UNSCRAMBLING CODES: FROM HIEROGLYPHS TO MARKET NEWS

Emilio Barone[1] and Gaia Barone[2]

[1]Department of Economics and Finance, Luiss Guido Carli, Rome, Italy

[2]School of Business, National College of Ireland, Dublin, Ireland

ABSTRACT

This paper reviews some of the steps that paved the way for the development of sentiment analysis (or opinion mining), a technique apparently used by Jim Simons' Medallion fund for scoring an 'impossible' performance: a 66% annual average rate of return in the 31 years between 1988 and 2018. Sentiment analysis is a powerful tool that uses natural language processing (NLP), or computational linguistics, to determine whether a text about a company is positive, negative or neutral and, in a final analysis, to discover stock price patterns. Humans have always used symbols to communicate, plainly or secretively. Here we review some of the methods used in the past centuries, including Egyptians' hieroglyphs, Julius Caesar's cipher, Fibonacci's abbreviations, Leonardo da Vinci's Mirror Writing, Mary Stuart's code. The intention is to describe some passages of the long journey made by human beings to arrive at the current sophisticated IT tools for sentiment analysis.

KEYWORDS

Natural Language Processing, Hedge Funds, Cryptography

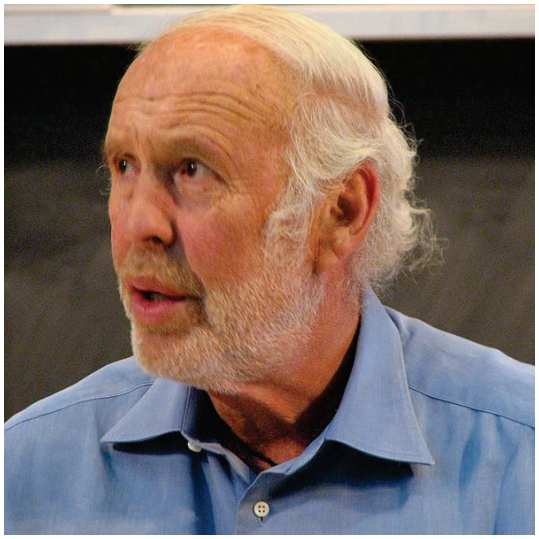
1. INTRODUCTION

What do Jean François Champollion, French archaeologist and Egyptologist, and James (Jim) Harris Simons, mathematician and hedge fund manager, have in common (Figure 1)? They have both been code breakers. Champollion was able to unscramble the message encrypted in the Rosetta stone. He had studied the Coptic language, which derived from ancient Egyptian, and this knowledge made it possible for him to understand the Demotic section of the Rosetta stone. Jim Simons received a bachelor's degree in mathematics from MIT, a PhD in mathematics from UC Berkeley and held a research position at Harvard. He later became a breaker of Russian codes at the Institute for Defense Analysis (IDA), during the Cold War.

John Hull [1] cites Jim Simons in the chapter on Natural Language Processing (NLP) of his book on *Machine Learning in Business*:

> *New data sources are becoming available all the time. One approach is to try and be one step ahead of most others in exploiting these new data sources. Another is to develop better models than those being used by others and then be very secretive about it. Renaissance Technologies, a hedge fund, provides an example of the second approach. It has been amazingly successful at using sophisticated models to understand stock price patterns. Other hedge funds have been unable to replicate its success. The average return of its flagship Medallion fund between 1988 and 2018 was 66% per year before fees. This included a return of close to 100% in 2008 when the S&P 500 lost 38.5%. Two senior executives, Robert Mercer and Peter Brown, are NLP experts and have been running the company following the retirement of the founder, Jim Simons, in 2009.* (p. 197)

DOI: 10.5121/ijnlc.2022.11603

Source: Wikipedia

Figure 1 Jean-François Champollion and James (Jim) Harris Simons.

2. MEDALLION FUND

Gregory Zuckerman [2] in his bestselling book on *The man who solved the market* reported that Simons quit Harvard in 1964 to join the Institute for Defense Analysis (IDA) "an elite research organization that hired mathematicians from top universities to assist the National Security Agency − the United States' largest and most secretive intelligence agency − in detecting and attacking Russian codes and ciphers. ... The IDA taught Simons how to develop mathematical models to discern and interpret patterns in seemingly meaningless data." (pp. 23-4).

While at IDA, Simons co-authored a paper on "Probabilistic Models for (and Prediction of) Stock Market Behavior". Detecting stock price patterns became one of his aims. However, he apparently did not care about their interpretation:

> *"I don't know why planets orbit the sun," Simons told a colleague, suggesting one needn't spend too much time figuring out why the market's patterns existed. "That doesn't mean I can't predict them."* {[2], p. 151}

Whatever techniques used by Jim Simons, his results are astonishing (Appendix A):

> *There were compelling reasons I was determined to tell Simon's story. A former math professor, Simons is arguably the most successful trader in the history of modern finance. Since 1988, Renaissance's flagship Medallion hedge fund has generated average annual returns of 66 percent, racking up trading profits of more than $100 billion (see Appendix 1 for how I arrive at these numbers). No one in the investment world comes close. Warren Buffett, George Soros, Peter Lynch, Steve Cohen, and Ray Dalio all fall short (see Appendix 2).* {[2], p. xvi}

