Machine learning and the Continuum Hypothesis

Hart, Klaas Pieter

Publication date
2019

Document Version
Final published version

Published in
Nieuw Archief voor Wiskunde

Citation (APA)
Hart, K. P. (2019). Machine learning and the Continuum Hypothesis. *Nieuw Archief voor Wiskunde*, *20*(3), 214-217.

K. P. Hart
Faculteit EWI
TU Delft
k.p.hart@tudelft.nl

Research

Machine learning and the Continuum Hypothesis

In January 2019 the journal *Nature* reported on an exciting development in Machine Learning: the very first issue of the journal *Nature Machine Intelligence* contains a paper that describes a learning problem whose solvability is neither provable nor refutable on the basis of the standard ZFC axioms of Set Theory. In this note K. P. Hart describes what the fuss is all about and indicates that maybe the problem is not so undecidable after all.

In the paper [1], in *Nature Machine Intelligence*, its authors exhibit an abstract machine-learning situation where the learnability is actually neither provable nor refutable on the basis of the axioms of ZFC. This was deemed so exciting that the mother journal *Nature* devoted *two* commentaries to this: [9] and [3].

The first of these, [9], is rather matter-of-fact in its description of the problem but the second, [3], manages, in just a few lines, to mix up Gödel's Incompleteness Theorems and the undecidability of the Continuum Hypothesis. It misstates the former — "Gödel discovered logical paradoxes" — and misinterprets the latter: "a paradox known as the Continuum Hypothesis".

No, Gödel did not discover paradoxes; he proved a (highly) technical result about formal proofs. That result shows that under certain circumstances a first-order theory will be incomplete, that is, there is a formula φ such that there is no formal proof of φ nor of its negation. The formula φ constructed by Gödel asserts, indirectly, "there is not formal proof for φ" and as such looks a bit like "this sentence is false", which can be construed as a version of the Liar's Paradox. There is a difference however: the formula φ does not refer directly to itself and this prevents it from being a paradox.

And, no, the Continuum Hypothesis is not a paradox either. It is 'simply' a statement about subsets of the real line that exhibits a concrete incompleteness of ZFC Set Theory. That theory is subject to Gödel's incompleteness theorem, hence it comes with its own version of the formula φ. Both φ and the Continuum Hypothesis show that ZFC is incomplete, the difference between these formulas is that the Continuum Hypothesis is interesting and φ is not. This is not meant in a pejorative way; as Gödel's construction applies to potentially very many different theories one would not expect φ to say something very specific in the theory that it is constructed for.

The set theory in [1] is related to Cantor's original formulation of the Continuum Hypothesis [2]: if one declares two sets to be equivalent if there is a bijection between them then the infinite subsets of $\mathbb{R}$ are divided into *two* equivalence classes, those of the sets equivalent to $\mathbb{N}$ and those of the sets equivalent to $\mathbb{R}$.

The learning problem from [1] is equivalent to a weaker version: the number of equivalence classes is finite. For the rest of this note we shall refer to that statement as the Weak Continuum Hypothesis. This statement is, like the Continuum Hypothesis neither provable not refutable on the basis of the axioms of ZFC.

For later use we abbreviate "there is a bijection between X and Y" as $X \equiv Y$. Thus, the Continuum Hypothesis states that if X is an infinite subset of $\mathbb{R}$ then either $X \equiv \mathbb{N}$ or $X \equiv \mathbb{R}$.

In the next section I will summarize the description from [1] of the so-called *EMX learning problem*. The section that follows contains the translation from [1] of the learning problem into a purely combinatorial problem about functions between powers of the unit interval and an explanation of why that translation is equivalent to the Weak Continuum Hypothesis. In the section thereafter we shall see that the combinatorial part is related to a result of Kuratowski from 1951 [6], that characterizes when, given $k \in \mathbb{N}$, a set has (at most) $k+1$ equivalence classes of infinite sets under the equivalence relation $\equiv$ discussed above. In the last section I will show why I think that the problem is not undecidable at all: there is no algorithm that solves this particular learning problem.

The learning problem

This is a summary of the parts of [1] that lead to the undecidability result.

The authors start with the following real-life situation as an instance of their general learning problem. A website has a collection of advertisements that it can show to its visitors; each advertisement, A, comes with a set, F_A, of visitors for whom it is of interest: say if A advertis-

es running shoes then F_A contains avid runners (or people who just like snazzy shoes). Choosing the optimal advertisement to display amounts to choosing a finite set from a population while maximizing the probability that the visitor is actually in that set. The problem is that the probability distribution is unknown.

