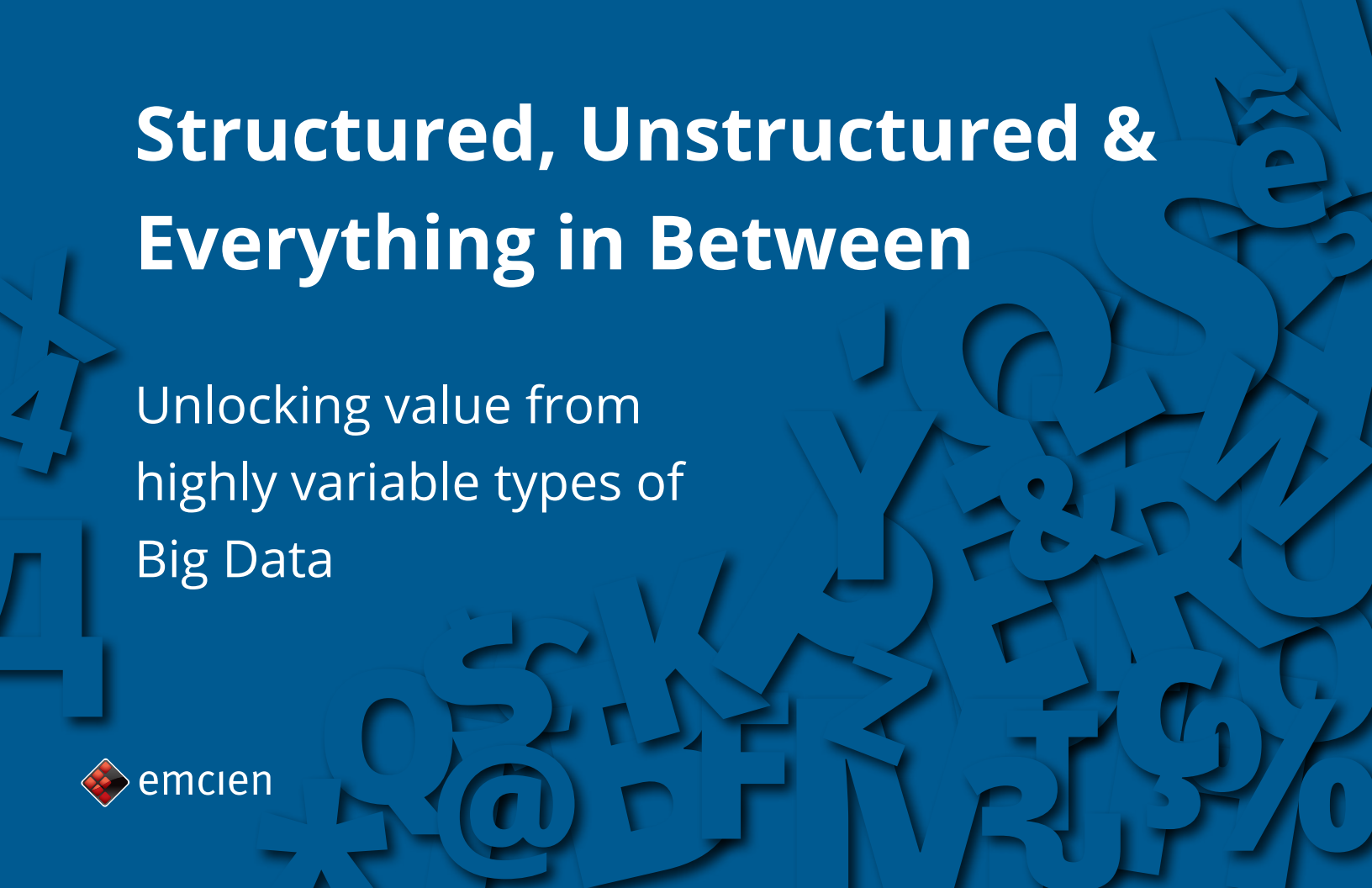
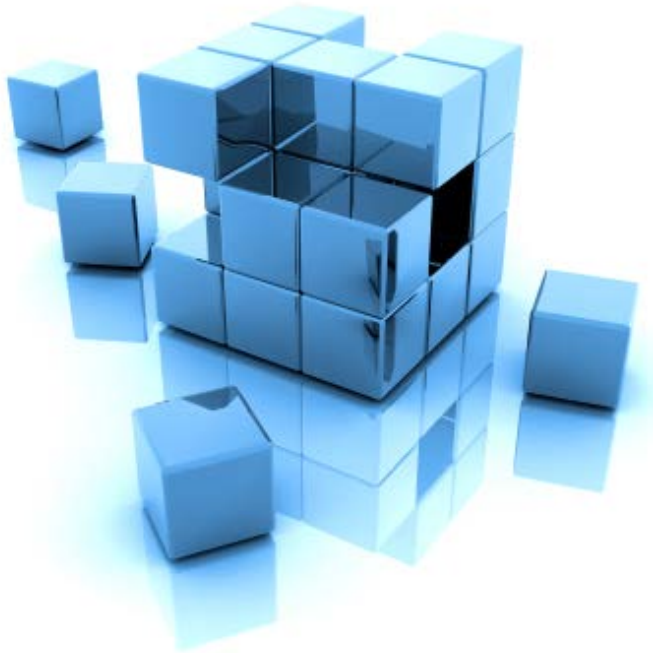
Structured, Unstructured & Everything in Between

