StackOverflow: A journey of bounty hunters

Social Media Mining WS 14/15

Tom Bocklisch & Tom Herold

Recap

8.4m questions

90k bounties

Only 1 bounty / question at a time

Bounty creation only after 48h

Research Questions

Data Cleansing

0. Understand the data and eliminate faulty data.

Analysis

1. What are the intrinsic factors and signals that are likely to influence a bounty's response time?

Prediction

2. Can we predict a successful claim of a bounty?
3. Can we predict whether an answer will be given in a certain timespan?

NEXT

Data Cleansing

What does the data look like?

What does the data look like?

- 80GB XML Data
- 2008-present
- All of StackExchange
- 7 SO Tables
- Including post history

Faulty data

2 + 2 = 5

Partially anonymized

WHAT WE DID

Removed 3000 elements

Cleaned up invalid references

Ensured semantic consistency of bounties

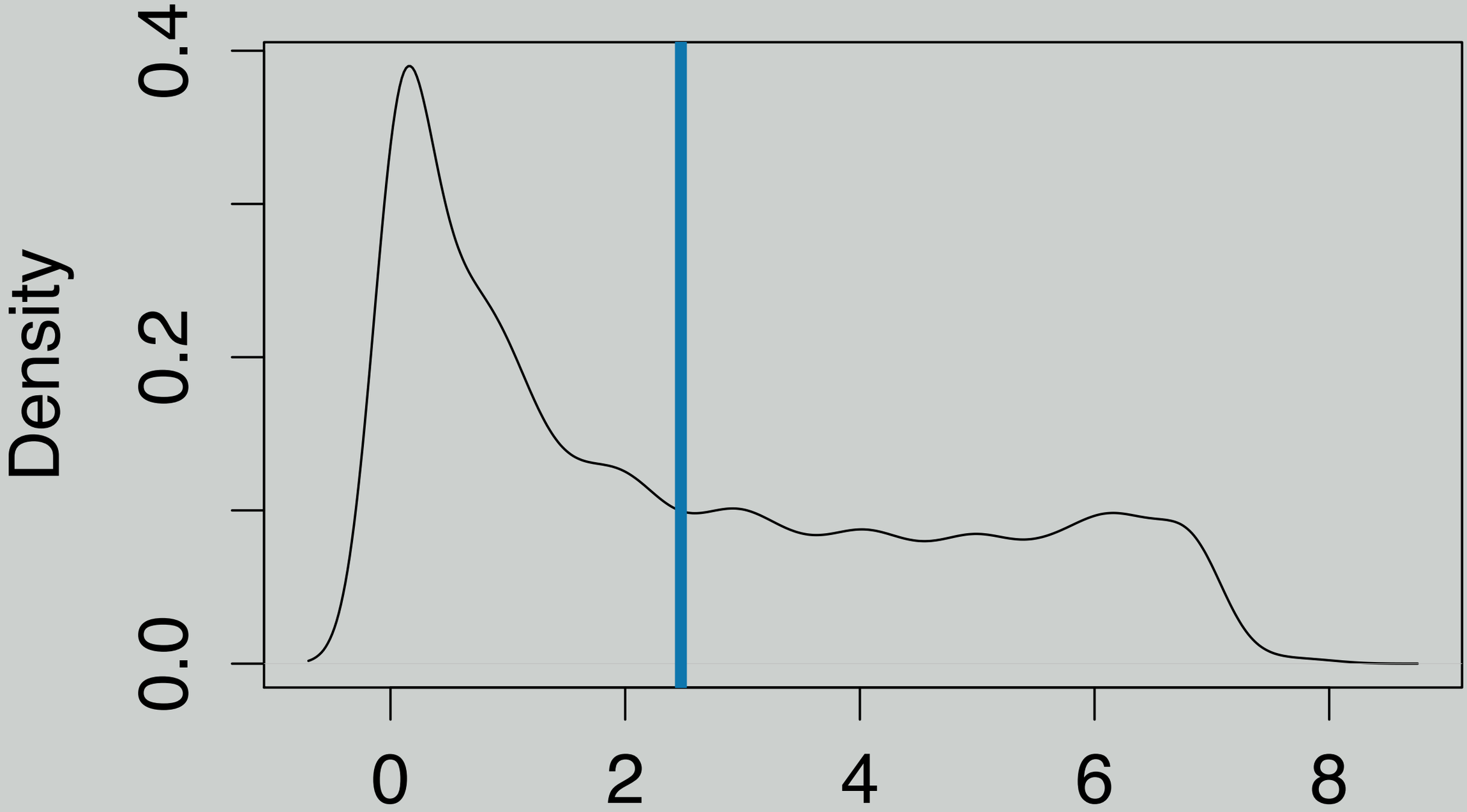
NEXT

Analysis

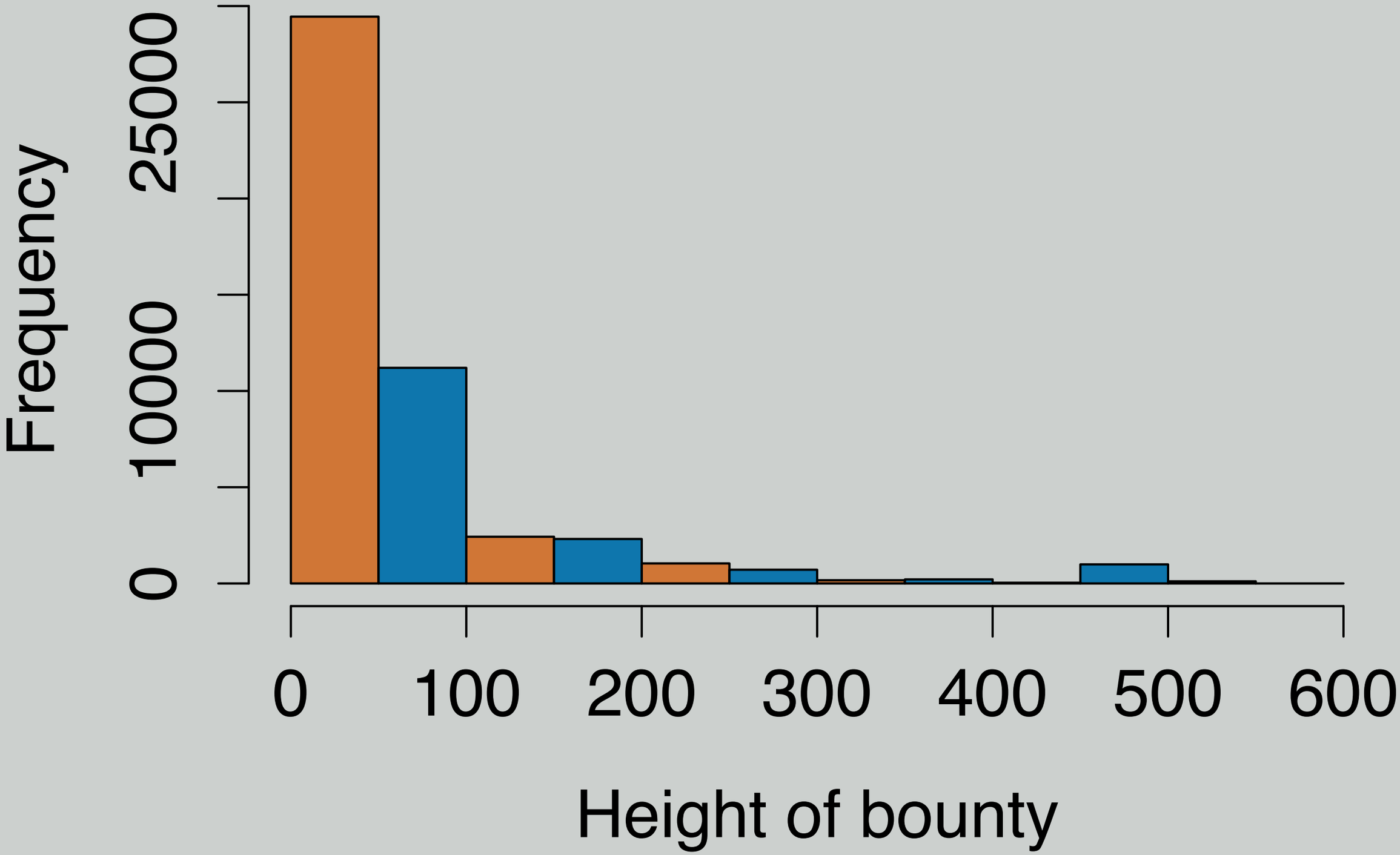
Statistics

# Questions with bounty	90,302
Average amount of bounty	90
Most bounties on a question	13
“Interesting tags”	code-reuse(15), vs2013-update-4(8), rgba(15)
Unsuccessful bounties	1/3

