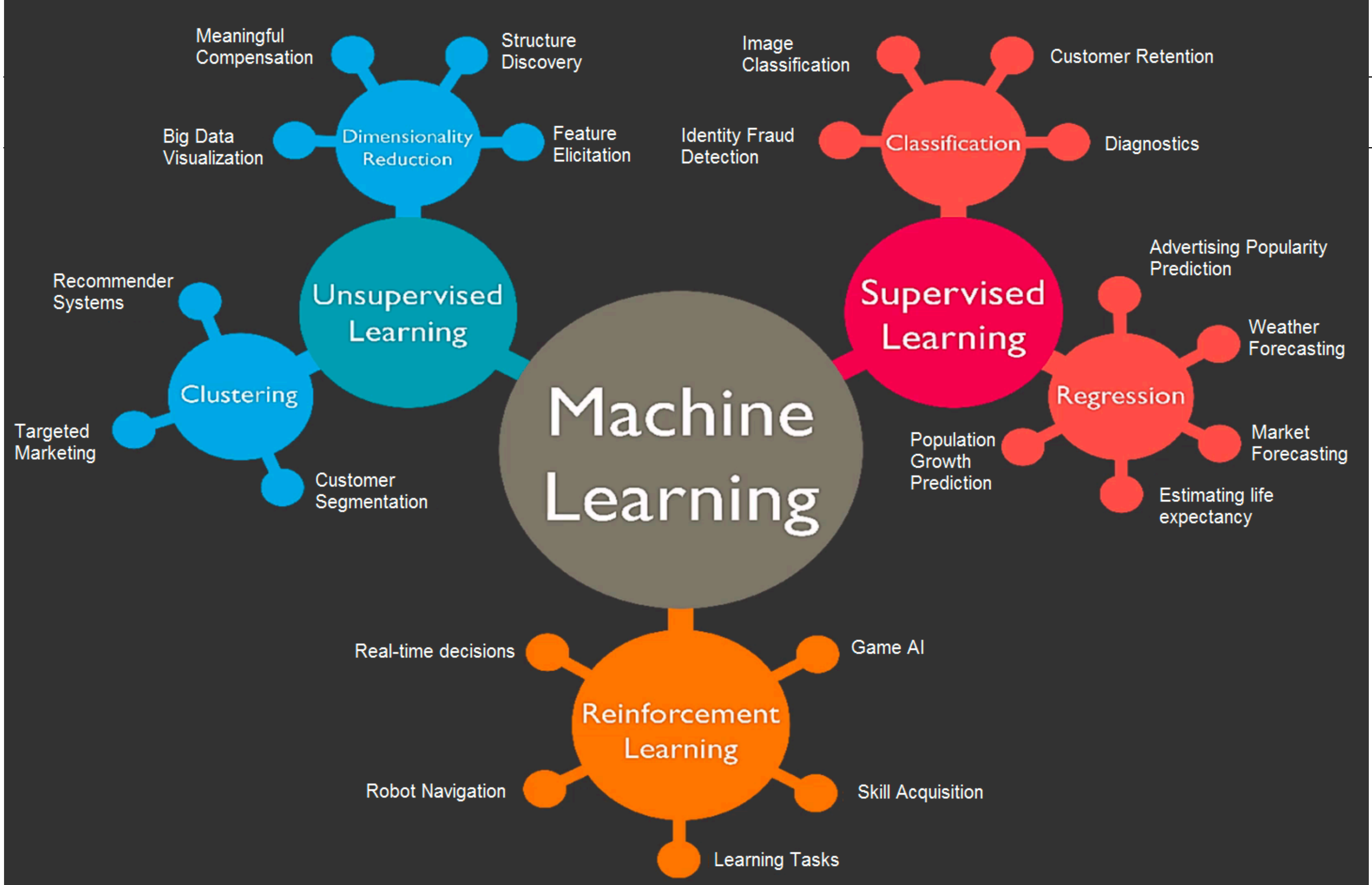
INTRODUCTION TO LOGISTIC REGRESSION

OPENING

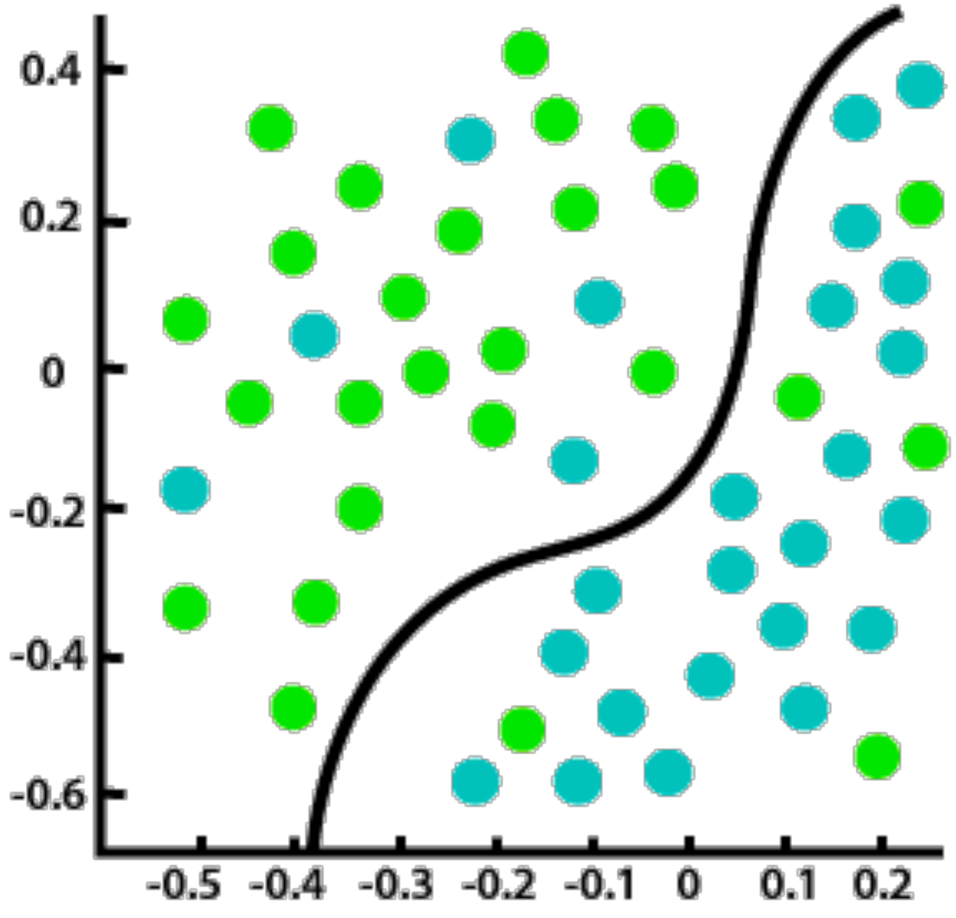
INTRO TO CLASSIFICATION


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compensation
Structure Discovery
Big Data Visualization
Feature Elicitation
Clustering
Recommender Systems
Targeted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Identity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

