Overview of ML approaches to modeling cognitive neuroscience data

Barrett et al. analyzing biological and artificial neural networks: challenges with opportunities for synergy?

Background

Background: Revolution in the area of ML in form of DNNs.

Millions of parameters	No engineered features	Very high performance

Produces an analogy with neuroscience that are *opportunities for synergy in analysis*

1. At an abstract level, both fields need to answer a similar question: “how do neural networks, consisting of large numbers of interconnected elements, transform representations of stimuli across multiple processing stages so as to implement a wide range of complex computations and behaviours, such as object recognition “.
2. Both need to describe and analyze very high dimensional data

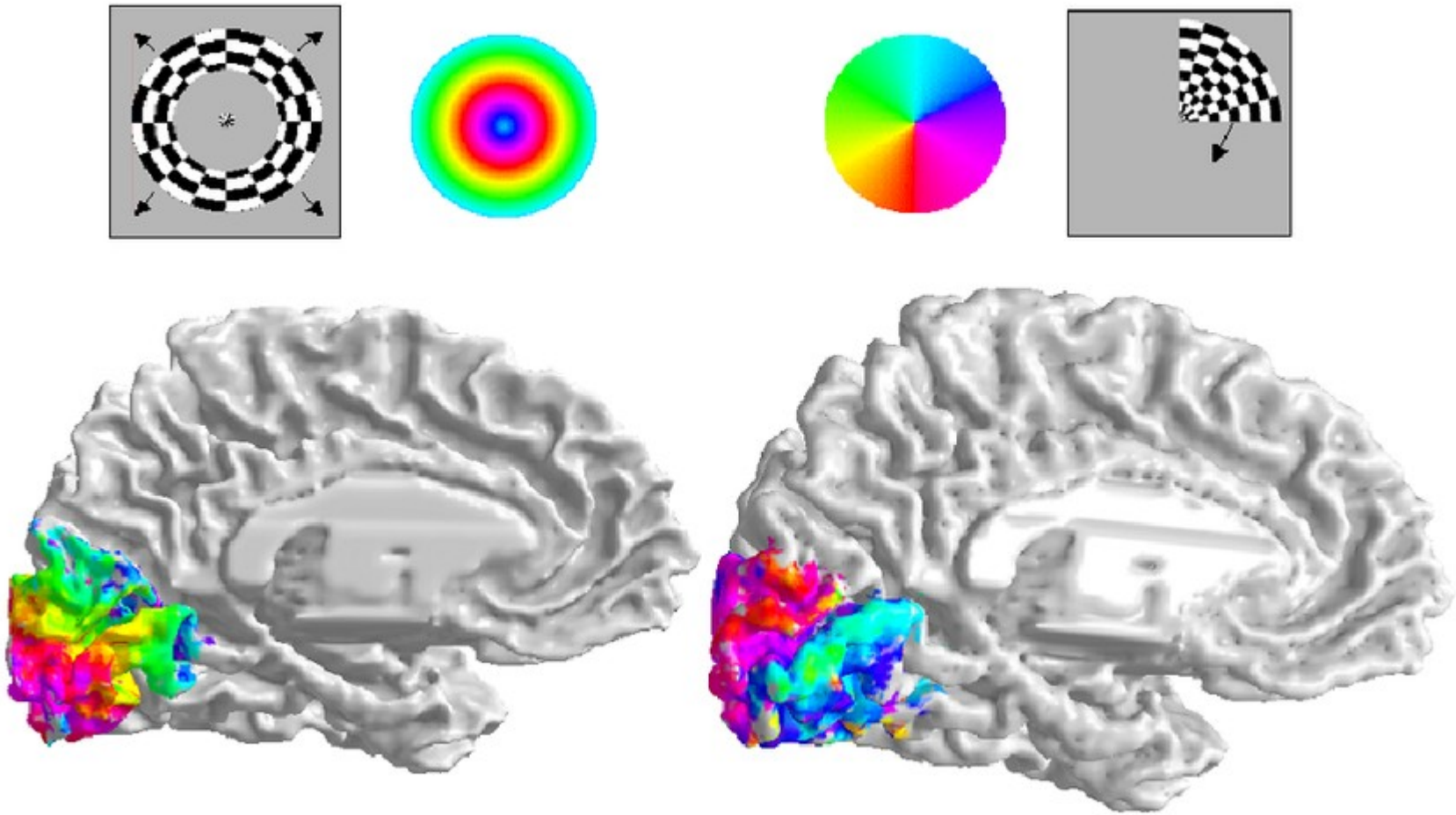
Their ‘analogies’

