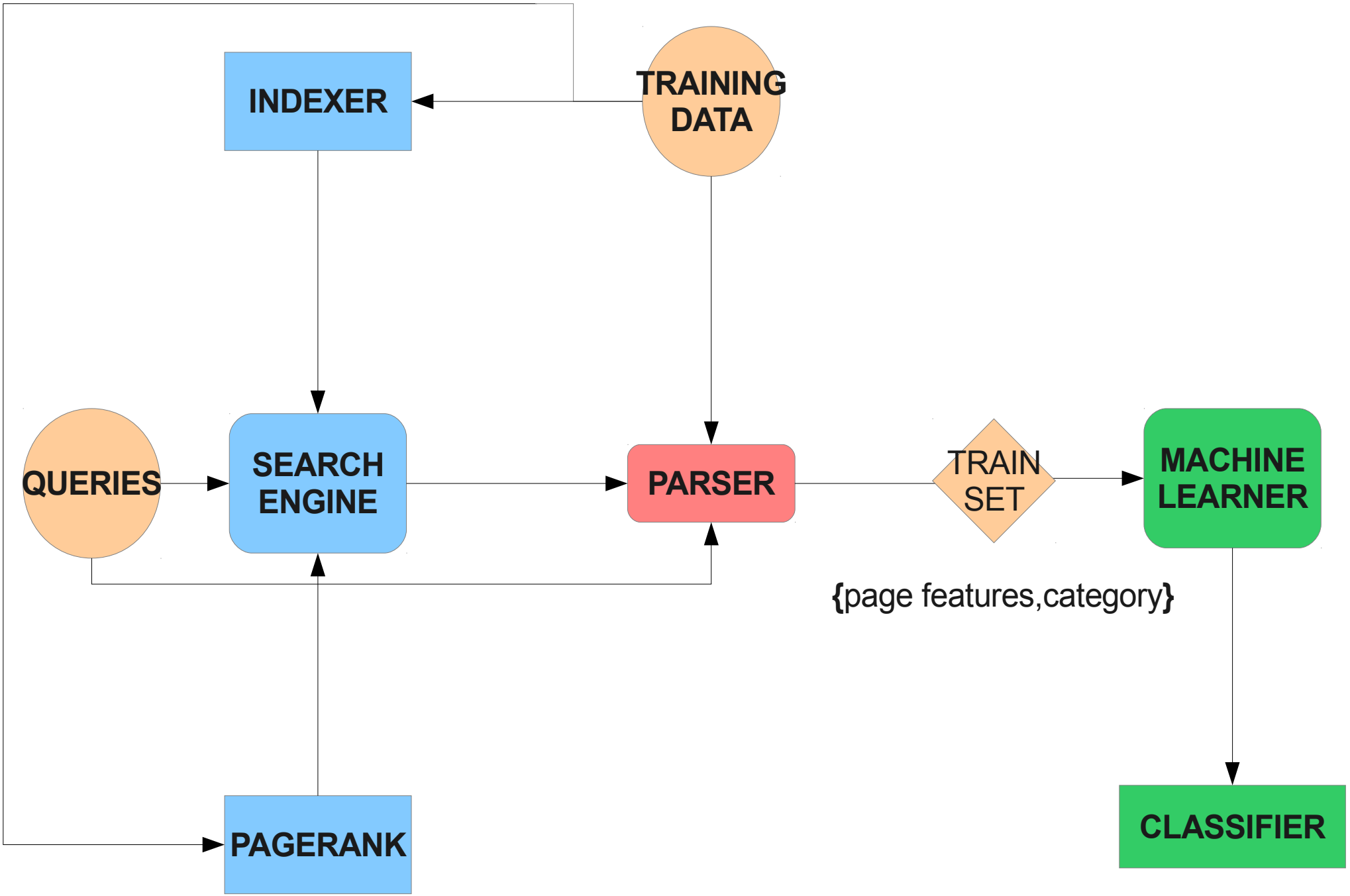
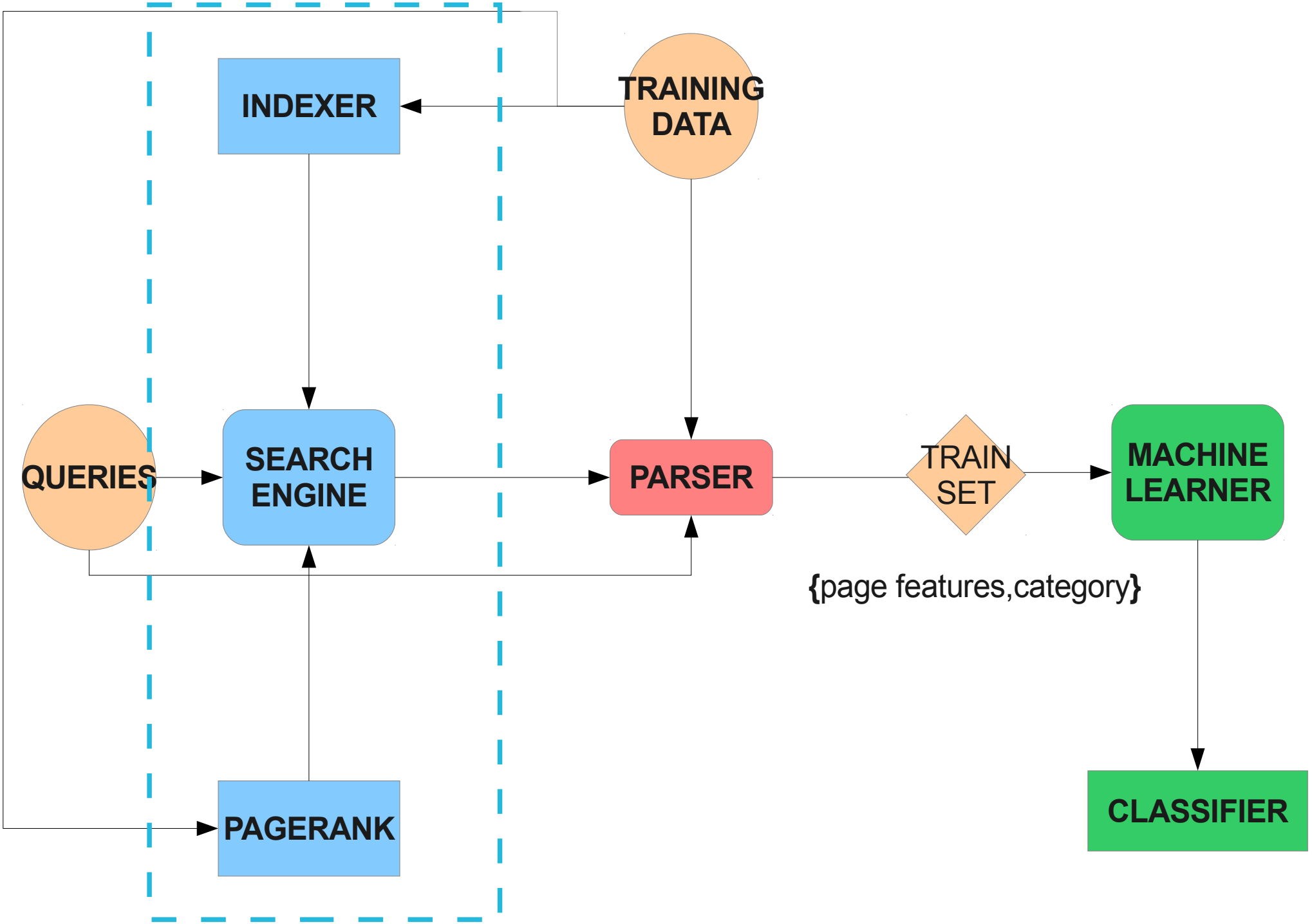
SEARCH ENGINE HEURISTICS VS MACHINE LEARNING

