ILLC Project Course in Information Theory

Crash course

13 January – 17 January 2014
12:00 to 14:00

Student presentations

27 January – 31 January 2014
12:00 to 14:00

Location

ILLC, room F1.15,
Science Park 107, Amsterdam

Materials

informationtheory.weebly.com

Contact

Mathias Winther Madsen
mathias.winther@gmail.com

Monday

Probability theory
Uncertainty and coding

Tuesday

The weak law of large numbers
The source coding theorem

Wednesday

Random processes
Arithmetic coding

Thursday

Divergence
Kelly Gambling

Friday

Kolmogorov Complexity
The limits of statistics

The **surprisal** associated with an event *A* is

$$s(A) \;==\; \log \frac{1}{\Pr(\mathrm{A})}.$$

Surprisal == ideal codeword length.

Entropy == expected surprisal.

x	1	2	3	4
$\Pr(X = x)$	1/2	1/4	1/8	1/8
$-\log \Pr(X = x)$	1	2	3	3

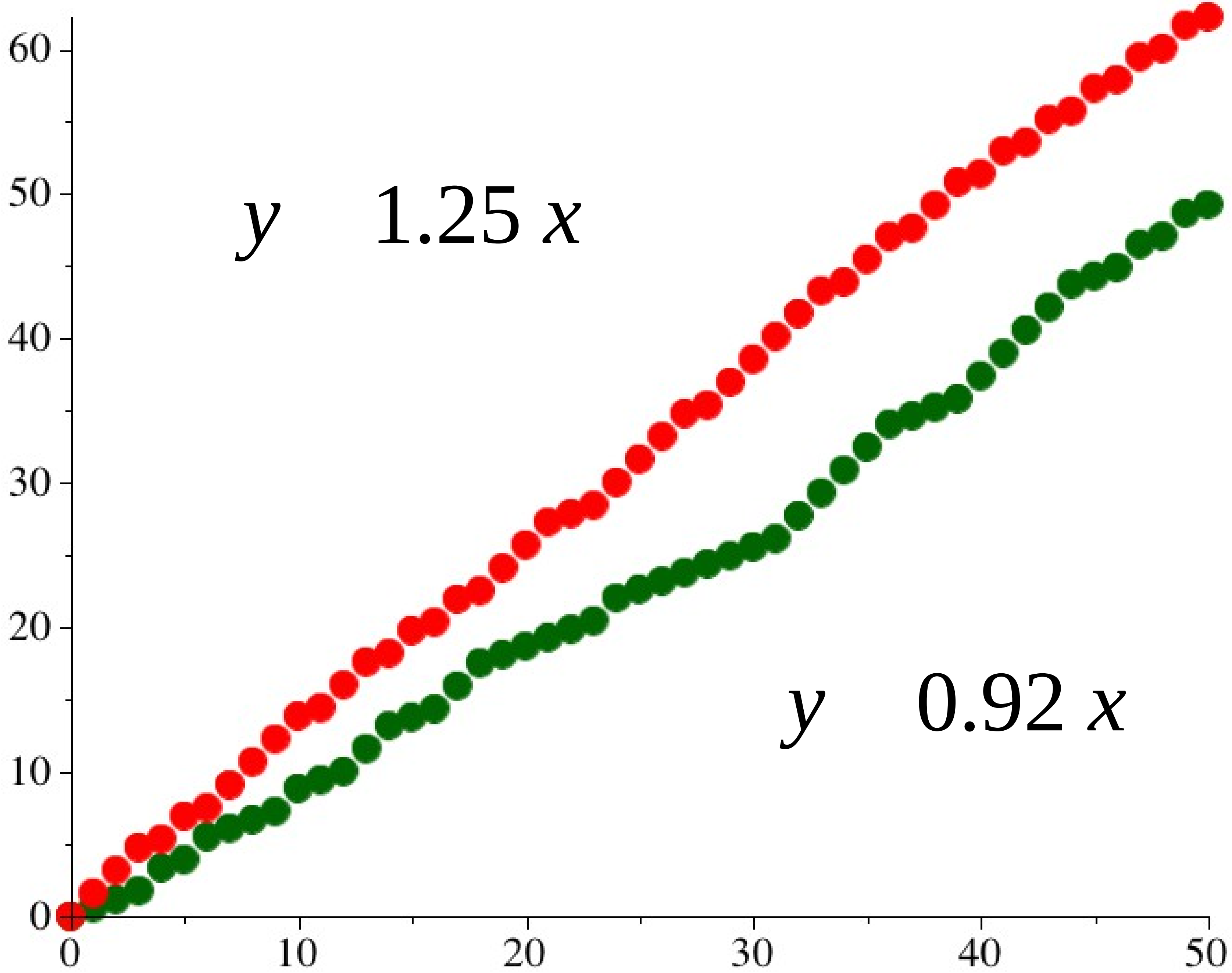
You have a bent coin with bias = 1/3 which you flip a large number of times, keeping track of your average surprisal so far.

What is the limit of your average surprisal?

Just before the experiment starts, however, I secretly secretly replace your coin by a different one with bias = 2/3.

What is the average of your surprisal values actually going to converge to?

How much higher is this average than what you could have achieved with better probability estimates?



$y \quad 1.25\,x$

$y \quad 0.92\,x$

Entropy of a probabilistic situation *P*:

$$H(P) \;==\; \mathrm{E}_P[\,-\log P(x)\,]$$

Average surprisal for model *Q* under *P*:

$$R(P \,\|\, Q) \;==\; \mathrm{E}_P[\,-\log Q(x)\,]$$

Divergence of *Q* from *P*:

$$D(P \,\|\, Q) \;==\; \mathrm{E}_P[\,-\log Q(x)\,] - H(P)$$

Solomon Kullback and Richard A. Leibler:
“On information and sufficiency,”
Annals of Mathematical Statistics, 1951.

