MIT Press • Open Encyclopedia of Cognitive Science

Supervised and Unsupervised Learning

Bradley C. Love[1]

[1]Los Alamos National Laboratory

MIT Press

Published on: Jan 22, 2026

DOI: https://doi.org/10.21428/e2759450.e4a8e626

Unsupervised learning is learning that occurs in the absence of feedback from an external teacher, which can be contrasted with *supervised learning*, in which an external teacher provides corrective labels during learning. Unsupervised learning is important across domains in cognitive science such as vision, language, and categorization. Although the distinction between supervised and unsupervised learning seems clear, closer examination suggests learning tasks fall along a continuum between these two approaches.

History

Unsupervised learning, the discovery of patterns and structure without explicit labels or teaching signals, has a rich history in both psychology and artificial intelligence. In psychology, early evidence came from Tolman and Honzik's (1930) latent learning experiments in which rats explored mazes without reinforcement for several days but, once food rewards were introduced, immediately demonstrated that they had learned the maze layout during the unrewarded trials (Tolman, 1948). In the 1960s, Arthur Reber's (1967) artificial grammar learning studies demonstrated that people could implicitly extract complex rules from unlabeled letter sequences, establishing that humans acquire structured knowledge without awareness of what they learned. In the 1990s, Saffran et al. (1996) found that infants could segment words from continuous speech by tracking transitional probabilities between syllables [see Statistical Learning]. Clapper and Bower's (1994) work on category invention demonstrated that people seize on presentation order to discover category structure in an unsupervised fashion rather than passively tracking correlations.

In parallel, artificial intelligence (AI) research developed formal methods for discovering structure; the term “k-means” (referring to a set of k clusters with distinct mean values) was first used by J. MacQueen (1967) and was followed by techniques like principal component analysis and Teuvo Kohonen's (1982) self-organizing maps in the 1970s to 1980s that created topology-preserving representations. Throughout this history, the fields of AI and cognitive science have cross-pollinated; Hebbian learning principles from neuroscience inspired computational algorithms, and AI’s clustering and dimensionality reduction techniques offered formal frameworks for testing psychological theories about how minds extract meaningful structure from experience.

Core concepts

In canonical supervised learning, the target output that the learner should predict is explicitly defined and provided by an external teacher. For instance, in a classification or category-learning task, the learner predicts a class or category label based on a stimulus input. Then, the external teacher provides the correct answer as feedback. For example, a radiology student could be presented with a mammogram, asked whether a tumor is present or not, respond, and then be told the correct answer. The goal in this case is solely to predict the presence or absence of a tumor. In effect, the learner’s task can be

formally described as calculating the probability of the class label (e.g., tumor vs. no tumor) given the features of the stimulus (Ng & Jordan, 2001).

In contrast, in purely unsupervised learning, there is no external teacher that provides a supervisory signal, and the learner is tasked with discovering the structure of the domain. For instance, the learner may use a clustering approach or dimensional reduction technique (Roads & Love, 2024) to characterize how the observable features interrelate with one another. In this case, rather than attempting to discriminate between different class labels (e.g., tumor present or not), the learner aims to discover patterns and relationships among all features without prioritizing any particular prediction task. In the mammogram example, this could correspond to clustering mammograms together based on shared visual patterns and characteristics, which could lead to the discovery of naturally occurring groupings that might or might not align with the presence of tumors.

Learning approaches can exist anywhere between these two extremes of supervised and unsupervised learning. For instance, we could view the category or class label as just another feature to predict from the other features. If the learner only cared about predicting the category label feature, then learning could be characterized as supervised. If the learner cared equally about predicting every feature, then learning could be characterized as unsupervised. Between these extremes, models can be designed to weight the importance of predicting different features to varying degrees, creating a continuous spectrum of learning approaches (Jones & Love, 2006).

For example, a Bayesian clustering model, the *rational model* (Anderson, 1991), operates along related principles [see Rational Analysis]. The rational model is unsupervised, but it can be made to organize its cluster along certain features, such as the category label, by specifying a narrower prior for these features. Another clustering model, SUSTAIN (Love et al., 2004), explicitly incorporates unsupervised and supervised learning to achieve related ends by organizing clusters around the learner's goal (i.e., what the learner is trying to predict). SUSTAIN uses unsupervised learning to update its clusters until the supervisory signal indicates an error that triggers the formation of a new cluster. By these means, SUSTAIN's internal representations (i.e., clusters) reflect both the structure of the environment and the learner's goals (Love, 2005).

Human learners also distort their internal representations in a manner consistent with error-driven learning models that seek to discriminate between responses as opposed to creating a veridical world model (Davis & Love, 2010; Ramscar et al., 2010). In models that incorporate aspects of structure discovery, such as incremental clustering models, the order of items can have strong effects on the representations acquired during learning. Likewise, human learners are strongly affected by item ordering in learning and sorting tasks (Clapper & Bower, 1994). Bayesian models that generalize Anderson's rational model can show varying degrees of sensitivity to order effects (Sanborn et al., 2010).

Self-supervised learning further blurs the distinction between unsupervised and supervised learning. In self-supervised learning, the learner creates their own teaching signal in the absence of an external

teacher. Modern large language models (LLMs) trained on massive text corpora are prominent examples of this approach (Devlin et al., 2019) [see Large Language Models]. During training, LLMs typically predict the next word based on the words in the preceding context. Notice the learning task that the LLM faces is inherently unsupervised—no teacher is instructing the model on what lessons to draw from the training text. Nevertheless, the training algorithm used by LLMs is supervised. In essence, self-supervision turns an unsupervised learning problem into a supervised learning problem. In psychology, similar approaches have been used to model how people master sequences (Gureckis & Love, 2010).