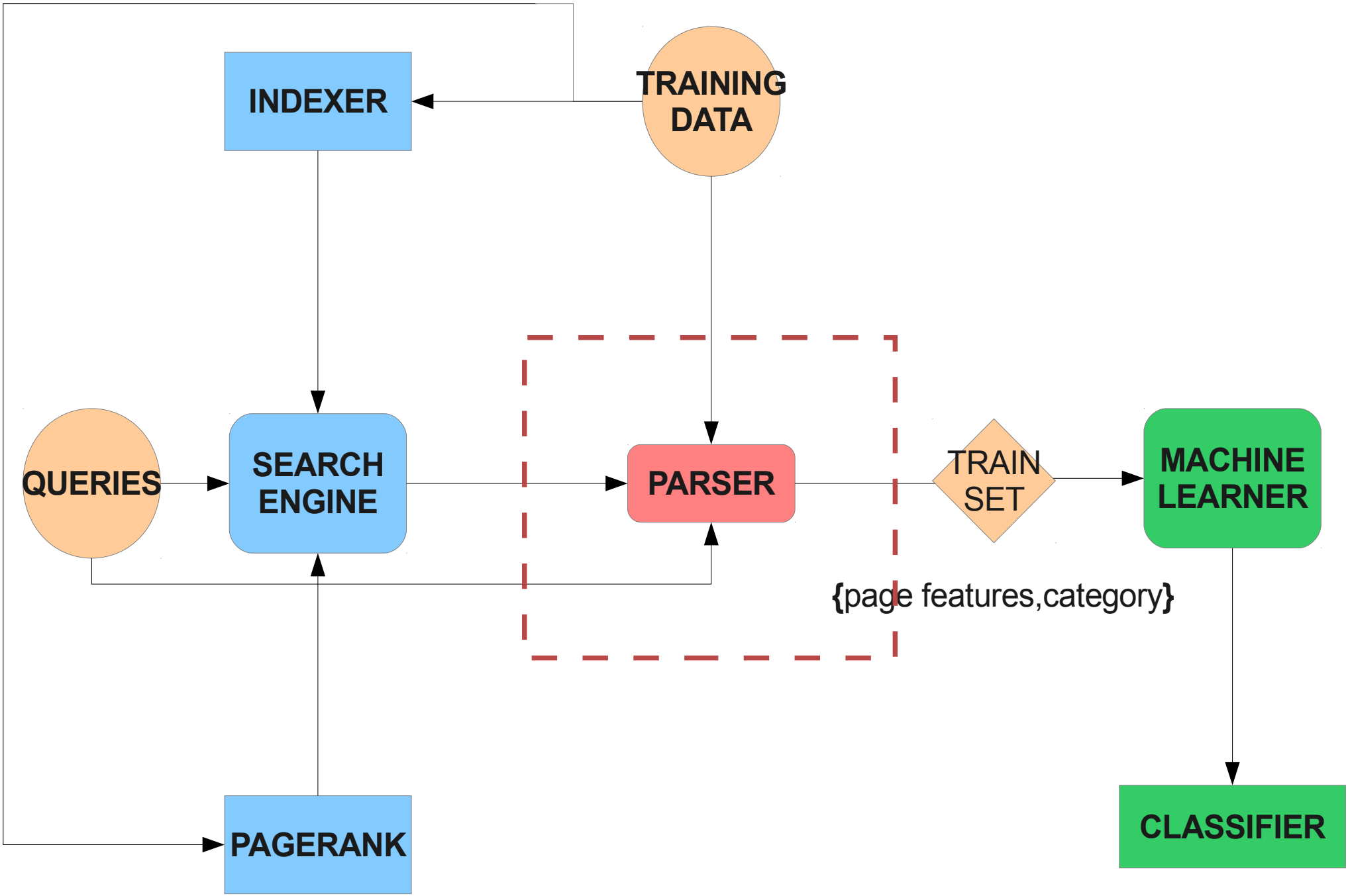
Karina Palyutina

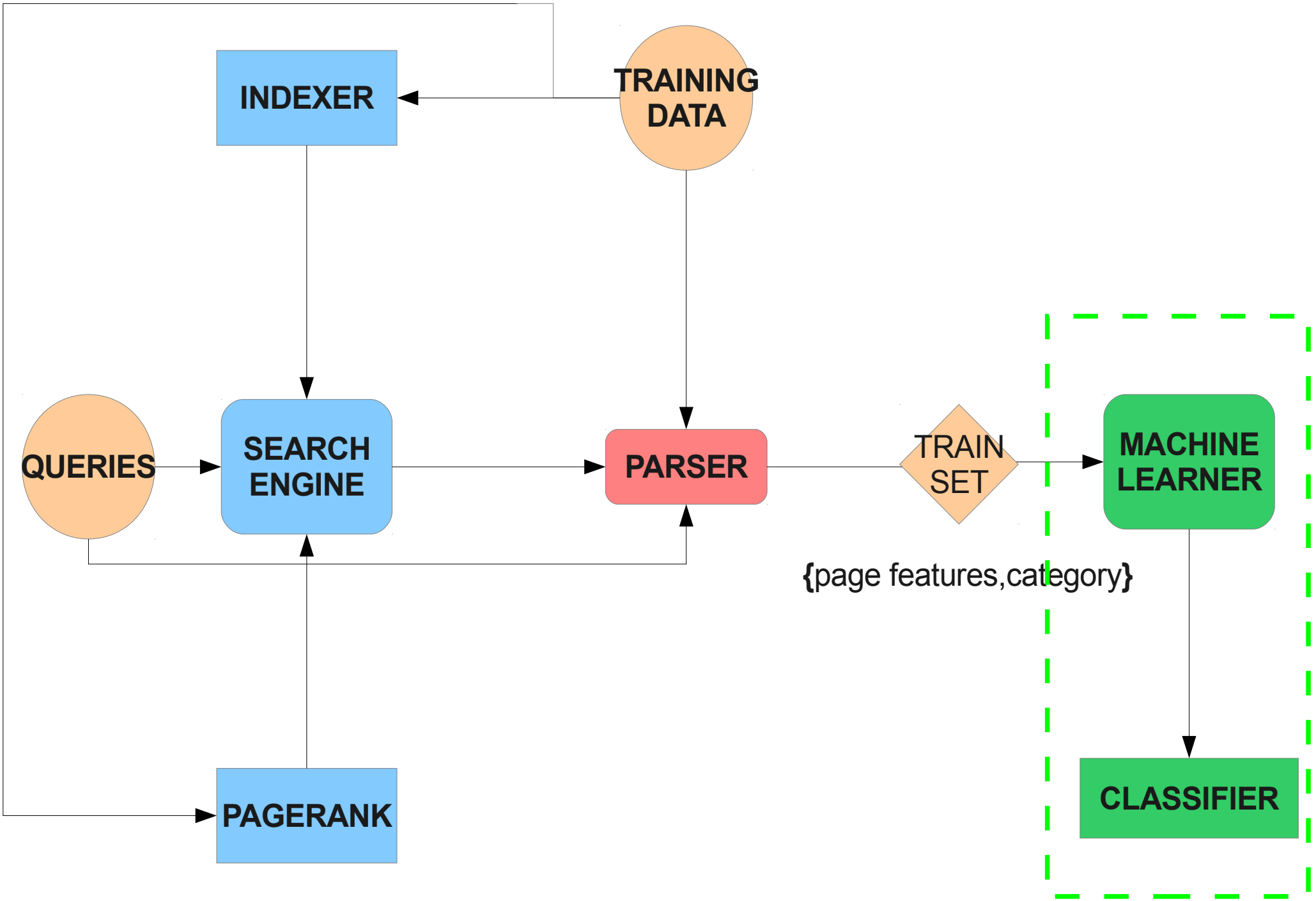
GOAL

“Find how effective machine learning techniques are in detecting heuristics”


