CHAPTER **A**

Analysis and Science

On Sunday morning, Henri is somewhat drowsy but not as much as could be expected after a heavy beer session. The main source for all the strange ideas he had last night was the inspiration created during a conversation with another mind. Henri wonders whether he would have noticed anything if he had participated a psychological study without knowing it at all. He considers a format in which some psychologists had agreed with the bar owner that during a certain period, say from 11 pm to 1 am, they are permitted to conduct an experiment.

When a customer orders a certain popular beer, the bartender would offer either authentic beer or a similar beer without any alcohol. The selection shall be made randomly. The experimenters shall follow the process and but they should not be aware of which customers drink alcohol-free beer. The glasses might, for instance, be marked by a small sign that can be used to track afterwards which glasses contained real beer. In an optimal arrangement, even the bartender should not know when he offers alcohol-free beer.

The objective of the experiment could be, Henri continues his thinking, to study the effect of alcohol on the emotions in a realistic environment. When the customer leaves the bar, the experimenter would ask a couple of short but clear questions about the feelings the customer have had during the evening. The experimenters would also ask permission to call the customer the next afternoon to ask additional questions about the customer's feelings during the next day.

Only when all the questions have been answered, the experimenters reveal the study arrangement. They may also offer a compensation for those customers that did not get the beer that they had ordered.

Henri is not sure whether this kind of experiment could be done in reality. It might be technically possible, but he was suspicious whether that kind of experiment was permitted. Still, he thinks that the results of the study might provide a lot of interesting and even surprising results. Henri was quite sure that after several pints of beer he would not notice anything strange with the alcohol-free beer. Henri is not sure whether he would be more or less happy, but probably he would be less tired if he drinks less alcohol, particularly on the day after.

To be sure, the reason to be happy now was not the beers drunk last night but the chain of messages sent between him and Irene. Or, actually, the main reason for happiness was the relationship with Irene in which the messages had a small but essential role. The other source of happiness was the fascinating discussions he had with the Umpire. In general, Henri ponders, whatsoever the source of happiness was, it was somehow related to social interactions.

However, Henri is not satisfied with the mere idea that social interactions create happiness; he wants to understand the fundamental phenomenon. Even though Henri is not a biologist, he sometimes thinks that literally every property of a living thing must be a result of evolution. If I happen to like Irene, that feeling must serve the purpose of evolution, because, Henri rationalizes; otherwise, the process that creates the feeling could not have survived in the minds of men.

If happiness is also created by evolution, then the purpose of happiness must be to increase the probability of survival and reproduction. Thus if someone believes that happiness is the purpose of life, he should actually come to the conclusion that the hidden purpose of life is to maintain life.

He had numerous times pondered the following dilemma. He imagines a situation in which he has two options to react to a tricky situation. In the first option, he certainly *knows* that it will increase his happiness while endangering his survival and reproduction capabilities. In contrast, the other option will certainly reduce his happiness while significantly increasing his survival probability. Which one of the two options shall he select?

How could he describe this dilemma in a scientific manner? Henri is not sure, but still has the vague impression that if there is any correct answer, it likely depends on some fundamental assumptions that shall be clarified first. What does happiness mean, anyway? Surely, some outstanding scientists must have thought about this question, Henri believes. He is sure that if scientists are able reveal the secrets of distant stars, they must also be able to reveal the secrets of happiness.

However, why, Henri wonders, do I need to think about all these scientific matters on Sunday evening? Does this serve the intention of a higher-level being? Henri has the vague impression that someone tries to put ideas in his head. Maybe everything that is happening just now is a part of bigger plan, including his beer session yesterday and his relationship problems with Irene. He tries to observe his own life from the perspective of an outsider without much success. Maybe he could write a brief story about his life, then forget it, and finally read the story without any a priori knowledge. Or draw a picture that includes himself, an observer, and a meta-level observer that is at the same time himself.

Science and reality

Should the studies of communications ecosystem be considered a sub-field of science? Obviously, any ecosystem can be studied scientifically. However, that is not the primary viewpoint of this book although I highly respect scientific studies conducted according to scientific methods. In the fields of mathematics and physics, it is reasonable to think that scientific knowledge is systematically built on the basis of previous knowledge—at least most of the time. In the case of ecosystems (including communications ecosystem), I would rather think of a new insight as a new plant that prospers for a while and afterwards provides fertile soil for new insights. Communications ecosystem is, also in this sense, an ecosystem rather than an edifice that is purposefully built according to a (pre)determined architecture.

This does not mean that we could not utilize the scientific knowledge to understand how ecosystems work whenever relevant knowledge is available. Figure A.1 illustrates the position of a "scientific core" consisting of mathematics and logic in a more general framework that is used in the process of understanding reality. Mathematics is, of course, inevitably a part of reality because human beings and all their activities are real. In Figure A.1, *reality* refers to everything that is outside the minds of human beings. The mind shown in Figure A.1 describes primarily the area that communications ecosystem experts (CEEs) have to cover: in order to tackle the challenges during their career they must able to carry out strict reasoning and to perceive reality in a holistic, open manner. Furthermore, they have to able to combine those two challenging tasks. The aim of this book is to support that endeavor.

Someone may suppose that in principle reality could be explained entirely by mathematics; the wormhole in Figure A.1 illustrates that possibility. In a way, that vision represents an extreme type of

> **reductionism:** a doctrine that maintains that all objects and events, their properties, and our experience and knowledge of them are made up of ultimate elements, indivisible parts.

Maybe that mathematical vision is the correct one in theory. Still it is somewhat hard to position all mental and human activities in the wormhole. For practical purposes, it is better to consider mathematics as an integral part of science.

Figure A.1 also provides a viewpoint on those specific events in which the general understanding of scientific knowledge in a certain field is radically transformed. A scientific revolution or a change of a scientific paradigm means that the development of scientific knowledge wanders briefly from scientific research and knowledge even to the area of personal beliefs. That sort of paradigm shift may deeply change the way we perceive the reality. General relativity proposed by Albert Einstein is perhaps the most famous example. Thomas Kuhn's prominent book (1963) about scientific revolutions illuminates the nature of the process. This book about communications ecosystems is hardly able to incite a general paradigm shift but still it may provide means for CEEs to perceive reality in a fresh way.

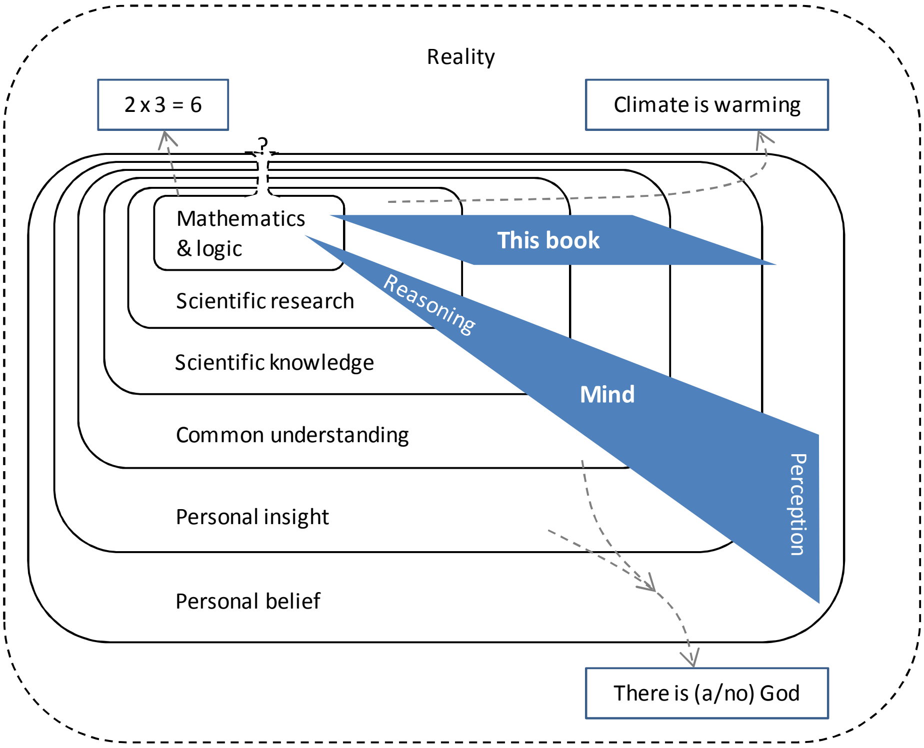


Figure A.1: Intermediate layers between mathematics & logic and reality. As to the structure of mind see also Figure H.1.

Terms

The main part of this chapter concentrates on various issues of statistical analysis. Thus, the central terms include probability and correlation. Despite—or maybe because of—their formal usage they are difficult to define in a useful way, that is, in a way that clarifies their role and relations to other concepts in the reality in which we are living. In this sense, information is even harder concept: it has a certain meaning in information theory, while its true role in reality is hard to comprehend. Nevertheless, it is better to have some definitions than no definitions at all. Here they are:

> **analysis:** the separation of a conceptual or material whole into its constituent parts and the study of the parts and their interrelationships,
>
> **correlation:** an interdependence of two or more variable quantities such that a change in the value of one is associated with a change in the value or the expectation of the others,
>
> **framework:** a basic conceptual structure used to solve or address complex issues,

hypothesis: a tentative explanation for an observation, phenomenon, or scientific problem that can be tested by further investigation,

information: a difference that makes a difference,

knowledge: theoretical or practical understanding about a subject,

method: a regular and systematic way of accomplishing something,

model: a system of assumptions, data, and conclusions presented as a mathematical description of an entity or state of affairs,

probability: a number between and inclusive of zero and one indicating the likelihood of an event,

science: a system of acquiring knowledge based on scientific method,

statistics: the science that collects, classifies, analyzes, and interprets data by means of theories of probability,

theory: a coherent group of general propositions used as principles of explanation for a class of phenomena,

truth: a statement that corresponds to fact or reality, and

variable: a measurable quantity that at every instant has a definite numerical value.

Furthermore, in Glossary you can find the following terms:

a priori	fact	paradigm	research
algorithm	factor	paradox	scenario
analogy	gross	parameter	scientific method
case study	Likert scale	pattern	simulation
causality	logic	prediction	survey
ceteris paribus	metasystem	proof	SWOT
criterion	net	psychology	symbol
cybernetics	null hypothesis	quantity	synthesis
discipline	observer	questionnaire	theorem
experiment	optimization	reductionism	trial

Understanding systems

In a major part of this chapter, we concentrate on the formal analysis based on experiments, probabilities, and statistical theories. However, there are many other valuable methods to gain knowledge about communications ecosystems. Figure A.2 presents some of the possible approaches from theories to real life. In the spirit of the sixth and seventh rules for CEE's I encourage any expert to recognize the value of all these methods and whenever feasible to develop his or her skills in all areas.

A strong theoretical foundation is one of the capabilities that distinguishes a genuine CEE from common consultants. Some modeling skills are necessary to be successful even when an expert uses models designed by others. As to modeling itself and as to the interpretation of the results of models, a CEE can apply many specific skills and methods.

Simulation is often used method to evaluate complex systems. I have made simulations regularly for 25 years and written thousands of lines of code for simulation programs. For instance, about half of my doctoral dissertation was directly based on simulations. Nonetheless, it is hard for me to give any wise advice about how to develop simulation tools, how to use the tools, or how to utilize the results of simulations. When thinking of those simulation activities afterwards, they were an integral part of a learning process rather than a separate activity that was able to create scientific results. Within a scientific community it is sometimes more respectable to prove a theorem—even when it is difficult to see any relevance between the theorem and real problems—than to state something more relevant and based on simulations. In the context of communications networks, formal theories are able to make explicit statements about the performance of a separate part of the system, such as a link or a queue. It might even be possible to develop accurate theories about network performance, but only under simplified assumptions. In contrast, on the level of ecosystems it is hardly ever possible to prove any formal theorems.

Therefore, I would recommend considering simulations as a way to *understand* the behavior of a system as part of a more general activity that includes many other methods. In order to understand the system, it is important to evaluate what kinds of systems are more efficient than other systems at accomplishing an objective—and simulations can facilitate that effort.

