DLR

Information Theory based Inference: applications to image mining and semantic coding

Mihai Datcu

Remote Sensing Technology Institute

Stochastic modelling: telecommunications, geosciences, bio-medicine, multimdeia, social and economic sciences, etc.

Image information: acquisition, processing, visualization, and analysis of spatial information. Thus modelling and understanding.

Statistical modeling: discover regularities in observed data. The more regularities are found in the data, the more data can be compressed. Finding regularities in the data means nothing else but to learn about the data.

Links between stochastic modelling, information theory, and machine learning, provided new basic theories for information processing and understanding.

DLR

Remote Sensing Technology Institute

TOPICS

applications and extensions of Rate-Distortion theory
the methods of information bottleneck
MDL and related methods
information and complexity based estimation and inference

image understanding
image semantic coding
image indexing and information mining
search engines in large (satellite) image archives
picture archiving and communication systems (PACS)

A Mathematical Theory of Communication

By C. E. SHANNON

Introduction

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

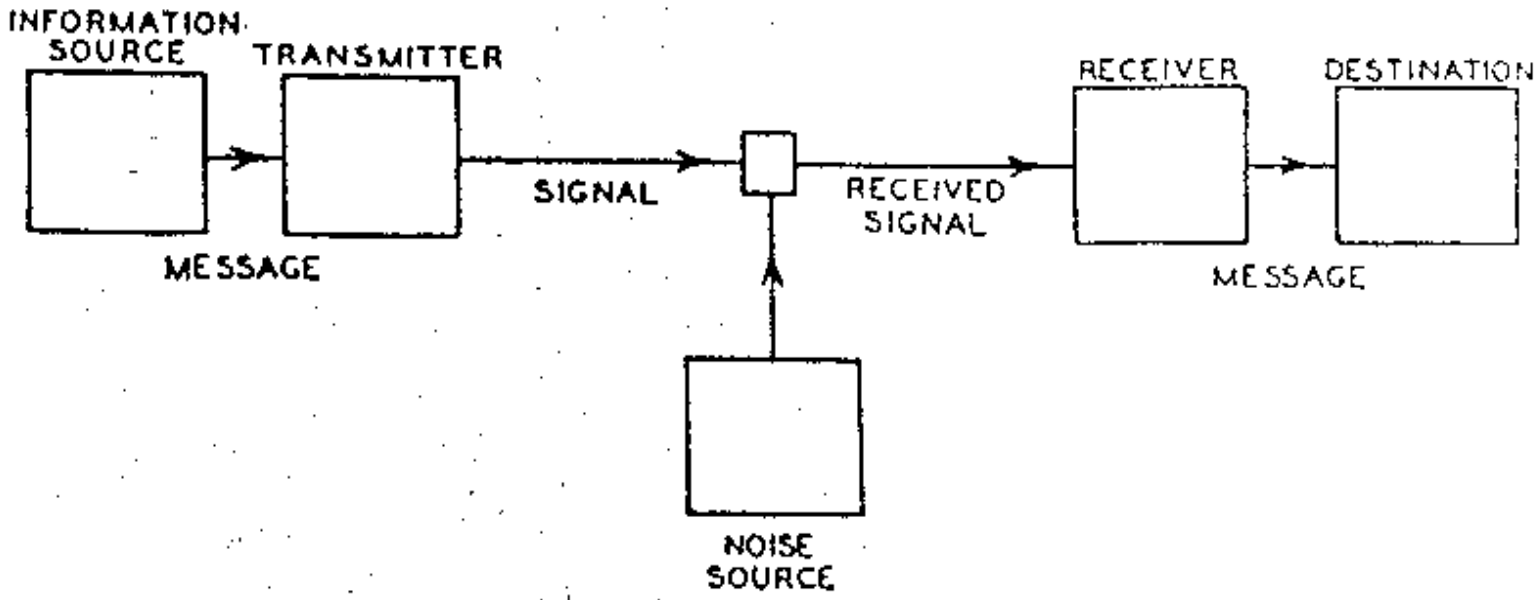


Fig. 1—Schematic diagram of a general communication system.

3. The Series of Approximations to English

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27-symbol "alphabet," the 26 letters and a space.

1. Zero-order approximation (symbols independent and equi-probable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ
   FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI
   ALHENHTTPA OOBTTVA NAH BRL

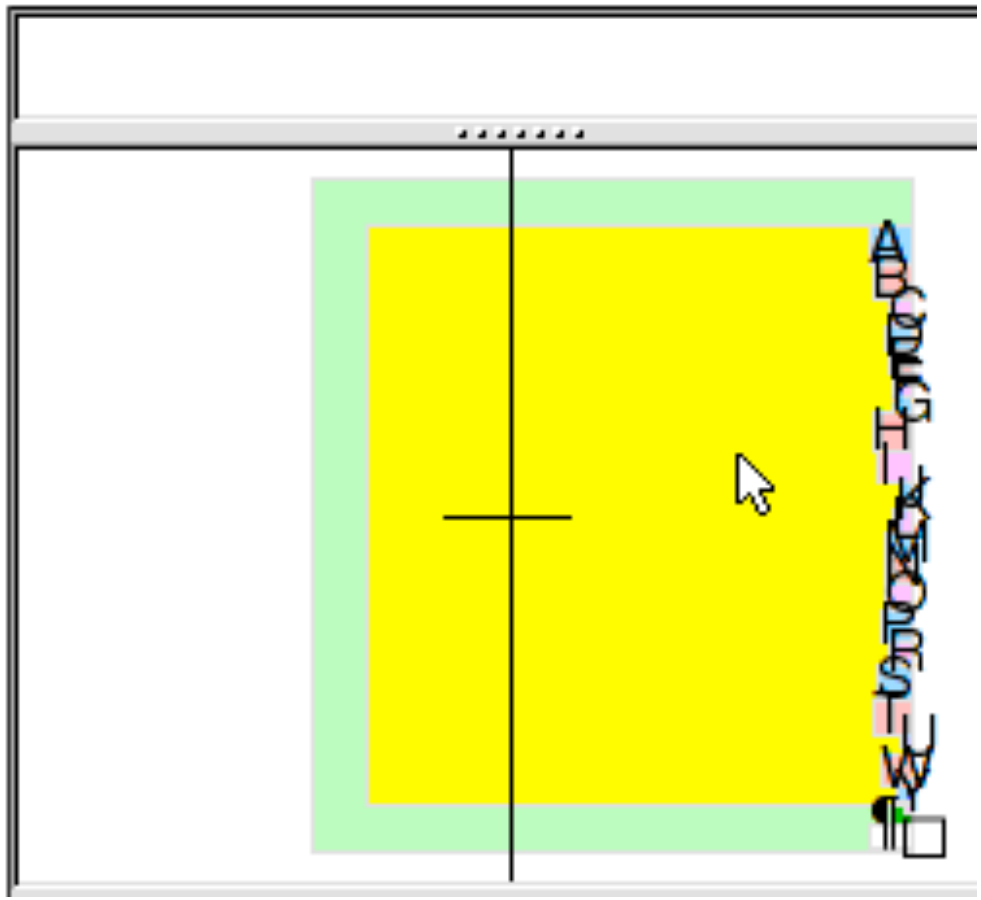
3. Second-order approximation (digram structure as in English).

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE

4. Third-order approximation (trigram structure as in English).

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

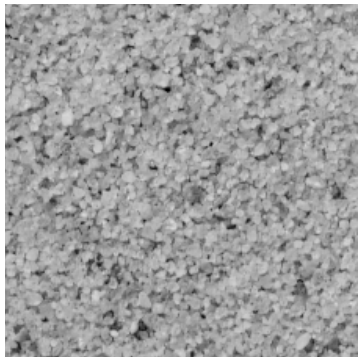
McKay´s *dasher*

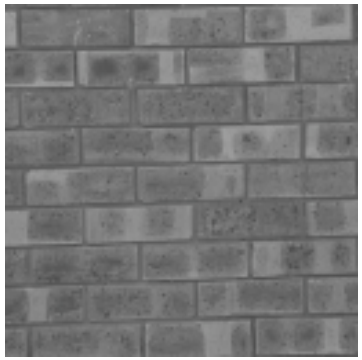
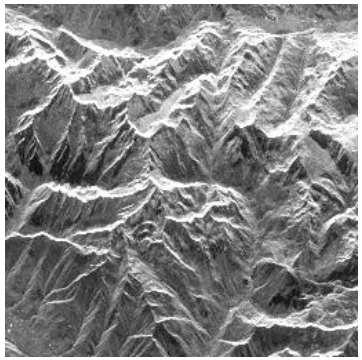
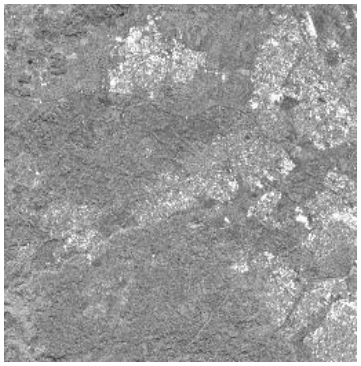


DLR

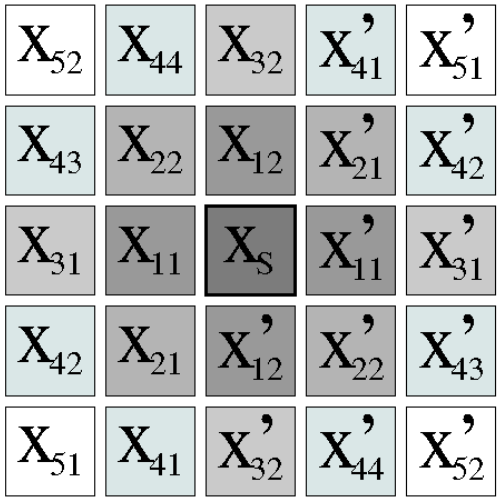
Remote Sensing Technology Institute

Questions: How to **define** and how to **describe** a texture?




Graphical Definition of a GRF



Neighbor-hood		
Cliques		
Parame-ters	θ_0 θ_{12} θ_{11}	θ_0 θ_{12} θ_{11} θ_{21} θ_{22}
Potential	V_1^0 V_2^{12} V_2^{11}	V_1^0 V_2^{12} V_2^{11} V_2^{21} V_2^{22}
Energy	H^1	H^2

$$p(x_S|\partial x_S, \boldsymbol{\theta}) = \frac{1}{Z_S} \exp\left(-H(x_S, \partial x_S; \boldsymbol{\theta})\right)$$

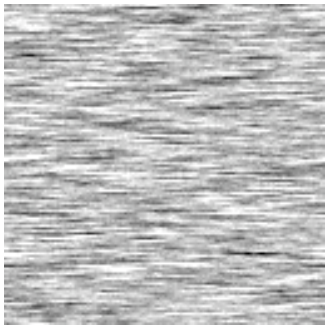
$$\boldsymbol{\theta} = (\theta_0, \underbrace{\theta_{11}, \theta_{12}}_{\text{1st order}}, \underbrace{\theta_{21}, \theta_{22}}_{\text{2nd order}}, \underbrace{\theta_{31}, \theta_{32}}_{\text{3rd order}}, \dots)$$

Gauss-Markov Random Fields (GMRF)

Definition:

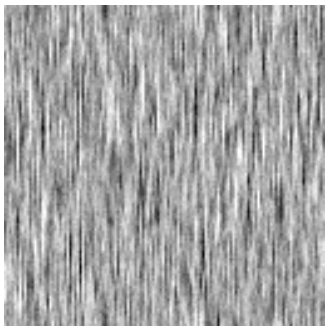
Probability density function $p(x_S|\partial x_S;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_S - \sum_{kl} \theta_{kl}(x_{kl} + x'_{kl}))^2}{2\sigma^2}\right)$

Synthesis:



