Cryptodiagnosis of “Kryptos K4”

Richard Bean

HistoCrypt 2022 (paper from HistoCrypt 2021)

Tuesday 21 June 2022, 9am, Amsterdam Trippenhuis

Summer Solstice

Cryptanalytic diagnosis

Given ciphertext and context such as known plaintext or a set of possible methods…

What type of cipher is in use here?

Cryptanalytic diagnosis is a classification problem

Systematized over time as volumes of ciphertext increased and classification techniques improved.

Automation and the invention of the computer

Machine learning techniques especially since ~2005.

Let's look at the history of books and techniques (20th century)

"You see, the diagnosis, the most difficult part of the cryptanalysis, seemed to me to be on a different plane.

In fact I came up with a suggestion once that cryptanalysts should be in a special world."

Brigadier John Tiltman (1978)

Gaines (1939)

Final chapter XXIII on “Investigating the Unknown Cipher”

The “Bible” of the American Cryptogram Association (ACA) by PICCOLA known as “ELCY”.

A sequence of questions – like a *decision tree*

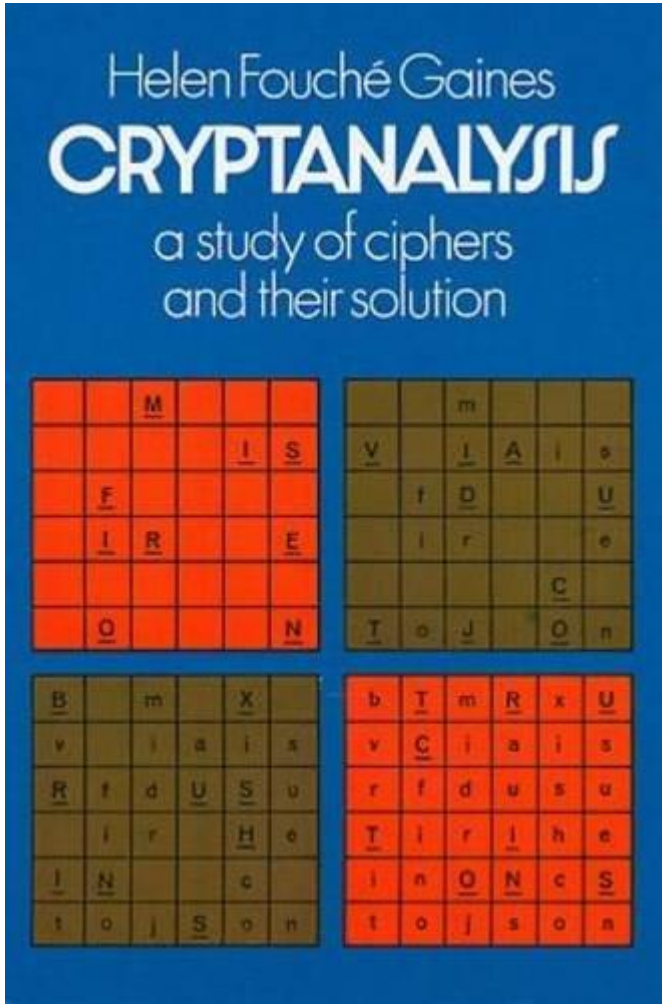
What is the probable language? Consider source and history of cryptogram.

Then – is it *transposition, substitution, or a combination?*

“Great evenness in frequencies” may suggest autokey or progressive key

“Absence of repeated sequences” usually means substitution plus transposition

Digram counts – could this be Playfair?



Sinkov (1970)

