H_2O

WORLD

2017

Matt Dancho

Founder / CEO of ***Business Science***

Twitter: @bizScienc // @mdancho84

Github: business-science // mdancho84

Email: mdancho@business-science.io

Website: www.business-science.io

About Business Science

Business & Finance Consultancy

- Executive Leadership
- Bolt-on **Data Science** Team

Education & Tools

- Blog: **www.business-science.io**
- Open Source Software
- Courses: Coming soon!!!

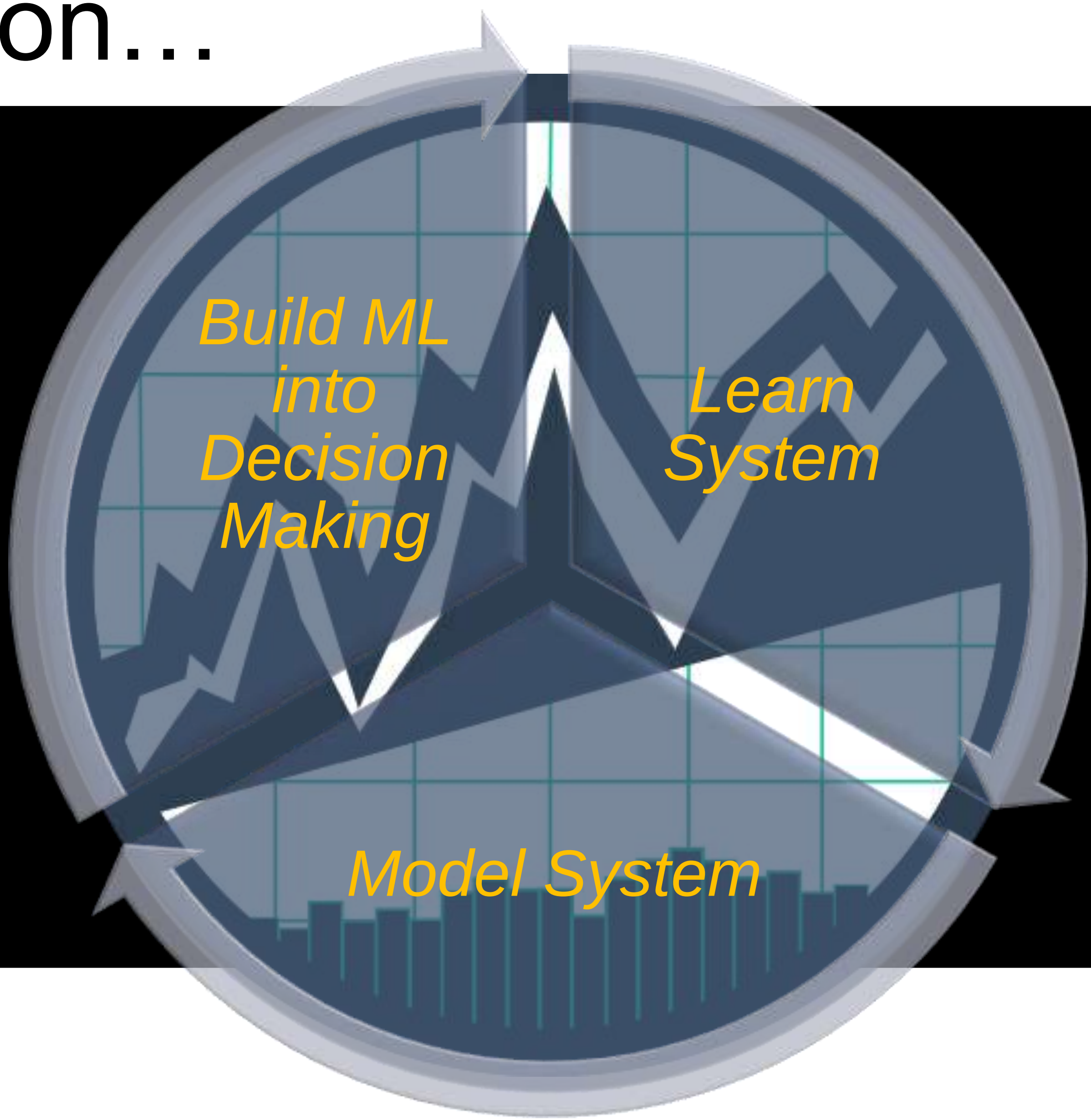
