

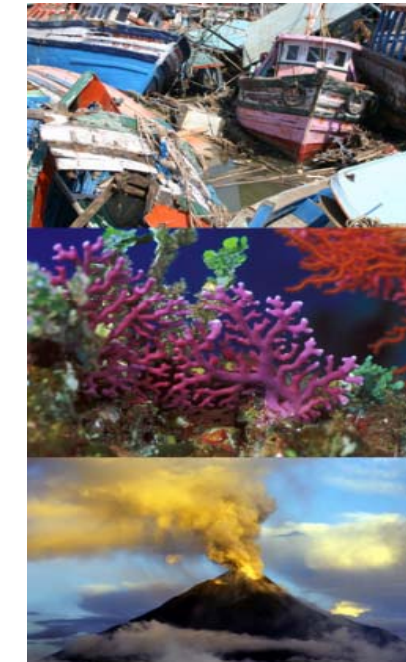
AMADEUS

Axe Qualité des données scientifiques :

*Détection des anomalies, chaînes de prétraitements
et application aux données du CEREGE*

UMR ESPACE DEV :

IRD, Univ. Montpellier 2, Univ. Antilles Guyanne,
Univ. de la Réunion



L'espace au service du développement

Expertise et spatialisation des connaissances pour l'aide à la décision

ESPACE DEV : Participants

- Laure Berti-Équille (DR, IRD)
- Isabelle Mougenot (MC, UM2)
- Emmanuel Roux (CR, IRD)
- Frédérique Seyler (DR, IRD)
- Thérèse Libourel (Pr, UM2)
- Yuan Lin (CDD, IRD)

Plan suivi *(cf. ce qui est attendu pour le 5 décembre)*

- ✓ Challenges scientifiques du projet à 4 ou 5 ans et affinage éventuel de ces challenges depuis la soumission
- ✓ Organisation du projet, modalités de travail collaboratif, agenda des réunions techniques de l'année 2012
- ✓ Premiers résultats scientifiques obtenus ou identification des premières pistes de recherche explorées et positionnement par rapport à l'état de l'art
- ✓ Objectifs scientifiques pour 2013 : vue détaillée
- ✓ Utilisation du budget alloué à ESPACE DEV

Challenges Scientifiques

ESPACE DEV

1. Améliorer la détection des anomalies

- Corroboration des méthodes de détection
- Benchmark des méthodes
- Stratégies de recommandation de méthodes de détection

2. Automatiser les chaines de prétraitements

- Formalisation des workflows de prétraitements des données du LAM et du CEREGE
- Générer d'autres alternatives de prétraitement
- Modèle de coût des différentes alternatives de prétraitement et optimisation
- Automatiser la préparation des données et son ajustement selon l'analyse

Organisation: Réunions ESPACE DEV

- 20 juin, Marseille : CEREGE, ESPACE DEV, LAM, LIF, LIRMM, LABRI
S.Maabout, N.Hanusse, P.Poncelet, D.lenco, T.Libourel, L. Berti-Equille, V.Moron, C.Surace, N.Novelli, S.Nedjar,
- 10-11 septembre, Bordeaux : ESPACE DEV, LIF, LAM, LABRI, LIRMM
S.Maabout, N.Hanusse, E.Garnaud, R.Bourqui, M.Teissere, D.lenco, Y.Lin, C.Surace, N.Novelli
- 13 septembre, Marseille : ESPACE DEV, LIF, LAM
C. Surace, N. Novelli, P-V. Chabaud, Y.Lin
- 13 septembre : ESPACE DEV, LAM, LIF
Y. Lin, N. Novelli, C. Surace
- 17 septembre, visio conf : ESPACE DEV, LIF, LAM, LIRMM, LABRI
C.Surace, N.Novelli, P-V.Chabaud, P.Poncelet, M.Teissere, D.lenco, T.Libourel, Y.Lin, S.Maabout, N.Hanusse, R.Bourqui
- 19 septembre, visio conf : ESPACE DEV, LIF, LAM, LIRM, LABRI
S. Mc Connell, C.Surace, T.Libourel, L. Berti-Equille, Y.Lin, M.Teisseire, D.lenco, N.Hanusse
- 2 novembre, visio conf : ESPACE DEV
Y. Lin, L. Berti-Equille
- 9 novembre, Marseille : ESPACE DEV, CEREGE
L. Berti-Equille, V. Moron

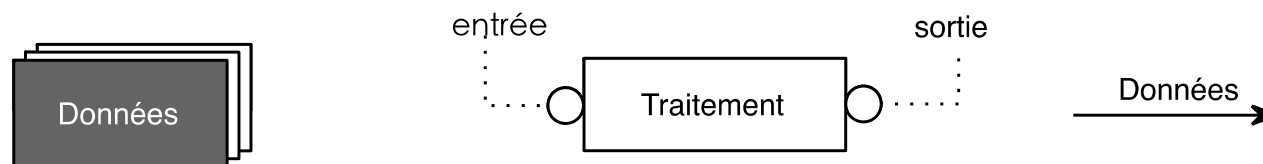
Premiers résultats

- Publication avec V. Moron et al., Journal of Applied Meteorology and Climatology
- Modélisation des workflows par Yuan Lin de traitements et analyses menées sur les données astro et climato
- Profilage des jeux de données (en cours)

