Use of Generalized Hough Transform on interpretation of memory dumps

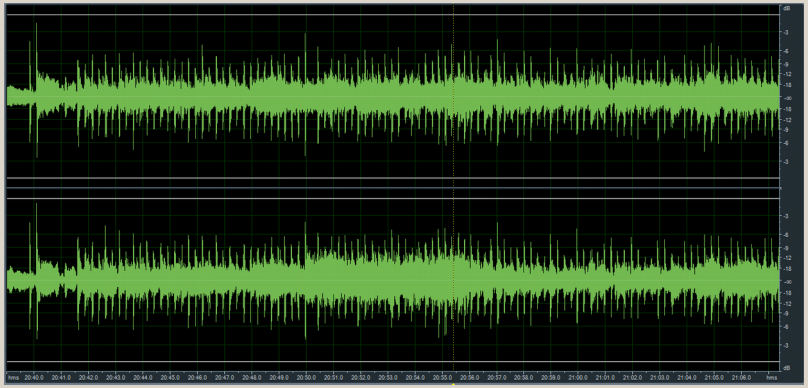
Paulo R. Nunes de Souza

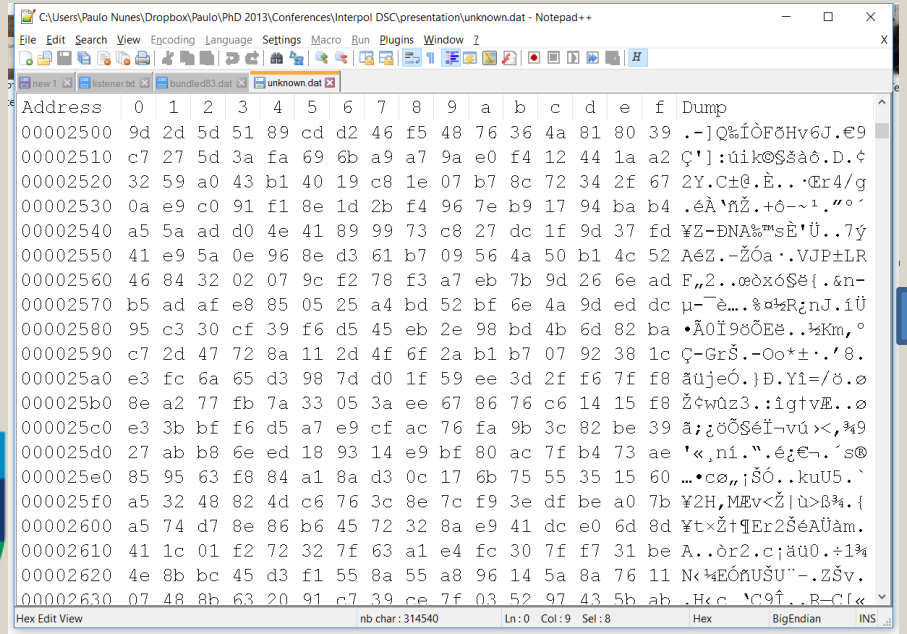
Pavel Gladyshev

What is partial data?

- Incomplete data
- Partially corrupted data
- Data without required code/algorithm
- Data of unknown source

What does it look like?





Word puzzle

SCIENCE OF SECRETS

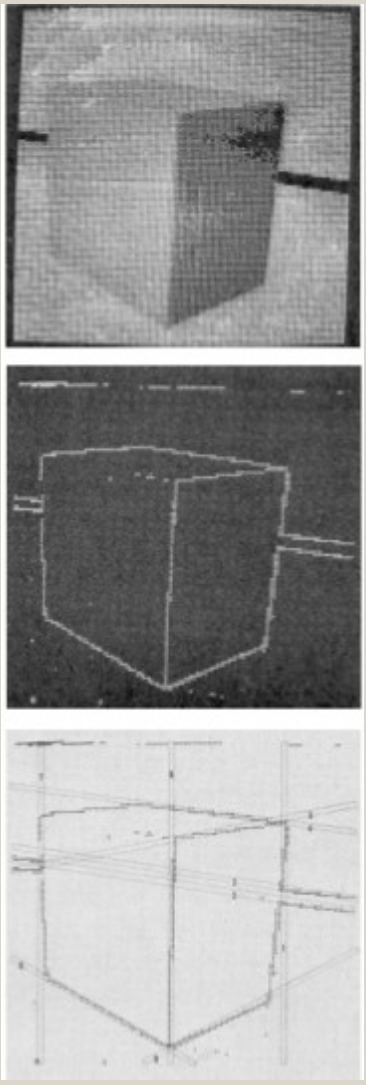
C O E L C J Y V F O D K G U L G K P T D
M K V U C O R S A T E L L I T E O B H N
R Z I J U U W U O N W W U F Y L G G I W
T M D X M V I O L E N T C R I M E L T F
E L E C T R O N I C E N O H P T R A M S
T A N R E I N I V S J E M W G E C M A X
Q F C J E S P T E E G G M U M Y M T A Z
U O E H D Z N A L P S A U L E Y C P K S
Z R G D A L B L O O D T N V V Q I S T Y
Q E S C K O Y E T U A E I L M M T R R H
T N N B C F O G S L U R G G R U O E E N
Z S U O I C O R E F D C A O A L I I D F
F I W S H B A N V D E E T Q L T B D W X
X C H A C P S Q R V N S E N V A I L O A
K W P S H G O A A D N M C G C I T O P H
T B B P S P H X H I P A T T E R N S N H
W G S O N O G R A M U I E W O L A C U V
L C T E R R O R I S M R M E C O L I G I
T P U J H Q B S L N I X I H D E C O D E
I B J O F P R E D A T O R V F Q L U C G