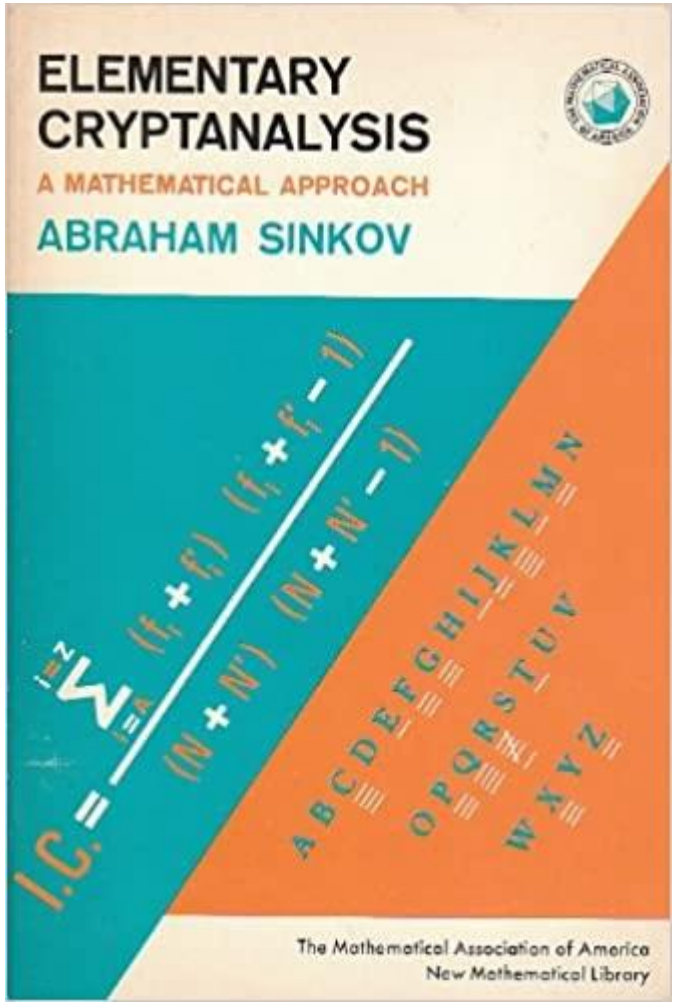
Introducing the index of coincidence

A mathematical approach from a deputy director of the NSA

Explaining the *Index of Coincidence* and *Digraphic Index of Coincidence* of Friedman to a wider audience.

Examples are possibly a bit contrived

e.g. for a flat digraphic system, the DIC is 1/676 = 0.015, but at what value of DIC should we say a system is certainly digraphic?



Callimahos (1977)

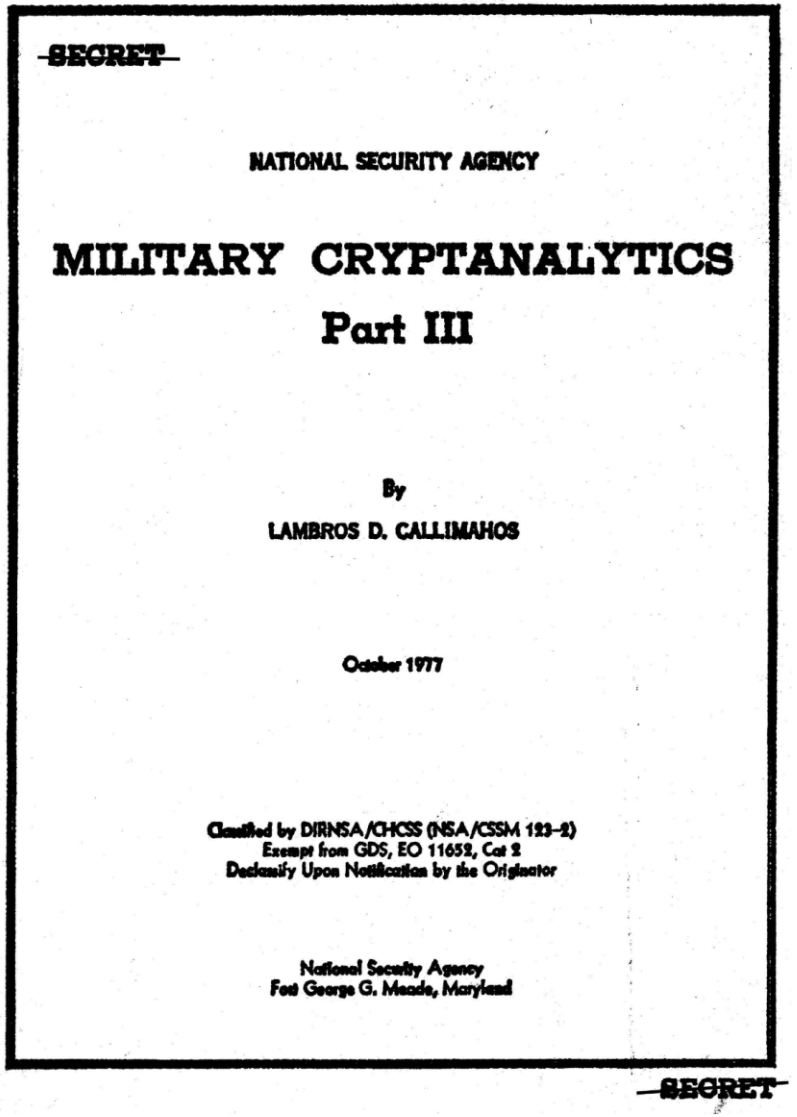
Declassified 2020 – “Cryptodiagnosis” chapter

NSA cryptanalysis teacher

Military Cryptanalysis and Military Cryptanalytics were a series of NSA cryptanalysis books published up until 1977.

A systematic approach to “Cryptodiagnosis” developed over several decades.

