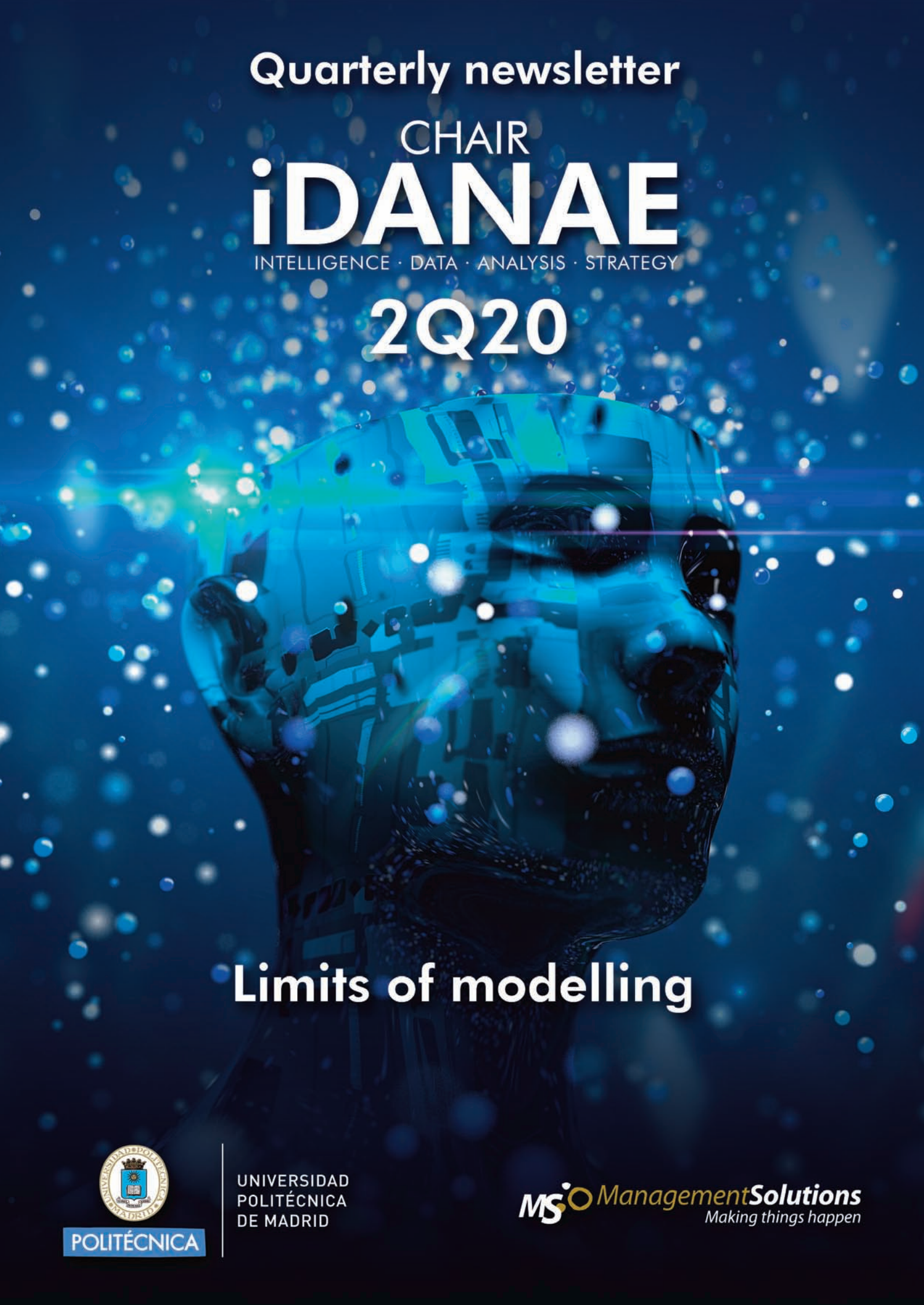
Quarterly newsletter

CHAIR iDANAE

INTELLIGENCE · DATA · ANALYSIS · STRATEGY

2Q20



Limits of modelling

UNIVERSIDAD POLITÉCNICA DE MADRID

ManagementSolutions
Making things happen

Analysis of meta-trends

The iDanae Chair for Big Data and Analytics (where iDanae stands for intelligence, data, analysis and strategy in Spanish), created within the framework of a collaboration between Universidad Politécnica de Madrid (UPM) and Management Solutions, aims to promote the generation and dissemination of knowledge, the transfer of technology, and the furthering of R&D in the Analytics field.

