# Semantic Flow Augmentation for the Automated Discovery of Organizational Relationships

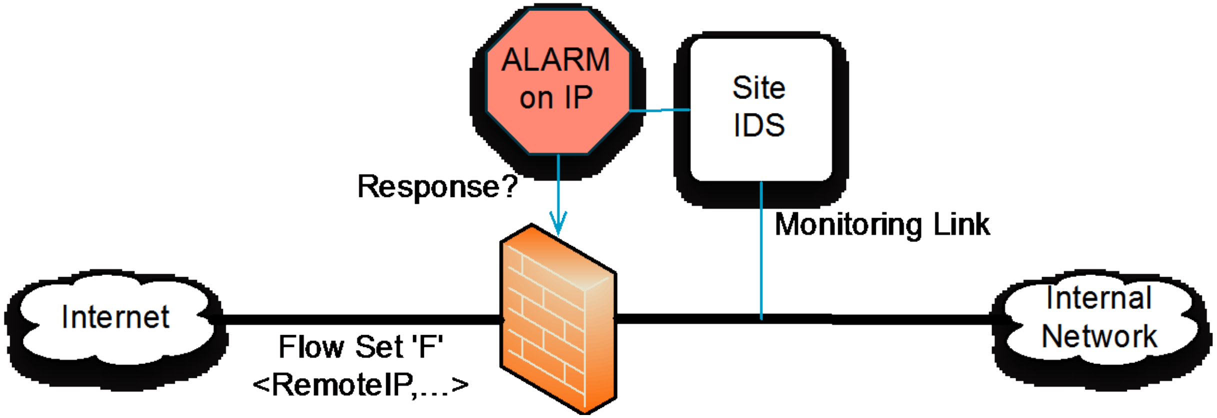
Chris Strasburg*, Harris T Lin, Nikolas Kinkel

The Ames Laboratory

{cstras,htlin,nskinkel}@ameslab.gov

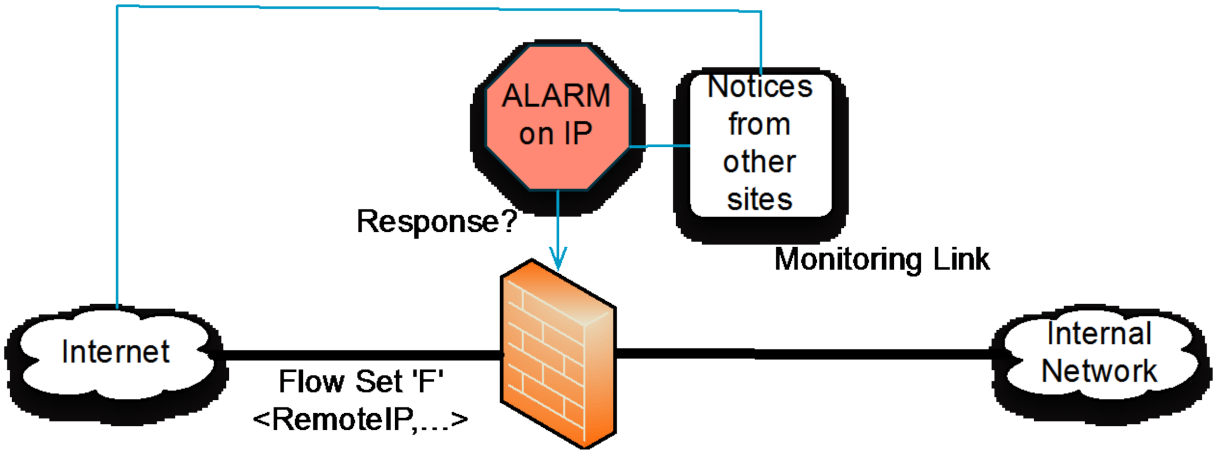
* - Presenting

# Relationship Discovery – Why does it matter?



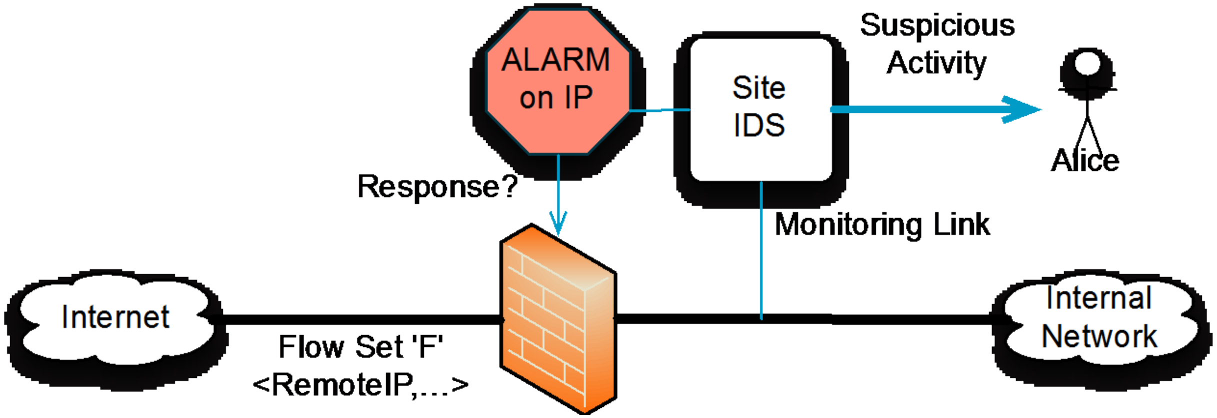
- What is the impact of disrupting communication associated with flow set ‘F’?

# Relationship Discovery – Why does it matter?



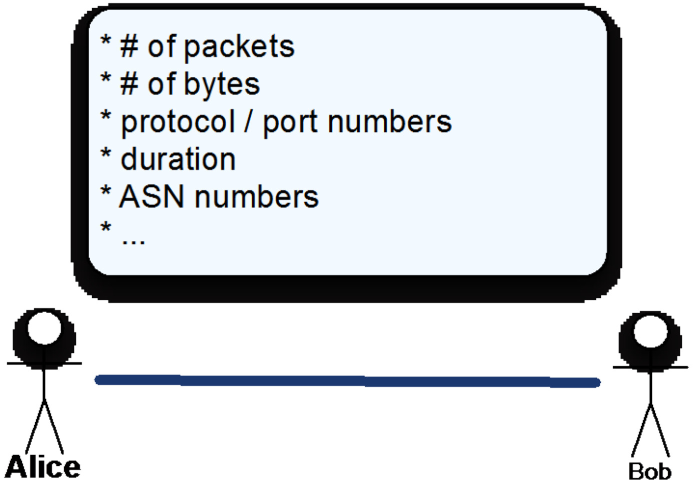
- What is the impact of disrupting communication associated with flow set ‘F’?

