UNPHU
Universidad Nacional
Pedro Henríquez Ureña

Dirección de
Investigación

Semana Dominicana de Ciencia y Tecnología 2023 y XVIII Congreso Internacional de Investigación Científica (XVIII CIC-2023)

Taller: Implementando algoritmos de Aprendizaje Automático sin la necesidad de programar

Ing. Christian López, PhD

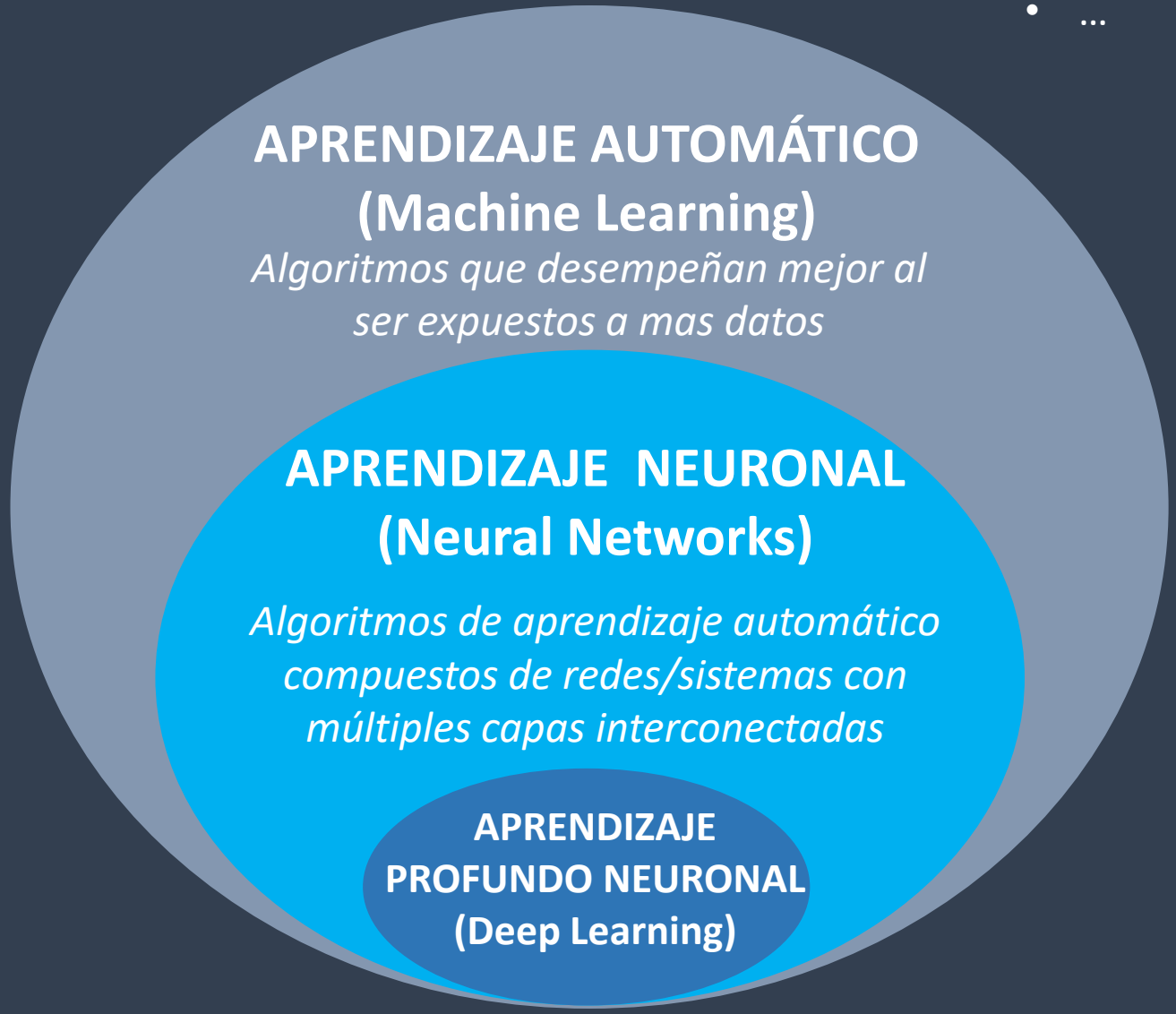
Profesor de Ciencias Computacionales, con afiliación en Ingeniería Mecánica, Lafayette College, USA

sites.lafayette.edu/lopezbec lopezbec@Lafayette.edu @C_LopezB github.com/lopezbec

UNPHU
Universidad Nacional
Pedro Henríquez Ureña

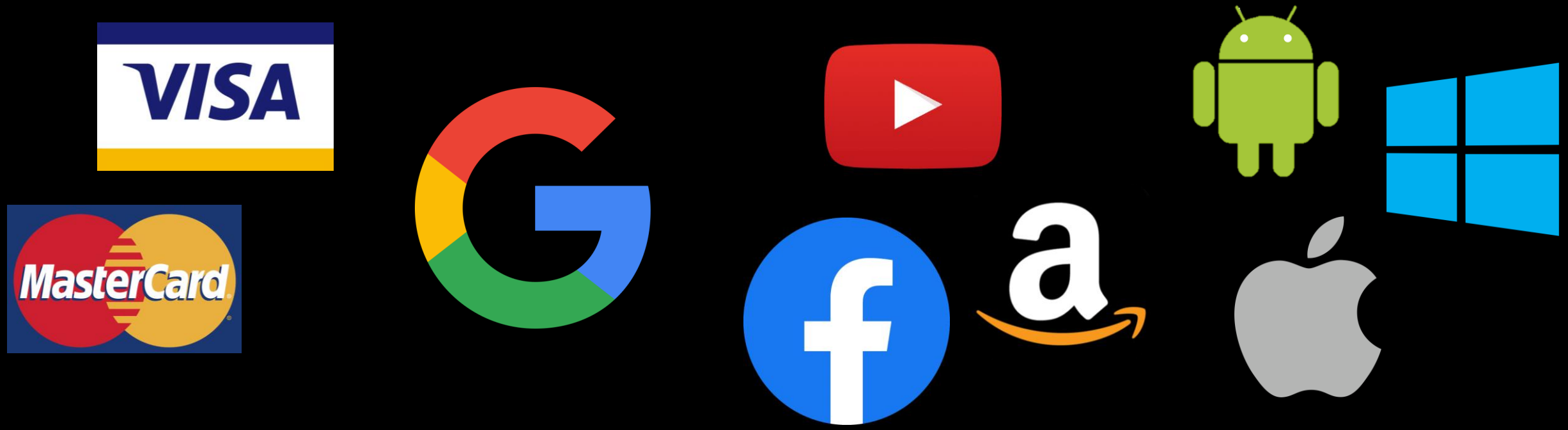
MESCYT

LAFAYETTE
COLLEGE



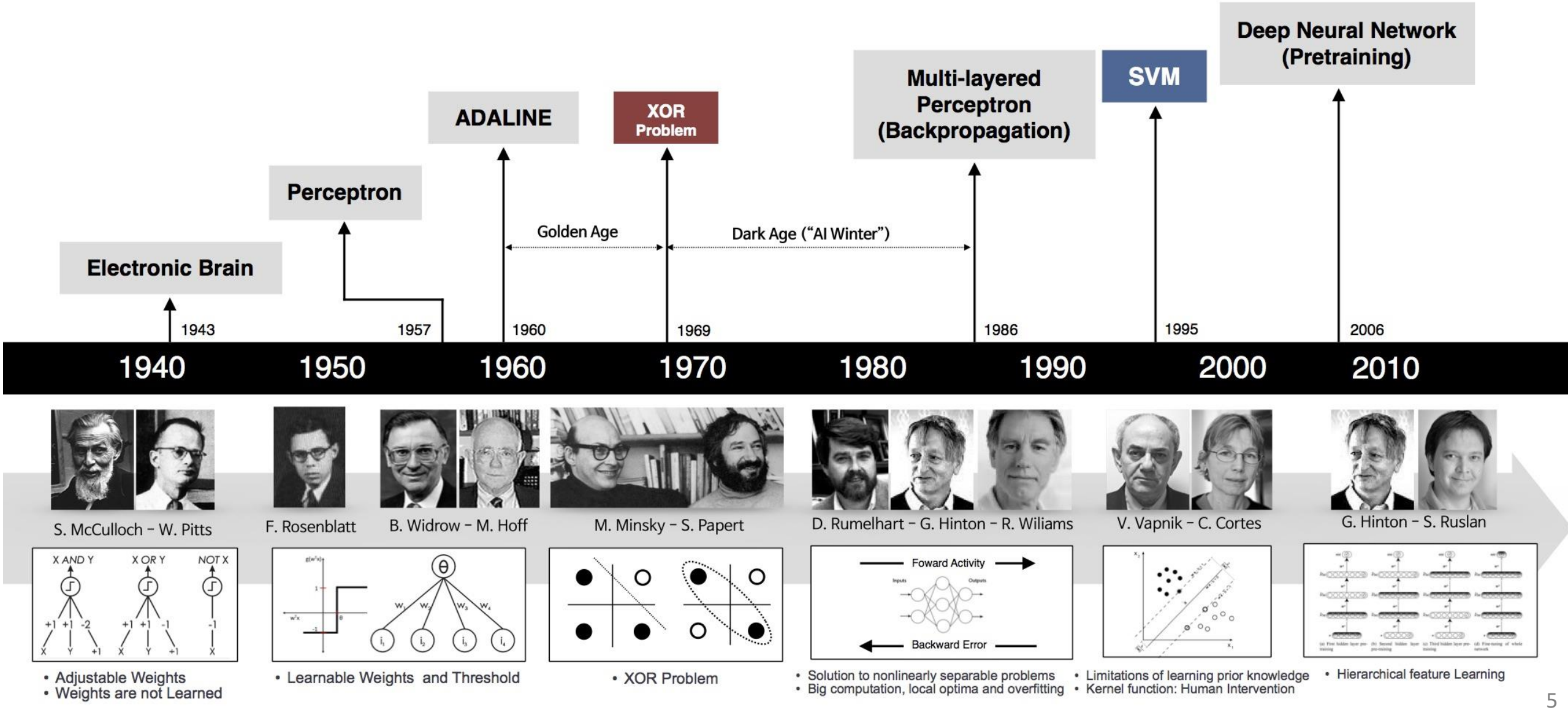
- Visión Artificial
- Procesamiento de Lenguaje Natural
- Investigación de Operaciones
- …

¿En esta semana, has interactuado con algo que usa o implementa algoritmos de Aprendizaje Automático?



*"No necesitas un Doctorado para **implementar** algoritmos de Aprendizaje Automático"*

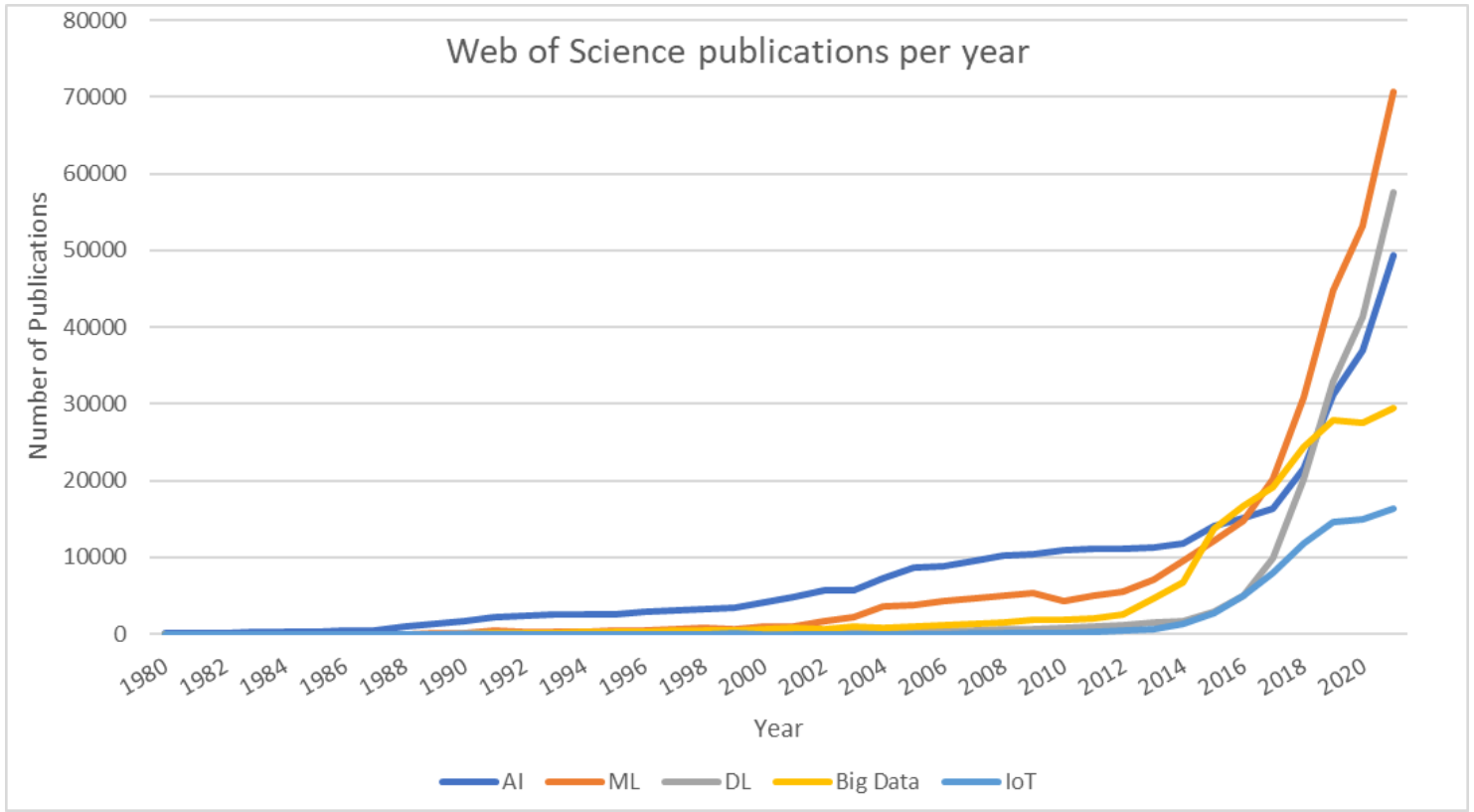
Aprendizaje Automático ha existido durante bastante tiempo, y su interés ha fluctuado a lo largo de las décadas



Pero en la última década hemos visto un crecimiento en su popularidad, la cual continúa creciendo

*"Programa informático que aprende de la experiencia **E** con respecto a una tarea **T** con rendimiento **P**, si su desempeño en la tarea **T**, medido por **P**, mejora con la experiencia **E**."*

Mitchell, T. [1997].