Rather than dwell on this particular example the authors make an abstraction: Given a set X and a family $\mathcal{F}$ of subsets of X find a member of $\mathcal{F}$ whose measure with respect to an unknown probability distribution is close to maximal. This should be done based on a finite sample generated i.i.d. from the unknown distribution.

The undecidability manifests itself when we let X be the unit interval $\mathbb{I}$ and $\mathcal{F}$ the family $\operatorname{fin}\mathbb{I}$ of finite subsets of $\mathbb{I}$.

Learning functions

In the general situation the abstract problem described above is made more explicit and quantitative as follows.

For the unknown probability distribution P on X find $F\in\mathcal{F}$ such that $E_P(F)$ is quite close to $\operatorname{Opt}(P)$, which is defined to be $\sup_{Y\in\mathcal{F}}E_P(Y)$.

To quantify this further a *learning function* for $\mathcal{F}$ is defined to be a function

$$G\colon \bigcup_{k\in\mathbb{N}} X^k \to \mathcal{F}$$

with certain desirable properties.

Say, if $S\in X^k$ represents a sample of visitors then $G(S)$ would be a set of visitors from which the next visitor is very likely to come. As $G(S)$ belongs to $\mathcal{F}$, there is an advertisement A such that $G(S)=F_A$ and this A will be displayed on the website.

The desirable properties, e.g., the 'very likely' in the example above, are captured in the following definition of an (ϵ,δ)-*EMX learner* for $\mathcal{F}$. This is a function G as above such that for some $d\in\mathbb{N}$, depending on ϵ and δ, the following inequality holds

$$\Pr_{S\sim P^d}\left[E_P(G(S))\leqslant \operatorname{Opt}(P)-\varepsilon\right]\leqslant\delta$$

for all distributions P with finite support.

The letters EMX abbreviate 'estimating the maximum'.

Combinatorics

It seems a nigh on hopeless task to say anything sensible when there are so many possible probability distributions to consider. However, as we shall see, the existence of EMX learning functions is equivalent to the existence of maps between finite powers of X that mention no probabilities at all but instead are required to satisfy a few simple inclusion relations. These will prove to be much more amenable to set-theoretic investigations.

A combinatorial translation

The authors of [1] do not waste a lot of time and formulate, without much ado, the combinatorial statement equivalent to the existence of an EMX learner.

This statement involves what the authors call monotone compression schemes. Their formulation needs the following piece of notation: For a set X and a natural number n we use $[X]^n$ to denote the family of n-element subsets of X.

Definition 1. Let m and d be two natural numbers with $m>d$. An $m\to d$ *monotone compression scheme* for a family $\mathcal{F}$ of finite subsets of a set X is a function $\eta\colon[X]^d\to\mathcal{F}$ such that whenever A is an m-element subset of X it has a d-element subset B such that $A\subseteq\eta(B)$.

This is slightly different from the formulation of Definition 2 in [1], which leaves open the possibility that $|A|<m$ and that $|B|<d$, as it uses indexed sets. It is clear from the results and their proofs that our definition captures the essence of the notion.

The idea here is that someone, Alice say, thinks of an m-element set A and provides their friend Bob with a d-element subset B of A. The function η helps Bob to recover some information about A, namely that it is a subset of the member $\eta(B)$ of the family $\mathcal{F}$.

There is a second unnamed function implicit in Definition 2: the choice of the subset B of A; this function we shall call σ.

So a scheme consists of a pair of functions: $\sigma\colon[X]^m\to[X]^d$ and $\eta\colon[X]^d\to\mathcal{F}$; these should satisfy $A\subseteq(\eta\circ\sigma)(A)$ for all A. In fact, as we shall see in the next section, the function σ is more convenient to work with.

The translation is now as follows.

Lemma 1 [1, Lemma 1.1]. *For an upward-directed family $\mathcal{F}$ of finite sets the existence of a $(\frac{1}{3},\frac{1}{3})$-EMX learning function is equivalent to the existence of a natural number m and an $(m+1)\to m$ monotone compression scheme for $\mathcal{F}$.*

The proof of necessity takes the natural number d in the learning function and produces a monotone $(m+1)\to m$ compression scheme with $m=\left\lceil\frac{3}{2}d\right\rceil$.

