Systems Engineering Meets Life Sciences: (Compositionality)

Prepared by:
Prof. Dr. Visvanathan Ramesh

Previous Lectures:

- Recap – Greiffenhagen Thesis / Systems Engineering Methodology
- Model-Based Recognition Overview (Mann, 1996, Dissertation)
- What is Context ? (Slides based on Derek Hoeim)
- Link to Systems Engineering Methodology
- Simulation for Cognitive Vision (Subbu Veerasavarappu)

Today's Lecture:

- Compositionality (Based on Slides from Borenstein et al, Stuart Geman)
- Compositional Models (P. Felzenswalb)
- Pattern Grammars Introduction (Song-Chun Zhu, Mumford)

Compositionality and Heirarchy (Geman, 2006)

Parsing Images with Context/Content Sensitive Grammars

Eran Borenstein, Stuart Geman, Ya Jin, Wei Zhang

I. Structured Representation in Neural Systems

II. Vision is Hard

III. Why is Vision Hard?

IV. Hierarchies of Reusable Parts

V. Demonstration System: Reading License Plates

VI. Generalization: Face Detection

Artificial Intelligence

- **Knowledge Engineering**

  engineer everything, learn nothing

- **Learning Theory**

  engineer nothing, learn everything

- **Both Lack Model**

Natural Intelligence

FIAS Frankfurt Institute for Advanced Studies

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

- **Strong Representation**

  simulation and semantics

- **Hierarchy and Reusability**

  ventral visual pathway, linguistics, compositionality

I. Structured Representation in Neural Systems

II. Vision is Hard

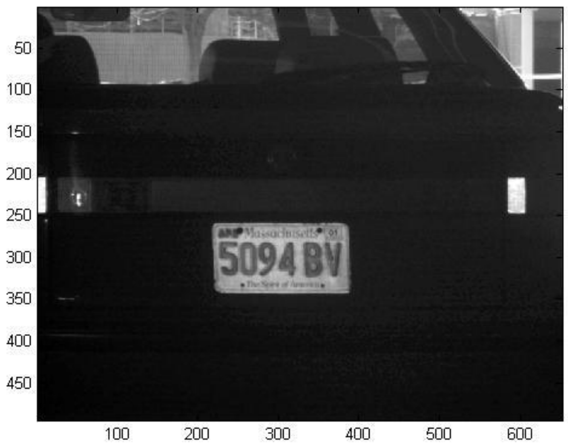
III. Why is Vision Hard?

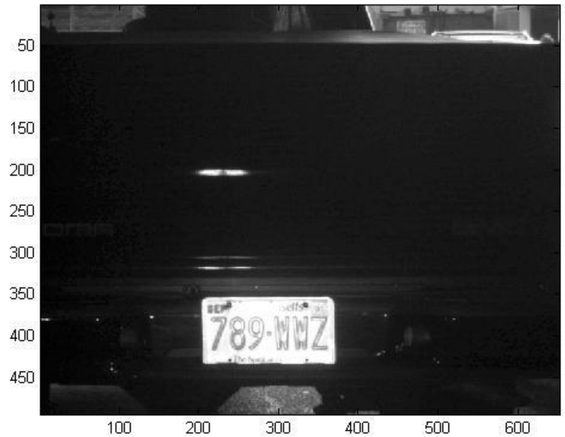
IV. Hierarchies of Reusable Parts

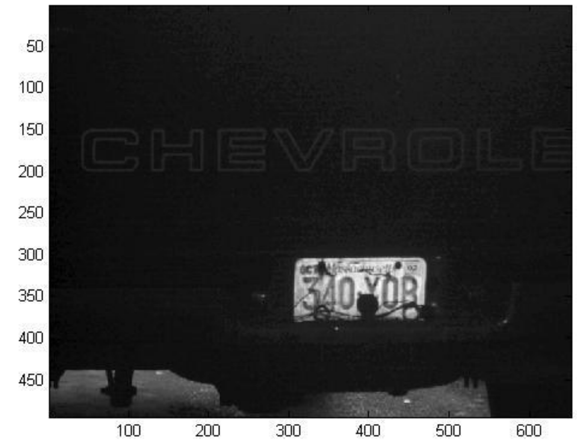
V. Demonstration System: Reading License Plate

VI. Generalization: Face Detection

License plate images from Logan Airport





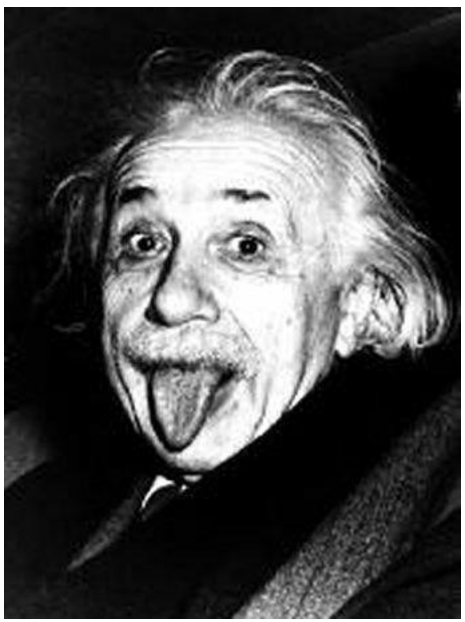


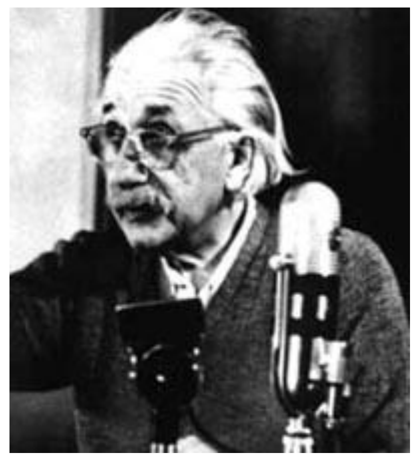
Machines *still* can’t reliably read license plates

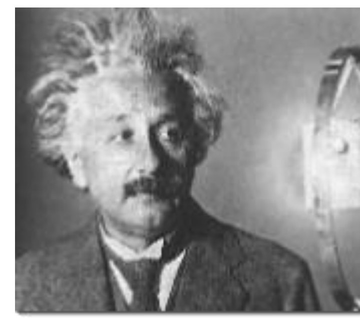
Wafer ID’s

Machines can’t read fixed-font fixed-scale characters as well as humans

Super Bowl







Machines can't find the bad guys at the Super Bowl

I. Structured Representation in Neural Systems

II. Vision is Hard

III. Why is Vision Hard?

IV. Hierarchies of Reusable Parts

V. Demonstration System: Reading License Plates

VI. Generalization: Face Detection

Instantiation

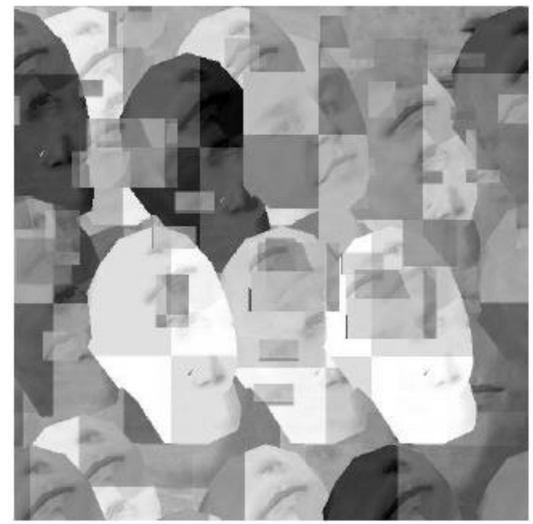
same

Empire style table

twins

Vision is *content sensitive*

“Clutter”



Human Interactive Proofs

Background is *structured,* and made of the *same stuff!*

I. Structured Representation in Neural Systems

II. Vision is Hard

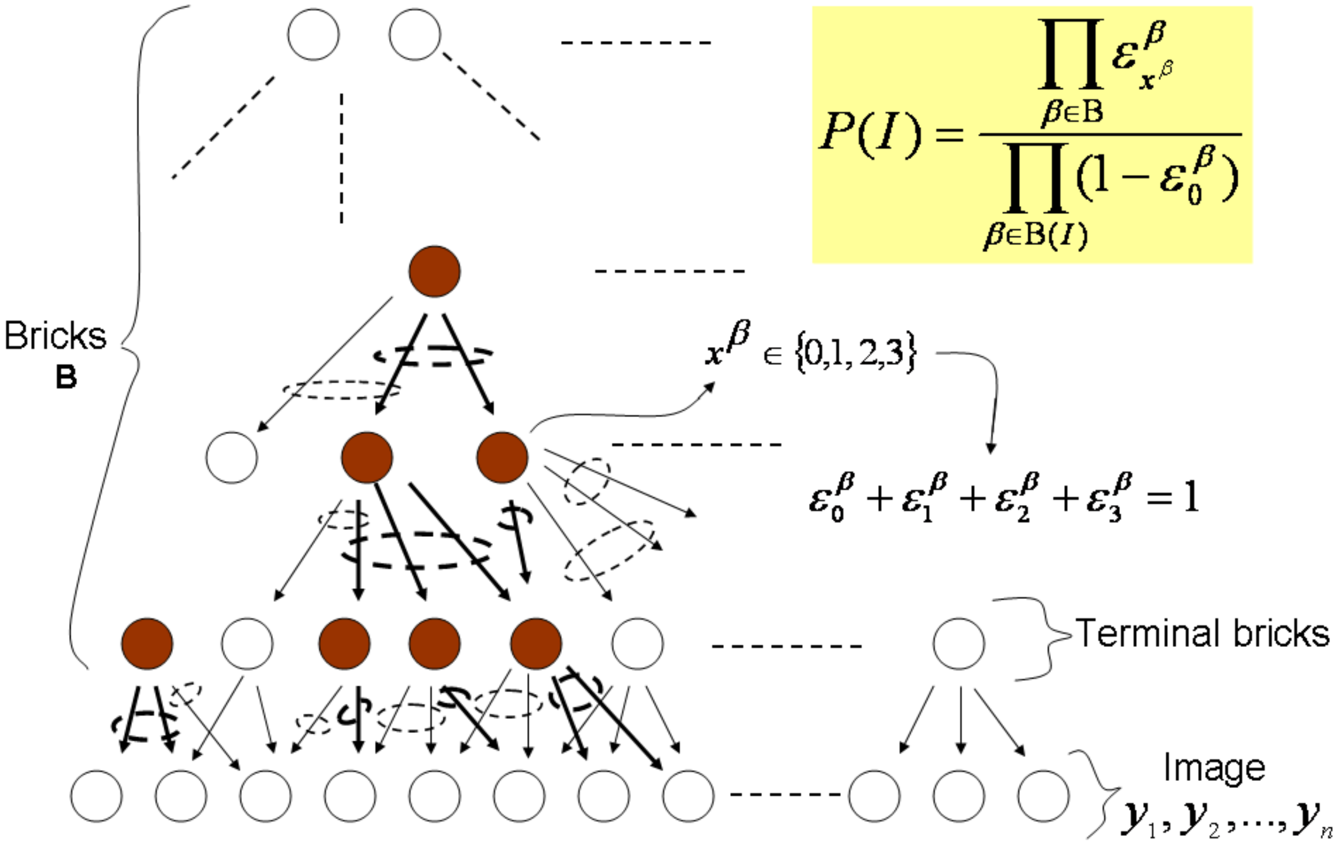
III. Why is Vision Hard?