Histogram: Answer speed



Histogram: Bounty height



WHAT WE DID

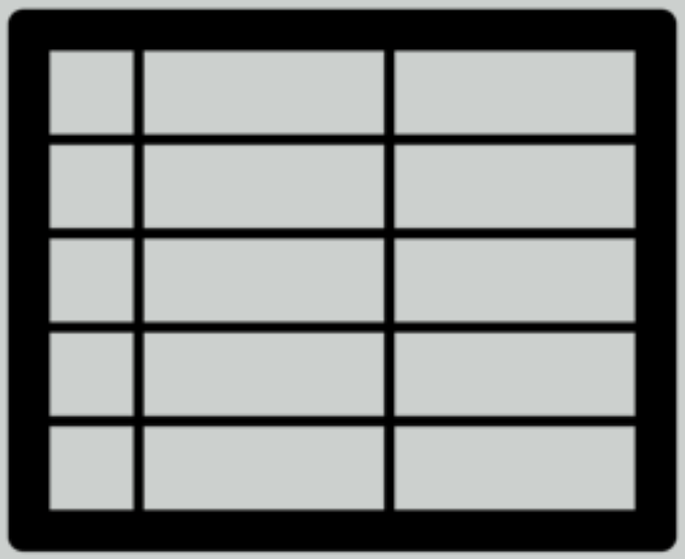
Created basic statistics

Found features that influence response time

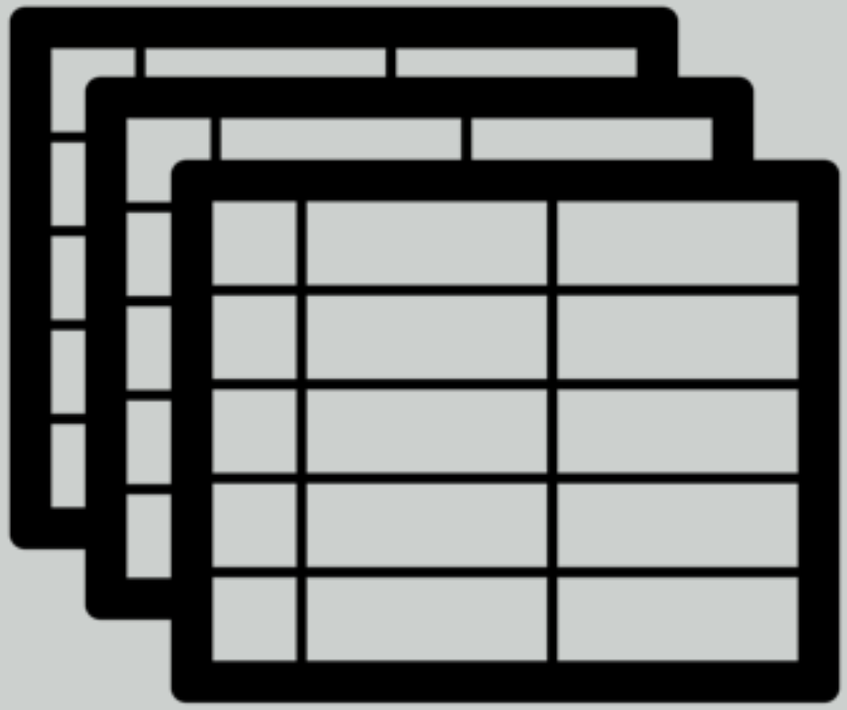
NEXT

Prediction

Prediction setup



Features



HANA / MySQL

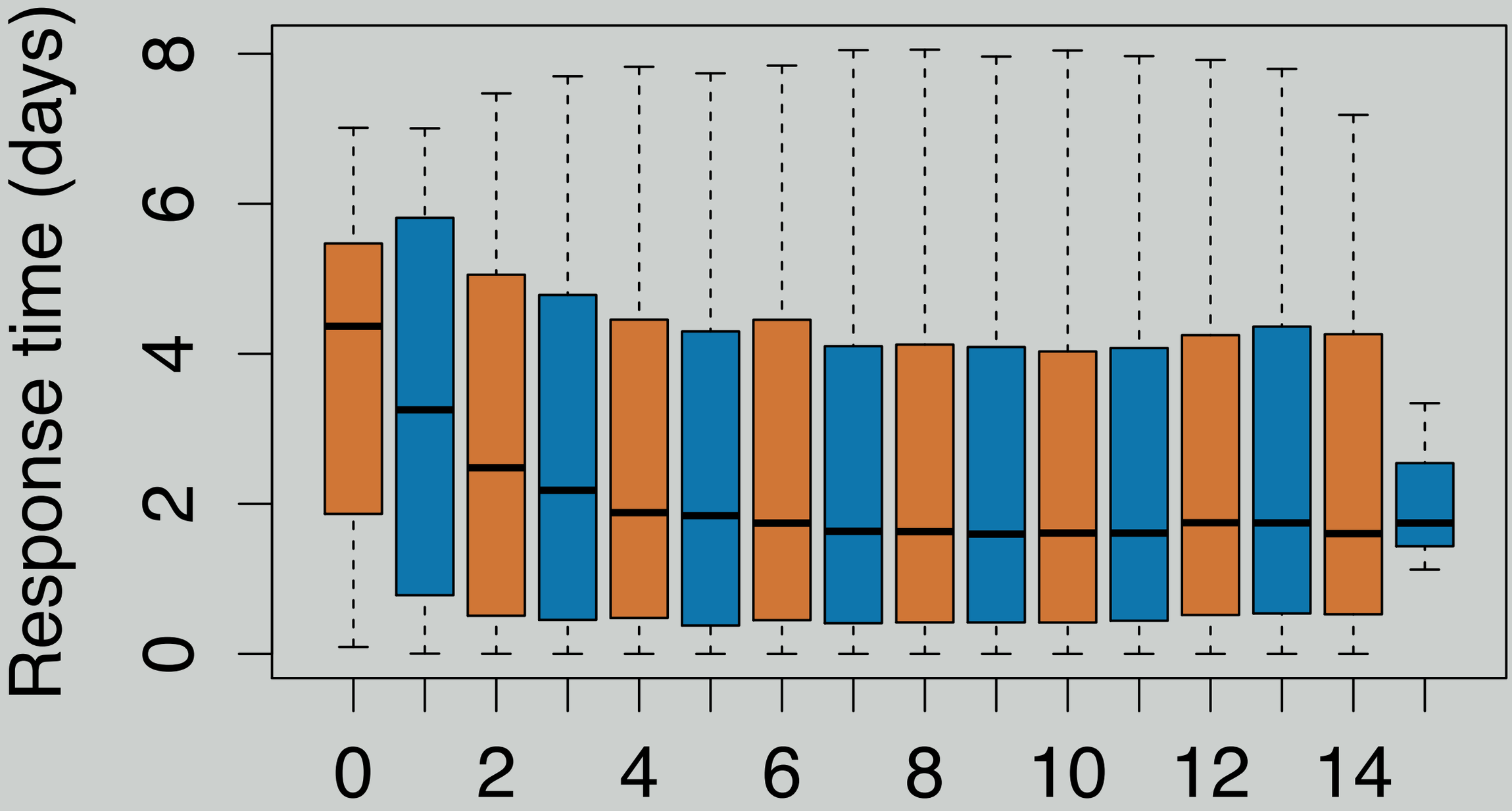
SVM with RBF

Prototype / Prediction

Feature categories

- Text features
- Shallow linguistic features
- Comment features
- Tag features
- Bounty features
- Topic features

Number of code lines



Feature categories

- Text features
  - # of code lines
  - # of code snippets
  - # of images
  - # of self references
  - # of active verbs
  - title length
  - ends with question mark
  - starts with question word
- Shallow linguistic features
- Comment features
- Tag features
- Bounty features
- Topic features

Ponzanelli et al. *Improving Low Quality Stack Overflow Post Detection*

Feature categories

positive
negative
influence

- Text features
- Shallow linguistic features
- Comment features
- Tag features
- Bounty features
- Topic features

- # of code lines
- # of code snippets
- # of images
- # of self references
- # of active verbs
- title length
- ends with question mark
- starts with question word

