CS6604 Project
Final Presentation

Ensemble Classification

Project Team:
Kannan, Vijayasarathy
Soundarapandian, Manikandan
Alabdulhadi, Mohammed
Hamid, Tania

Advisor:
Dr. Edward A. Fox

Project Client:
Yinlin Chen

Virginia Tech, Blacksburg
05/01/2014

VirginiaTech

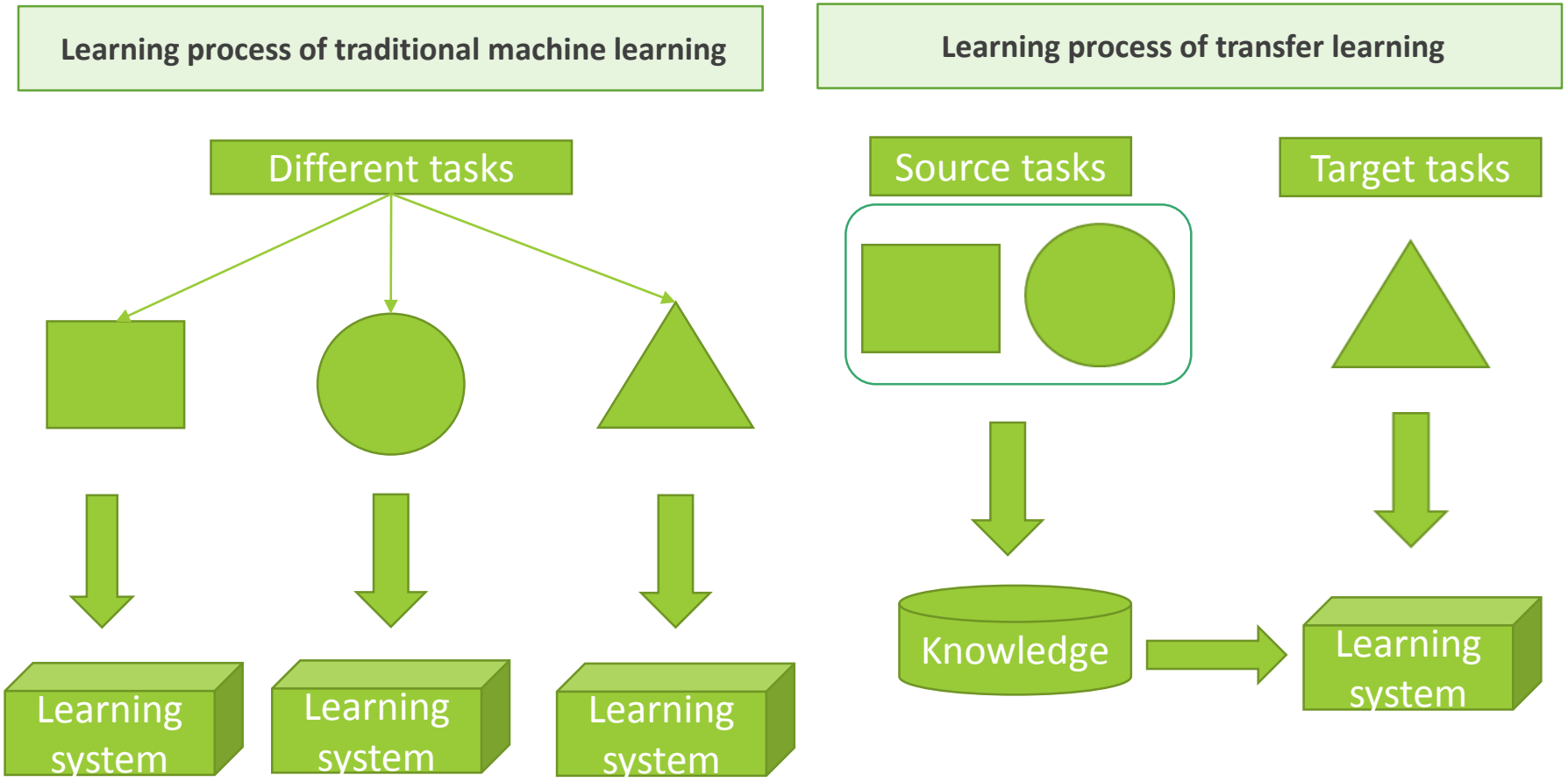
Outline

- Introduction
- Project big picture
- Workflow
- Tuning training data
- Multi-class vs. Single-class classification
- ACM taxonomy
- Methods and approaches
- Evaluation
- Challenges
- Lessons learned
- Future work
- Questions

Introduction

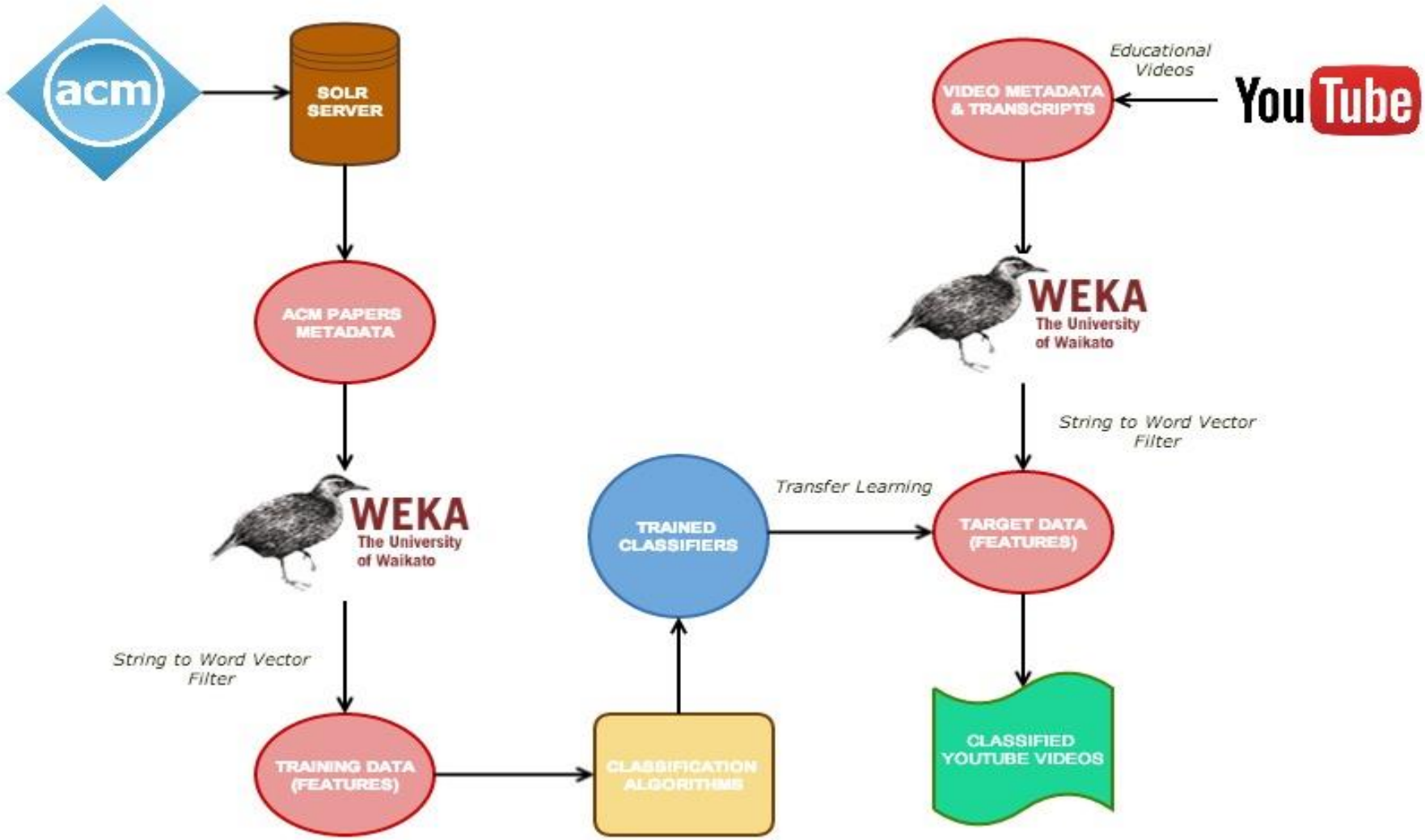
- Project objective:
  - Developing classifiers to aid in
    - Transfer learning
    - Classify educational resources for the Ensemble portal.
- Machine learning (Text classification)
- Transfer learning
  - Source data: 2012 ACM CCS
  - Target data: CS YouTube videos

