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	Categorical Variables	Causes
$p(y)$	y	x

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

Hitting Deadlines

Probability of hitting deadline

p(y)

Deadline: Hit or miss?

y = 1 or 0

Time of starting work

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?

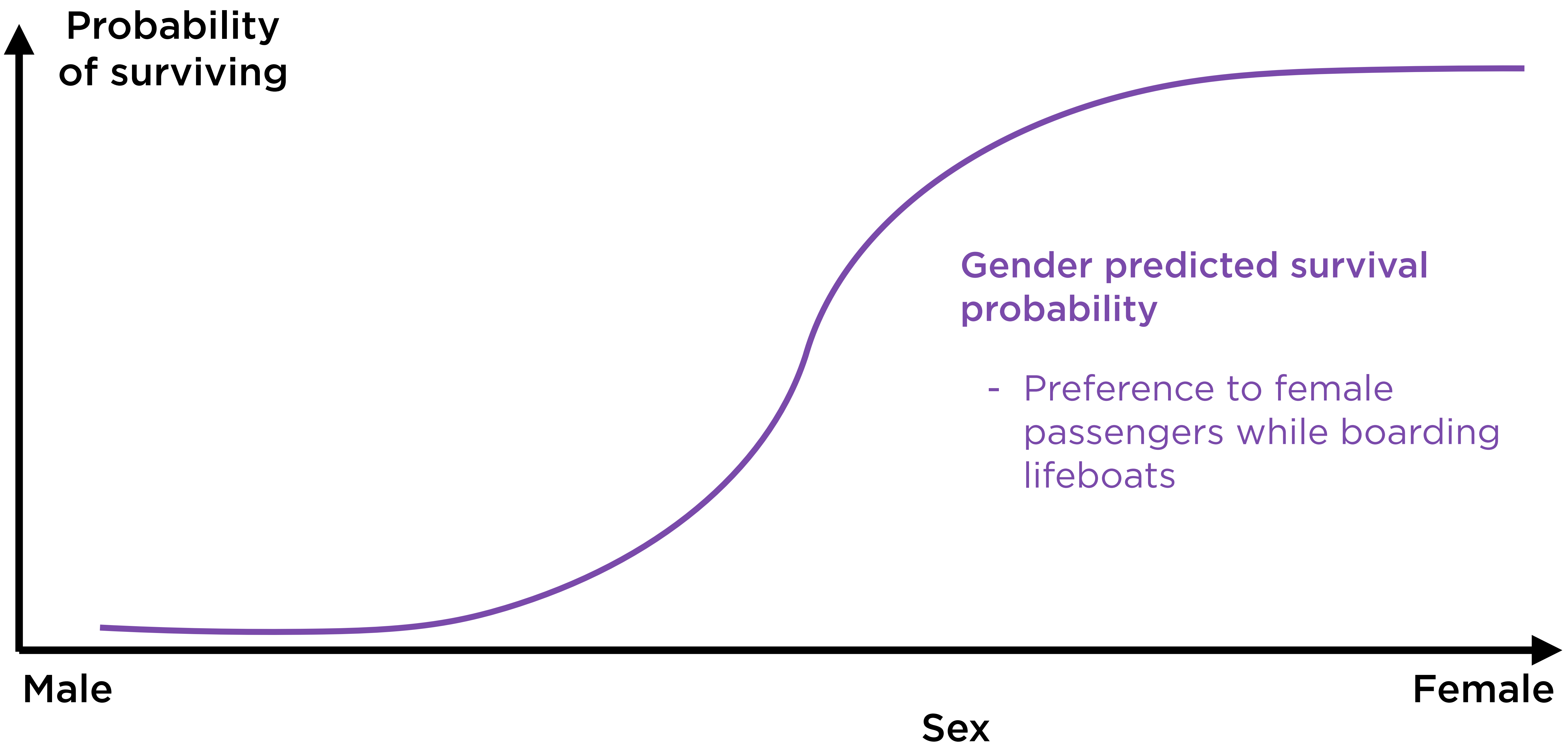
Who Would Survive the Titanic Shipwreck

Sex

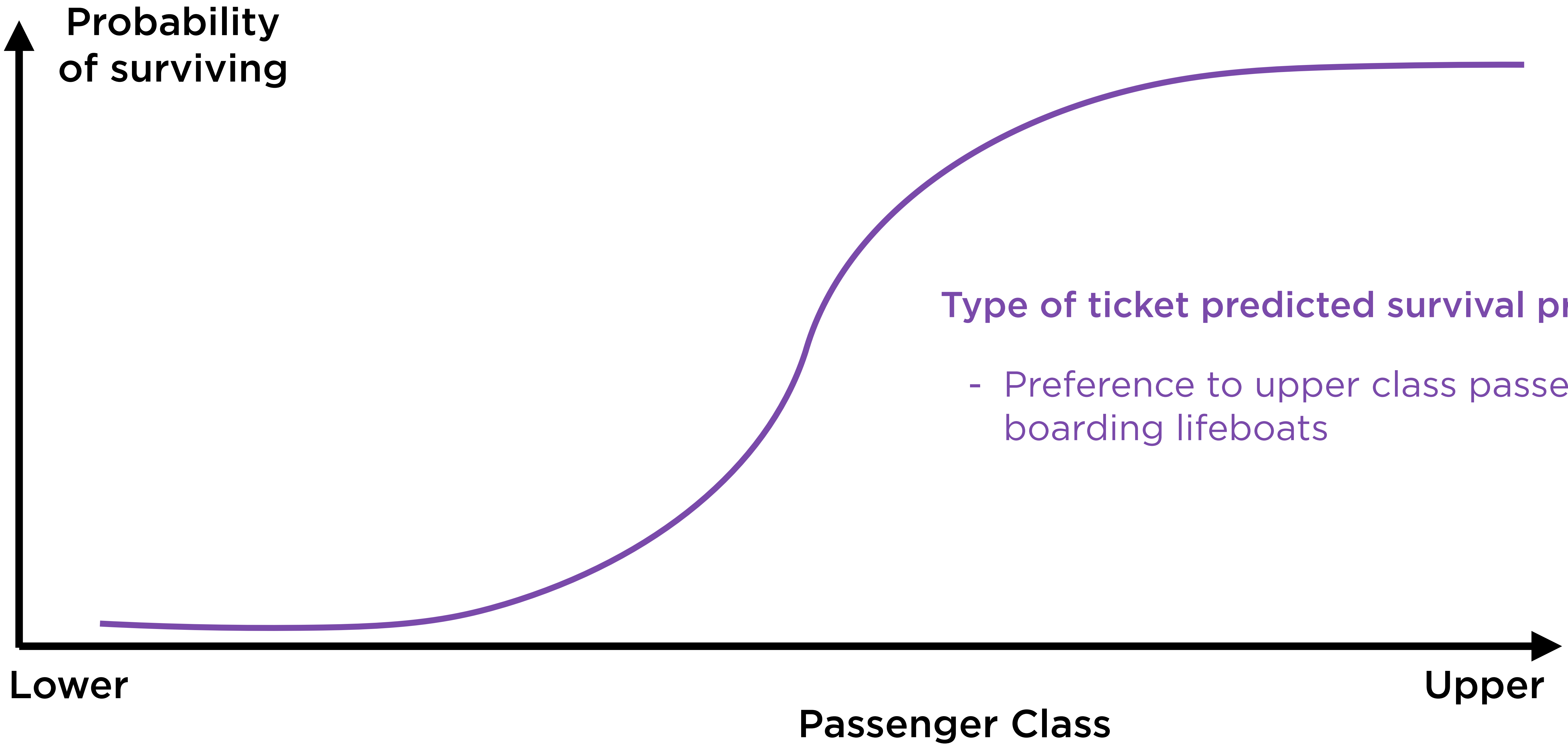
Age

Passenger class

Surviving the Titanic



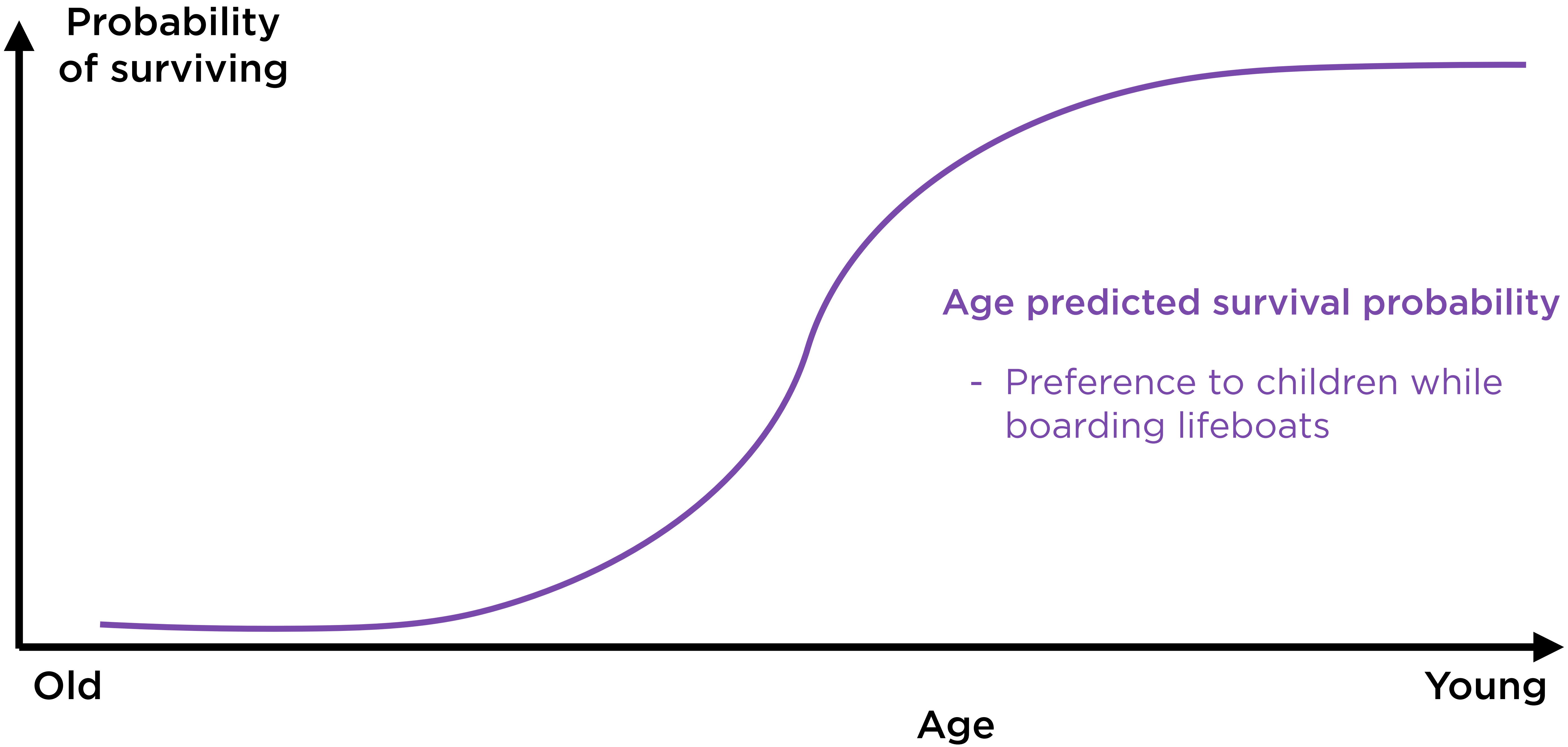
Surviving the Titanic



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

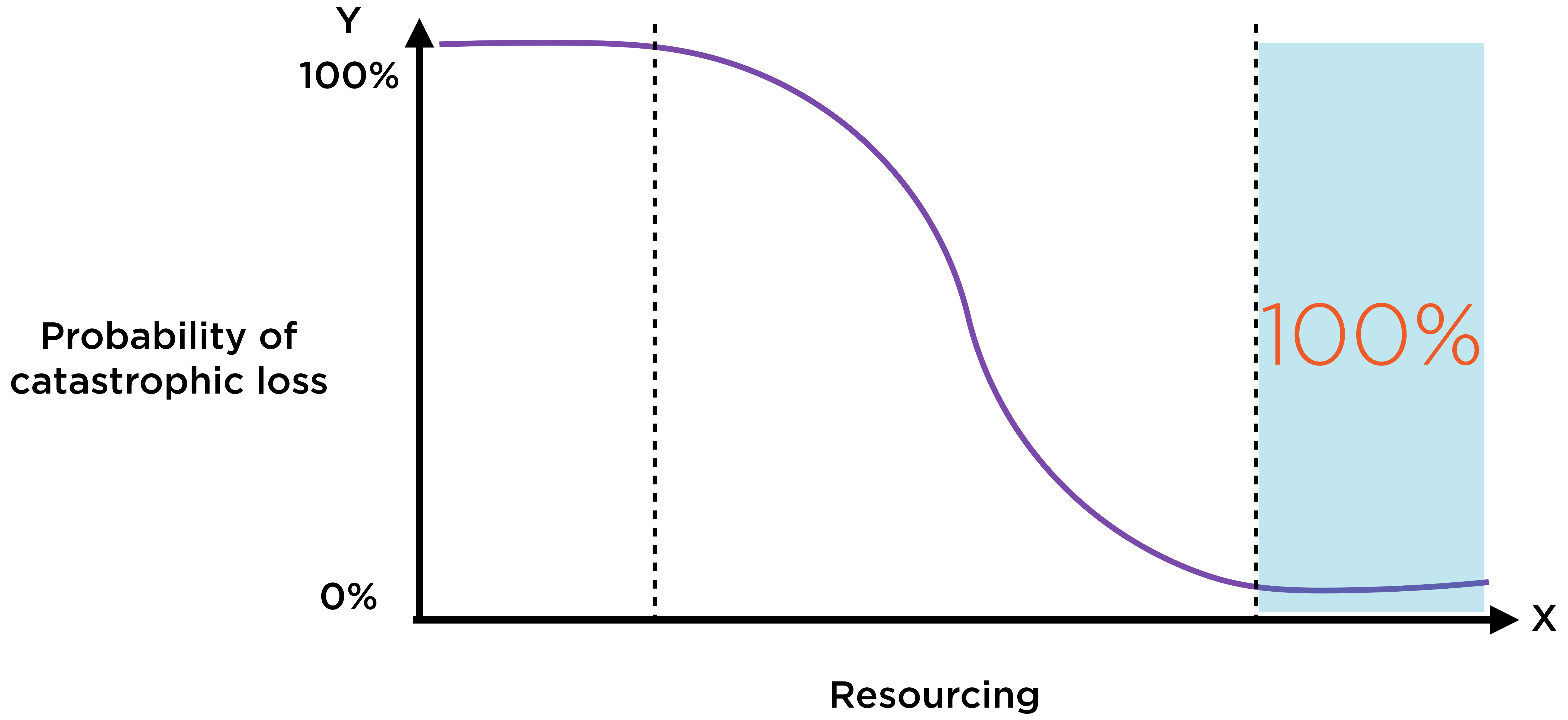
Start as late as possible to be sure to make it

Work hard

Start very early and do little else

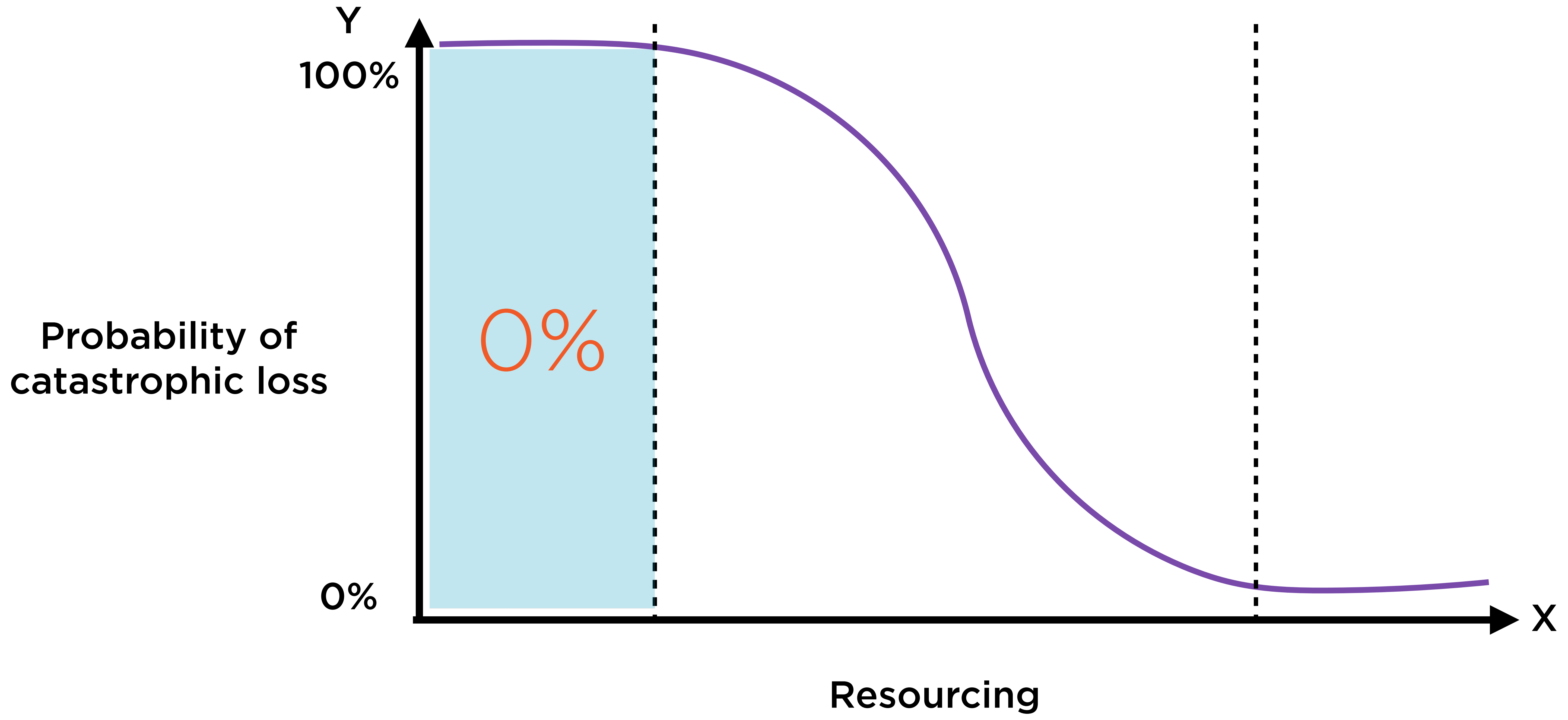
As usual, the middle path is best

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

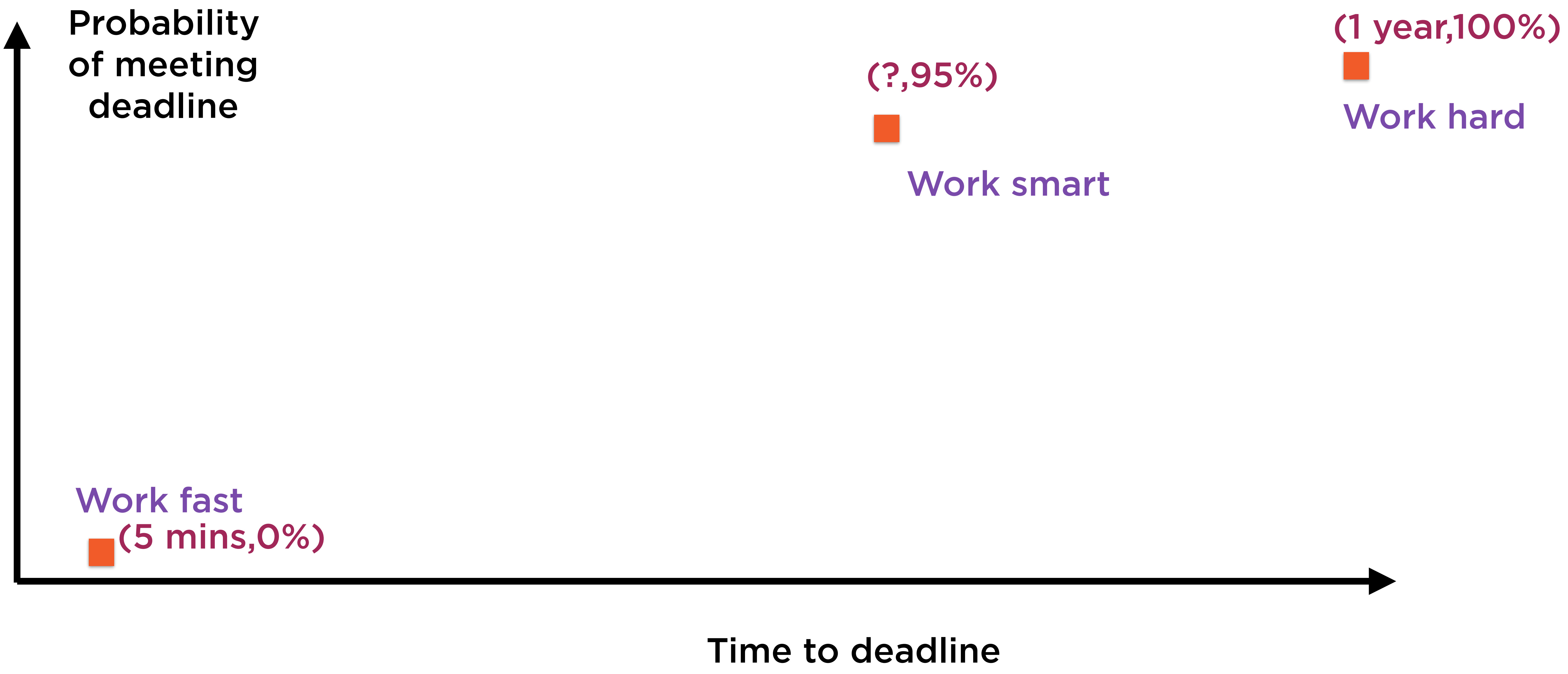
Probability of meeting the deadline

95%

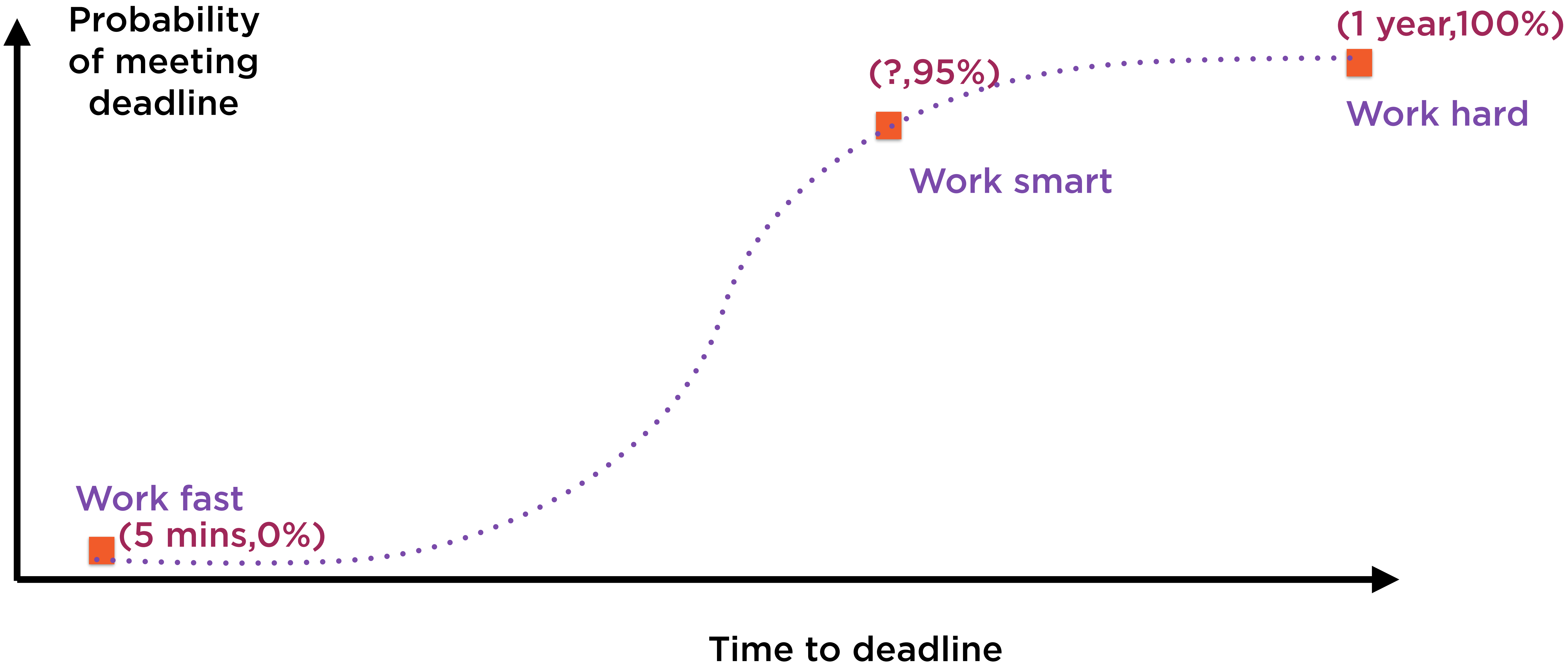
Probability of getting other important work done

