End-to-End Active Learning for Computer Security Experts

Anaël Beaugnon, Pierre Chifflier, Francis Bach

anael.beaugnon@ssi.gouv.fr

ANSSI, ENS Paris, INRIA

AICS 2018

Outline

Outline

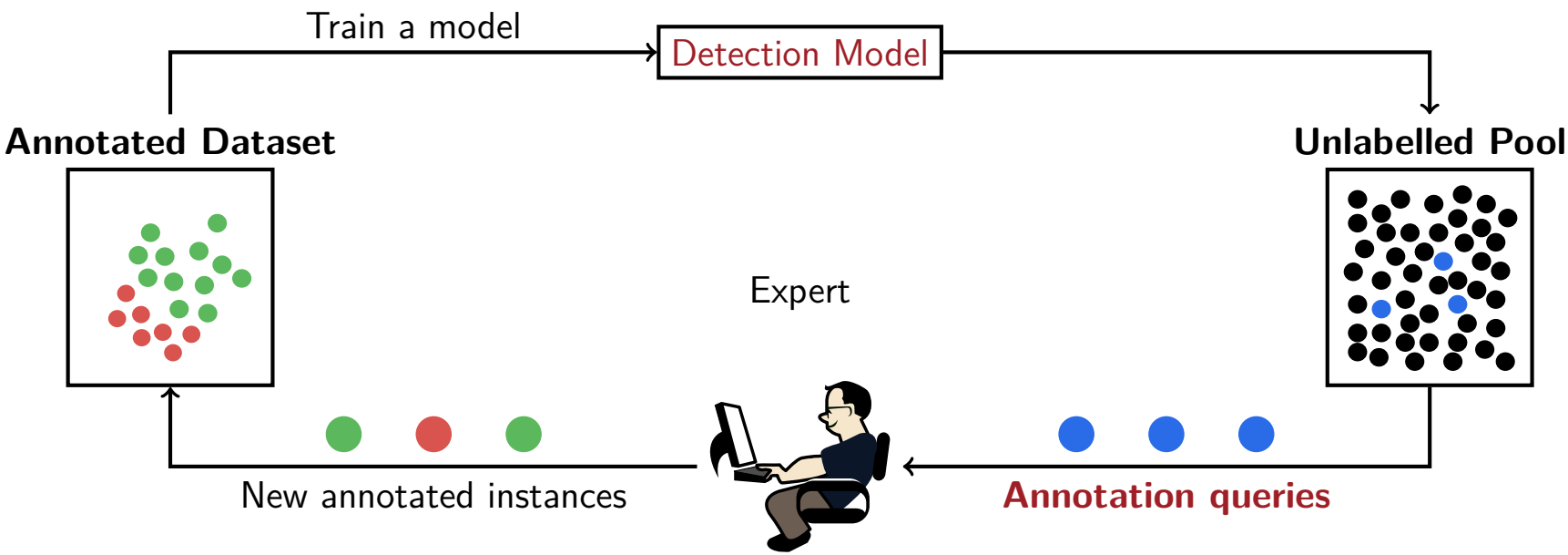
Lack of Representative Training Data !

- ✗ Public datasets $\neq$ deployment environments
- ✗ Crowd-sourcing is not suited for Computer Security