Unlocking value from highly variable types of Big Data



emcien

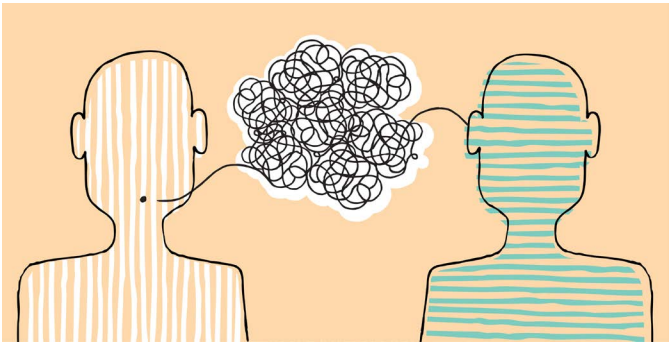
Table of Contents



Do you know ELIZA? Who is ELIZA? How does ELIZA make you feel?

These are a few of the plausible questions that MIT scholar Joseph Weizenbaum's ELIZA program could have asked while running its DOCTOR script. This program was born between 1964 and 1966, when Weizenbaum wrote many scripts for the project, the most famous of which was DOCTOR—the question-forming script that became famous for its ability to enrapture unwitting people in questions that mimicked Rogerian psychology, a popular form of psychotherapy at the time. If the human interacting with DOCTOR through screen and keyboard mentioned that his brother hated him through his typed text, the DOCTOR script would know to type back a question such as "does anyone else in your family hate you?" The DOCTOR would continue to prod based on collected information.

This early attempt at what we now know as Natural Language Processing arguably passed the Turing Test, a theory established by early Artificial Intelligence scholar Alan Turing. The Turing Test is passed if a human interacts in a natural language conversation with both a human at another computer and a computer following a script and the test human cannot distinguish between the two.



ELIZA and the Turing Test expose one of the greatest challenges in data that we have seen, and that we continue to see: **words.**

So much of what we have recorded in history and have built the mechanics for dealing with has been numbers. Why? Because they behave themselves. While the number 7 might not have always looked the way it does on a fairly universal scale now, the concept of 7 has always been the same.

When something is as steadfast as the concept of a number, the theory around it does not change much. Language is, however, a type of information that has always been spoken or written down. While individual letters and numbers can be translated into the 1s and 0s of binary code, the daily twist and turn that is the living beast of individual languages does not behave so beautifully as any kind of number. Linguistics maps language, but the cartographers can hardly keep up.