95%

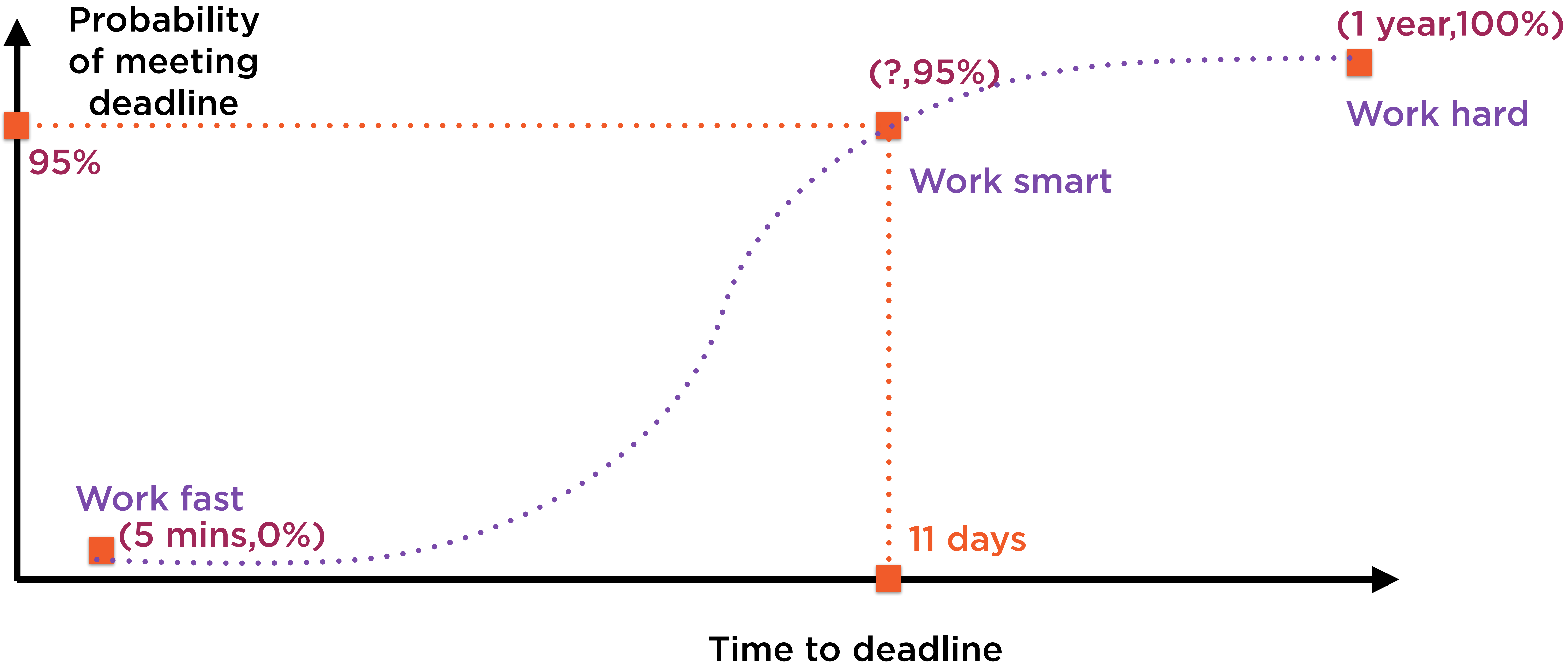
Working Hard, Fast, Smart



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

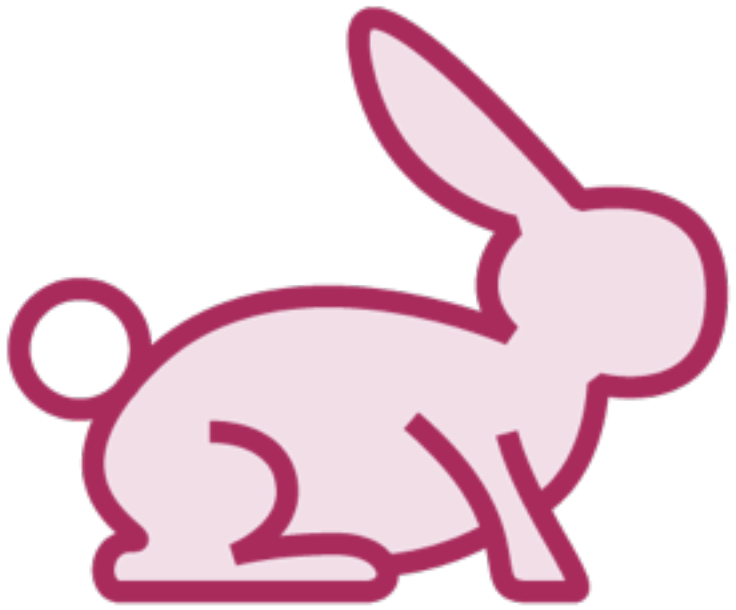
Analyse

Allocate

Predict

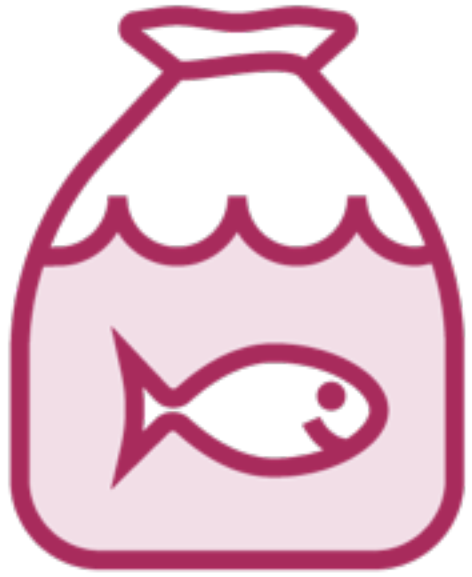
Classify

Whales: Fish or Mammals



Mammal

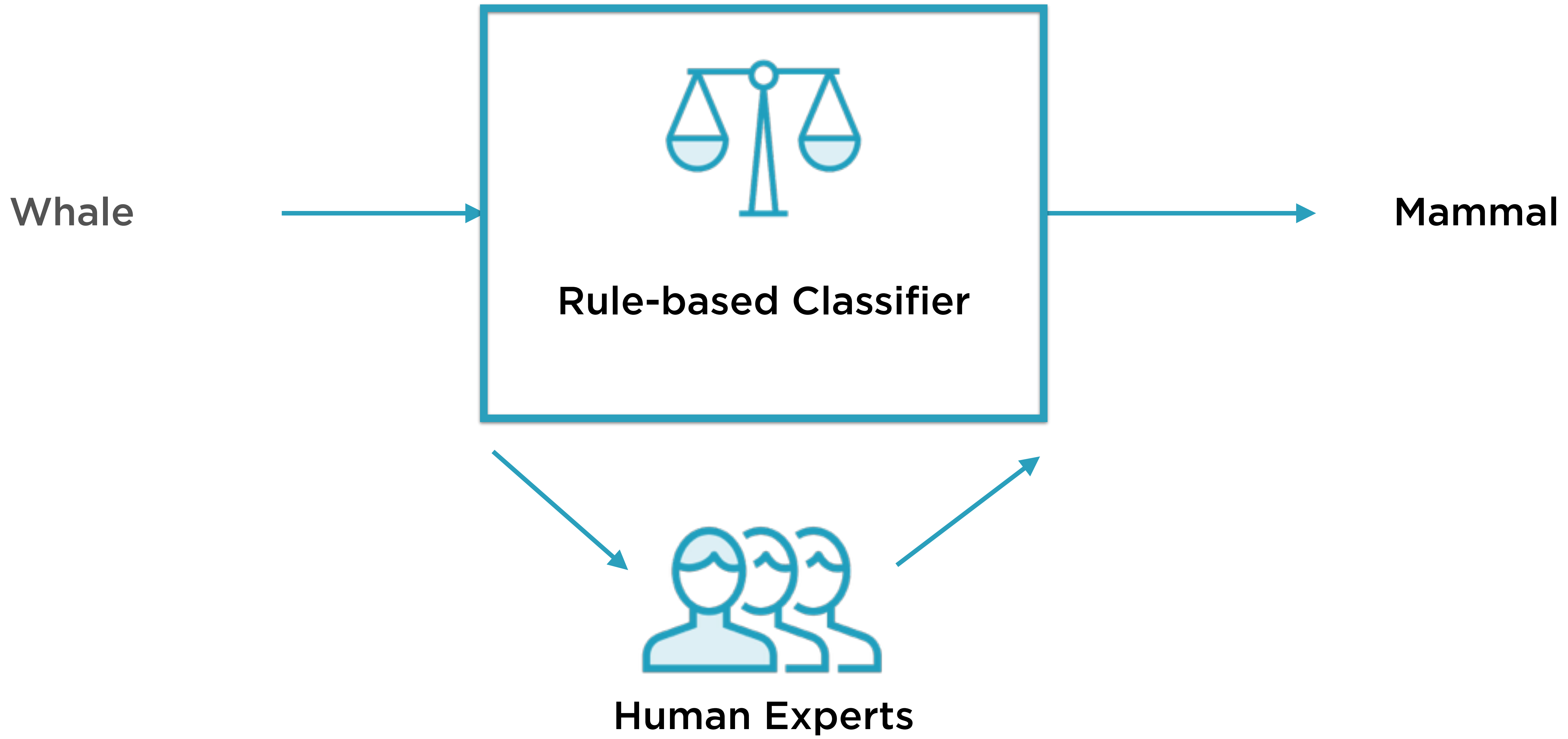
Member of the infraorder *Cetacea*



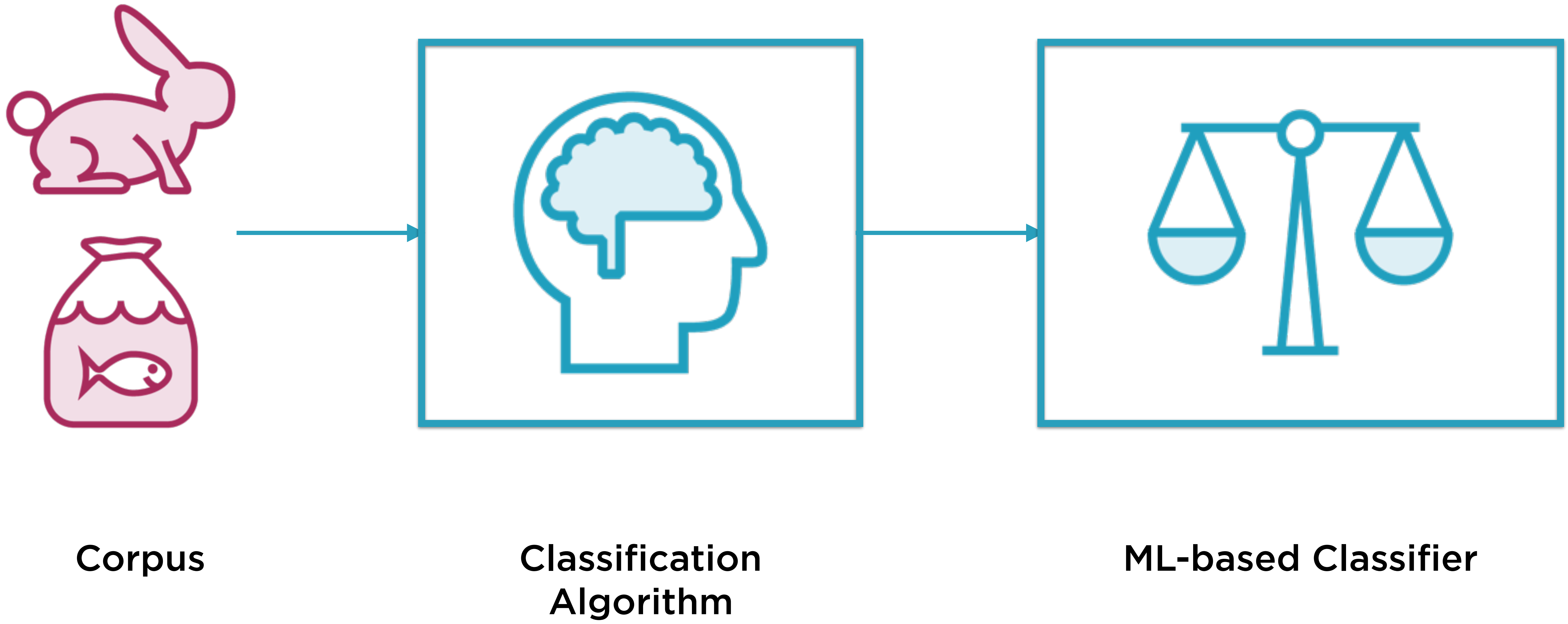
Fish

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

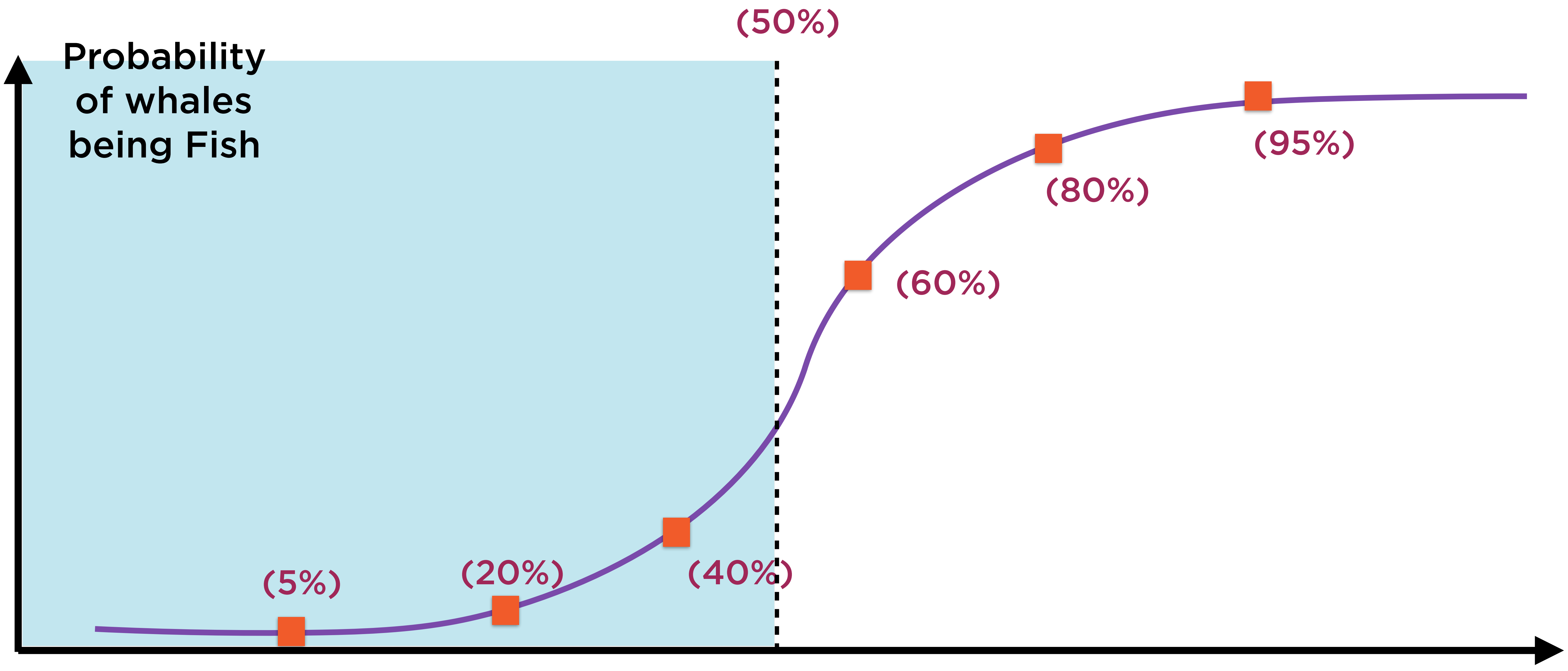
Rule-based Binary Classifier



ML-based Binary Classifier

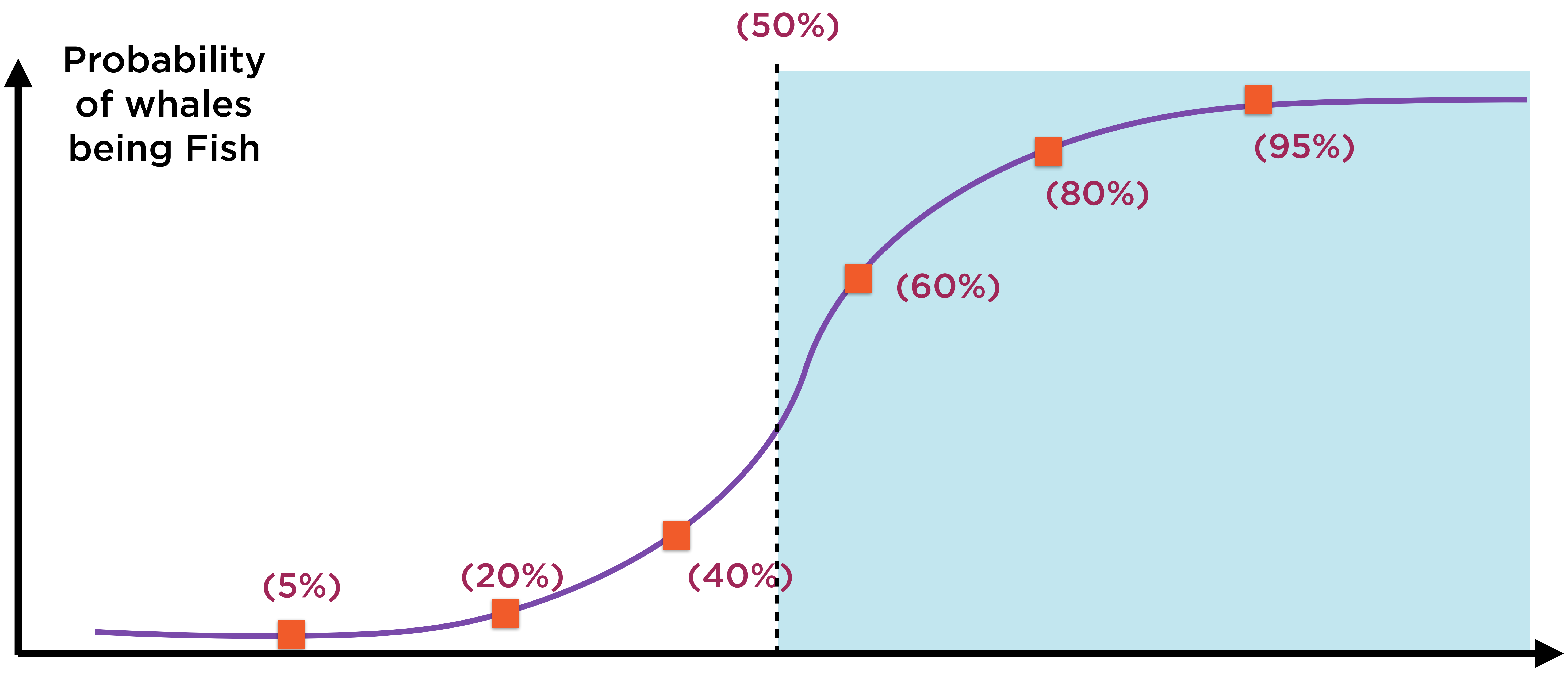


Applying Logistic Regression



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

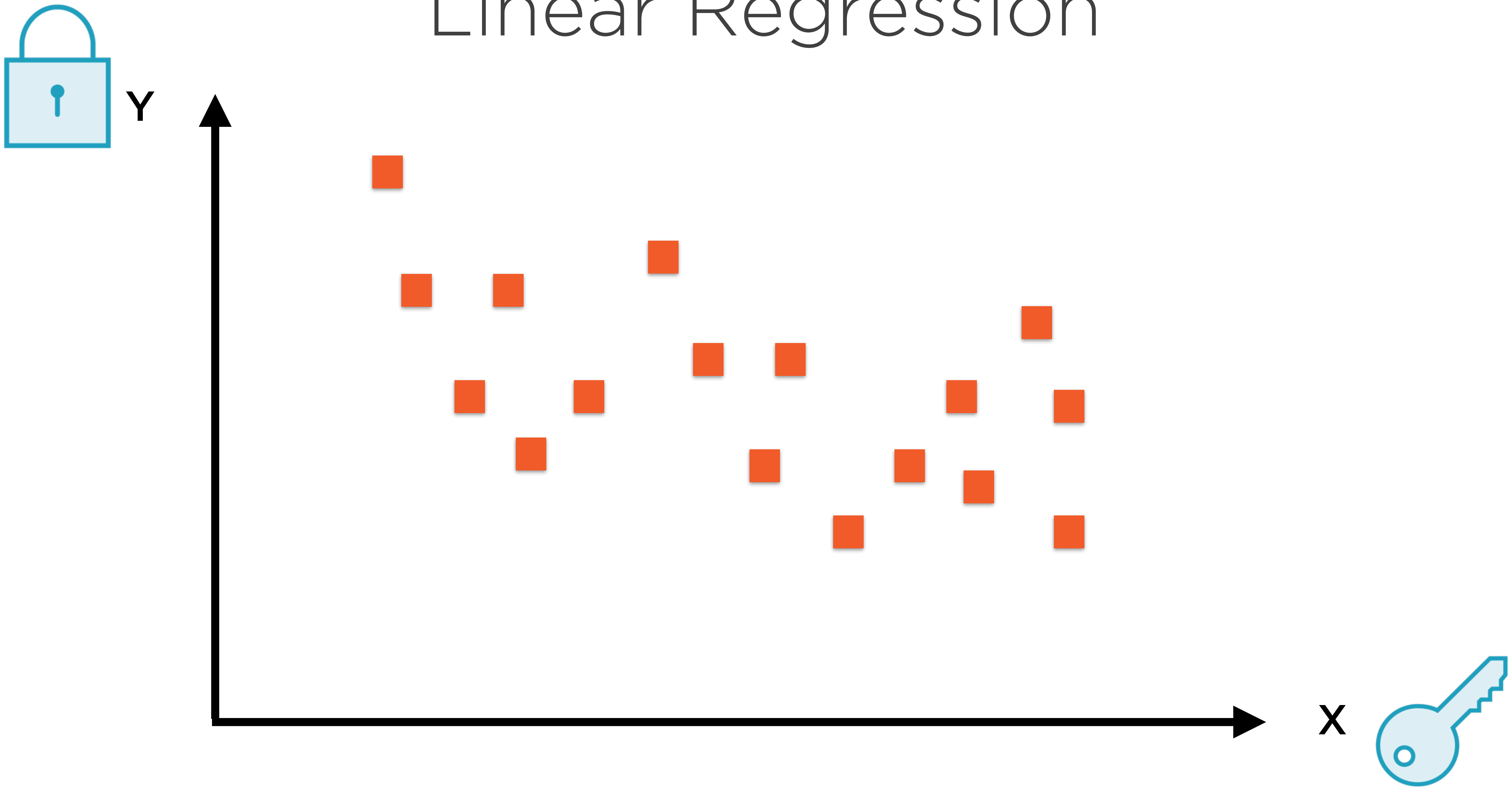
Cause

Explanatory variable

Effect

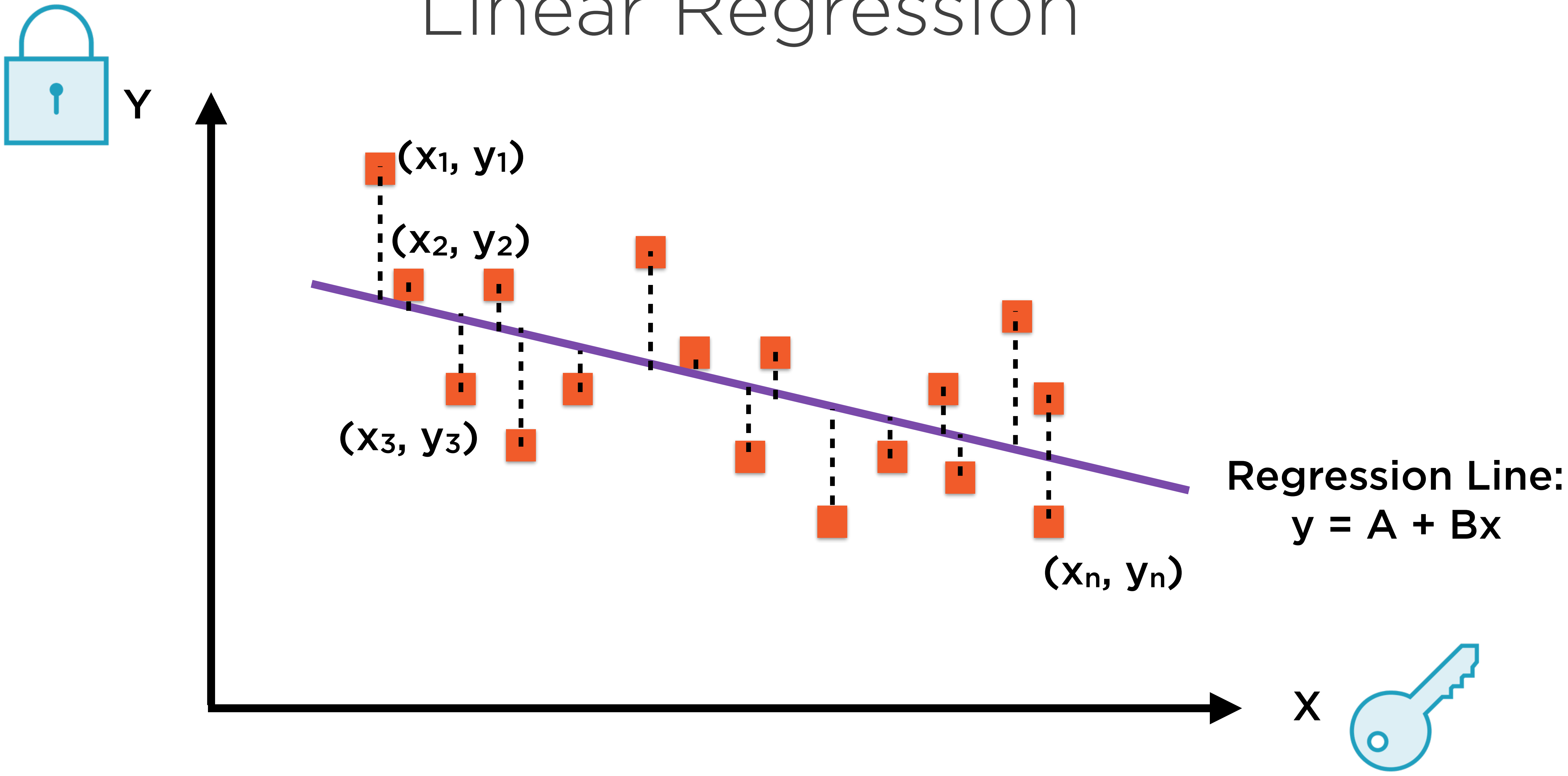
Dependent variable

Linear Regression



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

Linear Regression

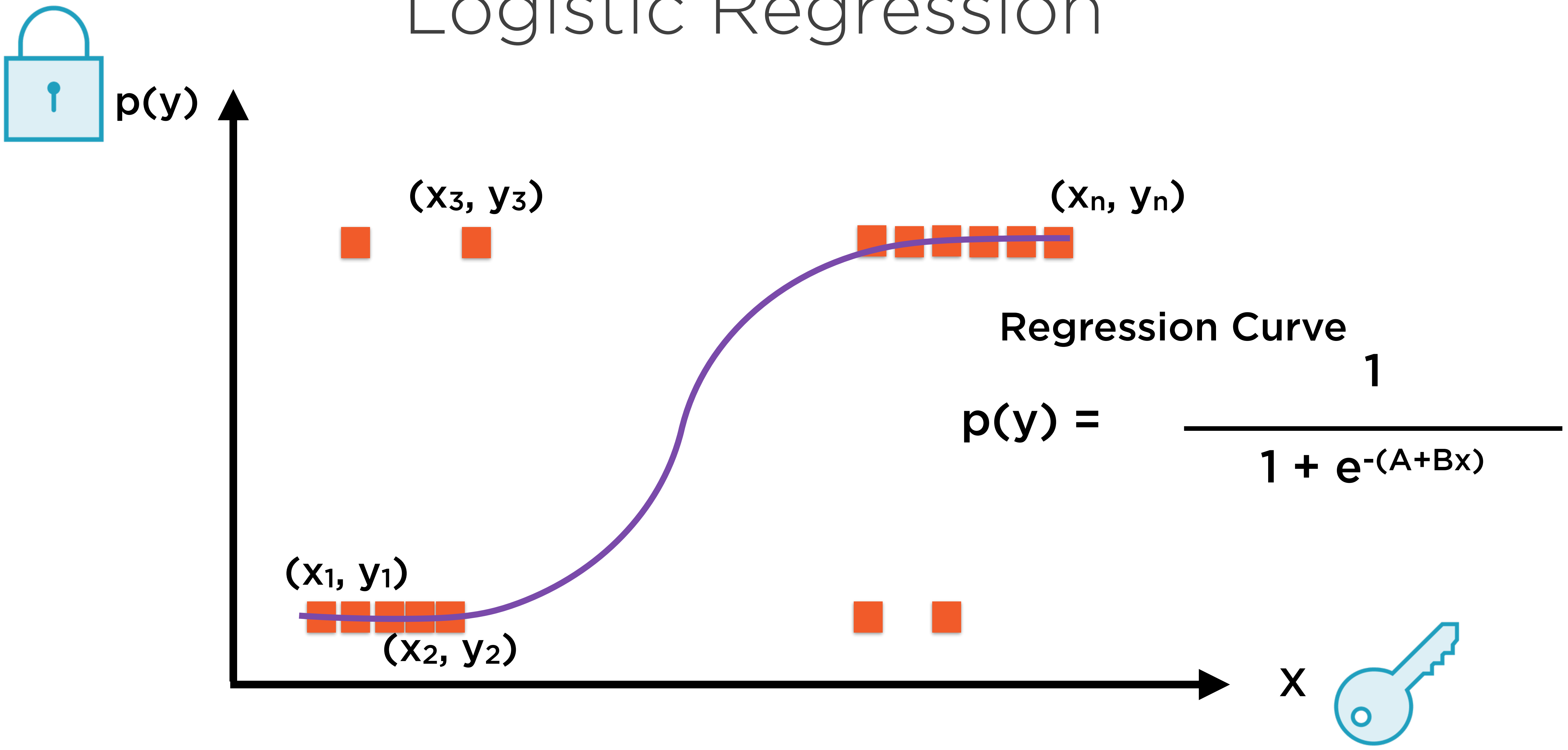


Represent all n points as
(x_i, y_i), where i = 1 to n

Logistic Regression

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

Logistic Regression

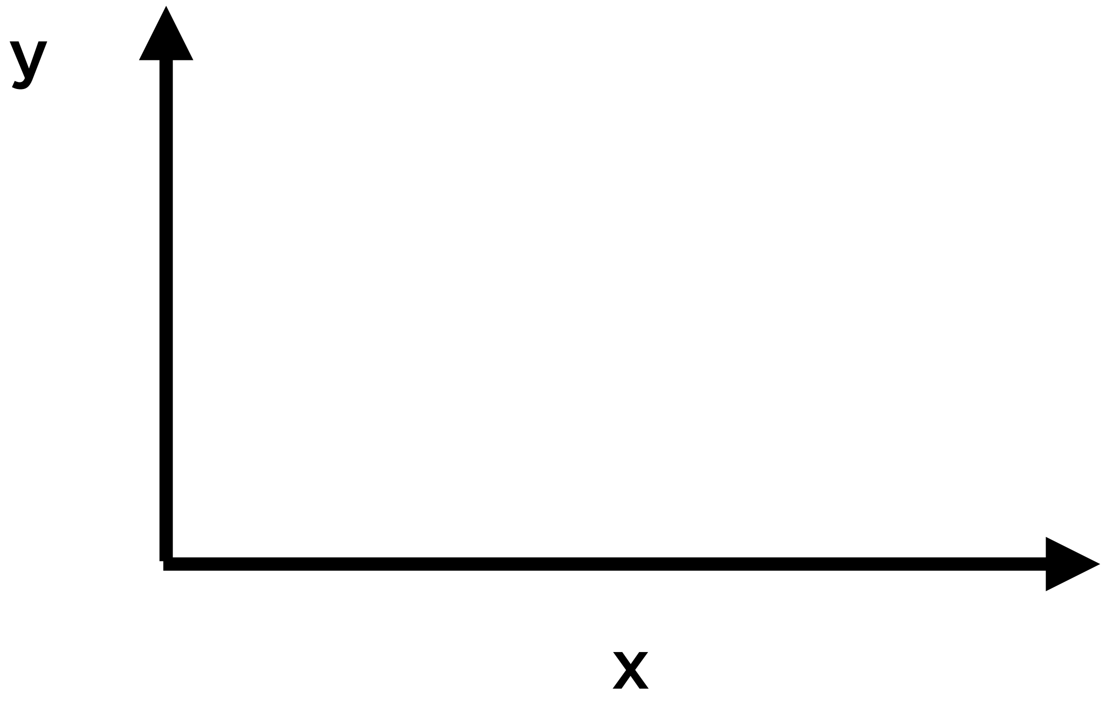


Represent all n points as
(x_i, y_i), where i = 1 to n

Similar, yet Different

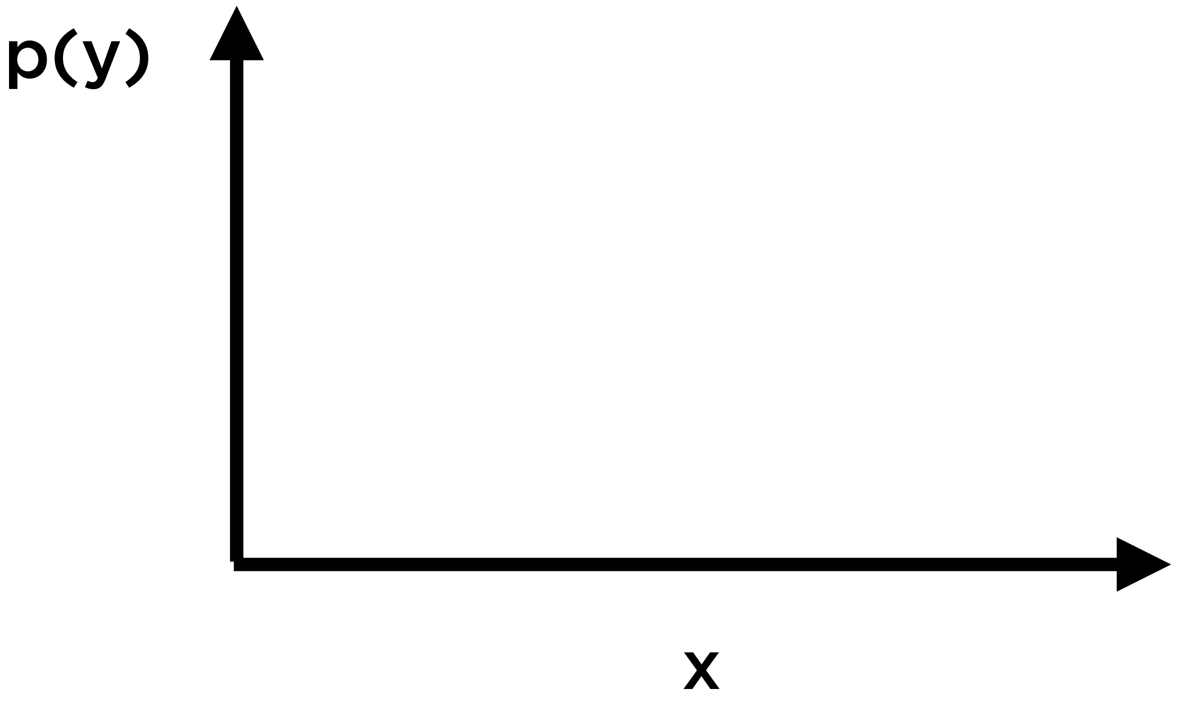
Linear Regression

Given causes, predict effect



Logistic Regression

Given causes, predict probability of effect



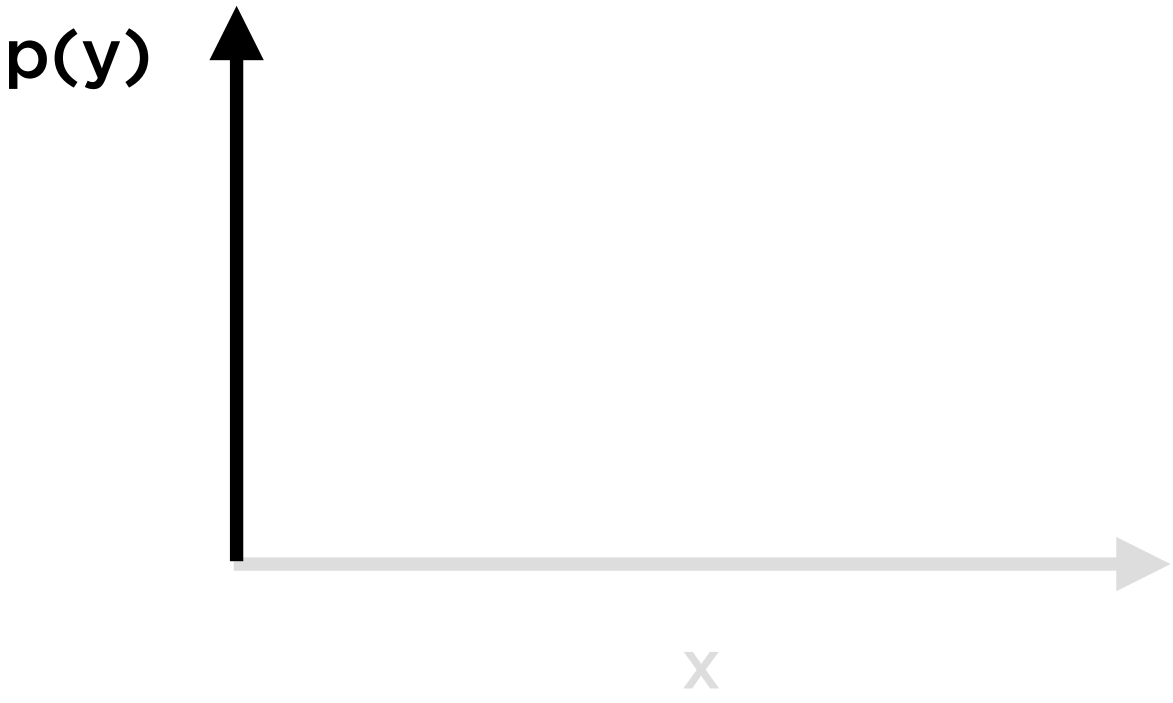
Similar, yet Different

Linear Regression

Effect variable (y) must be continuous

Logistic Regression

Effect variable (y) must be categorical



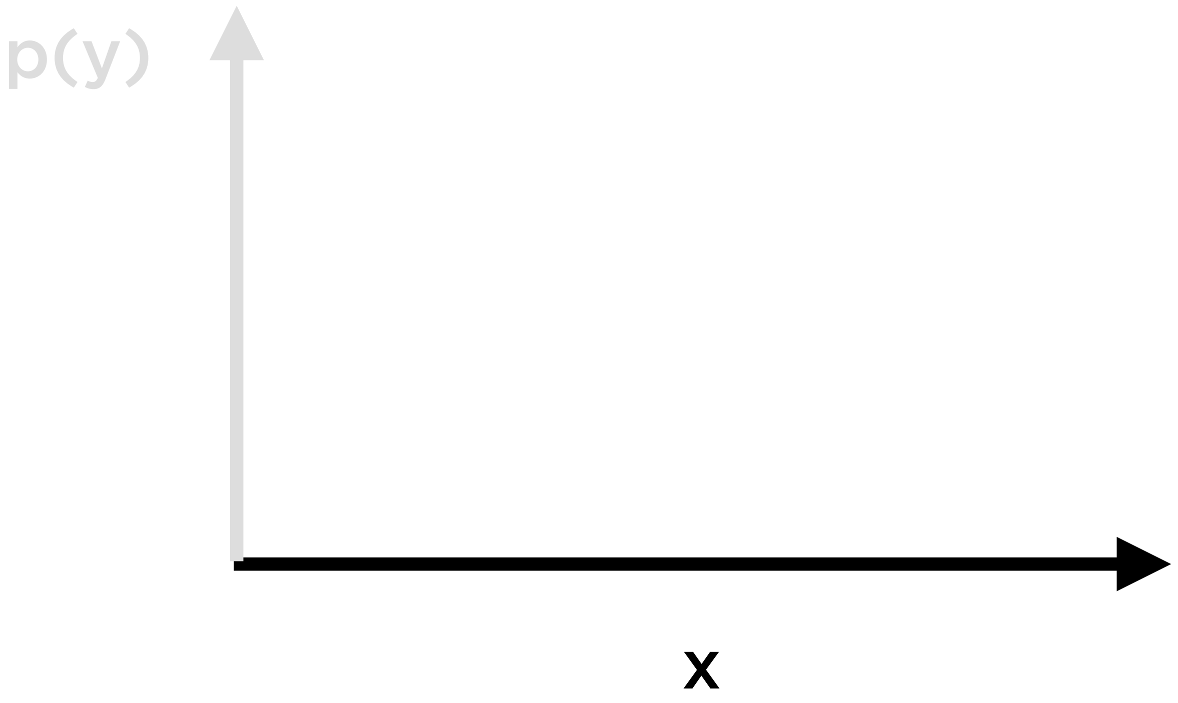
Similar, yet Different

Linear Regression

Cause variables (x) can be continuous or categorical

Logistic Regression

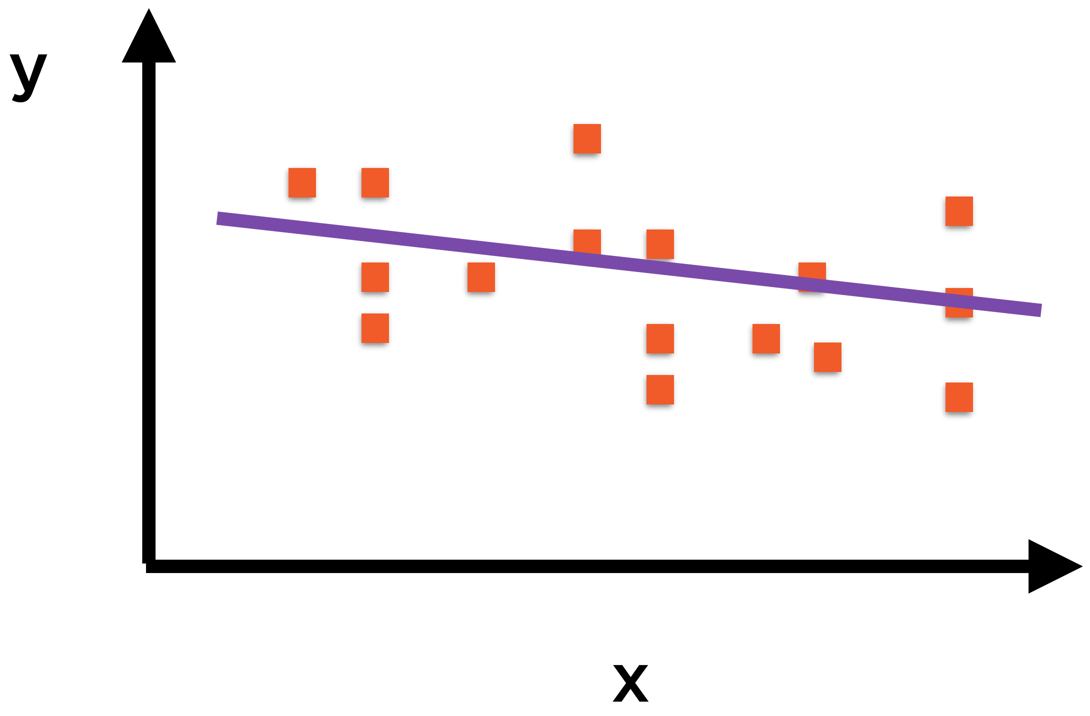
Cause variables (x) can be continuous or categorical



Similar, yet Different

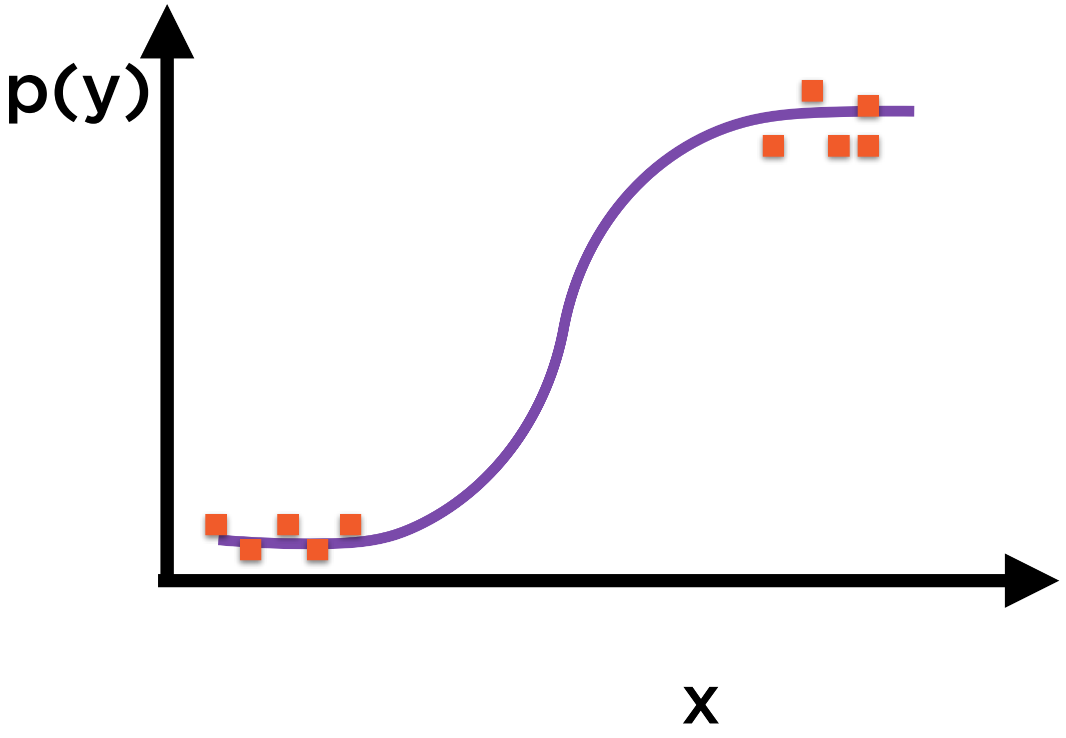
Linear Regression

Connect the dots with a straight line



Logistic Regression

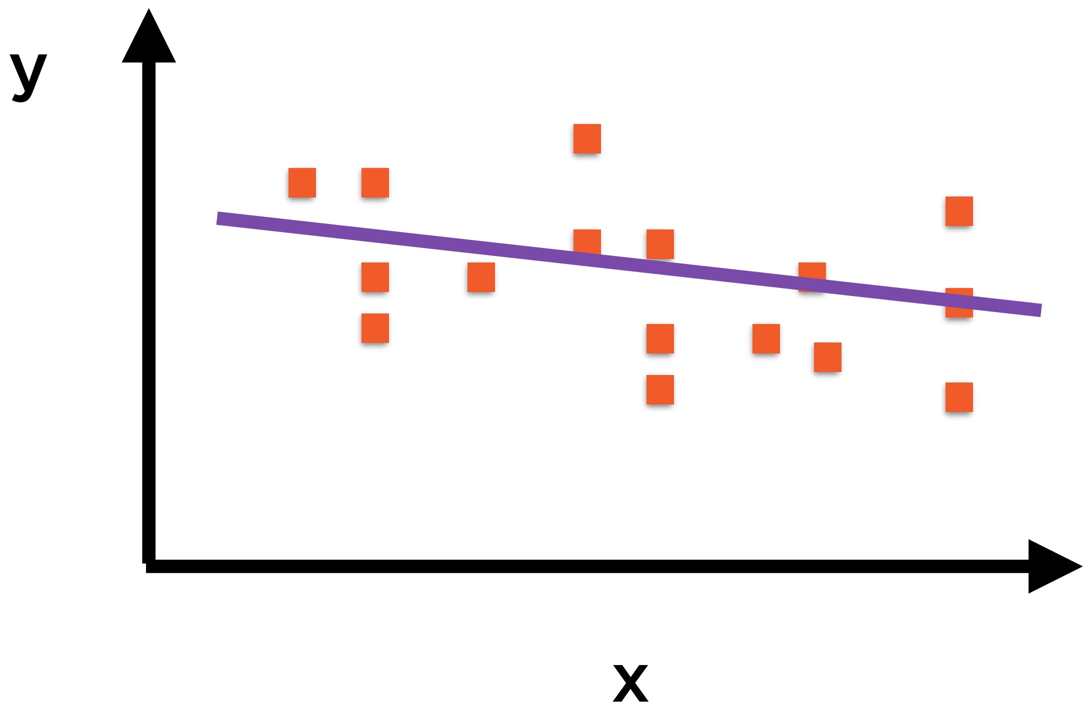
Connect the dots with an S-curve



Similar, yet Different

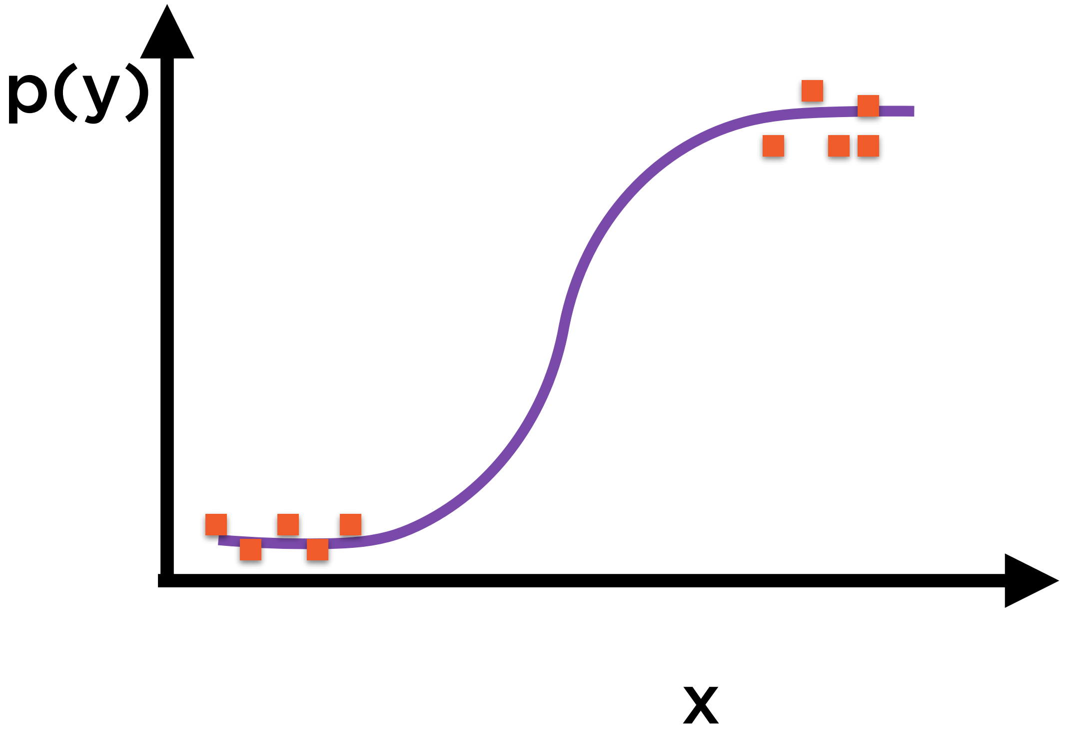
Linear Regression

$$y_i = A + Bx_i$$



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

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

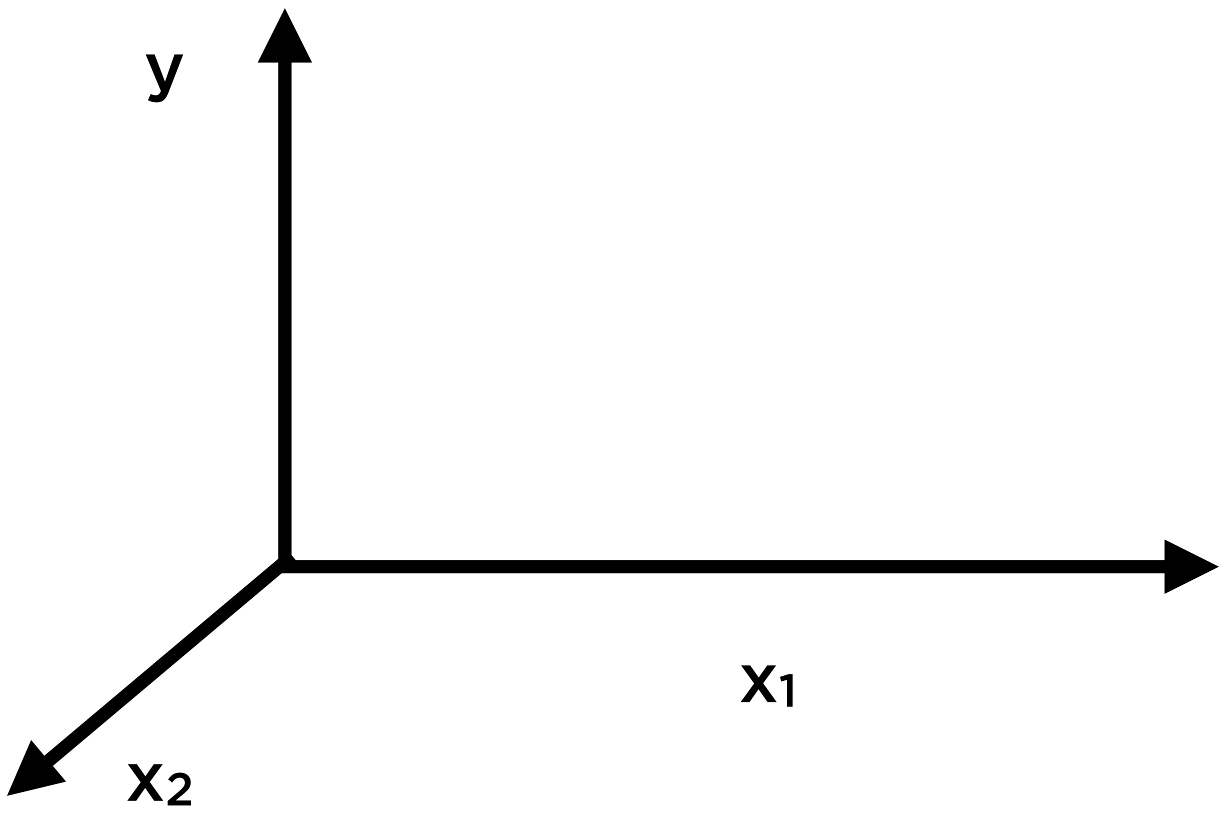
$$\text{logit}(p) = \ln\left(\frac{p}{1 - p}\right)$$

