Lesson: Differentiating prediction and explanation

Module: Prediction and Explanation in Linear Regression Analysis

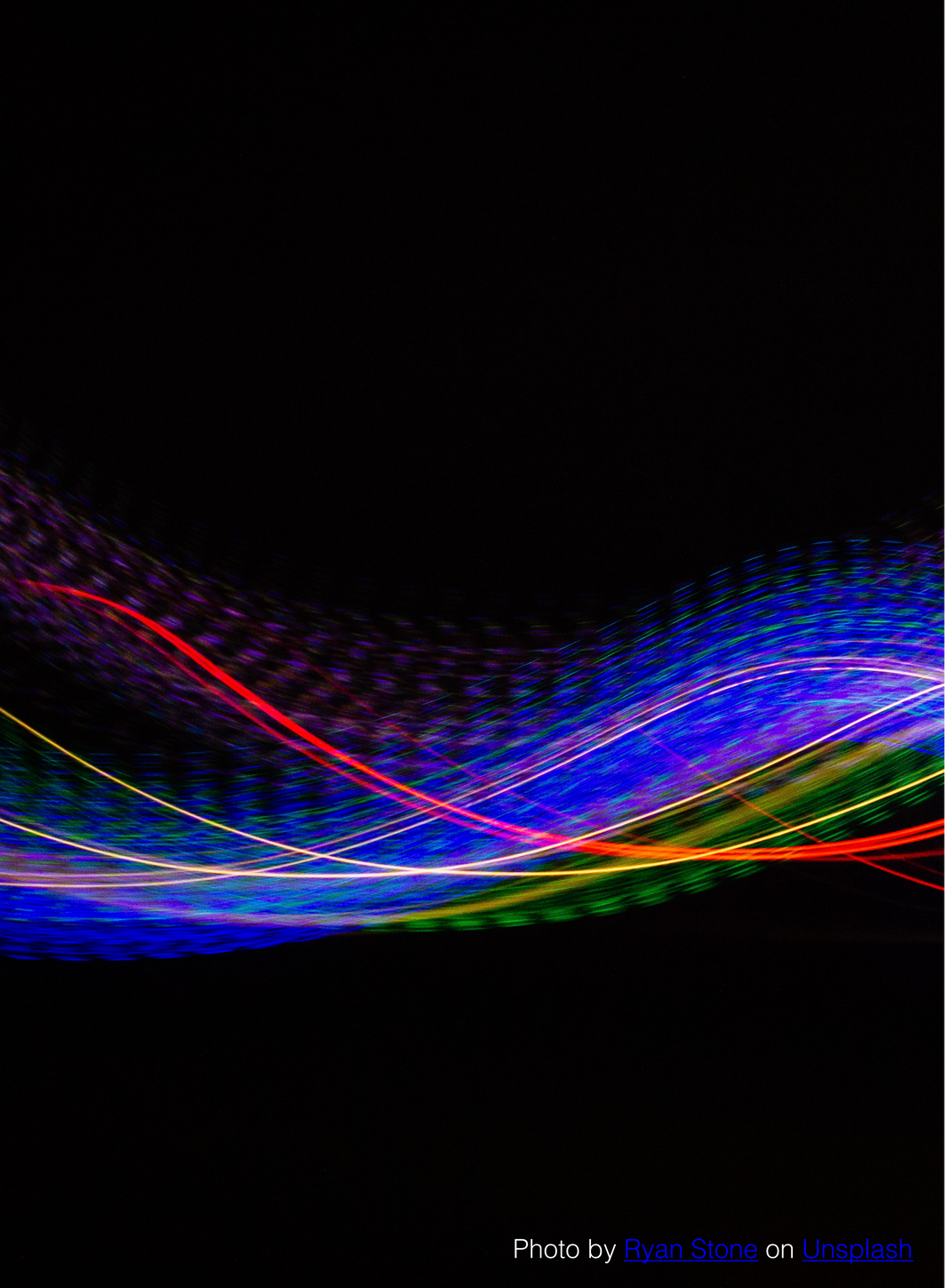


Photo by Ryan Stone on Unsplash

An *explanandum* is a set of propositions that refer to the thing being explained.

An *explanans* is the set of propositions that do the explaining.

Explanatory modeling is an explanans that makes use of statistical models for testing and providing causal explanations.

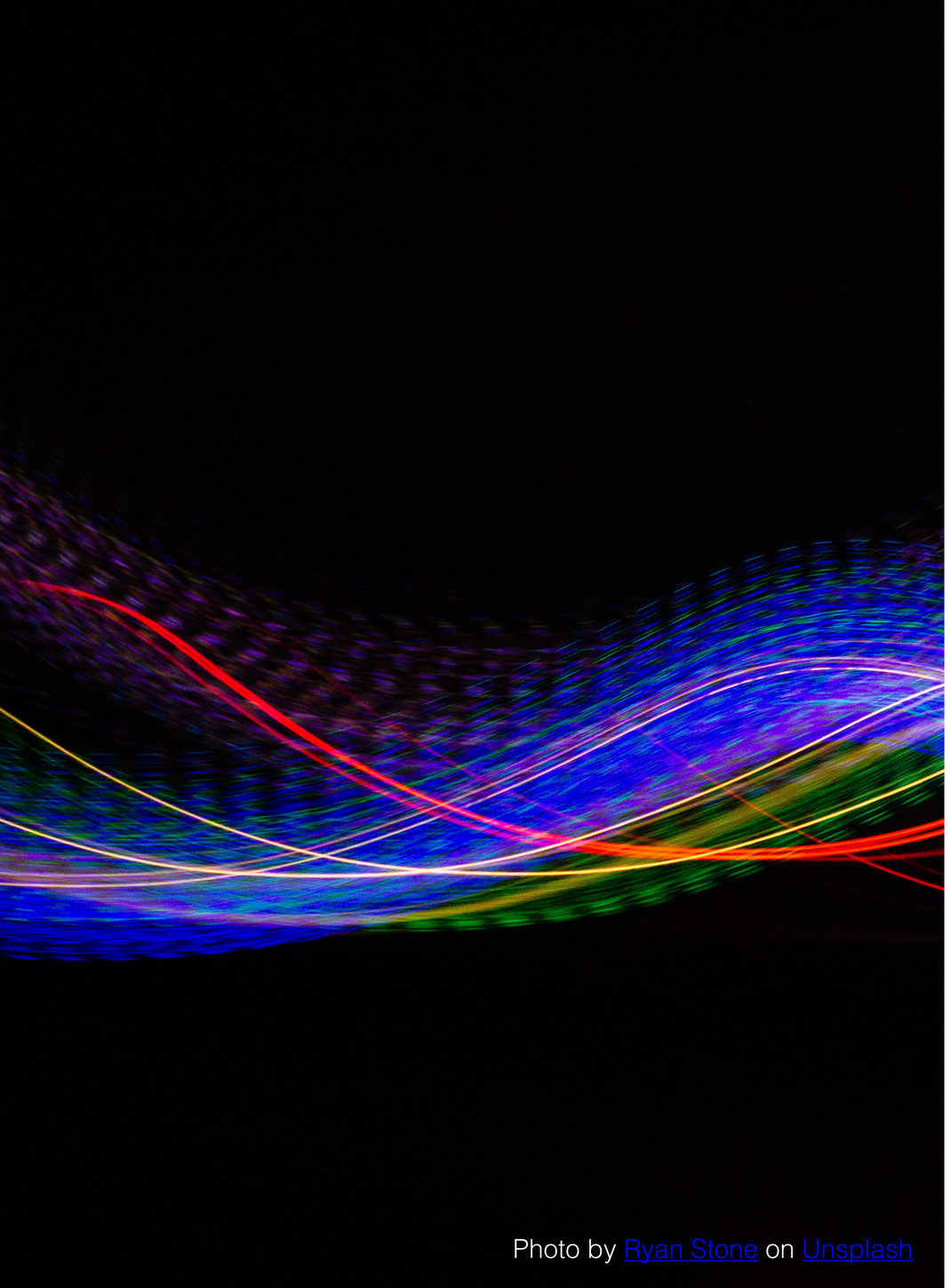


Photo by Ryan Stone on Unsplash

Explanatory models often:

1. Have parameters with a **physical meaning** (e.g., Hooke's Law).

2. Arise in the context of an **experiment design**.

3. Require additional assumptions if the data come from an **observational study**.

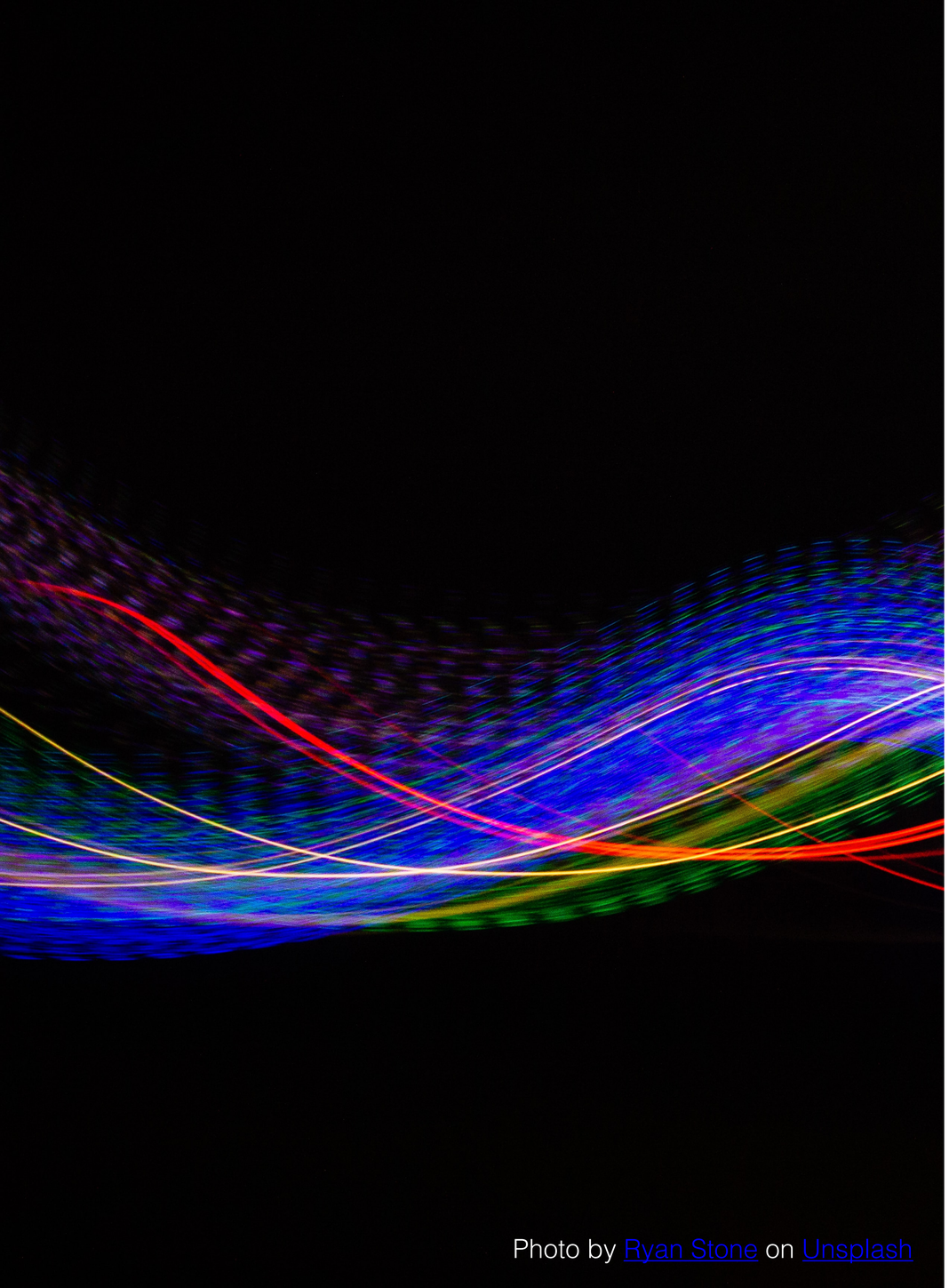


Photo by Ryan Stone on Unsplash

A *prediction* is an empirical consequence of a theory that has not yet been verified at the time the theory was constructed.

A *predictive model* is a statistical model used to provide a value (or range of values) of the response based on values of the predictors that were not used to train the model.

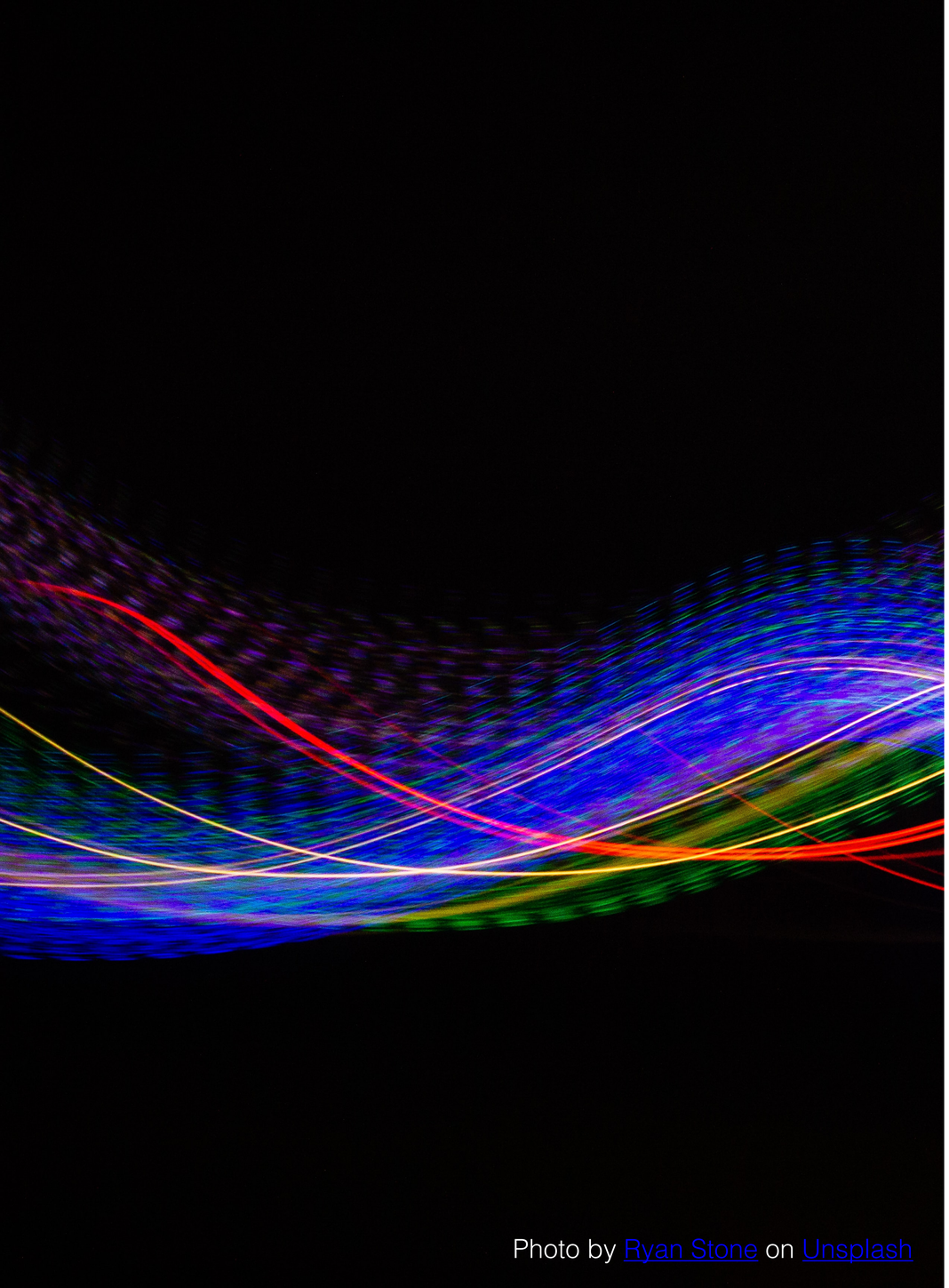


Photo by Ryan Stone on Unsplash

Difference between prediction and explanation?

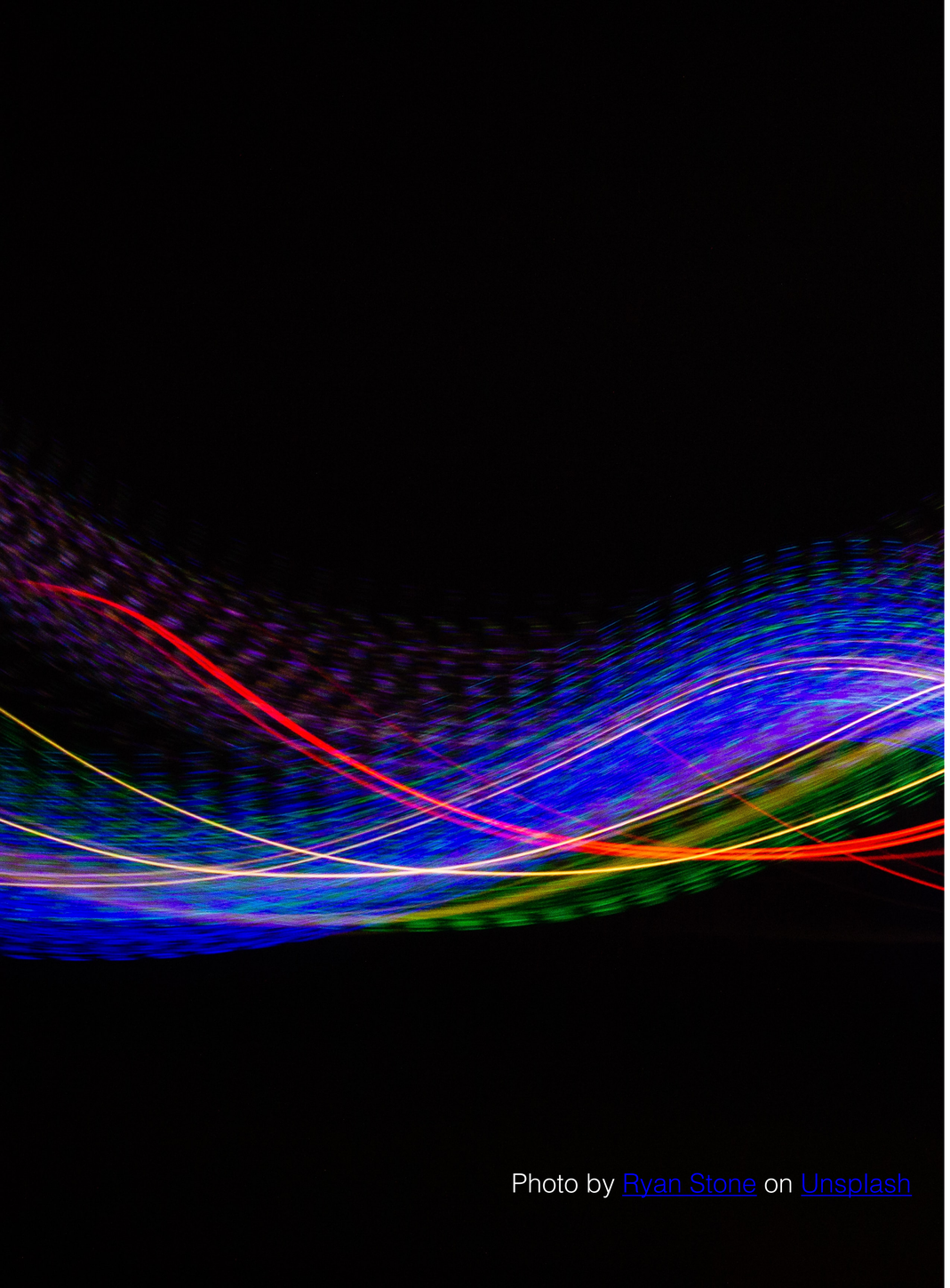
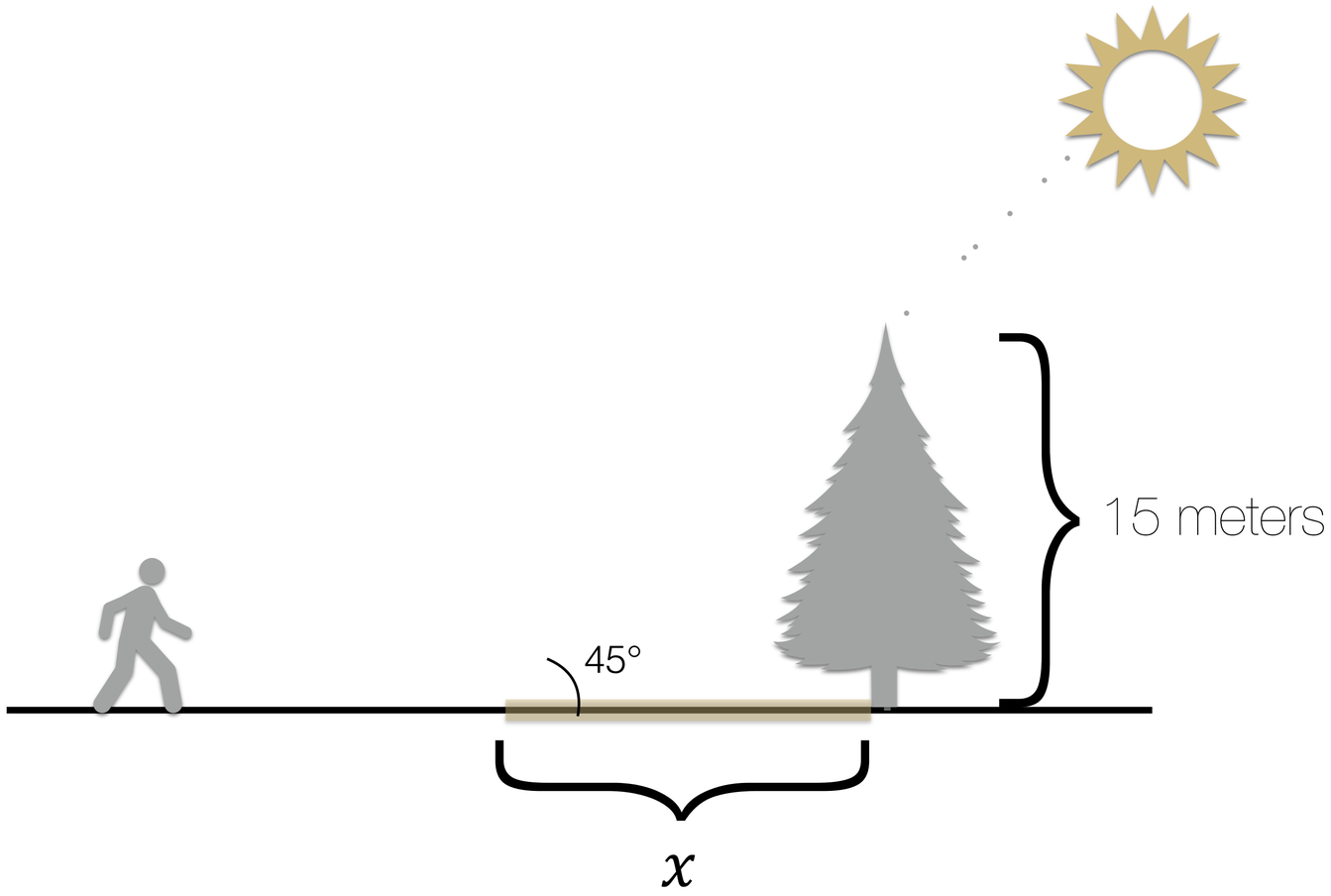


