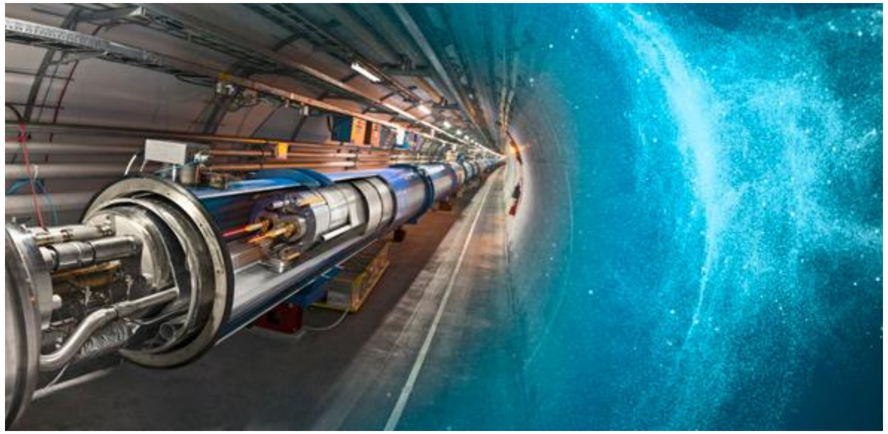
Flowzilla: A Methodology for Detecting Data Transfer Anomalies in Research Networks

Anna Giannakou, Daniel Gunter, Sean Peisert

BERKELEY LAB

Research Networks

- Scientific applications that process large amounts of data
- Frequent large data transfers between endpoints
- Research networks get attacked, just like every other information system



Anomaly-based Intrusion Detection

- How do we detect attacks/intrusions?

- Signature-based intrusion detection (known attacks)

- Anomaly-based intrusion detection (novel attacks)
    - *Anomaly: **Significant** deviation from **normal** profile*
    - Original idea from Dorothy Denning in 1986

- How do we perform anomaly intrusion detection?

Anomaly-based Intrusion Detection Limitations

- Bad news ☹
  - Defining a normal profile is hard
    - Too many individual sessions, unpredictable behavior
    - Feature distributions are very dynamic *(e.g. packet sizes, IP addresses, session size, duration, volume, payload patterns, etc)
    - Generic internet traffic exhibits high variability
  - Lack of ground truth

- Too many false positives
  - Not very operationally appealing

- Are the detected events real anomalies?

- A.K. Marnerides, D.P. Pezaros, D. Hutchison, “Internet traffic characterisation: Third-order statistics & higher-order spectra for precise traffic modelling”, in Computer Networks, Volume 134,2018
- R. Sommer and V. Paxson, “Outside the Closed World: On Using Machine Learning for Network Intrusion Detection,” in Proceedings of the 31st IEEE Symposium on Security and Privacy May 2010