On a more general level, cybernetics can be seen as early attempt build a link between scientific studies and real systems as ingeniously discussed by Andrew Pickering (2010). In brief, the main difference between cybernetics and conventional science is that cybernetics approaches systems and ecosystems primarily from the viewpoint of real life rather than formal theories. That is something that a communications ecosystem expert shall also cultivate. Note also that the system archetypes discussed in Chapter M can be seen as a continuation of the cybernetic undertaking.

The upper levels in Figure A.2 depend much more on perception and intuition than pure reasoning. For instance,

> **SWOT:** an analysis in which internal strengths and weaknesses, and external opportunities and threats are closely examined

and case studies are typically based more on the personal insight of a group people than a formal evaluation of the system under examination. Scenarios are even more intuitive because no one can know exactly how the future will evolve. Scenarios can be useful in many situation, for instance, when an organization selects between possible strategies (see also Figure M.2). Finally, a story is perhaps as close to reality as what can be included in a book. Stories are, anyway, an important way to understand our lives.

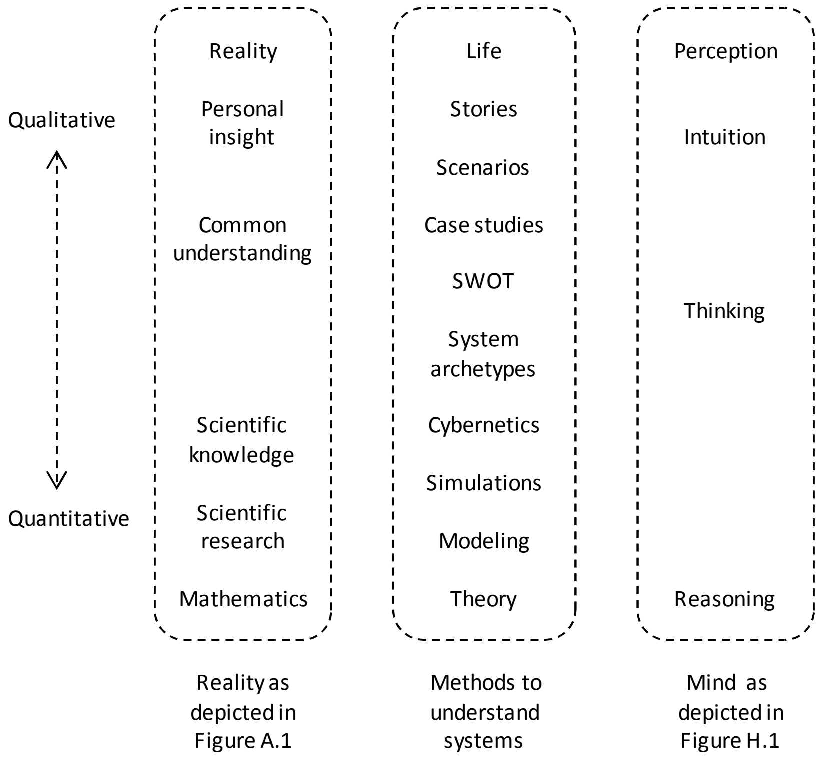


Figure A.2: Different methods to understand reality compared to the frameworks illustrated in Figures A.1 and H.1.

Formal reasoning

The stance of this book is that whatsoever question we consider we cannot avoid interpreting it through our beliefs and understanding of the world. One may even argue that for a human being there cannot be absolute certainty about anything outside the conscious mind: "Cogito, ergo sum," as the famous statement by René Descartes says. Another question is whether the concept of certainty is relevant in the unconscious part of mind. Unconscious "concepts" and "statements" are either less or more useful for the organism as a whole instead of being generally false or true.

Thus, in this framework, concepts like proofs and probabilities belong to the area of mathematics, measurements and models to the area of scientific research, paradigms and analogies to the area of scientific knowledge, and symbols and signs to the area of common understanding. Many terms are naturally used in many regions, which sometimes seriously complicates communication. Truth might refer to a strong personal belief or it might be a result of strict logical reasoning.

The most certain knowledge we possess, in addition to Descartes's statement "Cogito, seems to concern simple mathematical equations, like $2 \times 3 = 6$. You may think there is absolutely no doubt about its correctness. If we put together • • • and • • •, we surely get • • • • • •. Moreover, according to the established notation system, (• • •) and (• • •) means 2 times (• • •), • • • is equal to 3, and • • • • • • is equal to 6. The whole reasoning is so obvious that it requires an ingenious philosopher to cast any serious doubt over the self-evidence of the equation. Ludwig Wittgenstein did exactly that.

Even if it might be considered easy to explain the meaning of small integers, like 2, 3 and 6, it is much harder to explain the meaning of signs "×" and "=". How do we put together abstract concepts? What does it mean to state that something is equal to something else? What does multiplication mean? If you happen to be philosophically oriented, I encourage you to spend some time to reflect on these fundamental, philosophical issues. It is worthy to accept the truth that our understanding is limited even when we feel absolute certainty. Certainty about something can be thought of as an indication that according to evolution it is useful to believe something rather than to believe something else or not to believe anything at all. It is without a doubt useful to believe that $2 \times 3 = 6$.

If you are interested in these questions, you may start from Wittgenstein's Philosophical Investigations, which provides an amazing collection of thought-provoking argumentations and reflections. As an example related to the topic of equality, Wittgenstein writes (1954, item 216):

"A thing is identical with itself" – There is no finer example of a useless proposition, which yet is connected with a certain play of the imagination. It is as if in our imaginations we put a thing into its own shape and saw that it fit.

Then if you prefer more mathematically oriented text, you may read Hofstadter's Gödel, Escher, Bach (1999). As to the simple equation above, Hofstadter takes a huge effort to demonstrate what it requires to build a system that is able to prove number theoretic theorems based on a set of seemingly evident axioms without using any human intuition at all. In addition to the inspiring text, the book demonstrates that when a mathematical formula is applied, there is always an interpretative layer between mathematics and reality. Mathematics per se does not say anything about reality outside the realm of mathematics. This fundamental problem cannot be solved by developing a metalanguage or by constructing a metasystem to describe the relationship, because metalanguage and metasystem remain in the original area of mathematics, models, and communication.

I am surely unable to understand the limitations of mathematics perfectly, but it seems for me that self-reference and the introduction of metasystems are the major sources of controversies and paradoxes. If a system refers to itself, there always seems to be something that defies systematic analysis, because the system tries to serve as its own metasystem. You may think of a room that is almost empty at the beginning; it has a door but no windows. You want to document the contents of the room completely by means of various instruments including a camera. However, when you bring a camera in the room, the camera becomes a part of the room and, thus, must be documented, too. You may try to invent all kinds of arrangements to document all the additional instruments used to document the other instruments. This process seems to create an endless chain of tasks without any certainty that you would ever be able to finish the documentation task.

Now you may wonder, "What does this empty room have to do with ecosystems?"The point is that whenever we observe any important human ecosystem we are part of it and need to observe also ourselves. That process of self-observation cannot ever be perfect. As to communications ecosystem, communications ecosystem experts are a more or less significant part of the ecosystem, and must be included in a comprehensive analysis. Similarly, as to the empty room example, if you think that it is possible (in theory) to perfectly document technical devices, you have to notice that when you are inside the room, you had to document also yourself, including your aim to document the room and yourself.

Think, for instance, of a situation in which you give as a consultant justified advice for a service provider to replace usage based pricing by flat rate pricing. In order to make the advice justified you need to conduct an analysis that predicts how the ecosystem would evolve during the next couple of years. If you are not planning to cease your consultation work, you will give new advices later either for the service provider that you are consulting now or perhaps for some other services providers. Thus, the future of the ecosystem depends on your own forthcoming advice, and so on. That part of the analysis can be based either on your intuition, or perhaps, on some ingenious mathematical model that is able to handle the difficulties of self-reference. If the mathematical model is used in reality, it must also be able to solve the problems of its own self-references. As long as we are waiting for emergence of that superior mathematical model, we need to rely on our own insight and intuition (as illustrated in Figure A.1).

Further, Andrew Pickering provides in his renowned book "The Mangle of Practice" (1995, p. 126 - 138) an illustrative example about the intricate relationship between mathematical formulations and reality. The example is related to the meaning of multiplication. In 1843, Sir W. R. Hamilton formulated a novel theory of *quaternions*. His original aim was to extend the complex arithmetic ($x + iy$, where $i^2 = -1$) to a three-dimensional system ($x + iy + jz$) in a way that the system maintains the geometric association when complex numbers are added or multiplied. It turned out be an exceedingly challenging project. Finally, instead of a three-dimensional system he invented a four-dimensional system ($a + ib + jc + kd$) with special rules for multiplication, namely: $i^2 = j^2 = k^2 = ijk = -1, ij = k, ji = -k, ik = -j, ki = j, jk = i, kj = -i.$

The most important property of the quarternion system is that multiplication is non-commutative, that is, the result of multiplication depends on the order of the factors. This was a mental breakthrough in the area of mathematical thinking, although the quaternion system by itself has no major use, for instance, in theoretical physics. However, those of you who are familiar with electromagnetic theory likely notice the similarity between the quaternion system and the modern vector analysis used in Maxwell equations.

Even though the multiplication rules are counterintuitive, they made it possible for Hamilton to provide a consistent geometric interpretation of multiplication of vectors. In a way, geometrical manipulation can been considered more real than abstract arithmetic operations. Geometrical concepts are obviously more natural for us to think about than imaginary numbers, because we are living in a space that seems to obey simple geometrical laws. Few of us believe we are living in the realm of numbers.

Lessons for CEE

Lessons if any: mathematics can also be a construction that is either useful or not, rather than right or wrong. Moreover, there is a need to adjust mathematical constructions to the capabilities of the human mind. Still, all mathematic formulations must be internally consistent. I am not sure even about that requirement after reading Hofstadter's *Gödel, Escher, Bach*, because an absolute consistency is extremely difficult to preserve when new frontiers like infinity in mathematics are pursued.

Based on this brief philosophical discussion, it seems evident that there is no direct connection between the strict and definite region of mathematics and reality (as we are able to observe it). There always remain subjective interpretations and finally it is up to what each of us subjectively believes. Thus whenever you encounter a mathematical formulation, you are entitled (and often required) to ask what is the specific interpretation of the formulas when applied to reality. Mathematics may provide exceedingly strong proof that if a set of assumptions (B in Figure A.3) are correct then another mathematical statement is also true (C in Figure A.3). Still, a proof by itself made in the mathematical realm does not strictly prove anything in reality.

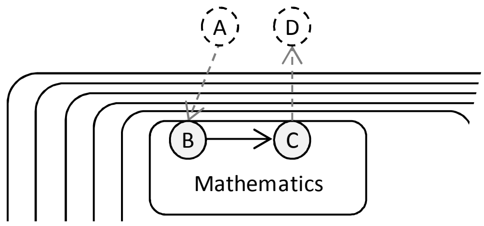


Figure A.3: The relationship between statements A and D interpreted through mathematical statements B and C.

"Climate is warming" is a statement that you may or may not agree with. My belief is that careful and thorough scientific studies have provided strong evidence in favor of the statement. If you disagree, the likely reason is that you doubt the credibility of scientific research in general. The history of science is full of errors, and it may even turn out that scientists have made some grave errors in the models or some important data has been ignored that has lead to misleading conclusions. Even in that case, it is unlikely that an individual antagonist has better knowledge about that matter than thousands of scientists conducting systematic research. The antagonist surely may believe he possesses superior knowledge but why should anyone else believe him rather than the scientific community? Still, someone may indeed believe the antagonist, probably because the person and the antagonist have similar common beliefs based on their cultural background, religious beliefs, or work community. Furthermore, conspiracy theories are popular in many fields of human activities.

At this point, you might wonder, what does climate warming have to do with communications ecosystem? In addition to the factual relevance of green technology in the area of communications ecosystem, my argument is that when you as a CEE try to convince your customers or other audience that your statements and inferences are correct, you will encounter all kinds of resistance. You may tend to think that the resistance is irrational, because your statements are based on credible research results, as far as you can assess. As a result, you may try to change the opinion of "the irrational part of the audience" by conducting even more research and scientific studies. That sort of approach rarely has a desirable effect.