A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devices - are hard to comprehend, particularly when looked at in the context of one day

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	1,000² bytes	1,000,000 bytes
GB	gigabyte	1,000³ bytes	1,000,000,000 bytes
TB	terabyte	1,000⁴ bytes	1,000,000,000,000 bytes
PB	petabyte	1,000⁵ bytes	1,000,000,000,000,000 bytes
EB	exabyte	1,000⁶ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	1,000⁷ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	1,000⁸ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

463EB

of data will be created every day by 2025

500m

tweets are sent every day

Twitter

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

95m

photos and videos are shared on Instagram

Instagram Business

320bn

emails to be sent each day by 2021

294bn

billion emails are sent

Radicati Group

306bn

emails to be sent each day by 2020

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

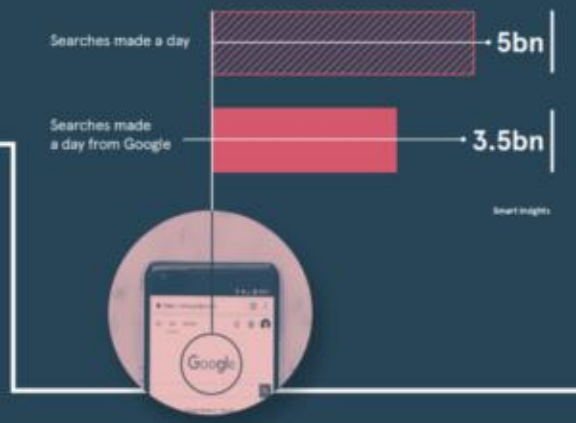
3.9bn

people use emails

4TB

of data produced by a connected car

Intel



Smart Insights

28PB

to be generated from wearable devices by 2020

Statista

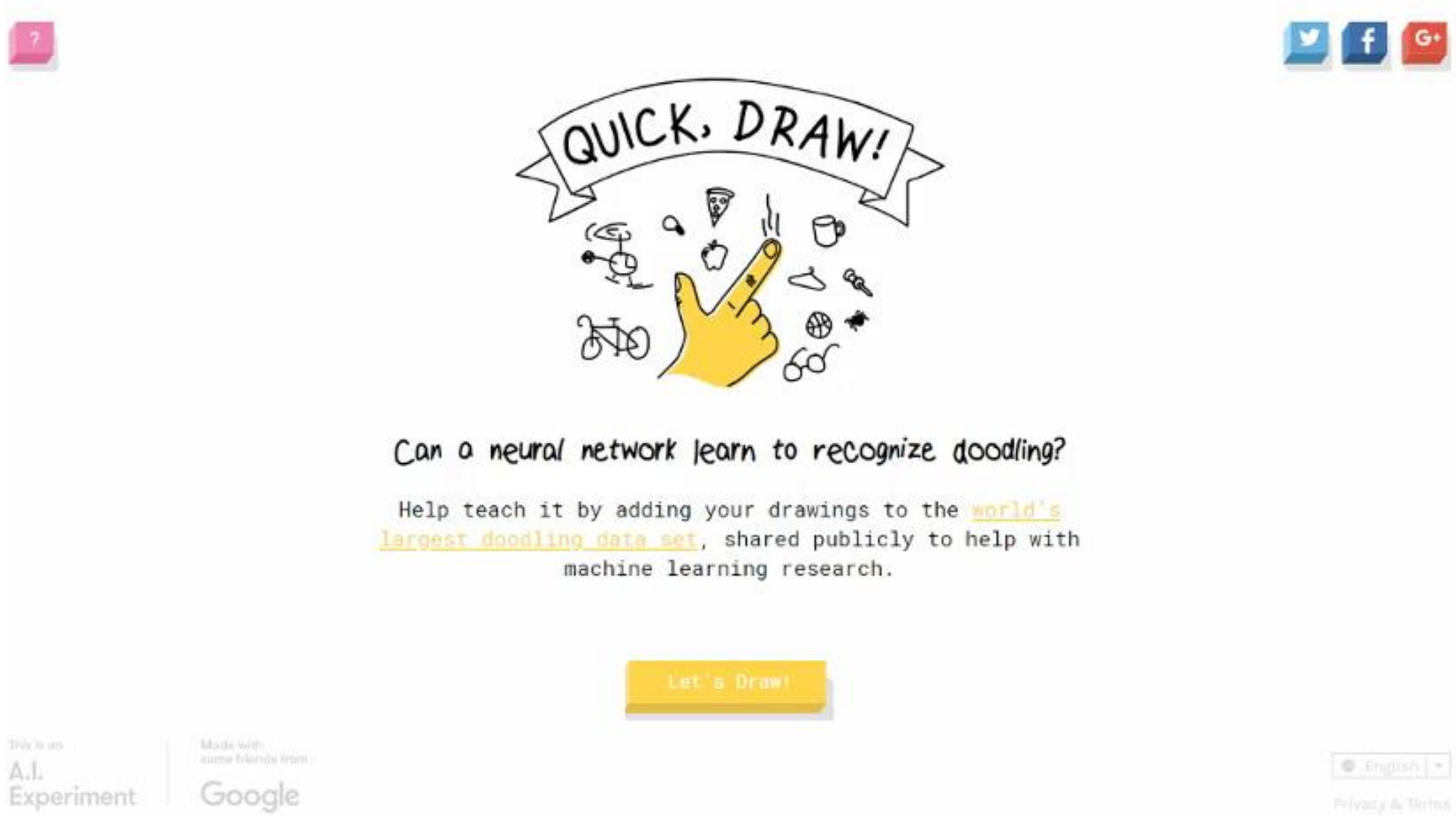
ACCUMULATED DIGITAL UNIVERSE OF DATA

IDC

https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

RACONTEUR

Quick, Draw



https://quickdraw.withgoogle.com/#

Existen muchos tipos de sistemas de Aprendizaje Automático, por lo tanto, es útil clasificarlos en diferentes categorías

- Si están o no entrenado con supervisión humana **(datos utilizados)**
  - Supervisado (*Supervised*)
  - No Supervisado (*Unsupervised*)
  - Semi-supervisado (*Semi-Supervised*)
  - Aprendizaje reforzado (*Reinforcement Learning*)

La clasificación más común de sistemas de Aprendizaje Automático es basado en los "datos de entrenamiento”

Datos Etiquetados

$(x^1, y^1), \ldots$
$, (x^m, y^m)$

Supervisado

No Supervisado

Datos No Etiquetados

$x^1, x^2, \ldots, x^m$

Supervisado es el tipo de algoritmos más común

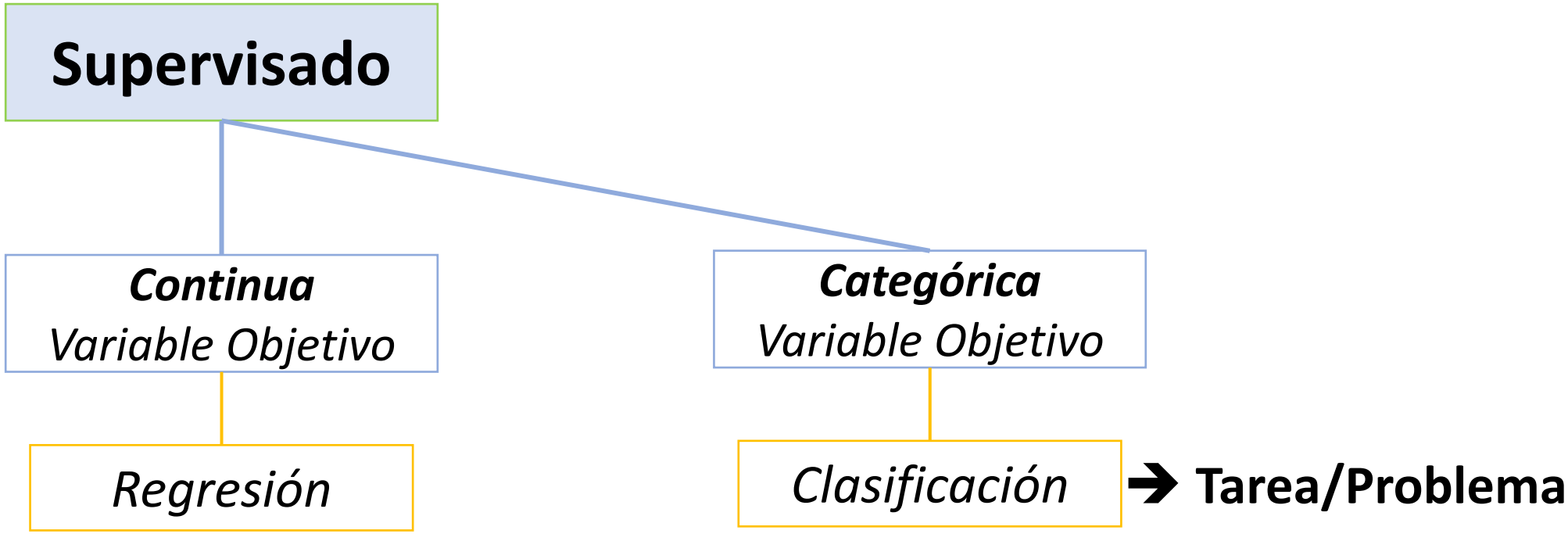
Datos Etiquetados

$(x^1, y^1), \ldots$
$, (x^m, y^m)$

Objetivo :

x➔Y

y=f(x)



- *Predicción del precio de casas*
- *Predicciones mercado de valores*
- *Predicciones de uso de la red eléctrica*

- *Imagenes medicas*
- *Detección de objetos*
- *Predicciones mercado de valores*

En la mayoría de los casos, los datos no etiquetados son más predominantes que los datos etiquetados

Datos Etiquetados

$(x^1, y^1), \ldots, (x^m, y^m)$

Supervisado

No Supervisado

Datos No Etiquetados

$x^1, x^2, \ldots, x^m$

No existe una Variable Objetivo

Agrupación	Asociación	Detección de anomalías	Reducción de dimensionalidad
• *Segmentación de clientes*	• *Análisis de la canasta de mercado*	• *Detecciones de fraudes*	• *Visualización* • *Ingeniería de Atributos*

¿Qué tipo de sistemas de ML puedo usar para responder las siguientes preguntas y qué tipo de tarea es?

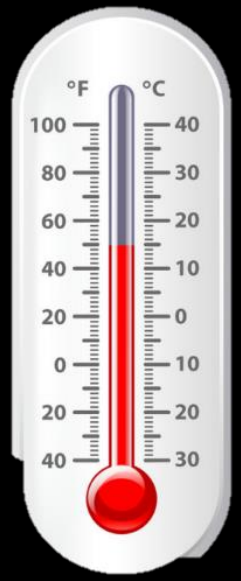
Supervisado

Clasificación

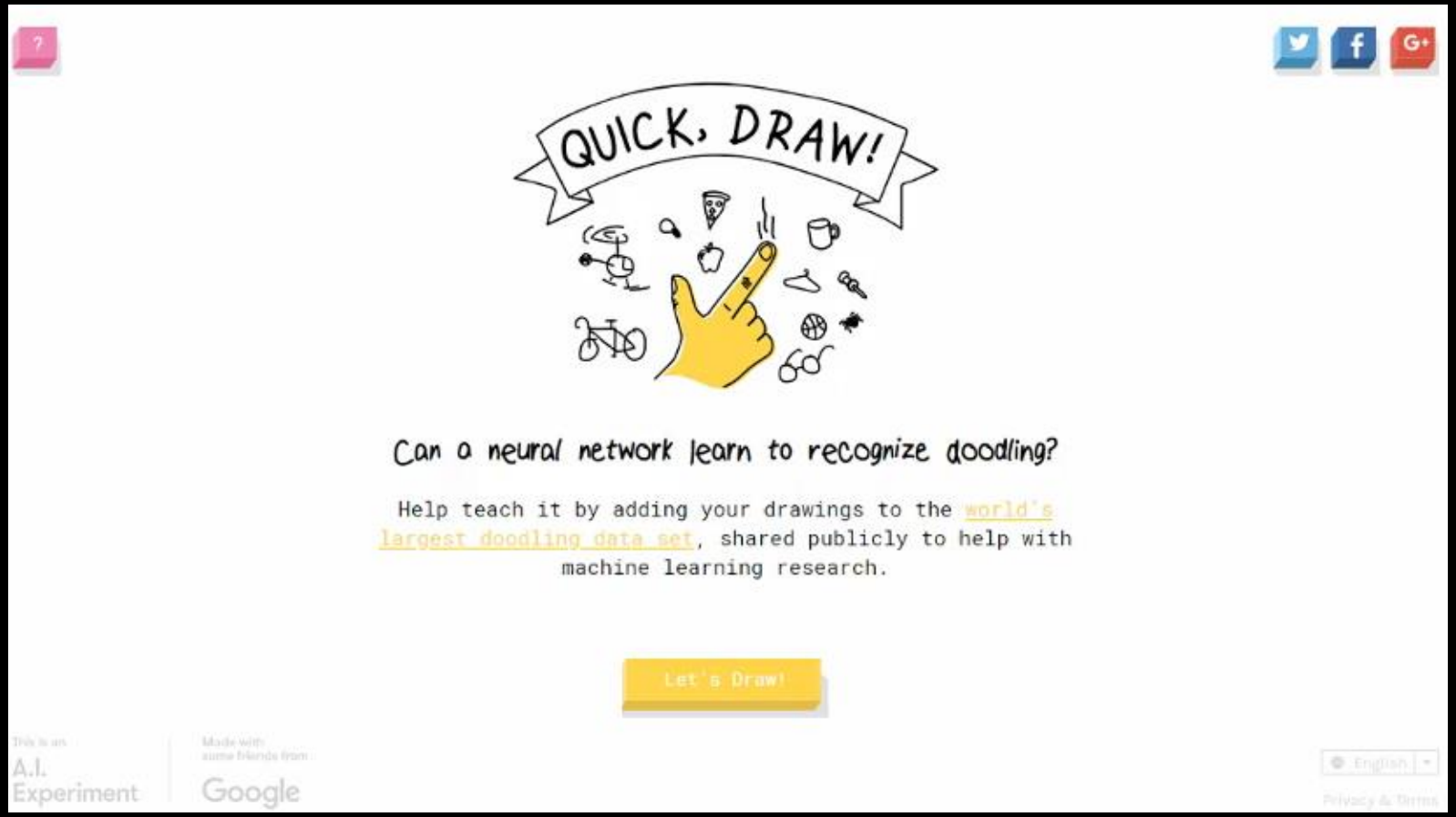
1. ¿Va a estar soleado o no el día de mañana?

Regresión

2. ¿Cuál sería la temperatura máxima el día de mañana?



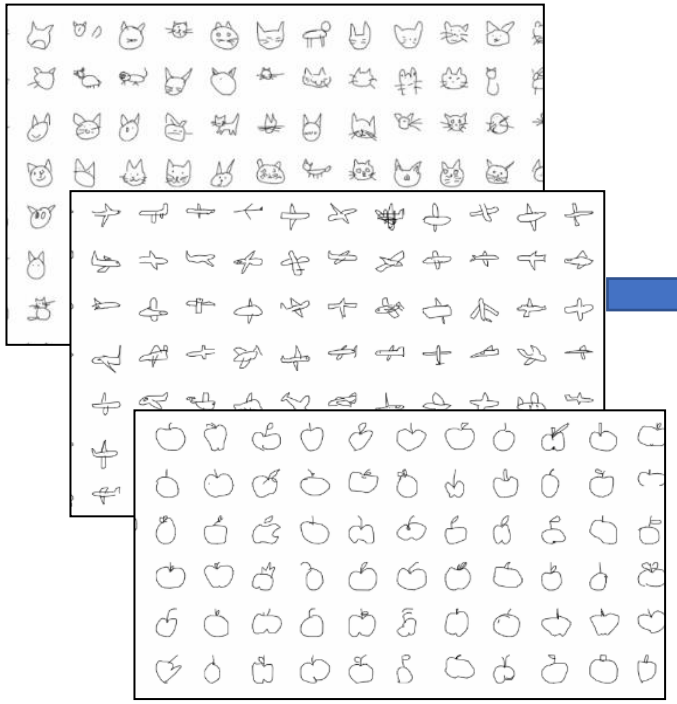
¿Qué tipo de sistemas de ML utilizaron, qué tipo de tarea es, y qué tipo de datos utilizaron?