Ponzanelli et al. *Improving Low Quality Stack Overflow Post Detection*

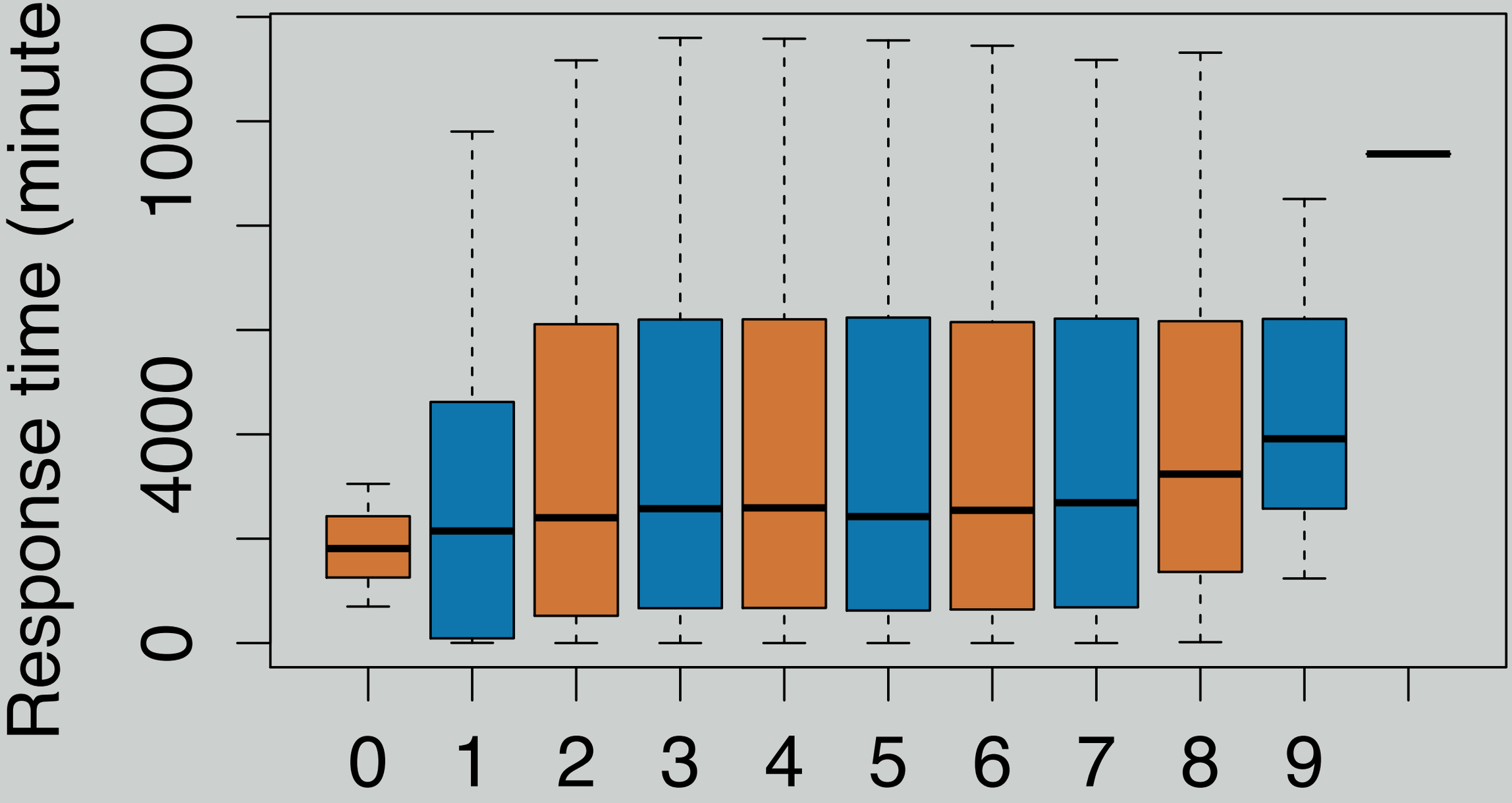
Feature categories

- Text features
- Shallow linguistic features
- Comment features
- Tag features
- Bounty features
- Topic features

- # of words
- avg. characters / word
- avg. words / sentence
- Automatic readability idx
- Coleman Liau idx
- Gunning fog idx
- Flesch reading ease

Gkotsis et al. *It's all in the content: state of the art best answer prediction based on discretisation of shallow linguistic features*

Flesch−Reading−Ease



Feature categories

positive
negative
influence

- Text features
- Shallow linguistic features
- Comment features
- Tag features
- Bounty features
- Topic features

# of words

avg. characters / word

avg. words / sentence

Automatic readability idx

Coleman Liau idx

Gunning fog idx

Flesch reading ease

Gkotsis et al. *It's all in the content: state of the art best answer prediction based on discretisation of shallow linguistic features*

Feature categories

- Text features
- Shallow linguistic features
- Comment features
- Tag features
- Bounty features
- Topic features

- # of comments
- # chars in comments
- avg. chars in comment

Feature categories

positive
negative
influence

- Text features
- Shallow linguistic features
- Comment features
- Tag features
- Bounty features
- Topic features

# of comments

chars in comments

avg. chars in comment

Feature categories

- Text features
