalt, CognitOps

Big Data Discovery

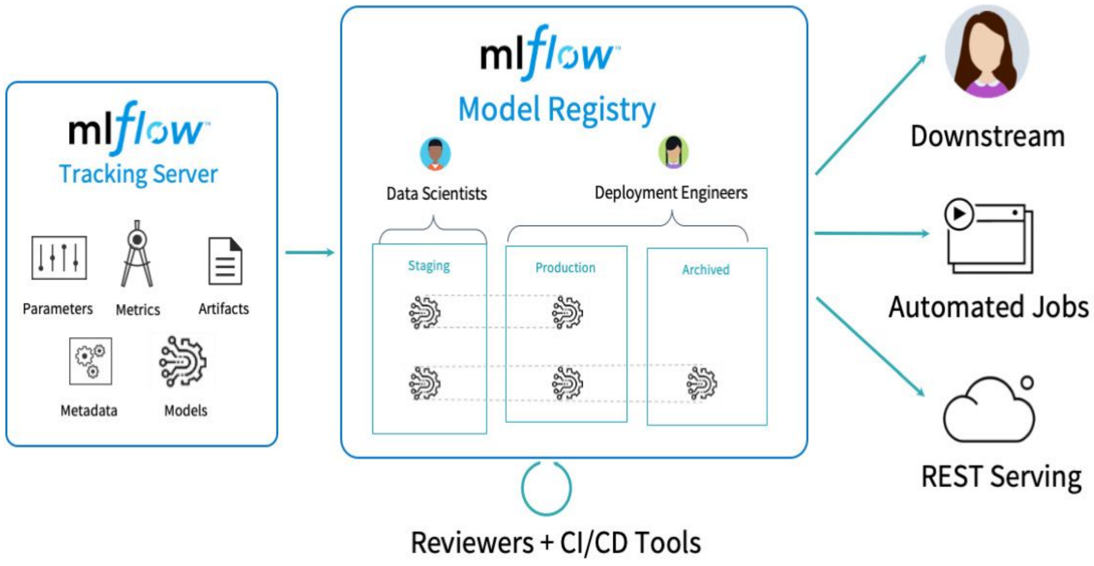
Unified Machine Learning Development

Two key components: model tracking and model registry.

Model Tracking:

UI that logs features, parameters, models, and metrics for ML models.

Multiple different models can easily be compared and reproduced.



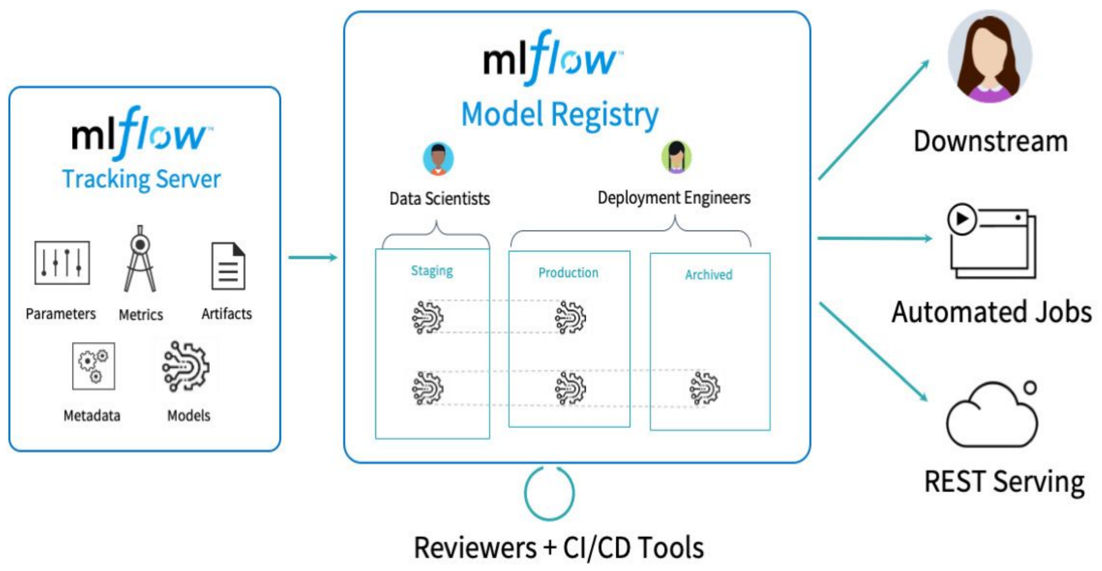
Big Data Discovery

Unified Machine Learning Development

Model Registry:

A centralized, cloud storage system for machine learning models built in Python, R, and AutoML frameworks.

All previously stored versions of a model are saved and can be promoted to development, staging, and production.