# Relationship Discovery – Why does it matter?



- Which alarms are most critical to manually investigate?

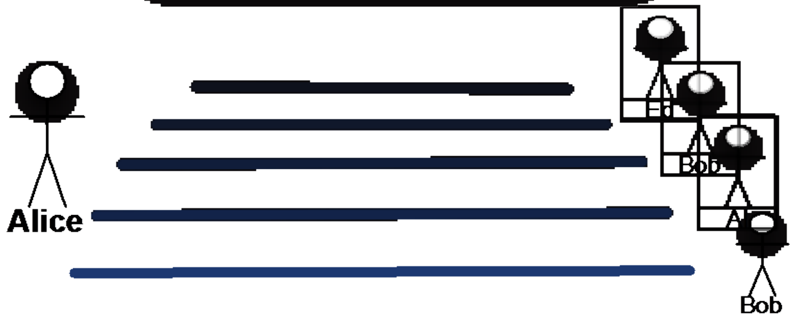
# What is Semantic Flow Augmentation



* # of packets
* # of bytes
* protocol / port numbers
* duration
* ASN numbers
* ...

# What is Semantic Flow Augmentation

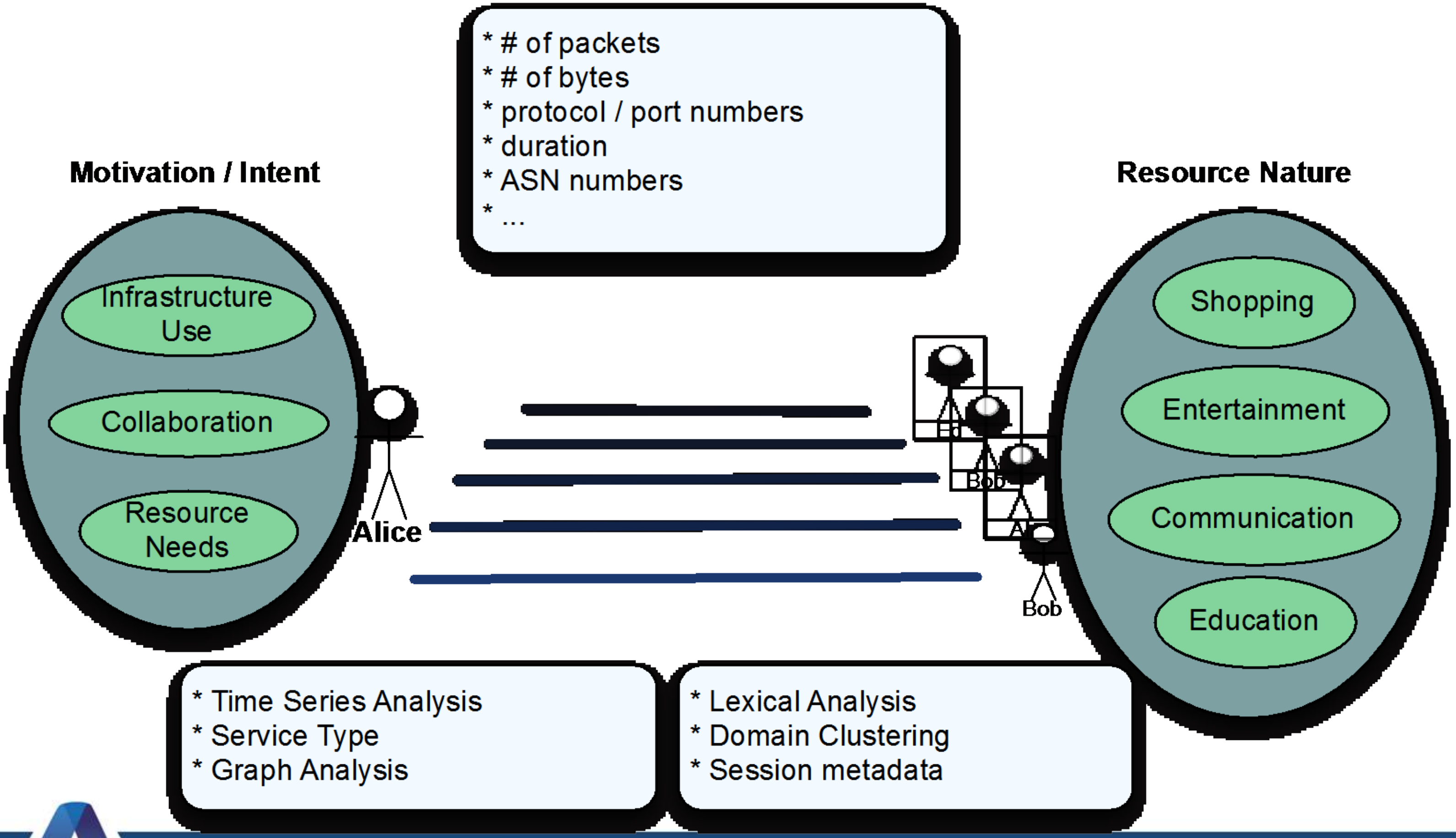
* # of packets
* # of bytes
* protocol / port numbers
* duration
* ASN numbers
* ...



