The Concept of REDUNDANCY

Introduction

The word REDUNDANCY stems from the Latin verb „redundare“, which means overflow, be available in ample supply. REDUNDANCY can mean both something positive, like the overflow in the sense of wealth, and something negative, like ballast. This double-entendre makes the expression interesting from the point of view of information theory. We will see that REDUNDANCY has both connotations at the same time, it can mean both wealth and ballast, depending on who is the receiver of information.

Definition of INFORMATION

The following definitions are taken (in a very much simplified form) from the contribution of Warren Weaver, to be found in the classical book written by C.E. Shannon: "The Mathematical Theory of Communication", Univ. of Illinois Press, 1949.

- **Definition 1a:** the INFORMATION is the minimum number of bits, necessary to encode the subject message.
- **Definition 1b:** the INFORMATION is the length of the shortest possible code, used to transport the message.

Examples for INFORMATION

Example 1.1: Weather report, limiting itself to the Sun/no Sun alternatives has the INFORMATION of 1 Bit:

Sunny?	Code
yes	1
no	0

Example 1.2: Weather report, a little more detailed - trying to answer two simple yes/no questions: will it be sunny or clouded, and will it be warm or cold ?

Sunny?	Warm?	Code
yes	yes	11
no	yes	01
yes	no	10
no	no	00

The INFORMATION content of the answer is 2 Bits.

From the information theoretical point of view the word „INFORMATION“ means something rather different from what our everyday language uses it for. It has nothing to do with the ***content*** of the message, and everything with its length, in a more concise fashion, with the minimal length necessary to convey the answer to a complex question - which, as we have seen, can be reduced to a series of simple questions (Q1: will it be sunny?, Q2....) The length of the message - in other words the INFORMATION - will depend on how complex the question is. The larger the number of possible answers, the bigger the INFORMATION.

Example 1.3: There's more telephones on the world than credit cards. So the information of the telephone number must be greater than the one available in the credit card number. How come credit card numbers are substantially longer than the telephone numbers?

Hint: why would anybody want a long telephone number? What about credit cards? The extra length, above the minimum length - i.e. INFORMATION - the REDUNDANCY, may be either wealth or ballast, depending on what the information is used for.

Definition of REDUNDANCY

If the message is longer than strictly necessary, i.e. it is longer than the INFORMATION of the message, then the code contains REDUNDANCY.
Definition 2: REDUNDANCY is the 2-logarithm of the quotient between the code length and the message information:

REDUNDANCY = $\log_2$ (Code length / INFORMATION)

The REDUNDANCY can never be negative. It is zero, when the code is as short as it can be (Code length is identical to the INFORMATION, which is seldom the case). It is 1.0, when the code length is 2 x INFORMATION. The REDUNDANCY present in different messages is eventually a compromise between efficiency concerns (one may want to keep the length and thus the time needed for transmission down) and security issues (if we know things may happen on the way to the message, for instance somebody may want to intercept and possibly change it, we will try to counteract this ahead of time).

Example for REDUNDANCY

Example 2.1: We decide to send our weather report via teletype, using two 8 bit characters:

Weather forecast	Code
Sunny & Warm	SW
Sunny & Cold	SC
Cloudy & Warm	CW
Cloudy & Cold	CC

The INFORMATION did not change and still amounts to 2 Bits.
The code length however increased from 2 to 16 bits.
The REDUNDANCY of this code is $\log_2(16/2) = \log_2(8) = 3$.

Example 2.2: SW above means Sunny and Warm, if you read weather forecasts on a regular basis. It could just as well mean South West, if you are backpacking while reading the weather report. And just as well as Sunny and Cold SC could mean you may be in South Carolina. So why not write out the message in simple English, as shown in the left side column?
The INFORMATION is still 2 bits, but by providing 16 characters per message the code length increases to 16*8 bit. The REDUNDANCY of this code is $\log^2(16*8/2) = \log_2(64) = 6$. Note that given the practical experience, there's usually more than 4 types of weather to be expected, so to talk about redundancy in this example is strictly speaking incorrect.

Example 2.3: Instead of writing out the message we decide to use icons. They are stored as 32x32x8bit images. The code length thus increases to 8192 bits. The REDUNDANCY of this code is $\log_2(8192/2) = \log_2 4096 = 12$.