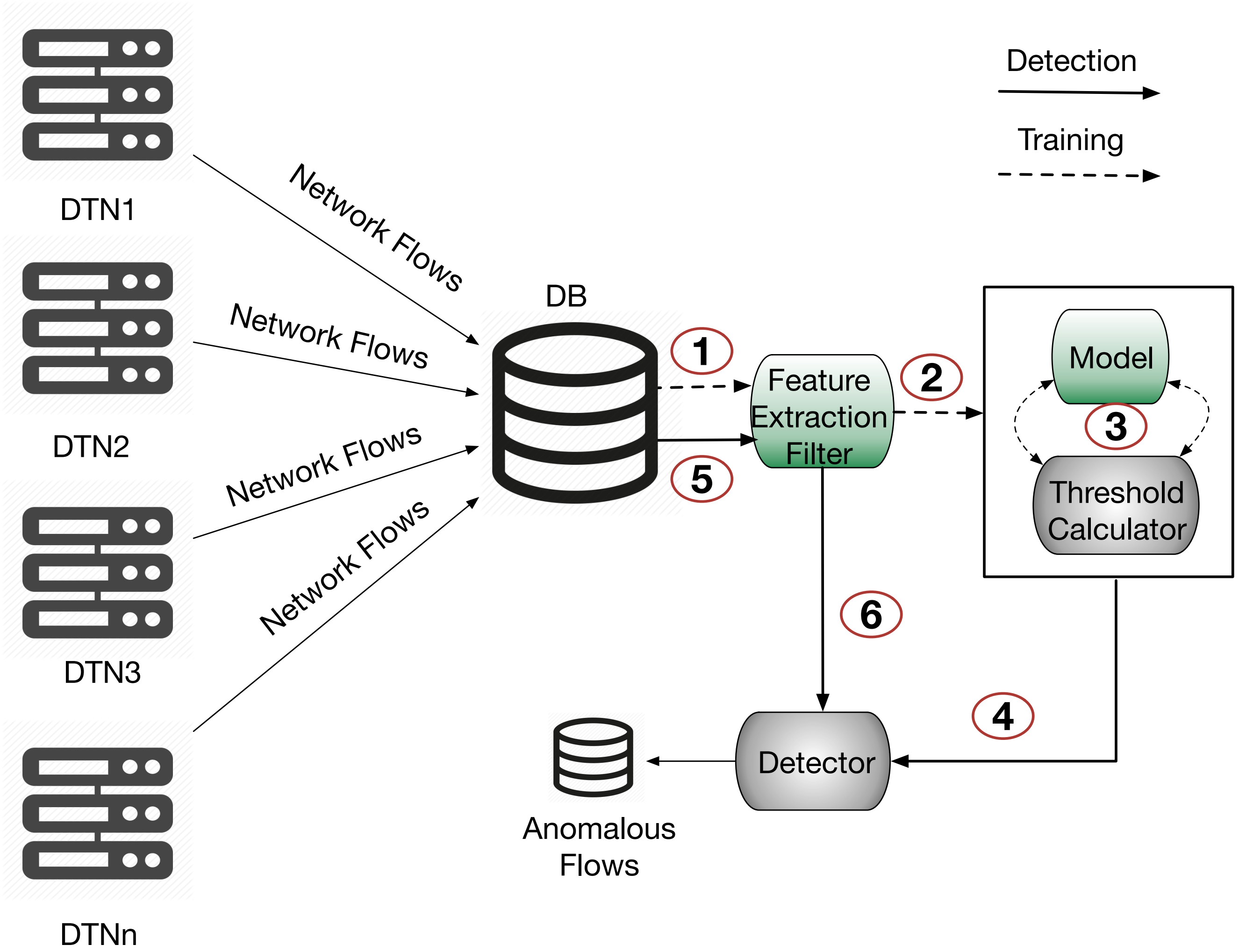
Anomaly-based Intrusion Detection Feasibility

- Is anomaly detection feasible?
  - Our hypothesis is that anomaly-based intrusion detection is feasible if…
    - Requires network domain with lower feature variability
    - Easier to establish reliable normal profile
    - Easier to detect deviations
  - Good candidate domain:
    - Research Networks

Flowzilla: General Principles

- Technique for detecting anomalies in traffic volumes
  - Significant changes in the size of scientific data transfers
- Use machine learning to establish normal profile
  - Train on past data transfers
- How do we define significant?
  - With an adaptive technique for establishing a threshold