Solve regression problem using cookie-cutter solvers

Similar, yet Different

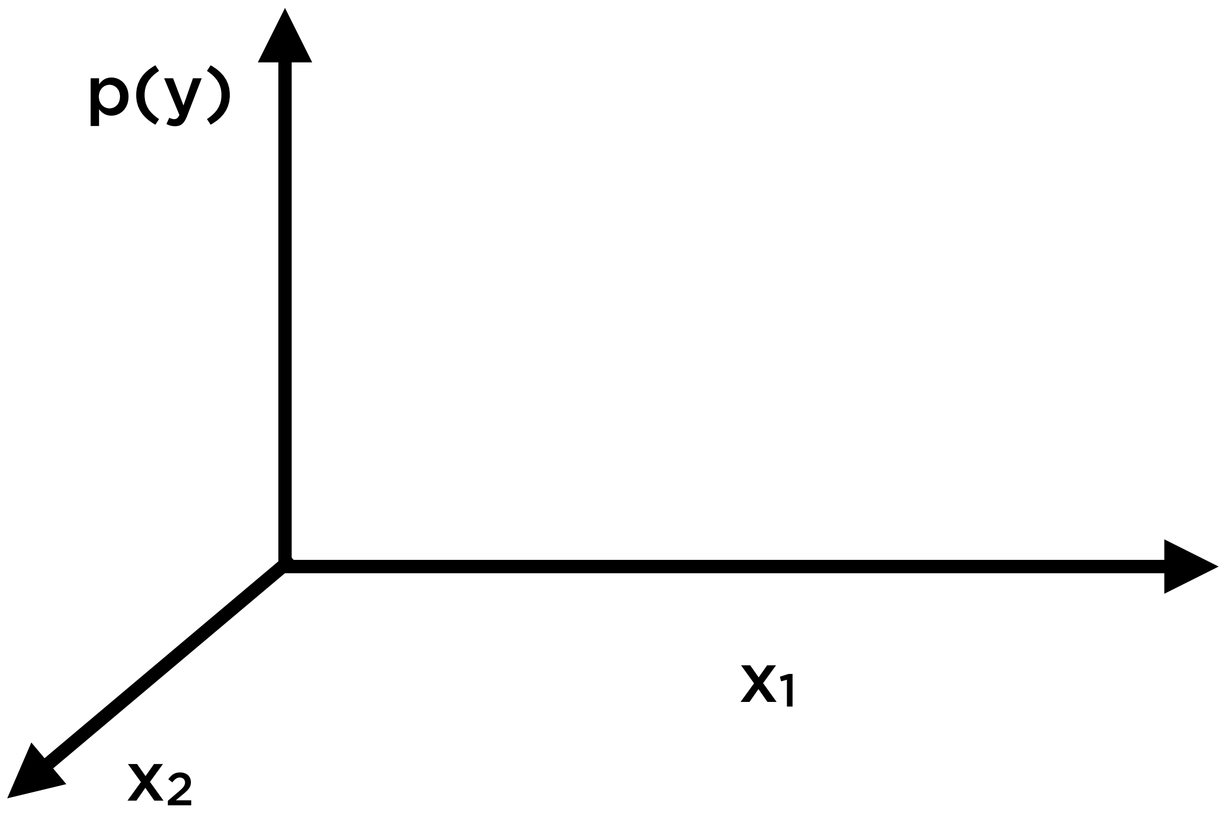
Linear Regression

Easily extended to multiple dimensions



Logistic Regression

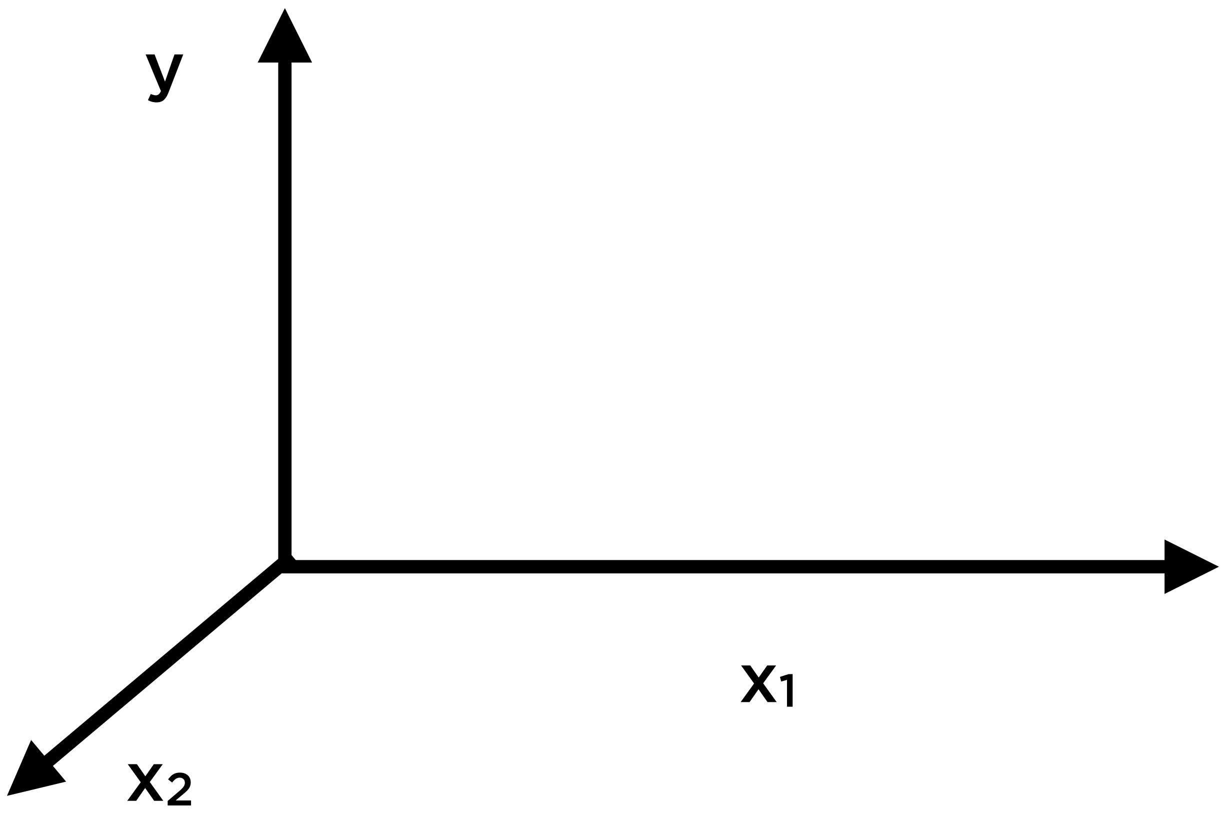
Easily extended to multiple dimensions



Similar, yet Different

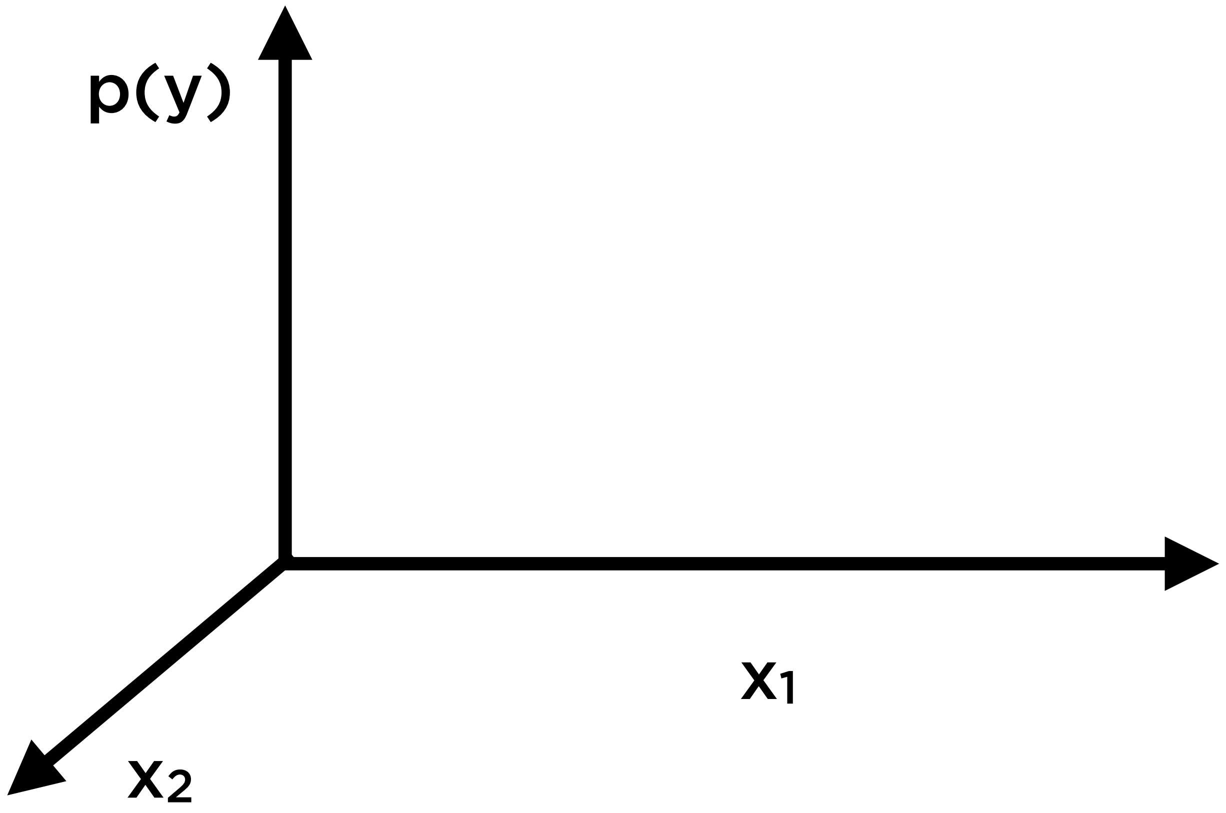
Linear Regression

Easily extended to multiple dimensions



Logistic Regression

Easily extended to multiple dimensions



Connecting the Dots with Regression

Linear Regression Equation:

$$y = A + Bx$$

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

Connecting the Dots with Regression

Linear Regression Equation:

$$y = A + Bx$$

$$y_1 = A + Bx_1 + e_1$$

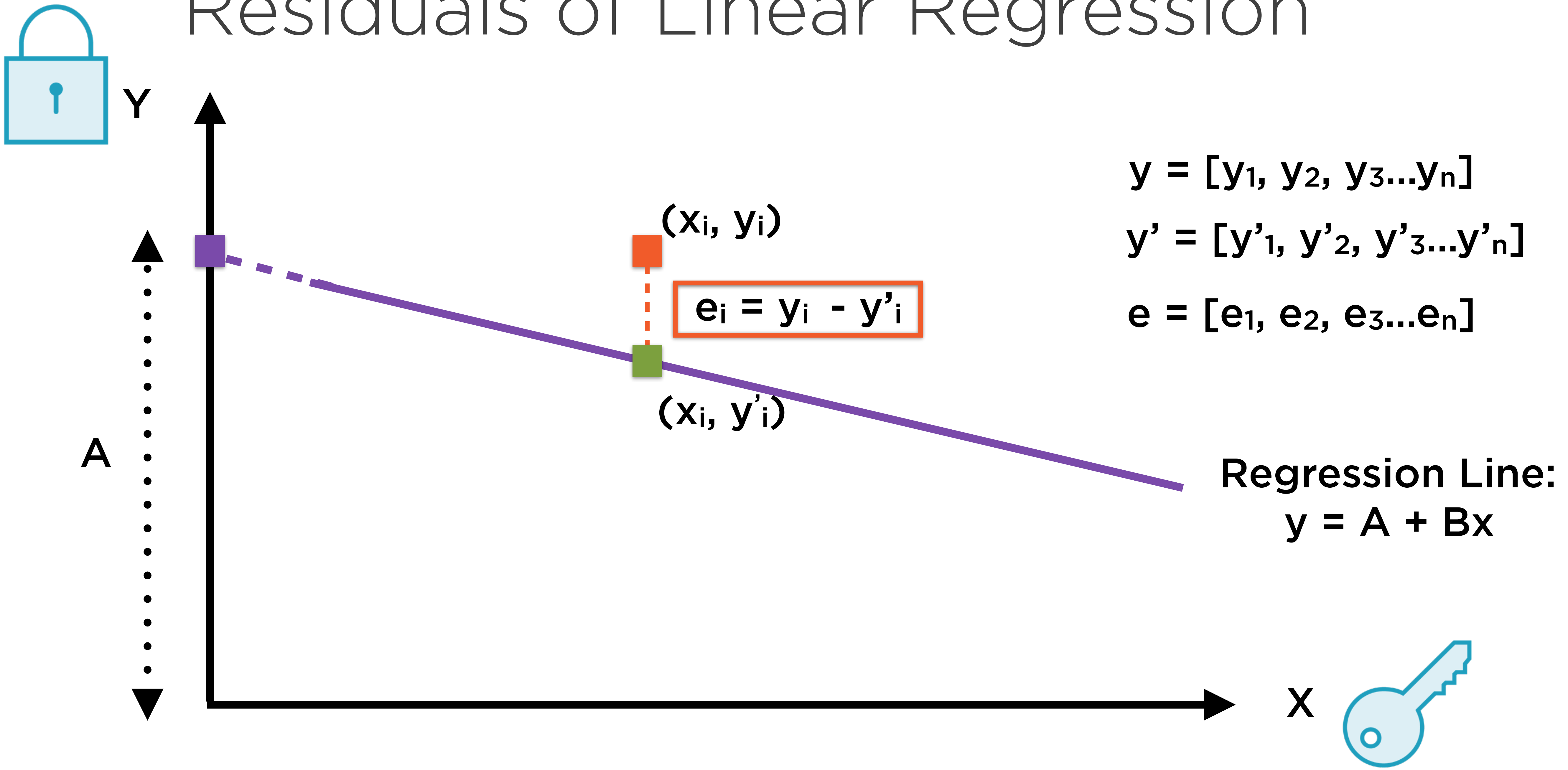
$$y_2 = A + Bx_2 + e_2$$

$$y_3 = A + Bx_3 + e_3$$

$$\ldots \qquad \ldots$$

$$y_n = A + Bx_n + e_n$$

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:

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

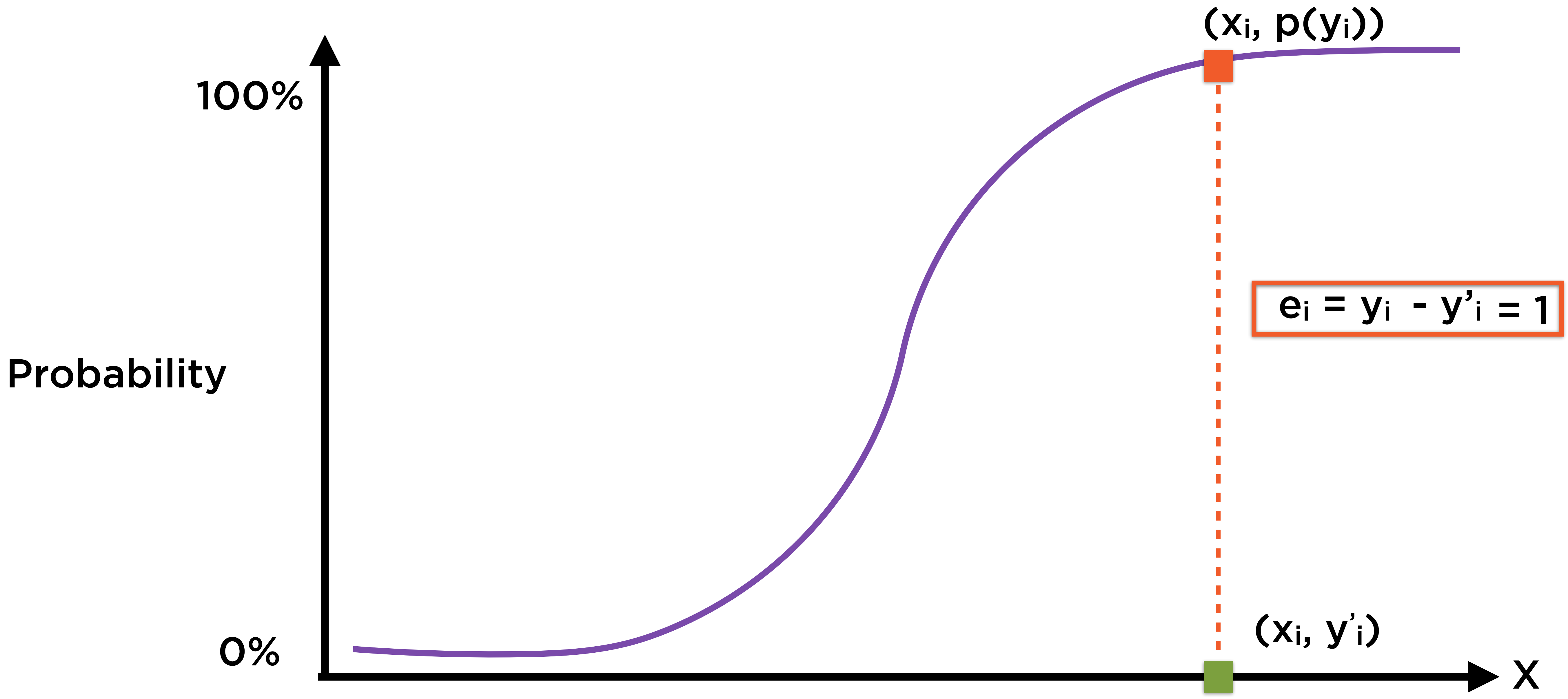
$$p(y_1) = \frac{1}{1 + e^{-(A+Bx_1)}}$$

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

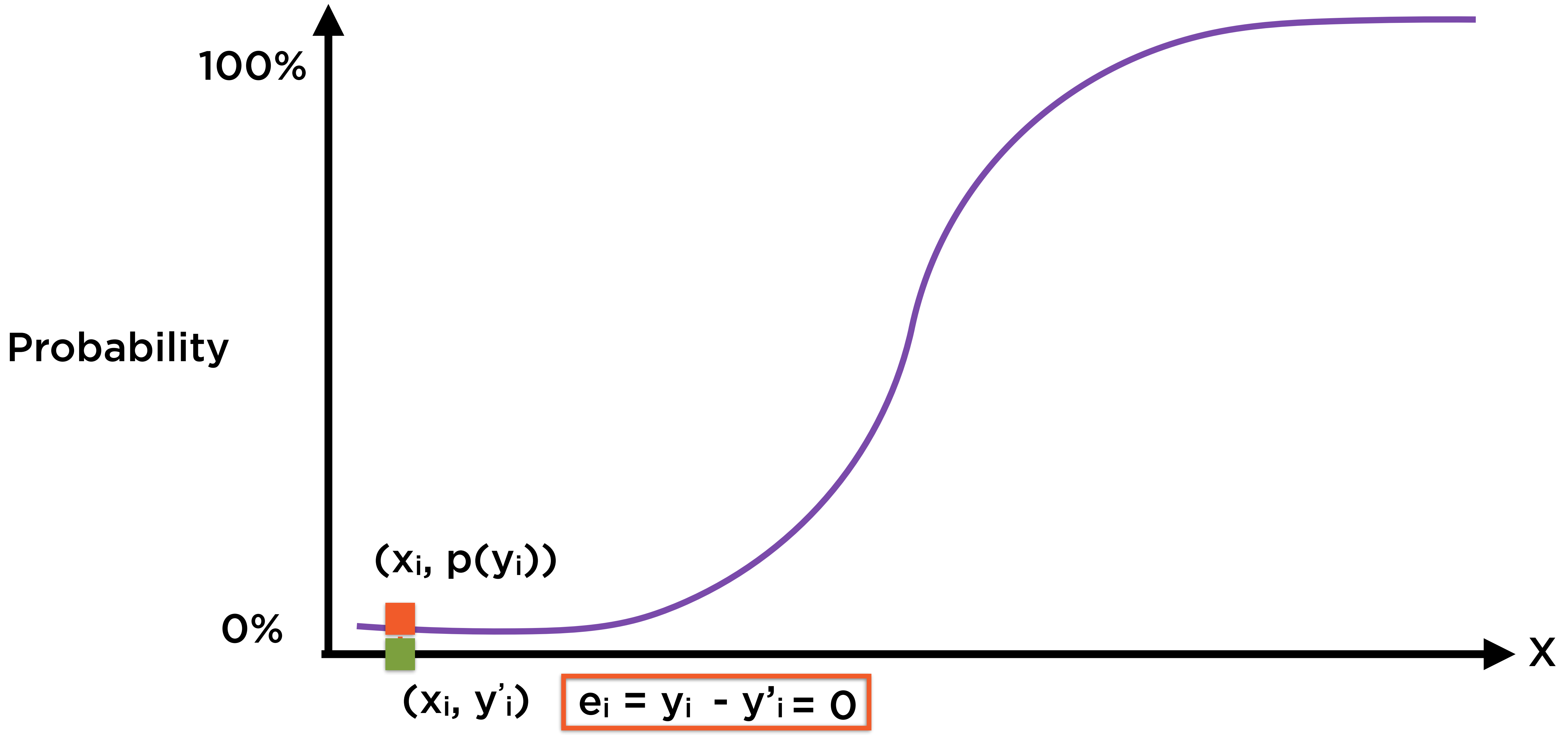
...

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

Residuals of Linear Regression



Residuals of Linear Regression



Similar, yet Different

Linear Regression

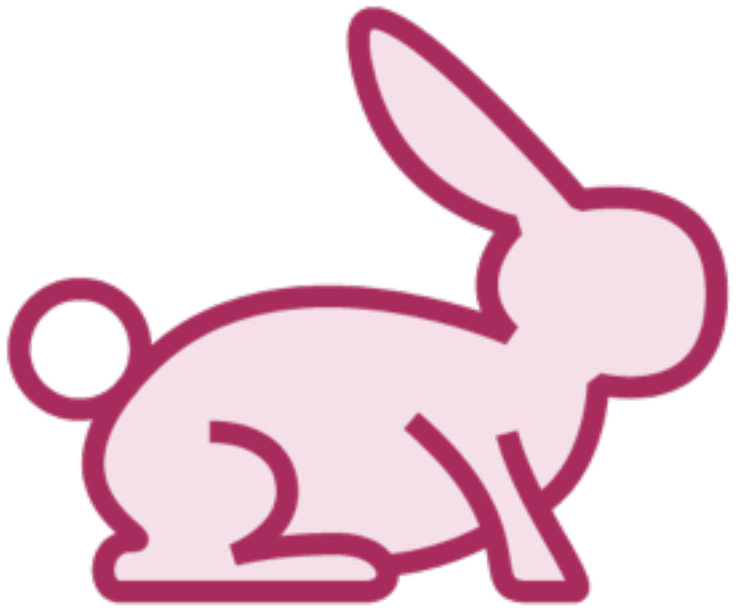
Residuals assumed to be normally distributed

Logistic Regression

Residuals cannot be normally distributed

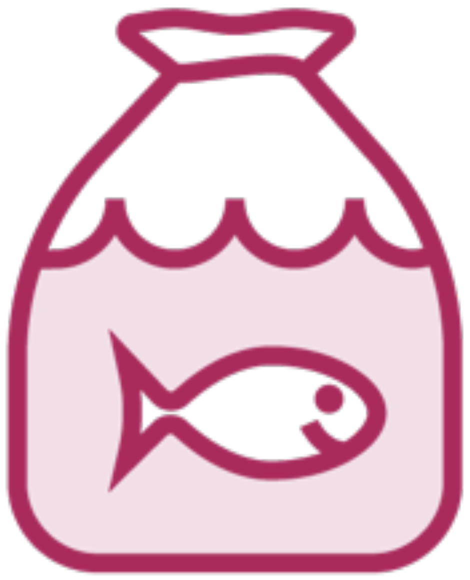
Logistic Regression and Machine Learning