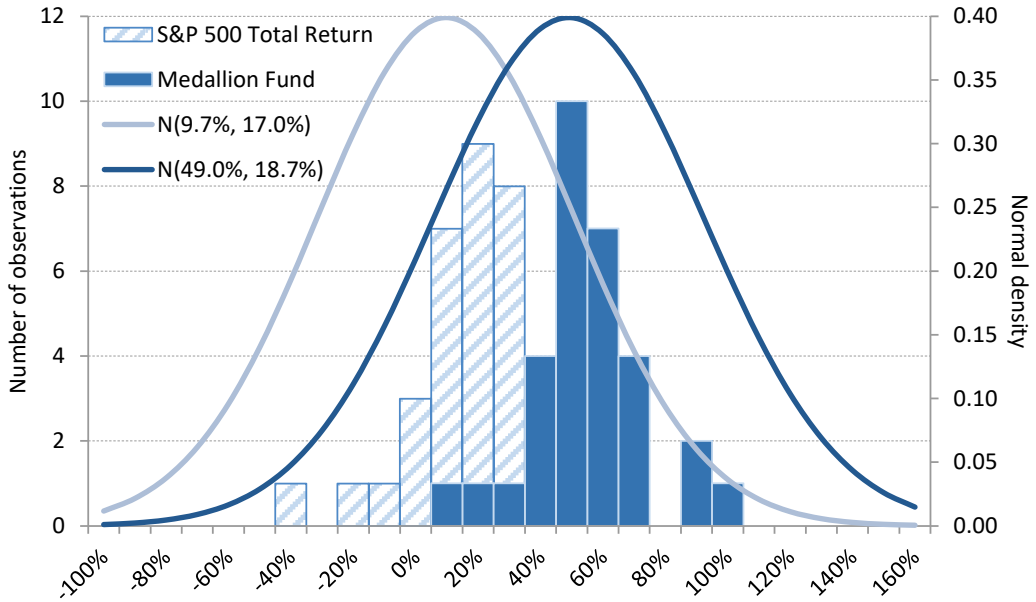
Gross log-returns consistent with the data of Appendix A are reported in Table 1, together with the correspondent data for the S&P 500 Total Return Index. The Medallion Fund (MF) outperformed the S&P 500 in 28 out of 31 years and never reported a loss. A $1 amount invested in the Medallion Fund at the end of 1987 grew to $3,995,683 after 31 years!

Figure 2 reports the performance distributions of both investments (MF and S&P 500). The average log-return, μ, of the Medallion Fund (49.0%) is 5 times greater than the average log-return of the S&P 500 (9.7%), while the standard deviations, σ, are similar (18.7% and 17.0%, respectively). It will be impossible to beat this incredible performance!

Table 1 Medallion Fund vs. S&P 500 Total Return Index: gross log-returns per year.

End of year	*Medallion Fund*	*S&P 500 Total Return*	*End of year*	*Medallion Fund*	*S&P 500 Total Return*
1988	15.1%	15.4%	2005	45.6%	4.8%
1989	1.0%	27.5%	2006	61.0%	14.7%
1990	57.5%	-3.2%	2007	86.1%	5.3%
1991	43.4%	26.6%	2008	92.5%	-46.2%
1992	38.5%	7.3%	2009	55.7%	23.5%
1993	43.1%	9.6%	2010	45.4%	14.0%
1994	66.0%	1.3%	2011	53.7%	2.1%
1995	42.5%	31.9%	2012	45.0%	14.8%
1996	36.7%	20.7%	2013	63.6%	28.1%
1997	27.4%	28.8%	2014	56.0%	12.8%
1998	45.2%	25.1%	2015	52.7%	1.4%
1999	30.5%	19.1%	2016	52.2%	11.3%
2000	82.5%	-9.5%	2017	61.7%	19.7%
2001	44.9%	-12.7%	2018	56.8%	-4.5%
2002	41.3%	-25.0%			
2003	36.5%	25.2%	μ	49.0%	9.7%
2004	40.2%	10.3%	σ	18.7%	17.0%

Note: The linear regression of Medallion Fund on S&P 500, based on normalized log-returns, gives $\alpha = 0$ (t-stat = 0) and $\beta = -0.5$ (t-stat = −3.3).



Source: Zuckerman {[[2], Appendix 1} and Yahoo Finance (^SP500TR).

Figure 2 Medallion Fund vs. S&P 500 Total Return.

2.1 The Impossible

"The Impossible" is the title of a movie based on the 2004 Indian Ocean tsunami, an extreme observation among the low-frequency high-severity natural disasters. Sometimes very rare phenomena as the S&P 500 fall of October 19, 1987 (-20.5%, from 282.70 to 224.84) and March

16, 2020 (-12.0%, from 2,711.02 to 2,386.13) are called "6-sigma" or "six standard deviation" events. Actually, if X is a standardized normal variable, the probability of $X > 6.4$ [=NORM.S.INV(1-1E-10) in Excel™] is equal to 0.0000000001, i.e. 1 out of 10 billion.

What is the probability of earning a 49.0% average annual log-return for 31 consecutive years without never reporting a loss?

2.2 Leverage and Roulette Wheel

Stellar performances are not possible without leverage. A brilliant example of leverage's effect on performance has been offered by Mark Rubinstein [3]:

> *Let's suppose, for the sake of this discussion, that the roulette wheel only has numbers one through 36 on it, and that it doesn't have a zero or a double zero. If you place one chip on each of the numbers from one through 35, you will win a small amount most of the time. Now suppose you borrowed most of the money to finance the 35 chips you bet. You'd find that most of the time you'd earn a high rate of return on your capital, but that, on average, one out of 36 times you would lose it all. ... The general problem with one–35 and 36 is one of evaluating investment strategies. A lesson we can draw is that when we look at the performance of an investment manager, we now have to ask ourselves, Is this a strategy that wins 35 out of 36 times, but really blows it the 36th time? Because of this problem, it's easy to be deceived by a manager's track record. (pp. 3-4)*

