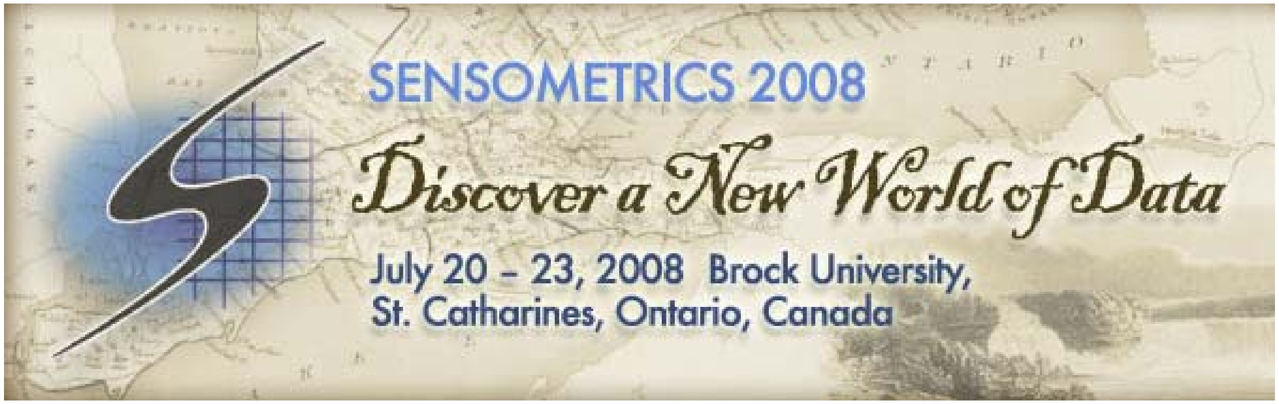
Studying consumer drivers with Bayesian Networks

Sensometrics meeting – St Catharines, Ontario – July 21st 2008


SENSOMETRICS 2008
Discover a New World of Data
July 20 – 23, 2008 Brock University,
St. Catharines, Ontario, Canada

BAYESIA
Your Decision Partner

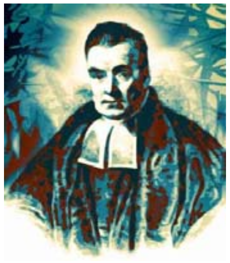
repères
passion for research

Bayesian Networks in a nutshell

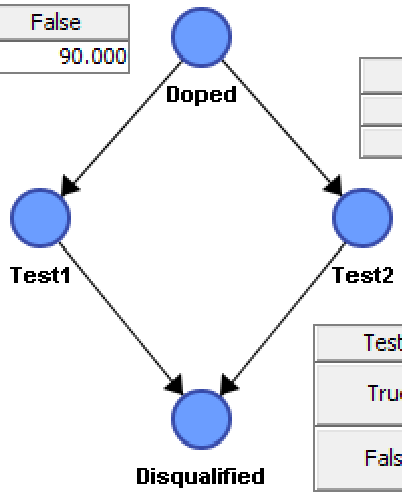
- ✓ A definition : a mathematical tool to model **PROBABILISTIC RELATIONS.**

- ✓ The basis **: BAYES THEOREM (1763)**

$$P(A|B) = P(A) * \frac{P(B|A)}{P(B)}$$



- ✓ Formalism : 2 distinctive parts ⇨ **GRAPH / PARAMETERS**



True	False
10.000	90.000

Doped	True	False
True	95.000	5.000
False	1.000	99.000

Doped	True	False
True	90.000	10.000
False	0.000	100.000

Test1	Test2	True	False
True	True	100.000	0.000
	False	100.000	0.000
False	True	100.000	0.000
	False	0.000	100.000

Real case study

Product testing survey

- ✓ Baby food tested amongst mothers
- ✓ 15 products tested
- ✓ Monadic blind test
- ✓ Standardized questionnaire
  - “LOOK” stage : mother handles the food before feeding her baby
  - “USE” stage : mother feeds her baby

What are the consumer drivers of liking ?
How do they relate to each other ?

Data presentation

- ✓ **1770 consumers**

- ✓ **17 variables**

  - Overall liking (score / 10)

