CSC321: 2011
Introduction to Neural Networks and Machine Learning

Lecture 5: Distributed Representations

Geoffrey Hinton

Localist representations

- The simplest way to represent things with neural networks is to dedicate one neuron to each thing.
  - Easy to understand.
  - Easy to code by hand
    - Often used to represent inputs to a net
  - Easy to learn
    - This is what mixture models do.
    - Each cluster corresponds to one neuron
  - Easy to associate with other representations or responses.
- But localist models are very inefficient whenever the data has componential structure.

Examples of componential structure

- Big, yellow, Volkswagen
  - Do we have a neuron for this combination
    - Is the BYV neuron set aside in advance?
    - Is it created on the fly?
    - How is it related to the neurons for big and yellow and Volkswagen?
- Consider a visual scene
  - It contains many different objects
  - Each object has many properties like shape, color, size, motion.
  - Objects have spatial relationships to each other.

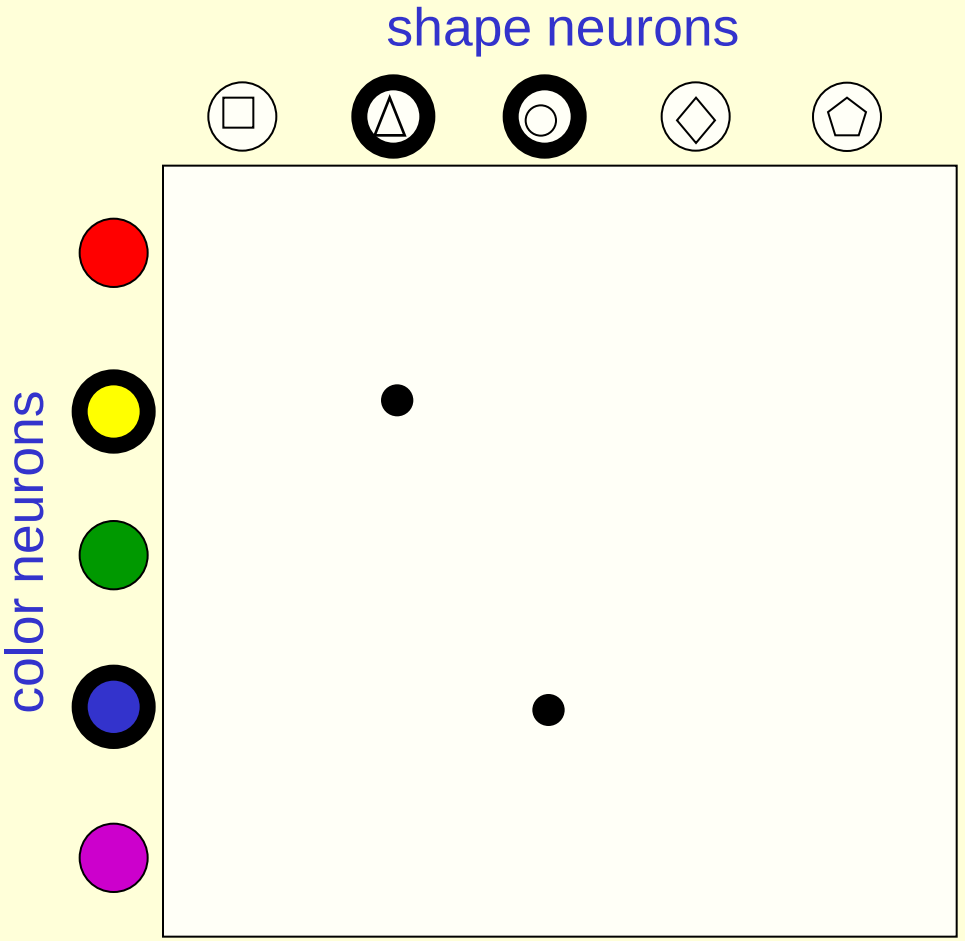
Using simultaneity to bind things together

Represent conjunctions by activating all the constituents at the same time.

- This doesn't require connections between the constituents.
- But what if we want to represent yellow triangle and blue circle at the same time?