Langage graphique utilisé

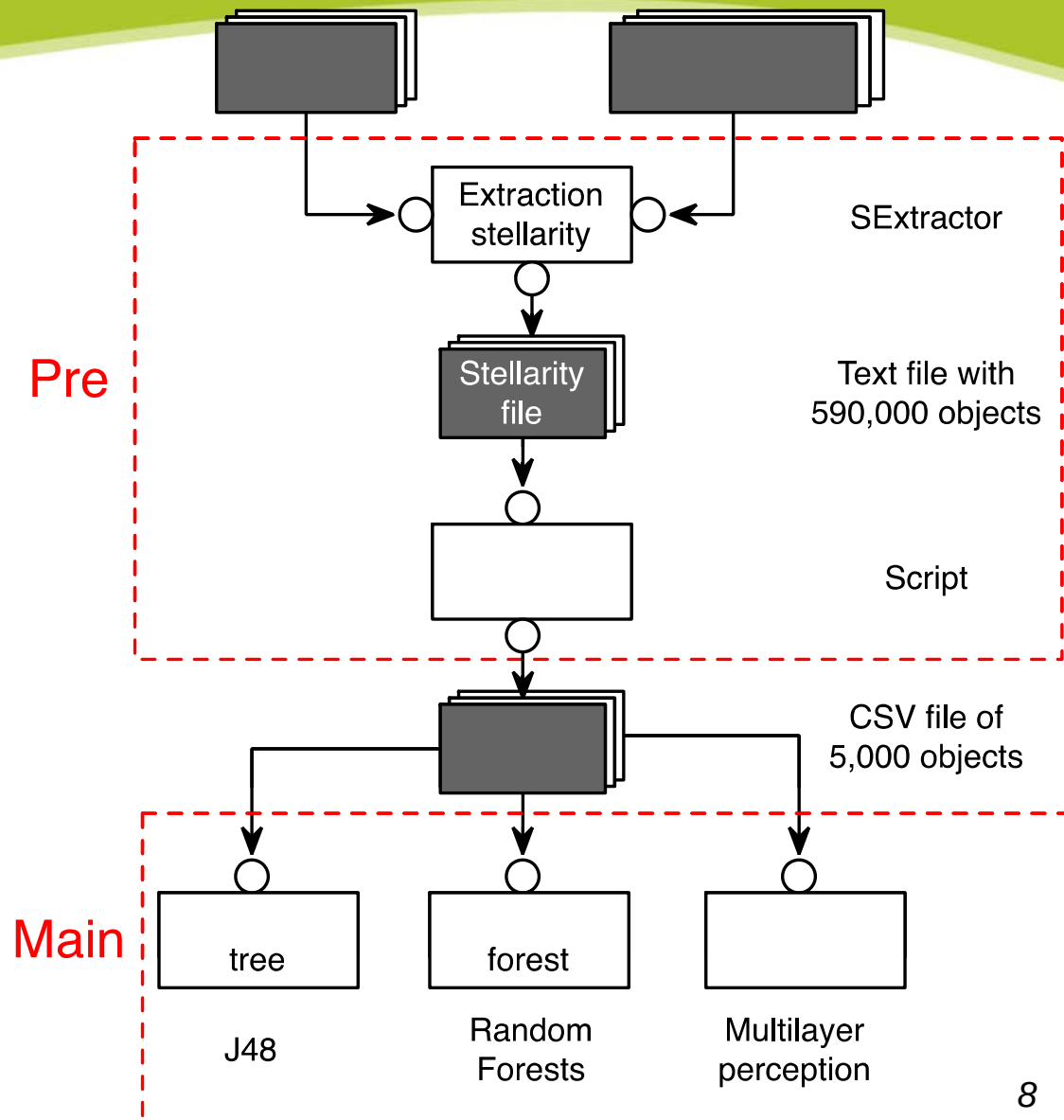
Trois symboles principaux :

