

# Machine Learning From Scratch

Collin Prather

April 25th, 2018

## **Make sure that everyone launches binder or colab.google first!!**

Also, give them a walkthrough of the whole presentation and tell them what we'll be building and what we'll be predicting!

\* we'll be talking about how to apply machine learning to problems!

## Machine Learning Overview

### Steps in the Machine Learning Process

Step 0: Identify The Problem

Step 1: Get the Data

Step 2: Data Exploration

Step 3: Data Preparation

Step 4: Model Selection

Step 5: Cross-validation/Hyper-parameter tuning

### Building a Support Vector Machine from Scratch

### Exploring Scikit-Learn and applying to GR Crash dataset

# Machine Learning From Scratch

## └ Machine Learning Overview

### Machine Learning Overview

#### Steps in the Machine Learning Process

- Step 0: Identify The Problem
- Step 1: Get the Data
- Step 2: Data Exploration
- Step 3: Data Preparation
- Step 4: Model Selection
- Step 5: Cross-validation/Hyper-parameter tuning

#### Building a Support Vector Machine from Scratch

Exploring Scikit-Learn and applying to GR Crash dataset

Pull a graph of google search trends indicating how terms like "Data Science" and "Machine Learning" have blown up.

Try to form talk around hitting on the theoretical mathematical side of ML as well as the difficulties/complexities faced in Applied ML

data + algorithms = predicting the future (it's really a lot more than this – understanding context and how to frame the question (usually) from a business perspective is huge)

classification v. regression

supervised/unsupervised/reinforcement learning

when talking on reinforcement learning, mention and recommend

AlphaGo documentary (it's on netflix!)

considerations/complexities in building ML models

# What is Machine Learning?



# Machine Learning

Arthur Samuel:

Machine learning is “Field of study that gives computers the ability to learn without being explicitly programmed”.

# Machine Learning From Scratch

## └ Machine Learning Overview

### └ What is Machine Learning?

What is Machine Learning?



Machine Learning

Arthur Samuel:

Machine learning is "Field of study that gives computers the ability to learn without being explicitly programmed".

Even according to the experts, the exact definition of the field of machine learning is a bit fuzzy, but As early as 1959, Arthur Samuel quote. .

also, include Ng's explanation from MLYearning!

# What is Machine Learning

data + algorithms = predicting the future  
combo of statistics, calculus, etc...

# Machine Learning From Scratch

## └ Machine Learning Overview

### └ What is Machine Learning

What is Machine Learning

data + algorithms = predicting the future  
combo of statistics, calculus, etc...

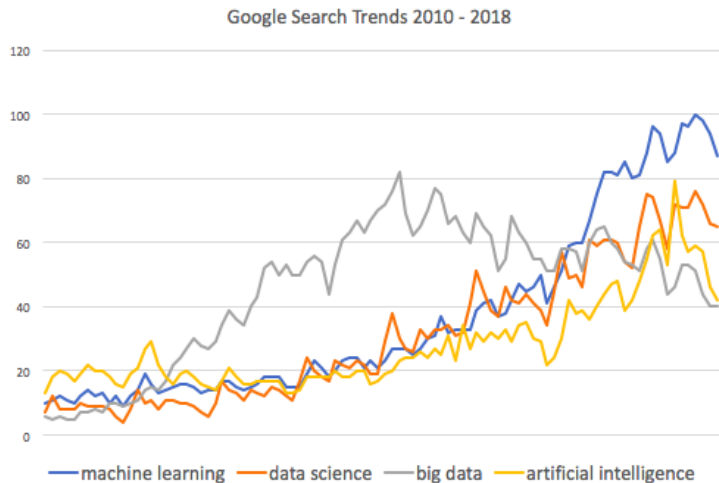
ML techniques can be applied to a wide range of problems in diverse industries. In fact, ML has become ubiquitous in our everyday lives \*

Siri/ Amazon Alexa

- \* Recommendation systems (amazon, netflix)
- \* Fraud Detection
- \* Disease diagnosis
- \* Supply Chain Optimization



# According to Google...



# What has caused this spike?

## 1. Data Availability

- ▶ ecommerce
- ▶ lot (sensor data)

## 2. Computational Scale

- ▶ Moore's Law

# Machine Learning From Scratch

## └ Machine Learning Overview

### └ What has caused this spike?

What has caused this spike?

