

#### **Review**



# Labels in a haystack: Approaches beyond supervised learning in biomedical applications

Artur Yakimovich,<sup>1,5,\*</sup> Anaël Beaugnon,<sup>2,5,\*</sup> Yi Huang,<sup>3,5,\*</sup> and Elif Ozkirimli<sup>4,\*</sup>

<sup>1</sup>Roche Pharma International Informatics, Roche Products Limited, Welwyn Garden City, UK

<sup>2</sup>Roche Pharma International Informatics, Roche, Boulogne-Billancourt, France

<sup>3</sup>Roche Pharma International Informatics, Roche (China) Holding Ltd., Shanghai, China

<sup>4</sup>Roche Pharma International Informatics, F. Hoffmann-La Roche AG, Kaiseraugst, Switzerland

<sup>5</sup>These authors contributed equally

\*Correspondence: artur.yakimovich@roche.com (A.Y.), anael.beaugnon@roche.com (A.B.), anael.beaugnon@roche.com (Y.H.), elif. ozkirimli@roche.com (E.O.)

https://doi.org/10.1016/j.patter.2021.100383

**THE BIGGER PICTURE** As machine learning models become more complex, requirements for large annotated datasets grow. Annotating data for machine learning applications is especially challenging in the biomedical domain as it requires domain expertise of highly trained specialists to perform the annotations. Several strategies to either increase efficiency of label utilization or improve the annotation process have been proposed by the machine learning community. In this review we explore these strategies, including semi-supervised learning, active learning, data augmentation, transfer learning, self-supervision, weak-supervision, and zero- or few-shot learning. We show successful examples of research that has applied these strategies to multi-modal biomedical data. We conclude that raising awareness of these strategies in the biomedical community may contribute to further adoption of machine learning techniques in this research field.

#### SUMMARY

Recent advances in biomedical machine learning demonstrate great potential for data-driven techniques in health care and biomedical research. However, this potential has thus far been hampered by both the scarcity of annotated data in the biomedical domain and the diversity of the domain's subfields. While unsupervised learning is capable of finding unknown patterns in the data by design, supervised learning requires human annotation to achieve the desired performance through training. With the latter performing vastly better than the former, the need for annotated datasets is high, but they are costly and laborious to obtain. This review explores a family of approaches existing between the supervised and the unsupervised problem setting. The goal of these algorithms is to make more efficient use of the available labeled data. The advantages and limitations of each approach are addressed and perspectives are provided.

#### INTRODUCTION

Learning a task is easier when you have examples. In the absence of examples, humans can leverage other strategies to learn new material, such as generalization based on examples from similar tasks or trial and error. Most of these strategies aim to make efficient use of the available examples, if there are any. Such strategies work best when the data (or examples) come from a representative and unbiased sampling of the underlying data landscape.

In life sciences and health care, finding labeled data that sample the whole distribution is a major challenge. For example, in microscopy, data scarcity, cross-equipment compatibility, resolution limitations, and image quality are challenges in building consensus-labeled datasets.<sup>1,2</sup> Another example is related to

the drug discovery domain; protein-compound interaction datasets are limited to highly studied proteins or compounds leading to an oversampling of privileged protein-compound pairs.<sup>3</sup> In health care, access to patient data and finding consensuslabeled electronic health records (EHR) remain a huge challenge.<sup>4,5</sup> Furthermore, the population is often not sampled with a fair or representative distribution, leading to bias in the available datasets.<sup>6,7</sup> Another area of focus in life sciences and health care is finding information in scientific publications,<sup>8</sup> where labeled data are long-tailed due to the flexible queries and differences in research topic popularity. Accelerated by COVID-19 research,<sup>9</sup> the field is experiencing a huge increase in the number of publications, making information retrieval a major bottleneck in keeping up with literature.<sup>10</sup>



#### Figure 1. A schematic depiction of supervised and unsupervised machine learning approaches

Patterns

(A) An example of a supervised approach. Here, data points (circles) in ×1 and ×2 dimensions are labeled in magenta and green categories, allowing the model (dashed line) to be fitted.
(B) An example of an unsupervised approach. Here an algorithm attempts to detect patterns (clusters; dashed line) in unlabeled data (light blue circles). For simplicity of graphical illustration, all the data points are depicted in a 2D space of ×1 and ×2. However, similar concepts apply to n-dimensional space. Dashed line represents an abstraction for a model or a decision boundary.

The lack of labeled data is a critical challenge that must be overcome to train supervised learning models in the biomedical field. Some solutions to deal with this problem have been reviewed in biomedical imaging<sup>11</sup> and clinical text data analysis.<sup>12</sup> In this paper, we summarize several machine learning strategies that can be used with no or limited labeled data, with a special focus on life sciences and health-care-related domains. These strategies are blurring the boundary between supervised and unsupervised learning.

Machine learning (ML) approaches are traditionally separated into supervised and unsupervised paradigms.<sup>13</sup> Additionally, many researchers single out reinforcement learning as a third paradigm, which is not discussed in this review. In the supervised paradigm, the machine learning algorithm learns how to perform a task from data manually annotated by a person. It is worth mentioning that for the sake of this review we define supervision as manual, although such definition is strict. This is aimed to concretely define assisted approaches to labeling in the respective sections below. In contrast, the unsupervised paradigm aims to identify the patterns within the data algorithmically, i.e., without human help (Figure 1). Supervised learning delivers more desirable results for automation, aiming to mimic human behavior, but it falls short in scalability, due to the laborious and often expensive process of data annotation.<sup>14,15</sup> Unsupervised learning can also be used to automate some tasks (e.g., anomaly detection), but it usually performs much more poorly than supervised learning because it is not guided by manually annotated data. Unsupervised learning is therefore more suited to data exploration tasks such as clustering and association mining.

The scarcity of annotated data hinders the use of supervised learning in many practical use cases. This is especially pronounced for the application of higher expressive capacity models used in representation learning and deep learning (DL). Such models are much more data-hungry (often by a factor of a 100) compared with feature-generation-based ML.<sup>16–18</sup> This means that even more annotated examples are required to learn relevant representations from the data than in traditional ML.

In this paper, we review some techniques that have been proposed by the ML community to address the annotated data scarcity problem. First, we discuss the value each data point brings to the resulting model. Next, we discuss *Semi-supervised Learning* and *Active Learning*, which fall between the supervised and the unsupervised paradigms by leveraging both labeled and unlabeled data. *Data Augmentation* and *Self-supervised Learning* generate reliable annotated data in an automatic fashion that can later be used by classical supervised learning models. *Transfer Learning* and *Zero/One/Few-Shot Learning* techniques are able to leverage models pre-trained for a similar context (Table 1). Finally, *Weakly Supervised Learning* techniques learn predictive models from inexpensive weak labels that may be wrong. Table 2 summarizes the references to biomedical applications for each technique and Table 3 depicts which approaches are relevant according to the amount of labeled and unlabeled data available.