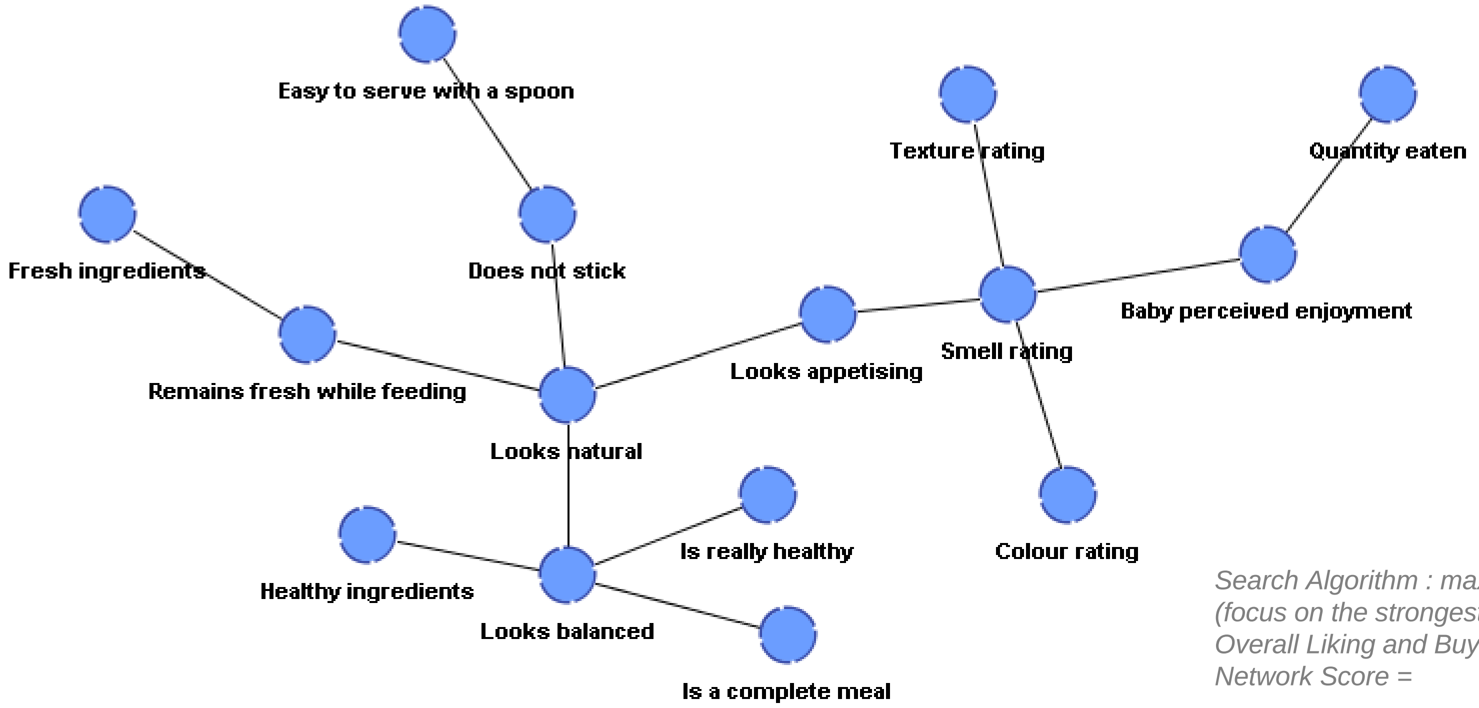
  - Consumer statements :

    - colour, texture, smell rating by the mother
    - perceived quantity eaten by the baby, did the baby enjoy the food ?
    - perceived benefits

Use this data to build a model explaining overall liking

Discovering relations between variables

Unsupervised learning



Search Algorithm : maximum spanning tree (focus on the strongest relations)
Overall Liking and Buying intention are let aside
Network Score =

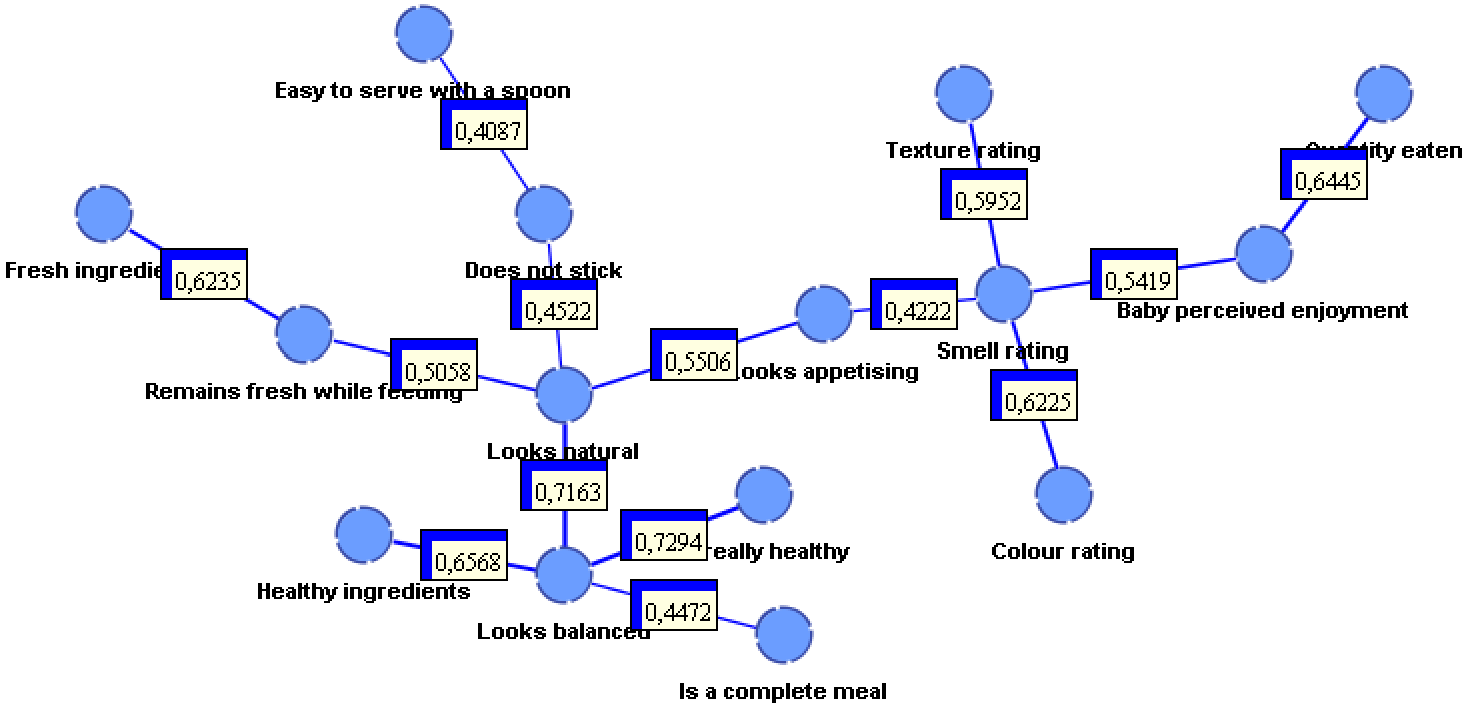
- ✓ **Heuristic Search Algorithm** to find the best representation of the joint probability distribution.