Undecidability

At this point the authors turn to the aforementioned special case of the unit interval $\mathbb{I}$ and its family $\operatorname{fin}\mathbb{I}$ of finite subsets and prove the following.

Theorem 1. *There is a monotone $(m+1)\to m$ compression scheme for $\operatorname{fin}\mathbb{I}$ for some $m\in\mathbb{N}$ if and only if the Weak Continuum Hypothesis holds.*

As the Weak Continuum Hypothesis is both consistent with and independent of the axioms of ZFC the same holds for the existence of a compression scheme and for the existence of a $(\frac{1}{3},\frac{1}{3})$-EMX learning function. Theorem 1 is an immediate consequence of the set of equivalences in the following theorem.

Theorem 2 [1, Theorem 1]**.** *Let $k\in\mathbb{N}$ and let X be a set. There is a $(k+2)\to(k+1)$ monotone compression scheme for the finite subsets of X if and only the infinite subsets of X are divided into $k+1$ (or fewer) equivalence classes by the relation $\equiv$.*

Indeed, the Weak Continuum Hypothesis holds iff the infinite subsets of $|\mathbb{I}|$ are divided into $k+1$ equivalence classes for some $k\in\mathbb{N}$.

In the next section we take a closer look at monotone compression schemes and point out a connection with an old result of Kuratowski's.

Compression schemes and decompositions

In the general case considered above the function η is important because of its codomain $\mathcal{F}$: Bob is required to choose a member of that family. It turns out that in the case considered by the authors of [1], namely the family of *all* finite subsets of a set X, it is the function σ that is more interesting. This is borne out by the following proposition.

Proposition 1. *Let m and d be natural numbers and let X be a set. There is an $m\to d$ monotone compression scheme for the finite subsets of X if and only if there is a finite-to-one function $\sigma\colon[X]^m\to[X]^d$ such that $\sigma(x)\subseteq x$ for all x.*

Proof. If the pair $\langle \eta, \sigma \rangle$ determines an $m \to d$ monotone compression scheme then σ is finite-to-one. For let $y \in [X]^d$ then $\sigma(x) = y$ implies $x \subseteq \eta(y)$, hence there are at most $\binom{M}{m}$ such x, where $M = |\eta(y)|$.

Conversely, if σ is as in the statement of the proposition then we can let $\eta(y) = \bigcup \{x : \sigma(x) = y\}$. □

Kuratowski's decompositions

To do justice to Kuratowski's results and because the proofs require it we will use standard set theoretic notions and notations. We shall need the first countably many infinite cardinal numbers $\aleph_k$ ($k \in \mathbb{N}$) and the ordinal numbers ω_k. What we also need to know is that ω_k is the 'standard' well-ordered set of cardinality $\aleph_k$.

Above I formulated Kuratowski's 1951 result in terms of the equivalence relation "there is a bijection" but as the title of [6] indicates the original formulation involved the cardinal numbers $\aleph_k$. To be precise the papers characterizes when a set has cardinality at most $\aleph_k$ in terms of its $(k+2)$-nd power. The very definition of the cardinal numbers $\aleph_k$ makes it clear that a set has cardinality at most $\aleph_k$ if and only if there are at most $k+1$ equivalence classes under the equivalence relation $\equiv$. From now on we let $|X|$ denote the cardinality of the set X, so that $|X| \leqslant \aleph_k$ abbreviates that the cardinality of X is at most $\aleph_k$.

It should come therefore as no big surprise that Kuratowski's results and Theorem 2 are related.

We start by quoting the following theorem from [6], it provides one direction in the aforementioned characterization.

Theorem 3. *The power ω_k^{k+2} can be written as the union of $k+2$ sets, $\{A_i : i < k+2\}$, such that for every $i < k+2$ and every point $\langle x_j : j < k+2 \rangle$ in ω_k^{k+2} the set of points y in A_i that satisfy $y_j = x_j$ for $j \neq i$ is finite.*

In Kuratowski's words "A_i is finite in the direction of the ith axis".

Sketch of the proof. The case $k = 0$ is easy: ω_0 is the first infinite ordinal, therefore $A_0 = \{\langle m, n \rangle : m \leq n\}$ and $A_1 = \{\langle m, n \rangle : m > n\}$ are as required.

The rest of the proof proceeds by induction on k. We give the step from $k = 0$ to $k = 1$ in some detail and leave the other steps to the reader.

