

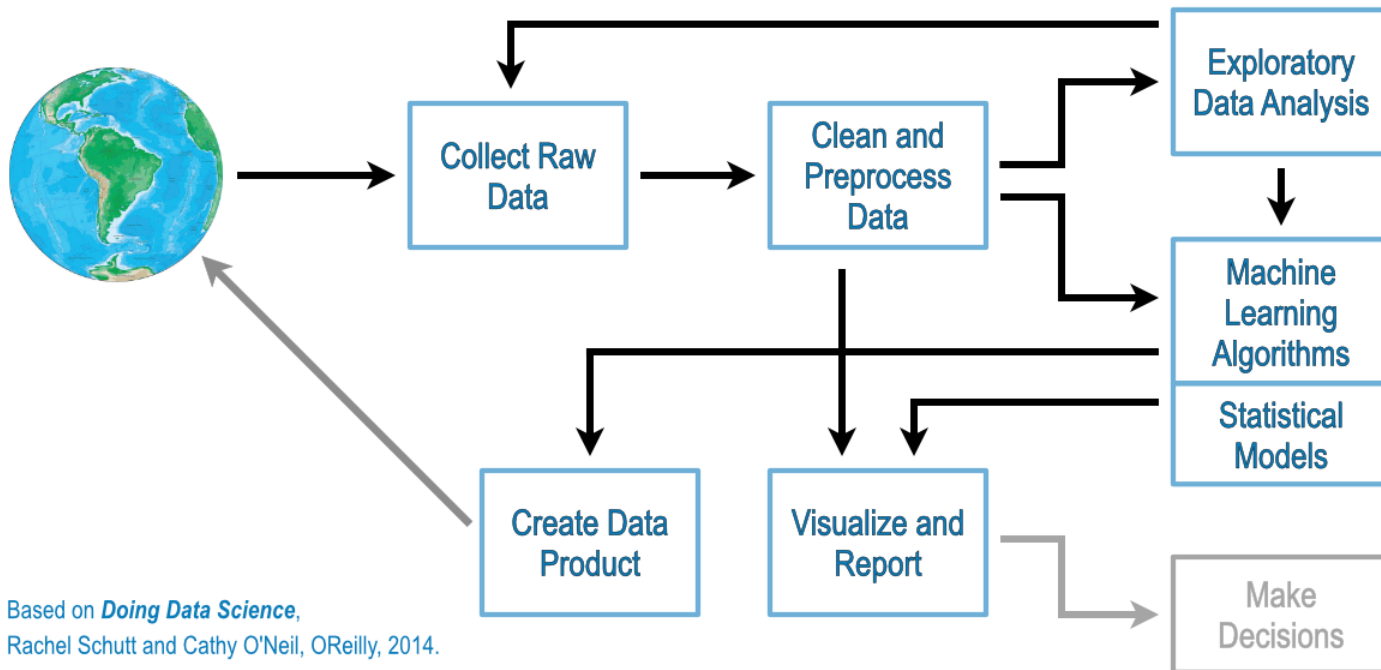
# CAP-359 PRINCIPLES AND APPLICATIONS OF DATA MINING

Rafael Santos – [rafael.santos@inpe.br](mailto:rafael.santos@inpe.br)  
[www.lac.inpe.br/~rafael.santos/](http://www.lac.inpe.br/~rafael.santos/)

Updated in 2019

# About

- What is Data Mining? What is it good for?
- Examples of Data Mining algorithms and applications.



# About: Resources

- [www.lac.inpe.br/~rafael.santos/cap359.html](http://www.lac.inpe.br/~rafael.santos/cap359.html)
  - Schedule, presentations, etc.
- Meetings about the projects: ask beforehand ([rafael.santos@inpe.br](mailto:rafael.santos@inpe.br))

## Principles and Applications of Data Mining

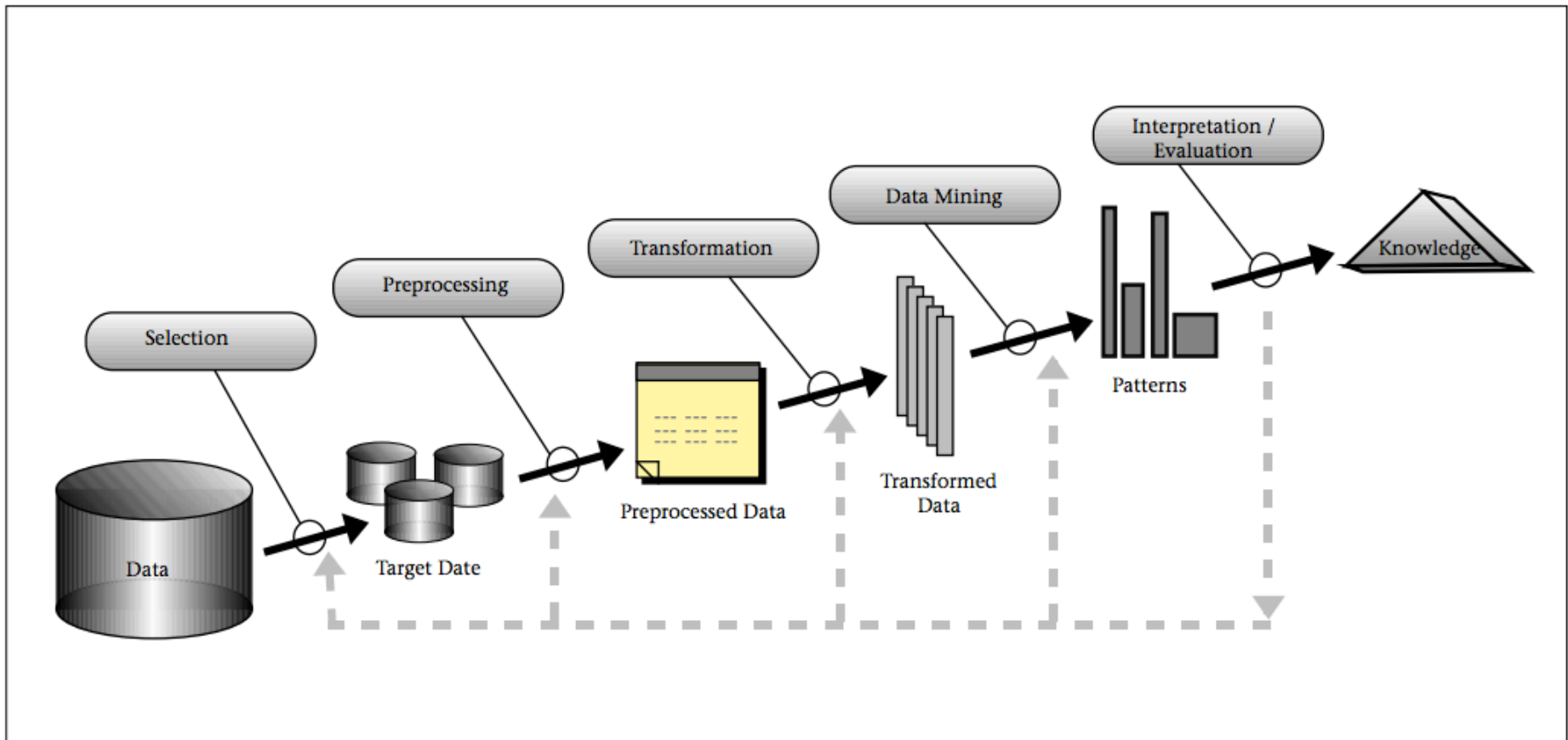
# What is Data Mining?

# What is Data Mining?

- First, what is *Knowledge Discovery in Databases (KDD)*?
  - “... the overall process of discovering useful knowledge from data”.
  - KDD is the nontrivial process of identifying **valid**, **novel**, **potentially useful**, and **ultimately understandable** patterns in data.

# What is Data Mining?

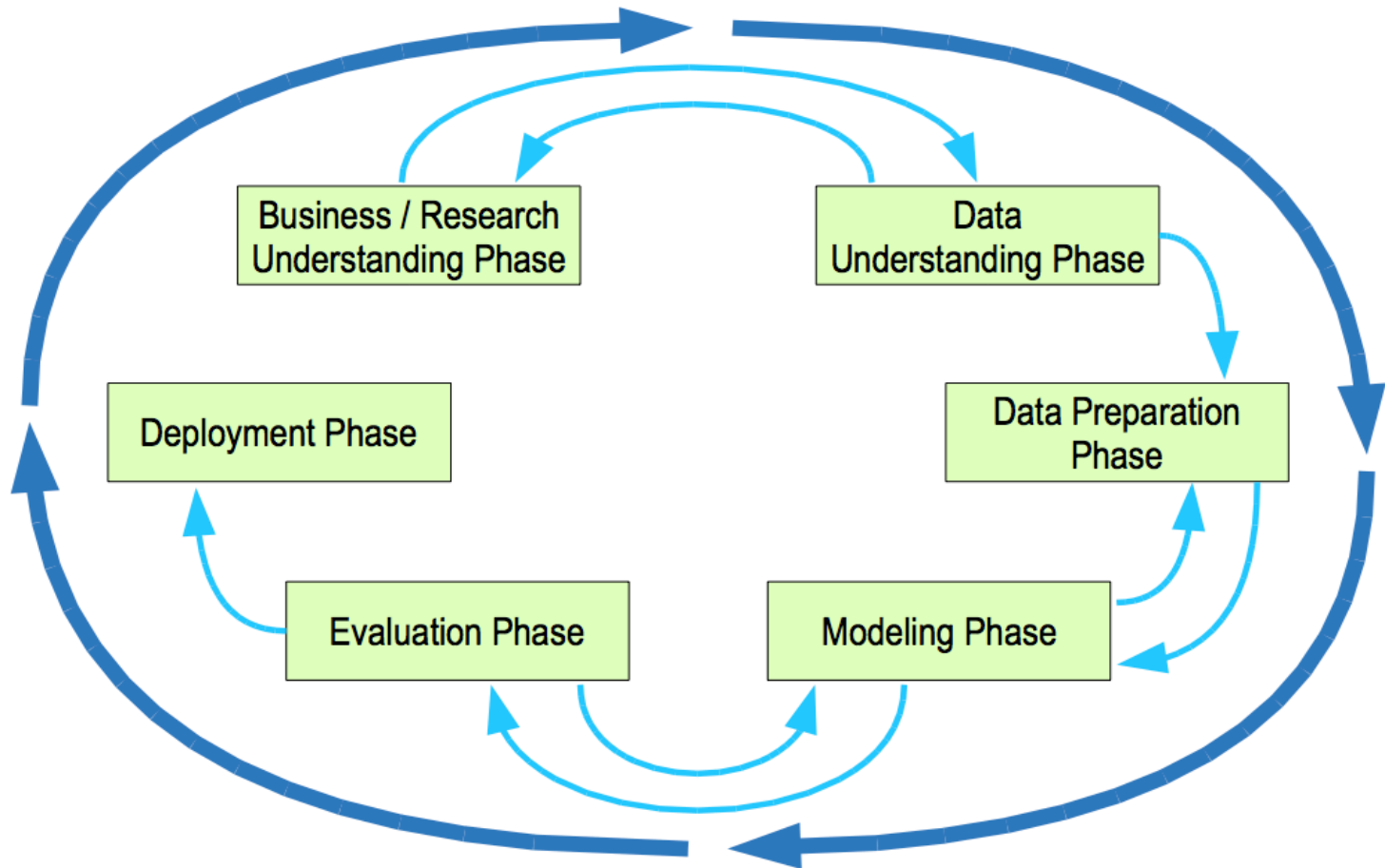
- Data Mining is a part of the KDD process:



From *Data Mining to Knowledge Discovery in Databases*; Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth, *AI Magazine*, 1996.

# What is Data Mining?

- CRISP-DM framework: Cross-Industry Standard Process for Data Mining.



# What is Data Mining?

- For most purposes, Data Mining = KDD.
  - Focus on algorithms.
  - We assume that data is already selected, filtered, transformed, preprocessed, normalized, etc.
- Input: set of data, knowledge about the data.
- Processing: application of algorithms, evaluation.
- Output: rules, classes, groups, metadata, etc.
- Post-processing: evaluation, repeat if needed.



# Data Mining Tasks

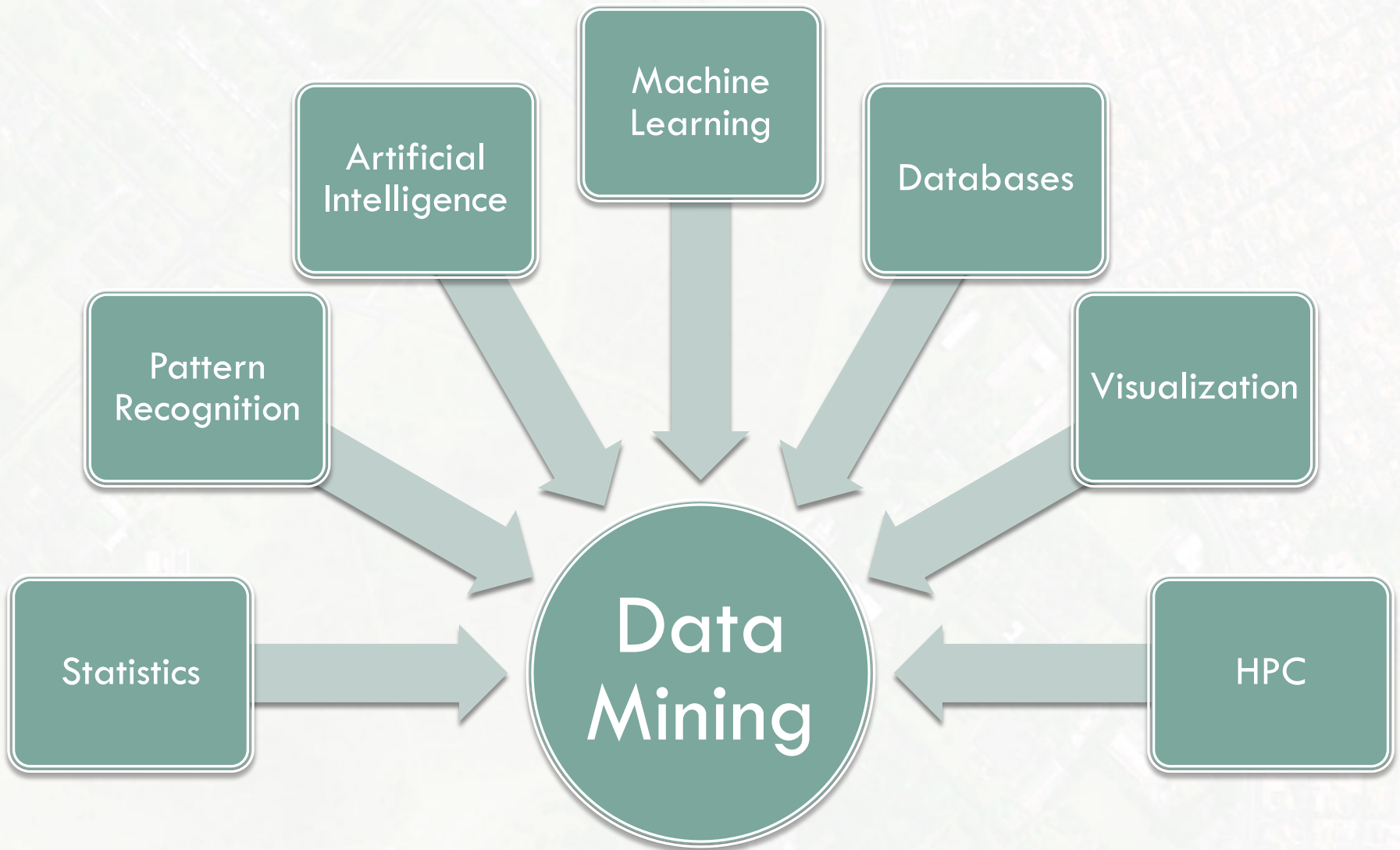
- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.
- Exploration
  - Visualization: use tools and techniques that help identify patterns

# Data Mining Tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]
- Visualization [Exploratory]
- Exploratory Data Analysis [Exploratory]

Based on *Introduction to Data Mining*; Pang-Ning Tan, Michael Steinbach and Vipin Kumar (2005).

# What is Data Mining?

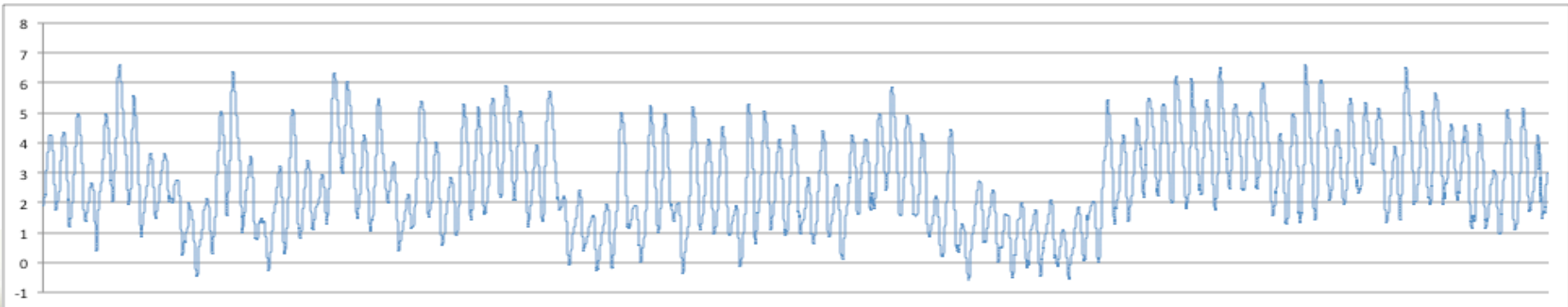


# What is **not** Data Mining?

- Look up a number in a phone directory
  - .. But match features/information from phone directories to geographic regions is!
- Query a search engine about “Amazon”
  - .. But group documents about amazon.com and Amazon rainforest based on context is!
- Segment a database into two explicit categories
  - .. But find implicit categories, groups and outliers is!

# What is Scientific Data Mining?

- Same as Data Mining, but...
  - Applications inherently scientific.
  - Emphasis on understanding the data (and the phenomena it represents): not necessarily *actionable results*.
  - Related to automated scientific discovery.
- Harder than plain DM?
  - Data representation can be different, interpretation of results require knowledge about the phenomena.