Supervisado

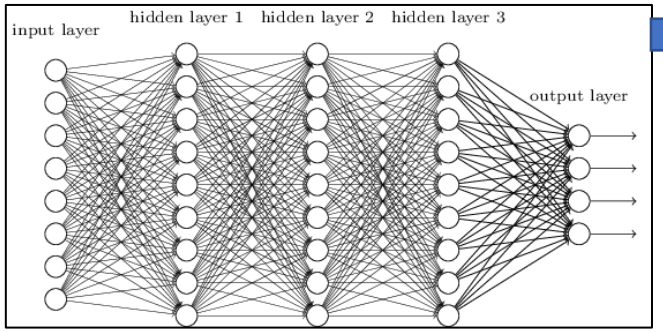
Clasificación

“Humanos provén el conocimiento”



345 clases/objetos

Algoritmos identifican patrones en la data

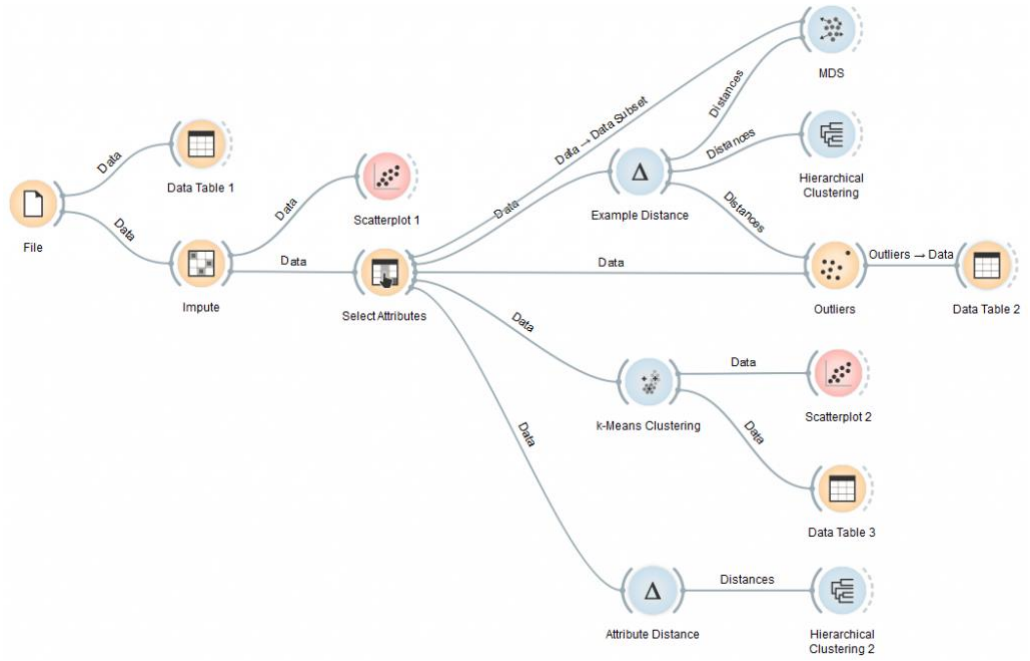


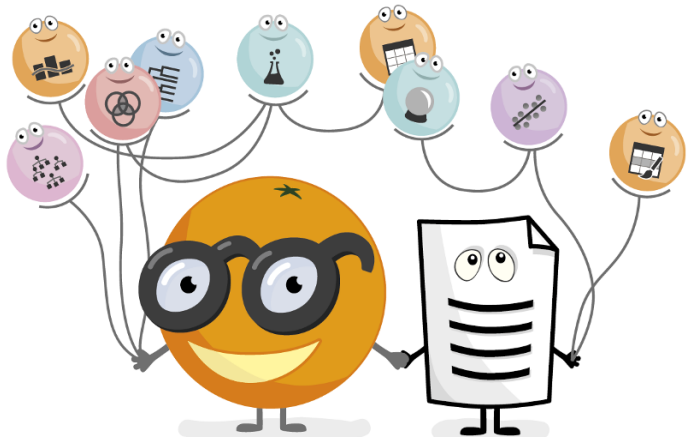
Despliegue del producto

Implementando algoritmos de Aprendizaje Automático sin la necesidad de programar*

orange

DATA MINING
FRUITFUL&FUN



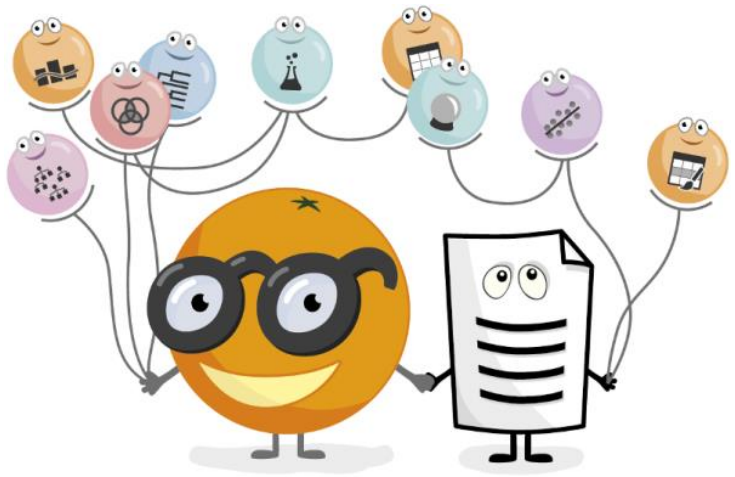


Lenguaje de Programación visual

1er paso: Instalar programa Orange

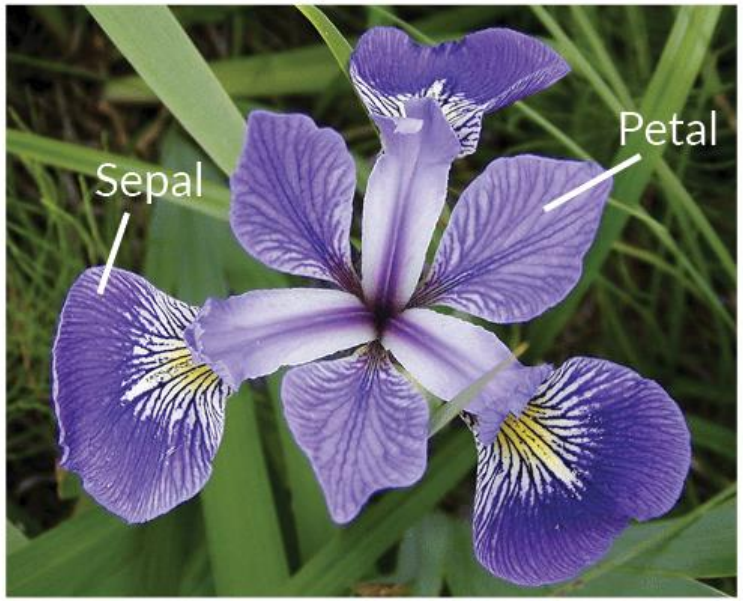
"Orange Data mining"

https://orangedatamining.com/download/



GitHub

Hoy vamos a trabajar con la base de datos *Iris* de Ronald Fisher



Iris Versicolor

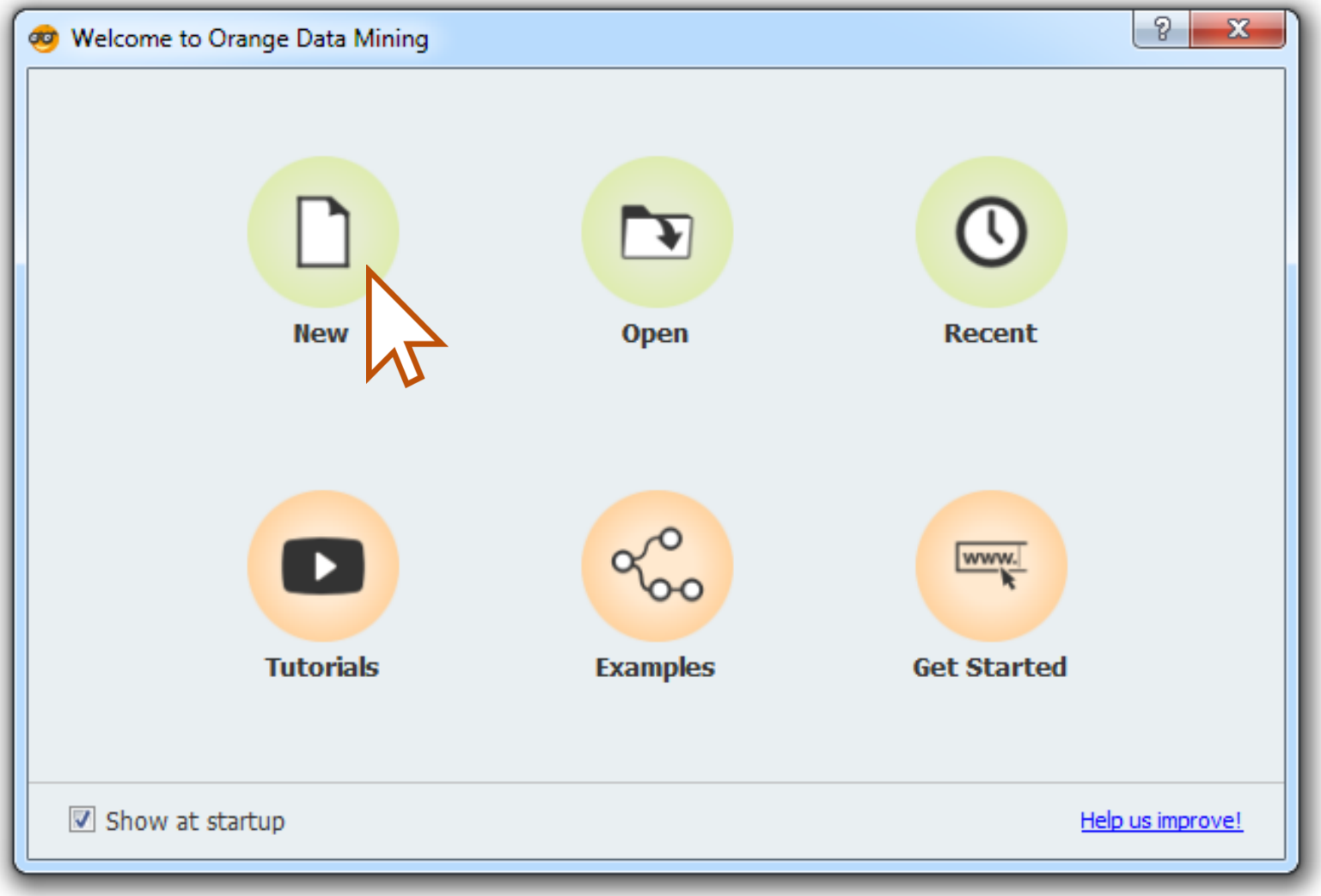
Iris Setosa

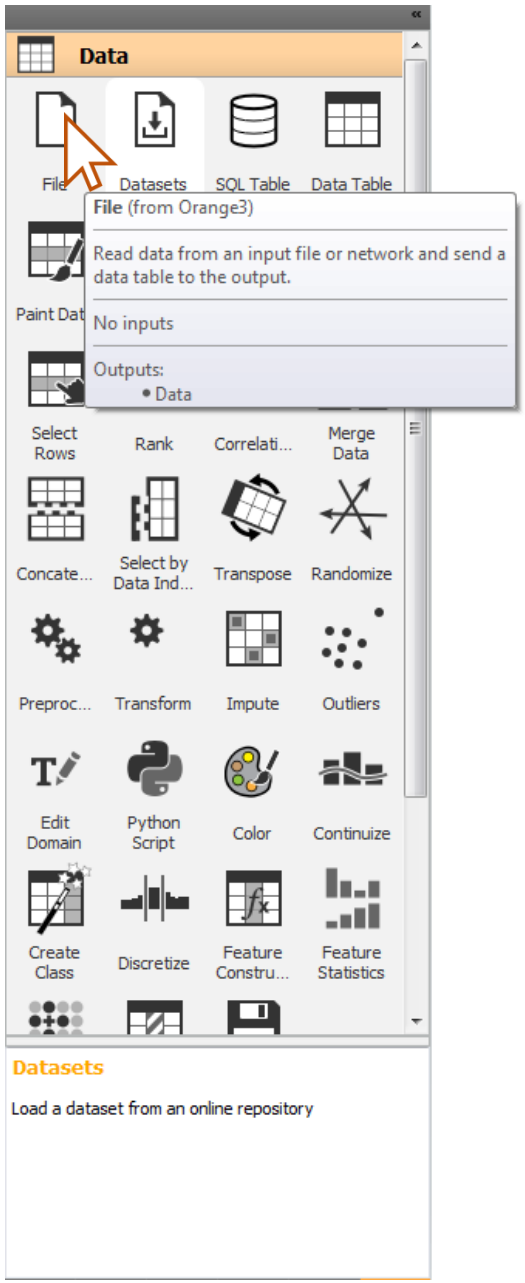
Iris Virginica

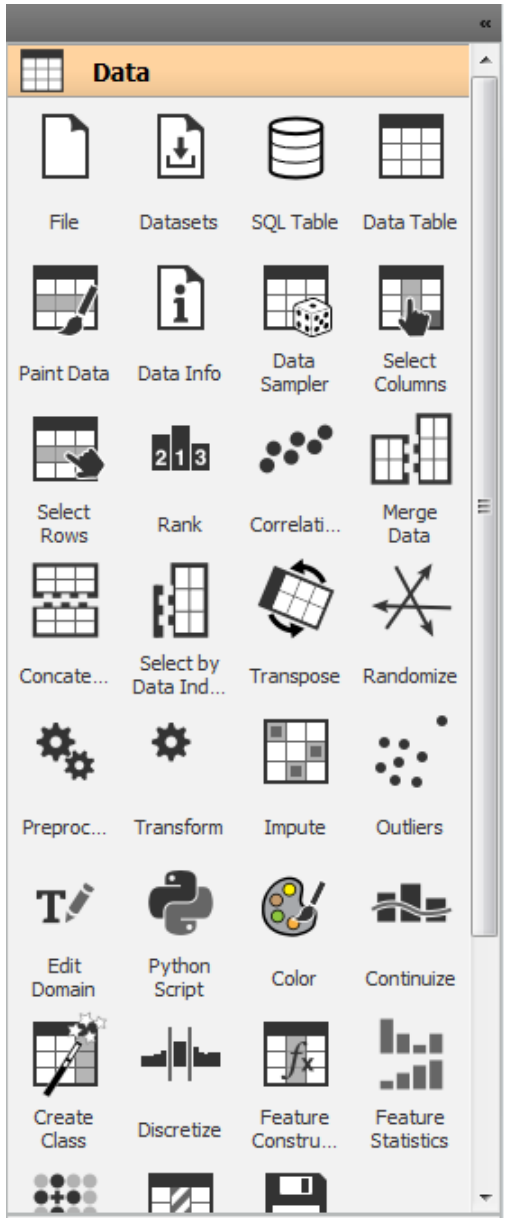
	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
…	…	…	…	…	…
51	7	3.2	4.7	0.2	versicolor
…	…	…	…	…	…
150	5.9	3	5.1	1.8	virginica

