1. Machine Learning Overview

This report is meant to serve as an intro to a few of the possibilities that machine learning opens up at Carimus. **CariML** has the potential to be our all encompassing solution to clients machine learning needs. Similarly to how we offer *design* and *development* services, *machine learning* and *data science* could be another set of services on that list.

This means that while **CariML** could be a specific product or service, it doesn't even need to be limited to that. We could implement machine learning in an unlimited number of ways, under the CariML umbrella. In other words, it could be marketed to users as the **CariML** service, but that may slightly change under the hood from project to project.

With that said, lets get started with a quick overview, followed by specific use cases that could be applicable already.

1.1 What is Machine Learning

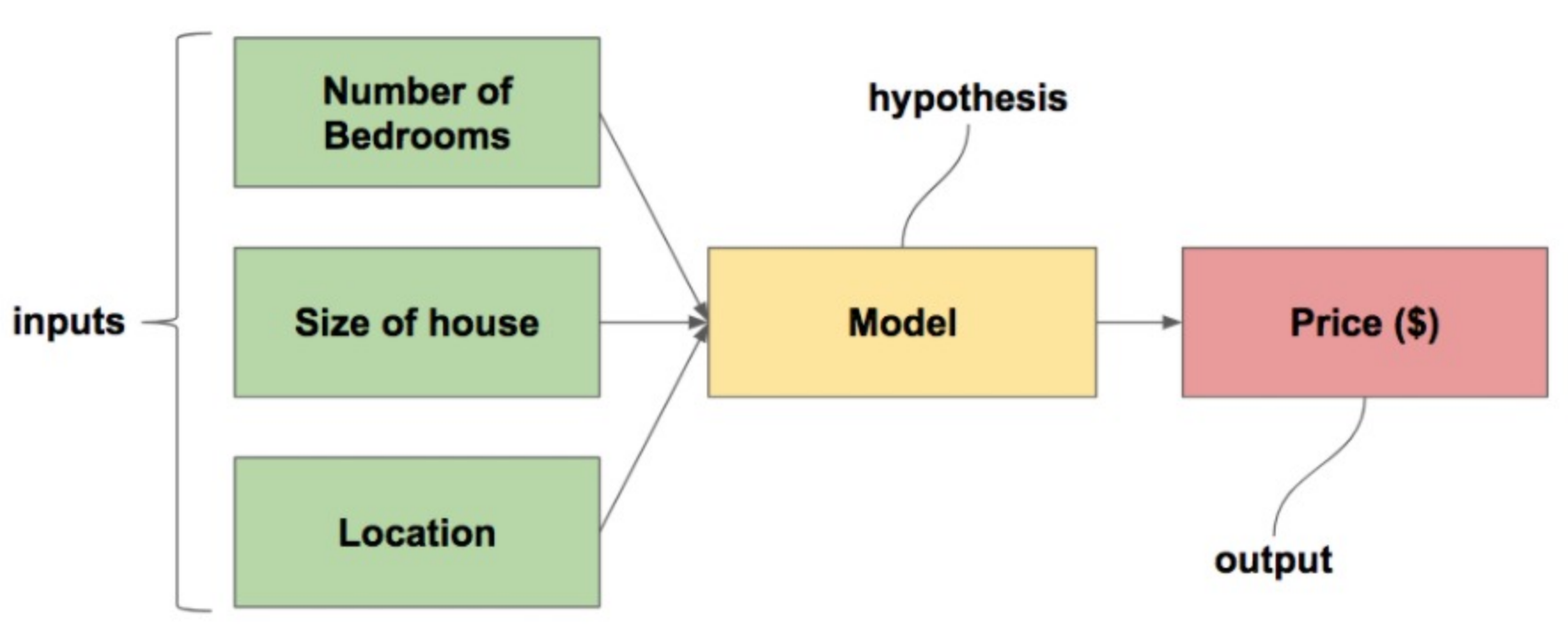
In the context of this report, machine learning can be defined as a way in which we create **models** that can map **inputs** to **outputs**. Mathematically, just think of a **function**. For instance, say we are trying to predict how much we can sell a house for; we have 3 different inputs:

> - **Number of Bedrooms**
> - **Size of House**
> - **Location**

And our output would be:

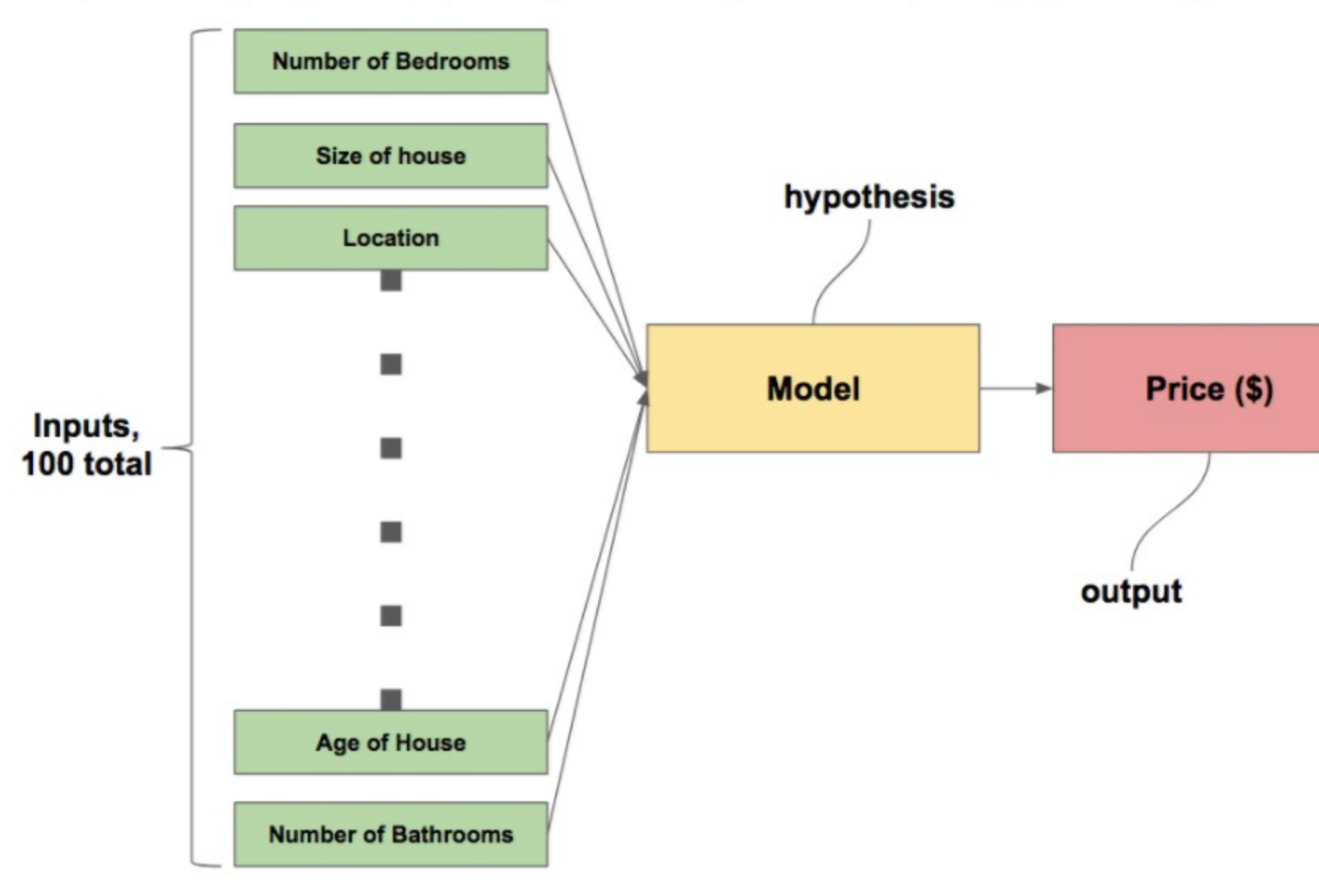
> - **Price to sell house for ($)**

This can easily be seen in the diagram below.



