Implementing Statistical Models

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Fitting and interpreting models

Understanding linear regression

Interpreting results of linear regression using diagnostic plots

Understanding logistic regression

Interpreting logistic regression using accuracy, sensitivity and specificity

Computing and interpreting odds ratio

Visualizing odds ratio using Forest plot

Types of Data

Categorical

Male/Female, Month of year

Numeric (Continuous)

Weight in lbs, Temperature in °F

Use regression to predict numeric (continuous) y-variables

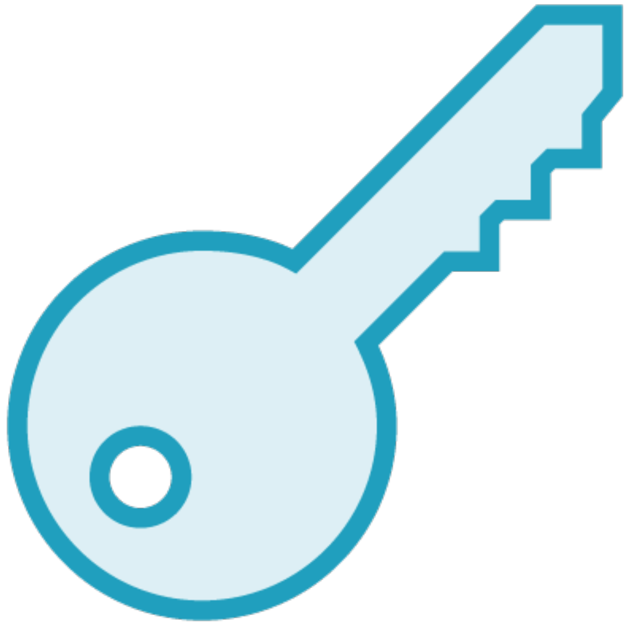
Use classification to predict categorical (discrete) y-variables

Numeric (Continuous) vs. Categorical Data

Numeric (Continuous)	Categorical
E.g. height or weight of individuals	E.g. day of week, month of year, gender, letter grade
Can take any value	Finite set of permissible values
Predicted using regression models	Predicted using classification models
Always can be sorted on magnitude	Categories may or may not be sortable