Big Data Discovery

Unified Machine Learning Development

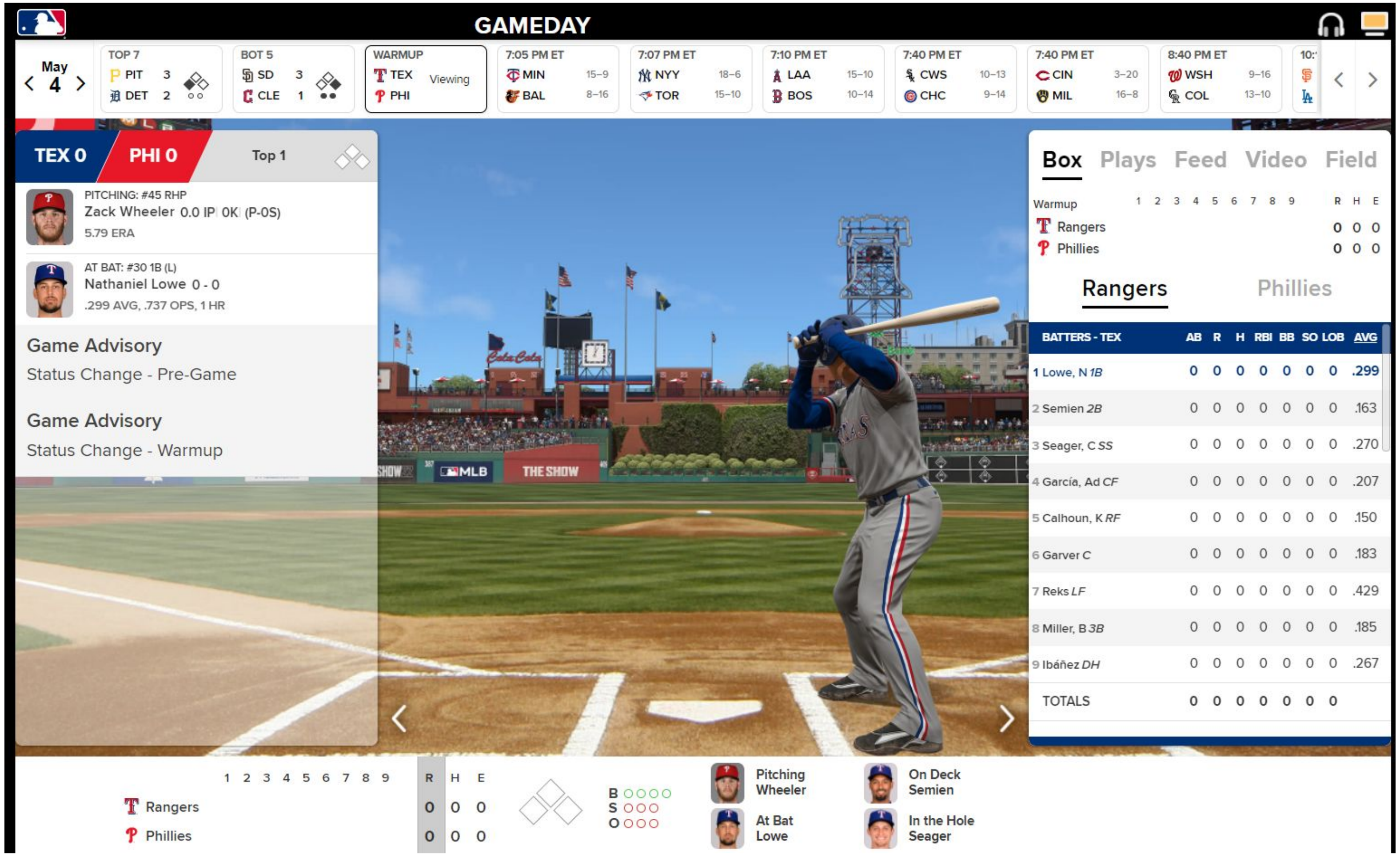
By using MLFlow within Databricks, the Texas Rangers R&D department have created a **centralized machine learning repository** to host models.

Centralizing our models across teams helped us **identify duplicated models** as well as provide a **constant source of truth**. One model for pitch evaluation, strike probability, or hit effectiveness could be used by everyone, across player development, advance reporting, and amateur.

These models can be integrated into our unified data pipeline.

Big Data Discovery

Unified Streaming Platform

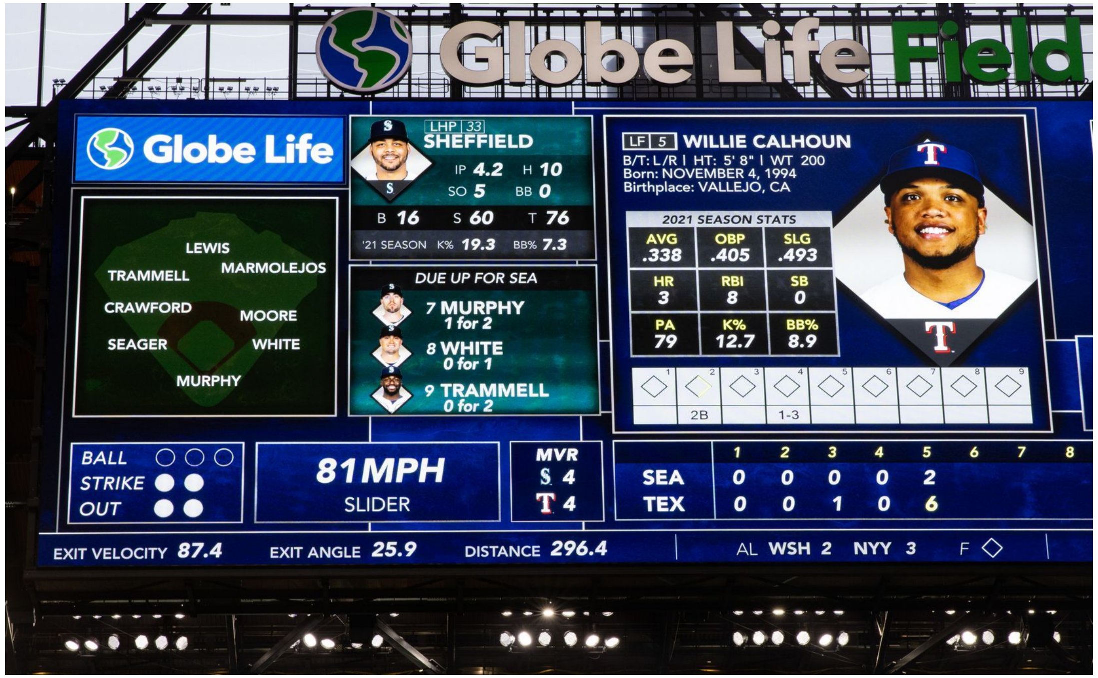


Big Data Discovery

Unified Streaming Platform

During games, bullpens, batting practice, and other data generating events, tracked pitch information can be streamed.

Think about the numbers that you hear during a modern broadcast. Exit velo, horizontal movement, sprint speed. We receive this information as it happens.