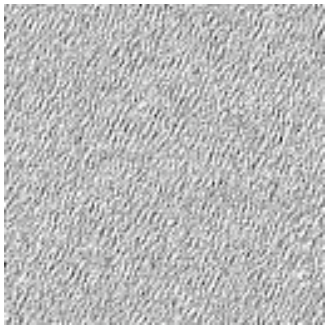
2nd order

$\boldsymbol{\theta} = (0.5, 0, 0, 0)$



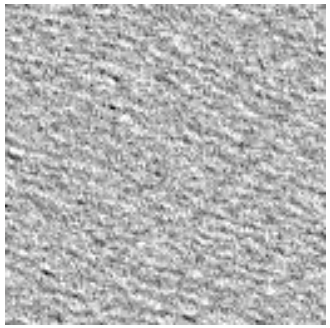
2nd order

$\boldsymbol{\theta} = (0, 0.5, 0, 0)$



2nd order

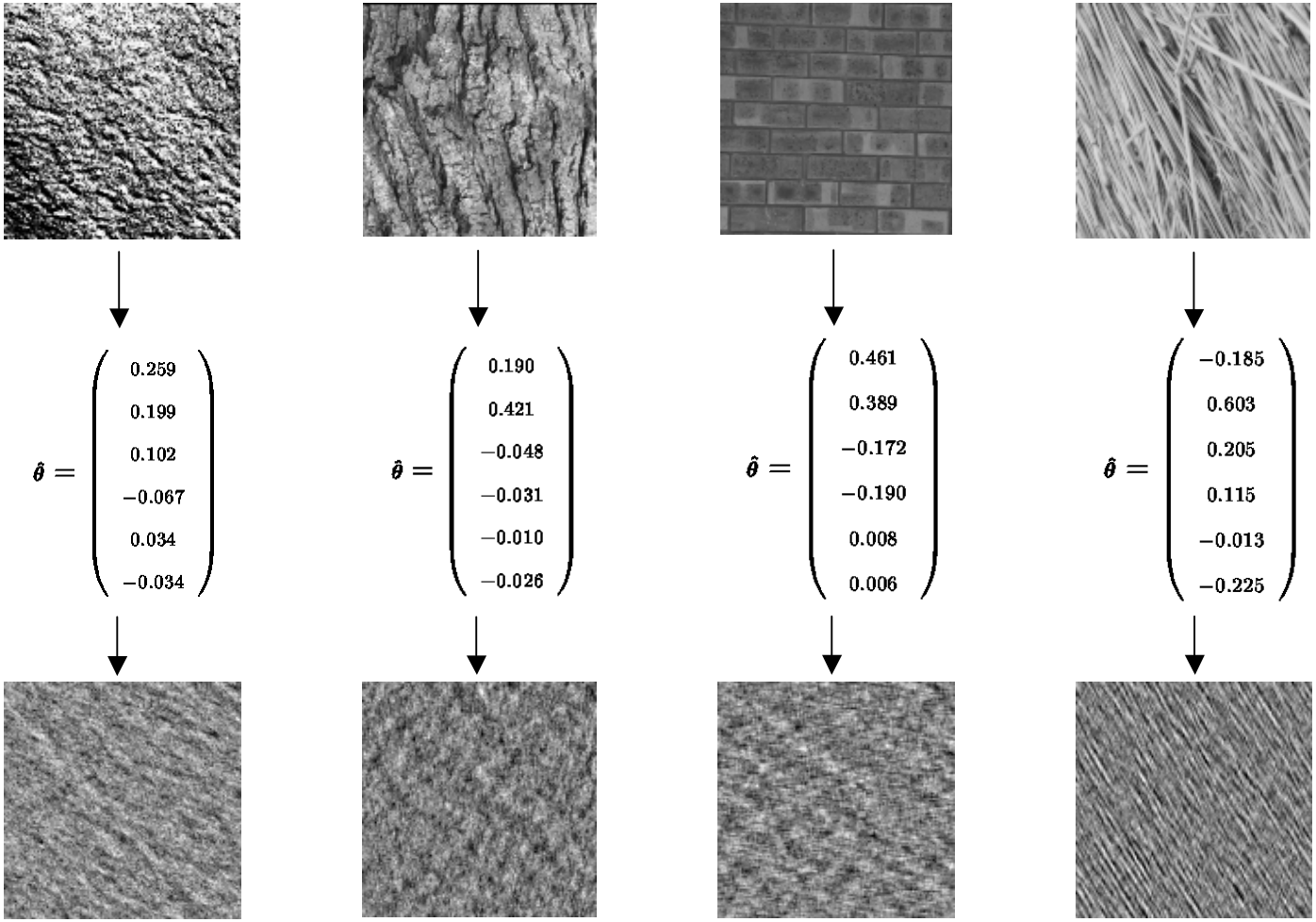
$\boldsymbol{\theta} = (-0.2, 0.1, -0.15, 0.15)$



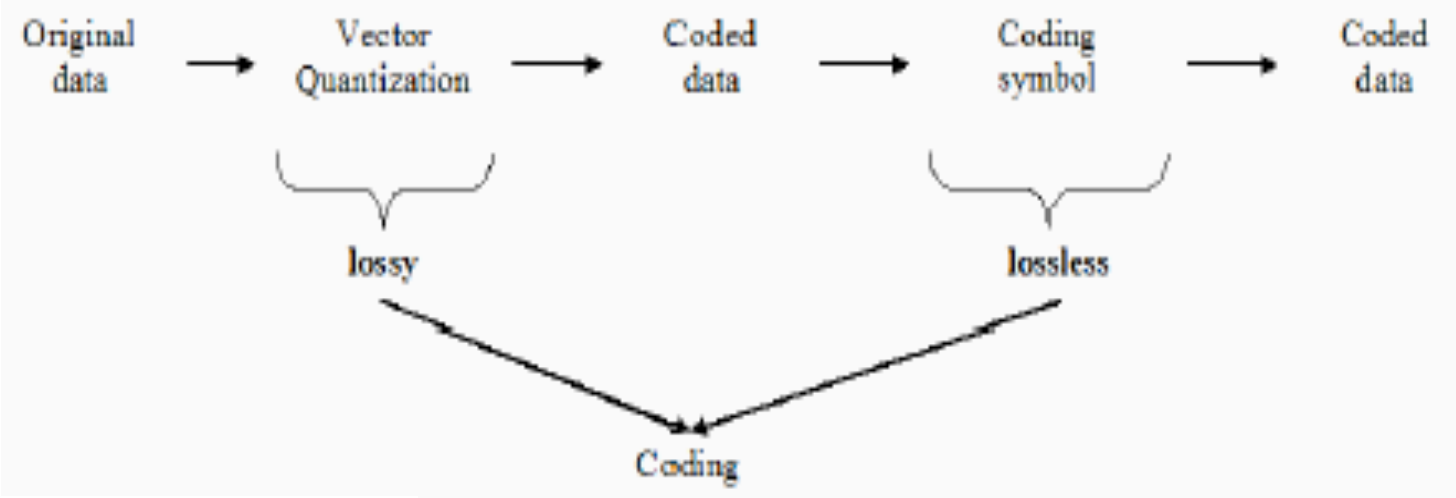
3rd order

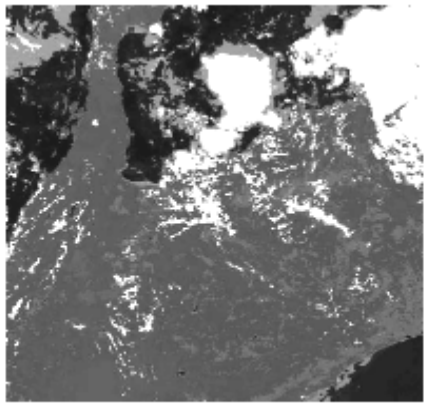
$\boldsymbol{\theta} = (0.2, 0.2, 0.1, -0.1, 0.1, -0.1)$

Parameter Estimation

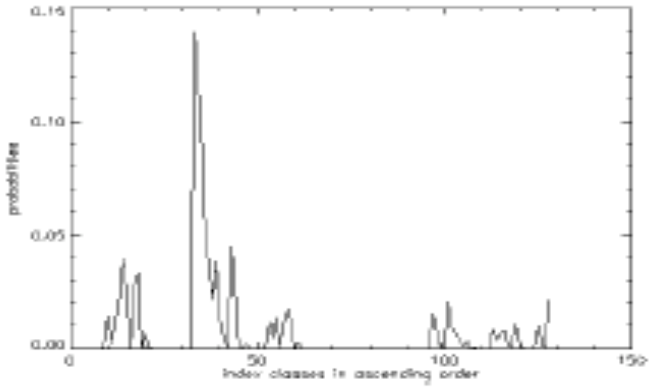


Coding



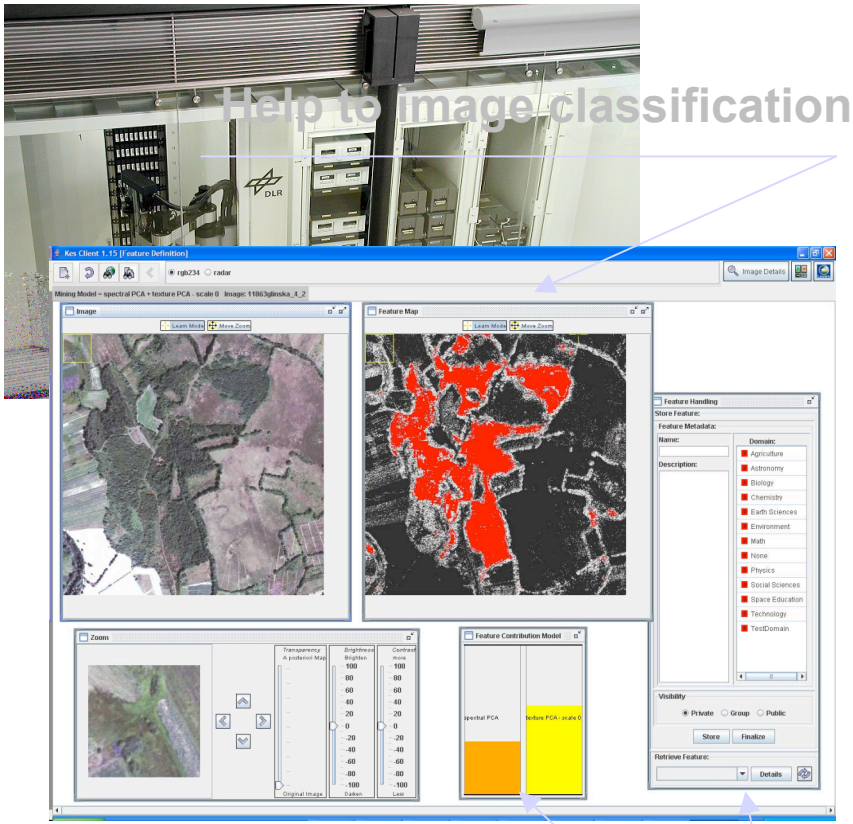


Coded Image



(b) Class probability $p(\omega \mid I)$) of the coded image with index class in ascending order

Interactive, user adapted, EO data content access



Mine Fields (Daedalus)

Access to:
information
knowledge

Represent degrees of belief by probabilities

$0 \le Pr \le 1$ 0 = FALSE, 1 = TRUE.

obeying

$$Pr(X) + Pr(\overline{X}) = 1$$
$$Pr(A,B) = Pr(A)\,Pr(B|A)$$

‖ I

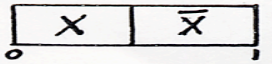
Sum & product rules

Rules plausible — obeyed by proportions.

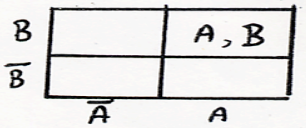
Cox sanity checks:

① A believed more than B, B believed more than C } ⇒ A believed more than C

② Knowledge of X ⇒ Knowledge of $\overline{X}$

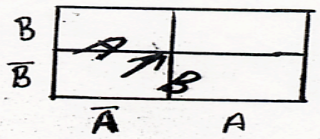


③ Knowledge of A, Knowledge of B|A } ⇒ Knowledge of (A,B)