1. Data Availability
  - ecommerce
  - IoT (sensor data)
2. Computational Scale
  - Moore's Law

The math that powers machine learning algorithms has been around for quite a few years... so what's changed?

1. Data Availability
2. Computational Scale (NG MLY 01 pg 10)

The rise of the big data era has given us access to astounding amounts of data. That phenomenon paired with the exponential growth we've experienced in computational advances, has created the perfect storm for the emergence of the field of machine learning.

## Machine Learning Overview

### Steps in the Machine Learning Process

Step 0: Identify The Problem

Step 1: Get the Data

Step 2: Data Exploration

Step 3: Data Preparation

Step 4: Model Selection

Step 5: Cross-validation/Hyper-parameter tuning

## Building a Support Vector Machine from Scratch

## Exploring Scikit-Learn and applying to GR Crash dataset



# Machine Learning From Scratch

## └ Steps in the Machine Learning Process

### └ Step 0: Identify The Problem



Let's say that you work for the city of Grand Rapids, and you find that there are an increased number of hit and runs when the driver 1 was drinkin.

**Do some research, maybe find a way to graph this?? You can do it!**

# Get the Data

This may look like:

- ▶ SQL query
- ▶ CSV download
- ▶ Web-scraping
- ▶ Designing experiments/surveys and collecting data yourself

# Get the Data

This may look like:

- ▶ SQL query
- ▶ CSV download
- ▶ Web-scraping
- ▶ Designing experiments/surveys and collecting data yourself

In our case, we'll head to [GRData](#).



# Machine Learning From Scratch

## └ Steps in the Machine Learning Process

### └└ Step 1: Get the Data

#### └└└ Get the Data

Get the Data

This may look like:

- SQL query
- CSV download
- Web-scraping
- Designing experiments/surveys and collecting data yourself

In our case, we'll head to [GRData](#).

Obtaining the data you'll need looks very different depending on what domain you're working in. In some instances, it can be fairly simple and straightforward, for example, In a business context, most often it will require querying some sort of internal database. Could also be downloading a csv file. In other instances, it may require a bit more creativity – For particular social research, you may need to scrape the web. In some cases, you may even need to collect some data yourself! Here are two examples:

1. You're developing a new data product at your company and are collecting data to fuel it
2. You're in public health and are working to make healthcare accessible to all residents of the greater GR area. You may need to conduct your own research to identify what may be inhibiting people from reaching healthcare.

In our case, we're lucky enough to have access to a meticulously maintained public database on the city of GR: GRData. Scroll through,

# Get the Data

In our case, we'll head to [GRData](#)

```
In [33]: crash = pd.read_csv('Data/CGR_Crash_Data.csv')
         crash.head()
```

```
Out[33]:
```

	X	Y	OBJECTID	ROADSOFTID	BIKE	CITY	CRASHDATE	CRASHSEVER	CRASHTYPE	WORKZNEACT	...
0	-85.639647	42.927216	6001	929923	No	Grand Rapids	2007-02-16	Property Damage Only	Side-Swipe Same	Uncoded & Errors	...
1	-85.639487	42.927213	6002	935745	No	Grand Rapids	2007-06-22	Property Damage Only	Side-Swipe Same	Uncoded & Errors	...
2	-85.639387	42.927212	6003	926813	No	Grand Rapids	2007-01-08	Property Damage Only	Head-on	Work on Shoulder / Median	...
3	-85.639288	42.927210	6004	943813	No	Grand Rapids	2007-11-12	Property Damage Only	Side-Swipe Same	Uncoded & Errors	...
4	-85.639288	42.927210	6005	943791	No	Grand Rapids	2007-11-09	Property Damage Only	Parking	Uncoded & Errors	...