Comparing REDUNDANCY alternatives

Let's compare 1.2 with Example 2.3:

	Example 1.2 with 2-Bit-Code	Example 2.3 with Icon-Code
INFORMATION	2 Bit	2 Bit
Code length	2 Bit	8*32*32 = 8192 Bit
REDUNDANCY	$\log_2(1) = 0$	$\log_2(4096) = 12$
Transmission time Telephone costs	minimal	excessive
a small transmission error...	destroys everything	makes no difference
when used between people	no good, you need a handbook to be able to communicate	optimal, self-evident, no reading (analphabets!) or language experience required
when used between computers	optimal, just two IF statements needed	very difficult, there are so many possibilities to design an icon

Summary

The lower the number of possible answers to the given question, the lower the amount of INFORMATION needed. On the other hand the higher the uncertainty, the more INFORMATION is needed.
The amount of REDUNDANCY needed or required is one of the prime factors influencing the decision about what code to use for communicating. Null REDUNDANCY is nearly always unfavorable, because it's extremely error prone and unreadable. With increasing REDUNDANCY the code becomes more and more fault-tolerant and it takes less effort for us to use the code and understand the message. We pay for the increased REDUNDANCY by the increase in the bandwidth required and the additional effort, needed to make computers handle the code.

Humans love REDUNDANCY, computer hate it.

People do not like codes with low REDUNDANCY (Telephone numbers, car license plates, account numbers). Computers do not like codes with high REDUNDANCY (spoken language, pictures, music). They do not know how to destroy the REDUNDANCY and retain the INFORMATION, something that is very natural thing to do for a man. This is also the point, where the difficulties in the man-machine dialogue arise: in the transformation between the codes with radically different REDUNDANCIES. When talking to a machine, for instance via keyboard, the human operator must destroy a sizeable amount of his or her natural REDUNDANCY. When giving the answers the machine must try to overcome its natural poverty in REDUNDANCY and use redundant-rich codes (Monitor instead of teletype, color monitor instead of a black and white one etc).

Categorization of information science

Using REDUNDANCY as a guideline we can see the following fields of informatics:

- **Data Processing - converting numbers to numbers**

Transforming from a low-REDUNDANCY code to another low-REDUNDANCY code
Examples: book keeping, administration, statistics

- **Computer Graphics - converting numbers into pictures**

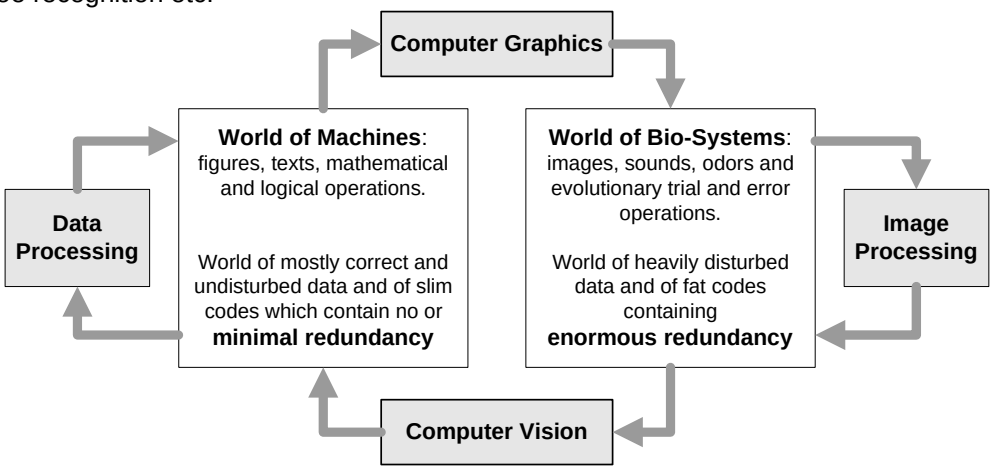
Transforming from a low-REDUNDANCY code to high-REDUNDANCY code
Examples: Graphical user interfaces such as McIntosh, Window, SunOS.
Computer Aided Design, plays and cartoons, visualizing tools

- **Image Processing - changing pictures to pictures**

Transforming from a high-REDUNDANCY code to another high-REDUNDANCY code
Examples: Scanners, copiers, Digital video

- **Computer Vision - understanding images**

Transforming from a high-REDUNDANCY code to a low-REDUNDANCY code
Destroying REDUNDANCY
Examples: Bar code reader, OCR, Fingerprint recognition, Chromosome- & Tumor recognition, Friend-or-Foe recognition etc.



Meanings of INFORMATION and REDUNDANCY

1. Problem: By any of the two definitions above the meaning of INFORMATION is far away from what we use this word for in our everyday conversation. The messages "The little green men landed in my backyard." and "The grass in my backyard needs to be cut." can not possibly carry the same INFORMATION. For the minimal encoding however we probably will need about the same number of bits. So by the above definitions the INFORMATION of the two messages should be about the same, which, as already said, is questionable.
Solution: Shannon (see above) suggests the complement of the probability, i.e. the improbability of the message, as a measure of INFORMATION. The suggestion is brilliant, but it is practically impossible to implement: it presupposes we are able to define and evaluate the complete set of possible answers to a given question. For some classes of cases this is possible. Trying to do this for the question "What will be the weather tomorrow?" - well, one can see a problem our weatherman has every evening.

