Large-Scale Machine Learning at PayPal Risk

Zhang Pengshan

TO DECLINE, OR NOT DECLINE?

7:15pm: Card holder lives in US - **His wife paid a bill online using their home laptop**
4:16am CEST: Card holder's son studies in Europe - **He bought Angry Birds on iPad instead of studying**
12:18pm AEST: Card holder travels to Australia - **He just paid for lunch at a POS**

AGENDA

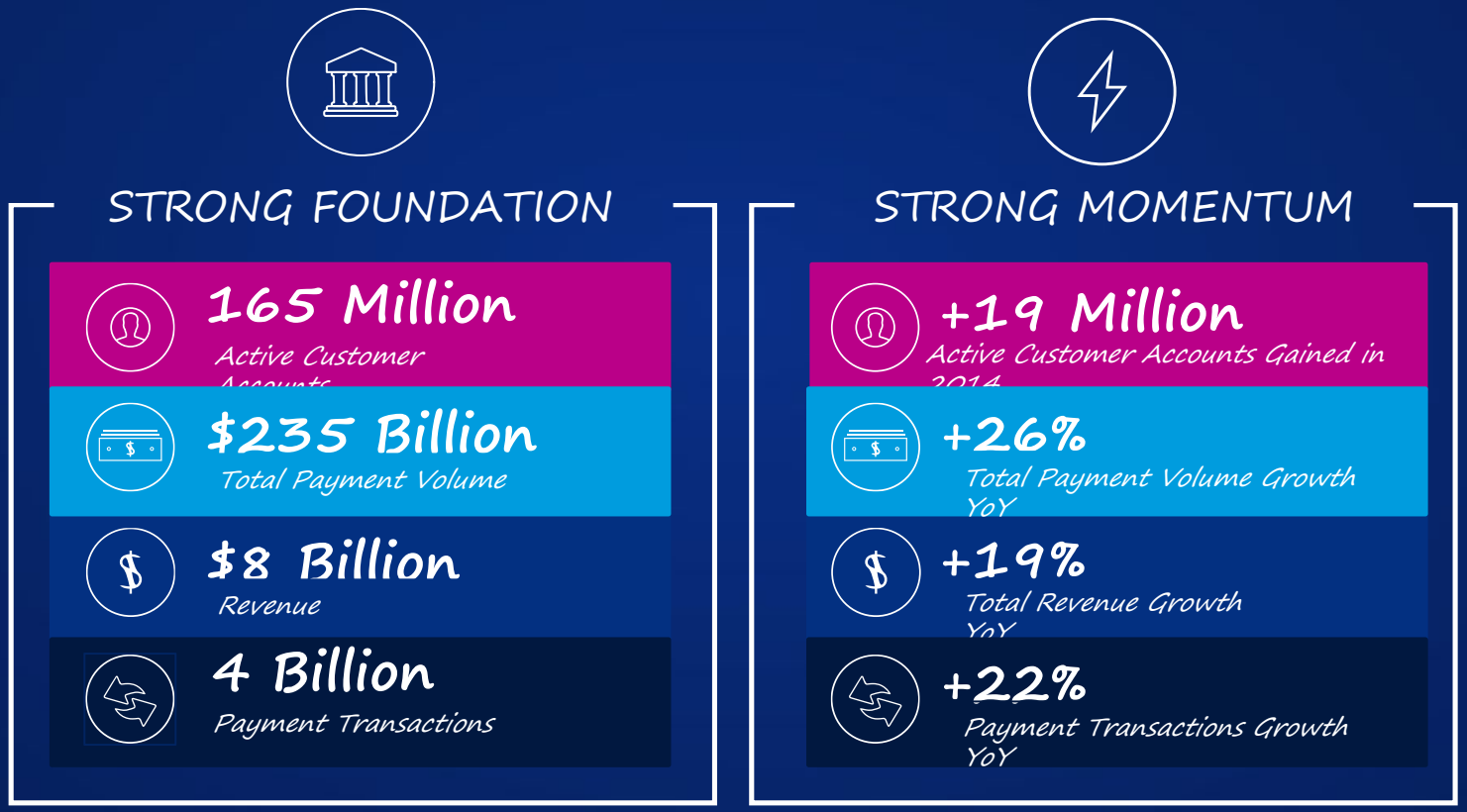
- PayPal & PayPal Risk
- Large Scale Machine Learning Solution
- Feature Engineering & Modeling
- Future Plan

AGENDA

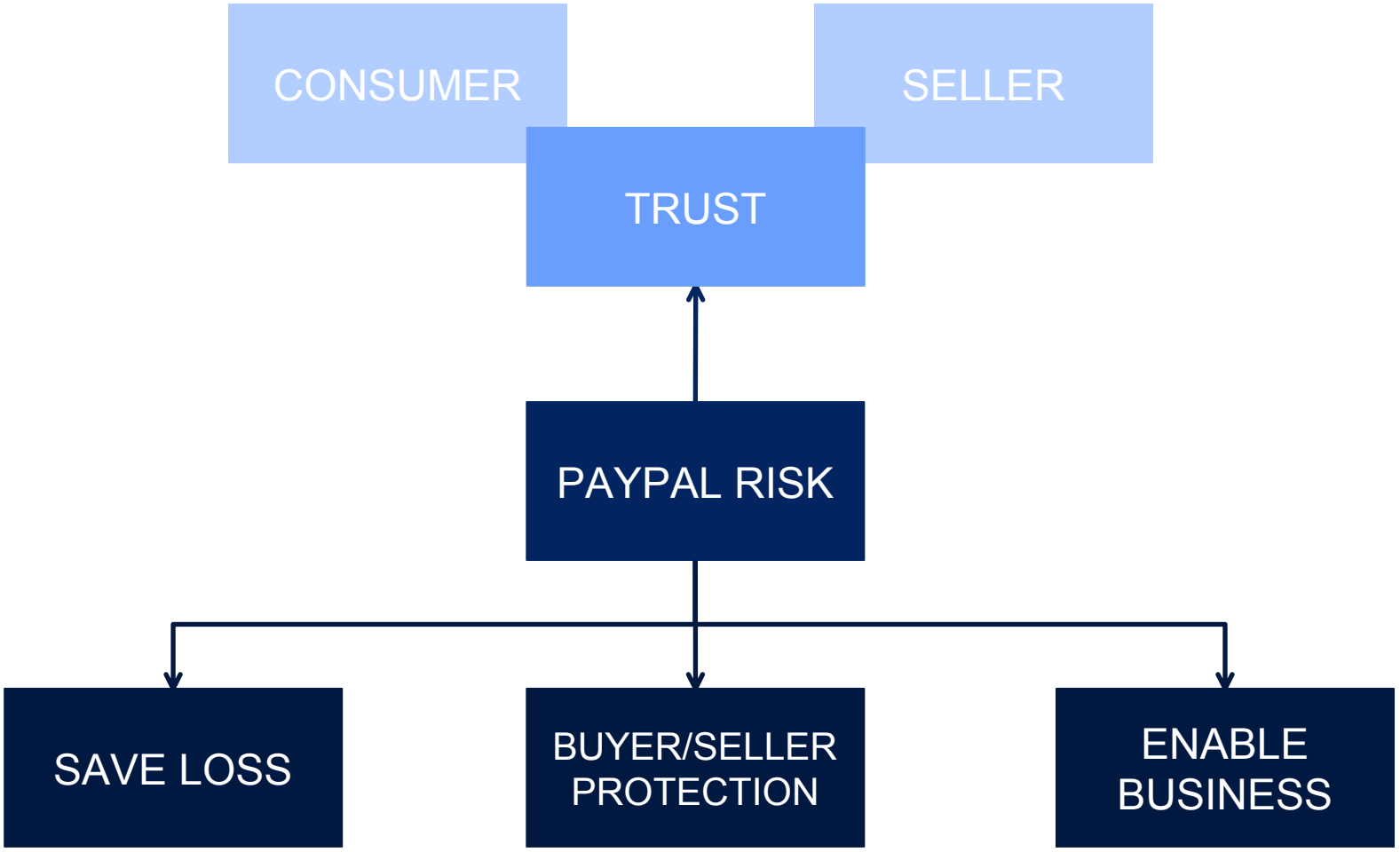
- PayPal & PayPal Risk
- Large Scale Machine Learning Solution
- Feature Engineering & Modeling
- Future Plan


