Abstract—The success of statistical machine learning (ML) methods made the field of Artificial Intelligence (AI) so popular again, after the last AI winter. Meanwhile deep learning approaches even exceed human performance in particular tasks. However, such approaches have some disadvantages besides of needing big quality data, much computational power and engineering effort; those approaches are becoming increasingly opaque, and even if we understand the underlying mathematical principles of such models they still lack explicit declarative knowledge. For example, words are mapped to high-dimensional vectors, making them unintelligible to humans. What we need in the future are context-adaptive procedures, i.e. systems that construct contextual explanatory models for classes of real-world phenomena. This is the goal of explainable AI, which is not a new field; rather, the problem of explainability is as old as AI itself. While rule-based approaches of early AI were comprehensible "glass-box" approaches at least in narrow domains, their weakness was in dealing with uncertainties of the real world. Maybe one step further is in linking probabilistic learning methods with large knowledge representations (ontologies) and logical approaches, thus making results re-traceable, explainable and comprehensible on demand.

I. INTRODUCTION

This talk is divided into six sections: 1) I will start with explaining the HCI–KDD approach towards integrative machine learning (ML); 2) I will continue with discussing the importance of understanding intelligence and 3) show very briefly our application domain health, where it becomes clear why 4) dealing with uncertainty is important. A quick journey through some successful applications of 5) automatic machine learning (aML) will bring us to the limitations of these and to find novel connections between data objects and the world. Graph theory provides powerful tools to map data structures and to perceive information in dimensions down to the lower dimensions, ultimately to \( \mathbb{R}^2 \). It is a hard task to map the results, gained in arbitrarily high dimensional spaces, down to the lower dimensions, ultimately to \( \mathbb{R}^2 \).

II. WHAT IS THE HCI–KDD APPROACH?

ML is a very practical field. Algorithm development is at the core, however, successful ML requires a concerted effort of various experts with diverse background. Such a field needs an integrated approach: Integrative Machine Learning (iML) is based on the idea of combining the best of the two worlds dealing with understanding intelligence, which is manifested in the HCI–KDD approach:[2, 3, 4]: Human–Computer Interaction (HCI), rooted in cognitive science, particularly dealing with human intelligence, and Knowledge Discovery/Data Mining (KDD), rooted in computer science particularly dealing with artificial intelligence. This approach fosters a complete machine learning pipeline beyond algorithm development. It includes knowledge extraction, ranging from issues of data preprocessing, data mapping and data fusion of heterogeneous and high-dimensional data sets up to the visualization of the results in a dimension accessible to a human end-user and making data interactively accessible and manipulable. Thematically, these Machine Learning & Knowledge Extraction (MAKE) pipeline encompasses seven sections ([5], [6]):

Section 1: Data: data preprocessing, integration, mapping, fusion. This starts with understanding the physical aspects of raw data and fostering a deep understanding of the data ecosystem, particularly within an application domain. Quality of the data is of utmost importance.

Section 2: Learning: algorithms. The core section deals with all aspects of learning algorithms, in the design, development, experimentation, testing and evaluation of algorithms generally and in the application to application domains specifically.

Section 3: Visualization: data visualization, visual analysis. At the end of the pipeline there is a human, who is limited to perceive information in dimensions \( \leq 3 \). It is a hard task to map the results, gained in arbitrarily high dimensional spaces, down to the lower dimensions, ultimately to \( \mathbb{R}^2 \).

Section 4: Privacy: Data Protection, Safety & Security. Worldwide increasing demands on data protection laws and regulations (e.g., the new European Union data protection directions), privacy aware machine learning becomes a necessity not an add-on. New approaches, e.g., federated learning, glass-box approaches, will be important in the future. However, all these topics need a strong focus on usability, acceptance and also social issues.

Section 5: Network Science: Graph-Based Data Mining. Graph theory provides powerful tools to map data structures and to find novel connections between data objects and the inferred graphs can be further analyzed by using graph-theoretical, statistical and ML techniques.

Section 6: Topology: Topology-Based Data Mining. The

From Machine Learning to Explainable AI

Andreas Holzinger
Holzinger Group, HCI-KDD, Institute for Medical Informatics, Statistics & Documentation
Medical University Graz, Austria
and
Institute of Interactive Systems & Data Science, Graz University of Technology, Austria
a.holzinger@hci.kdd.org

Keynote Talk at IEEE DISA 2018 Conference, Kosice, August, 23, 2018
most popular techniques of computational topology include homology and persistence and the combination with ML approaches would have enormous potential for solving many practical problems.

**Section 7: Entropy: Entropy-Based Data Mining.** Entropy can be used as a measure of uncertainty in data, thus providing a bridge to theoretical and practical aspects of information science (e.g., Kullback–Leibler Divergence for distance measure of probability distributions).

