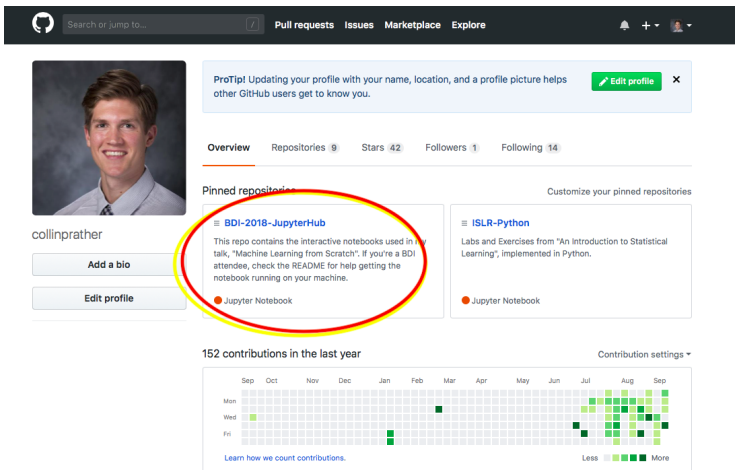
Welcome!

While you're waiting... I've prepared a Jupyter notebook that we will use to explore our data and build a machine learning algorithm from scratch. In order to get the notebook up and running on your computer:

1.) Head to *https://github.com/collinprather*
2.) Click on the "BDI-2018-JupyterHub" repository
3.) Scroll down and follow the step-by-step instructions in the readme.md



Machine Learning From Scratch

Collin Prather

September 21st, 2018

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

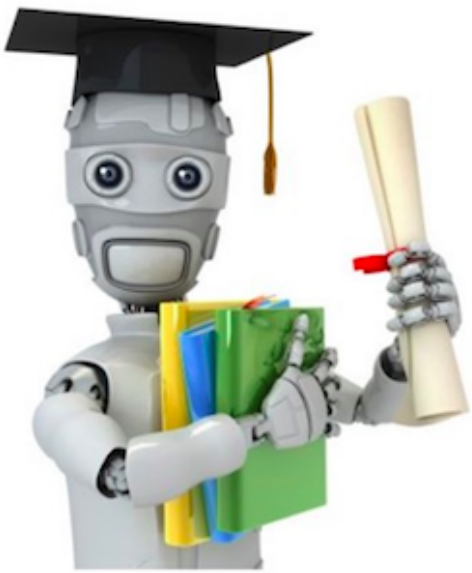
Building a Support Vector Machine from Scratch

- Representation
- Evaluation
- Optimization

Exploring Scikit-Learn and applying to GR Crash dataset

Machine Learning Overview

What is Machine Learning?