Nasdaq
S&P 500 INDEX
NASDAQ
+0.1%
2,126.64
+2.35
DOWJONES
INDUSTRIAL AVERAGE
NASDAQ
-0.2%
18,086.45
-33.80
NASDAQ COMPOSITE
NASDAQ
+0.9%
5,210.14
+46.96
NASDAQ Nasdaq NDAQ
NASDAQ LISTED
PRE-MARKET
UNCHANGED
$51.84
0.00
NASDAQ PayPal PYPL
NASDAQ LISTED
PRE-MARKET
+4.6%
$40.15
+1.76
DOWJONES
INDUSTRIAL AVERAGE
NASDAQ
-0.2%
18,086.45
-33.80
PayPal
Nasdaq
PYPL
NasdaqListed

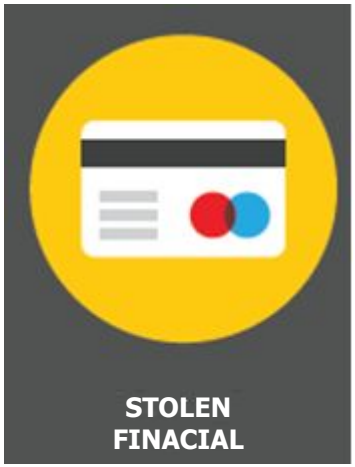
PayPal: Leading the Digital Payments Revolution

