



CAP-359 PRINCIPLES AND APPLICATIONS OF DATA MINING

Rafael Santos – rafael.santos@inpe.br
www.lac.inpe.br/~rafael.santos/

Overview

- So far...
 - What is Data Mining?
 - Applications, Examples.
- Let's think about your project: Your Data
 - What is data?
 - Raw and Tidy data.
 - Data Preprocessing.
 - Examples.

Principles and Applications of Data Mining

What is data?

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
 - Examples: eye color of a person, temperature, etc.
 - Attributes are also known as variable, field, characteristic, or **feature**
- A collection of attributes describe an object
 - Object is also known as **record**, point, row, observation, case, sample, entity, or **instance**

Attributes

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objects

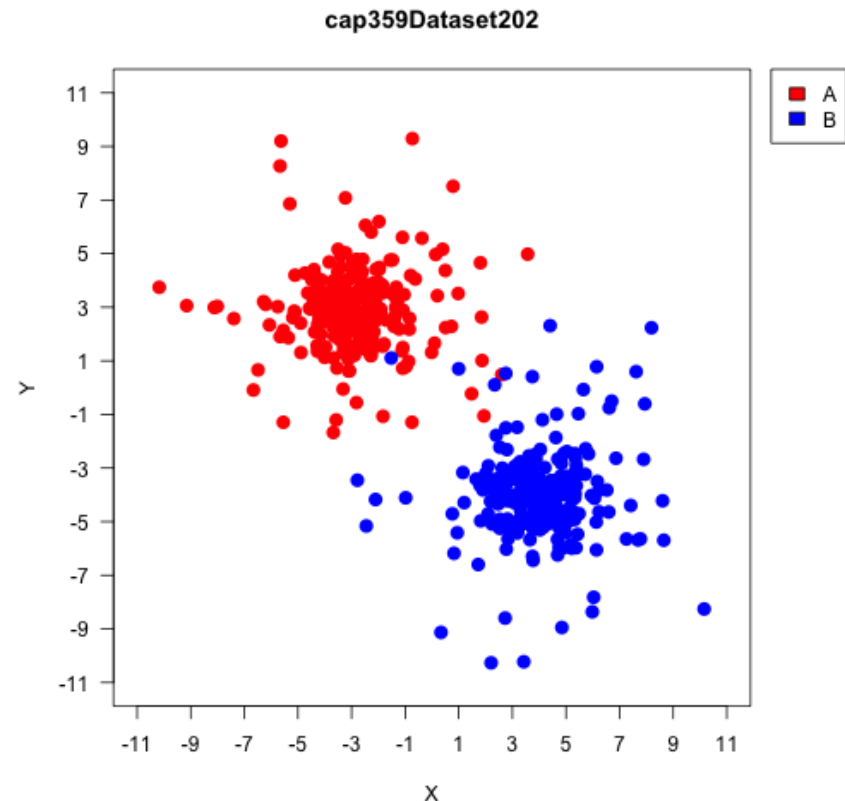
Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