INDEXER
TRAINING DATA
QUERIES
SEARCH ENGINE
PARSER
TRAIN SET
MACHINE LEARNER
{page features,category}
PAGERANK
CLASSIFIER


INDEXER
TRAINING DATA
QUERIES
SEARCH ENGINE
PARSER
TRAIN SET
MACHINE LEARNER
{page features,category}
PAGERANK
CLASSIFIER


INDEXER
TRAINING DATA
QUERIES
SEARCH ENGINE
PARSER
TRAIN SET
MACHINE LEARNER
{page features,category}
PAGERANK
CLASSIFIER


INDEXER
TRAINING DATA
QUERIES
SEARCH ENGINE
PARSER
TRAIN SET
MACHINE LEARNER
{page features,category}
PAGERANK
CLASSIFIER

Naive Bayes

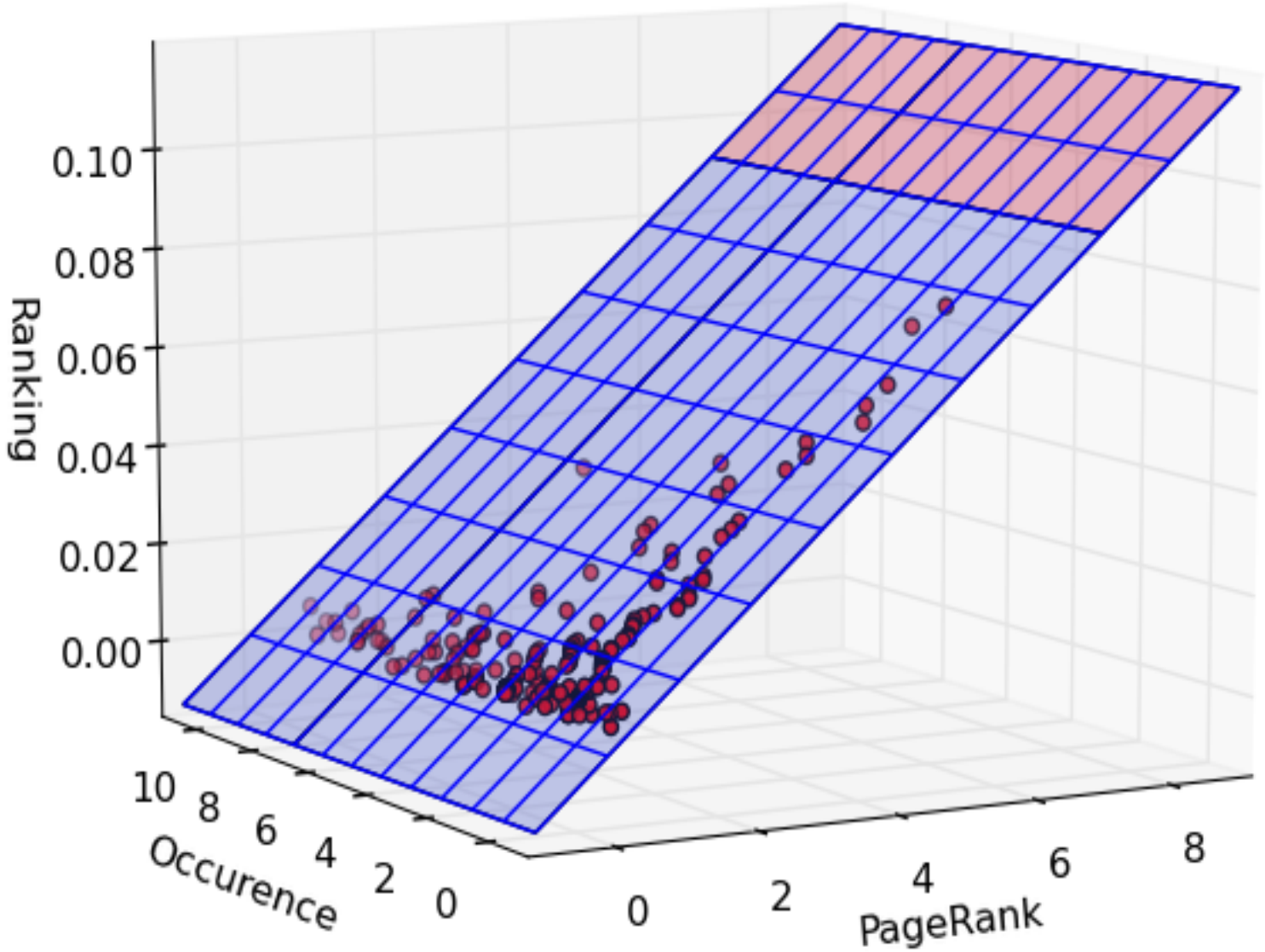
$$p(C|F_1, \ldots, F_n) = \frac{p(C)\, p(F_1, \ldots, F_n|C)}{p(F_1, \ldots, F_n)}$$

