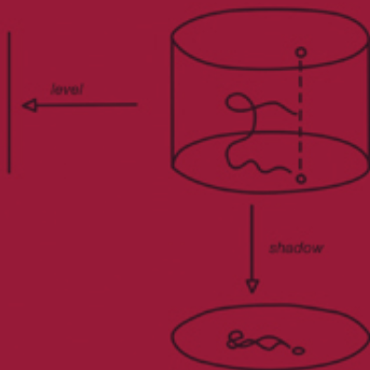
Conceptual Mathematics

A first introduction to categories

Second Edition

F. William Lawvere
Stephen H. Schanuel



CAMBRIDGE

Conceptual Mathematics, 2nd Edition

A first introduction to categories

F. WILLIAM LAWVERE
SUNY at Buffalo

STEPHEN H. SCHANUEL
SUNY at Buffalo

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo,
Delhi, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

www.cambridge.org
Information on this title: www.cambridge.org/9780521894852

First published 1997
Second edition 2009
3rd printing 2012

Printed and bound by MPG Books Group, UK

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data

Lawvere, F.W.
Conceptual mathematics : a first introduction to categories / F. William
Lawvere, Stephen H. Schanuel. – 2nd ed.
p. cm.
Includes index.
ISBN 978-0-521-71916-2 (pbk.) – ISBN 978-0-521-89485-2 (hardback)
1. Categories (Mathematics) I. Schanuel, S. H. (Stephen Hoel), 1933– II. Title.
QA169.L355 2008
512′.62–dc22 2007043671

ISBN 978-0-521-89485-2 Hardback
ISBN 978-0-521-71916-2 Paperback

Contents

Organisation of the book

The reader needs to be aware that this book has two very different kinds of 'chapters':

The **Articles** form the backbone of the book; they roughly correspond to the written material given to our students the first time we taught the course.

The **Sessions**, reflecting the informal classroom discussions, provide additional examples and exercises. Students who had difficulties with some of the exercises in the Articles could often solve them after the ensuing Sessions. We have tried in the Sessions to preserve the atmosphere (and even the names of the students) of that first class. The more experienced reader could gain an overview by reading only the Articles, but would miss out on many illuminating examples and perspectives.

Session 1 is introductory. Exceptionally, Session 10 is intended to give the reader a taste of more sophisticated applications; mastery of it is not essential for the rest of the book.

Each Article is further discussed and elaborated in the specific subsequent Sessions indicated below:

Article I	Sessions 2 and 3
Article II	Sessions 4 through 9
Article III	Sessions 11 through 17
Article IV	Sessions 19 through 29
Article V	Sessions 30 and 31
Article VI	Sessions 32 and 33
Article VII	Sessions 34 and 35

The **Appendices**, written in a less leisurely manner, are intended to provide a rapid summary of some of the main possible links of the basic material of the course with various more advanced developments of modern mathematics.

Acknowledgements

First Edition
This book would not have come about without the invaluable assistance of many people:

Emilio Faro, whose idea it was to include the dialogues with the students in his masterful record of the lectures, his transcriptions of which grew into the Sessions;
Danilo Lawvere, whose imaginative and efficient work played a key role in bringing this book to its current form;
our students (some of whom still make their appearance in the book), whose efforts and questions contributed to shaping it;
John Thorpe, who accepted our proposal that a foundation for discrete mathematics *and* continuous mathematics could constitute an appropriate course for beginners.

Special thanks go to Alberto Peruzzi, who provided invaluable expert criticism and much encouragement. Many helpful comments were contributed by John Bell, David Benson, Andreas Blass, Aurelio Carboni, John Corcoran, Bill Faris, Emilio Faro, Elaine Landry, Fred Linton, Saunders Mac Lane, Kazem Mahdavi, Mara Mondolfo, Koji Nakatogawa, Ivonne Pallares, Norm Severo, and Don Schack, as well as by many other friends and colleagues. We are grateful also to Cambridge University Press, in particular to Roger Astley and Maureen Storey, for all their work in producing this book.

Above all, we can never adequately acknowledge the ever-encouraging generous and graceful spirit of Fatima Fenaroli, who conceived the idea that this book should exist, and whose many creative contributions have been irreplaceable in the process of perfecting it.

Thank you all,

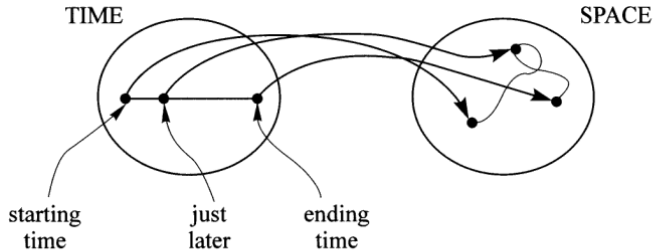
Buffalo, New York — F. William Lawvere
2009 — Stephen H. Schanuel

Second Edition
Thanks to the readers who encouraged us to expand to this second edition, and thanks to Roger Astley and his group at Cambridge University Press for their help in bringing it about.

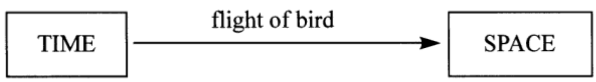
2009 — F. William Lawvere
Stephen H. Schanuel

Preview

The flight of a bird as a map from time to space



Schematically:

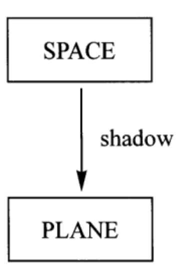


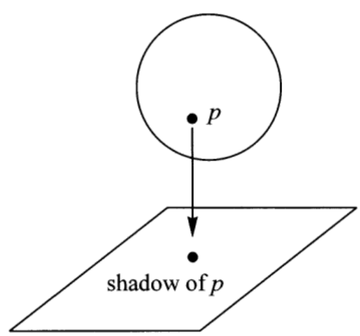
You have no doubt heard the legend; Galileo dropped a heavy weight and a light weight from the leaning tower of Pisa, surprising the onlookers when the weights hit the ground simultaneously. The study of vertical motion, of objects thrown straight up, thrown straight down, or simply dropped, seems too special to shed much light on general motion; the track of a dropped rock is straight, as any child knows. However, the motion of a dropped rock is not quite so simple; it accelerates as it falls, so that the last few feet of its fall takes less time than the first few. Why had Galileo decided to concentrate his attention on this special question of vertical motion? The answer lies in a simple equation:

$$\text{SPACE} = \text{PLANE} \times \text{LINE}$$

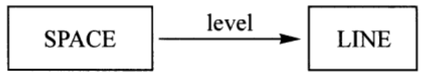
but it requires some explanation!

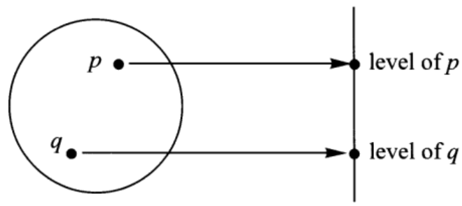
Two new maps enter the picture. Imagine the sun directly overhead, and for each point in space you'll get a shadow point on the horizontal plane:



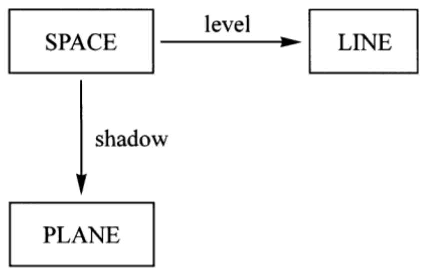


This is one of our two maps: the 'shadow' map from space to the plane. The second map we need is best imagined by thinking of a vertical line, perhaps a pole stuck into the ground. For each point in space there is a corresponding point on the line, the one at the same level as our point in space. Let's call this map 'level':



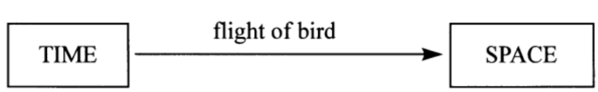


Together, we have:

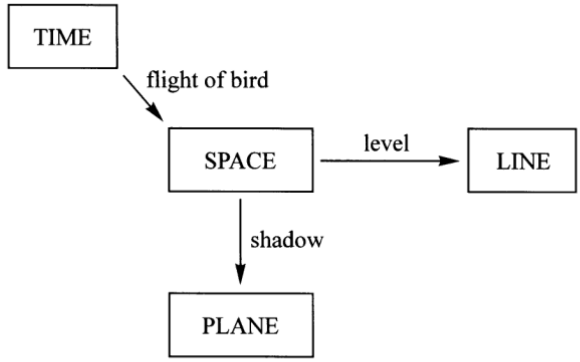


These two maps, 'shadow' and 'level', seem to reduce each problem about space to two simpler problems, one for the plane and one for the line. For instance, if a bird is in our space, and you know only the shadow of the bird and the level of the bird, then you can reconstruct the position of the bird. There is more, though. Suppose you have a motion picture of the bird's shadow as it flies, and a motion picture of its level – perhaps there was a bird-watcher climbing on our line, keeping always level with the bird, and you filmed the watcher. From these two motion pictures you can reconstruct the entire flight of the bird! So not only is a position in space reduced to a position in the plane and one on the line, but also a motion in space is reduced to a motion in the plane and one on the line.

