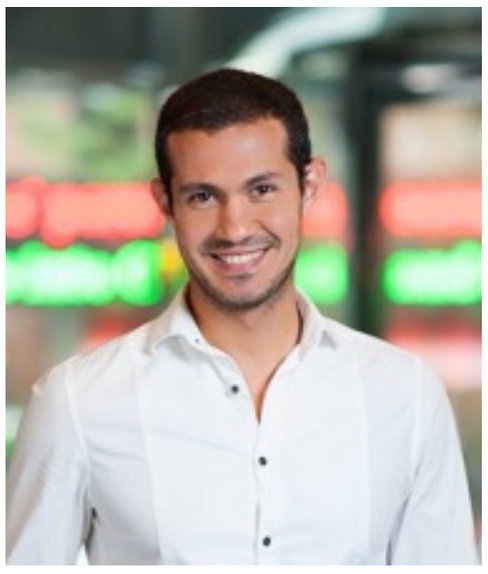
ThoughtWorks®

(Machine) Learning To Detect Fraudsters



Hany Elemary

Sarah LeBlanc

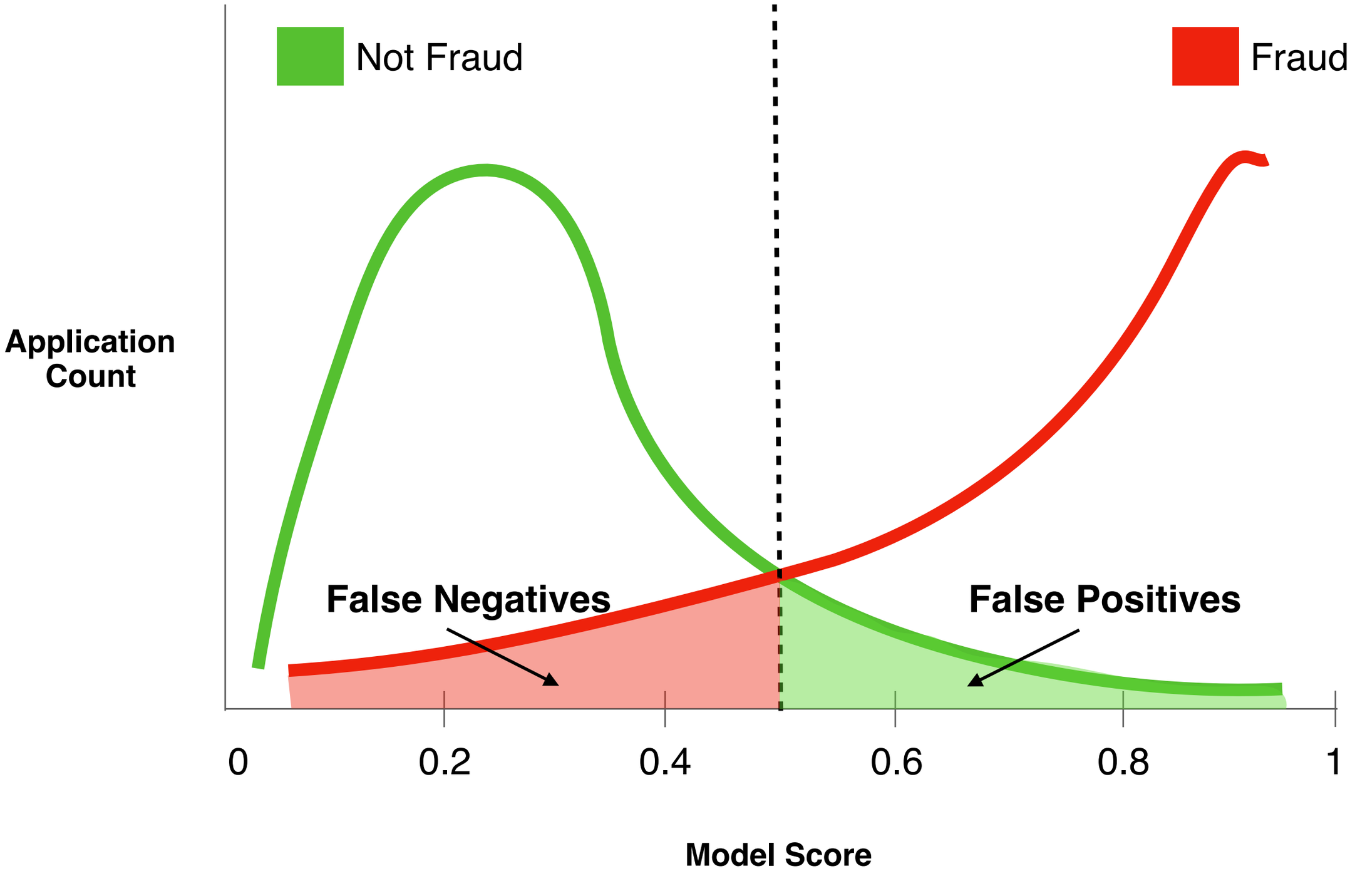
CREDIT CARD FRAUD

TRANSACTION

APPLICATION

CARD NOT FOUND

FRAUD DETECTION MODEL PLOT



