# Data Mining
# Algorithms and Applications

**Eduardo Ogasawara**

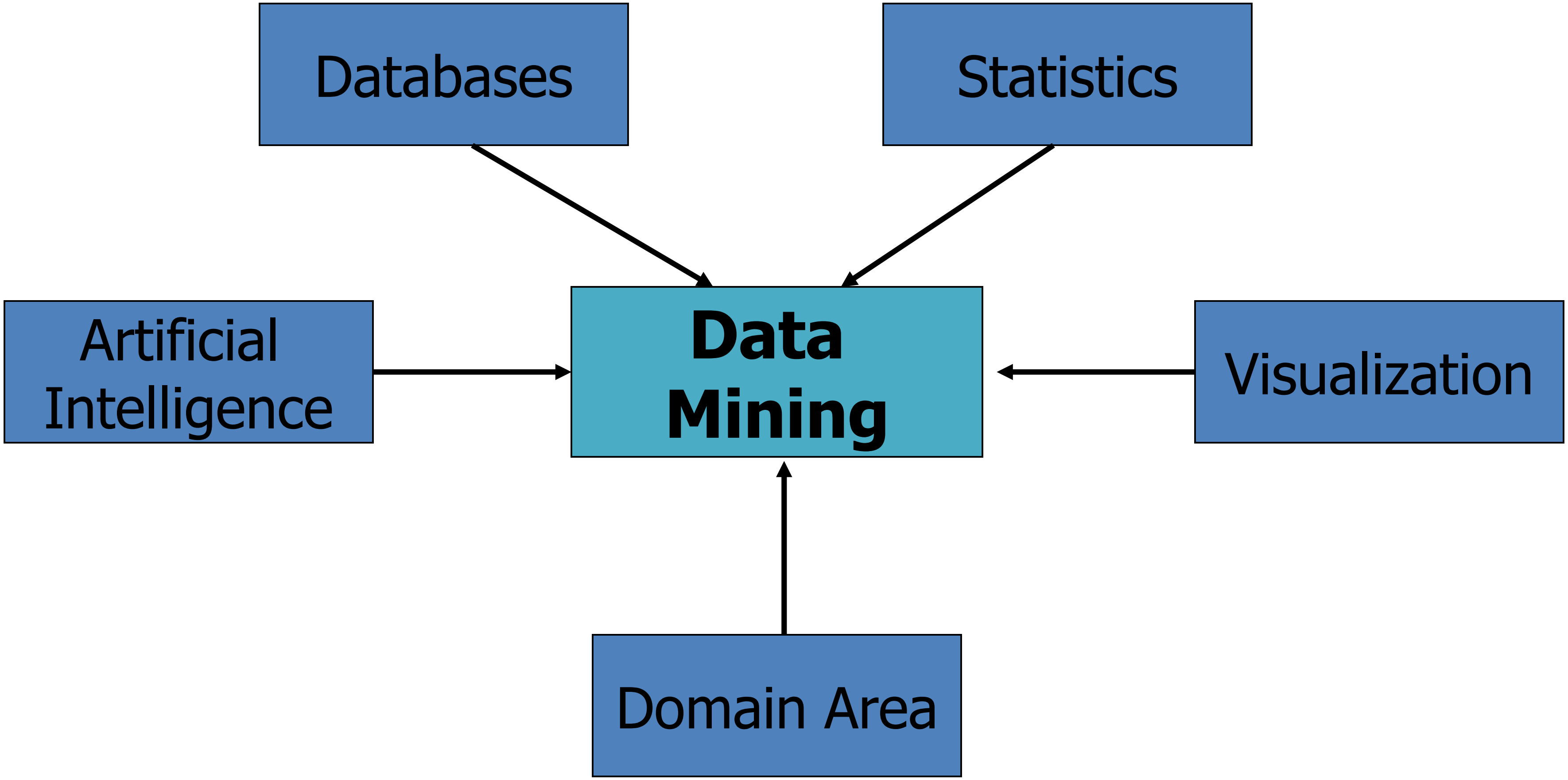
**eogasawara@cefet-rj.br**

**Centro Federal de Educação Tecnológica**
**Celso Suckow da Fonseca**
**CEFET/RJ**

# Data Mining

- **Extraction knowledge (rules, patterns, constraints) from "large" databases**
- **Data Deluge & Big Data introduce more complexity**
  - **We are immersed in data, but hungry for knowledge!**