- ✓ **Minimum Description Length Score** to evaluate the quality of the network based on **fitness** and **compactness**.

MDL = DL(network) + DL (data | network)

Discovering relations between variables

Quantifying the probabilistic relations 1/2

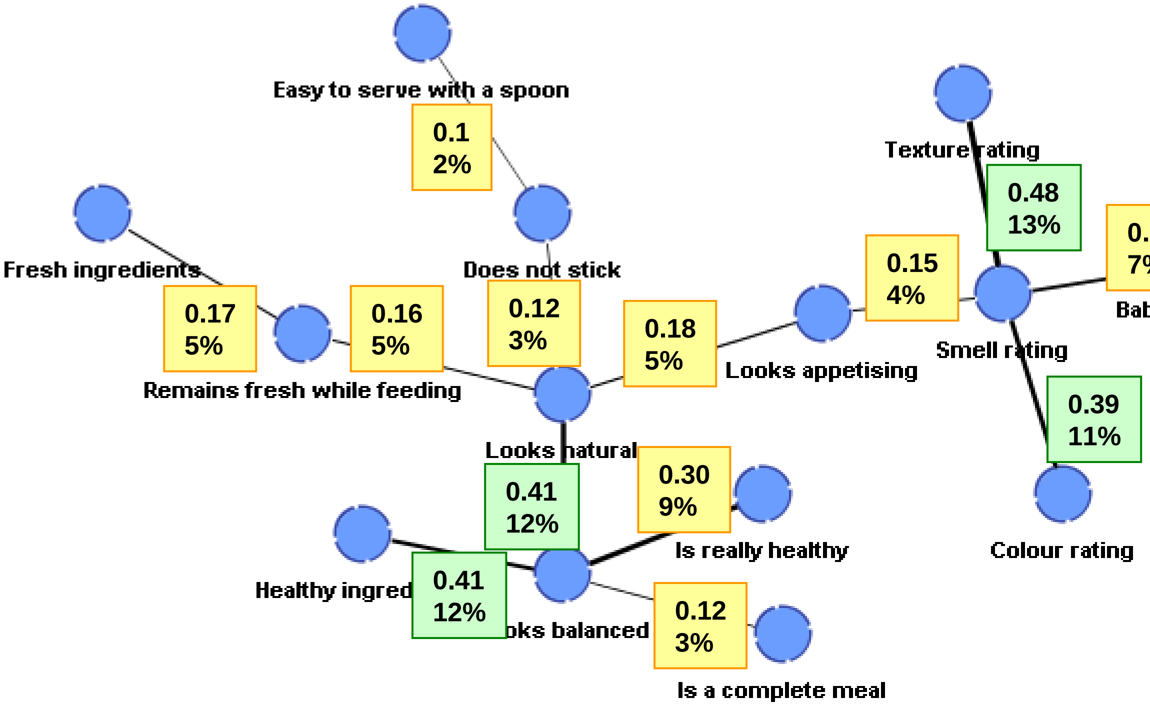


- ✓ Possible to compute the **Pearson Correlation Coefficient**

➡ Efficient in terms of **COMMUNICATION**

Discovering relations between variables

Quantifying the probabilistic relations 2/2



0.16 : K-L Divergence
6% : contribution of the relation to the network

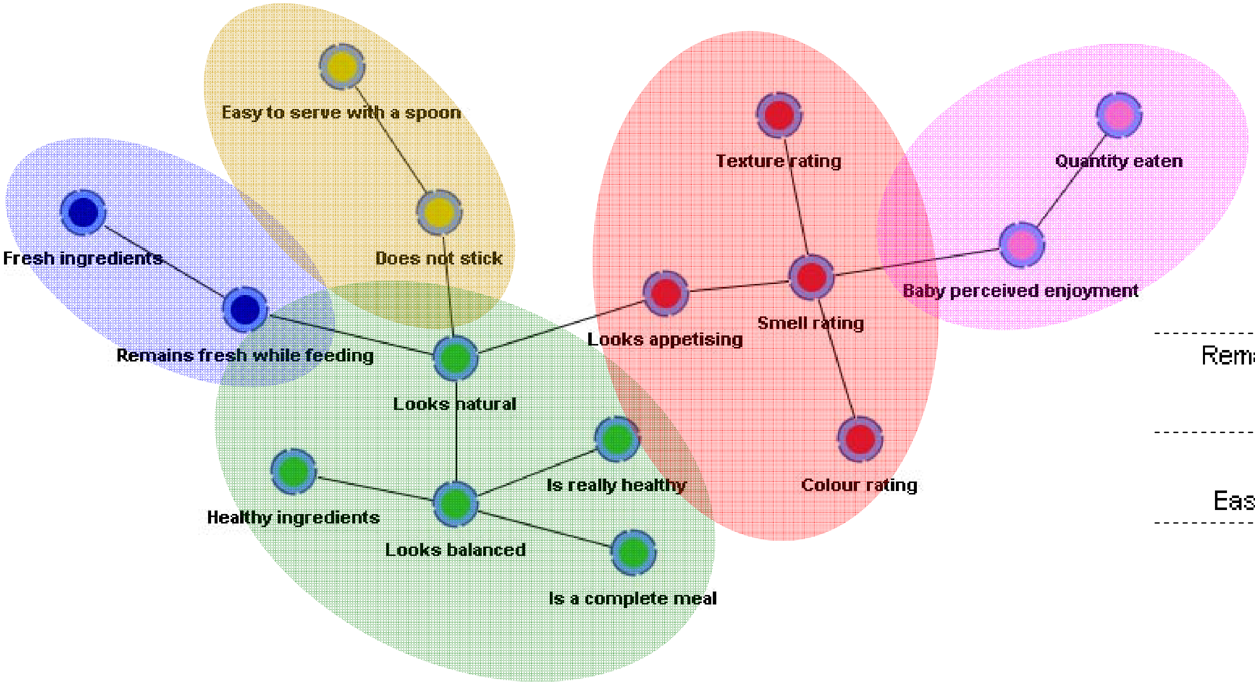
- ✓ More likely to use :
  Kullback Leibler divergence
  Non linear and **global** measure - Contribution of the relation to the network.

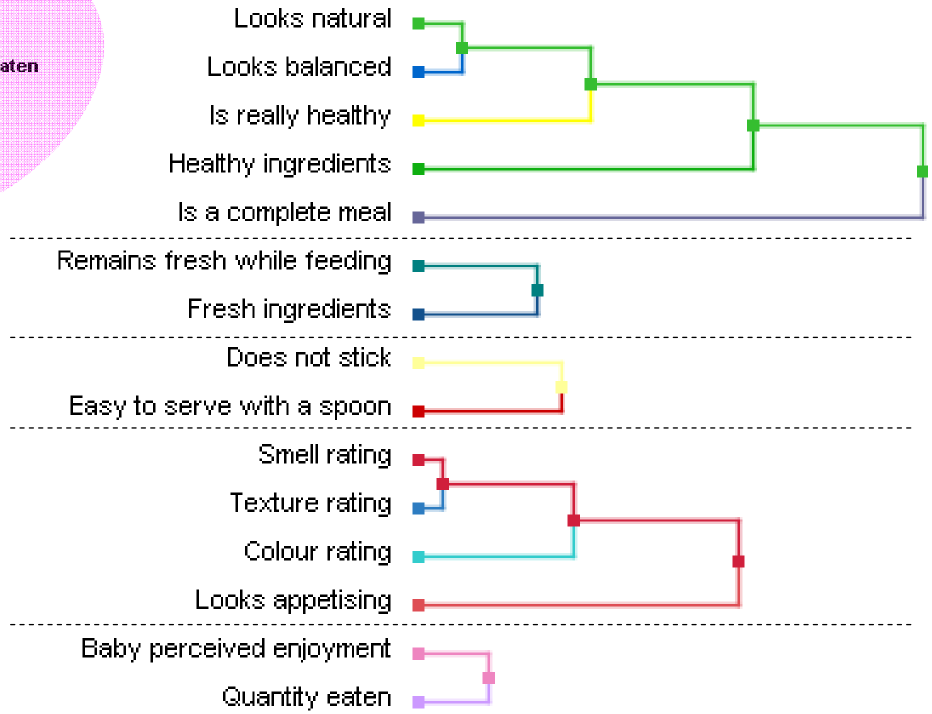
