Hit a Home Run

Discover and operationalize data-analytics for winning

Artificial Intelligence and Machine Learning

IBM + aginity + H_2O.ai

The Moneyball All-Stars

David Kearns
Data Science

@IBMAnalytics

Ari Kaplan
Mr. Moneyball

@Aginity

Chris Coad
Story Teller

@Aginity

Erin LeDell
Chief ML Scientist

@H20.ai

IBM + aginity + H_2O.ai

What is Moneyball?

Moneyball is the practice of using data, analytics, artificial intelligence and machine learning to find the best baseball players and build the optimal team to win championships and increase organizational revenue

Organizational AI Challenges

Data-related challenges are hindering 96% of organizations from taking full advantage of AI

Source: InfoWorld

Organizational AI Challenges

80% of organizations face collaboration challenges due to silos

Source: InfoWorld

Organizational AI Challenges

Data systems don’t “do AI” and AI technologies don’t “do data.” Organizations use 7 disparate tools

Source: InfoWorld

How Moneyball Relates to Business?

- Diverse datasets living in different systems and formats
- Collaboration across consistent data and analytics to power predictive insights and capabilities
- AI & ML is built on engineered features and “analytics-ready” data
- Technology drives more data and competition drives need for innovation and fast and clear ways to consume these insights

Moneyball Project Goals

- Create a predictive player performance model
- Work with a distributed set of data sources in different data repositories
- Build consistent data and derived attributes (analytics) across that can be shared across the world
- Share domain expertise
- Operationalize ML/AI

Combine Public and Proprietary Databases

Lahman Database

- Public database from 1871 to 2017
- Aggregate pitching and batting statistics

Attribute	Description
playerID	Player ID code
AB	At Bats
R	Runs
H	Hits
SO	Strike Outs
+ More	…