One of the lines of work developed by the iDanae Chair is the analysis of meta-trends in the field of Analytics. A meta-trend can be defined as a value-generating concept or area of interest within a particular field that will require investment and development from governments, companies and society in the near future[1].

This report is focused on the limits that exist in the modelling of a real system.

[1] For further information, please refer to iDanae, 2019.

Introduction

Artificial intelligence is based on the development of mathematical models capable of simulating the behavior of a system through the discovery of patterns[2]. This process of modelling is not unique to artificial intelligence, but is also used in other sciences such as physics, chemistry, economics and medicine. Usually, the mathematical representation of a real system cannot be complete[3] due to the large number of parameters and interactions involved, and therefore it is necessary to work with idealized models that simplify reality, which generates some limits in its representation capacity. These limits will be bigger the more complex the system to be modelled or the more simplified the model used. This has repercussions on the modelling capacity of artificial intelligence systems, where normally simple models are used, for reasons of interpretability, while attempts are made to model highly complex systems, resulting in a limitation in their predictive capacity.

The degree of development and use of artificial intelligence techniques nowadays is very different depending on the field of application, the sector, or the geography, among others. Four main causes of these differences can be highlighted: (i) the uneven interest, even within each organisation, in the possible applications in which artificial intelligence models and systems can be used; (ii) the difficulty in obtaining enough quality data to allow the correct development or operation of the algorithms; (iii) the limitations of interpretability, which lead to the simplification or discarding of certain algorithms; and (iv) regulatory limits, which may rule out some methodological options.

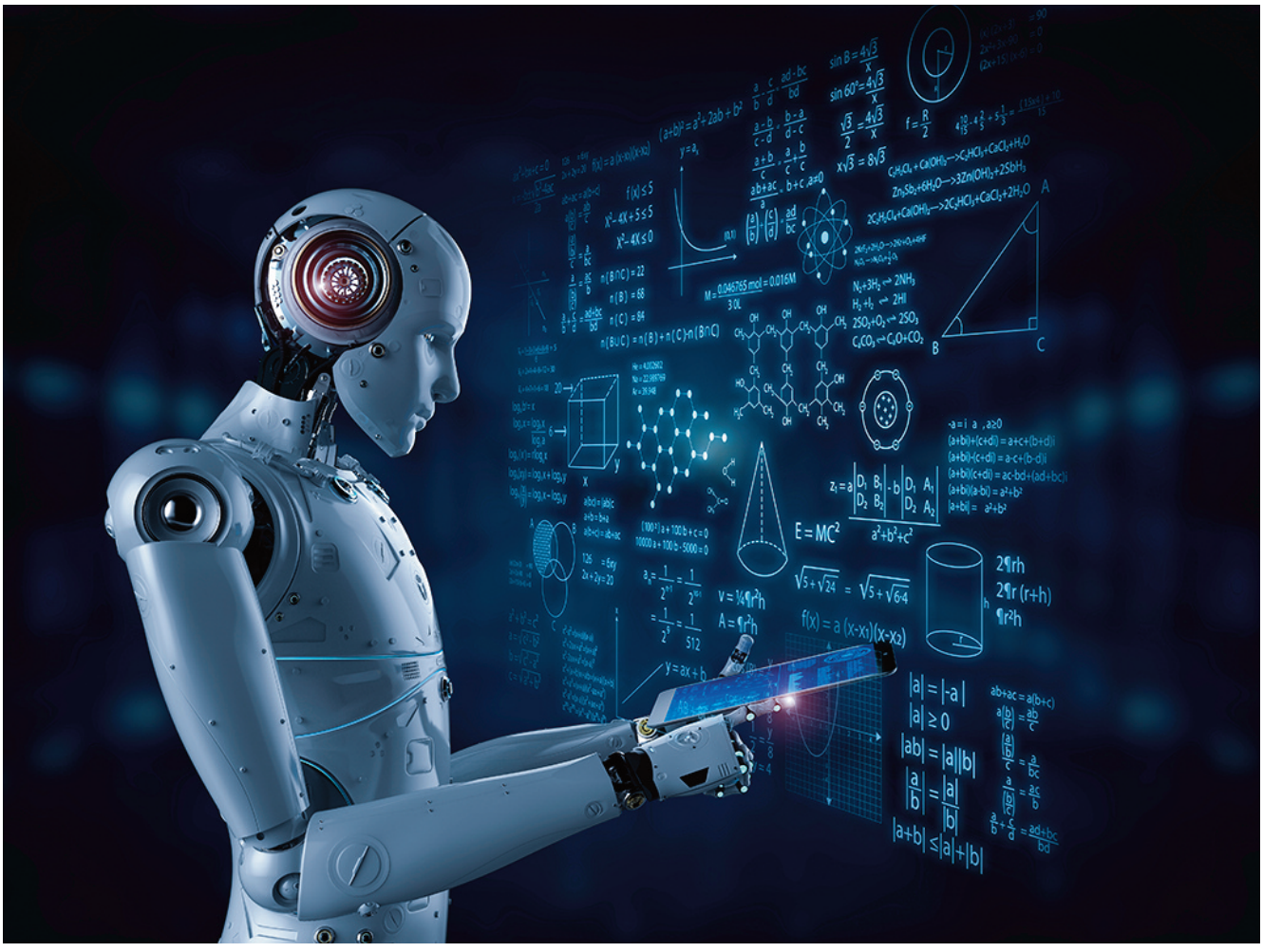
Behind all these reasons there is a common element: the inherent difficulty of the problem posed which, depending on the assumptions and complexity of the model formulated, will generate both data and computational requirements and inherent limitations on the capacity of the model and possible problems of interpretability.