- Arrange data to disclose non-random characteristics
- Recognize non-random characteristics
- Explain the recognized characteristics (experience and imagination)
- “Luck plays a very important part”


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

Lewis (1992)

Chapter 12 on “Diagnostics”

Another WW2 / NSA cryptanalyst

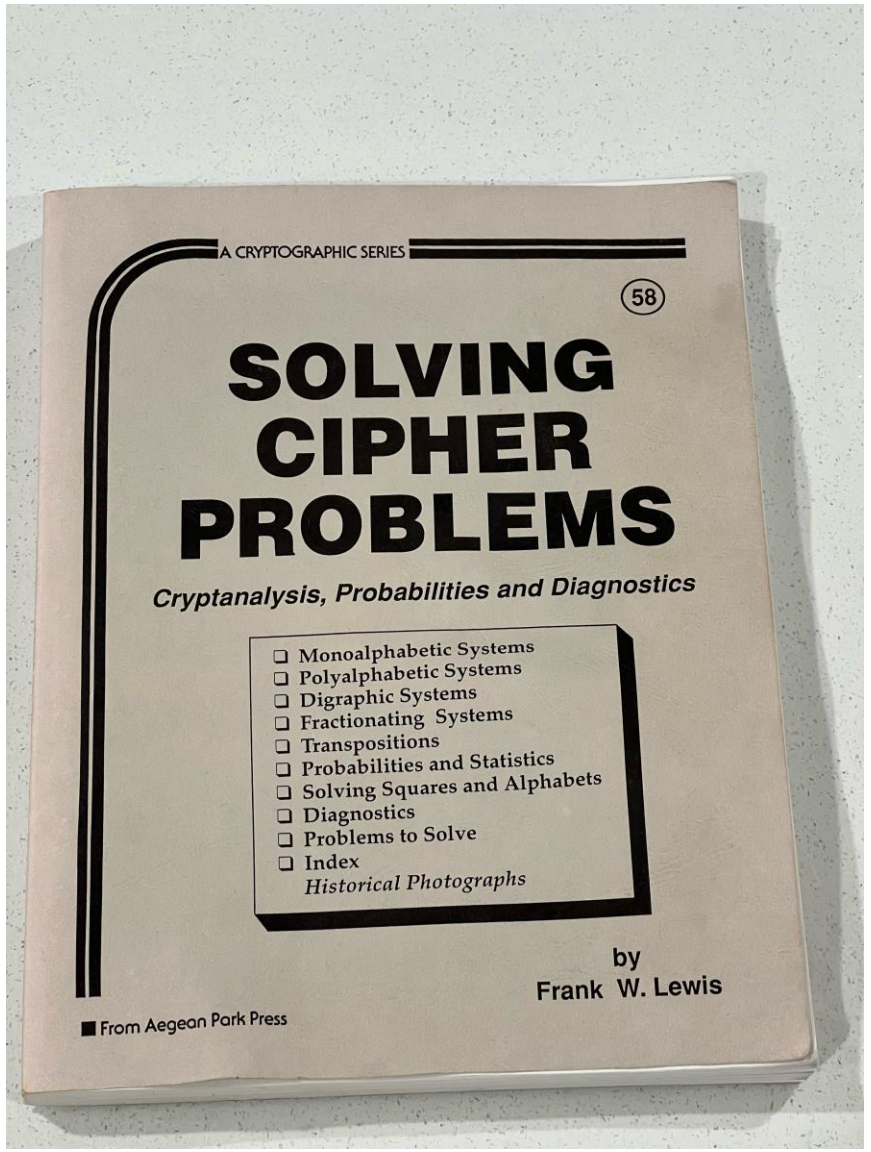
But book written in context of “American Cryptogram Association” (ACA) ciphers.

All the ACA ciphers to that point are described, and a systematic approach is given for identifying the “unknown cipher”.

Aegean Park Press – long out-of-print.

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”.



Kumar (1997)

Chapter 12 on “Diagnostics”