- Shallow linguistic features
- Comment features
- Tag features
- Bounty features
- Topic features

popularity of tags

specificity of tags

# of popular tags

# of subscribers for tags

# of responsive subscribers

min subscribers

max subscribers

Bhat et al. *Min(e)d Your Tags: Analysis of Question Response Time in StackOverflow*

Feature categories

positive
negative
influence

- Text features
- Shallow linguistic features
- Comment features
- Tag features
- Bounty features
- Topic features

popularity of tags

specificity of tags

# of popular tags

# of subscribers for tags

# of responsive subscribers

min subscribers

max subscribers

Bhat et al. *Min(e)d Your Tags: Analysis of Question Response Time in StackOverflow*

Feature categories

- Text features
- Shallow linguistic features
- Comment features
- Tag features
- Bounty features
- Topic features

height of bounty

# of answers

# of up votes

# of down votes

time till creation

# of other bounties

view count

Feature categories

positive
negative
influence

Text features

Shallow linguistic features

Comment features

Tag features

Bounty features

Topic features

height of bounty

# of answers

# of up votes

# of down votes

time till creation

# of other bounties

view count

Feature categories

Text features

Shallow linguistic features

Comment features

Tag features

Bounty features

Topic features

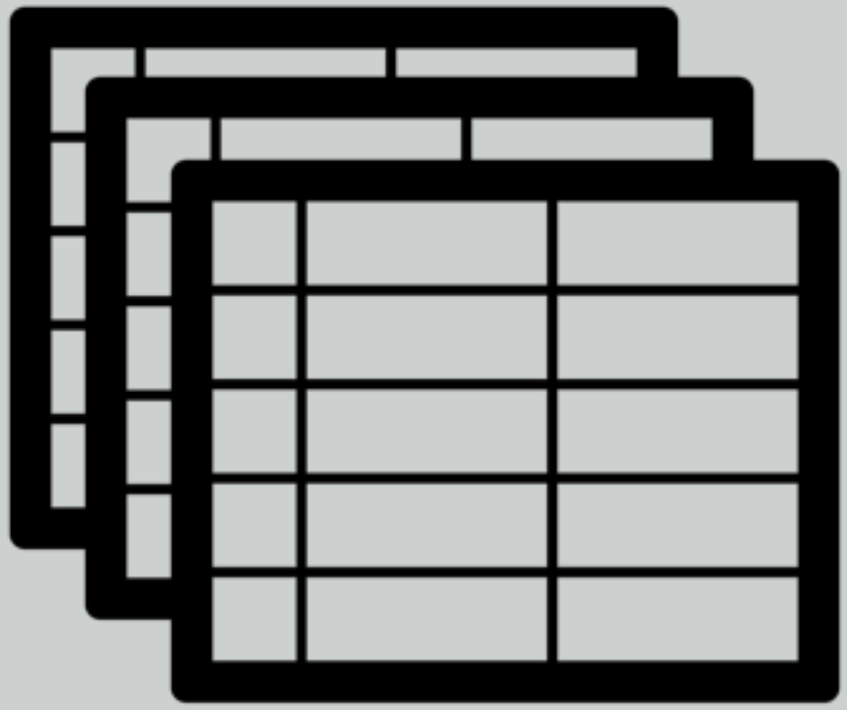
best matching topic

[feature for every topic]

