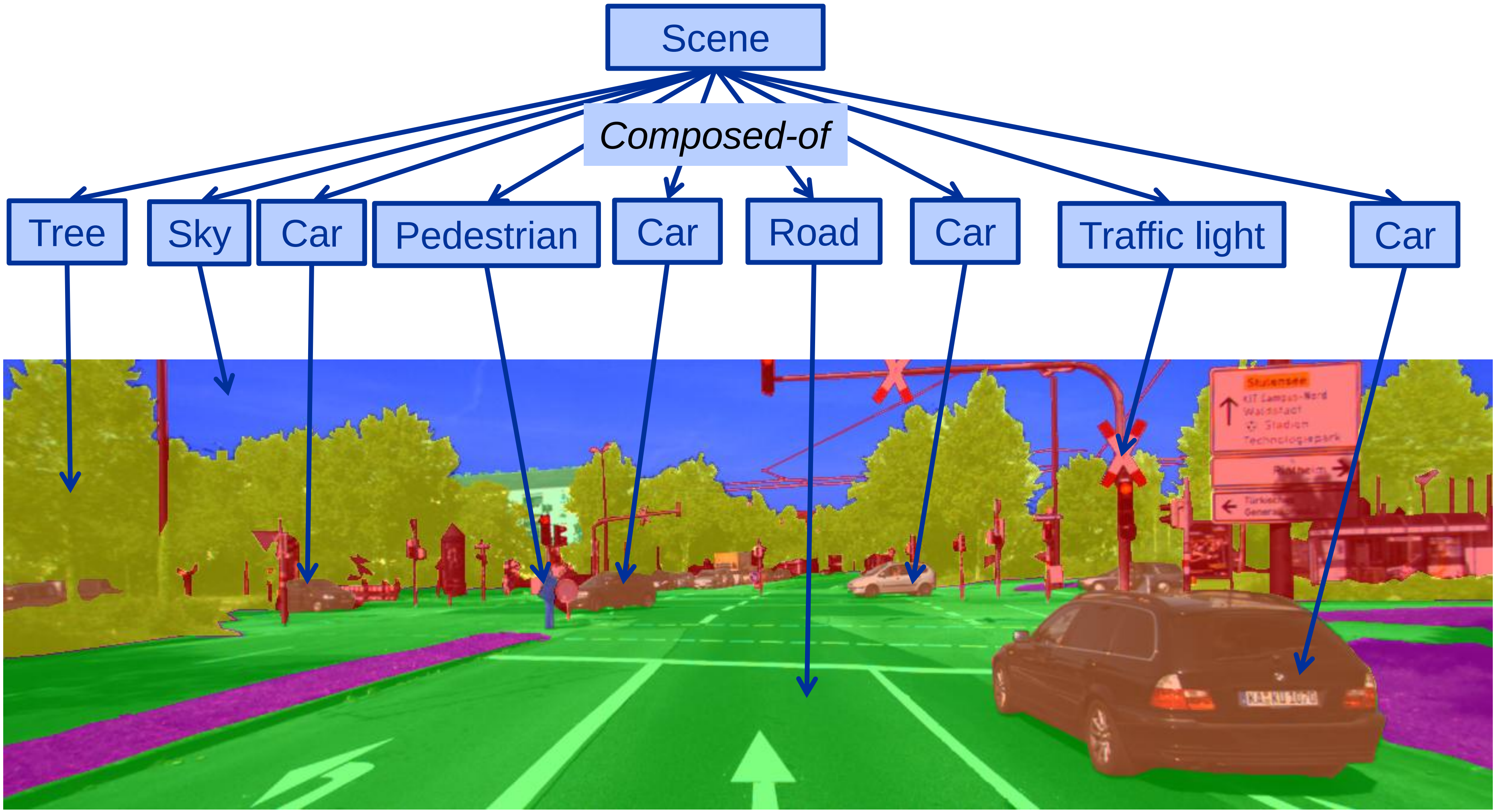
A Framework for Multimodal Urban Scene Understanding

Philippe XU
Jean-Baptiste BORDES
Franck DAVOINE
Thierry DENOEUX
Huijing ZHAO

Multi-sensors context

Expected Result

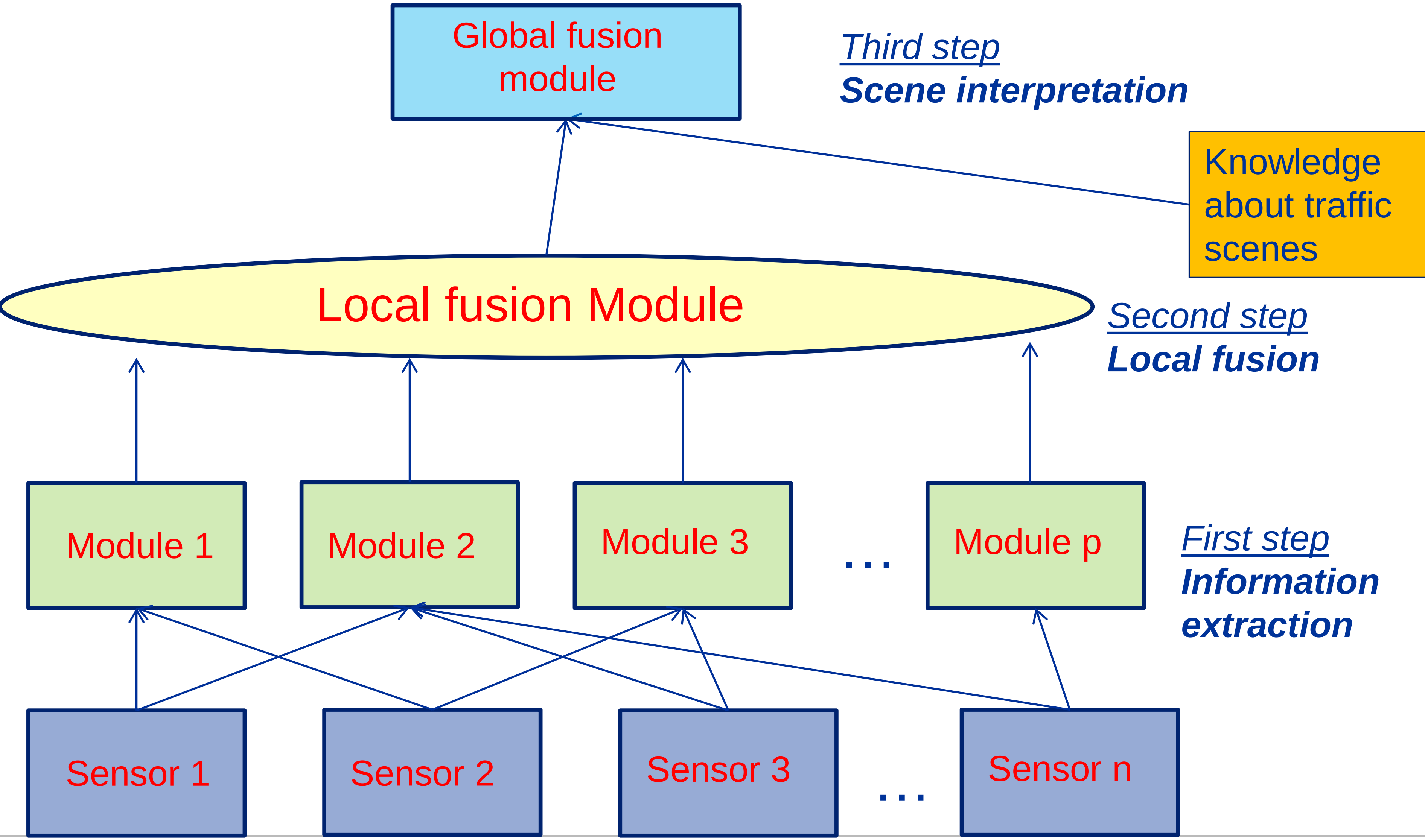


Requirements of the targeted framework

- **Flexibility:**
  - **Cope with sensor failure**
  - **Add new classes easily**

- **Multimodal:**
  - **Fuse the output incoming from various kinds of sources of information**
  - **Take into account easily new sources of information**

