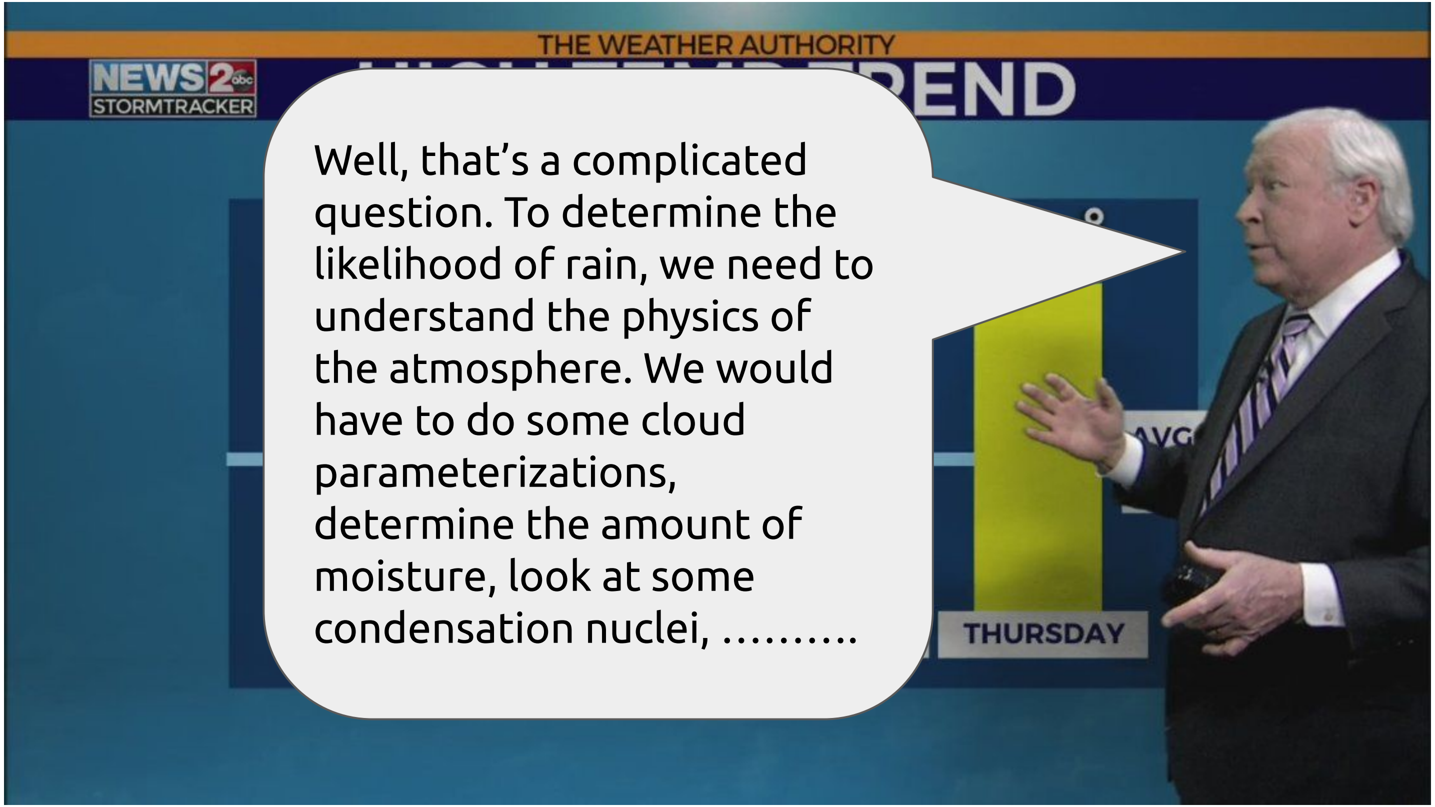
Introduction to Supervised Learning

Question - Is it going to rain today? Do I need my umbrella?

Question - Is it going to rain today? Do I need my umbrella?



Question - Is it going to rain today? Do I need my umbrella?

You look outside and it looks like this:

Question - Is it going to rain today? Do I need my umbrella?

You look outside and it looks like this:

Question - Is it going to rain today? Do I need my umbrella?

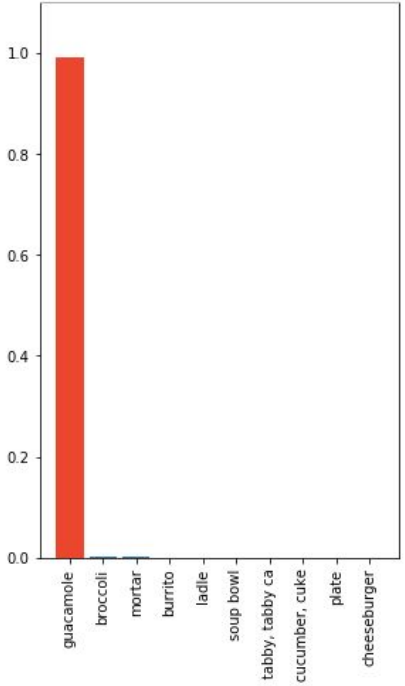
You look outside and it looks like this:

Question - Is it going to rain today? Do I need my umbrella?

You look outside and it looks like this:

Moral: Given a good set of predictor variables and enough “experience” (training data), we can often make good predictions.

Before we get started, a word of caution: machine learning is more than just dumping a lot of data into a magic black box to make predictions.



https://www.theverge.com/2017/11/2/16597276/google-ai-image-attacks-adversarial-turtle-rifle-3d-printed

Before we get started, a word of caution: machine learning is more than just dumping a lot of data into a magic black box to make predictions.

EXCLUSIVE

STAT+

IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show

By Casey Ross and Ike Swetlitz July 25, 2018

Reprints

https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

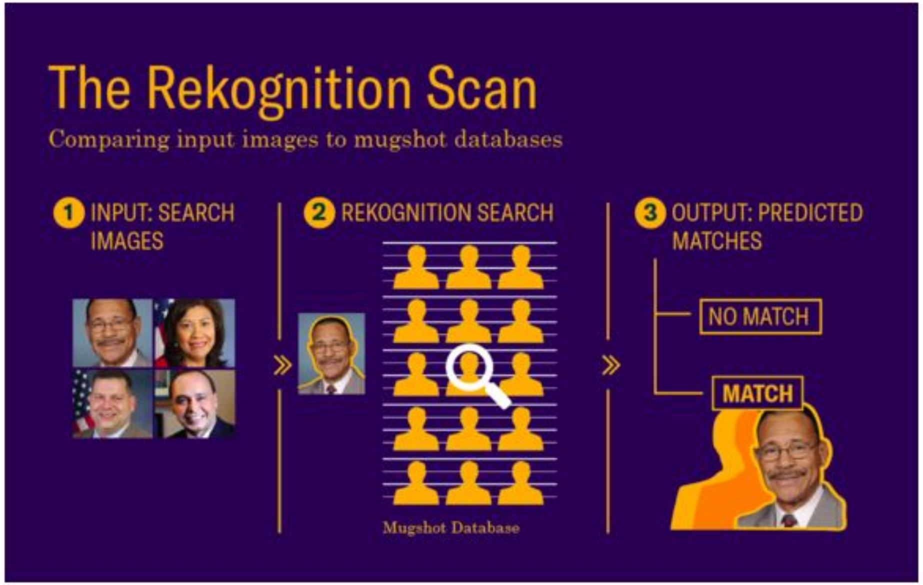
Before we get started, a word of caution: machine learning is more than just dumping a lot of data into a magic black box to make predictions.