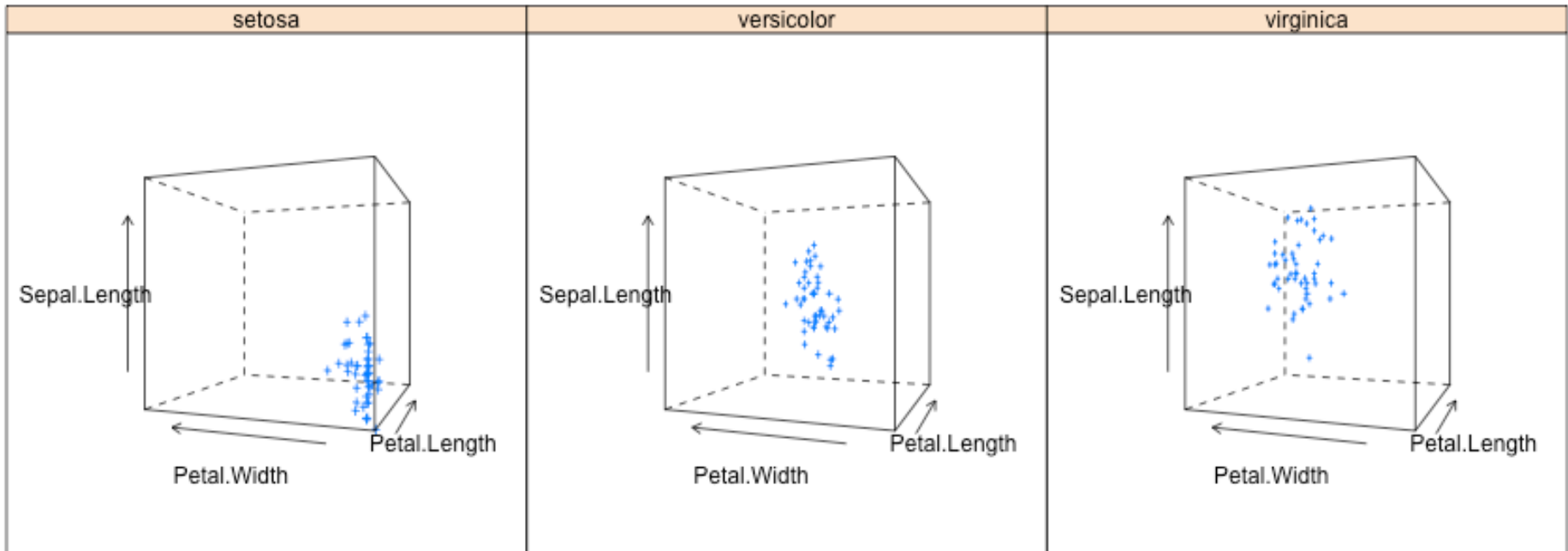
Feature Space

- N-dimensional space where our data can be represented.
- Each record (instance, etc.) is a point in the feature space.
- All records share the same attributes.
- Important concepts:
 - ▣ Distance.
 - ▣ Separability.



Feature Space

- Dimensions (attributes) may not be numeric.
- Limitations for visualization in higher dimensions.
 - ▣ The math is the same!



Types of Attributes

- **Nominal**

- Examples: ID numbers, eye color, zip codes

- **Ordinal**

- Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}

- **Interval**

- Examples: calendar dates, temperatures in Celsius or Fahrenheit.

- **Ratio**

- Examples: temperature in Kelvin, length, time, counts

Properties of Attributes' Values

- The type of an attribute depends on which of the following properties it possesses:
 - Distinctness: = ≠
 - Order: < >
 - Addition: + -
 - Multiplication: * /

- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & addition
- Ratio attribute: all 4 properties

Attribute Type	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., $new_value = f(old_value)$ where f is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	$new_value = a * old_value + b$ where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	$new_value = a * old_value$	Length can be measured in meters or feet.

Discrete and Continuous Attributes

□ Discrete Attribute

- Has only a finite or countable infinite set of values
- Examples: zip codes, counts, or the set of words in a collection of documents
- Often represented as integer variables.
- Note: binary attributes are a special case of discrete attributes

□ Continuous Attribute

- Has real numbers as attribute values
- Examples: temperature, height, or weight.
- Practically, real values can only be measured and represented using a finite number of digits.
- Continuous attributes are typically represented as floating-point variables.