Community Driven

- Powered by

Adaptive Solution…

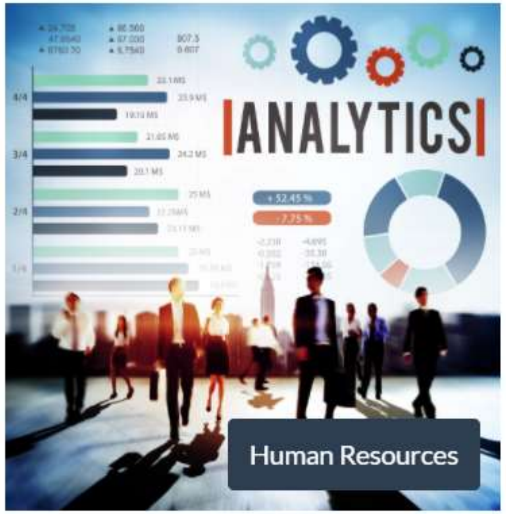
ML + Leadership = Good Decision Making

AI: Learning Built Into System



…Applied To Any Problem

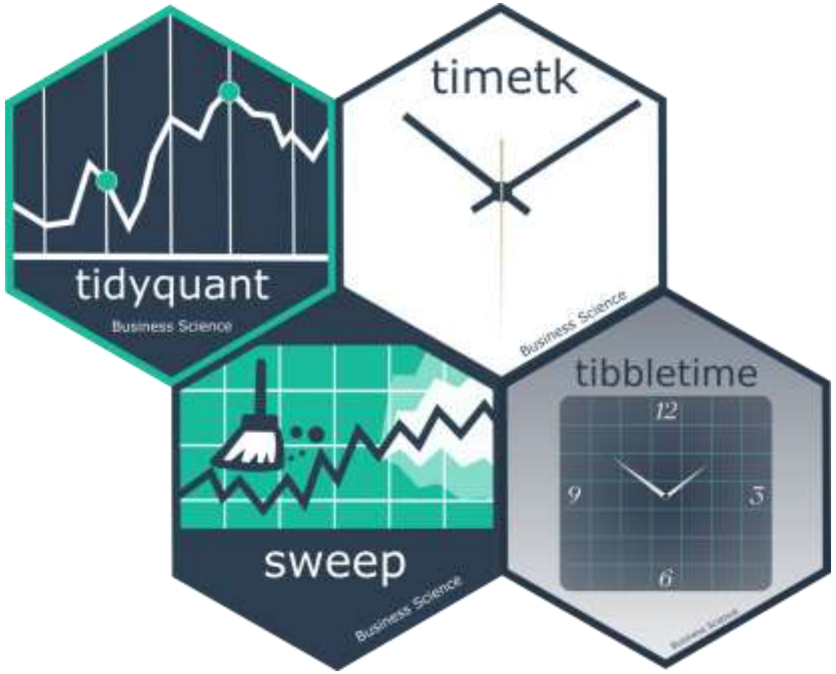
OUR EXPERTISE



Business Science: Not Just For Executives

Data Scientists

- **Open Source Software**



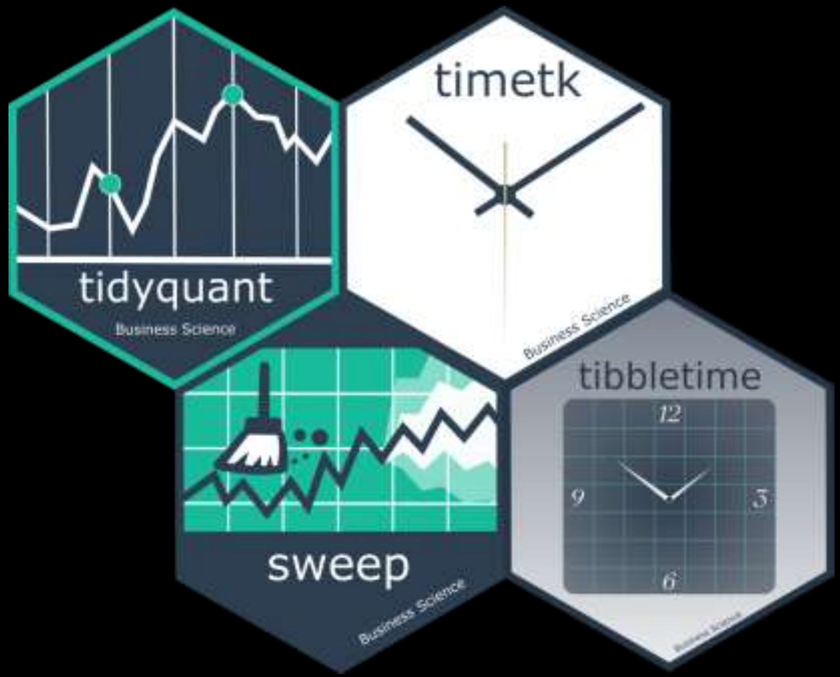
- **Courses (Coming Soon!)**
  - High Demand Applications
  - High Demand Tools
  - Integrated Solutions

