Learning the Semantics of Words and Pictures

Kobus Barnard, Pinar Duygulu, Nando de Freitas, and David Forsyth

UC Berkeley and UBC

Funding provided by NSF Digital Library Initiative II.
Kobus Barnard also receives funding from NSERC (Canada)
Pinar Duygulu also receives funding from TUBITAK (Turkey)

The Battle Plan

Survey the domain

Introduce the approach

Apply to browsing, searching, auto-illustrate

Attach words to pictures (auto-annotate)

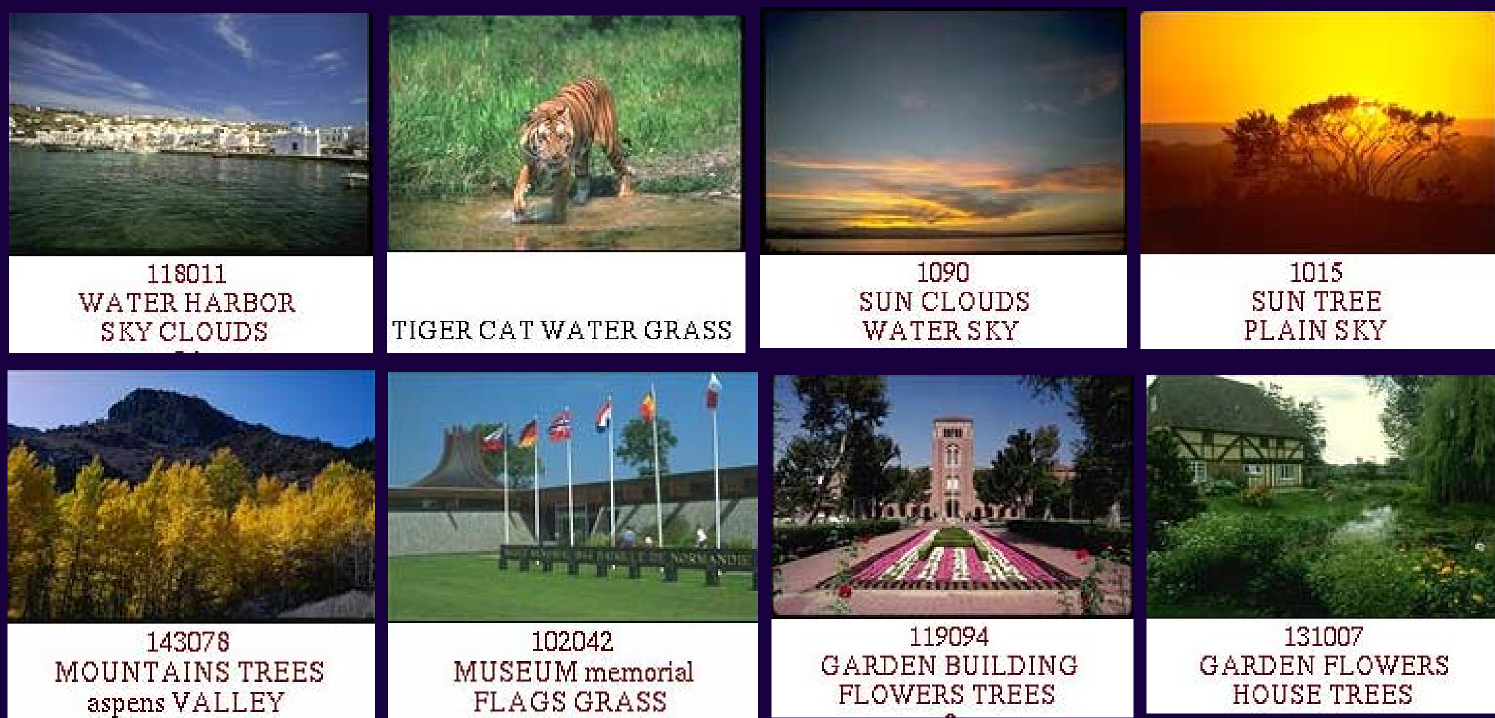
Compare image segmentation methods

Attach words to image regions (recognition)

Data Examples

Corel Image Data	40,000 images
Fine Arts Museum of San Francisco	83,000 images online
Cal-flora	20,000 images, species information
News photos with captions (yahoo.com)	1,500 images per day available from yahoo.com
Hulton Archive	40,000,000 images (only 230,000 online)
internet.archive.org	1,000 movies with no copyright
TV news archives (televisionarchive.org, informedia.cs.cmu.edu)	Several terabytes already available
Google Image Crawl	>330,000,000 images (with nearby text)
Satellite images (terrarserver.com, nasa.gov, usgs.gov)	(And associated demographic information)
Medial images	(And associated with clinical information)

Corel Database



392 CD’s, each consisting of 100 annotated images.

FAMSF Data (83,000 images online)

Web number: 4359202410830012

rec number: 2

Description:
serving woman stands in a dressing room, in front of vanity with chair, mirror and mantle, holding a tray with tea and toast

Title: Le Matin

Primary class: Print

Display date: 1886

Artist: Tissot

Country: France

Approaches to Finding Pictures

Meta-data indexing (keywords)

Content based image retrieval (query by example using global features, e.g. colour histograms)

Many papers, including [Flickner et al., 95; Carson et al., 99; Wang, 00]

Query by example with relevance feedback

Many papers including[Cox et al 00; Santini 00; Schettini, 02]

Keywords: rose flower plant leaves

Query on

“Rose”

Example from Berkeley Blobworld system

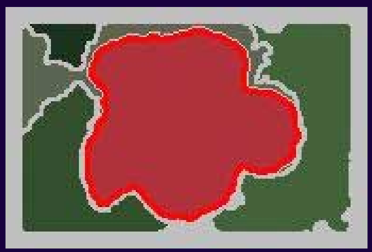
Query on

Example from Berkeley Blobworld system

Query on

“Rose”

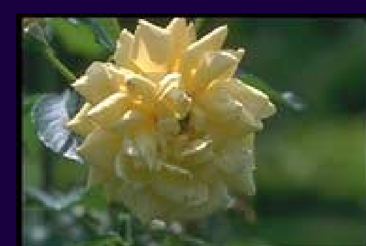
and



Example from Berkeley Blobworld system

Appearance counts!

Semantics counts!



Difficulties arising in more “real” applications

Images may not have keywords

(An image is worth … how many key-words?)

Real user queries are not easily satisfied using keywords

What will users pay for?

Work by Enser and others on real queries collected by photo librarians

Sample queries [Armitage and Enser, 97]

"… images of Native Americans or others murdering colonists' children especially babies …"

"The depiction of vanity in painting, the depiction of the female figure looking in the mirror, etc."

"Cheetahs running on a greyhound course in Haringey in 1932"

Approach

It looks like we need to solve the AI problem? (too ambitious)

Philosophy--move in this direction but in manageable steps with useful intermediate results

The Battle Plan

~~Survey the domain~~

Introduce the approach

Apply to browsing, searching, auto-illustrate

Discuss probabilistic inference and model fitting

Attach words to pictures (auto-annotate)

Compare image segmentation methods

Attach words to image regions (recognition)

Input

Image
processing*

Each blob is a large vector of features

"This is a picture of the sun setting over the sea with waves in the foreground"

Language
processing

sun sky waves sea

* Thanks to Blobworld team [Carson, Belongie, Greenspan, Malik], N-cuts team [Shi, Tal, Malik]

Image Features

- Region size
- Position
- Colour
- Oriented energy (12 filters)
- Simple shape features

Natural Language Processing

- Parts of speech* (prefer nouns for now)
- Expand semantics using WordNet†
- Sense Disambiguation

* We use Eric Brill's parts of speech tagger (available on-line)

† WordNet is an on-line lexical reference system from Princeto

Multiple Senses

212001 bank buildings trees city

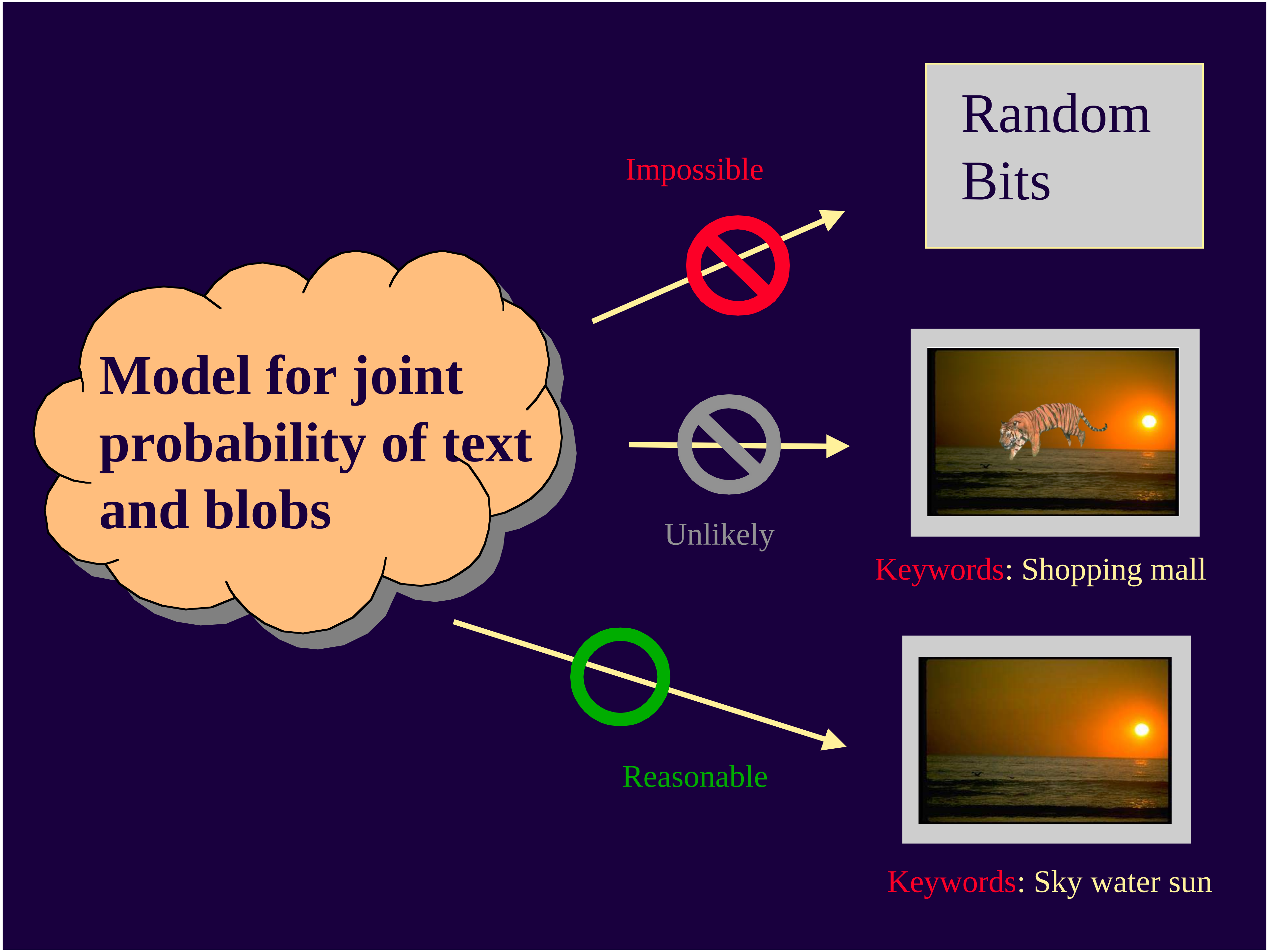
125090 bank machine money currency bills

125084 piggy bank coins currency money

26078 water grass trees bank

173044 mink rodent bank grass

151096 snow banks hills winter


Model for joint probability of text and blobs
Impossible
Random Bits
Unlikely
Keywords: Shopping mall
Reasonable
Keywords: Sky water sun

Model for joint probability of text and blobs

- Clustering models
- Aspect models
- Hierarchical models
- Bayesian models
- Co-occurrence models

Many of these based on models proposed for text [Brown, Della Pietra, Della Pietra & Mercer 93; Hofmann 98; Hofmann & Puzicha 98]

Model for joint probability of text and blobs

Hierarchical model based on Hofmann's hierarchical aspect model for text

[Hofmann 98; Hofmann & Puzicha 98]