IV. Hierarchies of Reusable Parts

V. Demonstration System: Reading License Pla

VI. Generalization: Face Detection

Composition Machine



$$P(I) = \frac{\prod_{\beta \in \mathrm{B}} \varepsilon^{\beta}_{x^{\beta}}}{\prod_{\beta \in \mathrm{B}(I)} (1 - \varepsilon_0^{\beta})}$$

Markov backbone

Hierarchical of Reusable Parts

FIAS Frankfurt Institute for Advanced Studies

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

“Bricks”

e.g. animals, trees, rocks

e.g. contours, intermediate objects

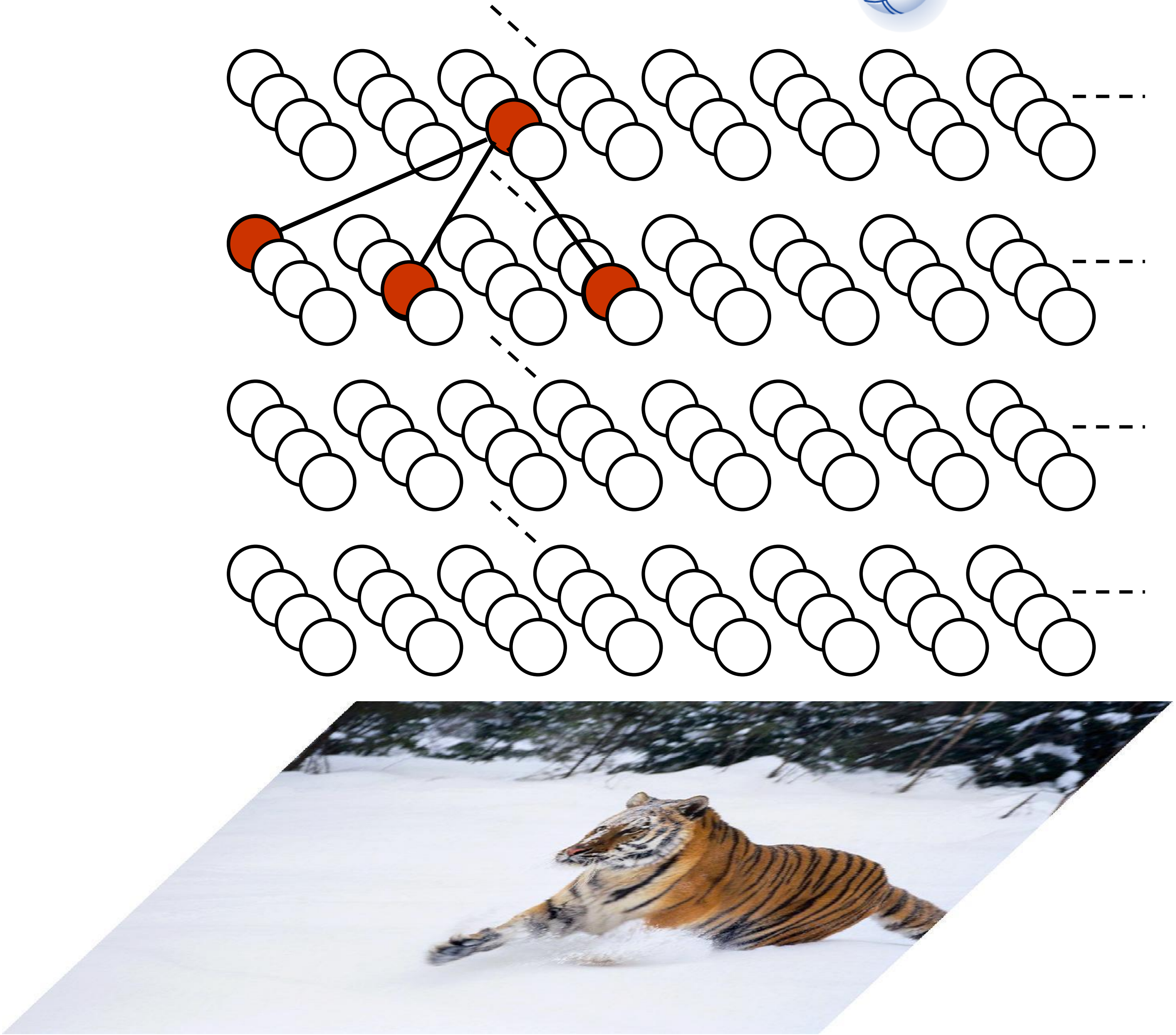
e.g. linelets, curvelets, T-junctions

e.g. discontinuities, gradient

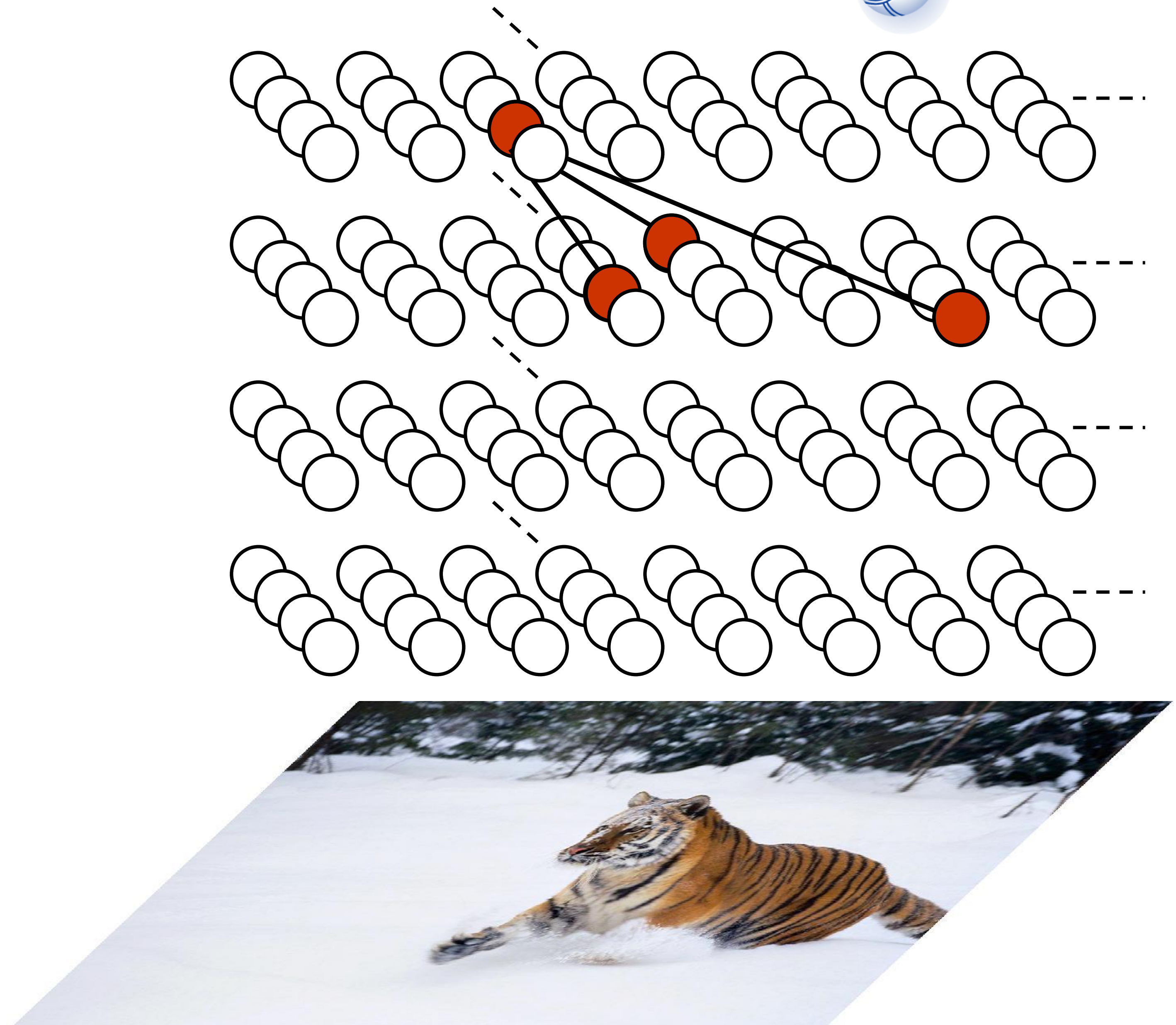
Hierarchy of Disjunctions of Conjunctions