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

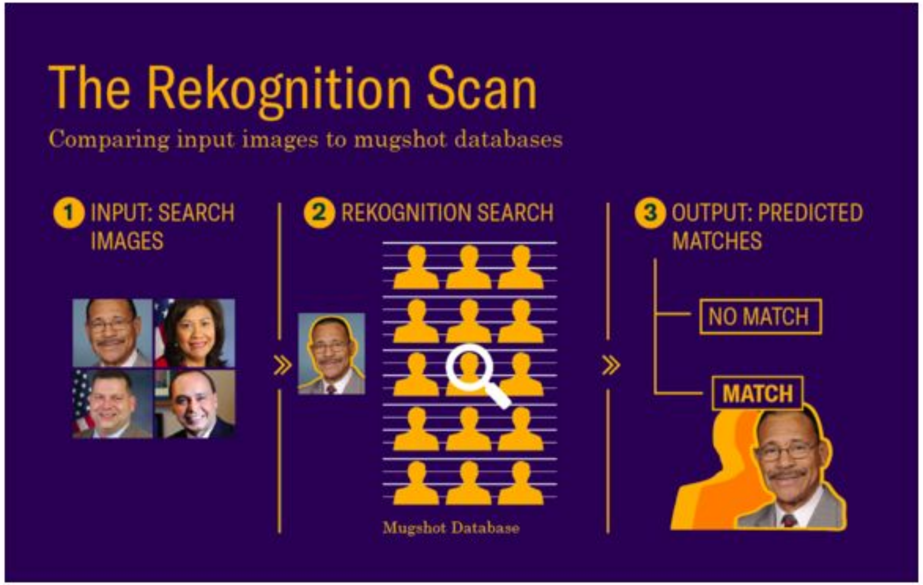
Before we get started, a word of caution: machine learning is more than just dumping a lot of data into a magic black box to make predictions.



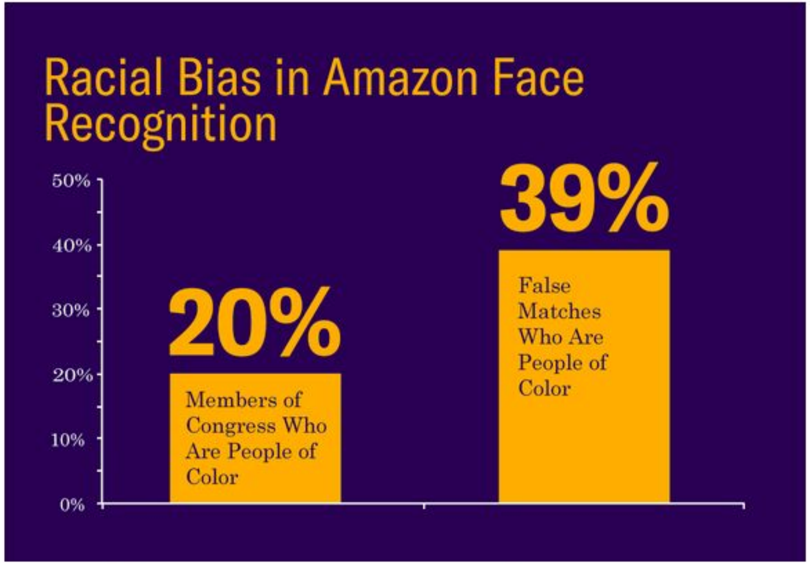
Rep. Sanford Bishop (D-Ga.) was falsely identified by Amazon Rekognition as someone who had been arrested for a crime.

https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

Before we get started, a word of caution: machine learning is more than just dumping a lot of data into a magic black box to make predictions.



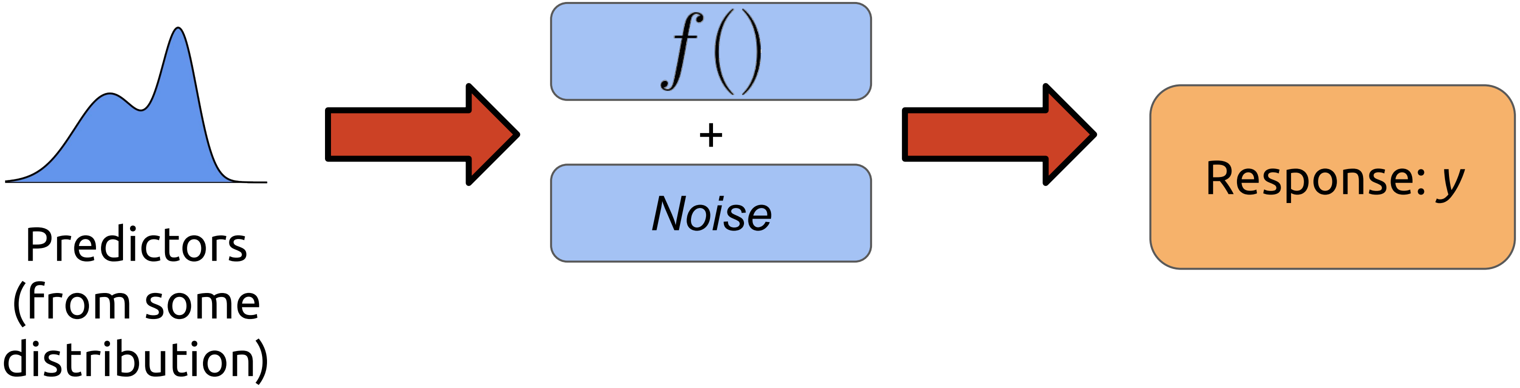
Rep. Sanford Bishop (D-Ga.) was falsely identified by Amazon Rekognition as someone who had been arrested for a crime.