DATA+AI SUMMIT 2022

Big Data Discovery

Unified Streaming Platform

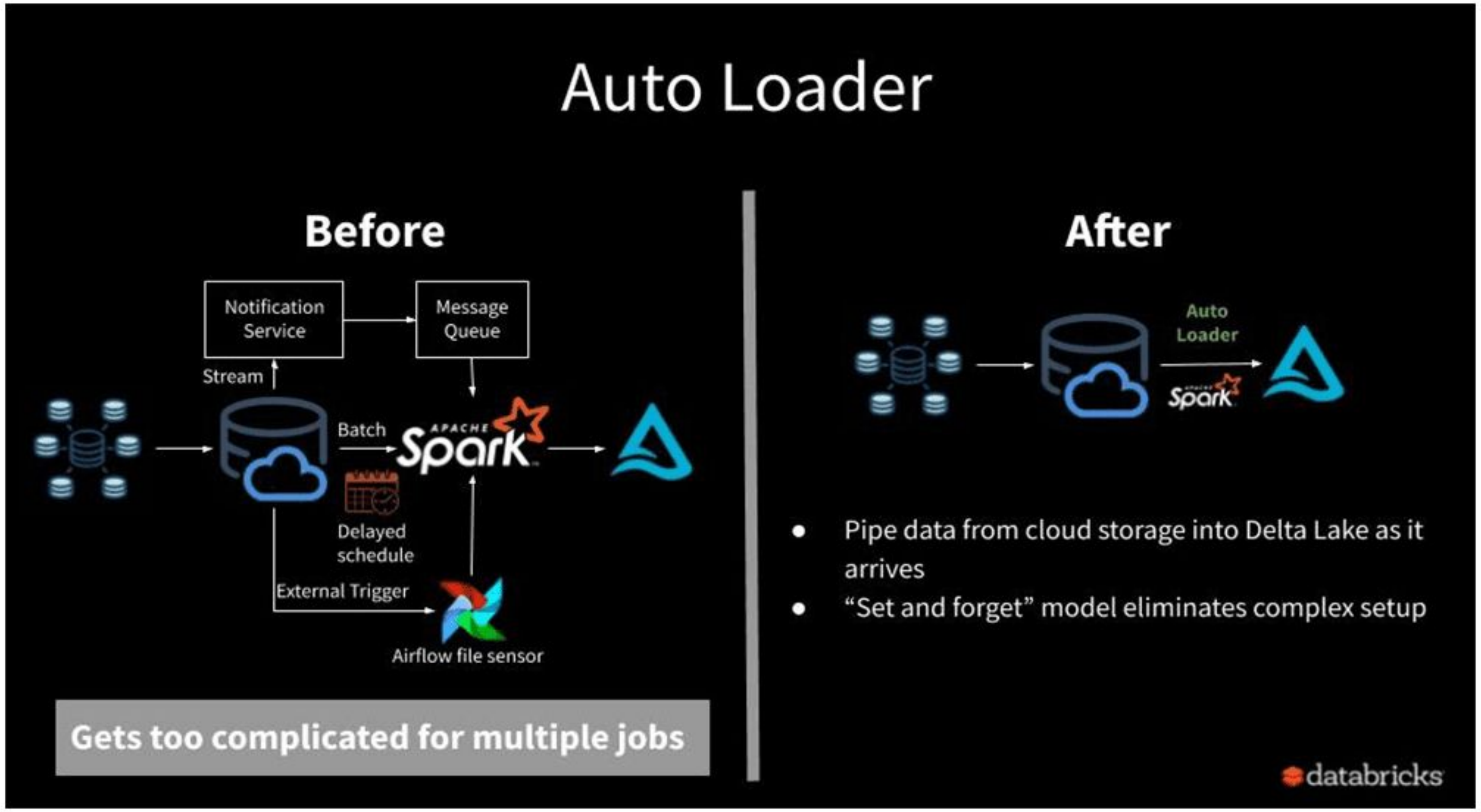
“Auto Loader is an optimized cloud file source for Apache Spark that loads data continuously and efficiently from cloud storage as new data arrives”

Prakash Chockalingam

Databricks Engineering Blog

Big Data Discovery

Unified Streaming Platform

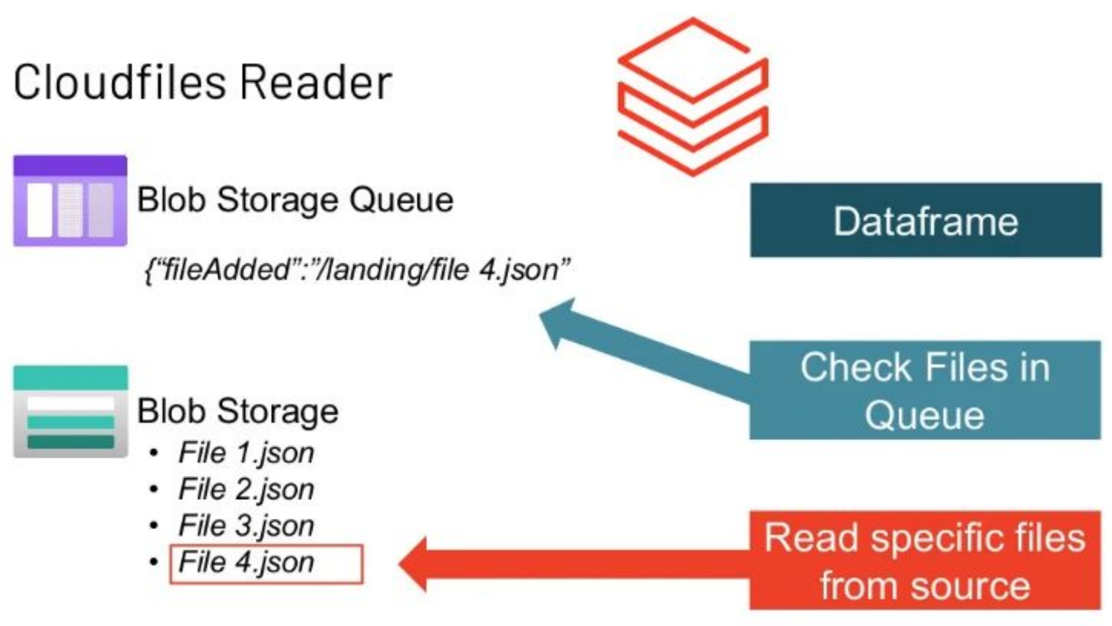


Big Data Discovery

Unified Streaming Platform

Our live data originates from API sources in JSON format. Other streaming data comes through as CSVs.

With Autoloader, we can put together a script to load these files into Cloud Storage, where they are then scored and pulled automatically into our data lake.