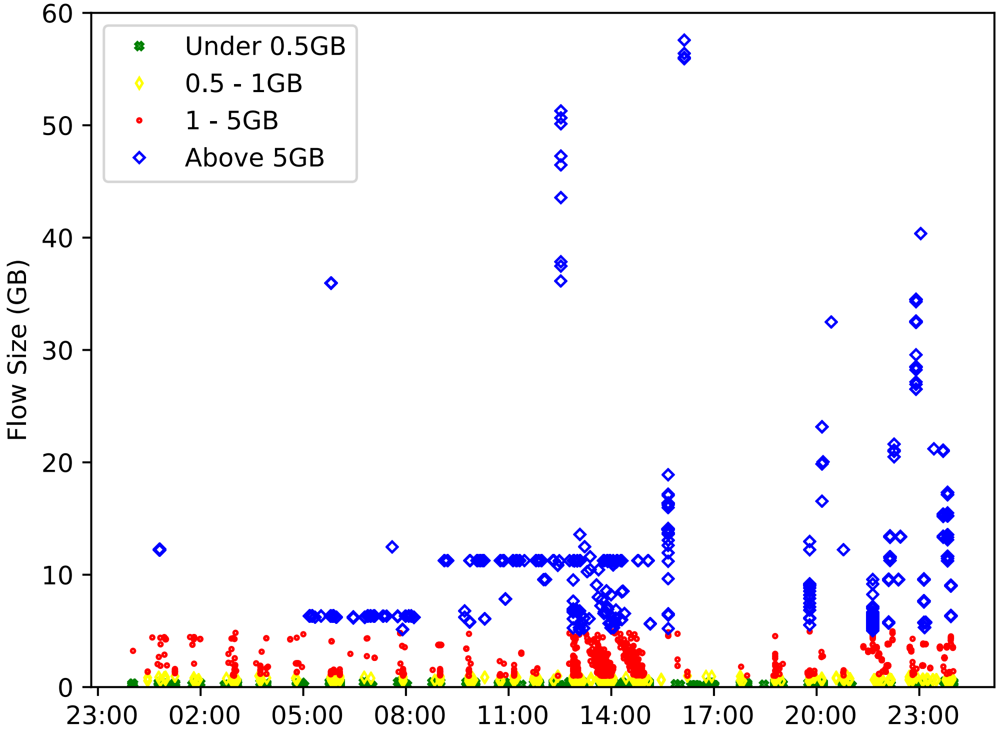
Flowzilla: Architecture

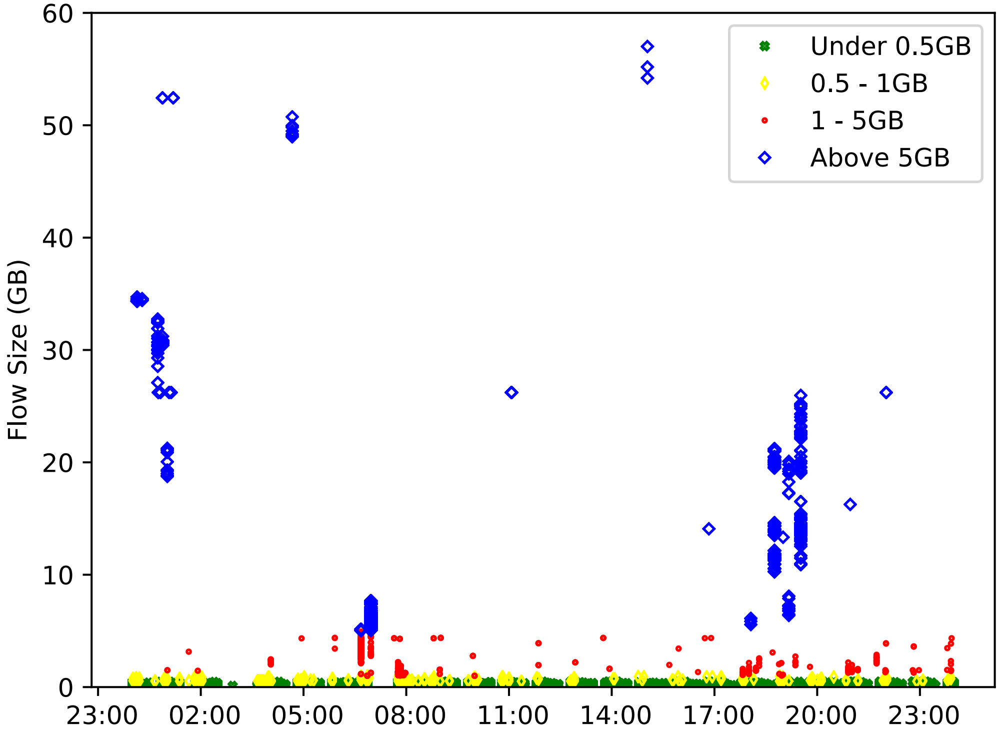


BERKELEY LAB

Flowzilla: Adaptive Threshold

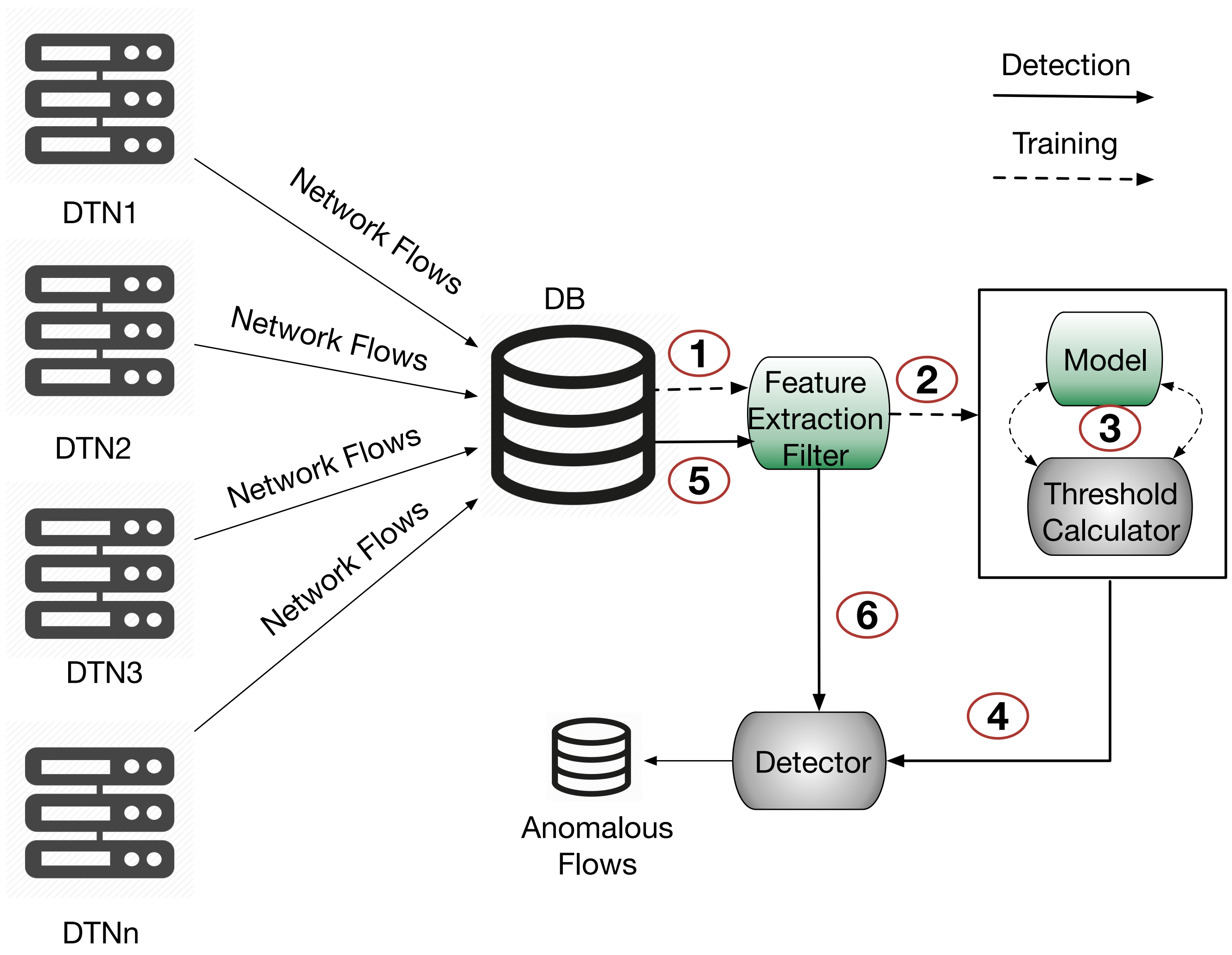
- Threshold definition can be tricky
  - Too high → False negatives
  - Too low → False positives
  - Constant value does not account for seasonal trends





Data transfers are one week apart

Flowzilla: Architecture



Flowzilla: Components

- Model
  - Predict flow size based on:
    - Network throughput
    - Flow duration
    - Source/Destination IP
- Threshold calculator
  - V_{real} real traffic volume
  - V_{pred} predicted traffic volume
  - $\mu =$ mean value of $|V_{real} - V_{pred}|$
  - $T = \mu + x$
  - x such that 90% of the flows are legitimate

Flowzilla: Training

- Training Data
  - Flows from 10 NERSC DTNs
    - Flows between 10/01/2017 - 11/30/2017
    - More than 350,000 flows
  - Collected through tstat
  - Originally 52 features per flow
  - Feature Extraction Filter to extract necessary fields for model training

Flowzilla: Evaluation

- Questions:
  1. How well does Flowzilla detect volume anomalies?
  2. Does it detect anomalies regardless of size/time of occurrence?
  3. Does the quality of predictions degrade after a certain time?
- Lack of ground truth in training dataset (which flows are actually malicious?)

