Stevens, Hallam (2018) ***Hans Peter Luhn and the birth of the hashing algorithm.*** **IEEE Spectrum, 55 (2) pp. 44-49.**

Hans Luhn and the Birth of Hashing Algorithms

[Accepted version]

IEEE Spectrum, 30th January 2018.
https://spectrum.ieee.org/hans-peter-luhn-and-the-birth-of-the-hashing-algorithm

Hallam Stevens

At the International Conference for Scientific Information in Washington, D.C. in November 1958, the inventor Hans Peter Luhn, an employee of IBM, demonstrated a series of his electro-mechanical devices. The machines *looked* rather ordinary [**figure**]. Much like many other clunky computing devices of the time, they were designed to operate on punched cards, scooping and sorting them into slots and bins.

Unlike most other computers, however, Luhn's devices were not designed to work with numbers and calculations, but rather with words and sentences. One machine, for instance, implemented a scheme Luhn called KWIC: Key Words In Context [**sidebar**]. It could quickly and automatically construct a kind of index of a large set of texts. KWIC resulted widespread recognition for Luhn, with newspapers across the United States reporting on his invention. Extracting keywords was usually a painstaking process that required human eyes and brains. With the amount of information in many fields growing too fast for most people to keep up, means for abstracting and summarizing were in high demand.

We now take for granted the fact that computers can deal in information, providing us with restaurant reviews, sports scores, or stock prices at a click. But computers were not always "information machines." Luhn's attempts to manipulate texts marks the beginning of a new way of thinking about computers.

But his ideas, in the form of the “hash,” also underpin modern-day algorithms that we use for everything from online shopping to automatic translation.

By the early 1960s, KWIC became central to the design of hundreds of computerized indexing systems including those used by the Chemical Abstracts Service, Biological Abstracts, and the Institute for Scientific Information. Luhn also developed an “intelligence system” for businesses, able to deliver relevant information to specific individuals within large organizations. KWIC was the 1950s equivalent of a search engine: it allowed users to rapidly locate the information they needed. But it also held the intellectual key to a different way of thinking about what “electronic brains” could achieve.

The 1950s were formative years in the development of electronic computers. The first modern electronic computers were developed during World War II and put to work crunching numbers for ballistics and atomic weapons. In the postwar years, cold war tensions ensured ample funding for a wide variety of military-related hardware, including the development of electronic computers. Computers were being made larger, faster, and more accurate. But their main uses – crunching and storing numbers – changed little.

Within this nascent computer world, Luhn cut a somewhat unusual figure. His many inventions [**sidebar**] seem to belong to another, perhaps quainter, pre-digital era of punched cards, mechanical calculators, and slides rules. Tall and elegantly dressed in a dark suit and thick-framed glasses, Luhn knew more about the textiles than computer science [**figure**]. Even in the 1950s, Luhn’s electro-mechanical devices were rapidly being outmoded by digital computers. Nevertheless, his ideas, transformed and remixed for a variety of purposes, are now embedded in almost every kind of software we can think of.

Luhn was born in Barman – one of the towns that later merged to become Wuppertal – in Germany in 1896. His father Johann was a master printer and the younger Luhn was sent to Switzerland to learn the family trade. After his printing career was disrupted by a stint in the German army in World War I, he ended up in the textile trade. Luhn came to New York in 1924 as an agent for a German firm. Even in textiles, Luhn's inventive bent soon became apparent. In 1927, he developed a ruler-like device that could be used to gauge the threadcount of cloth. The "Lunometer" is still in use [**figure**].

Luhn absorbed information from a wide variety of fields, becoming a proficient mountain climber, a gourmet cook, and an expert painter, in addition to his engineering prowess. During the 1930s, Luhn patented a remarkable range of devices including a foldable raincoat, a device for shaping women's stockings, a game table, and a recipe guide for cocktails (called the Cocktail Oracle) that told the user what could be made from the ingredients on hand [**figure**]. But, perhaps because of his early introduction to the printing trade, Luhn's real interest was in the storage, communication, and retrieval of information, especially textual information. It was largely in this capacity that he joined IBM in 1941. Given simply the title of "inventor," Luhn had a wide scope to tackle whatever problems he liked. He was prolific – he ended up holding seventy patents for IBM – but many of his inventions focused on problems of using machines, including new electronic computers, for manipulating information.

