Understanding and Applying Logistic Regression

MODELLING RELATIONSHIPS BETWEEN VARIABLES USING REGRESSION

Overview

Given causes, predict probability of effects - that’s logistic regression

Linear regression and logistic regression are similar, yet quite different

Unlike linear regression, logistic regression can be used for categorical y-variables

Forecasting and classifying are important applications of logistic regression

Playing the Odds with Logistic Regression

"I love deadlines. I love the whooshing noise they make as they go by."

Douglas Adams

Two Approaches to Deadlines

Start 5 minutes before deadline

Good luck with that

Start 1 year before deadline

Maybe overkill

Neither approach is optimal

Starting a Year in Advance

Probability of meeting the deadline

100%

Probability of getting other important work done

0%

Starting Five Minutes in Advance

Probability of meeting the deadline

0%

Probability of getting other important work done

100%

The Goldilocks Solution

Work fast

Start very late and hope for the best

Work smart

Start as late as possible to be sure to make it

Work hard

Start very early and do little else

As usual, the middle path is best

Working Smart

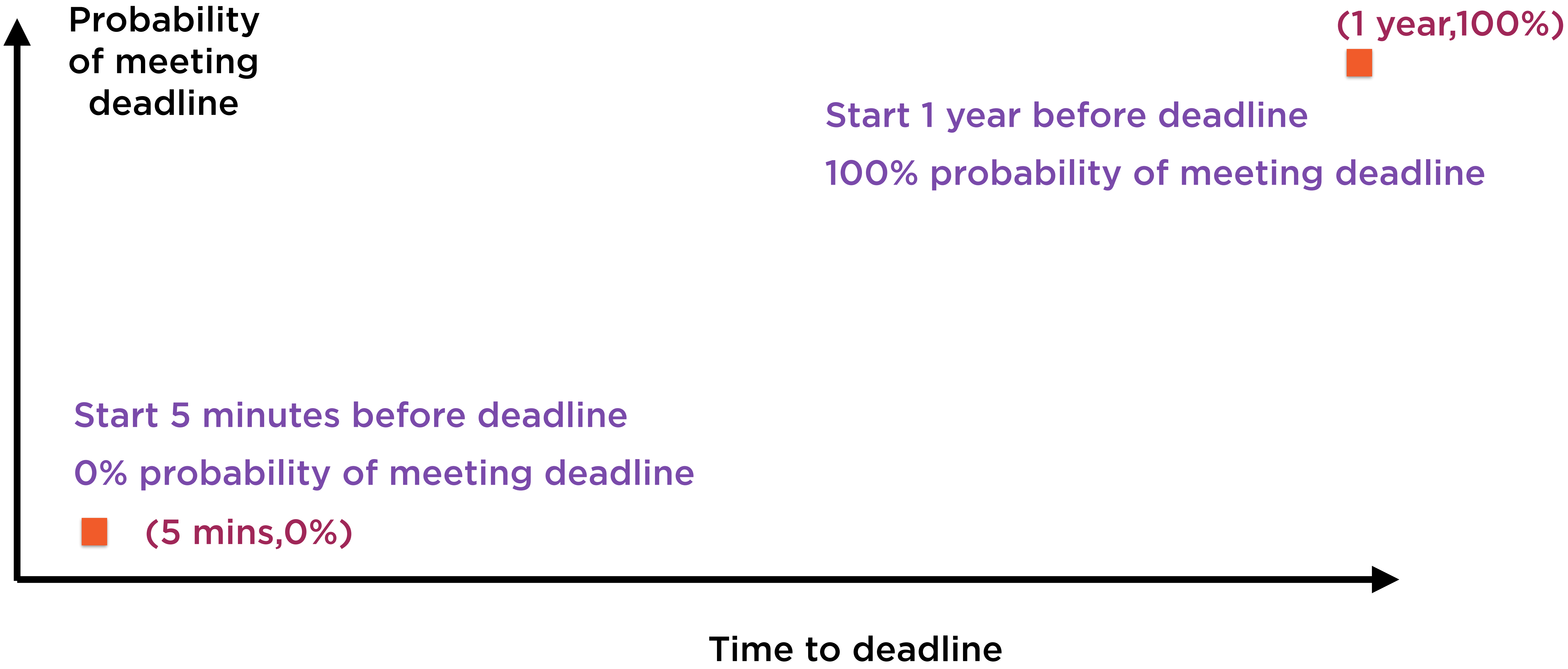
Probability of meeting the deadline

95%

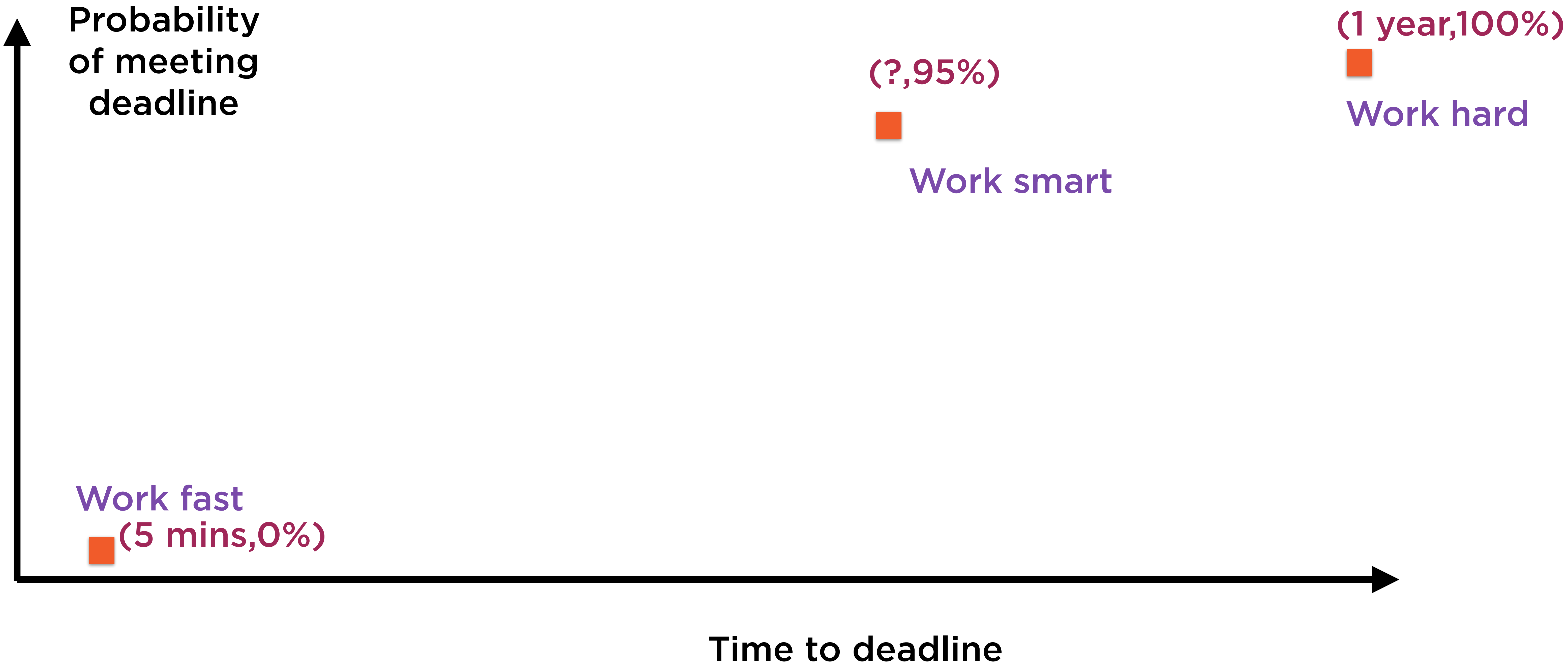
Probability of getting other important work done