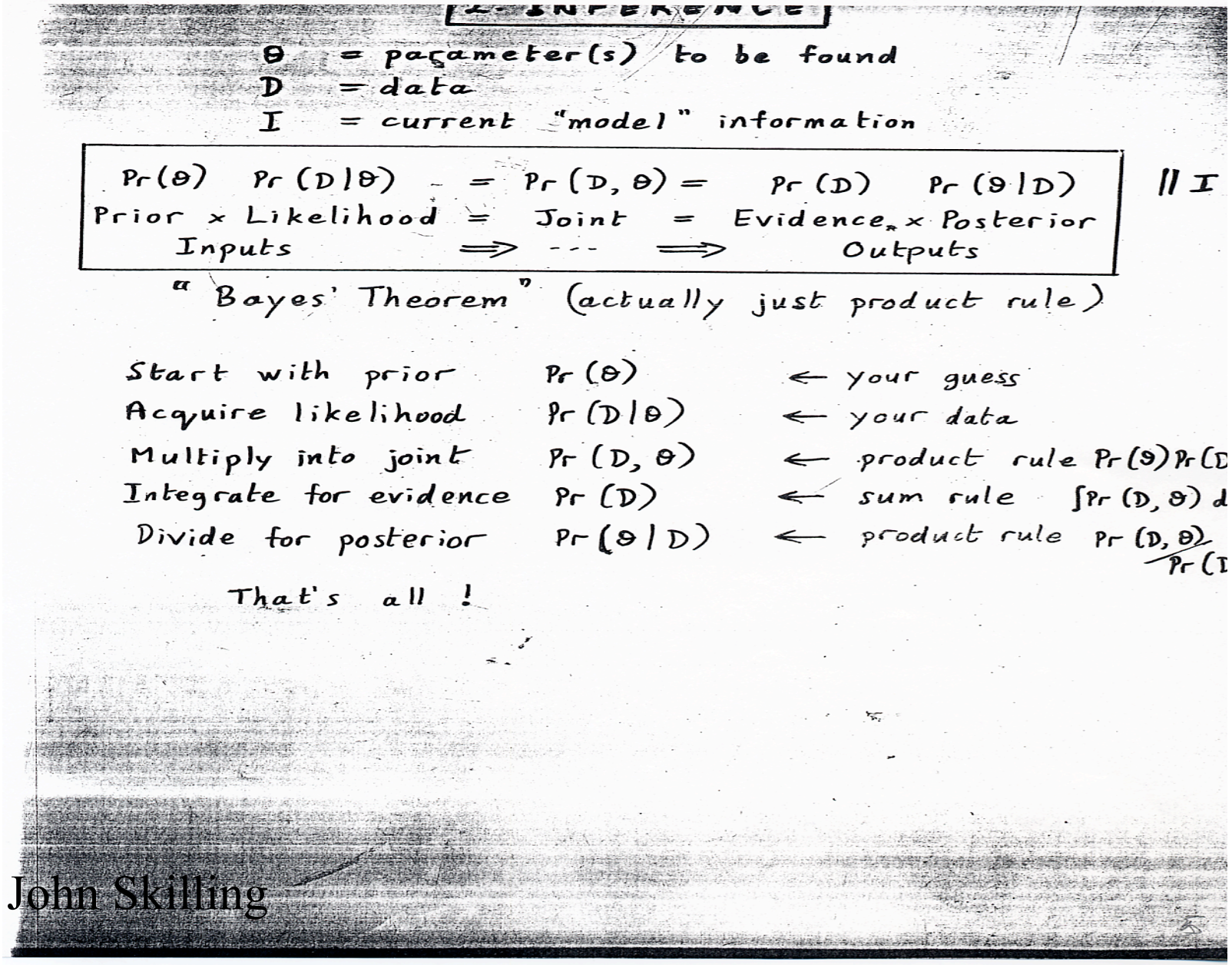


OR Garrett's NAND simplification

(2&3) Knowledge of $\overline{A}$, Knowledge of B|A } ⇒ Knowledge of $A \uparrow B$



John Skilling



2. INFERENCE

θ = parameter(s) to be found
D = data
I = current "model" information

$$Pr(\theta)\ Pr(D|\theta) = Pr(D,\theta) = Pr(D)\ Pr(\theta|D) \quad \| I$$

Prior × Likelihood = Joint = Evidence × Posterior
Inputs ⟹ --- ⟹ Outputs

"Bayes' Theorem" (actually just product rule)

Start with prior	$Pr(\theta)$	← your guess
Acquire likelihood	$Pr(D|\theta)$	← your data
Multiply into joint	$Pr(D,\theta)$	← product rule $Pr(\theta)Pr(D$
Integrate for evidence	$Pr(D)$	← sum rule $\int Pr(D,\theta)\,d$
Divide for posterior	$Pr(\theta|D)$	← product rule $\frac{Pr(D,\theta)}{Pr(D$

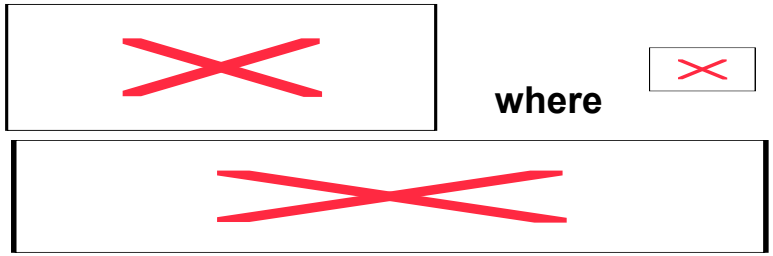
That's all !

John Skilling

Theoretical Premises

- **Rate Distortion Based Analysis of Image Parameters Estimation for Information Mining**
  - **Parameter estimation and theoretical accuracy bound: Cramer/Rao**

  - **Elements of R/D theory**

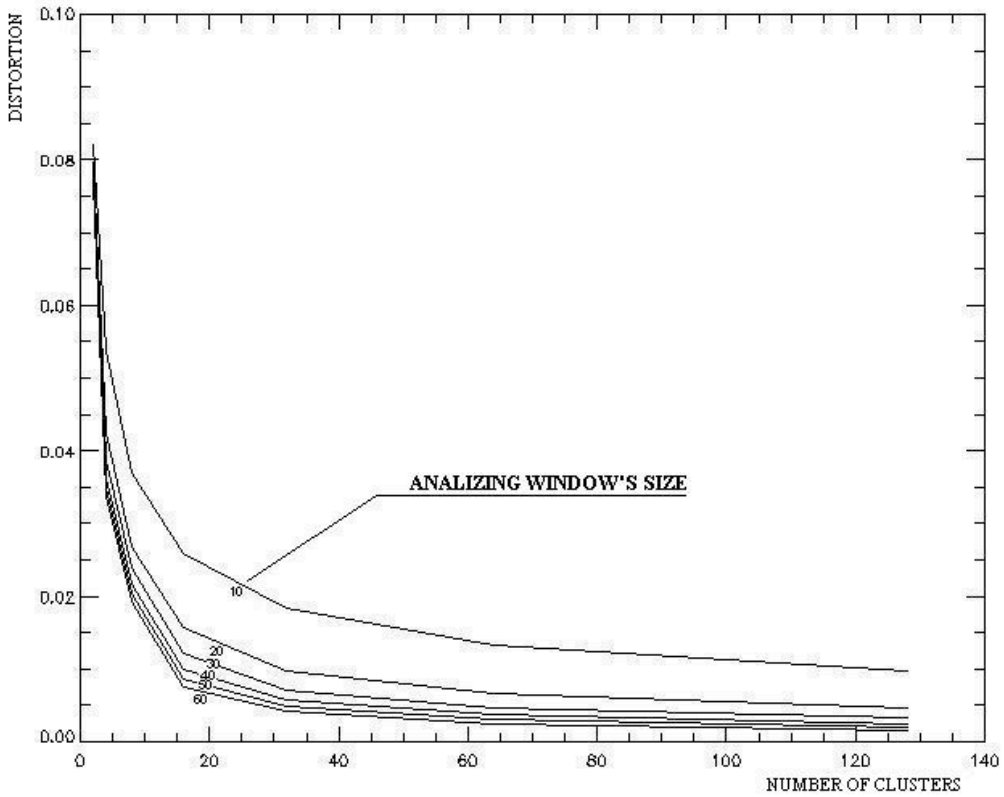


where **is the squared error distortion** 

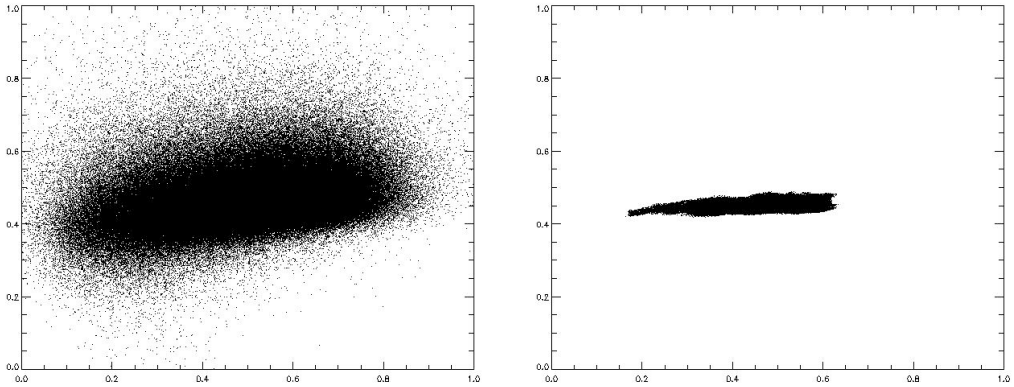
  - **Evaluation on algorithm libraries and sensor data sets:**
    - **GMRF algorithm**
    - **Dyadic K-means**
    - **Evaluation of the optimal size of the analyzing window for Landsat and Daedalus data**

Concept

- **R/D based concept for evaluation of data models and estimation algorithms:**
  - **Theoretical behavior for homogeneous texture (Brodatz data set)**



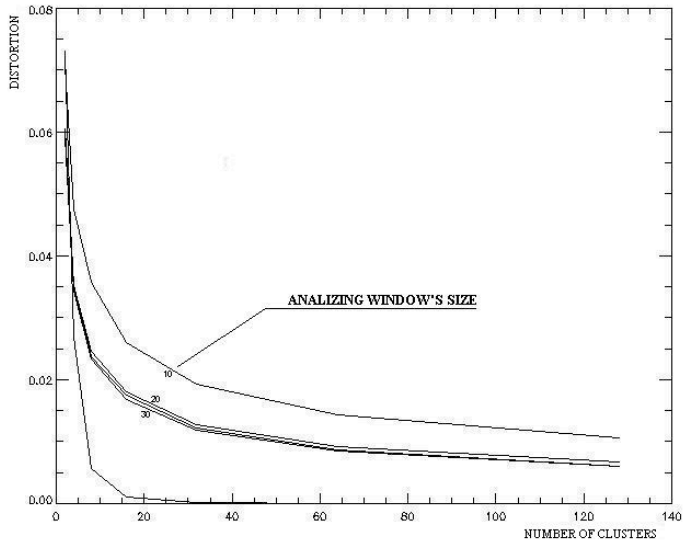
- **Visualization in the feature space of one sample image of the dataset :**
  - **For analyzing window of size 10 (left-plot)**
  - **For analyzing window of size 60 (right-plot)**

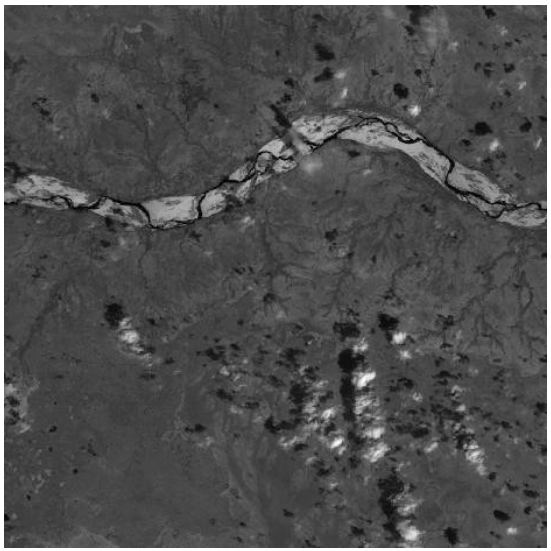


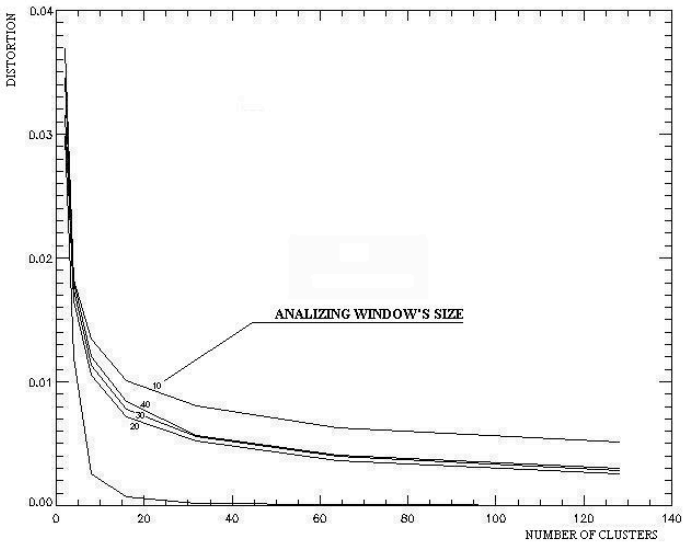
Results

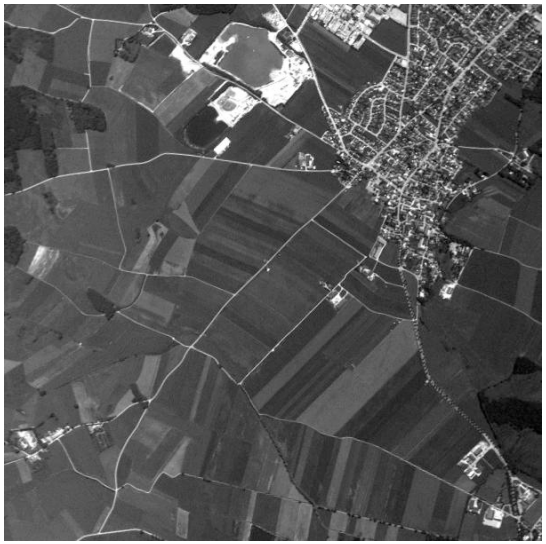
- **Results for assessment of non-stationarity data:**
  - **Landsat data set (W_L=30)**
  - **Daedalus data set (W_D=20)**

Where W is the estimated optimal size of the analyzing window







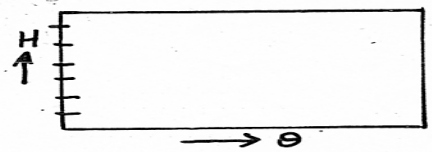


Remote Sensing Technology Institute

3. MODEL SELECTION

Modern focus of Bayesian programming:

II: Variables are model H as well as θ:



Determine joint pdf of everything, then marginalise.

$$\underbrace{\Pr(H)\,\Pr(\theta\,|\,H)}_{\text{Priors}} \times \underbrace{\Pr(D\,|\,\theta,H)}_{\text{Likelihood}} = \underbrace{\Pr(\theta,H,D)}_{\text{Joint}}$$

→ $\underbrace{\Pr(D)}_{\text{Evidence}}$ and $\underbrace{\Pr(\theta,H\,|\,D)}_{\text{Posteriors}}$ → $\Pr(H\,|\,D)$, $\Pr(\theta\,|\,D)$, $\Pr(\theta\,|\,H,D)$ ---

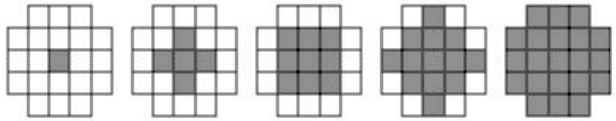
Similar to ordinary inference, (θ, H) replaces (x, y).