Maybe this explains the serial nature of consciousness.

- And maybe it doesn't!



Using space to bind things together

- Conventional computers can bind things together by putting them into neighboring memory locations.
  - This works nicely in vision. Surfaces are generally opaque, so we only get to see one thing at each location in the visual field.
    - If we use topographic maps for different properties, we can assume that properties at the same location belong to the same thing.

The definition of “distributed representation”

- Each neuron must represent something, so this must be a local representation.
- “Distributed representation” means a many-to-many relationship between two types of representation (such as concepts and neurons).
  - Each concept is represented by many neurons
  - Each neuron participates in the representation of many concepts

Coarse coding

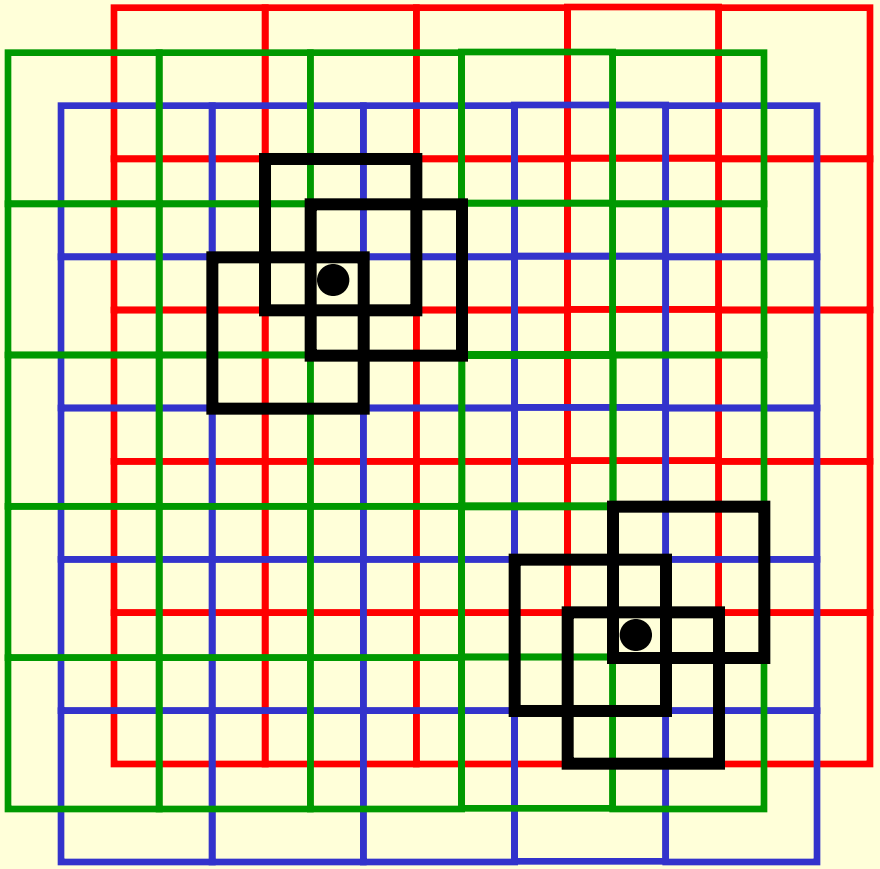
- Using one neuron per entity is inefficient.
  - An efficient code would have each neuron active half the time.
    - This might be inefficient for other purposes (like associating responses with representations).
- Can we get accurate representations by using lots of inaccurate neurons?
  - If we can it would be very robust against hardware failure.

Coarse coding

Use three overlapping arrays of large cells to get an array of fine cells

- If a point falls in a fine cell, code it by activating 3 coarse cells.

- This is more efficient than using a neuron for each fine cell.
  - It loses by needing 3 arrays
  - It wins by a factor of 3x3 per array
  - Overall it wins by a factor of 3



How efficient is coarse coding?

