The Data Mining Methods Conceptual map

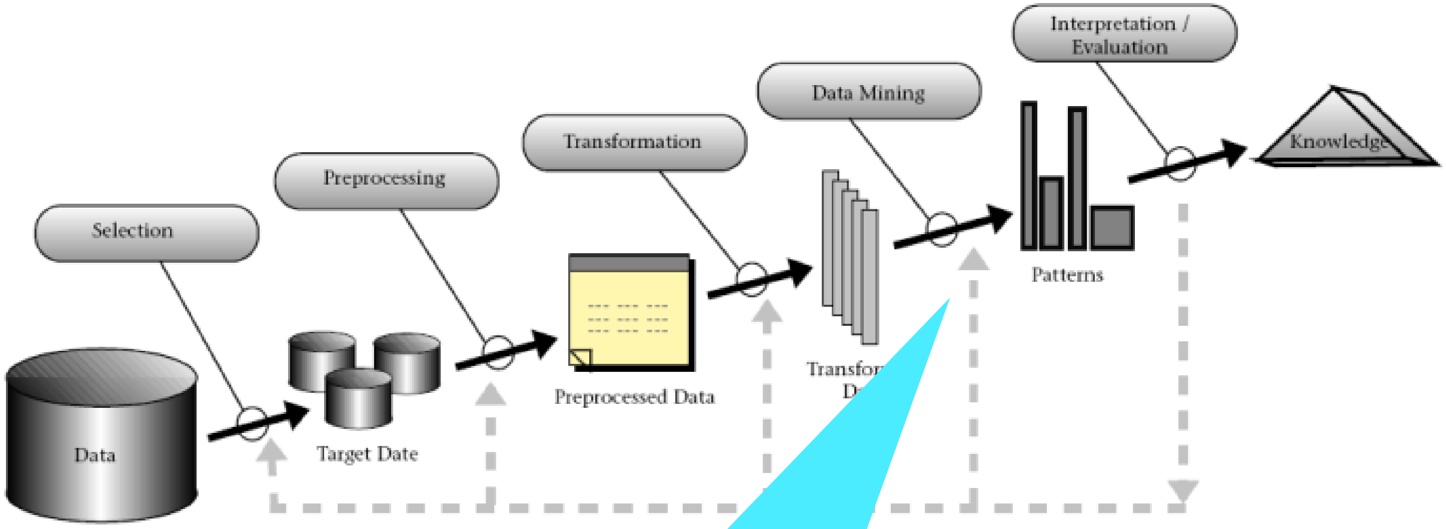
K. Gibert

Department of Statistics and Operation Research

Knowledge Engineering and Machine Learning group at
Intelligent Data Science and Artificial Intelligence Research Center
Science and Technology for Sustainability Research Institute
Universitat Politècnica de Catalunya, Barcelona
karina.gibert@upc.edu
http://www.eio.upc.edu/en/homepages/karina

Data Mining and Knowledge Discovery

- Knowledge Discovery System [Fayy96]:



DM method choice

- Critical
  - Suitable tools increase
  - Methodological expertise required

Choosing DM-technique *[iEMSs10][EMSO18]*

- Computational DM systems: Catalogues with many options
- Abbundant Literature
- Choices grouped under different criteria:
  - Technical proximity *[Gibert et alt, 2008]*
  - Research area, …
- Expert decision criteria (EDC):
  - Goal of problem to be solved
  - Structure of available data set
  - Method’s properties
  - Future use of the resulting model

Evidenced by real experiences

- Conceptual map of DM methods based on EDC *[Gibert et alt, 2015]*
  - Modelling decision process itself
  - Good-practice guidelines for non-expert users
  - Building intelligent recomenders

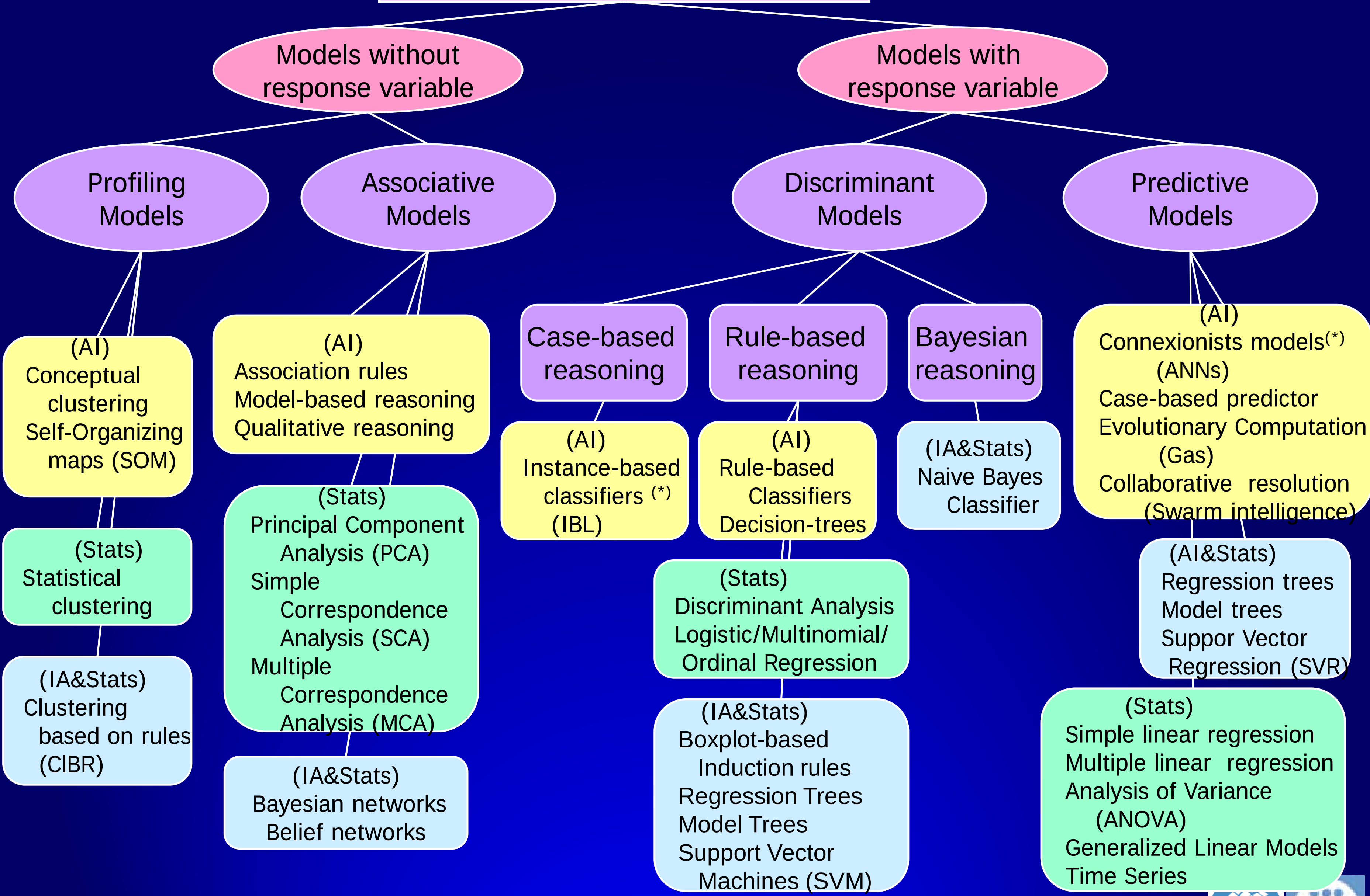
Towards Integral KDD systems construction

Gibert K, J. Spate, M. Sànchez-Marrè, I. Athanasiadis, J. Comas (2008): Data Mining for Environmental Systems. In Environmental Modeling, Software and Decision Support. State of the art and New Perspectives. IDEA Series v3 (Jackeman, A. J., Voinov, A., Rizzoli, A., and Chen, S. eds), pp 205-228. Elsevier NL.

Gibert K, M. Sànchez-Marrè (2015) Improving ontological knowledge with reinforcement in recommending the data mining method for real problems. In Proceedings of Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA) 2015:769-778, Albacete, nov 2015

Knowledge Models
Is there a response variable?

[Gibert et alt 2010, iEMSS]
[Gibert et alt 2010b, CAEPIA]



Models without response variable

Profiling Models

(AI)
Conceptual clustering
Self-Organizing maps (SOM)

(Stats)
Statistical clustering

(IA&Stats)
Clustering based on rules (CIBR)

Associative Models

(AI)
Association rules
Model-based reasoning
Qualitative reasoning

(Stats)
Principal Component Analysis (PCA)
Simple Correspondence Analysis (SCA)
Multiple Correspondence Analysis (MCA)

(IA&Stats)
Bayesian networks
Belief networks

Models with response variable

Discriminant Models

Case-based reasoning

(AI)
Instance-based classifiers [(*)]
(IBL)

Rule-based reasoning

(AI)
Rule-based Classifiers
Decision-trees

(Stats)
Discriminant Analysis
Logistic/Multinomial/ Ordinal Regression

(IA&Stats)
Boxplot-based Induction rules
Regression Trees
Model Trees
Support Vector Machines (SVM)

Bayesian reasoning

(IA&Stats)
Naive Bayes Classifier

Predictive Models

(AI)
Connexionists models[(*)]
(ANNs)
Case-based predictor
Evolutionary Computation (Gas)
Collaborative resolution (Swarm intelligence)