[feature for every topic without NP]

Allamanis et al. *Why, when, and what: Analyzing Stack Overflow questions by topic, type, and code*

Topic features



HANA / MySQL

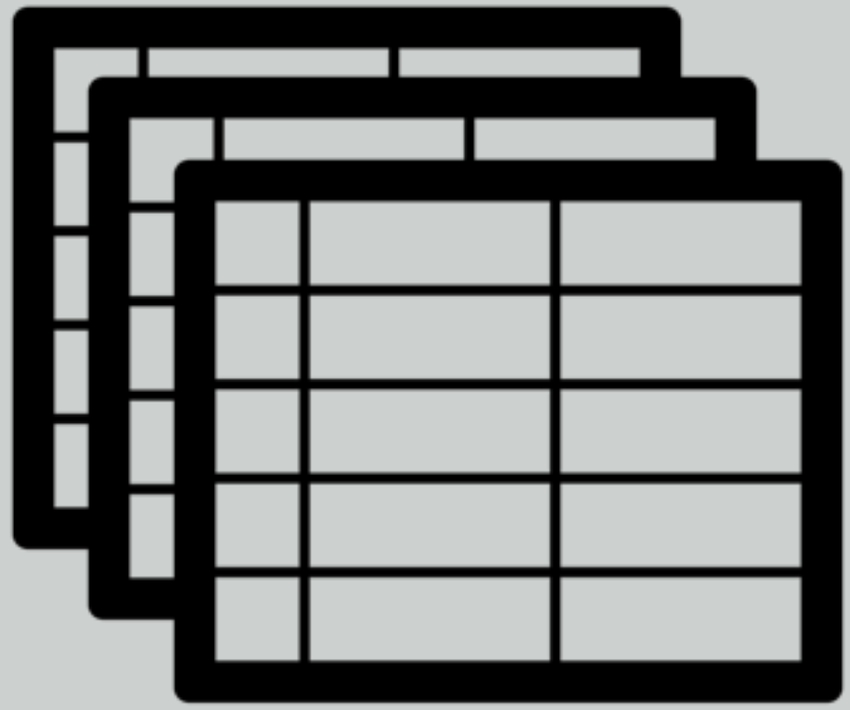
Latent Dirichlet Allocation

#1	facebook, post, upload, grid, posts, drag, share, drop, uploaded, condition
#2	api, google, map, location, docs, maps, engine, apis, chat, documentation
#3	access, login, authentication, token, password, security, widget, username, permissions, credentials
…	

VP Topic features

POS Tagger

Latent Dirichlet Allocation



HANA / MySQL

#1	doesn’t work, work, try, didn’t, won’t, isn’t, wrong
#2	hope, make, understand, give, to make, work, read, explain, check
#3	create, to create, is creating, call, can create, add, want to create
	…

Feature categories

Text features

Shallow linguistic features

Comment features

Tag features

Bounty features

Topic features

best matching topic

[feature for every topic]

[feature for every topic without NP]

Allamanis et al. *Why, when, and what: Analyzing Stack Overflow questions by topic, type, and code*