150 Flores medidas
(tuplas=filas)
X
4 atributos
(*variables=columnas*)

2do paso: Cargar los datos

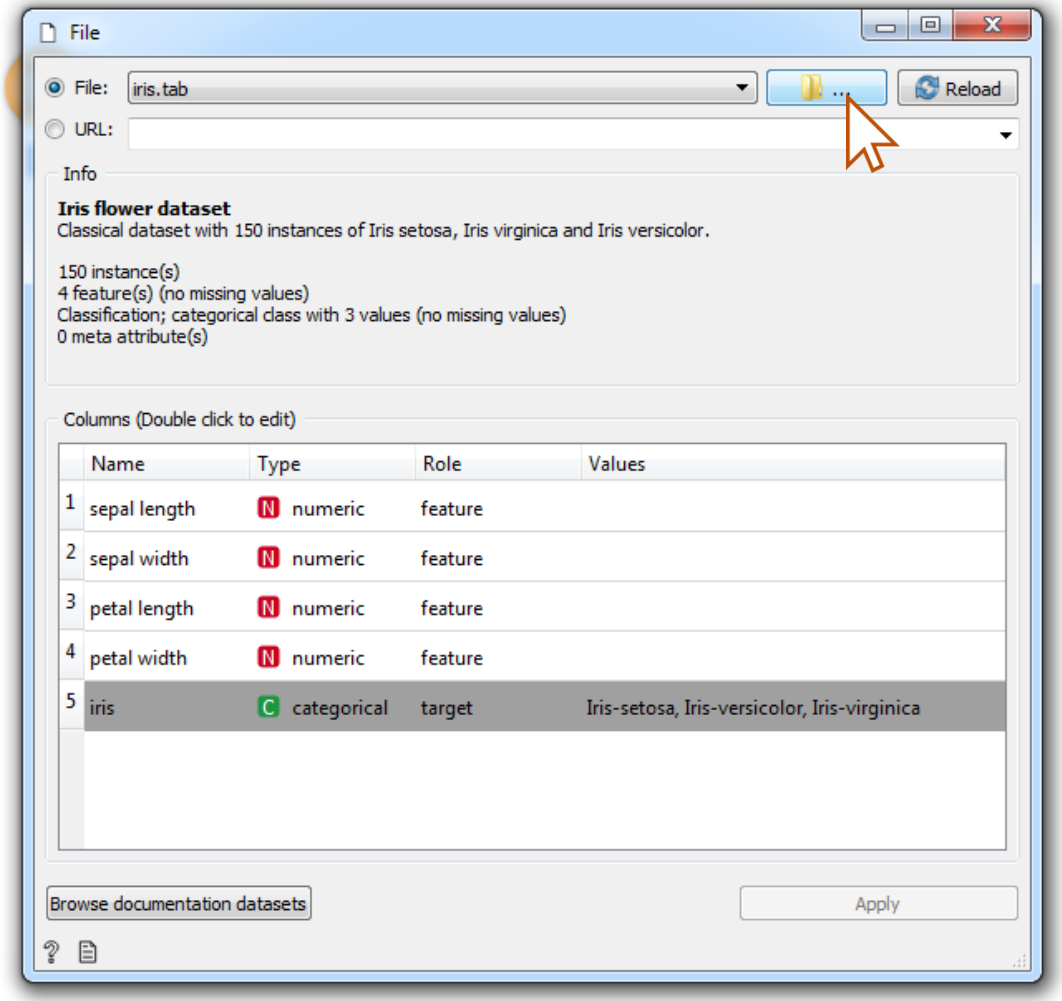


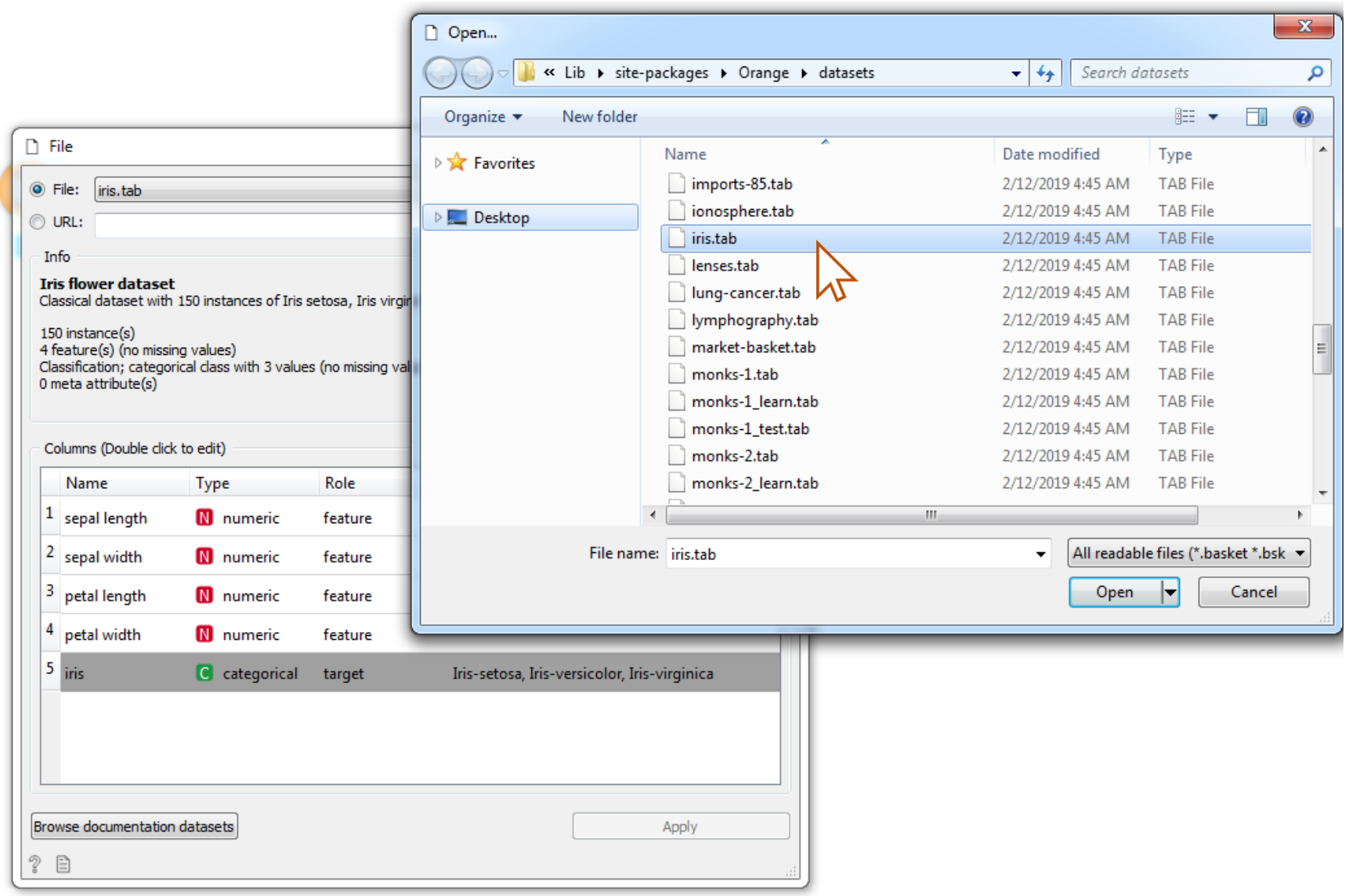

Data
File
Datasets
SQL Table
Data Table
File (from Orange3)
Read data from an input file or network and send a data table to the output.
No inputs
Outputs:
• Data
Paint Dat
Select Rows
Rank
Correlati...
Merge Data
Concate...
Select by Data Ind...
Transpose
Randomize
Preproc...
Transform
Impute
Outliers
Edit Domain
Python Script
Color
Continuize
Create Class
Discretize
Feature Constru...
Feature Statistics
Datasets
Load a dataset from an online repository

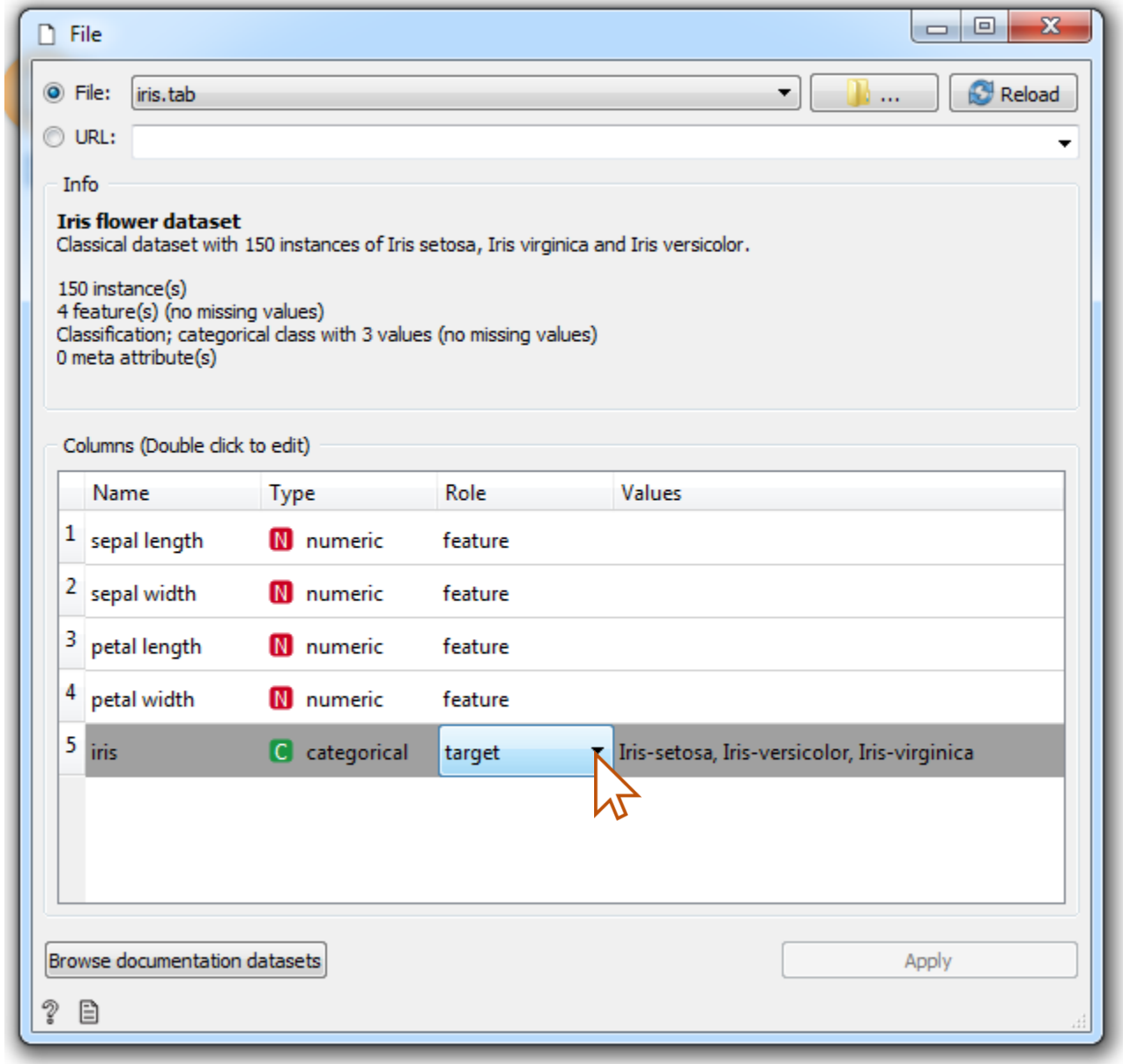

Data
File
Datasets
SQL Table
Data Table
Paint Data
Data Info
Data Sampler
Select Columns
Select Rows
Rank
Correlati...
Merge Data
Concate...
Select by Data Ind...
Transpose
Randomize
Preproc...
Transform
Impute
Outliers
Edit Domain
Python Script
Color
Continuize
Create Class
Discretize
Feature Constru...
Feature Statistics


File

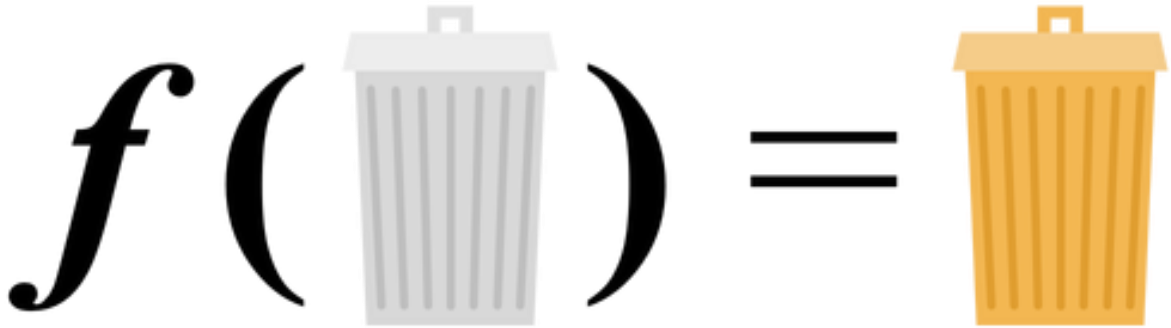

Data
File
Datasets
SQL Table
Data Table
Paint Data
Data Info
Data Sampler
Select Columns
Select Rows
Rank
Correlati...
Merge Data
Concate...
Select by Data Ind...
Transpose
Randomize
Preproc...
Transform
Impute
Outliers
Edit Domain
Python Script
Color
Continuize
Create Class
Discretize
Feature Constru...
Feature Statistics
File
Read data from an input file or network and send a data table to the output.
more...


File
File: iris.tab
Reload
URL:
Info
Iris flower dataset
Classical dataset with 150 instances of Iris setosa, Iris virginica and Iris versicolor.
150 instance(s)
4 feature(s) (no missing values)
Classification; categorical class with 3 values (no missing values)
0 meta attribute(s)
Columns (Double click to edit)
Name Type Role Values
1 sepal length numeric feature
2 sepal width numeric feature
3 petal length numeric feature
4 petal width numeric feature
5 iris categorical target Iris-setosa, Iris-versicolor, Iris-virginica
Browse documentation datasets
Apply


File
File: iris.tab
URL:
Info
Iris flower dataset
Classical dataset with 150 instances of Iris setosa, Iris virgi
150 instance(s)
4 feature(s) (no missing values)
Classification; categorical class with 3 values (no missing val
0 meta attribute(s)
Columns (Double click to edit)
Name Type Role
1 sepal length numeric feature
2 sepal width numeric feature
3 petal length numeric feature
4 petal width numeric feature
5 iris categorical target Iris-setosa, Iris-versicolor, Iris-virginica
Browse documentation datasets
Apply
Open...
Lib ▸ site-packages ▸ Orange ▸ datasets
Search datasets
Organize
New folder
Favorites
Desktop
Name
Date modified
Type
imports-85.tab 2/12/2019 4:45 AM TAB File
ionosphere.tab 2/12/2019 4:45 AM TAB File
iris.tab 2/12/2019 4:45 AM TAB File
lenses.tab 2/12/2019 4:45 AM TAB File
lung-cancer.tab 2/12/2019 4:45 AM TAB File
lymphography.tab 2/12/2019 4:45 AM TAB File
market-basket.tab 2/12/2019 4:45 AM TAB File
monks-1.tab 2/12/2019 4:45 AM TAB File
monks-1_learn.tab 2/12/2019 4:45 AM TAB File
monks-1_test.tab 2/12/2019 4:45 AM TAB File
monks-2.tab 2/12/2019 4:45 AM TAB File
monks-2_learn.tab 2/12/2019 4:45 AM TAB File
File name: iris.tab
All readable files (*.basket *.bsk
Open
Cancel


File
File:
iris.tab
...
Reload
URL:
Info
Iris flower dataset
Classical dataset with 150 instances of Iris setosa, Iris virginica and Iris versicolor.
150 instance(s)
4 feature(s) (no missing values)
Classification; categorical class with 3 values (no missing values)
0 meta attribute(s)
Columns (Double click to edit)
Name
Type
Role
Values
1 sepal length numeric feature
2 sepal width numeric feature
3 petal length numeric feature
4 petal width numeric feature
5 iris categorical target Iris-setosa, Iris-versicolor, Iris-virginica
Browse documentation datasets
Apply

