Preventing botnets

- Traditional
  - code analysis and finding malware fingerprints
- Code/binary analysis is mostly manual and increasingly harder
  - Code obfuscation
  - Encryption
  - Self-modifying
- Behavior-based analysis is much harder to thwart
  - Bots need to communicate!

Preventing botnets

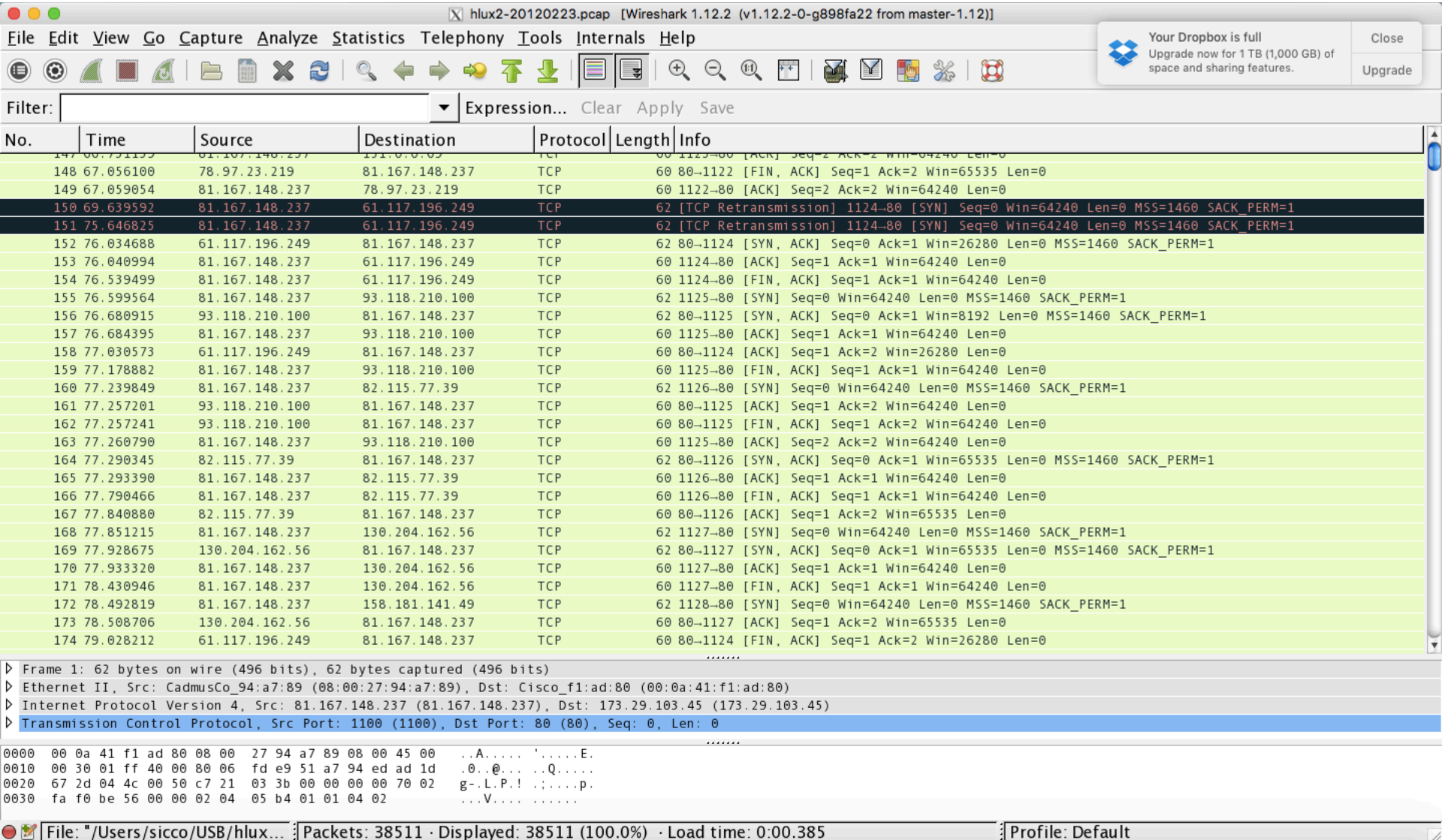
- Traditional
  - code analysis and finding malware fingerprints
- Code/binary analysis is mostly manual and increasingly harder
  - Code obfuscation
  - Encryption
  -
- B
  h
  - bots need to communicate!

