

# Probabilistic Models of the Visual Cortex Fall 2020

Alan Yuille

Note: This lecture had 14 parts in 2019,  
but some parts are removed for this online version.

# Part 1. Purpose of this course:

## Towards a Unified Theory of Vision for both Artificial and Natural Intelligence (AI & NI).

- The relationship between the study of AI & NI is fascinating. AI is booming, but many scientists think that taking AI to the level of human performance will require better understanding of NI
- Marr and Poggio conjectured (1978) that AI & NI could be studied jointly by distinguishing between three different levels of analysis (i) computational, (ii) algorithmic, (iii) hardware.
- Connectionists speculated that AI & NI could be modelled together.
- The goal of unifying AI and NI should yield: (i) scientific understanding of intelligence and (ii) the development of more advanced AI systems that can augment human intelligence, cooperate with humans, and in some situations replace humans.

# Part 2. What is Vision?

- To extract information from the environment in order to take action. To estimate the physical properties of the 3D world from light rays that reach our eyes (or cameras).
- These physical properties vary from coarse interpretations of an image (e.g., a horse in a field) , to more detailed (e.g., describing the hair on the horse, is the horse young or old, sick or healthy, and what is it doing).
- Images are generated from the 3D physical world. Computer graphics described how they are generated in terms of the lighting, the geometry and material properties of objects.
- Vision can be subdivided – for ease of study – into many different tasks (object recognition, object detection, depth estimation), but these sub-divisions are “fictions”, and the human visual systems performs many tasks at the same time.
- Vision is arguably the full AI problem (Poggio). It starts with processing images but also involves language, reasoning, analogy, action, and almost all aspects of intelligence.

## Part 2: What is Vision?

The more you look the more you see.

- Humans can extract a lot of information from a single image.
- “There is a fox in the garden” (coarse).
- “There is a young fox emerging from behind the base of a tree not far from the view point, it is heading right, stepping through short grass, and moving quickly. Its body fur is fluffy, reddish-brown, light in color, but with some variation. It has darker colored front legs and a dark patch above the mouth. Most of the body hairs flow from front to back.” (detailed).



# Part 2. What is Vision?

## The Full AI problem

- Understanding of objects, scenes, and events. Describing them in language.
- Reasoning about the functions and roles of objects, the goals and intentions of agents, and predicting the outcomes of events