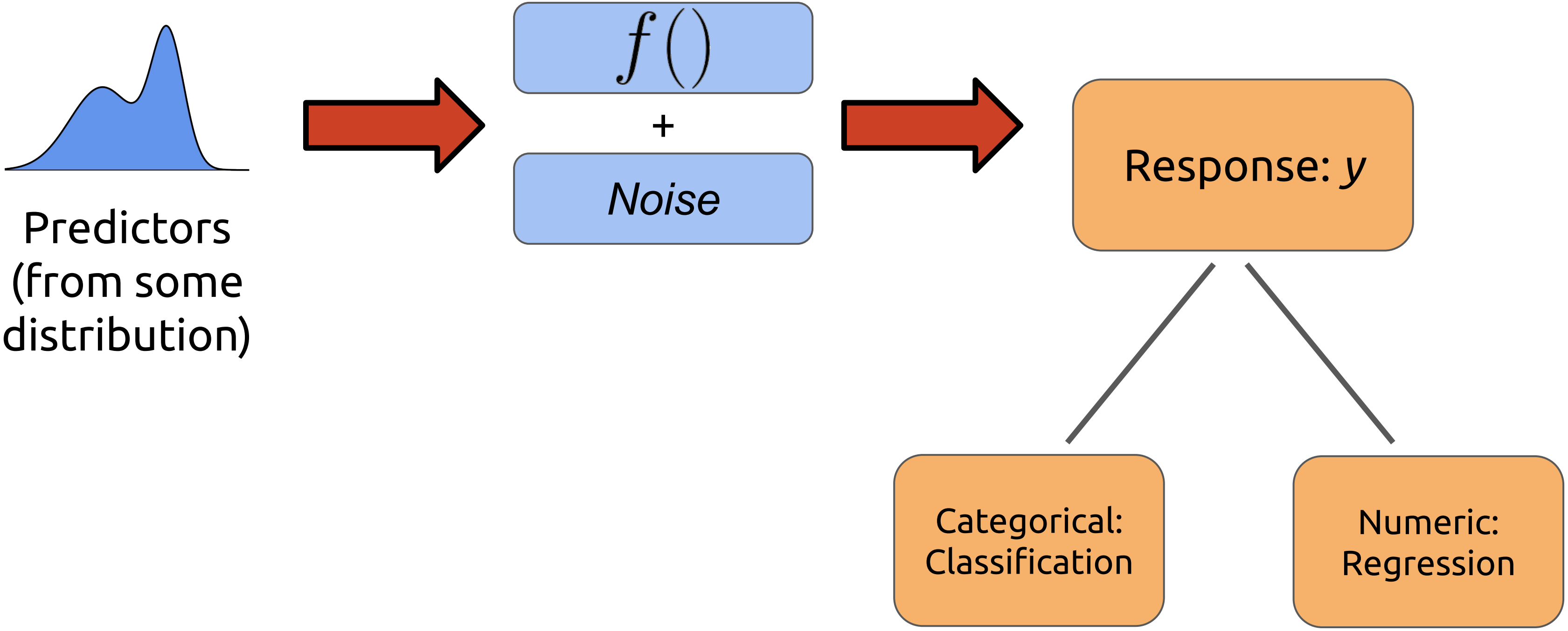
People of color were disproportionately falsely matched in our test.