5 rows × 77 columns

2018-08-16

# Machine Learning From Scratch

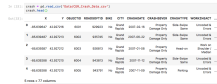
## Steps in the Machine Learning Process

### Step 1: Get the Data

#### Get the Data

[Get the Data](#)

In our case, we'll head to [CRData](#)



The screenshot shows a Jupyter Notebook interface. At the top, there's a code cell with the command `df = pd.read_csv('baseball_data.csv')`. Below it, a pandas DataFrame is displayed, showing columns: DATE, TIME, VISITOR, HOME, VISITOR\_WINS, HOME\_WINS, VISITOR\_RUNS, HOME\_RUNS, VISITOR\_HITS, HOME\_HITS, VISITOR\_ERRORS, HOME\_ERRORS, VISITOR\_LEFTS, HOME\_LEFTS, VISITOR\_STRIKES, HOME\_STRIKES, VISITOR\_Pitches, HOME\_Pitches, VISITOR\_Pitches, HOME\_Pitches, VISITOR\_Pitches, HOME\_Pitches. The first few rows of data are visible, showing game details for various dates and teams.

	DATE	TIME	VISITOR	HOME	VISITOR_WINS	HOME_WINS	VISITOR_RUNS	HOME_RUNS	VISITOR_HITS	HOME_HITS	VISITOR_ERRORS	HOME_ERRORS	VISITOR_LEFTS	HOME_LEFTS	VISITOR_STRIKES	HOME_STRIKES	VISITOR_Pitches	HOME_Pitches	VISITOR_Pitches	HOME_Pitches
0	1961-04-15	19:00	Red Sox	Yankees	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1961-04-16	19:00	Red Sox	Yankees	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1961-04-17	19:00	Red Sox	Yankees	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1961-04-18	19:00	Red Sox	Yankees	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1961-04-19	19:00	Red Sox	Yankees	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

After downloading the csv file, we can read the file into a pandas dataframe and explore it in our Jupyter notebook.

\* note something about how we'll often refer to each row as an observation and each column as a feature

# Explore the Data

- ▶ Verify data
- ▶ Visualize data
- ▶ Identify patterns
- ▶ Give direction to analysis

# Machine Learning From Scratch

## └ Steps in the Machine Learning Process

### └ Step 2: Data Exploration

#### └ Explore the Data

[Explore the Data](#)

- Verify data
- Visualize data
- Identify patterns
- Give direction to analysis

This is kind of an unstructured approach to understanding initial patterns in the data and potentially points of interest. This process isn't meant to reveal every bit of information a dataset holds, but rather give you direction in your analysis and potentially give you clues in how to process/model the data.[1] Now, if you're just emailed a csv file, this step is especially crucial, and it may take you some time to explore the data, get a feeling for what you're dealing with. If you are analyzing data that you work with day in and day out, this "exploration" process may be a bit more implicit.

The main idea here is to build an understanding of your data. Without an appreciation for the context of the data, it's just numbers. But when you see the data in context, it's fascinating, it's a story.

More often than not, your exploration of the data leads to more questions than answers.