Antes de iniciar cualquier análisis hay que entender bien la data

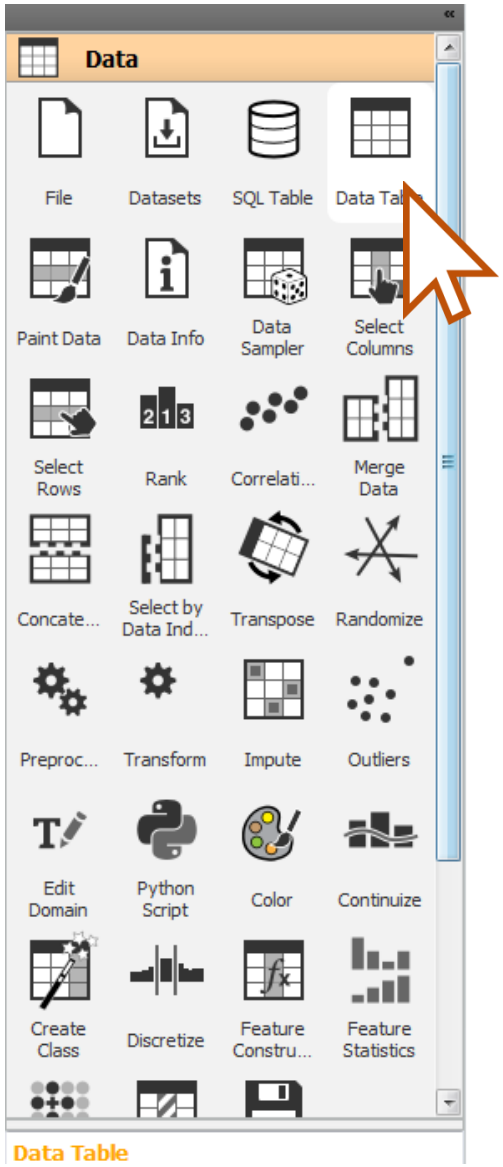


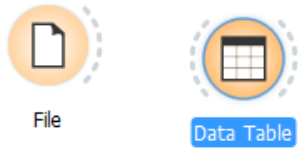
"Garbage in, Garbage out"

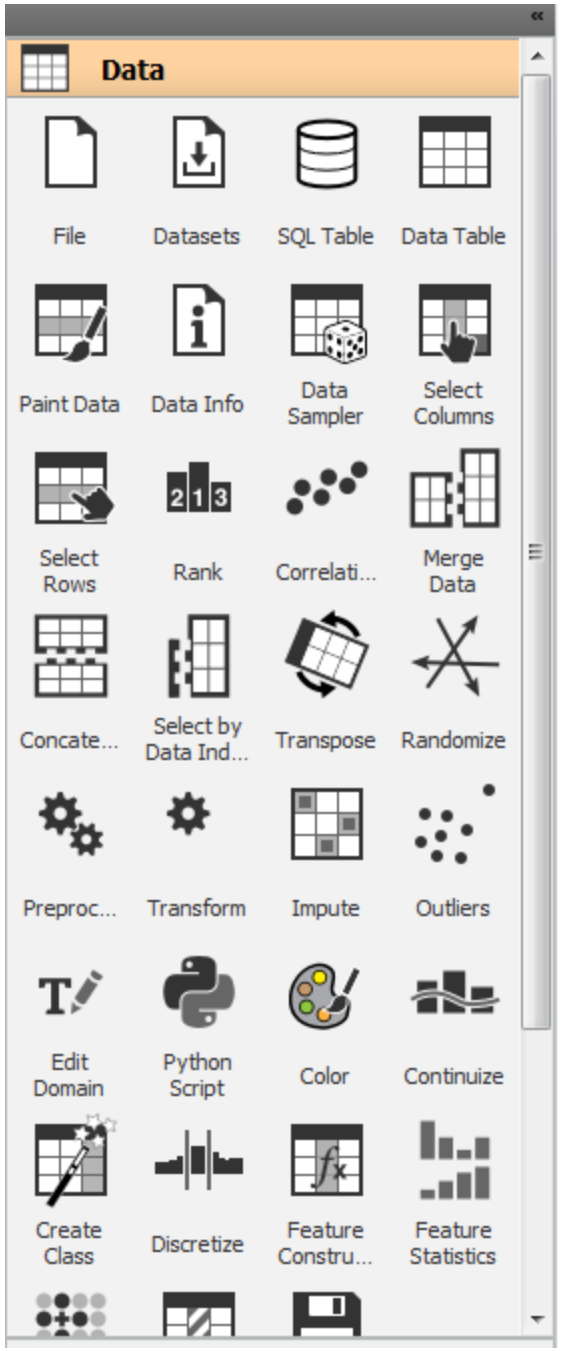
- **Cuántas variables y tuplas hay?**
  (filas y columnas)
- **Qué tipo de variables tengo?**
  (nominal, ordinal, intervalo, proporción, categóricas)
- **Qué información me falta ?**
  (NAs, celdas vacias)
- **Cuál es la distribución de mis variables?**
  (promedio, rango, moda,….)

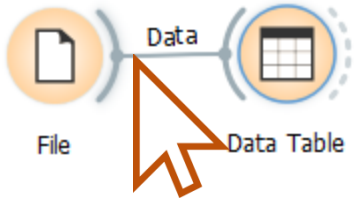
…

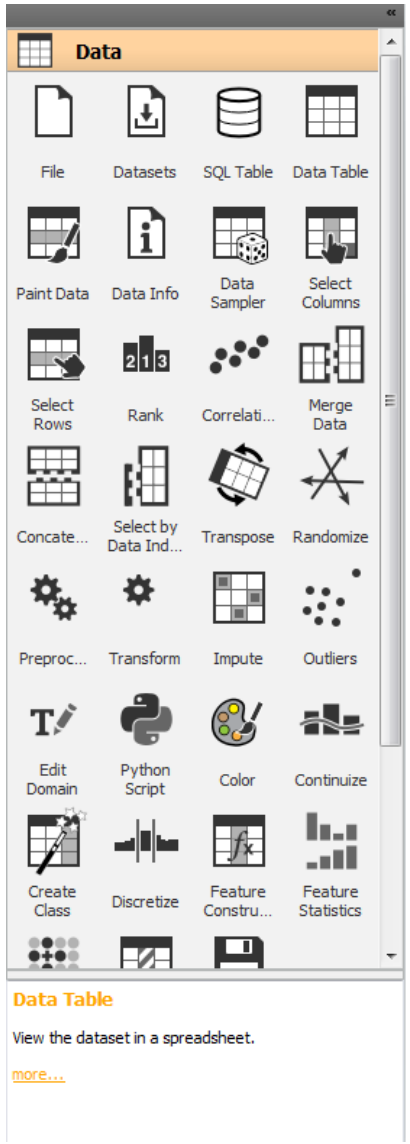
3er paso: Explorar los datos

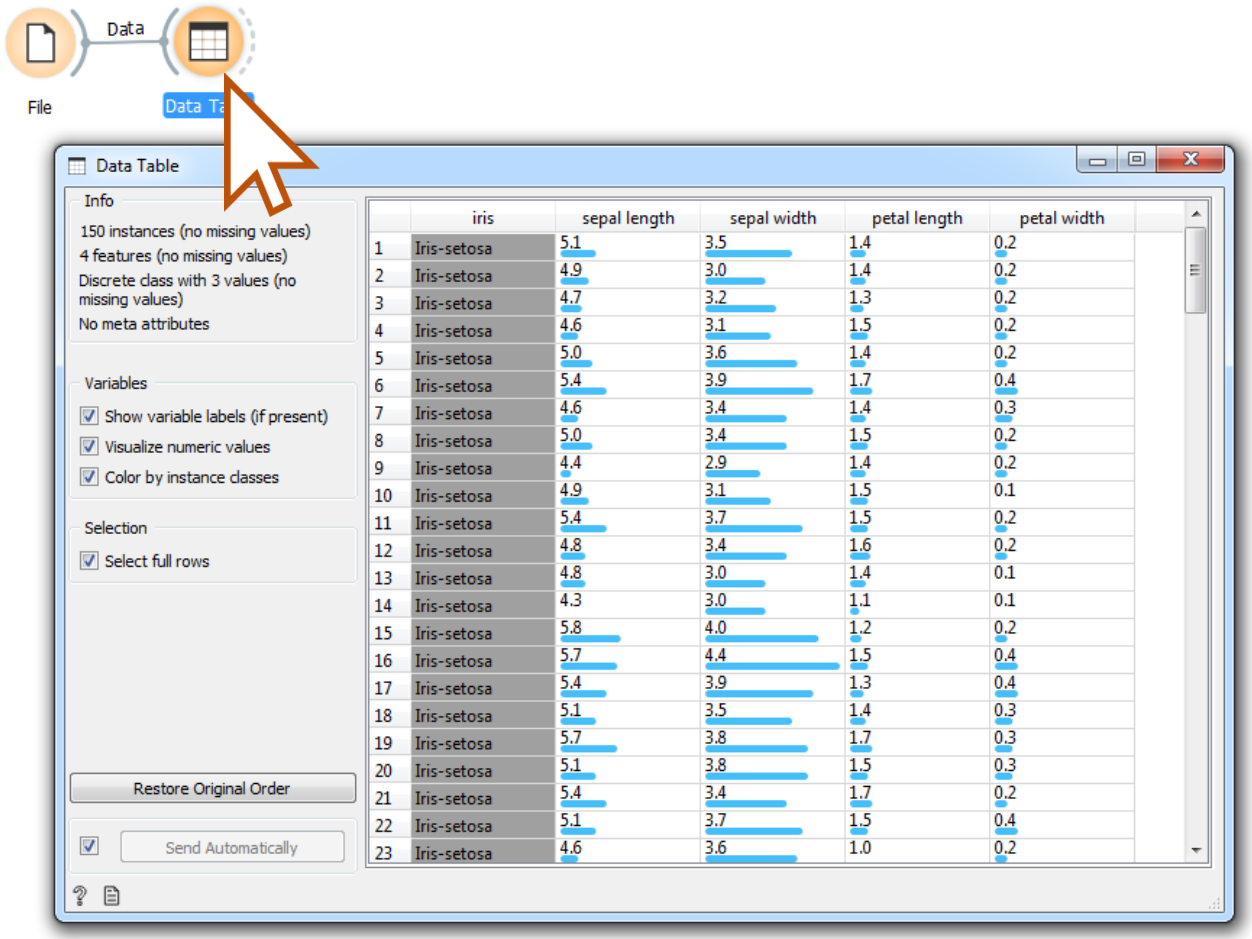


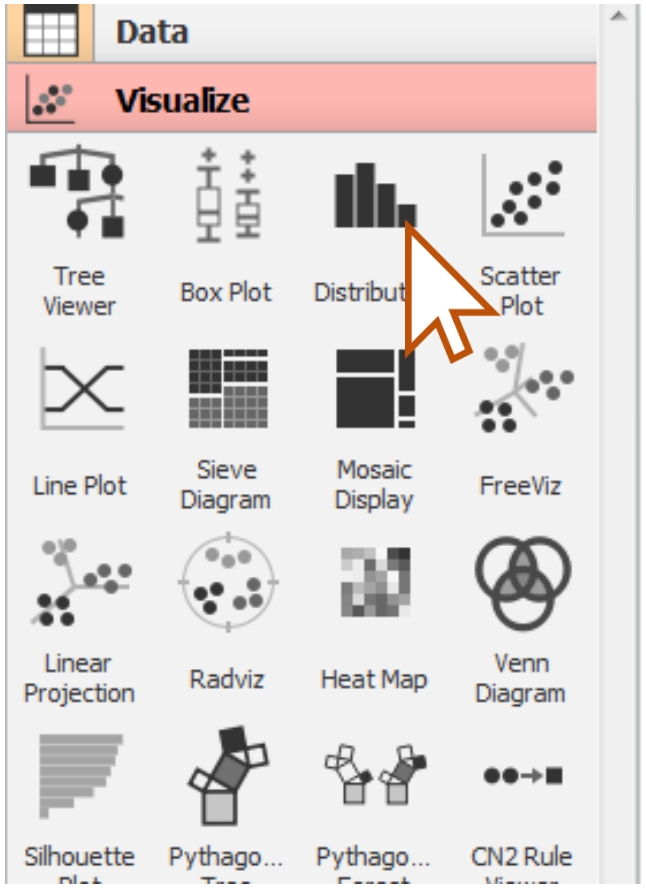


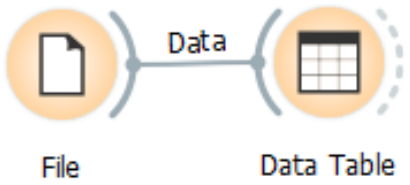

Data
File
Datasets
SQL Table
Data Table
Paint Data
Data Info
Data Sampler
Select Columns
Select Rows
Rank
Correlati...
Merge Data
Concate...
Select by Data Ind...
Transpose
Randomize
Preproc...
Transform
Impute
Outliers
Edit Domain
Python Script
Color
Continuize
Create Class
Discretize
Feature Constru...
Feature Statistics

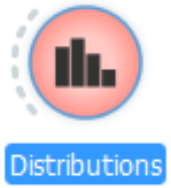

Data
File
Data Table

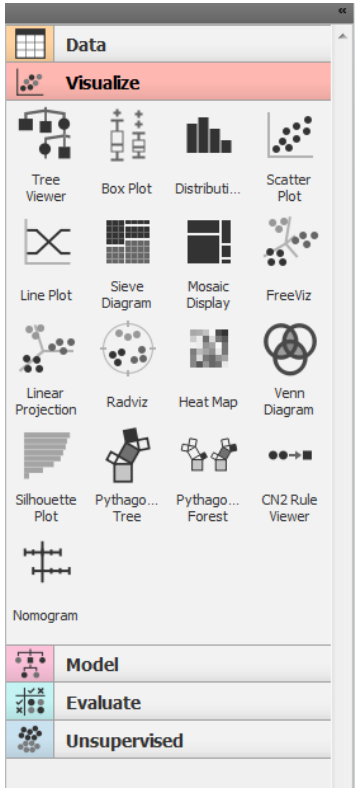

Data
File
Datasets
SQL Table
Data Table
Paint Data
Data Info
Data Sampler
Select Columns
Select Rows
Rank
Correlati...
Merge Data
Concate...
Select by Data Ind...
Transpose
Randomize
Preproc...
Transform
Impute
Outliers
Edit Domain
Python Script
Color
Continuize
Create Class
Discretize
Feature Constru...
Feature Statistics
Data Table
View the dataset in a spreadsheet.
more...

