Building a Simple Machine Learning Model with scikit-learn

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Classic problems in machine learning

Regression for predicting continuous data

Classification for predicting categorical data

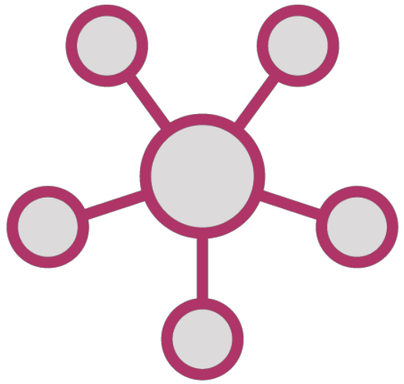
Implementing simple linear and logistic regression in scikit-learn

Building Regression Models

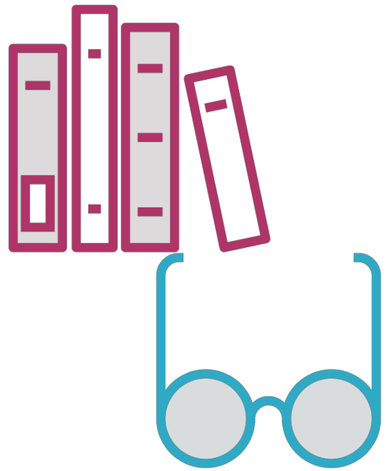
Types of Machine Learning Problems

Classification

Regression



Clustering



Dimensionality reduction

Types of Machine Learning Problems

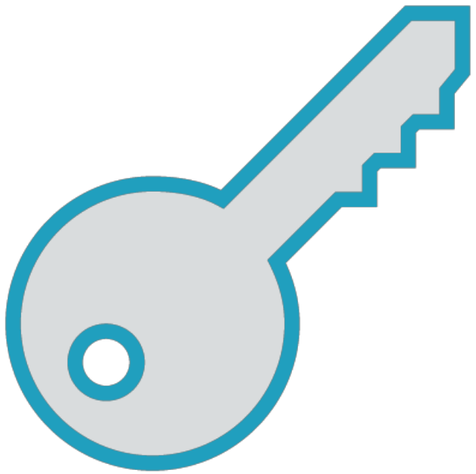
Classification

Regression

Clustering