MINIMIZE LOSSES

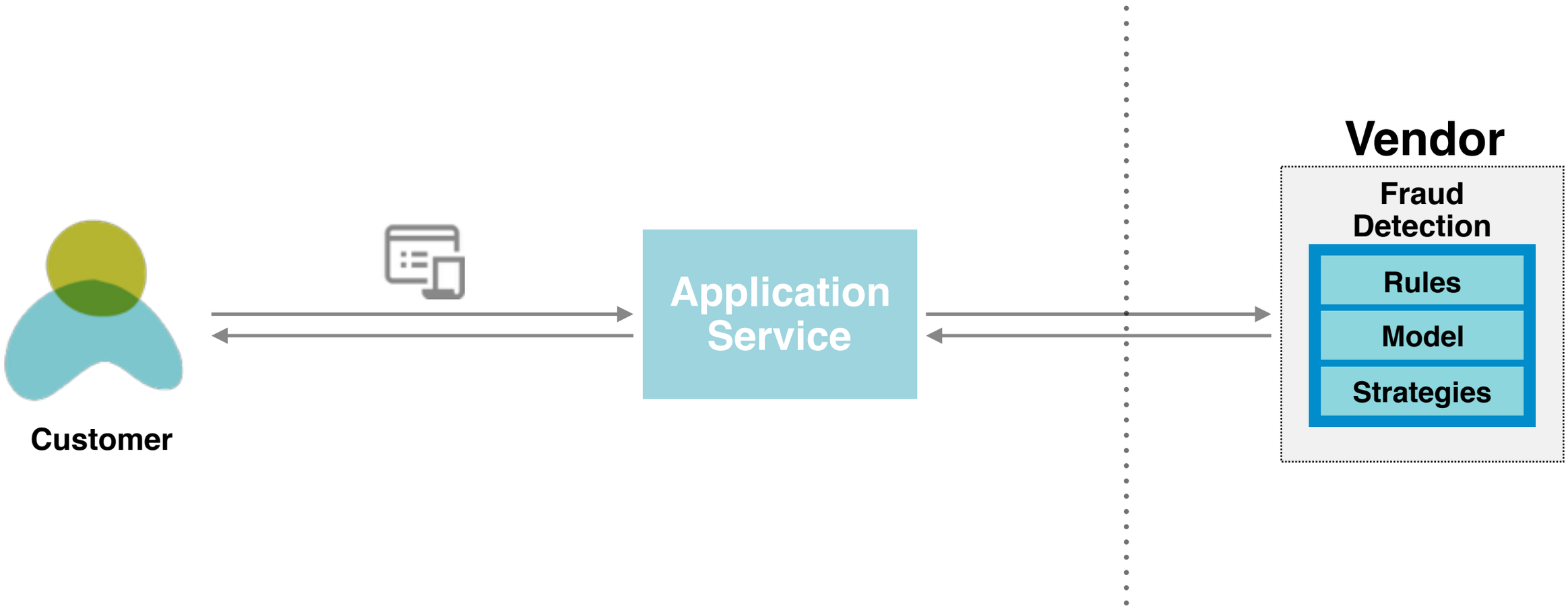
$$\text{Lost Profitability} = (\text{Fraud Cost} * \textit{FN}) + (\text{Opportunity Cost} * \textit{FP})$$

Legend:

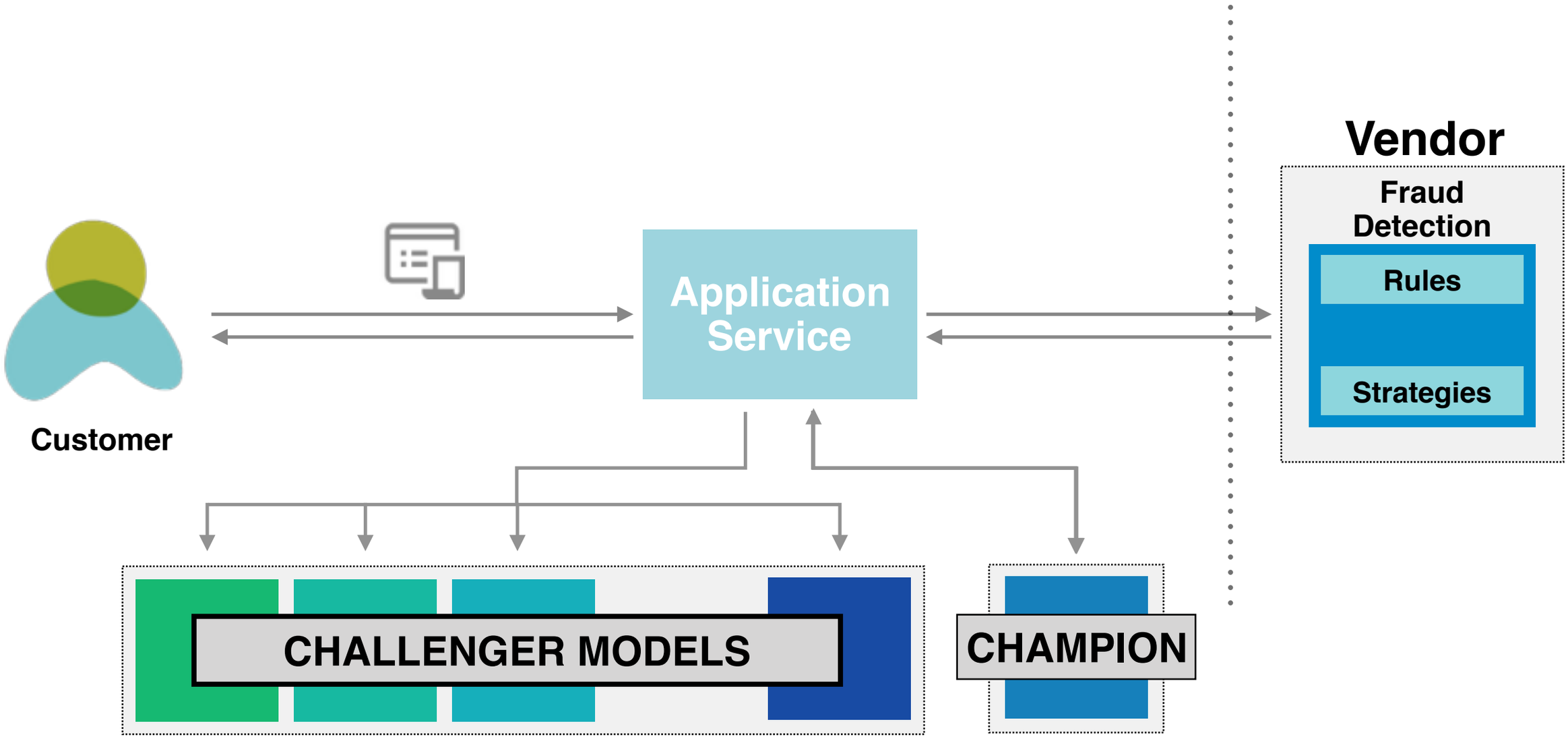
FN (Fraud missed)

FP (Mistaken fraud)