- Données
- Traitement (processus)
- Lien de données



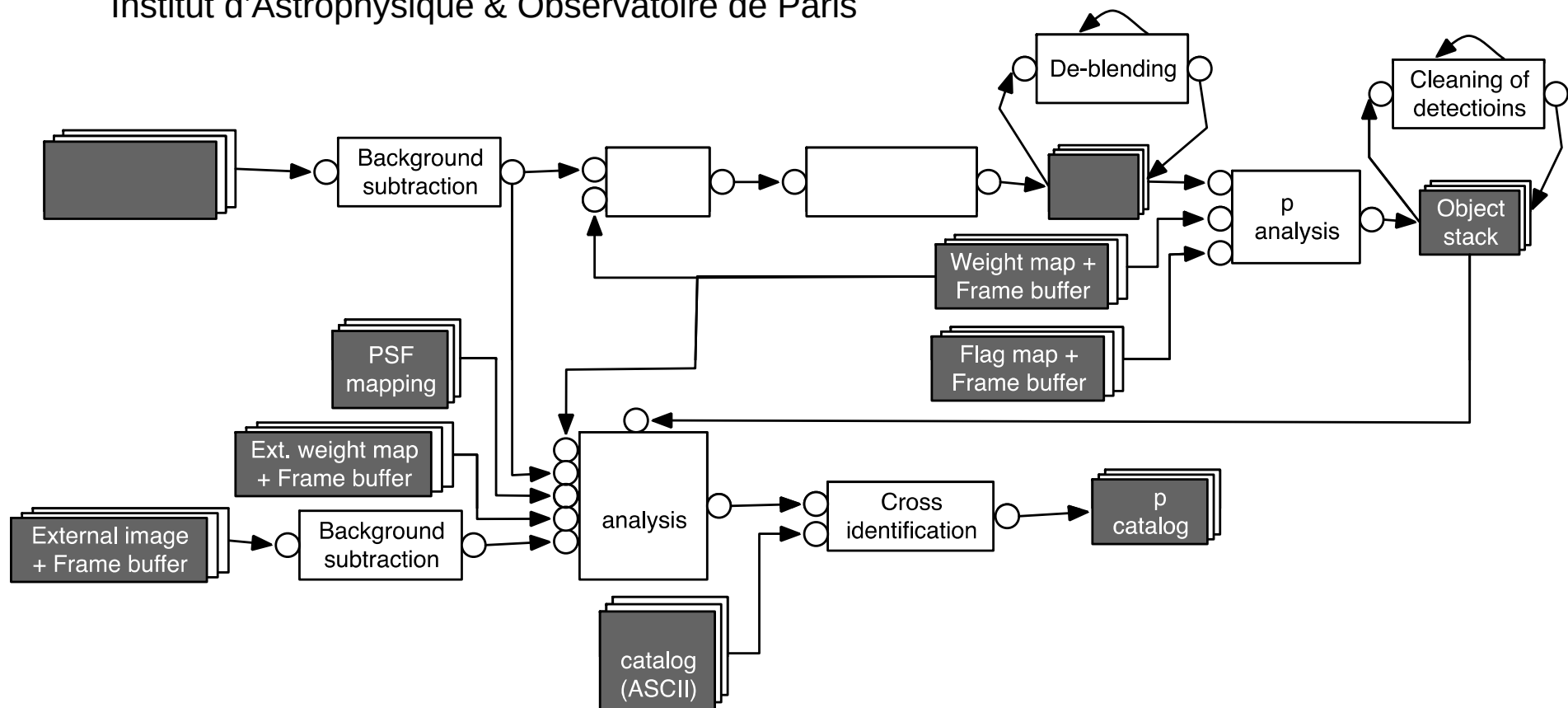
Les modèles de workflow (astro) (1)

*D'après [1] : Star-Galaxy
Classification Using Data Mining
Techniques with Considerations for
Unbalanced Datasets. Peter J.
O'Keefe and al.*



Les modèles de workflow (astro) (2)

D'après [2]: SExtractor SExtractor, v2.13, User's manual. E. BERTIN.
Institut d'Astrophysique & Observatoire de Paris