# *Overview: Confluence of various disciplines*



Databases

Statistics

Artificial Intelligence

Data Mining

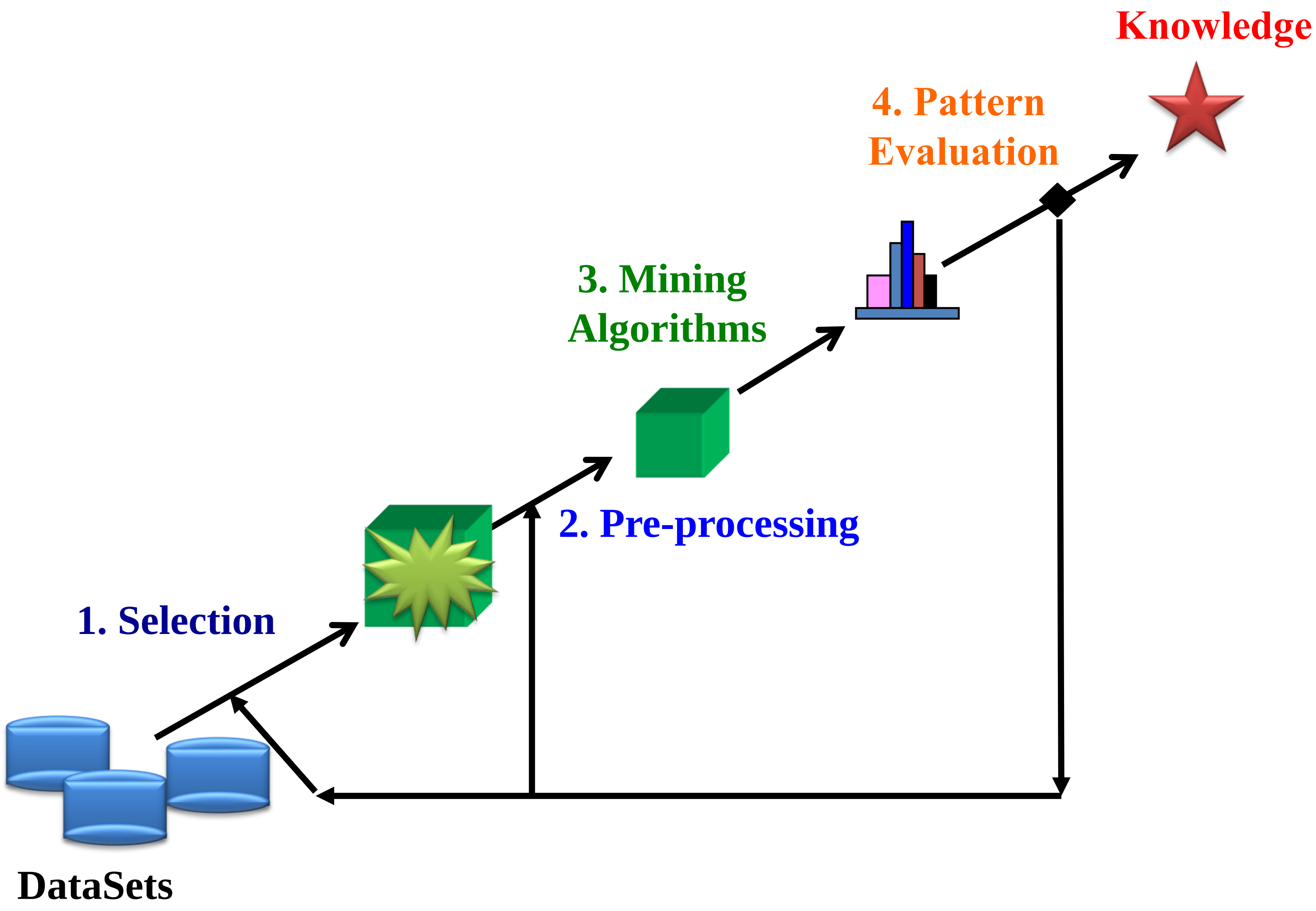
Visualization

Domain Area

# *Goal: Answer key questions*

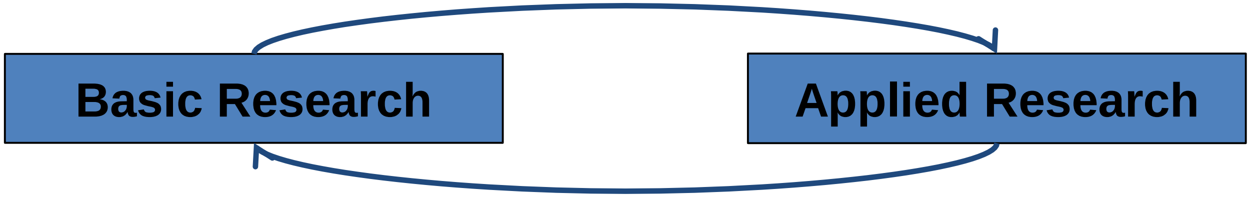
- **Consider an example of a internet sales system**
- **Do we have the data?**
  - **Transactions involving credit cards, loyalty cards, central customer service, lifestyle customer**
- **Can we identify groups?**
  - **Identified clusters for "models" of customers who share certain characteristics: interest, income, spending habits**
- **Can we classify or predict?**
  - **Identified patterns of consumption?**
  - **Can establish a pattern of consumption for single and married?**
- **Can we associate patterns?**
  - **Association / correlation between product sales**
  - **Prediction based on association information**