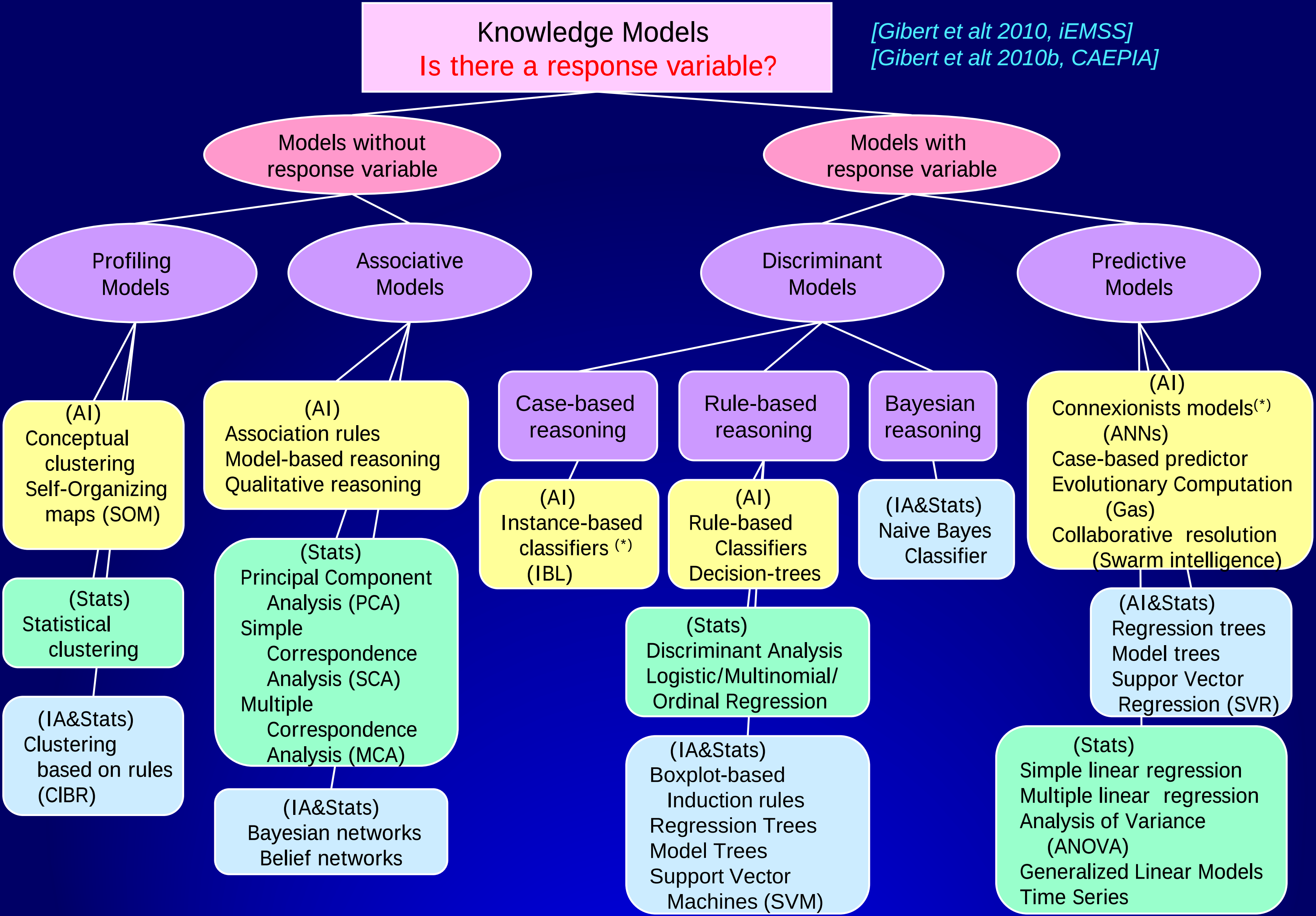
(AI&Stats)
Regression trees
Model trees
Suppor Vector Regression (SVR)

(Stats)
Simple linear regression
Multiple linear regression
Analysis of Variance (ANOVA)
Generalized Linear Models
Time Series

© K. Gibert

IDEAI

UPC



Knowledge Models
Is there a response variable?

[Gibert et alt 2010, iEMSS]
[Gibert et alt 2010b, CAEPIA]

Models without response variable

Profiling Models

(AI)
Conceptual clustering
Self-Organizing maps (SOM)

(Stats)
Statistical clustering

(IA&Stats)
Clustering based on rules (ClBR)

Associative Models

(AI)
Association rules
Model-based reasoning
Qualitative reasoning

(Stats)
Principal Component Analysis (PCA)
Simple Correspondence Analysis (SCA)
Multiple Correspondence Analysis (MCA)

(IA&Stats)
Bayesian networks
Belief networks

Models with response variable

Discriminant Models

Case-based reasoning

(AI)
Instance-based classifiers (*)
(IBL)

Rule-based reasoning

(AI)
Rule-based Classifiers
Decision-trees

(Stats)
Discriminant Analysis
Logistic/Multinomial/ Ordinal Regression

(IA&Stats)
Boxplot-based Induction rules
Regression Trees
Model Trees
Support Vector Machines (SVM)

Bayesian reasoning

(IA&Stats)
Naive Bayes Classifier

Predictive Models

(AI)
Connexionists models(*) (ANNs)
Case-based predictor
Evolutionary Computation (Gas)
Collaborative resolution (Swarm intelligence)

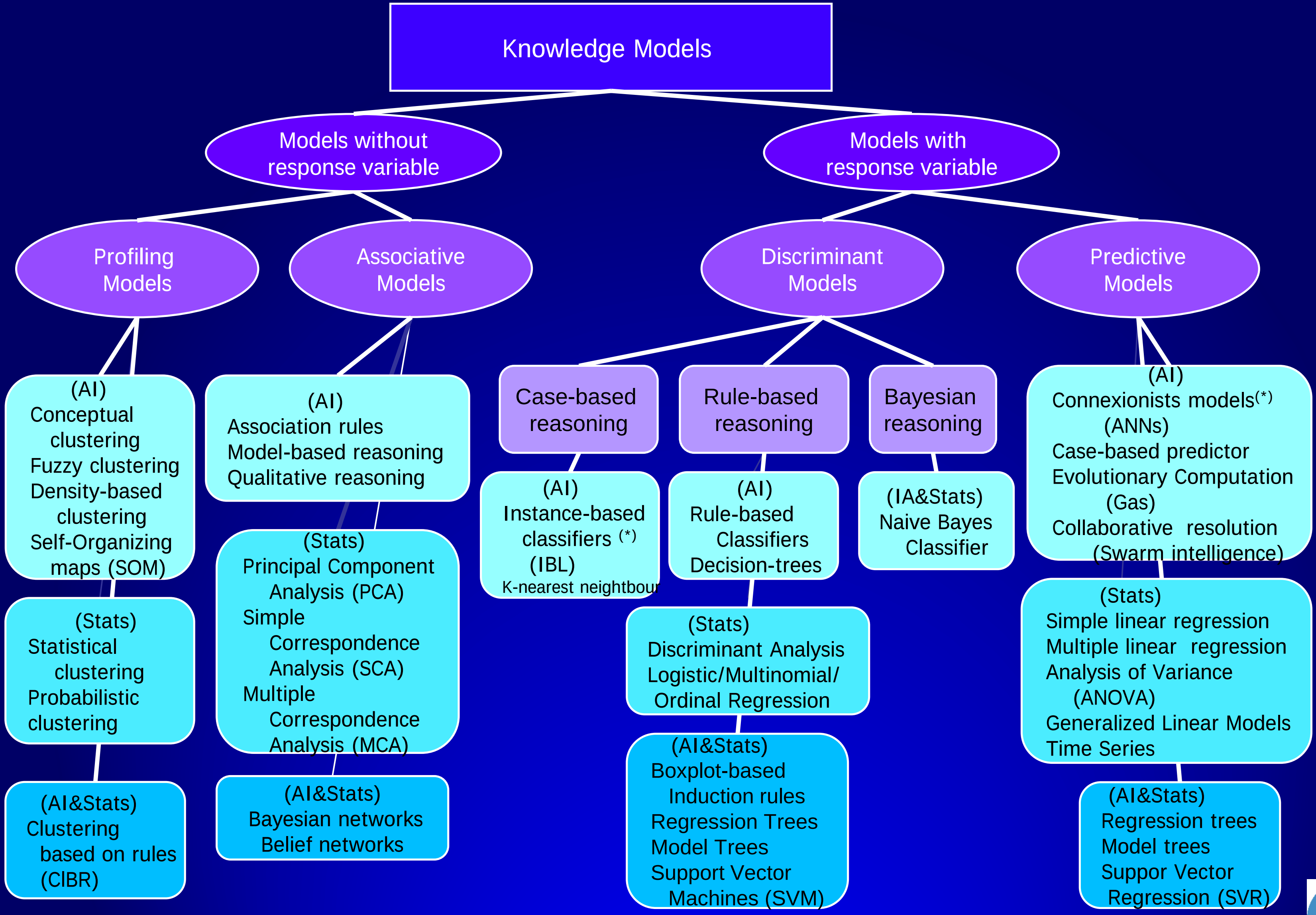
(AI&Stats)
Regression trees
Model trees
Suppor Vector Regression (SVR)

(Stats)
Simple linear regression
Multiple linear regression
Analysis of Variance (ANOVA)
Generalized Linear Models
Time Series

Gibert, K, M. Sànchez-Marrè, V. Codina (2010) Choosing the right data mining technique: classification of methods and intelligent recommenders. Proc. of the IEMSs'10, 5th biennial meeting (III DMTES Workshop), S23.03.1-S23.03.9. 2010

Gibert K, M. Sànchez-Marrè (2010b) Elección de la técnica de minería de datos: Mapa conceptual de técnicas. Actas del V simposio de teoría y aplicaciones de minería de datos: TAMIDA 2010. pp: 37—44. Ibergaceta.2010

The Data Mining Methods Conceptual Map DMMCM map *[Gibert et al 2018]*



Knowledge Models

Models without response variable

Models with response variable

Profiling Models

Associative Models

Discriminant Models

Predictive Models

(AI)
Conceptual clustering
Fuzzy clustering
Density-based clustering
Self-Organizing maps (SOM)

