Machine Learning for Network Traffic Analysis and Scientific Metadata Extraction

Anna Giannakou

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

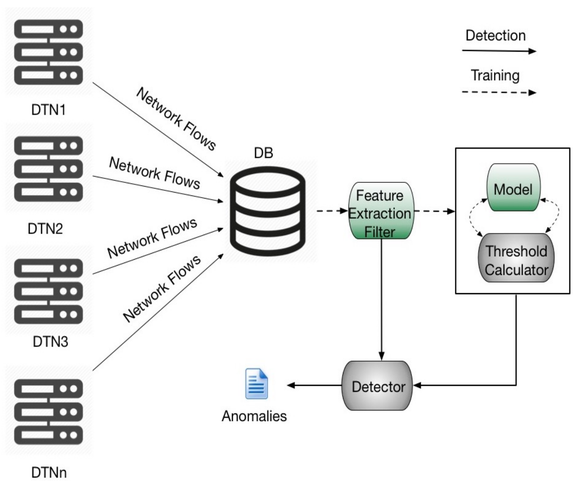
Flowzilla: A Methodology for Detecting Data Transfer Anomalies in Research Networks

Problem Statement

- Research networks enable large data transfers between endpoints
- Research networks get attacked just like any other network
- Establishing a profile for “normal” network behavior is hard
- Lack of ground truth generates many false positives

Approach

- Focus on volume anomalies, unexpected changes in the volume of data transfers
- Use machine learning to establish notion of normality (**Random Forest Regression**)
- Detect candidate anomalies (i.e. outliers) based on **distance** from normal profile
- Calculate distance based on adaptive threshold mechanism
  - Threshold definition is tricky
  - Too low → Increases the # of false negatives
  - Too high → Increases the # of false positives
  - Constant value does not account for seasonal trends
- Adapt threshold value based on detection results

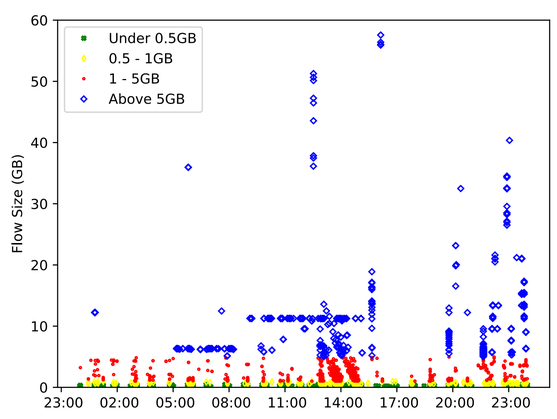


Flowzilla’s components for model training, threshold calculation and detection of anomalous flows

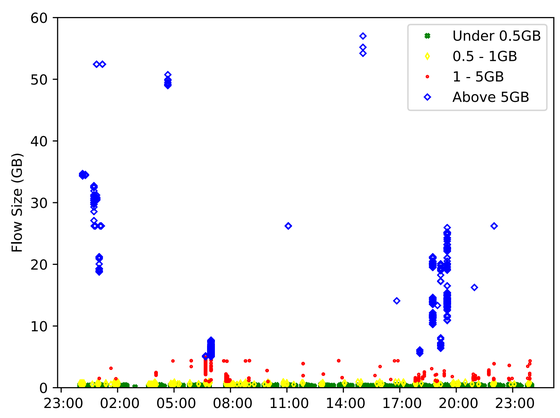
Next Steps

- Expand to other types of anomalies
- Detect anomalies that span across multiple flows

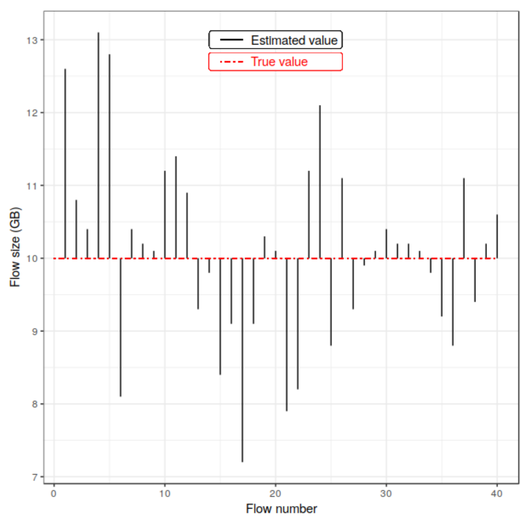
Evaluation - Results



Volume of data transfers on 29/05/2018



Volume of data transfers on 05/06/2018



Flowzilla’s prediction on anomalous flows sizes

- **Detection rate up to 92.5%**
- **Accuracy above 80%** in predicting anomaly sizes

Acknowledgements

Special thanks to Daniel Gunter and Sean Peisert. This work is part of Scientific Integrity for Exascale Scientific Data project supported by the Office of Advanced Scientific Computing Research (ASCR) program under contract number DE-AC02-05CH11231.