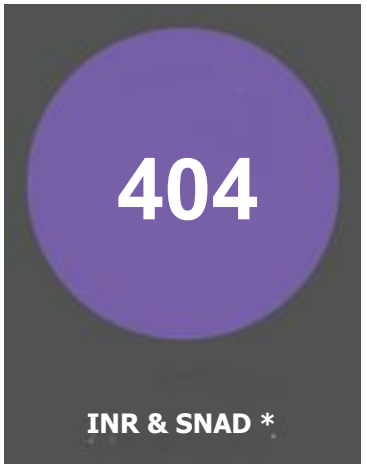


PAYPAL RISK



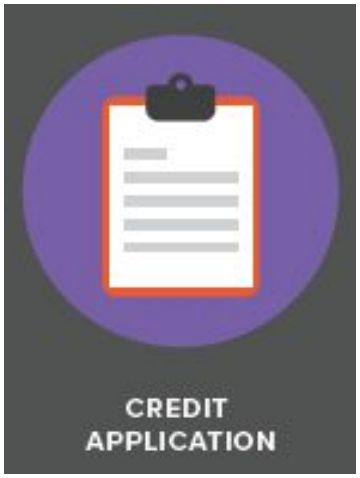
TRADITINAL FRAUD TYPES IN PAYPAL

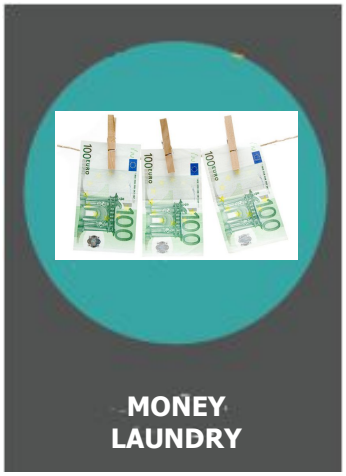


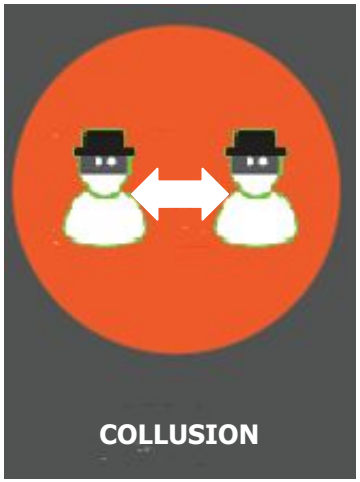


INR: Item Not Received
SNAD: Significantly Not as Described

“NEW” FRAUD TYPES IN PAYPAL





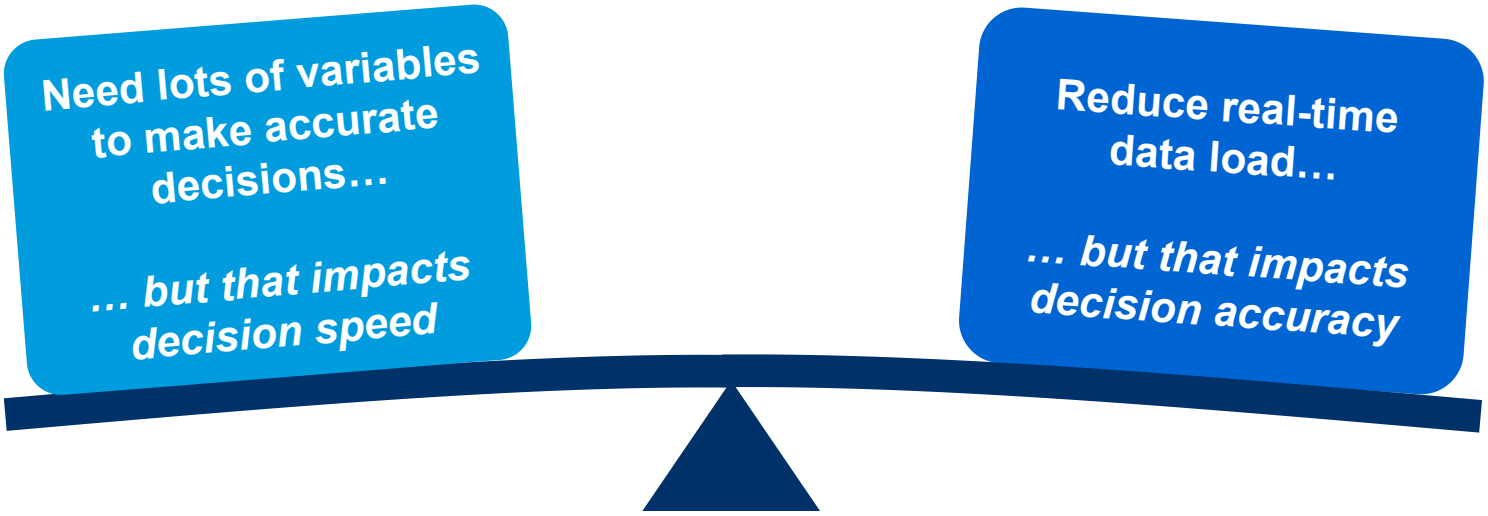


AGENDA

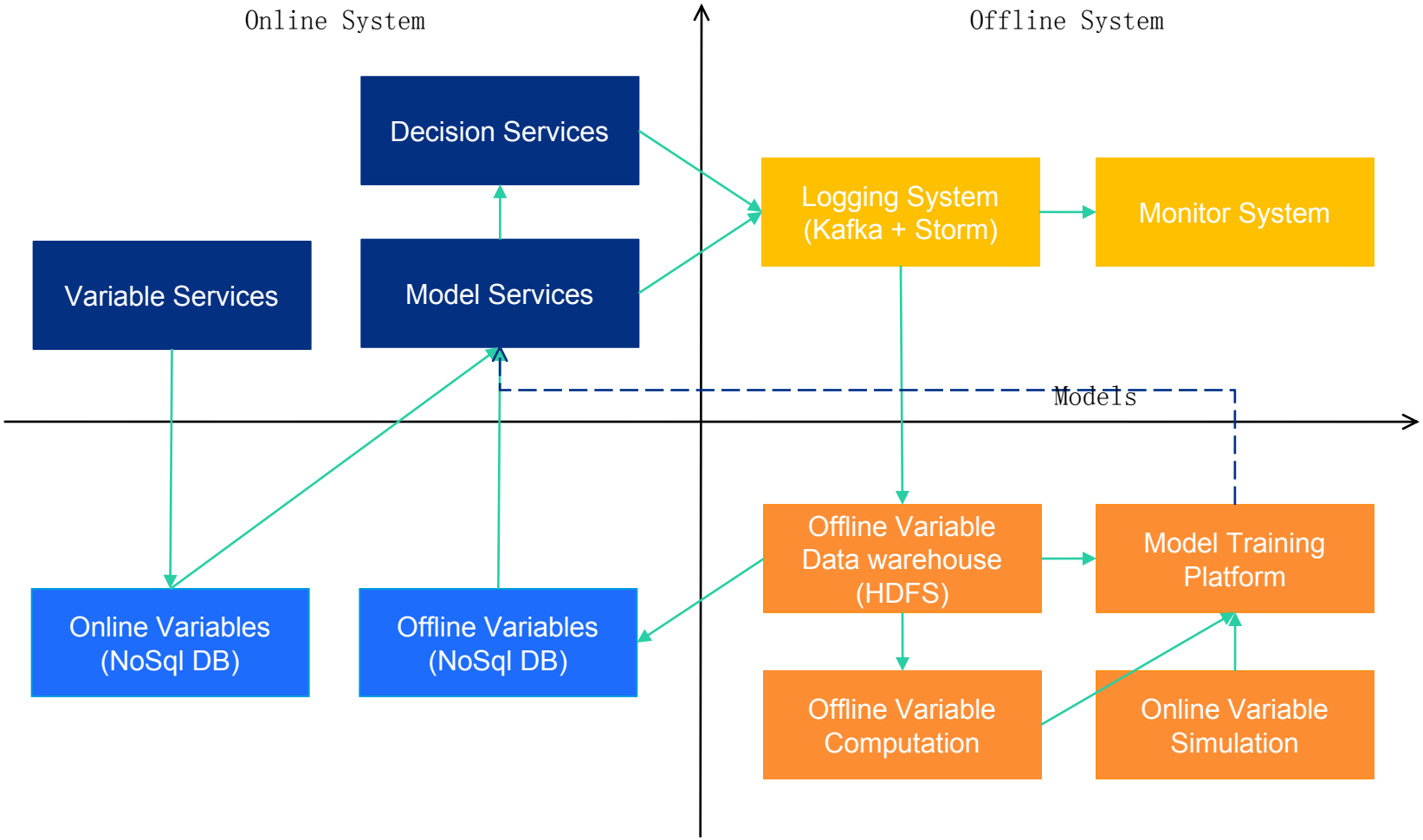
- PayPal & PayPal Risk
- Large Scale Machine Learning Solution
- Feature Engineering & Modeling
- Future Plan

TRADE-OFF IN RISK DECISION PLATFORM

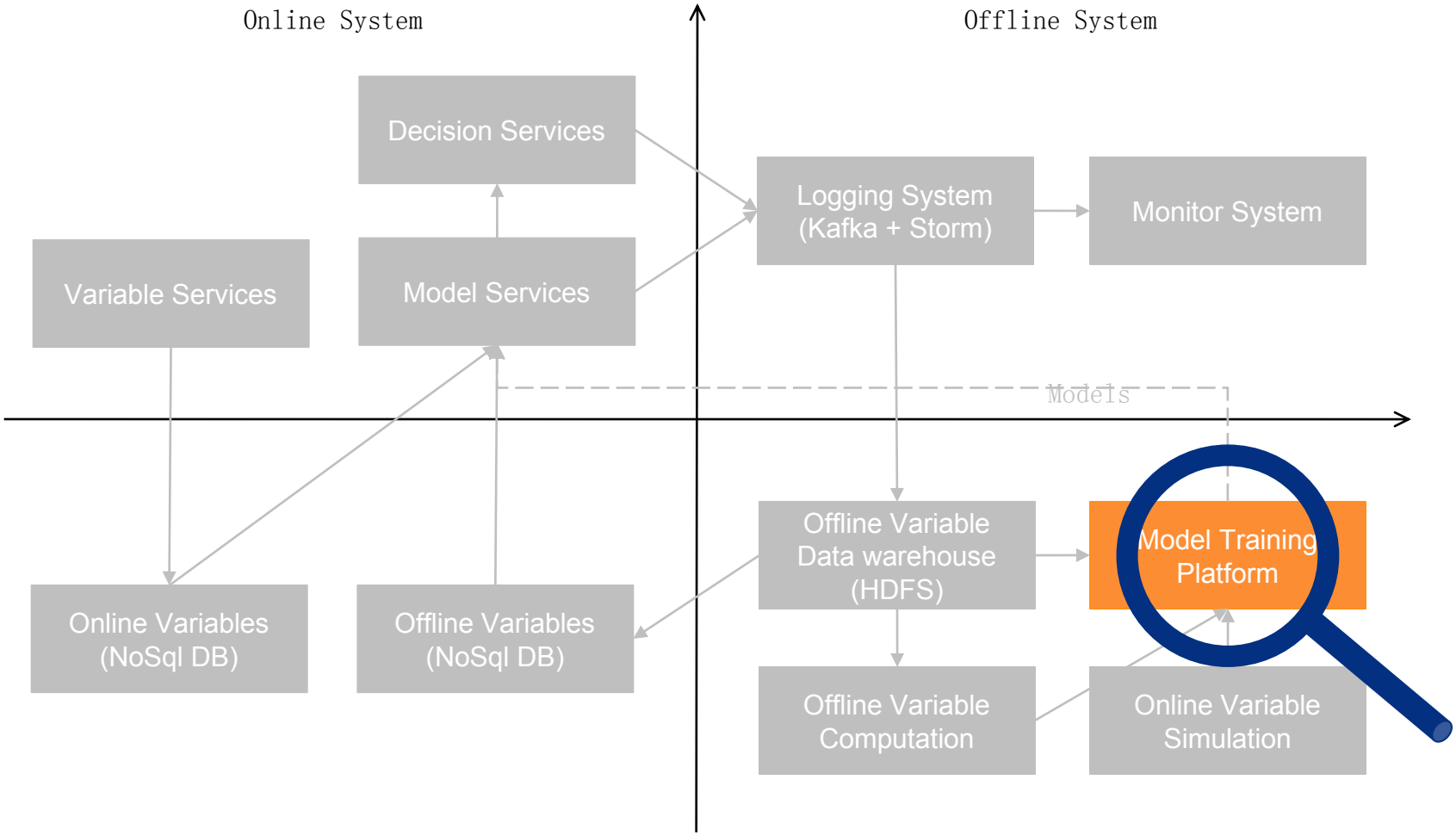
Accuracy, speed & robustness are conflicting requirements!