2.3 Natural Language Processing

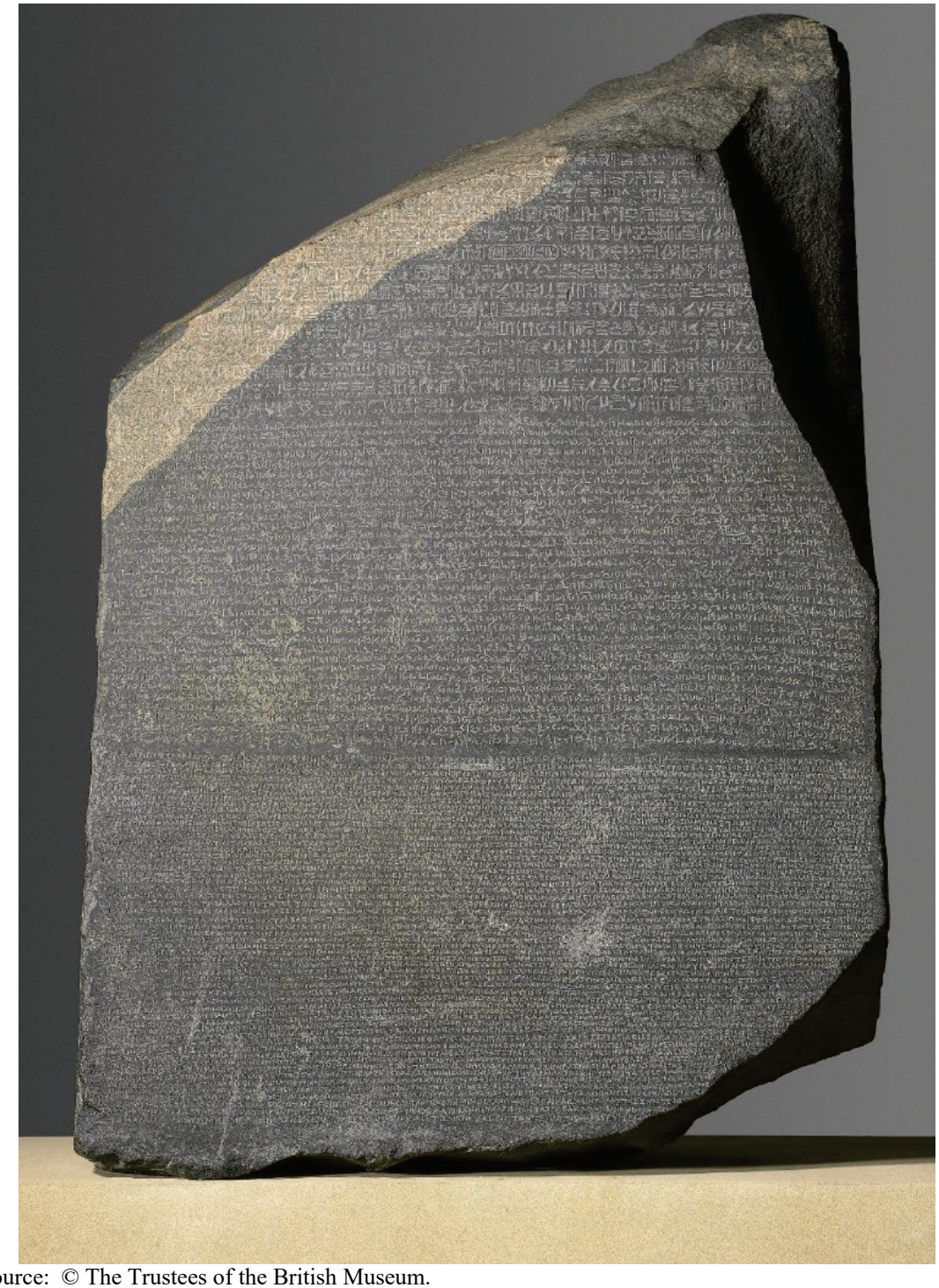
Our *a priori* is that there is much more than "good luck" behind the stellar performance of Medallion Fund: best human capital, massive usage of computer power, and natural language processing (NLP). Robert Mercer and Peter Brown, who joined forces with Jim Simons in 1993, were both working at IBM "searching for ways to get computers to do a better job transcribing speech into text and even translate language" {Zuckerman [2], p. 4}. The search for a common language, or binarisation, was the ultimate goal. Probably, natural language processing (NLP) or computational linguistics has been the key breakthrough that helps to explain most of MF's performance. Much of the *big data* generated in the world is in the form of written or spoken words. Quotes and volumes are basic statistics that summarize a great deal of information, but market news is just as important. The rules of language are difficult to communicate to a machine and words can have several meanings. To crack the market's code was not an easy job.

In the text that follows we will describe some passages of the long journey made by human beings to communicate with each other, in a manifest or secret way. We will deal with Egyptian Hieroglyphs (§3), Julius Caesar's cipher (§4), Fibonacci's Liber Abaci (§5), Leonardo da Vinci's Mirror Writing (§6), Mary Stuart's Code (§6), Monte Carlo Descrambling (§8), Shazam and Google Lens (§9), Kaggle (§10). Then we will conclude (§11).

3. EGYPTIAN HIEROGLYPHS

3.1 Rosetta Stone

The Rosetta Stone is a black granite stele inscribed with three versions of a decree issued in Memphis, Egypt, in 196 BC during the Ptolemaic dynasty on behalf of King Ptolemy V Epiphanes. It was discovered in July 1799 by French officer Pierre-François Bouchard during the Napoleonic campaign in Egypt, while digging the foundations of an addition to a fort in the Nile Delta, near el-Rashid (Italianized in Rosetta, which means "little rose"). It is an irregularly shaped grey and pink granite stone [3 feet 9 inches (114 cm) long and 2 feet 4.5 inches (72 cm) wide], exhibited in the British Museum since 1802 (Figure 3).

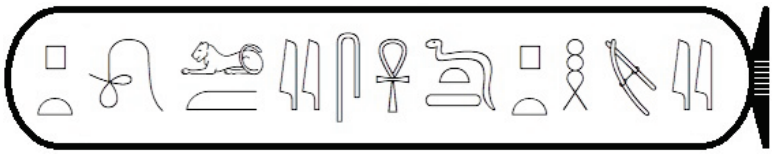


Source: © The Trustees of the British Museum.

Figure 3 Rosetta Stone.

The Rosetta Stone bears a priestly decree concerning Ptolemy V in three blocks of text: Hieroglyphic (the "language of the gods", 14 lines), Demotic (the "language of acts", 32 lines) and Ancient Greek (53 lines). Translation of the Ancient Greek was relatively easy. Translation of Demotic (the ancient Egyptian script preceding Coptic) was harder, but the proper names (Ptolemy, Alexander and Alexandria) were quickly deciphered. Translation of hieroglyphic text was even harder, but it was soon established that names of kings or pharaohs were contained within elongated ovals ("cartouches").