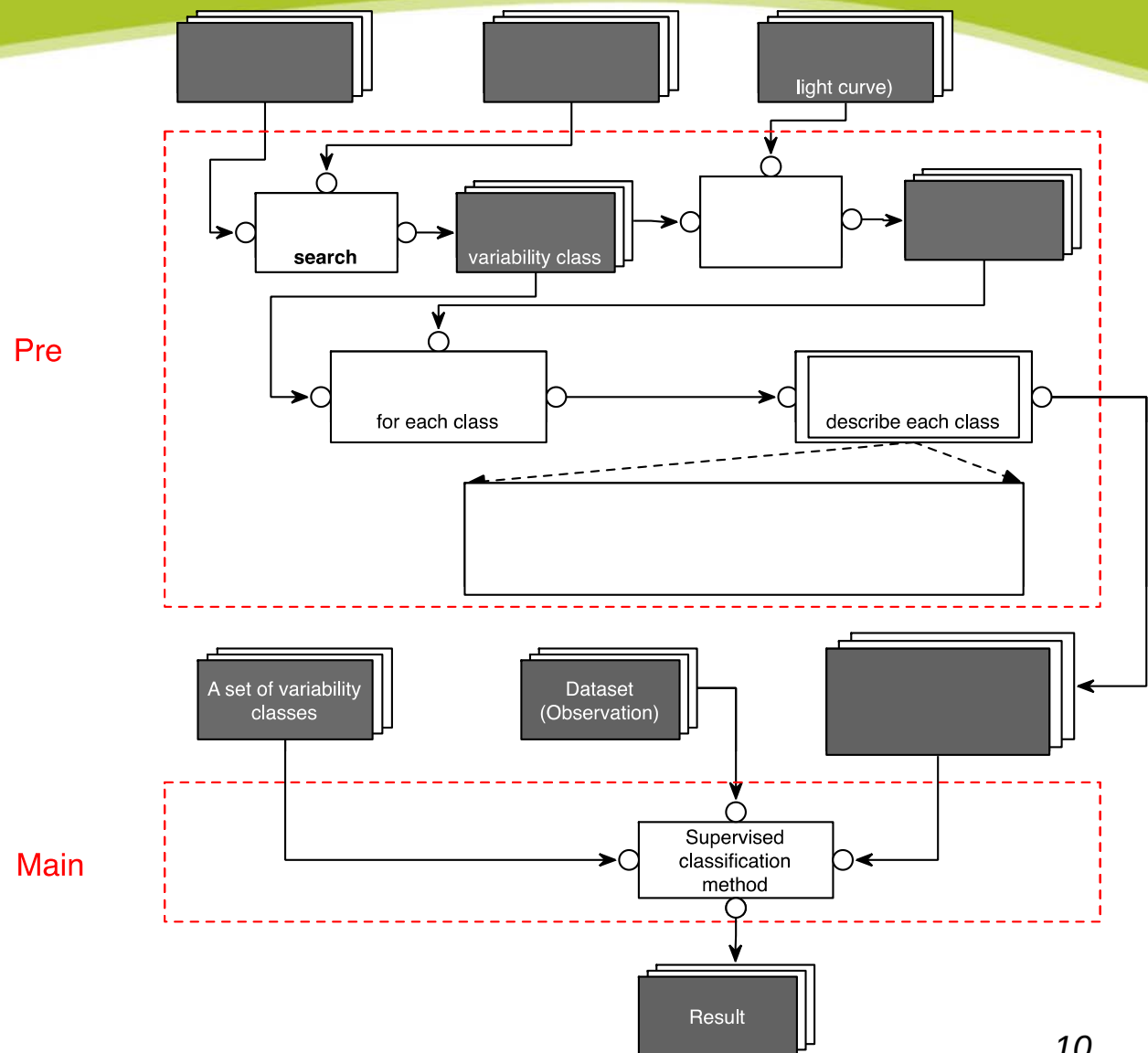
Les modèles de workflow (astro) (3)

D'après:

[3] Automated classification of variable stars in the asteroseismology program of the Kepler space mission. J.BLOMME and al.

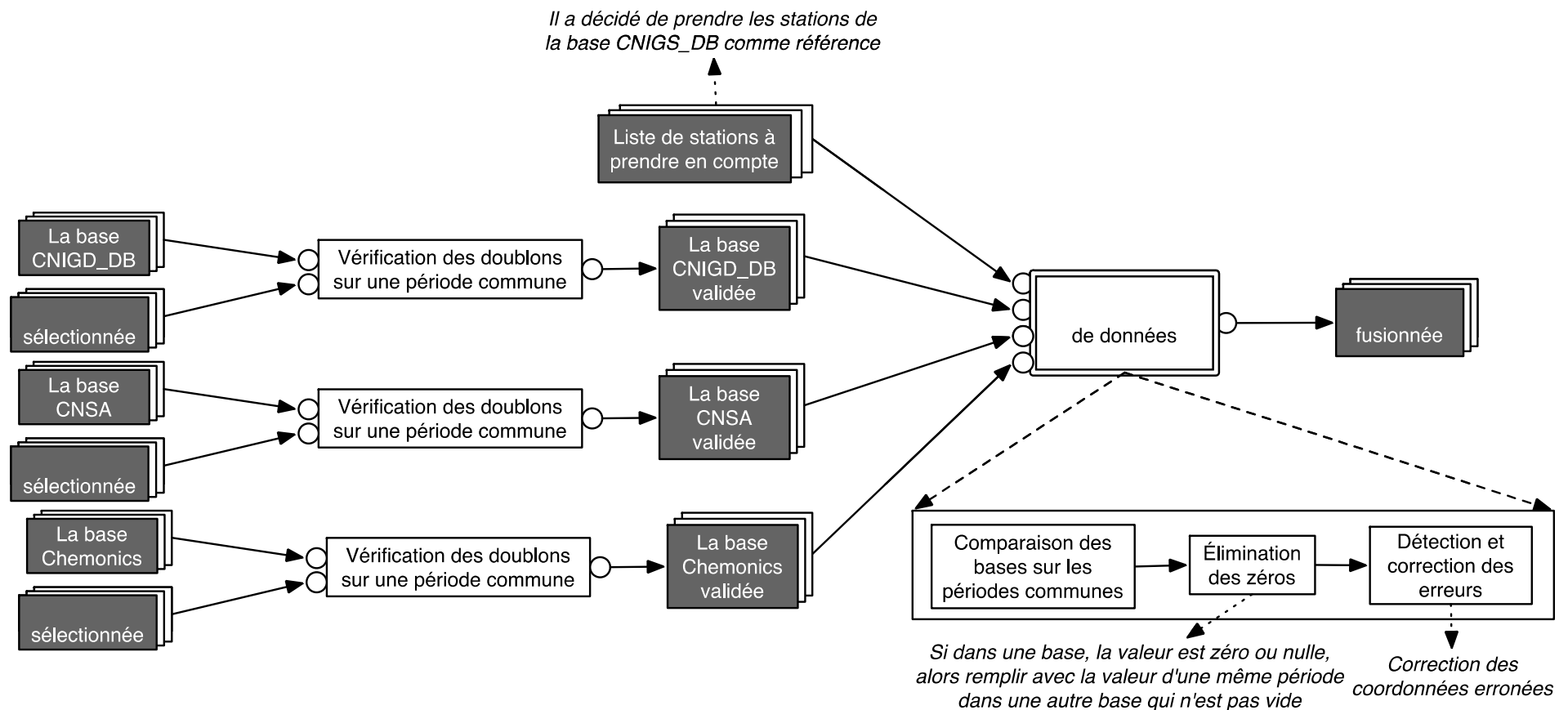
[4] Automated supervised classification of variable stars in the CoRoT programme. Method and application to the first four exoplanet fields. J.DEBOSSCHER and al.