In-situ labelling with Active Learning

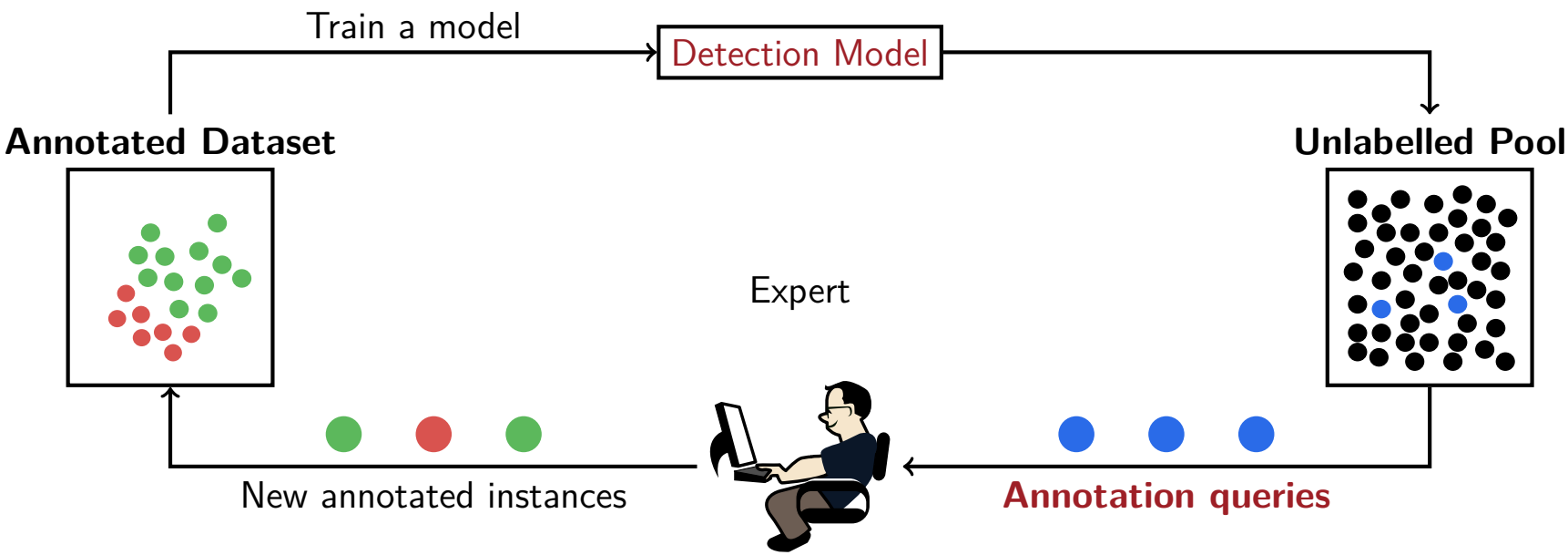
Annotate data from the deployment environment

Active Learning



Active Learning

Don't forget the expert !



Active learning is not only a query strategy ...

Outline

ILAB: an End-to-End Active Learning System

End-to-end active learning system

=

Active learning strategy + Annotation system

Active Learning Strategy

Selects cleverly the instances to be annotated

RAID'17 **Beaugnon et al., ILAB: An Interactive Labelling Strategy for Intrusion Detection**

Annotation System

Displays the annotation queries and gathers the answers

ILAB: an End-to-End Active Learning System

End-to-end active learning system

=

Active learning strategy + Annotation system

Active Learning Strategy

Selects cleverly the instances to be annotated

RAID'17 **Beaugnon et al., ILAB: An Interactive Labelling Strategy for Intrusion Detection**

Annotation System

Displays the annotation queries and gathers the answers

How to assess properly a whole active learning system ?