I will explain our HCI–KDD logo in more detail later on, but before we shall talk briefly about understanding intelligence.

III. UNDERSTANDING INTELLIGENCE

"Solve intelligence – then solve everything else" ... if I would say this, my students would not believe it; therefore not I am saying this, but it is the official motto of Google Deepmind (see e.g. the talk by Demis Hassabis from May, 22, 2015).

Now let me explain our HCI-KDD logo in more detail: Augmenting human intelligence (left) with artificial intelligence (right) means mapping results from high-dimensional spaces into the lower dimensions [3]. The logo shall indicate the connection between Cognitive Science and Computer Science: Cognitive Science studies the principles of human intelligence and human learning [7]. Our natural surrounding is in $\mathbb{R}^3$ and humans are excellent in perceiving patterns out of data sets with dimensions of $\leq 3$. In fact, it is amazing how humans learn and extract so much knowledge even from little or incomplete data [8]. This is a strong motivator for the concept of interactive Machine Learning (iML), i.e., using the experience, knowledge, even the intuition of humans to help to solve problems which would otherwise remain computationally intractable. However, in most application domains, e.g., in health informatics, we are challenged with data of arbitrarily high dimensions [9]. Within such data, relevant structural patterns and/or temporal patterns ("knowledge") are often hidden, knowledge is difficult to extract, hence not directly accessible to a human. There is need to bring the results from these high dimensions into the lower dimension for the human end user – the "customer" of ML/AI.

Computer Science studies the principles of computational learning from data to understand artificial intelligence [10]. Computational learning has been of general interest for a very long time, but we are far away from solving intelligence: facts are not knowledge and descriptions are not insight, and new approaches are needed. A challenge is to interactively discover unknown patterns within high-dimensional data sets. Computational geometry and algebraic topology may be of great help here [11]. For example, if we define $M$ as hidden parameter space, and we define $\mathbb{R}^D$ as an observation space, and let $f: M \rightarrow \mathbb{R}^D$ be a continuous embedding; $X \subset M$ be a finite set of data points, and $Y = f(X) \subset \mathbb{R}^D$ shall be the image of these points under the mapping $f$. Consequently, we may refer to $X$ as the hidden data, and $Y$ as the observed data. If we suppose that $M, f$ and $X$ are unknown, but $Y$ is known, the question remains if we can identify $M$? [12]. Such questions are studied in section 6 of the MAKE-approach [13].

Consequently, to reach a level of usable intelligence, we need (1) to learn from prior data, (2) to extract knowledge, (3) to generalize, (4) to fight the curse of dimensionality, (5) to disentangle the underlying explanatory factors of the data [14] and (6) to understand the data in the context of an application domain. One grand challenge still remains open: to make sense of the data in the context of the application domain. The quality of data and appropriate features matter most, and previous work has shown that the best-performing methods typically combine multiple low-level features with high-level context [15].

We compare a DQN-agent (Deep Q-learning, Q-learning is a kind of model-free reinforcement learning [16]) with the best reinforcement learning methods, where we normalize the performance of the DQN-agent with respect to a professional human games tester (that is, 100% level) and random play (that is, 0% level). It can be seen that DQN outperforms competing methods in almost all the games, and performs at a level that is broadly comparable with or superior to a professional human games tester (that is, operationalized as a level of 75% or above) in the majority of games. However, still humans are much better in certain games, and the question remains open why [7].

Always robotics is seen as the most feared threat of AI for mankind. In this talk, I want to emphasize that humanoid AI ≠ human-level AI. Achievement of human-level machine intelligence was the basic objective since the early days of AI. It was actually started by Alan Turing with his question Can machines think? [17]. Exaggerated expectations led to a bitter AI-winter, and recently we have been feeling a real AI-spring [18].

IV. APPLICATION AREA: HEALTH

Why is the application domain health complex? In medicine we have two different worlds: we have the science of medicine: mathematics, physics, physiology, biology, chemistry, etc. at the bench; and there is the clinical medicine focusing on the patient at the bed side. The main problem is that there is a (big) gap between those two (and some say not only a big gap, but there is an ocean between those two worlds). How can we bridge this gap? Our central hypothesis is: information may bridge this gap. Not data. Not knowledge. It is the quality of information what both sides need for making decisions [19]. Optimally designed workflows thereby integrating sophisticated AI/ML with appropriate visualization [20] directly into the workplaces of medical professionals may therefore be a great help for future medicine.