FIAS Frankfurt Institute for Advanced Studies

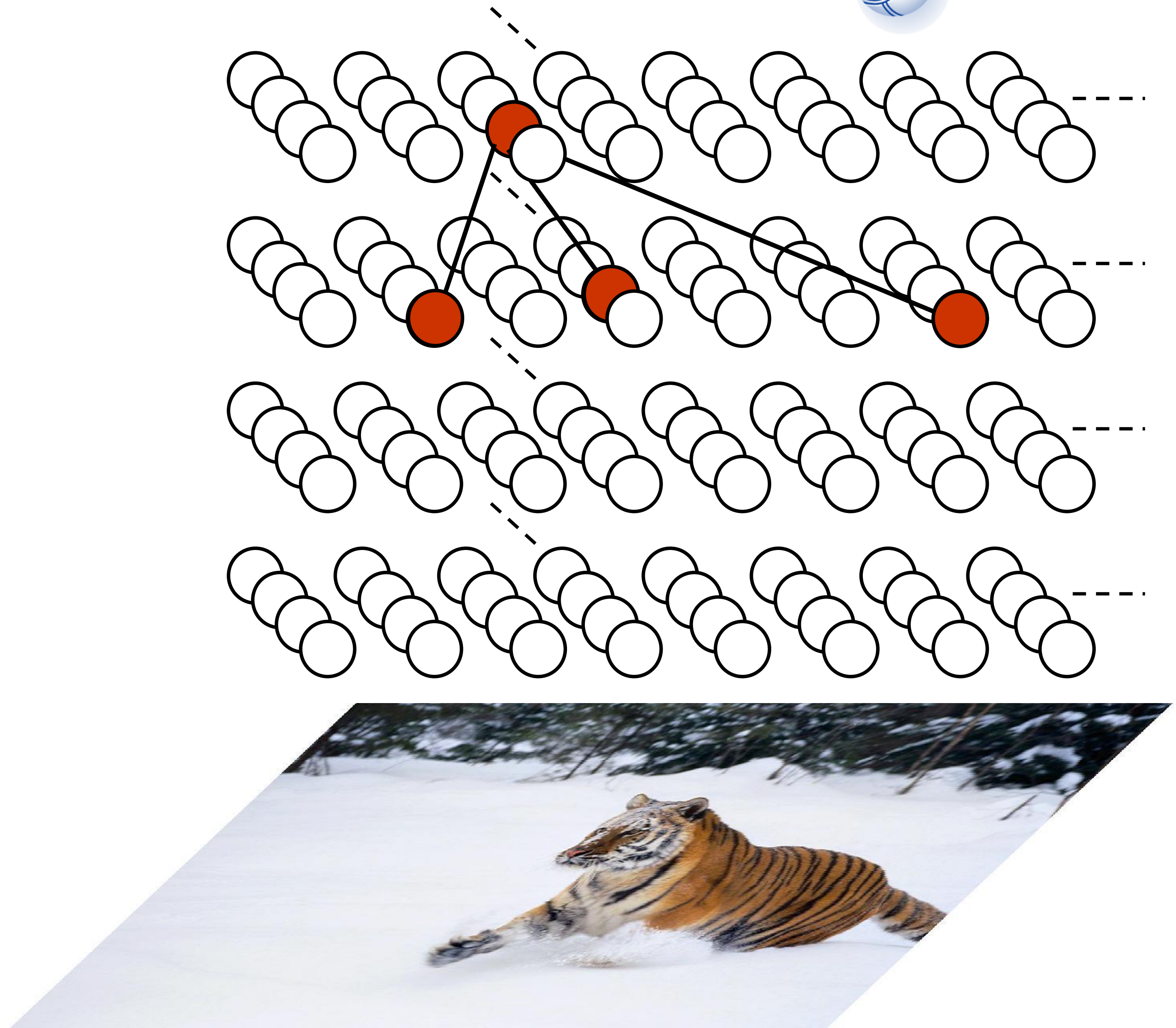
GOETHE UNIVERSITÄT FRANKFURT AM MAIN



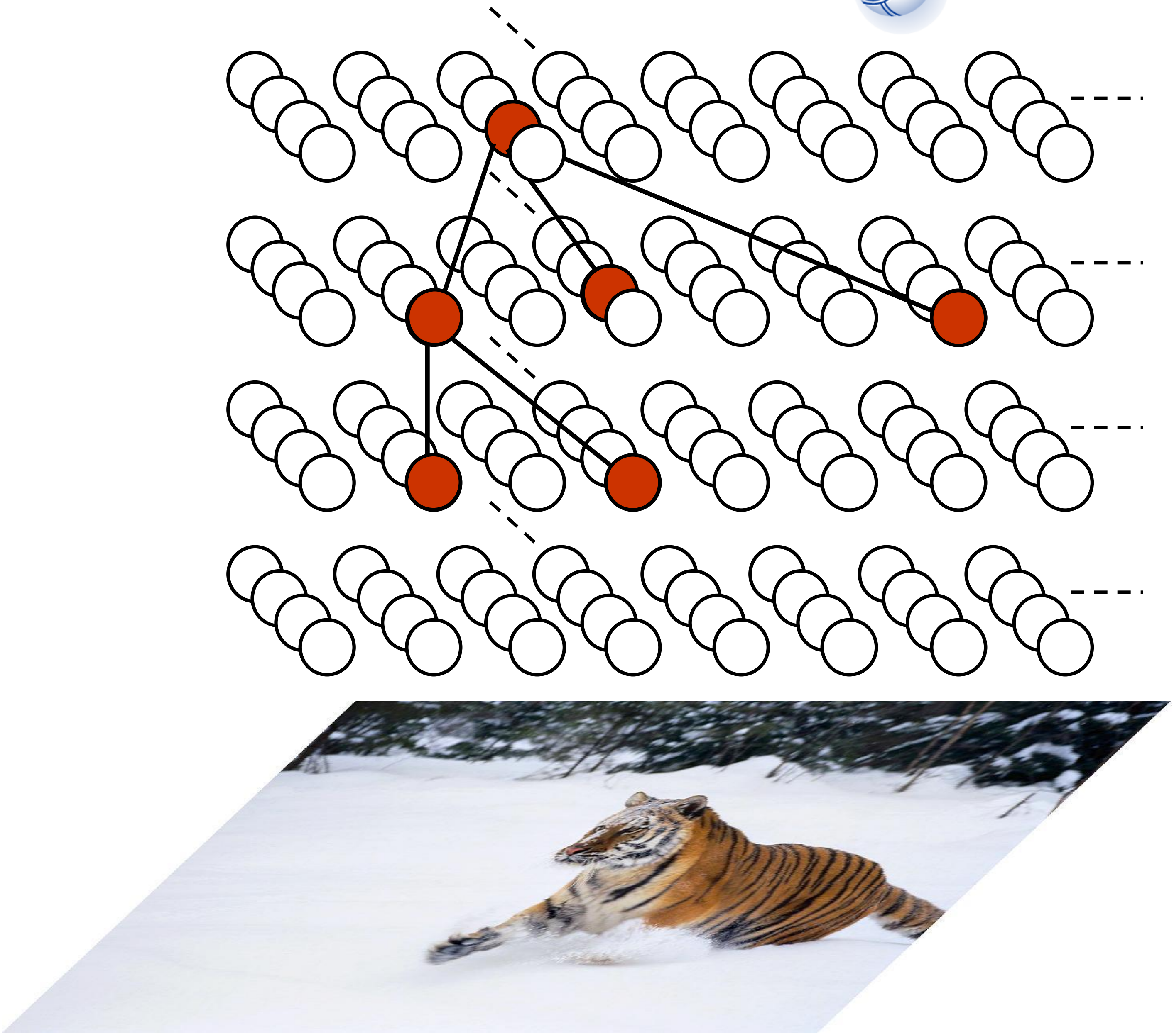
Hierarchy of Disjunctions of Conjunctions



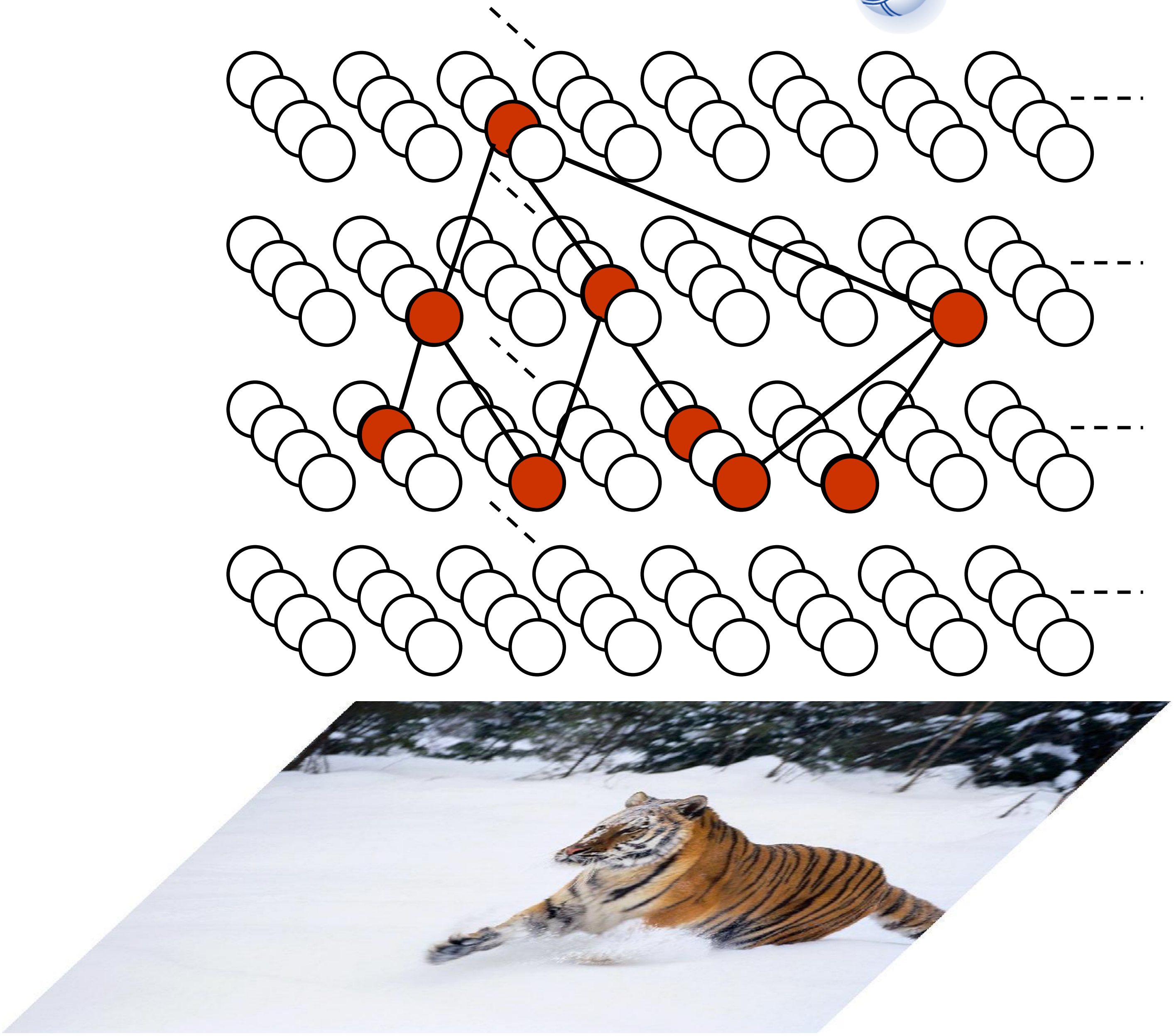
Hierarchy of Disjunctions of Conjunctions



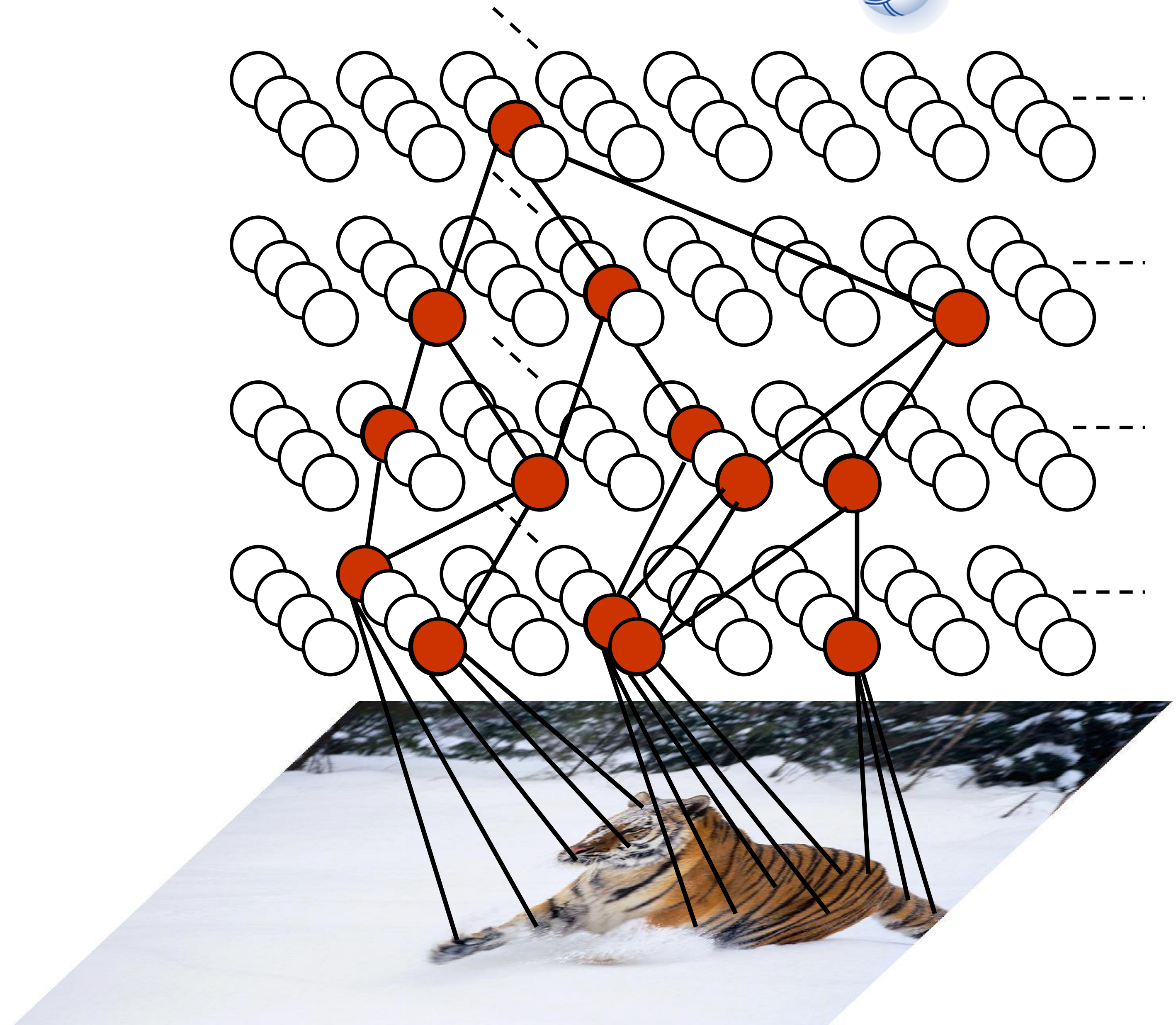
Hierarchy of Disjunctions of Conjunctions