RISK MODELING ARCHITECTURE

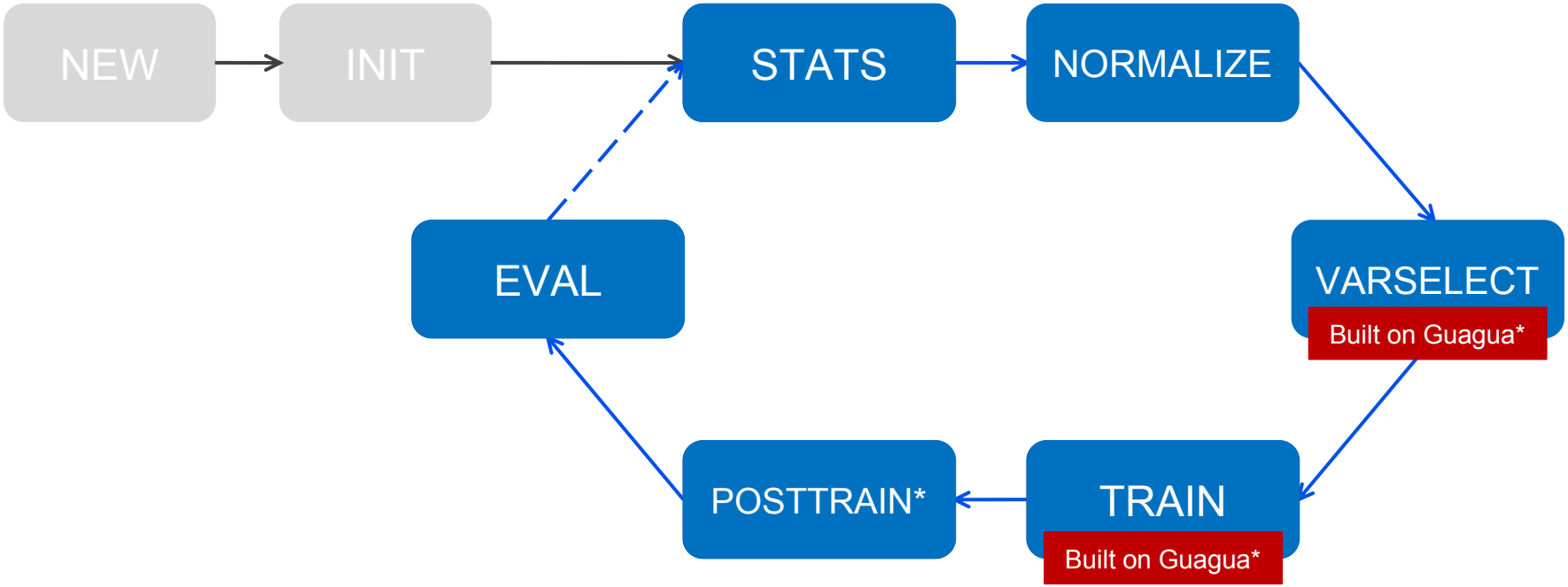


RISK MODELING ARCHITECTURE



SHIFU: COMBINING FEATURE ENGNEERING AND DATA MODELING

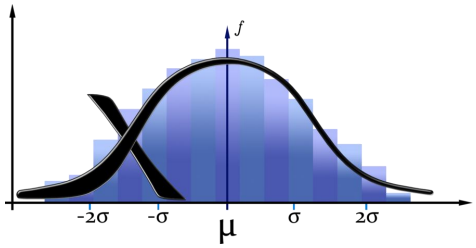
Shifu is an open-source, end-to-end machine learning and data mining framework built on top of Hadoop.



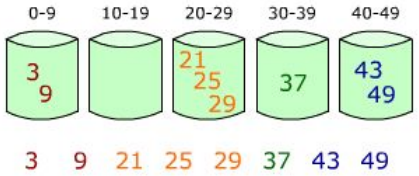
*Guagua is an iterative computing framework for both Hadoop MapReduce and Hadoop YARN

