CSci 8980

Edge-based Discovery of Training Data for Machine Learning

CMU authors

Deep Learning Recipe

- Collect a large amount of data and label it
- Select a model and train a DNN
- Deploy the DNN for inference

Labelled Data

- Some data are easy to label …

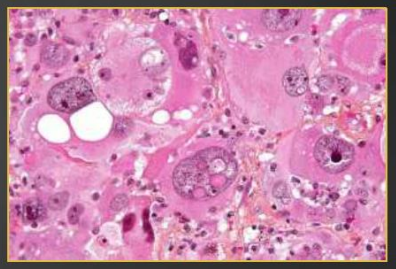
- Some require domain expertise

Valuable in ecology, military intelligence, medical diagnosis, etc.

- Low base rate (prevalence) in the data
- Requires expertise to identify

Masked palm civet (*Paguma larvata*). Transmitter of SARS during its 2003 outbreak in China.

BUK-M1. Believed to have shot down MH17 and killed 298, 2014.



Nuclear atypia in cancer.

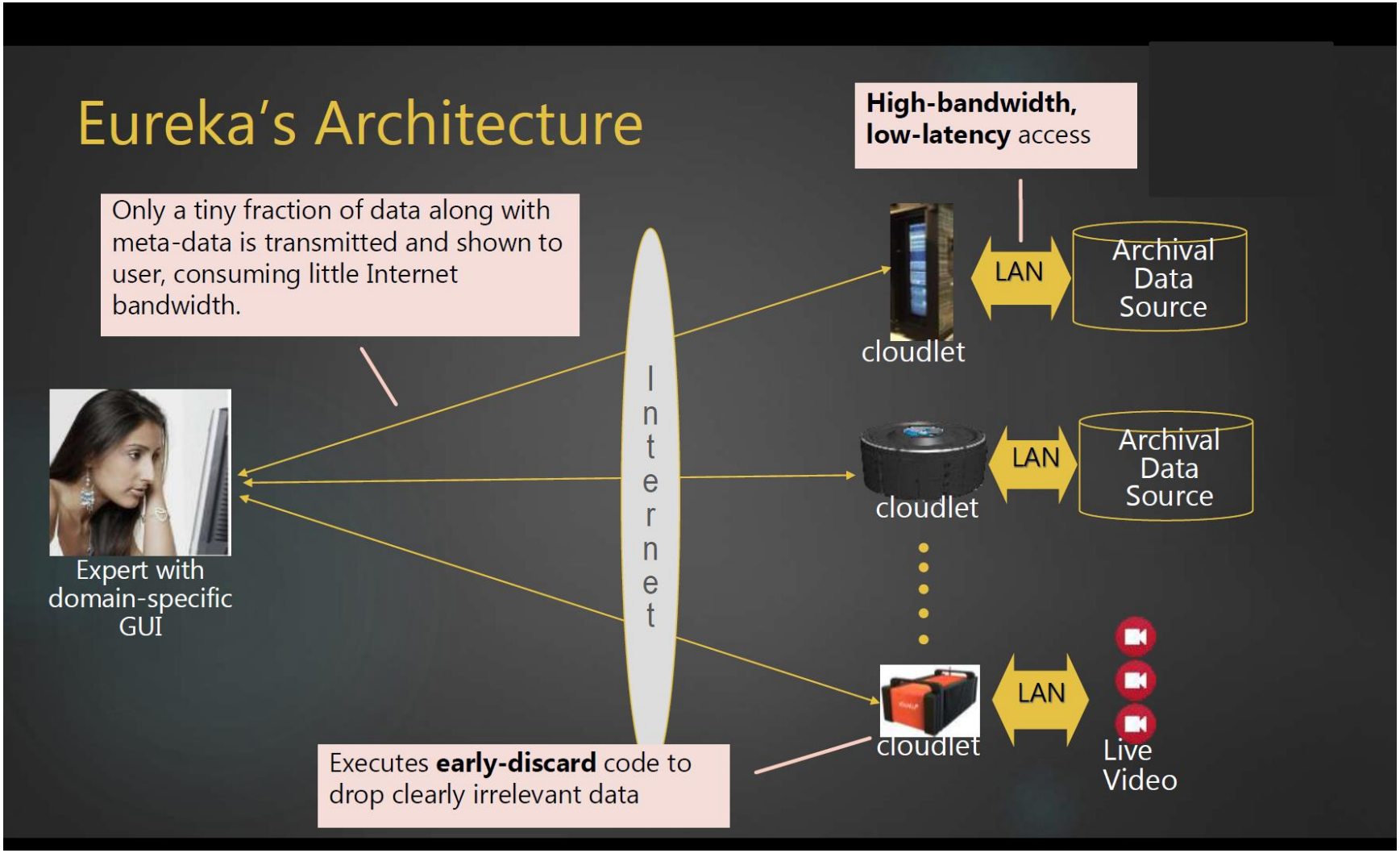
Building a test set is hard

- Non-expert crowd-sourcing won't work
- Data may have privacy or other restrictions
- Need 10^x or more training samples for DNN
- Expert may need to shift through 10^y , y>>x samples; experts are $$
- Goal: make expert's life easier
  - Optimize "human-in-the-loop" time