http://www.seanlahman.com/baseball-archive/statistics/

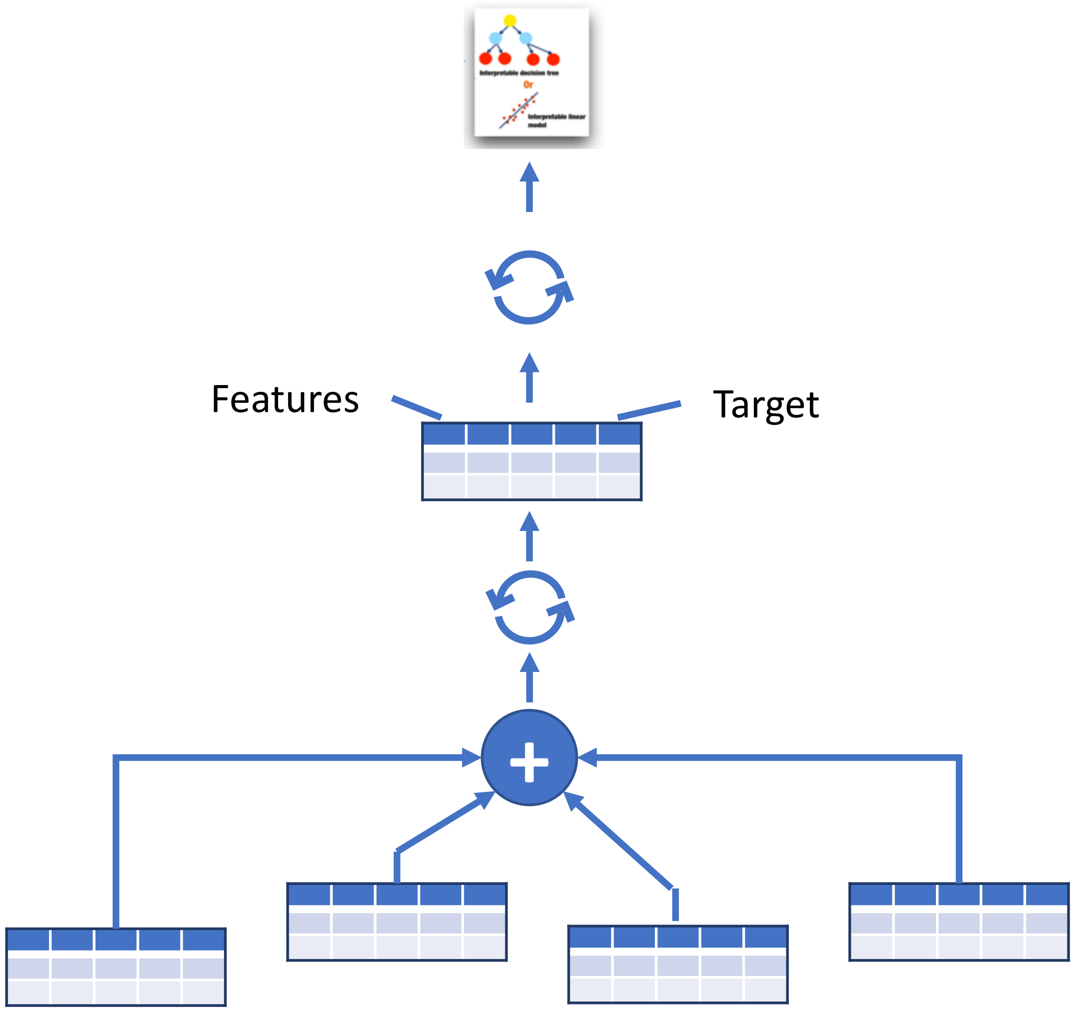
Ari Database

- Private database from 2012-2017
- Pitch-by-pitch play for each MLB game

Attribute	Description
Pitch_Type	Two - character code of type of pitch
Spin_rate	Spin of the pitch in rotations per minute.
Start_speed	The velocity of the pitch in MPH
End_speed	The velocity of the pitch when it arrives at the plate in MPH
Spray_des	Classification of type of hit
+ More	…