K-L Divergence for a probabilistic relation is a measure of the difference between :

- Joint probability distribution with the relation.
- And the joint probability distribution without the relation.

Summarizing information

Variable Clustering





Ascendant Hierarchical Clustering Results
5 groups automatically identified

- ✓ **Ascendant Hierarchical Clustering** based on Kullback Leibler measures.

- ✓ 5 groups of homogeneous variables have been identified : **5 “concepts”** that have to be seen as the main dimensions of a Factorial Analysis.

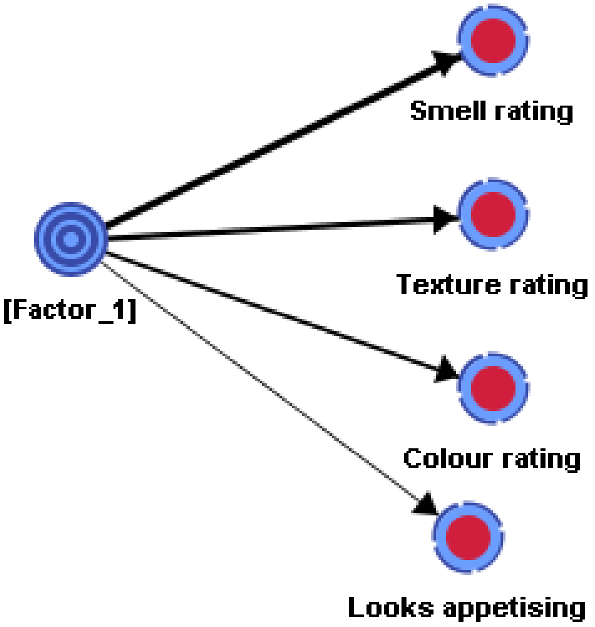
Summarizing information

Computing latent variables

FOR EACH CLUSTER :

✓ Introducing a new variable which is the hidden cause of the manifest variables.