In the medical domain the Number 1 problem is the bad quality of data along with the heterogeneity of data [21]. Consequently data integration, data fusion and data mapping

1https://hci-kdd.org

2Although this can range from tablet computers to large wall-displays the representation is always limited to $\mathbb{R}^2$.

3Intelligence does not need a metal body to be a thread.
is one of the most important issues of data science and the combination of various data would enable to get new kinds of information, hence novel insights.

However, most of the data are in arbitrarily high dimensions, therefore we are always confronted with the curse of dimensionality [22].

A further problem is complexity, of all sort: clinical practice, organization and information management are interdependent and built around multiple self adjusting and interacting systems. Here we are always confronted with unpredictability, non-linearity and non-homogeneity in time [23].

All these lead us to the fourth main problem: uncertainty. V. DEALING WITH UNCERTAINTY: STATISTICAL LEARNING FROM BIG DATA

We now briefly rehearse the very basics of what makes current ML so successful. The principles are so fascinating simple: Bayesian learning, optimization and prediction (inverse probability), which go back to Thomas Bayes and Richard Price [24]. I emphasize here that it was Pierre Simon Laplace who did the pioneering work and delivered us the foundations of statistical machine learning [25], [26].

Let us consider \( \mathcal{D} \) data contained in a set \( \mathcal{D} = \{x_1, x_2, ..., x_n\} \)

and let us write down the expression for the likelihood:

\[
p(D|\theta) = \sum_y p(x, y)
\]

Next we specify a prior:

\[
p(\theta)
\]

Finally we can compute the posterior:

\[
p(\theta|\mathcal{D}) = \frac{p(D|\theta) * p(\theta)}{p(D)}
\]

The inverse probability allows to learn from data, infer unknowns, and make predictions [27].

Most fascinating is the simplicity of this approach; we can add probabilities (sum rule):

\[
p(x) = \sum_y p(x, y)
\]

By introducing ”repeated adding” (adding multiple times), we can write (product rule):

\[
p(x, y) = p(y|x) * p(y)
\]

Laplace (in 1773 !) showed that we can write:

\[
p(x, y) * p(y) = p(y|x) * p(x)
\]

and introduced a third operation (division):

\[
\frac{p(x, y) * p(y)}{p(y)} = \frac{p(y|x) * p(x)}{p(y)}
\]

now we can reduce this fraction by \( p(y) \) and we receive what is today called Bayes rule (actually it was Laplace)\(^4\):

\[
p(x, y) = \frac{p(y|x) * p(x)}{p(y)}
\]

And this is now the basis for machine learning and applied in all sort of advanced techniques, based on adding, and repeated adding, and this is what our Von Neumann machines can do good. However, it is not that easy in high-dimensional spaces, and a grand challenge is: How to add efficient.

In the following the large \( \mathcal{H} \) is the hypothesis space, and e.g. decision making is searching for an optimal solution in an arbitrarily high dimensional search space. However, in medicine we need not always the optimal solution often a good solution in short time is better because time is a very critical aspect!

If we denote \( d \) as the data and \( h \) as the hypothesis and with \( \mathcal{H} = \{H_1, H_2, ..., H_n\} \) then \( \forall(h, d) \)

\[
P(h|d) = \frac{P(d|h) * P(h)}{\sum_{h \in \mathcal{H}} P(d|h)P(h)}
\]

\[
P(h|d) = \frac{P(d|h) * P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}
\]

\[
J = \int f(\theta) * p(\theta|\mathcal{D})d\theta
\]

These astonishingly simple fundamentals led to the current success in automatic machine learning.

VI. AUTOMATIC MACHINE LEARNING (AML)

The ML community today is concentrating on automatic machine learning (aML) approaches, with the grand goal of bringing humans-out-of-the-loop [28], resulting in fully autonomous solutions. Maybe the best practice real-world example of today is autonomous driving [29].

This automatic machine learning (aML) works well when having large amounts of training data [30]. That means the often debated "big data" issue is not bad, instead the large amount of data is beneficial for automatic approaches (and I emphasize again the need of good quality data!).

However, sometimes we do not have large amounts of data, and/or we are confronted with rare events and/or hard problems. The health domain is a representative example for a domain with many such complex data problems [31, 32]. In such domains the application of fully automatic approaches ("press the button and wait for the results") seems elusive in the near future.

Again, a good example are Gaussian processes, where aML approaches (e.g., kernel machines [33]) struggle on function extrapolation problems, which are astonishingly trivial for human learners [34].

A famous example was given by [35] where they considered the problem of building high-level, class-specific feature detectors from unlabeled for detecting a cat automatically on the

\(^4\) to be completely correct it should be called Bayes-Price-Laplace: BPL.
basis of 10 million 200x200 pixel internet images. They trained a deep autoencoder on a cluster with 1,000 machines (16,000 cores) for three days. The results impressively demonstrated that it is possible to train an image detector without labelling the images - but at what price.