Whales: Fish or Mammals



Mammal

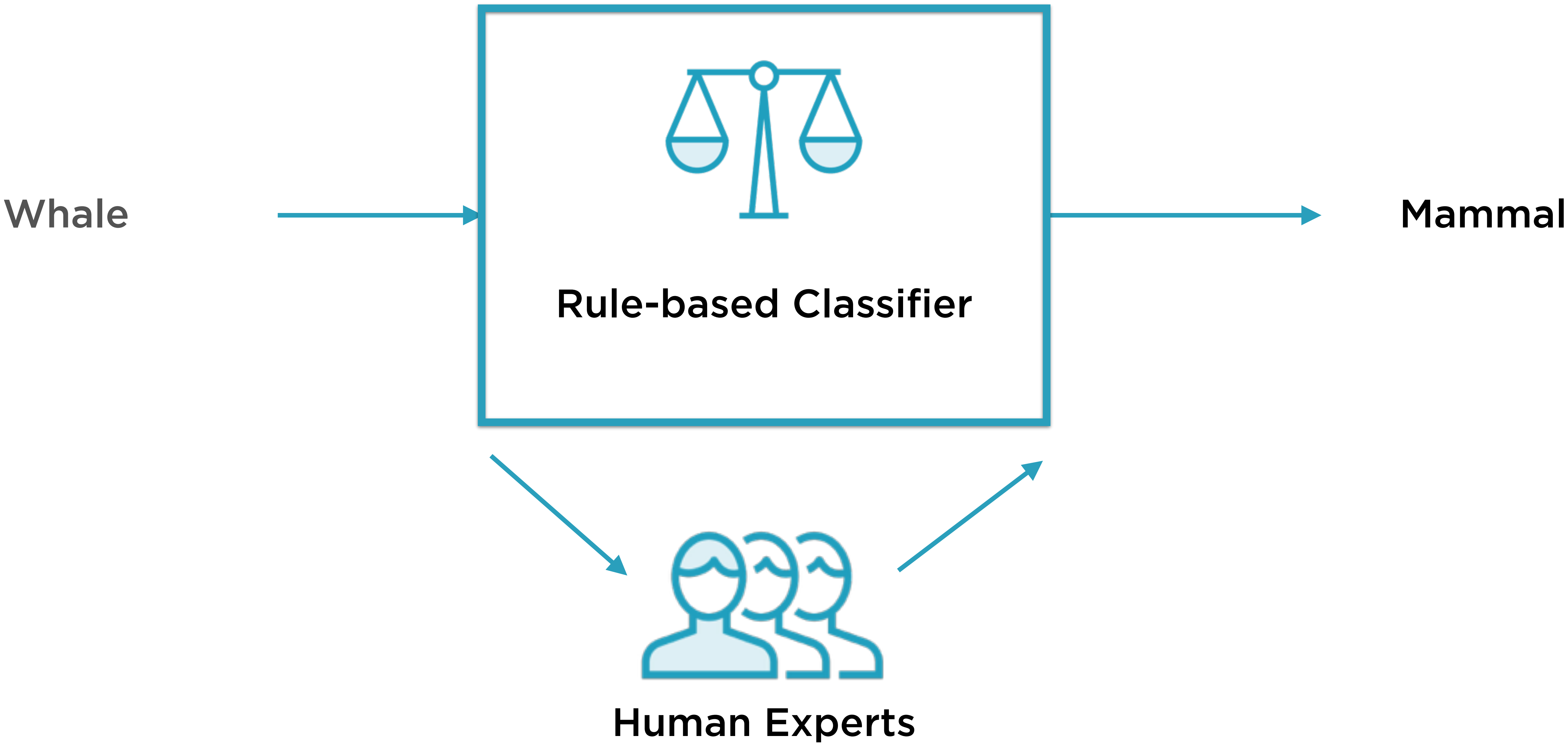
Member of the infraorder *Cetacea*



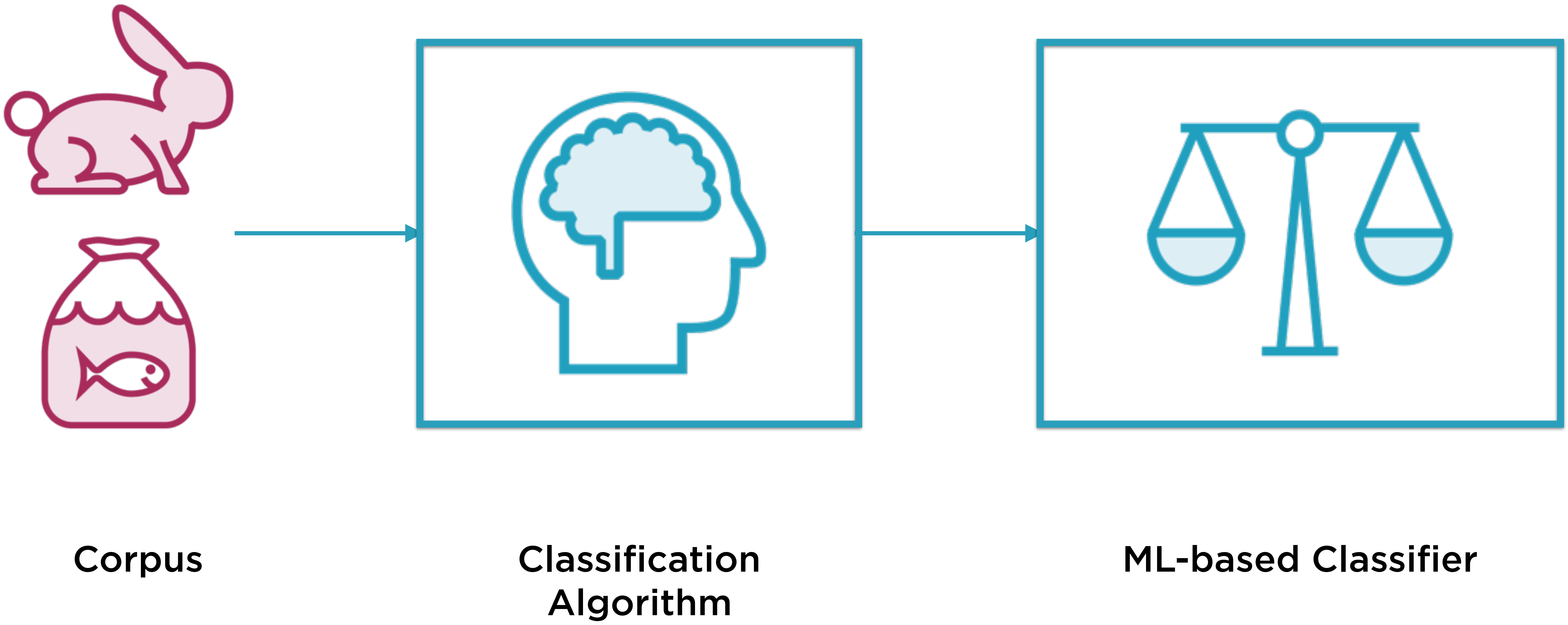
Fish

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

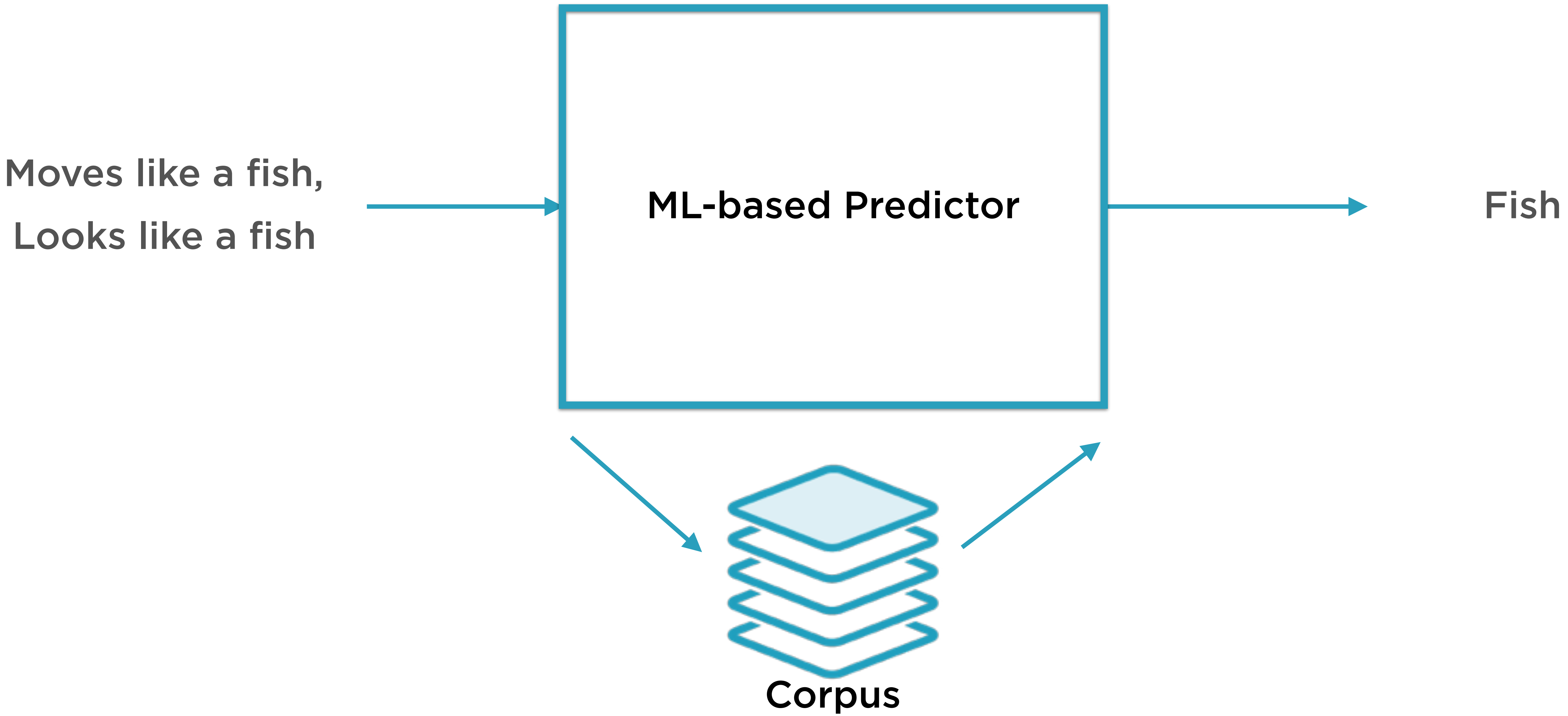
Rule-based Binary Classifier



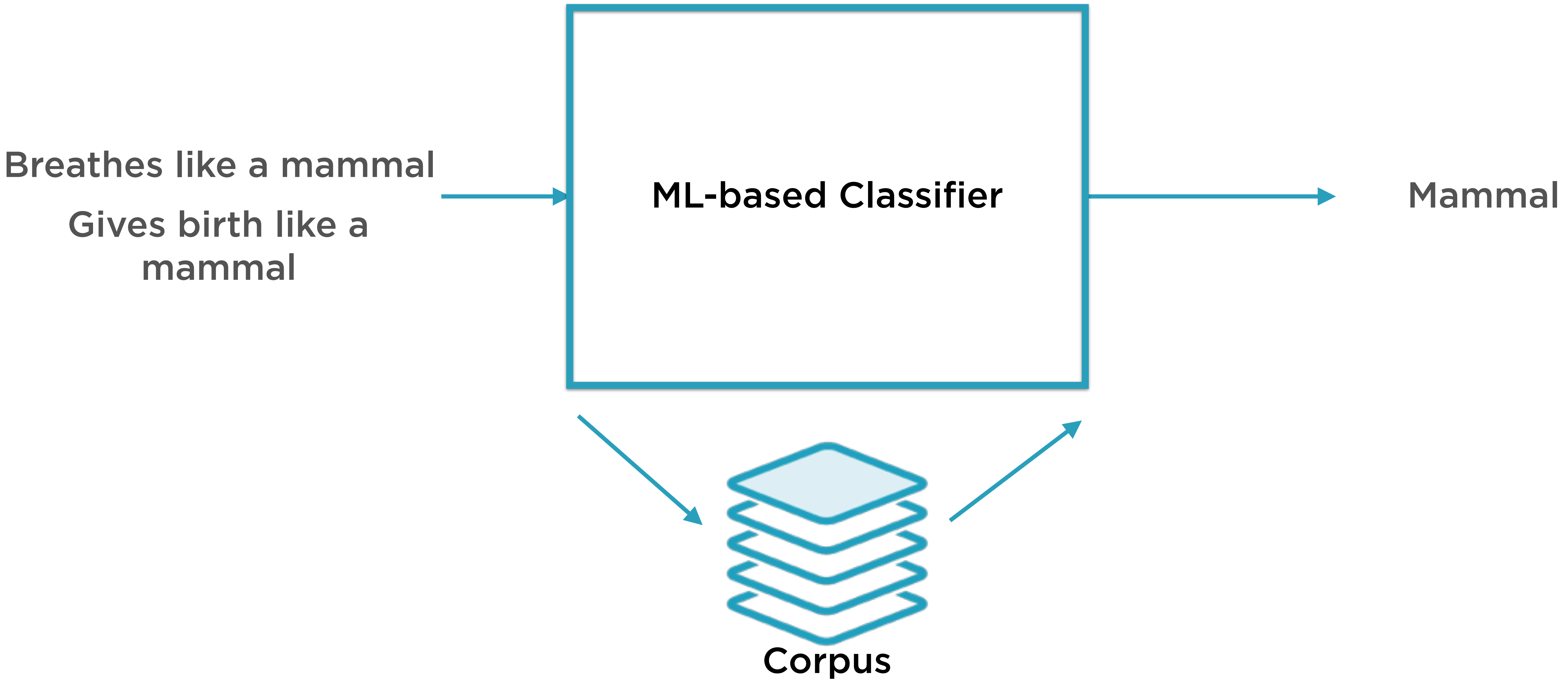
ML-based Binary Classifier



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

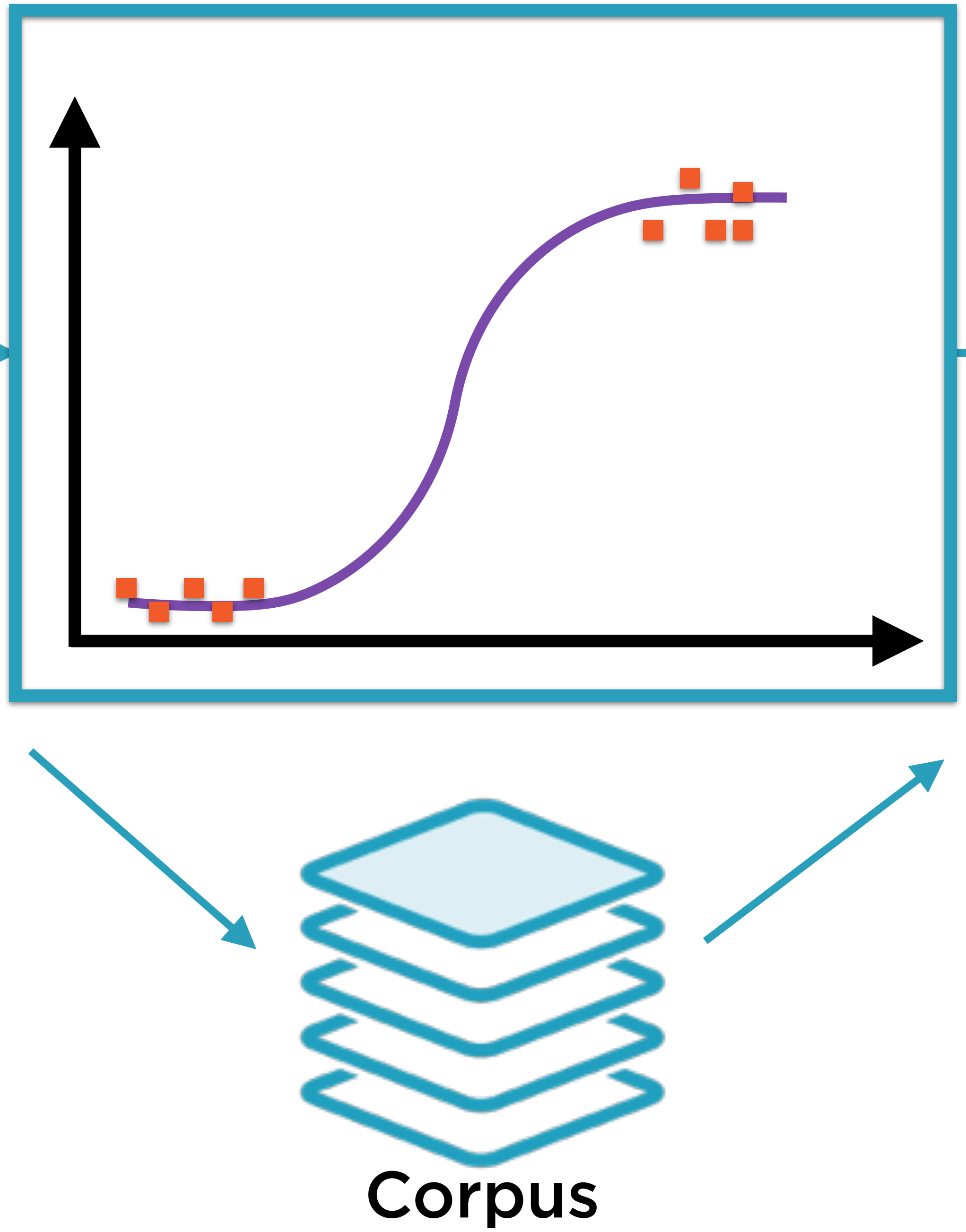
Logistic Regression

ML-based Predictor

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

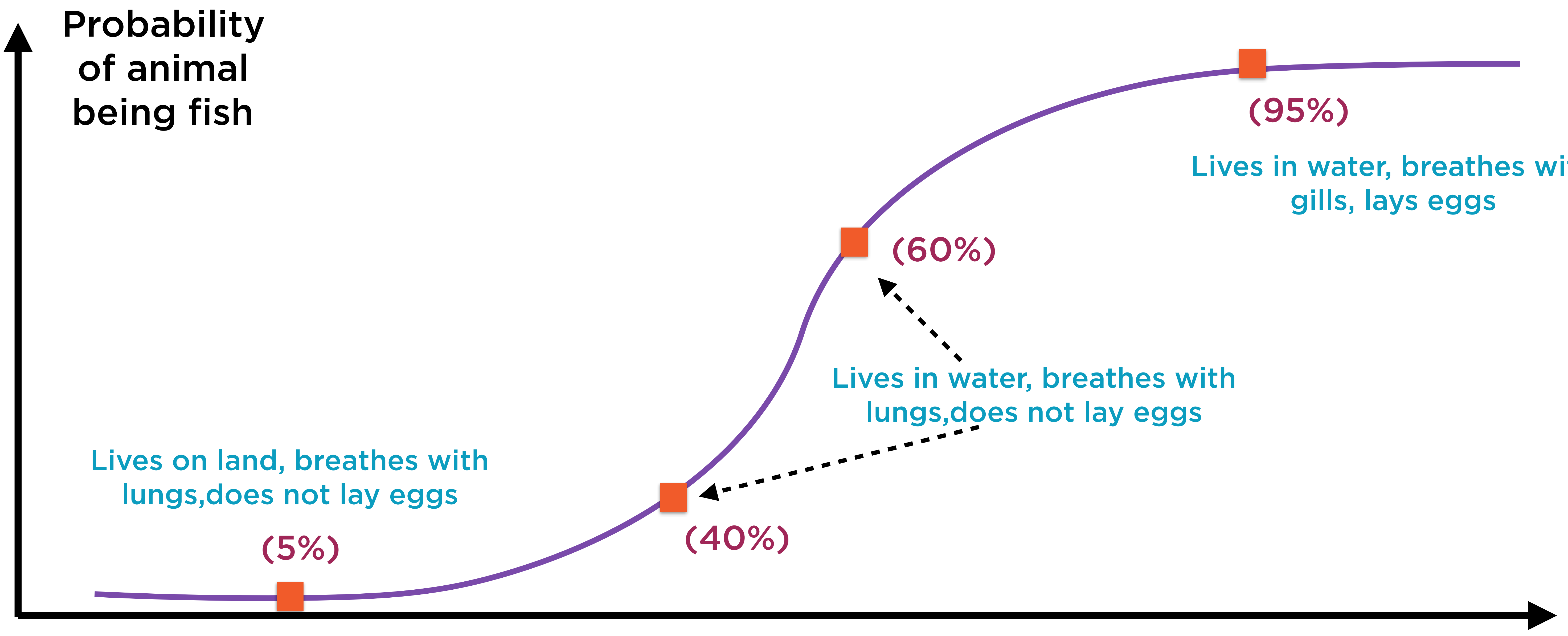
ML-based Predictor

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



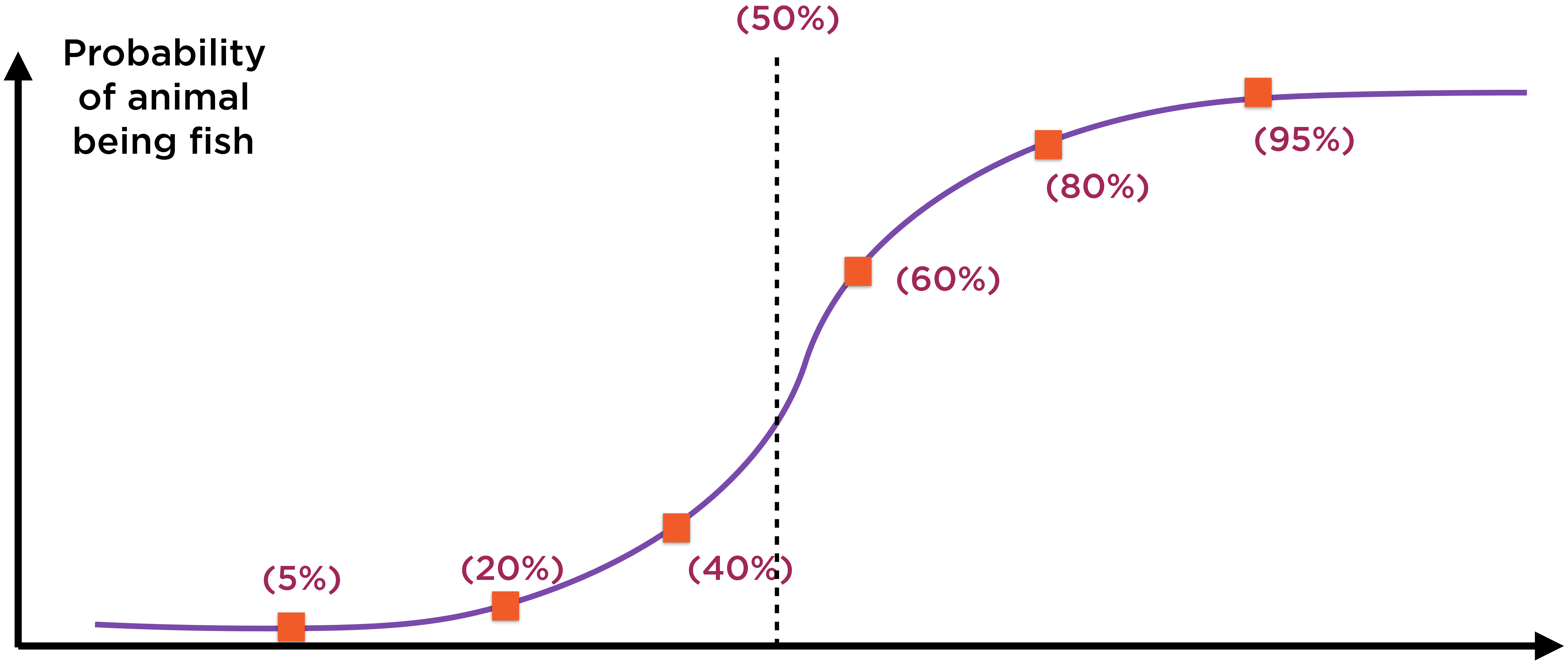
P(mammal) = 0.55

Applying Logistic Regression



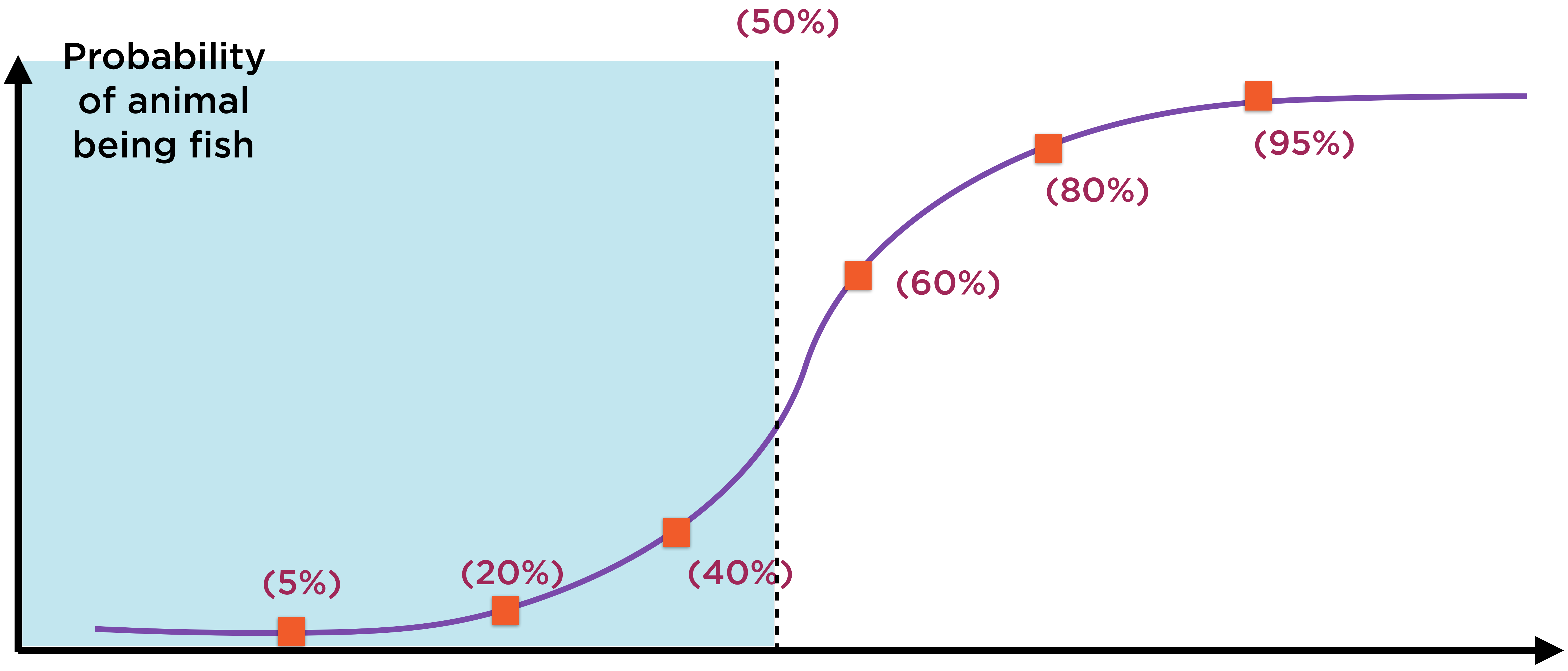
Whales: Fish or Mammals?

Applying Logistic Regression



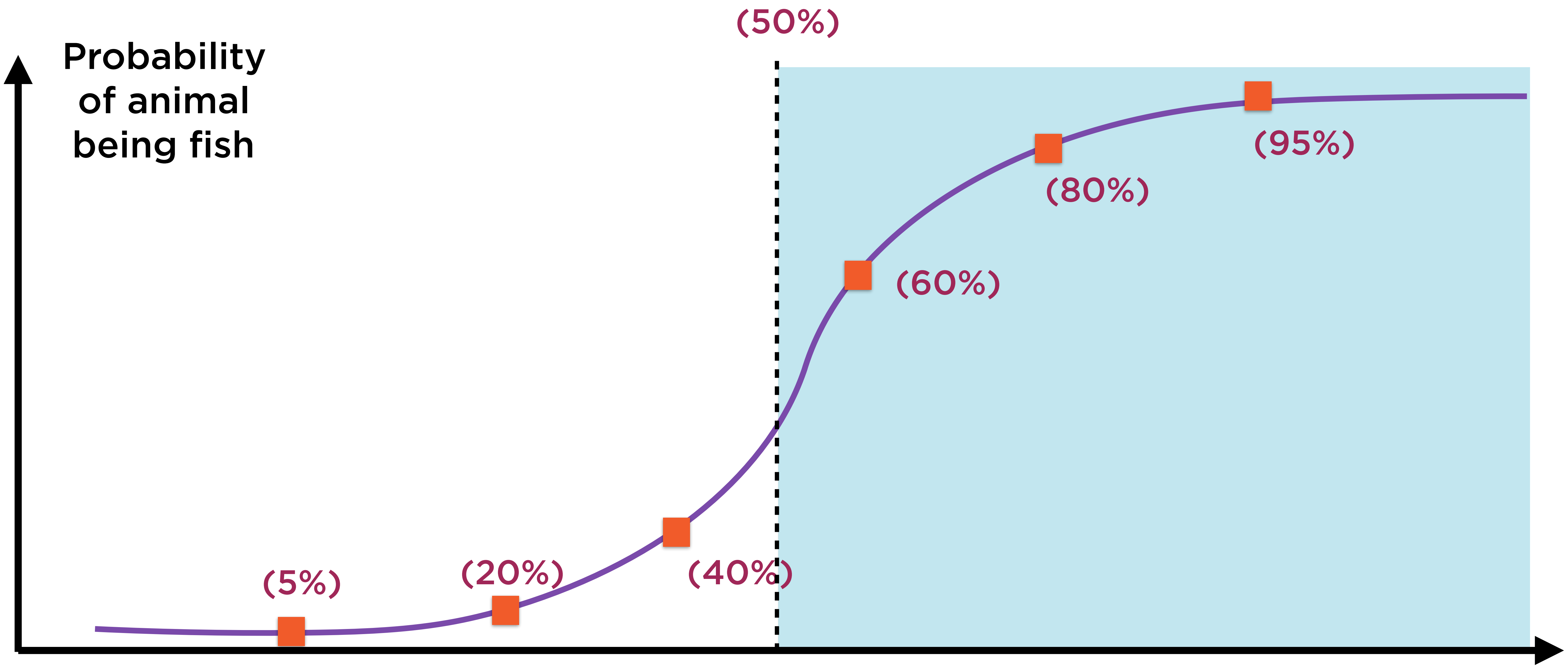
Rule of 50%

Applying Logistic Regression



If probability < 50%, it's a mammal

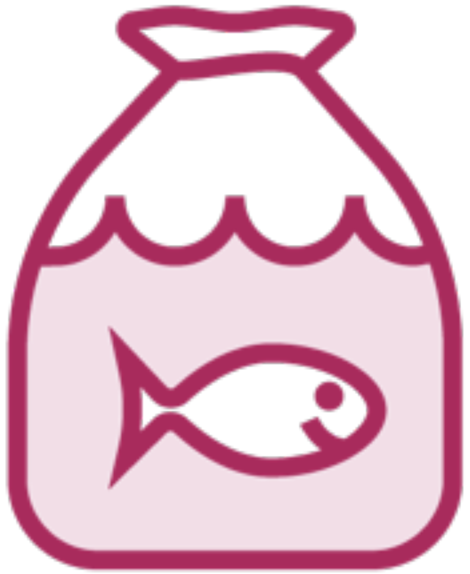
Applying Logistic Regression



If probability > 50%, it’s a fish

Applying Logistic Regression

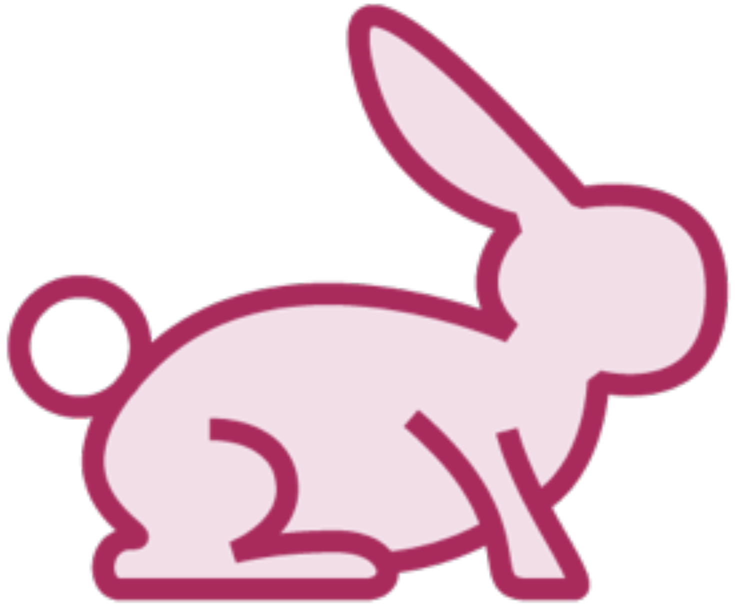
Mammal



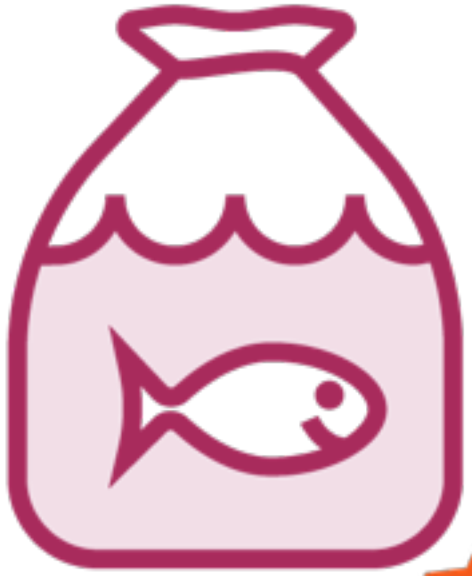
Fish

Probability of whales being Fish < 50%

Applying Logistic Regression



Mammal



Fish

Probability of whales being Fish > 50%

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

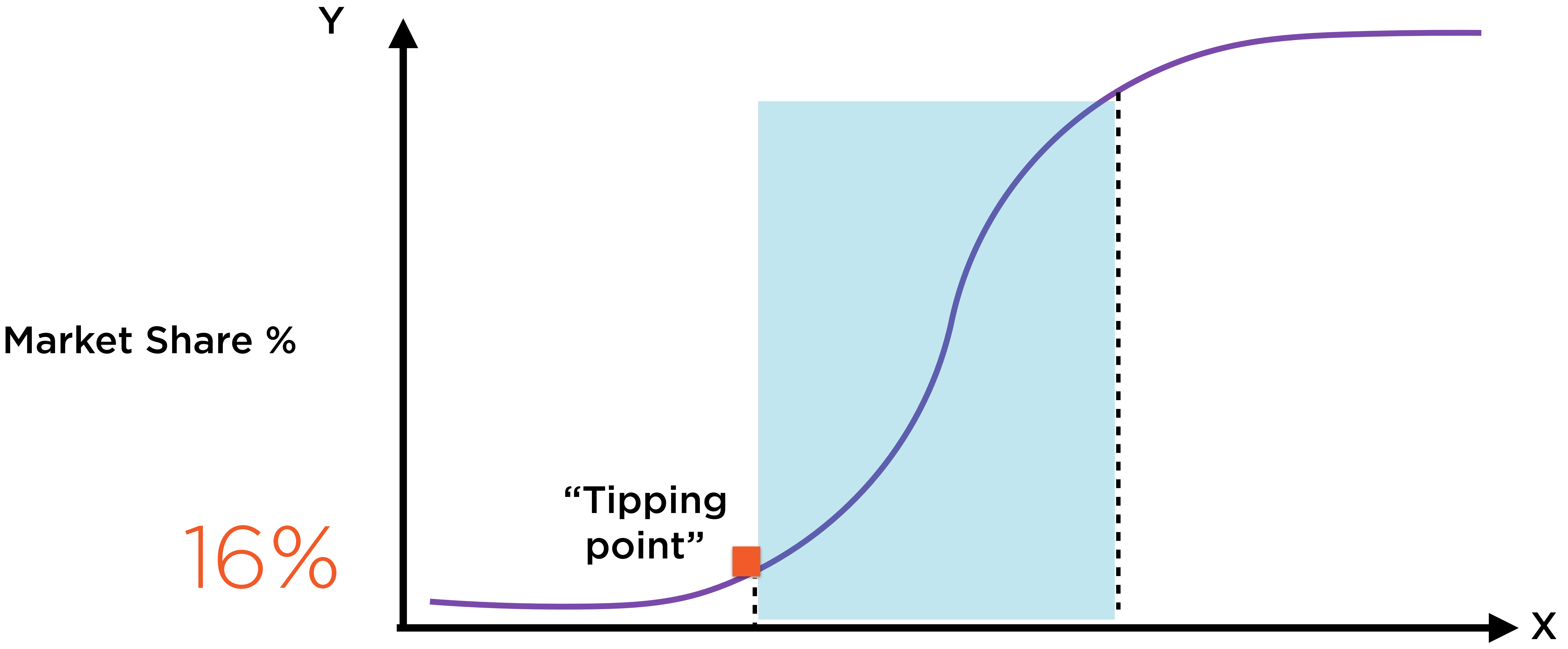
Logistic regression involves finding the “best fit” such curve

- A is the intercept

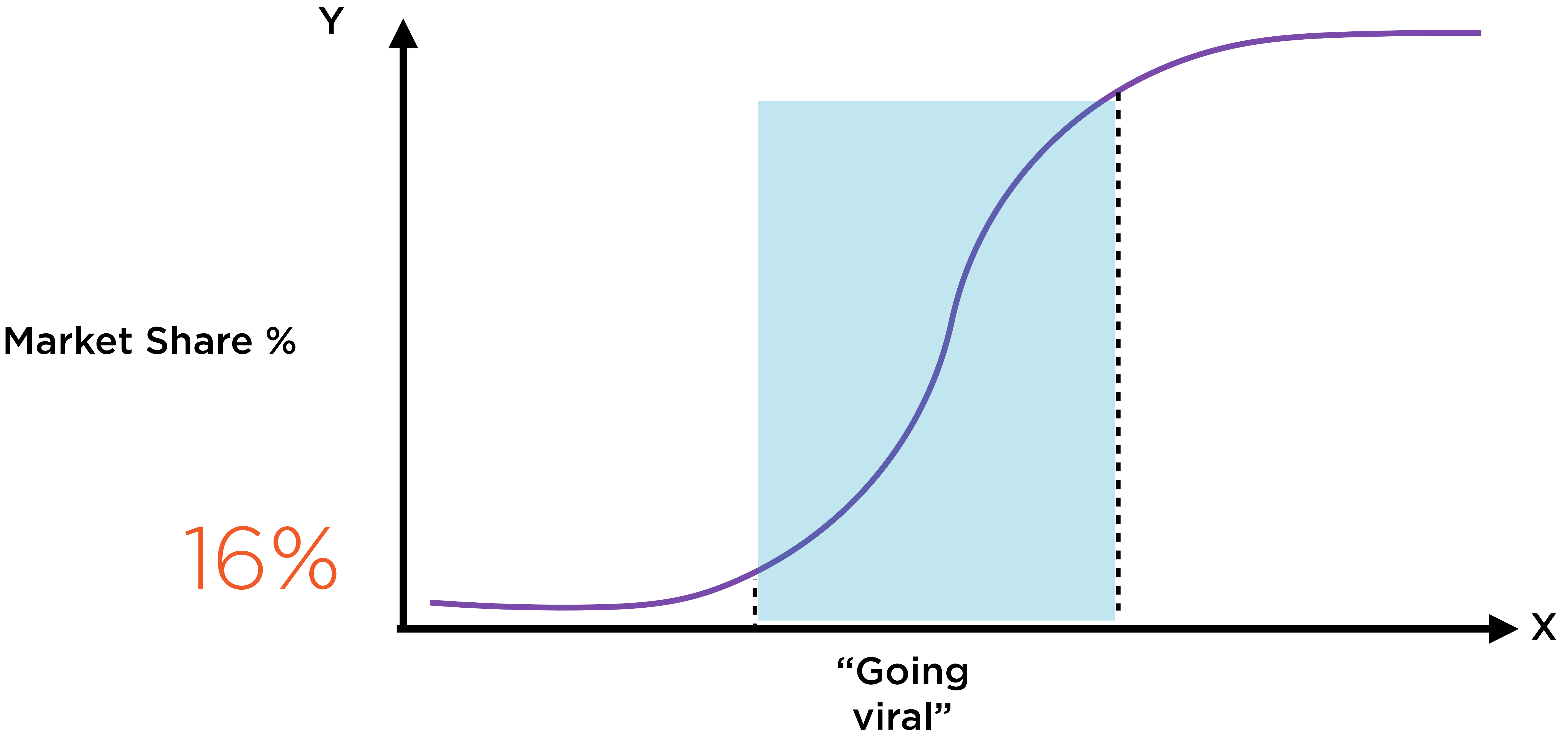
- B is the regression coefficient

(e is the constant 2.71828)

Diffusion of Innovation



Diffusion of Innovation

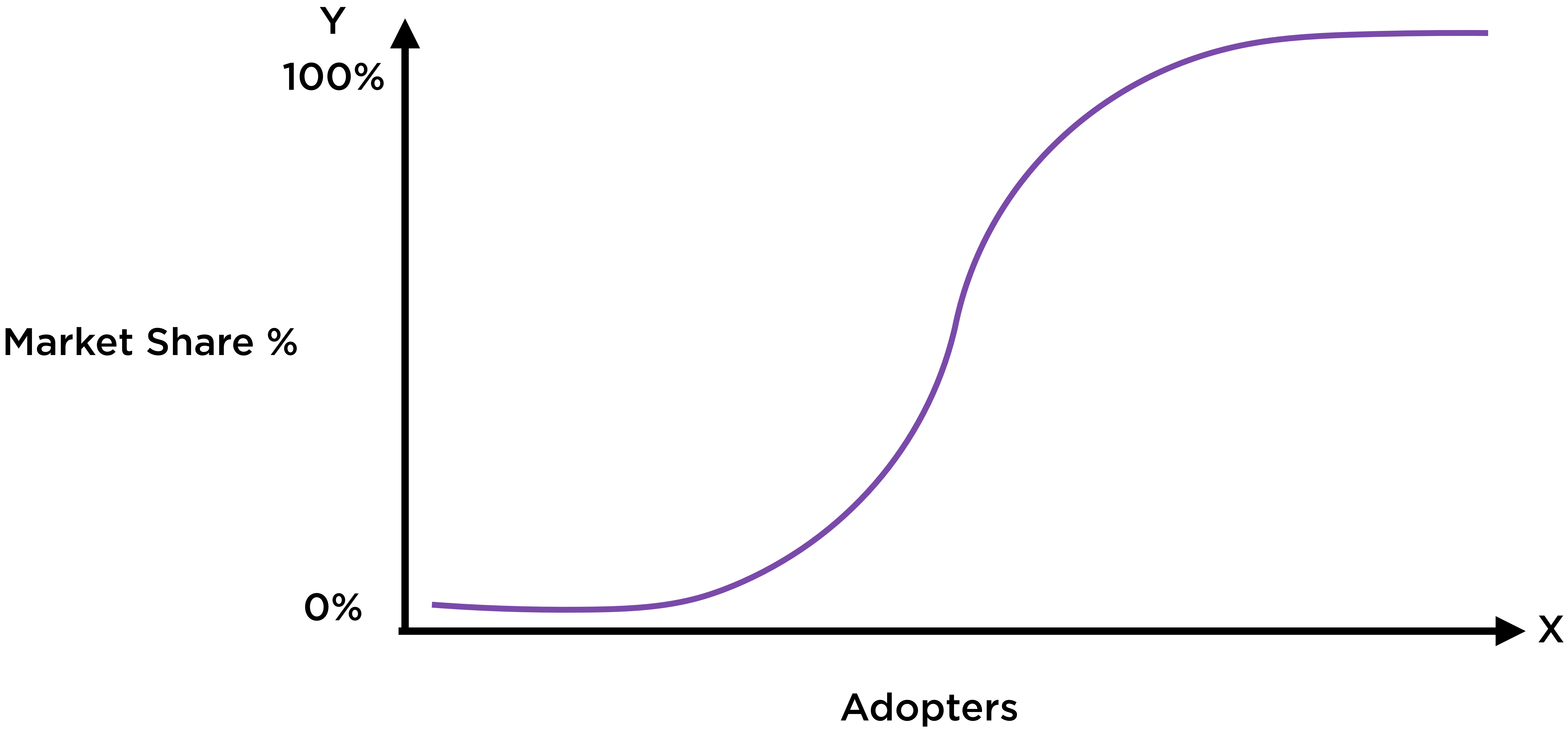


The Intuition Behind Logistic Regression

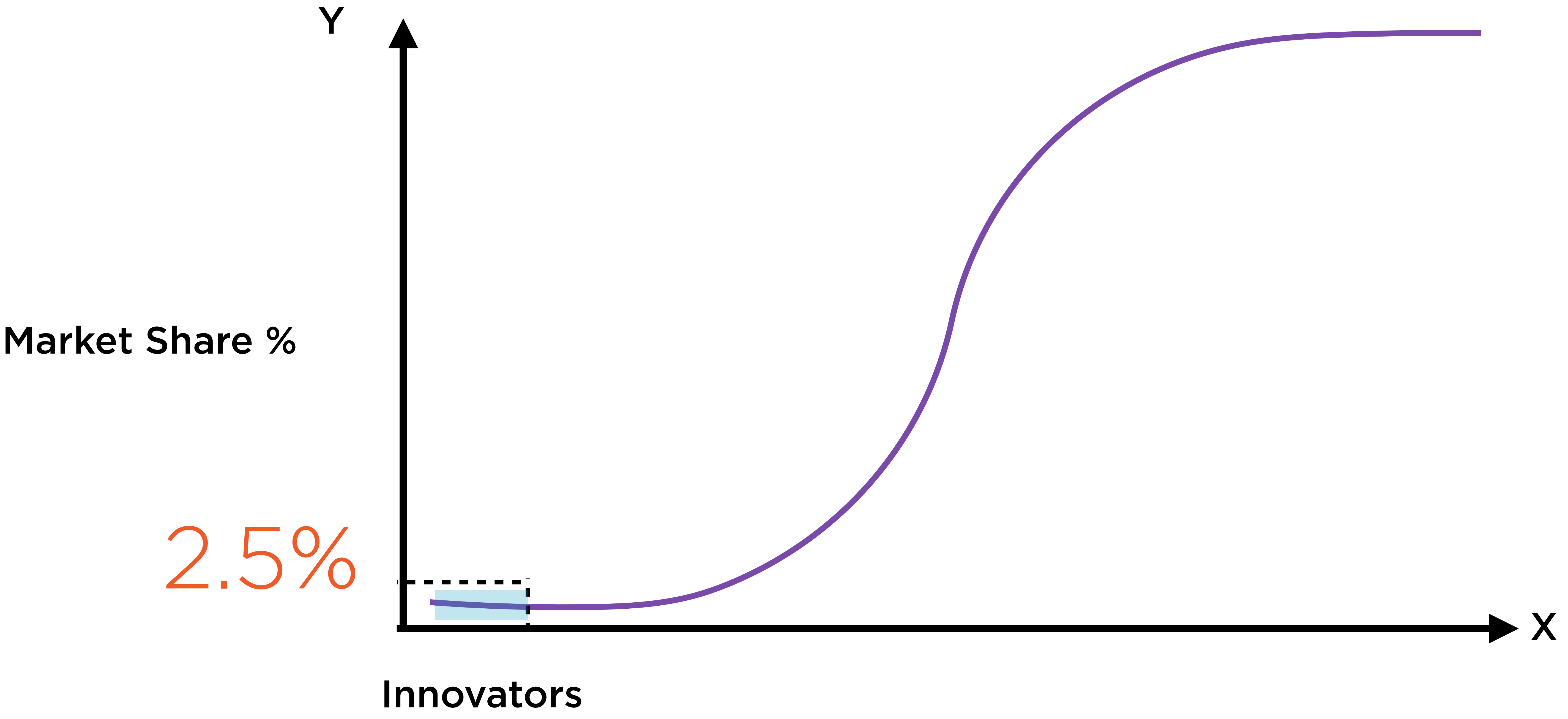
Tipping Point

A point in time when a group—or a large number of group members—rapidly and dramatically changes its behavior by widely adopting a previously rare practice