- The efficiency depends on the dimensionality
  - In one dimension coarse coding does not help
  - In 2-D the saving in neurons is proportional to the ratio of the fine radius to the coarse radius.
  - In k dimensions , by increasing the radius by a factor of r we can keep the same accuracy as with fine fields and get a saving of:

$$saving = \frac{\#\ fine\ neurons}{\#\ coarse\ neurons} = r^{k-1}$$

Coarse regions and fine regions use the same surface

- Each binary neuron defines a boundary between k-dimensional points that activate it and points that don't.
  - To get lots of small regions we need a lot of boundary.

fine coarse

$$total\ boundary = cnr^{k-1} = CNR^{k-1}$$

saving in neurons without loss of accuracy

$$\frac{n}{N} = \left(\frac{C}{c}\right)\left(\frac{R}{r}\right)^{k-1}$$

ratio of radii of fine and coarse fields

constant

Limitations of coarse coding

- It achieves accuracy at the cost of resolution
  - Accuracy is defined by how much a point must be moved before the representation changes.
  - Resolution is defined by how close points can be and still be distinguished in the represention.
    - Representations can overlap and still be decoded if we allow integer activities of more than 1.
- It makes it difficult to associate very different responses with similar points, because their representations overlap
  - This is useful for generalization.
- The boundary effects dominate when the fields are very big.

Coarse coding in the visual system

- As we get further from the retina the receptive fields of neurons get bigger and bigger and require more complicated patterns.
  - Most neuroscientists interpret this as neurons exhibiting invariance.
  - But its also just what would be needed if neurons wanted to achieve high accuracy
  - For properties like position orientation and size.
- High accuracy is needed to decide if the parts of an object are in the right spatial relationship to each other.

The Effects of Brain Damage

- Performance deteriorates in some unexpected ways when the brain is damaged. This can tell us a lot about how information is processed.
  - Damage to the right hemisphere can cause neglect of the left half of visual space and a lack of a sense of ownership of body parts.
  - Damage to parts of the infero-temporal cortex can prevent face recognition.
  - Damage to other areas can destroy the perception of color or of motion.
- Before brain scans, the performance deficits caused by physical damage were the main way to localize functions in the human brain
  - recording from human brain cells is not usually allowed (but it can give surprising results!).

Acquired dyslexia

- Occasionally, damage to the brain of an adult causes bizarre reading deficits
  - Surface dyslexics can read regular nonsense words like “mave” but mispronounce irregular words like “yacht”.
  - Deep dyslexics cannot deal with nonsense words at all. They can read “yacht” correctly sometimes but sometimes misread “yacht” as “boat”. They are also much better at concrete nouns than at abstract nouns (like “peace”) or verbs.