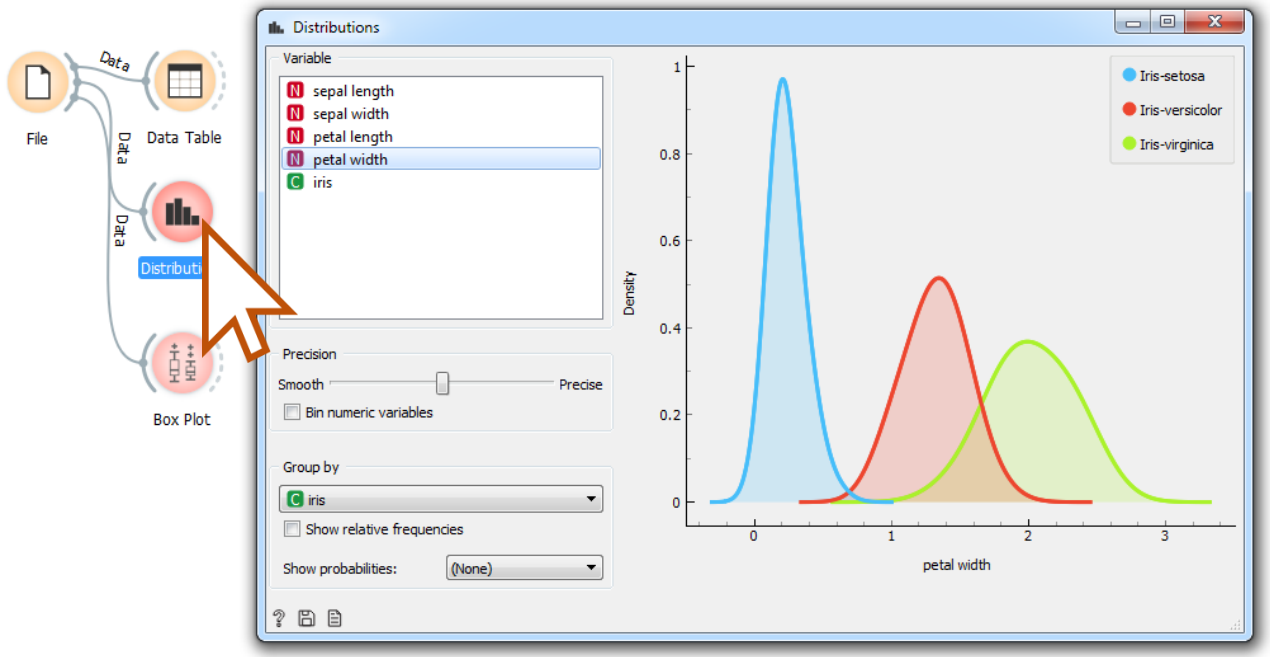

Data
File
Data Table
Info
150 instances (no missing values)
4 features (no missing values)
Discrete class with 3 values (no missing values)
No meta attributes
Variables
Show variable labels (if present)
Visualize numeric values
Color by instance classes
Selection
Select full rows
Restore Original Order
Send Automatically
iris
sepal length
sepal width
petal length
petal width
1 Iris-setosa 5.1 3.5 1.4 0.2
2 Iris-setosa 4.9 3.0 1.4 0.2
3 Iris-setosa 4.7 3.2 1.3 0.2
4 Iris-setosa 4.6 3.1 1.5 0.2
5 Iris-setosa 5.0 3.6 1.4 0.2
6 Iris-setosa 5.4 3.9 1.7 0.4
7 Iris-setosa 4.6 3.4 1.4 0.3
8 Iris-setosa 5.0 3.4 1.5 0.2
9 Iris-setosa 4.4 2.9 1.4 0.2
10 Iris-setosa 4.9 3.1 1.5 0.1
11 Iris-setosa 5.4 3.7 1.5 0.2
12 Iris-setosa 4.8 3.4 1.6 0.2
13 Iris-setosa 4.8 3.0 1.4 0.1
14 Iris-setosa 4.3 3.0 1.1 0.1
15 Iris-setosa 5.8 4.0 1.2 0.2
16 Iris-setosa 5.7 4.4 1.5 0.4
17 Iris-setosa 5.4 3.9 1.3 0.4
18 Iris-setosa 5.1 3.5 1.4 0.3
19 Iris-setosa 5.7 3.8 1.7 0.3
20 Iris-setosa 5.1 3.8 1.5 0.3
21 Iris-setosa 5.4 3.4 1.7 0.2
22 Iris-setosa 5.1 3.7 1.5 0.4
23 Iris-setosa 4.6 3.6 1.0 0.2

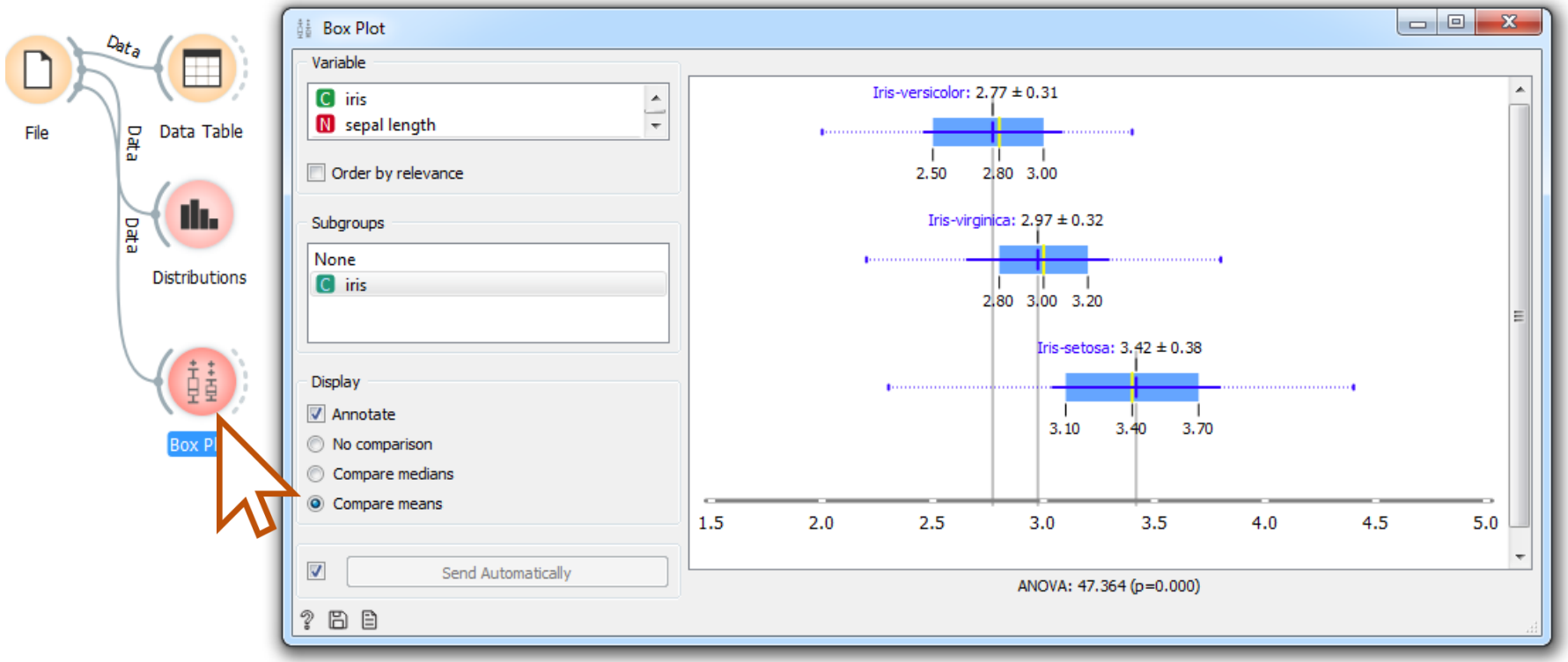

Data
Visualize
Tree Viewer
Box Plot
Distribut
Scatter Plot
Line Plot
Sieve Diagram
Mosaic Display
FreeViz
Linear Projection
Radviz
Heat Map
Venn Diagram
Silhouette
Pythago...
Pythago...
CN2 Rule


Data
File
Data Table

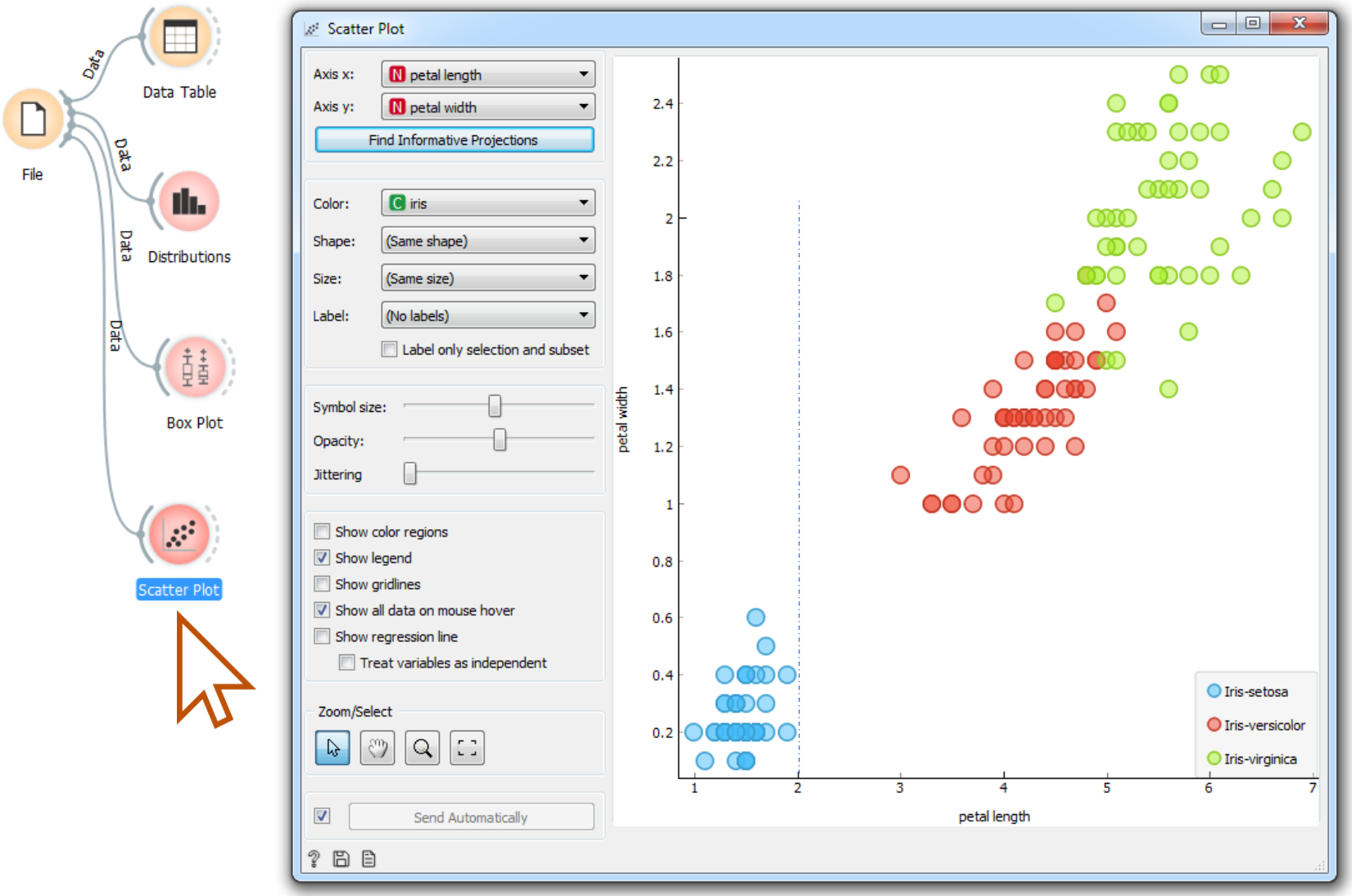

Distributions


Data
Visualize
Tree Viewer
Box Plot
Distributi...
Scatter Plot
Line Plot
Sieve Diagram
Mosaic Display
FreeViz
Linear Projection
Radviz
Heat Map
Venn Diagram
Silhouette Plot
Pythago... Tree
Pythago... Forest
CN2 Rule Viewer
Nomogram
Model
Evaluate
Unsupervised

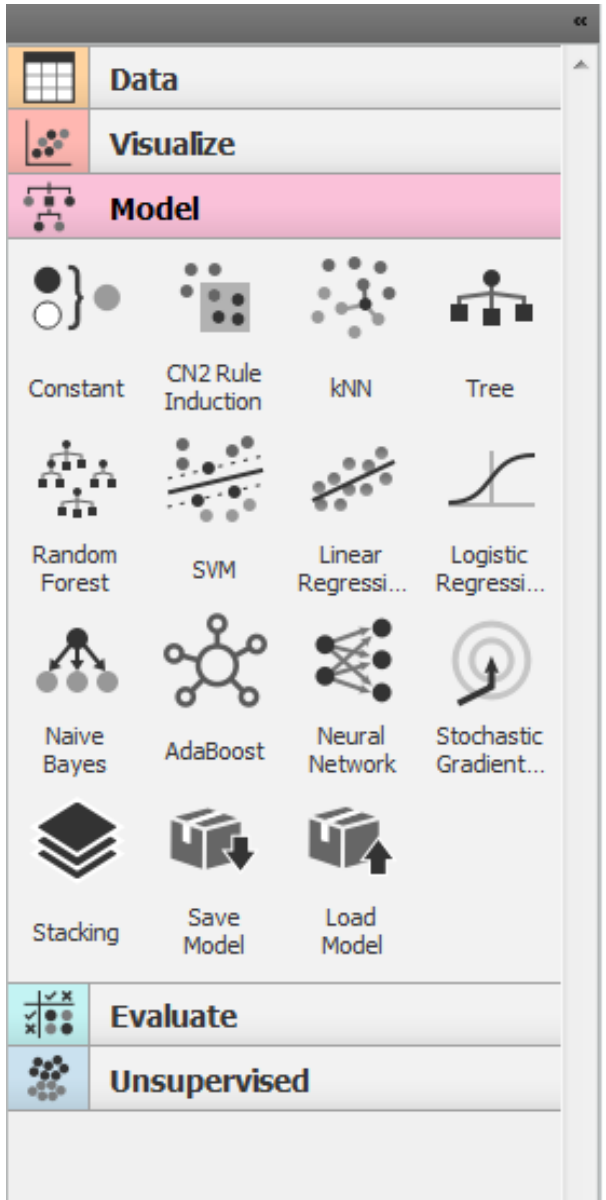

File
Data
Data Table
Data
Box Plot
Distributions
Variable
sepal length
sepal width
petal length
petal width
iris
Precision
Smooth
Precise
Bin numeric variables
Group by
iris
Show relative frequencies
Show probabilities:
(None)
Density
petal width
Iris-setosa
Iris-versicolor
Iris-virginica

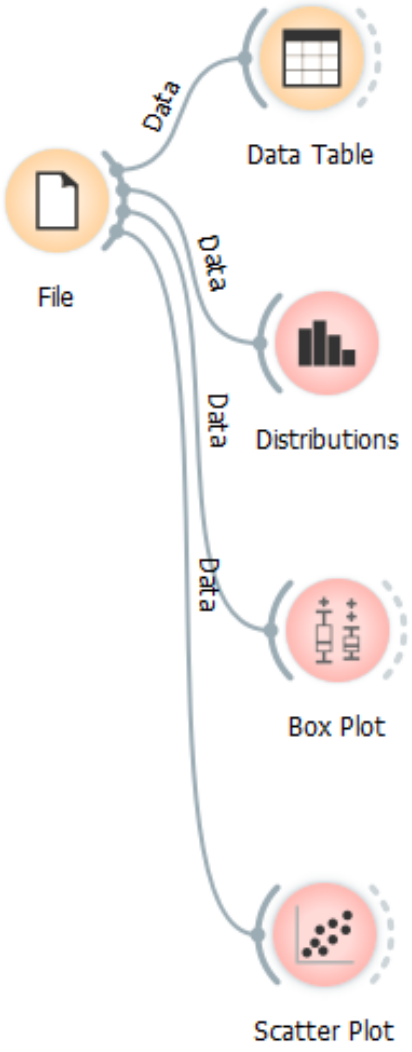

File
Data
Data Table
Data
Distributions
Box Pl
Box Plot
Variable
iris
sepal length
Order by relevance
Subgroups
None
iris
Display
Annotate
No comparison
Compare medians
Compare means
Send Automatically
Iris-versicolor: 2.77 ± 0.31
2.50
2.80
3.00
Iris-virginica: 2.97 ± 0.32
2.80
3.00
3.20
Iris-setosa: 3.42 ± 0.38
3.10
3.40
3.70
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
ANOVA: 47.364 (p=0.000)

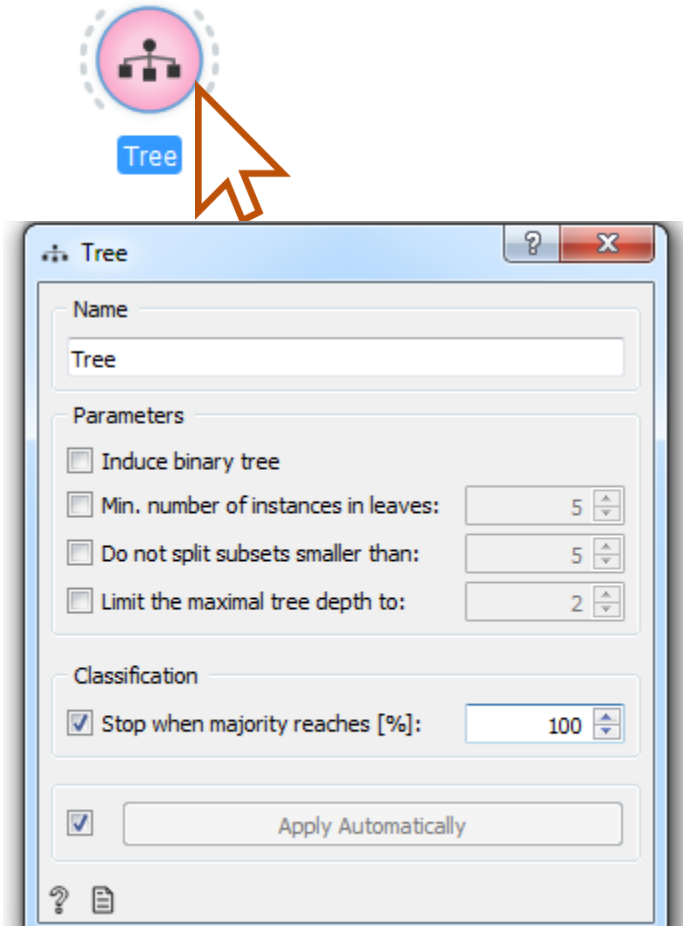
Qué tipo de reglas podemos crear para diferenciar entre las clases de Iris ?