95%

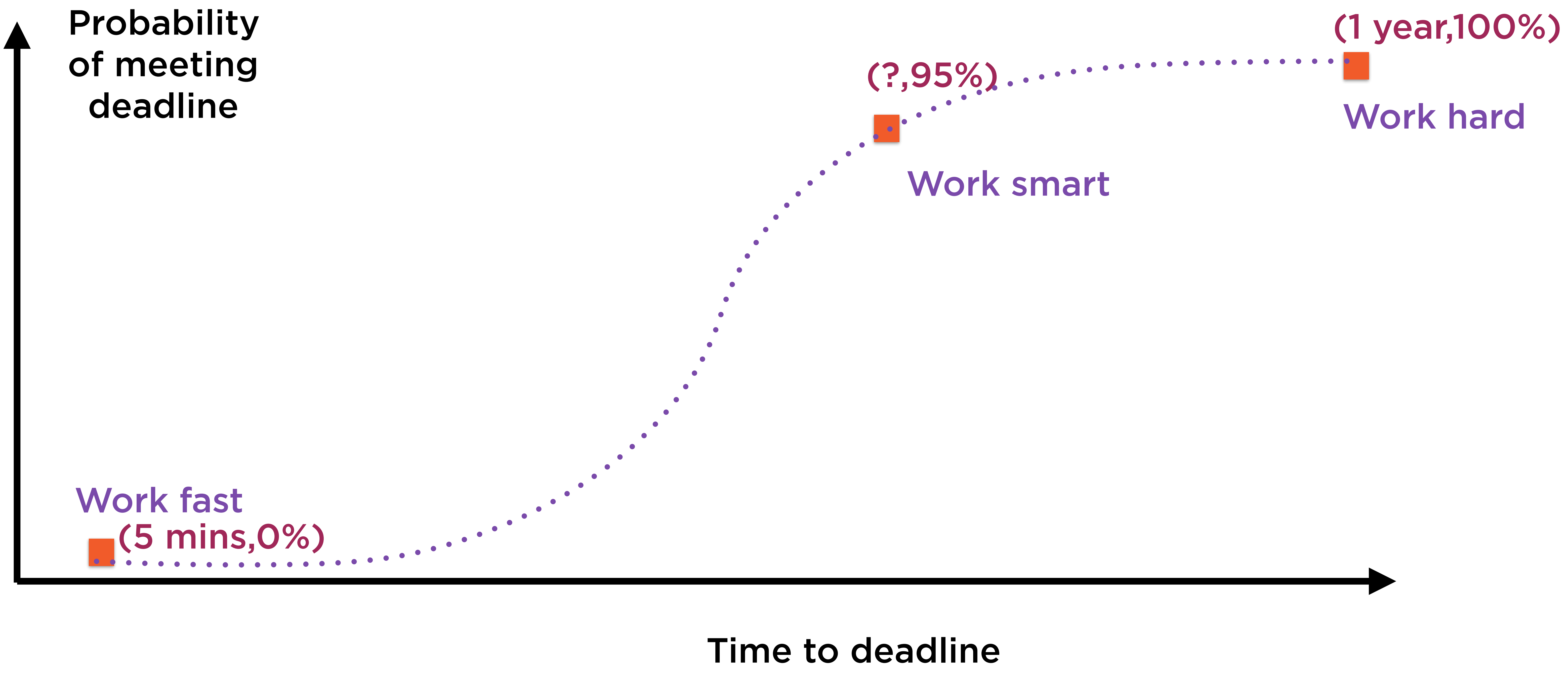
Working Hard, Fast, Smart



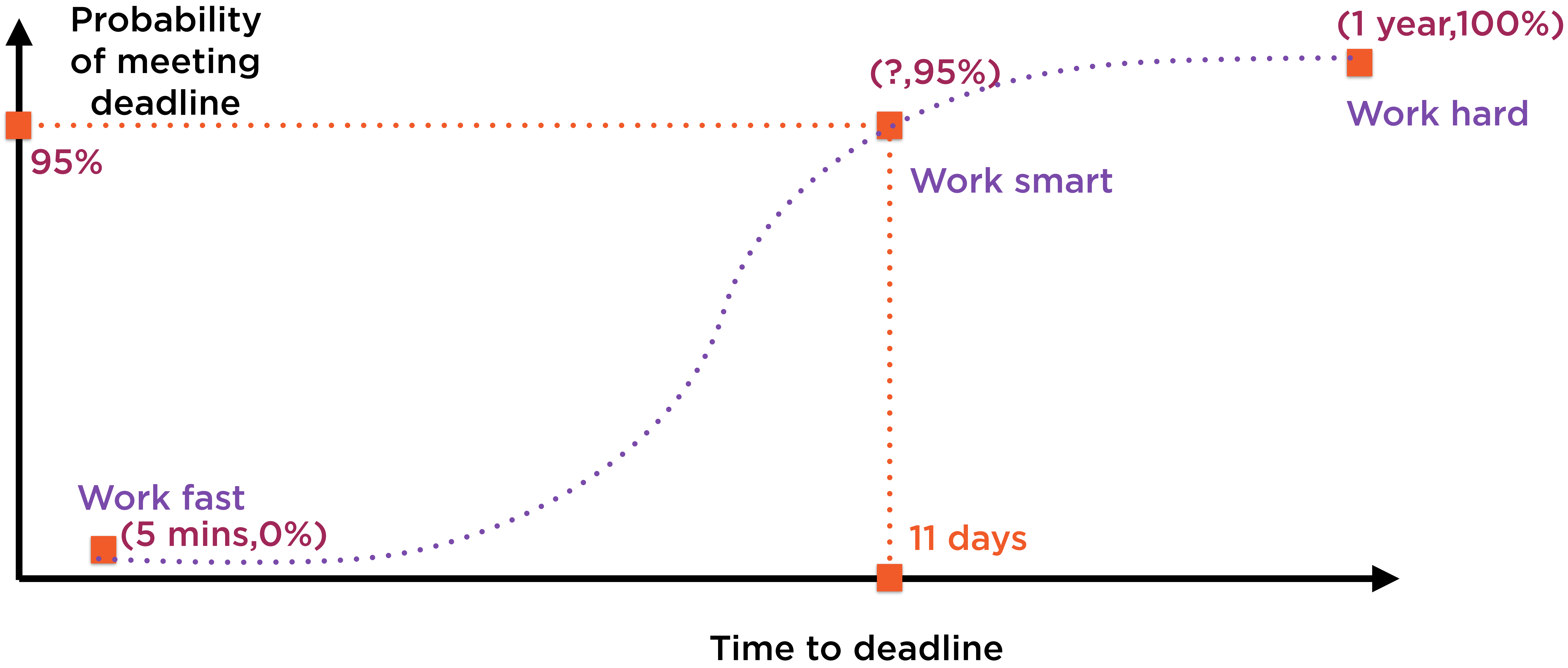
Working Hard, Fast, Smart



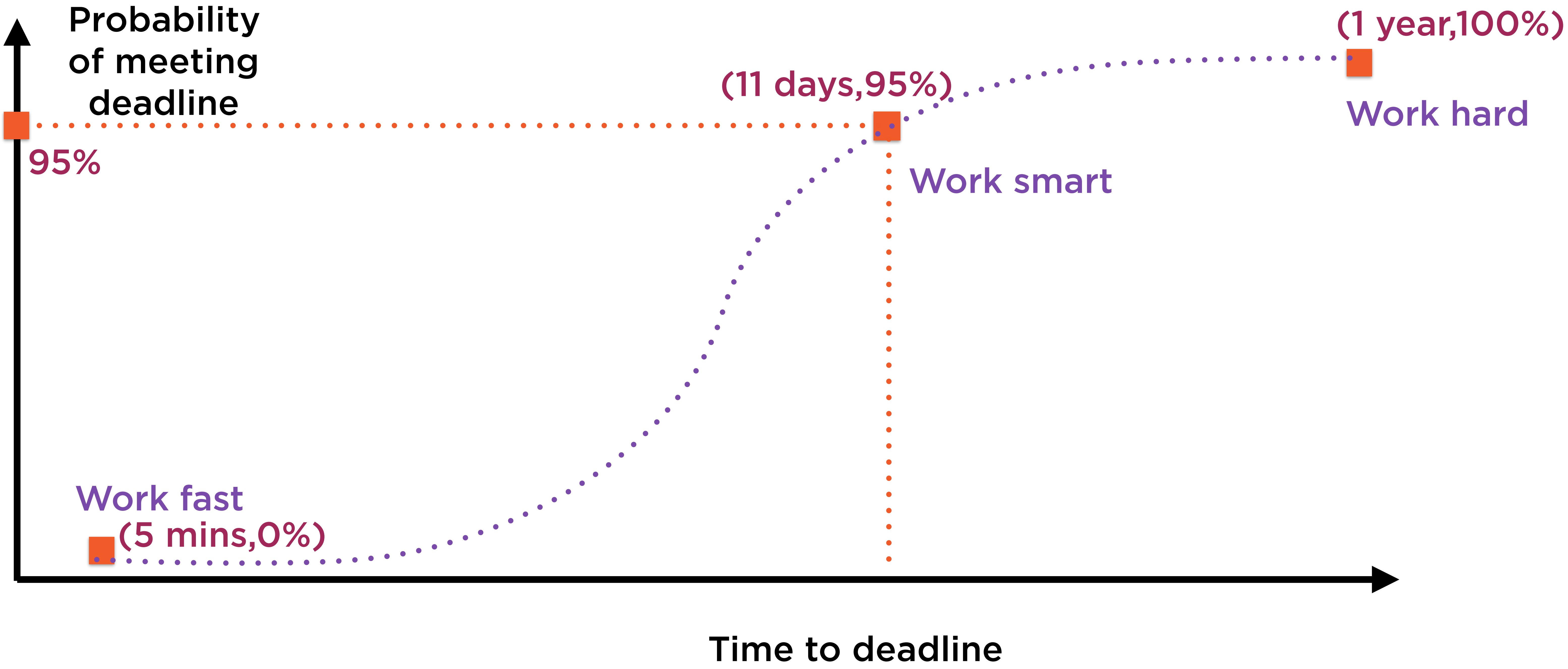
Working Hard, Fast, Smart



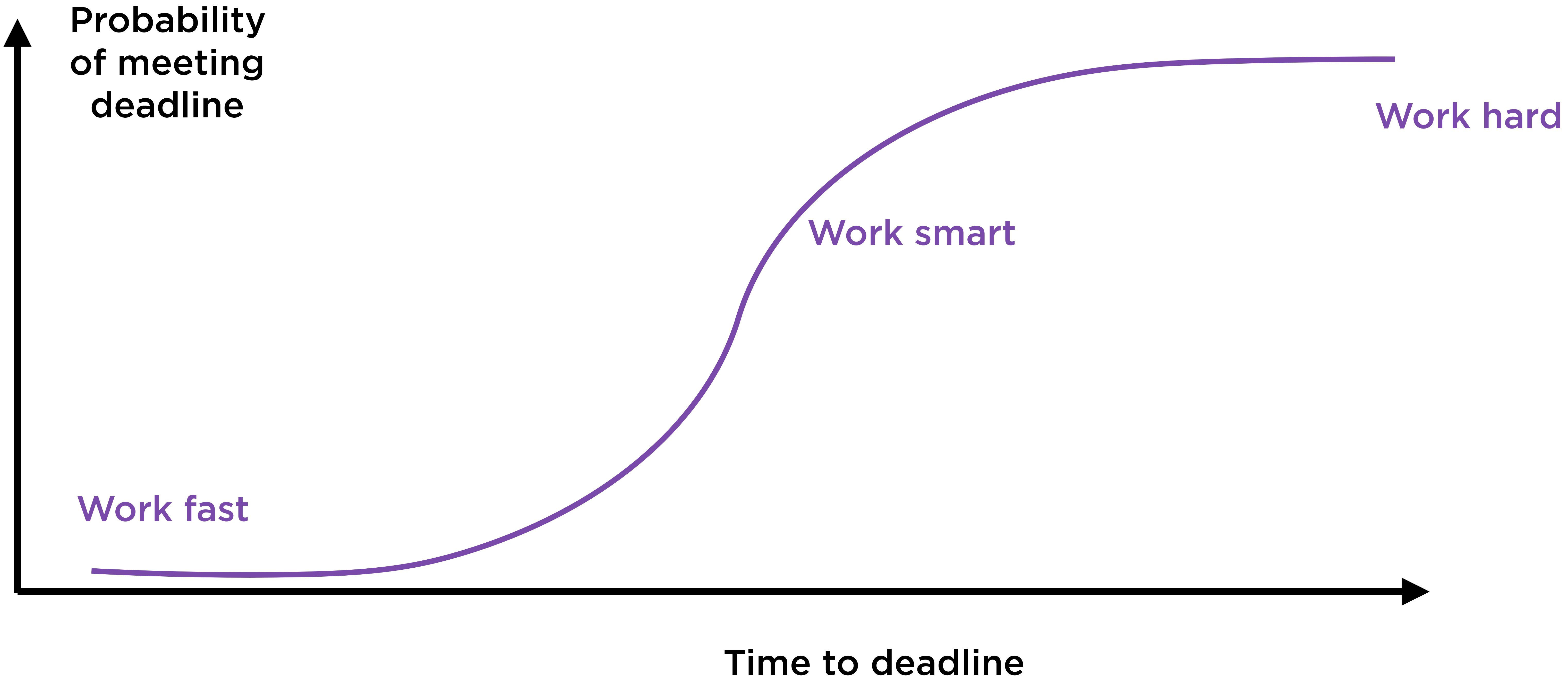
Working Hard, Fast, Smart



Working Hard, Fast, Smart

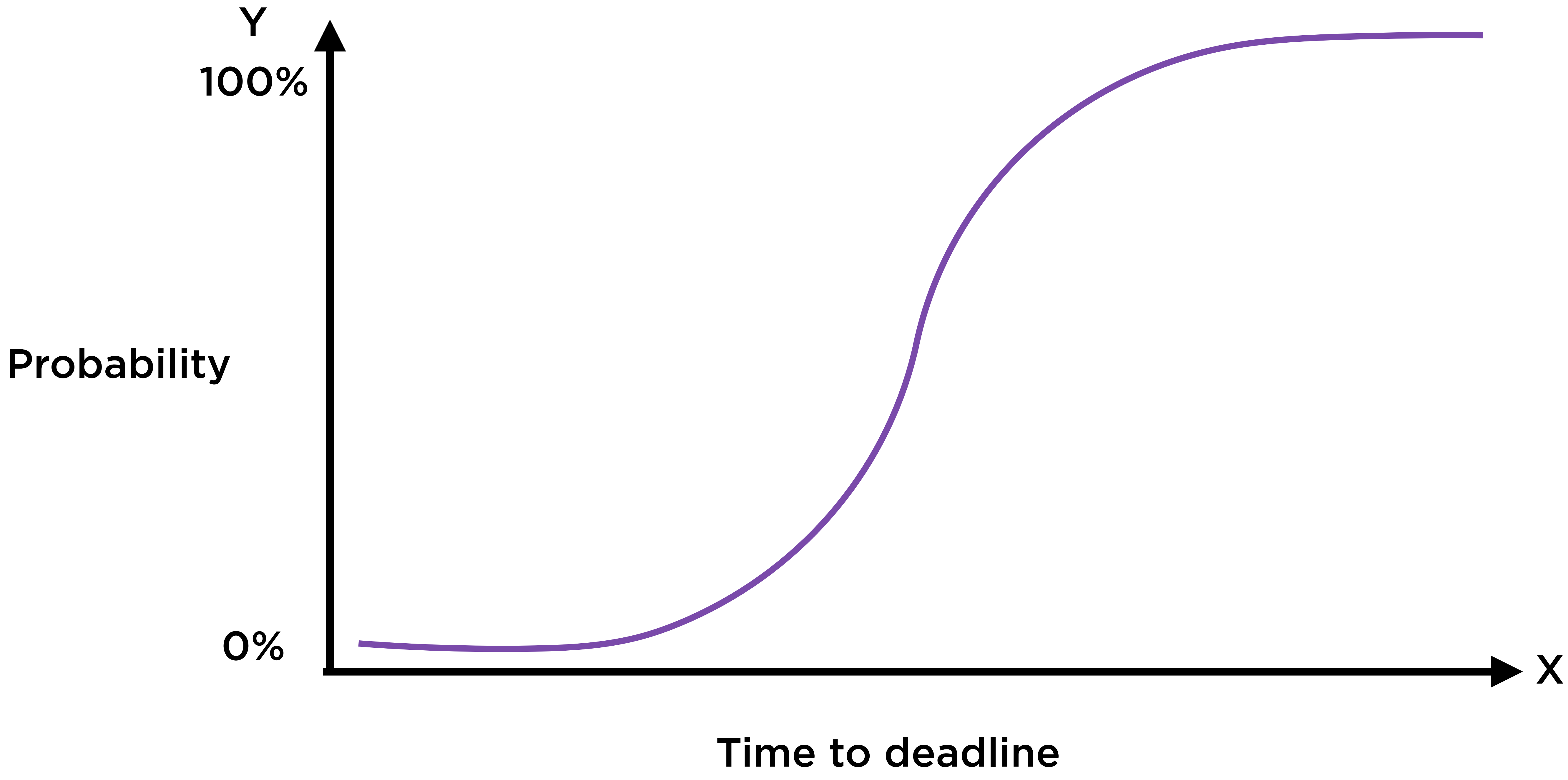


Working Hard, Fast, Smart

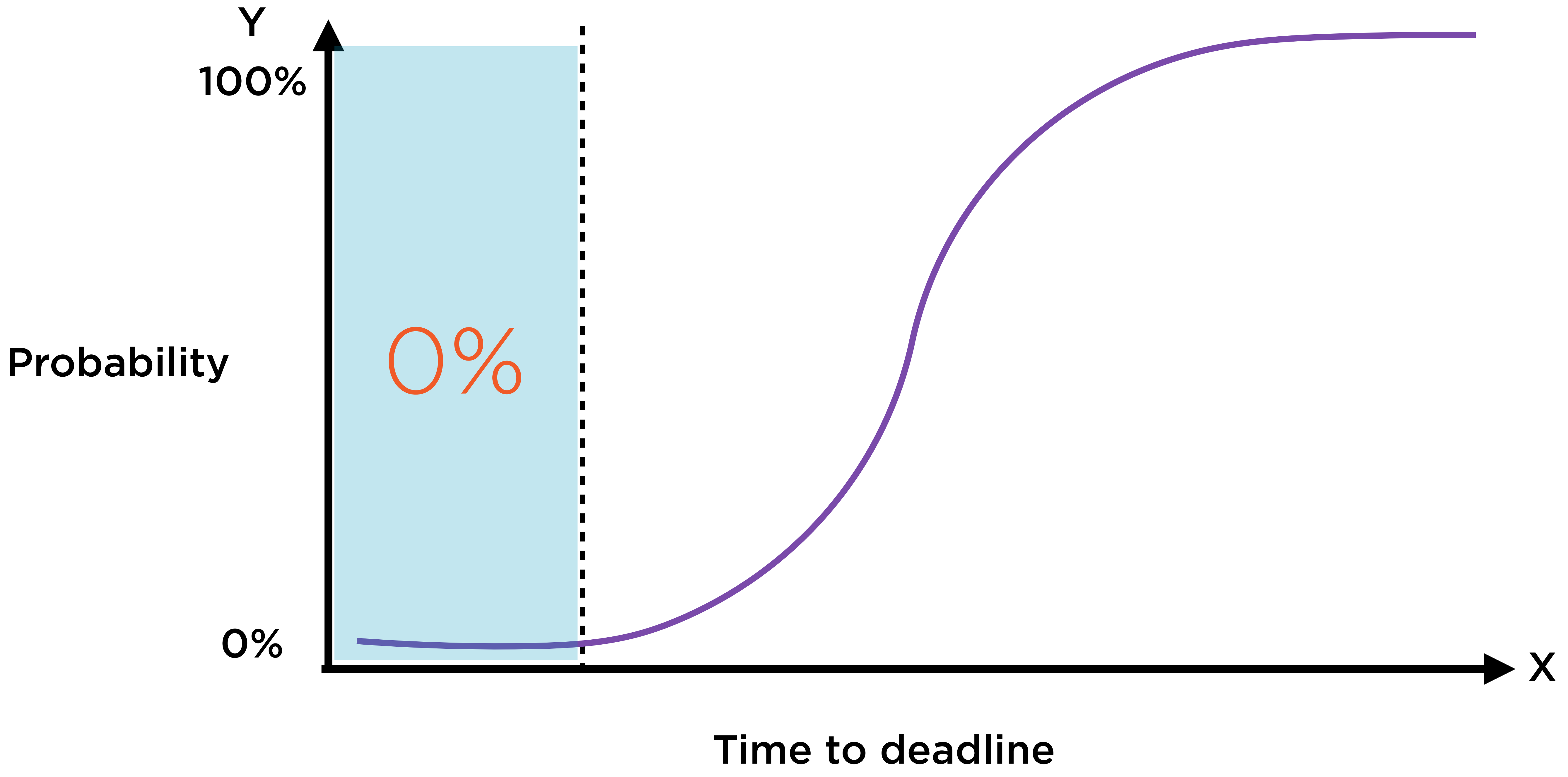


Logistic Regression helps find how probabilities are changed by actions

Working Smart with Logistic Regression

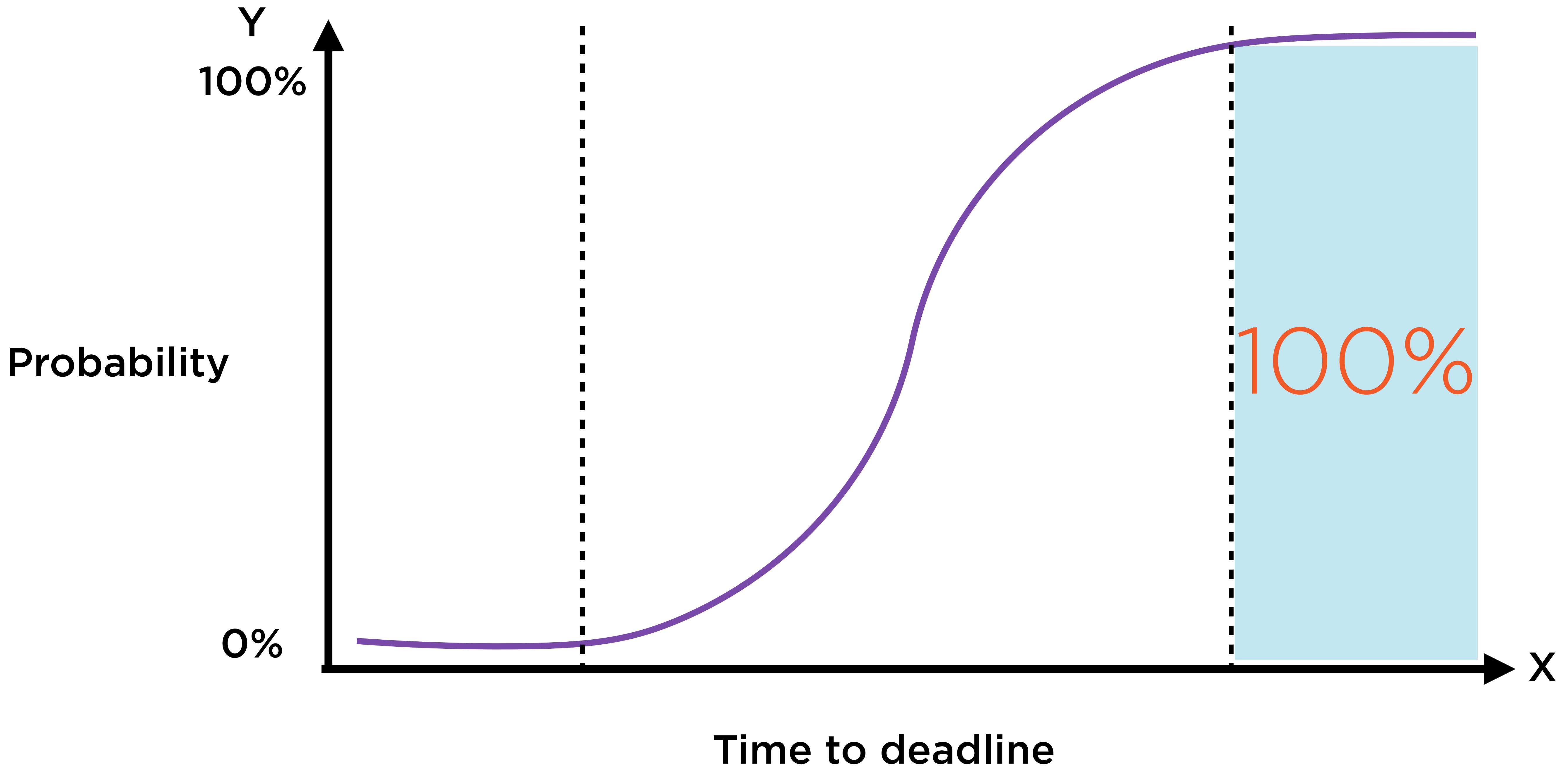


Working Smart with Logistic Regression



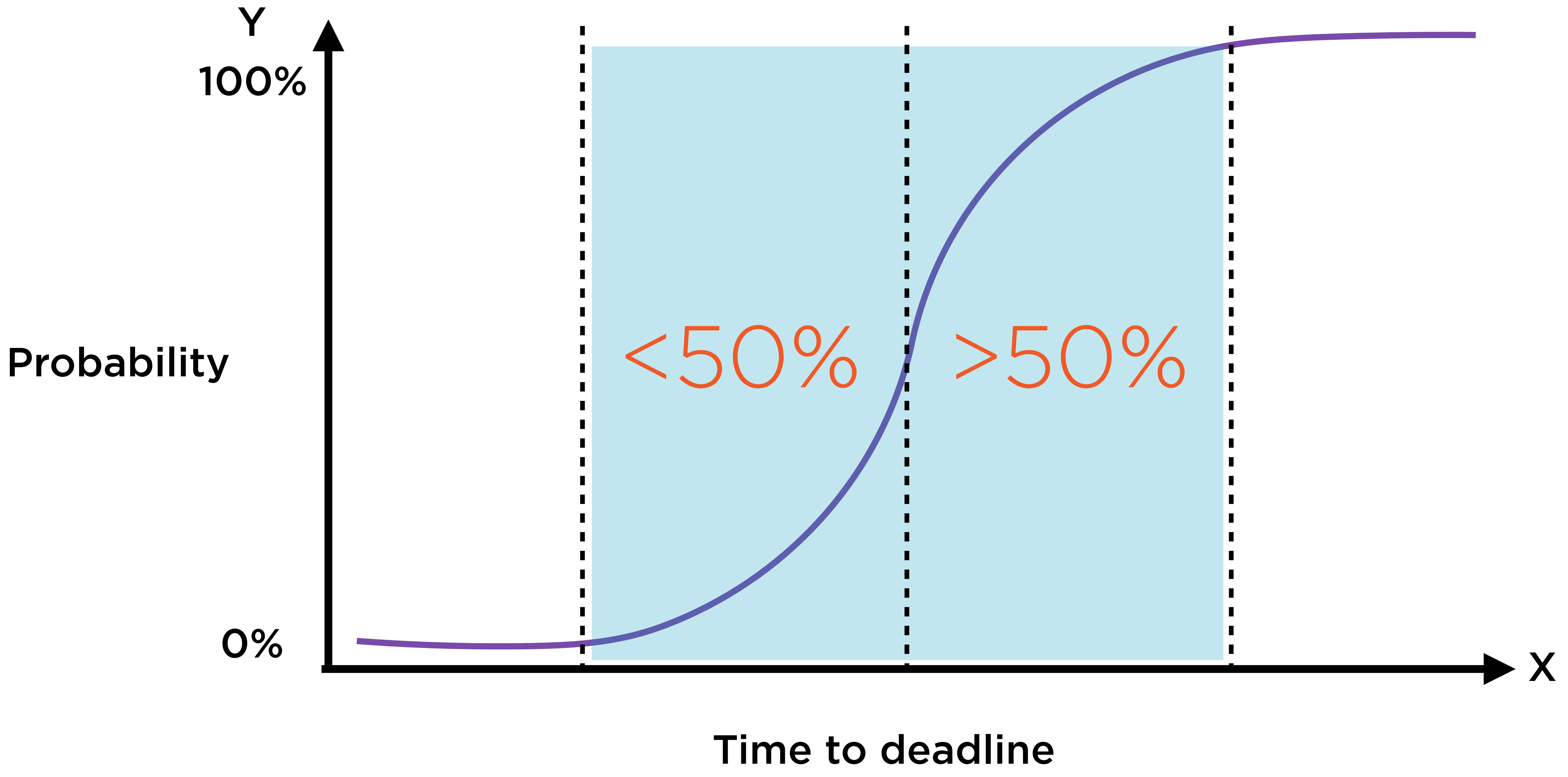
Start too late, and you'll definitely miss

Working Smart with Logistic Regression

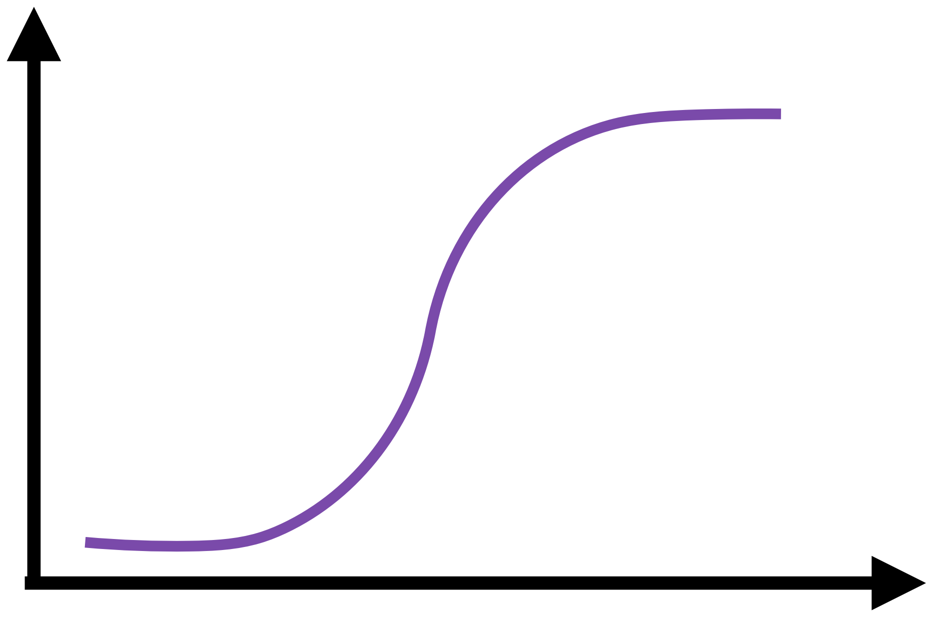


Start too early, and you'll definitely make it

Working Smart with Logistic Regression



Working smart is knowing when to start



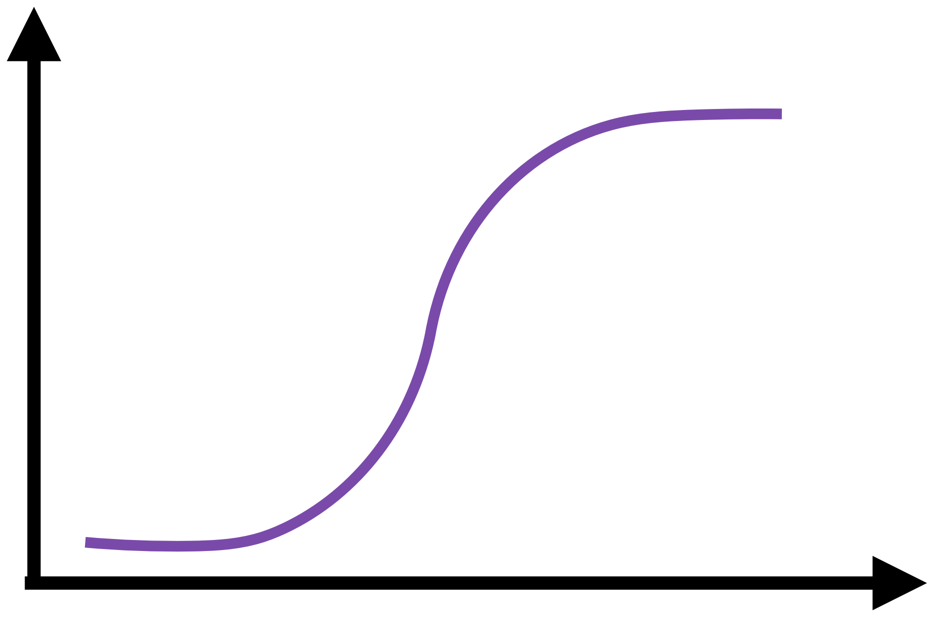
Y-axis: probability of meeting deadline

X-axis: time to deadline

Meeting or missing deadline is binary

Probability curve flattens at ends

- floor of 0
- ceiling of 1



y: hit or miss? (0 or 1?)

x: start time before deadline

p(y) : probability of y = 1

Categorical and Continuous Variables

Continuous

Can take an infinite set of values (height, weight, income...)

Categorical

Can take a finite set of values (Male/ Female, Day of week...)

Categorical variables that can take just two values are called binary variables

Logistic Regression helps estimate how **probabilities** of **categorical variables** are influenced by **causes**

Working Smart with Logistic Regression

Probabilities

p(y)

Categorical Variables

y

Causes

x

Logistic Regression helps estimate how probabilities of categorical variables are influenced by causes

Hitting Deadlines

Probability of hitting deadline	Deadline: Hit or miss?	Time of starting work
$p(y)$	$y = 1$ or 0	x

Logistic Regression helps estimate how probabilities of categorical variables are influenced by causes

Surviving the Titanic

Probability of surviving shipwreck $p(y)$	Survive or die? $y = 1$ or 0	Gender, age, class of ticket x_1, x_2, x_3

Logistic Regression helps estimate how probabilities of categorical variables are influenced by causes

Predicting Stock Markets

Probability of market rising tomorrow

$p(y)$

Up or down?

$y = 1$ or 0

Economic growth, oil prices, interest rates...

x_1, x_2, x_3...

Logistic Regression helps estimate how probabilities of categorical variables are influenced by causes

Applications of Logistic Regression

Common Applications of Logistic Regression

Analyse

Allocate

Predict

Classify

Common Applications of Logistic Regression

Analyse

Allocate

Predict

Classify

Analysing Consequences

Past events	Observed causes
Actual outcomes	Probabilities

Past events

- Sinking of the Titanic
- 2008-09 subprime mortgage crisis
- Software supplier’s history of meeting deadlines

Actual outcomes

- 1,514 deaths, 710 survivors on the Titanic
- Several banks, hedge funds collapsed
- Billions of dollars of cost overruns

Observed causes

- Sex, age, passenger class
- Interest rates, economic growth, oil prices
- Budget, leadership, technical know-how

Probabilities

- Survived or perished?
- Made or lost money?
- Ship or slip?

