Bad Neighborhoods - Learning malicious infrastructure at internet scale

Tamás Vörös – Sophos AI, Data Scientist
Rich Harang – Duo Security, Data Science Senior Tech Lead
Josh Saxe – Sophos AI, Chief Scientist
Konstantin Berlin – Sophos AI, Director

SOPHOS

We will talk about

A data driven ML approach to assign 'risk scores' to various blocks of the IPv4 space

Agenda

- Motivation
- Introduce/confirm structural bias
  - Review the hierarchical structure of the internet
  - Demonstrate non-uniform distribution of maliciousness over the IP space on two separate datasets
- Propose 3 architectures that support the structural bias
  - Random forest (RF)
  - Convolutional Neural Network (CNN)
  - Transformers
- Identify ISP IP space non contiguity as a weak spot of IP-input-only models
  - Generalize knowledge over non-contiguous IP spaces, by utilizing pretrained representation learning

High level threat landscape

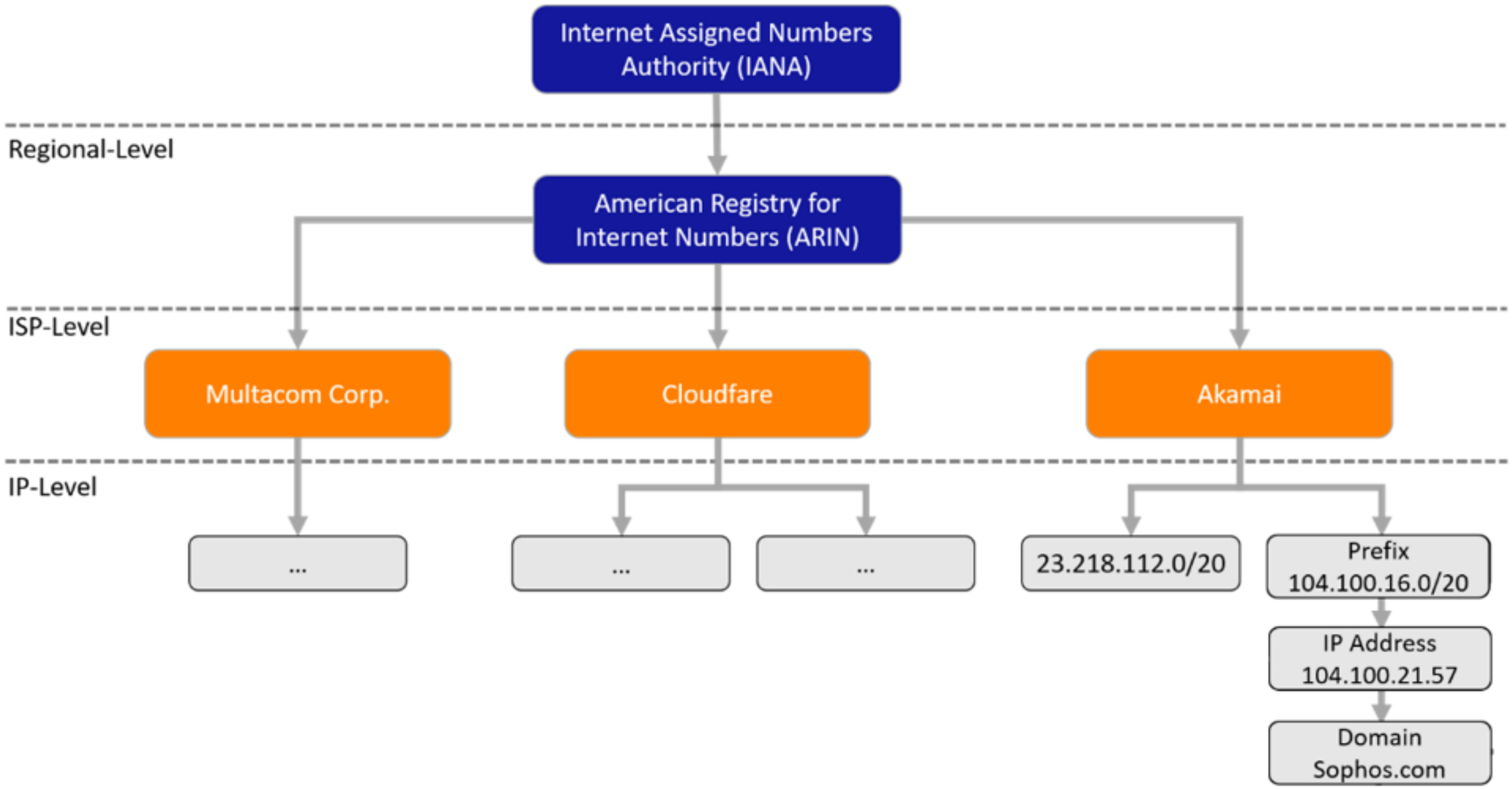
Malicious websites

Phishing emails

Command and
Control servers

Data exfiltration

IP addresses' structure lends itself to statistical detection modeling



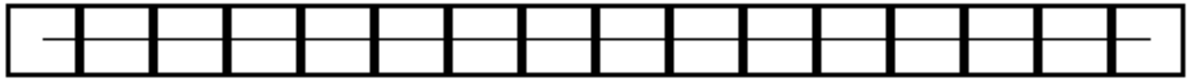
Datasets

- Malicious and benign websites' IP addresses
  - Benign:
    - social network infrastructures
    - search engines
    - online stores
    - video hosting providers, and so on.
  - Malicious:
    - malware repositories
    - phishing sites
    - callhomes
  - Total size: 455248