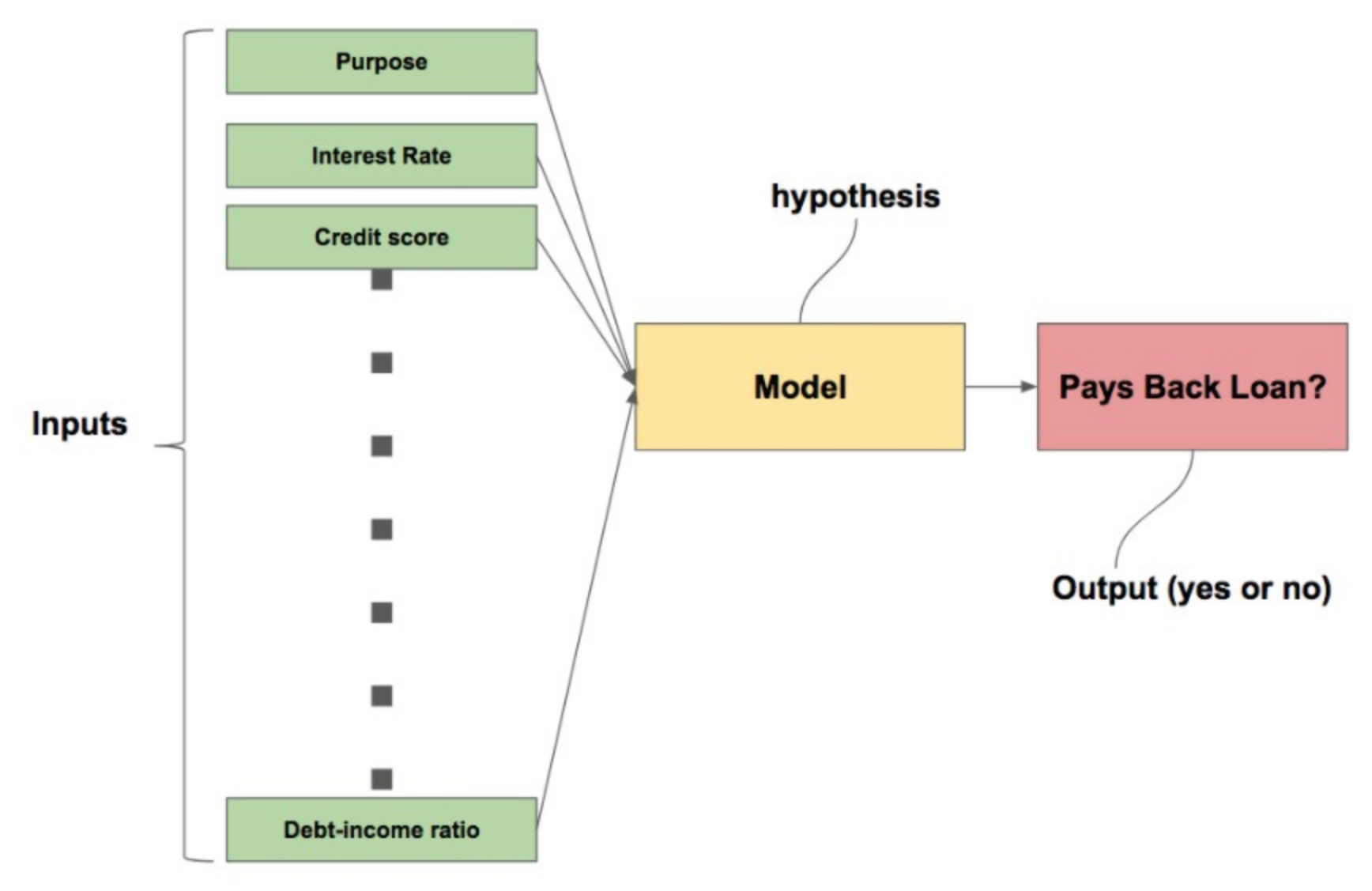
What we are saying is that given a certain **location**, **size**, and **number of bedrooms**, the house will be worth a **certain amount of money**. Our goal is to **determine the relationship between those inputs and how much money the house is worth**. This relationship is generally called a **model**.

Clearly this is a very simple example, but imagine if suddenly instead of only having 3 **predictors**, you had 100.