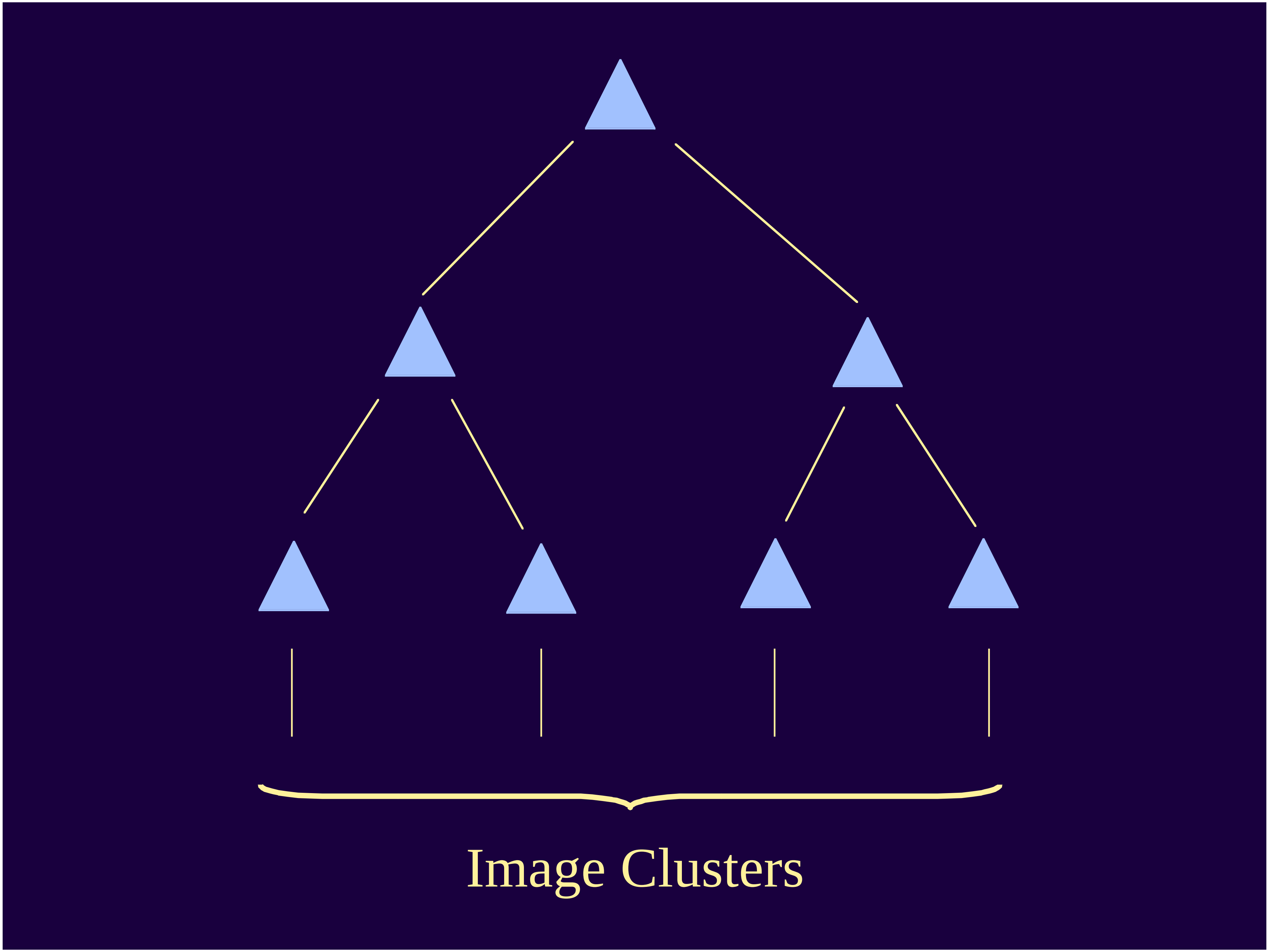
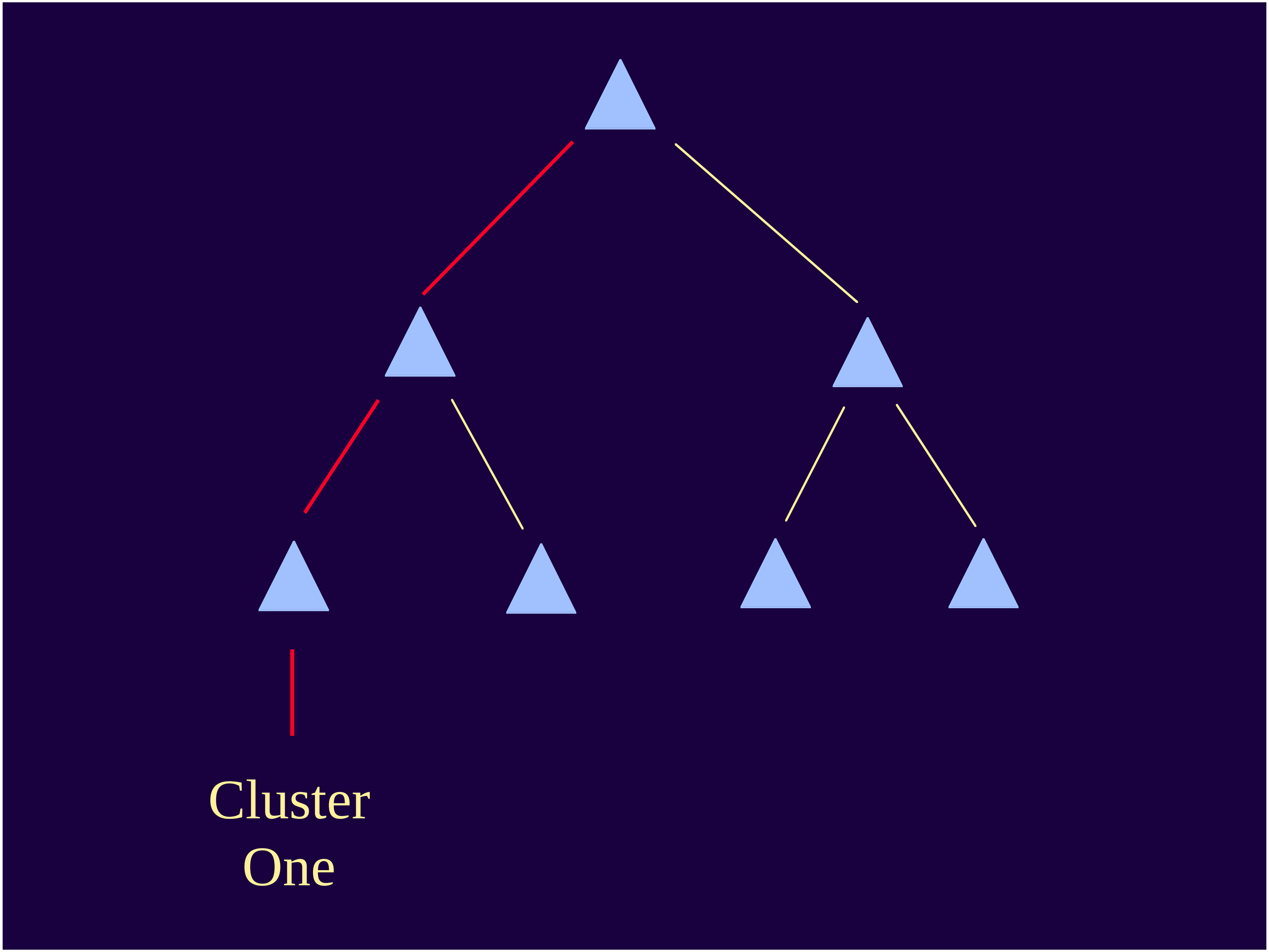
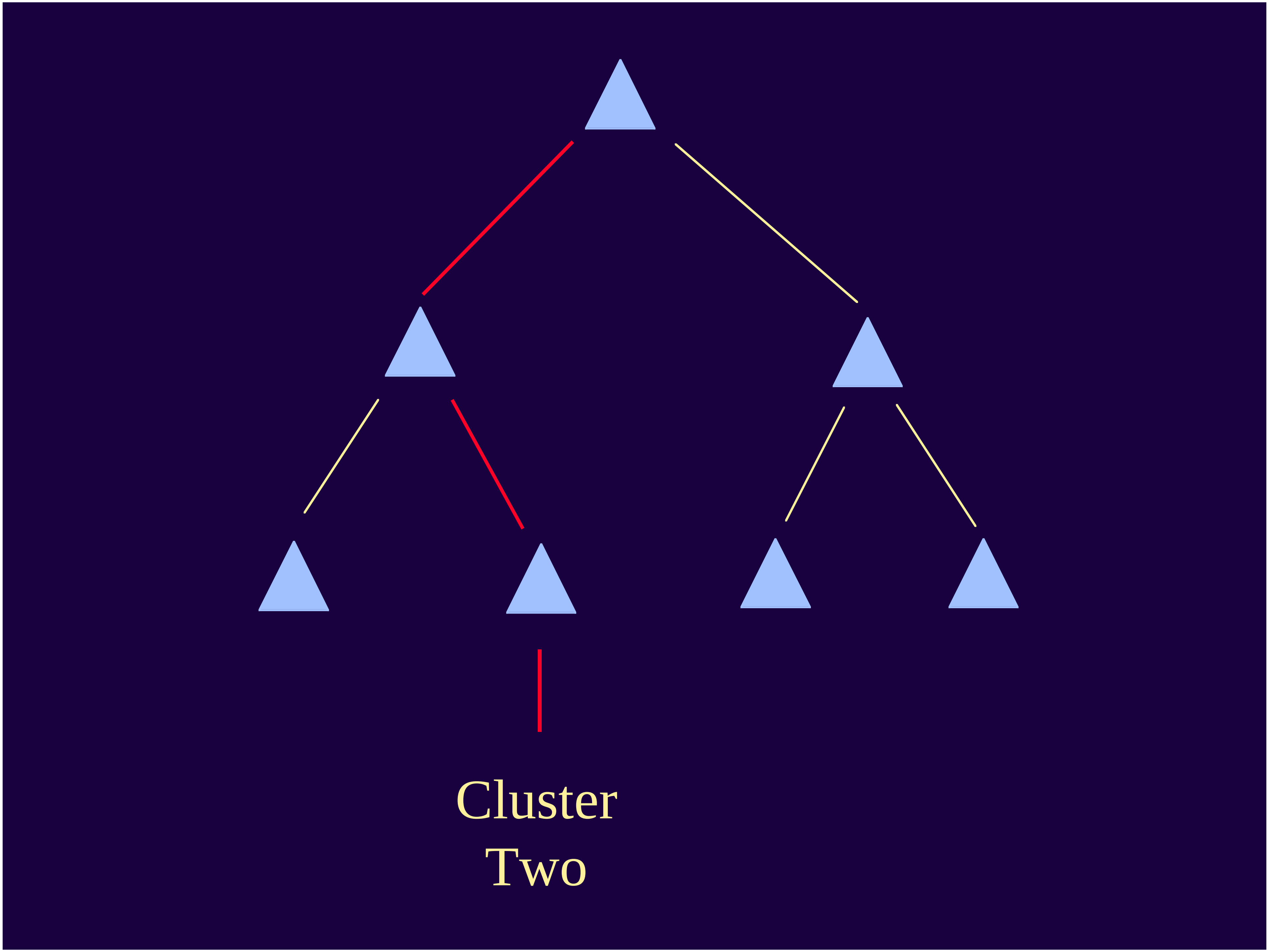
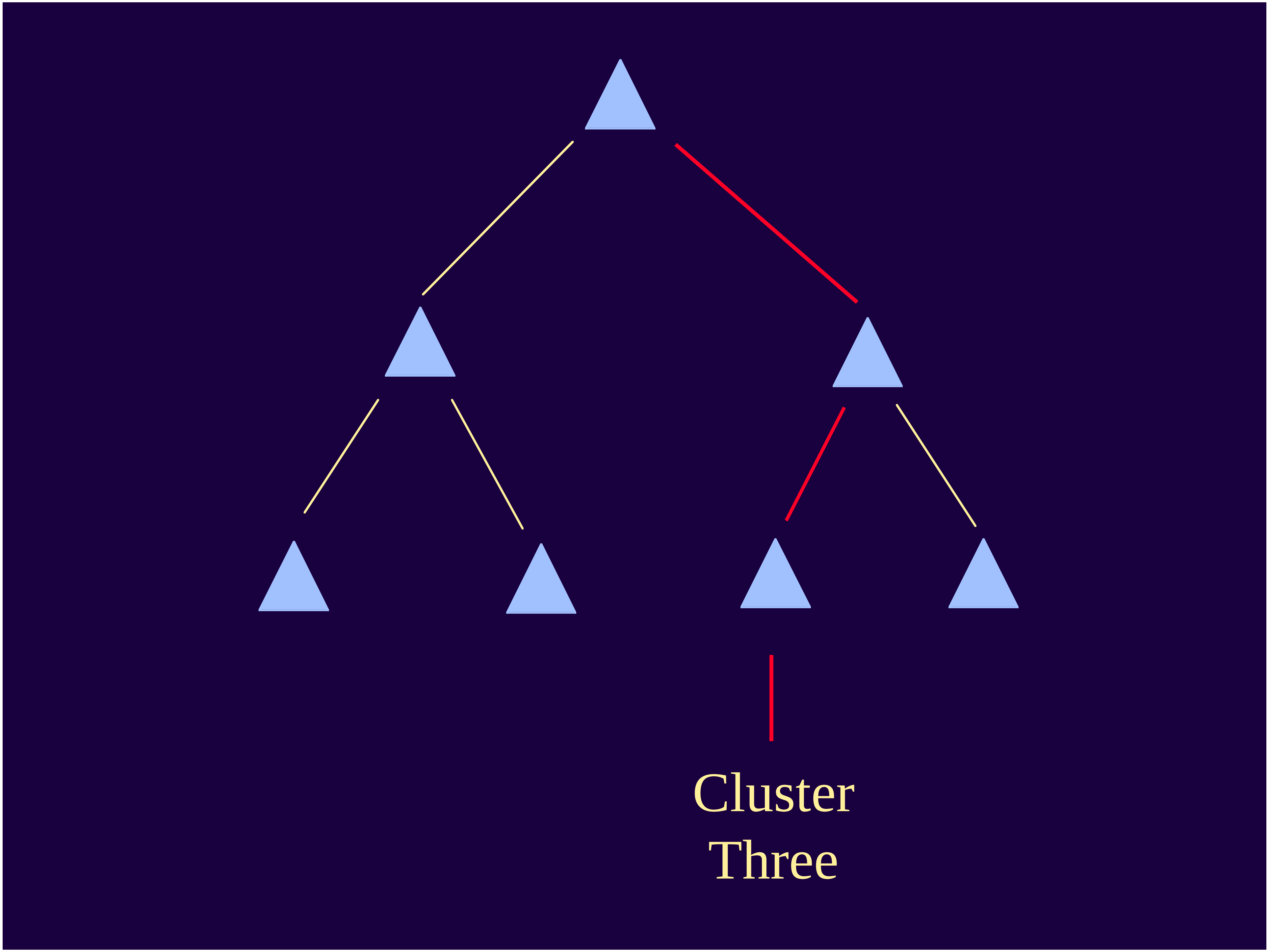
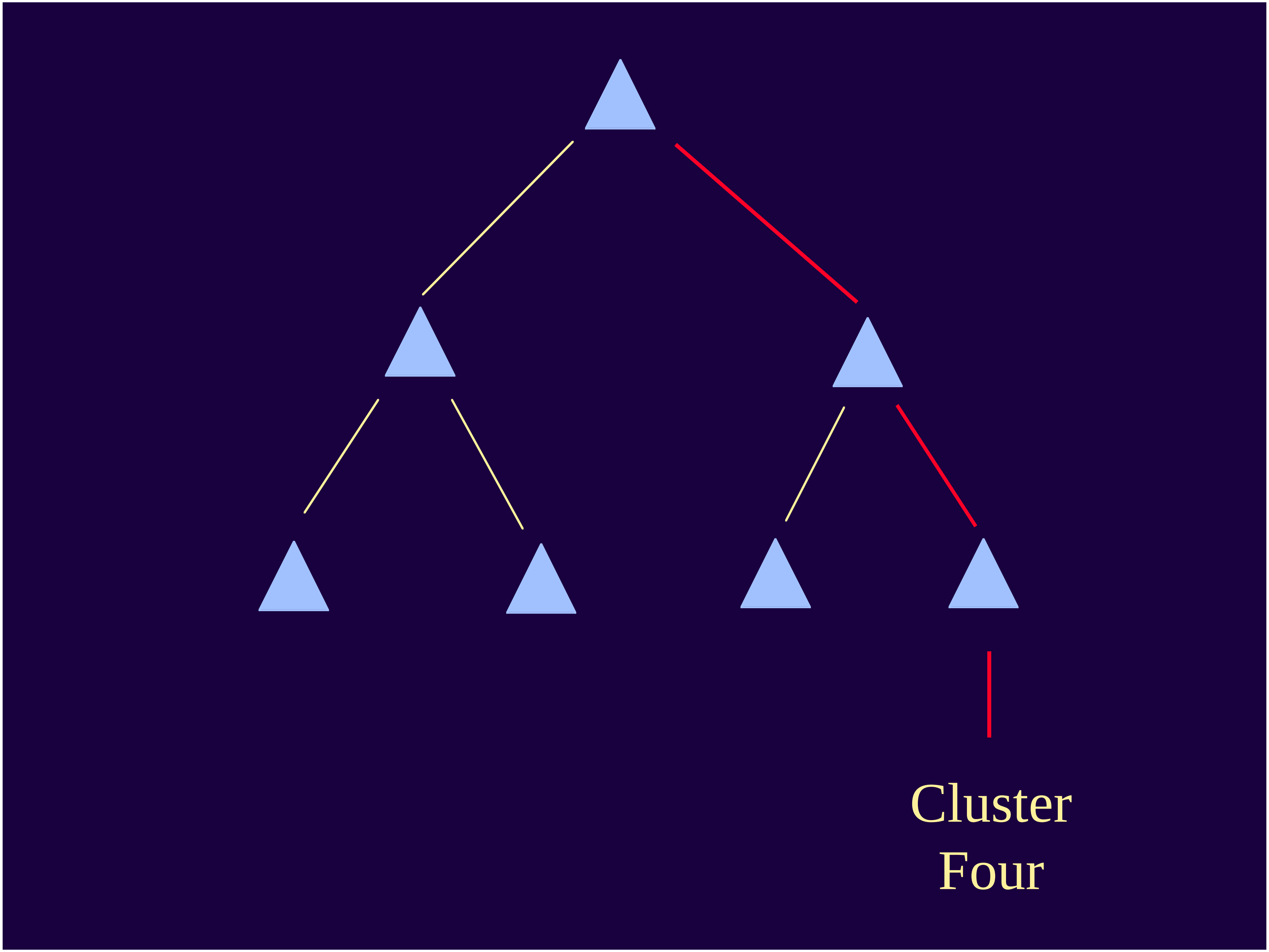


Image Clusters


Cluster
One

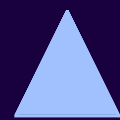

Cluster
Two


Cluster
Three


Cluster
Four

Node Behavior

Each node ....



Emits each modeled word, W_i , with some probability

Generates blobs according to a Gaussian distribution (parameters differ for each node).

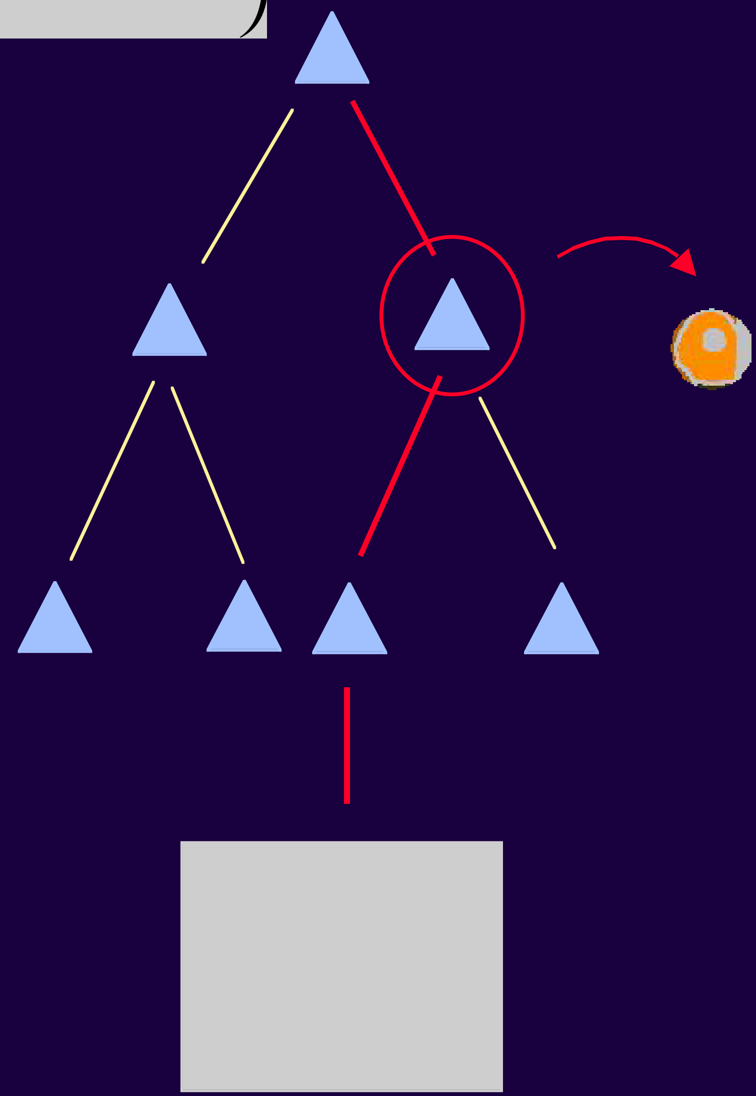
Nodes closer to the root emit more general/common words/blobs