Photo by Ryan Stone on Unsplash



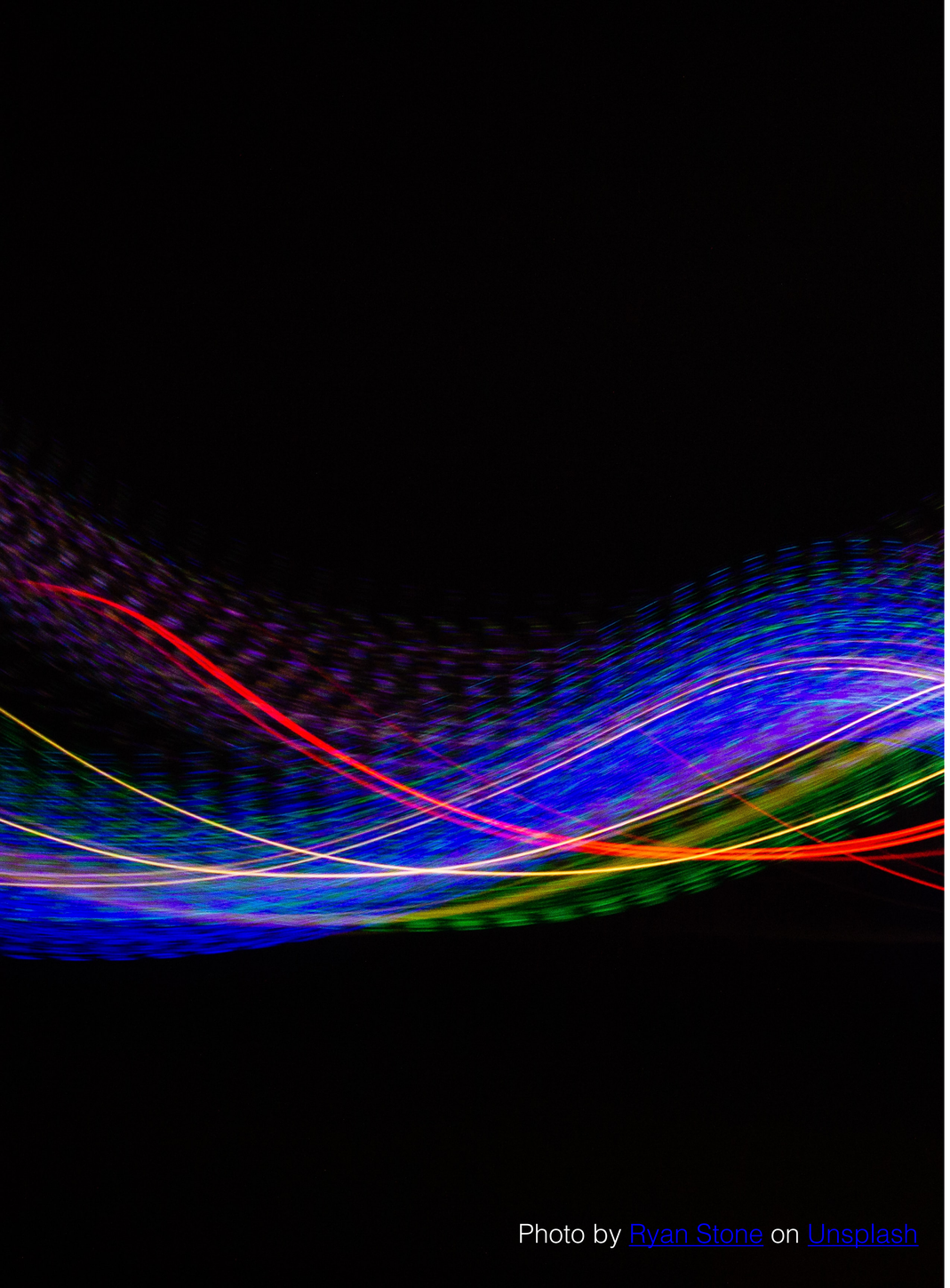
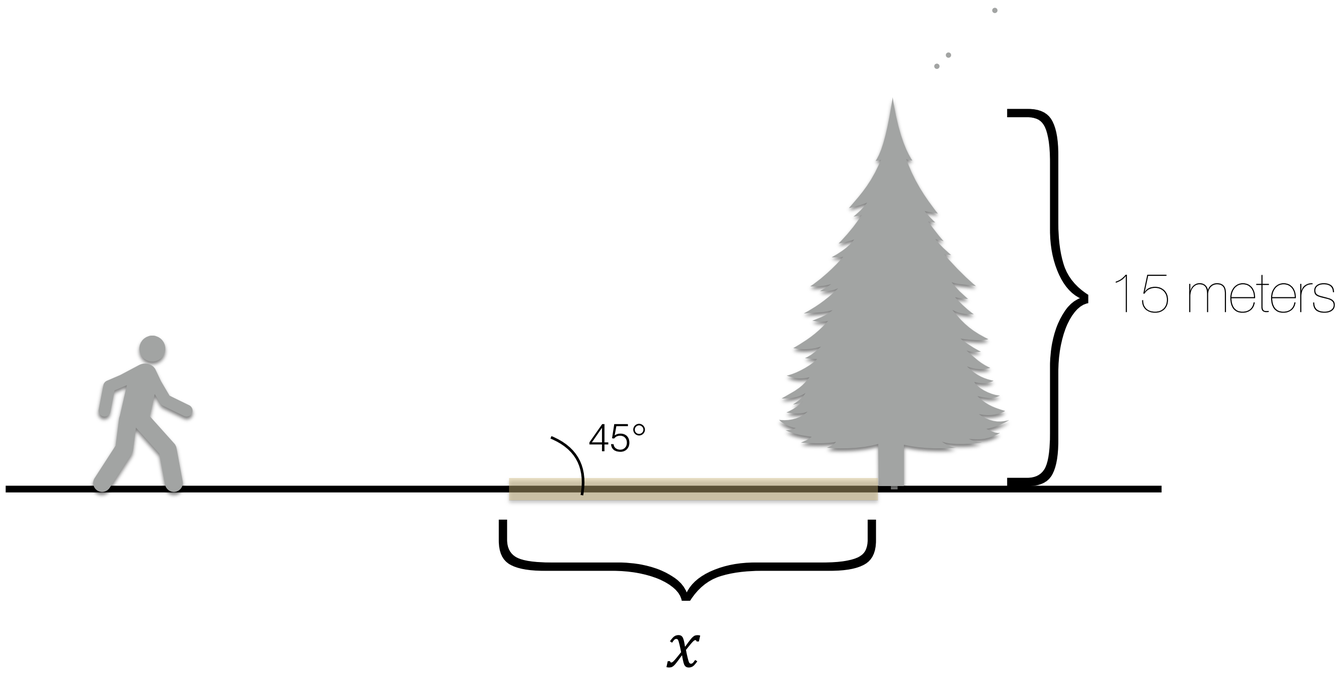


Photo by Ryan Stone on Unsplash

$$tan(45°) = 1 = x/15$$

$$\Longrightarrow x = 15$$



The tree height, angle of incidence, and laws of mathematics explain the shadow.