Example:

x	a	b
$P(x)$	1	0
$Q(x)$	1/2	1/2

$$D(P \,\|\, Q) \;==\; ?$$

$$D(Q \,\|\, P) \;==\; ?$$

Example:

x	a	b
$P(x)$	1	0
$Q(x)$	1/2	1/2

$$D(P \parallel Q) \;==\; 1$$

$$D(Q \parallel P) \;==$$

Character	'a'	'b'	'c'
Wrong probability	1/2	1/4	1/4
Bad codeword	'0'	'10'	'11'
Actual probability	1/4	1/2	1/4
Good codeword	'00'	'1'	'01'

Input stream	b	b	a	b	c	b	…
Bad encoding	10	10	0	10	11	10	…
Good encoding	1	1	00	1	01	1	…

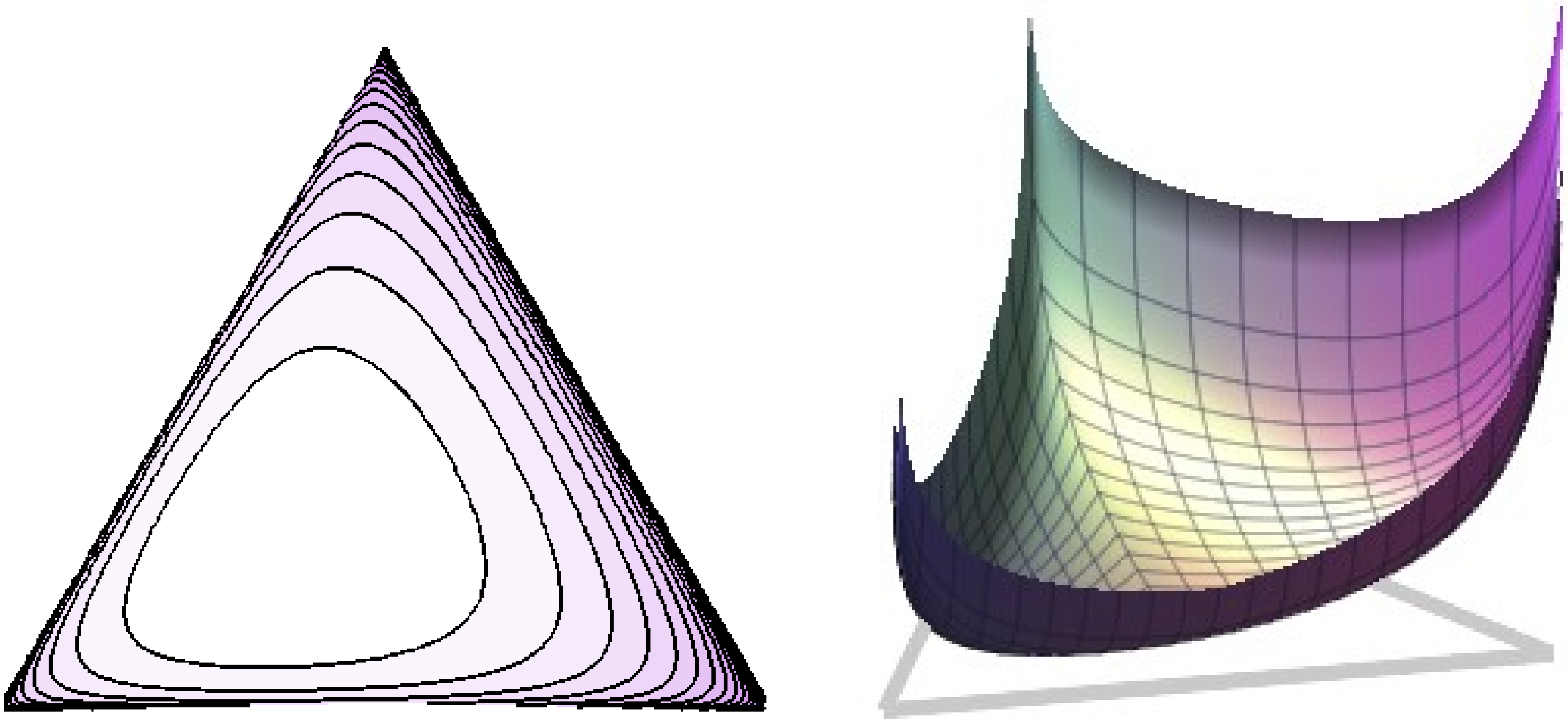
$$R(P \,\|\, Q) \qquad H(P),$$

and thus

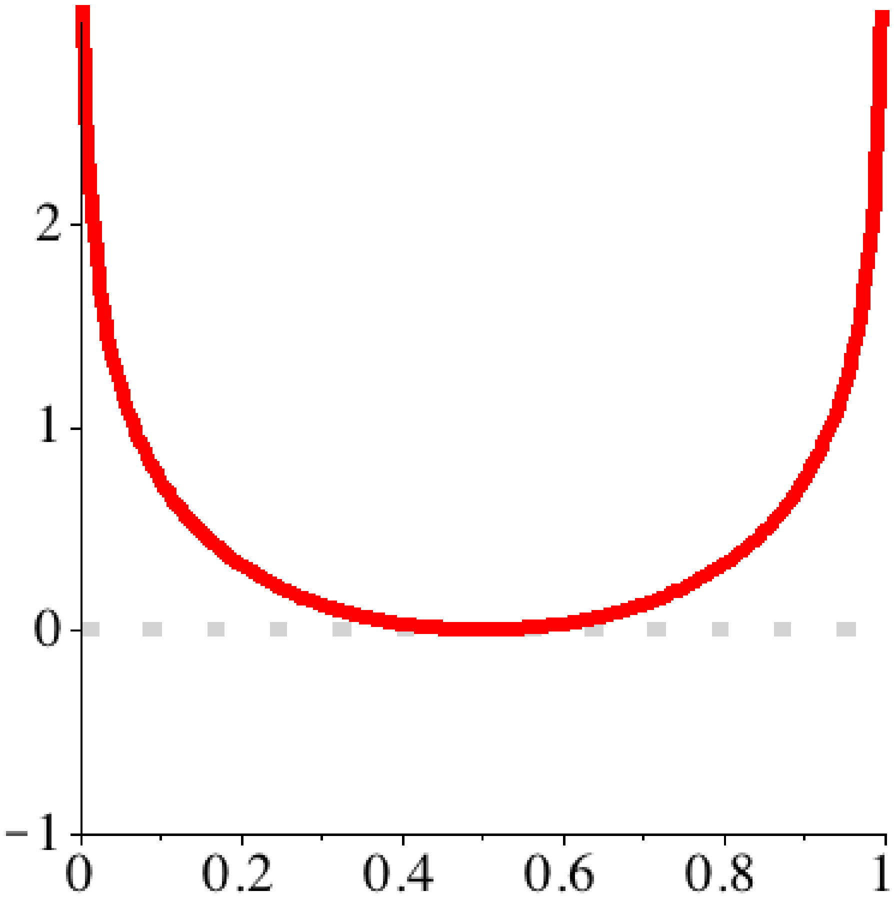
$$D(P \,\|\, Q) \qquad 0.$$

In fact, $D(P \,\|\, Q) = 0$ only if $P = Q$.

