Declassified Cold War code-breaking manual has lessons for solving ‘impossible’ puzzles

Dr Richard Bean

School of Information Technology and Electrical Engineering

10 March 2022

MATH3302 lecture

The University of Queensland

Jim Reeds: two kinds of code breaking (2000)

Real cryptanalysis

Actually trying to break *real* code messages sent by *real* people

Practice cryptanalysis

- Assessing algorithm strength
- Developing new analysis methods
- Doing homework
- Solving challenge puzzles (e.g. ***"Mystery Twister C3", ACA "American Cryptogram Association"***)

"Although the latter is every bit as intellectually challenging as the former, I think the former is by far the sexier.

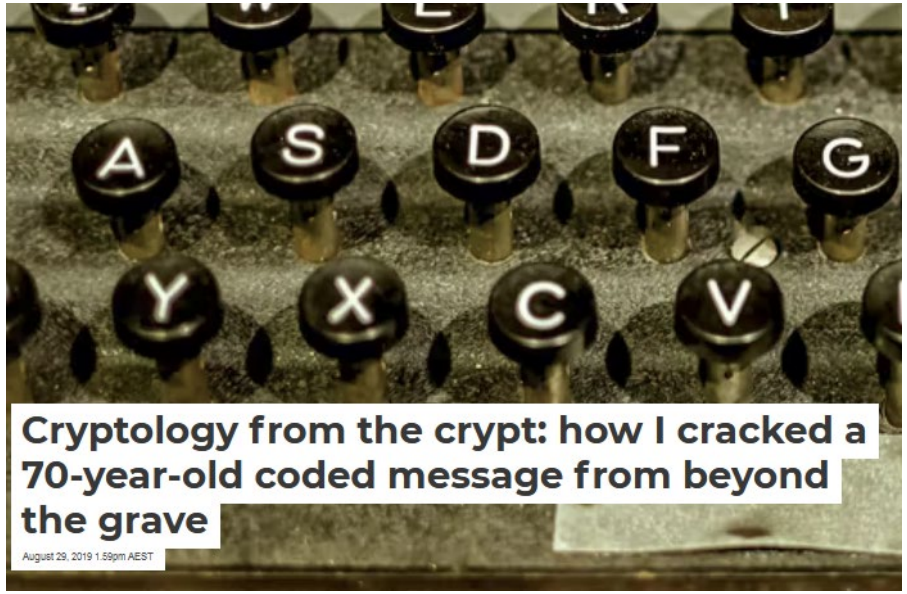
Alas, it is usually practiced by intelligence agencies, criminals, industrial espionage agents, detectives, etc, and is thus rarely described in the open literature until long after the fact.

I have been lucky in having done some "real" cryptanalysis that I can talk about, and lucky again because the examples are not too technical to explain here today."

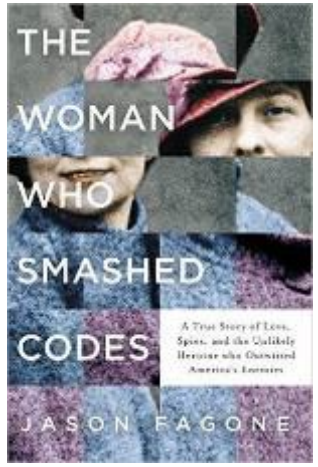
PERSONAL BACKGROUND

- Graduate of UQ (B.Sc. (Hons) 1997, Ph.D. 2001 in mathematics)
- Since then, worked in Iran and Australia in bioinformatics, combinatorics, energy, health in academia, government and private sector
- Recently had the opportunity to solve famous unsolved ciphers
  - Cipher from Irish Republican Army of 1920s and R. H. Thouless "Test for Survival" cipher of 1948 (*Histocrypt 2020*)
  - Biafran war ciphers of 1969 (presenting at NSA *Cryptologic History Symposium in May 2022*)
  - Arthur Dee's alchemical cipher c1635 (presenting at *Histocrypt in June 2022*)
- Lots of articles on *The Conversation* website and media appearances
- Empirical confirmation of Jim Reeds' thesis

The Conversation



US Government / National Security Agency



Reading the adversaries' mail / telegrams / radio signals

Black Chamber / Cipher Bureau of **Herbert Yardley** (1917-1929)
Signal Intelligence Service / Army Security Agency / Armed Forces Security Agency (1930-1949) **William F. Friedman**
NSA (National Security Agency) – signals intelligence organization of the US Government
Founded in 1952 after World War II

Intercepting and processing signals
Making and breaking codes

Published a book series "Military Cryptanalysis" and "Military Cryptanalytics" (1930s-1970s)
How to break codes in the "pre-modern-computer" era
Over time, emphasis at NSA changed from breaking ciphers to backdoors / plaintext interception

MILITARY CRYPTANALYSIS (FRIEDMAN, LATE 1930S)

~~Restricted~~

WAR DEPARTMENT
OFFICE OF THE CHIEF SIGNAL OFFICER
WASHINGTON

MILITARY CRYPTANALYSIS
Part I
MONOALPHABETIC SUBSTITUTION SYSTEMS

By
WILLIAM F. FRIEDMAN
Principal Cryptanalyst
Chief of Signal Intelligence Section
War Plans and Training Division

PREPARED UNDER THE DIRECTION OF THE CHIEF SIGNAL OFFICER

UNITED STATES
GOVERNMENT PRINTING OFFICE
WASHINGTON: 1938

~~CONFIDENTIAL~~

~~RESTRICTED~~

MILITARY CRYPTANALYSIS

PART II

SIMPLER VARIETIES OF POLYALPHABETIC SUBSTITUTION SYSTEMS

by

WILLIAM F. FRIEDMAN

Principal Cryptanalyst

Prepared under the direction of the Chief Signal Officer.

Record Taken from WFF's Home

1937

~~CONFIDENTIAL~~

~~CONFIDENTIAL~~

~~Restricted~~

WAR DEPARTMENT
OFFICE OF THE CHIEF SIGNAL OFFICER
WASHINGTON

MILITARY CRYPTANALYSIS
Part III
SIMPLER VARIETIES OF APERIODIC SUBSTITUTION SYSTEMS

By
WILLIAM F. FRIEDMAN
Principal Cryptanalyst
Signal Intelligence Service

PREPARED UNDER THE DIRECTION OF THE CHIEF SIGNAL OFFICER

~~RESTRICTED~~

~~Notice: This document contains information affecting the national defense of the United States within the meaning of the Espionage Act (U.S.C. 50: 31, 32). The transmission of this document or the revelation of its contents in any manner to any unauthorized person is prohibited.~~

UNITED STATES
GOVERNMENT PRINTING OFFICE
WASHINGTON: 1939

~~CONFIDENTIAL~~

~~Restricted~~

WAR DEPARTMENT
OFFICE OF THE CHIEF SIGNAL OFFICER
WASHINGTON

MILITARY CRYPTANALYSIS
Part IV
TRANSPOSITION AND FRACTIONATING SYSTEMS

By
WILLIAM F. FRIEDMAN
Principal Cryptanalyst
Signal Intelligence Service

PREPARED UNDER THE DIRECTION OF THE CHIEF SIGNAL OFFICER

UNITED STATES
GOVERNMENT PRINTING OFFICE
WASHINGTON 1941

MILITARY CRYPTANALYTICS SERIES (WFF AND LDC 1952-77)

~~CONFIDENTIAL~~

Copy No. ___

NATIONAL SECURITY AGENCY

MILITARY CRYPTANALYTICS
Part I

By
WILLIAM F. FRIEDMAN
and
LAMBROS D. CALLIMAHOS

NOTICE: This material contains information affecting the national defense of the United States within the meaning of the espionage laws, Title 18, U.S.C., Secs. 793 and 794, the transmission or revelation of which in any manner to an unauthorized person is prohibited by law.

National Security Agency
Washington 25, D. C.

December 1952

~~CONFIDENTIAL~~
Modified Handling Authorized

NATIONAL SECURITY AGENCY

MILITARY CRYPTANALYTICS
Part II

INTERIM EDITION
(Third Section)

By
LAMBROS D. CALLIMAHOS
and
WILLIAM F. FRIEDMAN

NOTICE: This material contains information affecting the national defense of the United States within the meaning of the espionage laws, Title 18, U.S.C., Secs. 793 and 794, the transmission or revelation of which in any manner to an unauthorized person is prohibited by law.

Office of Training Services
National Security Agency
Fort George G. Meade, Maryland

February 1958

~~CONFIDENTIAL~~
Modified Handling Authorized

~~SECRET~~

NATIONAL SECURITY AGENCY

MILITARY CRYPTANALYTICS
Part III

By
LAMBROS D. CALLIMAHOS

October 1977

Classified by DIRNSA/CHCSS (NSA/CSSM 123-2)
Exempt from GDS, EO 11652, Cat 2
Declassify Upon Notification by the Originator

National Security Agency
Fort George G. Meade, Maryland

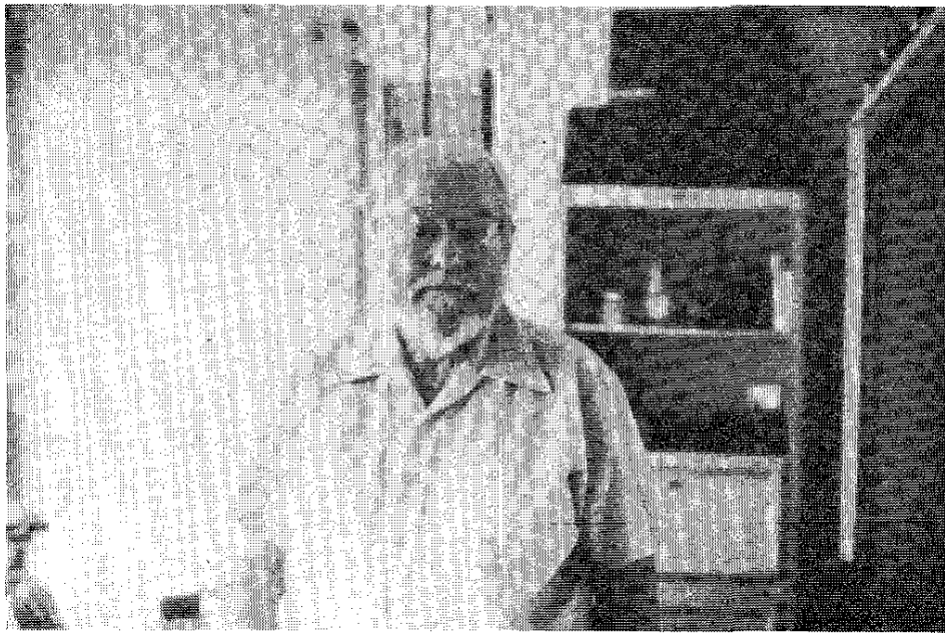
~~SECRET~~

CRYPTANALYSIS

Lambros D. Callimahos, cryptanalyst of
National Security Agency (1910-1977)

Frank B. Rowlett (1908-1998)
World War II cryptographer

"The day is past when one person, working alone, can break a sophisticated cipher system. The state of the cryptographic art is now so advanced that a well-designed system can be expected to resist all the efforts of a well-organized team supported by computers to solve it." – April 1979



Cryptography Timeline 1970s

- 1970 – CIA and *Bundesnachrichtendienst* purchase Swiss "Crypto AG" company
- 1972-1977 – *"Data Encryption Standard"* (DES)
- 1976 – Diffie-Hellmann Key Exchange (1974 GCHQ)
- 1976 – *"New Directions in Cryptography"* by Diffie and Hellman
- 1977 – Callimahos last book published
- 1977 – RSA public-key encryption developed (1973 GCHQ)
- 1979 – "cryptanalytic breakthrough against Soviet ciphers which involved a fundamental enough mathematical result that NSA still refused to declassify any details [in 2016] – *a last hurrah for the golden age of codebreaking*"
- 1979 – Rowlett says well-designed systems can't be broken

MILITARY CRYPTANALYTICS PART III

Colonel Friedman's still-classified work, volumes three and four of the first book and volume three of the second, was requested in June under the Freedom of Information Act by a California cryptographer, ***John Gilmore****.*

*He believes that widespread access to coding and code-breaking technologies will make it easier to protect personal privacy in the electronic information age. His logic is that the more that is known about code-breaking, the easier it will be for individuals to design computer codes that would be almost impossible to break. (*New York Times 1992)

- Released on “Government Attic” in January 2021

"The Net treats censorship as damage and routes around it."