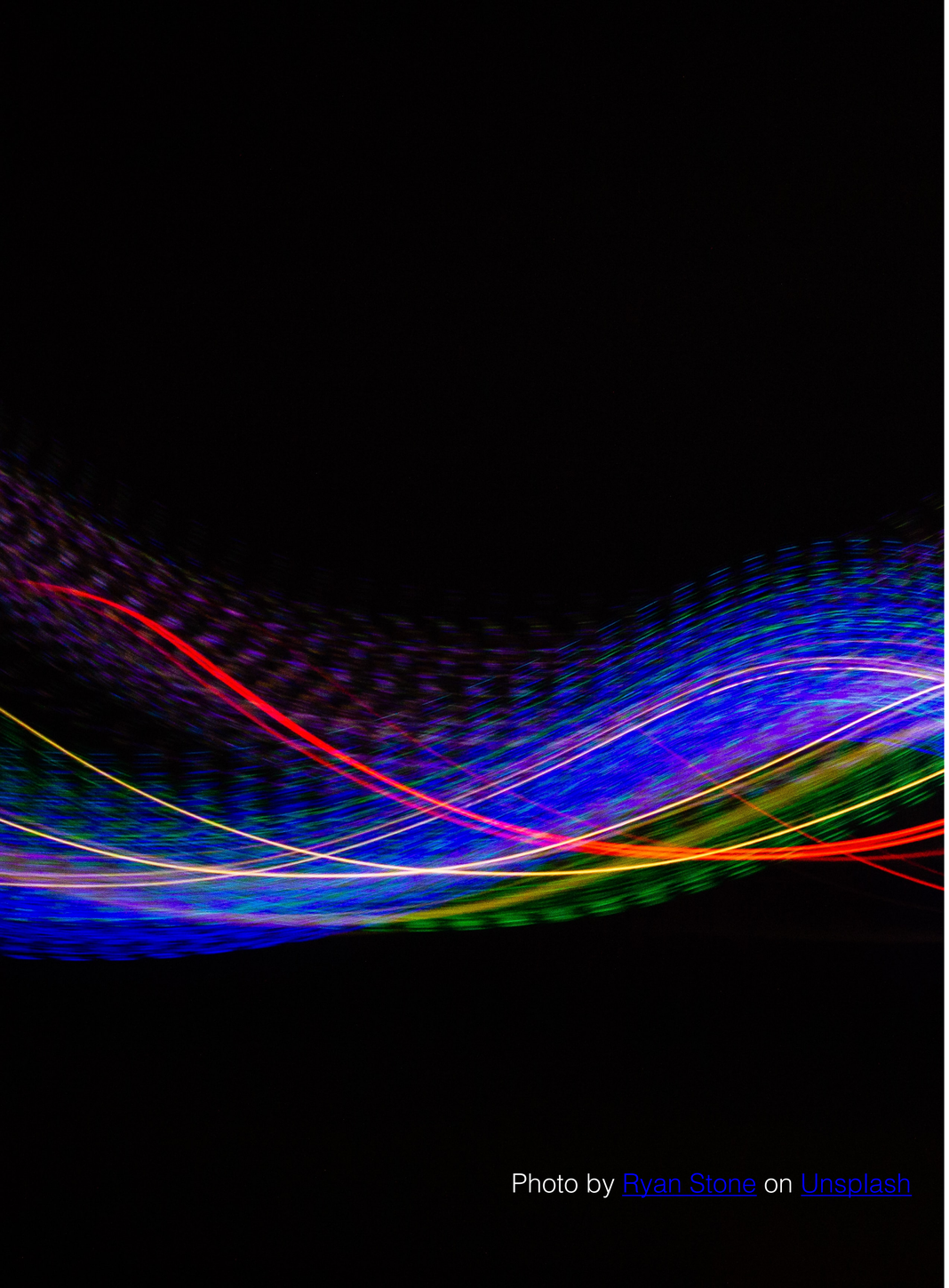
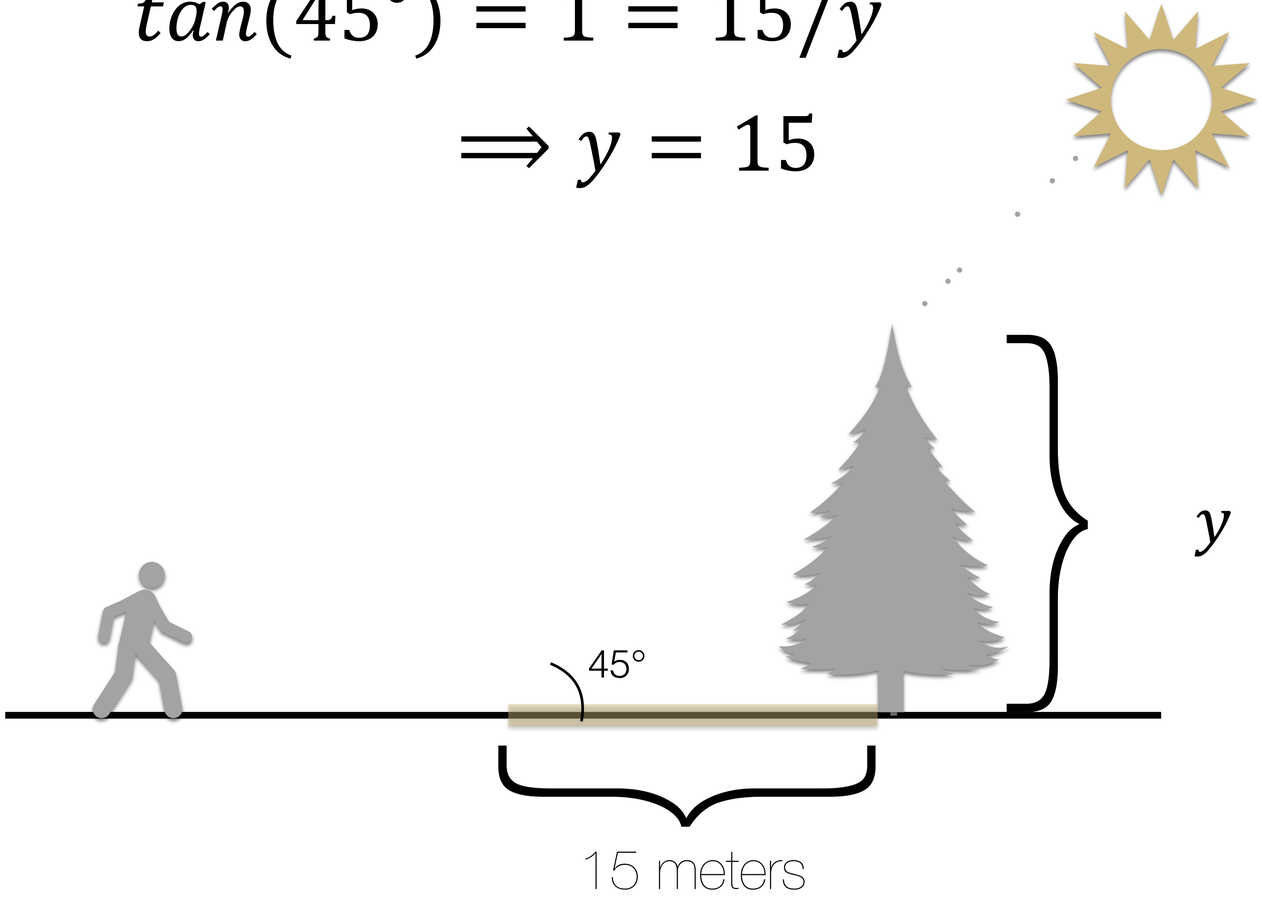


Photo by Ryan Stone on Unsplash

$$tan(45°) = 1 = 15/y$$

$$\Longrightarrow y = 15$$



The length of the shadow, angle of incidence, and laws of mathematics *do not explain* the tree height.

$$F = kx$$

Photo by James Allen on Unsplash

Lesson: Point estimates for prediction

Module: Prediction and Explanation in Linear Regression Analysis

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$$

$$\hat{\mathbf{y}}^{\star} = X^{\star}\hat{\boldsymbol{\beta}}$$

$$\hat{y}_i^\star = \mathbf{x}_i^\star \hat{\boldsymbol{\beta}}$$

$$MSPE = \frac{1}{k}\sum_{i=1}^{k}\left(y_i^\star - \hat{y}_i^\star\right)^2$$

$$= \frac{1}{k}\sum_{i=1}^{k}\left(y_i^\star - \mathbf{x}_i^\star\hat{\boldsymbol{\beta}}\right)^2$$

