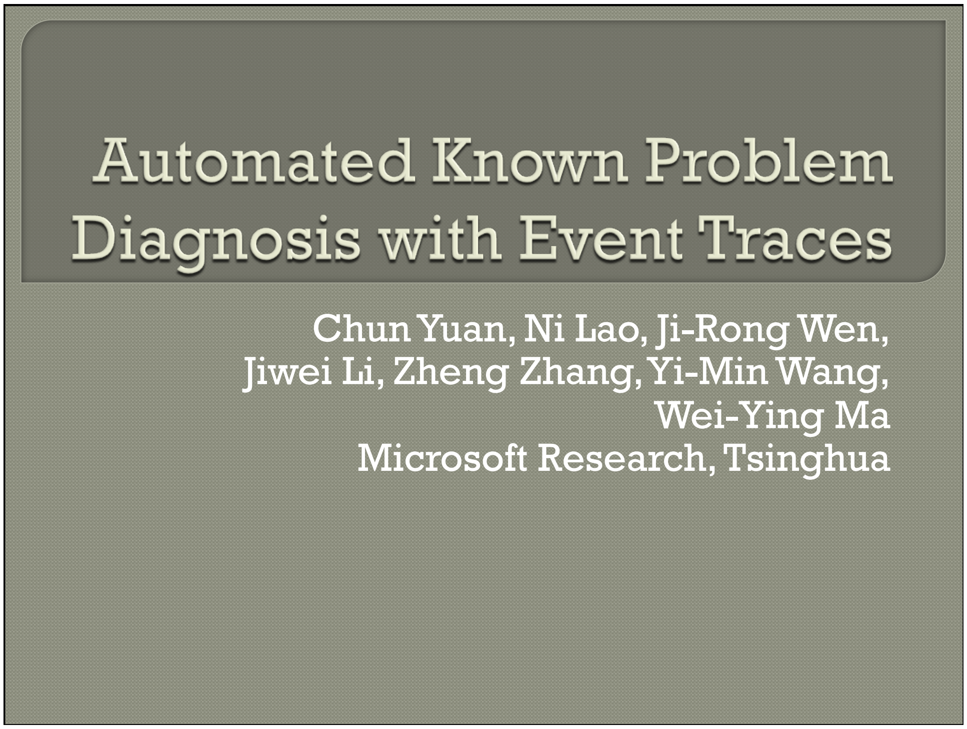
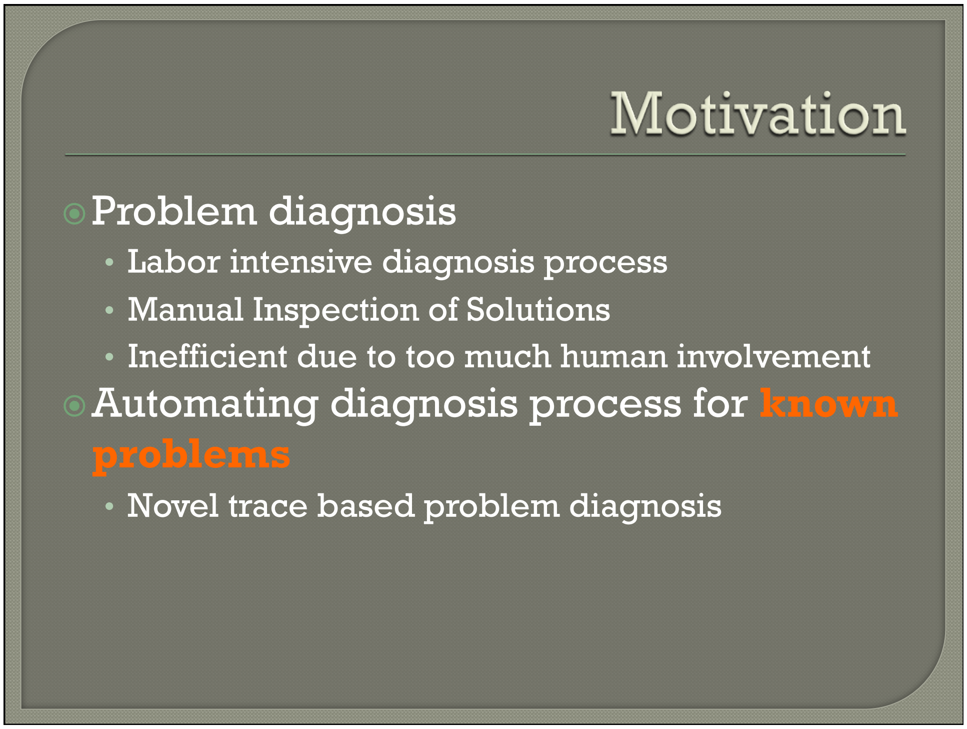