Proposed global Framework



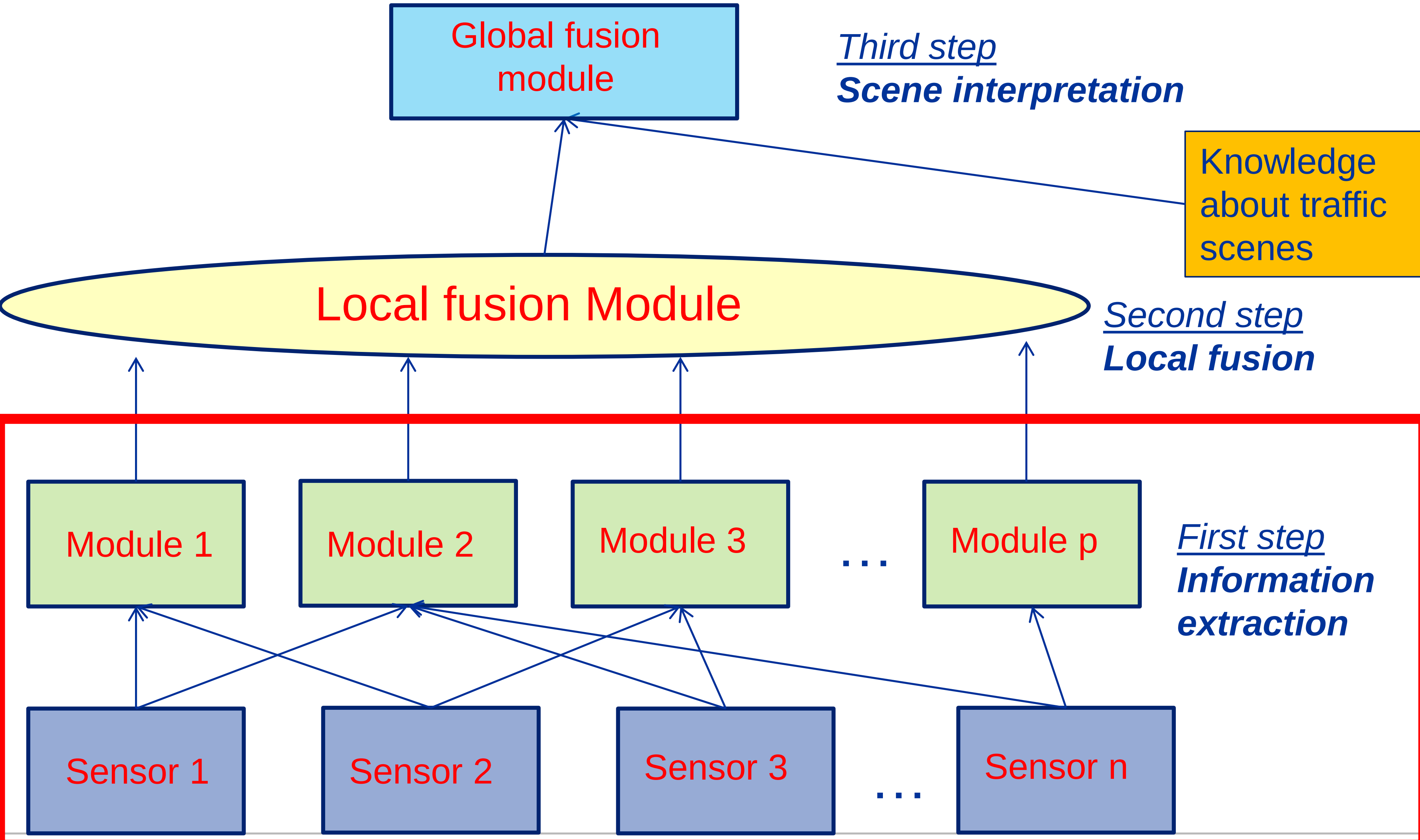
Outlines

1. **Information extraction**

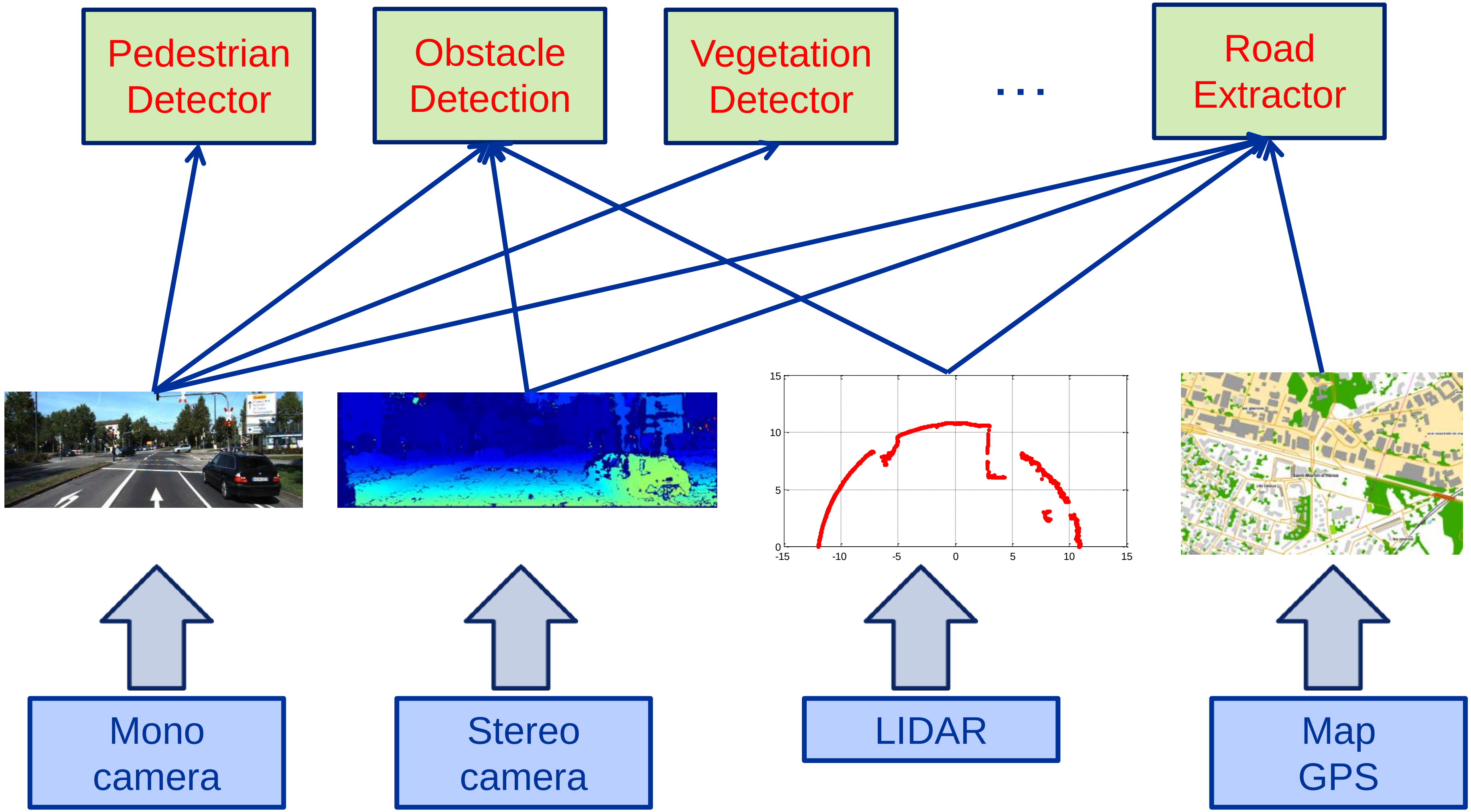
2. **Local Fusion using Dempster-Shafer theory**

3. **Global fusion using Evidential Grammar**

Proposed global Framework

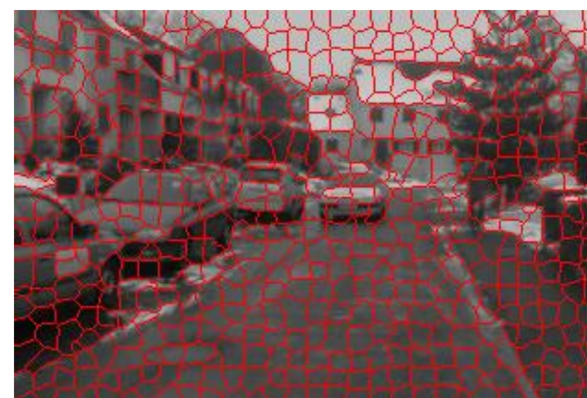


Global Framework



Decision space

- **Reasoning at the image level:**
  - **It is what the driver sees**
  - **Adapted for driver assistance systems (may not be the case for autonomous driving)**
- **Classification over an over-segmented image**
  - **Intermediate level between pixel level/object level**