Machine learning vs. Transfer Learning

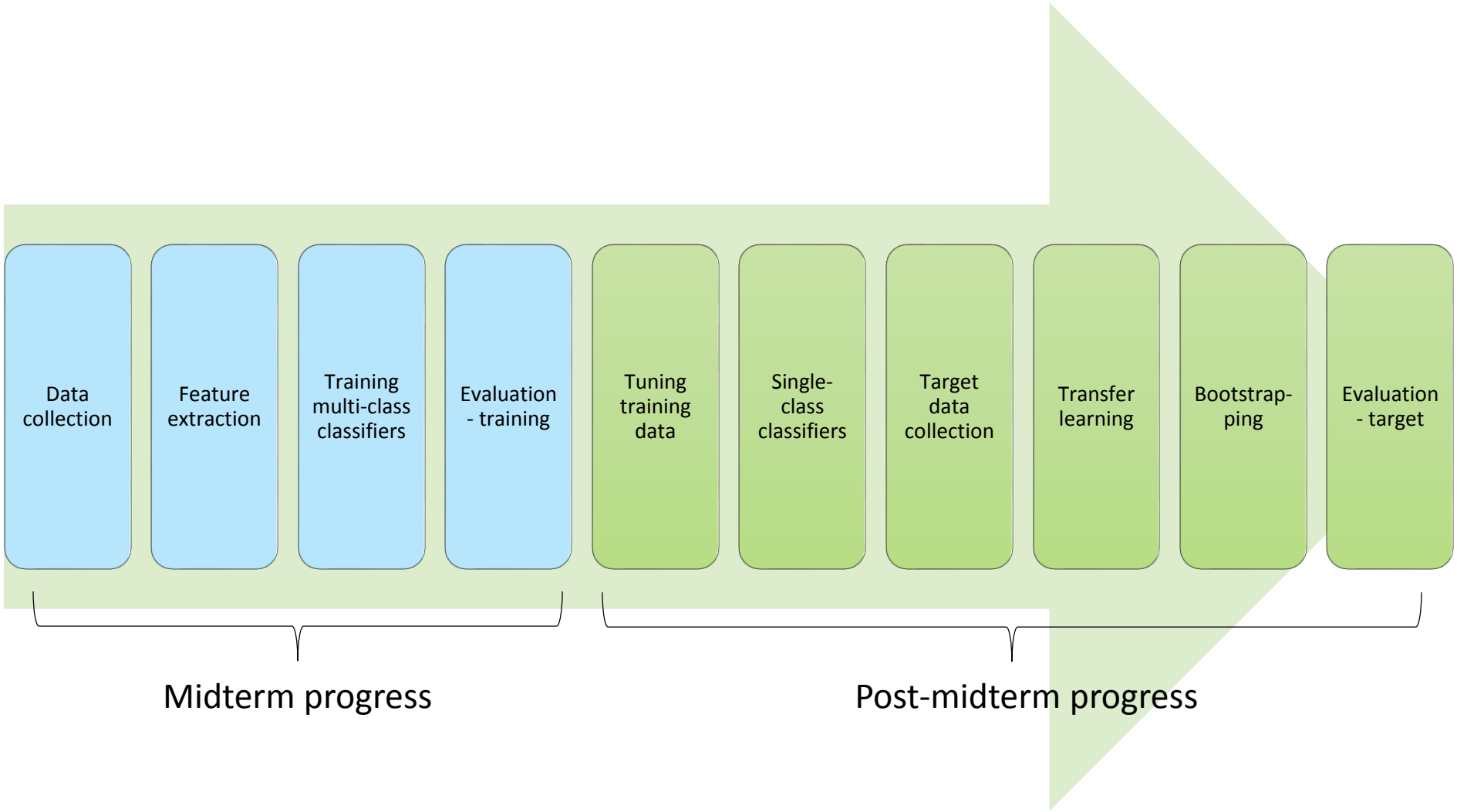


Adapted from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5288526&tag=1

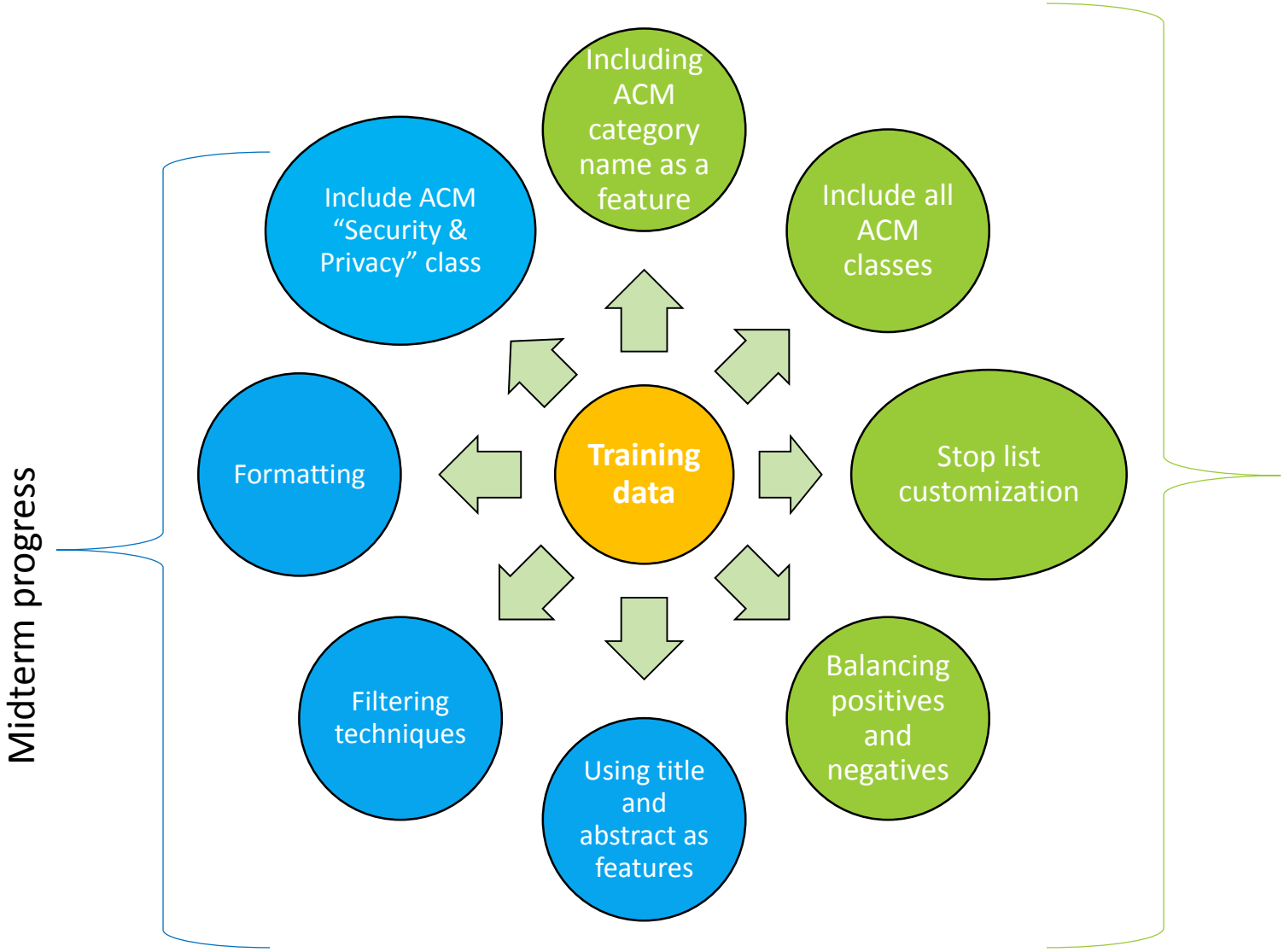
Big picture


acm
SOLR SERVER
ACM PAPERS METADATA
WEKA
The University of Waikato
String to Word Vector Filter
TRAINING DATA (FEATURES)
CLASSIFICATION ALGORITHMS
TRAINED CLASSIFIERS
Transfer Learning
Educational Videos
VIDEO METADATA & TRANSCRIPTS
YouTube
WEKA
The University of Waikato
String to Word Vector Filter
TARGET DATA (FEATURES)
CLASSIFIED YOUTUBE VIDEOS