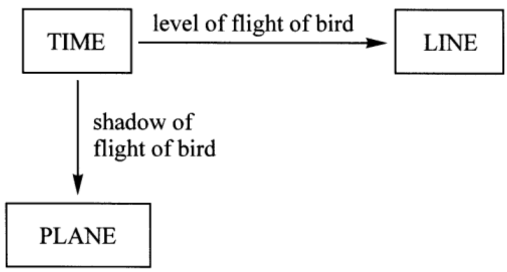
Let's assemble the pieces. From a motion, or flight, of a bird



we get two simpler motions by 'composing' the flight map with the shadow and level maps. From these three maps,



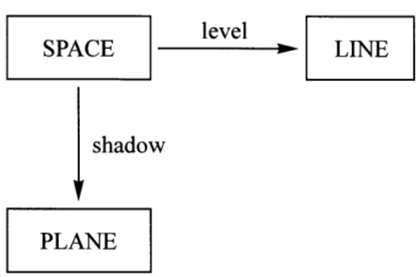
we get these two maps:



and now space has disappeared from the picture.

Galileo's discovery is that from these two simpler motions, in the plane and on the line, he could completely recapture the complicated motion in space. In fact, if the motions of the shadow and the level are 'continuous', so that the shadow does not suddenly disappear from one place and instantaneously reappear in another, the motion of the bird will be continuous too. This discovery enabled Galileo to reduce the study of motion to the special cases of horizontal and vertical motion. It would take us too far from our main point to describe here the beautiful experiments he designed to study these, and what he discovered, but I urge you to read about them.

Does it seem reasonable to express this relationship of space to the plane and the line, given by two maps,

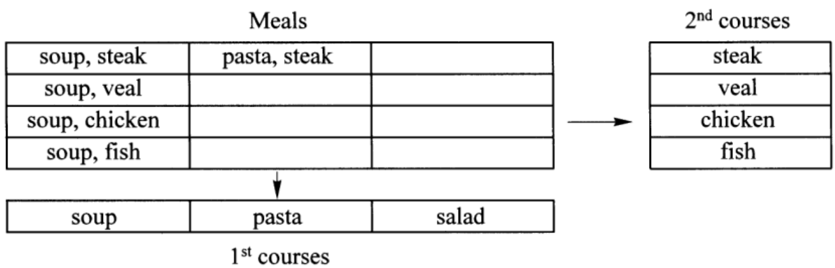


by the equation SPACE = PLANE × LINE? What do these maps have to do with multiplication? It may be helpful to look at some other examples.

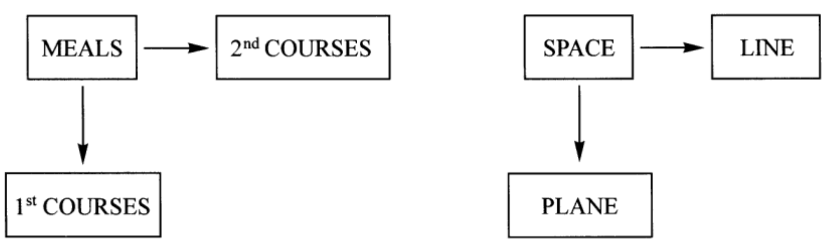
3. Other examples of multiplication of objects

Multiplication often appears in the guise of *independent choices*. Here is an example. Some restaurants have a list of options for the first course and another list for the second course; a 'meal' involves one item from each list. First courses: soup, pasta, salad. Second courses: steak, veal, chicken, fish.

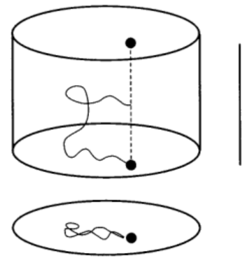
So, one possible 'meal' is: 'soup, then chicken'; but 'veal, then steak' is not allowed. Here is a diagram of the possible meals:



(Fill in the other meals yourself.) Notice the analogy with Galileo's diagram:

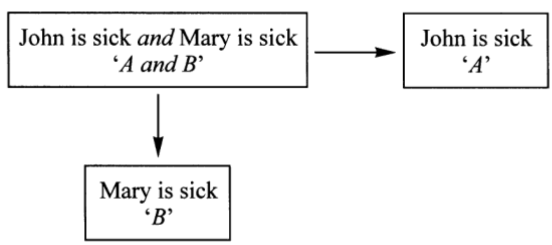


This scheme with three 'objects' and two 'maps' or 'processes' is the right picture of multiplication of objects, and it applies to a surprising variety of situations. The idea of multiplication is the same in all cases. Take for example a segment and a disk from geometry. We can multiply these too, and the result is a cylinder. I am not referring to the fact that the *volume* of the cylinder is obtained by multiplying the area of the disk by the length of the segment. The cylinder *itself* is the product, segment times disk, because again there are two processes or projections that take us from the cylinder to the segment and to the disk, in complete analogy with the previous examples.

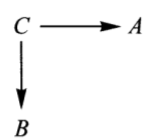


Every point in the cylinder has a corresponding 'level' point on the segment and a corresponding 'shadow' point in the disk, and if you know the shadow and level points, you can find the point in the cylinder to which they correspond. As before, the motion of a fly trapped in the cylinder is determined by the motion of its level point in the segment and the motion of its shadow point in the disk.

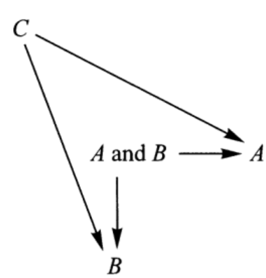
An example from logic will suggest a connection between multiplication and the word '*and*'. From a sentence of the form '*A and B*' (for example, 'John is sick *and* Mary is sick') we can deduce *A* and we can deduce *B*:



But more than that: to deduce the single sentence 'John is sick and Mary is sick' from some other sentence *C* is the same as deducing each of the two sentences from *C*. In other words, the two deductions



amount to one deduction $C \longrightarrow (A \text{ and } B)$. Compare this diagram



with the diagram of Galileo's idea.

PART I

The category of sets

A *map* of sets is a process for getting from one set to another. We investigate the *composition* of maps (following one process by a second process), and find that the algebra of composition of maps resembles the algebra of multiplication of numbers, but its interpretation is much richer.

ARTICLE I

Sets, maps, composition

A first example of a category

Before giving a precise definition of 'category', we should become familiar with one example, the **category of finite sets and maps**.

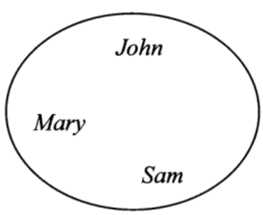
An object in this category is a finite *set* or *collection*. Here are some examples:

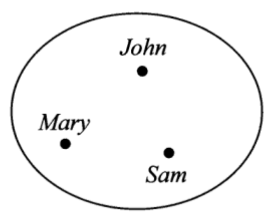
(the set of all students in the class) is one object,
(the set of all desks in the classroom) is another,
(the set of all the twenty-six letters in our alphabet) is another.

