성신여자대학교
SUNGSHIN UNIVERSITY

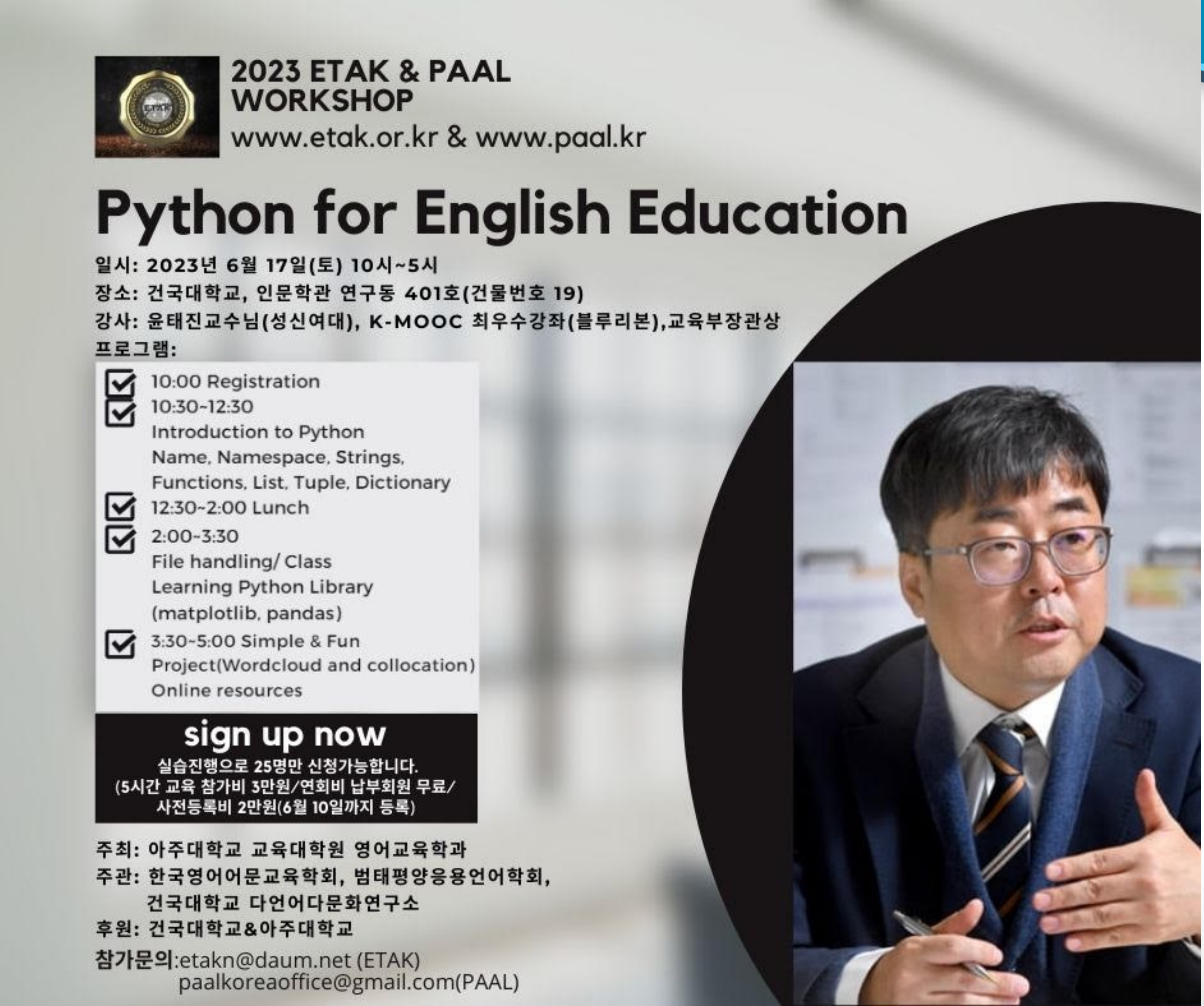
2023 ETAK & PAAL WORKSHOP
www.etak.or.kr & www.paal.kr

2023.06.17 (토) 10:00-17:00
건국대학교, 인문학관 연구동 401호

Python for English Education

Tae-Jin Yoon

Sungshin Women’s University



2023 ETAK & PAAL WORKSHOP
www.etak.or.kr & www.paal.kr

Python for English Education

일시: 2023년 6월 17일(토) 10시~5시
장소: 건국대학교, 인문학관 연구동 401호(건물번호 19)
강사: 윤태진교수님(성신여대), K-MOOC 최우수강좌(블루리본),교육부장관상
프로그램:

☑ 10:00 Registration
☑ 10:30~12:30
Introduction to Python
Name, Namespace, Strings,
Functions, List, Tuple, Dictionary
☑ 12:30~2:00 Lunch
☑ 2:00~3:30
File handling/ Class
Learning Python Library
(matplotlib, pandas)
☑ 3:30~5:00 Simple & Fun
Project(Wordcloud and collocation)
Online resources