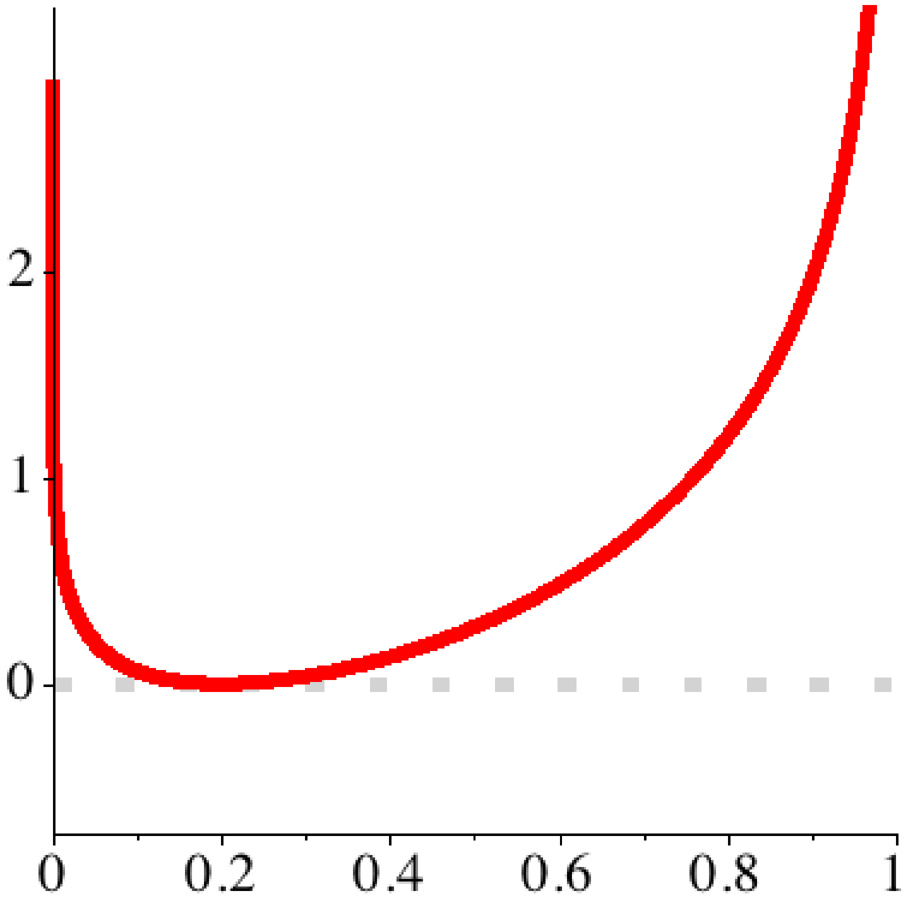
This follows from Jensen's inequality. (http://en.wikipedia.org/wiki/Jensen%27s_inequality)



Example: Divergence from (1/2, 1/4, 1/4).



From (1/2, 1/2)



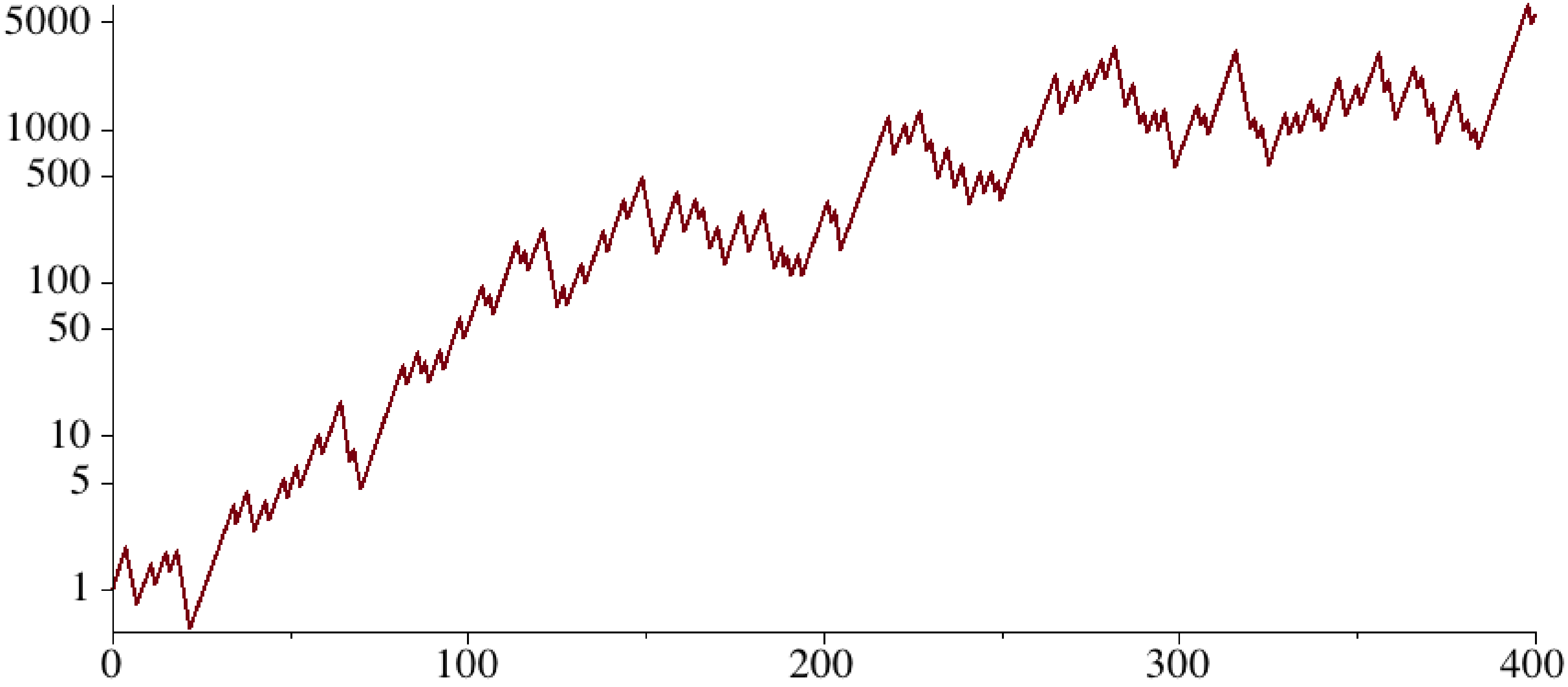
From (1/5, 4/5)