IBM + aginity + H2O.ai

The Manual Way…

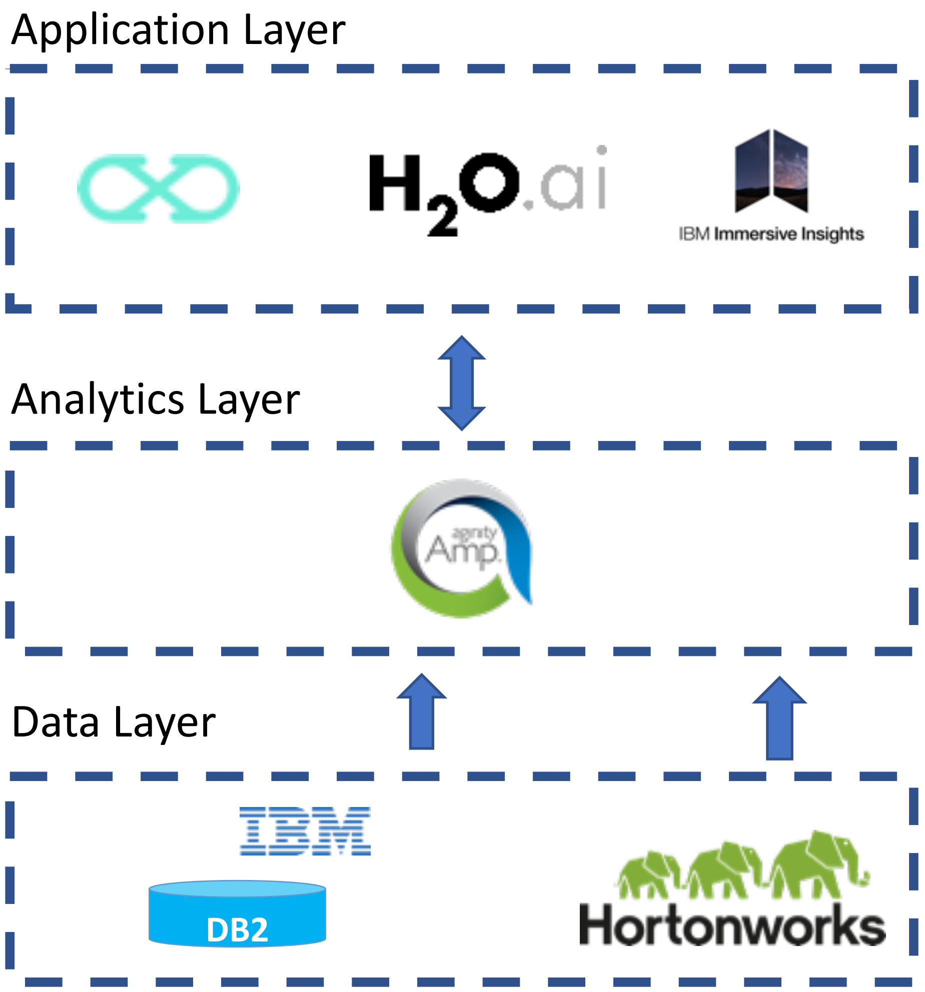


Phase	Challenge
Model	Limited traceability and full picture of how it was built
Model Building	Manual model tests & specific expertise
Modeling Table	
Data Quality & Transformation	Manual data prep, re-coding & no model reuse
Data Integration	Disparate data sources & small data sets

IBM + aginity + H2O.ai

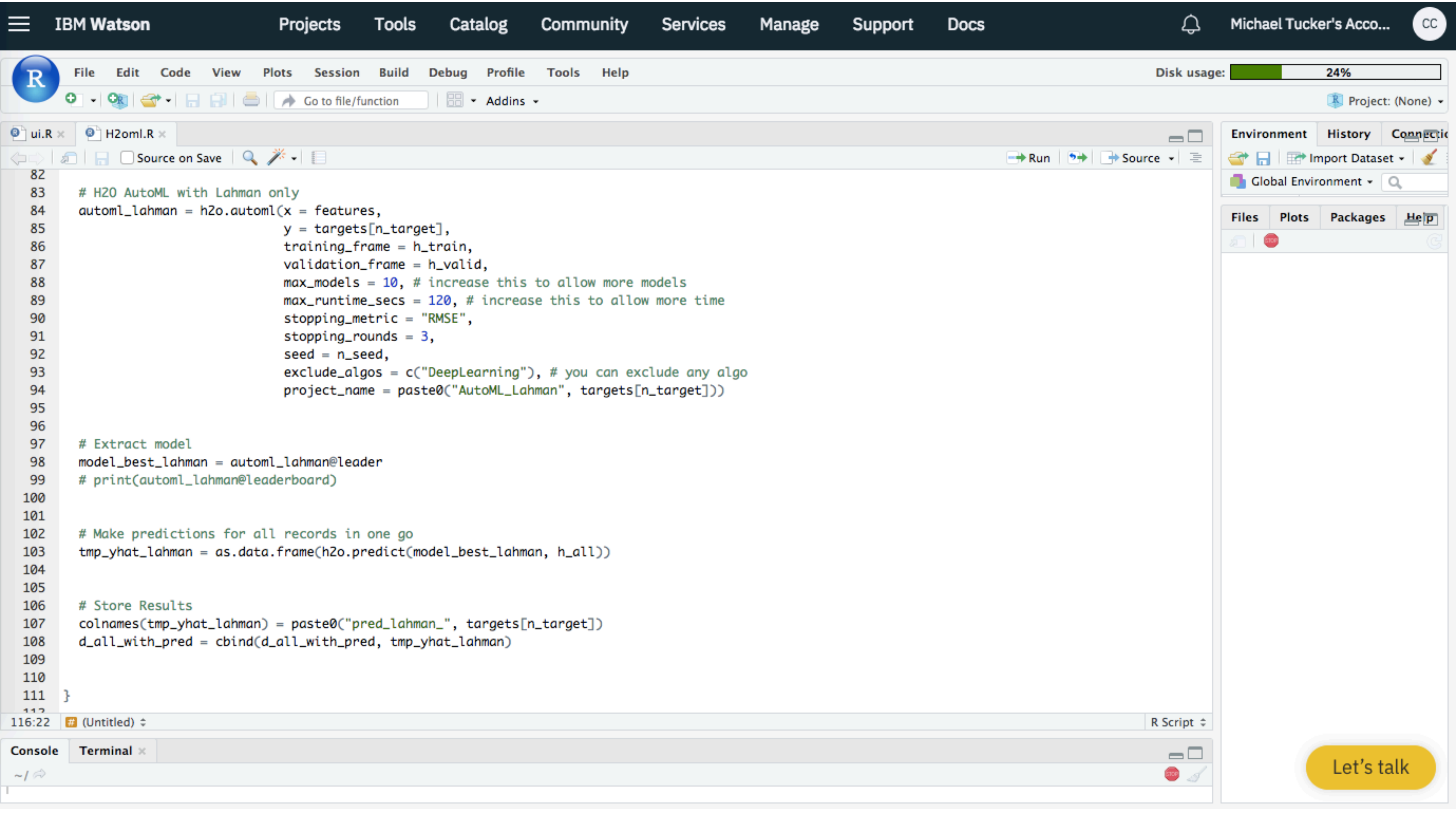
To the Enterprise Way: IBM + Aginity + H2O