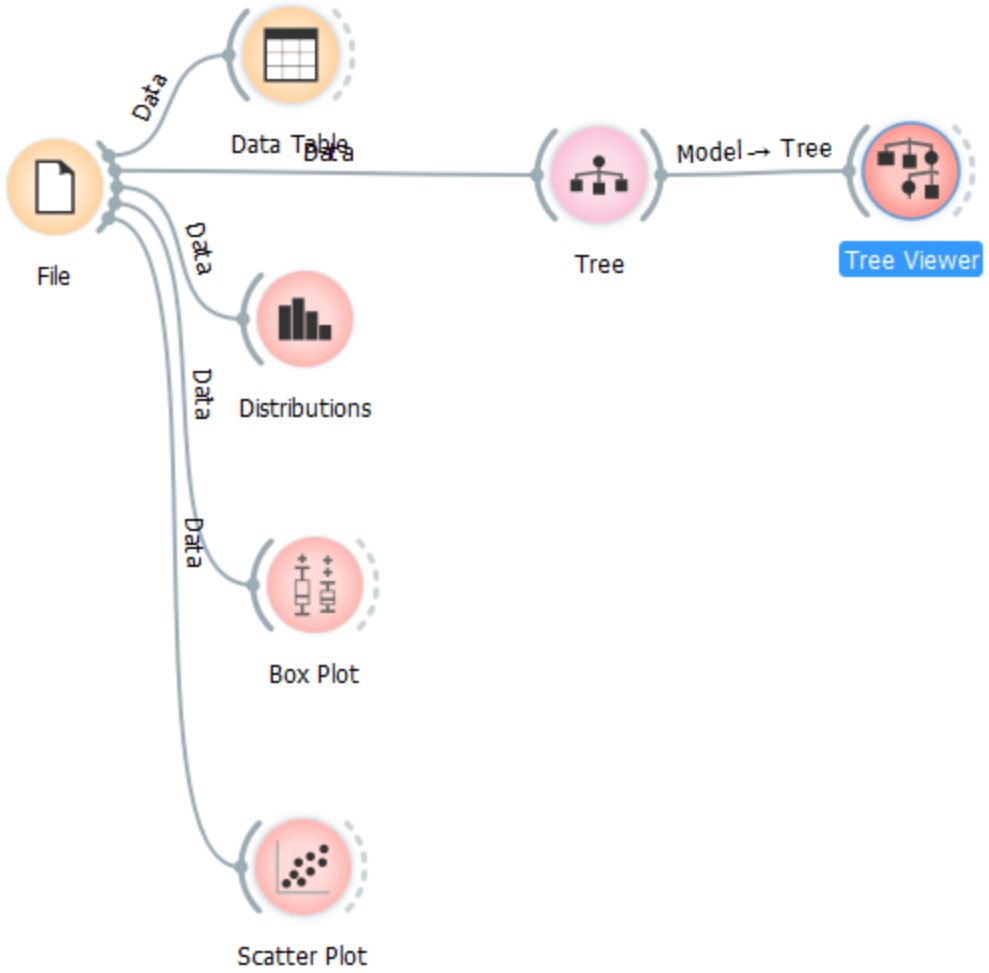
Puede un algoritmo generar estas reglas automáticamente?

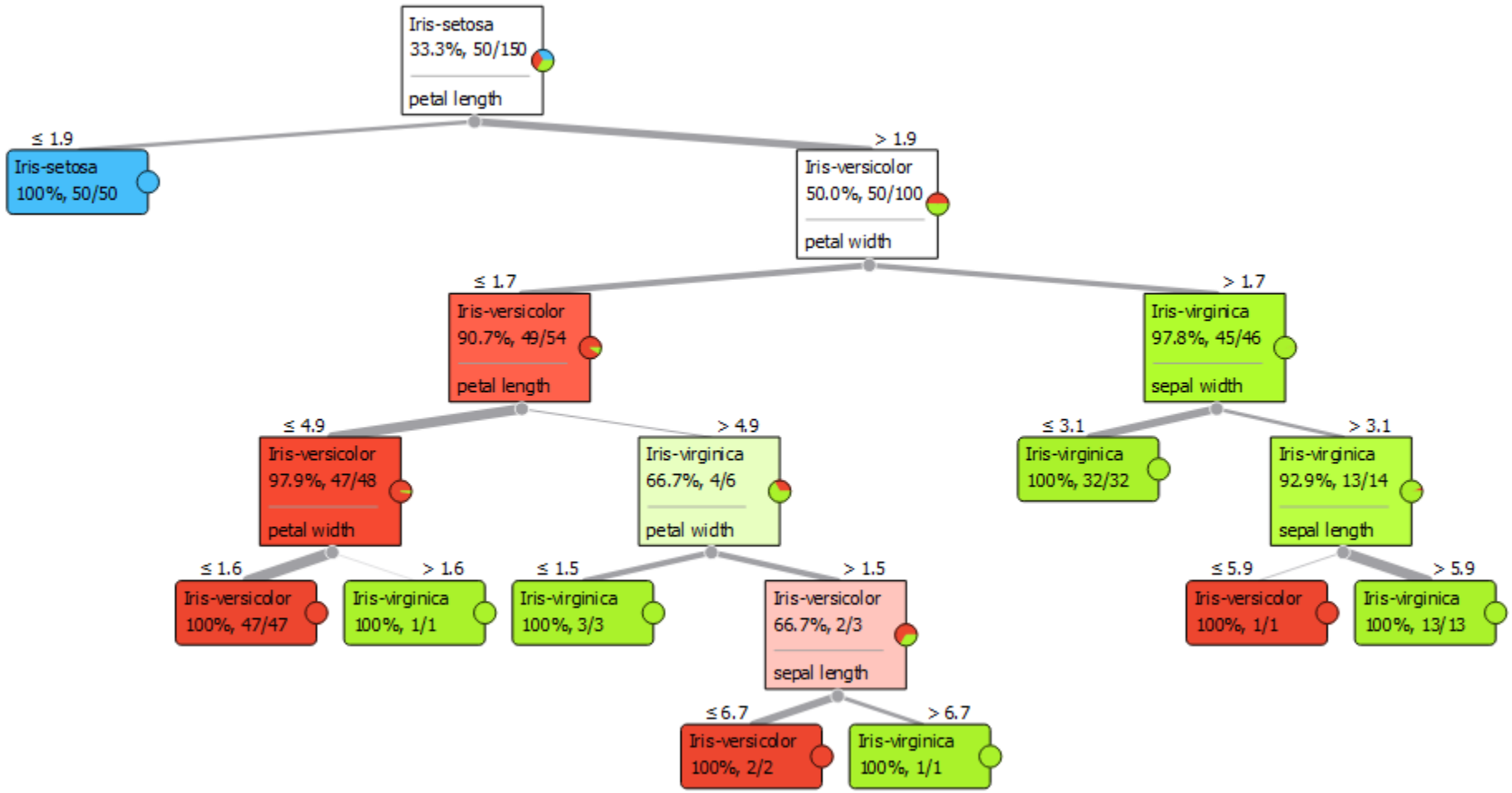




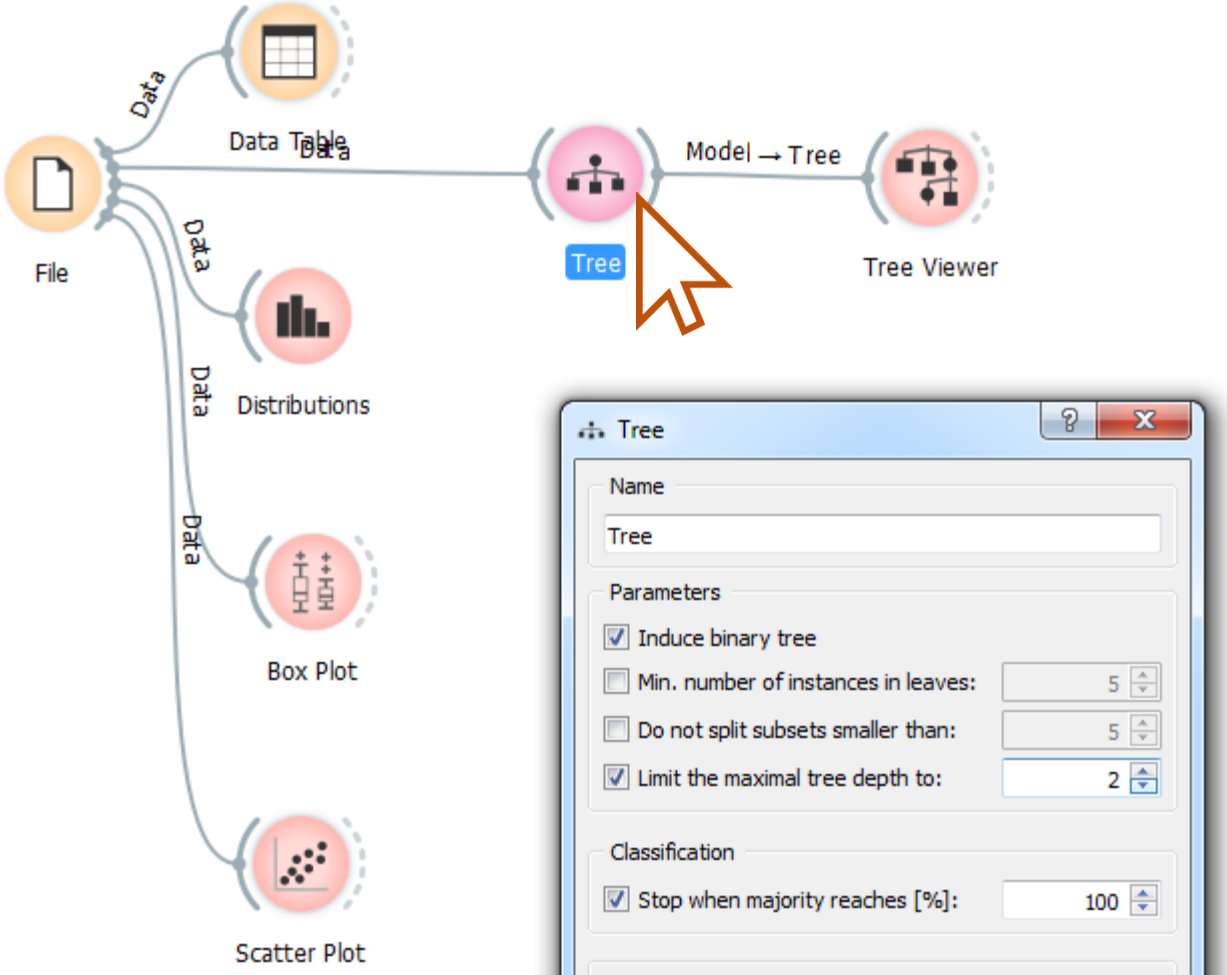


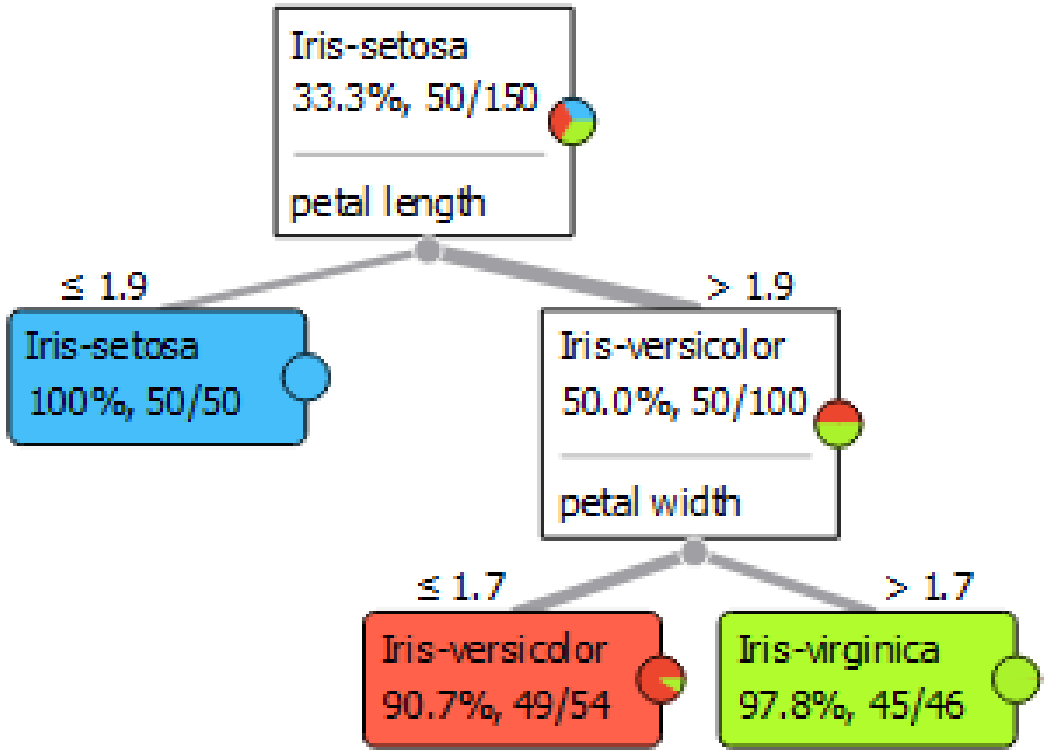

Data
Visualize
Tree Viewe
Box Plot
Distributi...
Scatter Plot
Line Plot
Sieve Diagram
Mosaic Display
FreeViz
Linear Projection
Radviz
Heat Map
Venn Diagram
Silhouette Plot
Pythago... Tree
Pythago... Forest
CN2 Rule Viewer
Nomogram
Model
Evaluate
Unsupervised


Data
Data Table
File
Data
Tree
Model → Tree
Tree Viewer
Data
Distributions
Data
Box Plot
Data
Scatter Plot