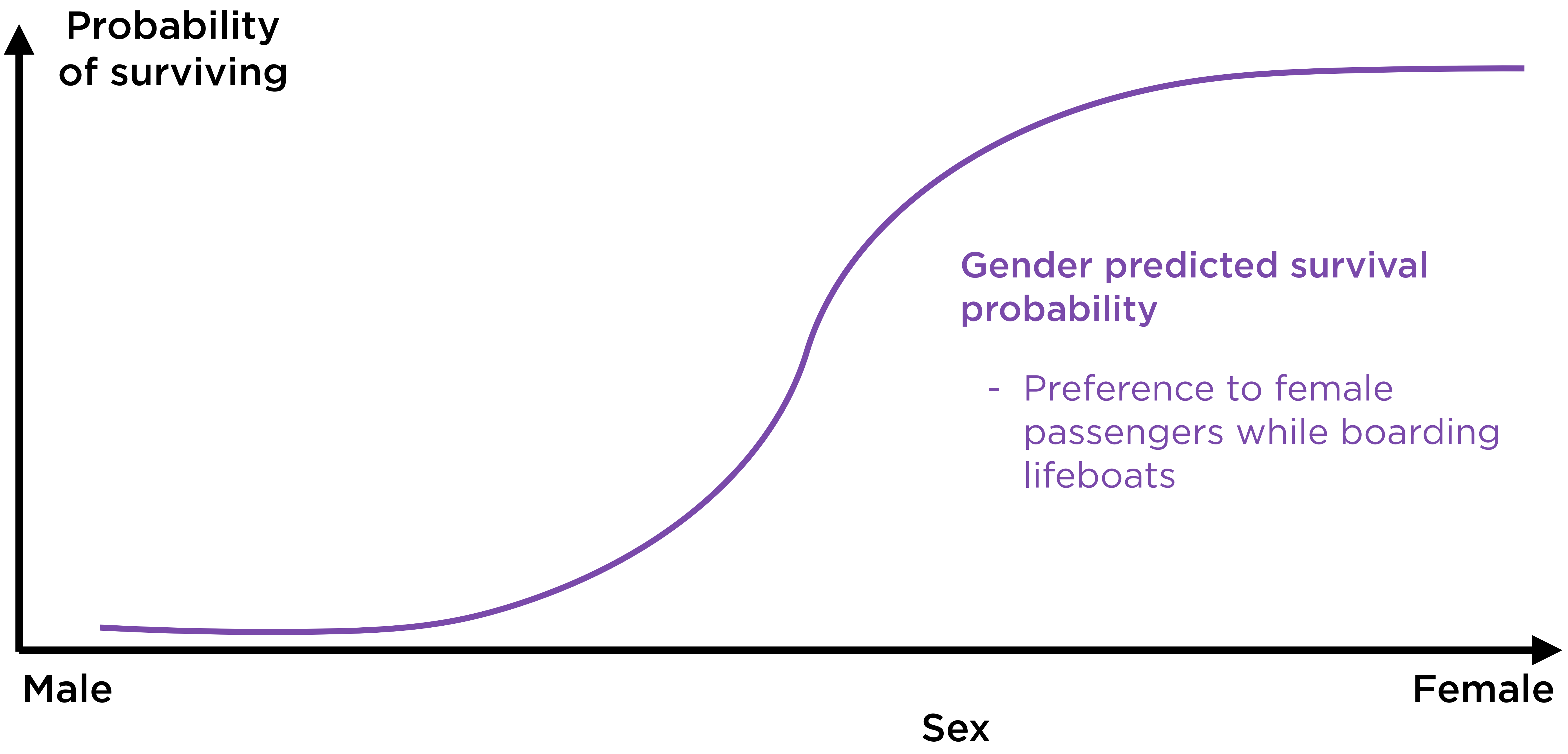
Who Would Survive the Titanic Shipwreck

Sex

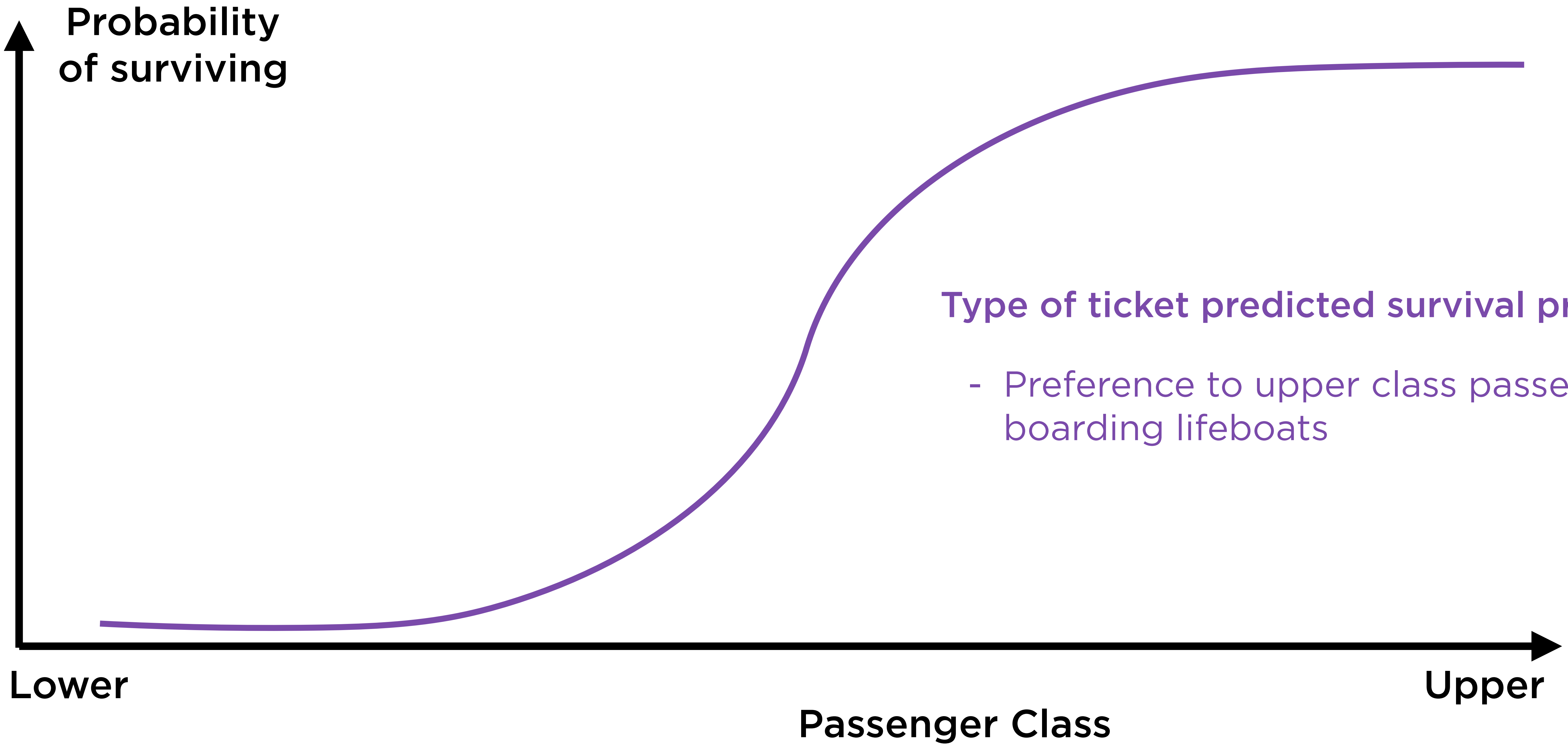
Age

Passenger class

Surviving the Titanic



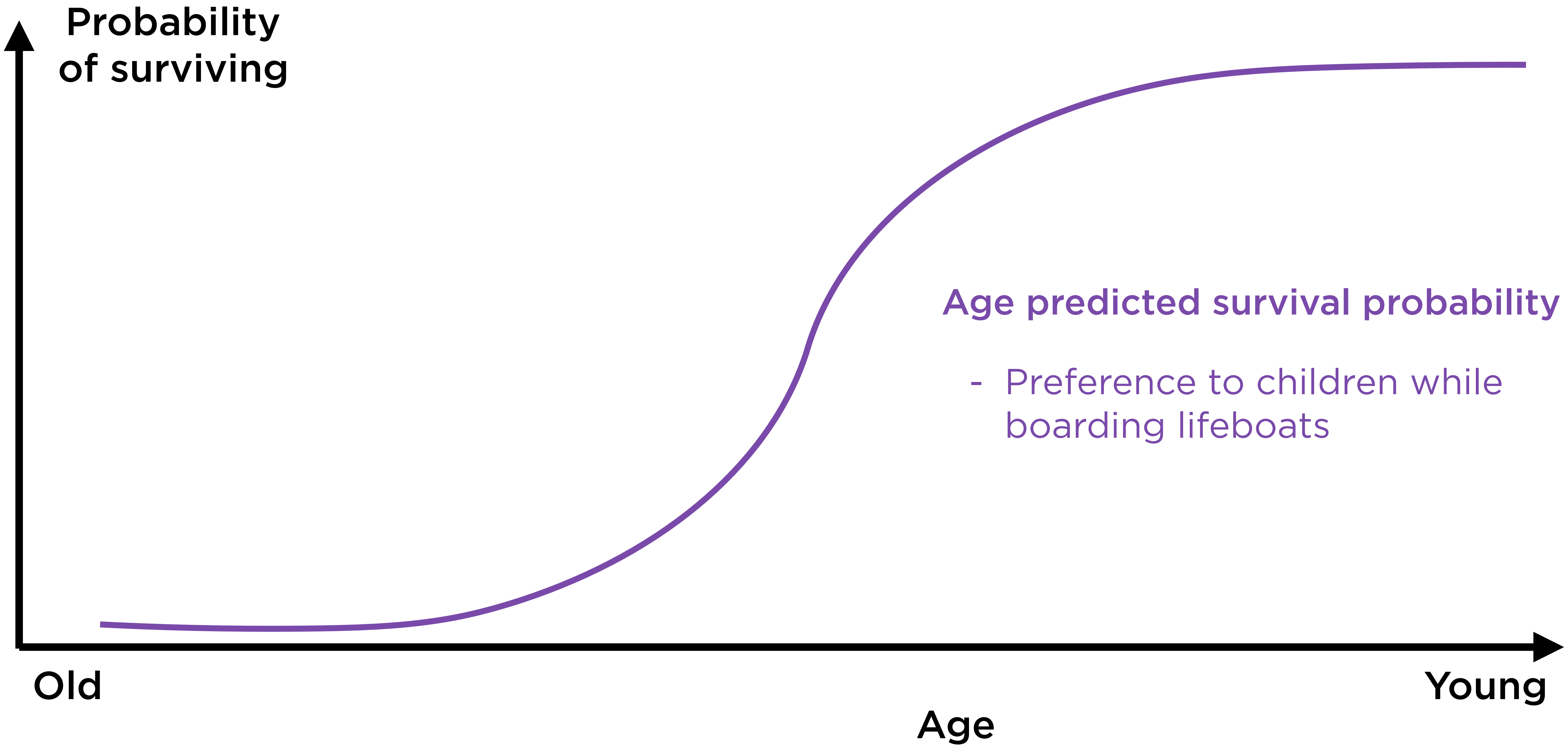
Surviving the Titanic



Type of ticket predicted survival probability

- Preference to upper class passengers while boarding lifeboats

Surviving the Titanic



Only 3% of women with first class tickets perished

92% of men with second class tickets perished

Common Applications of Logistic Regression

Analyse | **Allocate** | **Predict** | **Classify**

Allocating Resources

Economic opportunities

Catastrophic losses

Resources to avoid losses

Probabilities

The Goldilocks Solution

Work fast

Start very late and hope for the best

Work smart

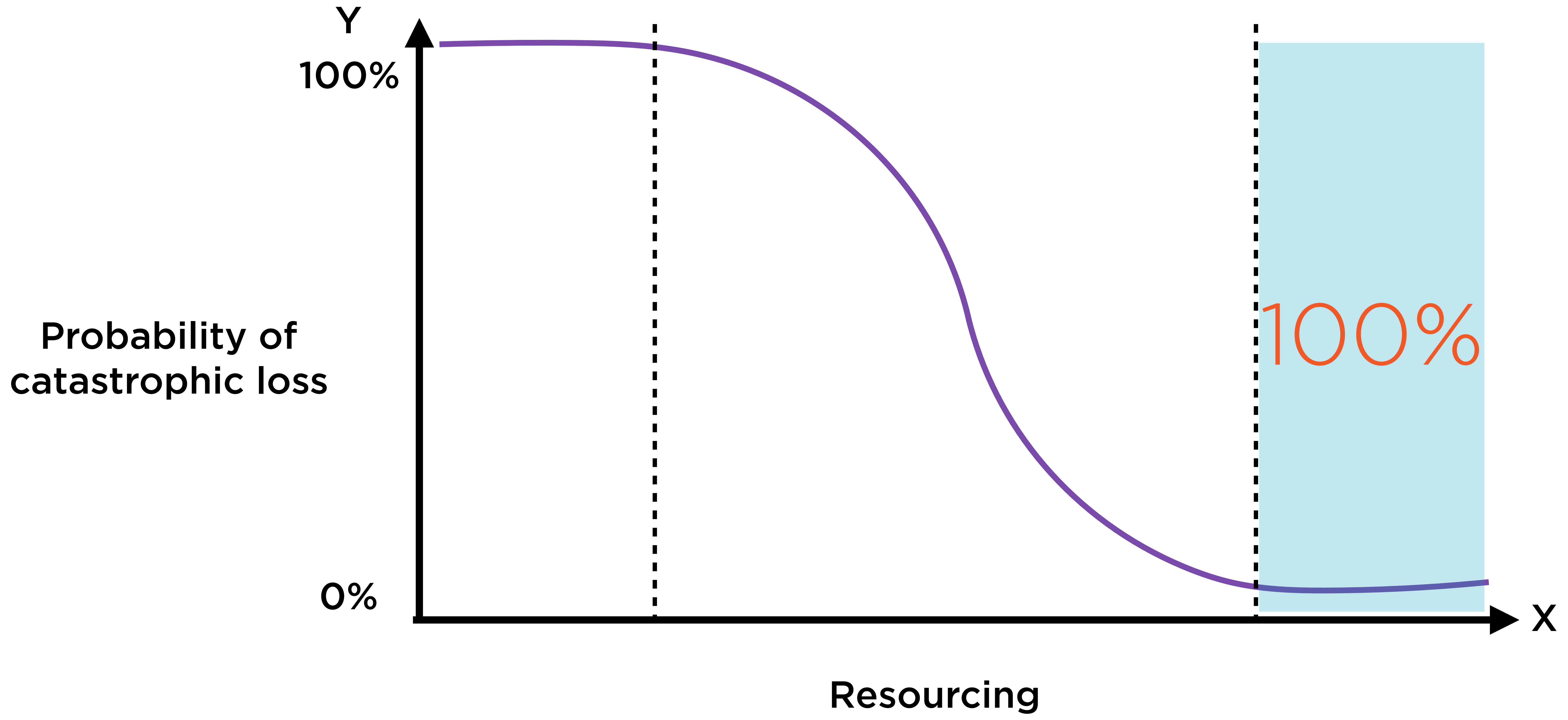
Start as late as possible to be sure to make it

Work hard

Start very early and do little else

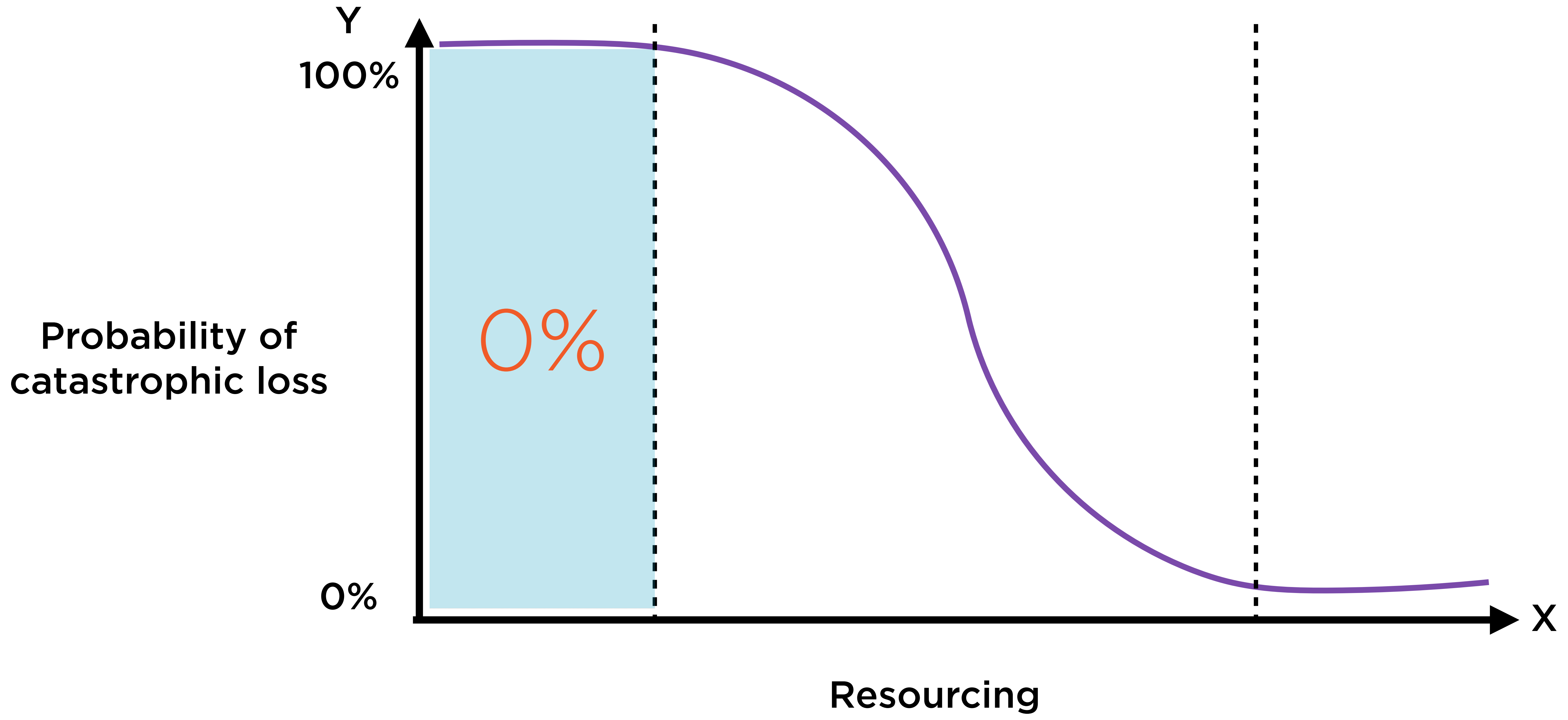
As usual, the middle path is best

Go Big or Go Home



Inadequate resource allocation

Nothing Ventured, Nothing Gained



Excessive resource allocation

Common Applications of Logistic Regression

Analyse

Allocate

Predict

Classify

Working Smart

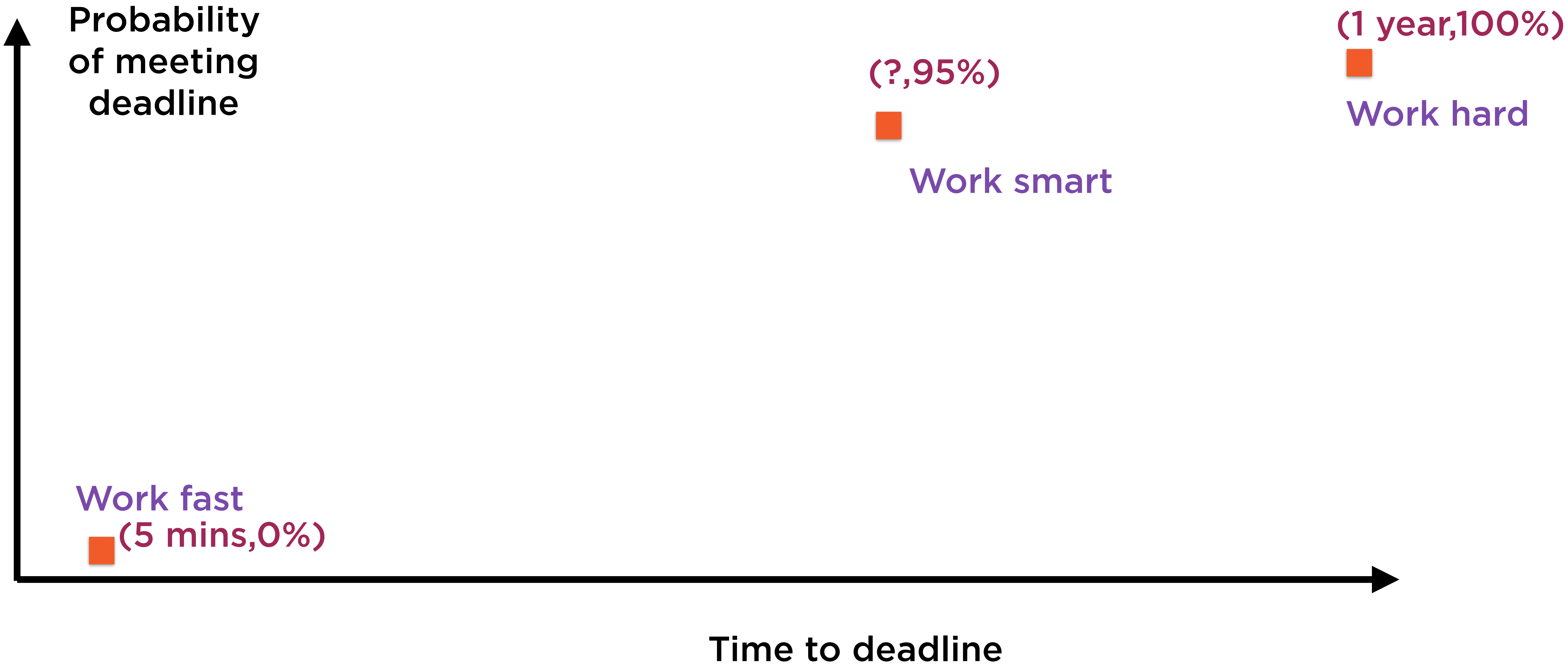
Probability of meeting the deadline

95%

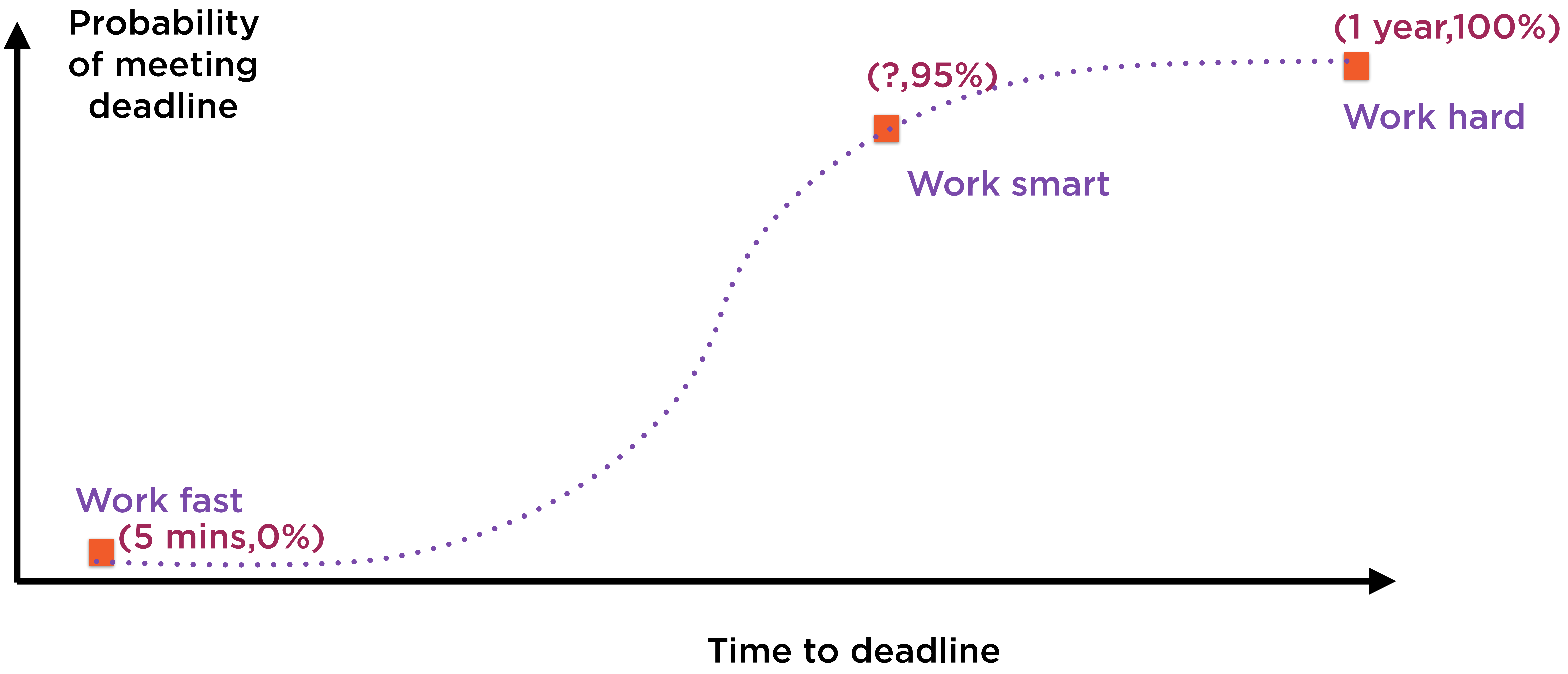
Probability of getting other important work done

95%

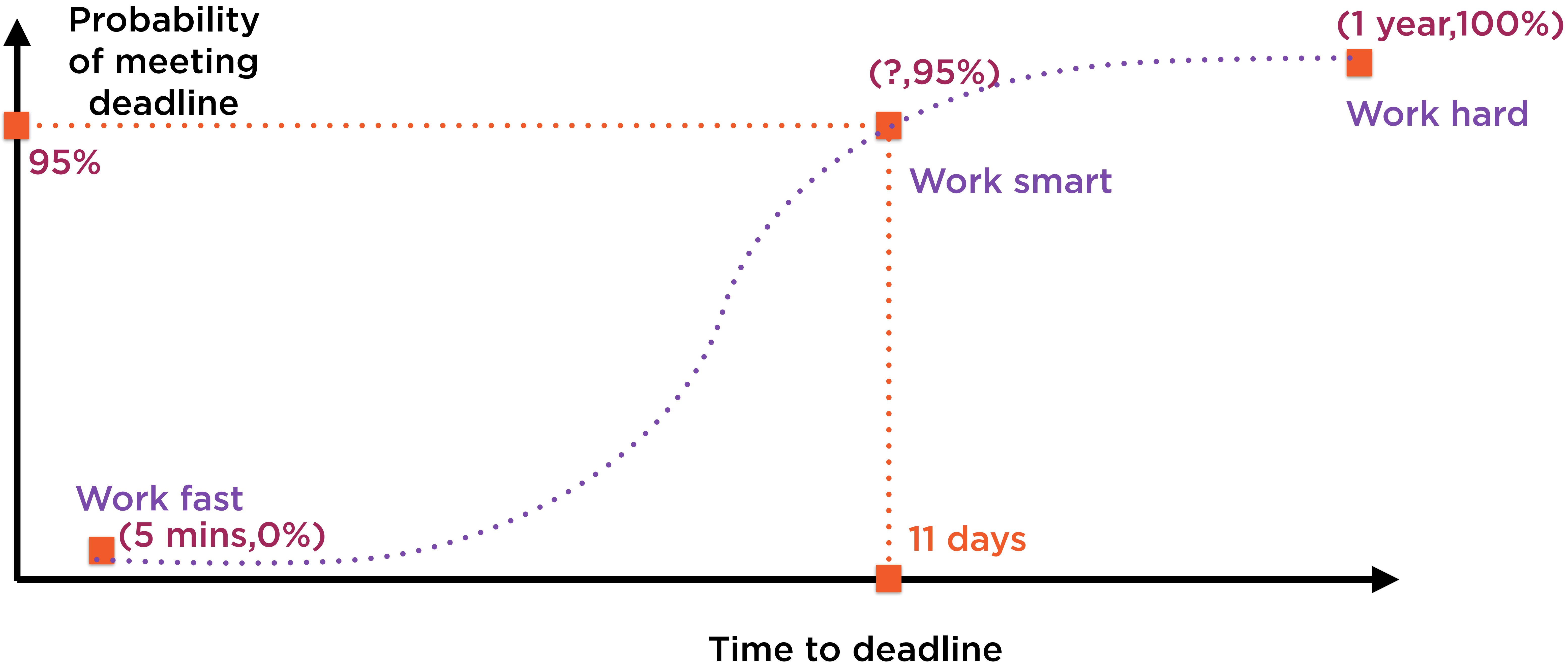
Working Hard, Fast, Smart



Working Hard, Fast, Smart



Working Hard, Fast, Smart



Predicting Future Events

Future events

Possible outcomes

Likely causes

Probabilities

Future events

- Investing savings in stocks
- Applying for a job at Google

Possible outcomes

- Make or lose money?
- Hired or not?

