

How To Catch A Thief

...

*Identifying Tax Evasion
with
Machine Learning*

About Me

Scripps Institution of Oceanography,
MS Geophysics

Science/AAAS,
Science Journalism Fellow

UC Berkeley,
BA Geophysics

+

Lauren DiPerna, Data Scientist @ H₂O.ai

What is Money Laundering?

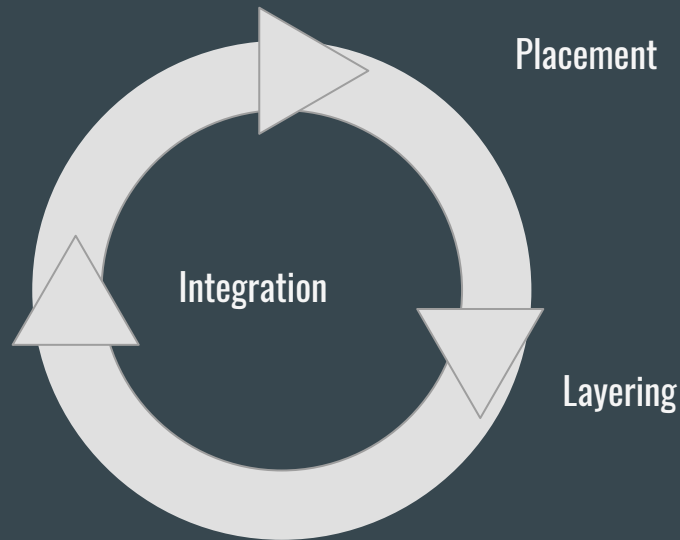
...

Making Dirty Money Look Clean

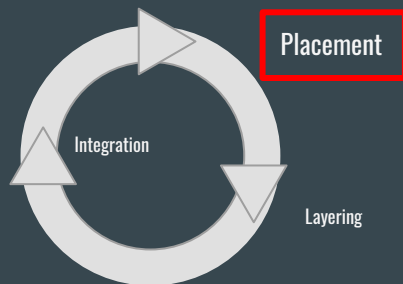
How it Works

CHECKLIST

- ✓ *Convert Dirty Money*
- ✓ *Deposit into Bank*
- ✓ *Withdraw from Bank*



Get Your Money Into a Financial Institution



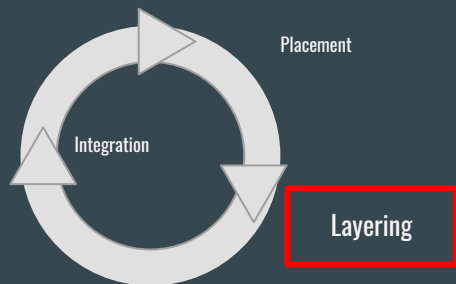
Pros

- ❏ Stop Stuffing Money into Mattress
- ❏ Stop Paying for Extra Guards

Cons

- ❏ Possibility of Getting Caught
- ❏ Hard to Find Trustworthy Smurfs

Use Multiple Accounts to Avoid Suspicion



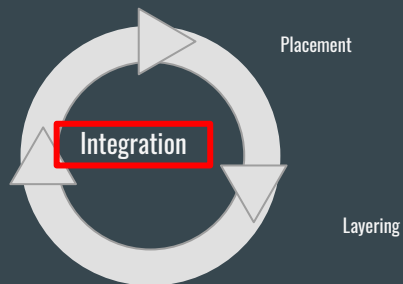
Pros

- ❏ Won't Lead Back to Me
- ❏ Hard for IRS to audit

Cons

- ❏ Headaches From Keeping Track of Everything
- ❏ Have to Figure out Loopholes in Different Languages

Get Your Money Back



Pros

- ❏ Can Access Money & Appear Innocent
- ❏ Excuse to Buy a House
- ❏ Excuse to Buy a Boat
- ❏ Excuse to Buy Expensive Art
- ❏ Money

Notorious Money Laundering



Shell Banks

No Record

*“Russian criminals laundered about \$70 billion through this shack in Nauru”
- NYTimes*

Why is Money Laundering Hard To Catch?



Untangle Transactions
& Identifying the
Source is Hard



Rule-Based Systems
Rely on Relatively
Stateless Checks



Criminals Find Ways
Around Regulatory
Detection

Could We Use ML to Help?

