Supervised by Mr.Sampath Deegalla

Feasibility of using Machine Learning to Access Control in Squid Proxy Server

- Kanchana Ihalagedara
- Rajitha Kithuldeniya
- Supun weerasekara

Internet in Educational Institutes

- Mainly for educational purposes.
- What happens if users priority is not the intended purpose.
  - Network congestions
  - Wastage of resources
  - Affects individual user performance negatively

Blocking Web Sites in Proxy Server

- Squid ACLs - Text file of blacklists
- SquidGuard - External databases
- DansGuardian - Content filter

World Wide Web is Growing

672,985,183 - 2013

968,882,453 - 2014

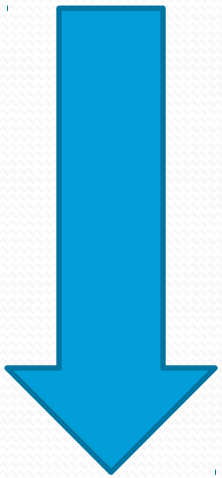
295,897,270

From www.internetlivestats.com

- Manually blacklisting web sites is impossible
- Related products are not updated with the growing web

Dynamic automated method

- Automated web classification is required



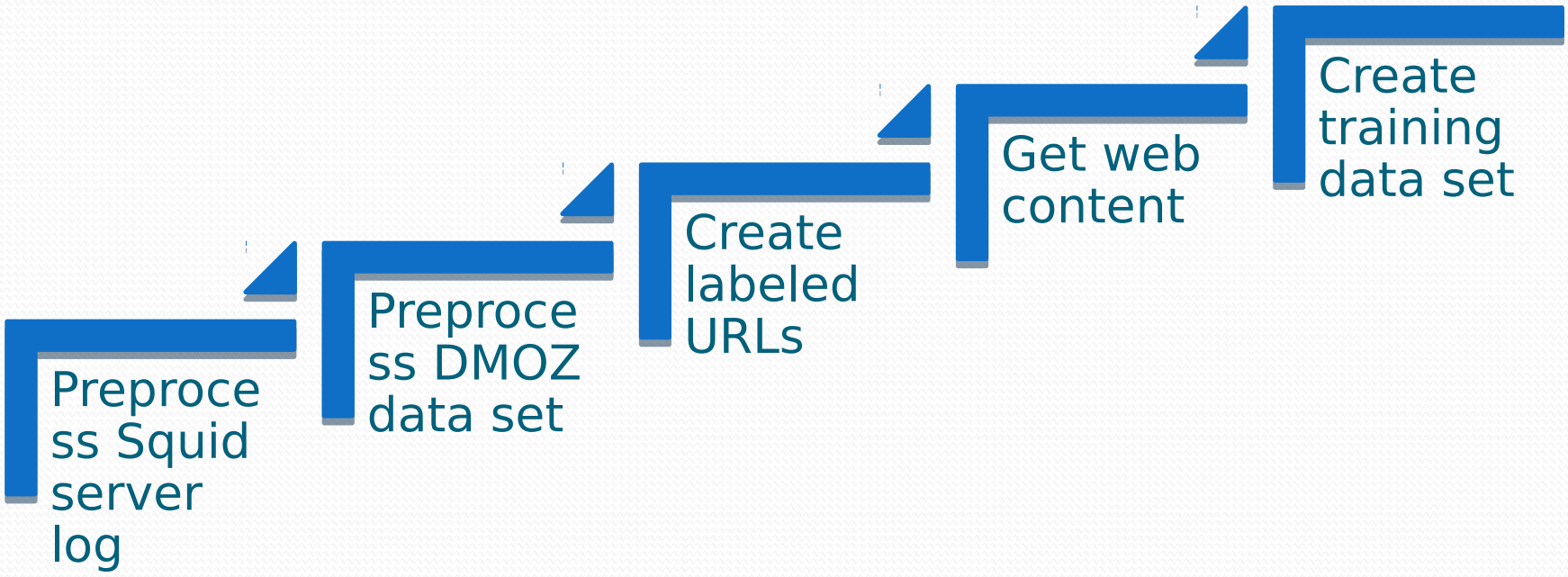
- Machine Learning is used in automated web classification

Over View of Our Solution

Machine Learning in Web Classification

- Several web classification researches can be found
- Frequently used algorithms
  - Naïve Byes
  - Support vector machine
  - Nearest neighbor
- Classification requires a data set
- Set of URLs labeled as educational or non educational