[5] Automated supervised classification of variable stars I. Methodology. J.DEBOSSCHER and al.



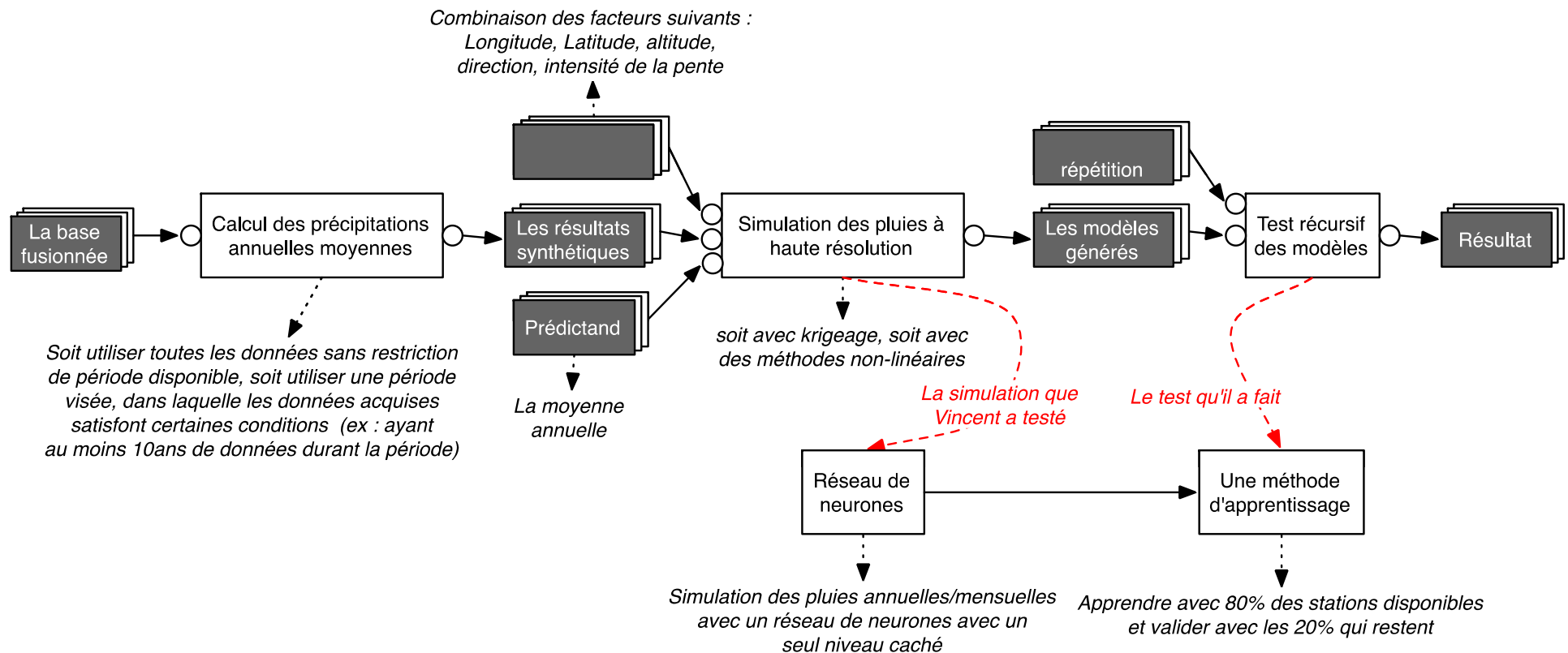
Les modèles de workflow (Climato) (4)