Eureka Approach

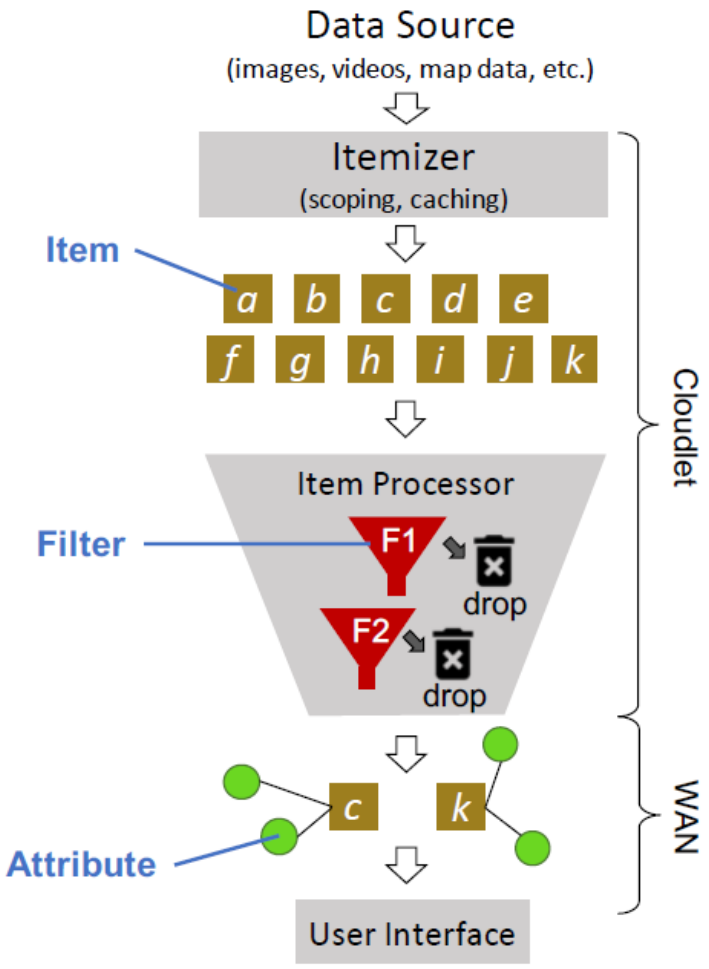
- Focus on image labelling
- Assume images are widely distributed and come from different sources
  - Even live streams, e.g. IoT
  - Can turn on/off data sources
- Support the expert in the labelling process
  - Early discard => filter or classifier that says “NO WAY”
  - Iterative discovery workflow
  - Edge computing

