Special Topics: Machine Learning (ML) for Networking

COL867

Holi, 2024

Week 2

Tarun Mangla

Recap

- High-level overview: What/Why of ML for Networking

- Introduction to network data
  - End-user/server: logs from application and network layer
  - Network operator: Firewall, router logs
    - Router state: SNMP logs
    - User traffic: flow statistics, packet traces

- Speedtest exercise

Data → Forward the data pack

Control → How to forward (Rules)

Management → N/w provisioning / configuration / operation

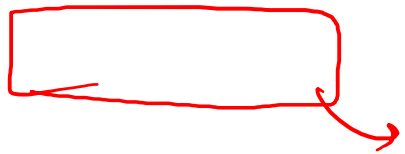
N/w management

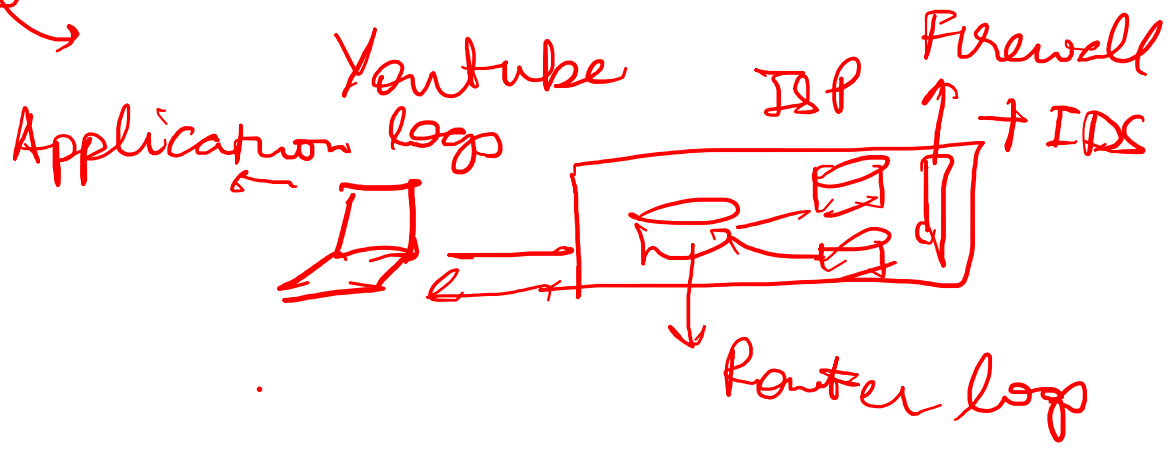
Monitor → Infer → control → (SDN)

(Prog in data plane) ↳ Human rules → ML

① N/w are becoming complex

② Technology push





Module 1: Network Learning Problems

- Traffic classification
- Performance inference
- Resource management
- Network security

Network Traffic Classification

What is it: Network operator wants to know which class does the traffic belong to?

- Class could be one of the following:
  - Application category: video streaming, P2P, video conferencing, web browsing etc.
  - Application: Netflix, YouTube, Google, Gmail
  - QoS category

Why is it important?

- Useful for various kinds of controls
  - Capacity planning
  - Service differentiation
  - Traffic engineering

- Preliminary step for other learning tasks
  - Performance monitoring
  - Intrusion detection

- Caveat: Use case determines some additional constraints:
  - Real-time vs offline

How to do traffic classification?

- Data: Network traffic

- How to use the data for traffic classification?

① Port - based
② IP - based
③. Payload - based
④ Traffic signature (ML-based)

[Modeling -based]

Port-based classification

- IANA keeps a registry of port to application

- Advantage:
  - Lightweight and easy to implement

- Disadvantage:
  - Multiple applications using the same port
  - Non-standardized ports for some applications: dynamic port negotiation
  - Easily misused

Content Distribution N/w (Akamai / LimelightN / AWS)

IP-based classification

(E.g. Google) → ① Content provider Many Apps
② CDN / Cloud

← ① Large database
② Same server different applications
③ Database can be challenging to construct

(sctp → 22
ssh → 22)

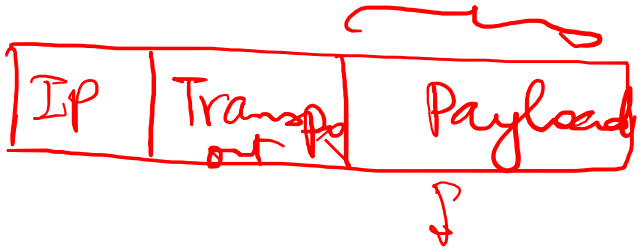
Web browsing : 80/443

Video Streaming : 80/443

↓
→

(P2P appns, Video conferencing)