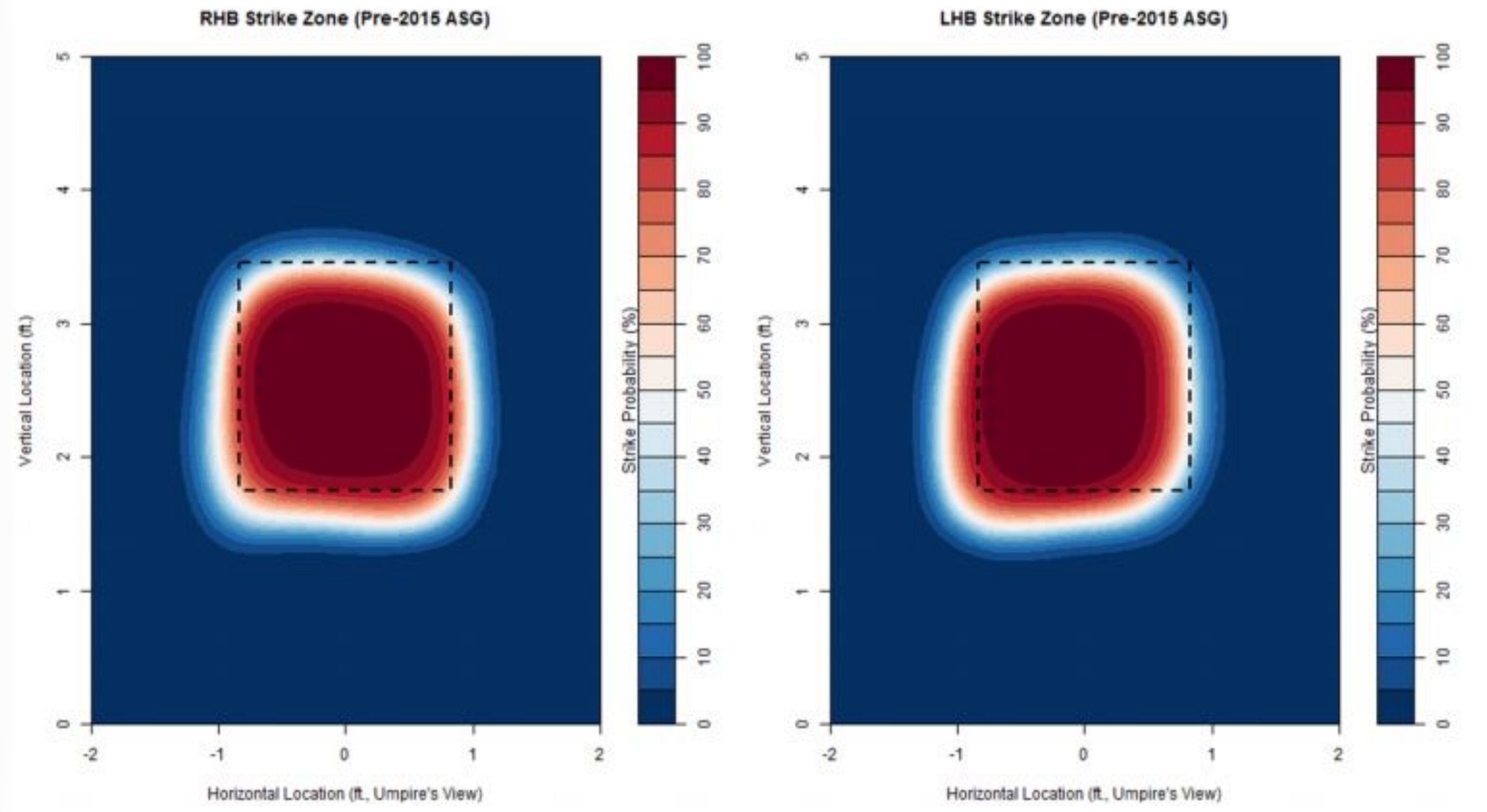
Big Data Discovery

Unified Streaming Platform

This streamed data can be predicted using models hosted in MLFlow.

Example:

By combining MLFlow and AutoLoader, we can visualize the current umpire's strike zone using a strike probability model in real time.



Case Study

The New Science of Hitting

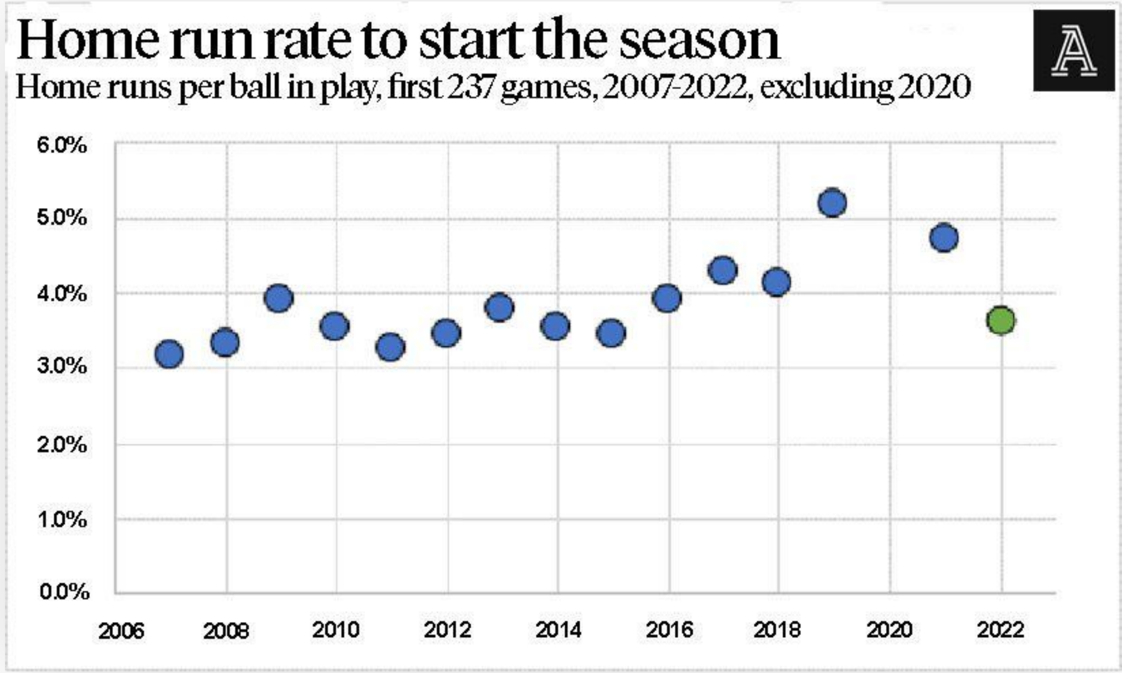
Case Study

The New Science of Hitting

In 2017, home run rates started to skyrocket across the league.

