AMD

Aprendizagem e Mineração de Dados

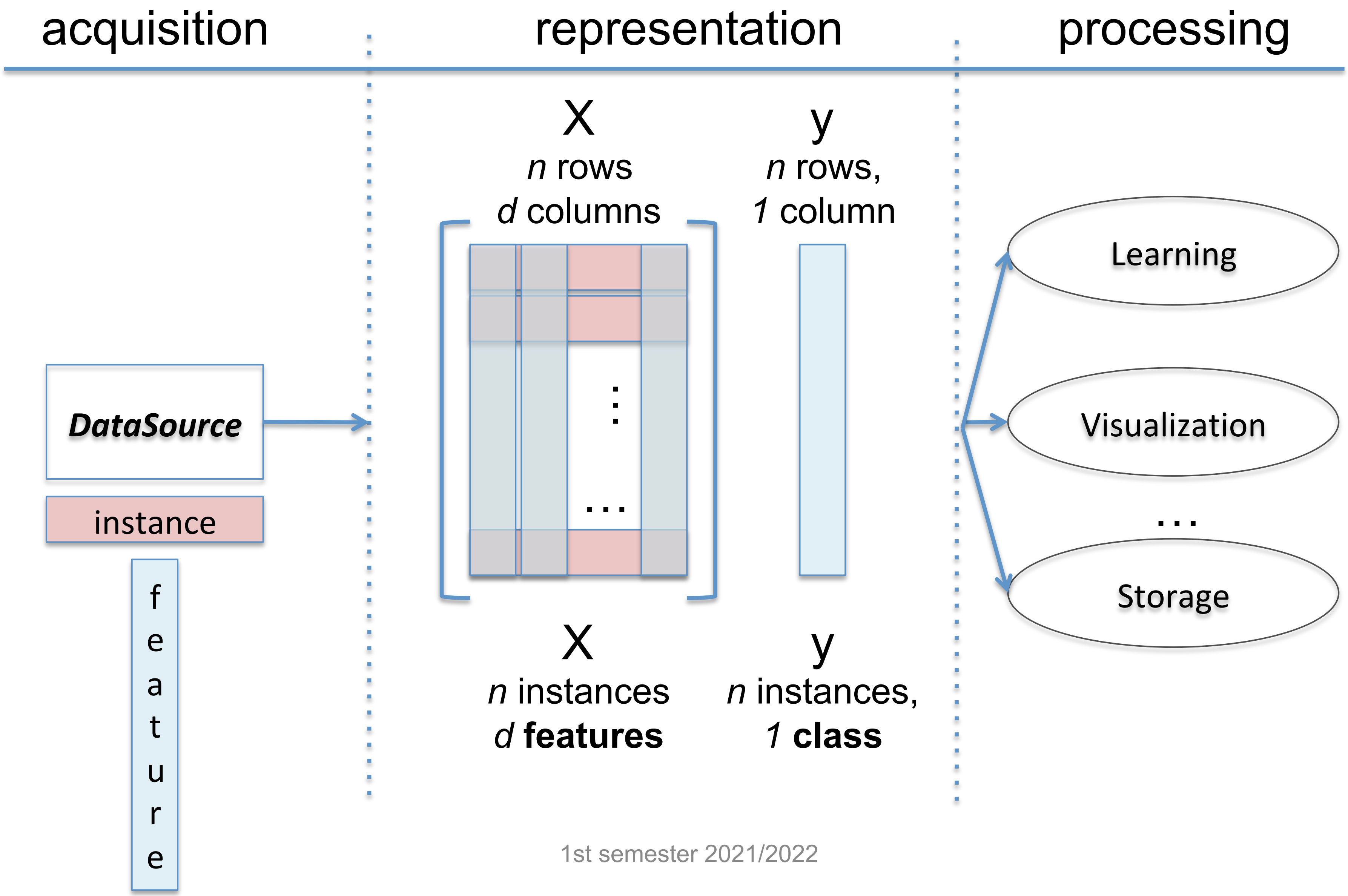
(Machine Learning and Data Mining)

Data and Data Representation

Summary

- Data and Data Mining Tasks
- Data Representation

Data and Data Mining Tasks



Data – Some Terminology

- The data is organized into instances/examples/individuals **x**
- Each instance is a vector with *d* elements **x=[x1, x2, …, x*d*]**
- Each element of **x** is the value of a **feature** (a.k.a an **attribute**)
- Each feature (attribute) represents a given measure
- A **dataset** is composed by *n* instances, each with *d* features