- Spam associated and non-spam associated IP addresses
  - Benign:
    - static reputation list containing IPs provided by organizations who have had prompt reactions to any kind of abuse of their mail servers. (DNSWL)
  - Malicious:
    - hijacked PCs worms/viruses with built-in spam engines. (Spamhaus)
  - Total size: 432792

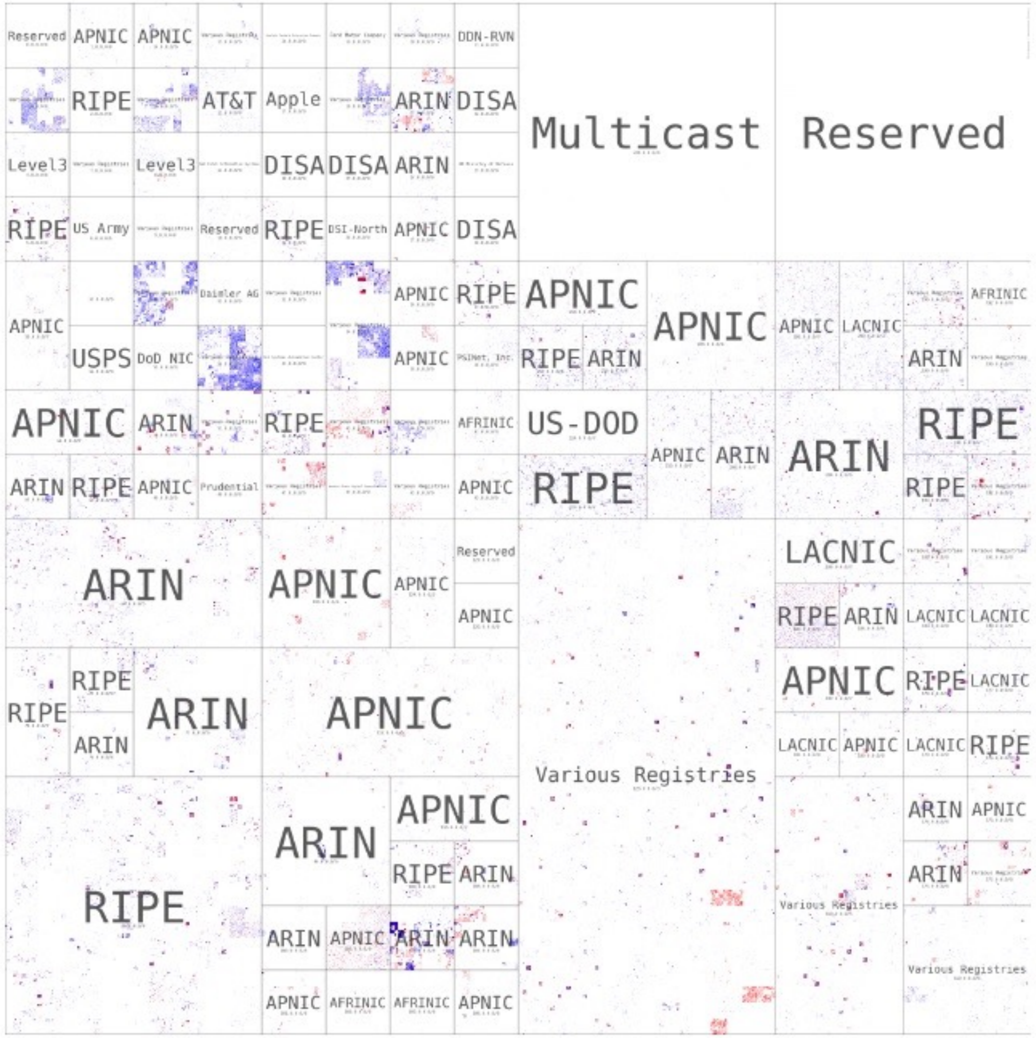
Hilbert curves - concept

- 1D and 2D space mapping
- Preserves some notion of closeness or locality
- Two data points that are close to each other in one-dimensional space are usually close to each other after the transformation.