# *Data Mining Process (DMP)*



# DMAA - Research Project

- **Goal**
  - **Develop Data Mining Algorithms and Application for practical problems**
- **Research Process**
  - **Basic Research**
    - **Produce data knowledge extraction algorithms or tools without immediate application for real problems**
  - **Applied Research**
    - **Usage of know or developed algorithms or tools for solving real problems**



# *Basic Research*

- **Data Management & Data Mining Process**
  - **Data Streaming**
  - **Data Sources Location**
  - **Uniform Data Model**
  - **Data Experiment Representation**
  - **Modeled through Workflows**
- **Algorithms and Methods**
  - **Research on specific Data Mining Activities**

# *DMP: Step #1 – Selection and Integration*

- Types of data
  - Relational, Spatial, Temporal, Textual, Graph
- Sources
  - Centralized
  - Distributed
  - Streaming
- Integration
  - Data integration from different sources
- Dataset Production
    - Projection
    - Selection

# *DMP: Step #2 – Pre-processing*

- Very important step
  - 60% of Data Mining effort!
- Data Reduction
  - Filtering or Sampling
- Attribution Reduction
  - Selection of important attributed
  - Reduction of Dimensionality
- Transformation
  - Combination
  - Derivation
  - Aggregation

# *DMP: Step #3 – Data Mining Algorithms and Methods*

- **Focus: Searching for patterns of interest**
- **Data Mining Methods**
  - **Clustering**
  - **Classification and Prediction**
  - **Association**
  - **New Approaches**

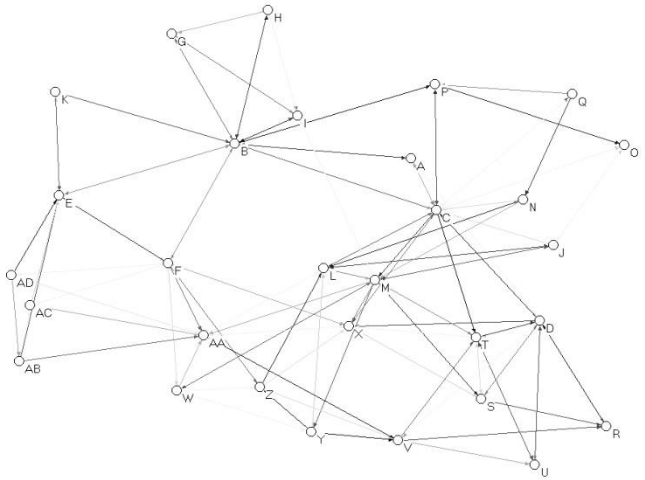
# *DMP: Step #4 – Pattern Evaluation*

- **Error Measurements**
  - **Error Metrics**
- **Evaluation of patterns and extracted knowledge**
  - **Visualization**
  - **Transformation**
  - **Redundant pattern removal**
- **Use of discovered knowledge**
- **Are all discovered patterns important?**

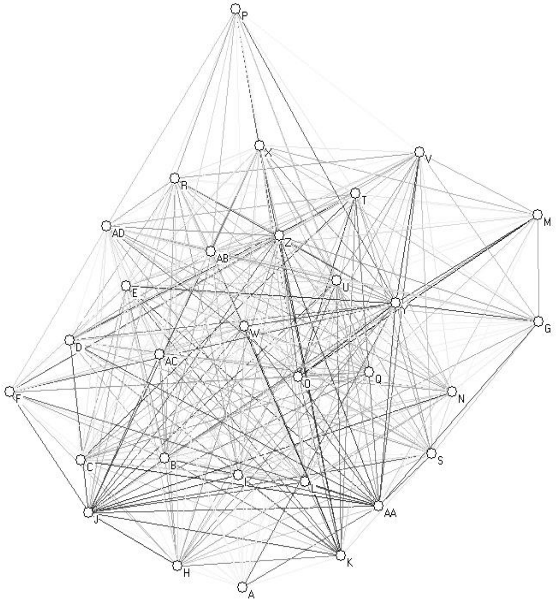
# *Applied Research*

- **Focus on solving target application areas (datasets)**
- **Basic domain of target area**
  - **data**
  - **Problem**
  - **Identification of goals**
- **Examples**
  - **Astronomy**
  - **Production engineering**
  - **Social Networks**
  - **Computing in Educational**