Some weird effects

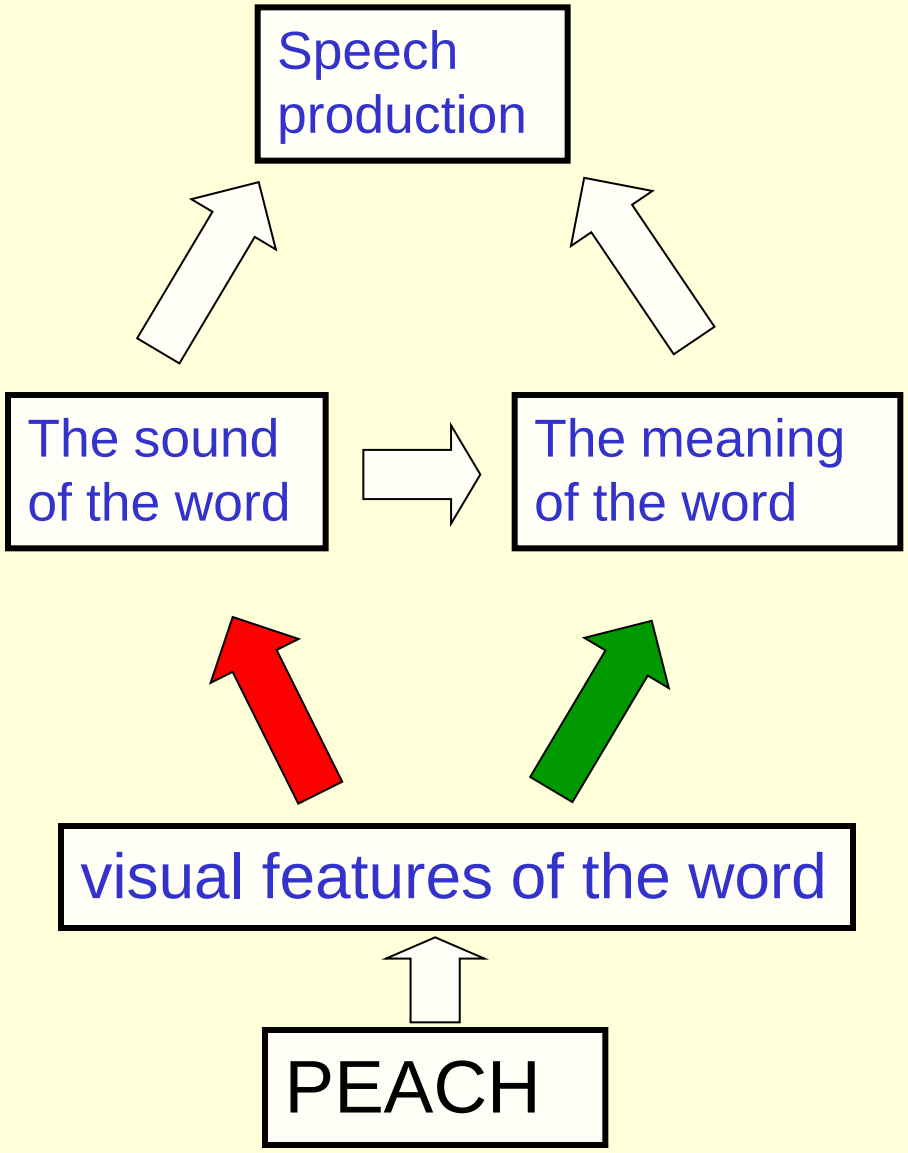
PEACH → “apricot”

SYMPATHY → “orchestra”

HAM → “food”

The dual route theory of reading

- Marshall and Newcombe proposed that there are two routes that can be separately damaged.
  - Deep dyslexics have lost the phonological route and may also have damage to the semantic route.
- But there are consistent peculiarities that are hard to explain this way.