Dimensionality reduction

X Causes Y



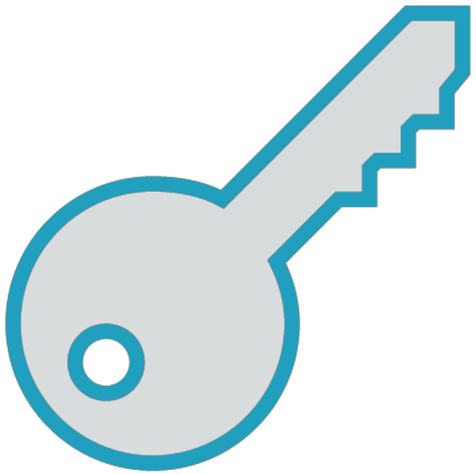
Cause

Independent variable

Effect

Dependent variable

X Causes Y



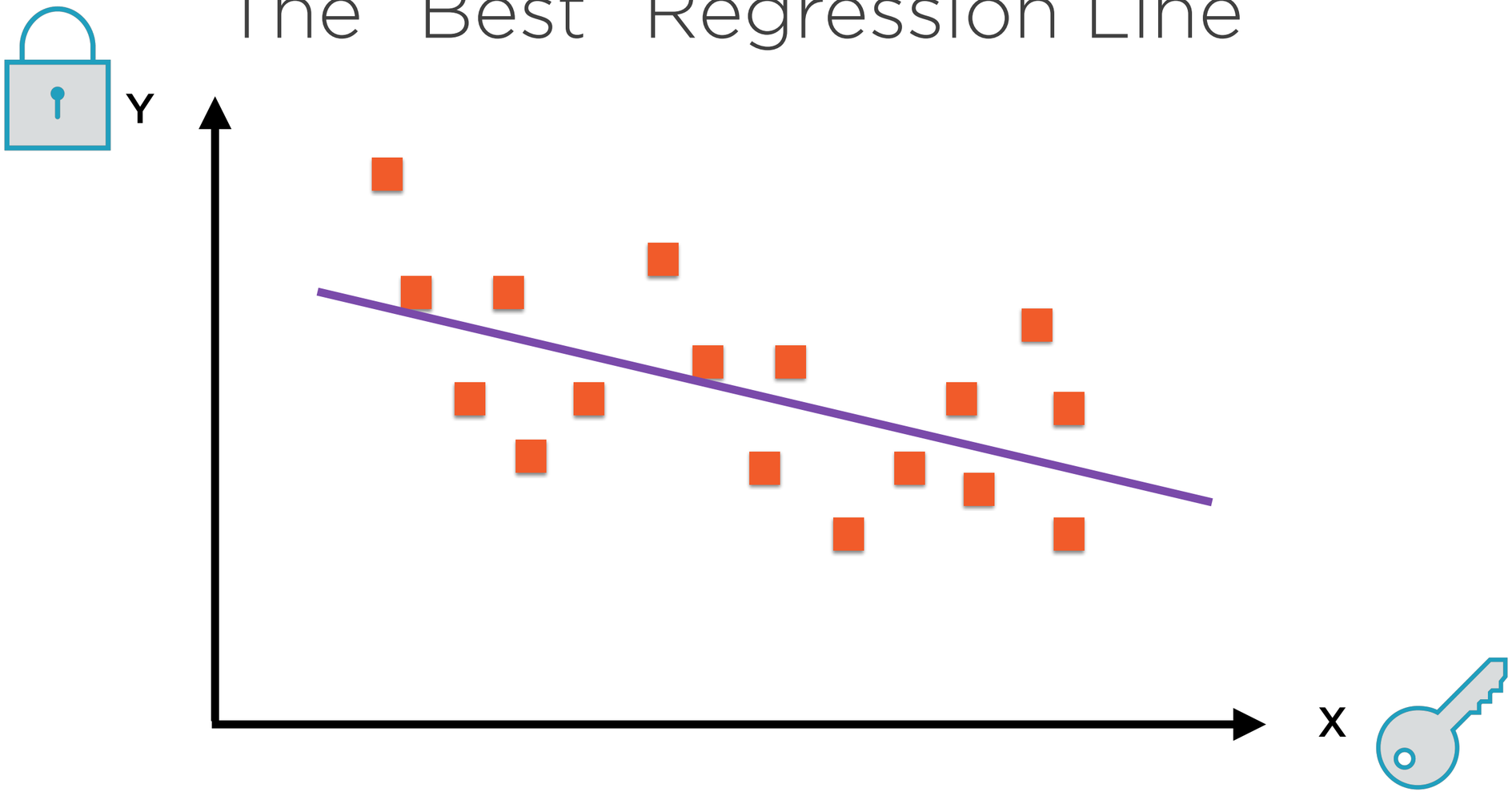
Cause

Explanatory variable

Effect

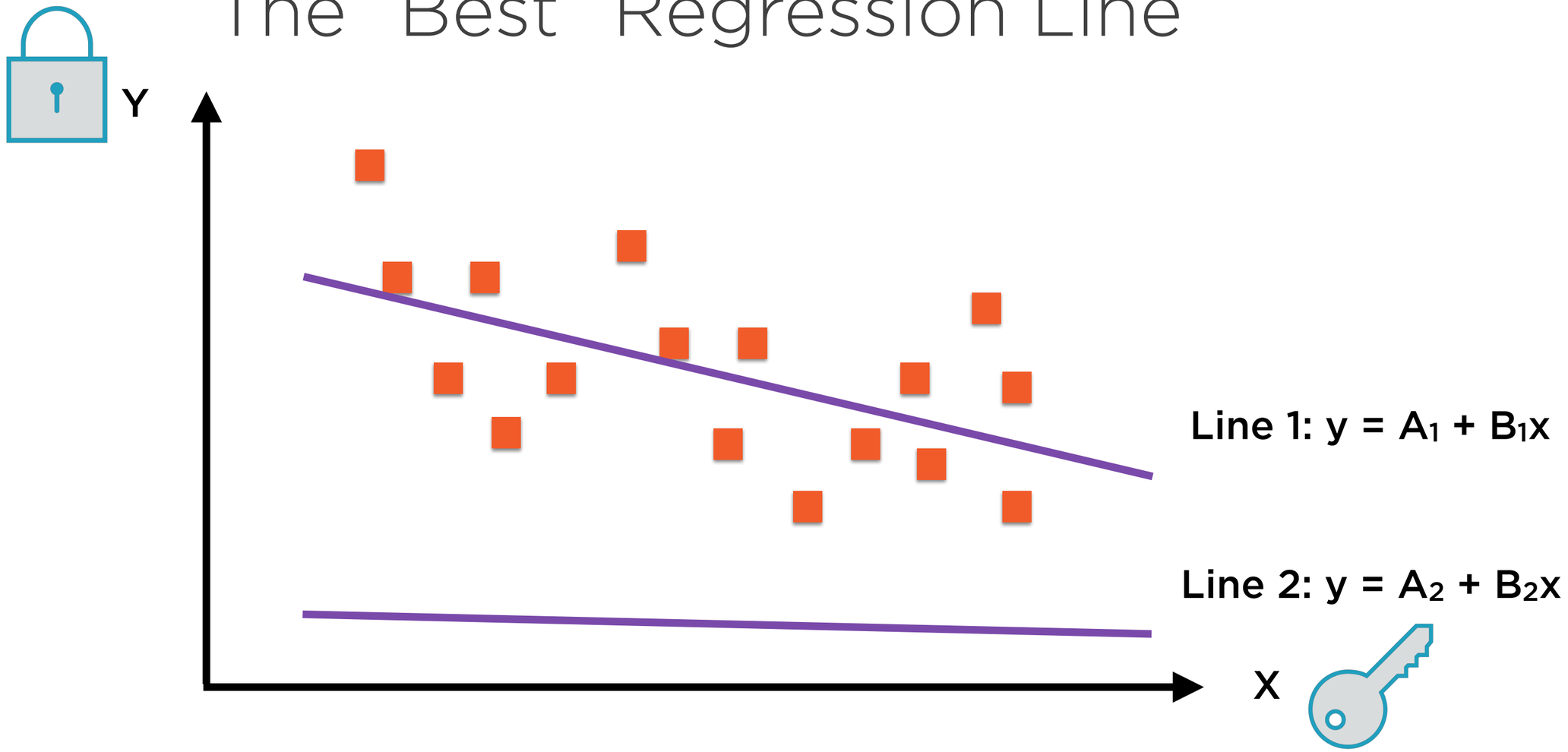
Dependent variable

The “Best” Regression Line



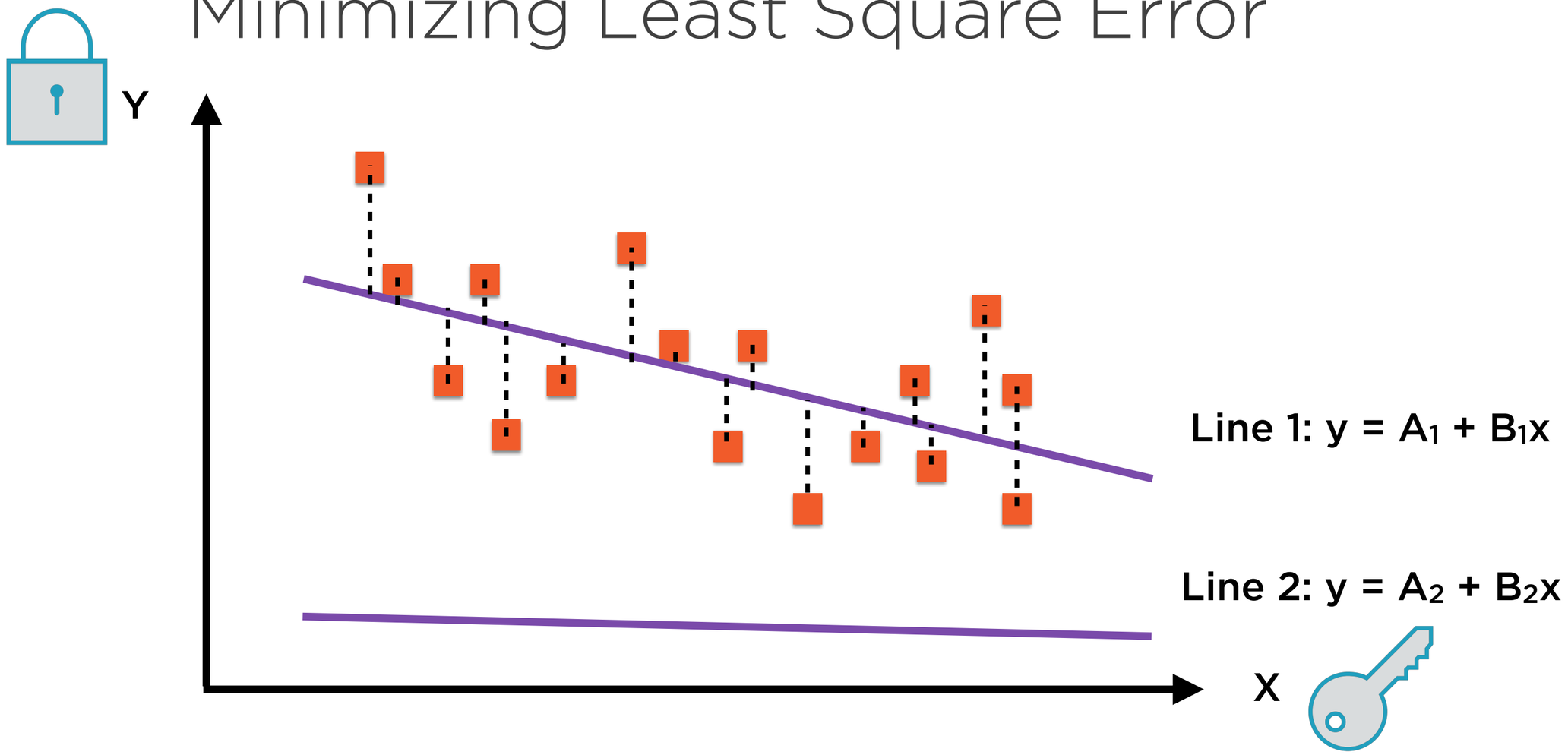
Linear regression involves finding the “best fit” line

The “Best” Regression Line



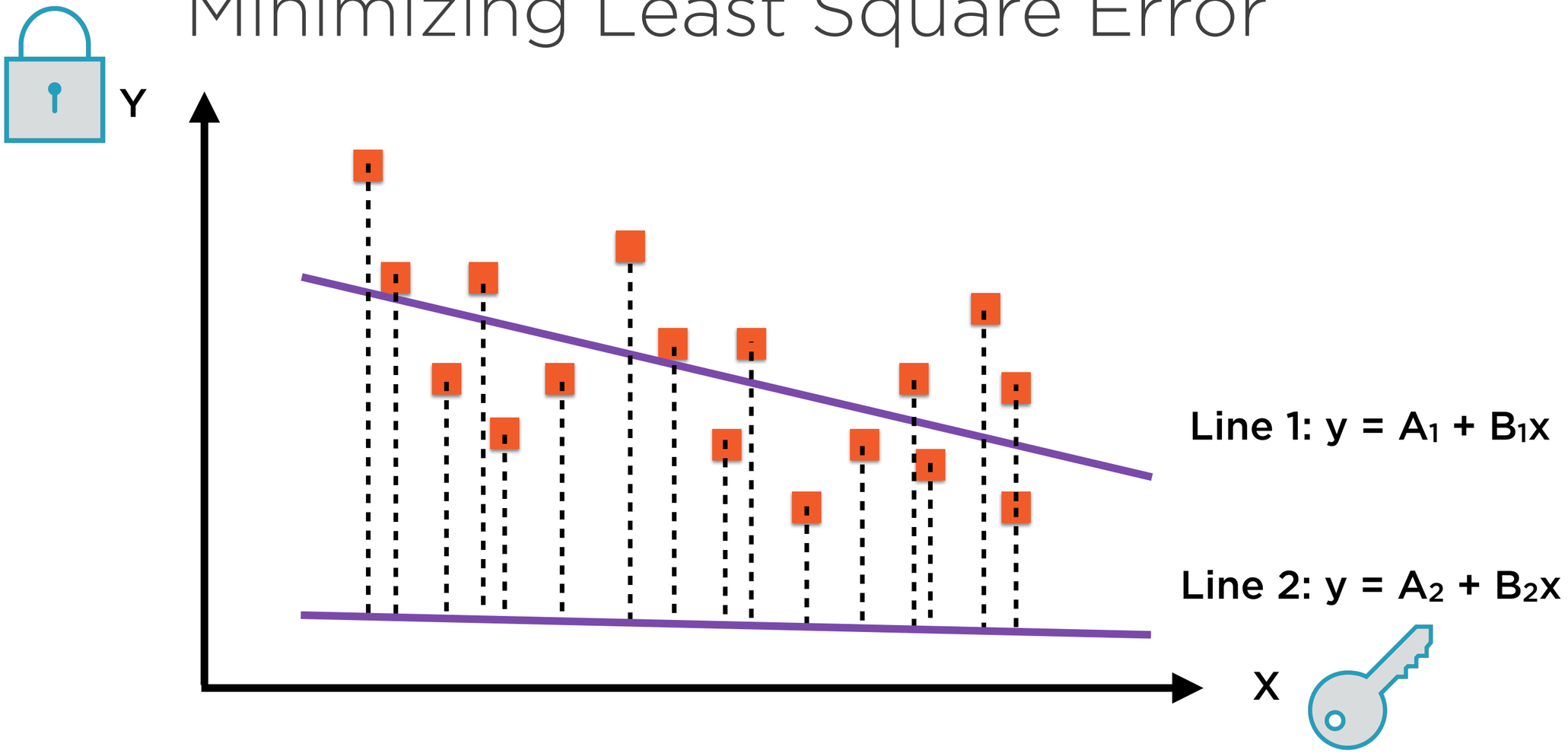
Let’s compare two lines, Line 1 and Line 2

Minimizing Least Square Error



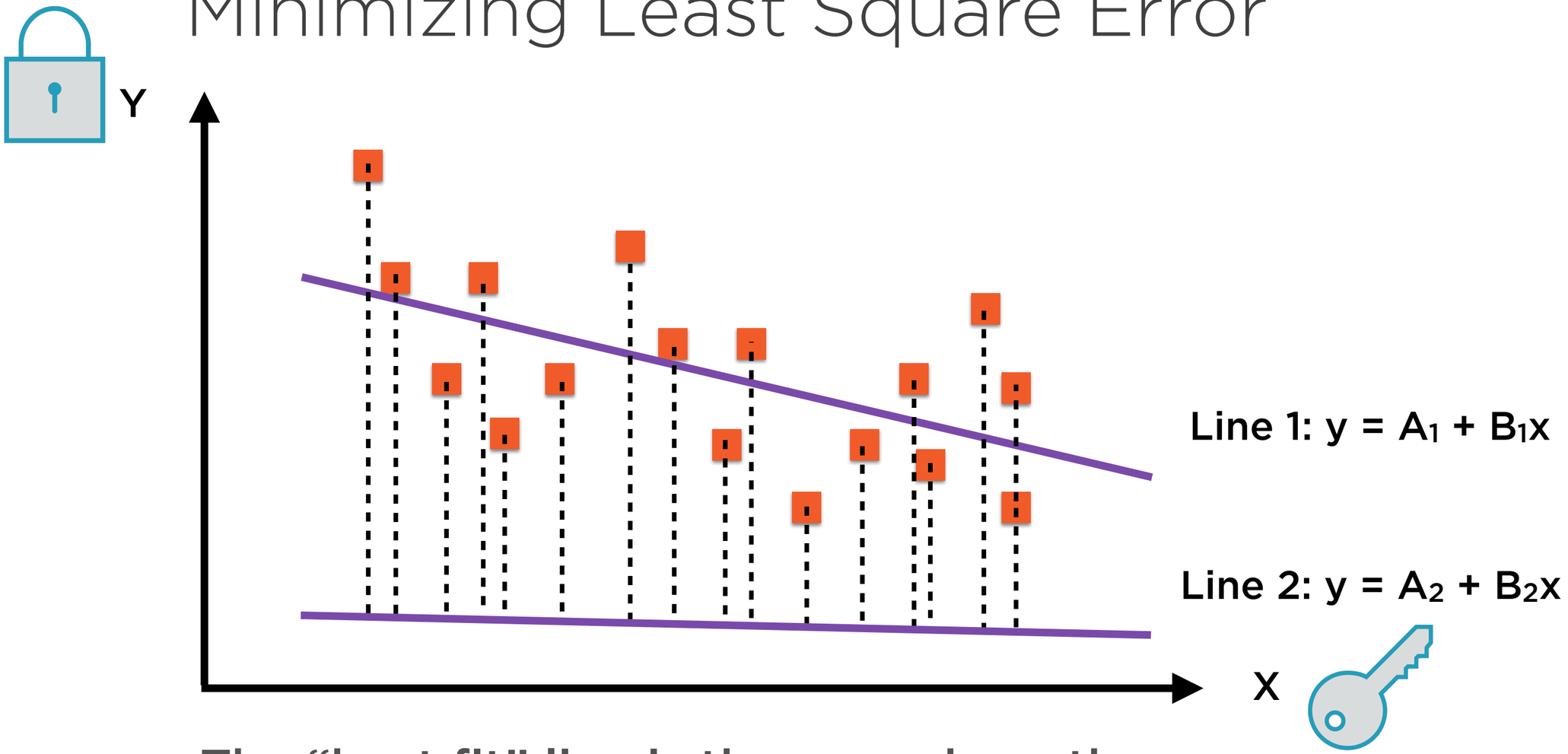
Drop vertical lines from each point to the lines 1 and 2