Stolen slides begin now

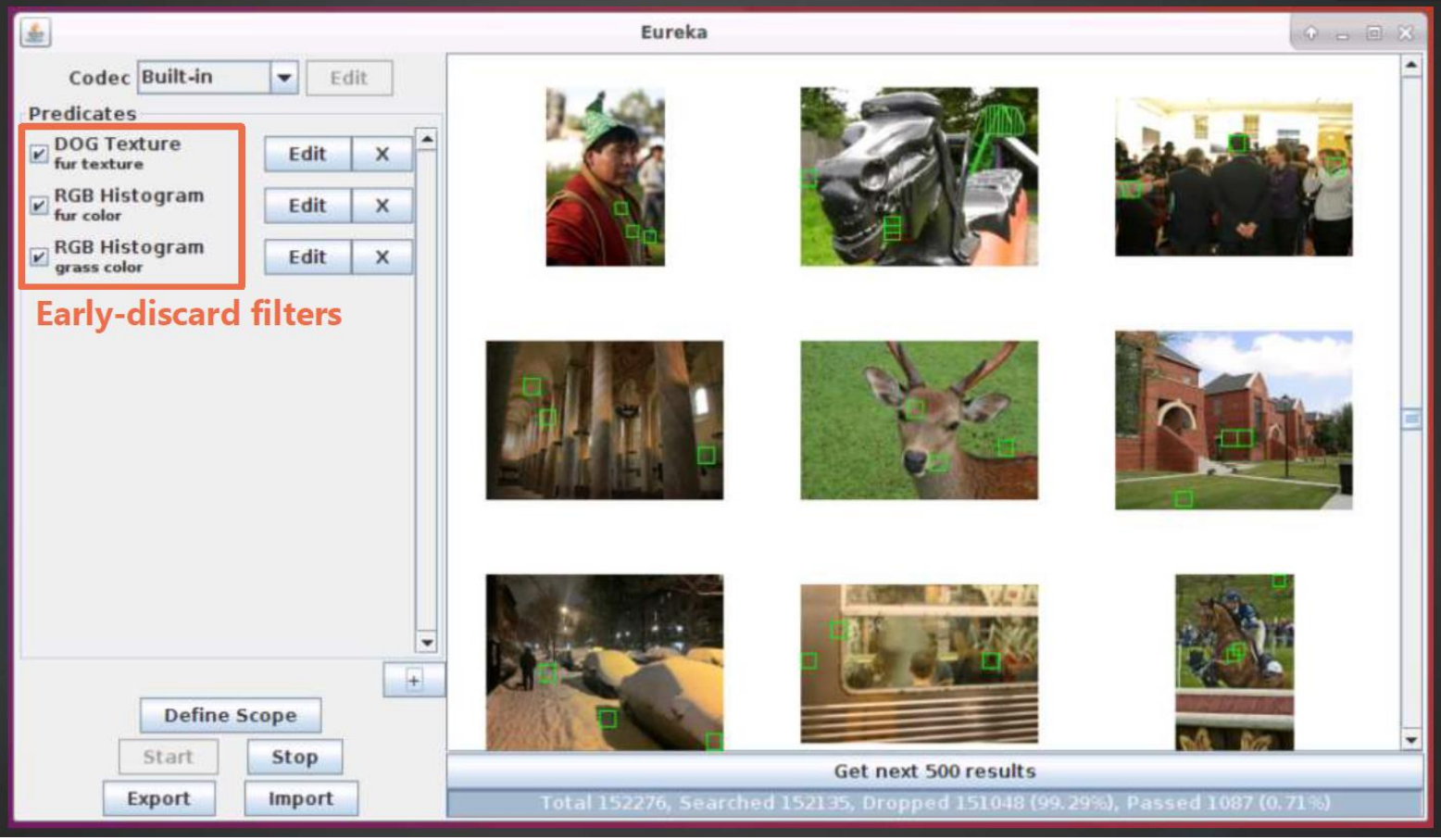


cloudlet = edge node near data source

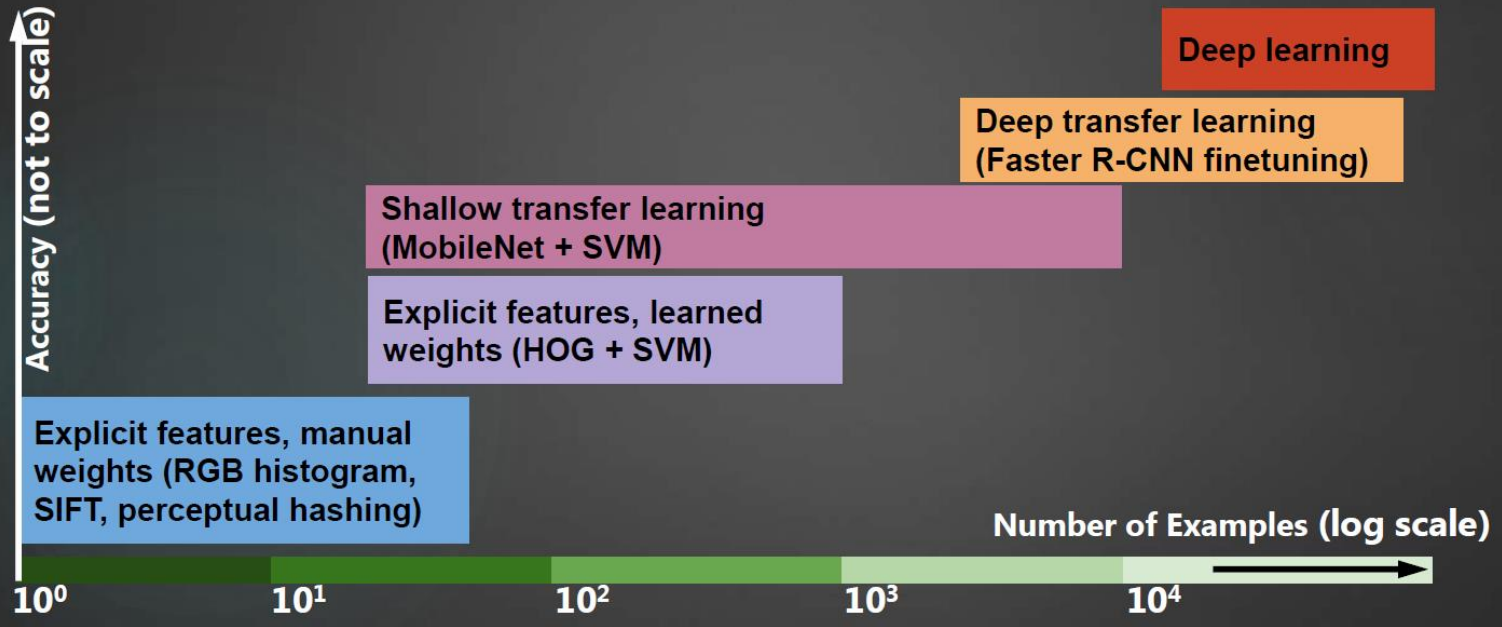
Edge node (cloudlets) run Filters