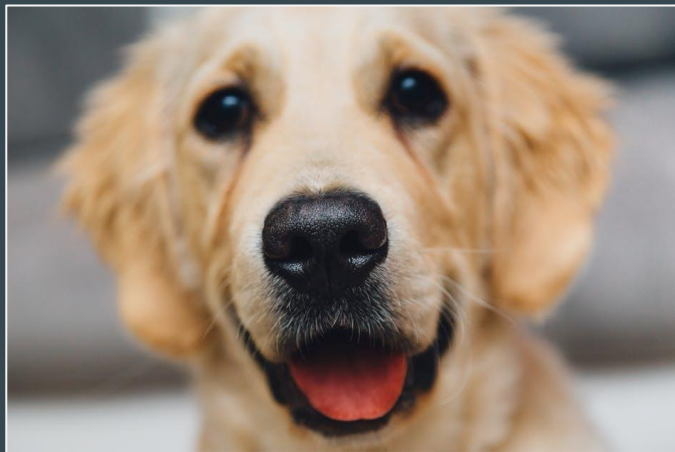


Machine	Human
2	0



Why We Should Use H2O Machine Learning

How Would you Teach a Child to Identify an Animal?

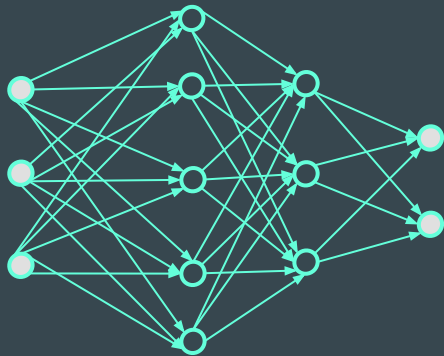


Why H2O Machine Learning Can Beat Rules

```
If deposit > 10,000 and  
Acct_age < 1 year then  
....
```

~~Hard Code Rules~~
~~Specify Thresholds~~
~~Deterministic~~

Learns Rules from Data
Optimizes Thresholds to Minimize Error
Probabilistic Scores



About H₂O.ai

Distributed in-memory platform

Easy to use SDK / APIs

Open Source

Can use ALL business data

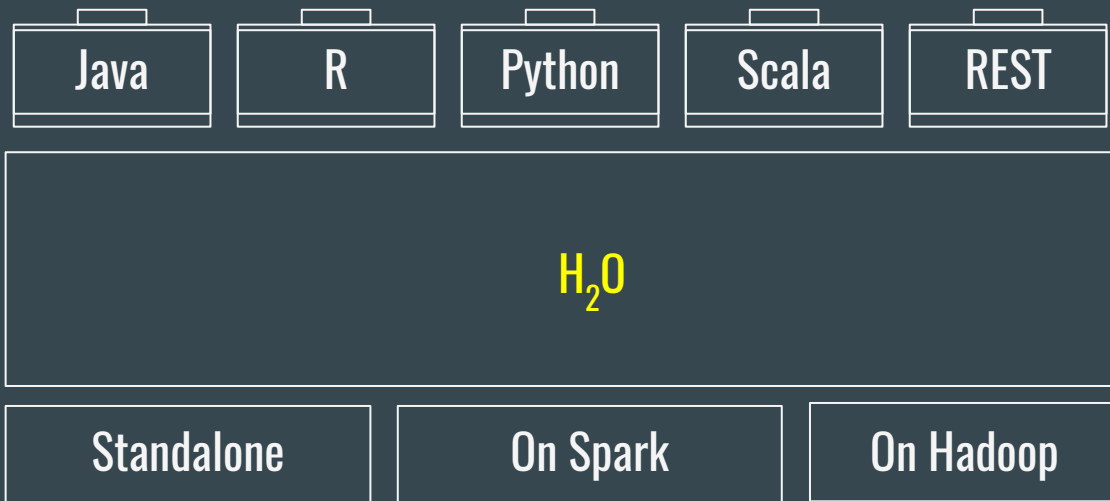
GLM, DRF, GBM, Deep Learning, K-means, PCA

Java, Python, R, Scala, JSON, Browser-Based GUI

Ownership of Methods

Modeling without Sampling

H₂O.ai Integration



H₂O Algorithm Overview

Statistical Analysis

Generalized Linear Models
Naïve Bayes

Clustering

K-means

Ensembles

Distributed Random Forest
Gradient Boosting Machine

Dimensionality Reduction

Principal Component Analysis
Generalized Low Rank Models

Deep Neural Networks

Deep learning

Anomaly Detection

Autoencoders

Distributed Random Forest

Original Dataset

	Current Balance	# Deposits	Age of Account
acct_1	50000	20000	100
acct_2	10	1000	2000
acct_3	1	5	10



Predictions

Acc. 1	Suspect
Acc. 2	Innocent
Acc. 3	Suspect
Acc. 4	Innocent
	.
	.
	.
	.
	.
	.