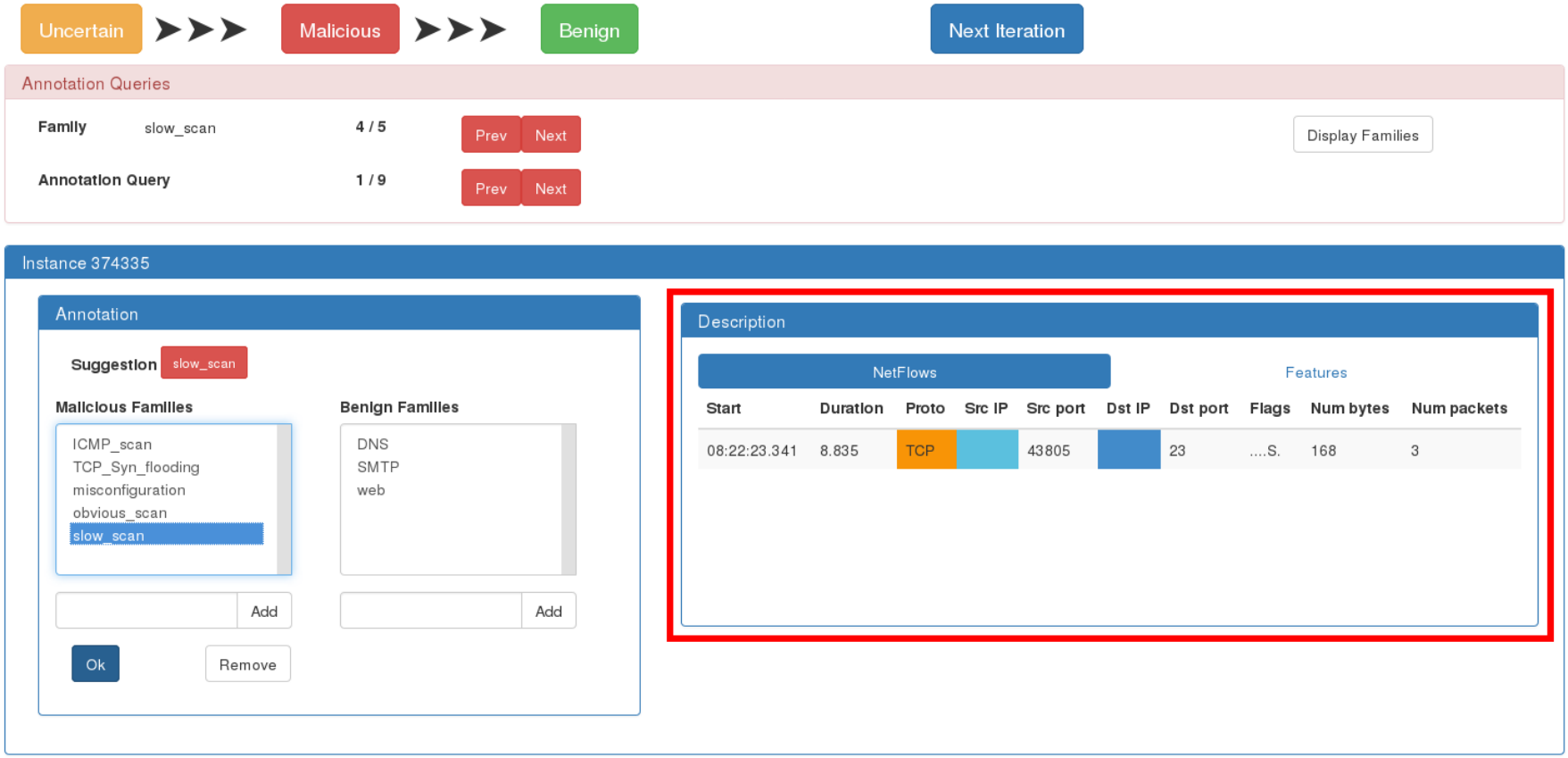
User Experiments

Anomaly detection from NetFlow data

Dataset	
Num. flows	$1.2 \cdot 10^8$
Num. IP	$463,913$
Num. features	134

How does ILAB help experts annotate ?

How does ILAB help experts annotate ?