In 1946 and 1947, for example, Luhn began to work on making machine-readable typewritten documents. In one device, a metallic ribbon was inserted into a typewriter, punching magnetic patterns onto paper that could be scanned by machine. In the early 1950s, the sciences experienced a massive explosion of information output, mostly in the form of published papers. Many worried that "information overload" would prevent scientists from doing their work effectively. This concern led Vannevar Bush, one of the leaders of America's massive wartime scientific bureaucracy, to propose a desk-sized electromechanical device he called the "Memex" for storing and linking together information.

In the early 1950s, Luhn began to work with two chemists at MIT, Malcolm Dyson and James Perry, on his own device. This was a system for sorting information about chemical compounds using punched cards. Each punch card was encoded with information about a specific compound. The user could then insert a “question card” into the machine; this card specified a set of criteria against which all the compound cards could be compared and sorted. At least for chemical information, Luhn devised a way for automatic searching.

On the 6th of January 1954, Luhn filed a US patent application for a “Computer for Verifying Numbers.” This handheld mechanical device would hardly pass for something we would call a computer today. Nevertheless, its gears and wheels provided the foundation for one of the most important algorithms of the digital age [**figure**].

Luhn’s “computer” aimed to solve a simple practical problem. In the early 1950s, various kinds of identification numbers – such as credit card numbers and social security numbers – were beginning to play a more important role in many aspects of public and private life. Numbers were convenient ways of managing people and information. But they also posed problems: they could be difficult to remember, they could be transcribed incorrectly, or deliberately falsified. What was needed was a means of quickly verifying whether a number – a credit card number provided by a customer for instance – was valid [**sidebar**].

The algorithm Luhn developed for this purpose, now known also as the “modulus 10” algorithm, is still utilized in a wide variety of contexts. For instance, IMEI numbers assigned to cellular phones are verified using a similar “checksum” method. More significantly, Luhn’s machine was the forerunner of the wide class of algorithms known as “hashes.” Hashing is a powerful means of organizing information so that it is easy for a computer to find. Like a culinary “hash,” a hash algorithm chops and mixes up data in various ways. Counterintuitively, such mixing – cleverly deployed – can actually speed up many kinds of computer operations.

In early 1953, Luhn had written an internal IBM memo in which he suggested putting information into "buckets" in order speed up a search. Let's say you wished to look up a telephone number in a large "white pages" database. Given a ten digit number such as 314-159-2652, a computer could simply search through the list until it found the relevant entry. With a large database, though, this could take a long time. Luhn's suggestion was to group digits into pairs (31, 41, 59, 26, 52), add those paired digits together (4, 5, 14, 8, 7), and then take only the last digit of any double digit number (leaving 45487). The original number (and the entry corresponding to it, such as a name or address) would then be put into a bucket labeled "45487." When given a telephone number to look up, its bucket number could quickly be calculated using Luhn's procedure. Although there might be more than one entry in each bucket, sequentially searching through a single bucket would be much faster than searching the whole list from beginning to end [**figure**].

Since the 1950s, computer scientists have improved on Luhn's simple procedure, devising different kinds of hash functions that are suitable for a range of different purposes. But the basic idea is the same: use a math problem to organize data into easily searchable buckets. Because organizing and search for data are such widespread problems in computing, hashing algorithms have been crucial to the development of cryptography, graphics, telecommunications, and biology. Every time you send a credit card number over the Web or even use your word processor's dictionary, hash functions are at work.

In the 1950s, of course, many of these applications were not even conceivable. What led Luhn to the solution of a problem that didn't yet exist?

Luhn was concerned with different kinds of problems from most of his contemporaries. Luhn saw his computers as sophisticated word processors, able to organize information in ways could solve practical problems in science and business. By deploying computers for reading, indexing, and understanding, Luhn connected machines to a tradition of textual analysis. These were the same kinds of problems that people in all sorts of fields – from philology to library science – had been trying to solve for a long time.

During the time that Luhn was working out his hashing scheme for checking numbers, he was also involved in a developing a range of more and more sophisticated machines for searching and organizing textual information. By 1958, his chemical card sorter had become the Universal Card Scanner and the 9900 Special Index Analyzer, which he demonstrated at the International Conference for Scientific Information. These were electromechanical devices that could search and sort punched cards according to logical criteria specified by the user.

What really caused a stir, though, was Luhn's computerized method of constructing concordances. Concordances are alphabetic lists of the most important words used in a book or a collection of writings, have a long history in theology and philology. However, in the era before full-text computerized search was available, constructing a concordance was extraordinarily time consuming and difficult and generally reserved only for the most important of works such as the Bible or the works of Shakespeare.