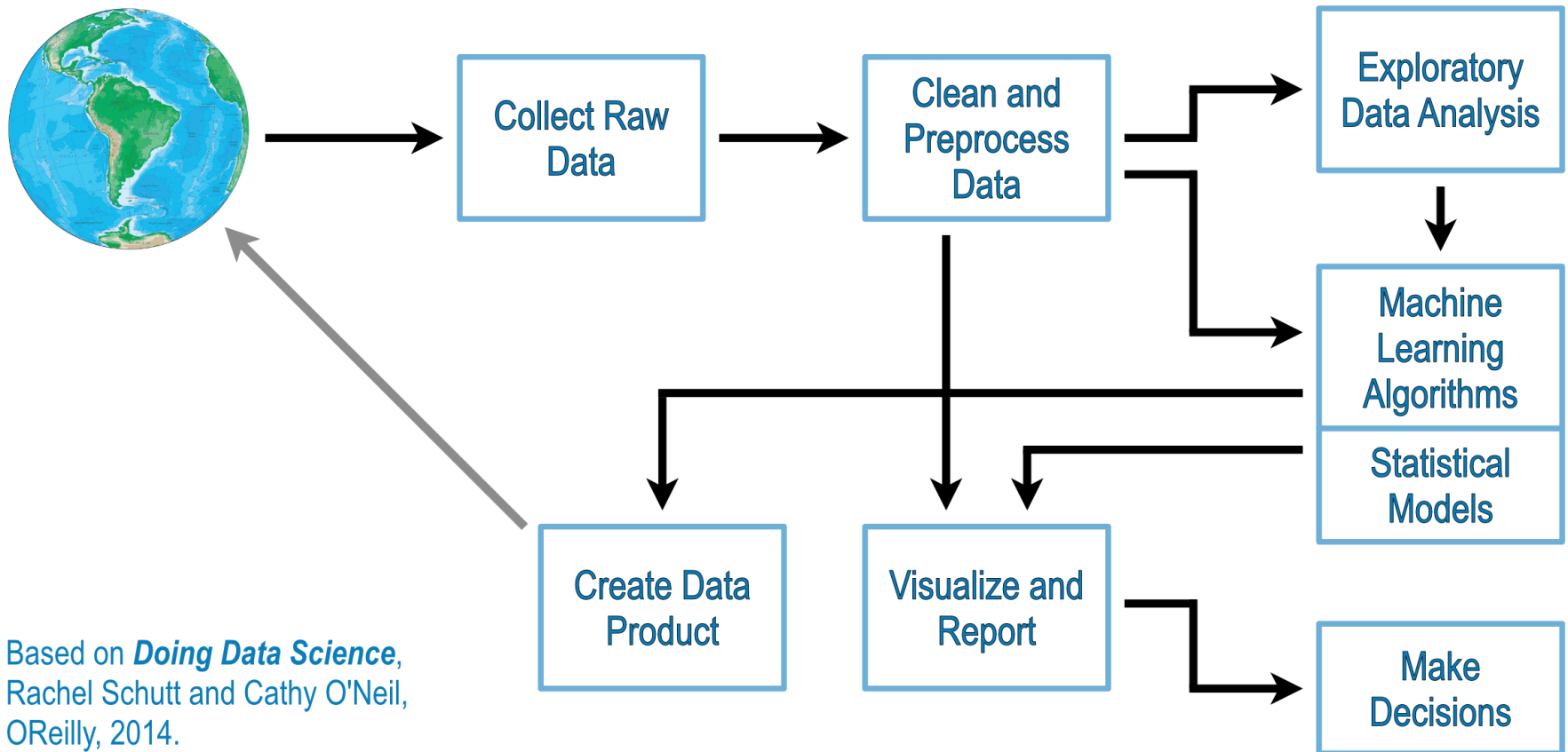
Why does it matter?

- Important because we may need to *preprocess* it.
- Preprocessing may be the most important step in Data Mining:
 - From *what data do I have* to *what data can I mine*.
 - Determine which data mining operation can be applied.

Principles and Applications of Data Mining

Raw and Tidy Data - Some Examples

Why Tidy Data?



Raw Data

- Data from the Real World:
 - Databases, spreadsheets...
 - Images, videos, audio...
 - Time series...
 - Logs, text, JSON files, XML files...



Based on Coursera's "Getting and Cleaning Data" course.

Record/Table Data





- Data that consists of a collection of records, each of which consists of a fixed set of attributes.
- Is it tidy?



<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Record/Table: Tidy Data

- One table with all the data (or linked tables).
 - Each variable in its column.
 - Each observation in its row.
 - Variable names in the first row, with good, clear names.
- “Tidiness” is not an absolute feature!
 - Depends on what we have and what we want to do.

  incidentLocation	  location
4000 MT PLEASANT AV	(39.2910056,-76.5636502)
1000 W BALTIMORE ST	(39.2889640,-76.6339440)
1000 STAMFORD RD	(39.2963480,-76.7101510)

Based on Coursera’s “Getting and Cleaning Data” course.

Record/Table: Tidy Data

- If our data is in a table and it is tidy we can easily* apply several data mining algorithms on it**:
 - Clustering.
 - Classification.
 - Regression.
 - Association.
 - Visualization.

* Maybe. It depends on many factors.

** Depending also on the type of the attributes!

Document Data

- Documents can be converted to term vectors.
 - Each term is a component (attribute) of the vector,
 - The value of each component is the number of times the corresponding term occurs in the document.
 - Post-processing, e.g. TF-IDF.