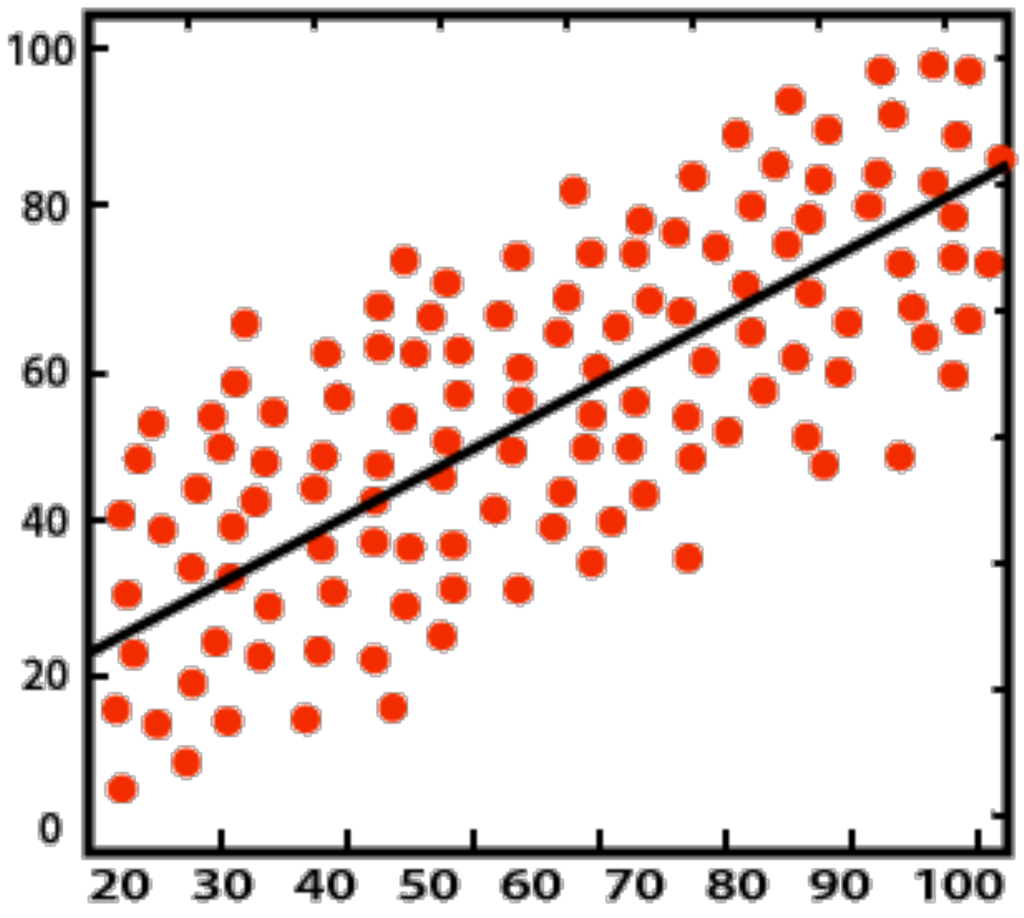
WHERE ARE WE?