There are, of course, numeric and non-numeric forms of data that have not, and do not, behave as well as numbers. Most of the data created since the advent of the consumer Internet has been generated in one of these less than easily analyzed, generally unstructured formats. Whereas numbers and other general information could previously be stacked into a neat database and analyzed with some ease, the massive amount of machine data and other less rigid data formats such as email, video, sound, and photo present a larger challenge.

With great challenge comes great opportunity.

In this ebook we'll approach the varied and sundry types of data that make up the data economy, explore their meaning in depth, and consider whether or not each type of raw resource is inherently valuable or would provide the strongest value for organizations if paired with a complementary data set.

The Information We've Collected Throughout History & Why

Nature vs. Nurture

In order to understand why data exists in the various forms that it currently takes, it's important to return to the very reasons humanity has chosen to record varied and sundry forms of data, specifically relating to numbers. What else is more compelling for most of us than money?

Trade is an essentially human, cooperative practice that is woven through the fabric of our history. Early artifacts from as far back as the Paleolithic period indicate that we have bartered with one another to trade a needed or desired good between two communities or entities. While the concept of money did not enter the world stage until much later, most ancient civilizations' artifacts contain some record of goods that were traded or purchased in a ranging complexity of numerical expression. With the advent of coin currency and the growing breadth and depth of international trade came greater interdependency and the requirement of keeping track of agreements, ledgers, and transactions. Thus the modern economy was born.

Many other calculations and records come to mind when considering the length of human history: astronomical measurements, the struggle to accurately measure time using calendars, tracking the growth of crops each harvest, and attempts by great civilizations like the Roman Empire to count their people, are just a few examples. While each category is important, **trade's use of the abstract concept of money was an important catalyst in the recording of highly structured, carefully audited tables**.

Before bills of sale were issued, a coin's value came literally by the weight of the metal they were made from. The British pound was just that, a piece of silver equal to the weight of one pound of the metal. As it became increasingly dangerous and impractical to transport massive amounts of metal-based wealth from one place to another, the evidence supporting the need for another system of payment was strong.

The advent of the bill of sale and the bank note ingrained the abstract concept of the economy into the rest of history. As Italian tradesmen saw a growing demand for their fineries, it became necessary to develop a document that guaranteed payment at a later date.

The interaction of these tradesmen with others from around the world introduced the concept of paper money to European states that were looking for heavy, cumbersome metal money. By the rise of the Industrial Age paper money and bills of sale became the primary source of payment in trade, replacing the antiquated systems of metal money and tallies used previously.

As printing and paper became less expensive and more practical, the previous methods of recording went by the wayside. The British system of using wood tallies fell symbolically when the Old Palace of Westminster burned to the ground in 1834, an event caused by the overheating of the stove that burned excess tally sticks. These sticks were an early way for the British to keep track of taxes, debt, and more when paper was more expensive. The current Parliament building is an incredible monument to what came after, and of the power of the British Empire around the world that was fueled by economic growth made possible by abstract money concepts.

Ledgers are some of the first truly structured data in existence. One of the most famous ledgers in history is the one in Luca Paciloli's Summa de arithmetica, geometria, proportioni et proportionalità, the Italian mathematician's book on the art of record keeping for merchants, published in 1494. The double ledger system prescribed by Paciloli to balance credits and debits to an account is also where the first representation of the plus and minus sign is introduced.

The double ledger allowed merchants to track the elusive state of paper-money trades made with individuals coming and going across the seven seas in a focused, balanced way. **The calculated entrance of exact figures by date, time, and name allowed merchants and others to more thoroughly analyze the comings and goings of money, goods, and promises unkept**.

The slow, steady advancement of time and the development of complex international trade also led to increasingly complex ledgers and data collection. Looking back at the humble beginnings of the monetary economy, it's easy to understand that most of the data that was collected before the advent of the Internet was information collected about business.

Without big data analytics, companies are blind and deaf, wandering out onto the web like deer on a freeway."

Geoffrey Moore
Author and Consultant at The McKenna Group

Business deals in the world of numbers, and numbers behave themselves. They can be stored in neat rows and columns that can be sorted with ease. With careful bookkeeping and attention to detail, nearly anyone with basic mathematical skills could manually track the income and expenditures of an enterprise with precision. Once personal computers and accounting programs were added into the mix, the amount and variety of data that could be stored in ledgers and other databases grew, along with the complexity of the type of analysis that was possible.