Hierarchy of Disjunctions of Conjunctions



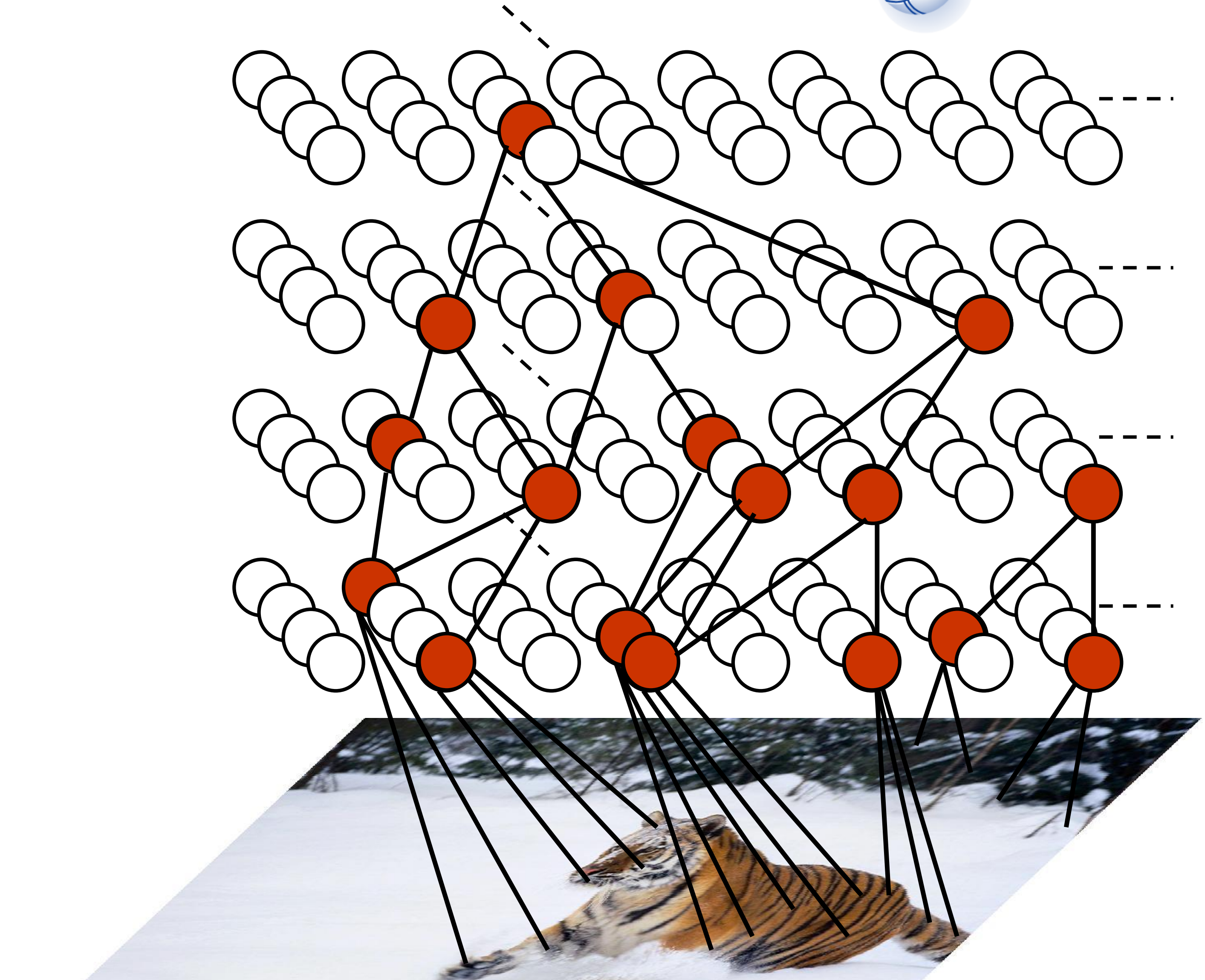
Hierarchy of Disjunctions of Conjunctions



Hierarchy of Disjunctions of Conjunctions

FIAS Frankfurt Institute for Advanced Studies

GOETHE UNIVERSITÄT FRANKFURT AM MAIN



Interpretations and Probabilities

FIAS Frankfurt Institute for Advanced Studies

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Interpretation

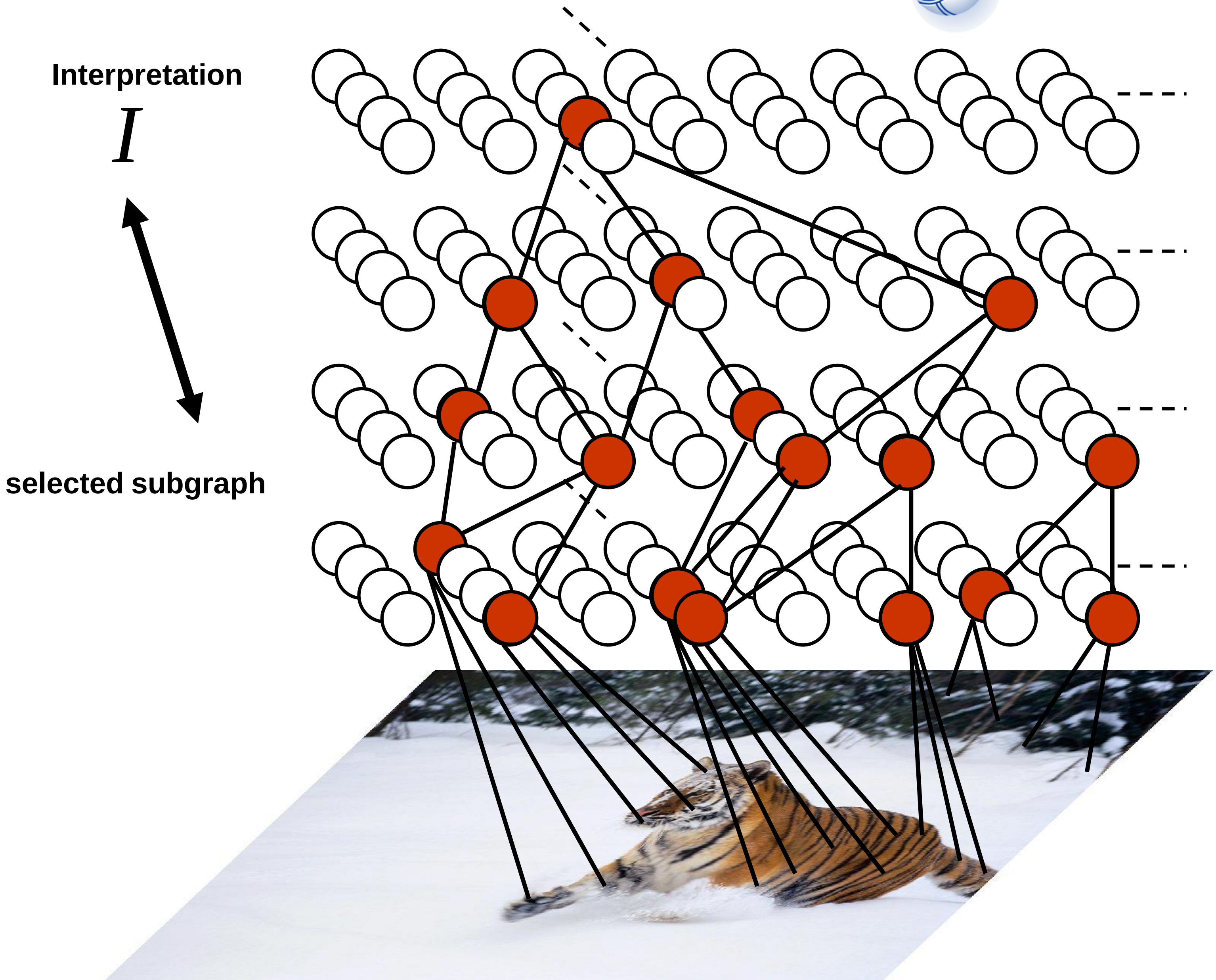
I

selected subgraph

Interpretations and Probabilities

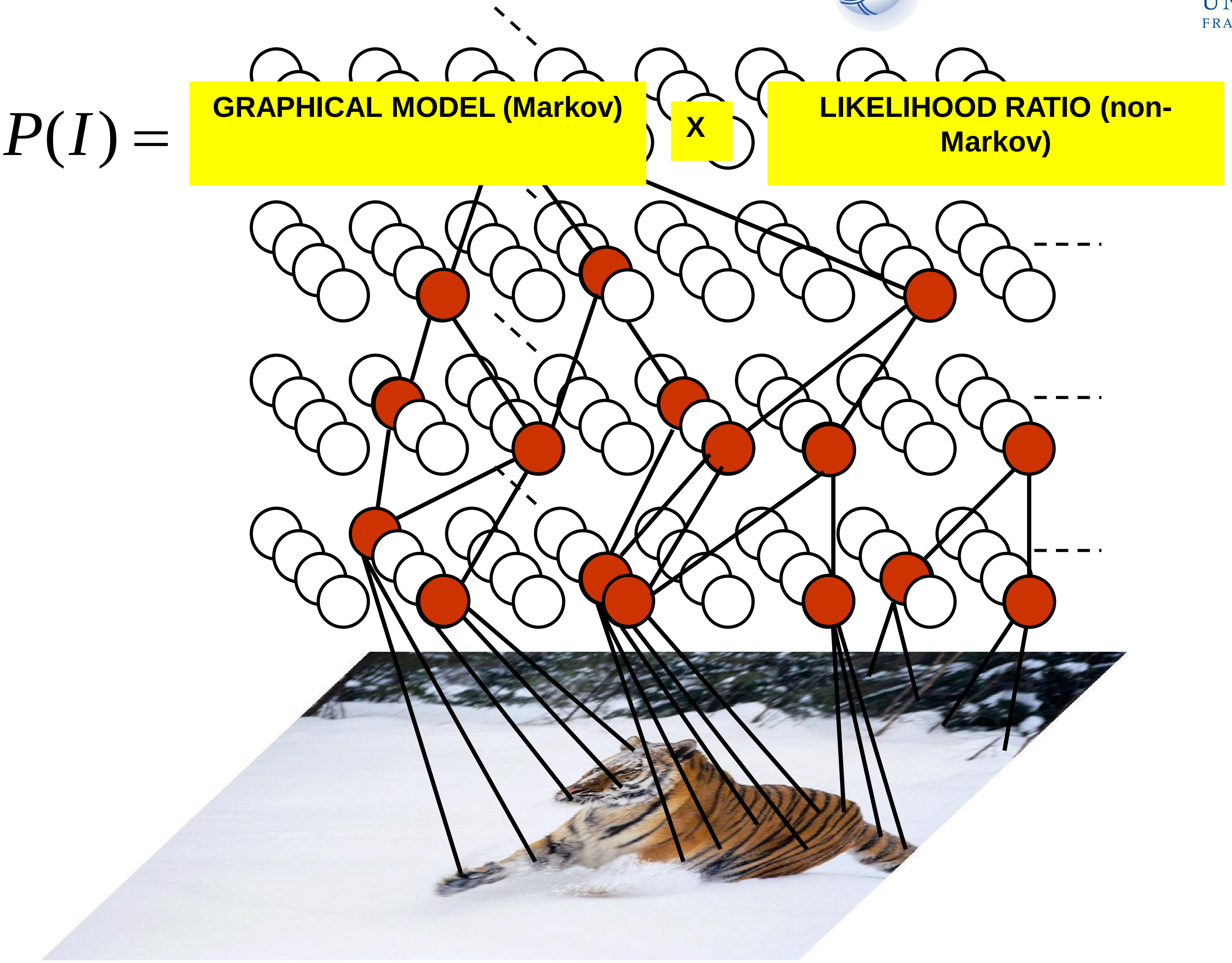
FIAS Frankfurt Institute for Advanced Studies