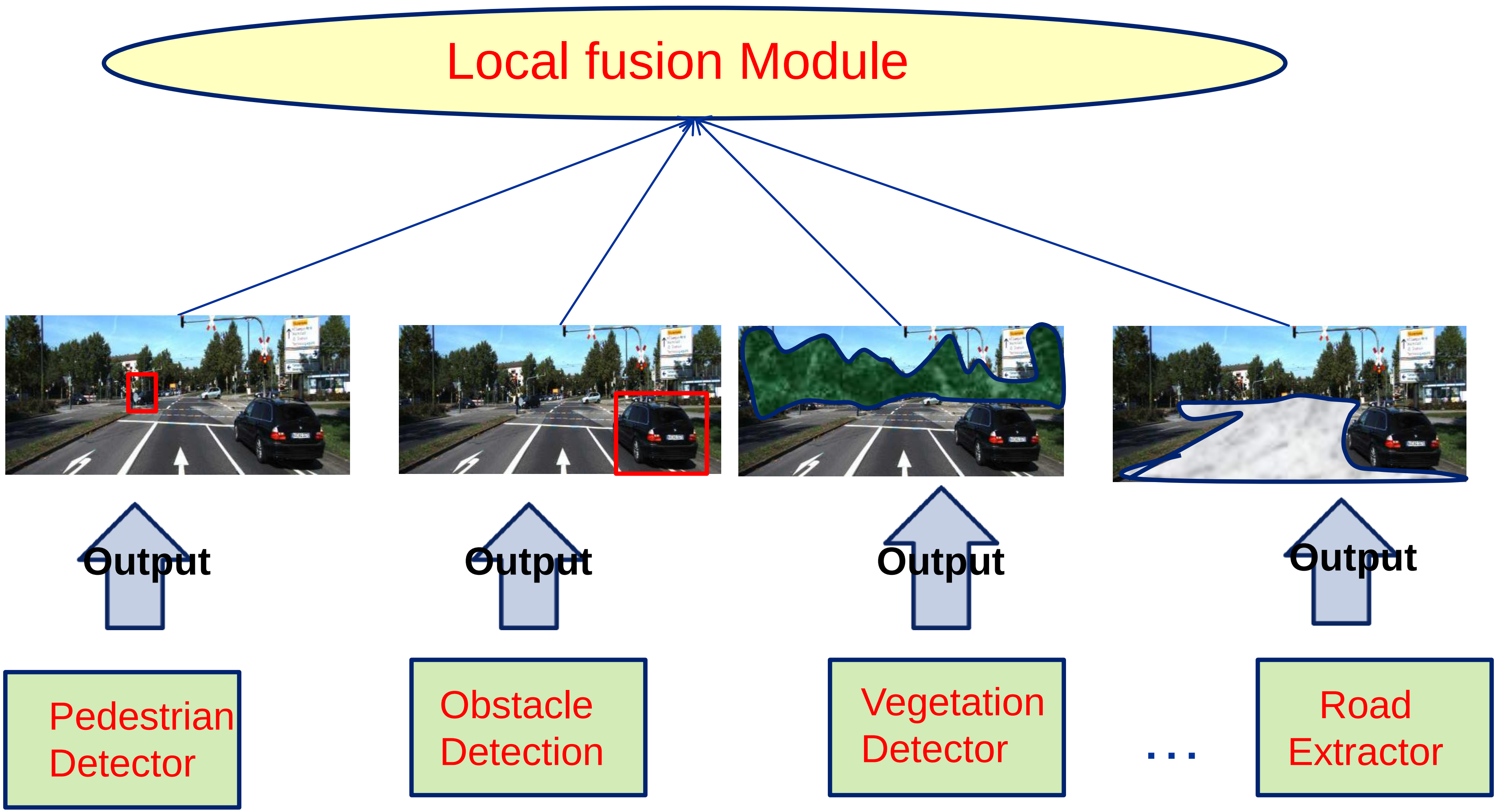
Outlines

1. Information extraction

2. Local Fusion using Dempster-Shafer theory

3. Global fusion using Evidential Grammar

Local fusion step



Class space

Several classes could be considered:

- Pedestrians, cyclists
- Cars, motorbikes, trucks
- Roads, buildings, trees
- Traffic signs
- …

Main issues:

- Which classes to choose?
- How can we make it flexible to add new classes?
- How to make common decision space to all classifiers?

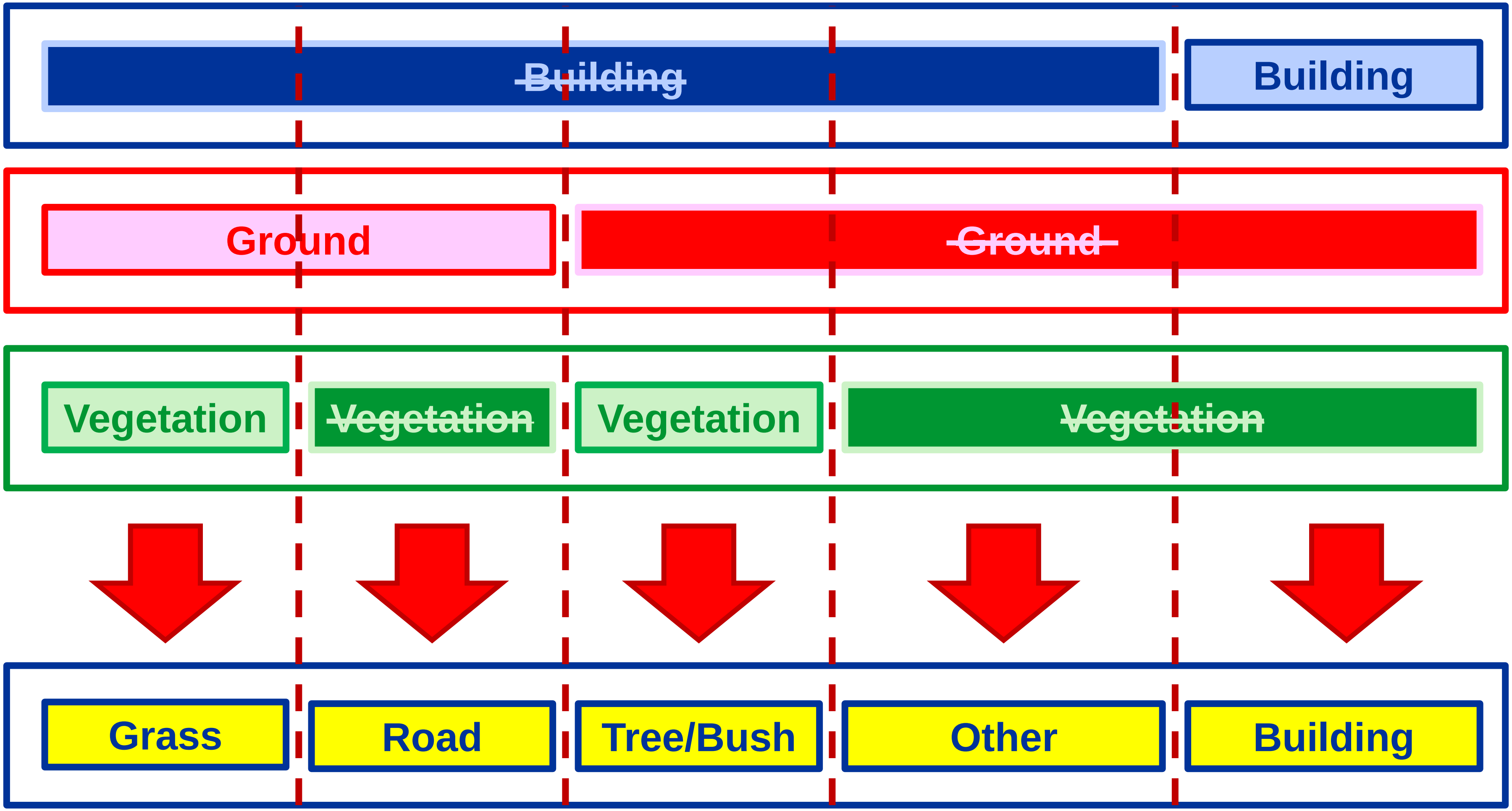
Belief functions work on sets

~~Building~~	Building

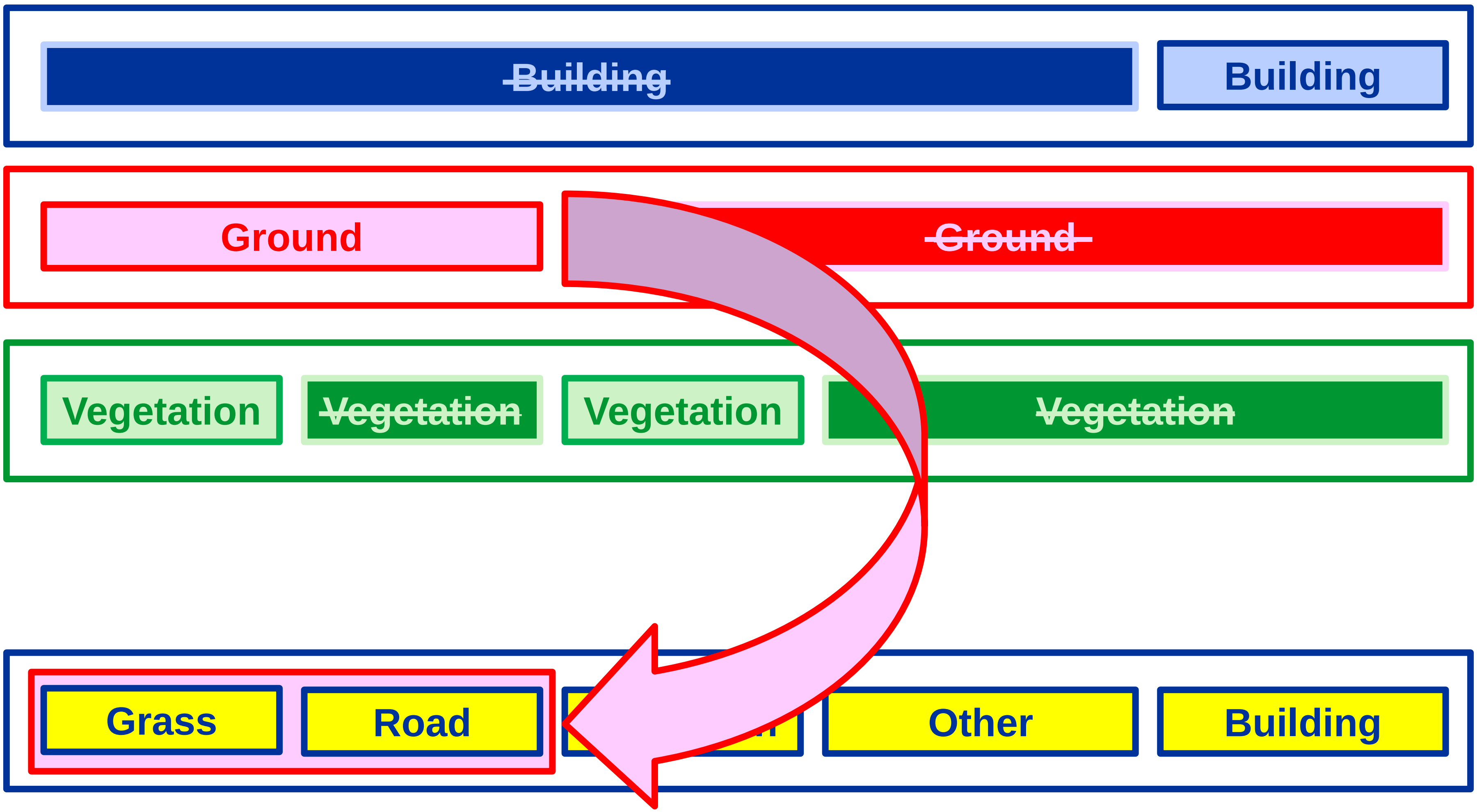
Ground	~~Ground~~

Vegetation	~~Vegetation~~	Vegetation	~~Vegetation~~

Belief functions work on sets



Belief functions work on sets



Ground detection

Disparity from stereo camera	Stereo based ground detector
Laser points (Velodyne)	Laser based free space detector
Optical flow from consecutive images	Optical flow based temporal propagation
Oversegmentation using Turbopixels	Demspter-Shafer fusion