https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

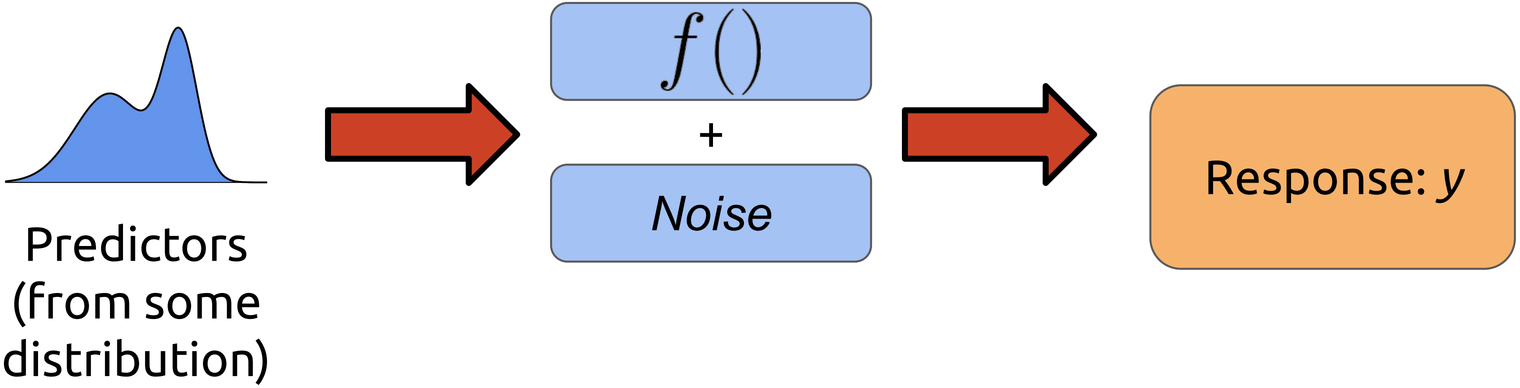
Supervised Learning - Setup



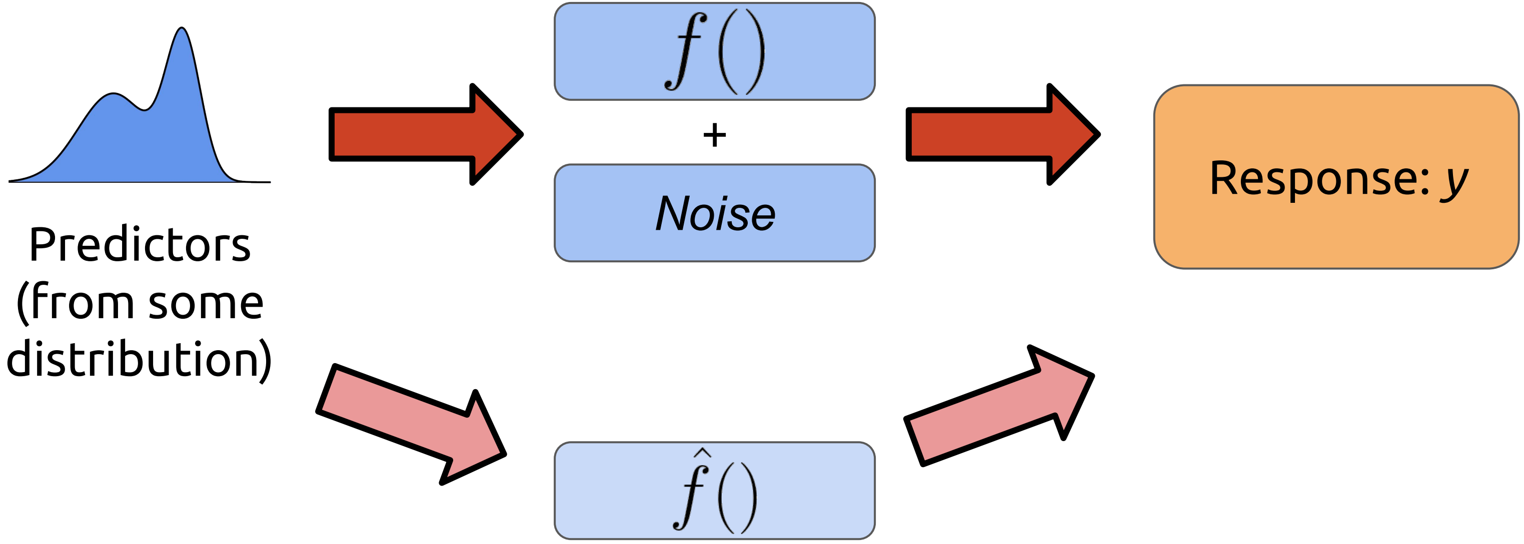
Supervised Learning - Setup



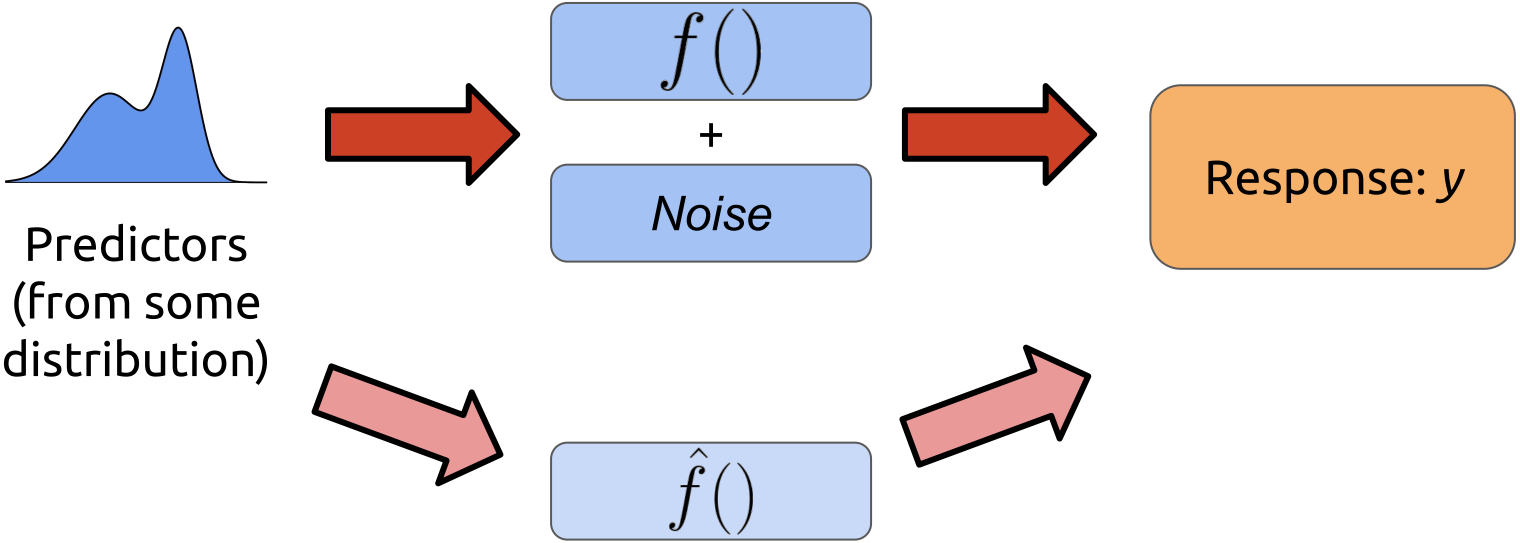
Supervised Learning - Setup



Supervised Learning - Goals