#### **DATA VALUE**

ML model performance often improves as more data are collected.<sup>70</sup> Increasing the size of the dataset serves two purposes simultaneously. First, it provides more information about the problem, making the solution likely to be more general. Second, it improves the performance of complex models. Increasing the dataset size often means an increase in the number of unannotated data points.<sup>15</sup> However, in a supervised learning setting, simply collecting observations is often insufficient. To be used as the training data for supervised learning, these observations must often be manually annotated. For instance, ImageNet,<sup>14</sup> an image dataset of around 1.3 million individual images, required manually annotating each of these images to belong to one or more of 1,000 classes in a routine and laborious process.<sup>71</sup>

However, not every data point is created equal. Some data points may be more useful to obtain a representative dataset. For example, for classification problems in cases when labeled data is scarce, it is preferable to have data points closer to the decision boundary (Figure 2). Several strategies for data point valuation have thus far been proposed, ranging from linear classification to decision trees and game theory (Shapley values).<sup>72</sup> Purposeful data collection may be achieved only when the goal is clearly defined. Cost-effective ways are therefore designed to weight the data points differently during training by the effective number (expected volume) of samples<sup>73</sup> or by focusing on hard examples.<sup>74</sup>

In recent years, a number of approaches to improve efficiency of labeled data utilization have gained traction.<sup>17,75,76</sup> Overall, these approaches attempt to improve the efficiency of labeled data utilization by high expressive capacity models. Such models often count millions of trainable parameters and hence require very large labeled datasets to avoid overfitting. Approaches tackling this are therefore indispensable for fields Table 1. Examples of zero-shot learning by formatting text data to fit models

Task	Input	Output	Reference
Multi-class classification	x	Y	
Question answering	X. Is this about Y?	Yes/No	Sun et al. <sup>19</sup>
Natural language inference	X. This is about Y.	Entailment/ Contradiction	Yin et al. <sup>20</sup>
Cloze	<b>X</b> . This is about	Choose from <b>Y</b>	Schick and Schütze, <sup>21</sup> Tam et al. <sup>22</sup>

where labeled data is scarce. In this review, the most popular examples are discussed.

#### SEMI-SUPERVISED LEARNING

Semi-supervised learning is halfway between supervised and unsupervised learning.<sup>77</sup> It trains predictive models from both labeled and unlabeled data (Figure 3A) in order to obtain better models than if they were trained with plain supervised learning on only the labeled data available.

There is an important prerequisite to obtain better models than supervised learning: the distribution of examples, which unlabeled data will help elucidate, must be relevant for the prediction problem at hand. If this condition is not met, using unlabeled data can degrade the prediction accuracy by misguiding the inference. In practice, semi-supervised learning strategies rely on assumptions about the data.

For instance, the generative semi-supervised method that consists in applying the expectation maximization (EM) algorithm on both labeled and unlabeled data<sup>78</sup> relies on the cluster assumption (if points are in the same cluster, they are likely to be of the same class). Another example is transductive support vector machine (SVM),<sup>79</sup> which implements the low-density separation assumption (the decision boundary should lie in a low-density region): it maximizes the margin for both labeled and unlabeled data. The reader may refer to Chapelle et al.<sup>77,80</sup> for more details about semi-supervised learning assumptions and strategies.

Semi-supervised learning will be most useful whenever there is far more unlabeled data than labeled. This is likely to occur if obtaining data points is cheap, but obtaining the labels costs a lot of time, effort, or money. For instance, semi-supervised learning is particularly suited to the classification of protein sequences.<sup>28</sup> Protein sequences are nowadays acquired at industrial speed, but determining the functions of a single protein may require years of scientific work.

#### **ACTIVE LEARNING**

Active learning<sup>81</sup> is an iterative human-in-the-loop process that starts by identifying the unlabeled data points that are considered as the most useful for training a predictive model. At each iteration, unlabeled data points are queried by the active learning



strategy and annotated by human experts, and the model is updated with the newly annotated data points (Figure 3B). Similar to semi-supervised learning, active learning aims to efficiently leverage unlabeled data during the learning process for performance promotion, while reducing the human expert workload. Both techniques can be combined to get the best out of the unlabeled data.<sup>82</sup>

The traditional active learning strategy, uncertainty sampling,<sup>83</sup> selects the data points closest to the decision boundary for annotation. The theory is that the model is uncertain about the prediction of these data points, so annotating them will significantly increase the accuracy. Other active learning strategies are more explorative and exploit the structure in data to annotate points from all of the data space.<sup>84</sup> Active learning strategies must find the right trade-off between exploitation and exploration, i.e., selecting data points both from sampled and unsampled areas of the data space.

Selecting the right query strategy and model for a given dataset is a real challenge. Indeed, active learning strategies perform differently on different datasets and there is no guarantee that they will outperform random sampling.<sup>35,85</sup> Meta-learning is investigated to address this issue:<sup>35,86</sup> it leverages reinforcement learning to learn the best active learning strategy.

Active learning is a human-in-the-loop process where the user experience should not be overlooked.<sup>87,88</sup> Taking humans into account and ensuring a good user experience is as important as choosing an optimal query strategy to effectively train a performant predictive model with reduced human workload. However, most studies do not evaluate their query strategy in a real-world setting: they leverage oracles that answer the queries automatically from fully annotated datasets and they leave the user experiments for future work. They assess the human workload with the number of manual annotations, but the time spent in the overall annotation process is a much more realistic metric. Indeed, data points do not all have the same cost to be annotated. For instance, the data points queried by uncertainty sampling, close to the decision boundary, are often tricky cases that are more costly to annotate.<sup>89</sup> In a real-world setting, even domain experts may not be able to annotate ambiguous data points. Moreover, studies often assume that a single data point is gueried and annotated at each active learning iteration, to always select the optimal query. This optimal setting is, however, not workable in realworld annotation systems,<sup>90</sup> as the annotators would spend more time waiting for annotation queries (while the predictive model is updated and the next annotation query is computed) than actually annotating data. Finally, most biomedical annotation tasks require specific domain knowledge, and crowdsourcing cannot be leveraged to annotate data at low cost. As annotating data is a cumbersome task, it is critical to show the biomedical experts that their annotations improve the accuracy of the predictive model, so that they continue annotating.<sup>18,87</sup> It is, however, not straightforward to provide such feedback:<sup>91</sup> most studies leverage a fully annotated validation dataset to assess the performance of the predictive model across iterations, but such a dataset is usually not available when deploying an annotation system.

There is currently little research that focuses on user experience in active learning processes, <sup>18,87,92,93</sup> and very few studies

# CellPress

Approach	Biomedical applications	
Supervised	<ul> <li>Protein-compound affinity pre- diction<sup>23</sup></li> <li>Omics<sup>24,25</sup></li> <li>Biomedical imaging<sup>1,2,26,27</sup></li> <li>EHR<sup>4,5</sup></li> </ul>	
Semi-supervised	• Omics <sup>28,29</sup>	
Active learning	<ul> <li>Drug discovery<sup>30,31</sup></li> <li>Omics<sup>32,33</sup></li> <li>Biomedical imaging<sup>11,34</sup></li> <li>Clinical text data<sup>35</sup></li> </ul>	
Data augmentation	<ul> <li>Omics<sup>36,37</sup></li> <li>Biomedical imaging<sup>38,39,40,41,42</sup></li> <li>Clinical text data<sup>43</sup></li> </ul>	
Transfer learning	<ul> <li>Omics<sup>44,45</sup></li> <li>Biomedical imaging<sup>46-49,26,50</sup></li> <li>Biomedical text data<sup>51-54,55</sup></li> </ul>	
Self-supervised	<ul> <li>Omics<sup>37,56</sup></li> <li>Biomedical imaging<sup>57,42</sup></li> </ul>	
Few/one/zero-shot and few-shot learning	<ul> <li>Omics and Affinity prediction<sup>58</sup></li> <li>Drug discovery<sup>59</sup></li> <li>Literature indexing<sup>60</sup></li> <li>Biomedical imaging<sup>61,62</sup></li> <li>Patient clustering<sup>63</sup></li> </ul>	
Weakly supervised	<ul> <li>Drug discovery<sup>64</sup></li> <li>Computer-aided diagnosis<sup>65</sup></li> <li>Biomedical imaging<sup>66,67,68</sup></li> <li>EHR<sup>69</sup></li> </ul>	

assess their method with user experiments.<sup>94</sup> Such research would significantly foster the use of this technique in real-world annotation systems.