Tax Evasion Use Case

...

How Regulators Deal with Tax Evasion

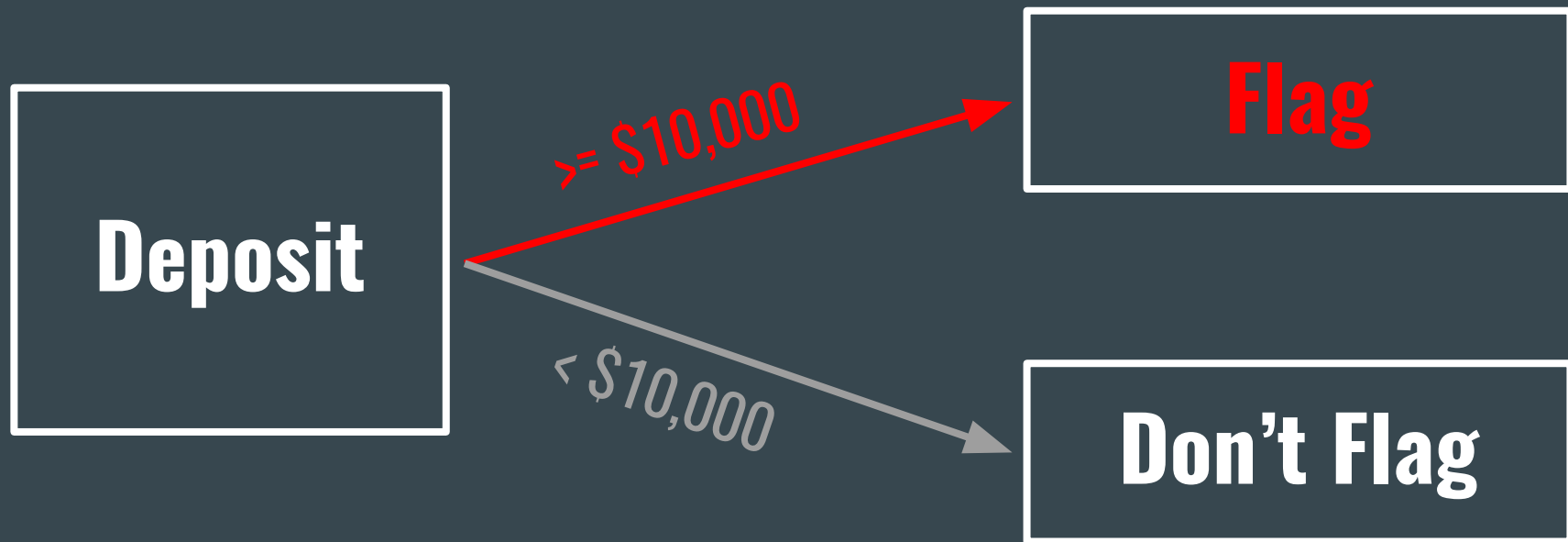
Bank Secrecy Act

THE BANKS



THE IRS/FinCEN

What Gets Sent to Regulators



Structuring: How Criminals Avoid Detection

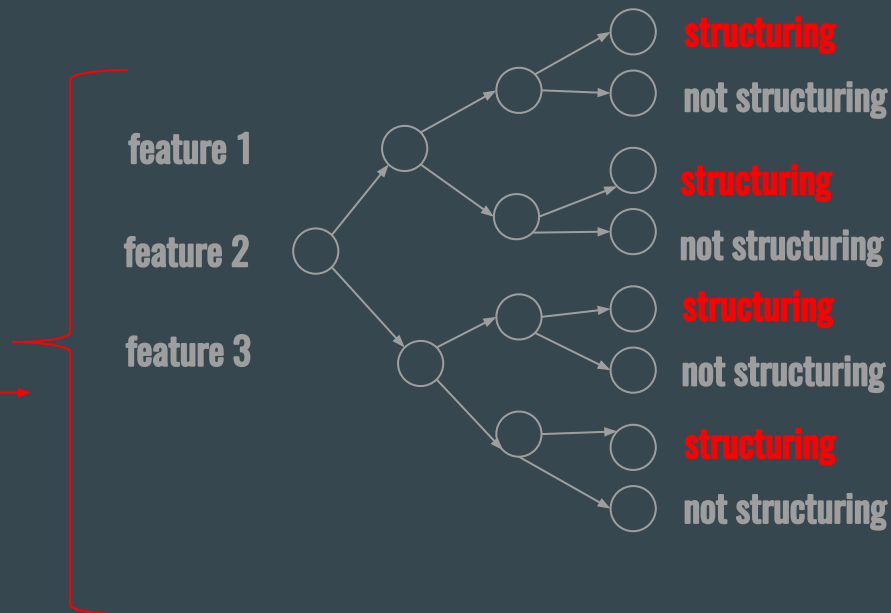
Spaced Out Deposits



\$9,999, ..., \$9,999



How H2O Machine Learning Can Help



But How is ML Different from Rule Systems?

	Threshold	Trends/Spikes	Memory	Train/Retrain
Machine Learning				
Rules				

Made for Machine Learning

...

What Features Should We Input to an Algorithm?