There have been other types of data recorded in history, from the breadth of knowledge collected in books to the music notes scribbled onto staff paper, to the mathematical equations to the priceless paintings and other art that make our world rich. Until recently, however, most of the collected data in the world could not be turned into numbers that were more easily analyzed and stored. The data simply didn't behave. It didn't fit itself into neat tables like ledgers that could be sifted through and balanced. Now that the age of machines allows us to scan and record virtually anything and turn it into some sort of numbered form, we're closer to uniting non-numeric data with the well-behaved world of numbers. Gone are the days of cumbersome paper logs, the need of physical photos, the morning paper, or a letter to Aunt Jane from your hotel room in Rio. All of that has floated into a data-filled cloud, a place filled with intricate structure and loads of information that hold little value other than they might one day be useful.

Businesses have always collected structured, numerical data, but we're beginning to realize that for the past century they have also been siloing more and more data that doesn't quite function as it did before. **The 21st Century is the age of data behaving badly and the struggle to figure out how to whip it into shape for the sake of value.**

The Variety and Variability of Data: Numbers, Words & Everything Else

Data collected by businesses is typically collected in a mostly clean, structured database by necessity. If you've ever seen an old-fashioned cash register, you've seen the record of sales recorded right before your eyes, the tiny inked digits representing the money flowing into the register and out to make change. If you've ever had to manage a budget, you know even more intimately the strict attention to detail that dictates the rows and columns of your spreadsheets or paper ledgers.

Modern business intelligence collects information than can populate a neatly ordered database, but since the Internet became an important part of everyone's lives, businesses have started to collect information that doesn't fit within such confines.

Information about the types of people that purchase products, when they purchase them, how they purchase them and how they receive their goods has been stored by businesses looking to dig down into the information.

Today much of it doesn't fit within the confines of our modern tools like the ledgers of the past. The rest of it that is in fact structured has also grown so large that it is often manifesting itself less as a tool and more as a hindrance.

Organizations everywhere are beginning to feel mounting pressure to collect and store any form of data they can get their hands on. From nonprofits collecting data on the populations they serve to social media giants looking ways to monetize, data of any kind is the raw good that most organizations this century will look to spin into piles of gold.

The challenge right now, however, is finding meaning and use from unstructured data that we don't currently have the tools for. So what exactly is the difference between structured and unstructured data? The divergent labels imply a definition of more neatness than really exists in either type.

Structured data is data that has been entered into a system of rows and columns that indicate a type or quantity. The most common example of structured data that the layperson encounters is a spreadsheet. Rows in your database indicate some type of thing that you're keeping track of, and columns provide some indicator of the attributes of the item in the row. These database spreadsheets can be gargantuan in size or can be as simple as the primitive, paper-based double ledger developed by Paciloli. A data-filled kitchen of structured data would have all its forks, knives, and spoons neatly stashed in their compartments, the knives sharpened and smartly stored in their block on the countertop, the pans stacked in order of size and stowed snugly under the range.

The world of unstructured data looks a lot more like the clutter-filled junk drawer, only the junk drawer has overflowed into the rest of the house. Unstructured data is a much broader category that exists in many forms. This data includes all of those emails stored in your account, the breadth of digitized music and movies floating around the Internet, the information silently generated by machines as they communicate with one another, and much more.

The potential benefits from analyzing data are limitless and three quarters of organizations predict that big data will be in mainstream use within the next three years."

Matthew Yorke
Chief Executive Officer at International Data Group (IDG)

The world of unstructured data can look like anything from the binary counterpart to specific words, to pages of IP addresses with no other information attached.