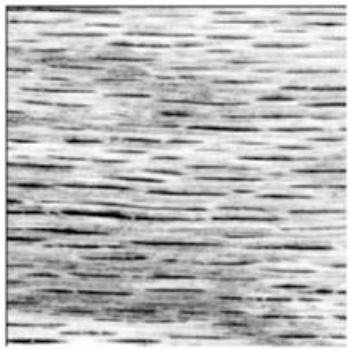
Equivalently, can accumulate results for each H in turn, then combine.

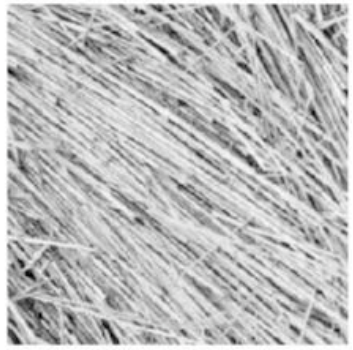
John Skilling

MORE ON MODEL SELECTION

Order of interaction

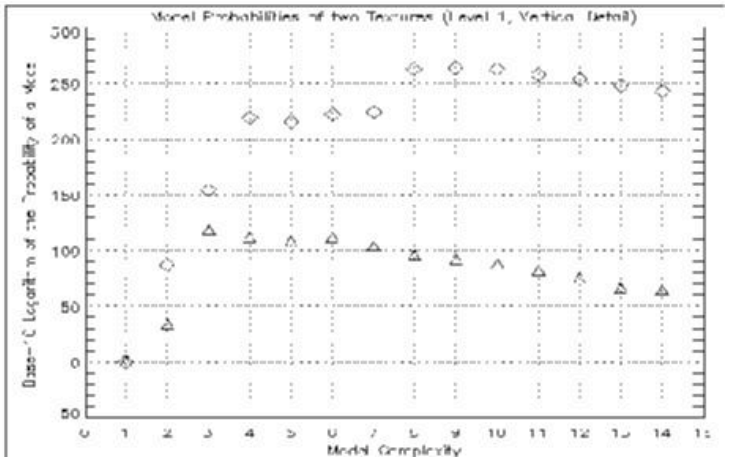






Type of interaction

$$p(x_s | x_{s+r}, r \in \mathcal{G}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{\left[x_s - \sum_{r \in \mathcal{G}} \theta_r x_{s+r} \right]^2}{2\sigma^2} \right\}$$



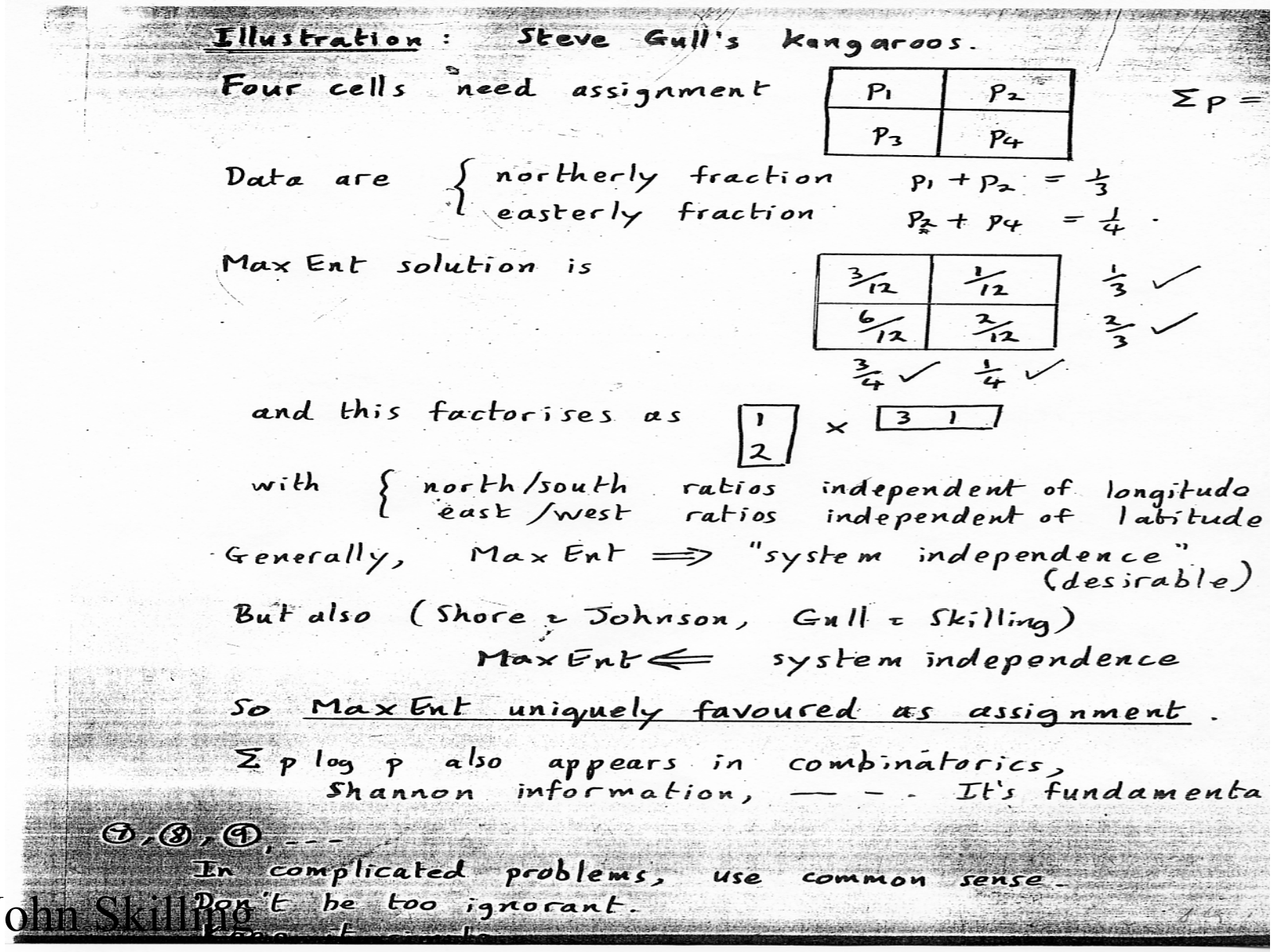


Illustration : Steve Gull's kangaroos.

Four cells need assignment

p_1	p_2
p_3	p_4

$\Sigma p =$

Data are { northerly fraction $p_1 + p_2 = \frac{1}{3}$
easterly fraction $p_2 + p_4 = \frac{1}{4}$.

Max Ent solution is

$\frac{3}{12}$	$\frac{1}{12}$	$\frac{1}{3}$ ✓
$\frac{6}{12}$	$\frac{2}{12}$	$\frac{2}{3}$ ✓
$\frac{3}{4}$ ✓	$\frac{1}{4}$ ✓	

and this factorises as $\begin{bmatrix}1\\2\end{bmatrix} \times \begin{bmatrix}3 & 1\end{bmatrix}$

with { north/south ratios independent of longitude
east/west ratios independent of latitude

Generally, Max Ent $\Longrightarrow$ "system independence" (desirable)

But also (Shore & Johnson, Gull & Skilling)
Max Ent $\Longleftarrow$ system independence

So <u>Max Ent uniquely favoured as assignment</u> .

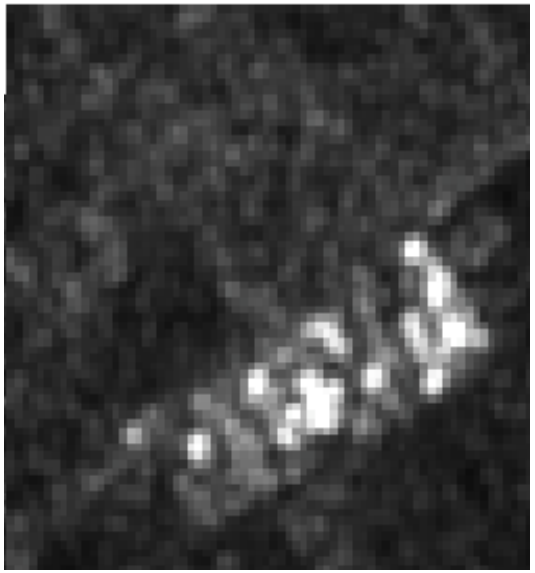
$\Sigma\, p \log p$ also appears in combinatorics, Shannon information, — — . It's fundamenta

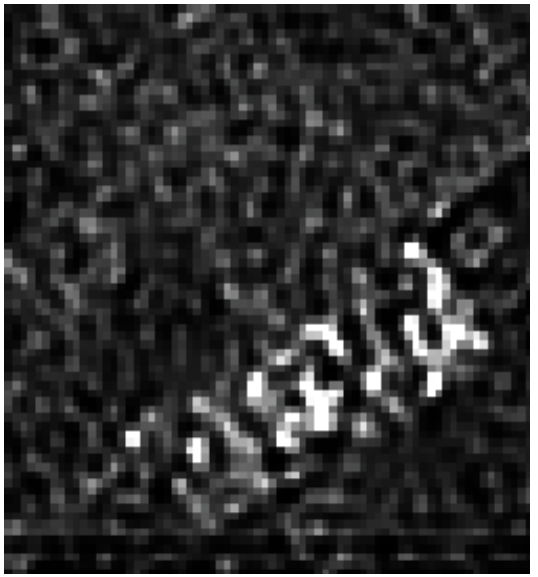
⑦, ⑧, ⑨, - - -

In complicated problems, use common sense.
Don't be too ignorant.

John Skilling

MaxEnt image enhancement





Méthodes entropiques

Divergence of Kullback-Leibler

$$D(p, q) = \int p(x) \log \frac{p(x)}{q(x)} dx \ ,$$

- **Akaike** Information Criteria
- minCross Entropy

Code lenghts of Shannon

$$L_c(x^n) = -\sum_{x_i} p(x_i) \log p(x_i) \ ,$$

- Minimum Description Length (MDL)
  - la complexité stochastique
  - ou la longueur de description en 2-parties

la **complexité algorithmique de Kolmogorov** (régularité d'une série particulière).

$K(x^n)$ = longueur minimal du programme à décoder par une machine pour reproduire la série x^n.

- le principe **MDL idéal** avec une longueur de description en 2-parties

Relations entre les méthodes entropiques et bayésiennes.

Divergence de Kullback-Leibler : extraction d'information a posteriori à partir d'a priori

MDL avec
-Complexité stochastique ⇔ évidence bayésienne
-Longueur de description en 2 parties ≈ évidence bayésienne

3) MDL idéal avec longueur de description en 2 parties ≠ évidence bayésienne

Problem formulation:

"Suppose source information yields a magnitude X and we are required to code the information obtained with precision [×]."

[×] **entropy** ⟶ the amount of **mutual information** of the magnitude X and another signal X' such that their joint probability density function is restricted to a certain family depending on the parameter [×] .