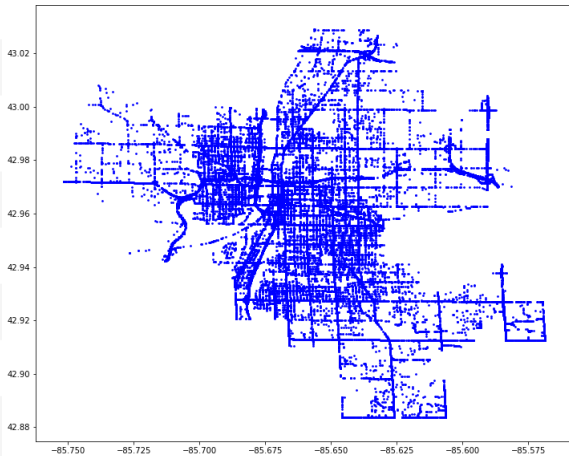
	X	Y
0	-85.639647	42.927216
1	-85.639487	42.927213
2	-85.639387	42.927212
3	-85.639288	42.927210
4	-85.639288	42.927210
5	-85.639188	42.927208
6	-85.639168	42.927208
7	-85.639108	42.927207

# Machine Learning From Scratch

## └ Steps in the Machine Learning Process

### └ Step 2: Data Exploration

	X	Y
0	-85.639647	42.927216
1	-85.639487	42.927213
2	-85.639387	42.927212
3	-85.639288	42.927210
4	-85.639288	42.927210
5	-85.639188	42.927208
6	-85.639168	42.927208
7	-85.639108	42.927207

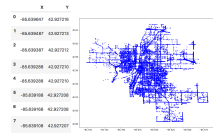


2018-08-16

# Machine Learning From Scratch

## └ Steps in the Machine Learning Process

### └ Step 2: Data Exploration



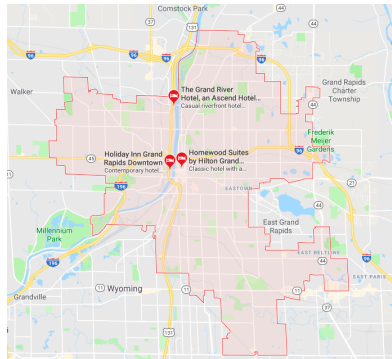
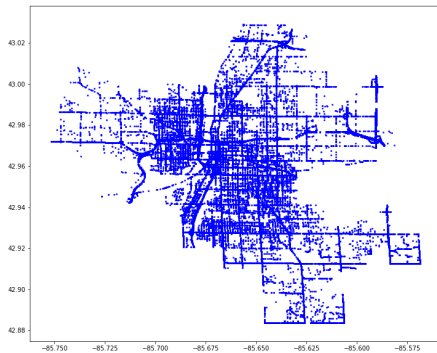
With our dataset, on car crashes, a logical place to begin would be the first two columns, containing latitudes and longitudes of each crash. This is just a snapshot of the data on the left side, and on the right, each dot represents a car crash.



## Machine Learning From Scratch

## Steps in the Machine Learning Process

## Step 2: Data Exploration

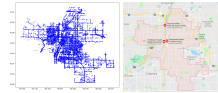


2018-08-16

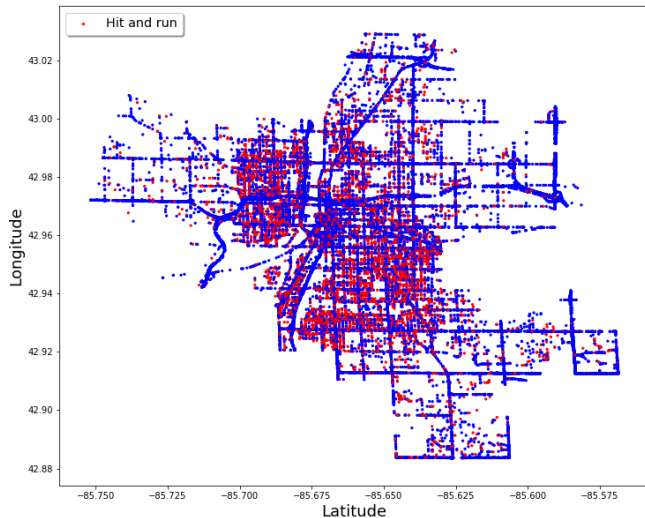
# Machine Learning From Scratch

## └ Steps in the Machine Learning Process

### └ Step 2: Data Exploration



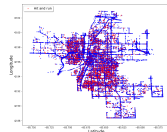
Now, this is pretty telling about our data, remember, there is nearly 73,000 crashes recorded, and if we juxtapose this plot with the city of GR, we actually see that the plot of crashes outline the city boundaries!