# *Evaluation of Enterprise Social Networks Adoption*

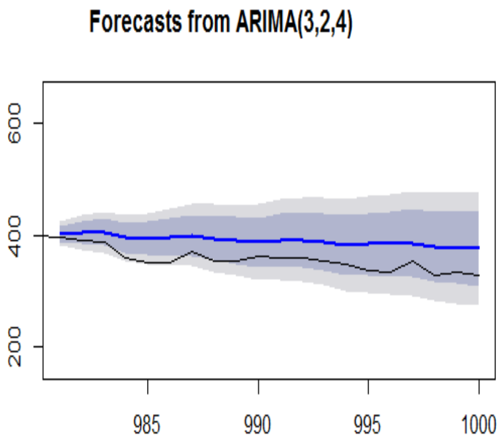


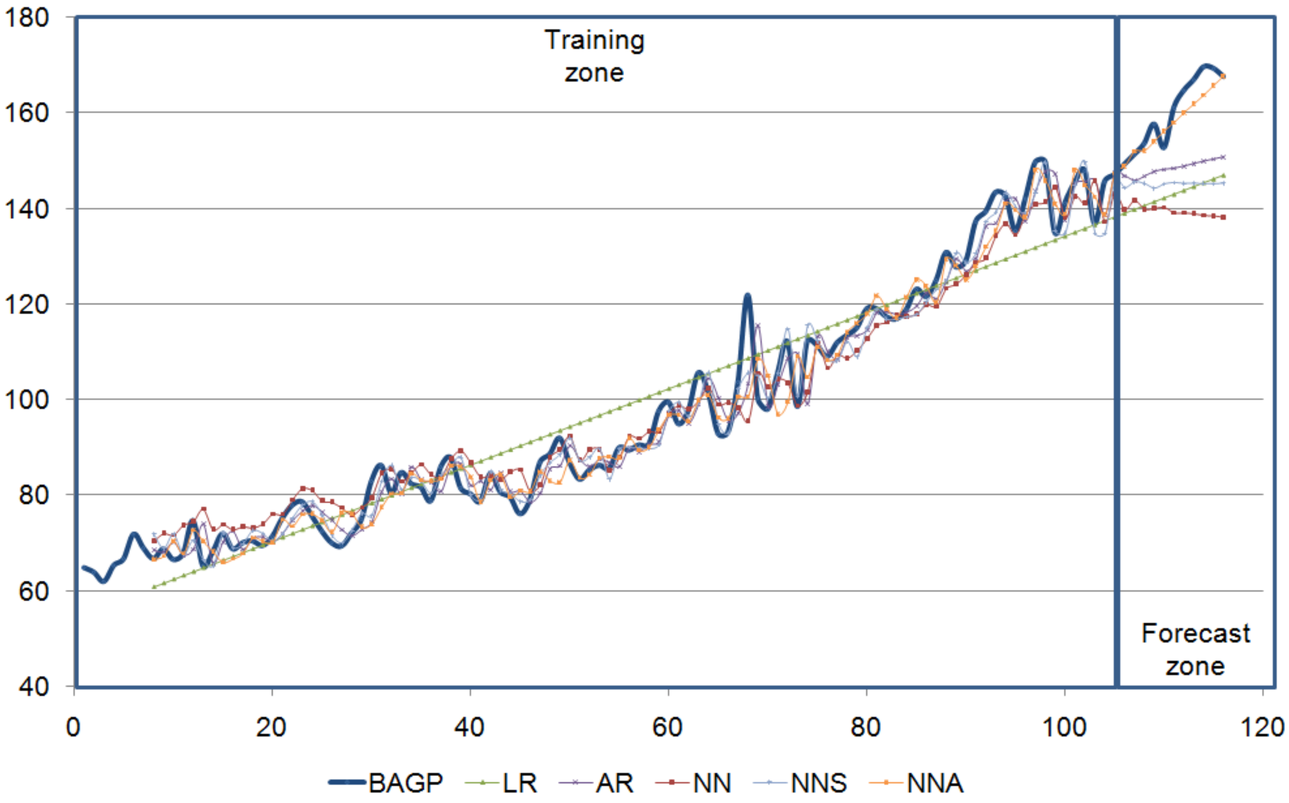
(GC)