	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

CLASSIFICATION VS REGRESSION

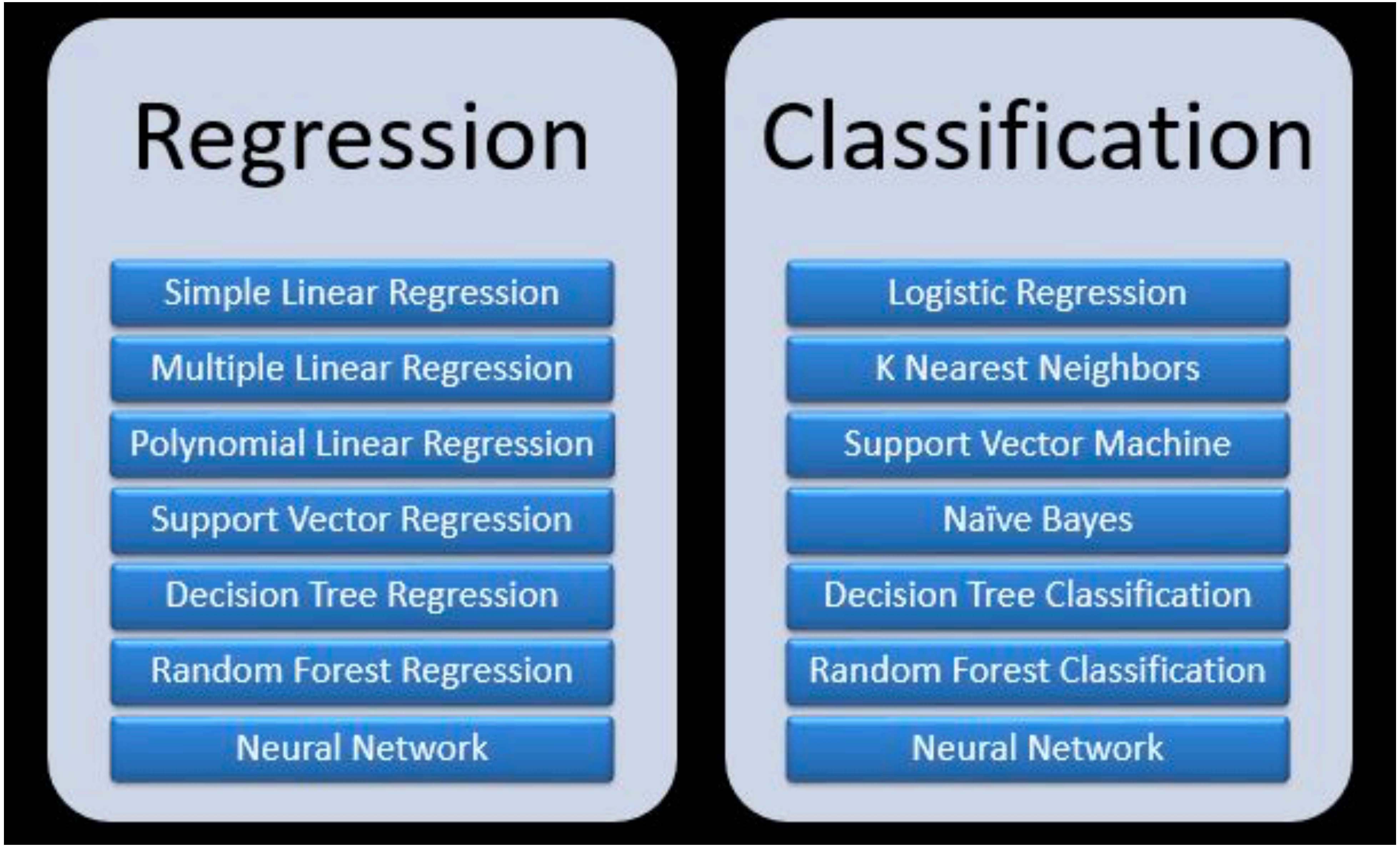


Classification



Regression

COMMON ALGORITHMS



WHAT IS THE OUTCOME VARIABLE?

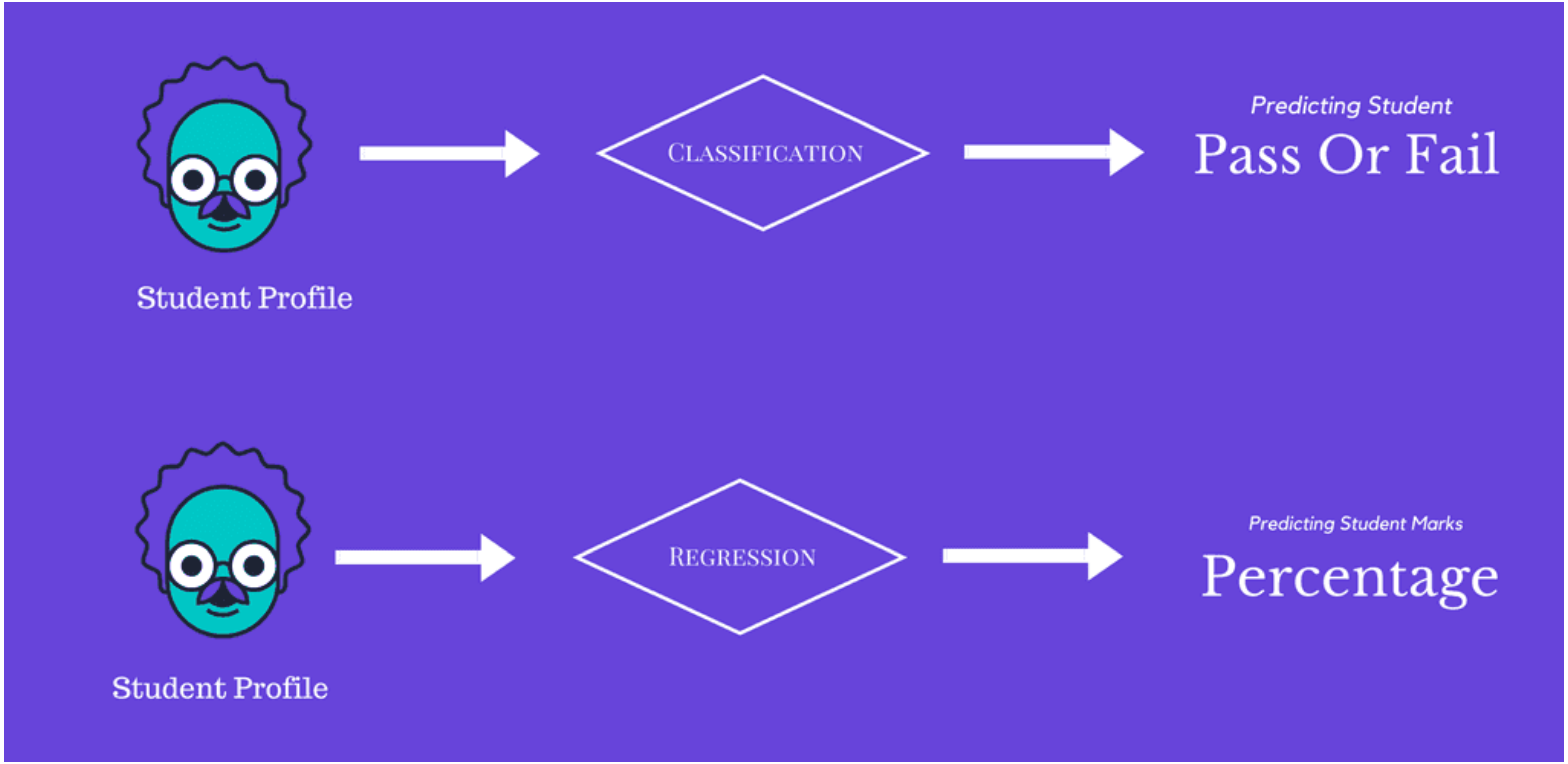
Regression

Age	Income	Loan Amount
21	20000	0
37	55000	150000
29	35000	120000
23	17000	550000
34	70000	250000
47	84000	0
25	30000	90000