You are probably familiar with some notations for finite sets:

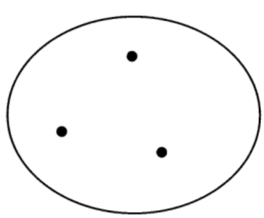
$$\{John,\ Mary,\ Sam\}$$

is a name for the set whose three elements are, of course, John, Mary, and Sam. (You know some infinite sets also, e.g. the set of all natural numbers: $\{0, 1, 2, 3, \ldots\}$.) Usually, since the order in which the elements are listed is irrelevant, it is more helpful to picture them as scattered about:





where a dot represents each element, and we are then free to leave off the labels when for one reason or another they are temporarily irrelevant to the discussion, and picture this set as:



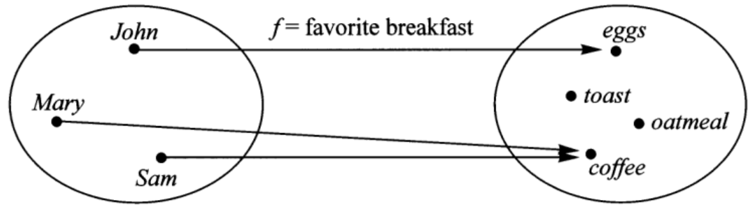
Such a picture, labeled or not, is called an *internal diagram* of the set.

A [m a p] f in this category consists of three things:

1. a set A, called the *domain* of the map,
2. a set B, called the *codomain* of the map,
3. a rule assigning to each element a in the domain, an element b in the codomain. This b is denoted by $f \circ a$ (or sometimes '$f(a)$'), read 'f of a'.

(Other words for map are 'function', 'transformation', 'operator', 'arrow', and 'morphism'.)

An example will probably make it clearer: Let $A = \{John, Mary, Sam\}$, and let $B = \{eggs, oatmeal, toast, coffee\}$, and let f assign to each person his or her favorite breakfast. Here is a picture of the situation, called the *internal diagram* of the map:

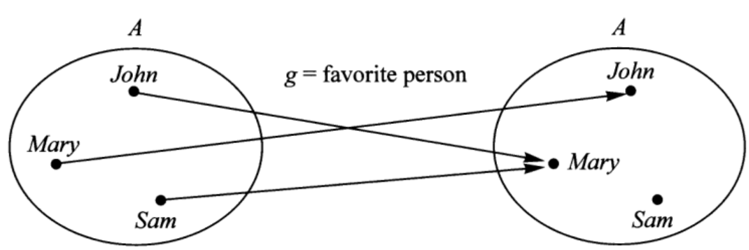


This indicates that the favorite breakfast of John is eggs, written $f(John) = eggs$, while Mary and Sam prefer coffee. Note some pecularities of the situation, because these are features of the internal diagram of any map:

(a) From each dot in the *domain* (here $\{John, Mary, Sam\}$), there is exactly one arrow leaving.
(b) To a dot in the *codomain* (here $\{eggs, oatmeal, toast, coffee\}$), there may be any number of arrows arriving: zero or one or more.

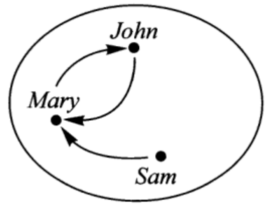
The important thing is: For each dot in the domain, we have exactly one arrow leaving, and the arrow arrives at some dot in the codomain.

Nothing in the discussion above is intended to exclude the possibility that A and B, the domain and codomain of the map, could be the *same* set. Here is an internal diagram of such a map g:

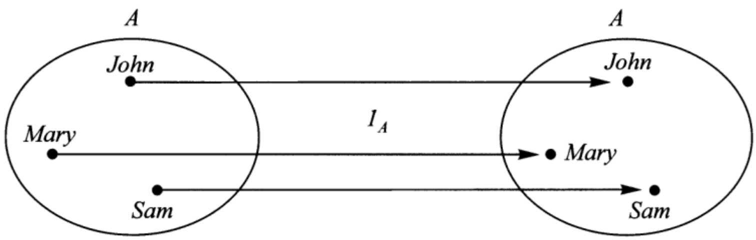


(Many 1950s movie plots are based on this diagram.)

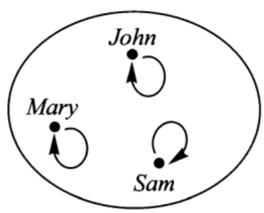
A map in which the domain and codomain are the same object is called an *endomap*. (Why? What does the prefix 'endo' mean?) For endomaps *only*, an alternative form of internal diagram is available. Here it is, for the endomap above:



For each object A, there is a special, especially simple, endomap which has domain and codomain both A. Here it is for our example:



Here is the corresponding special internal diagram, available because the map is an endomap:

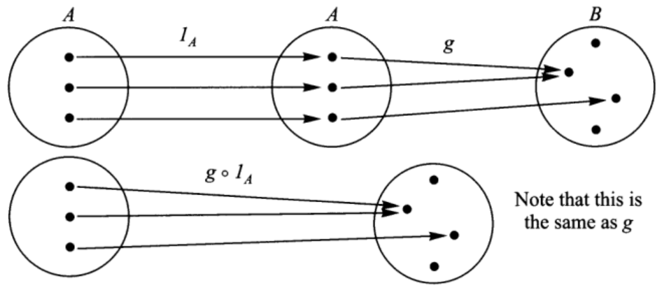


A map like this, in which the domain and the codomain are the same set A, *and* for each a in $A, f(a) = a$, is called an identity map. To state it more precisely, this map is 'the identity map from {*John*, *Mary*, *Sam*} to {*John*, *Mary*, *Sam*},' or 'the identity map on the object {*John*, *Mary*, *Sam*}.' (Simpler still is to give that object a short name, A = {*John*, *Mary*, *Sam*}; and then call our map 'the identity map on A', or simply '1_A'.)

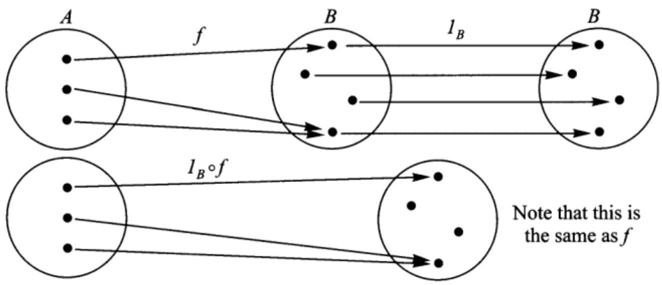
Sometimes we need a scheme to keep track of the domain and codomain, without indicating in the picture all the details of the map. Then we can use just a letter to stand for each object, and a single arrow for each map. Here are the *external diagrams* corresponding to the last five internal diagrams:

1. The identity laws:

(a)

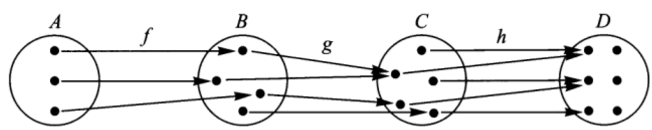


(b)

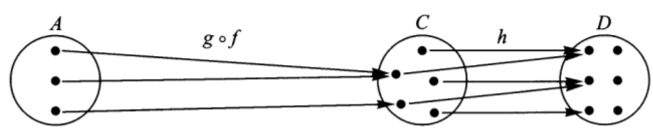


2. The associative law:

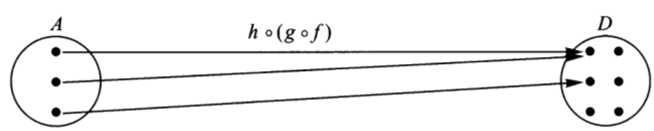
(i)



(ii)

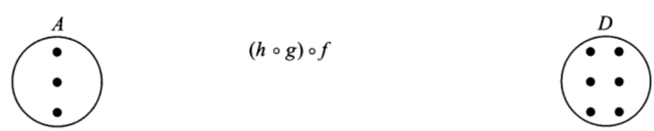


(iii)



(iv)

(v)



Exercise 1:
Check to be sure you understand how we got diagrams (ii) and (iii) from the given diagram (i). Then fill in (iv) and (v) yourself, starting over from (i). Then check to see that (v) and (iii) are the same.