## Principles and Applications of Data Mining

# A very quick introduction to Data Science

<http://www.lac.inpe.br/~rafael.santos/cap394.html>

# What is a Data Scientist?

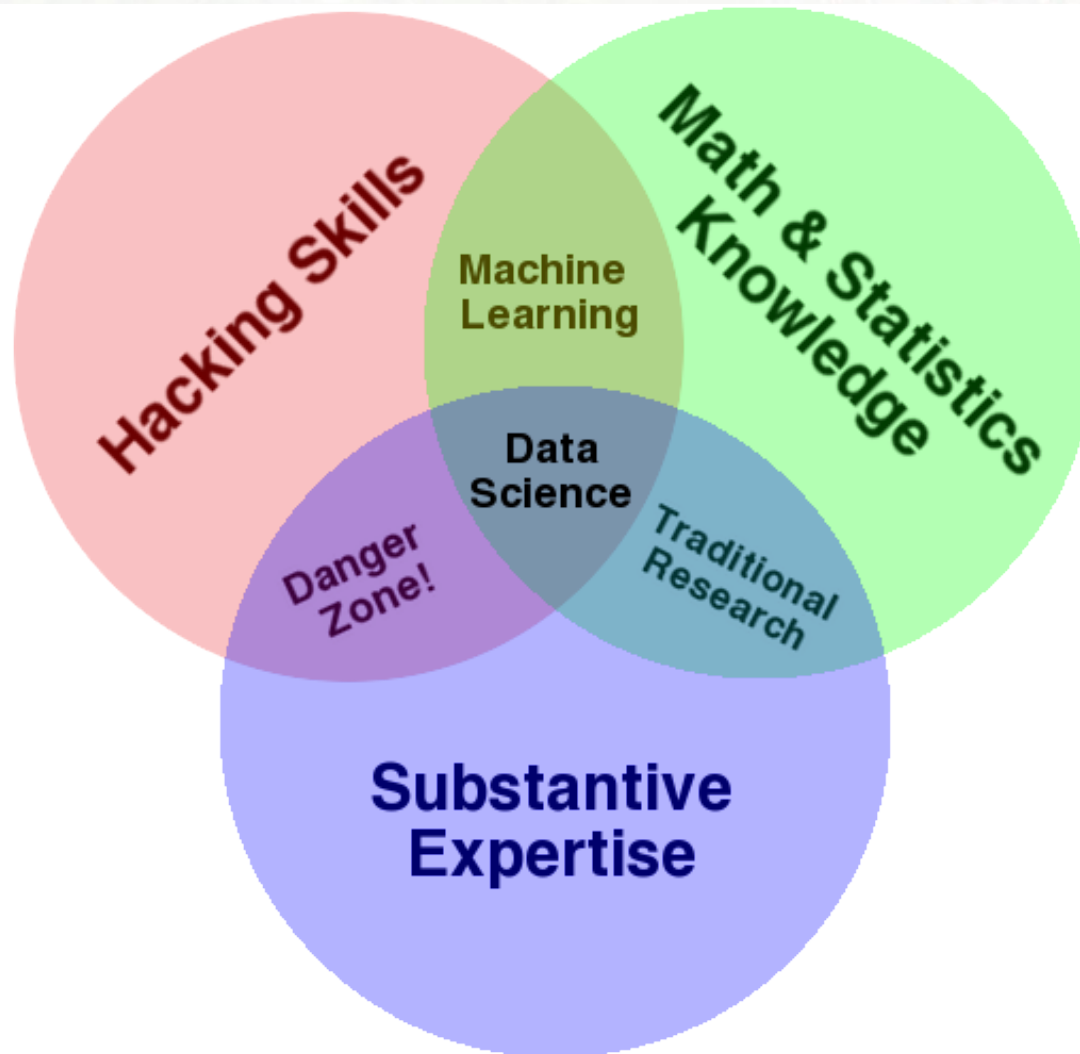
- “A data analyst who lives in California”
- ...almost everyone who works with data in an organization...
- ...a rare hybrid, a computer scientist with the programming abilities to build software to scrape, combine, and manage data from a variety of sources and a statistician who knows how to derive insights from the information within...
- ...someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning.

## For our purposes...

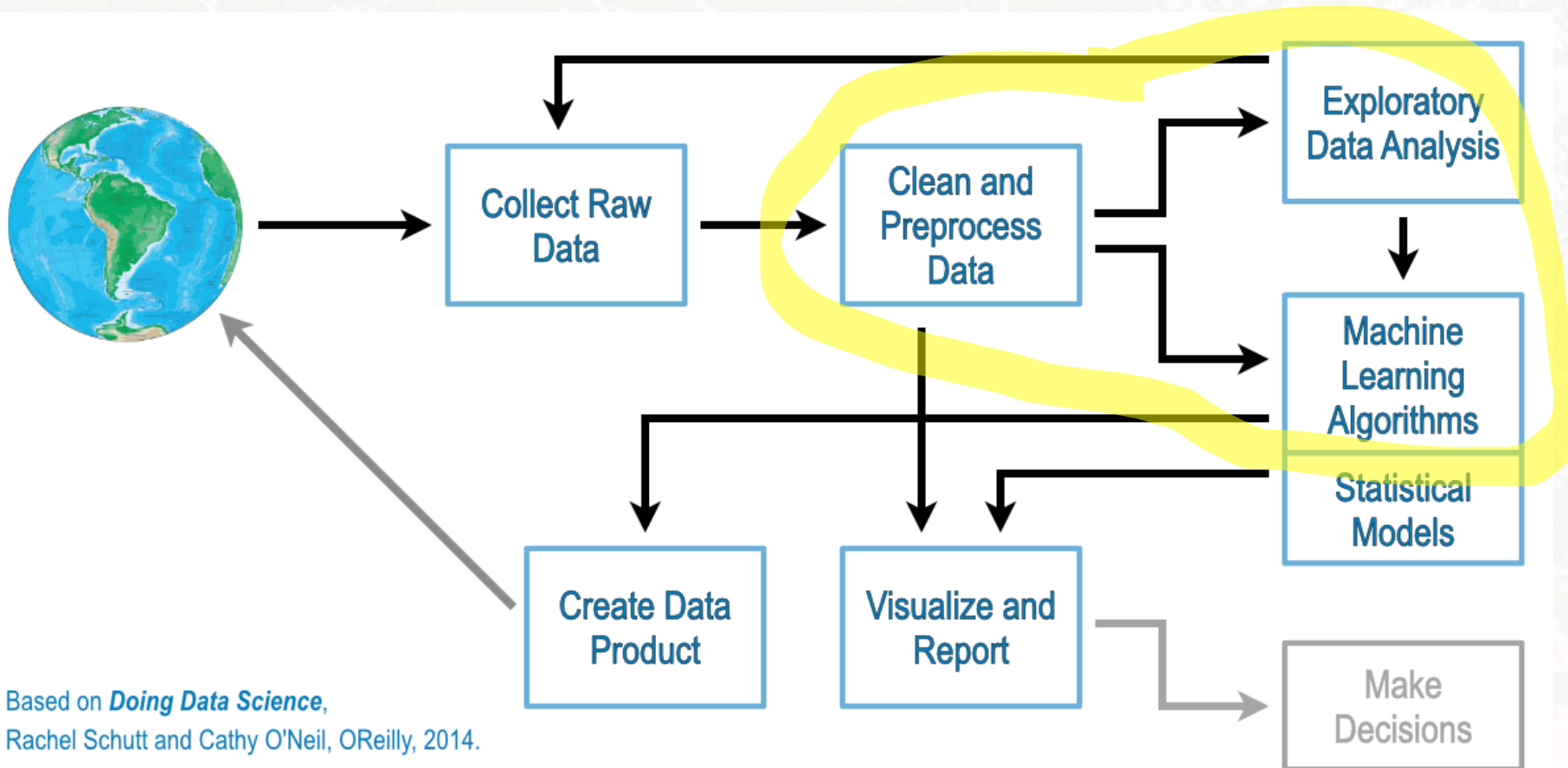
- *...an academic data scientist is a scientist, trained in anything from social science to biology, who works with large amounts of data, and must grapple with computational problems posed by the structure, size, messiness, and the complexity and nature of the data, while simultaneously solving a real-world problem.*



# What is Data Science?



# Data Mining vs Data Science



# Data Mining as part of Data Science

- We have the data.
- It can be easily *tidy*fyed.
- We need to know whether we are going to predict or descript it.
- We will explore algorithms and effects.

# Tidy Data

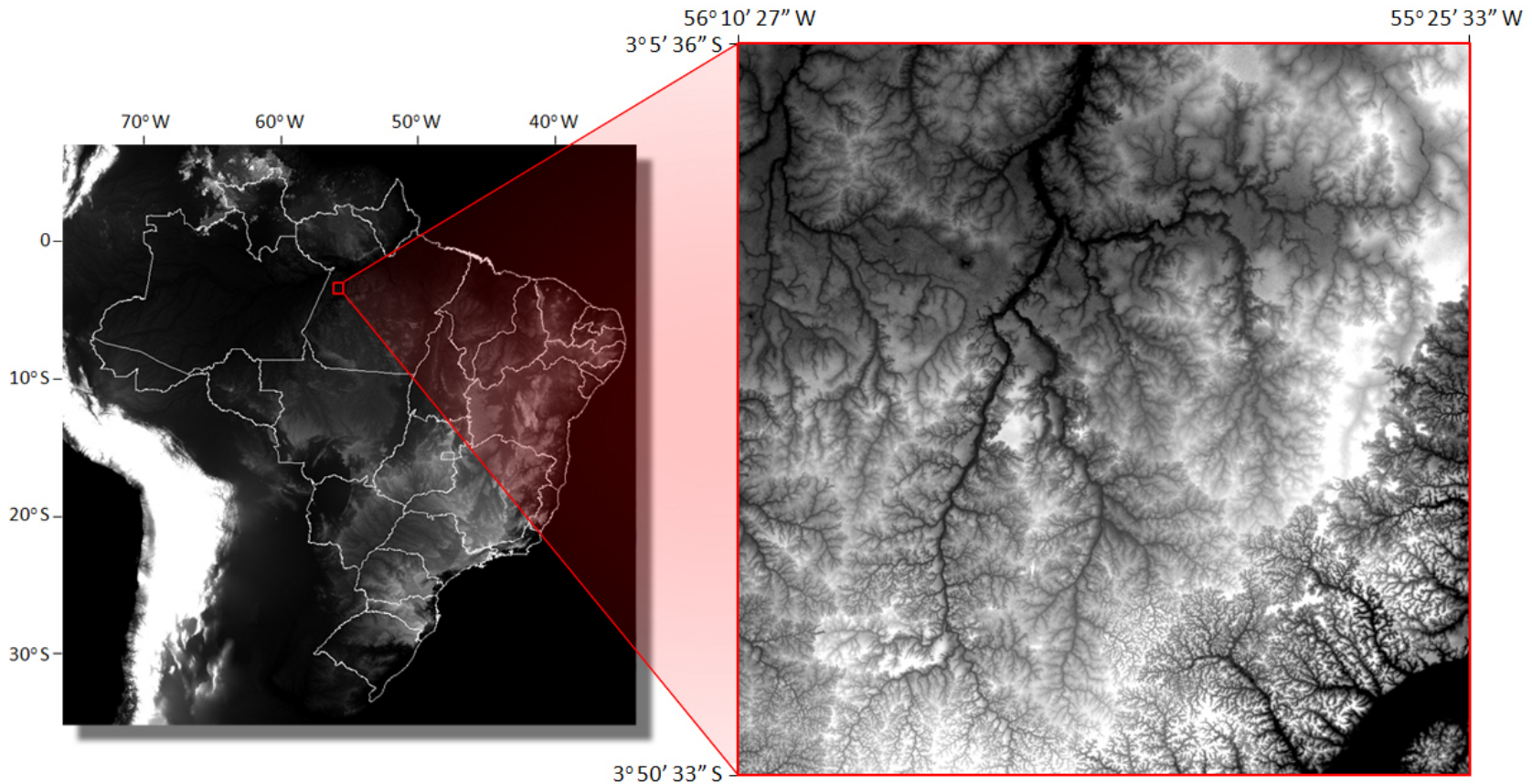
- For EDA, visualization, data mining, machine learning, etc. we will need **tidy data**.
  - Each variable measured should be in one column.
  - Each different observation of that variable should be in a different row.
  - There should be one table for each type of variable (purpose of measurement!)
  - *If there are multiple tables, they should include a column in the table to be linked.*

# Principles and Applications of Data Mining

## Examples

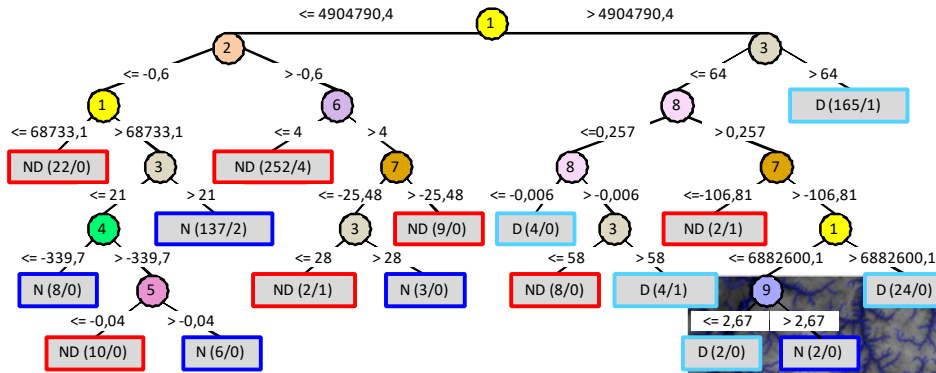
# Image Mining

Extraction of drainage networks from digital elevation images using decision trees.



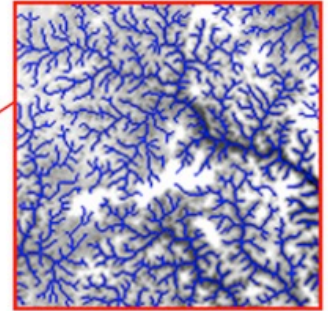
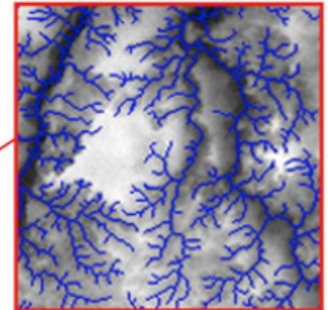
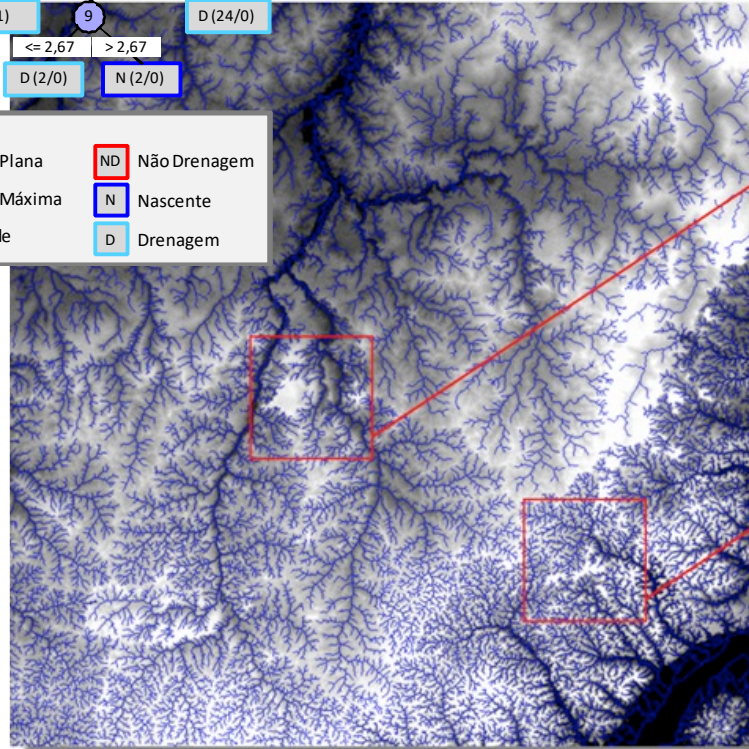
Banon, Lise Christine, et al. "Definição de critérios a partir da mineração de dados para a extração automática de redes de drenagem."

# Image Mining



## LEGENDA

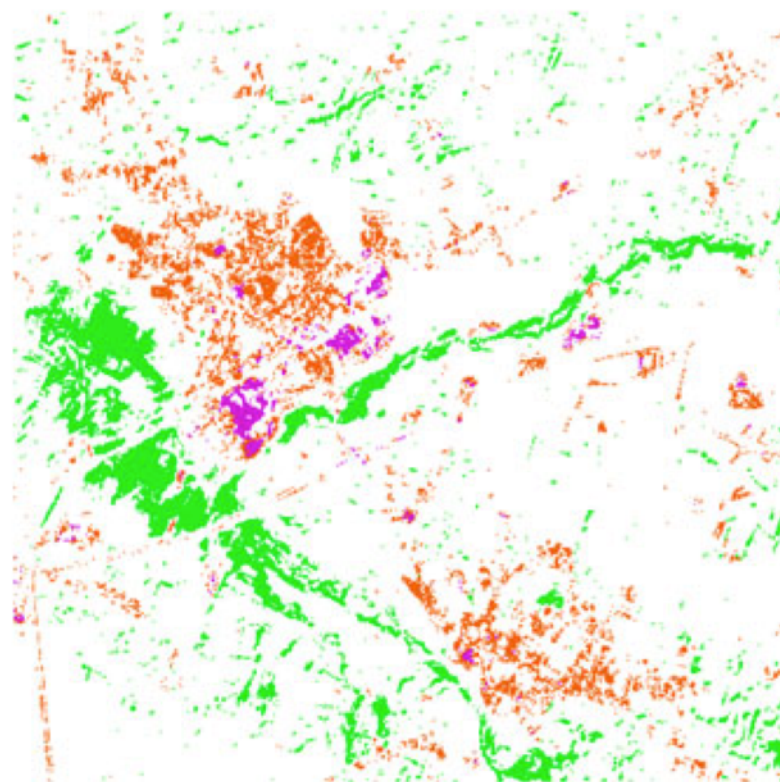
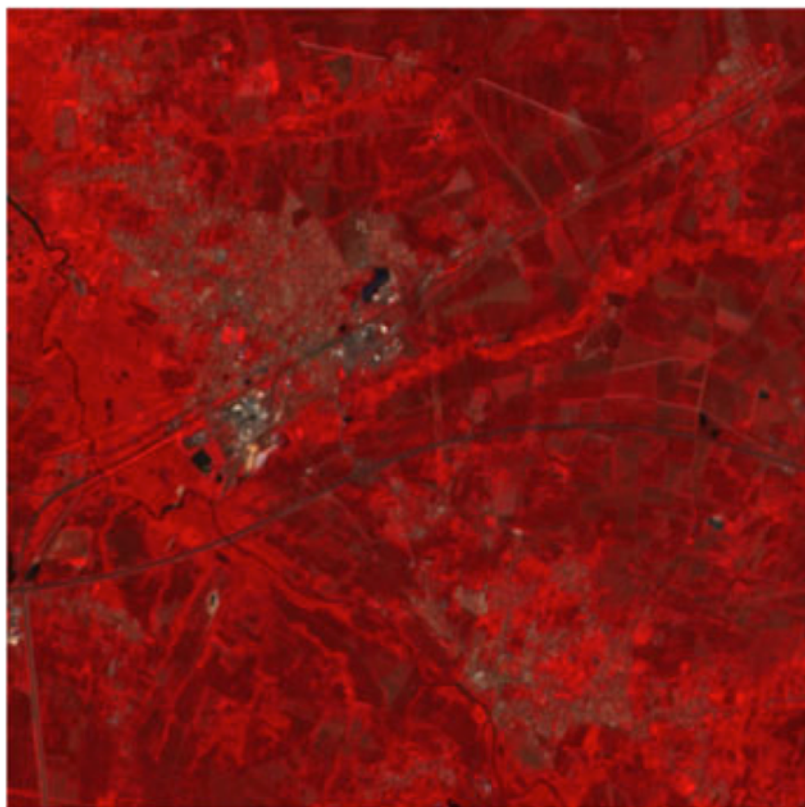
- |  |   |  |   |
|--|---|--|---|
| <span style="border: 1px solid black; border-radius: 50%; padding: 2px;">1</span> Área de Contribuição | <span style="border: 1px solid black; border-radius: 50%; padding: 2px;">4</span> Curvatura Circular Total      | <span style="border: 1px solid black; border-radius: 50%; padding: 2px;">7</span> Curvatura Plana  | <span style="border: 1px solid black; padding: 2px;">ND</span> Não Drenagem |
| <span style="border: 1px solid black; border-radius: 50%; padding: 2px;">2</span> Curvatura Mínima     | <span style="border: 1px solid black; border-radius: 50%; padding: 2px;">5</span> Curvatura de Acumulação Total | <span style="border: 1px solid black; border-radius: 50%; padding: 2px;">8</span> Curvatura Máxima | <span style="border: 1px solid black; padding: 2px;">N</span> Nascente      |
| <span style="border: 1px solid black; border-radius: 50%; padding: 2px;">3</span> Desnível ao Topo     | <span style="border: 1px solid black; border-radius: 50%; padding: 2px;">6</span> Ordem Máxima de Straler       | <span style="border: 1px solid black; border-radius: 50%; padding: 2px;">9</span> Declividade      | <span style="border: 1px solid black; padding: 2px;">D</span> Drenagem      |



Banon, Lise Christine, et al. "Definição de critérios a partir da mineração de dados para a extração automática de redes de drenagem."