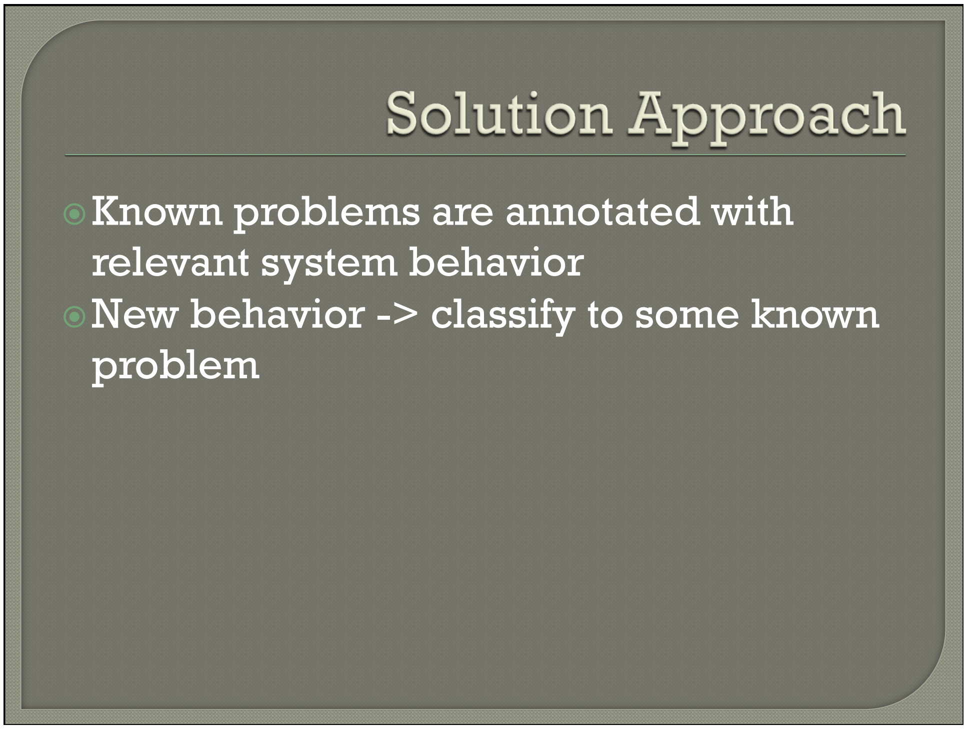
Automated Known Problem Diagnosis with Event Traces

Chun Yuan, Ni Lao, Ji-Rong Wen,
Jiwei Li, Zheng Zhang, Yi-Min Wang,
Wei-Ying Ma
Microsoft Research, Tsinghua



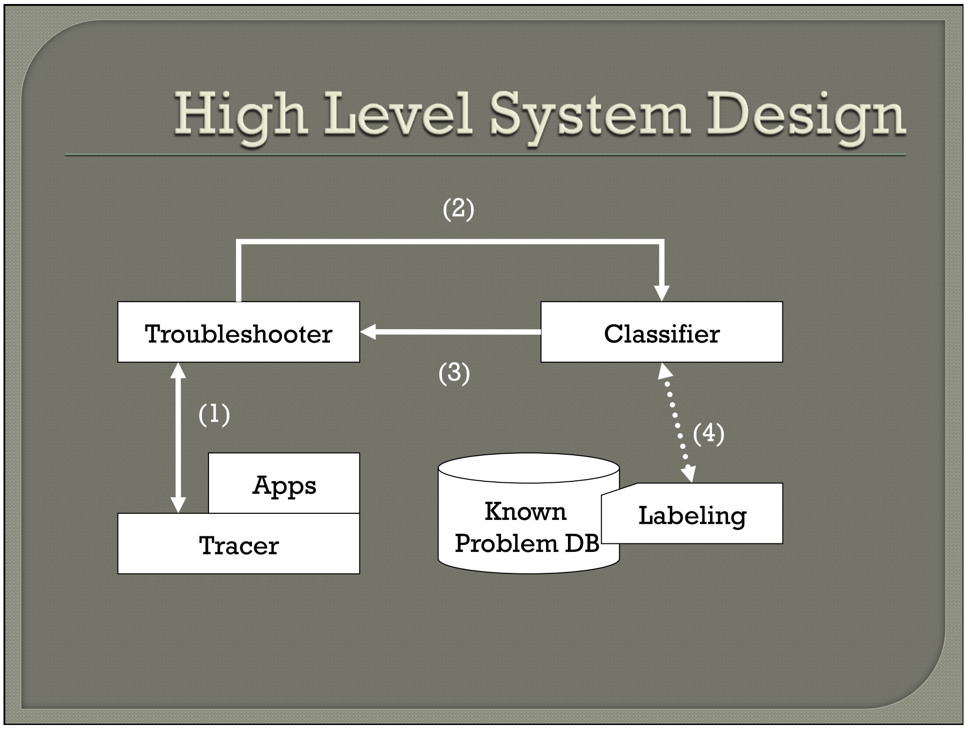
Motivation

- Problem diagnosis
  - Labor intensive diagnosis process
  - Manual Inspection of Solutions
  - Inefficient due to too much human involvement
- Automating diagnosis process for **known problems**
  - Novel trace based problem diagnosis



Solution Approach

- Known problems are annotated with relevant system behavior
- New behavior -> classify to some known problem



High Level System Design

(2)

Troubleshooter

Classifier

(3)

(1)

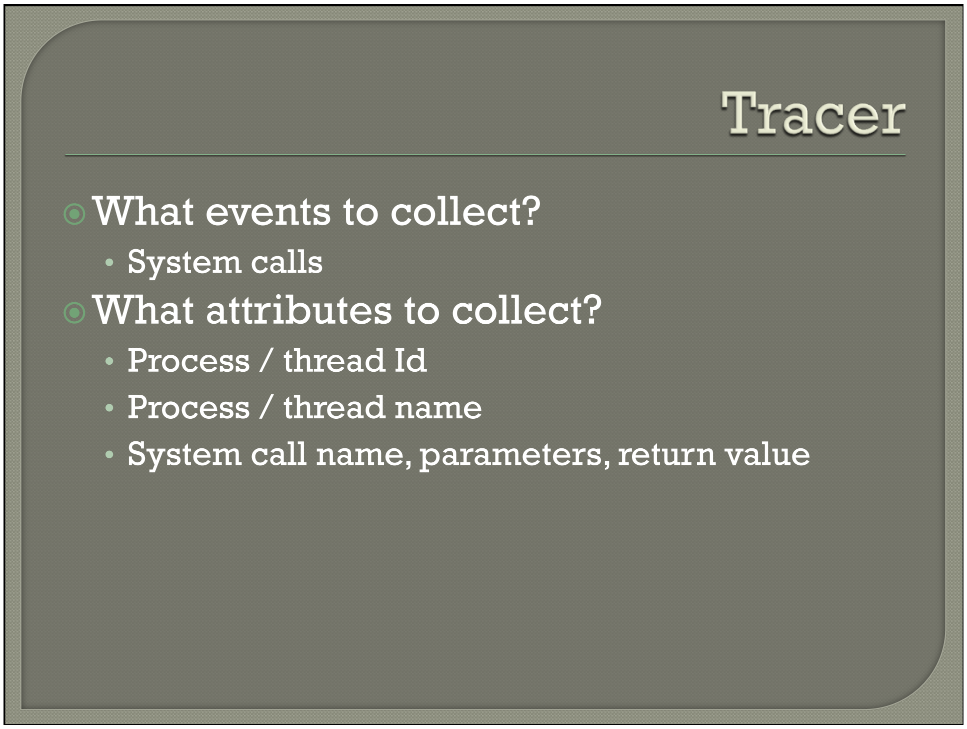
(4)