Is this an accident, or will this happen for any three maps in a row? Can you give a simple explanation why the results

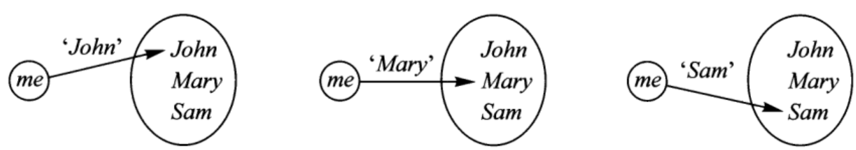
$$h \circ (g \circ f) \text{ and } (h \circ g) \circ f$$

will always come out the same, whenever we have three maps in a row

$$X \xrightarrow{f} Y \xrightarrow{g} Z \xrightarrow{h} W?$$

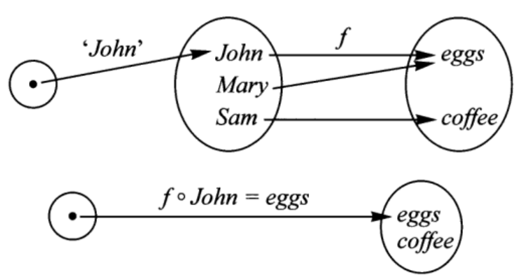
What can you say about *four* maps in a row?

One very useful sort of set is a 'singleton' set, a set with exactly one element. Fix one of these, say (*me*), and call this set '**1**'. Look at what the maps from **1** to {*John, Mary, Sam*} are. There are exactly three of them:



Definition: *A* **point** *of a set X is a map* $\mathbf{1} \longrightarrow X$.

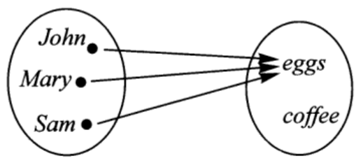
(If A is some familiar set, a map from A to X is called an 'A-element' of X; thus '**1**-elements' are points.) Since a point is a map, we can compose it with another map, and get a point again. Here is an example:



The equation $f \circ John = eggs$ is read 'f following *John* is *eggs*' or more briefly, 'f of *John* is *eggs*' (or sometimes 'f sends *John* to *eggs*').

To help familiarize yourself with the category of finite sets, here are some exercises. *Take* $A = \{John, Mary, Sam\}$, $B = \{eggs, coffee\}$ *in all of these.*

Exercise 2:
How many different maps f are there with domain A and codomain B? One example is



but there are lots of others: How many in all?

Exercise 3:
Same, but for maps $A \xrightarrow{f} A$

Exercise 4:
Same, but for maps $B \xrightarrow{f} A$

Exercise 5:
Same, but for maps $B \xrightarrow{f} B$

Exercise 6:
How many maps $A \xrightarrow{f} A$ satisfy $f \circ f = f$?

Exercise 7:
How many maps $B \xrightarrow{g} B$ satisfy $g \circ g = g$?

Exercise 8:
Can you find a pair of maps $A \xrightarrow{f} B \xrightarrow{g} A$ for which $g \circ f = 1_A$?
If so, how many such pairs?

Exercise 9:
Can you find a pair of maps $B \xrightarrow{h} A \xrightarrow{k} B$ for which $k \circ h = 1_B$?
If so, how many such pairs?

1. Guide

Our discussion of maps of sets has led us to the general definition of *category*, presented for reference on the next page. This material is reviewed in Sessions 2 and 3.

Definition of CATEGORY

A category consists of the DATA: ***. . . with corresponding notation***

(1) *OBJECTS* — $A, B, C, \ldots$

(2) *MAPS* — $f, g, h, \ldots$

(3) *For each map f, one object as DOMAIN of f and one object as CODOMAIN of f*

To indicate that f is a map, with domain A and codomain B, we write $A \xrightarrow{f} B$ (or $f\colon A \longrightarrow B$) and we say '$f$ is a map from A to B.'

(4) *For each object A an IDENTITY MAP, which has domain A and codomain A*

We denote this map by 1_A, so

$$A \xrightarrow{1_A} A$$

is one of the maps from A to A.

(5) *For each pair of maps*

$$A \xrightarrow{f} B \xrightarrow{g} C,$$

a COMPOSITE MAP

$$A \xrightarrow{g \text{ following } f} C$$

We denote this map by

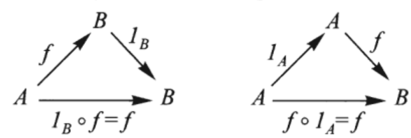
$$A \xrightarrow{g \circ f} C$$

(and sometimes say 'g of f').

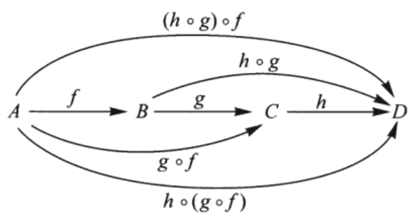
satisfying the following RULES:

These notations are used in the following external diagrams illustrating the rules:

(i) *IDENTITY LAWS: If* $A \xrightarrow{f} B$, *then* $1_B \circ f = f$ *and* $f \circ 1_A = f$



(ii) *ASSOCIATIVE LAW:*
If $A \xrightarrow{f} B \xrightarrow{g} C \xrightarrow{h} D$,
then $(h \circ g) \circ f = h \circ (g \circ f)$



The associative law allows us to leave out the parentheses and just write '$h \circ g \circ f$', which we read as 'h following g following f'. A longer composite like $h \circ g \circ f \circ e \circ d$ is also unambiguous; all ways of building it by composition of pairs give the same result.

Hidden in items (4) and (5) above are the BOOKKEEPING rules. Explicitly these are:

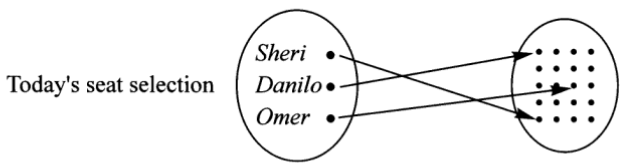
the domain and codomain of 1_A are both A;
$g \circ f$ is only defined if the domain of g is the codomain of f;
the domain of $g \circ f$ is the domain of f and the codomain of $g \circ f$ is the codomain of g.

SESSION 2

Sets, maps and composition

1. Review of Article I

Before discussing some of the exercises in Article I, let's have a quick review. A **set** is any collection of things. You know examples of infinite sets, like the set of all natural numbers, $\{0, 1, 2, 3, \ldots\}$, but we'll take most of our examples from finite sets. Here is a typical **internal diagram** of a **function**, or **map**:

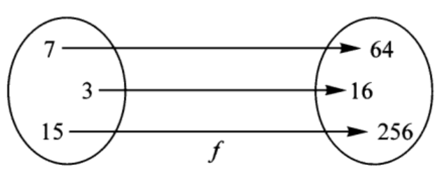


Other words that mean the same as *function* and *map* are **transformation**, **operator**, **morphism**, and **functional**; the idea is so important that it has been rediscovered and renamed in many different contexts.

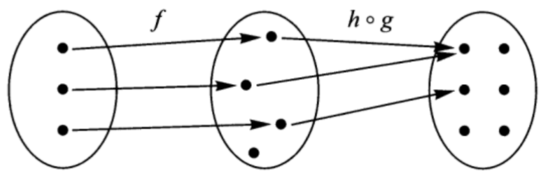
As the internal diagram suggests, to have a **map** f **of sets** involves three things:

1. a set A, called the **domain** of the map f;
2. a set B, called the **codomain** of the map f; and then the main ingredient:
3. a **rule** (or process) for f, assigning to each element of the domain A exactly one element of the codomain B.

That is a fairly accurate description of what a map is, but we also need a means to tell when two different rules give the same map. Here is an example. The first map will be called f and has as domain and as codomain the set of all natural numbers. The rule for f will be: '*add* 1 *and then square*'. (This can be written in mathematical shorthand as $f(x) = (x+1)^2$, but that is not important for our discussion.) Part of the internal picture of f is:



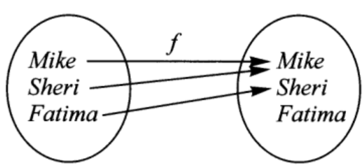
The second map will be called g. As domain and codomain of g we take again the set of all natural numbers, but the rule for g will be '*square the input*, *double the input*, which Chad has done like this:



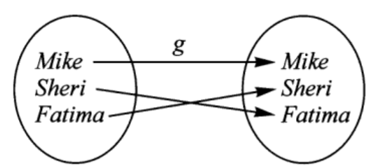
Is this correct? Not quite, because we are supposed to draw two maps, and the thing drawn for $h \circ g$ is not a map; one of the points of the domain of $h \circ g$ has been left without an assigned output. This deficiency won't matter for the next step, because that information is going to get lost anyhow, but it belongs in this step and it is incorrect to omit it. Chad's trouble was that in drawing $h \circ g$, he noticed that the last arrow would be irrelevant to the composite $(h \circ g) \circ f$, so he left it out.