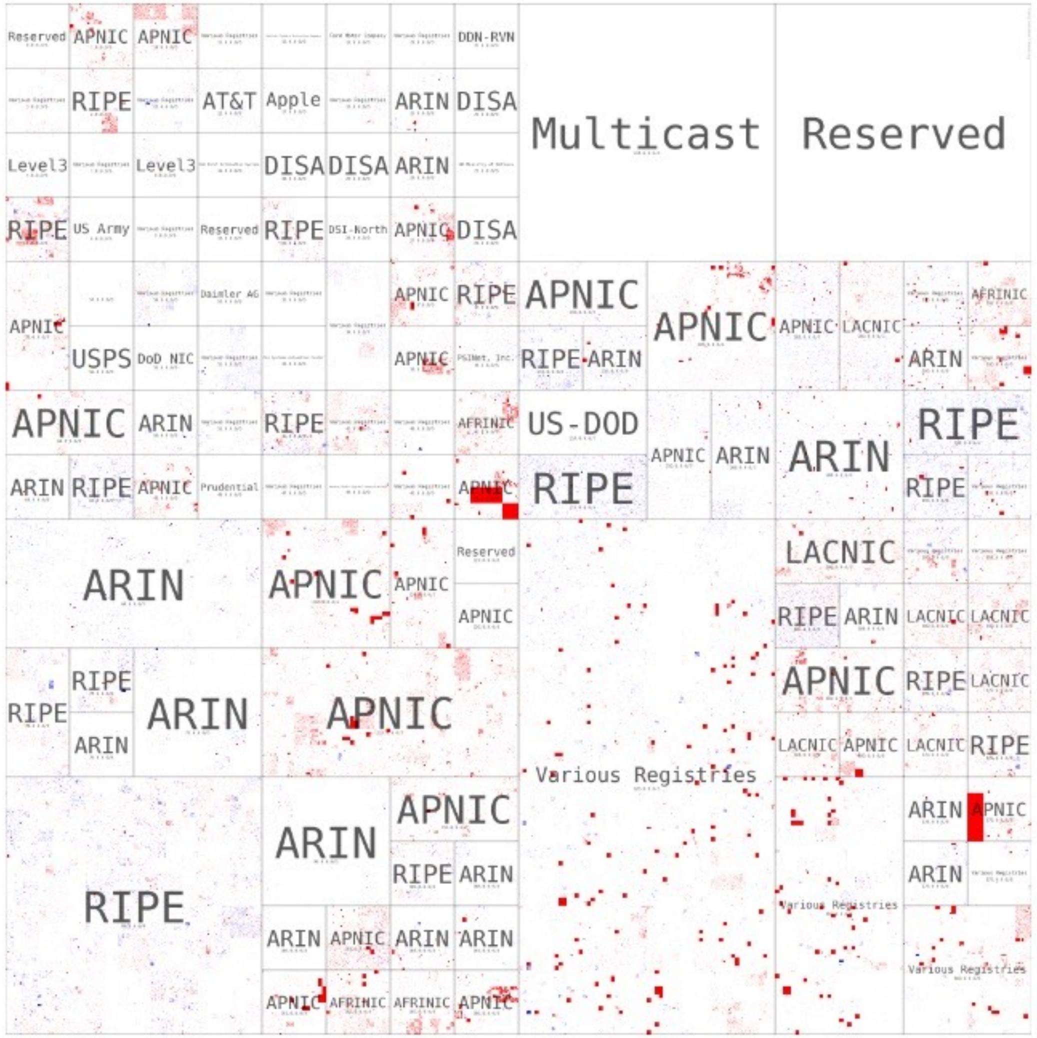


https://blog.benjojo.co.uk/post/scan-ping-the-internet-hilbert-curve

Hilbert curves



(a) Web ip addresses

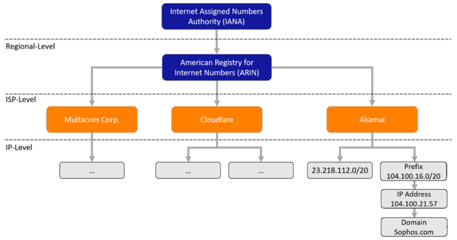


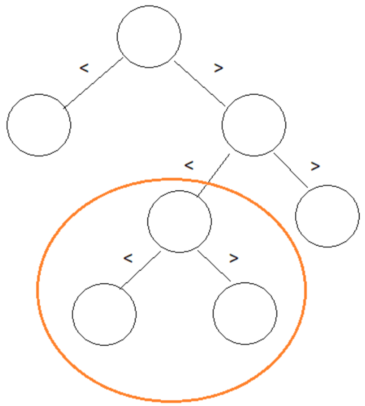
(b) Spam ip addresses

https://github.com/measurement-factory/ipv4-heatmap

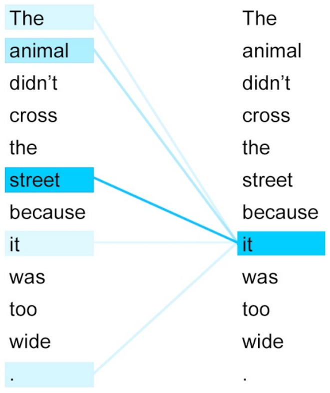
IP Address Modeling

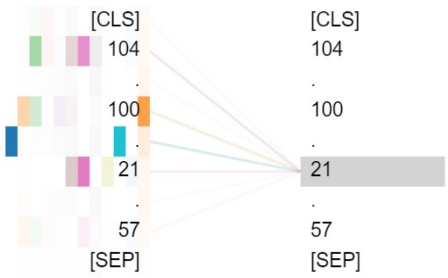
integer	binary	standard decimal dotted
1751389497	01101000.01100100.00010101.00111001	104.100.21.57
RF	**CNN**	**Transformer**