Minimizing Least Square Error



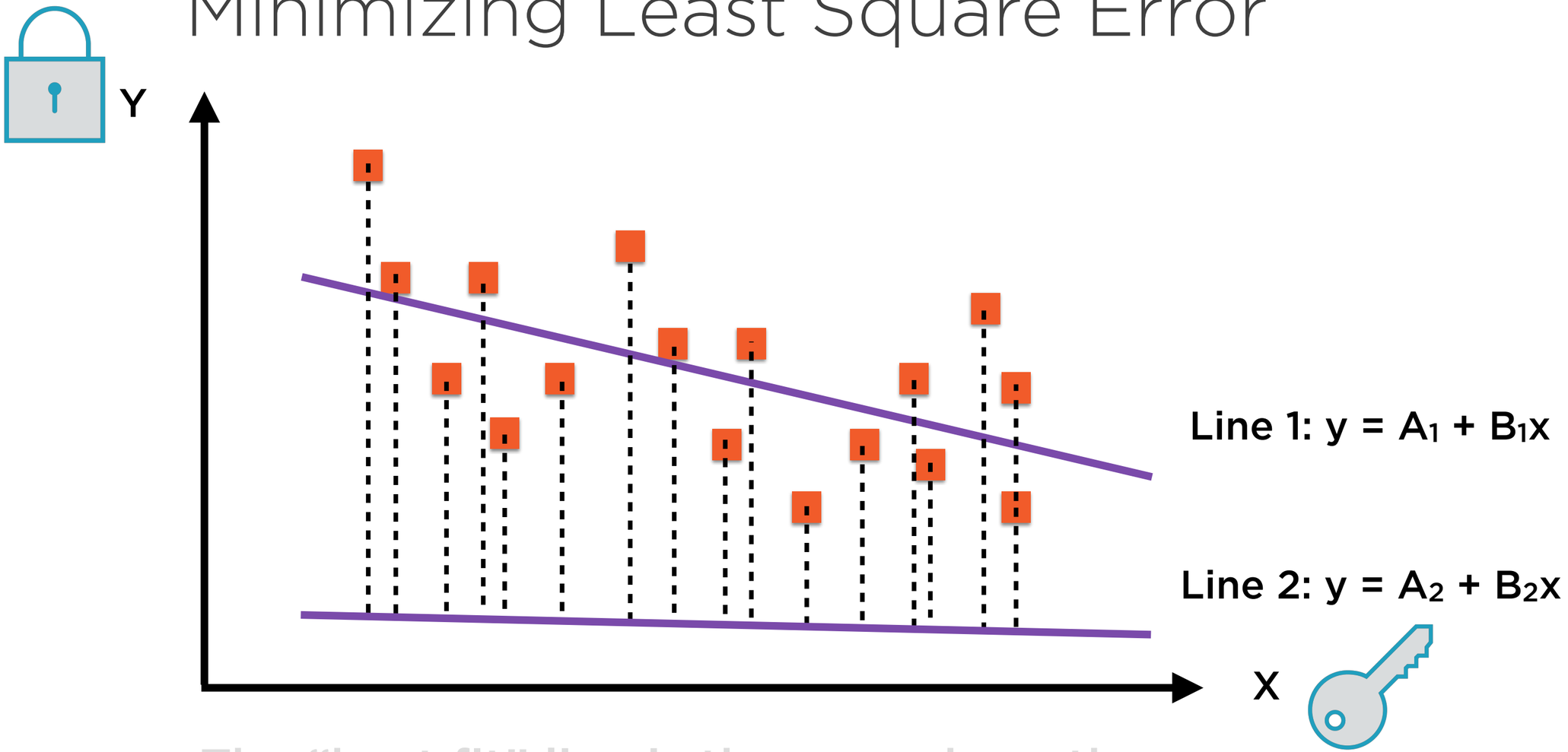
Drop vertical lines from each point to the lines 1 and 2

Minimizing Least Square Error



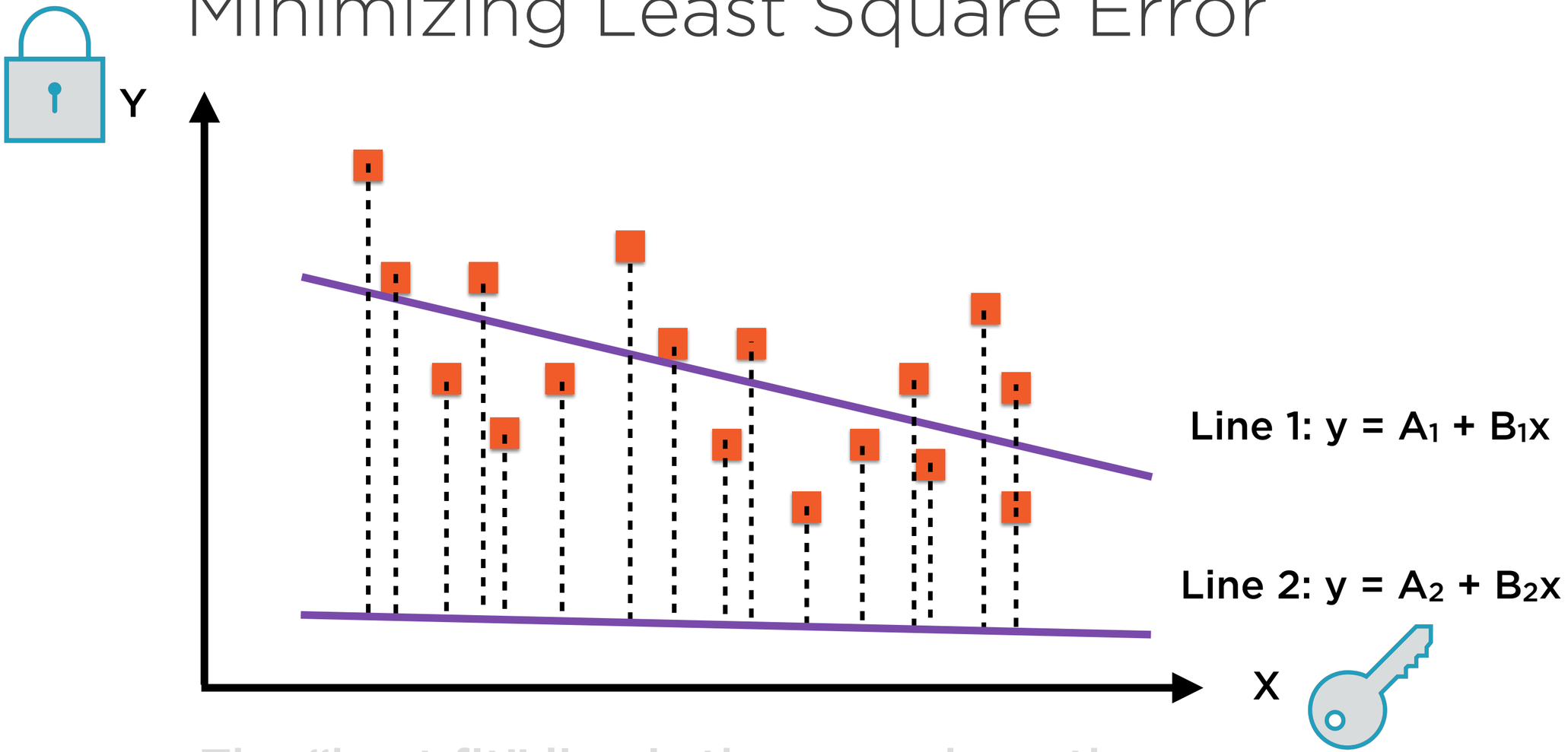
The “best fit” line is the one where the sum of the squares of the lengths of these dotted lines are minimum

Minimizing Least Square Error



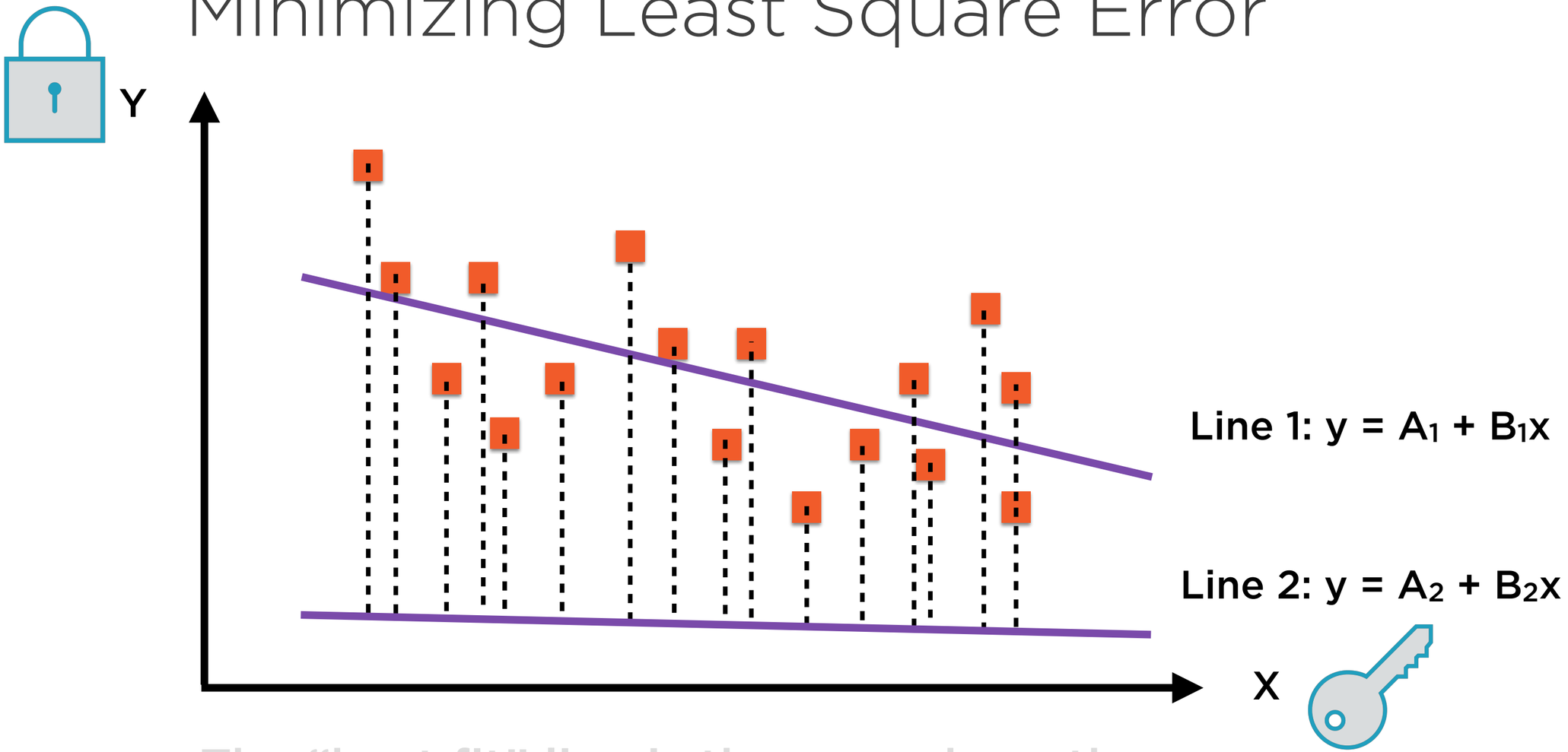
The "best fit" line is the one where the sum of the squares of the **lengths of these dotted lines** are minimum

