NAVAL
POSTGRADUATE
SCHOOL

What Are You Searching For? A Remote Keylogging Attack on Search Engine Autocomplete

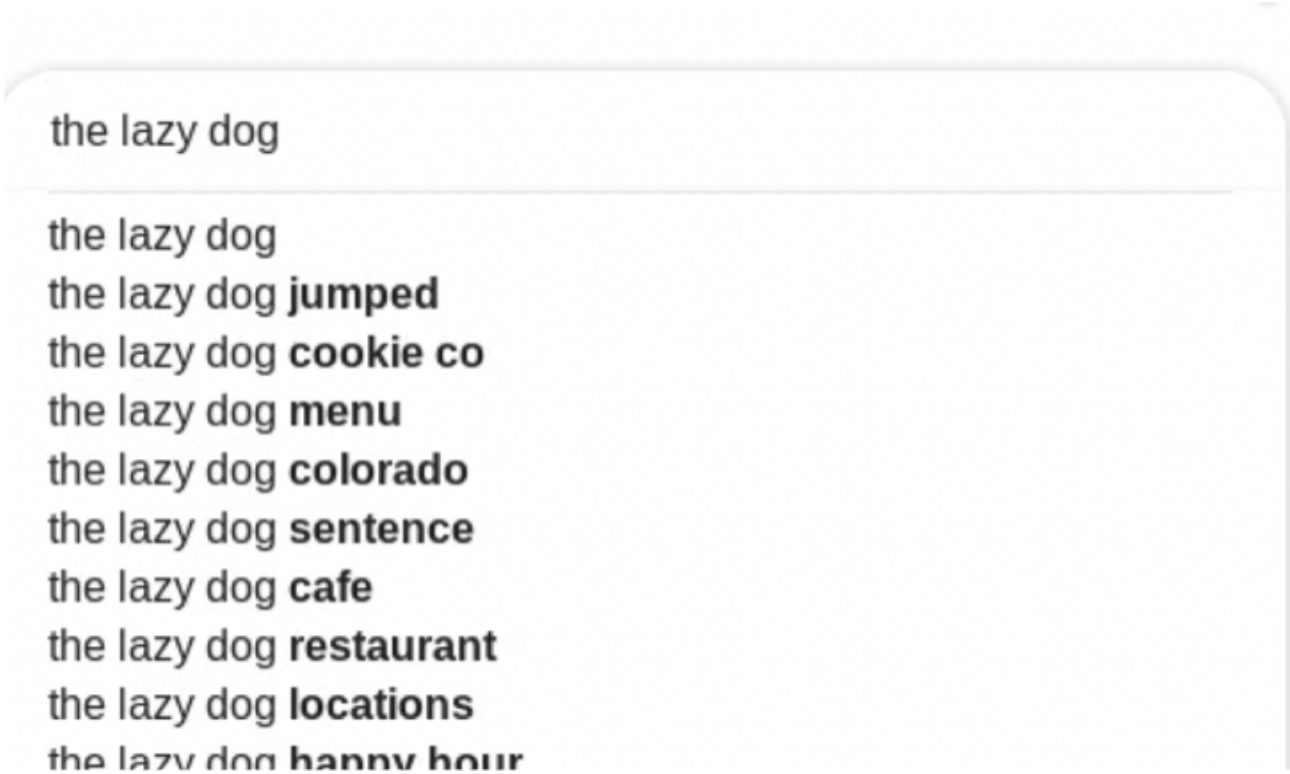
Vinnie Monaco

Naval Postgraduate School

Search engine autocomplete

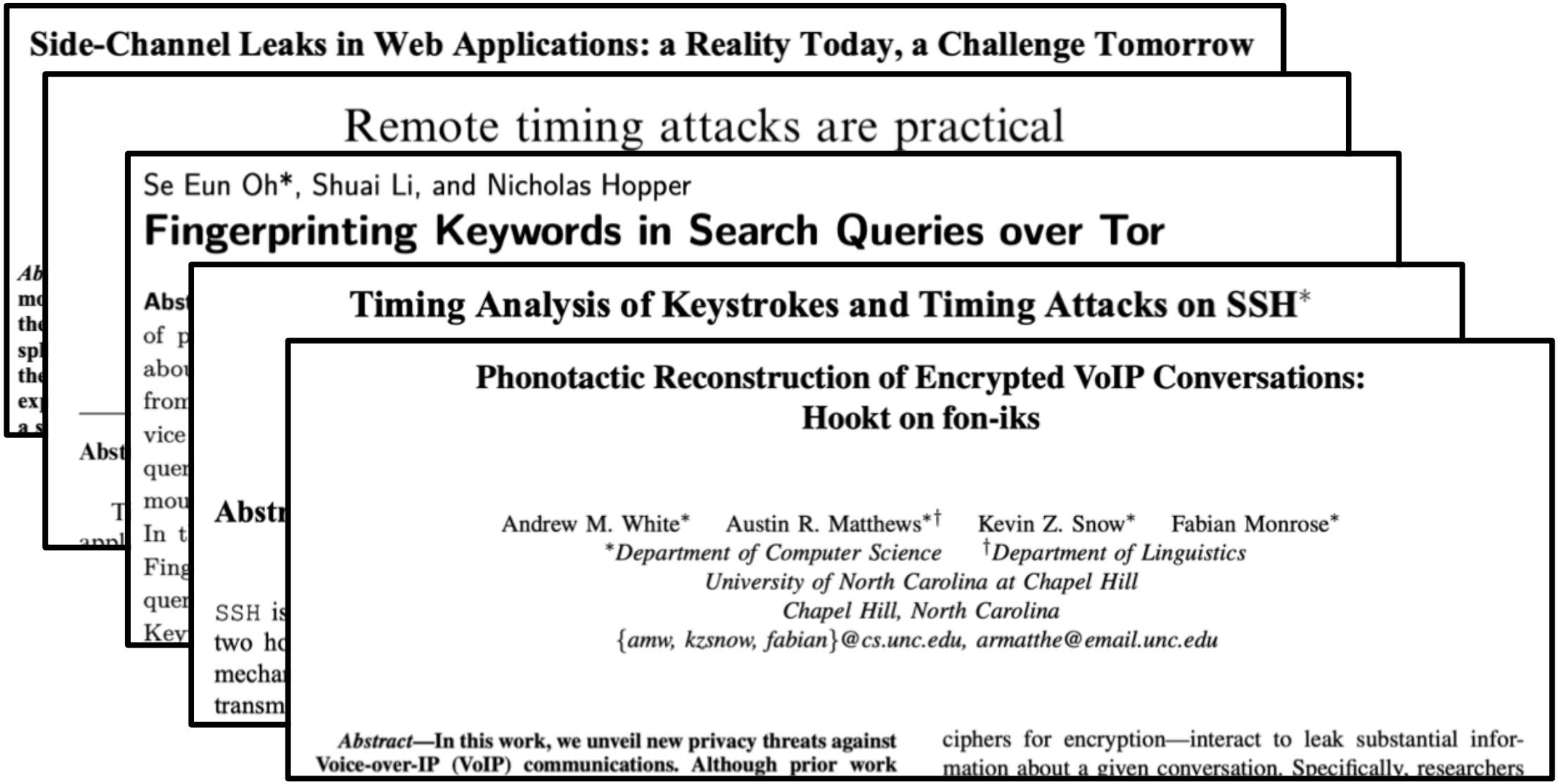
Search query

Packet capture



No.	Time	Protocol	TCP len	Info
8	8.865747280	TLSv1.2	151	Application Data
21	13.780190622	TLSv1.2	157	Application Data
22	13.782895588	TLSv1.2	182	Application Data,
39	14.680043369	TLSv1.2	157	Application Data
48	15.227565960	TLSv1.2	158	Application Data
58	15.873758188	TLSv1.2	160	Application Data
71	16.687042194	TLSv1.2	161	Application Data
82	17.746582385	TLSv1.2	162	Application Data
92	18.334356331	TLSv1.2	162	Application Data
101	18.910558934	TLSv1.2	163	Application Data
115	19.571843835	TLSv1.2	165	Application Data
125	20.564457628	TLSv1.2	167	Application Data
137	21.071393294	TLSv1.2	167	Application Data
147	21.627694121	TLSv1.2	168	Application Data