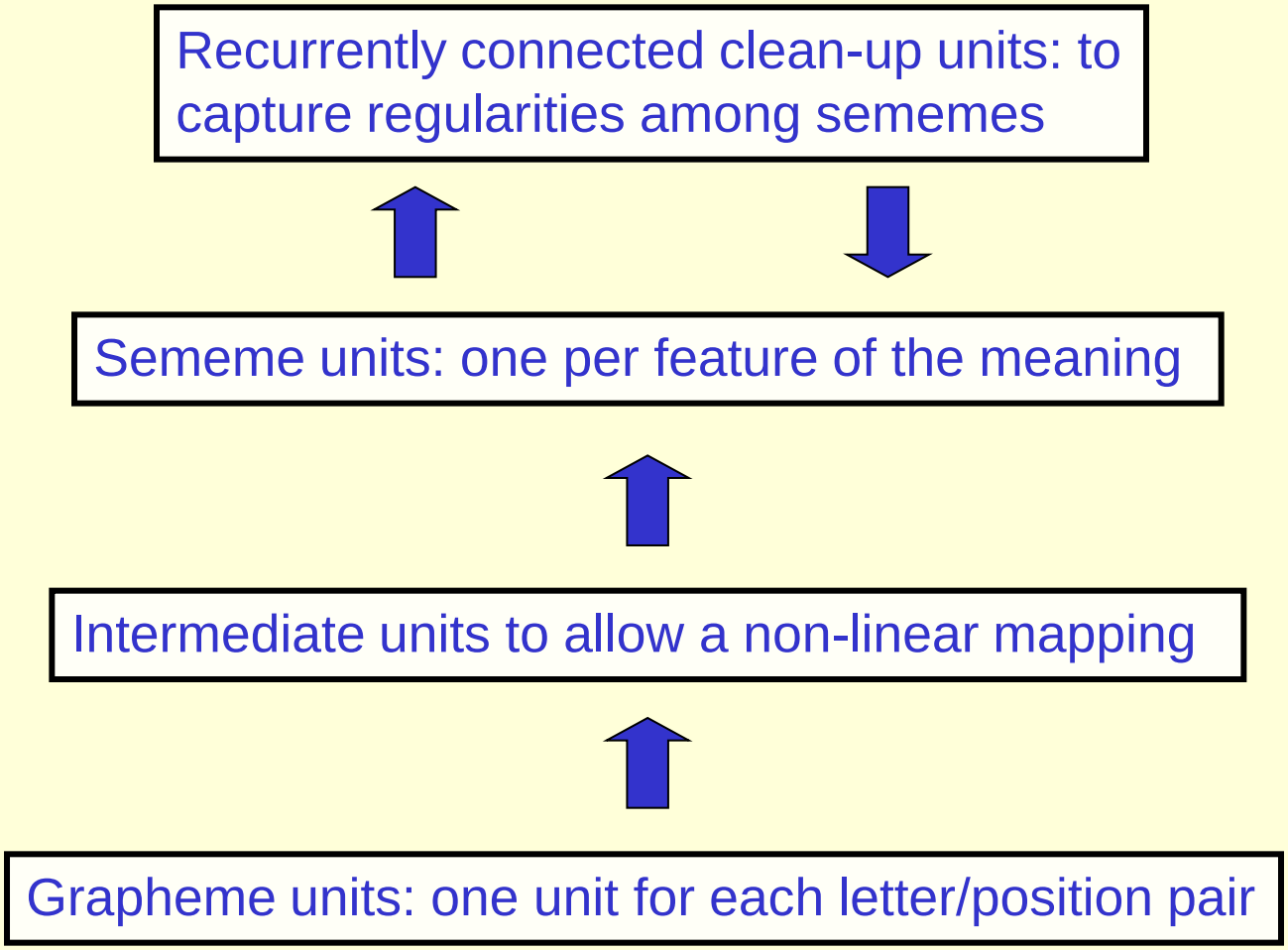
An advantage of neural network models

- Until quite recently, nearly all the models of information processing that psychologists used were inadequate for explaining the effects of damage.
  - Either they were symbol processing models that had no direct relationship to hardware
  - Or they were just vague descriptions that could not actually do the information processing.

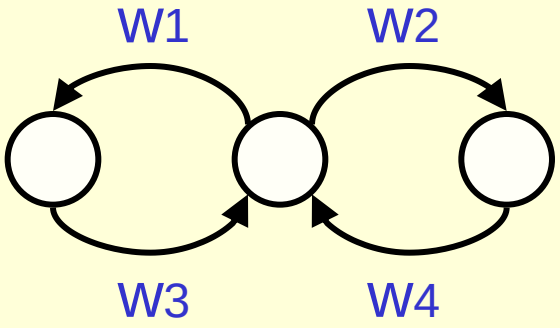
- There is no easy way to make detailed predictions of how hardware damage will affect performance in models of this type.

- Neural net models have several advantages:
  - They actually do the required information processing rather than just describing it .
  - They can be physically damaged and the effects can be observed.

