The development of mathematical ideas by collision: The Case of Categories and Topos theory

SAUNDERS MAC LANE
University of Chicago
Chicago, Il. 60637 USA

Recently, I prepared an article discussing the develoment of category theory (Mac Lane 1988b). Peter Johnstone, in commenting on its meaning, observed that category theory began in a collision between ideas: Algebra clashing with Topology. So it was: Eilenberg, coming from topology, met Mac Lane, coming from algebra; together they chanced upon a problem involving both these subjects. From this collision, of ideas and of individuals, both homological algebra and category theory developed.

1. Homological Algebra. In 1940, Norman Steenrod had described the "Regular cycles" of a compact metric space in order to define a new and more expressive homology group (MR 2, p. 73; note that most references in the present article will be to the volume and page of the corresponding review in Mathematical Reviews). However, Steenrod had not been able to compute this group for the case of the p-adic solenoid, for a prime number p. (This solenoid is the intersection of an infinite sequence of solid tori T_n where T_{n+1} winds p times around inside T_n.) On the other hand, the class field theory for a normal extension N of a base field K had used group extensions of the multiplicative group of N by the Galois group G acting on N. In this connection, Mac Lane had studied group extensions more generally, and in particular the group $\mathrm{Ext}(G, A)$ of all abelian group extensions of the group A by the group G. He had calculated a particular case which seemed of interest: That in which G is the abelian group generated by the list of elements a_n, where $a_{n+1} = pa_n$ for a prime p. After a lecture by Mac Lane on this calculation, Eilenberg pointed out that the calculation closely resembled that for the regular cycles of the p-adic solenoid; indeed, the group G above is the dual of the solenoid, regarded as a topological group. Joint examination of this surprising connection— a collision between group theory and homology theory— soon led Eilenberg-Mac Lane to formulate a "universal coefficient theorem": The cohomology $H^n(X, A)$ of a space X with coefficients in the abelian group A can be described in terms of the homology of that space by a short exact sequence (MR 4, p. 88):

$$0 \longrightarrow \mathrm{Ext}(H_{n+1}(X), A) \longrightarrow H^n(X, A) \longrightarrow \mathrm{Hom}(H_n(X), A) \longrightarrow 0.$$

Here the scond map p is that given by evaluating each cocycle of X on each integral cycle; i.e., on each element of the integral homology group $H(X)$. The development of this collision between algebra and topology has been described in more detail

(Mac Lane, 1988a in the list of references). In the hands of Cartan-Eilenberg, it soon turned out that it involved the construction of Ext as the first derived functor of Hom; in other words, this basic idea of homological algebra arose from the unexpected collision of class field theory with regular cycles in homology.

2. Categories. In order to treat the universal coefficient sequence under the inverse limits used in the definition of Čech cohomology one needed to study the behavior of this short exact sequence under a continuous mapping $f \quad X \to Y$ of spaces. Since Hom$(-, A)$ and $H(-, A)$ are both contravariant, this leads to a diagram of the form

$$\begin{array}{ccccc} H^n(Y,A) & \xrightarrow{p} & \mathrm{Hom}(H_n(Y),A) & \longrightarrow & 0 \\ \downarrow f^* & & \downarrow f^* & & \\ H^n(X,A) & \xrightarrow{p} & \mathrm{Hom}(H_n(X),A) & \longrightarrow & 0, \end{array}$$

The diagram commutes, and this was then regarded as a sort of expression of the fact that the mapping p described above is a "natural" one. But in order to make exact sense of this statement of naturality, one had to know how $H(X)$ depends on X; in present language, to know that it is a functor of X which goes from a suitable category of spaces to the category of abelian groups. Thus this clash of algebra with topology demanded that one define first the notion "category" and then "functor" so that the notion of natural transformation between functors could be properly described. From this collision arose category theory (Eilenberg-Mac Lane, 1945, MR 7. p. 109).

