Data And Analytics Semester 2

UCC Data and Analytics Society Newsletter

Semester 2 2021

Contents

Editor's Note

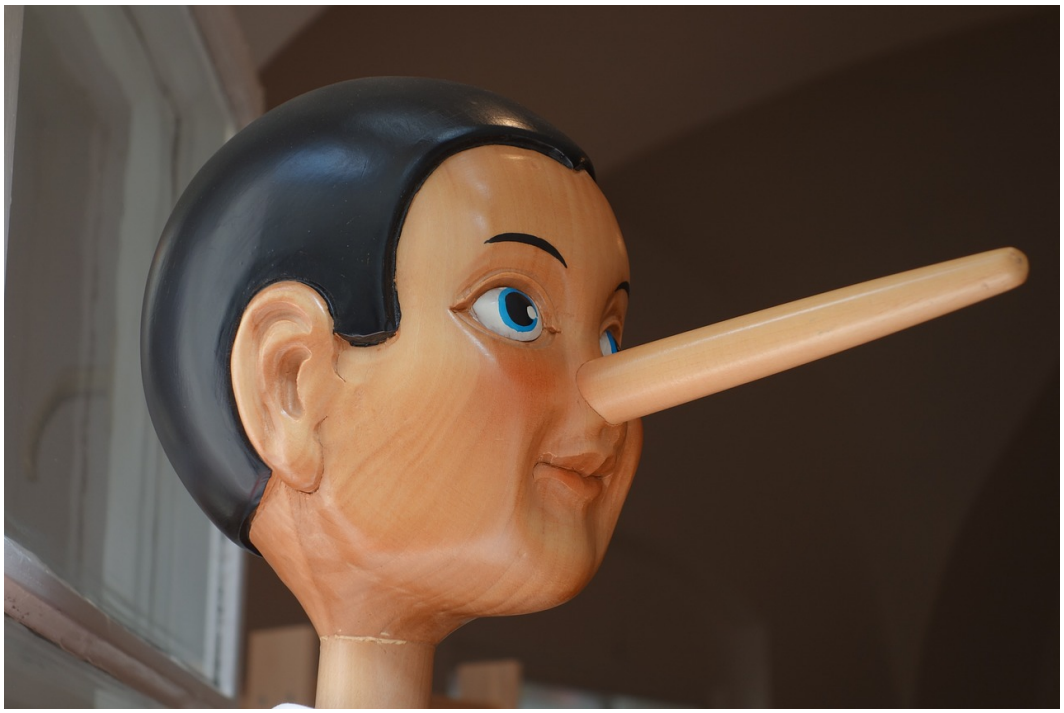
Hello and welcome to the second issue of the Data and Analytics Newsletter. In this issue there are a variety of fascinating articles touching on many areas relating to data science and analytics, contributed by Data and Analytics Society committee members. I hope that you enjoy reading them and learn a thing or two about data and analytics!

Laura Cosgrave, Publications Officer

Stats Never Lie. Or do they?

Aine Ginty

Imagine that you have just been appointed CEO of a prestigious, multinational organization. You are eager to make sure the company reaches its maximum potential under your leadership.



Having heard that the use of AI, predictive analysis and machine learning can dramatically improve decision making strategies, you decide to invest in the most modern, powerful and expensive software. It has been promised that these data analysis tools will improve your company's marketing techniques, finances, customer service and much more.
You are a diligent and careful decision maker and you make a huge effort to ensure all company decisions are based on the results obtained using the new predictive and analytical software.

One year later and there is a deterioration in the performance of your company. What could possibly have gone wrong? Statistics, facts and figures never lie. Or do they?

Data is driving critical decisions in our economy but the quality of those decisions is only as good as that of the data on which they are based. According to research done by IBM in 2016, the US economy loses an estimated $3.1 trillion dollars each year as a result of bad data. Poor quality data can have a detrimental effect, leading to erroneous decision making, missed opportunities and loss of revenue. The complex, high tech algorithms are only half the solution to improving decision making strategies. The other and arguably more important aspect is the data.

High quality data is characterized as that which is accurate, complete, relevant, valid and consistent. The Forbes Insights and KPMG "2016 Global CEO Outlook" study found that 84% of CEOs are concerned about the quality of the data upon which they base their decisions. Research carried out by Experian Data Quality in 2015 suggests that the average U.S. company loses 12% of its revenue as a result of poor-quality data. Data is driving critical decisions in our society but low-quality data leads to poor quality decisions.

Stats Never Lie. Or do they?