$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 2

item
i =1

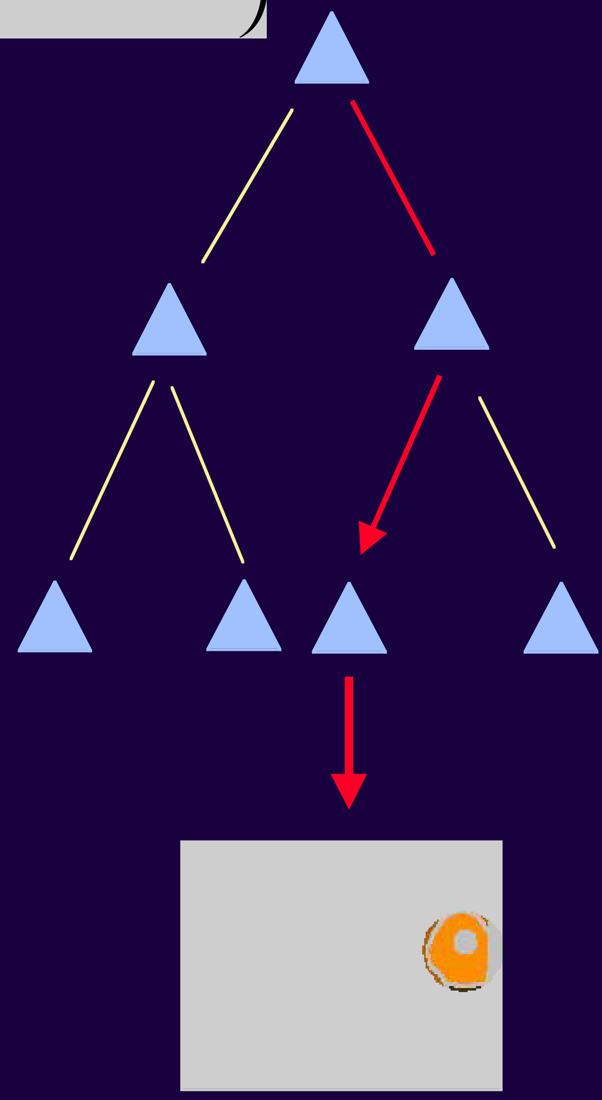


$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 2

item
i =1

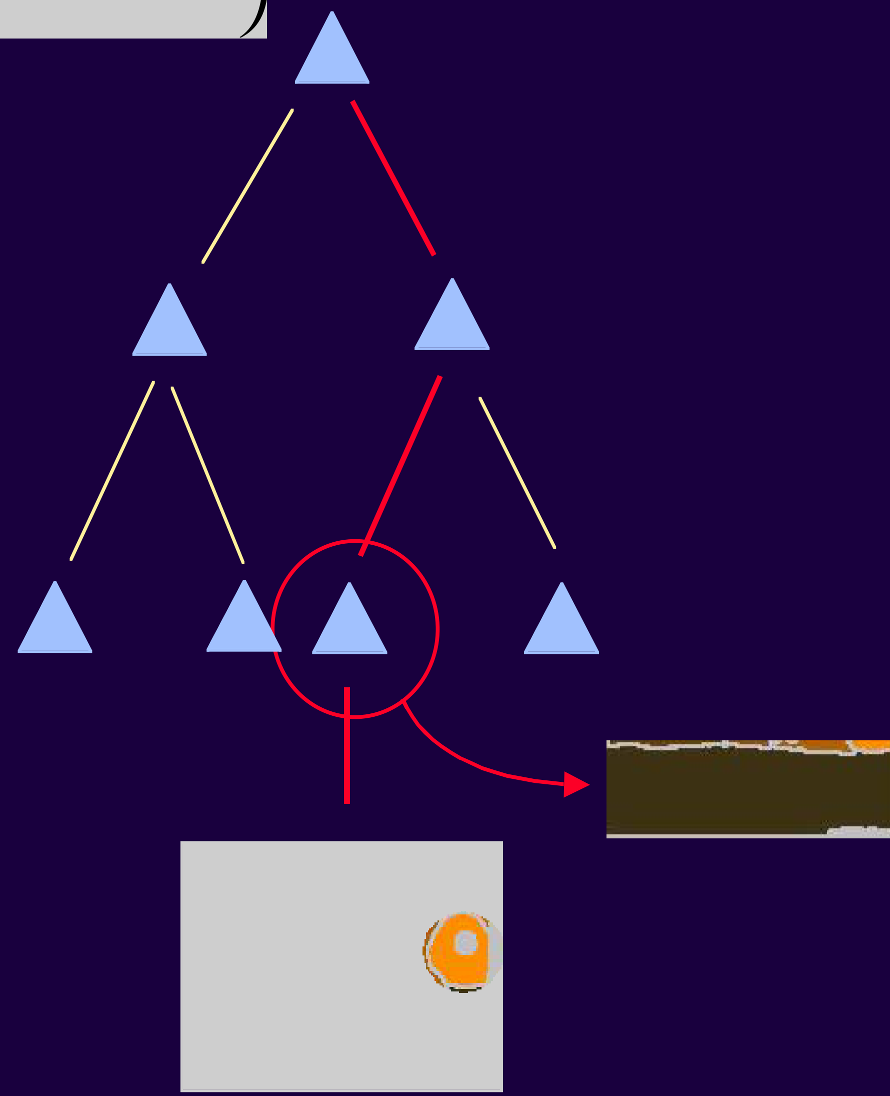


$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 3

item
i =2

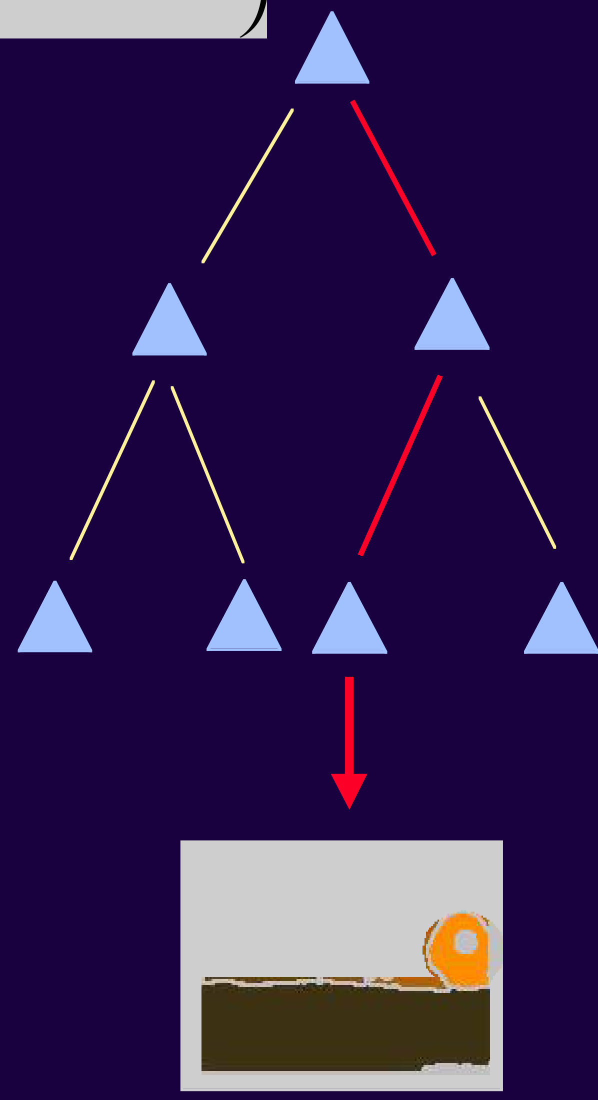


$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 3

item
i =2

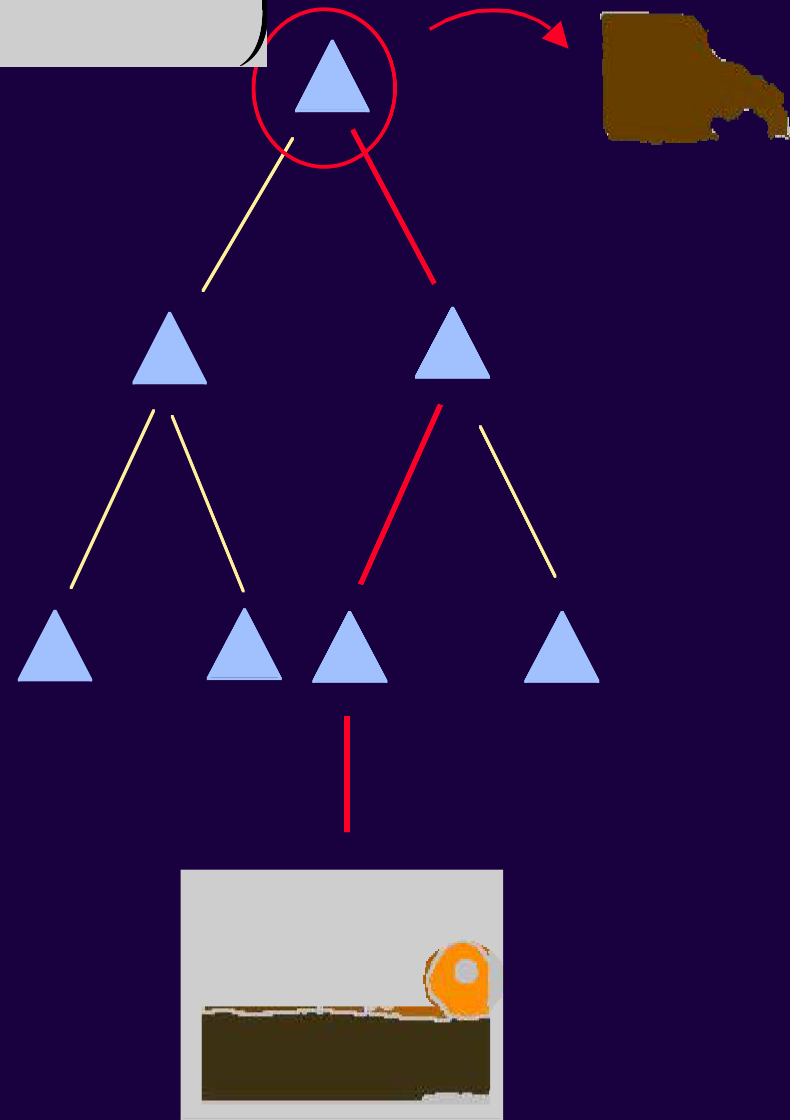


$$P(D \mid d) = \sum_c P(c) \prod_{i \in D} \left(\sum_l P(i \mid l, c) P(l \mid d) \right)$$