Example GUI: Finding Deer

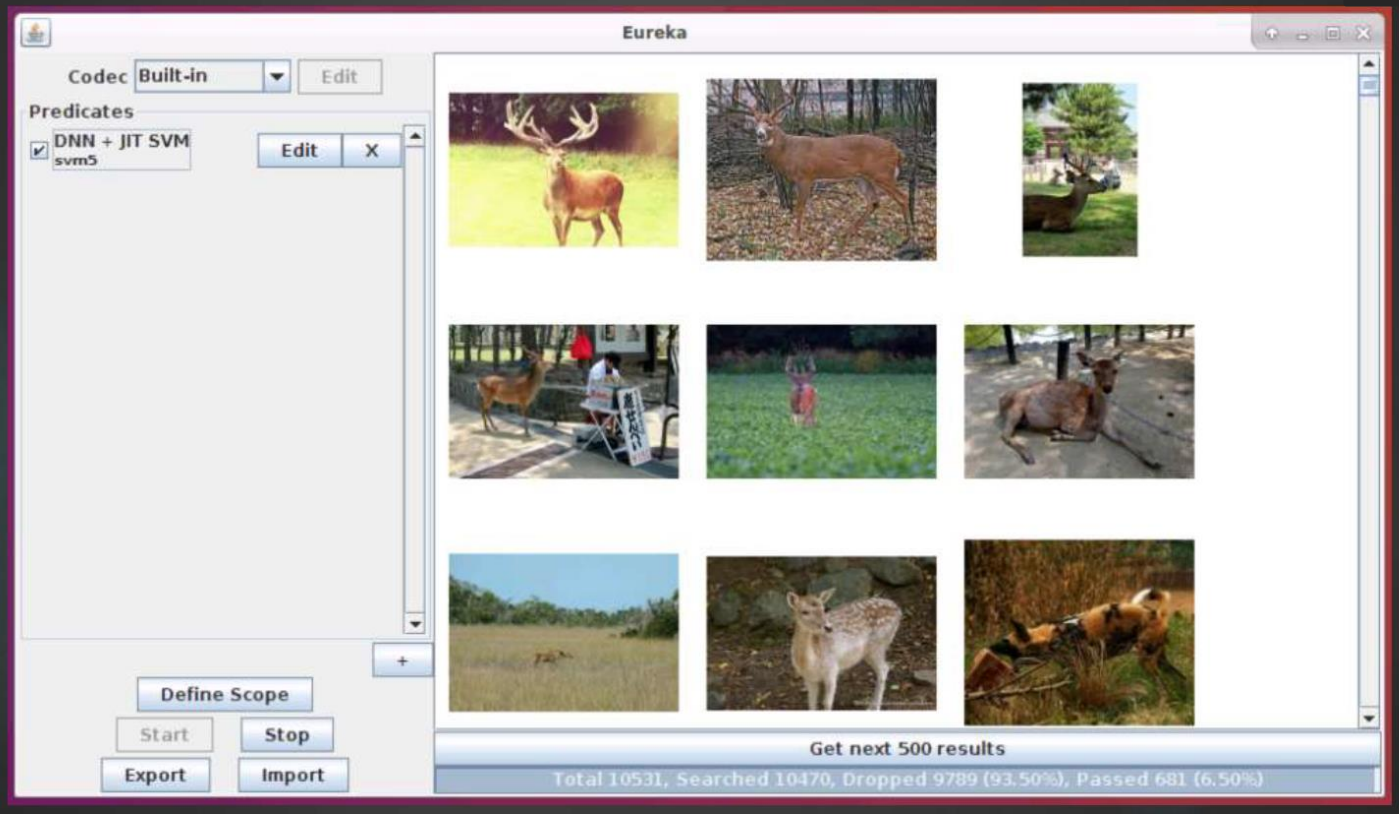


Iterative Discovery Workflow



=> More data … Better classifiers … Control false positives!