3. Group Cohomology. In a group extension of an abelian group A by a non-abelian group G the multiplication of representatives of the cosets x and y of G can be described by a factor set $f(x, y) \in A$ which satisfies a suitable identity expressing the associativity of multiplication. The set of all equivalence classes of such factor sets constitute an abelian group $H^2(G, A)$. These groups were much used in class field theory in the case when G is the Galois group of a normal field extension acting on the group of non-zero elements of the extension; in particular, this group arose in the description of the crossed product algebras split by the field extension A. Other questions about such field extensions had led the algebraist Oswald Teichmuller (MR 2, p. 122) to introduce three-dimensional factor sets satisfying a suitable identity, with equivalence classes forming an abelian group $H^3(G, A)$. Despite his remarkable talents, Teichmuller did not get beyond the dimension 3 for these cohomology groups, probably seeing no purpose in such an extension. But Eilenberg-Mac Lane, motivated by the use of $H_2(G, Z)$ and thus $H^2(G, A)$ in the Hopf description of the effect of the fundamental group G on the second homology (and second cohomology) of a space, thought that there might well be a similar influence upon cohomology groups in all the higher dimensions.

Hence, using singular simplicies to justify the boundary formulas, they defined the general cohomology groups. In other words, when there was no collision, nothing went beyond H^3; with a collision between crossed products in algebra and higher cohomology, the ideas for the cohomology of groups were forced. (Mac Lane: Origins of the cohomology of groups, MR 81j#01030)

4. Differential Geometry. A typical and speculative question is: Why was the idea X discovered at that particular time t? Thus the middle of the 20th century, with its emphasis on axiomatic methods and abstraction, was a favorable time for the discovery of categories; as someone has said, they would have been natural for Emmy Noether. But, absent Eilenberg-Mac Lane, would category theory have been developed? An alternative possible source lies in differential geometry; there the transport of geometrical structures from one tangent bundle to another along paths inevitably involves the groupoid of the homotopy classes of these paths, while with this groupoid other more general categories are at hand. In other words, a collision between algebra and ideas of differential geometry leads to category theory, and this collision was developed, 1952–1979, in the work of Ch. Ehresmann, for example in his "Local structures" (MR 20, #2392 and 84f #01118). The large body of Ehresmann's work is full of fertile ideas (such as double categories, structured categories or sketches), but their adequate development was unduly delayed because Ehresmann and his school for many years seem to have avoided contact and hence collision with the ideas and terminology of other workers in category theory.

5. Adjoint Functors. For a considerable time after their discovery in 1945, categories served chiefly just as a convenient language for axiomatic homology theory and for homological algebra. The next decisive notion, that of an adjoint functor, did not arise until 1958—although the closely associated notion of a universal construction had been recognized by several authors ten years before that. Hence the speculative question: Why was the discovery of adjoint functors so delayed? One possible answer is that a number of the most immediate adjoint functors—including those present in many universal constructions— were so trivial that they would hardly be named: for example the usual forgetful functors are often passed over in silence, without any notation. But there is another possible explanation. Adjoint functors did first appear in a paper of Daniel Kan (MR 24 #A1301). This paper came immediately after Kan's intensive study of homotopy via simplicial sets. (Then called semi-simplicial complexes)—and for this study Kan had to examine certain relevant adjoint situations: that between loop space and suspension and, even more important, the fact that the Milnor geometric realization of a simplicial set (MR 18 p.815) is left adjoint to the functor sending a space into its singular complex. In other words, this particular collision between geometry (spaces) and algebra (simplicial sets) directly involved the idea of adjunction, and so made the discovery of this idea almost inevitable.

6. Universal Algebra. When this subject began, in the work of Garrett

Birkhoff, an algebra was described as a set S equipped with a list of operations (unary, binary, ternary, and so on), together with identities between specified compositions of the given operations. Much later it was recognized that such an algebra really involves all of these composite operations from the beginning. This involvement was first explicit in (unpublished) lectures by Philip Hall on clones of operators; a clone being the set of all n-ary operations for each n. (This in lectures at Cambridge University about 1950). Independently, in 1963, F. W. Lawvere gave an even more elegant formulation of this invariant approach to universal algebra. He defined an algebraic theory as a category with objects the natural numbers n, each n represented as the product of n 1's, with the arrows $n \to 1$ as the n-ary operations of the theory (MR 34#1373).