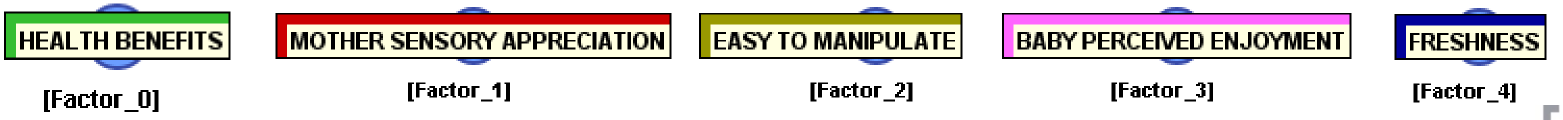
✓ Learning the probabilities with Expectation – Maximisation



Factor 1 summarizes **mother's sensory appreciation.**

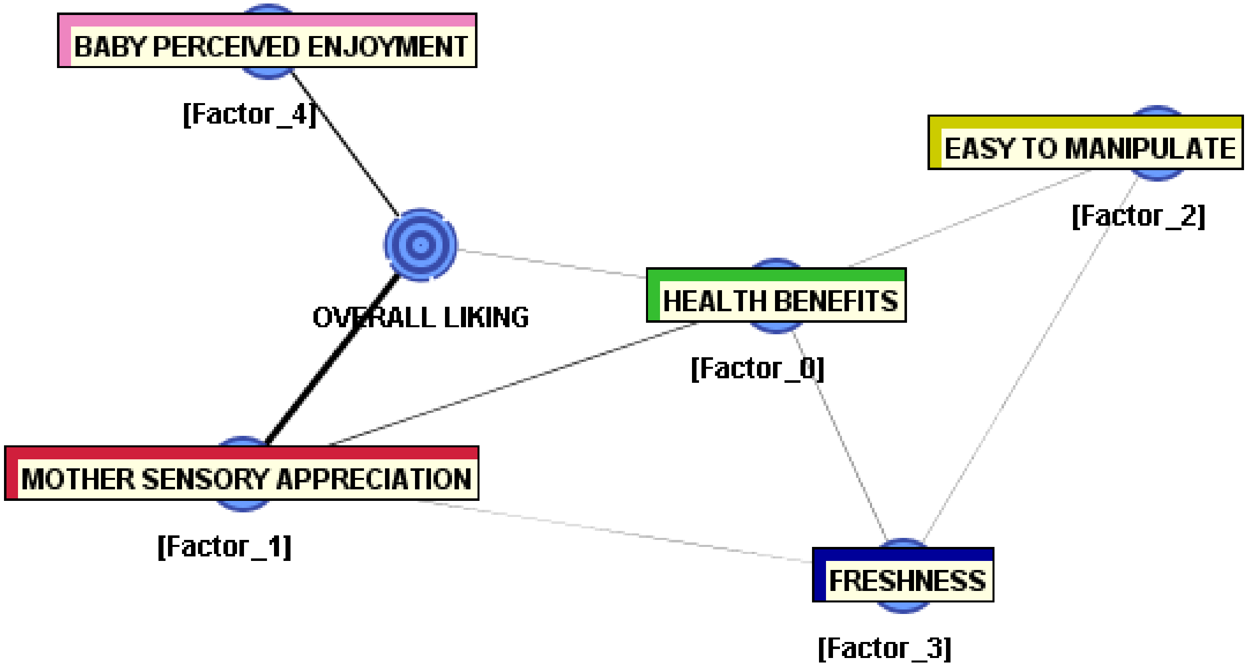
MOTHER SENSORY APPRECIATION
[Factor_1]

✓ **Each factor is then renamed by the analyst**



Modelling main dimensions and overall liking

✓ Modelling overall liking and latent variables with **automatic, unsupervised learning**