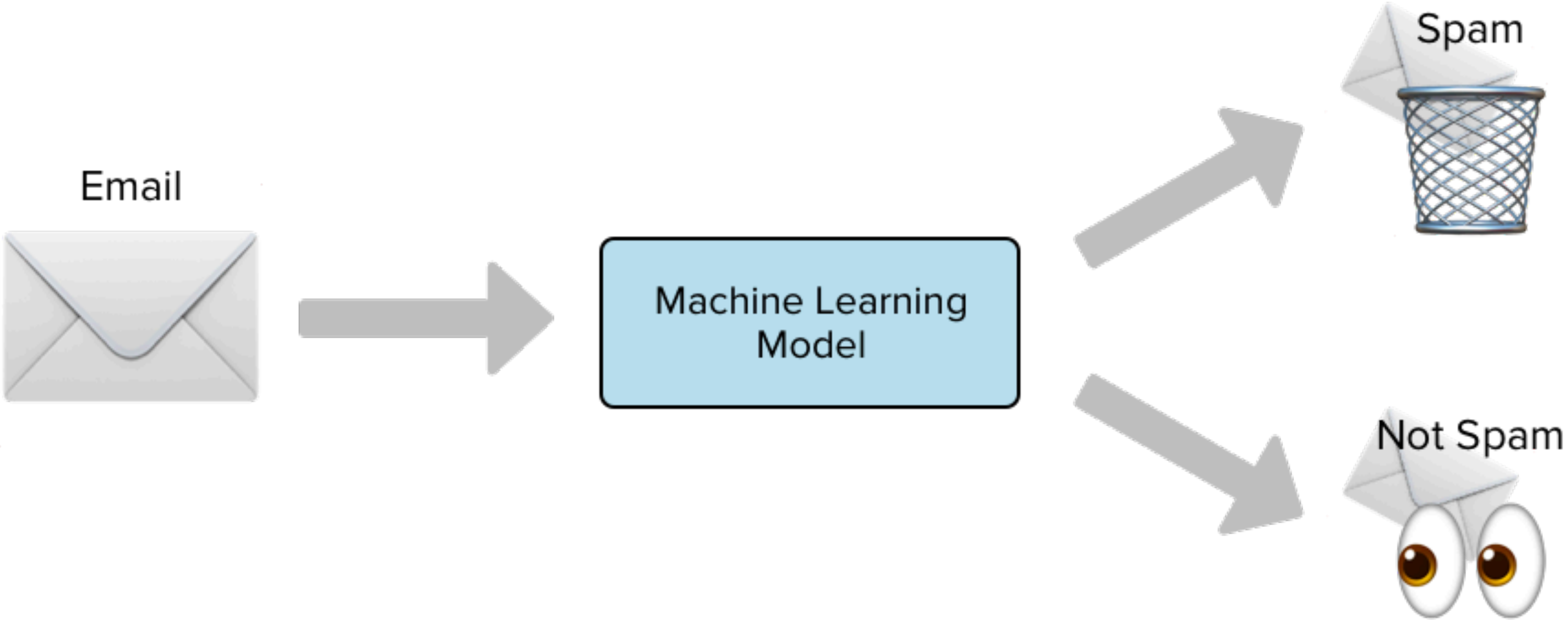
Classification

Age	Income	Loan Status
21	20000	Rejected
37	55000	Approved
29	35000	Approved
23	17000	Rejected
34	70000	Approved
47	84000	Rejected
25	30000	Approved