Active learning strategies are particularly useful to annotate datasets for unbalanced prediction problems where random sampling is not effective. For example, it has been widely applied to automate drug discovery,<sup>30,31</sup> where the active learning strategy identifies which experiments to perform next. Active learning is also relevant when crowdsourcing cannot be leveraged because expert knowledge is required to annotate<sup>11,35,34</sup> or the data are too sensitive to be shared.

#### DATA AUGMENTATION

An early attempt to address the mismatch on availability and requirement of labeled data in DL was data augmentation<sup>95</sup> (Figure 3C). Traditional techniques apply label-preserving transformations to the already-annotated data points to increase the amount of training data. For instance, image data can simply be flipped or rotated,<sup>17,38</sup> and text data can be slightly modified through deletion of random words and synonym substitution.<sup>43,96,97</sup>

Generative adversarial networks (GANs)<sup>98</sup> can be used to generate a much broader set of augmentations.<sup>99</sup> GANs learn the data distribution from some training data and then generate new samples that are as realistic as possible from this distribution. They use two competing neural networks: one that generates new samples from noise and one that discriminates sam-

Patterns

ples as real or synthetic. GANs have been used in medical imaging to generate additional realistic training data.<sup>100</sup> For example, Calimeri et al.<sup>39</sup> generated magnetic resonance imaging (MRI) slices of the human brain with GANs and human physicians were not able to distinguish the artificially generated examples from the real ones.

Augmentation strategies are often carefully selected and finetuned for each individual case, taking domain-specific understanding of medical image data into account to avoid noise and artifacts. For instance, Mok and Chung<sup>40</sup> have proposed a new data augmentation technique that generates images with the desired invariance and robustness properties for brain tumor segmentation. They have fine-tuned their GAN to generate images with realistic tumor boundaries. In another study, Gupta and colleagues used image-to-image translation GAN to enrich images with a broader pallet of readouts.<sup>101</sup> In contrast to conventional augmentation, however, such an approach did not lead to an increase of individual data points.

In contrast to semi-supervised and active learning techniques, data augmentation strategies are useful even when a large amount of unlabeled data is not available. Data augmentation can be used to address class imbalance by generating more examples of underrepresented classes. For example, Ollagnier and Williams<sup>43</sup> have compared several text data augmentation techniques to remedy class imbalance in clinical case classification. In the case of image data, Jin et al.<sup>41</sup> have used a GAN to increase the training data of their DL model for pathological lung segmentation of CT scans. They have particularly focused on generating examples where nodules lie on the lung border because these specific cases were previously segmented poorly and underrepresented in the original training dataset.

#### **TRANSFER LEARNING**

Transfer learning is an ML technique where a model that is trained on one task is then repurposed on a second related task<sup>102</sup> (Figure 3D). It is not specific to DL, but it has been widely leveraged in this field as the representation layers can be easily shared by different prediction tasks. In practice, the weights of the representation layers are set according to a model trained on a previous task. Then, these weights can be either used for initialization with subsequent update during the training process (fine-tuning) or completely frozen, and only the task-specific layers are trained.

Transfer learning is a relevant solution to deal with the lack of annotated data if you can identify a related task with abundant labeled data or if a pre-trained model is available. It has significantly fostered the use of representation learning in biomedical applications.

Manual feature engineering was often preferred over representation learning in the biomedical field because of the lack of training data. Thanks to transfer learning, generic DL models and generic annotated datasets can be leveraged to train biomedical prediction models with few domain-specific annotated examples. For instance, the natural language model BioBERT<sup>103</sup> has been created from the generic model BERT<sup>104</sup> through transfer learning to obtain better performance on biomedical text mining tasks. It has been initialized with the weights of BERT (trained on general domain corpora) and trained  
 Table 3. Relevant approaches according to the amount of labeled and unlabeled data available

Amount of available data	Some unlabeled data	No unlabeled data		
Enough labeled data to train a supervised model	Supervised learning	Supervised learning		
	Data augmentation Semi-supervised learning	Data augmentation		
	Active learning			
Some labeled data, but not enough to train a supervised model	Data augmentation	Data augmentation		
	Semi-supervised learning	Transfer learning <sup>a</sup>		
	Active learning			
	Transfer learning <sup>a</sup>			
	Self-supervised learning <sup>a</sup>			
	Few/one/zero-shot learning <sup>a</sup>			
No labeled data	Active learning Zero-shot learning <sup>a</sup>	Zero-shot learning <sup>a</sup>		
<sup>a</sup> Large labeled datasets are required for pre-training.				

on biomedical domain corpora. BioBERT has also been further fine-tuned through transfer learning for specific NLP tasks such as biomedical Named Entity Recognition.<sup>51</sup> A similar approach was applied to scientific text, clinical notes, and PubMed articles, and yielded SciBERT,<sup>52</sup> ClinicalBERT,<sup>53</sup> and PubMedBERT.<sup>54</sup> Transfer learning is also widely used in computer vision to create biomedical specific models.<sup>46</sup> For instance, Cheng et al.<sup>47</sup> have leveraged transfer learning from a model pre-trained on ImageNet<sup>14</sup> to identify specific patterns from abdominal radiographs with limited training data. Another group<sup>48</sup> also used ImageNet pre-trained ResNet-50 for a classification task in micrographs of virus-infected cells. While other examples<sup>49</sup> include transfer learning from models trained on, for example, an MNIST dataset, 105 overall transfer learning from ImageNet-pre-trained models remains the most-favored approach. Proteomics and genomics benefit from transfer learning as well. For example, Deznabi et al.<sup>58</sup> used ProtVec<sup>106</sup> to predict kinase-phosphosite association.

#### SELF-SUPERVISED LEARNING

Self-supervised learning (Figure 3E) learns relevant representations (or embeddings) without any manual annotation cost.<sup>75,107–112</sup> It relies on a two-step process: (1) a *pretext task* (or self-supervised task) is used to learn meaningful representations from annotations that are inherent to the data, and (2) the learned representations are leveraged to address the *downstream task*, the task at hand, with many fewer labeled data.

