

## Predictive Data Quality

Do you catch bad data? Or does bad data catch you by surprise?

Spark's Scale + Owl's Data Science = Scalable Insights

# Daily Reactive Scenario

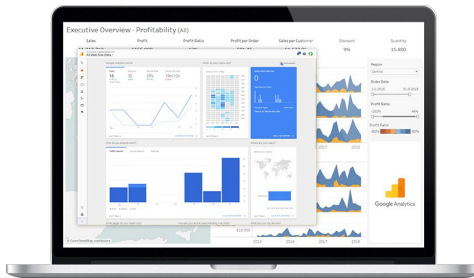


The Model is not right!

1



Analyst



It's hard to say what the model did, but I don't think it has changed.

2



Data Scientist

I'm investigating the database table and checking the 300 data quality rules. We might have missed something.

3



Data Engineer

# The Problem



10,000  
rules

The average number of **rules** needed to be manually written and maintained with a traditional approach.

---

2,000  
hrs

The number of **man hours** needed with a domain expert. Drafting rules, boundaries, and conditions that model the business.

---

30%  
coverage

The amount of **coverage** and **trust** actually achieved with a conventional approach.

# Predictive with Transparency



Data Engineer

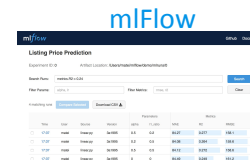
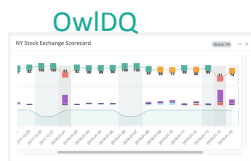


Data Scientist

The Model isn't wrong, the data suddenly changed



Analyst



Data Flow

Dataset  
Table/File/Kafka



Owl Scan  
REST/CMDLine/Spark

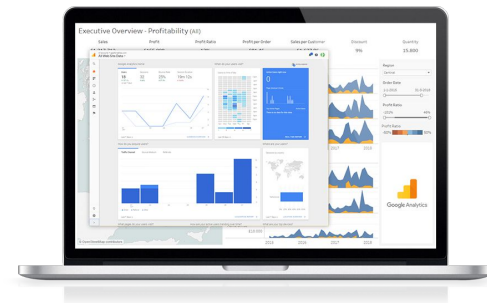


Reports  
Models  
Dashboards

Data Loading

OwlIDQ Suite

Trusted Analytics

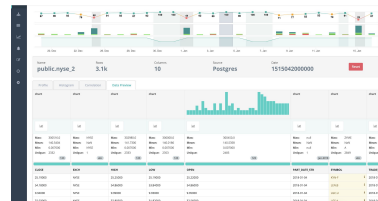


# A Complete DQ Suite



symbol ▾				
▼ symbol	String	xxx.x		
volume	symbol	high	low	exch
1,400	AGM.A	74	71.99	NYSE

Name	Date	Phone
John Doe	Jan-06	444-333-2222
Mark Smith	2015-04-13	111335555
jon doe	2016-01-06	444-333-2222



Profile & **Generate**  
thousand of rules

**Schema** Evolution &  
Data **Shape** issues

**Reconciliation**  
Source to Target

**Duplicates, Outliers,**  
**Pattern**

Discover data **Correlations,**  
**Segmentations** and **Behaviors**

Max: zwilshaw6c@	Max: n/a	Max: 99-9985119	Max: null
Mean: NaN	Mean: NaN	Mean: NaN	Mean: NaN
Min: aabramow9n	Min: Accident &lt;h	Min: 00-0340993	Min: null
Unique: 1000	Unique: 117	Unique: 1000	Unique: 1
merge EMAIL		merge EIV	
Abc email	Abc emplmt_industry	Abc empl_num	Abc emp_nut_col
skertessq2@bigcartel.com	Hotels/Resorts	56-2885659	
ddiversny@ezinearticles.com	n/a	86-7565104	
amedwellg@comsenz.com	Aerospace	59-7295249	

Automatic Rules

Custom Rule  
Builder

Builder

owl\_test\_loan\_customer | A.app\_id | B.app\_id | lake\_loan\_customer

NOT AND OR

intrst\_rate greater

+ Add rule + Add group

Delete

first_name	first_name
last_name	last_name
email	email
gender	gender
credit_card	credit_card
acct_number	
ssn_number	