In contrast, it might be a more useful for you to aim at understanding the motivation of the antagonists rather than to deem them as irrational persons. First, an antagonist might have personal reasons to hope that your statement is false. It is possible that the truth (as you see it) or the allegation or plot (as he may see it) evokes so unpleasant emotions that the easiest solution for him is to reject the statement. The more you try to convince the antagonist the more unpleasant emotions you evoke. In the worst scenario, those negative emotions are central to the antagonist while the positive emotions you may also be able to produce are peripheral for him (see discussion about emotions in Chapter H). For instance, a plausible scenario is that your statement, if true, endangers the current job or even the whole career of the antagonist. It may also be against what he and his colleagues have believed for twenty years.

What could you do? You need to operate more on the area of personal insight of the antagonist than on the areas of scientific knowledge or your own beliefs. It is unlikely that you would be able to change other person's core beliefs, but you may still give some ingredients for him to gradually develop and expand his insight. You may portray the situation as an opportunity rather than a threat: "Now it would be a perfect time to accept that the reality is changing and to utilize the capabilities of new technology before others. The adoption process may hurt at first, but because the change is inevitable, it is better to react immediately than to wait to see what will happen."

It may also be useful for the antagonist to believe something because that belief is highly appreciated within an important community for the person. If a person works at a small start-up company it definitely is important to believe the fundamental idea or story on which the start-up is building its future. In most cases, it is reasonable for an outsider to accept that kind of attitude without aggressively arguing against them, at least if you want to sell consultation service for the company as a CEE.

We may even go as far as considering the question of the existence of God (note, however, that it usually is advisable to avoid any unnecessary discussion about religious matters in the context of professional interactions). Although this topic might seem, once again, irrelevant, we shall be aware of the fact that a personal religious faith affects how a person interprets reality. If someone passionately believes in a god, he or she might interpret any event as a consequence of a plan of a higher, unfathomable power. In that case, it would be difficult to argue with the person by means of systematic reasoning, if the argumentation somehow touches his or her personal belief system. What could you do then?

If you want to work as a communications ecosystem expert, my recommendation is to start your own analysis with an assumption that there is no external source of intentions or plans (outside of the plans of the evident actors within the ecosystem). Remember also that real facts belong to the area of mathematics or scientific knowledge, which situation leaves room for different interpretations.

Even in cases where both the consultant and the customer believe on an external power, it is likely that they end up with different assumptions about the intentions of the power. How could anyone ever assume to understand or predict the true intentions of a higher power? It is hard enough to understand our own intentions. Besides, what a person typically expects about the intentions of a higher power is usually aligned with the interests of the person. Thus, the challenge once again returns to the challenge of understanding the motives of other persons.

Consequently, believing something is always a subjective affair that depends on a person's position in the society. There is always a thinner or thicker layer of subjective insight and belief between research and reality. Even when two persons have exactly the same information, they may make different interpretations due to their personal backgrounds. They may even use correct mathematical reasoning to justify two opposite statements about reality (D and ¬D), as illustrated in Figure A.4.

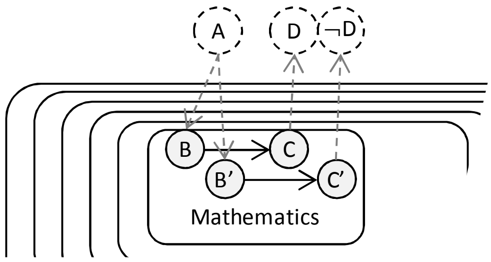


Figure A.4: A statement about reality (A) translated to two mathematical formulations (B and B') leading to two other formulations (C and C') that lead to two conflicting interpretation about reality (D and ¬D).

Probabilities

Probability is one of those concepts that are difficult to define in a persuasive way. The definition adopted here is:

> `**probability:**` a number between and inclusive of zero and one indicating the likelihood of an event within a mathematical framework.

But then one may ask: what is likelihood? Perhaps it is something that describes our expectations about the share of cases in which the event will occur among all similar situations. A probability of zero indicates an impossible event, a probability of one indicates a certain event, and a probability between zero and one means a contingent event.

A formal analysis is often based on

> **statistics:** the science that collects, classifies, analyzes, and interprets data by means of theories of probability.

Statistics answer questions such as: "If we have made certain observations, then what are the probabilities of certain events in the future?" Do we first need to make observations? Can we just rely on probability as a theory and say something about reality without any a priori knowledge? The standpoint of this book is that we cannot start any feasible analysis concerning reality from total ignorance. If one wants to say that he does not know anything about the possible values of a variable before observation, he cannot begin any analysis. Thus, we must first make some assumptions, say, that there are elementary outcomes with equal probability. An example of an elementary outcome is the number 4 in a game of dice: it is typically assumed that each number from 1 to 6 is equally probable. In reality, it is often difficult to define unambiguous elementary outcomes. Note also that the selection of elementary outcomes is a procedure not belonging to the area of mathematics; it lies somewhere between mathematics and reality in the framework described in Figure A.1. The selection of the elementary outcomes may also considerably affect the conclusions of the analysis.

Thus, a seemingly neutral selection of assumptions does not mean ignorance, but a well-defined starting point for an analysis. Let us take a small example to illustrate this issue. You decide to try something that you have never tried before, for instance, to drink 30 pints of beer during one evening. Now we may ask, what is the probability that you will succeed in the next try, if you have already failed the first n times? If we assume that each elementary probability (e.g., between 0.3 and 0.3+ε, where ε is an infinitesimally small number as shown in Figure A.5) between 0 and 1 to succeed is equally probable before any observation has been made, we can use probability theory. The result is that the probability that the next attempt will succeed is

$$\Pr(n+1{:}\,th\ attempt\ is\ successful) = 1/(n+2).$$

The fundamental assumption adopted in this analysis is that the reality can be described by a random process in which the probability of a successful attempt in every experiment is a constant *q*, where $0 \leq q \leq 1$. The question to be answered is: what is the expected value of *q* after a number of observations? Of course, reality is much more complex, because events rarely obey simple statistical rules: training may increase the probability of a successful attempt, whereas frustration may decrease the probability of success.

If we assume that each value of *q* is equally probable, it also means that the probability of q being zero is infinitesimally small, because there are an infinite number of possible elementary cases between 0 and 1. Consequently, the probability that you will never be successful is zero whatsoever you will try. You, indeed, may try to drink 30 pints of beer during one night. Now if you have tried that 8 nights in row and you rely on the model, you may infer that the probability to succeed on the 9th night is 10 percent and thus still worth trying (maybe this sort of use of probability theory demonstrates stupidity instead of ignorance).

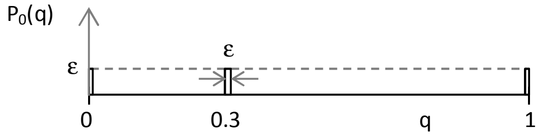


Figure A.5: Uniform distribution between 0 and 1 for parameter q. Parameter q describes the probability of a successful attempt in an experiment.

Let us construct a context-free setting for the question of success probabilities by considering a situation in which we can make as many identical experiments as we want and each experiment can yield one of two outcomes (A and B) which we can observe reliably. Now if the outcome has been A in the first n trials, what is the probability that the next experiment will result in the outcome B. What kind of analysis could we conduct, if someone says that there is no knowledge about the content and nature of the experiment itself but still requires that we must estimate the probability of the following outcome?

Nevertheless, we know something:

- there is a reality (outside mathematics) to be analyzed,
- there has been a series of experiments,
- the outcome of each experiment is either A or B, and
- there is someone who is interested in the question of probabilities.

Thus, we are not totally ignorant, because we cannot be totally ignorant because we are able to model ignorance. We can imagine various experiments that may result in two outcomes. The sun emits or does not emit a large amount of radiation on a given day, if a person puts two apples in an empty basket and then adds tree apples, he will or will not find 6 apples in the basket, a person will toss a coin and get heads or tails, tomorrow it will or will not rain in Mikkeli, Finland. There is obviously an enormous amount of possible experiments.

Let us assume that we were, regardless of the immense difficulties, able to make all possible experiments and each of them a huge number of times. As a result, we got results that indicate what will be the share of events with outcome B in each type of experiment. Let us designate the (a priori) probability that the share will be exactly q by $P_0(q)$. Is it reasonable to assume that this probability distribution is equally distributed between 0 and 1? Obviously, values below 0 and above 1 are impossible, and the distribution shall be symmetric as to 0.5, because names A and B are interchangeable. Other facts are harder to recognize.

The assumption of equality for any probability would mean that the likelihood that the probability of outcome A is exactly 0 is the same as the likelihood that the probability of outcome A is, say, exactly 0.19871672. This is hardly a reasonable assumption because anyone can imagine a huge amount of experiments that always, or almost always, result either in outcome A or B. A more reasonable assumption would be to assume an otherwise even distribution except that the probability of outcome A always happening is 1/3 and, correspondingly, that the probability of outcome B always happening is 1/3. This is illustrated in Figure A.6.

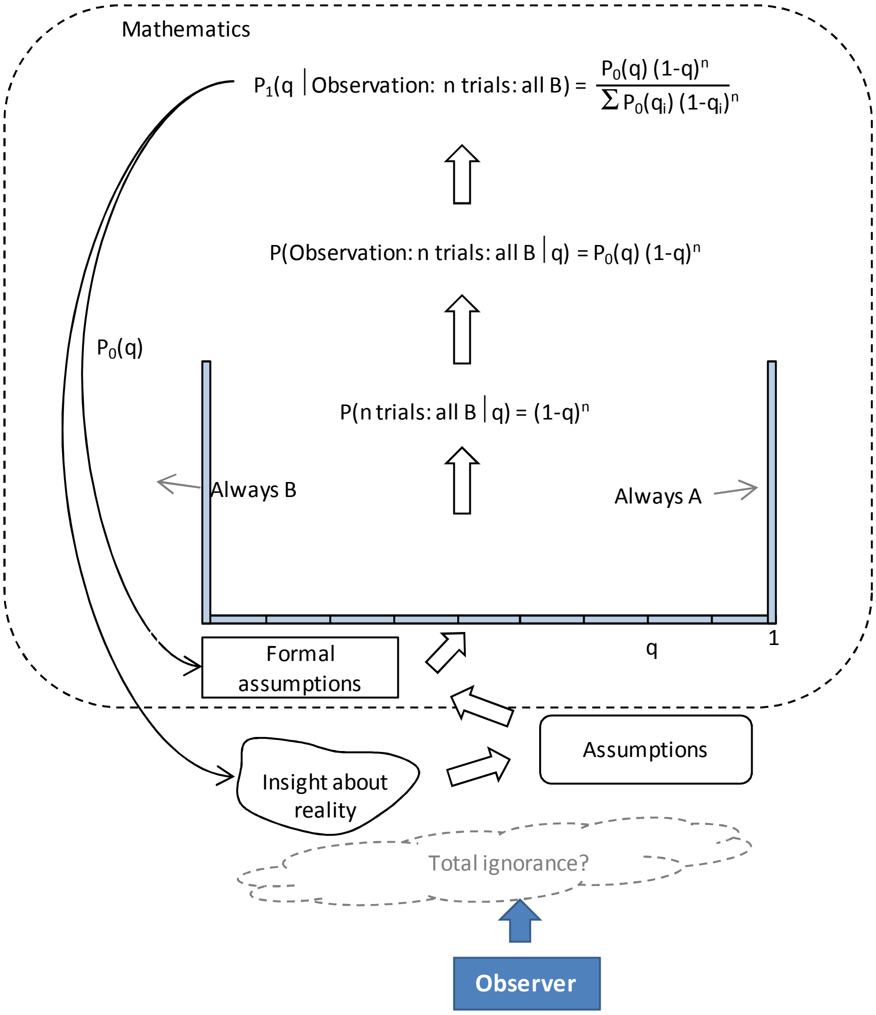


Figure A.6: A way of studying reality with experiments and formal models.

Does this additional assumption about the extreme probabilities (1/3 for q = 0 and q = 1) change anything? Yes, it does, but only if the previous experiments have always resulted in the same outcome, otherwise it does not change anything. Based on this (arbitrary) model the probability of outcome A after *n* outcomes of B can be estimated to be (when n ≥ 1):

$$P(A|n\ trials, all\ B) = \frac{1}{(n+2)^2}\,. \qquad (A.1)$$

According to this model if biologists had observed that all the known 30 species of swan were white, the likelihood that the next species of swan will not be white is about 0.1 percent. That is a small probability but anyway larger than zero. More about this in Black Swan by N. N. Taleb (2010). Note also that unless the possible outcomes are carefully selected, it might happen that neither A or B will occur, instead the observed outcome will be C. You could, of course, always define outcomes as A and not-A (or ¬A), but then the probability of not-A is typically larger than the probability of A.