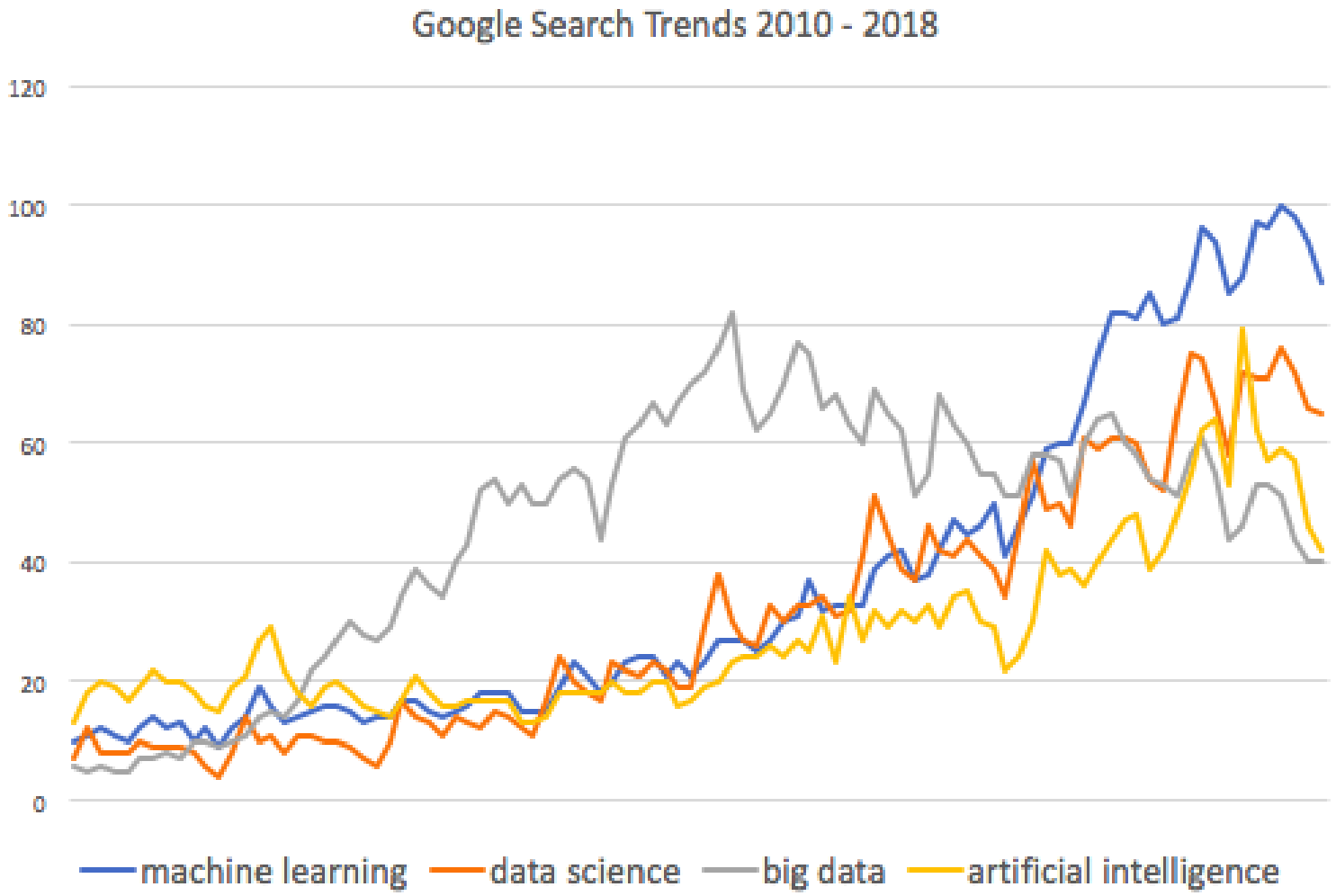


Machine Learning

Arthur Samuel:

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

According to Google...



Identify the Problem

Get the Data

In our case, we'll head to GRData

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	...