(a) original from right to left

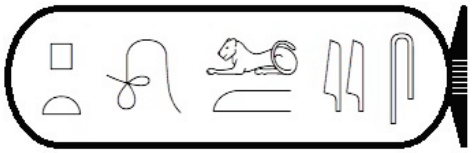


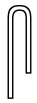
(b) transcribed from left to right

Source: British Museum (original).

Figure 4 The name Ptolemy in the Rosetta Stone.

There were six cartouches on the Rosetta Stone. According to the Greek translation, these cartouches clearly had to contain the name Ptolemy (Ptolemaios, in Greek). Three of them looked as in Figure 4 and featured the name Ptolemy along with an Egyptian honorific: "Ptolemy, living for ever, beloved of Ptah" {Robinson [4], p. 125}. The other three looked like this:



Hieroglyph							
Champollion's reading	P	T	O	L	M	E	S

The hieroglyphs of ancient Egypt are one of the earliest forms of writing. They can be read from right to left or left to right. You can distinguish the direction in which the text is to be read because the human or animal figures always face towards the beginning of the line. Also the upper symbols are read before the lower. There are no spaces between words, line breaks or punctuation.

3.2 Google Translator

On July 15, 2020 (exactly 221 years after the Rosetta Stone was discovered) Google has launched a hieroglyphics translator that uses machine learning to decode ancient Egyptian language: https://artsexperiments.withgoogle.com/fabricius/en (Table 2).

Google created tools and models for the three phases of the algorithm:

1) *Extraction* - taking hieroglyphic script and sequences from source images and creating workable facsimiles;
2) *Classification* - training a neural network to correctly identify over 1000 hieroglyphs;
3) *Translation* - matching sequences and blocks of text to available dictionaries and published translations.

Table 2 Ancient Egypt Hieroglyphs.

A E O	A	B	CH	D	F PH V
					
vulture 1313F	forearm 1309D	foot 130C0	hobble rope 1337F	hand 130A7	horned viper 13191
G	**H**	**H**	**I Y**	**J (G)**	**K (C)**
					
pot stand 133BC	rope 1339B	shelter 13254	reed leaf 131CB	cobra 13193	basket 133A1
K (C)	**L R**	**M**	**N**	**O U W**	**P**
					
hillside 133D8	open mouth 1308B	owl 13153	water 13216	quail chick 13171	stool 132AA
S (C)	**SH**	**T**	**TH**	**TH**	**Z**
					
folded cloth 132F4	lake 13219	bread loaf 133CF	(unknown) 1340D	cow's belly 13121	door bolt 13283

Note: 1,079 hieroglyphs are available at https://fonts.google.com/noto/specimen/Noto+Sans+Egyptian+Hieroglyphs.

4. JULIUS CAESAR'S CIPHER

Julius Caesar (12 July 100 BC - 15 March 44 BC) was well ahead of his time. At a time when illiteracy was the norm, he preferred to communicate with his generals by means of a secret code, Caesar's cipher. Suetonius, the private secretary to the Emperor Hadrian, tells us that Julius Caesar used a mono-alphabetical cipher where each letter was replaced by the letter that followed it by three places in the alphabet, for his confidential correspondence:

Extant et ad Ciceronem item ad familiares domesticis de rebus in quibus si qua occultius perferenda erant per notas scripsit id est sic structo litterarum ordine ut nullum verbum effici posset: quae si quis investigare et persequi volet quartam elementorum litteram id est D pro A et perinde reliquas commutet.	*There are also letters of his to Cicero, as well as to his intimates on private affairs, and in the latter, if he had anything confidential to say, he wrote it in cipher, that is, by so changing the order of the letters of the alphabet, that not a word could be made out. If anyone wishes to decipher these, and get at their meaning, he must substitute the fourth letter of the alphabet, namely D, for A, and so with the others.*

Source: Suetonius [5], *De vita Caesarum*, Ch. 1, §56.

A similar cipher was used by the Emperor Augustus (23 September 63 BC - 19 August AD 14):

Orthographiam, id est formulam rationemque scribendi a grammaticis institutam, non adeo custodit ac videtur eorum potius sequi opinionem, qui perinde scribendum ac loquamur existiment. Nam quod saepe non litteras modo sed syllabas aut permutat aut praeterit, communis hominum error est.	*He does not strictly comply with orthography, that is the say the theoretical rules of spelling laid down by the grammarians, seeming to be rather of the mind of those who believe that we should spell exactly as we pronounce. Of course his frequent transposition or omission of syllables as well as of letters are slips common to all mankind.*
Nec ego id notarem, nisi mihi mirum videtur tradidisse aliquos, legato eum consulari successorem dedisse ut rudi et indocto, cuius manu "ixi" pro ipsi scriptum animadverterit.	*I should not have noted this, did it not seem to me surprising that some have written that he cashiered a consular governor, as an uncultivated and ignorant fellow, because he observed that he had written ixi for ipsi.*
Quotiens autem per notas scribit, B pro A, C pro B ac deinceps eadem ratione sequentis litteras ponit; pro X autem duplex A.	*Whenever he wrote in cipher, he wrote B for A, C for B, and the rest of the letters on the same principle, using AA for X.*

Source: Suetonius [5], *De vita Caesarum*, Ch. 2, §88.

Figure 5 shows the Alberti cipher disk, an enciphering and deciphering tool developed in 1470 by Leon Battista Alberti. The device consists of two concentric circular plates mounted one on top of the other.