SCALABLE FEATURE ENGINEERING

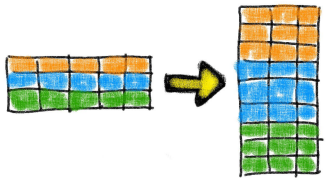
Feature Statistics

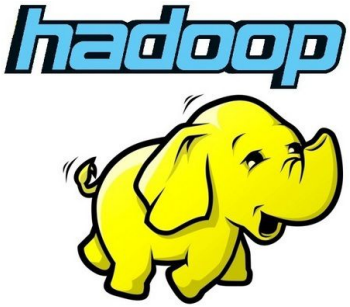


Feature Binning



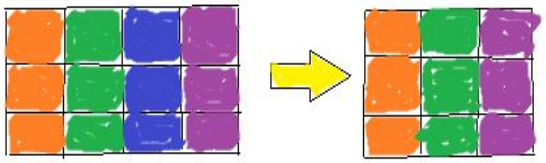
Feature Transform



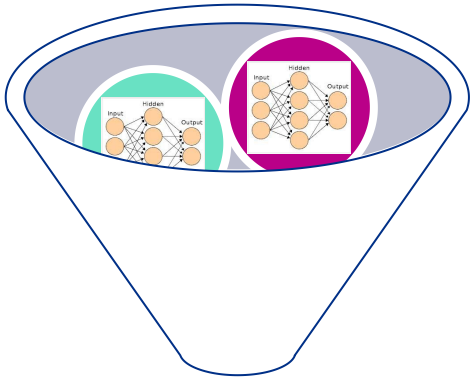


LARGE-SCALE MODELING

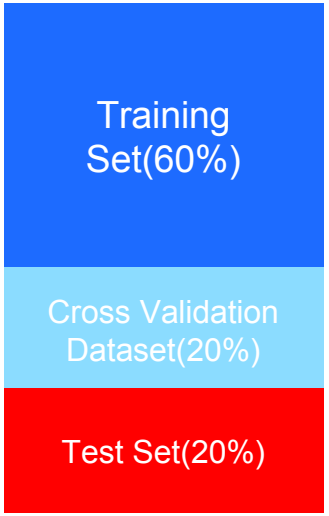
Feature Selection

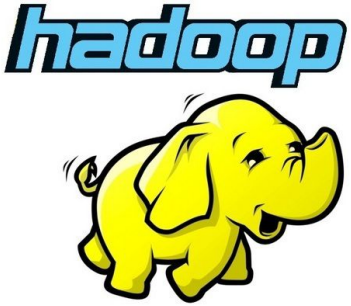


Model Ensemble



Cross Validation

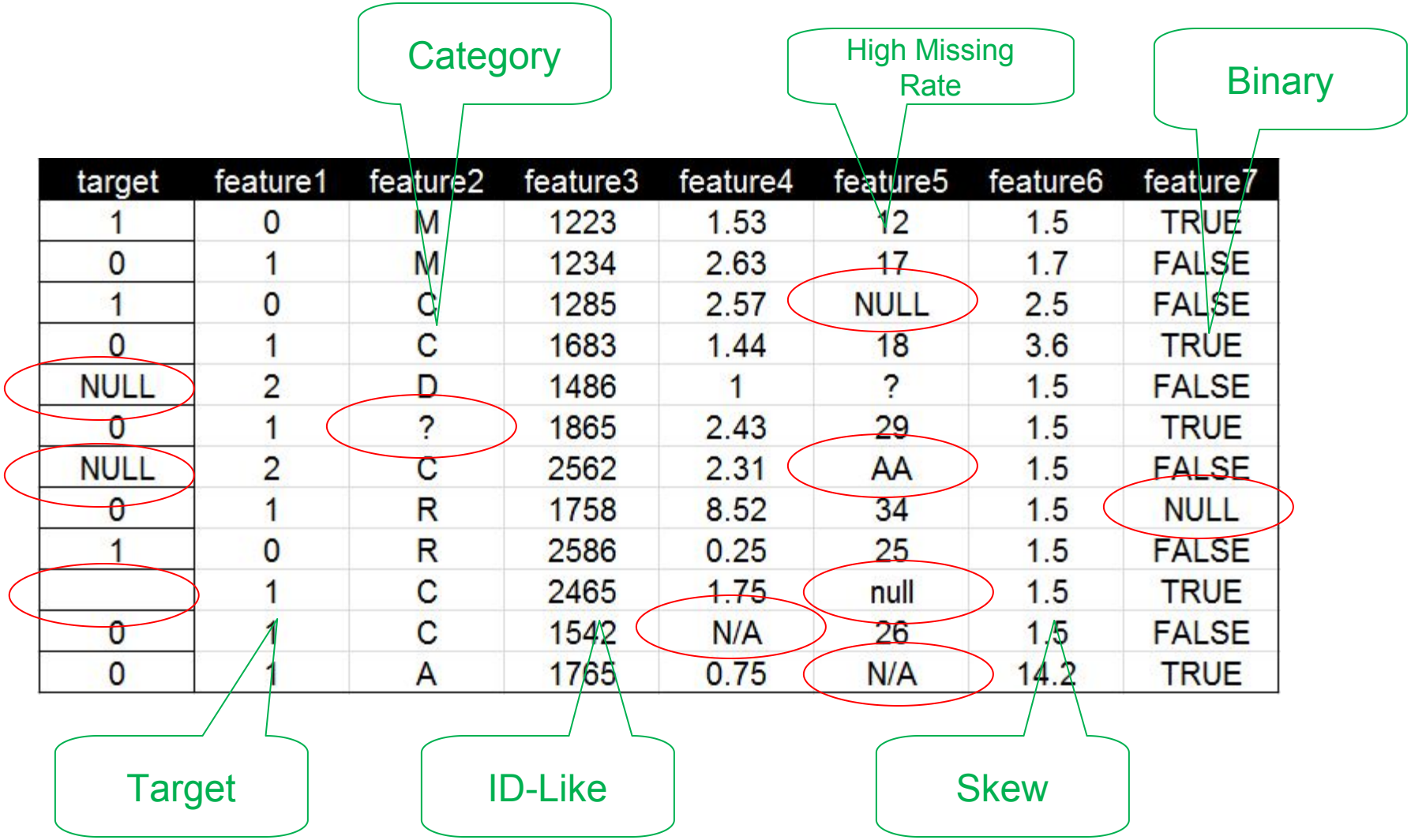




AGENDA

- PayPal & PayPal Risk
- Large Scale Machine Learning Solution
- Feature Engineering & Modeling
- Future Plan

A DATA EXAMPLE



target	feature1	feature2	feature3	feature4	feature5	feature6	feature7
1	0	M	1223	1.53	12	1.5	TRUE
0	1	M	1234	2.63	17	1.7	FALSE
1	0	C	1285	2.57	NULL	2.5	FALSE
0	1	C	1683	1.44	18	3.6	TRUE
NULL	2	D	1486	1	?	1.5	FALSE
0	1	?	1865	2.43	29	1.5	TRUE
NULL	2	C	2562	2.31	AA	1.5	FALSE
0	1	R	1758	8.52	34	1.5	NULL
1	0	R	2586	0.25	25	1.5	FALSE
	1	C	2465	1.75	null	1.5	TRUE
0	1	C	1542	N/A	26	1.5	FALSE
0	1	A	1765	0.75	N/A	14.2	TRUE

BASIC STATISTICS

Unit/Weighted Wised

Mean
Std-dev
Max
Min
Skewness*
Kurtosis*
Missing Rate

Unit/Weighted Binning

Boundaries
Counts
Positive Counts
Negative Counts

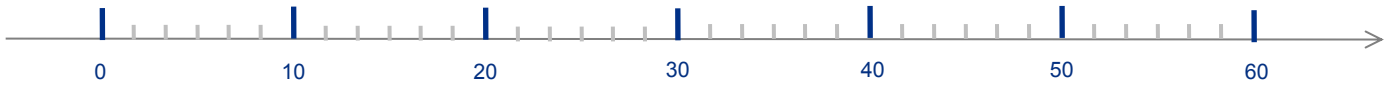
Level 2 Statistics

Positive Rates
Negative Rates
Kolmogorov – Smirnov Values
Information Values
Weight of Evidence Values

