Coherence Progress: A Measure of Interestingness Based on Fixed Compressors

Tom Schaul, Leo Pape, Tobias Glasmachers, Vincent Graziano,
Muammar “Dirty Old Man” Gaddafi[**], and Jürgen Schmidhuber

IDSIA, University of Lugano
6928, Manno-Lugano, Switzerland
{tom,pape,tobias,vincent,juergen}@idsia.ch

Abstract. The ability to identify novel patterns in observations is an essential aspect of intelligence. In a computational framework, the notion of a pattern can be formalized as a program that uses regularities in observations to store them in a compact form, called a compressor. The search for interesting patterns can then be stated as a search to better compress the history of observations. This paper introduces *coherence progress*, a novel, general measure of interestingness that is independent of its use in a particular agent and the ability of the compressor to learn from observations. Coherence progress considers the increase in coherence obtained by any compressor when adding an observation to the history of observations thus far. Because of its applicability to any type of compressor, the measure allows for an easy, quick, and domain-specific implementation. We demonstrate the capability of coherence progress to satisfy the requirements for qualitatively measuring interestingness on a Wikipedia dataset.

Keywords: compression, interestingness, curiosity, wikipedia

1 Introduction

The ability to focus on novel, yet learnable patterns in observations is an essential aspect of intelligence that has led mankind to explore its surroundings, all the way to our current understanding of the universe. When designing artificial agents, we have exactly this vision in mind. However, if an artificial agent is to exhibit some level of intelligence, or at least the ability to learn and adapt quickly in its environment, then it is essential to guide this agent to experience such patterns, a drive known as artificial curiosity. However, this approach requires a principled way to judge and rank data, in order to drive itself towards observations exhibiting novel, yet learnable patterns. This property is compactly captured by the subjective notion of interestingness.

Natural and artificial learning agents are equally dependent on the interestingness of their observations. Thus, in order to design intelligent agents, we need a formalization of interestingness. Such formalizations indeed exist, although some of these have shortcomings.

[**] contributed to this paper through his inspirational level of lived coherence.

We focus on compression progress, which is a successful formalization of interestingness. Our contribution is to decompose this measure into a data-dependent and a learning-related part. This decomposition is useful in a number of circumstances, such as when we care specifically about the interestingness of data, explicitly leaving learning effects aside. We propose *coherence progress* as a novel measure of the inherent interestingness of data, and we show in detail how it relates to the more general notion of compression progress.

2 Interestingness

The notion of interestingness as a subjective quality of information has been investigated in various ways in the literature, ranging from early work by Wundt [9] (see Figure 1), to the attempt of a full information theoretic formalization [6, 5, 8]. Based on its intuitive notion as the discovery of novel patterns, we can identify a number of qualitative requirements for any measure of interestingness:

1. Observations can be *trivial*, that is inherently uninteresting, such as a white wall. When observations have a simple structure and can be completely described by very simple rules they become boring very quickly.
2. The opposite of these are completely *random* observations. Completely random data contain no patterns at all, and are therefore not interesting neither. It is important to note that the same argument holds with information that *seems* random to the observer, e.g., the content of a mathematics textbook will appear random to most children.
3. Between these extremes of minimal and maximal complexity lies the domain of complex, yet structured observations. Here, the subjective nature of interestingness becomes patent. If the observer is already familiar with all the (*repeated*) patterns in the observations, no new patterns can be discovered, and the observations are no longer interesting.
4. Interesting observations can now be identified relative to the patterns the observer already knows. Observations with trivial, well-known, and overly complex patterns are not interesting. Instead, only observations that contain patterns that are not yet known, but can be learned by the observer are interesting (e.g., the same math book can be highly interesting when the reader has acquired the necessary background, given he does not already know it). As the observer discovers more patterns in its environment the interestingness of observations changes. Crucially also, patterns discovered by imperfect observers might be *forgotten* after a while, making a previously uninteresting observation interesting again.

To summarize, any quantitative measure of interestingness must assign low values to patterns the observer already knows, and patterns the observer cannot learn; and high values to patterns the observer does not yet know, but can still discover. Moreover, increasingly difficult patterns to learn should be assigned decreasing interestingness values. Given a choice of which observations to consider next, the observer could assign its resources to the next easiest pattern to learn.