The corresponding answer to the question "What is the probability that the next trial will result in outcome A?" is presented in Table A.1 and in Figure A.7. An experiment may include only one trial, or more typically, several similar trials. We may assume, for instance, that the observer makes first an experiment with one trial. The first model (Model 0 or Case 0,0 in Figure A.7) is a mathematical representation about the nature of reality. Note the following consequences of this framework:

- All probabilities stay in the realm of mathematics.
- Probability is a meaningful concept only in a well-defined model.
- There is always an interpretative or translation layer between reality and models, in both directions.

Table A.1: The probability of the next outcome being A (E(q)) after different combinations of earlier outcomes.

n	*n(A)*	*n(B)*	*Share(A)*	*E(q)*
0	0	0	-	0.500
1	0	1	0	0.111
8	0	8	0	0.010
9	1	8	0.111	0.182
19	1	18	0.053	0.095
20	2	18	0.100	0.136

Figure A.8 illustrates the process in the time dimension. Each new experiment or trial changes our understanding about reality. The process of understanding can be supported by formal probabilistic models.

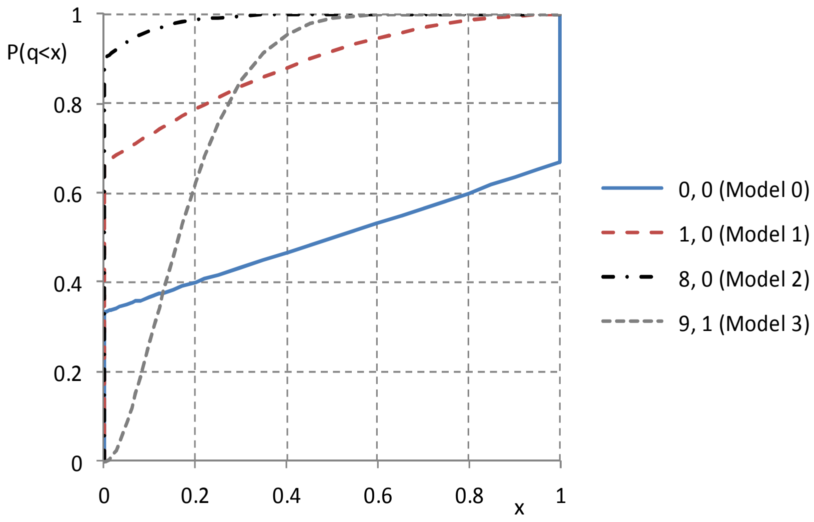


Figure A.7: The cumulative probability (q) that the next outcome will be A in situation (n, m) where n is the number of trials conducted before and m is the number of trials with outcome A in those trials. Case 0,0 represents the assumptions about reality before any trial is made.

For instance, E(P(A)) is a mathematical description that tells in the mathematical realm what the probability is that the next experiment (made in the mathematical realm) will produce outcome A. This result can be interpreted in a way that according to the model the "probability" that the next real experiment yields result A is E(P(A)). Nevertheless, if we speak about a probability we formally remain in the realm of mathematics.

A model that takes into account the practical limitations of real experiments is just a more complex model that includes mathematical formulations that try to explain the imperfections of reality. I would express the linkage between models and reality in a way that the results of a mathematical model can be *utilized* when making decisions. A model with certain interpretations is useful in a certain context if it turns out that after several experiments the advice based on the model is beneficial. A formally correct model can be harmful when interpreted in an erroneous way.

What can we say about reality itself? In this framework, "the share of outcome A among all outcomes after n experiments" is the closest point to reality we can reach in this context. Still the observed share of something is formally not a probability, because it is an observed share of something. A mathematical model (with probabilities) can yield prediction about those shares. Even though we often use the same term for the observed shares, and shares in mathematical models, in an exact analysis they have to be kept separate. The observations just indicate that certain models are more useful in making predictions about future events in reality. However, there always remain numerous distinct models able to explain any set of observations.

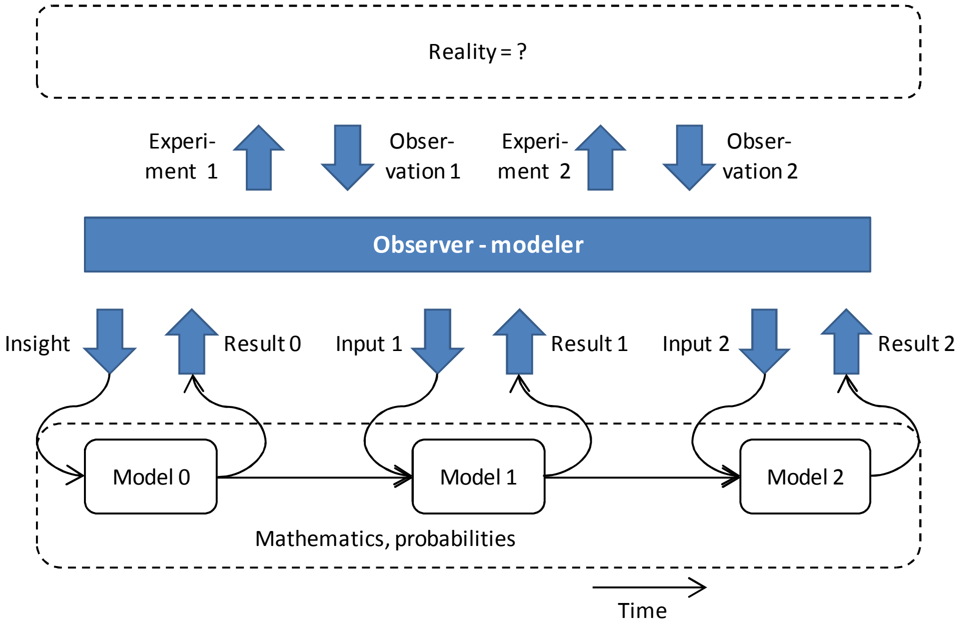


Figure A.8: Iterative process to develop insight about reality by means of experiments and mathematical models.

As to the probabilities, note that in the case where the ninth trial results in outcome A, the probability of the next outcome being A is 2/11 instead of 1/9 as could be expected. This difference can be "explained" by the fact that the probabilities are always limited to 0, which is much nearer to 1/n than to the other limit of 1. This difference demonstrates the distinction between share and probability. Surely, it is possible to construct a model in which the probability is the same as the observed share. However, according to that model if you have once observed outcome B (instead of A or something else), you should assume that all later experiments will produce the same outcome. That is hardly a useful assumption in reality.

We may also consider a situation in which an experiment with five trials has produced outcomes A, B, C, D, and E. Then we should rather believe that the probability of each of these outcomes is less than 1/5 because obviously it is also possible that the next outcome will be something other than A, B, C, D, or E (unless an outcome is defined as any outcome other than A, B, C, D, or E). What is the solution to this dilemma? The assumption about Model 0 for each possible outcome cannot be illustrated as in Figure A.6, because now we *know* that there are various (maybe equally probable outcomes) before we make any observations. We need to take that knowledge into account, which will considerable change the result of the analysis.

This also shows how difficult it is to make a formal set of assumptions that always lead to an intuitive outcome. Remember also that a model defined in advance shall be used to answer different questions. You shall not construct a specific model for answering each question if you want to preserve consistency between the answers. In general, one of the main benefits of modeling is the consistency of analysis, which cannot usually be expected if we apply pure intuition.

Do we need to care about the particular issue of rare events? I would say, yes. We encounter both in professional and civil life events that have always brought about the same result. We shall use our intuition and experience to assess whether the next result will once again be the same, or whether the result might be something else. Mathematical formulation may direct our thinking in this kind of case, both in favorable and unfavorable directions. If you really have no clue about the probabilities of the outcome, you may apply, with great caution, Formula A.1. Remember particularly, that the trials have to be independent of each other (which often is not the case) in order to make a simple statistical analysis justifiable.

Interpretation of numeric results

It would be valuable to have the ability to assess what realistic conclusions can be drawn from any numbers that are created by a random process. A Poisson distribution is applicable in various situations. Let us take as an example the number of people that die in traffic accidents in Finland: In 2008, the number was 344 people whereas in 2009 the number was only 279. This appears to be so large a change that there must be some good explanation behind it. Can we reliably deduce that some non-random factors have affected the decline, or can the difference be explained by the randomness of traffic accidents?

Let us first assume that all those fatal traffic accidents occurred randomly and only one person died in each accident. Under those assumptions, the number of persons per year should vary according to a Poisson distribution. Because of the nature of Poisson distributions, the standard deviation of the variation is the square root of the mean of the distribution. For 2008, we can thus assume variation of 18.5 and for 2009 variation of 16.7. As the difference between the years was 344 – 279 = 65, it seems likely that the change shall be mostly explained by something other than random variations. The expected variation in the difference between two years is defined by the combined number of accidents (344 + 279), which results in a standard deviation of 25.0. Standard statistical tests could be used to assess the statistical significance of the change.

However, the assumption of totally independent victims is unrealistic. In many accidents, several people die at the same time. Furthermore, some external factors, like bad weather conditions, may increase the overall probability of accidents on a given day. These kinds of phenomena may considerably increase the statistical variations of a type of event. In the case of traffic accidents, a rough estimation would be that the standard deviation of accidents per year is increased by 50 percent because of various interdependences. As a result, we may expect that the standard deviation of random variations between two years is 37.5 (instead of 25). This number is still much smaller than the observed difference (65), but only 1.73 times larger. Thus, the difference might be caused by random factors, although it seems likely that some external factors explain a considerable part of the change. In 2010, 272 people died in traffic accidents in Finland. This data supports the insight that some external reason caused part of the positive change from 2008 to 2009 while part of the positive change was the result of random variations.

Sixty five people represent a huge value. Although the value of life might be considered a profoundly spiritual matter, we may still estimate it on a monetary scale by using the value of time model. If we assume that the average value of time of a Finnish citizen is 50 Euros per hour and that each person killed in a traffic accident lost 40 years of life, the estimated total loss is about 760 million Euros. In addition, we shall take into account the decreased well-being of family members for a period of several years. The closest circle of significant persons typically consists of about five people. If we assume that the well-being of those close people is decreased by 30 percent for 20 years, the 65 saved people generate about 114 million Euros of value for other people. Those are huge amounts of value or lost well-being. Consequently, it is almost impossible to think of the saved lives and created well-being just as an accidental occurrence without any more meaningful reason than pure chance. There might be, indeed, some reason behind everything what happens in the world, but those reasons are out of the region of systematic analysis based on mathematical models.

Thus, the standpoint of this book is that when numbers are available, it is a useful practice to take them first as pure numbers created primarily by random processes. As a rule of thumb, you may use the following principles when you need to assess whether there is a change that needs an explanation other than random variation:

- For a number of observed events, N, assume random variability of $\sqrt{N}$.
- If you later observe M events in a similar situation, then if the change is smaller than $\sqrt{N+M}$, you do not need to look for any other explanation than random variation.
- If the change is larger than $2\sqrt{N+M}$ it is surely justified to expect and search for some other reason than random variation to explain the change.
- If the change is between $\sqrt{N+M}$ and $2\sqrt{N+M}$, then you may suspect that there is something to be explained although random variation might still be the main reason for the difference.

One possible explanation for larger variations than what random processes generate is that the observed events (during a limited period of time or within a limited region) do not occur independently of each other. As a simple example, in basketball 2 or 3 points are given at the same time, which means that we shall not assume that the total score of a game obeys a Poisson distribution. In contrast, in soccer goals occur more independently. So let us assume that two teams played two matches against each other, with results 1–0 and 3–5. Do we need any explanation for the change in the number of goals? As to the change from 1 to 3, we may safely assume that the change was caused by the intrinsic randomness of making goals. On the reverse, the change from 0 to 5 appears to be statistically significant, that is, a common factor is needed to explain at least partly the increased number of goals. Maybe the team spirit was improved.