Workflow


Data collection
Feature extraction
Training multi-class classifiers
Evaluation - training
Tuning training data
Single-class classifiers
Target data collection
Transfer learning
Bootstrap-ping
Evaluation - target
Midterm progress
Post-midterm progress

Tuning training data

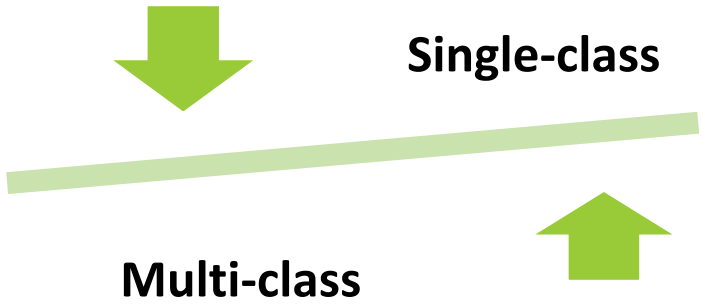


Multi-class vs. Single-class classification

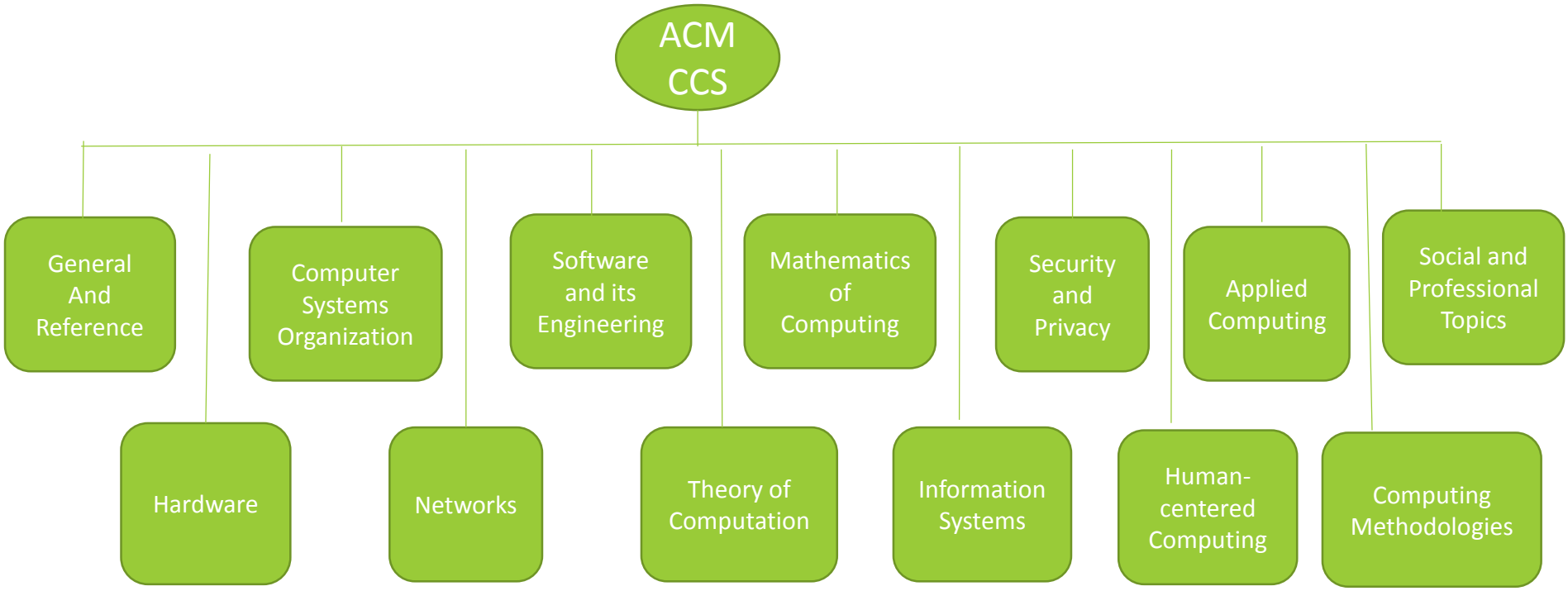
- **Multi-class classification**:
  - Each training point belongs to one of *N* different classes
  - Predict the class(es) to which a training point belongs to
  - 1 classifier

- **Single-class classification**:
  - Determine whether a training point belongs to a given class or not
  - *N* classifiers (one for each class)