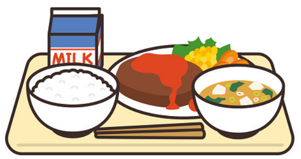
sign up now
실습진행으로 25명만 신청가능합니다.
(5시간 교육 참가비 3만원/연회비 납부회원 무료/
사전등록비 2만원(6월 10일까지 등록)

주최: 아주대학교 교육대학원 영어교육학과
주관: 한국영어어문교육학회, 범태평양응용언어학회,
건국대학교 다언어다문화연구소
후원: 건국대학교&아주대학교
참가문의:etakn@daum.net (ETAK)
paalkoreaoffice@gmail.com(PAAL)

Read map

시간	내용
10:00	Registration
10:30-12:30	Introduction to Python Name, Namespace, Strings, Functions, List, Tuple, Dictionary
12:30-2:00	Lunch
2:00-3:30	File handling/ Class Learning Python Library (matplotlib, pandas)
3:30-5:00	Simple & Fun Project(Wordcloud and collocation) Online resources

01 Introduction to Python and Orange



02 Machine Learning with Orange

03 Text Mining with Orange

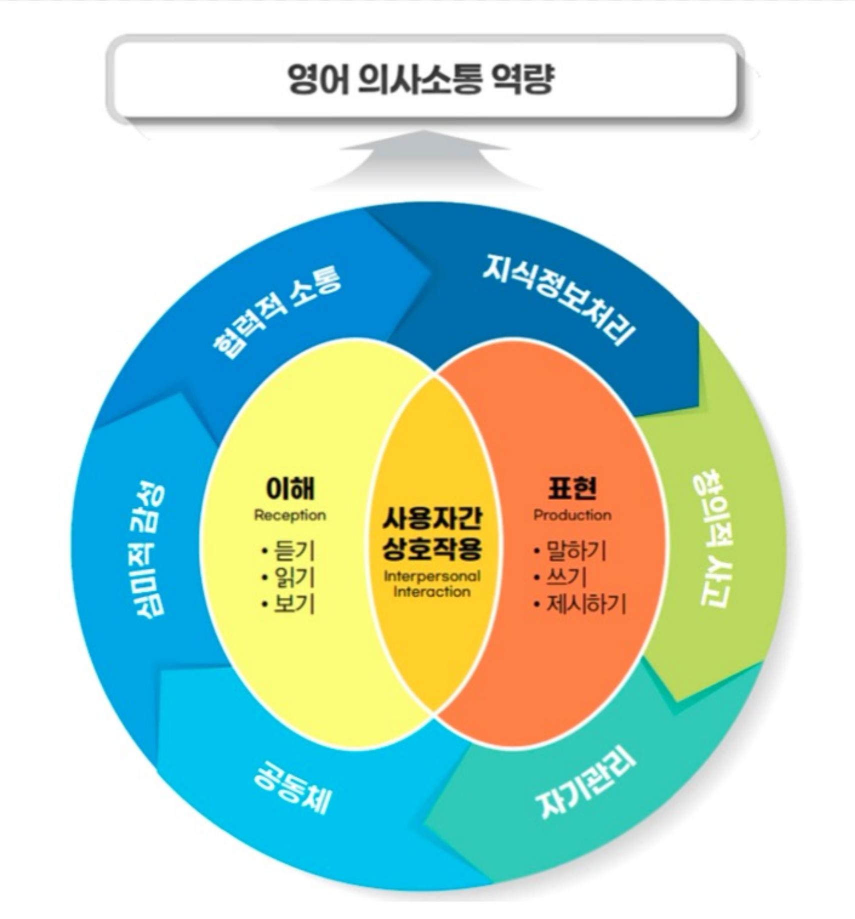
Part 1
Introduction to Python & Orange

10:30–11:20

English Language Teaching & Artificial Intelligence

- 2022개정 영어과 교육과정

(바) 다양한 디지털 평가 도구를 적극적으로 활용한다. 디지털 분석·평가 도구를 활용하여 실제적인 평가 맥락을 제공하고 다양한 학습자 데이터를 체계적으로 구축한다. 이를 토대로 다각적이고 신뢰할 만한 평가 결과를 도출할 수 있다.



영어 의사소통 역량

협력적 소통
지식정보처리
창의적 사고
자기관리
공동체
심미적 감성

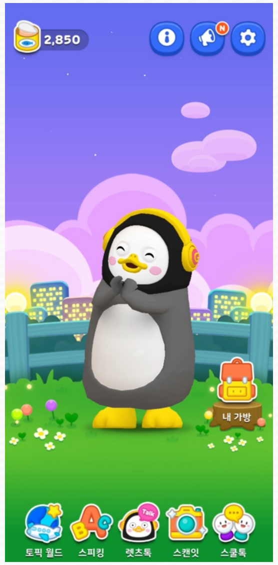
이해
Reception
- 듣기
- 읽기
- 보기

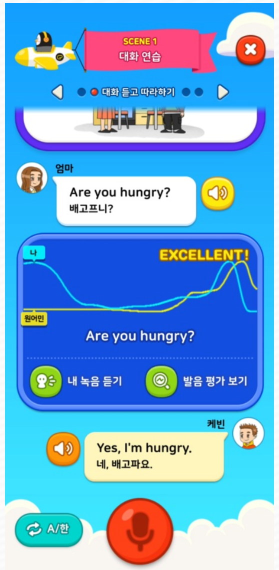
사용자간 상호작용
Interpersonal Interaction

표현
Production
- 말하기
- 쓰기
- 제시하기

[그림] 영어과 역량 및 영역 구성

교육부 고시 제2022-33호 [별책 1]