A more recent example is the work by [36], who presented fully automated classification of skin lesions using dermatological images. They trained deep convolutional neural networks with a data set of 129,450 clinical images. They used a GoogleNet Inception v3 network, pretrained on approximately 1.28 million images (1,000 object categories) from the 2014 ImageNet Large Scale Visual Recognition Challenge. The authors tested the performance against 21 board-certified dermatologists on biopsy-proven clinical images with two critical binary classification use cases: keratinocyte carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi. The automatic ML achieves performance on par with the 21 medical doctors, demonstrating that AI is actually capable of classifying skin cancer with a level of competence comparable to dermatologists.

Despite their impressive results such automatic approaches have some limitations:

1) they are very data intensive, need often millions of training samples of highest quality which is both hard to achieve;
2) they are non-convex, difficult to set up, difficult to train and to optimize, are very error prone and sensitive to adversarial examples, consequently need a lot of engineering effort;
3) the are affected by catastrophic forgetting, so when a problem changes only slightly, it looses the learned parameters, this calls urgently for transfer learning and multi-task learning;
4) they are very resource intensive, need much computational power and storage;
5) they are bad in dealing with uncertainties; but
6) most of all such approaches are considered to be black-box approaches, they lack transparency, do therefore not foster trust and acceptance, and legal aspects make such opaque models extremely difficult to use in certain situations.

So, sometimes we (still) need a human-in-the-loop. Sometimes we do not have big data where aML algorithms benefit, sometimes we have only small amount of data sets (see e.g. [37], or we have rare events or even no training samples, or we deal with NP-hard problems, e.g. subspace clustering, protein folding, or k-anonymization, to name three. This leads us directly to interactive machine learning.

VII. INTERACTIVE MACHINE LEARNING (iML)

However, we cannot compare car driving with the complexity of the biomedical domain. The main problem for automatic solutions is in the extremely poor quality of data in this domain. Biomedical data sets are full of uncertainty, incompleteness etc.; they can contain missing, wrong data, noisy data, dirty data, unwanted data, etc. However, many problems are computationally hard. All these constraints make the application of fully automated approaches difficult or even impossible. Also, the quality of results from automatic approaches might be questionable. Consequently, the integration of the knowledge, intuition and experience of a domain expert can sometimes be indispensable and the interaction of a domain expert with the data would greatly enhance the whole ML pipeline. Hence, interactive machine learning (iML) puts the “human-in-the-algorithmic-loop” to enable what neither a human nor a computer could do on their own.

We define iML-approaches as algorithms in a multi-agent-hybrid system, that can interact with both computational agents and human agents and can optimize their learning behaviour through these interactions [39].

Why should the integration of human intelligence be beneficial? One strength of humans is that they, even little children, can make inferences from little data (zero-shot learning). The greatest strength is that the are able to recognize the context. There is evidence that humans sometimes even outperform ML-algorithms. Humans can provide almost instantaneous interpretations of complex patterns, for example in diagnostic radiologic imaging: A promising technique to fill the semantic gap is to adopt an expert-in-the-loop approach, to integrate the physicians high-level expert knowledge into the retrieval process by acquiring his/her relevance judgments regarding a set of initial retrieval results [40].

Consequently, iML-approaches, by integrating a human-into-the-loop (e.g. a human kernel [41], or the involvement of a human directly into the machine-learning algorithm [39], thereby making use of human cognitive abilities, seems to be a promising approach. iML-approaches can be of particular interest to solve problems in health informatics, where we are lacking big data sets, deal with complex data and/or rare events, where traditional learning algorithms suffer due to insufficient training samples. Here the doctor-in-the-loop can help, where human expertise and long-term experience can assist in solving problems which otherwise would remain NP-hard.

A recent experimental work [42] demonstrates the usefulness on the Traveling Salesman Problem (TSP), which appears in a number of practical problems, e.g., the native folded three-dimensional conformation of a protein in its lowest free energy state; or both 2D and 3D folding processes as a free energy minimization problem belong to a large set of computational problems, assumed to be conditionally intractable [43]. As the TSP is about finding the shortest path through a set of points, it is an intransigent mathematical problem, where many heuristics have been developed in the past to find approximate solutions [44]. There is evidence that the inclusion of a human can be useful in numerous other problems in different application domains, see e.g., [45, 46]. However, for clarification, iML means the integration of a human into the algorithmic loop, i.e., to open the black box approach to a glass box. Other definitions speak also of a human-in-the-loop, but it is what 5In Active Learning such agents are referred to as so-called "oracles" [38]
we would call classic supervised approaches [47], or in a total
different meaning to put the human into physical feedback loops [48].

