

ICCA12

MS on Geometric Calculi and Deep Learning

Theoretical Resources for Deep Learning

Sebastià Xambó-Descamps

UPC/BSC

2020.8.04



Eduardo U. Moya

Ulises Cortés

Darío García-Gasulla

Ferran Parés

Sergio Álvarez

Armand Vilalta

Luis Oliva

Barcelona Supercomputing Center. Researchers at the HPAI unit.

Zeitgeist

Buzzwords and Prophecies From Life 3.0 Abstract

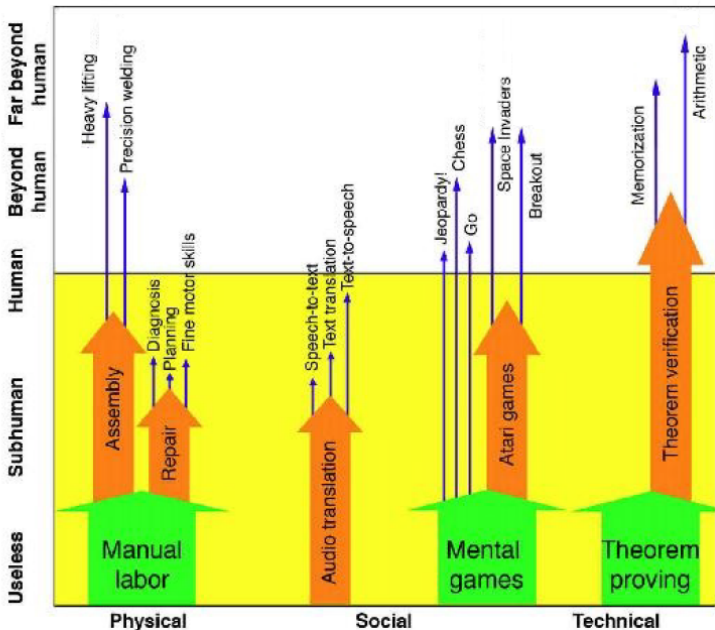
- *Lifelong learning* (a need), *learning to learn* (a mood), *continuous learning* (a pedagogical principle), ...
- *The ability to learn is arguably the most fascinating aspect of general intelligence* (tegmark-2017 [1], *Life 3.0*, page 71).
- *On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines* (hawkins-blakeslee-2004 [2])
- *How to create a mind: The secret of human thought revealed* (kurzweil-2012 [3])
- *The AI Spring of 2018* (olhede-wolfe-2018 [4]; racing for AI dominance).
- *AI superpowers: China, Silicon Valley, and the new world order* (lee-2018 [5])
- *Traffic Signs for AI* (garciagasulla-atiacortes-ulises-cortes-2020 [6]).

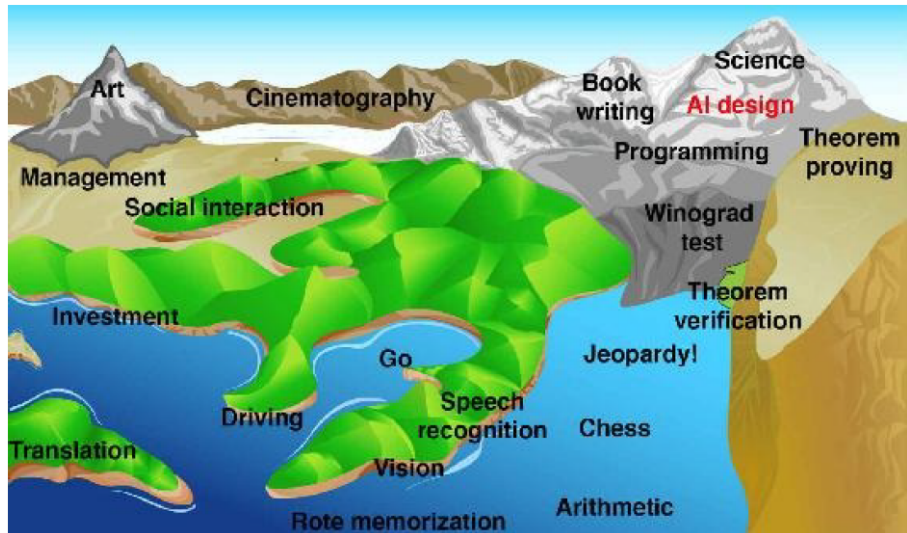
artificial intelligence (AI): Non-biological intelligence.

intelligence: Ability to accomplish complex goals.

narrow intelligence: Ability to accomplish a narrow set of goals.

SKILL LEVEL



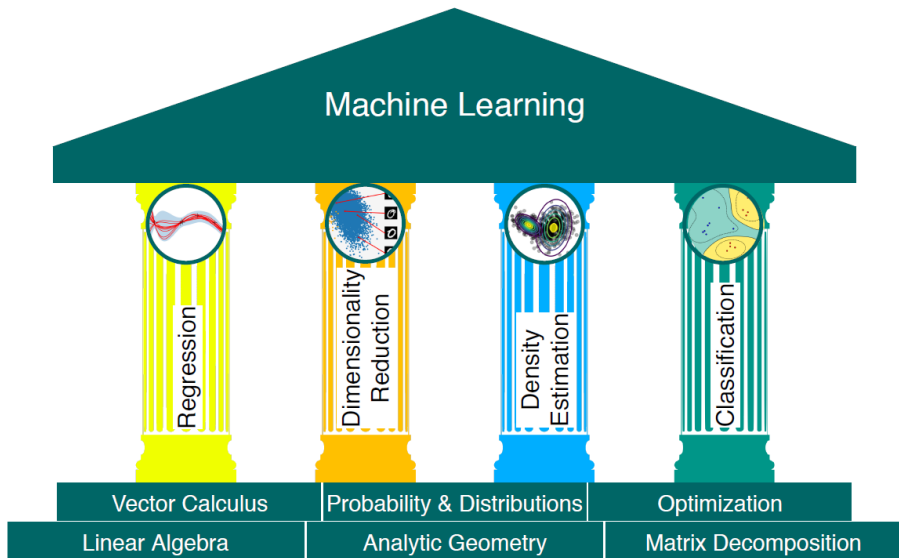


- brown-sandholm-2019 [7] (*Superhuman AI for multiplayer poker*)

Abstract

First, the frameworks that provide **theoretical support** to the main flavors of **automatic learning** (AL = ML) will be sketched. Then the focus will turn to **algebra-geometric neural networks** (their nature depends on the kind of inputs-outputs processed by their **neurons**) and the extensions of those frameworks to this kind of nets. The talk will finish by pointing out some challenges and opportunities for further research.

- **DL** \subset **ML** = **AL** \subset **AI**
- **ML**: “aims to understand computational mechanisms by which experience can lead to improved performance. [...] draws on ideas from many other fields, including , **cognitive psychology**, **information theory**, **logic**, **complexity theory**, and **operations research**, but always with the goal of understanding the computational character of learning” (dietterich-langley-2003 [8]).



“The foundations and four pillars of machine learning” (Figure 1.1 in *Mathematics for machine learning*, deisenroth-faisal-soon-2020 [9])

- **Everlasting jewels**
- **Inductive learning**
- **Bayesian technologies**
- **Neurons**
- **Neuron Nets**
- **Outlooks**

Neglect of mathematics works injury to all knowledge, since he who is ignorant of it cannot know the other sciences or the things of the world (Roger Bacon, 1214-1292).

This presentation can be downloaded from

<https://web.mat.upc.edu/sebastia.xambo/icca12/s-icca12.pdf>

Everlasting jewels

Bayes-Laplace formula

Pattern Theory

Principal Component Analysis

Singular Vector Decomposition

Let X, Y be events. Then

$$P(X, Y) = \begin{cases} P(X) \cdot P(Y|X) \\ P(Y) \cdot P(X|Y) \end{cases}$$

This is equivalent to

$$\frac{P(X|Y)}{P(X)} = \frac{P(Y|X)}{P(Y)} = \frac{P(X, Y)}{P(X) \cdot P(Y)}.$$

This quantity, which is symmetrical in X and Y , will be denoted $L(Y|X)$, and so

$$P(Y|X) = P(Y)L(Y|X),$$

which is the **Bayes-Laplace rule**. We see that $L(Y|X)$ is the factor that scales the **prior** probability $P(Y)$ of Y to the **posterior** probability $P(Y|X)$ (the probability of Y on having observed the occurrence of X). In other words, $L(Y|X)$ measures the **learning** about Y acquired by the **evidence** that X has occurred.

Readings for distressed times: mcgrayne-2011 [10] (*The theory that would not die*). **Modern Bayes revival: Alan Turing!** (and all that followed).

If an event X occurs, and it can be assigned to disjoint hypothesis Y_1, \dots, Y_r , the MAP rule selects the hypothesis Y_j such that $P(Y_j|X)$ is maximum. The Bayes-Laplace formula tells us that this is the same as selecting the Y_j such that $P(X|Y_j)P(Y_j)$ is maximum.

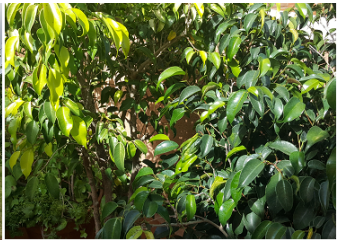
In the special case in which the Y_j have the same probability, this amounts to select the Y_j such that $P(X|Y_j)$ is maximum.