Minimizing Least Square Error



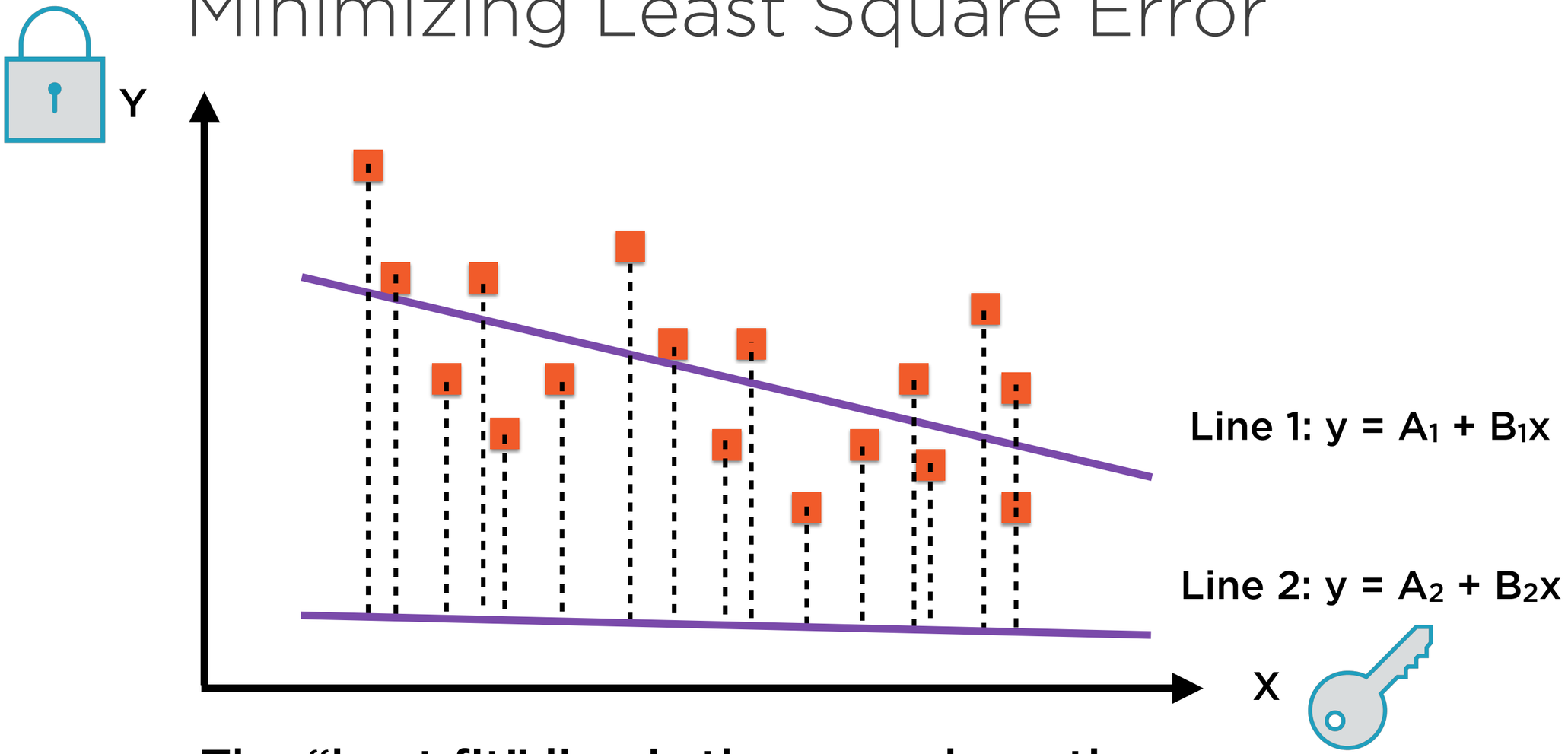
The "best fit" line is the one where the **sum of the squares** of the lengths of these dotted lines are minimum

Minimizing Least Square Error



The "best fit" line is the one where the sum of the squares of the lengths of **the errors are minimum**

Minimizing Least Square Error



The "best fit" line is the one where the sum of the squares of the lengths of the errors are minimum

Demo

Training a linear regression model and using it for prediction

Building Classification Models

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

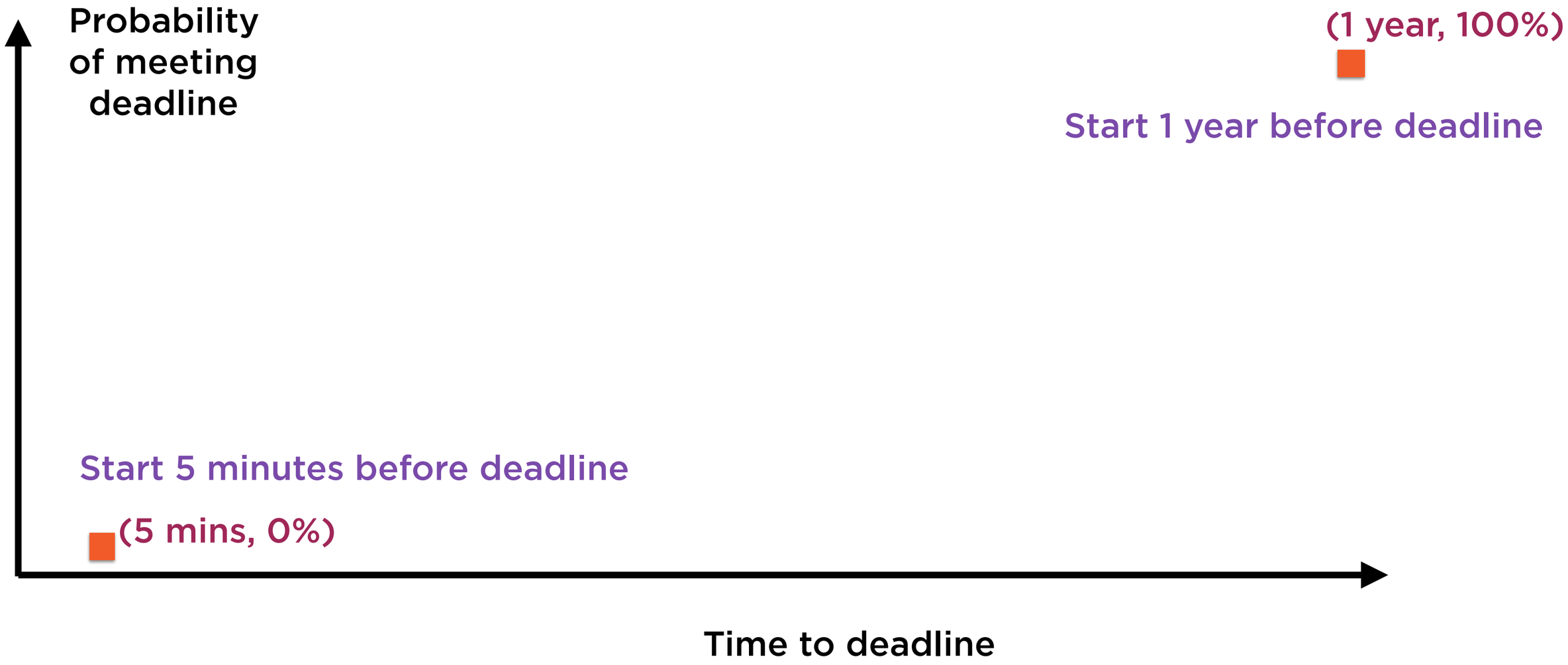
Probability of meeting the deadline

95%

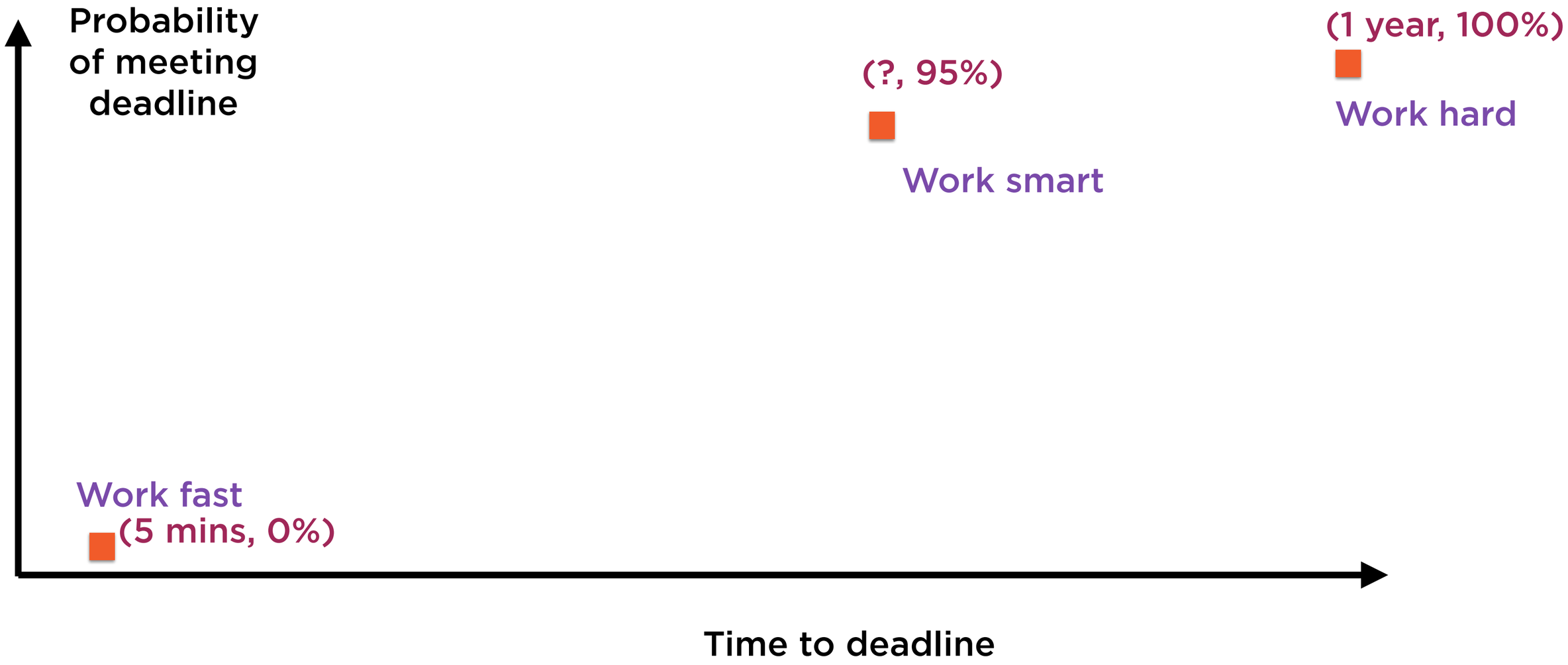
Probability of getting other important work done