Likely causes

- interest rates, global growth, politics
- interview preparedness, quality of resume, hiring environment

Probabilities

- portfolio - up or down?
- job application - hired or not?

Common Applications of Logistic Regression

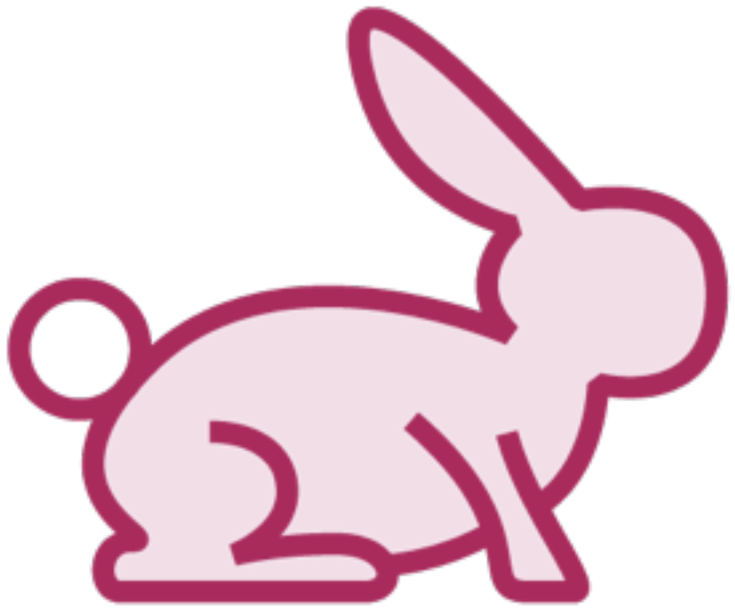
Analyse

Allocate

Predict

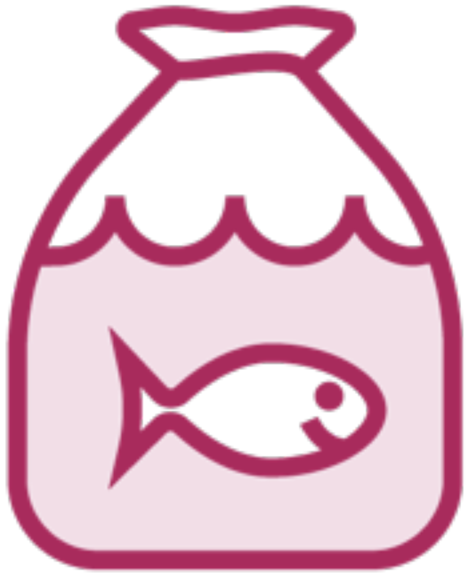
Classify

Whales: Fish or Mammals



Mammal

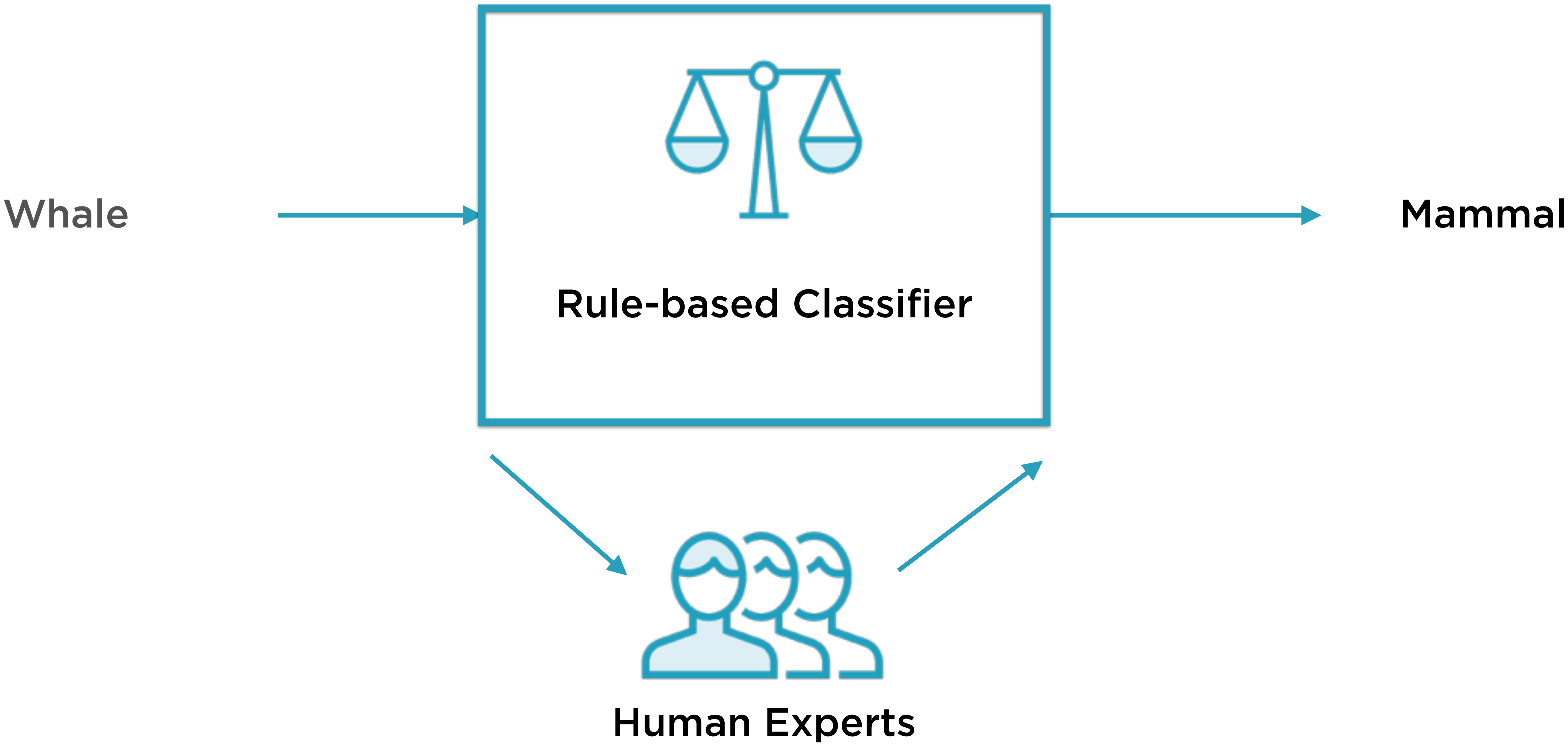
Member of the infraorder *Cetacea*



Fish

Looks like a fish, swims like a fish, moves like a fish

Rule-based Binary Classifier



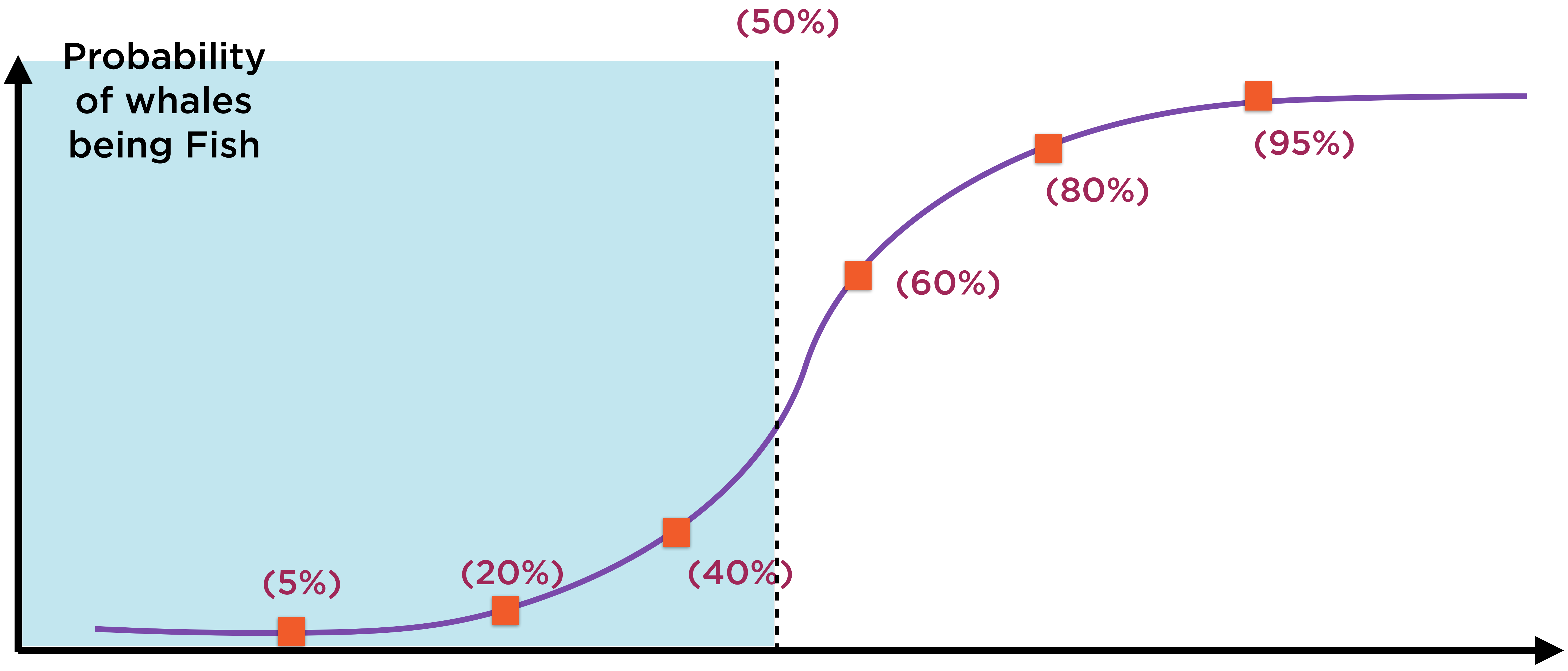
ML-based Binary Classifier

Corpus

Classification Algorithm

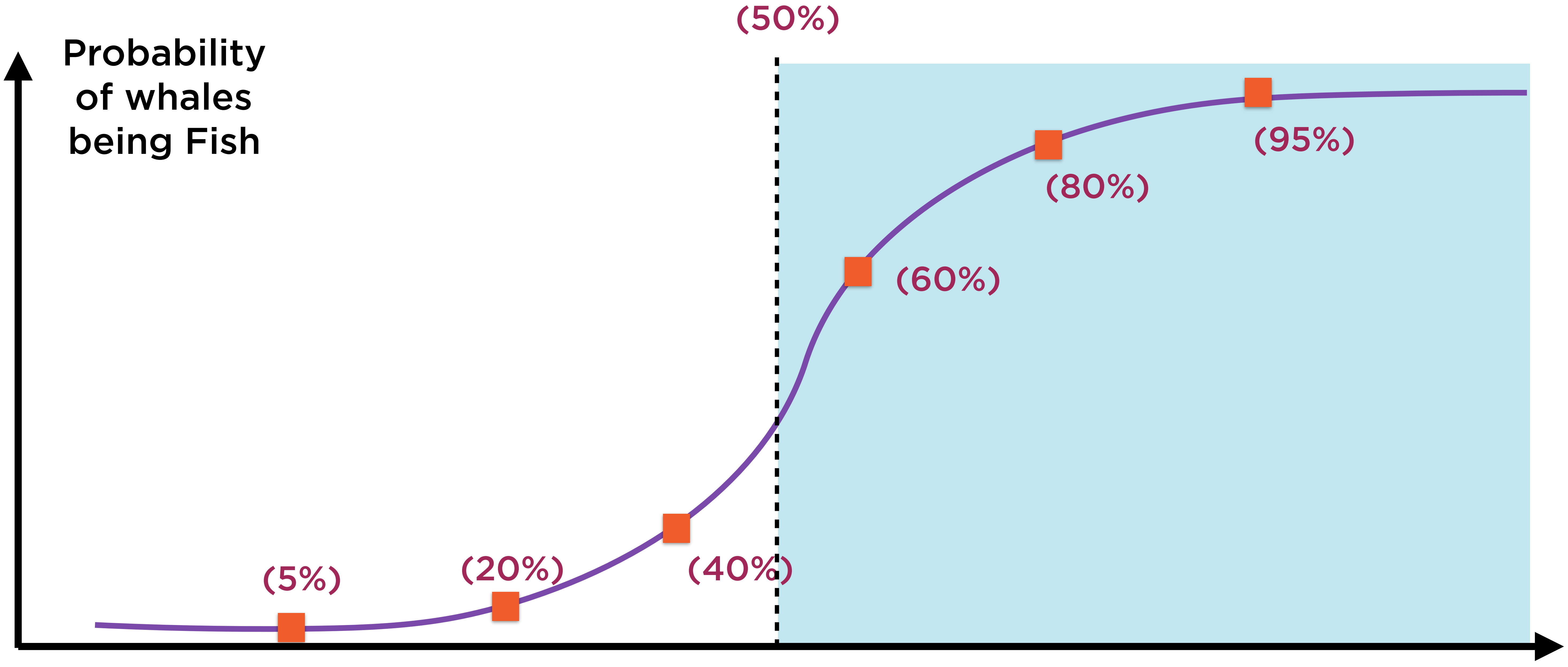
ML-based Classifier

Applying Logistic Regression



If probability < 50%, it's a mammal

Applying Logistic Regression



If probability > 50%, it’s a fish

Logistic Regression and Linear Regression

X Causes Y

Cause

Independent variable

Effect

Dependent variable

X Causes Y

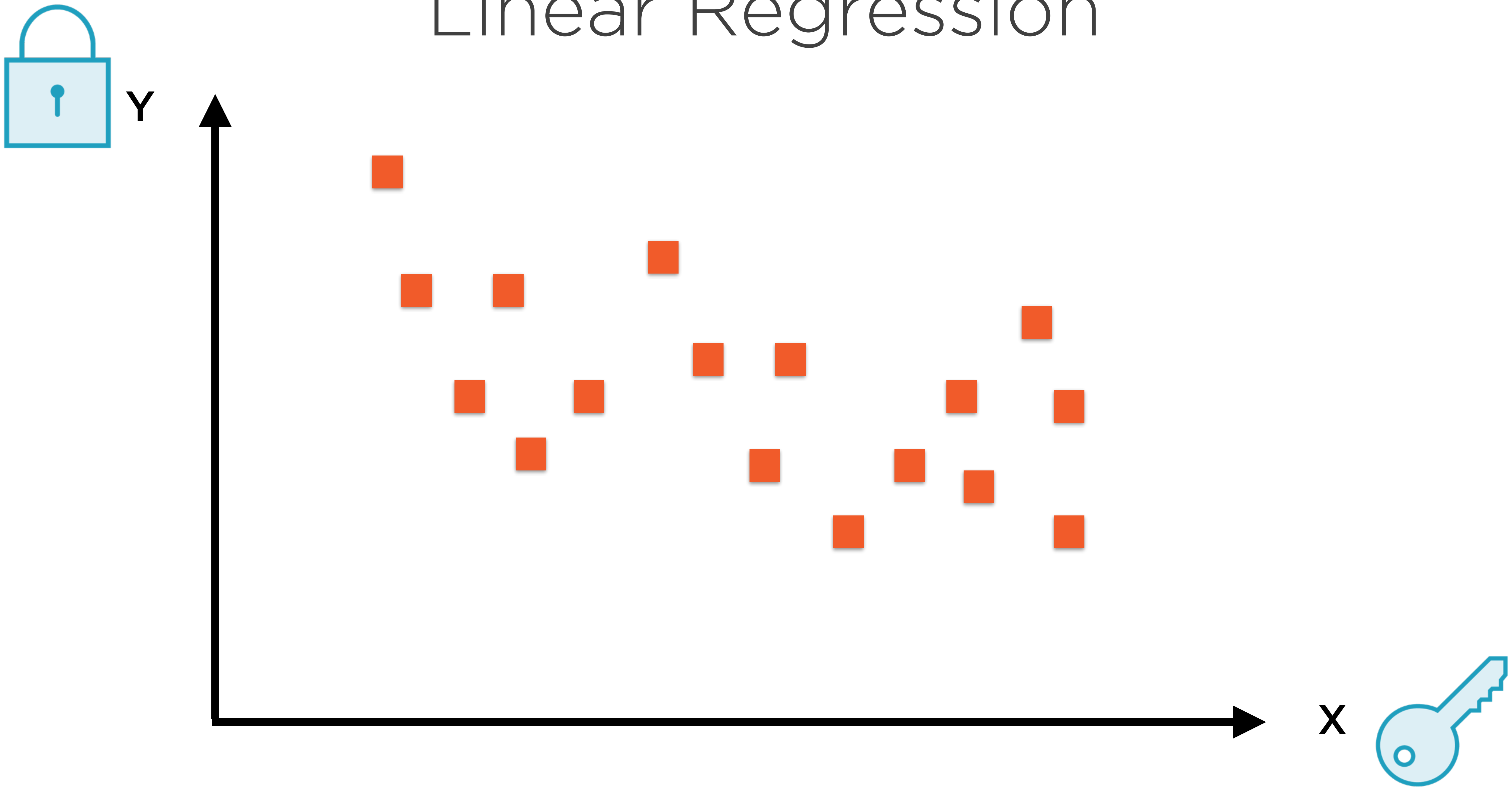
Cause

Explanatory variable

Effect

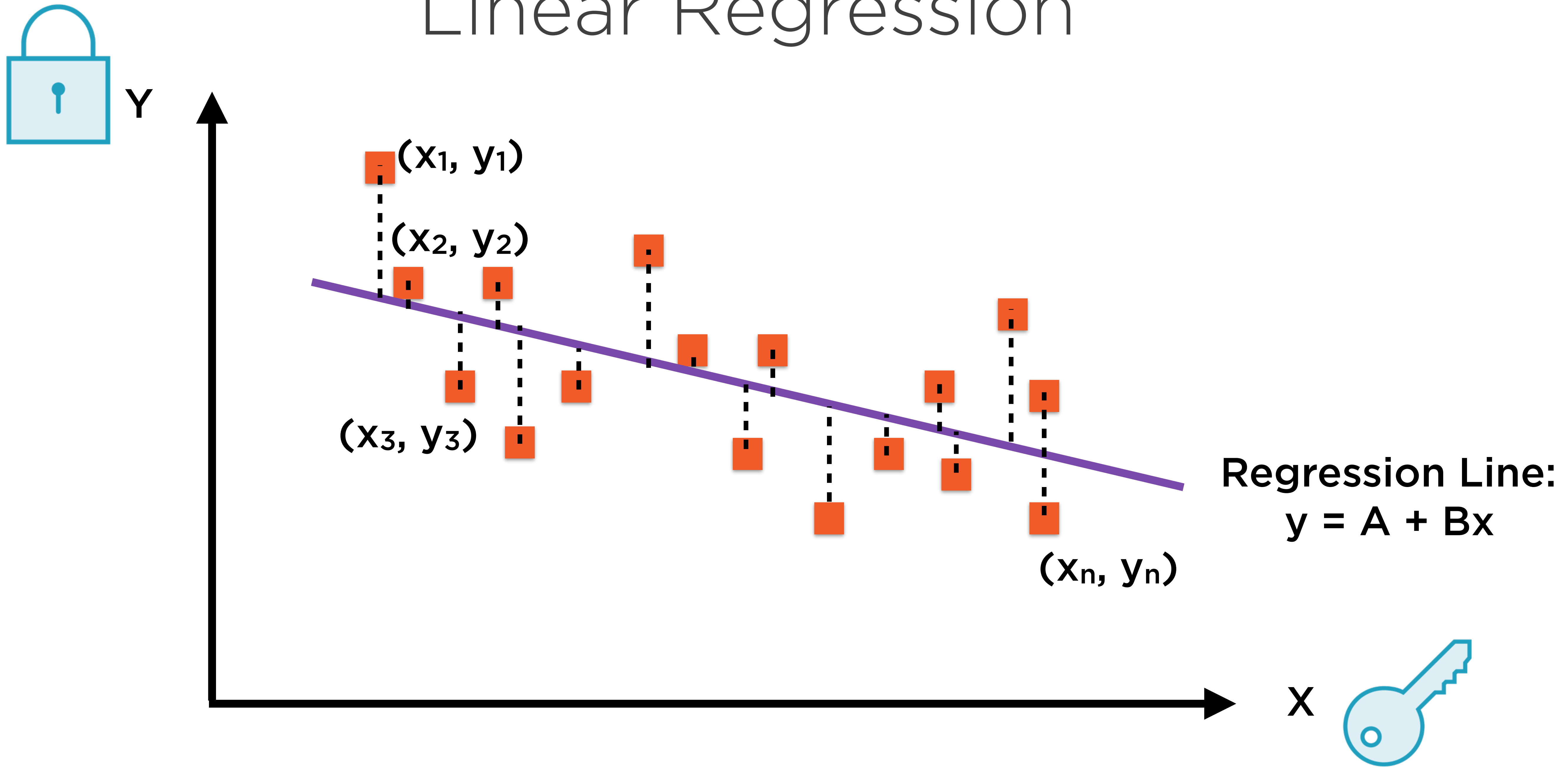
Dependent variable

Linear Regression



Represent all n points as (x_i, y_i), where $i = 1$ to n

Linear Regression

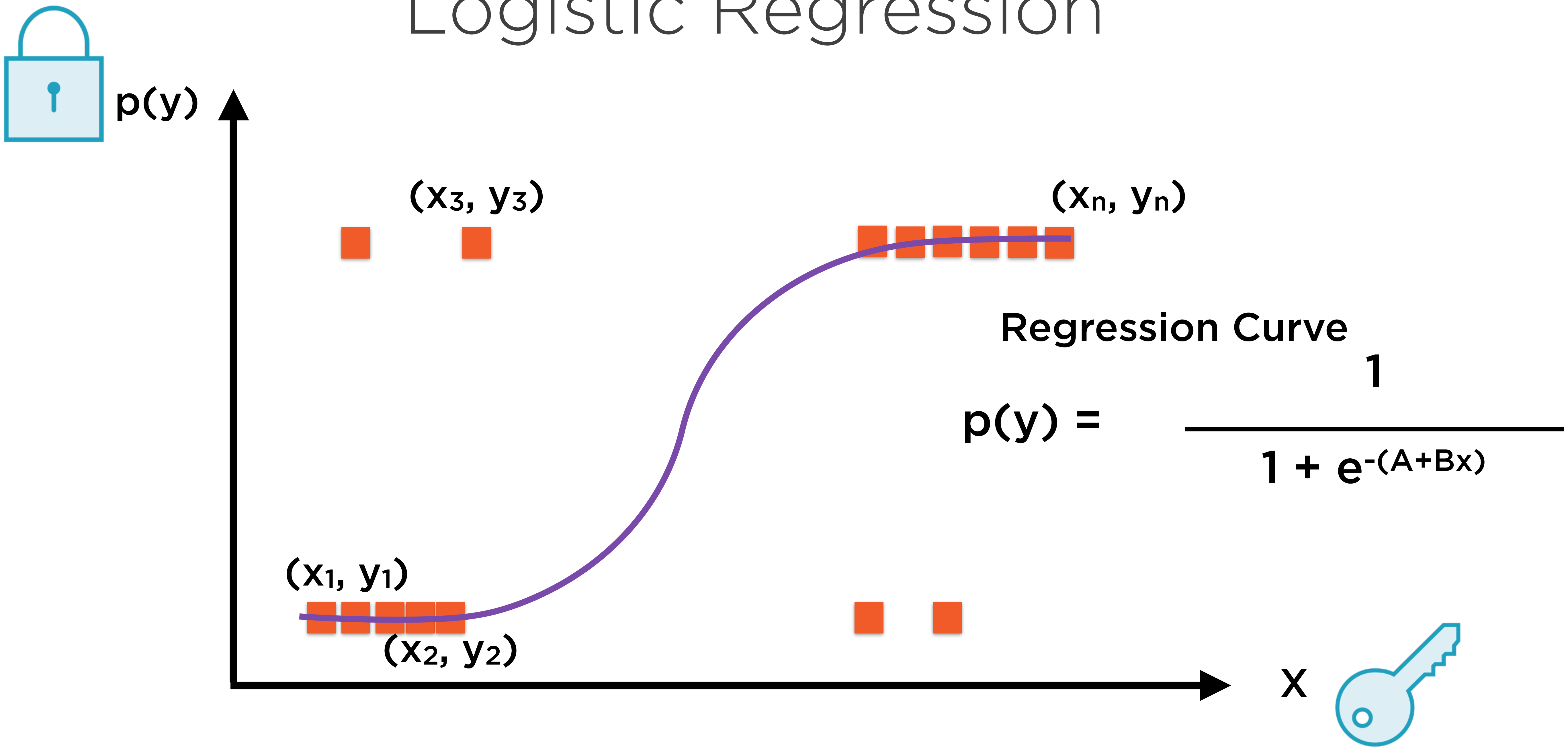


Represent all n points as (x_i, y_i), where i = 1 to n

Logistic Regression

Represent all n points as (x_i, y_i), where i = 1 to n

Logistic Regression

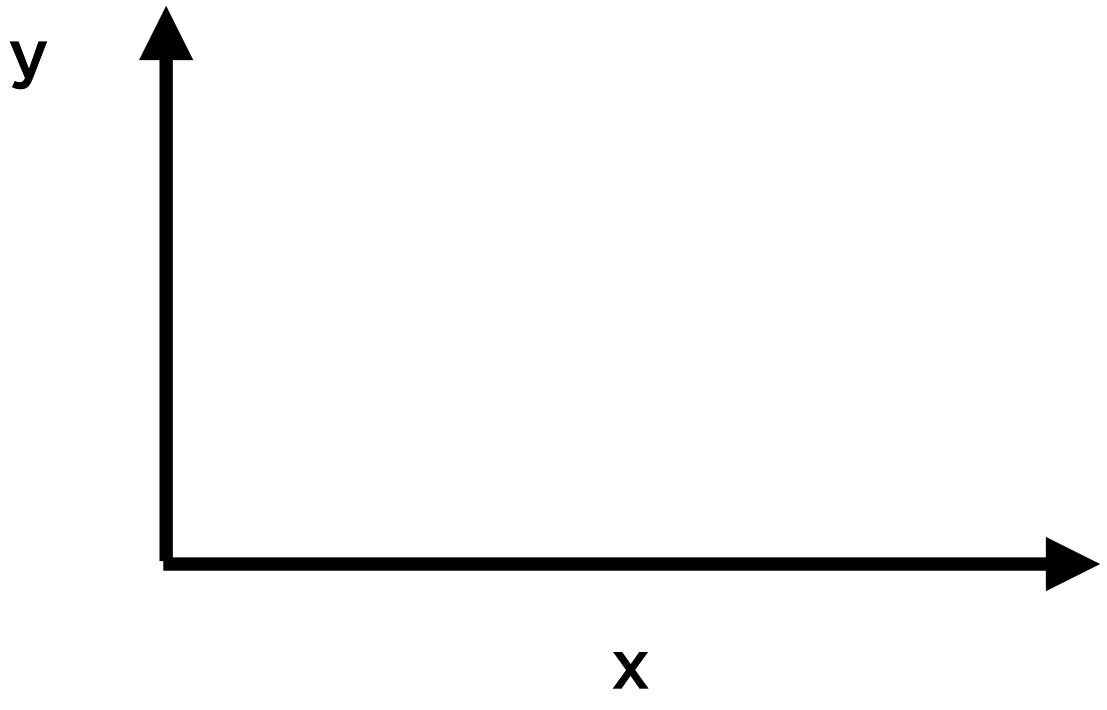


Represent all n points as
(x_i,y_i), where i = 1 to n