Flowzilla: Evaluation

- Insert artificial anomalies of different size
  - Data transfers between Grid5000 nodes and NERSC DTNs

Experiment	# of Nodes	# of Transfers per Node	Transfer Size	Time interval between transfers
1	8	5	1-5 GB	1-60 min
2	8	5	10 GB	60 min

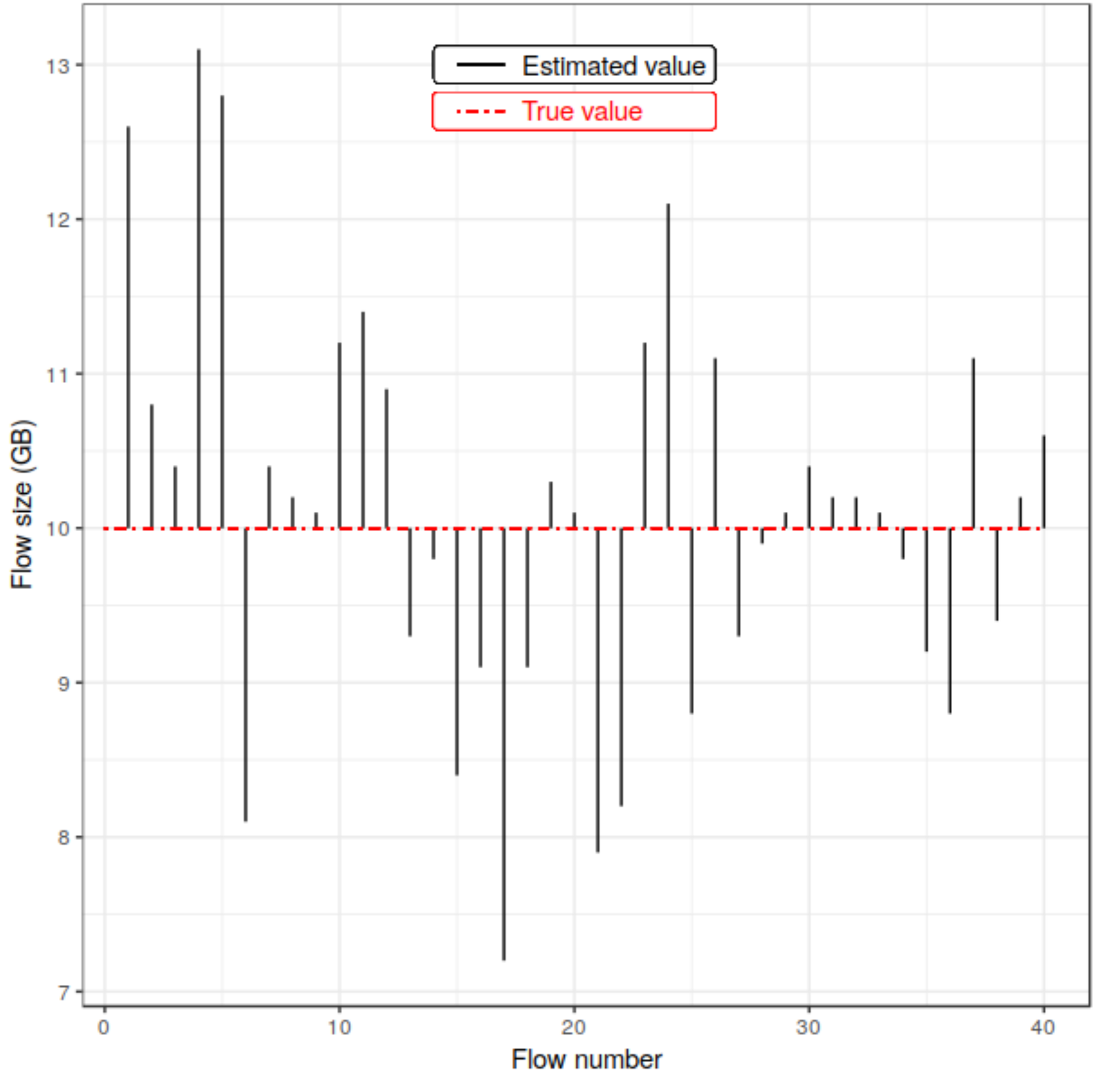
Flowzilla: Detection Results

- Model trained on data transfers between 01/10/2017 – 30/11/2017

Experiment	Total Anomalies	Anomaly Size	True Positives	False Negatives	Total # of Flows
1	40	1-5 GB	34	6	12810
2	40	10 GB	37	3	30595