Lots of data available…

Binary code (Zeus)

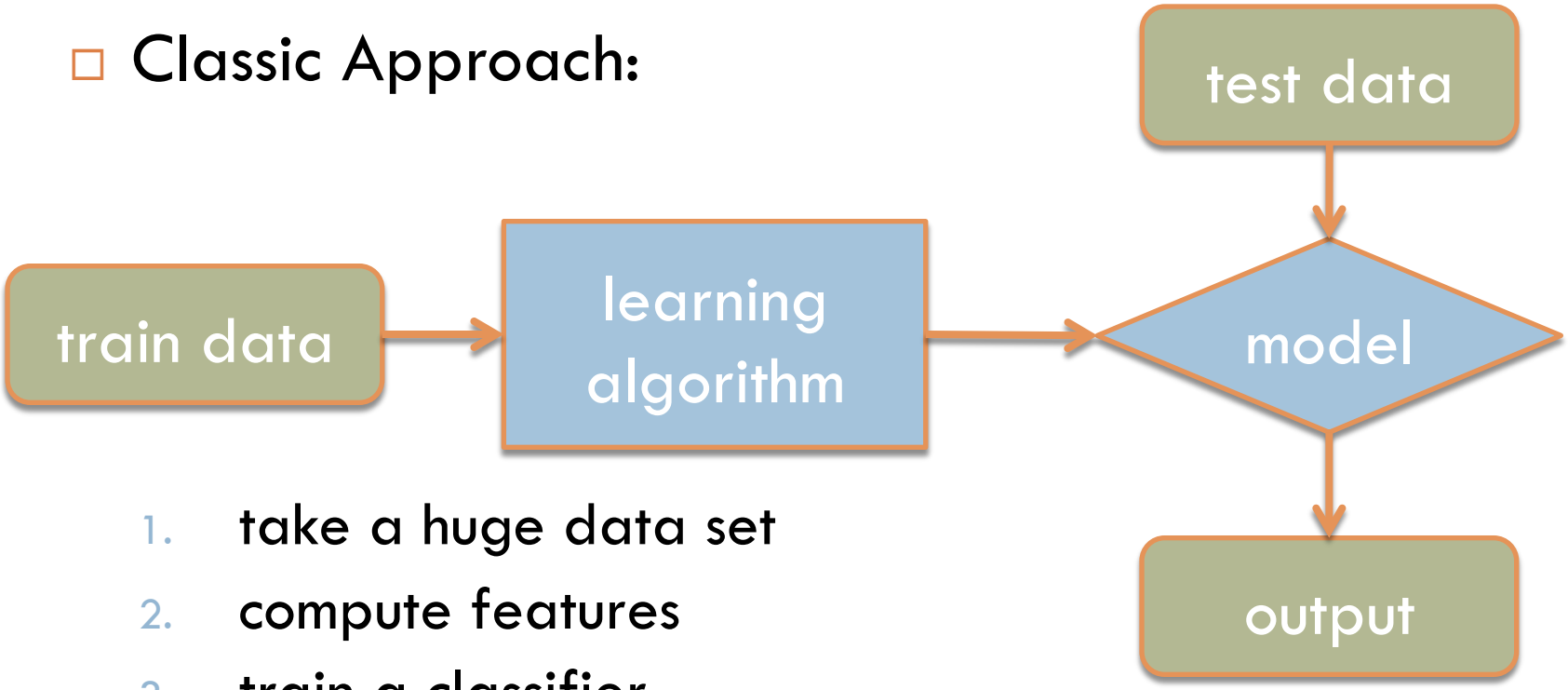
```
4D 5A 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 D8 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 50 45 00 00 4C 01 03 00 75
39 B2 52 00 00 00 00 00 00 00 00 E0 00 02 01 0B 01 0A 00 00 08 02 00 00 3A
00 00 00 00 00 00 48 30 01 00 00 10 00 00 00 20 02 00 00 00 40 00 00 10 00
00 00 02 00 00 05 00 01 00 01 00 00 00 05 00 01 00 00 00 00 00 00 70 02 00
00 04 00 00 00 00 00 00 02 00 00 81 00 00 10 00 00 10 00 00 00 00 10 00 00
10 00 00 00 00 00 00 10 00 00 00 00 00 00 00 00 00 00 00 A4 F7 01 00 18 01
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 50 02 00 AC 11 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 10 00 00 A0 05 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 2E 74 65 78 74 00 00 00 84 06 02
00 00 10 00 00 00 08 02 00 00 04 00 00 00 00 00 00 00 00 00 00 00 00 00 00
20 00 00 60 2E 64 61 74 61 00 00 00 50 20 00 00 00 20 02 00 00 04 00 00 00
0C 02 00 00 00 00 00 00 00 00 00 00 00 00 00 40 00 00 C0 2E 72 65 6C 6F 63
00 00 7C 16 00 00 00 50 02 00 00 18 00 00 00 10 02 00 00 00 00 00 00 00 00
00 00 00 00 00 40 00 00 42 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00
```