Similar, yet Different

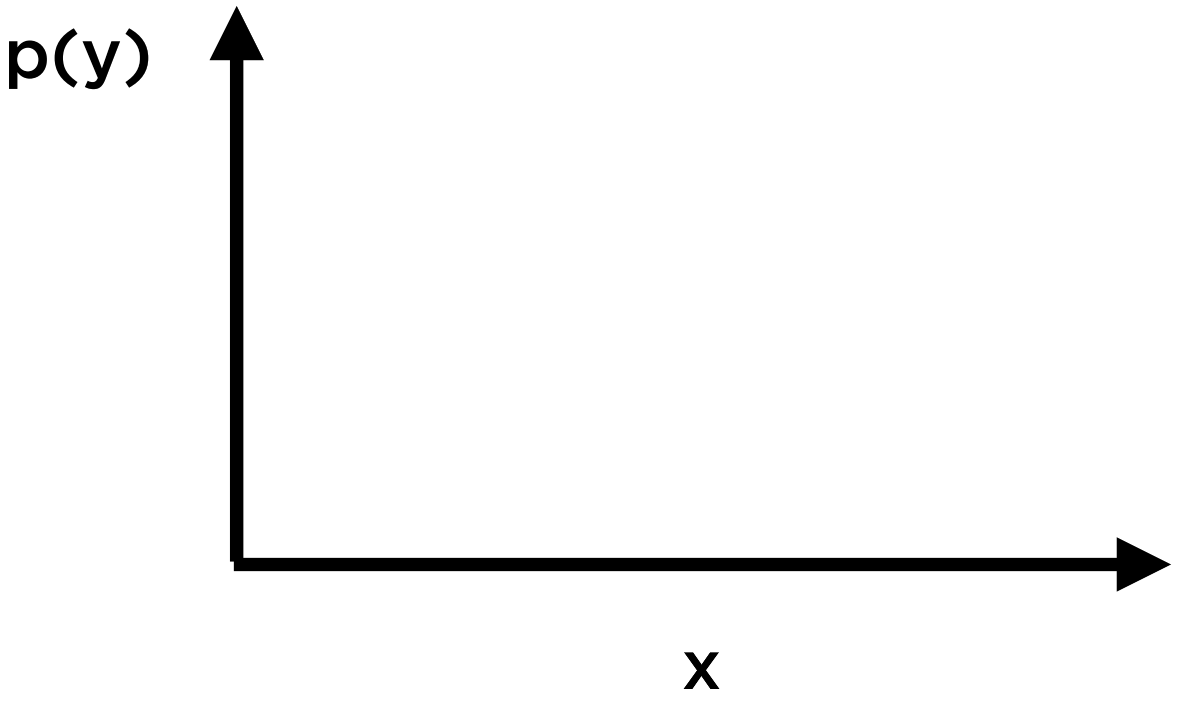
Linear Regression

Given causes, predict effect



Logistic Regression

Given causes, predict probability of effect



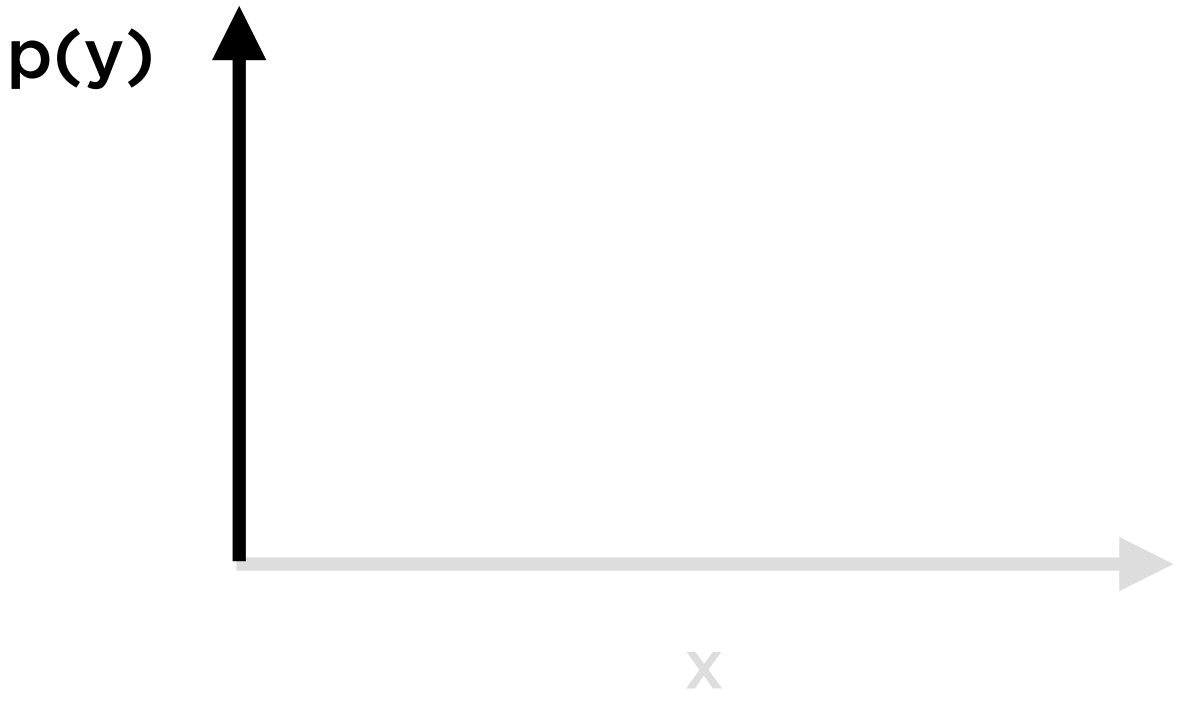
Similar, yet Different

Linear Regression

Effect variable (y) must be continuous

Logistic Regression

Effect variable (y) must be categorical



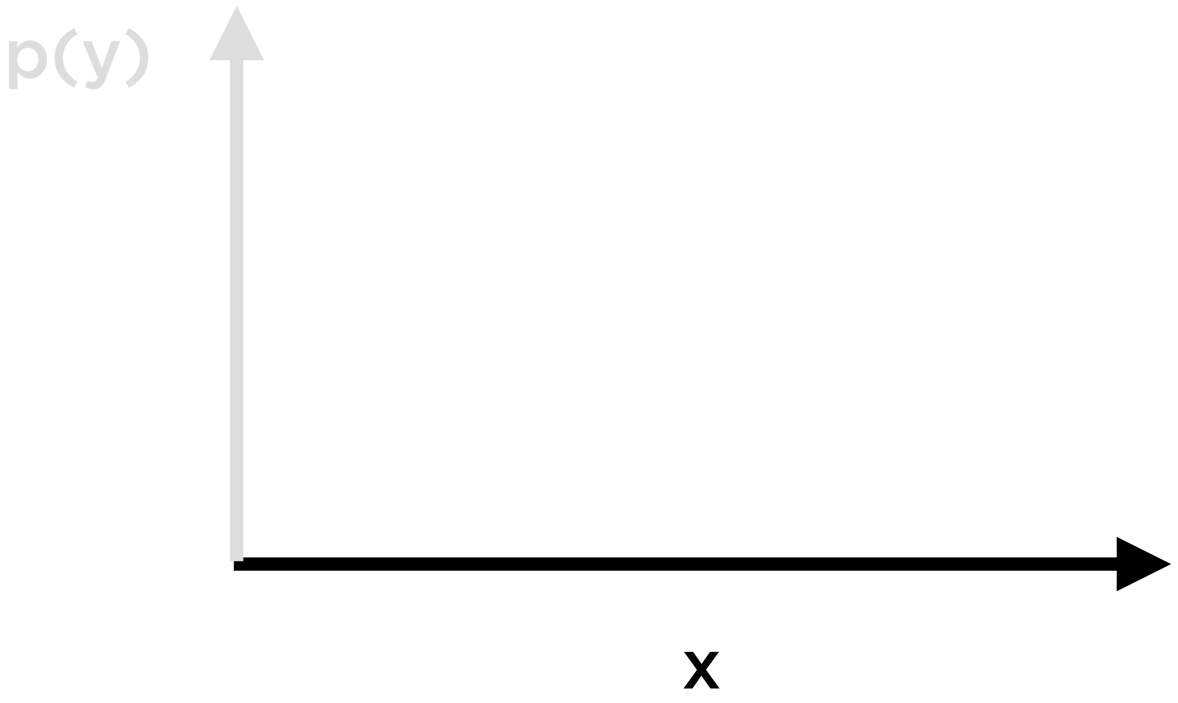
Similar, yet Different

Linear Regression

Cause variables (x) can be continuous or categorical

Logistic Regression

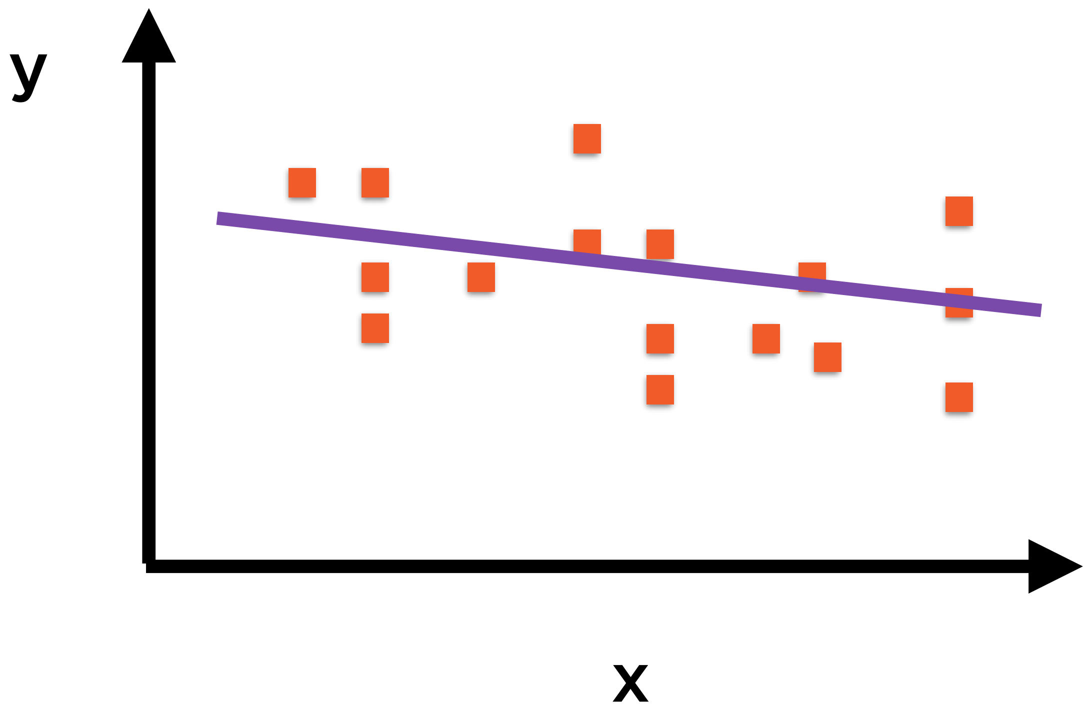
Cause variables (x) can be continuous or categorical



Similar, yet Different

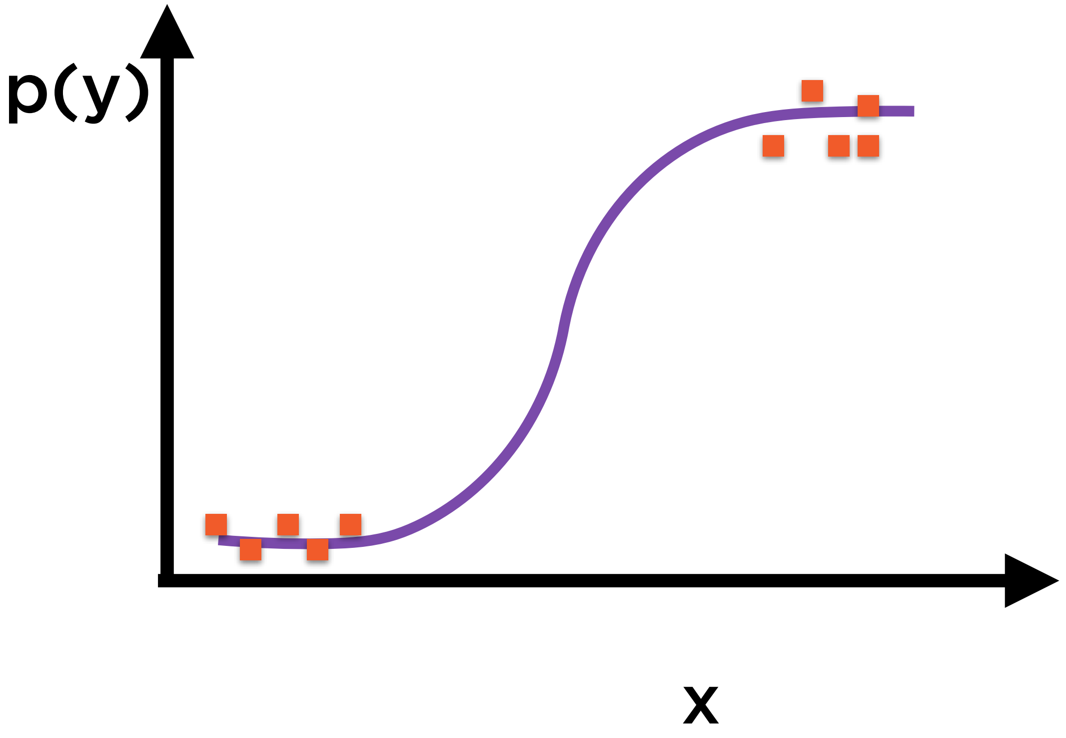
Linear Regression

Connect the dots with a straight line



Logistic Regression

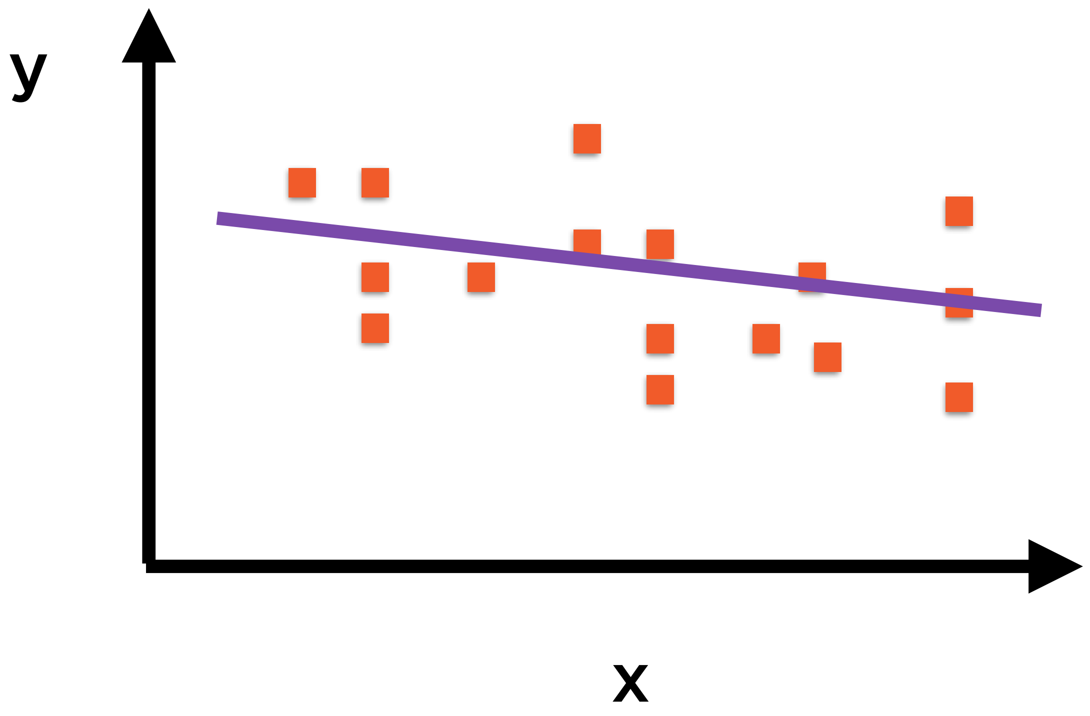
Connect the dots with an S-curve



Similar, yet Different

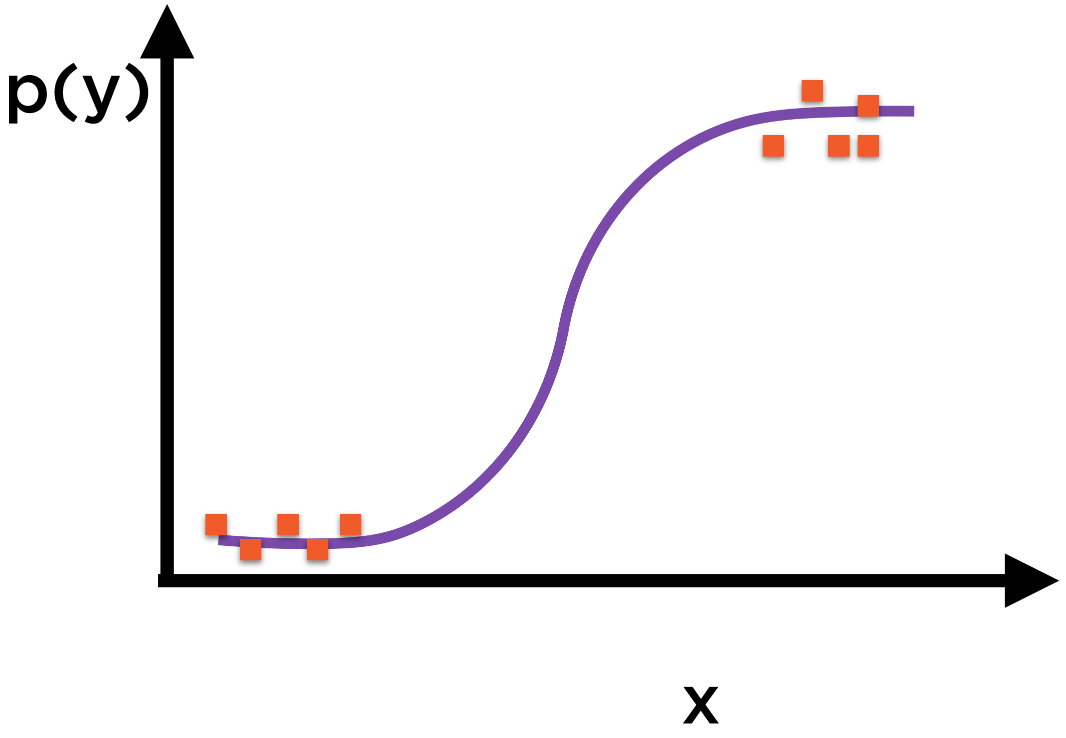
Linear Regression

$$y_i = A + Bx_i$$



Logistic Regression

$$p(y_i) = \frac{1}{1 + e^{-(A+Bx_i)}}$$



Similar, yet Different

Linear Regression

$$y_i = A + Bx_i$$

Objective of regression is to find A, B that "best fit" the data

Logistic Regression

$$p(y_i) = \frac{1}{1 + e^{-(A+Bx_i)}}$$

Objective of regression is to find A, B that "best fit" the data

Similar, yet Different

Linear Regression

$$y_i = A + Bx_i$$

Relationship is already linear (by assumption)

Logistic Regression

$$\ln\left(\frac{p(y_i)}{1 - p(y_i)}\right) = A + Bx_i$$

Relationship can be made linear (by log transformation)