Aine Ginty

Data cleaning is the process of identifying and fixing any issues in a dataset in order to increase the accuracy and quality of the data. This may involve replacing, modifying or even deleting data. Important steps to data cleaning include removing duplicates and irrelevant observations, fixing structural errors in the dataset, filtering unwanted outliers and handling missing data.

Duplications in datasets may arise when combining data from multiple sources. Differing structures and measurements can also lead to issues when collating datasets.
Even on a simple scale, we can think about how easy it is for data to become unhygienic. For example, combining datasets could lead to mislabeled categories, resulting in duplicated information. For instance, we could end up with two different classes “N/A” and “Not Applicable” in our dataset when in fact these actually represent the same collection of data.

Gathering data from different sources can also lead to inaccuracies in measurements. For example, consider a multinational company taking finance calculations from its headquarters across the world. A failure to standardize the currencies could lead to highly incorrect data in the company database.

Although these examples may seem straightforward and unsophisticated even the biggest and most prestigious organizations struggle to maintain consistent, accurate and complete datasets.

Take for example, NASA’s 1999 Mars Climate Orbiter Expedition, one of the most infamous examples of inconsistent data leading to disaster. While working on this project, the navigation team used the metric system units for their calculations while the team responsible for building and designing the spacecraft performed their calculations in the English units of inches, feet and pounds. This lack of consistency in the data led to a failure in the system which resulted in the combustion of the $125 million spacecraft after almost 10 months of travel. It was ultimately data that failed this space mission.

Some argue that data is more valuable than oil and gold. We all know that “all that glitters is not gold” and likewise we must learn to acknowledge that all data is not high-quality. Undeniably, there is an abundance of data in today’s modern online world but always remember, data is nothing if not high quality, relevant, accurate and consistent.

The Not-So-Distant Future of Data Storage

Jack O'Connor

In the age of the internet, data is king. However, a king without a kingdom is no king at all, and if there aren't data storage devices to physically house this data it may as well never have existed.

Every day vast quantities of data are produced by computers across the globe - IBM estimates up to 2.5 million Terabytes per day – including everything from pirated copies of movies for personal consumption to particle accelerator collision results for scientific research. The rate of data generation is only going to increase into the future and the long-term solution to physically storing all this data cannot be to exponentially build more and more data centres that depend on current magnetic storage device technology (such as the hard drives found in nearly all computers).

Luckily there are many promising data storage technologies on the near horizon to alleviate this issue. Two such technologies are 5D optical data storage and DNA digital data storage. It is quite likely that the vast majority of archived data in the future will be stored in data warehouses utilising one of these two technologies.

It must be pointed out though that these technologies will probably never be used in personal computers or supercomputers with an emphasis on fast computations as they simply cannot match the speeds of data operations supported by RAM and SSD, which are on the order of nanoseconds and microseconds respectively.

Hard drives only have an expected shelf life of 5-10 years before they start to fail and incur data loss. All data centres that use hard drives must undergo lengthy, expensive copying procedures every few years to ensure data integrity is maintained.

There are many companies that have data that they must store for periods of time far longer than this for archival purposes such as in the case of Disney, who keep every single frame of footage and animation for their entire library of productions, or other companies which must hold onto data for legal reasons.

Companies like this will pay big money for data solutions that allow them to safely and efficiently store their data far into the future. Two new technologies which could be the solution to their problems are detailed below.

The Not-So-Distant Future of Data Storage

Jack O'Connor

5D Optical Disk Data Storage

The word “digital” when talking about data is an adjective that means “expressed as a series of 1’s and 0’s”. All data stored on computers in its rawest state is a string of binary ON/OFF 1’s and 0’s, known as bits. 8 bits are known as a byte.

A standard CD uses a simple trick to store digital data: a tiny laser is shone on a CD, it follows a linear spiral track along its surface and whether or not the CD reflects the laser back is enough to encode a sequence of bits. Thus a CD can be thought of as using two spatial dimensions to store its data, these two dimensions being the surface of the disk.

5D optical disk data storage utilises three more dimensions than a standard CD to store vast quantities of information bits in a given volume of space. The first of these three extra dimensions is a standard spatial dimension, height, allowing many layers of data to be written onto the disc.

Next, using a femtosecond laser (each flash of the laser lasts around a quadrillionth of a second) structures called nanogratings can be etched into the internal layers of the quartz disk. The remaining two dimensions are the slow axis orientation and retardation, both properties of materials that affect how light passes through the material in unique ways. Each nanograting’s slow axis orientation and retardation of light can be changed by varying the femtosecond laser’s settings at the moment of etching.