Payload-based classification



- Look at the (unencrypted) packet payload
- Specific signatures in the traffic that can reveal application names
  - Whether RTP headers are being used? → Video conferencing
  - HTTP requests
  - DNS
- Use to identify application

Port
UDP - 53

www.google.com
↓
IP Address

DNS request
↓
response

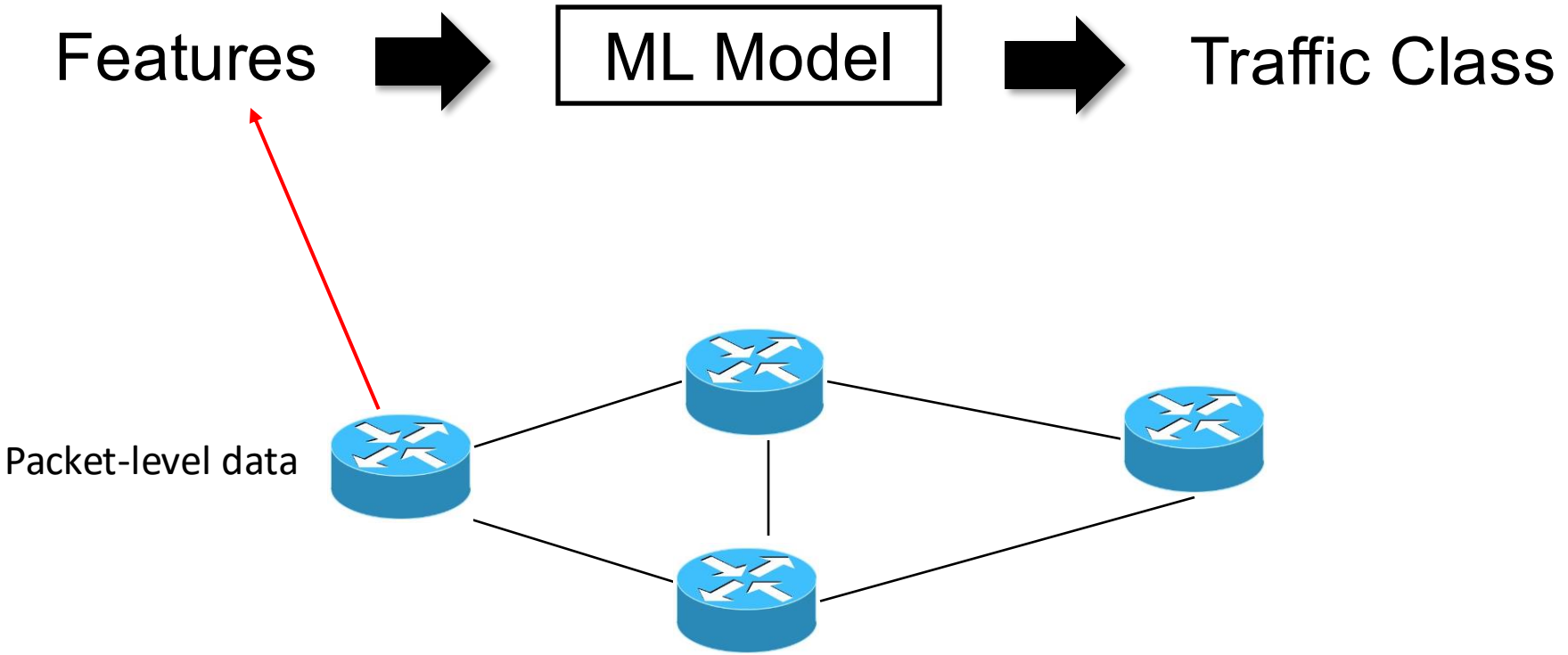
- Disadvantage:
  - Significant cost overhead (not true anymore?)
  - Increasing amount of traffic is encrypted
- What is still available in the network traffic?

Recap: Traffic Classification

- **What**: Classification of traffic into pre-defined categories
  - E.g., applications (Netflix, YouTube, Teams, Meet etc.) or application category (web, streaming, P2P etc.)
- **Why**: security, performance monitoring, capacity planning
- **How**:
  - Port-based (IANA)
  - Payload-based
- Can we use ML for this task?