Gaines (1939) and Sinkov (1970)

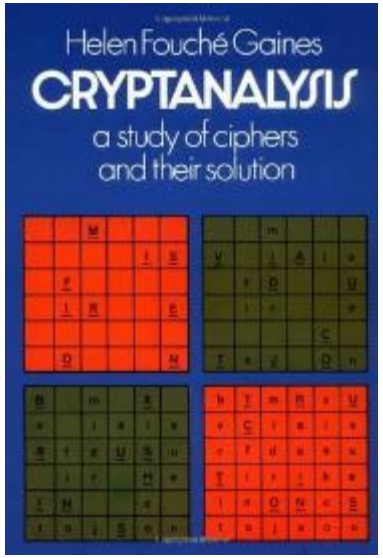
These books tell you about solving the classical cipher types (many are used in "ACA" puzzles) but not about how to identify a cipher type from a given ciphertext.

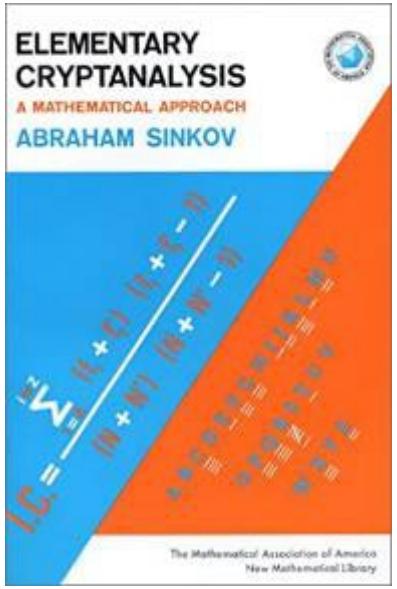
Gaines doesn't even mention **index of coincidence** (1939).

*"Great **evenness in frequencies** may suggest one of the key-lengthening devices, such as autokey and progressing key; and the practical absence of repeated sequences will usually mean that a transposition has been added to a substitution."*

Book focus is on "puzzle" cryptograms so these might be good diagnostic starting points.

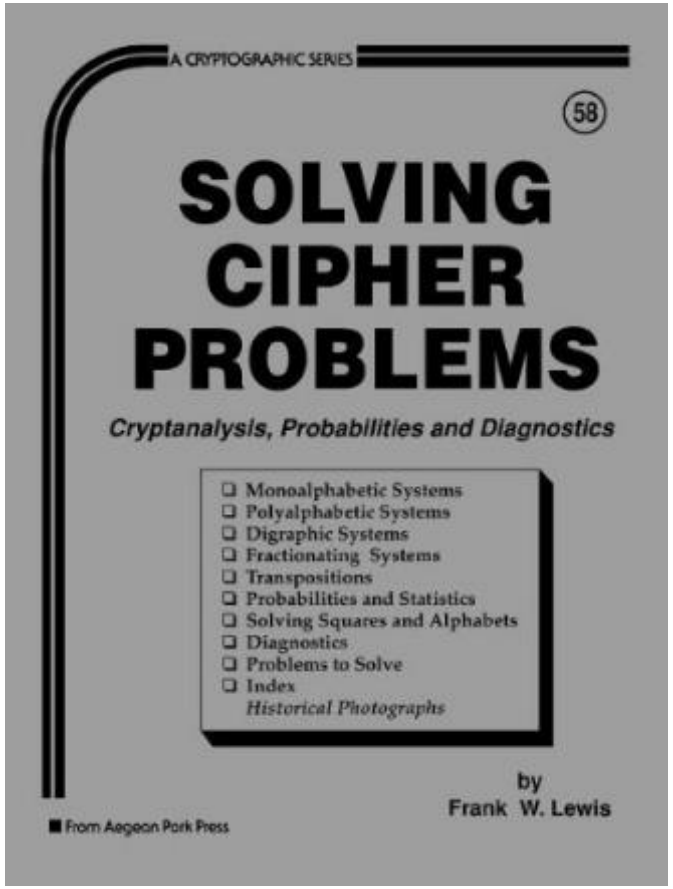
Sinkov (1970) covers the *Hill cipher* (i.e. using matrix multiplication) which depends on concepts of linear algebra.





Describing Cipher type Recognition informally

Frank W. Lewis, "Solving Cipher Problems" (1992)



The chapter on diagnostics demonstrates how a professional cryppie approaches a cryptanalytic problem.

Lewis states that the task of an analyst is finding, measuring, explaining, and exploiting a phenomenon (or phenomena). Writing about cipher type diagnosis, he describes the search for "something funny" or "finding the phenomena".

WHAT IS RANDOMNESS?

A random sequence of letters or digits …

(off the top of my head … time to ask the live audience!!)

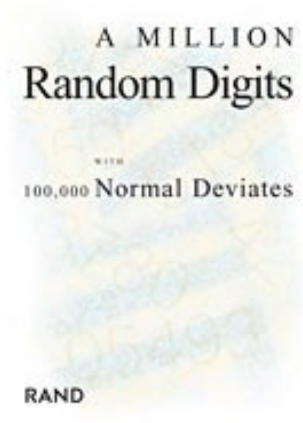
- At each position, each letter or digit must be equally likely (**"uniform distribution"**) – as opposed to following something like Benford's Law for leading digits / power-law distribution. *Chi-squared test*
  - *If this is true then the same is true for groups of digits / letters*
- At each position, each contiguous group of n letters / digits must be equally likely for all n > 1
- For each width m, if we look at digits / letters at position x and x+m, these must all be equally likely, and the generalization for groups of letters / digits
- Digits or letters in the sequence must have no correlation / connection to future digits / letters
- Runs test - runs of letters / digits must not occur at a "significantly" higher rate than we would expect; e.g. 999999 or 3456789
- Delta tests (related). The differences of adjacent letters / digits in the sequences, and differences of differences, and differences of letters / digits at a constant width, follow the same uniform distribution.
- "Diehard tests" of Marsaglia

ARE THE DECIMAL DIGITS OF PI RANDOM?

3.14159265358979323846264338327950288419716939937510
58209749445923078164062862089986280348253421170679
82148086513282306647093844609550582231725359408128
48111745028410270193852110555964462294895493038196
44288109756659334461284756482337867831652712019091
45648566923460348610454326648213393607260249141273
72458700660631558817488152092096282925409171536436
78925903600113305305488204665213841469519415116094
33057270365759591953092186117381932611793105118548
07446237996274956735188575272489122793818301194912
98336733624406566430860213949463952247371907021798
60943702770539217176293176752384674818467669405132
00056812714526356082778577134275778960917363717872
14684409012249534301465495853710507922796892589235
42019956112129021960864034418159813629774771309960
51870721134999999837297804995105973173281609631859
50244594553469083026425223082533446850352619311881
71010003137838752886587533208381420617177669147303
59825349042875546873115956286388235378759375195778
18577805321712268066130019278766119590921642019 89

Are numbers claimed to be random actually random?

- https://archive.is/8XrNy (WSJ 2020)
- RAND Corporation's book "A Million Random Digits" (1955)
- Gary Briggs of RAND (2020)



Gary Briggs with 'A Million Random Digits.' DIANE BALDWIN/RAND CORPORATION

A-HED

'A Million Random Digits' Was a Number-Cruncher's Bible. Now One Has Exposed Flaws in the Disorder.

A 1955 Rand Corp. book had a reputation as the go-to source for figures used by pollsters, analysts, researchers; engineer Gary Briggs has ruined it

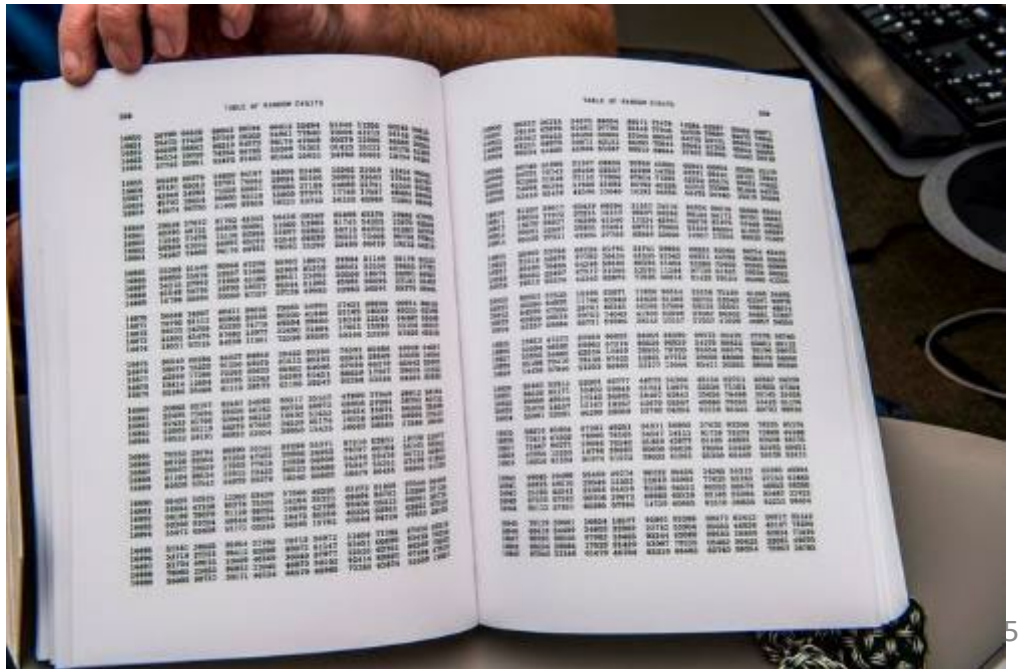
By Michael M. Phillips
Updated Sept. 24, 2020 12:43 pm ET

MOST POPULAR VIDEOS

Elated but cautious, Mr. Briggs examined sequences of repeated numbers as a final test.