Hitters were quoted as trying to optimize specific launch angle and exit velocity combinations, to achieve "barrels".

How can we tell this story using data?



Case Study

The New Science of Hitting

Using Spark and a python library called PyBaseball, we can bring in 1.8 million tracked pitches since the 2019 season. 300,000 hits were recorded from this data.

We can use this data to predict a hit probability.

```python
data = pybaseball.statcast(start_dt='2019-03-28', end_dt='2021-10-04')
print(f"Pitches: {len(data)}")
batters = data.batter.unique()
print(f"Batters: {len(batters)}")

# get and merge batter names
player_batter = pybaseball.playerid_reverse_lookup(batters)
player_batter_merge = player_batter.loc[:, ["name_last", "name_first", "key_mlbam"]]
player_batter_merge.columns = ["batter_name_last", "batter_name_first", "batter"]

data = pd.merge(data, player_batter_merge, on="batter")
data["batter_name"] = data["batter_name_last"].apply(lambda x: x.title()) + ", " + data["batter_name_first"].apply(lambda x: x.title())

# spray angle
data = add_spray_angle(data)
```

```
This is a large query, it may take a moment to complete
Skipping offseason dates
Skipping offseason dates
100%|██████████| 518/518 [03:09<00:00,  2.74it/s]
Pitches: 1774947
Batters: 1657
Gathering player lookup table. This may take a moment.
```

Case Study

The New Science of Hitting

Features:

- Hit Launch Angle
- Hit Exit Speed
- Hit Spray Angle
- Infield Positioning
- Outfield Positioning
- Batter Handedness
- Pitcher Handedness

	launch_angle	launch_speed	spray_angle	if_fielding_alignment	of_fielding_alignment	batter_stance	pitcher_throws
0	-13	95.2	-36.075133	Infield shift	Strategic	R	L
1	44	71.3	-27.056485	Standard	Standard	L	R
2	67	92.0	19.547047	Standard	Standard	L	R
3	27	94.3	-5.058637	Standard	Standard	L	R
4	31	101.3	1.433824	Standard	Standard	L	R

Case Study

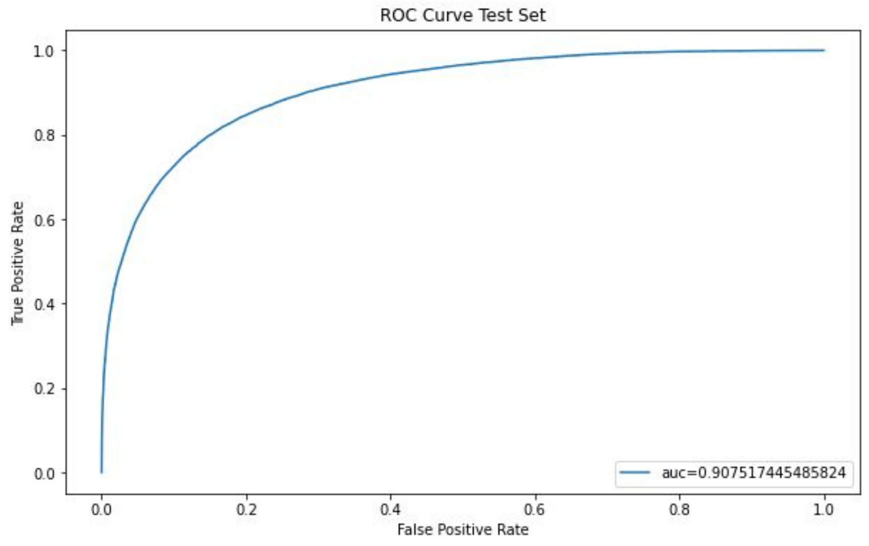
The New Science of Hitting

After performing one-hot encoding on the categorical variables, we split the data into a 75/25 train-test split.

An XGBoost Classifier was trained on this input data and registered with MLFlow.

This model now predicts hit probability.

	precision	recall	f1-score	support
0	0.86	0.90	0.88	48448
1	0.80	0.72	0.76	25390
accuracy			0.84	73838
macro avg	0.83	0.81	0.82	73838
weighted avg	0.84	0.84	0.84	73838