Lesson: Interval estimates for prediction

Module: Prediction and Explanation in Linear Regression Analysis

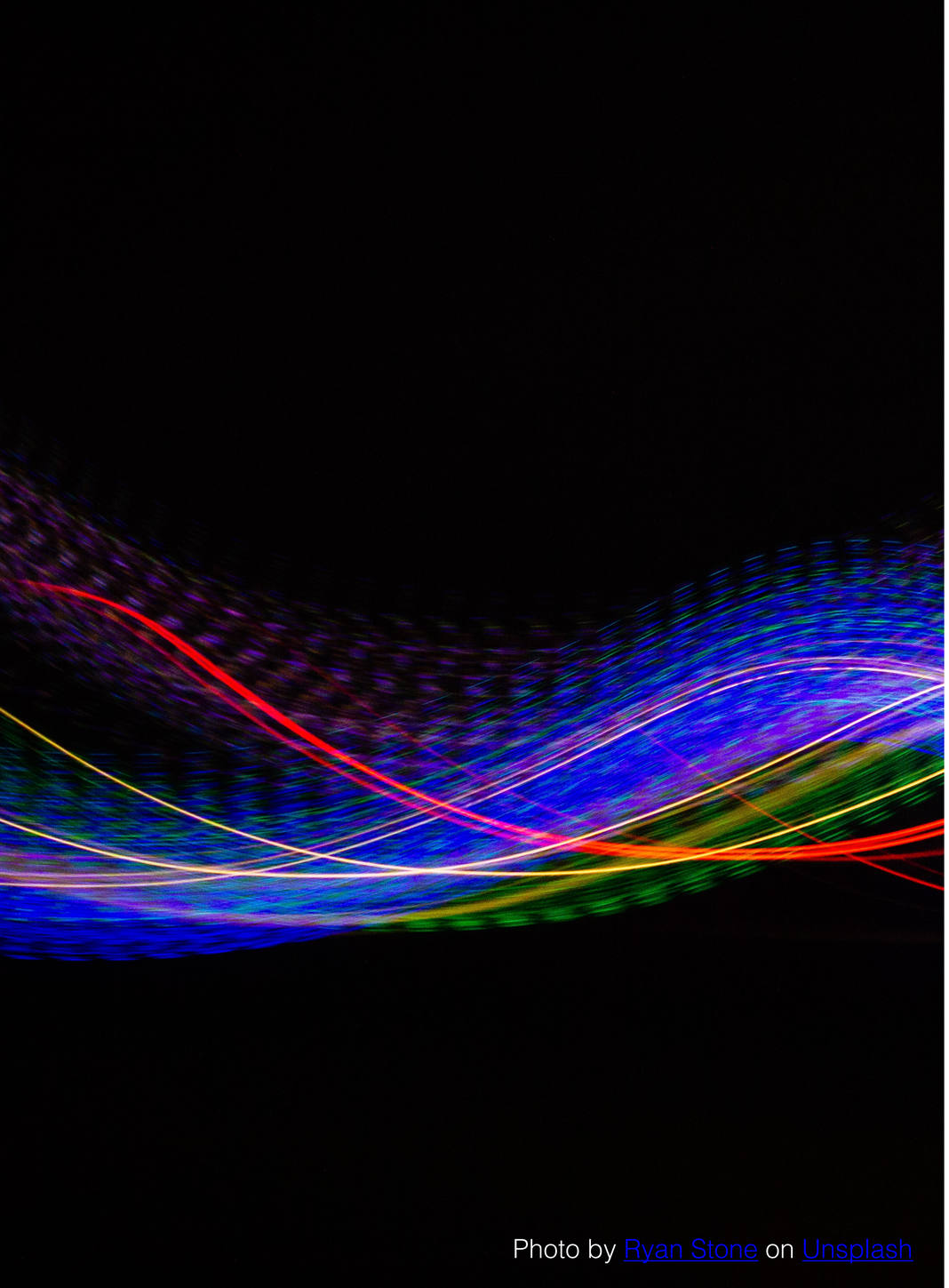


Photo by Ryan Stone on Unsplash

Prediction vs explanation

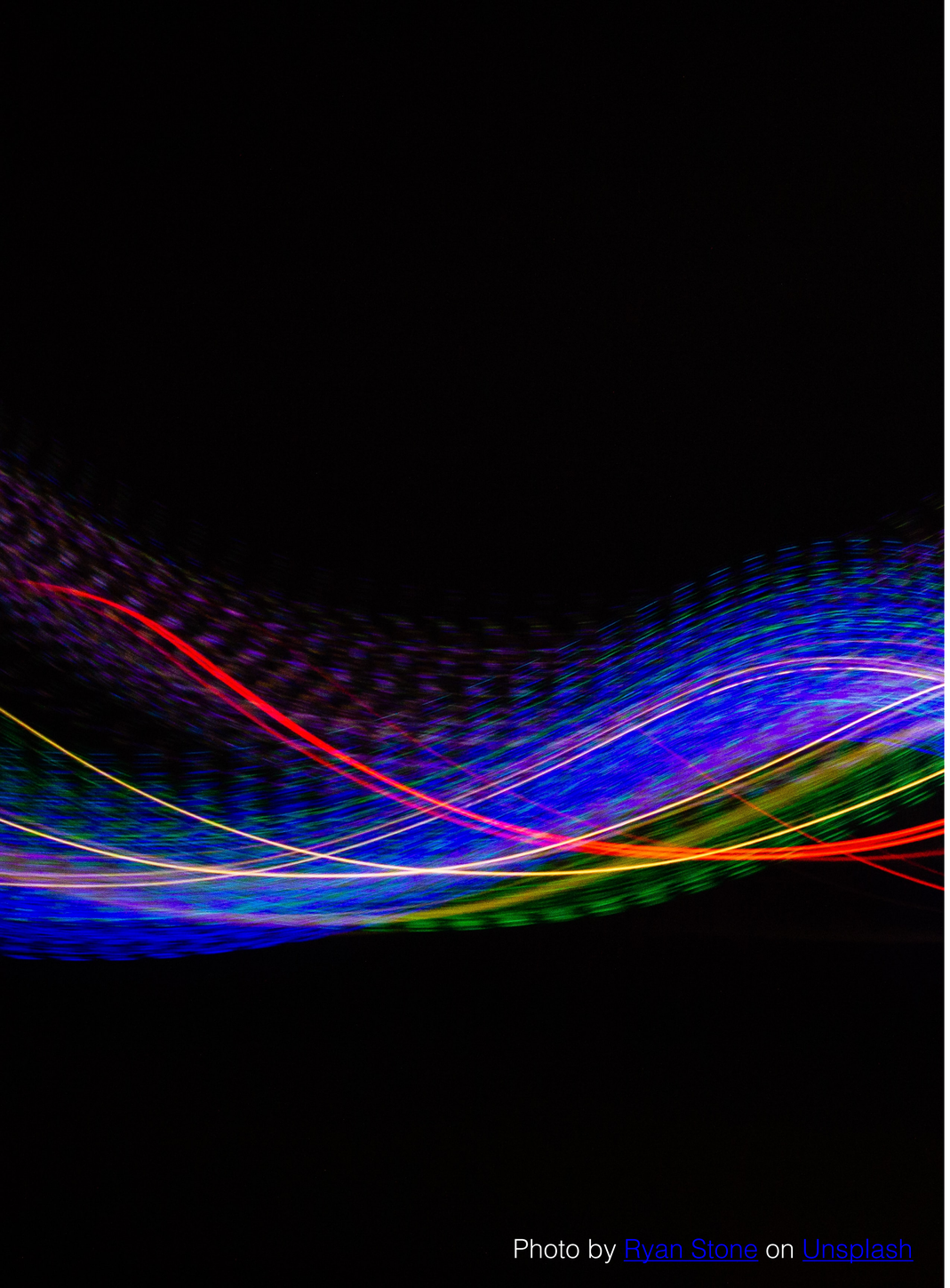


Photo by Ryan Stone on Unsplash

$$\hat{y}_i \pm t_{\alpha/2}(n-(p+1))\sqrt{\hat{\sigma}^2 \mathbf{x}_i^{\star}(X^T X))^{-1}\mathbf{x}_i^{\star T}}$$