In a group of 50,000 random digits, mathematicians would expect 4,050 sequences of two identical digits in a row—77, for instance. They would predict 405 spots with three identical digits in a row, such as 555. There would be about 40 cases of four identical digits in a row. And four or five places with five identical digits together.

His results were "soul crushing," Mr. Briggs says. The book contains 48 runs of four digits instead of 40, an astoundingly wide divergence in statistical terms that eluded any explanation he could conjure.



Are the Lottery Numbers random? (NY Times)

Here's how the Multi-State Lottery Association's random-number generators were supposed to work: The computer takes a reading from a Geiger counter that measures radiation in the surrounding air, specifically the radioactive isotope Americium-241. The reading is expressed as a long number of code; that number gives the generator its true randomness. The random number is called the seed, and the seed is plugged into the algorithm, a pseudorandom number generator called the Mersenne Twister. At the end, the computer spits out the winning lottery numbers.

Tipton's extra lines of code first checked to see if the coming lottery drawing fulfilled Tipton's narrow circumstances. It had to be on a Wednesday or a Saturday evening, and one of three dates in a nonleap year: the 147th day of the year (May 27), the 327th day (Nov. 23) or the 363rd day (Dec. 29). Investigators noticed those dates generally fell around holidays — Memorial Day, Thanksgiving and Christmas — when Tipton was often on vacation. If those criteria were satisfied, the random-number generator was diverted to a different track. Instead, the algorithm would use a predetermined seed number that restricted the pool of potential winning numbers to a much smaller, predictable set of numbers.

Is the roulette Wheel at the Casino Random?

Ms. Jarecki said in a telephone interview on Monday that she, Dr. Jarecki and a handful of other people helping them would record the results of every turn of a given roulette wheel to discover its biases, or tendency to land on some numbers more frequently than others, usually because of a minute mechanical defect caused by shoddy manufacturing or wear and tear.

“There was nothing original in his method — only in the successful way he employed his fanciful computer explanation to delude the managements of European casinos, the gaming police and the general public,” the gaming writer Russell T. Barnhart said in a chapter about Dr. Jarecki in his book “Beating the Wheel: Winning Strategies at Roulette” (1992).

Ms. Jarecki said that watching, or “clocking,” a wheel, as Mr. Barnhart described it, could mean observing more than 10,000 spins over as long as a month. Sometimes a wheel would yield no observable advantage. But when Dr. Jarecki and company did find a wheel with a discernible bias, he would have an edge over the house.

CASINOS CHANGE THE RULES (21ST CENTURY)

BRENDAN I. KOERNER BUSINESS AUG 5, 2017 10:00 AM

Meet Alex, the Russian Casino Hacker Who Makes Millions Targeting Slot Machines

A mathematician-turned-criminal unleashes his agents on casinos around the world. But there's money in the extortion racket, too.

gsl_rng_knuthran2

$X_n = 271828183\ X_{n-1} - 314159269\ X_{n-2}$ modulo $(2^{31} - 1)$

[A CS professor David Ackley] noticed that [constants in the Aristocrat PRNG algorithm] were familiar: One was an approximation of pi (31415926), one was an abbreviation of the mathematical constant e (271828), and one was a slightly ribald jest (69069).

… Ackley found that those exact numbers had also been used in a PRNG featured in SpaceOut, a 1988 program for the X Window System that simulated travel through a star field. When I contacted the author of SpaceOut, he recalled that he had cribbed his PRNG from the second volume of Donald Knuth's *The Art of Computer Programming*, a classic of the discipline.

What if you see a pattern where there is none?

Follow the Science... to God?

414 views • 23 Feb 2021

15 DISLIKE SHARE CLIP SAVE ...

Aishdotcom
51.9K subscribers

SUBSCRIBE

Harold Gans is a retired Senior Cryptologic Mathematician with the U.S. Department of Defense - an expert on codes. He has applied his knowledge to various esoteric properties of the Universe and to the complexity of biological structures and has concluded that the existence of a Creator is a mathematical certainty.

For more Jewish inspiration visit https://www.aish.com – the world's leading Judaism website.

Harold Gans, mathematician worked at NSA 1968-1996.

In late 1980s, read about "Torah codes" experiment finding the names of Rabbis *"near"* their dates of birth and death in Bible using "equidistant letter sequences".

(Just like the randomness test we just mentioned!)

Experiment comprehensively debunked in 1999 paper.

Why is randomness important in cryptography?

In an ideal encryption algorithm, with a plaintext P and a ciphertext C …

if we change one bit of P, 50% of the bits of C should be different

If you change any part of the input, every part of the output is affected – **Diffusion**

Each part of the output depends on several parts of the key – **Confusion**

Cryptanalysis of classical (pre-computer) ciphers depends on exploiting patterns in the plaintext that have continued through to the ciphertext. “Regularities” in the plaintext or key should not lead to “regularities” in the ciphertext.

Helps the *cryptographers* to build strong codes

- The “One Time Pad” method of encryption – a completely random key
  - All plaintexts and keys are equally likely

Helps the intelligence agencies insert backdoors

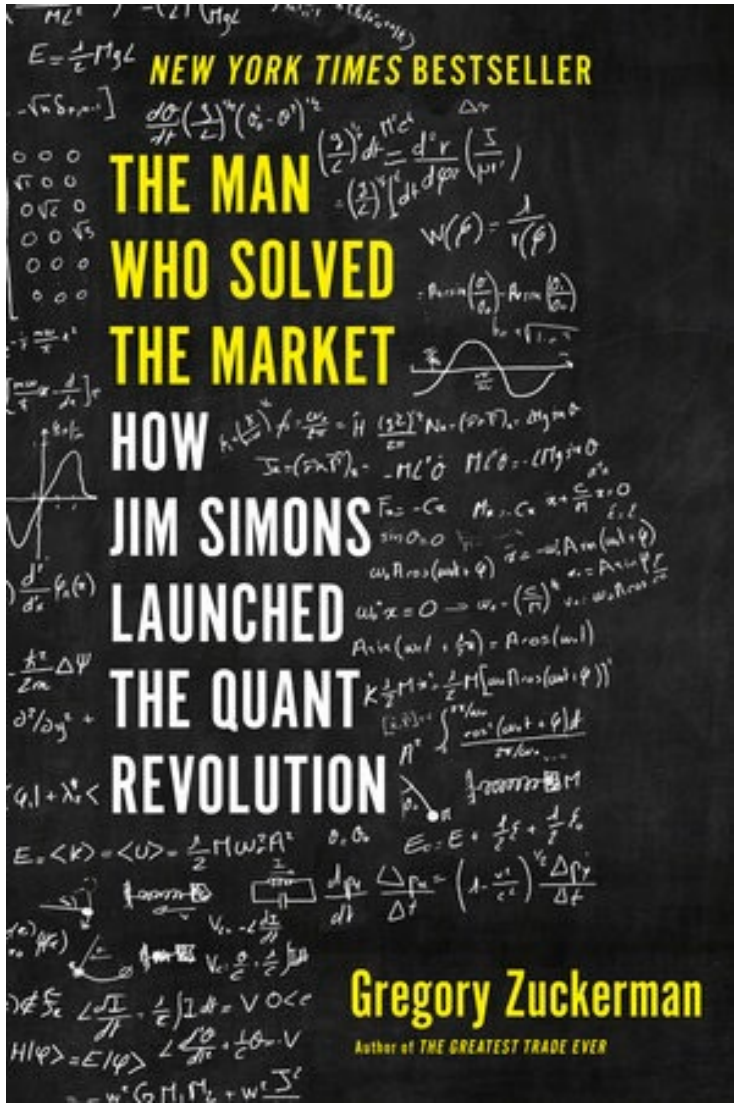
- Dual EC DRBG (Dual Elliptic Curve Deterministic Random Bit Generator)
- Backdoors in Crypto AG cipher machines
- Backdoor in PX-1000CR (pocket telex of 1980s) – DES removed, NSA algorithm replaced

"MATHEMATICIAN: AN INSIDER'S VIEW" (2003)

Diagnosis is the study of cipher, enciphering key, or cryptovariables (initial key settings) in an attempt to determine the cryptographic algorithms from which they were generated. I work on the Transnational (Target) Integrated Diagnosis Focus Team. The team is made up of twelve mathematicians/cryptanalysts. Our task is to perform diagnosis on indigenous cryptographic systems and simultaneously improve the health of cryptanalytic diagnosis.

During the course of a normal day I run cryptanalytic routines on UNIX desktop workstations, supercomputers, and special-purpose devices using available software tools. The routines employ standard cryptanalytic tests which search for patterns and non-random properties in data. If I devise a test for which no available tools exist, then I will write software to perform the test. If I detect a significant statistical property in data, I will immediately seek, expect, and receive help from team members.

Cryptanalysis / Finance Gurus



Jim Simons, founder of Renaissance Technologies and the Medallion Fund
Net worth $US24bn

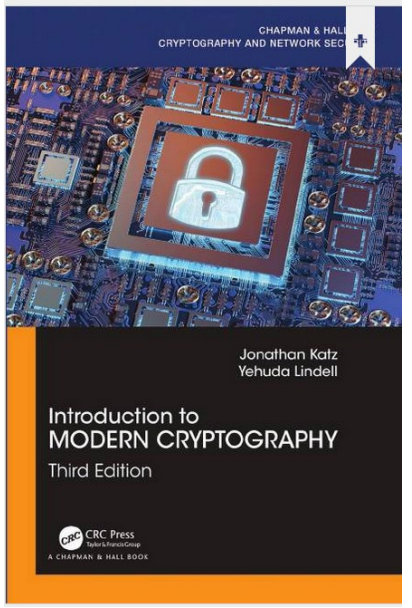
Leonard Baum, inventor of Baum-Welch Algorithm from IDA

Nick Patterson, GCHQ
to RenTech
to Broad Institute

Dr. Patterson said he also has a well-honed instinct about which data is important, after seeing "a lot of surprising stuff that turned out to be complete nonsense."
Dr. Lander of the Broad Institute describes him as a great skeptic, with the statistical insight to tell whether a signal is "simply random fluctuation or whether it's a smoking gun."