Likelihood ratios

$$D(P \parallel T) - D(P \parallel U) == \mathrm{E}_P\left(\log \frac{U(x)}{T(x)}\right)$$

Likelihood ratios

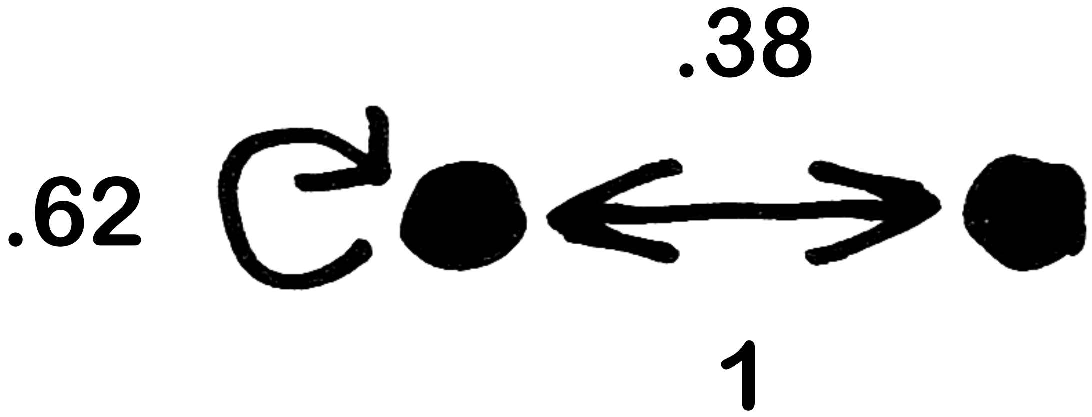
$$D(P \,||\, T) - D(P \,||\, U) \;==\; \mathrm{E}_P\left(\log \frac{U(x)}{T(x)}\right)$$



U claims:
= .700

T claims:
= .600

Actual:
= .667



$$\begin{pmatrix} .62 & 1 \\ .38 & 0 \end{pmatrix}$$

ALICE WAS BEGINNING TO GET VERY TIRED OF SITTING BY HER SISTER ON THE BANK, AND OF HAVING NOTHING TO DO: ONCE OR TWICE SHE HAD PEEPED INTO THE BOOK HER SISTER WAS READING, BUT IT HAD NO PICTURES OR CONVERSATIONS IN IT, 'AND WHAT IS THE USE OF A BOOK,' THOUGHT ALICE 'WITHOUT PICTURES OR CONVERSATION?'

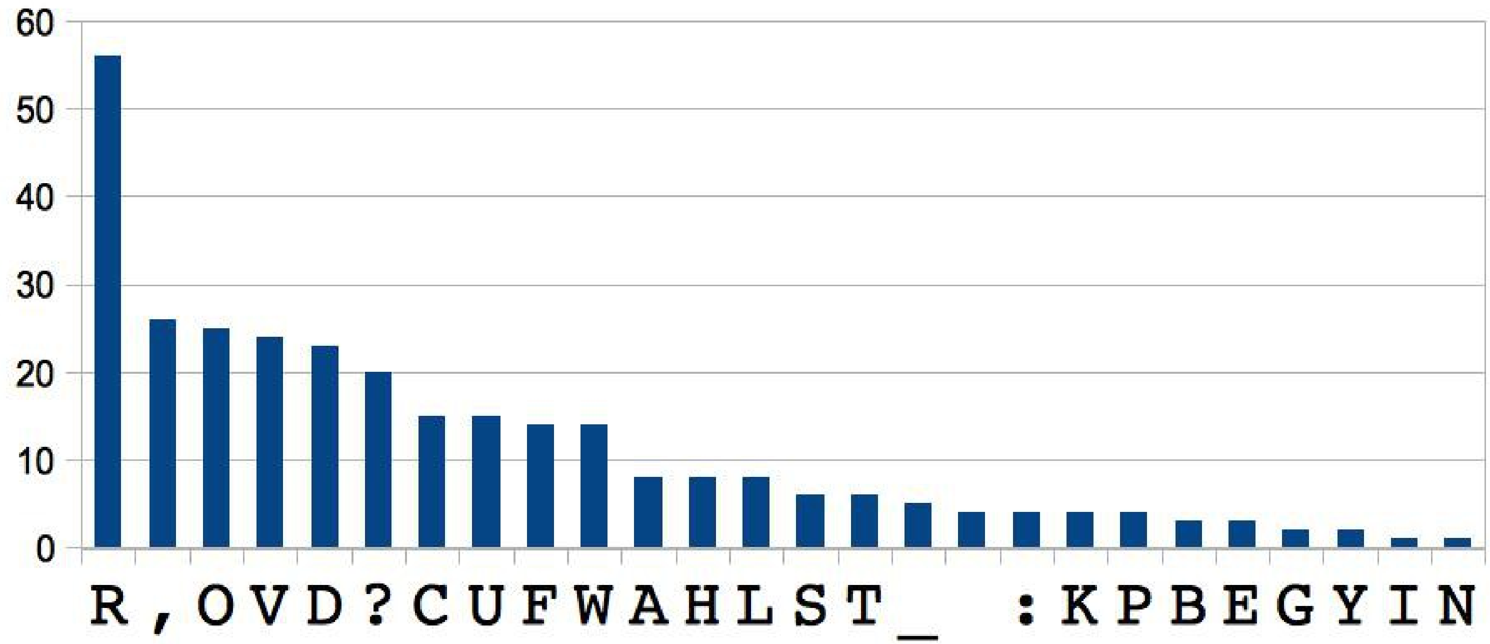
ALICE WAS BEGINNING TO GET VERY TIRED OF SITTING BY HER SISTER ON THE BANK, AND OF HAVING NOTHING TO DO: ONCE OR TWICE SHE HAD PEEPED INTO THE BOOK HER SISTER WAS READING, BUT IT HAD NO PICTURES OR CONVERSATIONS IN IT, 'AND WHAT IS THE USE OF A BOOK,' THOUGHT ALICE 'WITHOUT PICTURES OR CONVERSATION?'