95%

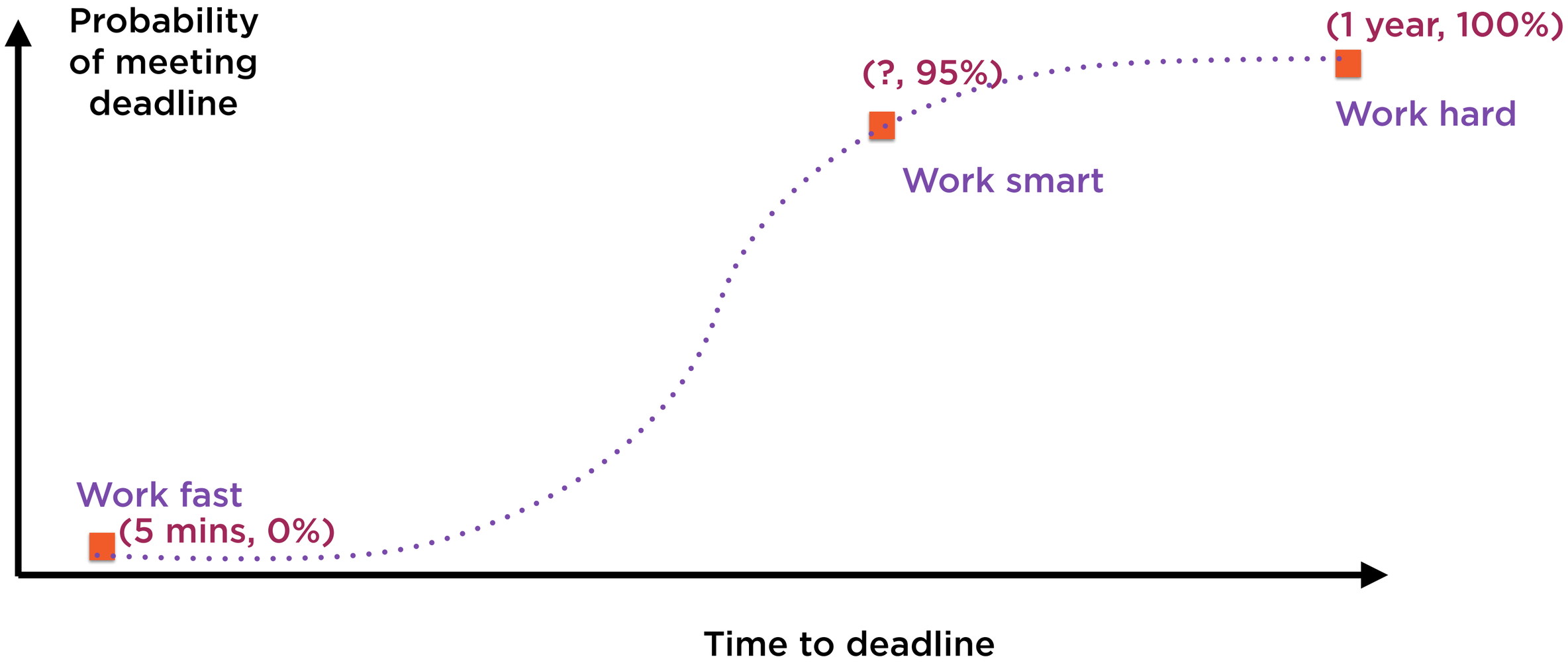
Working Hard, Fast, Smart



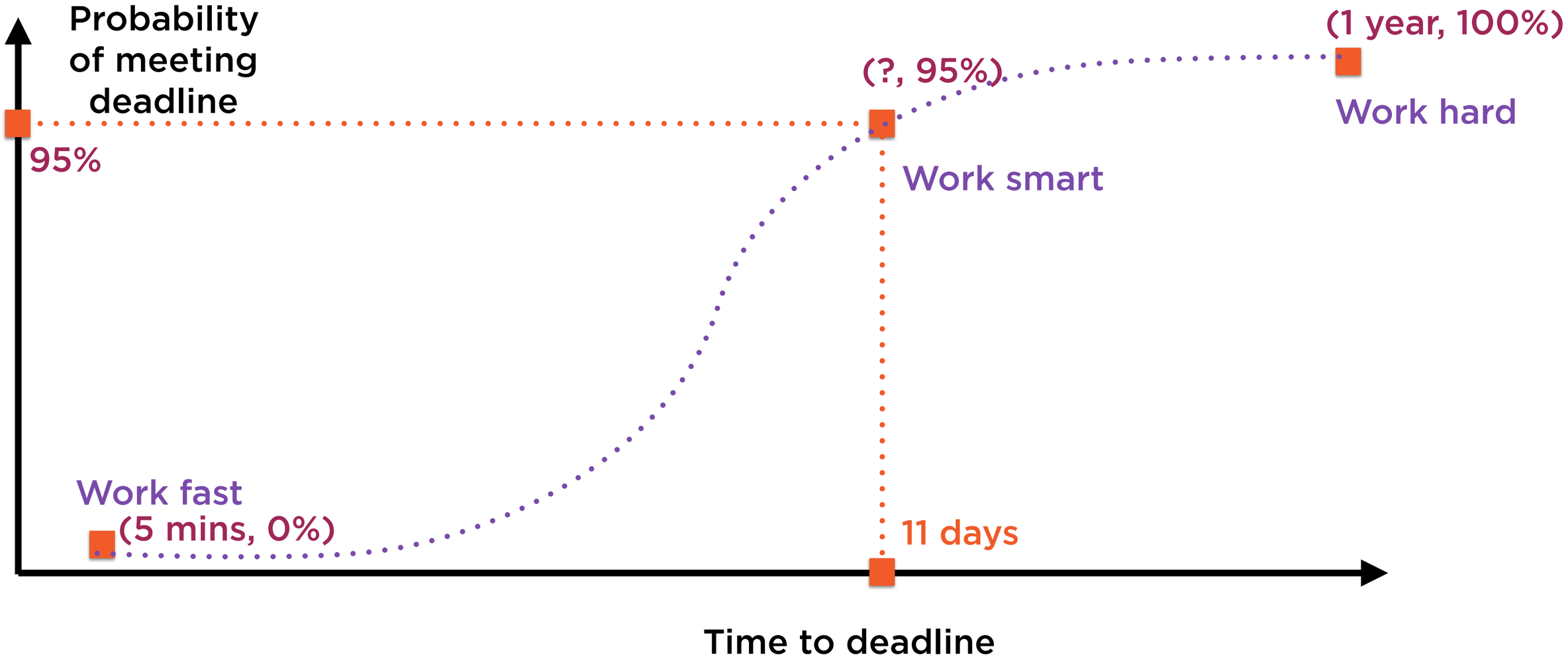
Working Hard, Fast, Smart



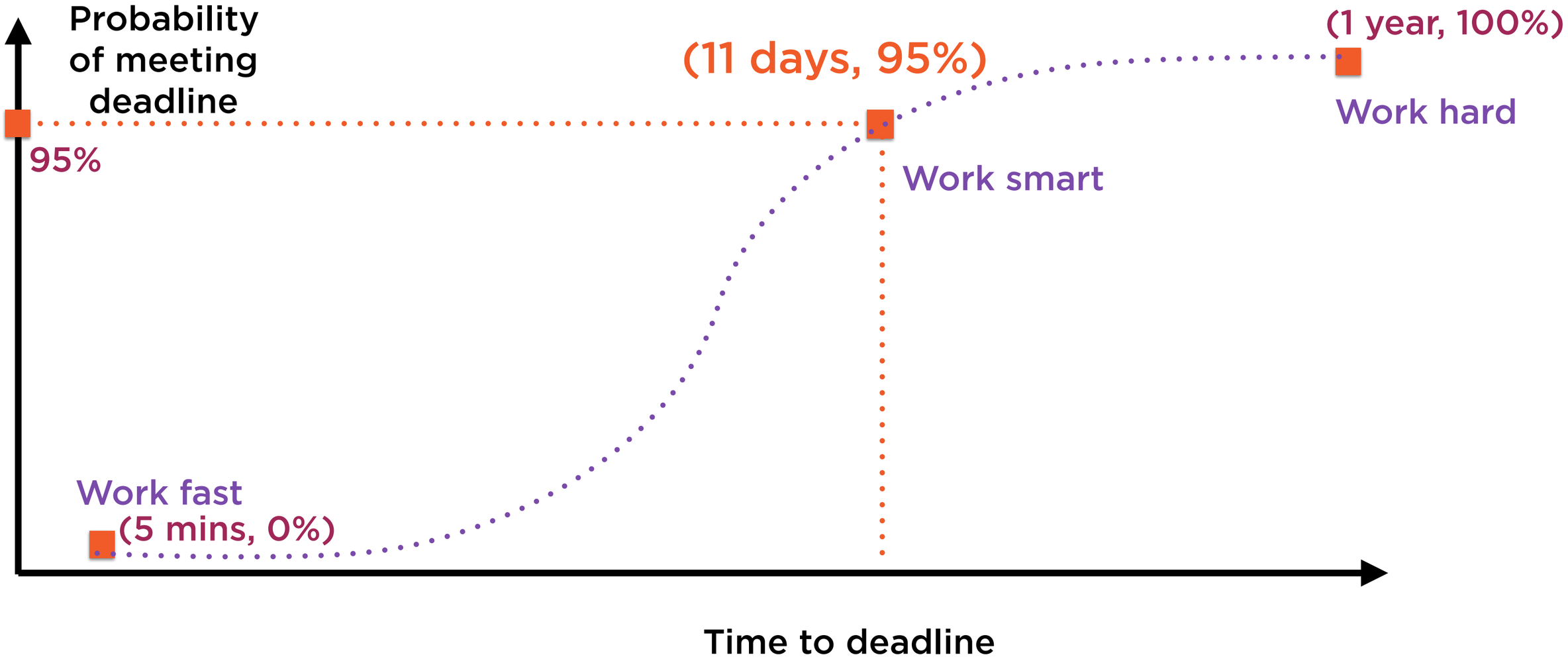
Working Hard, Fast, Smart



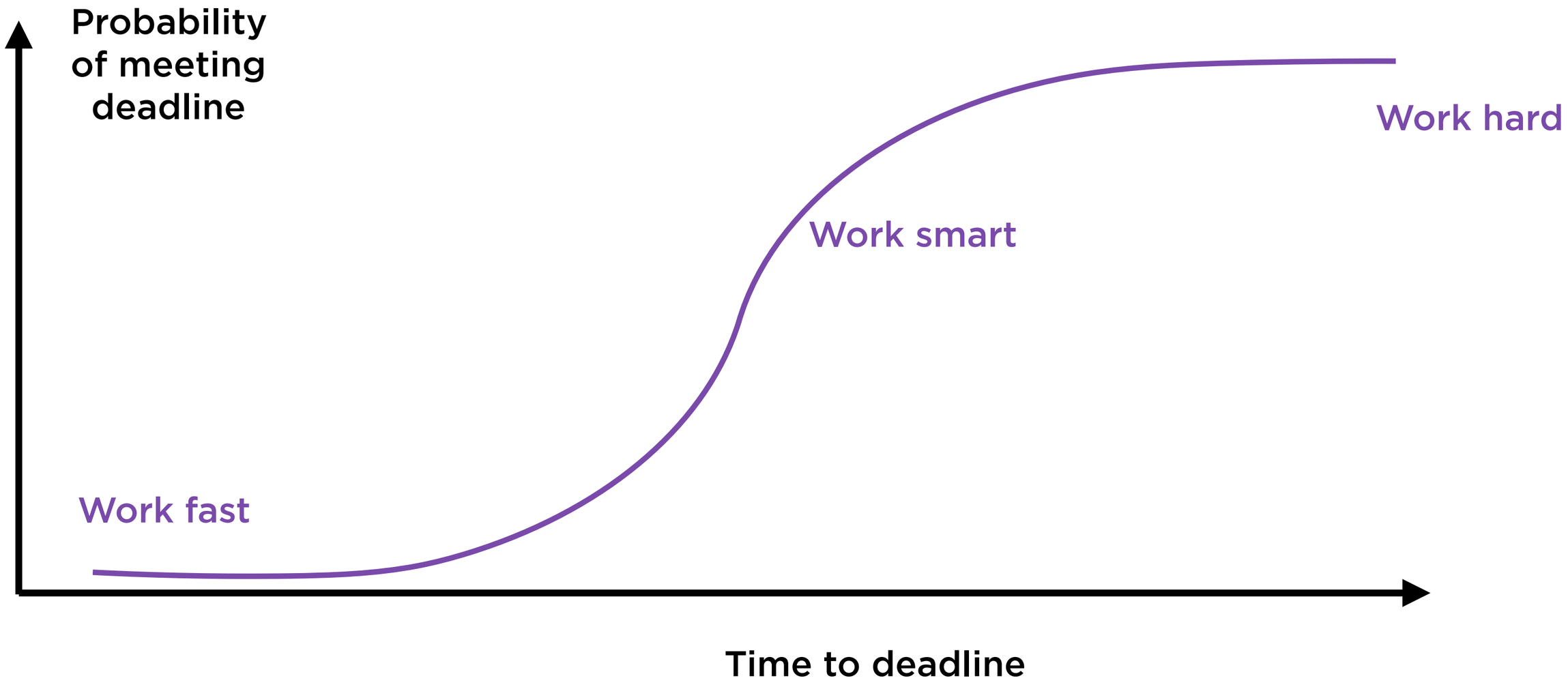
Working Hard, Fast, Smart



Working Hard, Fast, Smart

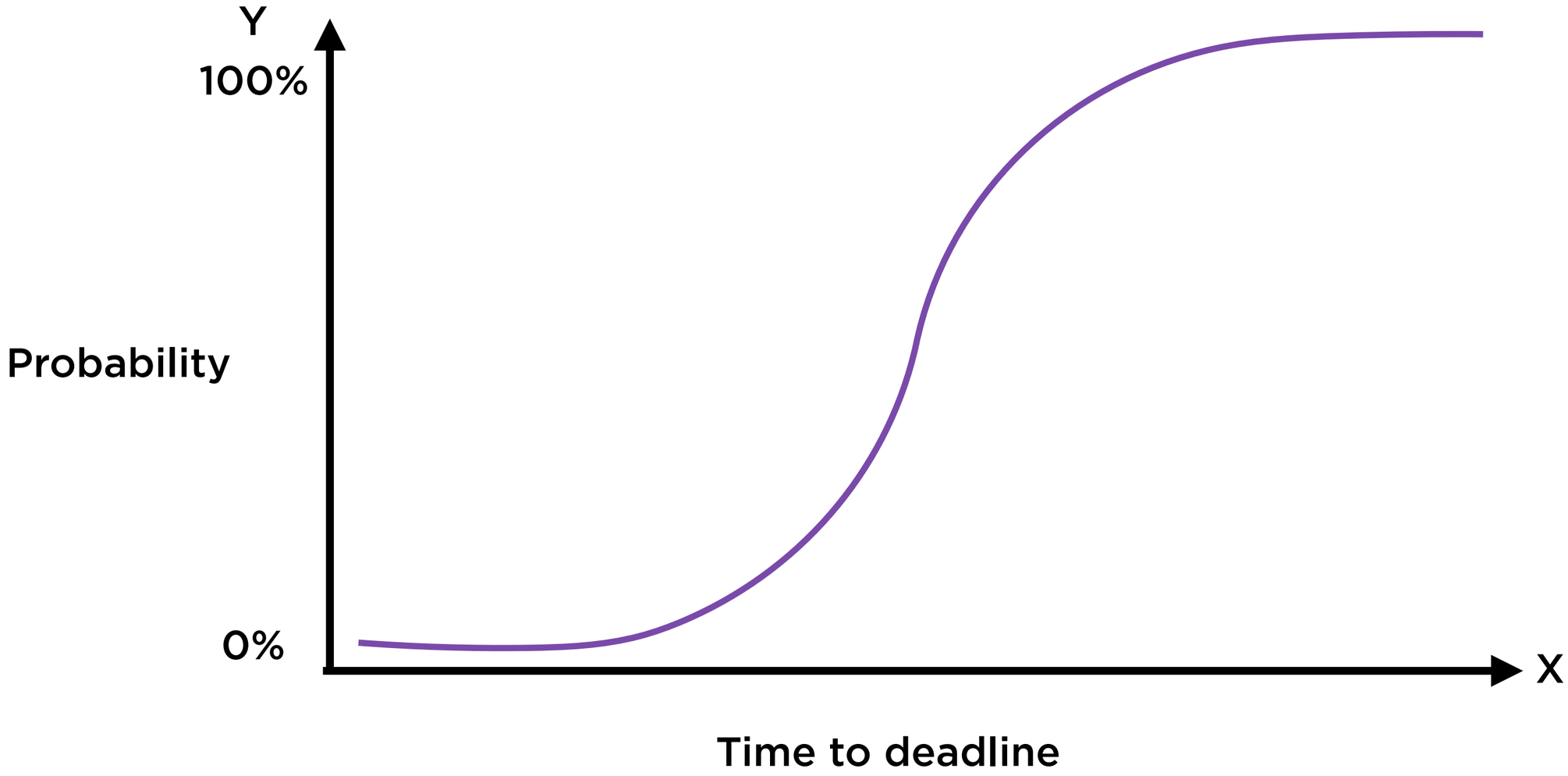


Working Hard, Fast, Smart

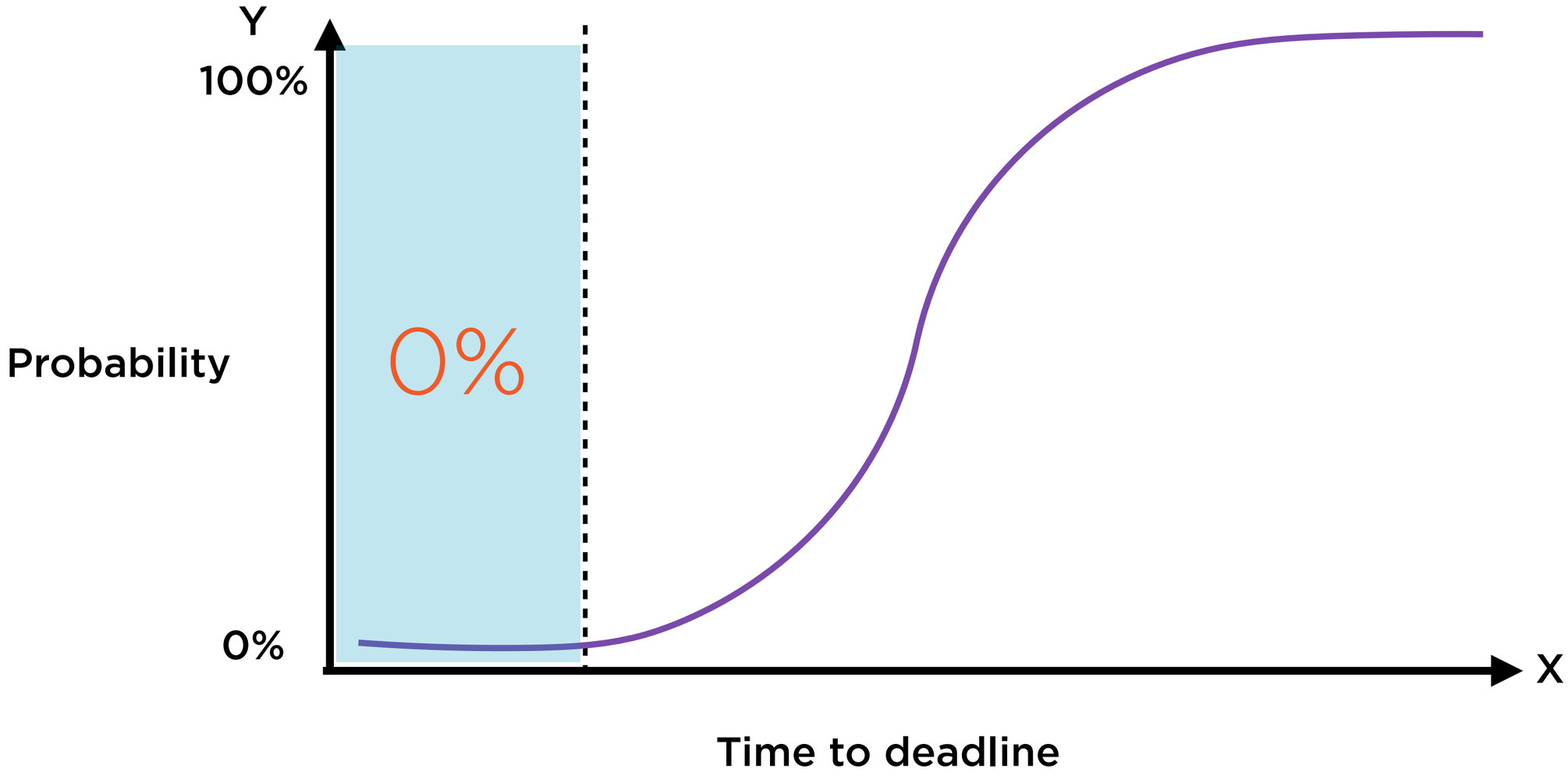