Research at India’s DRDO and Joint Cipher Bureau

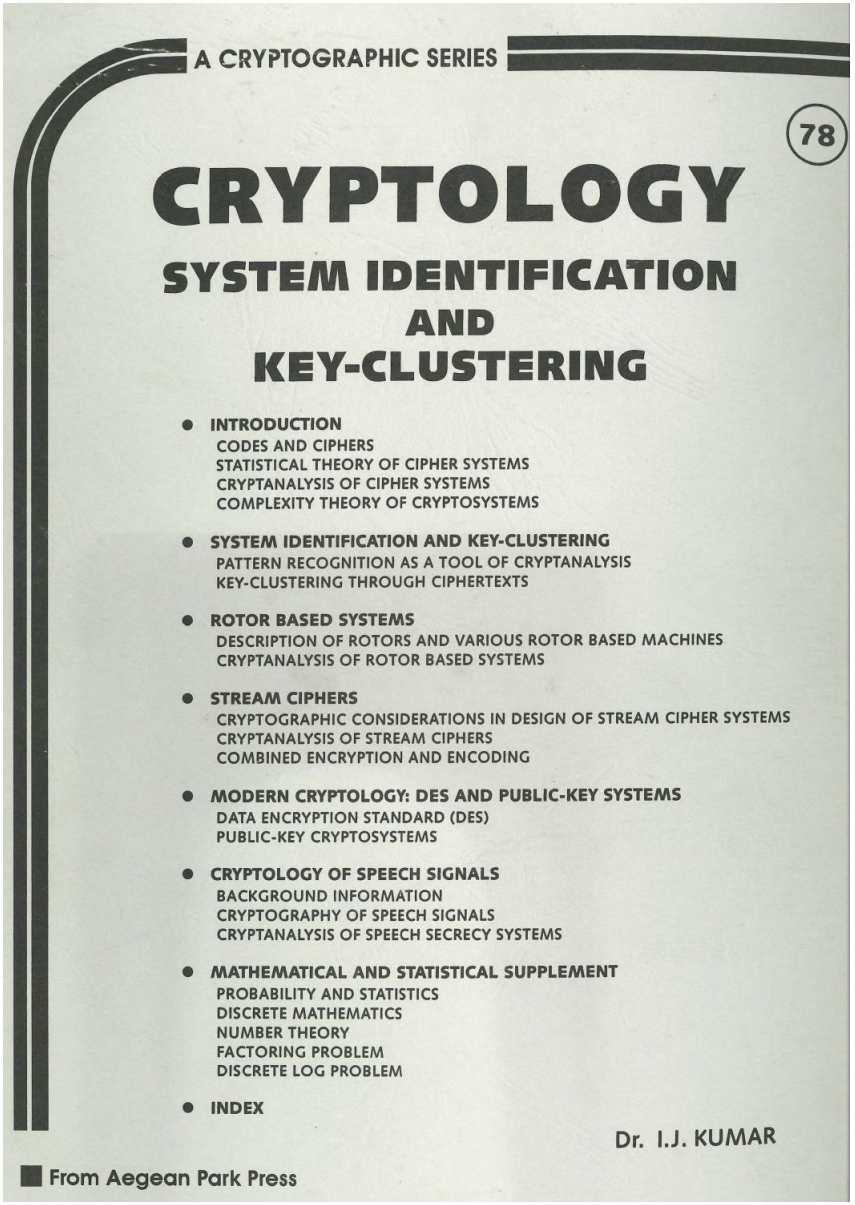
Another *Aegean Park Press book* out-of-print for 20+ years.

Pattern recognition as a tool of cryptanalysis

Linear discriminant functions to distinguish between sets of ciphers:

- Hill, Vigenere and Playfair cipher;
- Stream ciphers – Geffe, non-linear combiner function (NLC), non-linear feed-forward system with linear feedback (NLFFSR);
- Rotor systems – Hebern, Enigma, Typex

A chain of citations through to modern ML papers, but not cited much as long out-of-print



The age of machine learning

- **NSA SIDtoday** (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 run cryptanalytic routines [which] employ standard cryptanalytic tests which search for patterns and non-random properties in data. … If I detect a significant statistical property in data, I will immediately seek, expect, and receive help from team members.*

Projects to classify ACA ciphers

- **BION** 2005 to date – starting with random forest
  - https://williammason.github.io/rec-crypt/
- **Nuhn and Knight** 2014 *"Cipher Type Detection"* – support vector machine
- **Leierzopf, Kopal, Esslinger, Lampesberger and Hermann** 2021 *"A massive machine-learning approach for classical cipher type detection using feature engineering"* – random forest, FFNN, decision trees.
- **Leierzopf, Mikhalev, Kopal, Esslinger, Lampesberger and Hermann** 2021 "Detection of classical cipher types with machine-learning approaches" https://www.cryptool.org/en/cto/ncid

These classifiers are quite accurate but this comes at the cost of explainability (e.g. random forest "black box").

Kryptos

A sculpture at the CIA with four enciphered passages.

