

- 1.
 - Hi guys, we are team Crouching Tigers, and our project is about predictive maintenance on IOT devices.
- 2.
 - When approaching a problem, it's always a good idea to ask questions in the form of 5W1H. It typically means who, what, where, when, why and how. Today we will try to answer the following questions in 4 minutes.
 - What is the problem we want to solve? Why is the problem important? Who can benefit from the outcome of our project? And finally, how do we approach this problem? What have we done already and lastly what do we plan on doing next
- 3.
 - IoT, or Internet of Things is just more than just a fad. As you can see in this chart from Google Trends, we see the first interest spike back in early 2014. It has pretty much been growing non-stop ever since.
- 4.
 - More interestingly, the world had about 12.5 billion IoT devices in 2012, the number hit 25 billion on 2015. It's projected to be over 50 billion by 2020.
 - With that amount of growth, it definitely comes with some problems. Among with reliability, privacy, and other concerns, maintainability and longevity of the devices has always be the top issue. Manufacturer wants to know potential issues with a certain device before it comes out on the market. Service department wants to plan their labor and resources more efficiently. And we all know that it's more than just about 4,000 dollar Samsung smart fridge.
- 5.
 - The agriculture industry can benefit from putting more reliable sensors on livestock and plants for monitoring and

analyzing, dramatically increasing productivity.

- More vehicles are equipped with internet connectivity, sensors, GPS systems. It is estimated that there will be 381 million connected cars on the road by 2020, and a great amount of them will be at least semi-autonomous. It goes without saying the maintainability and reliability will be important.
- And there's construction, healthcare, retail, the list just goes on..
- 6.
 - Before diving into the problem itself, the original Kaggle challenge was deleted and unfortunately we didn't get the dataset from our contact person too. We spent about a week to get a usable dataset.
 - Overall we decided to approach this problem four ways.
 - The most intuitive way is to view it as a regression problem. Essentially, predicting how many days are left before the machine fails. We used Spark ML pipeline, using a regression model. We got a mean square error of 0.05.
 - The second approach is not as intuitive, we can also see it as the classification problem. Basically it's like saying is my smart fridge going to fail on March. We also used Spark ML pipeline, along with DecisionTreeClassifier and were able to achieve about 74% accuracy.
 - Next, we will apply what we just learned from the lab last week and try to do anomaly detection. Some challenge we foresee right now is anomalous behavior does not always lead to machine failure and anomaly detection does not tell you when it will happen. That's going to take us at least a week. We will also try doing better model selection and parameter tuning to yield better prediction results.
 - We will also build a streaming application to simulate real time data. We will either build it as a Kafka server or a NodeJS server. It will take us about a week as well.
 - We will need to find more datasets and generalize our solution