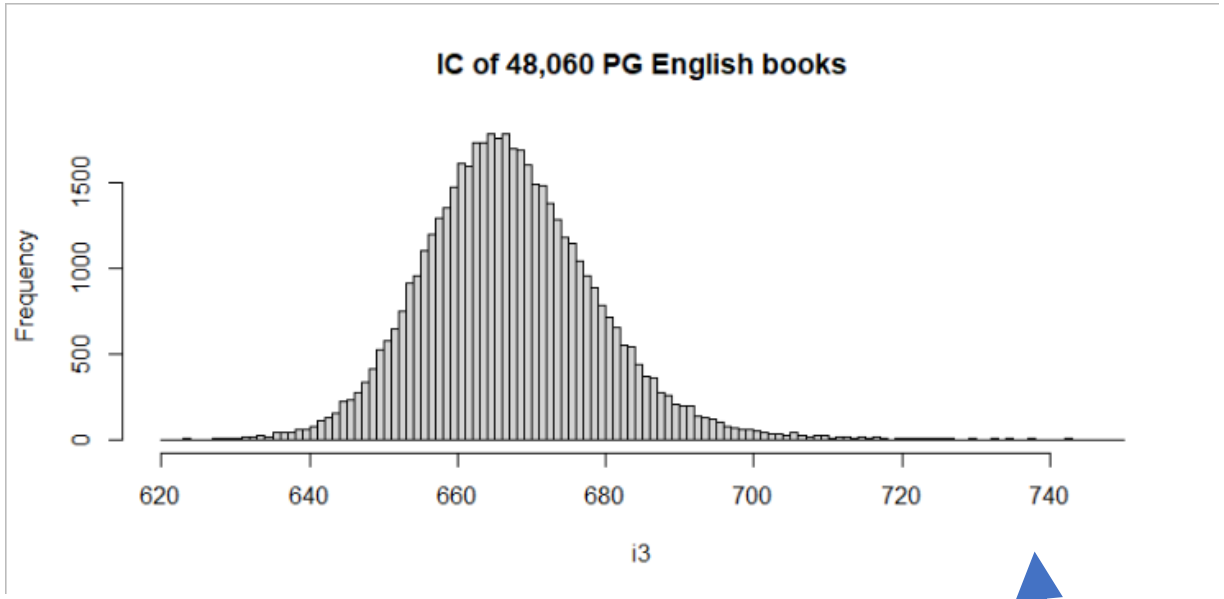
"PRINCIPLES OF CRYPTODIAGNOSIS"

- A chapter in Part III of "Military Cryptanalytics"
- Determining what method has been used to encipher a message given only the ciphertext
- Statistics of the plaintext. "Index of Coincidence" (IC) invented by William F. Friedman
- $\mathrm{IC} = \sum_{i=1}^{c} \frac{f_i(f_i-1)}{n(n-1)}$ or multiply by *c* to normalize value.
- For text with 26 possible letters:
  - Random text – the IC will be close to 1/26 ≈ 0.038462
  - English text – the IC will be close to 0.0667
  - The longer the text is – if IC deviates from 1/26 the less "random" it seems
  - Random 26 letters – s.d. = 0.010648 (1m iterations) – English +2.68 SD
  - 100 letters, s.d. = 0.002734 – English +10.4 SD
  - 300 letters, s.d. = 0.000908
  - 1000 letters, s.d. = 0.000271
- More complex techniques can distinguish subtle differences from "random" text or "random" bytes – particularly distinguishing cipher methods



INDEX OF COINCIDENCE

Language	# books	Mean	SD
english	46831	667	12
french	2972	795	15
finnish	2136	879	20
german	1689	787	25
dutch	796	806	27
italian	792	751	10
spanish	625	756	14
portuguese	496	784	12
swedish	201	693	16
hungarian	188	690	9
esperanto	117	685	18
latin	87	711	23
danish	67	791	18
tagalog	56	1149	58
catalan	32	765	16
polish	25	608	17
norwegian	21	757	25
czech	10	608	23
welsh	10	656	22
icelandic	7	706	14
friulian	6	733	20



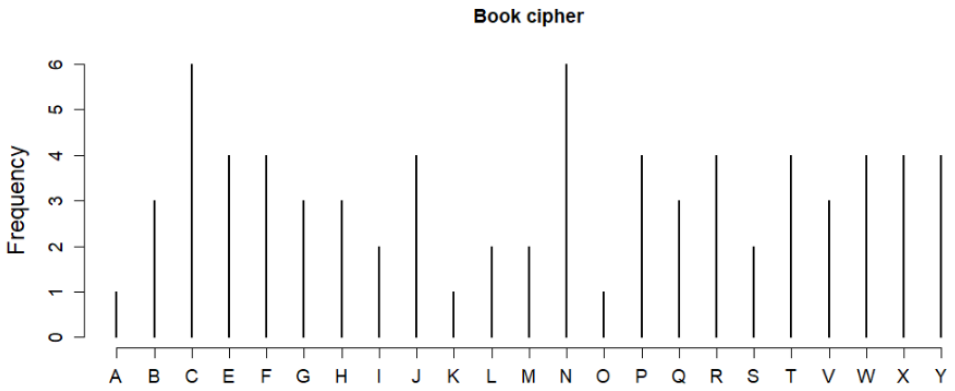
We can distinguish between English and French/German quite easily.

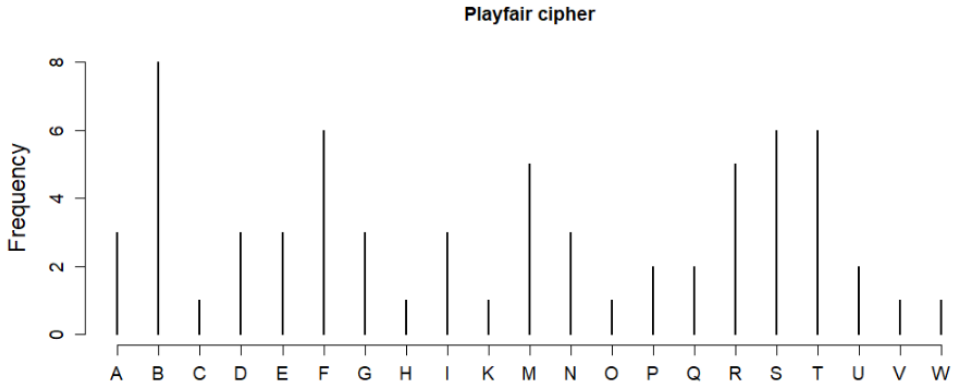
Older English books tend to have higher ICs like French.

Latin in between.

Index of coincidence – Thouless ciphers

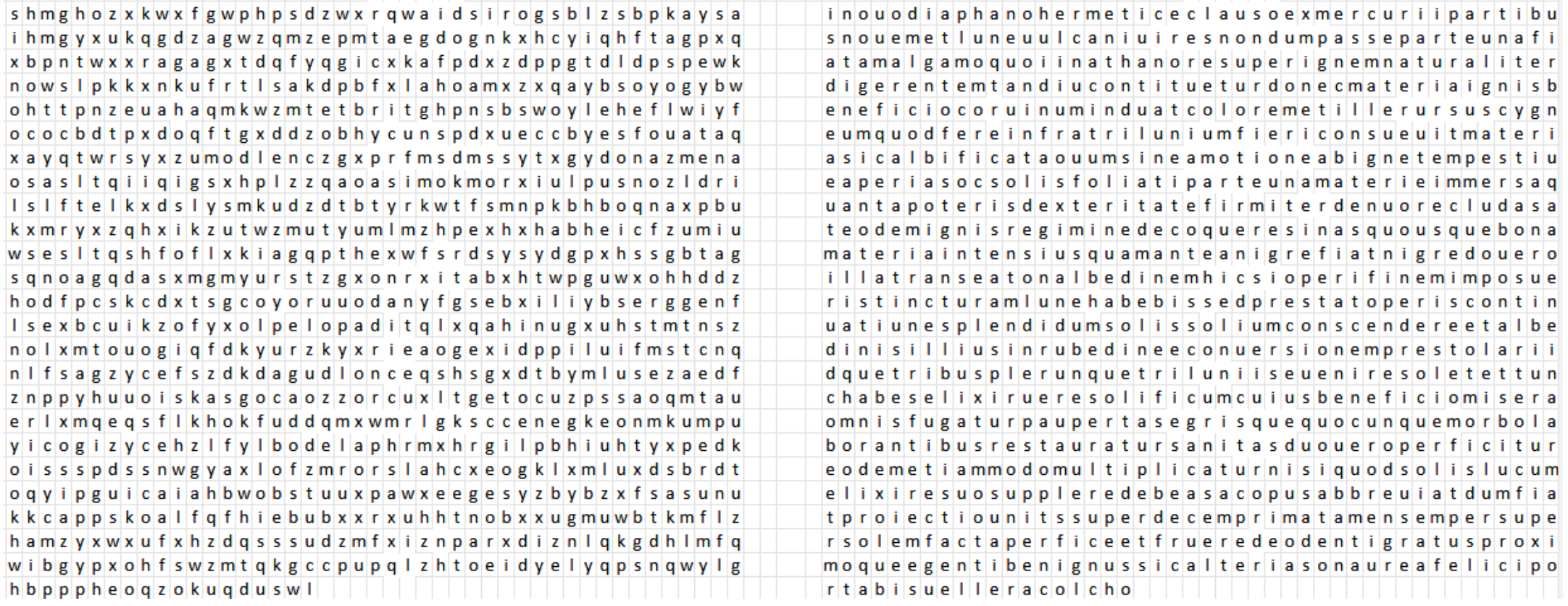
- Thouless single Playfair - IC 0.0531 – 66 letters. “Rough”
  - *CBFTM HGRIO TSTAU FSBDN WGNIS BRVEF BQTAB QRPEF BKSDG MNRPS RFBSU TTDMF EMA BIM*
- Thouless double Playfair – IC 0.0387 – 58 letters
  - *BTYRR OOFLH KCDXK FWPCZ KTADR GFHKA HTYXO ALZUP PYPVF AYMMF SDLR UVUB*
- Thouless Book Cipher – IC 0.03813 – 74 letters. “Flat”
  - *INXPH CJKGM JIRPR FBCVY WYWES NOECN SCVHE GYRJQ TEBJM TGXAT TWPNH CNYBC FNXPF LFXRV QWQL*





Index of coincidence (IRA, Dee ciphers)

- IRA cipher – IC 0.06274 – 52 letters. Incomplete columnar transposition.
  - *GTHOO RCSNM EOTDE TAEDI NRAHE EBFNS INSGD AILLA YTTSE AOITDE*
- Arthur Dee cipher – IC 0.04360 – 1099 letters; plaintext 0.07404 (Latin). CT probably polyalphabetic cipher. 45 letter key.