Document Data

```
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.



select_objid	1
select_ra	1
select_dec	1
select_u	1
select_g	1
select_r	1
select_i	1
select_z	1
select_plate	1
select_fiberid	1
select_elodiefeh	1
from_photoobj	1
from_fgetnearbyobjeq	1
from_specobj	1
from_platex	1
where_objid	3
where_logic	8
where_bestobjid	1
where_plateid	2
where_class	1
where_r	2
where_g	2
where_plate	1

*Text mining applied to SQL queries: a case study for SDSS
SkyServer.* Makiyama, V. H.

Document Data

- Term vectors are tables!
 - Each document is an instance, each term an attribute.
 - Often automatically tidy.
- Not limited to only terms and counts/frequencies.
- Additional attributes can represent:
 - Time, location.
 - Context.
 - IDs / Classes.
 - Hierarchies.

Transaction Data

- Variation of record data:
 - Metadata plus set of items.
 - Classic example: market basket analysis.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

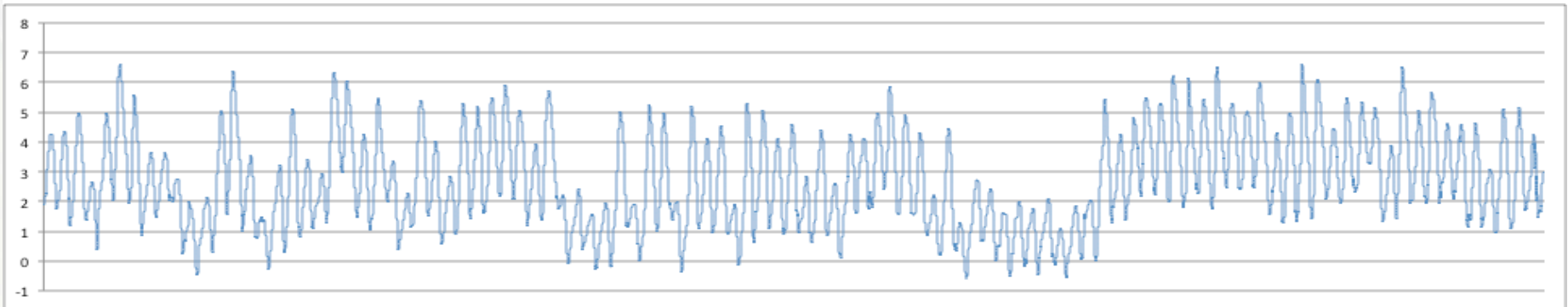