H_2O WORLD 2017

Business Science: Not Just For Executives

Data Scientists

- **Open Source Software**

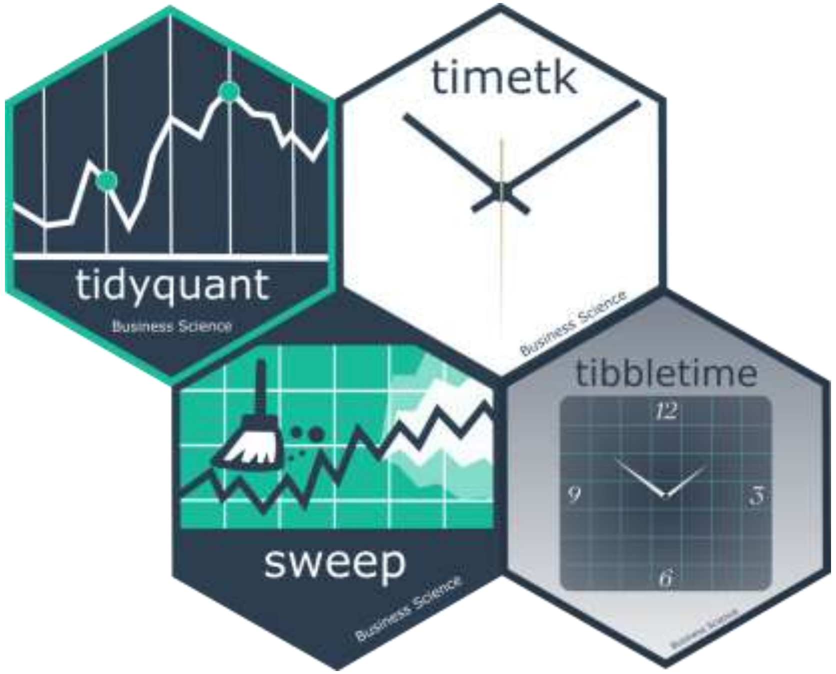


- **Courses (Coming Soon!)**
  - High Demand Applications
  - High Demand Tools
  - Integrated Solutions

Business Science: Not Just For Executives

Data Scientists

- **Open Source Software**



- **Courses (Coming Soon!)**
  - High Demand Applications
  - High Demand Tools
  - Integrated Solutions

Business Science: Not Just For Executives

Data Scientists

- **Open Source Software**

- **Courses (Coming Soon!)**
  - High Demand Applications

Learn from our blog

www.business-science.io/blog

Business + Data Science

www.business-science.io

HR Analytics

Using ML To Predict
Employee Turnover

Three Reasons

1. Employee Attrition: **A HUGE PROBLEM**
2. New Techniques To **Predict** & **Explain** Turnover
3. **Framework** For ML In Business Applications

Three Reasons

1. Employee Attrition: **A HUGE PROBLEM**

2. New Techniques To **Predict** & **Explain** Turnover

3. **Framework** For ML In Business Applications

Three Reasons

1. Employee Attrition: **A HUGE PROBLEM**

2. New Techniques To **Predict** & **Explain** Turnover

3. **Framework** For ML In Business Applications

Three Reasons

1. Employee Attrition: **A HUGE PROBLEM**

2. New Techniques To **Predict** & **Explain** Turnover

3. **Framework** For ML In Business Applications

One More Reason

4. Our Article Is Popular!

Just google:

“Predict Employee Turnover”

R-Bloggers ● KDNuggets ● LinkedIn

Code available in article:

http://www.business-science.io/business/2017/09/18/hr_employee_attrition.html

Competitive Advantage

"You take away our top 20 employees and overnight we become a mediocre company."

-Bill Gates

H2O WORLD 2017

Cost Of Turnover

Organizations face huge costs resulting from employee turnover

Cost Of Turnover

Organizations face huge costs resulting from employee turnover

Most important costs are intangible

New Product Ideas

Customer Relationships

Project Management

Engineering Talent

Machine Learning Is Evolving

- **H2O**
  - **Automated Machine Learning**
  - Predict at very high accuracy
  - Complex models can’t be explained

- **LIME**
  - Used to explain ML classifiers
  - **Deep learning, stacked ensembles now explainable**

IBM Watson HR Data Set

- Simulated HR Database
- Representative of real-world data
- Used for IBM Watson Case Study

Source: https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/

Feature Set

- HR Dataset
- 35 Features
- 1,470 Observations

Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	Education
41	Yes	Travel_Rarely	1102	Sales	1	2	Life Science
49	No	Travel_Frequently	279	Research & Development	8	1	Life Science
37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other
33	No	Travel_Frequently	1392	Research & Development	3	4	Life Science
27	No	Travel_Rarely	591	Research & Development	2	1	Medical
32	No	Travel_Frequently	1005	Research &	2	2	Life Science