Readings: stone-1975 [11] (*Theory of optimal search*) [1966, Palomares, lost bomb; 1968, Scorpion sub lost; ...]

mumford-desolneux-2010 [12] (*Pattern theory: the stochastic analysis of real-world signals*). Motto: Using Pattern Theory to create mathematical structures both in the natural and the man-made world (Ulf **Grenander**, 1923-2016).

silver-2012 [13] (*The signal and the noise*)

“Bayes’s theorem [...] implies that we must think differently about our ideas — and how to test them. We must become more comfortable with probability and uncertainty. We must think more carefully about the assumptions and beliefs that we bring to a problem” (page 15).



<p>Myth: Superintelligence by 2100 is inevitable</p> <p>Myth: Superintelligence by 2100 is impossible</p> <p>Myth: Only Luddites worry about AI</p> <p>Mythical worry: AI turning evil</p> <p>Mythical worry: AI turning conscious</p>	<table border="1"> <tr><td>Mon</td><td>Tue</td><td>Wed</td><td>Thu</td><td>Fri</td><td>Sat</td><td>Sun</td></tr> <tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td></tr> <tr><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td><td>13</td><td>14</td></tr> <tr><td>15</td><td>16</td><td>17</td><td>18</td><td>19</td><td>20</td><td>21</td></tr> <tr><td>22</td><td>23</td><td>24</td><td>25</td><td>26</td><td>27</td><td>28</td></tr> <tr><td>29</td><td>30</td><td>31</td><td></td><td></td><td></td><td></td></tr> </table> <p>Fact: It may happen in decades, centuries or never. AI experts disagree & we simply don't know</p> <p>Fact: Many top AI researchers are concerned</p> <p>Actual worry: AI turning competent, with goals misaligned with ours</p>	Mon	Tue	Wed	Thu	Fri	Sat	Sun	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31					  
Mon	Tue	Wed	Thu	Fri	Sat	Sun																																						
1	2	3	4	5	6	7																																						
8	9	10	11	12	13	14																																						
15	16	17	18	19	20	21																																						
22	23	24	25	26	27	28																																						
29	30	31																																										



- Statistics is concerned with estimating the parameters θ (*causes*) of the probability distribution that governs the generation of observations x (*effects*).

In generating data, we have the conditional probability, $P(x|\theta)$, where θ is fixed, and usually unknown. On the other hand, if we have data, then we can regard $P(x|\theta)$ as a function of θ , which is usually expressed by a *likelihood* function $L(\theta|x)$, which leads to the principle of *likelihood maximization*: To find the θ that maximizes $L(\theta|x) = P(x|\theta)$ given a set of observations x .

A Bayesian statistical model is made of a parametric statistical model, $f(x|\theta)$, and a prior distribution on the parameters, $\pi(\theta)$.

Probability distributions: robert-2007 [14] (*The Bayesian choice*, App. A).

[...] the range of possible applications of statistics is enormously widened so that we can deal with phenomena other than those of a repeatable nature (from D. V. Lindly's Foreword to de Finetti's landmark book [15]).

Let X a data matrix of size $m \times n$. We regard the rows X^i of X as *observations* on n objects, $X^i = (x_1^i, \dots, x_n^i)$, for m *features* ($i = 1, \dots, m$). Denote the mean value of X^i , $E(X^i)$, by μ^i .

Let $\sigma_{ij} = \text{Cov}(X^i, X^j) = E[(X^i - \mu^i)(X^j - \mu^j)] = E[X^i X^j] - \mu^i \mu^j$ and $\Sigma = (\sigma_{ij})_{1 \leq i, j \leq n}$. This is the *covariance matrix* of X , $\Sigma = \text{Cov}(X)$. Notice that $\sigma_{ii} = \sigma_i^2$, where σ_i is the *standard deviation* of X^i .

Given a unit m -vector u , it turns out that $\text{Var}(u^T X) = u^T \Sigma u$, and that this is maximum precisely when u is an *eigenvector of Σ with the highest eigenvalue*. This vector is the *principal component* of X , that is, the unit eigenvector $u = u_1$ of Σ whose eigenvalue λ_1 is largest (the eigenvalues of Σ are real). It accounts for the greatest variance of the data along a direction.

The second principal component of X is the eigenvector u_2 corresponding to the second eigenvalue λ_2 . It maximizes $\text{Var}(u^T X) = u^T \Sigma u$ for unit vectors u perpendicular to u_1 . And so on.

This frames an (unsupervised) approach to *dimension reduction* by means of the spectral decomposition $\Sigma = U \Lambda U^T$, where Λ is the diagonal matrix with the eigenvalues of Σ , ordered in non-increasing order, and U is the orthonormal matrix formed with the unit eigenvectors of Σ .

Let X be an $m \times n$ data matrix as in the preceding slide, and let r be its rank. Then XX^T and $X^T X$ have rank r and they have the same non-zero eigenvalues $\lambda_1^2, \dots, \lambda_r^2$, where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$. Moreover, if we let U and V denote the orthonormal matrices of eigenvectors of XX^T and $X^T X$, then $X = U\Lambda V^T$ where $\Lambda_{jj} = \lambda_j$ for $j = 1, \dots, r$ are the only non-zero values of Λ .

Note that $XX^T = U(\Lambda\Lambda^T)U^T$ and $X^T X = V(\Lambda^T\Lambda)V^T$, where the first r values of the diagonals of $\Lambda\Lambda^T$ and $\Lambda^T\Lambda$ are $\lambda_1^2, \dots, \lambda_r^2$ and all others 0 in both matrices (of sizes $m \times m$ and $n \times n$, respectively).

Since $U\Lambda = (\lambda_1 u_1, \dots, \lambda_r u_r)$, we get the *singular value decomposition* of X :

$$X = \lambda_1 u_1 v_1^T + \dots + \lambda_r u_r v_r^T.$$

Actually it turns out ([Eckart-Young theorem](#)) that for $k = 1, \dots, r$ the matrix

$$M_k = \lambda_1 u_1 v_1^T + \dots + \lambda_k u_k v_k^T$$

is the closest to X among the matrices of rank k .

Remark. The least-squares solution to $Xa = b$ is $a = X^\dagger b$, where $X^\dagger = V\Sigma^\dagger U^T$ (Moore-Penrose pseudo-inverse of X), $\Sigma^\dagger = \text{diag}(\lambda_1^{-1}, \dots, \lambda_r^{-1}, 0, \dots, 0)$.

Readings:

- *Latent semantic analysis*, particularly Ch. 2 (landauer-mcnamara-dennis-kintsch-2007 [16])
- *Image processing, analysis, and machine vision* (sonka-hlavac-boyle-2015 [17])
- *Linear algebra and learning from data* (strang-2019 [18])
- *Introduction to machine learning* (alpaydin-2020 [19])
- *Mathematics for ML* (deisenroth-faisal-soon-2020 [9])
- <https://en.wikipedia.org/wiki/Eigenface> (inspected August 1st, 2020)

Inductive Learning

Ingredients

The learning problem

Fundamental theorem of IL

After the LN of Joan Bruna at the MSRI, 2019

In general terms, a rough idea of *supervised machine learning* is to produce algorithms that output a function that

- **F** Gives good approximations of given values y^i for a set of given inputs x^i ($i = 1, \dots, N$);
- **G** Has good *generalization capacity*, which means that for any x (of a kind similar to that of the x^i) the value $y' = f(x)$ is a good approximation of the expected value y corresponding to x .

Data source: Data are drawn from a space \mathcal{X} , which may be assumed to be a subset of \mathbf{R}^N . Usually the dimension of \mathcal{X} is very large.

Data generation: Elements of \mathcal{X} are drawn according to a (usually unknown) probability density P . In symbols, $x \sim P$.

Hypotheses: A family $\mathcal{F} = \{f_w\}_{w \in W}$ of functions $f_w : \mathcal{X} \rightarrow \mathcal{Y}$ (*Inductive bias*). The elements of W are called *parameters* or *weights*. **Example:** For a polynomial, we can take its coefficients as parameters.

Complexity: The *complexity* of hypotheses is gauged by a map $\gamma : \mathcal{F} \rightarrow \mathbf{R}^+$. A common choice is the norm $\|f_w\|$.

For $\delta \in \mathbf{R}^+$, we set $\mathcal{F}_\delta = \{f \in \mathcal{F} : \gamma(f) \leq \delta\}$ (the set of hypotheses with complexity bounded by δ).

For a polynomial, it could be bounding the absolute value of its coefficients.

Expert or supervisor: A fixed map $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ (not necessarily in \mathcal{F}). For each $x \in \mathcal{X}$, it produces an *example*: the pair (x, y) , $y = f^*(x)$.

Training dataset: $\mathcal{D} = \{(x_i, y_i = f^*(x_i)) : x_i \in \mathcal{X}\}_{i=1, \dots, m}$, where $x_i \in \mathcal{X}$ are drawn according to P independently, $x_i \sim P$ in symbols.

Loss: The closeness of $y, y' \in \mathcal{Y}$ is given by a *loss function* $\ell(y, y') \geq 0$.

If $\mathcal{Y} = \mathbf{R}$ (*regression learning*), use $\ell(y, y') = |y - y'|^2$ or $\ell(y, y') = |y - y'|$.

If \mathcal{Y} is finite (*classification learning*), the natural loss is $\ell(y, y') = \delta(y, y')$, which is 1 when $y = y'$ and 0 otherwise.

The *loss* of $f \in \mathcal{F}$, denoted $L(f)$, it is a measure of how close the values $f(x)$ and $f^*(x)$ are on the average:

$$L(f) = E_P[\ell(f(x), f^*(x))] = \int_{\mathcal{X}} \ell(f(x), f^*(x)) P(x) dx.$$

In classification, $L(f) = P(f(x) \neq f^*(x))$.

The loss is also called *risc* or *error* in some contexts.