Logistic Regression helps find how probabilities are changed by actions

Working Smart with Logistic Regression

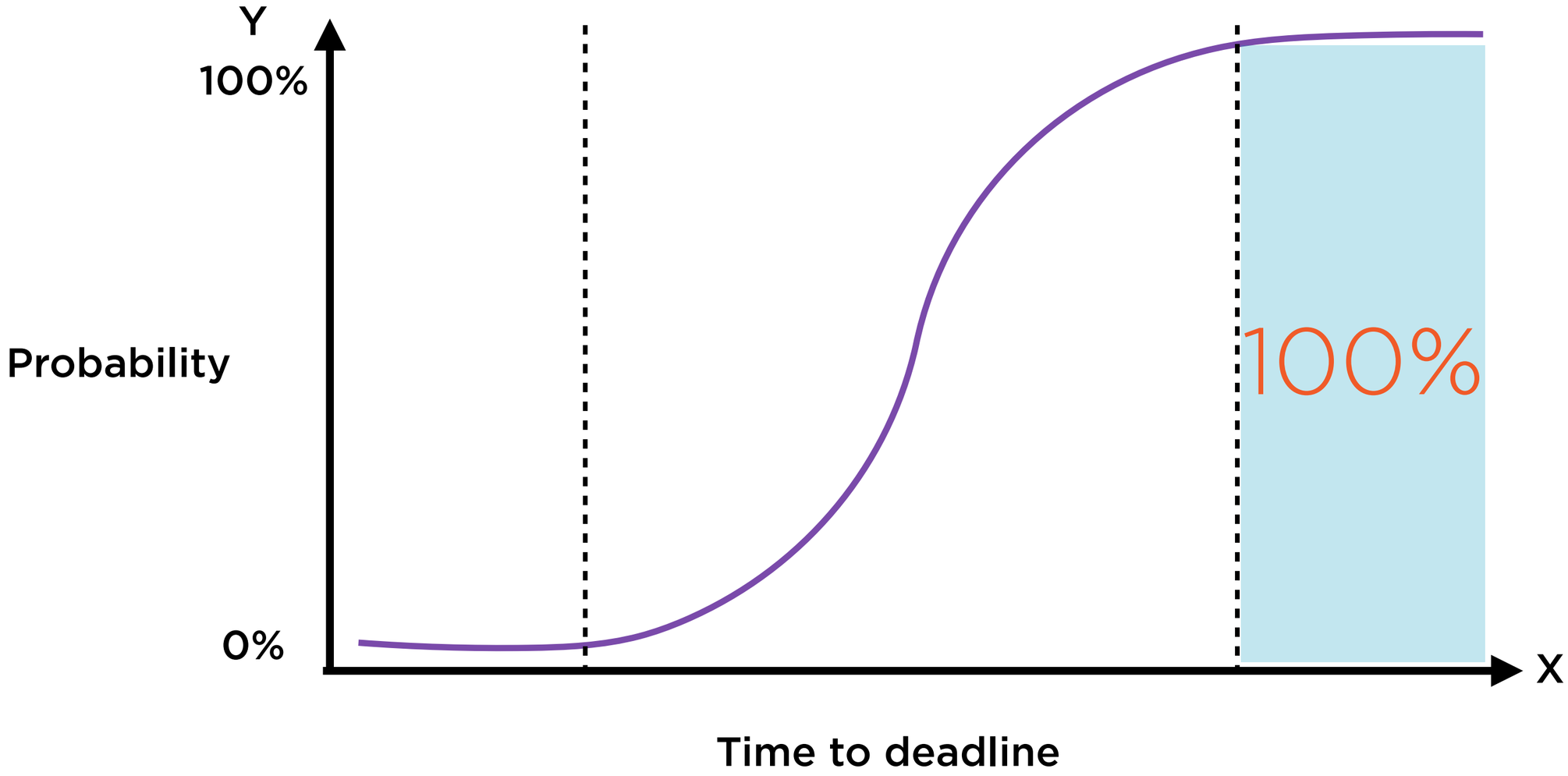


Working Smart with Logistic Regression



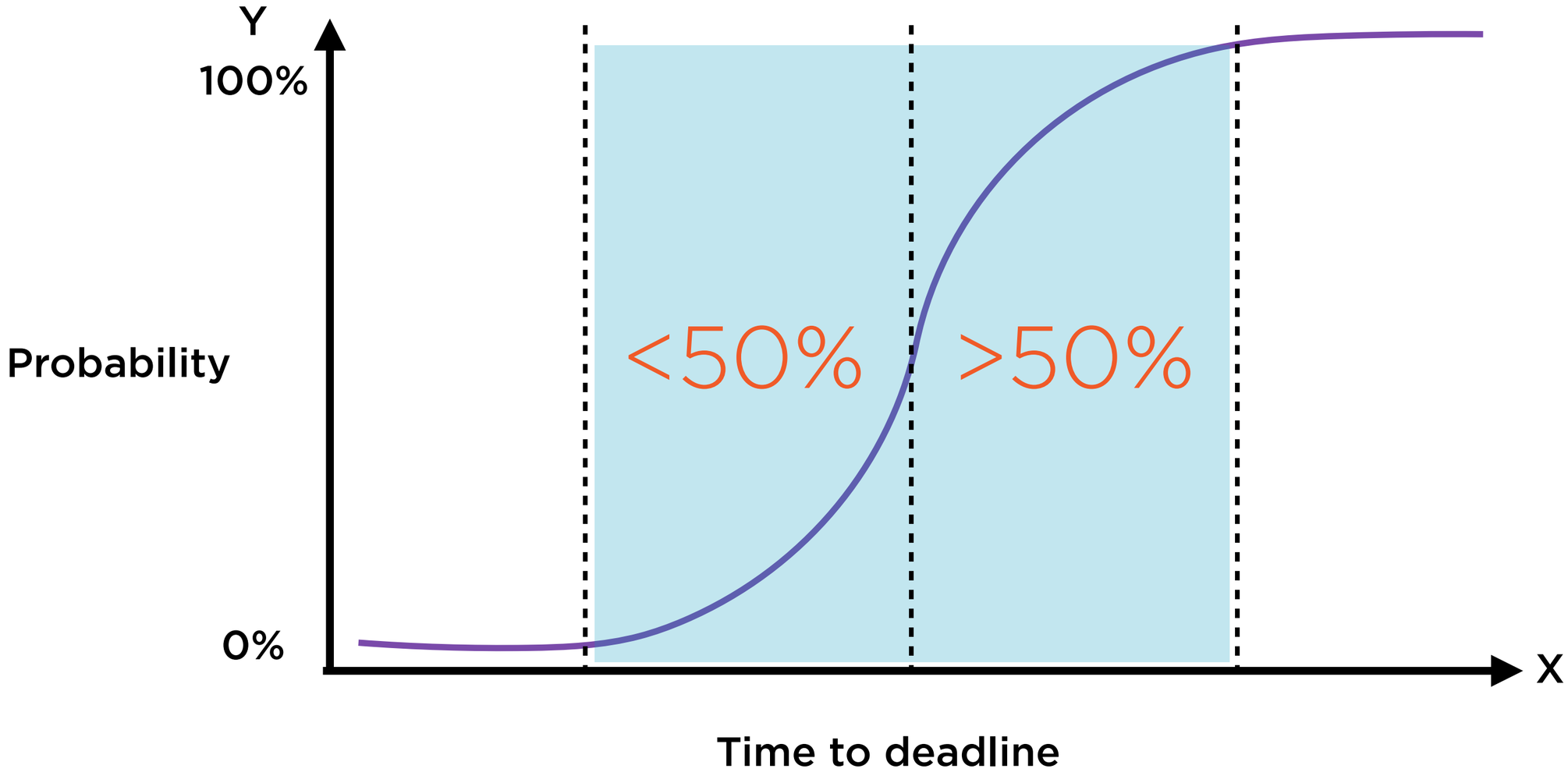
Start too late, and you'll definitely miss

Working Smart with Logistic Regression

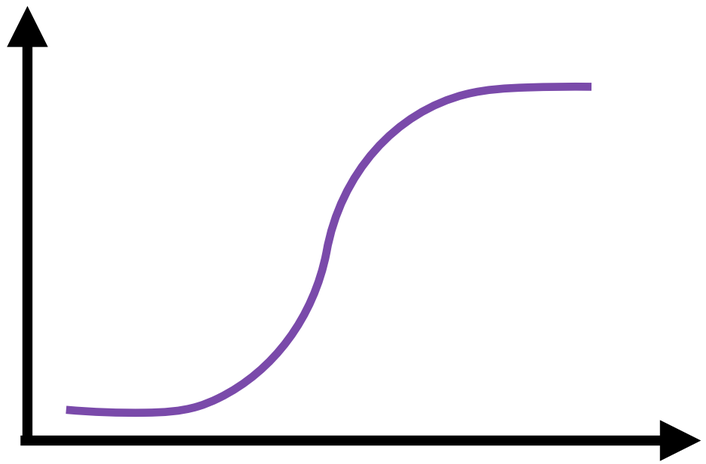


Start too early, and you'll definitely make it

