Proposal one:

Kaggle Competition:

https://inclass.kaggle.com/c/adcg-ss14-challenge-02-spam-mails-detection

In current society, anomaly detection gains its own importance from researchers day by day. This is a classification project based on spam emails. The data set is available on Kaggle.

I plan apply Spark and some machine learning methods including clustering, classification methods such as SVM to analyze the spam email data set. I would like to design and implement an automation of anomaly identification. I have been provided both the training data set and test data set. I plan to apply python or Scala in Spark to build such a machine learning project.

Some questions we might think about:

- Extract keywords and features from emails which are eml format.
- Determine features or keywords that are considered spam relevant
- Determine the spam keywords threshold of an email, which means how many spam keywords can classify a spam email.

Our deliverable will be:

- The machine learning models which classify the training data set.
- The source code of Python or Scala (TBD) which runs on Spark.
- The report and paper which describes the whole approach towards this problem.

Proposal two:

Kaggle Competition:

https://www.kaggle.com/c/yelp-restaurant-photo-classification

This is a machine learning project based on yelp restaurant photos. Based on photo labels, users are able to classify the restaurants according to their specific needs. This project is to classify restaurants with photos labels submitted by users.

I am considering using machine learning models such as clustering, k-NN algorithms in Spark. Maybe I can make use of SVM classifier same as proposal one. The data set is large as 2 GB.

Some questions we might think about:

- How to extract photos labels or reviews into summaries of each restaurant.
- Determine the keywords or features that in a label or review that classifies this label or review
- Which machine learning model should be applied.

Our deliverable will be:

- The classification result and tags of each photo.
- The machine learning model(s) that I chose to classify the data set.
- The report and paper which describes the whole approach towards this problem.