Goal of a LA: To find $f \in \mathcal{F}$, using only \mathcal{D} , such that $L(f)$ is a good approximation of the *minimum loss* $L_{\mathcal{F}} = \min_{h \in \mathcal{F}} L(h)$ achievable with \mathcal{F} .

Empirical risk: For $h \in \mathcal{F}$, the quantity $R_{\mathcal{D}}(h) = \frac{1}{m} \sum_{i=1}^m \ell(h(x_i), y_i)$.

Empirical risk minimization: Let $R_{\mathcal{D}, \mathcal{F}} = \min_{h \in \mathcal{F}} R_{\mathcal{D}}(h)$, i.e. the minimum of the empirical risks achievable with \mathcal{F} .

Variant: $R_{\mathcal{D}, \mathcal{F}, \lambda} = \min_{h \in \mathcal{F}} R_{\mathcal{D}}(h) + \lambda \gamma(h)$, $\lambda > 0$ fixed, is the *λ -regularized*, or *λ -penalized*, empirical risk minimum.

Problem: to find $\hat{f} \in \mathcal{F}$ such that $R_{\mathcal{D}}(\hat{f}) = R_{\mathcal{D}, \mathcal{F}}$, and $\hat{f}_{\delta} \in \mathcal{F}_{\delta}$ such that $R_{\mathcal{D}}(\hat{f}_{\delta}) = R_{\mathcal{D}, \mathcal{F}_{\delta}}$. With similar notations for the penalized versions.

Approximation error: Defined as the difference $A_{\delta} = L_{\mathcal{F}_{\delta}} - L_{\mathcal{F}}$. It is non-negative and decreases on increasing δ . It measures how close we can get to the minimum loss $L_{\mathcal{F}}$ with functions from \mathcal{F}_{δ} .

Statistical error: Given $f \in \mathcal{F}_{\delta}$, $|L(f) - R_{\mathcal{D}}(f)|$ is the error undergone on replacing the loss $L(f)$ by the empirical loss $R_{\mathcal{D}}(f)$. It is clearly bounded above by $S_{\mathcal{D}, \delta} = \sup_{h \in \mathcal{F}_{\delta}} |L(h) - R_{\mathcal{D}}(h)|$, which will be called *statistical error*.

- Let $\hat{f} \in \mathcal{F}_\delta$ be such that $R_{\mathcal{D}}(\hat{f}) \leq \epsilon + R_{\mathcal{D}, \mathcal{F}_\delta}$. This means that the empirical risk of \hat{f} differs in not more than ϵ from the minimum $R_{\mathcal{D}, \mathcal{F}_\delta}$ of the empirical risks of functions of \mathcal{F}_δ (ϵ is called *optimization error*). Then

$$L(\hat{f}) - L_{\mathcal{F}} \leq A_\delta + 2S_{\mathcal{D}, \delta} + \epsilon.$$

The approximation, statistical and optimization errors give an upper bound to the error produced on replacing the minimum loss $L_{\mathcal{F}}$ by the empirical loss $L(\hat{f})$.

Tradeoffs. (1) Decreasing ϵ may entail that we have to increase δ in order to guarantee that \hat{f} exists. (2) On increasing δ , A_δ decreases (or at least does not increase), but $S_{\mathcal{D}, \delta}$ increases. (3) In general, the statistical error decreases on increasing the dataset.

Let w be an unknown vector of **weights** (one weight for each of the n components of the x vectors) and $f_w(x) = w \cdot x = w_1x_1 + \dots + w_nx_n$ (a **weighted sum** of the components of x). Let $\mathcal{F} = \{f_w\}$.

A way of fulfilling condition **F** is to pick a w that achieves $\min_w \sum_{i=1}^N (w \cdot x^i - y^i)^2$ (**least squares** optimization). This can be obtained by standard procedures.

Regularized linear regression (improves generalization capacity):

$$\min_w \sum_i (w \cdot x^i - y^i)^2 + \lambda \|w\|_2^2$$

“where λ is a scalar ... discovered by experimenting with the data” (arora-2018 [20]).

This is related to the phenomena of **overfitting** and **underfitting** while learning.

The general steps followed by a supervised learning algorithm are as follows:

Givens: \mathcal{D} , $\mathcal{F} = \{f_w\}_{w \in W}$, a loss function ℓ .

1. Split \mathcal{D} in two sets: \mathcal{D}' (*training dataset*) and \mathcal{D}'' (*testing dataset*).
2. Find $w \in W$ such that the empirical risk of f_w on \mathcal{D}' , $R_{\mathcal{D}'}(f_w)$, is minimum (this is usually accomplished by an optimization iterative procedure, and each step in the loop is called a *training epoch*).
3. Return f_w together with *accuracy measures* of how often $f_w(x^j) \simeq y^j$ for the training and testing sets.

- *The Nature of Statistical Learning Theory*, summarized in vapnik-1999 [21] (vapnik-1995 [?])
- *Learning theory: an approximation theory viewpoint* (cucker-zhou-2007 [22])
- *An elementary introduction to statistical learning theory*, summarized in [23] (kulkarni-harman-2011-SLT [24])
- *Optimization for ML*, especially Chapter 13, *The tradeoffs of large-scale learning*, by L. Bottou and O. Bousquet (sra-nowozin-wright-2012 [25]).
- *Mathematical foundations of supervised learning* (wolf-2018 [26])
- *Mathematics for machine learning*: Ch. 9, Regression; Ch. 10, Dimensionality reduction; Ch. 12, Classification; Ch. 11, Density estimation. (deisenroth-faisal-soon-2020 [9])
- *Foundations of Data Science* (blum-hopcroft-kannan-2020 [27])

Bayesian technologies

Gibbs lemma and the KL divergence

Sources, with emphasis on causal learning-reasoning

Suppose $p = p_1, \dots, p_n$ and $q = q_1, \dots, q_n$ are probability distributions. Then $-\sum_j p_j \log_2(p_j) \leq -\sum_j p_j \log_2(q_j)$, with equality if and only if $q = p$.

It follows that $\sum_j p_j \log_2(p_j/q_j) \geq 0$, with equality precisely when $q = p$.

This expression is usually called the *Kullback-Leibler divergence* (KL) of p and q , and denoted $KL(p, q)$.

To recap, $KL(p, q) \geq 0$ with equality if and only if $p = q$. Note, however, that in general $KL(q, p) \neq KL(p, q)$.

The KL divergence is an important tool in the theory of Bayesian networks for comparing the network probability distributions at successive times.

Since $H(p) = -\sum_j p_j \log_2(p_j)$ is the *entropy* of the distribution p , which is the average information provided by a trial of p , it is only natural that KL is also significant in information theory.

- *Probabilistic reasoning in intelligent systems: Networks of plausible inference* (pearl-1988 [28])
- *Perception as Bayesian Inference* (knill-richards-1996 [29])
- *The art and science of cause and effect* (pearl-1996 [30])
- *Probability theory: The logic of science* (jaynes-2003 [31])
- *Learning Bayesian networks* (neapolitan-2004 [32])
- *Data analysis: a Bayesian tutorial* (sivia-skilling-2006 [33])
- *Pattern recognition and machine learning*, Ch. 5 (bishop-2006 [34])
- *Causality. Models, Reasoning, and Inference* (pearl-2009 [35])
- *Probabilistic Graphical Models—Principles and Techniques* (koller-friedman-2009 [36]*)
- *Modeling and Reasoning with Bayesian Networks* (darwiche-2009 [37]*)
- *Learning Hidden Markov Models using Non-Negative Matrix Factorization.* (cybenko-crespi-2011 [38]). A nice application of **SVD**.

- *Markov random fields for vision and image processing* (blake-kohli-rother-2011 [39])
- *Causality and Statistical Learning*, a review (gelman-2011 [40])

“All becomes more difficult when we shift our focus from **What** to **What-if** and **Why**”

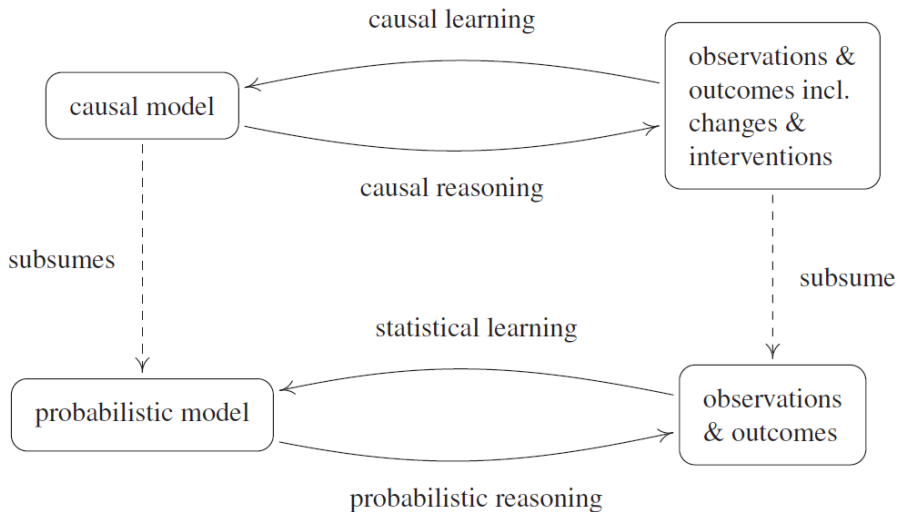
Consider two broad classes of inferential questions:

1. *Forward causal inference*. What might happen if we do X? What are the effects of smoking on health, the effects of schooling on knowledge, the effect of campaigns on election outcomes, and so forth?
2. *Reverse causal inference*. What causes Y? Why do more attractive people earn more money, why do many poor people vote for Republicans and rich people vote for Democrats, why did the economy collapse?