As another example, let us assume that when a person drives to his office on a certain day he observes 10 cyclists, whereas on the next day he observes 15 cyclists. Do we need any specific explanation for the increase? No, because $\sqrt{10+15} = 5 = (15-10)$. On the contrary, if he had observed 26 cyclists, there probably was some external cause that affected the popularity of cycling, because $\sqrt{10+26} = 6$ while $(26-10) = 2.67 \cdot 6$. The reason for the increase might be better weather conditions on the second day. However, that analysis is valid only if all cyclists appear alone. On the contrary, if they typically bike in bigger groups, the analysis must take that fact into account. As a result, even an apparently large change from 10 to 26 might be caused by the random traffic variations without any essential difference between the days.

As a CEE, you will encounter a lot of numerical data. Often someone has his own motive to convince you that something important has happened, because some numbers indicate an increase (or decrease). Probably he has his own favorite explanation for the change as well. Be careful in these cases and conduct quickly your own assessment about the relevance of the change. In that task, the hardest part is typically to find out the size or unit of an independent observation.

Correlation

This section provides some remarks on statistical correlation. First of all, it should be stressed that correlation is not a sufficient condition for causality and that other variables need to be accounted for or controlled to say with some confidence that true causality exists. Sometimes the chain of events is clear enough to enable a statement that a true causality exists. For instance, if a large number of people drink a variable amount of beer, there surely will be a high correlation between the amount of beer consumption and the drunkenness of a person. We may say that drunkenness is caused by drinking.

A more interesting question is: what reasons trigger the drinking in the beginning? A study may find, for instance, a negative correlation between alcohol consumption and life satisfaction. That finding does not indicate causality in either direction unless the study is made in a controlled manner. Unfortunately, a controlled study is almost impossible because alcohol consumption in a real environment is very difficult to control without affecting significantly the life of a person. Any manipulation of person's alcohol consumption, therefore, likely decreases life satisfaction.

The following discussion concentrates mainly on the statistical significance of observed correlation. In any case, we shall remember that statistically significant correlation is only an indication that two variables have something in common in the observed system, as depicted in Figure A.9: factor Y affects both alcohol consumption and life satisfaction. At the same time there can be processes through which alcohol consumption affects life satisfaction (R in Figure A.9) and vice versa (S in Figure A.9).

The interventions may include an increase of alcohol tax, restrictions on the availability of alcohol, and education about the negative effects of alcohol on health. Although the interventions may even have the same immediate effect on the observable alcohol consumption, the interventions might have different effects on life satisfaction. Thus the method by which alcohol consumption is changed most probably affects the observable relationship between alcohol consumption and life satisfaction. Because there is no objective or neutral method to change alcohol consumption, there cannot be any unequivocal correlation between alcohol consumption and life satisfaction. The observed correlation always depends on the specific methods used in the studies. In other words, in this kind of case it is not possible to change one essential variable and leave all other things unchanged. In other words,

ceteris paribus: other things being equal

is a great principle in theory but cannot be realized in any complex ecosystem. Any action inevitably generates reactions that change the system in a complicated way.

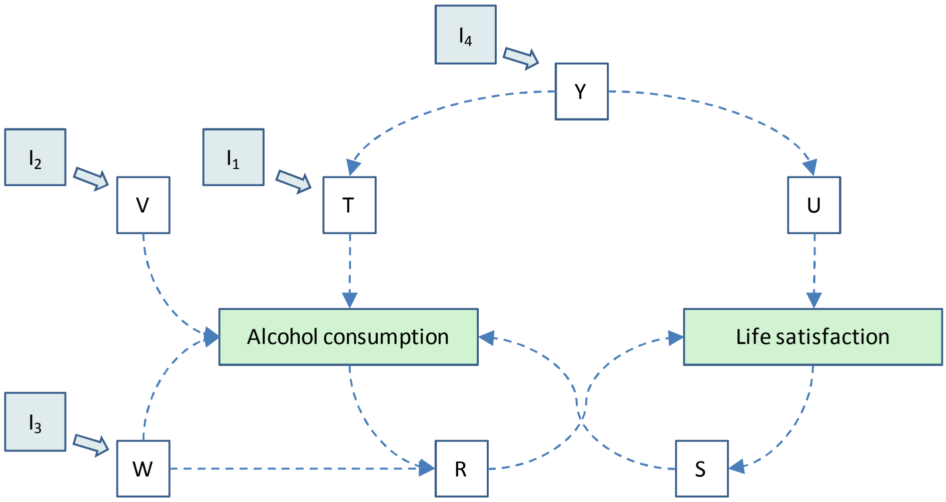


Figure A.9: Relationships between two observable phenomena (alcohol consumption and life satisfaction). R, S, T, U, V, W, and Y are factors affecting the observable phenomena. I_1, I_2, I_3,and I_4 are interventions used to influence alcohol consumption.

As a CEE, you can replace alcohol by communications services and make similar inferences. The usage of communications services can be changed in many ways. Each of those ways may result in a different change in the observed happiness of users.

Still in many cases, it is practical to make systematic studies in which the correlation between two variables is estimated. If we have sample of paired data (x_i, y_i) the *sample correlation coefficient*, commonly denoted *r* is defined as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where *n* is the number of paired variables, and $\bar{x}$ and $\bar{y}$ are the average values of *x* and *y*, respectively.

The main aim of this brief account about correlation is to give some additional insight about the use and interpretation of correlation coefficients. In order to gain deeper insight into statistical analysis, you shall consult basic textbooks about statistical methods or other relevant material

Figure A.10 illustrates the strength of the observable correlation for a small sample of 20 items. Actually, it is difficult to infer visually whether the correlation of a sample is larger than 0 if the real correlation is below 0.3. This is illustrated in Figure A.10: samples with correlation of 0.1, 0.2, and 0.3 do not really show signs of clear correlation. If one assesses visually the second row of samples in Figure A.10, the likely conclusion is that there is a positive correlation between the variables although the correlation is weak. Only if the correlation exceeds 0.7 we can say that the there seems to be strong correlation between the variables.

The main message of this discussion is that whenever you encounter analysis with correlation coefficients, you shall not overestimate the significance of the observed correlation. For instance, all the samples in Figure A.10 up to r = 0.7 were randomly generated (although with small tuning) without any “true” correlation between the variables. Above 0.7, a more complex but still random procedure was adopted.

Let us consider the case with r = 0.4 more closely. Because of the randomness of the sample, the observed correlation was purely due to the selection process in which the selection criterion was that correlation lies between 0.39 and 0.41. Now there seems to be two extraordinary points marked by *a* and *b*. In real cases it may appear that you could or should remove either (or both) of these are anomalies; it is often easy to find some justification for the removal. What will then happen to the correlation coefficient? If we remove point *a*, the correlation will decrease to 0.206. Correspondingly, if we remove point *b*, the correlation will increase to 0.519.

This example demonstrates the following important fact that you need to remember as a CEE: In the case of a small sample size, correlation values are very sensitive to any manipulation (conscious or unconscious).

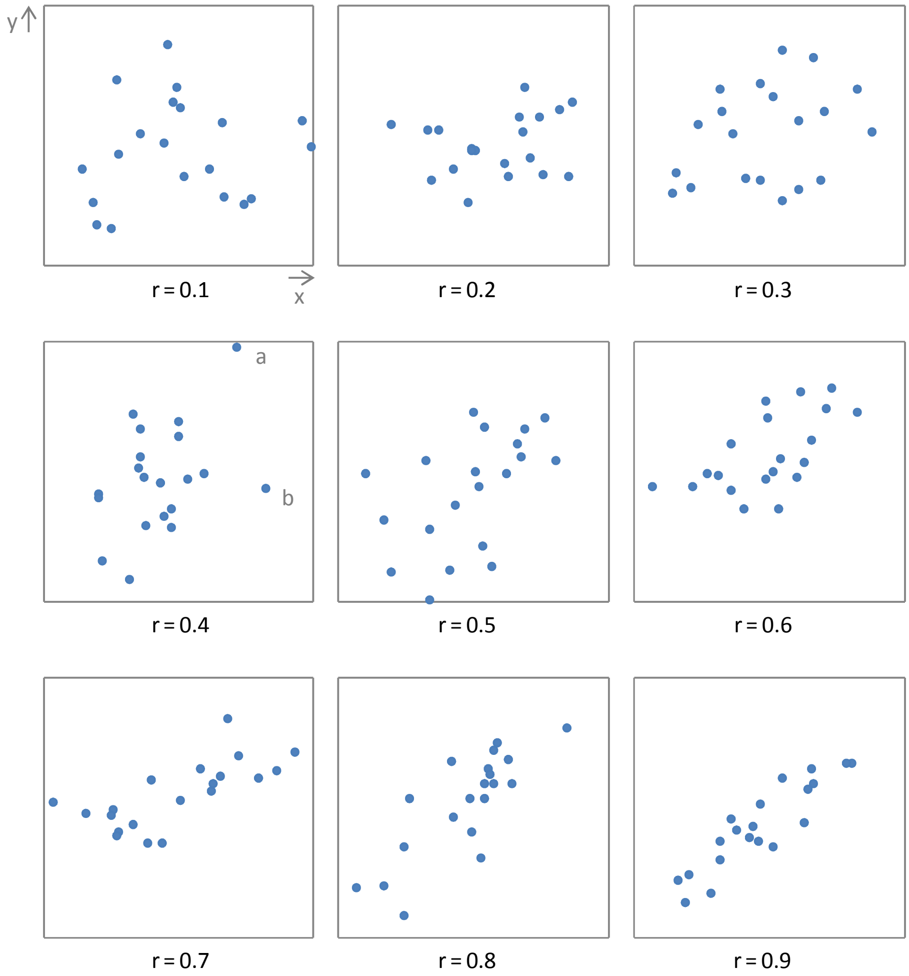


Figure A.10: Correlation coefficient (r) for samples with n = 20, normal distributed variables x and y.

Using statistical models

Finally, let us consider a situation in which a consultation company gives advice to a service provider about the business potential of an innovation based on a questionnaire study. Although the number of respondents that answered the questionnaire study was large enough to conduct a credible statistical analysis, it seems for the consultant responsible for the project that the respondents did not understand the real benefits of the innovation. Furthermore, she is aware of the fact that three companies in a similar market have already failed to convert the innovation into a profitable business. What should she do?

To start with, the last piece of information, she may assess that the odds are against any new provider: according to Formula A.1, the probability of success is only 4 percent. Unfortunately, she can hardly send a bill based purely on that reasoning. Still the starting point for a practical analysis seems to be that unless she is able to show that something essential has been changed compared to the three failure cases, there is not much hope for a successful business.

One troublesome aspect of consulting is that there might be pressure to achieve a positive result, that is, consultants may prefer to give advice to start something new, rather than to reject it totally, unless the analysis can provide a very clear, negative conclusion. You may think about a consultation project from both the viewpoint of consultant and the viewpoint of the person making the decision at the service provider. It is likely that both persons deem a decision to adopt the innovation more favorable for them and for their closest colleagues than a decision to abandon the innovation, particularly if they have expertise in that particular area. Both of them might already have gathered some knowledge and skills in the area of the innovation, and both of them might benefit in future years if the innovation would become a success story. You may notice that the service provider's interest is to gain objective insight as a company trying to make profitable business. However, we shall not confuse the motivation of a person (or group of persons) and the interest of the whole company.

All organizations have procedures that try to reduce the effect of personal motives in the decisions made by the employees. Those procedures work more or less efficiently, but companies will never be able to remove the phenomenon completely. The phenomenon also exists in the realm of scientific research. Positive research results are always preferred, because it is easier to get them published and because they can be used to make new research applications more credible. Now the actual questions in the consultation case are:

- Should the consultant take as a starting point of the analysis the fact that three companies have already failed (as a Model 1) or should she conduct her analysis purely based on the questionnaire study (starting from Model 0)?
- Should the consultant rely solely on what respondents have answered or is she allowed to add her personal insight about the "real" potential of the innovation?

In reality, the approach a consultant takes depends on his or her own motives. The consultant may even assess the questions systematically based on a well-defined metric. If her motive were to promote her career in the short term, she likely would dismiss the failed companies (as losers) and concentrate on the questionnaire results. If she wants to interpret the results in a favorable manner, there are numerous possibilities to achieve that without violating any explicit rule. She also needs to keep her own opinions separate from the questionnaire results, because of the risk of losing her credibility as a consultant. Thus, she likely applies widely used, standard methods and interpretations for analyzing the questionnaire results, even if she is aware of the problems of the underlying statistical models. Professional and appropriately complex treatment of data provides a lot of credibility as a consultant.