Challenge	Solve
Limited traceability and full picture of how it was built	Who, what, when and how the analytic was created
Manual model tests & specific expertise	AutoML
Manual data prep, re-coding & no model reuse	Re-usable analytics operationalized and universally shareable
Disparate data sources & small data sets	Rich data sets in from multiple enterprise data solutions

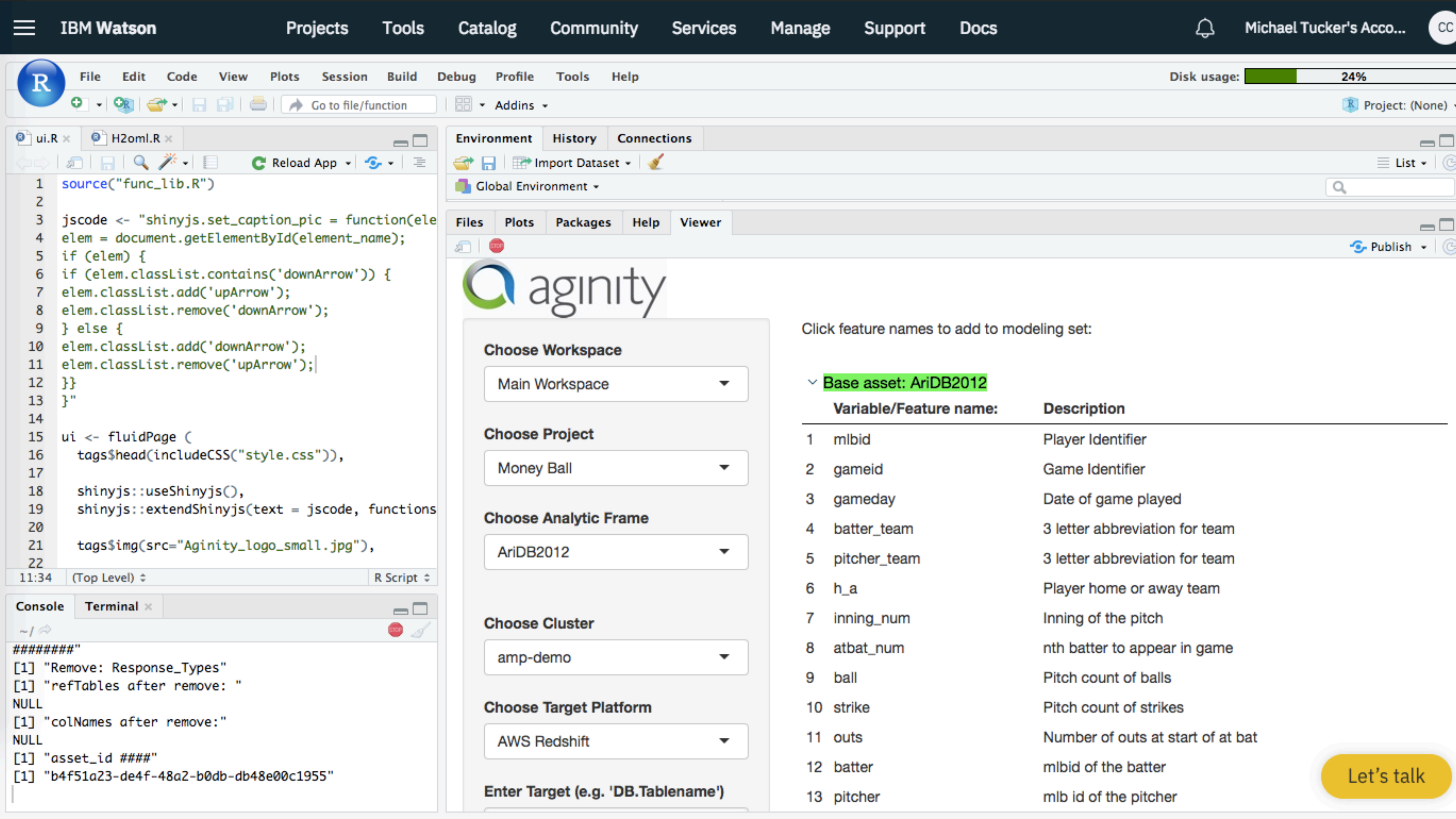


IBM + aginity + H2O.ai

Demos



```

  # H2O AutoML with Lahman only
  automl_lahman = h2o.automl(x = features,
                             y = targets[n_target],
                             training_frame = h_train,
                             validation_frame = h_valid,
                             max_models = 10, # increase this to allow more models
                             max_runtime_secs = 120, # increase this to allow more time
                             stopping_metric = "RMSE",
                             stopping_rounds = 3,
                             seed = n_seed,
                             exclude_algos = c("DeepLearning"), # you can exclude any algo
                             project_name = paste0("AutoML_Lahman", targets[n_target]))


  # Extract model
  model_best_lahman = automl_lahman@leader
  # print(automl_lahman@leaderboard)


  # Make predictions for all records in one go
  tmp_yhat_lahman = as.data.frame(h2o.predict(model_best_lahman, h_all))


  # Store Results
  colnames(tmp_yhat_lahman) = paste0("pred_lahman_", targets[n_target])