GOETHE UNIVERSITÄT FRANKFURT AM MAIN



Interpretations and Probabilities

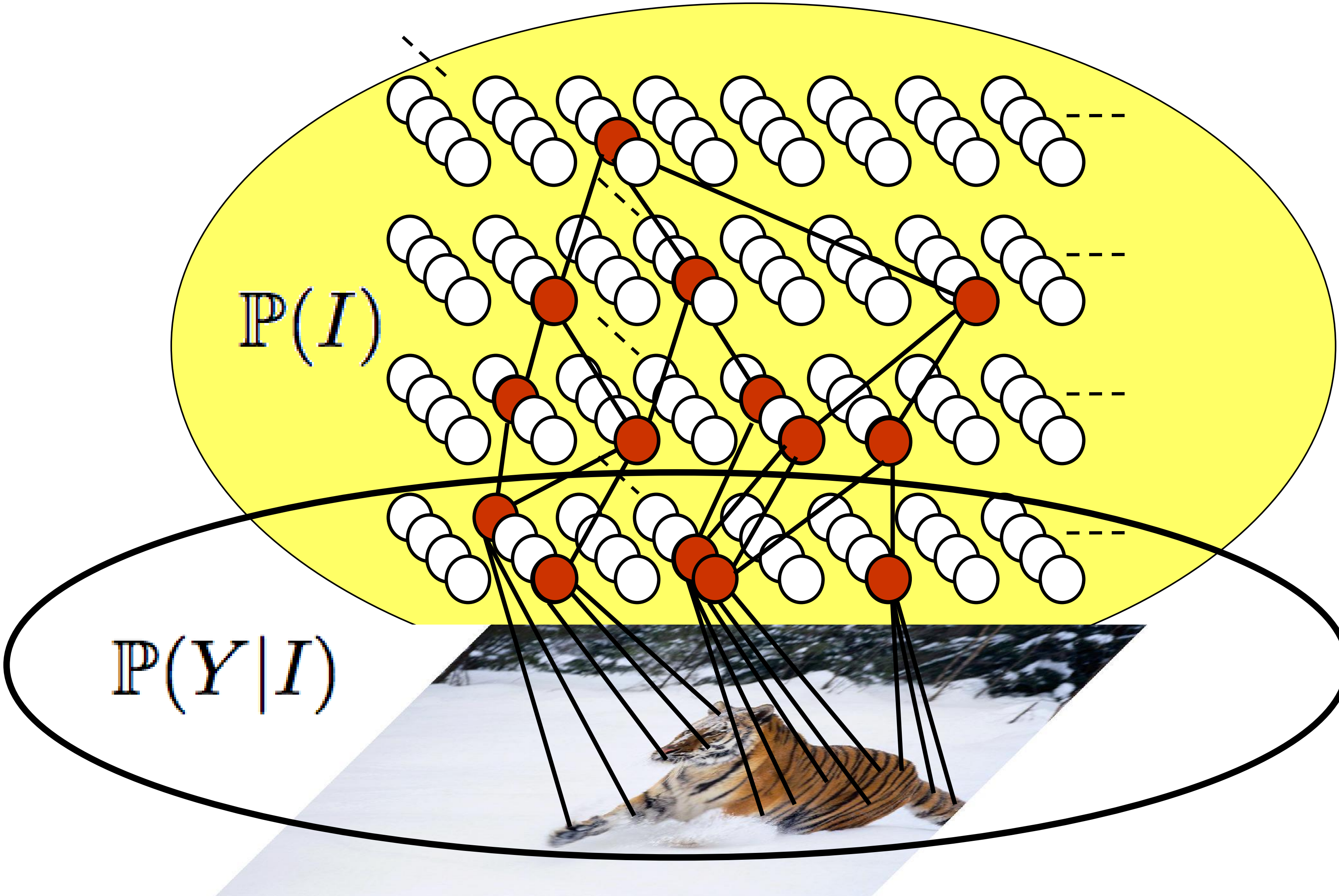
$$P(I) =$$



Generative (Bayesian) Model

FIAS Frankfurt Institute for Advanced Studies
GOETHE UNIVERSITÄT FRANKFURT AM MAIN

$\mathbb{P}(Y, I)$



FIAS Frankfurt Institute for Advanced Studies

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Formulation:

The state of a brick, say the brick $\beta \in \mathbf{B}$, is a random variable, $x^\beta \in \{0, 1, \ldots, n^\beta\}$, with $x^\beta = 0$ representing *off*, and $x^\beta = 1, 2, \ldots, n^\beta$ representing the selected set of children in Figure 1. The pixels themselves (actually, their grey levels) are represented by a vector of intensities, $\vec{y}$.

Markovian distribution on $\mathcal{I}$. Each brick $\beta \in \mathbf{B}$ is assigned a probability vector $(\epsilon_0^\beta, \epsilon_1^\beta, \ldots, \epsilon_{n^\beta}^\beta)$. In terms of these parameters, the probability $P(I)$ of an interpretation (i.e. a complete subgraph) I is

$$P(I) = \frac{\prod_{\beta \in \mathbf{B}} (\epsilon_{x^\beta}^\beta)}{\prod_{\beta \in \mathbf{B}(I)} (1 - \epsilon_0^\beta)} \qquad (1)$$

Formulation: Non-Markov Part

bone. Briefly, the derivation is as follows: Associate with each brick $\beta \in \mathbf{B}$ a (possibly vector-valued) attribute function $a^{\beta}(I)$, which measures the "fit" among the "parts" that instantiate β, as it appears in the particular interpretation $I \in \mathcal{I}$. If β is a "4-digit-string" brick, specifically, then

In a compositional distribution, the *null* attribute distributions are compared to their *composed* counterparts: given $I \in \mathcal{I}$,

$$P(I) \propto \frac{\prod_{\beta \in \mathbf{B}}(\epsilon^{\beta}_{x^{\beta}})}{\prod_{\beta \in \mathbf{B}(I)}(1-\epsilon^{\beta}_{0})} \prod_{\beta \in \mathbf{A}(I)} \frac{p^{c}_{\beta}(a^{\beta}(I))}{p^{0}_{\beta}(a^{\beta}(I))} \qquad (2)$$

where $\mathbf{A}(I)$, the "above set", is the set of non-terminal *on*

I. Structured Representation in Neural Systems

II. Vision is Hard

III. Why is Vision Hard?

IV. Hierarchies of Reusable Parts

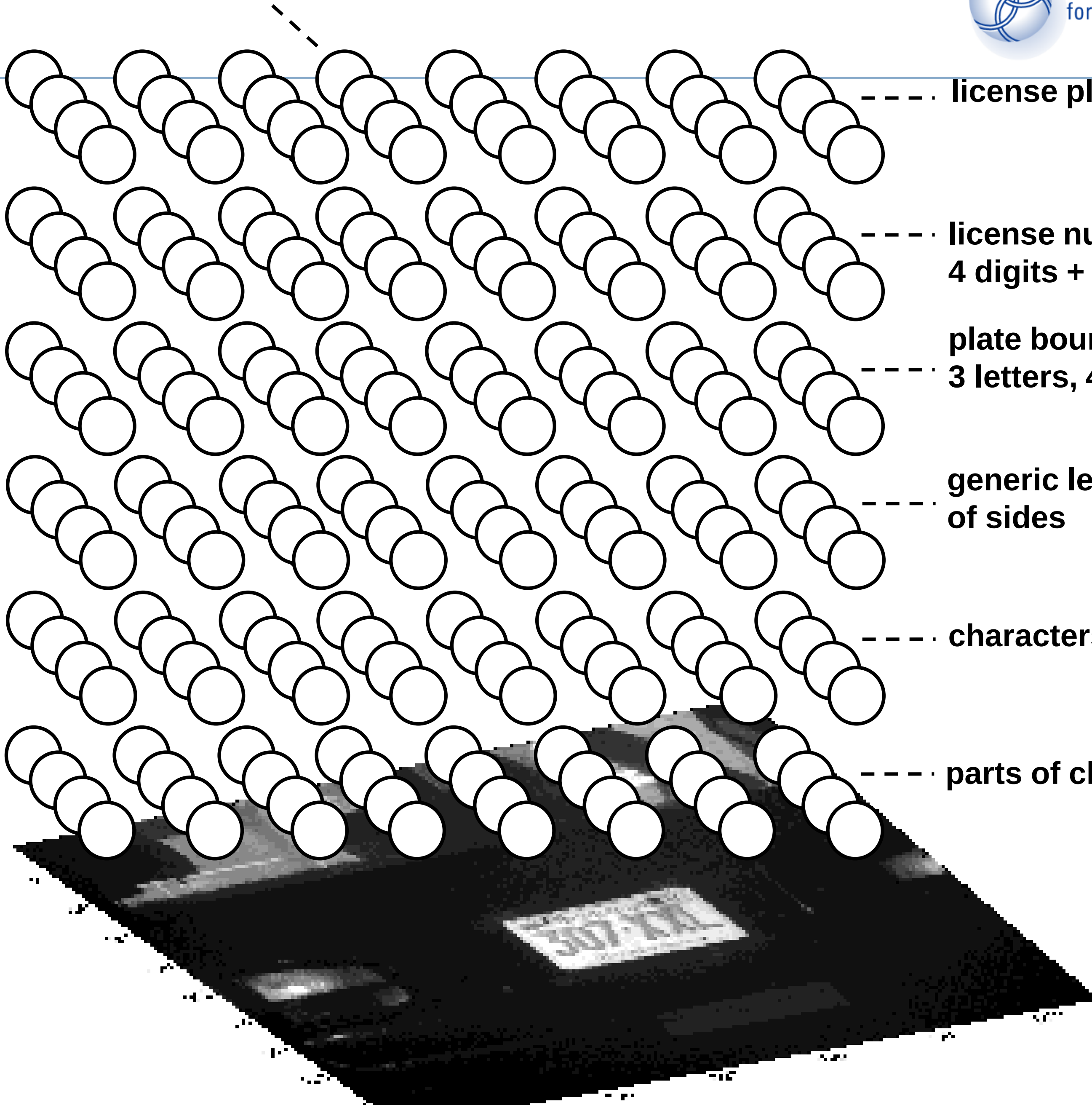
V. Demonstration System: Reading License Plates

VI. Generalization: Face Detection

Test set: 385 images, mostly from Logan Airport

Courtesy of Visics Corporation

Architecture



license plates

license numbers (3 digits + 3 letters, 4 digits + 2 letters)

plate boundaries, strings (2 letters, 3 digits, 3 letters, 4 digits)

generic letter, generic number, L-junctions of sides

characters, plate sides

parts of characters, parts of plate sides

Image interpretation

Original Image

Top object

Top 10 objects

Top 25 objects

Image interpretation

FIAS Frankfurt Institute for Advanced Studies

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Test image

Top objects

Performance

- 385 images
- Six plates read with mistakes (>98%)
- Approx. 99.5% characters read correctly
- Zero false positives

Efficient discrimination: Markov versus Content-Sensitive dist.