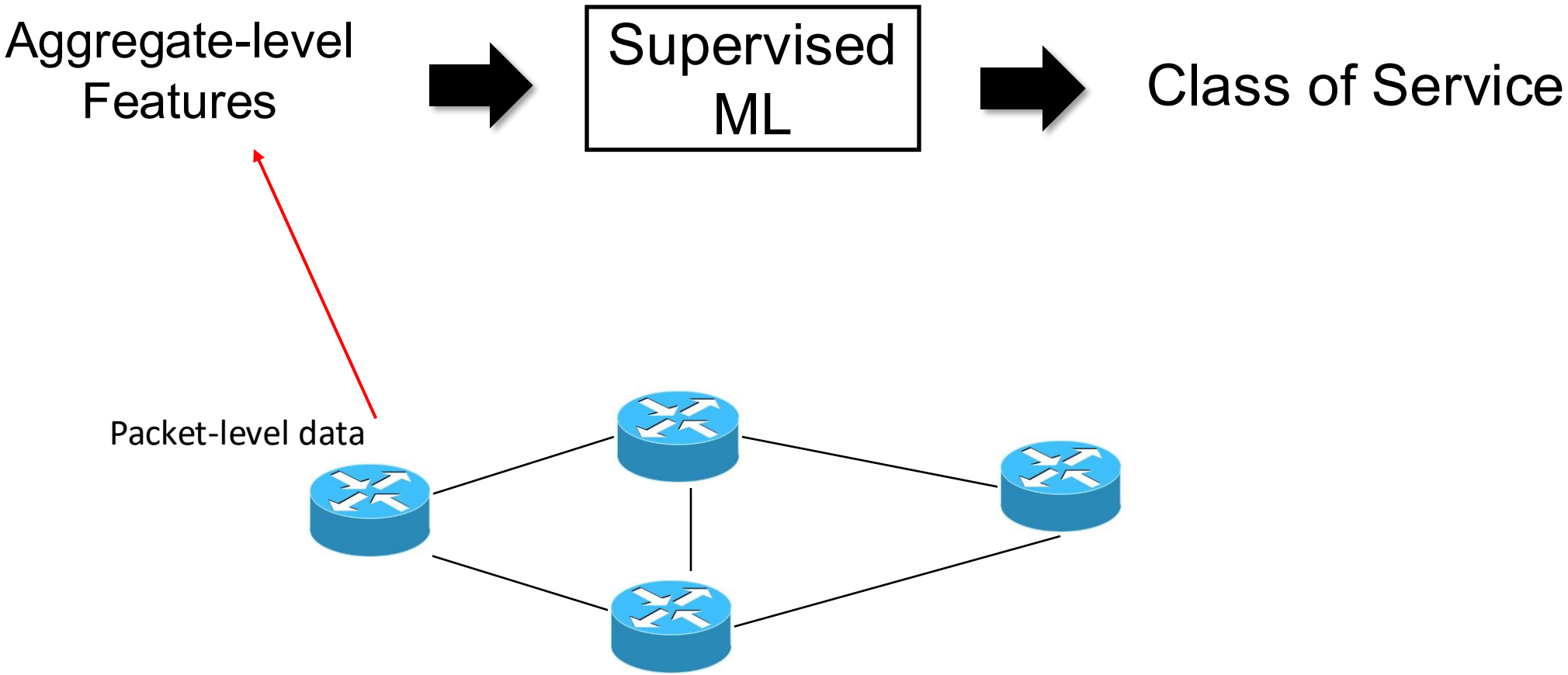
General Framework for Traffic Classification