[×] ⟶ posterior probabilities of signal classes given a certain image from the archive

[×] and [×] ⟶ prior probabilities for signal classes and images

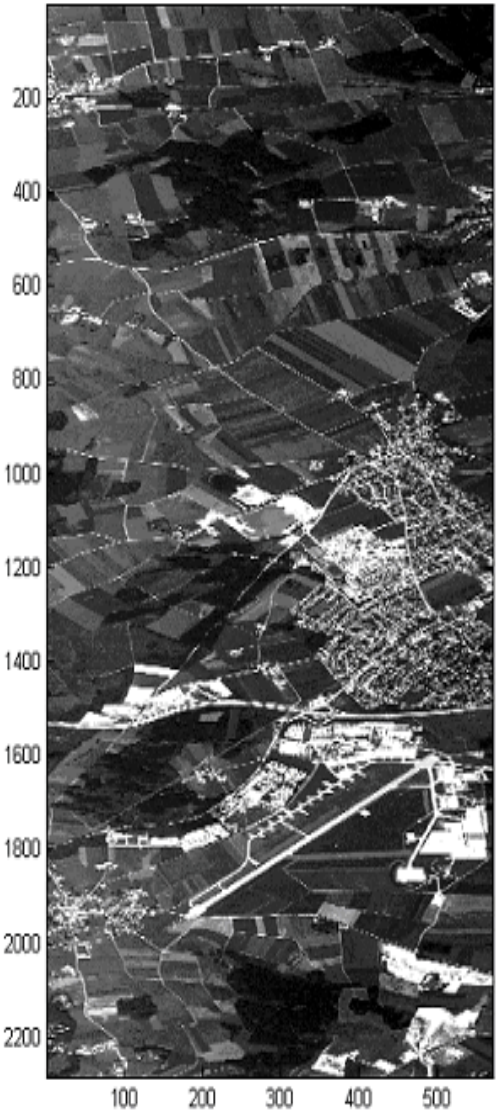
Rate distortion theory

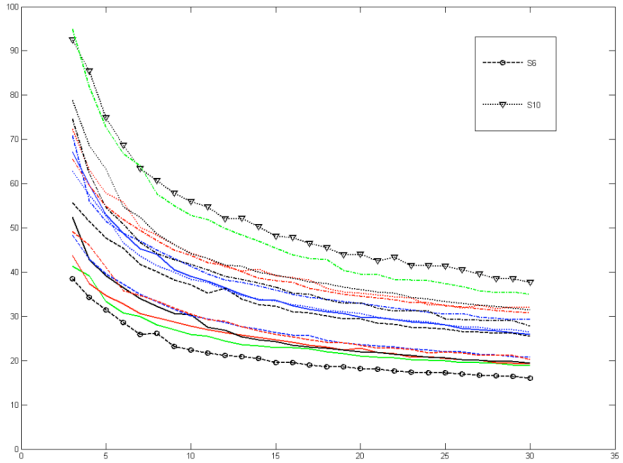
Continuous source output coding -> compression (quantization) ->loss of information.

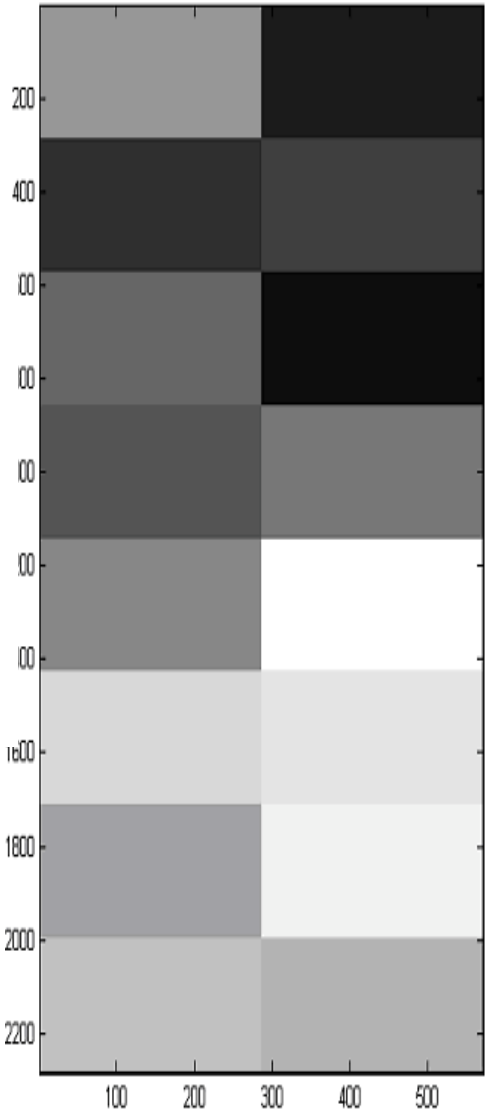
A finite set of codewords must be chosen to approximate the numbers of source symbols as well as possible.

Rate distortion function (R(D)) gives the minimum set of codewords for a given distortion.

Image complexity







DLR

Remote Sensing Technology Institute

INFORMATION BOTTLENECK PRINCIPLE

- emerged from Rate-Distortion theory.

- a formalism to express the trade-off between compression (short summary) and the relevant information contained in the summary.

$$\min_{p(\tilde{x}|x)} I(\tilde{X}, X) - \beta I(\tilde{X}, Y)$$

Information Bottleneck can be viewed as a Rate-Distortion problem based on KL divergence:

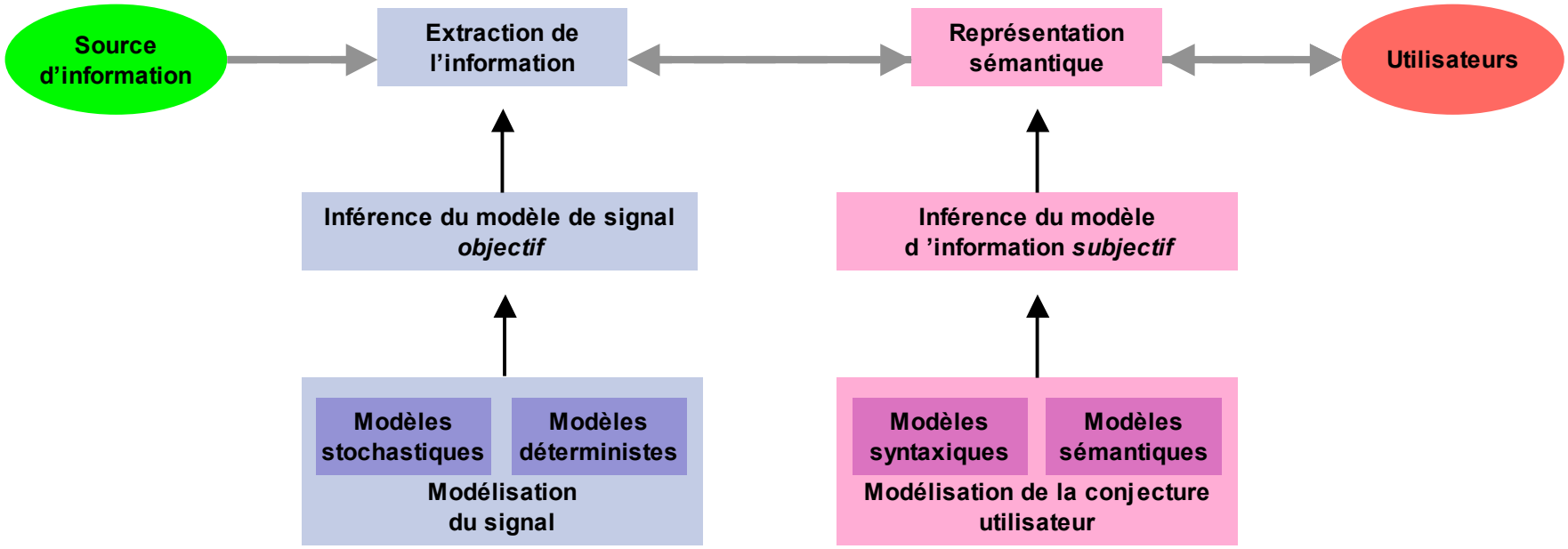
$$\min_{\tilde{\mathcal{Z}}_s, p(\tilde{z}|z)} I(Z, \tilde{Z}) - \beta E_{Z,\tilde{Z}} \left[d(Z, \tilde{Z}) \right]$$

$$d(z, \tilde{z}) = \sum_y p(y \mid x) \log \frac{p(y \mid x)}{p(y \mid \tilde{x})}$$

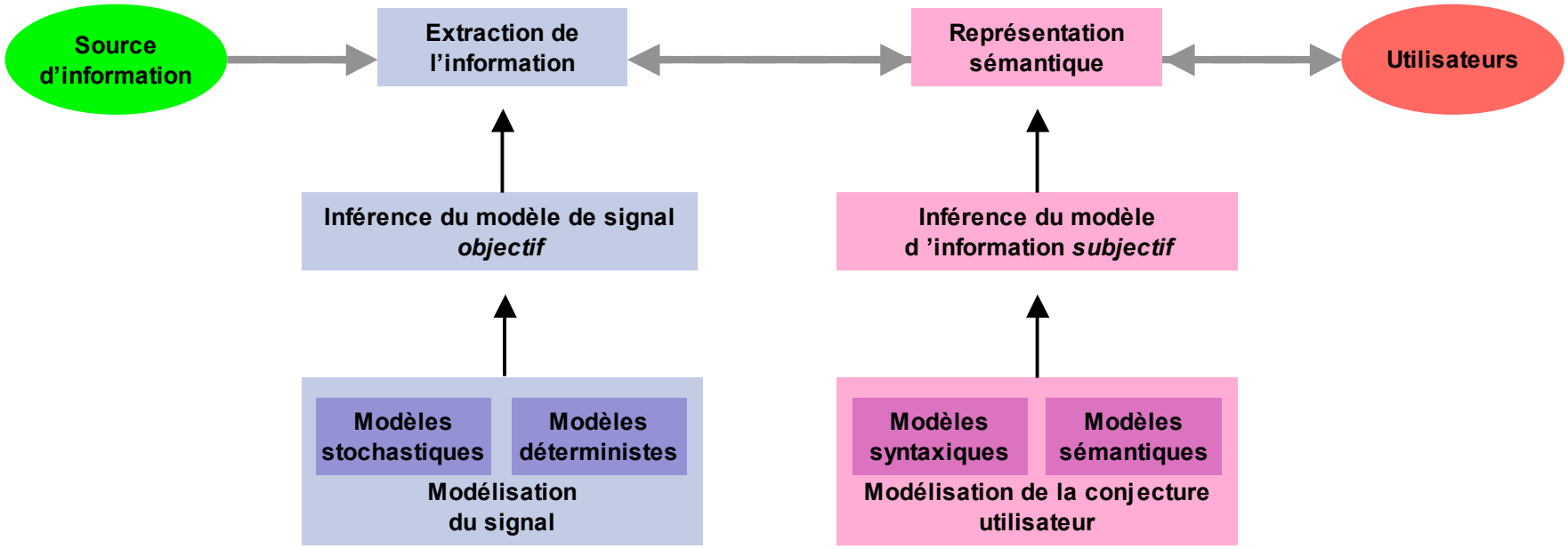
$$p(\tilde{z} \mid z) = \frac{p(\tilde{z})}{N(z, \beta)} e^{-\beta d(z, \tilde{z})}$$