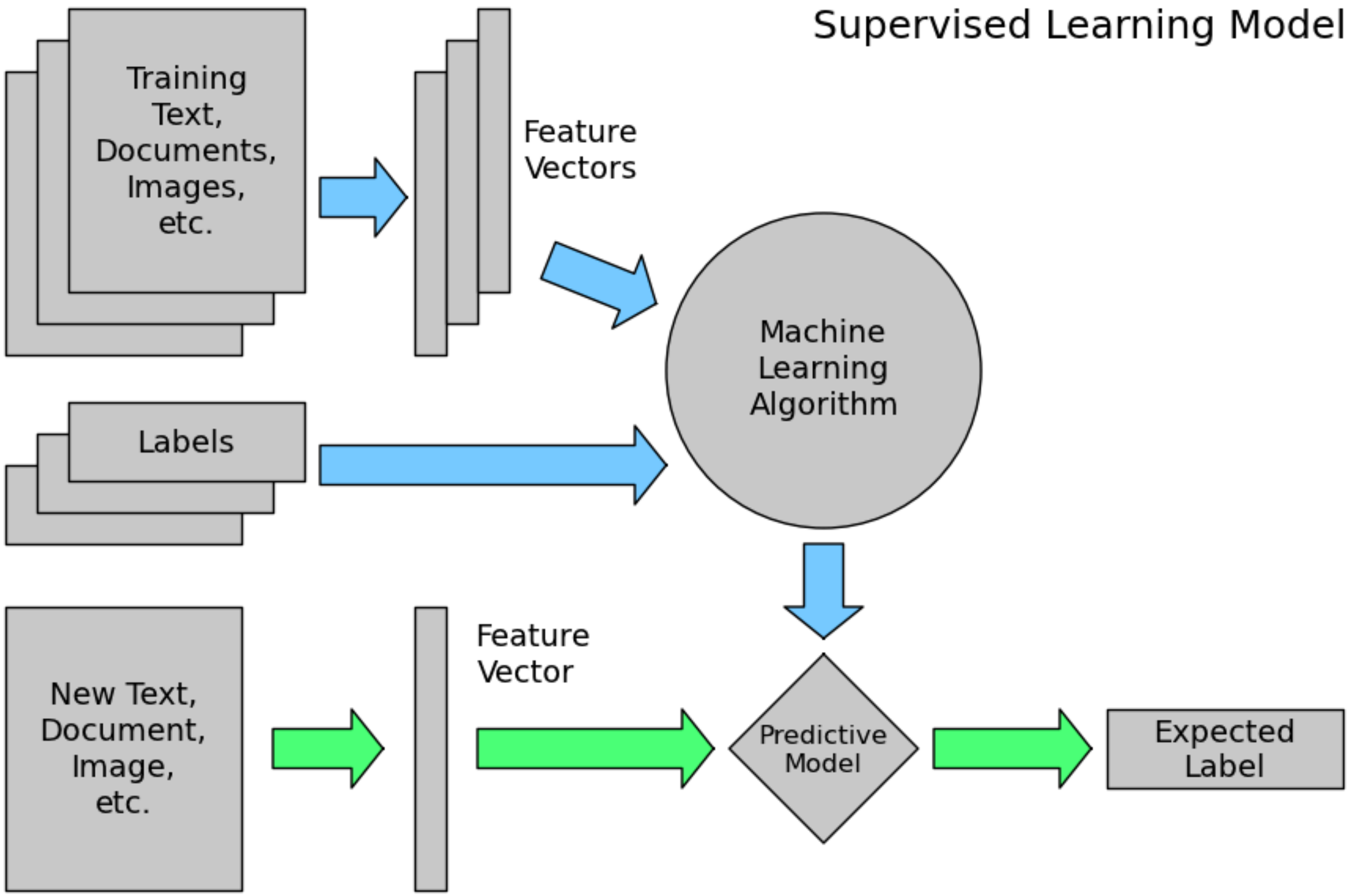
CLASSIFICATION vs REGRESSION



CLASSIFICATION



CLASSIFICATION

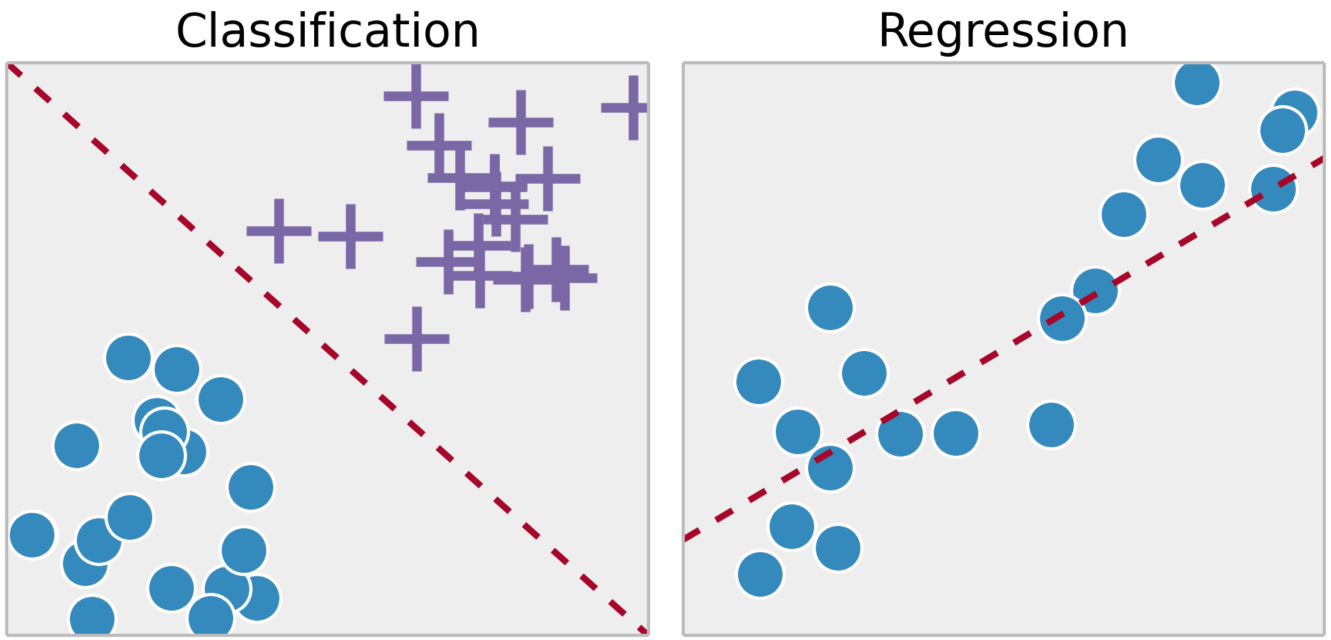


INTRODUCTION

LOGISTIC REGRESSION

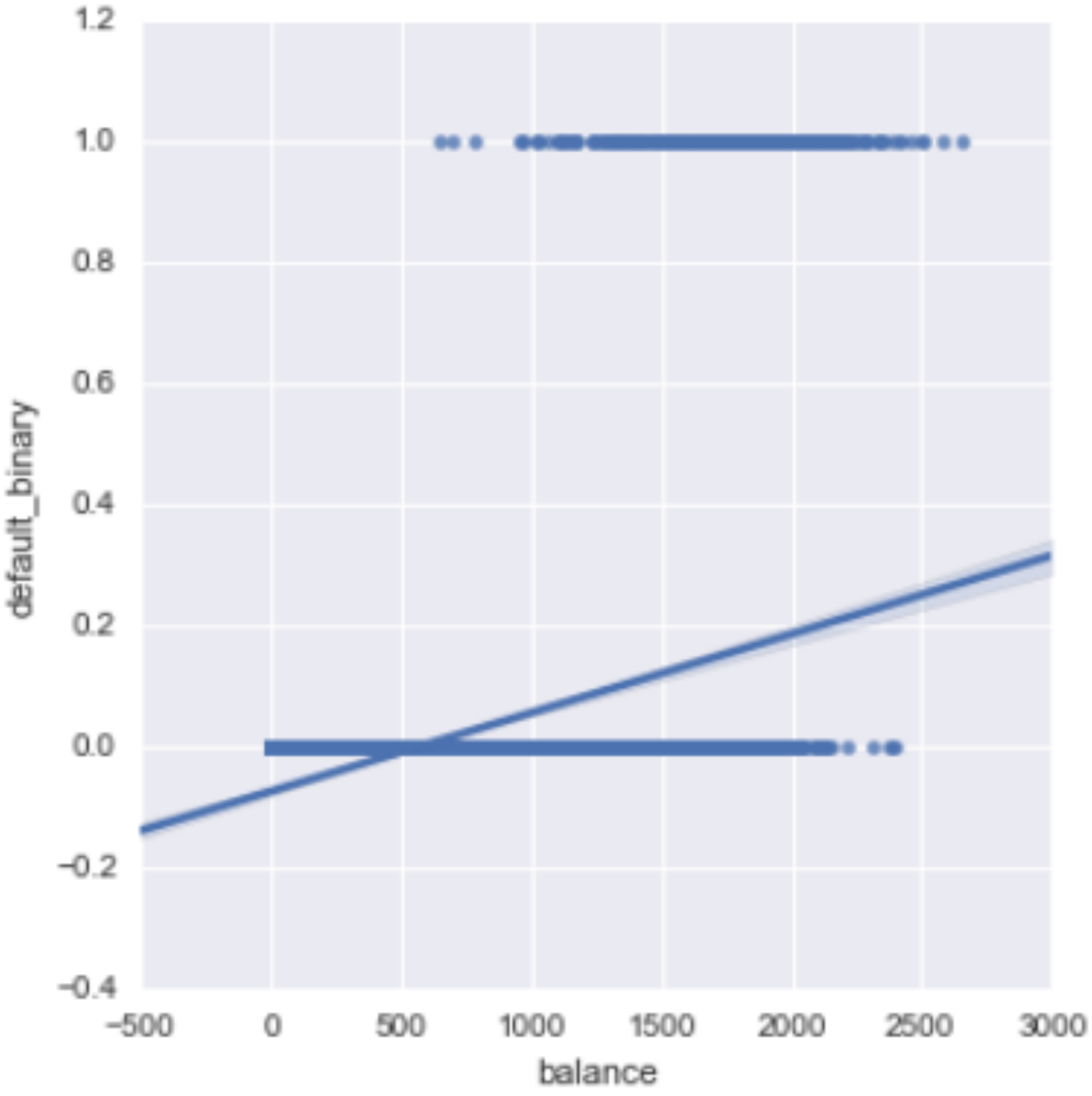
LOGISTIC REGRESSION

- Logistic regression is a linear approach to solving a classification problem. It will use a linear regression *style* approach to predict the class of an item, but retain the interpretability of linear regression model.