  - Better accuracy and performance

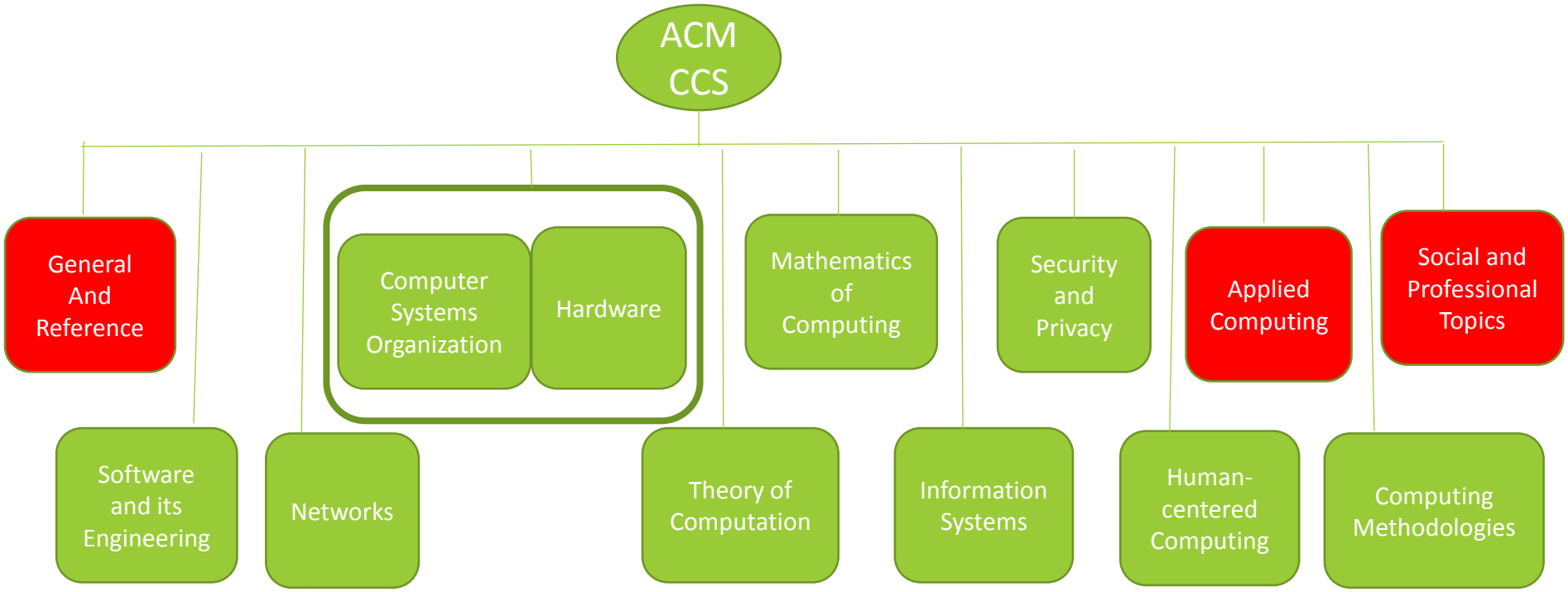


ACM taxonomy tree

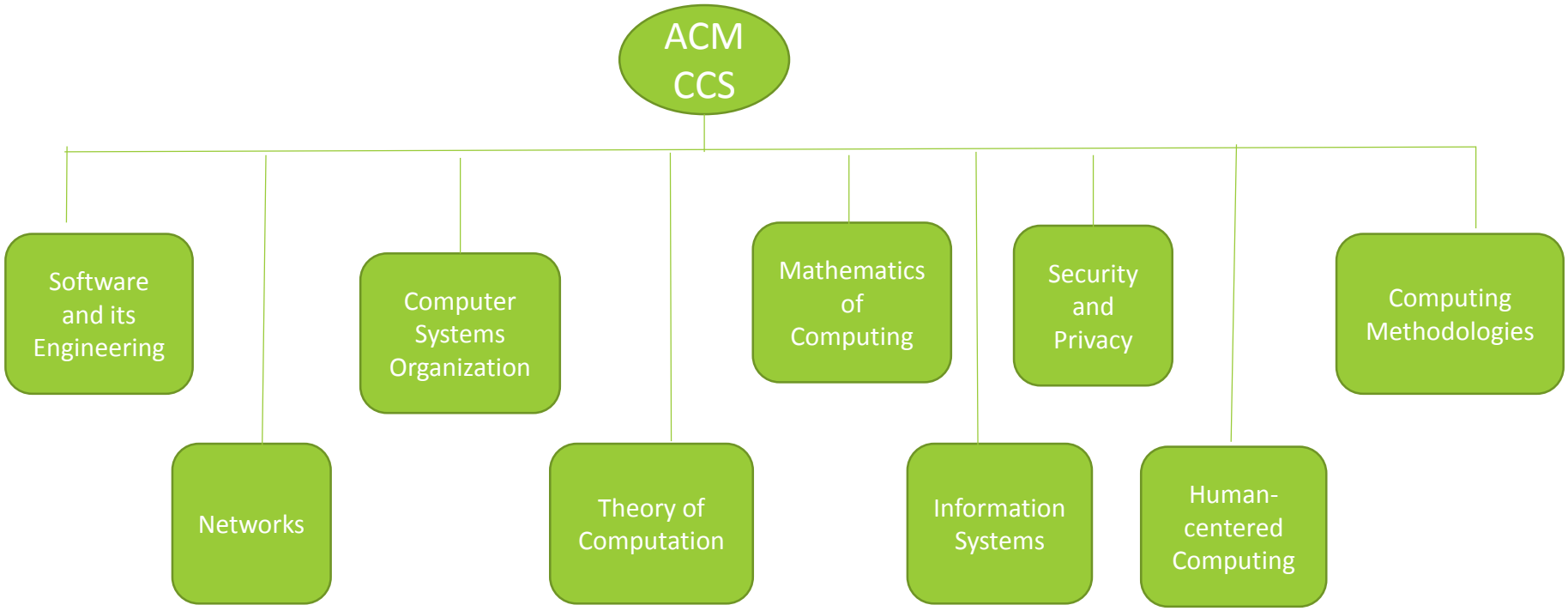


Level-2 (L2): 13 topics
Level-3 (L3): 84 topics

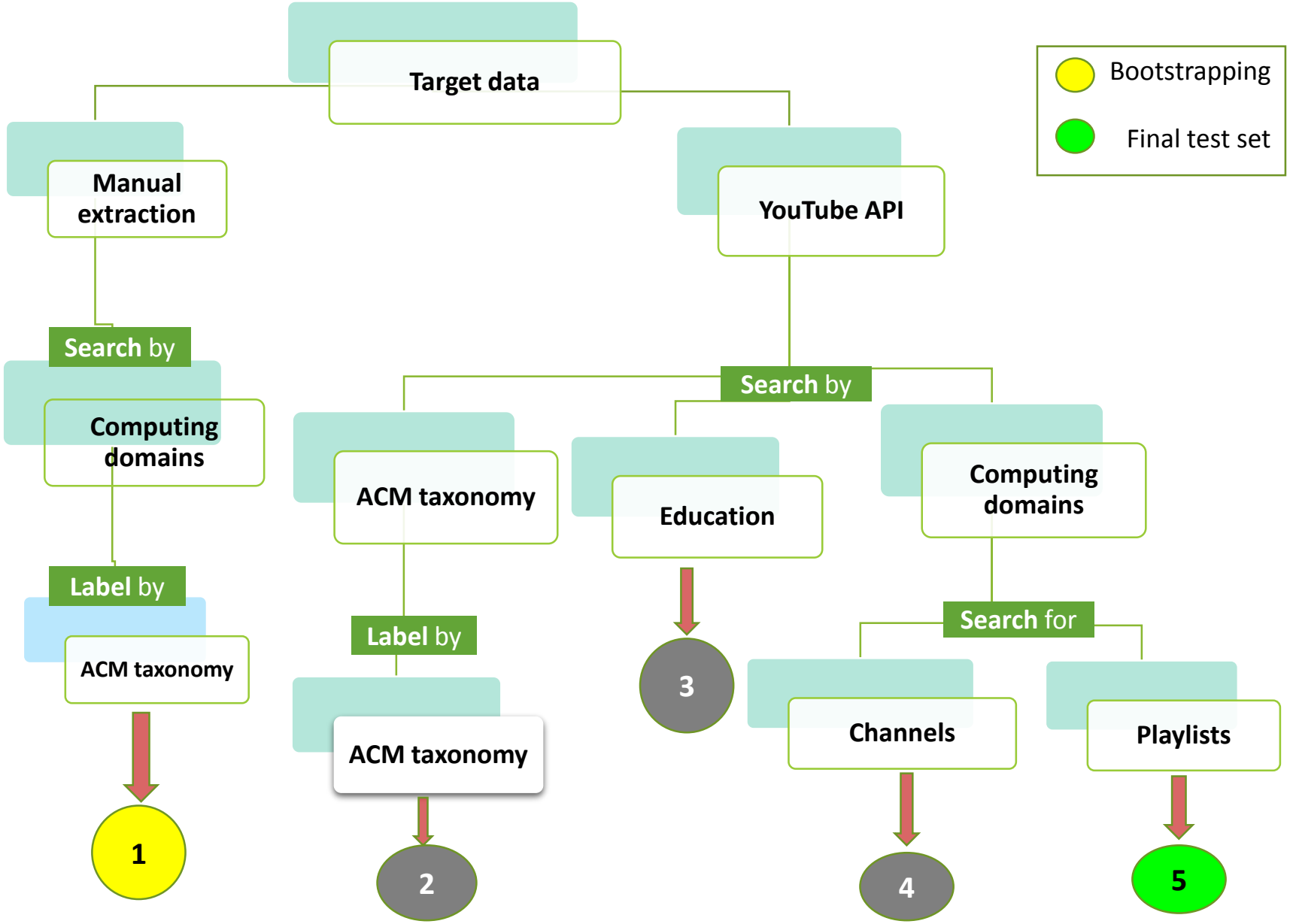
Pruning ACM taxonomy tree



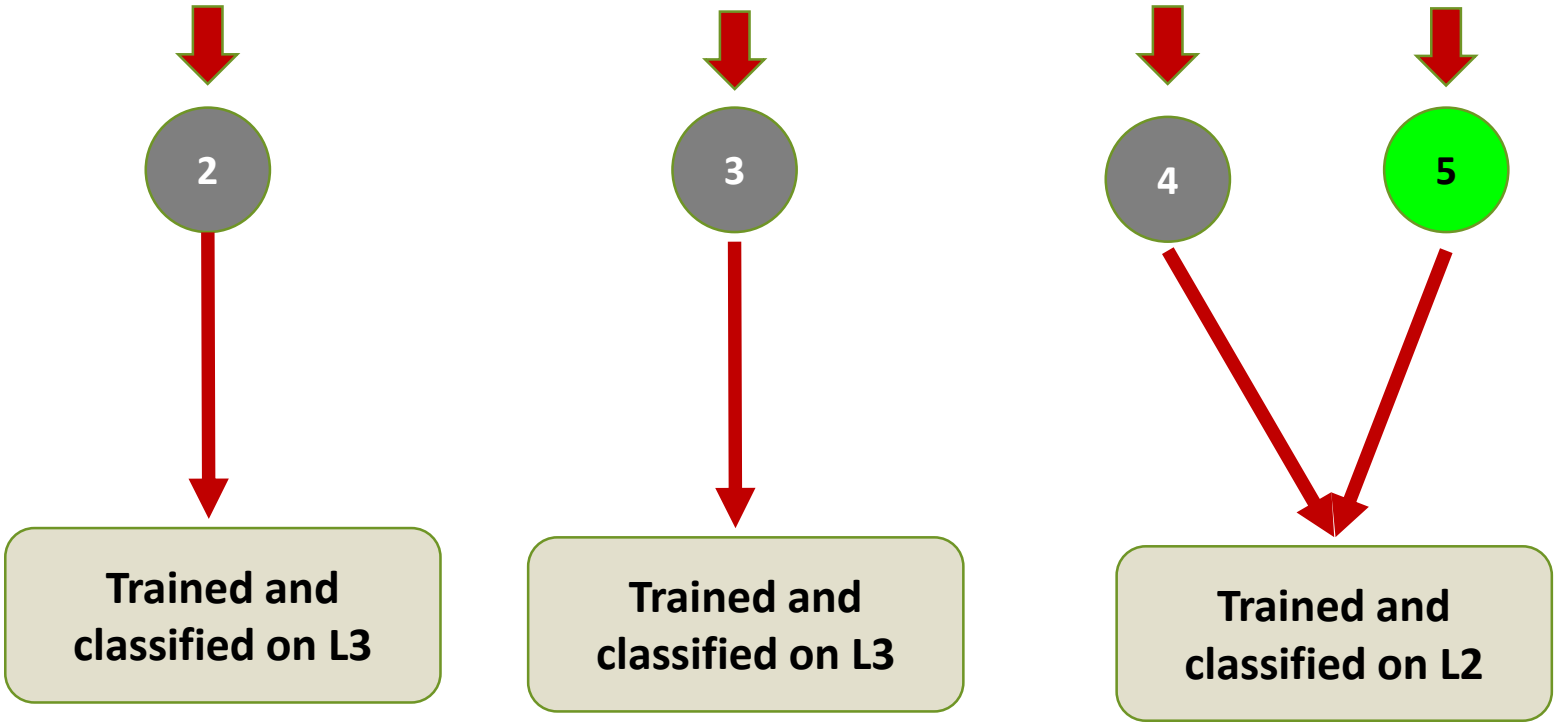