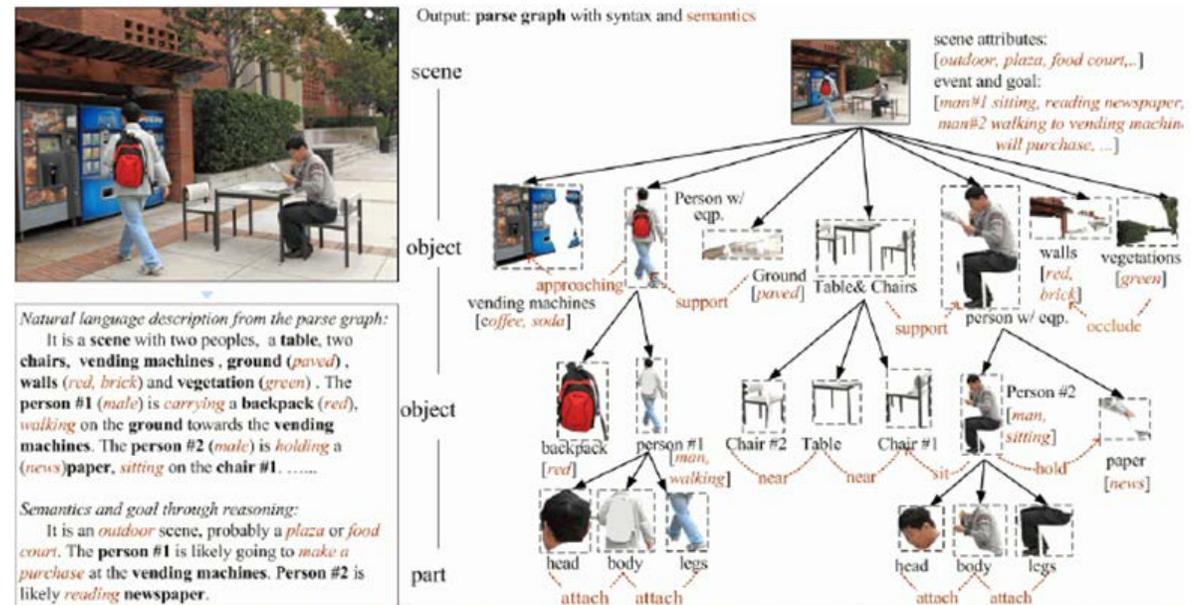


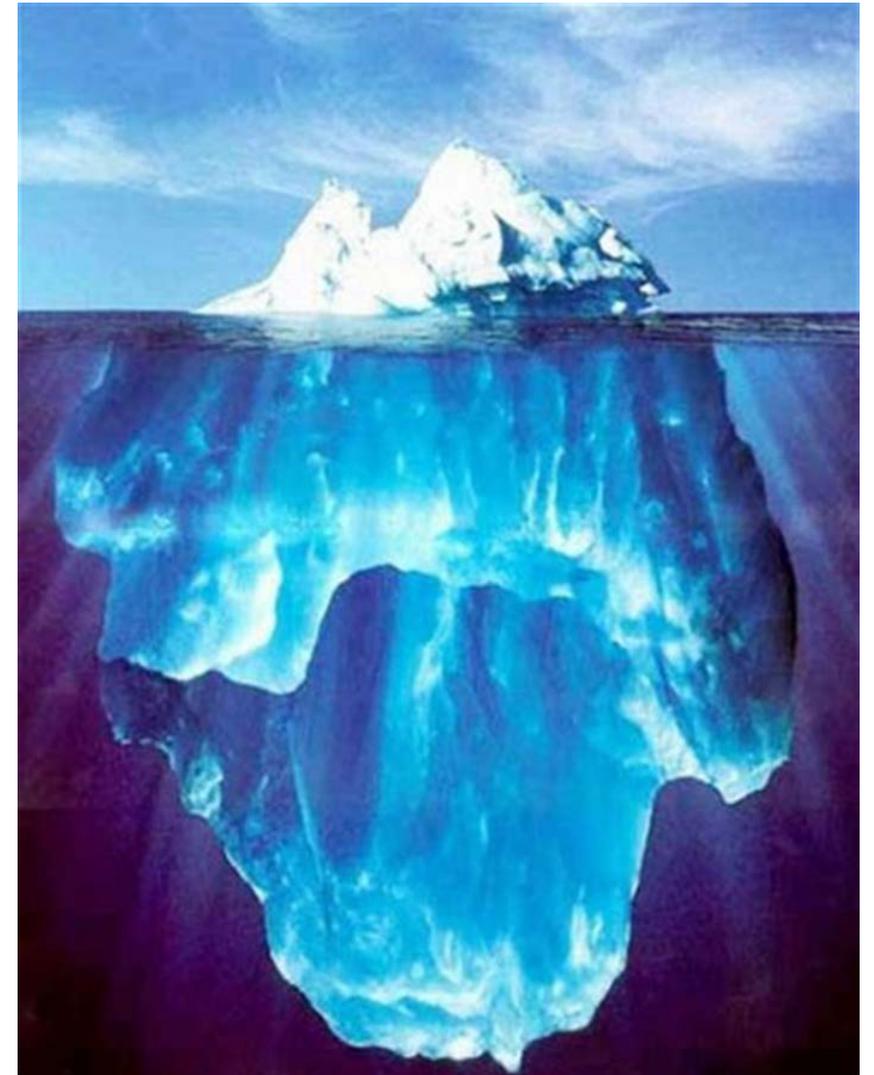
Figure 1. Example of image understanding. Analysis of the image (top-left) produces a parse graph (right) representing hierarchically objects, contextual relations, and semantic associations (in italic orange font) for attributes, functions, roles, and intents. The parse graph maybe converted to a description in natural language (bottom-left).

# Part 4A. Why is Vision Hard? Complexity.

- Vision is extremely hard due to complexity and ambiguity.
- The complexity of the set of all images. The set of images is infinite. If we restrict each pixel value to take 256 possible values (as in a digital camera), then there are more  $10 \times 10$  images than have been seen by all mankind over all history and pre-history. Humans see at most  $10^9$  each year.
- The complexity due to physical viewing conditions. For a single object – there are 13 viewing factors – and if we allow 1,000 values for each dimension, then we reach  $10^{39}$  images for a single object.
- The complexity of scene compositions. A scene can be composed in a combinatorial number of ways – placing  $N$  possible objects into  $M$  possible positions – yielding  $M^N$  possible ways to build a scene (more if you also include lighting, texture patterns, material).
- The complexity increases further for image sequences.

# Part 4A. Why is Vision Hard? Complexity.

- The set of images in any dataset are only an infinitesimal fraction of all images. The tip of the iceberg.
- Image of a single object is a function of 13 parameters – camera pose (4), Lighting (4), material (1), scene (3).



# Part 4A. Why is Vision Hard? Complexity.

- This combinatorial complexity puts a challenge on machine learning methods, like deep networks.
- Machine learning assumes that we have training and testing datasets which are big enough to be representative of the underlying problem domain. Otherwise the methods will be biased to the datasets and will perform badly on rare events (those underrepresented in the datasets).
- But if the problem domain is combinatorially complex, like some computer vision tasks, then it is impossible to have training and testing datasets which are big enough.
- This gives new challenges -- How to train models, if your datasets are too small to be unrepresentative of the real world? How to test models and guarantee performance if you can only test on a tiny fraction of possible images?
- The Human Visual system knows how to do this.

## Part 4B. Why is Vision Hard? Ambiguity.

- There are several types of ambiguity.
- Ambiguity in how images are generated from the 3D world:  
Images are functions of the geometry and material properties of the objects (and the lighting). This can be ambiguous. Sometimes we can confuse material properties for geometry. And geometry for material properties (Demo later).
- Ambiguity without context – images are often locally ambiguous and need context to disambiguate them.

# Part 4B. Why is Vision Hard?

- Ambiguity – geometry, material properties, lighting. C. von der Malsburg.



Felice Varini

# Part 4B. Why is Vision Hard

- Ambiguity – Toyota Video – C. von der Malsburg.



# Part 4B. Why is Vision Hard. Local Ambiguity of Images

Airplane  
Car  
Boat  
Sign  
Building



# Part 5. How can Humans do Vision?

- Humans and Primate devote an enormous amount of neural resources to do vision. Roughly 40-60% of the cortex is involved in vision.
- The human/primate visual system is extremely complex.
- The number of neurons is enormous (the number of trees in the Amazon rain forest) and the number of connections between neurons is even bigger (the number of leaves in the Amazon rainforest).
- Real neurons, and neural circuits, are also very complex. They have many different types of cells (but majority are pyramidal) with a large diversity of morphology (shapes). They have complex dendritic structures, possible internal states (in the cell body), and ways to change synaptic strength. They could be an order of magnitude (or more) complicated than artificial neurons.
- The human brain the most complicated physical system that we know about.

# 10. What are the key properties of human vision that distinguish it from AI Vision?

- Some aspects of Human visions worth copying for AI, but some are not.
- Marr & Poggio's three levels of analysis. Consider birds and airplanes. Wings are necessary for birds and airplanes. But airplanes do not need feathers.
- Brains versus Machines.
- The Brain uses 1 watt for computation (20 more watts to stay alive) while a computer with 4GPUS uses 1,500 watts.
- Real neurons get tired and need food (blood), artificial neurons do not.
- The brain is a product of evolution (a sequence of kludges?) but AI Vision is designed by humans (doesn't mean it is optimal).
- The brain has adapted to perform visual tasks in specific environments – we are good at reading facial expressions, but not at recognizing 100,000,000 faces (viewed front-on), or interpreting Computer Tomography images.