Skewness & Kurtosis: http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

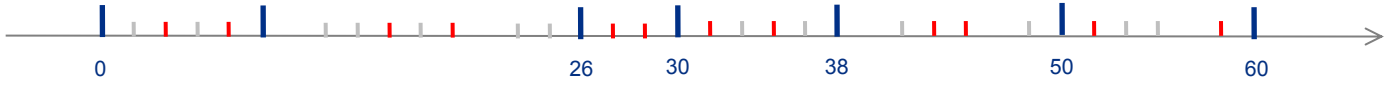
MULTIPLE BINNING METHODS

Equal Interval



Equal Total (Each bin with three elements)

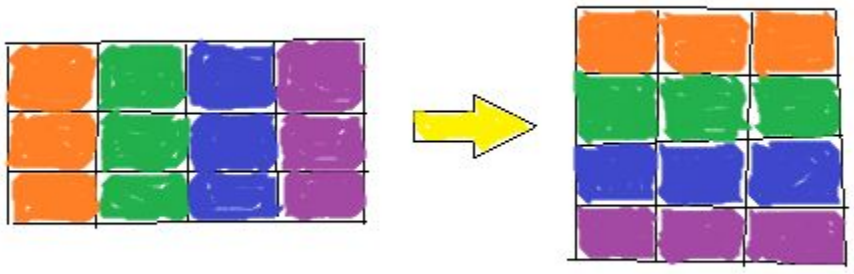
Equal Positive (Each bin with two positive (red) elements)



Equal Negative (Each bin with two negative (green) elements)

DEFAULT BINNING ALGORITHM (SORT)

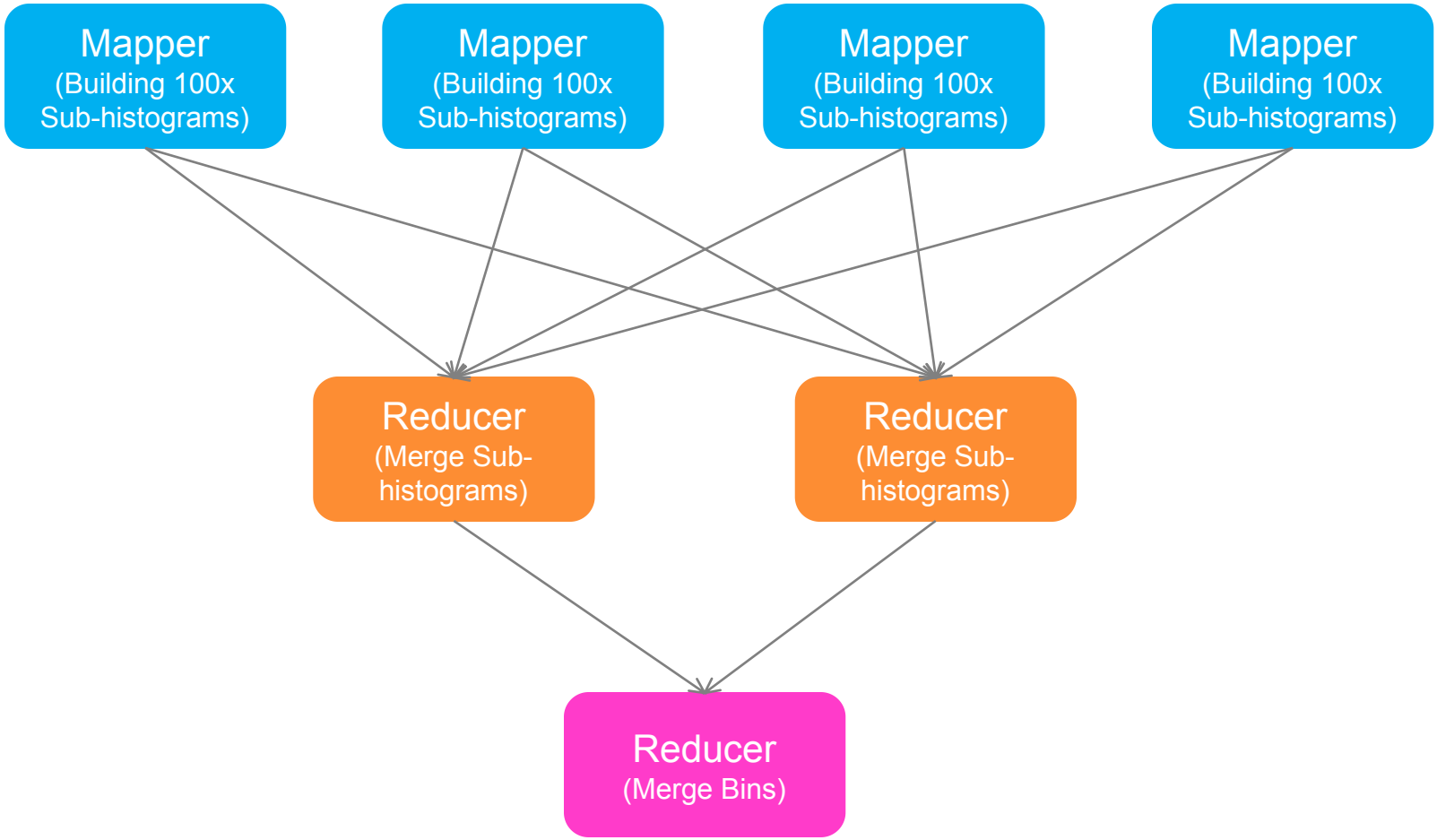
1. Rotate Columns to Rows



2. Sort Each Row & Scan to Pick Binning Boundaries

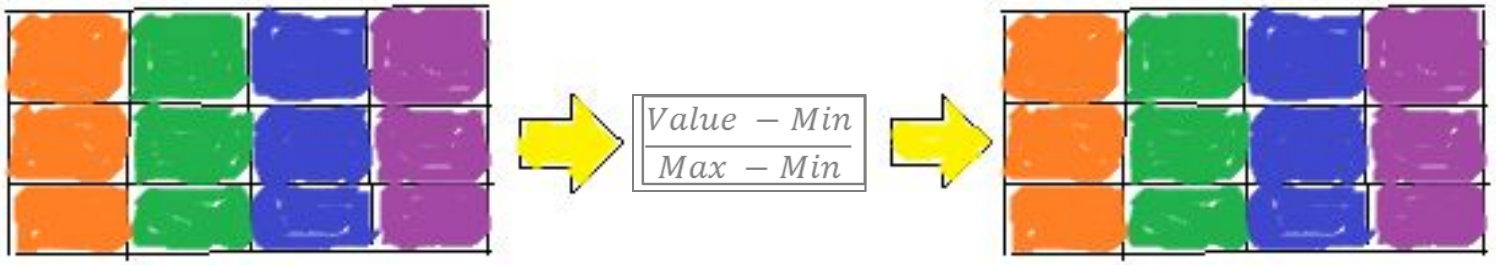
Issue: Not scalable on each row. Sometimes sampling must be enabled.