- Given traffic X_i, predict its class Y_j

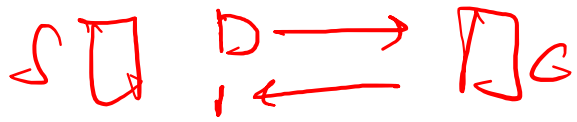


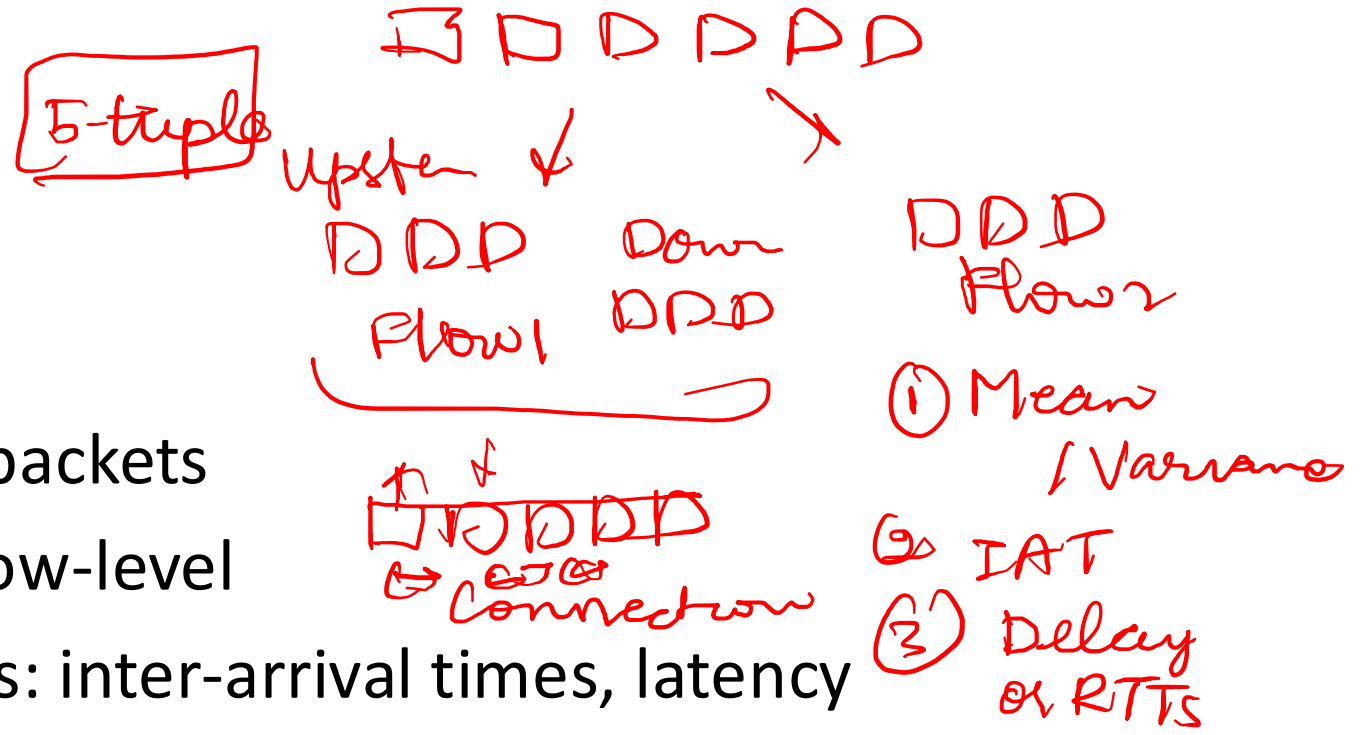
Paper: Class-of-Service Mapping for QoS.. [Roughan2004]

- Given an aggregate (server IP or port), predict its Class of Service Y_j



Candidate Features





Statistics (mean, variance,)

- Packet-level: packet size
- Flow-level: flow volume, # of packets
- Connection-level: similar to flow-level
- Intra-flow/connection features: inter-arrival times, latency
- Multi-flow: aggregate multiple connections (# connections, mean size per connection)

Flow statistic:

Features vs Network Data Collection Methods

AT&T

data source	features				
	packet level	flow-level	connection-level	intra-flow	multi-flow
packet trace	yes	yes	yes	yes	yes
sampled packets	yes	biased	no	biased	biased
flow-data →	some	yes	no	no	yes
SNMP	no	no	no	no	no

Classification Method

- Classical supervised ML methods
  - k-NN
  - Linear Discriminant Analysis
  - Naïve-bayes
  - Random Forest
  - …