Network traffic (HLUX2)



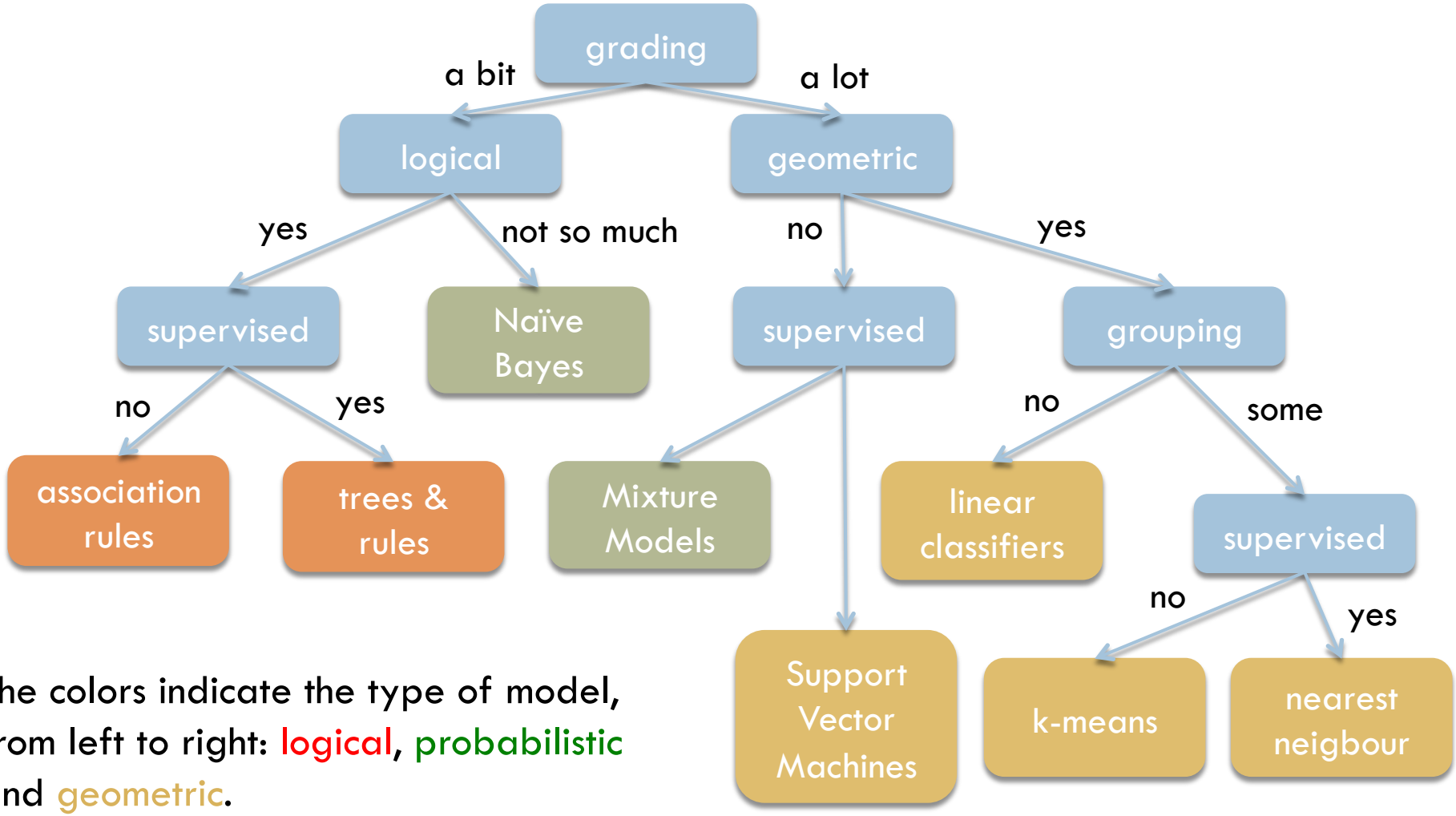
Machine learning

- Classic Approach:



1. take a huge data set
2. compute features
3. train a classifier
4. deploy the classifier on test

ML Taxonomy – many approaches



The colors indicate the type of model, from left to right: logical, probabilistic and geometric.