Search Algorithm : EQ
Latent variables and Overall Liking
Network Score = 8178

3 Dimensions have a direct impact on Overall Liking :

- Mother sensory perception

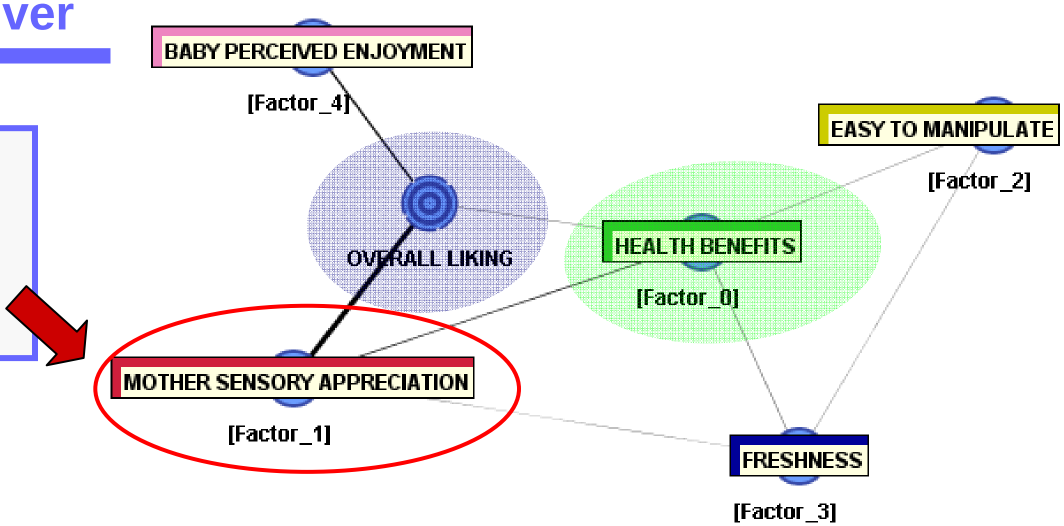
- Baby perceived enjoyment

- Perception of health benefits

Using the model…

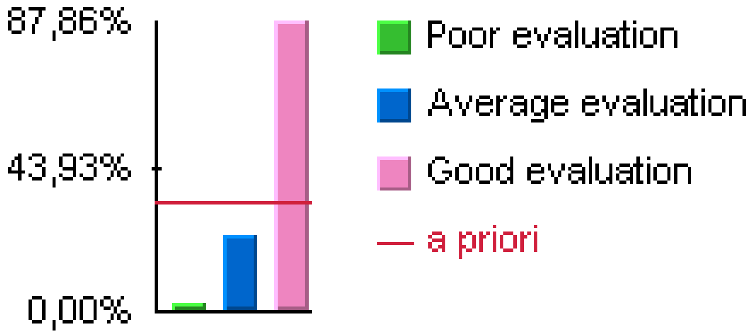
to understand the **precise role** of each driver

The experience of the product by the mother,
even before the baby eats the product,
will impact …



1. Overall Liking

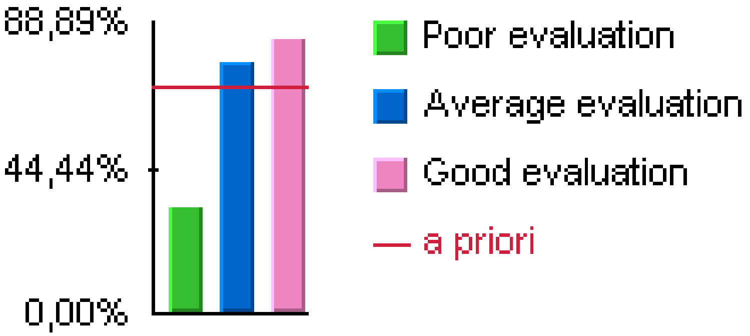
Probability that overall opinion >= 7



→ Mother sensory evaluation

2. Also perceived health benefits

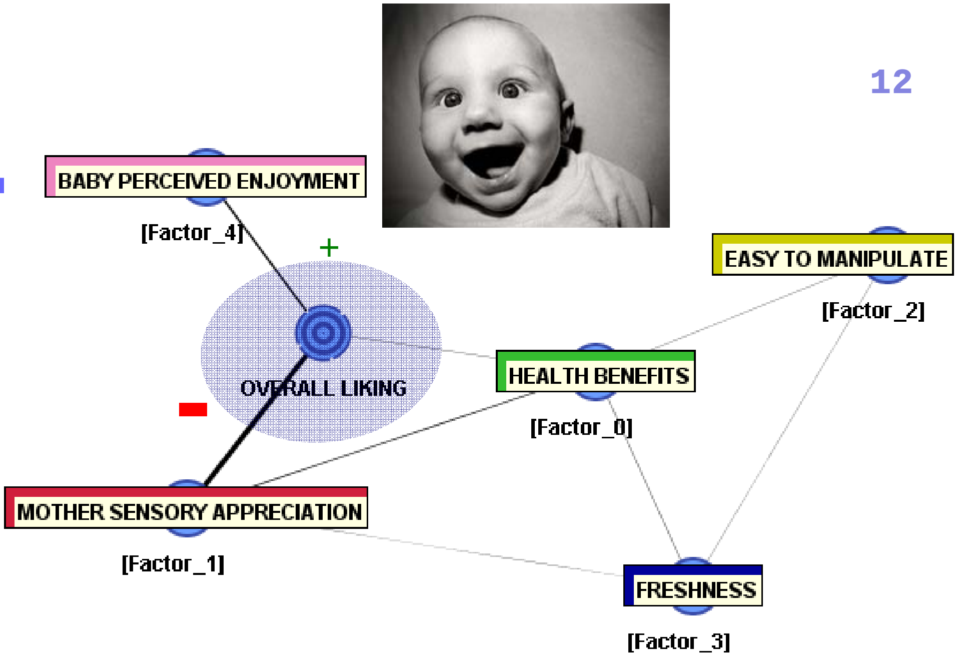
Probability that health benefits are perceived



→ Mother sensory evaluation

Using the model…

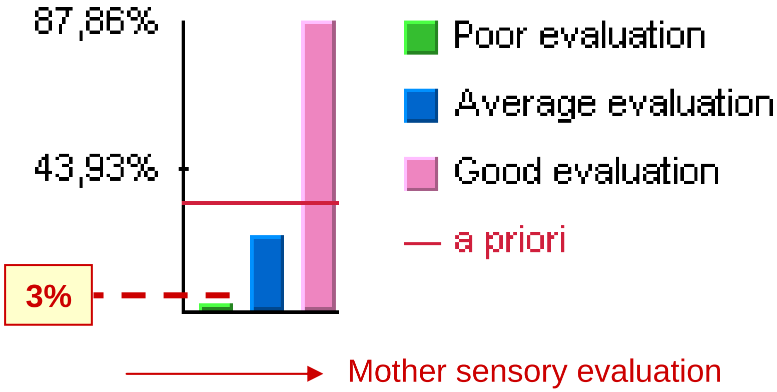
to understand the combination of drivers



Imagine the mother is not satisfied by the sensory properties. But what happens if the baby seems happy though ?