There are enough combinations of orientation and retardation that each combination can represent a unique byte of information. Compared to a CD which can only represent only two states of information - a bit - on its surface layer, 5D disks can represent up to 256 states of information - byte - on each of its many internal layers. It is this difference that accounts for the huge disparity in the medias’ respective storage potential.

In order to read back data from a 5D optical disk special equipment is needed. First you need a polarise, laser, and a microscope. When reading back info from a specific section of a disk layer, the mentioned equipment is focused on the relevant section and the images formed on the microscope are captured digitally. It is then up to complex machine learning algorithms to parse out the stored bits of information from these images. Currently, 5D optical data storage is still on the same order of magnitude in terms of storage capacity as hard drives but it is estimated that they will eventually hold up to 360 TB/disc.

The Not-So-Distant Future of Data Storage

Jack O'Connor

The main draw of 5d optical disks is not its huge storage capacity. In terms of pure bytes of information per volume (B/mm3) 5D optical disk is outmatched by DNA data storage. What 5D far excels at is its durability. Each disk is about 3 inches in diameter, relatively thick when compared to a CD and made of crystalline materials such as fused quartz - as such it is easy to imagine they are not the easiest things in the world to break

Even better than its physical durability is its chemical durability. Fused quartz is a very stable material. It strongly resists deformation by heat and maintains its exact state without change over time. Using extreme testing conditions it has been predicted that at optimal storage conditions can give 5D disks a working lifetime on the scale of the current age of the universe, a whopping 13 plus billion years. For all intents and purposes this may as well be categorised as data which lasts forever – completely uncorrupted.

One final benefit of the almost indestructible nature of 5D optical disks when it comes to data archiving is that these disks are write-once. Unlike hard drives and DNA data storage whose contents can be overwritten, once a piece of data is lasered into place in a 5D disk, it is there to stay. This is perfect for data archiving because there is no worry that the data you will be looking at in a 100 years time will be any different to what was stored by the original owner.

DNA Digital Data Storage

Nucleotides are organic molecules which form the basic structural unit of all DNA. Naturally occurring DNA found in living organisms is restricted to certain templates of nucleotide combinations since these templates represent functions which are necessary for life, such as how to create energy from glucose. It is possible, under lab conditions, to chain together nucleotides in arbitrary orders that do not necessarily match anything found in nature. It is even possible to create new synthetic nucleotides using base molecules outside the standard ACGT bases, upon which all naturally occurring DNA depends exclusively on. Since synthetic DNA can be made to take on any combination of nucleotides desired by the scientists creating it, it stands to reason that our digital data can be encoded within it.

DNA data storage promises unprecedented storage by volume rates at a staggering 1 x 1018 B/mm3 – over 1,000 times the information density of the most advanced magnetic tapes. It also supports the essential data operation (although so far only shown at small scales) of random access retrieval. This means it is possible to start the data reading process at any point along the DNA without having to go through the entire DNA sequence every time you want to retrieve a piece of information. And while not being nearly as durable as 5D optical storage, DNA storage still has a potential lifetime on the scale of millenia, far exceeding current hard drives which struggle to reach a single decade without any deterioration.

The Not-So-Distant Future of Data Storage

Jack O'Connor

DNA storage is unique in using biological reactions to store information and it comes with its own unique challenges. When reading or writing information to DNA storage it is not at all uncommon for bits to become corrupted as the necessary chemical reactions take place. For now this problem is not an issue as due to the astronomically high information density of DNA one can simply store several copies of each piece of information without worrying about efficiency. This is not a particularly satisfying solution though and it is to be expected that as the technology matures these errors will be reduced. While the biological nature of DNA storage presents unique challenges, it also comes with new opportunities. Recent studies have shown it is possible to encode the phrase “Hello world!” into the DNA of living bacteria. Better yet, this message can then be passed onto succeeding generations of that bacteria. This would be a hugely novel approach to creating backup copies of data as bacteria could simply be left in a petri dish by themselves and no more action would need to be taken to create copies of the desired data.

Conclusion

Both of the aforementioned technologies sound like they were ripped straight from a science fiction novel and it is very exciting to think that they may someday become the standard. But it would be remiss of me to not point out that the near future of large-scale data storage may just be an improved version of the hard drives we already see today. Hard drive technology has decades of research and development behind it as well as tried and tested production lines already in place. This head start over 5D and DNA storage will not be going away anytime soon. Even though 5D and DNA storage are so much more space efficient than hard drives, it is not unless they also become more cost efficient and practical that we will begin to see these technologies become commercially viable and used by organisations to store archival data.