As an example, algorithms have been developed that are capable of winning over the best players in various games such as chess or GO, while the development of the autonomous car or the detection of emotions seem to be in a more initial state. This may be mainly due[4] to the nature and characteristics of the problem: in the first case, there is a defined set of rules, the elements to be evaluated are limited and, therefore, the simplifications of the model are reduced, and the cost associated with an error in this system is limited. However, in the second case, the number of inputs, the heterogeneity in their nature, the complexity of their interpretation and the cost associated with an error in the prediction of this type of system is much higher. In this context, problems related to the scarcity

or quality of data, the need for greater computing power, or challenges in interpretability, are constantly assessed and both the sources of information and the algorithms and systems are improved.

The usual way to try to overcome these limits is through substantial modifications to the model used, or by including new variables that provide more information. Even so, it is necessary to consider the limits that come from ethics: these appear in situations where an artificial intelligence model generates ethical dilemmas related to the use of personal data, biases or errors in decision-making[5].

[2]Frigg, Roman, Hartmann, & Stephan, 2020.
[3]Podnieks, 2010.
[4]Stipić, Bronzin, Prole, & Pap, 2019.
[5]iDanae, 2020.



Finally, it is worth noting the limitation regarding the difficulty of interpreting the model. It is currently possible to have a complex model capable of solving a problem that the human being cannot understand the process by which a concrete prediction is reached, which makes the use of these models difficult in certain circumstances (for example, when there is a regulation that requires this interpretability[6]).

This publication will therefore focus on the limits of modelling inherent to the type of problem being addressed in each case. These limits have been conceptualized in two types: those of theoretical origin and the limits generated by the current development context, which are limitations of practical origin.

In any case, this classification is made for expository purposes, since both groups are strongly related: when a practical limitation arises, it is possible to revise the hypothesis on which the mathematical model used has been built, so that a new approach can be taken that can overcome that limitation. On the other hand, when a theoretical limitation arises, it is sometimes possible to overcome it through practical solutions, such as the use of greater computing power.

[6]iDanae, 2019.

Theoretical constraints

Complexity classes

When using an algorithm to model the observable behaviour in a data set, the training of the model usually occurs in an iterative way: in each iteration improvements are added with respect to the previous one (for example, in terms of the error made by the model). When the process converges, for a large number of iterations the improvement between one and another iteration is smaller and smaller, establishing a higher level that a model cannot surpass in terms of predictive power. This higher level on the ability to solve a problem through algorithms is part of the theory of computational complexity[7], which studies the inherent difficulty of computer problems and the classification of these difficulties. The resources needed to solve a problem through an algorithm can be temporal (how much execution time is needed to solve the problem) or spatial (how much storage you need). Complexity analyses how many of these resources are needed to solve a problem, and how the number of operations required increases as the size of the input data increases. This allows to categorize problems in classes, according to their different complexity, being the most known the P class (when a problem can be solved efficiently[8]), the NP (when there is not known an algorithm that solves the problem efficiently, but a solution can be verified efficiently), or the PSPACE (the spatial analogue of the P class). The relationships between the complexity classes are not clear (Figure 1), the most famous being the P=NP[9] assumption.