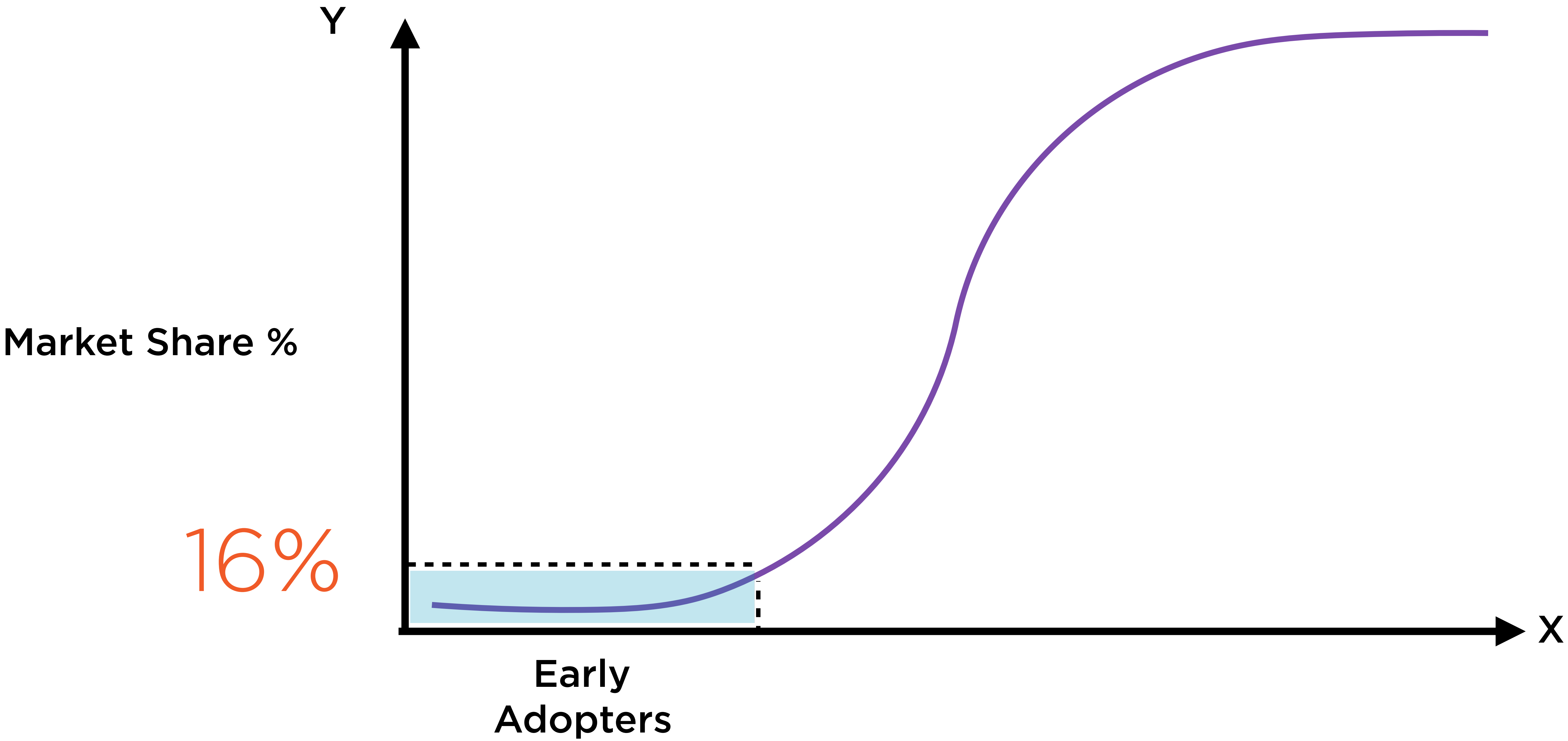
Diffusion of Innovation

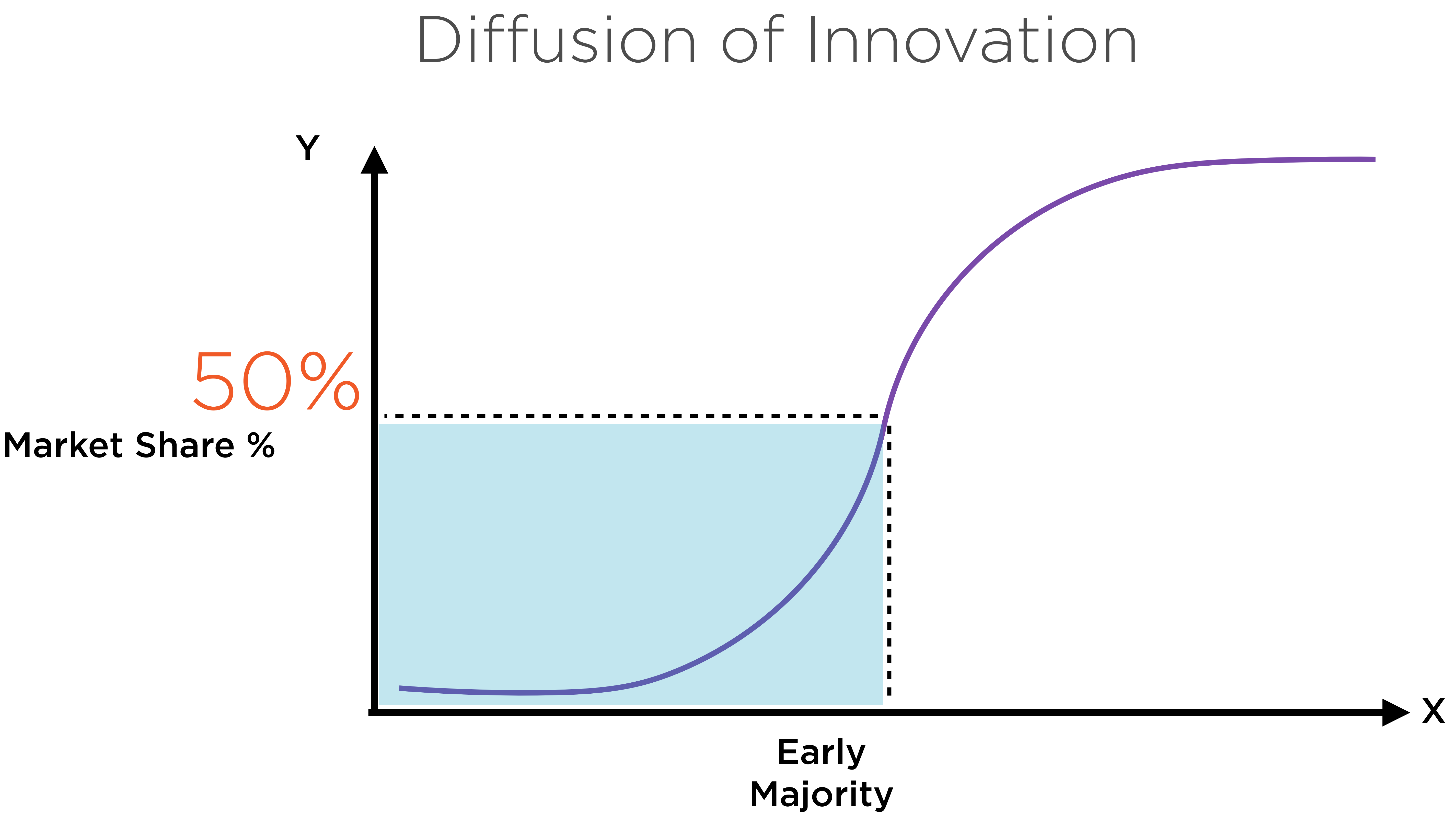


Diffusion of Innovation

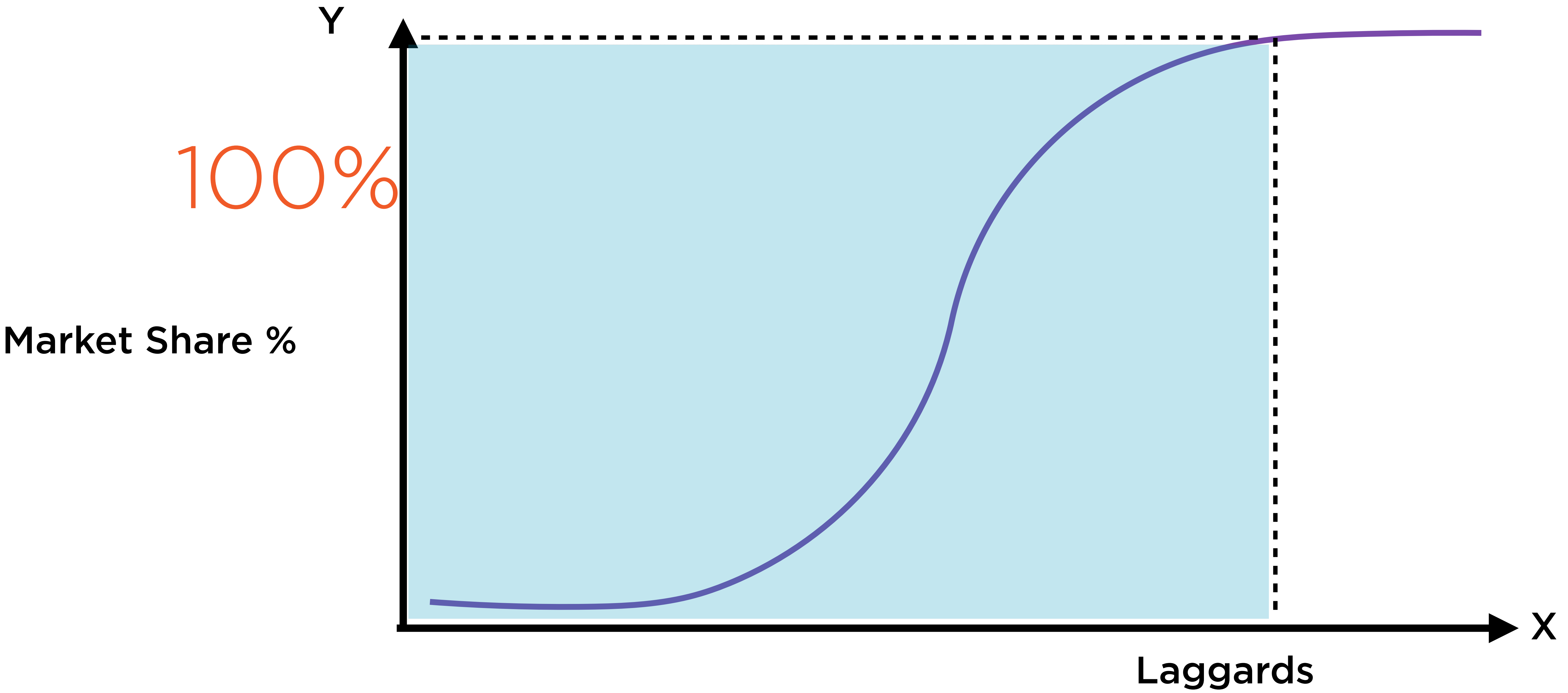


Diffusion of Innovation

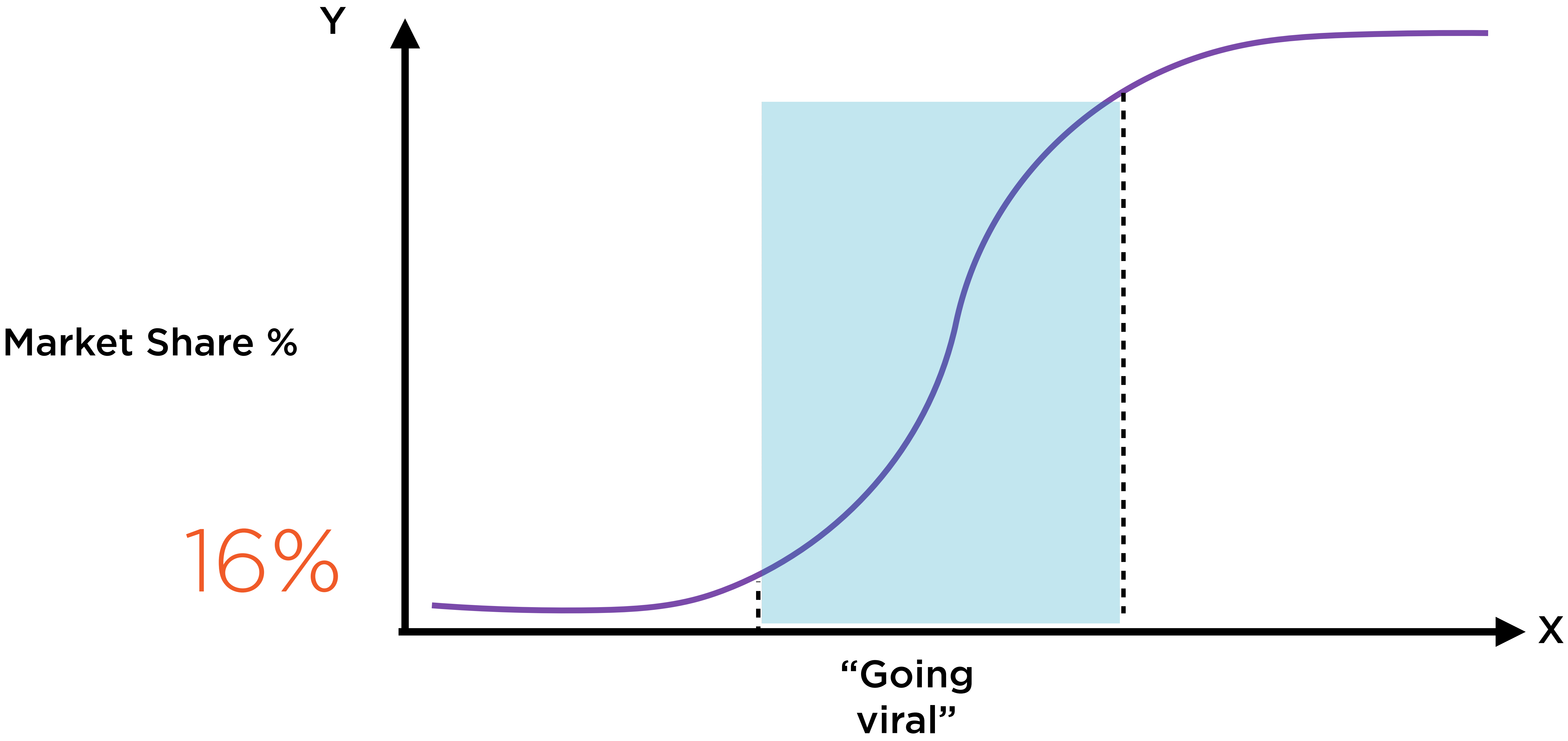



Diffusion of Innovation
Y
50%
Market Share %
X
Early
Majority

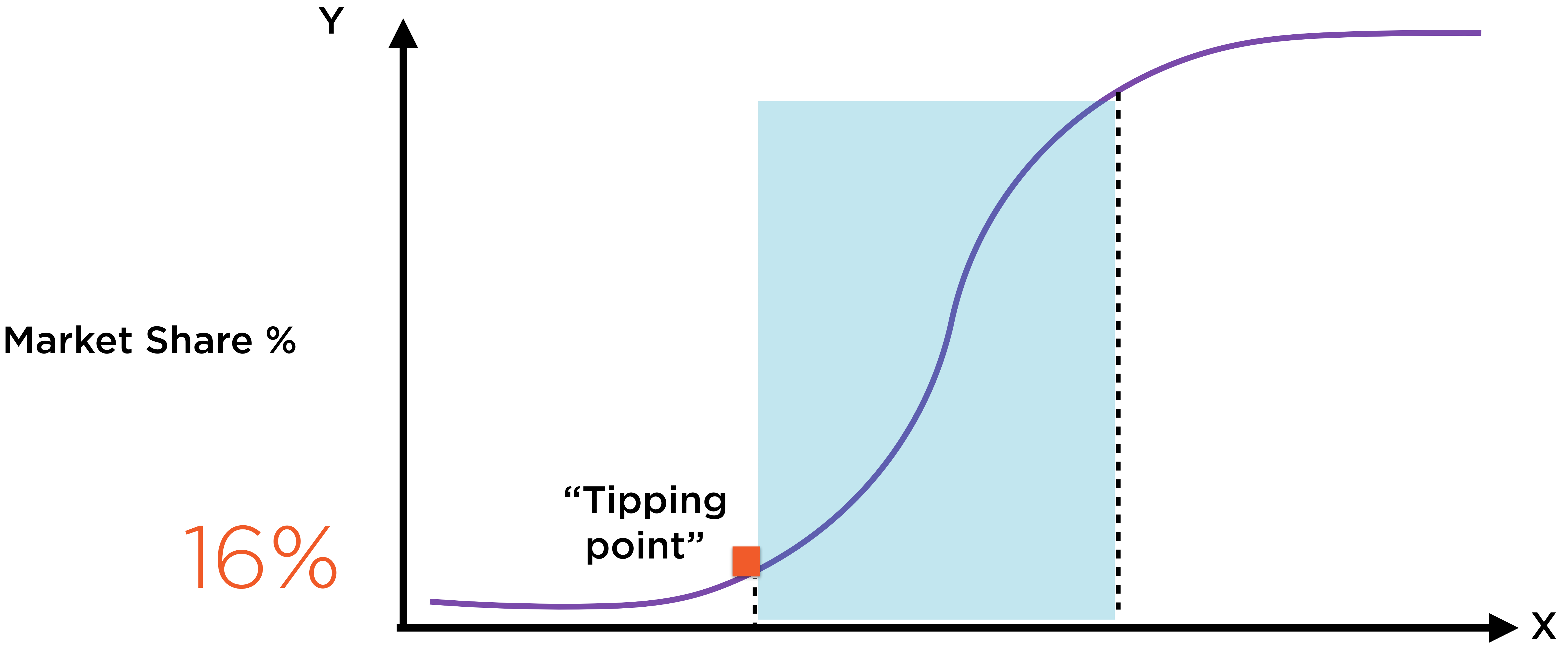
Diffusion of Innovation

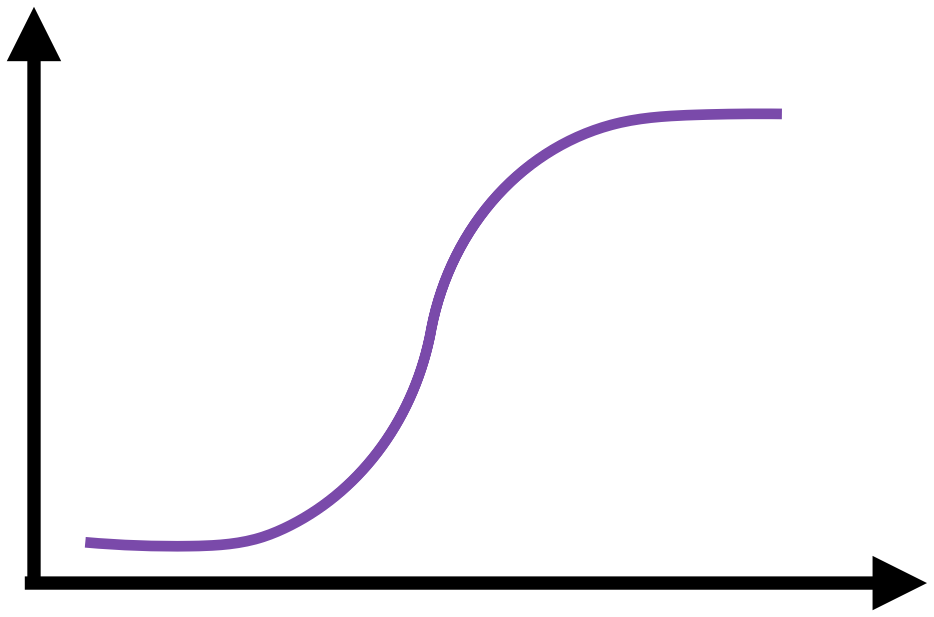


Diffusion of Innovation



Diffusion of Innovation





S-curves are widely studied, well understood

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

Logistic regression uses S-curve to estimate probabilities

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

Logistic Regression

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

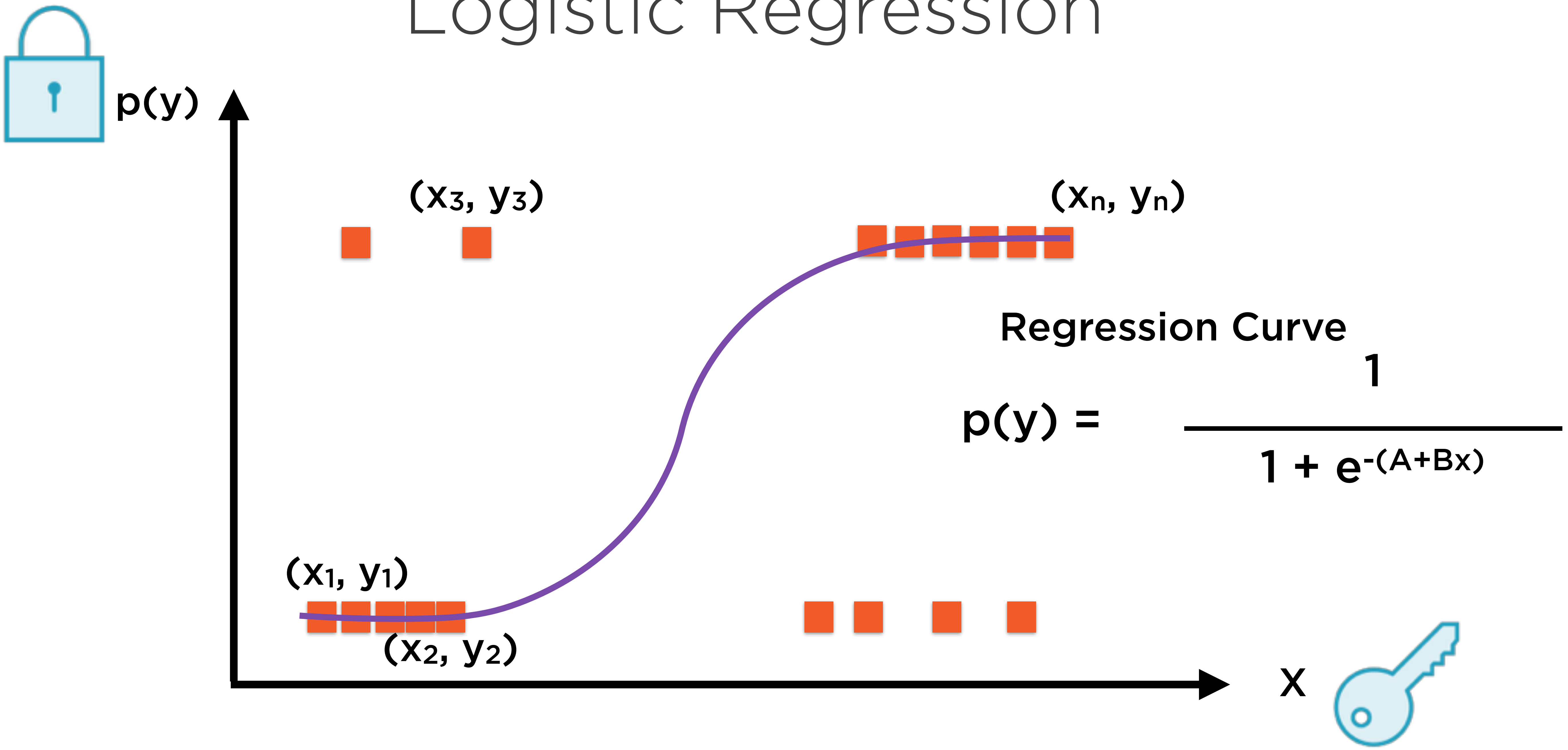
Logistic Regression

Regression Equation:

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

Given a set of points where x "predicts" probability of success in y, use logistic regression

Logistic Regression



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

Two Approaches to Deadlines

Start 5 minutes before deadline

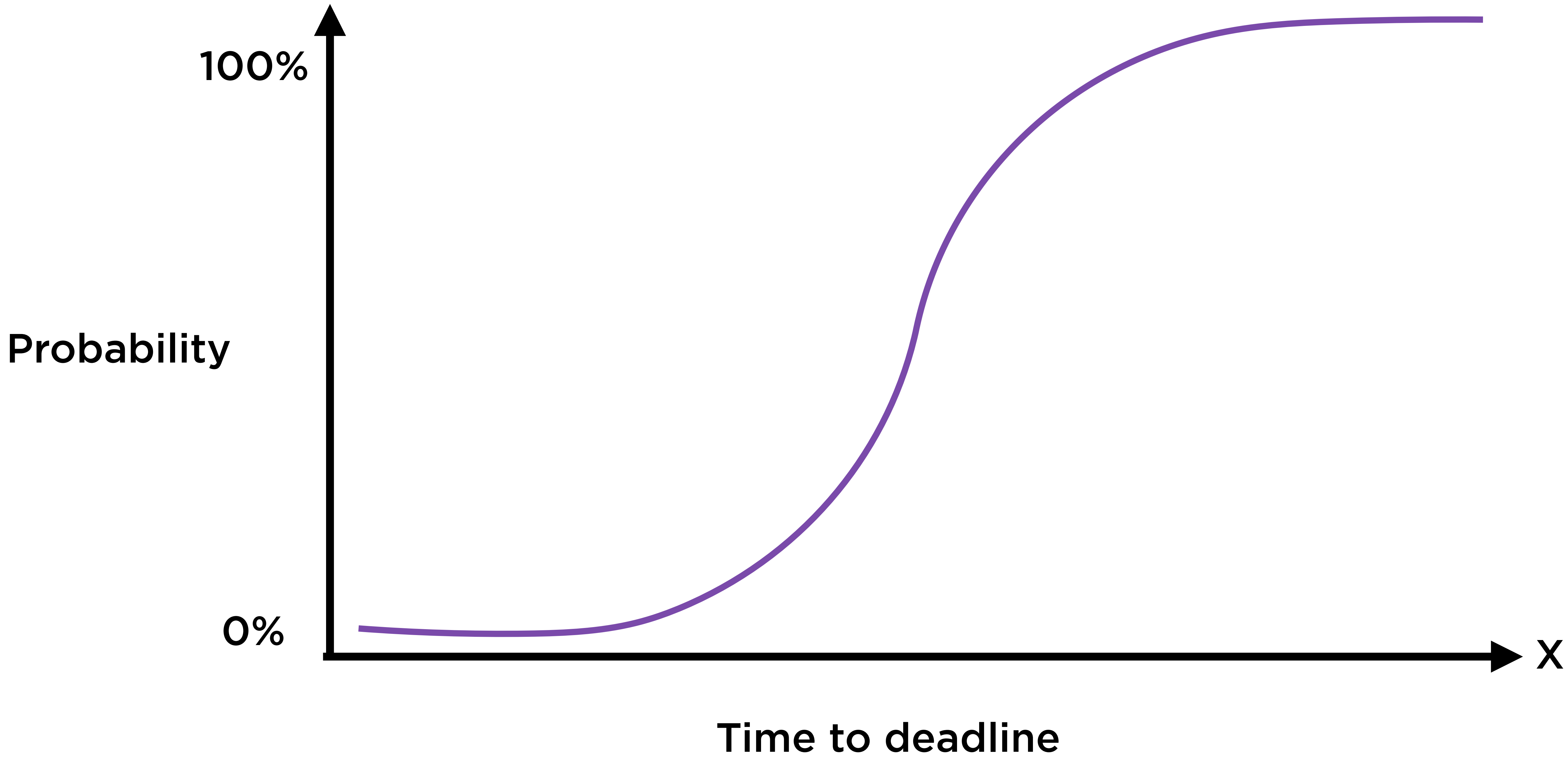
Good luck with that

Start 1 year before deadline

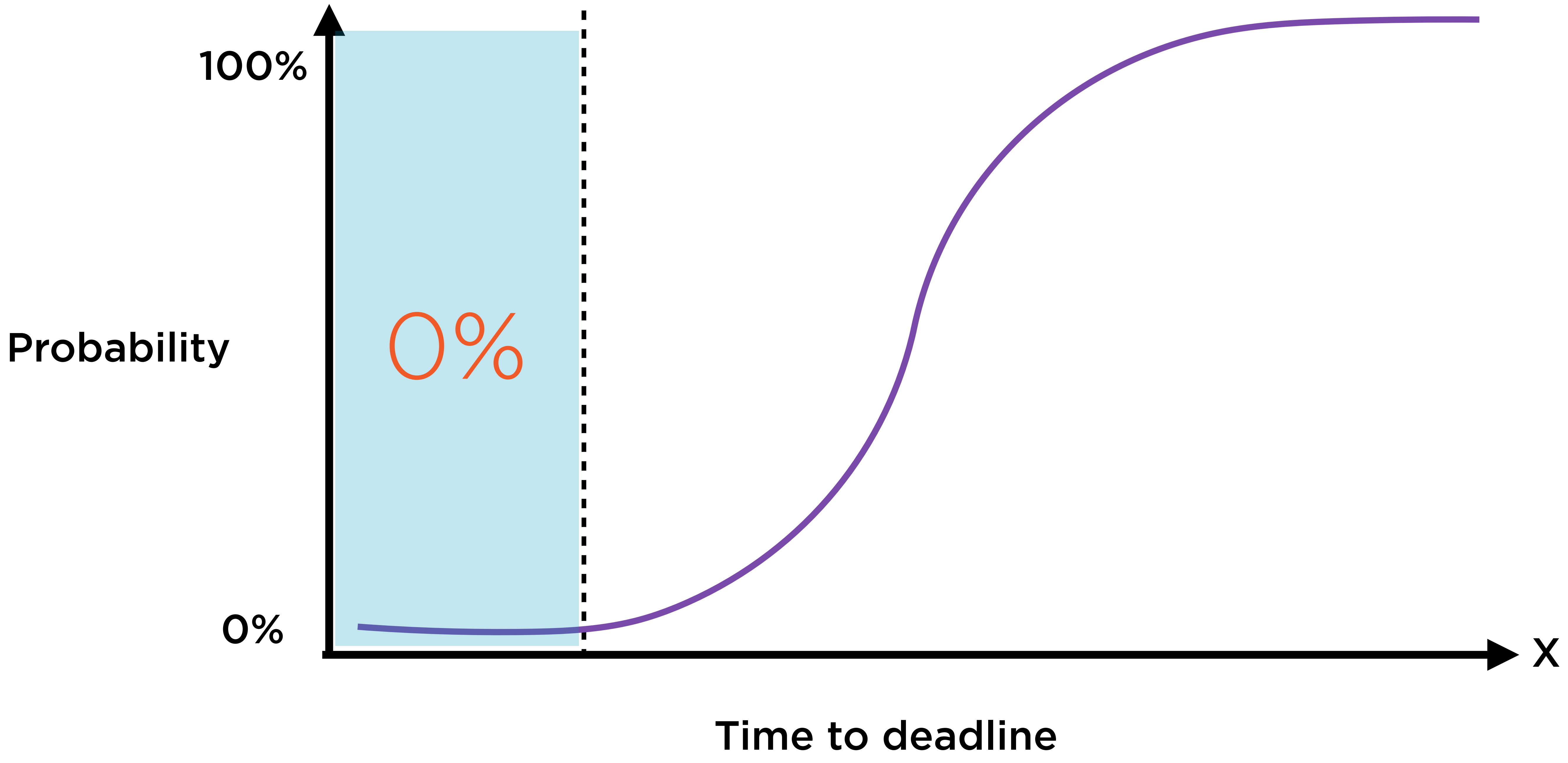
Maybe overkill

Neither approach is optimal

Working Smart with Logistic Regression

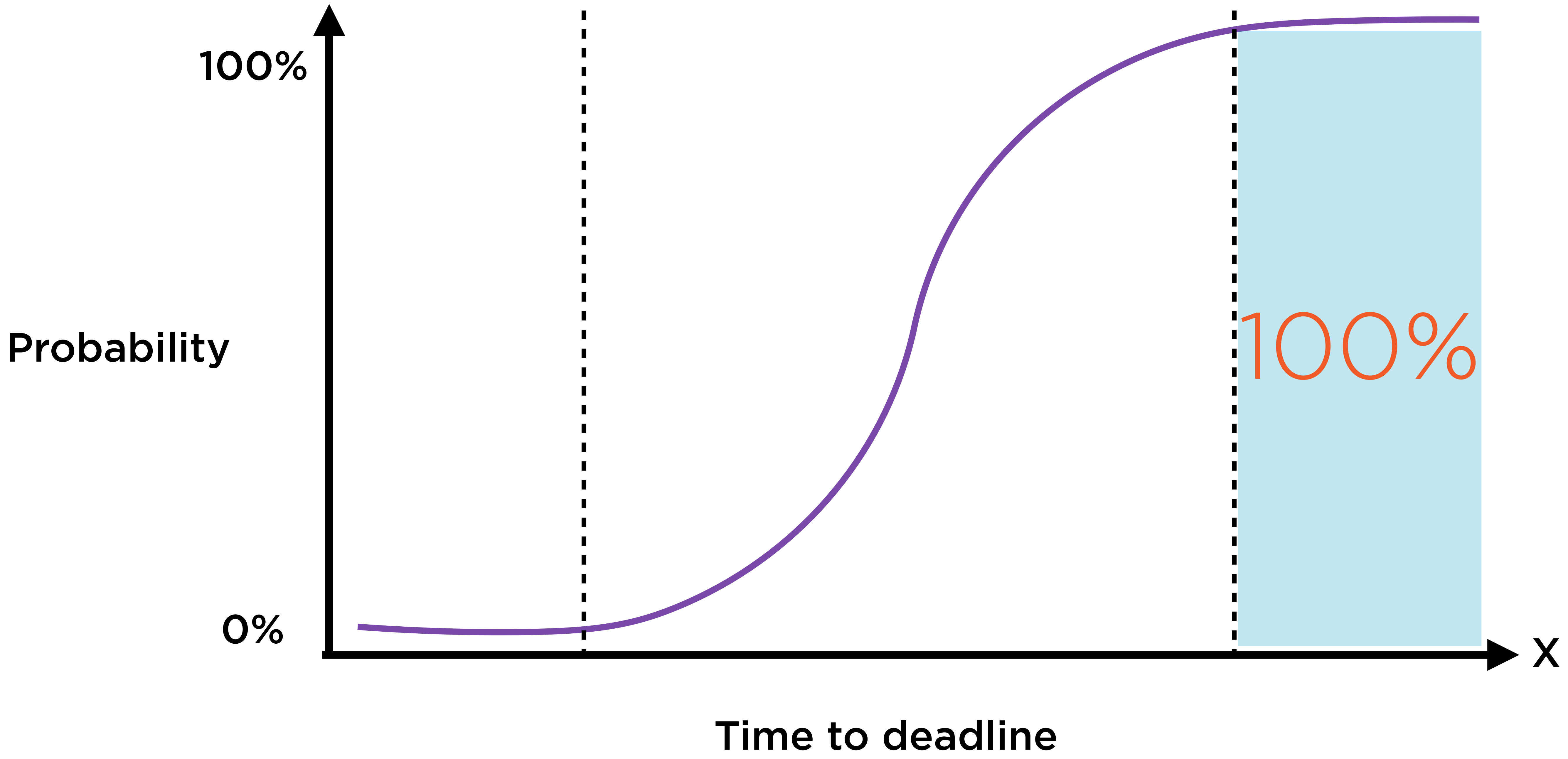


Working Smart with Logistic Regression



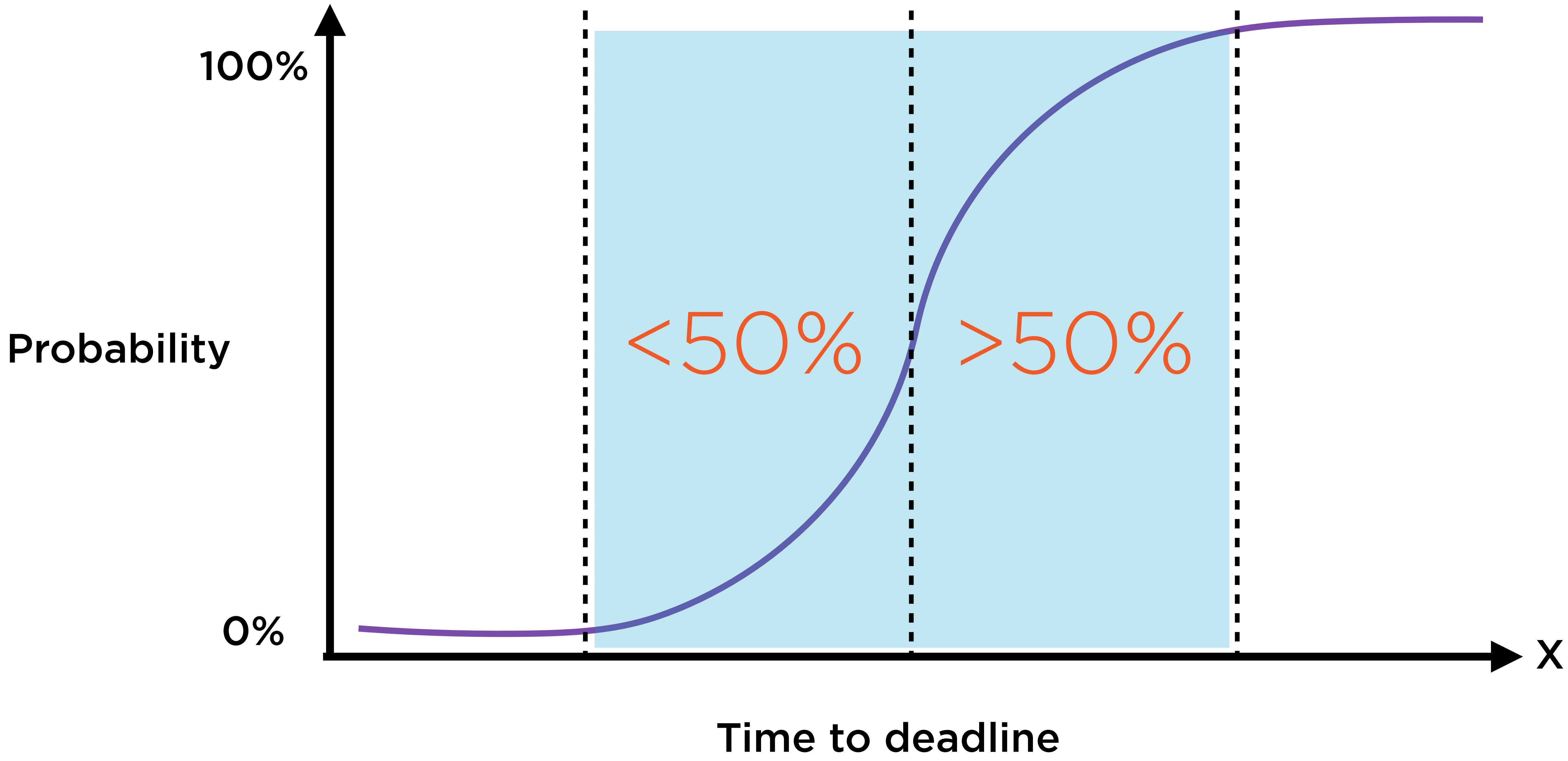
Start too late, and you'll definitely miss

Working Smart with Logistic Regression

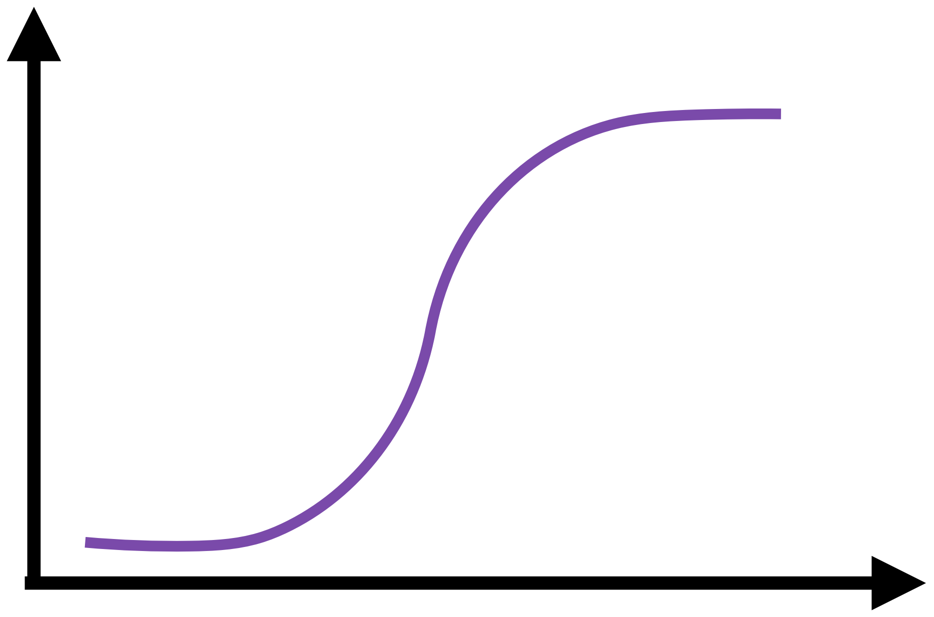


Start too early, and you'll definitely make it

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

Working Smart with Logistic Regression

Probabilities	Categorical Variables	Causes
p(y)	y	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

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

$y_i = 1$ or 0 (hit or miss)

x_i = time spent working on deadline

$p(y_i)$ = probability that $y_i = 1$

$1 - p(y_i)$ = probability that $y_i = 0$

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

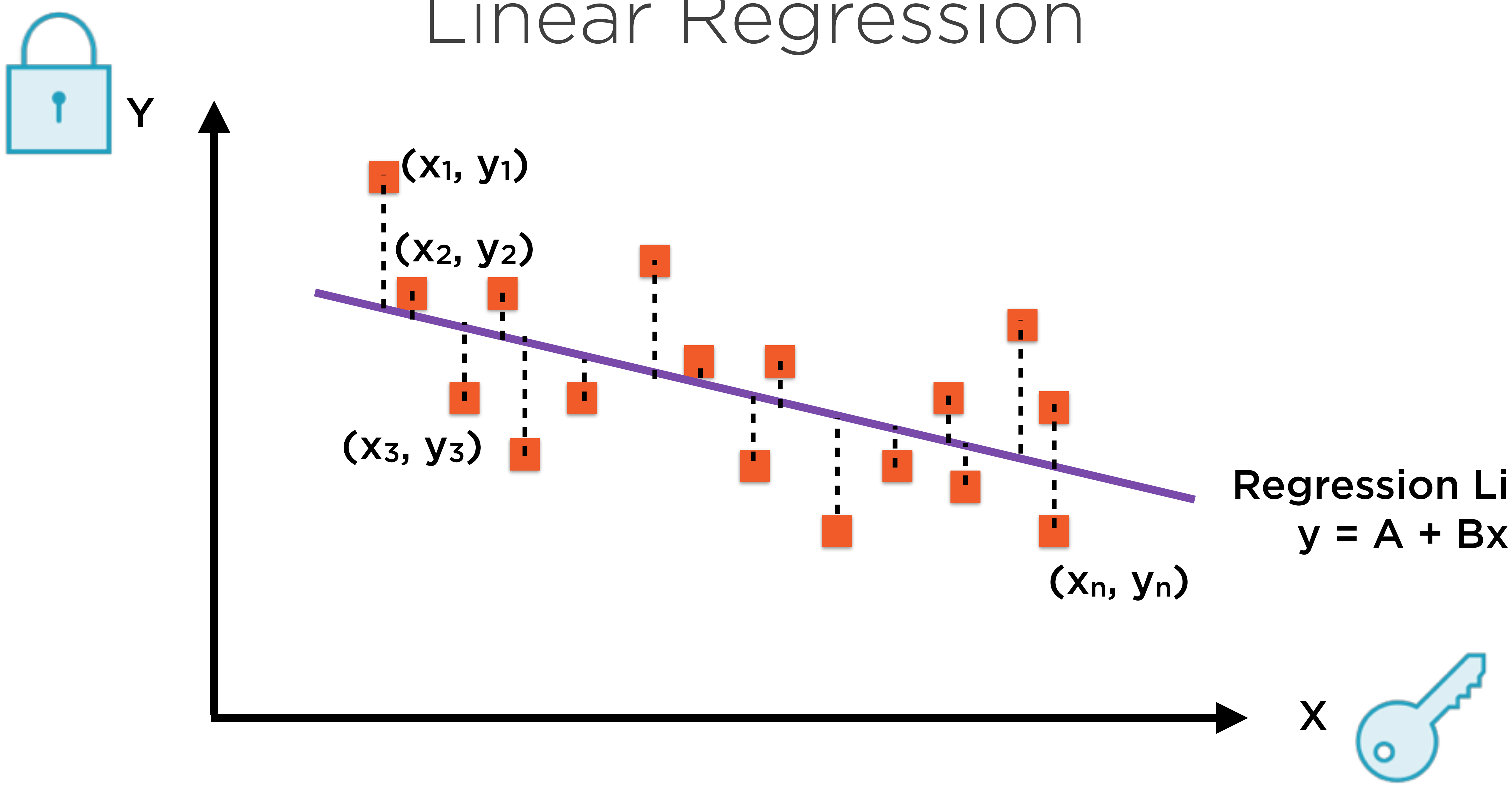
Logistic regression involves finding the "best fit" such curve

