Constellations of Meaning: Emergent Symbolic Categorization from Labeled Data

Collaboration

Hiram Calvo
Center for Computing Research,
National Polytechnic Institute (CIC-IPN)

Prof. Albert Newen
Institute of Philosophy II, Ruhr University Bochum

1 Background and Motivation

Neuroimaging studies suggest that the human brain organizes semantic concepts into continuous and partially hierarchical representational spaces [5] (e.g. animal–mammal–cat) and use diverse strategies for categorization. Classical theories (Aristotelian) defined categories by necessary and sufficient features, but this fails to explain typicality effects and fuzzy boundaries. Prototype and family-resemblance accounts emphasize best examples and graded membership: Rosch (1975) [9] showed that categories can be represented by an "ideal member" or prototype (items more similar to that prototype are considered better examples). Wittgenstein (1953) [11] similarly noted that many categories (e.g. "game") have no single defining feature, instead exhibiting family-resemblance. Exemplar and similarity-based models suggest categorization depends on comparing new instances to stored examples.

Connectionist (distributed) models represent categories as patterns of activation across neural networks [2]. Bayesian or predictive-coding approaches cast categorization as probabilistic inference with hierarchical generative models [1]. Embodied and enactive theories argue that an agent's bodily and environmental interactions partially constitute conceptual structure [10].

Some theorists (e.g. Lakoff 1987 [7]; Gärdenfors 2000 [4]) propose that conceptual categories can be organized geometrically (as regions in a conceptual space) or as relational networks.

Modern machine learning often uses statistical clustering or deep networks for categorization, but these models are usually opaque. Meanwhile, large labeled datasets (e.g. Open Food Facts, Spotify genre tags) often contain latent structure that can be explored to derive conceptual groupings. We propose to combine these approaches: using supervised hierarchical learning together with symbolic or algebraic representations to recover interpretable structural descriptions of categories — whether as symbolic grammars, partial feature constellations, or relational overlap networks. This will allow a rigorous meta-analysis of how the discovered structures align with different theories of concepts.

2 Objectives

This project has four main goals: (1) to develop methods for extracting interpretable structural descriptions of categories from labeled data; (2) to evaluate these methods on real-world datasets; (3) to analyze the results in light of cognitive theories of concepts; and (4) to disseminate the findings through publications and software. Key objectives include:

- Formulate a hybrid symbolic learning approach that enables the bottom-up discovery of conceptual structures from labeled data — including possible emergent hierarchies, overlapping feature groupings, or other interpretable forms.

- Develop techniques for representing category structures as symbolic grammars, constellations of partially overlapping features, or relational networks, inspired by the principles of Bayesian Program Learning [6], particularly its focus on inferring generative symbolic structures from minimal data, as well as conceptual space theory [4].

- Apply these methods to datasets (e.g. food categories with nutritional features, music genres with audio/user features) and evaluate both classification performance and interpretability of the resulting structures.

- Analyze the emergent categories in light of philosophical theories: do

they reflect classical definitions, prototypical gradience, exemplar-based clusters, predictive structures, or family resemblance patterns?

- Publish results in AI and cognitive science venues and develop open-source code and visualizations to support broader dissemination and reuse.

3 Methodology

The project will proceed by iterative development of computational models alongside conceptual analysis. First, we will gather labeled datasets with hierarchical taxonomies (for example, the Open Food Facts database with food category labels, and music datasets with genre/subgenre tags from Spotify or Last.fm). For each item we will extract feature vectors (e.g. nutritional attributes for foods, audio or usage statistics for music).

The core algorithm will interleave hierarchical clustering and symbolic inference. We will use hierarchical agglomerative clustering and Formal Concept Analysis [3] to build a concept lattice consistent with the given label hierarchy. For each node (category) in the hierarchy, we will induce a symbolic description (such as a logical rule or grammar) that captures its defining features. Grammar induction techniques—analogous to Bayesian program learning [6]—will be used to generate category-defining rules or programs.

We will evaluate the learned category hierarchies both quantitatively and qualitatively. Quantitatively, we will measure classification accuracy (recovering the original labels) and generalization to held-out data. Qualitatively, we will analyze the structure of learned categories: for example, do they form sharp boundaries or graded clusters, or even distributed constellations of partially overlapping members in different dimensions — a structure we hypothesize as indicative of Familienähnlichkeit (see Section 4). We will identify prototype exemplars (e.g. cluster centroids) and examine similarity relationships among items (testing alignment with conceptual spaces [4] or exemplar models). We will also test the models' predictive performance on new data. Throughout, we will compare the computational results to the predictions of different concept theories. For instance, if the output behaves more like classical logic rules versus similarity-based clusters, or how well it matches expectations from predictive-coding theories [1].