Modeling with H2O

- Training the model

```r
# Split data into Train/Validation/Test Sets
hr_data_h2o <- as.h2o(hr_data)

split_h2o <- h2o.splitFrame(hr_data_h2o, c(0.7, 0.15), seed = 1234 )

train_h2o <- h2o.assign(split_h2o[[1]], "train" ) # 70%
valid_h2o <- h2o.assign(split_h2o[[2]], "valid" ) # 15%
test_h2o  <- h2o.assign(split_h2o[[3]], "test" )  # 15%
```

```r
# Run the automated machine learning
automl_models_h2o <- h2o.automl(
    x = x,
    y = y,
    training_frame    = train_h2o,
    leaderboard_frame = valid_h2o,
    max_runtime_secs  = 30
    )
```

Automated ML:
-Deep Learning
-Ensembles
-GBM

Modeling with H2O

- Prediction: Test Data (Unseen)

```
# Predict on hold-out set, test_h2o
pred_h2o <- h2o.predict(object = automl_leader, newdata = test_h2o)
```

- Performance: 88% Accuracy

```
## [[1]]
## [[1]]$accuracy
## [1] 0.8767773
##
## [[1]]$misclassification_rate
## [1] 0.1232227
##
## [[1]]$recall
## [1] 0.6206897
##
## [[1]]$precision
## [1] 0.5454545
##
## [[1]]$null_error_rate
## [1] 0.7914692
```

Important for Goal

Important for Business Case

Puts Accuracy Into Perspective

Business Implications

- Recall = 62%
  - Will correctly classify those at risk of turnover 62 of 100 times
  - Critical to the business
    - 62% of at risk employees that can be targeted preemptively

- Precision = 54%
  - Will avoid incorrectly assigning “Yes” 54 of 100 times
  - Better to target incorrectly than miss
    - Should not sacrifice Recall

Understanding Drivers Is Key

Have a ***great model****, but…*

how do we ***prevent turnover****?*

LIME

- **L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanation

- Theory
  - LIME approximates model locally as logistic or linear model
  - Repeats process 5000X
  - Outputs features that are important to local models

- Result: Data Scientists Understand Why Model Predicts

LIME

- Complex classification models can now be interpreted
  - Black Box Models
  - Neural Networks, Ensembles, Random Forests

- **H2O and LIME now integrated!**
  - https://github.com/thomasp85/lime

LIME

- Step 1: Create explainer using lime()

```
# Run lime() on training set
explainer <- lime::lime(
    as.data.frame(train_h2o[,-1]),
    model           = automl_leader,
    bin_continuous = FALSE)
```

Create explainer object

LIME

- Step 2: Create explanation using explain()

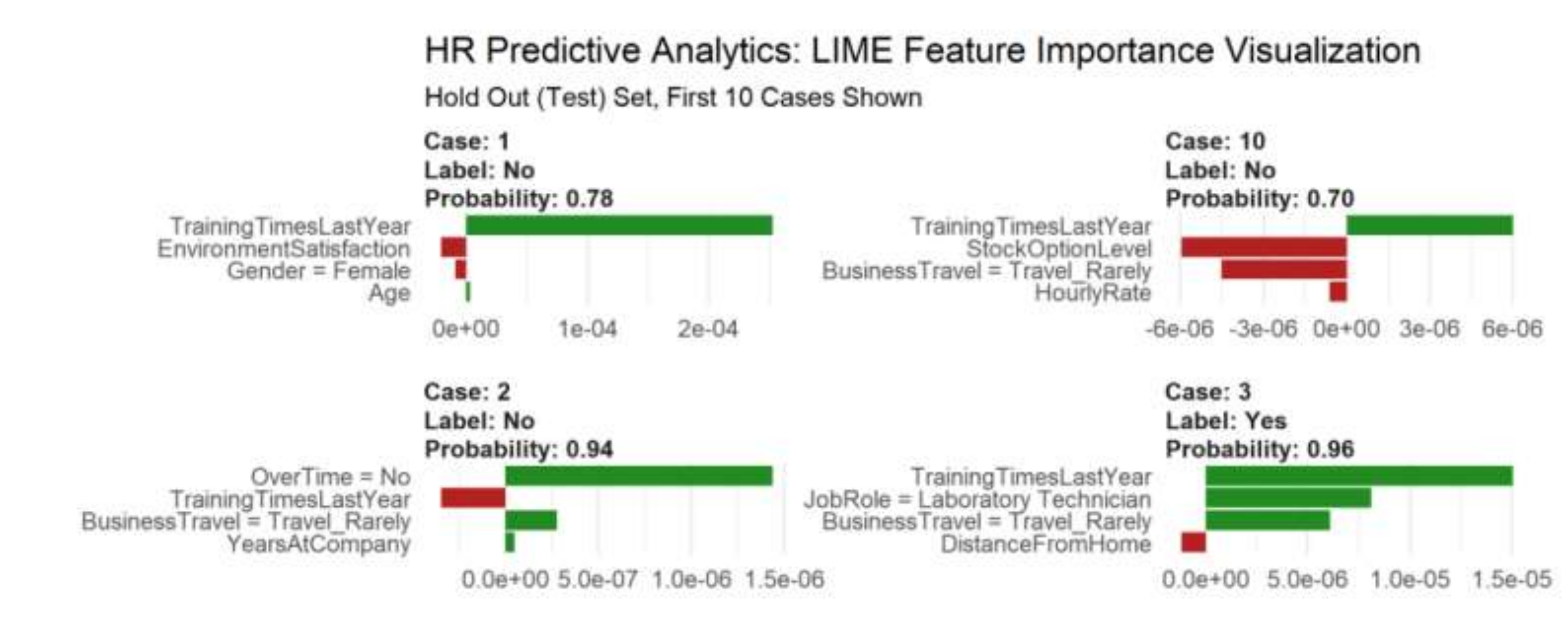
```
# Run explain() on explainer
explanation <- lime::explain(
    as.data.frame(test_h2o[1:10,-1]),
    explainer    = explainer,
    n_labels     = 1,
    n_features   = 4,
    kernel_width = 0.5)
```