Apps

Tracer

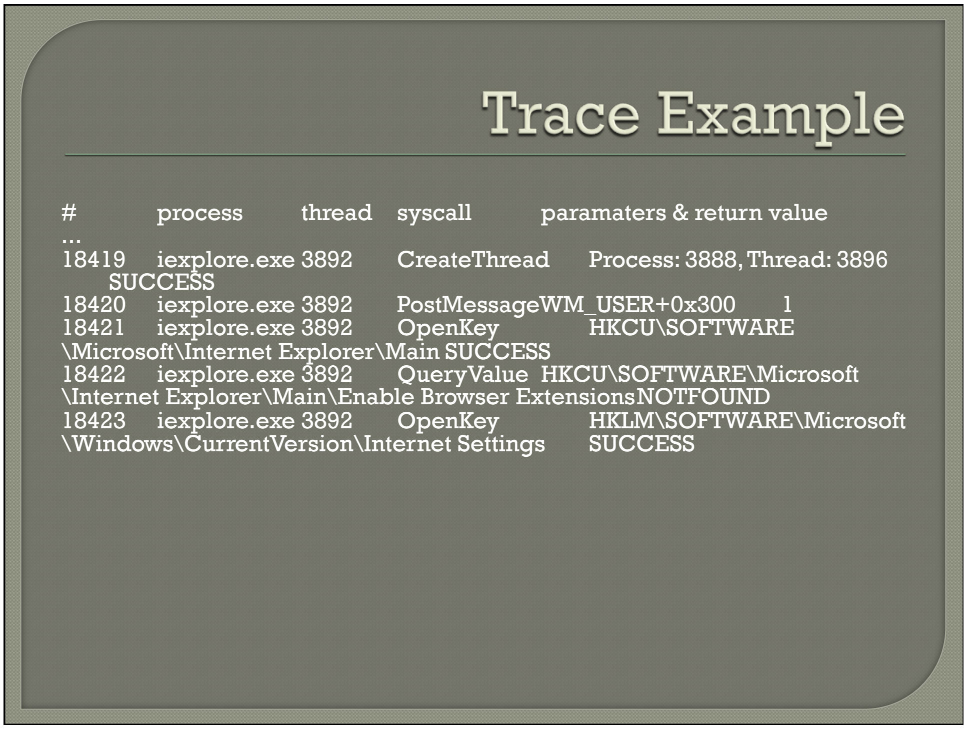
Known Problem DB

Labeling



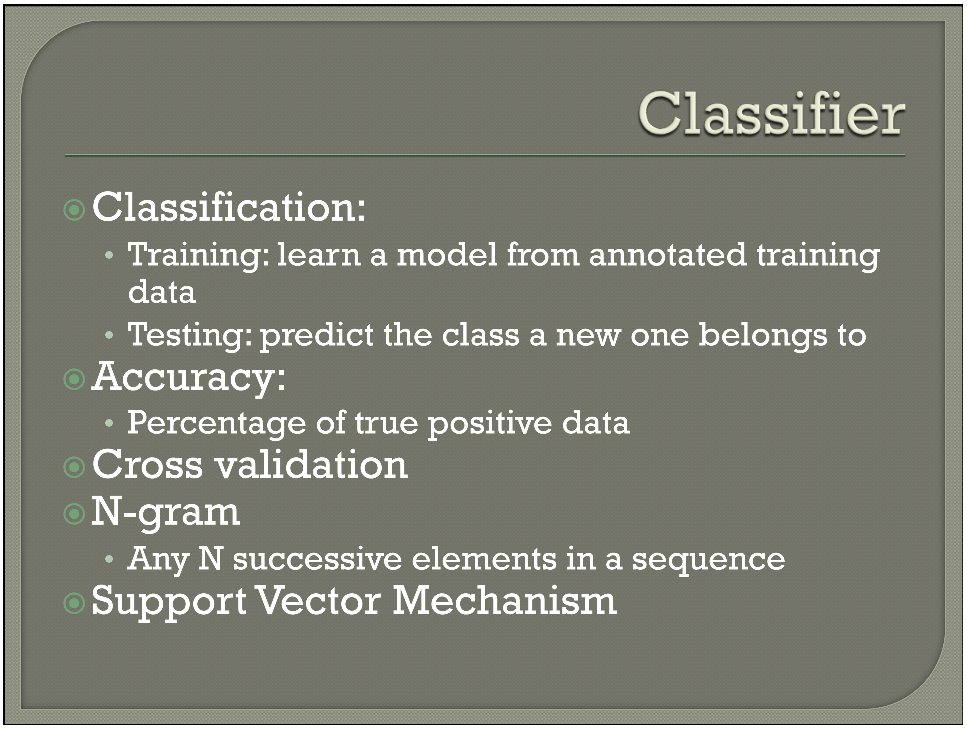
Tracer

- What events to collect?
  - System calls
- What attributes to collect?
  - Process / thread Id
  - Process / thread name
  - System call name, parameters, return value