cluster
$c = 3$

level
$l = 1$

item
$i = 3$

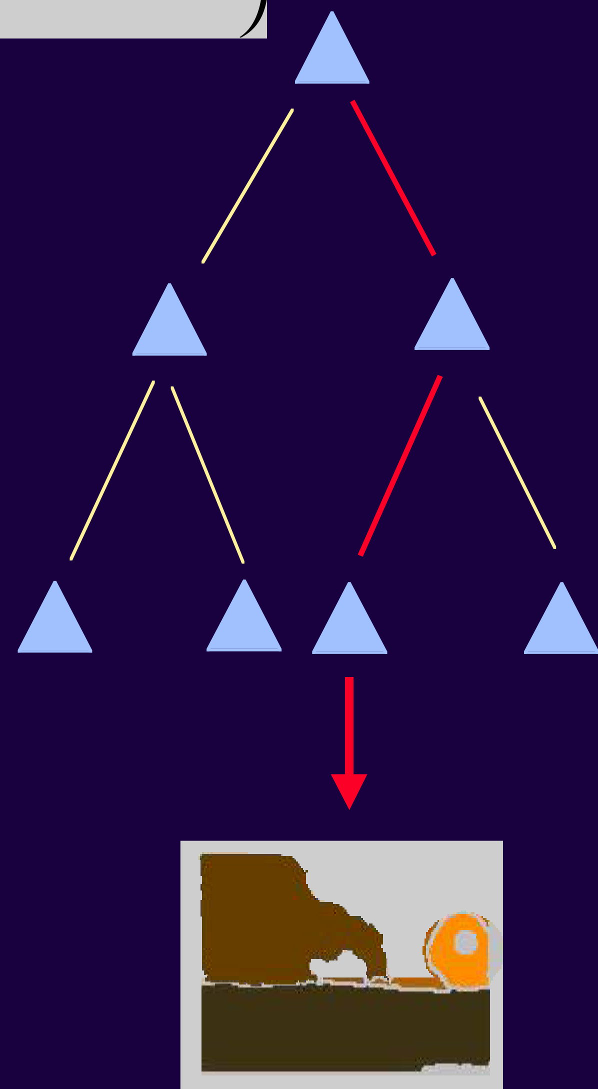


$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

item
i = 3

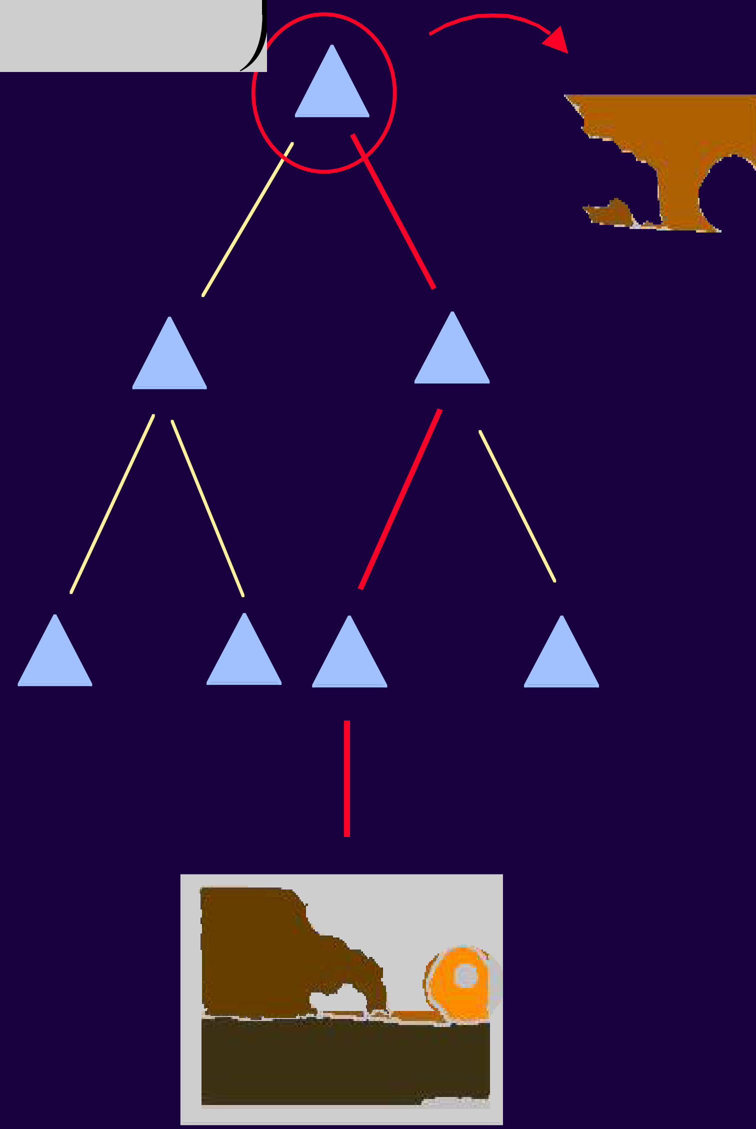


$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

item
i = 4

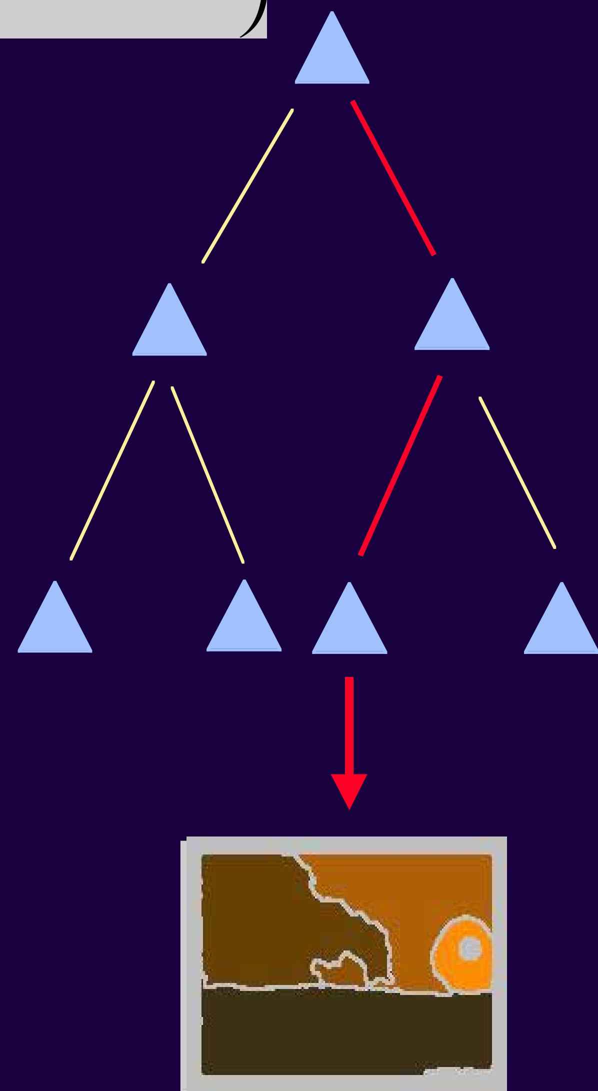


$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

item
i = 4

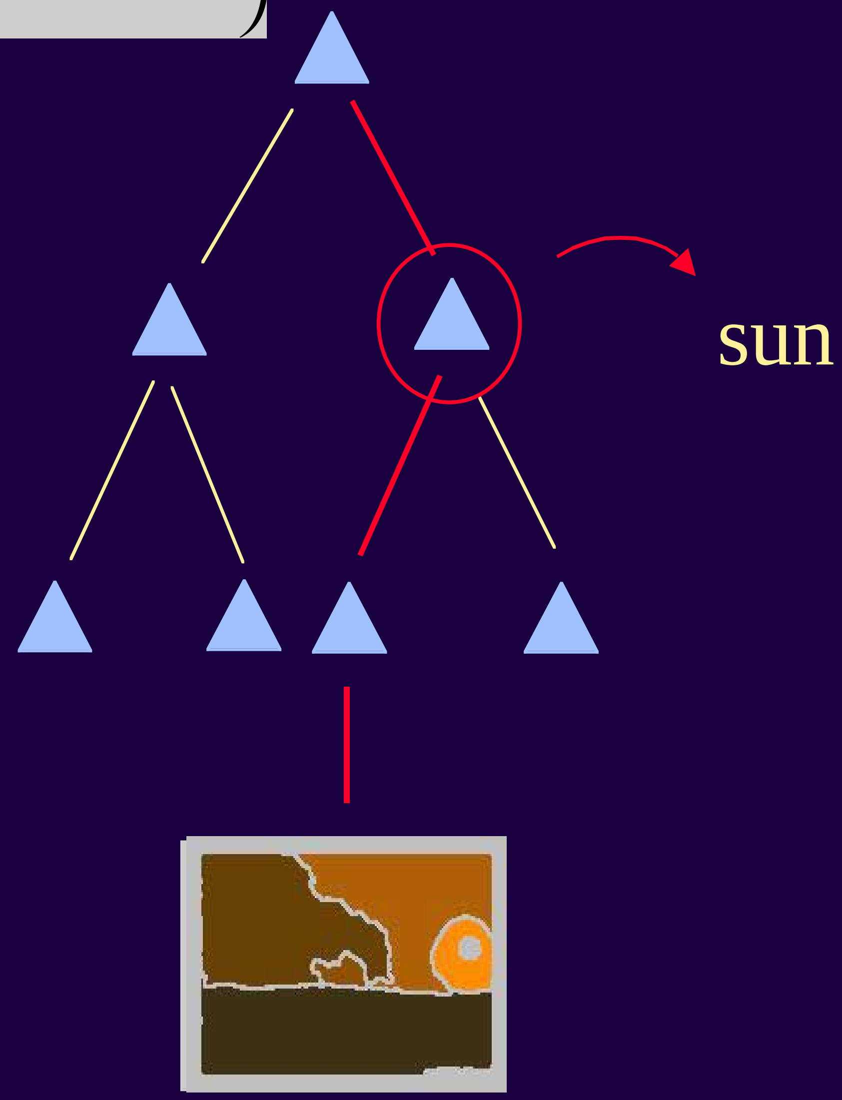


$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
$c = 3$

level
$l = 2$

item
$i = 5$

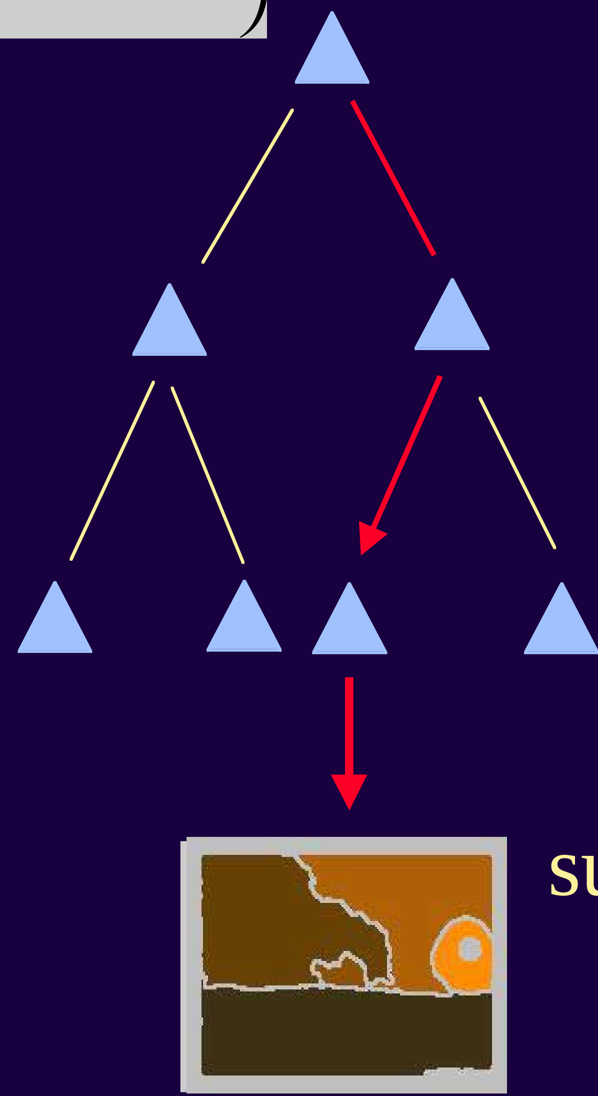


$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 2

item
i = 5



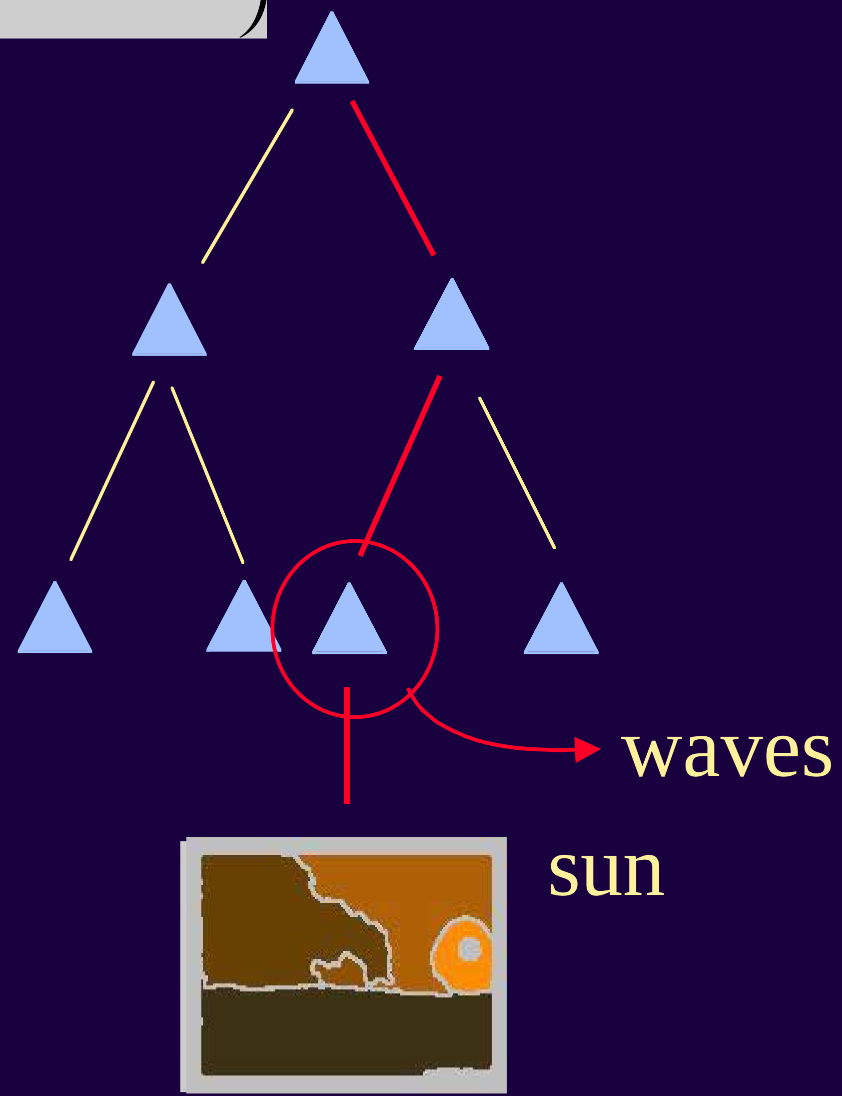
sun

$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 3

item
i = 6

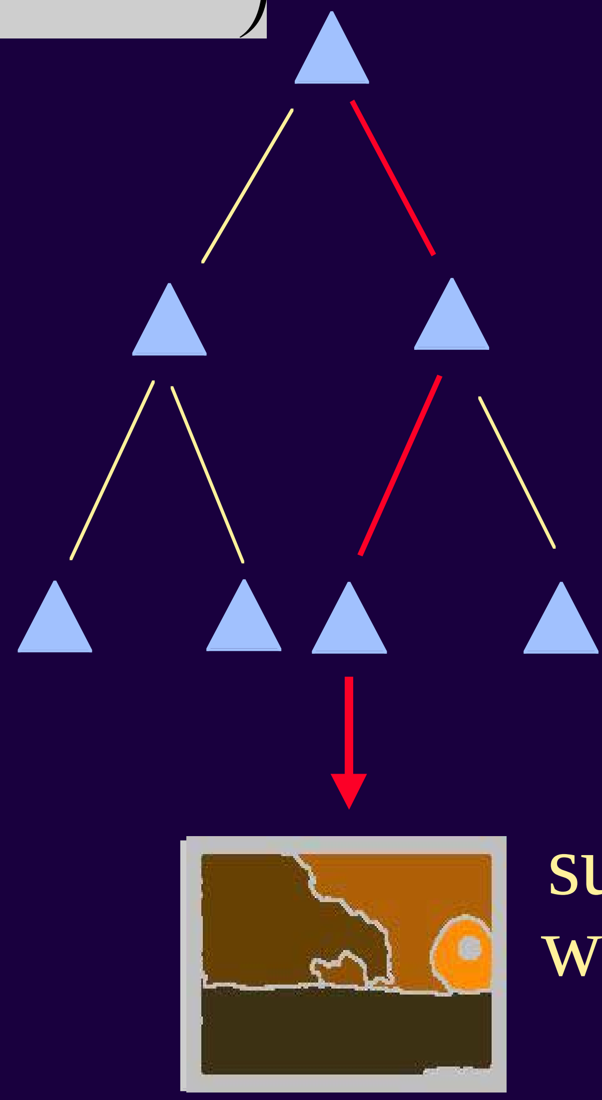


$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 3

item
i = 6



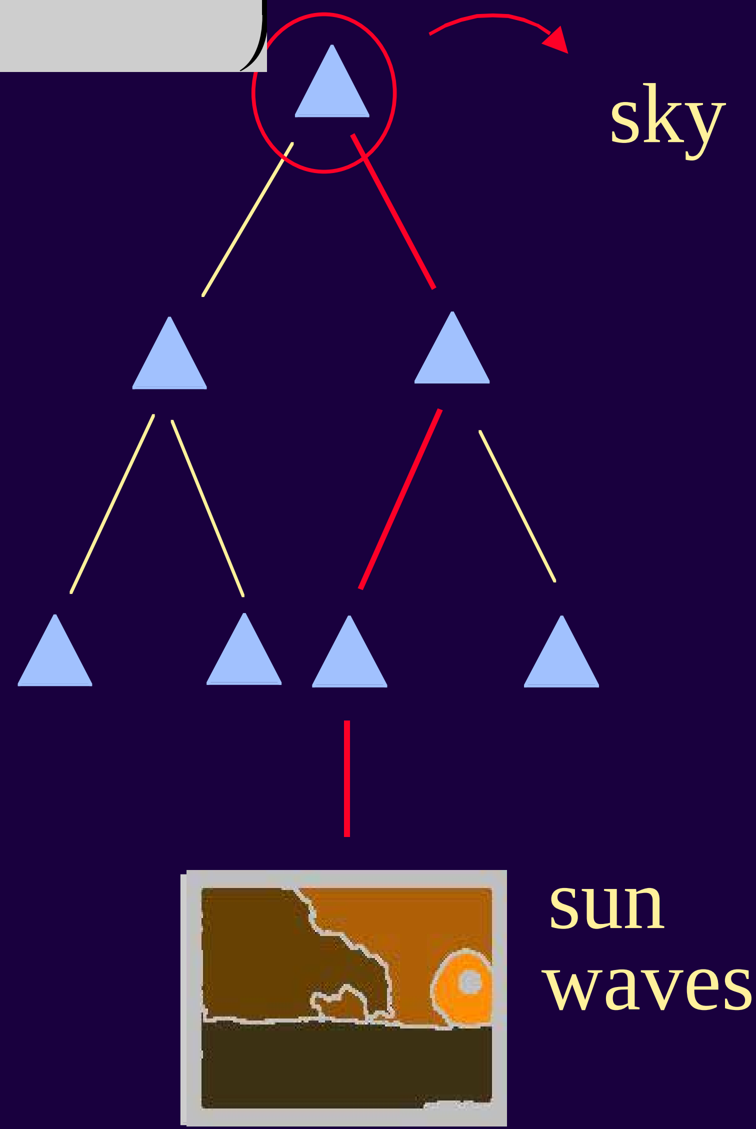
sun
waves

$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

item
i = 7

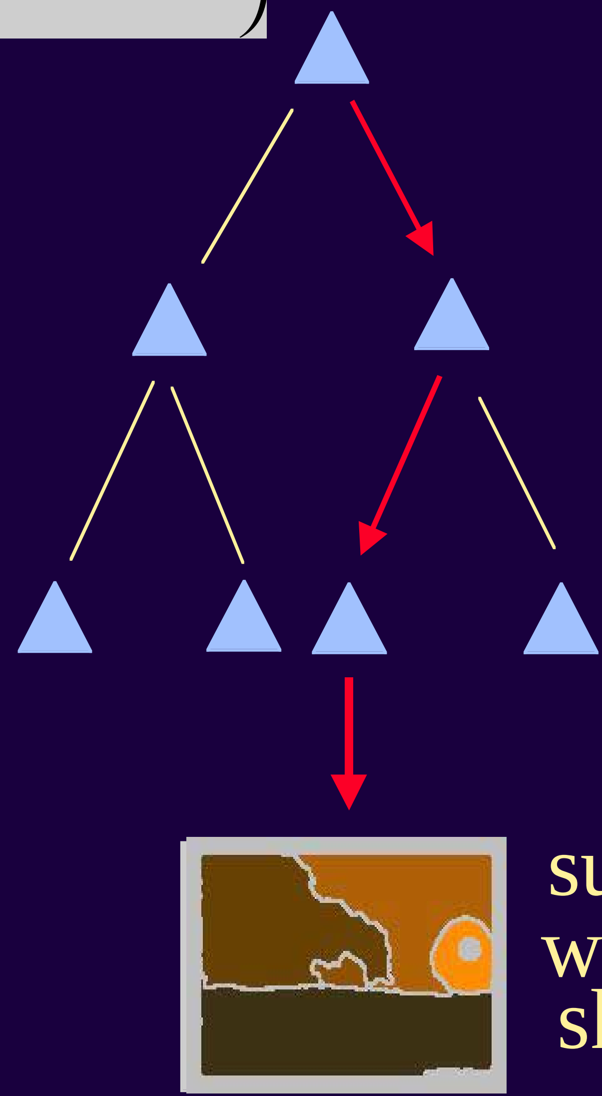


$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

item
i = 7



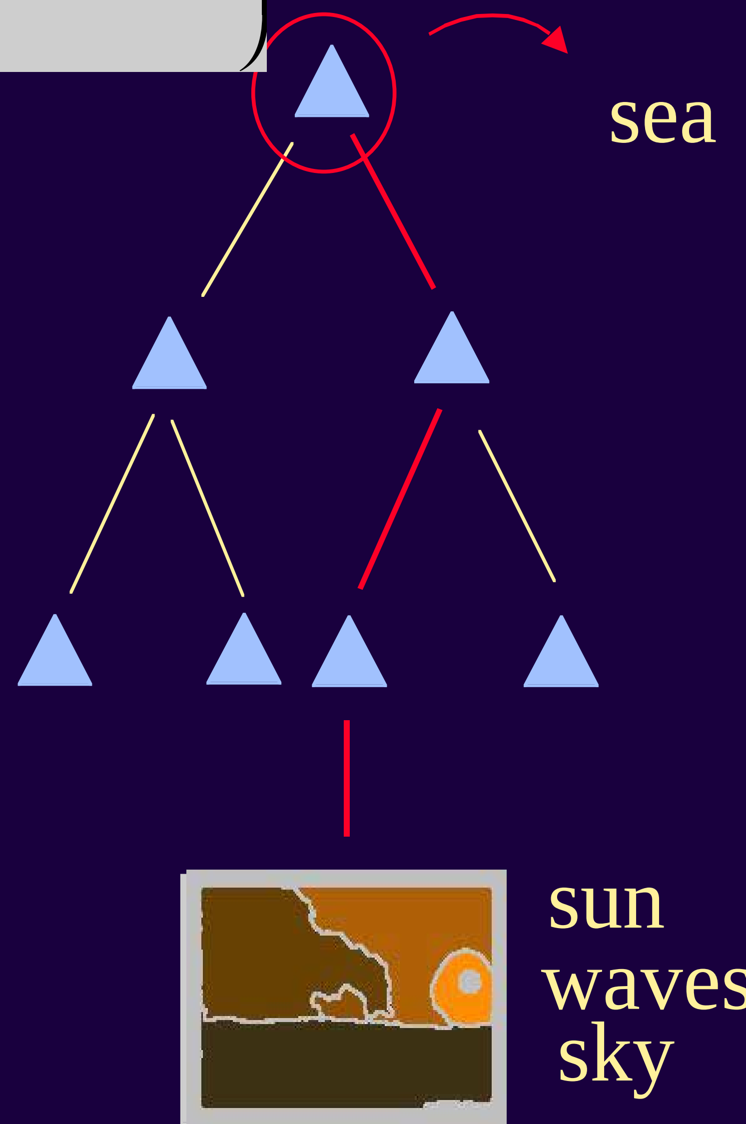
sun
waves
sky

$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

item
i =8

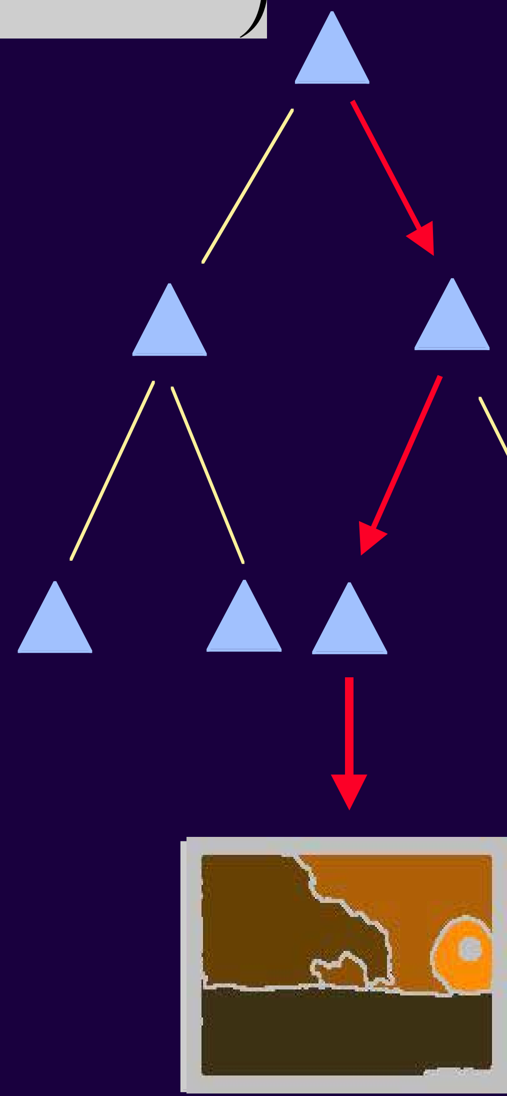


$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

item
i = 8



sun
waves
sky
sea

Motivation for Model Structure

Need to generate items (tigers, grass, water) in arbitrary combinations

Intractable to model all combinations

But want to exploit context (jungle, city)