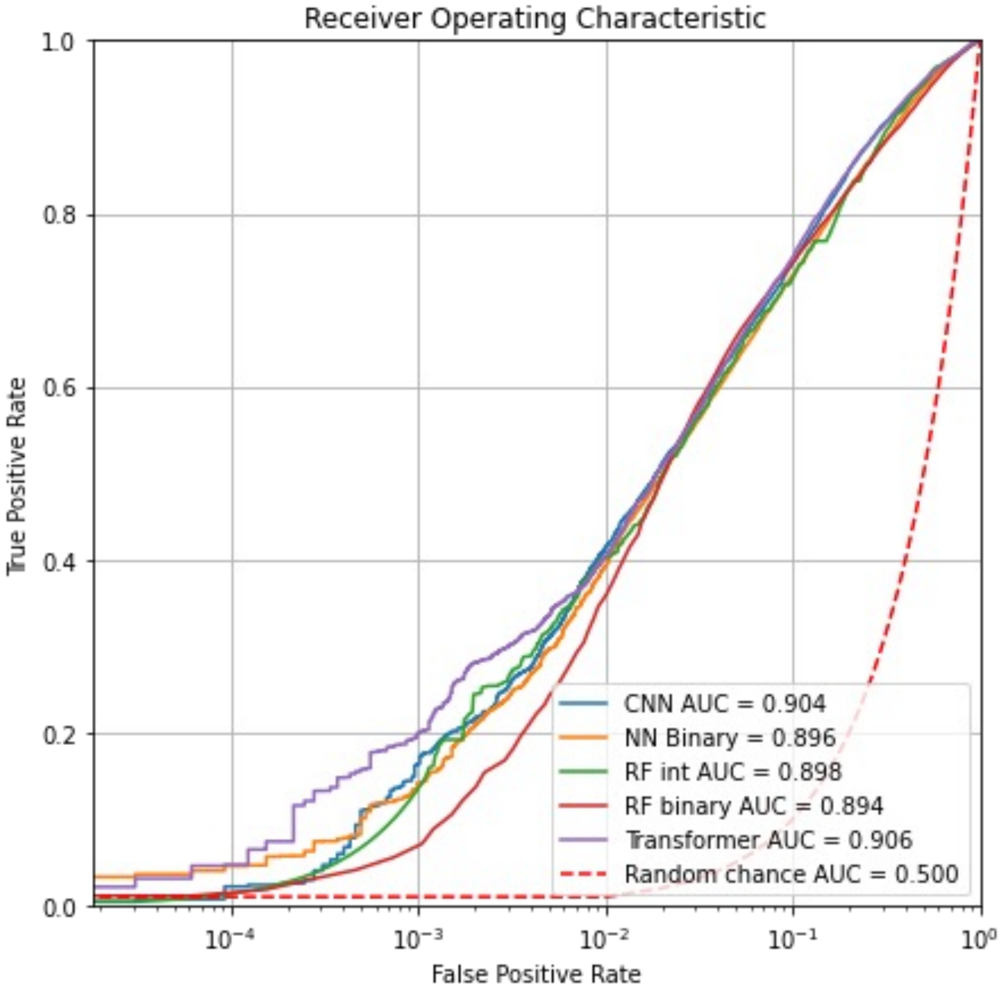


104. 100 . 21 . 57	Prefix
01101000.01100100.00010101.00111001	104.100.21.57/32
01101000.01100100.00010101.00111000	104.100.21.57/31
01101000.01100100.00010101.00111000	104.100.21.57/30
…	…
01101000.01100100.00010101.00111000	104.100.21.57/20
01101000.01100100.00010101.00111000	104.100.21.57/19
…	…
01000000.00000000.00000000.00000000	104.100.21.57/2
00000000.00000000.00000000.00000000	104.100.21.57/1