Original image

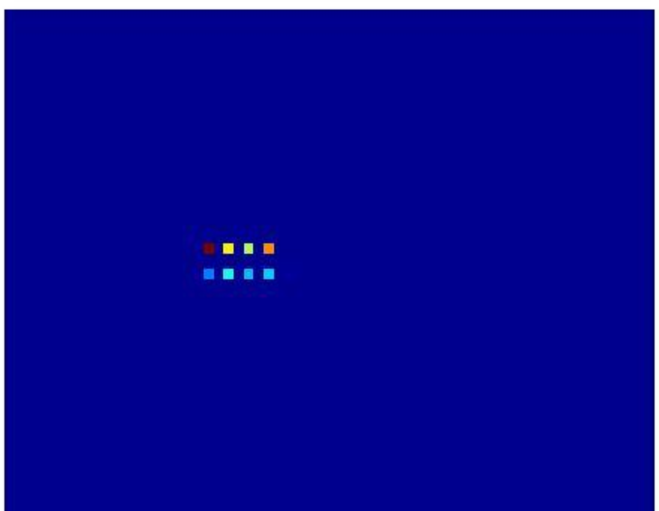
Zoomed license region

Top object under Markov distribution

Top object under content-sensitive distribution

Efficient discrimination: testing objects against their parts

Test image



9 active “8” bricks under whole model

1 active “8” brick under parts model

Summary

Vision is Content Sensitive

Non-Markovian probability models

Background is Structured, and Made of the Same Stuff

Objects come equipped with their own background models

I. Structured Representation in Neural Systems

II. Vision is Hard

III. Why is Vision Hard?

IV. Hierarchies of Reusable Parts

V. Demonstration System: Reading License Plate

VI. Generalization: Face Detection

Plates ⟶ Face Detection

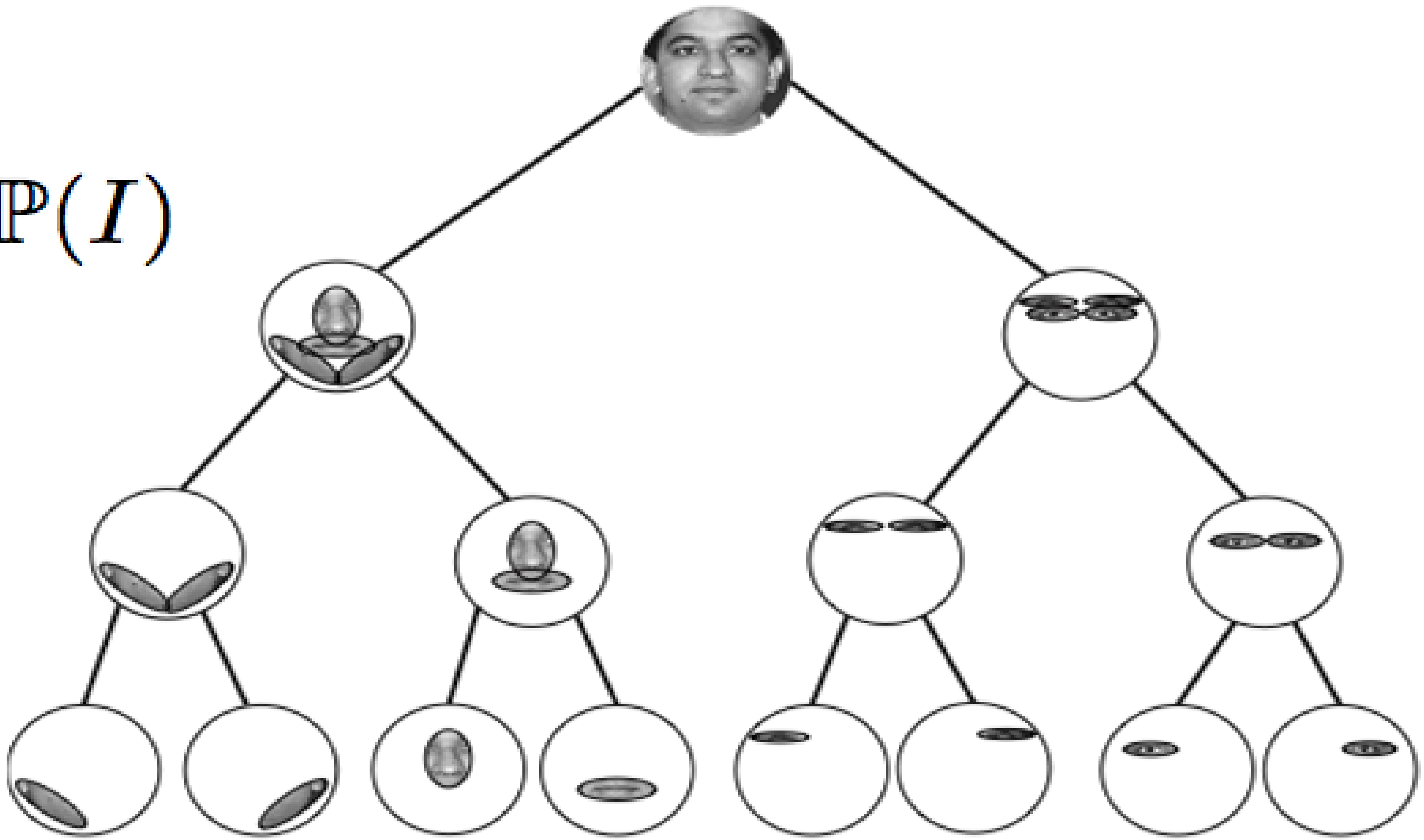
Rigid ⟶ Deformable

“Black/White” Data Model ⟶ Intensity Model

Hand-Crafted Probabilities ⟶ Learned Probabilities

Face Hierarchy

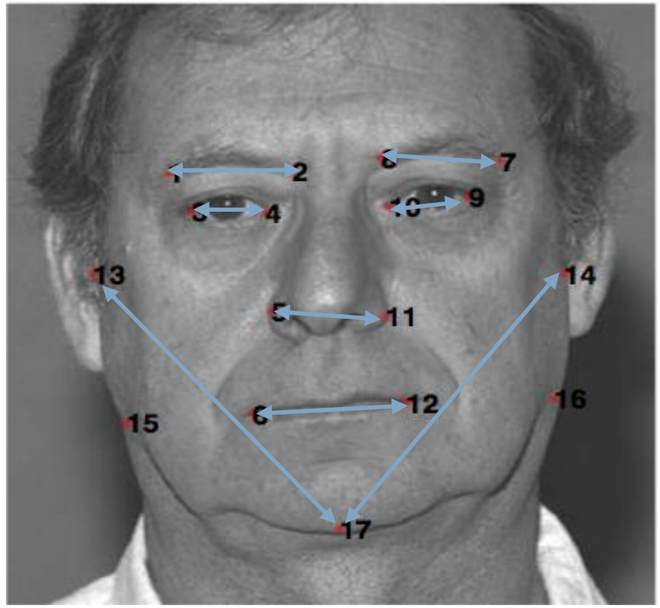
$$\mathbb{P}(I)$$

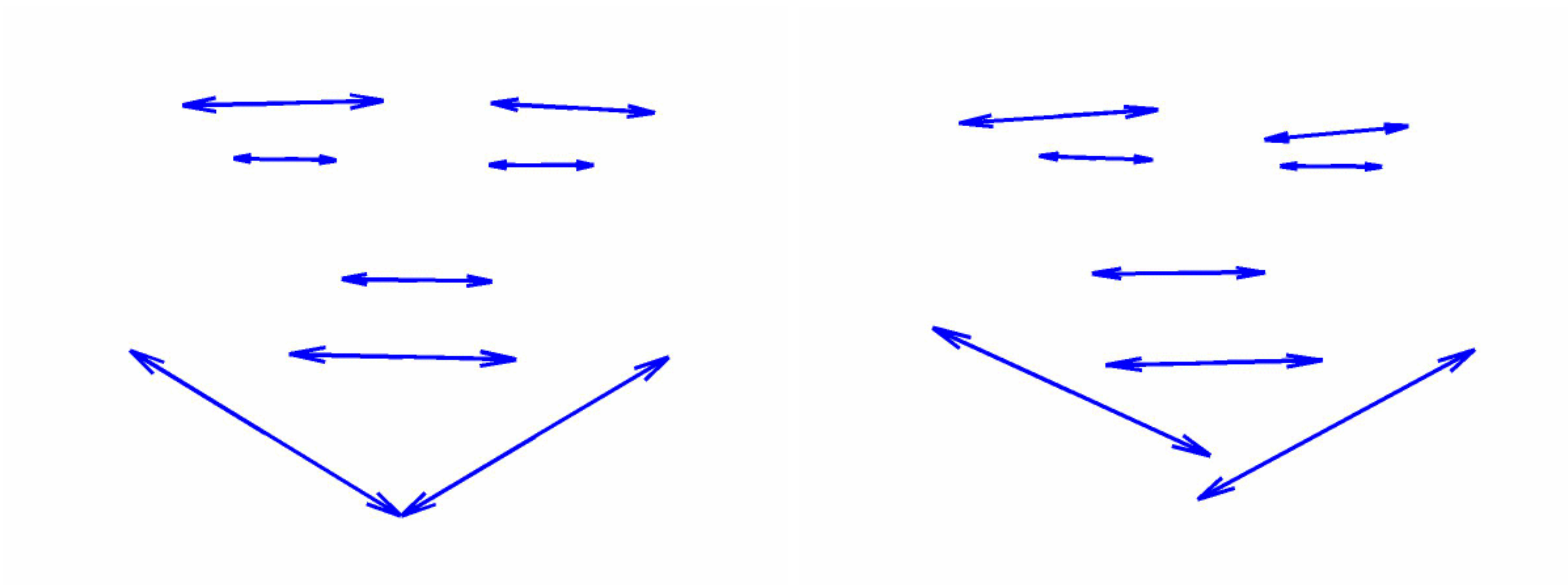


FIAS Frankfurt Institute for Advanced Studies

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

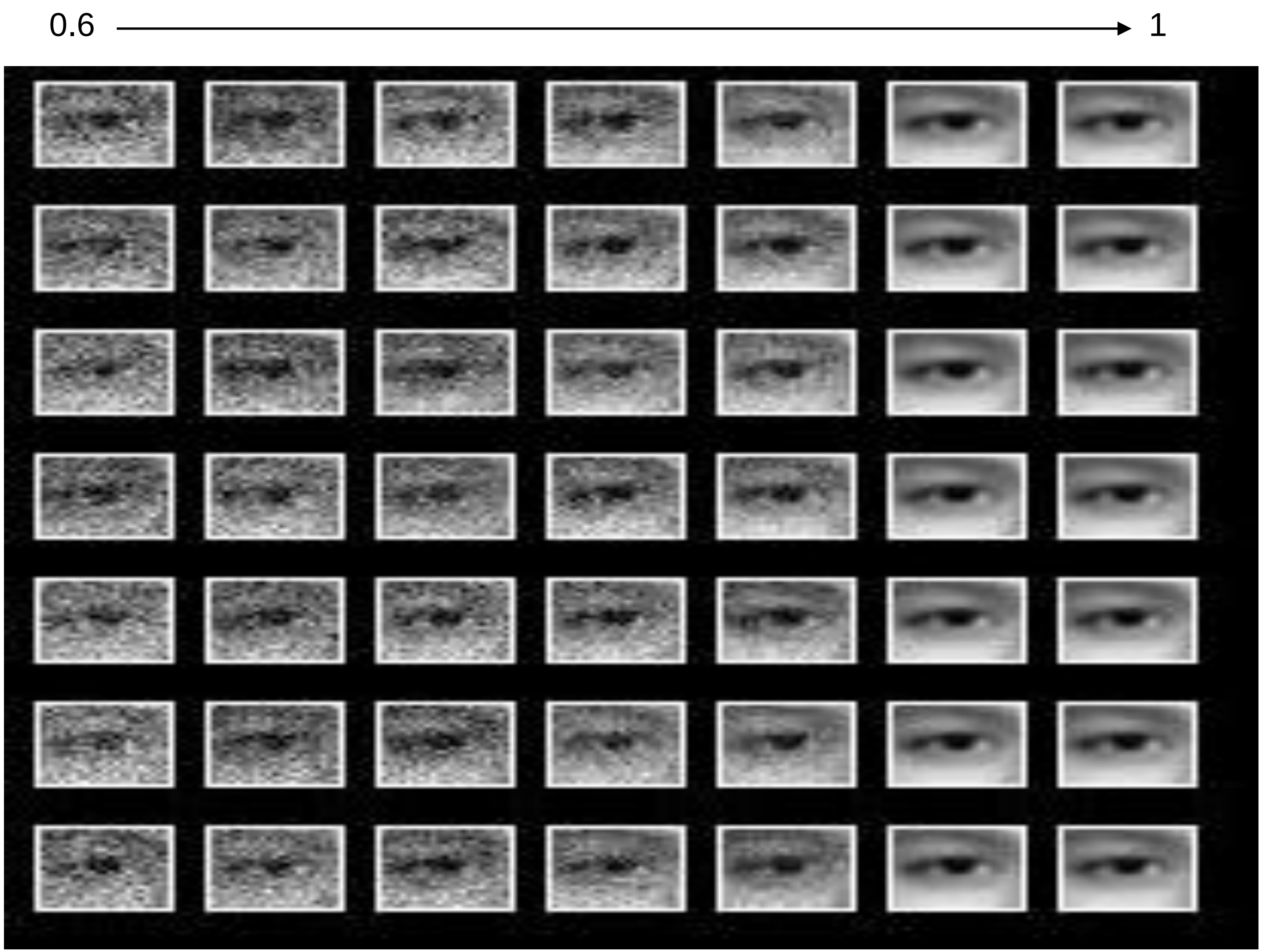
$$\mathbb{P}(I)$$





Sampling from Data Model

$$\mathbb{P}(Y|I)$$



Sampling faces from the distribution

$$\mathbb{P}(Y, I)$$

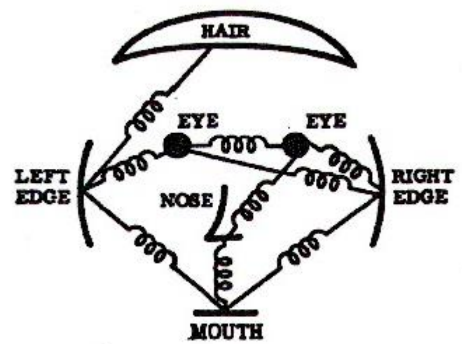
PATTERN SYNTHESIS

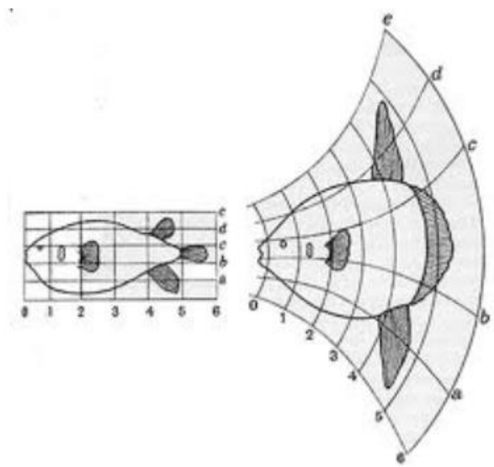
= PATTERN RECOGNITION

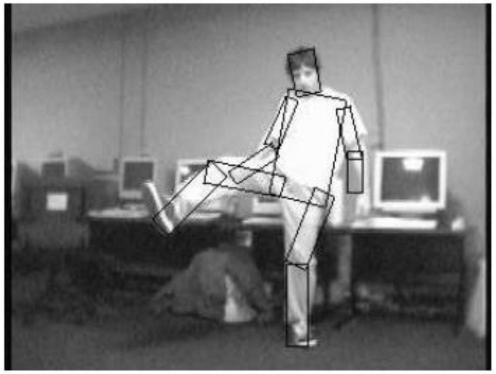
Ulf Grenander

Deformable models

- Can take us a long way...
- But not all the way

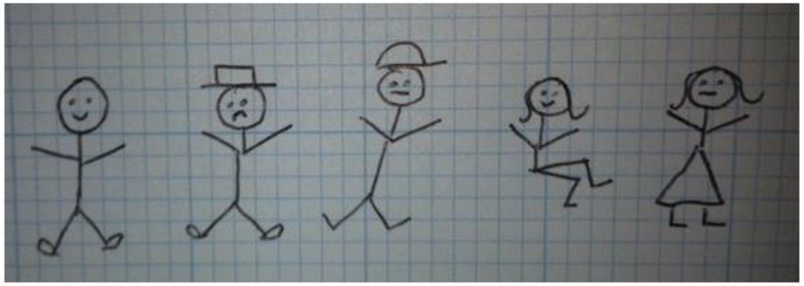






Structure Variation

- Object in rich categories have variable structure