Complexity classes directly reflect a limit on the ability to model systems so that they can be solved efficiently, regardless of computational capacity.

Correlation and causality

Correlation[10] is a statistical technique used to estimate whether there is any correspondence between any two variables. This relationship can be of different types: linear, quadratic, logarithmic, exponential, etc. Correlation does not explain why this relationship exists or how it occurs, nor does it have to imply that one variable is the cause of the other, regardless how strong the relationship between the two may be.

[7]Papadimitriou, 2003.
[8]That is, there is an algorithm that solves the problem in a polynomial time.
[9]This is one of the millenial problems (Clay Mathematics Institute, s.f.).
[10]Rodgers & Nicewander, 1988.
[11]BGSMath, 2020.
[12]Hanza, 2020.

Figure 1: *relationships between different classes of complexity[11]. Although it is known which classes of complexity contain others, it is not yet known whether these classes can be equivalent or not.*

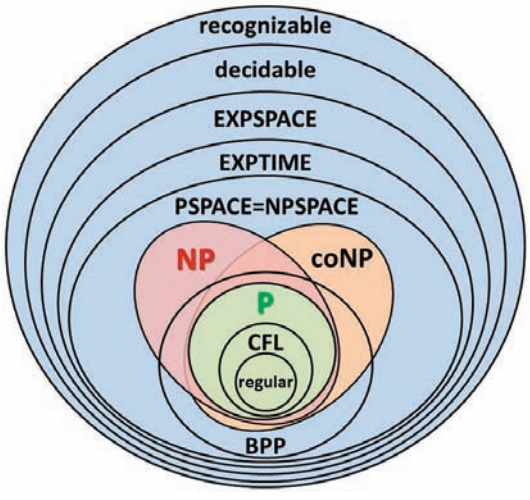
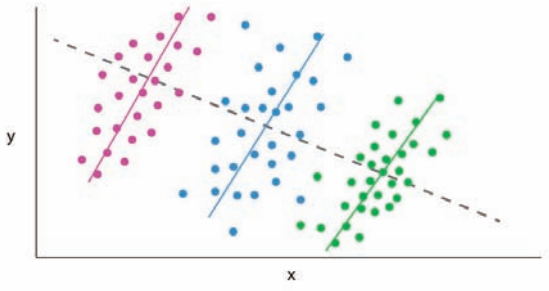


Figure 2: *Simpson's Paradox[12] . As can be seen, by studying the correlation between variables X and Y, an inverse proportionality relationship is obtained, i.e. the more X increases, the more Y decreases. However, if a third variable is taken into account, such as colour, which allows the data to be grouped, a positive correlation is obtained, where an increase in X implies an increase in Y.*



This is important, as there may be situations where two variables are correlated and, by including the effect of a third variable, the correlation is opposite. This often occurs in social and medical research and is known as the Simpson's Paradox (figure 2).

Causality[13], on the other hand, establishes a cause and effect relationship, making a change in the cause variable implies a change in the effect variable. As in the case of correlated variables, there can be a causal relationship between two variables without any correlation between them.

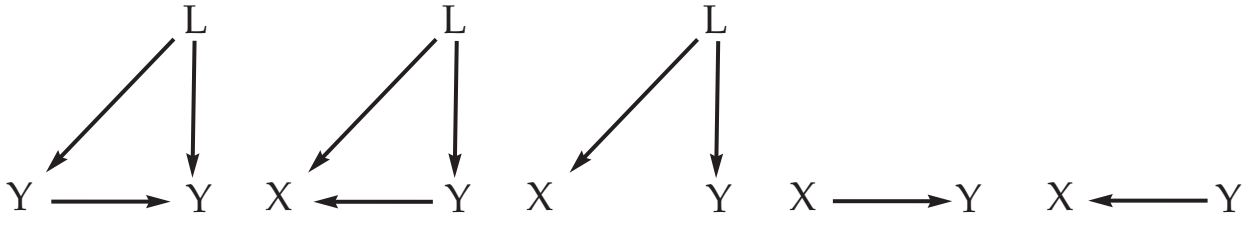
The conditions[14] for speaking of causality between event X and event Y involve the assumption of the Markov condition, which establishes that an event is independent of all events not directly related to it; the condition of fidelity, which establishes that the independence between events is due to the Markov condition and not to other possible effects; and the assumption of causal sufficiency, which establishes that there is no common cause in the system that has not been measured. In the event that the latter assumption is not fulfilled, other cases may occur, such as those shown in figure 3, where L is an unobserved event that generates spurious correlations between observed events.