CHAD: It seems the principle is like in multiplication, where the order in which you do things doesn't matter; you get the same answer.

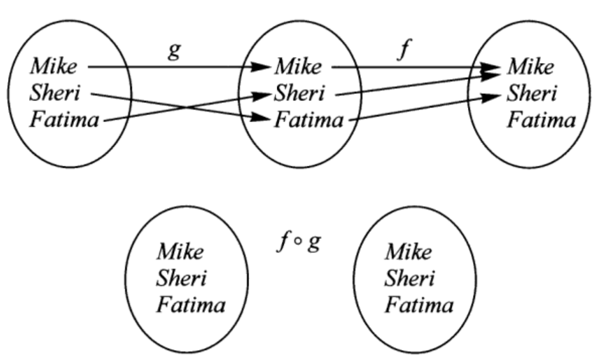
I am glad you mention order. Let me give you an example to show that the order *does* matter. Consider the two maps



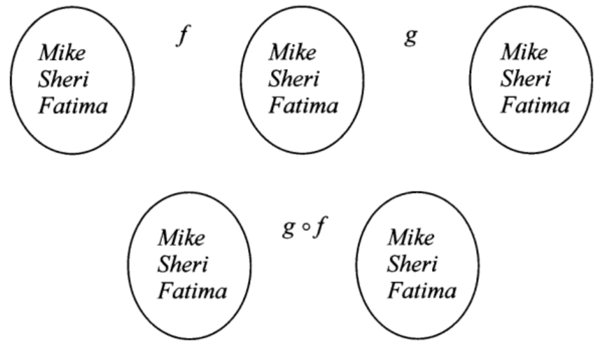
and



Work out the composite $f \circ g$, and see what you get:



Now work out the composite in the opposite order:



The two results are different. In composition of maps the order matters.

When I was little I had a large family, and in large families there are always many small chores to be done. So my mother would say to one of us: 'Wouldn't you like to wash the dishes?' But as we grew, two or more tasks were merged into one, so that my mother would say: 'Wouldn't you like to **wash and then rinse** the dishes?' or: '**scrape and wash** and then **rinse and dry** the dishes?' And you can't change the order. You'll make a mess if you try to dry before scraping. The 'associative law for tasks' says that the two tasks:

(**scrape** then **wash**) then (**rinse** then **dry**)

and

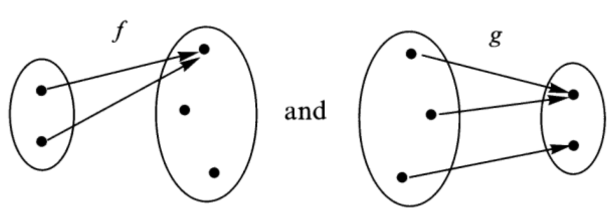
scrape then [(**wash** then **rinse**) then **dry**]

accomplish the same thing. All that matters is the order, not when you take your coffee break. All the parentheses are unnecessary; the composite task is:

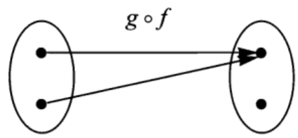
scrape then **wash** then **rinse** then **dry**

Think about this and see if it suggests an explanation for the associative law. Then look back at the pictures, to see how you can *directly* draw the picture for a composite of several maps without doing 'two at a time'.

Several students have asked why some arrows disappear when you compose two maps, i.e. when you pass from the diagrams

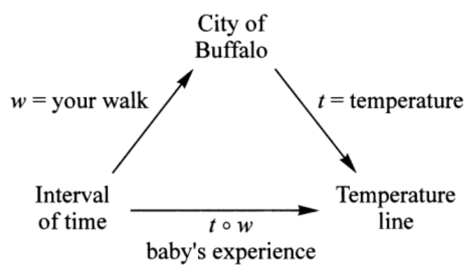


to the diagram for 'g following f'



To understand this you should realize that the composite of two maps is supposed to be another map, so that it just has a domain, a codomain and a rule. The pasting together of two diagrams is not the composite map, it is just a rule to find the composite map, which can be done easily by 'following the arrows' to draw the diagram of the resulting (composite) map. The point of erasing all the irrelevant detail (like the extra arrows) is that the simplified picture really gives a different rule which defines the same map, but a simpler rule.

Suppose you carry a sleeping baby on a brief walk around town, first walking in the hot sun, then through the cool shade in the park, then out in the sun again.



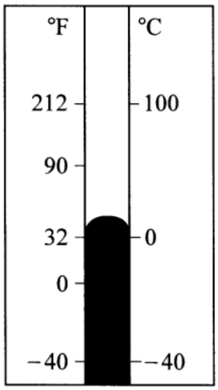
The map w assigns to each instant your location at that time, and the map t assigns to each spot in Buffalo the temperature there. ('Temperature line' has as its points physical temperatures, rather than numbers which measure temperature on some scale; a baby is affected by temperature before learning of either Fahrenheit or Celsius.) The baby was hot, then cool, then hot again, but doesn't know the two maps that were composed to get this one map.

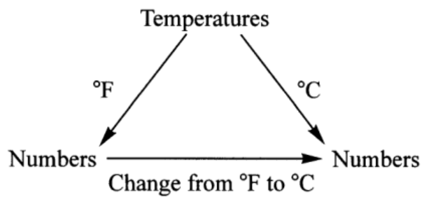
2. An example of different rules for a map

The measurement of temperature provides a nice example of different rules for a 'numerical' map. If one looks at a thermometer which has both scales, Celsius and Fahrenheit, it becomes obvious that there is a map,

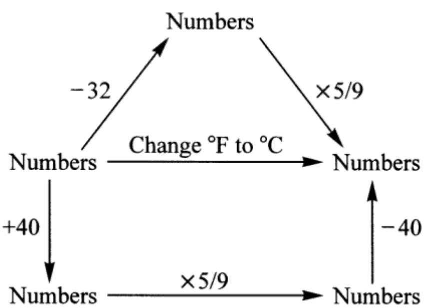
$$\text{Numbers} \xrightarrow{\text{change from Fahrenheit to Celsius}} \text{Numbers}$$

which sends the measure in degrees Fahrenheit of a temperature to the measure in degrees Celsius of the same temperature. In other words, it is the map that fits in the diagram





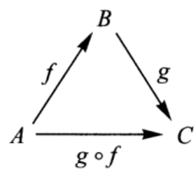
How is this map calculated? Well, there are several possible *rules*. One of them is: 'subtract 32, then multiply by 5/9.' Another is: 'add 40, multiply by 5/9, then subtract 40.' Notice that each of these rules is itself a composite of maps, so that we can draw the following diagram:



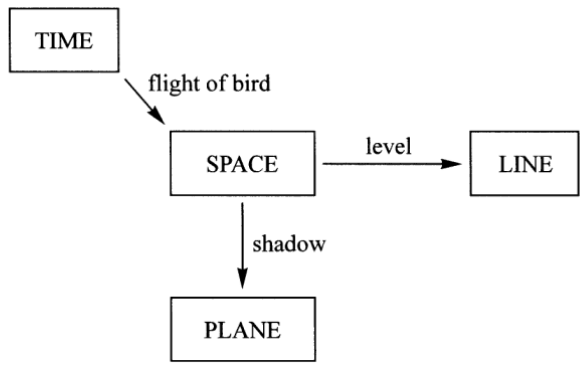
The above example illustrates that a single map may arise as a composite in several ways.