Concaténation des bases de données pluviométriques



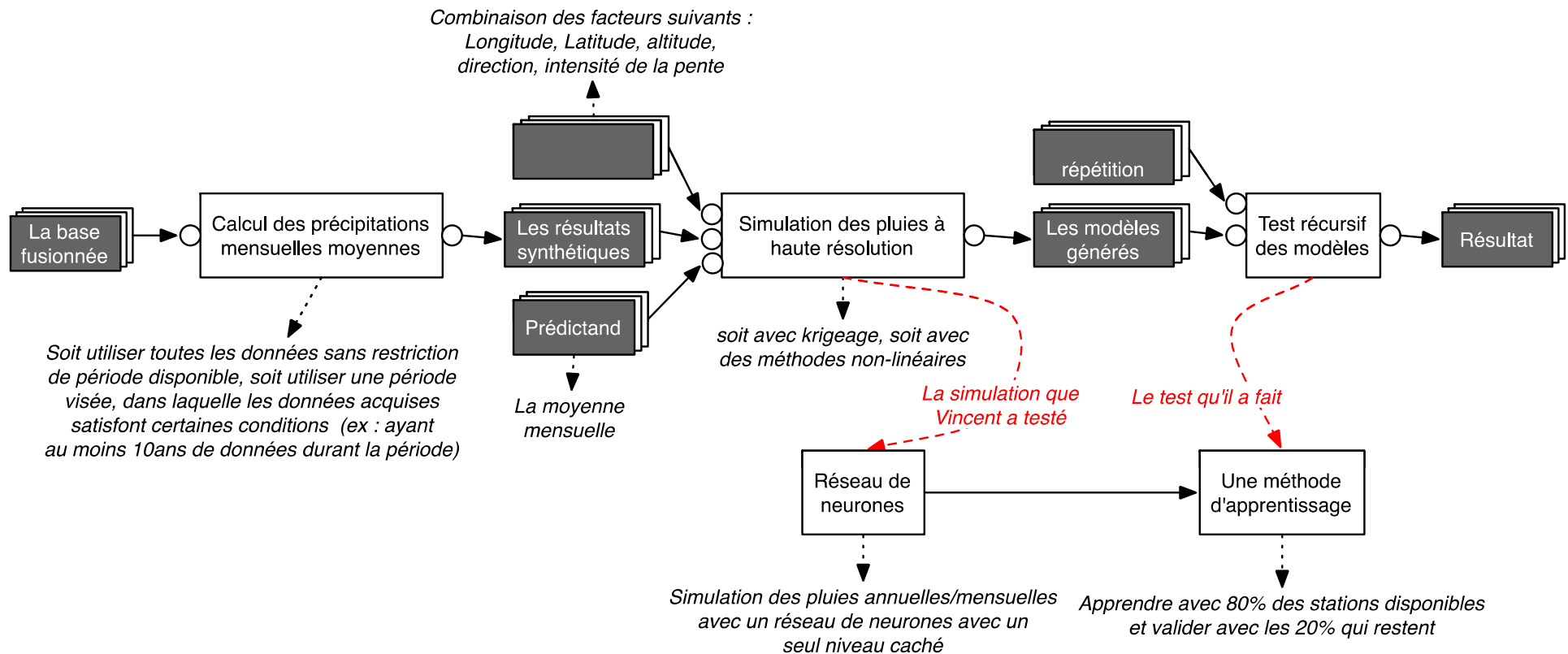
Les modèles de workflow (Climato) (5)

Analyse des précipitations annuelles



Les modèles de workflow (Climato) (6)

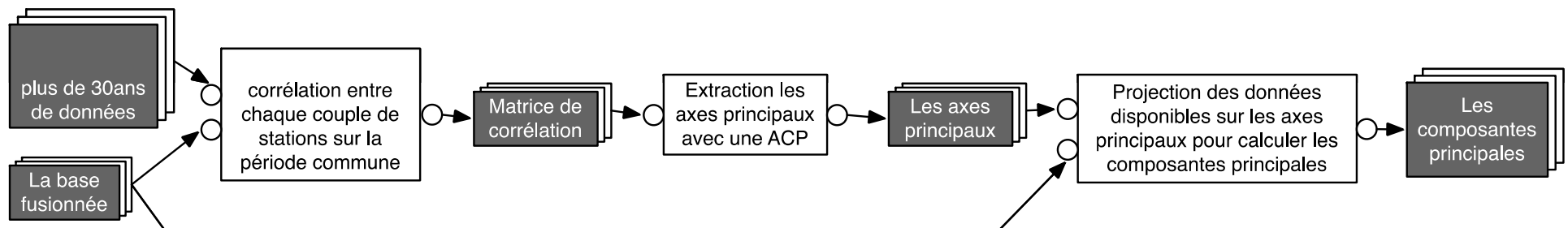
Analyse des précipitations mensuelles



Les modèles de workflow (Climato) (7)

Analyse des variabilité des pluies

Approche 1 : Considérer l'échelle annuelle qui est la plus intégrative



Approche 2 : Considérer une ACP "étendue" selon le modèle de Camberlin et al. (2003)

Approche 3 : Considérer de façon indépendante la variabilité interannuelle des totaux trimestriels glissants

Objectifs 2013 — ESPACE DEV

1. Compléter la modélisation des workflows
2. Générer les alternatives équivalentes
3. Mener une étude comparative des alternatives (coût, robustesse, précision)
4. Co-analyse sur les données climato

Etat de l'art: Positionnement

- État de l'art « prétraitement » et « préparation des données »
- État de l'art sur les workflows scientifiques
- Les outils existants (ETL, PMML-compliant):
knime, Zementis, PMML