Pretext tasks themselves do not usually provide any useful application. Their learning objective should be set properly to receive supervision from the data themselves. This way, a large amount of annotated data can be generated automatically to learn high-quality representations. Several pretext tasks have been proposed for images; Doersch et al.<sup>113</sup> have proposed a context prediction task where the objective is to predict the rela-



tive location of patches extracted from a given image. A similar pretext task consists of solving jigsaw puzzles where the tiles have been generated automatically by splitting images.<sup>107,114</sup> Pretext tasks can also rely on generative models. For instance, in-painting<sup>115</sup> is a generative pretext task where an arbitrary fraction of an image is removed, and the objective is to reconstruct it from the remaining context. In-painting has recently been demonstrated<sup>57</sup> to be capable of predicting unseen fluorescence signals on a single-cell level in confocal microscopy. Colorization<sup>116</sup> is another example of a generative pretext task: a gray-scale filter is applied to the input images and the objective is to recover the colors. Pretext tasks for language data are posed more natively, e.g., with next-word prediction<sup>117,118</sup> or masked language modeling and next-sentence prediction used by BERT.<sup>104</sup>

This approach may not immediately make intuitive sense. After all, what practical use would such an ML model present for solving, e.g., a classification problem? However, in the sense of representation learning, performing such a task achieves the main part of the goal: it allows the system to learn meaningful representation from a large dataset. Once done, these representations can be repurposed through transfer learning (weights transfer) and fine-tuning to address an actual task at hand with fewer labeled data. Furthermore, these two steps can be merged into one through shared weights.<sup>19,119–121</sup>

Representations learned through self-supervised learning can be leveraged for both supervised and unsupervised downstream tasks. The learned representations can be used directly for unsupervised learning tasks such as clustering<sup>122</sup> or similarity computation for search and retrieval tasks.<sup>123</sup> The model trained for the pretext task can also be repurposed through transfer learning<sup>57</sup> to address supervised tasks with many fewer labeled data. Self-supervised learning bridges the gap between supervised and unsupervised learning by focusing on learning high-quality representations. The pretext task must be chosen carefully to get representations as meaningful as possible for the downstream task. Multi-task learning can also be leveraged to train relevant representations from several pretext tasks.<sup>124</sup>

Self-supervised learning is proving incredibly powerful across various domains and types of data. This set of techniques refocuses the field on learning high-quality representations bridging domains of generative and discriminative modeling.<sup>125</sup>

#### FEW/ONE/ZERO-SHOT LEARNING

When there are only a limited number of valuable labeled data points, ML solutions have to adapt and generalize to tackle few-shot or one-shot learning, especially to avoid overfitting.<sup>126,127</sup> In some extreme situations, there are no labeled data at all for training (for example, when there are too many labels to annotate), and the model has to deal with unseen labels, leading to a challenge called zero-shot learning.<sup>128</sup> The solutions to the few/one/zero-shot learning challenges essentially imitate how humans recognize and learn by analogy when there is limited evidence. These solutions often fall into three strategies: relationship similarity, task reduction, and prior knowledge.

First, in most cases, with proper embedding or vector representation, the few/one/zero-shot learning models can locate



Figure 2. Example of data point importance for model selection

Patterns

Magenta and green circles are data points that correspond to respective class labels.

(A) An example of a full dataset with 8 data points belonging to magenta class and 14 belonging to the green class; dashed line represents a selected model.

(B) A subset of six data points selected from the full dataset is less valuable for accurate model defini-

tion, as multiple models (e.g., two dotted lines of incorrect models and one dashed line for correct model) can be fitted to this subset, but not the full dataset. (C) A subset of six data points selected from the full dataset are more valuable for accurate model definition (single dashed line). For simplicity of graphical illustration, all the data points are depicted in a 2D space of ×1 and ×2. However, similar concepts apply to n-dimensional space. Dashed line represents an abstraction for a model or a decision boundary.

the data or labels in the high-dimensional "semantic space", measure the similarity of their relationship to those in the training set, and provide proper predictions accordingly.<sup>129,20</sup> The embedding approach has proved its effectiveness in multiple biomedical domains, such as affinity prediction,<sup>58</sup> drug discov-

ery,<sup>59</sup> literature indexing,<sup>60</sup> image recognition for cancer detection,<sup>61,62</sup> and patient clustering.<sup>63</sup> To generate suitable embeddings or vector representations in each biomedical domain, self-supervised learning and transfer learning methods can be leveraged.



#### Figure 3. Strategies between supervised and unsupervised approaches

Magenta and green colors correspond to respective class labels, blue circles represent unlabeled data, dashed lines represent the learned decision boundary. (A) In a semi-supervised learning approach, clustering and manual annotation of few points is performed.

(B) In an active learning approach, an active request to the user to obtain annotation is performed; here it is depicted in a gray bubble with a "select" call to action. (C) In the data augmentation approach, light circles with dashed borders represent data points obtained from the original through augmentation (e.g., linear transformation).

(D) A transfer learning approach uses a model pre-trained on one dataset (regular circles) and fine-tuned on another dataset (light circles with dashed borders). Here the trained model parameters transfer is symbolized by the dashed model line "transferred" from left to right as indicated by the blue arrow.

(E) Self-supervised learning approach. Here, a jigsaw in-painting task, which is not related to labels (so-called pretext task) is depicted as an example. The jigsaw pretext task is formulated automatically and allows learning of representations from the data.

(F) Weakly supervised learning approach. Here, magenta and green inequalities represent coarse heuristic rules used for data annotation.

For simplicity of graphical illustration, all the data points are depicted in a 2D space of ×1 and ×2. However, similar concepts apply to n-dimensional space. Dashed line represents an abstraction for a model or a decision boundary.



Second, when there are well-established models in the same domain, it is more straightforward to format data to fit the models than to tune the models to fit the data. For example, with auxiliary sentences, the input and output of a text classification task can be compatible with a question-and-answer model,<sup>19</sup> a natural-language-inference model,<sup>20</sup> or a cloze model<sup>21,22</sup> (Table 1). Last, but not least, prior structured knowledge can provide additional domain knowledge to the models<sup>130</sup> to support "reasoning." For example, the International Classification of Diseases (ICD) hierarchy improves the classification model performance of infrequent labels in electronic medical records (EMR).<sup>131</sup>

#### WEAKLY SUPERVISED LEARNING

Many challenging prediction tasks require complex models (i.e., with large numbers of parameters) that need large training datasets. However, it is often difficult to acquire strong supervision such as large fully annotated datasets with perfectly accurate labels due to the high cost of the data-labeling process. In some cases, only weak supervision can be collected, and some training algorithms, called weakly supervised, can handle this imperfect supervision to build predictive models.<sup>132</sup> The goal of weak supervision is to train predictive models on large imperfectly labeled training datasets to get better performance than fully supervised models trained on small datasets annotated with perfectly accurate labels.

There are many settings of weak supervision. For instance, distant supervision leverages knowledge bases to derive some weak supervision.<sup>133</sup> In other cases, only partial labels, coarse-grained labels, may be available. Only the partial ordering may be provided when learning user preferences over items, or the coarse-grained label "flower" can be provided for a picture of an arum lily instead of spending a consequent amount of time to find the exact taxonomy.<sup>134</sup>

Multi-instance learning is another setting of weak supervision where the learner receives a set of labeled bags, each containing many instances. In the simple case of multiple-instance binary classification, a bag may be labeled negative if all the instances in it are negative. On the other hand, a bag is labeled positive if there is at least one instance in it which is positive. Multi-instance learning has been applied to drug-activity prediction,<sup>64</sup> computer-aided diagnosis,<sup>65</sup> and pathology.<sup>66</sup>

Weak supervision can also rely on labels with a low accuracy coming from rule-based heuristics<sup>135,136</sup> (Figure 3F) or crowd-sourcing annotations.<sup>137</sup> In this setting, each instance can be associated with several labels, and the key technical challenge is how to unify and de-noise them, given that they are each noisy, may disagree with one another, may be correlated, and may have arbitrary (unknown) accuracies that may depend on the subset of the dataset being labeled.<sup>136</sup> This type of weak supervision is gaining rapid adoption by the biomedical research community to assist in labeling electronic health records<sup>69</sup> or MRI images.<sup>67</sup>

#### **CONCLUSIONS: BIOMEDICAL APPLICATIONS**

The recent advances in data-intensive computing have been accelerated by improvements in performance due to techniques such as representation learning that address data curation and



feature engineering problems. Yet, as we improve the performance, models tend to get larger and the requirement for annotated datasets grows.