- *Bayesian reasoning and machine learning* (barber-2012 [41])
- *Causes of effects and effects of causes* (pearl-2015 [42])
- *Trygve Haavelmo and the emergence of causal calculus* (pearl-2015-Haavelmo [43])
- *Efficient Algorithms for Bayesian Network Parameter Learning from Incomplete Data* (vandenbroeck-mohan-choi-pearl-2015 [44])

“In contrast to textbook approaches such as EM and the gradient method, our approach is non-iterative, yields closed form parameter estimates, and eliminates the need for inference in a Bayesian network.”

- *Causal inference and the data-fusion problem* (bareinboim-pearl-2016 [45])
- *Elements of causal inference. Foundations of learning algorithms* (peters-janzing-sholkopf-2017 [46]*)



- *Elements of causal inference. Foundations of learning algorithms* (peters-janzing-sholkopf-2017 [46]*)

- *Theoretical impediments to machine learning with seven sparks from the causal revolution* (pearl-2018 [47])
- *The book of why: The new science of cause and effect* (pearl-mackenzie-2018 [48])
- *Probability Theory And Statistical Inference: Empirical Modeling With Observational Data*; for a shorter account, see [49] (spanos-2019 [50])
- *Graphical models for processing missing data* (mohan-pearl-2019 [51])
- *Markov blankets in the brain* (hipolito-ramstead-convertino-bhat-friston-parr-2020 [52])

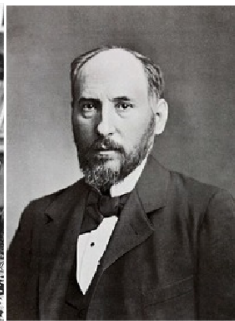
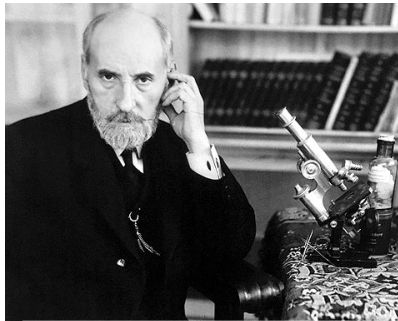
“ ‘Markov blanket’: a statistical boundary that mediates the interactions between what is inside of and outside of a system.”

Neurons

Biological neurons (in homage to S. Ramon y Cajal)

Artificial neurons

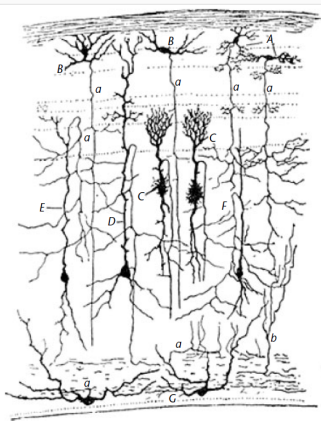
A-neurons



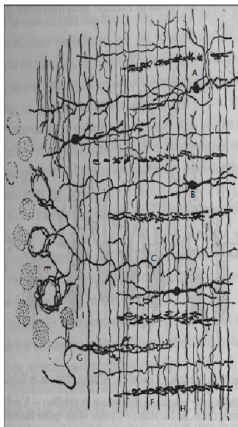
Santiago Ramón y Cajal (1882-1934). Nobel Prize of Physiology and Anatomy (1906, shared with Camillo Golgi) for his discoveries about the structure of the nervous system and the role of the neuron.

In 1887, he moved to **Barcelona** to occupy the chair of Histology created at the Faculty of Medicine of the University of Barcelona.¹³ It was in **1888**, defined by Cajal himself as his '**peak year**', when he discovered the mechanisms that govern the morphology and connective processes of gray matter nerve cells of the cerebrospinal nervous system.

- Every man can be, if he wants to, a sculptor of his own brain.
- Nothing inspires me more reverence and awe than an old man who is willing to change his mind.
- The car of Spanish culture lacks the wheel of science.



Santiago Ramón y Cajal: different types of neurons in the optic tectum of a bird (Cajal Institute, CSIC)



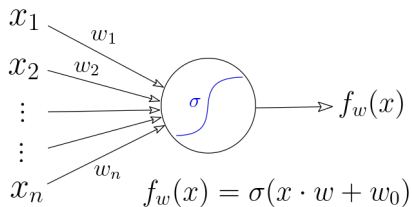
Cajal: circuitry of the cerebellum. The cell F is the dendrite of a Purkinje cell.



Cajal: Single Purkinje cell

- Hundreds of his drawings illustrating the delicate arborizations of brain cells are still in use for educational purposes.
- He conjectured that learning is related to variations of the synaptic connections.

- *A logical calculus of the ideas immanent in nervous activity* (mcculloch-pitts-1943 [53])
 - *The perceptron: a probabilistic model for information storage and organization in the brain* (rosenblatt-1958 [54])
 - *Adaptive switching circuits* (widrow-1960 [55])
 - *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms* (rosenblatt-1962 [56])
 - *Perceptrons* (minsky-papert-1969 [57])
 - *Learning representations by back-propagating errors* (rumelhart-hinton-williams-1986 [58])
-
- *Encyclopedia of Cognitive Science* (nadel-2003 [59])



Ordinarily, the quantities x and w are real numbers, but instead we can envision various natural generalizations.

The components of x and w could belong to a given algebra \mathcal{A} , as the expression $x \cdot w = x_1 w_1 + \dots + x_n w_n \in \mathcal{A}$ is well defined.

For this to work, we need an *activation function* $\sigma : \mathcal{A} \rightarrow \mathcal{A}$, which in principle (an in practice) can be implemented by applying an ordinary $\sigma : \mathbf{R} \rightarrow \mathbf{R}$ ‘component-wise’ to elements of \mathcal{A} .

For example, beyond the real numbers \mathbf{R} , \mathcal{A} could be \mathbf{C} (complex numbers), \mathbf{H} (quaternions), \mathbf{CH} (commutative quaternions), $2\mathbf{H}$ (biquaternions), \mathbf{O} (octonions), a matrix algebra $\mathbf{R}(n)$, or a geometric algebra $\mathcal{G} = \mathcal{G}_{r,s}$.

Another generalization direction is replacing x and w by more general data structures, as for example \mathcal{A} -arrays, and $x \cdot w$ by a suitable operation $x \star w$.

Among these operations, the most common are *cross-correlations* or *convolutions* (in this case the w are often called *filters*) or *max-pooling* operations. In sum, we arrive at a general notion of a neuron, that we may call \mathcal{A} -neuron, by specifying:

- The algebra \mathcal{A} ;
- The shapes of the \mathcal{A} -arrays x (*input*) and w (*weights* or *filters*);
- The operation \star ;
- The activation function σ .

The shape of the output array x' is determined by the above elements, and the map $f_w : x \mapsto x'$ is the *functional signature* of the \mathcal{A} -neuron.

\mathcal{A} -neurons can learn by modifying w in suitable ways.

Neuron nets

Standard NNs

\mathcal{A} -NNs

R -, C -, Q -, O -, \mathcal{G} -, ... -NNs

A NN can be thought of a **composition of neurons** according to some *architecture* (a **graph of connections**).

Standard NNs are *layered* and their functional signature is like this:

$$\mathcal{N}: \text{Input} \rightarrow L_1 \rightarrow L_2 \rightarrow \cdots \rightarrow L_m \rightarrow \text{Output}$$

- A layer takes an input x and yields an output x' .
- The map $f: x \mapsto x'$ depends on *parameters* (or *weights*) associated to the layer and whose nature depends on the *kind of layer*.
- The input to the first layer is the *signal* to be processed.
- The last layer is the *output layer*, and its output is the *result* produced by the net on the input signal.
- The net is *shallow* if $m = 1$ and *deep* if $m > 1$.

- In practice, x, x' , and the *layer parameters* are **multidimensional arrays** (or *tensors*) whose nature is chosen according to the processing that has to be achieved.

Write $[n_1, n_2, \dots, n_d]$ to denote the type of a d -dimensional (real) array with axis dimensions n_1, \dots, n_d .

Thus $[n]$ is the type of n -dimensional vectors and $[n_1, n_2]$ the type of matrices with n_1 rows and n_2 columns. Matrices are useful to represent monochrome images, but for **RGB** images we need arrays of type $[n_1, n_2, 3]$, or $[n_1, n_2, n_3]$ if it is required that the image be represented by n_3 channels, as for example $n_3 = 6$ for a pair of color stereoscopic images.

The parameters associated to *convolutional* and *fully connected* layers are represented by an *array of weights*, W , and a bias array, b . In these cases, the expression of f has the form

$$f_{W,b}^{\pi}(x) = \sigma(x \star_{\pi} W + b) \quad (1)$$

where \star_{π} is a pairing specific of the layer and σ is an activation function that is applied component-wise to arrays (e.g. $\text{ReLU}(t) = \max(t - \beta, 0)$).

For *convolutional layers*, $\star_{\pi} = \star$ is *array cross-correlation*, while for *fully connected layers*, \star_{π} is *matrix product*, which is denoted by juxtaposition of its factors, xW .

- Given weight arrays and biases W_k and b_k ($k = 1, \dots, m$), the net \mathcal{N} computes the function

$$f = f_{W_m, b_m}^{\pi_m} \circ \dots \circ f_{W_1, b_1}^{\pi_1}$$

that is **continuous** and **pice-wise affine**.