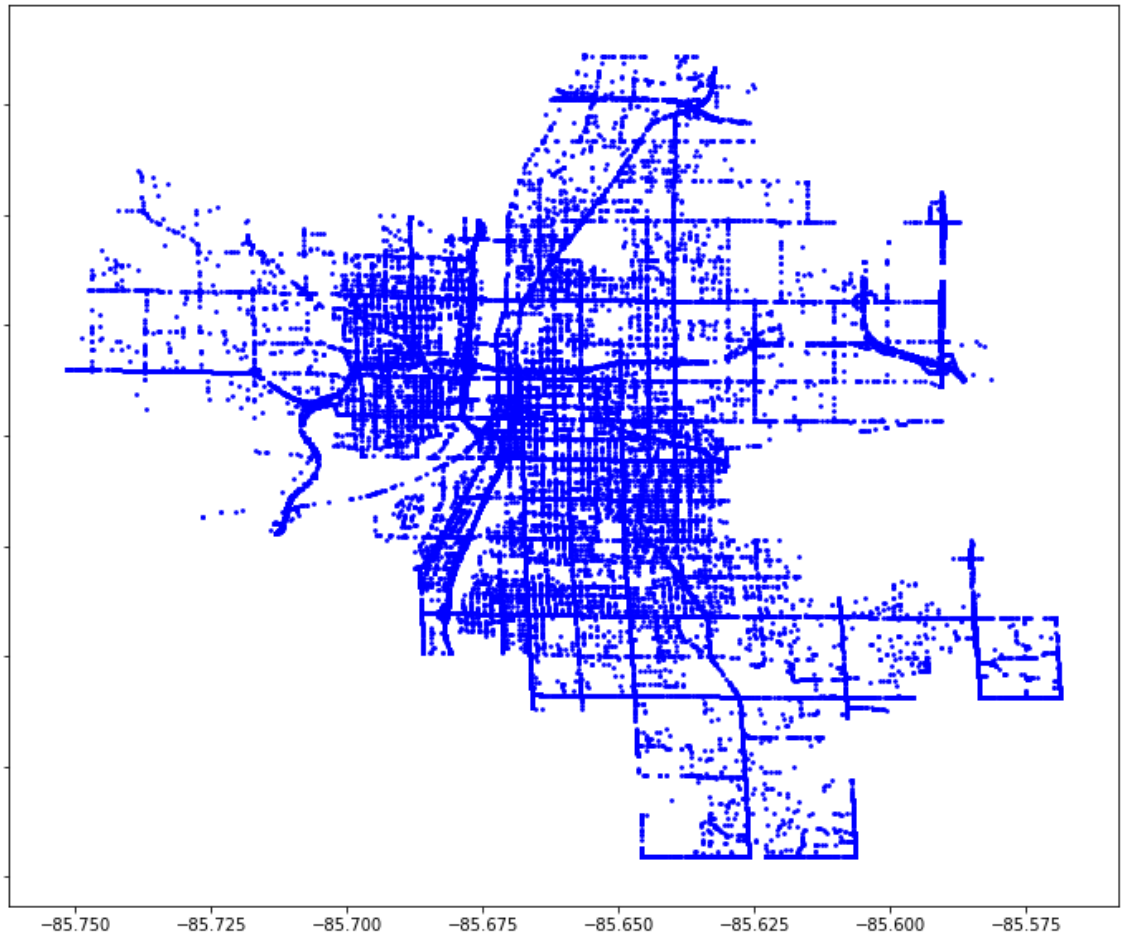
5 rows × 77 columns

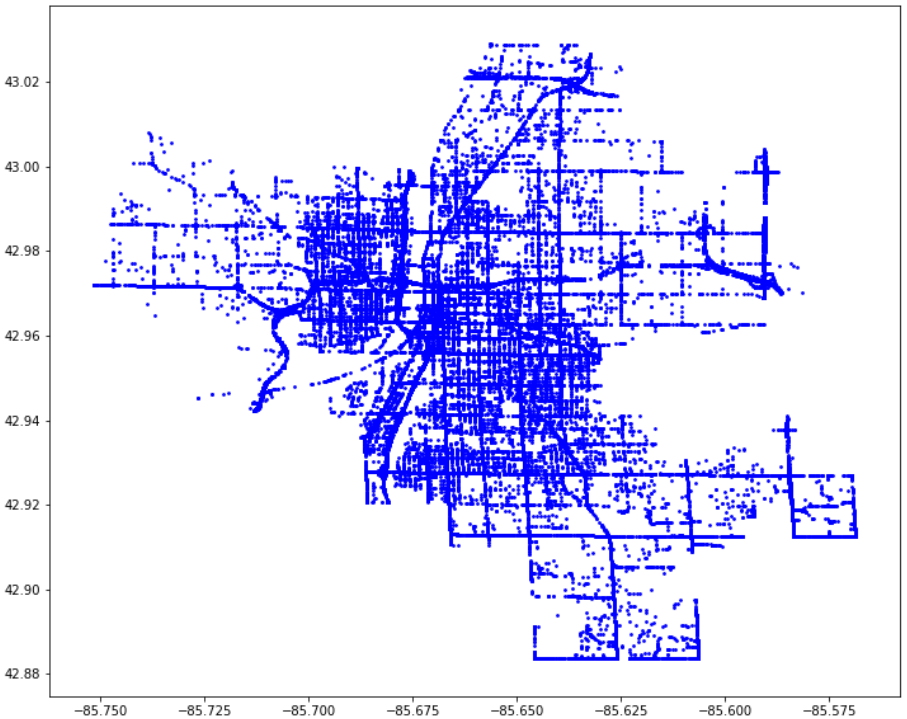
Explore the Data

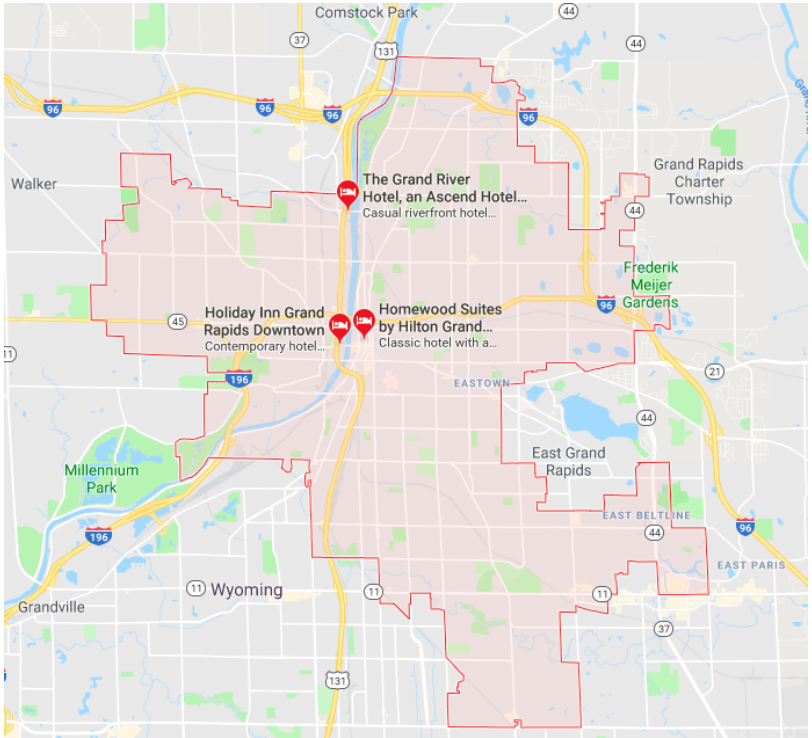
- Verify data
- Visualize data
- Identify patterns
- Give direction to analysis