Features that Provide Context



Profile Summary

Current balance: \$50,000

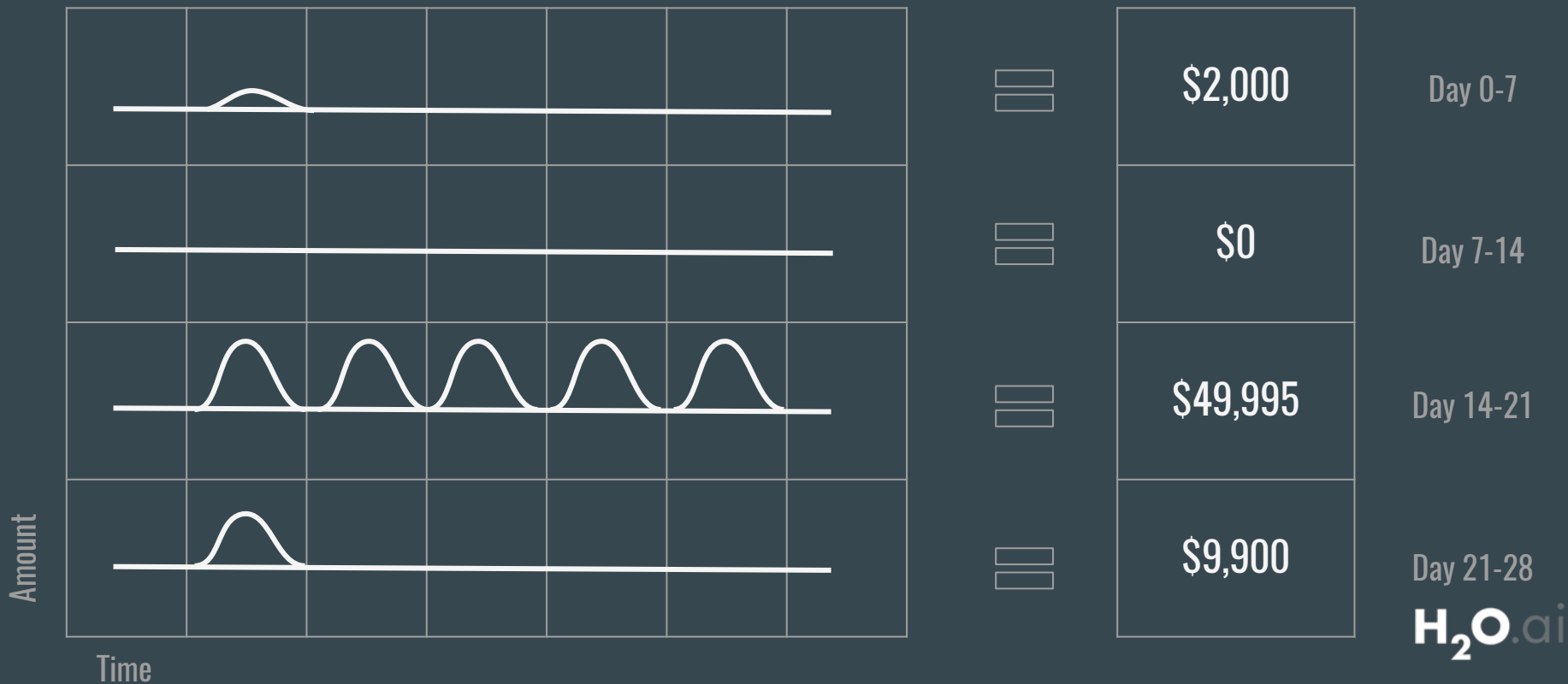
Deposits: 1

Withdrawals: 3

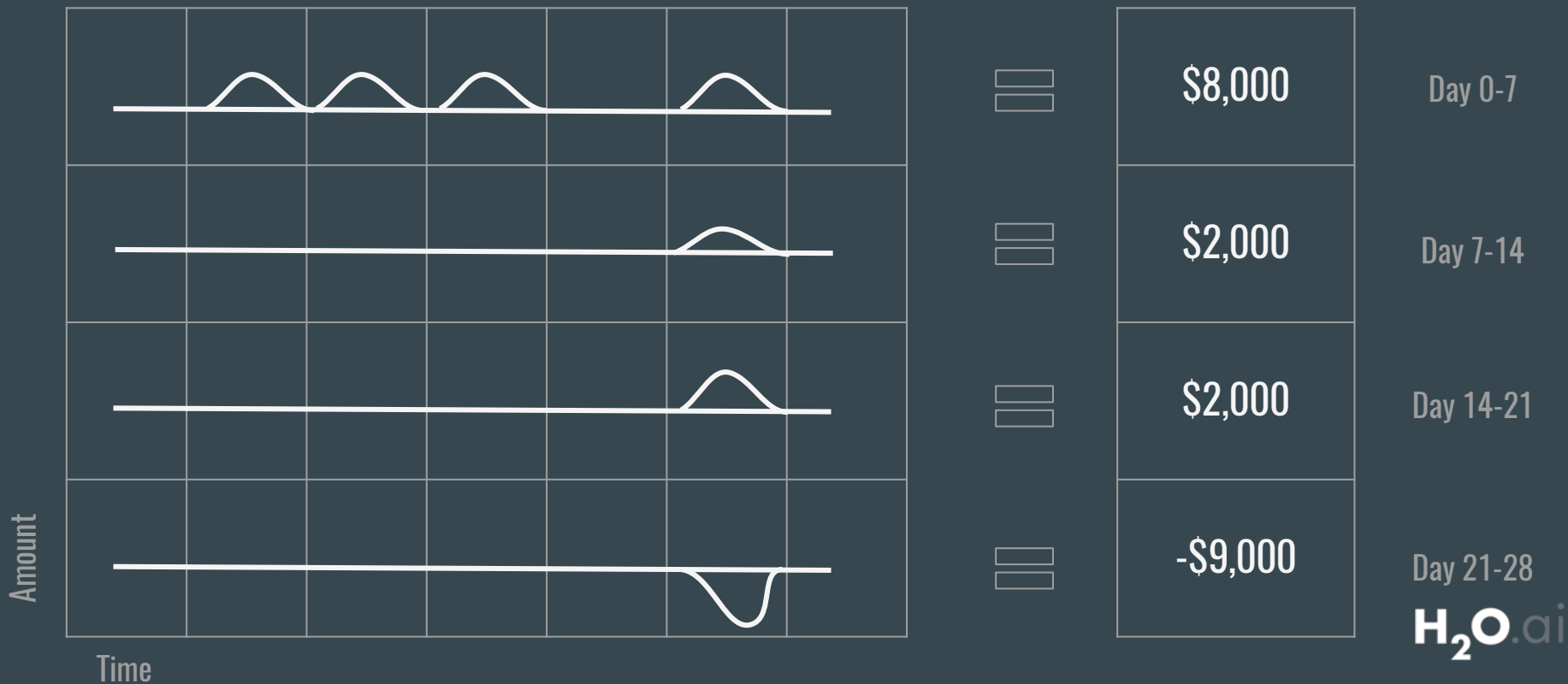
Age of account: 10 years

Linked accounts: 3