20 years of network side channels



Attack overview

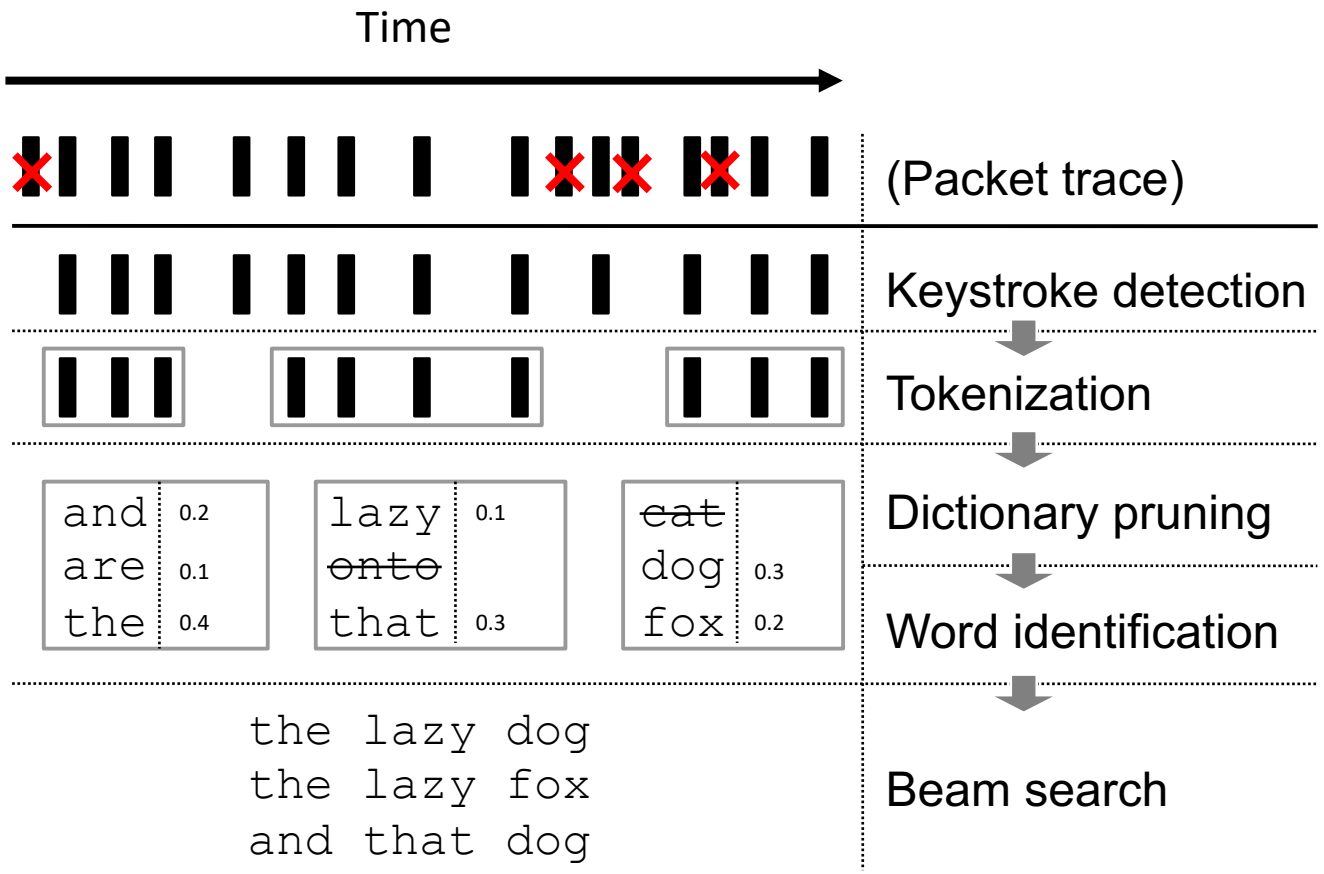
- Predict search queries using only client traffic

- Combine multiple independent weak predictors
  - Escaped URL characters
  - HTTP2 header compression
  - Key-press time intervals
  - Natural language

Threat model

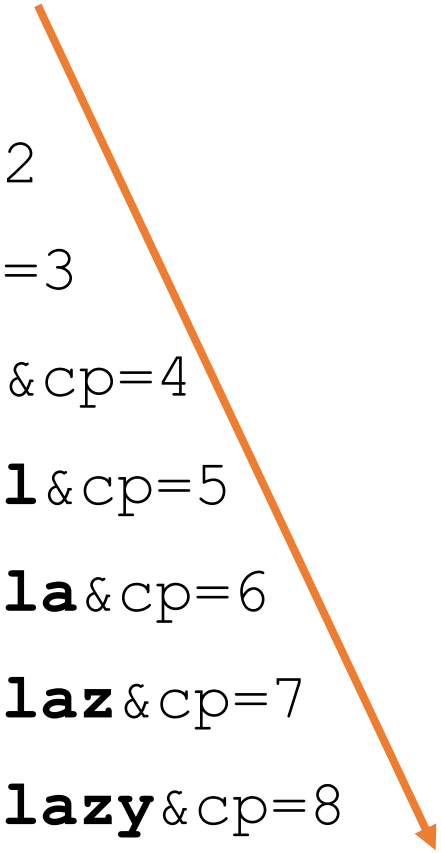
- Capture encrypted traffic at the NIC
- Victim types lowercase English letters + Space
  - No typos/backspace
- Autocomplete requests triggered by keydown events