Transaction Data

- Real cases are way more complex...
 - Preprocessing / item mapping and annotations may be required.
 - Temporal indication may be very relevant.

Loja	Caixa	Transação	Compras
03	05	011672	'PAO FRANCES'
03	05	011673	'PAO FRANCES'
03	06	010169	'PAO FRANCES'
03	05	011674	'PAO FRANCES', 'FLV PIMENTAO VERDE', 'LEITE PAST. SERRAMAR S', 'DANONE DANETTE CHOCOLA', 'ADOCANTE DOCE MENOR LI'
01	14	003752	'PAO FRANCES', 'LEITE PAST. SERRAMAR S'
01	14	003758	'BEB. REF.COCA COLA 1L', 'PAO FRANCES', 'LEITE PAST. PAULI TIPO'
01	13	003001	'LEITE PAST. PAULI TIPO', 'PAO FRANCES'
03	05	011685	'PAO FRANCES', 'PAO FRANCES'
01	14	003764	'ACUCAR REFINADO UNIAO', 'FEIJAO PRETO TARUMA 1K', 'PAO FRANCES'
03	05	011688	'PAO FRANCES'
01	14	003765	'BISC. TRIUNFO C.CRACKE', 'BISC. BAUD.GULOSOS 170', 'PAO FRANCES', 'ACUCAR REFINADO A. ALE', 'MORTADELA MARBA'
03	06	010188	'PAO FRANCES', 'ACUCAR REFINADO A. ALE'

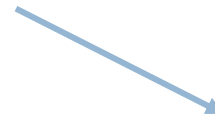