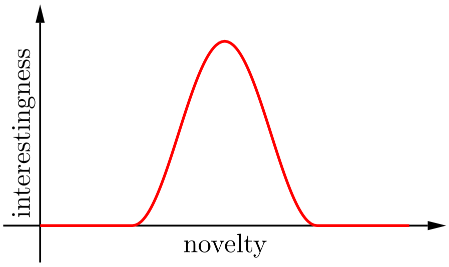

Fig. 1. Wundt Curve. The x-axis represents the *novelty* of the information. Novelty depends on the relationship between the information and the person observing it. Trivial patterns quickly lose their novelty, noise is always novel. Well-known patterns lack novelty and these too are not interesting. As learning proceeds, the complexity of the patterns which are most interesting increases, but the converse does not hold: As learning progresses, less complex patterns do not necessarily become less interesting, nor do more complex patterns necessarily become more interesting. The change in interestingness is a function of that which has been learned, and cannot be simplified to a general shift towards patterns of higher complexity.

Measures closely related to interestingness are commonly found in data mining. In large databases one often wants to mine for association rules between sets of items, which may return an intractable number of rules. Measures of interestingness are valuable for reducing this set to a smaller subset. Typical criteria in common use are conciseness, reliability, diversity, novelty, surprisingness, and utility. Some of these measures lack subjectiveness, and the subjective measures typically only fulfill one or two of the criteria above. A review on such measures can be found in [1].

Bayesian surprise [4], a measure of the difference between the prior beliefs and the posterior beliefs (the term actually refers to the earlier idea presented in [8]), is an approach closely related to the formalization of interestingness, as both surprise and interestingness are cognitive concepts attributed subjectively to information of observations. Because of their subjective nature, both concepts judge information in the context of an observation history. However, surprise and interestingness as cognitive concepts are not fully equivalent: Any interesting observation is to some extent surprising, because completely predictable information is not interesting. However, surprising data are most interesting if they exhibit novel patterns, while it is possible to be surprised by inherently random and thus uninteresting events.

A general approach is to base interestingness on the very general concept of data *compression*. For example, the entropy of data (which is related to complexity, not necessarily interestingness) can be expressed in terms of compression by relying on a purely statistical compressor, such as Huffman [3] codes or entropy-based encoding. *Compression progress* [7] was the first attempt to capture interestingness using compression. Its guiding principle is that any process that increases compressibility is interesting. This allows for a measure of interestingness based on a well-defined information theoretic concept: the negative of the time derivative of the length of the compressed data. This paper moves beyond compression progress in that it introduces a distinction between two separate, orthogonal components.

3 Fixed, Adaptive and Ideal Compressors

Data compression refers to the process of encoding information by means of a shorter code. Typically we understand a compressor as a program that, given an input string $x = (x_1 x_2 \dots x_n)$, outputs a (shorter) output string $y = C(x)$, where $y = (y_1 y_2 \dots y_m)$, such that there exists another program, the decompressor, for reconstructing x from y: $C^{-1}(y) = x$. This function may depend on additional parameters w, in which case we write $C_w(x)$.

Many different types of compressors can be distinguished, some are more or less fixed programs, while others methods can adjust by learning from observations, such as neural networks. While many of these approaches involve adjustable parameters, here we introduce a clear distinction between *fixed* and *adaptive* (or *learning*) compressors.

Essentially, we treat a compressor as fixed, if each time it is invoked it starts with the same w (and thus $C(x)$ keeps producing the same encoding for identical inputs). For this distinction to be clear, it may be helpful to think of a compressor as a program that makes predictions of the next observation it will see, based on the the observations so far At any point t in the sequence x a predictor f predicts the subsequent observation x_t based on the seen part of x, i.e., the history $h = (x_1 x_2 \dots x_{t-1})$. (This directly allows for compression, in that high-probability observations can be encoded with short codes.) So, if the predictor for the next symbol $\hat{x_t} = f(h)$ is a fixed function that depends only on the history, the corresponding compressor is fixed. However, because a compressor is essentially equivalent to a predictor, it is tempting to replace the fixed function $f(h)$ with a learning machine that takes advantage of experience. For example, the predictor may learn that in English texts, there is a high chance for the letter ‘y’ to follow the sequence ‘happ’. And this kind of knowledge may well *transfer*, i.e., be useful for compressing other sequences x' (e.g., shortening the code of the first occurrence of ‘happy’ in x'). This transfer (stored in changing parameters w) is precisely the essence of an adaptive compressor.