Similar, yet Different

Linear Regression

$$y_i = A + Bx_i$$

Solve regression problem using cookie-cutter solvers

Logistic Regression

$$\text{logit}(p) = A + Bx_i$$

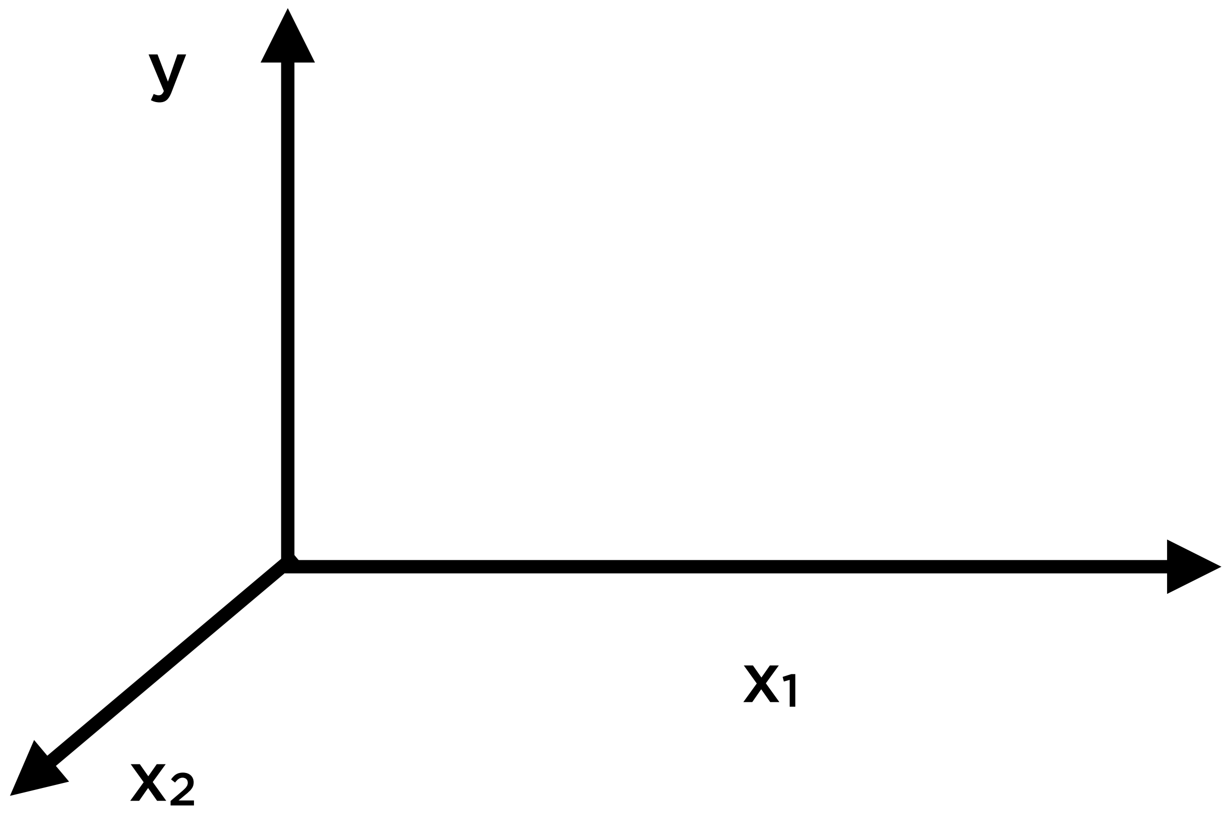
$$\text{logit}(p) = \ln\left(\frac{p}{1 - p}\right)$$

Solve regression problem using cookie-cutter solvers

Similar, yet Different

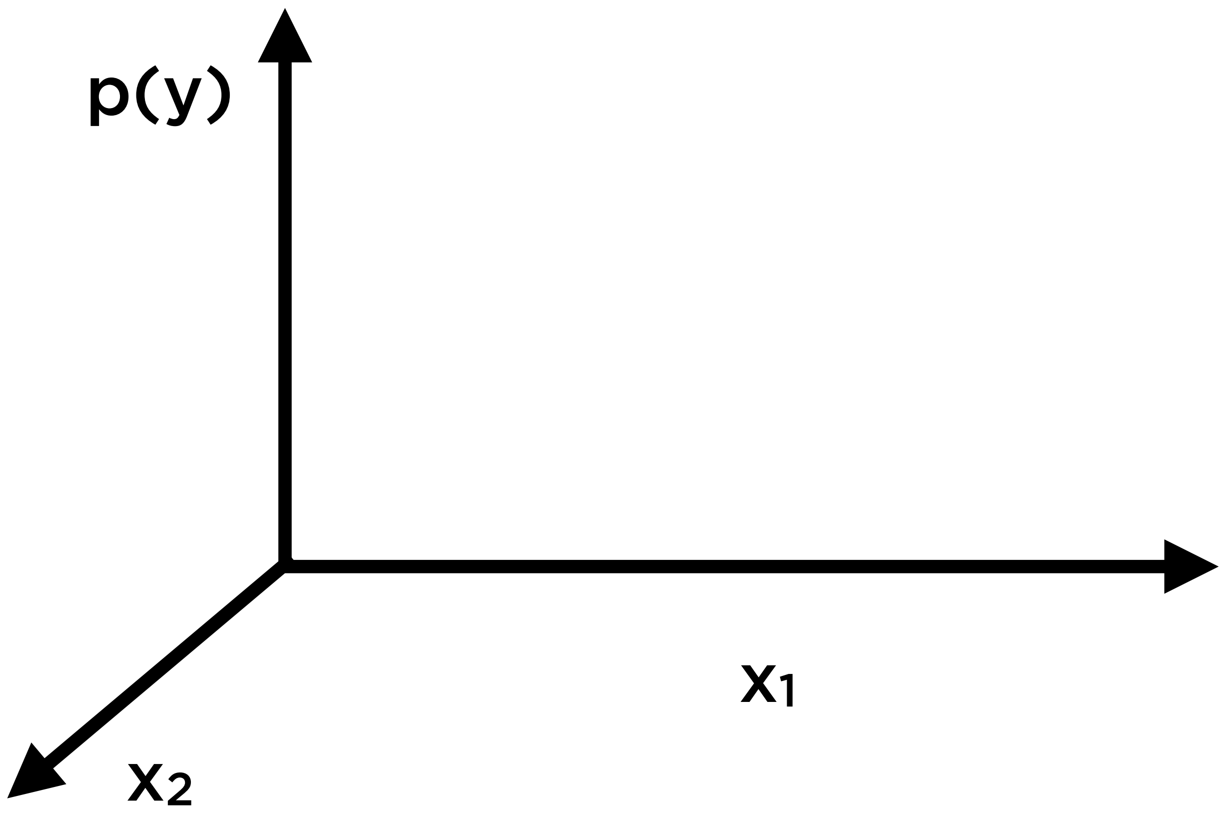
Linear Regression

Easily extended to multiple dimensions



Logistic Regression

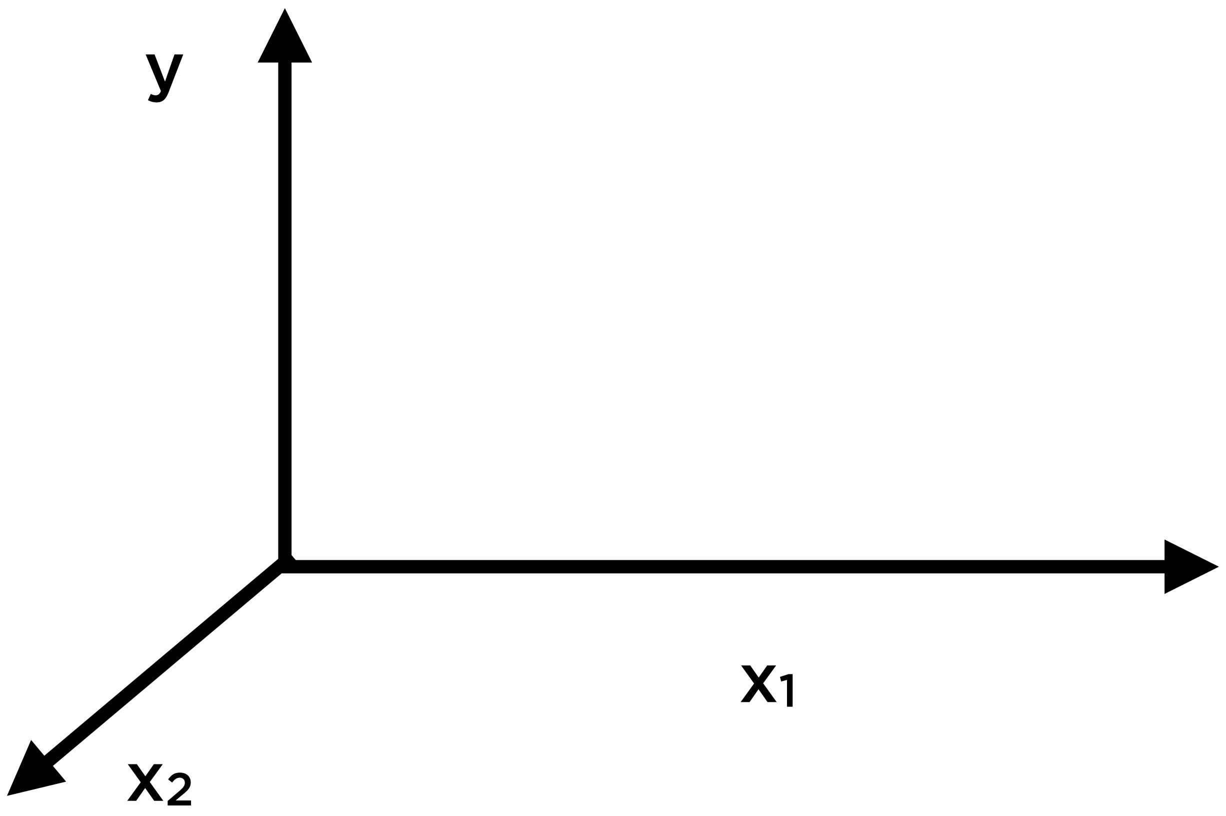
Easily extended to multiple dimensions



Similar, yet Different

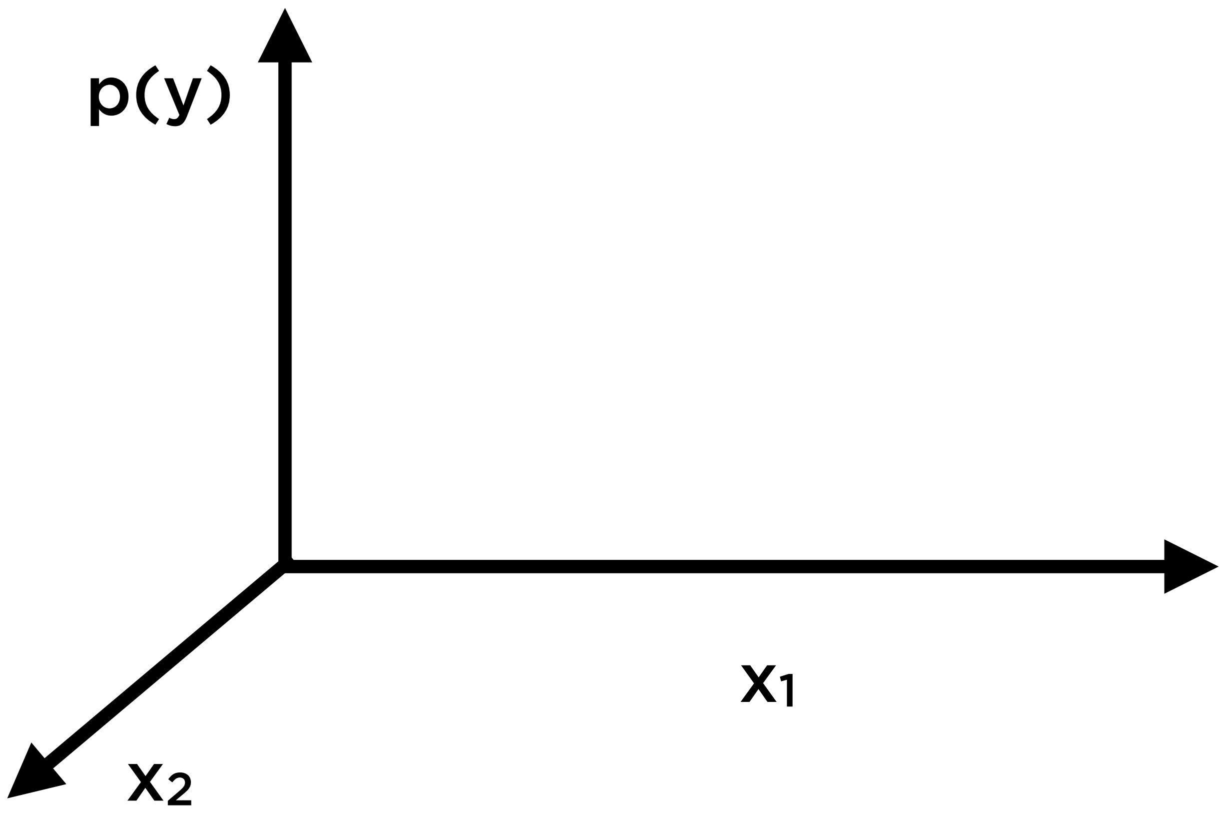
Linear Regression

Easily extended to multiple dimensions



Logistic Regression

Easily extended to multiple dimensions



Connecting the Dots with Regression

Linear Regression Equation:

$$y = A + Bx$$

$$
\begin{aligned}
y_1 &= A + Bx_1 \\
y_2 &= A + Bx_2 \\
y_3 &= A + Bx_3 \\
\ldots & \quad \ldots \\
y_n &= A + Bx_n
\end{aligned}
$$

Connecting the Dots with Regression

Linear Regression Equation:

$$y = A + Bx$$

$$y_1 = A + Bx_1 + e_1$$

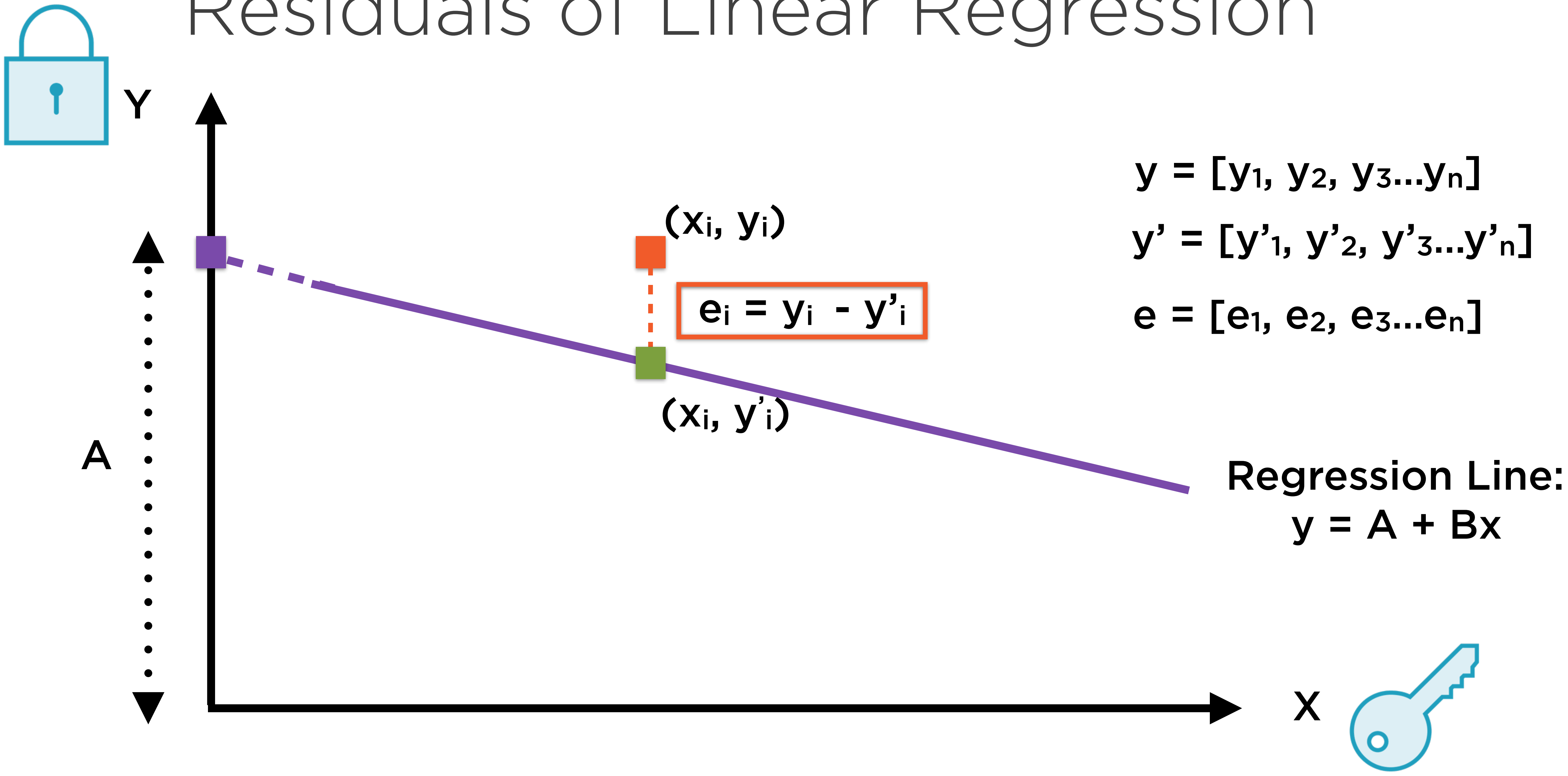
$$y_2 = A + Bx_2 + e_2$$

$$y_3 = A + Bx_3 + e_3$$

$$\ldots \qquad \ldots$$

$$y_n = A + Bx_n + e_n$$

Residuals of Linear Regression



Residuals of a regression are the difference between actual and fitted values of the dependent variable

Logistic Regression

Logistic Regression Equation:

$$p(y) = \frac{1}{1 + e^{-(A+Bx)}}$$

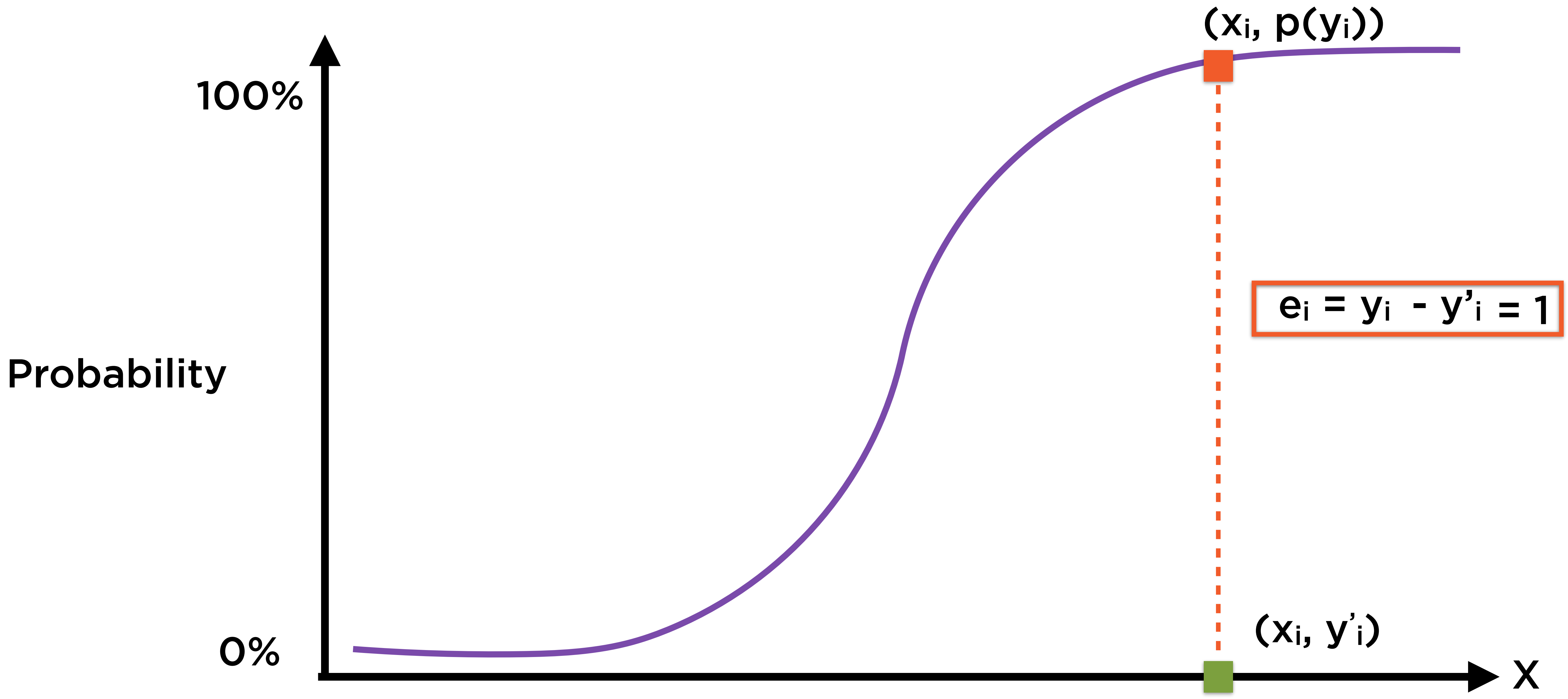
$$p(y_1) = \frac{1}{1 + e^{-(A+Bx_1)}}$$

$$p(y_2) = \frac{1}{1 + e^{-(A+Bx_2)}}$$

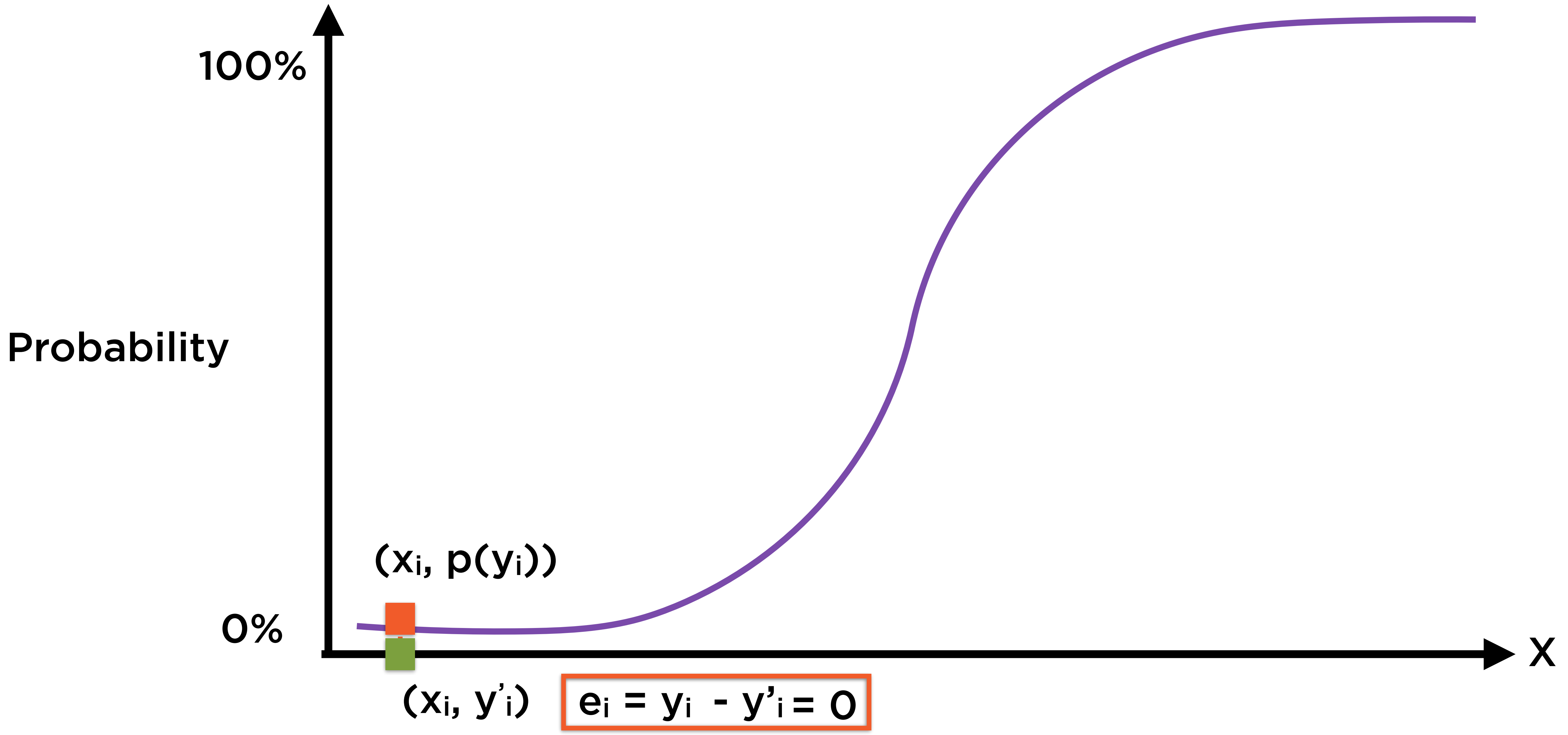
...