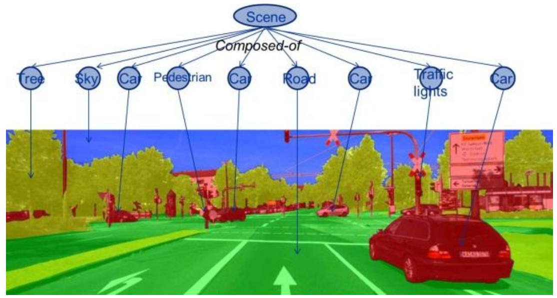
Outlines

1. **Information extraction**

2. **Local Fusion using Dempster-Shafer theory**

3. **Global fusion using Evidential Grammar**

Global fusion

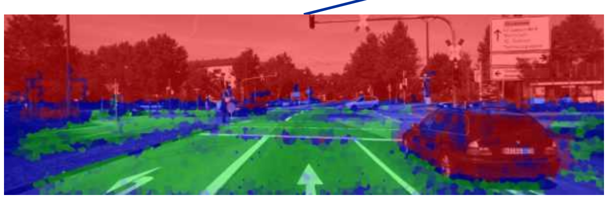


Output

Global fusion module

Third step

Scene interpretation



Output

Knowledge about traffic scenes

Local fusion Module

Evidential grammars

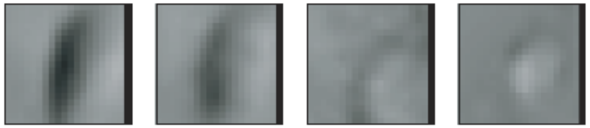
- **Input : Classification at the segment level**
  - **Objects or parts of objects**
  - **Contain uncertainty**
- **Output: Image understanding**
  - **Identification of the objects**
  - **Relationships between the objects**
- **Visual Grammars**
  - **Model of the decomposition of a scene into objects, parts of objects and primitives**

Stochastic grammars

- **A stochastic grammar is a 5 tupple (S,V_N,V_T, Γ, P) where:**
  - **S is a starting symbol**
  - **V_N is a set of non-terminal nodes**
  - **V_T is a set of terminal nodes**
  - **Γ is a set of production rules augmented with a set of probabilities P:**
    - **A -> A_1 with probability p_1**
    - **A -> A_2 with probability p_2**
    - **…**
    - **A -> A_n with probability p_n**

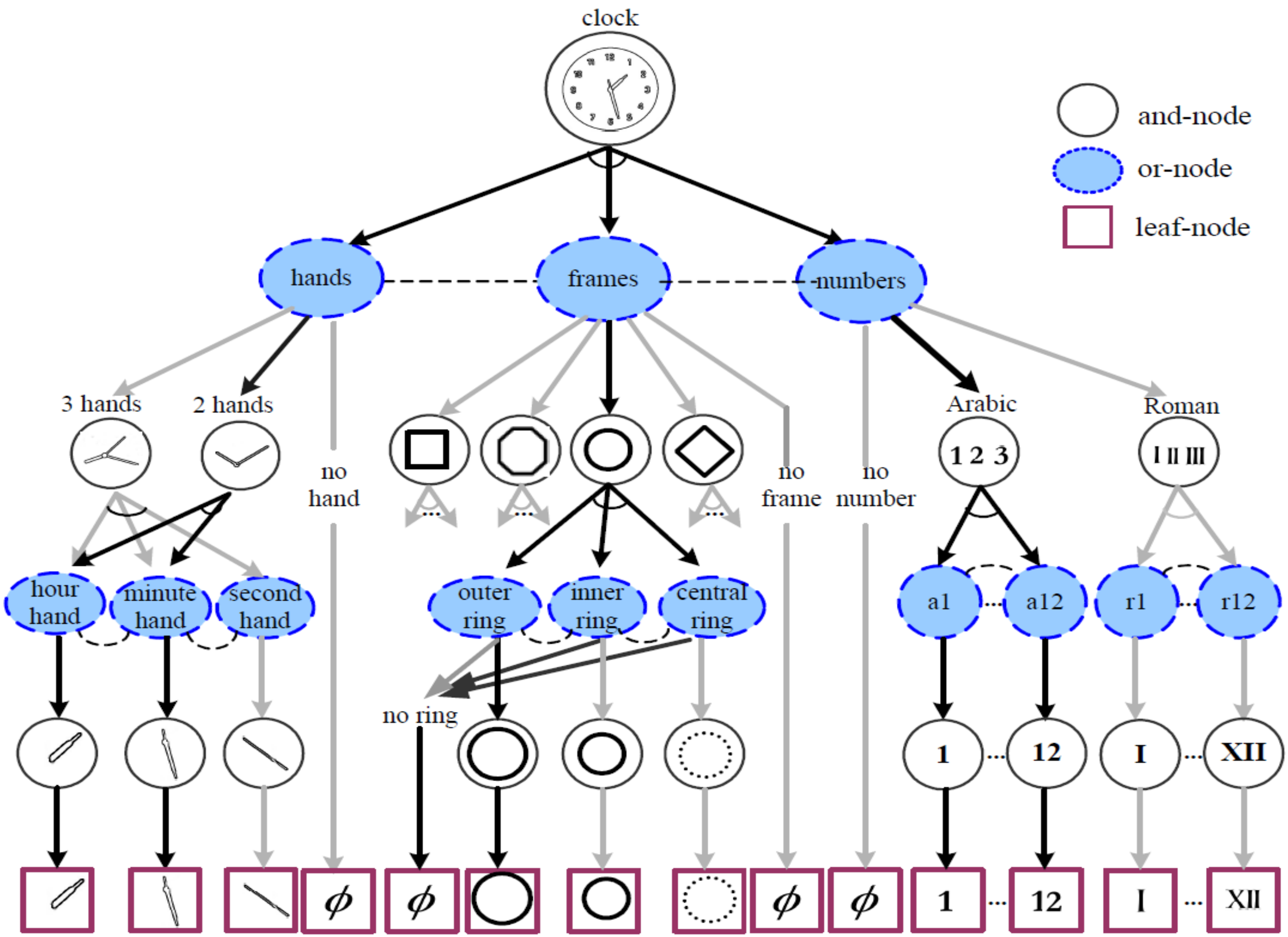
Visual Grammars

- **Extension of the notion of stochastic grammars for the image**
  - **The natural left-to-right ordering of words is replaced by spatial relationships:**
    - **Ex: Pedestrian -> Head “over” Body**
  - **The terminal nodes are called “visual primitives”**



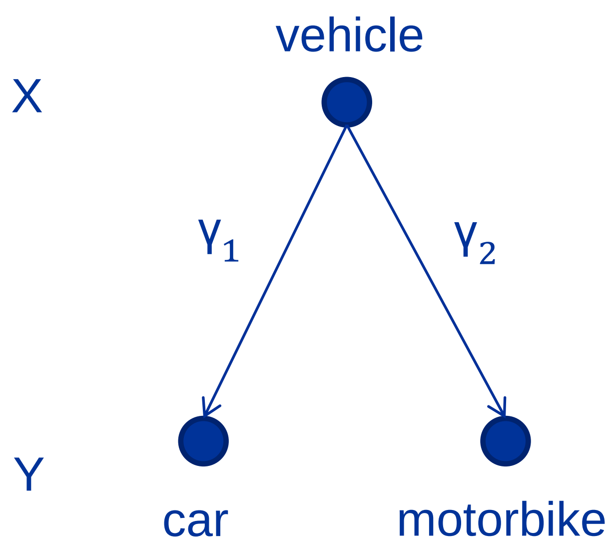
  - **The relationships may depend on the semantic level of the vocabulary**
    - **Low-level: “adjacent”, “disjoint”**
    - **Middle-level: ”radial”, ”bordering”, “hinge”**
    - **High-level: “support”, “occlude”**

Visual grammars



Evidential grammars

- **An evidential grammar is a 5 tupple (S,V_N,V_T, Γ, M)**
- **M contains a set of conditional mass functions defining the grammar rules**



Assumption of complete ignorance:
m(Yϵ{car,motorbike}|X=vehicle})=1

Why using evidential grammars?