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	70	96	False	Yes
Rainy	68	80	False	Yes
Rainy	65	70	True	No

Supervised Learning

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	70	96	False	Yes
Rainy	68	80	False	Yes

- Each instance may also be described by a corresponding **class** label
  - or, in other words, we have prior-knowledge (an "oracle") that tells us,
  - ... of **what the output values** for our samples should be
- So (in this sense) **supervised** learning is done using a **ground-truth**
- A **class** label may have a
  - binary domain (binary problem); e.g., {0, 1} or {-1, 1}, or {yes, no}, or ...
  - domain with more than two values (multiclass problem); e.g., 0, 1, 2, ..., M-1

Unsupervised Learning

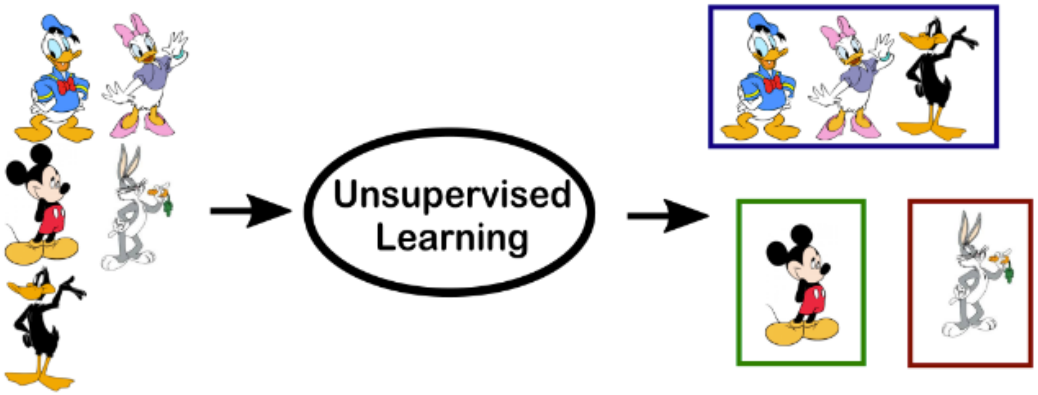
- If each instance is **NOT** described by a corresponding class label
  - we have **NO** prior-knowledge,
  - ... the **instances** themselves represent the “structure” within the data
- So (in this sense) unsupervised learning aims to **extract inherent structure** of data **without using explicitly provided labels**
- … therefore, unsupervised learning is useful in **exploratory analysis** because it can automatically **identify structure** in data
  - e.g., if an analyst were trying to segment consumers, unsupervised clustering methods would be a starting point
  - … where it is impossible or impractical for a human to propose trends in data, unsupervised learning can provide initial insights that can then be used to test individual hypotheses

Unsupervised Learning

search for "structural" resemblances (similarities)

among

instances and group them together...