(Stats)
Statistical clustering
Probabilistic clustering

(AI&Stats)
Clustering based on rules (ClBR)

(AI)
Association rules
Model-based reasoning
Qualitative reasoning

(Stats)
Principal Component Analysis (PCA)
Simple Correspondence Analysis (SCA)
Multiple Correspondence Analysis (MCA)

(AI&Stats)
Bayesian networks
Belief networks

Case-based reasoning

Rule-based reasoning

Bayesian reasoning

(AI)
Instance-based classifiers (*)
(IBL)
K-nearest neightbour

(AI)
Rule-based Classifiers
Decision-trees

(IA&Stats)
Naive Bayes Classifier

(Stats)
Discriminant Analysis
Logistic/Multinomial/ Ordinal Regression

(AI&Stats)
Boxplot-based Induction rules
Regression Trees
Model Trees
Support Vector Machines (SVM)

(AI)
Connexionists models(*)
(ANNs)
Case-based predictor
Evolutionary Computation (Gas)
Collaborative resolution (Swarm intelligence)

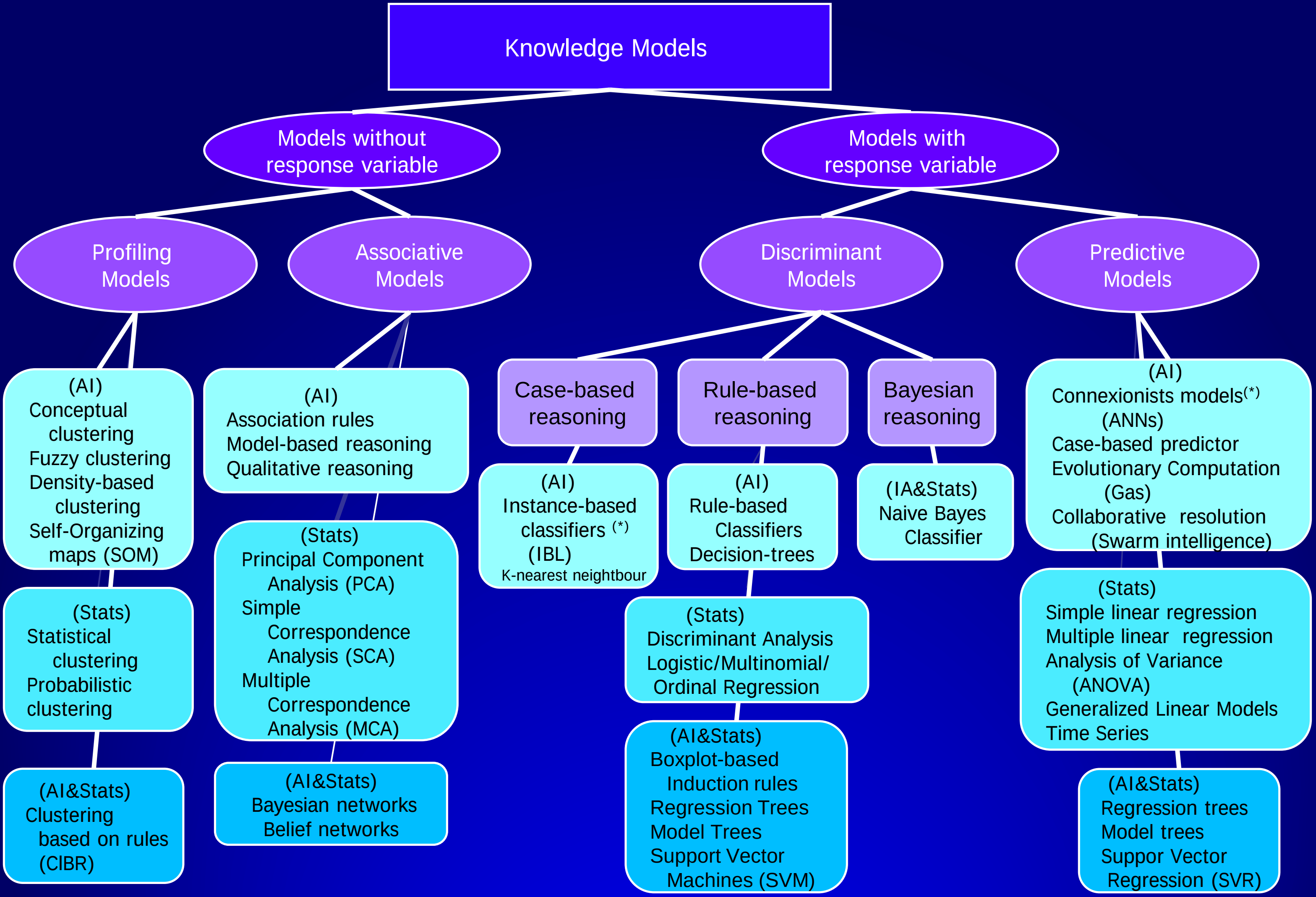
(Stats)
Simple linear regression
Multiple linear regression
Analysis of Variance (ANOVA)
Generalized Linear Models
Time Series

(AI&Stats)
Regression trees
Model trees
Suppor Vector Regression (SVR)

© K. Gibert

The Data Mining Methods Conceptual Map

DMMCM map *[Gibert et al 2018]*



Knowledge Models

Models without response variable

Profiling Models

(AI)
Conceptual clustering
Fuzzy clustering
Density-based clustering
Self-Organizing maps (SOM)

(Stats)
Statistical clustering
Probabilistic clustering

(AI&Stats)
Clustering based on rules (CIBR)

Associative Models

(AI)
Association rules
Model-based reasoning
Qualitative reasoning

(Stats)
Principal Component Analysis (PCA)
Simple Correspondence Analysis (SCA)
Multiple Correspondence Analysis (MCA)

(AI&Stats)
Bayesian networks
Belief networks

Models with response variable

Discriminant Models

Case-based reasoning

(AI)
Instance-based classifiers (*)
(IBL)
K-nearest neightbour

Rule-based reasoning

(AI)
Rule-based Classifiers
Decision-trees

(Stats)
Discriminant Analysis
Logistic/Multinomial/ Ordinal Regression

(AI&Stats)
Boxplot-based Induction rules
Regression Trees
Model Trees
Support Vector Machines (SVM)

Bayesian reasoning

(IA&Stats)
Naive Bayes Classifier

Predictive Models

(AI)
Connexionists models(*) (ANNs)
Case-based predictor
Evolutionary Computation (Gas)
Collaborative resolution (Swarm intelligence)

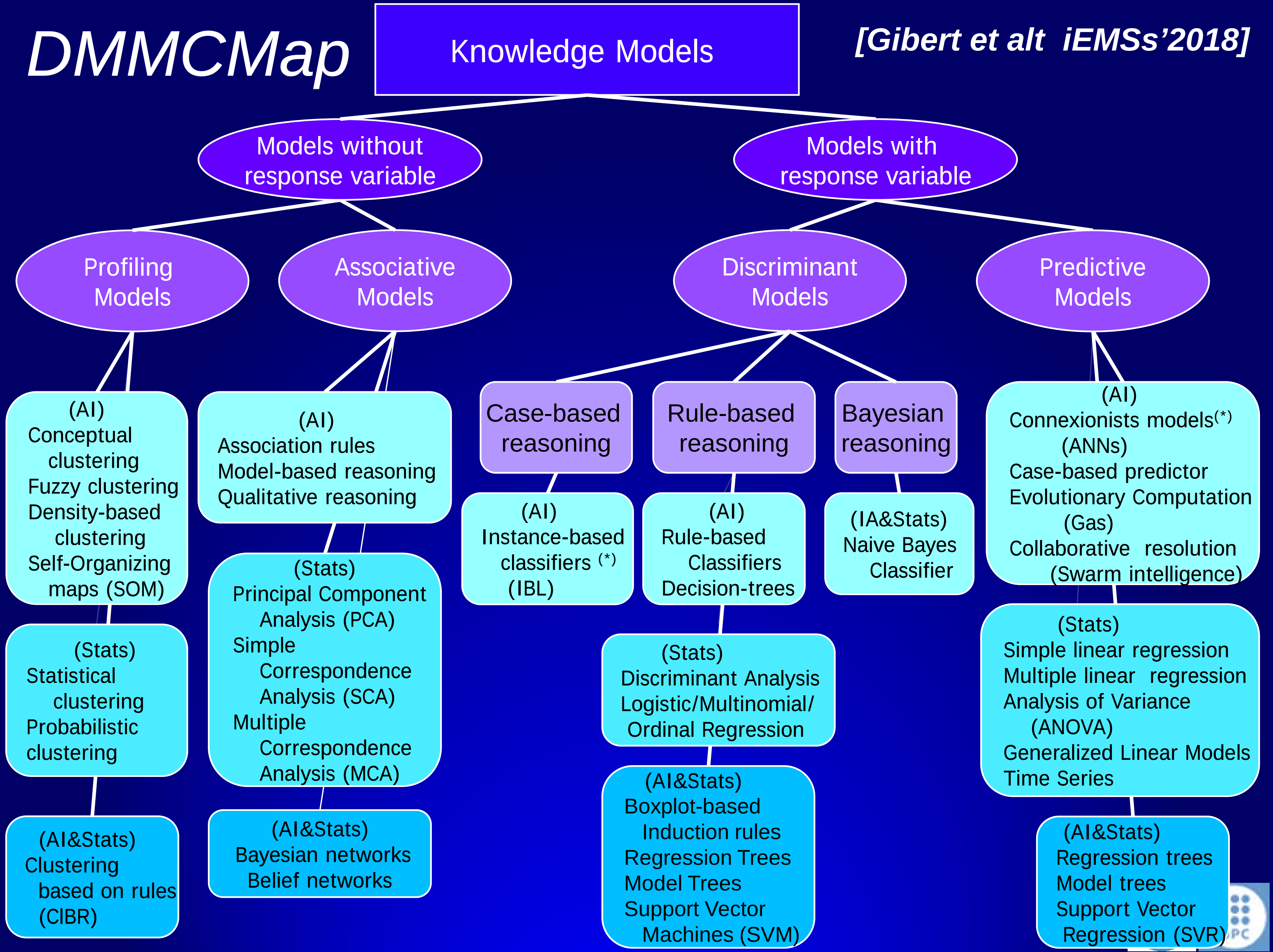
(Stats)
Simple linear regression
Multiple linear regression
Analysis of Variance (ANOVA)
Generalized Linear Models
Time Series