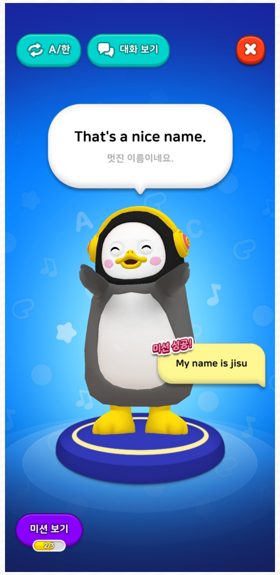




엄마
Are you hungry?
배고프니?

Are you hungry?

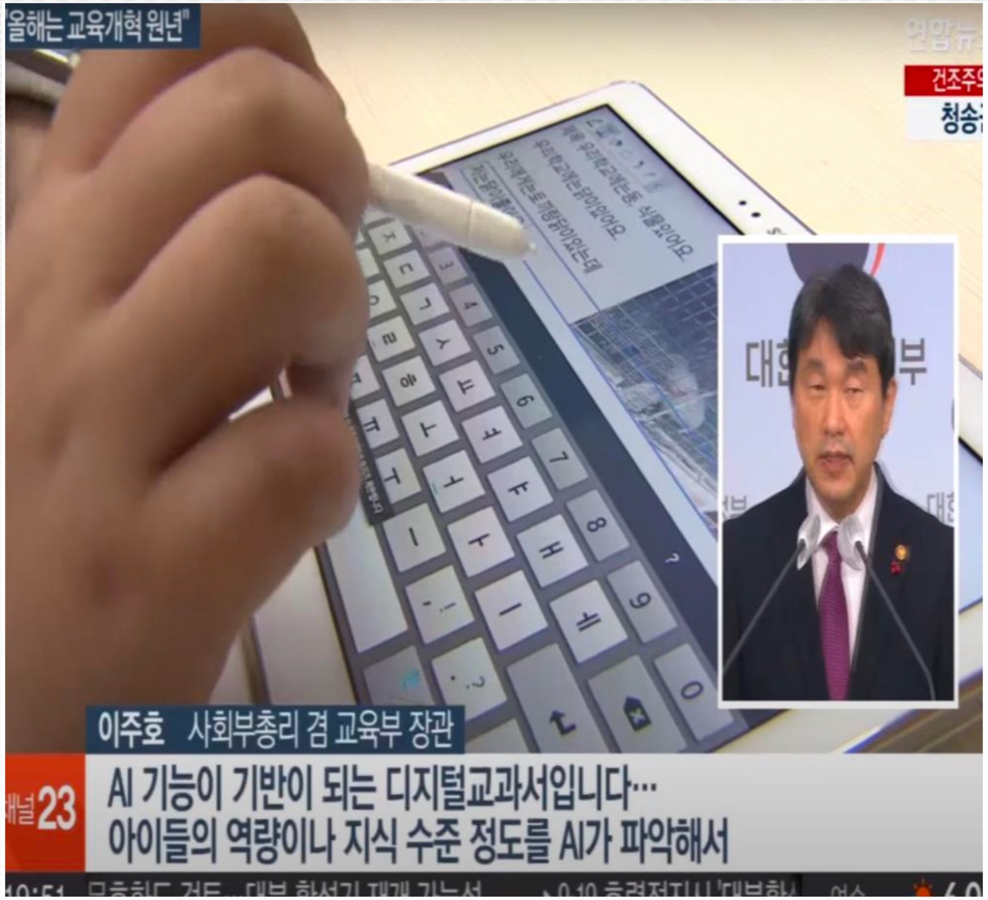
케빈
Yes, I'm hungry.
네, 배고파요.



That's a nice name.
멋진 이름이네요.