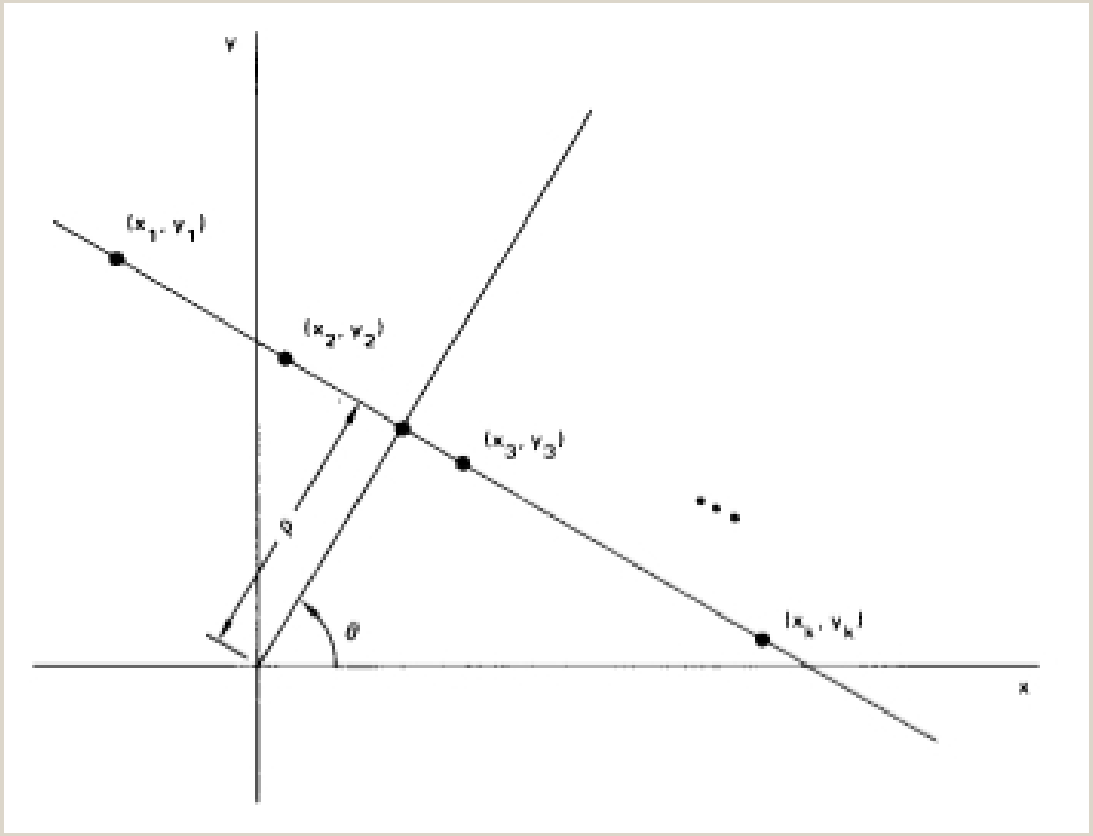
ANTIBIOTIC
BACTERIA
BLOOD
BRAINSTORM
CHICKADEE
COLLEGE
COMMUNICATE
DECODE
E COLI
ELECTRONIC
EVIDENCE
FEROCIOUS
FLUORESCENT
FORENSIC
GELATINOUS
GUNPOWDER
HARD DRIVE
HARVEST
INTERNET
INVESTIGATION
JELLYFISH
MERLIN
PATTERNS
PREDATOR
SATELLITE
SAXOPHONE
SECRET AGENT
SMART PHONE
SOLDIERS
SONOGRAM
STROLL
TERRORISM
VIOLENT CRIME
WIGGLY