(AI&Stats)
Regression trees
Model trees
Suppor Vector Regression (SVR)

Gibert, K. J. Izquierdo, M. Sànchez-Marrè, S. (H) Hamilton, I. Rodríguez-Roda, G. Holmes (2018) Which method to use? An assessment of Data Mining Methods in Environmental Data Mining. Environmental Modelling and Software 110: 3-27. Elsevier. DOI: doi.org/10.1016/j.envsoft.2018.09.021

DMMCMap

[Gibert et alt iEMSs'2018]



Knowledge Models

Models without response variable

Models with response variable

Profiling Models

Associative Models

Discriminant Models

Predictive Models

(AI)
Conceptual clustering
Fuzzy clustering
Density-based clustering
Self-Organizing maps (SOM)

(Stats)
Statistical clustering
Probabilistic clustering

(AI&Stats)
Clustering based on rules (ClBR)

(AI)
Association rules
Model-based reasoning
Qualitative reasoning

(Stats)
Principal Component Analysis (PCA)
Simple Correspondence Analysis (SCA)
Multiple Correspondence Analysis (MCA)

(AI&Stats)
Bayesian networks
Belief networks

Case-based reasoning

(AI)
Instance-based classifiers $^{(*)}$ (IBL)

Rule-based reasoning

(AI)
Rule-based Classifiers
Decision-trees

(Stats)
Discriminant Analysis
Logistic/Multinomial/ Ordinal Regression

(AI&Stats)
Boxplot-based Induction rules
Regression Trees
Model Trees
Support Vector Machines (SVM)

Bayesian reasoning

(IA&Stats)
Naive Bayes Classifier

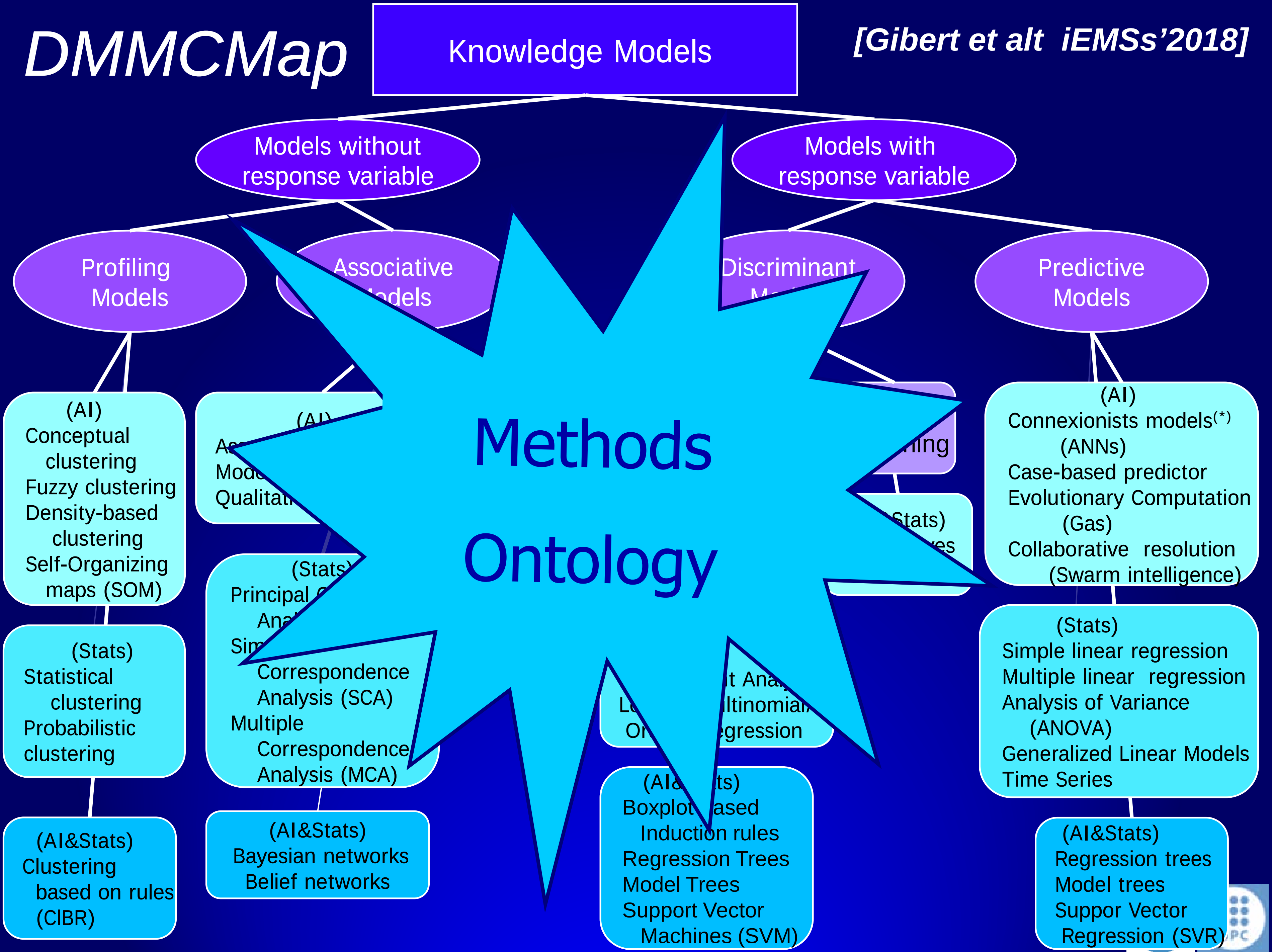
(AI)
Connexionists models$^{(*)}$ (ANNs)
Case-based predictor
Evolutionary Computation (Gas)
Collaborative resolution (Swarm intelligence)

(Stats)
Simple linear regression
Multiple linear regression
Analysis of Variance (ANOVA)
Generalized Linear Models
Time Series

(AI&Stats)
Regression trees
Model trees
Support Vector Regression (SVR)

DMMCMap

[Gibert et alt iEMSs'2018]



Knowledge Models

Models without response variable

Models with response variable

Profiling Models

Associative ...odels

Discriminant M...

Predictive Models

(AI)
Conceptual clustering
Fuzzy clustering
Density-based clustering
Self-Organizing maps (SOM)

(Stats)
Statistical clustering
Probabilistic clustering

(AI&Stats)
Clustering based on rules (ClBR)

(AI)
As...
Mode...
Qualitat...

(Stats)
Principal C...
Ana...
Sim...
Correspondence Analysis (SCA)
Multiple Correspondence Analysis (MCA)

(AI&Stats)
Bayesian networks
Belief networks

...ning

...Stats)
...es

...t Ana...
L... ...ltinomial
Or... ...egression

(AI& ...ts)
Boxplot ...ased Induction rules
Regression Trees
Model Trees
Support Vector Machines (SVM)

(AI)
Connexionists models(*) (ANNs)
Case-based predictor
Evolutionary Computation (Gas)
Collaborative resolution (Swarm intelligence)

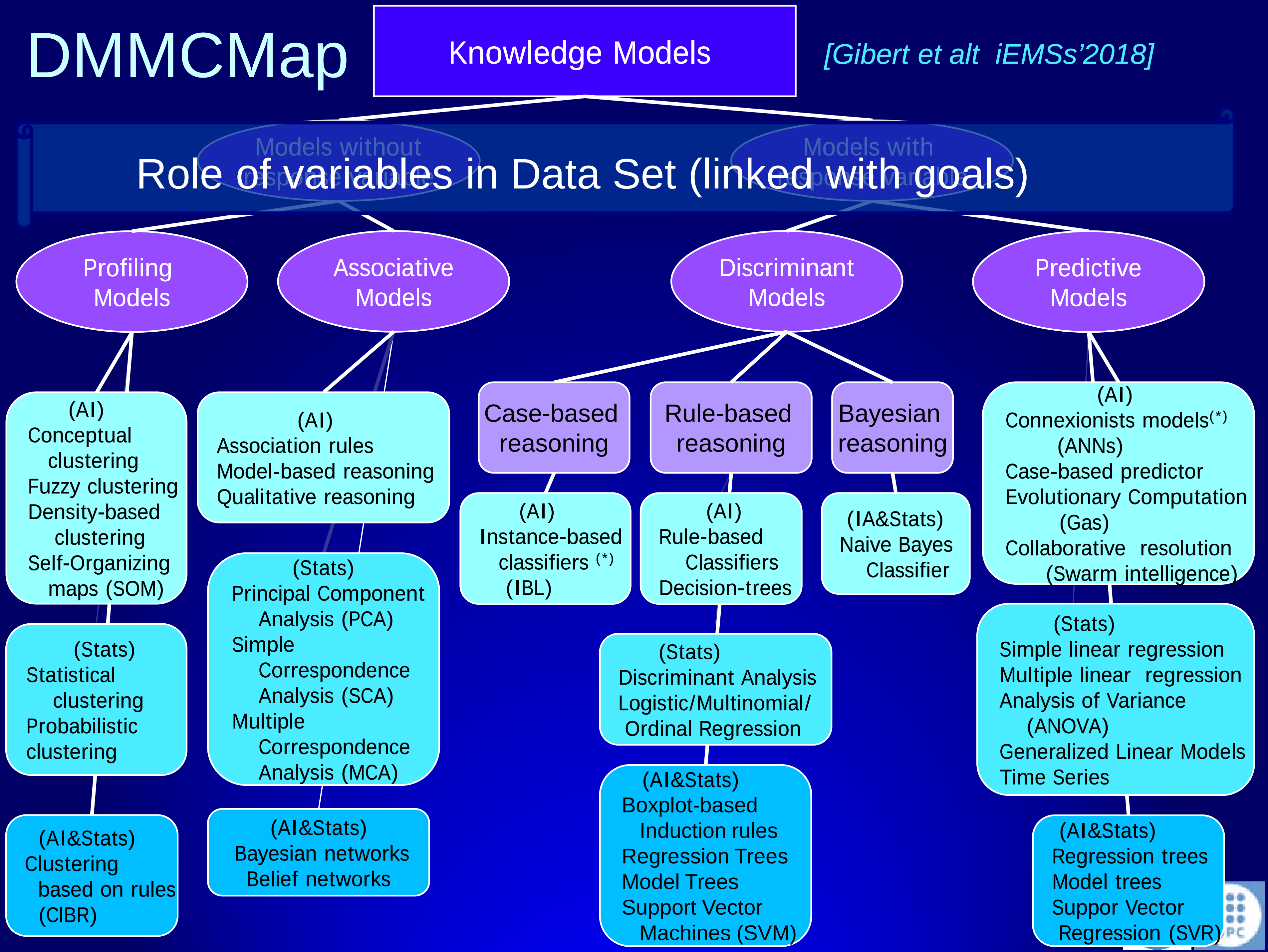
(Stats)
Simple linear regression
Multiple linear regression
Analysis of Variance (ANOVA)
Generalized Linear Models
Time Series