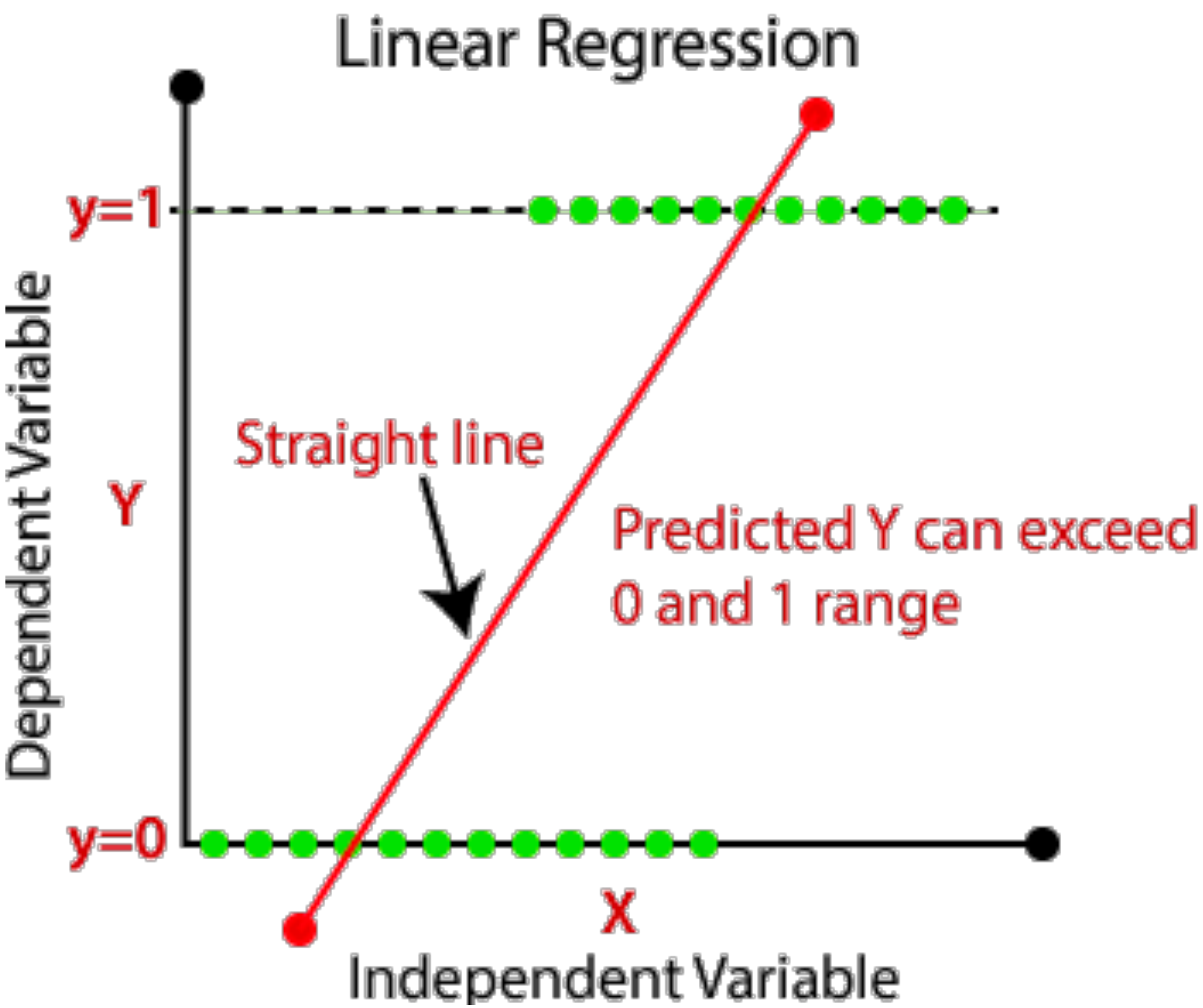


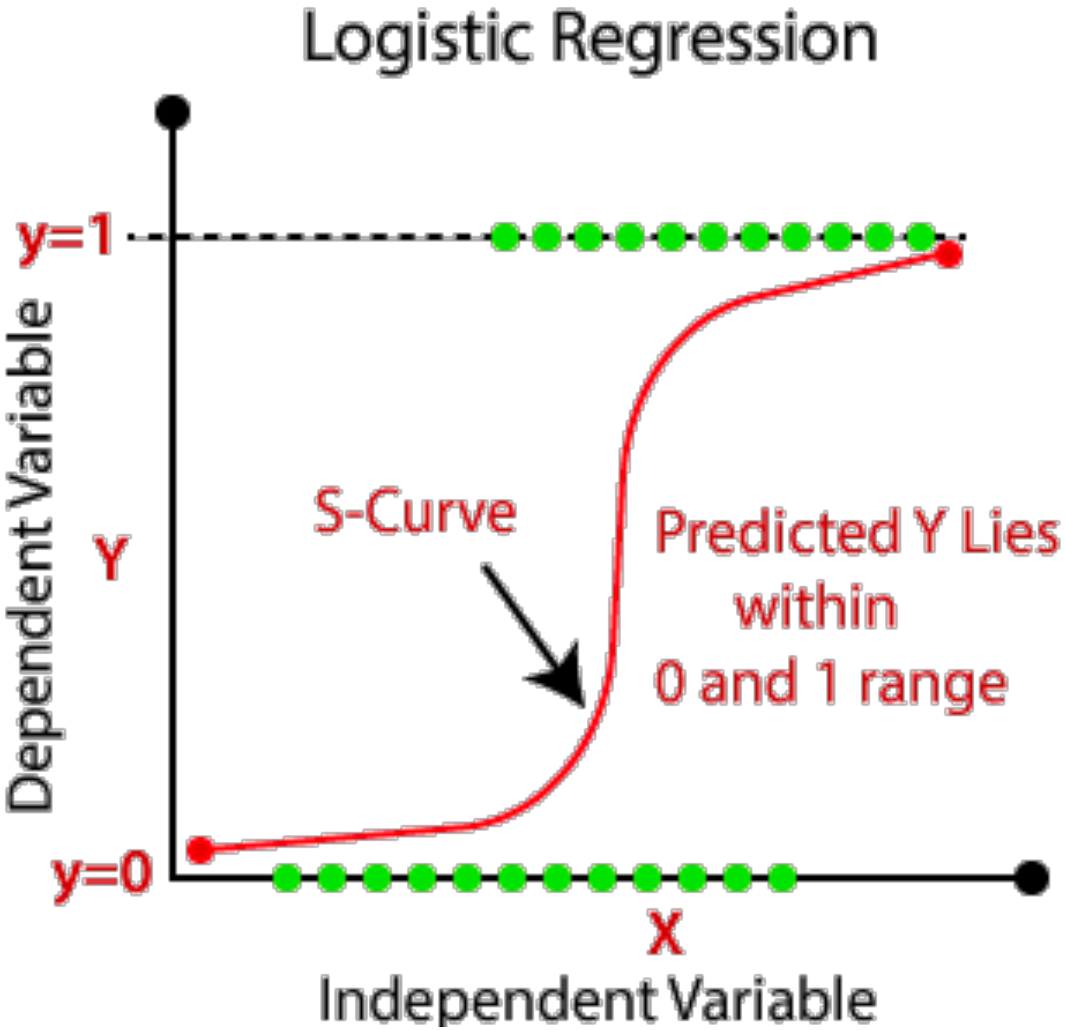
LINEAR REGRESSION can’t model a binary outcome

- We need a way to *transform* our regression model so that its range changes from [-∞, ∞] to [0, 1].



LINEAR REGRESSION can't model a binary outcome