Note that this distinction is more subtle than is appears, because they rely on distinguishing w and h by their role, even though in principle one could be incorporated into the other one (e.g., presenting `gzip` with dictionary D containing words like ‘happy’ before we start compressing x). So the same compressor (`gzip`) can be seen as adaptive if we keep adapting (learning) D, or fixed otherwise.

Interestingly, the *ideal* compressor is non-adaptive. Ideal, or *Kolmogorov* compression amounts to encoding the input string x by the shortest program y in a Turing complete language that outputs x. Per definition, this ideal compression scheme is theoretically optimal, even if incomputable (because when searching for the program y for a given x we run into the halting problem).

4 Coherence Progress

In order to formally introduce coherence progress, we first define a couple of auxiliary concepts. We call *compression similarity* between two sets a and b, the difference between their length when compressed together, and the sum of their individual compressed lengths:

$$s_C(a,b) = l_C(a) + l_C(b) - l_C(a \cup b)$$

where $l_C(x)$ is the length of the resulting string when compressing a set x with a (fixed) compressor C.[1] This measure clearly depends on (the quality of) the compressor used, and is measured in bits. Furthermore, for reasonable compressors, we have $s_C(a,b) \geq 0$ and $s_C(a|\emptyset) \approx 0$.

Next, *compression coherence*, is a measure on sets: For any partitioning of a set or sequence h into a and b ($a = h \setminus b$), we can compute the compression similarity $s_C(a,b)$, and if we average over this, the resulting value is a measure of how closely the elements (and subsets) of h are related to each other:

$$\overline{s_C}(h) = \frac{1}{|\mathcal{P}(h)|} \sum_{b \subset h} s_C(h \setminus b, b)$$

Here, $\mathcal{P}(x)$ denotes a set of subsets of x, for example the power set of x, or in case of sequential data a set of sub-sequences, such as $\{h_{1:1}, \ldots, h_{1:t}\}$, where $h_{t_1:t_2}$ denotes the history from time t_1 to t_2, inclusively. The choice of $\mathcal{P}(x)$ depends mostly on the types of relations we want to capture, and depending on the context several choices will result in a reasonable measure of interestingness.

So if all elements of h are unrelated, $\overline{s_C}(h) = 0$, whereas if they are highly related (e.g. all images of donkeys), $\overline{s_C}(h)$ is high. Note that if h contains a single element, then $\overline{s_C}(h) = 0$.

We now consider the case where we incrementally have more data available, the history h_t (at time t). The history is a set of observations o_t and $h_{t+1} = h_t \cup \{o_t\}$. We want to determine the *coherence progress*, that is the amount by which the coherence of the history h_t increases when a new observation o_t becomes available:

$$P_C(t) = P_C(o_t|h_t) = \overline{s_C}(h_{t+1}) - \overline{s_C}(h_t).$$

An alternative formulation is

$$\begin{aligned}
P_C(t) = P_C(o_t|h_t) &= \overline{s_C}(h_{t+1}) - \overline{s_C}(h_t) \\
&= \frac{1}{|\mathcal{P}(h_t)|} \sum_{b \subset h_t} \Big[[l_C(h_{t+1} \setminus b) + l_C(b) - l_C(h_{t+1})] - [l_C(h_t \setminus b) + l_C(b) - l_C(h_t)] \Big] \\
&= \frac{1}{|\mathcal{P}(h_t)|} \sum_{b \subset h_t} \Big[[l_C(h_{t+1} \setminus b) - l_C(h_t \setminus b)] - [l_C(h_{t+1}) - l_C(h_t)] \Big] \\
&\approx -\frac{\partial}{\partial t} l_C(h)\Big|_t + \frac{1}{|\mathcal{P}(h)|} \sum_{b \subset h} \frac{\partial}{\partial t} l_C(h \setminus b)\Big|_t .
\end{aligned}$$

[1] We use set notation such as $a \cup b$ and $h \setminus b$ in this section for both sets and sequences. The obvious meaning for sequences is that the original order of the symbols is preserved. This does not directly affect the question whether the order of symbols is of importance.