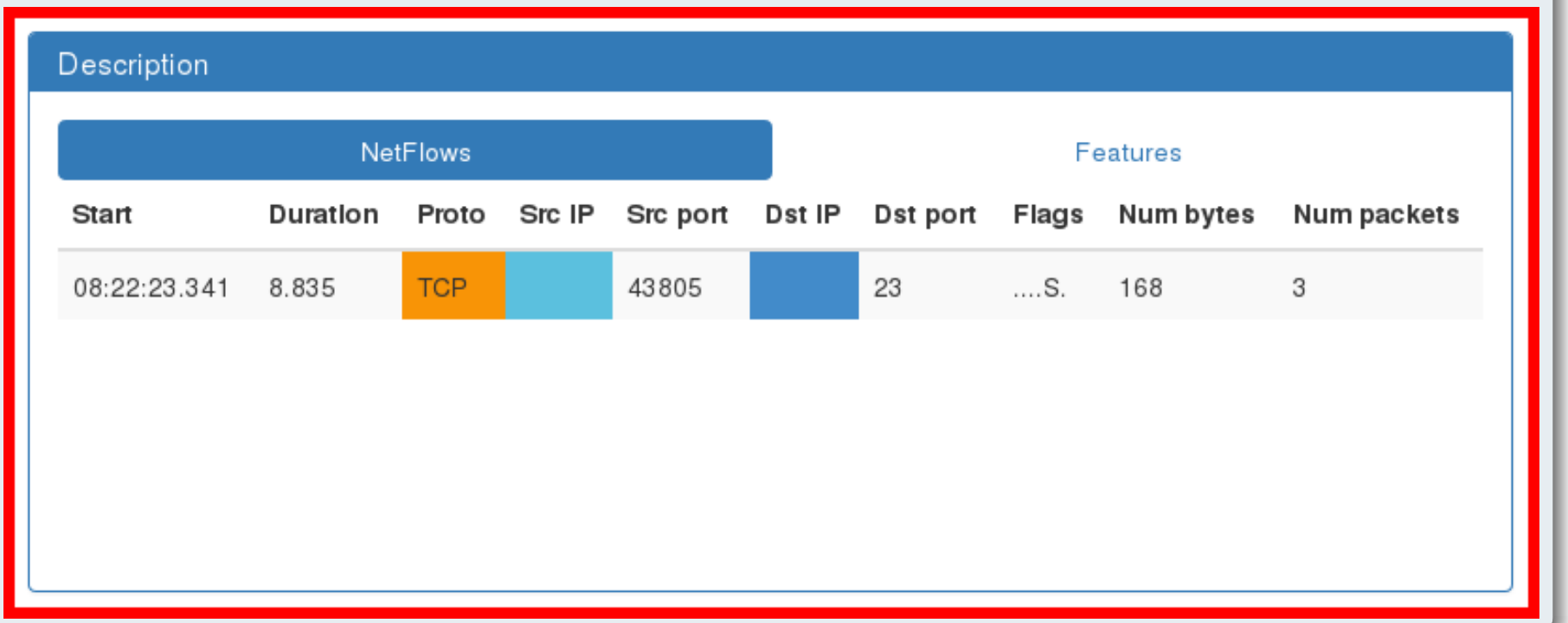


How does ILAB help experts annotate ?

Problem

Need for contextual informations to annotate

Solution: Problem-specific visualization

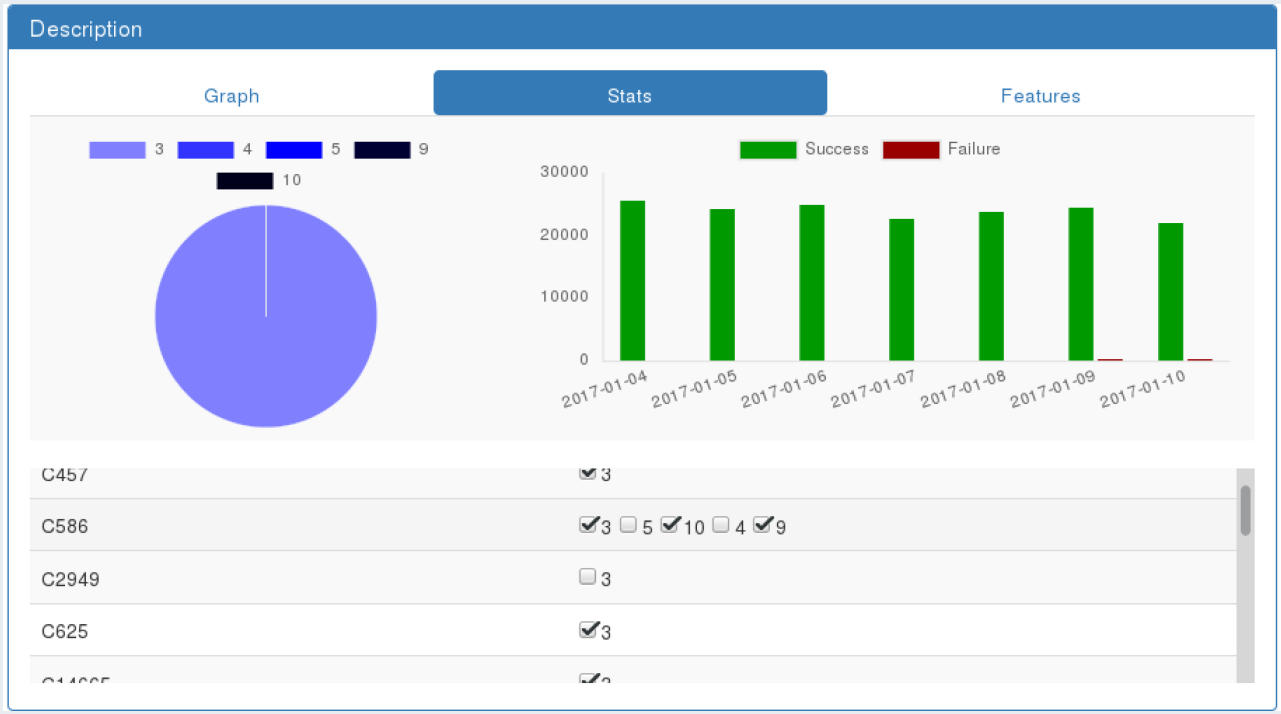


How does ILAB help experts annotate ?