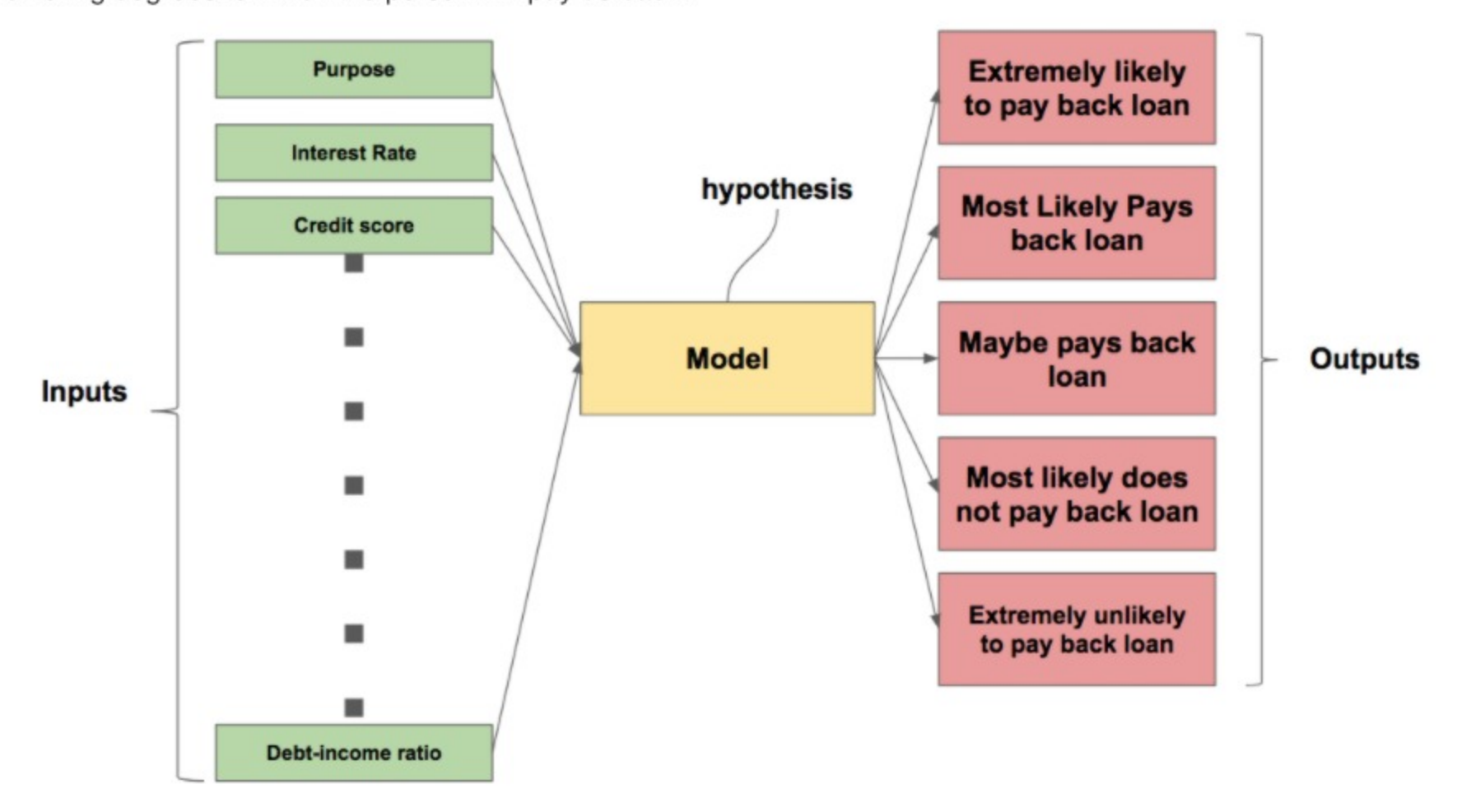


Now the value of being able to determine the relationship between 100 predictors and the corresponding price is far more concrete. We can find **KPI's** that a human may have never have even considered.