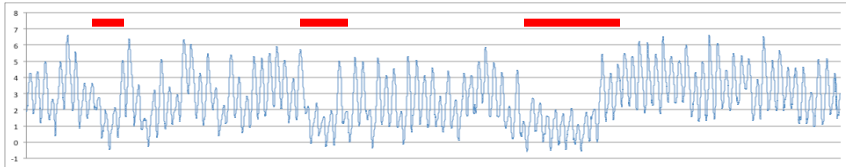
Time Series Data

- Can be considered as a table, with an explicit temporal attribute.
 - ▣ Other attributes: what was measured at that time.
- Consider richness of data: how many attributes associated to the time?



Time Series Data

- Often used for prediction and association.
- Different approaches for different problems:
 - Windowing (each object is a slice of the time series).
 - Change in representation (ex. Fourier descriptors, Wavelets).



id	min	max	avg	event
1	8	45	17	Y
2	7	32	12	N

Time Series (Coverage) Data

- Metadata.

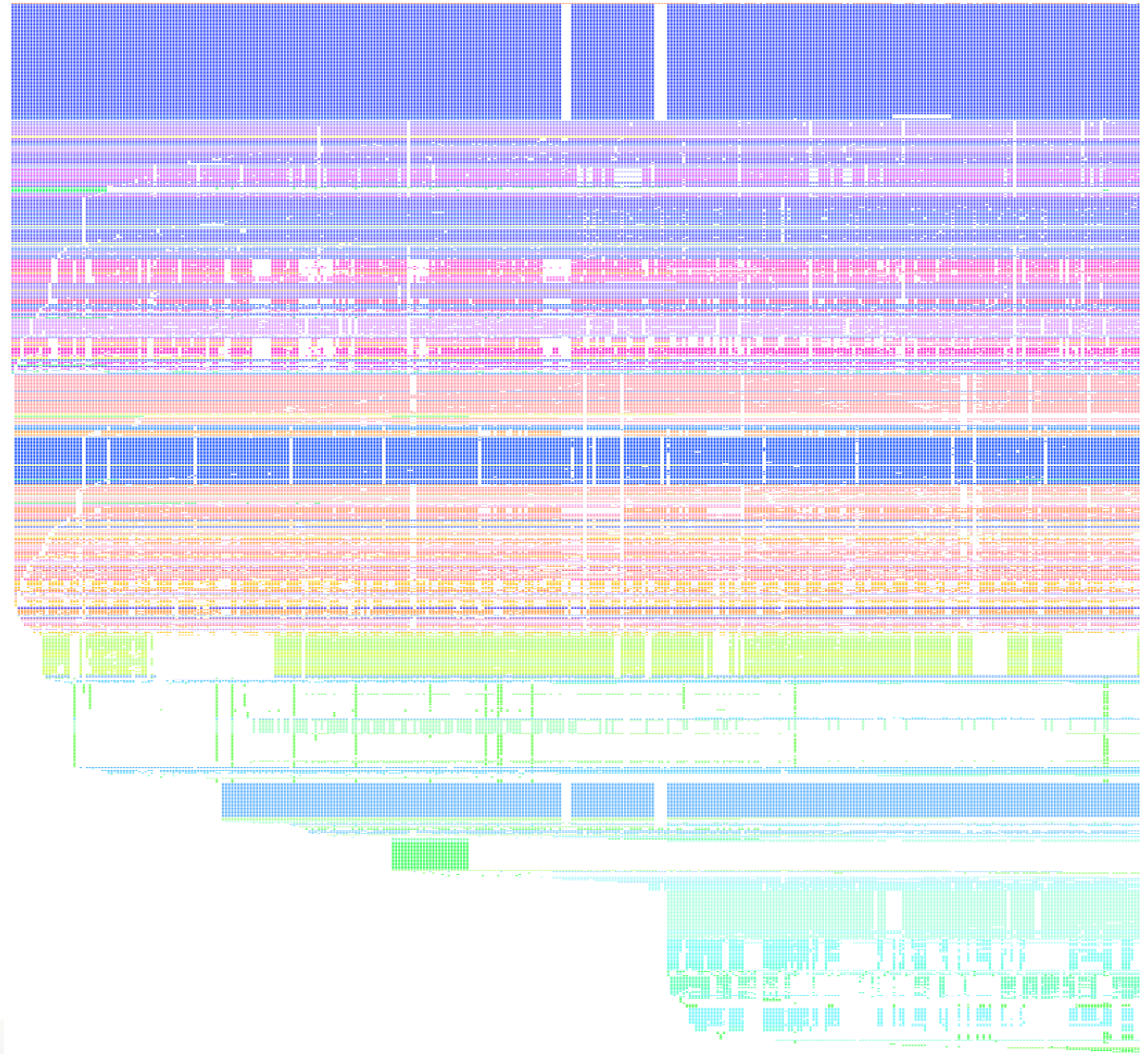


Image Data

- There is no “pixel data mining”: we need to extract features from the images.
- Many different approaches:
 - ▣ Local descriptors, texture, Hough, snakes, etc.
 - ▣ Signal processing (quantization, wavelets, Fourier, etc).
 - ▣ Bag-of-words (then problem is similar to text mining!)

Image Data

- Keypoints from images:
 - ▣ SIFT: Scale-invariant Feature Transform