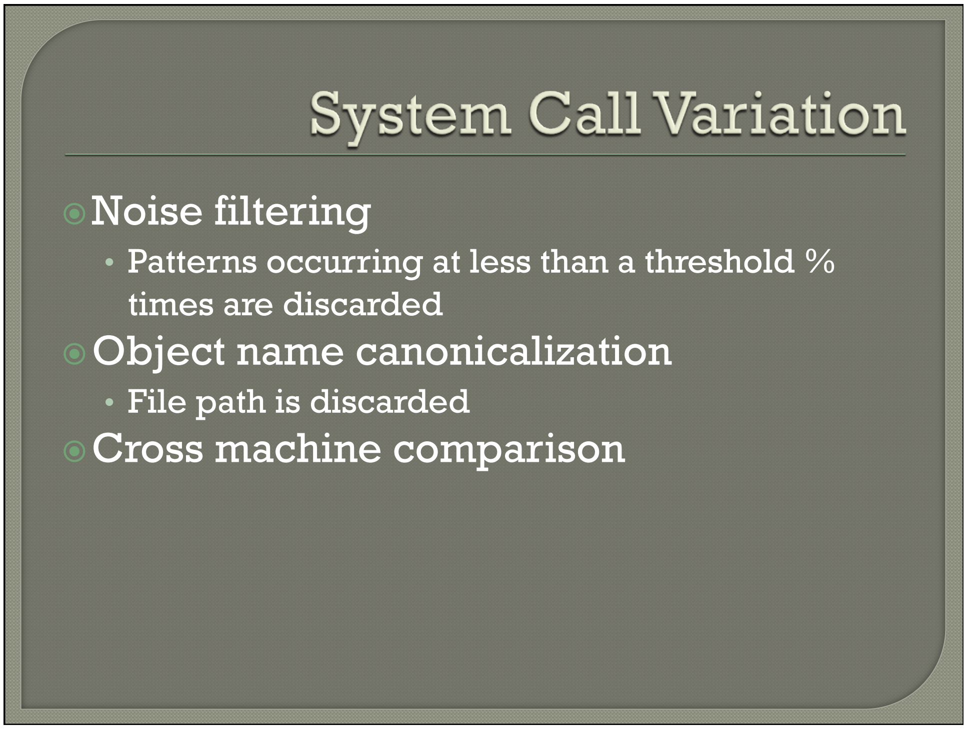
Trace Example

```
#         process         thread   syscall          paramaters & return value
...
18419     iexplore.exe 3892        CreateThread     Process: 3888, Thread: 3896
     SUCCESS
18420     iexplore.exe 3892        PostMessageWM_USER+0x300       1
18421     iexplore.exe 3892        OpenKey          HKCU\SOFTWARE
\Microsoft\Internet Explorer\Main SUCCESS
18422     iexplore.exe 3892        QueryValue  HKCU\SOFTWARE\Microsoft
\Internet Explorer\Main\Enable Browser Extensions NOTFOUND
18423     iexplore.exe 3892        OpenKey          HKLM\SOFTWARE\Microsoft
\Windows\CurrentVersion\Internet Settings       SUCCESS
```



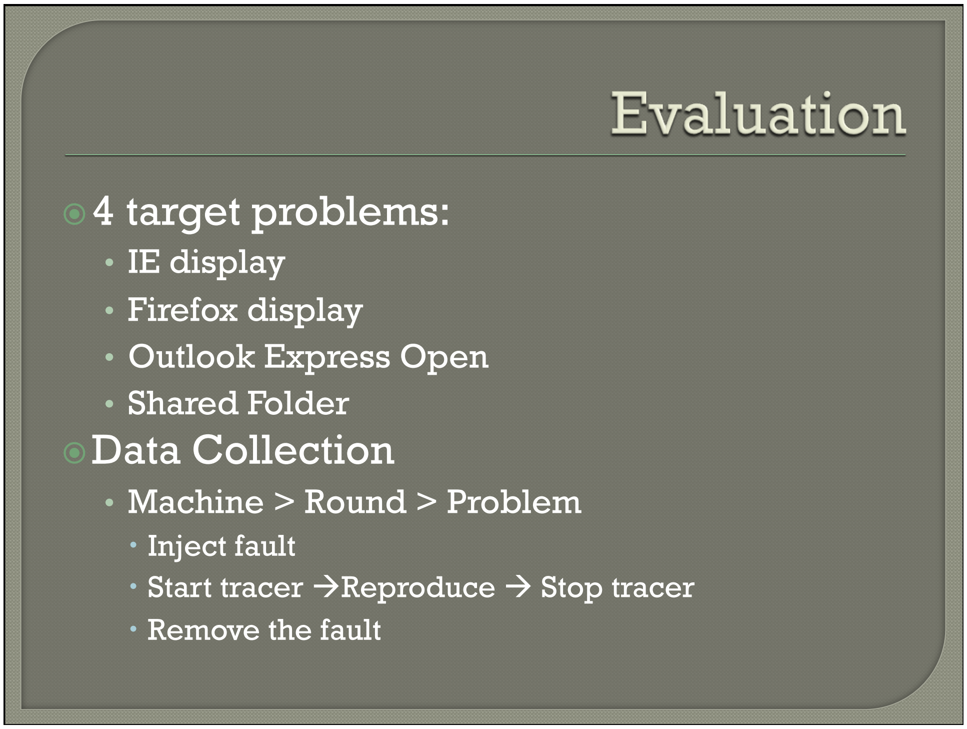
Classifier

- Classification:
  - Training: learn a model from annotated training data
  - Testing: predict the class a new one belongs to
- Accuracy:
  - Percentage of true positive data
- Cross validation
- N-gram
  - Any N successive elements in a sequence
- Support Vector Mechanism