The current techniques used to develop artificial intelligence models are specialized in finding patterns in the data, that is, correlations. However, it is very important to be able to discern whether the correlation found is also a causal relationship, since a model based on correlations will have much less predictive capacity in situations that are very different from the data it has been trained with, since the generalization of the model is much more limited. A model based on causal relationships, on the other hand, will be able to correctly predict situations on which it has not been directly trained.

The above limits, which are strongly theoretical, establish an upper limit above which it is not possible to design more efficient algorithms even in an idealised situation where all possible data are available and their quality is optimal. These limitations, either because of their complexity or when establishing relationships between various effects, prevent the design of a model that exactly replicates the system, preventing efficient resolution.

[13]Pearl, 2000.
[14]Eberhardt, 2009.

Figure 3: *possible causality options, in which the assumed case may occur (X causes Y), but in which it may also be the case that Y causes X, that X and Y are caused by L, or that X causes Y (or vice versa) only when L occurs.*



Practical constraints

Having studied the limitations that exist at the theoretical level, it is also interesting to study the limitations in modelling in a practical situation. The fundamental problem arises both because of the nature of the problem to be solved (which influences the degree of perfection achieved) and because it is impossible to model all the variables that act in a system, because the number is too large, because it cannot be measured correctly, because the existence of a variable is unknown, or because the quality of the data obtained is not optimal. All this contributes to uncertainty in the estimation.

Current degree of improvement of artificial intelligence models

Different practical problems can be analysed according to their degree of evolution. Arvind Narayanan, professor at Princeton University, establishes three levels of difficulty in the tasks performed by an artificial intelligence system[15], with a different degree of perfection:

- **Perception tasks - rapid progress:** this includes the tasks of content identification, facial recognition, medical diagnostics by image analysis and voice-text transcription. In this category, artificial intelligence is at a point where it has already reached, and even surpassed human capacity[16] and continues to improve steadily. This is because there is no uncertainty or ambiguity when performing this type of task, and data volume and computational capacity are key elements.

- **Automatic ratings - imperfect, but improving:** this includes spam identification tasks, detection of copyright violations, evaluation of essays, detection of hate speech or recommendation of content. In this category there are debates about the level of perfection that an artificial intelligence can achieve, because judgments are made that people may not agree on what the right decision is. Despite this, progress is quite satisfactory[17].

- **Predicting behaviour - unsatisfactory results:** this includes predicting the recidivism of criminals, job performance or predicting terrorist risks, among many others. Very little progress has been made by artificial intelligence in this category, and no further progress[18] is expected in the near future, given the inability to predict such issues for the future.

[15] Narayanan, 2019.
[16] Arcadu, and others, 2019.
[17] Google, 2019.
[18] Hao, 2020.

Scarcity of data or large number of parameters

Regardless of the subject matter to be modelled, another major problem that limits the predictive ability of an algorithm is systems where few observations are available with respect to the number of parameters to be modelled.

In this situation, the main problem is incurring discoveries of patterns in the data that do not really exist in a larger population. This is known as overfitting, where the main concern lies in the complexity of the model to be evaluated. This complexity is measured in terms of the degrees of freedom of the model, and sets a higher benchmark when it comes to modeling the data[19]. For example, if the number of degrees of freedom is the same as the number of observations, a model will be obtained that fits the observations perfectly, even if they respond to purely random behaviour.