$$N(z, \beta) = \sum_{\tilde{z}} p(\tilde{z}) e^{-\beta d(z, \tilde{z})}$$

CONCEPT DE COMMUNICATION AVANCE

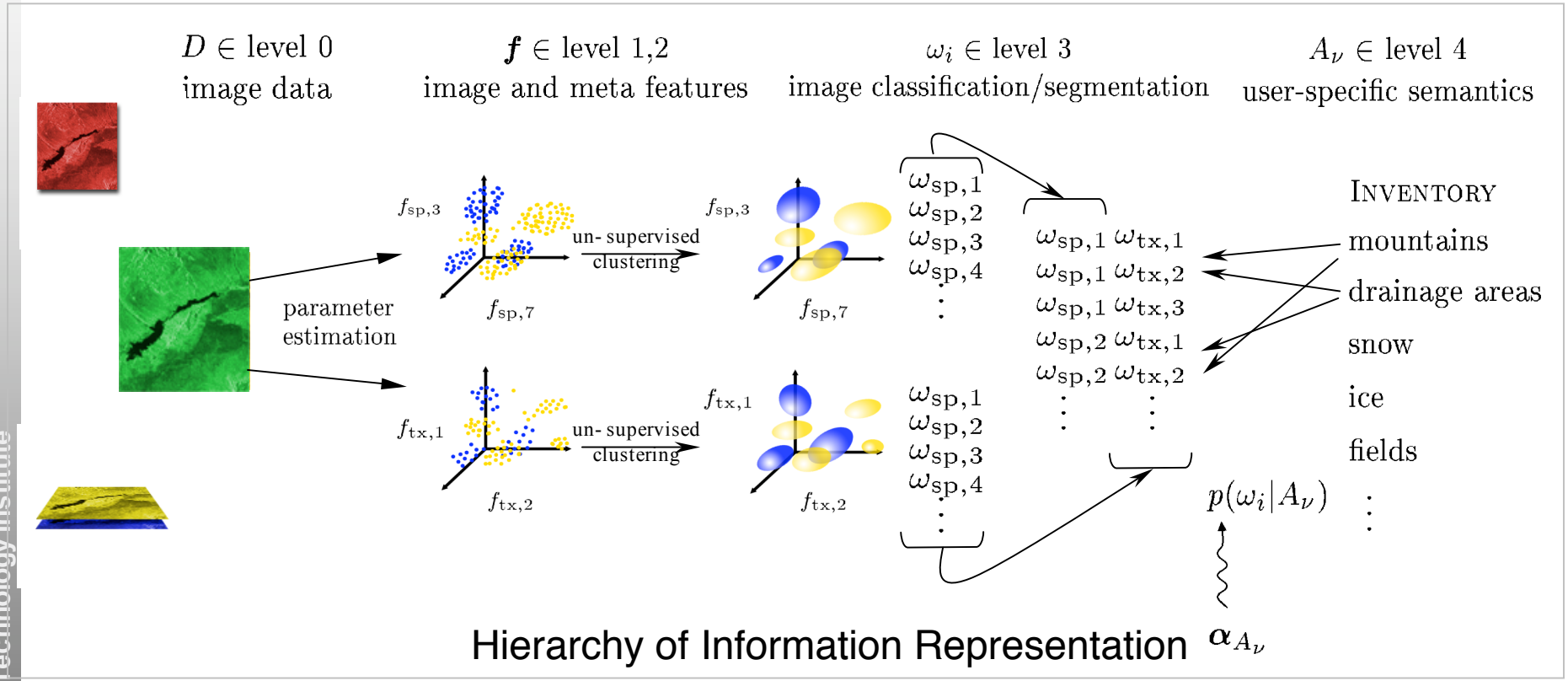


CONCEPT DE COMMUNICATION AVANCE



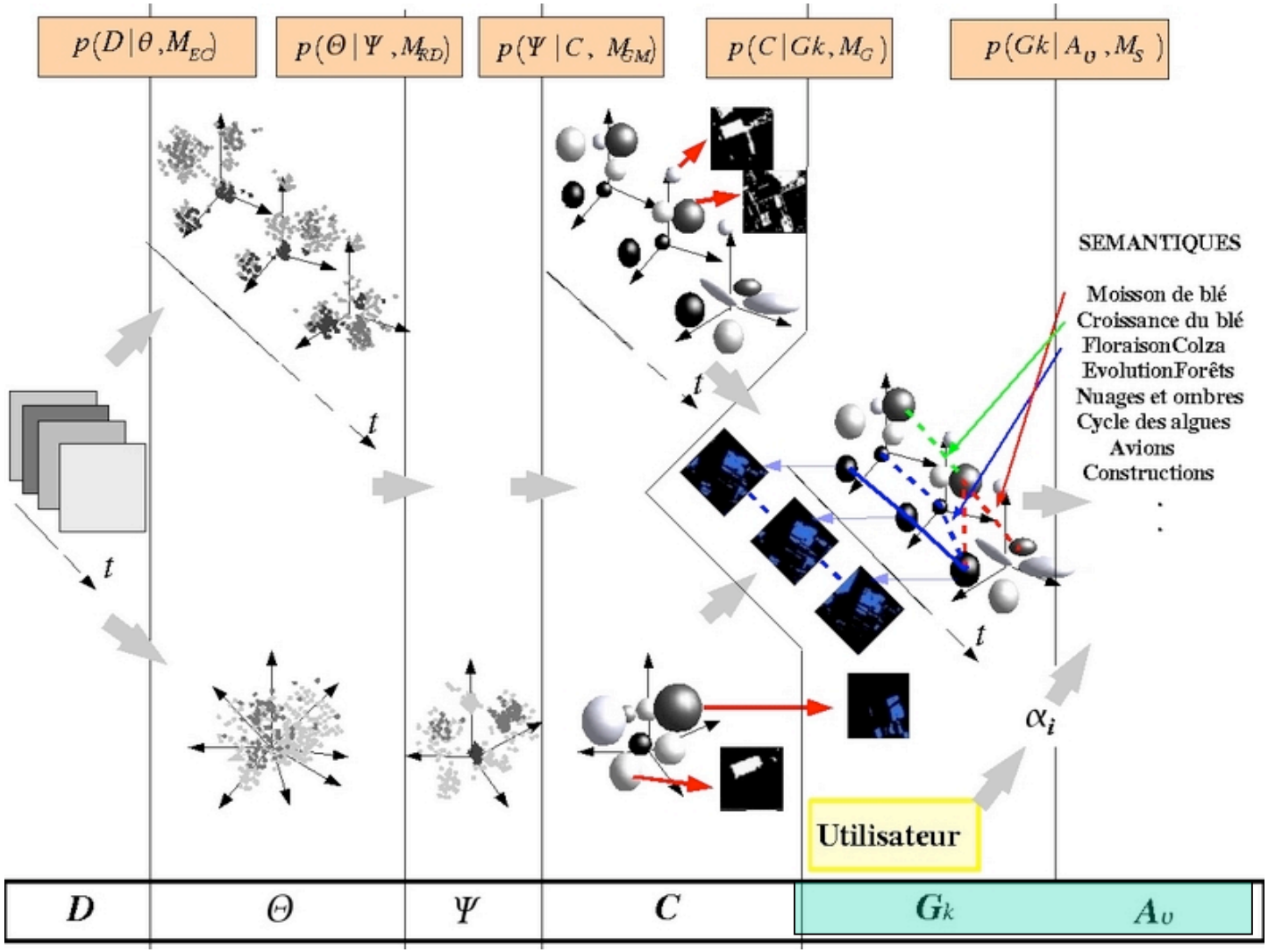
An observation, strictly, is only a sensation. … But as soon as we go beyond sensations we are making inferences. Jeffreys

Spatial data modelling



Hierarchy of Information Representation

Spatio-temporal data modelling



Coding

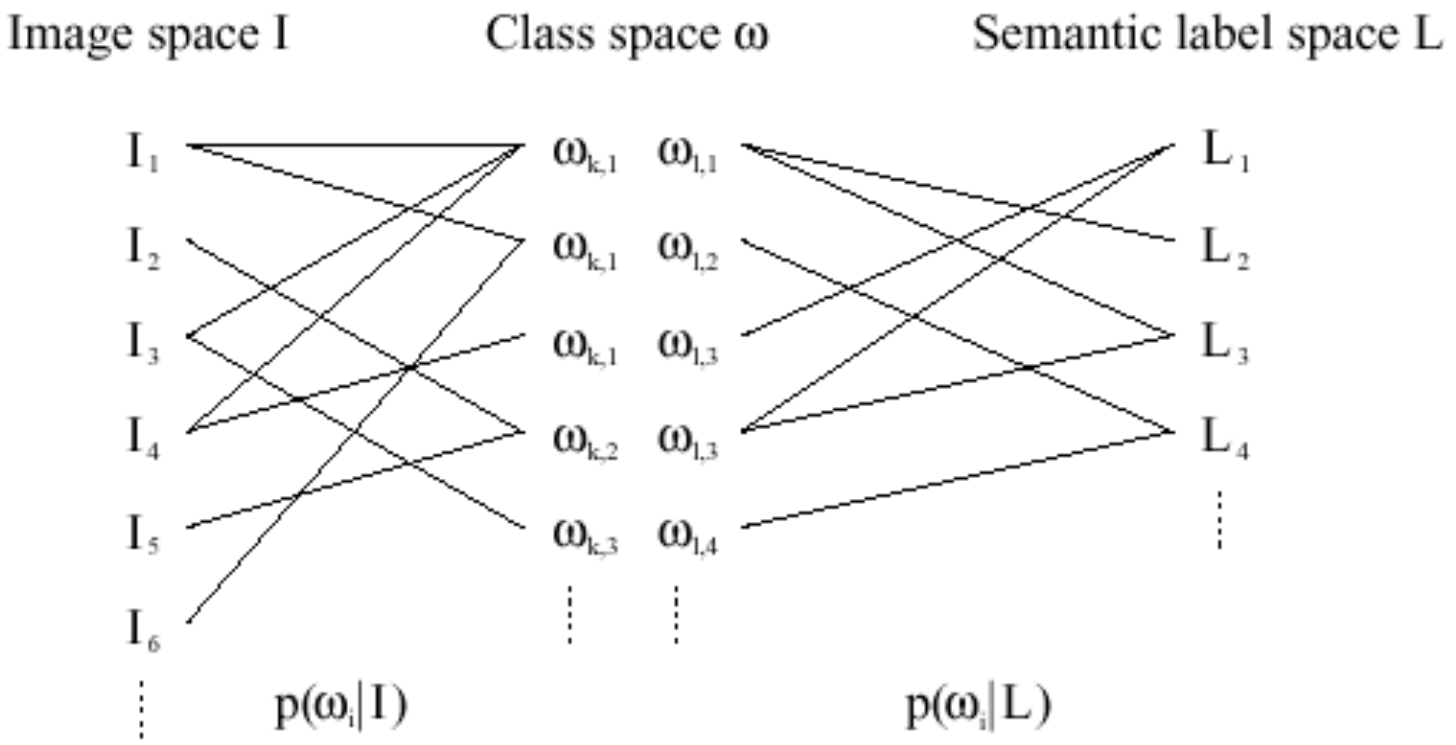
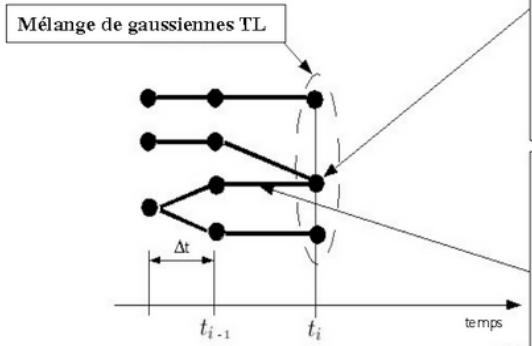


Image space I
Class space ω
Semantic label space L
I1
I2
I3
I4
I5
I6
ωk,1
ωk,1
ωk,1
ωk,2
ωk,3
ωl,1
ωl,2
ωl,3
ωl,3
ωl,4
L1
L2
L3
L4
p(ωi|I)
p(ωi|L)

MDL used to code object dynamic: spatio-temporal clustering

➢ Graph of trajectories



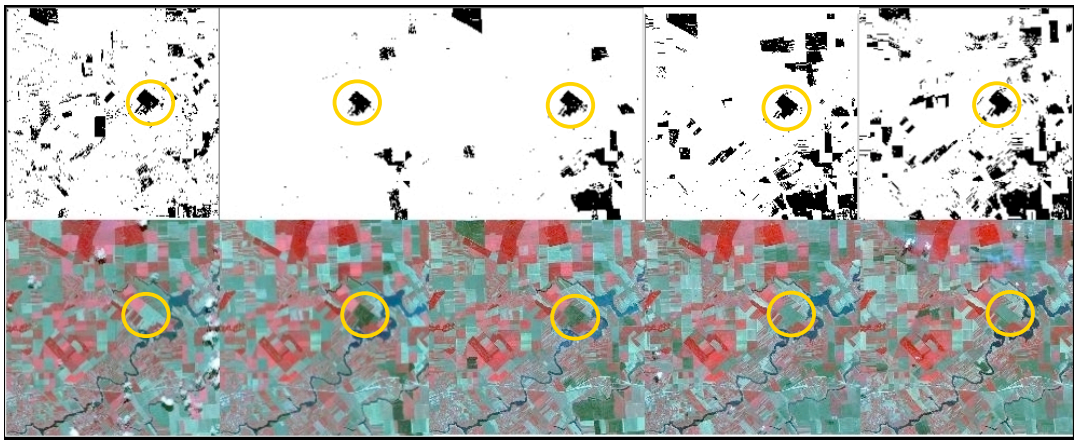
Noeud

- pour n'importe quel cluster MT :
  - paramètres du cluster TL
  - classeTL
- par cluster MT associé :
  - la divergence
  - la surface de recouvrement