The set of techniques described in this review aim to make more efficient use of limited annotated data. These techniques focus on generating synthetic data from existing annotations, automating data annotation processes, or lowering the amount of annotated data needed to train the machine learning systems. Even though they were proposed independently, these methods show increasing overlap. For example, weak supervision and self-supervision in DL ultimately serve the same purpose, exploit similar implicit properties of the learning systems, and often require fine-tuning for the best performance. Weak supervision also has some conceptual overlap with data augmentation: both techniques generate labeled data automatically with a rule-based approach to provide some supervision at low cost. Another example is the similarity between semi-supervision, weak supervision, and active learning: they all leverage unlabeled data to enrich or create an annotated dataset with different levels of involvement from domain experts. While semi-supervised learning pseudo-labels unlabeled data automatically, weakly supervised learning requires domain experts to provide pseudo-labeling rules, and active learning requires even more involvement of domain experts to manually annotate some unlabeled data points. Finally, transfer learning is closely related to self-supervised learning and few/one/zero-shot learning. Selfsupervised learning and transfer learning are often applied jointly: self-supervised learning allows the collection of relevant representations at low cost, which are then fine-tuned through transfer learning to answer the problem at hand with fewer annotated data points. As for few/one-shot learning, it often leverages a model previously trained for another task such as transfer learning, while zero-shot learning often leverages a model pretrained with self-supervised learning without fine-tuning.

Several dawning and established techniques have been omitted from this review due to their broader focus on the way systems learn. They include meta-learning,<sup>138</sup> and universal representations.<sup>139</sup> The former relates to approaches aimed at optimizing the representation learning algorithms using training meta-data, i.e., learning to learn. Recent work in this field shows further improvement of learning algorithms is still possible.<sup>140</sup> The latter seeks to identify the best algorithmic way to obtain reusable (i.e., "universal") representations valid across multiple domains building upon self-supervised learning approaches. Perhaps it is those, or the methods reviewed in this work, that will one day allow us to stop searching for the labels in a haystack of data.

#### **DECLARATION OF INTERESTS**

The authors declare no competing interests.

#### REFERENCES

- Sanchez-Garcia, R., Segura, J., Maluenda, D., Carazo, J.M., and Sorzano, C.O.S. (2018). Deep Consensus, a deep learning-based approach for particle pruning in cryo-electron microscopy. IUCrJ. 5, 854–865.
- Wollmann, T., and Rohr, K. (2021). Deep Consensus Network: Aggregating predictions to improve object detection in microscopy images. Med. Image Anal. 70, 102019.



- Xiao, X., Min, J.-L., Lin, W.-Z., Liu, Z., Cheng, X., and Chou, K.-C. (2015). iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach. J. Biomol. Struct. Dyn. 33, 2221–2233.
- Ghassemi, M., Naumann, T., Schulam, P., Beam, A.L., Chen, I.Y., and Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. AMIA Jt. Summits Transl. Sci. Proc 2020, 191–200.
- Griffith, S.D., Tucker, M., Bowser, B., Calkins, G., Chang, C.-H.J., Guardino, E., Khozin, S., Kraut, J., You, P., Schrag, D., and Miksad, R.A. (2019). Generating real-world tumor burden endpoints from electronic health record data: comparison of RECIST, radiology-anchored, and clinician-anchored approaches for abstracting real-world progression in non-small cell lung cancer. Adv. Ther. *36*, 2122–2136.
- Geneviève, L.D., Martani, A., Shaw, D., Elger, B.S., and Wangmo, T. (2020). Structural racism in precision medicine: leaving no one behind. BMC Med. Ethics 21, 17.
- Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science 366, 447–453.
- Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J., and Valencia, A. (2017). Information retrieval and text mining technologies for chemistry. Chem. Rev. 117, 7673–7761.
- Köksal, A., Dönmez, H., Özçelik, R., Ozkirimli, E., and Özgür, A. (2020). Vapur: a search engine to find related protein - compound pairs in COVID-19 literature. bioRxiv, 2020.09.05.284224.
- Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., et al. (2020). CORD-19: the covid-19 open research dataset. ArXiv https://www.ncbi.nlm.nih.gov/ pmc/articles/PMC7251955/.
- Sahiner, B., Pezeshk, A., Hadjilski, L.M., Wang, X., Drukker, K., Cha, K.H., Summers, R.M., and Giger, M.L. (2019). Deep learning in medical imaging and radiation therapy. Med. Phys. 46, e1–e36.
- Spasic, I., and Nenadic, G. (2020). Clinical text data in machine learning: systematic review. JMIR Med. Inform. https://www.ncbi.nlm.nih.gov/ pmc/articles/PMC7157505/.
- 13. Hinton, G.E., and Sejnowski, T.J. (1999). Unsupervised Learning: Foundations of Neural Computation (MIT Press), p. 420.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, Kai, and Fei-Fei, Li (2009). ImageNet: a large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.
- Sorokin, A., and Forsyth, D. (2008). Utility data annotation with Amazon mechanical Turk. In 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8.
- Hinton G.E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R.R.. Improving neural networks by preventing co-adaptation of feature detectors. ArXiv http://arxiv.org/abs/1207.0580
- Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, 25, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, eds. (Curran Associates, Inc.), pp. 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deepconvolutional-neural-networks.pdf.
- Sun C., Shrivastava A., Singh S., Gupta A.. Revisiting unreasonable effectiveness of data in deep learning era. ArXiv http://arxiv.org/abs/ 1707.02968
- Sun, Y., Tzeng, E., Darrell, T., and Efros, A.A. (2019). Unsupervised domain adaptation through self-supervision. ArXiv http://arxiv.org/abs/ 1909.11825.
- 20. Yin, W., Hay, J., and Roth, D. (2019). Benchmarking zero-shot text classification: datasets, evaluation and entailment approach. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), S. Padó and R. Huang, eds. (Association for Computational Linguistics), pp. 3914–3923.