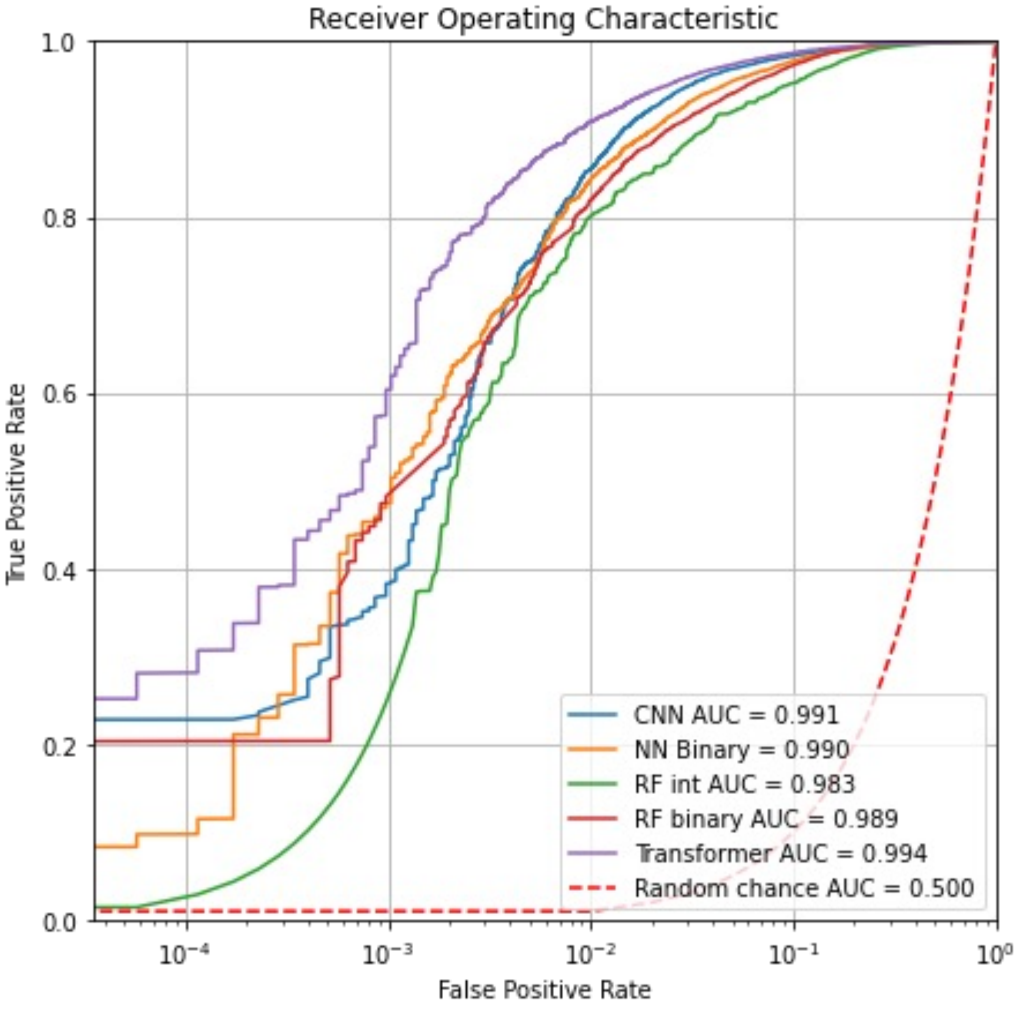




Results



Web dataset

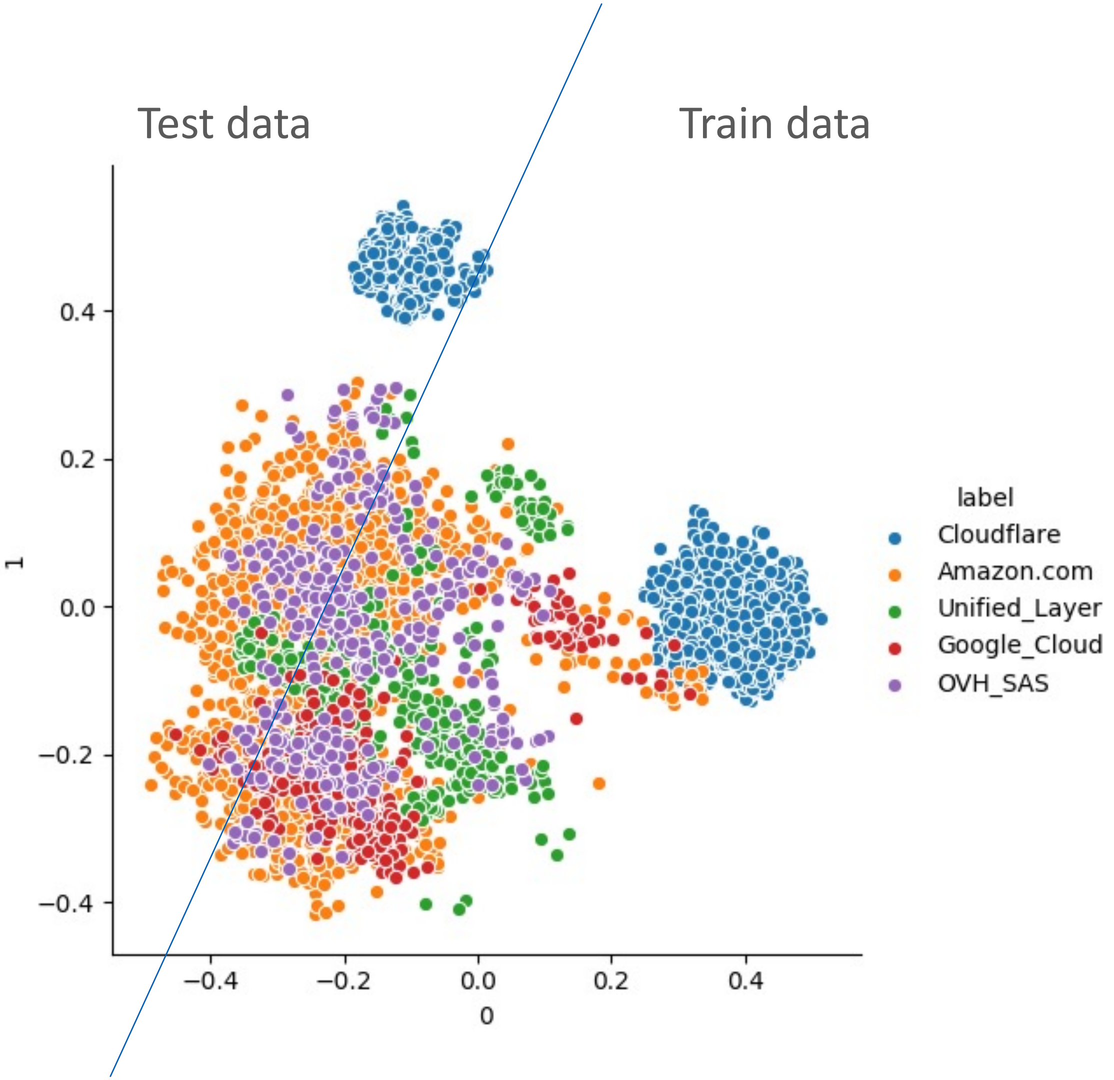