$$p(y_n) = \frac{1}{1 + e^{-(A+Bx_n)}}$$

Residuals of Linear Regression



Residuals of Linear Regression



Similar, yet Different

Linear Regression

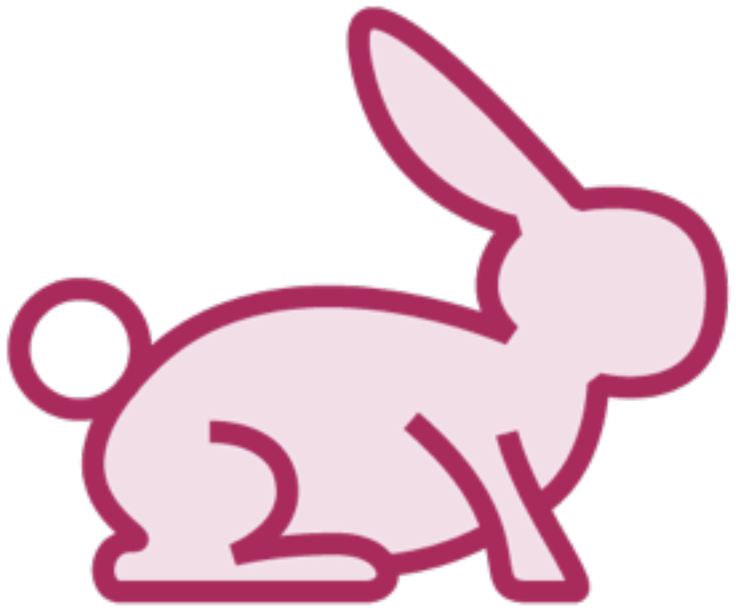
Residuals assumed to be normally distributed

Logistic Regression

Residuals cannot be normally distributed

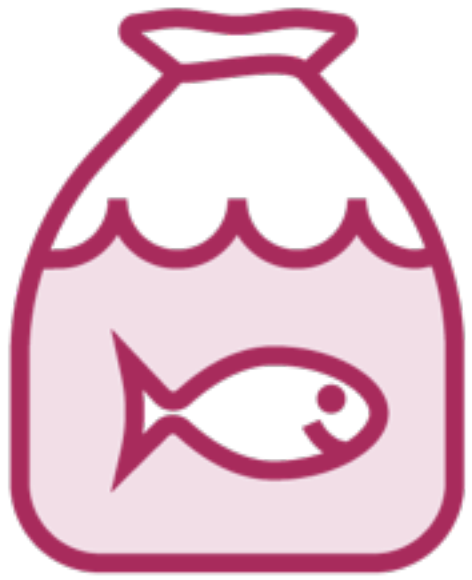
Logistic Regression and Machine Learning

Whales: Fish or Mammals



Mammal

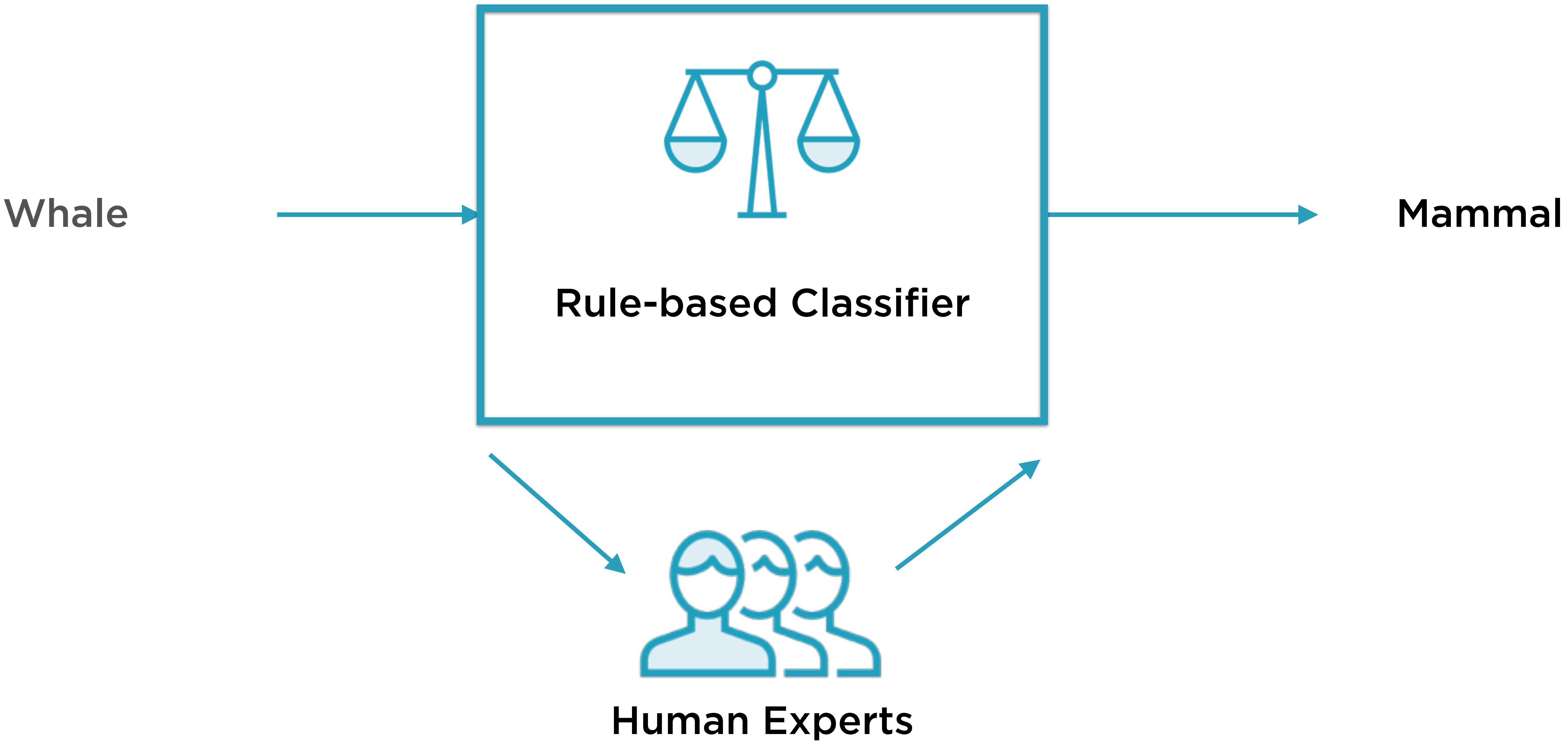
Member of the infraorder *Cetacea*



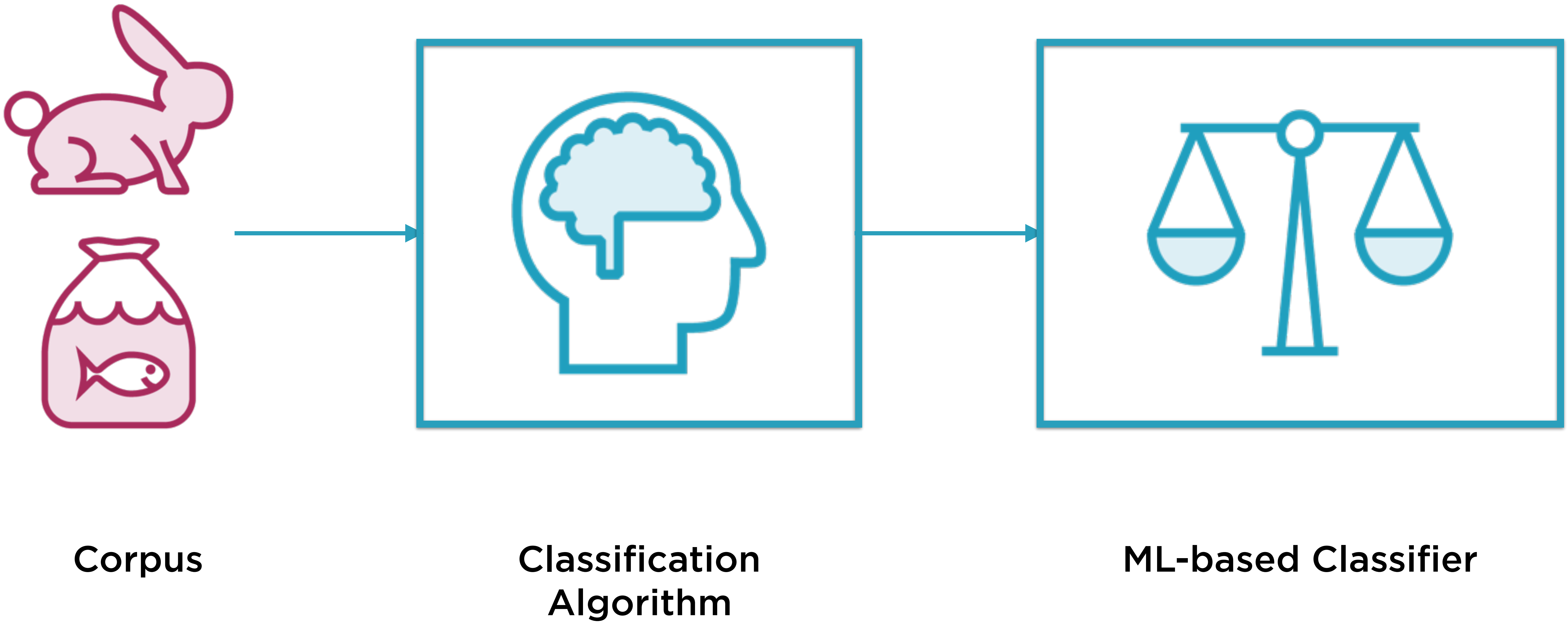
Fish

Looks like a fish, swims like a fish, moves like a fish

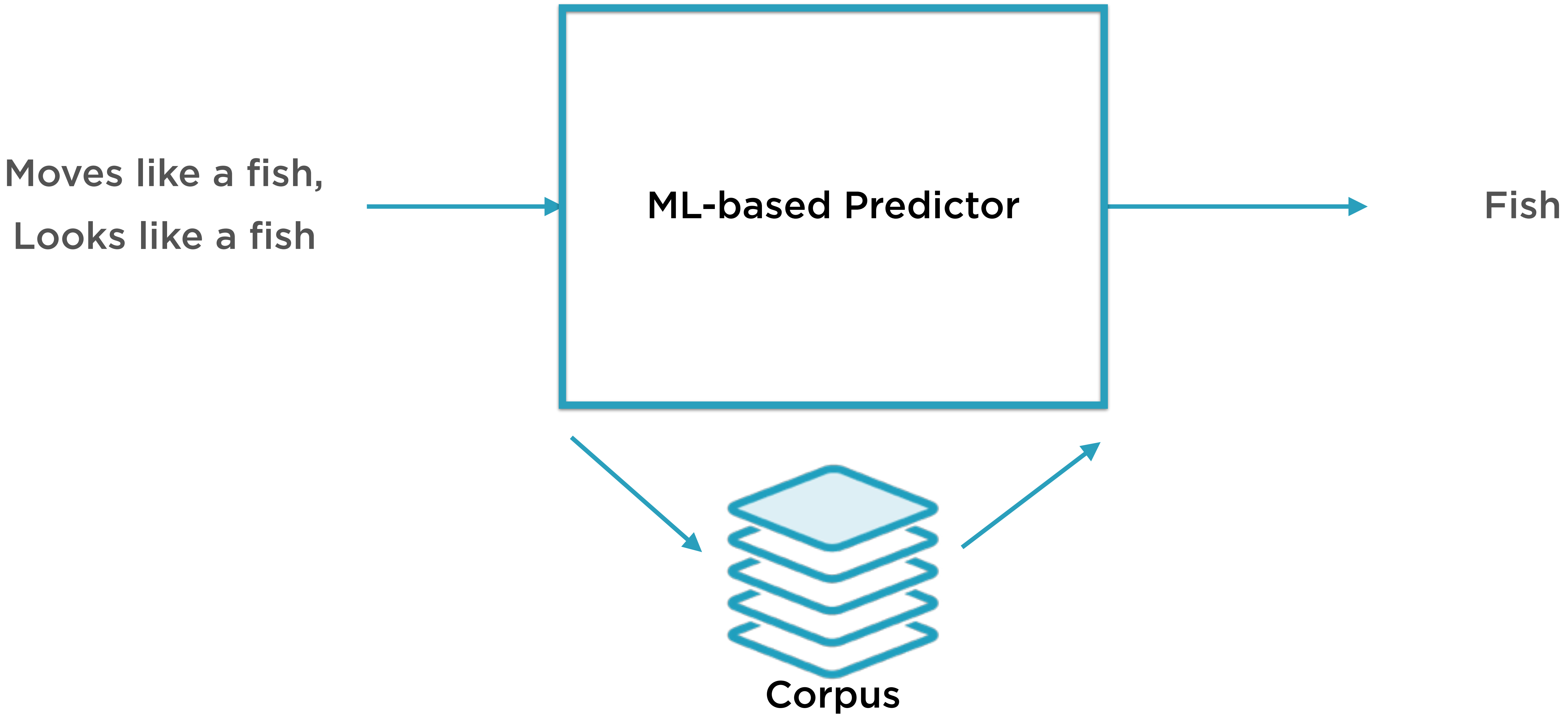
Rule-based Binary Classifier



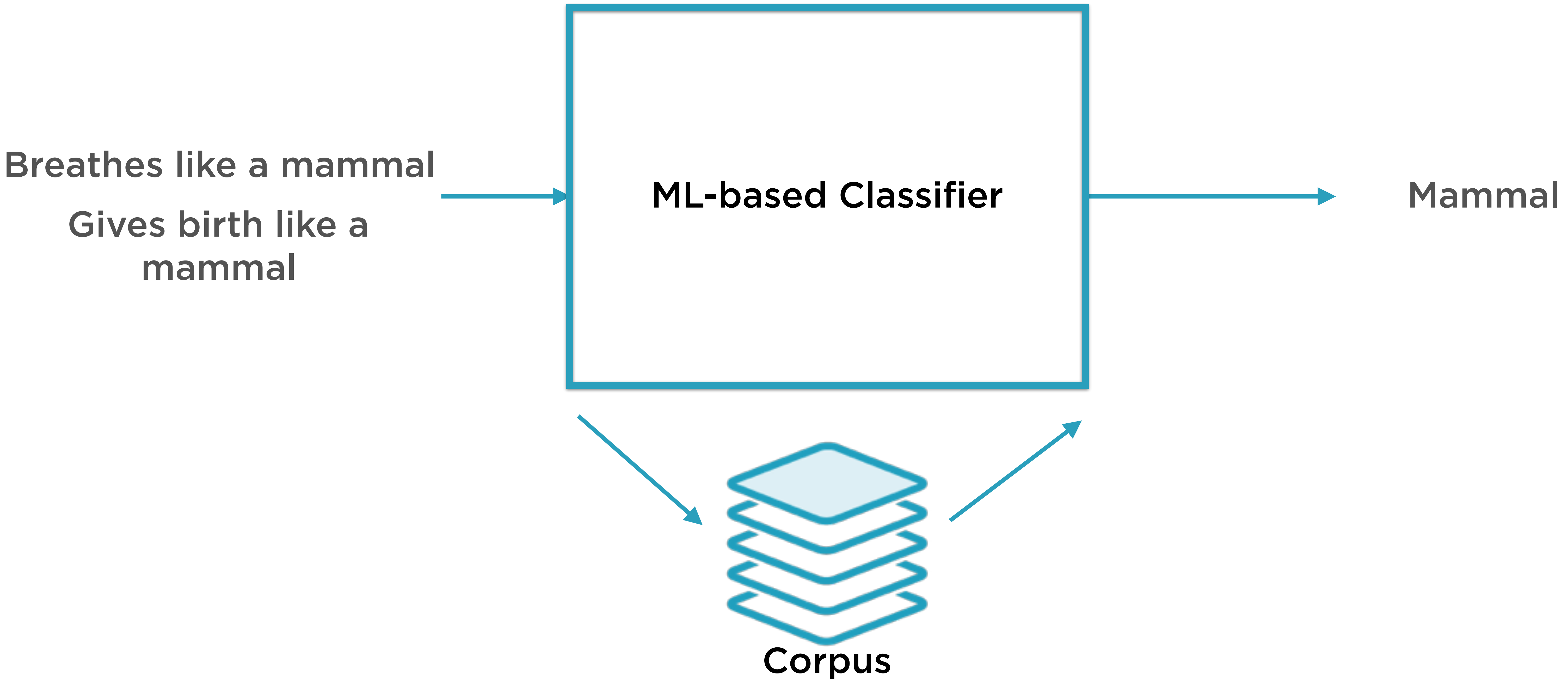
ML-based Binary Classifier



ML-based Binary Classifier



ML-based Binary Classifier



Rule-based or ML-based?

ML-based	Rule-based
Dynamic	Static
Experts optional	Experts required
Corpus required	Corpus optional
Training step	No training step

ML-based Predictor

Corpus

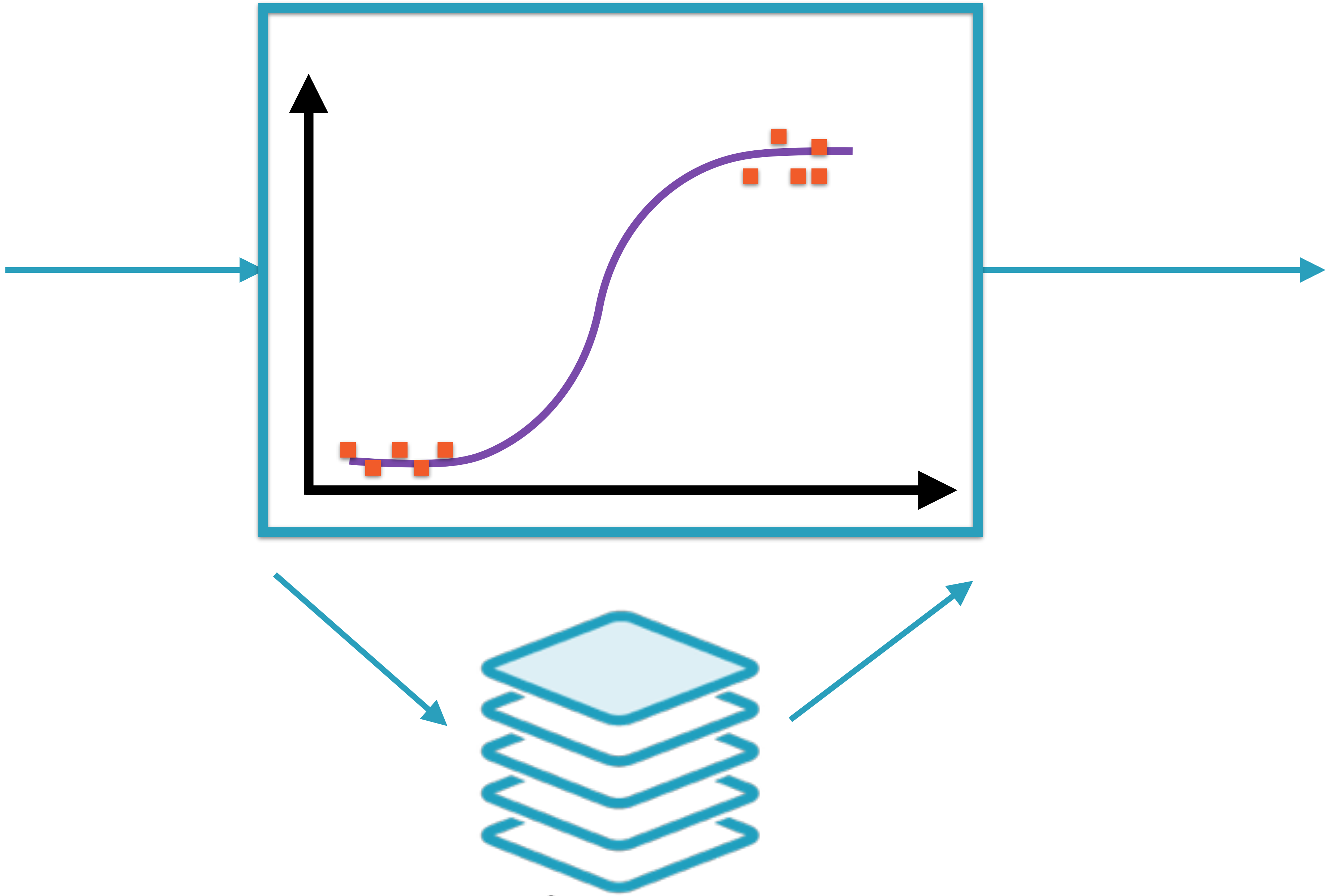
Logistic Regression

ML-based Predictor

$$p(y_i) = \frac{1}{1 + e^{-(A+Bx_i)}}$$

ML-based Predictor

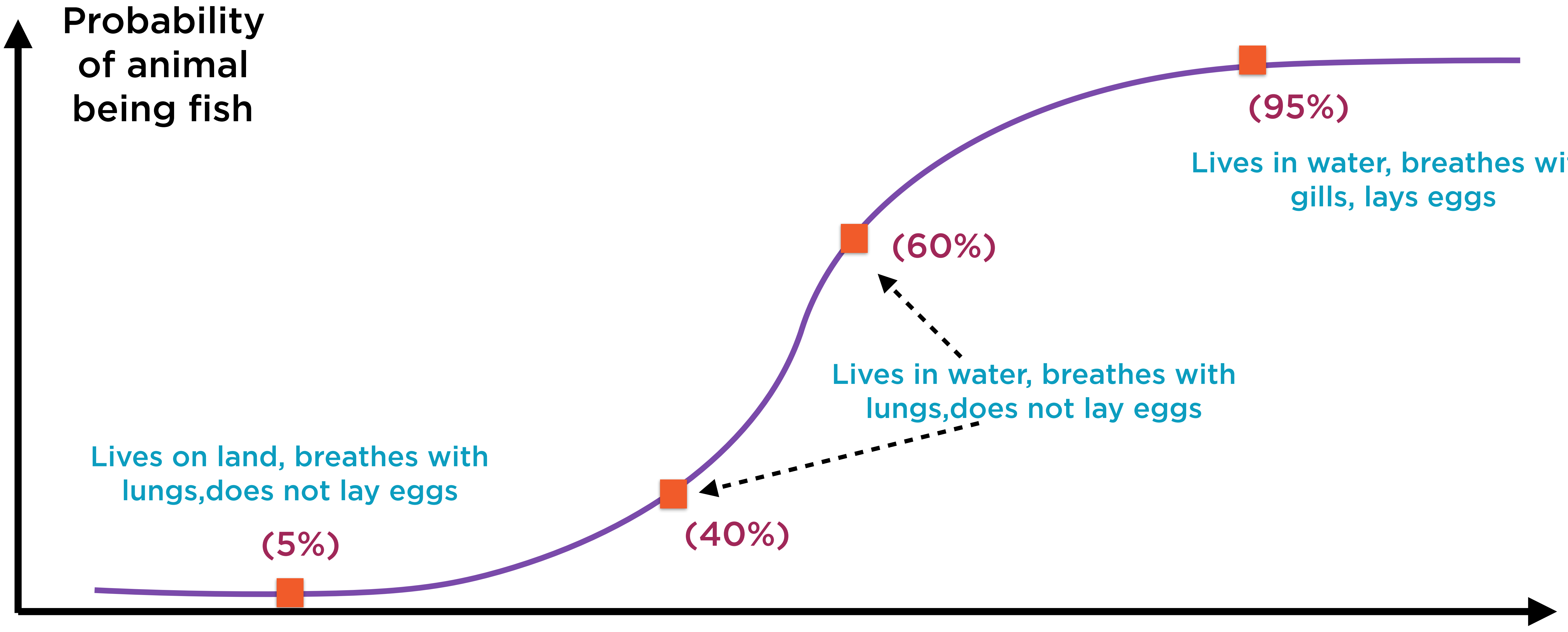
Lives in water, breathes with lungs,does not lay eggs