- A is the intercept

- B is the regression coefficient

(e is the constant 2.71828)

Linear Regression



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

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

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

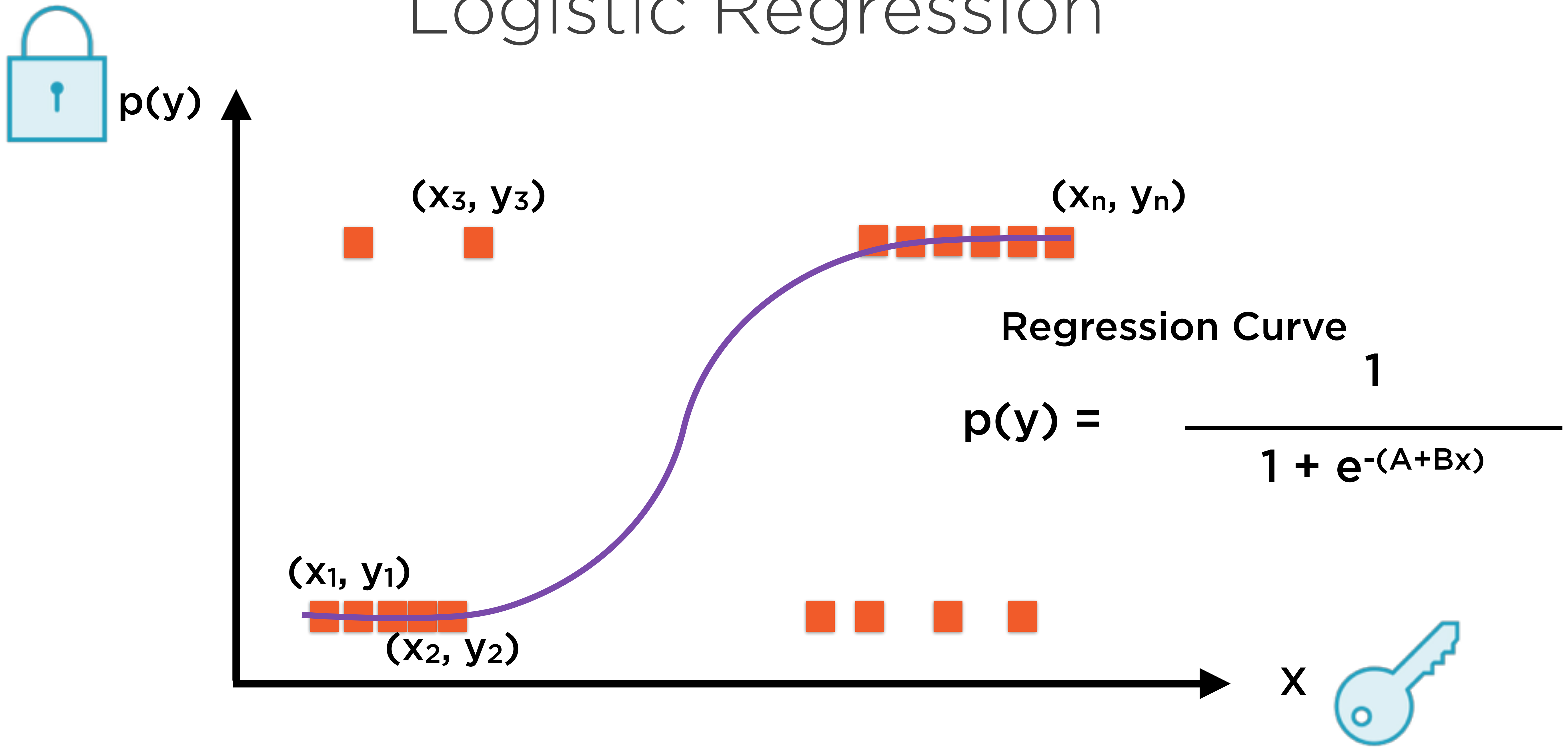
$$\text{logit}(p) = \ln\left(\frac{p}{1 - p}\right)$$

Solve regression problem using cookie-cutter solvers

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

Linear Regression

$$y = A + Bx$$

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

Logistic Regression

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

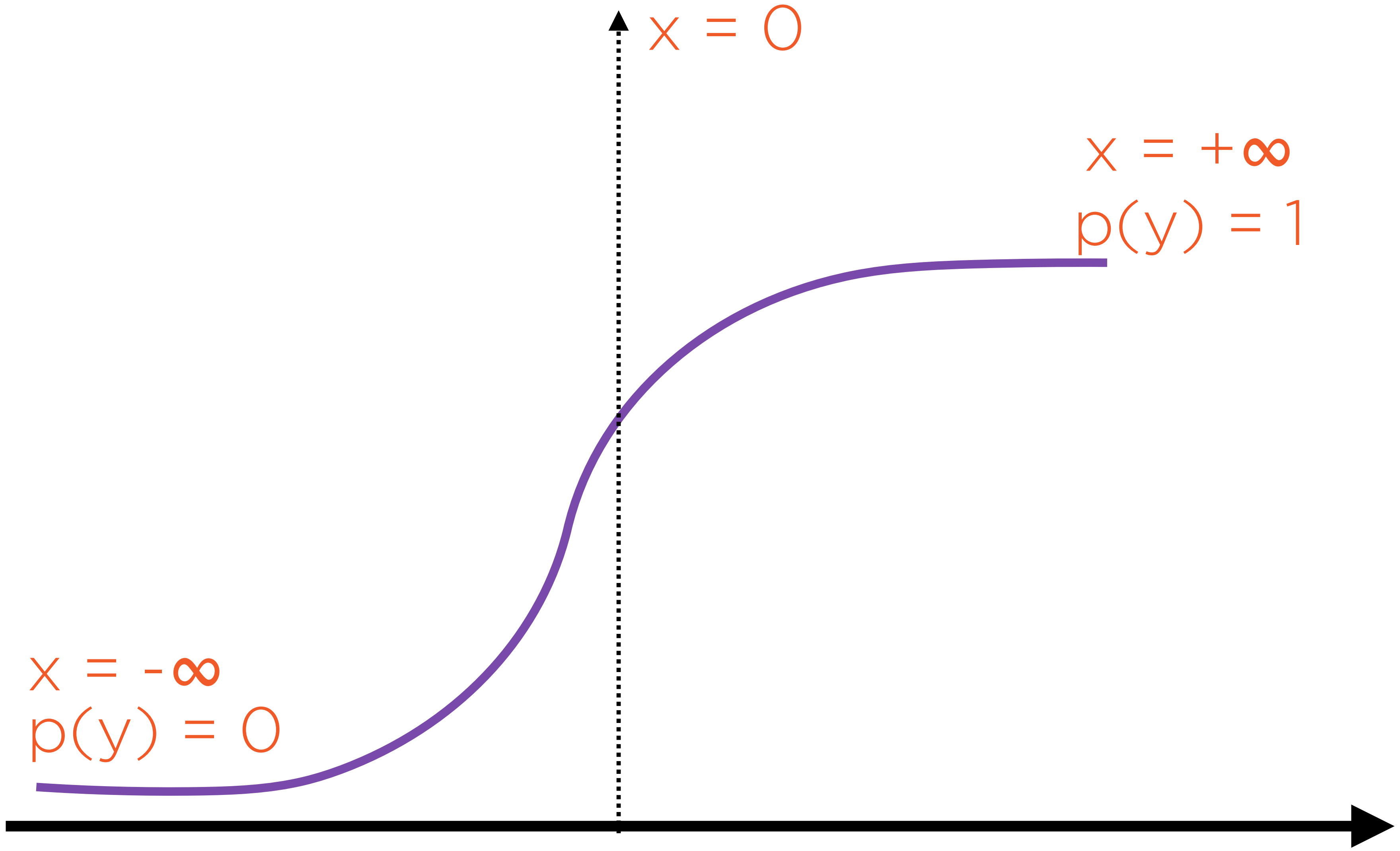
$$p(y_i) = \frac{1}{1 + e^{-(A+Bx_i)}}$$

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

...

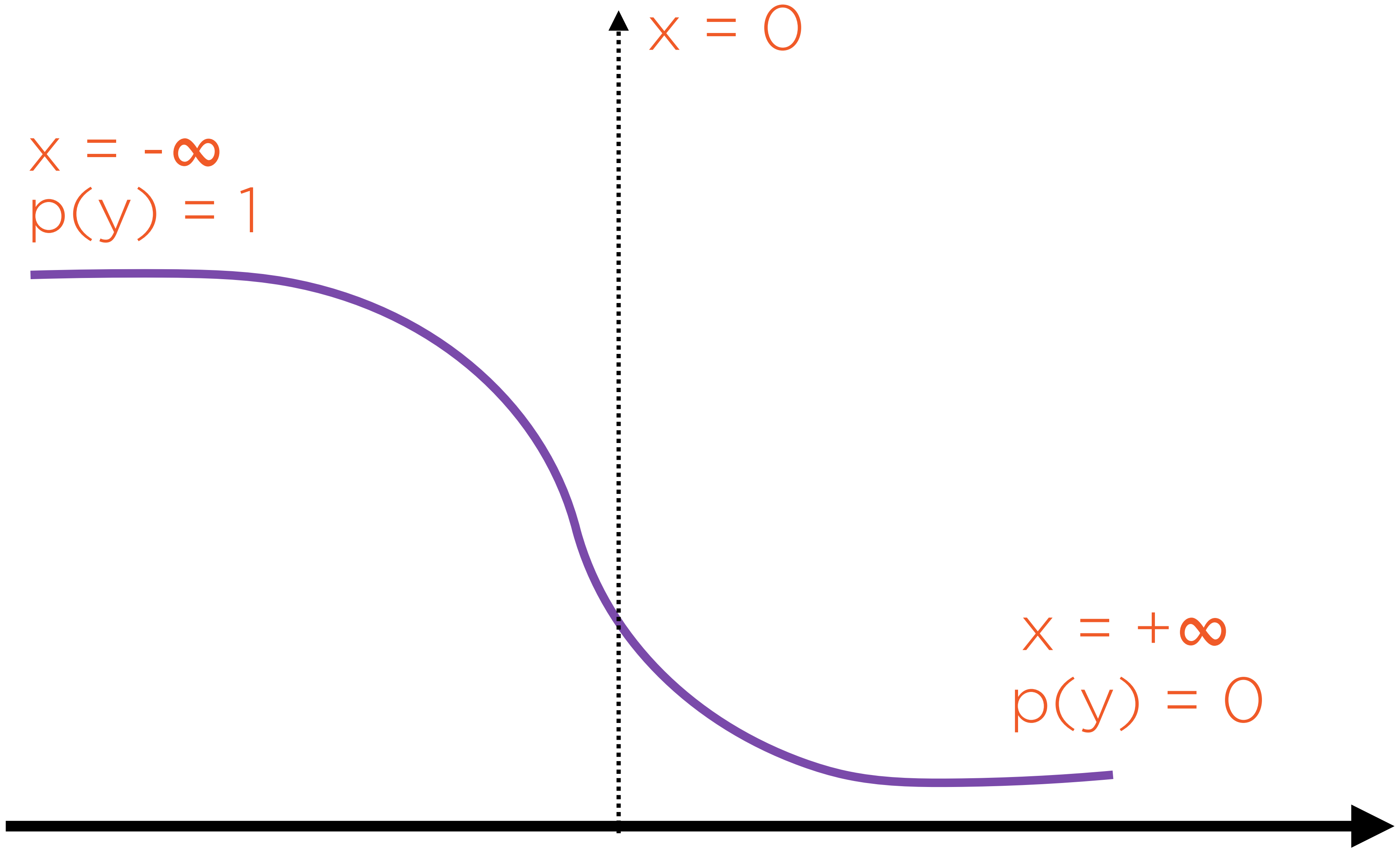
$$p(y_n) = \frac{1}{1 + e^{-(A+Bx_n)}}$$

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



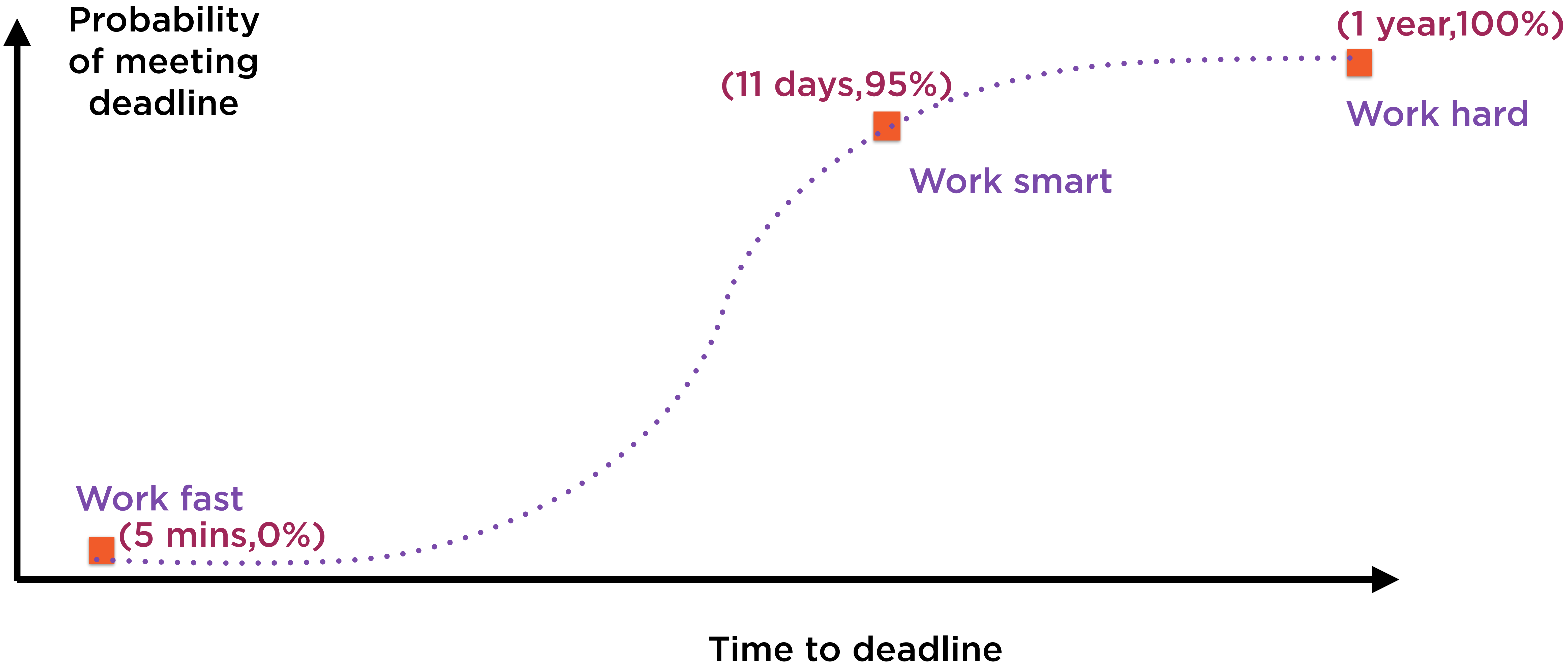
If A and B are **positive**

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

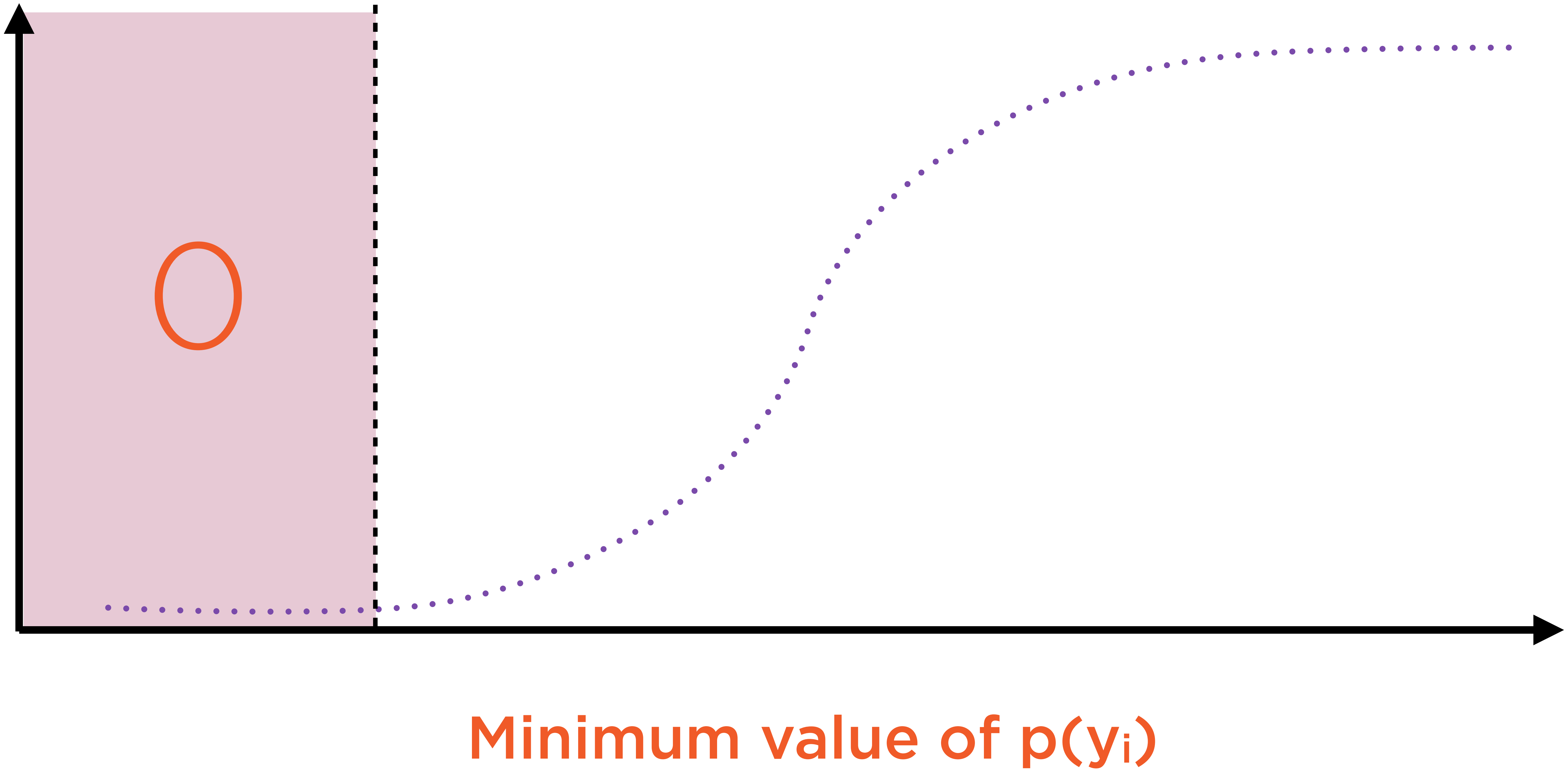


If A and B are **negative**

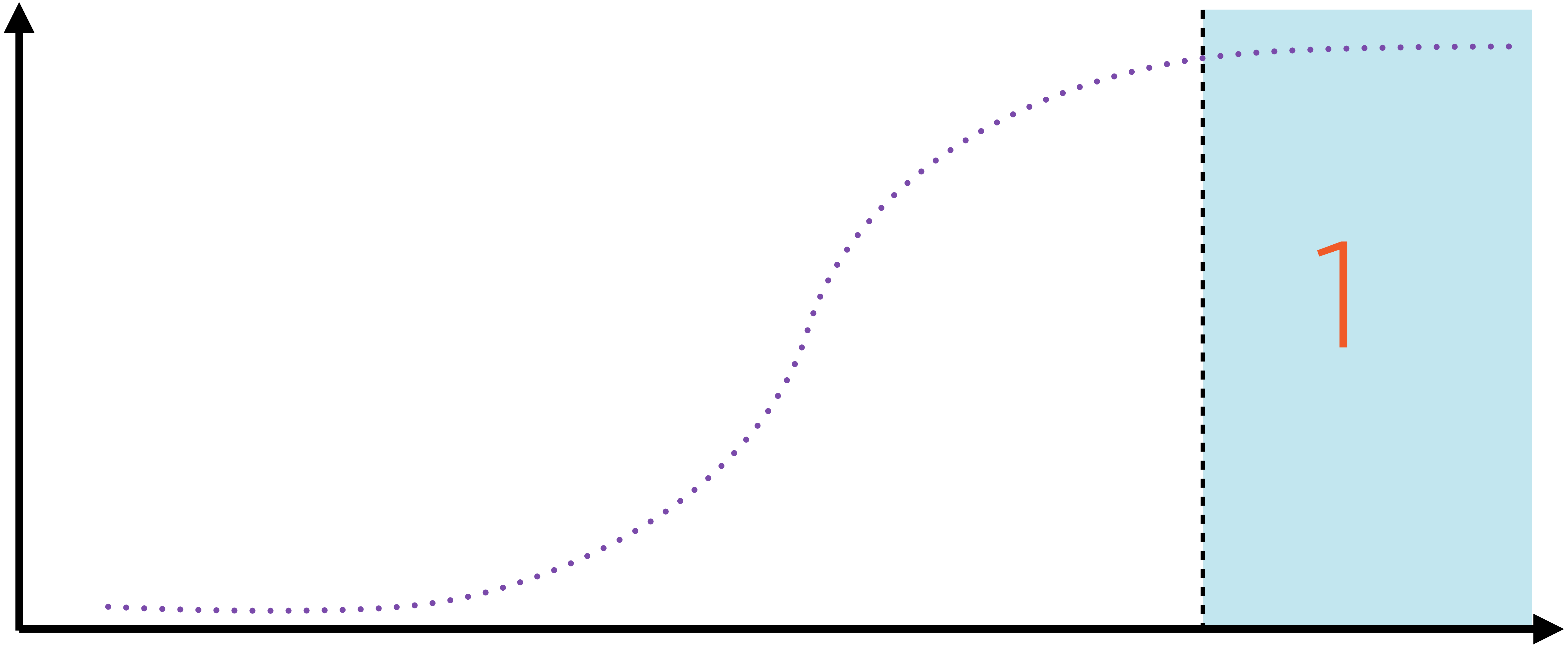
Working Hard, Fast, Smart



Working Hard, Fast, Smart

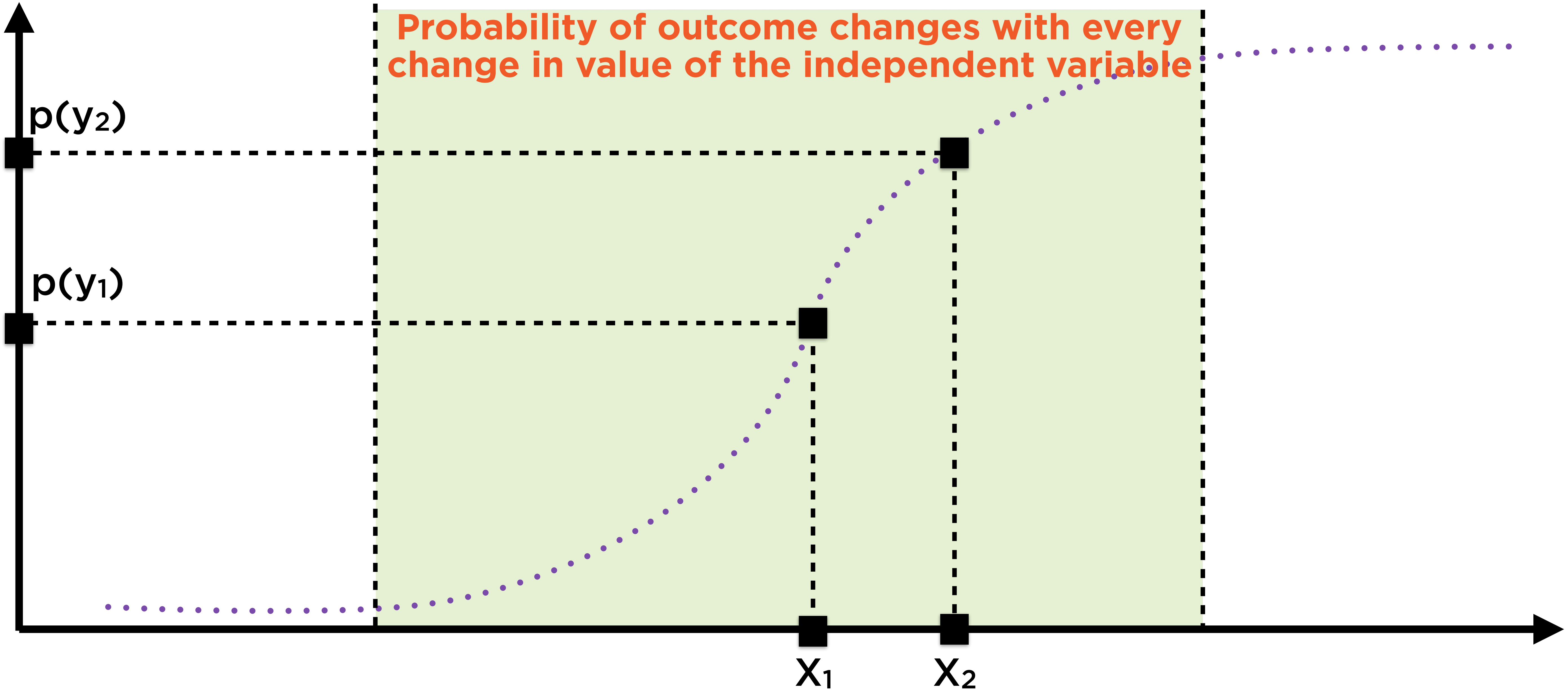


Working Hard, Fast, Smart



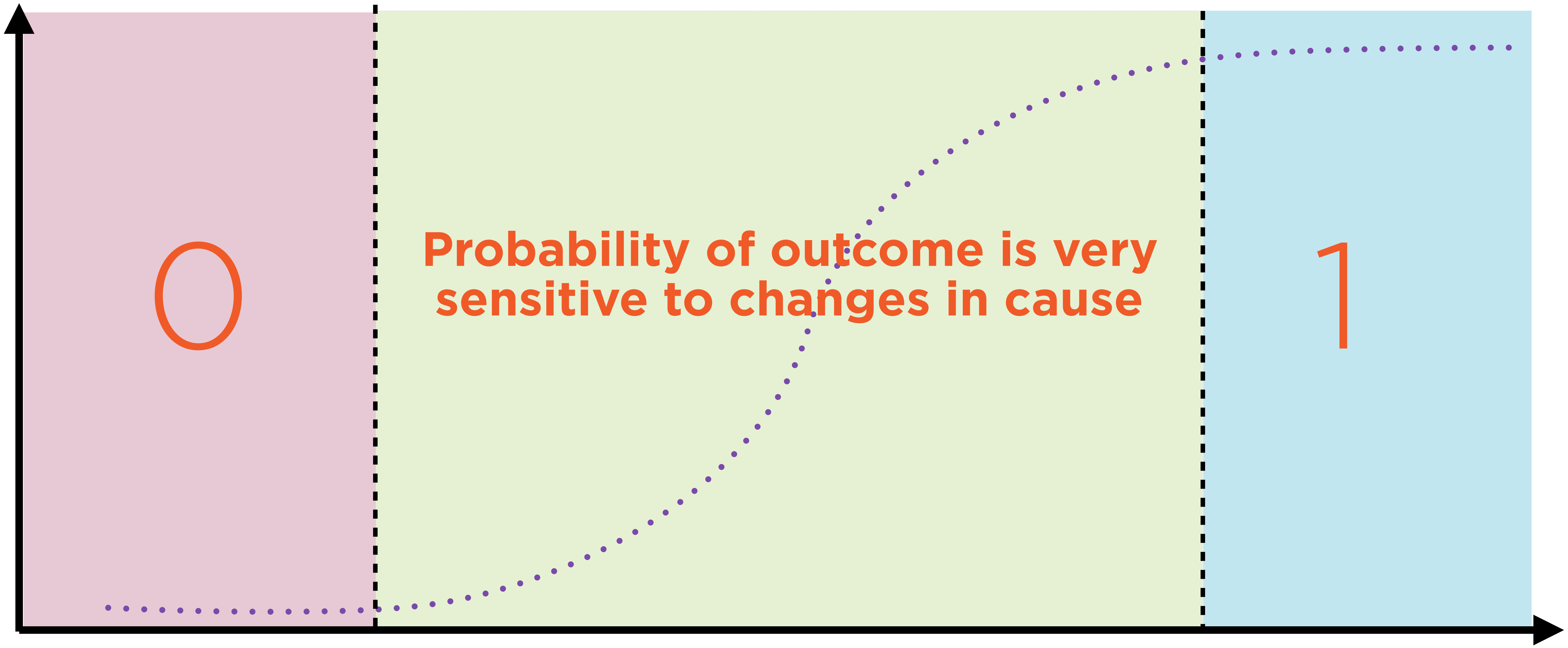
Maximum value of $p(y_i)$

Working Hard, Fast, Smart



Between maximum and minimum values of $p(y_i)$

Logistic Regression



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

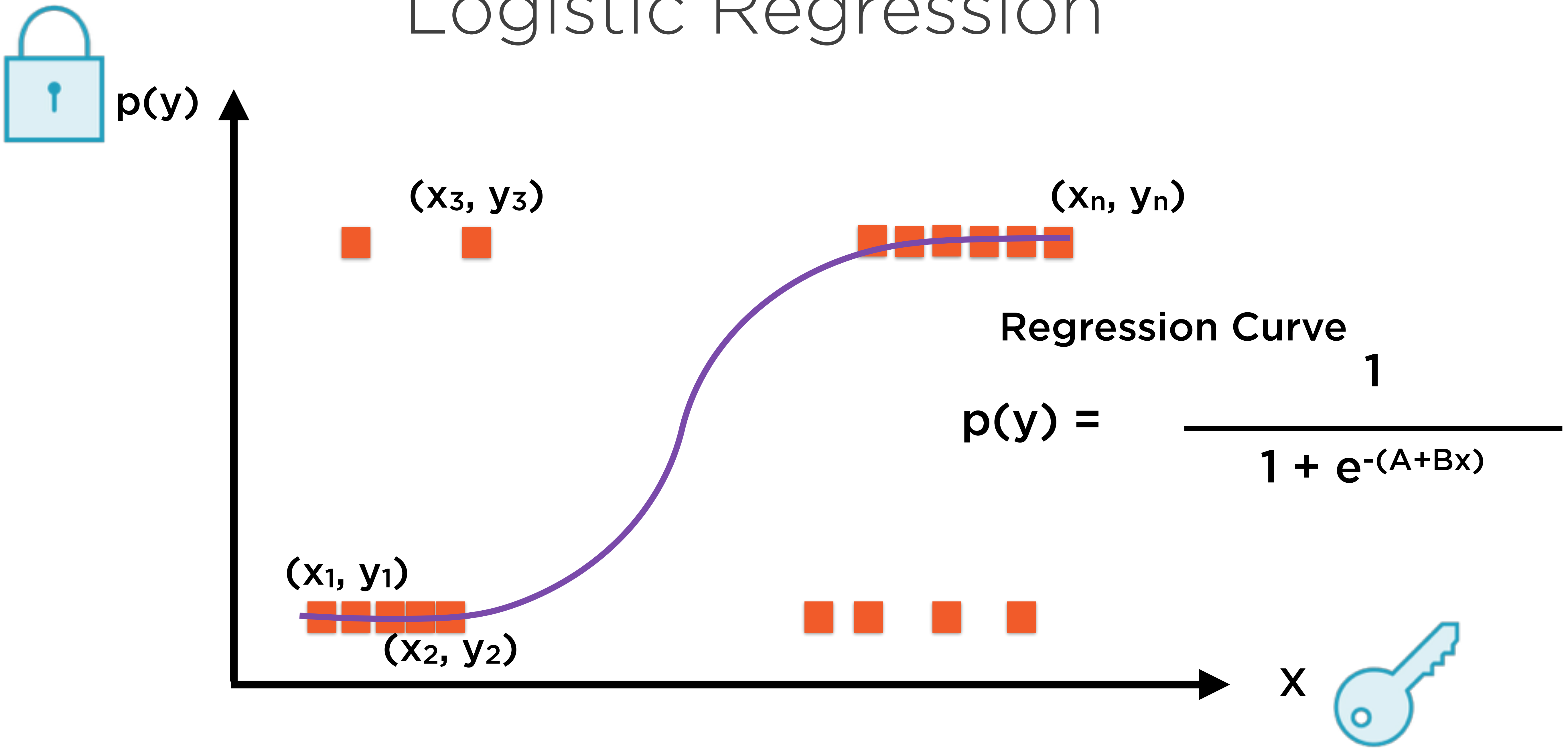
Logistic Regression fits an **S-curve** to estimate how probabilities of categorical variables are influenced by causes

Solving the Logistic Regression Problem via Maximum Likelihood Estimation (MLE)

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

Logistic Regression

Regression Equation:

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

Solve for A and B that “best fit” the data

A Thought Experiment

Toss n coins

Head: $y_i = 1$

Tail: $y_i = 0$

Probability of Head = p_i

Probability of Tail = $1-p_i$

A Thought Experiment

Coin i	Result	y_i	Probability
1	Heads	1	p_1
2	Tails	0	$1-p_2$
3	Heads	1	p_3
4	Heads	1	p_4
5	Tails	0	$1-p_5$
6	Tails	0	$1-p_6$
7	Heads	1	p_7
8	Heads	1	p_8
9	Heads	1	p_9
...	...	...	...
n	Tails	0	$1-p_n$

A Thought Experiment

Probability of independent events = product of individual probabilities

A Thought Experiment

Overall likelihood of getting these results

$$L = (p_1)*(1-p_2)*(p_3)*(p_4)*(1-p_5)\ldots*(1-p_n)$$

A Thought Experiment

Conveniently combine probabilities of head or tail into one expression

Outcome of coin i = $p_i^{y_i}(1-p_i)^{1-y_i}$

If outcome = Head

$y_i = 1$

$$p_i^{y_i}(1-p_i)^{1-y_i} = p_i^{1}(1-p_i)^{0} = p_i$$

If outcome = Tail

$y_i = 0$

$$p_i^{y_i}(1-p_i)^{1-y_i} = p_i^{0}(1-p_i)^{1} = 1 - p_i$$

Tossing n Coins

$$L = (p_1) * (1-p_2) * (p_3) * (p_4) * (1-p_5) \ldots * (1-p_n)$$

$$p_i^{y_i}(1-p_i)^{1-y_i}$$

$$L = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

Π denotes product of multiple terms

Tossing n Coins

$$L = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

Transform equation by taking natural log (ln)

$$LL = \ln L = \sum_{i=1}^{n} [y_i \ln(p_i) + (1-y_i)\ln(1-p_i)]$$

$\sum$ denotes sum of multiple terms

Logistic Regression

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

Solve for A and B that “best fit” the data

Tossing n Coins

$$LL = \ln L = \sum_{i=1}^{n} [y_i \ln(p_i) + (1-y_i)\ln(1-p_i)]$$

The "best fit" values of A and B are those that maximise this likelihood

Maximum Likelihood Estimation (MLE)

Solving the Logistic Regression Problem via Linear Regression

A Thought Experiment

Toss n coins

Head: $y_i = 1$

Tail: $y_i = 0$

Probability of Head = p_i

Probability of Tail = $1-p_i$

A Thought Experiment

Coin i	Result	y_i	Probability
1	Heads	1	p_1
2	Tails	0	$1-p_2$
3	Heads	1	p_3
4	Heads	1	p_4
5	Tails	0	$1-p_5$
6	Tails	0	$1-p_6$
7	Heads	1	p_7
8	Heads	1	p_8
9	Heads	1	p_9
…	…	…	…
n	Tails	0	$1-p_n$

A Thought Experiment

Coin i	Result	y_i	x_i	Probability
1	Heads	1	x_1	p_1
2	Tails	0	x_2	$1-p_2$
3	Heads	1	x_3	p_3
4	Heads	1	x_4	p_4
5	Tails	0	x_5	$1-p_5$
6	Tails	0	…	$1-p_6$
7	Heads	1	…	p_7
8	Heads	1	…	p_8
9	Heads	1	…	p_9
…	…	…	…	…
n	Tails	0	x_n	$1-p_n$

A Thought Experiment

Coin i	Result	y_i	x_i	Probability
1	Heads	1	x_1	p_1
2	Tails	0	x_2	$1-p_2$
3	Heads	1	x_3	p_3
4	Heads	1	x_4	p_4
5	Tails	0	x_5	$1-p_5$
6	Tails	0	...	$1-p_6$
7	Heads	1	...	p_7
8	Heads	1	...	p_8
9	Heads	1	...	p_9
...	...	...	...	...
n	Tails	0	x_n	$1-p_n$

A Thought Experiment

Coin i	Result	y_i	x_i	Probability
1	Heads	1	x_1	p_1
2	Tails	0	x_2	$1-p_2$
3	Heads	1	x_3	p_3
4	Heads	1	x_4	p_4
5	Tails	0	x_5	$1-p_5$

A Thought Experiment

Coin i	Result	y_i	x_i	Probability
1	Heads	1	x_1	p_1
2	Tails	0	x_1	$1-p_1$
3	Heads	1	x_1	p_1
4	Heads	1	x_1	p_1
5	Tails	0	x_1	$1-p_1$

A Thought Experiment

Unique x_i	Frequency(y = 1)	Frequency(y = 0)	$p(y_i)$	$1 - p(y_i)$
x_1	3	2	$p_1 = 3/(3+2) = 3/5$	2/5

Collapse these 5 rows into a single row, where the probabilities "fit" the data

If x is continuous, we will need to create ranges of x-values

Frequency Table

Unique x_i	Frequency(y = 1)	Frequency(y = 0)	$p(y_i)$	$1 - p(y_i)$
x_1	3	2	$p_1 = 3/(3+2) = 3/5$	2/5
x_2	8	12	$p_2 = 8/(8+12) = 2/5$	3/5
...	...	...	...	...