Feature 1. Total Amount Deposited Last 7 Day



Feature 2. Total Amount Withdrawn Last 7 Day






Feature 3. % of Monetary Instruments Last 30 Days

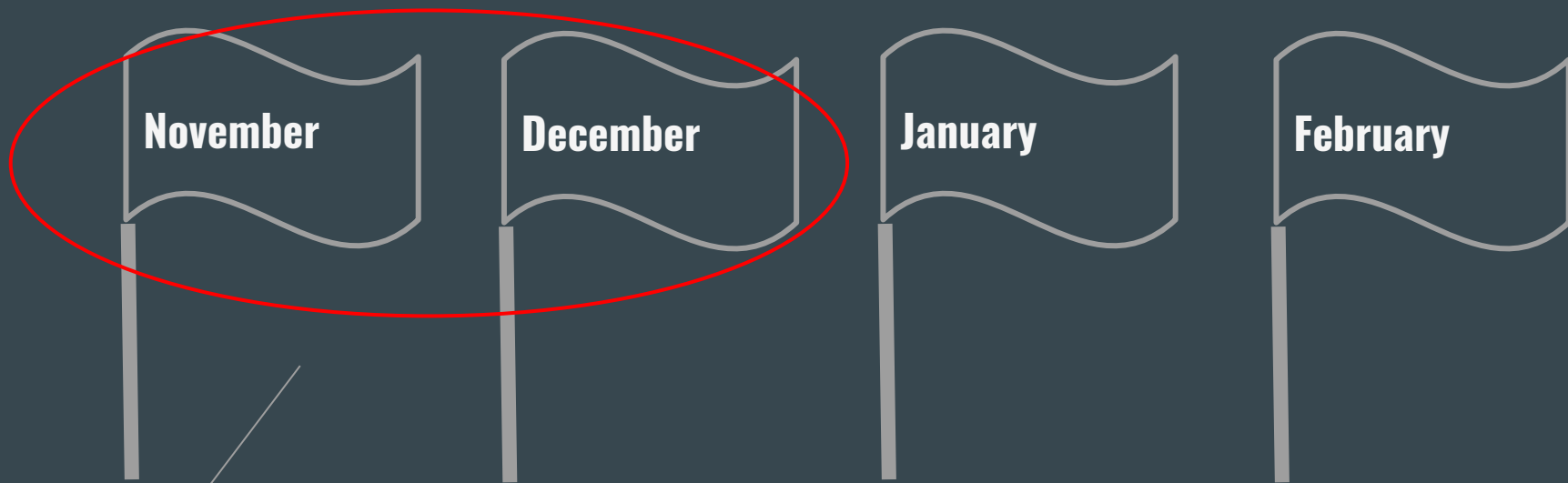
100% Cash

50% Cash

20% Cash

-  Suspicious
-  Moderately Suspicious
-  Normal

Feature 4. Tax Day Behavior

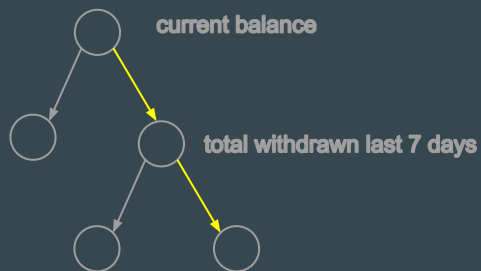


Add weight to transactions closer to tax season

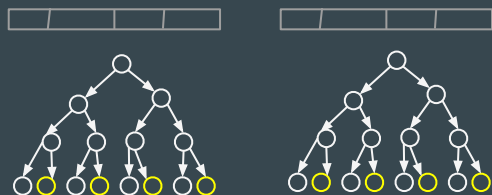
How the Algo Uses Features to Identify Structuring