We may also consider the question about the general objectives more profoundly. If you think of consultation project as an activity, what could be a proper metric for it? Customer satisfaction, maybe. Then we may add something about the development of credibility as a provider of consultation services. The customer (or at least some persons within the organization) might become dissatisfied if the consultation company had immediately declared that they would not offer any extensive consultation service related to an innovation that certainly will become a failure. Still, if the future development of the innovation will prove that the consultant's claim about likely failure was justified, her credibility as an expert will be improved significantly. If you continue this reasoning, you may end up with a metric that measures the correctness and worthiness of your advice, independent of possible short-term pros and cons. This is what I would recommend in all kinds of cases including scientific research, consultation, and internal research and development projects.

About null hypothesis

From the viewpoint of the discussion illustrated in Figure A.8 the practice of accepting or rejecting a null hypothesis is somewhat peculiar. This peculiarity—that has occupied my mind during my entire career—is the main reason to discuss about it in this chapter. In the conceptual framework of this book illustrated in Figure A.1 the null hypothesis belongs to the area of mathematics. For many readers this issue might appear inconsequential. However, it is likely that a communications ecosystem expert comes across situations in which he or she needs to evaluate whether data from questionnaires or measurements support the proposition that certain variables affect each other. Will a change in the design of a device affect the willingness to pay? Standard statistical tests could be used to answer this question. Yet, that sort of analysis does not necessarily give all the insight that questionnaires and measurement could provide.

If a researcher has collected data that reveals some correlation between two variables, she may use a standard procedure to determine whether she shall reject the null hypothesis (which means that there is no correlation between the variables). For instance, conventional statistical analysis may state that:

If the number of data points is 20, we can infer with 95 percent certainty that there is some correlation ($r \neq 0$) between the variables if the observed correlation is above 0.444 or below −0.444. Then if we require 99 percent certainty, the observed correlation shall be larger than 0.561 or smaller than -0.561 to justify the rejection of the null hypothesis of zero correlation.

In a way, this might be an appropriate way to interpret the meaning of observations, but it is also somewhat peculiar. Particularly, the reasoning does not take into account the possibility that zero correlation is more common than any other correlation in reality. Actually, a correct statement is that under certain mathematical assumptions if two variables are independent of each other, the probability that the observed correlation with n = 20 is above 0.444 is 2.5 percent. We cannot say anything for sure about the probability that the "true" correlation is 0 on the condition that the observed correlation is 0.444 unless we make some assumptions about the other possible "true" correlations (that is, unless we define Model 0 in Figure A.8).

These days, we have a lot of computing power and we are not limited to practices that rely on pre-calculated tables. Thus, I would rather prefer an approach similar to the approach used to analyze the situation in which we had a small number of observations entirely or mostly with the same outcome (illustrated in Figures A.5, A.6 and A.7). Initially, we can be certain that the correlation coefficient lies somewhere between -1 and 1. This is the basis for defining the Model 0 illustrated in Figure A.8. In addition, we can be sure that in many real cases there is no correlation between two variables. As a result, there are three special values for the correlation coefficient: -1, 0, and 1. Values -1 and 1 are rare in reality. Besides, they are not relevant in typical cases, because even if we have only a few data points, we can usually be sure that the true correlation is neither 1 nor -1.

Thus Model 0 in Figure A.6 (describing the situation before any experiment) can be defined as follows:

$$P(r = 0) = 0.5,$$

$$P(r < x) = 0.25 + 0.25x \ for \ -1 \leq x < 0, and$$

$$P(r < x) = 0.75 + 0.25x \ for \quad 0 < x \leq 1.$$

Now any experiment that is made changes this distribution according to the rules of probability theory. A possible result of the analysis is shown in Figure A.11 in a case where the observed correlation is either 0.20 or 0.30. In this way of thinking, the objective of the statistical analysis is not to reject or accept a null-hypothesis, or any other hypothesis, but to present how the experiment changes the insight about the relationship between the variables by means of formal reasoning. Unfortunately, this kind of reasoning makes it more difficult to claim that the results of an experimental study support the assumption of significant correlation.

Some results of the analysis may appear surprising: if we observe a correlation of 0.2, it seems that the estimated probability of the "true" correlation being zero is increased because of the experiment. However, there is an explanation for this. Let us imagine 20 small boxes each of which is filled with 100 balls. The first box represents correlations from -1 to -0.9, the second box represents correlations from -0.9 to -0.8, and finally the twentieth box represents correlations from 0.9 to 1. In addition there is one big (Zero) box filled with 2000 balls representing zero correlation. Thus, these boxes illustrate roughly the Model 0 shown in Figure A.8.

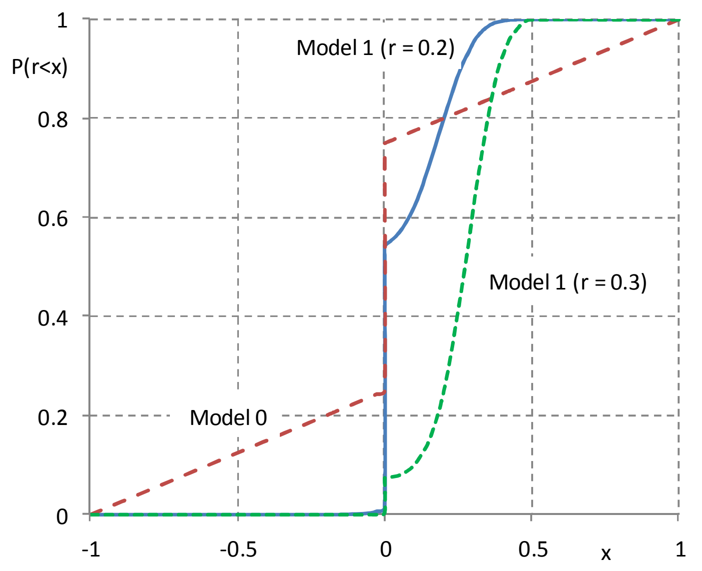


Figure A.11: Cumulative probability distribution for correlation coefficient (r) before (Model 0) and after (Model 1) an experiment with an observed correlation of 0.2 and 0.3 (N = 100).

Each ball denotes a different correlation coefficient to be observed in an experiment with 100 data points. Let us assume that each ball that denotes a correlation being between 0.195 and 0.205 is painted red. The number of red balls corresponds to the probability that the observed correlation would be between those limits. As a result, there will be one red ball in the eleventh box, four red balls in the twelfth and thirteenth boxes, one red ball in fourteenth box, and eleven red balls in Zero box. Now an experiment means that you pick randomly a ball among all the 4 000 balls. If you get a red ball, we can ask: what is the probability that the ball was taken from the Zero box? The answer is: the probability is 11/21 = 52 percent. Why is it so large a probability even though the share of red balls in Zero box is below 1 percent? The reasoning is that the selection of red balls leaves out sixteen out of twenty boxes. Although the share of red balls is smaller in the Zero box than in the boxes from eleven to fourteen, that difference is compensated by the larger size of Zero box and by the omission of the other sixteen boxes.

As a consequence of this reasoning, if the number of data points is 100, the observed correlation has to be above 0.35 (or below -0.35) in order to justify a statement that there certainly is positive (or negative) correlation between the variables. Note also that the reasoning depends essentially on the assumptions we made before the experiment (Model 0). The assumption that the probability of zero (or very small) correlation is significant affects the interpretation of the correlation results. Yet it is not reasonable to assume that a CEE would make a detailed analysis in every case she needs to deal with correlations. Notwithstanding it would be useful to have a rough insight about the strength of the results. You may use the following rules of thumb:

If the number of data points is *N* and the observed correlation is *r* then:

- If $-2/\sqrt{N} < r < 2/\sqrt{N}$, primarily assume that there is no real correlation.
- Only if $r < -3.5/\sqrt{N}$ or $r > 3.5/\sqrt{N}$ you can be (relatively) certain that there is real correlation between the variables.
- This also means that if the sample size is smaller than twenty, it is almost impossible to state anything certain about the correlation between two variables.

Table A.2 provides some examples based on this rule. Remember also that the "Model 0" assumption was that zero correlation between two randomly selected variables is more likely than other values of correlation.

Table A.2: Significance of correlation between two variables: "+" strong support on positive correlation, "?": weak support on positive correlation, "-": supports assumption about zero correlation.

	Number of data points							
r	*20*	*30*	*50*	*100*	*200*	*300*	*500*	*1000*
0.05	-	-	-	-	-	-	-	-
0.10	-	-	-	-	-	-	-	?
0.15	-	-	-	-	-	?	?	+
0.20	-	-	-	?	?	?	+	+
0.25	-	-	?	?	?	+	+	+
0.30	-	-	?	?	+	+	+	+
0.35	-	?	?	+	+	+	+	+
0.40	?	?	?	+	+	+	+	+
0.50	?	?	+	+	+	+	+	+
0.60	?	+	+	+	+	+	+	+
0.70	+	+	+	+	+	+	+	+

Note also that these rules result in somewhat different conclusions than the standard statistical methods with significance levels and null-hypothesis (see, for instance, the Wikipedia article about correlation). If you want to write a journal paper, it is recommended to use standard methods.

Wisdom of crowds

In many cases, the only acceptable way to conduct a scientific study is to make carefully designed, objective measurements, and then to analyze the results according to established principles. However, there are various phenomena that are still out of the reach of objective measurements, most notably, emotions. One important question also from the viewpoint of this book is whether people are able to assess and to give reliable answers about their emotions. Let us consider this issue from the viewpoint of

> **wisdom of crowds:** a situation in which the average opinion of a large group of people is better than the opinion of an individual or a small group of experts.

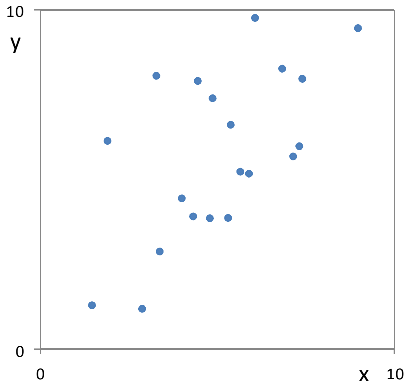
The main point of this brief discussion is to show that if something mathematical can be assessed by this method of wisdom of crowds, then it gives some credibility for assessing softer matters, particularly emotions and experiences.

Figure A.12 shows an example about the wisdom of crowds. In an experiment made at Aalto University 37 students were asked to estimate the correlation between two variables in six different cases. The result of the experiment in the case with a true correlation of 0.6 is shown in Figure A.12. The students were not taught anything about correlation coefficients before the experiment except that the correlation may vary between -1 and 1. They were also informed that the correlation in each figure was either -1, -0.9, -0.8, … 0.8, 0.9, or 1. Although a majority of the students had learned something about correlations during their previous courses, there were considerable variations in the estimations between students. The average estimation was 0.58 and the median estimation was 0.60, while the correct correlation also was 0.60. In five other similar cases, the difference between the median estimation of the students and the true correlation was either 0 or 0.1. When all six estimations about different correlations were taken into account, the median estimation of all students provided a better result than the best student did individually.

There are many examples demonstrating the feasibility of the wisdom of crowds. Gary Hamel (2007, p. 229 - 237) explains a case in which the ordinary employees at Best Buy tried to predict the gift-card sales for the following month. Normally, the estimation used by the company was made within a small team of experts that knew the special properties of the gift-card business. In the first experiment, the expert team's estimation deviated by 5 percent from the true sales while the crowd's estimation deviated by less than 0.5 percent. The results of other experiments made at other departments were at least as surprising. It seems, indeed, that the wider range of data and insight among a large group of people almost always compensates for the lower expertise and knowledge of individual persons compared to small group of experts. In a way, this is the logic of making survey studies: although the answers of an individual might be quite unreliable, a large enough group of people can provide reliable and credible results.

A modest incentive for the participants of this kind of experiment could be useful both in private companies and in an academic environment. In the case of Best Buy, the monetary incentive in the first gift-card experiment was only $50. In the case of the correlation experiment shown in Figure A.12 the only incentive was that the students with the best estimations got an extra point, while the total number of points available during the course including the exam was about 100.