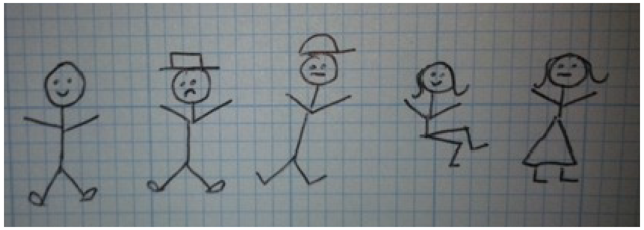


- These are NOT deformations
- There is always something you never saw before
- Mixture of deformable models? too many combined choices
- Bag of words? not enough structure
- Non-parametric? doesn't generalize

Object Detection Grammars

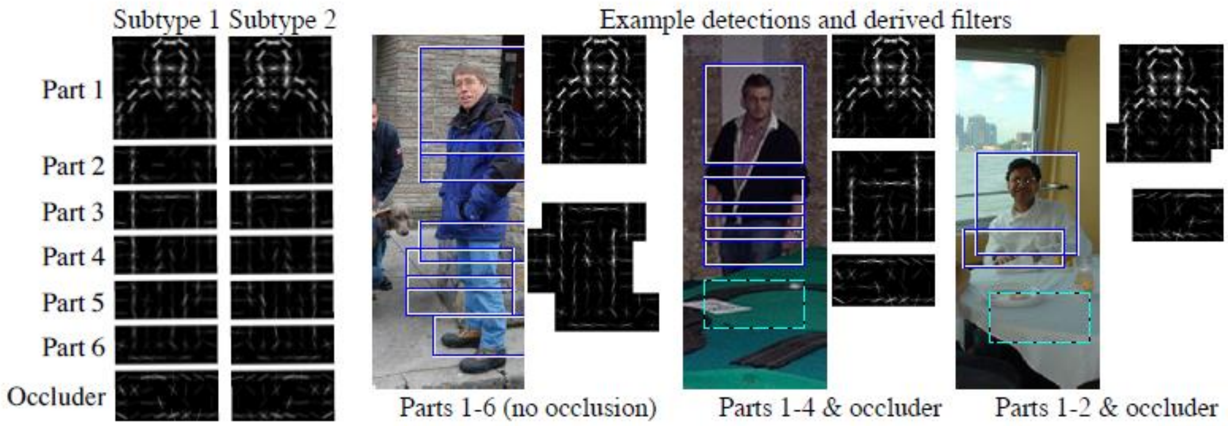
- Pictorial structure model with variable structure

- Stochastic context-free grammar
  - Generates tree-structured model
  - Springs connect symbols along derivation tree
  - Appearance model associated with each terminal



- person -> face, trunk, arms, lower-part
- face -> hat, eyes, nose, mouth
- face -> eyes, nose, mouth
- hat -> baseball-cap
- hat -> sombrero
- lower-part -> shoe, shoe, legs
- lower-part -> bare-foot, bare-foot, legs
- legs -> pants
- legs -> skirt

Person Detection Grammar



- Instantiation includes a variable number of parts
  - 1,...,k and occluder if k < 6
- Parts can translate relative to each other
- Parts have subtypes
- Parts have deformable sub-parts (not shown)
- Beats all other methods on PASCAL 2010 (49.5 AP)

Pattern Grammars as And-Or trees (Zhu, Mumford)

- Universal And-Or Tree can have an infinite size (as in the example)
- Rules are explicitly named (r1 , r2 , …)
- Each or-node A have one child for each rule having A at its left side
- A parsing tree is a sub-graph of a universal and-or tree

Grammar - And-Or trees

Grammar

$V_T = \{a, b\}$
$V_N = \{S\}$
$R = \{r_1 : S \to aS,\ r_2 : S \to b\}$

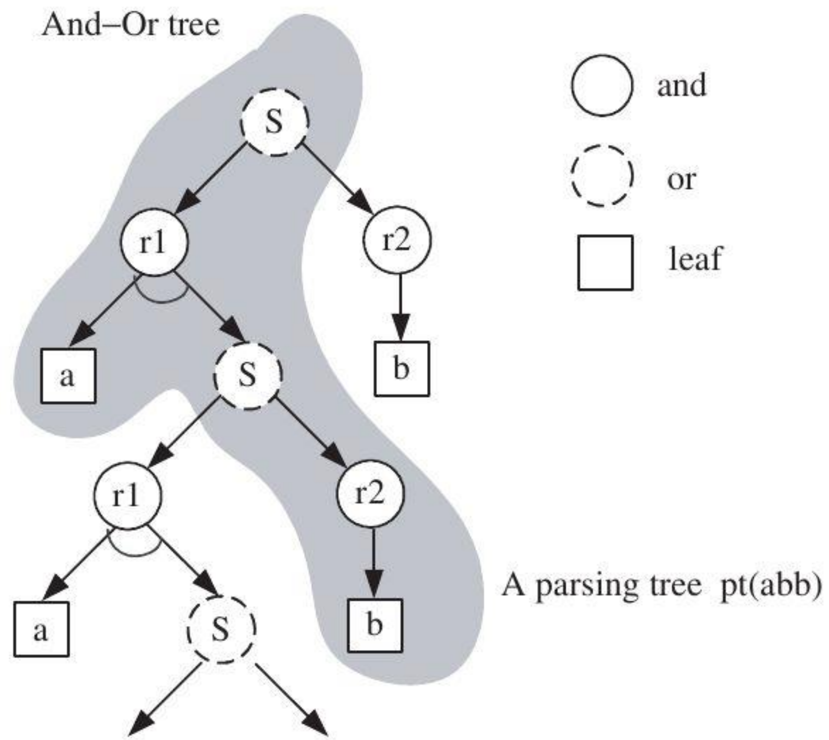


Fig. 2.2 A very simple grammar, its universal And–Or tree and a specific parse tree in shadow.