(AI&Stats)
Regression trees
Model trees
Support Vector Regression (SVR)

Methods Ontology

DMMCMap

[Gibert et alt iEMSs'2018]



Knowledge Models

Models without response variable

Models with response variable

Role of variables in Data Set (linked with goals)

Profiling Models

(AI)
Conceptual clustering
Fuzzy clustering
Density-based clustering
Self-Organizing maps (SOM)

(Stats)
Statistical clustering
Probabilistic clustering

(AI&Stats)
Clustering based on rules (ClBR)

Associative Models

(AI)
Association rules
Model-based reasoning
Qualitative reasoning

(Stats)
Principal Component Analysis (PCA)
Simple Correspondence Analysis (SCA)
Multiple Correspondence Analysis (MCA)

(AI&Stats)
Bayesian networks
Belief networks

Discriminant Models

Case-based reasoning

(AI)
Instance-based classifiers [(*)]
(IBL)

Rule-based reasoning

(AI)
Rule-based Classifiers
Decision-trees

(Stats)
Discriminant Analysis
Logistic/Multinomial/ Ordinal Regression

(AI&Stats)
Boxplot-based Induction rules
Regression Trees
Model Trees
Support Vector Machines (SVM)

Bayesian reasoning

(IA&Stats)
Naive Bayes Classifier

Predictive Models

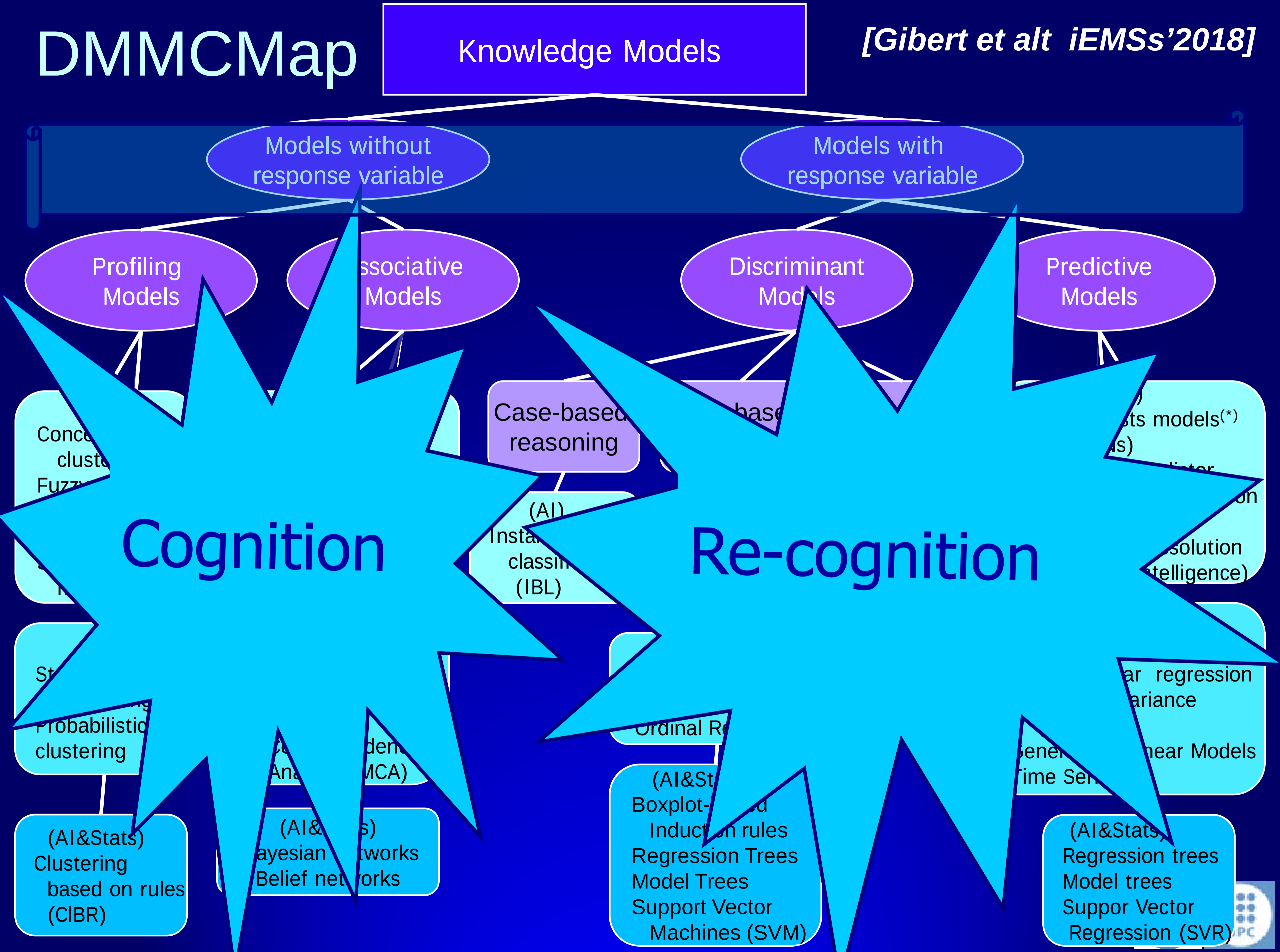
(AI)
Connexionists models[(*)]
(ANNs)
Case-based predictor
Evolutionary Computation
(Gas)
Collaborative resolution
(Swarm intelligence)

(Stats)
Simple linear regression
Multiple linear regression
Analysis of Variance
(ANOVA)
Generalized Linear Models
Time Series

(AI&Stats)
Regression trees
Model trees
Support Vector Regression (SVR)

DMMCMap

[Gibert et alt iEMSs'2018]



Knowledge Models

Models without response variable

Models with response variable

Profiling Models

ssociative Models

Discriminant Mod ls

Predictive Models

Conce
cluste
Fuzzy

Case-based reasoning

base

)
ts models(*)
s)
lictor
on
solution
telligence)

(AI)
Insta
classi
(IBL)

Cognition

Re-cognition

St
Probabilistic
clustering

Co dent
Ana MCA)

Ordinal Re

ar regression
ariance
Gener near Models
Time Se

(AI&Stats)
Clustering
based on rules
(ClBR)