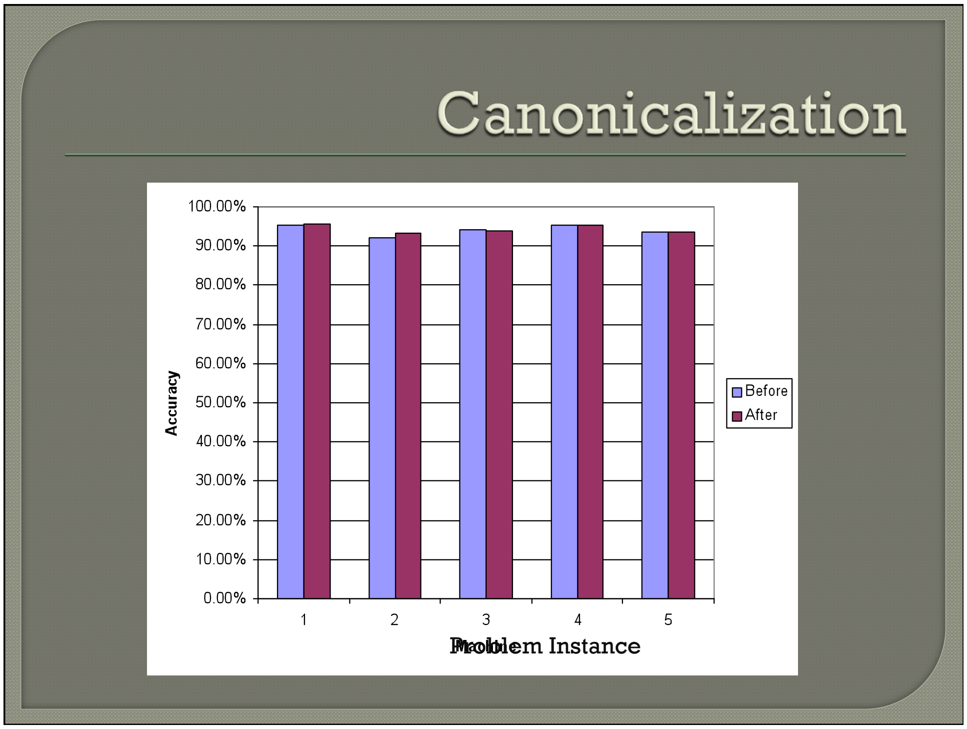
System Call Variation

- Noise filtering
  - Patterns occurring at less than a threshold % times are discarded
- Object name canonicalization
  - File path is discarded
- Cross machine comparison