Problem

Need for contextual informations to annotate

Solution: Problem-specific visualization

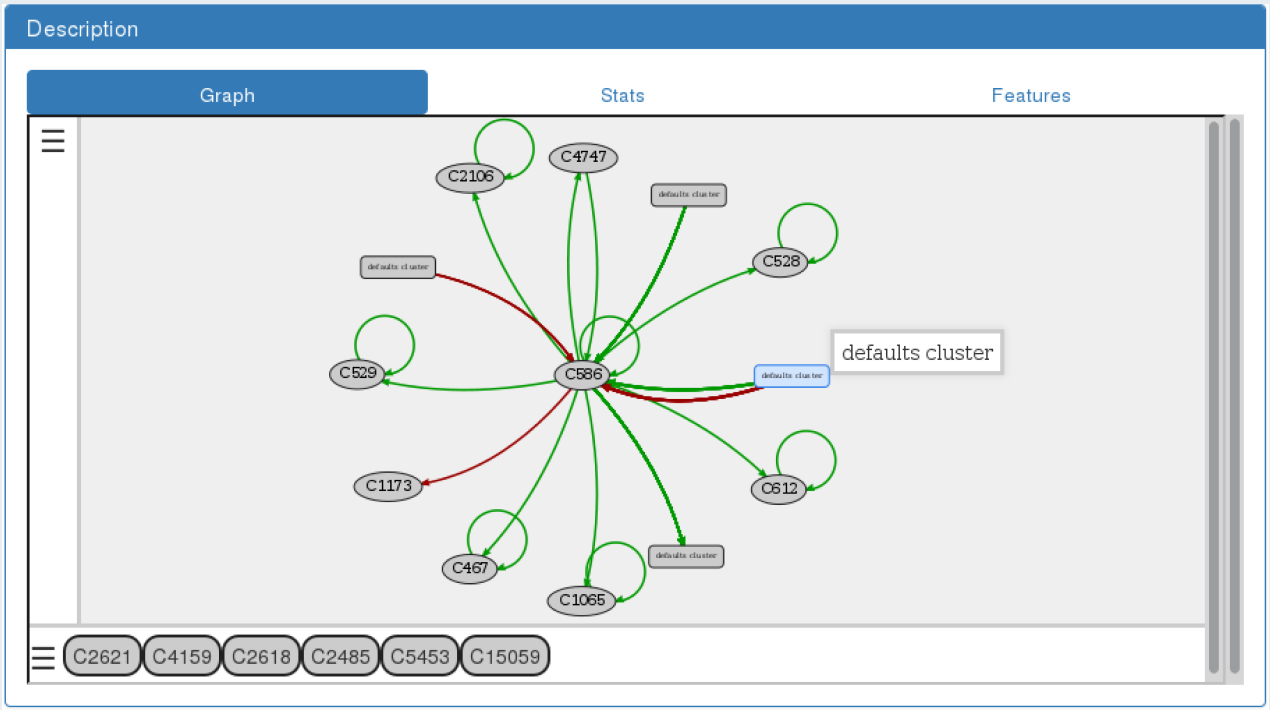


How does ILAB help experts annotate ?

Problem

Need for contextual informations to annotate

Solution: Problem-specific visualization



How does ILAB help experts annotate ?

Problem

- What families should I create ?
- How should I group the data ?

Families definitions may evolve across iterations ...