- Receptive fields
- Ablation
- Dimensionality reduction
- Representational geometries

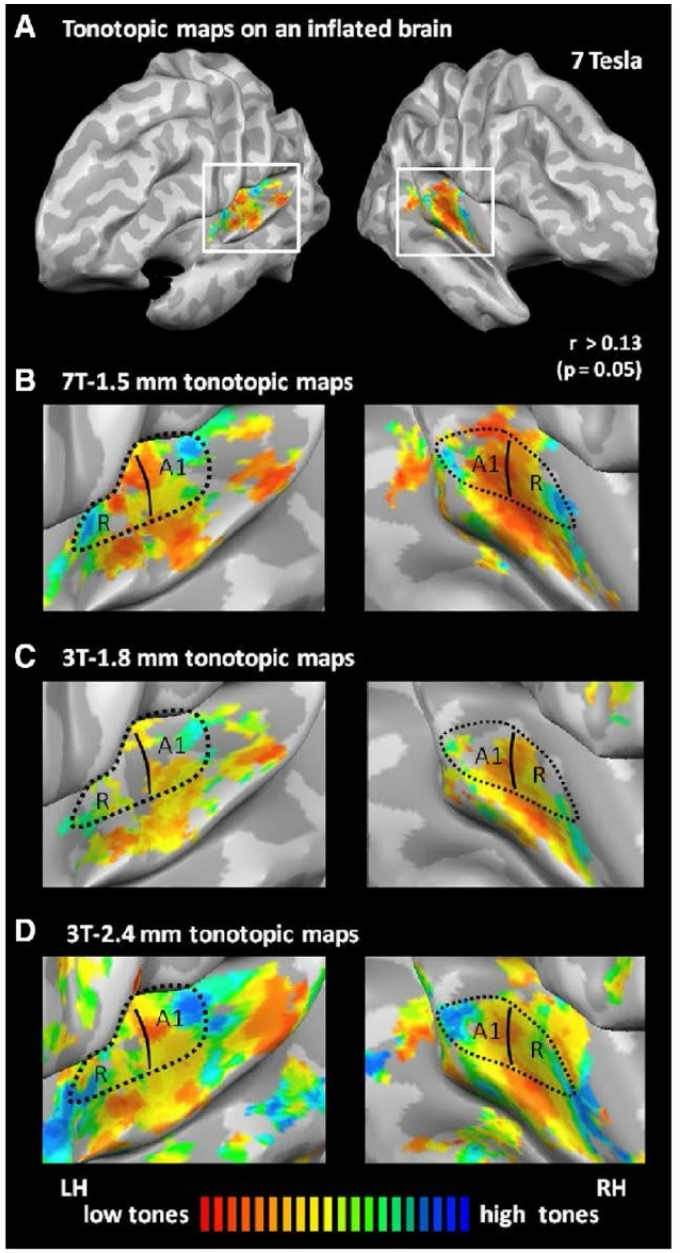
Analogy: receptive fields

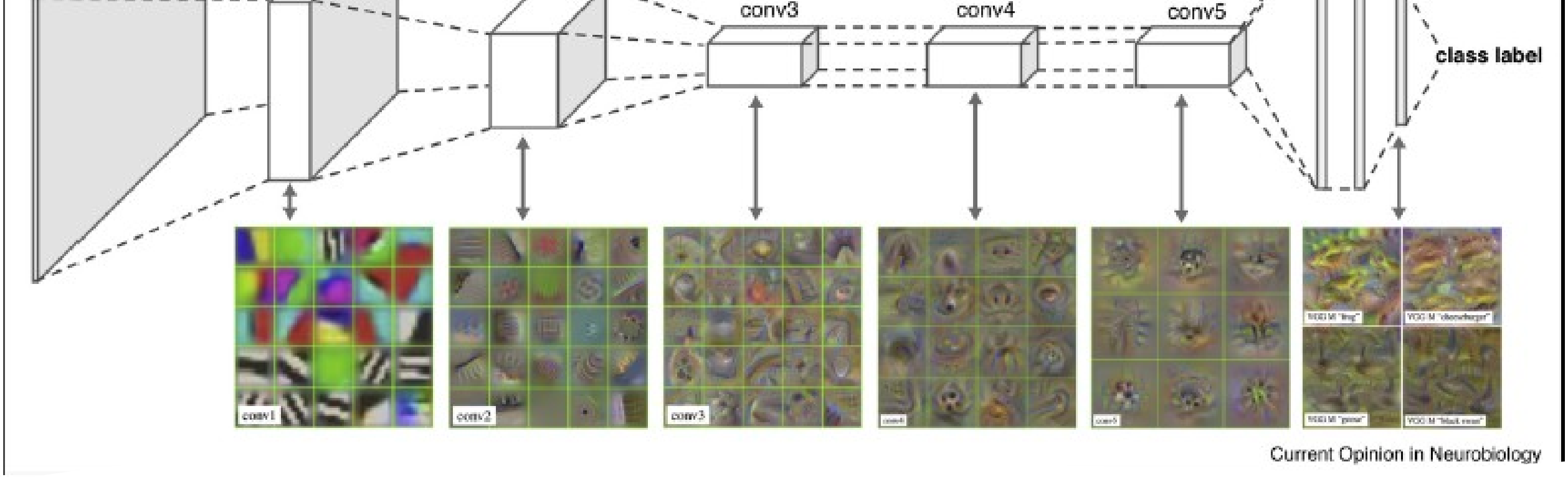
- Neurons in the human visual cortex are specialized to process stimuli in specific spatial areas (retinotopy) or certain types of features.
- The neurons in the initial processing regions of the visual cortex have small receptive fields; sensitive to stimuli in small areas of visual space.
- As information is transmitted to higher level areas of visual processing, receptive fields become larger, enabling sensitivity to larger areas of space. These regions also encode more complex features, and there is evidence of "abstract" coding with invariance to small transformations.
  - “concept cells” sensitive to identity of objects but not to appearance. Example: simple ‘repetition priming’ effect for same faces repeated exactly, or with different orientation (but not different faces)

Retinotopy map



Tonotopy map





Current Opinion in Neurobiology

Analogy: receptive fields

- (Some) AI researchers also think that DNN neurons may code for specific information, which can be studied via receptive field analysis.
- This produces experiments investigating which types of images maximally activate a neuron. It also motivates studies examining how receptive fields change in deeper layer
  - Pooling just means that the receptive fields will become larger; but in what way are they more complex?

Analogy: Ablation

Brain lesions offer much information about potential function of brain areas.