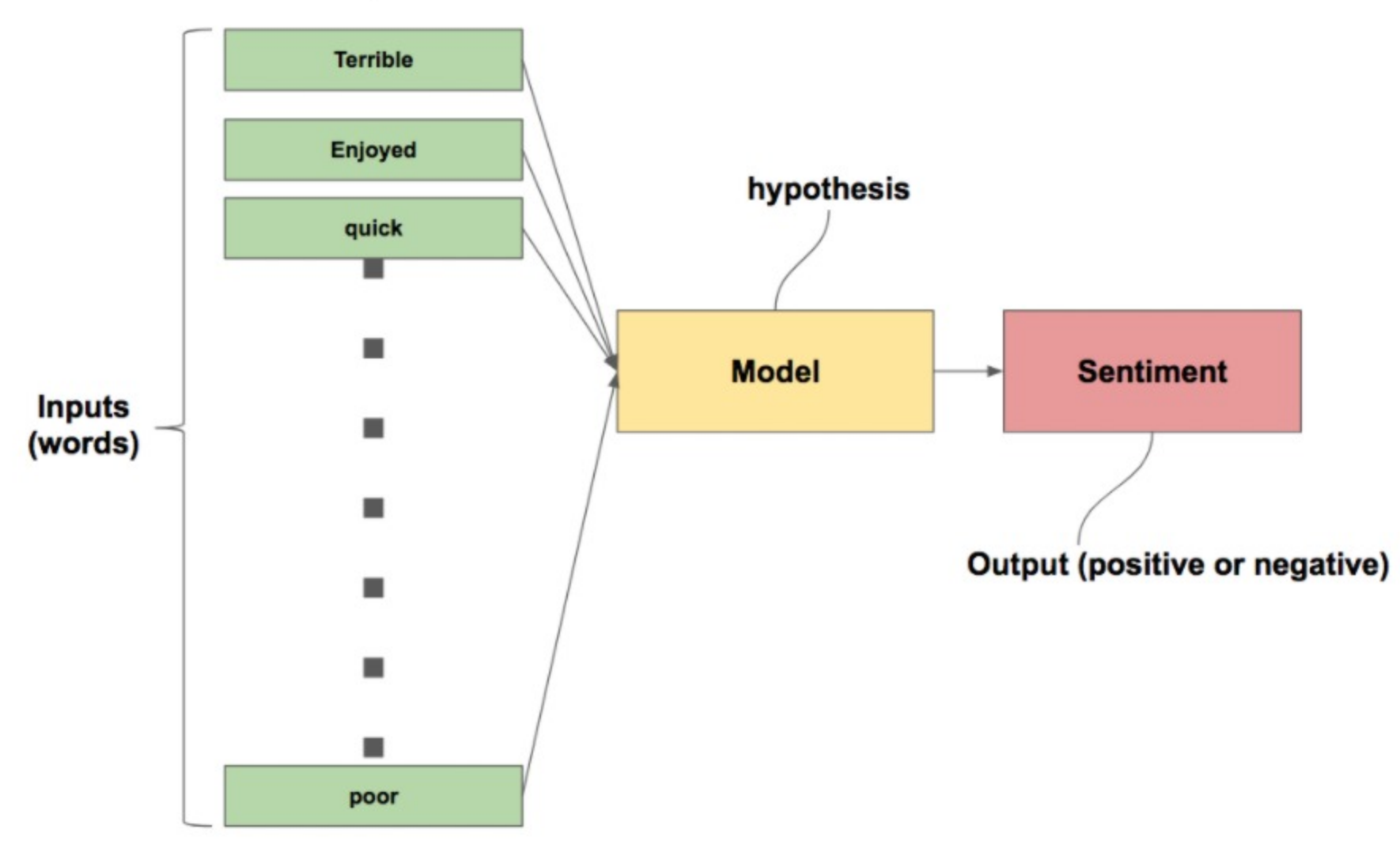
We can also create models that map **inputs** to **classes**. As an example, say we are an investor, and we are trying to determine if we should lend someone money. We could create a model that can predict whether the person will pay us back or not:



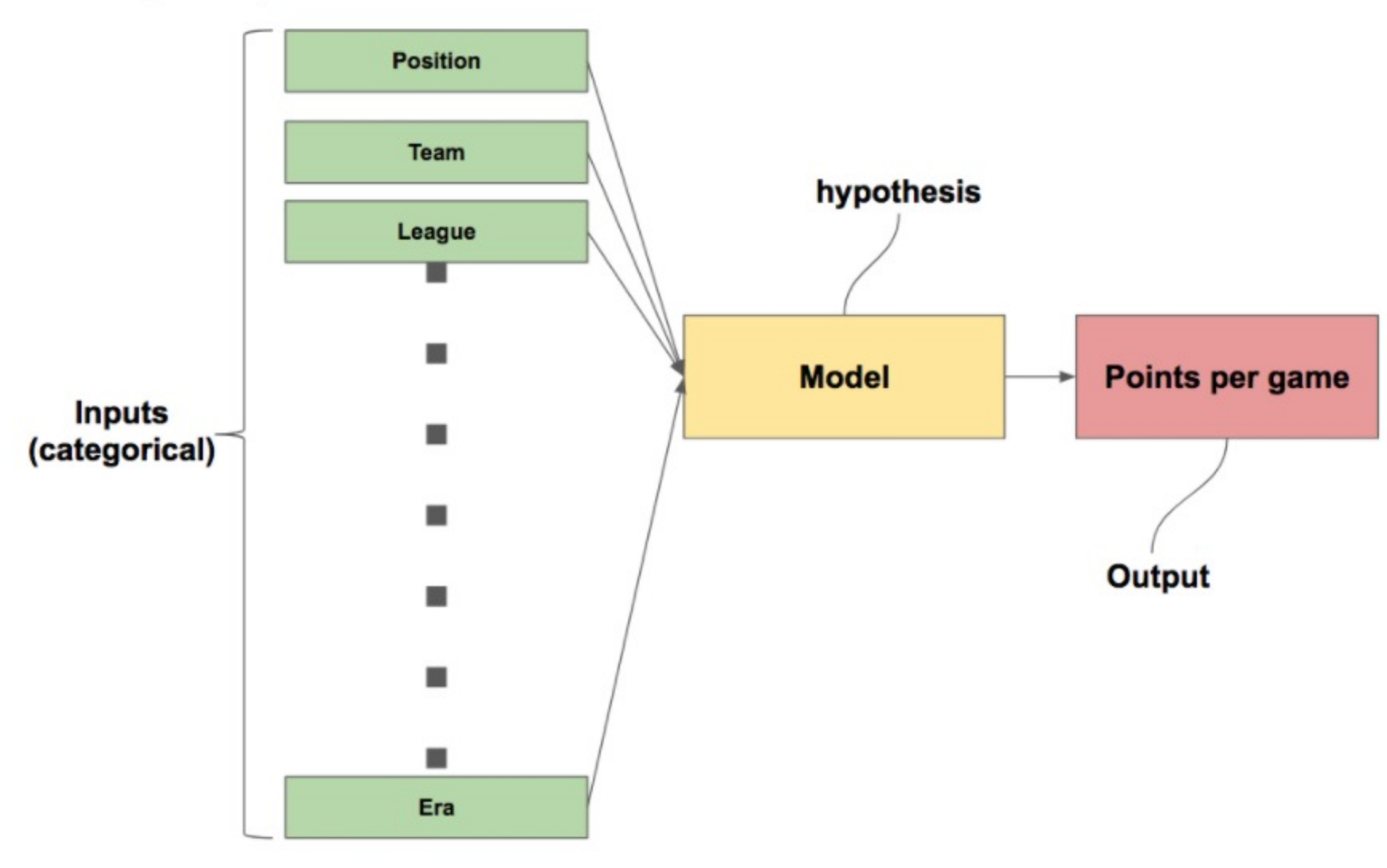
Note, this prediction does not need to be limited to two outputs. We could also have our output be composed of differing degrees to which the person will pay us back:



And we should mention, our inputs don't have to be limited to numerical data, words can be utilized as well:



As can categorical inputs:



You can begin to imagine the ways in which this can be applied at Carimus:

PPS:

> Predict whether a **domain** is valuable or not valuable, *or*, predict its value on a scale of 1-10 with 1 being **least valuable** and 10 being **most valuable**

Nextlot:

> Based on user actions during auctions (number of bids placed, number of auctions attended, time on site, and so on), try and classify users who are on the cusp of bidding, but have not yet. Use that insight in order to send an email to those bidders, or have the client offer a discount or some sort of incentive to bid, driving up sales. Could also optimize pricing, promotion, etc.