The swath of information covered under the unstructured category is the catalyst for the dichotomy of structured vs. unstructured data. Think about your last doctor's visit. If you've been visiting the same doctor for some time, he or she probably has a wealth of information about you: blood test results, height and weight at your given age, your blood pressure range, and lots of other little notes about your comments or complaints and the nature of your visit. Though these charts and records are all written into forms and stored either in a paper file or a computer file, there is a lot of information in forms that today's number-centric analysis tools simply can't tackle. If your records have been entered into an electronic database, traditional analytics tools could parse through information about your weight and other numbers listed, but the wealth of doctor's notes (the unstructured data) would be left behind.

Another useful way to understand the grey areas between structured and unstructured data is through the concept of business intelligence. Over the course of decades, businesses have entrenched themselves in more and more technology to collect and store transaction data. Businesses carry plenty of structured information about a range of things: product flow, the customer demographics, changes in the prices of raw goods, income, and expenses. These numbers are well-behaved. **As businesses became ever more entrenched in machine-based transactions, including internet purchases, businesses started storing more and more extraneous, less numerical information about transactions, simply because it existed**. All of the data, structured or not, cleaned up or not, has been siloed into large collections of information that businesses were scared not to collect, simply because it could one day prove valuable.

These are not the only instances of unstructured data being stored in some sort of database to be labored over at a later date.

The United States Census Board might have been a pioneer in making sense of massive amounts of data, but there are many unstructured forms of collected information that Census workers collected systematically, but that don't fit neatly into rows and columns of a spreadsheet.

While the ultimate delineation between structured and unstructured data might appear to be a structure, that's only half of the truth, as many kinds of data that we are currently grappling with also have structures: words fit into linguistic patterns, photos have a composition, IP addresses have a logical order, and machines speak in script that has an order.

There are databases out there that these things fit into, but they simply don't behave the way numeric databases did. To understand why, let's look at the difference between numbers and everything else.

Numbers and Everything Else

Simply saying that numbers are easily analyzed would be to grossly underestimate the complexity and range of existing number types. There are four kinds of numbers that are frequently encountered when analyzing data, and only one of them has been studied at length and has intrinsic value. The set of mathematical tools that we have to analyze data were not developed over thousands of years to reign in the meaning of the other three.

“More than the amount of data itself, the unstructured data from the Web and sensors is a much more salient feature of what is being called Big Data.”

Thomas Davenport
Author and Professor at Harvard University

Nominal numbers: these are any sets of numbers that have no other meaning attached to them other than being an identifier. For example, postal codes in the United States are usually composed of 5 numbers. A postal code in Atlanta, GA, is 30312. The number itself does not mean the value 30,312—it is a code number to indicate a geographic area. Similar categories are license plate numbers, Social Security numbers, and bank account numbers. These numbers cannot be added, subtracted, multiplied, or divided, as they have no inherent numeric value.

Ordinal numbers: these are numbers indicating a range of values. If you have ever completed a survey that requested you to rank something from 1 to 5, with 1 being the lowest in value and 5 being the highest, you have seen ordinal numbers. A linguistic version of ordinals is first, second, third, fourth, fifth, and so on. In complex mathematics, ordinals are part of a well-ordered set that are an extension of natural numbers. As indicators of rank, ordinal numbers only have the meaning attached to them in context: one being lowest, three being neutral, five being highest, and so on. Ordinals in the mathematical context can be added, multiplied, and exponentiated.

Interval numbers: Interval numbers are more rare than the numbers listed above. An example of an interval scale is that of temperature—numbers that indicate a point that is more or less than a fixed point.

Real numbers: real numbers indicate the value that is actually attached to them. 7 means 7. The number 1 indicates the value that the symbol traditionally implies when seen. Real numbers indicate a value on a line of numbers that we conceive of as infinite, either negative or positive. While there are other types of numbers that fit into the realm of mathematical knowledge, real numbers are the largest swath of value points that we have a set of tools to understand.

While a column full of real numbers could be added, averaged, turned into a graph, or further manipulated for result with relative ease of calculation, a column full of nominal numbers would not provide the same results. Nominal numbers act much more like words. So are the varying types of other data that have been turned into numbers that don't have numeric meaning.

One of the reasons why data of varying types have exploded the collective growth of information is the ability to digitize almost anything.