PATTERN RECOGNITION

Defence Science Journal, Vol. 62, No. 1, January 2012, pp. 38-45, DOI: 10.14429/dsj.62.1440

Classification Models for Symmetric Key Cryptosystem Identification

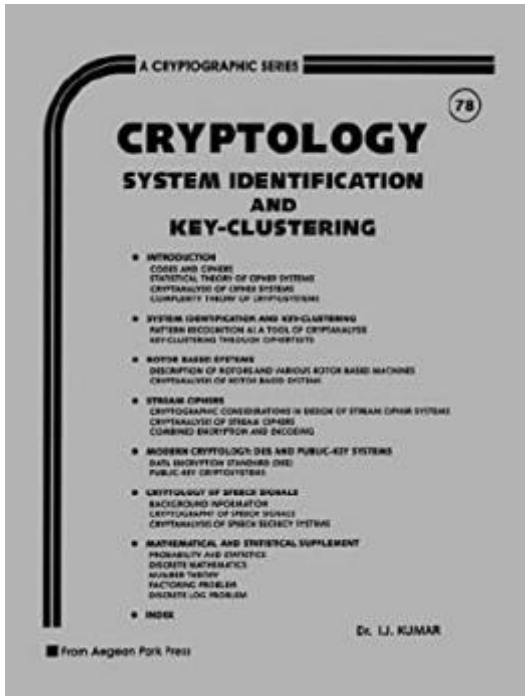
Shri Kant

Joint Cipher Bureau, Delhi – 110 054, India
E-mail: shrikant.ojha@gmail.com

ABSTRACT

The present paper deals with the basic principle and theory behind prevalent classification models and their judicious application for symmetric key cryptosystem identification. These techniques have been implemented and verified on varieties of known and simulated data sets. After establishing the techniques the problems of cryptosystem identification have been addressed.

Keywords: Classification, supervised pattern recognition, feature extraction, decision surfaces, discriminant functions



"Discrimination between Classical Systems"

Kumar, I.J. and Rao, T.L. (1983) *Parametric Discrimination of Three Systems*, Defence Research and Development Organisation (DRDO) report (unpublished)
from Kumar, I.J. (1997) *Cryptology: System Identification and Key-clustering*
Discriminant functions
(1) Book cipher system
(2) Rotor based system with complex rotation regime
(3) Non-linear shift register based system

(1) Vigenere Square
(2) Playfair
(3) Hill Cipher

or: Hebern / Enigma / Typex rotor-based systems
or: DES / non-linear shift register (Geffe generator)
or: Geffe generator / non-linear combiner function / non-linear feed-forward system with linear feedback

BION STATS (2005)

Type	IC	MIC	MKA	DIC	EDI	LR	ROD	LDI	SDD
Plaintext	63 ±5	73 ±11	95 ±19	72 ±18	73 ±24	22 ±5	50 ±6	756 ±13	303 ±23
Rand Digit	100 ±2	108 ±8	132 ±16	100 ±8	98 ±15	21 ±3	50 ±3	0 ±0	0 ±0
Rand Text	38 ±1	44 ±5	60 ±12	14 ±2	14 ±5	5 ±3	50 ±10	428 ±23	109 ±14
6x6 Bifid7	35 ±4	47 ±9	62 ±16	14 ±5	14 ±8	4 ±3	49 ±12	298 ±53	71 ±16
6x6Playfair	42 ±4	51 ±9	67 ±15	32 ±9	72 ±24	11 ±5	25 ±9	243 ±57	63 ±19
Amsco	63 ±5	72 ±10	94 ±19	44 ±10	43 ±13	11 ±4	50 ±8	688 ±15	188 ±17
Bazeries	64 ±4	74 ±13	94 ±20	60 ±15	61 ±20	17 ±5	49 ±5	477 ±44	112 ±21
Beaufort7	42 ±3	67 ±9	78 ±17	23 ±5	23 ±9	9 ±4	50 ±10	443 ±32	113 ±15
Bifid6	47 ±4	58 ±10	75 ±15	24 ±6	24 ±8	7 ±4	48 ±10	510 ±36	119 ±16
Bifid7	47 ±4	58 ±9	77 ±17	24 ±6	23 ±8	7 ±4	49 ±9	517 ±37	118 ±17
Cadenus	63 ±5	74 ±11	95 ±17	40 ±9	41 ±13	10 ±4	49 ±9	657 ±17	134 ±18
Cmbifid7	47 ±4	57 ±9	75 ±15	23 ±5	23 ±9	6 ±4	50 ±10	493 ±31	114 ±16
Columnar	63 ±5	73 ±11	96 ±18	41 ±8	41 ±12	11 ±4	50 ±7	653 ±16	128 ±15
Digrafid5	41 ±3	53 ±7	67 ±13	17 ±4	20 ±7	5 ±3	43 ±11	469 ±33	112 ±15
Dbl Ck Bd	90 ±13	133 ±18	149 ±23	110 ±30	207 ±58	25 ±5	13 ±7	609 ±44	133 ±19
4 Square	48 ±3	58 ±9	76 ±15	36 ±8	72 ±24	11 ±4	28 ±8	507 ±33	114 ±16
FracMorse	47 ±2	53 ±8	70 ±15	42 ±9	43 ±13	16 ±3	50 ±7	444 ±32	107 ±17
Grandpre	128 ±3	136 ±7	158 ±15	179 ±15	227 ±39	33 ±3	43 ±3	0 ±0	0 ±0
Grille	63 ±5	74 ±12	91 ±16	42 ±9	43 ±14	10 ±4	49 ±7	679 ±16	173 ±17
Gromark	39 ±1	46 ±7	63 ±13	15 ±3	15 ±6	4 ±3	50 ±12	431 ±26	109 ±15
Gronsfeld	40 ±2	66 ±8	76 ±19	21 ±5	25 ±11	9 ±4	42 ±14	444 ±27	111 ±15
Homoph	101 ±1	108 ±6	127 ±13	116 ±7	160 ±15	24 ±2	42 ±2	0 ±0	0 ±0
Mon Din	124 ±7	134 ±11	169 ±19	249 ±36	252 ±43	45 ±5	49 ±2	0 ±0	0 ±0
Morbit	122 ±4	129 ±7	156 ±16	193 ±15	194 ±25	38 ±2	49 ±2	0 ±0	0 ±0
Myszk	63 ±5	72 ±10	95 ±18	41 ±8	41 ±13	11 ±4	49 ±7	657 ±18	135 ±18
Nicodemus	42 ±3	50 ±7	73 ±14	18 ±4	18 ±7	5 ±3	50 ±10	442 ±35	112 ±15
Nihil Sub	144 ±11	201 ±23	195 ±30	218 ±33	266 ±42	38 ±4	40 ±6	0 ±0	0 ±0

Distinguishing between American Cryptogram Association (ACA) cipher types using various statistics

IC = index of coincidence
MIC = max index of coincidence (mean by cols)
MKA = kappa stat
DIC = digraphic index of coincidence
EDI = even position DIC

(values are multiplied by 1000, mean +/- SD given)

MACHINE LEARNING (2021)

A Massive Machine-Learning Approach For Classical Cipher Type Detection Using Feature Engineering

Ernst Leierzopf[1]*, Nils Kopal[2], Bernhard Esslinger[2], Harald Lampesberger[1], Eckehard Hermann[1]
[1]University of Applied Sciences Upper Austria, Hagenberg, Austria
[2]University of Siegen, Germany
*e.leierzopf@gmail.com

Abstract

Cryptanalysis of enciphered documents typically starts with identifying the cipher type. A large number of encrypted historical documents exists, whose decryption can potentially increase the knowledge of historical events. This paper investigates whether machine learning can support the cipher type classification task when only ciphertexts are given. A selection of engineered features for historical ciphertexts and various machine-learning classifiers have been applied for 56 different cipher types specified by the American Cryptogram Association. Different neuronal network models were empirically evaluated. Our best-performing model achieved an accuracy of 80.24% which improves the current state of the art by 37%. Accuracy is calculated by dividing the total number of samples by the number of true positive predictions. The software-suite is published under the name "Neural Cipher Identifier (NCID)".

plaintext into a quasi-random order. There are also ciphers combining substitution and transposition like ADFG(V)X (Friedman, 1941).

A typical cryptanalysis method for classical substitution ciphers is frequency analysis. Here, the frequencies of single or groups of multiple ciphertext symbols are counted and then compared to the frequencies of the assumed plaintext language. Then, based on the different frequencies in the plaintext language, assumptions of which letter was replaced by which symbol, can be made. Knowledge of the used cipher type allows the application of cipher-specific and heuristic algorithms to find the plaintexts more precisely. For example the Kasiski examination (1863) of the Vigenère cipher takes advantage of the fact that, by chance, repeated words are sometimes encrypted using the same key letters and therefore give indication for the possible key lengths.

The goal of this research is to determine how cipher type detection can be improved with machine learning approaches like feedforward neu-

Distinguishing between the 56 cipher types of the ACA.

amsco	grandpre	per. gromark	ragbaby
autokey	grille	phillips	railfence
baconian	gromark	phillips rc	redefence
bazeries	gronsfeld	plaintext	route transp.
beaufort	headlines	playfair	running key
bifid	homophonic	pollux	seriatedpfair
cadenus	key phrase	porta	slidefair
checkerboard	mnmedinome	portax	swagman
col. transp.	morbit	progkey	tridigital
condi	myskowski	quagmire1	trifid
cmbifid	nicodemus	quagmire2	trisquare
digrafid	nihilist subst.	quagmire3	two square
foursquare	nihilist transp.	quagmire4	variant
fract. morse	null	numbered key	vigenère

Leierzopf (2021) vs Callimahos (1977)

Abbr	Term	Description
SDD	Average Single Letter – Digraph Discrepancy Score	This feature uses a table of the differences between unigrams and bigrams. The score is calculated by adding each value at the position of the first letter in the alphabet times 26 plus the position of the second letter in the alphabet from the SDD table. The score is then divided by the length of the text minus 1. For normalization the scores are divided by 10.
CHI^2	Chi Square	With the Chi^2 function, the deviation from the distribution of English letters, which is known, can be calculated. This value is divided by 100 to be normalized.
DIC	Digraphic Index of Coincidence	Sum of all probabilities of the occurrence of two identical pairs of characters in a text times 1000.
IoC	Index of Coincidence	Sum of all probabilities of the occurrence of two identical characters in a text.
NOMOR	Normal Order	The frequency of each character is calculated and sorted by size. The normal order is the sum of the distances of all characters from their normal position divided by 1000.
PHIC	Phillips IC	Calculates the IC using a fixed column size = 5 and a fixed period = 8. The result is multiplied by 10. For ciphertexts that contain characters other than letters, this feature is 0.
REP	Repetition Feature	This feature is adopted from Nuhn and Knight (2014). It consists of the normalized number of exactly n times occurring identical characters for $2 \leq n \leq 5$. The normalization is calculated by dividing through the total number of repetitions.

a. The uniliteral frequency distribution, the monographic I.C., and its sigmage.
b. The over-all digraphic I.C., as well as the digraphic I.C.'s on cut "A" and cut "B" and their sigmages.
c. The over-all trigraphic I.C., and the trigraphic I.C.'s on cuts "A", "B", and "C" and their sigmages.
d. The local roughness (in terms of the observed and expected number of hits, and sigmage), when the message is offset against itself at offsets of 1 to 33.
e. Width tests, giving average columnar I.C.'s and sigmages of the message if it were written out on widths from 2 to 51.
f. The observed and expected number of tetragraphic and pentagraphic repetitions, and their I.C.'s and sigmages.
g. A listing of polygraphic repetitions of length 4 or longer.
h. If desired, the same categories of statistical information on the delta stream.