Nextlot:

> Predict an opening bid based on similar, past lots, prebidding, viewers, etc.

Goodbookey:

> Utilize **A/N** testing in order to compare a variety of webpages/UI configurations/etc, at the same time

Goodbookey:

> Optimal push notifications. Based on user location, previous bets, time zone, nearby sports teams, etc, determine the best time to send them a custom push notification to pump up retention and engagement

Tom - lets add more here

1.2 The CariML Strategy

With all of these different use cases in mind, we feel that the goal should be to slowly integrate Machine Learning services as something that Carimus offers. It could look similar in form to the development strategy of Carimus at this point: There is a suite of tools that we prefer to stick by (AWS, Google Analytics, React, etc), but in certain cases that will change to fit the project/client needs). We could do something similar by mainly using Google's machine learning suite, but generally performing data wrangling and preprocessing on a case by case basis.

It can be marketed as one service, **CariML**, but under the hood that implementation will be malleable.

1.3 Use Cases

Lets take a quick look at a few use cases. These have all been heavily parsed down, and all of the heavy lifting has been done prior to this report. Each example is also relatively basic in order to keep things as clear as possible.

1.3.1 Use Case: Determine KPI's

Say we are working with a client, and they are looking for development help. They plan on improving both their mobile app experiene as well as their website. However, they are not sure which one should tackled first. They are able to give us access to the data that they have on both applications. We could take a guess and go with our gut, or we could utilize the data in order to make the most informed decision possible, given the client the best value for their dollar.

The data is a csv file from the company. It has Customer info, such as **Email** and **Address** Then it also has numerical value columns:

- **Avg. Session Length**: Average amount of time user spends in store
- **Time on App**: Average time spent on App in minutes
- **Time on Website**: Average time spent on Website in minutes
- **Length of Membership**: How many years the customer has been a member

For reference the data looks like:

	Email	Address	Avatar	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
0	mstephenson@fernandez.com	835 Frank Tunnel\nWrightmouth, MI 82180-9605	Violet	34.497268	12.655651	39.577668	4.082621	587.951054
1	hduke@hotmail.com	4547 Archer Common\nDiazchester, CA 06566-8576	DarkGreen	31.926272	11.109461	37.268959	2.664034	392.204933
2	pallen@yahoo.com	24645 Valerie Unions Suite 582\nCobbborough, D...	Bisque	33.000915	11.330278	37.110597	4.104543	487.547505
3	riverarebecca@gmail.com	1414 David Throughway\nPort Jason, OH 22070-1220	SaddleBrown	34.305557	13.717514	36.721283	3.120179	581.852344
4	mstephens@davidson-herman.com	14023 Rodriguez Passage\nPort Jacobville, PR 3...	MediumAquaMarine	33.330673	12.795189	37.536653	4.446308	599.406092

By utilizing machine learning and data science, we can take in the above data and deliver the following to the client:

	Coeffecient
Avg. Session Length	25.981550
Time on App	38.590159
Time on Website	0.190405
Length of Membership	61.279097

Which in english can be interpreted as:

> - Holding all other features fixed, a 1 unit increase in **Avg. Session Length** is associated with an **increase of 25.98 total dollars spent**.
> - Holding all other features fixed, a 1 unit increase in **Time on App** is associated with an **increase of 38.59 total dollars spent**.
> - Holding all other features fixed, a 1 unit increase in **Time on Website** is associated with an **increase of 0.19 total dollars spent**.
> - Holding all other features fixed, a 1 unit increase in **Length of Membership** is associated with an **increase of 61.27 total dollars spent**.

We have been able to quantitatively define how the each variable effects the bottom line: how much money the company is bringing in. At this point, the information could then be used in one of two ways: Decide to develop the Website to catch up to the performance of the mobile app, or develop the app more since that is what is working better.

Key Takeaway

This type of analysis is a service that we could provide in two different types of scenarios:

1. A one off statistical/machine learning analysis for a client, done quickly in order to help make a pressing decision. Just having the ability to say "we at Carimus can do this" is a big plus. It only strengthens us as an organization.
2. A service we offer that features a dash board to display the results to the client, and a machine learning algorithm on the back end that takes in their data, and outputs the KPI's to the UI. A note: each client would most likely have data in different format that is full of all sorts of complexities (missing values, improper data types, the list goes on). We would most likely need to build a small service on a case by case basis that performs data engineering on the client data, before feeding it into the Machine learning algorithm.