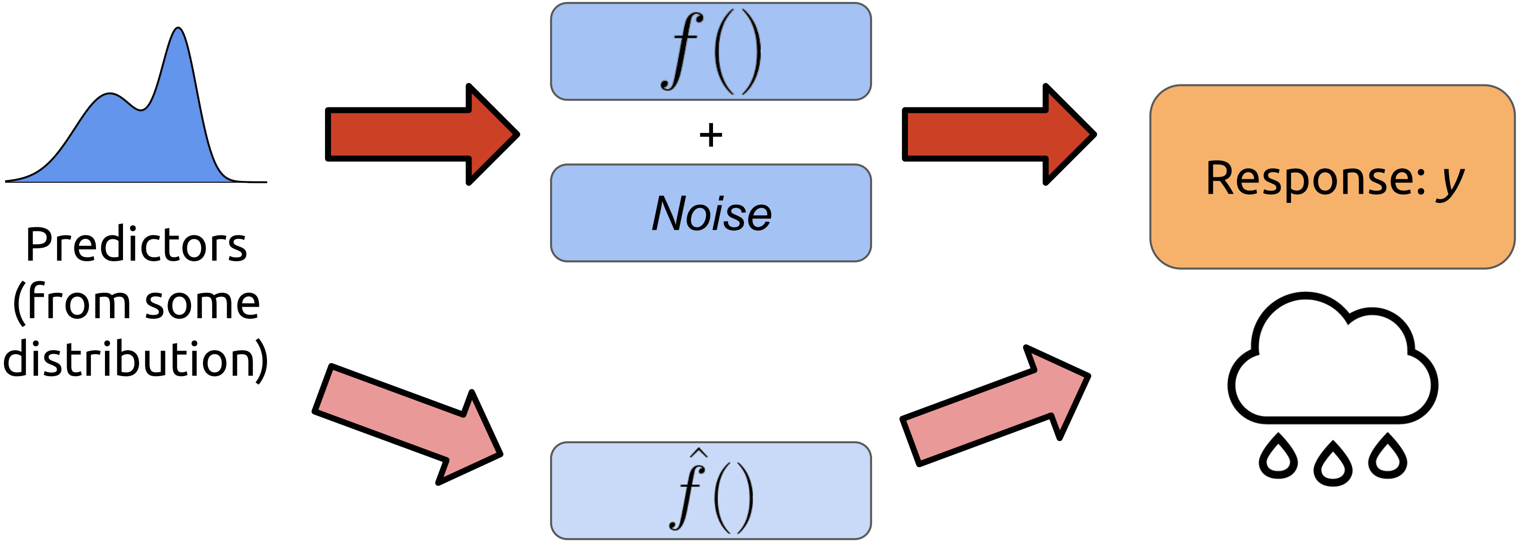


Supervised Learning - Goals

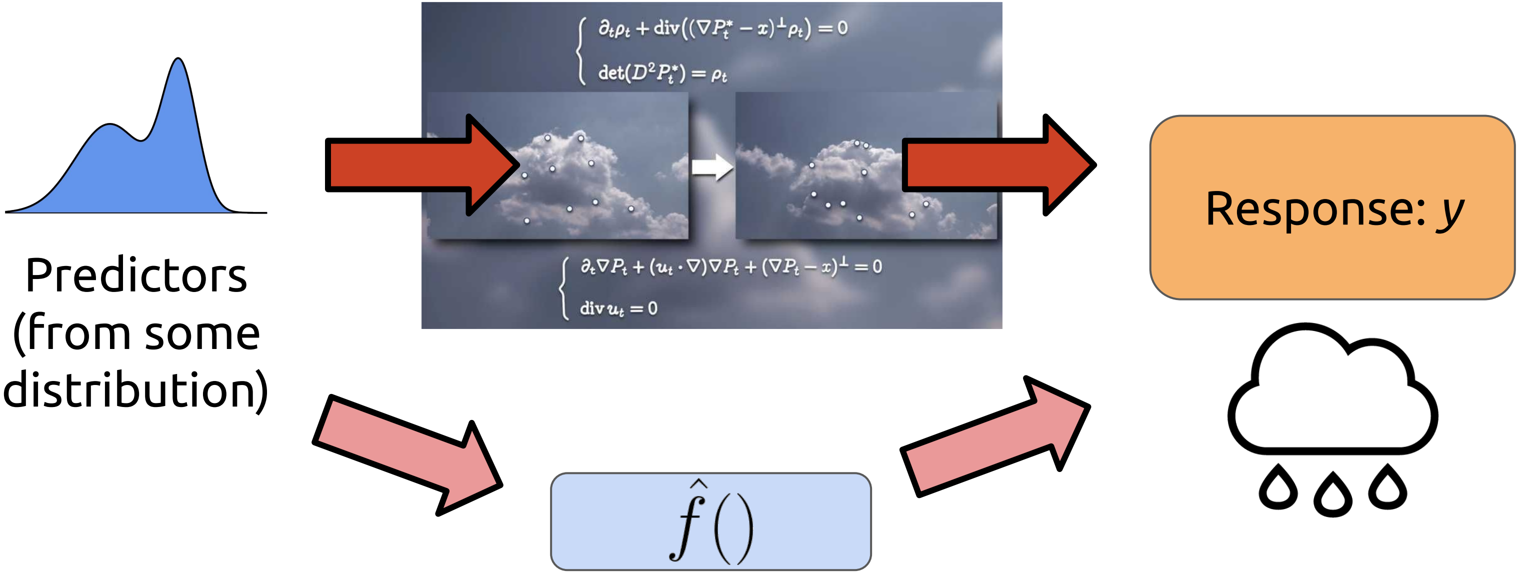


Goal: Choose a function so that the our predictions are close (on average) to the true values.

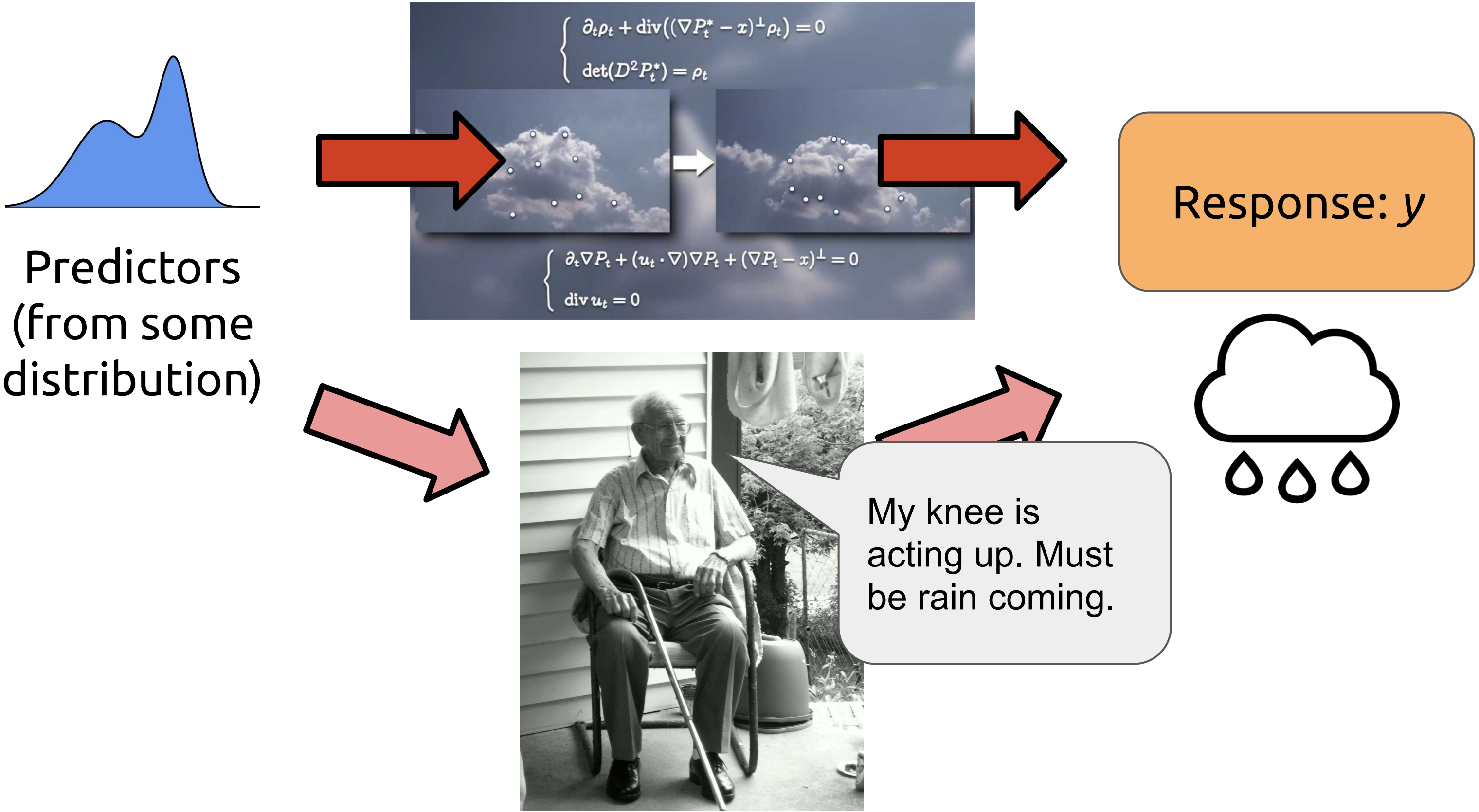
Supervised Learning - Grossly Oversimplified