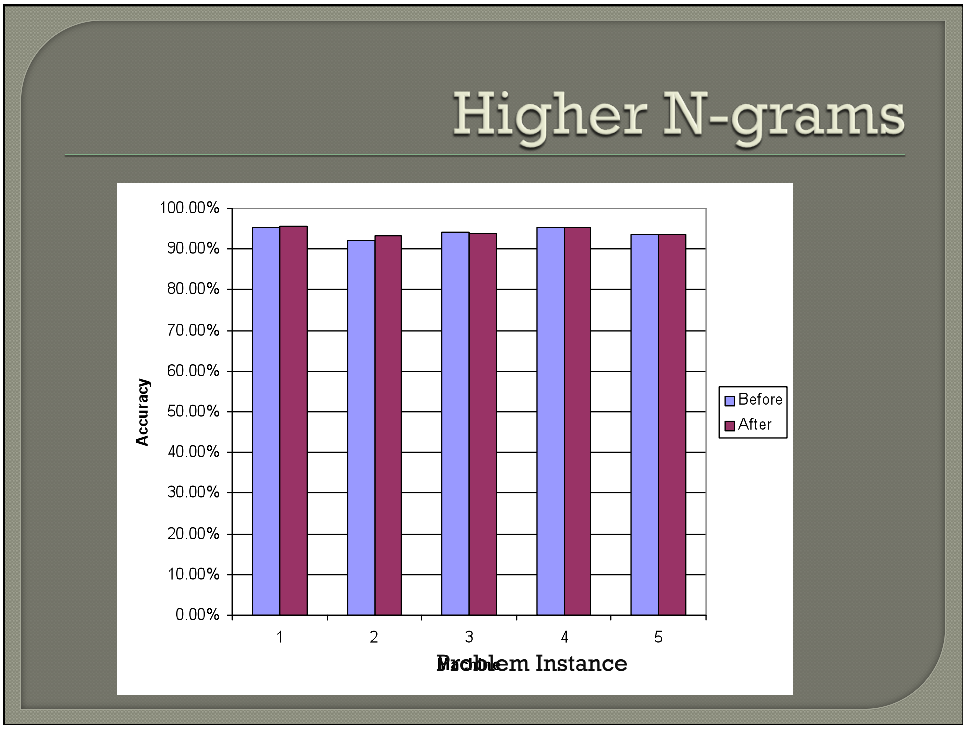


Evaluation

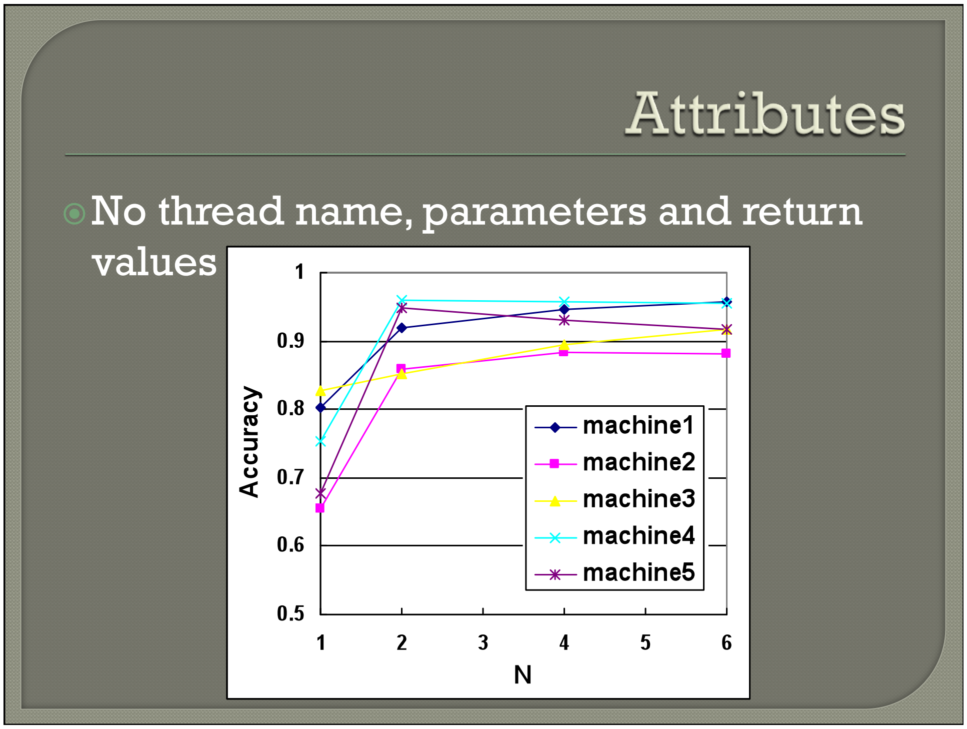
- 4 target problems:
  - IE display
  - Firefox display
  - Outlook Express Open
  - Shared Folder
- Data Collection
  - Machine > Round > Problem
    - Inject fault
    - Start tracer →Reproduce → Stop tracer
    - Remove the fault