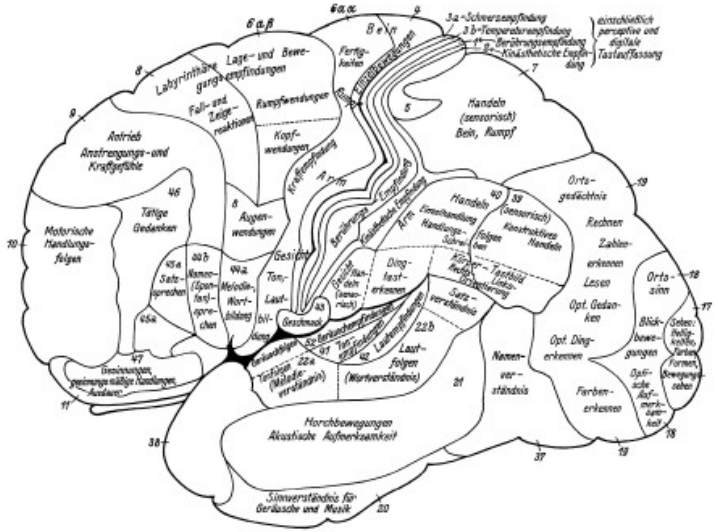
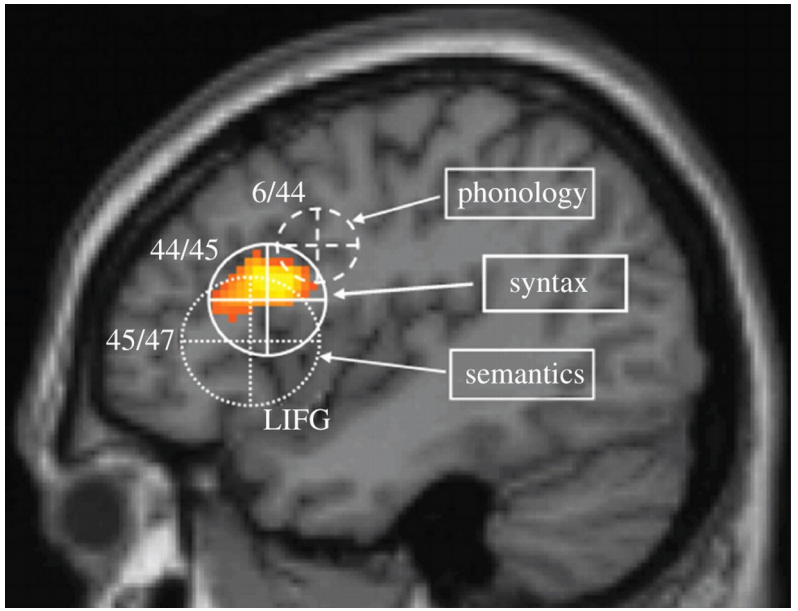
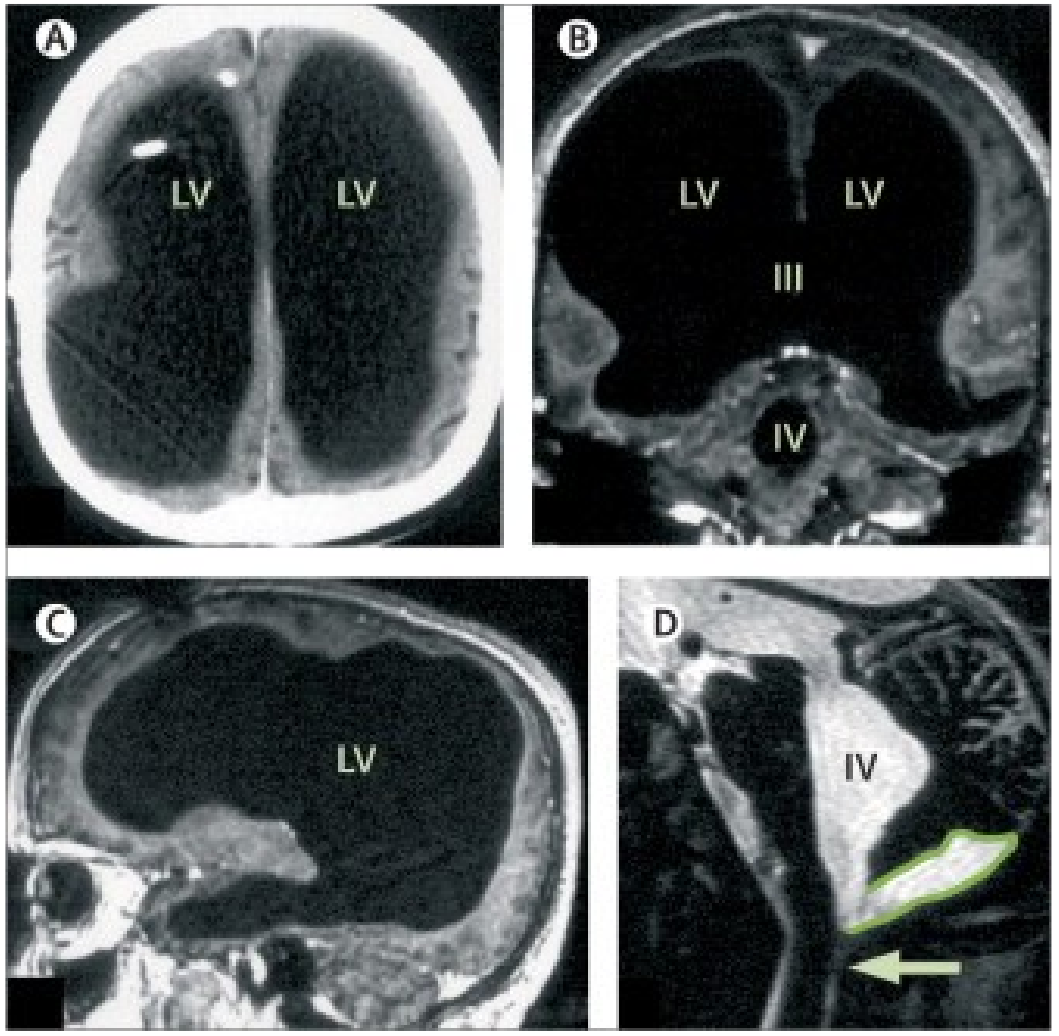


Fig. 11.5 Kleist's functional brain map. Reproduced from Kleist, K., *Gehirnpathologie*, p. 1365, © 1934, Johann Ambrosius Barth.