P(mammal) = 0.55

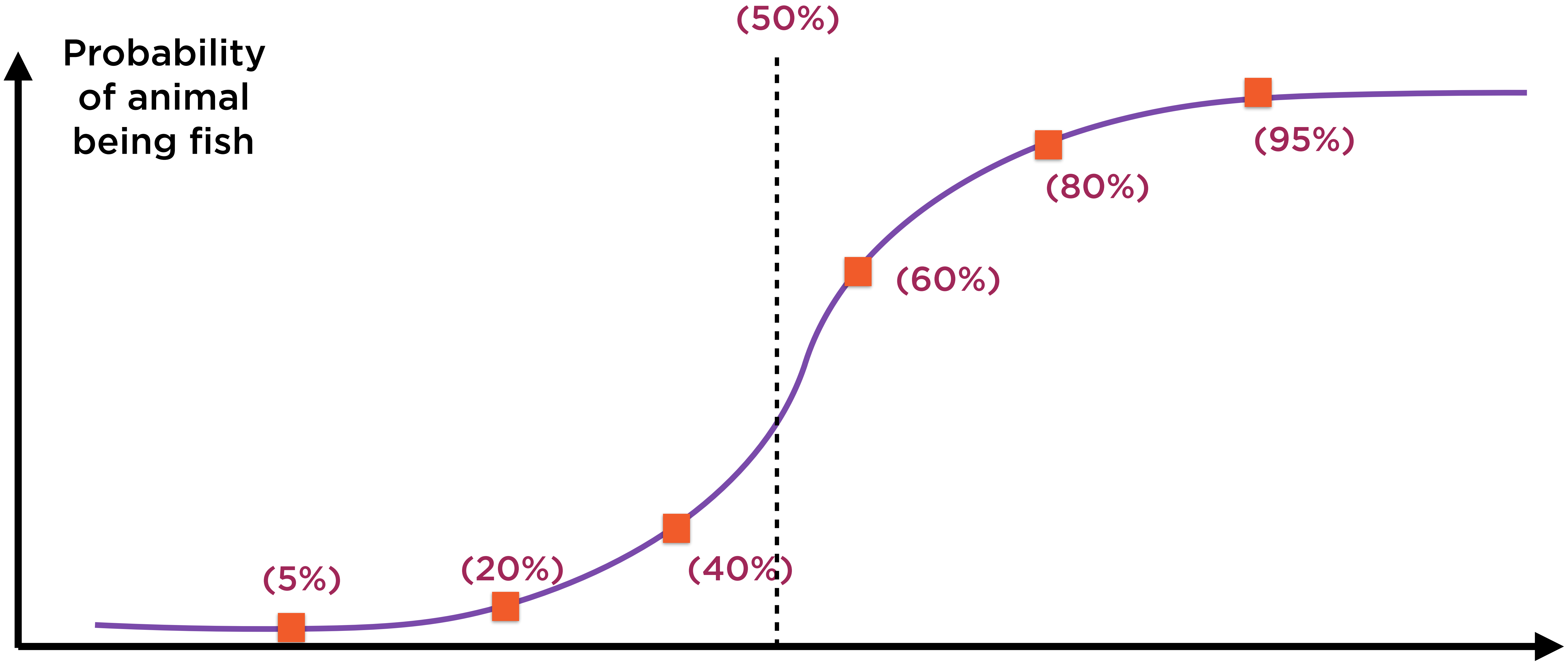
Corpus

Applying Logistic Regression



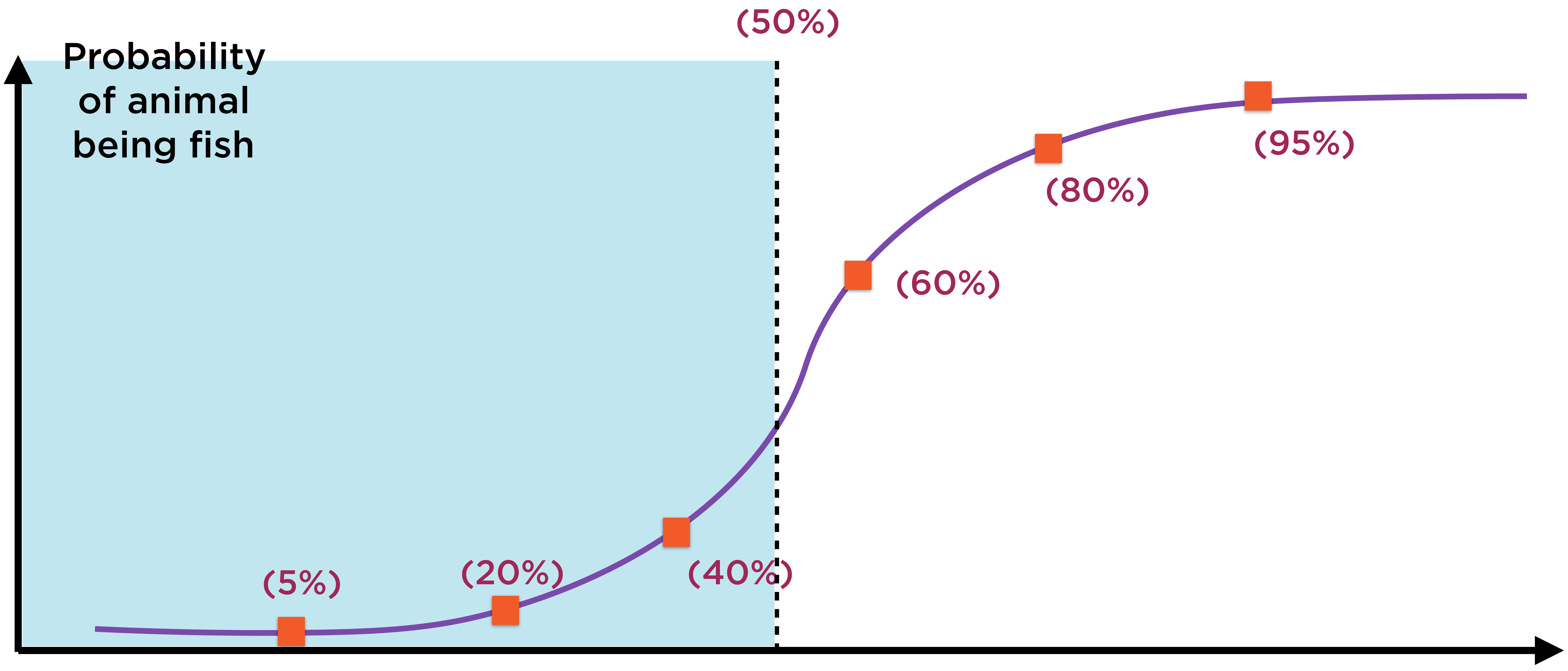
Whales: Fish or Mammals?

Applying Logistic Regression



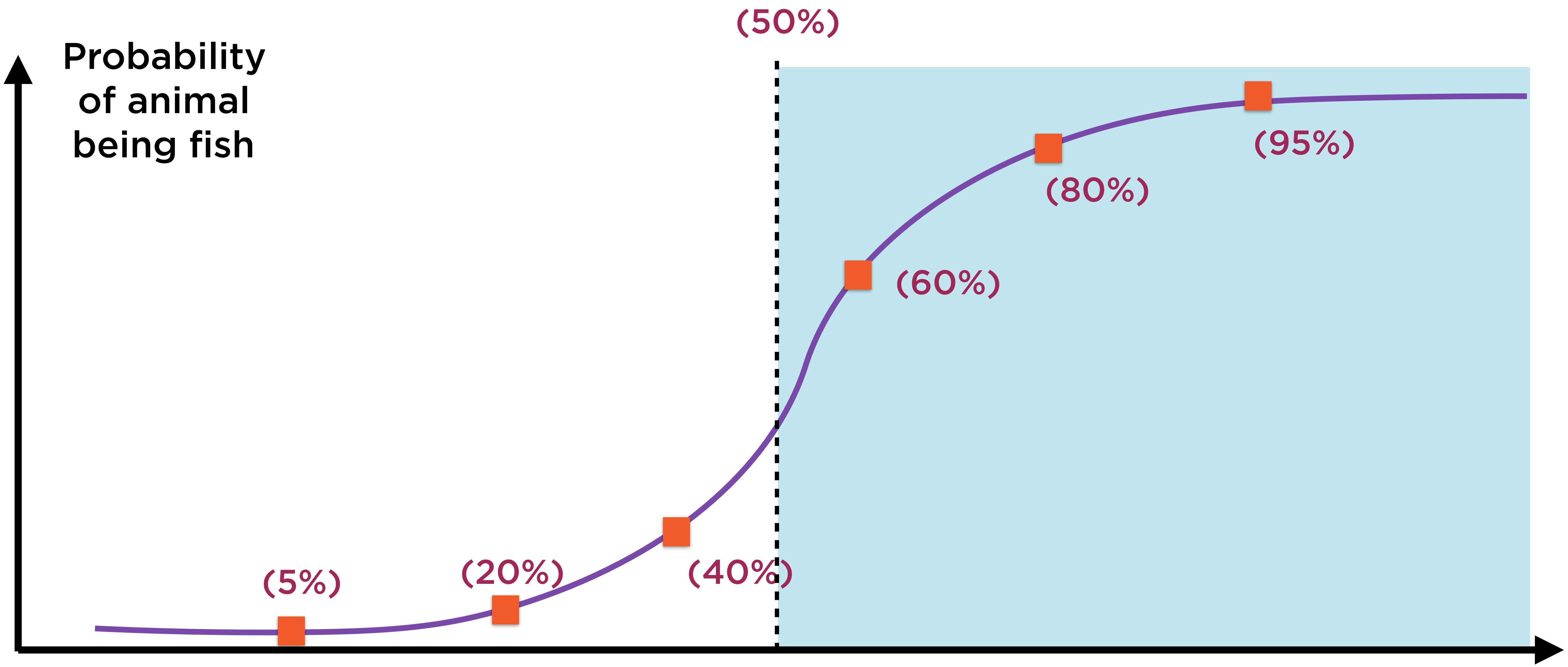
Rule of 50%

Applying Logistic Regression



If probability < 50%, it’s a mammal

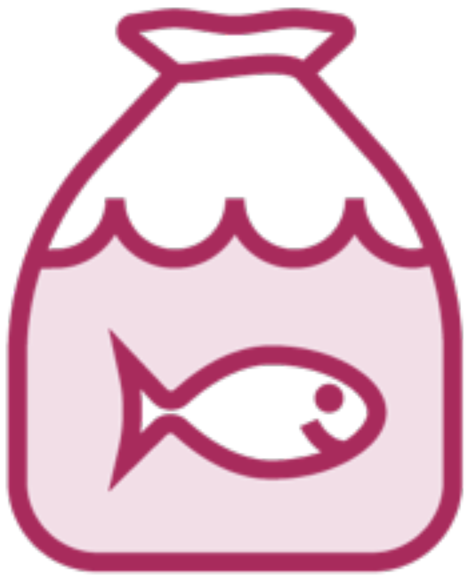
Applying Logistic Regression



If probability > 50%, it’s a fish

Applying Logistic Regression

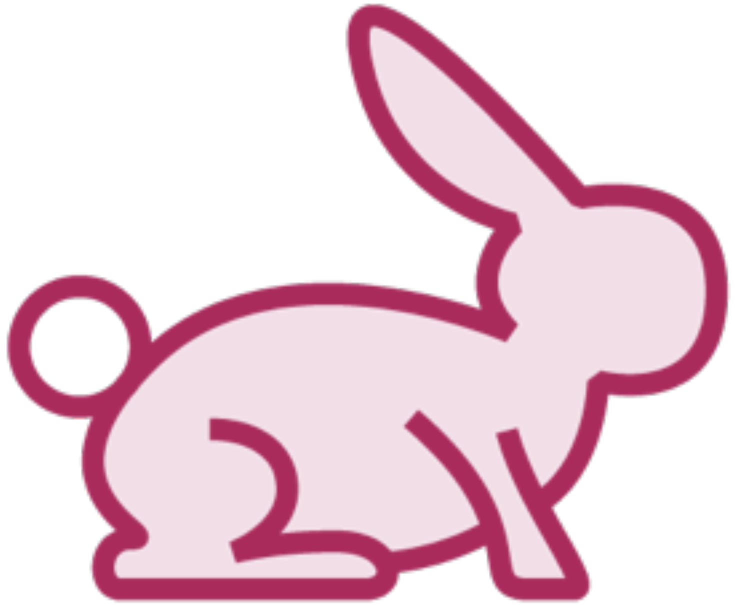
Mammal



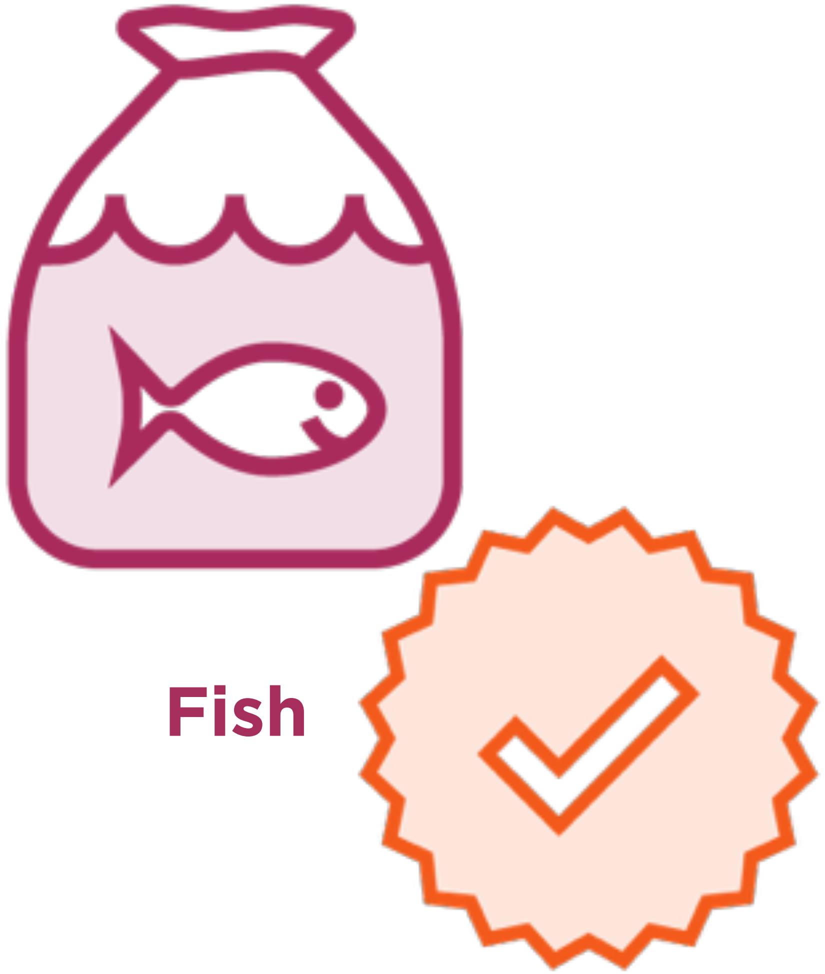
Fish

Probability of whales being Fish < 50%

Applying Logistic Regression



Mammal



Fish

Probability of whales being Fish > 50%

$$p(y_i) = \frac{1}{1 + e^{-(A+Bx_i)}}$$

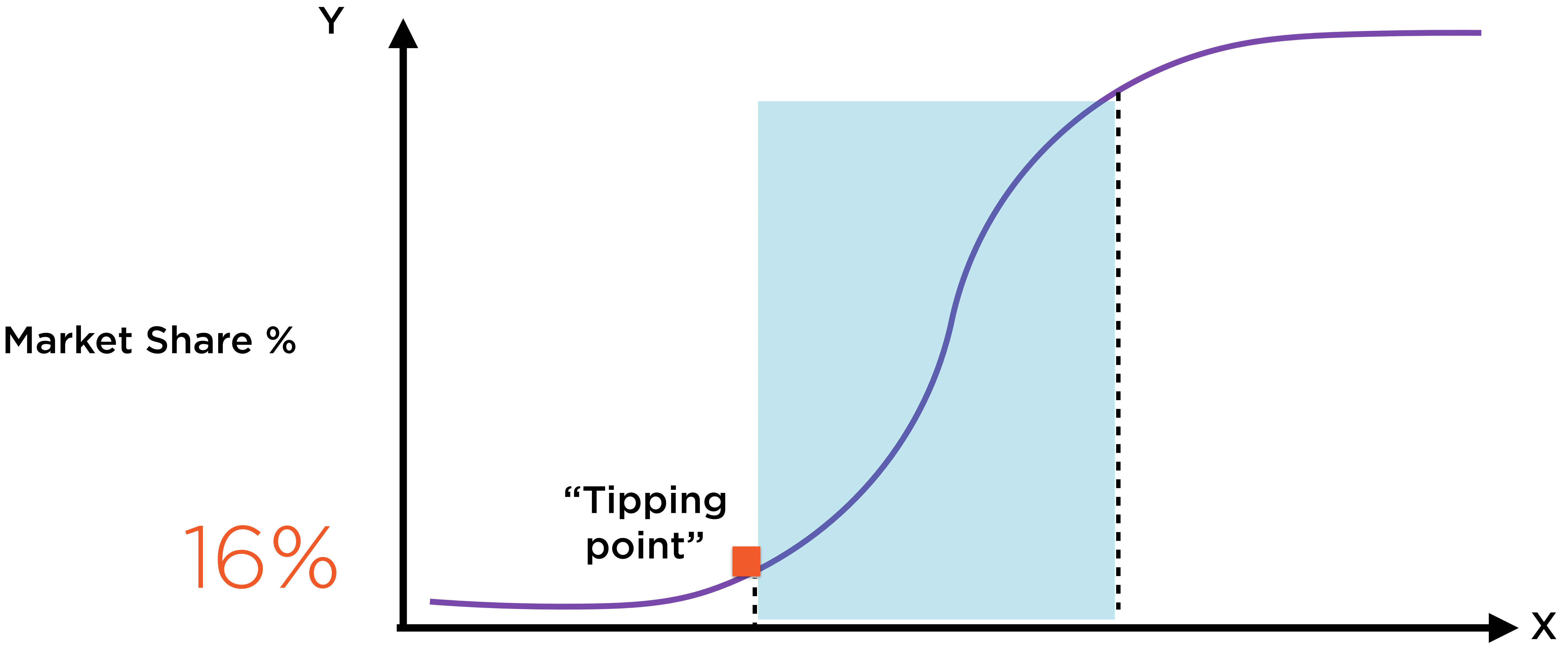
Logistic regression involves finding the “best fit” such curve

- A is the intercept

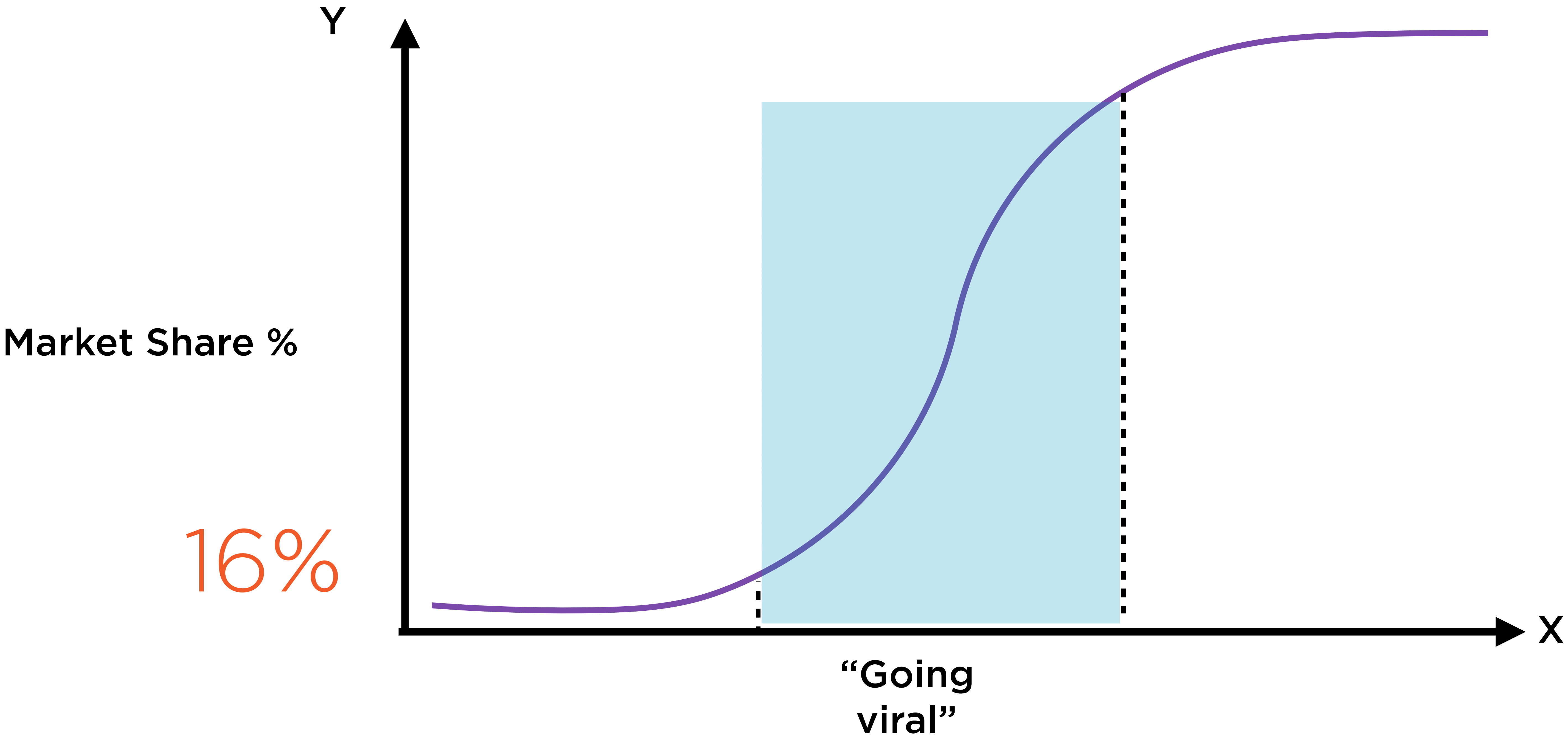
- B is the regression coefficient

(e is the constant 2.71828)

Diffusion of Innovation



Diffusion of Innovation



Summary

Logistic regression is a way to predict probabilities from causes

Linear regression and logistic regression are similar, yet quite different

Unlike linear regression, logistic regression can be used for categorical y-variables

Forecasting and classifying are important applications of logistic regression