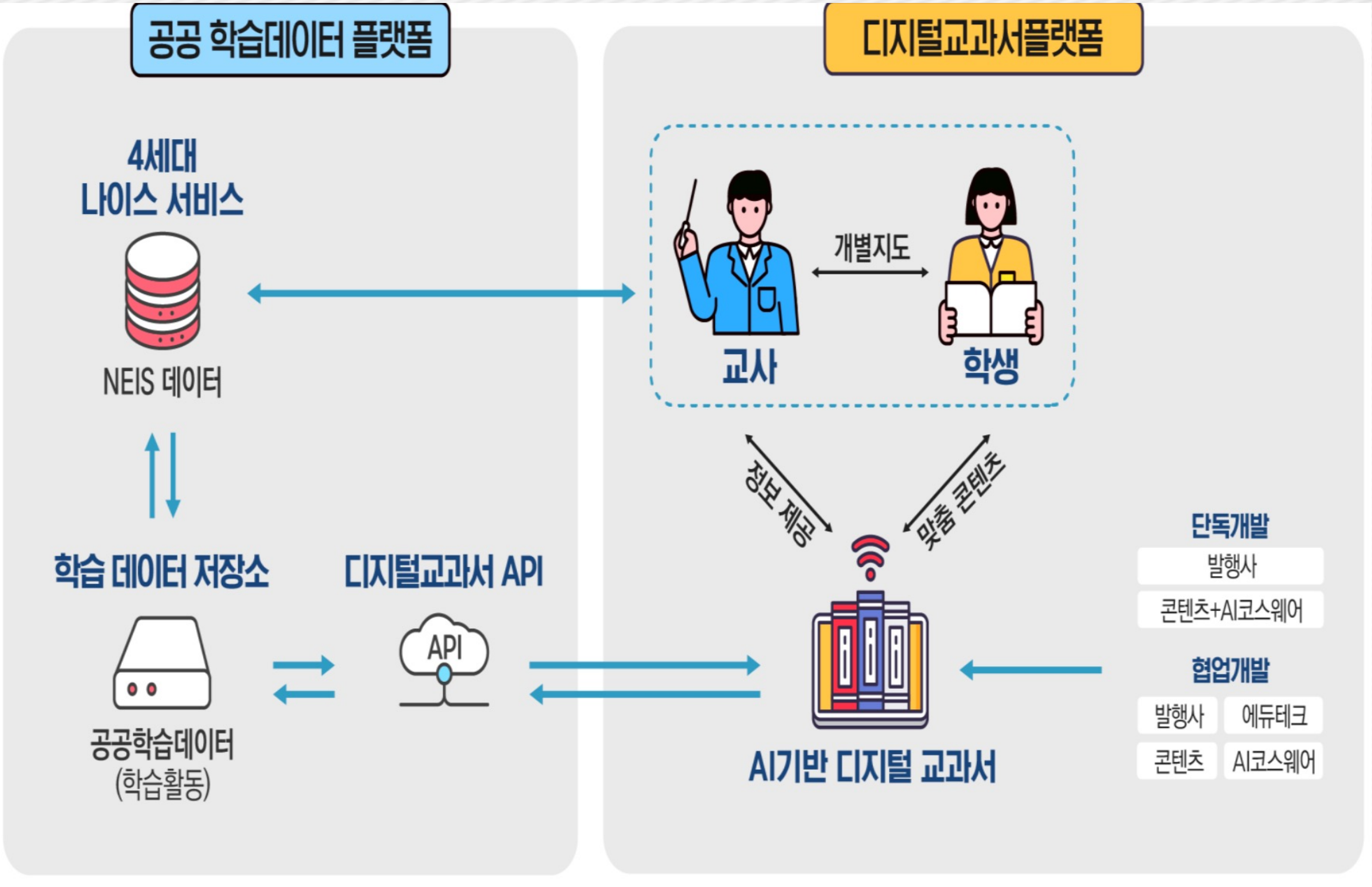
My name is jisu

English Language Teaching & Artificial Intelligence

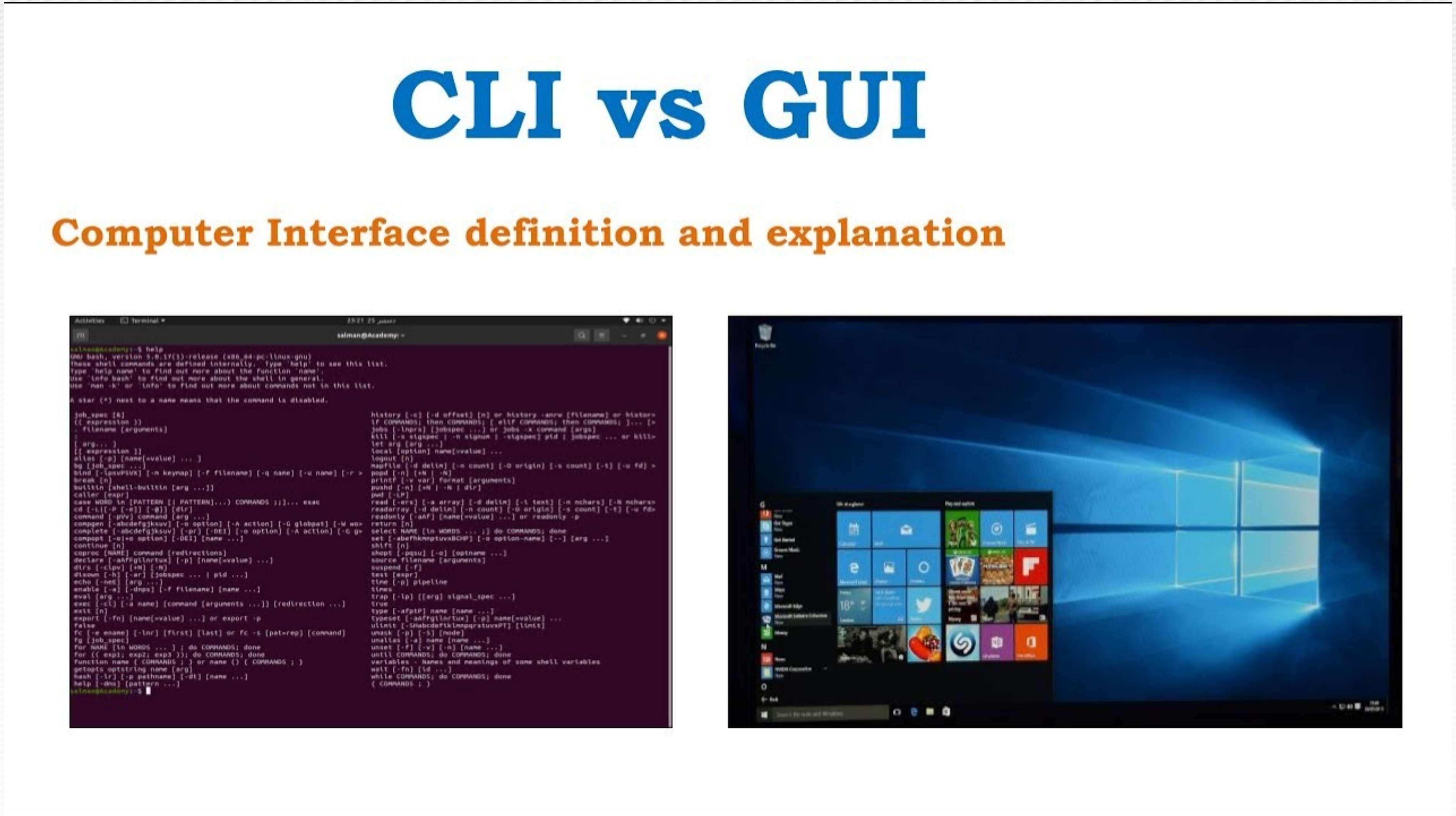

올해는 교육개혁 원년"