2018-08-16

# Machine Learning From Scratch

- └ Steps in the Machine Learning Process
  - └ Step 2: Data Exploration



We may be interested in hit and runs...

I should add a caption or title to this plot

# Check the variable's distribution

```
In [34]: crash = pd.read_csv('Data/CGR_Crash_Data.csv')
         crash.head(3)
```

```
Out[34]:
```

	X	Y	OBJECTID	ROADSOFTID	BIKE	CITY	CRASHDATE	CRASHSEVER	CRASHTYPE	WORKZNEACT	...
0	-85.639647	42.927216	6001	929923	No	Grand Rapids	2007-02-16	Property Damage Only	Side-Swipe Same	Uncoded & Errors	...
1	-85.639487	42.927213	6002	935745	No	Grand Rapids	2007-06-22	Property Damage Only	Side-Swipe Same	Uncoded & Errors	...
2	-85.639387	42.927212	6003	926813	No	Grand Rapids	2007-01-08	Property Damage Only	Head-on	Work on Shoulder / Median	...

3 rows x 77 columns

```
In [32]: crash.VEH3TYPE.value_counts()
```

```
Out[32]:
```

Uncoded & Errors	67212
Passenger Car, SUV, Van	4788
Pickup Truck	503
Motorhome	327
Truck Under 10,000 lbs	63
Truck / Bus (Commercial)	62
Other Non-Commercial	10
Motorcycle	10
Go-cart / Golf Cart	2

Name: VEH3TYPE, dtype: int64

## Steps in the Machine Learning Process

## Step 2: Data Exploration

- Check the variable's distribution

[illegible]

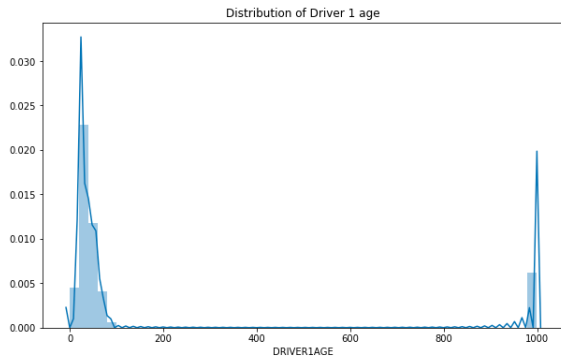
It's also often very helpful to check the distribution of the different variables. For example, our dataset contains many characteristics about what it defines as "DRIVER1", "DRIVER2", "DRIVER3". When I first came across that, I was impressed, like that's some seriously accurate data! But upon further examination, we find that many of the "DRIVER2" and "DRIVER3" columns are empty or contain errors. (This makes sense... not all crashes involve 3 driver!)

When we check the counts of the different values found in the "VEH3" column, we see that over 67,000 of them are errors! Now, this doesn't mean that the column is useless, but in terms of building a predictive model, this column probably won't be much help, so as we'll see in the data processing step, we'll end up dropping it.