Timeline

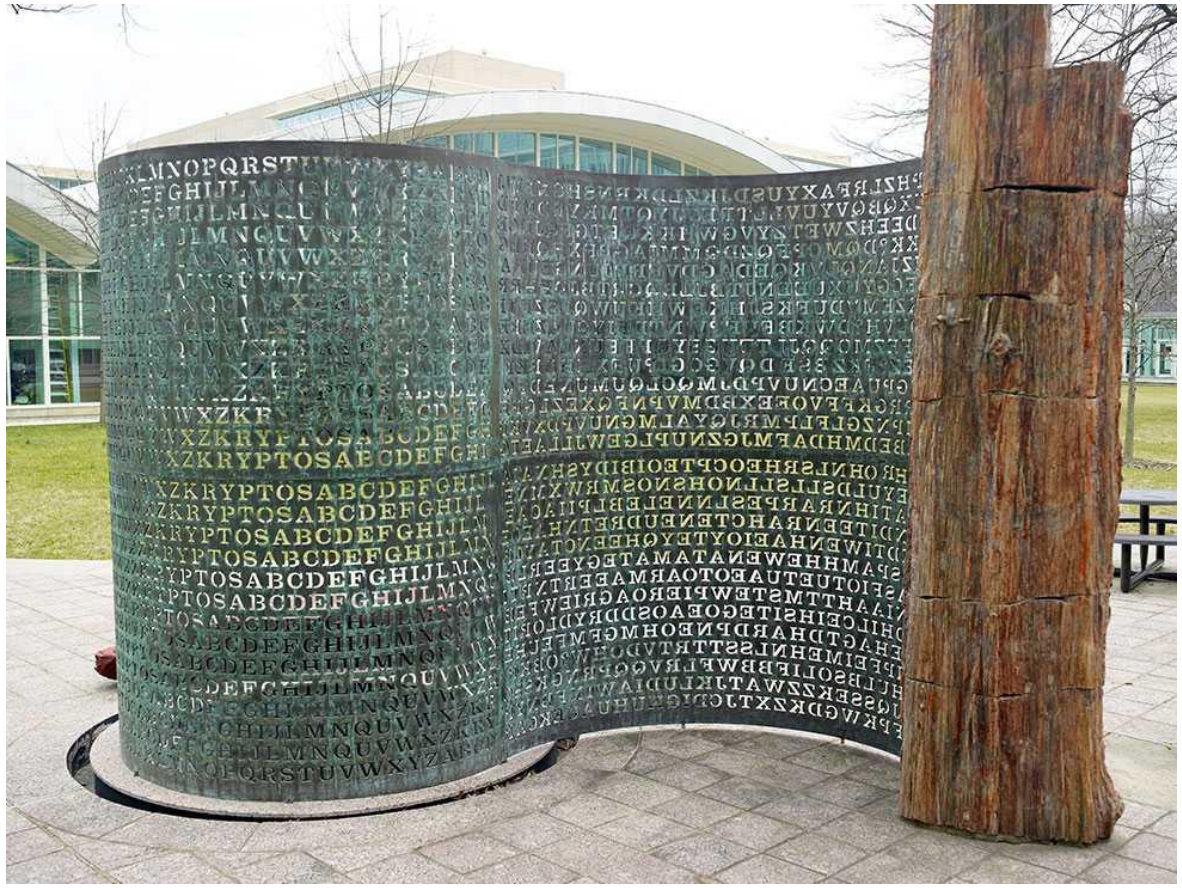
- May 1984. New Headquarters Building (NHB) construction begins
- Oct 1987. Artwork commissioned to “reflect life in all its positive aspects (truth, justice, courage, liberty etc) which “should not produce feelings of futility”
- Mar 1988. DC sculptor Jim Sanborn selected
- Dec 1989. Sanborn letter to CIA employees. “The right side is a text that can be partly deciphered by using the [left-side Vigenere] table and partly by using a ***potentially challenging encoding system*****.**”
- Jan 1990. Sanborn: (Non-Vigenere section) “will be encoded in a ***modern system created for the project by an expert cryptographer*** whom Sanborn would not identify"
- Nov 1990. Dedication. A *“game for viewers”* and *“tribute to information”.*
- Apr 1991. Sanborn: (Non-Vigenere section) is “a ***whole different ballgame of multiple codes*** written by a retired CIA cryptographer”

Kryptos

A sculpture at the CIA with four enciphered passages.

Timeline

- May 1991. Cryptographer named as Ed Scheidt in *Der Spiegel*
- 1992. First solution – first three passages solved by NSA cryptanalysts: Ed Hannon, Lance Estes, Denny McDaniels
- 1998. Second solution – CIA physicist David Stein
- 1999. Third solution – Jim Gillogly (knew of Scheidt involvement)



Kryptos

A sculpture at the CIA with four enciphered passages.