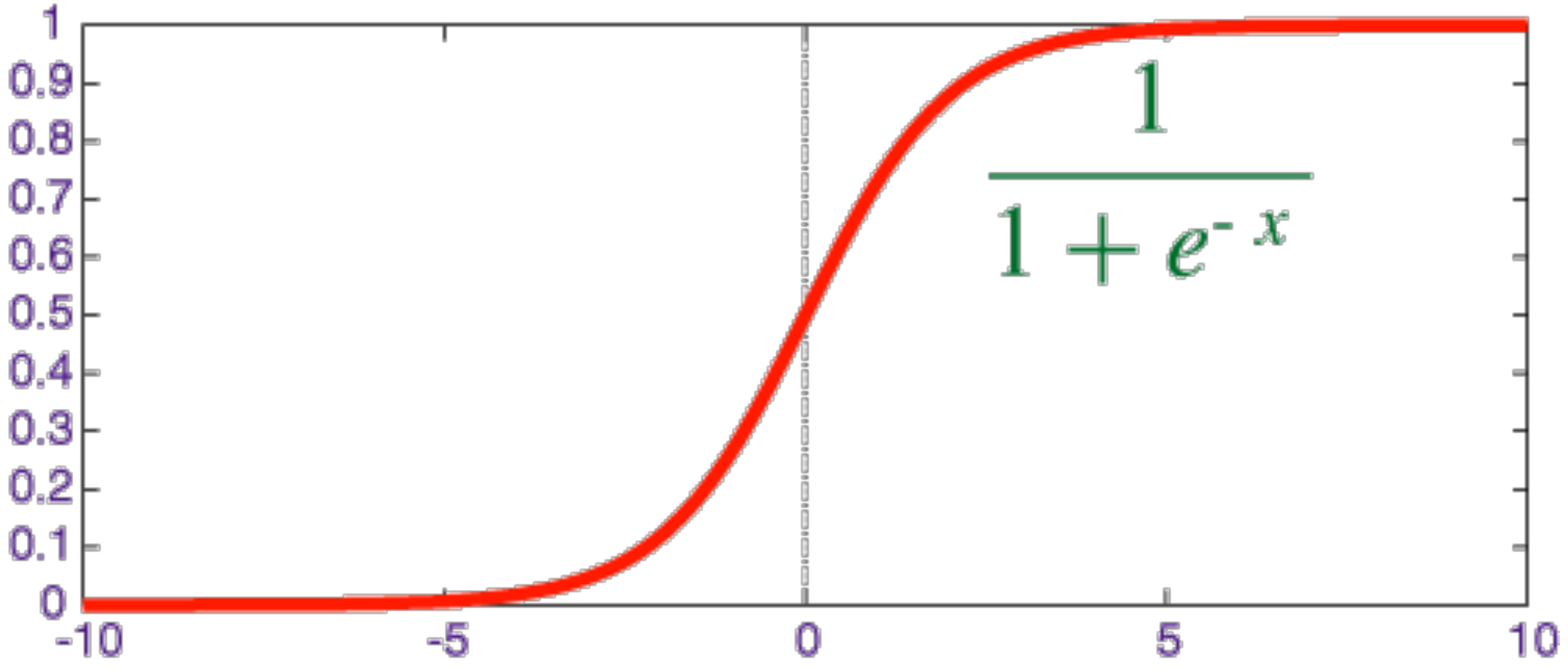




LOGISTIC REGRESSION

- To do this, we'll use a log-based transformation called the **logit function**.

- It will limit our range to [0,1] and create the right shape for our regression line to match the categorical outcome variable.