3. External diagrams

The pasting of the diagrams to calculate composition of maps is nice because from it you can read what f does, what g does, and also what the composite $g \circ f$ does. This is much more information than is contained in $g \circ f$ alone. In fact internal diagrams aren't always drawn. We use schematic diagrams like those in our 'temperature' example, or this:



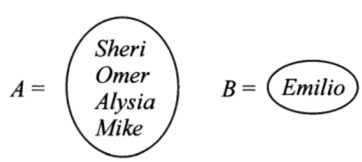
These are called **external diagrams** because they don't show what's going on inside. In Session 1 we met an external diagram when discussing Galileo's ideas:



4. Problems on the number of maps from one set to another

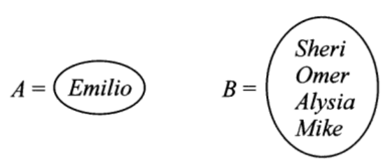
Let's work out a few problems that are not in Article I. How many maps are there from the set A to the set B in the following examples?

(1)



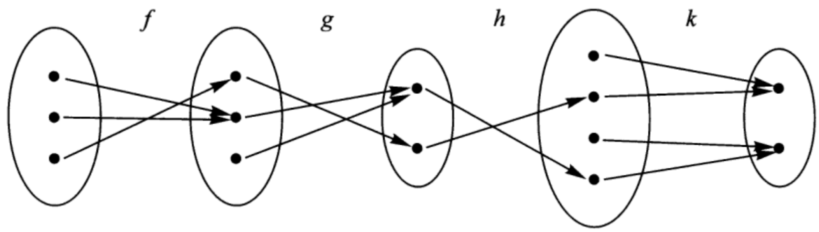
Answer: Exactly one map; all elements of A go to *Emilio*.

(2)

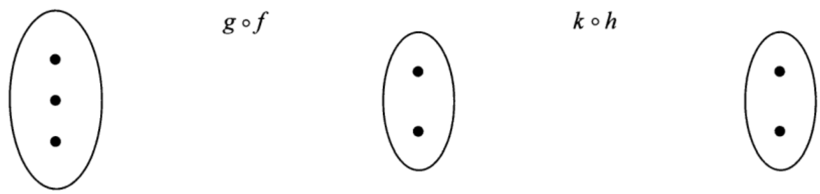


Answer: There are four maps because all a map does is to tell where Emilio goes, and there are four choices for that.

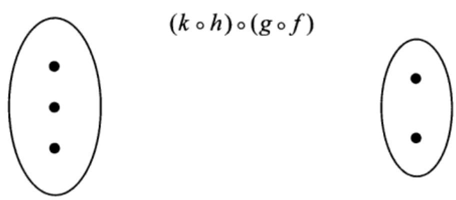
(3) Now the set A is ... What shall I say? Ah! The set of all purple people-eaters in this room, and B is as before:



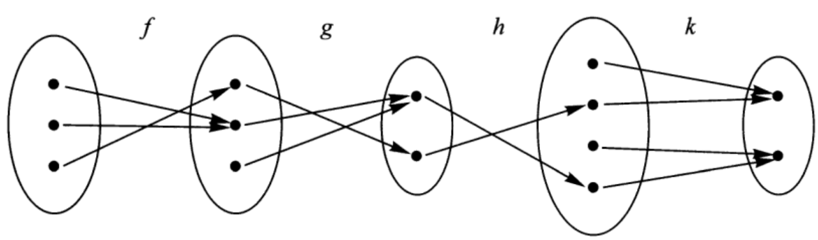
to



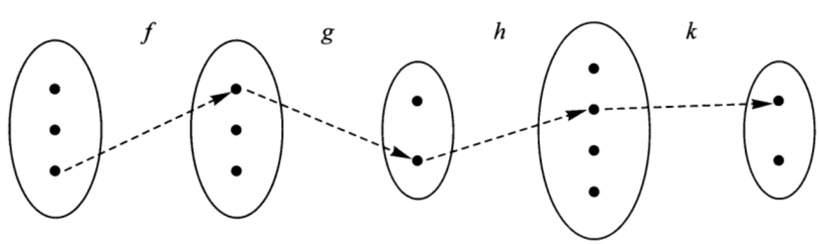
(Fill in any missing arrows yourself.) Then, repeating the process, we get



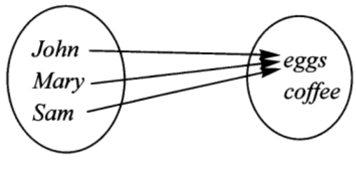
But this piecemeal work is unnecessary. The analogy of *scrape*, then *wash*, then *rinse*, then *dry* is meant to suggest that we can go from the beginning to the end in one step, if we stick to the idea that the diagram



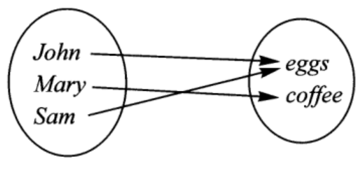
itself gives a good rule for calculating the composite $k \circ h \circ g \circ f$. Just 'look at the whole diagram and follow the arrows'; for example:



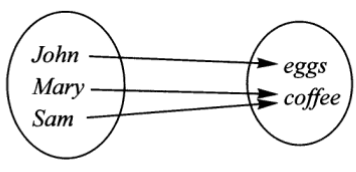
Now let's see if we can find a way to tell the number of maps between any two finite sets. For that we should start by working out simple cases. For example, Exercise 4 is to find the number of maps from a three-element set to a two-element set. How can we do this? The most immediate way I can think of is to draw them (taking care not to repeat any and not to omit any), and then count them. Say we begin with



Then we can do something else,



and then perhaps



and let's see .... Do we have all the maps that send John to eggs? Right, we need one more, sending Mary to eggs and Sam to coffee. So there are four maps that send 'John' to 'eggs', and I hope it is clear that there are also four maps that send 'John' to 'coffee', and that their diagrams are the same as the four above, but changing the arrow from 'John'. Thus the answer to this exercise is 8 maps. The same method of drawing all possibilities should give you the answers to Exercises 5, 6, and 7, so that you can start to fill in a table like this:

Number of DOMAIN	3	3	2	2	
Number of CODOMAIN	2	3	3	2	
Number of MAPS	8	27	9	4	

hoping to find a pattern that may allow you to answer other cases as well.

ALYSIA: It seems that the number of maps is equal to the number of elements of the codomain raised to a power (the number of elements of the domain.)

That's a very good idea. One has to discover the reason behind it. Let's see if it also works with the extreme cases that we found at the end of last session.

Adding those results to our table we get:

Number of DOMAIN	3	3	2	2	4	1	0	4	0
Number of CODOMAIN	2	3	3	2	1	4	4	0	0
Number of MAPS	8	27	9	4	1	4	1	0	1
	2^3	3^3	3^2	2^2	1^4	4^1	4^0	0^4	0^0

and

n	1	0	$n \neq 0$
1	n	n	0
1	n	1	0
1^n	n^1	n^0	0^n

where n is any natural number, with the only exception that in the last column it must be different from zero. Now you should think of some reason that justifies this pattern.

CHAD: For every element of the domain there are as many possibilities as there are elements in the codomain, and since the choices for the different elements of the domain are independent, we must multiply all these values, so the number of maps is the number of elements of the codomain multiplied by itself as many times as there are elements in the domain.

Chad's answer seems to me very nice. Still we might want a little more explanation. Why multiply? What does 'independent' mean? If John has some apples and Mary has some apples, aren't Mary's apples independent of John's? So, if you put them all in a bag do you add them or multiply them? Why?

Going back to Alysia's formula for the number of maps from a set A to a set B, it suggests a reasonable notation, which we will adopt. It consists in denoting the set of maps from A to B by the symbol B^A, so that our formula can be written in this nice way

$$\#(B^A) = (\#B)^{(\#A)} \quad \text{or} \quad |B^A| = |B|^{|A|}$$

where the notations $\#A$ and $|A|$ are used to indicate the number of elements of the set A. The notation $\#A$ is self-explanatory since the symbol # is often used to denote 'number', while $|A|$ is similar to the notation used for the absolute value of a number. The bars indicate that you forget everything except the 'size'; for numbers you forget the sign, while for sets you forget what the elements are, and remember only how many of them there are. So, for example, if

$P =$ (*Ian*, *Chad*) $\qquad R =$ (*living room*, *dining room*)

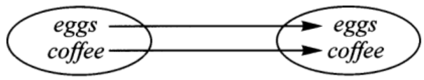
then we wouldn't say $P = R$, but rather $|P| = |R|$. To remember which set goes in the base and which one in the exponent you can imagine that the maps are lazy, so that they go down from the exponent to the base. Another way to remember this is to think of an especially simple case, for instance the case in which the codomain has only one element, and therefore the set of maps has also only one element (and, of course, remember that $1^n = 1$).