SCALABLE BINNING ALGORITHM(HISTOGRAM)

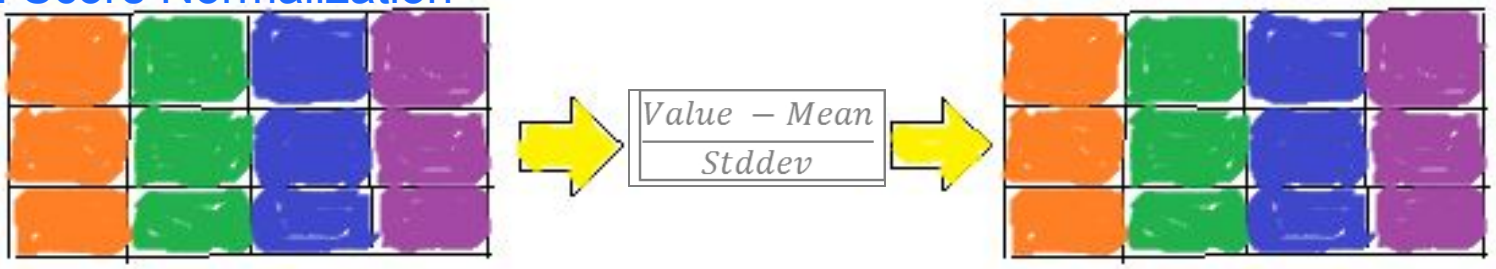


MULTIPLE FEATURE NORMALIZATION

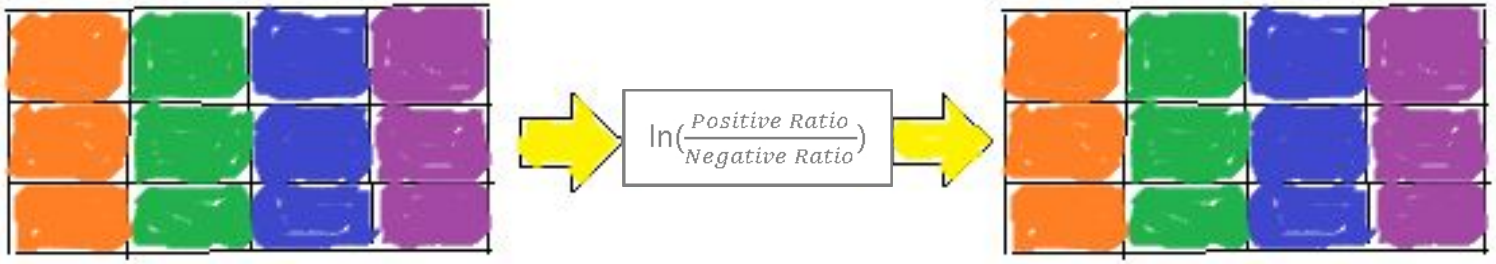
Max-Min Normalization



Z-Score Normalization

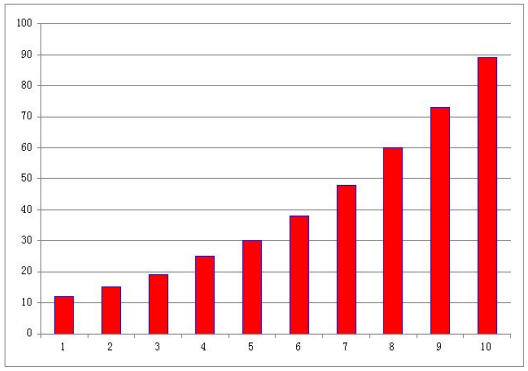


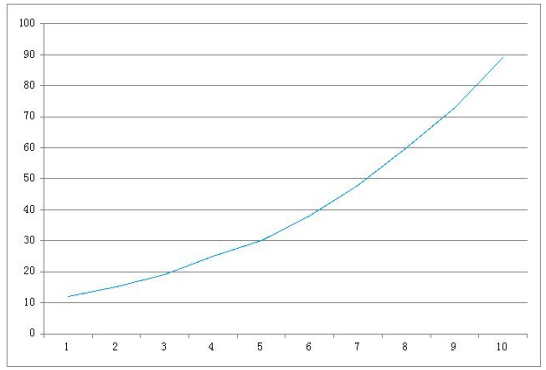
WoE* (Z-Score) Normalization



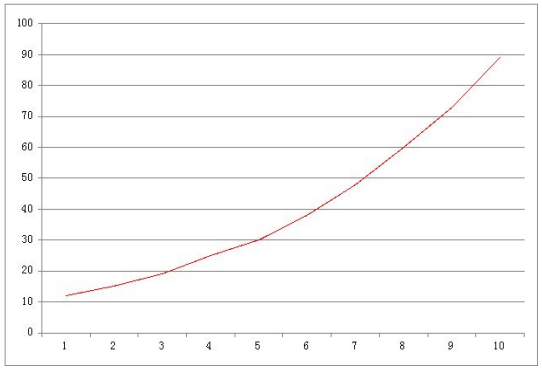
*WoE: Weight of Evidence http://support.sas.com/resources/papers/proceedings13/095-2013.pdf

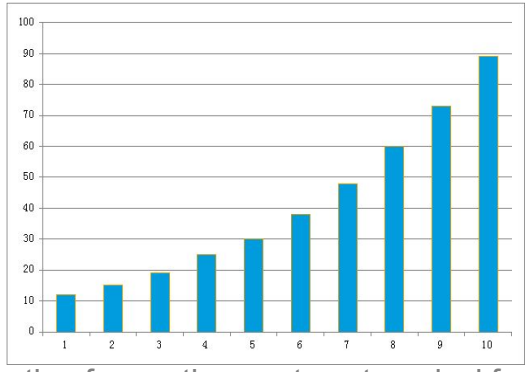
CATEGORY FEATURE ⟺ CONTINUOUS FEATURE





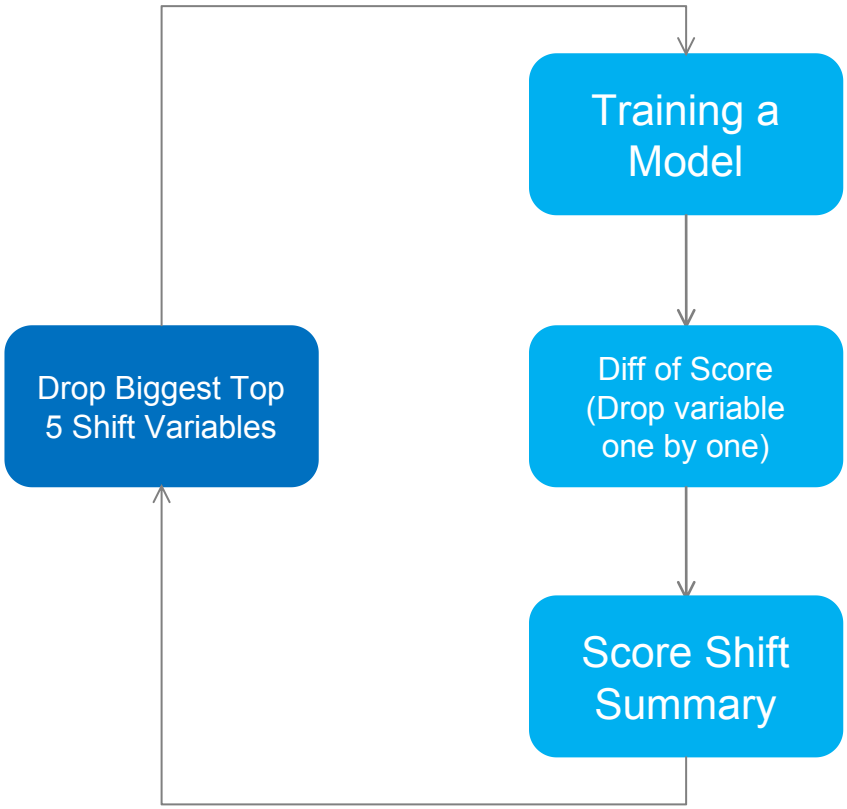
One-hot encoding, Negative rate or WoE value are used in normalization for categorical to continuous feature.