  d_all_with_pred = cbind(d_all_with_pred, tmp_yhat_lahman)


}
```



```
source("func_lib.R")

jscode <- "shinyjs.set_caption_pic = function(ele
elem = document.getElementById(element_name);
if (elem) {
if (elem.classList.contains('downArrow')) {
elem.classList.add('upArrow');
elem.classList.remove('downArrow');
} else {
elem.classList.add('downArrow');
elem.classList.remove('upArrow');
}}
}"

ui <- fluidPage (
  tags$head(includeCSS("style.css")),

  shinyjs::useShinyjs(),
  shinyjs::extendShinyjs(text = jscode, functions

  tags$img(src="Aginity_logo_small.jpg"),
```

```
########"
[1] "Remove: Response_Types"
[1] "refTables after remove: "
NULL
[1] "colNames after remove:"
NULL
[1] "asset_id ####"
[1] "b4f51a23-de4f-48a2-b0db-db48e00c1955"
```

aginity

Choose Workspace

Main Workspace

Choose Project

Money Ball

Choose Analytic Frame

AriDB2012

Choose Cluster

amp-demo

Choose Target Platform

AWS Redshift

Enter Target (e.g. 'DB.Tablename')

Click feature names to add to modeling set:

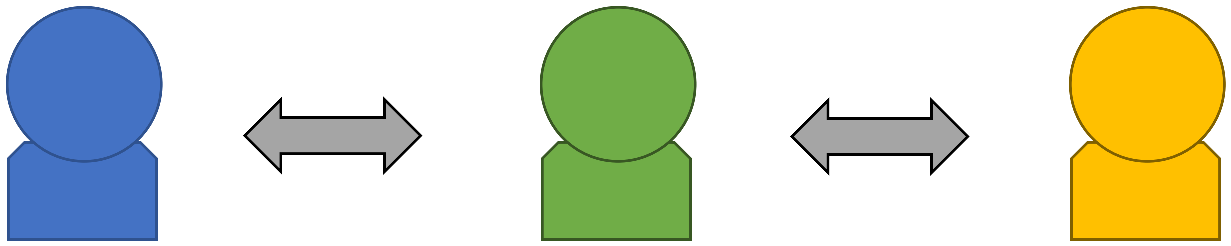
Base asset: AriDB2012

	Variable/Feature name:	Description
1	mlbid	Player Identifier
2	gameid	Game Identifier
3	gameday	Date of game played
4	batter_team	3 letter abbreviation for team
5	pitcher_team	3 letter abbreviation for team
6	h_a	Player home or away team
7	inning_num	Inning of the pitch
8	atbat_num	nth batter to appear in game
9	ball	Pitch count of balls
10	strike	Pitch count of strikes
11	outs	Number of outs at start of at bat
12	batter	mlbid of the batter
13	pitcher	mlb id of the pitcher