# Spatiotemporal/Image Mining

Finding patterns of temporal changes in remote sensing images (patterns from NIR,R,G,NDVI bands in a 20-year sequence).

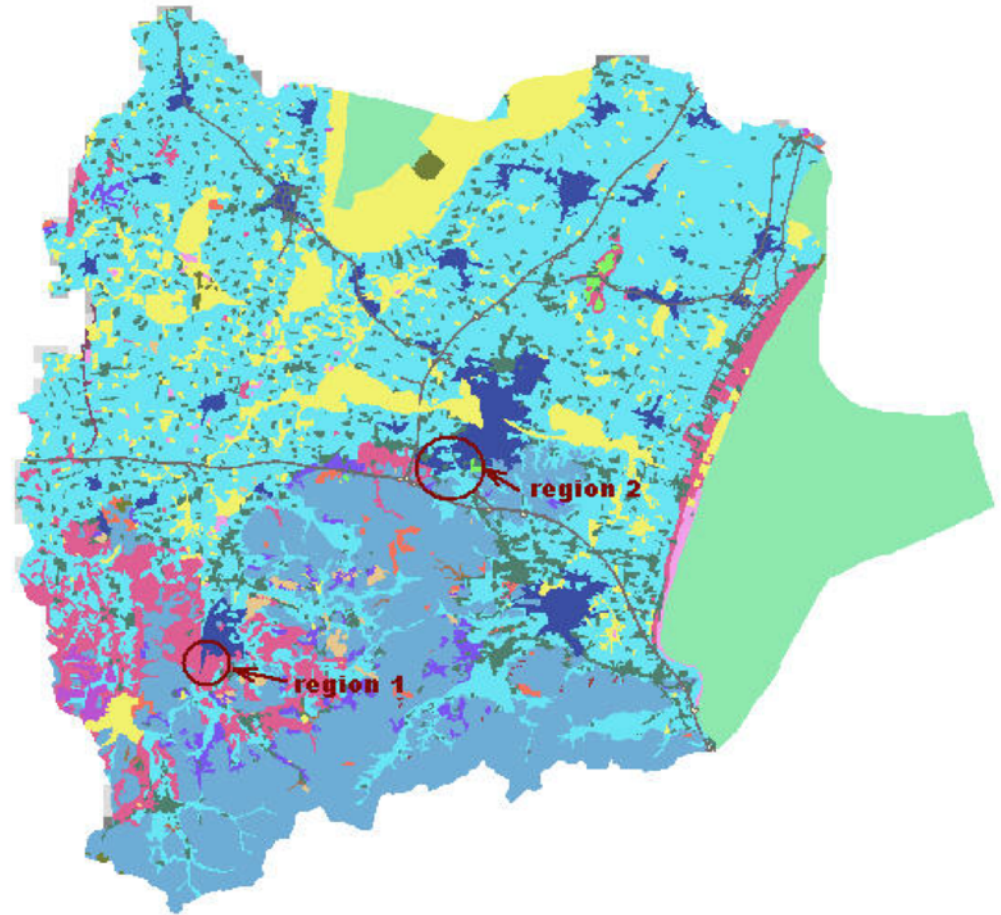


Petitjean, François, et al. "Analysing satellite image time series by means of pattern mining." *Intelligent Data Engineering and Automated Learning–IDEAL 2010*. Springer Berlin Heidelberg, 2010. 45-52.



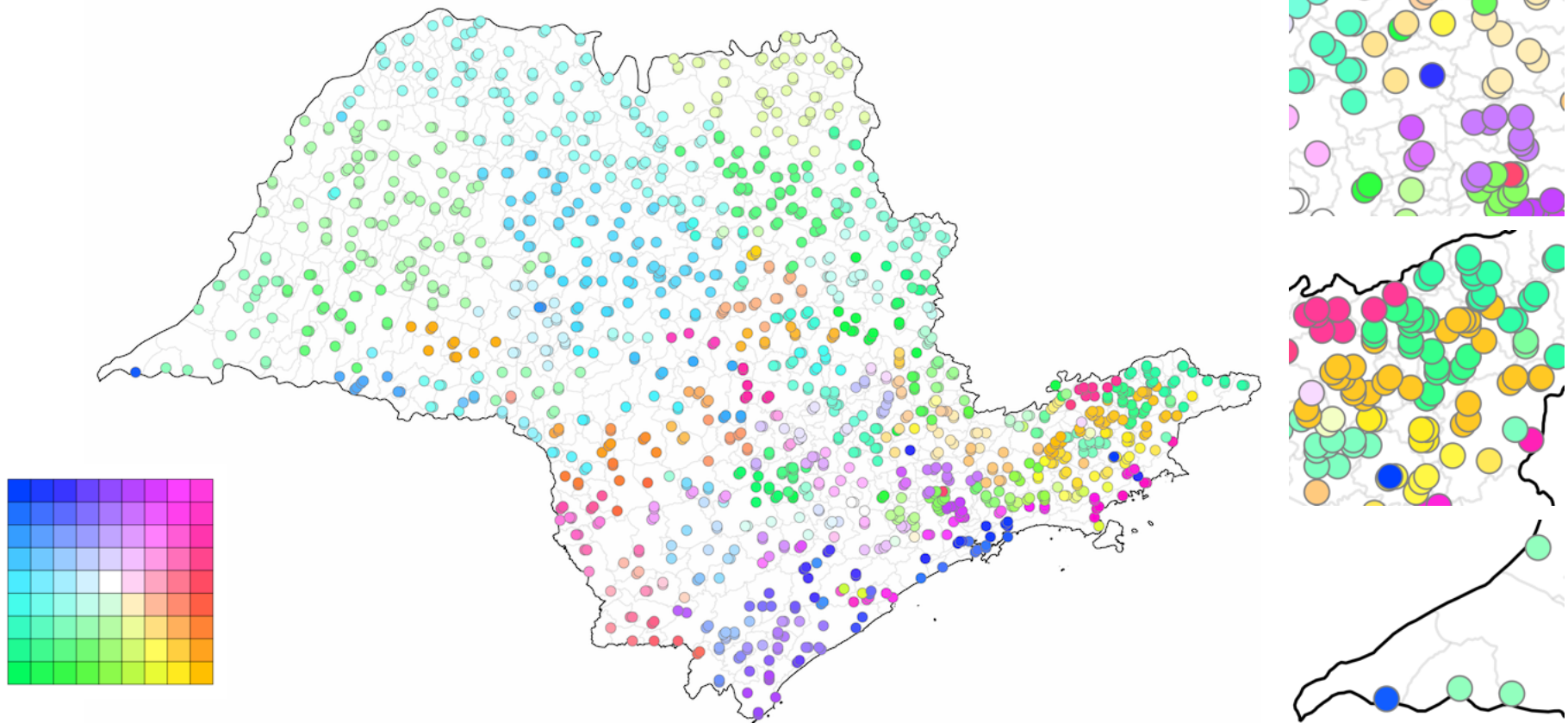
# Mining Spatial Association Rules

Identification of interesting patterns in land use change using association rules.



# Spatiotemporal Outliers

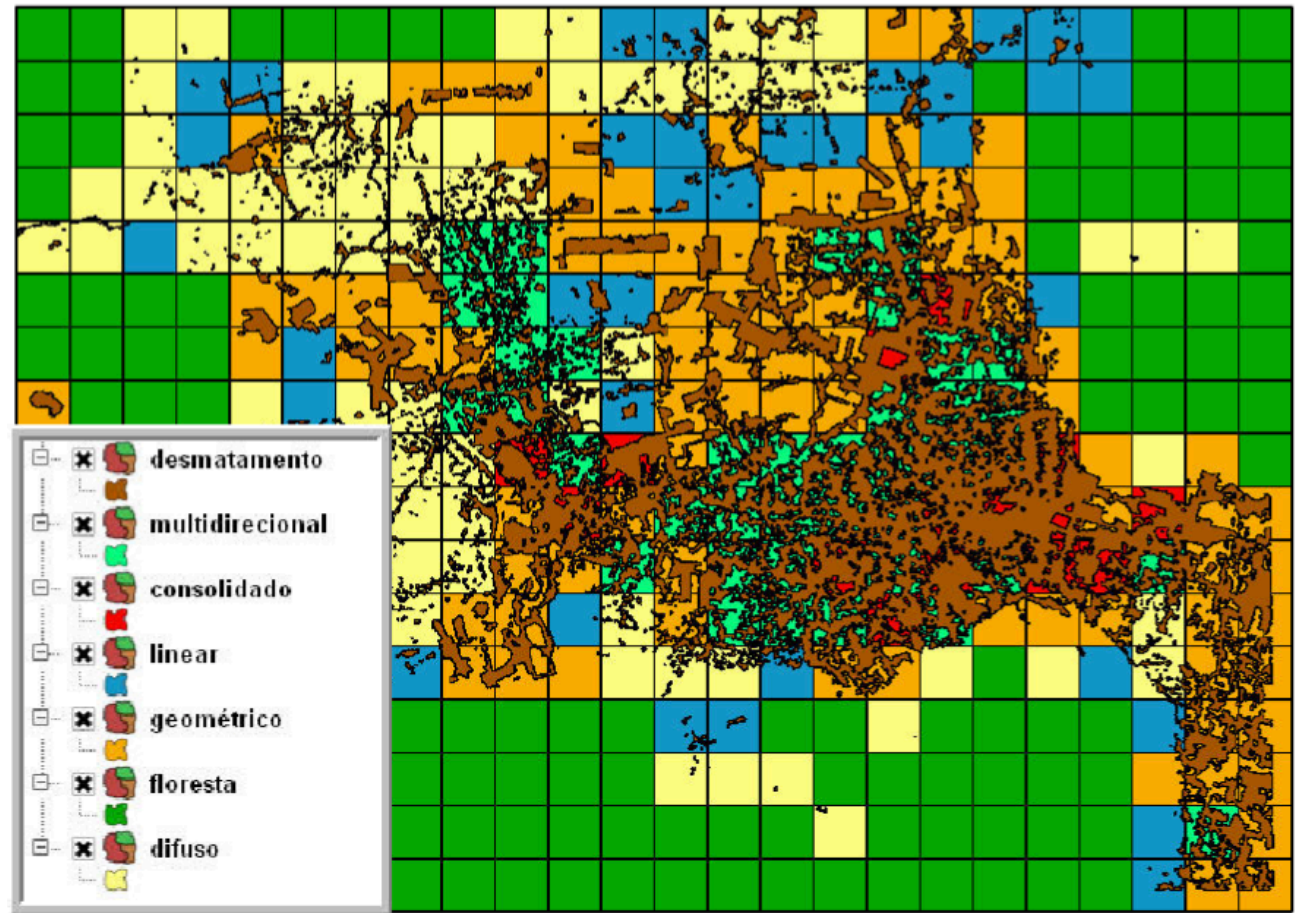
Map time series to a color map using SOM: visual identification of outliers.



Garcia, José Roberto M., Antônio Miguel V. Monteiro, and Rafael DC Santos. "Visual data mining for identification of patterns and outliers in weather stations' data." *Intelligent Data Engineering and Automated Learning-IDEAL 2012*. Springer Berlin Heidelberg, 2012. 245-252.

# Mining Spatial Patterns

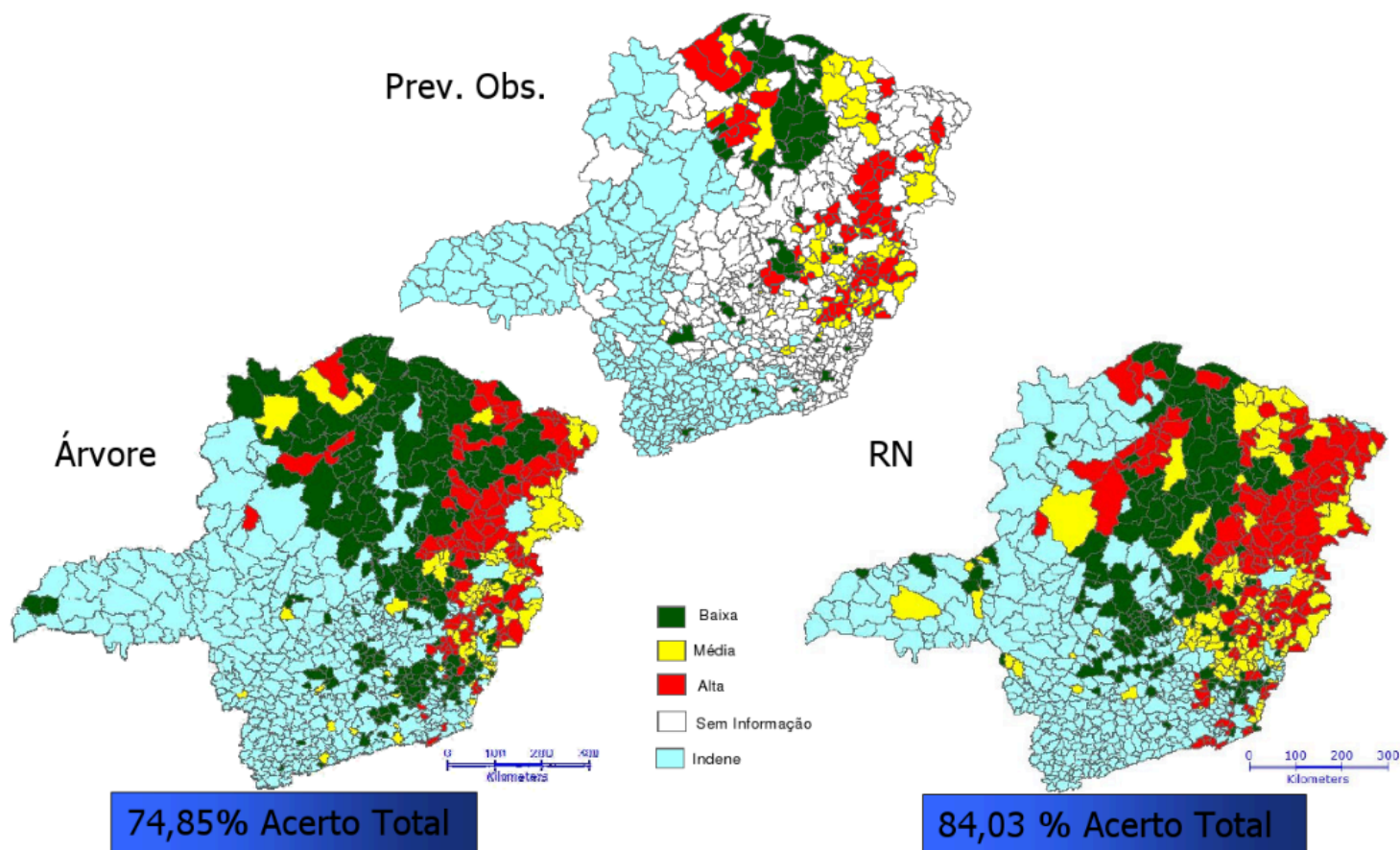
Patterns of deforestation in the Amazon Rainforest.



Márcio Azeredo report for this course.

# Spatial Data Mining

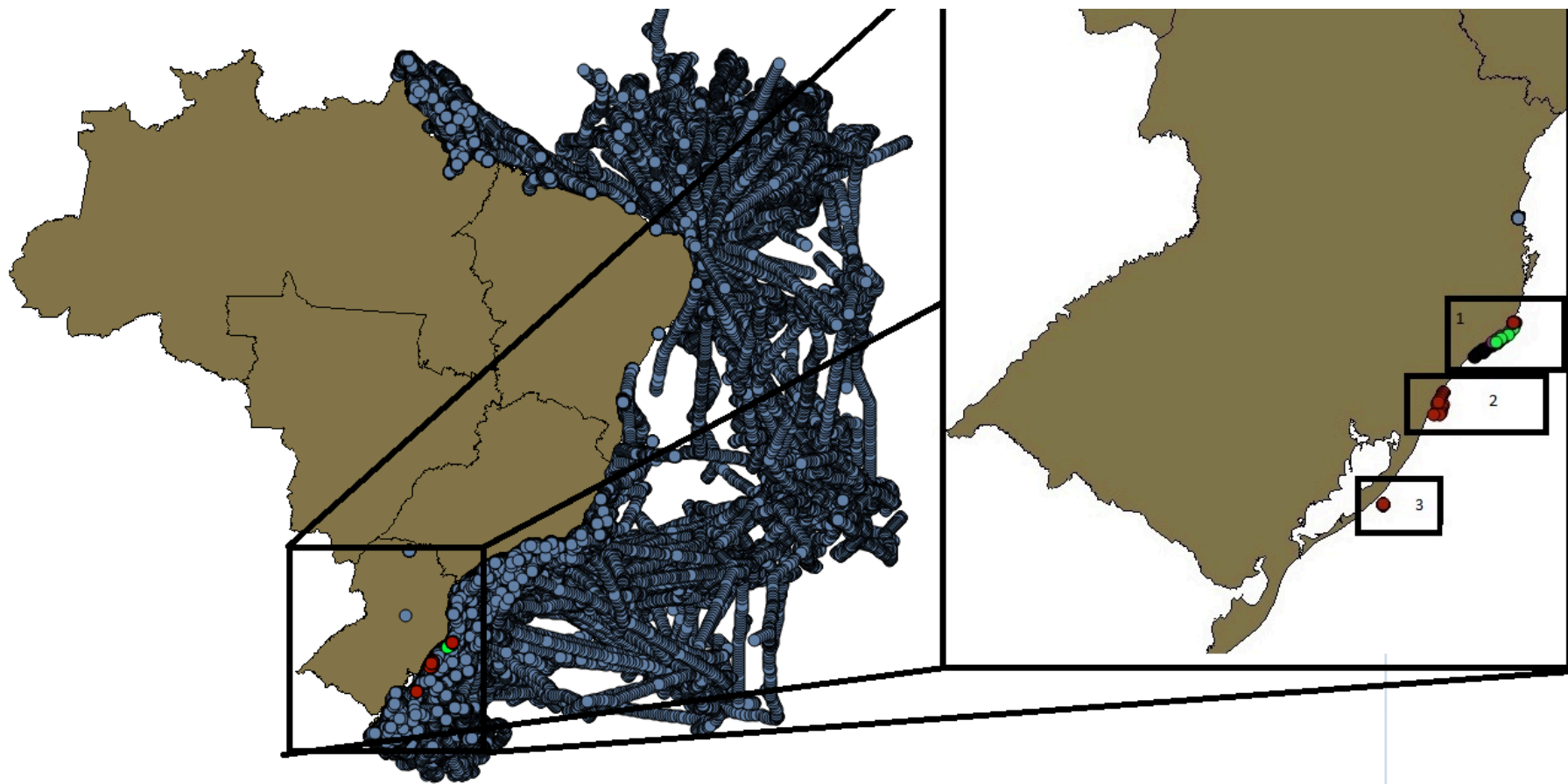
- Schistosomiasis risk estimation using data mining.



Flávia Toledo Martins report for this course.