Requirements

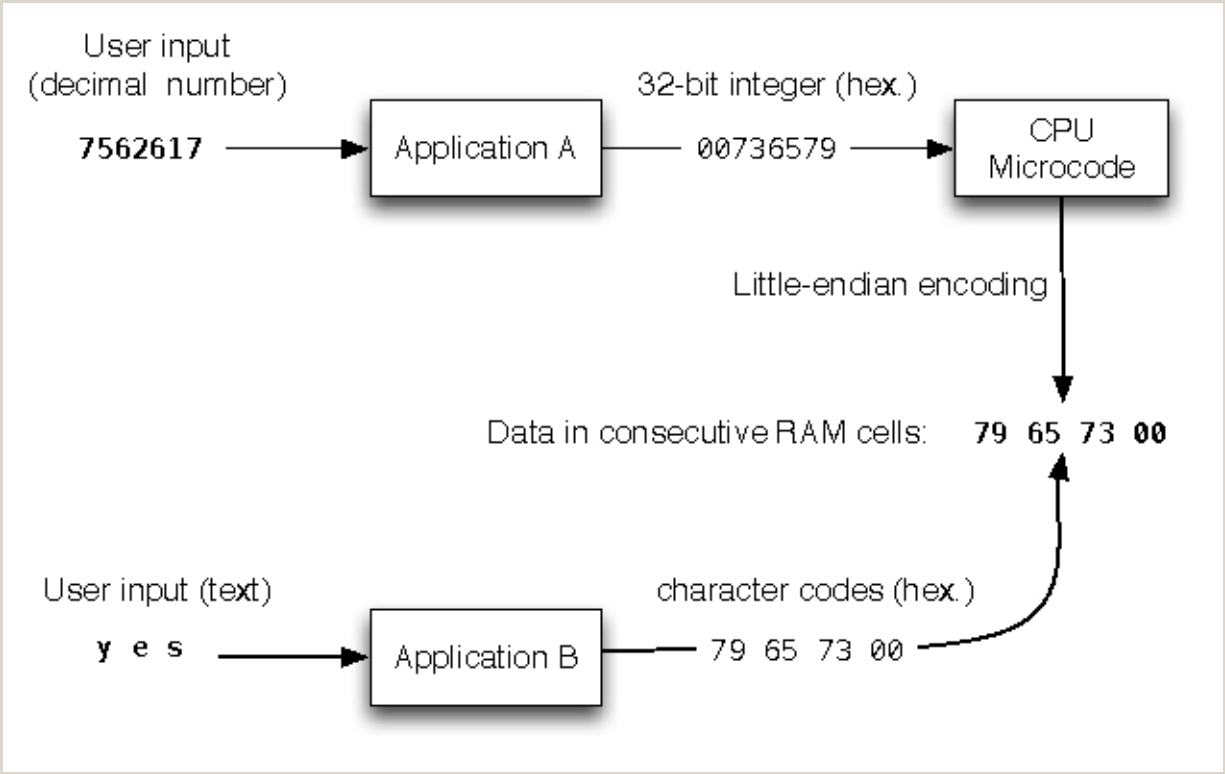
- Tolerance to noisy data;
- Tolerance to partial data corruption;
- Flexibility to define the structure being searched.
- Suitable candidate: **Hough Transform**

Hough Transform

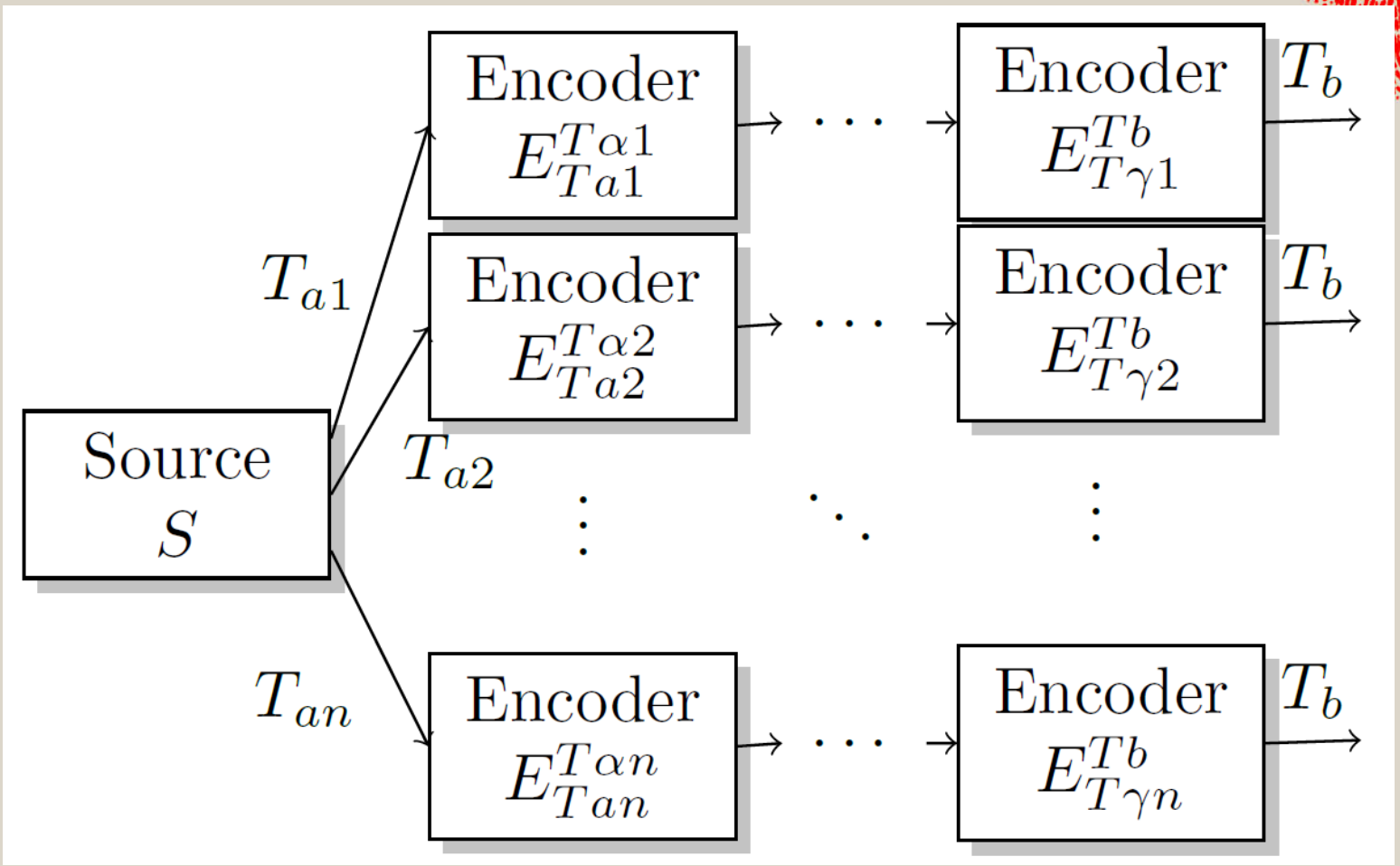




Encoding


User input
(decimal number)
7562617
Application A
32-bit integer (hex.)
00736579
CPU
Microcode
Little-endian encoding
Data in consecutive RAM cells:
79 65 73 00
User input (text)
y e s
Application B
character codes (hex.)
79 65 73 00