Linear Regression

X Causes Y



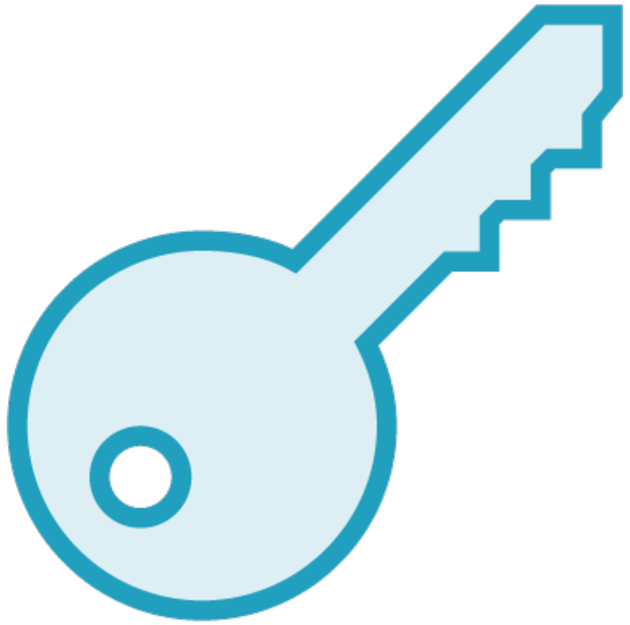
Cause

Independent variable

Effect

Dependent variable

X Causes Y



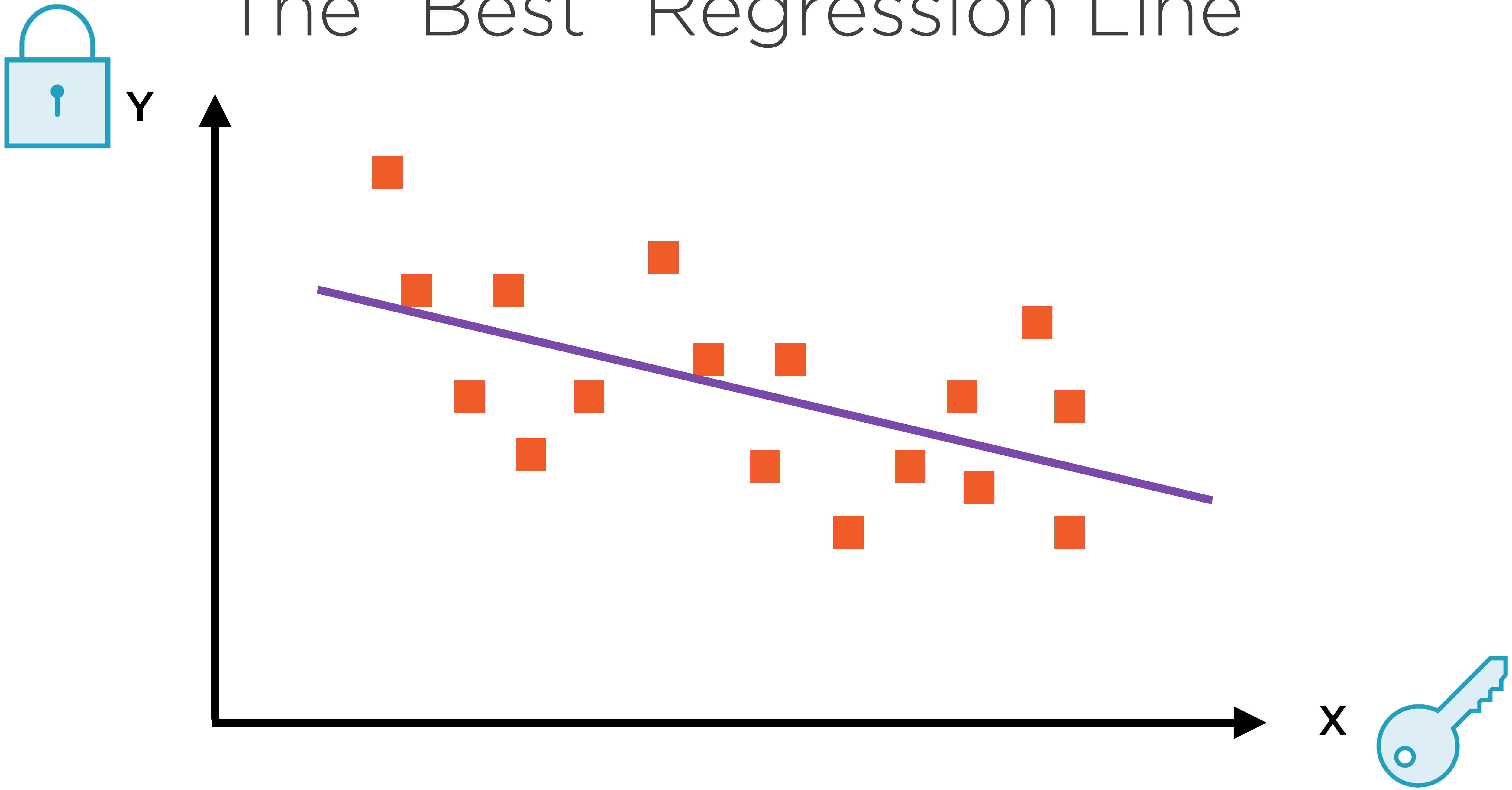
Cause

Explanatory variable

Effect

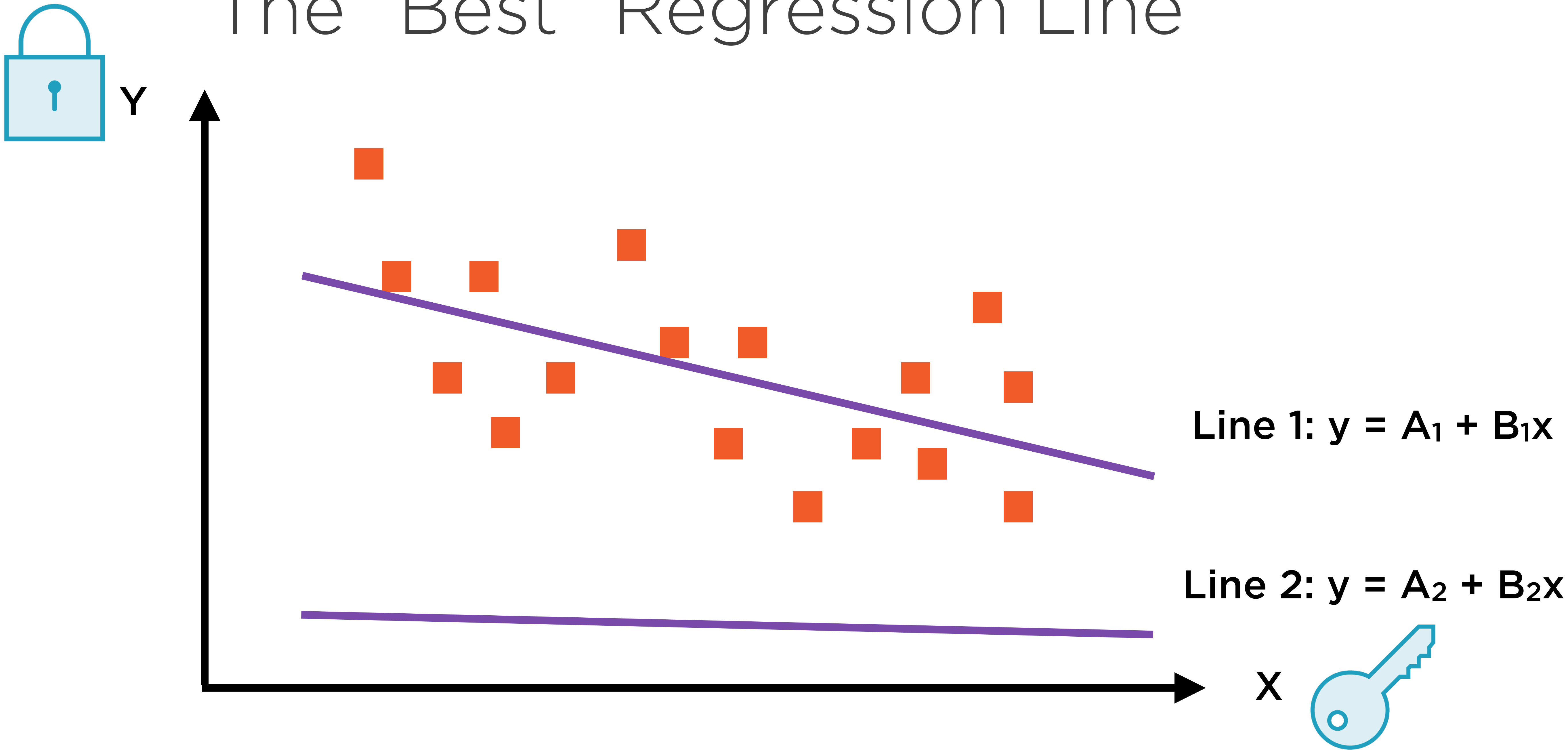
Dependent variable

The “Best” Regression Line



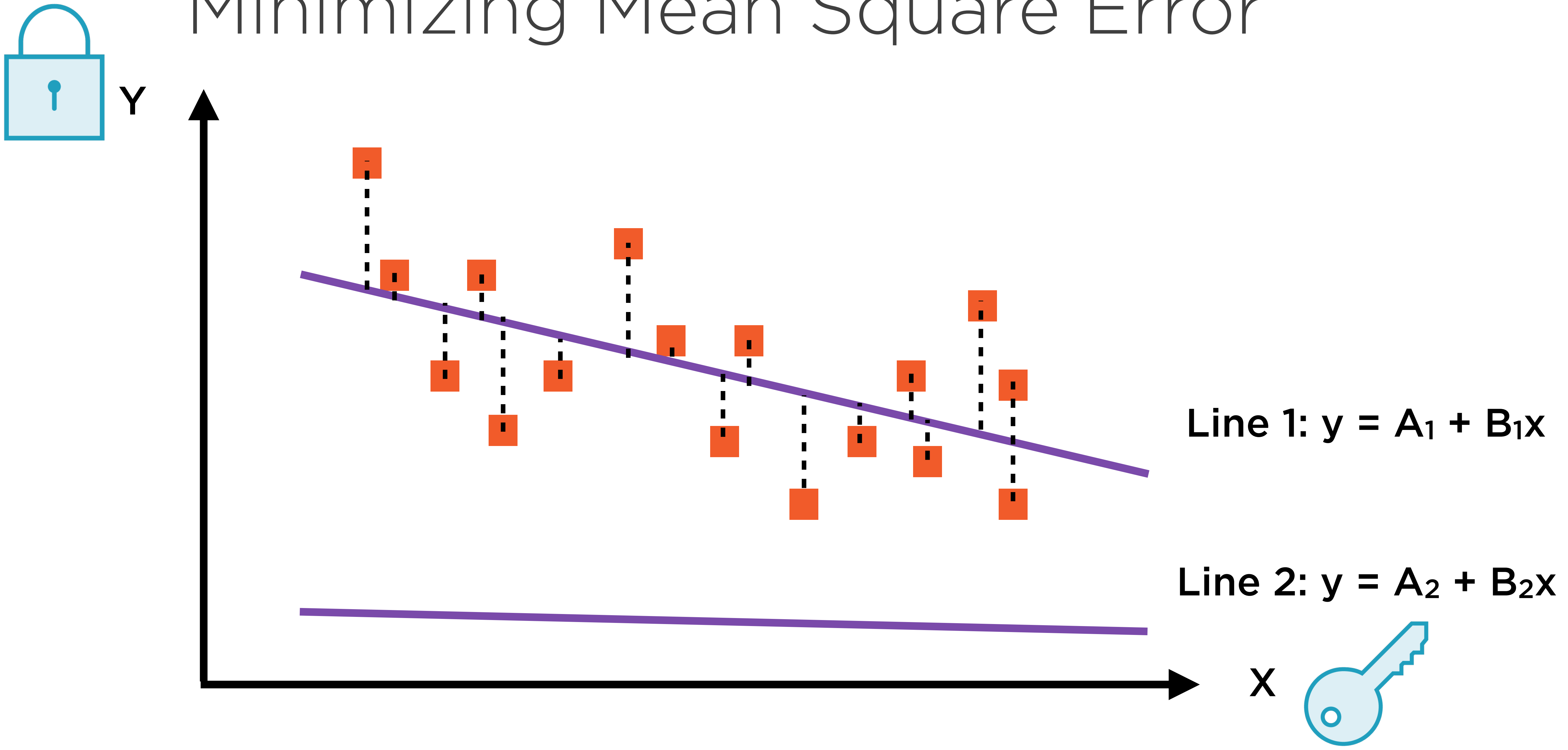
Linear Regression involves finding the “best fit” line

The "Best" Regression Line



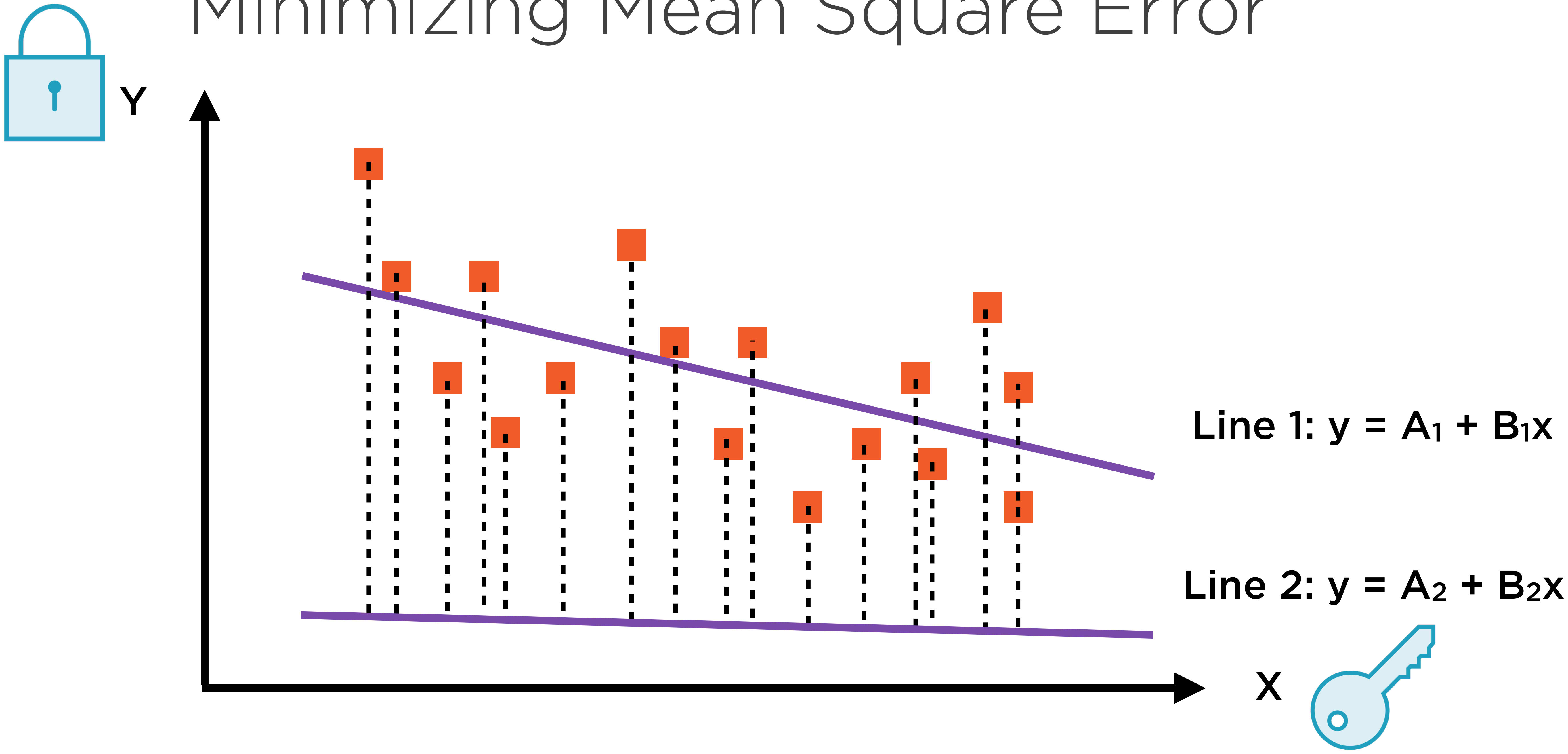
Let's compare two lines, Line 1 and Line 2