* Time Series Analysis
* Service Type
* Graph Analysis

# What is Semantic Flow Augmentation

- Semantic – *Of or relating to meaning…*



* # of packets
* # of bytes
* protocol / port numbers
* duration
* ASN numbers
* …

Motivation / Intent

Infrastructure Use

Collaboration

Resource Needs

Alice

Ed

Bob

Al

Bob

Resource Nature

Shopping

Entertainment

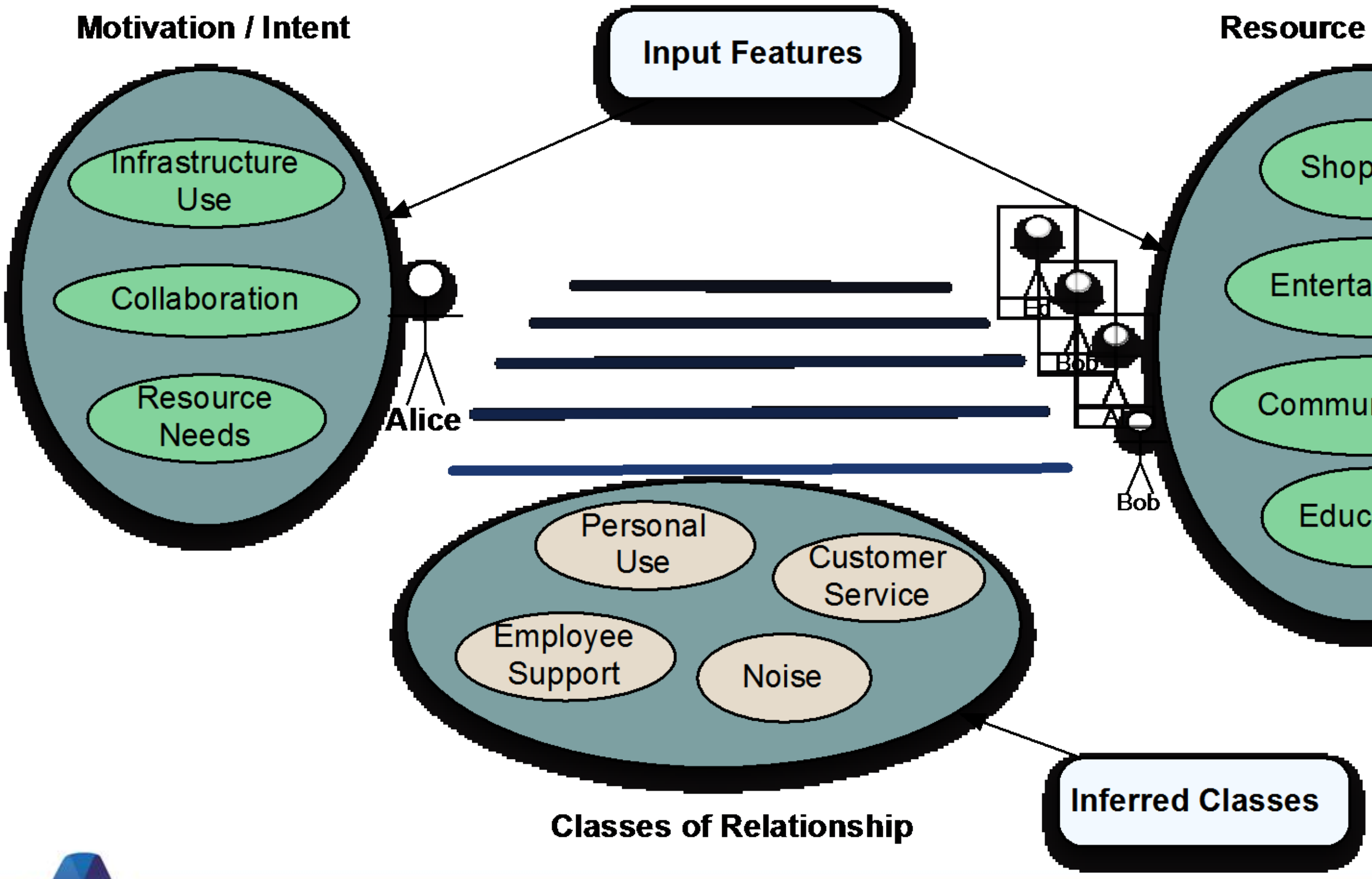
Communication

Education

* Time Series Analysis
* Service Type
* Graph Analysis

* Lexical Analysis
* Domain Clustering
* Session metadata

# Why Semantic Augmentation



Motivation / Intent

Infrastructure Use

Collaboration

Resource Needs

Input Features

Resource Nature

Shopping

Entertainment

Communication

Education

Personal Use

Customer Service

Employee Support

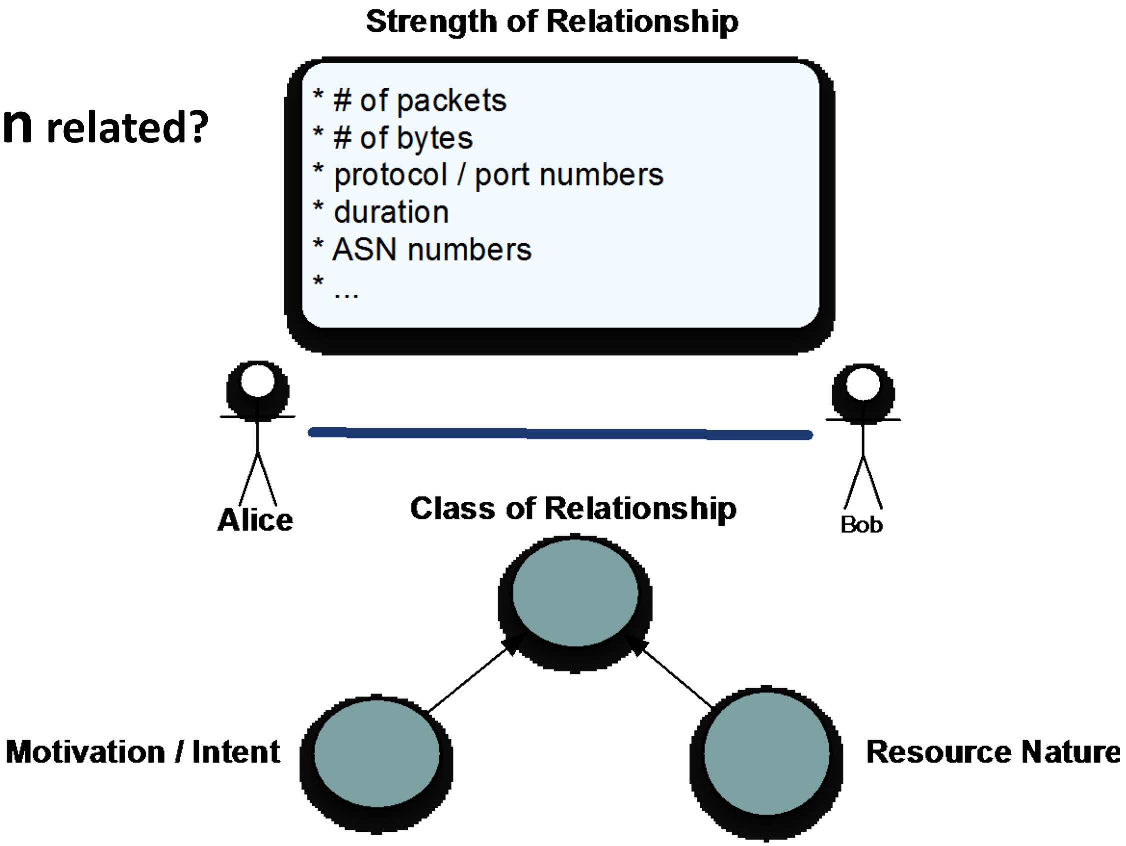
Noise

Classes of Relationship

Inferred Classes

# Why Semantic Augmentation

Is it **mission** related?