For the choice of $\mathcal{P} = \{o_{t-1}\}$, averaging only over the previous observation $b = o_{t-1}$ instead of over all subsets b, we get

$$\hat{P}_C(t) \approx -\frac{\partial}{\partial t} l_C(h)\Big|_t + \frac{\partial}{\partial t} l_C(h)\Big|_{t-1} \approx -\frac{\partial^2}{\partial t^2} l_C(h)\Big|_t .$$

So, in another possible intuitive understanding, we can say that coherence progress is the negative second derivative of the compressed length of the history, except more robust, because of the averaging over all the partitions.

4.1 Qualitative Correctness

We now return to the qualitiative intuitions introduced in section 2, and show how our formalization of coherence progress is indeed a good candidate for interestingness.

1. If an observation is uninteresting per se, i.e., if $l_C(o)$ is vanishingly small for any reasonable compressor, then clearly $\forall h, l_C(h \cup o) \approx l_C(h)$ and thus $P_C(o|h) \approx 0$.
2. If an observation is random, then it will also be virtually uncompressible, which means that $l_C(o) \approx |o|$, and not help compress other observations: $\forall h, l_C(h \cup o) \approx l_C(h) + l_C(o)$, and therefore $P_C(o|h) \approx 0$.
3. If an observation is well-known, i.e., very similar to many of the past observations, that means that the coherence is high, but the coherence progress will be small $\overline{s_C}(h \cup o) \approx \overline{s_C}(h) \gg 0$,
4. In all other cases, the compression similarity $s_C(o, b)$ will be non-zero, for at least some subsets $b \in h$, and thus probably $P_C(o|h) > 0$.

4.2 Oversimplified Alternatives

Occam's razor entices us to choose a measure of interestingness that is as simple as possible, so in this section we show a few alternatives that are simpler than our suggested coherence progress in formulation, and why they do not satisfy the criteria of a good measure of interestingness.

1. The compression similarity $s_C(o_t, h_t)$ does not work, because it is maximal for repetitions of previous observations.
2. The so-called compression distance $l_C(h_{t+1}) - l_C(h_t)$ does not work, because random, unrelated observations always have a positive (and maximally high) value.
3. The normalized compression distance $\frac{l_C(h_{t+1})}{|h_{t+1}|} - \frac{l_C(h_t)}{|h_t|}$ has similar problems to the previous one, and an additional problem because now appending long blanks (that are easily compressible) to some observations changes the outcome significantly.
4. The derivative of compression distance, that is, the second derivative of compressed length $\hat{P}_C(o|h)$, which we introduced above as an approximation of $P_C(o|h)$. This is a more interesting case, but it can still be problematic,

because the robustness from the averaging is lost. To illustrate how this can lead to an unintuitive result, consider the case where each observation is random, but their size increases (decreases) by some amount at each step: then $\hat{P}_C(o|h)$ is a positive (negative) constant, although all observations are unrelated (compression similarity of 0).

5. A normalized form of the above does not solve the problem neither, rather it adds the issue of padding with blanks (3, above) to the case.

4.3 Coherence Progress versus Learning Progress

The classical framework of compression progress is more general than ours, because it assumes an adaptive compressor instead of a fixed one. We can separate its two components: coherence progress, as a measure based purely on the data, and *learning progress*, a measure of what has been learned from experience, as encoded in the changes of the parameters w.[2] In short, while coherence progress is purely a measure of interestingness of the new observation (given h), pure learning progress does not require a new observation, and instead is a measure of how interesting (useful) a change of the compressor's parameters w has been.

For example, consider the case of an adaptive compressor based on a learning algorithm, say, an auto-encoder neural network, the predictive power of which is used to compress the data. In this case, the parameters of the network w can be trained on a sequence x, e.g., though back-propagation, becoming w', which then may improve the compression: $l_{C_{w'}}(x) < l_{C_w}(x)$ (say, if both are English texts). This difference in compressed lengths therefore is (one form of) pure learning progress, as it captures the interestingness inherent to the learning process itself: we can relate it to 'thinking through' of past experience, an activity that is interesting to the degree where we gain new insights about it.