1. Mother NOT satisfied by the sensory properties

Probability that overall opinion >= 7



2. BUT the baby seems happy

Probability that overall opinion >= 7 **= 9% (+ 6 points only !)**

PERFORM LOOK STAGE AS A SCREENING PROCESS !

Using the model…

to predict product optimization benefits 1/2

✓ Imagine a product X which is **deficient** in terms of **sensory appreciation**, because of **colour and smell shortcomings**.

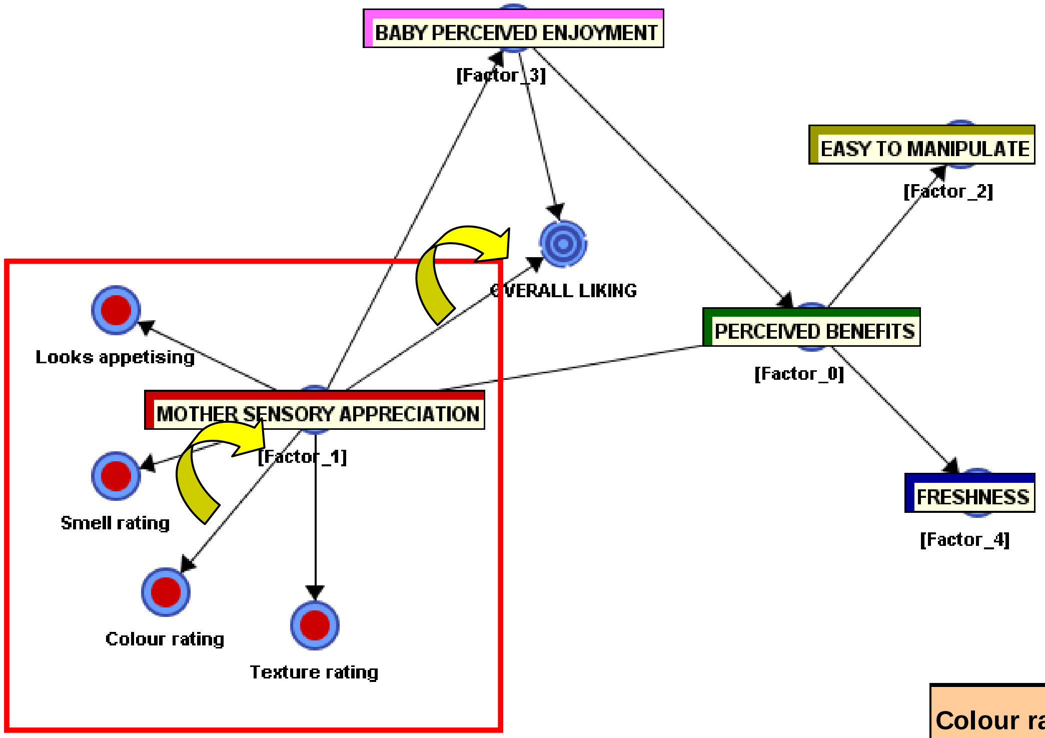
	Product X	Average of all products
Overall Liking probability that score >=7	**28%**	**34%**
Mother sensory appreciation probability that mother is satisifed	**22%**	**27%**
Colour rating probability that score >=7	**28%**	**34%**
Smell rating probability that score >=7	**27%**	**33%**
Texture rating probability that score >=7	**27%**	**31%**
Looks appetising probability of Total Agree	**73%**	**83%**

✓ **What would happen if colour was optimized ?**
Feasible optimization : reaching a satisfaction level on colour equal to products average.