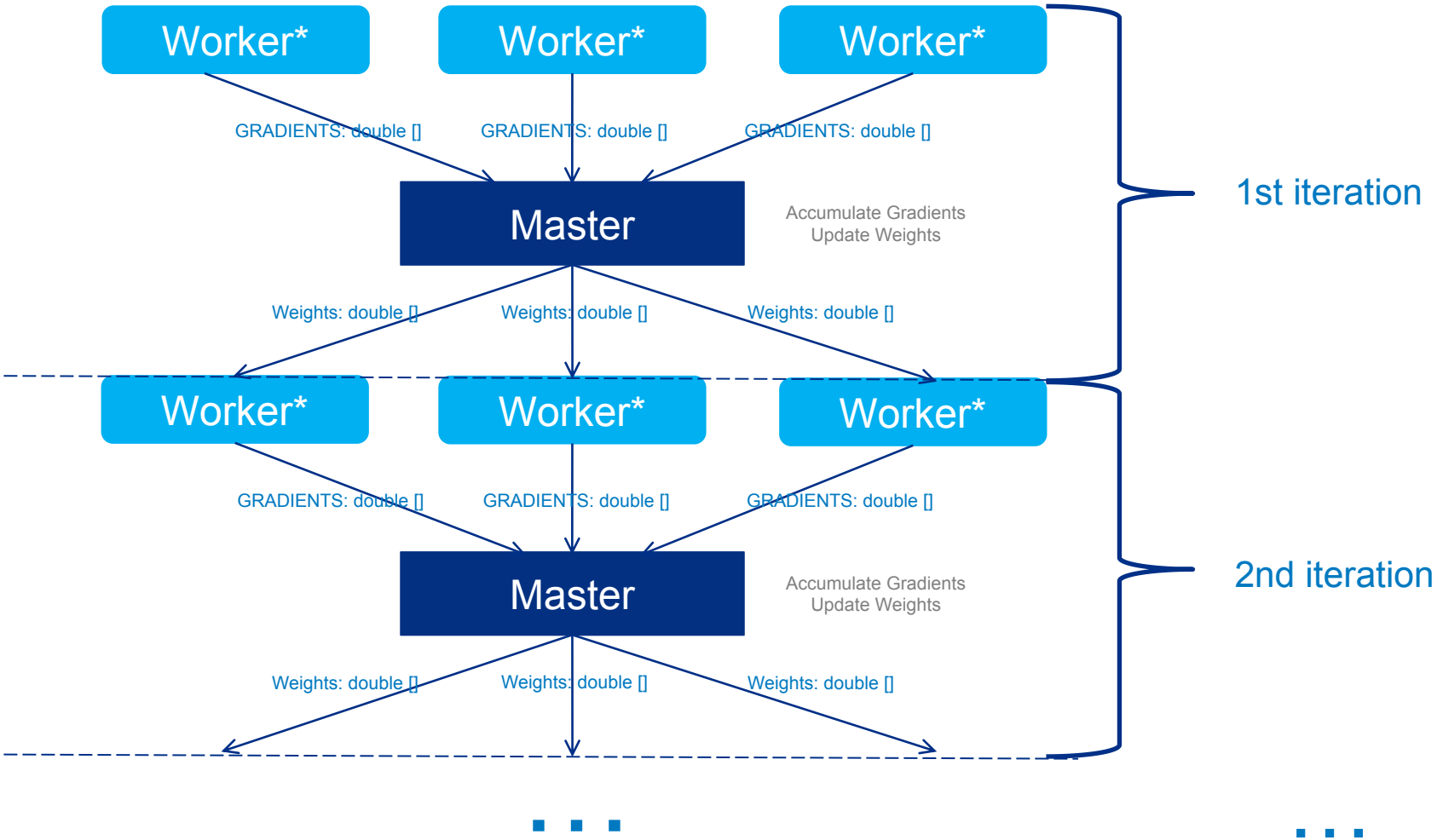
WoE value or Binning threshold are used in normalization for continuous to categorical feature.

SENSITIVITY ANALYSIS FOR VARIABLE SELECTION



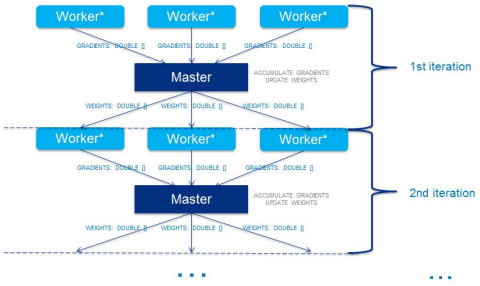
* Sensitivity analysis report are also saved as a file in each round.

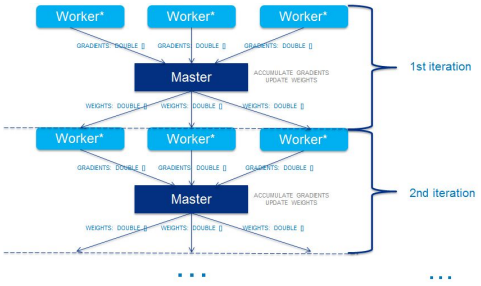
DISTRIBUTED NEURAL NETWORK TRAINING*

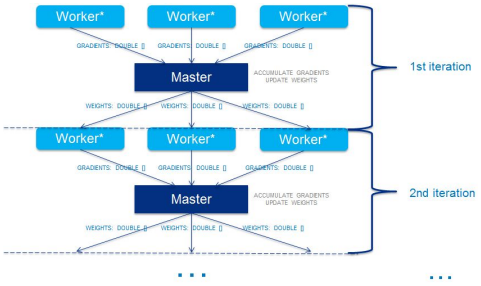


* Distributed batch gradient descent algorithm

BAGGING & CROSS VALIDATION & EVALUATION







AGENDA

- PayPal & PayPal Risk
- Large Scale Machine Learning Solution
- Feature Engineering & Modeling
- Future Plan

FUTURE PLAN

- Better Usability
- Clear & Pluggable Model Ensemble Module
- Spark Migration for ‘stats’ and ‘eval’ steps (WIP)
- Restricted Boltzmann Machine (WIP)
- Hash Variable Encoding (WIP)

Thank You