Minimizing Mean Square Error



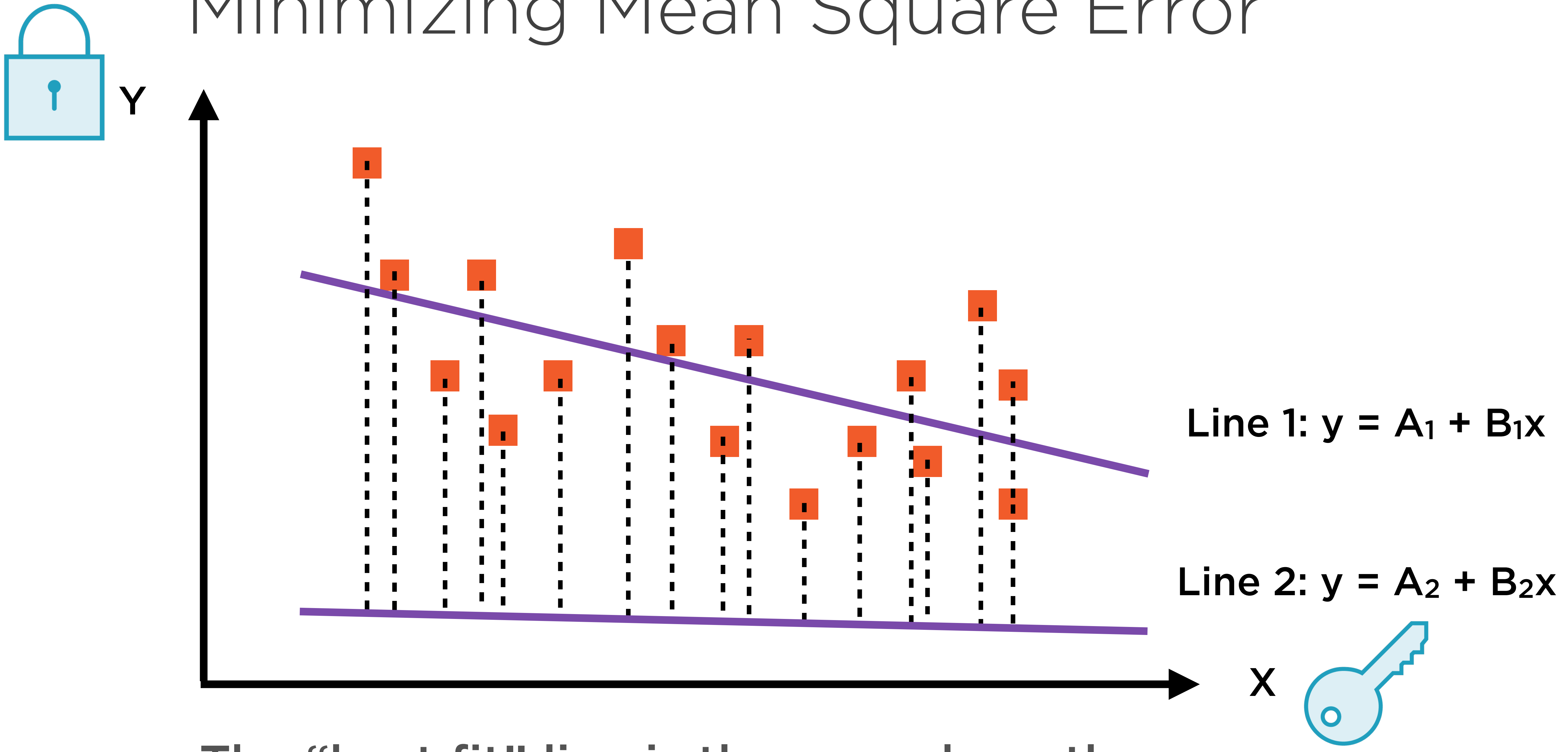
Drop vertical lines from each point to the lines 1 and 2

Minimizing Mean Square Error



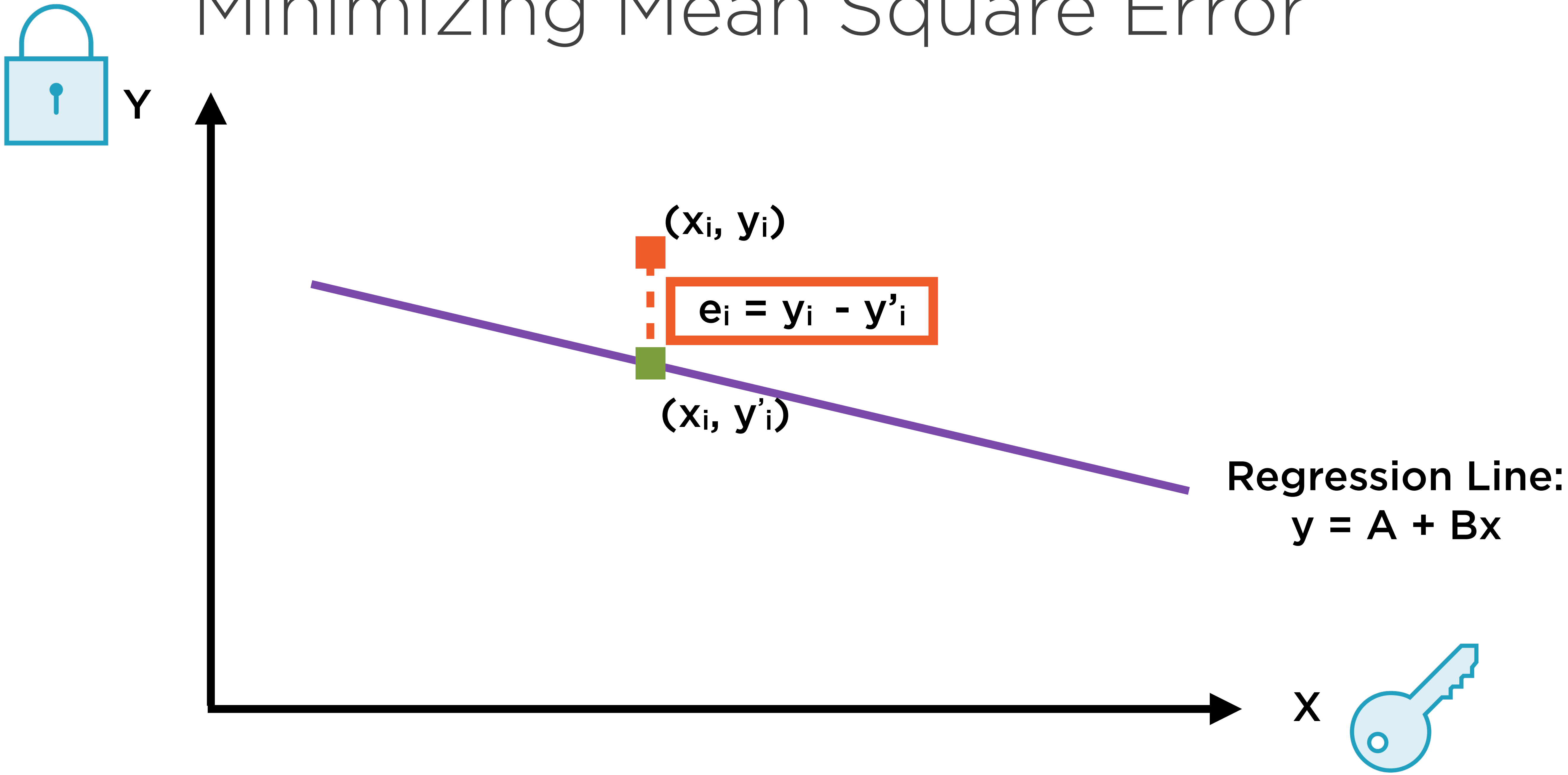
Drop vertical lines from each point to the lines 1 and 2

Minimizing Mean Square Error



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

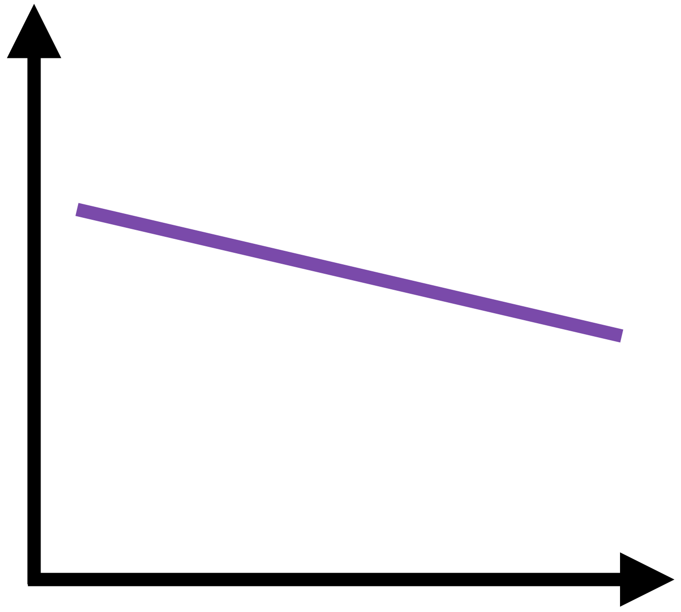
Minimizing Mean Square Error



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

The regression line is that line which minimizes the variance of the residuals (MSE)