Spam dataset

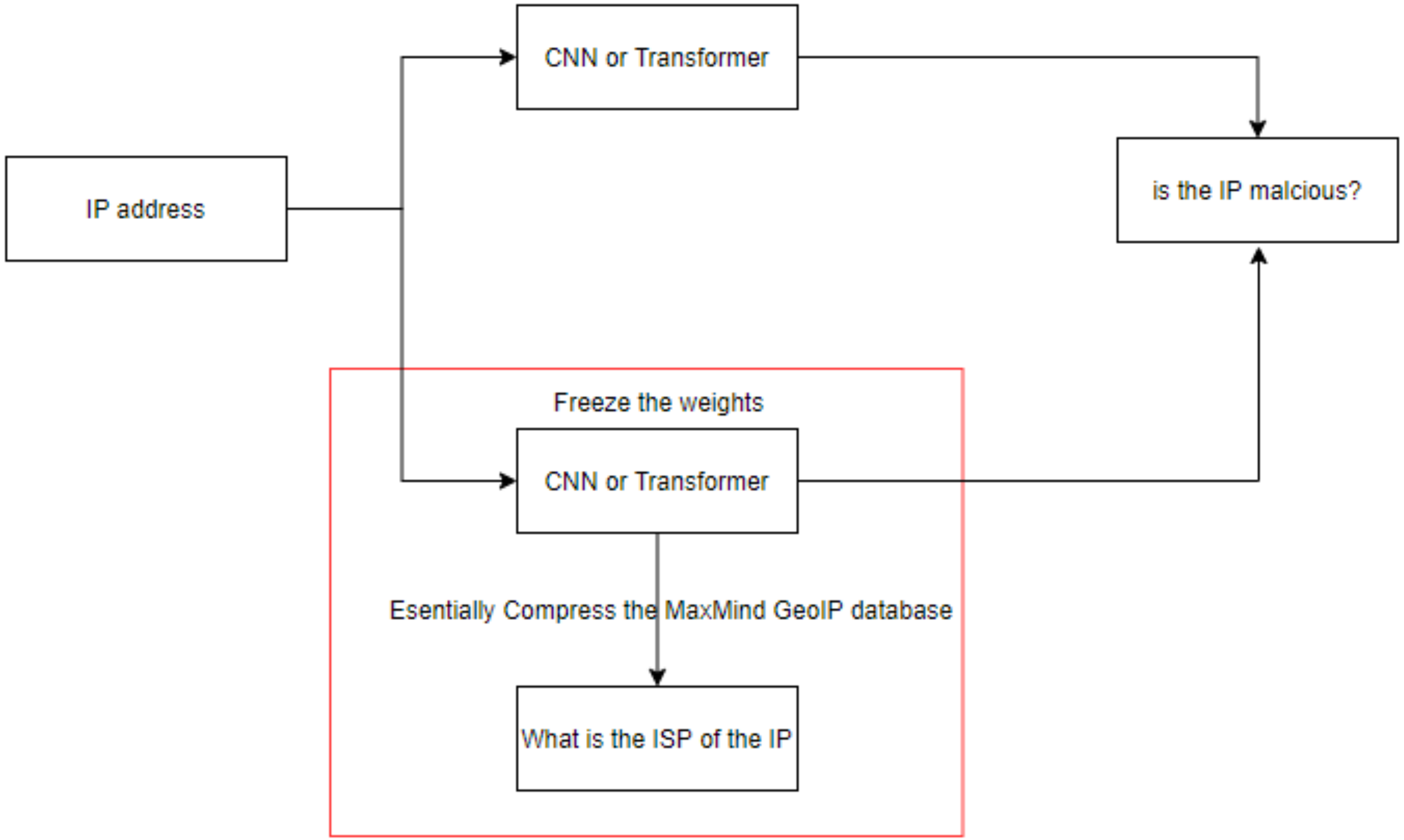
Utilizing ISP

Fixed ISP: Cloudflare

