# Predicting coding question quality using Stack Overflow ratings

## Slide 2: Overview

* Stack Overflow (SO)
* What is a question?
* Support Vector Machine (SVM)
* Methodology
* Experiments and Results
* Summary

Demo

## Slide 3: Stack Overflow

* Released late 2008
* A Question-Answering (QnA) community for programmers
* Model which Stack Exchange is built on
* Uses gamification to reward users for participation
* Reputation
* Votes
* Badges
* Accepted answer
* Who are the experts?
* What defines a good Question and Answer on SO?

**TODO:** Look into *"... how much it would have cost to pay people to rate the questions if you were doing it on a time spend reading and rating*"

## Slide 4: Picture of good SO question

"How do you undo the last commit?" - 10,493 votes, Community Wiki  
In right bottom corner, number is reference to source (last slide)

## Slide 5: Picture of bad SO question

"Forcing function to return if false" [locked, closed, off-topic], -154

## Slide 6: Picture of Badges on SO

Picture of question badges

## Slide 7: What is a question?

* Factoid vs. Broad questions
* In education: Learn something new, or evaluate knowledge
* Could be the goal of your research
* The quality of a question can be equal to the quality of the answer
* Question classification: Categorizing questions
  + WH-words
  + Bag-of-Words and N-grams
  + Word mapping and processing

## Slide 8: Support Vector Machine (SVM)

* Good for regression and classification problems
* Main focus is binary classification
* Often used for text classification
* Separates classes by using a hyperplane
* Four kernels:
  + Linear
  + Radial Basis Function (RBF)
  + Polynomial
  + Sigmoid

## Slide 9: Methodology

* Data set based on data dump from Stack Exchange Archive
  + Contains XML files based on table content
  + Imported data into MySQL database
  + Imported data from MySQL database into Pandas.DataFrame
* Development: Python 3.5 and Scikit-learn (0.18.dev0)
* Question processing
* Selecting questions and features
* Selecting estimator and parameters for classification

## Slide 9: Experiments and Results

* 6 different features
  + Code samples
  + Hexadecimal
  + Homework (synonyms for homework)
  + Links
  + Numerical
  + Tags
  + All features
* 4 different experiments
  + Unprocessed data set vs. all singular feature, and all questions
  + Unprocessed data set vs. all singular features, and question occurence only
  + Unprocessed data set vs. selected set of features only
  + Stochastic Gradient Descent (SGD) as classifier

## Slide 11: Conclusion

* Stack overflow as a question quality metric
* Limitations and issues
* Further work
  + Code blocks, Links and Numerical as a feature set
  + Code analysis
  + Sentiment analysis
  + Version numbering

## Slide 12: Demo

Show demo (if time)

## Slide 13: Thanks for listening

## Slide 14: References