- **The knowledge about the situations which can occur in traffic scenes has to be used**
  - The lack of training data can be balanced by supplying knowledge to the system

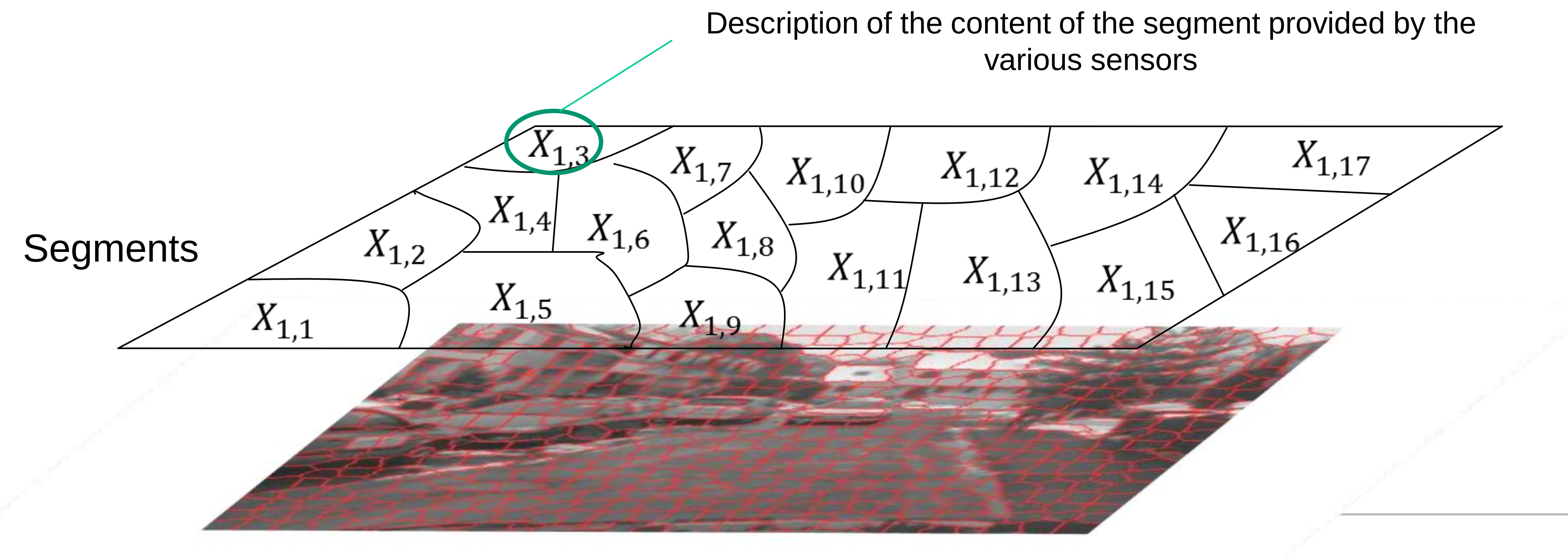
- **Learn part-of objects can be used to detect several objects**
  - A wheel can belong to a truck, a car, a bike, a moto ...

- **The informations provided by the various sensors will provide belief on the different spaces**
  - It is of highest importance for us to handle the uncertainty in the interpretation process

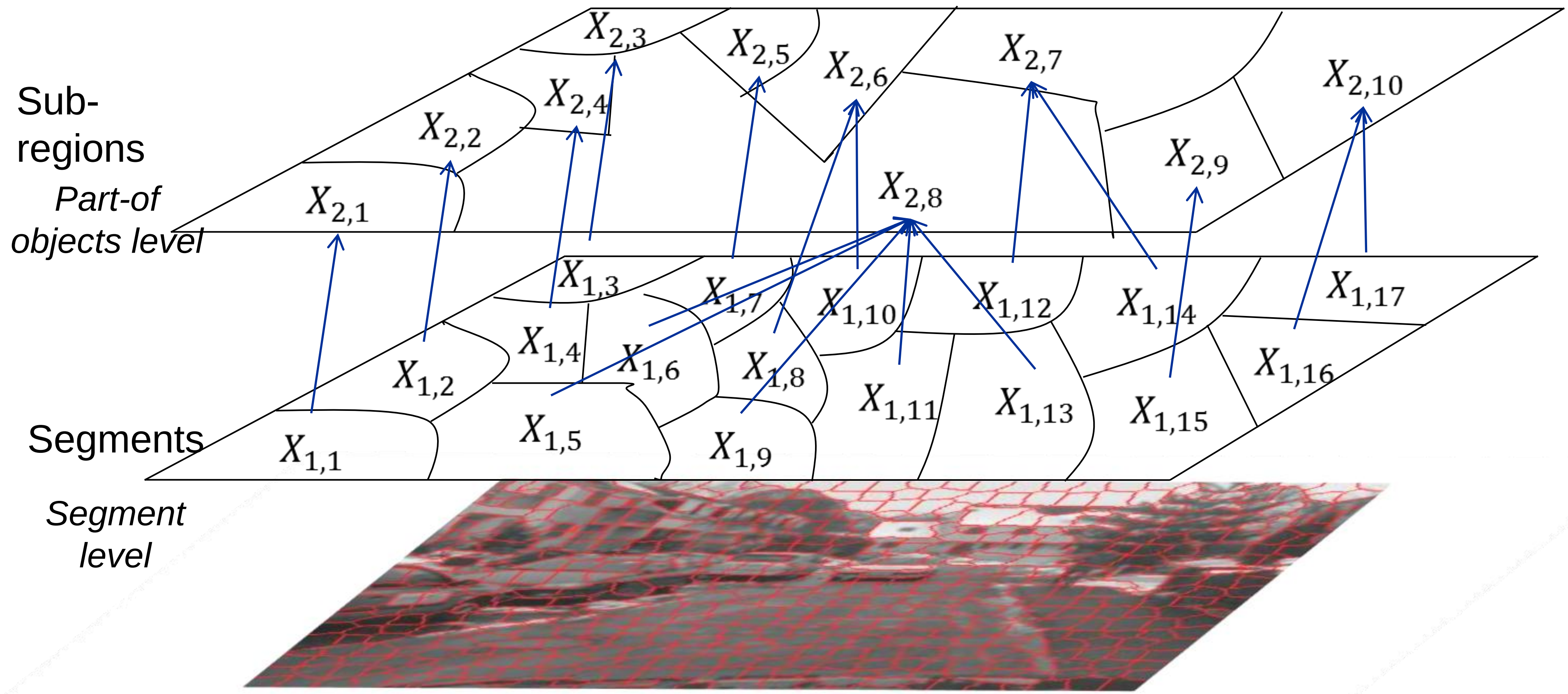
=> A precise framework to handle visual grammars and belief functions has to be defined

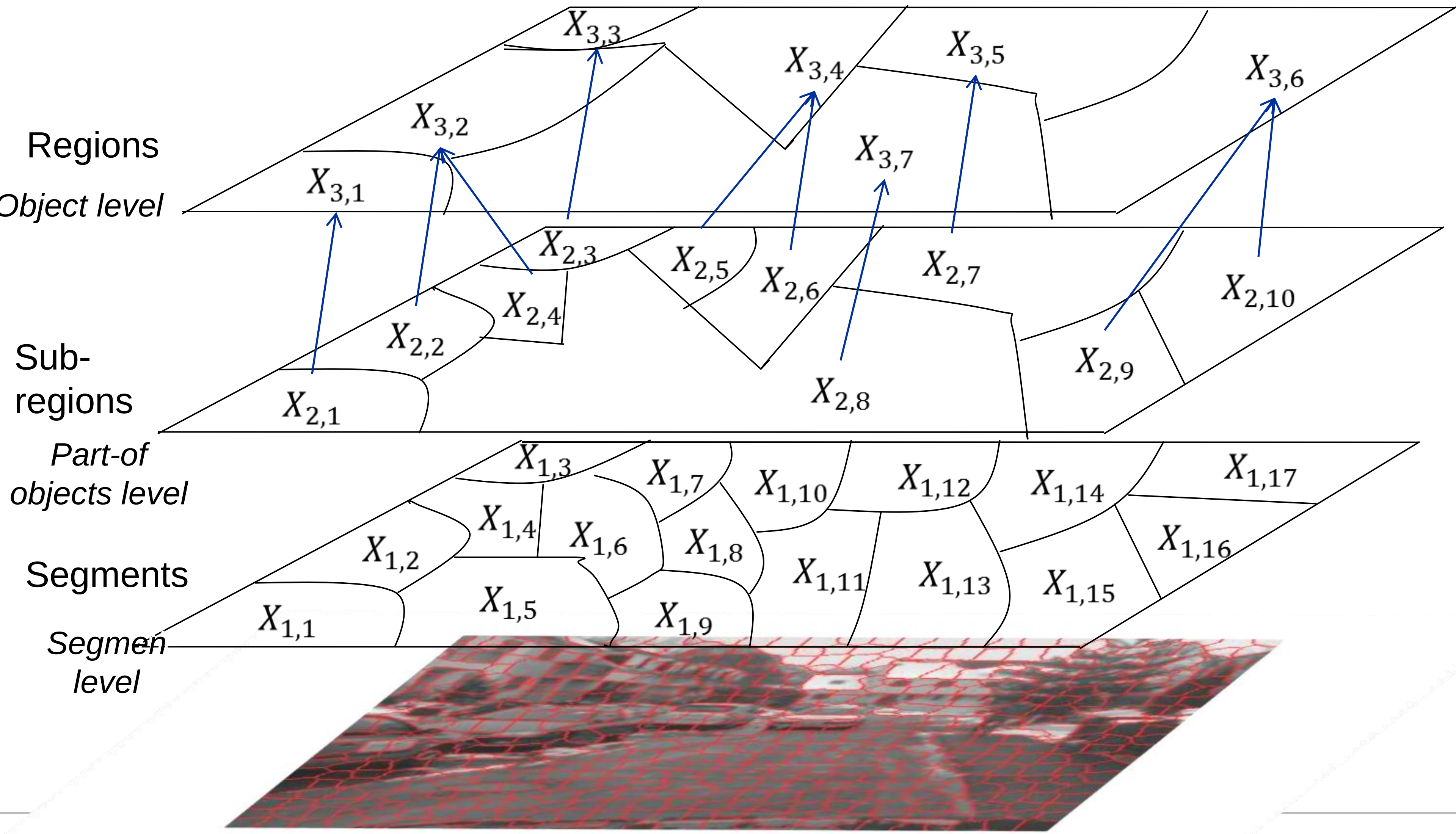
Interpretation Tree

- **The initial image is oversegmented and the information about the class contained in each one of these segments is described by a belief function and modelled by a random variable**