Split along features that minimize prediction error



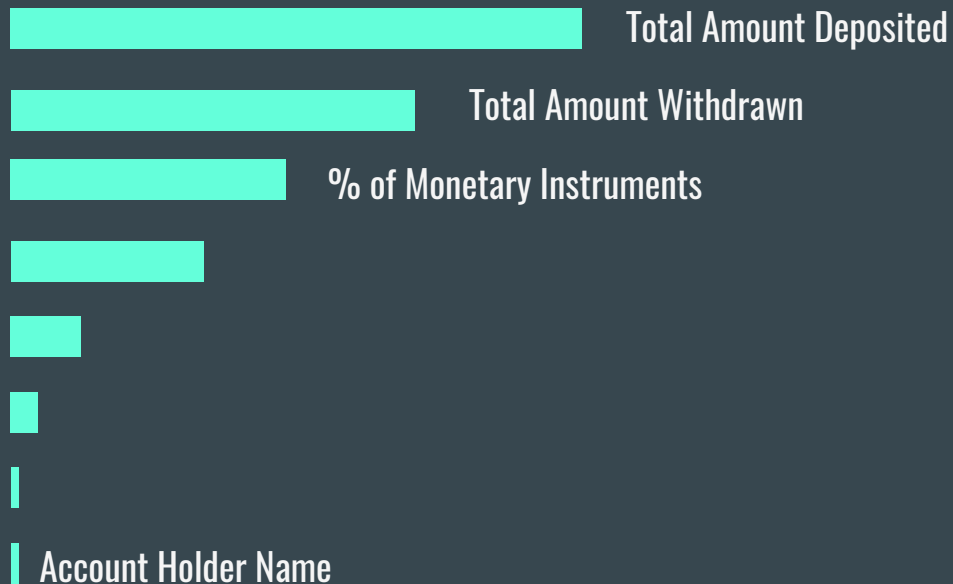
Ability to include feature interactions



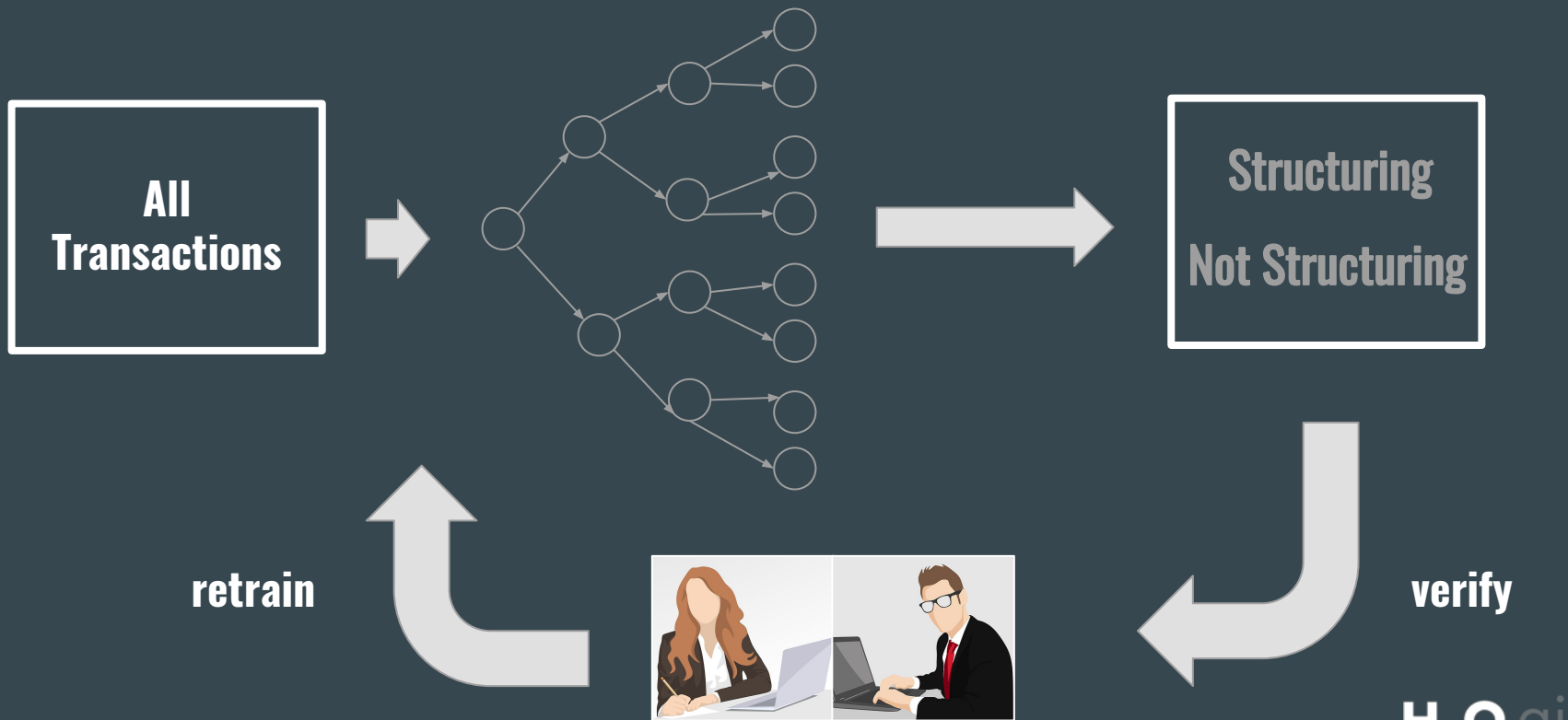
Composition of features yields a probability score,
used to identify structuring