In such cases the inclusion of a “doctor-into-the-loop” [49]
can play a significant role in support of solving hard problems
(see the examples in the next paragraph), particularly in combi-
nation with a large number of human agents (crowdsourcing).
From the theory of human problem solving it is known that,
for example, medical doctors can often make diagnoses with
great reliability - but without being able to explain their rules
explicitly. Here iML could help to equip algorithms with such
“instinctive” knowledge and learn thereof. The importance of
iML becomes also apparent when the use of automated
solutions due to the incompleteness of ontologies is difficult
[50].

In the following I provide three examples where the human-
in-the-loop is beneficial.

*iML-Example 1: Subspace Clustering*

Clustering is a descriptive task to identify homogeneous
groups of data objects based on the dimensions. Cluster-
ing of large high-dimensional gene expression data sets has
widespread application in -omics [51].

Unfortunately, the underlying structure of these natural data
sets is often fuzzy, and the computational identification of data
clusters generally requires domain expert knowledge about e.g.
cluster number and geometry. The high-dimensionality of data
is a huge problem in health informatics, because of the curse
of dimensionality: with increasing dimensionality the volume
of the space increases so fast that the available data becomes
sparse, hence, it becomes impossible to find reliable clusters;
also the concept of distance becomes less precise as the num-
ber of dimensions grows, since the distance between any two
points in a given data set converges. Last but not least different
clusters might be found in different subspaces, so a global
filtering of attributes is not sufficient. Given that large number
of attributes, it is likely that some attributes are correlated,
therefore clusters might exist in arbitrarily oriented affinity
subspaces. Moreover, high-dimensional data often includes
irrelevant features, which can obscure to find the relevant ones,
thus increases the danger of modeling artifacts (i.e. undesired
outcomes or errors which can be misleading or confusing)
[52]. The integration of a human-in-the-loop can be of help
[53].

*iML-Example 2: Protein Folding*

Proteins\(^6\) are very important for all life sciences. In protein
structure prediction there is much interest in using amino acid
interaction preferences to align (thread) a protein sequence
to a known structural motif. The protein alignment decision
problem (does there exist an alignment (threading) with a
score less than or equal to \(K^*\)) is NP-complete and the related
problem of finding the globally optimal protein threading is
NP-hard. Therefore, no polynomial time algorithm is possible
(unless \(P = NP\)). Consequently the protein folding problem is
NP-complete [55].

Many such problems (still) require an expert-in-the-loop,
e.g., genome annotation, image analysis, knowledge-base pop-
ulation and protein structure. In some cases, humans are
needed in vast quantities (e.g. in cancer research), whereas
in others, we need just a few very specialized experts in
certain fields (e.g., in the case of rare diseases). Crowdsourc-
ing encompasses an emerging collection of approaches for
harnessing such distributed human intelligence. Recently, the
bioinformatics community has begun to apply crowdsourcing
in a variety of contexts, yet few resources are available that
describe how these human-powered systems work and how to
use them effectively in scientific domains. Generally, there are
large-volume micro-tasks and highly difficult mega-tasks [56].

A good example of such an approach is foldit, an experimental

game which takes advantage of crowdsourcing for category
discovery of new protein structures [57]. Crowdsourcing and
collective intelligence (putting many experts-into-the-loop)
would generally offer much potential to foster translational
medicine (bridging biomedical sciences and clinical applica-
tions) by providing platforms upon which interdisciplinary
workforces can communicate and collaborate [58].

*iML-Example 3: \(k\)-anonymization of patient data*

Privacy preserving machine learning is an important issue,
fostered by anonymization, in which a record is released
only if it is indistinguishable from \(k\) other entities in the
data. \(k\)-anonymity is highly dependent on spatial locality
in order to effectively implement the technique in a statistically
robust way, and in high dimensionalities data becomes sparse,
hence, the concept of spatial locality is not easy to define.
Consequently, it becomes difficult to anonymize the data
without an unacceptably high amount of information loss [59].
Consequently, the problem of \(k\)-anonymization is on the one
hand NP-hard, on the other hand the quality of the result
obtained can be measured at the given factors: \(k\)-anonymity
means that attributes are suppressed or generalized until each
row in a database is identical with at least \(k – 1\) other rows
[60] [61]; \(l\)-diversity as extension of the \(k\)-anonymity model
reduces the granularity of data representation by generalization
and suppression so that any given record maps onto at least \(k\)
other records in the data [62]; \(t\)-closeness is a refinement of
\(l\)-diversity by reducing the granularity of a data representation,
and treating the values of an attribute distinctly by taking into
account the distribution of data values for that attribute [63];
and \(delta\)-presence, which links the quality of anonymization
to the risk posed by inadequate anonymization [64]), but not
with regard to the actual security of the data, i.e., the re-
deidentification through an attacker. For this purpose certain
assumptions about the background knowledge of the hypo-
thesetical enemy must be made. With regard to the particular
demographic and cultural clinical environment this is best done
by a human agent. Thus, the problem of (\(k\)-)anonymization
represents a natural application domain for iML.