Pruned ACM taxonomy tree



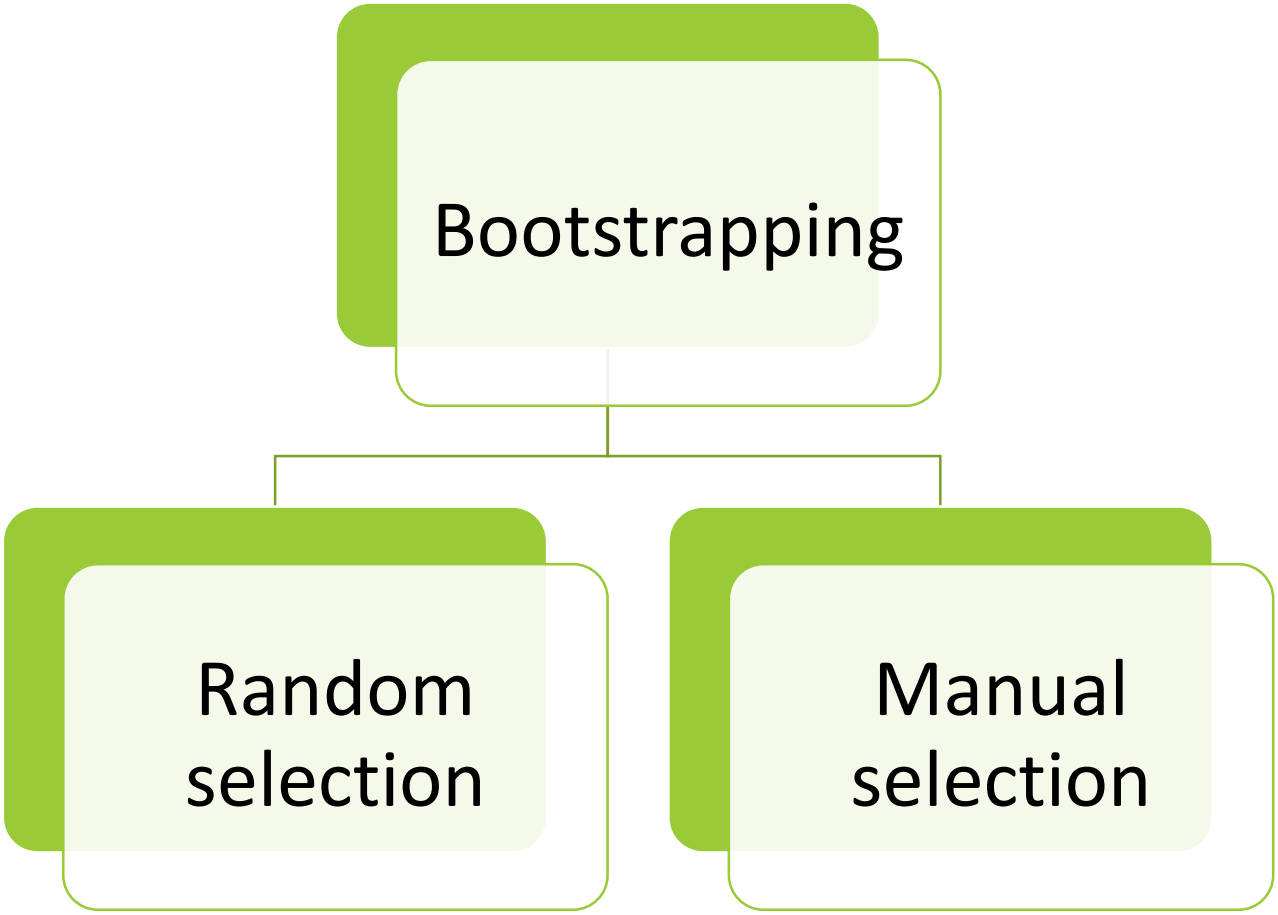
Target data collection approaches



Transfer learning approaches



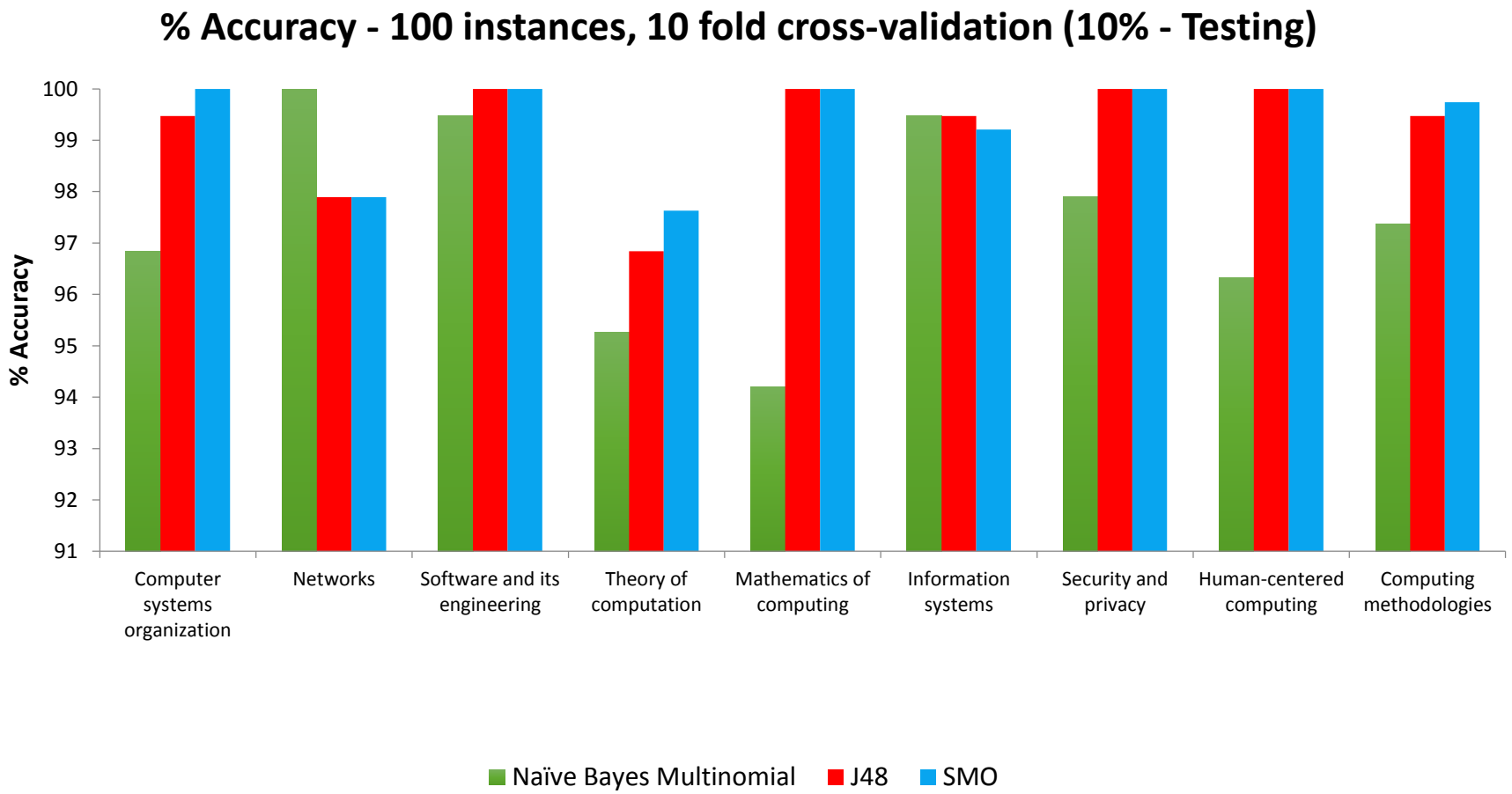
Bootstrapping



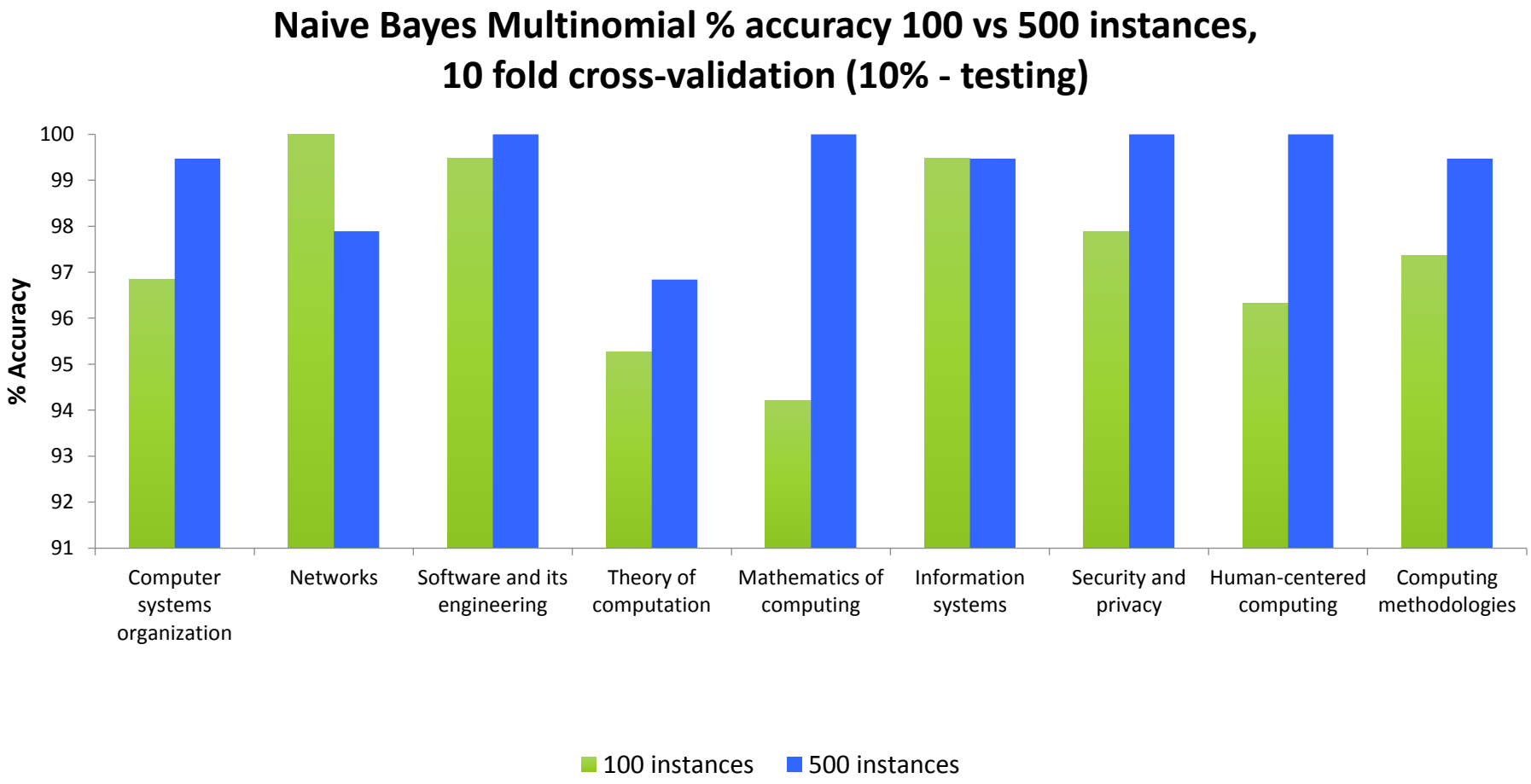
Evaluation - training