This invariant description of an algebra was clearly a great step forward—fully analogous to the description of a group as a set with a suitable multiplication and not just in terms of some set of generators and relations. Thus the related discoveries of clones and algebraic theories provided the possibility of a productive collision between the original notion of a universal algebra and ideas from category theory. It is to be greatly regretted that the specialists in universal algebra have stubbornly ignored this possibility. Philip Hall's notion of a clone appeared in book form in 1965, in the monograph by Paul Cohn (MR 31#224)— but without the connection to categories. George Grätzer's encyclopedic book on Universal Algebra (MR 40, #1320 and second edition MR 80g#08001) does remark that the Lawvere Algebraic Theories do give "the most elegant treatment"— but the book unfortunately wholly avoids this treatment. The result is that universal algebraists still use a completely obsolete notion of the "type" of an algebra; namely, the list of generating operations and their arities (see e.g., p. 23 in S. Burris and H. P. Sankappanavar, MR 83k#08001). This ridiculous limitation means that a group described by the operations multiplication, inverse, and unit is of a different type from a group described by the operation $(x, y) \to xy^{-1}$. The sad result of this over-specialization and avoidance of collision is that universal algebra today is one of the most backward and isolated branches of mathematics.

7. Categories in Prague. The development of a new "school" of mathematics in a given locality may often come about because of the collision of ideas: Some local inputs clashing with some ideas from outside. This appears to be the case with the very active development of categories in Prague. Thanks to information given me by Jiří Adámek, I may try to describe this development tentatively as follows (See also Mac Lane 1988b, p. 361). First of all the pioneering work of the influential topologist Eduard Čech involved a background of precategorical concepts—as for example, the use of inverse limits in the definition of the Čech cohomology groups. Then in about 1960 A. G. Kurosh from Moscow University lectured on categories in Prague. This in turn led to further categorical discussions in the major seminars on general topology conducted in Prague by Miroslav Katětov; this involved a study of the systematic paper on categories by Kurosh, Lifshits, and Sulgeifer (MR 22 #9526) as well as the 1951 paper by Katětov "On a category of spaces" (MR 32 #4644). This activity stimulated the initial research on category theory by V. Trnková in 1962 (MR 26 #3627) and by Miroslav Hušek in 1964 (MR 30 #4234). There were also effective contacts with the topological studies of J. de Groot in Amsterdam; visits there led to the work of Z. Hedrlín and of A. Pultr on the representation of small categories within other specific categories (MR 30 # 3123)— a subject of continued interest today in Prague. Clearly this development represented an international exchange and collision of ideas.