Conclusion: the index of coincidence (first described by Friedman in 1922) and its modifications are still useful in the diagnosis of (classical) ciphers. For modern ciphers, we might need to use other techniques.

Modern machine learning techniques really help.

FRIEDMAN AND CALLIMAHOS (1952)

---

[1] Although the foregoing four steps represent the classical or ideal approach to cryptanalysis, the art may be reduced to the following:

	Procedures in cryptanalysis	Requirements
1.	Arrangement and rearrangement of data to disclose non-random characteristics or manifestations (i.e., in frequency counts, repetitions, patterns, symmetrical phenomena, etc.).	Experience or ingenuity, and time (which latter may be appreciably lowered by the use of machine aids in cryptanalysis).
2.	Recognition of the non-random characteristics or manifestations when disclosed.	Experience or statistics.
3.	Explanation of the non-random characteristics when recognized.	Experience or imagination, and intelligence.

In all of the foregoing, the element of luck plays a very important part, as it is possible to side-step a large amount of labor and effort, in many cases, if "hunches" or intuition lead the analyst forthwith to the right path. Therefore, the phrase "or luck" should be added to each of the requirements above.

In fact, it all boils down to the simple statement: "Find something significant, and attach some significance thereto."

Zodiac Killer Cipher (Z408) Jul/Aug 1969

Homophonic substitution cipher

That is, one plaintext letter can be represented by several different plaintext letters. Harder than a MASC (monoalphabetic substitution cipher).

The Zodiac's 408 Character Cipher

Cipher Text & Plain Text Equivalents

S = A
G = A
J = A
▲ = A
Δ = A

V = B

Ǝ = C

ㄱ = D
⊕ = D

Z = E
W = E
N = E
E = E
9 = E
+ = E
O = E

Q = F
J = F

R = G

⊖ = H
M = H

P = I
ﾘ = I
Δ = I
U = I
Δ = I

/ = K

B = L
■ = L
◪ = L

ᘐ = M

O = N
D = N
Λ = N
ϕ = N

I = O
Q = O
X = O
T = O

π = P

Я = R
⊥ = R
\ = R

▲ = S
◘ = S
K = S
F = S

L = T
H = T
I = T
● = T

Y = U

Ɔ = V

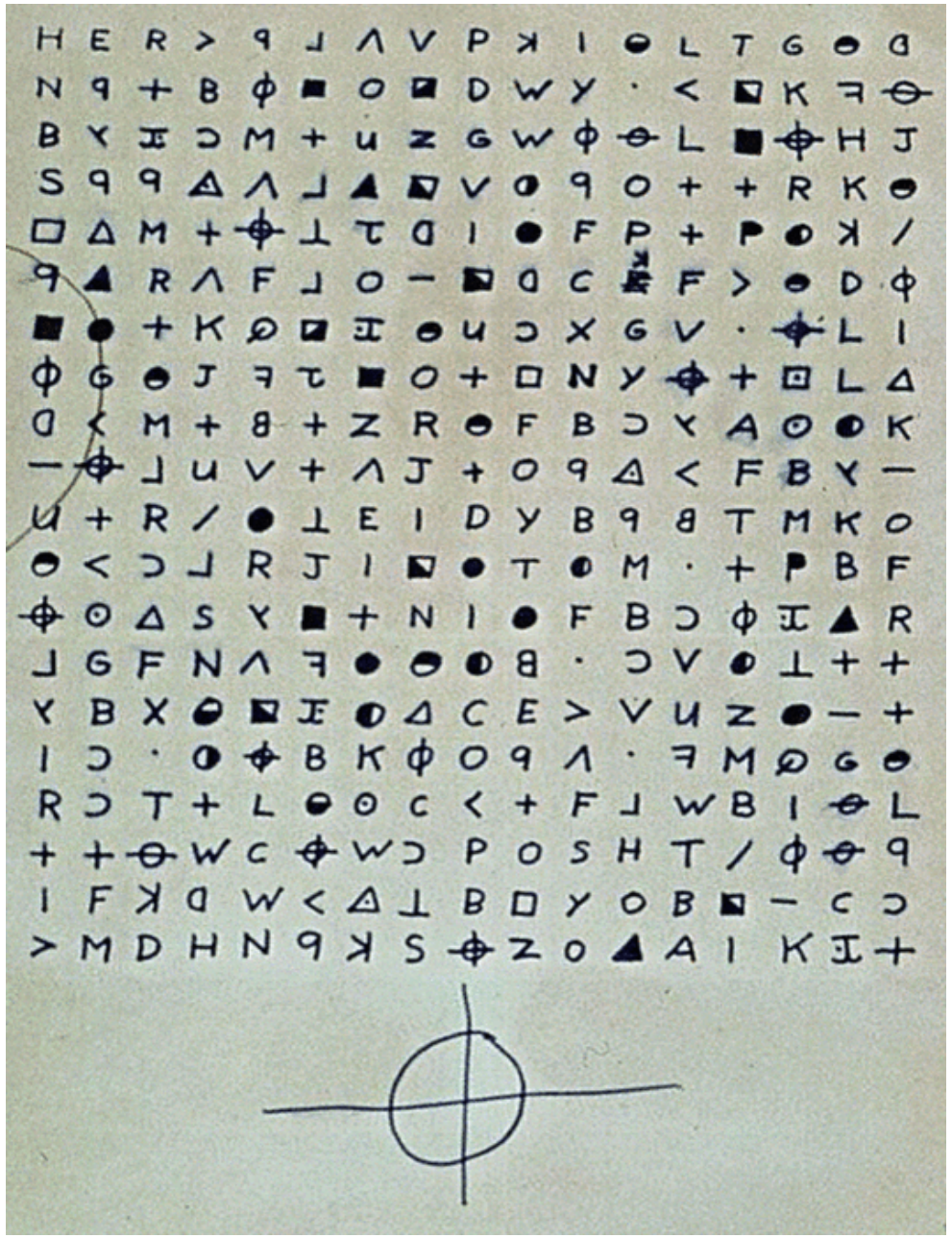
A = W

ㄊ = X

◘ = Y

Note: The Zodiac used the symbol "▲" for both "A" and "S"

Zodiac Killer Cipher (Z340) November 1969

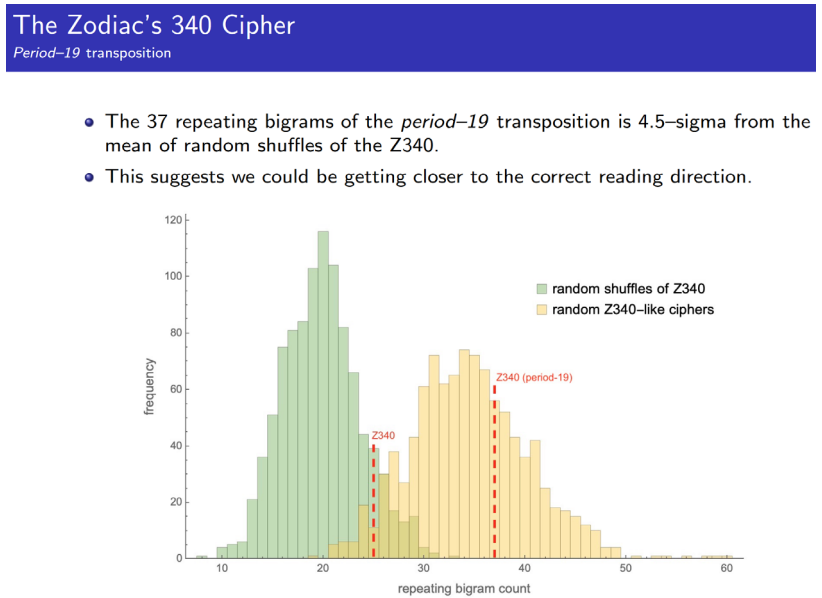


Z340 – IC 0.0192 – 340 letters / 63 symbols
http://zodiackillerciphers.com/wiki/index.php?title=Cipher_comparisons

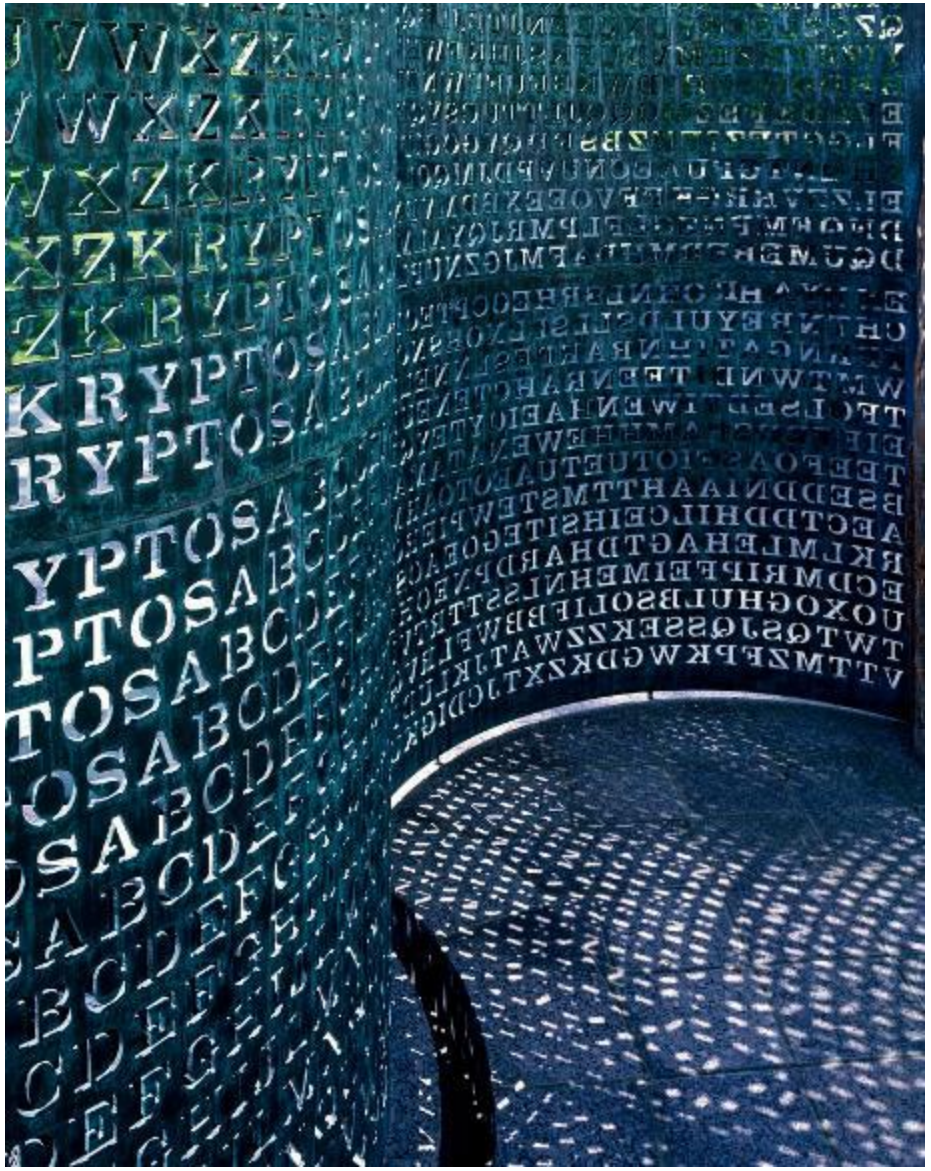
A relatively high number of repeated bigrams was seen at width 19 in the ciphertext. The cipher was found to be a combination of transposition and homophonic substitution (December 2020 – Oranchak, Blake, van Eycke).