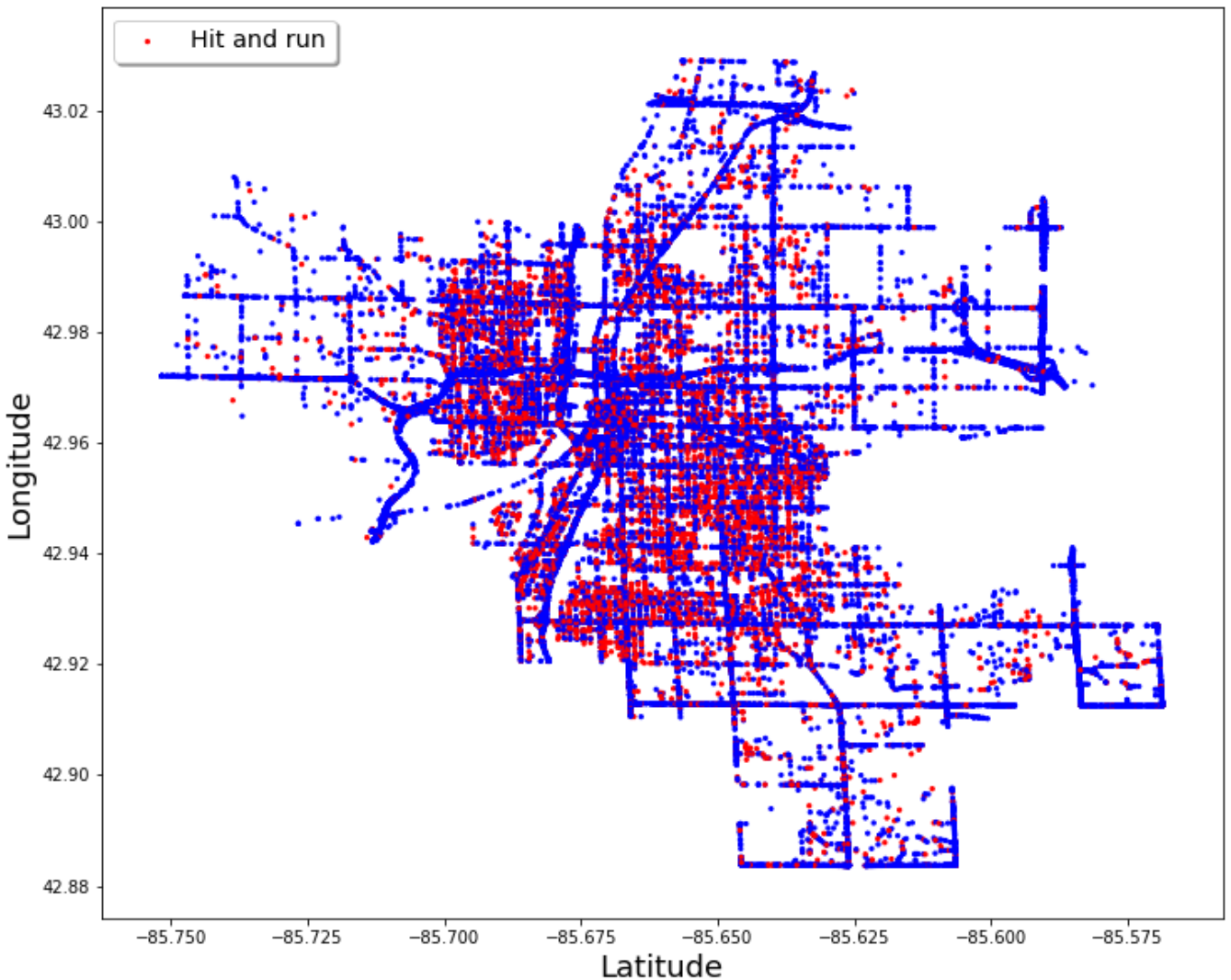
	X	Y
0	-85.639647	42.927216
1	-85.639487	42.927213
2	-85.639387	42.927212
3	-85.639288	42.927210
4	-85.639288	42.927210
5	-85.639188	42.927208
6	-85.639168	42.927208
7	-85.639108	42.927207