Unsupervised learning can set the stage for successful supervised learning, particularly because unsupervised training data is often more plentiful. Using data that does not require human annotation, unsupervised learning can discover useful representations that facilitate subsequent supervised learning. For example, LLMs are initially trained on next word prediction, which does not require a teacher. Later training tunes these representations using reinforcement or supervised learning procedures to ensure that LLMs follow instructions and are helpful and polite (Rafailov et al., 2023). Likewise, in object recognition tasks, unsupervised learning may initially be used to learn visual features that are then fine-tuned for the target task (Chen et al., 2020).

Questions, controversies, and new developments

For both people and models, in order for unsupervised learning to complement supervised learning, the representations it acquires must align with those needed for subsequent supervised tasks (Bröker et al., 2022, 2024; Gibson et al., 2013). Intuitively, the distinctions captured in unsupervised learning need to relate to those needed by subsequent supervised learning tasks to be useful. For example, if unsupervised clustering ignores color (i.e., items with different colors are clustered together) and color is vital for subsequent supervised learning, then unsupervised learning may hinder supervised learning. In human and machine learning in real-world scenarios at scale, unsupervised and supervised learning appear to be in concert.

Broader connections

Research into learning models—such as Bayesian models, support vector machines, and artificial neural networks—is relevant to both unsupervised and supervised learning, as are data visualization and dimensional reduction techniques. Unsupervised learning also connects with perceptual learning, in which exposure alone improves discrimination abilities without explicit feedback. In developmental psychology, statistical learning mechanisms discovered through unsupervised learning research help explain how infants acquire foundational knowledge about language, objects, and events [see Language Acquisition]. The relationship between unsupervised learning and memory consolidation is also of growing interest, as the brain may extract statistical regularities during sleep and offline periods [see Memory]. Additionally, debates about innateness versus learning in cognitive development often hinge on the power and limitations of unsupervised learning mechanisms [see Cognitive Development].

Further reading

- Bröker, F., Holt, L. L., Roads, B. D., Dayan, P., & Love, B. C. (2024). Demystifying unsupervised learning: How it helps and hurts. *Trends in Cognitive Sciences*, 28(11), 974–986. https://doi.org/10.1016/j.tics.2024.09.005
- Ng, A., & Jordan, M. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* (Vol. 14). MIT Press.
- Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440. https://doi.org/10.1007/s10994-019-05855-6

References

- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

  ↩
- Bröker, F., Holt, L. L., Roads, B. D., Dayan, P., & Love, B. C. (2024). Demystifying unsupervised learning: How it helps and hurts. *Trends in Cognitive Sciences*, *28*(11), 974–986. https://doi.org/10.1016/j.tics.2024.09.005

  ↩
- Bröker, F., Love, B. C., & Dayan, P. (2022). When unsupervised training benefits category learning. *Cognition*, *221*, 104984. https://doi.org/10.1016/j.cognition.2021.104984

  ↩
- Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. (2020). Big self-supervised models are strong semi-supervised learners. *arXiv*. https://doi.org/10.48550/ARXIV.2006.10029

  ↩
- Clapper, J. P., & Bower, G. H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *20*(2), 443–460. https://doi.org/10.1037//0278-7393.20.2.443

  ↩
- Davis, T., & Love, B. C. (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science*, *21*(2), 234–242. https://doi.org/10.1177/0956797609357712

  ↩
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://doi.org/10.48550/arXiv.1810.04805

  ↩

- Gibson, B. R., Rogers, T. T., & Zhu, X. (2013). Human semi-supervised learning. *Topics in Cognitive Science*, *5*(1), 132–172. https://doi.org/10.1111/tops.12010

  ↩
- Gureckis, T. M., & Love, B. C. (2010). Direct associations or internal transformations? Exploring the mechanisms underlying sequential learning behavior. *Cognitive Science*, *34*(1), 10–50. https://doi.org/10.1111/j.1551-6709.2009.01076.x

  ↩
- Jones, M., & Love, B. C. (2006). The emergence of multiple learning systems. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *28*. https://escholarship.org/uc/item/0rv7d3hx

  ↩
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59-69. https://doi.org/10.1007/BF00337288

  ↩
- Love, B. C. (2005). Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science*, *14*(4), 195–199. https://doi.org/10.1111/j.0963-7214.2005.00363.

  ↩
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309

  ↩
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281-297). University of California Press.

  ↩
- Ng, A., & Jordan, M. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* (Vol. 14). MIT Press.

  ↩
- Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *arXiv*. https://doi.org/10.48550/ARXIV.2305.18290

  ↩
- Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957. https://doi.org/10.1111/j.1551-

6709.2009.01092.x

↩

- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855-863. https://doi.org/10.1016/S0022-5371(67)80149-X

↩

- Roads, B. D., & Love, B. C. (2024). The dimensions of dimensionality. *Trends in Cognitive Sciences*, *28*(2), 1118-1131. https://doi.org/10.1016/j.tics.2024.07.005

↩

- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science, 274(5294), 1926-1928. https://doi.org/10.1126/science.274.5294.1926

↩

- Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167. https://doi.org/10.1037/a0020511

↩

- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189-208. https://doi.org/10.1037/h0061626

↩

- Tolman, E. C., & Honzik, C. H. (1930). Introduction and removal of reward, and maze performance in rats. *University of California Publications in Psychology*, *4*, 257-275.

↩