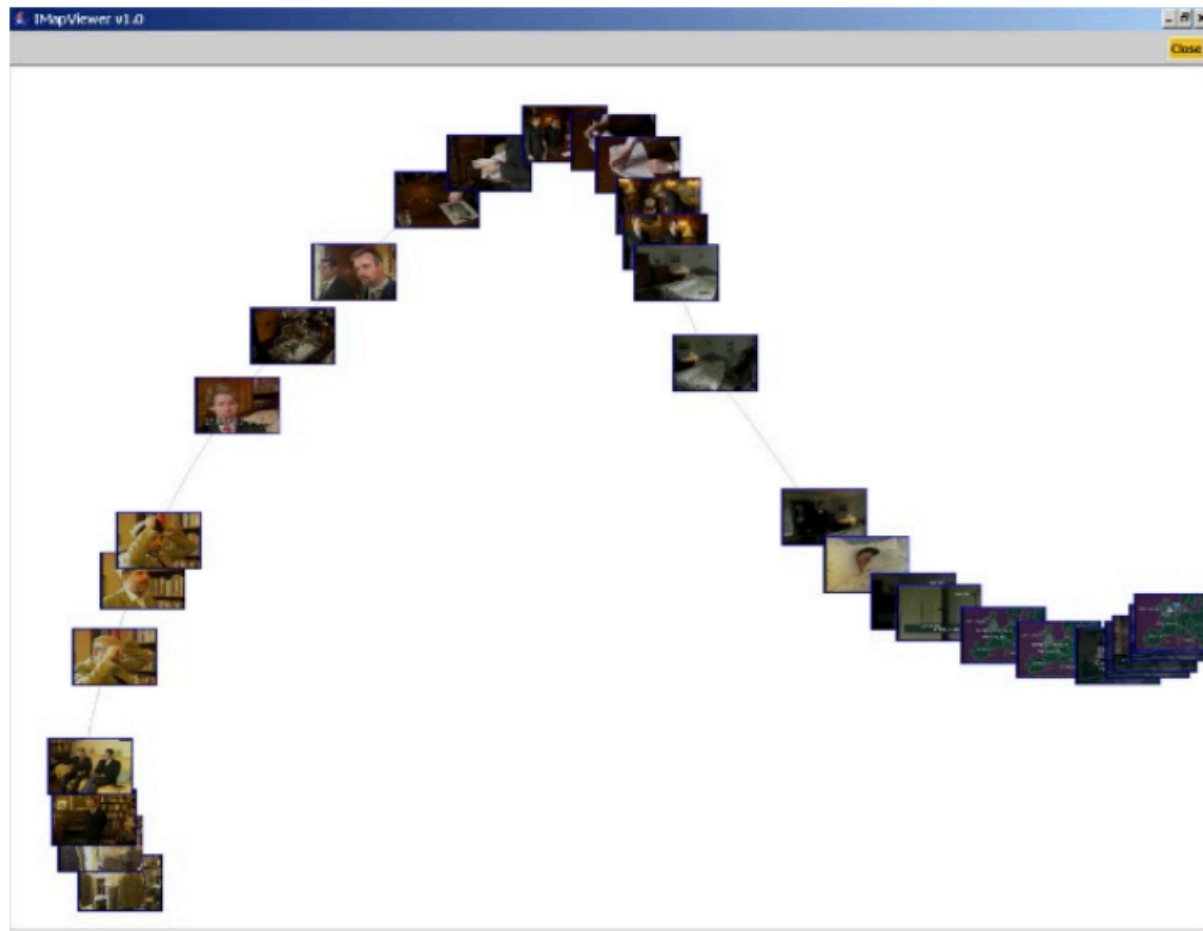
# Spatiotemporal Patterns Mining

“Partners”



# Video/Image Mining

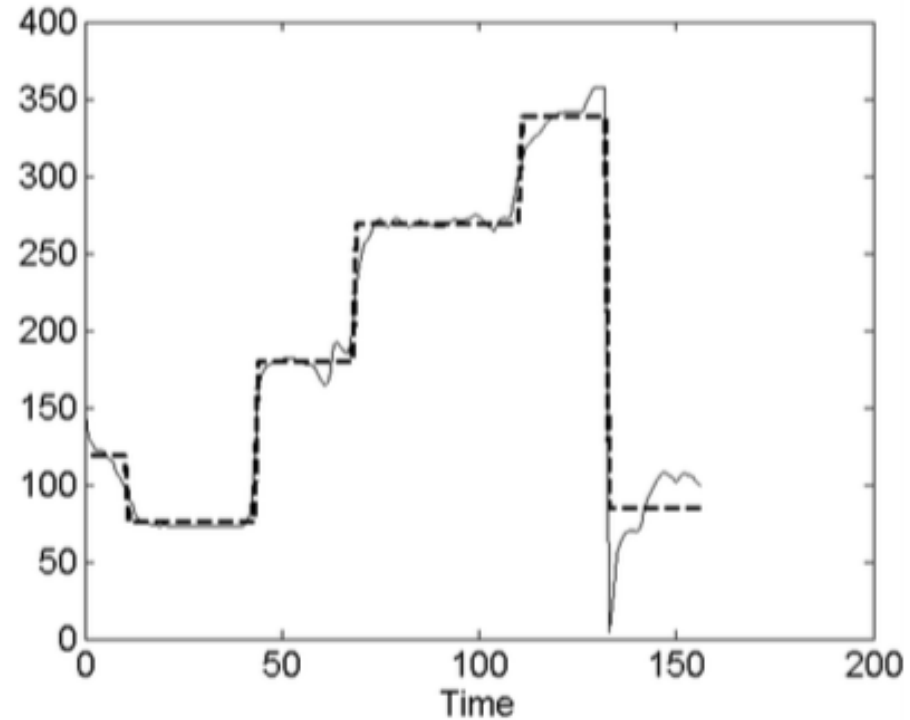
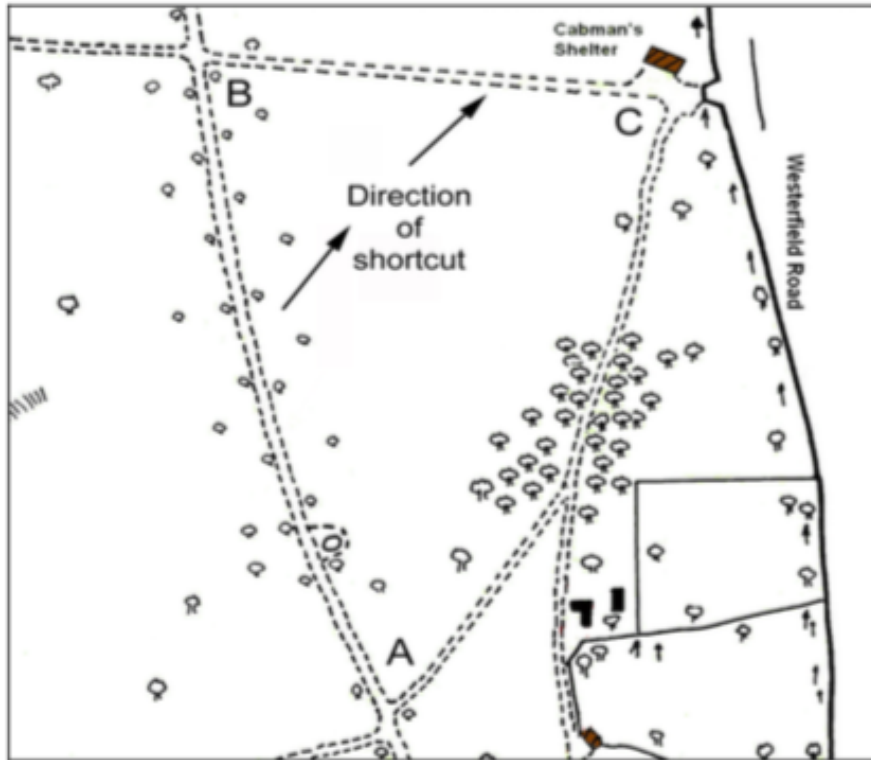
Creation of video summaries from key frames and color layout descriptors.



Deng, Da. "Braving the semantic gap: Mapping visual concepts from images and videos." *Advances in Data Mining*. Springer Berlin Heidelberg, 2005. 50-59.

# Trajectory Mining

Clustering of journey paths to determine behavior.



Hunter, Julia, and Martin Colley. "Feature extraction from sensor data streams for real-time human behaviour recognition." *Knowledge Discovery in Databases: PKDD 2007*. Springer Berlin Heidelberg, 2007. 115-126.

# Trajectory Mining

Identification of outliers (anomalous behavior) from data collected by taxis' GPSs.

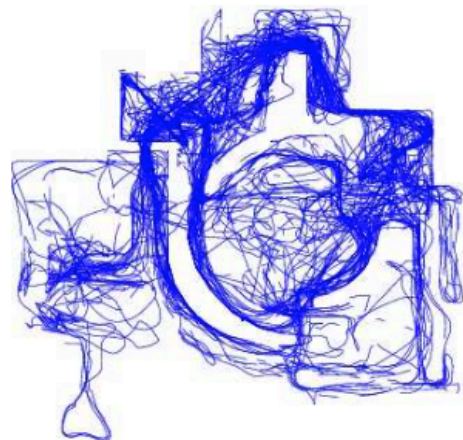


Pang, Linsey Xiaolin, et al. "On mining anomalous patterns in road traffic streams." *Advanced Data Mining and Applications*. Springer Berlin Heidelberg, 2011. 237-251.

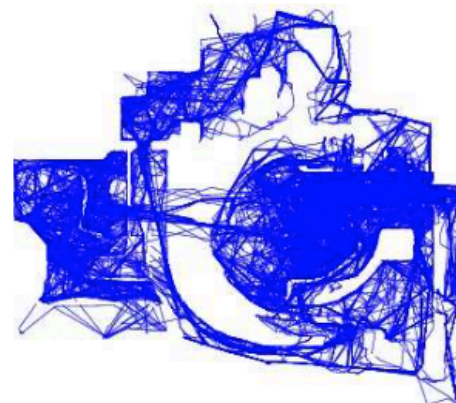


# Trajectory Mining

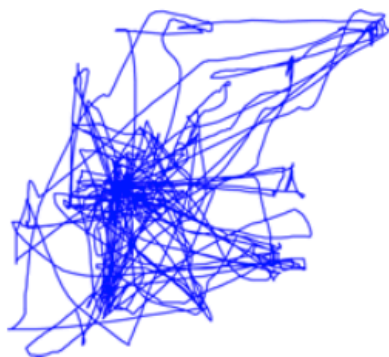
Generic trajectory mapping for identification and verification.



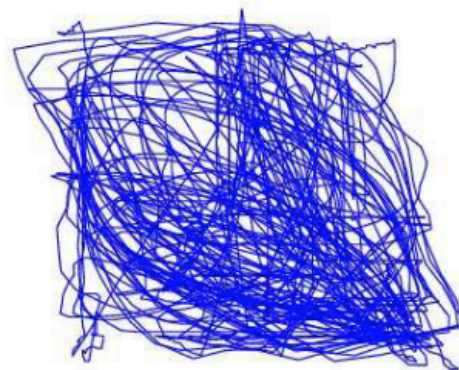
(a) On-line game: human



(b) On-line game: bot



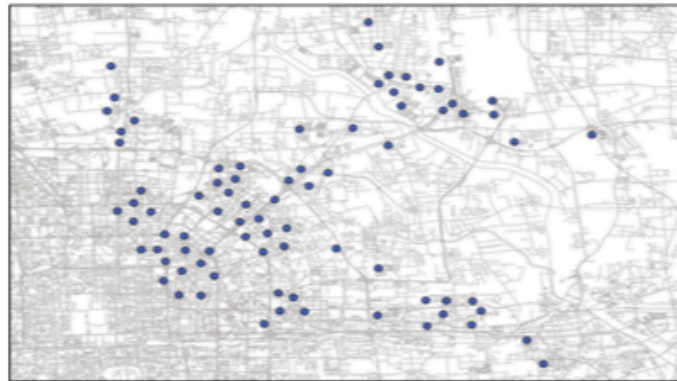
(c) Mouse trace: a left-handed user



(d) Mouse trace: a right-handed user

# Trajectory Mining

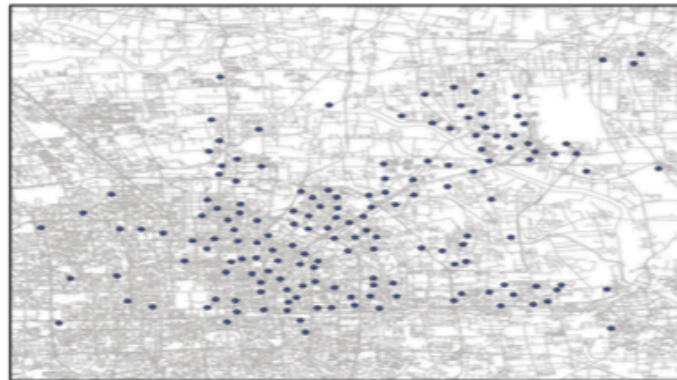
Clustering algorithm to identify “hotspots” for bus routes



(a) Eps = 0.5, MinPt = 30



(b) Eps = 0.5, MinPts = 50



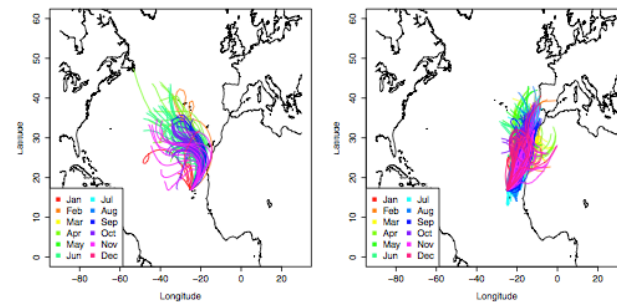
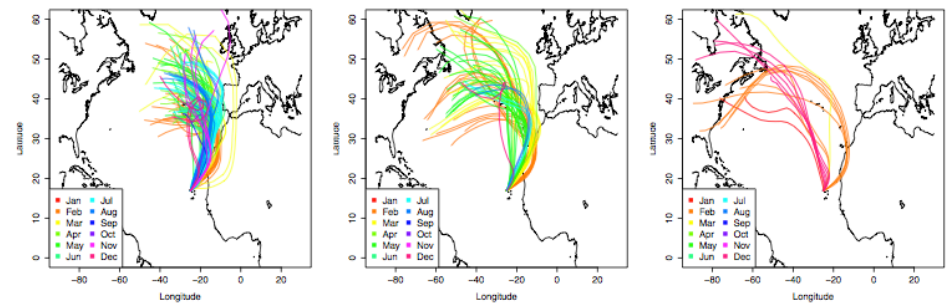
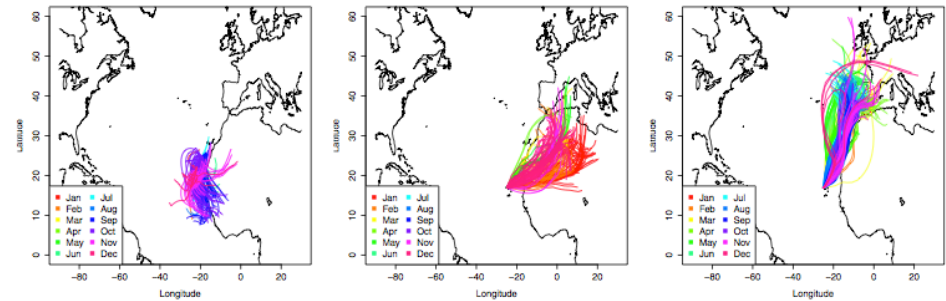
(c) Eps = 0.5, MinPt = 20



(d) Eps = 0.2, MinPts = 20

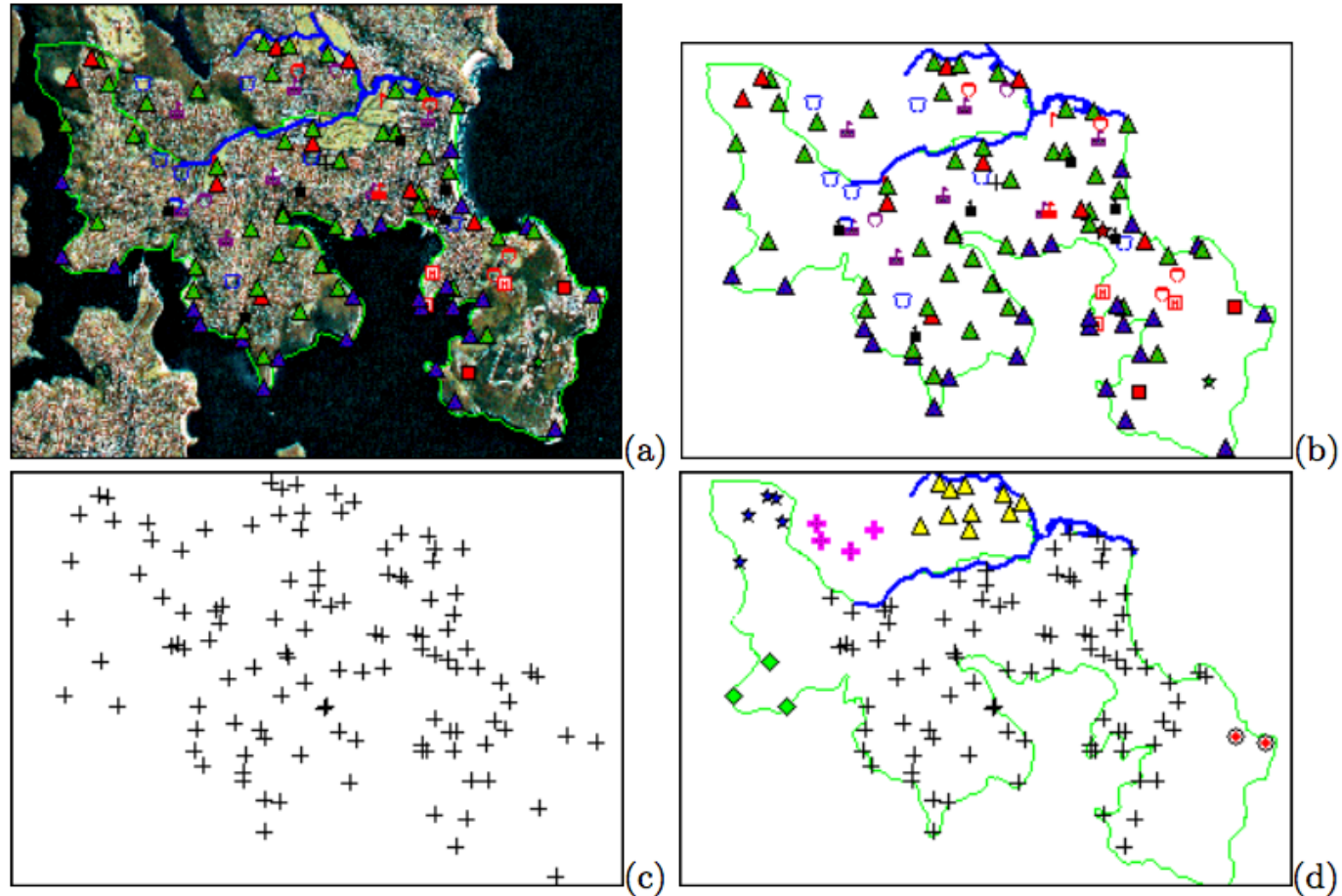
# Air Mass Trajectories

Grouping of air mass trajectories into similar shapes, using also chemical composition for differentiation.