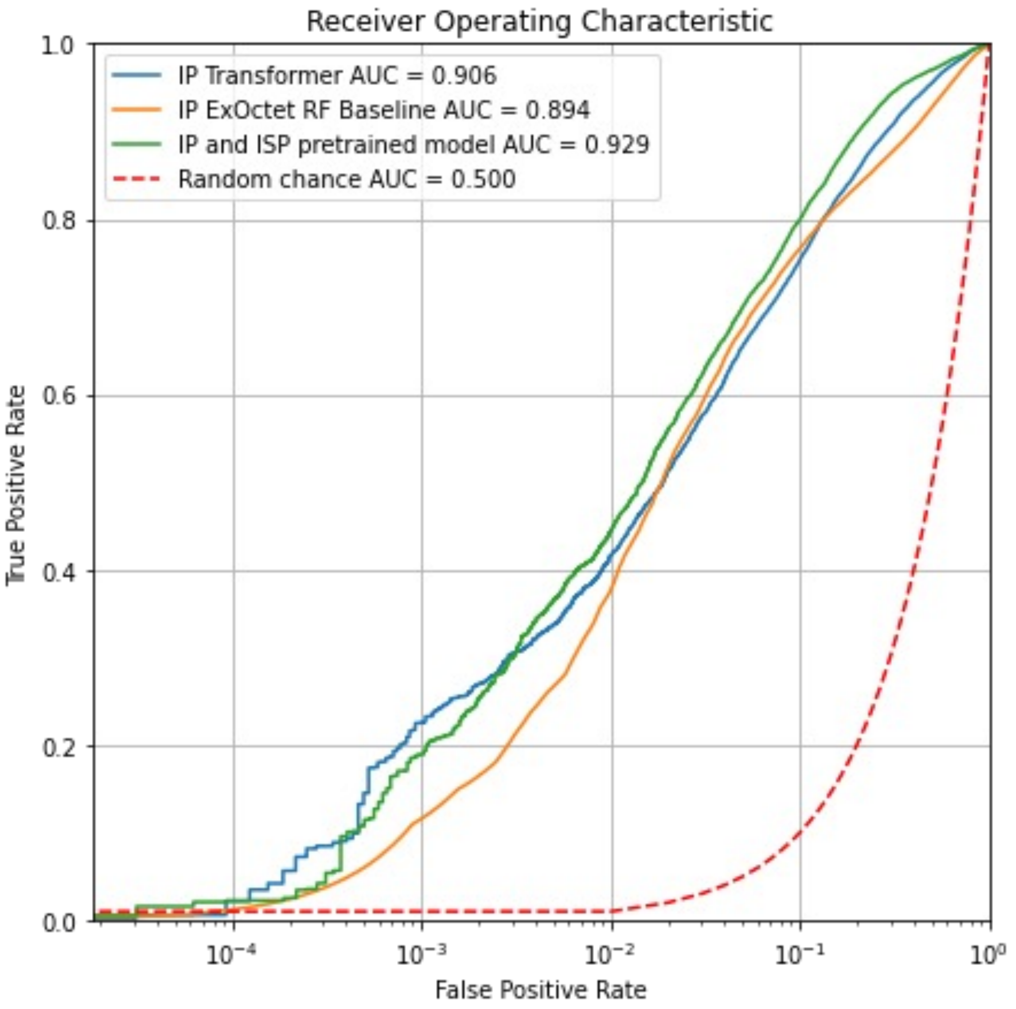
Prefix	# Training data	# Test data
103.21.244.0/22	5000	0
23.15.11.0/24	3	5000