To decompose ω_1^3 into three sets A_0, A_1 and A_2 we apply the Axiom of Choice to choose (simultaneously) for each infinite ordinal α in ω_1 a decomposition $\{X(\alpha,0), X(\alpha,1)\}$ of $(\alpha+1)^2$, say by choosing well-orders of type ω and then using the decomposition obtained for $k = 0$.

- One puts $\langle \alpha, \beta, \gamma \rangle$ into A_0 if β is the largest coordinate and $\langle \alpha, \gamma \rangle \in X(\beta, 0)$ or if γ is the largest coordinate and $\langle \alpha, \beta \rangle \in X(\gamma, 0)$.
- One puts $\langle \alpha, \beta, \gamma \rangle$ into A_1 if α is the largest coordinate and $\langle \beta, \gamma \rangle \in X(\alpha, 0)$ or if γ is the largest coordinate and $\langle \alpha, \beta \rangle \in X(\gamma, 1)$.
- One puts $\langle \alpha, \beta, \gamma \rangle$ into A_2 if α is the largest coordinate and $\langle \beta, \gamma \rangle \in X(\alpha, 1)$ or if β is the largest coordinate and $\langle \alpha, \gamma \rangle \in X(\gamma, 0)$.

To see that A_0 is finite in the direction of the 0th coordinate take $\langle \beta, \gamma \rangle \in \omega_1^2$, then $\langle \alpha, \beta, \gamma \rangle \in A_0$ implies β is largest and $\langle \alpha, \gamma \rangle \in X(\beta, 0)$, or γ is largest and $\langle \alpha, \beta \rangle \in X(\gamma, 0)$; in either case α belongs to a finite set.

A similar argument works for A_1 and A_2 of course.

The inductive steps for larger k are modeled on this step. □

We now show how Theorem 3 can be used to prove sufficiency in Theorem 2.

Here and in later sections it will be convenient to identify $[X]^m$, the family of m-element subsets of X, with a subset of the product X^m. In the cases of interest the set X has a (natural) linear order $\prec$; we use this to let a set correspond with its monotone enumeration:

$$[X]^m = \{x \in X^m : (i < j < m) \to (x_i \prec x_j)\}$$

Constructing a compression scheme from a decomposition. From a decomposition as in Theorem 3 we construct a finite-to-one function $\sigma : [\omega_k]^{k+2} \to [\omega_k]^{k+1}$ such that $\sigma(x) \subseteq x$ for all x. We assume, without loss of generality, that the sets A_i are disjoint.

Let $x \in [\omega_k]^{k+2}$ (so $i < j < k+2$ implies $x_i < x_j$). Take (the unique) i such that $x \in A_i$ and let $\sigma(x)$ be the point in ω_k^{k+1} that is x but without its coordinate x_i. In terms of sets we would have set $\sigma(x) = x \setminus \{x_i\}$.

This function is finite-to-one: if $y \in [\omega_k]^{k+1}$ then for each $i < k+2$ there are only finitely many x in A_i with $y = \sigma(x)$. □

As mentioned above Kuratowski's result works both ways: if X^{k+2} admits a decomposition as above for ω_k^{k+2} then $|X| \leq \aleph_k$. This suggests that the necessity in Theorem 2 is related to the converse of Theorem 3. This is indeed the case: one can construct a Kuratowski-type decomposition from a compression scheme, but because of our definition of the schemes we only get a decomposition of the subset $[\omega_k]^{k+2}$ of the whole power. This can be turned into one for the whole power but the process is a bit messy so we leave it be.

The proof of necessity from [1] closes the circle of implications that proves the following.

Theorem 4. *For a set X and a natural number k the following are equivalent:*

1. $|X| \leq \aleph_k$;
2. *X^{k+2} admits a Kuratowski-type decomposition into $k+2$ sets*;
3. *there is a $(k+2) \to (k+1)$ monotone compression scheme for the finite subsets of X.*

For completeness sake I sketch the proof of that last implication. Both it and Kuratowski's necessity proof use a form of the following lemma. Its proof uses some elementary cardinal arithmetic for infinite cardinals numbers.