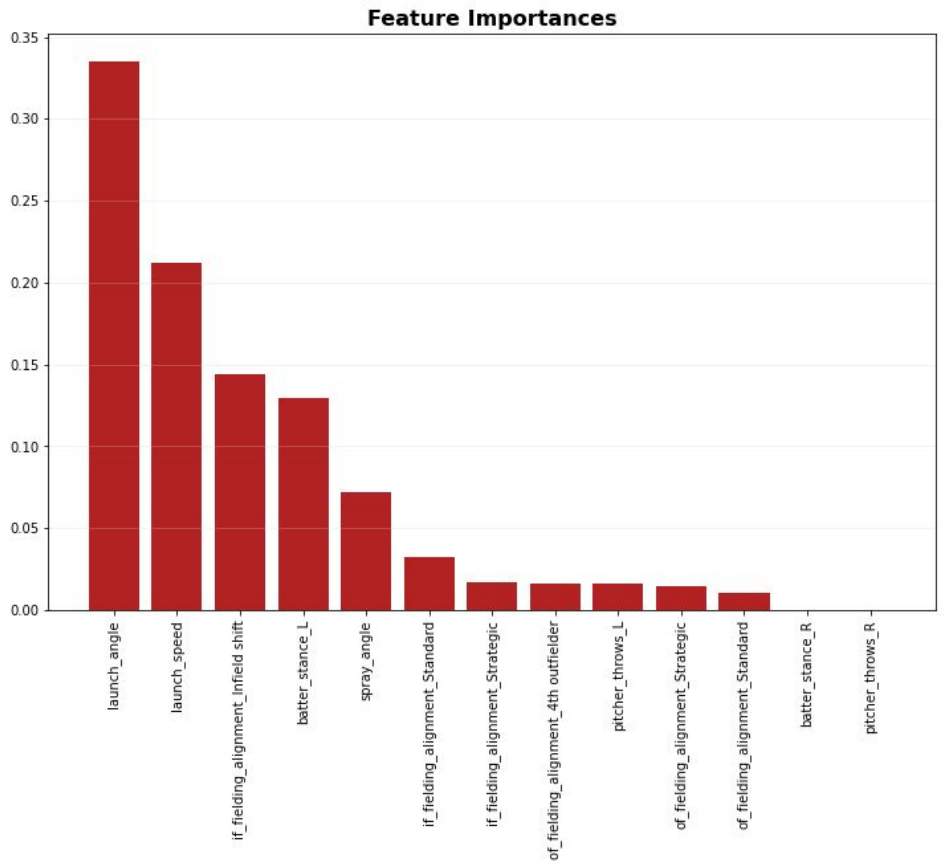
Case Study

The New Science of Hitting

Launch Angle and Exit Velocity are the two most important features.

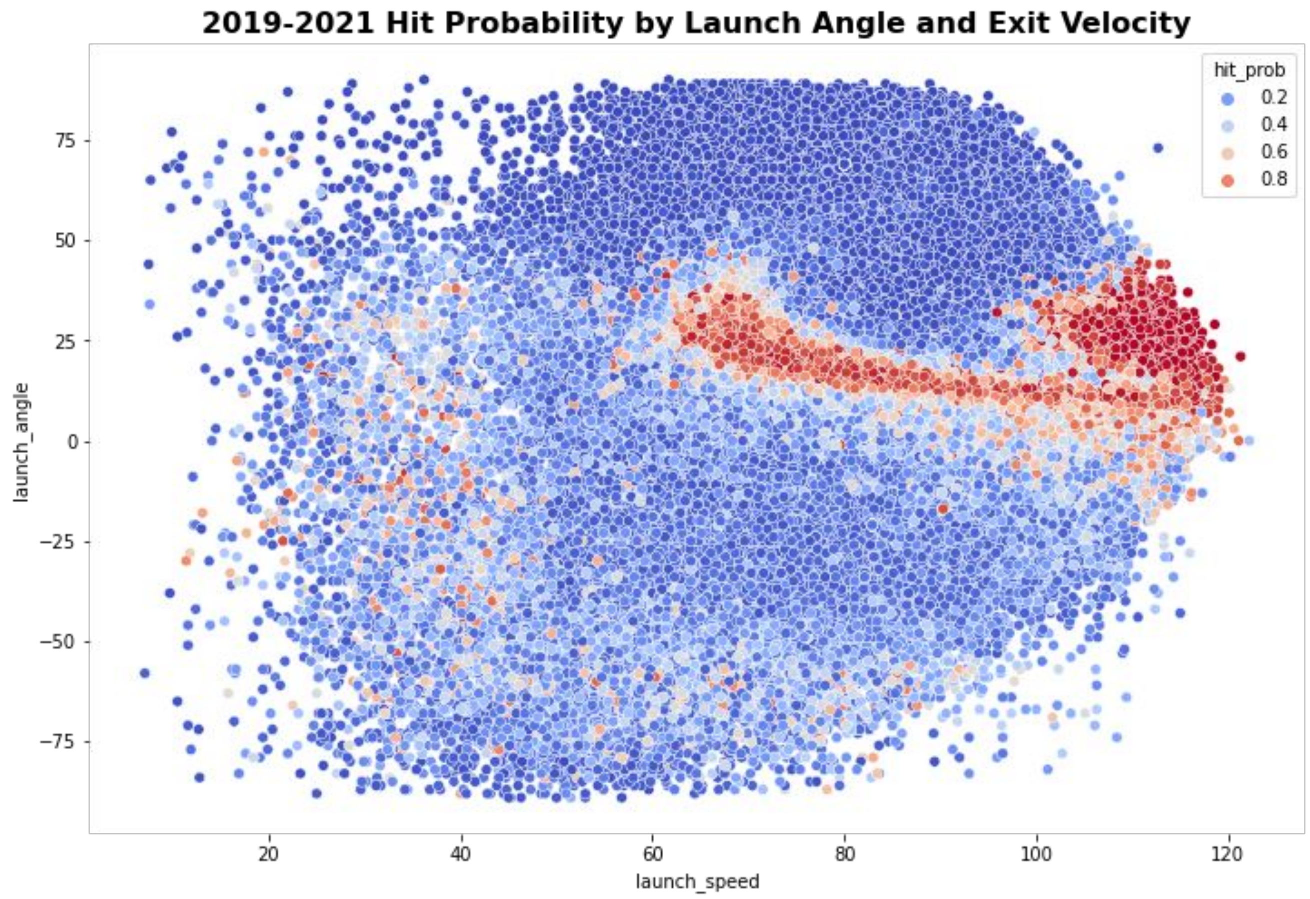
However, our model also detected the **significance of the shift**, especially coupled with a left-handed hitter.

MLB is exploring banning the shift next season to increase the probability of a hit.