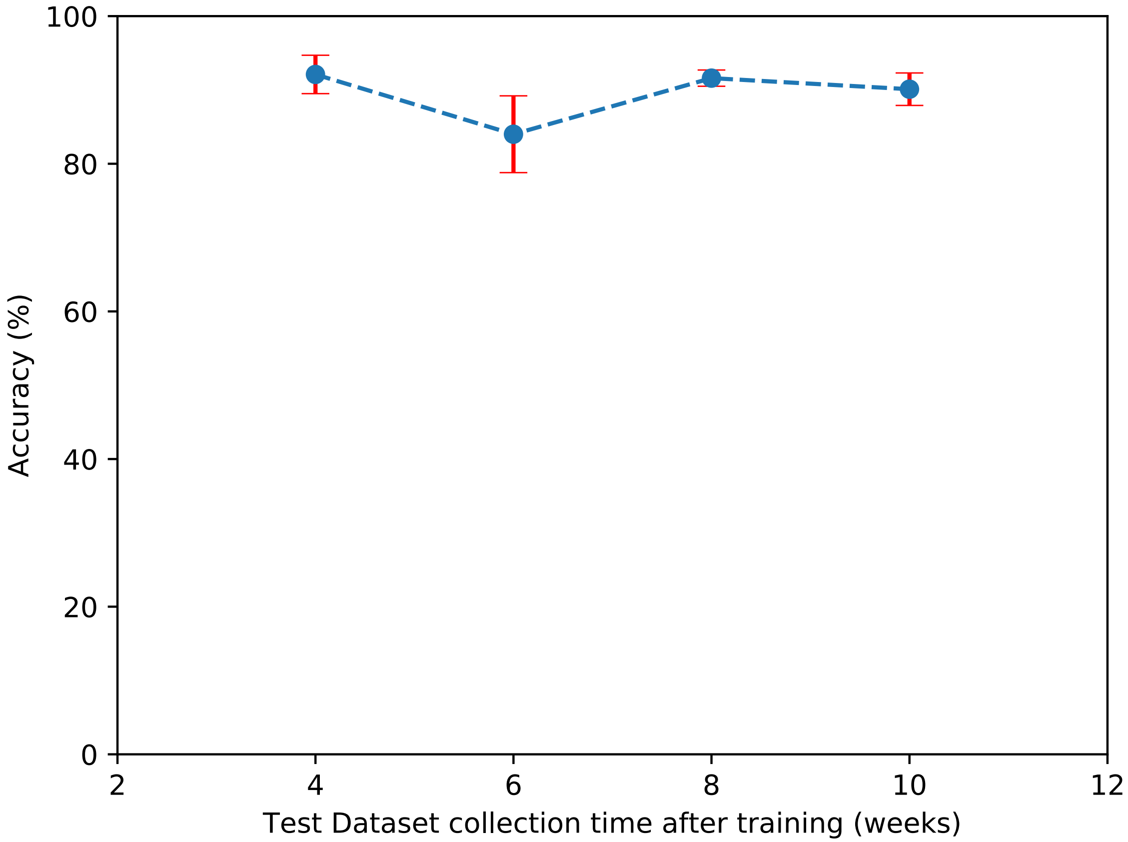
- Detection rate remains above 80% in both experiments

Flowzilla: Anomalous Flows Size Prediction



BERKELEY LAB

Flowzilla: Quality of Prediction Weeks after Training



- Accuracy remains above 80% even 10 weeks after training

Flowzilla: Conclusion

- We have developed a technique for detecting volume anomalies in network transfers on research networks using machine learning
- Adaptive thresholds are predictably helpful for reducing false positives
- Acceptable detection rate (up to 92.5%)
- Model is temporally stable in predicting scientific flow sizes

Flowzilla: Future Work

- Expand to other types of anomalies
- Detect anomalies that span across multiple flows
- Incorporate additional tstat metrics in our prediction
- Experiment with different retraining strategies (confidence intervals)

Questions?