Model



$$T = \{Tbin, Tint8, Tascii, \cdots\}$$

Model

$$P_{E_{T_a}^{T_b}}(o|i) = \sum_{\substack{\sigma \in T_\alpha \\ \eta \in T_\beta \\ \mu \in T_\gamma}} P_{E_{T_a}^{T_\alpha}}(\sigma|i) \cdot P_{E_{T_\alpha}^{T_\beta}}(\eta|\sigma) \ldots P_{E_{T_\gamma}^{T_b}}(o|\mu)$$

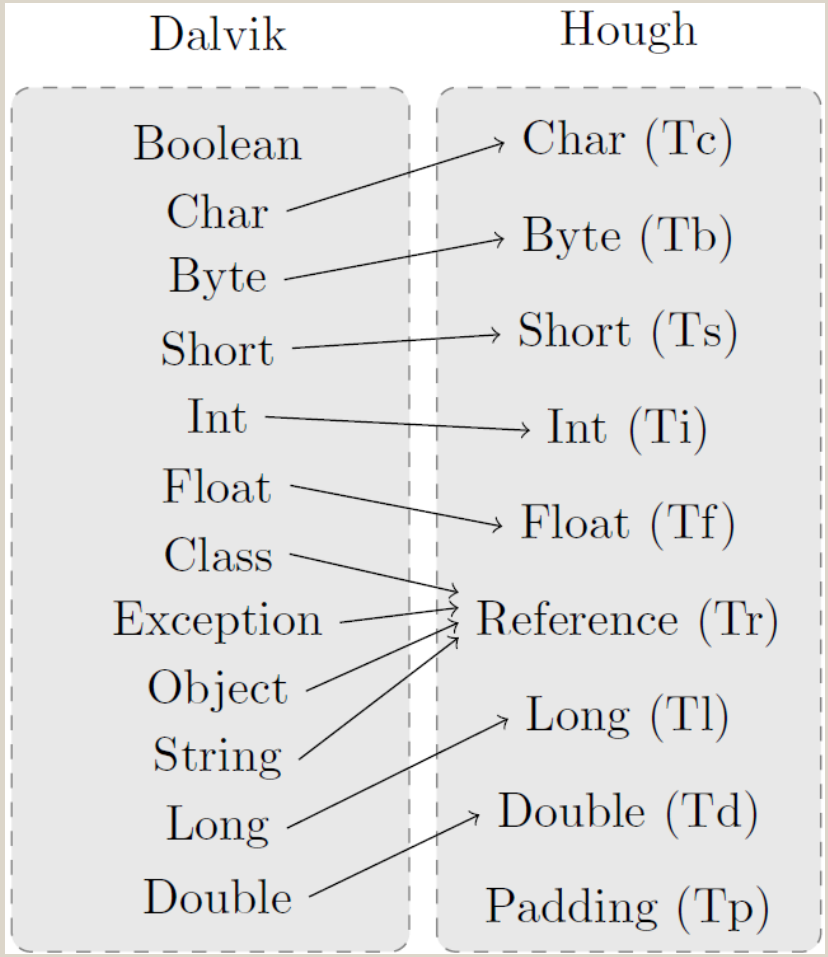
$$P_{E^{T_b}}(o|i,t) = \begin{cases} P_{E_t^{T_b}}(o|i) & \text{if } i \in t \\ 0 & \text{if } i \notin t \end{cases}$$

$$P_{SE^{T_b}}(w|k) = \prod_{j=0}^{q} \sum_{\forall \lambda \in k(j)} P_{E^{T_b}}(w(j)|\lambda, k(j))$$