With the help of numbers that represent things, photos can be turned into coded numbers that tell us the intensity and variety of color in every pixel of a given image. The same is true for reproducing digital video and sound recording. While words will always carry their own nuances of meaning in their whole form and will continue to inform the meaning of other words within a given linguistic tradition, even words and letters can be translated into binary code, stored, and reconstructed later.

While nominal, ordinal, and interval numbers combine forces with binary code to create big headaches for analytics, the approach of mere words as words is presenting what might possibly be the biggest analytics challenge yet. **ELIZA was mentioned in the introduction of this ebook as a primitive form of artificial intelligence that mimicked the natural language patterns of the human interacting with it**. This early attempt at processing natural language was a breakthrough at the time, but it didn't take into account spoken language.

Modern natural language processing efforts run up against the complex problem that spoken words are really difficult to turn into numbers that can be understood by computers. Why?

In the new world of data analysis your questions are going to evolve and change over time and as such you need to be able to collect, store and analyze data without being constrained by resources."

Dr. Werner Vogels
Chief Technology Officer and Vice President at Amazon

Words are flexible. They shift shape and form, take on new meanings rapidly, and many languages borrow words from other languages that might even have another implication than the initial meaning in the original language. Think of the word “touché,” which in French means “touch,” while the English meaning implies acknowledgement of the success or appropriateness of an argument or implication in the context of a conversation.

Better yet, some languages possess words to describe emotions, actions, or feelings that are completely missing in other languages. What is a computer to do if an Italian asks a computer to translate “ti voglio bene” to English when we have no succinct idiom to express the joy and attachment one feels for the circle of their friends and family? What about a Dane trying to communicate the meaning of the word “hygge” when standard English doesn’t have a more positive word for the warm, snuggly feeling one feels after good food and drink with their loved ones other than “food coma”?

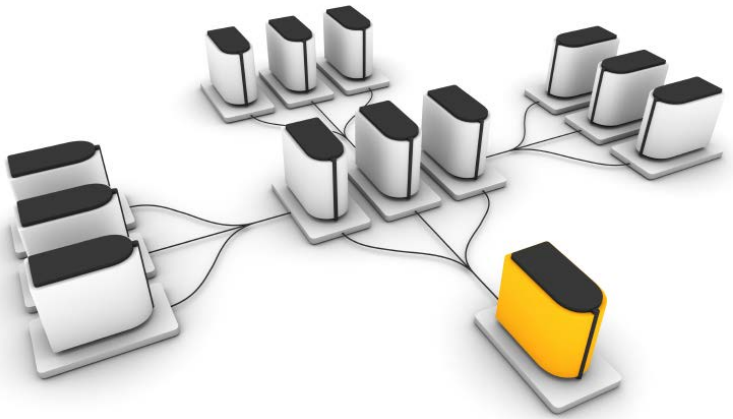
The imprecise, ever-changing nature of how we communicate most ideas simply does not align well with the stolidly unchanging world of real numbers. While we can assign numbers to words for the sake of finding patterns within a larger collection of words, the analytic structures that we have in place to manage numbers simply cannot be layered over the massive amount of data that exists in words and expect it to perform.

Computers are not yet sentient, but if they were, they probably wouldn’t tell us, just to engage in a little data-frustration schadenfreude (that’s German for the giddy feeling of taking pleasure in another’s pain and suffering).

Similarly, **machines might be able to analyze movement, pick out patterns across images, refine satellite imagery to obtain license plate numbers and sense the movements of an algorithm-control robot, but the analysis of whole images for patterns that the human visual cortex easily understands simply has not happened**.

The density of numbers present in data around a single photograph allow a computer to understand some interesting attributes about it, but a computer cannot yet scan through hundreds of thousands of images and find a classifiably "beautiful person" based on data on the culturally normative assumptions of the person telling it to do so.

Google completed a project that united thousands of computers together on the mission to identify cats' faces across millions of YouTube videos, but it took the machines days to successfully teach themselves what a cat's face looks like.