# Spatial Clustering

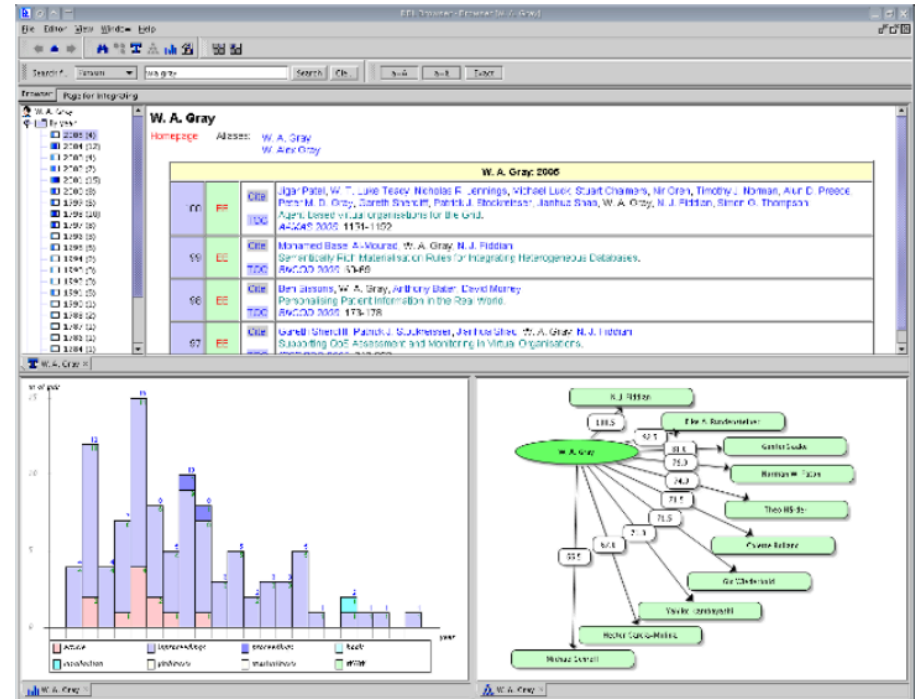
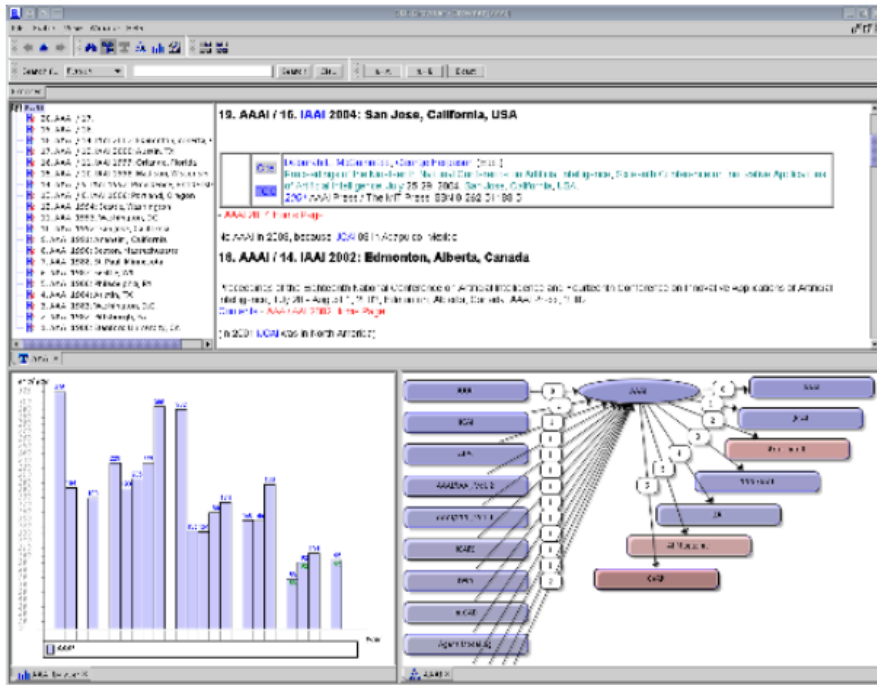
Algorithm for clustering of spatial data with obstacles.



Estivill-Castro, Vladimir, and Ickjai Lee. "Autoclust+: Automatic clustering of point-data sets in the presence of obstacles." *Temporal, Spatial, and Spatio-Temporal Data Mining*. Springer Berlin Heidelberg, 2001. 133-146.

# Social Networks Mining

Tools to help discover implicit social networks from references.



Klink, Stefan, et al. "Analysing social networks within bibliographical data." *Database and Expert Systems Applications*. Springer Berlin Heidelberg, 2006.



# Multiple Features Clustering

Clustering Twitter data considering location, time, contents and social data.

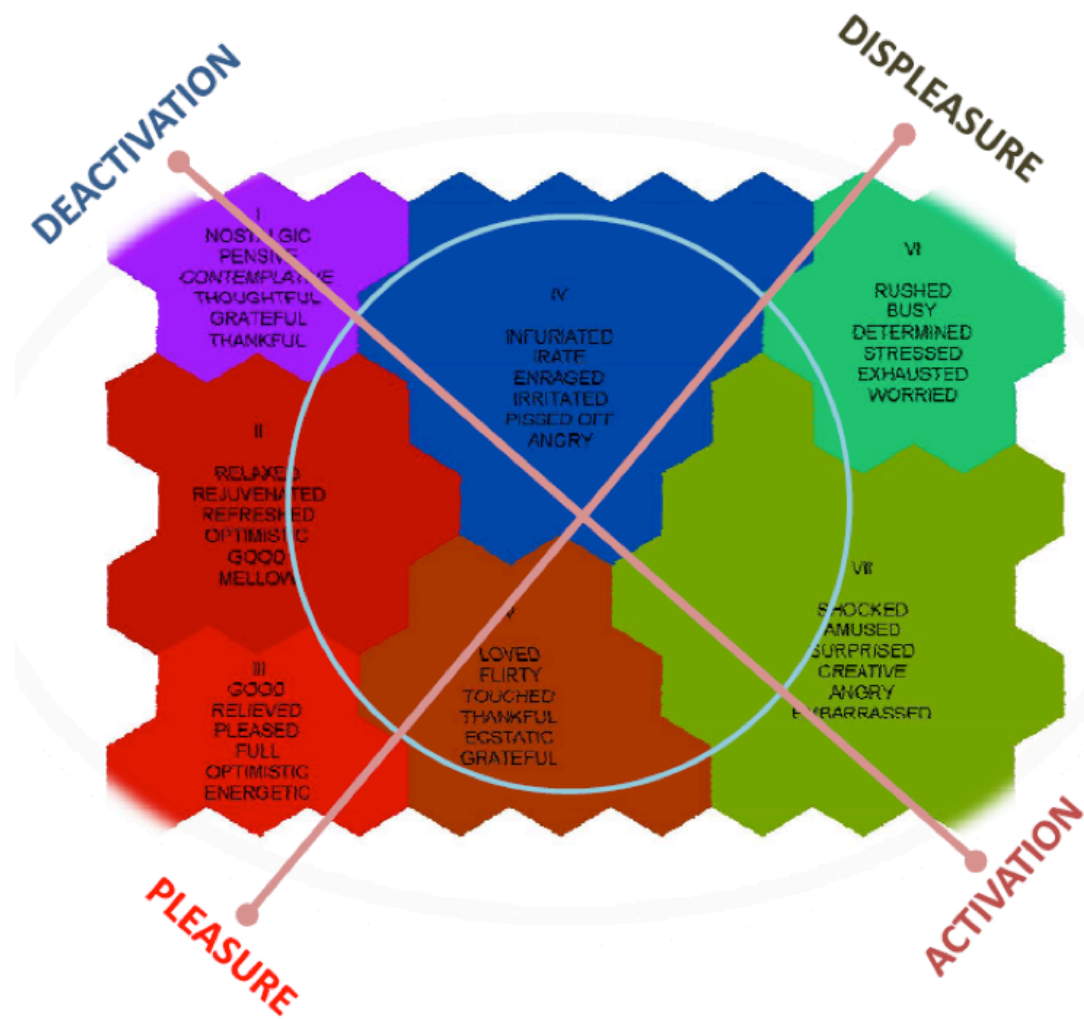


Tiago Cunha, Carlos Soares, and Eduarda Mendes Rodrigues. "TweProfiles: detection of spatio-temporal patterns on Twitter." *Advanced Data Mining and Applications*. Springer International Publishing, 2014. 123-136.

# Text Mining (Blogs)

Mood analysis from blogs post (with ground truth).

SOM and clustering to find mood groups.

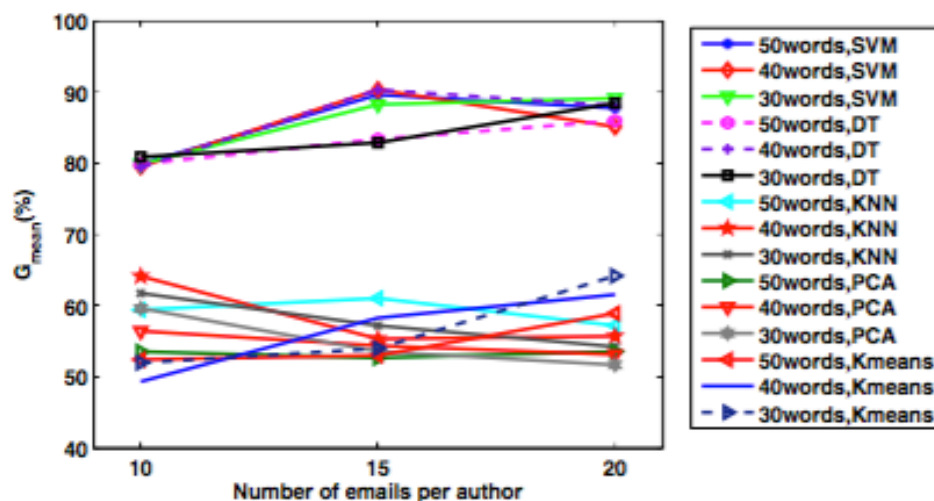




# Text Mining (E-mails)

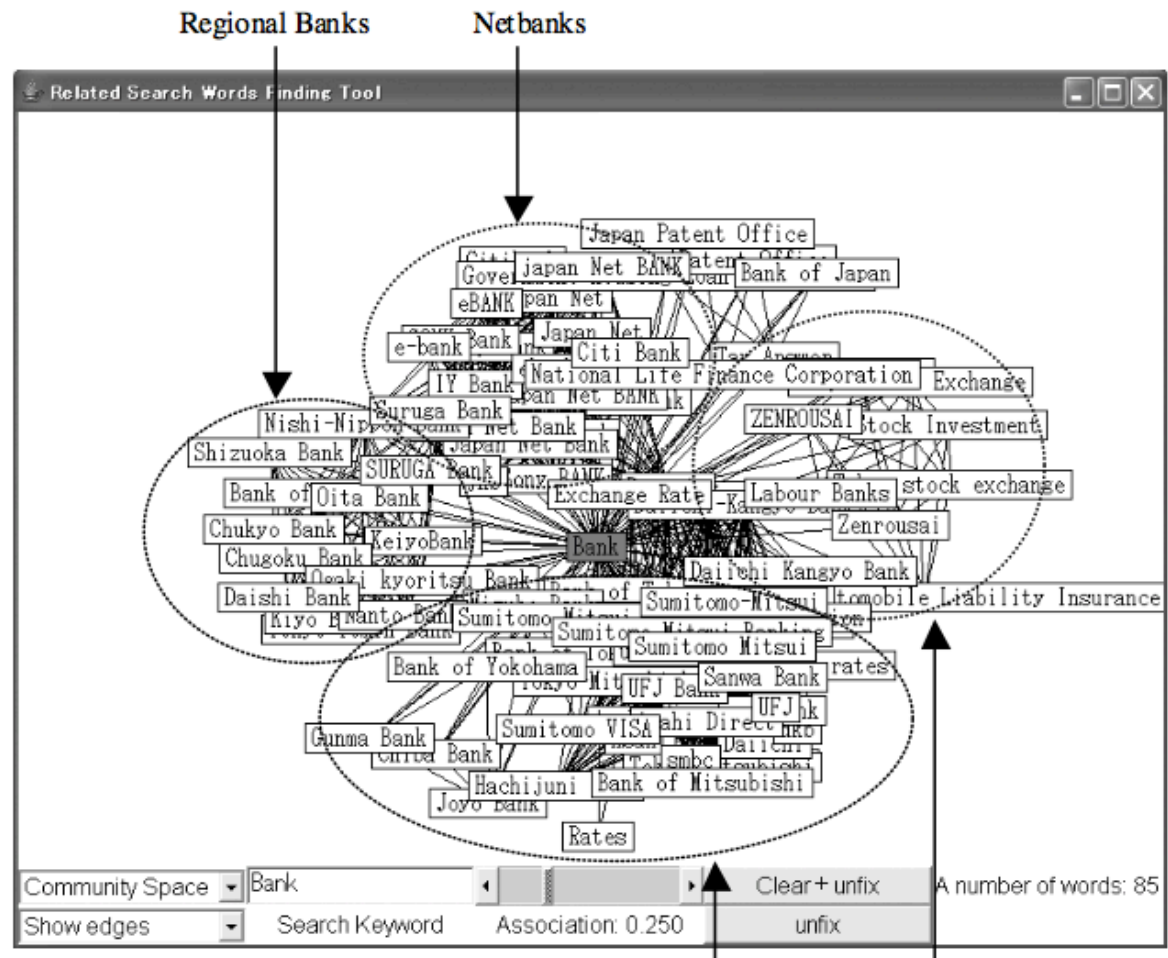
## Detecting e-mail authorship.

Category	Features
Lexical	Total number of characters in words(Ch) Total number of letters (a-z)/Ch Total number of digital characters/Ch Total number of upper characters/Ch Average length per word (in characters) Word count (C) Average words per sentence Word length distribution (1-30)/N (30 features) Unique words/C Words longer than 6 characters/C Total number of short words (1-3 characters)/C
syntactical	Total number of punctuation characters/Ch Number of each punctuation (31 features) /Ch 44 function features from LIWC
Structural	Absence/present of greeting words Absence/present of farewell words Number of blank lines/total number of lines Average length of non blank line Number of paragraphs Average words per paragraph Number of sentences (S) Number of sentences beginning with upper case/S Number of sentences beginning with lower case/S
Content-specific	24 content-specific features from LIWC The number of net abbreviation/C



# Analytics/Log Mining

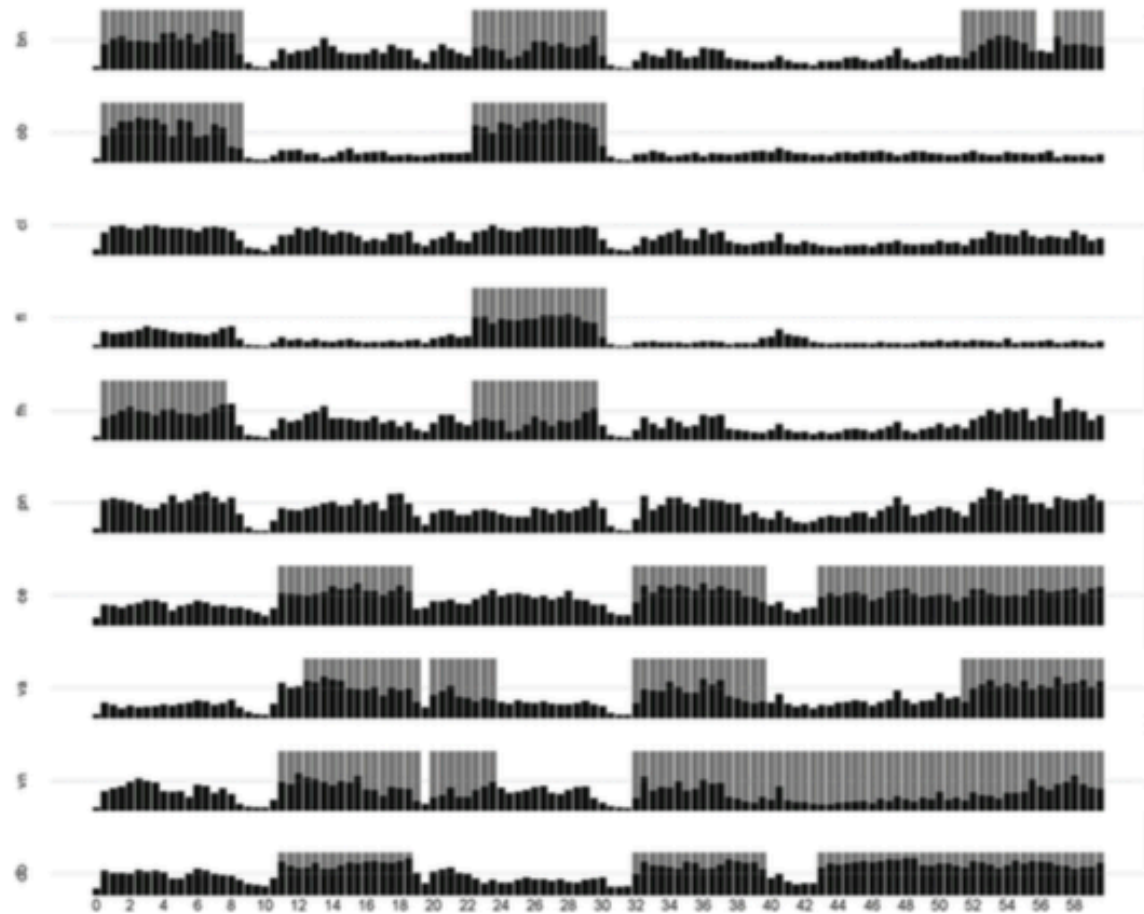
## Keyword clustering in search engines.



Otsuka, Shingo, and Masaru Kitsuregawa. "Clustering of search engine keywords using access logs." *Database and Expert Systems Applications*. Springer Berlin Heidelberg, 2006.

# Sound/Music Mining

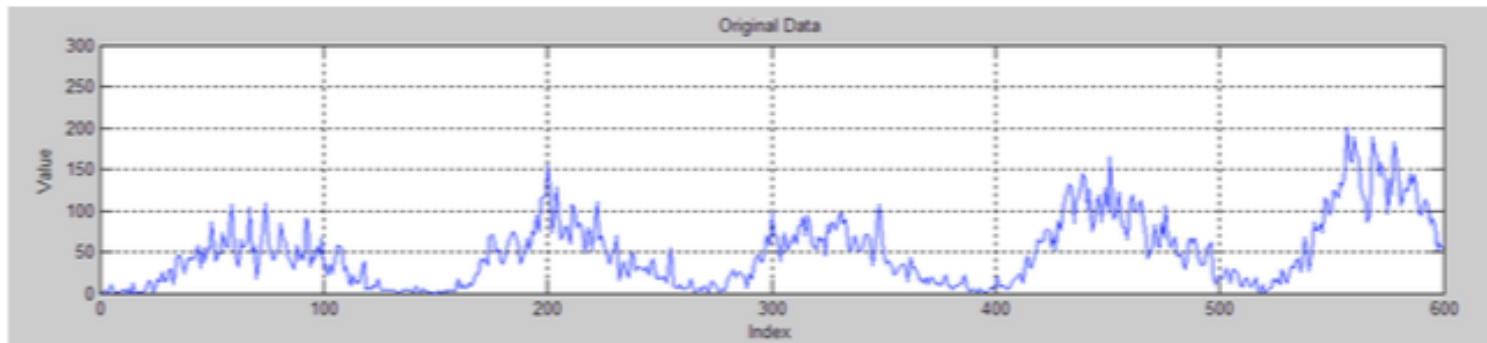
Recognizing instruments in classical music.



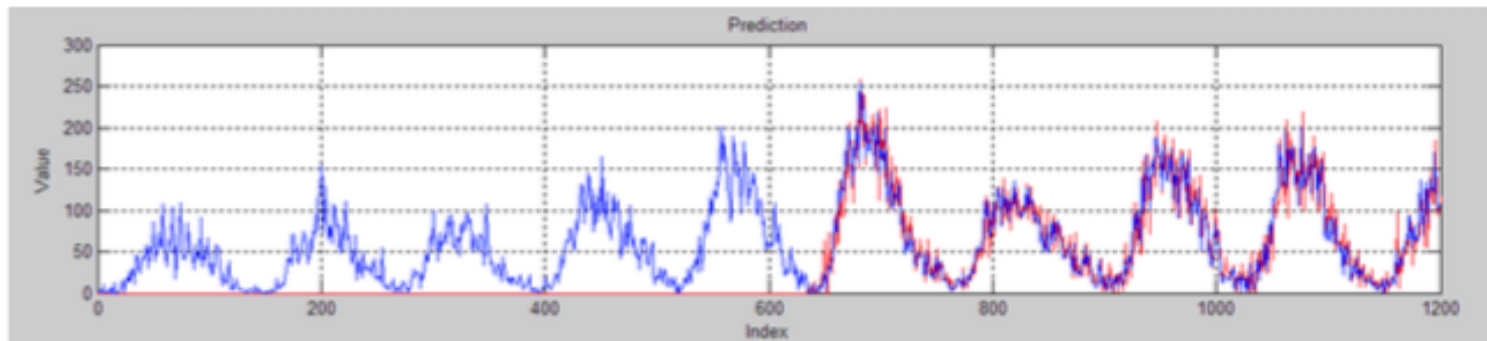
Kubera, Elżbieta, and Alicja A. Wiczorkowska. "Mining audio data for multiple instrument recognition in classical music." *New Frontiers in Mining Complex Patterns*. Springer International Publishing, 2014. 246-260.