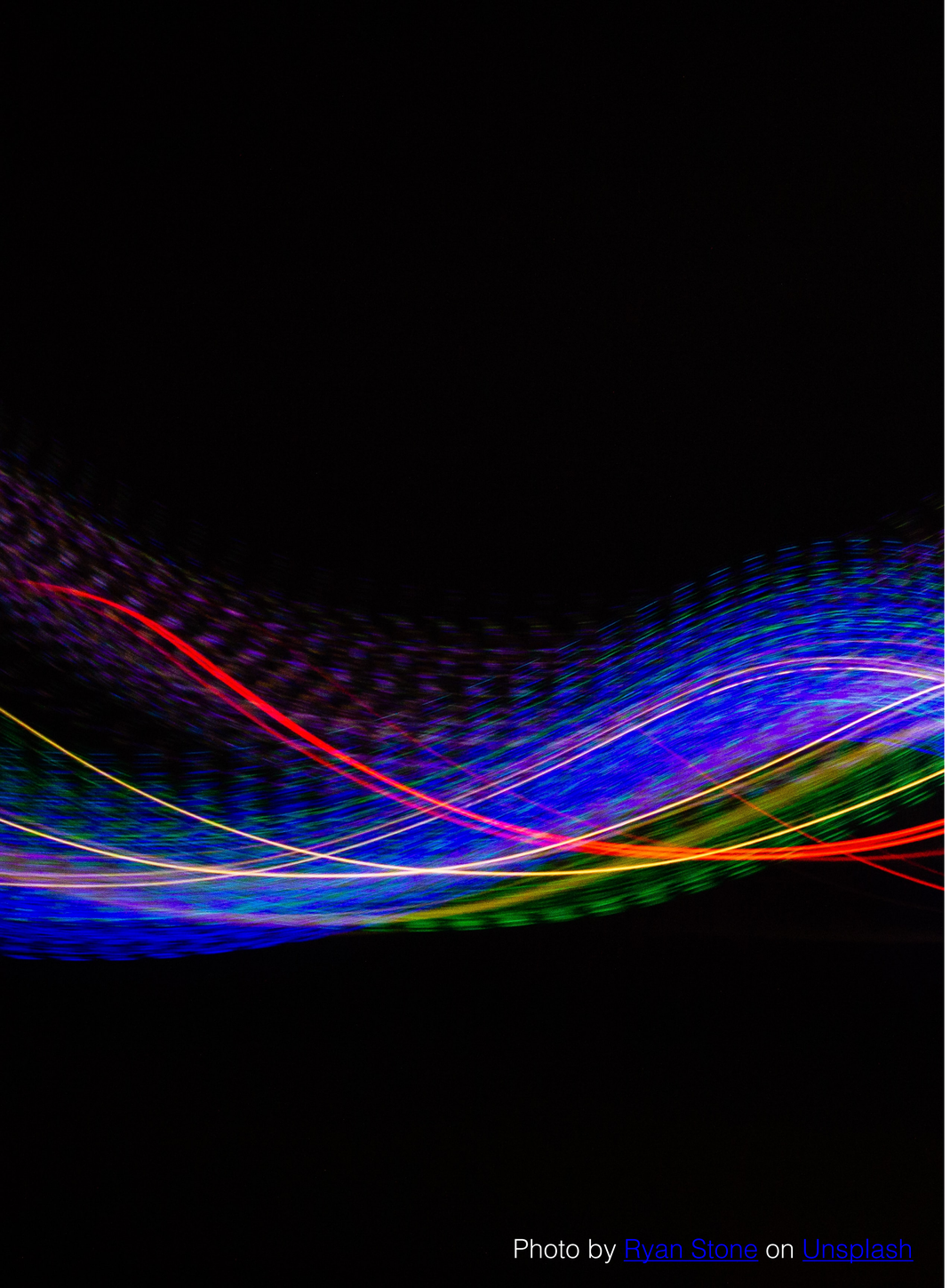
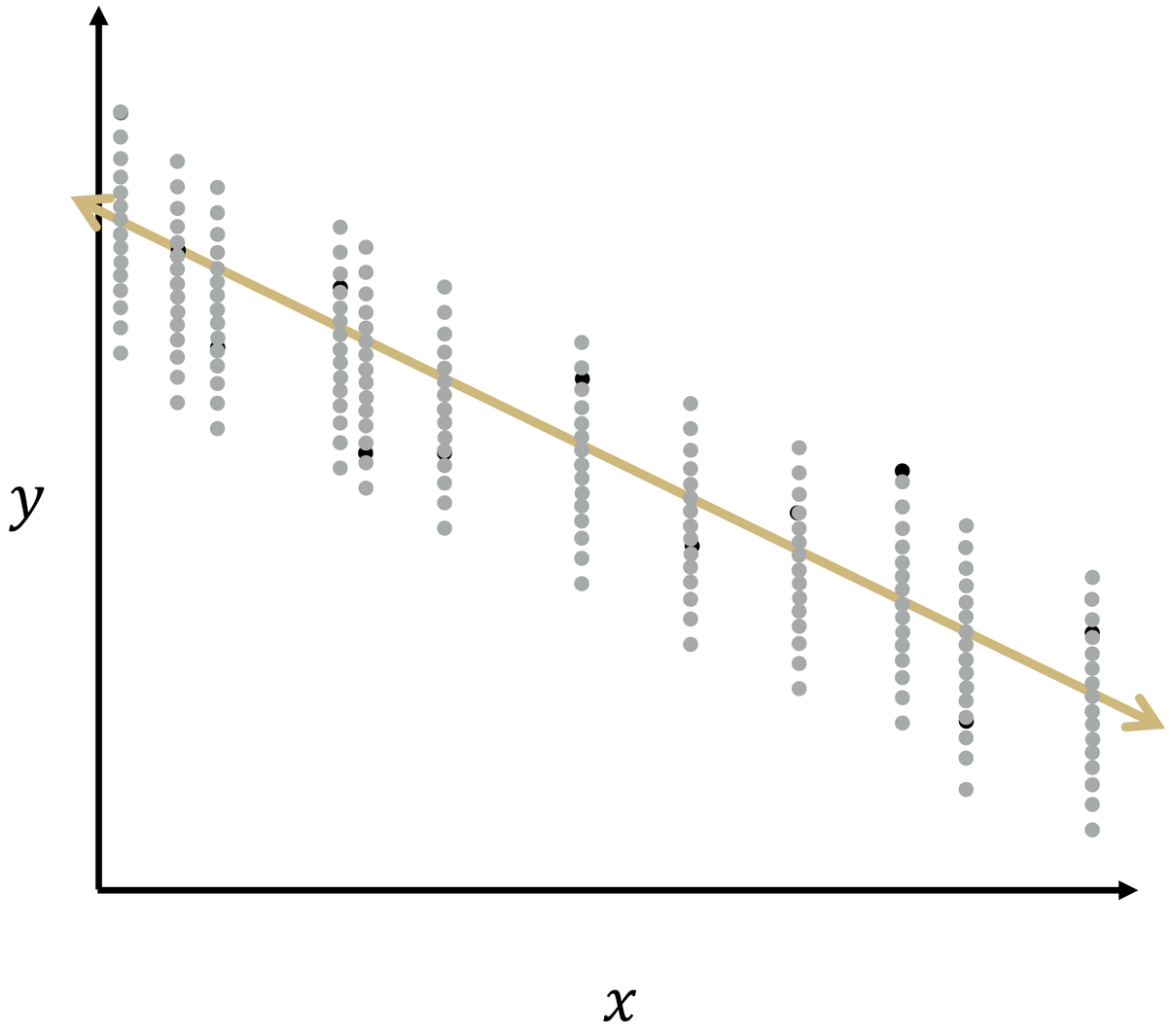


Photo by Ryan Stone on Unsplash



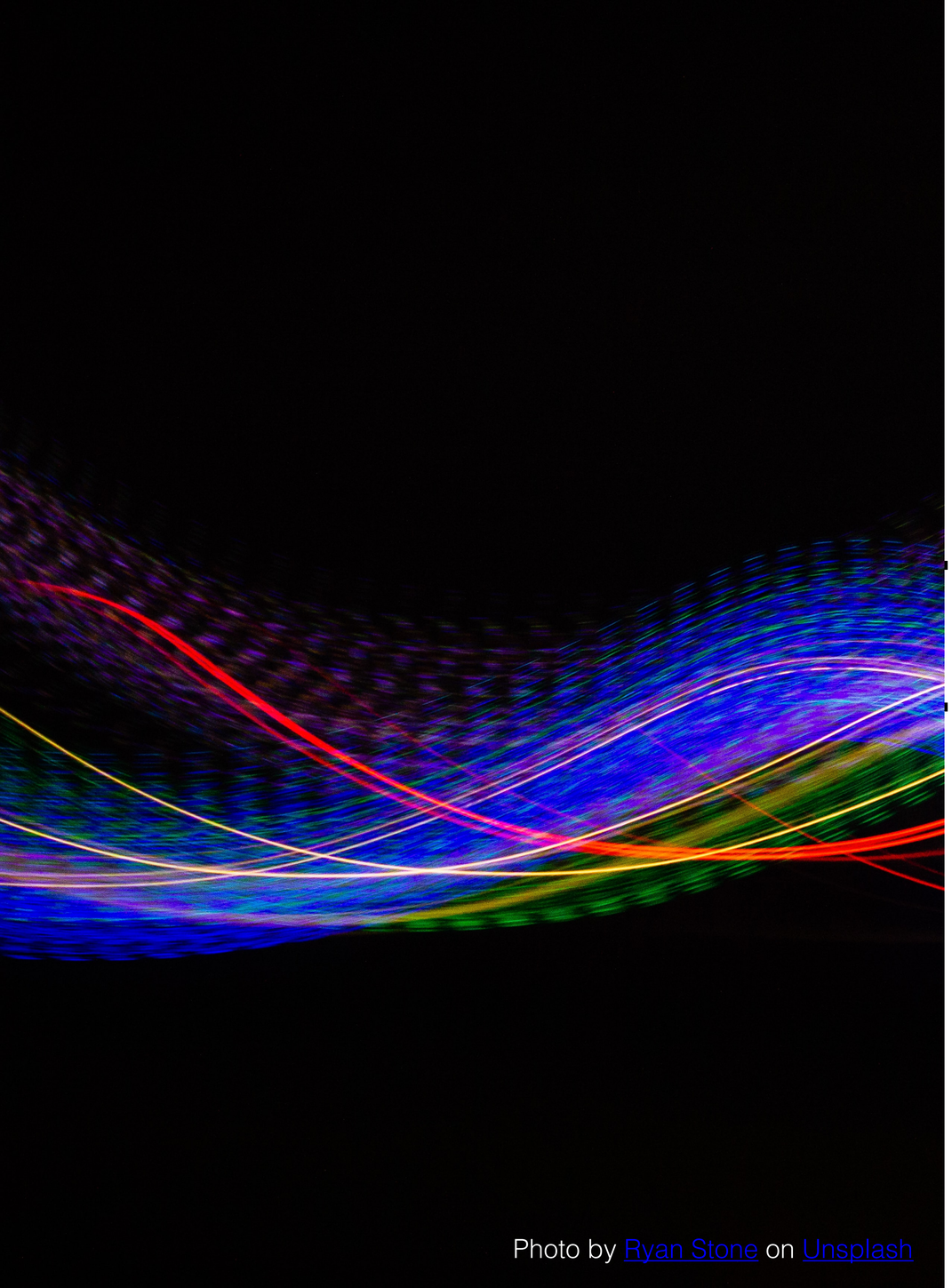


Photo by Ryan Stone on Unsplash

- **Confidence interval** for the *mean response*.
- **Prediction interval** for a *future observation* or *observation not used to train the model*.