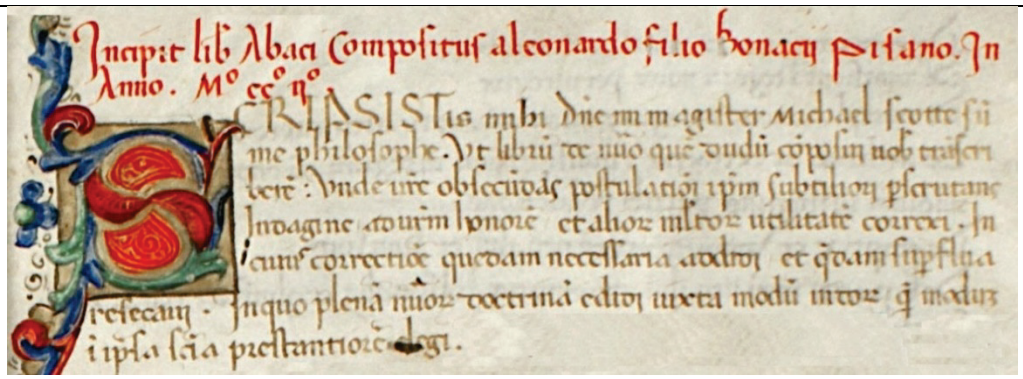
Source: Wikipedia.

Figure 5 Alberti cipher disk.

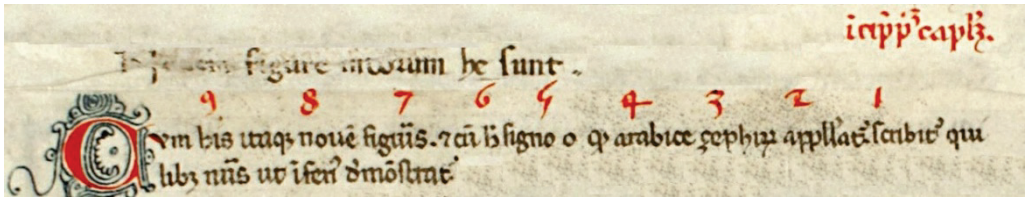
5. FIBONACCI'S LIBER ABACI

Before Gutenberg invented the printing press, around 1440, texts were handwritten. Even if there is no secret code to decrypt, it is sometimes hard to "translate" characters in machine-readable text. This is the case of Liber Abaci, the book written in 1202 by Leonardo Pisano, known today to mathematicians and scientists all over the world by the name Fibonacci (Table 3).

Table 3 Fibonacci's Liber Abaci.



...



Manuscript| Liber Abaci, Leonardo Pisano, 1202 | Codice Magliabechiano, C. I, 2616, Badia Fiorentina, n. 73.

Incipit liber Abaci Compositus a leonardo filio Bonacij Pisano In Anno. M° cc° ij.°

SCRIPSISTis mihi domine mi magister Michael Scotte, summe philosophe, vt librum de numero, quem dudum composui, uobis transcriberem: vnde uestrae obsecundans postulationi, ipsum subtiliori perscrutans Indagine ad uestrum honorem et aliorum multorum utilitatem correxi. In cuius correctione quedam necessaria addidj, et quedam superflua resecaui. In quo plenam numerorum doctrinam edidj, iuxta modum indorum, quem modum in ipsa scientia prestantiorem elegi.

...

Incipit primum capitulum. Nouem figure indorum he sunt

9 8 7 6 5 4 3 2 1 .

Cvm his itaque nouem figuris, et cum hoc signo 0, quod arabice zephirum appellatur, scribitur quilibet numerus, ut inferius demonstratur.

"Decrypted" Latin | Boncompagni, B. [6], *Il Liber Abbaci di Leonardo Pisano* | Rome, 1857.

Here begins the Book of Calculation Composed by Leonardo Pisano, Family Bonaci, In the Year 1202.

You, my Master Michael Scott, most great philosopher, wrote to my Lord about the book on numbers which some time ago I composed and transcribed to you; whence complying with your criticism, your more subtle examining circumspection, to the honor of you and many others I with advantage corrected this work. In this rectification I added certain necessities, and I deleted certain superfluities. In it I presented a full instruction on numbers close to the method of the Indians, whose outstanding method I chose for this science.

...

Here Begins the First Chapter. The nine Indian figures are:

9 8 7 6 5 4 3 2 1.

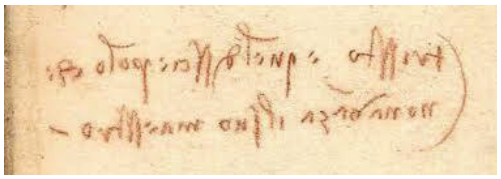
With these nine figures, and with the sign 0 which the Arabs call zephir any number whatsoever is written, as is demonstrated below.

English | Sigler, L. E. [7], *Fibonacci's Liber Abaci,* pp. 15 and 17, 2002.

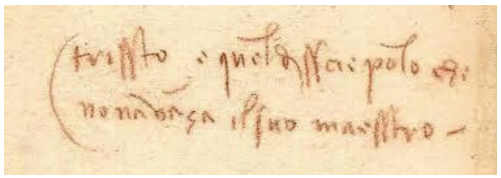
6. LEONARDO DA VINCI'S MIRROR WRITING

Leonardo da Vinci (15 April 1452 – 2 May 1519) wrote backward, from right to left, most of his personal notes, in an attempt to keep them illegible to anyone other than him. For instance, he wrote:

[Leonardo, in Italian, encrypted]	[Leonardo, in Italian, decrypted]
ehc olopecsid leuq 'e otsirt _ ortseam ous li aznava non	(tristo e' quel discepolo che non avanza il suo maestro _



flipped to



[Text in English, encrypted]	[Text in English, decrypted]
taht lipup eht si roop _ retsam sih ssaprus ton seod	(poor is the pupil that does not surpass his master _

Source: 1493, Forster III, South Kensington Museum III 24b, http://t.co/ON3xNdrOhb.

7. MARY STUART CODE

On February 8, 1587, Mary Queen of Scots was beheaded for high treason after nineteen years of imprisonment in England. She was found guilty of her involvement in the Babington plot, a conspiracy to assassinate Elizabeth I organized by Anthony Babington. Mary's encoded correspondence had been decrypted {Singh, [8]}. The alphabet used is shown in Table 4.