Interlude

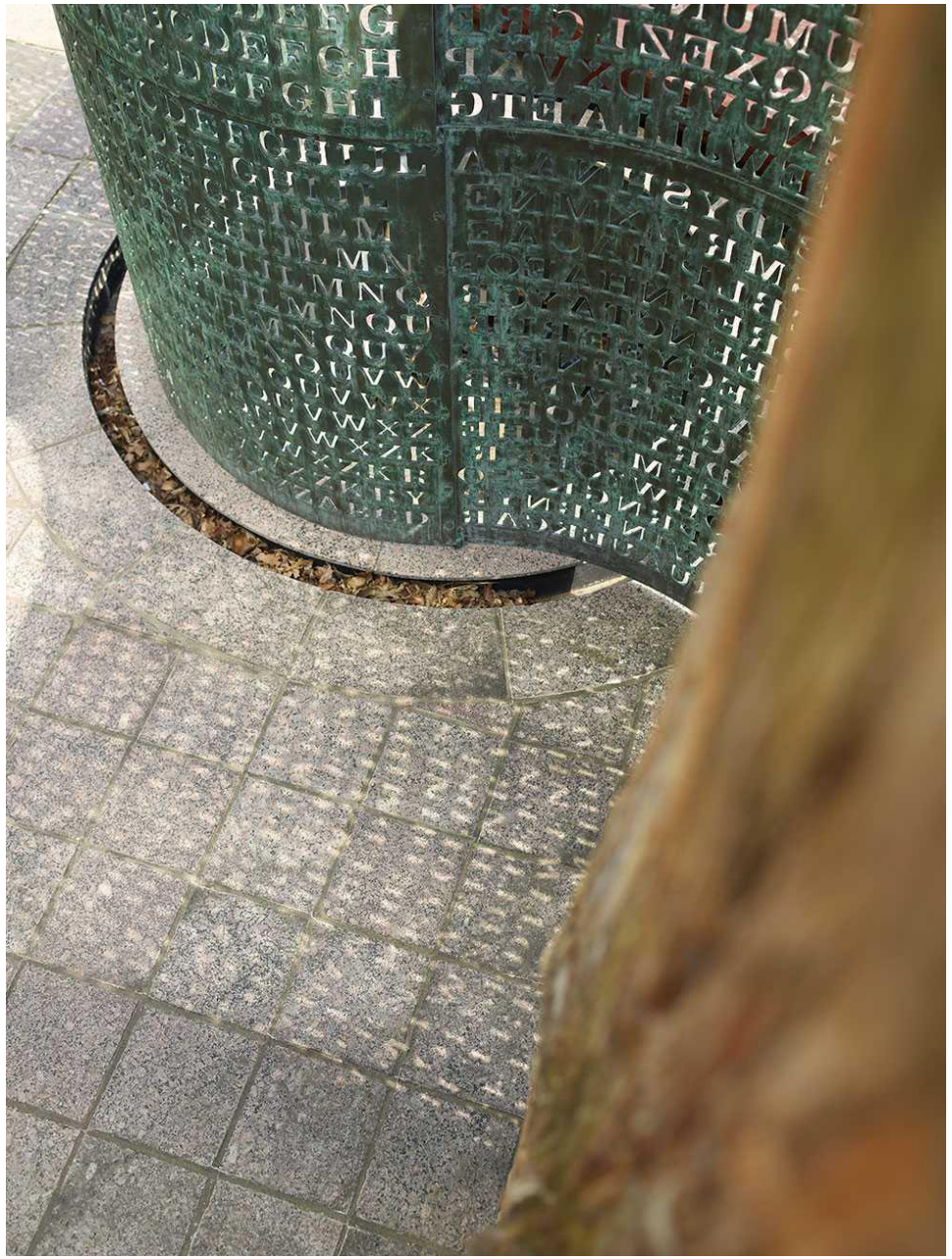
- The fourth passage of 97 letters (“K4”) remains unsolved