Key Findings

1. Diverse Team with Specific Skills

- Important have a team with a wide variety of skills so the full context of what you're trying to accomplish is understood
- Provide the context first



Domain Expert

Project Manager

Data Scientist

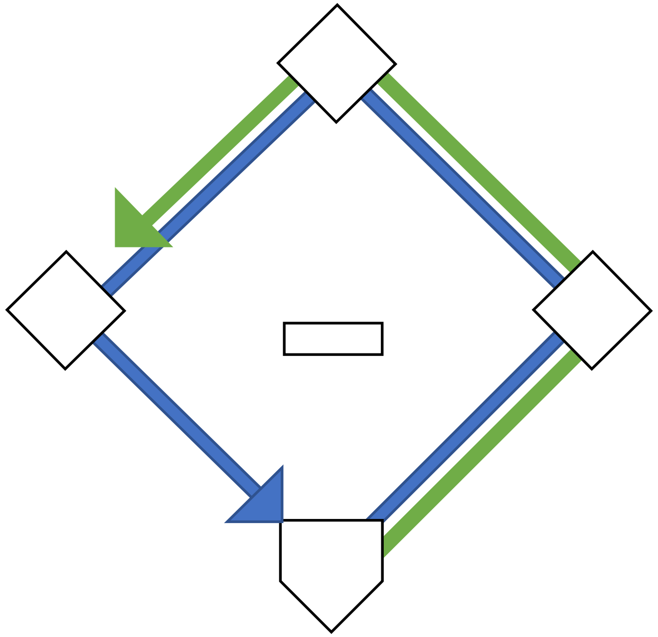
IBM + aginity + H_2O.ai

2. Encapsulate Domain Expert Knowledge

- Who do you constantly rely for domain expertise?
- How can you encapsulate their knowledge…
- …Make it easily accessible to the rest of the team, department or enterprise?