Lemma 2. *Let k, l, and m be natural numbers with $m > l$. Assume $\sigma : [\omega_{k+1}]^{m+1} \to [\omega_{k+1}]^{l+1}$ determines an $(m+1) \to (l+1)$ monotone compression scheme. Then there is an $m \to l$ monotone compression scheme for the finite subsets of ω_k.*

Proof. We start by determining an ordinal δ as follows. Let $\delta_0 = \omega_k$. Given δ_n use the fact that σ is finite-to-one to find an ordinal $\delta_{n+1} > \delta_n$ such that every $x \in [\omega_{k+1}]^{m+1}$ that satisfies $\sigma(x) \in [\delta_n]^{l+1}$ is in $[\delta_{n+1}]^{m+1}$.

In the end let $\delta = \sup_n \delta_n$. Then δ satisfies: every $x \in [\omega_{k+1}]^{m+1}$ that satisfies $\sigma(x) \in [\delta]^{l+1}$ is in $[\delta]^{m+1}$.

We define an $m \to l$ monotone compression scheme for δ. If $x \in [\delta]^m$ then $y = x \cup \{\delta\}$ is in $[\omega_{k+1}]^{m+1}$ and so $\sigma(y) \subseteq y$. By the choice of δ it is not possible that $\sigma(y) \subseteq x$ hence $\delta \in \sigma(y)$ and so setting $\varsigma(x) = \sigma(y) \setminus \{\delta\}$ defines a map $\varsigma : [\delta]^m \to [\delta]^l$. This map is finite-to-one and satisfies $\varsigma(x) \subseteq x$ for all x. □

To finish the proof of necessity we argue by induction and contradiction. If $|X| = \aleph_{k+1}$ and there is a finite-to-one $\sigma : [X]^{k+2} \to [X]^{k+1}$ with $\sigma(x) \subseteq x$ for all x then there is a subset Y of X with $|Y| = \aleph_k$ and a finite-to-one $\varsigma : [Y]^{k+1} \to [Y]^k$ with $\varsigma(x) \subseteq x$ for all x. This would contradict the obvious inductive assumption. We leave it as an exercise to the reader to ponder what absurdity would arise in the case $k = 0$ and provide the basis for the induction.

Algorithmic considerations

In this section we address a point already raised by the authors in [1]: the functions that are used in the previous sections are quite arbitrary and not related to any recognizable algorithm. Indeed, the constructions of the compression schemes for uncountable sets blatantly applied the Axiom of Choice: once by assuming that the underlying sets were well-ordered and again when in every step of the induction a choice of well-orders of type ω_k needed to be made.

One may therefore wonder what happens if we impose some structure on the maps in question. One possible way of separating out 'algorithmic' functions is by requiring them to have nice descriptive properties. If 'nice' is taken to mean 'Borel measurable' then the desired functions do not exist.

Continuity and Borel measurability

Here we show, for arbitrary $m \in \mathbb{N}$, that there does not exist an $(m+1) \to m$ monotone compression scheme for the finite subsets of $\mathbb{I}$ where the function σ is Borel measurable. Remember that we identify $[\mathbb{I}]^k$ with the open subset of the k-cube $\mathbb{I}^k$ consisting of its strictly increasing elements. As such it inherits a metric and a Borel structure from that cube. We consider continuity and Borel measurability with respect to these structures. Let m be a natural number and let $\sigma : [\mathbb{I}]^{m+1} \to [\mathbb{I}]^m$ be a function such that $\sigma(x) \subseteq x$ for all x.

If σ is continuous then σ is not finite-to-one To see this let $x \in [\mathbb{I}]^{m+1}$ and assume for notational convenience that $\sigma(x) = \langle x_i : i < m \rangle$, i.e., that the coordinate x_m is left out of x when forming $\sigma(x)$.

Let $\varepsilon = \frac{1}{3}\min\{x_{i+1} - x_i : i < m\}$ and let $\delta > 0$ be such that $\delta \leq \varepsilon$ and for all $y \in [\mathbb{I}]^{m+1}$ with $\|y - x\| < \delta$ we have $\|\sigma(y) - \sigma(x)\| < \varepsilon$.

Now if $y \in [\mathbb{I}]^{m+1}$ and $\|y - x\| < \delta$ then $|y_i - x_i| < \varepsilon$ for all $i \leq m$. Also, when $i < j$ we have $x_j - x_i > 3\varepsilon$. It follows that $y_m - x_i > \varepsilon$ for all $i < m$. This implies that $\sigma(y) = \langle y_i : i < m \rangle$ for all y with $|y - x| < \delta$.