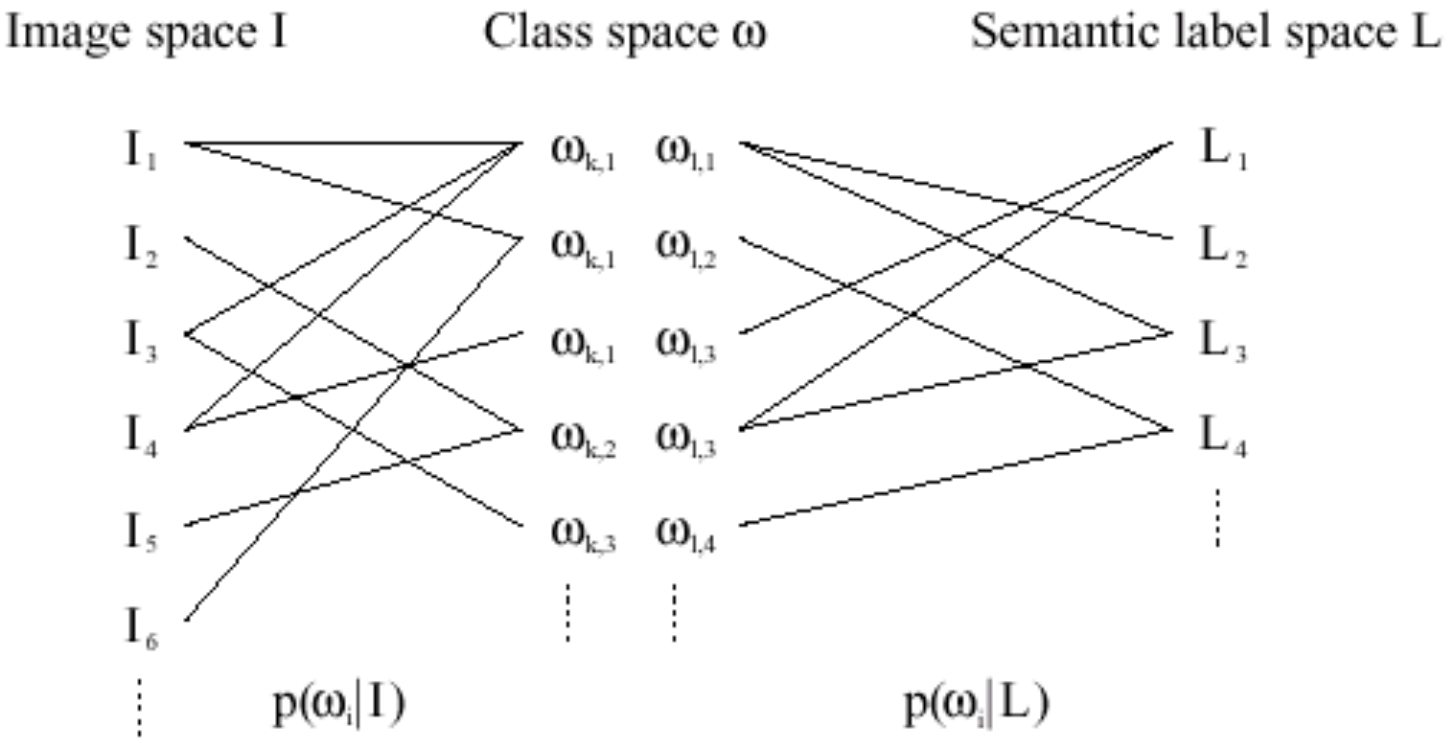
Arc

- pour n'importe quel cluster MT :
  - intervalle de temps
  - évolution du cluster TL
- par cluster MT associé :
  - le flux de pixels
  - l'information mutuelle
  - fusion & scission

➢ Dynamic classification



Clustering and coding



DIRICHLET MODEL

- **after N_i instances the likelihood is**

$$
\begin{aligned}
p(\boldsymbol{\theta}|T) &= \frac{p(T|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(T)} \\
&= \frac{\Gamma(r+N)}{\prod_i \Gamma(1+N_i)} \prod_i \theta_i^{N_i} \\
&= \mathrm{Dir}(\boldsymbol{\theta}|1+N_1, \ldots, 1+N_r)
\end{aligned}
$$

DIRICHLET MODEL

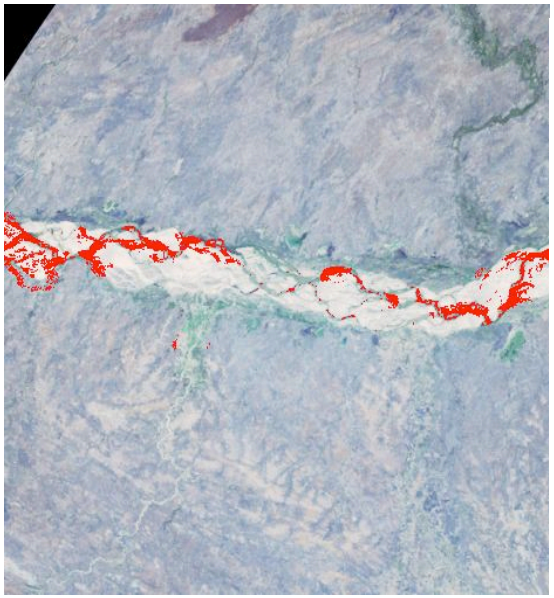
- **after a new training data set**

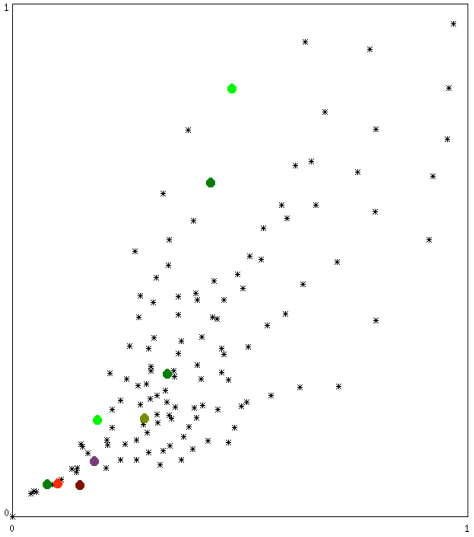
$$p(\boldsymbol{\theta}|T', T) = \frac{p(T'|\boldsymbol{\theta}, T)\, p(\boldsymbol{\theta}|T)}{p(T', T)}$$

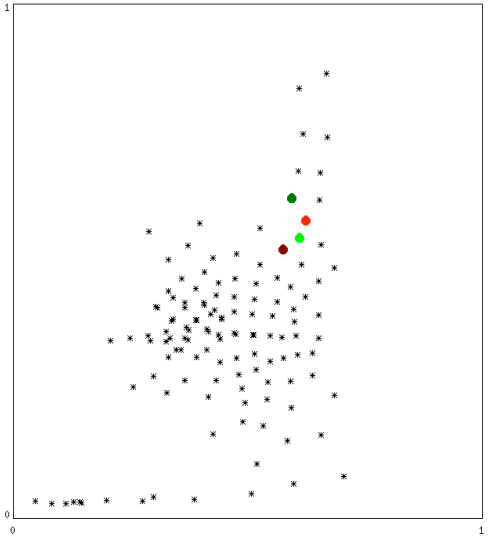
$$= \mathrm{Dir}(\boldsymbol{\theta}|\alpha_1 + N'_1, \ldots, \alpha_r + N'_r)$$

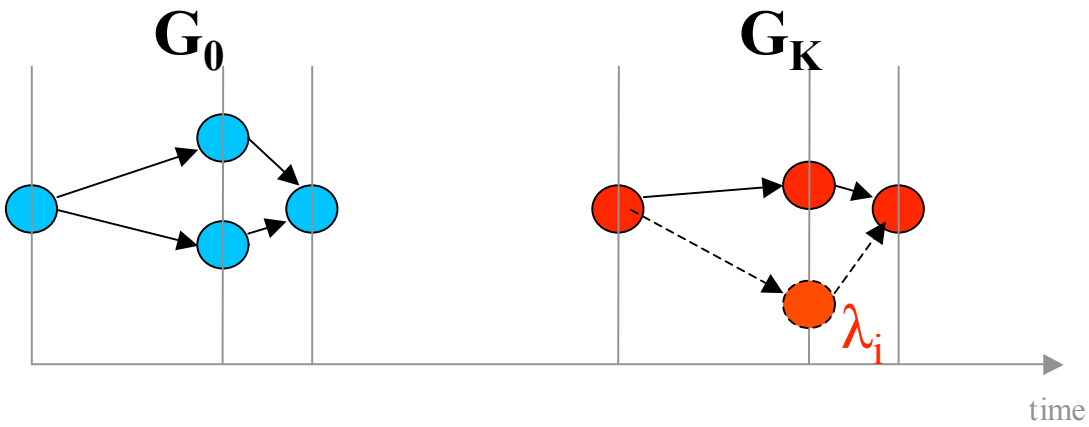
$$p(\omega_i|A_\nu, T) = \frac{\alpha_i}{\alpha}$$

Semantic coding: 3









- On définit une **transformation composée d'opérations élémentaires**

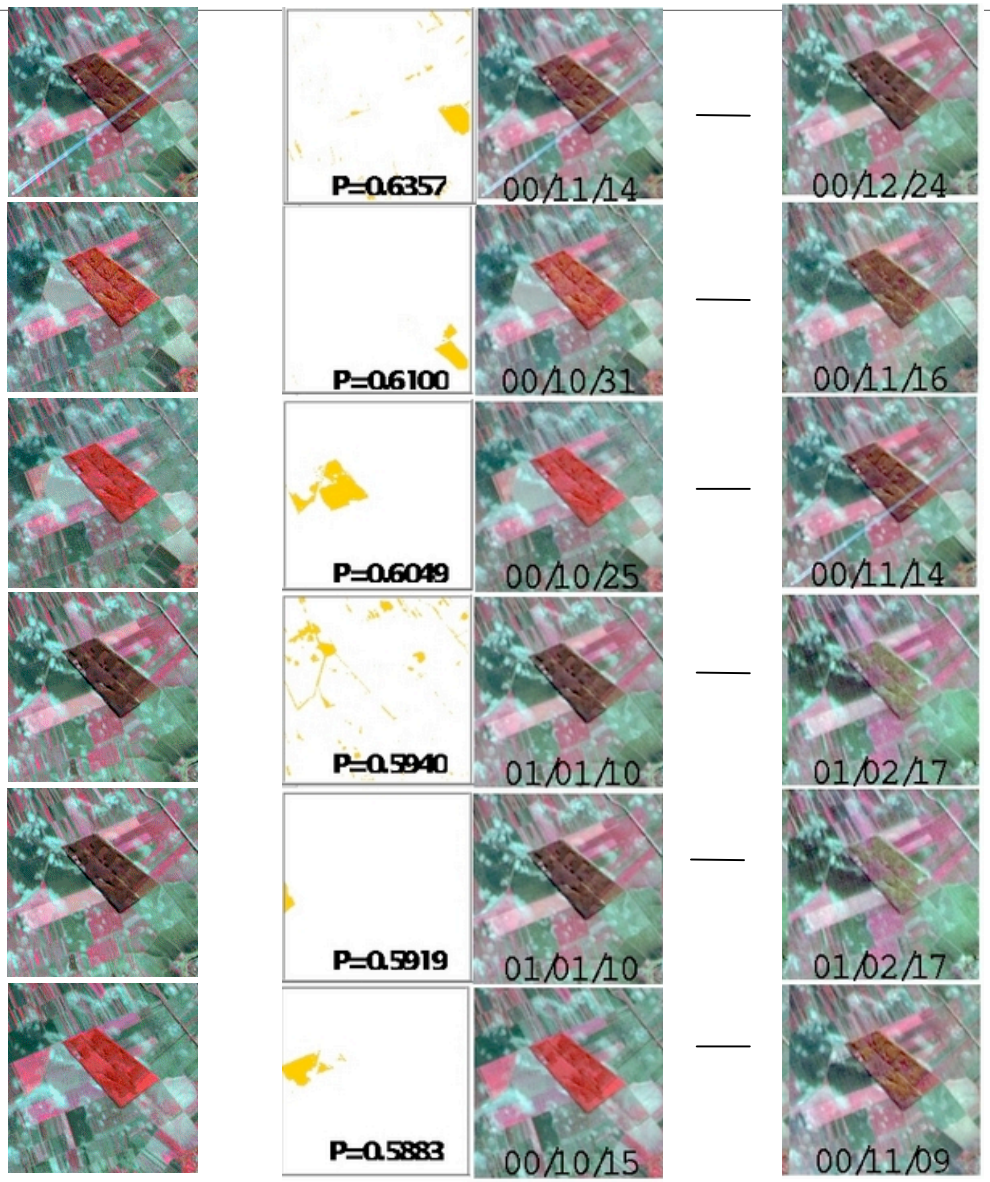
$$f : v^0 \rightarrow v^{k\lambda} = v^k \cup \lambda$$

de coût égale à une **somme pondérée de coût partiels** relatifs aux similarités entre les différents attributs de graphes (différence, divergence de Kullback-Leibler) :