Explain new observations

LIME

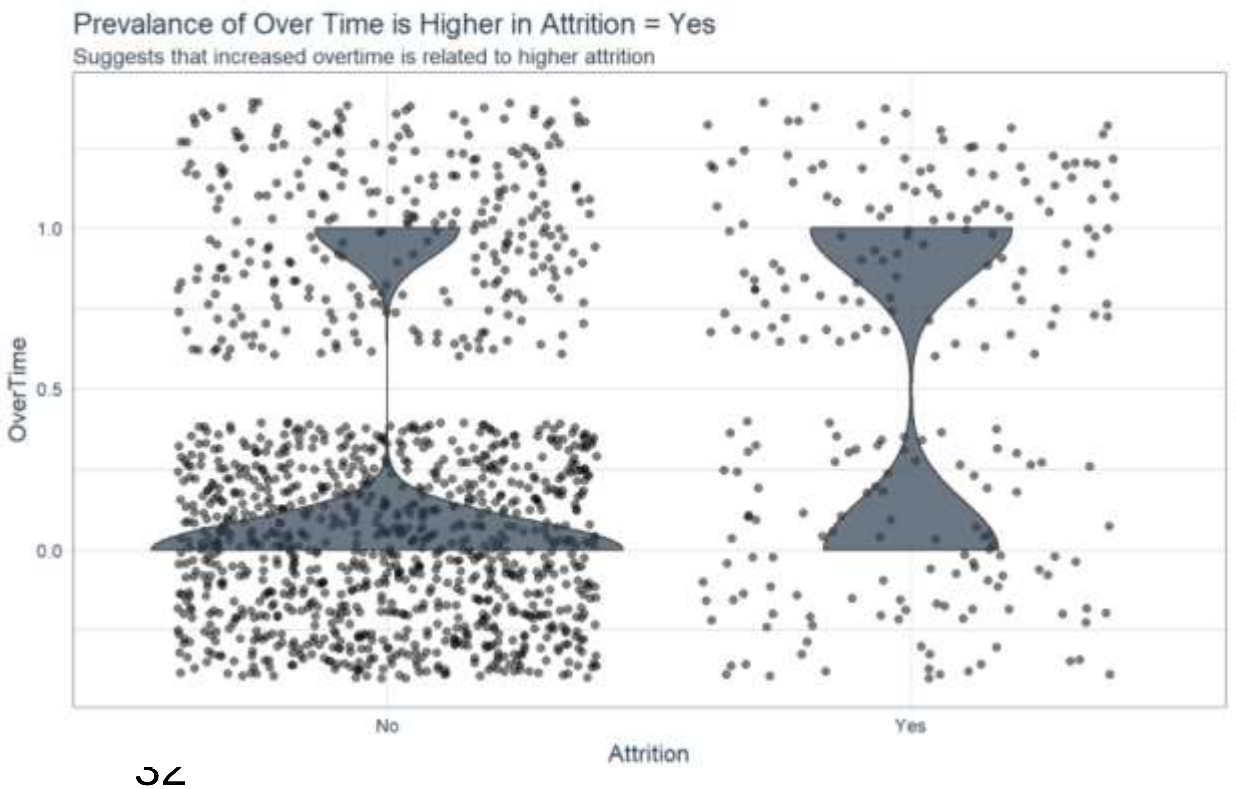
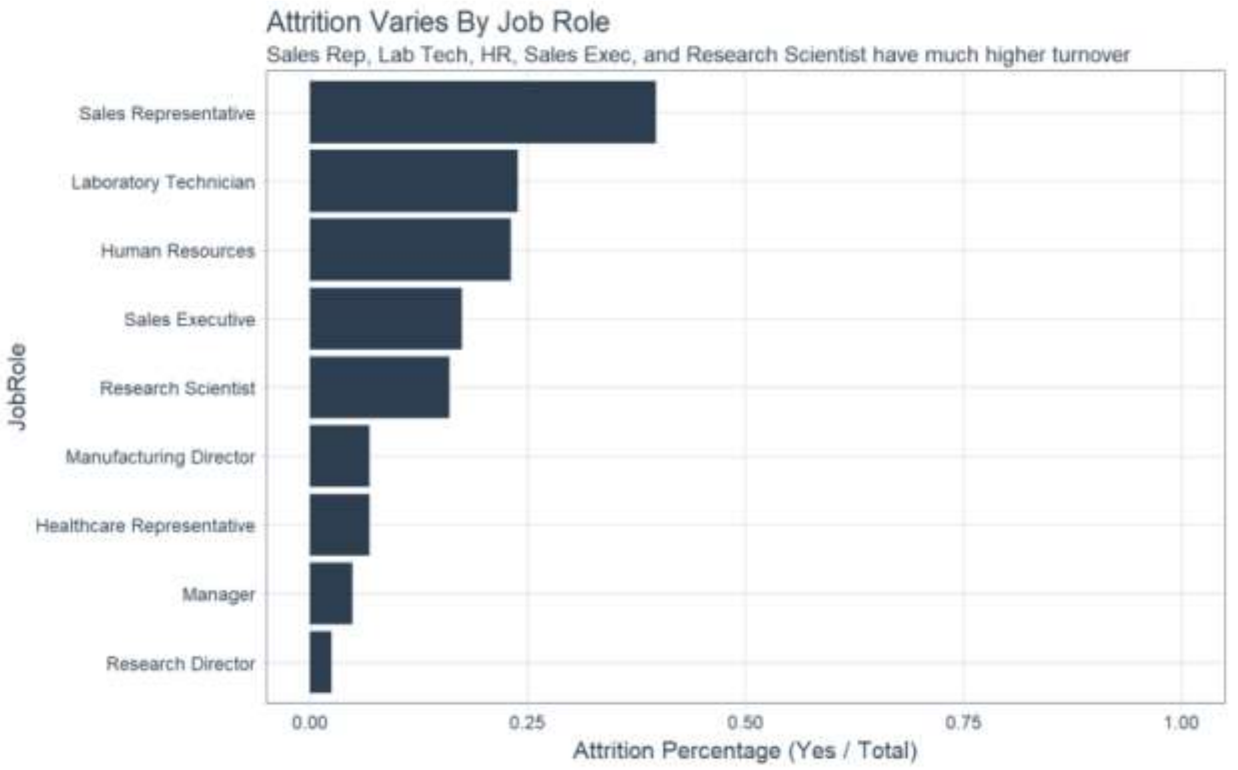
- Step 3: Plot Feature Importance

```r
plot_features(explanation) +
    labs(title = "HR Predictive Analytics: LIME Feature Importance Visualization",
         subtitle = "Hold Out (Test) Set, First 10 Cases Shown")
```



LIME

- Step 4: Inspect Important Features

Overtime	Job Role
	

Real World: Solves Real Problems

Client Case Study

- Fortune 500 firm
- Modeled **executive potential** using more sophisticated process
- Our **algorithm identified 16 employees** that predicted as executive potential but were not targeted by client

Real World: Solves Real Problems

Client Case Study

- Fortune 500 firm
- Modeled **executive potential** using more sophisticated process
- Our **algorithm identified 16 employees** that predicted as executive potential but were not targeted by client

Conclusions

Machine Learning Predicted Turnover

- 88% Accuracy
- 62% Recall ← Important!

Machine Learning Explained Turnover

- Turnover greater based on Job Role & Overtime

Framework For High Accuracy & Explainability

Conclusions

Machine Learning Predicted Turnover

- 88% Accuracy
- 62% Recall ← Important!

Machine Learning Explained Turnover

- Turnover greater based on Job Role & Overtime

Framework For High Accuracy & Explainability

Conclusions

Machine Learning Predicted Turnover

- 88% Accuracy
- 62% Recall ← Important!