- There exist **training algorithms** of \mathcal{N} , particularly those of *back-propagation* type, achieving trained weights and biases for which f is 'optimal' in the sense of the conditions **F** and **G**.
- Approximation superpositions of a sigmoidal function* (cybenko-1989 [60])

For a **max pooling** layer, the parameters are represented by a triple of positive integers $(w_1, w_2, s = 1)$, where (w_1, w_2) is the shape of the pooling window and s is the stride (1 by default). In this case $\star_{\pi} = \star_{\text{mp}}$ is given by the rule

$$(x \star_{\text{mp}} W)[i, j, k] = \max(x[is : is + w_1 - 1, js : js + w_2 - 1, k]) \quad (2)$$

where we use the standard slicing conventions for arrays. The shape of the array $x \star_{\text{mp}} W$ is $[n'_1, n'_2, n_3]$, where n'_1 and n'_2 are the greatest integers such that $n'_1 \leq (n_1 - w_1) / s$ and $n'_2 \leq (n_2 - w_2) / s$.

In the **cross-correlation product** $y = x \star W$, x is an array of type $[n_1, n_2, n_3]$ and W (the filter) is an array of type $[w_1, w_2, n_3, m_3]$. The pair (n_1, n_2) is the shape of the space dimensions of x and n_3 the number of channels. The pair (w_1, w_2) denotes the window dimensions of the filter and m_3 the number of channels of the array y . The definition is given by the following formula:

$$y[i, j, k] = \sum_{m=0}^{w_1-1} \sum_{n=0}^{w_2-1} \sum_{r=0}^{n_3-1} x[i+m, j+n, r] W[m, n, r, k] \quad (3)$$

which can be expressed more compactly as

$$y[i, j, k] = \sum_{r=0}^{n_3-1} x[i : i + w_1 - 1, j : j + w_2 - 1, r] * W[:, :, r, k] \quad (4)$$

where $*$ denotes the ordinary scalar product of matrices. Notice that the shape of y is $[n_1 - w_1 + 1, n_2 - w_2 + 1, m_3]$.

There is also a **downsampled cross-correlation** $y = x \star_s W$ by a stride s :

$$\begin{aligned}
 y[i, j, l] &= \sum_{k, m, n} x[is + m, js + n, k] W[m, n, k, l] \\
 &= \sum_k x[is : is + w_1 - 1, js : js + w_2 - 1, k] \\
 &\quad * W[:, k, l]
 \end{aligned} \tag{5}$$

The shape of the array $x \star_s W$ is $[n'_1, n'_2, n_3]$, where n'_1 and n'_2 are the greatest integers such that $n'_1 \leq (n_1 - w_1) / s$ and $n'_2 \leq (n_2 - w_2) / s$.

- *Neural networks and learning machines* (haykin-2009 [61])
- *ImageNet classification with deep convolutional neural networks* (krizhevsky-sutskever-hinton-2012 [62])
- *Deep learning* (lecun-bengio-hinton-2015 [63])
- *Neural networks and deep learning* (nielsen-2015 [64])
- *Deep learning in neural networks: an overview* (schmidhuber-2015 [65]; 54 pages of references)
- *Deep learning* (goodfellow-bengio-courville-2016 [66])
- *Understanding deep convolutional networks* (mallat-2016 [67])
- *Neural networks and deep learning* (aggarwal-2018 [68])
- *Universality of deep convolutional neural networks* (zhou-2019 [69])
- *A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects* (li-yang-peng-liu-2020 [70])
- *Deep Learning architectures applied to wind time series multi-step forecasting* (manero-2020 [71], PhD thesis)

- *Complex-valued neural networks: Theories and applications* (hirose-2003 [72])
- *Complex-valued neural networks: The merits and their origins* (hirose-2009 [73])
- *Complex-Valued Neural Networks: Utilizing High-Dimensional Parameters* (nitta-2009 [74])
- *Complex-valued neural networks with multi-valued neurons* (aizenberg-2011 [75])
- *Complex-valued neural networks* (hirose-2012 [76]; second edition of [77]).
- *Complex-valued neural networks: Advances and applications* (Hirose-2013 [78]) An interesting collection of ten papers of which the first four are about C-NN's. The most outstanding is the first, by Hirose (the editor of the volume): *Application fields and fundamental merits of complex-valued neural networks*.
- *A mathematical motivation for complex-valued convolutional networks* (bruna-chintala-lecun-piantino-szlam-tygert-2015 [79])

- *On complex valued convolutional neural networks* (guberman-2016 [80])
- *Complex-valued convolutional neural networks for real-valued image classification* (popa-2017 [81])
- *Deep complex networks* (trabelsi-2017 [82])
- *Evaluation of complex-valued neural networks on real-valued classification tasks* (monning-nils-manandhar-suresh-2018 [83])

- *Quaternionic Gabor filters for local structure classification* (bulow-sommer-1998 [84])
- *Quaternionic spinor MLP* (buchholz-sommer-2000 [85])
- *Quaternion wavelets for image analysis and processing* (chan-choi-baraniuk-2004 [86])
- *Quaternionic neural networks: Fundamental properties and applications* (isokawa-matsui-nishimura-2009 [87])
- *Quaternion atomic function wavelet for applications in image processing* (moya-bayro-2010 [88])
- *Quaternionic multilayer perceptron with local analyticity* (isokawa-nishimura-matsui-2012 [89])
- *Quaternion and Clifford Fourier transforms and wavelets* (hitzer-sangwine-2013 [90])
- *Rotational invariance of quaternionic Hopfield neural networks* (kobayashi-2016 [91])

- *Quaternion neural networks for spoken language understanding* (parcollet-titouan-et-10-2016 [92])
- *Design of quaternion-neural-network-based self-tuning control systems* (takahashi-hasegawa-hashimoto-2017 [93])
- *Quaternion convolutional neural networks for end-to-end automatic speech recognition* (parcollet-et-6-2018 [94])
- *Deep quaternion networks* (gaudet-maida-2018 [95])
- *Quaternion convolutional neural networks* (zhu-xu-xu-chen-2018 [96])
- *Neural ordinary differential equations* (chen-rubanova-bettencourt-duvenaud-2018 [97])
- *Quaternion Equivariant Capsule Networks for 3D Point Clouds* (zhao-birdal-lenssen-menegatti-guibas-tombari-2020 [98])
- *A bio-inspired quaternion local phase CNN layer with contrast invariance and linear sensitivity to rotation angles* (moya-xambo-perez-salazar-mzortega-cortes-2020 [99])

- *Neural networks in the Clifford domain* (pearson-bisset-1994 [100])
- *Geometric computing with Clifford algebras: theoretical foundations and applications in computer vision and robotics* (sommer-2001 [101])
- *Clifford algebra multilayer perceptrons* (buchholz-sommer-2001 [102], a chapter in the preceding reference)
- *The monogenic signal* (felsberg-sommer-2001 [103])
- *Hypercomplex signals – a novel extension of the analytic signal to the multidimensional case* (bulow-sommer-2001 [104])
- *Clifford convolution and pattern matching on vector fields* (ebling-scheurmann-2003 [105])
- *Design of kernels for support multivector machines involving the Clifford geometric product and the conformal geometric neuron* (bayro-arana-vallejo-2003 [106])
- *A theory of neural computation with Clifford algebras* (buchholz-2005 [107], PhD thesis)

- *Clifford Fourier transform on vector fields* (ebling-scheurmann-2005 [108])
- *Medical image segmentation using a self-organizing neural network and Clifford geometric algebra* (rivera-bayro-2006 [109])
- *On Clifford neurons and Clifford multi-layer perceptrons* (buchholz-sommer-2008 [110])
- *Coordinate independent update formulas for versor Clifford neurons* (buchholz-hitzer-tachibana-2008 [111])
- *Clifford support vector machines for classification, regression, and recurrence* (bayro-arana-2010 [112])
- *Geometric computing: for wavelet transforms, robot vision, learning, control and action* (bayro-2010 [113])
- *Geometric algebra computing: in engineering and computer science* (bayro-scheuermann-2010 [114])
- *Clifford algebra based edge detector for color images* (franchini-gentili-sorbello-vassallo-vitabile-2012 [115])
- *The Clifford Fourier transform in real Clifford algebras* (hitzer-2013 [116])

- *A specialized architecture for color image edge detection based on Clifford algebra* (franchini-gentile-vassallo-sorbello-vitabile-2013 [117])
- *Overviews of optimization techniques for geometric estimation* (kanatani-2013 [118])
- *Understanding geometric algebra: Hamilton, Grassmann, and Clifford for computer vision and graphics* (kanatani-2015 [119])
- *A geometric algebra co-processor for color edge detection* (mishra-wilson-wilcock-2015 [120])
- *A conformal geometric algebra based clustering method and its applications* (pham-tachibana-2016 [121])
- *Outlook for Clifford algebra based feature and deep learning AI architectures* (yin-hadjiloucas-zhang-2017 [122])
- *Geometric Algebra Applications Vol. I: Computer Vision, Graphics and Neurocomputing* (bayro-2018 [123])
- *Feature extraction using conformal geometric algebra for AdaBoost algorithm based inplane rotated face detection* (pham-doan-hitzer-2019 [124])

- *GA-ORB: A new efficient feature extraction algorithm for multispectral images based on geometric algebra* (wang-zhang-shi-wang-cao-2019 [125])
- *GA-SURF: A new speeded-up robust feature extraction algorithm for multispectral images based on geometric algebra* (wang-shi-cao-2019 [126])
- *Geometric-algebra adaptive filters* (lopes-lopes-2019 [127])
- *Generalizing convolutional neural networks for equivariance to Lie groups on arbitrary continuous data* (finzi-stanton-izmailov-wilson-2020 [128])

- *AlgebraNets* (hoffmann-schmitt-osindero-simonyan-elsen-2020 [129])