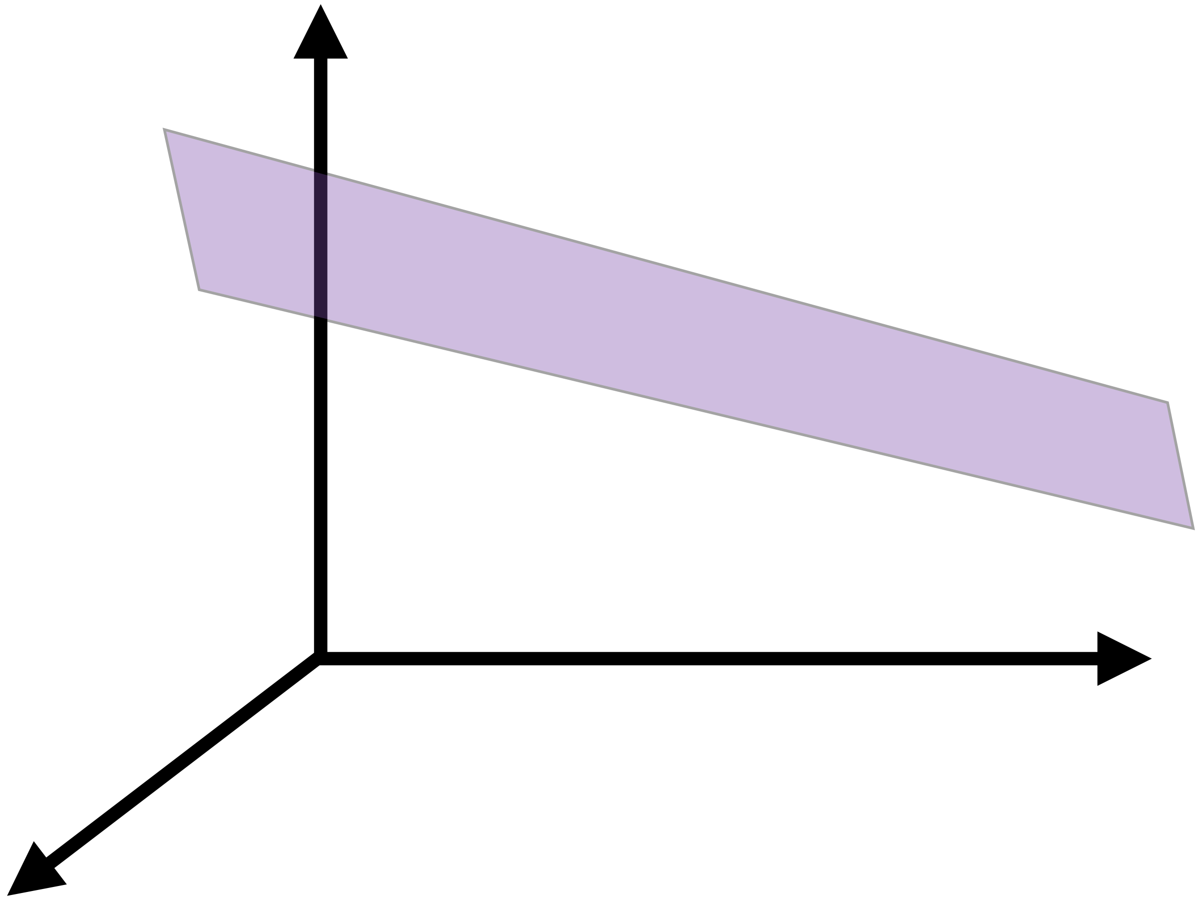
Simple and Multiple Regression



Simple Regression

One independent variable

$y = A + Bx$

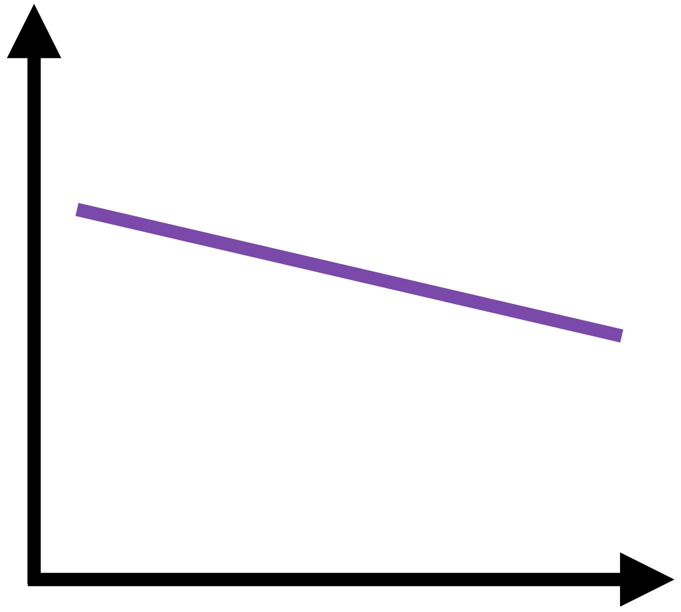


Multiple Regression

Multiple independent variables

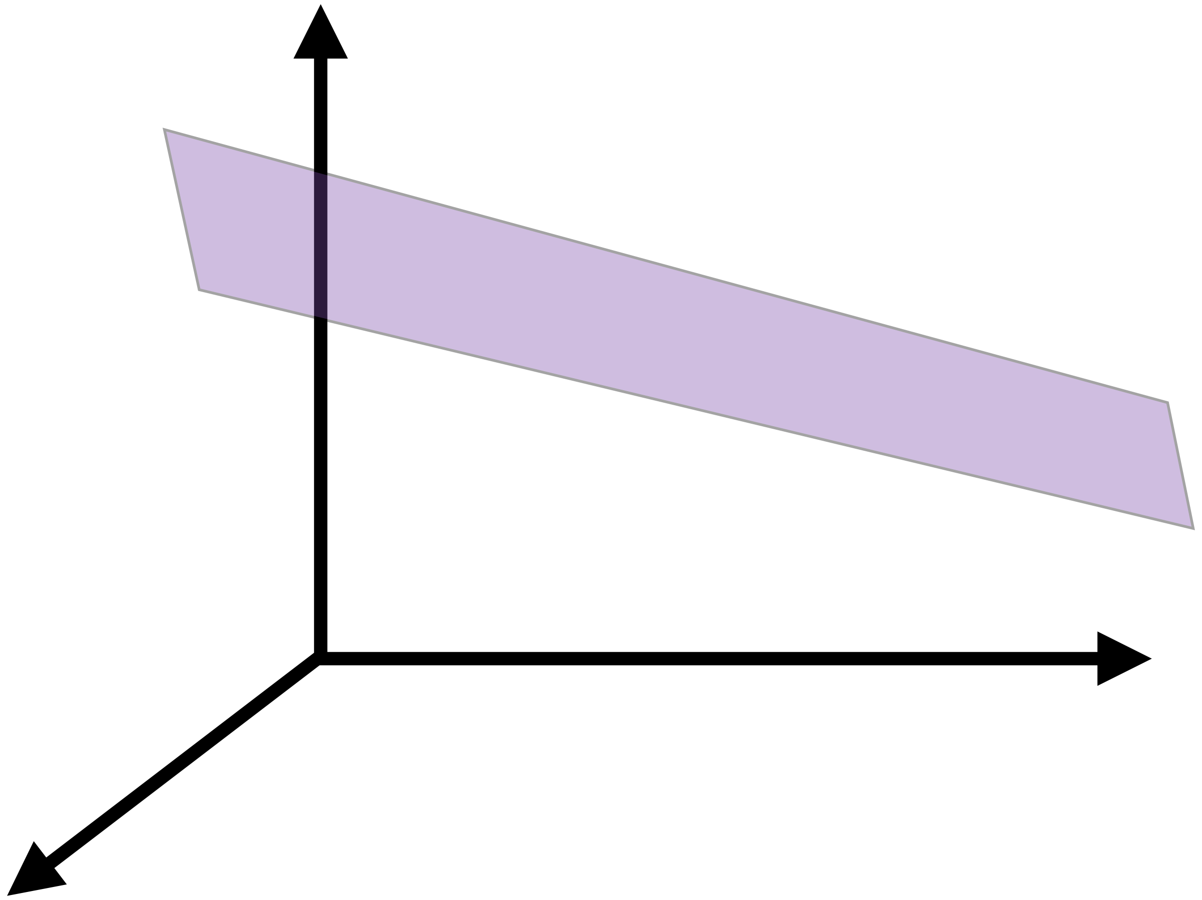
$y = A + B_1x_1 + B_2x_2 + B_3x_3$

MSE Minimization Extends To Multiple Regression



Simple Regression

One independent variable



Multiple Regression

Multiple independent variables

R^2 = ESS / TSS

R^2

R^2 = Explained Sum of Squares / Total Sum of Squares

R^2

ESS - Variance of fitted values

TSS - Variance of actual values

R^2 = Explained Sum of Squares / Total Sum of Squares

R^2

The percentage of total variance explained by the regression. Usually, the higher the R^2, the better the quality of the regression (upper bound is 100%)

$R^2 = ESS / TSS$

R^2

How much of the original variance is captured in the fitted values?