Data Collection & Preprocessing



Model Creation & Testing

- Four models were created from WEKA(small data set)
  - Data set with two hundred records
  - 10

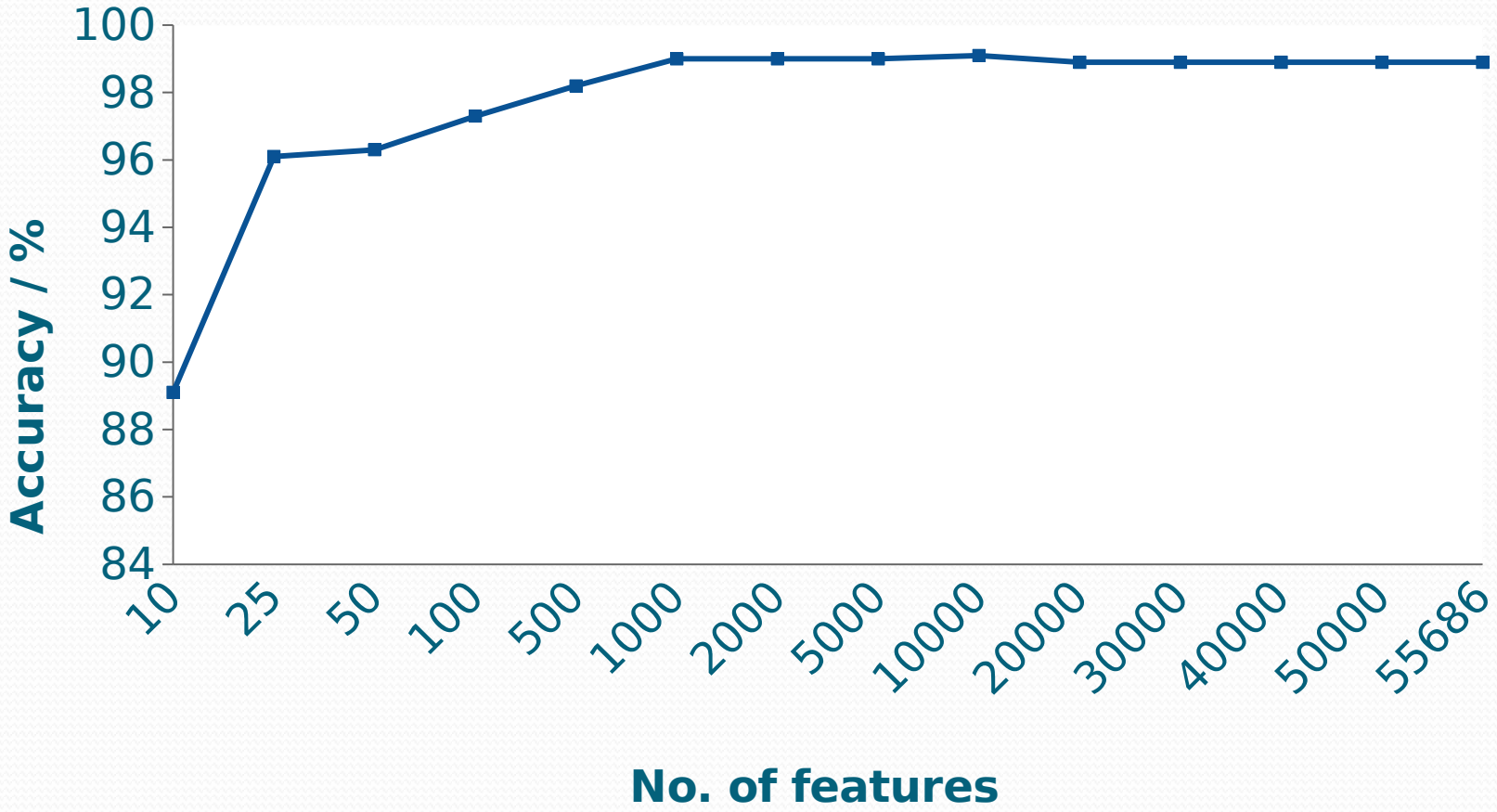
Algorithm	Accuracy(%)
PRISM	74.5
C4.5 (J48 in WEKA)	83.0
Naïve bayes	**95.0**
Support Vector Machines	**95.5**