References

DNA: https://www.biorxiv.org/content/10.1101/114553v1.full.pdf
https://www.sciencemag.org/news/2021/01/scientists-program-living-bacteria-store-data
5D: https://www.makeuseof.com/tag/5d-data-discs-can-outlast-sun-whats-catch/
https://www.mdpi.com/2072-666X/11/12/1026
https://www.researchgate.net/publication/312605376_Eternal_5D_data_storage_by_ultrafast_laser_writing_in_glass
https://news.microsoft.com/innovation-stories/ignite-project-silica-superman/?utm_source=stories&utm_campaign=1639
https://www.youtube.com/watchv=6XnlTwuION0&feature=emb_title&ab_channel=InsideHPCReport
https://www.youtube.com/watch?v=cgWZ_g0BkeE&ab_channel=Seeker 1:19 for a picture of the femtosecond laser

Interview with Dr Michael Cronin
Clinical Trial Statistics

Dr Michael Cronin is a lecturer in statistics in UCC. He is particularly interested in oral health research. He is involved with clinical trials in the University Dental School and Hospital.

How did you become involved in oral health research?

At the time that I did my BSc, you did a double honours degree and you chose two subjects - maths and statistics, maths and applied maths, applied maths and statistics, or I think some people also chose statistics and computer science. It was kind of the precursor to what you now know as mathematical sciences. I did maths and statistics.

After graduating with my BSc, I started a master's in UCC. It was a one year program similar to the master's by research that we run at the minute, but towards the end of it, I got an opportunity to work as a research statistician in the oral health services research centre, which is attached to the University Dental School and Hospital. I worked there for what was supposed to be two months, but I think it worked out to be a year in the end, with the result that my master's was put on the back burner. I don't advise anyone to do this, but I did put it on the back burner for what turned out to be five years.

While that job was still going, a role came up in Cork in a company that was analysing clinical trials. I applied for that and worked in that job for five years. During that time I realised that I really liked working in clinical trials. I liked research and I liked working in oral health, which I had been doing in the research centre. I was in contact the whole time with the directors of that research centre. At the time there were very valuable health research board scholarships available, I think the scheme was relatively new. I applied for and successfully got a HRB fellowship to do a PhD in the area of survival analysis of tooth fillings.

Most people have one or more fillings in their teeth, and those fillings don't last forever. There was a very large database available that recorded when a person got a filling, when a person was examined again and again, or perhaps got the filling replaced. There was a vast amount of data there. It was government data and I put together a proposal to use survival analysis, which is a technique in statistics, to see what are the factors that influence the survival of your fillings.

Interview with Dr Michael Cronin
Clinical Trial Statistics

What does being a clinical trial statistician involve? Is the statistician just involved in the data analysis at the end?

The statistician gets involved much earlier [than the data analysis in the clinical trial] because a clinical trial needs to be designed. An aspect that a statistician would have input into would be the sample size - how many patients do we need in our clinical trial. There's no point in running a clinical trial where you have too few patients to show the thing you want to show, and at the other end of the scale, there's no point having too many patients in your clinical trial.

Another aspect that comes into that then would be the length of time that you need to run your clinical trial. For example, in dental restorations or dental fillings, there's no point in having a clinical trial for six months because no fillings will fail within six months, and so you won't have any data.

Another aspect that a statistician would be involved in would be determining the primary measures. You might want to, for example, show that your headache tablet is good. That's a very abstract concept. We need to measure how you're going to demonstrate how your tablet, your pharmaceutical product, or your medical device actually works. A statistician has a lot of input into what are called the primary outcomes. These are the variables that are measured and used to determine how well the product - it could be a drug, it could be some other kind of intervention, or it could be a medical device - works.

The statistician would also have input into how the data will be collected. Again, it seems a very straightforward thing, how will we collect the data, but data needs to be recorded, then it needs to be coded, verified, and validated, and so very often the statistician gets involved in that aspect. This would happen before the data gets collected at all. Some research units would have people dedicated to data management, this would be where they design either paper forms or electronic forms to capture the data.

There would be validation of the data. For example, if you were recording someone's age and if you had in your clinical trial some criteria whereby your patients must be between the ages of say, 18 and 65, which is a typical age range, then you would build in checks that the data that is being recorded on age fits within that. You're checking that you haven't entered an illegal patient, or that you haven't mistakenly recorded that they were 16 when in fact they were 26.

.

Interview with Dr Michael Cronin
Clinical Trial Statistics

The statistician in the clinical trial has a lot of input into the design stage, and then at the other end, when the data has been collected, the statistician then has to analyse that data, but before analysing the data there would be a lot of reviewing of the data to make sure that what was collected makes sense, that you haven't collected illegal values. Then you would analyse the data as per the protocol. The protocol is the plan for a clinical trial, and because most clinical trials have to be registered and are subsequently submitted to the likes of the FDA or the EMA, it's important that you have everything down on paper before it happens, so that there's no suggestion that you were doing things or making decisions based on the data you were collecting.