EQUATIONS

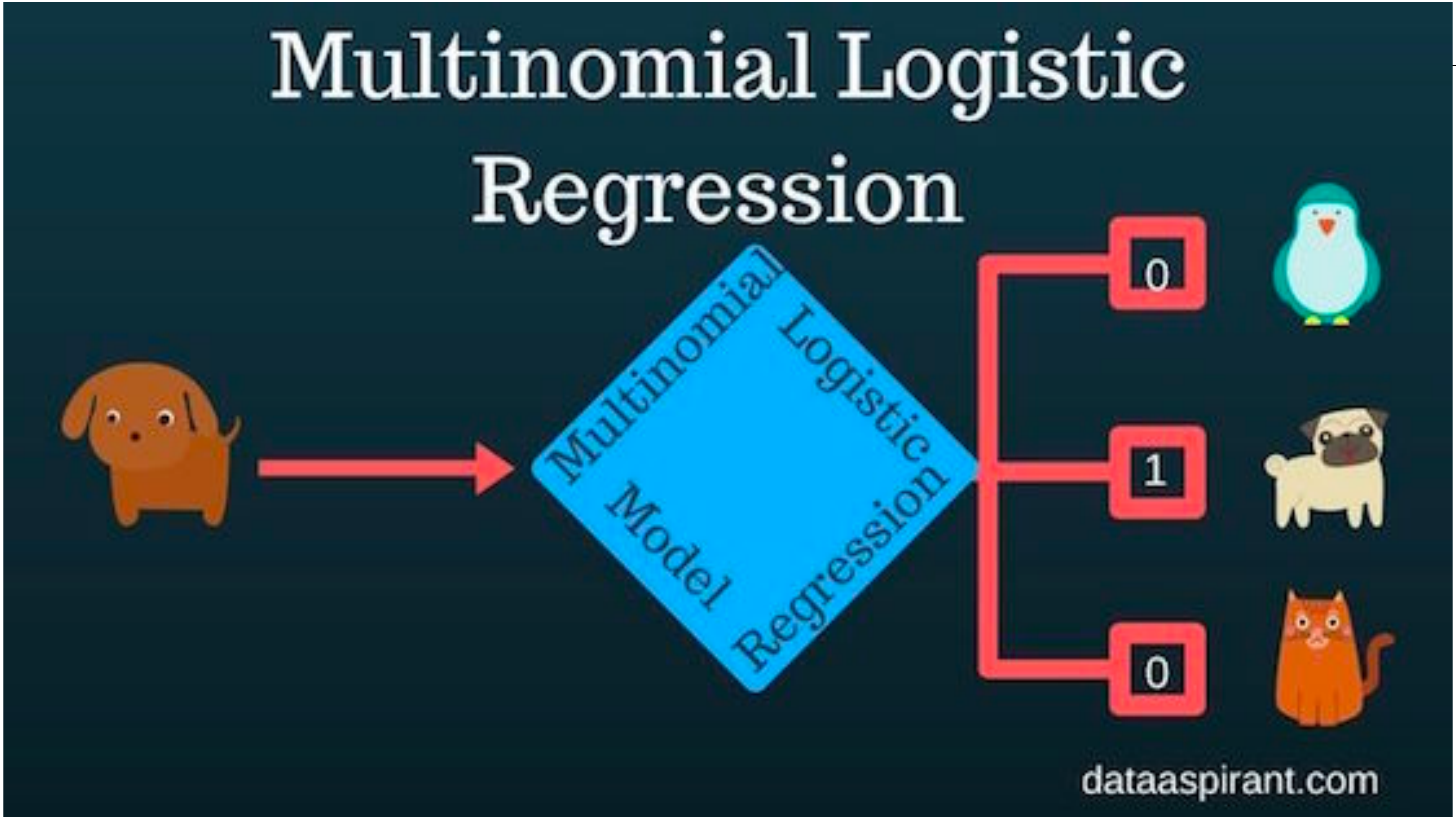
- Linear regression equation:

$$y = \beta_1 X + \beta_0$$

- Logistic regression equation:

$$p = P\left(y \mid X\right) = \frac{1}{1 + e^{-\beta_1 X + \beta_0}}$$

$$\operatorname{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$



Multinomial Logistic Regression
Multinomial Logistic Regression Model
0
1
0
dataaspirant.com

INTRODUCTION

INTERPRETING COEFFICIENTS

COEFFICIENTS

- Coefficients are expressed as logodds
- Positive coefficients increase the log odds of the response (and thus increase the probability)
- Negative coefficients decrease the log odds of the response (and thus decrease the probability).

	probability	odds	logodds
0	0.10	0.111111	-2.197225
1	0.20	0.250000	-1.386294
2	0.25	0.333333	-1.098612
3	0.50	1.000000	0.000000
4	0.60	1.500000	0.405465
5	0.80	4.000000	1.386294
6	0.90	9.000000	2.197225

LOGISTIC REGRESSION COEFFICIENTS

- The intercept is the log of the odds when all predictors are zero
- Coefficients are the log of the odds of each predictor

Binary Logit: Churn

	Estimate	Standard Error	z	p
(Intercept)	-1.41	0.16	-8.73	< .001
Senior Citizen: Yes	0.41	0.11	3.60	< .001
Tenure	-0.03	0.00	-11.38	< .001
Internet Service: DSL	0.92	0.21	4.39	< .001
Internet Service: Fiber optic	1.82	0.32	5.66	< .001
Contract: One year	-0.88	0.14	-6.25	< .001
Contract: Two year	-1.68	0.24	-7.02	< .001
Monthly Charges	0.00	0.00	1.11	.266

n = 3,522 cases used in estimation (Training sample); R-squared: 0.1898; Correct predictions: 79.05%; McFadden's rho-squared: 0.2564; AIC: 3,065.1; multiple comparisons correction: None

COEFFICIENTS EXAMPLE: TITANIC SURVIVAL

- These coefficients are "log odds"

- log odds = 0 indicates even probability
- log odds < 0 indicates less likely to occur
- log odds > 0 indicates more likely to occur

	Log Odds	odds_ratios
Pclass	-0.796128	0.451072
Sex_male	-0.637771	0.528469
Sex_female	0.637771	1.892258
Age	-0.441080	0.643341
SibSp	-0.324210	0.723098
Parch	-0.109567	0.896222
Fare	0.165687	1.180204
Embarked_S	-0.094984	0.909388
Embarked_C	0.093482	1.097991
Embarked_Q	0.022137	1.022384

INTERPRET THE COEFFICIENTS

Changing the β_0 value shifts the curve horizontally, whereas changing the β_1 value changes the slope of the curve.