Giannakou Anna, Gunter Daniel, Peisert Sean. (2018). Flowzilla, A Methodology for Detecting Data Transfer Anomalies in Research Networks . INDIS

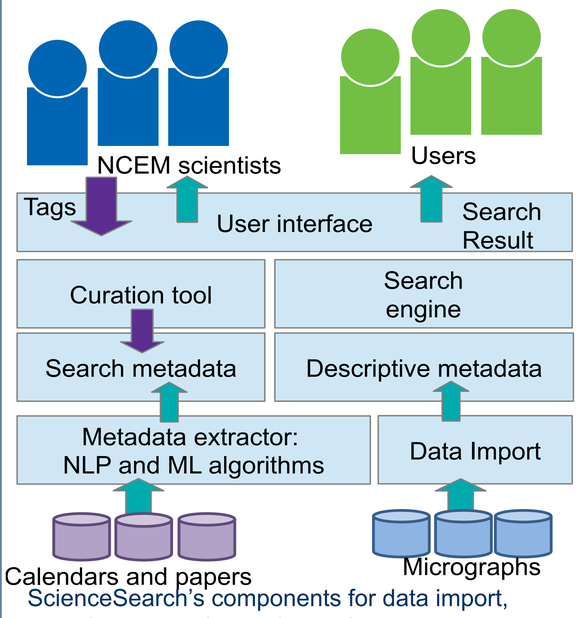
ScienceSearch: Enabling Search through Automatic Metadata Generation

Problem Statement

- Scientific discovery depends on exploring large amounts of data
- Adding the right metadata to large datasets is is done manually
- Automatic metadata creation requires tuning

Approach

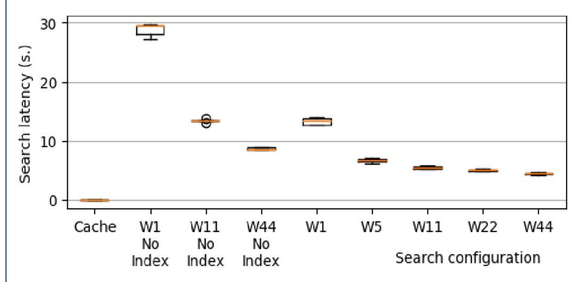
- Automate scientific data exploration through scalable search engine
- Incorporate user feedback in result verification and metadata analysis recalibration
- ScienceSearch features four components: **data import, metadata extraction, search engine** and **user feedback**
  - Data import: identify searchable data and related metadata
  - Metadata extraction: generate metadata that characterize searchable data. Metadata include information about the purpose and production conditions (e.g. experiment) of the data.
  - Search engine: enables search based on the generated metadata based on a hit score. Hit score represents the similarity between the search query and the metadata.
  - User feedback: users can modify and invalidate existing metadata or add new ones through a specific **tagging tool.**
- Metadata is extracted by using NLP tools **(TextRank, Spacy, Tf-idf).** A curated **feature detection model** describes relationships between data and metadata



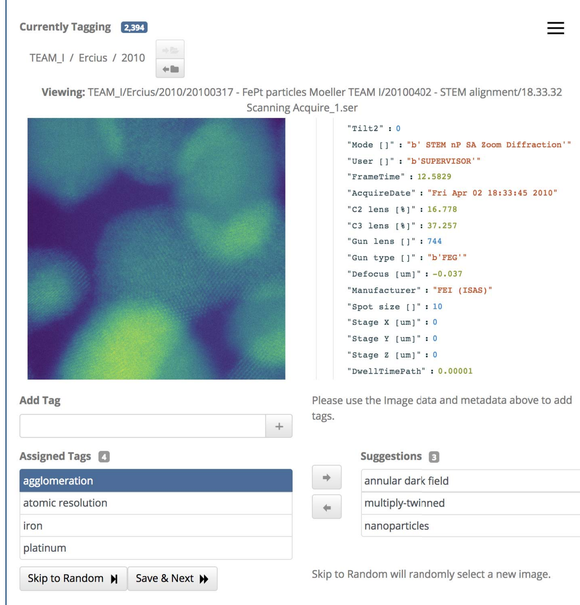
ScienceSearch’s components for data import, metadata extraction and search

Evaluation - Results

Measure search result latency for cached and not cached queries
Approve/invalidate generated tags through tagging tool



Search latency of a query is under 30s (worst case scenario). Parallelizing the hit score calculation and hit aggregation reduces latency up to 2/3 (under 10s).



ScienceSearch’s user tagging tool

Next Steps

- Expand metadata generation to other sources (e.g. file system structure) using machine learning **(unsupervised clustering, weighted clustering)** and natural language processing **(named entity recognition, chemical element extraction, word embedding creation)**
- Use deep learning to assign value scores to new metadata sources (**weighted classifier**)
- Explore relationship between data based on file system structure

Acknowledgements

Special thanks to Lavanya Ramakrishnan, Gonzalo Rodrigo, Matt Henderson, Gunther Weber, Colin Ophus and Katie Antypas.

This work is supported by the Office of Advanced Scientific Computing Research (ASCR) program under contract number DE-AC02-05CH11231

P. Rodrigo, Gonzalo & Henderson, Matt & Weber, Gunther & Ophus, Colin & Antypas, Katie & Ramakrishnan, Lavanya. (2018). ScienceSearch: Enabling Search through Automatic Metadata Generation. eScience

U.S. DEPARTMENT OF ENERGY

ScienceSearch

CRD