# Time Series

Solar activity forecasting based on ANNs and moving averages.



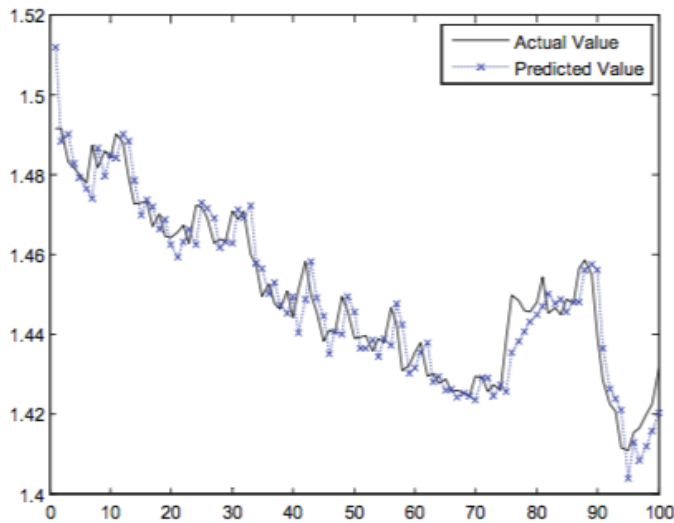
**Fig. 1.** The Original Value of Monthly Average of Sunspot Numbers



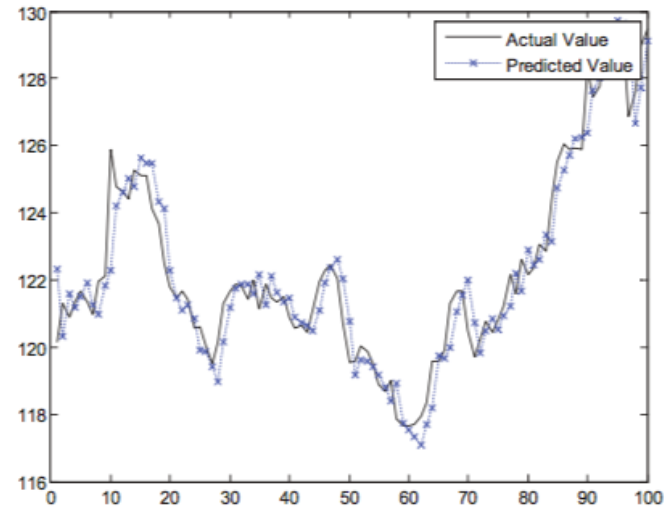
**Fig. 9.** The Prediction of Values for Monthly Average of Sunspot Number Time Series based on the proposed algorithm

# Time Series

## Forecasting of Financial Time Series with a SOM/Neural Gas.



**Fig. 3.** Actual and predicted test series for GBP/EUR time series



**Fig. 4.** Actual and predicted test series for GBP/JPY time series

# Stock Market Mining

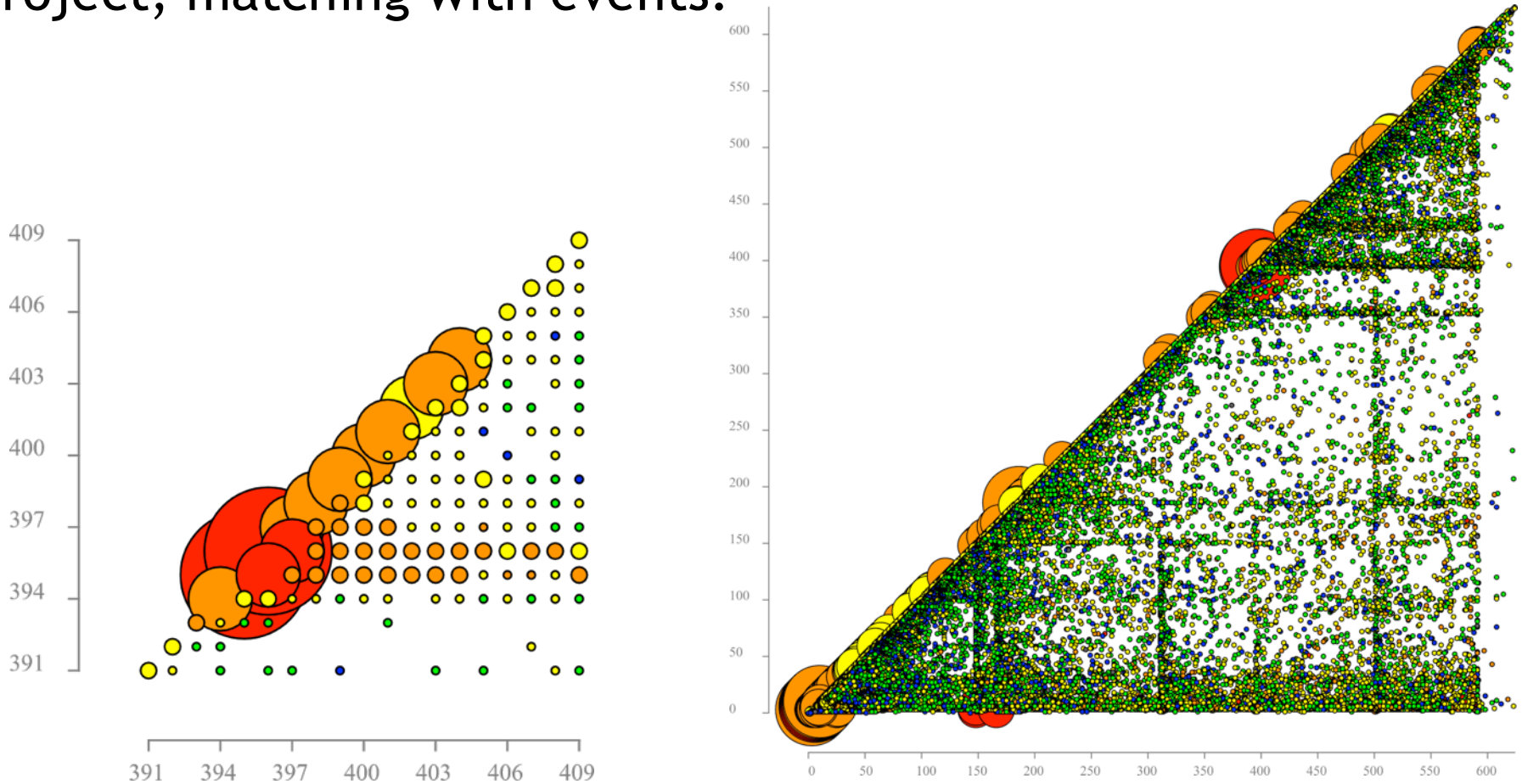
Discovering association rules in stock data variation.

No.	Classification Rules
1	<i>if AMMB=UP and RHBCAP=UP and TIME=UP then COMPOSITE=UP (191, 8)</i>
2	<i>if AMMB=DOWN and MBB=DOWN and MAS=DOWN then COMPOSITE=DOWN (118, 5)</i>
3	<i>if AMMB=DOWN and MBB=SAME and COMMERZ=DOWN then COMPOSITE=DOWN (80, 12)</i>
4	<i>if AMMB=DOWN and MBB=UP and MMBCORP=DOWN and TA=DOWN then COMPOSITE=DOWN (38,4)</i>
5	<i>if AMMB=UP and RHBCAP=UP and TIME=SAME then COMPOSITE=UP (36, 3)</i>
6	<i>if AMMB=SAME and MAS=UP and YTL=UP then COMPOSITE=UP (34, 1)</i>
7	<i>if AMMB=UP and RHBCAP=DOWN and TENAGA=UP and GENTING=UP then COMPOSITE=UP (28)</i>
8	<i>if AMMB=UP and RHBCAP=UP and TIME=DOWN and BERNAS=UP then COMPOSITE=UP (22)</i>
9	<i>if AMMB=SAME and MAS=DOWN and GUTHRIE=DOWN then COMPOSITE=DOWN (18)</i>
10	<i>if AMMB=UP and RHBCAP=DOWN and TENAGA=DOWN and DRBHCOM=DOWN and COMMERZ=DOWN then COMPOSITE=DOWN (18)</i>

Soon, Lay-Ki, and Sang Ho Lee. "Explorative Data Mining on Stock Data—Experimental Results and Findings." *Advanced Data Mining and Applications*. Springer Berlin Heidelberg, 2007. 562-569.

# Visual Data Mining

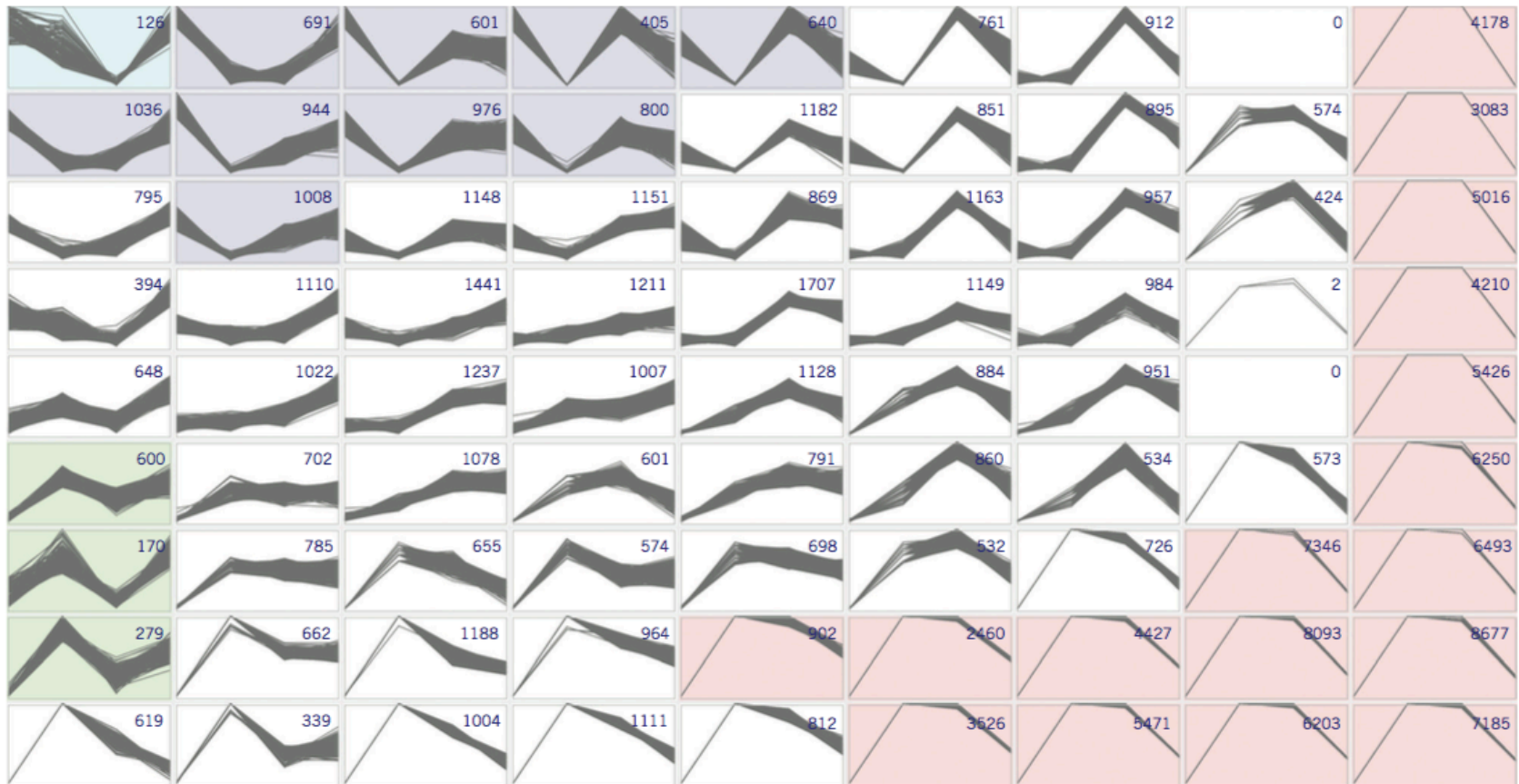
Identification of patterns of collaboration in a citizen science project, matching with events.



Morais, Alessandra M. M., Jordan Raddick, and Rafael D. C. dos Santos. "Visualization and characterization of users in a citizen science project." *SPIE Defense, Security, and Sensing. International Society for Optics and Photonics, 2013.*

# Logs Data Mining

## Clustering of behaviors of volunteers with a SOM.

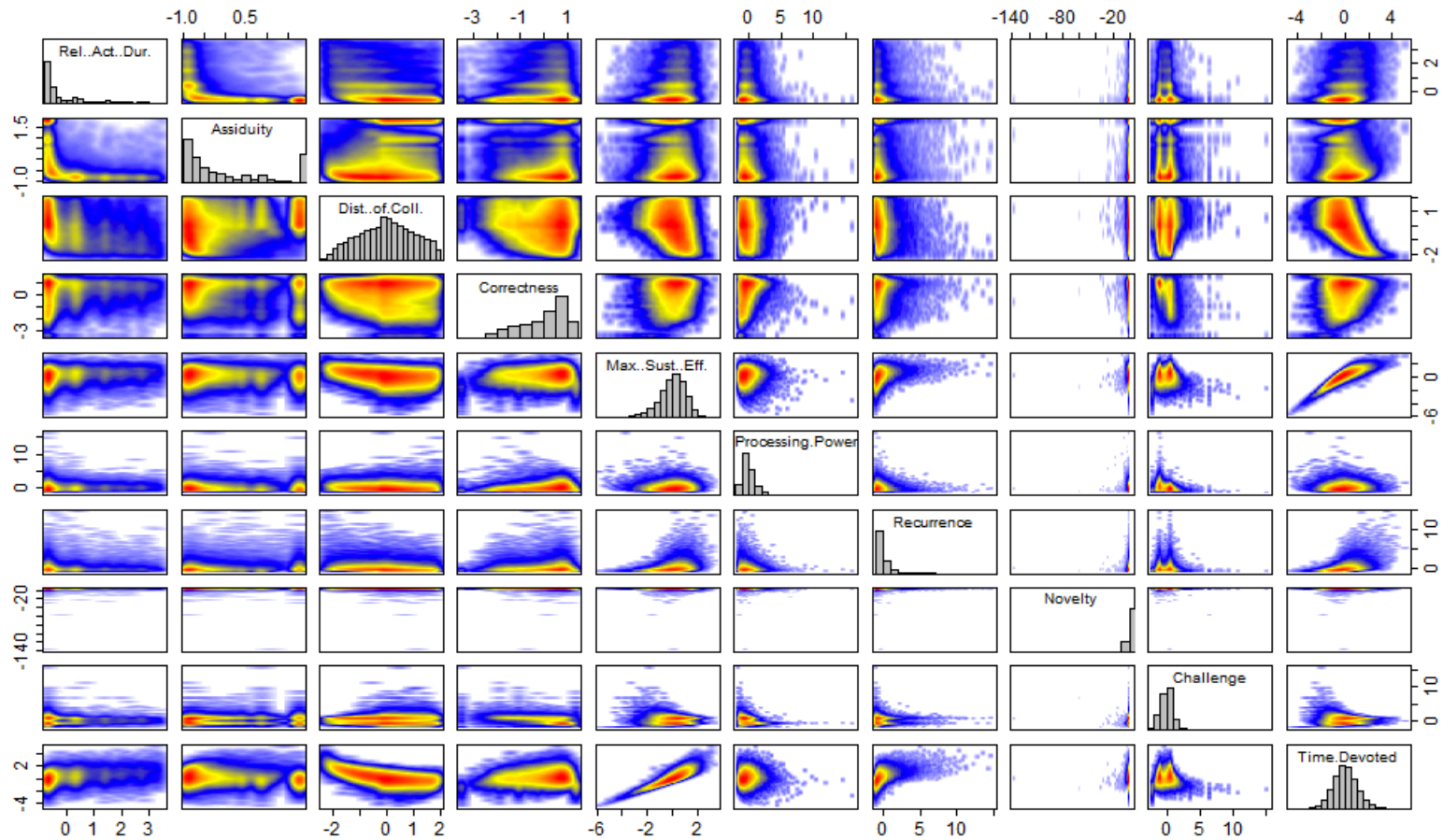


Morais, Alessandra M. M., Jordan Raddick, and Rafael D. C. dos Santos. "Visualization of Citizen Science Volunteers' Behaviors with Data from Usage Logs." to appear in *Computing in Science and Engineering*.



# Logs Data Mining

Cluster analysis of behaviors of volunteers.



Alessandra M. M. Morais' master thesis, to be submitted.

# SQL Logs Data Mining

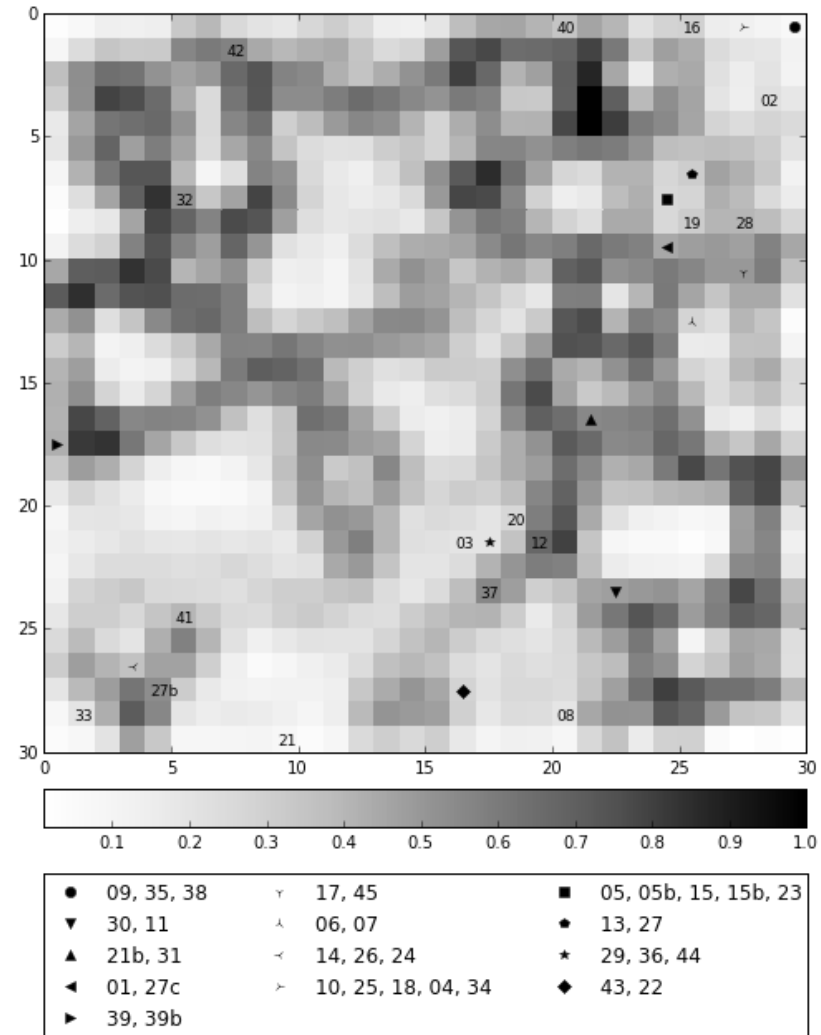
Mining SQL queries in a scientific database to identify deviations from templates.

```
SELECT p.objid, p.ra, p.dec,
       p.u, p.g, p.r, p.i, p.z,
       platex.plate, s.fiberid,
       s.elodiefeh
FROM   photoobj p,
       dbo.fgetnearbyobjeq(1.62917,
                           27.6417, 30) n,
       specobj s, platex
WHERE  p.objid = n.objid
       AND p.objid = s.bestobjid
       AND s.plateid =
         platex.plateid
       AND class = 'star'
       AND p.r >= 14
       AND p.r <= 22.5
       AND p.g >= 15
       AND p.g <= 23
       AND platex.plate = 2803
```

(a) Raw SQL query.