A model of the semantic route

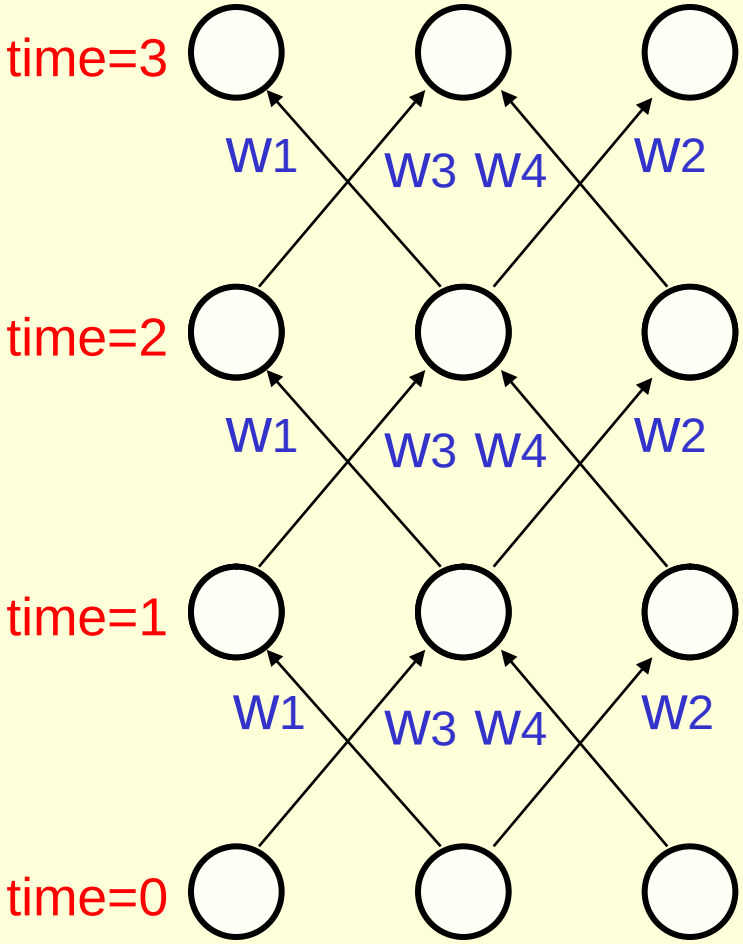


The equivalence between layered, feedforward nets and recurrent nets



Assume that there is a time delay of 1 in using each connection.

The recurrent net is just a layered net that keeps reusing the same weights.

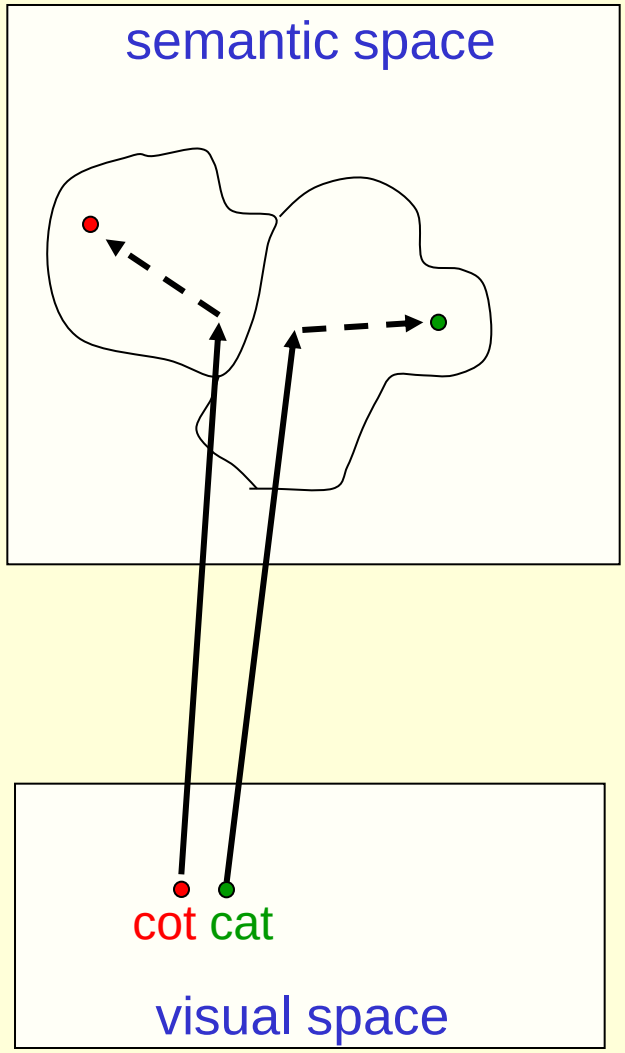


What the network learns

- We used recurrent back-propagation for six time steps with the sememe vector as the desired output for the last 3 time steps.
  - The network creates semantic attractors.
  - Each word meaning is a point in semantic space and has its own basin of attraction.
  - Damage to the sememe or clean-up units can change the boundaries of the attractors.
    - This explains semantic errors. Meanings fall into a neighboring attractor.
  - Damage to the bottom-up input can change the initial conditions for the attractors.
    - This explains why early damage can cause semantic errors.