Visual vs Text Grammars

- No left-to-right ordering in language

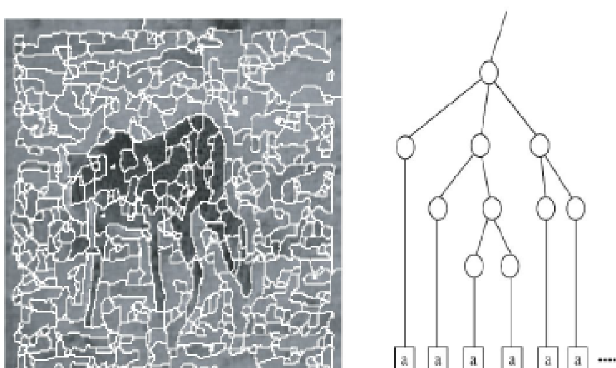
Solution: explicitly add horizontal edges to represent adjancy

- Objects appear in arbitrary scales

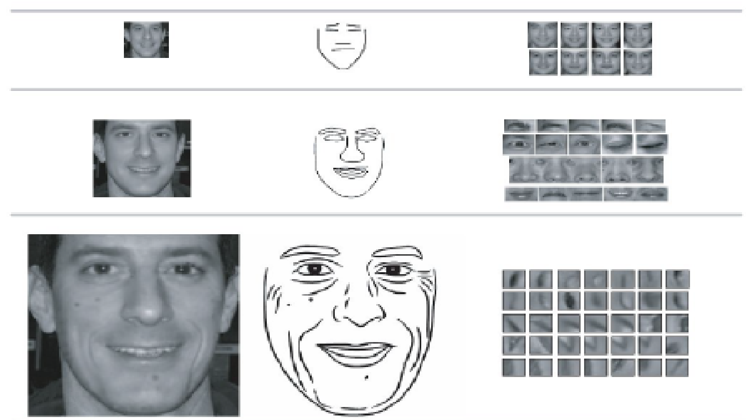
Solution: termination rules at different levels (higher leaves)

- Much wider spectrum of quite irregular local patterns

Solution: combine Markov random fields with stochastic grammars



(a) (b)



images sketches primitives

Contextual Information

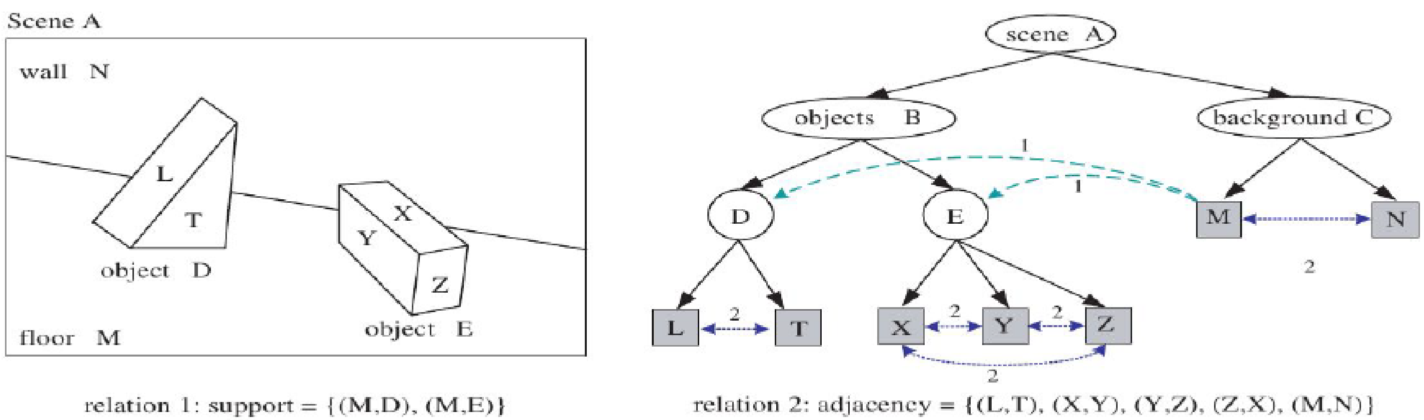


Fig. 2.11 A parser tree for a block world from [22]. The ellipses represents non-terminal nodes and the squares are for terminal nodes. The parse tree is augmented into a parse graph with horizontal connections for relations, such as one object supporting the other, or two adjacent objects sharing a boundary.

- Horizontal lines to represent relations and constraints:
  - Bonds and connections (more dense)
  - Joints and junctions
  - Interactions and semantics (less dense). E.g.: person eating an apple

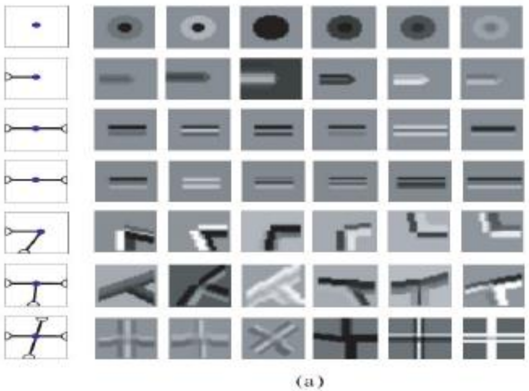
Stochastic Model

- Probabilities for rules (stochastic grammars). One local probability at each Or-node to account for the relative frequency of each alternative
- Probabilities of relations (Markov random fields). Local energies associated with each horizontal link.
- A Configurations is a “word” of the “visual language”.

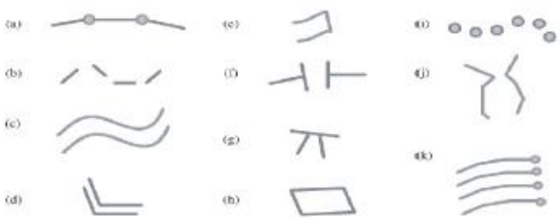
Visual Vocabulary

- Bonds – topological information
- Three hierarchical and connected levels

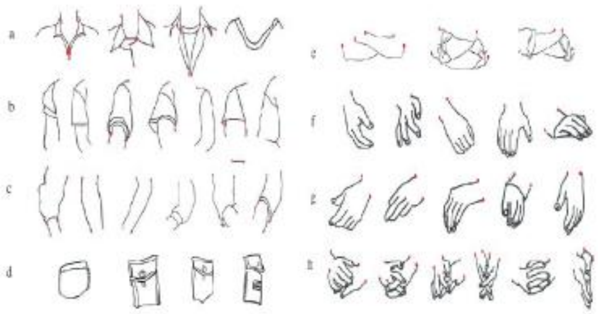
Image **primitives**: Textons (blobs, bars, terminators and crosses)



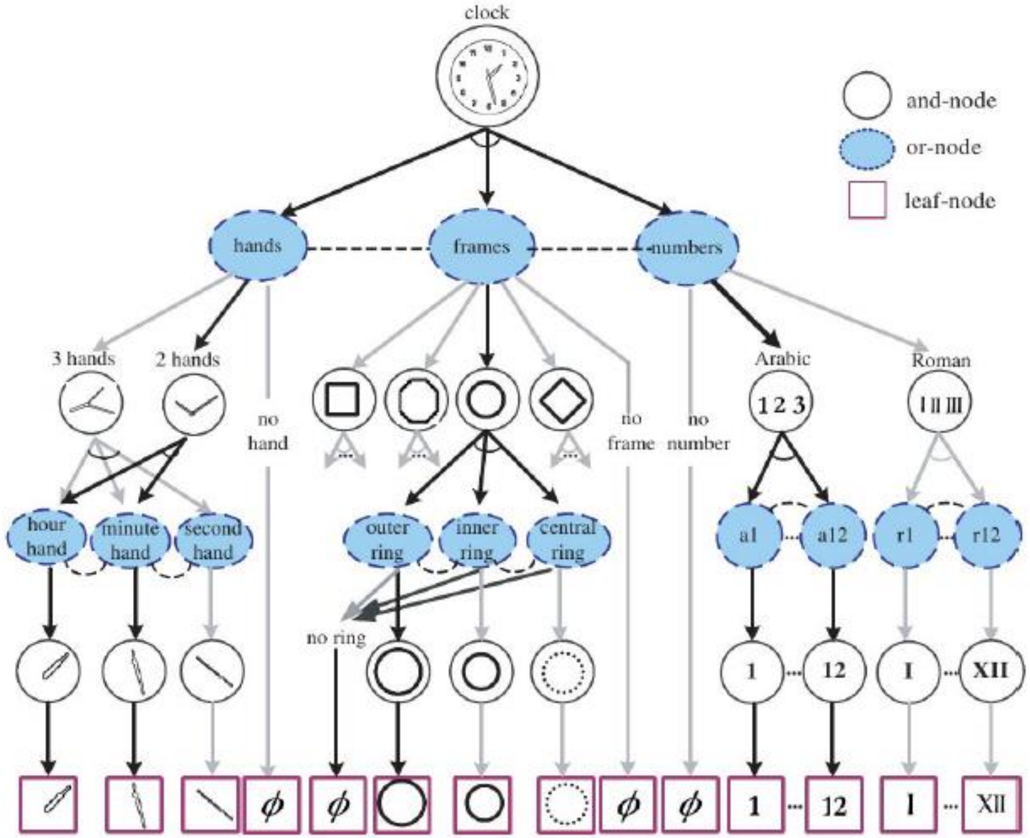
Geometric **groupings**: Graphlets



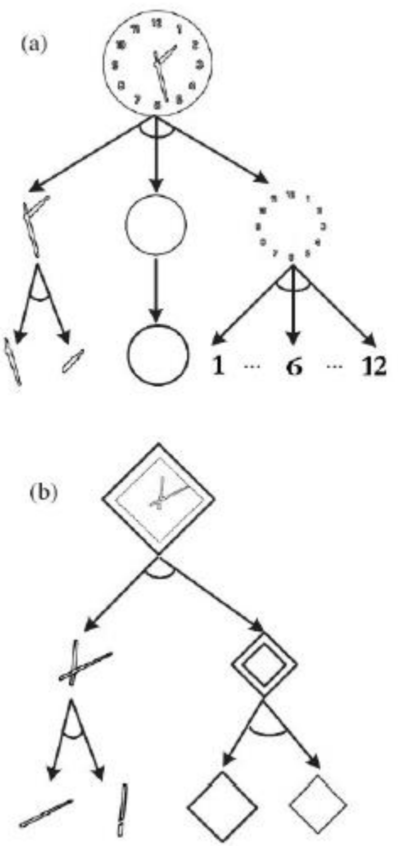
Object parts



Clock Example



And-Or Graph (Grammar)



And-Or Parse Graphs

Learning and Estimation with And-OR graphs

- Main elements to be learned: (1) Vocabulary and And-Or tree, (2) Relations – Horizontal Line and (3) Parameters
- What is available (training data): Images and parse trees (manually constructed ground-truths)
- Three phases:
  - Learning parameters from training data given relations and vocabulary (gradient method)
  - Learning news relations given vocabulary and learned parameters (inspired in texture synthesis)
  - Learning vocabulary and And-Or tree

Image Parsing??

Backup