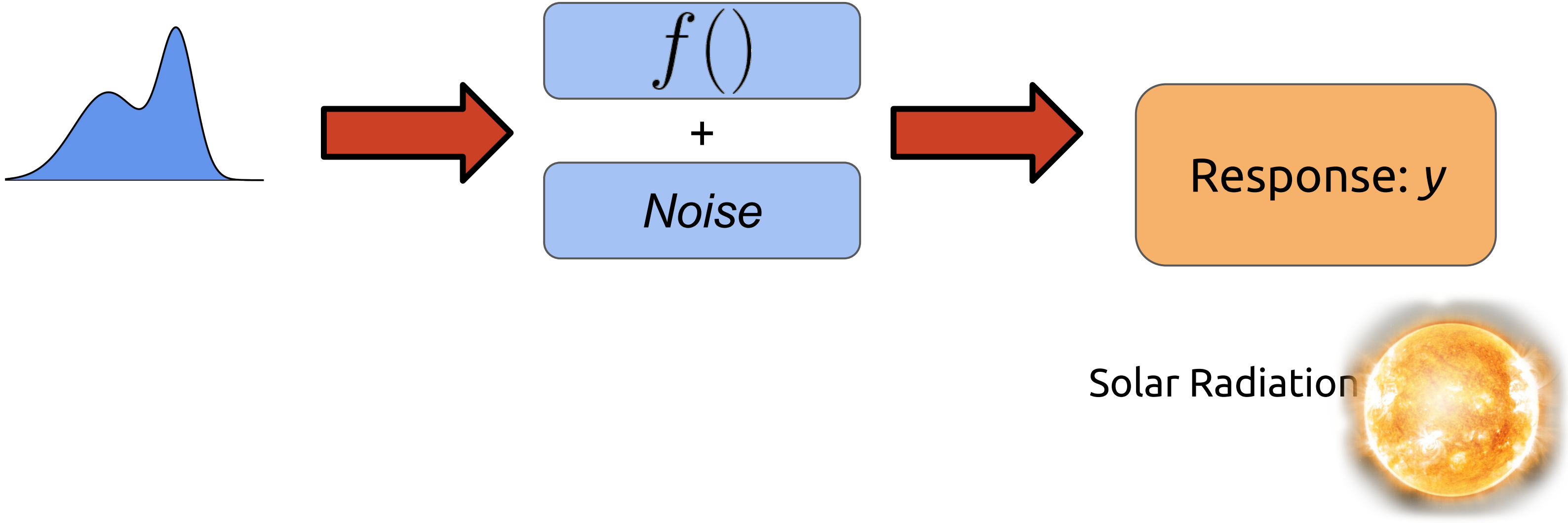
Supervised Learning - Grossly Oversimplified



Supervised Learning - Grossly Oversimplified

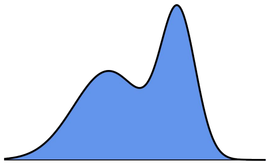


Example - Weather Prediction



https://www.ai.uga.edu/sites/default/files/inline-files/theses/sanders_william_s_201712_ms.pdf

Example - Weather Prediction



air temperature
precipitation rate
visibility
wind speed
wind direction
dew point
temperature
air pressure
relative humidity

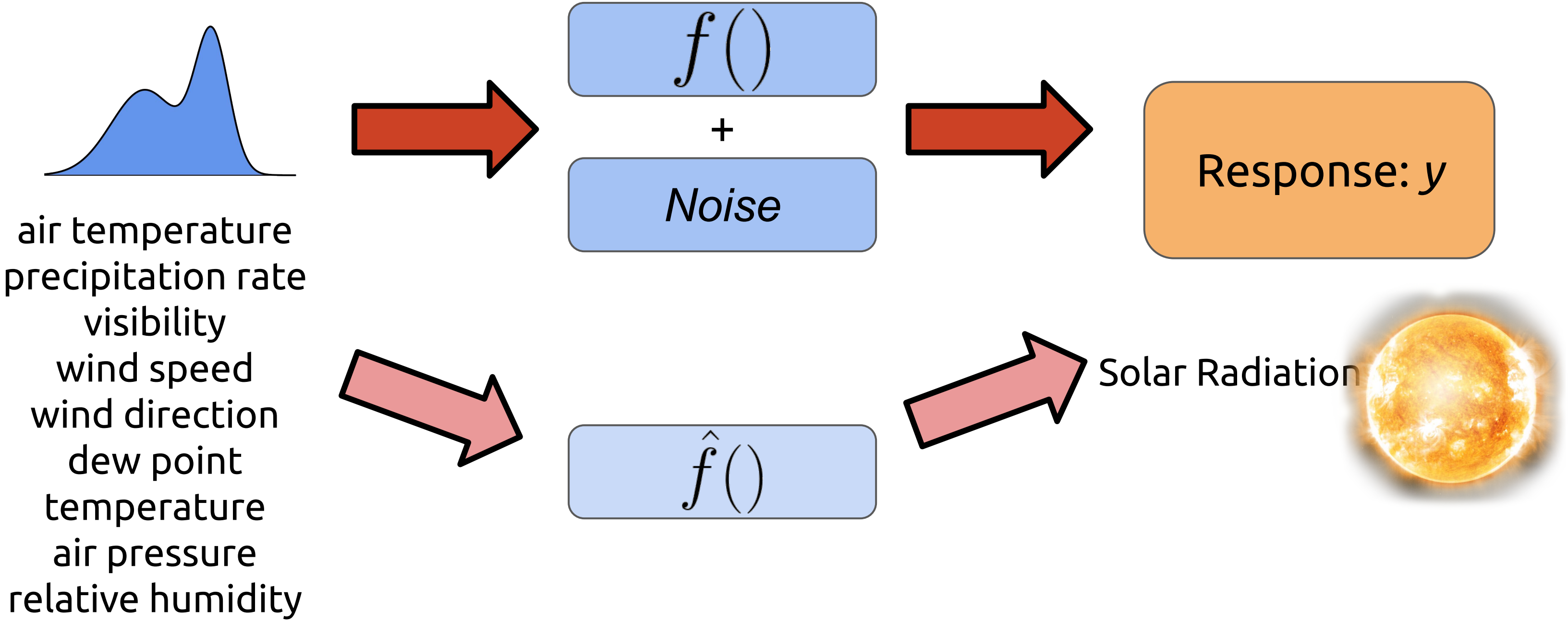
$f()$
+
Noise

Response: y

Solar Radiation

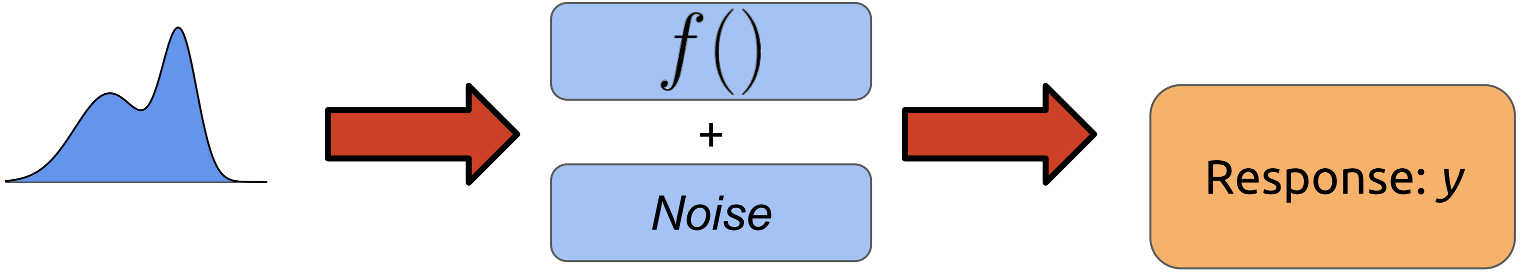
https://www.ai.uga.edu/sites/default/files/inline-files/theses/sanders_william_s_201712_ms.pdf