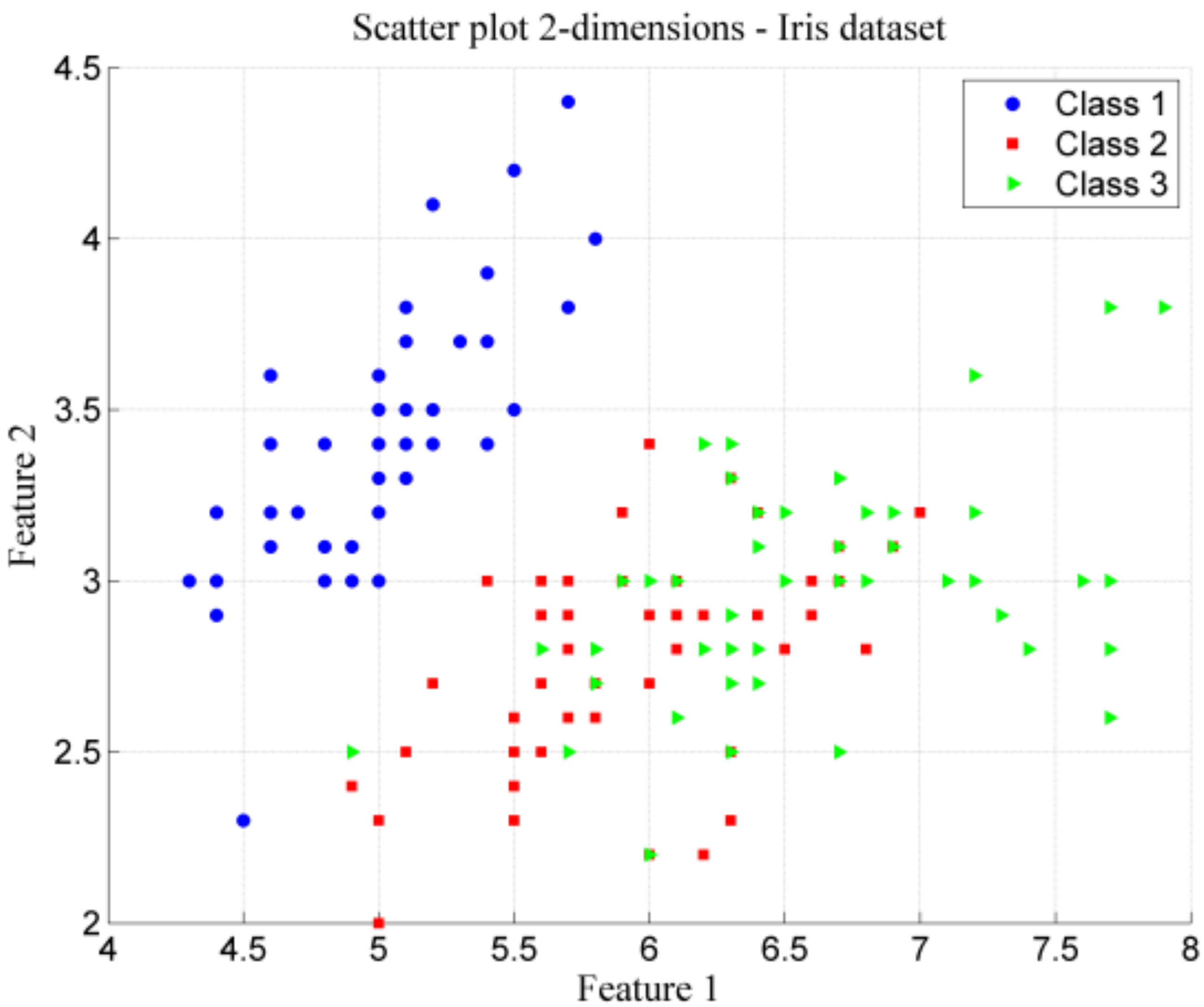


	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
...				
51	7.0	3.2	4.7	1.4
52	6.4	3.2	4.5	1.5
53	6.9	3.1	4.9	1.5
54	5.5	2.3	4.0	1.3
55	6.5	2.8	4.6	1.5
...				
101	6.3	3.3	6.0	2.5
102	5.8	2.7	5.1	1.9
103	7.1	3.0	5.9	2.1
104	6.3	2.9	5.6	1.8
105	6.5	3.0	5.8	2.2