8. Sheaves. The development of sheaf theory provides an especially striking example of a multiple collision of ideas— from function theory, topology, algebraic geometry, and differential equations. The idea of a sheaf is implicitly present in complex analysis, as the sheaf of germs of analytic functions on C and in the accompanying notion of analytic continuation. Indeed, this notion of continuation by patching together different power series expansions foreshadows the notion of constructing more general functions by patching together local pieces (e.g., H. Weyl, "Die Idee der Riemannsche Flasche", MR 16, p. 1097). For functions of several complex variables the intensive study of the Cousin problems (e.g., K. Stein, MR 13, p. 224 and H. Cartan MR 7, p. 290) concerned functions defined differently over several parts of a space and so developed a background for sheaf theory. In algebraic topology, obstruction theory required the use of coefficient groups such as the homotopy groups $\pi_n(X, x)$ based at all the different points x of the space X; this led Norman Steenrod to his consideration of homology with local coefficients (MR 5, p. 104). At about the same time, J. Leray's profound study of the cohomology of fiber spaces involved the first formal definition of a sheaf (a faisceau; MR 12, p. 272). At that time, he defined sheaves on a space in terms of pieces given on closed subsets of the space; the present definition in terms of pieces given on open subsets was worked out by H. Cartan in his seminars (1950–51, exposé 14ff, MR 14, p. 670, 1951–52, exposé 15ff, MR 16, p.235). The description of a sheaf on X as an "espace étale" over X was developed in this connection by M. Lazard. The basic notion of cohomology with coefficients in a sheaf of modules then included Steenrod's cohomology with local coefficients. In 1955, J. P. Serre in his famous paper "Faisceau algébrique cohérent" (MR 16, p. 953) showed decisively that sheaf theory was needed to get suitable cohomology groups for algebraic geometry. Shortly thereafter, A Grothendieck, in his study of the extent of homological algebra, observed that the sheaves of modules on a space form an abelian category; this showed clearly the important role of this type of category. (Tôhoku, MR 21, #1328). For a full discussion, we refer to the historical article by John W. Gray on sheaf theory (MR 82j, #01060)—but the brief indications above show that the rapid development of sheaf theory resulted from a multiple collision of ideas—Riemann surfaces, several complex variables, local coefficients for cohomology, algebraic geometry and abelian categories. The collision continues to reverberate today—as in the study of D-modules and perverse sheaves in geometry (see e.g., R. Macpherson and K. Vilonen, MR 87m,#32028).

9. Grothendieck Topologies. In SGA 1, the first of his notable series of the "Séminaire de Géometrie Algébrique", (MR 50, #7129), Alexander Grothendieck contributed another decisive idea which arose from a collision between Galois theory and the study of covering spaces in topology. If $p : Y \to X$ is a covering space, then according to the usual definition each point x of X has a neighborhood $U \to X$ such that the pullback $U \times_X Y$ is a coproduct of copies of U (i.e., seen locally, X consists of a stack of homeomorphic copies of neighborhoods U.) Moreover, the covering group of p consists of those homeomorphisms $a \quad Y \to Y$ which preserve p ($pa = p$). On the other hand, in the Galois theory, a normal extension of fields is a monomorphism $m \quad K \to N$ of fields, and the Galois group consists of all the automorphisms $a \quad N \to N$ with $am = m$. Also, if $K \to L$ is another finite field extension of K, the tensor product (the pushout) $N \otimes_K L$ is a commutative algebra over N, and as such a direct sum of field extensions of N. The normality of N means that when one summand is N, then all the summands are N, so that $N \otimes_K L \cong \sum N$. The analogy with the stacked pieces of a covering space is a strong one (involving a duality, pushout to pullback, but the analogy misses at one point, in that a neighborhood $U \to X$ in covering spaces is a monomorphism, while the corresponding $K \to N$ is not the dual; i.e., not an epimorphism. Now at that time (1971) topology and sheaf theory was couched in terms of coverings by monos from neighborhoods. This analogy and its partial failure led Grothendieck to understand that one could also define sheaves for "coverings" of X by families of maps $C \to X$ which were not necessarily mono. This led to his definition of a "topology" in terms of suitable axioms for such "coverings". This was a step which was decisive for the introduction of the étale topology and other topologies in algebraic geometry. This idea of a topology by way of such coverings first appeared in M. Artin's Harvard lecture notes on Grothendieck topologies. It is a penetrating example of an idea which arose, in part, from a collision between topology (open coverings) and algebra (Galois theory). My colleagues in algebraic geometry still view this discovery as wonderful.