CURRENT STATE

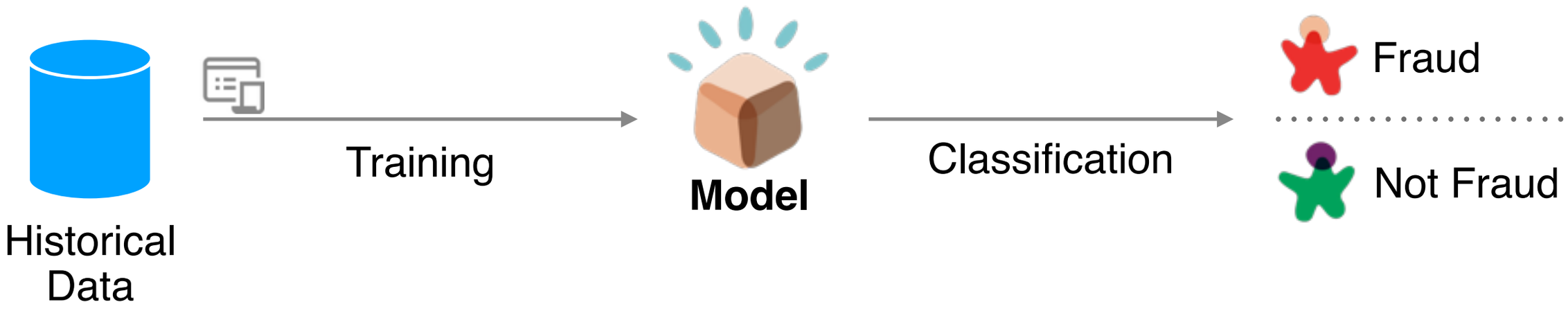


PROPOSED STATE

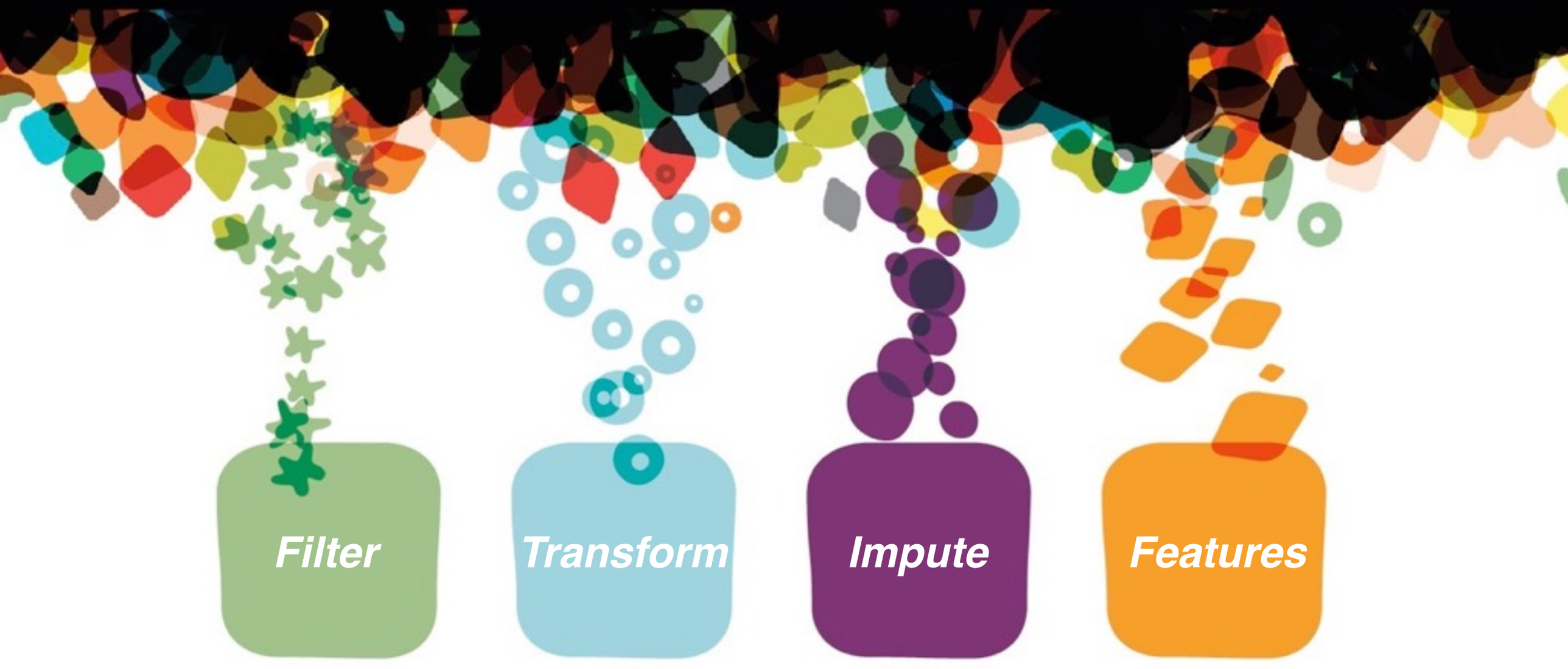


MODEL TRAINING

Supervised Learning

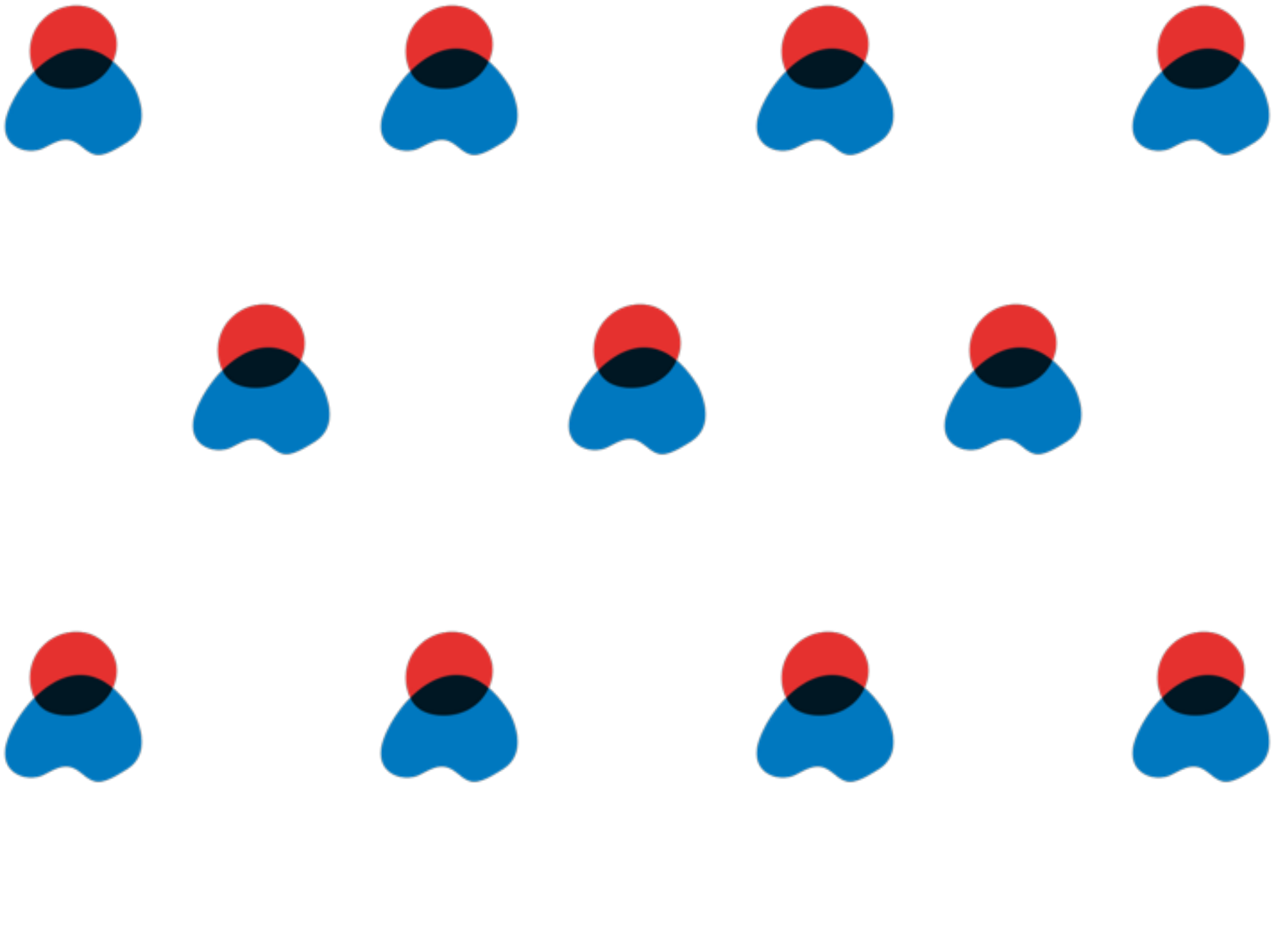


DATA PATTERNS



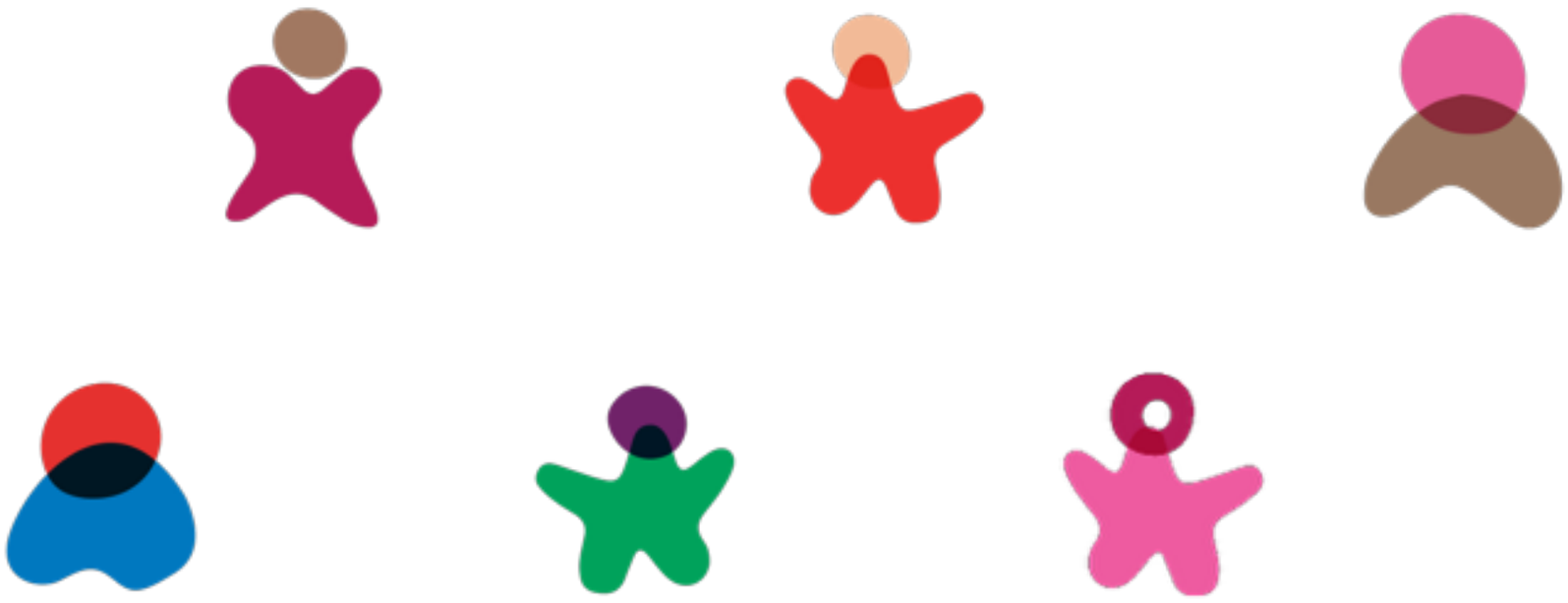
DATA FILTERING

Low Cardinality



DATA FILTERING

High Cardinality

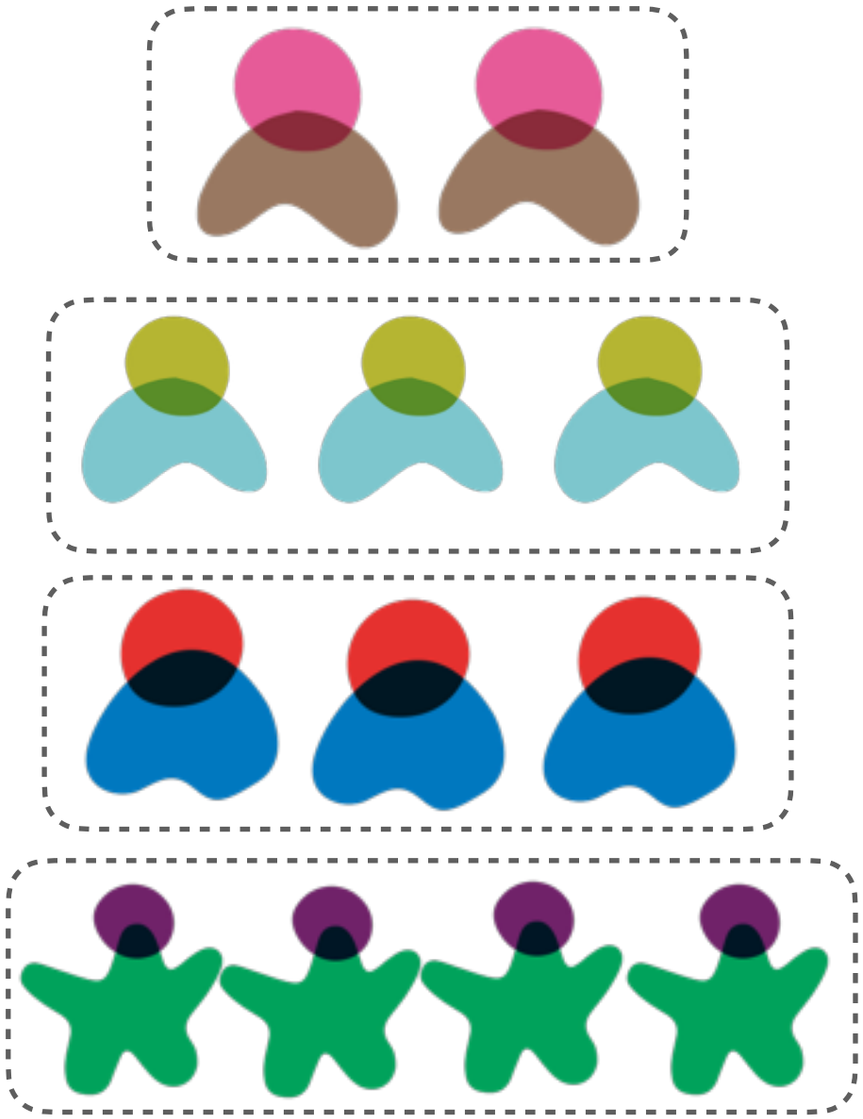