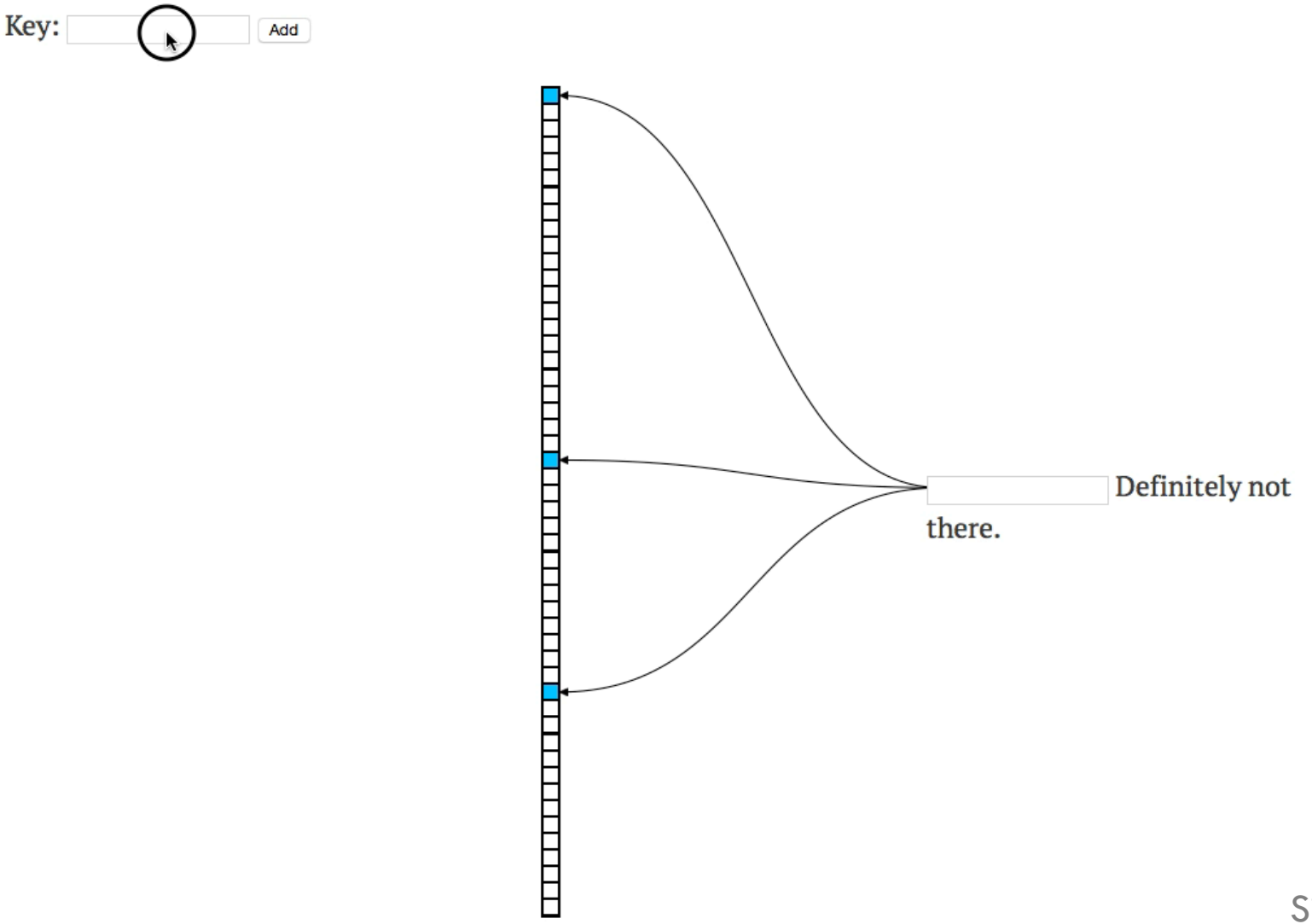
based on Peter Flach

Classic ML fails in cyber security

- Large majority (> 99%) of cases are benign!
  - adapt data/models, otherwise no positives
- Data is massive and keeps coming in!
  - need to count quickly, reduce false positives
- There is an opponent, they learn too!
  - avoid using generic fingerprints/simple rules
- Privacy makes data inaccessible…

http://www.jasondavies.com/bloomfilter/

Bloom filter: dealing with massive data



Course

- Read scientific papers, use techniques on real data:
  - Credit fraud data from Adyen
  - Botnet traffic from HLUX2
  - NetFlow traces, perhaps from EEMCS…
  - …
- Learn by doing in labs, an exam on content