ECDMRIPFEIMEHNLSSTTRTVDOHW?**OBKR**

UOXOGHULBSOLIFBBWFLRVQQPRNGKSSO

TWTQSJQSSEKZZWATJKLUDIAWINFBNYP

VTTMZFPKWGDKZXTJCDIGKUHUAUEKCAR

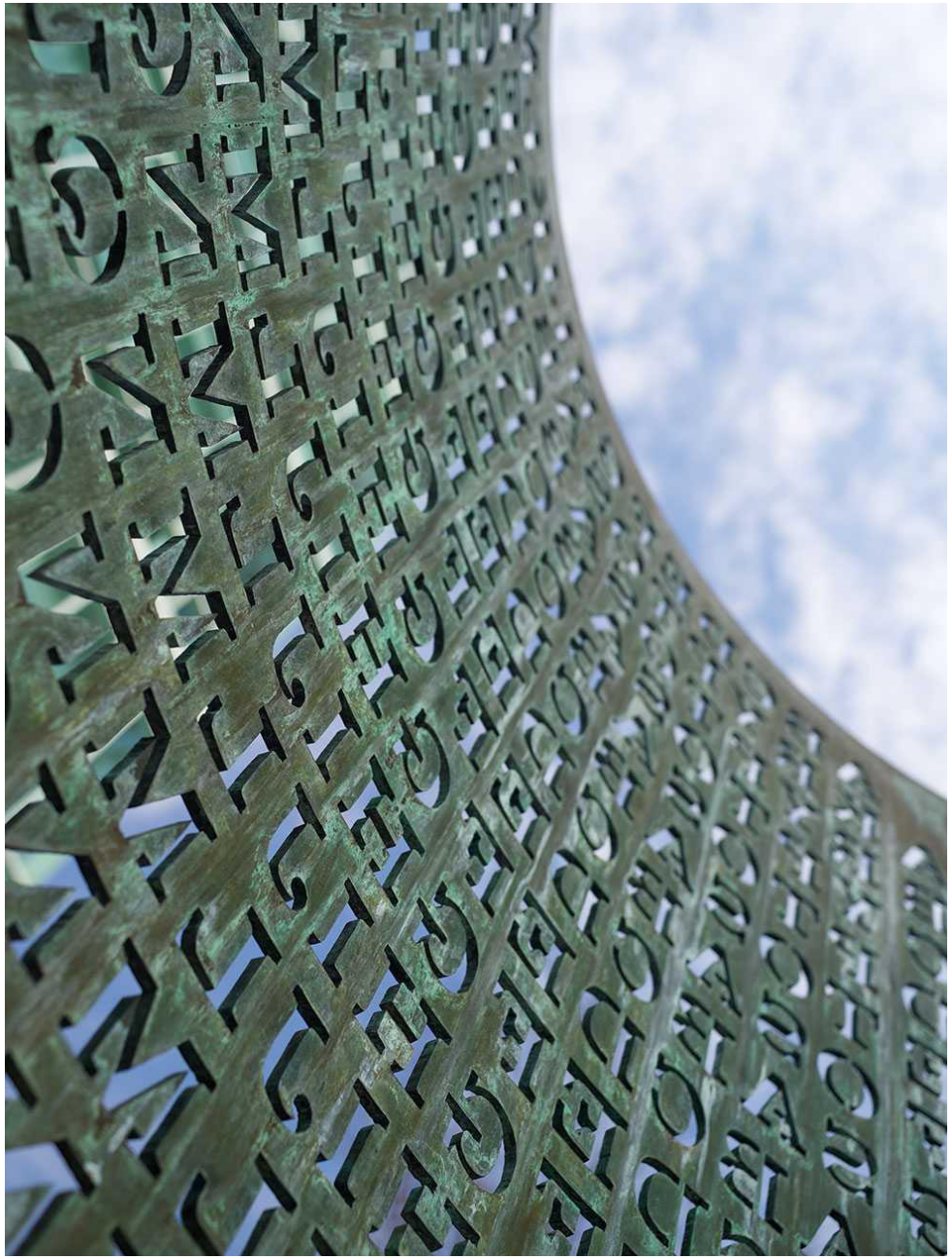


Kryptos

A sculpture at the CIA with four enciphered passages.

Back to timeline – Sanborn plaintext releases

- 2010. First plaintext released – BERLIN letters 64-69
- 2014. Second plaintext released – CLOCK letters 70-74
- Jan 2020. Third plaintext relased – NORTHEAST letters 26-34
- Apr-Aug 2020. Fourth plaintext released – EAST letters 22-25
- In total: 24 known plaintext letters and positions released



Kryptos

Why is it worth giving a talk on this?

Sculpture as a history of cryptography

- Lessons learned will help explain cryptographic history
- “Scheidt’s underlying goal was to [present] the sculpture, in some sense [as] a history of cryptography” – Gillogly 1999
- After so many plaintext releases, needs systematic summary.
- Main online “discussion group” has many posts but very little measurement going on

Starting point

Learning from books presented

Gaines

- Decision tree – look at index of coincidence and letter frequency; a *random forest* is like a generalization of a decision tree
- IC = 0.03608 – very flat; all 26 letters occur and K, T, O, S are among most frequent
- No significant CT repetitions (e.g. 10 repeated bigrams, no repeated trigrams as in K1, K2)
- Perhaps autokey, progressive key, combined transposition and substitution

Lewis

- *Find phenomenon (or phenomena), measure it, explain it, exploit it*
- Extraordinary number of digrams when CT written at width 21 noted by independent observers
  - Similar to Zodiac cipher Z340 which had large number of digrams when CT written at width 19
  - **Explained** by combination of substitution and transposition
  - **Exploited** by large scale computer search (Oranchak, Blake, van Eyck 2020)

Repeated bigrams

Measure and explain an unusual phenomena

OBKRUOXOGHULBSOLIFBBW

FLRVQQPRNGKSSOTWTQSJQ

SSEKZZWATJKLUDIAWINFB

NYPVTTMZFPKWGDKZXTJCD

IGKUHUAUEKCAR

e.g. (AZ BS IT LS LW PK QZ SN WA ZT KK) repeated.

Monte Carlo sampling of CT – a 1 in 6,750 chance.

But, should measure over all possible widths or DIC.

Does this reflect **underlying repeated plaintext**?

If so, a **seriated digraphic cipher** is a possibility.