It's a highly regulated environment, so you analyse the data according to the protocol, you present the results in a report and very often then, particularly if it's a clinical trial within academia, you'd be involved in the preparation of the publications associated with the research. In the middle of the process you might also get involved in assessing patient recruitment, because very often, because you're dealing with people, things happen and you might need to revise the sample size, or you might need to revise the data, the particular questions that you're asking, or the particular observations that you're making. A statistician has a lot of input into the whole process of the clinical trial.

Has the work of statisticians in clinical trials changed much over recent decades?

If we go back maybe thirty years, to the advent of statistical software, in clinical trials the methods of analysis haven't changed much. That's because of the regulated environment you're working in. The regulatory agencies such as the FDA and the EMA and other, local or regional entities, are slow to embrace very new methods. So you would find in a clinical trial the methods have been used for ten, twenty, thirty years. In fact, some of these methods are methods that a lot of undergraduate students would be familiar with, so while there's a lot of advances in statistics as a discipline, it's fair to say that in clinical trials they're always lagging slightly behind.

When you're using software, the software purports to do something, it purports to run a particular method. But a lot of software, particularly R, is not validated. You can use thousands of packages that people around the world, particularly academics, develop to do certain things in R. Nobody has checked these robustly, and with the result that the regulatory agencies are reluctant to trust these methods. Whereas, if you're using a validated software, something like SAS, that's something that we don't teach at undergraduate level much, that package is all but recommended by the FDA for example. It's recognised that because it's a commercial package, and it's a very expensive package, it does what it says it does.

Interview With Dr Michael Cronin
Clinical Trial Statistics

When you run a particular routine, or a proc as they're called in SAS, you can be confident that it has been validated and that it has been tested robustly, it's been tested with unusual data, or it's been tested with challenging data, outliers, missing values, etc. Packages in R may not have been tested as robustly as that.

What was the most interesting clinical that you have been involved with?

Some are more challenging than others. If you have a clinical trial where it is necessary to follow people for a longer period of time, then people drop out of clinical trials. People move away, people lose interest, people don't follow the treatment that they're supposed to be given or do the things that they're supposed to be doing. Those clinical trials as a result become quite challenging. Generally the longer the clinical trial, the lower the quality of the data that you get at the end, through no fault of your own. It's just human nature, patients lose interest. When I worked in industry for five years, there were lots of different types of clinical trials.

I think the type of clinical trial that was most interesting involved generic drugs. A generic drug is where some company has developed and patented a particular formulation, a particular drug that does something, and when these drugs run out of patent, which I think is after ten years, then any other company can replicate and sell those drugs - if they're able to replicate the drug. But, to do that, you have to show that your product does the same thing as the original drug. These generic drugs need to be tested in a different way to the original drug, in a method called equivalence or bioequivalence. This is a concept whereby there's no point in trying to show that your drug is better than the market leader, the market leader is the market leader for a reason, it's obviously doing a good job. You can't show that your drug is better. Instead, you need to show that your drug is equivalent. It's coming at it from a different angle, instead of starting off trying to show that you're better, you assume that it is worse and you have to prove that it's as good, or equivalent. Equivalence testing is an interesting area of statistics.

How safe is your communication?

Laura Cosgrave

Every day billions of messages are sent between people over the internet. From confidential business details to private thoughts, this personal data is protected to different degrees. How do you stop people other than the intended recipients from reading your messages?

There are a few different options. One, don't encrypt it and let everyone read your messages. Two, encrypt the message while it is in transit from you to the messaging service provider's servers and then from the servers to the recipient. A third option is end-to-end(E2E) encryption, where the data is encrypted all the way from your device to the recipient's device, and nobody in between can access it.

In the previous edition, an article discussed the debate in the EU over forcing messaging services with end-to-end encryption to insert a back door. But what does end to end encryption mean? In what other ways can your data be protected? Much of what I am saying is applicable to many different types of data, but for simplicity, I will focus on messaging services such as Whatsapp and Messenger.

What is encryption?

Encryption is the process of taking 'plaintext', or readable, data and scrambling it using a key and an algorithm. A key is a piece of information used to encrypt or decrypt data, and the algorithm is the process used to encrypt or decrypt the data. One big problem in the history of encryption is key exchange - how can both parties have access to the key(s) needed? Historically, governments hired couriers to physically bring keys to people they needed to communicate securely with, but this is expensive and doesn't scale well.

How safe is your communication?