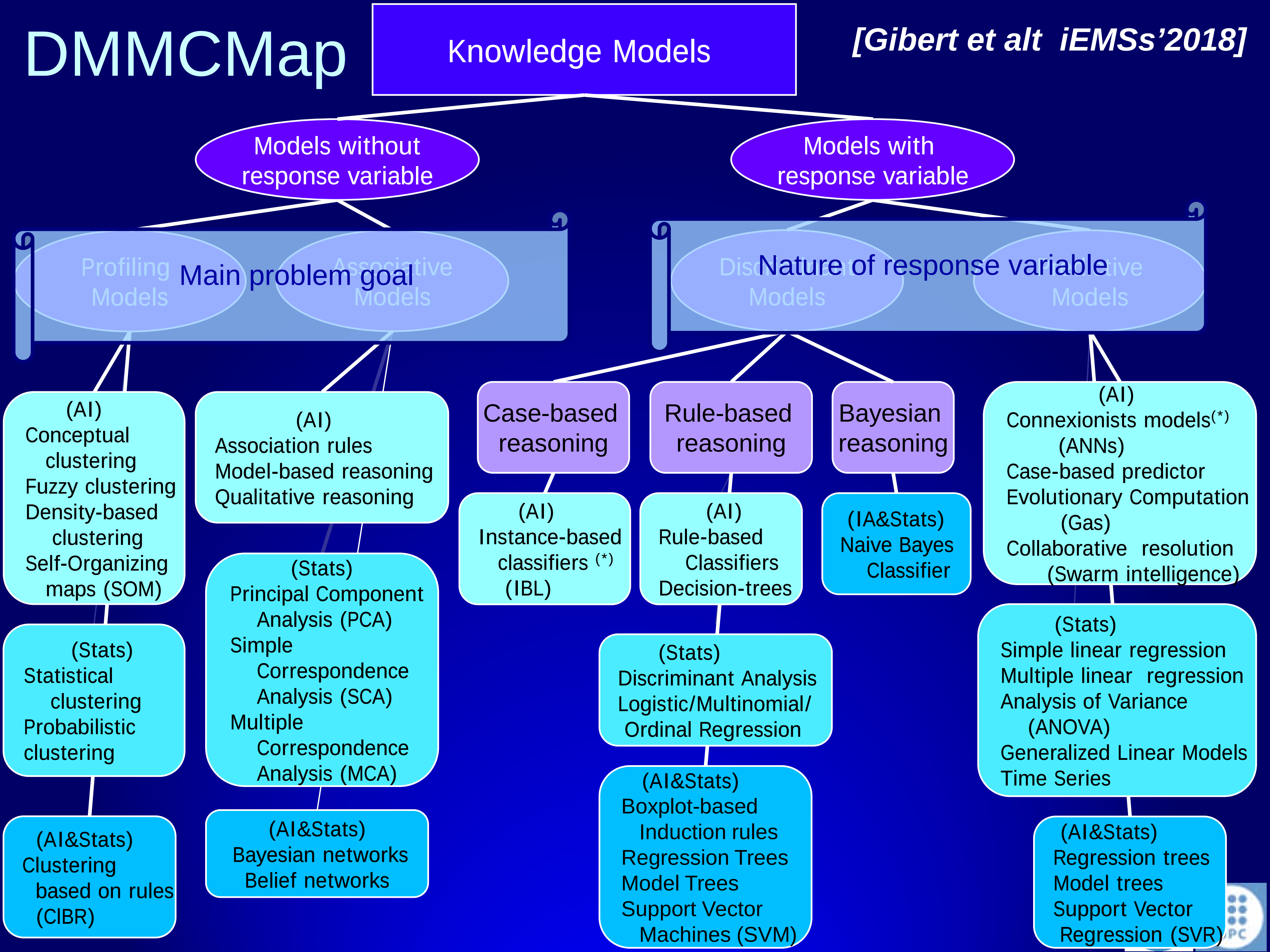
(AI& s)
ayesian tworks
Belief net orks

(AI&St
Boxplot- d
Induct on rules
Regression Trees
Model Trees
Support Vector
Machines (SVM)

(AI&Stats)
Regression trees
Model trees
Support Vector
Regression (SVR)

DMMCMap

[Gibert et alt iEMSs'2018]



Knowledge Models

Models without response variable

Main problem goal

Profiling Models

Associative Models

(AI)
Conceptual clustering
Fuzzy clustering
Density-based clustering
Self-Organizing maps (SOM)

(Stats)
Statistical clustering
Probabilistic clustering

(AI&Stats)
Clustering based on rules (ClBR)

(AI)
Association rules
Model-based reasoning
Qualitative reasoning

(Stats)
Principal Component Analysis (PCA)
Simple Correspondence Analysis (SCA)
Multiple Correspondence Analysis (MCA)

(AI&Stats)
Bayesian networks
Belief networks

Models with response variable

Nature of response variable

Discriminant Models

Predictive Models

Case-based reasoning

(AI)
Instance-based classifiers (*)
(IBL)

Rule-based reasoning

(AI)
Rule-based Classifiers
Decision-trees

(Stats)
Discriminant Analysis
Logistic/Multinomial/ Ordinal Regression

(AI&Stats)
Boxplot-based Induction rules
Regression Trees
Model Trees
Support Vector Machines (SVM)

Bayesian reasoning

(IA&Stats)
Naive Bayes Classifier

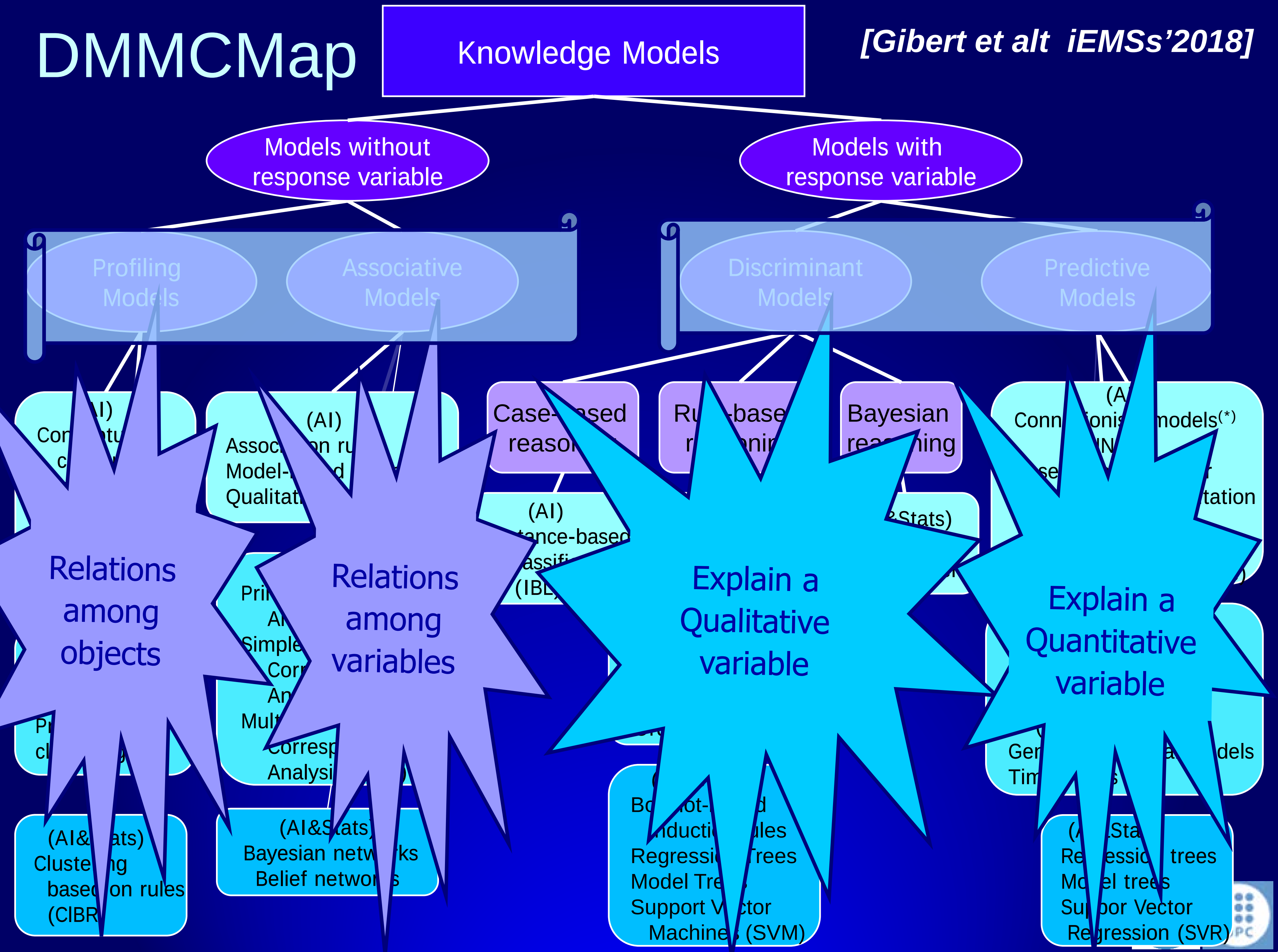
(AI)
Connexionists models(*)
(ANNs)
Case-based predictor
Evolutionary Computation
(Gas)
Collaborative resolution
(Swarm intelligence)

(Stats)
Simple linear regression
Multiple linear regression
Analysis of Variance (ANOVA)
Generalized Linear Models
Time Series

(AI&Stats)
Regression trees
Model trees
Support Vector Regression (SVR)

DMMCMap

[Gibert et alt iEMSs'2018]



Knowledge Models

Models without response variable

Models with response variable

Profiling Models

Associative Models

Discriminant Models

Predictive Models

(AI)
Con tu
c

(AI)
Assoc on ru
Model- d
Qualitat

Case- sed
reaso

Ru -base
r nin

Bayesian
rea ning

(A
Conn ionis models(*)
IN
se r
tation

(AI)
tance-based
assifi
(IBL)

Stats)

Pri
A
Simple
Corr
An
Mult
Corresp
Analysi

Pr
cl

(AI& ats)
Cluste ng
base on rules
(CIBR

(AI&Stats)
Bayesian netw ks
Belief networ s

Bo ot- d
nducti ules
Regressi Trees
Model Tre
Support V ctor
Machine (SVM)

Ger a dels
Tim s

(A Sta
Re essio trees
Mo el trees
Sup por Vector
Regression (SVR)