$$P_{SE}(k|w) = \frac{P_{SE}(w|k) \cdot P_{SE}(k)}{P_{SE}(w)}$$

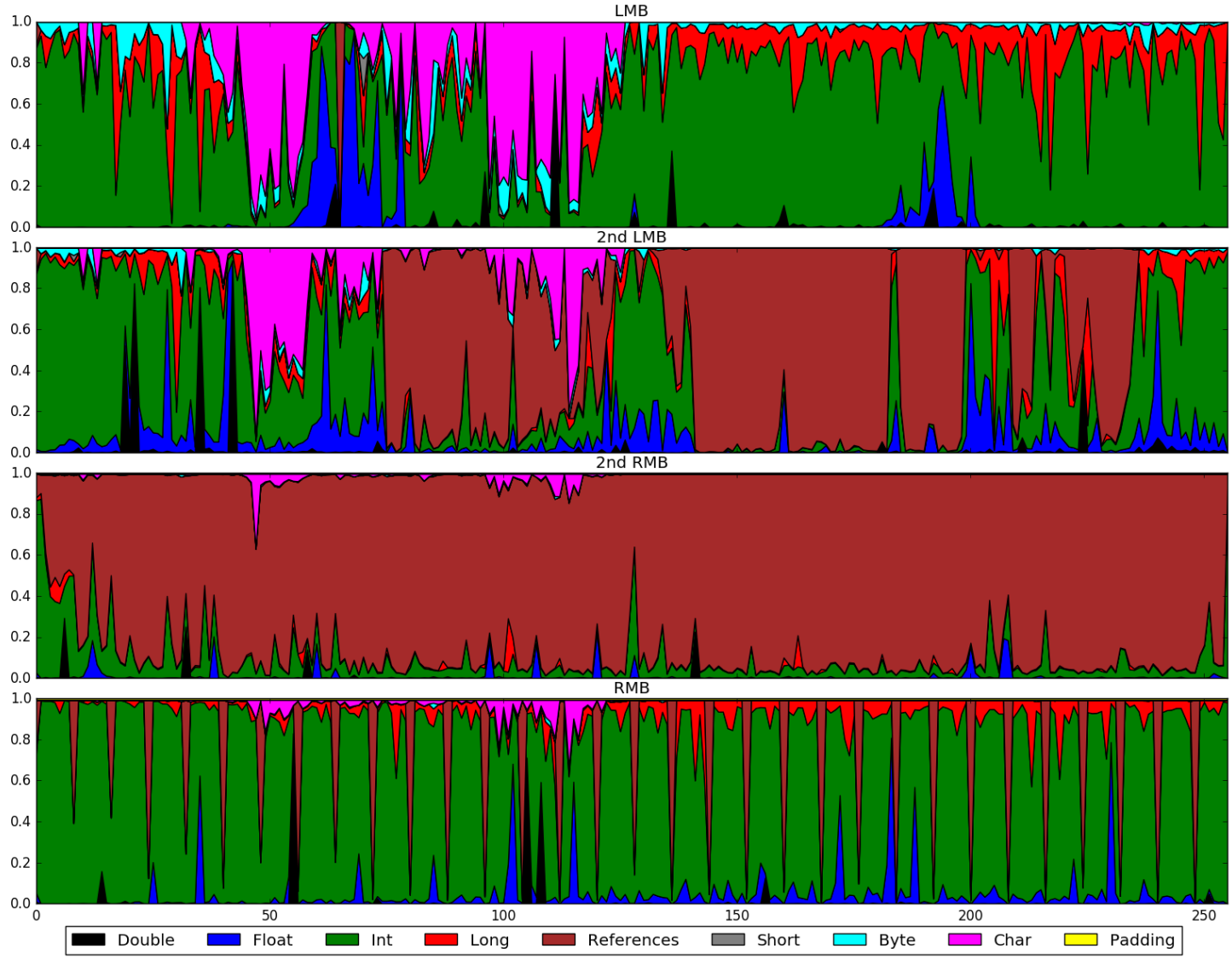
UCD DUBLIN

Data types

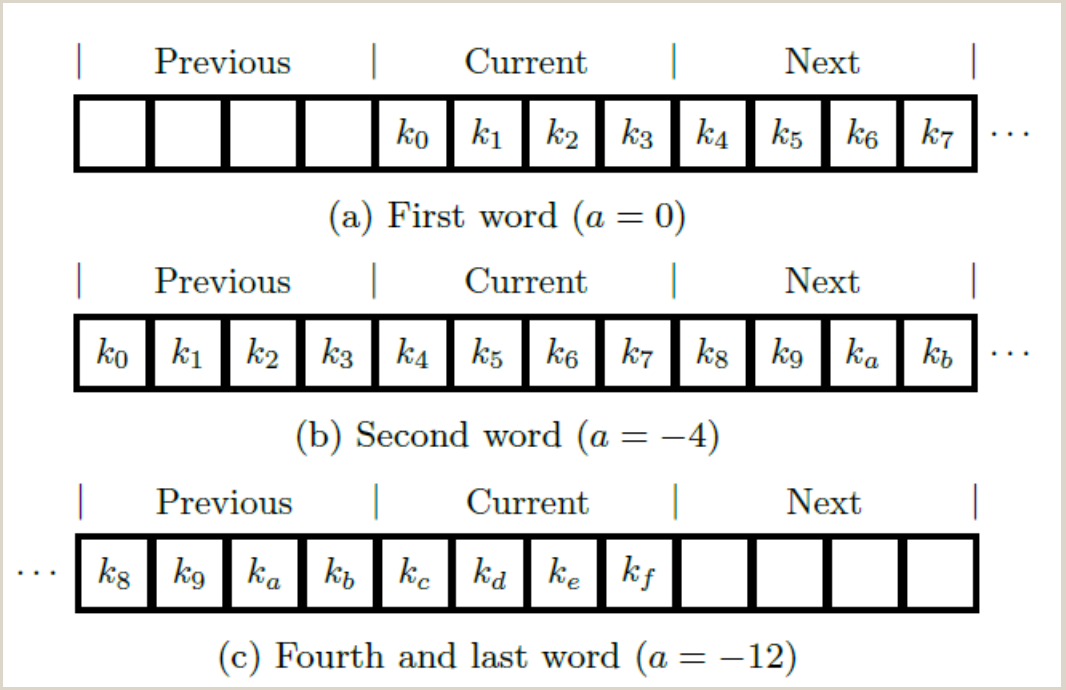


$$T = \{Tbin, Tb, Tc, Td, Tf, Ti, Tl, Tp, Tr, Ts\}$$

Probabilities $P_{se}(k|w,b)$

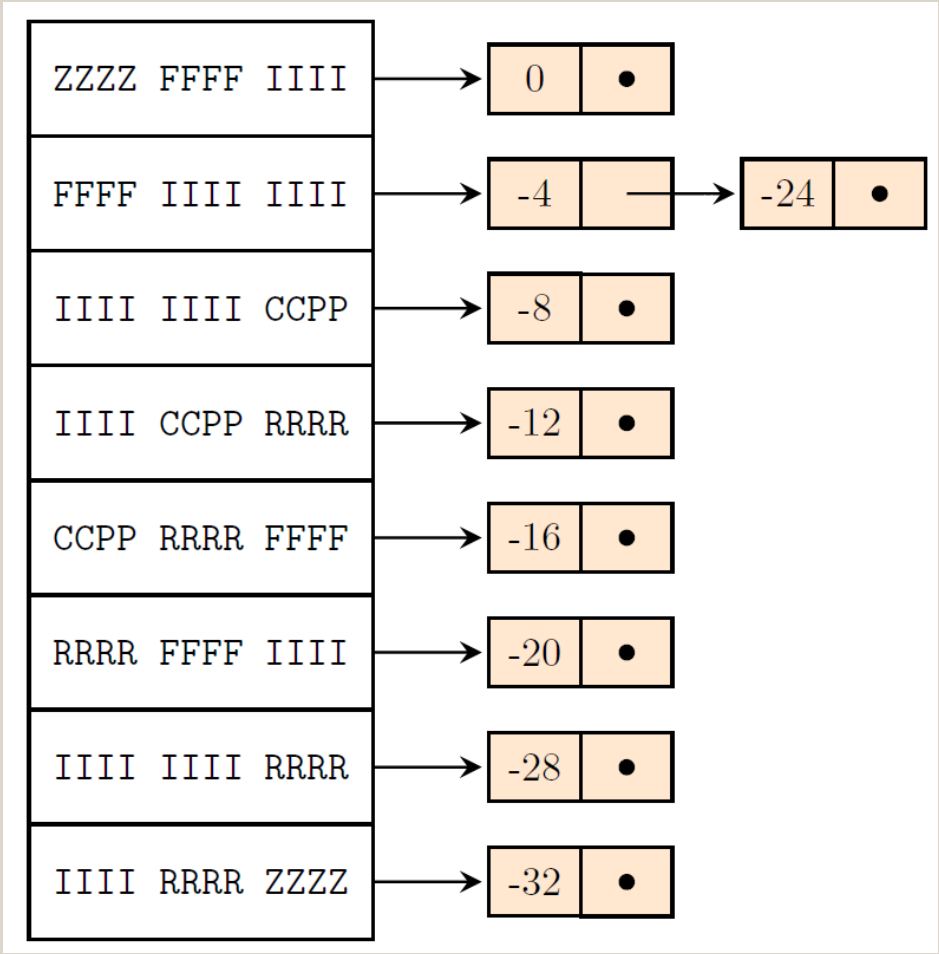


R-table

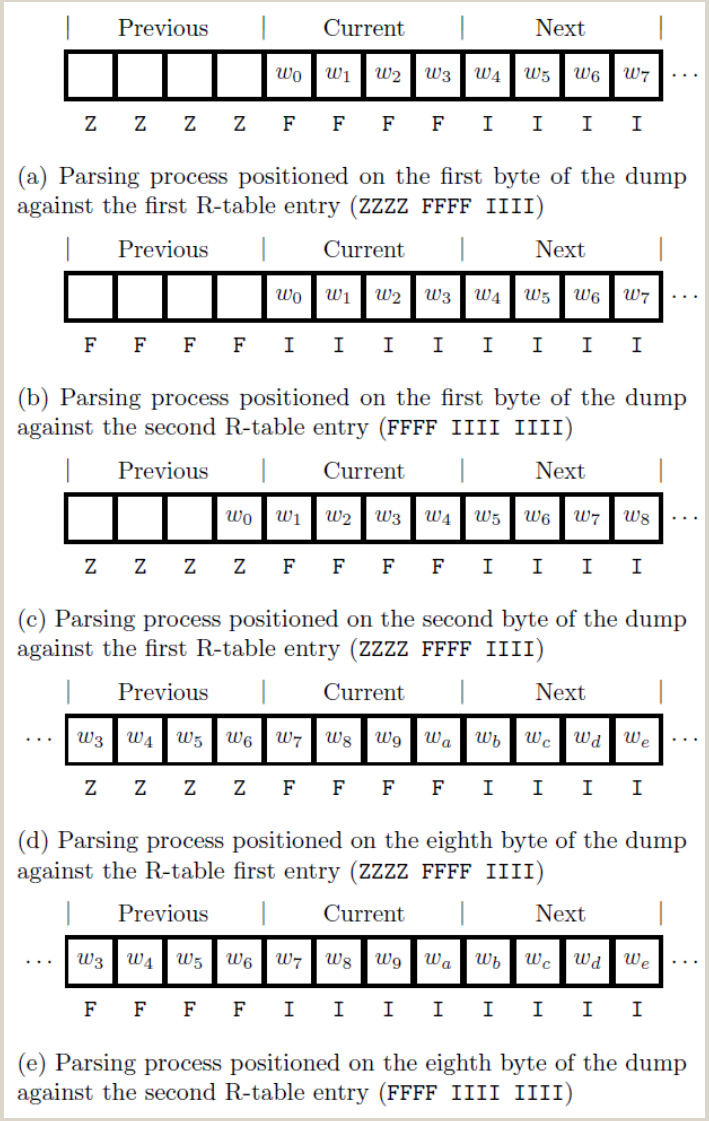

Previous
Current
Next
k_0 k_1 k_2 k_3 k_4 k_5 k_6 k_7 ⋯
(a) First word ($a = 0$)
Previous
Current
Next
k_0 k_1 k_2 k_3 k_4 k_5 k_6 k_7 k_8 k_9 k_a k_b ⋯
(b) Second word ($a = -4$)
Previous
Current
Next
⋯ k_8 k_9 k_a k_b k_c k_d k_e k_f