# Part 11. Human Visual Failures.

- *Lack of Attention*– the Gorilla in the Room, *Change Blindness*, inability to realize an image is an impossible scene, the Gorilla in the CT image. (sensible “short-cut” strategies to avoid computations).
- *Accidental Alignments* – sensible assumption that works most of the time.
- *Failures of Consistency* – inconsistent interpretation of entire image.
- More general effects:
  - After-effects – demo -- these may be due to “tired neurons” or might be a sensible strategy (e.g., for a system that has to repeatedly self-calibrate itself). Seeing motion in static images.
  - Visual crowding in the periphery – know what the objects are, but get their order wrong (“short-cut” strategy so that only fovea is high resolution).
  - Memory/resource limitations – inability to track more than five objects, to remember details of pictures.

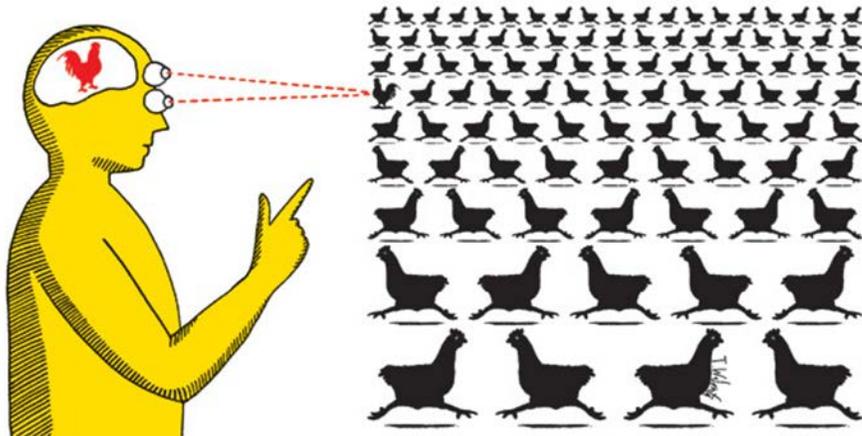
# Part 11A Failure of Attention

- Gorilla in the Room. We fail to see gorilla's if our visual attention is directed elsewhere. This may be due to a visual strategy that is very efficient (requires few computational resources) and correct most of the time. (Skilled illusionists perform tricks by diverting attention).



# Part 11B Change Blindness

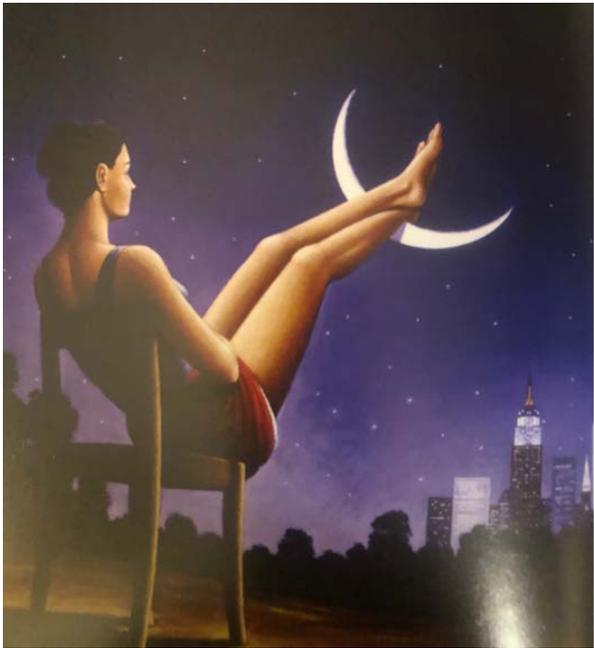
- We are bad at noticing differences between images (provided they have similar semantic content). We are bad at noticing changes outside our center of gaze.