Clusters are images drawn from the same set of nodes

The Battle Plan

~~Survey the domain~~

~~Introduce the approach~~

Browsing, searching, and auto-illustrate

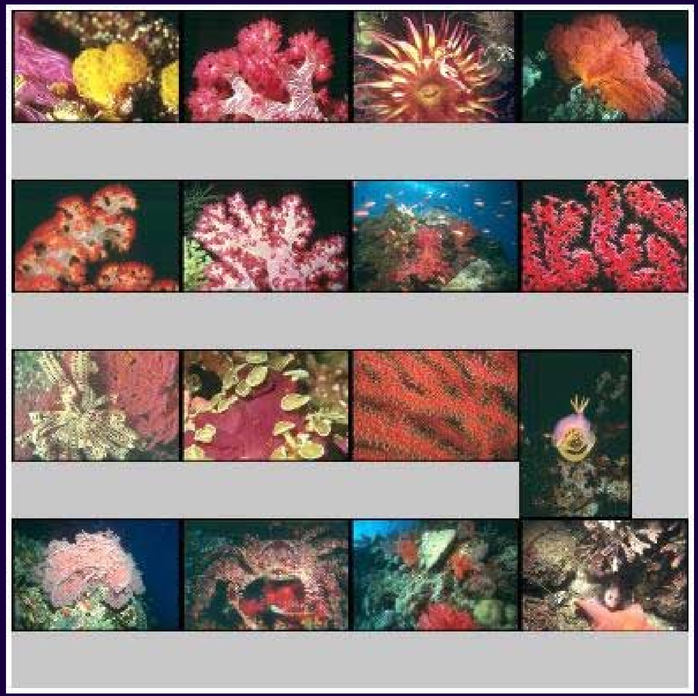
Attach words to pictures (auto-annotate)

Compare image segmentation methods

Attach words to image regions (recognition)

Cluster found using only text

Cluster found using only blob features



Adjacent clusters found using both text and blob features

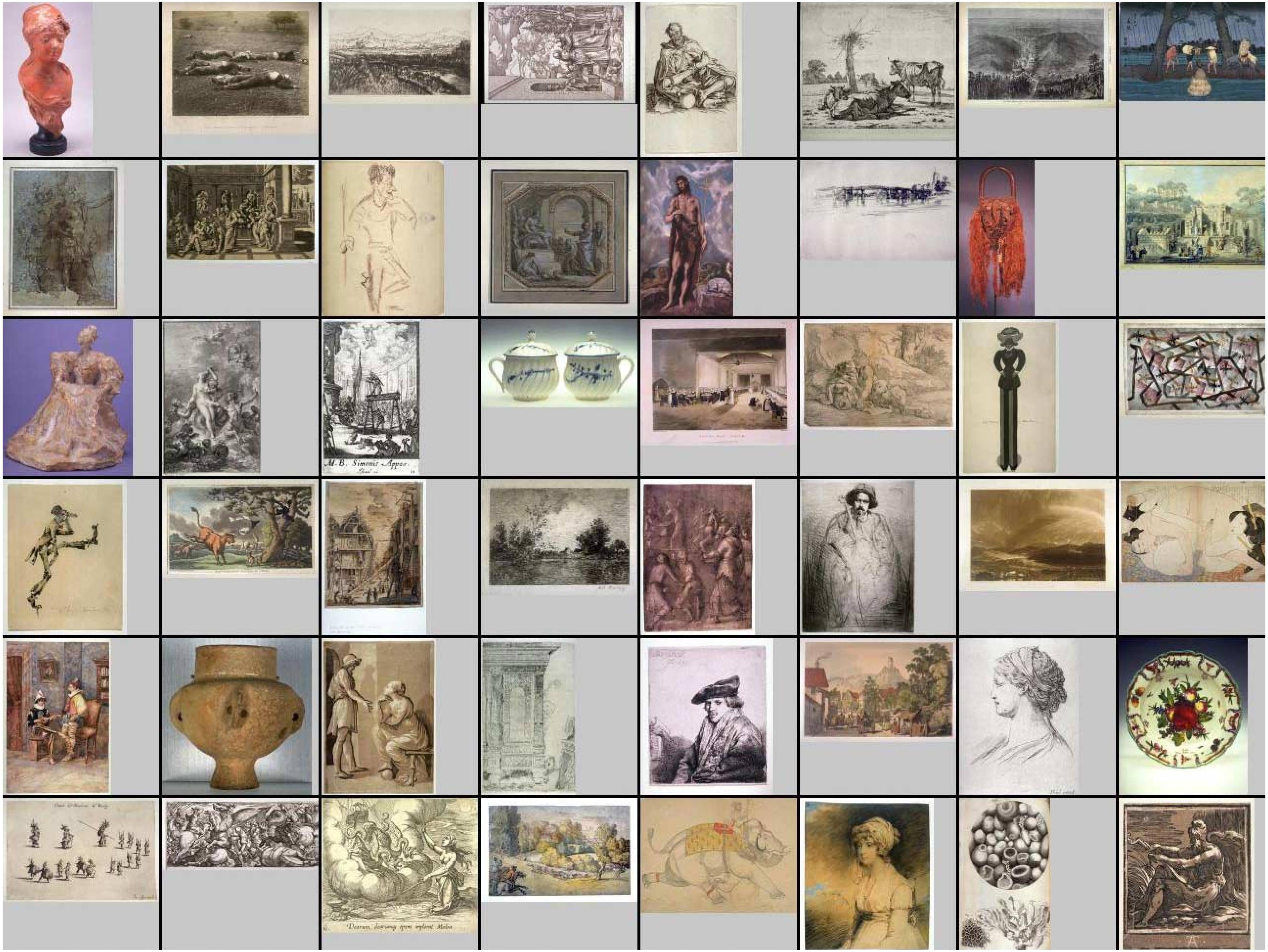
Browsing

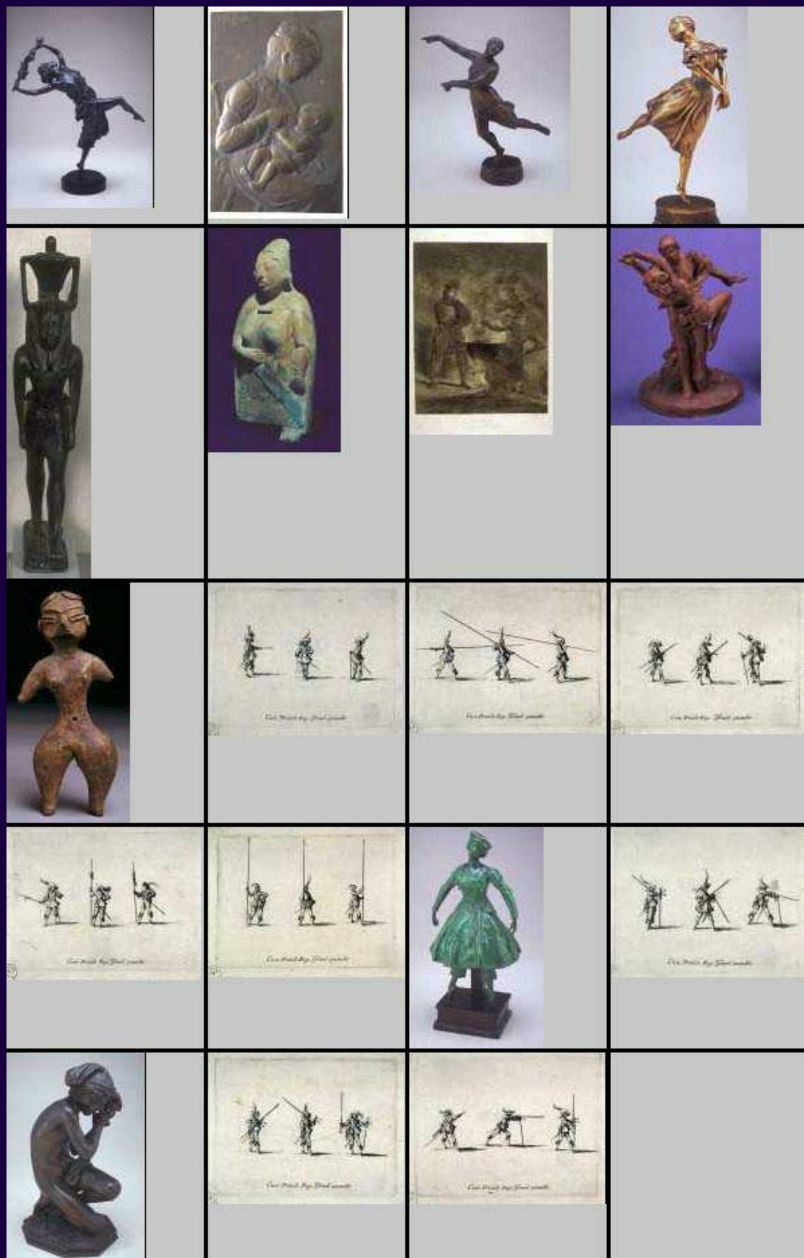
Browsing gives users an overall understanding of what is in a collection--a prerequisite for effective searching.

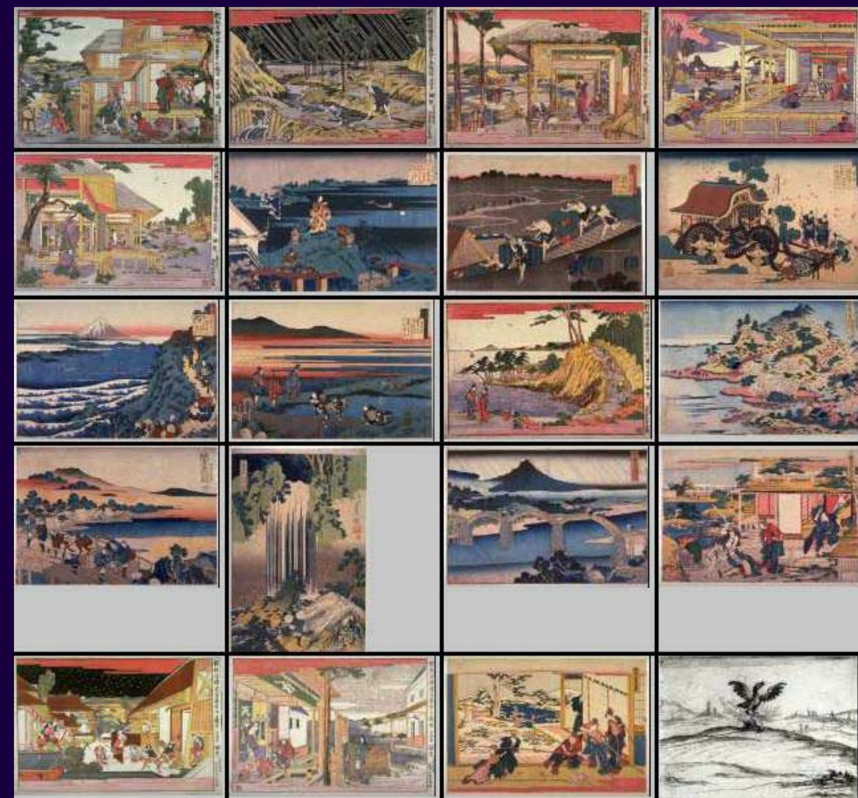
Browsing is not often provided for image databases, partly because it is really hard*.

Need to organize images in a way that is relevant to humans

*Notable exceptions ---Sclaroff, Taycher, and La Cascia, 98; Rubner, Tomasi, and Guibas, 00; Smith Kanade, 97.


M.B. Simonis Appar.





FAMSF Demo

(Based on GIS Viewer from UC Berkeley digital library project)

Searching

Compute P(document | query_items)

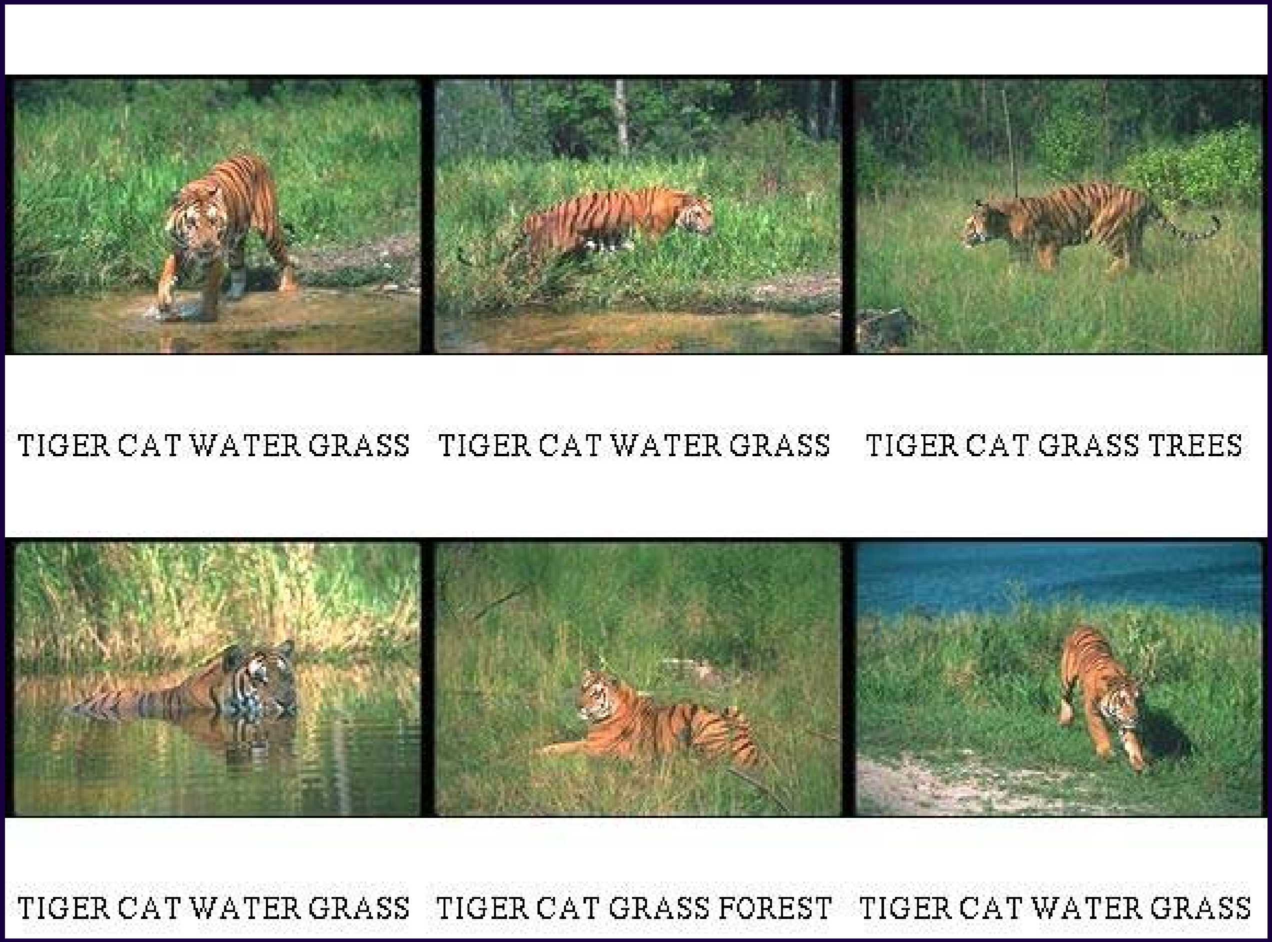
query_items can be words, features, or both

Natural way to express “soft queries”

Related retrieval work: Cascia, Sethi, and Sclaroff, 98; Berger and Lafferty, 98; Papadimitriou et al., 98

Query: "river tiger" from 5,000 Coral images
(The words never occur together.)