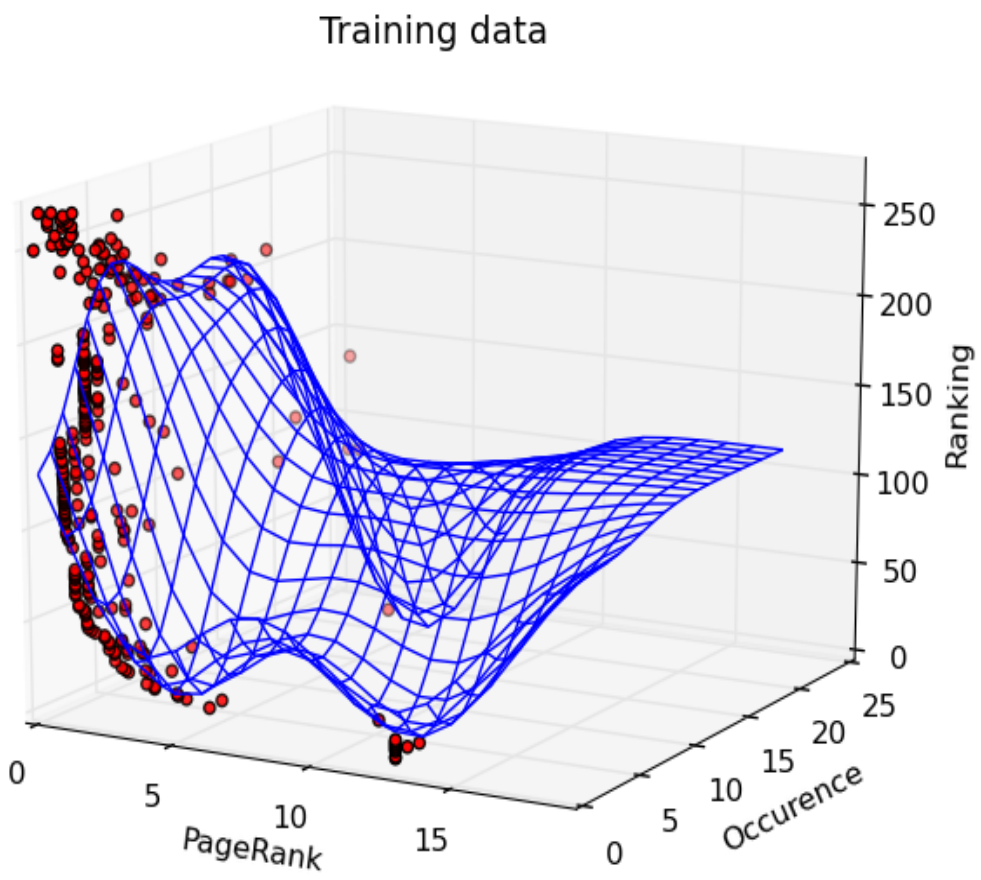
99% Static Features

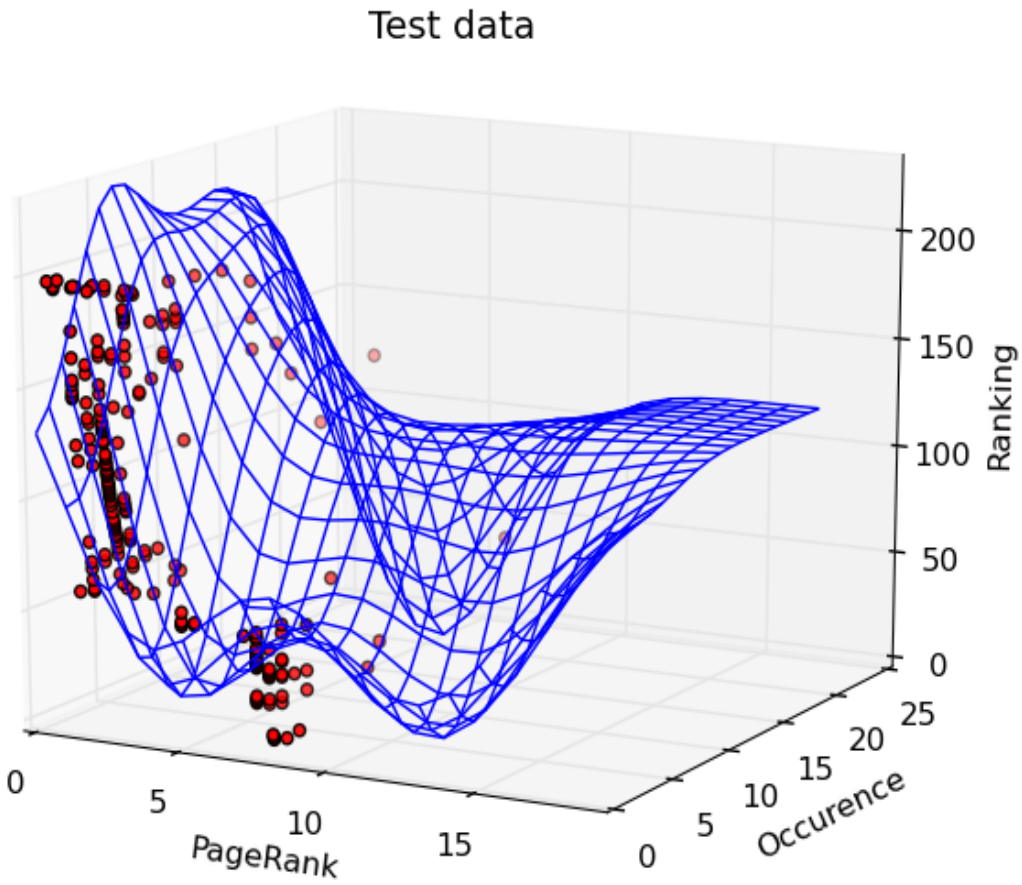
55% PageRank

3000 Pages

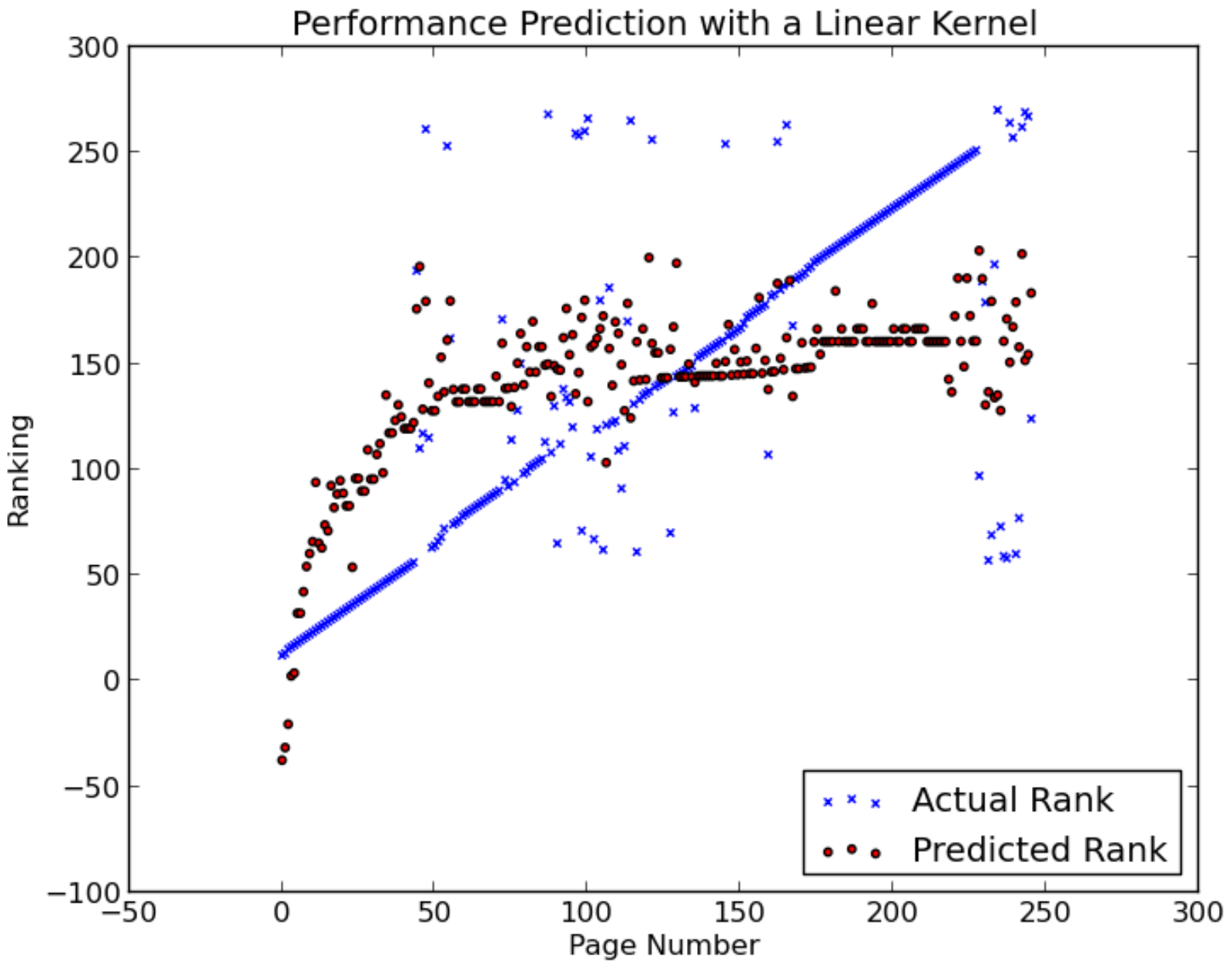
Support Vector Regression


Ranking
0.10
0.08
0.06
0.04
0.02
0.00
Occurence
10
8
6
4
2
0
PageRank
0
2
4
6
8


Training data
Ranking
PageRank
Occurence
0
5
10
15
0
5
10
15
20
25
0
50
100
150
200
250


Test data
Ranking
PageRank
Occurence
0
5
10
15
0
5
10
15
20
25
0
50
100
150
200

Evaluation


Performance Prediction with a Linear Kernel
Ranking
Page Number
Actual Rank
Predicted Rank
300
250
200
150
100
50
0
−50
−100
−50
0
50
100
150
200
250
300

MSE ~ 50

Score or Rank?

Kernel	Kernel function
Linear	$x \circ y + c$
Polynomial	$(ax \circ y + c)^d$
Gaussian	$exp(-\gamma \, \|\|x - y\|\|^2)$
Sigmoid	$tanh(ax \circ y + c)$

Score Function
Sum
Product
Exponential
Weighted Sum
Arbitrary
Polynomial