Bonus Machine Learning Treat

Feature Importance



Putting It All Together



Machine Learning for Anti-Money Laundering

~~LOOPHOLES~~

Enhance Rules

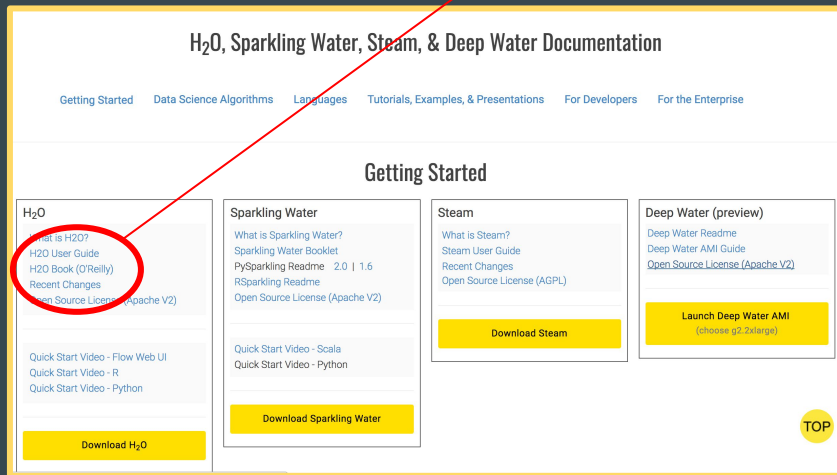
Help Generate Better Rules

Identify New Criminal Patterns

Replace Rule-Based Systems

H₂O Machine Learning Resources

H2O User Guide



H₂O, Sparkling Water, Steam, & Deep Water Documentation

Getting Started | Data Science Algorithms | Languages | Tutorials, Examples, & Presentations | For Developers | For the Enterprise

Getting Started

H₂O

- What is H₂O?
- H₂O User Guide**
- H₂O Book (O'Reilly)
- Recent Changes
- Open Source License (Apache V2)

Quick Start Video - Flow Web UI
Quick Start Video - R
Quick Start Video - Python

Download H₂O

Sparkling Water

- What is Sparkling Water?
- Sparkling Water Booklet
- PySparkling Readme 2.0 | 1.6
- RSparkling Readme
- Open Source License (Apache V2)

Quick Start Video - Scala
Quick Start Video - Python

Download Sparkling Water

Steam

- What is Steam?
- Steam User Guide
- Recent Changes
- Open Source License (AGPL)

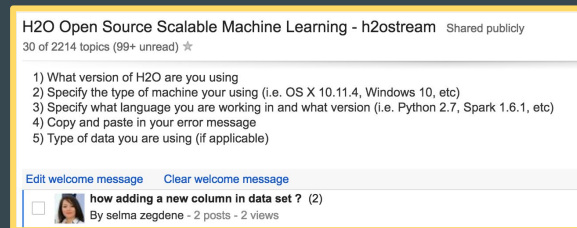
Download Steam

Deep Water (preview)

- Deep Water Readme
- Deep Water AMI Guide
- Open Source License (Apache V2)

Launch Deep Water AMI
(choose g2.xlarge)


TOP

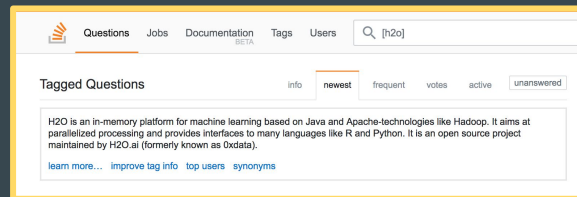


H₂O Open Source Scalable Machine Learning - h2ostream Shared publicly
30 of 2214 topics (99+ unread) ☆

- 1) What version of H₂O are you using
- 2) Specify the type of machine your using (i.e. OS X 10.11.4, Windows 10, etc)
- 3) Specify what language you are working in and what version (i.e. Python 2.7, Spark 1.6.1, etc)
- 4) Copy and paste in your error message
- 5) Type of data you are using (if applicable)

[Edit welcome message](#) [Clear welcome message](#)

 **how adding a new column in data set ?** (2)
By selma zegdene - 2 posts - 2 views



Questions | Jobs | Documentation | Tags | Users |

Tagged Questions

H₂O is an in-memory platform for machine learning based on Java and Apache-Technologies like Hadoop. It aims at parallelized processing and provides interfaces to many languages like R and Python. It is an open source project maintained by H₂O.ai (formerly known as Oxdelta).

[learn more...](#) [improve tag info](#) [top users](#) [synonyms](#)

<http://docs.h2o.ai/>

<https://groups.google.com/forum/#!forum/h2ostream>
<http://stackoverflow.com/questions/tagged/h2o>

laurend@h2o.ai

H₂O.ai

Don't Forget to File your Taxes (on all your income)