Create a frequency table with 1 row for each unique value of x

Frequency Table

Unique x_i	Frequency(y = 1)	Frequency(y = 0)	$p(y_i)$	$1 - p(y_i)$
x_1	3	2	$p_1 = 3/(3+2) = 3/5$	2/5
x_2	8	12	$p_2 = 8/(8+12) = 2/5$	3/5
...	...	...	...	...

Now, unlike with the MLE approach, each p_i is a continuous variable

Odds from Probabilities

$$\text{Odds}(p) = \frac{p}{1-p}$$

Odds of an Event

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

$$p = \frac{e^{A + Bx}}{1 + e^{A + Bx}}$$

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

Odds of an Event

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

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

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

Odds of an Event

$$p = \frac{e^{A + Bx}}{1 + e^{A + Bx}}$$

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

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

Logit Is Linear

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

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

ln(Odds(p)) is called the logit function

Logit Is Linear

$$\ln \text{Odds}(p) = \ln(p) - \ln(1-p)$$

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

$$\text{logit}(p) = \ln \text{Odds}(p) = A + Bx$$

This is a linear function!

Logit Is Linear

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

$$\text{logit}(p_1) = A + Bx_1$$

$$\text{logit}(p_2) = A + Bx_2$$

$$\text{logit}(p_3) = A + Bx_3$$

$$\ldots \qquad \ldots$$

$$\text{logit}(p_n) = A + Bx_n$$

Tossing n Coins - Linear

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

$$\text{logit}(p_1) = A + Bx_1 + \varepsilon_1$$

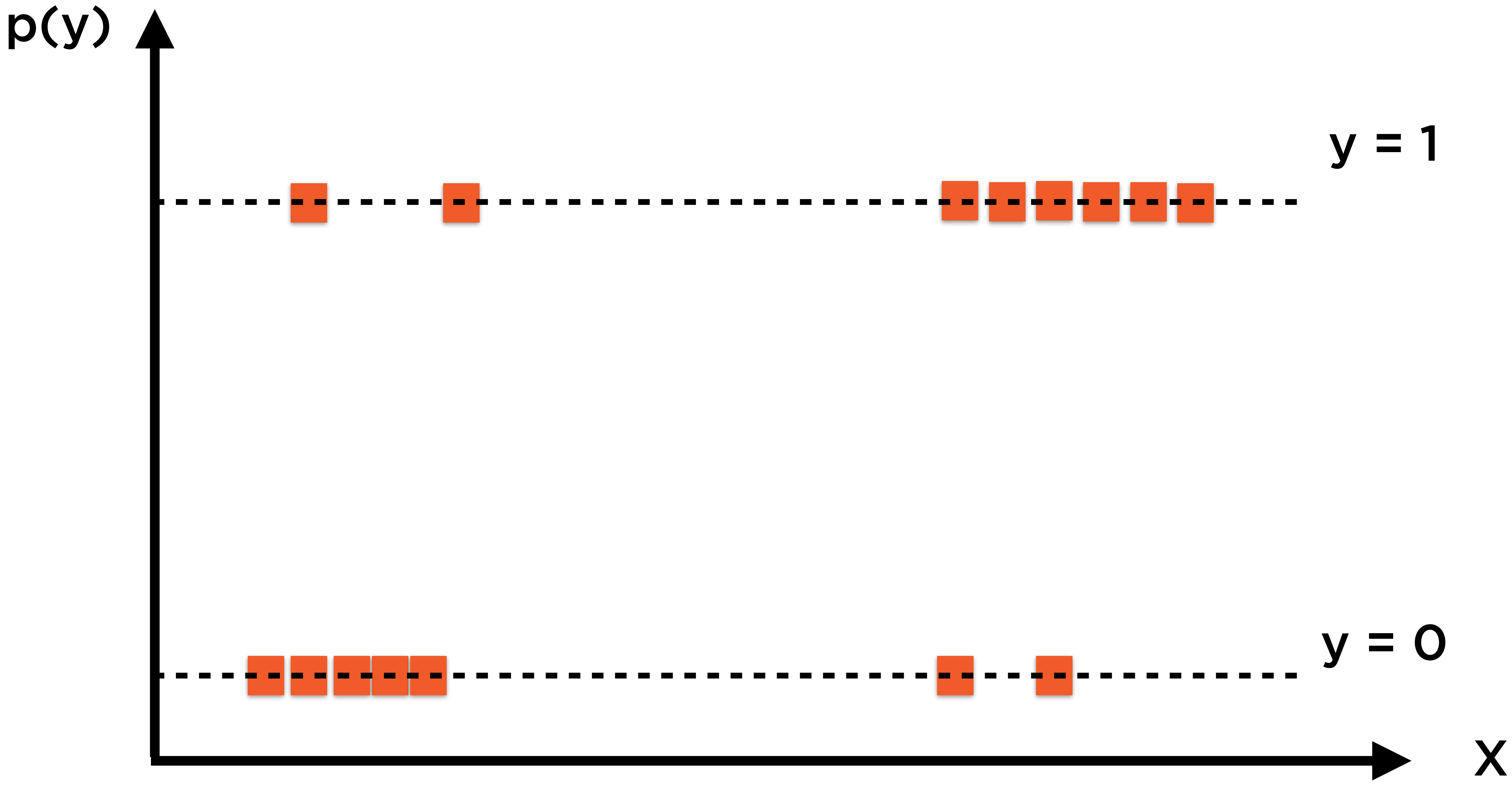
$$\text{logit}(p_2) = A + Bx_2 + \varepsilon_2$$

$$\text{logit}(p_3) = A + Bx_3 + \varepsilon_3$$

$$\ldots \qquad \ldots \qquad \ldots$$

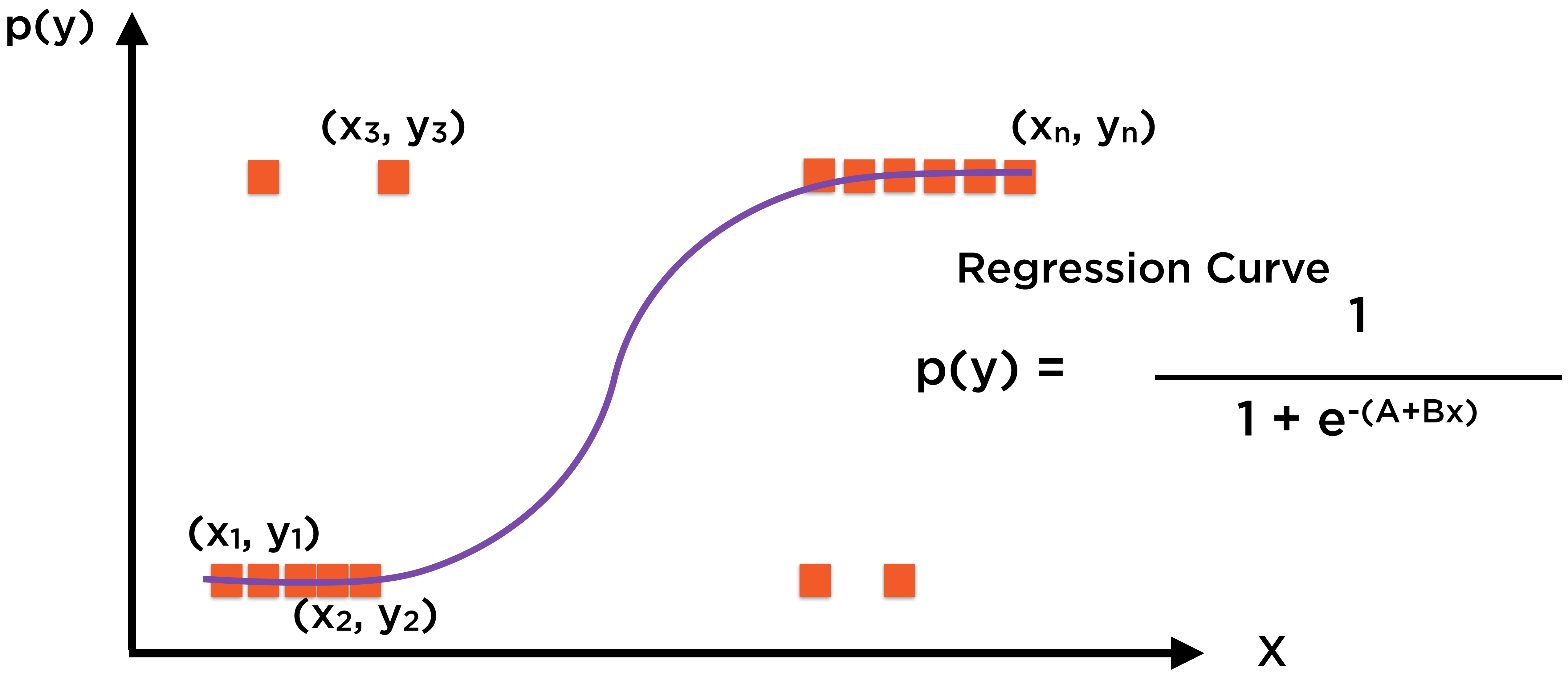
$$\text{logit}(p_n) = A + Bx_n + \varepsilon_n$$

Logistic Regression



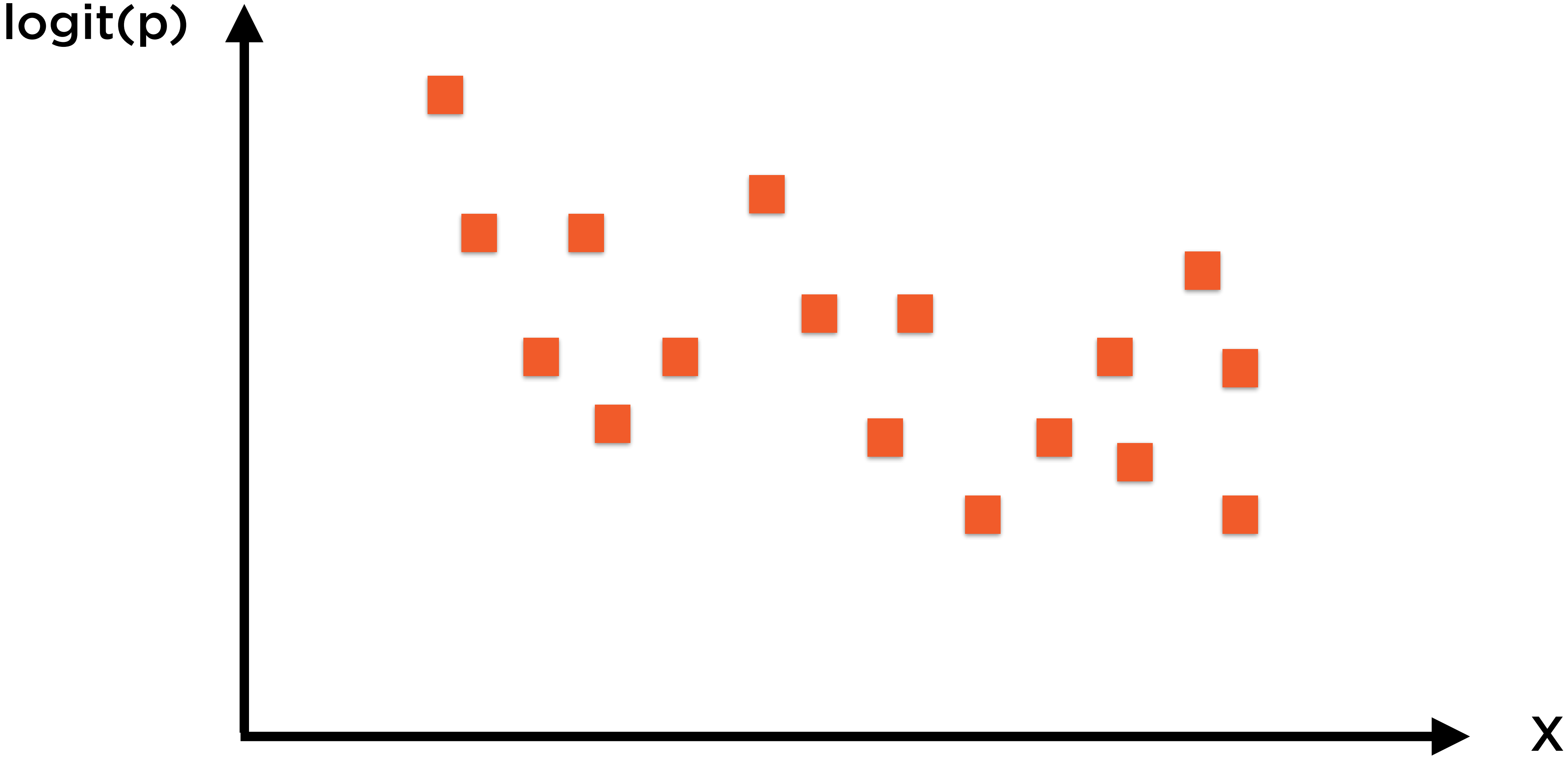
Represent all n points as
(x_i, y_i), where i = 1 to n

Logistic Regression

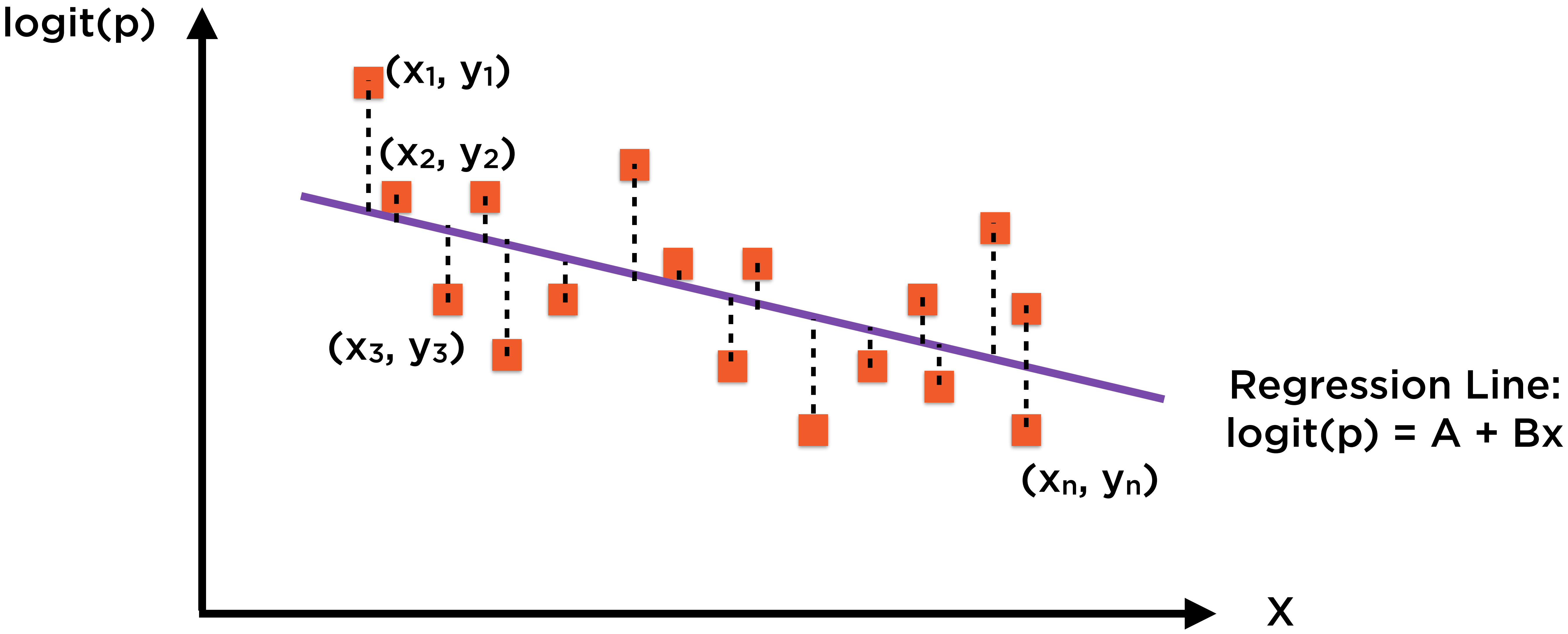


Represent all n points as
(x_i, y_i), where i = 1 to n

Linear Regression



Linear Regression



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

Logistic Regression can be solved via **linear regression on the logit function** (log of the odds function)

Linear Regression Estimation Methods

Method of moments

Method of least squares

Maximum likelihood estimation

Cookie cutter techniques to determine the values of A and B (regression coefficients)

Binomial and Multinomial Logistic Regression

Binomial and Multinomial

Binomial

Two categorical outcomes

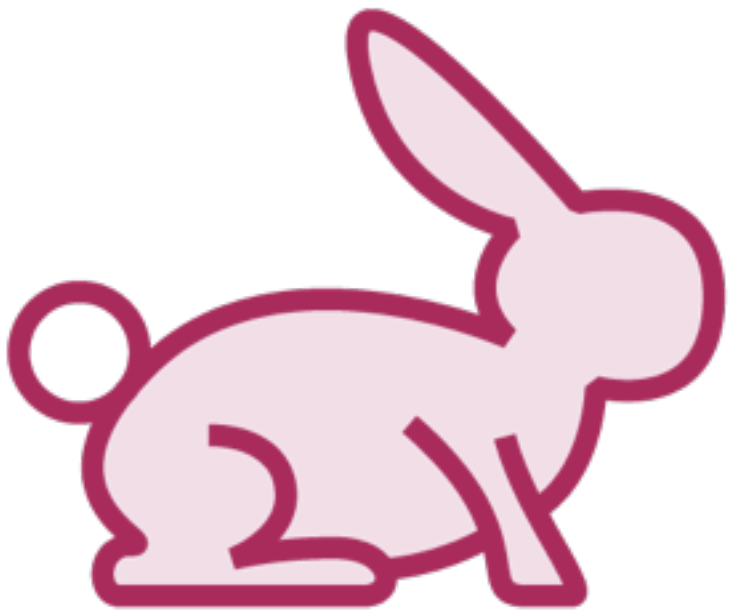
(Head/Tail; True/False)

Multinomial

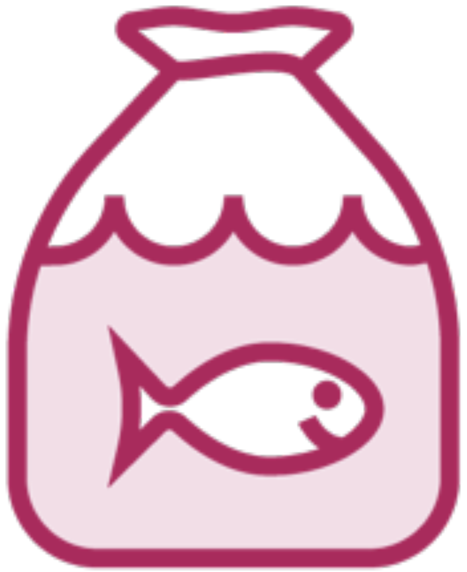
>Two categorical outcomes

(Days in a week; Months in a year)

Binomial Is Binary



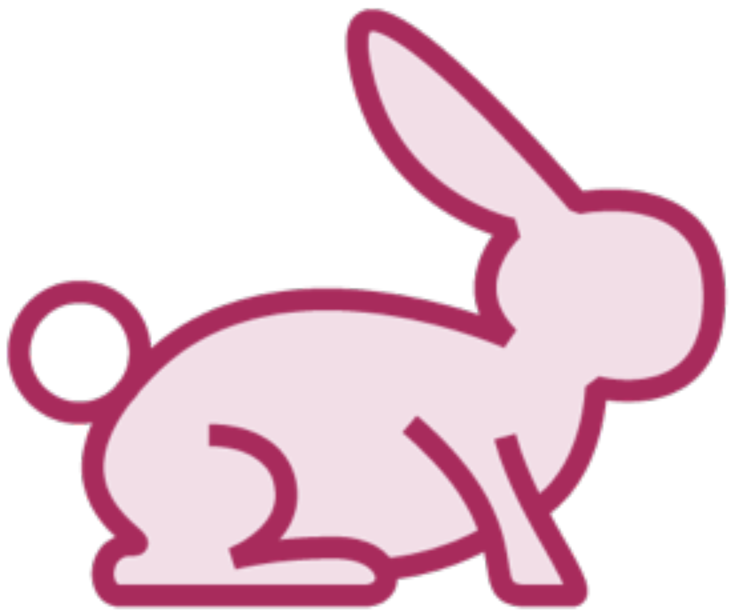
Mammals



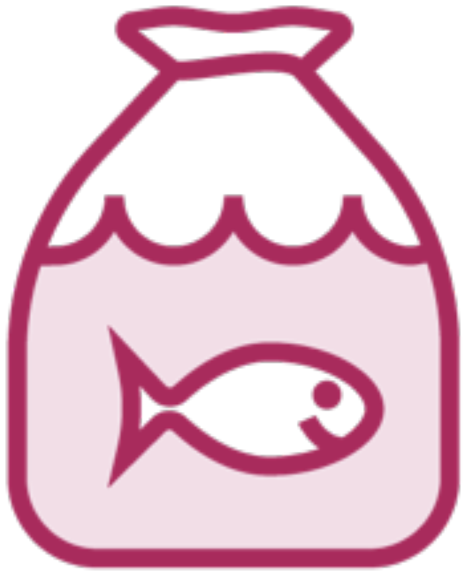
Fish

Whales: Mammals or Fish?

Binomial Is Binary



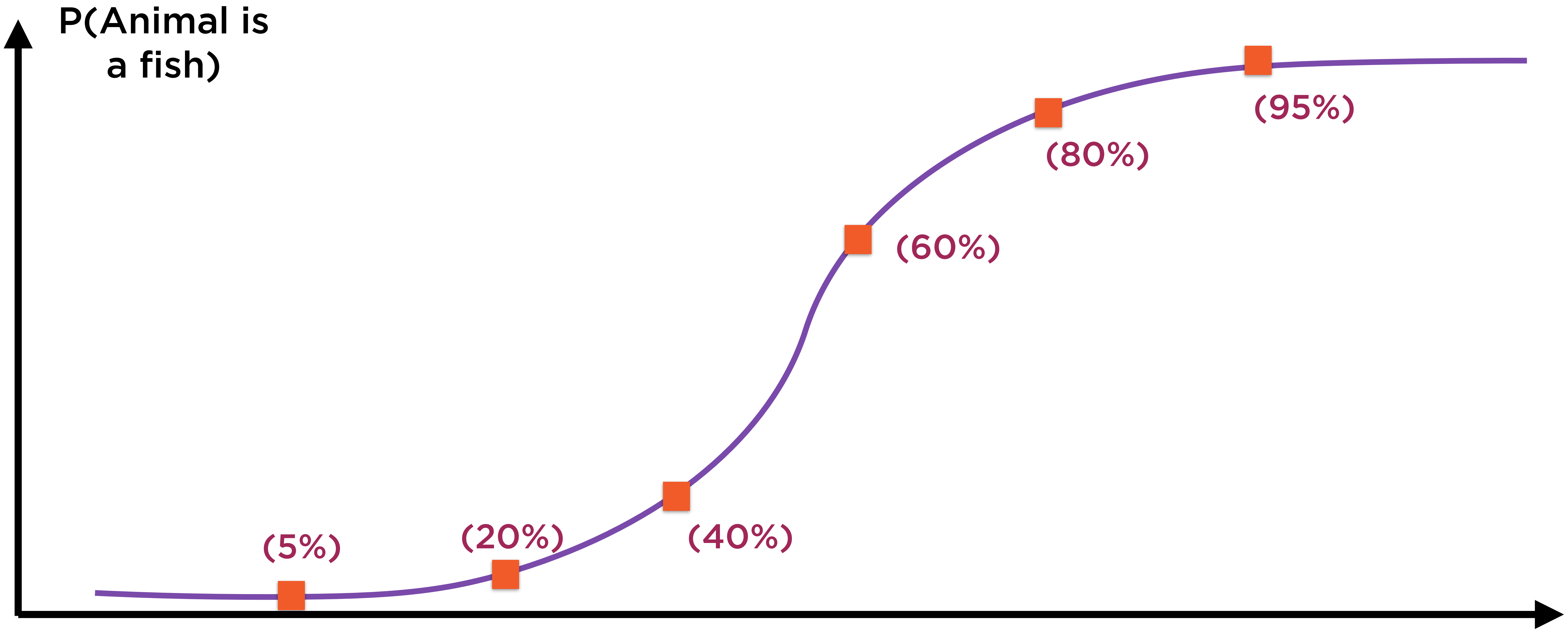
Mammals



Fish

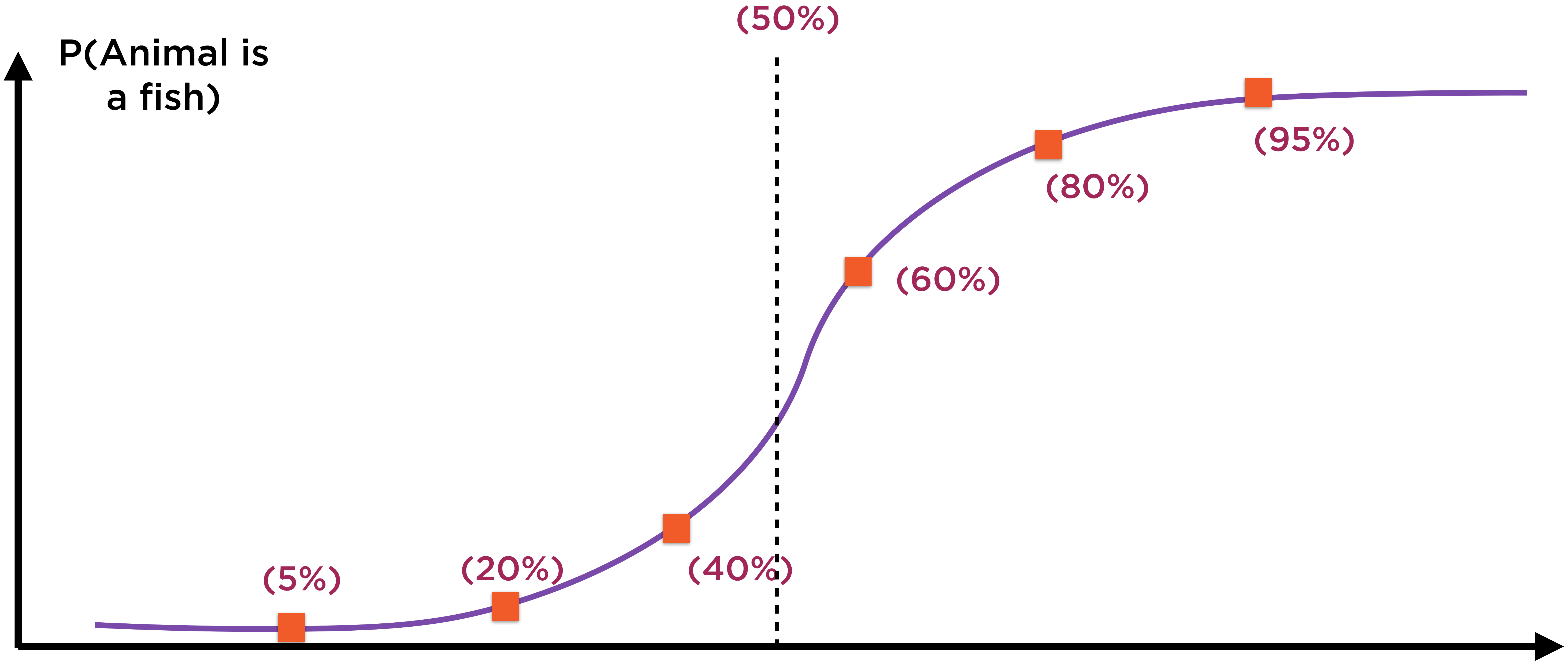
Result	Mammals	Fish
Label (y_i)	0	1
Probability ($p(y_i)$)	p	1-p

Binomial Is Binary



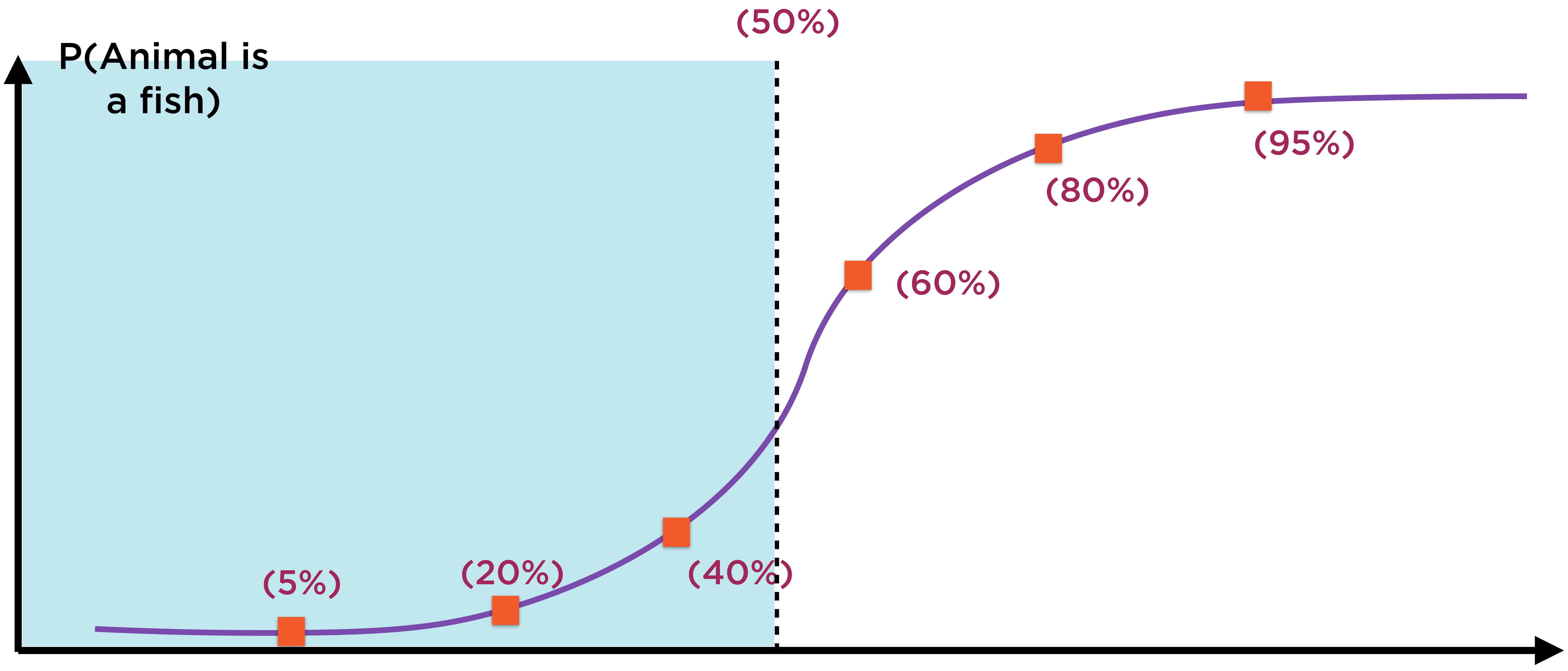
Whales: Mammals or Fish?

Binomial Is Binary



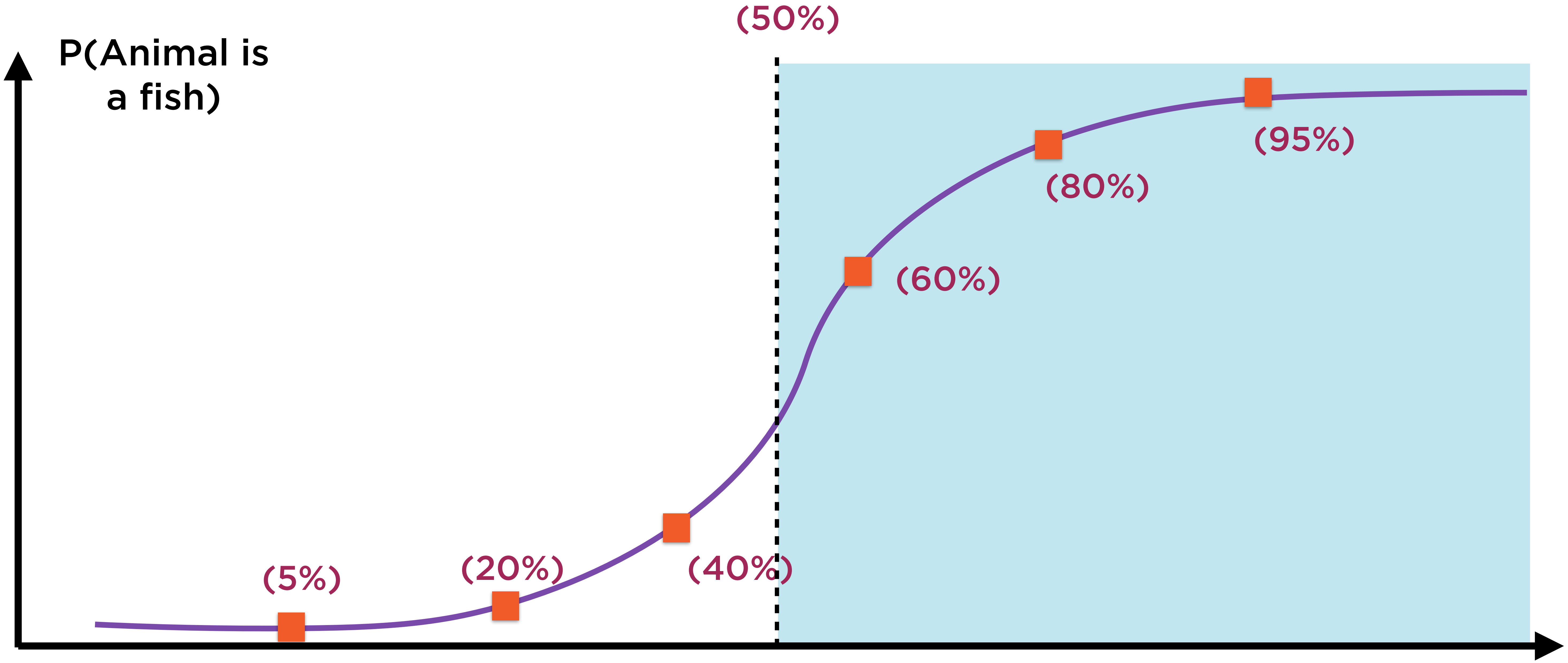
Rule of 50%

Binomial Is Binary



If probability < 50%, it’s a mammal

Binomial Is Binary

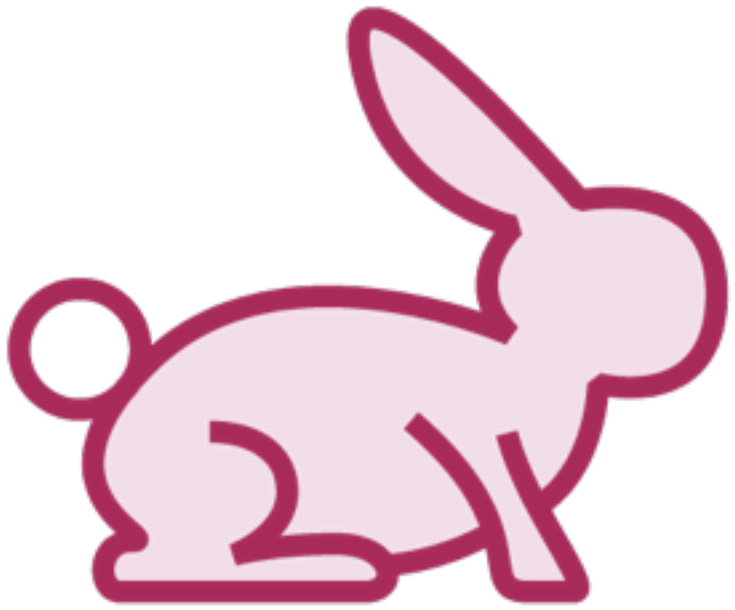


If probability > 50%, it's a fish

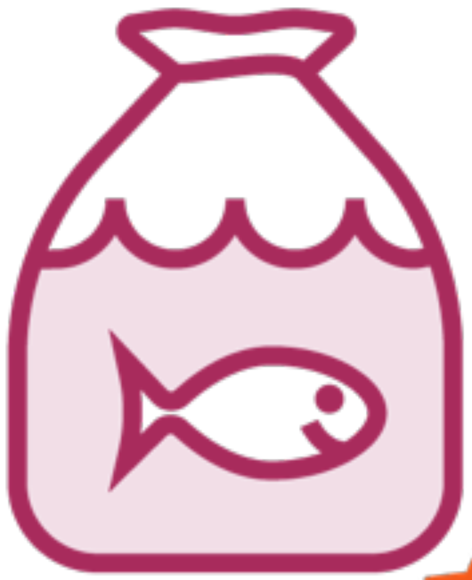
Binomial Is Binary

Probability of whales being Fish < 50%

Binomial Is Binary



Mammals



Fish

Probability of whales being Fish > 50%

Binomial and Multinomial

Binomial

Two categorical outcomes

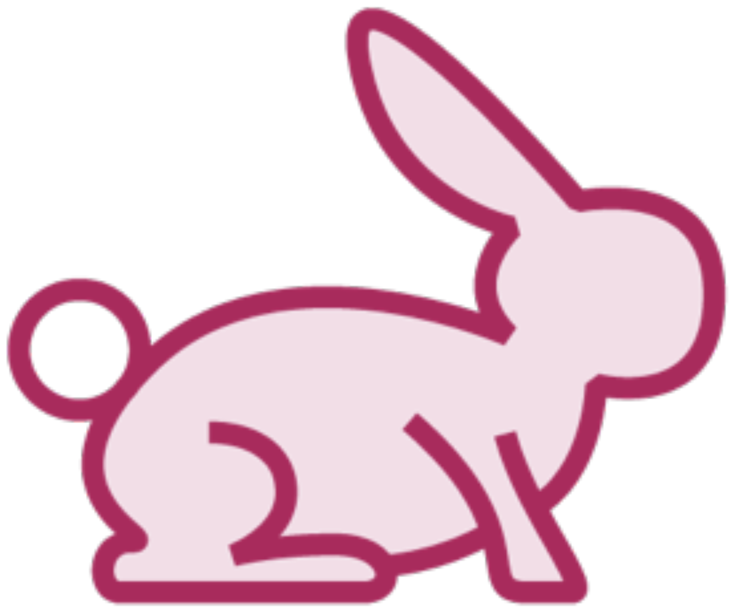
(Head/Tail; True/False)

Multinomial

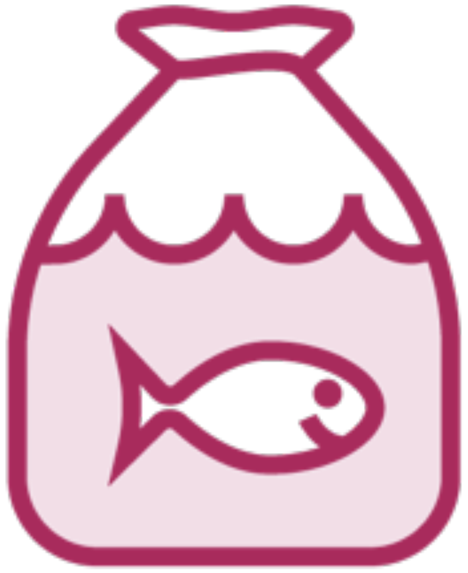
>Two categorical outcomes

(Days in a week; Months in a year)

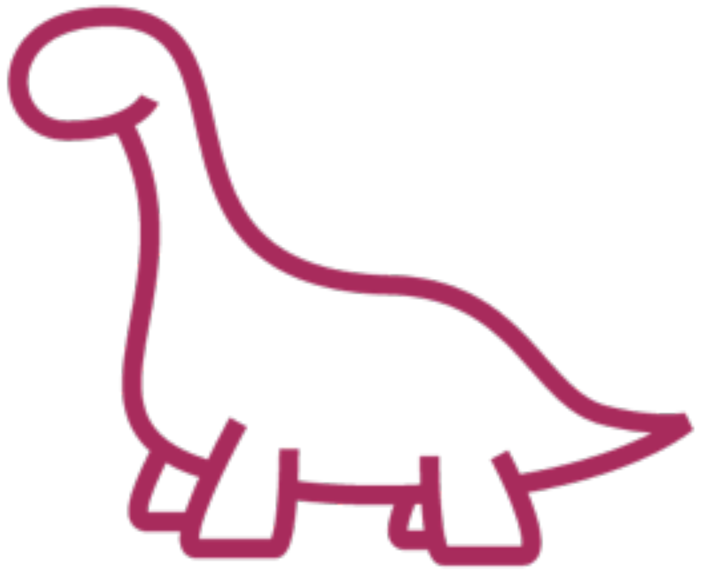
Multinomial Is Non-binary



Mammals



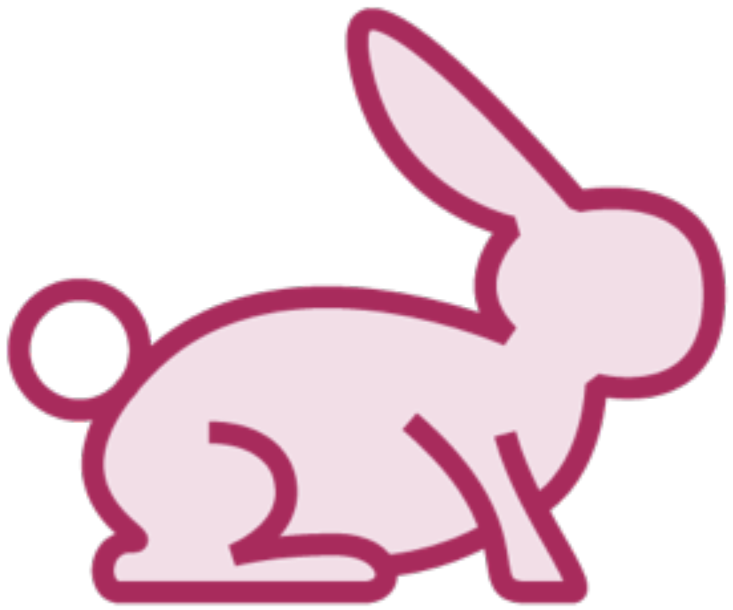
Fish



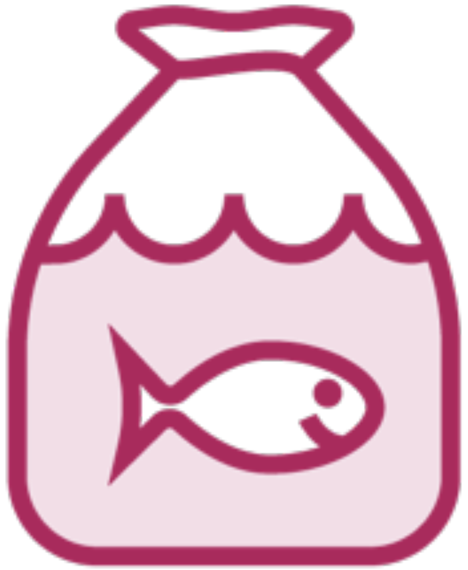
Reptiles

Whales: Mammals or Fish or Reptiles?