Disentangling compression progress, and separating it into its data-dependent an learning-dependent components are helpful. On the one hand, it allows us to explicitly analyze and trade off the two types of progress, which might have different cost scales (learning is usually measured in its computational cost, whereas getting new observations might involve a substantial monetary cost). On the other hand, if data acquisition and learning are realized by different mechanisms it permits us to disambiguate the success of the different units.

5 Experiments

We start with an illustrative general example, which we can handle analytically. Suppose all observations o_i are identical and uncompressible strings of length n. Assuming a reasonable compressor and n large allows us to disregard any small constant effects, and we have $\forall i, j$: $l_C(o_i) = n$, $l_C(o_i \cup o_j) = n$, $s_C(o_i, o_j) = n$,

[2] Note that the ideal (Kolmogorov) compressor is a fixed compressor, which precludes it from making any learning progress.

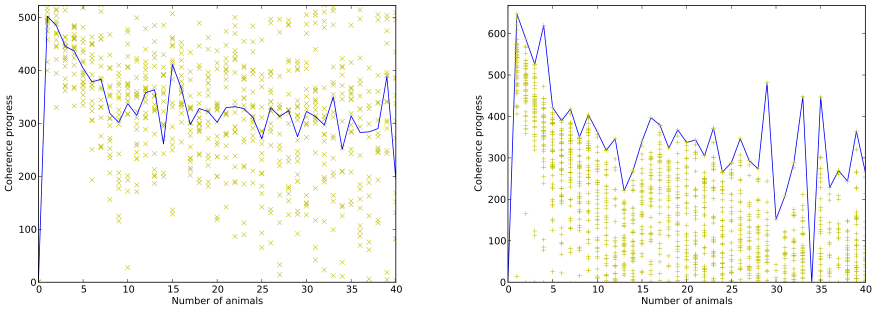

Fig. 2. Coherence progress on the animal dataset. Left: randomly choosing the next animal to add to the set (average over 10 runs, each shown as crosses). Right: greedily adding the article which maximally increases coherence (or, equivalently, coherence progress), at each step (suboptimal choices shown as yellow crosses). The choice of articles was among all animals, (repetition allowed), an empty article, and a randomly scrambled one. The latter two did never get picked (for reasons described in the previous sections), nor did any repetitions become more interesting than new animals. Starting from ‘Human’, the next animals picked were ‘Chimpanzee’, ‘Hippopotamus’, ‘Jaguar’ and ‘Leopard’. We also notice that in the greedy case, the fist few additions give a significantly higher coherence progress than in the random case (left).

and even $s_C(o_i, h_t) = n$, for any $t > 0$.

$$\overline{s_C}(h_t) = \frac{1}{|\mathcal{P}(h_t)|} \sum_{b \subset h_t} s_C(h_t \setminus b, b) = \frac{1}{2^t} \left[2 \cdot 0 + (2^t - 2)n\right] = n(1 - 2^{1-t})$$

because in 2 cases b or its complement are empty, and in all other cases the similarity is constant. Thus, we see that coherence progress follows an exponential decay trend:

$$P_C(o_t|h_t) = \overline{s_C}(h_{t+1}) - \overline{s_C}(h_t) = n(1 - 2^{1-(t+1)}) - n(1 - 2^{1-t}) = 2^{-t}n.$$

In a more realistic setting, we investigate whether coherence progress gives us a reasonable measure of interestingness when the observations are Wikipedia articles. We chose articles in the class of animals (the 50 with the largest entries) and movies (the 50 with most Oscar wins). As averaging over all possible partitions is intractable for large sets, in the remainder of this section, we approximate the true coherence (progress) by averaging over 200 random partitions.

In a first experiment, we show how coherence progress evolves as more and more of the articles of a class (animals here) get added to the history without any particular order (Figure 5, left). Similarly, we can make a greedy choice before each addition to pick the animal article that will maximally increase coherence (see Figure 5, right).