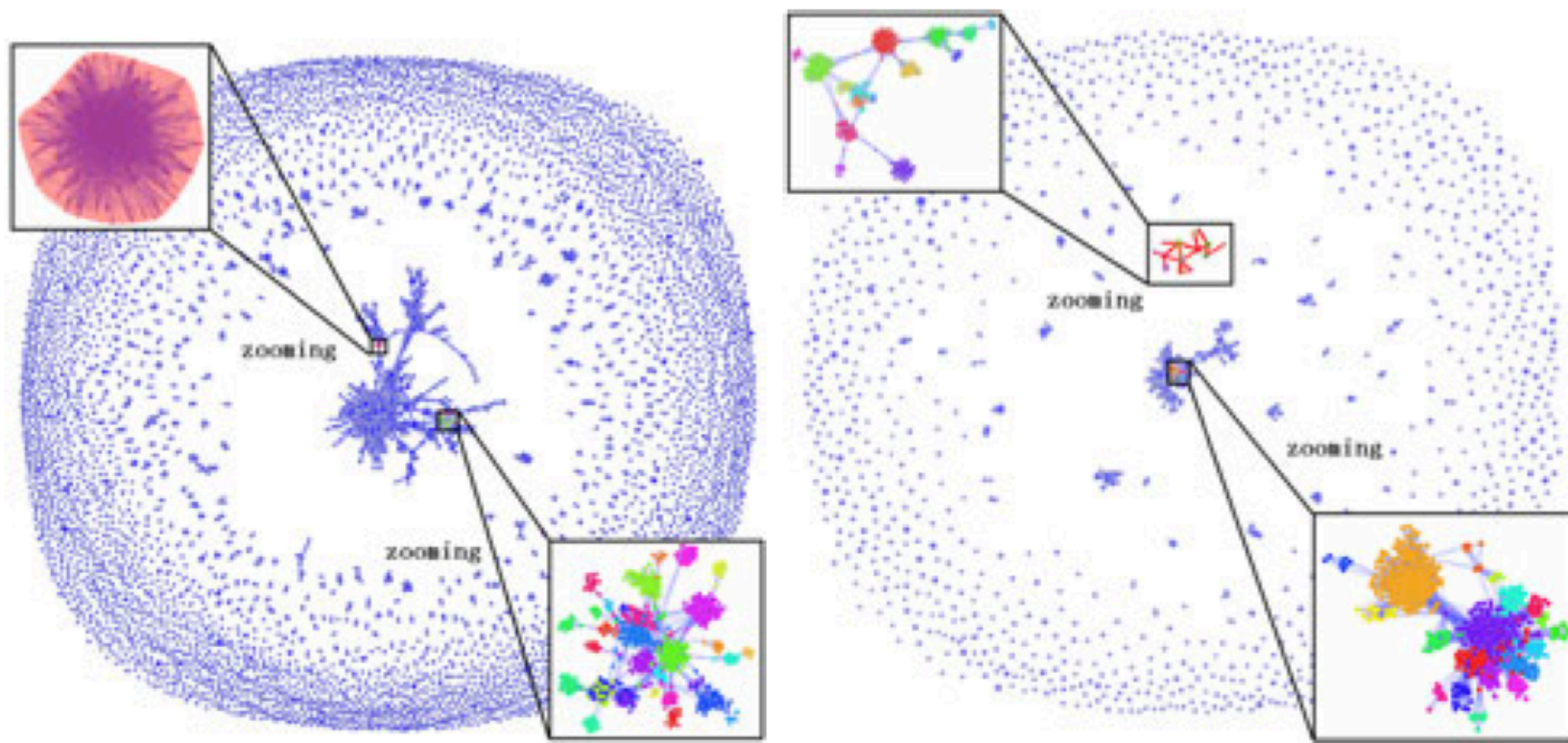
```
select objid ra dec u g r i z
       plate fiberid elodiefeh
from   photoobj fgetnearbyobjeq
       specobj platex
where  objid objid logic objid
       bestobjid logic plateid
       plateid logic class logic
       r logic r logic g logic g
       logic plate
```

(b) Tokenized SQL.



# Mining Call Graphs

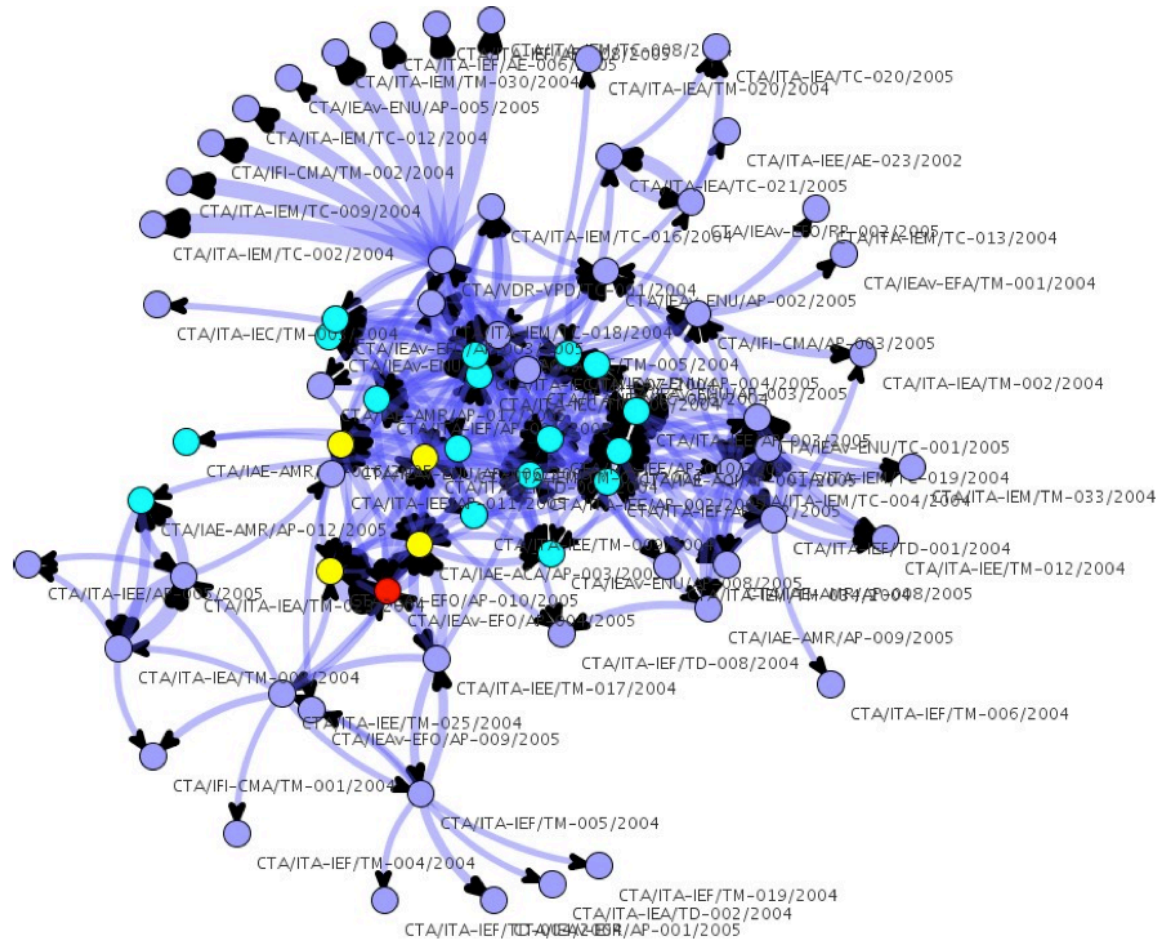
Graphs of reciprocated phone calls (time, duration, callers).



Ye, Qi, Bin Wu, and Bai Wang. "Multiple level views on the adherent cohesive subgraphs in massive temporal call graphs." *Advanced Data Mining and Applications*. Springer Berlin Heidelberg, 2010. 441-452.

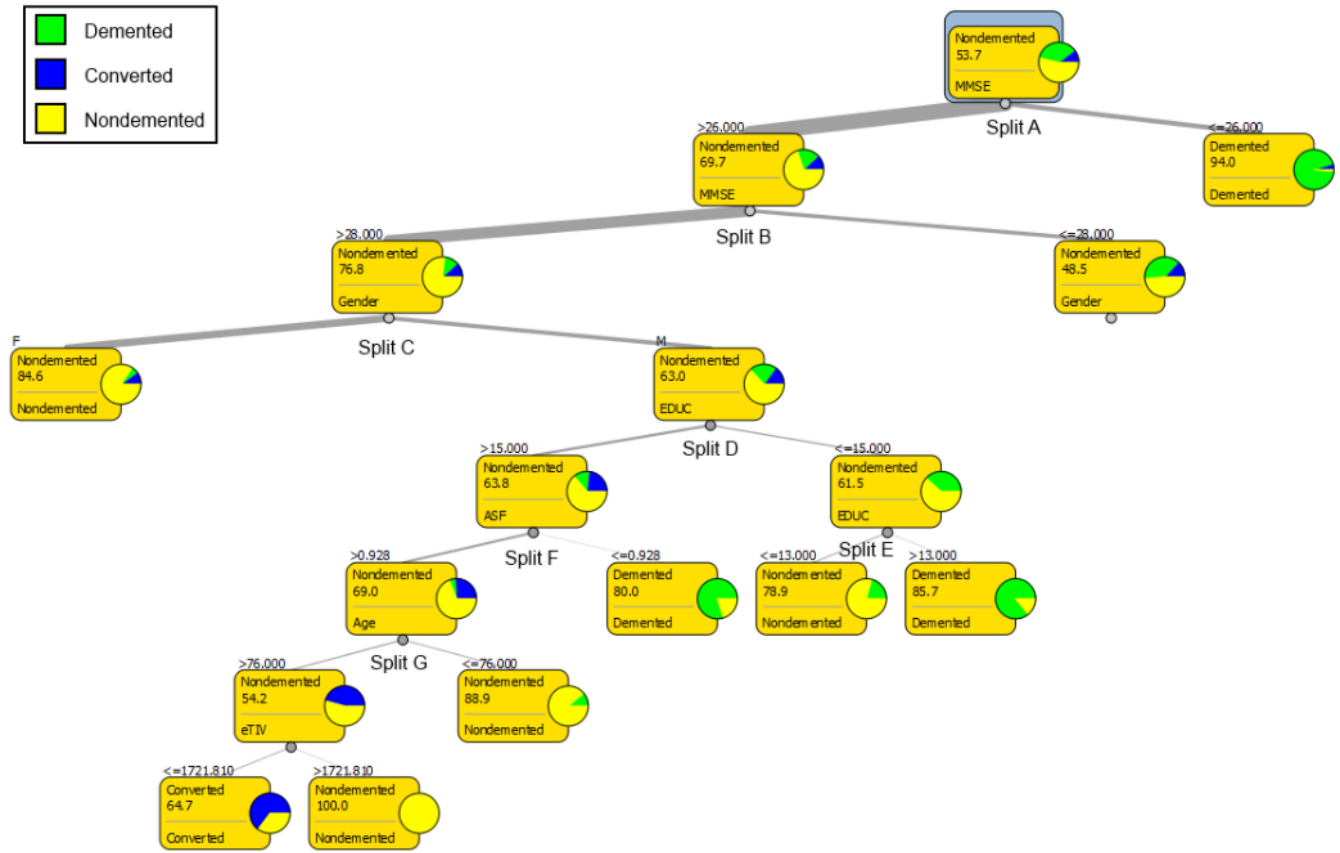
# Visualizing Citation Data

Visual exploration of publications related through citations.



# Mining Medical Data

Decision trees and visualization to identify factors that influence Alzheimer's Disease



# Graph Mining

Finding herbs with associated uses in traditional Chinese medicine.

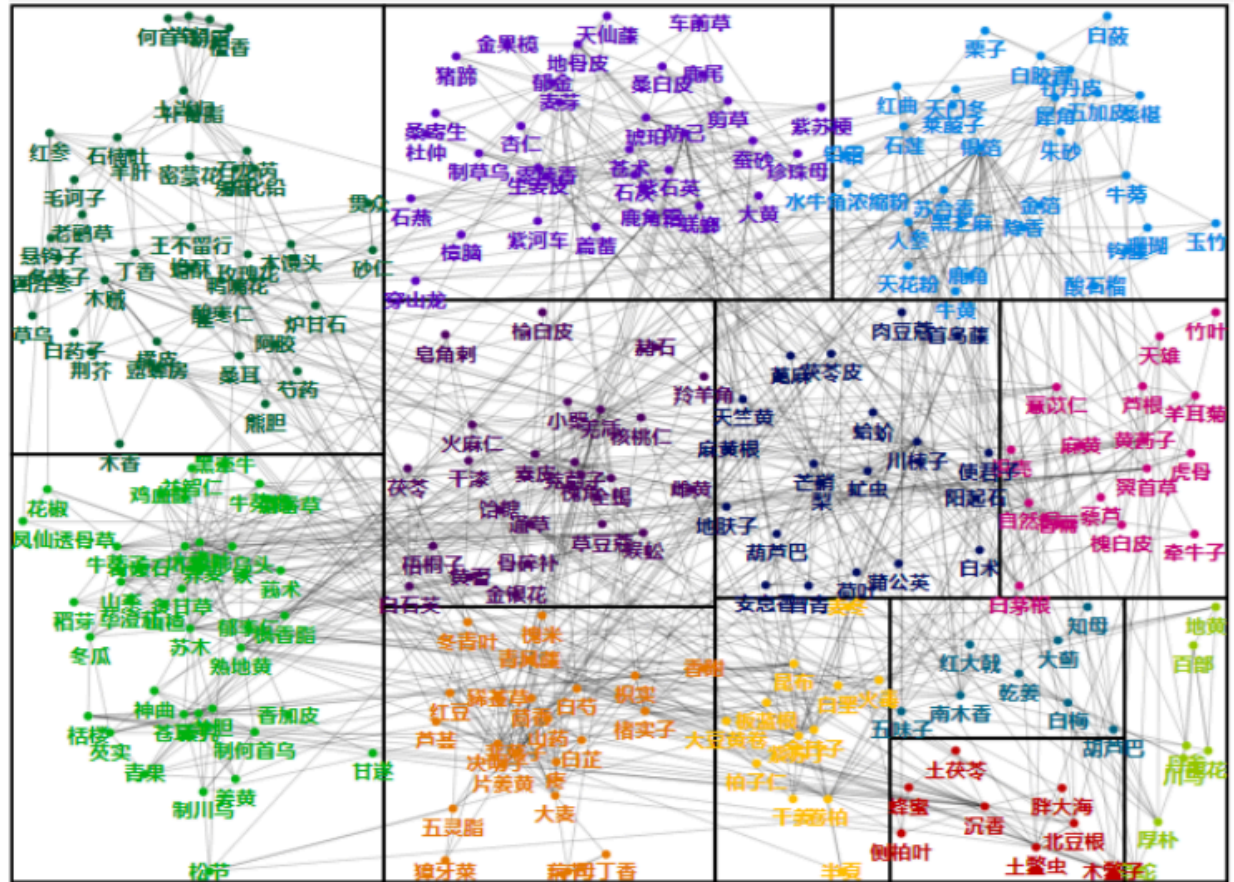
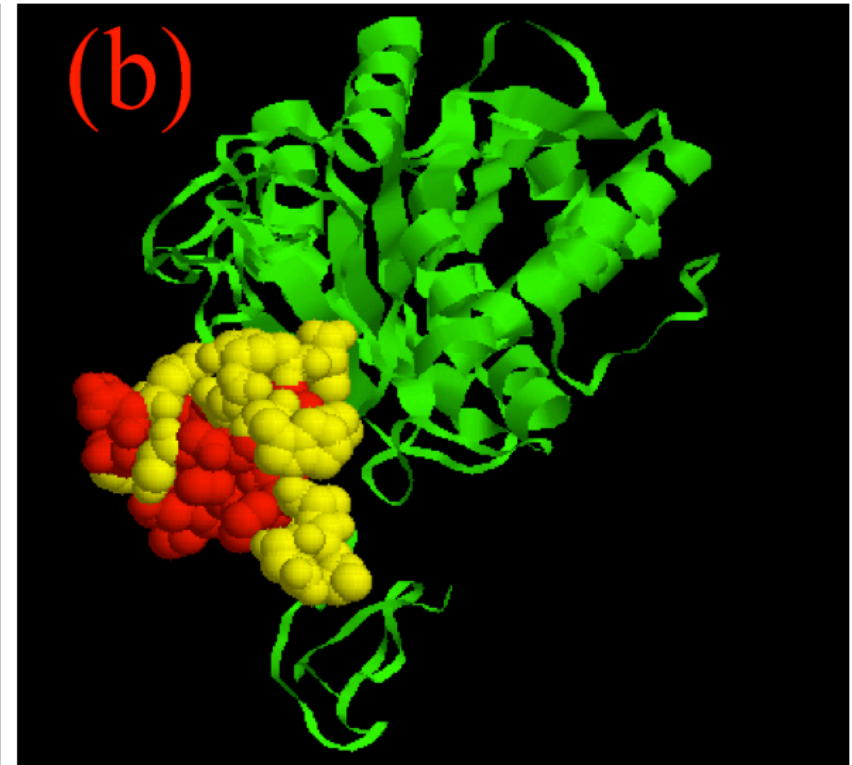
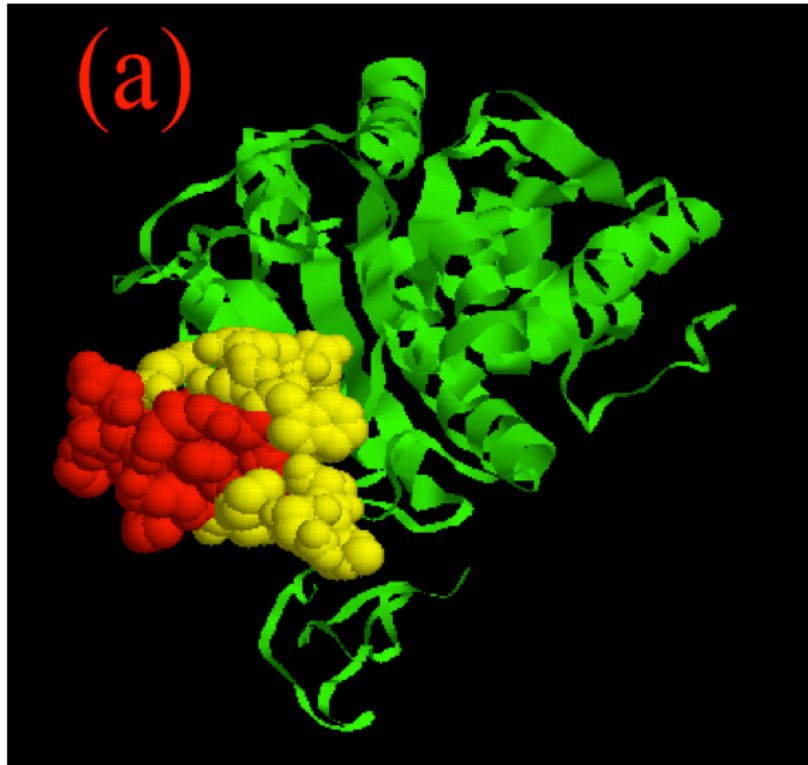


Fig. 5. Groups detected by our method from herbal network

Wang, Lidong, et al. "A method for finding groups of related herbs in traditional chinese medicine." *Advanced Data Mining and Applications*. Springer Berlin Heidelberg, 2011. 55-68.