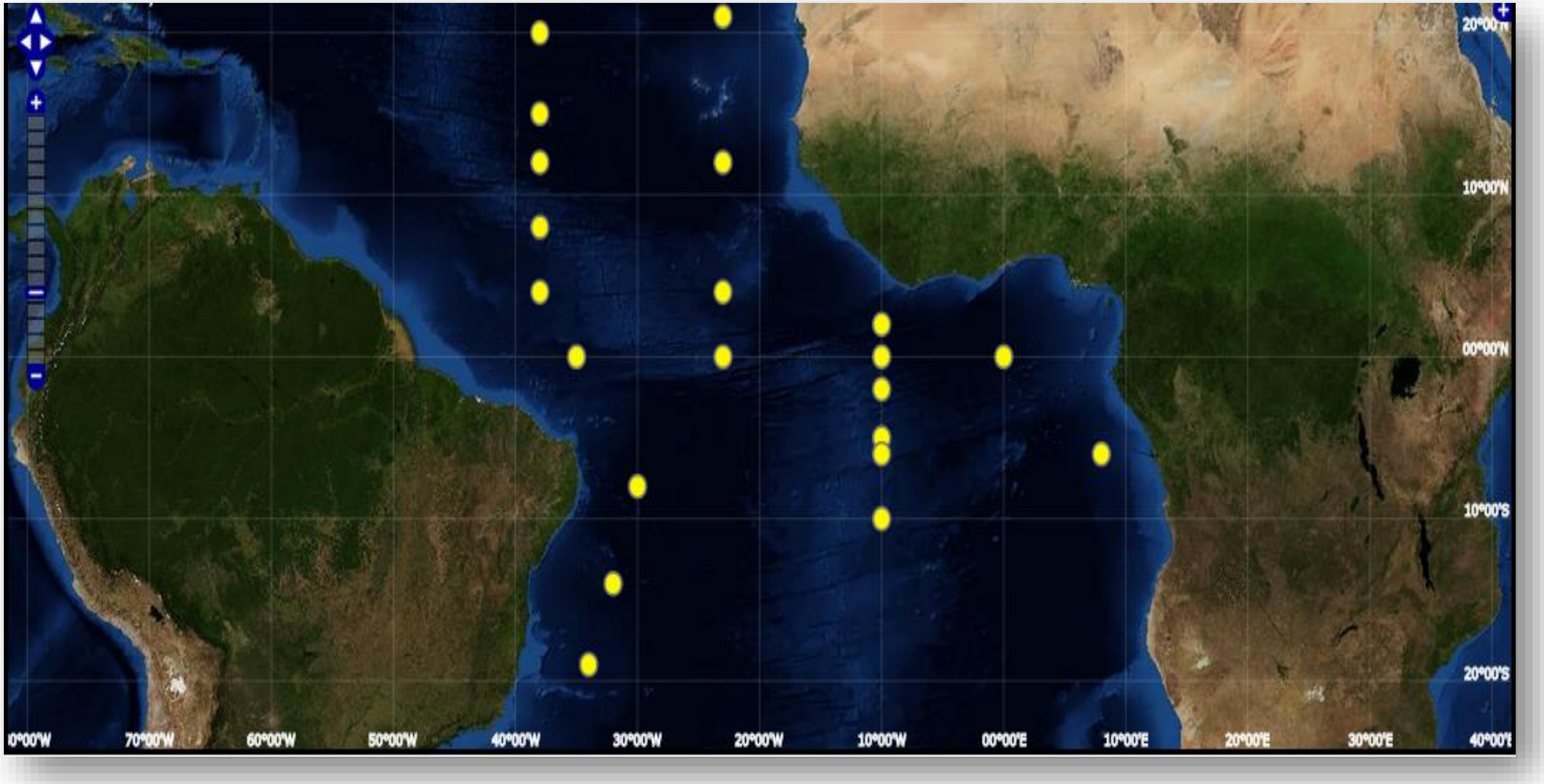
(GR)