# Check the variable's distribution

```
In [41]: fig, ax = plt.subplots(figsize=(10,6))  
         ax.set_title('Distribution of Driver 1 age')  
         sns.distplot(crash.DRIVER1AGE)
```

```
Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1a742080>
```

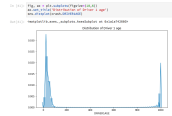


2018-08-16

# Machine Learning From Scratch

- └ Steps in the Machine Learning Process
  - └ Step 2: Data Exploration
    - └ Check the variable's distribution

Check the variable's distribution



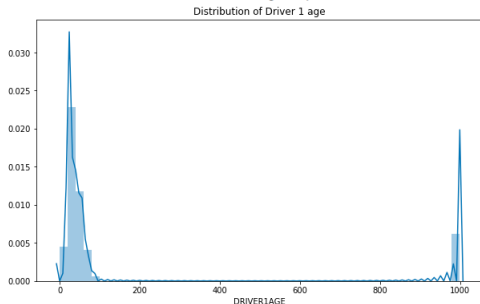
Continuing on, here we check the distribution of the age's of all the "DRIVER1"s recorded in the dataset. In my head, I would think that age would be an important feature in car crashes. And so we check it out, and it appears that there are a ton of instances bunched up between 1 and 100... that makes sense... then there is also a decent cluster of observations around 1000 years old... that does not make sense.



## Check the variable's distribution

```
In [45]: fig, ax = plt.subplots(figsize=(10,6))  
         ax.set_title('Distribution of Driver 1 age')  
         sns.distplot(crash.DRIVER1AGE)
```

There are 8979 Driver 1's recorded as being 999 years-old.



```
In [46]: print('There are', crash.DRIVER1AGE[crash.DRIVER1AGE == 999].count(), "driver 1's recorded as being 999 years-old.")
```

There are 8979 driver 1's recorded as being 999 years-old.

# Machine Learning From Scratch

- └ Steps in the Machine Learning Process
  - └ Step 2: Data Exploration
    - └ Check the variable's distribution

Check the variable's distribution



Nearly 9,000 driver 1's are recorded as being 999 years-old.. That is a good thing to know before trying to build a predictive model, as it seriously skews the data. We'll address that in our data processing stage.

Next, we move onto data preparation. This is the stage where we make the final manipulations to our data before feeding it into our ML algorithm. Now, you could make an argument that preparing the data is the most important part of the machine learning workflow. After all it's the data that fuels the algorithm, garbage in, garbage out! It's been shown time and time again that more/bigger data beats a better algorithm everytime. Though it usually appears to be straightforward, this step can often require a lot of creativity.

# Data Selection

Use as minimal features as possible

1. Computationally efficient
2. Easier to interpret
3. Simpler is better

# Data Selection

Use as minimal features as possible

1. Computationally efficient
2. Easier to interpret
3. Simpler is better

```
In [8]: crash = pd.read_csv('Data/CGR_Crash_Data.csv')
crash = crash[['X', 'Y', 'CRASHSEVER', 'DRIVER1AGE', 'DRIVER1SEX',
               'EMRGVEH', 'HITANDRUN', 'SPEEDLIMIT', 'HOUR', 'MOTORCYCLE',
               'NUMOFINJ', 'D1COND', 'D1DRINKIN']]
crash.columns

Out[8]: Index(['X', 'Y', 'CRASHSEVER', 'DRIVER1AGE', 'DRIVER1SEX', 'EMRGVEH',
               'HITANDRUN', 'SPEEDLIMIT', 'HOUR', 'MOTORCYCLE', 'NUMOFINJ', 'D1COND',
               'D1DRINKIN'],
              dtype='object')
```

## Steps in the Machine Learning Process

### Step 3: Data Preparation

## └ Data Selection

Use as minimal features as possible

1. Computationally efficient
2. Easier to interpret
3. Simpler is better

[illegible]

The first step in preparing the data is simply choosing which features (you can think of as columns in the spreadsheet) you'll want to use in your model. As seen in our EDA some columns have a lot of missing data, and we'll drop them entirely.

It's generally regarded as best practice to use as minimal amount of features as possible, such that your predictive model still predicts as accurately as you need it to.

- \* computationally efficient
- \* more easily interpretable
- \* simpler is better

So how do we actually determine which features to use?, we can use various statistical tests, and even some algorithms to determine which features are going to be most relevant to predicting our target variable (which for us is whether or not the driver who caused the crash was drinking). We won't go into detail here.