Prétraitement et préparation

- Nettoyage des données
Remplacement de valeurs manquantes, erronées,
lissage détection et élimination des doublons,
incohérences et outliers
- Intégration des données
Intégration de multiple sources de données
- Transformation des données
Normalisation, discrétisation, aggrégation
- Réduction des données
Sélection de variables, résumés

References

Books

- BATINI, Carlo, SCANNAPIECO, Monica. Data Quality Concepts, Methodologies and Techniques. Data-Centric Systems and Applications. Springer-Verlag, 2006.
- BARNETT, V., LEWIS, T., Outliers in statistical data. John Wiley, Chichester, 1994.
- DASU, Tamraparni, JOHNSON, Theodore. Exploratory Data Mining and Data Cleaning. John Wiley, 2003.
- HAWKINS, D., Identification of Outliers. Chapman and Hall, London, 1980.
- HERZOG, Thomas N., SCHEUREN, Fritz J., WINKLER, William E., Data Quality and Record Linkage Techniques, Springer, May 2007.
- KIMBALL, Ralph, CASERTA, Joe. The Data Warehouse ETL Toolkit, Wiley, 2004.
- NAUMANN, Felix Quality-Driven Query Answering for Integrated Information Systems. Lecture Notes in Computer Science, vol. 2261. Springer-Verlag, 2002.
- Tukey, John Wilder. Exploratory Data Analysis. Addison-Wesley, 1977
- WANG, Richard Y., ZIAD, Mostapha, LEE, Yang W. Data Quality. Advances in Database Systems, vol. 23. Kluwer Academic Publishers, 2002.

Surveys

- CHANDOLA, Varun, BANERJEE, Arindam, KUMAR, Vipin, Anomaly Detection A Survey. ACM Computing Surveys, September 2009.
- ELMAGARMID, Ahmed K., IPEIROTIS, Panagiotis G., VERYKIOS, Vassilios S., Duplicate Record Detection A Survey, IEEE Transactions on Knowledge and Data Engineering (TKDE) Vol. 19 No. 1 January 2007, pp. 1-16.
- HELLERSTEIN, Joseph, Quantitative Data Cleaning for Large Databases. White paper, United Nations Economic Commission for Europe, February, 2008.
<http://db.cs.berkeley.edu/jmh/cleaning-unece.pdf>
- NAVARRO, Gonzalo. A Guided Tour to Approximate String Matching. ACM Comput. Surv., 33(1), pp. 31-88, 2001.
- WINKLER, William E., Overview of Record Linkage and Current Research Directions, Tech. Rep. of U.S. Census Bureau, February. 2006
<http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf>

References

Tutorials

- BATINI, Carlo, CATARCI, Tiziana, SCANNAPIECO, Monica. A Survey of Data Quality Issues in Cooperative Systems. Tutorial ER 2004.
- KOUDAS, Nick, SARAWAGI, Sunita, SRIVASTAVA, Divesh. Record Linkage Similarity Measures and Algorithms. Tutorial SIGMOD 2006.
- BANERJEE, Arindam, CHANDOLA, Varun, KUMAR, Vipin, SRIVASTAVA Jaideep, LAZAREVIC, Aleksandar. Anomaly Detection A Tutorial. Tutorial SIAM Conf. on Data Mining 2008.
- KRIEGL, Hans-Peter, KROGER, Peer, ZIMEK, Arthur. Outlier Detection Techniques. Tutorial, PAKDD 2009. http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/tutorial_slides.pdf

Data Profiling

- CARUSO, FRANCESCO, COCHINWALA, MUNIR, GANAPATHY, UMA, LALK, GAIL, MISSIER, PAOLO. 2000. Telcordia's Database Reconciliation and Data Quality Analysis Tool. Proc. VLDB 2000, pp. 615-618, 2000.
- DASU, TAMRAPARNI, JOHNSON, THEODORE, S. Muthukrishnan, V. Shkapenyuk, Mining Database Structure; Or, How to Build a Data Quality Browser, Proc. SIGMOD 2002.

Data Preparation

- HIPPE, J., GUNTZER, U., GRIMMER, U. Data Quality Mining - Making a Virtue of Necessity. Proc. Workshop DMKD 2001.
- LUBBERS, D., GRIMMER, U., JARKE, M. Systematic Development of Data Mining-Based Data Quality Tools. Proc. VLDB 2003, pp. 548-559, 2003.
- KLINE, R.B., Data Preparation and Screening, Chapter 3. in Principles and Practice of Structural Equation Modeling, NY Guilford Press, pp. 45-62, 2005.
- PEARSON, Ronald K. Surveying Data for Patchy Structure. SDM 2005.
- STATNOTES Topics in Multivariate Analysis. Retrieved 10/17/2008 from <http://www2.chass.ncsu.edu/garson/pa765/statnote.htm>

Préparation des données

- Fusion/Intégration des sources de données
 - Structurée / Non-structurée / Semi-structurée
 - Locale / Distribuée
 - Hétérogénéité
- Nettoyage
 - Élimination des doublons
 - Détection et correction des erreurs
 - Valeurs manquantes