DATA FILTERING

Medium Cardinality

DATA FILTERING

Medium Cardinality



Training

Predictive Model

DATA TRANSFORMATION

Email	Fraud Status
jack.smith@gmail.com	
annie.may@fraudster.com	
freddy.jr@gmail.com	
nicole.jack@fraudster.com	
jon.johnston@gmail.com	
claudia.penns@us.gov	
walter.carson@gmail.com	
ben.benjamin@fraudster.com	

DATA TRANSFORMATION

Domain name	Fraud Status
gmail.com	✔
fraudster.com	⚠
gmail.com	✔
fraudster.com	⚠
gmail.com	✔
us.gov	✔
gmail.com	✔
fraudster.com	⚠

DATA IMPUTATION

Handling Missing Data

Column 1	Column 2	Column 3	Column 4
			
			
			
			

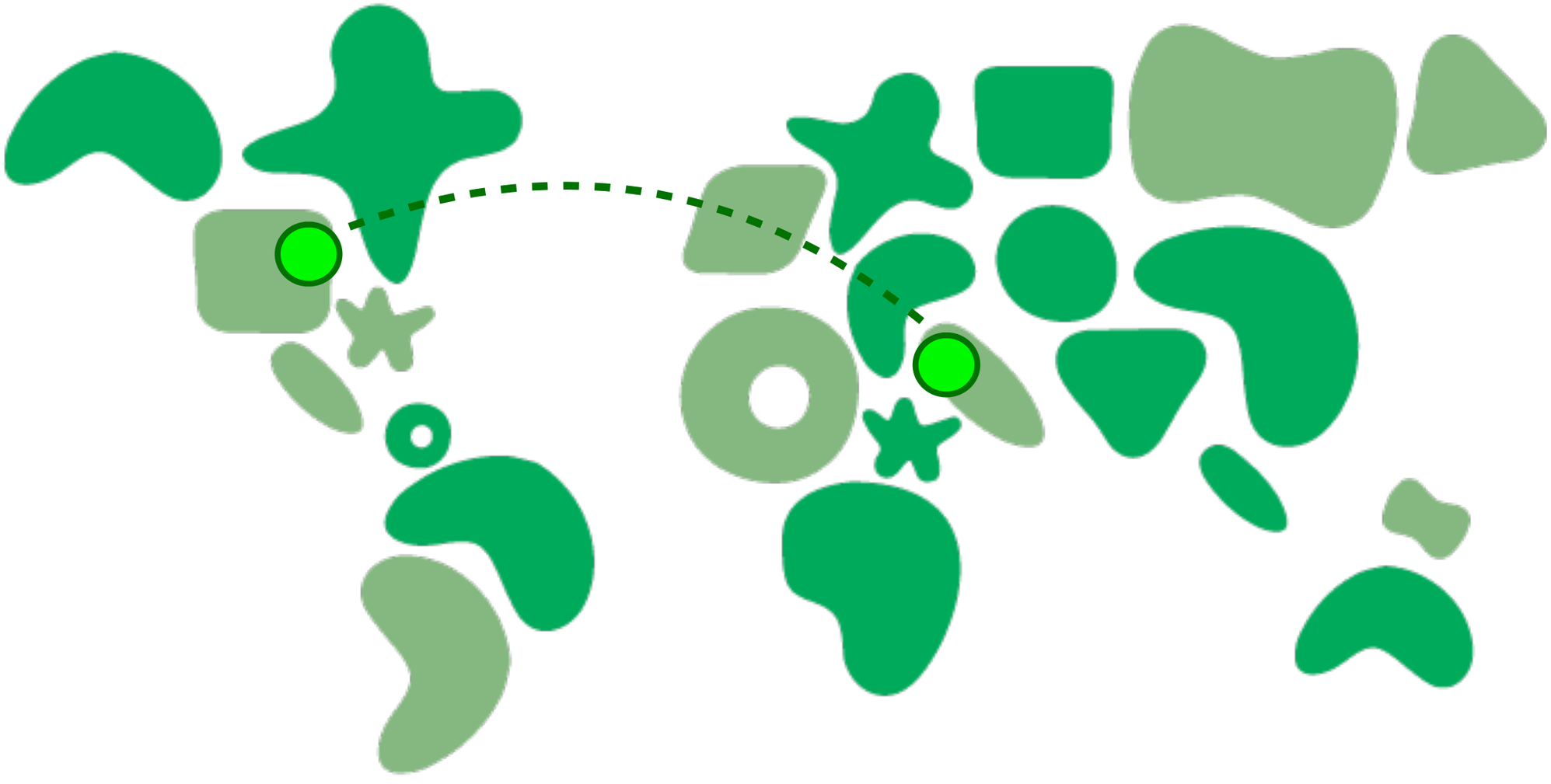
DATA IMPUTATION

Handling Missing Data

Column 1	Column 2	Column 3	Column 4
			
			
			
			