2. Problem: The meaning of INFORMATION and REDUNDANCY is well-defined only within the communication industry. Everywhere else it's very seldom that the INFORMATION and REDUNDANCY can be quantified. Of course this does not mean, they are worthless. Even when used in qualitative, non-quantitative fashion, they are irreplaceable, when describing the variety of transformations (=re-codings) of messages between people and media.

Example: The big novel "War and piece" by Lev Tolstoy has often been made to a film with different degrees of success. Fact is, one would need much less time to transmit (for instance via Internet) the original novel than any of its film copies. Would that mean that even the worst of the film contains more INFORMATION than the novel? If at all, it's the INFORMATION, that has literally "nothing to do with the novel". The truth is that by moving to the film medium the real, original INFORMATION gets destroyed, because (depending of course on the screen writer) the central themes of the novel get presented as fragments, which are blown up and expanded beyond their original intent. Moving from the written word to the film thus causes a destruction of INFORMATION, which is counteracted by an enormous buildup in REDUNDANCY. After all, one needs to keep people happy for two hours.

3. Problem: There's a deep seated notion that computers work pretty much the same way as the human brain. The truth is the analogy is nothing but a platitude - the differences are enormous. The bio-systems have been optimized by the evolution for the processing of data flows, containing a huge amount of noise. Computers on the other hand have designed to handle relatively small data flows (max 10 MByte/sec), which however are supposed to be error-free.

Example: A human eye contains about 120 Million light sensitive cells, which eventually create an awesome data flow of about 2000 MByte/sec per eye. It is a powerful river of information, that carries on its surface all kinds of debris, the imprint of blood vessels, covering the detectors, our short- and farsightedness, many other optical distortions, our halfway coordinated camera platform in a form of our cranium, to name just a few. And this flow, flowing day and night, night and day, is our umbilical cord to the surrounding world, that provokes us, asks us questions and gives us answers. That, to put is simply ,communicates with us.

The two words, INFORMATION and REDUNDANCY, can help us to understand the basic difference: Computer process the INFORMATION, so they are helpless in face of errors. Biosystems on the other side process the REDUNDANCY and are thus in a completely different league, when it comes to errors and noise. One can say - looking back at evolution, for instance - that biosystems thrive on errors.

On the sober note: the chances for cooperation between such basically different information processing systems is not exactly overwhelming. A man will never have a computer as a friend.

Genetic Code

Now that the sequence of the 3 billion bases of human genome has been determined, the question arises: what codes for what? Where in the genome is the place, where some given property has been encoded? The main problem here is the high REDUNDANCY of the genome, which is at least $\log_2(128) = 7$ if not $\log_2(1024) = 10$.

The REDUNDANCY can be assessed from the following, still not completely clear properties of code: 90% of the code seems to contain nothing but insensible, pure noise. The separation between useful and senseless code is at the moment still not sharp. Most of the properties are coded several times (at least twice) at several distinct places in the genome. Furthermore the codes often differ from place to place. From the length of a sensible genome section one can not extrapolate neither its importance nor its complexity. Very complex proteins have often rather shorter code than one would expect, Complicated controlling steps in the embryonic development can be very short and free of redundancy, some other less important enzymes surprisingly long. Within a given gene long identical sequences can be repeated on one side, on the other side these repeats can be missing without influencing the functionality of the outcome. Genes shift positions, they may split up, join back, move and jump. The expression of majority of genes is influenced by other genes, it can be turned off, enhanced or impeded. During the chromosome replication a number of surprising errors can happen. The repeats of identical code sequences do not per se mean anything - even nonsense may need to be redundant -. About 80% of the human genome seems to be identical to the genome of Drosophila fruit fly. The length of code evidently does not tell anything about the message delivered. Given the fact the code used is identical, we can identify the 20% difference as the time it took evolution to move past the Drosophila to human species. And if we want to understand the starting 80% of the message, Drosophila should do fine.

The unraveling of the genetic information, the identification and isolation of its functionality is one of the central scientific themes on this century. The name of the game is: remove the REDUNDANCY from the Code, distill out the INFORMATION. The job will be finished, when the REDUNDANCY reaches zero. This is probably not possible to achieve, and anyway unnecessary. But: the sharper the separation of REDUNDANCY is, the clearer will be our understanding and the easier it will be to build on it.

We are probably one of the last generations of the original human race. Our descendants, few generations down the road, may well with God's help, expect the redemption from the animal kingdom and the salvation from disease, stupidity, aging or even death. Hundreds of years of health, youth and intelligence beyond our imagination may be waiting for them. Let's hope that future generations will not forget their parents and great parents, which with the limited knowledge and experience set the foundations for their fruitful lives.