**Change Blindness (using flicker)**  
(from J. Kevin O'Regan -- <http://nivea.psych.univ-paris5.fr>)

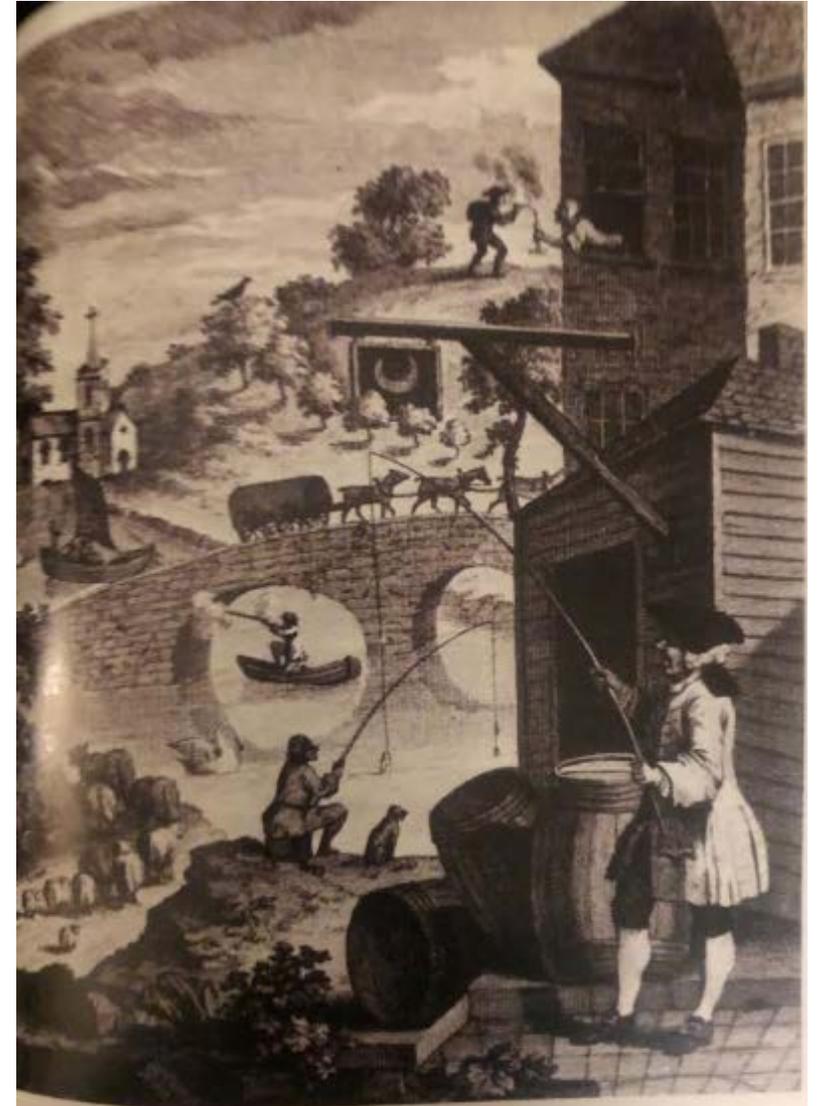
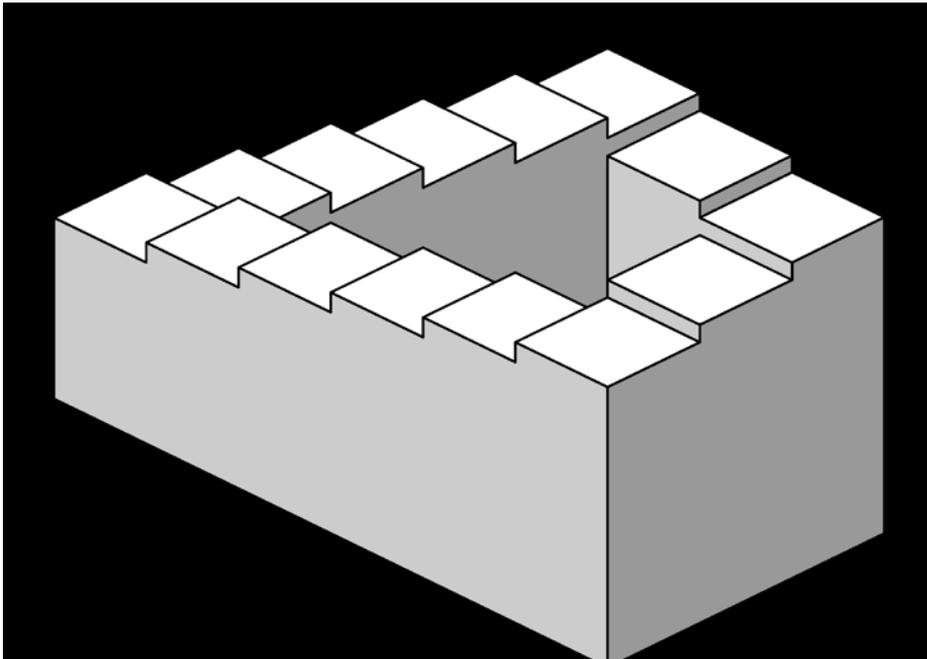
# Part 11C Accidental Alignment

- Accidental Alignment. This may result from a sensible visual strategy that is correct most of the time. (Far right – eye in the kitchen sink).



# Part 11D Failure of Consistency

- Failures of consistency. Without careful attention we may fail to notice the inconsistency. Perhaps a sensible efficient (i.e. lazy) strategy



WILLIAM HOGARTH

# Part 11E Human Visual Failures

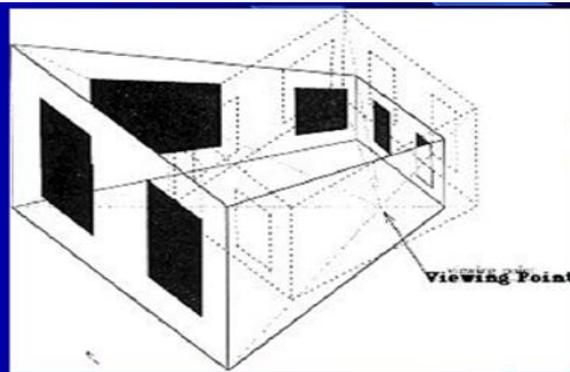
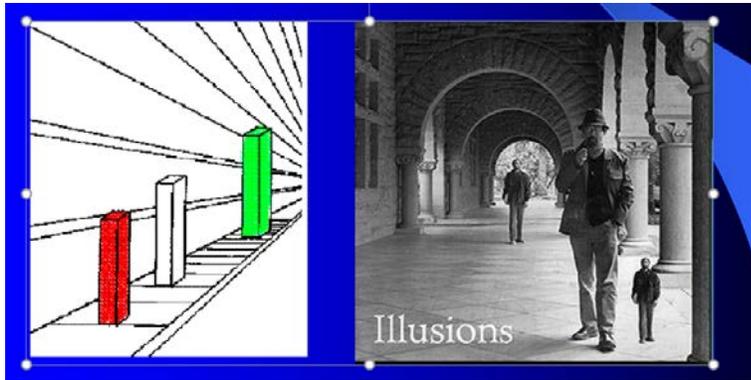
- Michael Bach's webpage gives a huge set of visual illusions with discussions and possible explanations. <http://www.michaelbach.de/ot/>
- Many of these can be explained as sensible visual strategies – e.g., in Bayesian terms (see Pavan later) -- which sometimes make mistakes.
- Consciousness, Visual Awareness, and Blindsight.
- Sometimes we see things but are not consciously aware of it. E.g, we say we cannot see something but if asked to guess what it is, we will guess right most of the time. This relate to the problem of consciousness, which may or may not be relevant to AI-Vision.