# Statistical Features

- Flow Statistics
  - # of Flows
  - # of Bytes
  - Peer count

- Timeseries Analysis
  - First seen
  - Last seen
  - Fourier Transform Coefficient

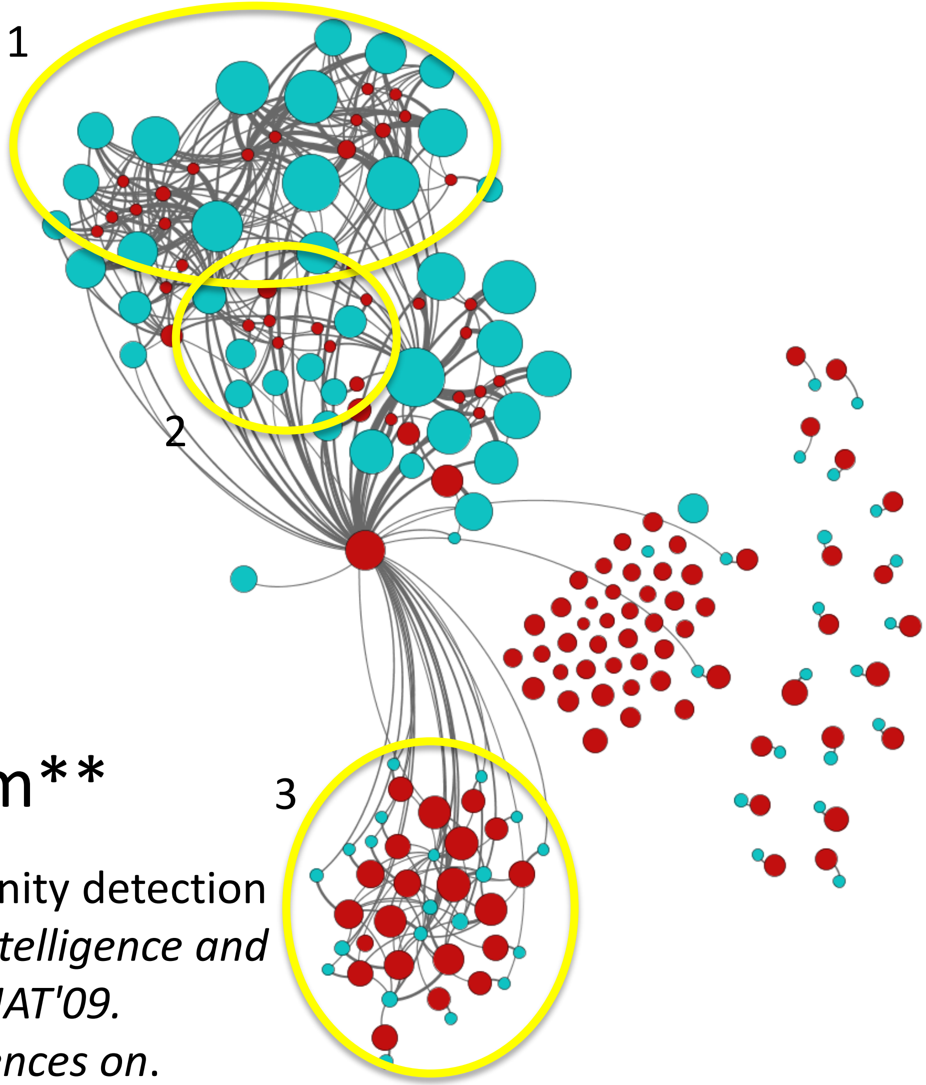
# Semantic Features

- Lexical Analysis (Mallet)
  - – Cluster according to web page contents from:
    - Reverse DNS Lookups
    - WHOIS Org Searches
- Session Metadata
  - – Requested URLs

- Service Distribution
  - – Interactive / Authenticated (SSH, IMAP, POP)
  - – Interactive / Non-Authenticated (STMP, HTTP/S)
  - – Non-Interactive (NTP, DNS)