(And reorganization)



Analogy: Ablation

- Ablation (lesioning) analysis applicable to DNNs. “Silencing neurons” and seeing how this impacts the network output.
- Reversible or non-reversible silencing of neurons in a brain structure can be accomplished using various tools in humans.
- In DNNs: structural pruning (removing of entire neuron with all its outgoing weights)
- Networks trained for generalization (out of sample prediction) are more robust to ablation than those trained on memorizing labels.
  - The importance of a neuron is determined not by its class selectivity but perhaps by the sum of its effects
- Pruning and fine tuning are active areas of research.

Analogy: dimensionality reduction for characterizing distributed representations

- The brain codes information in a distributed manner, necessitating multivariate analysis
  - Multiple units encode information in the brain, leading to redundancy where two neurons may fire almost identically; OR
  - Information is coded in a distributed manner among multiple units (e.g., coding for 4 classes among 2 neurons, each coding 1/0); OR
  - Correlation among units could indicate that the activity can be described in a lower dimensional space.
- In DNNs: An object-by-feature matrix from fully connected layer can be compressed by more than 80% while maintaining almost all variance. This means, few low dimensions explain differences between images.

Analogy: studying representational geometries

- How are representations represented in different layers or change over time? How are different embedding spaces related to each other?
- Given and object-by-feature matrices (2 layers, 2 networks, etc') we can use Canonical Correlation Analysis or PLScorrelation to 'compare representations across networks'. These methods identify lower-level factors that capture and maximize the correlations/covariance between the datasets. (Note: These methods can be seen as "supervised" as they reweight columns in both tables to maximize similarity.)
- Representational Similarity Analysis involves comparing two similarity matrices, often constructed from object-by-feature matrices. This yields an object-by-object similarity matrix.
- Neurobiological activation vectors can be predicted from DNN embeddings using linear regression ("brain score").

Spicer and Sanborn

Methods: spatial methods, logical methods and ANNs

Spatial methods

- Spatial methods involve placing items in a multidimensional space and using their location to draw conclusions about categorization
- Classification based on spatial methods can be determined by an item's location relative to a hyperplane or its similarity to different prototypes (means) or exemplars (centroids).
  - Prototype approaches assume that learning is based on similarity to the center of a category (mean), which is stored after training
  - Exemplar approaches calculate similarity as a ratio between the similarity of an item to all items within a class (i...n) and the similarity of that item to all other items. This provides a fit per class and requires storing item-level information.
  - Clustering organizes items into groups (cohorts), with quality often quantified by the distance between items within and between clusters. Clustering can be either hard (each item belongs to only one cluster) or soft (items can have multiple memberships, potentially fuzzy).

Spatial method: example. Nosofsky 1986. Generalized Context Model (GCM)