“... our results demonstrate that there are alternative algebras which deliver better parameter and computational efficiency compared with \mathbb{R} . We consider \mathbb{C} , \mathbb{H} , $M_2(\mathbb{R})$, $M_2(\mathbb{C})$, $M_3(\mathbb{R})$, $M_4(\mathbb{R})$, dual numbers and the \mathbb{R}^3 cross product. Additionally, we note that multiplication in these algebras has higher compute density than real multiplication, a useful property in situations with inherently limited parameter reuse such as auto-regressive inference and sparse neural networks. We therefore investigate how to induce sparsity within AlgebraNets. We hope that our strong results on large-scale, practical benchmarks will spur further exploration of these unconventional architectures which challenge the default choice of using real numbers for neural network weights and activations.” (from the Abstract)

- *Deep octonion networks* (wu-xu-wu-kong-senhadji-shu-2020 [130])

“This paper constructs a general framework of deep octonion networks [...] and provides the main building blocks [...] such as octonion convolution, octonion batch normalization and octonion weight initialization [...] used in image classification tasks for CIFAR-10 and CIFAR-100 data sets. [...] have better convergence and higher classification accuracy.” (from the Abstract)

Outlooks

- *Geometric deep learning: going beyond Euclidean data* (bronstein-bruna-lecun-szlam-vandergheynst-2017 [131])

“*Geometric deep learning* is an umbrella term for emerging techniques attempting to generalize (structured) deep neural models to **non-Euclidean domains such as graphs and manifolds**. The purpose of this paper is to overview different examples of geometric deep learning problems and present available solutions, key difficulties, applications, and **future research directions in this nascent field.**”

- *Geometric deep learning: A Quick Tour* (kosasih-2020 [132])

Explainability and interpretability in ML models

- *Explainable and Interpretable Models in Computer Vision and Machine Learning* (escalante-escalera-guyon-baro-et-3-2018 [133]*).
- *GNNExplainer: Generating explanations for graph neural networks* (ying-bourgeois-you-zitnik-lescovec-2019 [134])
- *One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques* (arya-bellamy-chen-et-17-2019 [135])
- *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI* (arrieta-et-11-2020 [136])

- *Learning algebraic structures: preliminary investigations* (he-kim-2019 [137])
- *Machine learning meets number theory: The data science of Birch-Swinnerton-Dyer* (alessandretti-baronchelli-he-2019 [138])
- *Deep learning for symbolic mathematics* (lample-charton-2019 [139])

“Neural networks have a reputation for being better at solving statistical or approximate problems than at performing calculations or working with symbolic data. In this paper, we show that they can be surprisingly good at more elaborated tasks in mathematics, such as [symbolic integration](#) and [solving differential equations](#). We propose a syntax for representing mathematical problems, and methods for generating large datasets that can be used to train sequence-to-sequence models. We achieve [results that outperform commercial Computer Algebra Systems such as Matlab or Mathematica](#).”

- *Machine learning and the physical sciences*
(carleo-cirac-cranmer-daudet-schuld-tishby-vogtmaranto-zdeborova-2019 [140])
- *Graph Laplacians, Riemannian Manifolds and their Machine-Learning*
(he-yau-2020 [141])

- *Discovering Symbolic Models from Deep Learning with Inductive Biases* (cranmer-et-5-2020 [142])
- *Graph Neural Networks Meet Neural-Symbolic Computing: A Survey and Perspective* (lamb-garcez-gori-prates-avelar-varidi-2020 [143])

The Quantum internet

- *From Long-distance Entanglement to Building a Nationwide Quantum Internet: Report of the DOE Quantum Internet Blueprint Workshop* (osti-2020 [144])

Philosophy, Ethics

- *Human-level intelligence or animal-like abilities?* (darwiche-2018 [145])
- *On the relative expressiveness of Bayesian and neural networks* (choi-wang-darwiche-2019 [146])

Dimensionality reduction

- *Visualizing data using t-SNE* (vdmaaten-hinton-2008 [147])
- *Accelerating t-SNE using tree-based algorithms* (vdmaaten-2014 [148])
- *Studying the impact of the full-network embedding on multimodal pipelines* (vilalta-garciagasulla-et-5-2019 [149])
- *Overview and comparative study of dimensionality reduction techniques for high dimensional data* (ayesha-hanif-talib-2020 [150])

CapsNets

- *Dynamic routing between capsules* (sabour-frosst-hinton-2017 [151])
- *Matrix capsules with EM routing* (hinton-sabour-frosst-2018 [152])
- *Examining the Benefits of Capsule Neural Networks* (punjabi-schmid-katsaggelos-2020 [153])

Neuroscience

- *Backpropagation and the brain* (lillicrap-et-4-2020 [154])

Invariance and covariance

- *Group equivariant convolutional networks* (cohen-welling-2016 [155])

Physics

- *Toward an AI physicist for unsupervised learning* (wu-tegmark-2018 [156])
- *Discovering physical concepts with neural networks* (iten-metger-wilming-delrio-renner-2020 [157])

References I

- [1] M. Tegmark, *Life 3.0: Being human in the age of artificial intelligence*. Knopf, 2017.
- [2] J. Hawkins and S. Blakeslee, *On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines*. Macmillan, 2007.
- [3] R. Kurzweil, *How to create a mind: The secret of human thought revealed*. Viking, 2013.
- [4] S. Olhede and P. Wolfe, "The ai spring of 2018," 2018. <http://discovery.ucl.ac.uk/10051919/1/AIspringv5.pdf>, 3 pages. Authors discuss the implications as nations race for AI dominance.
- [5] K.-F. Lee, *AI superpowers: China, Silicon Valley, and the new world order*. Houghton Mifflin Harcourt, 2018.
- [6] D. Garcia-Gasulla, A. Cortés, and U. Cortés, "Traffic signs for ai," 2020. Preprint.
- [7] N. Brown and T. Sandholm, "Superhuman AI for multiplayer poker," *Science*, vol. 10, no. 1126, p. 13 pages, 2019.
- [8] T. G. Dietterich and P. Langley, "Machine learning for cognitive networks: Technology assessment and research challenges," 2003.
- [9] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*. Cambridge University Press, 2020.

References II

- [10] S. B. McGrayne, *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale University Press, 2011.
- [11] L. D. Stone, *Theory of optimal search*, vol. 118 of *Mathematics in Science and Engineering*. Academic Press, 1975.
- [12] D. Mumford and A. Desolneux, *Pattern theory: the stochastic analysis of real-world signals*. A. K. Peters, 2010.
- [13] N. Silver, *The signal and the noise: the art and science of prediction*. Penguin UK, 2012.
- [14] C. Robert, *The Bayesian choice: from decision-theoretic foundations to computational implementation (second edition)*. Springer, 2007.
- [15] B. De Finetti, *Theory of probability: A critical introductory treatment*, vol. 6. John Wiley & Sons, 2017.
This new publication is Volume I and Volume II combined.
- [16] T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, *Handbook of latent semantic analysis*. Routledge, 2007.
- [17] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision (4th edition)*. CENGAGE Learning, 2015.
- [18] G. Strang, *Linear algebra and learning from data*. Wellesley-Cambridge Press, 2019.

References III

- [19] E. Alpaydin, *Introduction to machine learning (fourth edition)*. Adaptive computation and machine learning, MIT press, 2020. 1st edition: 2004; 2nd, 2010; 3rd, 2014.
- [20] S. Arora, "Mathematics of Machine Learning: An introduction," 2018.
- [21] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.
- [22] F. Cucker and D. X. Zhou, *Learning theory: an approximation theory viewpoint*, vol. 24. Cambridge University Press, 2007.
- [23] S. R. Kulkarni and G. Harman, "Statistical learning theory: a tutorial," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 3, no. 6, pp. 543–556, 2011.
- [24] S. Kulkarni and G. Harman, *An elementary introduction to statistical learning theory*, vol. 853. John Wiley & Sons, 2011.
- [25] S. Sra, S. Nowozin, and S. J. Wright (editors), *Optimization for machine learning*. Neural Information Processing Series, MIT Press, 2012.
- [26] M. M. Wolf, "Mathematical foundations of supervised learning," 2018.
- [27] A. Blum, J. Hopcroft, and R. Kannan, *Foundations of Data Science*. Cambridge University Press, 2020.
- [28] J. Pearl, *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.

References IV

- [29] D. C. Knill and W. Richards (editors), *Perception as Bayesian inference*. Cambridge University Press, 1996.
Paper #1 is *Pattern theory: a unifying perspective*, by D. Mumford.
- [30] J. Pearl, "The art and science of cause and effect," 1996.
A public lecture delivered November 1996 as part of the UCLA Faculty Research Lectureship Program.
- [31] E. T. Jaynes, *Probability theory: The logic of science*. Cambridge university press, 2003.
- [32] R. E. Neapolitan, *Learning Bayesian networks*, vol. 38 of *Artificial Intelligence*. Pearson Prentice Hall Upper Saddle River, NJ, 2004.
- [33] D. Sivia and J. Skilling, *Data analysis: A Bayesian tutorial*. Oxford University Press, 2006.
- [34] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.
- [35] J. Pearl, *Causality. Models, Reasoning, and Inference (Second edition)*. Cambridge University Press, 2009.
- [36] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
xxxvi+1233 p.
- [37] A. Darwiche, *Modeling and reasoning with Bayesian networks*. Cambridge University Press, 2009.