This shows that for every i the set $O_i = \{x \in [\mathbb{I}]^{m+1} : \sigma(x) = x \setminus \{x_i\}\}$ is open. Because $[\mathbb{I}]^{m+1}$ is connected there is one i such that $O_i = [\mathbb{I}]^{m+1}$. This shows that σ cannot be finite-to-one. □

The above proof can be used/adapted to show that if σ is Borel measurable it is not finite-to-one either.

If σ is Borel measurable then σ is not finite-to-one. There is a dense G_δ-set G in $[\mathbb{I}]^{m+1}$ such that the restriction of σ to G is continuous, see [7, Section 31 II].

Let $x \in G$. As in the previous proof we assume $\sigma(x) = \langle x_i : i < m \rangle$ and we obtain a $\delta > 0$ such that $\sigma(y) = \langle y_i : i < m \rangle$ for all $y \in G$ that satisfy $\|y - x\| < \delta$.

By the Kuratowski–Ulam theorem [8] we can find a point y in G with $\|y - x\| < \delta$ such that the set of points t in the interval $(x_m - \delta, x_m + \delta)$ for which $y_t = \sigma(y) * \langle t \rangle$ belongs to G is co-meager. But for every such point we have $\sigma(y_t) = \sigma(y)$ and this shows that σ is not finite-to-one. □

EMX learning is impossible

As we saw above a learning function is a function G from the union $\bigcup_{k \in \mathbb{N}} \mathbb{I}^k$ to the family of finite subsets of $\mathbb{I}$. We can call such a function continuous or Borel measurable if its restriction to each individual power is.

In the construction of an $(m+1) \to m$ compression scheme from a learning function the authors use its restriction to just one of these powers $\mathbb{I}^d$, where $d \leqslant m$. The definition of $\eta(S)$ involves taking the union of $G(T)$ for all d-element subsets T of S, hence a union of $\binom{m}{d}$ many sets.

The definition of σ involves choosing one m-element subset with a certain property from of a given $m+1$-element set.

The latter choice can be made explicit using a Borel linear order on the family of all finite subsets of $\mathbb{I}$, or just on $[\mathbb{I}]^m$.

An analysis of this procedure shows that if G is Borel measurable then so are σ and η. The results of this section then imply that a Borel measurable learning function does not exist. In this author's opinion that means that the title of [1] should be emended to "EMX learning is impossible".

On the other hand...

One may argue that the choice of the unit interval in [1] is a bit of a red herring. None of the arguments in the paper use the structure of $\mathbb{I}$ in any significant way.

In the step from the problem of the advertisements to the more abstract problem there is no real need to go to the unit interval. One may equally well use the set of rational numbers to code or rank the elements of the learning set.

In that case there is, as we have seen, a $2 \to 1$ monotone compression scheme for the finite subsets of $\mathbb{N}$: simply let $\sigma(x) = \max x$; the corresponding function η is defined by $\eta(n) = \{i : i \leqslant n\}$.

It is an easy matter to transfer this scheme to the family of finite subsets of the rational numbers. Whether this scheme gives rise to a useful EMX learning function remains to be seen.

References

1 Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka and Amir Yehudayoff, Learnability can be undecidable, *Nature Machine Intelligence* 1 (2019), 44–48.

2 Georg Cantor, Ein Beitrag zur Mannigfaltigkeitslehre, *Crelles Journal für Mathematik* 84 (1878) 242–258.

3 Davide Castelvecchi, Machine learning leads mathematicians to unsolvable problem, *Nature* 565 (2019), 277.

4 Ryszard Engelking, *General Topology,* Sigma Series in Pure Mathematics 6, Heldermann Verlag, 1989, 2nd ed..

5 Kenneth Kunen, *Set Theory. An Introduction to Independence Proofs,* Studies in Logic and the Foundations of Mathematics 102, North-Holland, 1980.

6 Casimir Kuratowski, Sur une caractérisation des alephs, *Fundamenta Mathematicae* 38 (1951), 14–17.

7 K. Kuratowski, *Topology. Vol. I*, Academic Press and Państwowe Wydawnictwo Naukowe, 1966.

8 C. Kuratowski and St. Ulam, Quelques propriétés topologiques du produit combinatoire, *Fundamenta Mathematicae* 19 (1932), 247–251.

9 Lev Reyzin, Unprovability comes to machine learning, *Nature* 565 (2019), 166–167.