Generally, higher this number the better the regression

Adjusted-R^2 = R^2 x (Penalty for adding irrelevant variables)

Adjusted-R^2

Increases if irrelevant* variables are deleted

(*irrelevant variables = any group whose F-ratio < 1)

The regression line found by minimizing variance of residuals (MSE) is the line with the **best R^2**

Demo

Performing linear regression and interpreting the results

Logistic Regression: Intuition

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 hard

Start very early and do little else

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

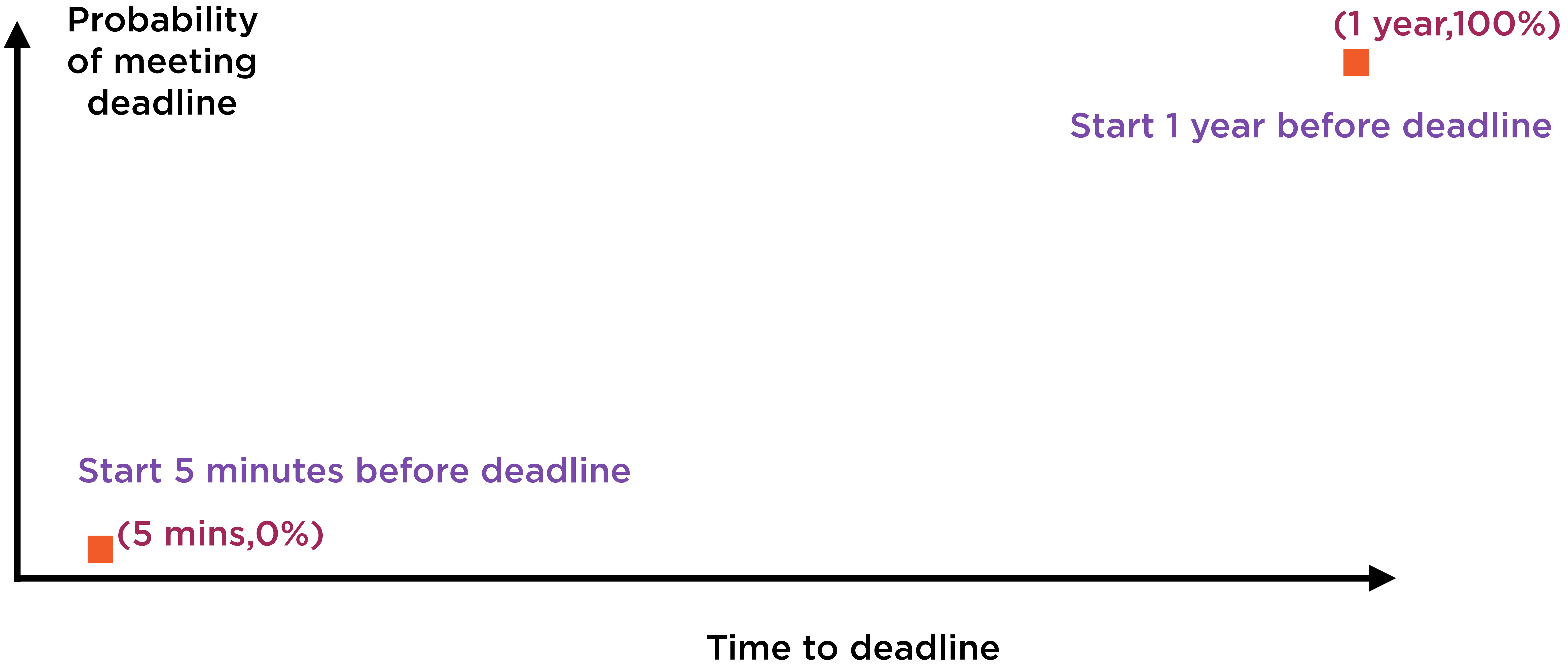
As usual, the middle path is best

Working Smart

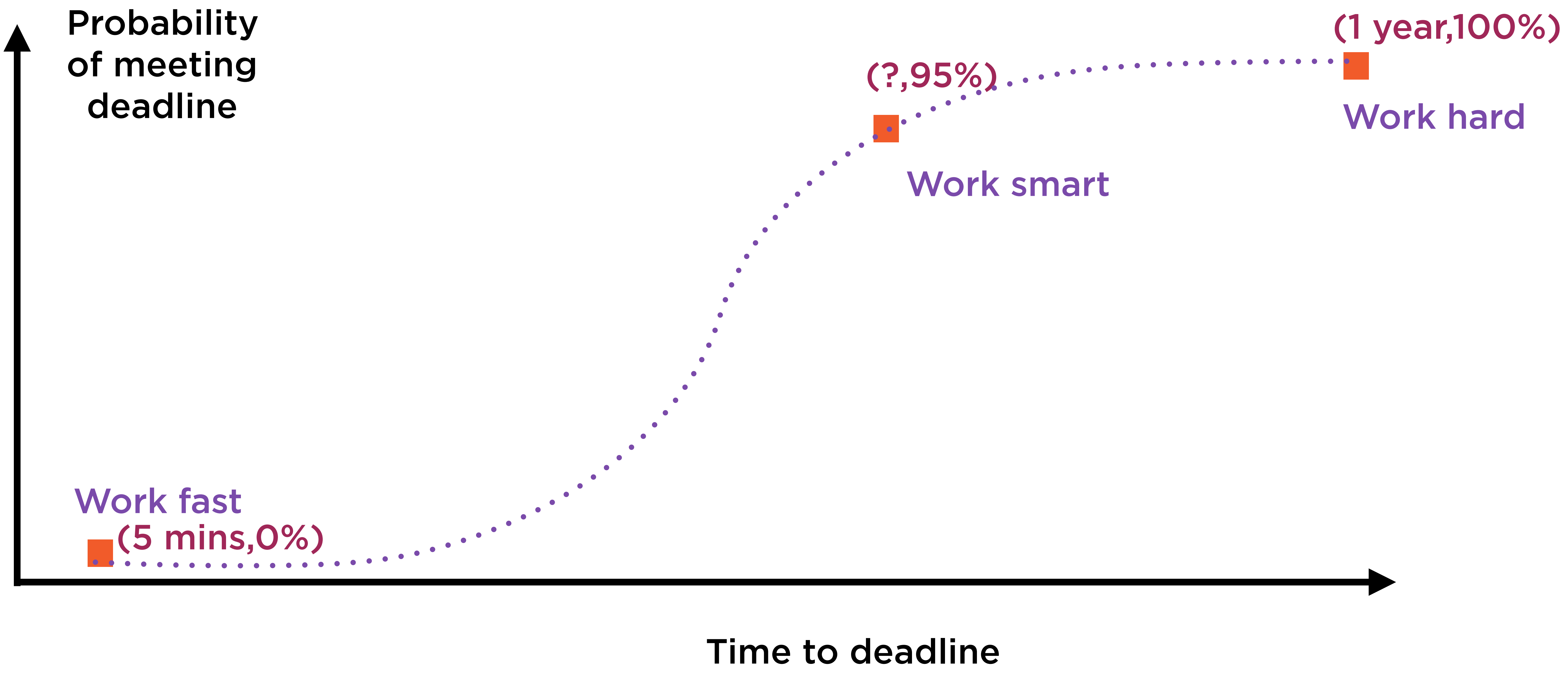
Probability of meeting the deadline

Probability of getting other important work done