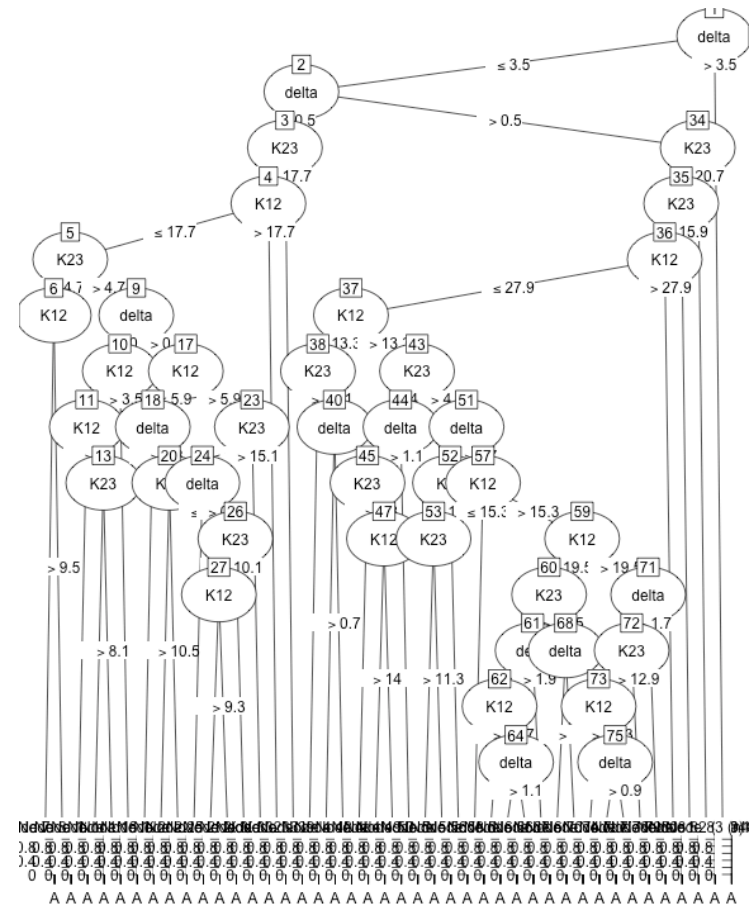
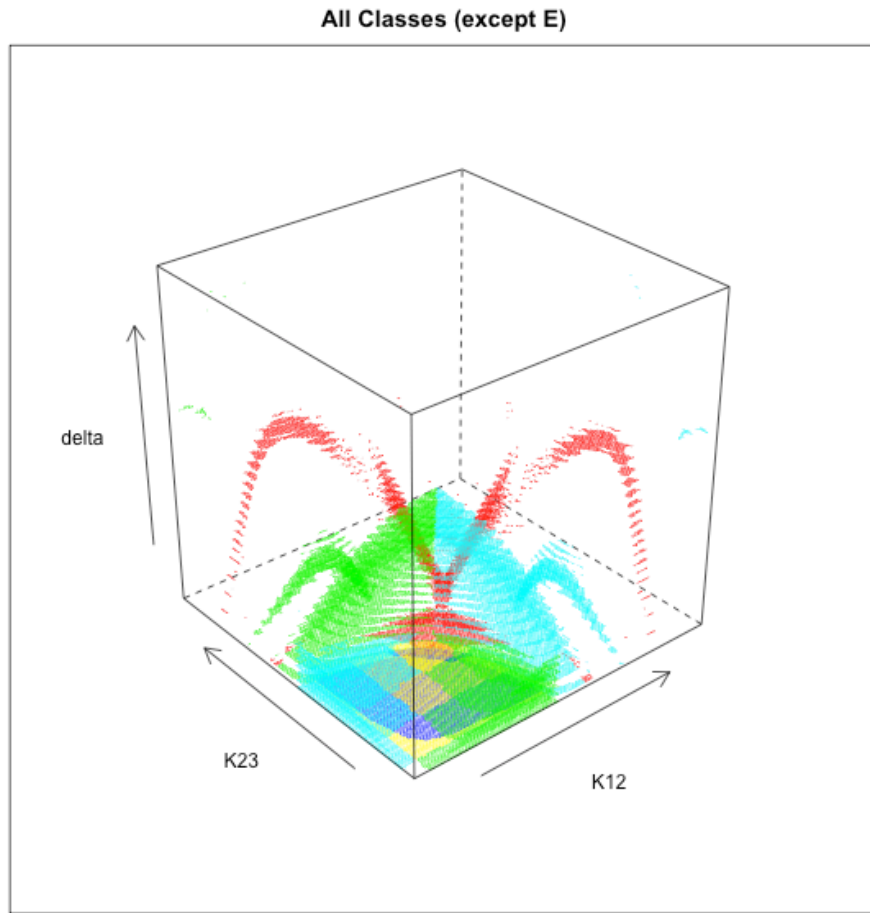
# Bioinformatics Data Mining

Using a classifier to identify protein-protein interfaces.



# Model Mining

Can we derive a simpler model from a complex one?



Juliana Cestari Lacerda, "Sincronização de três osciladores utilizando o modelo de Kuramoto", 2015.



# Principles and Applications of Data Mining

## Class Project

# Class Project

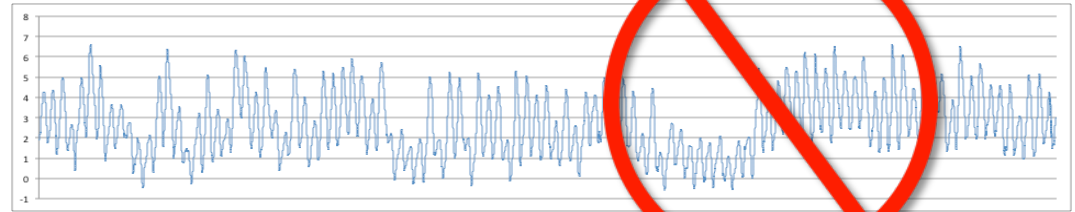
- Each student **must** develop an **individual** data mining project during this course.
- Evaluation and grading will consider the project, the development and a report about it.
- Projects will take some time (yours and mine!): **plan ahead!**
- We may have a seminar with presentations.

# Class Project

- Ask your advisor about a problem, dataset, algorithm, etc. - think about projects related to your thesis or dissertation.
- Projects will be developed with the lecturer, with several meetings that will count as lectures.
  - Check the schedule at <http://www.lac.inpe.br/~rafael.santos/cap359.html>.
  - **Start as soon as possible!**
  - You must **plan ahead** for enough meetings - avoid leaving work to the last week!

# Class Project: Data Checklist

- Do you have a reasonable amount of data?
  - ▣ Can you filter/select it so it becomes reasonable?
- Is it *rich*?
  - ▣ Can you improve it?
- Is it readily accessible?
  - ▣ Where is it?
  - ▣ What about privacy and disclosure issues?
- Do you understand what the data represents?
- Is it tidy?
- **We can mine it.**



# Class Project: Your Data

- First step: talk to your advisor about data.
  - Don't wait too long to do this!
- Second step: collect and organize the data!
  - Make it *ready for processing*.
  - Write a summary about the data and the problem: this will be used for grades (and in your report).
- Schedule a meeting for (brief) discussion around week 2 or 3.



# Class Project: Making it Work

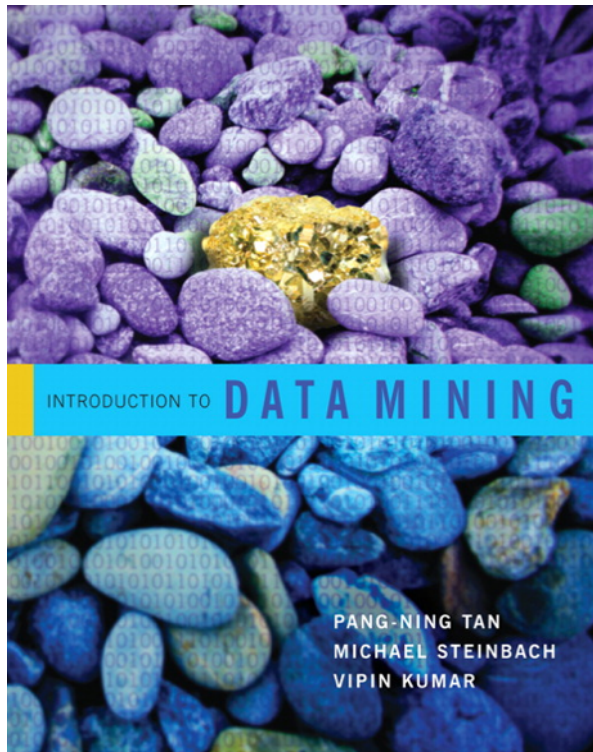
- What could possibly go wrong?
  - No data.
  - Too few data.
  - Too much data.
  - Poor data.
  - Non-mineable data.
  - Untidy data.
  - No advisor.
  - Ill-defined problem.

# Principles and Applications of Data Mining

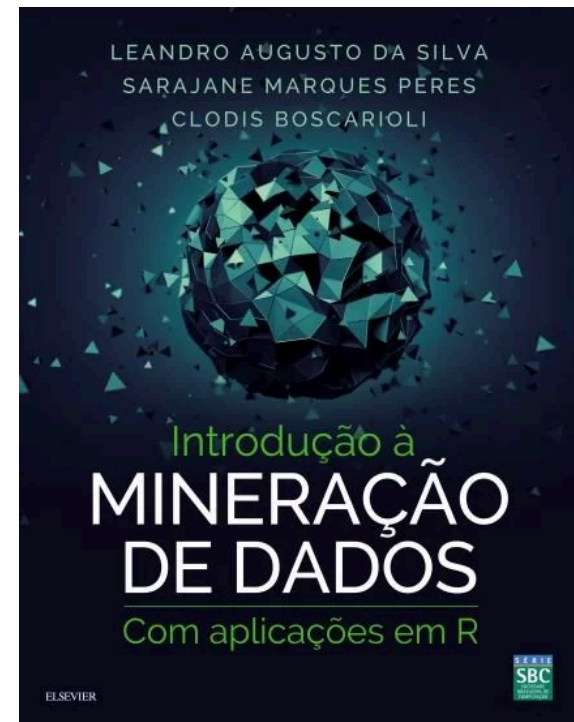
## References

# Recommended Books

*Introduction to Data Mining*; Pang-Ning Tan, Michael Steinbach and Vipin Kumar (2005).



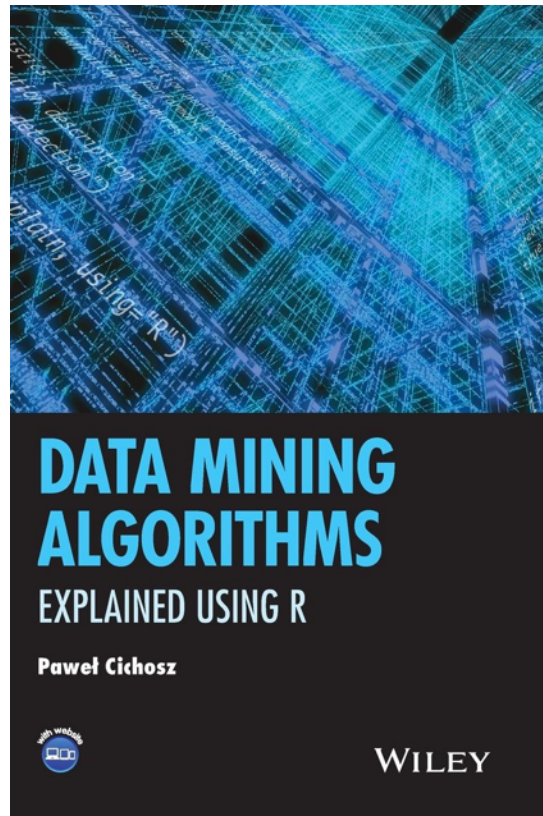
*Introdução à Mineração de Dados com aplicações em R*; Leandro Augusto da Silva, Sarajane Marques Peres e Clodis Boscarioli (2016).



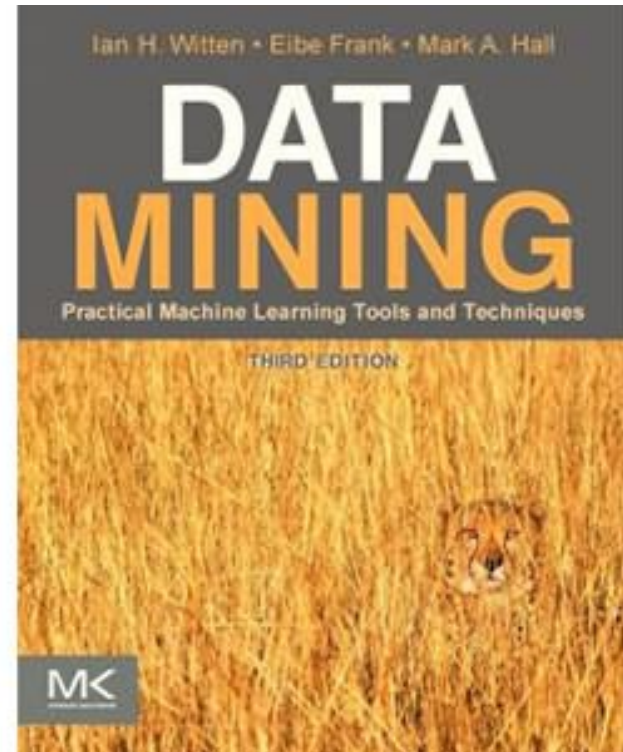


# Recommended Books

*Data Mining Algorithms: Explained Using R*; Pawel Cichosz (2015).

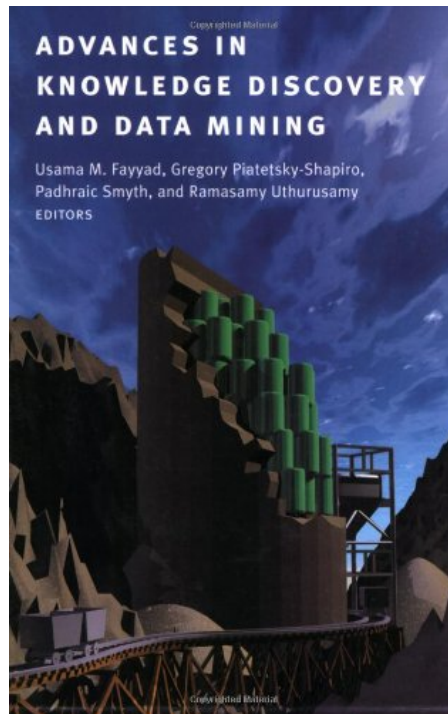


*Data Mining: Practical Machine Learning Tools and Techniques*; Ian H. Witten, Eibe Frank and Mark A. Hall (2011).

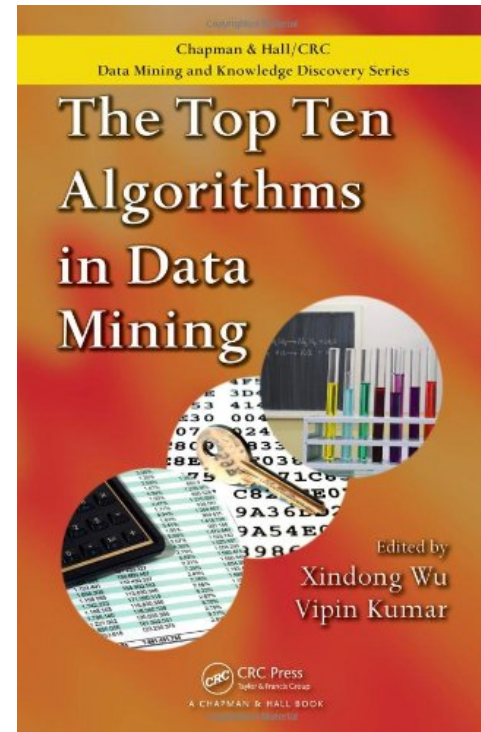


# Recommended Books

*Advances in Knowledge Discovery and Data Mining*; Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth and Ramasamy Uthurusamy (1996).

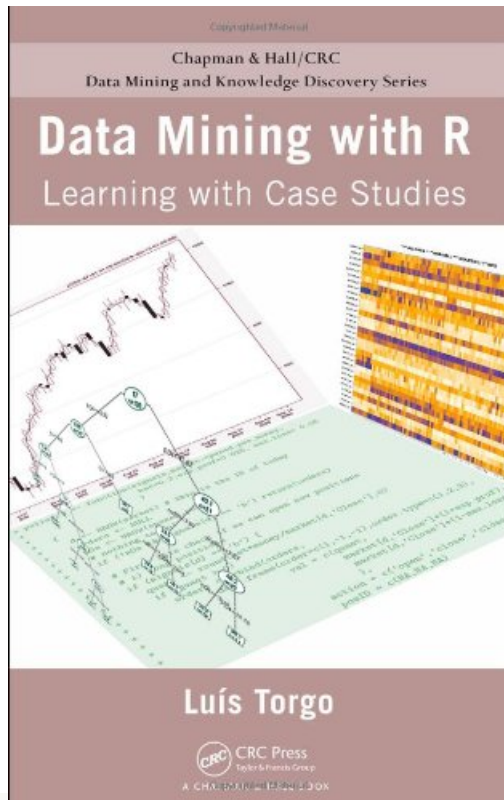


*The Top Ten Algorithms in Data Mining*; Xindong Wu and Vipin Kumar (2009).

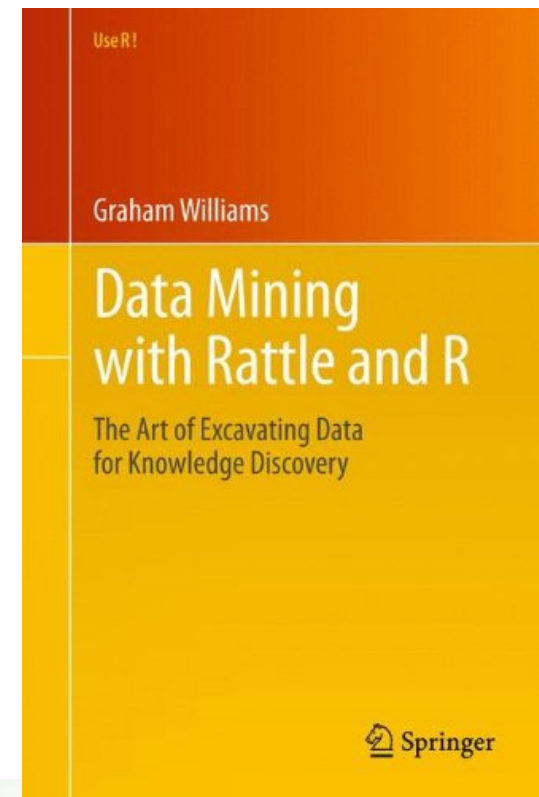


# Recommended Books

*Data Mining with R: Learning with Case Studies*; Luis Torgo (2010).



*Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*; Graham Williams (2011).



# Recommended Sites

- Data sets: <https://archive.ics.uci.edu/ml/datasets.html>
- Challenges/rewards: <https://www.kaggle.com/>
- DM information: <http://www.kdnuggets.com/>
- Coursera: <https://www.coursera.org/>
  
- Many more links and references at [www.lac.inpe.br/~rafael.santos/cap359.html](http://www.lac.inpe.br/~rafael.santos/cap359.html)