Attack workflow

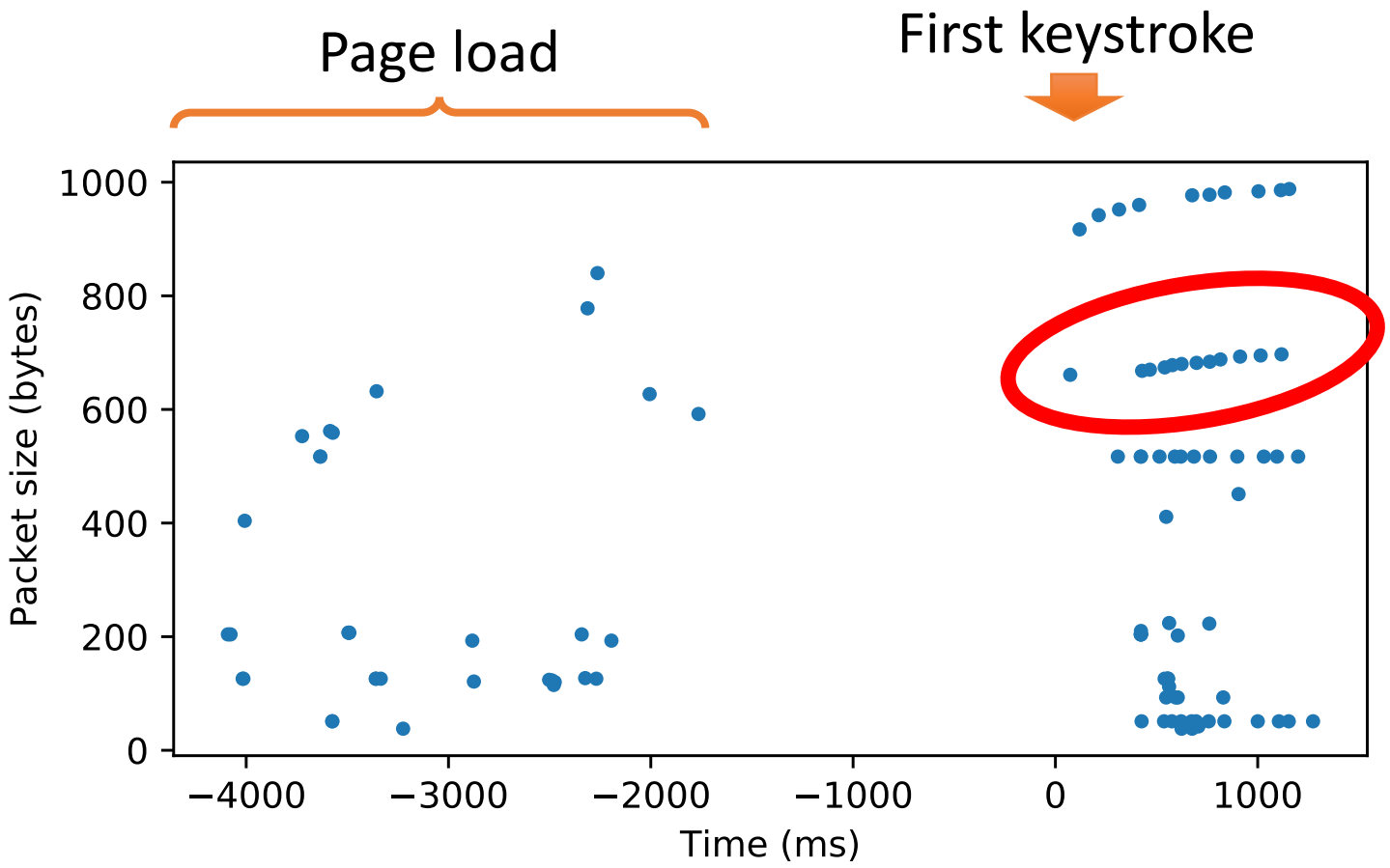


Autocomplete GET requests

```
GET /complete/search?q=t&cp=1
GET /complete/search?q=th&cp=2
GET /complete/search?q=the&cp=3
GET /complete/search?q=the%20&cp=4
GET /complete/search?q=the%20l&cp=5
GET /complete/search?q=the%20la&cp=6
GET /complete/search?q=the%20laz&cp=7
GET /complete/search?q=the%20lazy&cp=8
```



Keystroke detection



Baidu example: searching for “the lazy dog”