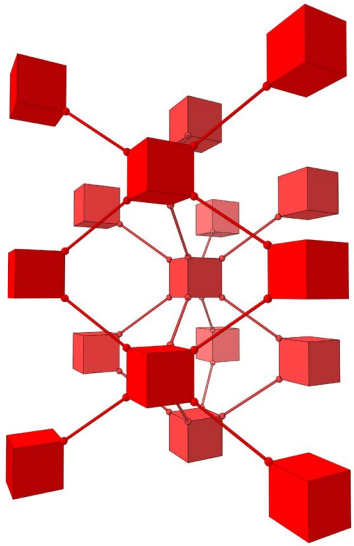
Until more processing power exists, more data is collected, and better solutions are created either by ourselves or computers, much of this non-numeric, unstructured data might rest in the collective store of human

The Many Data Microcosms

Various Types of Data

Now that the heavy lifting of understanding databases and the different classifications of data is done, a full snapshot of the many types of data that exist can be laid out and more easily understood.

A number of data types have already been mentioned, but they'll be explained here and provided a grouping based on typical structure and whether the data usually involved numbers or something else.



A fact one must accept and understand is that nearly everything can now be turned into a data point and stored: where you stop at a traffic light, the measurement of the wind's speed to guide sails, the density of the earth we stand on, and the invisible radio waves moving through time and space across the universe are potential pieces of information to be stored and possibly used at some point.

The following lists of data paint broad strokes about the various types of data that can be used to leverage some sort of real-world value—whether that be monetary, to solve one of the world's biggest challenges, to make a scientific breakthrough, to develop supremely efficient logistical operations, or to achieve some goal we haven't even thought of yet.

Website Data: this information is incredibly broad and complex. Website data can range from the unstructured, non-numeric data of the words you read to the nominal IP address numbers that act as a sort of calling card for who has been to the website and when. This information also includes metadata regarding web traffic statistics, where the website is linked to, the website's design, and any other images and videos on the website. Considering the complexity of information included in websites, it's no wonder that the Internet age has driven an explosion of data.

Email Data: Similar to website data, email data is primarily unstructured and non-numeric in nature. Emails include everything from words to images and videos, in HTML or not, and can include many recipients. Tack on the possibility of email attachments and the layers of complexity are dense. Even one more level up is how and when emails are sent, if they're classified as spam or not, and any given number of other crumbs of metadata.

Social Media Data: one of the Internet's most conspicuous data sources is the constant stream of social data provided by Facebook, Twitter, an infinite number of blogs, and a number of other social sharing mediums. This data is almost all unstructured, non-numeric data. Where this data's real value lies is in the tracking of what is shared, how it's shared, where it's shared from, and who it is shared with. Sharing habits reveal information from where a power outage is happening to the evolution of a word's meaning. Social Media data can be mined for patterns of human behavior, making it one of the more sought after value creators.

Video, Photo, and Sound Data: the data surrounding these mediums presents a unique challenge for analysis. All of these forms of information are broken down into the numbers needed to make them digital—indicators of intensity of color or tone, brightness, speed, and so on. This data is generally very unstructured. Like words mentioned previously, photos, videos, and sound, these data types have complex behavior in relation to real numbers. These types of data can be easily shared or sent, making the data around sharing valuable, but the complexity of recognizing patterns in these items leave standard analytics in need of much more complex solutions.

Mobile Data: information generated from connected mobile devices has exploded the amount of data available for analysis. If your tablet or smartphone is connected to a data source, the sensors in them are giving off cell, network, and geolocational data, tracking every action taken. Call times are tracked, as well as which numbers you called and which numbers called you, where they were calling from, and what type of connection was used. Mobile data also collects any information about interactions on your phone using any of the mediums listed above, including any gaming done on your phone.

Machine Data: all of our devices with sensors are "speaking" to one another through their own language. Machine generated data is sometimes structured into a database and many times is not. Machine data can include measurements taken about any number of conceivable topics or ideas, tokens generated based on user actions or interactions with other machines, recordings of sound or video that the machine has made and processed, IP addresses, transcription of human words, and much more. Machine data is one of the greatest forces in ballooning our pool of data, regardless of what form it is in, because as long as a machine has power and something to do it can function and generate more data. While machines do not speak in natural language, they do have their own language.