Retrieved items: rank order P(document | query)



Query: “water sky cloud  ”

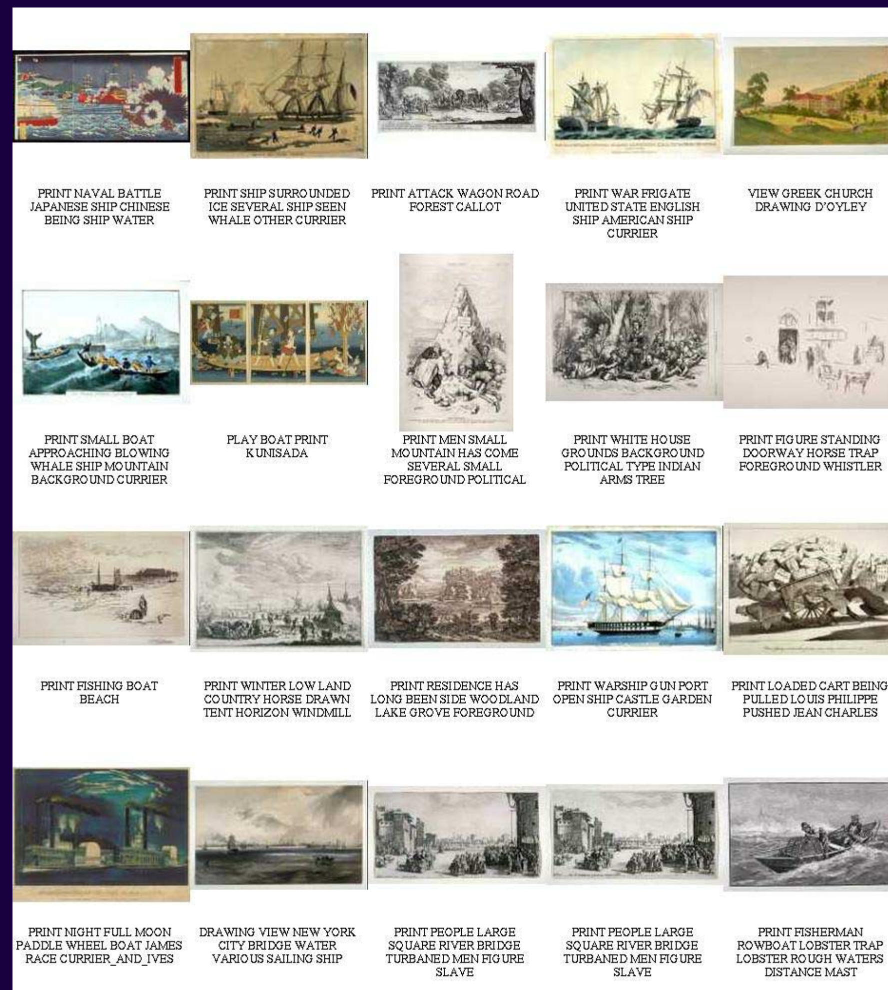
Retrieved items:

Pictures from Words (Auto-illustration)

Text Passage (Moby Dick)

"The large importance attached to the harpooneer's vocation is evinced by the fact, that originally in the old Dutch Fishery, two centuries and more ago, the command of a whale-ship …"

Retrieved Images



Extracted Query

large importance attached fact old dutch century more command whale ship was person was divided officer word means fat cutter time made days was general vessel whale hunting concern british title old dutch ...

PRINT NAVAL BATTLE JAPANESE SHIP CHINESE BEING SHIP WATER

PRINT SHIP SURROUNDED ICE SEVERAL SHIP SEEN WHALE OTHER CURRIER

PRINT ATTACK WAGON ROAD FOREST CALLOT

PRINT WAR FRIGATE UNITED STATE ENGLISH SHIP AMERICAN SHIP CURRIER

PRINT SMALL BOAT APPROACHING BLOWING WHALE SHIP MOUNTAIN BACKGROUND CURRIER

PLAY BOAT PRINT KUNISADA

PRINT MEN SMALL MOUNTAIN HAS COME SEVERAL SMALL FOREGROUND POLITICAL

PRINT WHITE HOUSE GROUNDS BACKGROUND POLITICAL TYPE INDIAN ARMS TREE

The Battle Plan

~~Survey the domain~~

~~Introduce the approach~~

~~Apply to browsing, searching, auto-illustrate~~

Attach words to pictures (auto-annotate)

Compare image segmentation methods

Attach words to image regions (recognition)

The Battle Plan

~~Survey the domain~~

~~Introduce the approach~~

~~Apply to browsing, searching, auto-illustrate~~

Attach words to pictures (auto-annotate)

Compare image segmentation methods

Attach words to image regions (recognition)

Words from Pictures (Auto-annotation)

Compute P(word | regions) on images without captions (or images held out from training)

$$P(w \mid R) \propto P(w,R) = \sum_{c} P(c) \prod_{i \in \{w\} \cup R} \left(\sum_{l} P(i \mid l,c) P(l) \right)$$

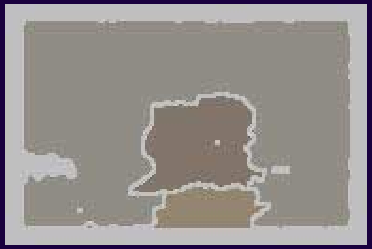
$$Where \quad R = \{regions\}$$

Keywords

GRASS TIGER CAT FOREST

Predicted Words (rank order)

tiger cat grass people water bengal buildings ocean forest reef



Keywords

HIPPO BULL mouth walk

Predicted Words (rank order)

water hippos rhino river grass reflection one-horned head plain sand

Keywords

FLOWER coralberry LEAVES PLANT

Predicted Words (rank order)

fish reef church wall people water landscape coral sand trees

Measuring Performance

Predicted Words → water hippos rhino river grass reflection one-horned head plain

Actual Keywords → HIPPO BULL

Measuring Performance (cont.)

Predicted Words

water hippos rhino
river grass reflection
one-horned head plain

Actual Keywords

HIPPO BULL

Applying Performance Measurement

- Model Selection
- Feature Selection
- Segmentation Comparison

The Battle Plan

~~Survey the domain~~

~~Introduce the approach~~

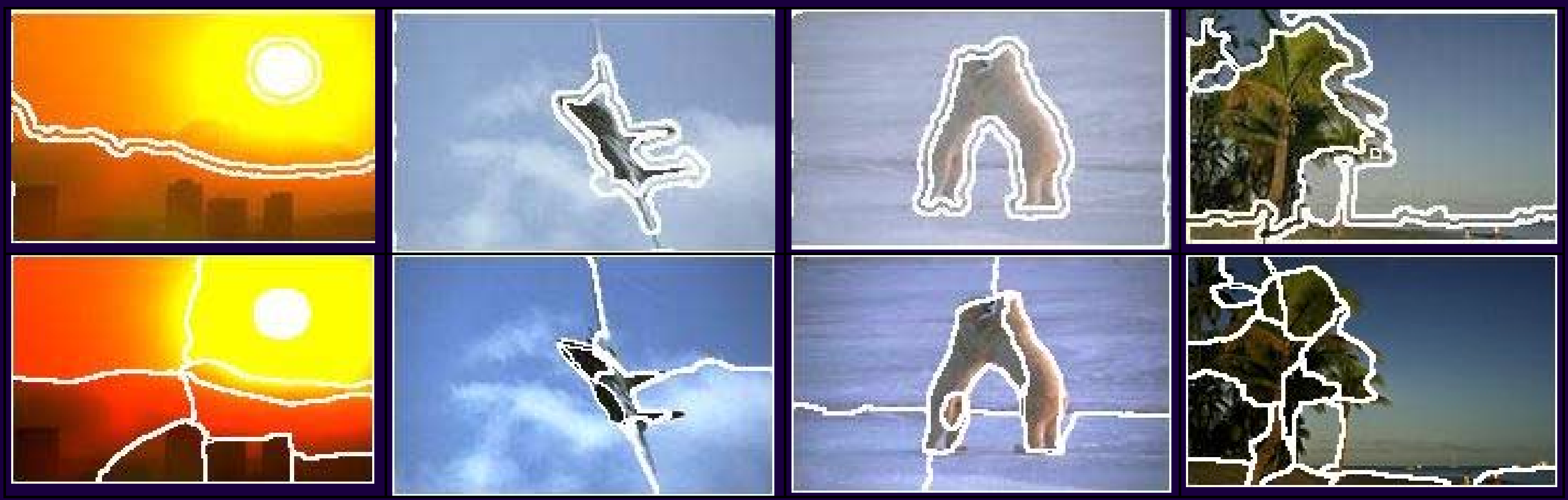
~~Apply to browsing, searching, auto-illustrate~~

~~Attach words to pictures (auto-annotate)~~

Compare image segmentation methods

Attach words to image regions (recognition)

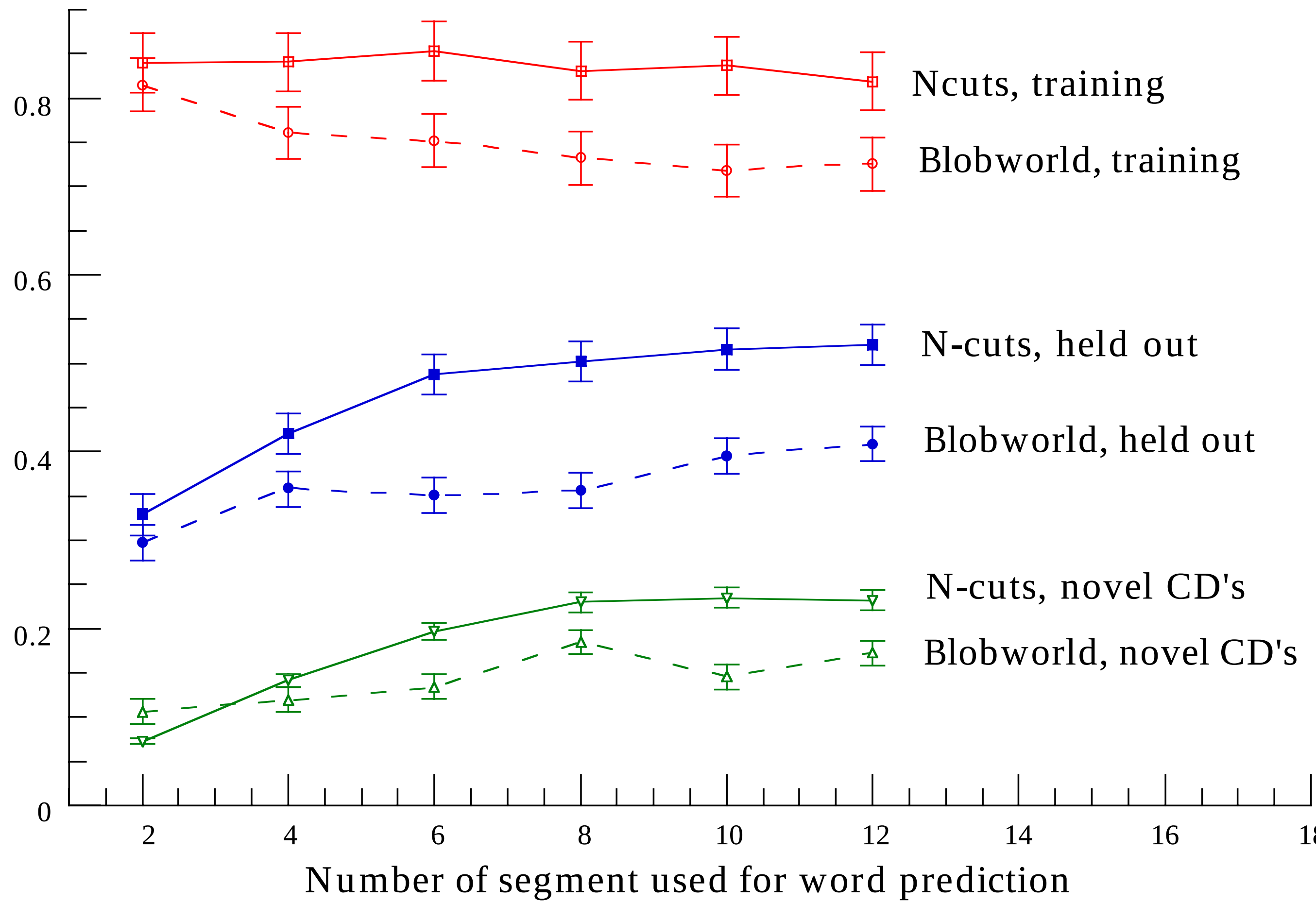
Blobworld segmentations



N-cuts segmentations

A comparison of two segmentation algorithms using word prediction performance

KL divergence based word prediction measure (compared with prior, bigger is better)



The Battle Plan

~~Survey the domain~~

~~Introduce the approach~~

~~Apply to browsing, searching, auto-illustrate~~

~~Attach words to pictures (auto-annotate)~~

~~Compare image segmentation methods~~

Attach words to image regions (recognition)

Annotation vs Recognition

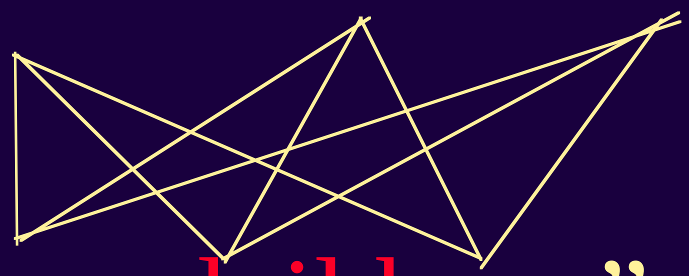
?

tiger cat grass

Statistical Machine Translation

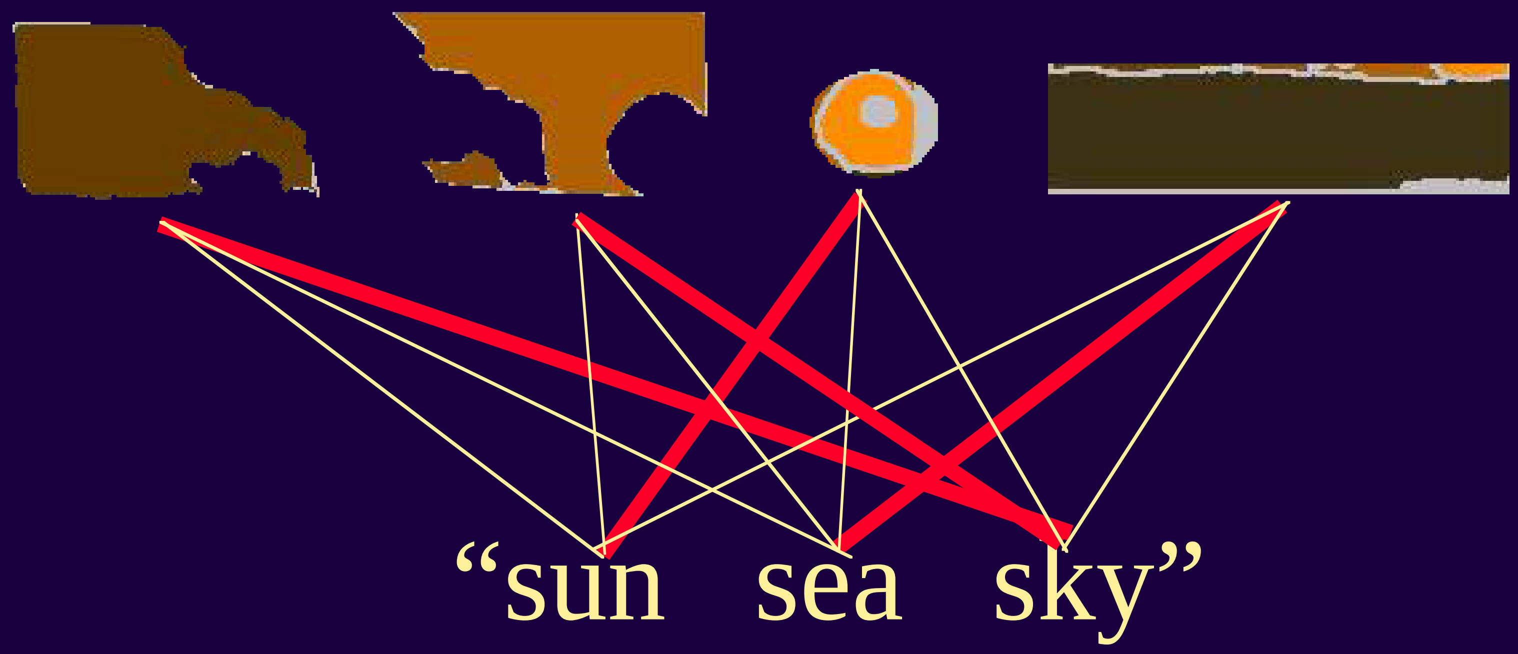
Data: Aligned sentences, but word correspondences are unknown

"the beautiful sun"



"le soleil beau"

Multimedia Translation



Statistical Machine Translation

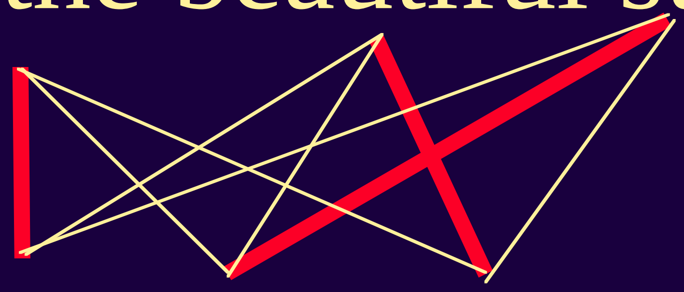
Given the correspondences, we can estimate the translation p(sun|soleil)