IBM + aginity + H2O.ai

3. Similar Analytics Across Use Cases

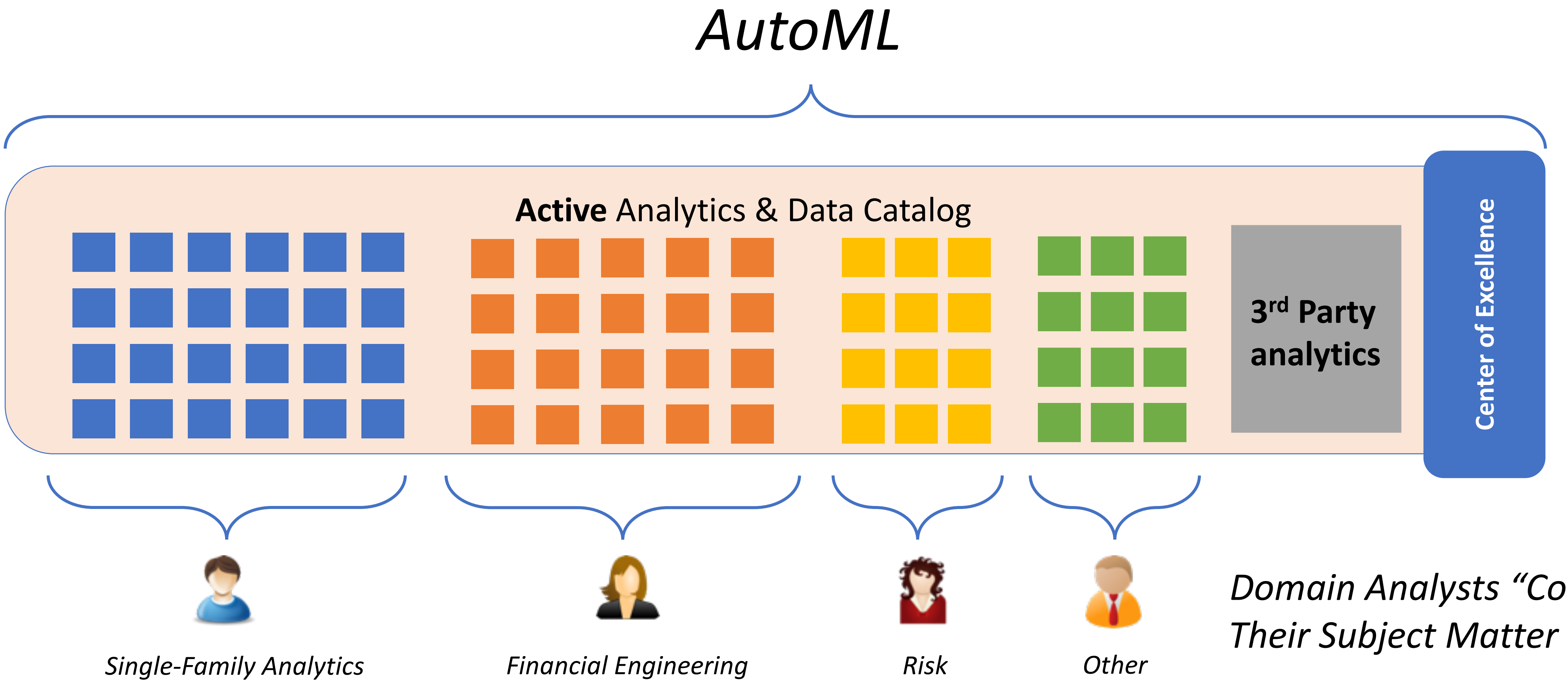


80% Of enterprise analytics **are highly similar.**

100% Of enterprise analytics **are recreated.**

IBM + aginity + H2O.ai

4. Enterprise-width Features Enables AI/ML



AutoML

Active Analytics & Data Catalog

3rd Party analytics

Center of Excellence

Single-Family Analytics

Financial Engineering

Risk

Other

Domain Analysts "Contribute" Their Subject Matter Expertise

IBM + aginity + H_2O.ai

Final Project Result

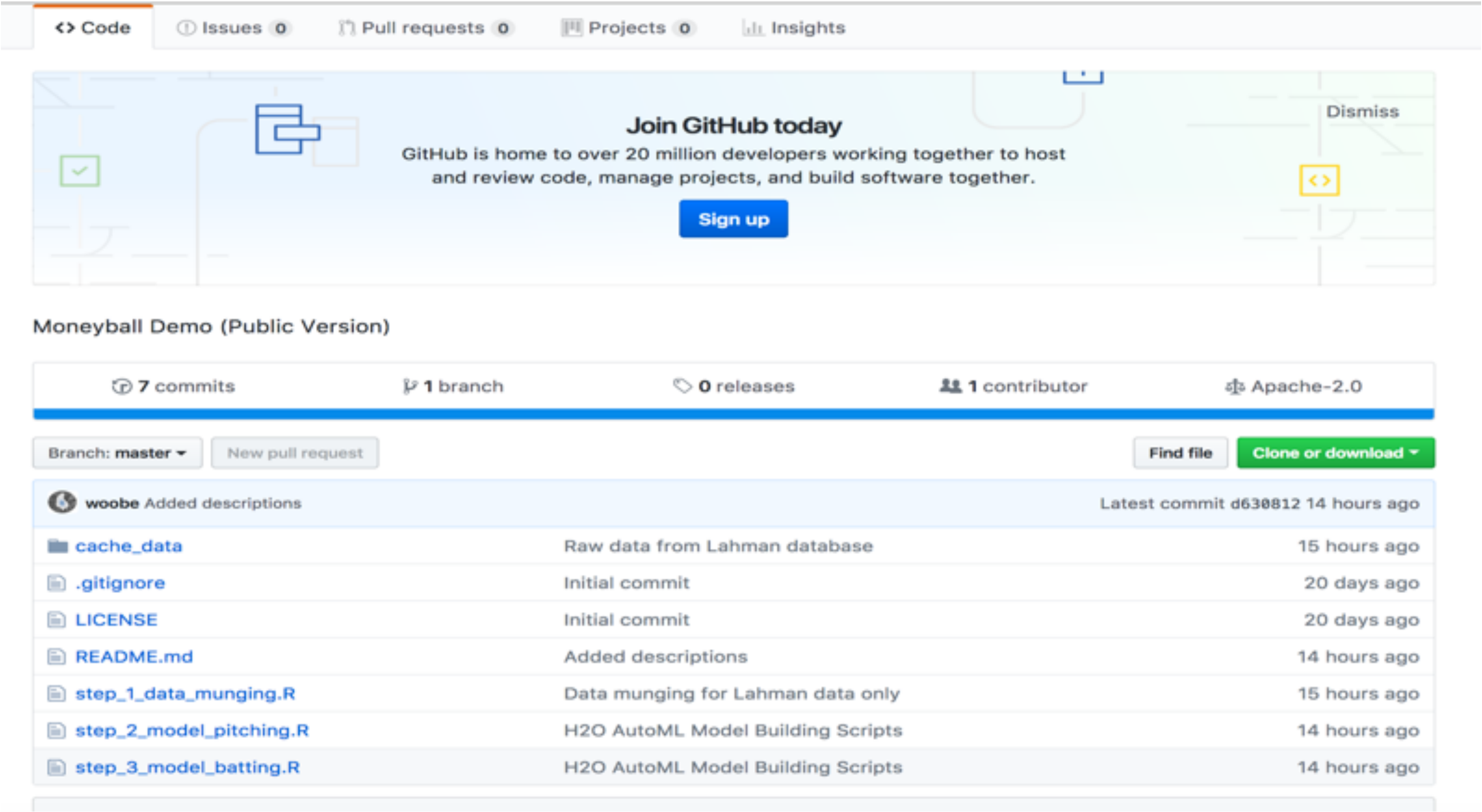
$20M

trade 2 weeks prior to the season beginning

IBM + aginity + H₂O.ai

Try it Yourself!

GitHub: https://github.com/woobe/moneyball



Join GitHub today

GitHub is home to over 20 million developers working together to host and review code, manage projects, and build software together.

Moneyball Demo (Public Version)

7 commits | 1 branch | 0 releases | 1 contributor | Apache-2.0

woobe Added descriptions — Latest commit d630812 14 hours ago

cache_data	Raw data from Lahman database	15 hours ago
.gitignore	Initial commit	20 days ago
LICENSE	Initial commit	20 days ago
README.md	Added descriptions	14 hours ago