In a second experiment we decided to determine to what degree knowing about objects in one class makes more observations in the same class interesting

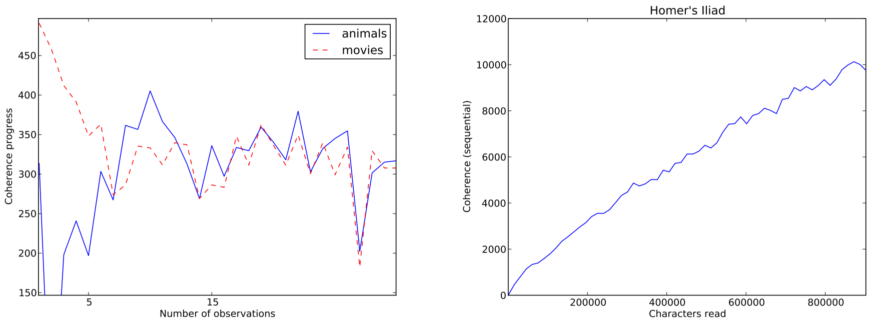

Fig. 3. Left: We plot the average (hypothetical) coherence progress for adding a movie (dashed, red) or an animal (full, blue) at each step of a sequence (which consists of 5 movies, followed by 10 animals, followed by 15 movies). We see that while the history contains only movies, those are more interesting, but after a few animals are added, those become more or less equally interesting. Right: Evolution of compression coherence when incrementally considering longer pieces of a sequential text (Homer's Iliad). Note the slowly diminishing returns, and that how the limited approximation introduces more noise, the longer the sequence (because it can capture only a shrinking fraction of the possible partitions).

– versus observations from a different class (see Figure 3, left). In our last experiment (Figure 3, right), we illustrate how the sequential variant of coherence progress can be employed to track the progress in a long sequential text (in our case, Homer's Iliad [2]).

Together, these experiments illustrate what values of interestingness coherence progress provides in practice, show the broad applicability and are (arguably) qualitatively on par with interestingness values a human would express.

6 Discussion

One use-case within the framework of artificial curiosity, which assumes an agent learning about the world, may be to disentangle coherence progress and learning progress (see section 4.3). However, coherence progress is also applicable to systems designed to explore, but without learning at the same time—i.e., classical compression progress is not applicable.

A possible direction of future work could be to validate our results also quantitatively, with data from humans (or primates) from psychological experiments. Another one would be to design a normalized version of coherence progress (e.g., taking values in the unit interval), removing the dependence on the size of the observations and the number of elements in the set, which may be useful in applications where those differ vastly over time.

More concretely, a measure like coherence progress could be a powerful addition to applications like recommender systems (say, Amazon or Netflix): they may provide a measure of how interesting an upcoming book or movie is to users *before* the first user has seen/rated it, based on the history of the user.

7 Conclusion

This paper has introduced coherence progress, a novel measure of interestingness that depends only on data, and is independent of any learning mechanism. It at once matches the qualitative requirements for such a measure, is formally specified for any type of (possibly domain-specific) compressor, and can be used effectively in practice, as shown in our experiments on Wikipedia data.

Acknowledgments We thank Giuseppe Cuccu and Jürgen Leitner for valuable discussions. This research was funded in part through the 7th framework programme of the European Union, under grant number 231576 (STIFF project), and SNF grant number 200021-113364/1.

References

1. L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38, September 2006.
2. Homer. Iliad (ca. 800 BC). Translated by Alexander Pope, London, 1715.
3. D. A. Huffman. A method for construction of minimum-redundancy codes. *Proceedings IRE*, 40:1098–1101, 1952.
4. L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems 19*, pages 547–554. MIT Press, Cambridge, MA, 2005.
5. J. Schmidhuber. Curious model-building control systems. In *Proceedings of the International Joint Conference on Neural Networks, Singapore*, volume 2, pages 1458–1463. IEEE press, 1991.
6. J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, pages 222–227. MIT Press/Bradford Books, 1991.
7. J. Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Anticipatory Behavior in Adaptive Learning Systems. From Psychological Theories to Artificial Cognitive Systems*, volume 5499 of *LNCS*, pages 48–76. Springer, 2009.
8. J. Storck, S. Hochreiter, and J. Schmidhuber. Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the International Conference on Artificial Neural Networks, Paris*, volume 2, pages 159–164. EC2 & Cie, 1995.
9. W. M. Wundt. *Grundzüge der Phvsiologischen Psychologie.* Leipzig: Engelmann, 1874.