Given the probabilities, we can estimate the correspondences

Statistical Machine Translation

Enough data + EM, we can obtain the translation p(sun|soleil)=1

“the beautiful sun”



“le soleil beau”

Hierarchical Clustering with Correspondence

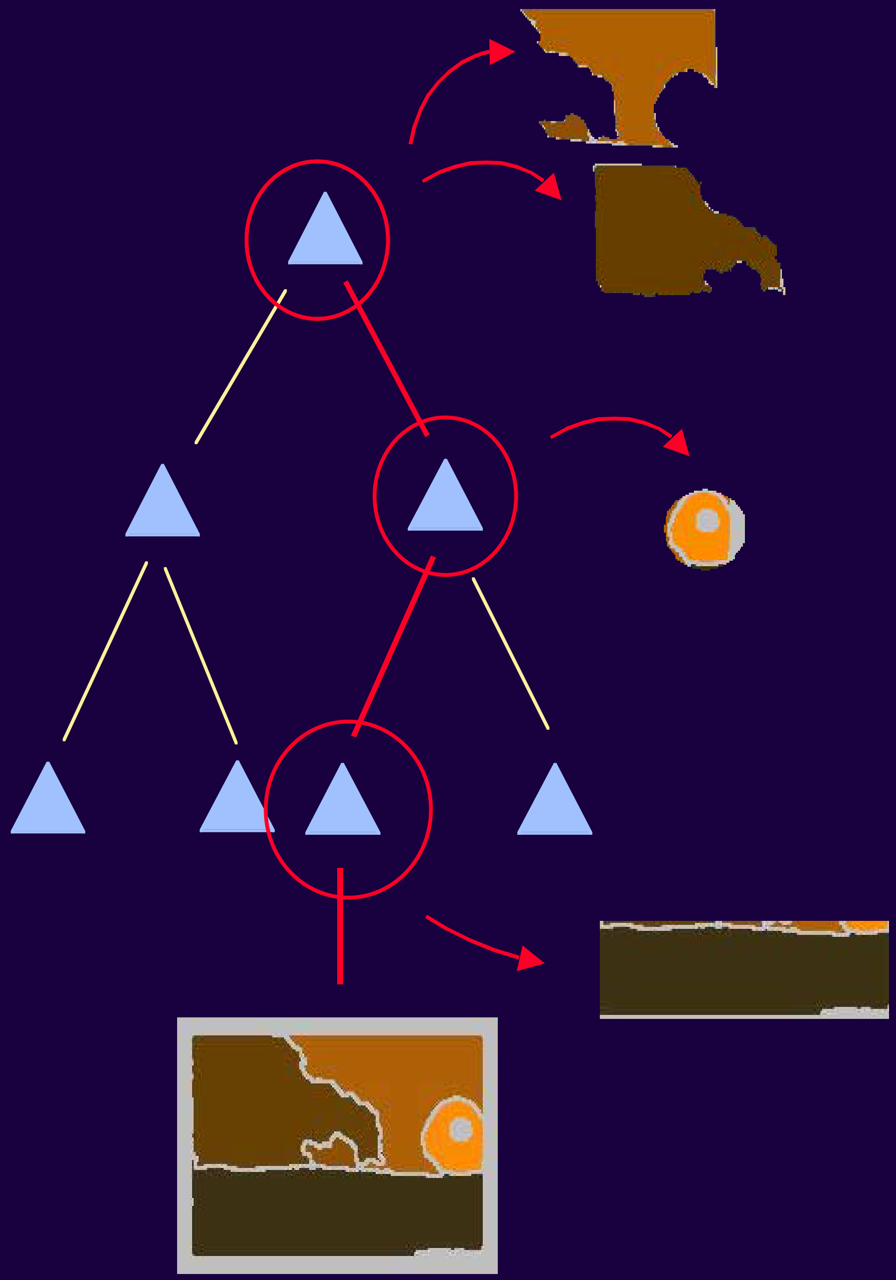
Can force original model to give correspondence (works OK) but better to incorporate it.

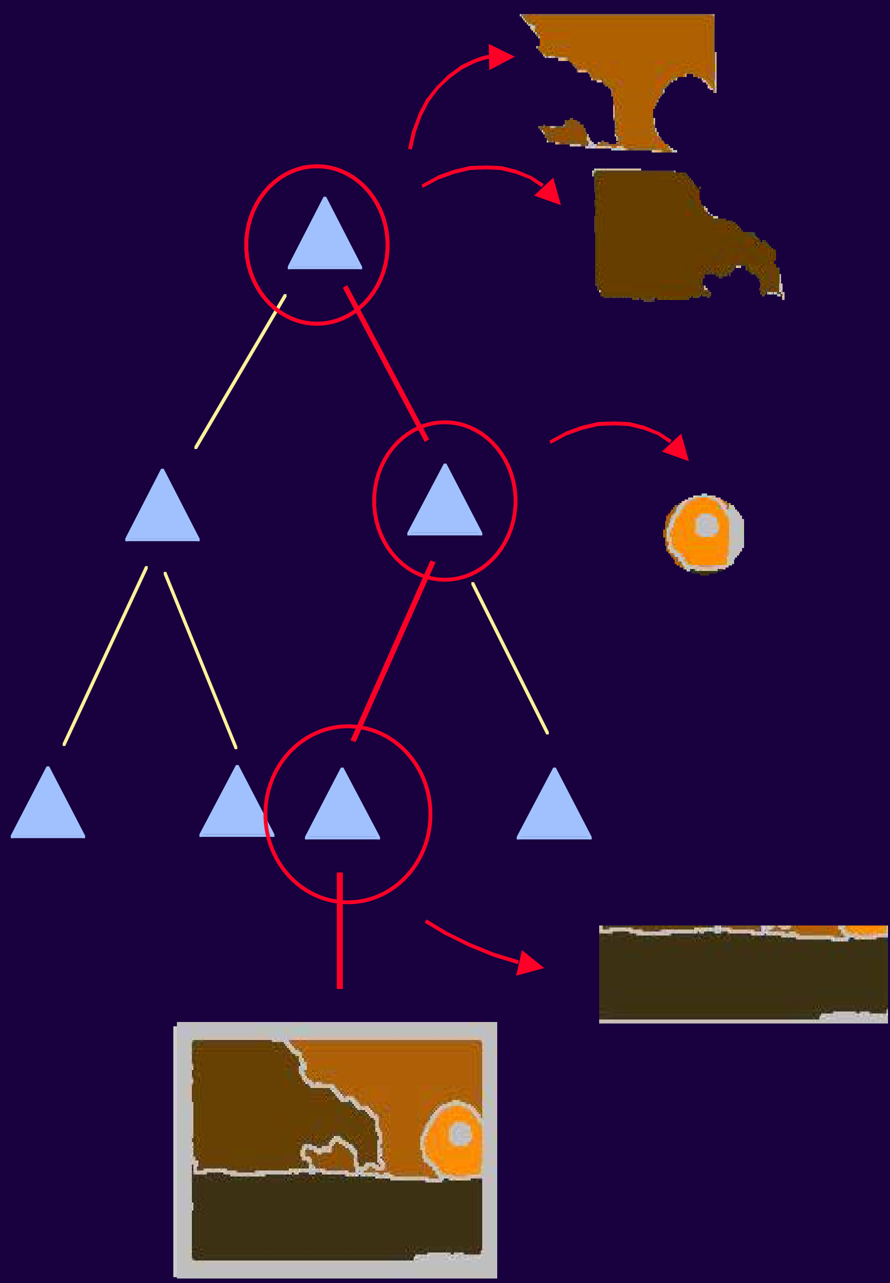
Change the assumption of conditional independence (words should be emitted conditioned on the regions).

Hierarchical Clustering with Correspondence

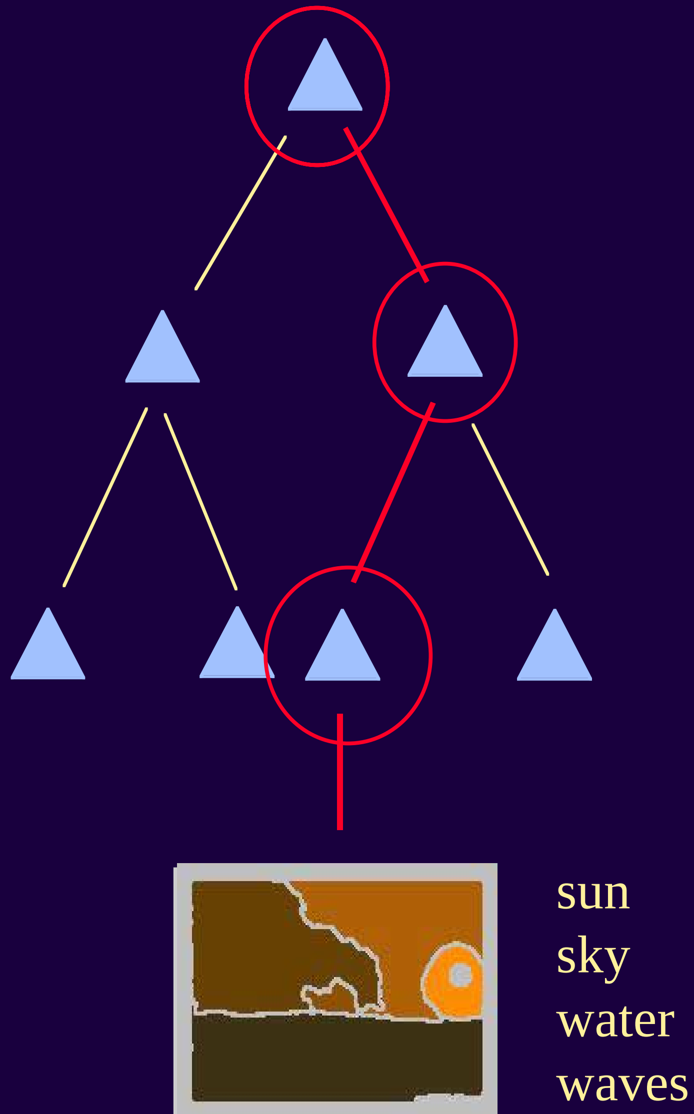
Method One:

Model regions as before, but compute P(word | regions, cluster)





Generate words from the distribution for blobs



Generate
words
from the
distribution
for blobs

Hierarchical Clustering with Correspondence

Method Two:

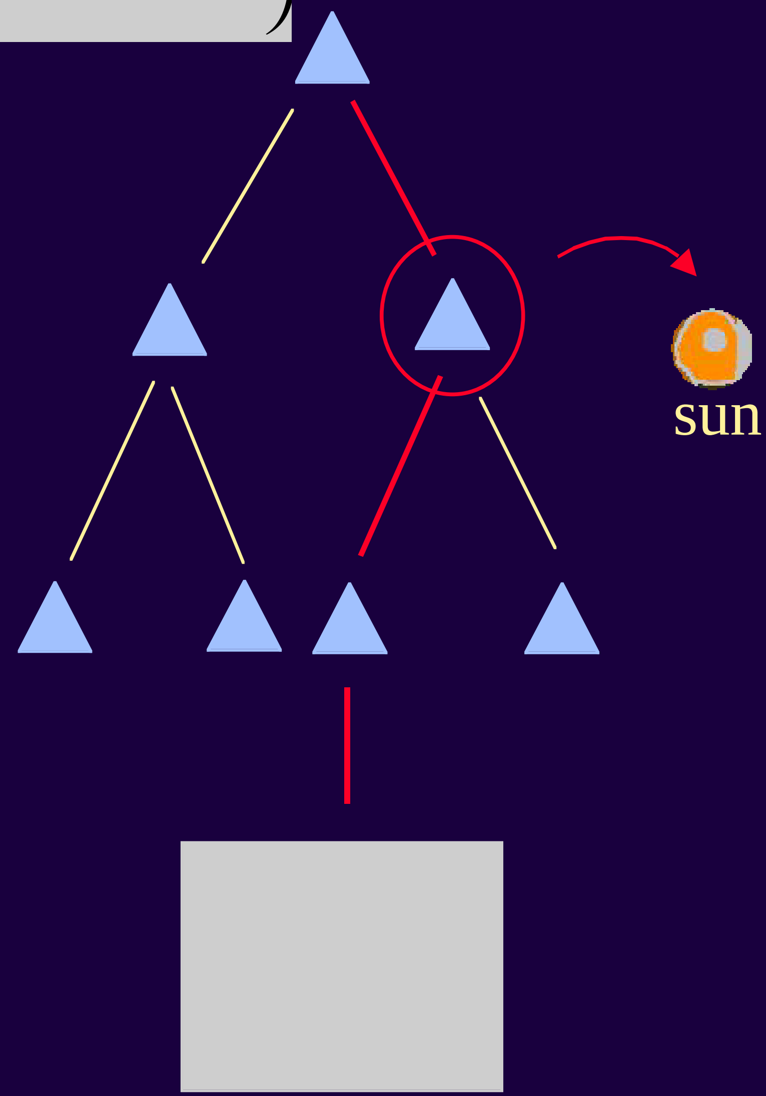
Words and regions are now generated as pairs from the same node (estimate correspondence in training with graph matching--algorithm and source code from Jonker and Volgenant).