	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




43.02
43.00
42.98
42.96
42.94
42.92
42.90
42.88
−85.750
−85.725
−85.700
−85.675
−85.650
−85.625
−85.600
−85.575


Comstock Park
Walker
The Grand River Hotel, an Ascend Hotel...
Casual riverfront hotel...
Grand Rapids Charter Township
Frederik Meijer Gardens
Holiday Inn Grand Rapids Downtown
Contemporary hotel...
Homewood Suites by Hilton Grand...
Classic hotel with a...
EASTOWN
East Grand Rapids
Millennium Park
EAST BELTLINE
EAST PARIS
Wyoming
Grandville


Hit and run
43.02
43.00
42.98
42.96
42.94
42.92
42.90
42.88
Longitude
−85.750
−85.725
−85.700
−85.675
−85.650
−85.625
−85.600
−85.575
Latitude

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 Domingos

Feature Engineering: *transforming "hour" variable*

$$\begin{bmatrix} 12 \\ 2 \\ 4 \\ 18 \\ 19 \\ 6 \\ \vdots \end{bmatrix}$$

Feature Engineering: *transforming "hour" variable*

$$\begin{bmatrix} 12 \\ 2 \\ 4 \\ 18 \\ 19 \\ 6 \\ \vdots \end{bmatrix} \quad \Longrightarrow \quad f(hour) = \frac{2 \cdot \pi \cdot (\text{hour})}{24}$$

Feature Engineering: *transforming "hour" variable*

$$\begin{bmatrix} 12 \\ 2 \\ 4 \\ 18 \\ 19 \\ 6 \\ \vdots \end{bmatrix} \quad \Longrightarrow \quad f(hour) = \frac{2 \cdot \pi \cdot (\text{hour})}{24} \quad \Longrightarrow \quad \begin{bmatrix} 3.14 \\ 0.52 \\ 1.03 \\ 4.71 \\ 4.98 \\ 1.57 \\ \vdots \end{bmatrix}$$

Feature Engineering: *transforming "hour" variable*

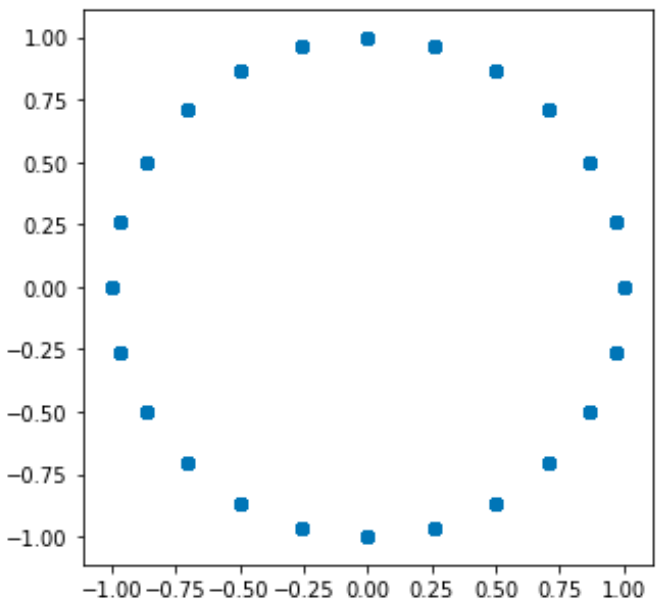
$$\begin{bmatrix} 3.14 \\ 0.52 \\ 1.03 \\ 4.71 \\ 4.98 \\ 1.57 \\ \vdots \end{bmatrix} \Longrightarrow \underbrace{\begin{bmatrix} 0.00 \\ 0.47 \\ 0.86 \\ -0.99 \\ -0.97 \\ 1.0 \\ \vdots \end{bmatrix}}_{sin(f(hour))}, \underbrace{\begin{bmatrix} -1.0 \\ 0.87 \\ 0.51 \\ -0.002 \\ -0.26 \\ 0.001 \\ \vdots \end{bmatrix}}_{cos(f(hour))}$$