Laura Cosgrave

Solving the problem of key exchange

In the 1970s, two methods of key exchange were invented. One was the Diffie-Hellman key exchange protocol, which involves both party A and party B each choosing a secret piece of information, putting it through a non-reversible mathematical function, and sending the results to each other. By combining the result that party A sent with the secret piece of information party B has, or vice versa, a secret key is generated that is the same for both parties. If the exchange is intercepted, it doesn't matter, because you need one of the secret pieces of information, which the interceptor doesn't have, to get the key.

The other way was through asymmetric encryption. In symmetric encryption, the same key is used to encrypt and decrypt the messages. In asymmetric encryption, however, different keys are used to encrypt and decrypt the encryption. For example, I might share a public key with someone, which they can use to encrypt a message for me. It doesn't matter if someone intercepts my public key - only I have the private key that can be used to decrypt the message.

So how do Whatsapp and Signal encrypt messages?

Let's say Alice and Bob want to send E2E encrypted messages to each other. They initiate a session, which lasts until one of them gets a new device, or something else goes wrong. Both Alice and Bob generate a number of public keys and private keys. They share the public keys with each other and keep the private keys secure on their devices. Alice combines her private keys with Bob's public keys using an algorithm to generate a 'master secret'. Bob combines his private keys with Alice's public keys to generate a corresponding 'master secret'. If the public keys are intercepted, it's no problem - the master secret can only be generated using both the private keys and public keys - and the interceptor only has access to the public keys! Each device then generates a 'root key' and 'chain keys' from the 'master secret'. These steps are usually just performed once, at the start of the session.

Alice and Bob each have a 'sending chain' and a 'receiving chain'. Bob's receiving chain matches Alice's sending chain, and vice versa. Each time Alice wants to send Bob a message, a new key is generated from her sending chain. This key matches the next one generated from Bob's receiving chain. To add an extra layer of security, each time Alice sends Bob a message, or vice versa, a Diffie-Hellman public key is sent with it. This means that new receiving and sending chain keys are derived each time a message is sent. This means that if I discover Alice's receiving chain key, I won't be able to derive any future or past keys, so I won't be able to read any of her future or past messages. This is called 'perfect forward secrecy' - and not all E2E encrypted messaging services have it.

How safe is your communication?

Laura Cosgrave

How do Gmail, Messenger, etc encrypt messages?

Many messaging services use Transport Layer Security. This means that your messages are encrypted going from your device to the messaging service's server, and from the server to the recipient's device. This means they can't be read by any servers or networks they pass through, or if they are intercepted. The data is usually separately encrypted 'at rest' too, so it can only be accessed by those with the encryption keys (i.e. the service provider). However, the messaging service provider can still read your messages. Furthermore, they may be required to share unencrypted copies of messages with law enforcement.

What are the Advantages and Disadvantages of End-to-End Encryption?

The main advantage of end-to-end encryption is that any servers the message passes through cannot read the data being sent - to them, it is just meaningless nonsense. If it wasn't E2E encrypted, it could be read by the messaging service provider, which could use it for many nefarious reasons or sell it to third parties.

Another advantage is that you can guarantee the message hadn't been tampered with - if someone tries to edit the encrypted message when it is decrypted it will just be gobbledygook. This is important for data integrity - you know that you are reading exactly what the sender has written.

End-to-end encryption is not perfect - it only protects your messages while they're going from the sender's device to the recipient's device. If someone gains access to your device, they will still be able to read all of your messages. You also must trust the recipient not to share the messages and to keep their device secure. E2E encryption only protects messages in transport, they must be separately encrypted if stored on the recipient's or sender's devices, or backed up elsewhere. It is a good idea to periodically delete messages to reduce the impact of someone accessing your device.

Another big risk is man-in-the-middle attacks. E2E encryption ensures your data is secure from endpoint to endpoint, but what if the message is sent to the wrong endpoint? One method of man-in-the-middle attack is substituting the intended recipient's public key with one of the hacker's public keys. Endpoint authentication is used to guard against man-in-the-middle attacks, such as using certification authorities or comparing 'fingerprints' generated using the sender and recipient's public keys.

How safe is your communication?

Laura Cosgrave

A potential disadvantage (or advantage) is that law enforcement can't get access to messages. Law enforcement can subpoena messaging companies to get access to the messages of potential criminals, but if the messages are end-to-end encrypted they can't decipher the messages to read them. However, some information is still available for authorities to access - depending on the messaging service, they can access information such as who the person has been messaging and when.

What's a back door - and why do people want one?