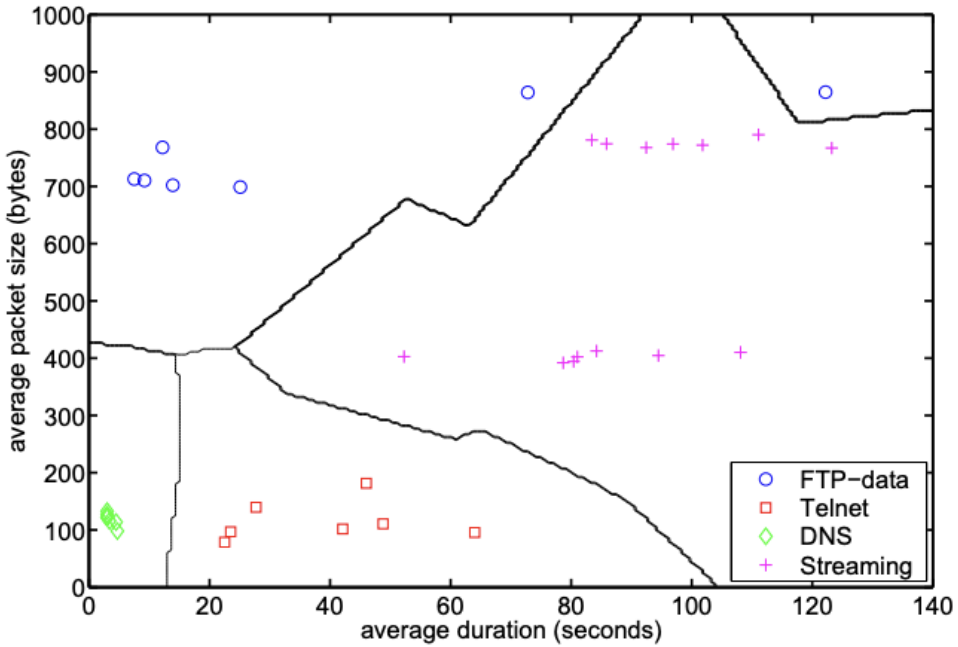
Training

IIT Delhi / Campus n/w

- Data collected from multiple vantage points (public dataset, specific-application, within AT&T network, AT&T labs network)
- Data labeling: using port numbers, application payload
- 10-fold cross validation

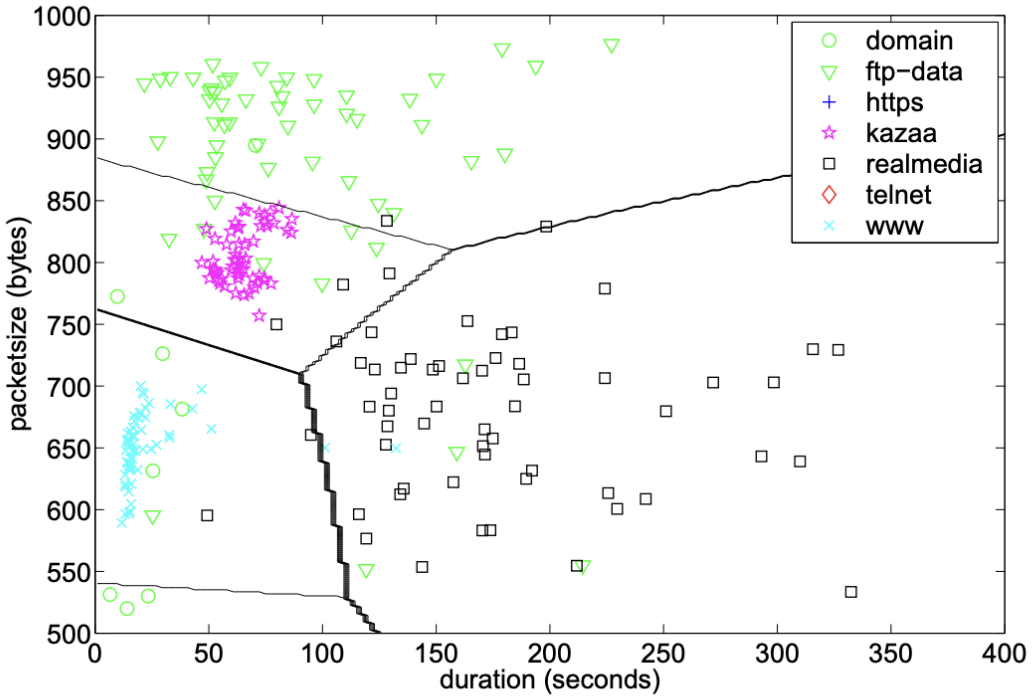
Which are the most important features?

- Candidate features: average packet size, flow duration, bytes per flow, packet per flow, and root mean square packet size
- **Most important features:** Average packet size and flow duration



(b) Nearest Neighbor.

With a different dataset.



(b) Zoom into upper left region.

- realmedia (streaming) confused with ftp-data (download)

How do we separate out these two?

- Use inter-arrival variability metric

Summary: Class-of-Service Mapping for QoS.. [Roughan2004]

- Given an aggregate (server IP or port), predict its Class of Service Y_j