10. Topos. A turbulent period of clashing ideas is often followed by more specialized study, because ideas formed by a collision normally lead to further internal developments. This was the case for the notion of a Grothendieck topology, defined by giving a category (a "site") and a notion of covering for that site. Then the category of sheaves for that site constituted the basic object of study for Grothendieck's drastic revision of algebraic geometry as it was developed systematically in the famous SGA IV of 1963, later published (MR 50, #7130 to 7132). The category of sheaves for a site was called a topos (now a Grothendieck topos, to distinguish it from the more general elementary topos). Giraud found a characterization of such categories of sheaves. In one version of SGA IV (page 3 of exposé 4) appears the slogan "The authors of the present seminar consider that the object of topology is the study of toposes and not just of topological spaces" The resulting rapid development of topos theory for a period changed the general view of the way in which algebraic geometry should be carried out and assisted in the solution of some famous problems, such as the Weil conjectures (Deligne MR 49 #5013). This done, there has been a general return to more traditional views of specific problems in algebraic geometry. Of such is the ebb and flow of the development of mathematics.

11. Elementary Topos. In 1969, F. W. Lawvere held the new Killam Professorship at Dalhousie University in Halifax and in this connection was able to bring a number of visitors, in particular Myles Tierney. During this year, Tierney planned to lecture on axiomatic sheaf theory, following the general lead of the Grothendieck SGA IV. Lawvere, on the other hand, beginning in 1963 and 1964, (MR 30 #2029, MR 53 #130) had worked to find an appropriate axiomatization of the category of sets. In this work, he proposed to use the composition of functions and not set membership as the primitive notion. These two proposals—axiomatic sheaves and axiomatic functions—collided with sparks. For sets, it presently appeared that the usual notion of a characteristic function could be written as a pullback from the two element set Ω of truth values, and that there was a similar such Ω—no longer two-valued—in categories of sheaves. During the year 1969–70 this collision led to the beautiful and expressive definition of an elementary topos, described by axioms including the existence of a subobject classifier. The axioms were all first order statements in the language of categories—unlike the Giraud characterization of Grothendieck topoi, which depended on the requirement of infinite coproducts. There was a further collision with the notion of forcing from set theory, since Lawvere and Tierney were able to show that Paul Cohen's proof of the independence of the continuum hypothesis in a suitable topos (a functor category, of functors to sets from the Cohen partially ordered set) by regarding a Grothendieck topology as given by a suitable idempotent operator $j \quad \Omega \to \Omega$ on the subobject classifier of this topos. The crucial observation was that sheafification for this topology was the essential step in Cohen's construction of a new model violating the continuum hypothesis (Tierney, MR 51, #10088).

12 Development of Topos Theory. Once the notion of an elementary topos had been revealed by collision and described by axioms, there ensued a rapid development of this idea, with many simplifications in the axioms and in the deduction of properties of an elementary topos. Lawvere presented the original axioms at the International Congress of Mathematicians at Nice in 1970 (MR 55 #3029). Soon afterwards, the notes of Tierney's lectures at Varenna were circulated (MR 50 #7277) independently of these lectures, Benabou's seminar in Paris in 1970–71 developed the properties of an elementary topos. Peter Freyd, lecturing in Australia, provided decisive representation theorems for topoi in his 1972 paper "Aspects of Topoi" (MR

53 #576). A systematic presentation of topoi had appeared in the 1971 Aarhus lecture notes of Kock and Wraith (MR 49 #7324). There was a subsequent treatment in a 1975 paper by Wraith (MR 52 #13989). Mikkelson and then Paré showed how colimits could be constructed from limits and the other axioms. Finally, the many developments were all pulled together in a definitive 1977 book by Peter Johnstone, "Topos Theory" (MR 57#9791). This listing covers only selected high points of the rapid internal development of the subject. It illustrates well the notion of an internal development following a collision of ideas.