WHAT WE DID

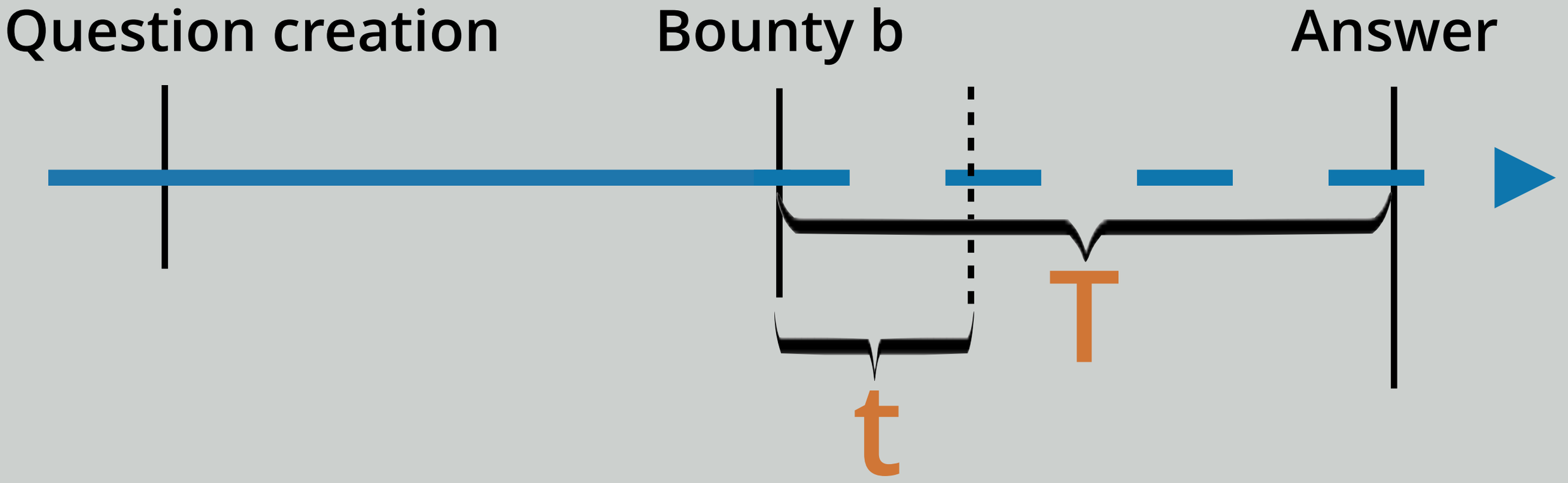
Established topics as features

Evaluated different feature categories

NEXT

Prediction Results

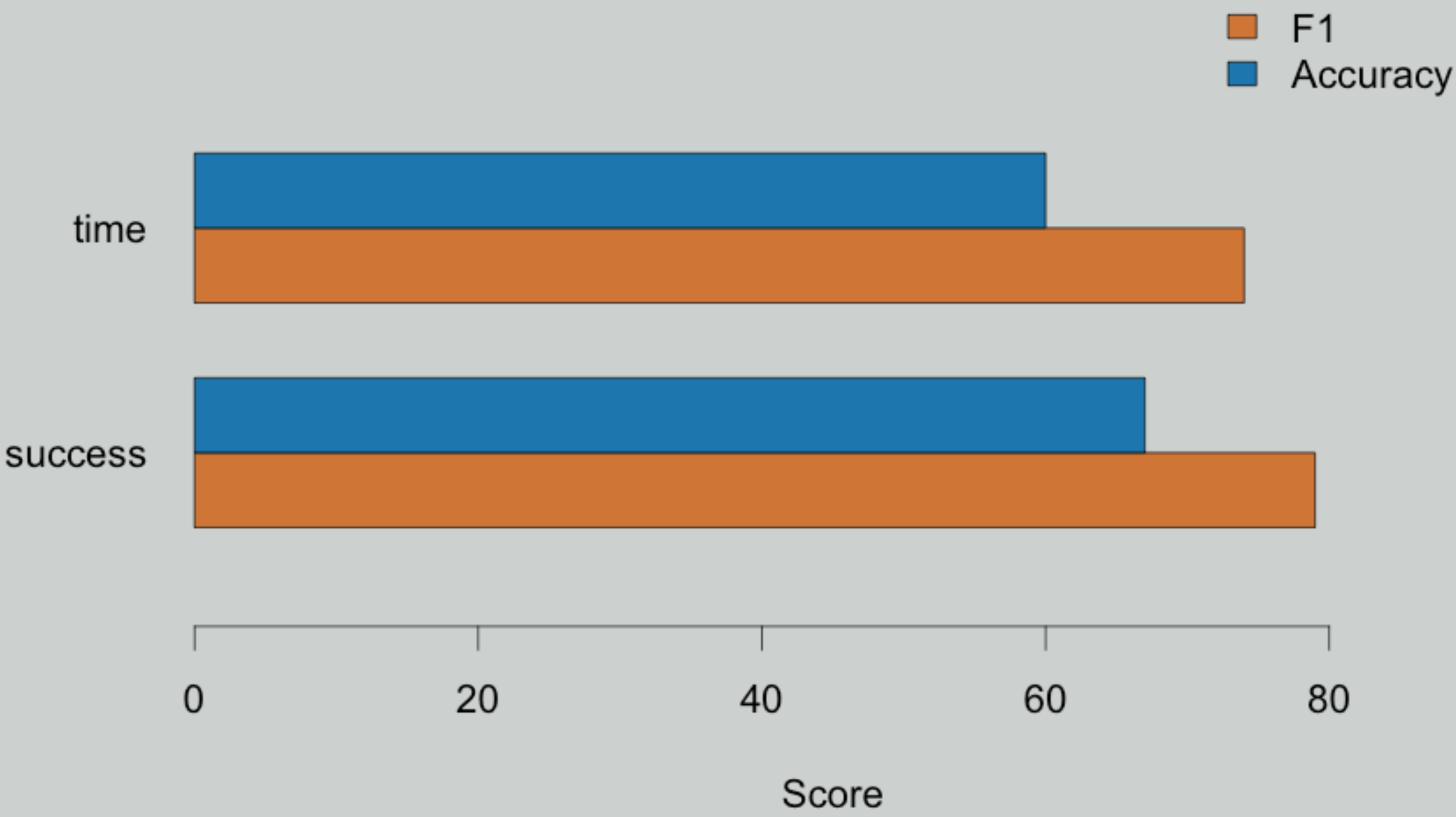
Prediction task



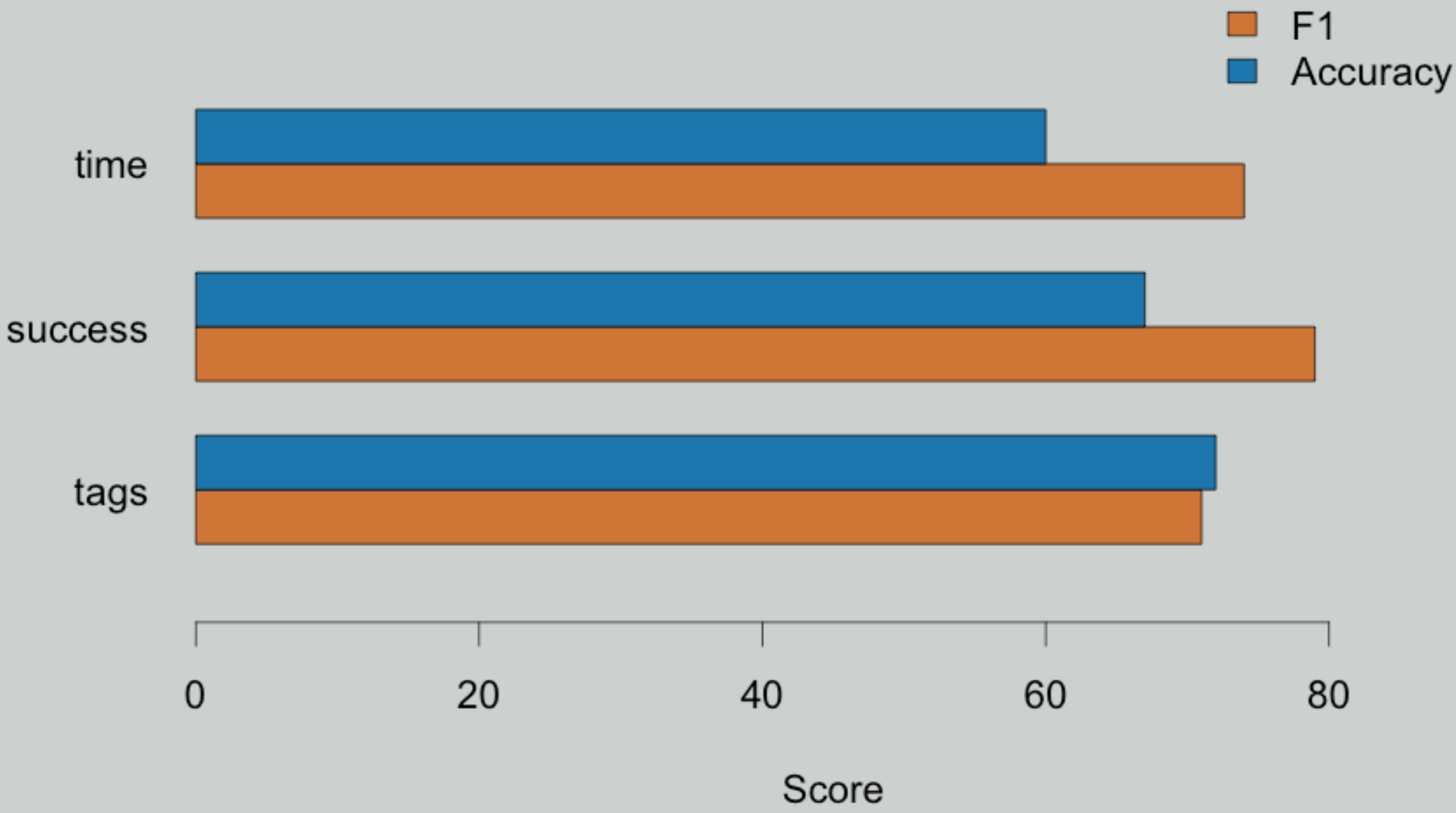
success(b) Will the bounty be successful?

$T \overset{?}{\leq} t$ Will the question be answered within the timespan t (e.g. 2,5 days) of setting a bounty?

Prediction scores



Comparison with 'Min(e)d your tags'



Insights

- Features for regular questions don’t work as well as hoped
- Bounty questions are different in nature:
  - very specific topics
  - very difficult
  - long discussions

NEXT

Prototype

Stackoverflow Bounty Predictions

Stackoverflow Question URL or Question Id

Submit

Results for Question 6824681

Predicition

The Bounty generates at least one answer: False

An answer is posted within 2.5 days: True

commentFeatures

- num_comments : 0
- comment_len : 0
- avg_comment_len : 0

textFeatures

- num_active_verb : 0
- num_images : 0
- begin_que_word : False
- num_code_snippet : 2
- title_len : 31
- body_len : 189
- end_que_mark : True
- num_selfref : 3
- postId : 6824681
- code_len : 41

WHAT WE DID

Store all trained instances

Webserver for live prediction

NEXT

Future

Future

1. Finish the prototype
2. Train LDA on all questions
3. Topic modeling for tags
4. Evaluation of prediction models
5. Be creative, evaluate new features (e.g. user features)

StackOverflow: A journey of bounty hunters

Social Media Mining WS 14/15

Tom Bocklisch & Tom Herold

Research Questions

Data Cleansing

0. Understand the data and eliminate faulty data.

Analysis

1. What are the intrinsic factors and signals that are likely to influence a bounty's response time?

Prediction

2. Can we predict a successful claim of a bounty?
3. Can we predict whether an answer will be given in a certain timespan?