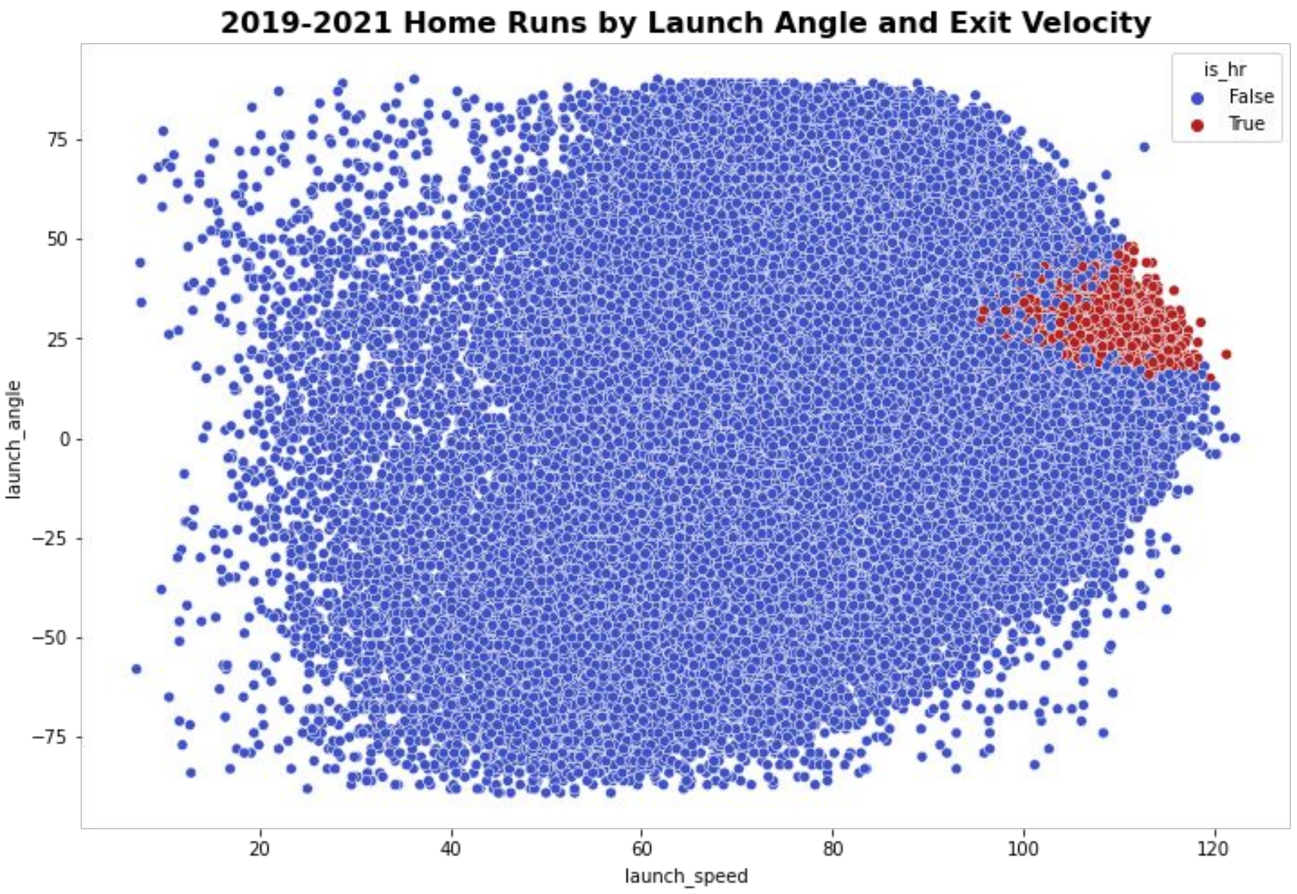
Case Study

The New Science of Hitting



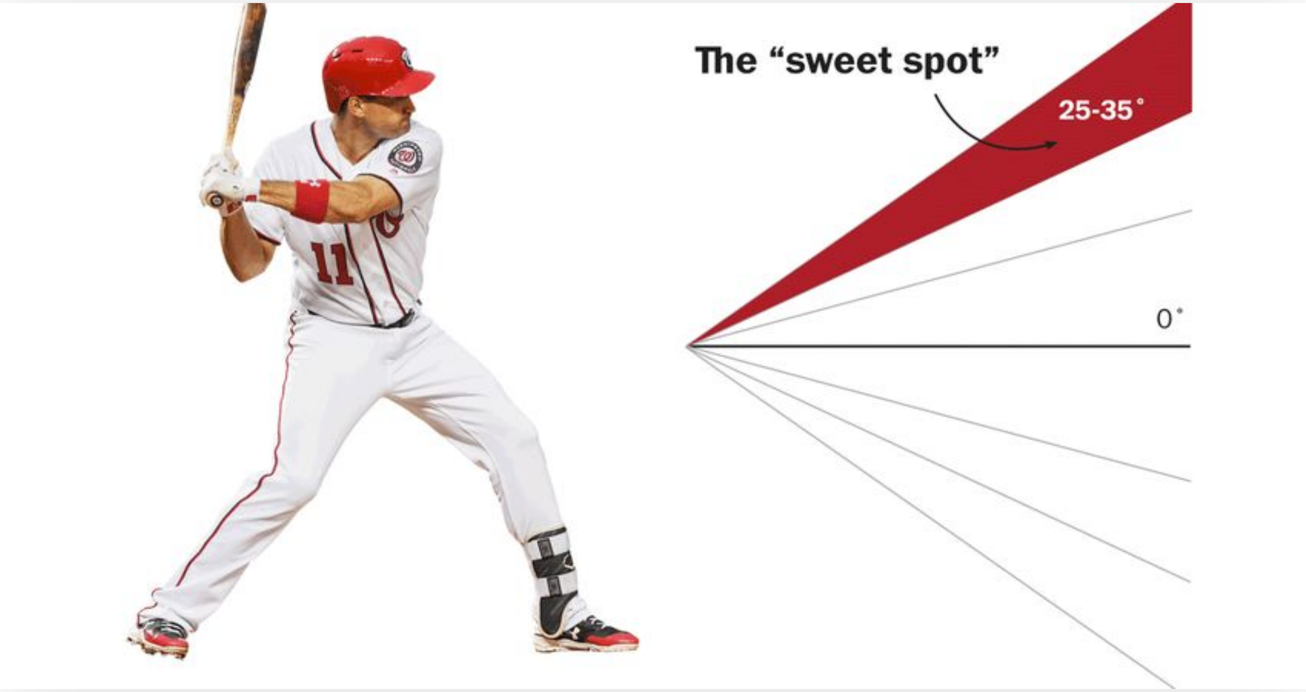
Case Study

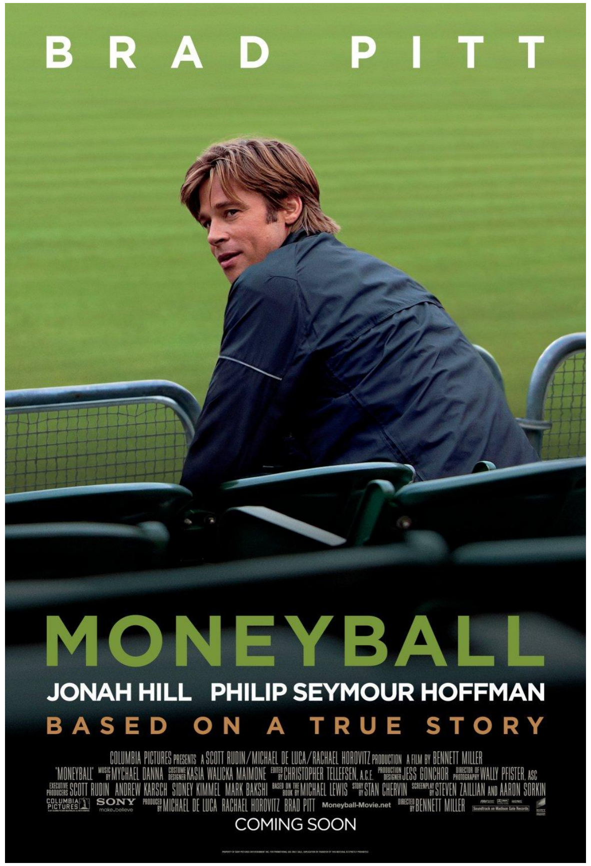
The New Science of Hitting



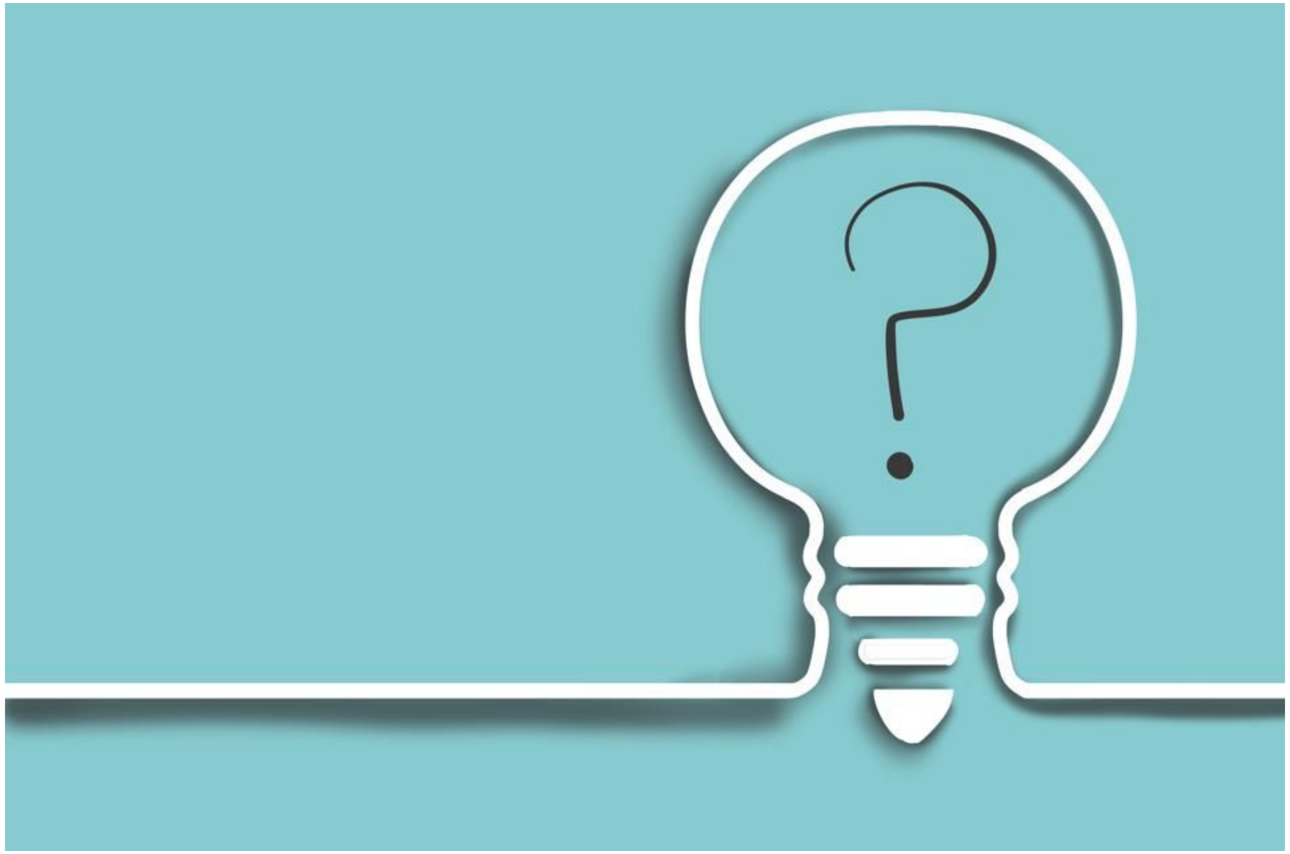
Case Study

"You get on base, we win. You don't, we lose. And I hate losing." - Brad Pitt/Billy Beane





Questions



References

- https://www.mlb.com/news/hawk-eye-statcast-pose-tracking-the-best-2020-postseason-moments
- https://en.wikipedia.org/wiki/Statcast
- https://databricks.com/session_na21/accelerating-data-ingestion-with-databricks-autoloader
- https://tht.fangraphs.com/are-the-umpires-at-it-again/
- https://theathletic.com/3272450/2022/04/26/baseballs-arent-flying-as-far-and-home-runs-are-down-across-mlb-is-it-the-ball-itself/
- https://github.com/jldbc/pybaseball
- https://databricks.com/product/managed-mlflow
- https://databricks.com/blog/2020/02/24/introducing-databricks-ingest-easy-data-ingestion-into-delta-lake.html

DATA+AI SUMMIT 2022

Thank you

Alexander Booth

Senior Analyst, Texas Rangers

Ryan Stoll

Data Engineer, Texas Rangers