# Semantic Features (2)

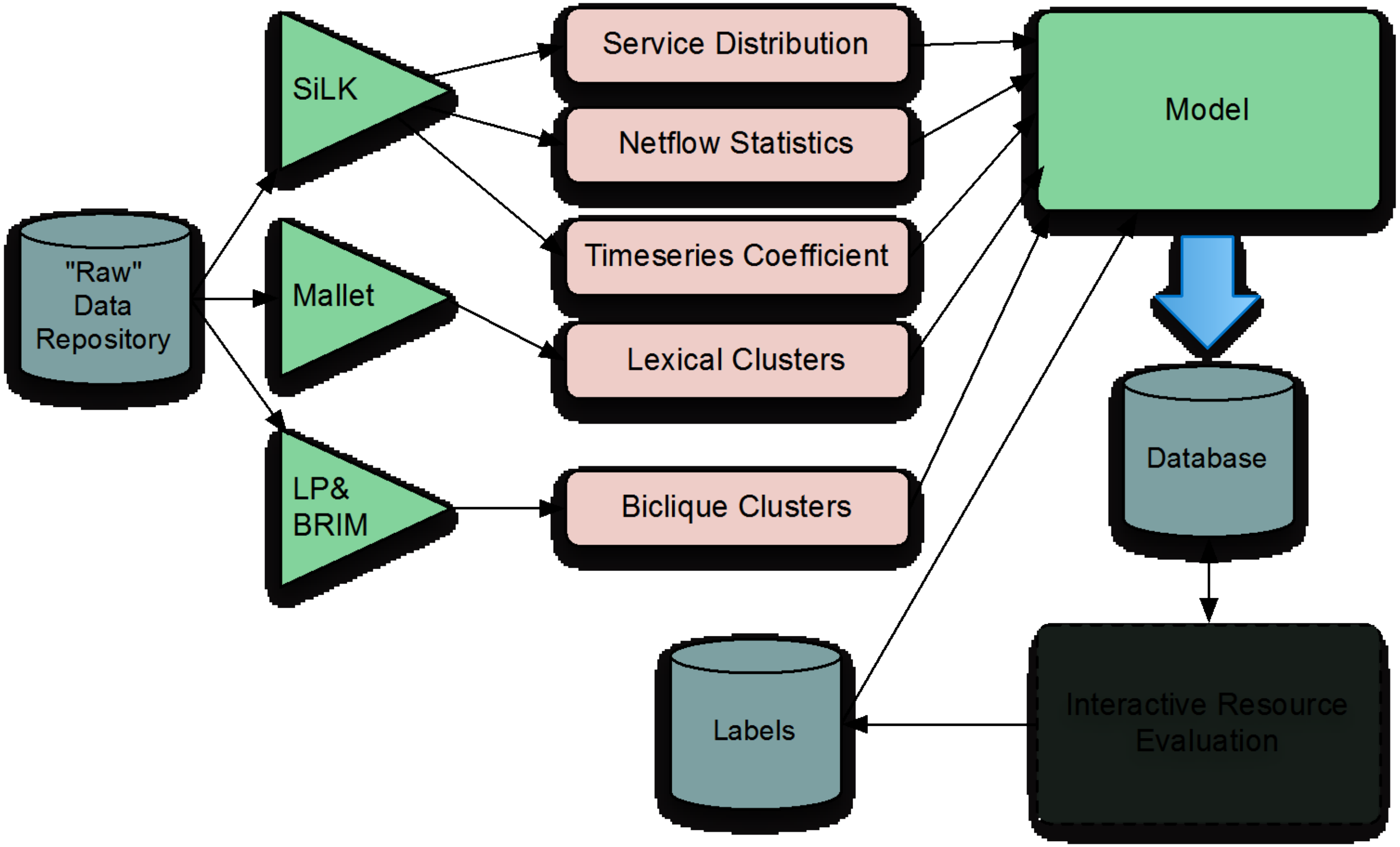
- Bi-clique Grouping
  - Red = Internal
  - Green = External
  - Edges pruned
  - LP & BRIM Algorithm**



**Liu, Xin, and Tsuyoshi Murata. "Community detection in large-scale bipartite networks." *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*. Vol. 1. IET, 2009.