Sharing the work between attractors and the bottom-up pathway

- Feed-forward nets prefer to produce similar outputs for similar inputs.
  - Attractors can be used to make life easy for the feed-forward pathway.
- Damaging attractors can cause errors involving visually similar words.
  - This explains why patients who make semantic errors always make some visual errors as well.
  - It also explains why errors that are both visually and semantically similar are particularly frequent.



Can very different meanings be next to each other in semantic space?

Take two random binary vectors in a high-dimensional space.

- Their scalar product depends on the fraction of the bits that are on in each vector.
- The average of two random binary vectors is much closer to them than to other random binary vectors!

1 1 0 0 0 0 1 1 0 1 0

0 1 1 0 0 1 1 0 1 0 0

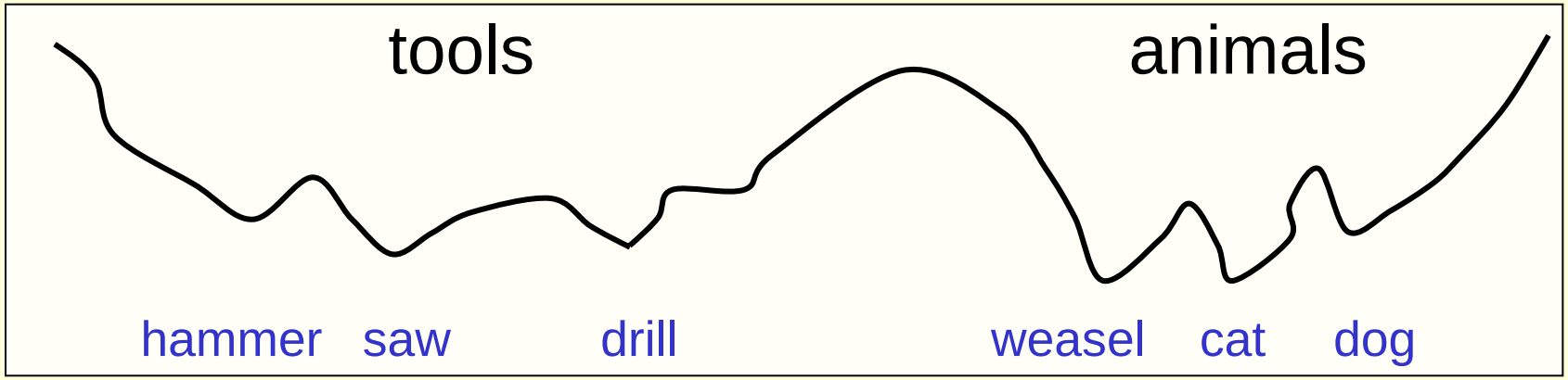
$$\mathbf{a}'\mathbf{c} \approx N\, p_{\mathbf{a}} p_{\mathbf{c}} = N\left(\frac{p_{\mathbf{a}} p_{\mathbf{c}}}{2} + \frac{p_{\mathbf{a}} p_{\mathbf{c}}}{2}\right)$$

$$\mathbf{a}'\left(\frac{\mathbf{a}}{2} + \frac{\mathbf{b}}{2}\right) \approx N\left(\frac{p_{\mathbf{a}}}{2} + \frac{p_{\mathbf{a}} p_{\mathbf{b}}}{2}\right)$$

$$p_{\mathbf{a}} > p_{\mathbf{a}} p_{\mathbf{a}}$$

Fractal attractors?

- The semantic space may have structure at several different scales.
  - Large-scale structure represents broad categories.
  - Fine-scale structure represents finer distinctions
- Severe damage could blur out all the fine structure
  - Meanings get cleaned-up to the meaning of the broad category. Complex features get lost.
  - May explain regression to childhood in senility?



The advantage of concrete words

- We assume that concrete nouns have many more semantic features than abstract words.
  - So they can benefit much more from the semantic clean-up. The right meaning can be recovered even if the bottom-up input is severely damaged.
- But severe damage to the semantic part of the network will hurt concrete nouns more because they are more reliant on the clean-up.