Table 4 Mary Stuart Code | Alphabet

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**	**and**	**for**	**but**	**with**	**that**	**if**	**as**
										
of	**the**	**by**	**so**	**not**	**when**	**from**	**this**	**is**	**in**	**say**
										
me	**my**	**I**	**you**	**what**	**where**	**which**	**there**	**send**	**receive**	**pray**

Note: The alphabet contains 5 blank characters that can be used as space or word separator (this makes the frequency analysis more complex):

Source: https://www.dcode.fr/mary-stuart-code.

8. MONTE CARLO DESCRAMBLING

The first step to decrypt a code is to guess the text's language. Then a frequency analysis can be applied. The current version of Google's translator, launched in April 2006, can recognize 108 languages, from Afrikaans to Xhosa, Yiddish, Yoruba, and Zulu {WU, Y., *et al.* [10]}.

The statistical distribution of letters in the English language, derived from some classic books, is reported in Table 5.

Table 5 Statistical Distributions of English Text: Letter Frequencies.

a	6.51738%	h	4.92888%	o	5.96302%	v	0.82903%
b	1.24248%	i	5.58094%	p	1.37645%	w	1.71272%
c	2.17339%	j	0.09033%	q	0.08606%	x	0.13692%
d	3.49835%	k	0.50529%	r	4.97563%	y	1.45984%
e	10.41442%	l	3.31490%	s	5.15760%	z	0.07836%
f	1.97881%	m	2.02124%	t	7.29357%	Space	19.18182%
g	1.58610%	n	5.64513%	u	2.25134%		100.00000%

Source: Art Owen [11], Stat 362 (Monte Carlo Methods), Stanford University, Fall 2008.

Note: The above statistics are derived from these classics (available at www.literature.org): Origin of Species (Charles Darwin), The Voyage of the Beagle (Charles Darwin), Jane Eyre (Charlotte Bronte), Wuthering Heights (Emily Bronte), Tarzan of the Apes (Edgar Rice Burroughs), The Return of Tarzan (Edgar Rice Burroughs), Paradise Lost (John Milton). All upper-case letters were converted to lower-case. All numbers and all special characters were removed. A carriage return is treated as a space. There are exactly 5,086,936 characters in these files.

The Monte Carlo approach also requires the probabilities of adjacent letters, i.e. the (first-order) letter transition probabilities:

$$\underset{i\to j}{\text{Prob}} \equiv \text{Prob}(x_{\ell+1} = j \mid x_{\ell} = i) \text{ for } 1 \le i, j \le 27.$$

A real-life example of Monte Carlo descrambling is shown by Persi Diaconis {[9], pp. 1-3}:

> *One day, a psychologist from the state prison system showed up with a collection of coded messages. ... The problem was to decode these messages. Marc [Coram] guessed that the code was a simple substitution cipher, each symbol standing for a letter, number, punctuation mark or space. ... I like this example because a) it is real, b) there is no question the [Monte Carlo] algorithm found the correct answer, and c) the procedure works despite the implausible underlying assumptions. In fact, the message is in a mix of English, Spanish and prison jargon. The plausibility measure is based on first-order transitions only. A preliminary attempt with single-letter frequencies failed.*

9. SHAZAM AND GOOGLE LENS

Binarisation is the preprocessing technique used to transform characters, sounds and images in binary digits (0, 1). A great part of all the available information is binarised. Optical character recognition (OCR) has made the electronic conversion of typed, handwritten or printed text into machine-encoded text possible. Optical music recognition (OMR) has the same aim, but is based on printed sheet music. Shazam, owned by Apple, uses the microphone on the device to identify music, movies, advertising, and television shows, while Google Lens identifies images in a few instants (Table 6).

Table 6 Music and Image Recognition: Shazam and Google Lens.

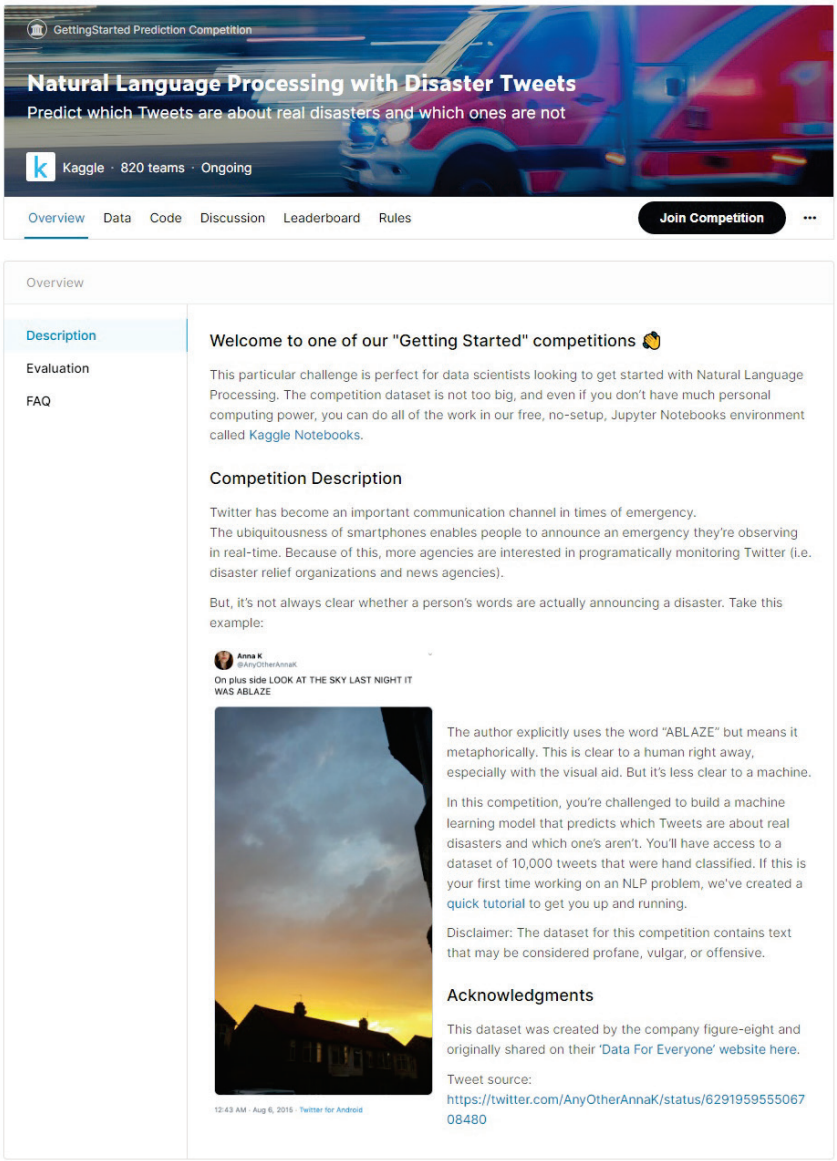
Source: YouTube (https://youtu.be/ck57LnYScNQ), Shazam (Burt Bacharach, Live in Sidney, 2008) and Google Lens.