Now known plaintext eliminates many possibilities

e.g. if seriation is at width 21, and PT BERLIN goes to CT NYPVTT, different bigrams ending in “I” and “N” must both encipher to “ZT”. Impossible.

If CLOCK goes to MZFPK, then method can’t be based on Playfair – letters can’t encipher to themselves.

Use known plaintext

Use revealed clues from 2010-2020

"Weight of evidence" changes with known PT. *Bayesian inference.*

For repeated PT letters, check corresponding CT letters

```
Plain  EAST NORTH EAST BERLIN CLOCK

Cipher FLRV QQPRN GKSS NYPVTT MZFPK
```

- A goes to L and K
- C goes to M and P
- L goes to V and Z.

CT letters are very close in A-Z alphabet. Mean distance = 3.6; MC sampling ~ 1 in 240 permutations.

If for all *i, j* look at PT letters where $d(P_i, P_j) \leq 1$ & calculate corresponding $d(C_i, C_j)$ mean = 3.8, ~1 in 5,000.

Materna observation: PT letters in {K,R,Y,P,T,O,S} move very little in CT, mean distance = 2.1… ~1 in 5,500 chance.

Can we explain or exploit these observations?

Gromark cipher

Described in ACA and NSA publications

Uses a PLAIN and CIPHER alphabet, and a KEY generated from a primer by repeated addition. E.g. primer length 5, base 10

e.g.

```
pt: a b c d e f g h i j k l m n o p q r s t u v w x y z

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

encipherment:

K:  2345257977266498203702307253797 8066

pt: thereareuptotensubstitutesperletter

CT: NFYCKBTIJCNWZYCACJNAYNLQPWWSTWPJQFL
```

- DUMBO (W. J. Hall) in “The Cryptogram” of the ACA - 1969
- Callimahos in “Military Cryptanalytics Part III” 1977.

Gromark cipher as potential K4 cipher

Arguments for and against

FOR

- Observations could be explained by PT and CT alphabets being “close” to A-Z e.g. mixed keyword based.
- Gromark one of few ACA ciphers where IC almost “flat” (i.e. 1/26)
- Even base and primer length 5 -> underlying CT pattern at width 21
- Fits “letter to letter” encipherment observed i.e. **no transposition**
- Fits hints “pencil-and-paper field cipher”, “more than one stage”, use of “base” arithmetic
- Open source method, difficulty level seems to be calibrated

e.g. keywords impounded, deubiquitylate

```
IMPOUNDEABCFGHJKLQRSTVWXYZ

DEUBIQTYLACFGHJKMNOPRSVWXZ

EAST NORTH EAST BERLIN CLOCK

4012 02104 5721 801661 69890

FLRV QQPRN GKSS NYPVTT MZFPK
```

Gromark cipher as potential K4 cipher

Arguments for and against

AGAINST

- “K4 cryptography is not mathematical although this does not preclude it being modelled mathematically”
- Sanborn “anathemath”
- PT alphabet often A-Z
- Observable pattern at width 21 perhaps implies low base which means fewer ciphertext alphabets, increasing IC
- Exhaustive search over many bases, key gen methods failed. E.g. base 10, primer length 5 – 39 of 99,999 primers
- Stats are strongest *within* the cribs but not *between* them → progressive key? Periodic Gromark?

Conclusions

BION and NCID

- BION’s tool and Leierzopf’s NCID give **Seriated Playfair** and **Gromark** as their top choices.
- ML “Black Box” output - explanation

Notes

- K4 “never checked” by Scheidt so possibility of mistake
- Original intent was for solution in 5, 7 or 10 years with no known plaintext
- Something has gone quite wrong, if no solution found after 30 years with 24 known plaintext letters
  - *“Chaocipher”* difficulty at 97 letters
  - *“Chaocipher” Cryptologia* challenge

Conclusions

Quotes and an alternative path

- “There's probably something that the author of Kryptos thought was an acceptably findable trick, but nobody working on the puzzle … has thought of it.” (Michael Hamburg, Quora.com Dec 2012)
- “There has been SO much time and effort spent on K4 without a solution, that I fear we will never get it unless we hit on EXACTLY the right encryption method.” (Lance Estes, Oct 2021, personal email)
- Calibrated path to a solution – the puzzle book “Masquerade” by Kit Williams; published Aug 1979, hint Dec 1980, solved Mar 1982.

Acknowledgements

Everyone

- Ed Hannon
- Jim Gillogly
- Doug Gwyn
- Lance Estes
- Bill Briere
- Bob Bogart
- Dan Ankuda
- Greg Materna
- Ernst Leierzopf
- Pete Ryland
- Jim Melichar
- Bernd Adameit
- Eleanor Joyner
- Seth Kintigh
- Frode Weierud
- KRYPTOS group participants

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Contact

Richard Bean

Research Fellow

R.Bean1@uq.edu.au

https://www.linkedin.com/in/richardbean