1.3.2 Use Case: Make Predictions

Now lets switch gears and imagine a client came to us, or we had a project internally (hint hint), where their was a goal of having a user take a particular action. For instance, goodbookey wants a user to make a bet that was recommended to them. Or nextlot wants a user to register to bid. What if we could predict which users are most likely to convert, and specifically target them, possibly with an incentive or push notification.

For example lets consider the following data set that is concerned with users clicking on advertisements:

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	Clicked on Ad
0	68.95	35	61833.90	256.09	Cloned 5thgeneration orchestration	Wrightburgh	0	Tunisia	2016-03-27 00:53:11	0
1	80.23	31	68441.85	193.77	Monitored national standardization	West Jodi	1	Nauru	2016-04-04 01:39:02	0
2	69.47	26	59785.94	236.50	Organic bottom-line service-desk	Davidton	0	San Marino	2016-03-13 20:35:42	0
3	74.15	29	54806.18	245.89	Triple-buffered reciprocal time-frame	West Terrifurt	1	Italy	2016-01-10 02:31:19	0
4	68.37	35	73889.99	225.58	Robust logistical utilization	South Manuel	0	Iceland	2016-06-03 03:36:18	0

So we have the following variables to work with: This data set contains the following features:

- **Daily Time Spent on Site**: consumer time on site in minutes
- **Age**: customer age in years
- **Area Income**: Avg. Income of geographical area of consumer
- **Daily Internet Usage**: Avg. minutes a day consumer is on the internet
- **Ad Topic Line**: Headline of the advertisement
- **City**: City of consumer
- **Male**: Whether or not consumer was male
- **Country**: Country of consumer
- **Timestamp**: Time at which consumer clicked on Ad or closed window
- **Clicked on Ad**: 0 or 1 indicated clicking on Ad

Our goal here is to use this data in order to predict which users, based on their associated data, are most likely to click on the advertisement. However, the action could just as easily be **making a bet**, or **registering to bid**, or any other thing we could think of.

By making use of machine learning we can offer the client a tremendous value in knowing which users are the most important to target, since they are the most likely to convert. What's awesome is that different machine learning algorithms can provide us with different benefits, depending on what the client is after.

Option 1

For instance, if the client wishes to obtain a good prediction accuracy as well as a gain insight into the what the **KPI's** are, we can give them that by utilizing a specific machine learning algorithm. The results are a **91% prediction accuracy**, and the following table:

	Coeffecient
Daily Time Spent on Site	-0.052400
Age	0.253437
Area Income	-0.000017
Daily Internet Usage	-0.026505
Male	0.021753

Which again in english can be interpreted as:

> - Holding all other features fixed, a 1 unit increase in **Daily time spent on site** is associated with a net decrease in **probability that user will click on advertisement**.
> - Holding all other features fixed, a 1 unit increase in **Age** is associated with a net increase in the probability that **user will click on advertisement**.
> - Holding all other features fixed, a 1 unit increase in **Area Income** is associated with a net decrease in the probability that **user will click on advertisement**.
> - Holding all other features fixed, a 1 unit increase in **Daily internet usage** is associated with a net decrease in the probability that **user will click on advertisement**.
> - Holding all other features fixed, a 1 unit increase in **Male** is associated with a net increase in the probability that a **user will click on advertisement**.

With this information the client has the ability to target certain users in the meantime (due to good prediction accuracy) and also has more knowledge about their key users that can be utilized to make better business decisions.

Option 2

Another option would be to use a machine learning algorithm that is less interpretable, but offers a higher prediction accuracy, in this case **96%**. This decision would be made on a case by case basis, but clearly can offer great value.

Key Takeaway

Here we **again** have the ability to take two routes/offer two things:

1. A one off statistical/machine learning analysis for a client, done quickly in order to help make a pressing decision.
2. A service we offer that features a dash board to display the results to the client, and a machine learning algorithm(s) on the back end that takes in their data, and outputs the KPI's to the UI.