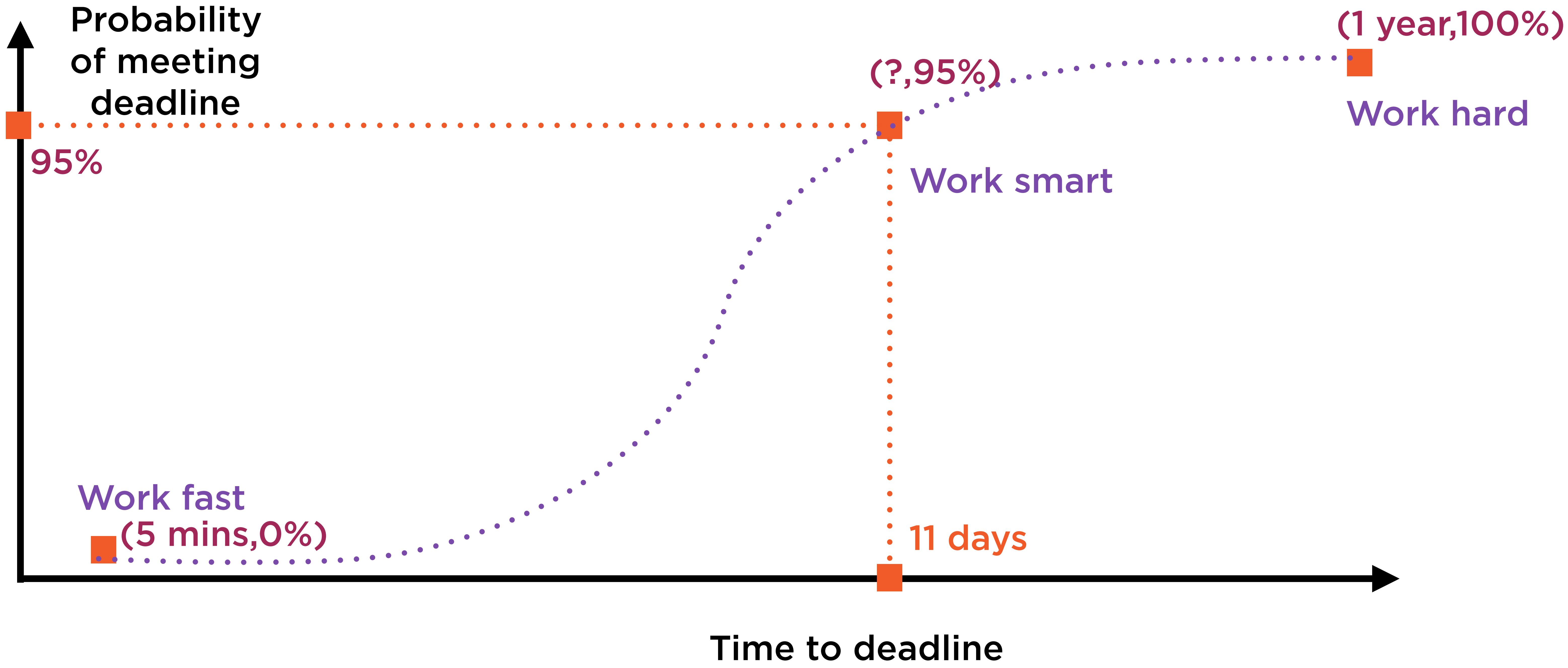
Working Hard, Fast, Smart



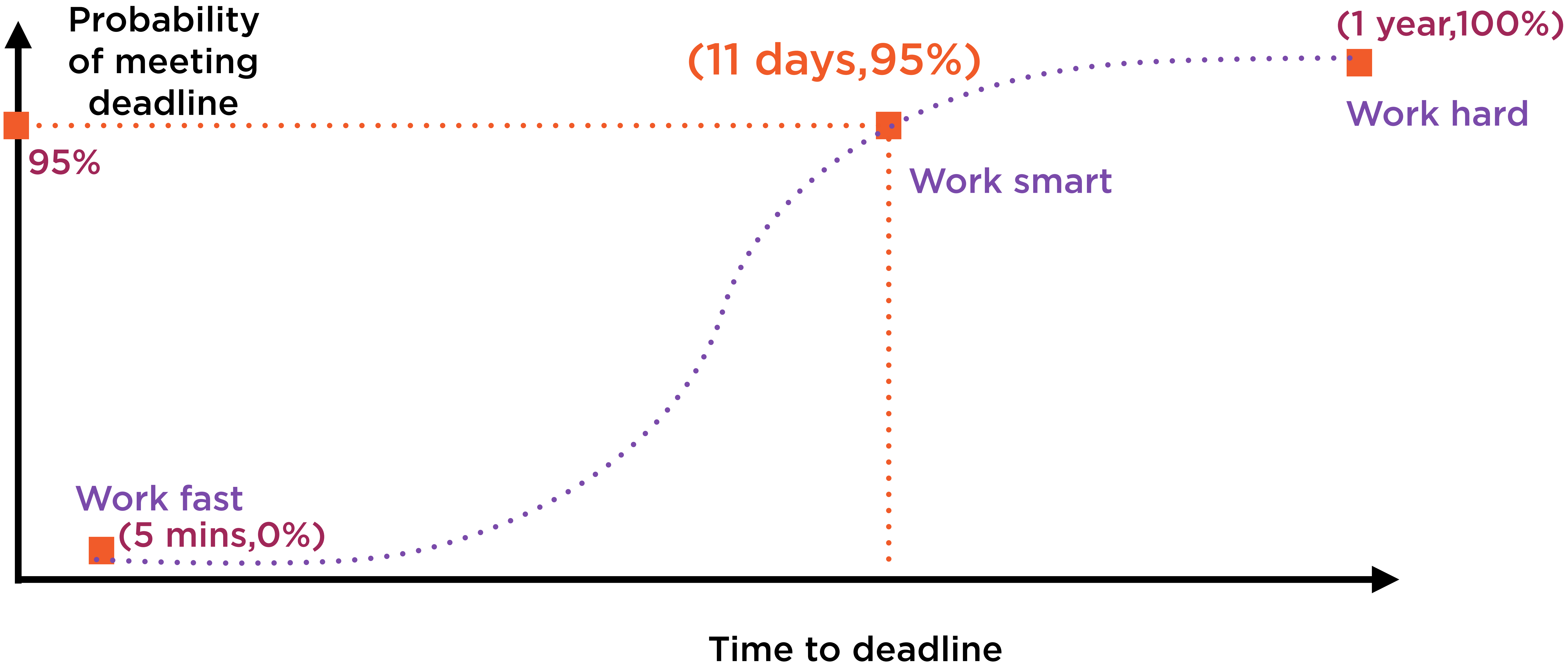
Working Hard, Fast, Smart



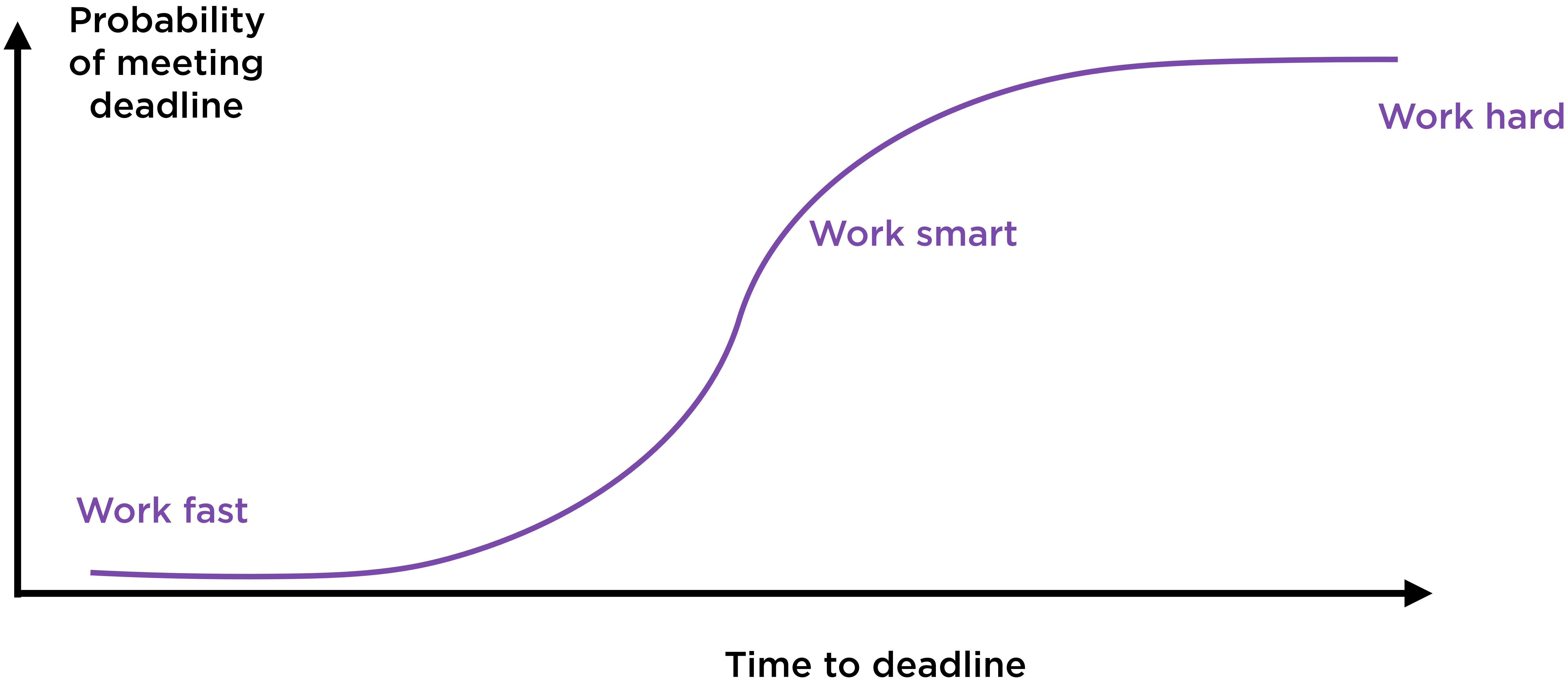
Working Hard, Fast, Smart



Working Hard, Fast, Smart

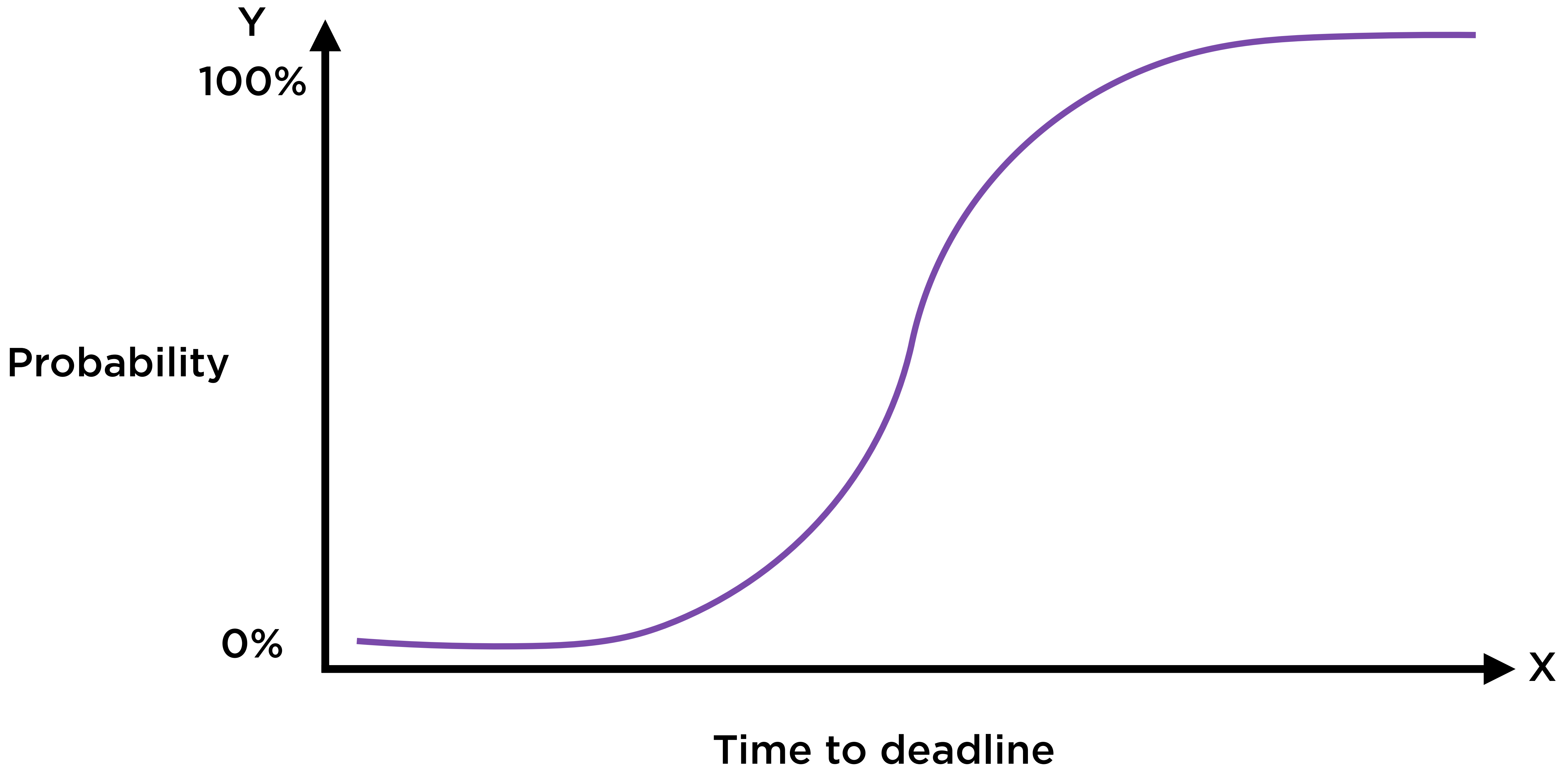


Working Hard, Fast, Smart

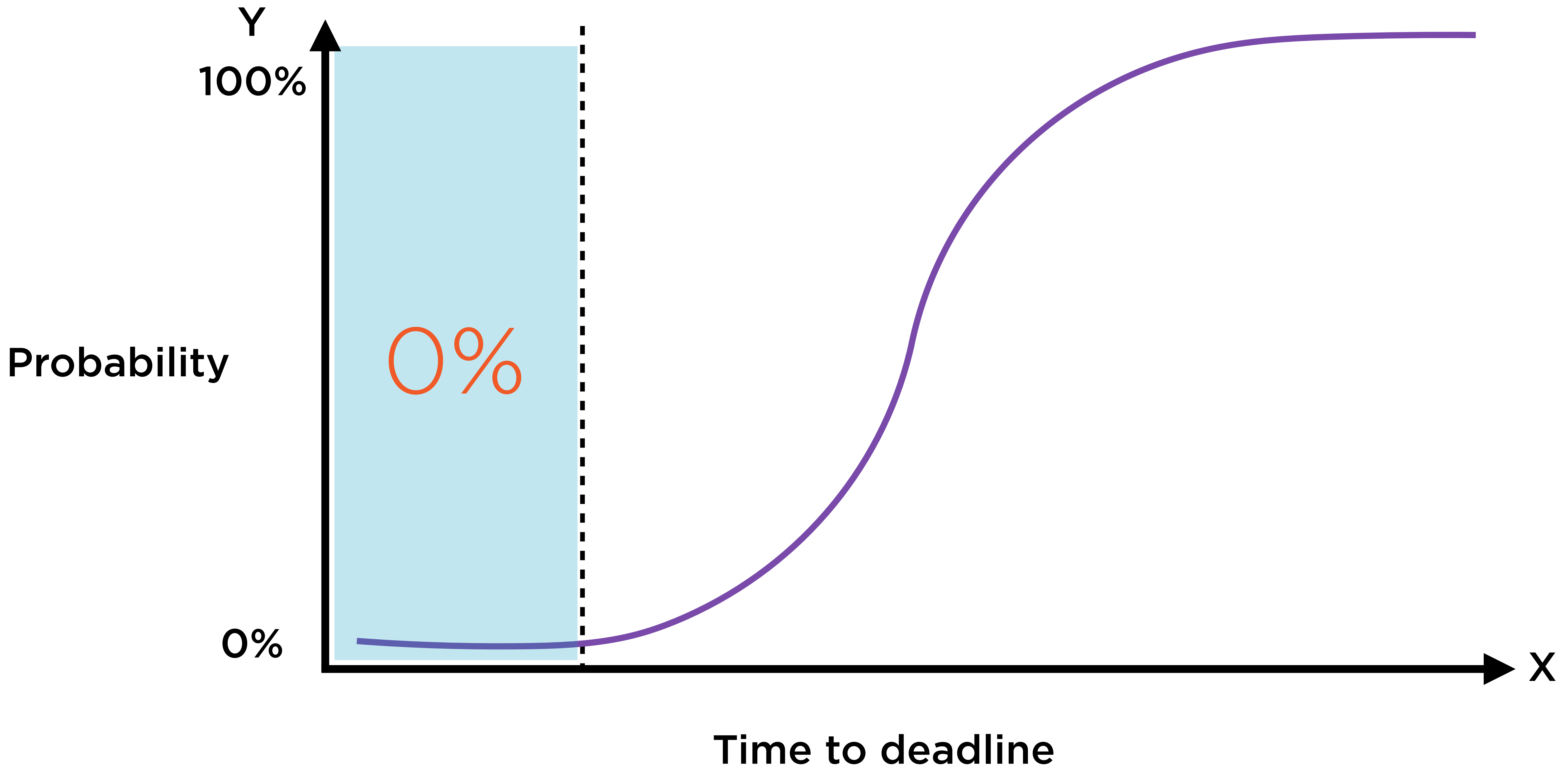


Logistic Regression helps find how probabilities are changed by actions

Working Smart with Logistic Regression

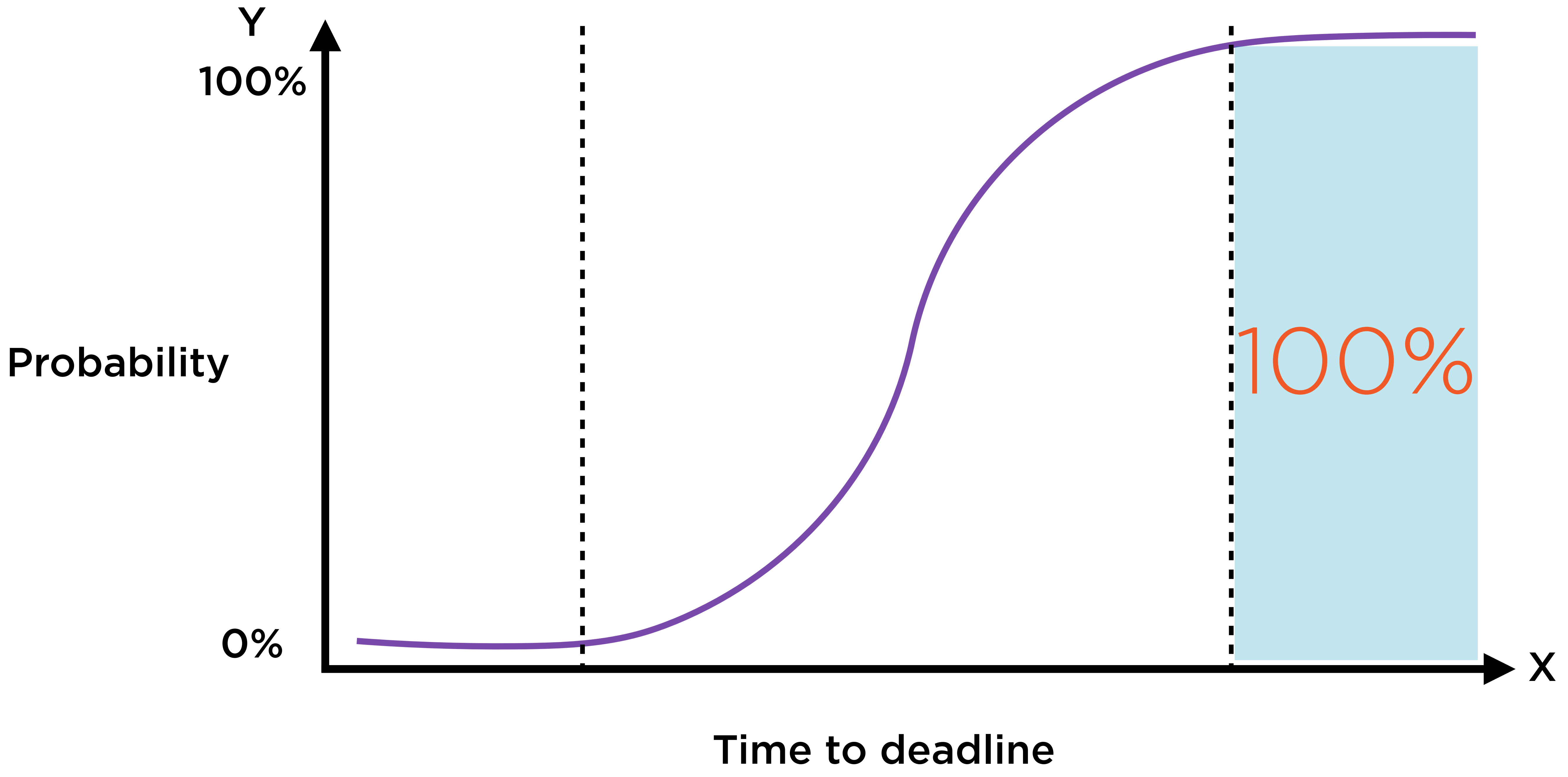


Working Smart with Logistic Regression



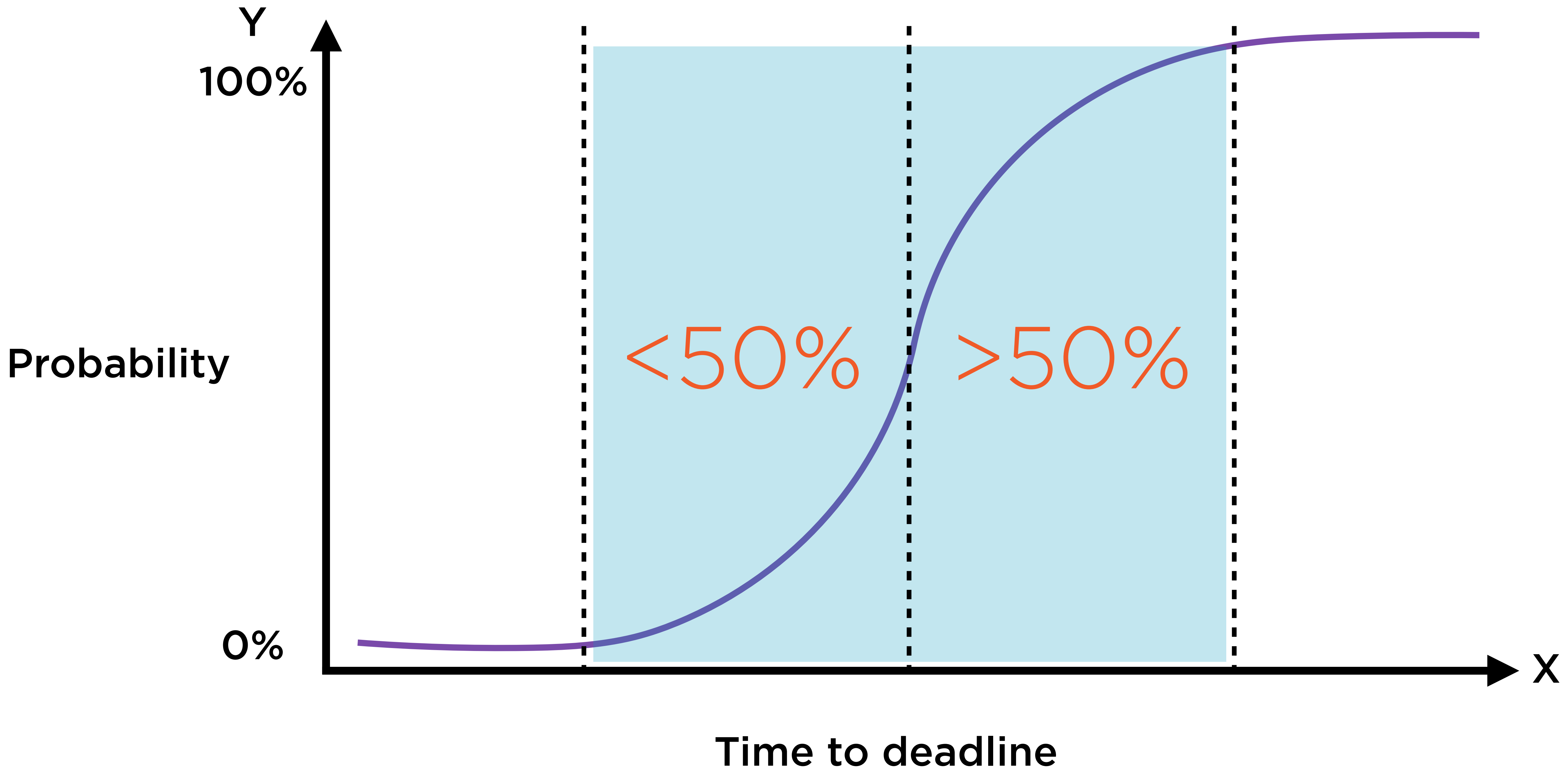
Start too late, and you'll definitely miss

Working Smart with Logistic Regression



Start too early, and you'll definitely make it

Working Smart with Logistic Regression