(c) Fourth and last word ($a = -12$)

R-table

FFFF IIII IIII CCPP RRRR FFFF IIII IIII RRRR



Accumulation



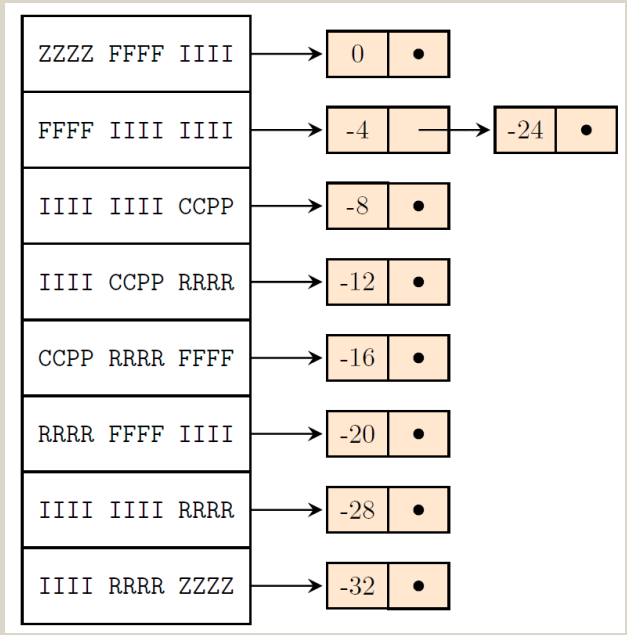
(a) Parsing process positioned on the first byte of the dump against the first R-table entry (`ZZZZ FFFF IIII`)

(b) Parsing process positioned on the first byte of the dump against the second R-table entry (`FFFF IIII IIII`)

(c) Parsing process positioned on the second byte of the dump against the first R-table entry (`ZZZZ FFFF IIII`)

(d) Parsing process positioned on the eighth byte of the dump against the R-table first entry (`ZZZZ FFFF IIII`)

(e) Parsing process positioned on the eighth byte of the dump against the second R-table entry (`FFFF IIII IIII`)



$$
\begin{aligned}
C = & P_{SE}(F|w(3),0) \cdot P_{SE}(F|w(4),1) \cdot P_{SE}(F|w(5),2) \cdot \\
& P_{SE}(F|w(6),3) \cdot P_{SE}(I|w(7),0) \cdot P_{SE}(I|w(8),1) \cdot \\
& P_{SE}(I|w(9),2) \cdot P_{SE}(I|w(a),3) \cdot P_{SE}(I|w(b),0) \cdot \\
& P_{SE}(I|w(c),1) \cdot P_{SE}(I|w(d),2) \cdot P_{SE}(I|w(e),3)
\end{aligned}
$$

If $C > C_t$, increment accumulation table at position *i+a*.

Peak detection

- Local maxima above threshold H
- Positions of dump where structure was identified