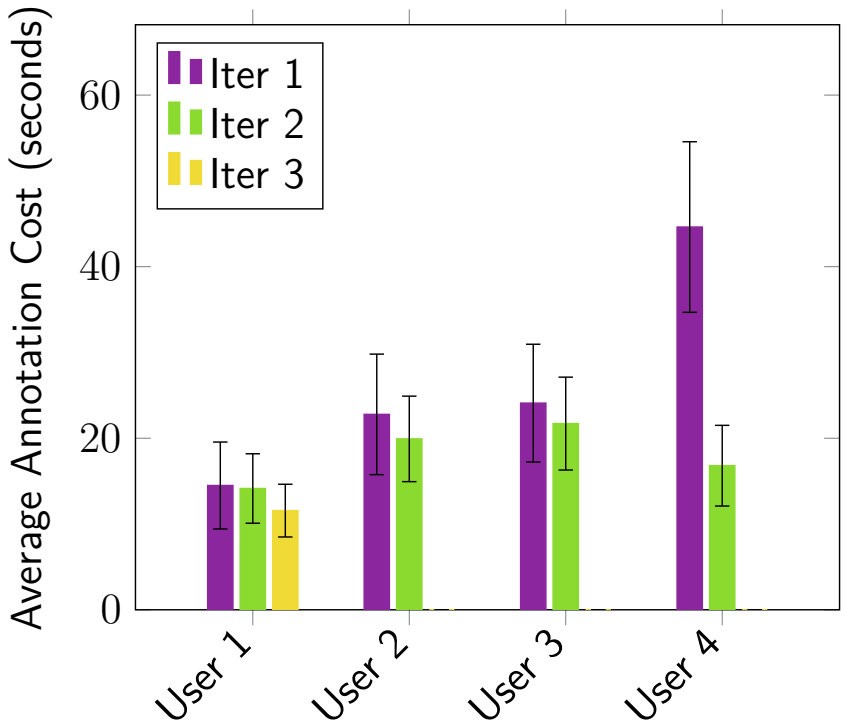
Solution: Family Editor

- Edit the name of a family
- Merge several families
- Change the label associated with a family

It is hard to annotate, but ILAB helps !

Solutions

- Problem-specific visualization
- Family Editor

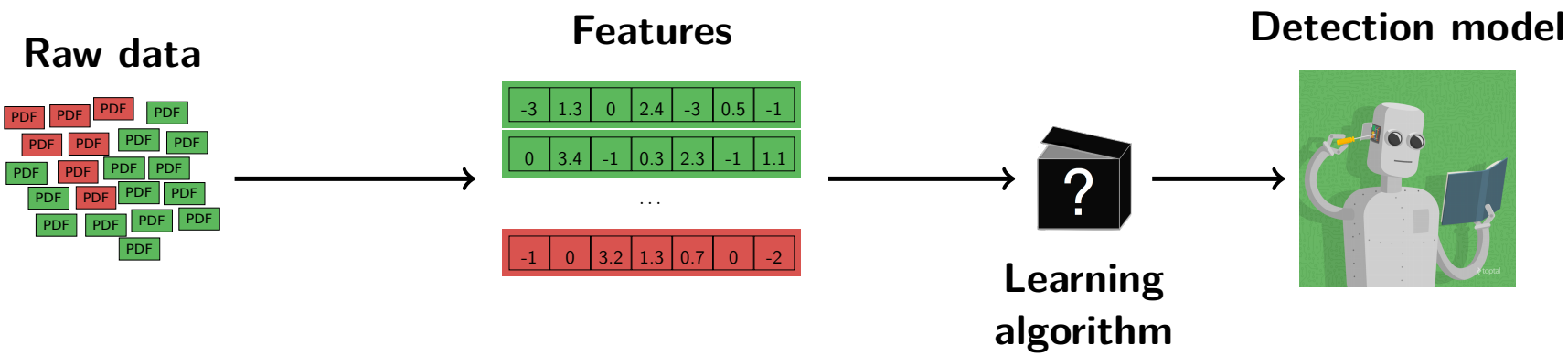


Outline

Importance of features

Problem

Features may be not expressive enough



Loss of information with feature extraction ...

Importance of features

Problem

Features may be not expressive enough

Features

Num. bytes sent/received:

- globally
- on port 80
- on port 53
- on port 25

User annotation

- anomalous
- trafic on port 1258

Importance of features

Problem

Features may be not expressive enough

Solutions

- Annotators must know the extracted features
- Make features evolve across iterations
  - manually
  - or even better, automatically

ECML'14 Boulle, Towards automatic feature construction for supervised classification

DSAA'15 Kanter et al., Deep feature synthesis: towards automating data science endeavors

EURASIP'16 Šrndić et al., Hidost: a static machine learning based detector of malicious files

ILAB: an End-to-End Active Learning System

https://github.com/ANSSI-FR/SecuML

ILAB helps experts to annotate

- Problem-specific visualization
- Family editor

Features Expressiveness

- Annotators must know the features
- Features should evolve across iterations (future work)