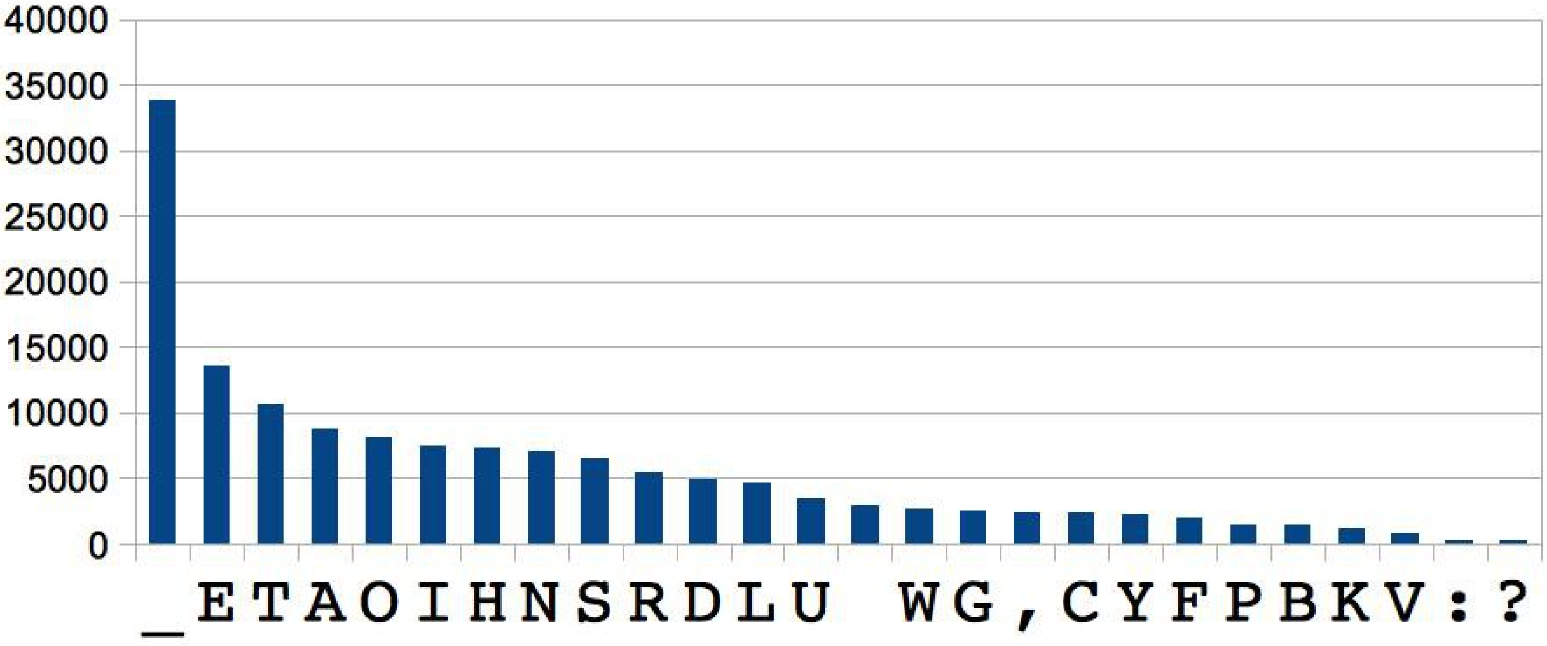
Plaintext letter:	A	B	C	D	E	F	G	H	…
Cryptocharacter:	C	T	H	A	O	E	L	R	…

```
ALICE WAS BEGINNING TO GET VERY TIRED OF
SITTING BY HER SISTER ON THE BANK, AND OF
HAVING NOTHING TO DO: ONCE OR TWICE SHE
HAD PEEPED INTO THE BOOK HER SISTER WAS
READING, BUT IT HAD NO PICTURES OR
CONVERSATIONS IN IT, 'AND WHAT IS THE USE
OF A BOOK,' THOUGHT ALICE 'WITHOUT
PICTURES OR CONVERSATION?'
```

▼

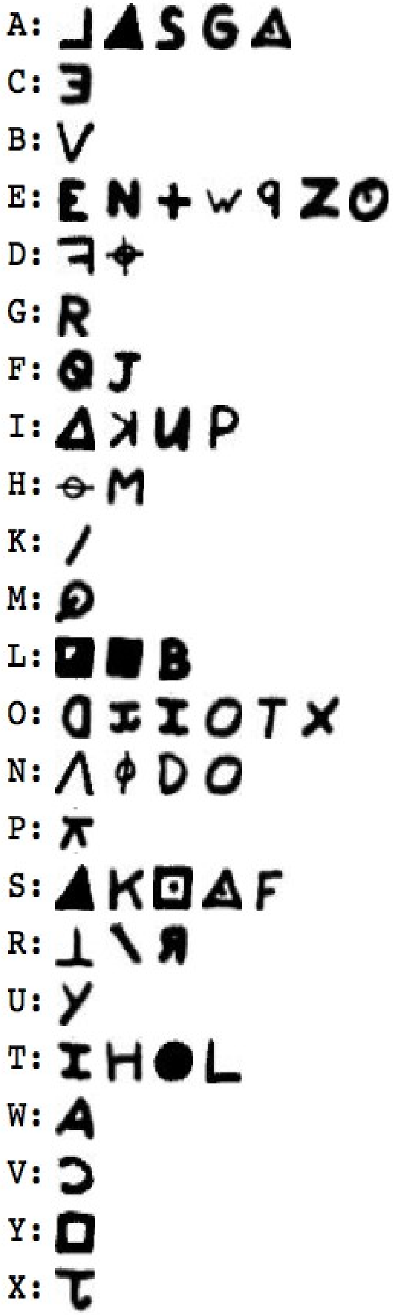
```
CYDHOR CURTOLD??D?LR,VRLO,R:OFGR,DFOARVE
RUD,,D?LRTGRWOFRUDU,OFRV?R,WORTC?B'RC?AR
VERWC:D?LR?V,WD?LR,VRAVNRV?HORVFR, DHORU
WORWCARKOOKOARD?,VR,WORTVVBRWOFRUDU,OFR
CURFOCAD?L'RTS,RD,RWCAR?VRKDH,SFOURVFRH
V?:OFUC,DV?URD?RD,'RPC?AR WC,RDUR,WORSUOR
VERCRTVVB'PR,WVSLW,RCYDHORP D,WVS,RKDH,S
FOURVFRHV?:OFUC,DV?IP
```