In Exercise 9, we don't ask for the total number of maps from one set to another, but only the number of maps g

from 

to 

such that $g \circ g = g$. Can you think of one? Right,



This is the first example anybody would think of. Remember from Article I that this map is called an **identity map**. Any set B has an identity map, which is denoted

$$B \xrightarrow{1_B} B$$

and sends each element of the domain to itself. This map certainly satisfies $1_B \circ 1_B = 1_B$. In fact it satisfies much more; namely, for any map $A \xrightarrow{f} B$, and any map $B \xrightarrow{g} C$,

$$1_B \circ f = f \quad \text{and} \quad g \circ 1_B = g$$

(These two equations give two different proofs of the property $1_B \circ 1_B = 1_B$: one by taking $f = 1_B$ and one by taking $g = 1_B$.) These properties of the identity maps are like the property of the number 1, that multiplied by any number gives the same number. So, identity maps behave for composition as the number 1 does for multiplication. That is the reason a '1' is used to denote identity maps. What's another map g

from 

to 

which satisfies $g \circ g = g$? What about the map

This map also has the property, since the composite

is

Now try to do the exercises again if you had difficulty before. One suggestion is to look back and use the special diagrams available only for endomaps explained in Article I.

Here are some exercises on the 'bookkeeping rules' about domains and codomains of composites.

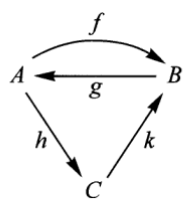
Exercise 1:
A, B, and C are three different sets (or even three different objects in any category); f, g, h, and k are maps with domains and codomains as follows:

$$A \xrightarrow{f} B, \quad B \xrightarrow{g} A, \quad A \xrightarrow{h} C, \quad C \xrightarrow{k} B$$

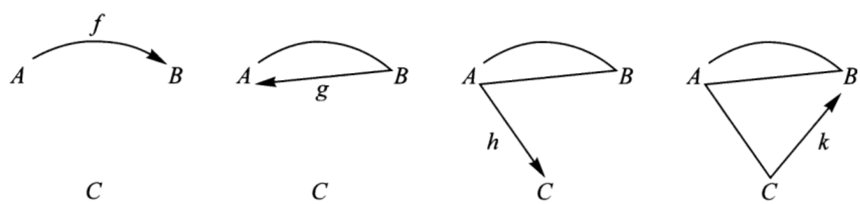
Two of the expressions below make sense. Find each of the two, and say what its domain and codomain are:

(a) $k \circ h \circ g \circ f$ (b) $k \circ f \circ g$ (c) $g \circ f \circ g \circ k \circ h$

Exercise 2:
Do Exercise 1 again, first drawing this diagram:



Now just read each expression from right to left; so (a) is 'f then g then h then k.' As you read, follow the arrows in the diagram with your finger, like this:



The composite makes sense, and goes from A to B. See how much easier this external diagram makes keeping track of domains, etc.

ARTICLE II

Isomorphisms

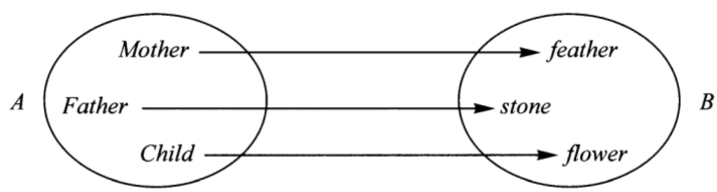
Retractions, sections, idempotents, automorphisms

1. Isomorphisms

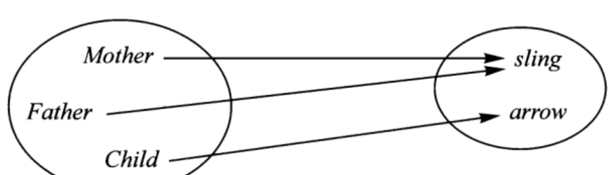
It seems probable that before man learned to count, it was first necessary to notice that sometimes one collection of things has a certain kind of resemblance to another collection. For example, these two collections

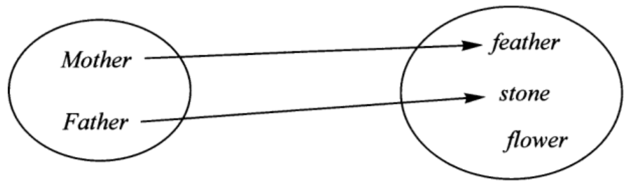
are similar. In what way? (Remember that numbers had not yet been invented, so it is not fair to say 'the resemblance is that each has three elements.')

After some thought, you may arrive at the conclusion that the resemblance is actually given by choosing a *map*, for instance this one:

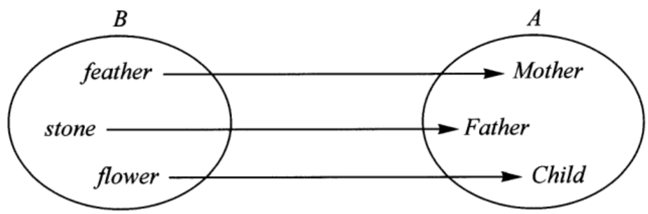


What special properties does this map f have? We would like them to be expressed entirely in terms of composition of maps so that we can later use the same idea in other categories, as well as in the category of finite sets. The properties should *exclude* maps like these:





The crucial property that f has, and the other two maps do not have, is that there is an ***inverse map*** g for the map f. Here is a picture of g:



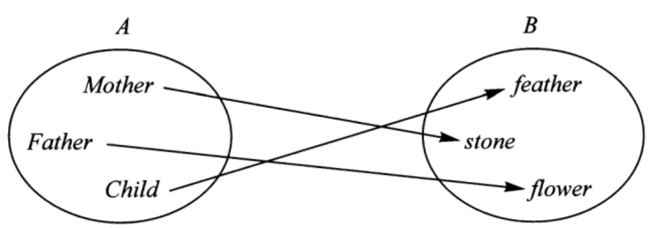
The important thing to notice is that g and f are related by *two equations*

$$g \circ f = 1_A \quad f \circ g = 1_B$$

As we will see, neither of these equations by itself will guarantee that A and B have the same size; we need both. This gives rise to the following concepts:

Definitions: *A map* $A \xrightarrow{f} B$ *is called an* **isomorphism**†, *or* **invertible map**, *if there is a map* $B \xrightarrow{g} A$ *for which* $g \circ f = 1_A$ *and* $f \circ g = 1_B$.
A map g related to f by satisfying these equations is called an **inverse for** *f.*
Two objects A and B are said to be **isomorphic** *if there is at least one isomorphism* $A \xrightarrow{f} B$

Notice that there are other isomorphisms from {*Mother*, *Father*, *Child*} to {*feather*, *stone*, *flower*}, for instance



but to show that these two sets are isomorphic, we only need to find *one* of the many – how many? – isomorphisms from A to B.

Once mankind had noticed this way of finding 'resemblance' between collections, it was probably not too long before some names for the 'sizes' of small collections – words like *pair*, or *triple* – came about. But first a crucial step had to be made: one

†The word *isomorphism* comes from Greek: *iso* = same; *morph* = shape, form; though in our category of finite sets same *size* might seem more appropriate.

had to see that the notion of *isomorphic* or 'equinumerous' or 'same-size', or whatever it was called (if indeed it had any name at all yet), has certain properties:

Reflexive: A is isomorphic to A.
Symmetric: If A is isomorphic to B, then B is isomorphic to A.
Transitive: If A is isomorphic to B, and B is isomorphic to C, then A is isomorphic to C.

Surprisingly, all these properties come directly from the associative and identity laws for composition of maps.