# Part 12A: Key Aspects of Human Vision

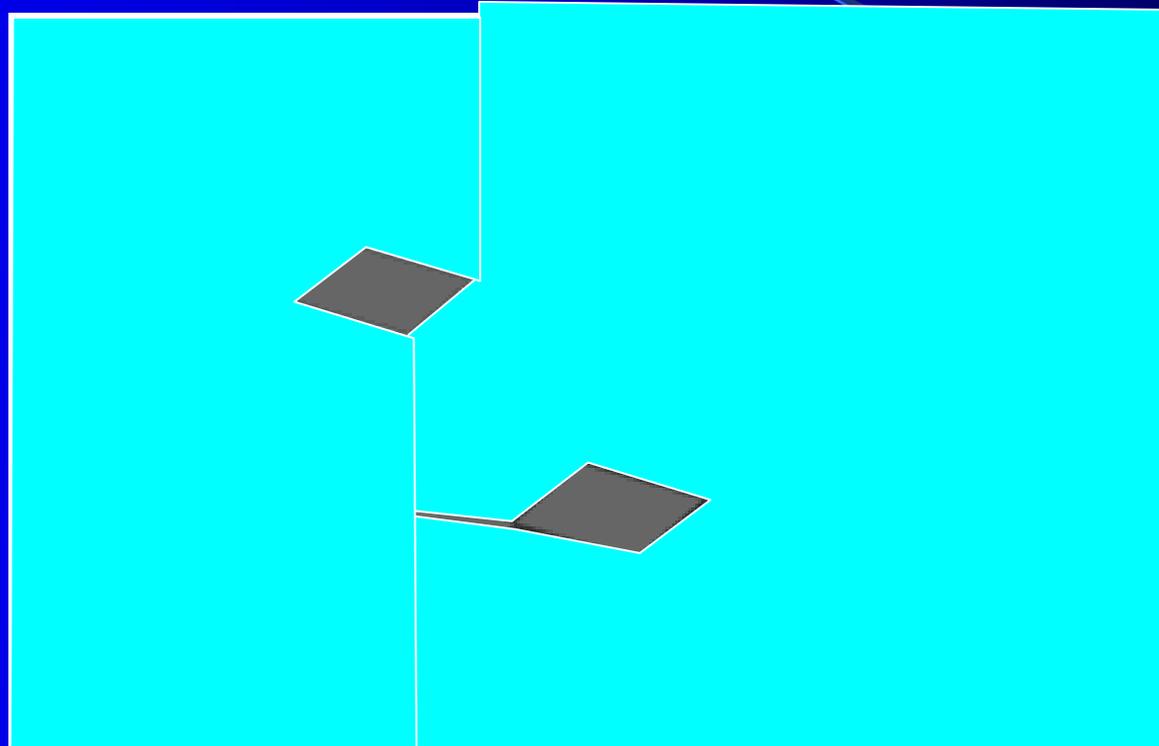
- ***Approximate and as-needed Analysis by Synthesis:*** Humans have the ability to perform approximate and as-needed inverse computer graphics. This involves the ability to “parse the image” and explain every pixel.
- We perform multiple visual tasks. As we can look round the room and see multiple objects (chairs, pictures, books, a robot dog, carpets, clothes, a sleeping cat, an empty wine glass, a projector, an exercise machine, and many others). For all of them, we can estimate their shape, their position in the world, and the positions and shapes of their parts. We can also describe their properties, like the age of the carpet and how recently it has been cleaned.
- This is a modern update to the classic ideas of “Vision as Inference” (Helmholz, Gregory). “Vision is not just a passive acceptance of stimuli, but an active process involving memory and other internal processes”. This leads to Bayesian Theories of Vision which combine likelihood functions and prior probabilities.
- Physical constraints, ecological constraints, natural constraints.

# Part 12A: Key Aspects of Human Vision

- We see in 3D by assuming normal 3D structure (ground planes, shadows, and 3D structure), but are fooled sometimes.

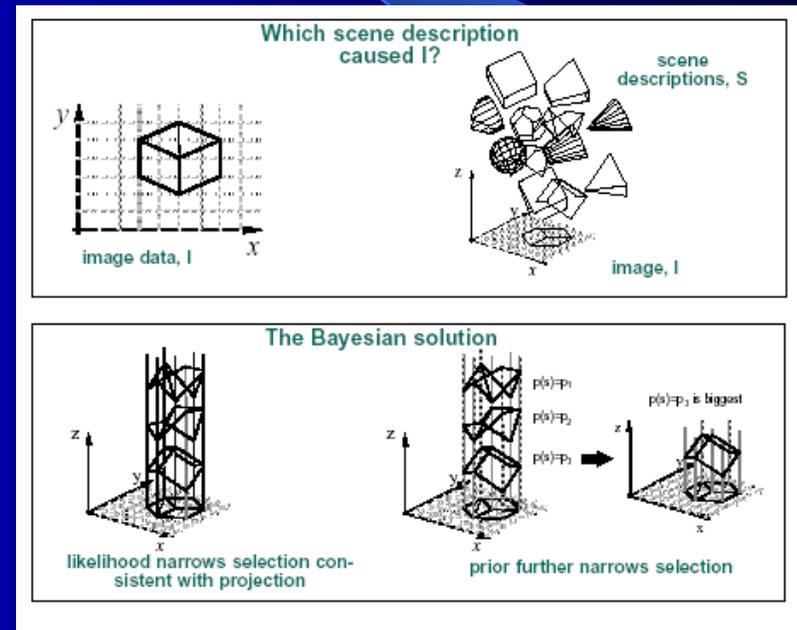


Part 12A: Which square is brighter?



# 12A Bayesian Perspective

- There are an infinite number of ways that images can be formed.
- Why do we see a cube?
- The likelihood  $P(I|S)$  rules out some interpretations  $S$
- Prior  $P(S)$ —cubes are more likely than other shapes consistent with the image.