# Time Series Prediction





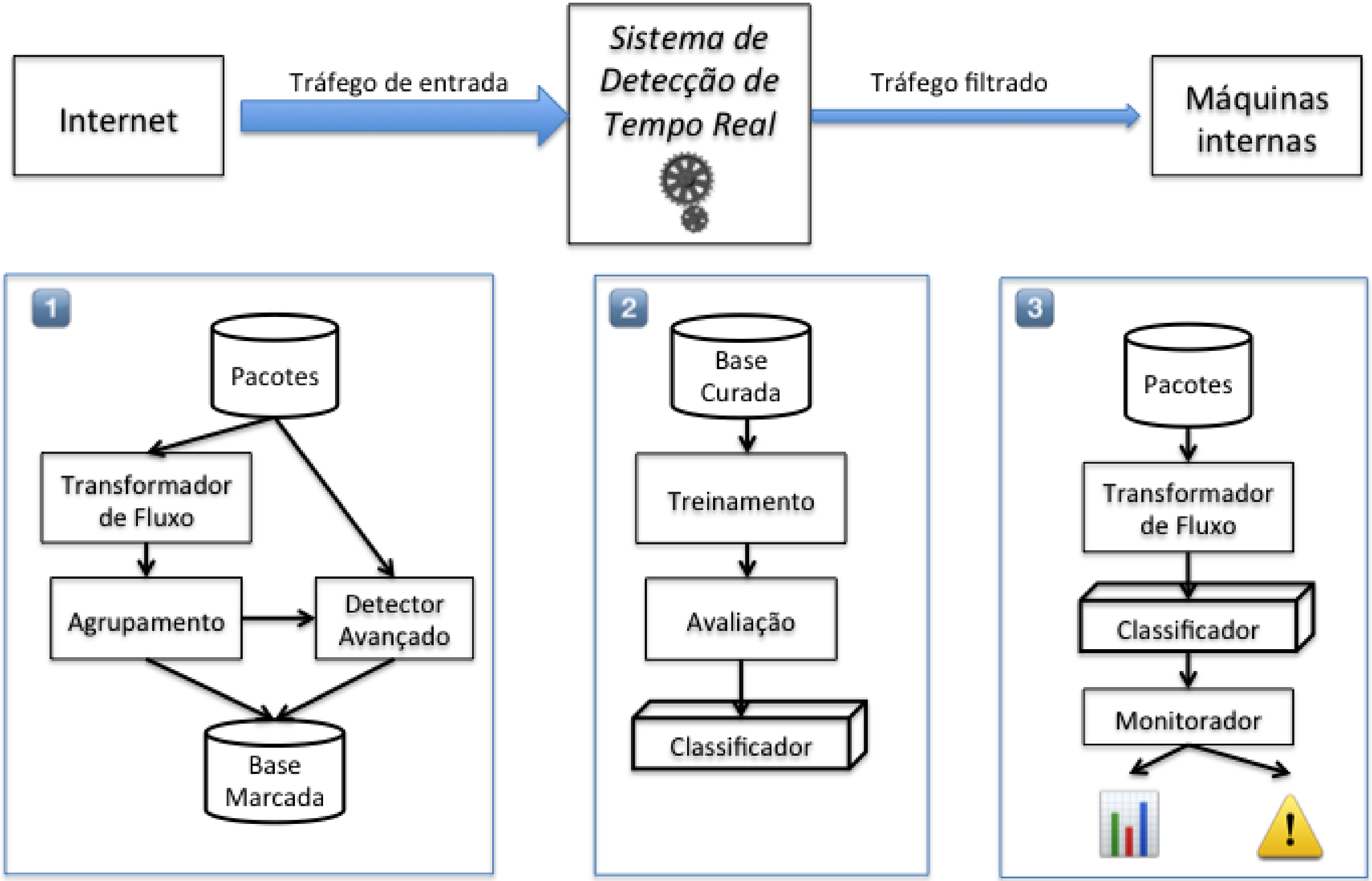
# *Forecast Sea Surface Temperature in Atlantic Pacific Ocean*