10. KAGGLE

A huge amount of market news available on social media and elsewhere needs to be worked out to discover if they will have a positive, negative, or neutral effect on a particular company. A good starting point to learn Natural Language Processing (NLP) is to participate in a competition organized by Kaggle, a subsidiary of Google. One of the Kaggle's "Getting Started" competitions is titled "Natural Language Processing with Disaster Tweets: Predict which Tweets are about real disasters and which ones are not." All of the work can be done in Kaggle's free, no-setup, Jupyter Notebooks environment, where you can run Python code.

The competition is ongoing. Table 7 shows a few steps of the current leader's code.

Table 7 Kaggle's Competition: Natural Language Processing with Disaster Tweets.



A few steps of the code used by the current leader (Shahules).

Importing required Libraries.

Using TensorFlow backend. [TensorFlow: an open-source machine learning platform developed by Google]

Loading the data and getting basic idea [Training set: 7613 rows and 5 columns; Test set: 3263 rows and 4 columns]

Class distribution [There are more tweets with class 0 (no disaster) than class 1 (disaster tweets)]

Exploratory Data Analysis of tweets [at character level, word level and sentence level]

Number of characters in tweets, number of words in a tweet, average word length in a tweet, common stopwords in tweets [stopwords: words that are so commonly used that they carry very little useful information (the, in, of, a, to, and, on, for, is, at, ...)], analyzing punctuations [=> - : ? . | + # !]) ' ; / = @ ~ * ...], Ngram analysis [Ngram: a sequence of N words]

Data Cleaning [Removing URLs, HTML tags, Emojis, Punctuations. Spelling Correction]

GloVe for Vectorization [GloVe: an unsupervised learning algorithm for obtaining vector representations for words.]

Number of unique words: 20342

Baseline Model

Source: www.kaggle.com/c/nlp-getting-started/overview, www.kaggle.com/shahules/basic-eda-cleaning-and-glove.

11. FINAL REMARKS

The average annual log return of Medallion Fund, the hedge fund of Jim Simons' Renaissance Technologies, has been 49.0% in the 31 years from 1988 to 2018. This stellar performance is in strong contrast to the Efficient Market Hypothesis (EMH). If markets were informationally efficient, none could perform as Medallion Fund (MF). So, what was the MF's key to success? It is not enough to process numerical data, such as stock prices and volumes. The MF's key to success seems to be Natural Language Processing (NLP):

> *"There are patterns in the market," Simons told a colleague, "I know we can find them."* {[2], p. 5}.

The daunting goal of NLP is to successfully perform human like language processing. In order to make investment/disinvestment decisions, people rely on news reports about companies, on quarterly earnings calls to analysts, on opinions posted on the web, etc. Baker and Wurgler [12] were among the first researchers to study how investor sentiment affects the cross-section of stock returns. Zhang and Skiena [13] exploited blog and news sentiment to construct market neutral portfolios that produced impressive returns. We believe that information retrieval techniques, as those shown in John and Govilkar [14], will make the difference between failure or success in the world of investing.

REFERENCES

[1] Hull, J. C., *Machine Learning in Business: An Introduction to the World of Data Science*, 3rd ed., KDP, May 26th, 2021.

[2] Zuckerman, G., *The man who solved the market: How Jim Simons launched the quant revolution*, Penguin Random House, 2019.

[3] Rubinstein, M., "The World According to Mark Rubinstein: Interview", *Derivatives Strategy Magazine*, July 1999.

[4] Robinson, A., "Cracking the Egyptian Code: The Revolutionary Life of Jean-François Champollion", Thames & Hudson, 2018.

[5] Suetonius, G., *The Twelve Caesars: The Lives of the Roman Emperors*, J. C. Rolfe (Trans.). St. Petersburg, FL: Red and Black Publications, 2008.

[6] Boncompagni, B., *Il Liber Abbaci di Leonardo Pisano*, Rome, 1857.

[7] Leonardo da Pisa, *Liber Abaci*, 1202. See SIGLER, L. E., *Fibonacci's Liber Abaci - A Translation into Modern English of Leonardo Pisano's Book of Calculation*, Springer, 2002.

[8] Singh, S., *The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography*, Fourth Estate, 1999.

[9] Diaconis, P., "The Markov Chain Monte Carlo Revolution," *Bulletin of the American Mathematical Society*, 46(2), 179-205, 2009.

[10] Wu, Y., *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," https://arxiv.org/pdf/1609.08144.pdf (2016).

[11] Owen, A., "Monte Carlo theory, methods and examples," https://artowen.su.domains/mc/.

[12] Baker, M., and Wurgler, J., "Investor Sentiment and the Cross-Section of Stock Returns", Journal of Finance, Vol. 61, No. 4, 2006.

[13] Zhang, W., and Skiena, S., "Trading Strategies to Exploit Blog and News Sentiment", *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[14] John, R., and Govilkar, S., "Information Retrieval Technique for Web Using NLP", *International Journal on Natural Language Computing*, Vol. 6, No. 5, 2017.