What Luhn's "bucket" scheme did for numbers, his concordance system did for texts. Both were ways of making an index of a large body of information in order to be able to search it more quickly. Although "hashing" involved numbers, its intellectual origins lie in Luhn's thinking about texts, language, and meaning.

This was a very different, and more ambitious, way of imagining and understanding what computers could do, and what they could be. Luhn was trying to find ways to help people understand the larger and larger amounts of written information being thrown at them. Solving these "information overload" problems was much harder than doing most kinds of numerical calculations and required different kinds of machines and different ways of using them. Luhn's unique inventions – from hashing to KWIC – were the result of bringing different modes of reasoning to bear on machines.

The historian of technology Michael Mahoney once called the computer "a protean machine." That is, a computer is not just one thing, but many things at once, a machine waiting to be shaped to many different kinds of purposes. Most often, when we think of computers, we think of them as giant number-crunchers. We usually think of the power of computers in terms of calculation*s* and operations per second. But the story of hashing suggests that we should be open to the multiplicity of computers, to their ability to be used in different ways and for different purposes. Luhn's story shows how we can imagine vastly different uses for our machines that can open up new territories for exploration.

Today, hashing plays a wide range of roles in managing our everyday digital lives. The passwords we use every day for email, Facebook, other online services are usually stored as hashes. When you enter your password on a website, it is hashed and then compared to the hashed version in the website's database. If the hashes match, then you are granted access. Storing a hash, rather than your actual password, has the advantage that anyone who might gain unauthorized access to the database of passwords would still not be able to obtain your password, only its hash [**figure**].

This security of such systems relies on the fact that hashes are one way streets: it is easy to go from the password to the hash, but very hard to go from the hash to the password. This property means that hashes have a range of uses in cryptography. Digital signatures, for example, use hashes. If you wish to send a document securely (ensuring no-one has tampered with it en route), you can send the document along with a hash of itself. In this case, the hash is often called a "digest," since it is a condensed version of the text. For example, hashing the text, "The quick brown fox jumped over the lazy dog," might produce the hash "12345." If someone modifies the document even slightly ("The quick brown fox jumped over the lazy dop") this is very likely to produce a very different hashed result, say "67890." By comparing the has generated by the sender to the hash generated at the receiving end, it is possible to detect any tampering [**figure**].

Large file storage and backup systems such as Dropbox and Google Drive also use the unique properties of hashes to save space on their servers. Many of the files you upload to these services are probably identical to files already stored by other users (the latest Lady Gaga tune, for instance). Dropbox doesn't want to waste space storing the same file over and over. Instead, when you attempt to upload a file, Dropbox makes a hash of your file and compares it to the hashes of all the files it already stores. If the hash is the same, chances are the file is the same. This saves you time too: no need to wait for the file to actually be uploaded – Dropbox just provides you with a link to the file it already has.

Hashes are at work all around us, helping us to manage the vast amounts of data that we rely on for work and play. More often than not, hashing draws us into text: words, sentences, concordances, abstracts, indexes, and digests. Hashes have allowed us to think about computers, more and more, as textual tools – reasoning with letters and words. Luhn's story suggests how "computer thinking" is tied not only to mathematics, statistics, and logic, but also to traditions of thought in philology, literature, and textual analysis.

In 1964, the New York Times reported the passing of a "data process specialist": "Mr. Luhn, in a demonstration, took a 2,326-word article on hormones of the nervous system from The Scientific American, inserted it in the form of magnetic tape into an I.B.M. computer, and pushed a button. Three minutes later, the machine's automatic typewriter typed four sentences giving the gist of the article, of which the machine had made an abstract."

In an era when many of the most important applications of computers are textual, Luhn's abstracting does not seem especially impressive. Now, "data processing" seems a natural way to think about what computers do; Google Translate, Google N-gram viewer, Google AdWords, and Google Search are all devoted to determining, in one way or another, the meanings of texts. The explosion of information on the Web has made automated reading and understanding of central importance to business, to science, and to almost everyone.

In the 1950s, however, this was a revolutionary way for Luhn to think about machines and their function. The history of computing is often depicted as a straight line: ever faster and ever smaller machines taking us from the military-industrial complex to Steve Jobs to Google. But Luhn's story, and the story of "hash," suggests a more complicated and interesting history of multiple origins, multiple disciplines, and competing visions of what a computer could be for.