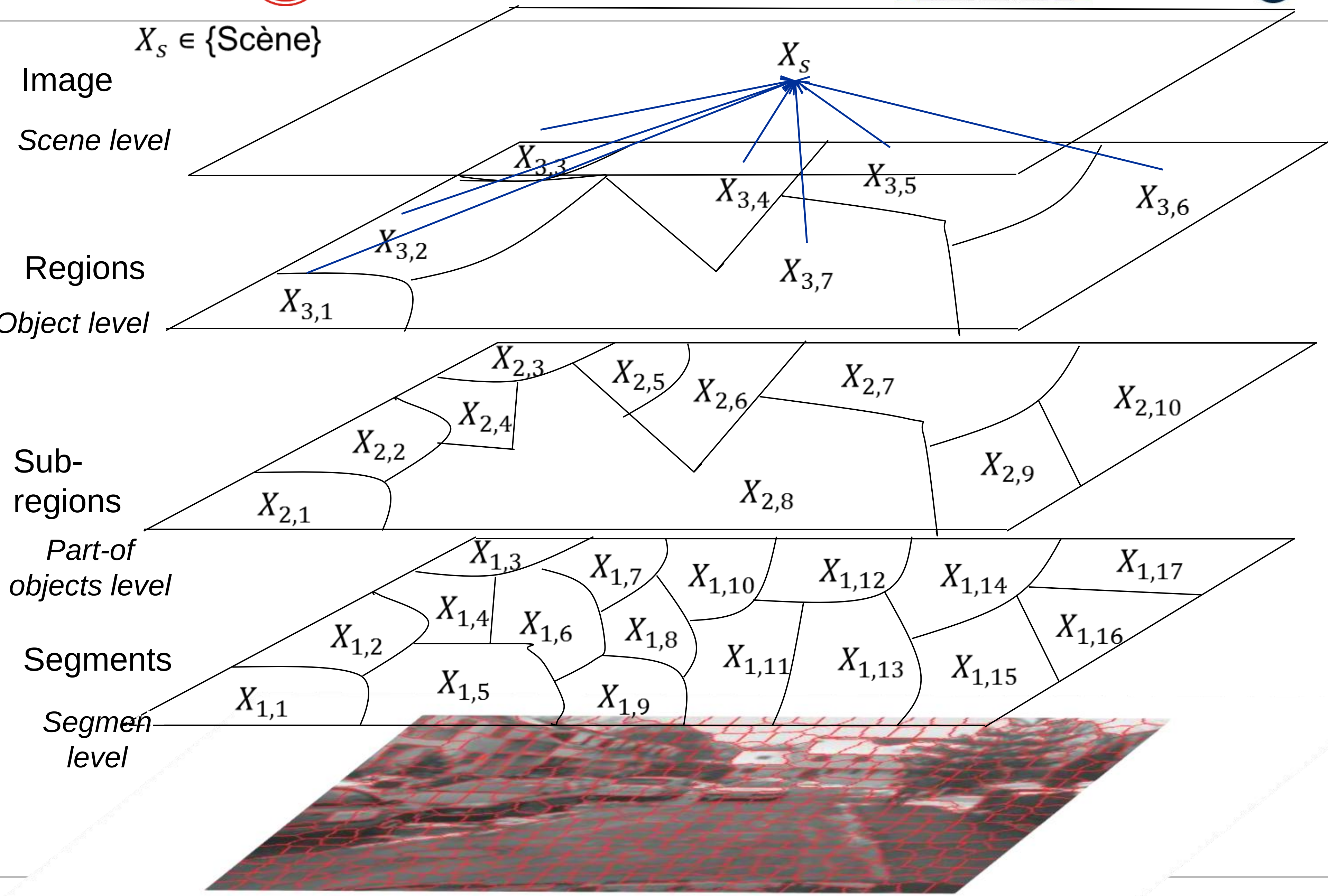


Interpretation Tree





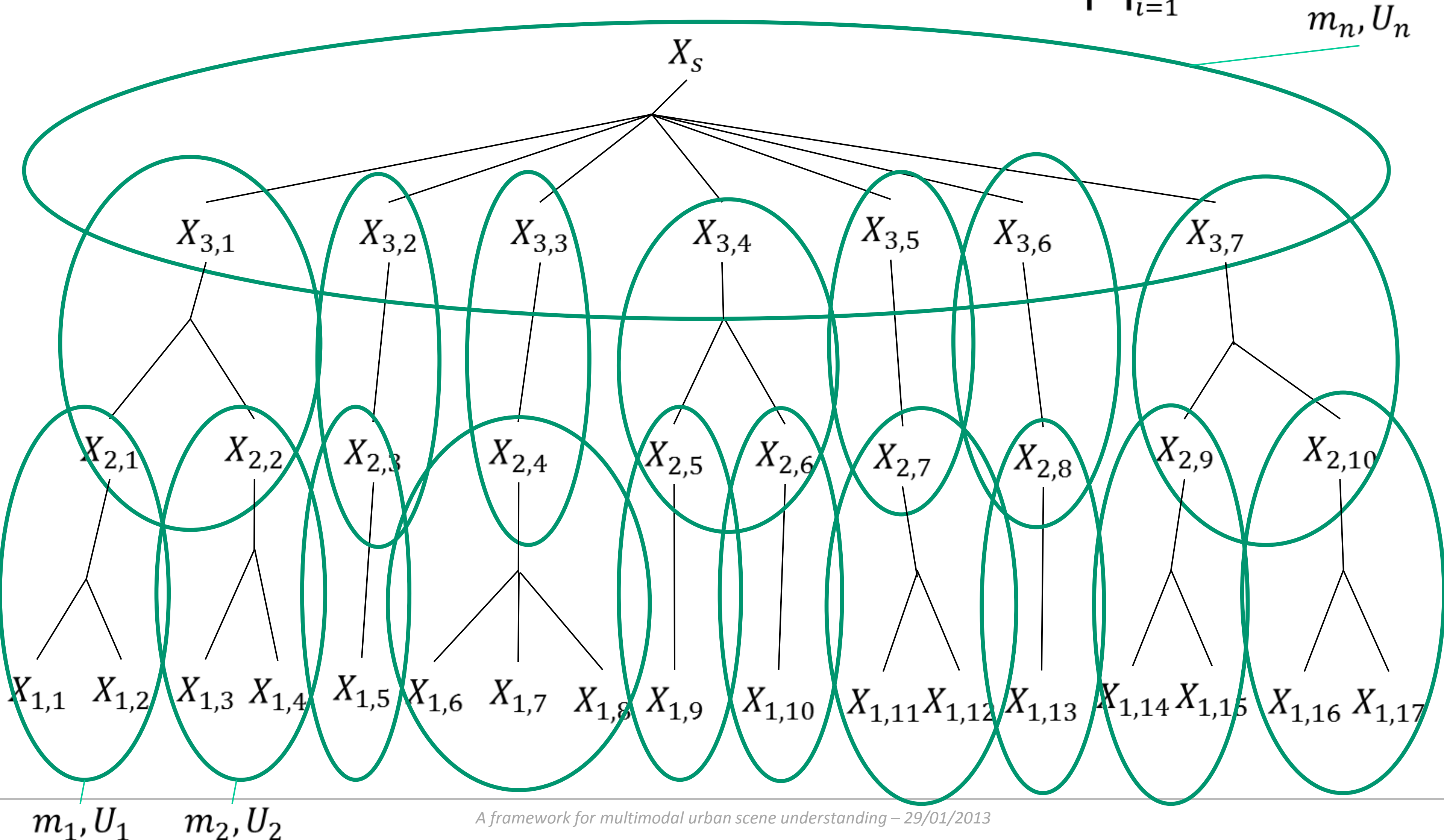
Regions
Object level
Sub-regions
Part-of objects level
Segments
Segment level
$X_{3,1}$
$X_{3,2}$
$X_{3,3}$
$X_{3,4}$
$X_{3,5}$
$X_{3,6}$
$X_{3,7}$
$X_{2,1}$
$X_{2,2}$
$X_{2,3}$
$X_{2,4}$
$X_{2,5}$
$X_{2,6}$
$X_{2,7}$
$X_{2,8}$
$X_{2,9}$
$X_{2,10}$
$X_{1,1}$
$X_{1,2}$
$X_{1,3}$
$X_{1,4}$
$X_{1,5}$
$X_{1,6}$
$X_{1,7}$
$X_{1,8}$
$X_{1,9}$
$X_{1,10}$
$X_{1,11}$
$X_{1,12}$
$X_{1,13}$
$X_{1,14}$
$X_{1,15}$
$X_{1,16}$
$X_{1,17}$