The width 19 property can thus, after the fact, be deemed "causal" as the enciphering process caused this property to appear.

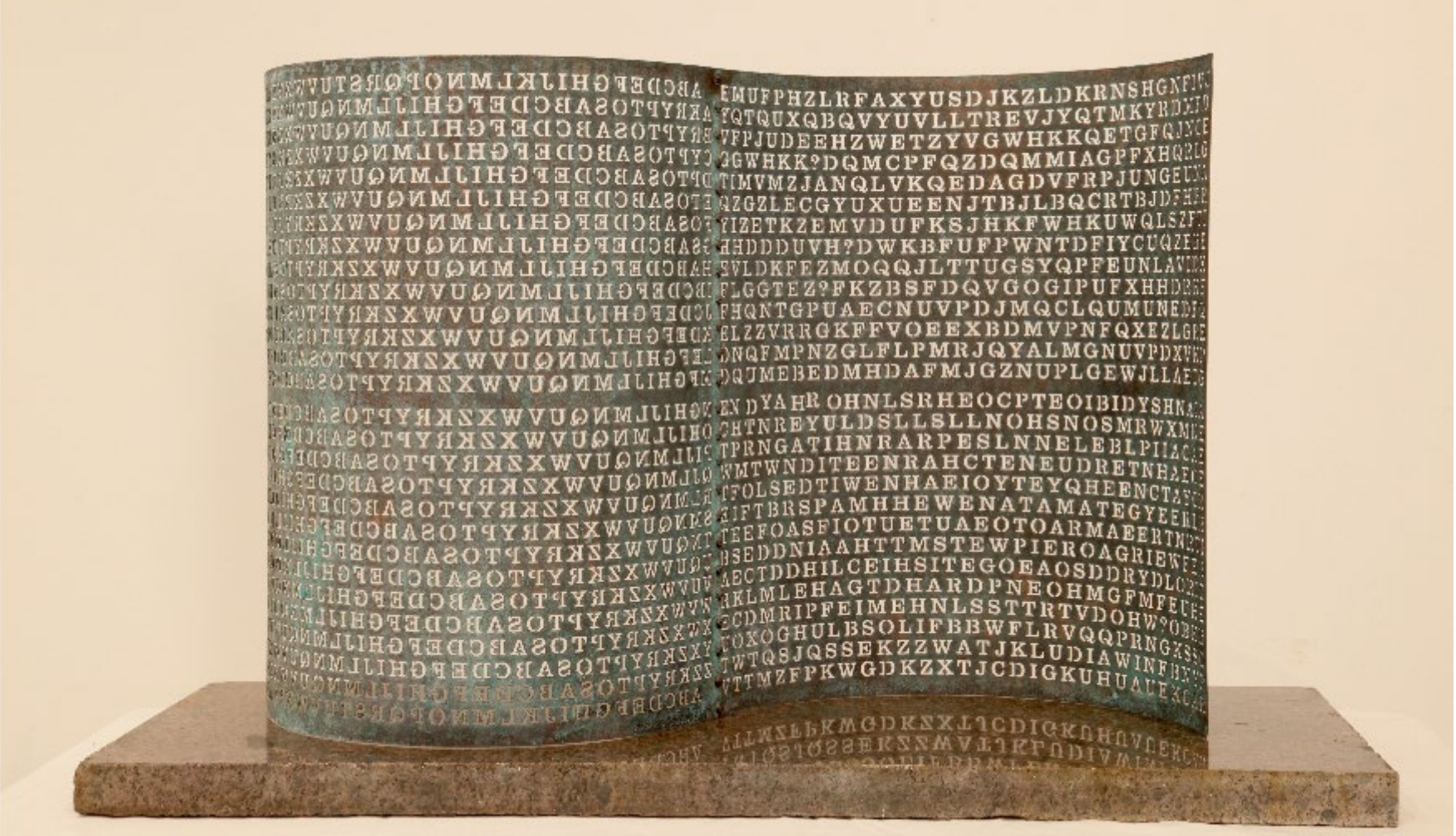
One hour talk by Sam Blake at https://youtu.be/D36gCEpGDU4



Kryptos, CIA Sculpture (1990)



KRYPTOS, "ALIAS" TV SHOW VERSION (2006)



https://www.artsy.net/artwork/jim-sanborn-the-kryptos-model … price in Sep 2020: $US40k

KRYPTOS, TRANSCRIPT

EMUFPHZLRFAXYUSDJKZLDKRNSHGNFIVJ
YQTQUXQBQVYUVLLTREVJYQTMKYRDMFD
VFPJUDEEHZWETZYVGWHKKQETGFQJNCE
GGWHKK?DQMCPFQZDQMMIAGPFXHQRLG
TIMVMZJANQLVKQEDAGDVFRPJUNGEUNA
QZGZLECGYUXUEENJTBJLBQCRTBJDFHRR
YIZETKZEMVDUFKSJHKFWHKUWQLSZFTI
HHDDDUVH?DWKBFUFPWNTDFIYCUQZERE
EVLDKFEZMOQQJLTTUGSYQPFEUNLAVIDX
FLGGTEZ?FKZBSFDQVGOGIPUFXHHDRKF
FHQNTGPUAECNUVPDJMQCLQUMUNEDFQ
ELZZVRRGKFFVOEEXBDMVPNFQXEZLGRE
DNQFMPNZGLFLPMRJQYALMGNUVPDXVKP
DQUMEBEDMHDAFMJGZNUPLGEWJLLAETG

EN $D^{YA}H^{R}$ OHNLSRHEOCPTEOIBIDYSHNAIA
CHTNREYULDSLLSLLNOHSNOSMRWXMNE
TPRNGATIHNRARPESLNNELEBLPIIACAE
WMTWNDITEENRAHCTENEUDRETNHAEOE
TFOLSEDTIWENHAEIOYTEYQHEENCTAYCR
EIFTBRSPAMHHEWENATAMATEGYEERLB
TEEFOASFIOTUETUAEOTOARMAEERTNRTI
BSEDDNIAAHTTMSTEWPIEROAGRIEWFEB
AECTDDHILCEIHSITEGOEAOSDDRYDLORIT
RKLMLEHAGTDHARDPNEOHMGFMFEUHE
ECDMRIPFEIMEHNLSSTTRTVDOHW?OBKR
UOXOGHULBSOLIFBBWFLRVQQPRNGKSSO
TWTQSJQSSEKZZWATJKLUDIAWINFBNYP
VTTMZFPKWGDKZXTJCDIGKUHUAUEKCAR

KRYPTOS, PART 1, PERIODIC POLYALPHABETIC REDUCTION

EMUFPHZLRFAXYUSDJKZLDKRNSHGNFIVJYQTQUXQBQV
YUVLLTREVJYQTMKYRDMFD

EMUFPHZLRF
AXYUSDJKZL
DKRNSHGNFI
VJYQTQUXQB
QVYUVLLTRE
VJYQTMKYRD
MFD
IC 0.03789

Mean(10 Column ICs) = **0.0914**

KRYPTOSABCDEFGHIJLMNQUVWXZ PT
IJLMNQUVWXZKRYPTOSABCDEFGH CT (col 4)
SABCDEFGHIJLMNQUVWXZKRYPTO CT (col 9)

FUNQUQ -> WSTOSO
RZFQRR -> SIGHIH

BETWEENSUB
TLESHADING
ANDTHEABSE
NCEOFLIGHT
LIESTHENUA
NCEOFIQLUS
ION

KRYPTOS, PART 2, PERIODIC POLYALPHABETIC REDUCTION

```
VFPJUDEEHZWETZYVGWHKKQETGFQJNCEGGWHKKDQMCPFQZDQMMIAGPFXHQRLGTIM
VMZJANQLVKQEDAGDVFRPJUNGEUNAQZGZLECGYUXUEENJTBJLBQCRTBJDFHRRYIZ
ETKZEMVDUFKSJHKFWHKUWQLSZFTIHHDDDUVHDWKBFUFPWNTDFIYCUQZEREEVLDK
FEZMOQQJLTTUGSYQPFEUNLAVIDXFLGGTEZFKZBSFDQVGOGIPUFXHHDRKFFHQNTG
PUAECNUVPDJMQCLQUMUNEDFQELZZVRRGKFFVOEEXBDMVPNFQXEZLGREDNQFMPNZ
GLFLPMRJQYALMGNUVPDXVKPDQUMEBEDMHDAFMJGZNUPLGEWJLLAETG

VHGGGCMQMKFUEEQHKKKTVFYEMUUFKGHQCQEVOPGPMGKEJEG
FZWFWPIRZQRNCNCRZSUIHPCVOGNLZOHNNCDRENRNRNPDGW
PWHQHFALJEPAGJRREJWHDWULQSLGBGDTULFREFEZJUDMZJ
JEKJKQGGADJQYTTYMHQHWNQDQYAGSIRGVQQGXQDGQVQHNL
UTKNKZPTNAUZUBBIVKLDKTZKJQVTFPKPPUEKBXNLYPUDUL
DZQCDDFIQGNGXJJZDFSDBDEFLPIEDUFUDMLFDEQFADMAPA
EYEEQQXMLDGZULDEUWZDFFRETFDZQFFAJUZFMZFLLXEFLE
EVTGMMHVVVELEBFTFHFUUIEZTEXFVXHEMNZVVLMPMVBMGT
IC 0.04547
Mean(8 Column ICs) = 0.0688

KRYPTOSABCDEFGHIJLMNQUVWXZ PT
SABCDEFGHIJLMNQUVWXZKRYPTO CT (col 7)
```