# *Identification of Stars and Gallaxies*

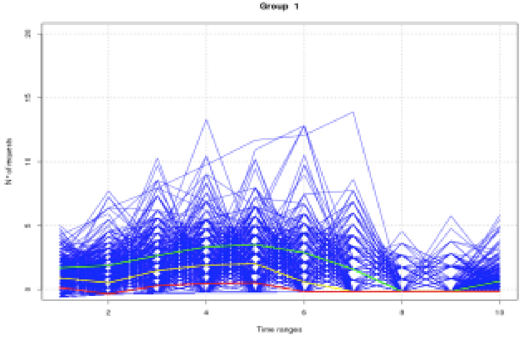
**NGC 7793: a spiral galaxy**

# *Production of Access Flow Dataset for Identification of Deny of Service (DoS) Attacks*

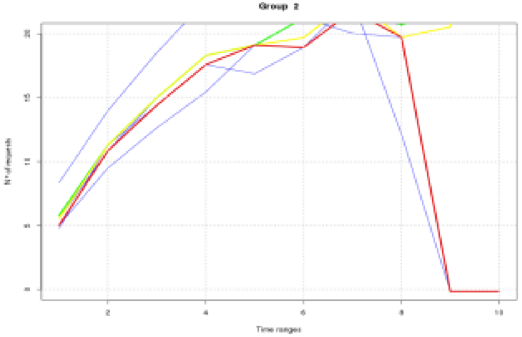


# *Clustering of Access Flow Dataset (Clustering of Time Series)*

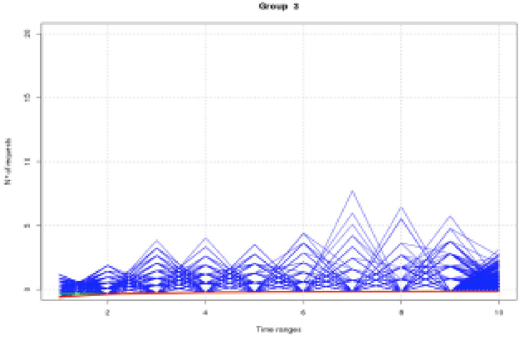
Grupo 1



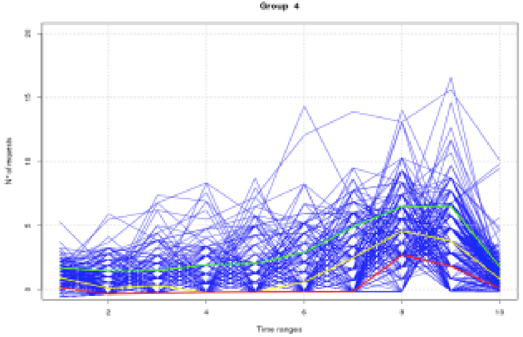
Grupo 2



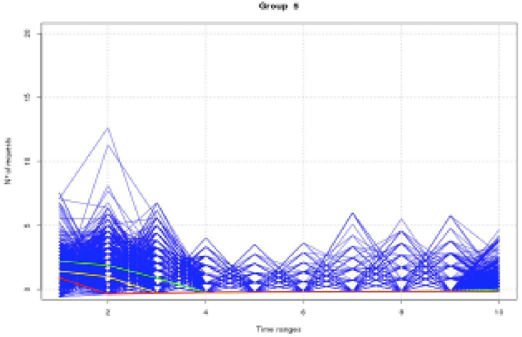
Grupo 3



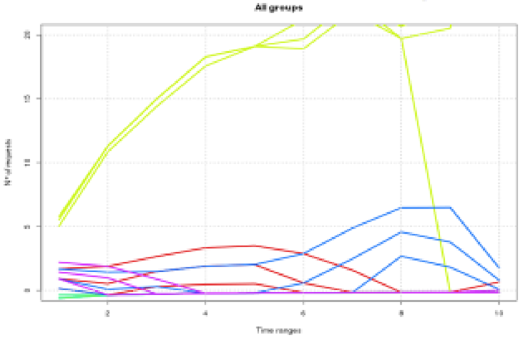
Grupo 4



Grupo 5

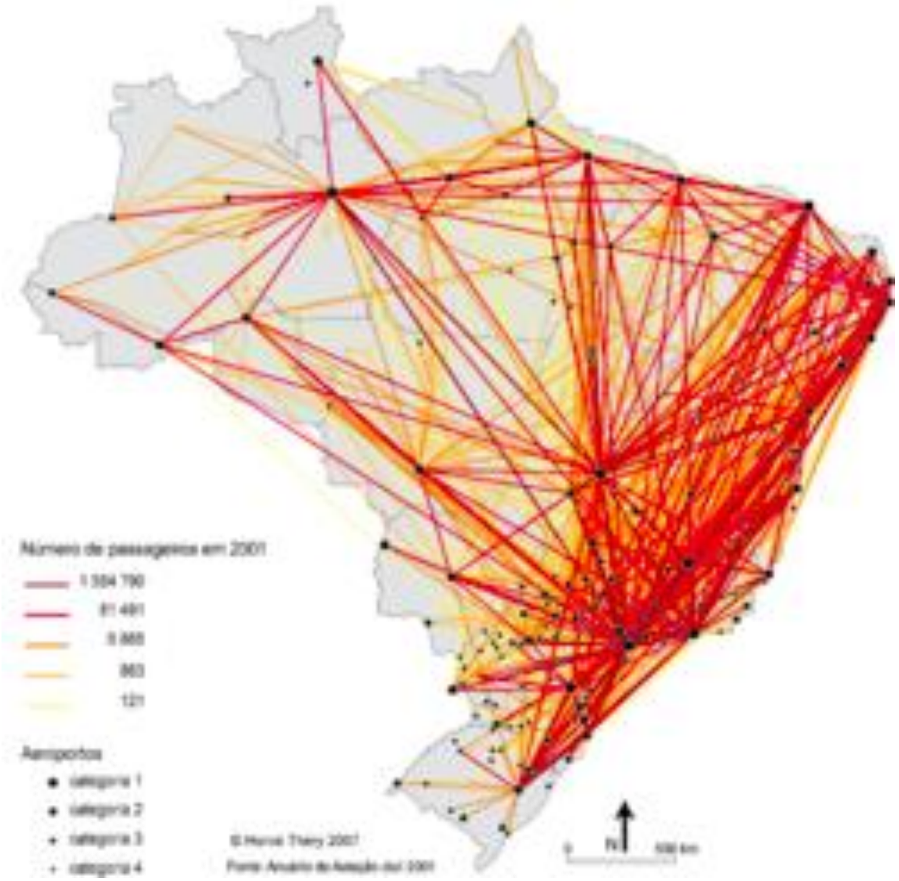


All medians and Qs



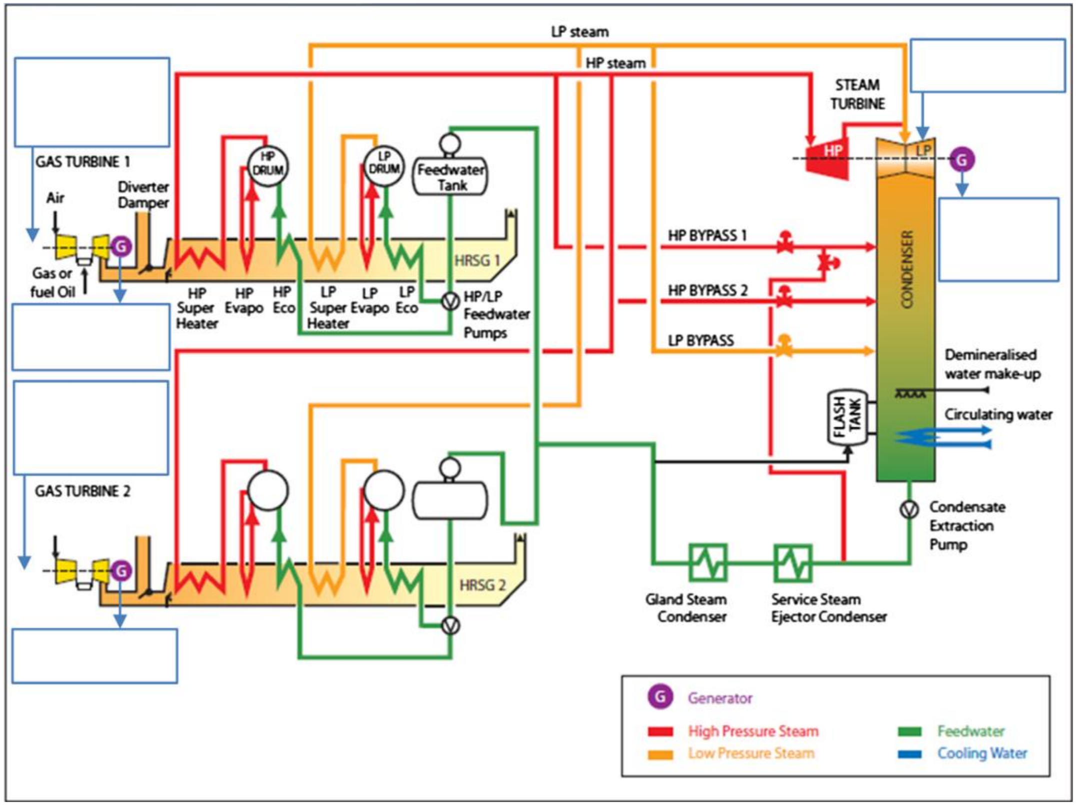