The encryption vs law enforcement debate has been going on since encryption became available for use by non-governmental organisations. It essentially comes down to the question: which is more important, the right to privacy or the ability of law enforcement organisations to intercept the communication of potential criminals? There has been a lot of debate over whether or not law enforcement should be allowed to access people's messages - in services that are not end-to-end encrypted, messages have been subpoenaed, but with E2E encryption, there is no way to read the

Whatsapp's white paper on end-to-end encryption defines end-to-end encryption as ' communications that remain encrypted from a device controlled by the sender to one controlled by the recipient, where no third parties, not even WhatsApp or our parent company Facebook, can access the content in between.' Some countries including Australia, the US, the UK, and others have called for a 'back door' into end to end encrypted messages -essentially they want law enforcement to be able to access them in readable form, although then these messages are then by definition not actually end to end encrypted (although this doesn't mean the messaging service provider won't say messages are E2E encrypted, even if a backdoor exists).

Essentially, many governments want to ban end-to-end encryption, so that they have the ability to read all messages. This would also however leave the messages far more vulnerable to misuse and access by malicious actors. However, even if E2E encryption was banned, this wouldn't necessarily stop serious criminals from using it. Which is more important, the right to privacy or law enforcement's ability to access potential criminals' messages? That is the question that the world is grappling with.

References

Whatsapp white paper
https://policies.google.com/terms/information-requests
https://signal.org/docs/
https://www.justice.gov/opa/pr/international-statement-end-end-encryption-and-public-safety#_ftnref1

Misinterpreted COVID-19 Data in the News

Mark Healy

Data is the future of journalism. Although Data Science is a relatively new term, it first came about in 1962 when John W. Tukey released his book “The Future of Data Analysis”. It is widely known that data is seen all around us, but what about in the news? According to RTE, almost 4 million people tuned into the RTE news service between 2nd March 2020 and 17th May 2020. Of course, this was around the time when COVID-19 was coming to the forefront and some people's eyes were glued to the news throughout the day to learn about lockdowns and restrictions around the world.

However, this opinion piece looks at how data is used in the news and specifically how data was reported at the beginning of the COVID-19 pandemic. At this stage we are used to hearing about case numbers and death numbers and recently, even vaccine numbers, but these numbers are sometimes misinterpreted to mean something slightly different than they actually do. For one, death numbers are in fact reported death numbers, not death numbers recorded on a specific day like case numbers.

For example, if 14 deaths are reported on a specific day, this does not mean there were 14 deaths that day, but rather there were 14 deaths recorded or reported on that day. Of course, this makes sense as it would be virtually impossible for deaths to be reported the minute they happen.

Misinterpreted COVID-19 Data in the News

Mark Healy

Another excellent example of how the death rate during the initial months of COVID-19 can be taken out of context is a table posted by the Department of Employment and Social Protection (and reported by thejournal.ie) in July 2020 which showed a total of 717 deaths in July of that year. Many people commented at how the death rate in previous years in July was 2000+ but again the July 2020 number of deaths is the number reported at the time of the report, rather than the actual number of deaths that month. It can take up to three months to register a death in Ireland and this is the reason for fluctuating and "incorrect" numbers.

The reporting of "incorrect" data often leads to more har, than good. Not that these figures are necessarily wrong but they do not represent the whole picture. It takes time for this to happen. Data that is not interpreted correctly can lead to some disastrous consequences including being the foundation of events like protests (Personally I don't come down on either side of whether protests in relation to covid are a good thing or a bad thing, it is the facts that this article is about).

Data can often change over a period of time and due to this change it can appear to portray a better or worse situation. The amount of COVID-19 cases is one of those numbers that dictates whether we are in a good or bad situation in terms of COVID-19. By the end of March 2020 Ireland had conducted just over 30,000 tests for COVID-19 and had a positivity rate of 15%. The positivity rate is as important if not more important than the number of cases as it dictates whether the virus is spreading or being suppressed. Of course the total number of tests has increased as well as the daily number of covid cases. In March 2021, Ireland was conducting more than 100,000 tests a week. The positivity rate in March was in and around 4% with obvious small changes throughout the month.

A major consequence that has emerged from the rapid increase of testing is that of course the number of positive tests increased. That makes sense, if you test more people there is naturally going to be an increase in positive tests.

Famously, former US President Donald Trump, regularly pointed to the increased testing as the reason for more positive cases against all evidence that said the contrary. In fact, COVID-19 was rapidly spreading throughout the communities during this time in the US. Or in other worlds the positivity rate was drastically increasing. This example most certainly shows the importance of the positivity rate number and its importance to showing the spreading or suppression of the virus.