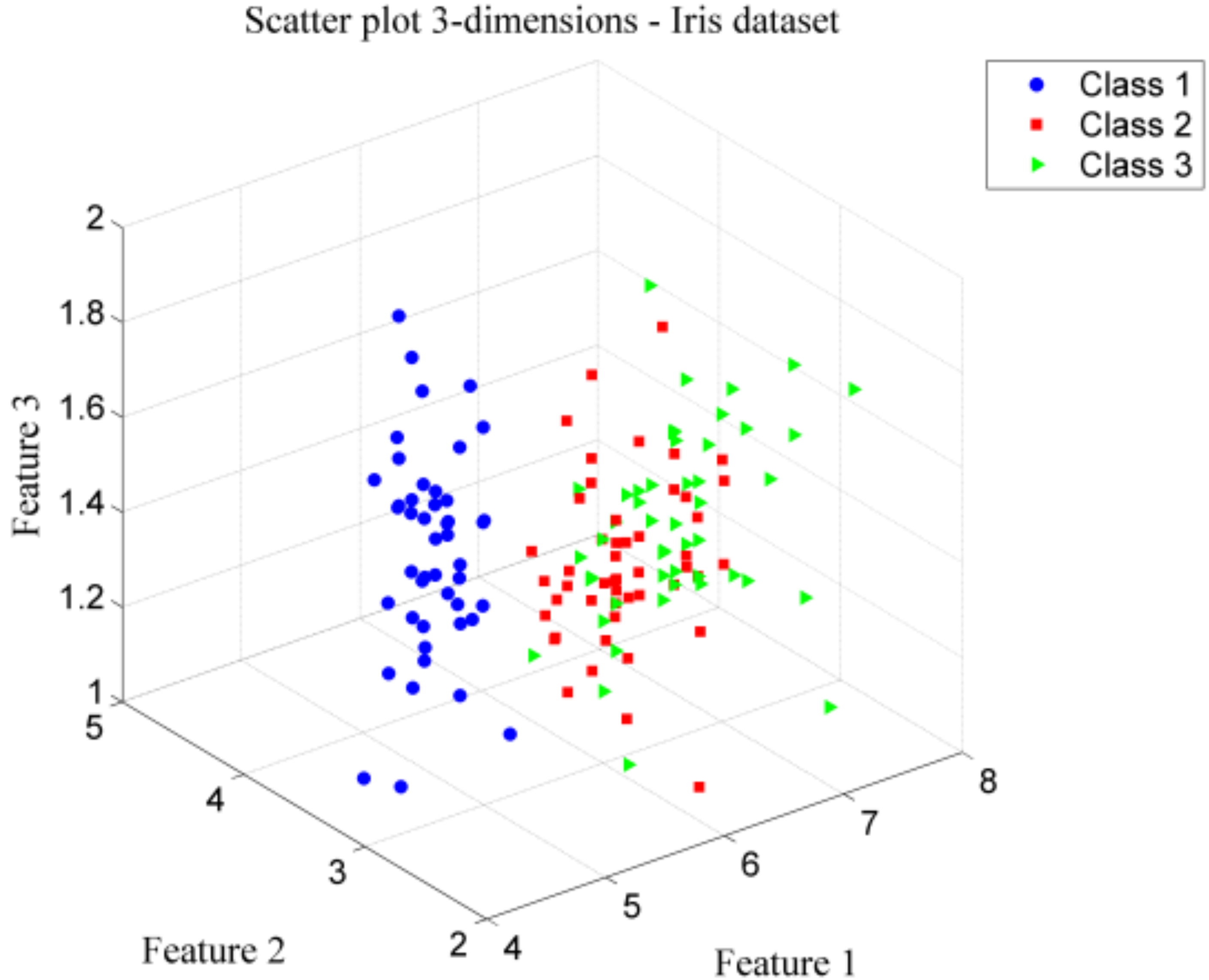
... Data – about the Features (Attributes)

- Should be in lowest number as possible
- Should be as discriminative as possible
- Should be relevant
- Should not be redundant, on the presence of others

Data – about Visualization (2 features, 3 classes)