AUTHORS

Emilio Barone is a Stephen A. Ross Professor of Financial Economics, Intesa Sanpaolo Chair, in the Department of Economics and Finance, Luiss Guido Carli University of Rome. He has been an Executive Director at Intesa Sanpaolo, Head of the Financial Risks Analysis Department and Head of Research at Istituto Mobiliare Italiano (I.M.I.). He has previously worked as an economist in the Research Department of the Bank of Italy. He has followed postgraduate studies at Yale University after receiving his B.S./M.S. in Economics from "La Sapienza" University of Rome. He is the author of many articles on derivatives pricing and risk measurement / management.

Gaia Barone is an Assistant Professor in "Economics and Finance" at the School of Business at National College of Ireland. She is Program Director for the M.Sc. in Finance. Before September 2018, she was an Assistant Professor in "Mathematical Methods for Economics", and Chair of "Financial Mathematics" and "Quantitative Methods for Management" at Luiss Guido Carli University of Rome. In July 2011 she received her Ph.D. in Money and Finance from "Tor Vergata" University of Rome and in June 2009 her M.Sc. in Financial Mathematics from Stanford University. She has published two books on arbitrages and articles on derivatives and credit risk.

APPENDIX A

THE MAN WHO SOLVED THE MARKET BY GREGORY ZUCKERMAN

Appendix 1

	Net Returns	*Management Fee**	*Performance Fee*	*Returns Before Fees*	*Size of Fund*	*Medallion Trading Profits***
1988	9.0%	5%	20%	16.3%	$20 million	$3 million
1989	-4.0%	5%	20%	1.0%	$20 million	$0 million
1990	55.0%	5%	20%	77.8%	$30 million	$23 million
1991	39.4%	5%	20%	54.3%	$42 million	$23 million
1992	33.6%	5%	20%	47.0%	$74 million	$35 million
1993	39.1%	5%	20%	53.9%	$122 million	$66 million
1994	70.7%	5%	20%	93.4%	$276 million	$258 million
1995	38.3%	5%	20%	52.9%	$462 million	$244 million
1996	31.5%	5%	20%	44.4%	$637 million	$283 million
1997	21.2%	5%	20%	31.5%	$829 million	$261 million
1998	41.7%	5%	20%	57.1%	$1.1 billion	$628 million
1999	24.5%	5%	20%	35.6%	$1.54 billion	$549 million
2000	98.5%	5%	20%	128.1%	$1.9 billion	$2,434 million
2001	33.0%	5%	36%	56.6%	$3.8 billion	$2.149 million
2002	25.8%	5%	44%	51.1%	$6.24 billion	$2.676 billion
2003	21.9%	5%	44%	44.1%	$5.09 billion	$2.245 billion
2004	24.9%	5%	44%	49.5%	$5.2 billion	$2.572 billion
2005	29.5%	5%	44%	57.7%	$5.2 billion	$2.999 billion
2006	44.3%	5%	44%	84.1%	$5.2 billion	$4.374 billion
2007	73.7%	5%	44%	136.6%	$6.2 billion	$7.104 billion
2008	82.4%	5%	44%	152.1%	$5.2 billion	$7.911 billion
2009	39.0%	5%	44%	74.6%	$5.2 billion	$3.881 billion
2010	29.4%	5%	44%	57.5%	$10 billion	$5.750 billion
2011	37.0%	5%	44%	71.1%	$10 billion	$7.107 billion
2012	29.0%	5%	44%	56.8%	$10 billion	$5.679 billion
2013	46.9%	5%	44%	88.8%	$10 billion	$8.875 billion
2014	39.2%	5%	44%	75.0%	$9.5 billion	$7.125 billion
2015	36.0%	5%	44%	69.3%	$9.5 billion	$6.582 billion
2016	35.6%	5%	44%	68.6%	$9.5 billion	$6.514 billion
2017	45.0%	5%	44%	85.4%	$10 billion	$8.536 billion
2018	40.0%	5%	44%	76.4%	$10 billion	$7.643 billion
	39.1% average net returns			66.1% average returns before fees		$104.530.000.000 total trading profits

* Fees are charged by the Medallion fund to its investors, which in most years represents the firm's own employees and former employees.

** Gross returns and Medallion profits are estimates − the actual number could vary slightly depending on when the annual asset fee is charged, among other things. Medallion's profits are before the fund's various expenses.

Average Annual Returns: 66.1% gross, 39.1% net

The above profits of $104.5 billion represent those of the Medallion fund. Renaissance also profits from three hedge funds available to outside investors, which managed approximately $55 billion as of April 30, 2019. (*Source*: Medallion annual reports; investors)

Appendix 2

Returns Comparison

Investor	*Key Fund/Vehicle*	*Period*	*Annualized Returns**
Jim Simons	Medallion Fund	1988-2018	39.1%
George Soros	Quantum Fund	1969-2000	32.0%†
Steven Cohen	SAC	1992-2003	30.0%
Peter Lynch	Magellan Fund	1977-1990	29.0%
Warren Buffett	Berkshire Hathaway	1965-2018	20.5%‡
Ray Dalio	Pure Alpha	1991-2018	12.0%

* All returns are after fees.
† Returns have fallen in recent years as Soros has stopped investing money for others.
‡ Buffett averaged 62% gains investing his personal money from 1961 to 1957, starting with less than $10,000, and saw average gains of 24.3$% for a partnership managed from 1957 to 1969.

Appendix 2 does not report the performance of Princeton/Newport Partners, the hedge fund of Edward Thorp. However, Zuckerman mentions it elsewhere {[2], pp. 127-9}:

> *During the 1970s, Thorp helped lead a hedge fund, Princeton/Newport Partners, recording strong gains and attracting well-known investors ... by the late 1980s, Thorp's fund stood at nearly $300 million, dwarfing the $25 million Simons's Medallion fund was managing at the time ... Over its nineteen-year existence, the hedge fund featured annual gains averaging more than 15 percent (after charging investors various fees), topping the market's returns over that span.*