References V

- [38] G. Cybenko and V. Crespi, "Learning hidden Markov models using nonnegative matrix factorization," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3963–3970, 2011.
- [39] A. Blake, P. Kohli, and C. Rother (editors), *Markov random fields for vision and image processing*. MIT Press, 2011.
- [40] A. Gelman, "Causality and statistical learning," 2011.
- [41] D. Barber, *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.
- [42] J. Pearl, "Causes of effects and effects of causes," *Sociological Methods & Research*, vol. 44, no. 1, pp. 149–164, 2015.
- [43] J. Pearl, "Trygve haavelmo and the emergence of causal calculus," *Econometric Theory*, vol. 31, no. 1, pp. 152–179, 2015.
- [44] G. Van den Broeck, K. Mohan, A. Choi, A. Darwiche, and J. Pearl, "Efficient algorithms for Bayesian network parameter learning from incomplete data," in *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI)*, p. 15, 2015.
- [45] E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7345–7352, 2016.
- [46] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.
- [47] J. Pearl, "Theoretical impediments to machine learning with seven sparks from the causal revolution," 2018. [arXiv:1801.04016](https://arxiv.org/abs/1801.04016). Keynote Talk at the WSDM'18 (Web Search and Data Mining), February 5-9, 2018, Marina Del Rey, CA, USA.

References VI

- [48] J. Pearl and D. Mackenzie, *The book of why: The new science of cause and effect*. Basic Books, 2018.
- [49] A. B. Patel, T. Nguyen, and R. G. Baraniuk, “A probabilistic theory of deep learning,” 2015.
- [50] A. Spanos, *Probability Theory and Statistical Inference: Empirical Modeling with Observational Data (second edition)*. Cambridge University Press, 2019.
- [51] K. Mohan and J. Pearl, “Graphical models for processing missing data,” 2019. <https://arxiv.org/pdf/1801.03583.pdf>, v2.
- [52] I. Hipolito, M. J. D. Ramstead, L. Convertino, A. Bhat, K. Friton, and T. Parr, “Markov blankets in the brain,” 2020. <https://arxiv.org/ftp/arxiv/papers/2006/2006.02741.pdf>, 25 pages.
- [53] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943.
- [54] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain,” *Psychological review*, vol. 65, no. 6, p. 386, 1958.
- [55] B. Widrow and M. E. Hoff, “Adaptive switching circuits,” tech. rep., Stanford Electronics Labs, Stanford University, CA, 1960.
- [56] F. Rosenblat, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, 1962.
- [57] M. Minsky and S. Papert, “Perceptrons,” *MIT Press*, 1969.
- [58] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

References VII

- [59] L. Nadel (editor), *Encyclopedia of cognitive science*. MacMillan London, 2003.
2006: <https://onlinelibrary.wiley.com/doi/book/10.1002/0470018860>, Wiley Online Library.
- [60] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.
- [61] S. Haykin, "Neural networks and learning machines," 2009.
- [62] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the Conference Neural Information Processing Systems (NIPS 2011)*, pp. 1097–1105, 2012.
<http://www.image-net.org/>.
- [63] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
- [64] M. A. Nielsen, *Neural networks and deep learning*, vol. 25. Determination Press, San Francisco, CA, USA, 2015.
- [65] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.
- [66] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [67] S. Mallat, "Understanding deep convolutional networks," *Phil. Trans. R. Soc. A*, vol. 374, no. 2065, p. 20150203, 2016.
- [68] C. C. Aggarwal, *Neural networks and deep learning*. Springer, 2018.
- [69] D.-X. Zhou, "Universality of deep convolutional neural networks," *Applied and Computational Harmonic Analysis*, 2019.

References VIII

- [70] Z. Li, W. Yang, S. Peng, and F. Liu, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," 2020.
<https://arxiv.org/ftp/arxiv/papers/2004/2004.02806.pdf>.
- [71] J. Manero Font, *Deep Learning architectures applied to wind time series multi-step forecasting*. PhD thesis, Computer Science Department, Universitat Politècnica de Catalunya - BarcelonaTECH, 2020. xxvi+242 p.
- [72] A. Hirose (ed), *Complex-valued neural networks: Theories and applications*, vol. 5 of *Series on Innovative Intelligence*. World Scientific, 2003.
- [73] A. Hirose, "Complex-valued neural networks: The merits and their origins," in *2009 International joint conference on neural networks*, pp. 1237–1244, IEEE, 2009.
- [74] T. Nitta, *Complex-Valued Neural Networks: Utilizing High-Dimensional Parameters*. IGI Global, 2009.
- [75] I. Aizenberg, *Complex-valued neural networks with multi-valued neurons*, vol. 353. Springer, 2011.
- [76] A. Hirose, *Complex-valued neural networks (second edition)*. Springer, 2012.
Japanese edition 2004, first English edition 2006.
- [77] A. Hirose, *Complex-Valued Neural Networks*. Springer-Verlag, 2006.
- [78] A. Hirose (editor), *Complex-valued neural networks: Advances and applications*. IEEE Computational Intelligence, John Wiley & Sons, 2013.

References IX

- [79] J. Bruna, S. Chintala, Y. LeCun, S. Piantino, A. Szlam, and M. Tygert, "A mathematical motivation for complex-valued convolutional networks," *arXiv: 1503. 03438*, 2015.
- [80] N. Guberman, "On complex valued convolutional neural networks," 2016.
arXiv:1602.09046.
- [81] C.-A. Popa, "Complex-valued convolutional neural networks for real-valued image classification," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 816–822, IEEE, 2017.
- [82] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," *arXiv: 1705. 09792*, 2017.
- [83] N. Mönning and S. Manandhar, "Evaluation of complex-valued neural networks on real-valued classification tasks," *arXiv: 1811. 12351*, 2018.
- [84] T. Bulow and G. Sommer, "Quaternionic Gabor filters for local structure classification," in *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, vol. 1, pp. 808–810, IEEE, 1998.
- [85] S. Buchholz and G. Sommer, "Quaternionic spinor MLP," in *ESANN'2000 Proceedings*, pp. 377–382, D-Facto, 2000. European Symposium on Artificial Neural Networks, Bruges (Belgium), 26-28 April 2000.
- [86] W. L. Chan, H. Choi, and R. Baraniuk, "Quaternion wavelets for image analysis and processing," in *IEEE International Conference on Image Processing*, vol. 5, pp. 3057–3060, 2004.
- [87] T. Isokawa, N. Matsui, and H. Nishimura, "Quaternionic neural networks: Fundamental properties and applications," in *Complex-valued neural networks: utilizing high-dimensional parameters*, pp. 411–439, IGI Global, 2009.
- [88] E. U. Moya-Sánchez and E. Bayro-Corrochano, "Quaternion atomic function wavelet for applications in image processing," in *Iberoamerican Congress on Pattern Recognition*, pp. 346–353, Springer, 2010.

References X

- [89] T. Isokawa, H. Nishimura, and N. Matsui, "Quaternionic multilayer perceptron with local analyticity," *Information*, vol. 3, no. 4, pp. 756–770, 2012.
<https://arxiv.org/pdf/1901.09342>.
- [90] E. Hitzer and S. J. Sangwine, *Quaternion and Clifford Fourier transforms and wavelets*. Springer, 2013.
- [91] M. Kobayashi, "Rotational invariance of quaternionic Hopfield neural networks," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 11, no. 4, pp. 516–520, 2016.
- [92] T. Parcollet, M. Morchid, P.-M. Bousquet, R. Dufour, G. Linarès, and R. De Mori, "Quaternion neural networks for spoken language understanding," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, pp. 362–368, IEEE, 2016.
- [93] K. Takahashi, Y. Hasegawa, and M. Hashimoto, "Design of quaternion-neural-network-based self-tuning control systems," *Sensors and Materials*, vol. 29, no. 6, pp. 699–711, 2017.
- [94] T. Parcollet, Y. Zhang, M. Morchid, C. Trabelsi, G. Linarès, R. De Mori, and Y. Bengio, "Quaternion convolutional neural networks for end-to-end automatic speech recognition," *arXiv: 1806. 07789*, 2018.
- [95] C. J. Gaudet and A. S. Maida, "Deep quaternion networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2018.
- [96] X. Zhu, Y. Xu, H. Xu, and C. Chen, "Quaternion convolutional neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 631–647, 2018.
- [97] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *Advances in neural information processing systems*, pp. 6571–6583, 2018.

References XI

- [98] Y. Zhao, T. Birdal, J. E. Lenssen, E. Menegatti, L. Guibas, and F. Tombari, "Quaternion Equivariant Capsule Networks for 3D Point Clouds," 2019.
<http://arxiv.org/abs/1912.12098>, v2.
- [99] E. U. Moya-Sánchez, S. Xambó-Descamps, A. S. Pérez, S. Salazar-Colores, J. Martínez-Ortega, and U. Cortés, "A bio-inspired quaternion local phase cnn layer with contrast invariance and linear sensitivity to rotation angles," *Pattern Recognition Letters*, vol. 131, pp. 56–62, 2020.
- [100] J. K. Pearson and D. L. Bisset, "Neural networks in the Clifford domain," in *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, vol. 3, pp. 1465–1469, IEEE, 1994.
- [101] G. Sommer (ed.), *Geometric computing with Clifford algebras: theoretical foundations and applications in computer vision and robotics*. Springer, 2001.
- [102] S. Buchholz and G. Sommer, "Clifford algebra multilayer perceptrons," in *Geometric computing with Clifford algebras*, pp. 315–334, Springer, 2001.
- [103] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE Transactions on Signal Processing*, vol. 49, no. 12, pp. 3136–3144, 2001.
- [104] T. Bülow and G. Sommer, "Hypercomplex signals – a novel extension of the analytic signal to the multidimensional case," *IEEE Transactions on signal processing*, vol. 49, no. 11, pp. 2844–2852, 2001.
- [105] J. Ebling and G. Scheuermann, "Clifford convolution and pattern matching on vector fields," in *IEEE Visualization, 2003. VIS 2003.*, pp. 193–200, IEEE, 2003.
- [106] E. Bayro-Corrochano, N. Arana, and R. Vallejo, "Design of kernels for support multivector machines involving the Clifford geometric product and the conformal geometric neuron," in *Proceedings of the International Joint Conference on Neural Networks, 2003.*, vol. 4, pp. 2893–2898, IEEE, 2003.