이주호 사회부총리 겸 교육부 장관

AI 기능이 기반이 되는 디지털교과서입니다…
아이들의 역량이나 지식 수준 정도를 AI가 파악해서



GUI vs. CLI



CLI vs GUI

Computer Interface definition and explanation

Guido van Rossum
귀도 반 로섬

89 크리스마스 연휴

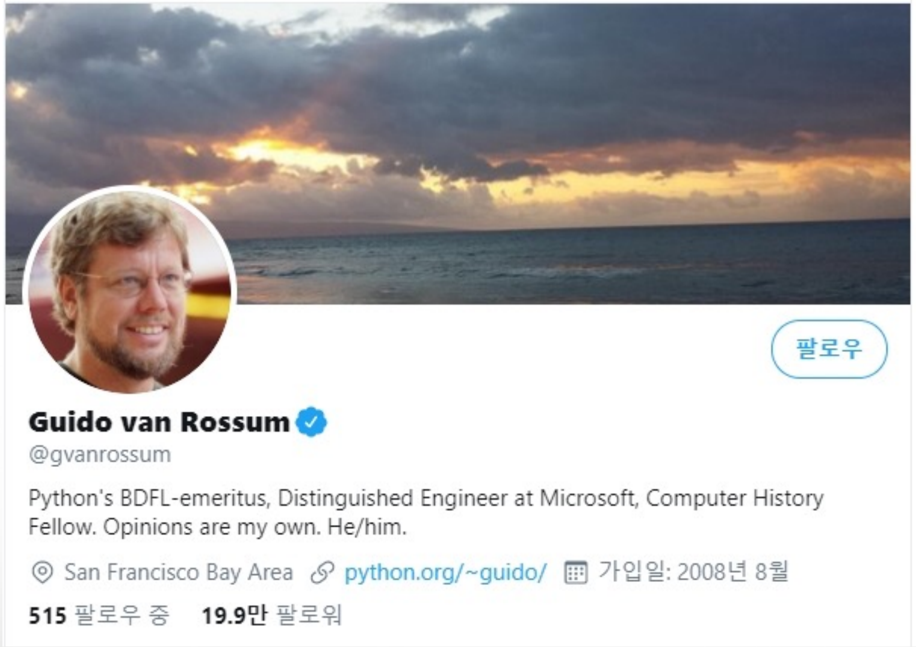
ent

e

Guido Van Rossum

99 DARPA

Computer Programming for Everybody



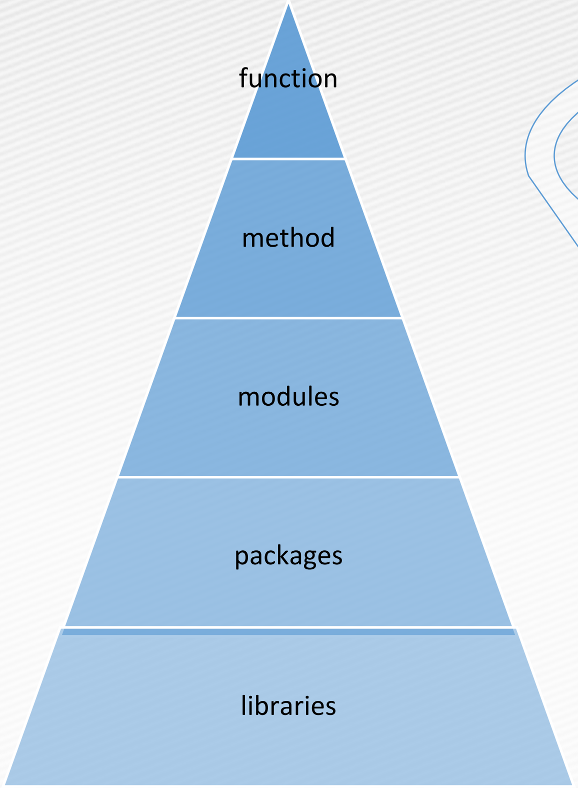
Guido van Rossum
@gvanrossum

Python's BDFL-emeritus, Distinguished Engineer at Microsoft, Computer History Fellow. Opinions are my own. He/him.