60
50
40
30
20
10
0
R,OVD?CUFWAHLST_ :KPBEGYIN

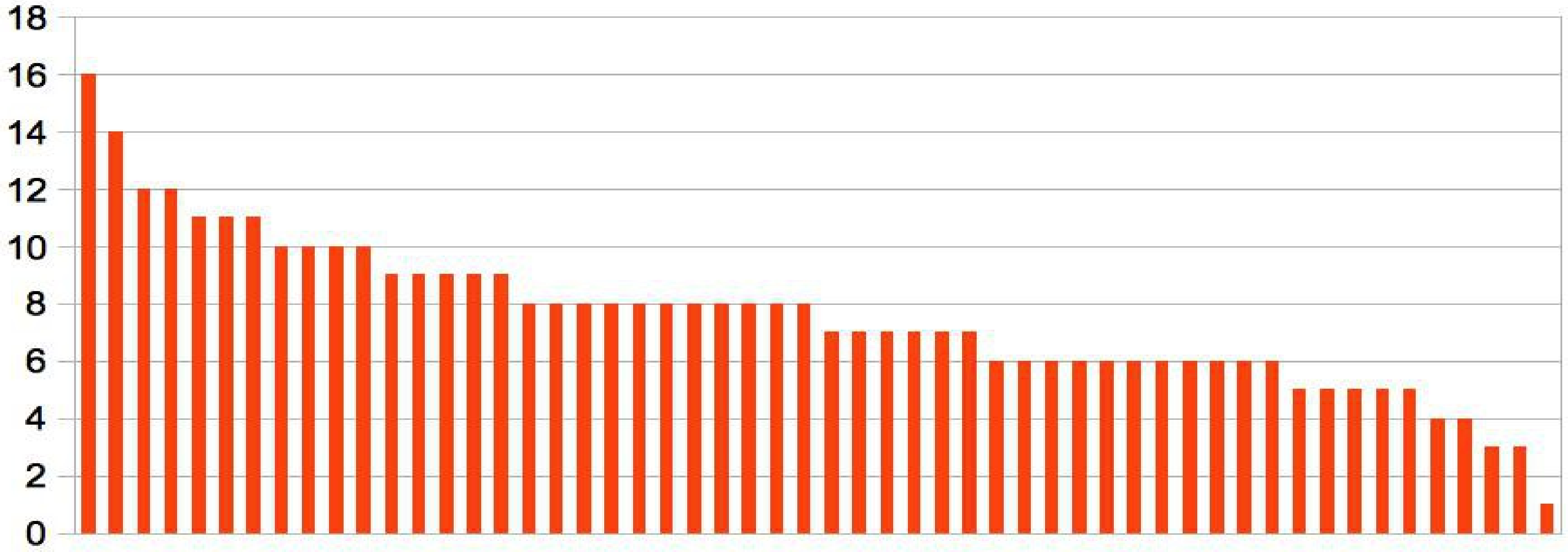

40000
35000
30000
25000
20000
15000
10000
5000
0
_ETAOIHNSRDLU WG,CYFPBKV:?

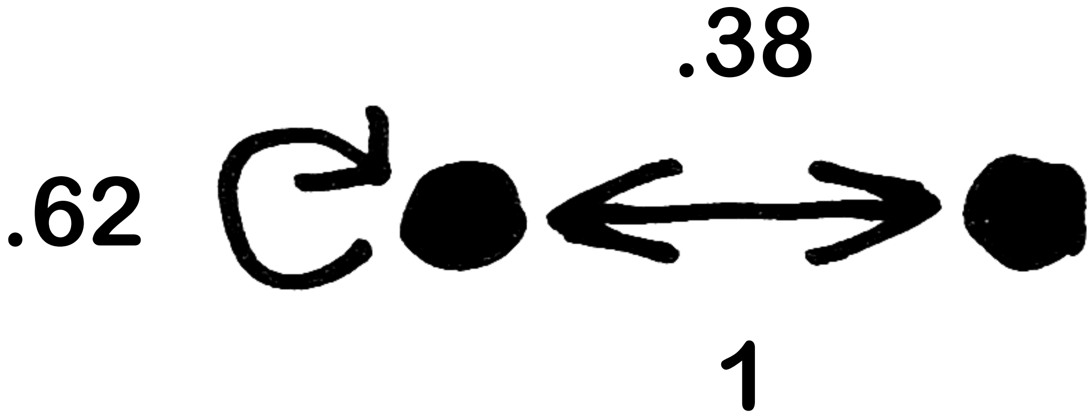
Homophonic cipher
(one-to-many cipher)

```
for each letter in your file:
   if no codeword exists:
      choose one
   else:
      either come up with a new
      one, or reuse an old one,
      depending on how often
      you have already used the
      other cipher symbols.
```