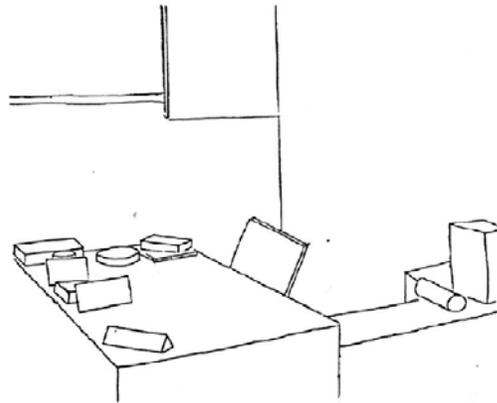
# Part 12B. Key Aspects of Vision

- ***How humans acquire knowledge: Development and Learning.***
- Human vision brings to bear an enormous amount of knowledge acquired/learnt over a lifetime (unlike current AI-Vision systems).
- This knowledge is initially learnt, during development, by an orchestrated procedure where certain visual abilities are learnt first to enable the learning of more complex ones.
- This learning relies (at least initially) on exploiting image sequences, searching for causal structure, taking actions in the world, and exploiting other senses. “Learn like a child”.
- What is a surrogate for this knowledge? Computer Graphics?

# Part 12C: Key Aspect of Human Vision

- ***Humans have the ability to use context.*** This results from our knowledge about the world. C, von der Malsburg.

Object recognition:  
50% by context



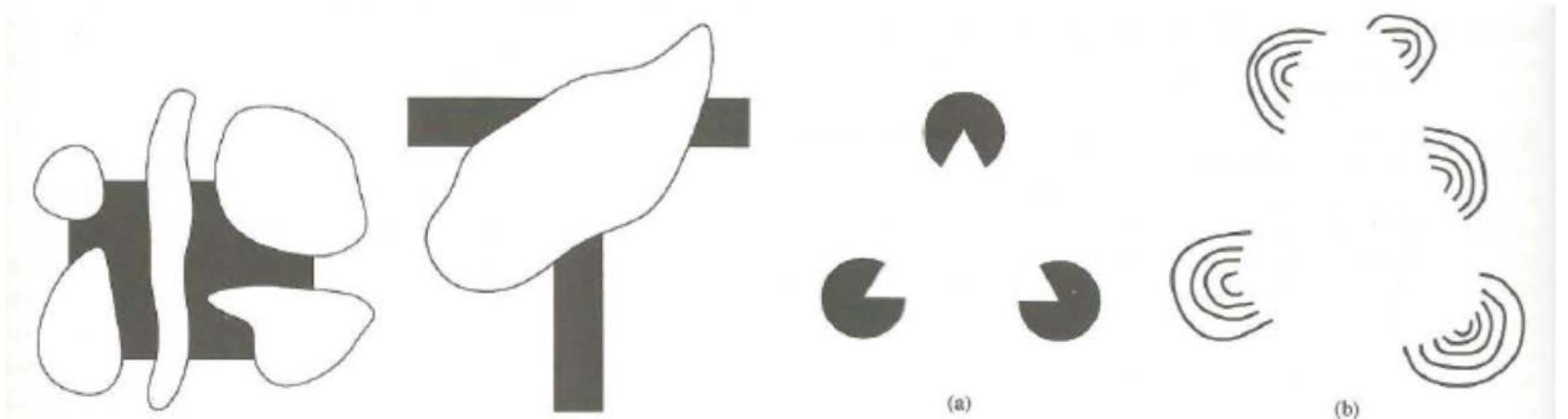
- Object recognition: 50% by context. Search guided by context.
- Context may be less relevant for real images (more information available).

# Part 12D. Key Aspects of Vision

- ***Perceptual Organization. Gestalt Grouping/.***
- Humans have also the ability to see patterns and to group basic elements into more complex structures. This was studied by Gestalt psychologists (e.g., Wertheimer, Kanisza).
- This can be illustrated by various grouping properties – accidental alignment, common fate, etc. (the phenomena are so strong – everybody gets the same perception) that demonstrations are sufficient.
- The ability to group patterns, of highly variable components, shows that human vision can deal with abstraction.
- Certain types of grouping (e.g., Kanisza) shows that human vision is aware of geometry and occlusion (independent of object knowledge)

# Part 12E. Key Aspects of Vision

- Examples from Kanisza's book. Organization in Vision.