- Schick, T., and Schütze, H. (2020). It's not just size that matters: small language models are also few-shot learners. ArXiv http://arxiv.org/abs/ 2009.07118.
- Tam, D., Menon, R.R., Bansal, M., Srivastava, S., and Raffel, C. (2021). Improving and simplifying pattern exploiting training. ArXiv http://arxiv. org/abs/2103.11955.
- Öztürk, H., Özgür, A., and Ozkirimli, E. (2018). DeepDTA: deep drugtarget binding affinity prediction. Bioinformatics 34, i821–i829.
- Sun, S., Dong, B., and Zou, Q. (2020). Revisiting genome-wide association studies from statistical modelling to machine learning. Brief. Bioinform. https://doi.org/10.1093/bib/bbaa263.
- 25. Zrimec, J., Börlin, C.S., Buric, F., Muhammad, A.S., Chen, R., Siewers, V., Verendel, V., Nielsen, J., Töpel, M., and Zelezniak, A. (2020). Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nat. Commun. 11, 1–16.
- Fisch, D., Yakimovich, A., Clough, B., Wright, J., Bunyan, M., Howell, M., Mercer, J., and Frickel, E. (2019). Defining host-pathogen interactions employing an artificial intelligence workflow. eLife 8, e40560.
- Lucas, A.M., Ryder, P.V., Li, B., Cimini, B.A., Eliceiri, K.W., and Carpenter, A.E. (2021). Open-source deep-learning software for bioimage segmentation. Mol. Biol. Cell. 32, 823–829.
- Weston, J., Leslie, C., Ie, E., Zhou, D., Elisseeff, A., and Noble, W.S. (2005). Semi-supervised protein classification using cluster kernels. Bioinformatics 21, 3241–3247.
- Krogel, M.-A., and Scheffer, T. (2004). Multi-relational learning, text mining, and semi-supervised learning for functional genomics. Mach. Learn. 57, 61–81.
- Reker, D. (2019). Practical considerations for active machine learning in drug discovery. Drug Discov. Today Technol. 32–33, 73–79, (Artificial Intelligence).
- Schneider, G. (2018). Automating drug discovery. Nat. Rev. Drug Discov. 17, 97–113.
- 32. Farid, D.M., Nowé, A., and Manderick, B. (2016). Combining boosting and active learning for mining multi-class genomic data. In 25th Belgian-Dutch Conference on Machine Learning (Benelearn), pp. 1–2.
- Liu, Y. (2004). Active learning with support vector machine applied to gene expression data for cancer classification. J. Chem. Inf. Comput. Sci. 44, 1936–1941.
- Hoi, S.C.H., Jin, R., Zhu, J., and Lyu, M.R. (2006). Batch mode active learning and its application to medical image classification. In Proceedings of the 23rd International Conference on Machine Learning (Association for Computing Machinery), pp. 417–424. (ICML '06). https://doi.org/ 10.1145/1143844.1143897.
- De Angeli, K., Gao, S., Alawad, M., Yoon, H.-J., Schaefferkoetter, N., Wu, X.-C., Durbin, E.B., Doherty, J., Stroup, A., Coyle, L., et al. (2021). Deep active learning for classifying cancer pathology reports. BMC Bioinformatics 22, 113.
- Chaudhari, P., Agarwal, H., and Bhateja, V. (2021). Data augmentation for cancer classification in oncogenomics: an improved KNN based approach. Evol. Intell. 14, 489–498.
- Chen J., Mowlaei M.E., Shi X.: Population-scale Genomic Data Augmentation Based on Conditional Generative Adversarial Networks. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. 2020. p. 1–6.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 234–241.
- Calimeri, F., Marzullo, A., Stamile, C., and Terracina, G. (2017). Biomedical data augmentation using generative adversarial neural networks (In proceedings of the International Conference on Artificial Neural Networks). https://link.springer.com/chapter/10.1007/978-3-319-68612-7\_71.

## Patterns

Review

- Mok, T.C.W., and Chung, A.C.S. (2019). Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks. ArXiv 11383, 70–80.
- 41. Jin, D., Xu, Z., Tang, Y., Harrison, A.P., and Mollura, D.J. (2018). CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, A. Frangi, J. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, eds. (Springer), pp. 732–740.
- Horlava, N., Mironenko, A., Niehaus, S., Wagner, S., Roeder, I., and Scherf, N. (2020). A comparative study of semi- and self-supervised semantic segmentation of biomedical microscopy data. ArXiv http://arxiv. org/abs/2011.08076.
- Ollagnier, A., and Williams, H. (2020). Text Augmentation Techniques for Clinical Case Classification9.
- 44. Schwessinger, R., Gosden, M., Downes, D., Brown, R.C., Oudelaar, A.M., Telenius, J., Teh, Y.W., Lunter, G., and Hughes, J.R. (2020). DeepC: predicting 3D genome folding using megabase-scale transfer learning. Nat. Methods 17, 1118–1124.
- Taroni, J.N., Grayson, P.C., Hu, Q., Eddy, S., Kretzler, M., Merkel, P.A., and Greene, C.S. (2019). MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. Cell Syst 8, 380–394.
- Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). Transfusion: understanding transfer learning for medical imaging. ArXiv http://arxiv. org/abs/1902.07208.
- 47. Cheng, P.M., Tejura, T.K., Tran, K.N., and Whang, G. (2018). Detection of high-grade small bowel obstruction on conventional radiography with convolutional neural networks. Abdom. Radiol. N. Y. 43, 1120–1127.
- Andriasyan, V., Yakimovich, A., Petkidis, A., Georgi, F., Witte, R., Puntener, D., and Greber, U.F. (2021). Microscopy deep learning predicts virus infections and reveals mechanics of lytic-infected cells. Iscience 24, 102543.
- Yakimovich, A., Huttunen, M., Samolej, J., Clough, B., Yoshida, N., Mostowy, S., Frickel, E.-M., and Mercer, J. (2020). Mimicry embedding facilitates advanced neural network training for image-based pathogen detection. Msphere 5, e00836–20.
- Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C.S., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al. (2018). Identifying medical diagnoses and treatable Diseases by image-based deep learning. Cell *172*, 1122–1131.e9.
- Symeonidou, A., Sazonau, V., and Groth, P. (2019). Transfer learning for biomedical named entity recognition with BioBERT. In SEMANTICS Posters&Demos.
- Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: a pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Association for Computational Linguistics), pp. 3613–3618. https://www.aclweb.org/anthology/D19-1371.
- Huang, K., Altosaar, J., and Ranganath, R. (2020). ClinicalBERT: modeling clinical notes and predicting hospital readmission. ArXiv http://arxiv.org/abs/1904.05342.
- Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ArXiv http://arxiv.org/ abs/2007.15779.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., et al. (2019). Bio-BERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics.
- Kung, S.-Y., Luo, Y., and Mak, M.-W. (2010). Feature selection for genomic signal processing: unsupervised, supervised, and self-supervised scenarios. J. Signal. Process. Syst. 61, 3–20.



- Lu, A.X., Kraus, O.Z., Cooper, S., and Moses, A.M. (2019). Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. PLoS Comput. Biol. 15, e1007348.
- Deznabi, I., Arabaci, B., Koyutürk, M., and Tastan, O. (2020). DeepKin-Zero: zero-shot learning for predicting kinase–phosphosite associations involving understudied kinases. Bioinformatics 36, 3652–3661.
- Altae-Tran, H., Ramsundar, B., Pappu, A.S., and Pande, V. (2017). Low data drug discovery with one-shot learning. ACS Cent. Sci. 3, 283–293.
- Mylonas, N., Karlos, S., and Tsoumakas, G. (2020). Zero-shot classification of biomedical articles with emerging MeSH descriptors. In 11th Hellenic Conference on Artificial Intelligence (ACM), pp. 175–184. https://dl. acm.org/doi/10.1145/3411408.3411414.
- Kim, M., Zuallaert, J., and De Neve, W. (2017). Few-shot learning using a small-sized dataset of high-resolution FUNDUS images for glaucoma diagnosis. In Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care (Association for Computing Machinery), pp. 89–92. (MMHealth '17). https://doi.org/10. 1145/3132635.3132650.
- 62. Medela, A., Picon, A., Saratxaga, C.L., Belar, O., Cabezón, V., Cicchi, R., Bilbao, R., and Glover, B. (2019). Few shot learning in histopathological images:reducing the need of labeled data on biological datasets. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1860–1864.
- **63.** Ma, T., and Zhang, A. (2018). Affinity network fusion and semi-supervised learning for cancer patient clustering. Methods *145*, *16–24*, (Data mining methods for analyzing biological data in terms of phenotypes).
- 64. Dietterich, T.G., Lathrop, R.H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. 89, 31–71.
- Fung, G., Dundar, M., Krishnapuram, B., and Rao, R.B. (2007). Multiple instance learning for computer aided diagnosis. Adv. Neural Inf. Process. Syst. 19, 425.
- 66. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., and Fuchs, T.J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25, 1301–1309.
- 67. Fries, J.A., Varma, P., Chen, V.S., Xiao, K., Tejeda, H., Saha, P., Dunnmon, J., Chubb, H., Maskatia, S., Fiterau, M., et al. (2019). Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. Nat. Commun. 10, 3111.
- Doan, M., Barnes, C., McQuin, C., Caicedo, J.C., Goodman, A., Carpenter, A.E., and Rees, P. (2021). Deepometry, a framework for applying supervised and weakly supervised deep learning to imaging cytometry. Nat. Protoc. 1–24.
- Callahan, A., Fries, J.A., Ré, C., Huddleston, J.I., Giori, N.J., Delp, S., and Shah, N.H. (2019). Medical device surveillance with electronic health records. Npj Digit. Med. 2, 1–10.
- Halevy, A., Norvig, P., and Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intell. Syst. 24, 8–12.
- Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L.M. (2021). "Everyone wants to do the model work, not the data work": data Cascades in High-Stakes Al. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Association for Computing Machinery), pp. 1–15. (CHI '21). https://doi.org/10. 1145/3411764.3445518.
- 72. Tideman, L.E.M., Migas, L.G., Djambazova, K.V., Patterson, N.H., Caprioli, R.M., Spraggins, J.M., and Van de Plas, R. (2021). Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized shapley additive explanations. Anal. Chim. Acta, 338522.
- Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). Classbalanced loss based on effective number of samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9268–9277.