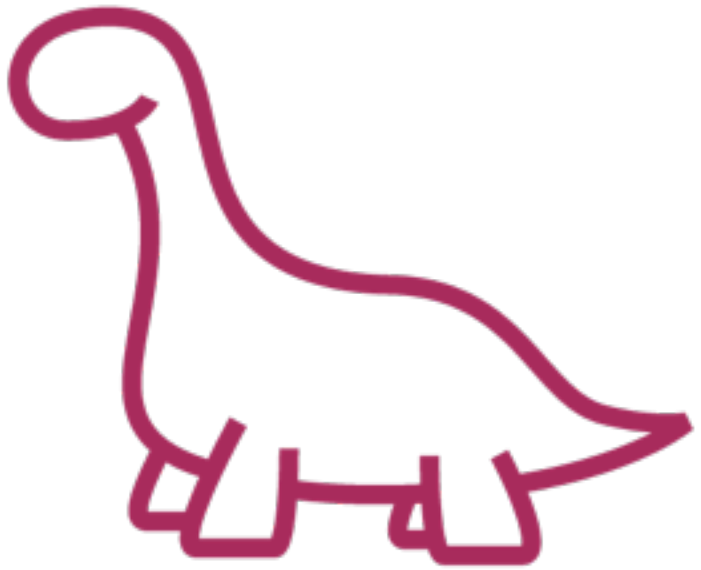
Multinomial Is Non-binary



Mammals



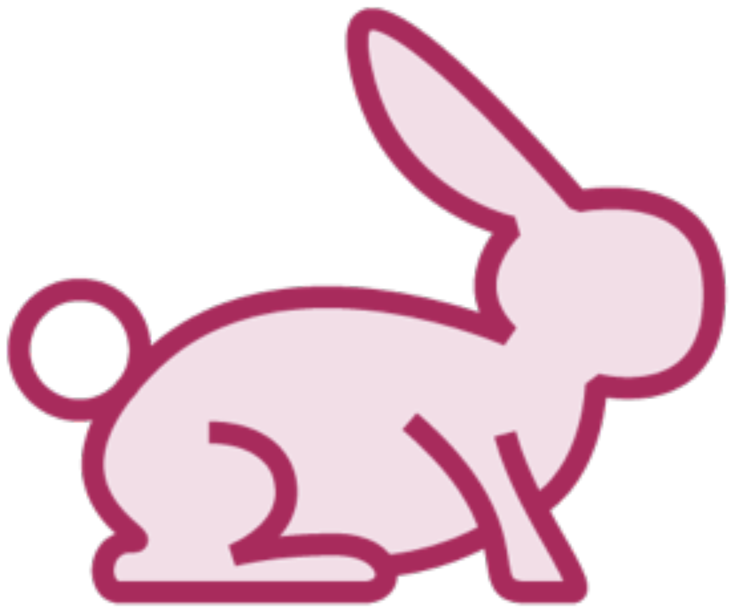
Fish



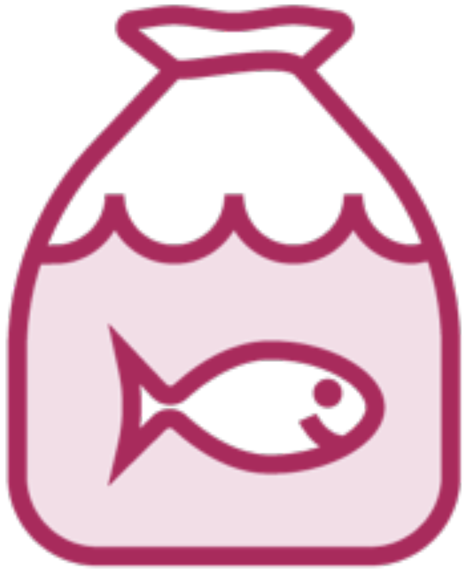
Reptiles

Result	Mammals	Fish	Reptiles
Label (y_i)	0	1	2
Probability ($p(y_i)$)	p_0	p_1	p_2
Probability ($p(y_i')$)	$1-p_0$	$1-p_1$	$1-p_2$

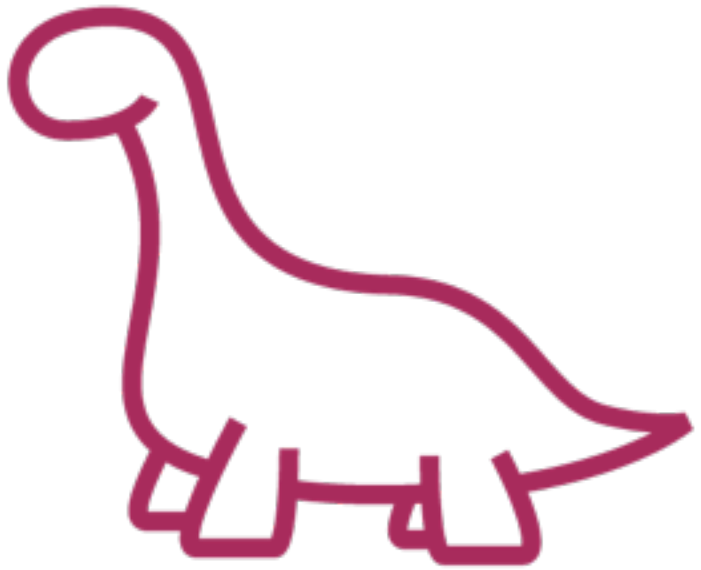
Multinomial Is Non-binary



Mammals



Fish

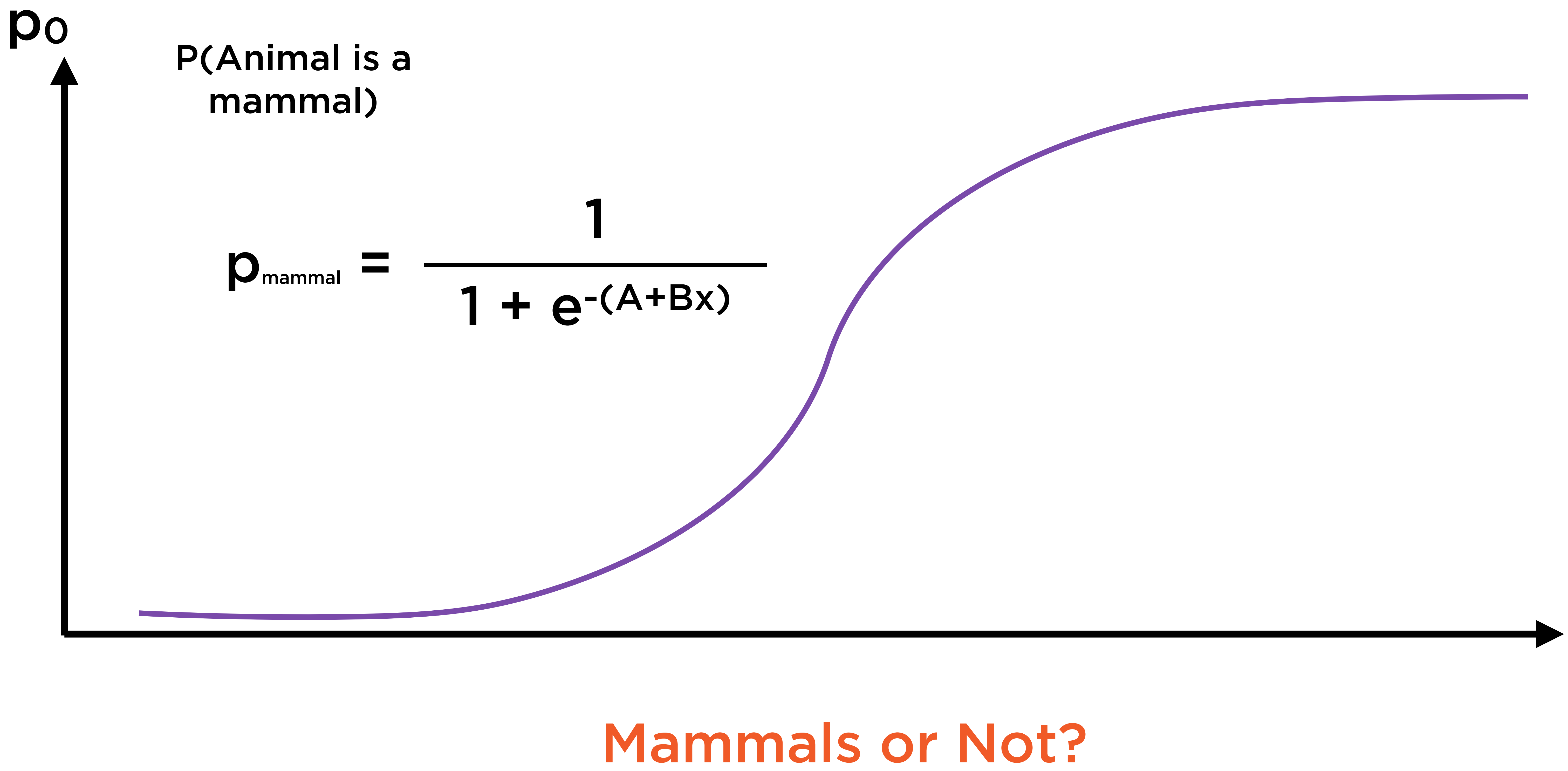


Reptiles

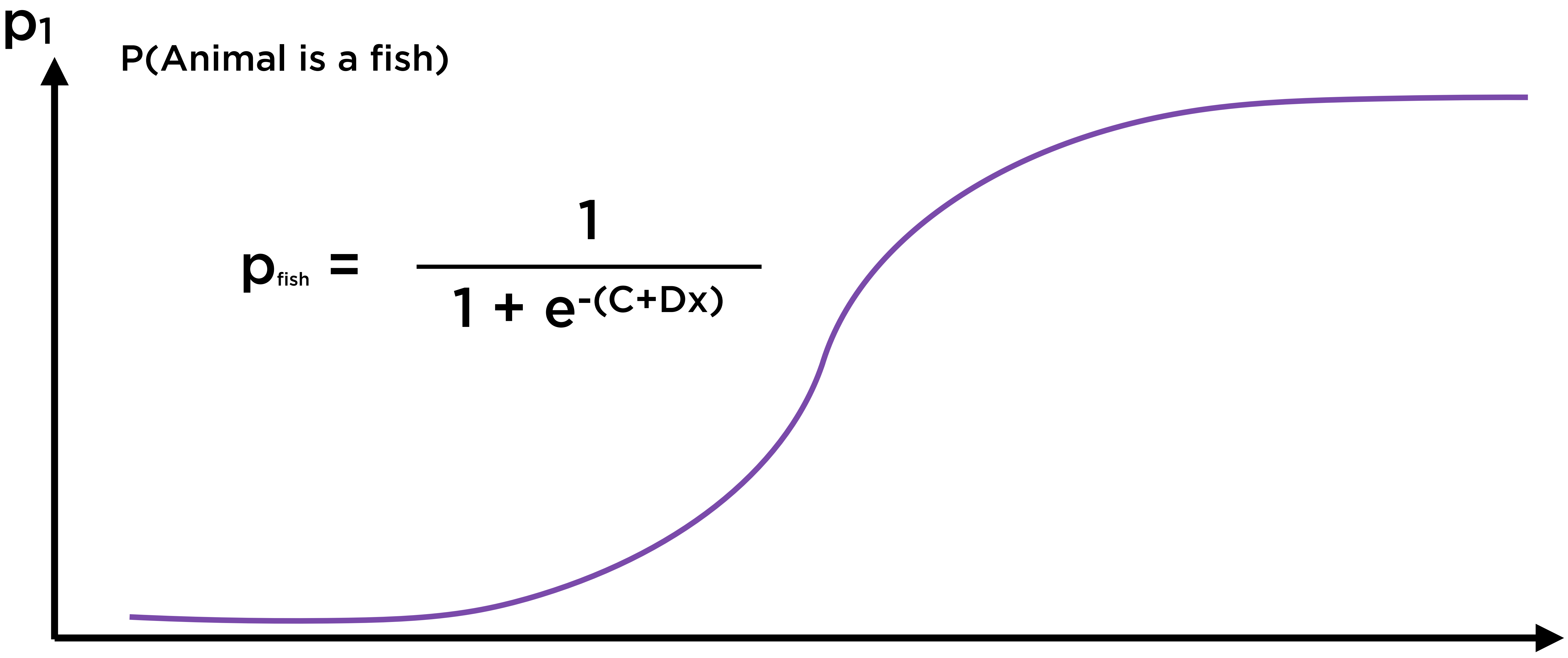
Whales: Mammals or Fish or Reptiles?

Run three logistic regressions

Multinomial Is Non-binary

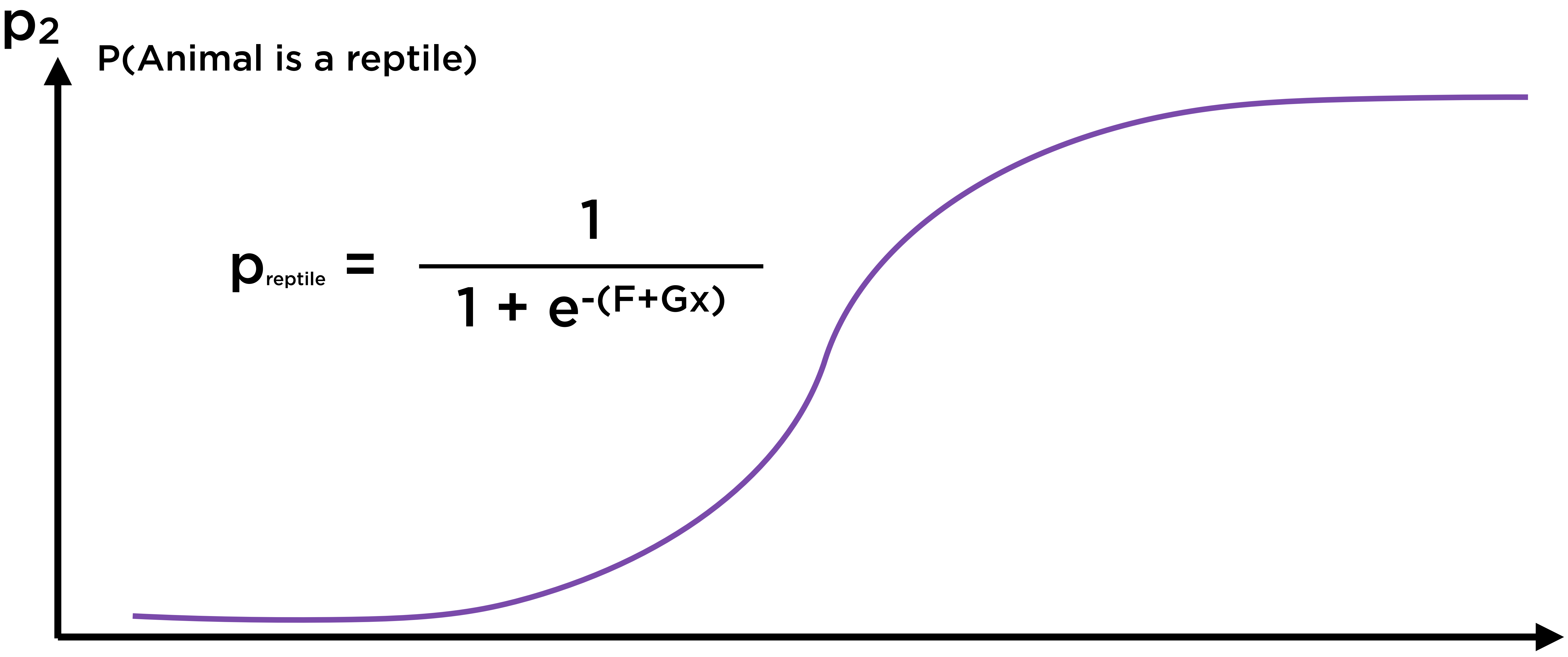


Multinomial Is Non-binary



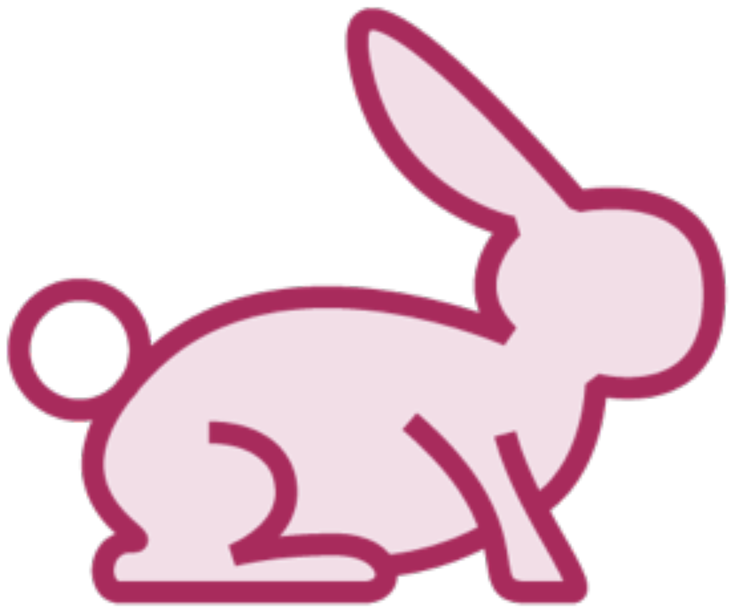
Fish or Not?

Multinomial Is Non-binary

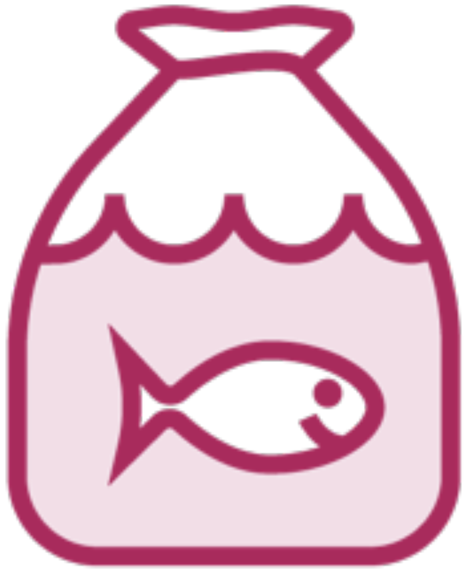


Reptiles or Not?

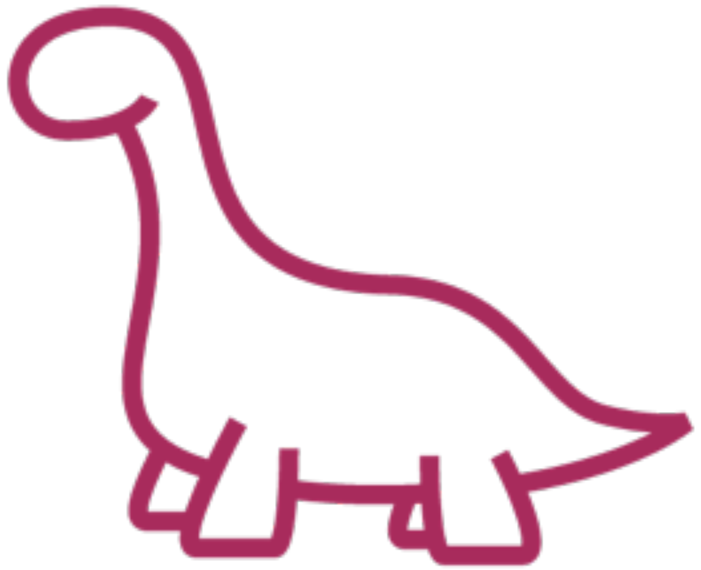
Multinomial Is Non-binary



Mammals



Fish

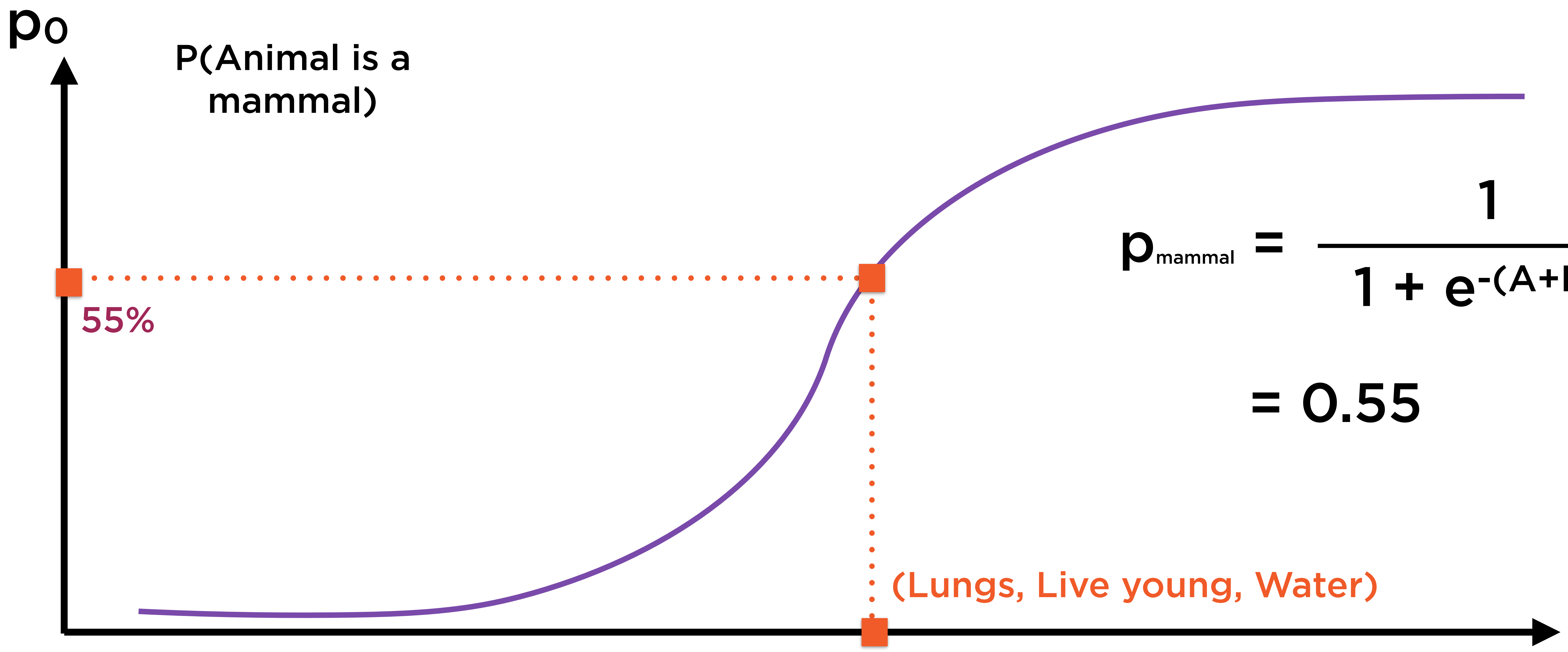


Reptiles

Whales: Mammals or Fish or Reptiles?

Choose the highest probability

Multinomial Is Non-binary

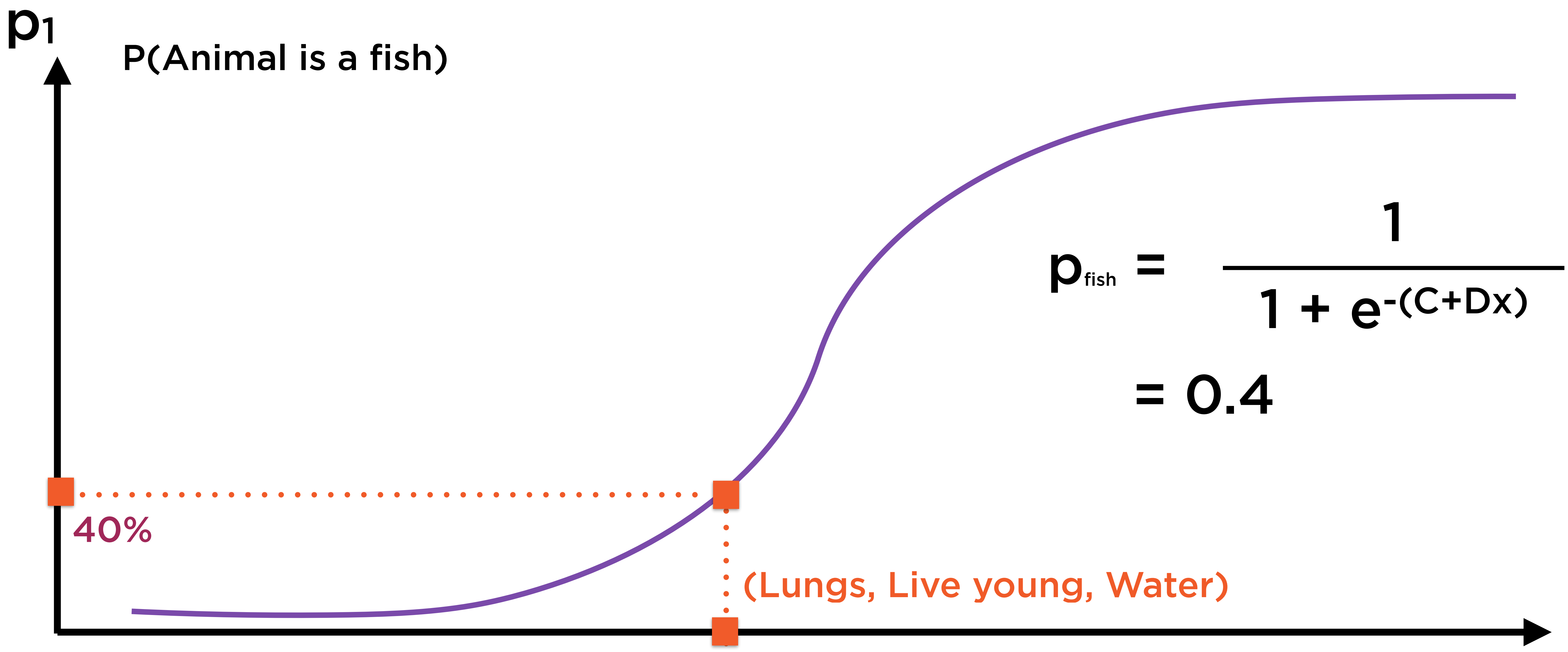


$$p_{mammal} = \frac{1}{1 + e^{-(A+Bx)}}$$

$$= 0.55$$

Mammals or Not?

Multinomial Is Non-binary

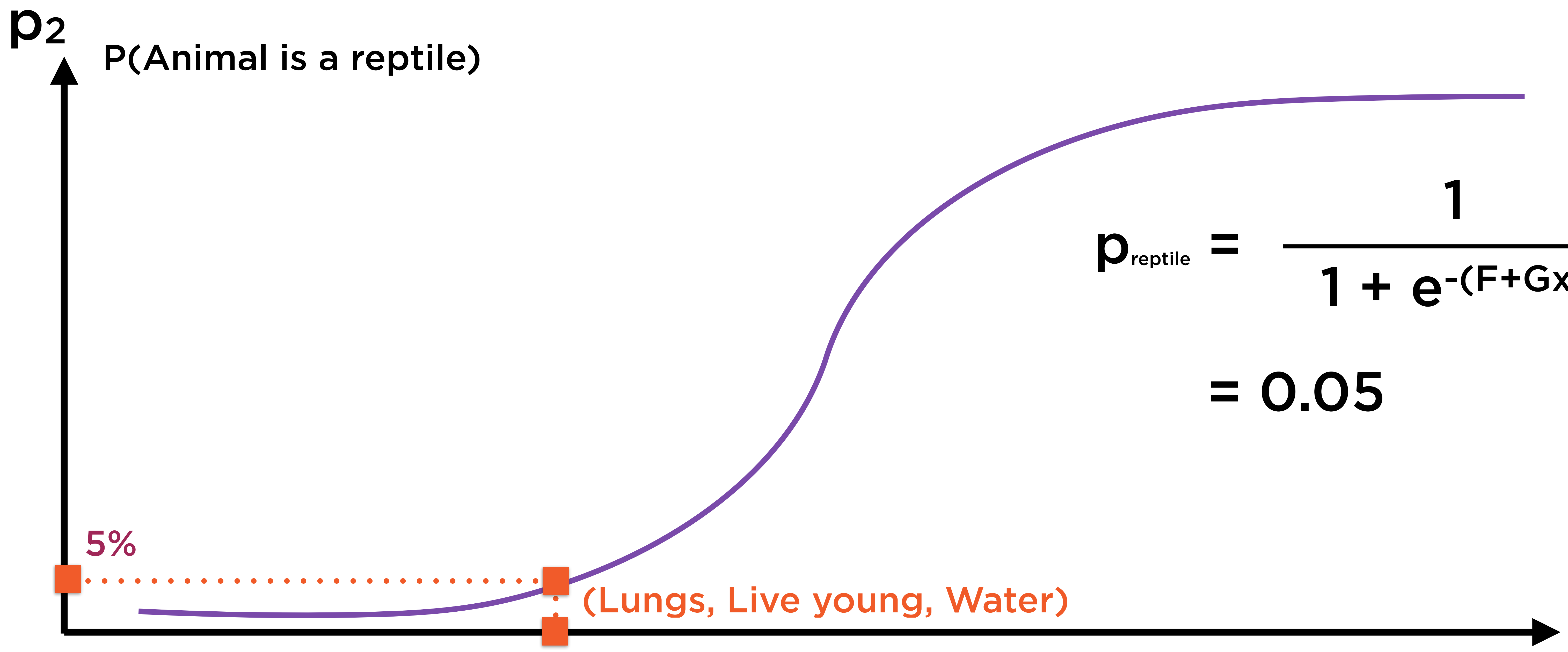


$$p_{fish} = \frac{1}{1 + e^{-(C+Dx)}}$$

$$= 0.4$$

Fish or Not?

Multinomial Is Non-binary



$$p_{reptile} = \frac{1}{1 + e^{-(F+Gx)}}$$

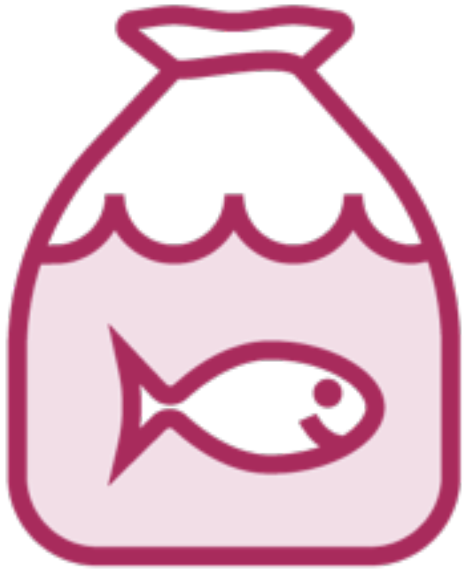
$$= 0.05$$

Reptiles or Not?

Multinomial Is Non-binary

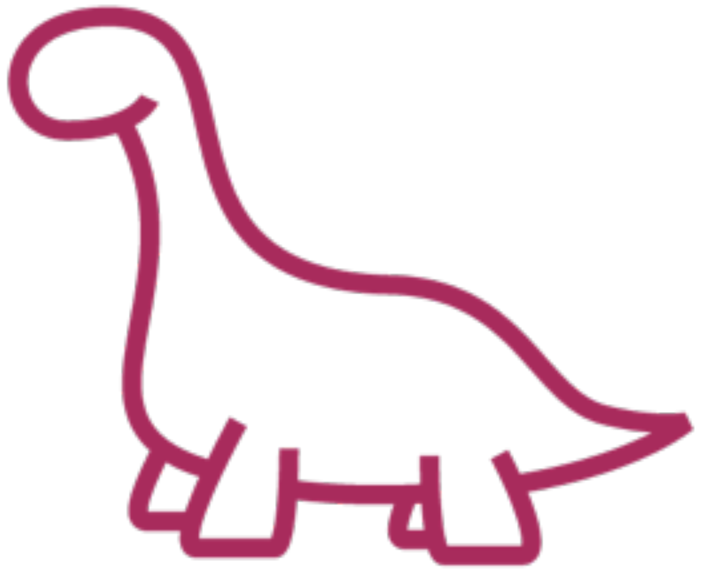
Mammals

$p_{mammal} = 0.55$



Fish

$p_{fish} = 0.4$



Reptiles

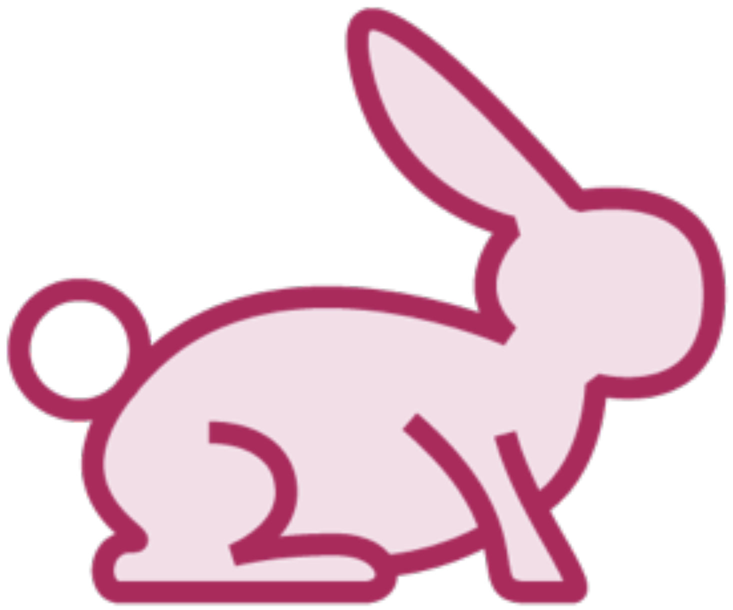
$p_{reptile} = 0.05$

$p_{mammal} > p_{fish} > p_{reptile}$

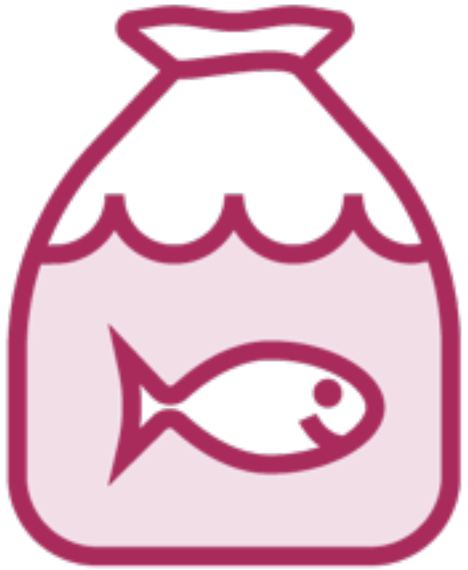
Multinomial Is Non-binary

$$p_{mammal} < p_{fish} > p_{reptile}$$

Multinomial Is Non-binary



Mammals



Fish

Reptiles

$$p_{mammal} < p_{fish} < p_{reptile}$$

Binomial and Multinomial

Binomial	Multinomial
Two categorical outcomes	N categorical outcomes
One logistic regression	N logistic regressions

Regression: Excel, R or Python

Excel

Create a regression slide for an important presentation

R

Create a regression case study for a seminar

Python

Build trading model that scrapes websites, combines sentiment analysis and regression

A simple multinomial logistic regression technique uses N logistic models for N categories