San Francisco Bay Area python.org/~guido/ 가입일: 2008년 8월

515 팔로우 중 19.9만 팔로워

https://www.python.org/doc/essays/cp4e/



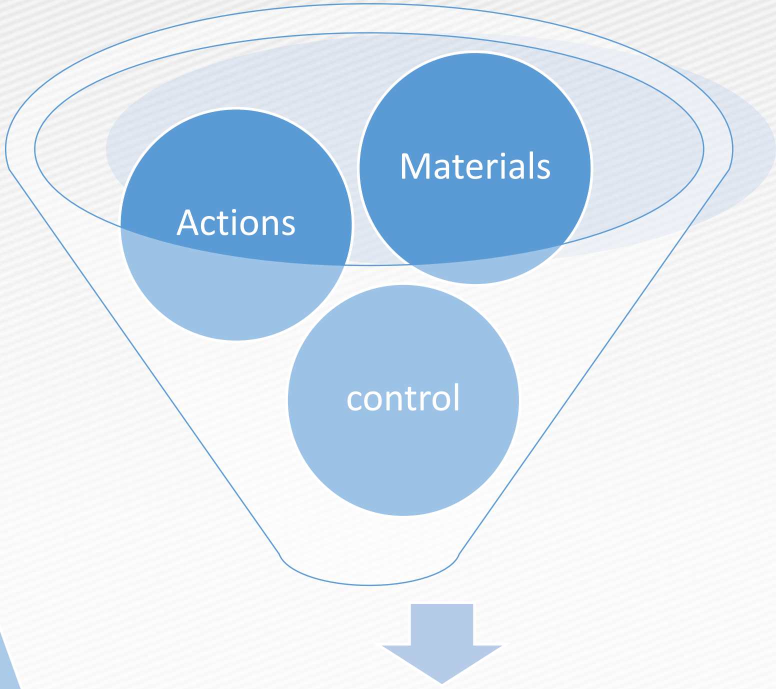
function

method

modules

packages

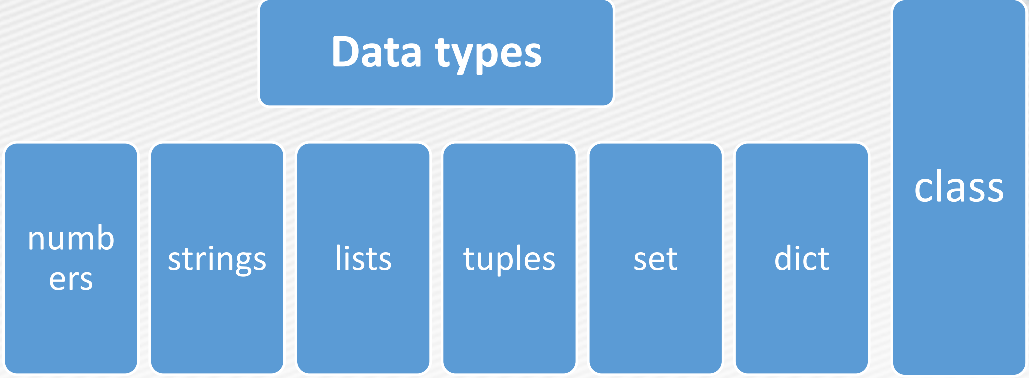
libraries



Materials

Actions

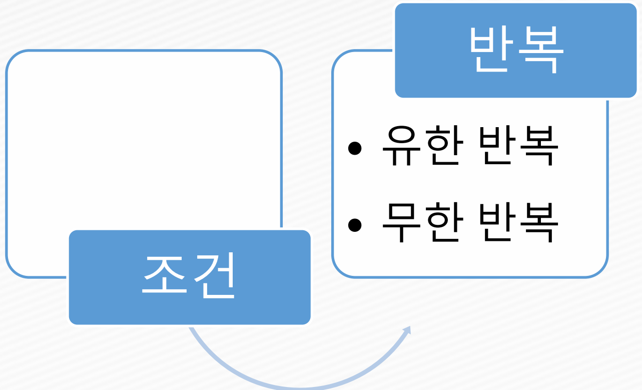
control



Data types

- numbers
- strings
- lists
- tuples
- set
- dict

class



조건

반복

- 유한 반복
- 무한 반복

출처 https://realpython.com/python-basics/

실습

```
pip3 install wordcloud
```

library

```
text = “Sungshin Sungshin Sungshin Sungshin University University”
```