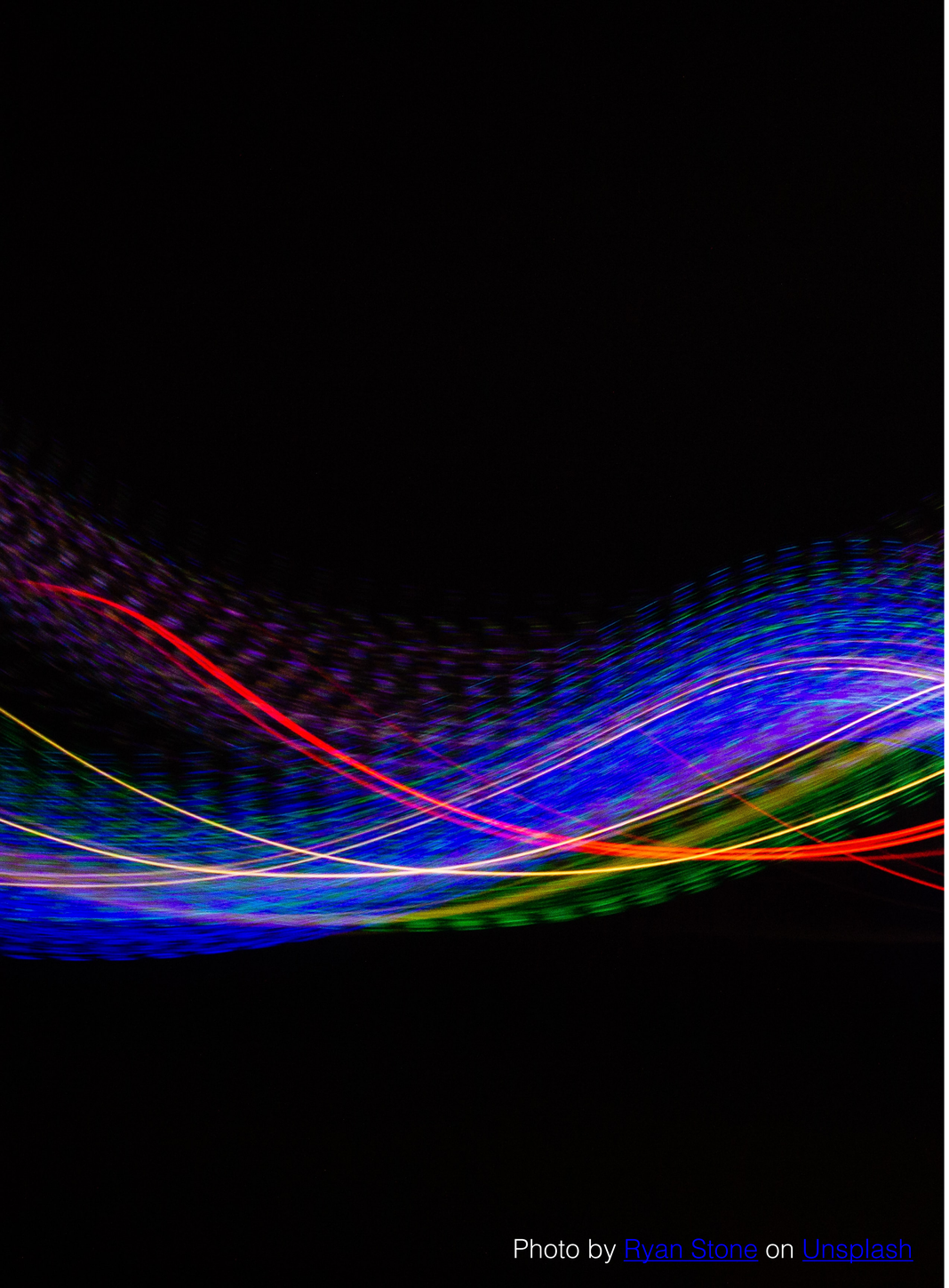


Photo by Ryan Stone on Unsplash

For our prediction interval, we'll need:

$$Var(\mathbf{x}_i^\star \hat{\beta} + \varepsilon_i) =$$

Thus, a prediction interval for a ***future observation*** or ***observation not used to train the model,*** given predictor $\mathbf{x}_i^\star$ is:

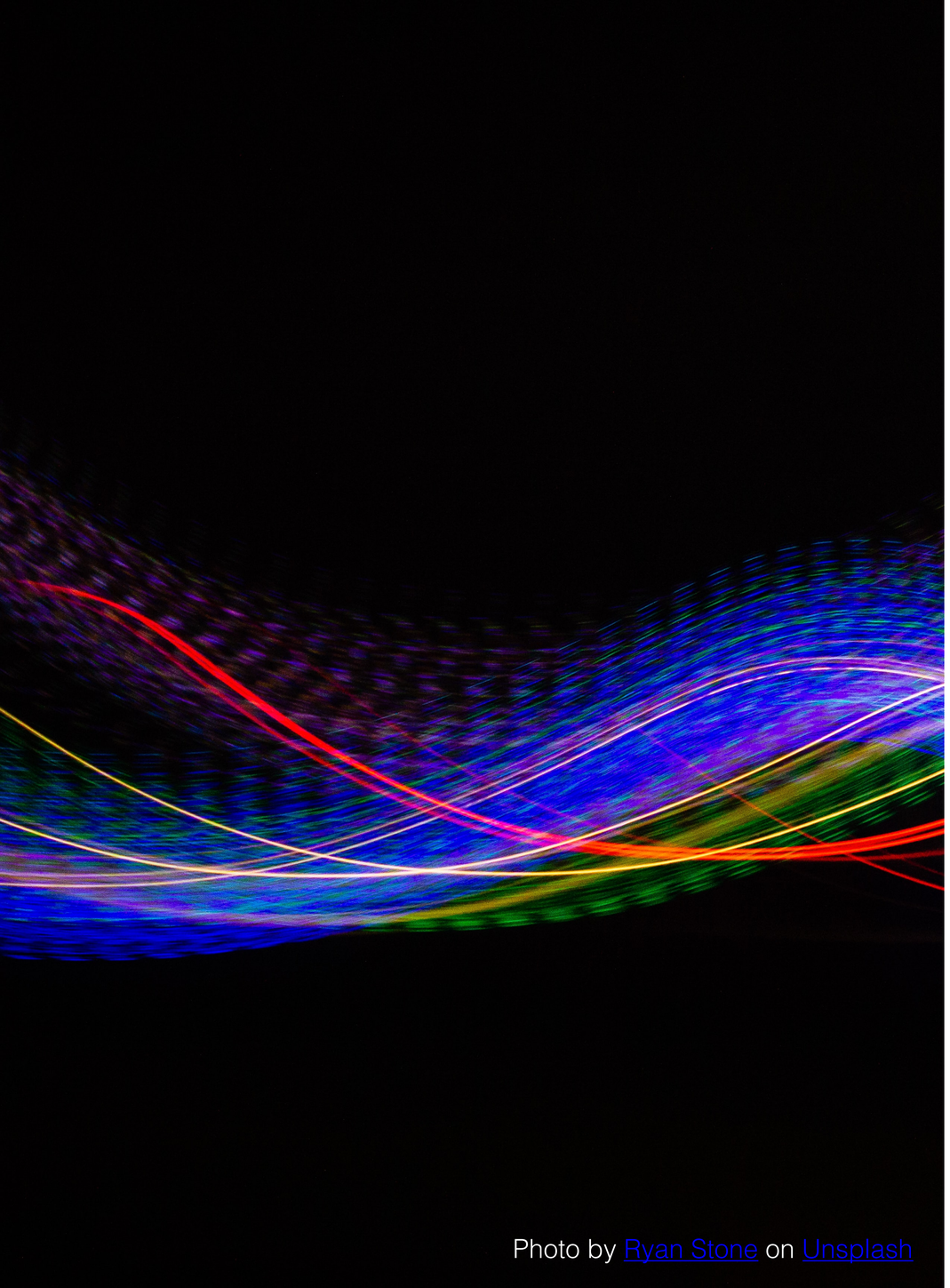


Photo by Ryan Stone on Unsplash

Prediction interval for a future observation in R:

predict(lm_data, new = $\mathbf{x}_i^\star$, interval = “prediction”)

Lesson: When prediction goes wrong

Module: Prediction and Explanation in Linear Regression Analysis

Photo by Jen Theodore on Unsplash

What can go wrong with predictions:

1. Poorly fitting model (overfitting, not enough data to fit properly).

2. Quantitative extrapolation.

3. Qualitative extrapolation.

Photo by Jen Theodore on Unsplash

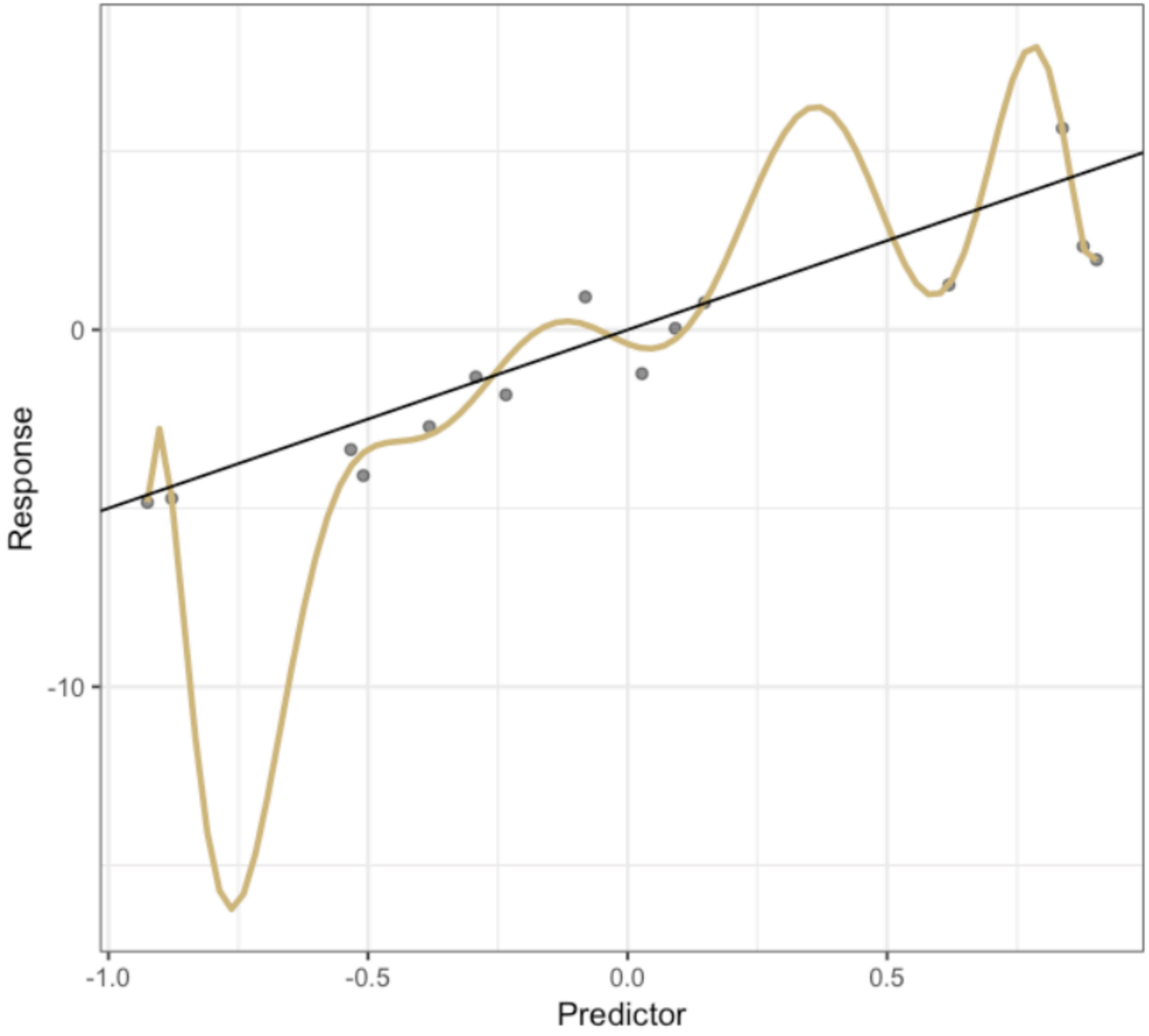


Photo by Jen Theodore on Unsplash

What can go wrong with predictions:

1. Poorly fitting model (overfitting, not enough data to fit properly).

2. Quantitative extrapolation.

3. Qualitative extrapolation.

Photo by Jen Theodore on Unsplash

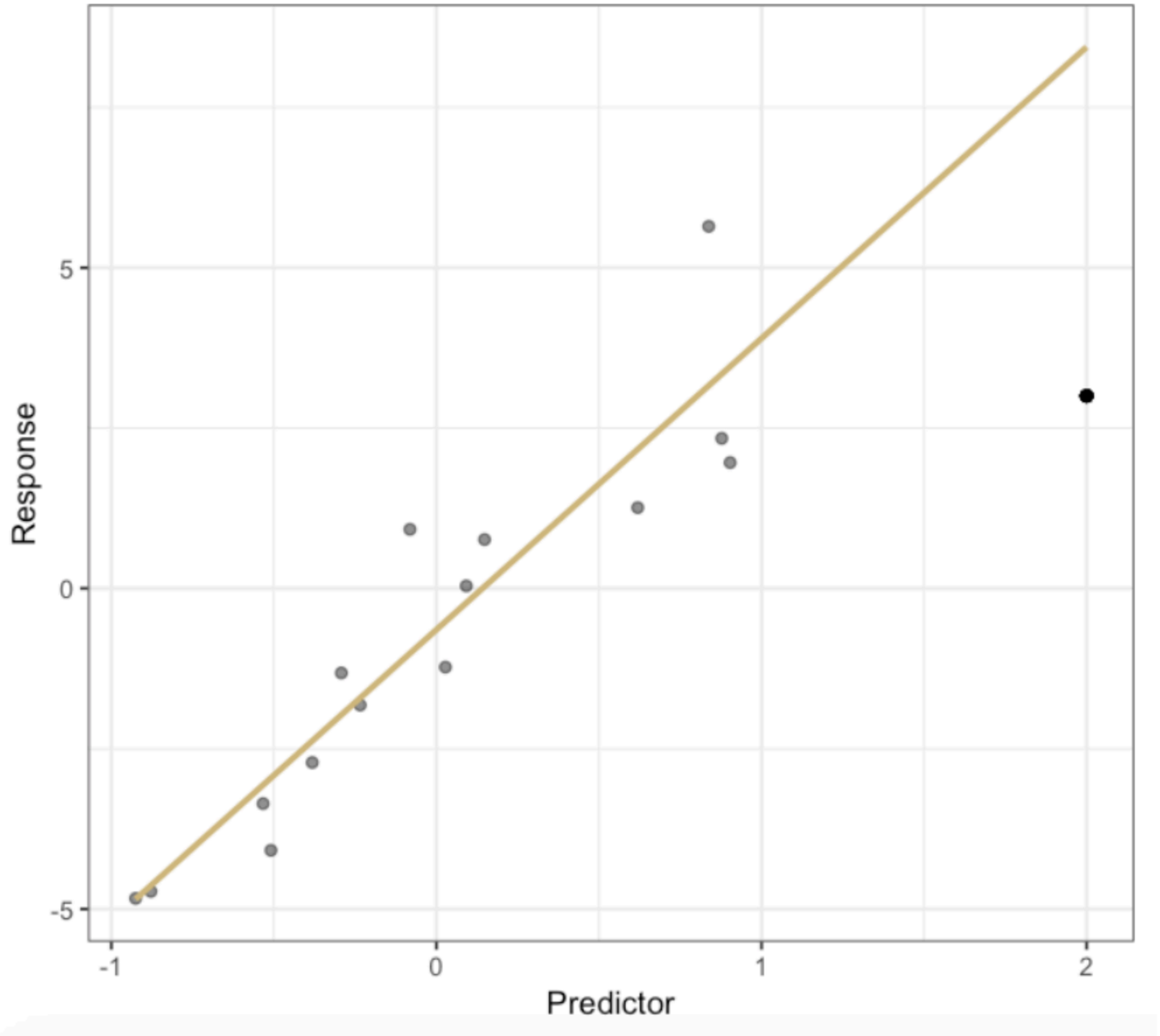


Photo by Jen Theodore on Unsplash

What can go wrong with predictions:

1. Poorly fitting model (overfitting, not enough data to fit properly).

2. Quantitative extrapolation.

3. Qualitative extrapolation.

Lesson: Defining causality

Module: Prediction and Explanation in Linear Regression Analysis

Recall: explanation typically means *causal explanation.* So:

1. What is a cause?
   (metaphysical/ontological question)

2. How can we learn about causes?
   (epistemological question)

Definition: The *relata* are the terms that show up in the causal relation.

Ontological questions:

1. What are the *relata* in a causal relationship?

2. How many *relata* are there?

Is a causal relationship ***deterministic*** or ***probabilistic***?

- A theory about the nature of causality is ***deterministic*** if the effect necessarily follows from the cause.

- A theory about the nature of causality is ***probabilistic*** if the existence of a cause impacts the probability of an effect.

Counterfactual theory of causality posits that C causes E if, in the absence of C, E would not have occurred (or would have been less likely to occur).

The fundamental problem of causal inference: at most, only one of many *potential outcomes* can be observed for a given experimental or observational study, and thus, causal effects cannot be measured directly.

This is a problem for a scientific of statistical philosophy that posits that *observations* drive the generation of knowledge!

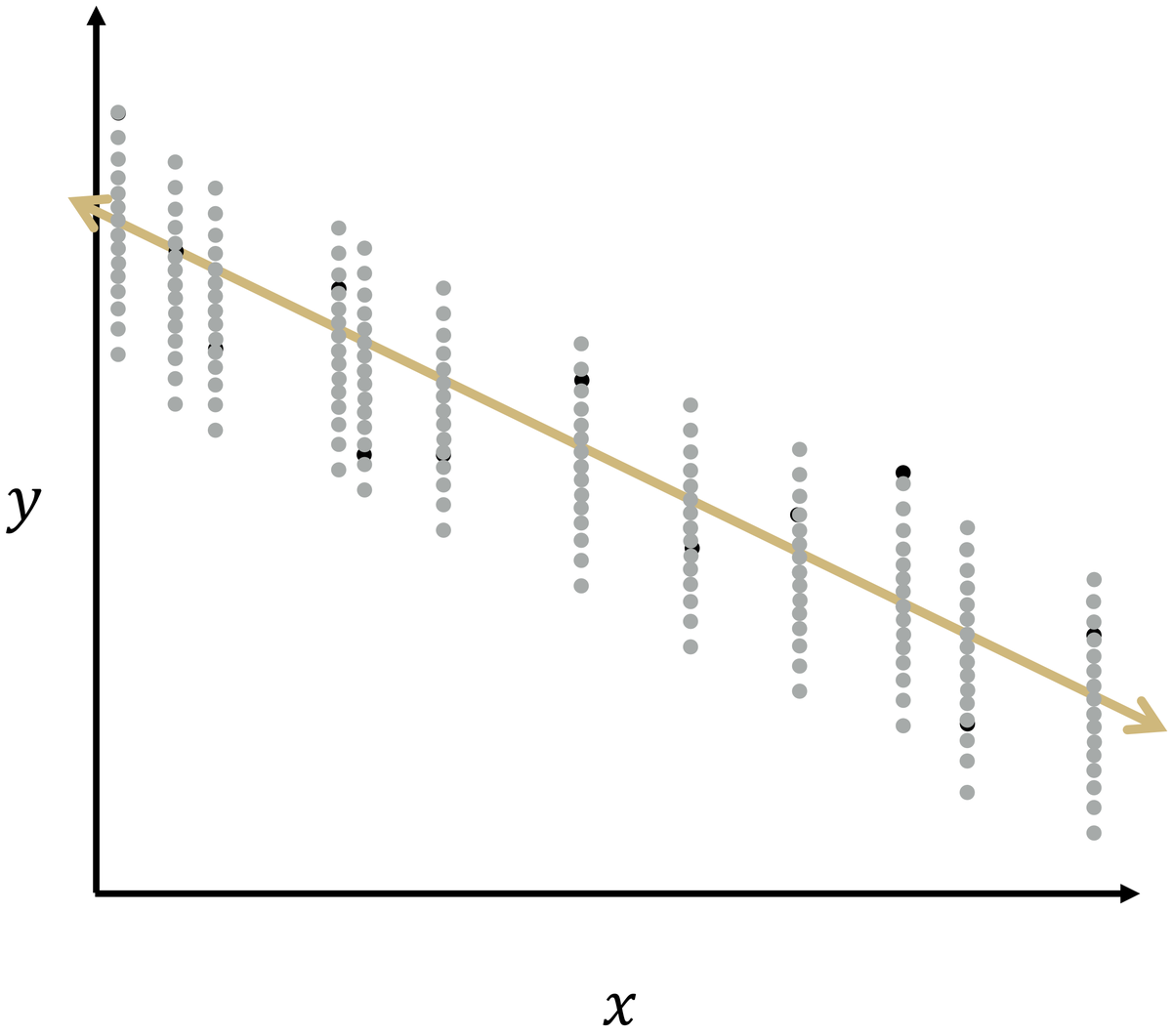
Ways of circumventing the fundamental problem of causal inference:

1. Find close substitutes

2. Conduct randomized experiments

3. Make additional assumptions about observational data

Photo by Jr Korpa on Unsplash


y
x

Recall that prediction is one of the main purposes of using regression models. Suppose for now that we have a model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ that fits the data well (we'll say more later about checking wither the model fits the data).

Given a new set of predictors, $\mathbf{x}^\star = (1, x_1^\star, \ldots, x_p^\star)^T$, the estimated average value of the response is:

Suppose we'd like to know the *average sales in a market with advertising budgets* $x_1^\star, x_2^\star, and x_3^\star$. The point estimate is:

But what about an interval estimate?