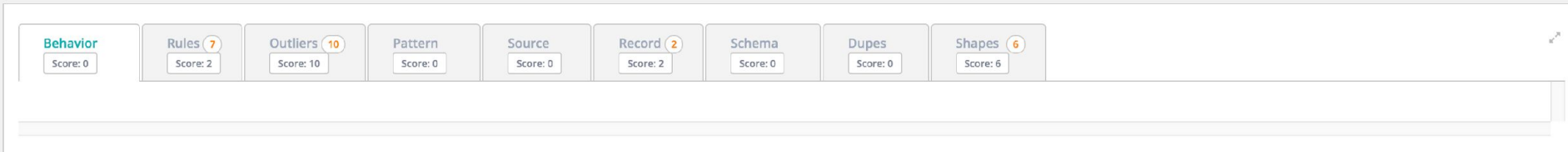
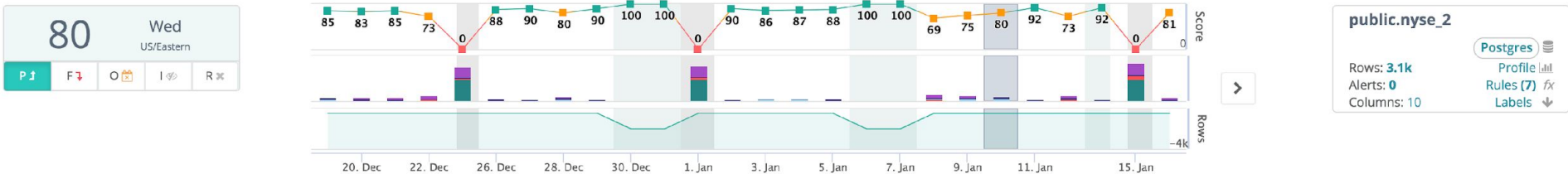
Name (pk)	age	date
Mark	34	1986-03-02
Kevin	22	1995-01-05
Janice	50	1970-12-20

Name (pk)	age	date
Mark	34	1986-03-02
Kevin	2	1995-01-05



- There is a strong relationship between age and empl\_pn
- There is a strong relationship between age and salary
- There is a strong relationship between empl\_pn and salary

# Owl automatically learns the patterns in your data, drill in for unsupervised discovery



Search:  Copy CSV Excel PDF Print Whitespace OFF

Max: 308350.0 Mean: 143.3041 Min: 0.090000 Unique: 2397	Max: NYSE Mean: NaN Min: NYSE Unique: 1	Max: 308599.0 Mean: 143.7881 Min: 0.106000 Unique: 2536	Max: 303930.0 Mean: 141.5091 Min: 0.090000 Unique: 2302	Max: 304625.0 Mean: 142.1981 Min: 0.095000 Unique: 2525	Max: null Mean: NaN Min: null Unique: 1	Max: ZYME Mean: NaN Min: A Unique: 2837	Max: null Mean: NaN Min: null Unique: 1	Max: 1075129 Mean: 1152843 Min: 0 Unique: 2533	Max: 3105 Mean: 1553.0 Min: 1 Unique: 2834
123	123	123	123	123	123	123	123	123	123

CLOSE	EXCH	HIGH	LOW	OPEN	PART_DATE_STR	SYMBOL	TRADE_DATE	VOLUME	owl_id
1.60000	NYSE	2.95000	1.60000	2.35000	2018-01-10	KOD.W	2018-01-10	1190400	1628
21.01000	NYSE	21.75000	20.50000	21.28000	2018-01-10	AFS-A	2018-01-10	59600	58
21.07000	NYSE	21.33000	20.89000	21.25000	2018-01-10	AFS-F	2018-01-10	152500	63
22.18000	NYSE	22.56000	22.08000	22.30000	2018-01-10	AFS-B	2018-01-10	13500	59

# OwlDQ Helps Companies to...



AI

Reach your AI goals within weeks by applying OwlDQ's proven machine learning approach that is purpose-built to find erroneous data.

---

Save

Achieve costs savings and lessen DQ efforts up to 70% after integrating OwlDQ's auto-discovery and adaptive rules.

---

Gain

Gain end-user trust, unified coverage, team collaboration through feedback loops, and derive deeper insights.