On the other hand, the use of a model that does not have enough degrees of freedom may not fit correctly to the data to be modeled, generating, in an analogous way to the previous one, an underfitting problem. Mathematically, this translates into the terms bias and variance:

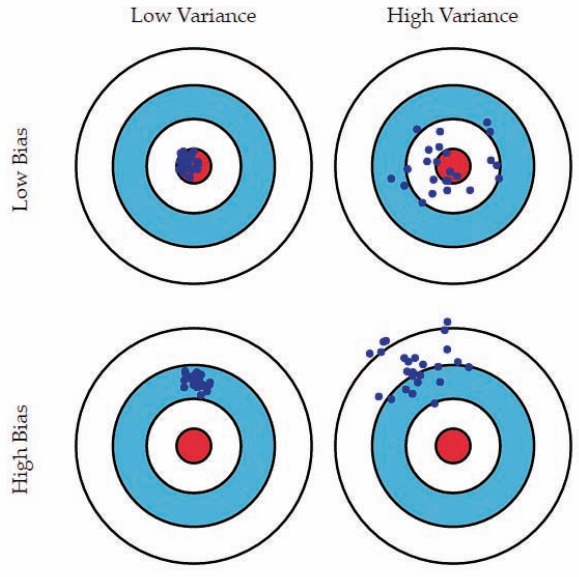
- Bias measures the difference between the model's average prediction and the real value .
- Variance measures the variability of the model's prediction when modifying training data.

A process of overfitting corresponds to a high variance and a low bias, while an underfitting process corresponds to a low variance and a high bias (figure 4).

On the other hand, there are the so-called chaotic systems, also known as complex systems, in which a variation in the starting data can imply great changes in the model output. In these systems, the problems lie in the conditions of stability. In the analysis of the application of partial derivatives to physics problem solving, Hadamard indicated that a problem is well defined if there is a solution, it is unique, and the behavior of the solution depends continuously on the initial values[21].

It can be extrapolated from this that the predictive capacity of a model will be reduced if some of the conditions mentioned are not met, which could result in the same initial configuration resulting in two totally different predictions. Many interpretability techniques, such as PDP charts or LIME methods, rely on these conditions to try to explain the output of a model[22], for example, by analyzing the variation of the output due to slight modifications of the initial values to study the effect of each variable on the prediction.

Figure 4: *results of a variance-biased model[20]. When a problem is complex to model and the hypothesis is complex, both theoretical and practical problems arise. If a problem is complex to model and a simplified working hypothesis is used, an underfitting case can occur, where the hypothesis does not fully capture the behaviour of the model. If a problem is simple, but a too complex hypothesis is used, a case of overfitting can occur. Finally, a simple problem treated with a simple hypothesis is what currently produces the best results.*



[19]Babyak, 2004.
[20]James, Hastie, Witten, & Tibshirani, 2013.
[21]Hadamard, 1902.
[22]For further information of this methods, please refer to iDanae, 2019.

A case study

This section aims to illustrate the aforementioned limits with a case study. Currently, the outbreak of COVID-19 has generated the need to develop projection models, with the aim of anticipating the pandemic and foreseeing its possible effects in order to determine mitigation and management measures. In this context, some limitations have been revealed when estimating some models for the Spanish geography.

One of the theoretical limits that found in the development of these models is the limitation in terms of sample sufficiency and its representativeness: in the field of genomics, this limitation can be a challenge when the number of available samples is quite small compared to the number of actual cases, which makes the generation of the models more complex.

On the other hand, the lack of knowledge about the virus may lead to the questions underlying the generation of models by researchers not being appropriate; likewise, the lack of certain types of relevant information means that not all the important variables are available, which may lead to correlations being found that do not necessarily imply causality.

Regarding the practical limits, the special circumstances at the peak of the pandemic implied that data collection by hospitals could not be performed with the usual level of completeness in traditional medical visits: on the one hand, an adequate management information system and a unified protocol for the centralization of information and its subsequent application in management were not designed and prepared; on the other hand, data that could have been of interest or variables whose importance in the modeling was unknown were not collected.

Therefore, despite the fact that the data currently available may be a good starting point, the existence of incomplete data may mean that the current information is not sufficient to generate statistically significant results or to draw conclusions that could be relevant in the management and fight against the pandemic.

Proposed solutions

One solution to overcome these limits is to improve computational resources by designing increasingly powerful processors[23] through the creation of new architectures and miniaturization, allowing more transistors on the same surface. In addition, to achieve more power, industry has typically turned to the use of supercomputers or, more economically, distributed computing solutions[24].

However, there are some obstacles in the short and medium term. Supercomputers have historically been very expensive to obtain and maintain, and it seems that miniaturization is reaching its limits, beyond which, due to the effects of quantum mechanics, transistors stop working in the same way. This is making distributed computing, thanks to the emergence of cloud computing, a solution to the problems of lack of computing power.