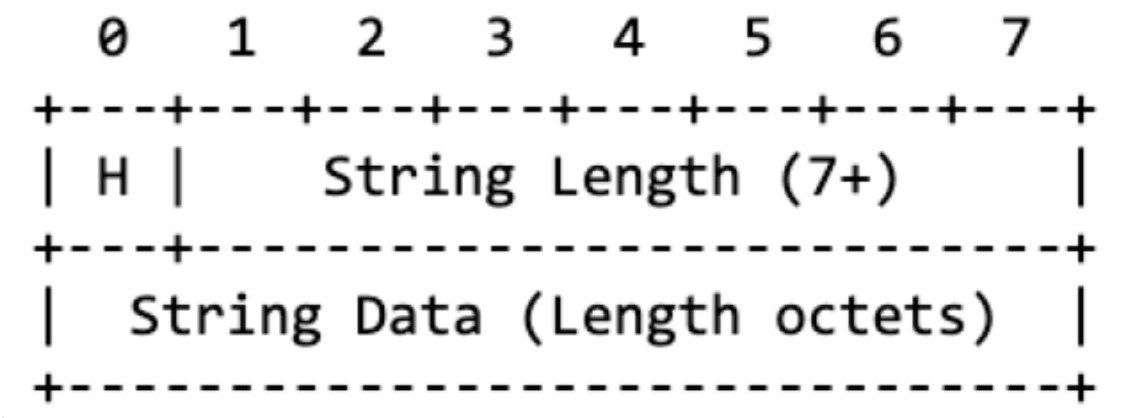
- Find the longest increasing subsequence (LIS) of packet sizes

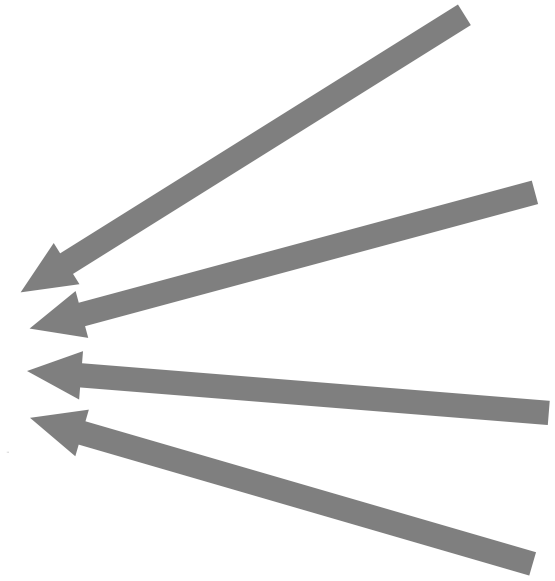
Tokenization

	Packet size difference
GET /complete/search?q=**t**&cp=1	
GET /complete/search?q=**th**&cp=2	+1
GET /complete/search?q=**the**&cp=3	+1
GET /complete/search?q=**the%20**&cp=4	**+3**
GET /complete/search?q=**the%20l**&cp=5	+1
GET /complete/search?q=**the%20la**&cp=6	+1
GET /complete/search?q=**the%20laz**&cp=7	+1
GET /complete/search?q=**the%20lazy**&cp=8	+1

HPACK (HTTP2 header compression)