Iris-setosa
33.3%, 50/150
petal length
≤ 1.9
Iris-setosa
100%, 50/50
> 1.9
Iris-versicolor
50.0%, 50/100
petal width
≤ 1.7
Iris-versicolor
90.7%, 49/54
petal length
> 1.7
Iris-virginica
97.8%, 45/46
sepal width
≤ 4.9
Iris-versicolor
97.9%, 47/48
petal width
> 4.9
Iris-virginica
66.7%, 4/6
petal width
≤ 3.1
Iris-virginica
100%, 32/32
> 3.1
Iris-virginica
92.9%, 13/14
sepal length
≤ 1.6
Iris-versicolor
100%, 47/47
> 1.6
Iris-virginica
100%, 1/1
≤ 1.5
Iris-virginica
100%, 3/3
> 1.5
Iris-versicolor
66.7%, 2/3
sepal length
≤ 5.9
Iris-versicolor
100%, 1/1
> 5.9
Iris-virginica
100%, 13/13
≤ 6.7
Iris-versicolor
100%, 2/2
> 6.7
Iris-virginica
100%, 1/1

Tenemos que entender los parámetros de los diferentes algoritmos



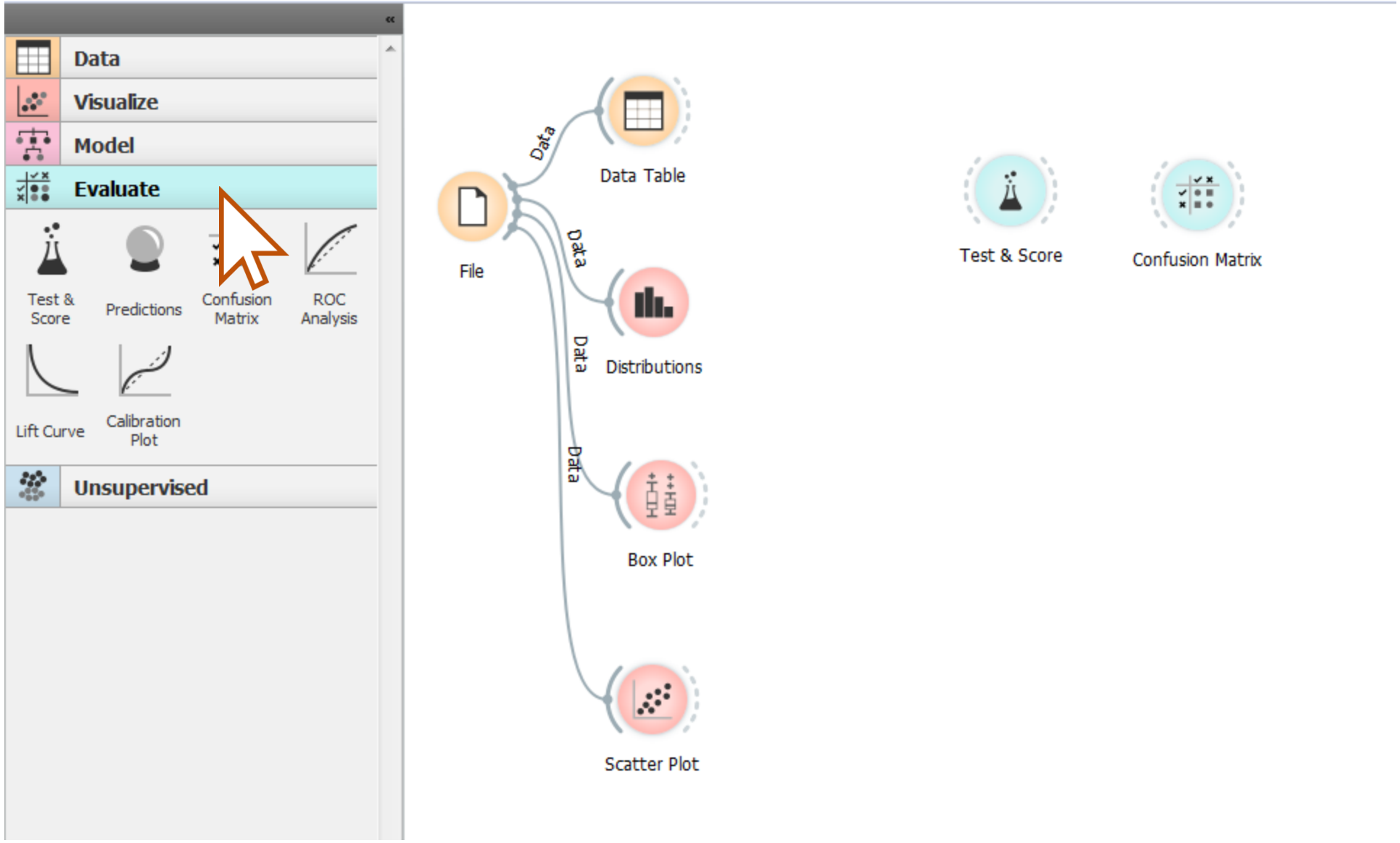


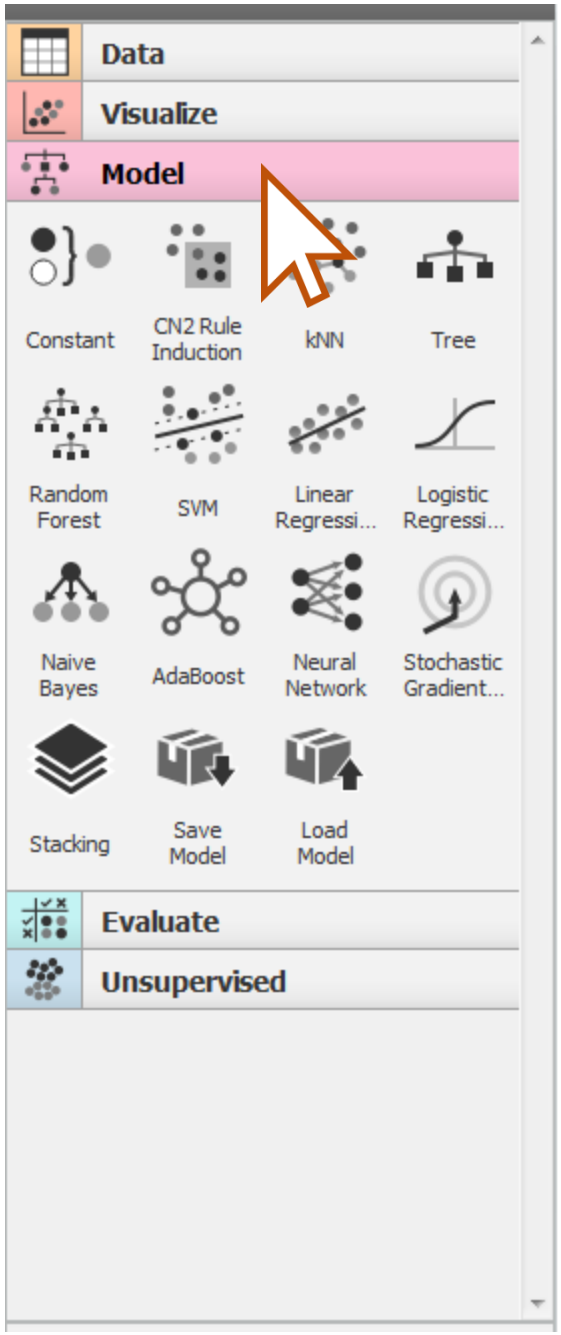
Predicted

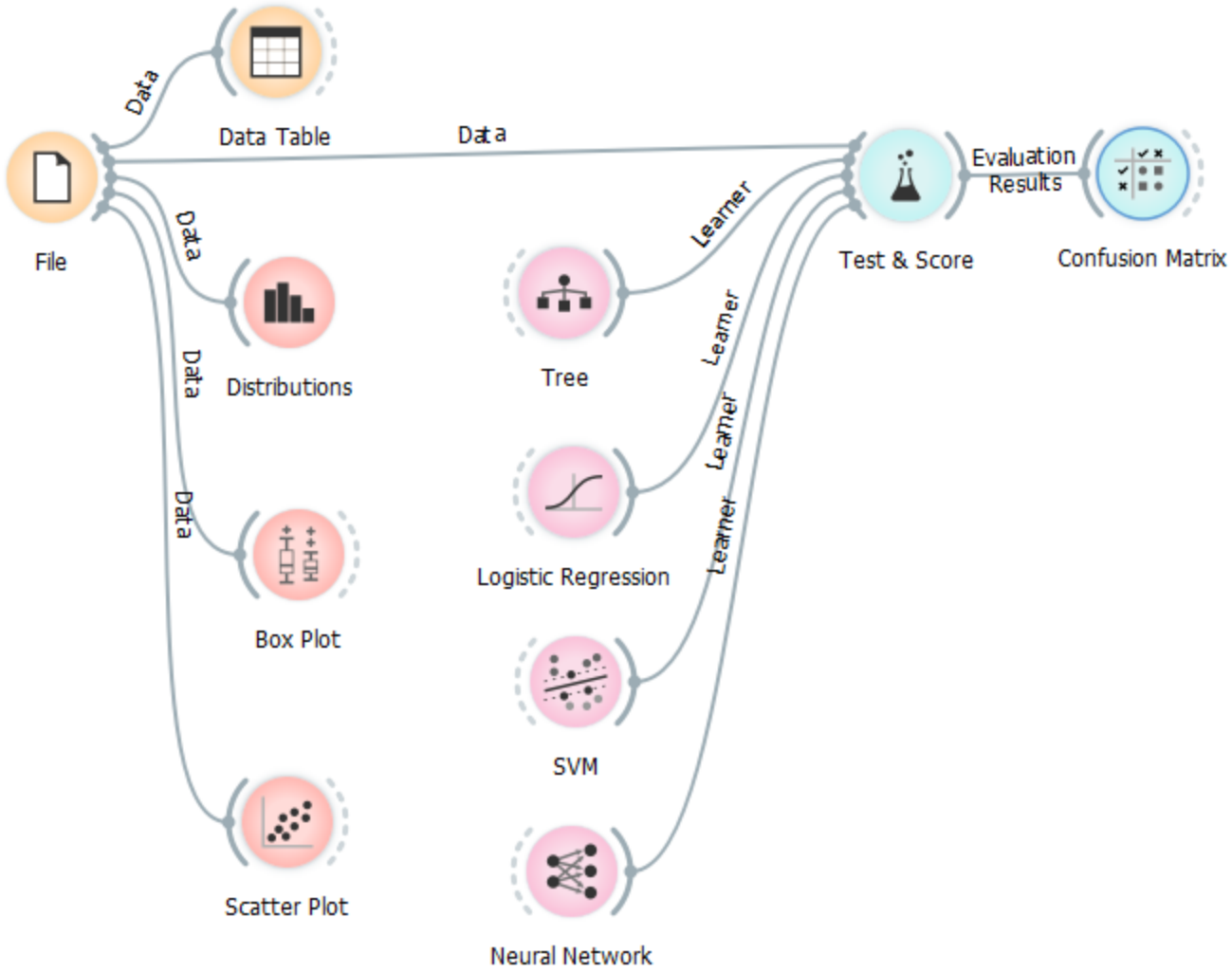
Actual	Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Iris-setosa	50	0	0	50
Iris-versicolor	0	49	1	50
Iris-virginica	0	5	45	50
Σ	50	54	46	150

"96% de clasificaciones correctas"

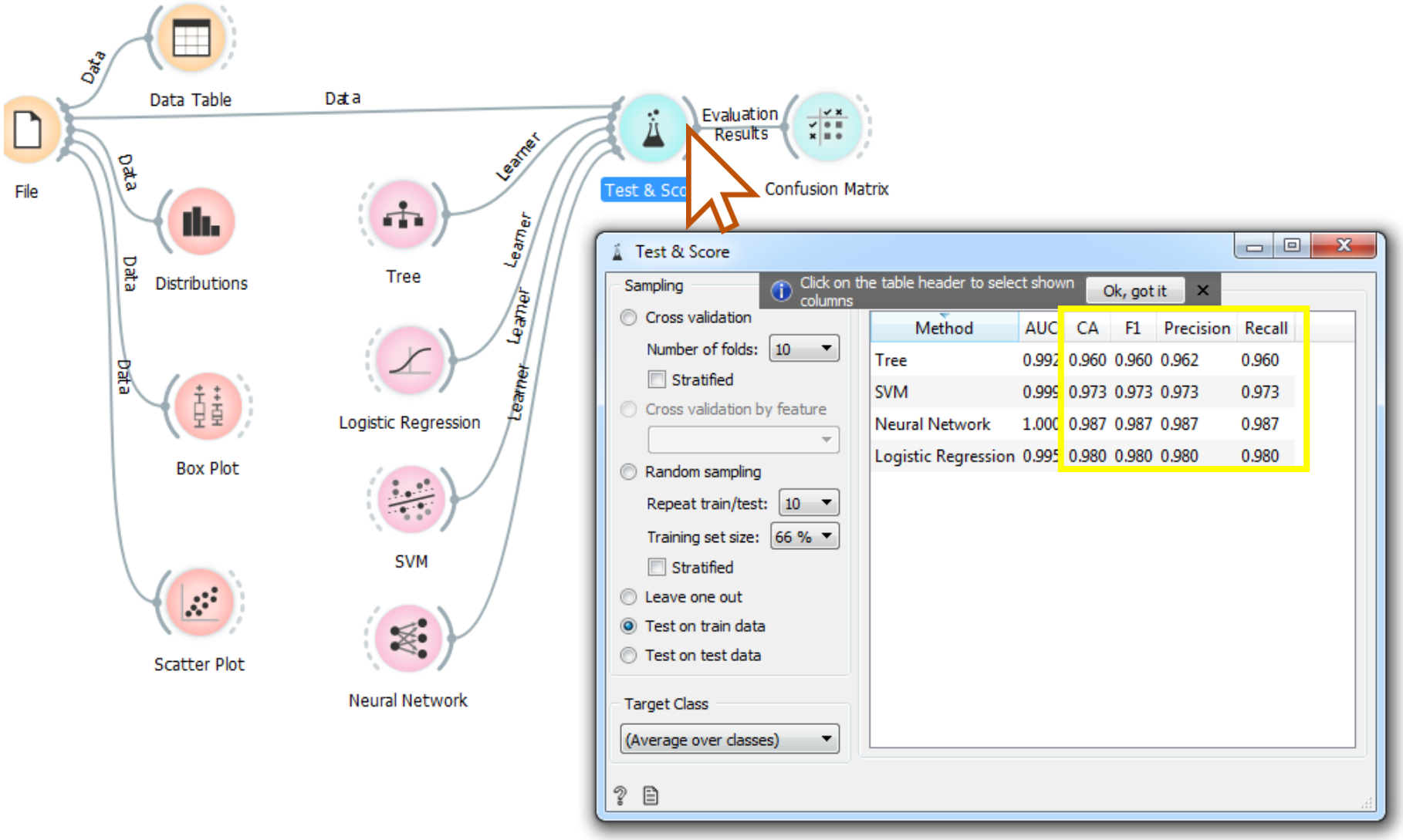
4to paso: Evaluar múltiples algoritmos




Data
Visualize
Model
Constant
CN2 Rule Induction
kNN
Tree
Random Forest
SVM
Linear Regressi...
Logistic Regressi...
Naive Bayes
AdaBoost
Neural Network
Stochastic Gradient...
Stacking
Save Model
Load Model
Evaluate
Unsupervised


File
Data
Data Table
Data
Data
Distributions
Data
Box Plot
Data
Scatter Plot
Tree
Logistic Regression
SVM
Neural Network
Learner
Learner
Learner
Learner
Test & Score
Evaluation Results
Confusion Matrix

Métricas de desempeño facilitan comparar modelos generados por varios algoritmos



Programas de fuente abierta

Existen un sinnúmero de recursos y oportunidades para el aprender sobre este *Aprendizaje Automático*!