FEATURE SELECTION

IP to Zip Proximity



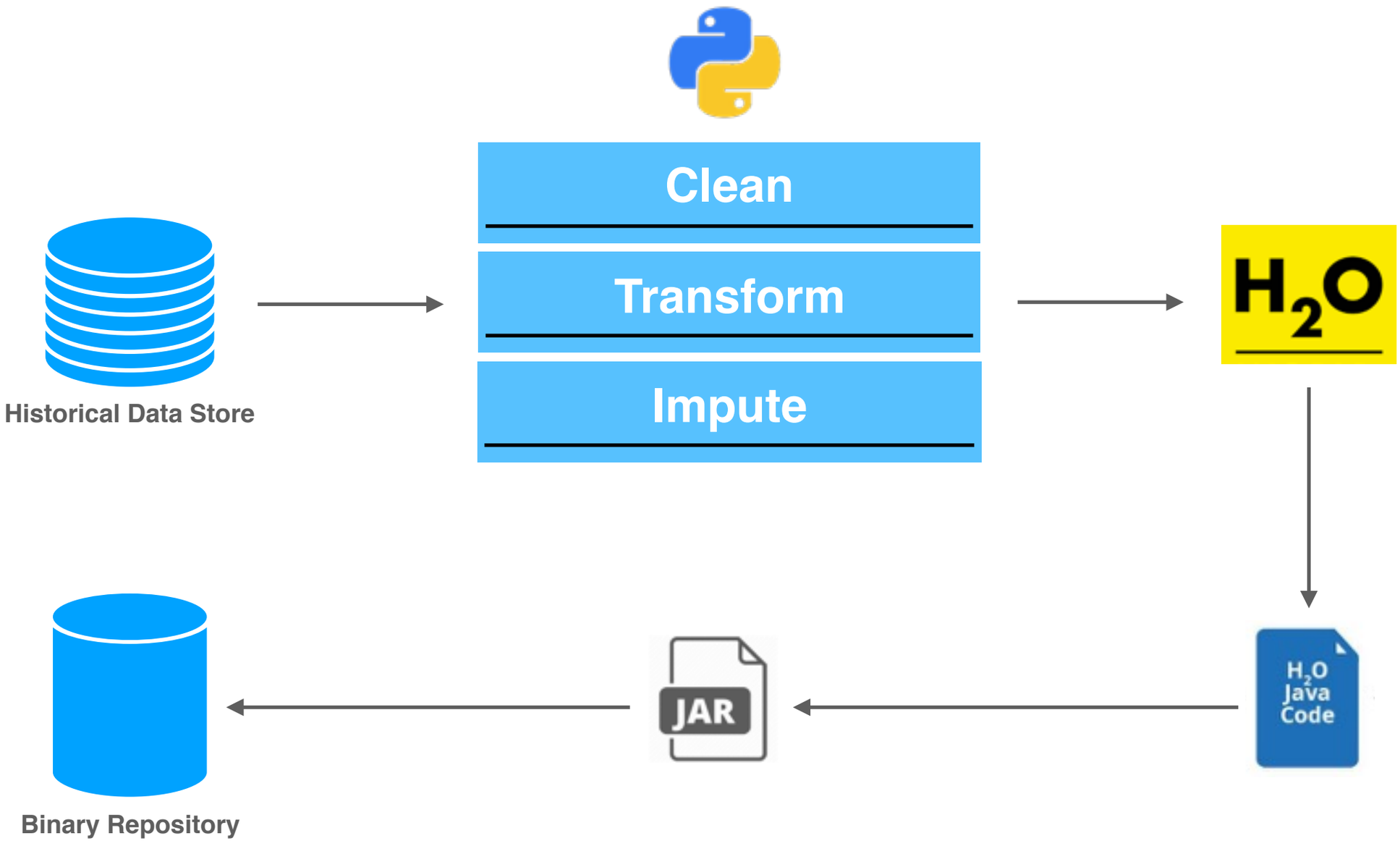
ARCHITECTURE

DATA SCIENTIST WORKFLOW

DEVELOPER WORKFLOW

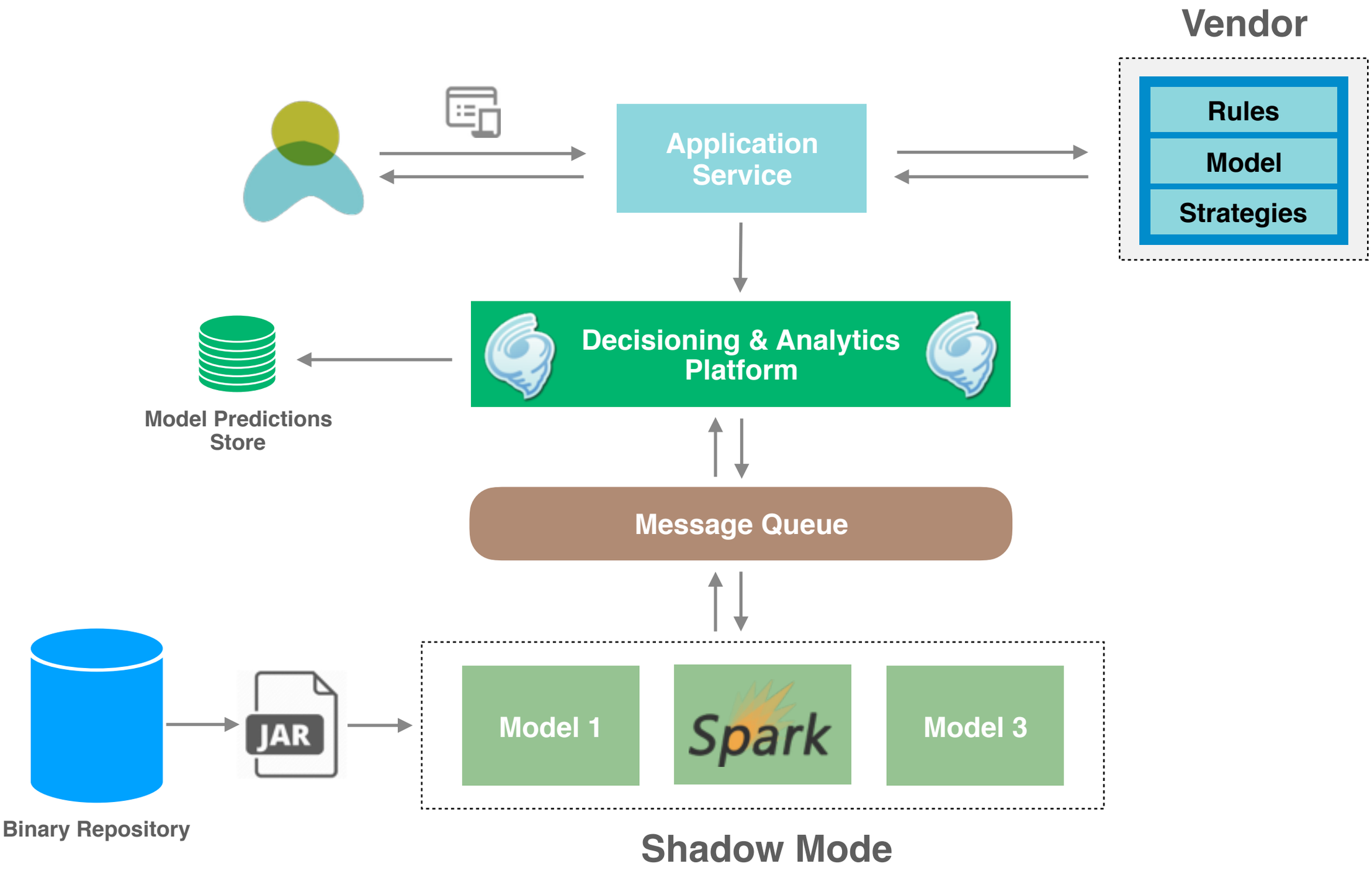
DATA SCIENTIST WORKFLOW

Raw Data → Transformed Data → Trained Model



