# AMAZON MACHINE LEARNING

BINARY CLASSIFICATION
"IS IT SPAM OR NOT"
TRUE - OR - FALSE

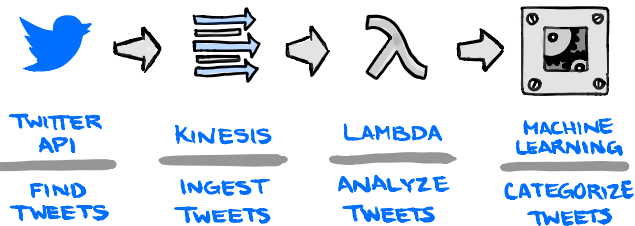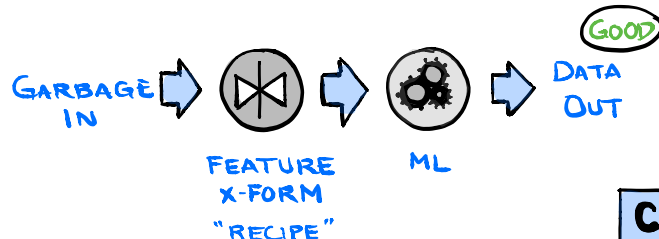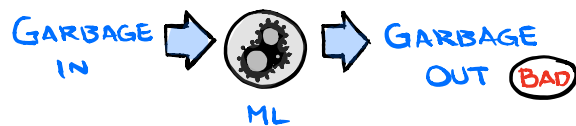$\frac{\text{MORE DATA}}{\text{BETTER MODEL}}$

Models

REGRESSION
"WHEN WILL IT HAPPEN AGAIN?"

MULTI-CLASS CLASSIFICATION
"WHICH ORG SHOULD HANDLE THIS TWEET?"

ML TRAINING DATA IS KEY DIFFERENTIATOR

## USE CASES
- FRAUD DETECTION
- DOC CLASSIFICATION
- SENTIMENT ANALYSIS
- DEMAND FORCASTING
- CLICK PREDICTION
- CUSTOMER CHURN
- MARKETING INTEL

GARBAGE IN → ML → GARBAGE OUT  BAD

GARBAGE IN → FEATURE X-FORM "RECIPE" → ML → DATA OUT  GOOD

TWITTER API — FIND TWEETS
KINESIS — INGEST TWEETS
LAMBDA — ANALYZE TWEETS
MACHINE LEARNING — CATEGORIZE TWEETS

## EXAMPLE: CATEGORIZE TWEETS

US EAST (N. VIRGINIA)

COST: PAY AS YOU GO

### COMPUTE FEES
DATA ANALYSIS & MODEL BUILDING ........ 42¢/HR

### PREDICTION FEES
| BATCH | OR | REAL-TIME |
|-------|----|-----------|
| 10¢/1K | | $\frac{1}{100}$¢/EA + $\frac{1}{10}$¢/HR/10 MB MEMORY |

No PhD Required

## SUPERVISED LEARNING
- HISTORICAL DATA
- KNOWN INPUTS
- KNOWN OUTCOMES

UNSEEN INPUT → ML → SAME OUTCOME
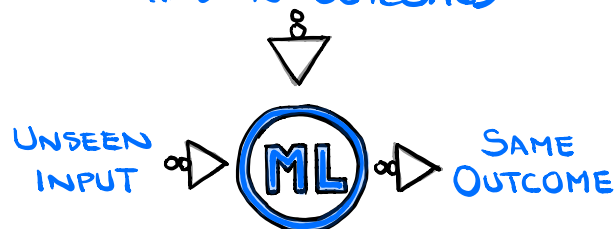
WHAT IS COST OF FALSE POSITIVE?

WHAT IS COST OF FALSE NEGATIVE?

S3   RDS   REDSHIFT

DATASET
70% TRAINING | 30% EVAL
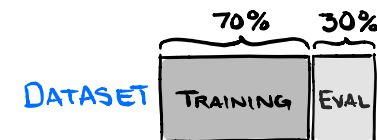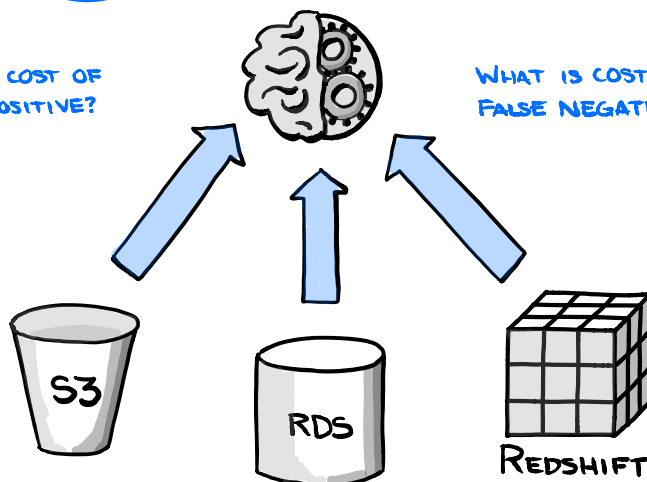
PREVENT MODEL "OVERFIT"

@awsgeek 9/22/17