Machine Learning Explained Turnover

- Turnover greater based on Job Role & Overtime

Framework For High Accuracy & Explainability

Conclusions

Machine Learning Predicted Turnover

- 88% Accuracy
- 62% Recall ← Important!

Machine Learning Explained Turnover

- Turnover greater based on Job Role & Overtime

Framework For High Accuracy & Explainability

We're Done, Right? No!!

Data Science Risk

How do we know model is right?

- Model not back-tested

Time: Cross-sectional analysis

- Model not adaptive

What do we do when model breaks down?

- Only certainty: **CHANGE**

We’re Done, Right? No!!

Data Science Risk

How do we know model is right?

- Model not back-tested

Time: Cross-sectional analysis

- Model not adaptive

What do we do when model breaks down?

- Only certainty: **CHANGE**

We’re Done, Right? No!!

Data Science Risk

How do we know model is right?

- Model not back-tested

Time: Cross-sectional analysis

- Model not adaptive

What do we do when model breaks down?

- Only certainty: **CHANGE**

We're Done, Right? No!!

Data Science Risk

How do we know model is right?

- Model not back-tested

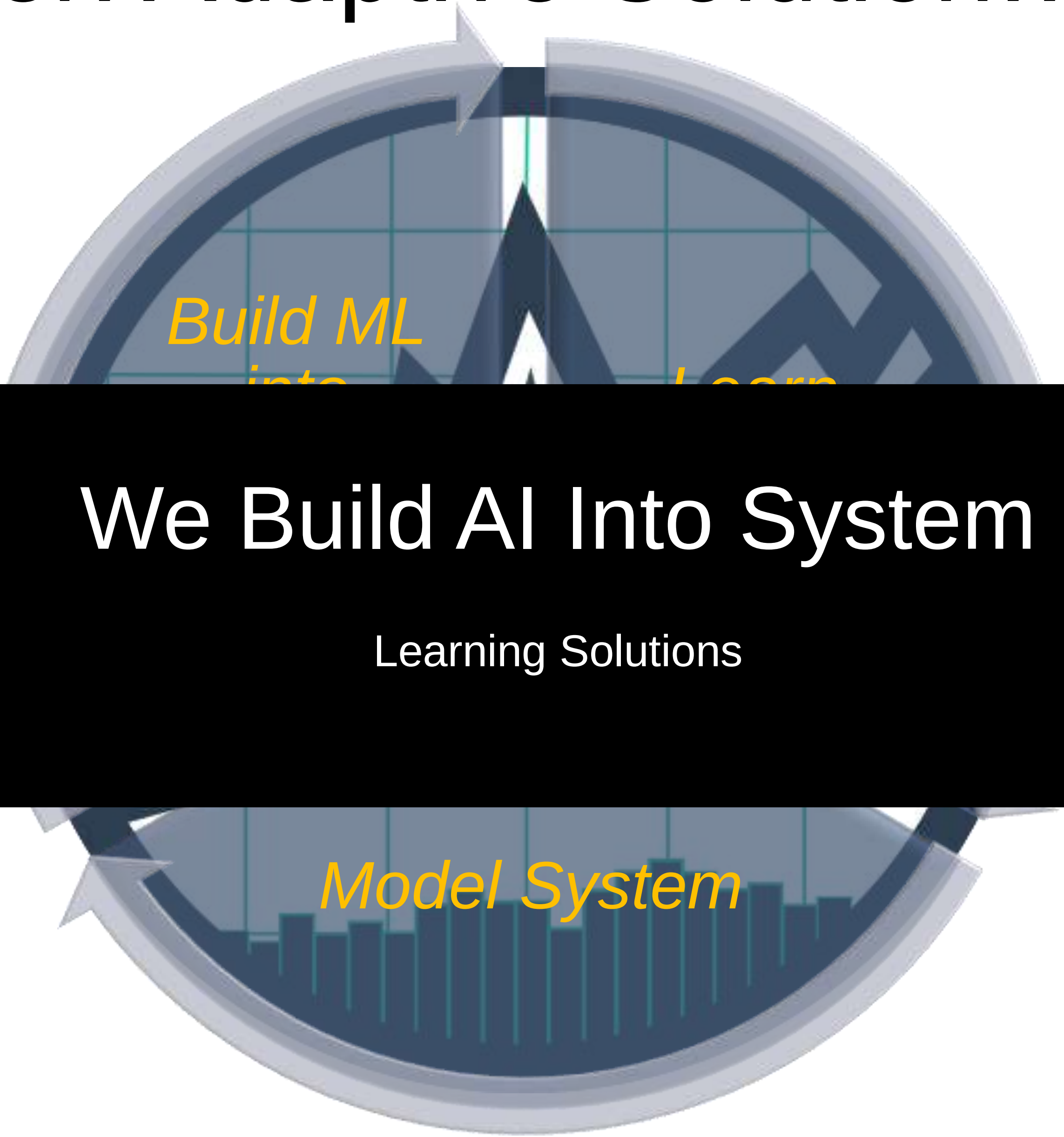
Time: Cross-sectional analysis

- Model not adaptive

What do we do when model breaks down?