- Lin, H., Gao, S., Gotz, D., Du, F., He, J., and Cao, N. (2017). Rclens: interactive rare category exploration and identification. IEEE Trans. Vis. Comput. Graph. 24, 2223–2237.
- Jing, L., and Tian, Y. (2019). Self-supervised visual feature learning with deep neural networks: a survey. ArXiv http://arxiv.org/abs/1902.06162.
- Zhu, X.(Jerry) (2005). Semi-Supervised Learning Literature Survey (University of Wisconsin-Madison Department of Computer Sciences). https://minds.wisconsin.edu/handle/1793/60444.
- Chapelle, O., Scholkopf, B., and Zien, A. (2009). Semi-supervised learning (Chapelle, O. et al., Eds.; 2006) [Book Reviews]. IEEE Trans. Neural Netw. 20, 542.
- Titterington, D.M., Afm, S., Smith, A.F., Makov, U., and others. (1985). Statistical Analysis of Finite Mixture Distributions (John Wiley & Sons Incorporated).
- 79. Vapnik, V. (1998). Statistical Learning Theory, 1 (Wiley. N. Y.), p. 2.
- van Engelen, J.E., and Hoos, H.H. (2020). A survey on semi-supervised learning. Mach. Learn. 109, 373–440.
- Settles, B. (2009). Active Learning Literature Survey. CS Technical Reports 67. https://minds.wisconsin.edu/handle/1793/60660.
- 82. Zhu, X., Lafferty, J., and Ghahramani, Z. (2003). Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In ICML 2003 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, *3*, T. Fawcett and N. Mishra, eds. (In proceedings of the Twentieth International Conference on Machine Learning), p. 1000.
- Lewis, D.D., and Gale, W.A. (1994). A sequential algorithm for training text classifiers. In SIGIR'94, B.W. Croft and C.J. van Rijsbergen, eds. (Springer), pp. 3–12.
- Dasgupta, S., and Hsu, D. (2008). Hierarchical sampling for active learning. In Proceedings of the 25th International Conference on Machine Learning - ICML '08 (ACM Press), pp. 208–215. http://portal.acm.org/ citation.cfm?doid=1390156.1390183.
- Duros, V., Grizou, J., Xuan, W., Hosni, Z., Long, D.-L., Miras, H.N., and Cronin, L. (2017). Human versus robots in the discovery and crystallization of gigantic polyoxometalates. Angew. Chem. Int. Ed. 56, 10815–10820.
- Bachman, P., Sordoni, A., and Trischler, A. (2017). Learning algorithms for active learning. In International Conference on Machine Learning, D. Precup and Y.W. Teh, eds. (PMLR), pp. 301–310.
- Amershi, S., Cakmak, M., Knox, W.B., and Kulesza, T. (2014). Power to the people: the role of humans in interactive machine learning. Al Mag. 35, 105–120.
- Wagstaff, K. (2012). Machine learning that matters. In Proceedings of the 29th International Conference on Machine Learning (ICML-12), J. Langford and J. Pineau, eds. (Omnipress), pp. 529–536, (ICML '12).
- Settles, B., Craven, M., and Friedland, L. (2008). Active learning with real annotation costs. In Proceedings of the NIPS Workshop on Cost-Sensitive Learning, *Vol. 1*, B.R. Krishnapuram, S. Yu, Y. Oksana, R. Bharat Rao, and L. Carin, eds. (Curran Associates).
- Settles, B. (2011). From theories to queries: active learning in practice. In Active Learning and Experimental Design Workshop in Conjunction with AISTATS 2010 (JMLR Workshop and Conference Proceedings), pp. 1–18. http://proceedings.mlr.press/v16/settles11a.html.
- Kottke, D., Schellinger, J., Huseljic, D., and Sick, B. (2019). Limitations of assessing active learning performance at runtime. CoRR, 10338, abs/1901.
- Choi, M., Park, C., Yang, S., Kim, Y., Choo, J., and Hong, S.R. (2019). AlLA: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Association for Computing Machinery), pp. 1–12. https://doi.org/10.1145/ 3290605.3300460.
- Kulesza, T., Amershi, S., Caruana, R., Fisher, D., and Charles, D. (2014). Structured labeling for facilitating concept evolution in machine learning.



In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (ACM), pp. 3075–3084. https://dl.acm.org/doi/10. 1145/2556288.2557238.

- Reker, D., Bernardes, G., and Rodrigues, T. (2018). Evolving and nano data enabled machine intelligence for chemical reaction optimization. https://chemrxiv.org/articles/Evolving\_and\_Nano\_Data\_Enabled\_Machine\_ Intelligence\_for\_Chemical\_Reaction\_Optimization/7291205/1.
- Dyk, D.A.van, and Meng, X.-L. (2001). The Art of data augmentation. J. Comput. Graph. Stat. 10, 1–50.
- 96. Giridhara, P., Mishra, C., Venkataramana, R., Bukhari, S., and Dengel, A. (2019). A study of various text augmentation techniques for relation classification in free text. In Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (SCITEPRESS Science and Technology Publications), pp. 360–367. http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0007311003600367.
- 97. Wang, W.Y., and Yang, D. (2015). That's so Annoying!!!: a lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 2557–2563. https://www.aclweb.org/anthology/D15-1306.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27, 2672–2680.
- Antoniou, A., Storkey, A., and Edwards, H. (2018). Augmenting image classifiers using data augmentation generative adversarial networks. In International Conference on Artificial Neural Networks (Springer), pp. 594–603.
- 100. Yi, X., Walia, E., and Babyn, P. (2019). Generative adversarial network in medical imaging: a review. Med. Image Anal. *58*, 101552.
- 101. Gupta, L., Klinkhammer, B.M., Boor, P., Merhof, D., and Gadermayr, M. (2019). GAN-based image enrichment in digital pathology boosts segmentation accuracy. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 631–639.
- Pratt, L.Y. (1993). Discriminability-based transfer between neural networks. In Advances in Neural Information Processing Systems 5, S.J. Hanson, J.D. Cowan, and C.L. Giles, eds. (Morgan-Kaufmann), pp. 204–211. http://papers.nips.cc/paper/641-discriminability-basedtransfer-between-neural-networks.pdf.
- 103. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. In Bioinformatics, J. Wren, ed., p. btz682.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: pretraining of deep bidirectional transformers for language understanding. ArXiv http://arxiv.org/abs/1810.04805.
- 105. Deng, L. (2012). The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal. Process. Mag. 29, 141–142.
- 106. Asgari, E., and Mofrad, M.R.K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One 10, e0141287.
- 107. Kim, D., Cho, D., and Kweon, I.S. (2019). Self-supervised video representation learning with space-time cubic puzzles. Proc. AAAI Conf. Artif. Intell. 33, 8545–8552.
- Kolesnikov, A., Zhai, X., and Beyer, L. (2019). Revisiting self-supervised visual representation learning. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 1920–1929. https://ieeexplore.ieee.org/document/8953672/.
- 109. Korbar, B., Tran, D., and Torresani, L. (2018). Cooperative learning of audio and video models from self-supervised synchronization. In Advances in Neural Information Processing Systems, 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds. (Curran Associates, Inc.), pp. 7763–7774. http://papers.nips.cc/paper/ 8002-cooperative-learning-of-audio-and-video-models-from-selfsupervised-synchronization.pdf.