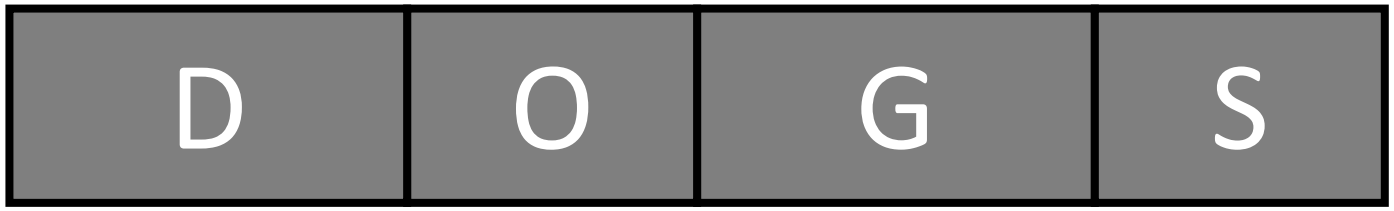
Static Huffman Encoding



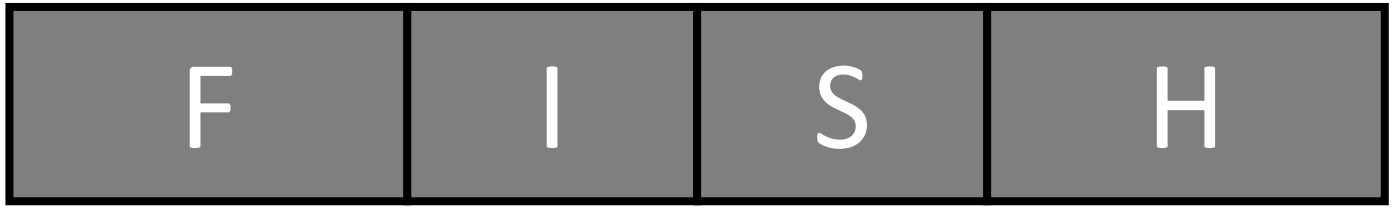


```
'a' ( 97)  |00011
'b' ( 98)  |100011
'c' ( 99)  |00100
'd' (100)  |100100
'e' (101)  |00101
'f' (102)  |100101
'g' (103)  |100110
'h' (104)  |100111
'i' (105)  |00110
'j' (106)  |1110100
'k' (107)  |1110101
'l' (108)  |101000
'm' (109)  |101001
'n' (110)  |101010
'o' (111)  |00111
'p' (112)  |101011
'q' (113)  |1110110
'r' (114)  |101100
's' (115)  |01000
't' (116)  |01001
'u' (117)  |101101
'v' (118)  |1110111
'w' (119)  |1111000
'x' (120)  |1111001
'y' (121)  |1111010
'z' (122)  |1111011
```

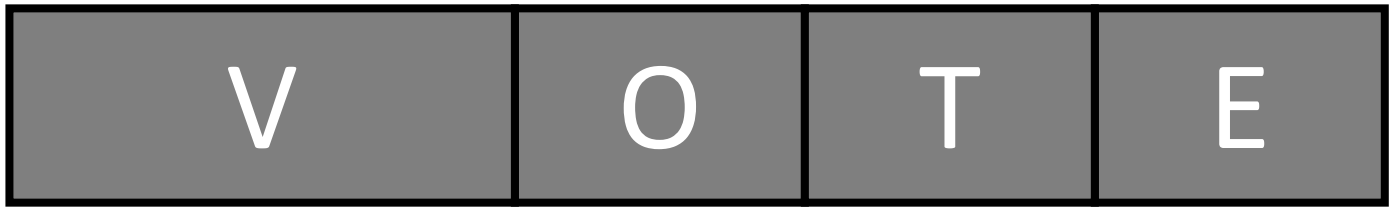
PETAL
(Preset Encoding Table Information Leakage)