On the other hand, other proposals have emerged to solve the problem of not being able to continue increasing computational capacity through miniaturization. Some of these solutions propose the replacement of silicon by other materials[25]; the use of photons as a unit of information[26], instead of electrons; or even changes in computing paradigms such as the use of edge computing[27], where data processing is not centralized in the cloud but carried out in the devices themselves (Figure 5).

Other more innovative proposals regarding the changes in the computing paradigm include quantum computers, in which bits (with values of 0 and 1) are no longer used, but qubits, which are the superposition of the two values. This generates its own classes of complexity, different from those of classical computing, such as the BPQ[29] class. The relationships between classical and quantum complexity classes are not clear, but it is believed that there are problems that can be efficiently solved by quantum computers that cannot be efficiently solved by deterministic classical computers[30] (Figure 6). In 2019, Google, in collaboration with NASA, announced that it had achieved quantum supremacy[31], solving in 200 seconds a task that would have taken a classic computer more than ten thousand years.

[23]Schaller, 1997.
[24]Peleg, 2011.
[25]Passian & Imam, 2019.
[26]Caulfield & Dolev, 2010.
[27]Shi, Cao, Zhang, Li, & Xu, 2016.
[28]IEEE Innovation, 2020.
[29]BPQ is the quantum polynomial time complexity class with bounded error, which are decision problems that can be solved by a quantum computer in a polynomial time with an error probability of at most 1/3 for all instances.
[30]Nielsen & Chuang, 2002.
[31]See Arute, Arya, Babbush, & al, 2019. Quantum supremacy is the moment when a quantum computer performs a task that a classical computer cannot perform in a reasonable time. This announcement by Google generated a lot of controversy, as IBM, its main competitor in the area of quantum computing, announced that the test performed did not demonstrate quantum supremacy. However, the international community seems to have accepted Google's demonstration, Hartnett, 2019.
[32]Bellaiche, 2020.

Figure 5: *Edge Computing scheme of operation[28].*

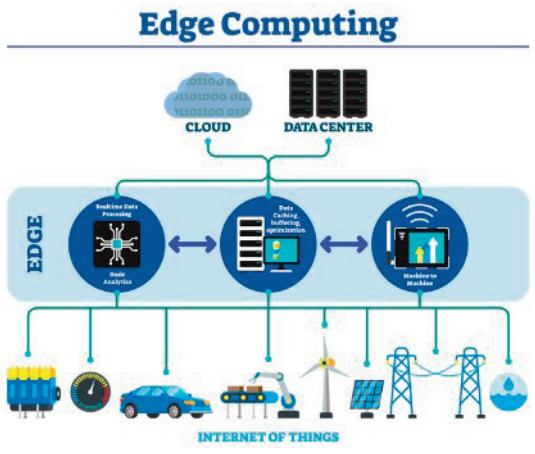
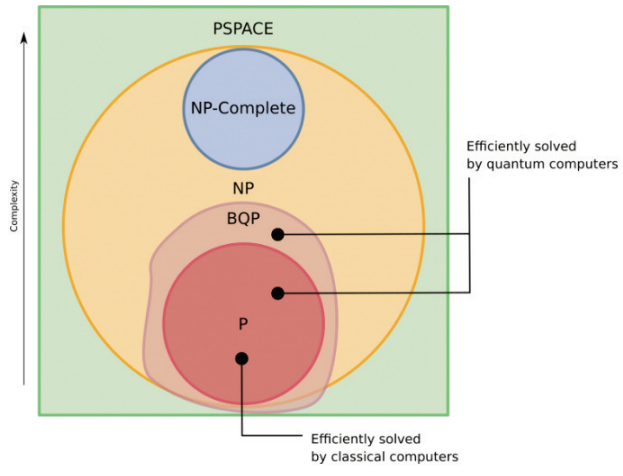


Figure 6: *classes of quantum computing complexity [32]. The relationship between BQP and NP is still unknown.*



Conclusions

The limitations of the modelling capabilities have been conceptualised in two main blocks: a block on theoretical limits, whose possible resolution could be achieved in the long term through the advance of science, and a block on practical limits, where advances for their solution are expected to occur in the short and medium term, thanks to the advance of technology. In the short term, the options for increasing the capacity of the designed algorithms are based on a correct approach to the problem to be modelled, obtaining greater amounts of data, improving the representativeness of the data and using techniques to avoid overfitting and underfitting. The development of solutions through the generation of new computing resources is also anticipated as an additional path that will be key in the resolution of these elements.