Préparation des données

- D'autres traitements nécessaires *[PG]*
 - Sélectionne des données
 - Échantillonnage
 - Sélection de sources
 - Réduction dimensionnalité
 - Sélection ou transformation d'attributs
 - Pondération
 - Codage
 - Agrégation, discrétisation, codage des attributs discrets, uniformisation d'échelle ou standardisation

PMML (Predictive Model Markup Language)

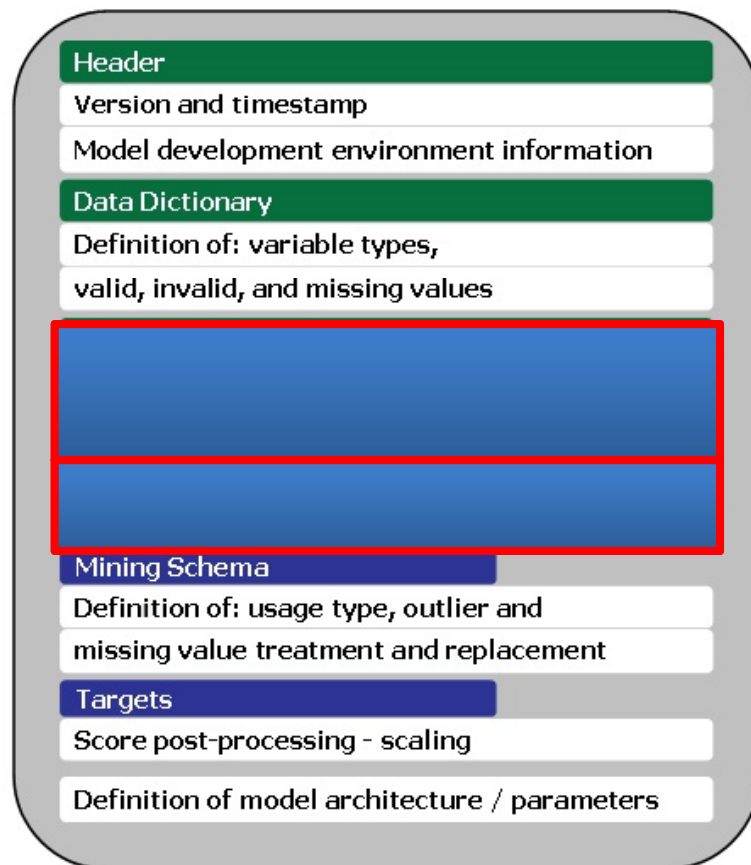
<http://www.dmg.org>



- DMG standard, Open source
- Objectif : Définir les modèles, faciliter le partage et la validation ces modèles entre partenaires
- Fichier XML contenant un ensemble de descriptions des :
 - Pré traitements (transformation, standardisation, etc.)
 - Modèles de fouille de données
 - Post traitements (évaluation, scaling, décision, etc.)

PMML (Predictive Model Markup Language)

- Structure générale



```

<AssociationModel column="longForm">
  <AssociationModelPair field="gender" column="shortForm"/>
</AssociationModel>
<ClusteringModel>
  GeneralRegressionModel
  MiningModel m </shortForm>
  MiningBayesModel </longForm>
  Normalisation
  NearestNeighborModel
  Discretisation
  NeuralNetwork
  Mapping des valeurs
  RegressionModel <shortForm>
  Functions
  RuleSetModel female </longForm>
  Aggregation
  SequenceModel
  Secret
  <Secret>
  </Secret>
  SupportVectorMachineModel
  TextModel
  TimeSeriesModel
  TreeModel
    
```

Les outils existants PMML-compliant

- Ad hoc Script
- Talend : outil ETL <http://www.talend.com>
- WEKA Knowledge Flow : Préparation, fouille, visualisation, validation <http://www.cs.waikato.ac.nz/ml/weka>
- KNIME : Utilisation des fonctions d'R et tous les modules d'analyse de Weka <http://www.knime.org>
- Zementis (The Transformations Generator) : phase préparatoire des données avec une interface web, exportation en PMML <http://www.zementis.com>

PMML (Predictive Model Markup Language)

- Outils (<http://www.dmg.org/products.html>)

Outil	Export	Import	Information
Zementis Transformations Generator	PMML 4.1		Pré traitements
R	PMML 3.2/4.1		Librairie PMML
KNIME	PMML 4.0	PMML 4.0	Fonctions d'R et des modules de WEKA
WEKA		PMML 3.2	Nombreux modules d'analyse et de fouille
IBM InfoSphere Warehouse V9.7	PMML 3.0/3.2	PMML 2.0	Payant

AMADEUS – Dépenses depuis l'attribution du budget à ESPACE DEV

Montant alloué : 15 000 euros

Dépenses : 14 917 euros

- **Frais de gestion IRD :** 1 363 euros
- **Missions passées :** 580 euros
- Mission prévue (5 déc.) :** ± 500 euros
- Participation achat Licence Matlab :** 800 euros
- **CDD 3 mois de Yuan Lin** 11 674 euros

Mission Sept.-Nov. 2012: Formaliser les chaînes de traitements
3 891,40 euros/mois (salaire + charges)