4 Beyond Convexity: Constellations of Partial Overlap in Concept Structures

While classical and geometric models of concepts often assume convex or well-bounded structures (e.g., definable by necessary and sufficient features, or as convex regions in conceptual spaces), our approach allows for more flexible, distributed structures.

We propose that, in many real-world datasets, conceptual categories emerge not as compact sets or convex regions, but rather as networks of partially overlapping feature clusters. These resemble constellations: subsets of items that share some features with others, but not all, forming a connected structure through localized similarities. This pattern reflects the philosophical idea of *Familienähnlichkeit* (family resemblance), where no single trait is common to all members, yet all are connected through a web of resemblances.

Computationally, this can be modeled by:

- analyzing subsets of features per pair of items;
- identifying overlapping "micro-clusters" in different dimensions;
- and tracing the emergent graph structure formed by these overlaps.

This structure is not only descriptively richer, but also provides a test case for distinguishing between conceptual theories: when convex categories fail to explain the grouping, partial-overlap networks may offer a more faithful and cognitively plausible account.

5 Expected Outcomes

We anticipate the following deliverables:

- A novel computational framework (algorithm and software) for hierarchical category learning with interpretable symbolic outputs.
- Case studies on real datasets (food, music) demonstrating the extracted symbolic category representations — including rules, feature constellations, or network structures.
- Academic publications bridging AI, cognitive science, and philosophy, plus an open-source code repository.

- Philosophical insights: analysis linking the results to theories of concepts (e.g. assessing whether categories are better captured by definitions or by prototypes).
- Strengthened research ties between CIC-IPN and Ruhr University, including potential follow-up projects on concept learning and explainable AI.

6 Relevance for Theory of Concepts

This interdisciplinary project directly addresses debates in the theory of concepts. By extracting explicit category structures from data, we will test claims of various frameworks. If our method yields clear definitional rules, that supports classical concept theories; if it yields graded memberships or overlapping families, that aligns with prototype/family-resemblance views [9]. Geometric or network-like category structures would connect with conceptual spaces or relational models [4]. We will also consider whether embodied or enactive aspects play a role (e.g. do sensorimotor features dominate the learned categories [10]?). Prof. Newen's expertise ensures that our analysis will engage with cutting-edge philosophical perspectives and 4E cognition approaches [8].

7 Proposed Timeline

1. **Months 1–3:** Literature review on categorization theories and symbolic ML; data collection (food, music, etc.); initial meetings at Ruhr University; design of the algorithmic framework.
2. **Months 4–6:** Implementation of learning algorithms (hierarchical clustering + grammar induction); pilot experiments on subset of data; interim evaluation and adjustments.
3. **Months 7–9:** Full-scale experiments on entire datasets; analysis of category structures and theoretical comparison; drafting of research papers.
4. **Months 10–12:** Final refinement of methods; preparation of joint publications and reports; remote collaboration for follow-up; organization

of a seminar or workshop at Ruhr University on categorization and concept theory.

8 Roles of Collaborators

- **Hiram Calvo (CIC-IPN)**: Lead the computational development, implement algorithms, and conduct data experiments.
- **Prof. Albert Newen (Ruhr Univ.)**: Provide expertise in cognitive science and philosophy of mind, shape the theoretical analysis, and interpret results in terms of concept theories.
- **Joint research group (CIC-IPN and Ruhr Univ. students**: Assist with data collection and analysis (e.g. students can help gather datasets or evaluate models) and contribute to the philosophical interpretation of results.
- **Joint activities**: Regular meetings, co-authoring papers, and organizing a workshop at Ruhr University on concept learning and cognitive theory.

References

[1] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

[2] Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Behavioral and Brain Sciences*, 13(1):1–74, 1988.

[3] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin, 1999.

[4] Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000.

[5] Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. A continuous semantic space describes the representation of thousands of

object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

[6] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[7] George Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press, Chicago, 1987.

[8] Albert Newen, Leon De Bruin, and Shaun Gallagher, editors. *The Oxford Handbook of 4E Cognition*. Oxford University Press, 2018.

[9] Eleanor Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192–233, 1975.

[10] Lawrence A. Shapiro. Embodied cognition. Stanford Encyclopedia of Philosophy, 2021. URL: `https://plato.stanford.edu/entries/embodied-cognition/`.

[11] Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, 1953. Translated by G. E. M. Anscombe.