Relations among objects

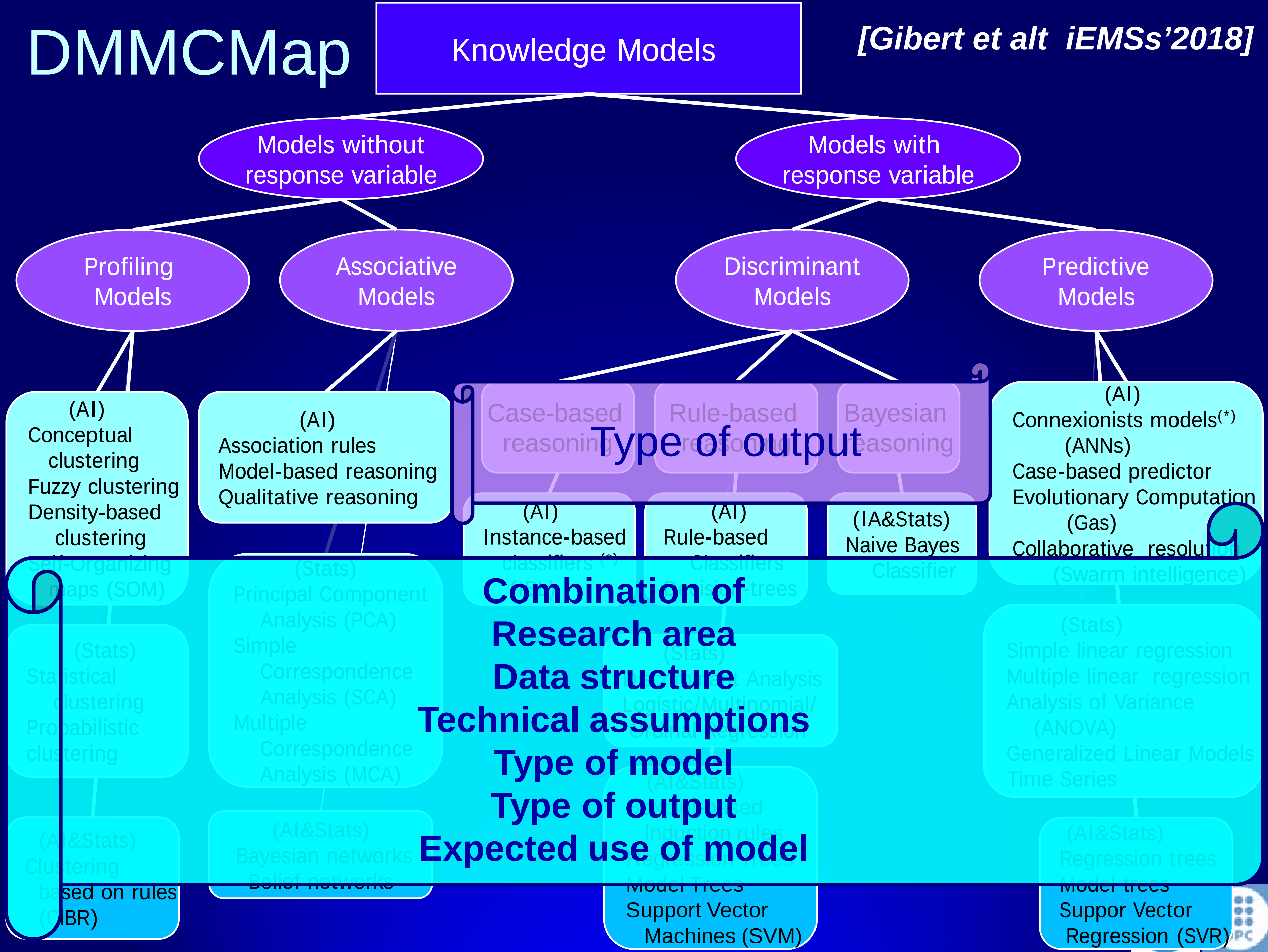
Relations among variables

Explain a Qualitative variable

Explain a Quantitative variable

DMMCMap

[Gibert et alt iEMSs'2018]



Knowledge Models

Models without response variable

Models with response variable

Profiling Models

Associative Models

Discriminant Models

Predictive Models

(AI)
Conceptual clustering
Fuzzy clustering
Density-based clustering
Self-Organizing maps (SOM)

(Stats)
Statistical clustering
Probabilistic clustering

(AI&Stats)
Clustering based on rules (ClBR)

(AI)
Association rules
Model-based reasoning
Qualitative reasoning

(Stats)
Principal Component Analysis (PCA)
Simple Correspondence Analysis (SCA)
Multiple Correspondence Analysis (MCA)

(AI&Stats)
Bayesian networks
Belief networks

Case-based reasoning

Rule-based reasoning

Bayesian reasoning

(AI)
Instance-based classifiers (*) (IBL)

(AI)
Rule-based Classifiers
Decision trees

(IA&Stats)
Naive Bayes Classifier

(Stats)
Discriminant Analysis
Logistic/Multinomial/ Ordinal Regression

(AI&Stats)
Boxplot-based Induction rules
Regression Trees
Model Trees
Support Vector Machines (SVM)

(AI)
Connexionists models(*)
(ANNs)
Case-based predictor
Evolutionary Computation (Gas)
Collaborative resolution (Swarm intelligence)

(Stats)
Simple linear regression
Multiple linear regression
Analysis of Variance (ANOVA)
Generalized Linear Models
Time Series

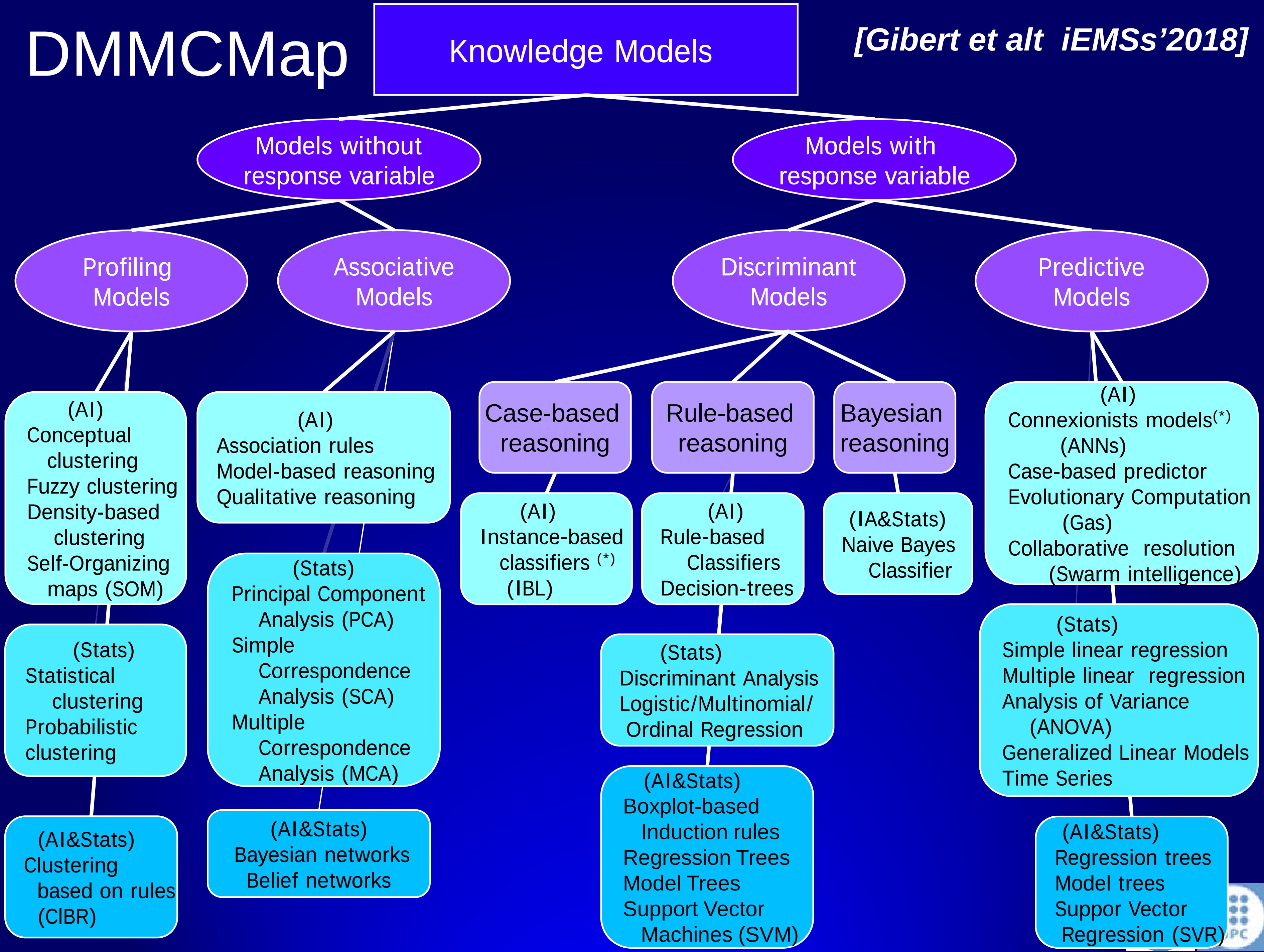
(AI&Stats)
Regression trees
Model trees
Suppor Vector Regression (SVR)

Type of output

Combination of
Research area
Data structure
Technical assumptions
Type of model
Type of output
Expected use of model

DMMCMap

[Gibert et alt iEMSs'2018]



Knowledge Models

- Models without response variable
  - Profiling Models
    - (AI) Conceptual clustering; Fuzzy clustering; Density-based clustering; Self-Organizing maps (SOM)
    - (Stats) Statistical clustering; Probabilistic clustering
    - (AI&Stats) Clustering based on rules (ClBR)
  - Associative Models
    - (AI) Association rules; Model-based reasoning; Qualitative reasoning
    - (Stats) Principal Component Analysis (PCA); Simple Correspondence Analysis (SCA); Multiple Correspondence Analysis (MCA)
    - (AI&Stats) Bayesian networks; Belief networks
- Models with response variable
  - Discriminant Models
    - Case-based reasoning
      - (AI) Instance-based classifiers (*) (IBL)
    - Rule-based reasoning
      - (AI) Rule-based Classifiers; Decision-trees
      - (Stats) Discriminant Analysis; Logistic/Multinomial/ Ordinal Regression
      - (AI&Stats) Boxplot-based Induction rules; Regression Trees; Model Trees; Support Vector Machines (SVM)
    - Bayesian reasoning
      - (IA&Stats) Naive Bayes Classifier
  - Predictive Models
    - (AI) Connexionists models(*) (ANNs); Case-based predictor; Evolutionary Computation (Gas); Collaborative resolution (Swarm intelligence)
    - (Stats) Simple linear regression; Multiple linear regression; Analysis of Variance (ANOVA); Generalized Linear Models; Time Series
    - (AI&Stats) Regression trees; Model trees; Support Vector Regression (SVR)