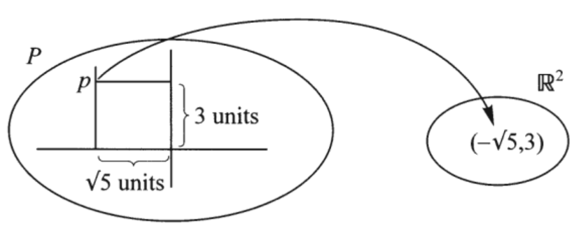
Exercise 1:
(R) Show that $A \xrightarrow{1_A} A$ is an isomorphism.
(Hint: find an inverse for 1_A.)
(S) Show that if $A \xrightarrow{f} B$ is an isomorphism, and $B \xrightarrow{g} A$ is an inverse for f, then g is also an isomorphism.
(Hint: find an inverse for g.)
(T) Show that if $A \xrightarrow{f} B$ and $B \xrightarrow{k} C$ are isomorphisms, $A \xrightarrow{k \circ f} C$ is also an isomorphism.

These exercises show that the three properties listed before them are correct, but the exercises are more explicit: solving them tells you not just that certain maps *have* inverses, but how actually to *find* the inverses.

All this may seem to be a lot of fuss about what it is that all three-element sets have in common! Perhaps you will be partially persuaded that the effort is worthwhile if we look at an example from geometry, due to Descartes. P is the plane, the plane from geometry that extends indefinitely in all directions. $\mathbb{R}^2$ is the set of all lists of two real numbers (positive or negative infinite decimals like $\sqrt{3}$ or $-\pi$ or 2.1397). Descartes' analytic approach to geometry begins with an isomorphism

$$P \xrightarrow{f} \mathbb{R}^2$$

assigning to each point its coordinate-pair, *after* choosing two perpendicular lines in the plane and a unit of distance:



The map f assigns to each point p in the plane a pair of numbers, called the 'coordinates of p in the chosen coordinate system'. (What does the *inverse* map g do? It must assign to each pair of numbers, like $(\pi, 7)$, a point. Which point?)

By systematically using this kind of isomorphism, Descartes was able to *translate* difficult problems in geometry, involving lines, circles, parabolas, etc., into easier problems in algebra, involving equations satisfied by the coordinate-pairs of the points on the curves. We still use this procedure today, and honor Descartes by calling these coordinate systems 'cartesian coordinates'. Our notion of 'isomorphism' is what makes this technique work perfectly: we can 'translate' any problem about a plane – i.e. apply the map f to it – to a problem about pairs of numbers. This problem about pairs of numbers may be easier to solve, because we have many algebraic techniques for dealing with it. Afterwards, we can 'translate back' – i.e. apply the inverse map for f – to return to the plane. (It should be mentioned that Descartes' method has also proved useful in the opposite way – sometimes algebraic problems are most easily solved by translating them into geometry!)

You will notice that we have sneaked in something as we went along. Before, we talked of *an* inverse for f, and now we have switched to *the* inverse for f. This is justified by the following exercise, which shows that, while a map f may not have any inverse, it *cannot* have two different inverses!

Exercise 2:
Suppose $B \xrightarrow{g} A$ and $B \xrightarrow{k} A$ are *both* inverses for $A \xrightarrow{f} B$. Show that $g = k$.

Since the algebra of composition of maps resembles the algebra of multiplication of numbers, we might expect that our experience with numbers would be a good guide to understanding composition of maps. For instance, the associative laws are parallel:

$$f \circ (g \circ h) = (f \circ g) \circ h$$
$$3 \times (5 \times 7) = (3 \times 5) \times 7$$

But we need to take some care, since

$$f \circ g \neq g \circ f$$

in general. The kind of care we need to take is exemplified in our discussion of inverses. For numbers, the 'inverse of 5', or $\frac{1}{5}$, is characterized by: it is *the* number x such that $5 \times x = 1$; but for the inverse of a map, we needed *two* equations, not just one.

More care of this sort is needed when we come to the analog of division. For numbers, $\frac{3}{5}$ (or $3 \div 5$) is characterized as *the* number x for which

$$5 \times x = 3;$$

but it can also be obtained as

$$x = \tfrac{1}{5} \times 3$$

Thus for numbers we really don't need division in general; once we understand inverses (like $\frac{1}{5}$) and multiplication, we can get the answers to more general division problems by inverses and multiplication. We will see that a similar idea can be used for maps, but that not all 'division problems' reduce to finding inverses; and also that there are interesting cases of 'one-sided inverses', where $f \circ g$ is an identity map but $g \circ f$ is not.

Before we go into general 'division problems' for maps, it is important to master isomorphisms and some of their uses. Because of our earlier exercise, showing that a map $A \xrightarrow{f} B$ can have at most *one* inverse, it is reasonable to give a special name, or symbol, to that inverse (when there is an inverse).

Notation: If $A \xrightarrow{f} B$ has an inverse, then the (one and only) inverse for f is denoted by the symbol f^{-1} (read 'f-inverse', or 'the inverse of f'.)

Two things are important to notice:

1. To show that a map $B \xrightarrow{g} A$ satisfies $g = f^{-1}$, you must show that

$$g \circ f = 1_A \quad \textit{and} \quad f \circ g = 1_B$$

2. If f does *not* have an inverse, then the symbol 'f^{-1}' *does not stand for anything*; it's a nonsense expression like 'grlbding' or '$\frac{3}{0}$'.

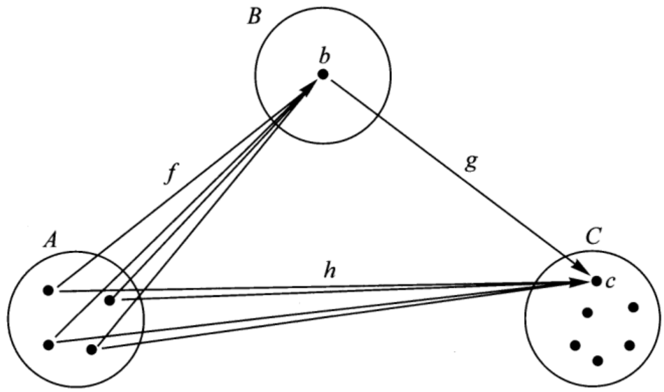
Exercise 3:
If f has an inverse, then f satisfies the two cancellation laws:

(a) If $f \circ h = f \circ k$, then $h = k$.
(b) If $h \circ f = k \circ f$, then $h = k$.

Warning: The following 'cancellation law' is *not* correct, even if f has an inverse.

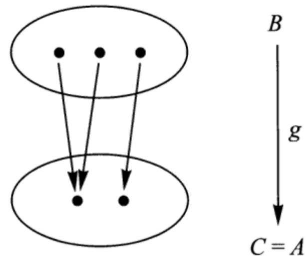
(c) (wrong): If $h \circ f = f \circ k$, then $h = k$.

When an exercise is simply a statement, the task is to prove the statement. Let's do part (a). We assume that f has an inverse and that $f \circ h = f \circ k$, and we try to show that $h = k$. Well, since $f \circ h$ and $f \circ k$ are the same map, the maps $f^{-1} \circ (f \circ h)$ and $f^{-1} \circ (f \circ k)$ are also the same:

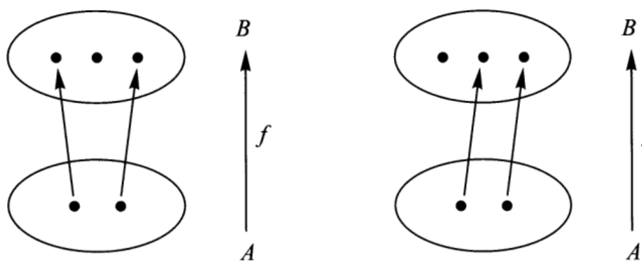


Example 2, a 'choice' problem

Now consider the following example in which B has three elements and $h = 1_A$ where $A = C$ has two elements, while $B \xrightarrow{g} C$ is a given map with the property that every element of C is a value of g, such as



How many maps f can we find with $g \circ f = 1_A$? Such an f must be a map from $A = C$ to B and satisfy $g(f(x)) = x$ for both elements x. That is, f must 'choose' for each x an element z of B for which $g(z) = x$. From the picture we see that this determines the value of f at one x but leaves two acceptable choices for the value of f at the other x. Therefore there are exactly two solutions f to the question as follows:



On the other hand, suppose the first of these f is considered given, and we ask for all maps g for which $g \circ f = 1_A$, a 'determination' problem. The equation $g(f(x)) = x$ can now be interpreted to mean that for each element of B which is of the form $f(x)$, g is forced to be defined so as to take it to x itself; there is one