Example - Weather Prediction



https://www.ai.uga.edu/sites/default/files/inline-files/theses/sanders_william_s_201712_ms.pdf

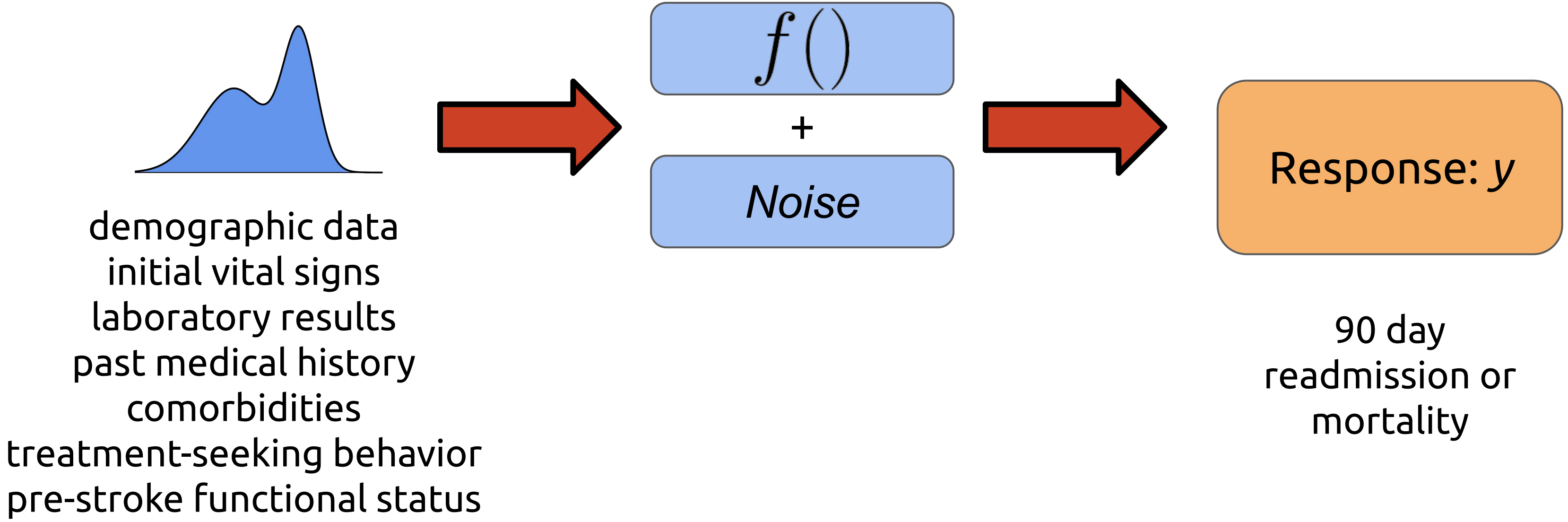
Example - Readmission or Death of Stroke Patients



90 day
readmission or
mortality

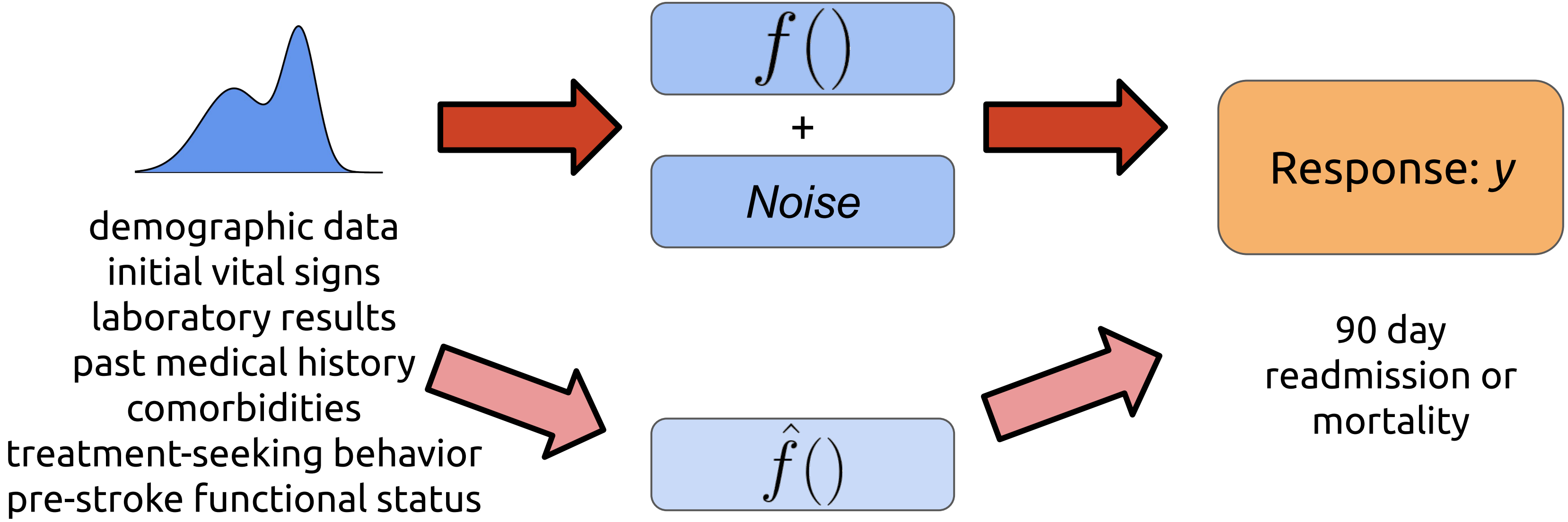
https://www.mdpi.com/2076-3417/10/18/6337/pdf