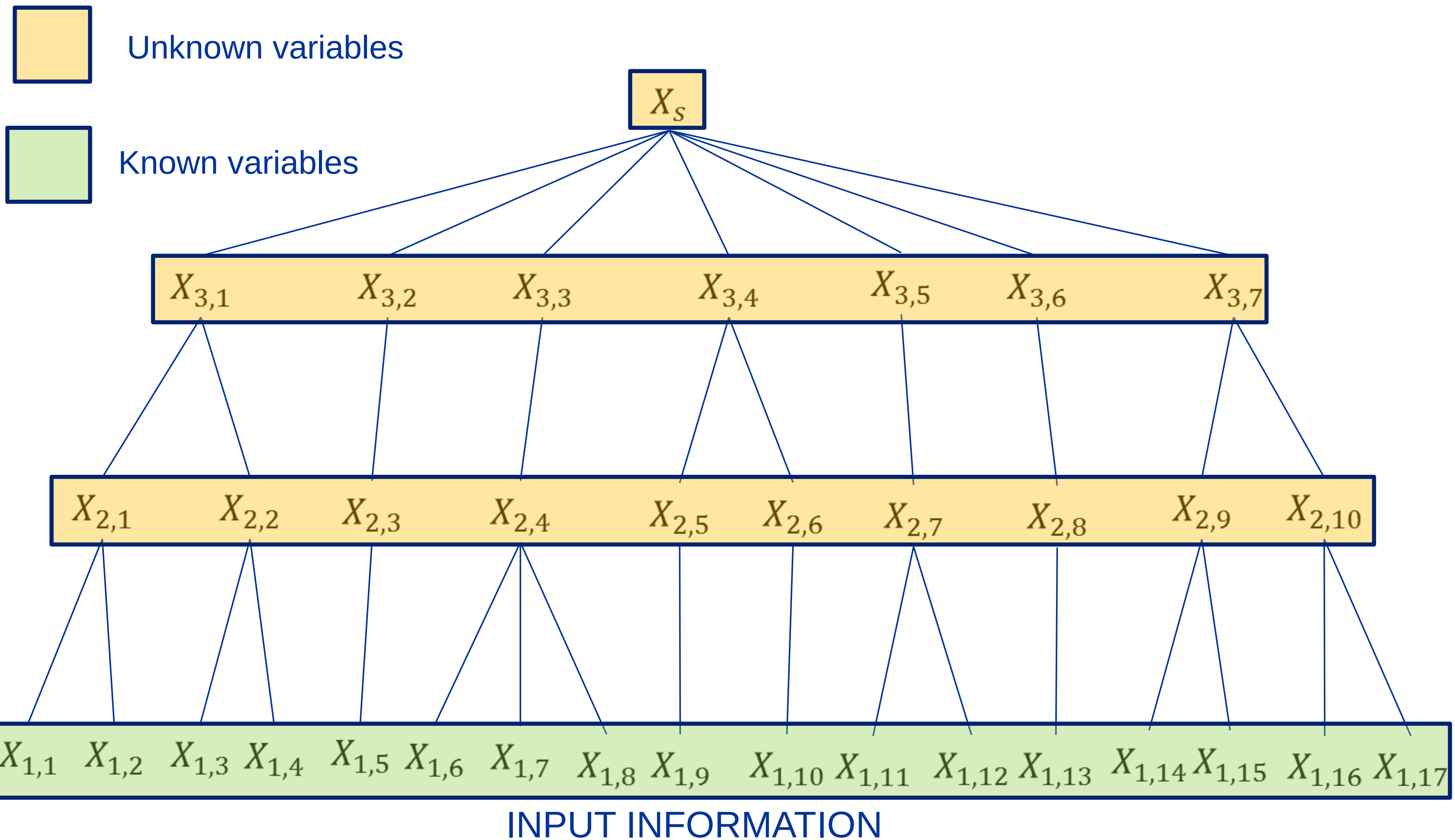
$X_s \in$ {Scène}
Image
Scene level
X_s
$X_{3,1}$ $X_{3,2}$ $X_{3,3}$ $X_{3,4}$ $X_{3,5}$ $X_{3,6}$ $X_{3,7}$
Regions
Object level
$X_{2,1}$ $X_{2,2}$ $X_{2,3}$ $X_{2,4}$ $X_{2,5}$ $X_{2,6}$ $X_{2,7}$ $X_{2,8}$ $X_{2,9}$ $X_{2,10}$
Sub-regions
Part-of objects level
$X_{1,1}$ $X_{1,2}$ $X_{1,3}$ $X_{1,4}$ $X_{1,5}$ $X_{1,6}$ $X_{1,7}$ $X_{1,8}$ $X_{1,9}$ $X_{1,10}$ $X_{1,11}$ $X_{1,12}$ $X_{1,13}$ $X_{1,14}$ $X_{1,15}$ $X_{1,16}$ $X_{1,17}$
Segments
Segment level