Using the model…

to predict product optimization benefits 2/2

✓ **Getting back to manifest variables**, like in Structural Equation Modelling

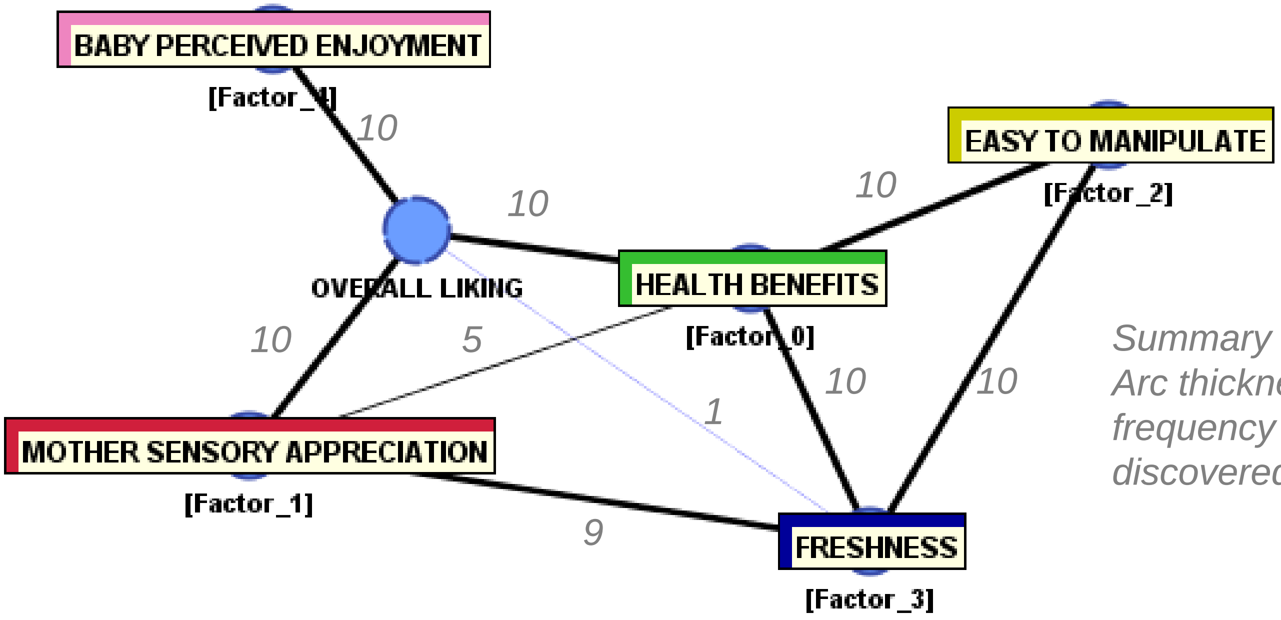


Effect of a reasonable colour optimization

	Product X	Reminder : before optimization
Colour rating probability that score >=7	34%	28%
Overall Liking probability that score >=7	32% ←	28%
Mother sensory appreciation probability that mother is satisifed	29% ←	22%

Model validation

- ✓ Structure validation : Jackknife method (10 times)



Summary of the 10 discovered structures. Arc thickness represents the relation's frequency : number of times the arc has been discovered in the 10 structures.

- ✓ Prediction validation : cross-validation using factor scores
  Global precision = 72,5%

- ✓ Going further : validating variable clustering

CONCLUSION

✓ **Good tool to UNDERSTAND and PREDICT (Diagnosis and Simulation)**

- **How consumer dimensions impact Liking**
- **How consumer dimensions relate to each other**
- **Product optimization effects**

✓ **SOUND and TRANSPARENT computations**

- **Everything relates to conditional probabilities**
- **Stable structures validated by Jackknife validation : no over fitting (conservative learning)**

✓ **Good COMMUNICATION tool**

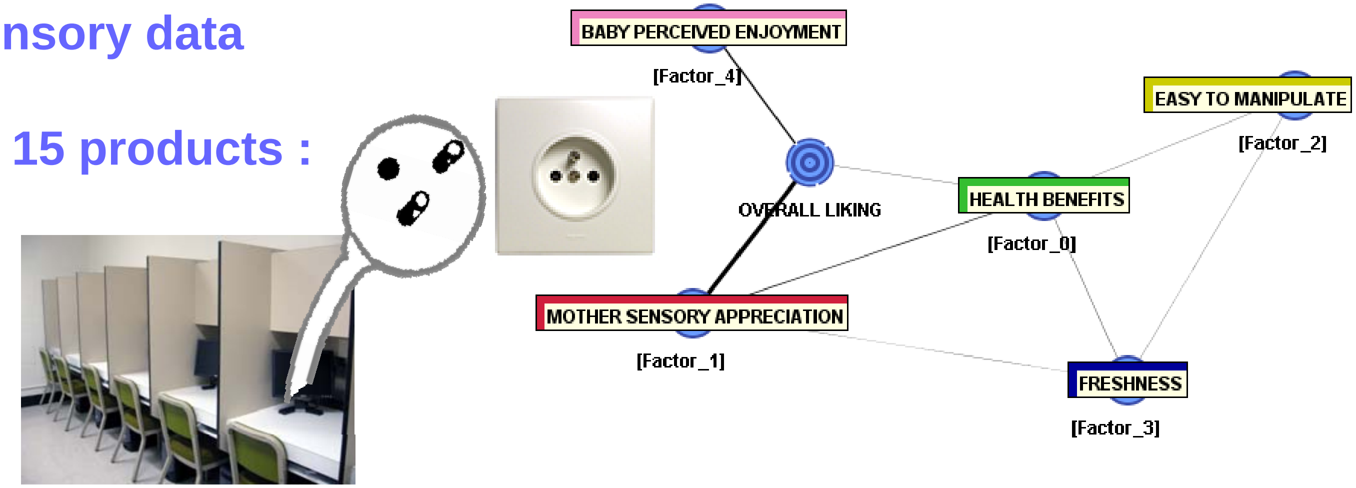
- **Graphical representation**
- **Probabilities are easy to understand**

CONCLUSION

- ✓ **To guarantee a RELEVANT model : MINIMUM requirements**
  - We recommend that at least 10 products have been tested
  - As representative of the market as possible
  - Following the same methodology

- ✓ **Going FURTHER**
  - Integrating sensory data
  - First test with 15 products : not enough ?

Sensory Data



THANK YOU FOR YOUR ATTENTION !

Jouffe Lionel
Managing Director

jouffe@bayesia.com

Craignou Fabien
Data Mining Department Manager

fcr@reperes.net