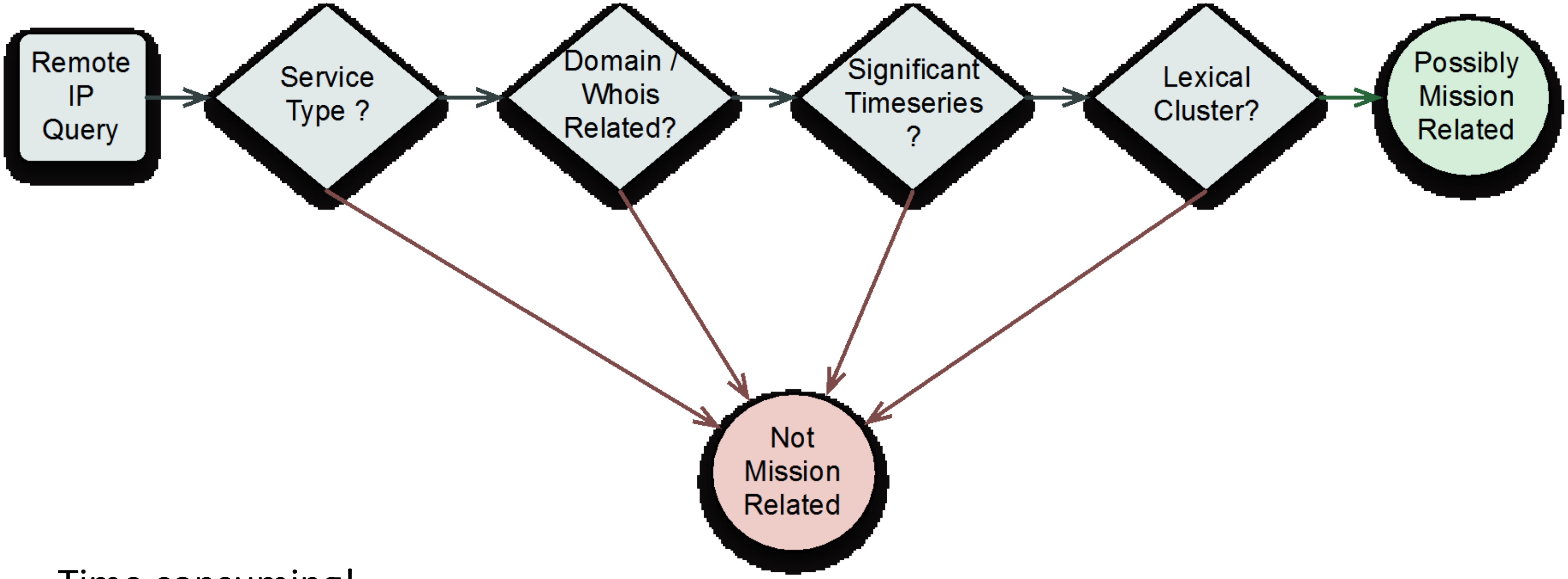
*Gephi http://gephi.org/

# Architecture Overview



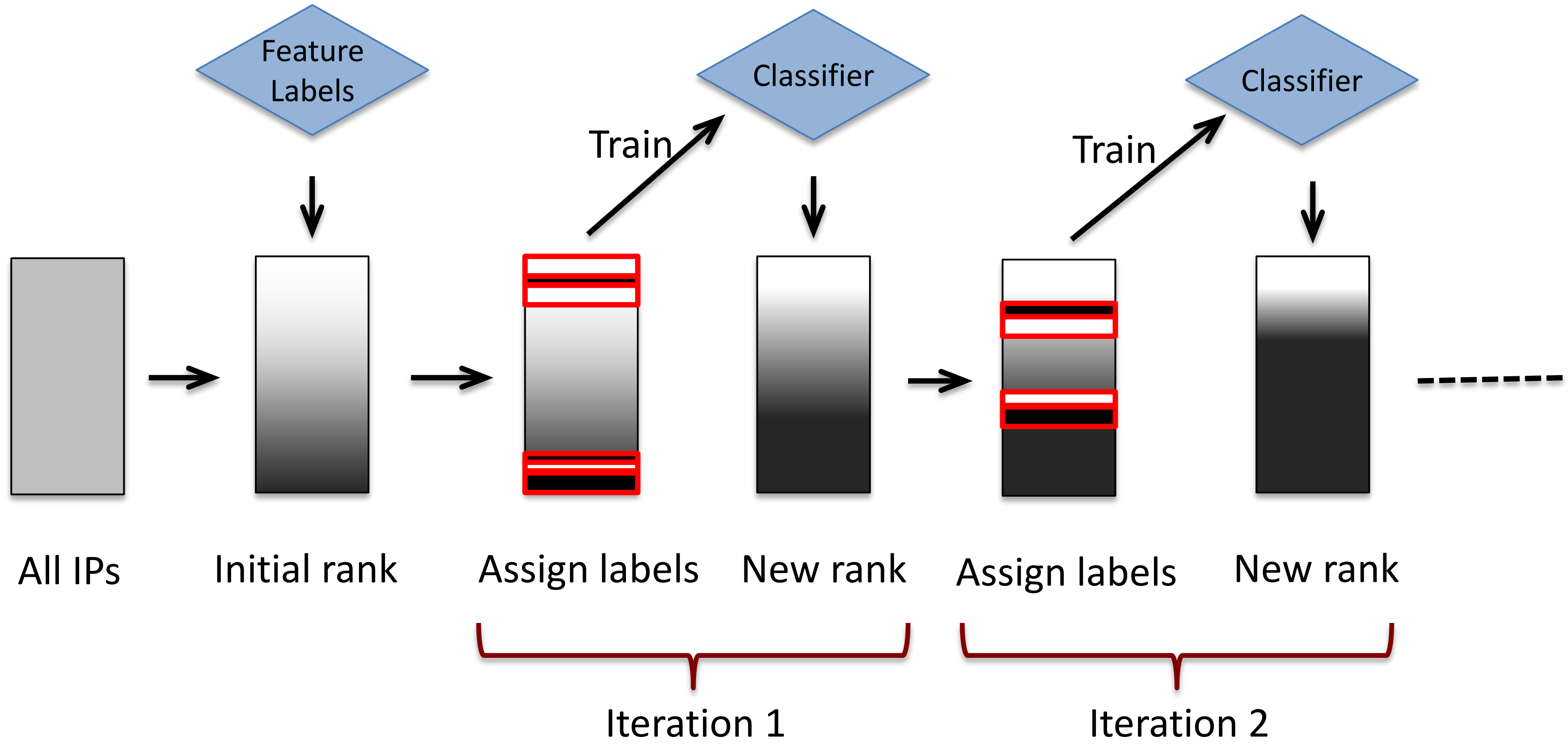
# How to Label / Train

Anecdotal Human Process



Time consuming!

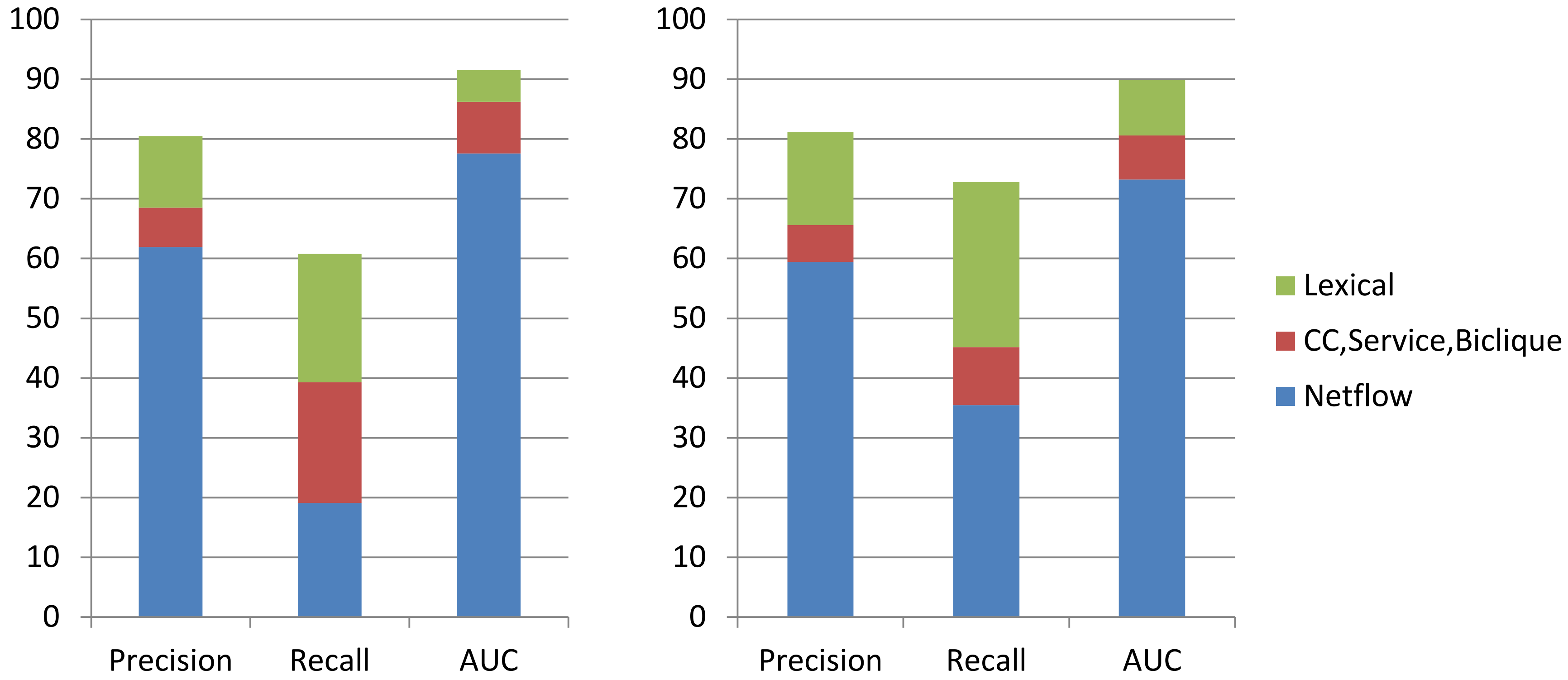
# Kick Start Labeling



# Anecdotal Validation – Ames Data

- Gathering Data
  - One month of NetFlow data in Ames Lab
- Preprocessing
  - 4 sets of features: simple NetFlow statistics, time series features, lexical analysis features (document topic distributions), biclique community features
- Labeling
  - 4242 IPs (801 white / 3441 black)
- Testing / verifying classifier
  - Weka (Logistic Regression, SVM, Bayesian Network, Decision Tree)
  - 10 cross-fold validation