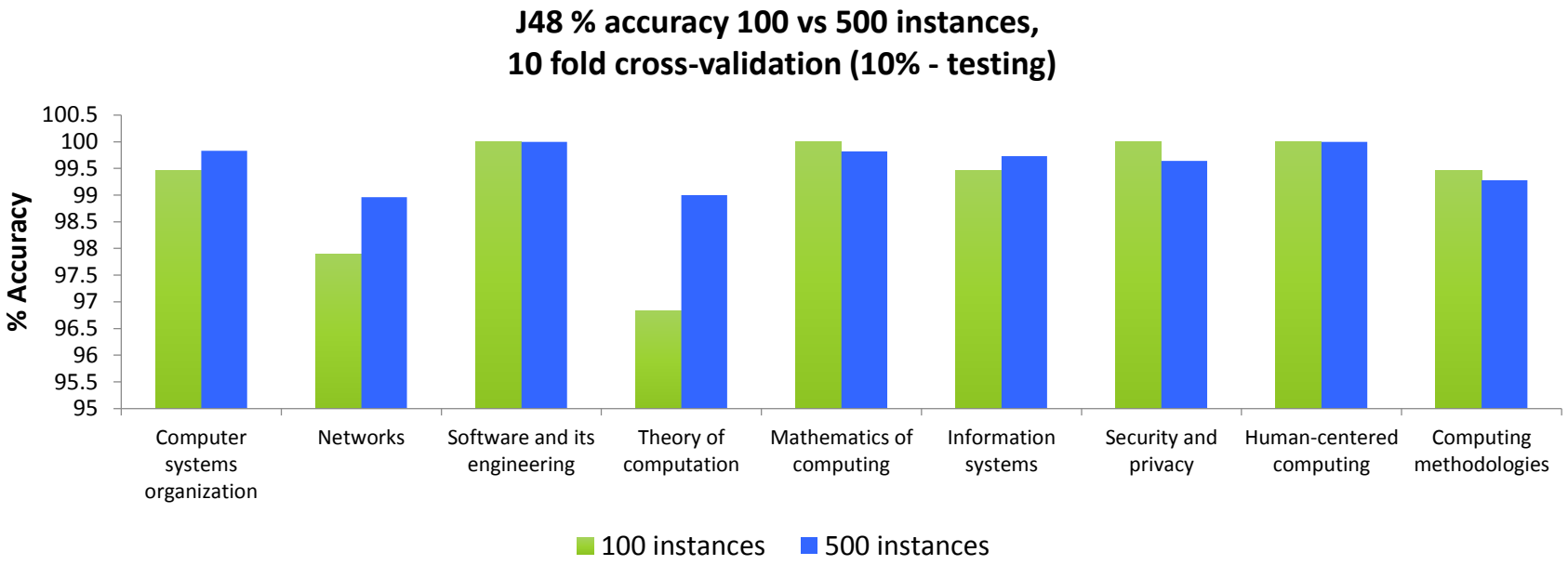
Naïve Bayes Multinomial preferred

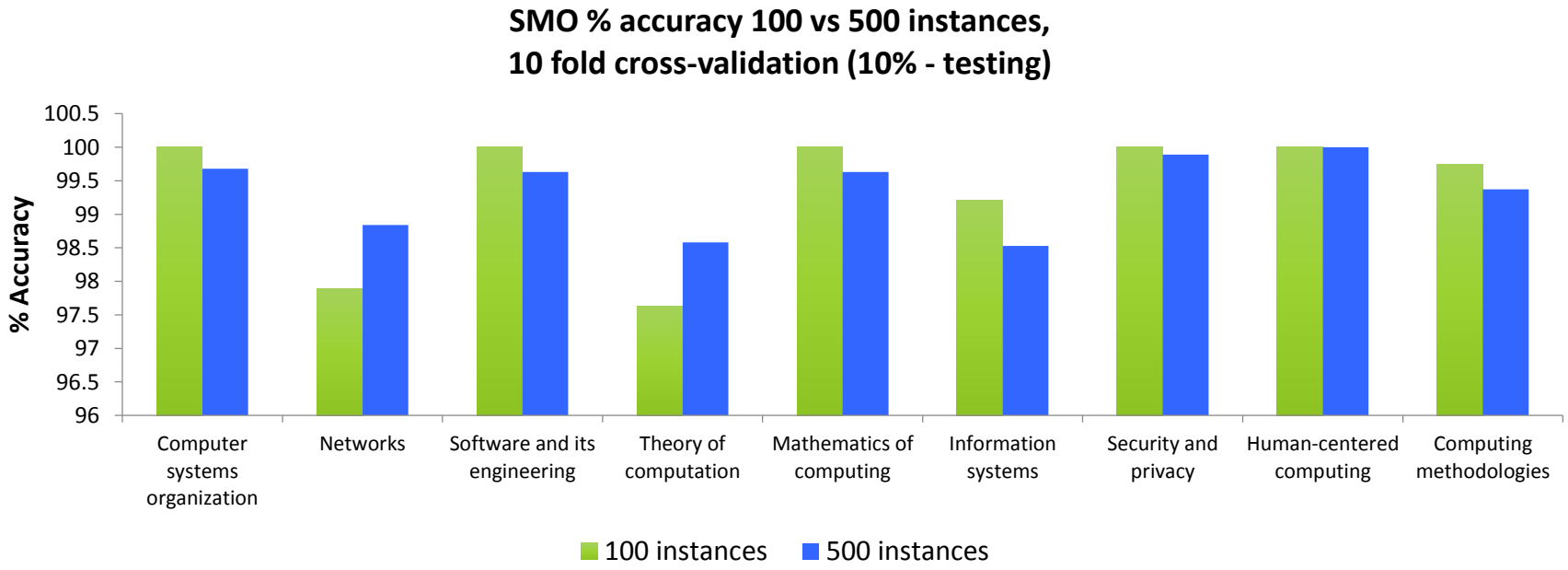
- Fast
- Reduce over-fitting



Evaluation - training (contd.)



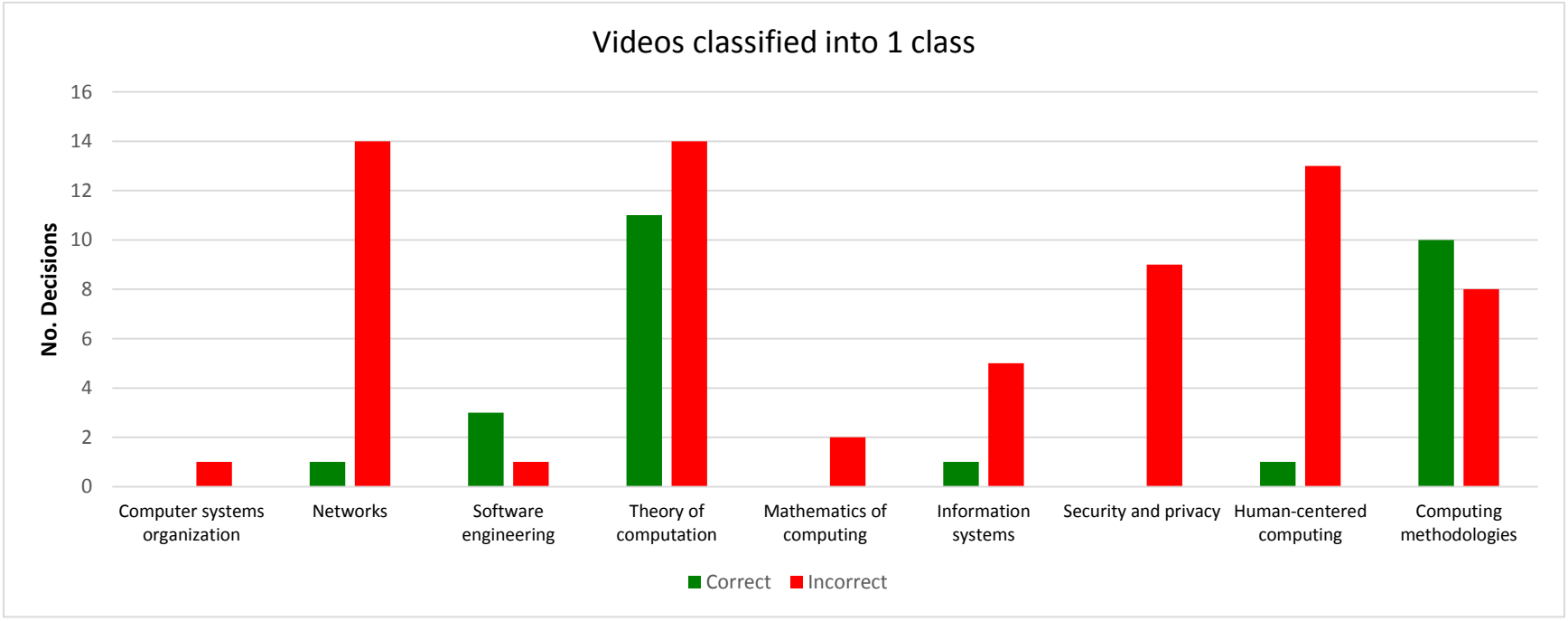

J48 % accuracy 100 vs 500 instances,
10 fold cross-validation (10% - testing)
% Accuracy
100.5
100
99.5
99
98.5
98
97.5
97
96.5
96
95.5
95
Computer systems organization
Networks
Software and its engineering
Theory of computation
Mathematics of computing
Information systems
Security and privacy
Human-centered computing
Computing methodologies
100 instances
500 instances