Working Smart with Logistic Regression



Working smart is knowing when to start



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

x: Start time before deadline

p(y) : Probability of y = 1

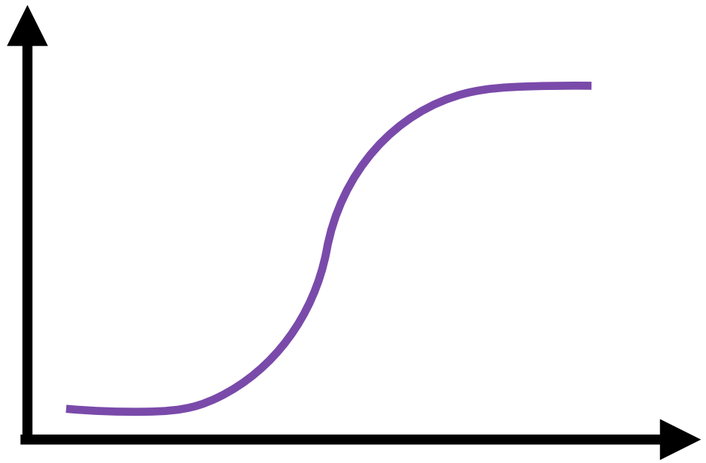
$$p(y_i) = \frac{1}{1 + e^{-(A+Bx_i)}}$$

Logistic regression involves finding the "best fit" such curve

- A is the intercept

- B is the regression coefficient

(e is the constant 2.71828)