In this piece, we have taken a brief look at how data is reported in the media, specifically during the COVID-19 pandemic. It is always important to know where your data is coming from and to have reliable sources of information. Some of these sources for COVID-19 related material include the gov.ie and HSE websites amongst many others.

Data Science and Analytics Roundup

UCC's Professor Barry O'Sullivan was awarded Nerode Prize

Police use car data to destroy alibis

Python top programming language while R popularity rises

EU's rights watchdog warns of risks of use of AI in predictive policing, medical diagnoses and more

App alerts users about data collection by IoT devices around them

New research highlights the impact of the digital divide

A new way to visualise biological data

Topological data analysis can help predict crashes

MIT study finds 'systematic' labeling errors in popular AI benchmark datasets

Watch, Read, Listen

The Code Book

Simon Singh's bestselling book gives a good overview of encryption.

Dublin Data Science: Soup to Nuts - Workshop 1

Mick Cooney presents a series of workshops working through a data science project.

Hannah Fry: The Role of Algorithms

In this episode of The Knowledge Project podcast, Dr Hannah Fry discusses the role of maths in society and what it means to be human in the age of algorithms

On Digital Literacy

Niall McCarthy

Undisclosed Location,

2021

Dear Reader,

I hope you are keeping well in these unprecedented times.

I write to you, in great alarm, that while we have used this horrific worldwide event to reassess our approach to many institutions - healthcare, hospitality, travel - we seem committed to our outdated and cookie-cutter approach to education and literacy.

Those damn kids these days couldn't write a letter to save their lives!

Power to them, there is little they will need letters for in their lives. In this day and age, I cannot think of anything that cannot be replaced with a digital equivalent, stored securely.

In a time period where we obsessed with the "New Normal", and change, we haven't quite taken the time to reflect on how we need to prepare people for a world that is mostly online. We talk about time spent on phones as if it is in disharmony with the natural order of things, then praise efforts to bring in more jobs that will involve much of the same activity in processing something via computer all day.

We currently make the students literate for a bygone society. I do not mean to belittle the skill of handwriting or its importance, but we are not liberating young people in focusing on a transition from pen to pencil, and from block to cursive with their writing, rather than from two-finger to all finger to touch typing. The knock-on effects of this are quite real and material.

For a large period of time, the technological revolution had a contradictory place in the discourse, and this led to a society that continued to increasingly expect technological knowledge from its workers, while the idea of literacy remained reading, 'riting and 'rithmetic. Much ink has been spilled over the class element in education, best put by David I. Backer in 'The False Promise of Education'. However, we do not need to go this deep to see clear issues staring us in the face.

There are many lessons to learn from this pandemic regarding our digital lives, I will try to briefly outline the main considerations we should be making, and how these relate to data science.

On Digital Literacy

In reading up on literacy, I noted that 'Digital Literacy' is still a young term. Despite the identified need for a program such as ECDL back in 1995; the Hamburg Declaration on education for UNESCO in '97, in points 19 and 20 identifying that key points in our approach to education must factor in "Transformation of the economy" and "Access to information"; the proliferation of technology - we have not truly updated our idea of literacy.

The only detailed account for the idea of 'Digital Literacy' I could find is from the Swiss linguist Gunther Kress, who views literacy in the way I think we should in his work 'Literacy in the New Media Age' - "Literacy, in all its aspects, is entirely social, cultural and personal" - which I doubt anyone would feel the need to argue with.

Our social, cultural, and personal experiences are increasingly being experienced through our screens. Our work is, too. We need to reflect this in our societal actions.

We must recognise some of the atrocities of the last year as extensions of great illiteracy. Denial of science - from both public and policymakers - has caused thousands of needless deaths. The lack of knowledge of online scams and disinformation has hurt many of our most vulnerable. We must hear the phrase 'do your own research' among conspiracy theorists and realise it is a consequence of literacy being critiquing Shakespeare and solving maths problems for little reason other than to get marks.

Most shockingly, many of the most 'literate' and skilled in charge of merely tracking cases, in Ireland and the U.K. were caught in not knowing the limitations of a spreadsheet. To harp on that point, many well-educated and assumedly quite literate people oversaw systems that have potentially led to preventable deaths because they cannot discern file formats and data-logging. Cynically, I assume that these people almost definitely could discern between formal and informal letters they have not written in years - but do not know the file types they deal with every day.

If we, as the beneficiaries of the place technology and science have in the current market, do not share our work and help more to understand and have input while we are relied on more for decisions affecting the lives of the public, we will continue to perpetuate this illiteracy. We are the proprietors of this age's Gutenburg Press, but merely printing notes to pass to each other.

Yours faithfully,

Niall McCarthy, PRO