# Part 12F: Key Aspects of Human Vision

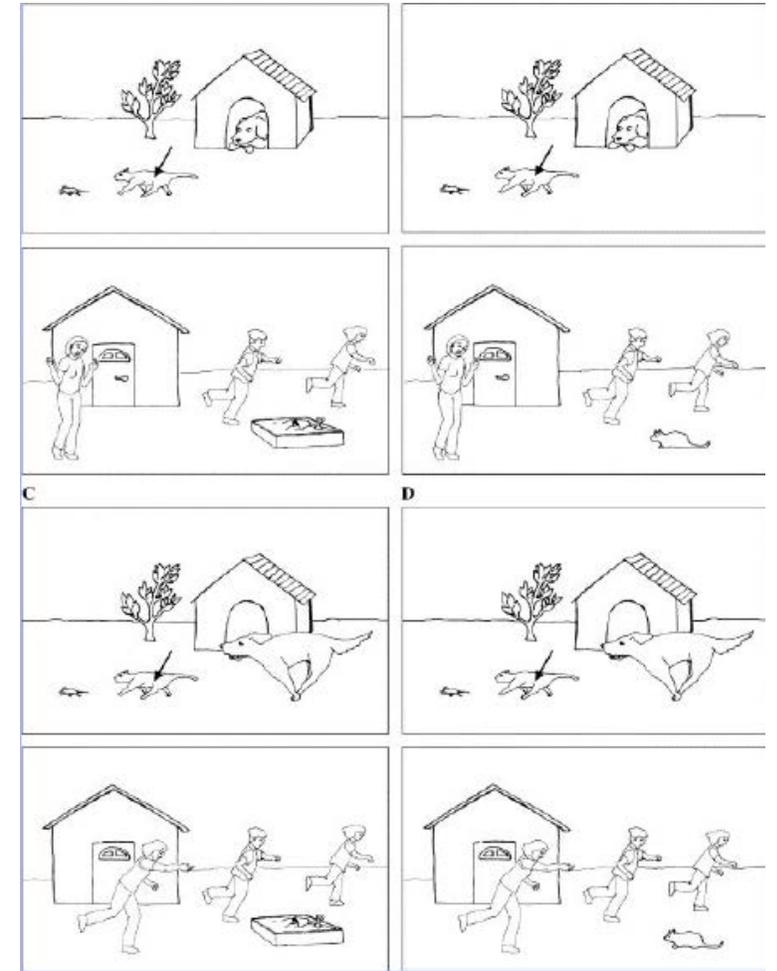
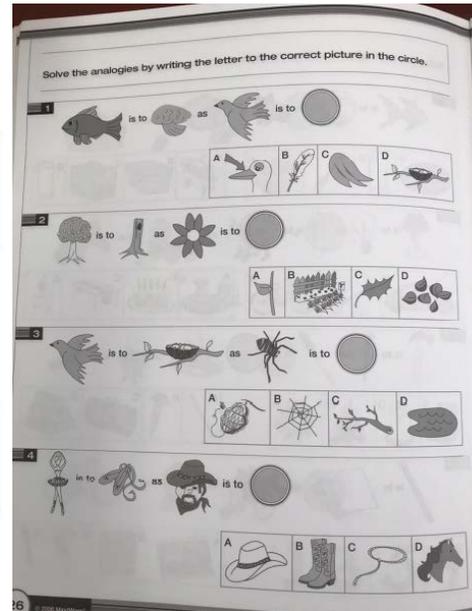
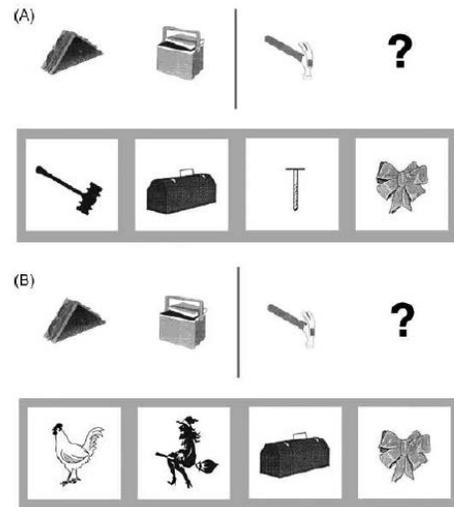
- ***Cues and Modularity.***
- The study of human vision has identified visual cues which are sufficient for performing visual tasks in restricted (toy environments).
- E.g., Shape from shading, texture, contour, focus, and perspective.
- These cues are effective in simplified domains (toy worlds) although extending them to work in the complexity of real images is often extremely challenging.
- Many of these cues are now embedded in AI-Vision models, but many are not.

# Part 12G. Key Aspects of Vision

- ***Abstraction and Domain Transfer.***
- Humans can understand an object from an image, from a drawing, from an highly abstract sketch. This is, in AI-vision terminology, an extreme form of domain transfer.
- Humans can factor shape and geometry – and recognize a blue tree, even if they have never seen one before.
- Humans can perform analogical reasoning. We can not only recognize visual similarity between objects, but also relationships (e.g., part-whole: paw to cat, hand to person), and functional relations (e.g., hammer is in toolbox, notebook is in backpack), and other relations (e.g., woman chases child is like cat chases mouse – but only in some ways!).

# Part 12G. Domain Transfer & Analogies

- Domain Transfer. Analogies.