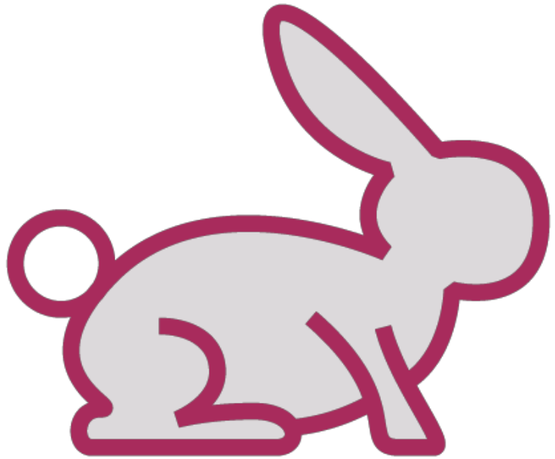


S-curves are widely studied, well understood

Logistic regression uses S-curve to estimate probabilities

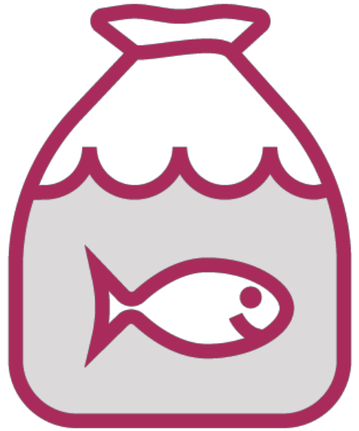
$$p(y) = \frac{1}{1 + e^{-(A+Bx)}}$$

Whales: Fish or Mammals



Mammal

Member of the infraorder *Cetacea*



Fish

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

ML-based Binary Classifier

Corpus

Classification algorithm

ML-based classifier

ML-based Predictor

Corpus

Logistic regression

ML-based predictor

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

ML-based Predictor

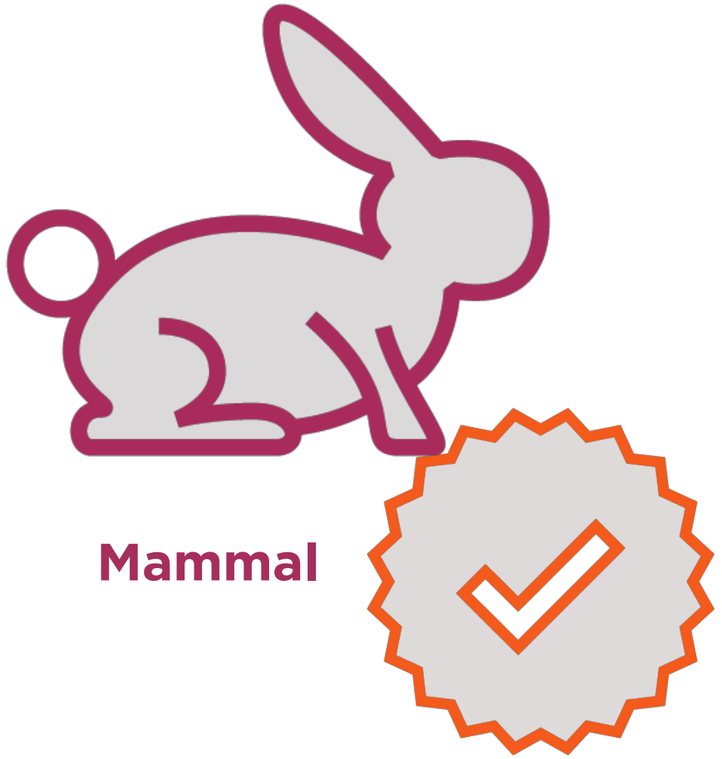
Corpus

Logistic regression

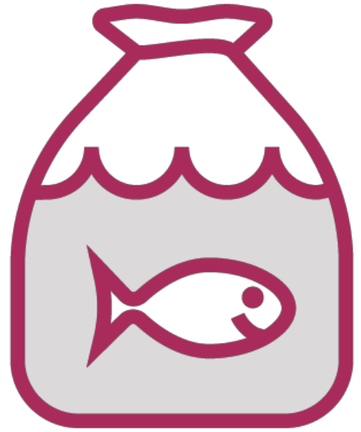
ML-based predictor

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

Applying Logistic Regression



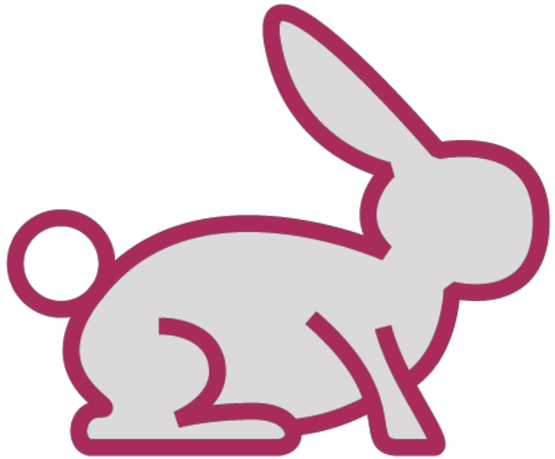
Mammal



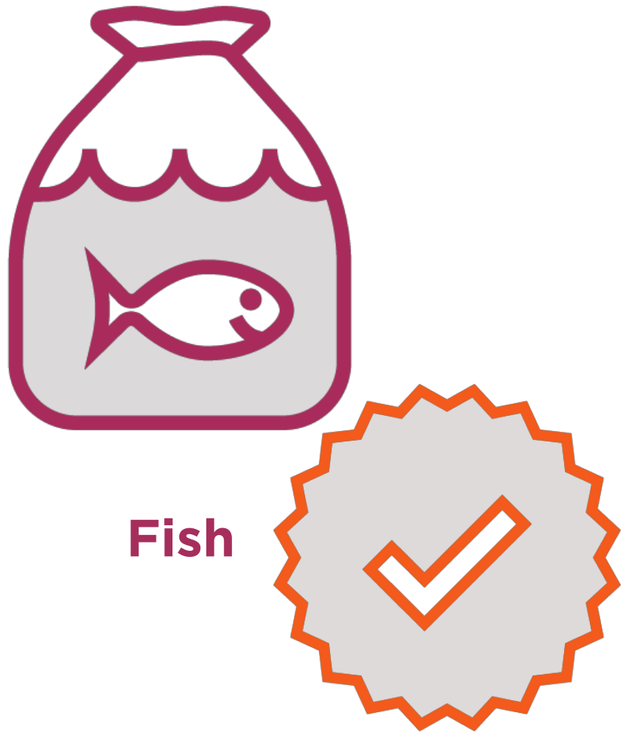
Fish

Probability of whales being fish < $P_{threshold}$

Applying Logistic Regression



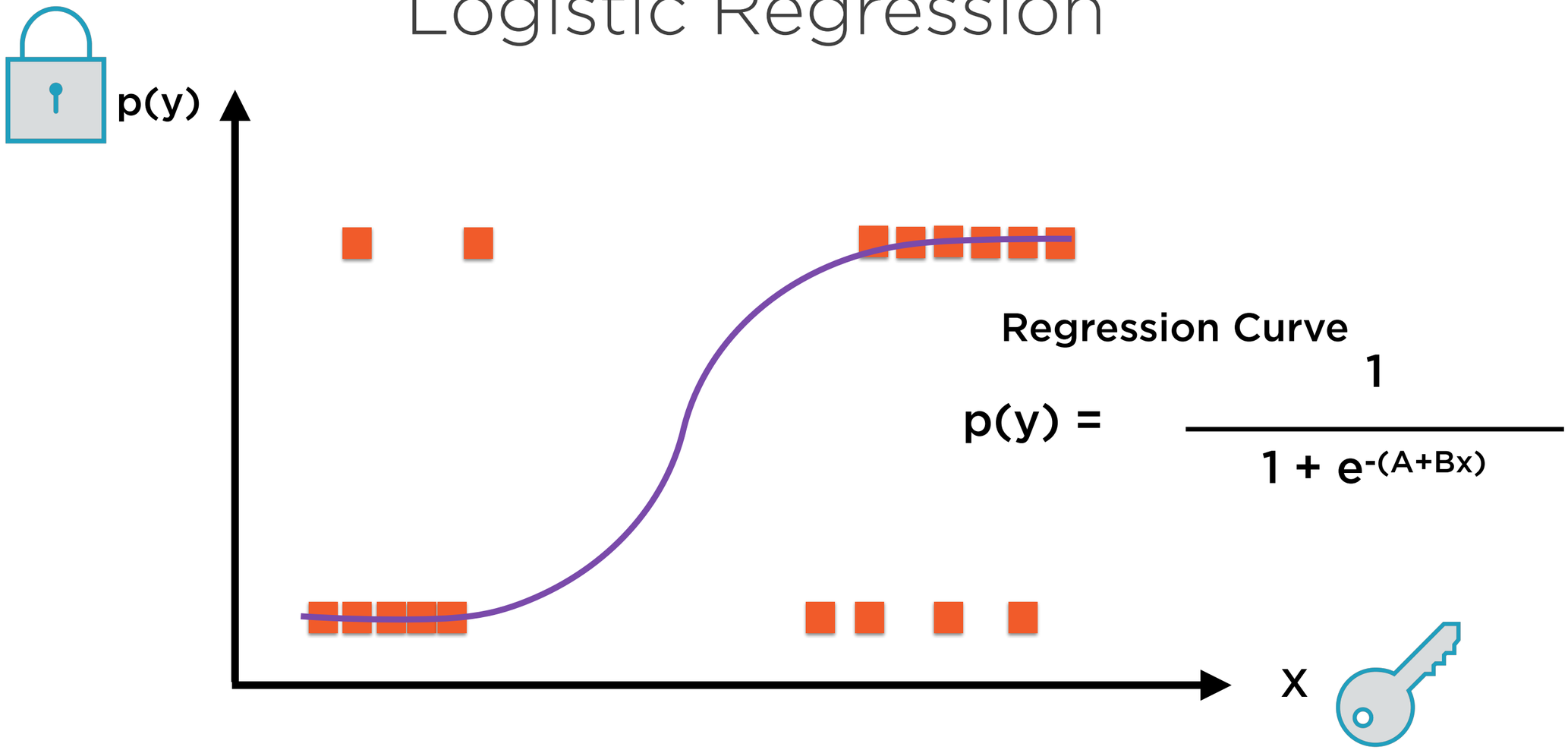
Mammal



Fish

Probability of whales being fish > $P_{threshold}$

Logistic Regression



Finding the best fit S-curve
through these points

Demo

Training a logistic regression model and using it for classification

Summary

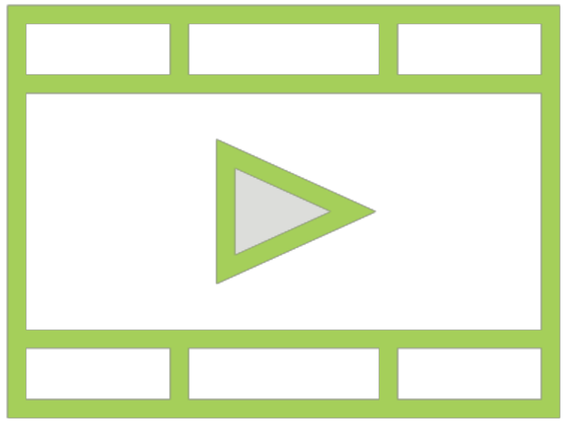
Classic problems in machine learning

Regression for predicting continuous data

Classification for predicting categorical data

Implementing simple linear and logistic regression in scikit-learn

Related Courses



Building Clustering Models with scikit-learn

Employing Ensemble Methods with scikit-learn

Building Neural Networks with scikit-learn