\(^6\)In my talk I provide an example from protein conformation, i.e. the x-
ray structure of Avian Pancreatic Polypeptide (APP), which is a medium-size
protein of 36 amino acids [54].
Humans are very capable in the explorative learning of patterns from relatively few samples, whilst classic supervised ML needs large sets of data and long processing time. In the biomedical domain often large sets of training data are missing, e.g., with rare diseases or with malfunctions of humans or machines. Moreover, in clinical medicine time is a crucial factor - where a medical doctor needs the results quasi in real-time, or at least in a very short time (less than 5 minutes), for example, in emergency medicine or intensive care. Rare diseases are often life threatening and require a rapid intervention - the lack of much data makes aML-approaches nearly impossible. An example for such a rare disease with only few available data sets is CADASIL (Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy), a disease, which is prevalent in 5 per 100,000 persons and is therefore the most frequent monogenic inherited apoplectic stroke in Germany.

Particularly in the patient admission, human agents have the advantage to perceive the total situation at a glance. This aptitude results from the ability of transfer learning, where knowledge can be transferred from one situation to another situation, in which model parameters, i.e., learned features or contextual knowledge are transferred.

The examples mentioned so far demonstrate that the application of iML-approaches in “real-world” situations sometimes can be advantageous. These examples demonstrate, that human experience can help to reduce a search space of exponential possibilities drastically by heuristic selection of samples, thereby help to solve NP-hard problems efficiently - or at least optimize them acceptably for a human end-user.

We focused in a recent work [42] on the Traveling Salesman Problem (TSP), because it appears in a number of practical problems in health informatics, e.g. the native folded three-dimensional conformation of a protein is its lowest free energy state and both two- and three-dimensional folding processes as a free energy minimization problem belong to a large set of computational problems, assumed to be very hard (conditionally intractable) [43].

The TSP basically is about finding the shortest path through a set of points, returning to the origin. As it is an intransigent mathematical problem, many heuristics have been developed in the past to find approximate solutions [44].

The Traveling Salesman Problem (TSP) is one of the most known and studied Combinatorial Optimization Problems. Problems connected to TSP were mentioned as early as the last eighteenth century [65]. During the past century, TSP has become a traditional example of difficult problems and also a common testing problem for new methodologies and algorithms in Optimization. It has now many variants, solving approaches and applications [66]. For example, it models in computational biology the construction of the evolutionary trees [67] and in genetics - the DNA sequencing [68] - to provide only a few examples.

The problem is an $NP$-hard problem, meaning that there is no polynomial algorithm for solving it to optimality. For a given number of $n$ cities there are $(n - 1)!$ different tours.

In terms of integer linear programming the TSP is formulated as follows [69].

The cities, as the nodes, are in the set $\mathcal{N}$ of numbers $1, \ldots, n$; the edges are $\mathcal{L} = \{(i, j) : i, j \in \mathcal{N}, i \neq j\}$.

There are considered several variables: $x_{ij}$ as in equation (12), the cost between cities $i$ and $j$ denoted with $c_{ij}$.

\[
x_{ij} = \begin{cases} 1 & \text{the path goes from city } i \text{ to city } j \\ 0 & \text{otherwise} \end{cases} \tag{12}
\]

The Traveling Salesman Problem is formulated to optimize, more precisely to minimize the objective function illustrated in equation (13).

\[
\min \sum_{i=1}^{n} \sum_{i \neq j, j=1}^{n} c_{ij} x_{ij} \tag{13}
\]

The TSP constraints follow.

- The first condition, equation (14) is that each node $i$ is visited only once.

\[
\sum_{i \in \mathcal{N}, (i, j) \in \mathcal{L}} x_{ij} + \sum_{j \in \mathcal{N}, (i, j) \in \mathcal{L}} x_{ji} = 2 \tag{14}
\]

- The second condition, equation (15), ensures that no subtours, $\mathcal{S}$ are allowed.

\[
\sum_{i, j \in \mathcal{L}, (i, j) \in \mathcal{S}} x_{ij} \leq |\mathcal{S}| - 1, \forall \mathcal{S} \subset \mathcal{N} : 2 \leq |\mathcal{S}| \leq n - 2 \tag{15}
\]

For the symmetric TSP the condition $c_{ij} = c_{ji}$ holds. For the metric version the triangle inequality holds: $c_{ik} + c_{kj} \geq c_{ij}, \forall i, j, k$ nodes.