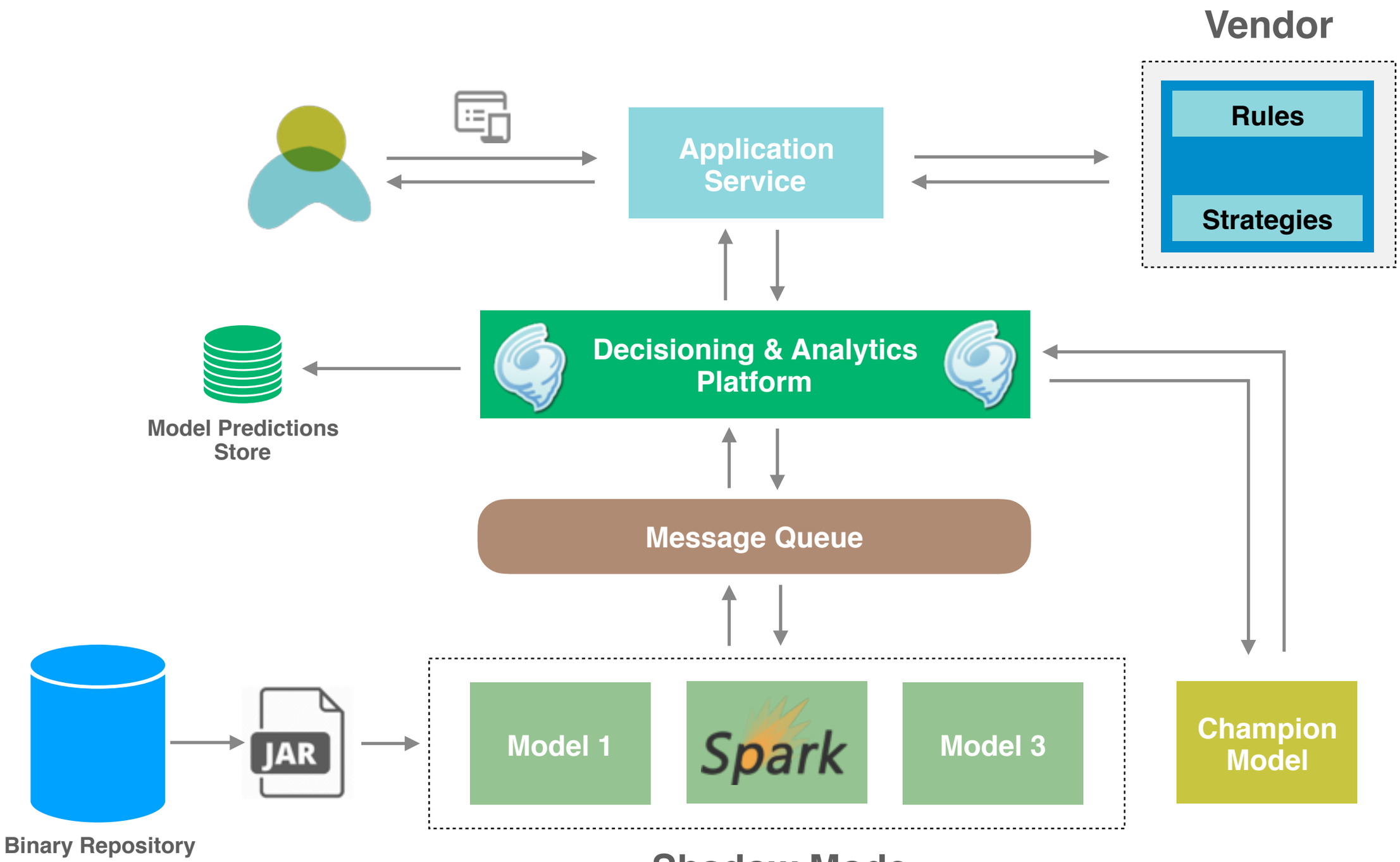
DEVELOPER WORKFLOW

Applications → Trained Model → Score



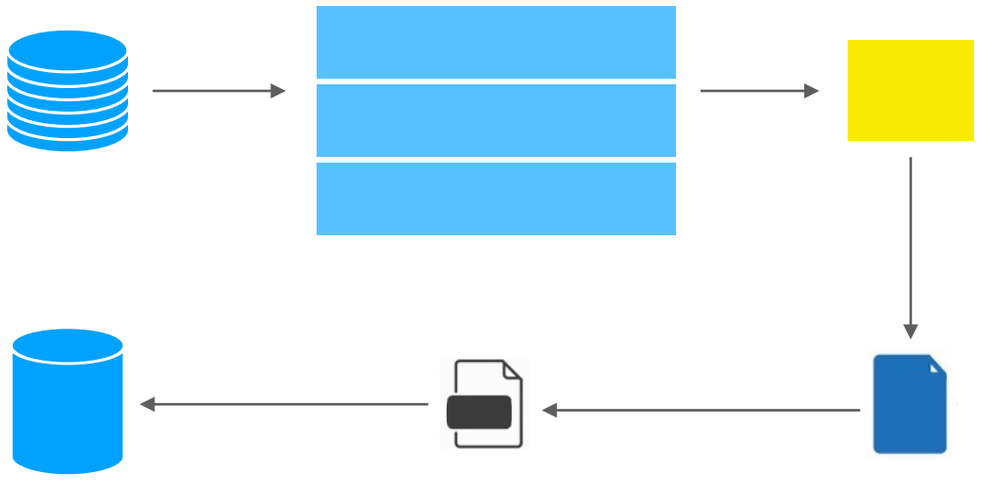
DEVELOPER WORKFLOW

Applications → Trained Model → Score

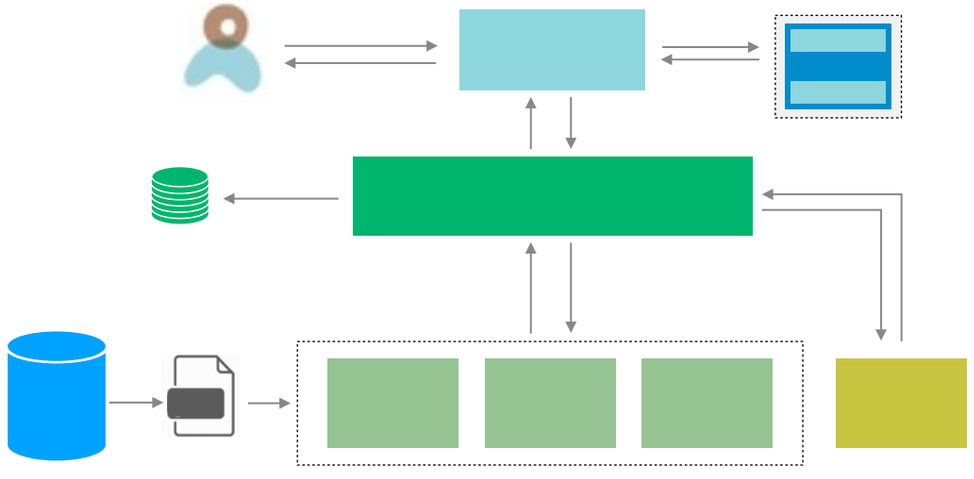