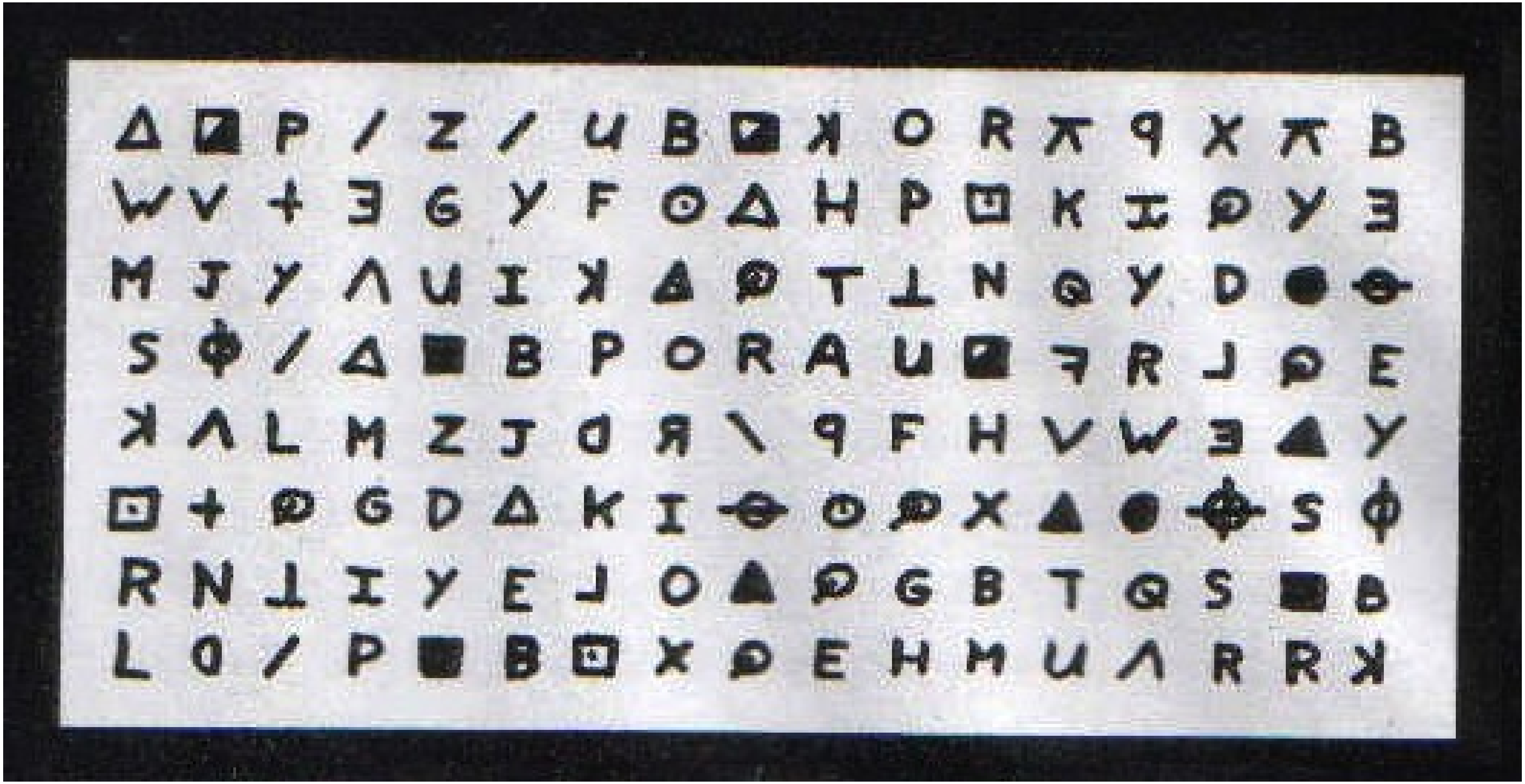


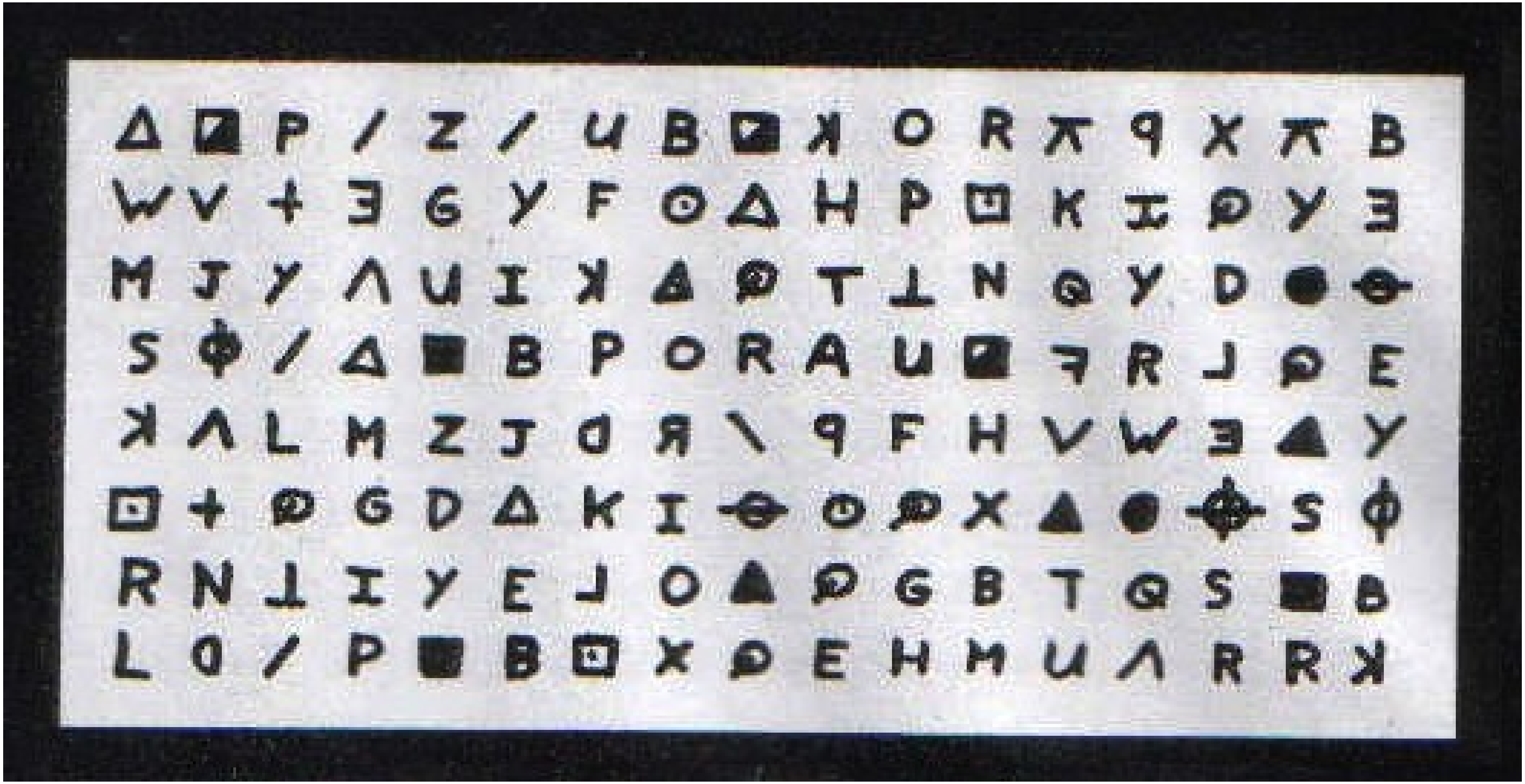
A:
C:
B:
E:
D:
G:
F:
I:
H:
K:
M:
L:
O:
N:
P:
S:
R:
U:
T:
W:
V:
Y:
X:


18
16
14
12
10
8
6
4
2
0

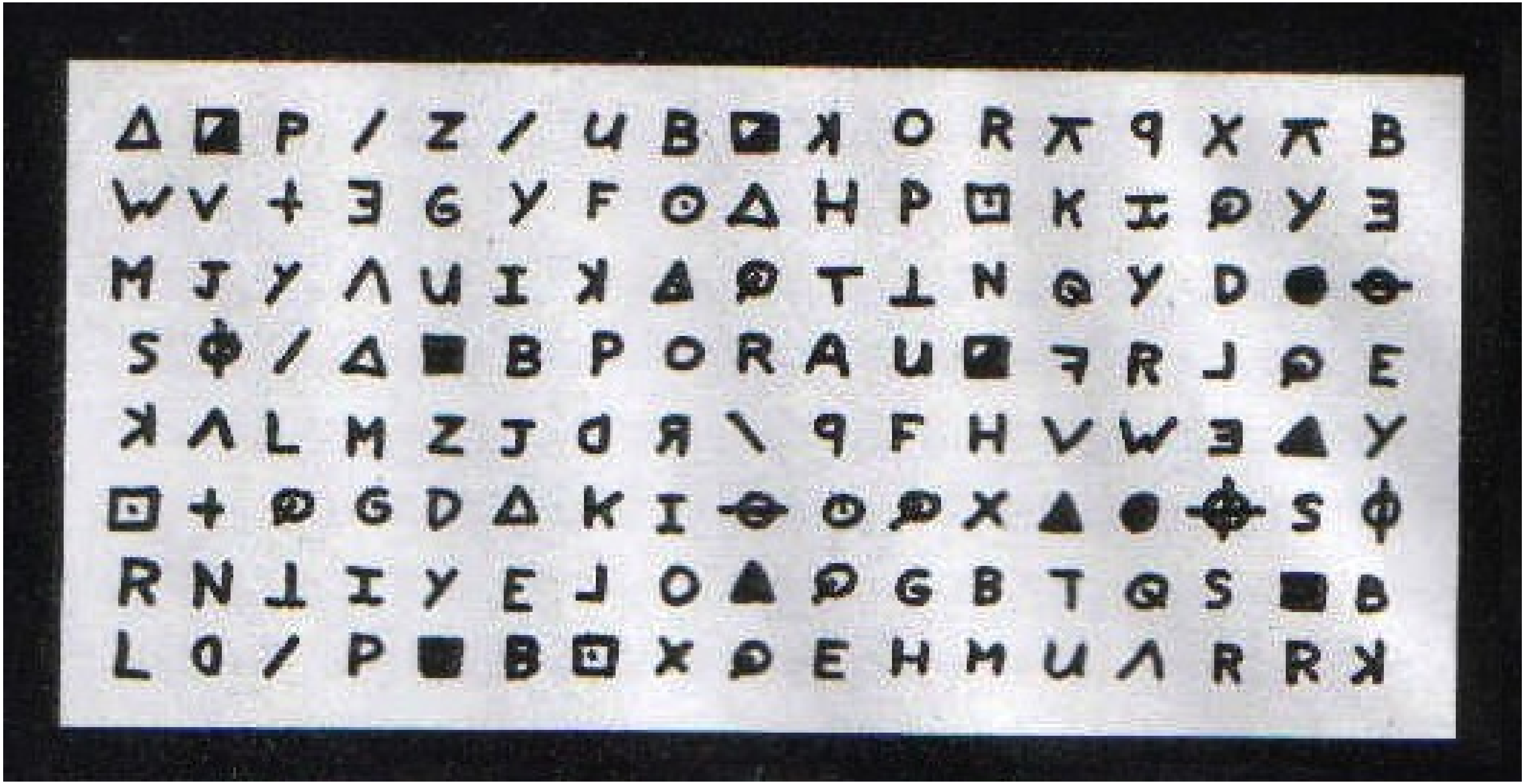


$$\begin{pmatrix} .62 & 1 \\ .38 & 0 \end{pmatrix}$$





ILIKEKILLINGPEOPLEB
ECAUSEITISSOMUCHFUN
ITISMOREFUNTHANKILL
INGWILDGAMEINTHEFO…



Sujith Ravi and Kevin Knight: "Bayesian Inference for Zodiac and Other Homophonic Ciphers," Proceedings of the conference of the Association for Computational Linguistics, 2011.

(aclweb.org/anthology/P/P11/P11-1025.pdf)

Encryption with relatively small sets of possible encryption schemes

$$D(P \,\|\, T) - D(P \,\|\, U) \;==\; \mathrm{E}_P\left(\log \frac{U(x)}{T(x)}\right)$$

Can we use our sophisticated statistical knowledge of English to crack the cipher, or did the encoder capture all statistical structure there is to describe?