step_1_data_munging.R	Data munging for Lahman data only	15 hours ago
step_2_model_pitching.R	H2O AutoML Model Building Scripts	14 hours ago
step_3_model_batting.R	H2O AutoML Model Building Scripts	14 hours ago

IBM + aginity + H2O.ai

Get in Touch with Us

David Kearns
Data Science

@DaithiOCiaran

Ari Kaplan
Mr. Moneyball

@arikaplan

Chris Coad
Story Teller

@chriscoad

Erin LeDell
Chief ML Scientist

@ledell

IBM + aginity + H_2O.ai

Sports Analytics Podcasts

- Selfish plug (in case you dozed off)
  - http://www.ibmbigdatahub.com/podcast/making-data-simple-hit-home-run-using-ai-machine-learning
- New York Yankees on injury prevention
  - https://itunes.apple.com/gb/podcast/4-optimising-player-training-treatment-strategies-in/id1327803354?i=1000411908873&mt=2
- ML Sports-focused Series
  - https://twimlai.com/aiinsports2018/

Find Out More About Our Products

- ICP for Data
  - https://www.ibm.com/analytics/cloud-private-for-data
- Watson Studio
  - https://www.ibm.com/cloud/watson-studio

- Aginity Amp
  - https://www.aginity.com/main/products/
- Aginity Workbench
  - https://www.aginity.com/main/workbench/

- AutoML
  - https://www.h2o.ai/products/h2o/
- Driverless AI
  - https://www.h2o.ai/products/h2o-driverless-ai/