ARCHITECTURE

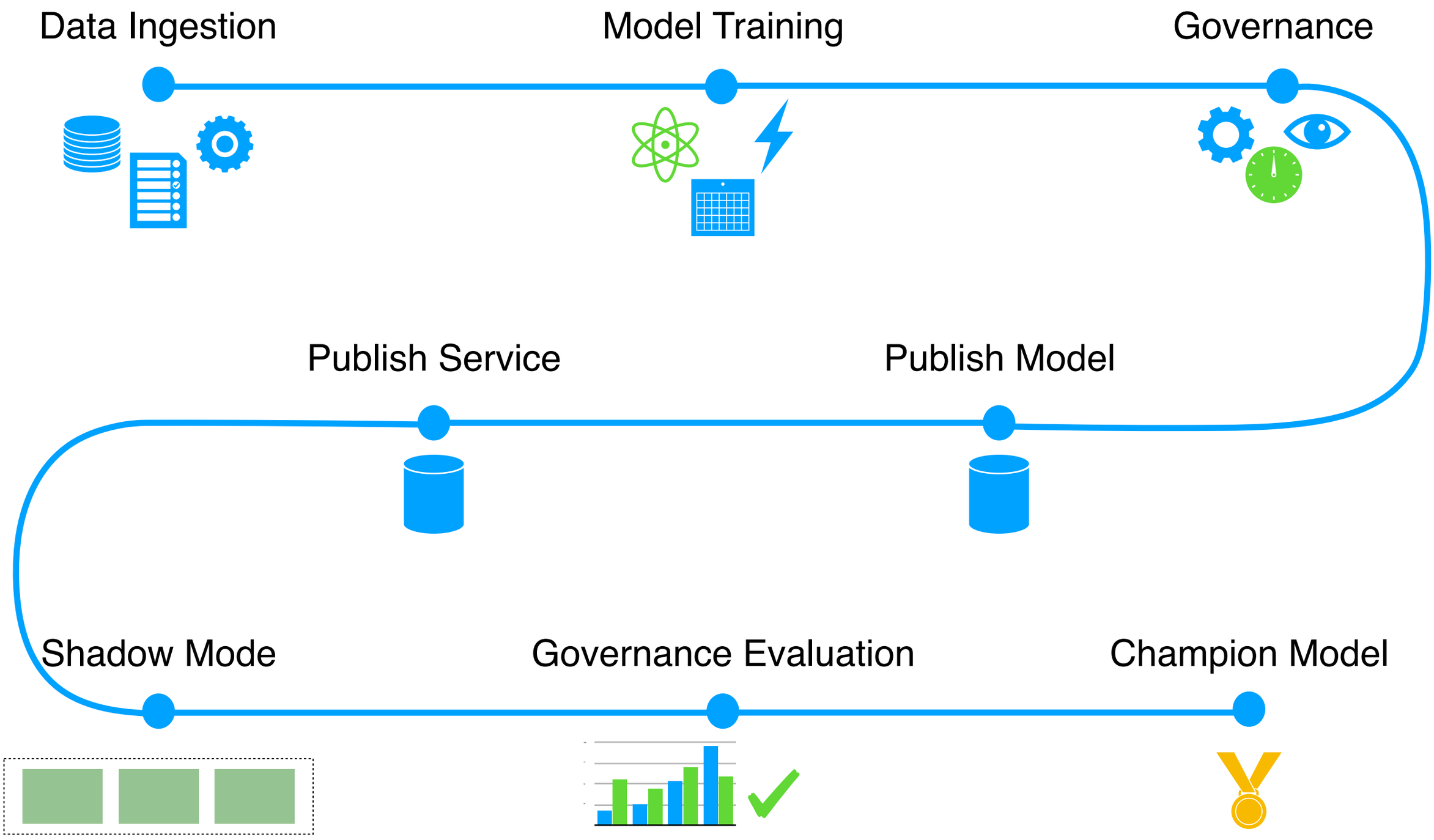
DATA SCIENTIST WORKFLOW



DEVELOPER WORKFLOW



VALUE STREAM



THANK YOU

Sarah LeBlanc

sleblanc@thoughtworks.com

@sarah_g_leblanc

Hany Elemary

helemary@thoughtworks.com

@hanyelemary

Questions?