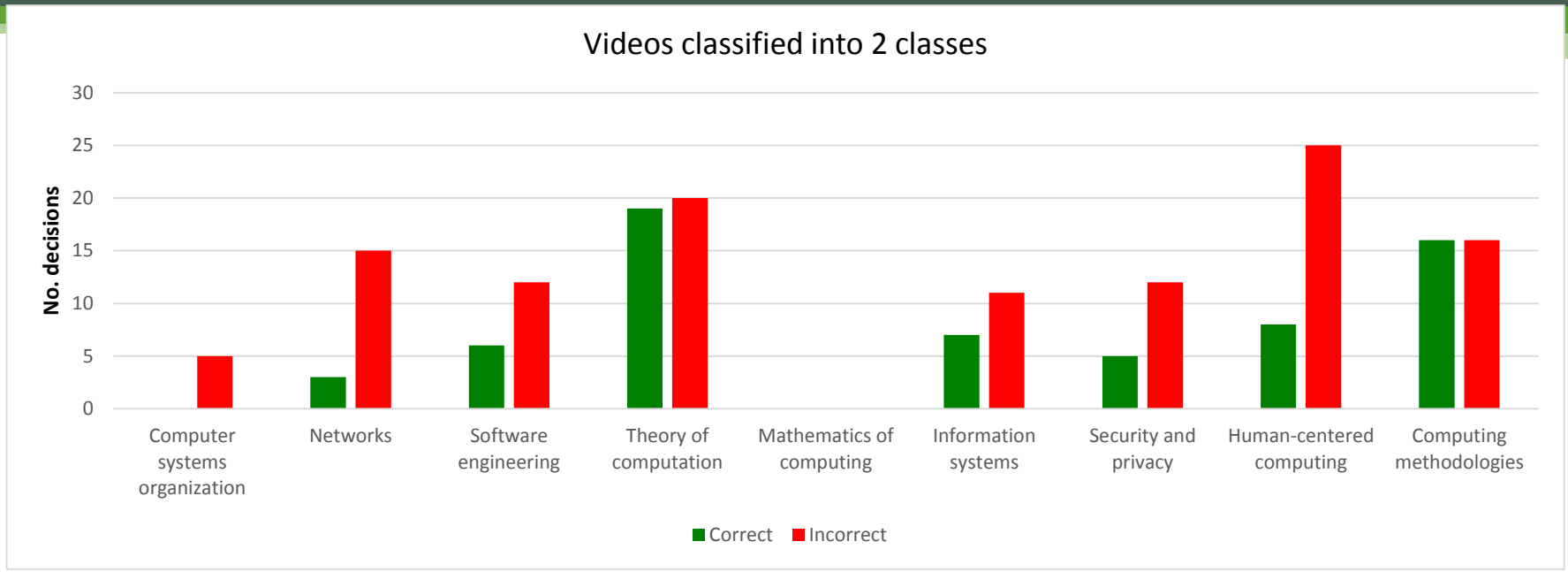

SMO % accuracy 100 vs 500 instances,
10 fold cross-validation (10% - testing)
% Accuracy
100.5
100
99.5
99
98.5
98
97.5
97
96.5
96
Computer systems organization
Networks
Software and its engineering
Theory of computation
Mathematics of computing
Information systems
Security and privacy
Human-centered computing
Computing methodologies
100 instances
500 instances

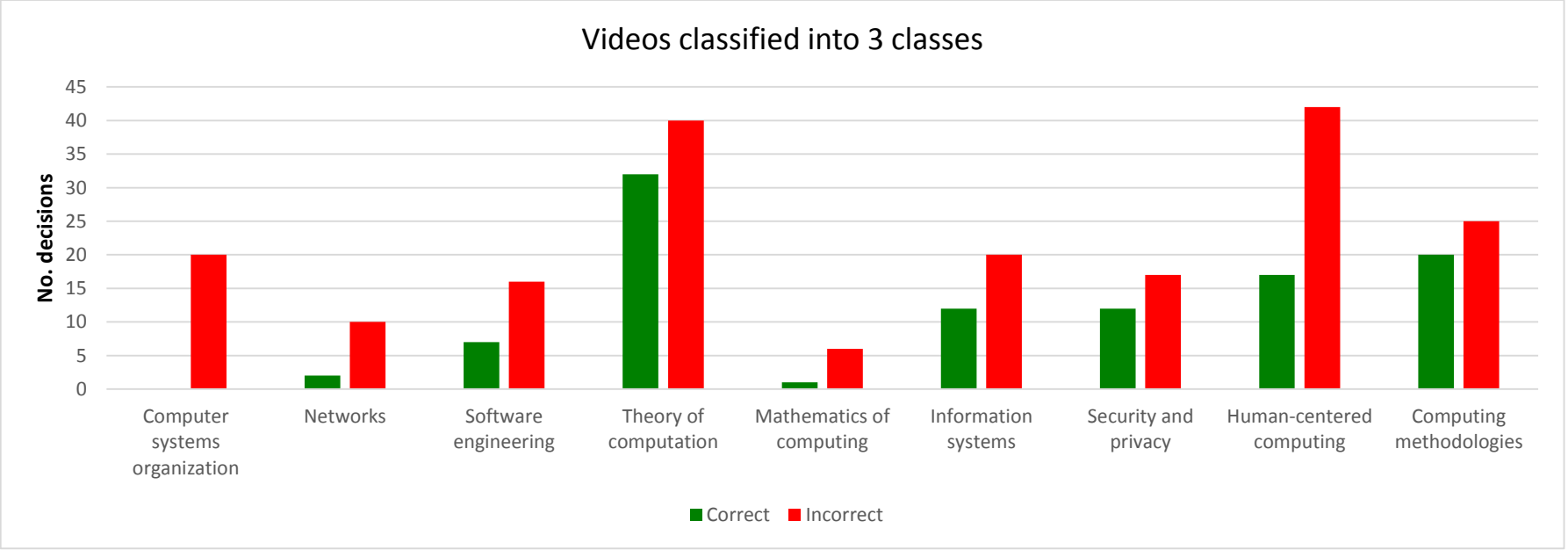
Evaluation - target

Included only videos classified into <= 3 classes

Number of classes	% Correct decisions
1	31
2	35
3	35




Videos classified into 2 classes
No. decisions
30
25
20
15
10
5
0
Computer systems organization
Networks
Software engineering
Theory of computation
Mathematics of computing
Information systems
Security and privacy
Human-centered computing
Computing methodologies
Correct
Incorrect


Videos classified into 3 classes
No. decisions
45
40
35
30
25
20
15
10
5
0
Computer systems organization
Networks
Software engineering
Theory of computation
Mathematics of computing
Information systems
Security and privacy
Human-centered computing
Computing methodologies
Correct
Incorrect

Challenges

- Target data collection
  - Availability and quality of target metadata.
  - Reliability of search.
- Mismatch in ACM and YouTube vocabulary.
- Limited features set for target data (YouTube).
- Interdisciplinary nature of data poses difficulty in classification.

Challenges (contd.)

- ACM CCS is generic and ambiguous

Theory of computation

- Design and analysis of algorithms
  - Graph algorithms analysis
    - Network flows
    - Sparsification and spanners
    - Shortest paths
    - Dynamic graph algorithms
  - Approximation algorithms analysis
    - Scheduling algorithms
    - Packing and covering problems
    - Routing and network design problems
    - Facility location and clustering
    - Rounding techniques
    - Stochastic approximation

Computing methodologies

- Symbolic and algebraic manipulation
  - Symbolic and algebraic algorithms
    - Combinatorial algorithms
    - Algebraic algorithms
    - Nonalgebraic algorithms
    - Symbolic calculus algorithms
    - Exact arithmetic algorithms
    - Hybrid symbolic-numeric methods
    - Discrete calculus algorithms
    - Number theory algorithms
    - Equation and inequality solving algorithms
    - Linear algebra algorithms
    - Theorem proving algorithms
    - Boolean algebra algorithms
    - Optimization algorithms
  - Computer algebra systems

Lessons learned

- “Do not trust anything!”
- Techniques and processes used in transfer learning and text classification.
- YouTube search by playlists – more relevant videos
- Identifying more relevant set of features
  - Voice-to-text conversion
- Classification in same domains is easier.

Future work

- Avoid classification into multiple classes – probability of correctness
- Extend the target set to different domains such as *slideshare*

- Enhancing features selection
  - NLP to refine the features
  - Voice-to-text transformation
  - Image processing
    - CBIR
    - Text extraction - subtitles, text embedded


THANK YOU
Questions ?