The main challenge of this method is the possibility of systematic bias. For instance, in the case of emotions people may systematically either underestimate or overestimate the strength of their emotions because of social reasons. Moreover, the participants of a study may interpret the scale used in the questionnaire systematically in different way than what the researcher expects. Thus, even though the quantity of answers may partly compensate the uncertainty of individual answers every researcher shall take into account the possibility of systematic errors.



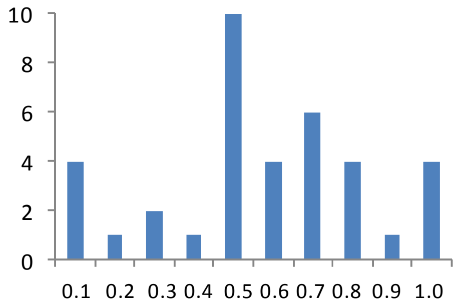


Figure A.12: Estimation of correlation as an example of wisdom of crowds: a) true correlation between x and y is 0.60, b) the distribution of estimated correlation by 37 students.

Lessons for CEE

Any model that aims to describe reality must start from assumptions that are based on the modeler's insight about reality. Those assumptions do not belong to the realm of mathematics, but they have to be converted to well-defined formal assumptions that can used to construct a tractable mathematical model (or simulation).

Because the results of the analysis are always sensitive to the underlying assumptions, a skillful person or a potent organization will always be able to produce "biased" results that are seemingly credible because they are based on a strong mathematical basis and on valid data about reality.

Even cases where something has not yet been observed can be analyzed by means of statistical model. However, it is important to remember that the results of the analysis are heavily dependent of the underlying assumptions.

Still, the results achieved with the model do not necessarily describe reality in a way that could be called "correct," because we can only have limited information about reality. Correctness is a valid term inside the realm of mathematics whereas usefulness is usually a better term to describe the relationship between reality and mathematical models (see Figure A.1).

Authors in communications ecosystems

Now let us apply the correlation method to analyze the field of communications ecosystem. First, we need enough data to calculate correlations, preferably at least 50 variables. Fortunately, search engines provide a great opportunity to gather relevant information without any expensive tools. Note that the main cost factor to the analysis described here is by far the time consumed to gather the data, unless the process is fully automated. However, an analyzer needs to be a bit cautious with too quick an automatization because then he loses the possibility to develop the model and idea during the collection of data. Besides, when you do something manually, you get a deeper and more solid insight into the data.

Thus, I started with an idea that I can use the authors cited in this book as the basis for an analysis of the structure of the field. We can easily check how many hits a name, like "Gregory Bateson," is able to generate. I used Google in this particular analysis in August 2011 and obtained a result of 520 000 (similar results can be obtained by other search engines as well). Remember also that the number of hits shall not be interpreted as the number of pages in which the searched term appears, because different pages have different weights in the formula that gives the approximate number of hits (if the number of hits is larger than ten). Still the number of hits provides a helpful indication of the *popularity* of the term or name in the web. Because Google's business is based on advertising, Google is more interested in the popularity of objects than the number occurrences of keywords in the web.

The hits for separate names do not provide, however, any basis for a correlation analysis. There are various options available to define a set of factors for the names. The most straightforward way is to use a pair of names, for instance, "Daniel Kahneman" and "Peter Senge" and to test the number of hits they generate together (in this analysis the result was 6590). I did this analysis for the 50 authors listed in Table A.3. Consequently, there were in total 50×49/2 = 1225 data points.

The research process was not as simple as could be imagined. I did not have any clear idea about the method to be applied in advance, and I was not sure about the usefulness of the results. What I had in my mind was to create a 2-dimensional map in which the authors of each field (economics, usability, technology, etc.) form a distinct group. I did a similar approach with the Systems Intelligence area in 2008. In that study, the map was based on the data about those authors that generated the highest number of hits together with the given author. Then I drew manually a map in which each author was located close to three other authors, if possible. Even though the result was useful, I was dissatisfied with the process of making the map, because it relied too much on my own intuition rather than the collected data. Intuition is often useful, but it is also bias-prone.

Table A.3: Authors for the field of Communications Ecosystems

Abbr.	*Author*	*Example of specialty*
ALB	Albert-László Barabási	Scale-free networks
AMa	Abraham Maslow	Hierarchy of needs
AOd	Andrew Odlyzko	Content is not the king
APi	Andrew Pickering	Mangle of practice
BFr	Barbara Fredrickson	Positivity
BGi	Bob Gilbreath	Marketing with meaning
BLi	Benjamin Libet	Unconscious brain activity before conscious decision
CAn	Chris Anderson	Long tail
CDw	Carol Dweck	Growth mindset
CFo	Claes Fornell	Customer satisfaction
DAN	Donald A Norman	User centered design
DDC	David D Clark	Design philosophy of Internet protocols
DHo	Douglas Hofstadter	Gödel-Escher-Bach
DKa	Daniel Kahneman	Prospect theory
DMe	Donella Meadows	Thinking in systems
ERo	Everett Rogers	Diffusion of innovations
ESa	Esa Saarinen	Systems intelligence
FVa	Francisco Varela	Autopoiesis
GBa	Gregory Bateson	Criteria of mind
GHa	Gary Hamel	The future of management
GHu	Geoff Huston	Internet architecture and management
GMK	Gordon MacKenzie	Orbiting the giant hairball
GWe	Gerald Weinberg	General systems thinking
HGa	Howard Gardner	Multiple intelligences
HMa	Humberto Maturana	Constructivist epistemology
HRV	Hal R Varian	Information economics
HSc	Henning Schulzrinne	Internet protocols
JNa	John Nash	Nash equilibrium
JSt	John Sterman	All models are wrong
LLe	Lawrence Lessig	Free culture
LWi	Ludwig Wittgenstein	*Wovon man nicht sprechen kann, darüber muß man schweigen*
MGl	Malcolm Gladwell	Tipping point
MHa	Marc Hassenzahl	Experience design
MSe	Martin Seligman	Flourish
NLu	Niklas Luhmann	Social systems
PBl	Pete Blackshaw	Angry customers tell 3000
PSe	Peter Senge	Fifth discipline

Table continues on the next page

Table A.3 continues.

Abbr.	*Author*	*Example of specialty*
RBa	Roy Baumeister	Bad is stronger than good
RDu	Robin Dunbar	Dunbar number
RLa	Richard Layard	Happiness as a new science
SBe	Stafford Beer	Management cybernetics
SBo	Samuel Bowles	Co-evolution of preferences, institutions, and behavior
SHa	Sam Harris	The moral landscape
STa	Steve Talbott	Technology, nature, and the human prospect
TBL	Tim Berners-Lee	World Wide Web
TKu	Thomas Kuhn	Scientific revolutions
VJa	Van Jacobson	TCP/IP
VKa	Victor Kaptelinin	Acting with technology
VLS	Vernon L. Smith	Experimental economics
WBA	W. Brian Arthur	The nature of technology

Thus, I tried to develop another approach bearing in mind that an efficient process to refine a large collection of data shall combine three phases. First, human insight is needed to define what type of data shall be collected. Secondly, the collected data shall be automatically processed to a format that can be easily perceived by a human being. Finally, human insight is necessary to interpret the results. Furthermore, human insight is needed to direct the construction of the process itself. In contrast, the results of the analysis are more credible when the human interaction is minimized during the data processing, because otherwise it would be almost impossible for an external observer to assess whether the person that has conducted the analysis, has tuned (either intentionally or subconsciously) the result to a specific direction.

In this example, I had to decide what authors should be used in the analysis. I started with the primary authors I use in this book. However, after obtaining the first results, I decided to add about 10 authors, mainly to clarify the position of Internet technology in the map and to strengthen some other fields. It should be stressed that with a different set of authors the map shown in Figure A.13 might appear different. For instance, an addition of a group of mathematicians or computer scientists might change the structure of the map. However, it seems unfeasible to add more fields to the map without removing some of the present fields, because a two-dimensional structure has its limitations.

As the results, I developed a practice in which the three phases can be clearly separated and the data processing phase could be automated as efficiently as possible. I ended up with the following procedure for creating an ecosystem map based on key authors in the field:

1. Select about 50 names to represent the various parts of the ecosystem. Most of the authors shall generate at least 50 000 hits individually to make the analysis reliable.
2. Determine the number of hits for each pair of authors by using any available search engine. For instance a search for “Daniel Kahneman” “John Nash” may produce 14700 hits.

3. Calculate the hit profile of each author. Because there are enormous variations in the number of hits, it is better to present the hit profile on a logarithmic scale. Thus the hit profile for author i is a vector in which the j:th term is $\log_{10}(\mathrm{hits}_{i,j} + 1)$. Thus for the case in which i is "Daniel Kahneman", and j is "John Nash", the result is $\log_{10}(14701) = 4.17$.
4. Calculate the correlation (r_{ij}) between the hit profiles of every author pair (i, j). For instance, Figure A.14 shows the correlation between the hit profiles of Daniel Kahneman and Martin Seligman.
5. Draw a map in which those authors that have a strong positive correlation are located near each other and those with a negative correlation are located far away from each other. For instance, you may minimize

$$\sum_{i,j=1}^{N}\left((1 - r_{ij}) - \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}\right)^2$$

   where x_i and y_i are the coordinates for author *i*, x_j and y_j are the coordinates for author *j*, and *N* is the number of authors.
6. Interpret the constructed map and give names for distinct group of authors.

The highest positive correlation (0.959) found in the study occurs between Humberto Maturana and Francisco Varela—which is not a surprise taking into account the intensive research cooperation between them. Some other high correlations were less expected. For instance, the correlation between Ludwig Wittgenstein and Howard Gardner is as high as 0.880, whereas there is no correlation between Samuel Bowles and David D. Clark as shown in Figure A.15. The strongest negative correlation is -0.227 between Henning Schulzrinne and Samuel Bowles. Once again, that is not a surprise if we consider their specific areas, human well-being and Internet protocols.

We can make interesting observations about the number of hits. There were many pairs of authors that Google could not find together on a single page in the whole web, including:

- Geoff Huston and Humberto Maturana,
- Henning Schulzrinne and Donella Meadows,
- Andrew Odlyzko and Bob Gilbreath,
- Victor Kaptelinin and Barbara Fredrickson, and
- Marc Hassenzahl and Benjamin Libet.

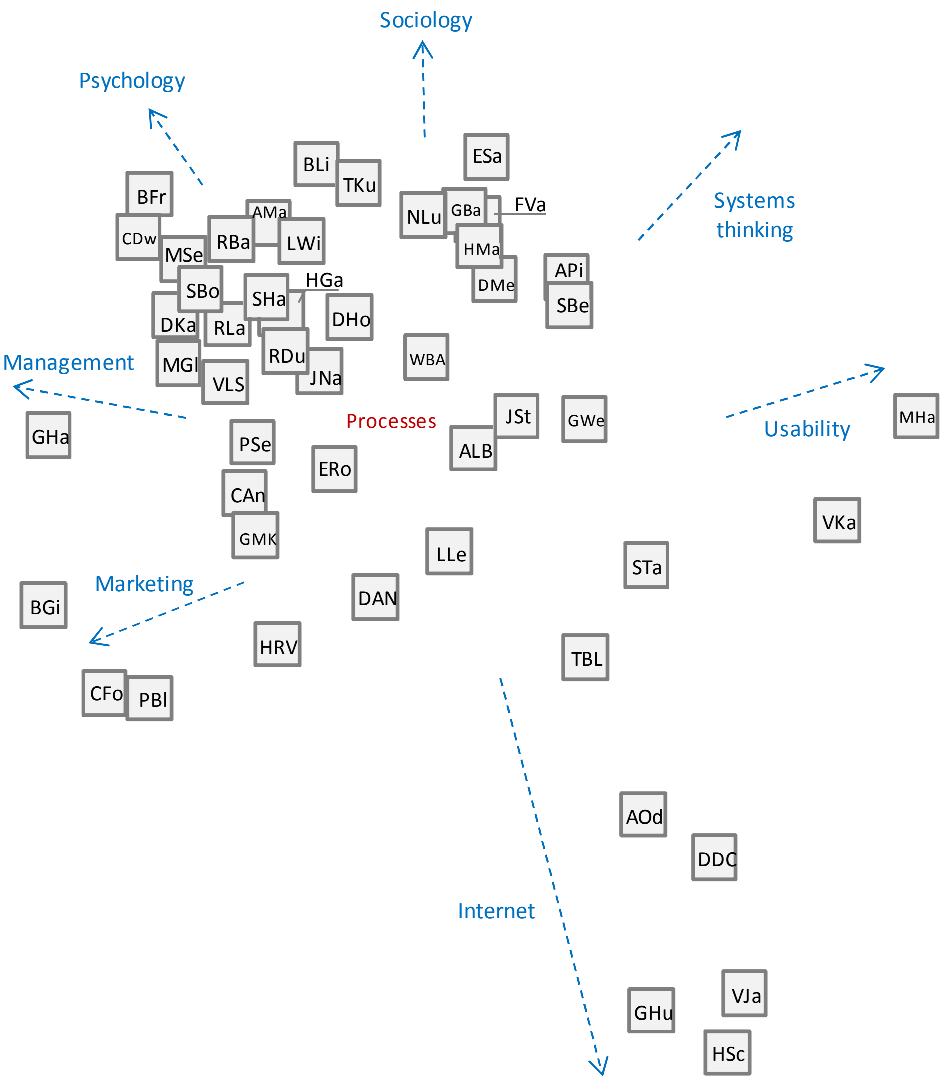


Figure A.13: Map of Communications Ecosystem based on the concurrent occurrence of authors in the Web.