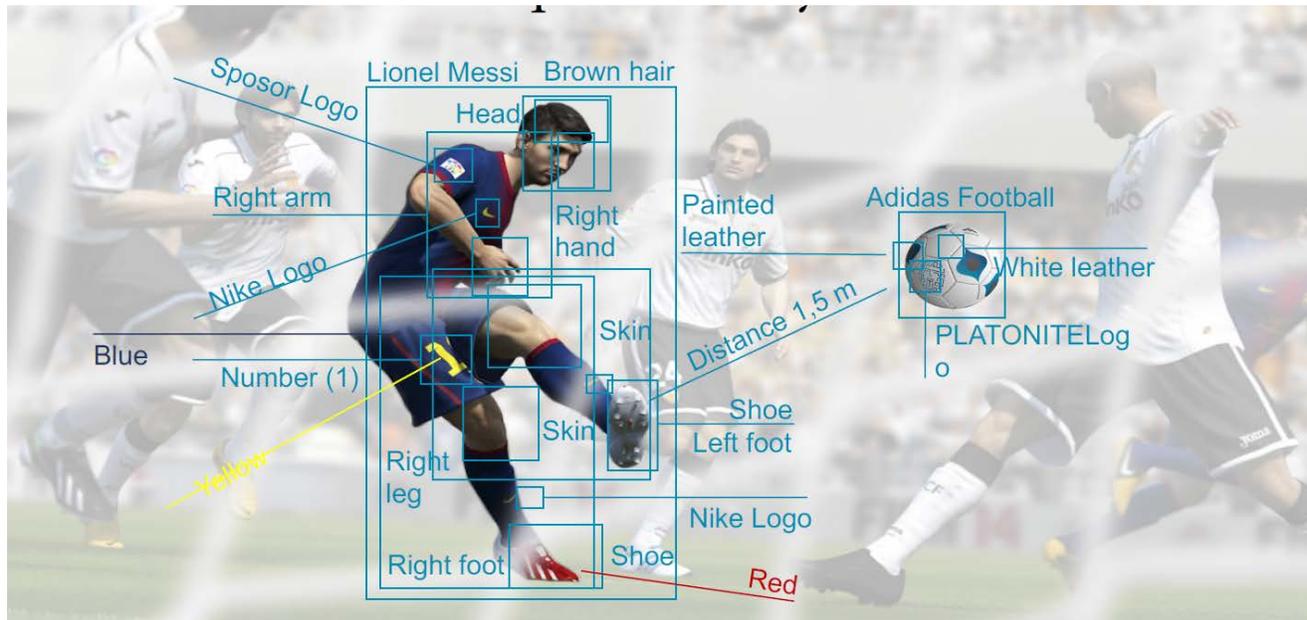


# Part 12H: Key Aspects of Vision

- ***Internal representations of parts. Vision is explainable.***
- These parts can be detected without context. They can also be described by language.
- Humans can explain why they have recognized an object – this is a car because I can see the wheels, the chassis, the doors, etc – and they satisfy the correct spatial relationships.
- These parts can also be abstract – i.e. we can recognize a fish even if it is constructed from bicycle parts.
- Humans have the ability to deal with the complexity of vision and to answer many different tasks using the same underlying representation

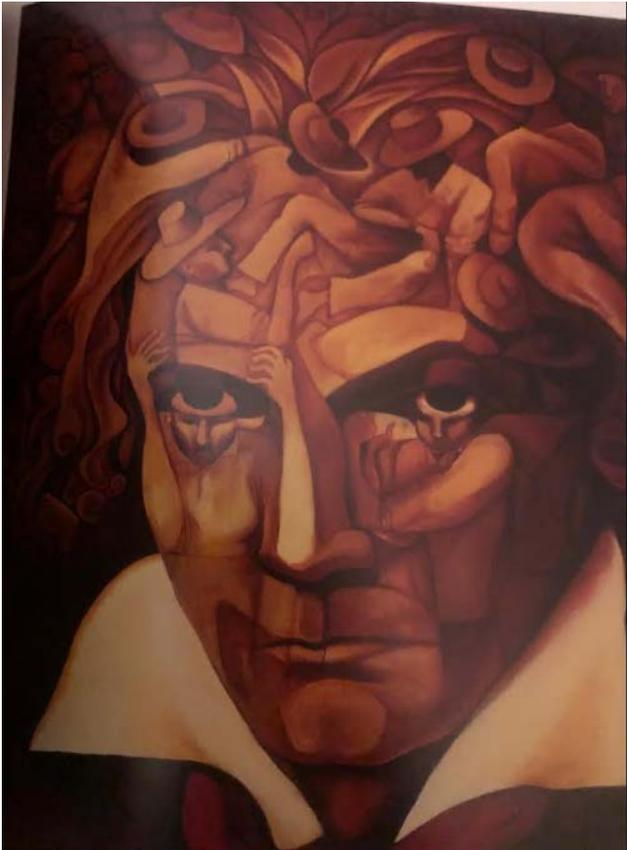
# Part 12H: Key Aspects of Human Vision

- Part Examples (from von der Malsburg)



# Part 12: Key Aspects of Vision

- ***Parts. And Abstractions.*** Beethoven constructed from Humans.



# Part 12: Key Aspects of Human Vision

- ***Humans have the ability to deal with the complexity of vision and to answer many different tasks using the same underlying representation.***
- ***Compositionality is an strategy for doing this. Objects are composed hierarchically in terms of basic elements. This gives the ability to construct an enormous number of objects from basic elements.***

## Part 13: *Hierarchical Architectures*

- The input to the visual system starts at the retina which captures the image and transmits it to the visual cortex (but the retina may also process it).
- The visual cortex is organized into visual areas. These start at visual areas V1 and V2, which are huge and contain roughly 70-80% of the neurons in the visual cortex. The ventral stream, where most object recognition is done, is organized hierarchically.
- The ventral stream (and the whole visual cortex) contains both feedforward neural pathways (up the hierarchy) and even more feedback neural pathways (down the hierarchy).

## Part 13. *Hierarchical Computational Theories*

- **Marr's Theory.** The visual system has three representations: (I) The primal sketch, which describes properties of the image, (II) The 2 1/2D sketch which captures the geometry, and (III) A 3D representation of objects & scene structure. Feedforward.
- **Fukushima. Hmax. Deep Networks.** A hierarchy with simple features at the lowest level and more complex and specific features higher up the hierarchy (but increasingly independent of position). Feedforward.
- **Mumford. Analysis by Synthesis.** Bayesian theory where feedforward processing activates high-level models which synthesizes images in a feedback process and compares with input image.

# Part 13: Key Aspects of Human Visual Architecture

- Other theories include analysis-by-synthesis which combines bottom-up and top-down processing (Mumford 1991). Some instantiations of this are the Helmholtz machine and DDMCMC. It is also clear that top-down processing is important for video sequences when we want to predict motion (Rao and Ballard).
- Another metaphor is that the vision system is organized like a hierarchical organization like an Army (or Corporation) where knowledge is distributed hierarchically. The General has an executive summary and can contact the Colonels for more details, who can contact the Major and so on down to the Privates. Here information flows up and down the hierarchy.
- More generally, most (according to Ed) neuroscientists think there must be both feedforward and feedback processing.

# Part 13: Other Perspectives

- A computational perspective is that the architecture is developed to solve the problem – object specific knowledge is the most powerful but is harder to apply because it requires recognizing the object. Instead it may be better to first perform “generic” processing (which exploits properties common to all objects) in order to obtain representations, like the 2 1/2D sketch, which can then be matched to objects.
- A related perspective, is the visual cortex must address the problem of representing an enormous number of objects, learning new objects from few examples, and rapidly processing images to determine which object is present. Toy-world models (Y&M) give one example.

# Part 14. This course.

- The course aims at describing visual phenomena and the types of computational models that are used to describe them.
- This involves techniques like linear/non-linear filtering, Bayesian decision theory, Markov random fields, neural network models, geometry, and radiosity.
- The hope is to introduce students to this fascinating area on the boundary between Artificial and Natural Intelligence.
- And to motivate researchers to take AI-Vision to the next level by developing models that can match human visual abilities.