Example - Readmission or Death of Stroke Patients



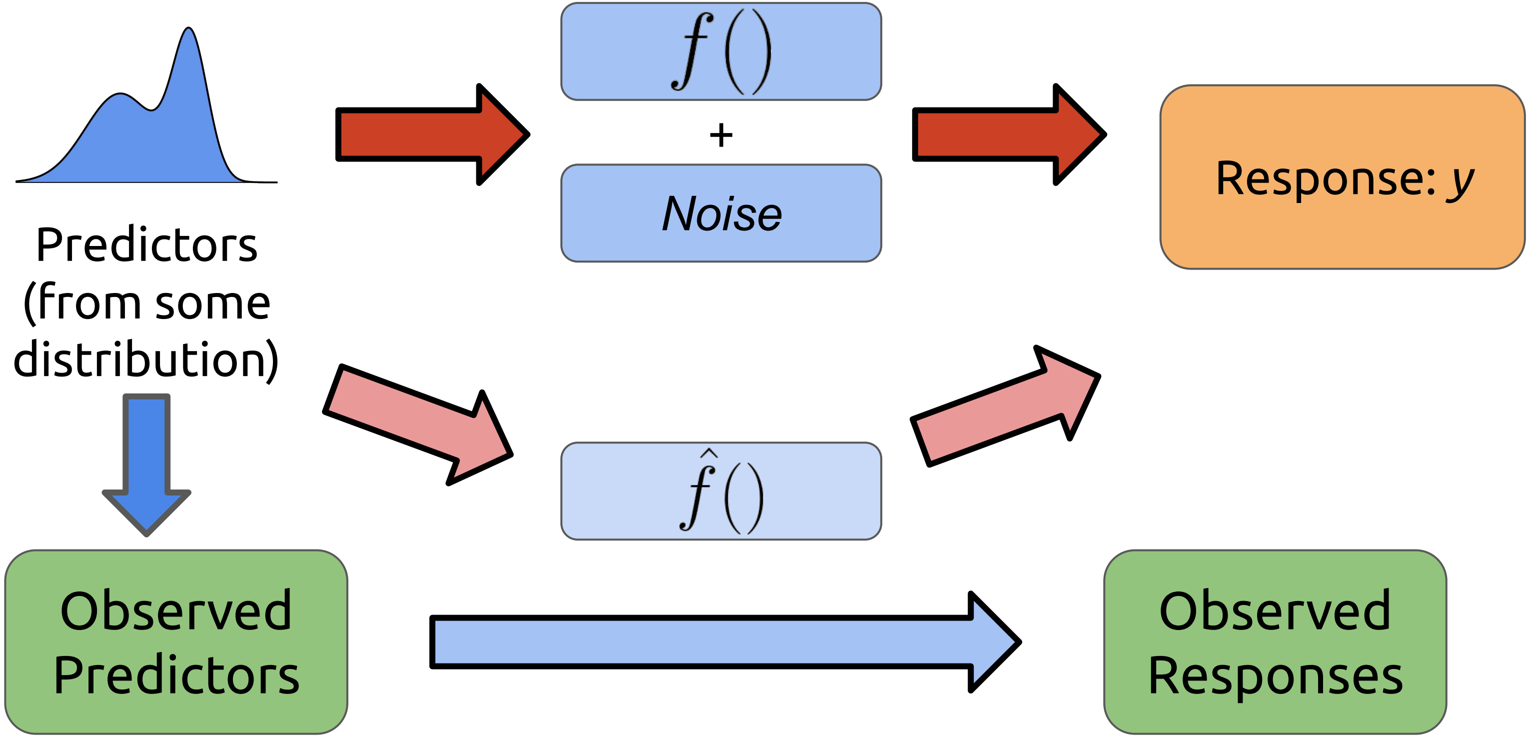
https://www.mdpi.com/2076-3417/10/18/6337/pdf

Example - Readmission or Death of Stroke Patients



https://www.mdpi.com/2076-3417/10/18/6337/pdf

Supervised Learning - How



Supervised Learning - Goals

To measure how “good” our model is, we need some way to measure “error” (eg. mean squared error).

Our goal is to minimize the expected loss over *new* data.

Important: We are not trying to minimize loss over the observed data (which is often very easy to do), but to minimize the *generalization error* - the performance on unseen data.

Measuring Generalization Data - How

If we only care about how well our model performs on unseen data, how do we measure that?

Measuring Generalization Data - How

If we only care about how well our model performs on unseen data, how do we measure that?

We can’t - it’s unseen!

Measuring Generalization Data - How

If we only care about how well our model performs on unseen data, how do we measure that?

We can’t - it’s unseen!

But, we can *estimate* it.

Measuring Generalization Error - How

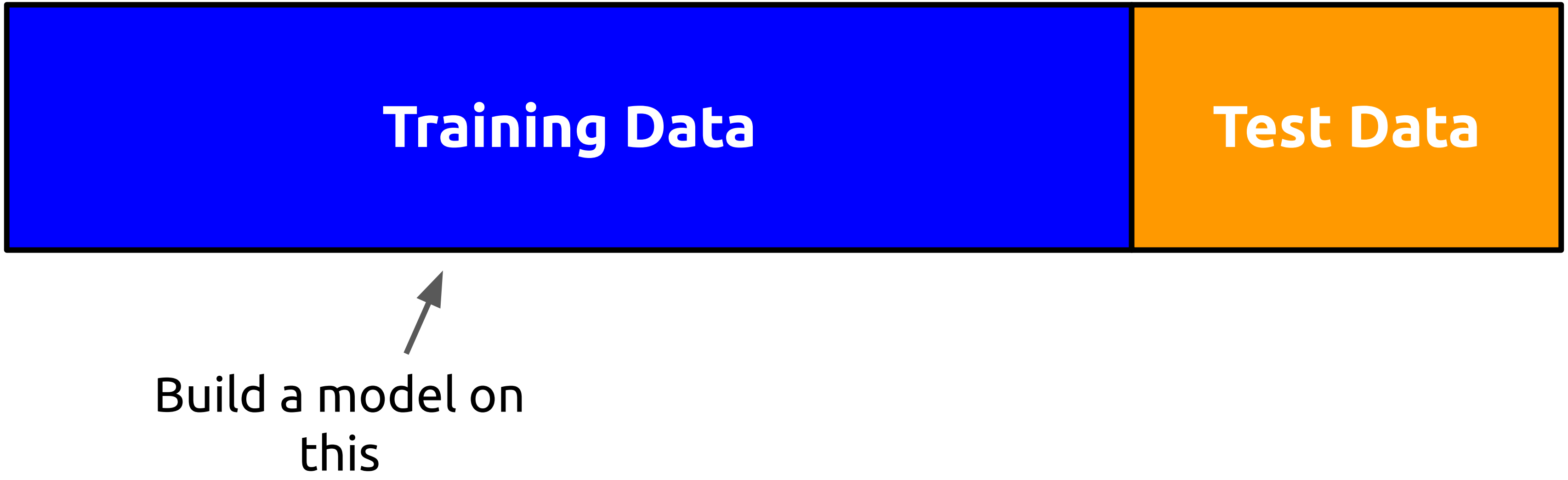
The most simple way to estimate generalization error is through employing a train/test split.

Measuring Generalization Error - How

The most simple way to estimate generalization error is through employing a train/test split.

Measuring Generalization Error - How

The most simple way to estimate generalization error is through employing a train/test split.



Measuring Generalization Error - How

The most simple way to estimate generalization error is through employing a train/test split.

Build a model on this

Estimate generalization error by seeing how it performs on this.