Such an internal development is generally in contrast to the immediate turmoil of a collision. However, in this case, there were also further collisions. For example, it turned out that the logic which is "internal" to a topos, with propositional operators acting on the subobject classifier, is in general an intuitionist logic, and this collision with intuitionism brought in other people and led to further advances. Again in 1972 the Grothendieck school introduced (M. Hakim, MR 51 # 500) a classifying topos for commutative rings—that is, a topos with a "universal" internal ring object such that every topos with such a ring could be obtained from this one by a pullback taking the universal ring to the other one. For algebraic geometry the commutative rings play a prominent role, but this result for them led to another collision calling for the construction of classifying topoi for other structures. Thus, in 1972 A. Joyal and G. E. Reyes described the classifying topos for any finitary algebraic theory (MR 50 #13182). Then in 1977, M. Makkai and G. E. Reyes did the same for possibly infinitary algebraic theories (MR 58 #21600). There were further contributions of related ideas by Diaconescu in 1975 (MR 52 #532) and by Benabou, also in 1975 (MR 52 #13990). Then definitive treatments of classifying topoi were given by Joyal (unpublished) and by Tierney (MR 53 #13357). Other results spanned both logic and algebraic geometry, as with Lawvere's 1975 observation (MR 52 #13384) that the Deligne theorem about enough "points" in a topos was essentially equivalent to Gödel's completeness theorem. Other such connections with logic have been emphasized by A. Joyal (often unpublished).

13. Computer Science. Not all collisions of ideas turn out to be productive. For some time now there have been proposed various uses of categories in computer science. About 1972 there was considerable enthusiasm for the study of finite state machines, not as usual just in sets but in more general categories. Thus J. A. Goguen in 1972 (MR 47 #302 and MR 48 #372) emphasized the idea that the minimal realization functor was adjoint to the behavior (of a machine). In the same year, H. Ehrig and M. Pfender wrote a book on categories and automata (MR 49 # 5119), while soon afterwards in 1974, M. Arbib and E. G. Manes wrote on machines in a category (MR 50 #16156 and 52 # 4714). This collision of category theory and automata has not subsequently prospered, though this may in part be due to the rapid change of fashions in computer science. A more lasting categorical notion for computers is that of initial algebra semantics, as developed for example by J. A.

Goguen and J. W. Thatcher in 1974 (MR 54 #4168). This direction makes good use of the notion of a universal (initial) construction.

Currently the active investigation of the "Polymorphic" data types appears to offer a much more productive collision. The essential observation appears to be that one often performs the same operation on several different types — and that ways of doing this can be effectively formulated with categorical techniques. There are, for example, recent investigations by Peter Freyd, John Gray, and Andre Scedrov. I think that this developing collision will be productive.

Conclusions. An examination of the history of mathematics can serve many purposes. One such is to illuminate the dynamical processes by which mathematics develops. The present brief analysis has shown that the development of category theory in the last 43 years has derived much stimulus from the interaction with other fields. At this time, when so much of mathematics tends to be highly specialized, this can be helpful. Thus the slogan: Watch out for profitable collisions.

OTHER REFERENCES

Artin, M. (1962) Grothendieck Topologies. Lecture Notes, Harvard University.

Mac Lane, S. (1988a) Group extensions for 45 years. The Mathematical Intelligencer, vol. 10. No. 2, pp. 29–35.

Mac Lane, S. (1988b) *Concepts and categories in perspective*, pp. 323–366 in "A Century of Mathematics in America, Part 1"; edited by Peter Duren, with the assistance of Richard A.Askey and Uta C. Merzbach, Providence R. I.: The American Mathematical Society.

Note added in proof on April 14, 1989

This essay is intended as a quite preliminary study of a historical question: What are the forces and dynamical processes which influence the development of mathematics? The thesis of this essay is that a "collision" of ideas from different fields is a significant such force. This thesis needs further study and a wider choice of examples. Now any examination of the history of late 20th century mathematics is complicated by the presence of an overwhelming number of publications of quite varied quality and scope. For this reason I have adopted the method of reference by the review number of each publication from Mathematical Reviews; this method is not intended as a recommendation of the quality of the review, but as a way of shortening the citations; I have previously used this method in 1988b.

As a referee has observed, some of the collisions here noted should be described with greater care. In particular, the discussion of category theory in Prague is too brief; research there went far beyond questions of the representation of categories, while these questions themselves concerned the representations of both small and large categories, and indeed arose in part from a collision with combinatorics and with questions of universal algebra (Isbell; MR29 #1238), as reflected in various papers, such as that by Hedrlin and Pultr (MR 33, #85).