# *Forecasting Delays in Airports*

Fluxos de passageiros



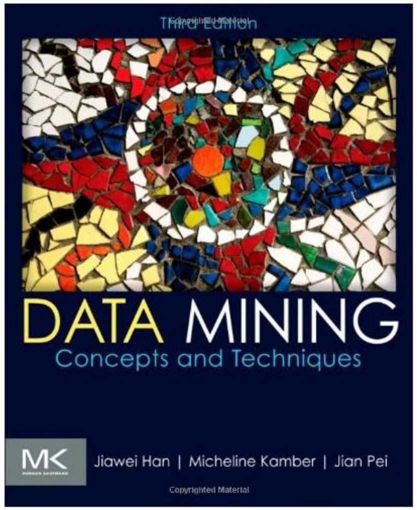
© Hervé Théry 2007
Fonte: Anuário do Aviação civil 2001

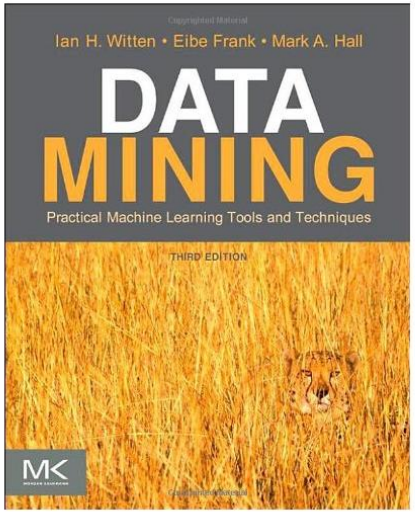
# *Prediction of electrical power output of a base load operated combined cycle power plan*

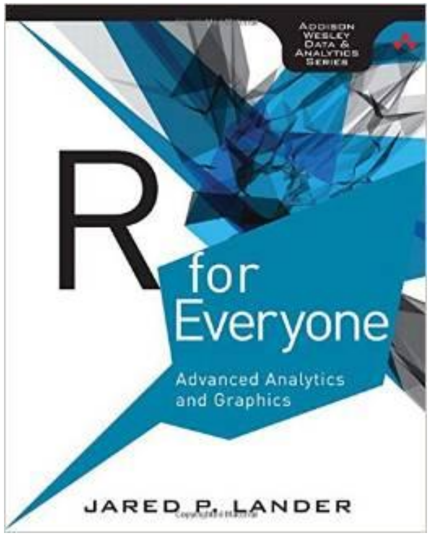


**Tüfekci, P. 2014**

# *Referências*




Copyrighted Material
Ian H. Witten • Eibe Frank • Mark A. Hall
DATA MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK


Addison Wesley Data & Analytics Series
R for Everyone
Advanced Analytics and Graphics
JARED P. LANDER