6 + 5 + 6 + 5 = 22 bits

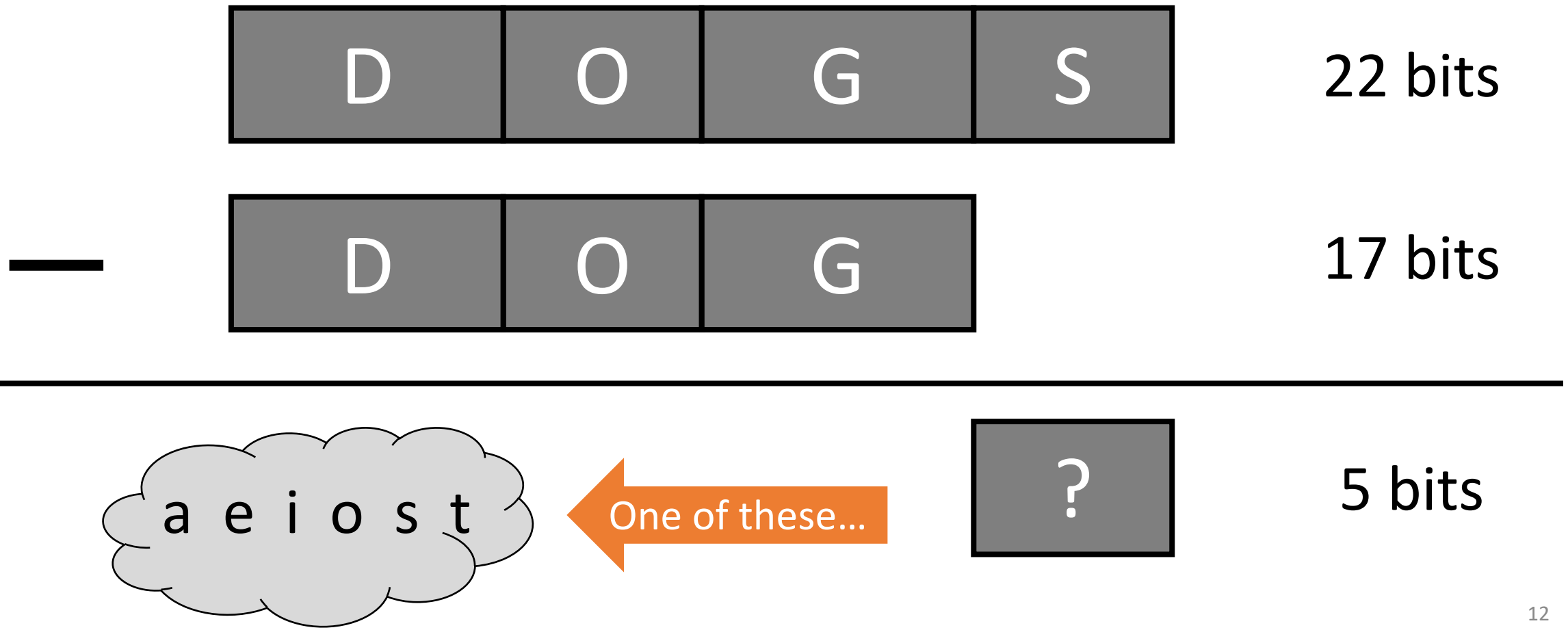


6 + 5 + 5 + 6 = 22 bits

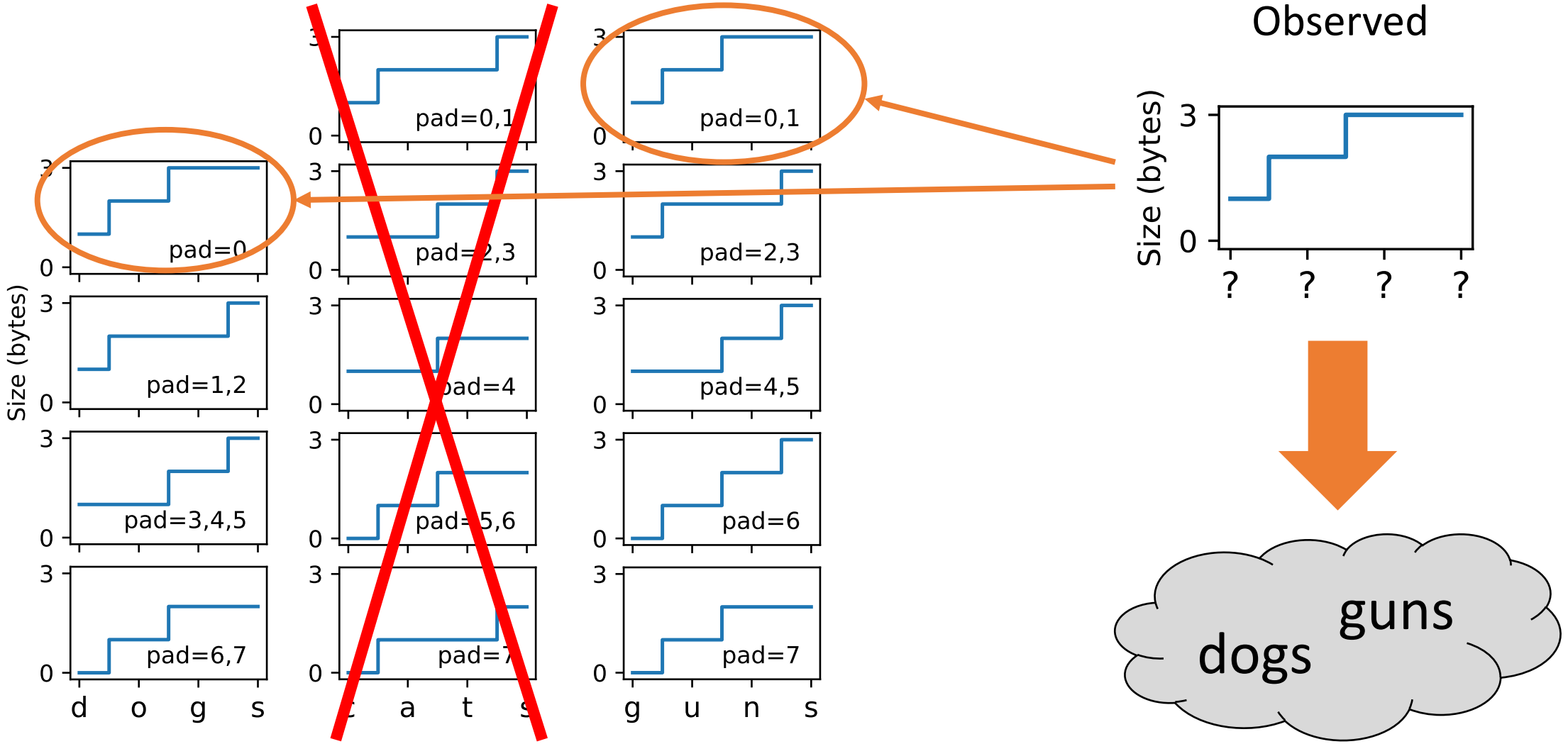


7 + 5 + 5 + 5 = 22 bits

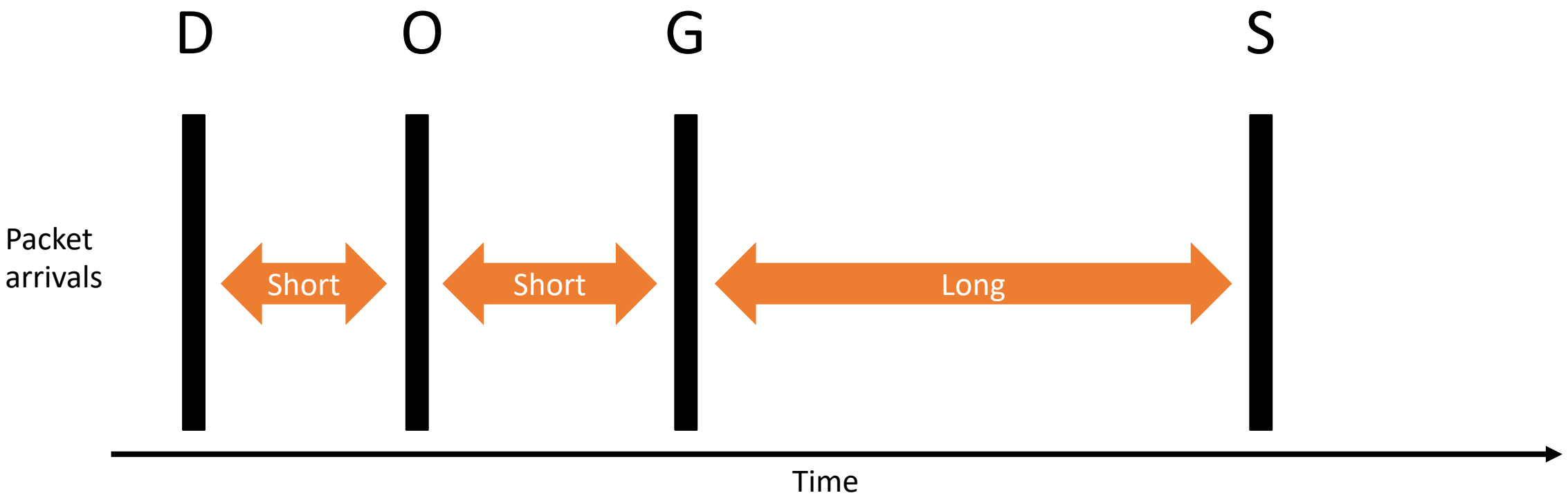
Incremental compression



Dictionary pruning



Word identification



- Use a BiRNN to predict keys

Language model and beam search

Which word comes next?

> the lazy ____

1) dog
2) car
3) hat
4) big

Top 50 hypotheses

```
the lazy dog
the blue car
and some fox
    …
how they run
```

Data collection and results

- Data collect
  - Browser automation with Selenium
  - Replay keystrokes with `uinput`
  - 4k unique queries
  - 2 search engines (Google, Baidu)
  - 2 browsers (Chrome, Firefox)
  - 16k total queries recorded
- Keystroke detection and tokenization accuracy
  - > 99% (Google and Baidu)
- Top-50 classification accuracy (entire query is correct)
  - 15% (Google)
  - 13% (Baidu)

Example

Truth
he is recovering from a sprained

Good hypotheses
he is recovering from a sprained
he is recovering from a strained

Bad hypotheses
to be president from a position
is to learn from such a position

Conclusions

- This attack has many of moving parts…
  - Several independent weak side channels combine to create a strong one
- Language modeling is key
  - The predictability of human behavior is difficult to mask
- Where else does incremental compression occur?
  - Thin clients/websites with autosave feature?
  - Mapping services (latitude/longitude changes incrementally)?

Thank you

- Source code

  kreep (keystroke recognition and entropy elimination program)
  https://github.com/vmonaco/kreep

- Contact me

  https://vmonaco.com

- Questions?