We implemented the travelling Snakesman7 in C#, which is also part of the .NET framework. The choice was made because it is supported by the game engine Unity [70].

If you are now smiling, or even skeptical that we use a game for our experiments, I would like to emphasize that Google is making enormous progress through the use of such games, see e.g. [7].

Gamification [71] is very powerful and we also could proof the concept of interactive machine learning with the human-in-the-loop with gamification experiments. Psychological research indicates that human intuition based on distinct behavioral and cognitive strategies that developed evolutionary over millions of years. For improving ML, we need to identify the concrete mechanisms and we argue that this can be done best by observing crowd behaviors and decisions in gamified situations (see also a classic example for the usefulness of games in [72].

VIII. TOWARDS EXPLAINABLE AI

Very interesting was the recent success in mastering the game of go without human knowledge [73].

When Google DeepMind won the (human) Go player exclaimed: "Why did it make this move ...?". The main problem

\[7]https://hci-kdd.org/project/impl
is that the best performing methods are "black boxes" and can not "explain" why they came up with a certain decision.

Unfortunately, always the "no free-lunch"-theorem [74] is true, and also in explainable AI we have a "trade-off between prediction performance and explainability: the most performing models are the least transparent. DARPA had last year initiated a funding initiative with the goal to create a suite of ML techniques that produce explainable models, while at the same time maintaining a high level of performance [75].

At first we should make us aware of what is understandable to a human; understanding is not only recognizing, perceiving and reproducing or simply a "re-presentation" of facts, but the intellectual understanding of the context in which these facts appear. Rather, understanding can be seen as a bridge between perceiving and reasoning. From capturing the context, without doubt an important indicator of intelligence, the current state-of-the-art AI is still many miles away. On the other hand, humans are able to instantaneously capture the context and make very good generalizations from very few data points – at least in the lower dimensions [76]. I present an example of an explanation interface from our own work: A heatmap for visualizing molecule properties; the central view is an interactive table view (1), i.e. the columns are molecules grouped by similarities using an hierarchical cluster algorithm. The dendrogram (2) shows the grouping, that means the end user can decide which molecules form a group. The rows are characteristics of the molecules, e.g. efficacy, chemical and other numerical measurements. The values per molecule are color-coded, so one can quickly see how groups differ. Exploration is done through interaction and drill-down, among other things to get structure explanations. In 3 we see the settings to configure the heatmap; between the left sidebar and the heatmap we find the legend for the values (4), on the right the names of the molecules (5); the controls (6) and a tree map (7) [51].

IX. Future Outlook

One possibility of how we might bridge the gap between artificial inference and human understanding is a combination of deep learning technologies with ontological approaches [77]. A good current example is Deep Tensor [78], which is a deep neural network, suited for data sets with meaningful graph-like properties and it is beneficial for us that the domains of biology, chemistry, medicine, drug design, etc. offer many such data sets (see for an overview [79]). Here the interactions between various entities (mutations, genes, drugs, disease, etc.) can be encoded via graphs. If we now consider a Deep Tensor network that learns to identify biological interaction paths that lead to a certain disease, we would be able now to automatically identify and to make understandable the influence factors that significantly influenced the classification results. These influence factors can further be used to filter a knowledge graph [80] constructed from publicly available medical research corpora (large ontologies). In addition, the resulting interaction paths are further constrained by known logical limitations of the domain (in this example: Biology). As a result, the classification is presented, thus can be made re-traceable, hence be explained by an human expert as an annotated interaction path, with annotations on each edge linking to specific medical texts that provide supporting evidence [81].

A framework for unsupervised learning of a hierarchical reconﬁgurable image template was presented by [82]. This AND-OR Template (AOT) for visual objects shows three interesting elements: 1) a hierarchical composition as AND nodes, 2) the deformation and articulation of parts as geometric OR nodes, and 3) multiple ways of composition as structural OR nodes. The terminal nodes are hybrid image templates (HIT) [83] that are fully generative to the pixels. Both structures and parameters of this model can be learned unsupervised from images using an information projection principle; which is an awesome technique known for a long time [84], [85]. The learning algorithm itself consists of two steps: 1) a recursive block pursuit procedure to learn the hierarchical dictionary of the primitives, and 2) a graph compression procedure to minimize the model structure for generalizability.