Canonicalization

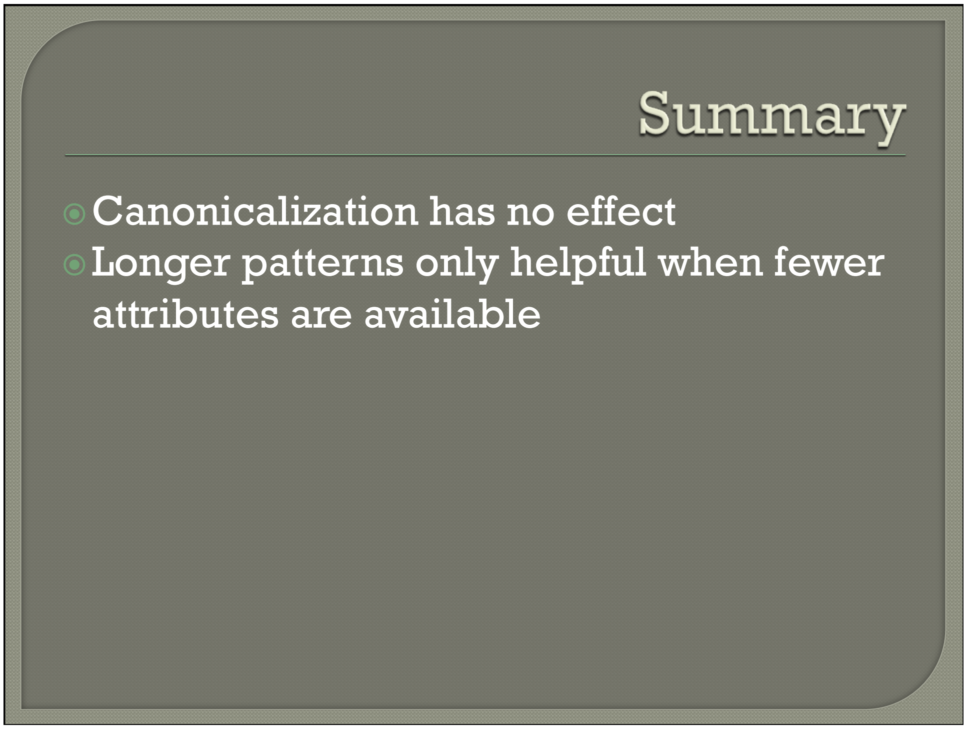


Higher N-grams



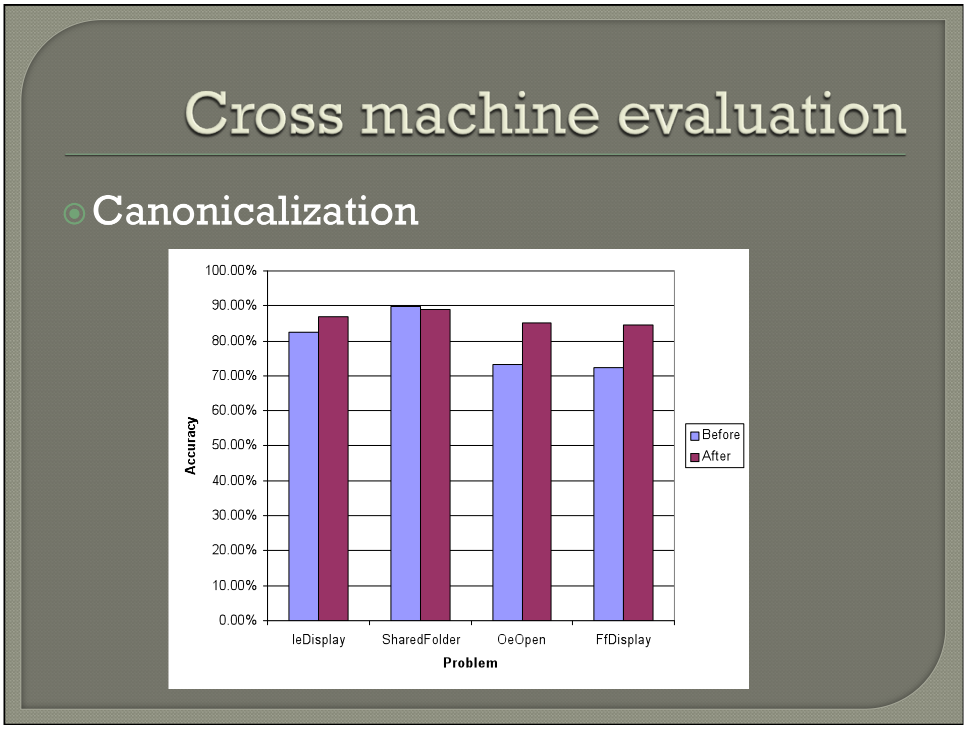
Attributes

- No thread name, parameters and return values



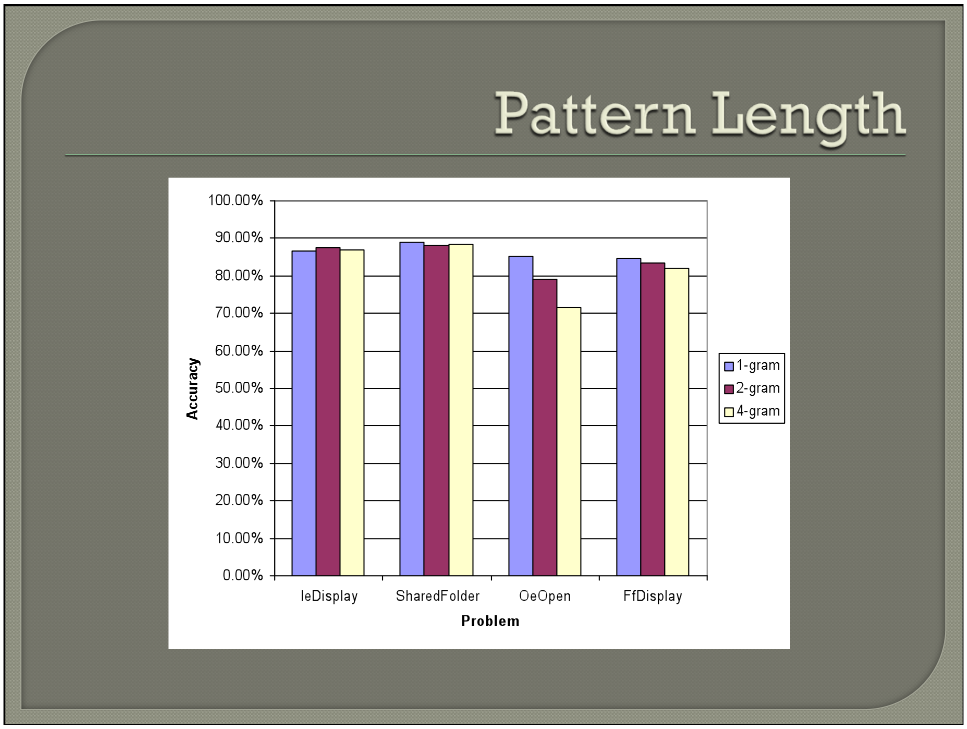
Summary

- Canonicalization has no effect
- Longer patterns only helpful when fewer attributes are available