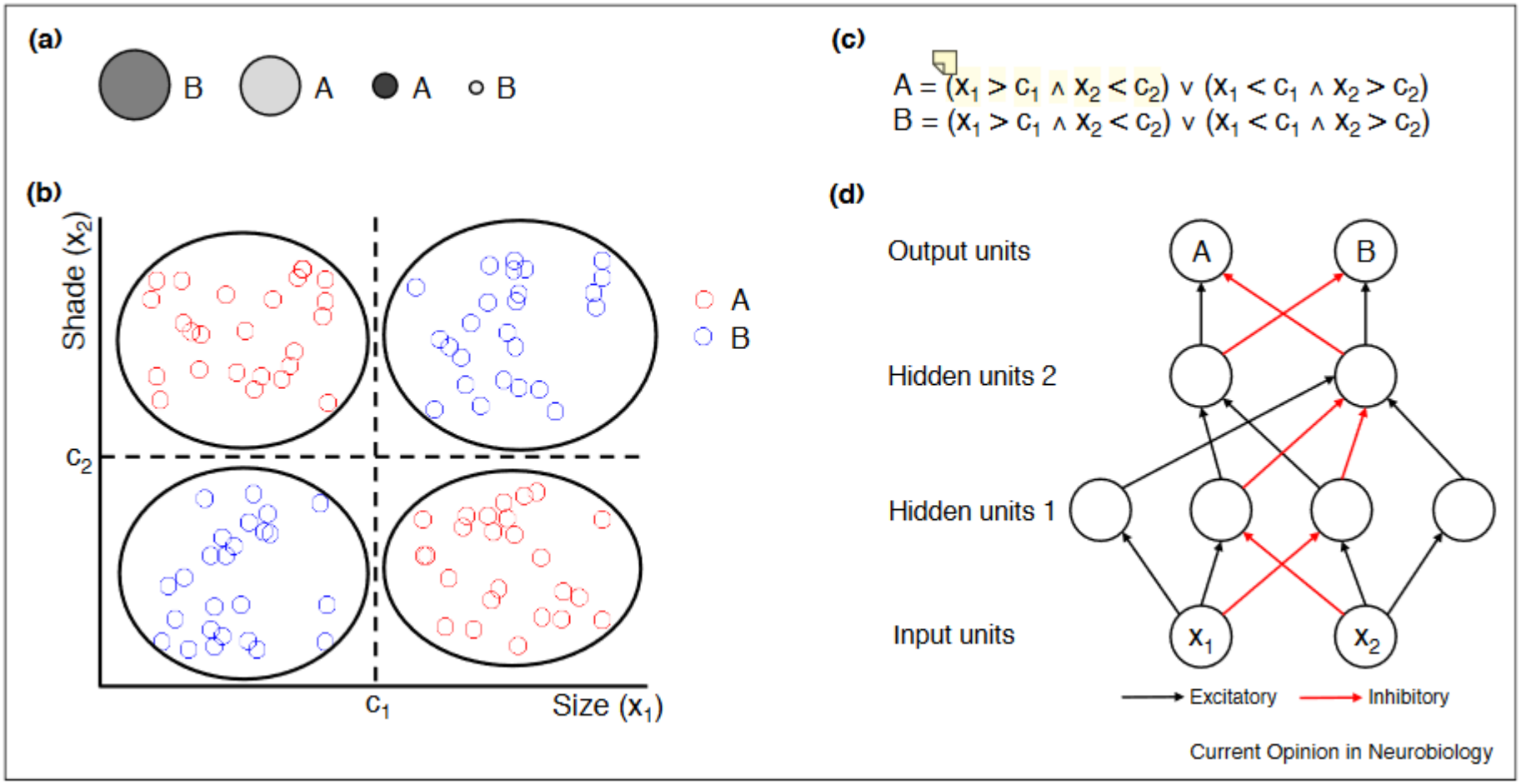
According to the GCM, the probability that stimulus i is classified into Category J (C_J) is found by summing the similarity of i to all training exemplars of C_J and then dividing by the summed similarity of i to all training exemplars of all categories:

$$P(C_J|i) = \left(\sum_{j \in J} s_{ij}\right)^{\gamma} / \sum_K \left(\sum_{k \in K} s_{ik}\right)^{\gamma} \qquad (1)$$

where s_{ij} denotes the similarity of test item i to exemplar j. The parameter γ is a response-scaling parameter that describes the degree of determinism in participants' response strategies: As γ grows larger, participants respond with higher probability with the category that yields the largest summed similarity (Ashby & Maddox, 1993; McKinley & Nosofsky, 1995; Nosofsky & Zaki, 2002; for a process-model interpretation, see Nosofsky & Palmeri, 1997). For simplicity in this article, however, we set $\gamma = 1$.

Logical methods and ANN

- Logical methods: concepts are based on a definition that is applied to the features of the object. One embodiment: search for rules that maximizes discrimination between stimuli; the rule can be probabilistic.
  - Allow compositionality (e.g., use “AND”). Draw Hard boundaries (no fuzziness).
- ANNs: do not make assumption about the representations involved, but offer an implementation method.


(a)
B
A
A
B
(b)
Shade (x_2)
c_2
c_1
Size (x_1)
A
B
(c)
$A = (x_1 > c_1 \wedge x_2 < c_2) \vee (x_1 < c_1 \wedge x_2 > c_2)$
$B = (x_1 > c_1 \wedge x_2 < c_2) \vee (x_1 < c_1 \wedge x_2 > c_2)$
(d)
Output units
Hidden units 2
Hidden units 1
Input units
A
B
x_1
x_2
Excitatory
Inhibitory
Current Opinion in Neurobiology

Points for thought.

- Is it useful to ask which model is the most accurate?
  - The authors: focus on whether they offer “useful explorations of the ways in which human learning operates”. The answer you prefer will depend on what area of science you work in.
- Value depends not just on match to human behavior, but whether there is a need to understand the underlying representations.
  - Not just accuracy; but what confusions occur