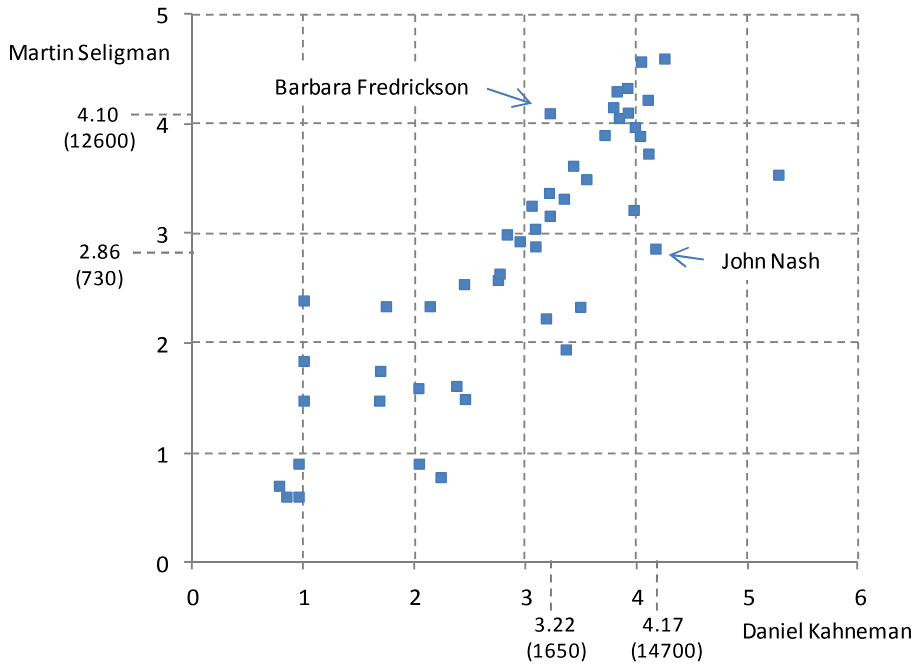


Figure A.14: Two authors close to each other on the map of communications ecosystem (Figure A.13). Each data point represents the number of hits with another author on a logarithmic scale: horizontal axis with Daniel Kahneman and vertical axis with Peter Senge. The correlation between Kahneman and Martin Seligman is 0.84.

What do these non-existences tell about the separation between fields? A person developing Internet technology might be highly successful without knowing anything about Humberto Maturana or Donella Meadows. Still, since the Internet is a system par excellence, system theory might provide an alternative viewpoint on many important issues emerging during its technical development. Can it really be so that no one has ever tried to apply systems thinking promoted by Donella Meadows when assessing the development of Internet protocols?

An even more surprising deficiency is the weak bond between usability studies and psychology, because someone may consider usability as a sub-field of psychology. Moreover, there does not seem to be much in common between marketing and usability. Nonetheless, a stronger relationship between these fields may exist in reality but the relationship is not observable by means of the authors used in this example.

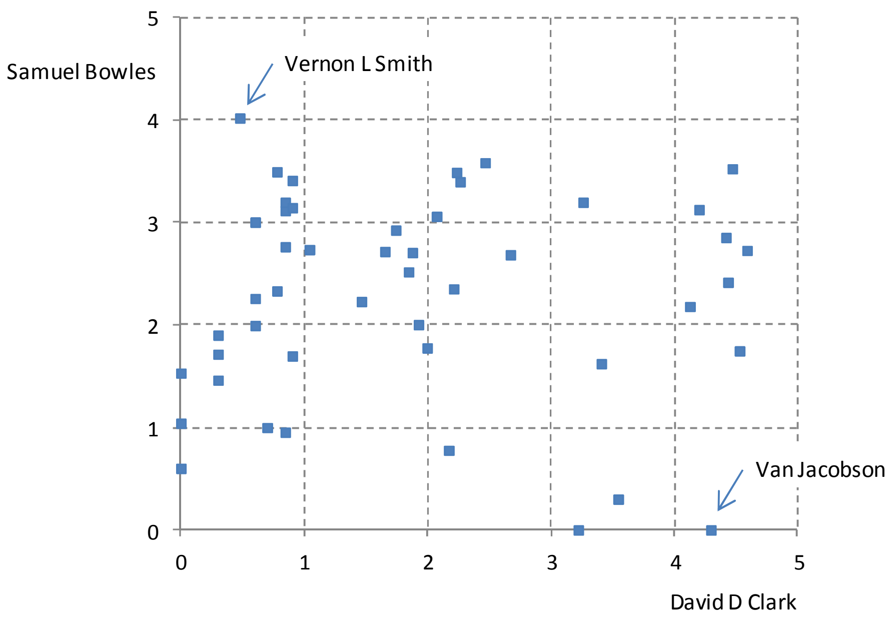


Figure A.15: Authors far away from each other on the map of communications ecosystem (Figure A.13), horizontal axis: David D Clark, vertical axis: Samuel Bowles, correlation = -0.005.

A comparison between Figure A.13 and Figure I.1 yields many interesting observations. Figure I.1 was designed already at the beginning of the writing process of this book. It is wholly based on my own intuition on how different fields discussed in this book are related. I assumed that the studies of human behavior and psychology form a discrete field that can be located quite far from economics (that I put on the other side of the map). However, according to the Figure A.13 there is no separate field of economics, as far as the selected authors are concerned. Maybe microeconomics shall considered nothing more (or less) than studies of human behavior in an economic context. Thus, it seems that psychology is a decent name for the upper-left corner of Figure A.13. Barbara Fredrickson excellently represents the outer corner of that field.

I also assumed before this correlation study that economics would be more closely related to system science (that, indeed, may be an appropriate assumption about macroeconomics, but that relationship is not investigated here). A group of systems scientists emerged unambiguously from the correlation study, but without a direct connection to economists. A less clear group (that might be called sociology) emerges between psychology (including microeconomics) and systems theory, although there are also two authors that can be classified as philosophers (Wittgenstein and Saarinen).

In the middle area of the figure there is an almost empty space surrounded by a variety of people. What could serve as a common denominator for names like Everett Rogers, John Sterman, and Albert-László Barabási? My tentative answer is that they all have studied many kinds of processes. Rather than forming a clear group, it seems that the middle region on the map represents processes or organizational aspects. Besides, it appears that the authors in the outer areas are typically more interested in solving practical problems than building theories.

In addition, the ecosystem map includes more scattered fields, named management, marketing, Internet, and usability. There are several possible reasons for the spreading of these fields. First, there might be too few authors to form a clear group. This probably explains partly the spreading, because the correlation coefficients are now defined in relation to authors in separate fields. As a result, the number of combined hits is small (often less than 10), which generates more randomness in the correlation figures. Increased randomness means both a smaller correlation and more separated locations. Furthermore, the authors located in the periphery of the ecosystem map (for instance, Marc Hassenzahl with 27 600 hits) tend to be have significantly less hits than the authors in the middle (for instance, Malcolm Gladwell with 3 360 000 hits). Finally, some fields might genuinely be scattered and separated. Particularly, if analyzed in this way, the whole field of modern technology is likely formed by tens of separate fields with minimal correlation with each other.

But what did we learn from this example, if anything? Although it is presented in the context of correlation analysis, most of the discussion is general and non-mathematical. If you think of the nature of reality in this respect, the small share of mathematical analysis appears natural. Even though correlation analysis is an integral part of the analysis, the results require a lot of non-mathematical interpretation. Furthermore, the reliability of the results depends essentially on the validity of the input data, most notably, on the selection of the authors. That part of the process definitely requires prior insight.

Lessons for CEE

Could we use this example to make some general (meta-level) observations that even might be useful for a CEE? At least, the example illustrates the difference between an intentional design (Figure I.1) and an evolutionary outcome (Figure A.13). Figure I.1 is highly structured and symmetric, as could be expected due to my engineering background, while Figure A.13 is less regular, as an ecosystem should be. Still they illustrate the same real, but ambiguous system. Note that because the selection of the authors was based on the same insight that was used to draw Figure I.1, there is an association between the figures.

Which one of these two figures is then better? The answer depends, of course, on the metric to be used to compare the benefits of the figures. If you want get a *realistic* picture of the current state of an ecosystem, then probably Figure A.13 is better, and in a way, more objective. Besides, anyone could make his or her own map by selecting a different set of authors and by conforming to the presented procedure. In contrast, if the objective is to design something more systematic, like an organized textbook, Figure I.1 might provide a more solid basis.

Is this kind of research really rational and justified? I spent maybe 15 hours just to gather the data (part of which is not included in the analysis) and many more hours to analyze it and finally write this text. If you think about the question in the framework of value of time, it might seem that the cost was high (noticeably above one thousand Euros).

Finally, as to my position in the ecosystem map, the problem is that I am not known enough to enable a reliable analysis. In September 2011, I generated less than 5 000 hits, and thirteen out of the fifty authors did not appear together with me in any web page. Owing to my Differentiated Services book, I am mostly known as a QoS expert, and thus I belong to the group of Internet experts consisting of Geoff Huston, Henning Schulzrinne, and Van Jacobson (I am the small dot below Van Jacobson in Figure A.13). I hope that I will gradually move towards the centre of the map.

Book recommendations

D. Hofstadter, 1999, *Gödel, Escher, Bach: an Eternal Golden Braid*, New York: Basic Books.

> This is an amazing volume with facts, fiction, formulas, and illustrations. If I had to select two books for a visit to a desert island, I would select this and Ulysses—partly because I could hardly ever understand them completely. Thus, if you want to challenge your mental capabilities try this book.

A. Pickering, 1995, *The Mangle of Practice*, Chicago: The University of Chicago Press.

> Andrew Pickering has undertaken the ambitious goal of understanding the nature of scientific, mathematical, and engineering practice. An expert may develop a high level of expertise on a specific field without thinking of the fundamental nature of the field. In contrast, a communications ecosystem expert must cope with many fields of science and engineering, and in that effort, Pickering's book may provide valuable insight.

L. Wittgenstein, 1958, *Philosophical Investigations*, Englewood Cliffs, NJ: Prentice Hall.

> Wittgenstein has certainly affected my way of thinking, although it is hard to say exactly how. Maybe some of the effects can be seen also in this book (that you are reading) as recurring reflections of type: What does it mean to say, "Something is important?" For me philosophical reflections are important. You, of course, may select another philosopher as a guide towards the realm of deep meanings.

References

Hamel, G., 2007, *The Future of Management*, Boston: Harvard Business Review Press.
Kuhn, T., 1962, *The Structure of Scientific Revolutions*, Chicago: The University of Chicago Press.
Pickering, A., 2010, *The Cybernetic Brain*, Chicago: The University of Chicago Press.
Taleb, N. N., 2010, *The Black Swan*, New York: Random House.