 - ▣ SURF: Speeded-up Robust Features

Automatic Remote-sensing Image Registration Using SURF. Bouchiha, R. & Besbes, K..

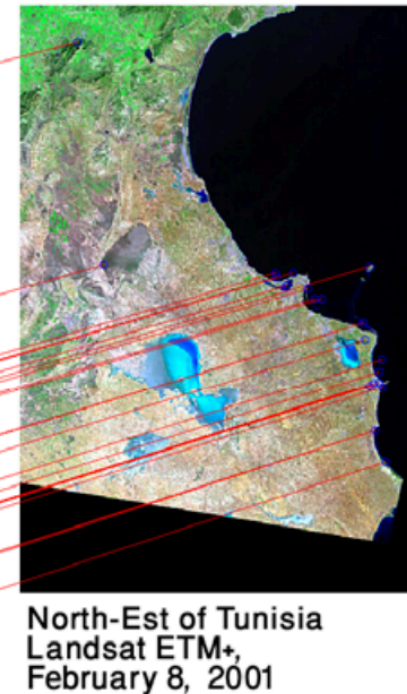
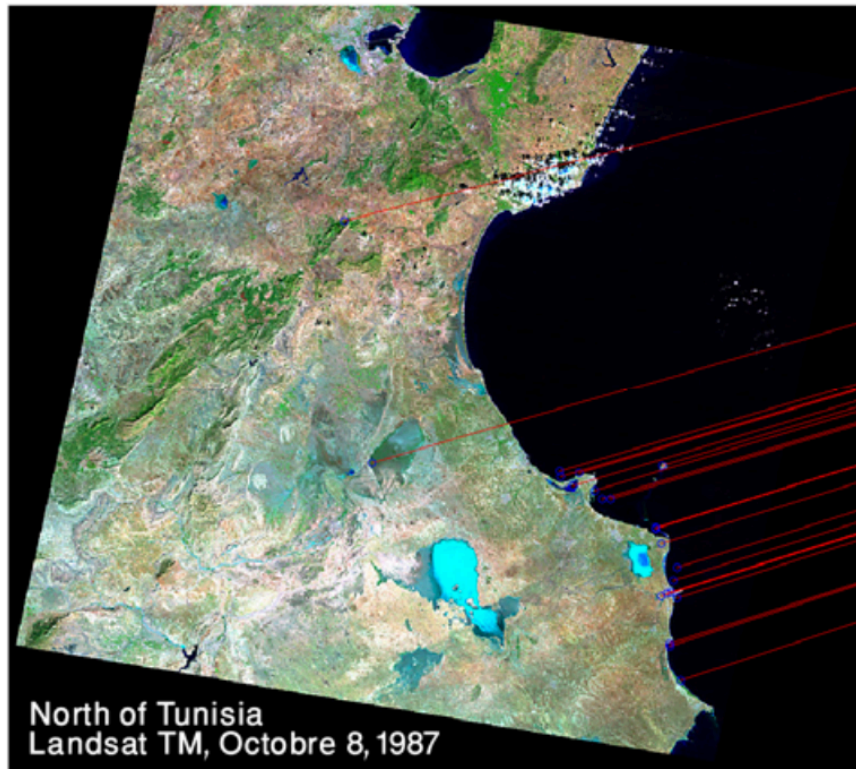
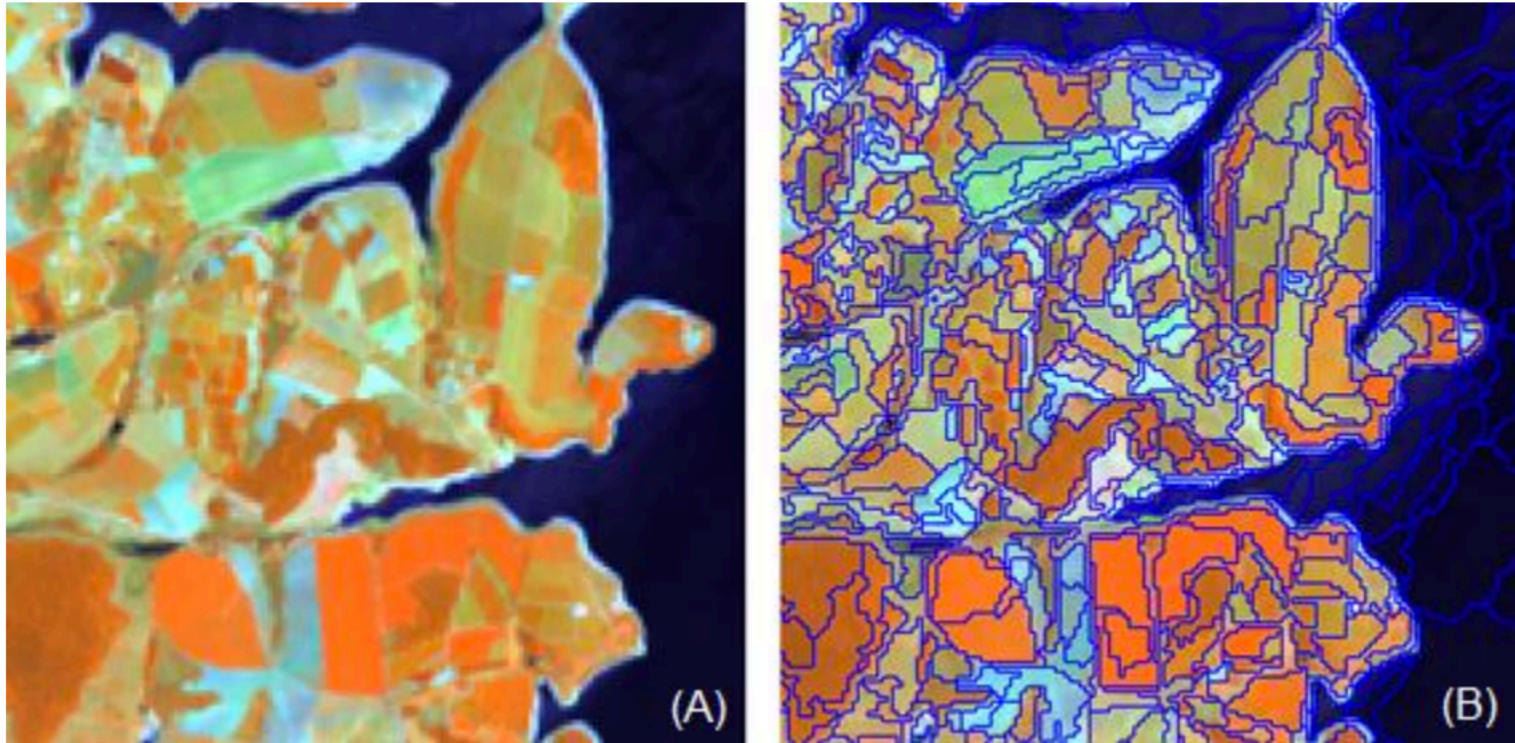


Image Data

□ Segmentation / OBIA



Influência dos atributos espectrais, texturais e fator de iluminação na classificação baseada em objetos de áreas cafeeiras, Marujo, R. F. B.

Graph Data

- Data represented as Graphs (Networks)
- Objects and their relations.
- Not covered on this course (**yet!**)

Issues with Data

- Data may not be easily accessible
 - E.g. Lattes CV Database
- Data may be unstructured or disorganized
 - E.g. IMDB database dump
- Data structure may be complex
 - E.g. software repositories
- Real nature may be much more complex
 - E.g. tertiary and quaternary structures in molecules

Principles and Applications of Data Mining

References

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