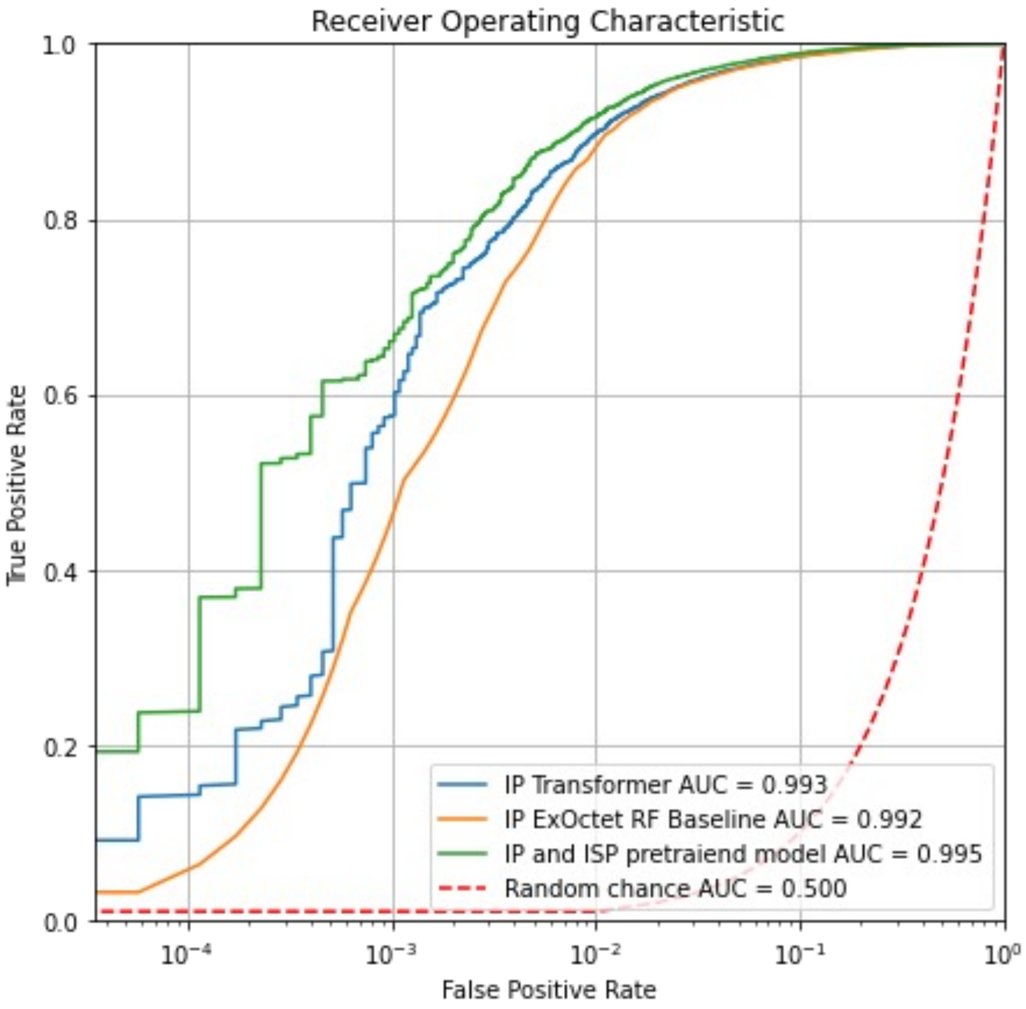
Let's add pretrained a component



Final results

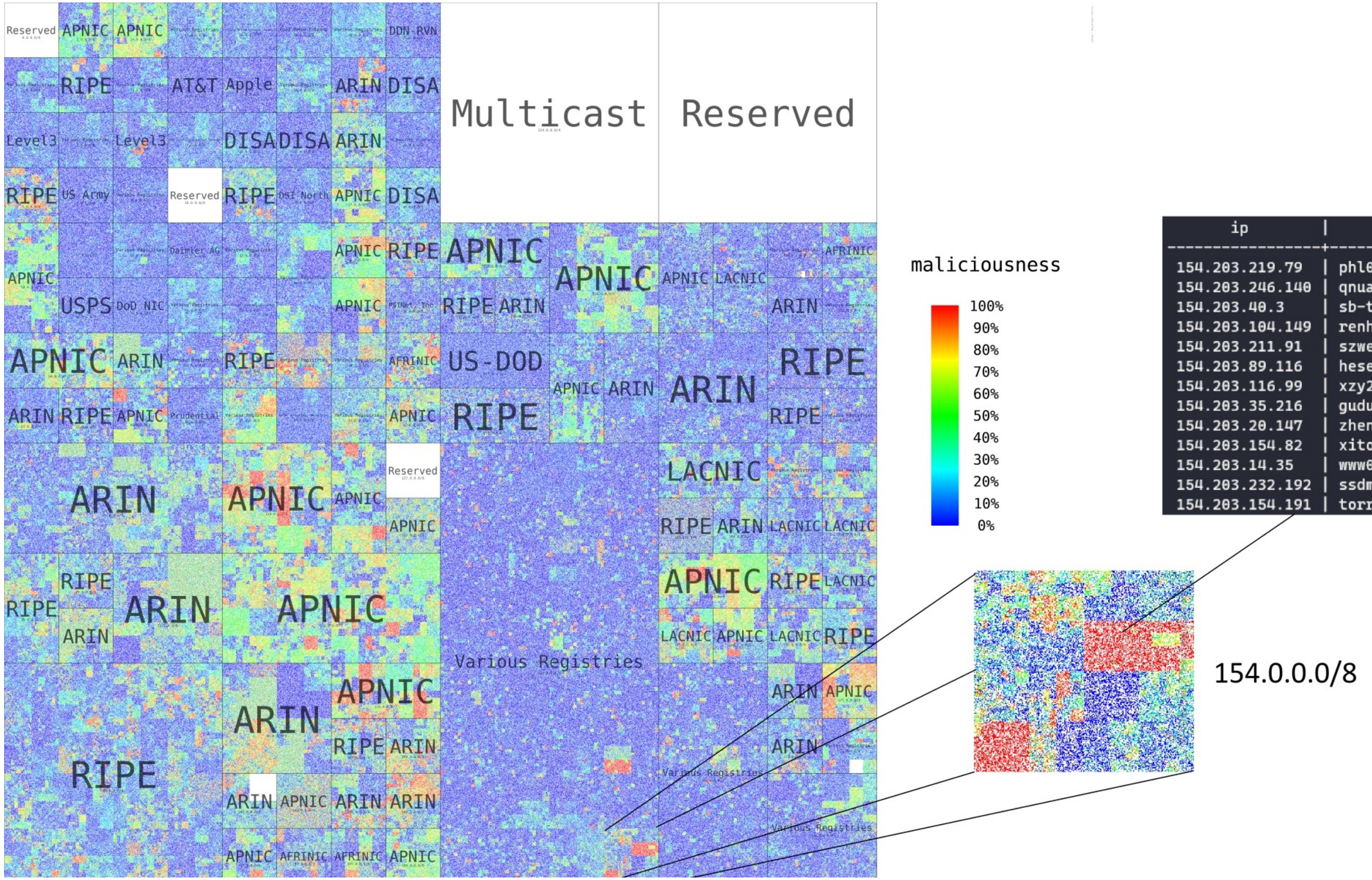


Web dataset



Spam dataset

Conclusion



ip	domain
154.203.219.79	phl0063.com
154.203.246.140	qnuan.cn
154.203.40.3	sb-tools.cn
154.203.104.149	renhuikj.com
154.203.211.91	szwebdesign.net
154.203.89.116	hesea.net
154.203.116.99	xzy2.com
154.203.35.216	guduopu.com
154.203.20.147	zhengxinsy.com
154.203.154.82	xitongdaren.com
154.203.14.35	www0ot.com
154.203.232.192	ssdmkj.com
154.203.154.191	torreonze.com