References XII

- [107] S. Buchholz, *A theory of neural computation with Clifford algebras*. PhD thesis, Christian-Albrechts Universität Kiel, 2005.
- [108] J. Ebling and G. Scheuermann, "Clifford Fourier transform on vector fields," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 4, pp. 469–479, 2005.
- [109] J. Rivera-Rovelo and E. Bayro-Corrochano, "Medical image segmentation using a self-organizing neural network and Clifford geometric algebra," in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pp. 3538–3545, IEEE, 2006.
- [110] S. Buchholz and G. Sommer, "On Clifford neurons and Clifford multi-layer perceptrons," *Neural Networks*, vol. 21, no. 7, pp. 925–935, 2008.
- [111] S. Buchholz, E. Hitzer, and K. Tachibana, "Coordinate independent update formulas for versor Clifford neurons," 2008. SCIS & ISIS 2008, Japan Society for Fuzzy Theory and Intelligent Informatics, 814-819.
- [112] E. J. Bayro-Corrochano and N. Arana-Daniel, "Clifford support vector machines for classification, regression, and recurrence," *IEEE transactions on neural networks*, vol. 21, no. 11, pp. 1731–1746, 2010.
- [113] E. Bayro-Corrochano, *Geometric computing: for wavelet transforms, robot vision, learning, control and action*. Springer, 2010.
- [114] E. Bayro-Corrochano and G. Scheuermann, *Geometric algebra computing: in engineering and computer science*. Springer, 2010.
- [115] S. Franchini, A. Gentile, F. Sorbello, G. Vassallo, and S. Vitabile, "Clifford algebra based edge detector for color images," in *2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems*, pp. 84–91, IEEE, 2012.
- [116] E. Hitzer, "The Clifford Fourier transform in real Clifford algebras," *Journal of Fourier analysis & applications*, vol. 2, no. 3, pp. 669–681, 2013.

References XIII

- [117] S. Franchini, A. Gentile, G. Vassallo, F. Sorbello, and S. Vitabile, "A specialized architecture for color image edge detection based on Clifford algebra," in *2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems*, pp. 128–135, IEEE, 2013.
- [118] K. Kanatani, "Overviews of optimization techniques for geometric estimation," *Memoirs of the Faculty of Engineering, Okayama University*, vol. 47, pp. 1–18, 2013.
- [119] K. Kanatani, *Understanding geometric algebra: Hamilton, Grassmann, and Clifford for computer vision and graphics*. CRC Press, 2015.
- [120] B. Mishra, P. Wilson, and R. Wilcock, "A geometric algebra co-processor for color edge detection," *Electronics*, vol. 4, no. 1, pp. 94–117, 2015.
- [121] M. T. Pham and K. Tachibana, "A conformal geometric algebra based clustering method and its applications," *Advances in Applied Clifford Algebras*, vol. 26, no. 3, pp. 1013–1032, 2016.
- [122] X.-X. Yin, S. Hadjiloucas, and Y. Zhang, "Outlook for Clifford algebra based feature and deep learning AI architectures," in *Pattern Classification of Medical Images: Computer Aided Diagnosis*, Health Information Science, pp. 165–177, Springer, 2017.
- [123] E. Bayro-Corrochano, *Geometric Algebra Applications Vol. I: Computer Vision, Graphics and Neurocomputing*. Springer, 2018.
- [124] T. M. Pham, D. C. Doan, and E. Hitzer, "Feature extraction using conformal geometric algebra for AdaBoost algorithm based inplane rotated face detection," *Advances in Applied Clifford Algebras*, vol. 29, no. 4, p. 19 pages, 2019.
- [125] R. Wang, W. Zhang, Y. Shi, X. Wang, and W. Cao, "Ga-orb: A new efficient feature extraction algorithm for multispectral images based on geometric algebra," *IEEE access*, vol. 7, pp. 71235–71244, 2019.

References XIV

- [126] R. Wang, Y. Shi, and W. Cao, "GA-SURF: A new speeded-up robust feature extraction algorithm for multispectral images based on geometric algebra," *Pattern Recognition Letters*, vol. 127, pp. 11–17, 2019.
- [127] W. B. Lopes and C. G. Lopes, "Geometric-algebra adaptive filters," *IEEE Transactions on Signal Processing*, vol. 67, no. 14, pp. 3649–3662, 2019.
- [128] M. Finzi, S. Stanton, P. Izmailov, and A. G. Wilson, "Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data," 2020.
[arXiv:2002.12880](https://arxiv.org/abs/2002.12880).
- [129] J. Hoffmann, S. Schmitt, S. Osindero, K. Simonyan, and E. Elsen, "Algebranets," 2020.
<https://arxiv.org/pdf/2006.07360.pdf>.
- [130] J. Wu, L. Xu, F. Wu, Y. Kong, L. Senhadji, and H. Shu, "Deep octonion networks," *Neurocomputing*, vol. 397, pp. 179–181, July 2020.
- [131] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond Euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.
- [132] E. E. Kosasih, "Geometric deep learning: A Quick Tour," 2020.
<https://towardsdatascience.com/geometric-deep-learning-a-quick-tour-12cef72492ca>.
- [133] H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, and M. Van Gerven (editors), *Explainable and interpretable models in computer vision and machine learning*. Challenges in Machine Learning, Springer, 2018.
- [134] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," in *Advances in Neural Information Processing Systems*, pp. 9240–9251, 2019.

References XV

- [135] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, *et al.*, "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques," 2019. [arXiv:1909.03012](https://arxiv.org/abs/1909.03012).
- [136] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.
- [137] Y.-H. He and M. Kim, "Learning algebraic structures: preliminary investigations," 2019. [arXiv:1905.02263](https://arxiv.org/abs/1905.02263).
- [138] L. Alessandretti, A. Baronchelli, and Y.-H. He, "Machine learning meets number theory: The data science of birch-swinnerton-dyer," 2019. <https://arxiv.org/pdf/1911.02008.pdf>.
- [139] G. Lample and F. Charton, "Deep learning for symbolic mathematics," 2019. <https://arxiv.org/pdf/1912.01412.pdf>.
- [140] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, "Machine learning and the physical sciences," *Reviews of Modern Physics*, vol. 91, no. 4, p. 045002, 2019. <https://arxiv.org/pdf/1903.10563.pdf>.
- [141] Y.-H. He and S.-T. Yau, "Graph laplacians, riemannian manifolds and their machine-learning," 2020. <https://arxiv.org/pdf/2006.16619.pdf>.
- [142] M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, and S. Ho, "Discovering symbolic models from deep learning with inductive biases," 2020. <https://arxiv.org/pdf/2006.11287v1.pdf>.

References XVI

- [143] L. Lamb, A. Garcez, M. Gori, M. Prates, P. Avelar, and M. Vardi, "Graph neural networks meet neural-symbolic computing: A survey and perspective," 2020.
[arXiv:2003.00330](https://arxiv.org/abs/2003.00330).
- [144] K. Kleese van Dam, I. Monda, N. Peters, and T. Schenkel, "From Long-distance Entanglement to Building a Nationwide Quantum Internet: Report of the DOE Quantum Internet Blueprint Workshop," 2020/2.
- [145] A. Darwiche, "Human-level intelligence or animal-like abilities?," *Communications of the ACM*, vol. 61, no. 10, pp. 56–67, 2018.
<https://dl.acm.org/doi/fullHtml/10.1145/3271625>.
- [146] A. Choi, R. Wang, and A. Darwiche, "On the relative expressiveness of Bayesian and neural networks," *International Journal of Approximate Reasoning*, vol. 113, pp. 303–323, 2019.
<https://arxiv.org/pdf/1812.08957.pdf>.
- [147] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [148] L. Van Der Maaten, "Accelerating t-SNE using tree-based algorithms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014.
- [149] A. Vilalta, D. Garcia-Gasulla, F. Parés, E. Ayguadé, J. Labarta, E. U. Moya-Sánchez, and U. Cortés, "Studying the impact of the full-network embedding on multimodal pipelines," *Semantic Web*, vol. 10, no. 5, pp. 909–923, 2019.
- [150] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, 2020.
- [151] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in neural information processing systems*, pp. 3856–3866, Springer, 2017.

References XVII

- [152] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with EM routing," in *ICLR 2018*, Springer, 2018. 15 pages.
- [153] A. Punjabi, J. Schmid, and A. K. Katsaggelos, "Examining the Benefits of Capsule Neural Networks," 2020. <http://arxiv.org/abs/2001.10964>.
- [154] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton, "Backpropagation and the brain," *Nature Reviews Neuroscience*, 2020.
- [155] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *International conference on machine learning*, pp. 2990–2999, 2016.
- [156] T. Wu and M. Tegmark, "Toward an AI Physicist for Unsupervised Learning," 2018. [arXiv:1810.10525](https://arxiv.org/abs/1810.10525), v2.
- [157] R. Iten, T. Metger, H. Wilming, L. Del Rio, and R. Renner, "Discovering physical concepts with neural networks," *Physical Review Letters*, vol. 124, no. 1, p. 010508, 2020.