When you are first beginning to iterate through different ml models, it is okay to kind of eyeball it or use what you know about the context of the

# Feature Engineering



*'Coming up with features is difficult, time-consuming, requires expert knowledge. Applied machine learning is basically feature engineering.'*

Prof. Andrew Ng



*'At the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used.'*

Prof. Pedro Domingo

# Machine Learning From Scratch

## └ Steps in the Machine Learning Process

### └ Step 3: Data Preparation

#### └ Feature Engineering

#### Feature Engineering



*"Coming up with features is difficult, time-consuming, requires expert knowledge. Applied machine learning is basically feature engineering."*  
Prof. Andrew Ng



*"At the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used."*  
Prof. Pedro Domingo

So now we have our 13 features that we've chosen to use, and we could just send these raw features into our algorithm, and it may perform well, but it may not. It's important to remember that our ultimate goal is to build a predictive model that can make accurate predictions on new/unseen observations. When all the data comes in from a car crash that just happened, we want to be able to accurately predict whether or not it was a drunk driver that caused it.

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. - Jason Brownlee [2]

We acknowledge, however, that the data that's being recorded (from those UD10 reports) isn't necessarily guaranteed to accurately represent reality. Which is an important thing if we want to build accurate, stable predictive models.

The examples we've shown here are absolutely not exhaustive when it

## Feature Engineering: *transforming "hour" variable*



## Feature Engineering: *imputing missing ages*

content...

# Data Processing

Two general types of data to deal with:

- ▶ Numerical variables (Quantitative)
  - ▶ Driver 1 age, number of injuries, etc
- ▶ Categorical variables (Qualitative)
  - ▶ Hit and run, motorcycle involved, etc

# Machine Learning From Scratch

## Steps in the Machine Learning Process

### Step 3: Data Preparation

#### Data Processing

#### Data Processing

Two general types of data to deal with:

- Numerical variables (Quantitative)
  - Driver 1 age, number of injuries, etc
- Categorical variables (Qualitative)
  - Hit and run, motorcycle involved, etc

Now, we move onto processing the data that we've selected. The data processing stage naturally diverges into two substeps: dealing with numerical variables, and dealing with categorical variables.

It's important to note that some of the preprocessing steps we'll talk about here may actually be necessary for you to do to get the data in a format where you can explore it.

Numerical variables are quantitative – it's something you can measure. In our case, some numerical variables are Driver1 age, and number of injuries.

Categorical variables are qualitative, in our case, we have some binary categorical variables: like whether or not the crash was a hit and run, whether or not a motorcycle was involved. It's either one or the other. Variables can of course have multiple categories, for example, which car insurance provider you choose: (Nationwide, Statefarm, BlueCross BlueShield) There is obviously a lot of them, but all drivers fit into one of those categories... or at least they should.

There is some grey area between the two – maybe mention speedlimits or

## Data Processing: *numerical variables*

## Machine Learning Overview

### Steps in the Machine Learning Process

Step 0: Identify The Problem

Step 1: Get the Data

Step 2: Data Exploration

Step 3: Data Preparation

Step 4: Model Selection

Step 5: Cross-validation/Hyper-parameter tuning

## Building a Support Vector Machine from Scratch

### Exploring Scikit-Learn and applying to GR Crash dataset

## Machine Learning Overview

### Steps in the Machine Learning Process

Step 0: Identify The Problem

Step 1: Get the Data

Step 2: Data Exploration

Step 3: Data Preparation

Step 4: Model Selection

Step 5: Cross-validation/Hyper-parameter tuning

### Building a Support Vector Machine from Scratch

### Exploring Scikit-Learn and applying to GR Crash dataset



<https://www.sisense.com/glossary/data-exploration/>



<https://towardsdatascience.com/understanding-feature-engineering-part-2-categorical-data-f54324193e63>