$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 2

pair
p =1

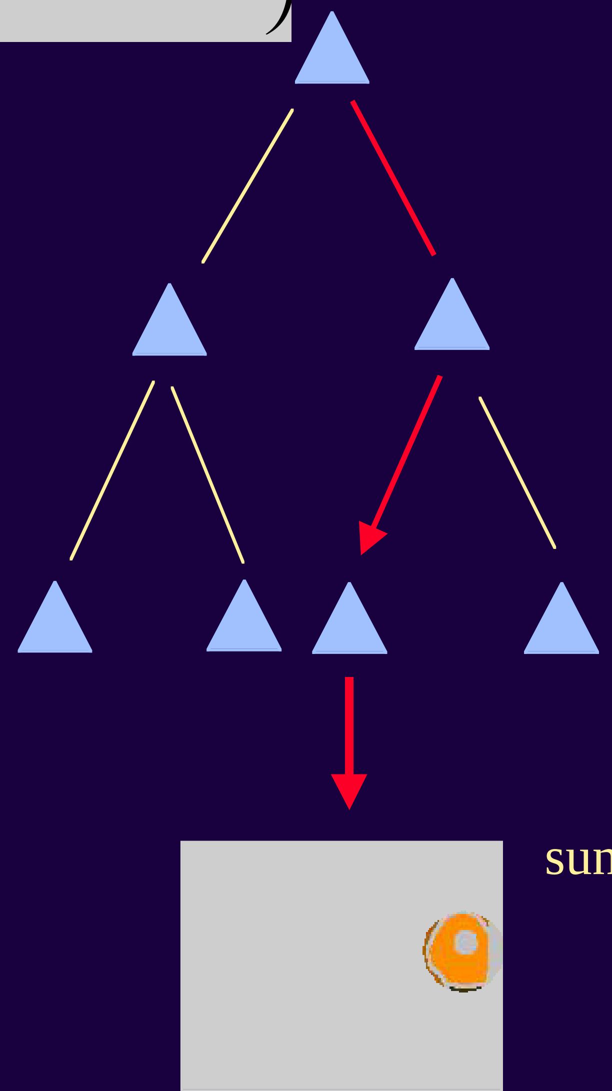


$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 2

pair
p =1

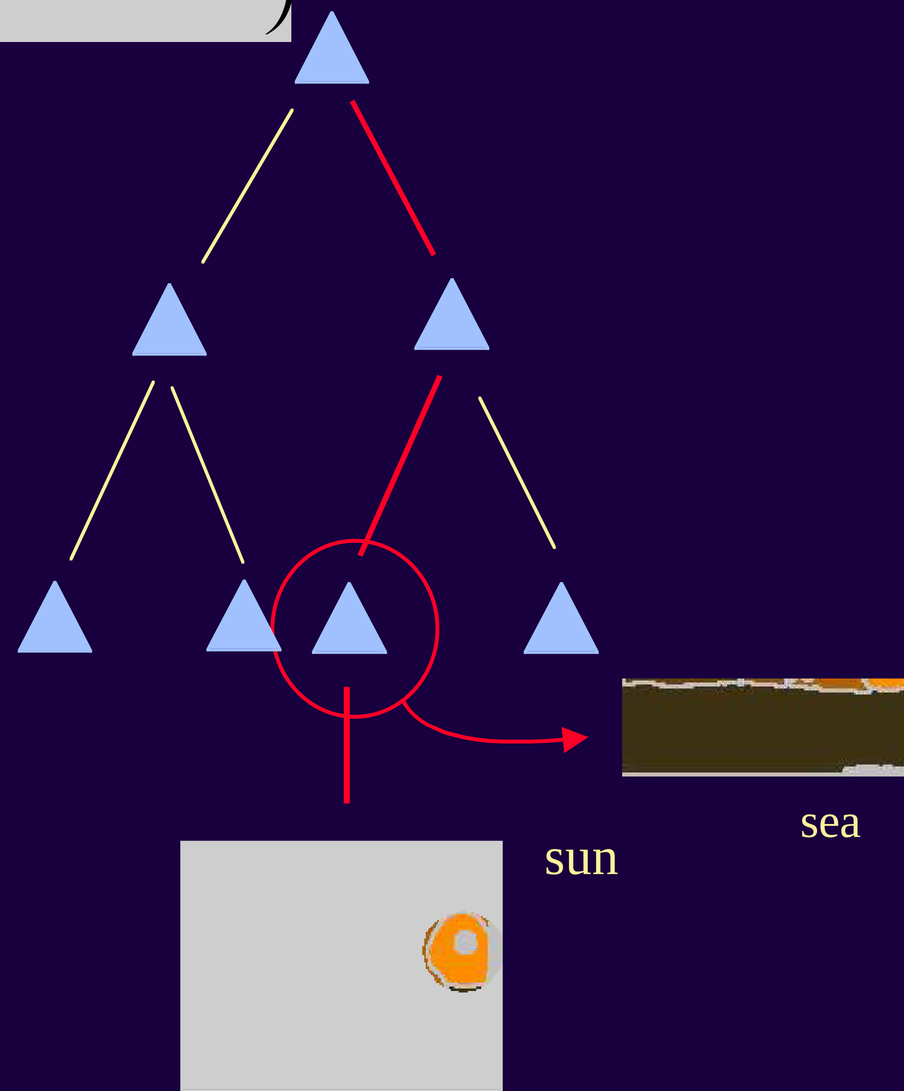


$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 3

pair
p =2

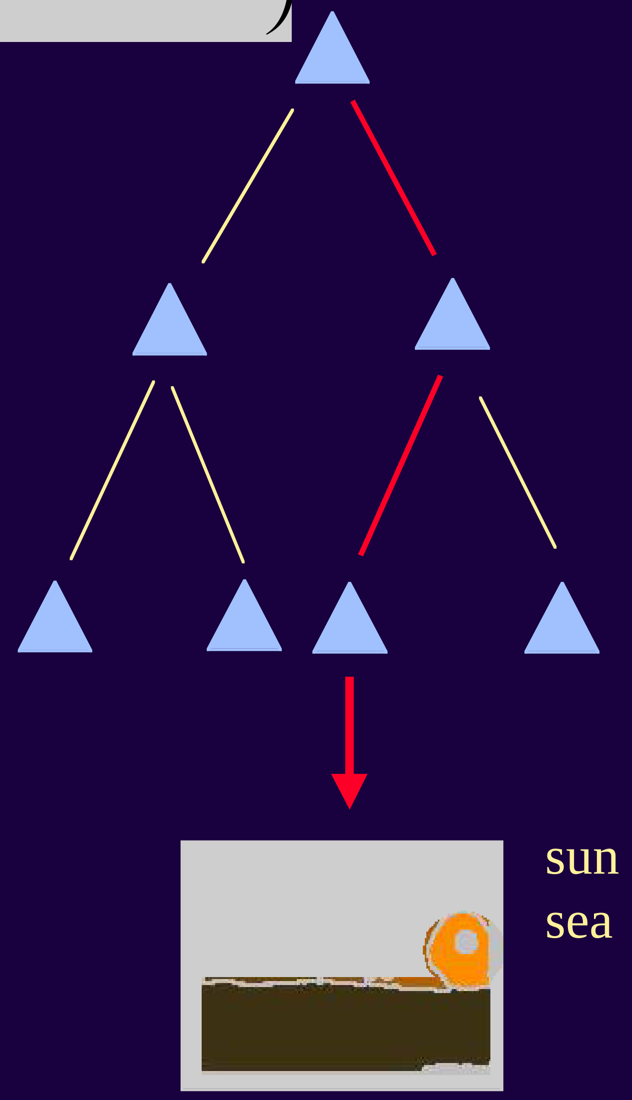


$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 3

pair
p =2

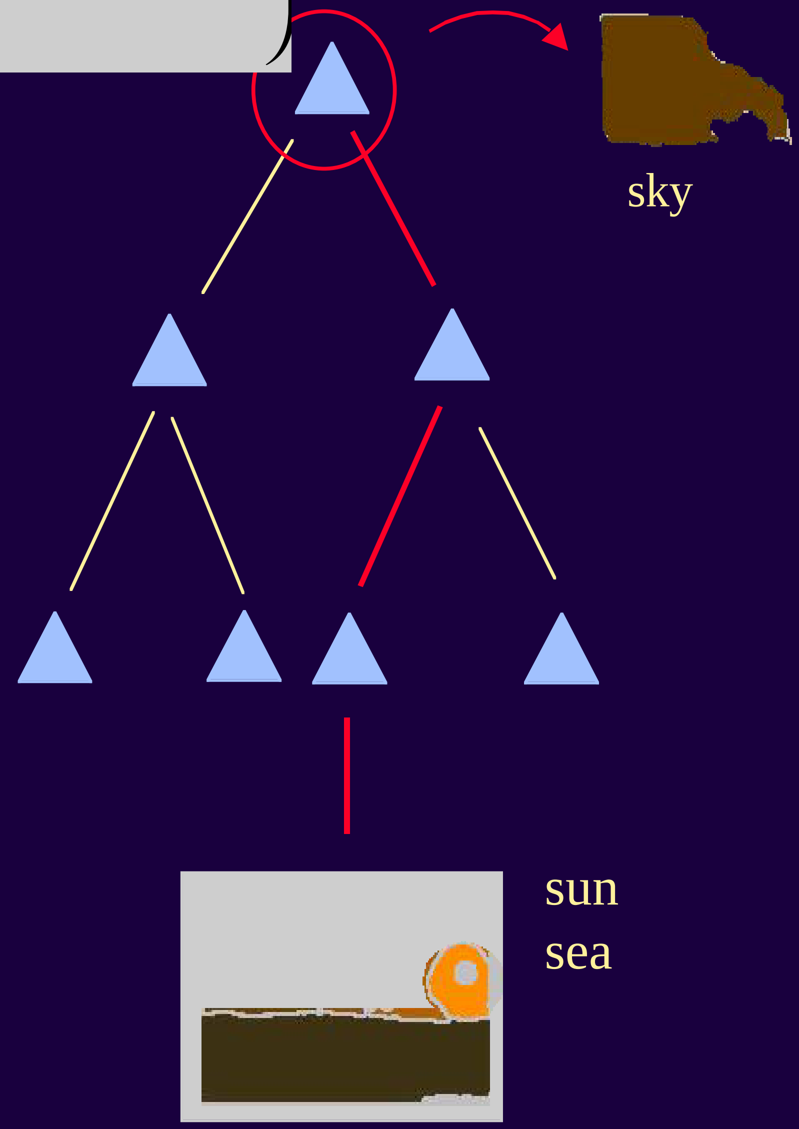


$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

pair
p =3

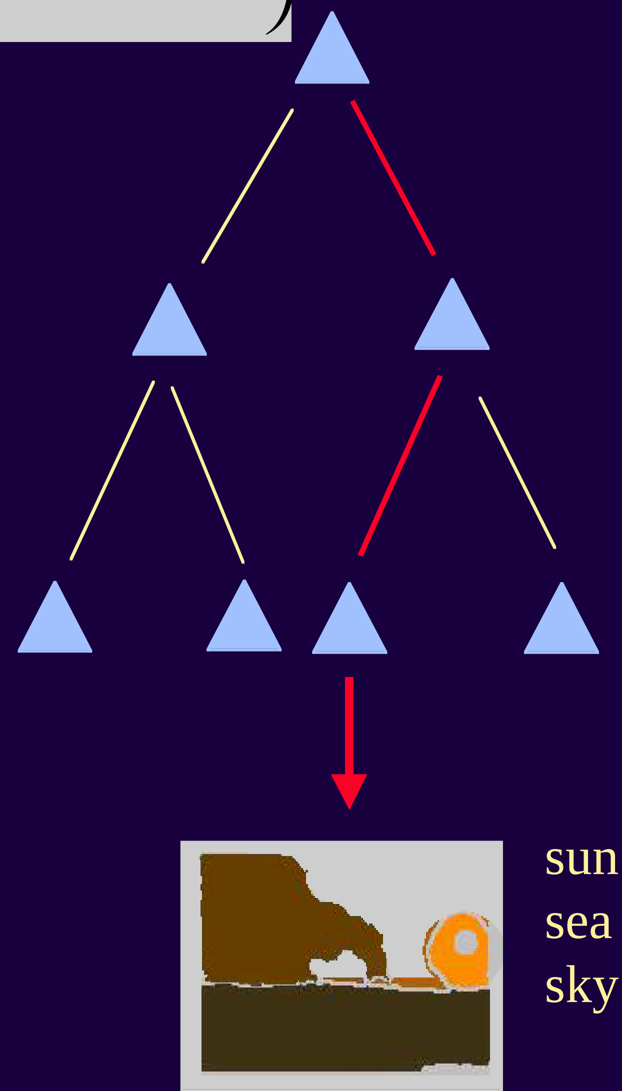


$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

pair
p =3

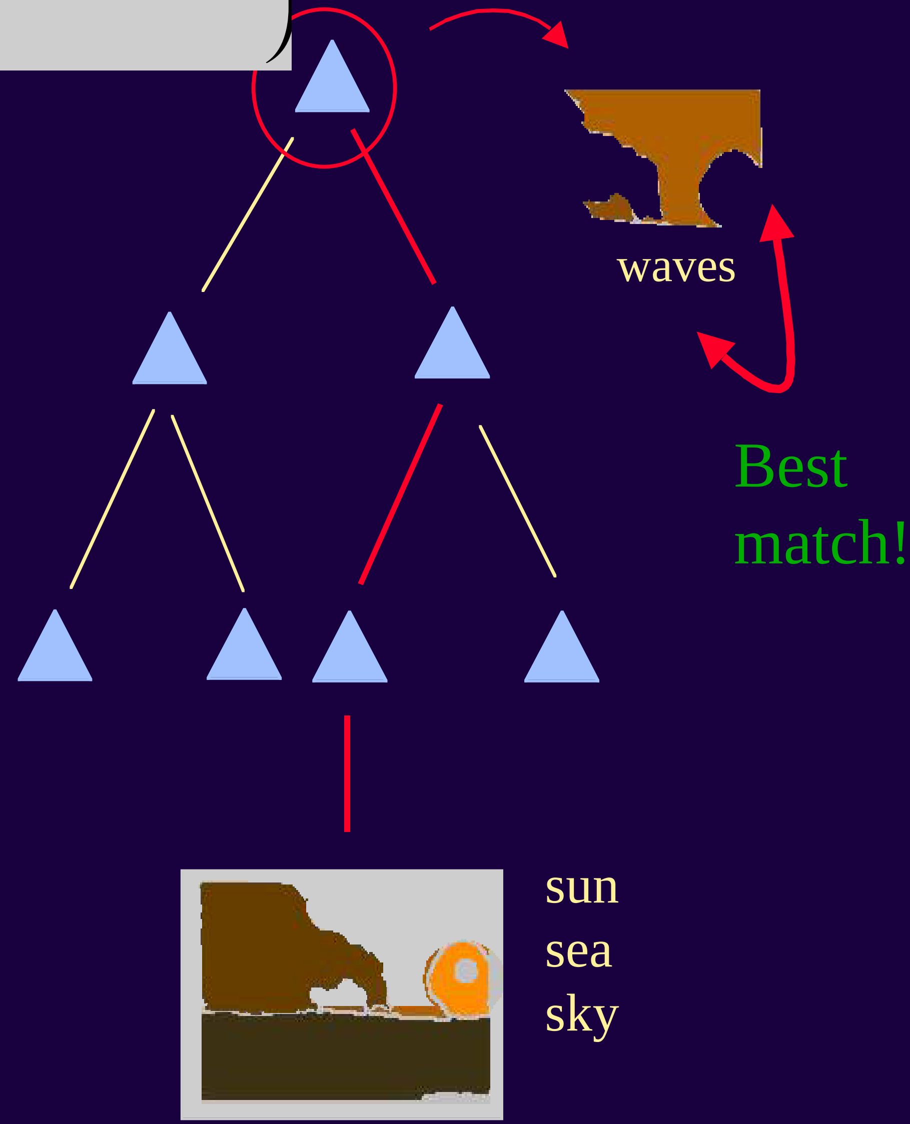


$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

pair
p =4

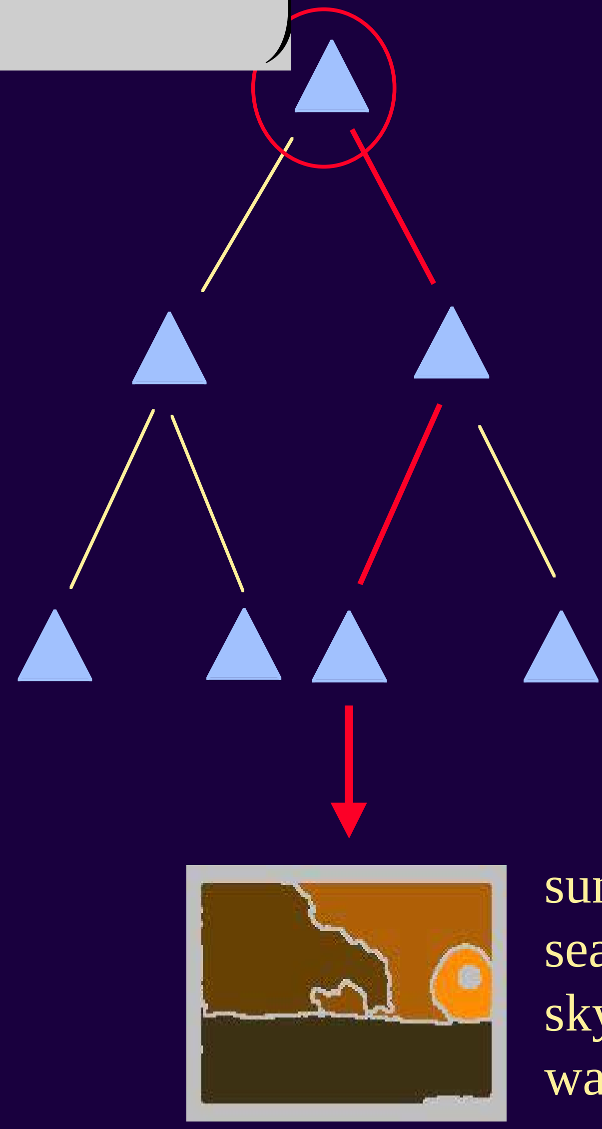


$$P(D \mid d) = \sum_{c} P(c) \prod_{p \in D} \left(\sum_{l} P(p \mid l, c) P(l \mid d) \right)$$

cluster
c = 3

level
l = 1

pair
p =4



sun
sea
sky
waves

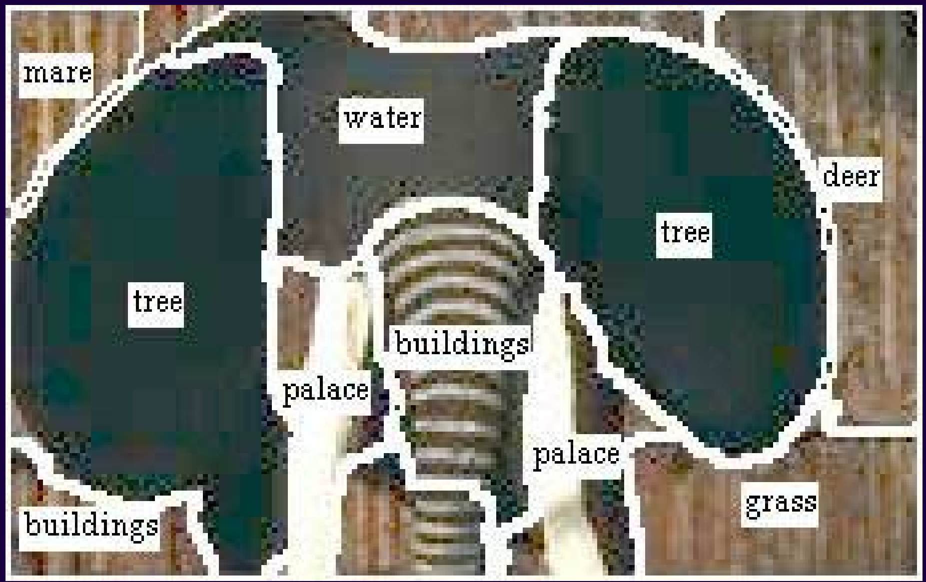
Recognition Approach

Learn to label without labels

Learn what to recognize

(Current vocabulary size--several hundred)


grass
cat
horses
buildings
tiger
grass
mare
mare
aterfalls
sky
row
building
buildings
buildings
house
people
tree
tree
ground
flowers
ground
antlers
shops
town
crab
tiger
town
ground


mare
water
deer
tree
tree
buildings
palace
palace
grass
buildings

Measuring Recognition Performance

First strategy--use annotation performance as a proxy.

Second strategy--score by hand.

Scoring rules for comparing models efficiently

Look only at maximal probable word

Ignore confidence (force prediction of something)

Recognition performance

Average performance is four times better than guessing the most common word (“water”)

Bottom Line

Recognition as machine translation

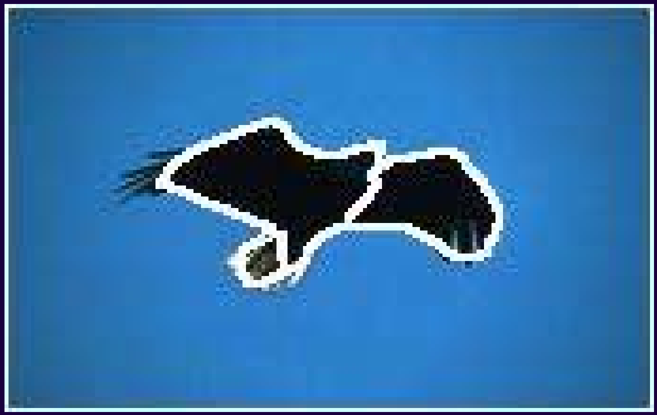
Machine vision as data-mining

Future Directions
(computer vision)

Propose region merging based on posterior word probabilities

Future Directions
(computer vision)

Propose good features to differentiate words that are not distinguishable (e.g., eagle and jet)



Future Directions
(machine learning)

Estimate where a minimal amount of supervision can be most helpful (and provide it)