OVOOHHMFETANIETOILNTSSUOXSTNHSSRDINSFNSEEMOSEO

KRYPTOS, PART 3, TRANSPOSITION

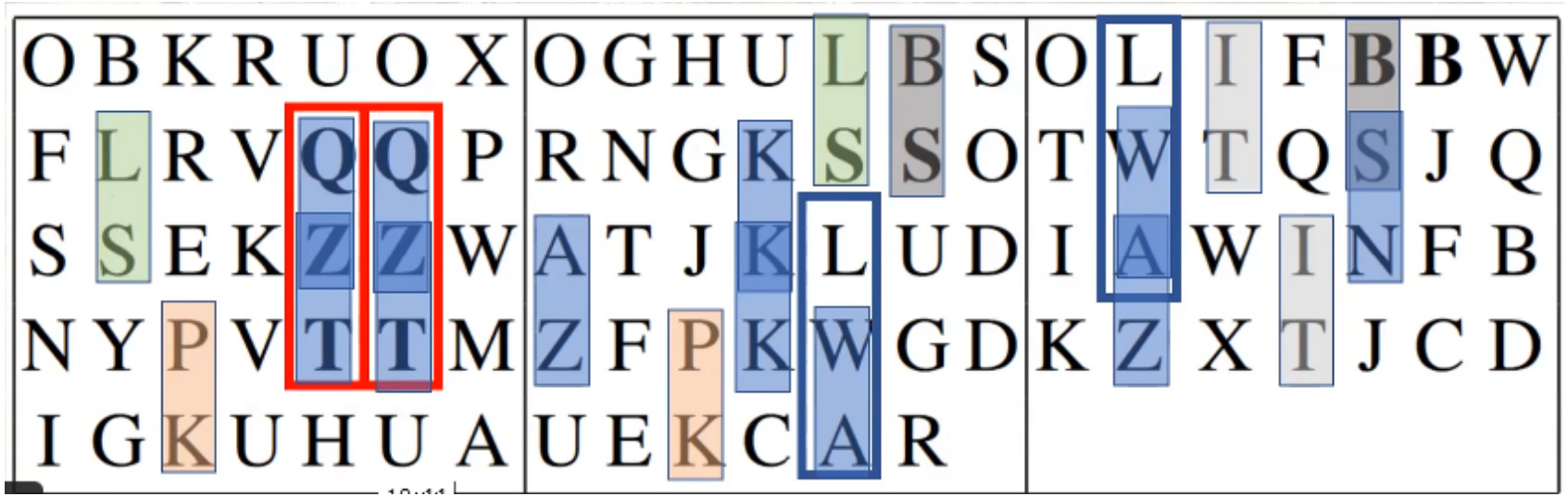
```
ENDYAHROHNLSRHEOCPTEOIBIDYSHNAIA
CHTNREYULDSLLSLLNOHSNOSMRWXMNE
TPRNGATIHNRARPESLNNELEBLPIIACAE
WMTWNDITEENRAHCTENEUDRETNHAEOE
TFOLSEDTIWENHAEIOYTEYQHEENCTAYCR
EIFTBRSPAMHHEWENATAMATEGYEERLB
TEEFOASFIOTUETUAEOTOARMAEERTNRTI
BSEDDNIAAHTTMSTEWPIEROAGRIEWFEB
AECTDDHILCEIHSITEGOEAOSDDRYDLORIT
RKLMLEHAGTDHARDPNEOHMGFMFEUHE
ECDMRIPFEIMEHNLSSTTRTVDOHW?

IC = 0.06615; freq order is
ETARNHIODSLMCYWPFGBUXVQK

336 or 337 characters
```

```
SLOWLY DESPARATLY SLOWLY THE
REMAINS OF PASSAGE DEBRIS THAT
ENCUMBERED THE LOWER PART OF
THE DOORWAY WAS REMOVED WITH
TREMBLING HANDS I MADE A TINY
BREACH IN THE UPPER LEFT HAND
CORNER AND THEN WIDENING THE
HOLE A LITTLE I INSERTED THE
CANDLE AND PEERED IN THE HOT
AIR ESCAPING FROM THE CHAMBER
CAUSED THE FLAME TO FLICKER BUT
PRESENTLY DETAILS OF THE ROOM
WITHIN EMERGED FROM THE MIST X
CAN YOU SEE ANYTHING Q
```

(from Howard Carter "Tomb of Tutankhamun")

KRYPTOS, PART 4, WHO KNOWS?

- A famous unsolved cipher, become famous if you solve it!
- IC = 0.03608 – 97 letters. Very flat letter distribution.
- Four clues or “known plaintext” given by sculptor BERLINCLOCK (letters 64-74) and EASTNORTHEAST (letters 22-34) – seem to correspond letter for letter
- Since 1990, people have observed many *possibly* “non-random” patterns
- But are the patterns “real” (or, more accurately, “causal”) and what do they mean?



The Hantman family

492.40, 448.01, 450.33, 446.19, 31010, 44619, 45500, 43100, 53209, 32210, 34827, 45003, 44716.
Washington: 52117, 30444, Fairfax, repeating, Fairfax -
52117, 35001, 40631, 34827, 41103, 53942, 45003, 31200, 36900, 47600, 42100, 36900, 42200, 36900, 56400, 51504, 49000, 51721, 45003, 55600, 55200.

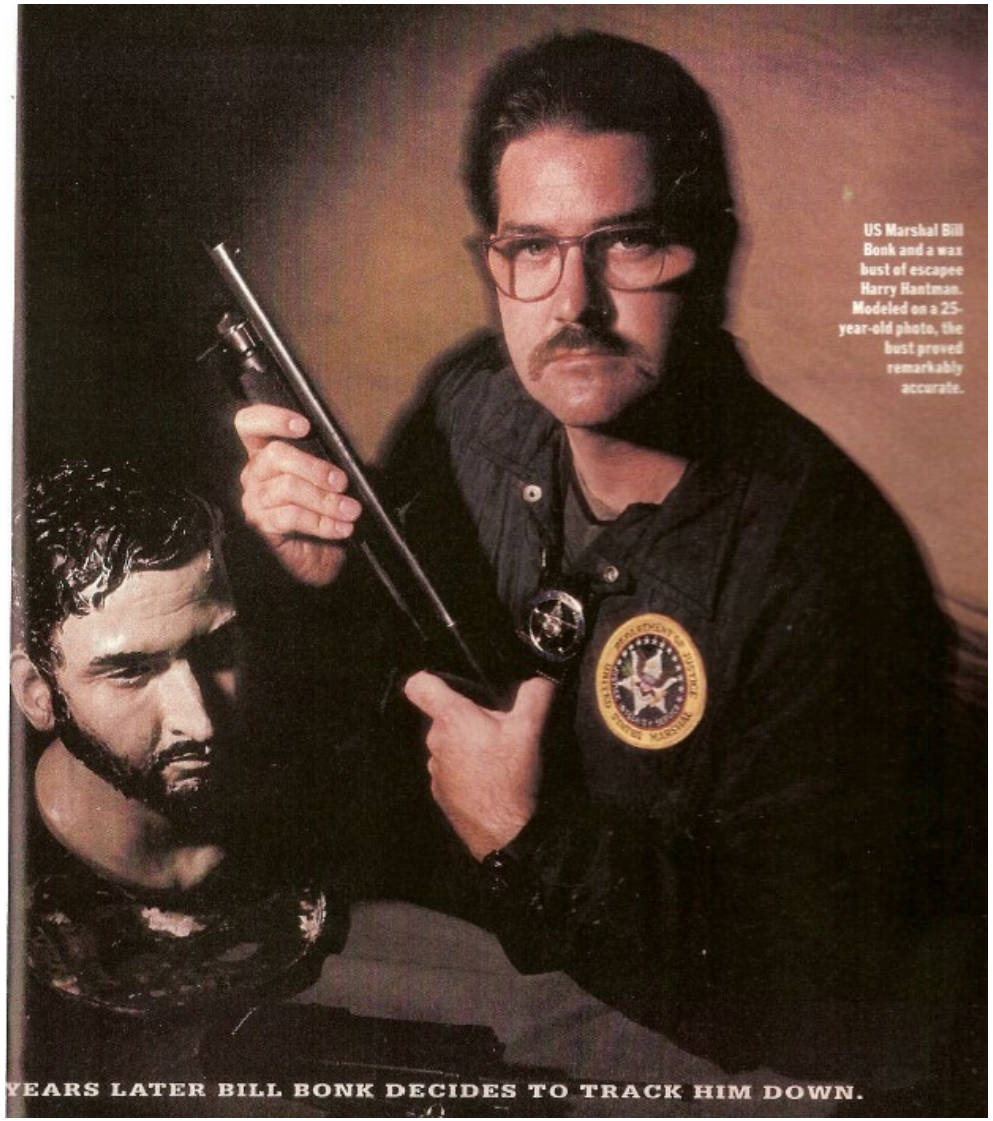
53203, 32210, 31010, 44819, 54523, 51139, 39900, 29700, 47600, 47600, 40800, 49000, 44800, 44400, 52117, 42200, 36900, 36900, 49000, 31200, 54500, 47600, 39100, 51721, 45003, 55200, 29700, 54338, 43508, 38328, 31010, 53226, 45500, 43100.

[US Marshal Bill Bonk] … shipped [the tape] to decoding specialists [at the NSA who'd previously been the first to break the first three passages of the "Kryptos" sculpture at the CIA].

It took months, but they broke enough of the code - it turned out to be based on pages and words in a French-English dictionary - to establish that the calls had been made in 1983.

"The Fugitive" by Harry Jaffe in *"The Washingtonian"* January 1994

Lance: *But without having the dictionary, you can still get far.*



A man called one of the most notorious of the nation's missing criminally insane was recaptured Friday morning by federal marshals outside a Lewiston motel. (26 Mar 1993)

CHALLENGE CRYPTOGRAMS - HANTMAN FAMILY *UNSOLVED*

824.02, 814.63, 1553.37, 895.50,
660.53, 1364.45, 1389.99, 1422.08,
804.06, 958.14, 1159
(Interruption to tape.)
--repeating, 1159.58, 868.36, 1316.11,
1112.03, 703.74, 1246.82, 1552.80,
706.25, 1312.70.

This is from book four, and page two,
volume two, number 220 through 239.
Here we go.
1128.47, 1113.26, 1468.28,
1028.71, 1402.28, 1057.87, 996.93,
1489.60, 985.49, 927.35, 924.94,
749.11, 1289.61, 953.91, 1131.53,
1194.45, 728.18, 1160.24, 975.93,
1243.71. Okay?

This is from book seven, and
page six, volume two, numbers 901
through 907. Message begins. If I
give you A1265.15, 971.33, and 759.31,
would you be able to 1149.64, 1055.44,
and 1091.00, absolutely without being
1151.79.

That's the end of the
message.

This is from book seven, and
page six, volume two, number 908
through 941.
Starting 571.40, 1202.35,
557.05, 1327.48, 1354.22, 619.03,
772.14, 767.13, 1141.01, 986.48,
1227.23, 812.00, 915.16, 1338.58,
1257.09, 987.31, 637.92, 957.95,
882.91, 9–
(Interruption to tape.)
MALE VOICE #1:
Are you still there?
FEMALE VOICE:
Yeah.
MALE VOICE #1:
Okay, I'm back. 988.98, 1239.91,
1291.04, 743.66, 861.80, 923.05,
1444.92, 1261.35, 1362.75, 1022.59,
608.51, 1013.32, 776.82, 622.96,
1461.40. End of message. Have you got
that?