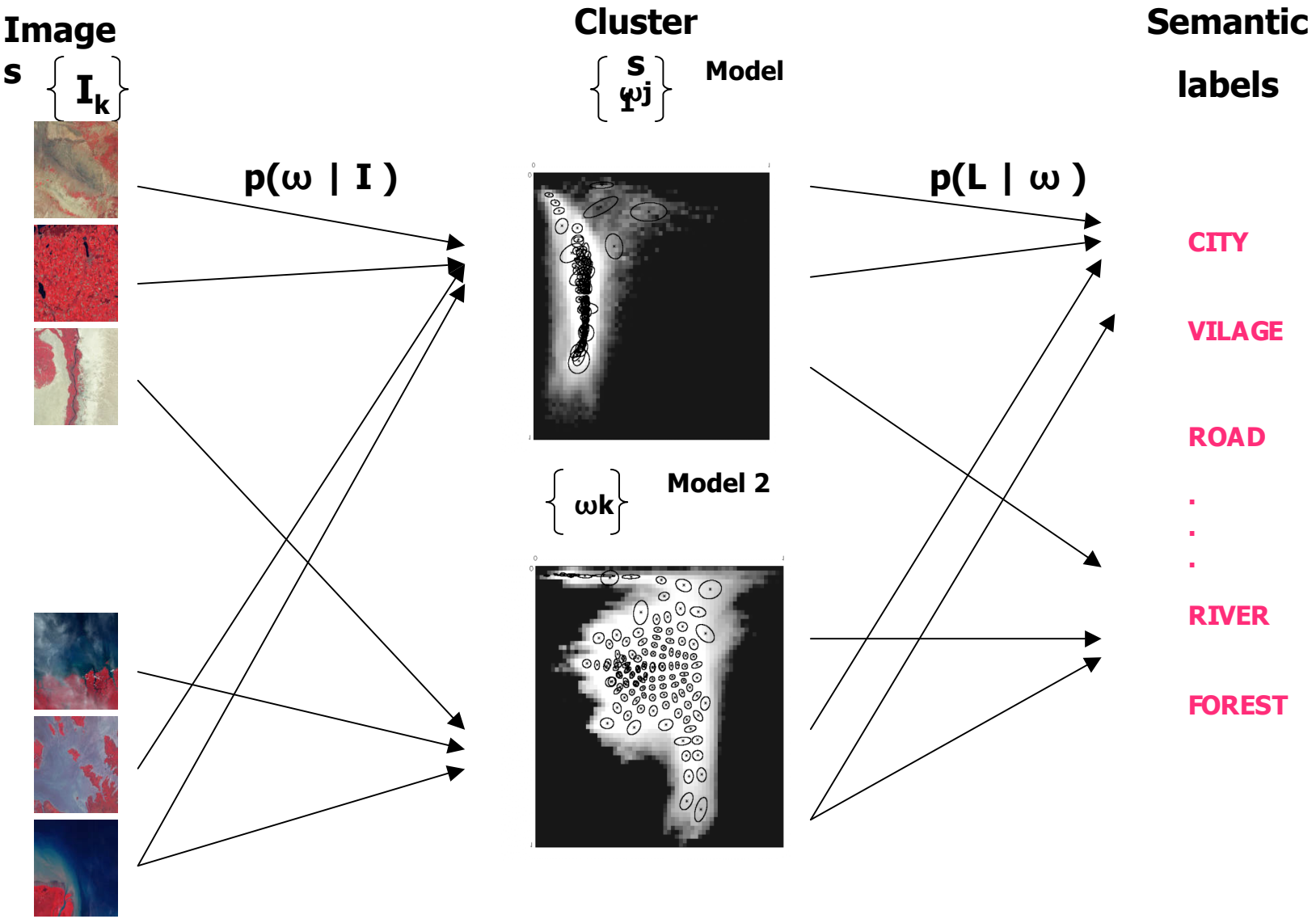
- **Apprentissage interactif de la distribution p(ω | T)**

a priori conjugué de **Dirichlet :** p(ω)

$$p(\omega|T^{(1)}) = \qquad = \mathrm{Dir}(\omega|1+N_1^{(1)}, \ \ldots \ ,1+N_r^{(1)}).$$


P=0.6357
00/11/14
00/12/24
P=0.6100
00/10/31
00/11/16
P=0.6049
00/10/25
00/11/14
P=0.5940
01/01/10
01/02/17
P=0.5919
01/01/10
01/02/17
P=0.5883
00/10/15
00/11/09

Advanced communication



Controlling the semantics

➢ **Problem**

In the system there may be

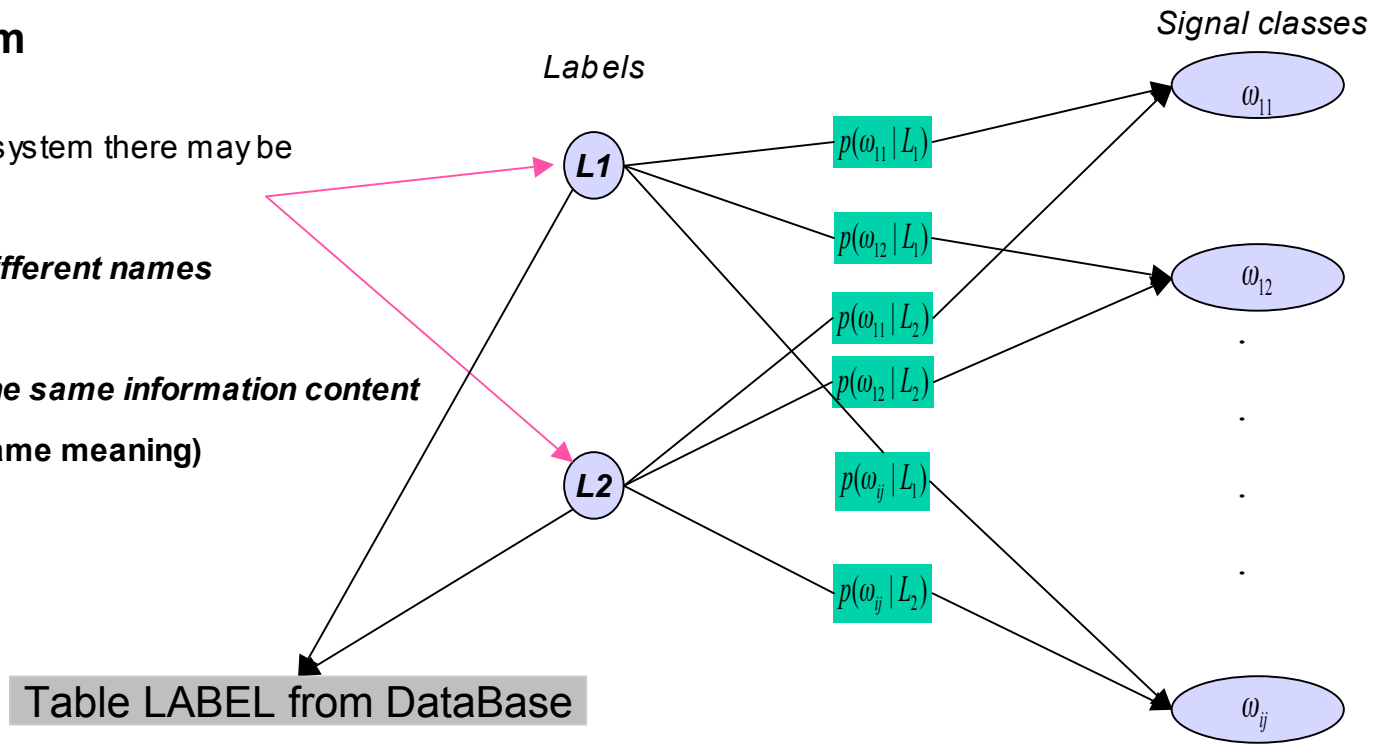
labels

with different names

and

with the same information content

(the same meaning)

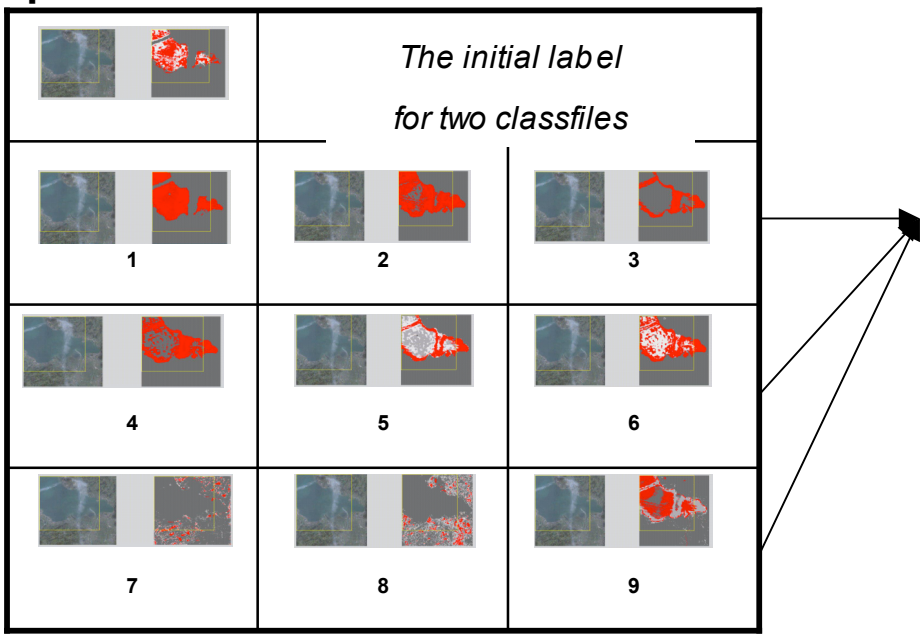


➢Solution : similarity measure , Kullback-Leibler divergence

$$KL(L_1 \mid L_2) = \sum_i \sum_j p(\omega_{ij} \mid L_1) * \log\left(\frac{p(\omega_{ij} \mid L_1)}{q(\omega_{ij} \mid L_2)}\right)$$

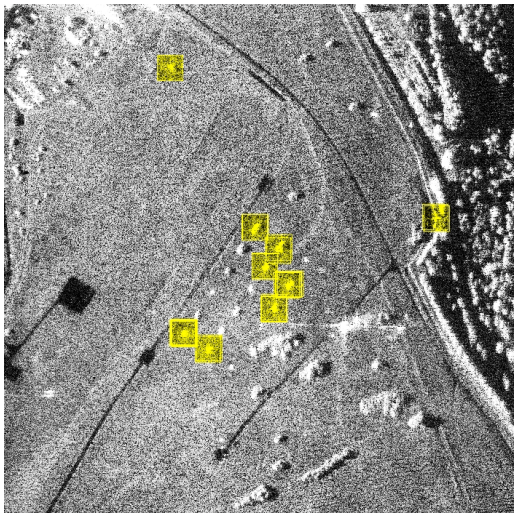
$$\begin{cases} p(\omega_{ij} \mid L_1) = p_1(\omega_i \mid L_1) * p_2(\omega_j \mid L_1) \\ q(\omega_{ij} \mid L_2) = q_1(\omega_i \mid L_2) * q_2(\omega_j \mid L_2) \end{cases}$$

➢Example

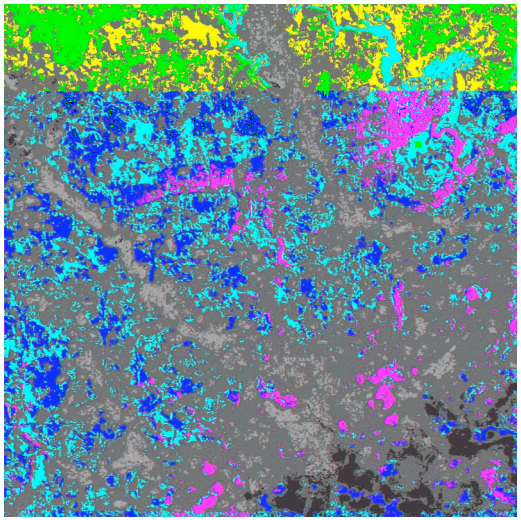


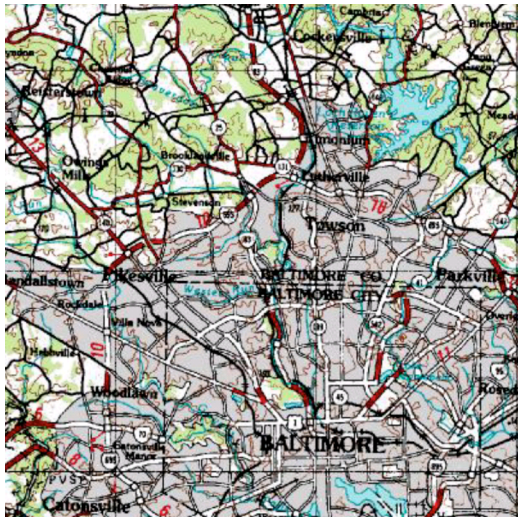
After Kullback-Leibler procedure the sorted labels list is : 2, 1, 6, 4, 5, 9, 7, 8. 3

Applications



Target discover and analysis
FGAN PolSAR





Mapping by fusion of SRTM DEM and SAR image

- dense urbanized
- medium urbanized
- light urbanized
- forest
- agricultural
- water

Target analysis