A good example is the hierarchical generative model presented by Lin et al. (2009), where objects are broken into their constituent parts (Yann LeCun always underlies in his talks that "our world is compositional" which is also often used by Alan Yuille, Jason Eisner, Stuart A Geman) and the variability of configurations and relationships between these parts are modeled by stochastic attribute graph grammars. These are embedded in an AND-OR graph for each compositional object category. It combines the power of a stochastic context free grammar to express the variability of part configurations and a Markov random field represents the pictorial spatial relationships between these parts. As a generative model, different object instances of a category can be realized as a traversal through the AND-OR graph in order to get a valid configuration. The inference is connected to the structure of the model and follows a probabilistic formulation consisting of bottom-up detection steps for the parts, which in turn recursively activate the grammar rules for top-down verification and searches for missing parts [86].

Coming to the conclusion, I want to emphasize that computational approaches can find patterns in arbitrarily high-dimensional spaces what no human would be able to see. Consequently, we need an augmentation of human intelligence with artificial intelligence - but also vice versa. In particular situations of problem solving to date only human experts are able to understand the context. Therefore we need solutions for effective mapping of results from high dimensional spaces into the lower dimensions to make it not only perceivable and manipulative to humans, but the raising challenge is that the best performing methods are the least transparent. Our best methods are not re-traceable, thus not understandable, hence it is not possible to explain why a decision has been made. However, current trends in privacy make transparent "glass box" solutions mandatory.

8 the image can be seen here: https://gi.de/informatiklexikon/explainable-ai-ex-ai
If you ask me now what the most interesting topics towards reaching context adaptivity are, I would recommend three major directions:

1) Multi-Task Learning to help to reduce catastrophic forgetting 2) Transfer learning, which is is not easy: learning to perform a task by exploiting knowledge acquired when solving previous tasks: A solution to this problem would have major impact to AI research generally and ML specifically! 3) Multi-Agent Hybrid Systems making use of collective intelligence and crowd-sourcing by integrating a human-in-the-loop, and fostering client side machine learning (federated learning) to ensure privacy, data protection, safety and security.

I would like to close this talk with a citation attributed to Albert Einstein (which surely is not from Albert Einstein):

"Computers are incredibly fast, accurate and stupid, humans are incredibly slow, inaccurate and brilliant, together they are powerful beyond imagination."

Thank you very much!

ABOUT THE KEYNOTE SPEAKER

Andreas Holzinger is lead of the Holzinger Group, HCI-KDD, Institute for Medical Informatics, Statistics at the Medical University Graz, and Associate Professor of Applied Computer Science at the Faculty of Computer Science and Biomedical Engineering at Graz University of Technology. He serves as consultant for the Canadian, US, UK, Swiss, French, Italian and Dutch governments, for the German Excellence Initiative, and as national expert in the European Commission. Andreas obtained a Ph.D. in Cognitive Science from Graz University in 1998 and his Habilitation in Computer Science from TU Graz in 2003. Andreas was Visiting Professor for Machine Learning & Knowledge Extraction in Verona, RWTH Aachen, University College London and Middlesex University London. Since 2016 Andreas is Visiting Professor for Machine Learning in Health Informatics at the Faculty of Informatics at Vienna University of Technology. He founded the Network HCI-KDD to foster a synergistic combination of methodologies of two areas that offer ideal conditions toward unraveling problems in understanding intelligence: Human-Computer Interaction (HCI) & Knowledge Discovery/Data Mining (KDD), with the goal of supporting human intelligence with artificial intelligence. Andreas is Associate Editor of Springer/Nature Knowledge and Information Systems (KAIS), Section Editor for Machine Learning of Springer/Nature BMC Medical Informatics and Decision Making (MIDM), and Editor-in-Chief of Machine Learning & Knowledge Extraction (MAKE). He is organizer of the IFIP Cross-Domain Conference Machine Learning & Knowledge Extraction (CD-MAKE) and Austrian representative for Artificial Intelligence in the IFIP TC 12 and member of IFIP WG 12.9 Computational Intelligence, the ACM, IEEE, GI, the Austrian Computer Science and the Association for the Advancement of Artificial Intelligence (AAAI). Since 2003 Andreas has participated in leading positions in 30+ R&D multi-national projects, budget 4+ MEUR, 300+ publications, 9600+ citations, h-Index = 45.

REFERENCES


[9] Sangkyun Lee and Andreas Holzinger. Knowledge discovery from complex high dimensional data. In Stefan


[34] Thomas L Griffiths, Chris Lucas, Joseph Williams, and Michael L Kalish. Modeling human function learning with gaussian processes. In Daphne Koller, Dale Schu...


[77] Andreas Holzinger, Peter Kieseberg, Edgar Weippl, and A Min Tjoa. Current advances, trends and challenges of machine learning and knowledge extraction: From