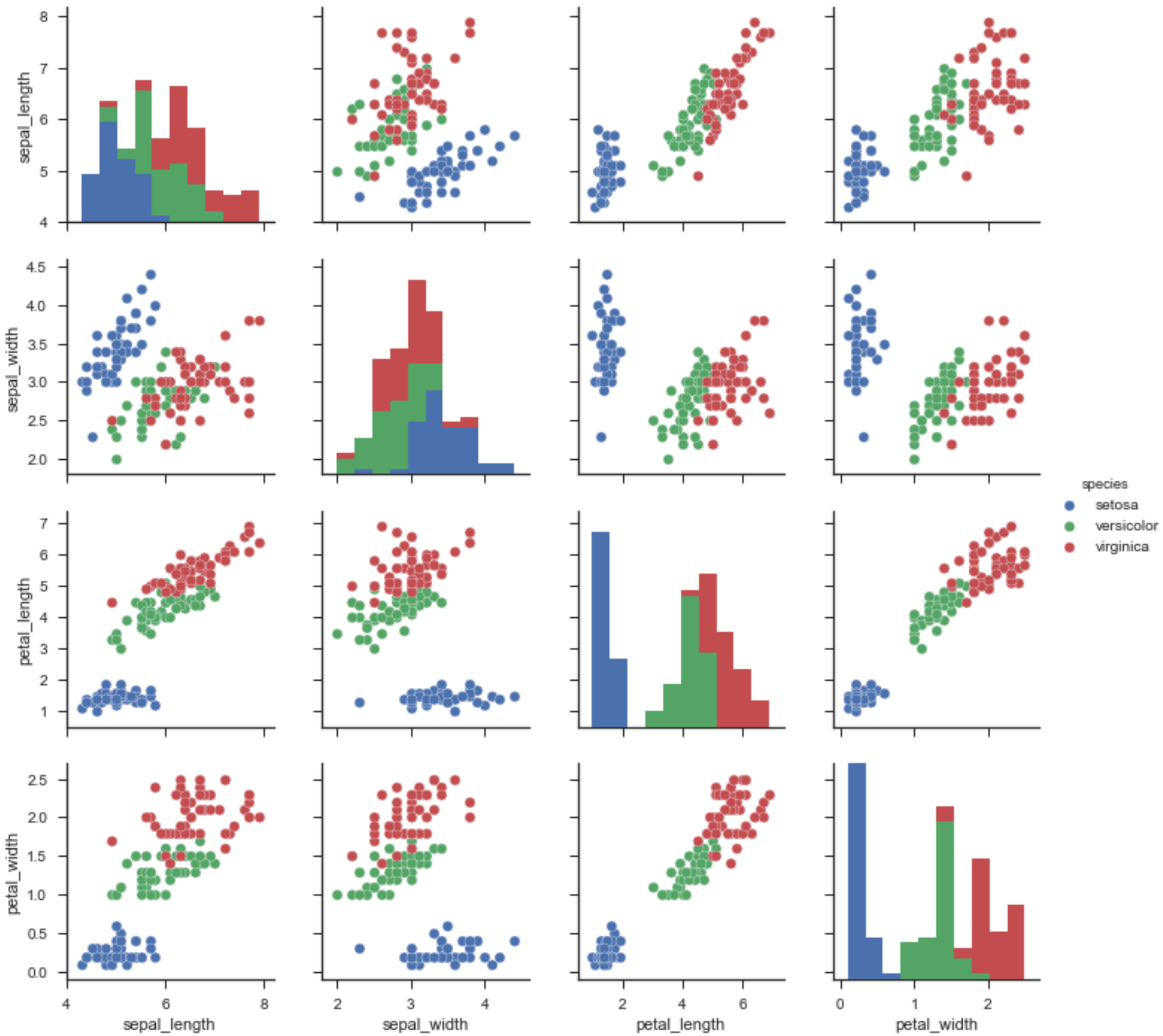


Data – about Visualization (3 features, 3 classes)



Data – about Visualization (all 2 features, 3 classes)

scatter
matrix



Data Representation (ARFF)

- ARFF (Attribute-Relation File Format) file is an **ASCII text** file that describes a list of **instances** sharing a set of **attributes**.
- ARFF files have two distinct sections:
  - the first section is the **Header** (*meta-data*)
  - the second section is the **Data**.

ARFF files were developed by the Machine Learning (ML) Project at the Department of Computer Science of The University of Waikato for use with the Weka ML software.

`https://www.cs.waikato.ac.nz/ml/weka/arff.html`

`https://waikato.github.io/weka-wiki/formats_and_processing/arff_stable/`

Data Representation (ARFF – header example)

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
% (a) Creator: R.A. Fisher
% (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
% (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}
```

Data Representation (ARFF – data example)

```
@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
...
```

Data Representation (Orange-DM)

- Orange-DM native data format is a tab-delimited text file with three header (*meta-data*) rows:
  - first row lists **attribute** names
  - second row defines their **domain** (continuous – c, discrete – d and string – s)
  - third row is an optional type (class, meta, or ignore)
- Orange-DM also supports a condensed single-line header format
  - feature names prefixed by an optional “<flags>#” string

Orange-DM (Data Mining) is an open source machine learning and data visualization.

Allows to build data analysis workflows visually, with a large, diverse toolbox.

`https://orangedatamining.com/`

`https://orange3.readthedocs.io/projects/orange-data-mining-library/en/latest/reference/data.io.html/`

Data Representation (Orange-DM example)

sepal length	sepal width	petal length	petal width	iris
c	c	c	c	d
				class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa

age	prescription	astigmatic	tear_rate	lenses
discrete	discrete	discrete	discrete	discrete
				class
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none