### **Patterns**

Review

- 110. Mahendran, A., Thewlis, J., and Vedaldi, A. (2019). Cross pixel opticalflow similarity for self-supervised learning. In Computer Vision – ACCV 2018, C.V. Jawahar, H. Li, G. Mori, and K. Schindler, eds. (Springer International Publishing), pp. 99–116, (Lecture Notes in Computer Science).
- 111. Owens, A., and Efros, A.A. (2018). Audio-visual scene analysis with selfsupervised multisensory features, p. 631–648 https://openaccess. thecvf.com/content\_ECCV\_2018/html/Andrew\_Owens\_Audio-Visual\_ Scene\_Analysis\_ECCV\_2018\_paper.html.
- 112. Sayed, N., Brattoli, B., and Ommer, B. (2019). Cross and learn: crossmodal self-supervision. In Pattern Recognition, T. Brox, A. Bruhn, and M. Fritz, eds. (Springer International Publishing), pp. 228–243, (Lecture Notes in Computer Science).
- 113. Doersch, C., Gupta, A., and Efros, A.A. (2015). Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1422–1430.
- 114. Noroozi, M., and Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In Computer Vision ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, eds. (Springer International Publishing), pp. 69–84, (Lecture Notes in Computer Science).
- 115. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A.A. (2016). Context encoders: feature learning by inpainting, p. 2536–2544 https://openaccess.thecvf.com/content\_cvpr\_2016/html/Pathak\_Context\_ Encoders\_Feature\_CVPR\_2016\_paper.html.
- **116.** Zhang, R., Isola, P., and Efros, A.A. (2016). Colorful image colorization. In European Conference on Computer Vision (Springer), pp. 649–666.
- 117. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. ArXiv 1310, arXiv:1310.4546.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301. 3781v3.
- Caron M., Touvron H., Misra I., Jégou H., Mairal J., Bojanowski P., Joulin A.. Emerging properties in self-supervised vision transformers. ArXiv Prepr. ArXiv210414294. 2021.
- 120. Chen, L., Zhai, Y., He, Q., Wang, W., and Deng, M. (2020). Integrating deep supervised, self-supervised and unsupervised learning for singlecell RNA-seq clustering and annotation. Genes 11, 792.
- Zhai, X., Oliver, A., Kolesnikov, A., and Beyer, L. (2019). S4L: self-supervised semi-supervised learning. ArXiv http://arxiv.org/abs/1905.03670.
- Zheltonozhskii, E., Baskin, C., Bronstein, A.M., and Mendelson, A. (2020). Self-supervised learning for large-scale unsupervised image clustering. ArXiv http://arxiv.org/abs/2008.10312.
- 123. Gildenblat, J., and Klaiman, E. (2020). Self-supervised similarity learning for digital pathology. ArXiv http://arxiv.org/abs/1905.08139.
- Doersch, C., and Zisserman, A. (2017). Multi-task self-supervised visual learning. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2051–2060.
- 125. Oord, A.van den, Li, Y., and Vinyals, O. (2019). Representation learning with contrastive predictive coding. ArXiv http://arxiv.org/abs/ 1807.03748.
- 126. Li, F.-F., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. 28, 594–611.

- CellPress OPEN ACCESS
- 127. Miller, E.G., Matsakis, N.E., and Viola, P.A. (2000). Learning from one example through shared densities on transforms. In Proceedings IEEE Conference on Computer Vision and Pattern Recognition, *1* (CVPR 2000 (Cat. No.PR00662)), pp. 464–471.
- Larochelle, H., Erhan, D., and Bengio, Y. (2008). Zero-data learning of new tasks. In Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2 (AAAI Press), pp. 646–651, (AAAI'08).
- 129. Socher, R., Ganjoo, M., Manning, C.D., and Ng, A. (2013). Zero-shot learning through cross-modal transfer, C.J.C. Burges, L. Bottou, and M. Welling, eds. (2013). Advances in Neural Information Processing Systems, Vol. 26 (Curran Associates, Inc.). https://papers.nips.cc/paper/ 2013/hash/2d6cc4b2d139a53512fb8cbb3086ae2e-Abstract.html.
- Lee, C.-W., Fang, W., Yeh, C.-K., and Wang, Y.-C.F. (2018). Multi-label zero-shot learning with structured knowledge graphs, p. 1576–1585 https://openaccess.thecvf.com/content\_cvpr\_2018/html/Lee\_Multi-Label\_Zero-Shot\_Learning\_CVPR\_2018\_paper.html.
- 131. Rios, A., and Kavuluru, R. (2018). Few-shot and zero-shot multi-label learning for structured label spaces. Proc. Conf. Empir. Methods Nat. Lang. Process 2018, 3132–3142.
- 132. Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. Natl. Sci. Rev. 5, 44–53.
- 133. Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (Association for Computational Linguistics), pp. 1003–1011. https:// www.aclweb.org/anthology/P09-1113.
- Cabannes, V., Rudi, A., and Bach, F. (2020). Structured prediction with partial labelling through the infimum loss. In International Conference on Machine Learning, N. Lawrence and M. Reid, eds. (PMLR), pp. 1230–1239.
- 135. Mann, G.S., and McCallum, A. (2010). Generalized expectation criteria for semi-supervised learning with weakly labeled data. J. Mach. Learn. Res. *11*, 955–984.
- 136. Ratner, A.J., De Sa, C.M., Wu, S., Selsam, D., and Ré, C. (2016). Data programming: creating large training sets, quickly. In Advances in Neural Information Processing Systems, 29, D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, and R. Garnett, eds. (Curran Associates, Inc.), pp. 3567–3575. http://papers.nips.cc/paper/6523-data-programmingcreating-large-training-sets-quickly.pdf.
- 137. Zhang, J., and Wu, X. (2018). Multi-label inference for crowdsourcing. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2738–2747.
- Finn, C., Xu, K., and Levine, S. (2018). Probabilistic model-agnostic metalearning. In Advances in Neural Information Processing Systems, *31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds. (Curran Associates, Inc.), pp. 9516–9527. http://papers. nips.cc/paper/8161-probabilistic-model-agnostic-meta-learning.pdf.
- Dvornik, N., Schmid, C., and Mairal, J. (2020). Selecting relevant features from a multi-domain representation for few-shot classification. ArXiv http://arxiv.org/abs/2003.09338.
- 140. Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. (2020). Metalearning in neural networks: a survey. ArXiv.