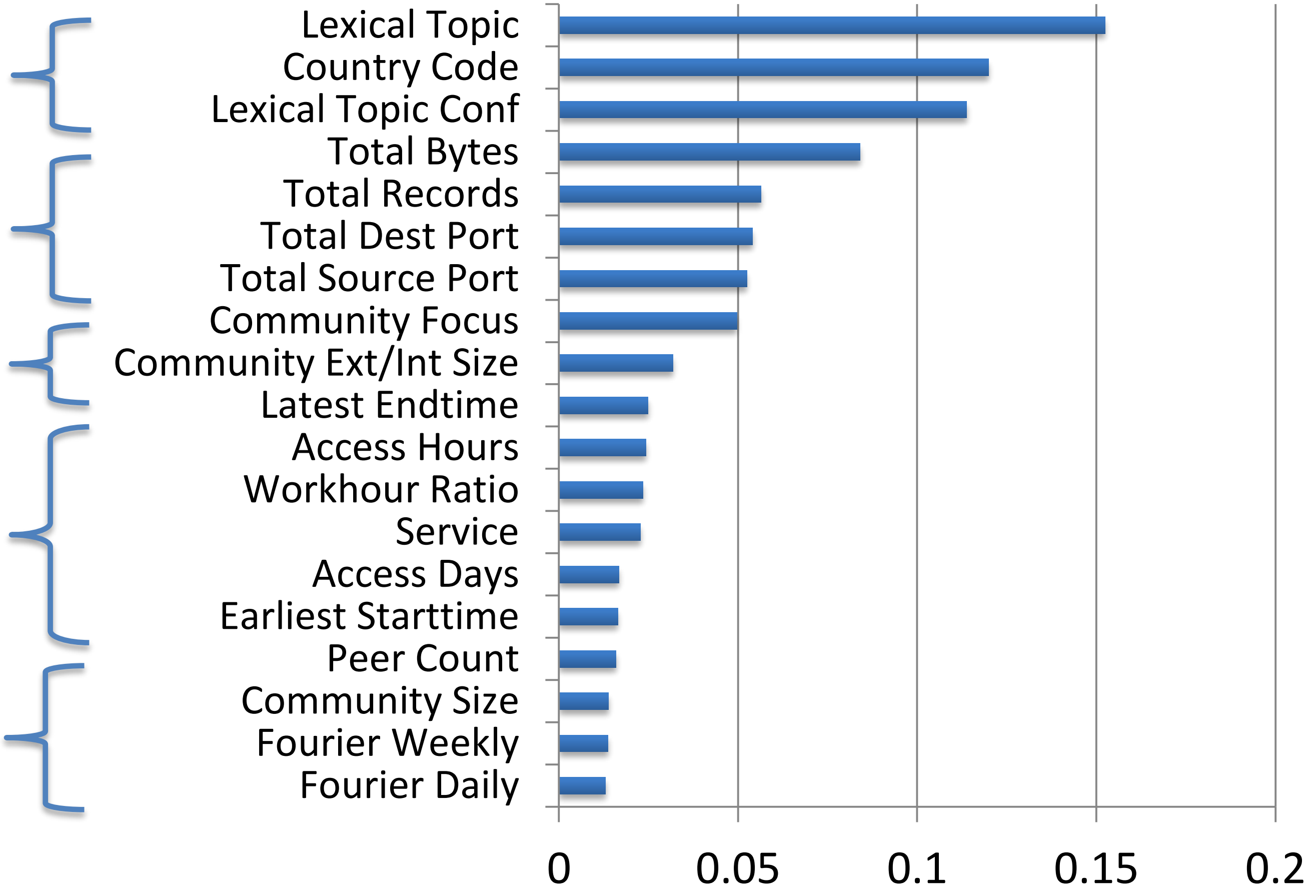
# Performance Results

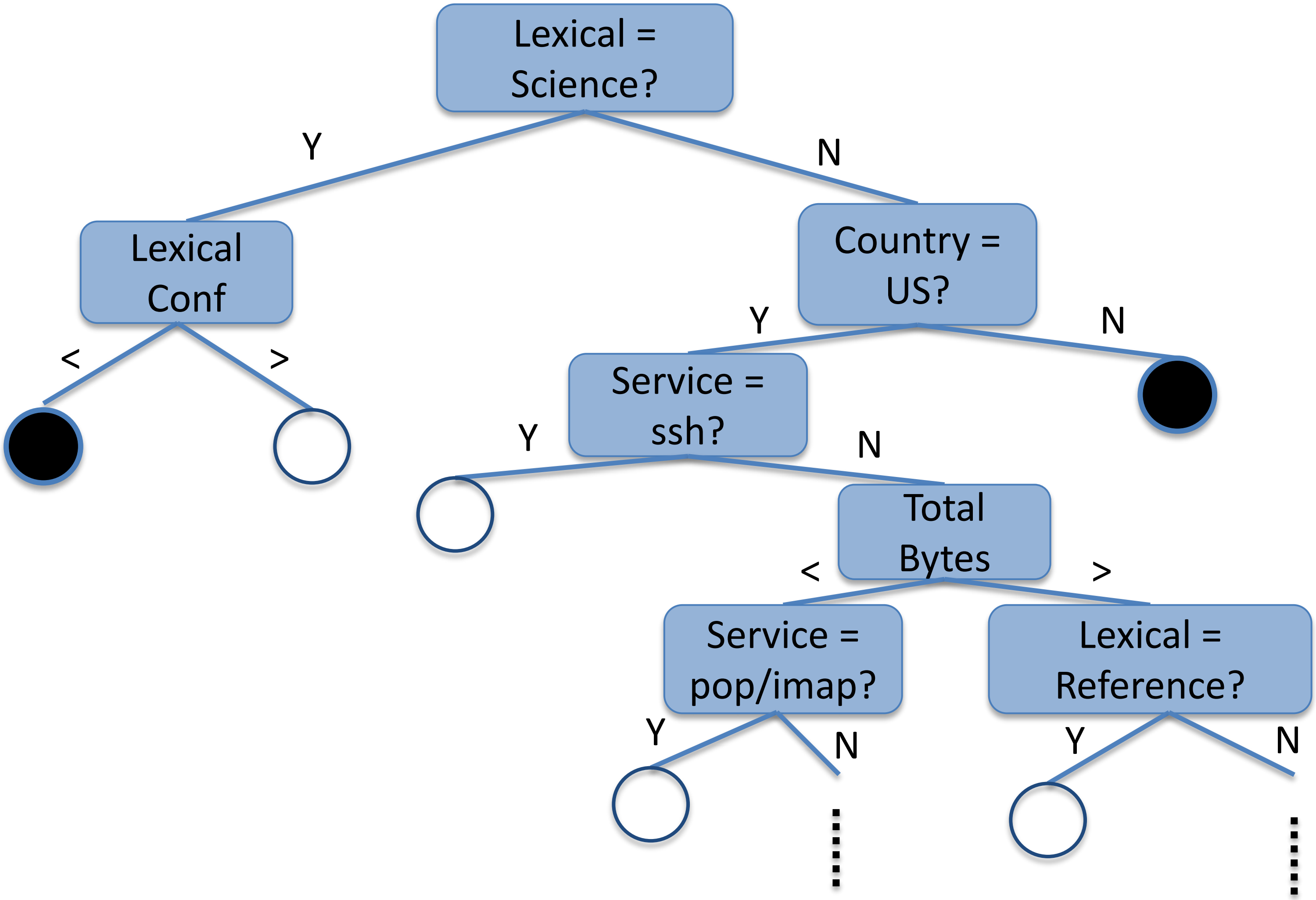


Logistic Regression

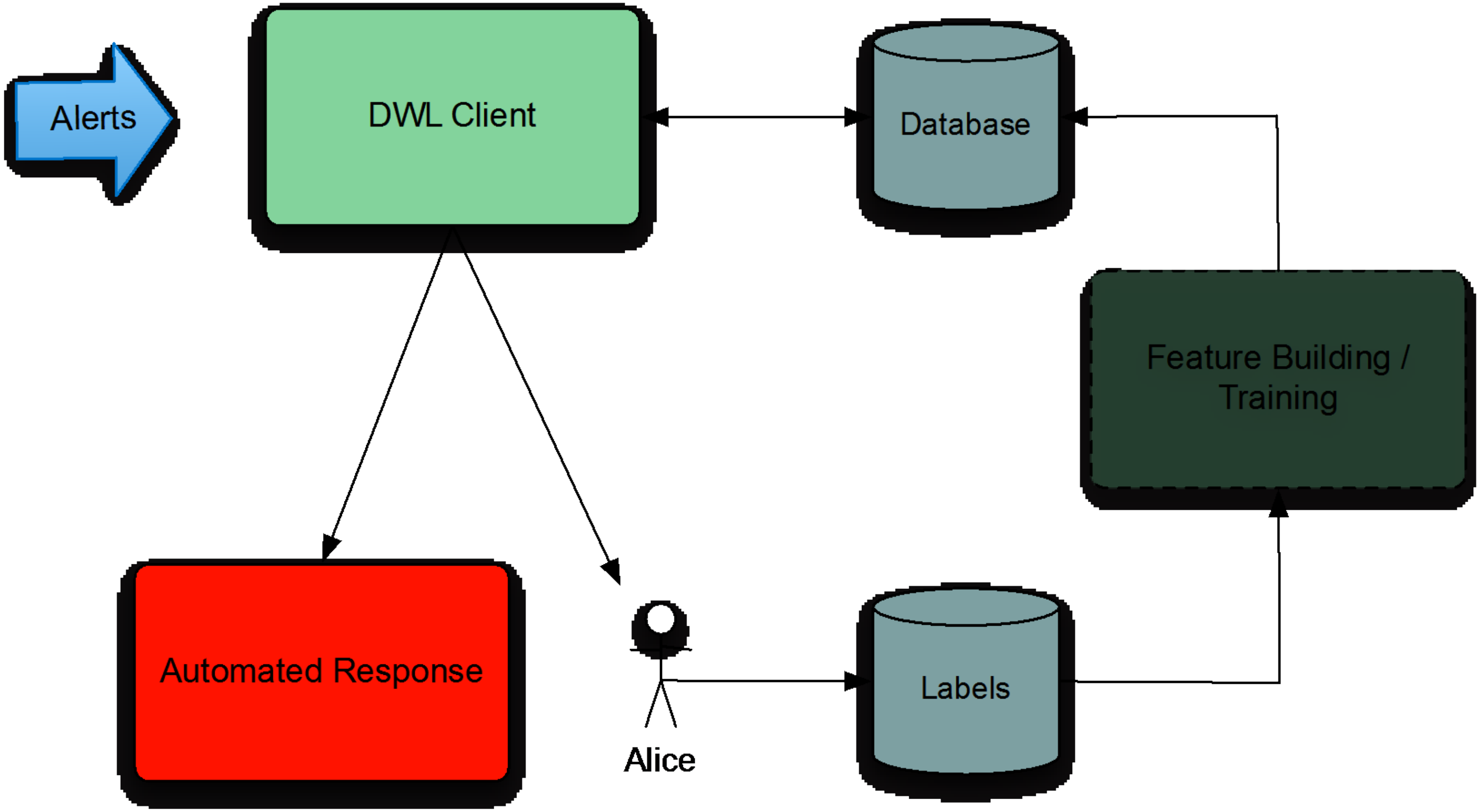
Decision Tree (C4.5)

# Info Gain by Features




Lexical = Science?
Y
N
Lexical Conf
<
>
Country = US?
Y
N
Service = ssh?
Y
N
Total Bytes
<
>
Service = pop/imap?
Y
N
Lexical = Reference?
Y
N

# Implementation at Ames Laboratory



# Challenges / Future Work

- Majority of IPs don’t have a web page
  - Automated query for WHOIS Organization
  - Use of AMP data; actual HTTP resources
- Speed / Streaming
  - Slow to gather features; currently batched daily.
- Searching
  - Search engines w/ free API (Faroo?)
- Production ‘burn-in’
  - Feedback from analysts into a growing set of labels
- Integration with other systems
  - BroIDS Module?
- Mining of graphical data
  - Second derivative clusters (clusters of clusters)
  - Internal resource categorization

# Summary

- Flow provides ‘how much’; a bit of semantics is required for mission relevance.
- Public tools:
  - SiLK – Flow Statistics
  - Crawler4J + Mallet – Lexical Analysis
  - Weka – Machine Learning SAK
  - Apache Commons Math – (Timeseries transforms)
  - A sprinkle of Java and a dash of Python