Evidential Network

$$m_U = \bigcap_{i=1}^{n} m_{U_i \uparrow U}$$



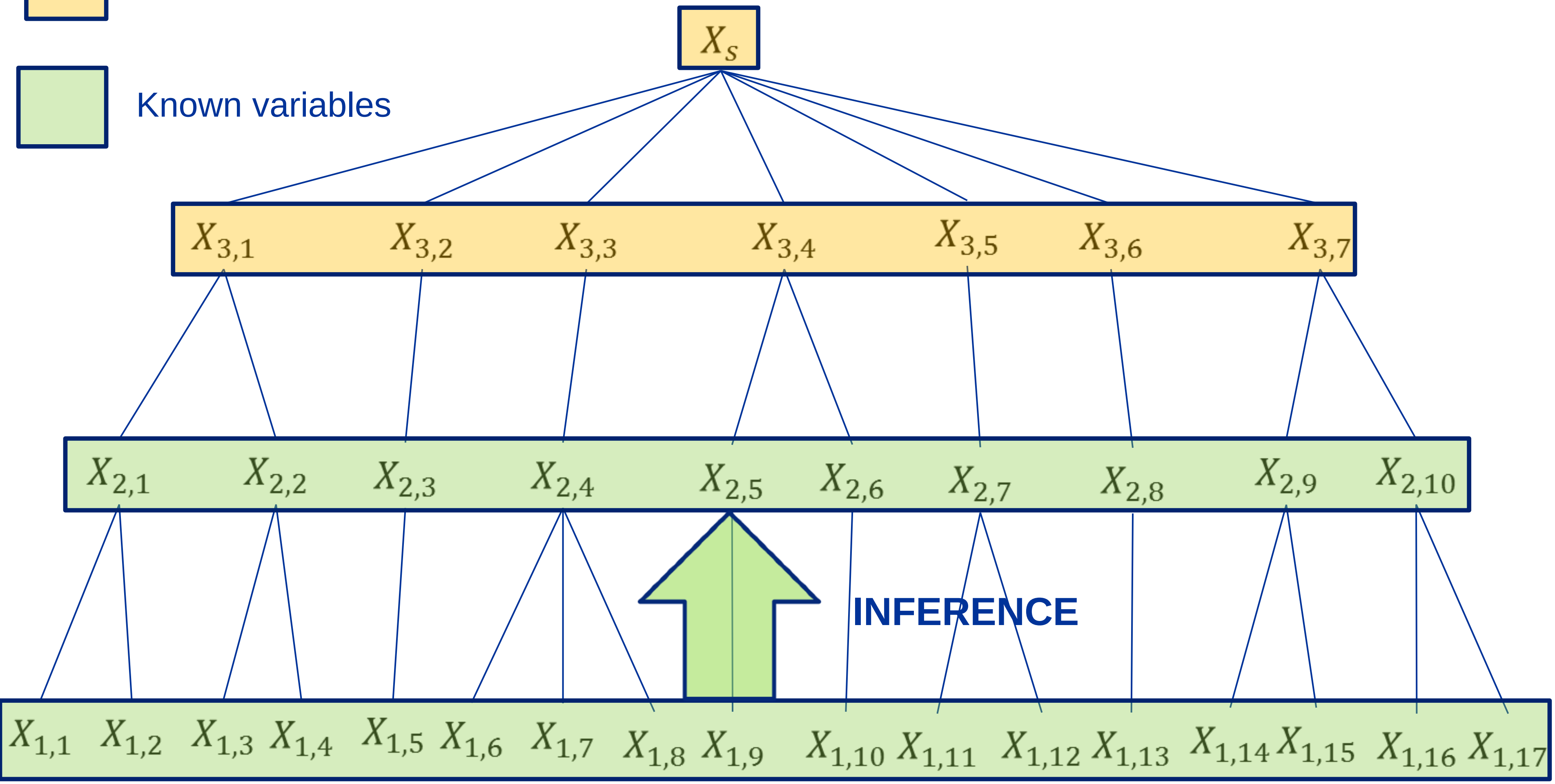
Bottom-up Propagation of the belief



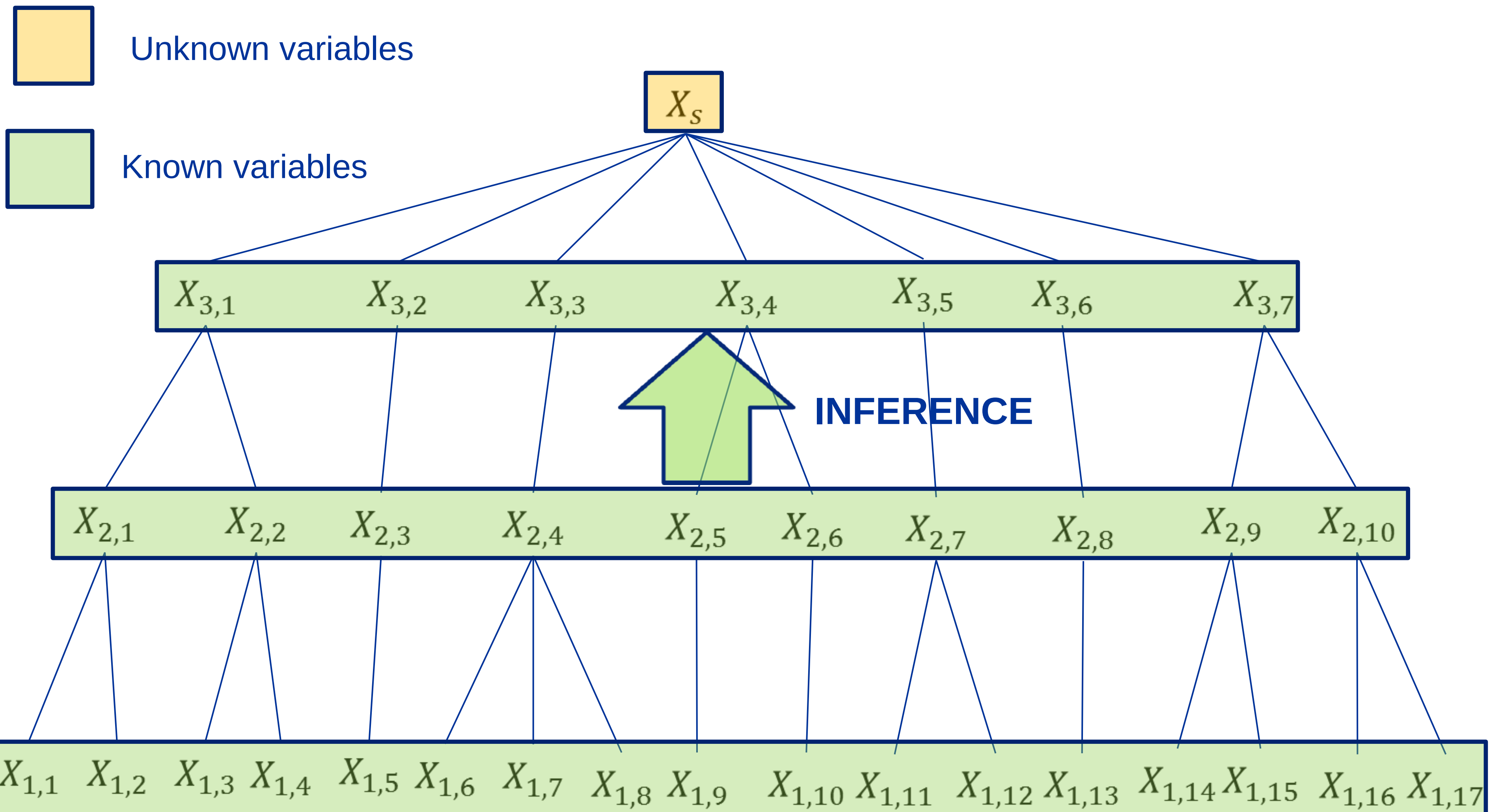
Bottom-up Propagation of the belief

Unknown variables

Known variables



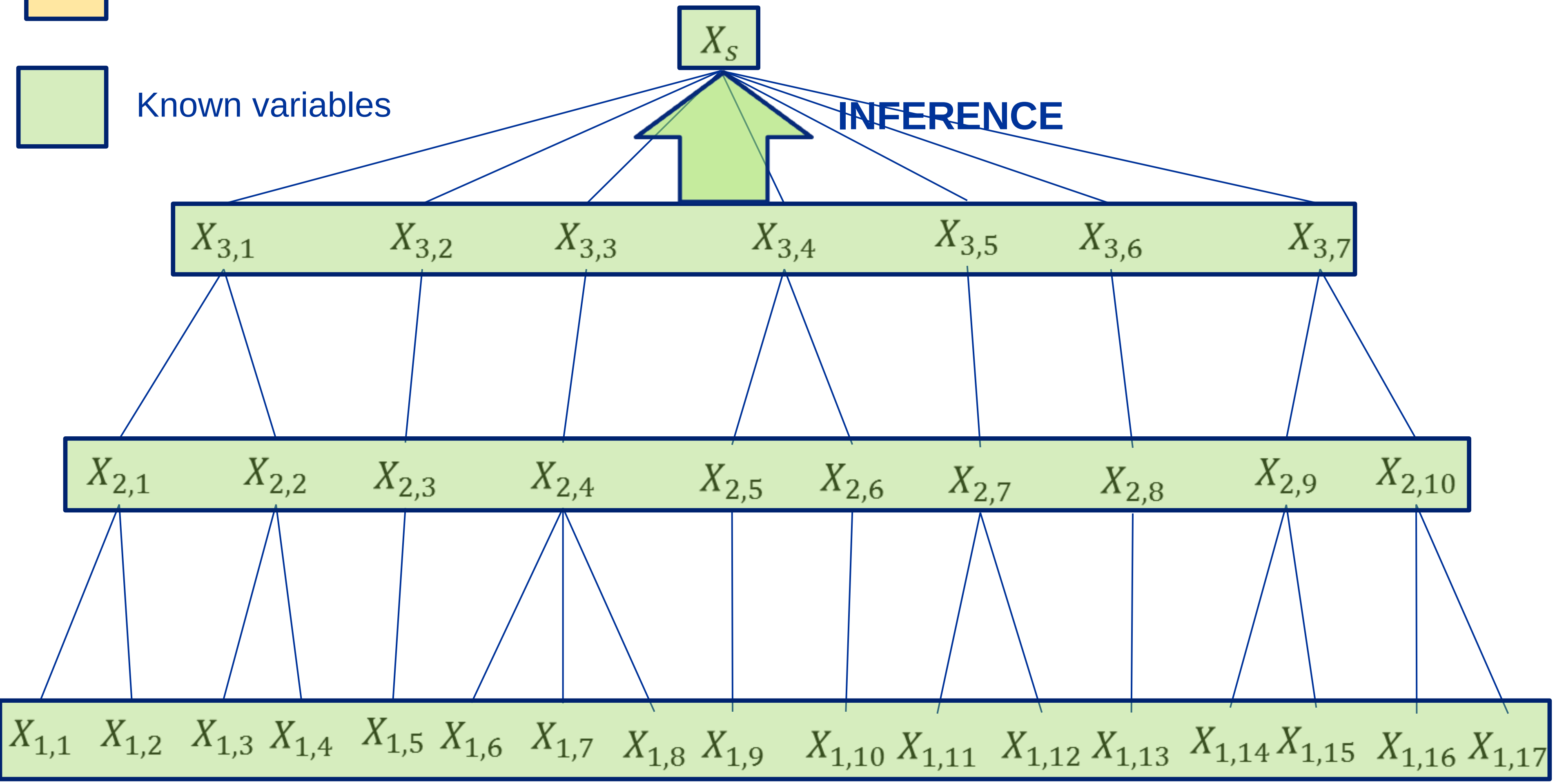
Bottom-up Propagation of the belief



Bottom-up Propagation of the belief

Unknown variables

Known variables



Research of the Parse tree

- X_S has only one possible value: S (Scene)
  - **m_{X_S} is defined on {S,Ø}**
  - **$m_{X_S}(\emptyset)$ measures the consistency of the image interpretation**

- Optimal interpretations of an image :
  - **Parse tree minimizing $m_{X_S}(\emptyset)$**

Experiments

- **Evidential grammars is a framework:**
  - **The 5-tuple (S,V_N,V_T, Γ, $\mathbf{M}$) has to be instantiated**
  - **For traffic scenes, we have to define:**
    - **The objects and parts of objects**
    - **Spatial relationships**
    - **The production rules**
- **First step of experiments:**
  - **input directly at the objects level**
- **Second step of experiments:**
  - **input at the part-of-objects level**

Ongoing student works

- **Semi-automatic data annotation**
  - **5 master students (Beihang University, Ecole Centrale)**
- **Local feature (texture) analysis**
  - **1 master student (Peking University)**
- **Part-based object detection**
  - **Master internship (Spring 2013)**
- **One class learning**
  - **Master internship (Spring 2013)**

Ongoing and future publications

- **IAPR MVA 2013:**
  - Information Fusion on Oversegmented Images: An Application for Urban Scene Understanding
- **ORASIS 2013:**
  - Fusion d’informations sur des images sursegmentées : Une application à la compréhension de scènes routières
- **IUKM 2013:**
  - Evidential Grammars Framework for Image Interpretation. Application to Multimodel Traffic Scene Understanding