Feature Engineering: *transforming "hour" variable*

In [193]:
```
crash['HOUR_X']=np.sin(2. * np.pi * crash.HOUR / 24.)
crash['HOUR_Y']=np.cos(2. * np.pi * crash.HOUR / 24.)
```

In [194]:
```
# Hence, the time of day is now cyclic (just as in reality)
plt.figure(figsize = (5,5))
plt.scatter(crash.HOUR_X, crash.HOUR_Y)
```

Out[194]:
```
<matplotlib.collections.PathCollection at 0x1a0daeca20>
```



Data Processing

Two general types of data to deal with:

Data Processing

Two general types of data to deal with:

- Numerical variables (Quantitative)
  - Driver 1 age, number of injuries, etc

Data Processing

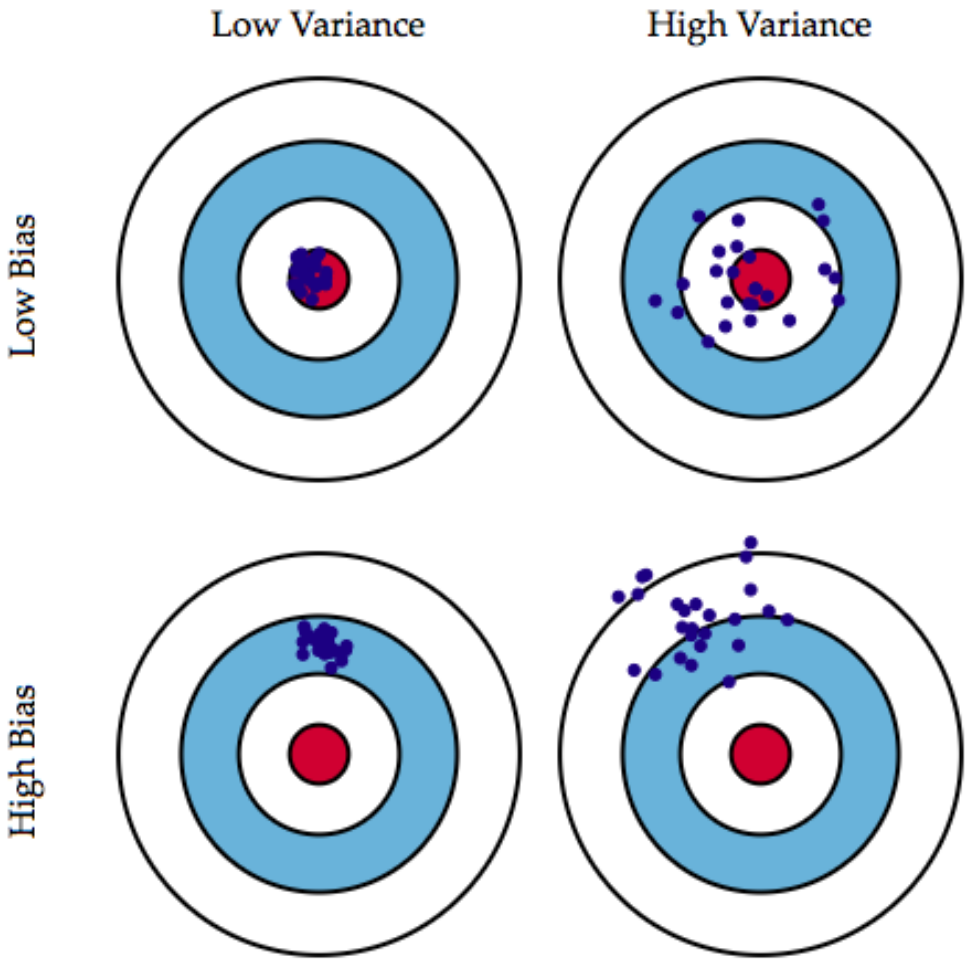
Two general types of data to deal with:

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

Choosing a Model/Representation

	Classification	**Regression**
Supervised	• Logistic Regression • Naive-Bayes • KNN • SVM	• Linear Regression • Decision Trees • Random Forests
Unsupervised	• Apriori • Hidden Markov Model	• PCA • K-means • SVD

Bias-Variance Tradeoff



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

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

 http://scott.fortmann-roe.com/docs/BiasVariance.html