Finding Deer (after a few iterations)



Matching

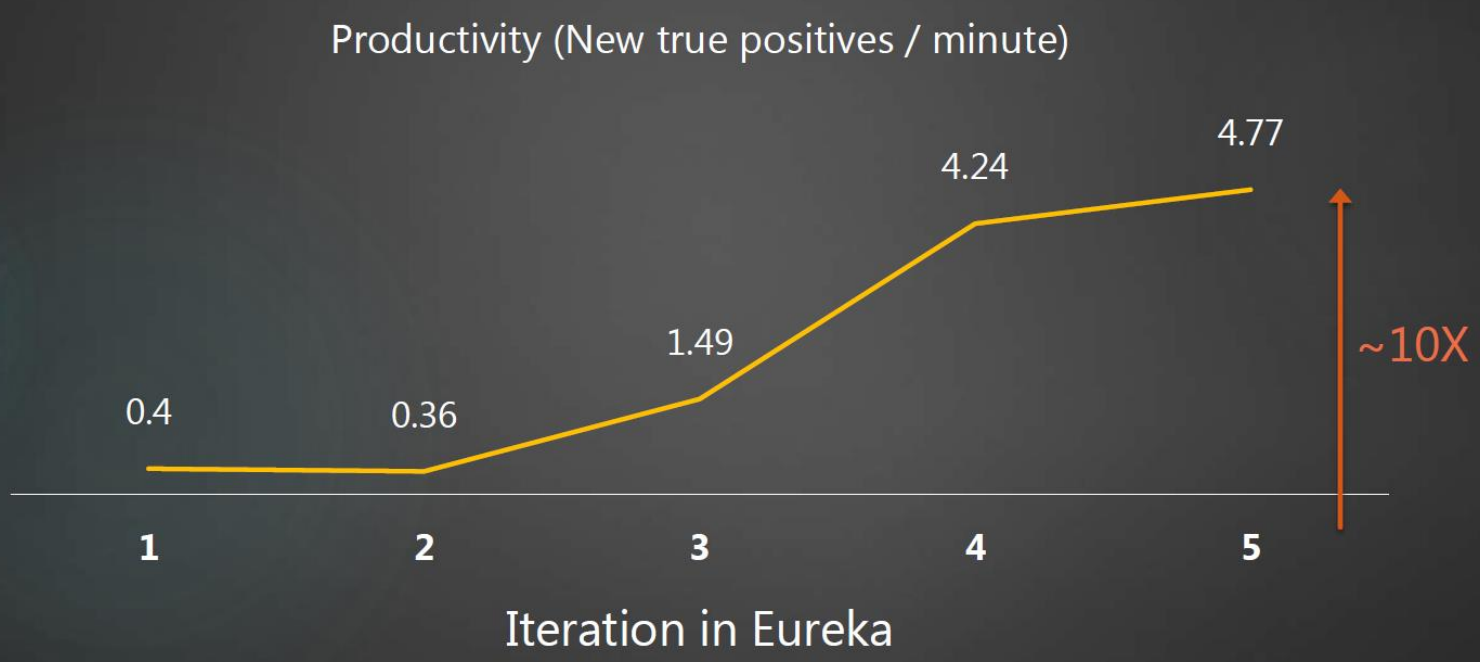
- Optimize user time/attention
- Deliver data to expert at a rate they can handle
  - Human labelling time >> Single filter time
- Too fast – overwhelmed with data
  - Fewer cloudlets (less data) or deeper filter
- Too slow – kept waiting
  - More cloudlets (Watch false positives)

Evaluation: Case Studies

	 Deer	 Taj Mahal	 Fire hydrant
Estimated base rate	0.07%	0.02%	0.005%
Collected positives in evaluation	111	105	74
Images viewed by user	7,447	4,791	15,379
Images discarded by Eureka	2,104,076	2,542,889	2,734,070

Iteratively Improving Productivity

The case of deer



Discussion

- Creating data labels is time-consuming

- Discussion
  - Assumptions: data can come from anywhere
  - Expert data: is this true?