These are only the types of information generated on the Internet through every day user interfaces and the hardware we interact with.

Human life also generates a lot of data. From business interactions to researching a cure for cancer, human enterprise has always created lots of data. We're just able to digitize it for ease of access now. That process comes with risks and benefits, but the collection of more easily accessible data is making for a richer understanding of everything in the world.

“

The goal is to turn data into information, and information into insight.”

Carly Florina
Former President and Chair at Hewlett-Packard

Here are some of the largest growing types of human generated data that can be stored and analyzed like the data listed above.

Medical Data: one of the greatest overhauls in the way we understand our health is the collection of medical records into electronic databases. Medical data is also being collected by specially designed sensors that record number of steps walked, heart rate, sleep conditions, and more. Medical data has traditionally been taken down by hand in a file, with doctor's notes scribbled in famously indecipherable handwriting. With the advent of digitized records, nurses and doctors can enter data about the patient directly into a secure medical database that allows for ease of sharing with other practitioners. Medical data, especially medical notes, have historically been unstructured, non-numeric information that is difficult to analyze. As mentioned before, we need better analytics to tackle this unstructured data and make sense of it. We are, however, building more and more knowledge about the human medical condition that can help doctors better understand how to eradicate diseases and improve human life.

Business Data: Businesses have siloed structured and unstructured data of words, numbers, and everything in between, for decades. Business intelligence professionals have attempted to mine data from transactions and other business functions to meet the bottom line more efficiently and profitably. More often than not, though, they've put this data in a bathtub, found no repeatable method for making sense of it, and have forgotten about it. The burden of this data has grown for businesses as they have begun to also track social media information around their products, their customers, and what a potential customer might look like based on competitors' data. The world of business data might be one of the more complex ones, as businesses have recognized the value of data collection for longer than other organizations.

Governmental Data: the complex federal, state, and local governments of this age collect an incredible amount of data about the people they are governing or ruling. From census data to intelligence information collected through spy satellites and data mining, government entities understand the importance of gathering metrics on everything that might demand attention in the course of decision making. The data that the government collects ranges from highly structured survey and tax information to unstructured legal contracts and policy notes. Governments everywhere are having to catch up to modern technology and figure out how to both fund the tools that will help make sense of data and how to write policy that governs the use of citizen data within the strictures of their respective constitutions and laws.

Scientific Data: much of today's scientific data is produced by machines that are monitoring actions and changes, but this data is also being produced by human observation and recording. Scientific data is traditionally very structured, as the scientific process dictates that an experiment be replicable. This data ranges from observations taken during Large Hadron Collider experiments at the European Center for Nuclear Research to the number of cells counted on a slide by a college lab research assistant. Perhaps the biggest example of scientific data that we can all relate to is genome mapping.

Historical Data: information about the history of the planet, from biological events to moments of triumph over systematic oppression, have been recorded. Granted, it has been stored in books, and discussed mostly in the highly unstructured form of words. Databases allow the transcription of events in history into more automatically analyzable forms that help thinkers like futurists review the past, to understand what's happening in the present, and help shape the future. As machines move closer to a better understand natural language, the corpus of all books ever written as part of history might be conceivably analyzed for their importance in time and place, as well as for an understanding of how written language and communication have evolved alongside the development of ever-deeper human intelligence.

Finding Value in a Variety of Data

Clearly these lists are far from complete. A separate book could be written about the nature and history of each one of these data types, and by the time the book is written the information would already be out of date. **One of the most exciting challenges of data's proliferation is the fact that it grows and changes at such profound rates.** As we grow in innovation and the data grows in density, we'll move closer and closer to being able to let that data speak to us. In the meantime, we will tackle the data we have with all that we've got.

There's lots of value to be found in the data we're able to make sense of, and we do have some pretty impressive analytical solutions for taking on massive waves of data and finding the sense in them. **Any raw good or resource takes on much more value when manipulated to become something more refined. It's time to chisel and hammer our way to maximum data value.**

Click here to see data analytics tools in action!