The associated decision process *[iEMSs18]*

Two steps decision process:

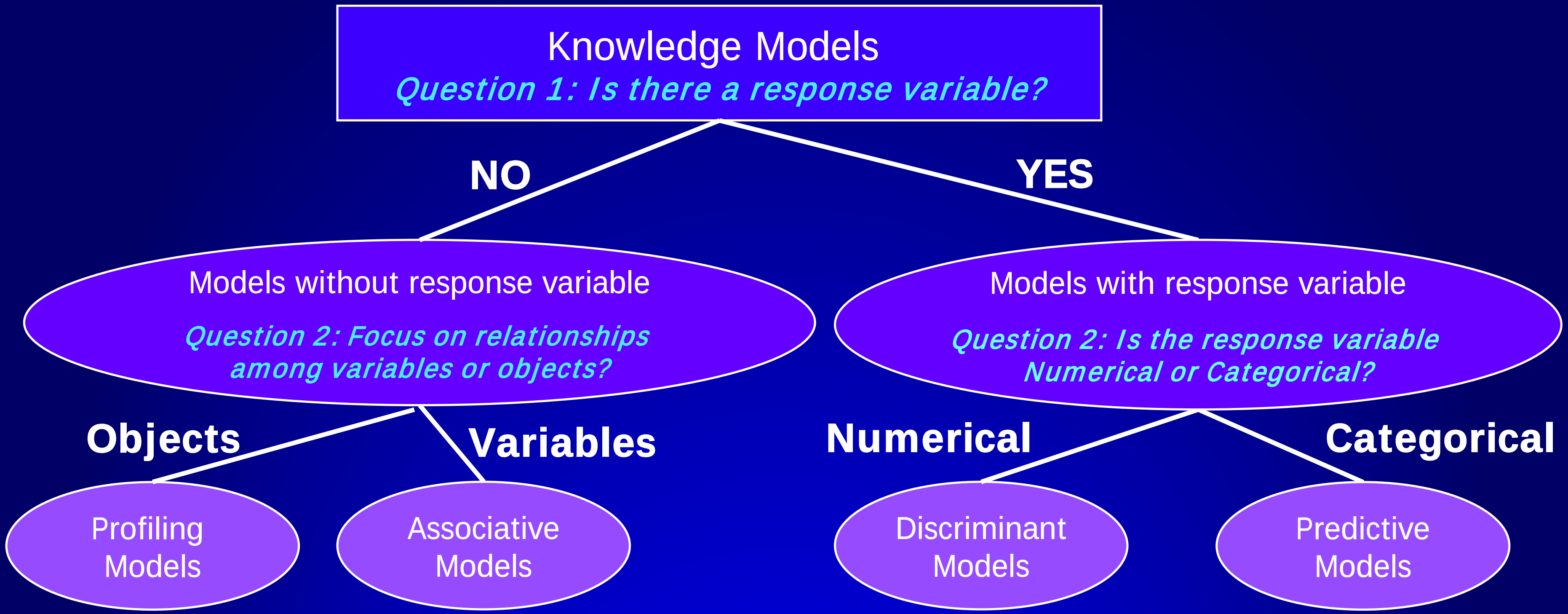
1. Identify the structure of the target problem:
   Determine the main branch of the DMMCM
   questions associated to nodes in the DMMCM

2. Identify appropriate technique within the DMMCM branch:

   i. Find DMMT of selected branch and identify a particular box in DMMCM map

The associated decision process *[iEMSs18]*

Three questions for Step 1



Knowledge Models

Question 1: Is there a response variable?

NO

YES

Models without response variable

Question 2: Focus on relationships among variables or objects?

Models with response variable

Question 2: Is the response variable Numerical or Categorical?

Objects

Variables

Numerical

Categorical

Profiling Models

Associative Models

Discriminant Models

Predictive Models

Select one brach of the DMMCM map

The associated decision process *[iEMSs18]*

Formal framework for methods

i) **Technical requirements**: related to dataset structure.

- Type of explanatory/response variable?

Numerical,ordered qualitative, non-ordered qualitative, all

Missmatch: incorrect use

ii) **Non restrictive technical properties**: related to data structure.

- Required data size
  - Variable independence required
  - Normality required
  - Outliers non acceptable
  - Recommended data size

Missmatch: loose of performance

iii) **Non restrictive preference properties**: user preferences/goals

- Is running speed a priority?
- Is interpretability of results a priority?
- Is machine readability required?

Missmatch: loose of preference

DMMT (Data Mining Method Template)

Formal framework for methods description

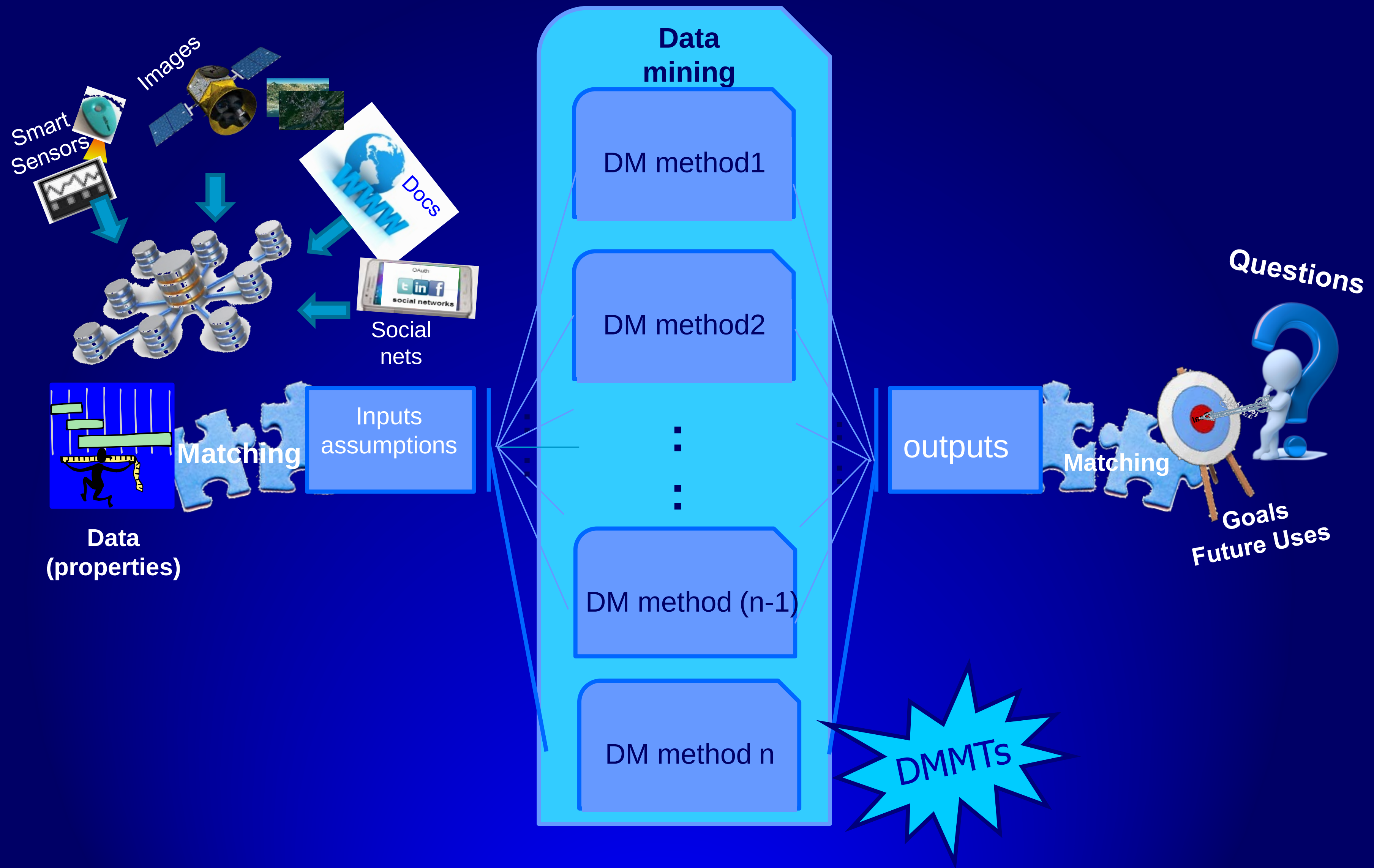
- Main goals of a family of methods
- Brief discussion of the main principles of the family
- Type of input required
- Technical assumptions to be assessed on data
- Requirements
- Type of output expected from the method
- Applications and references

DMMT (Data Mining Method Template)

- Profiling methods
- Associative
- Discriminant
  - CBR
- Predictive
  - ANN
  - Evolutionary Computation

The associated decision process *[iEMSs18]*

Browse on corresponding DMMT in Step 2



Smart Sensors

Images

Docs

Social nets

Data (properties)

Matching

Inputs assumptions

Data mining

DM method1

DM method2

⋮

⋮

DM method (n-1)

DM method n

outputs

Matching

Questions

Goals

Future Uses

DMMTs

Conclusions and future work

- DMMCM: (non-exhaustive) ontology of Data Mining methods
- A 2-steps decision process is proposed to choose a DM method for a real problem
  - Step 1 determines a family of problems with 3 simple questions
  - Step 2 determines a concrete suitable method in the family
- Properties of methods are critical to choose
- A formal framework to describe methods is proposed
  - DMMT
  - Restrictions system
  - Allows dynamic growth of DM methods according to SoTA
- Data structure and model uses are relevant criteria
- Currently an intelligent methodological recommender is built using DMMT and restrictions framework