Model Creation & Testing

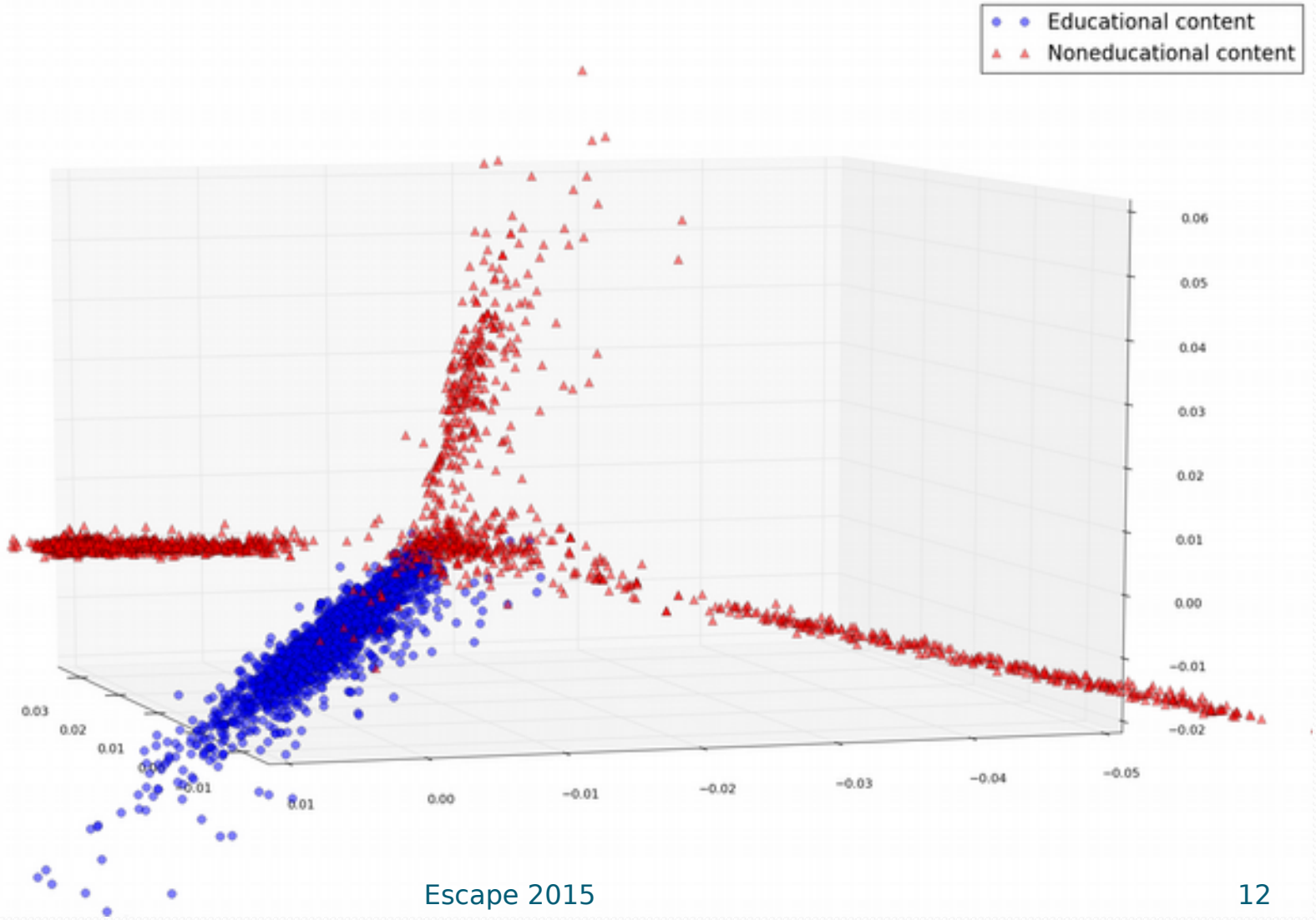
- Three models using Python (larger dataset)
  - Data set of 4000 records
  - Separate data set of 1000 records for Testing

Algorithm	Accuracy
Naïve Bayes multinomial	92.9%
SVC	77.5%
Linear SVC	98.9%

Feature Selection in Linear SVC



Principal Component Analysis



Future Work

- Consider more content (Meta data)
- Other Languages (Sinhala)
- Image processing can be added

Thank You!