Bibliography

Arcadu, F., Benmansour, F., Maunz, A., Willis, J., Haskova, Z., & Prunotto, M. (2019). Deep learning algorithm predicts diabetic retinopathy progression in individual patients. Nature.

Arute, F., Arya, K., Babbush, R., & al, e. (2019). Quantum supremacy using a programmable superconducting processor. Nature.

Babyak, M. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic Medicine.

Bellaiche, F. (2020). An update on quantum computing and complexity classes. Obtenido de Quantum Bits: https://www.quantum-bits.org/?p=2309

BGSMath. (2020). An introduction to parametrized complexity. Obtenido de BGSMath: https://bgsmath.cat/event/an-introduction-to-parameterized-complexity/

Caulfield, J., & Dolev, S. (2010). Why future supercomputing requires optics. Nature.

Clay Mathematics Institute. (s.f.). Obtenido de Millenium problems: http://www.claymath.org/millennium-problems

Eberhardt, F. (2009). Introduction to the Epistemology of Causation. Philosophy Compass.

Frigg, Roman, Hartmann, & Stephan. (2020). Models in Science. Obtenido de The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (ed.): https://plato.stanford.edu/archives/spr2020/entries/models-science/

Google. (2019). Ridding Gmail of 100 million more spam messages with TensorFlow.

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. Princeton University Bulletin.

Hanza, A. (2020). 'Statistic' has made my day, or has it? Obtenido de Towards Data Science: https://towardsdatascience.com/statistic-has-made-my-day-or-has-it-982c7d527410

Hao, K. (2 de 4 de 2020). AI can't predict how a child's life will turn out even with a ton of data. MIT Techonology Review.

Hartnett, K. (23 de 10 de 2019). Google and IBM Clash Over Milestone Quantum Computing Experiment. Quanta Magazine.

iDanae. (2019). Interpretabilidad de los modelos de Machine Learning.

iDanae. (2020). Ética e Inteligencia Artificial.

IEEE Innovation. (2020). Obtenido de Real-Life Use Cases for Edge Computing: https://innovationatwork.ieee.org/real-life-edge-computing-use-cases/

James, G., Hastie, T., Witten, D., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.

Malhotra, P. (2020). Machine Learning: regressions, regularization and learning curves. Obtenido de Predictive Programmer: https://predictiveprogrammer.com/machine-learning-regressions-regularization-and-learning-curves/

Narayanan, A. (2019). How to recognize AI snake oil.

Nielsen, M., & Chuang, I. (2002). Quantum Computation and Quantum Information. American Journal of Physics.

Papadimitriou, C. (2003). Computational Complexity. Encyclopedia of Computer Science.

Passian, A., & Imam, N. (2019). Nanosystems, Edge Computing, and the Next Generation Computing Systems. Sensors.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press.

Peleg, D. (2011). Distributed Computing. Springer.

Podnieks, K. (2010). The limits of modeling. PhilSci Archive.

Rodgers, J., & Nicewander, A. (1988). Thirteen Ways to Look at the Correlation Coefficient. The American Statistician.

Schaller, R. (1997). Moore's law: past, present and future. IEEE Spectrum.

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge Computing: Vision and Challenges. IEEE Internet of Things Journal.

Stipić, A., Bronzin, T., Prole, B., & Pap, K. (2019). Deep learning advancements: closing the gap. 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE.

Authors

Ernestina Menasalvas (UPM)

Alejandro Rodríguez (UPM)

Manuel Ángel Guzmán (Management Solutions)

Daniel Ramos (Management Solutions)

Carlos Alonso (Management Solutions)

UNIVERSIDAD POLITÉCNICA DE MADRID

The Universidad Politécnica de Madrid is a public-law organization of a multisectoral and multidisciplinary nature that is engaged in teaching, research, as well as science and technology development activities.

www.upm.es

Management Solutions is an international consulting firm whose core mission is to deliver business, risk, financial, organizational and process-related advisory services, with operations in more than 40 countries and a multidisciplinary team of 2,500 professionals working for over 900 clients worldwide.

www.managementsolutions.com

For more information, visit

blogs.upm.es/catedra-idanae/