- Only certainty: **CHANGE**

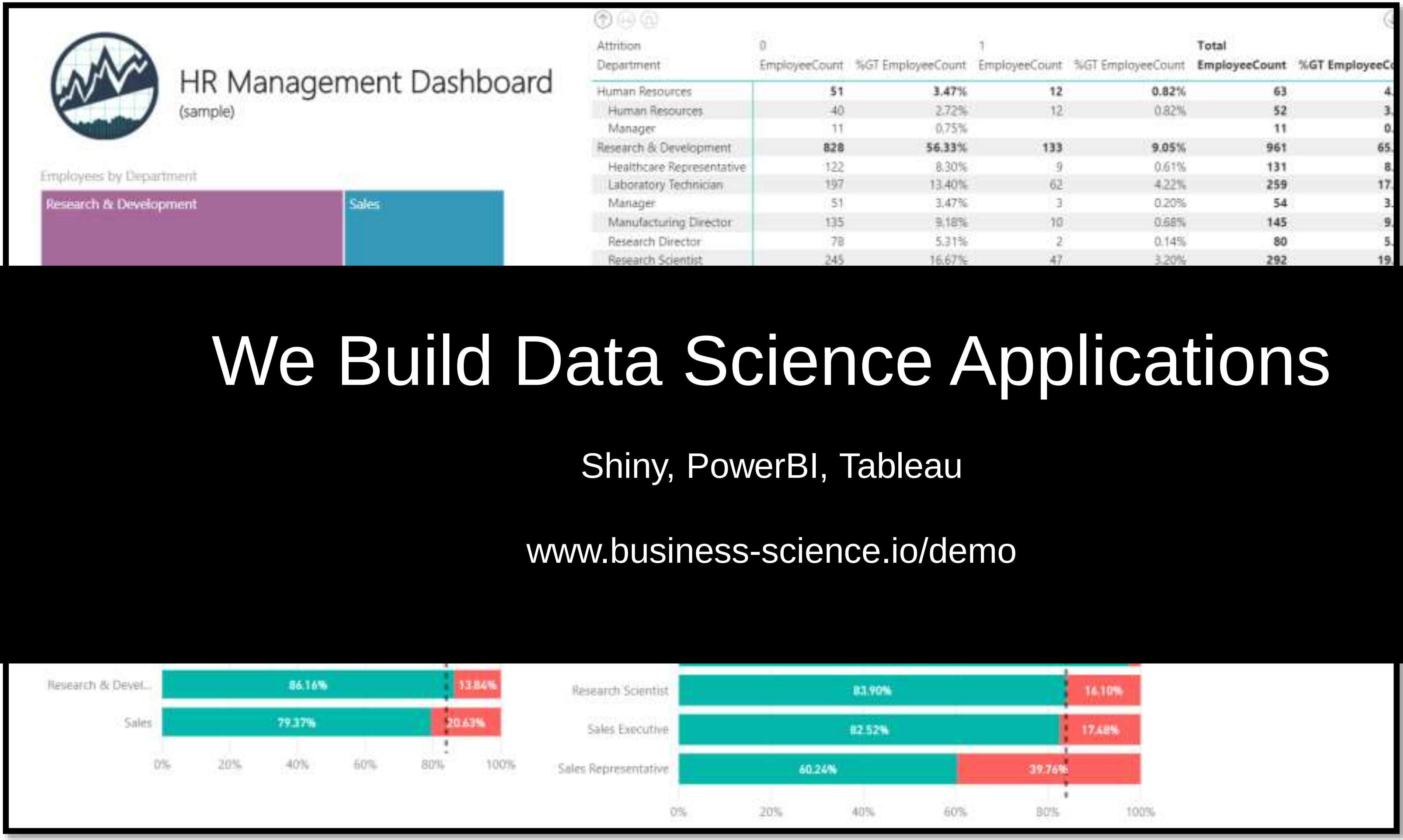
Remember: Adaptive Solution…



We Build AI Into System

Learning Solutions

What About Communication?



We Build Data Science Applications

Shiny, PowerBI, Tableau

www.business-science.io/demo

Business Science

Applying Data Science to Business and Financial Analysis

Twitter: @bizScienc // @mdancho84

Github: business-science // mdancho84

Email: mdancho@business-science.io

Website: www.business-science.io

H2O WORLD 2017