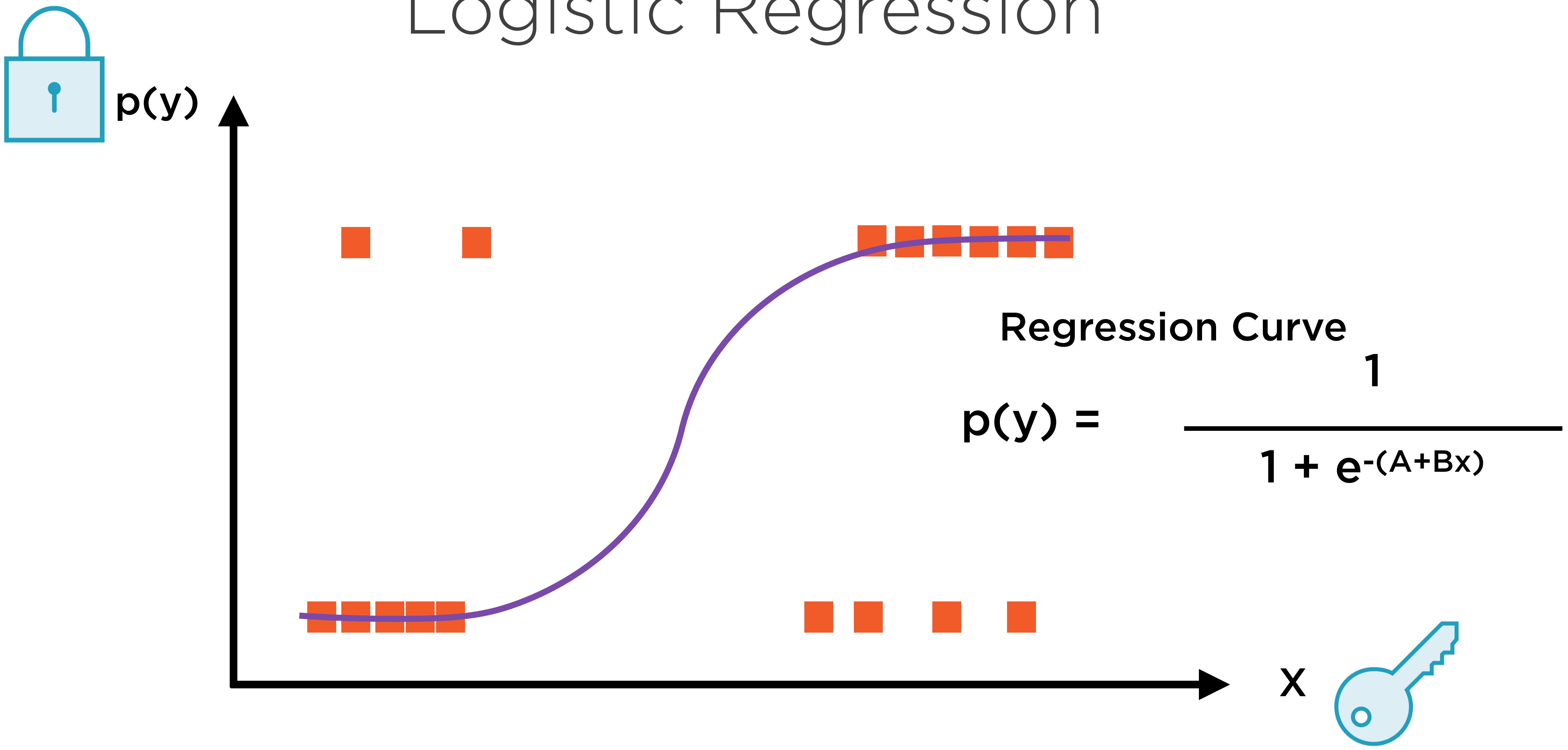


Working smart is knowing when to start

Logistic Regression

Finding the best fit S-curve through these points

Logistic Regression



Finding the best fit S-curve through these points

Logistic Regression

Regression Equation:

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

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

Odds Ratio

Measures strength of association between two events.

Odds Ratio

Consider two events A and B

Odds Ratio helps measure whether A and B are independent

Somewhat similar to correlation, but with important differences

Odds Ratio

Used for categorical events (not continuous variables)

Odds Ratio = 1 implies independence between events

Odds Ratio

If the odds ratio is greater than 1

The presence of B raises the odds of A

Symmetrically the presence of A raises the odds of B

Odds Ratio

If the odds ratio is less than 1

The presence of B reduces the odds of A

Symmetrically the presence of A reduces the odds of B

Forest Plot (a.k.a. Blobbogram)

Visualization used for meta-analysis of results of many different studies addressing the same question, such as different estimates of odds ratio.

Demo

Performing logistic regression and interpreting the results

Summary

Fitting and interpreting models

Understanding linear regression

Interpreting results of linear regression using diagnostic plots

Understanding logistic regression

Interpreting logistic regression using accuracy, sensitivity and specificity

Computing and interpreting odds ratio

Visualizing odds ratio using Forest plot