Tests

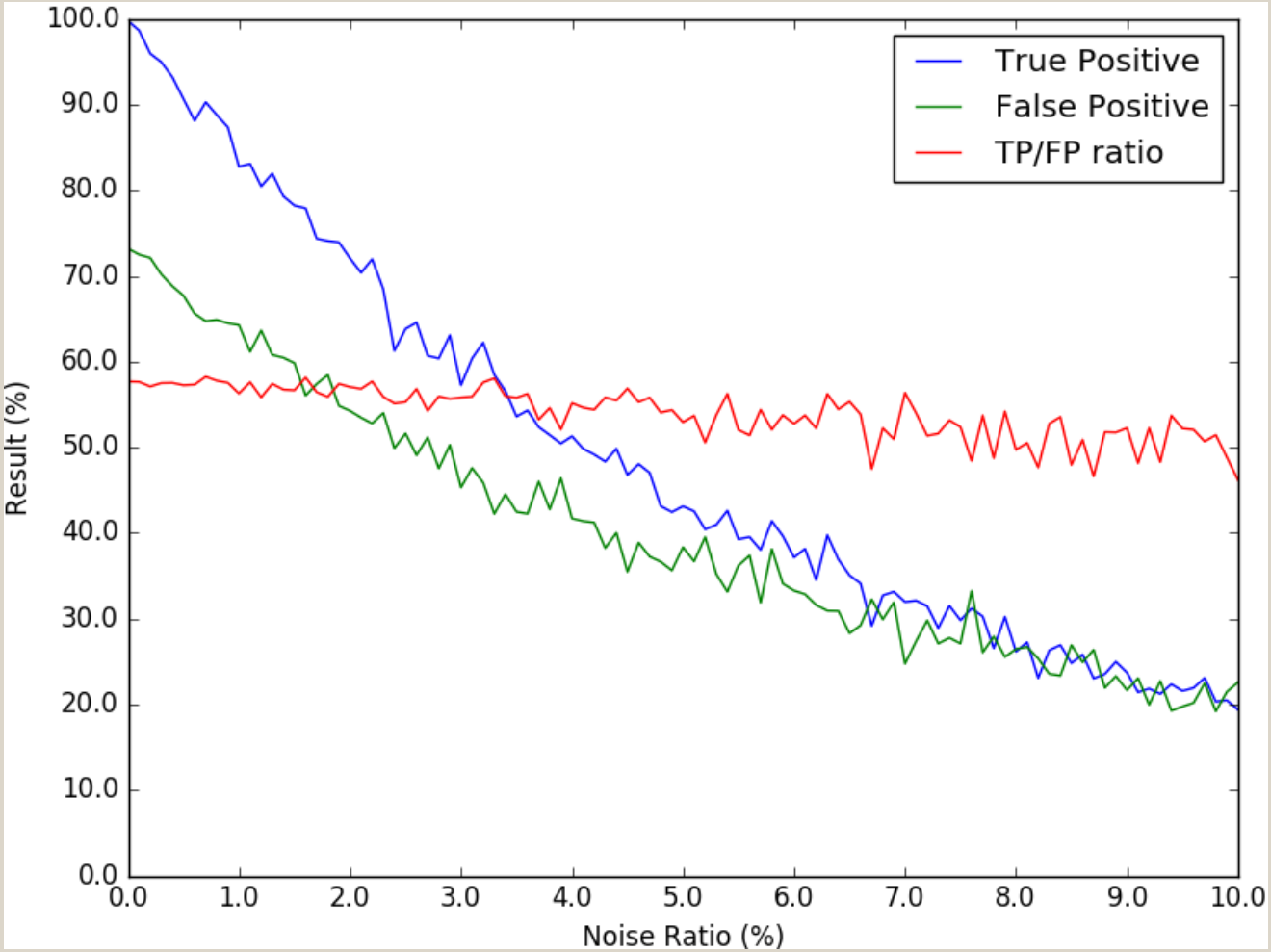
- Randomly created 100 distinct dumps
- Each dump with 4 kb in size
- Randomly choose one of the types in T
- Then a random value of the chosen type was inserted in the dump
- Repeat this until a 4kb dump was created
- The last step was to insert instances of the structure of interest across the previously created dumps.
- The position and quantity of those structures were also randomly chosen
- Each dump and the positions of the structure of interest were saved to respective files

Tests

- For each of the 100 dumps, another 100 versions of them were created
- One file for each level of added noise. Starting at 0% for each 0.1% step until reaching 10%
- At the end, we have 101 versions of each of the 100 dumps
- Total of 10100 test dumps

UCD DUBLIN

Results

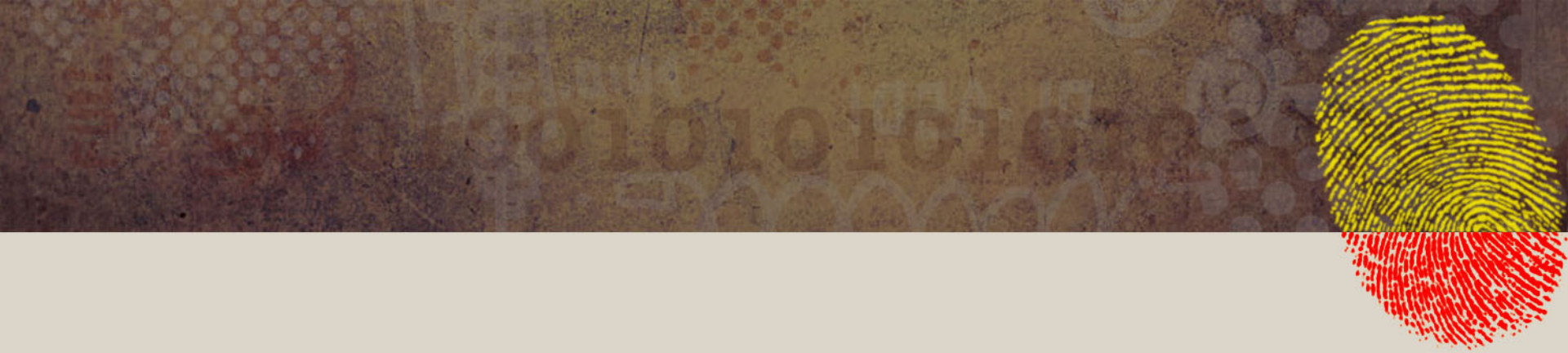

True Positive
False Positive
TP/FP ratio
Result (%)
Noise Ratio (%)
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0

Conclusions

- Tolerance to noisy data
- Flexibility to identify structure of interest
- The structure of interest was correctly spotted in 99.8% of the tests with no noise
- The structure of interest was correctly spotted in 20% of the cases with 10% of noise.
- The downside is the high false positive rate.
- Applicable beyond memory realm

Follow ups

- Hough-Forensic DSI
- PNG filecarving



Thank you

UCD DUBLIN