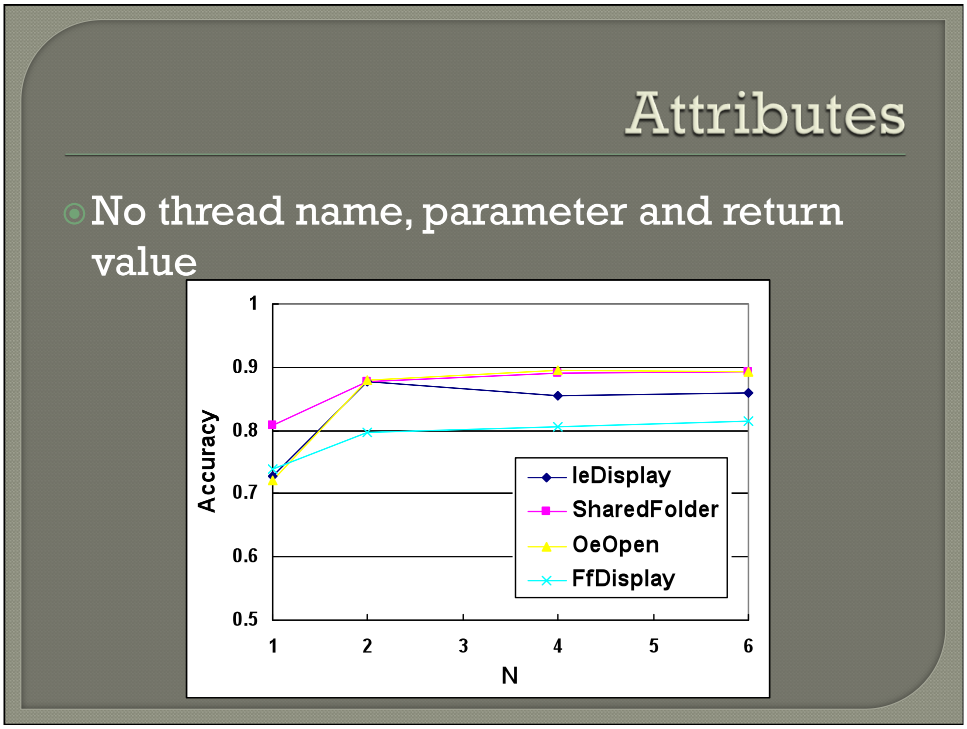


Cross machine evaluation

- Canonicalization

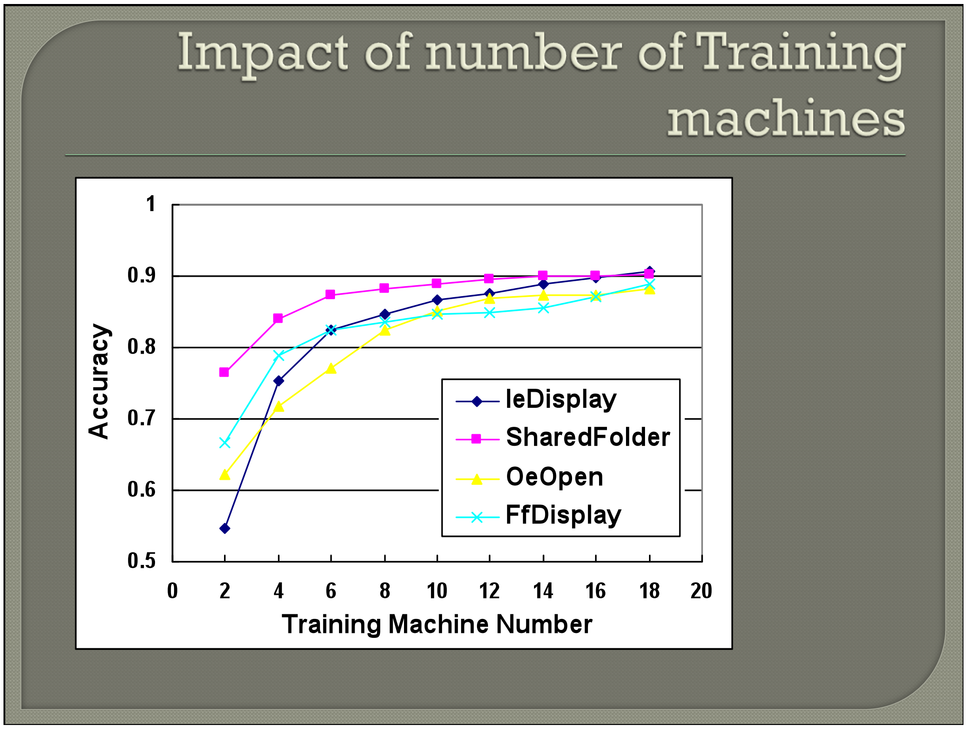


Pattern Length

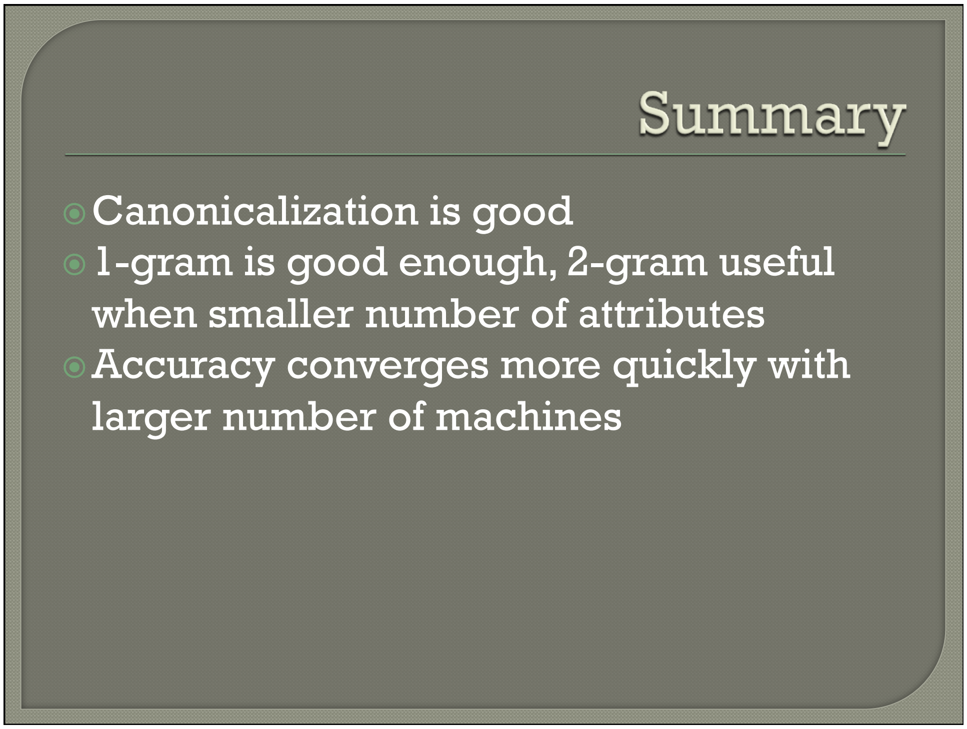


Attributes

- No thread name, parameter and return value



Impact of number of Training machines



Summary

- Canonicalization is good
- 1-gram is good enough, 2-gram useful when smaller number of attributes
- Accuracy converges more quickly with larger number of machines

Questions?

Thank You