variable

str(ing) data type

```
import wordcloud
```

library

```
wc_obj = wordcloud.WordCloud()
```

instance

class

```
wc = wc_obj.generate(text)
```

method

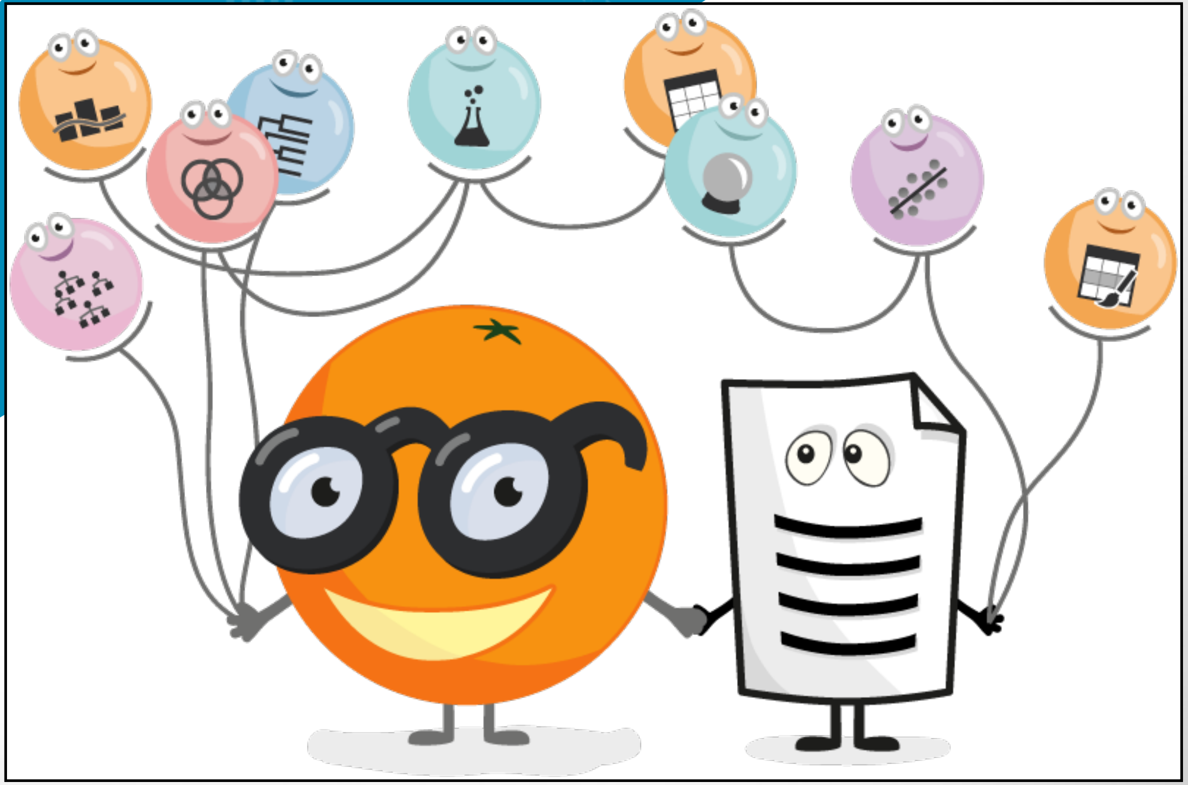
```
wc.words_
```

attribute

Part 2
Installing orange

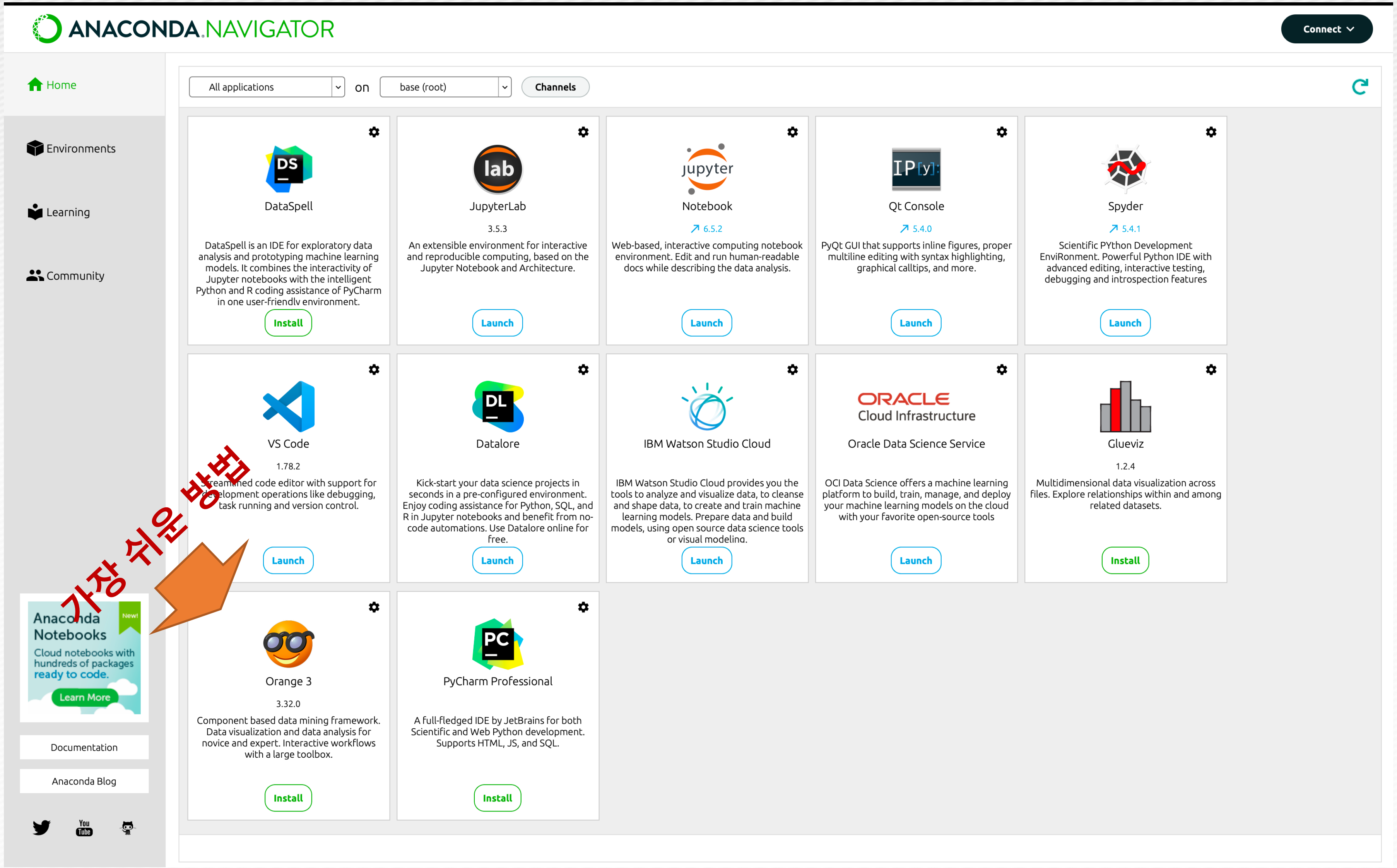
Orange is developed by Bioinformatics Lab at University of Ljubljana, Slovenia.

Initial release: 10 October 1996; 26 years ago



11:20–11:30

Orange 설치



ANACONDA.NAVIGATOR

DataSpell
DataSpell is an IDE for exploratory data analysis and prototyping machine learning models. It combines the interactivity of Jupyter notebooks with the intelligent Python and R coding assistance of PyCharm in one user-friendly environment.
Install

JupyterLab
3.5.3
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
Launch

Notebook
↗ 6.5.2
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Launch

Qt Console
↗ 5.4.0
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.
Launch

Spyder
↗ 5.4.1
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features
Launch

VS Code
1.78.2
reamlined code editor with support for elopment operations like debugging, task running and version control.
Launch

Datalore
Kick-start your data science projects in seconds in a pre-configured environment. Enjoy coding assistance for Python, SQL, and R in Jupyter notebooks and benefit from no-code automations. Use Datalore online for free.
Launch

IBM Watson Studio Cloud
IBM Watson Studio Cloud provides you the tools to analyze and visualize data, to cleanse and shape data, to create and train machine learning models. Prepare data and build models, using open source data science tools or visual modeling.
Launch

ORACLE Cloud Infrastructure
Oracle Data Science Service
OCI Data Science offers a machine learning platform to build, train, manage, and deploy your machine learning models on the cloud with your favorite open-source tools
Launch

Glueviz
1.2.4
Multidimensional data visualization across files. Explore relationships within and among related datasets.
Install

Orange 3
3.32.0
Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.
Install

PyCharm Professional
A full-fledged IDE by JetBrains for both Scientific and Web Python development. Supports HTML, JS, and SQL.
Install

가장 쉬운 방법

or

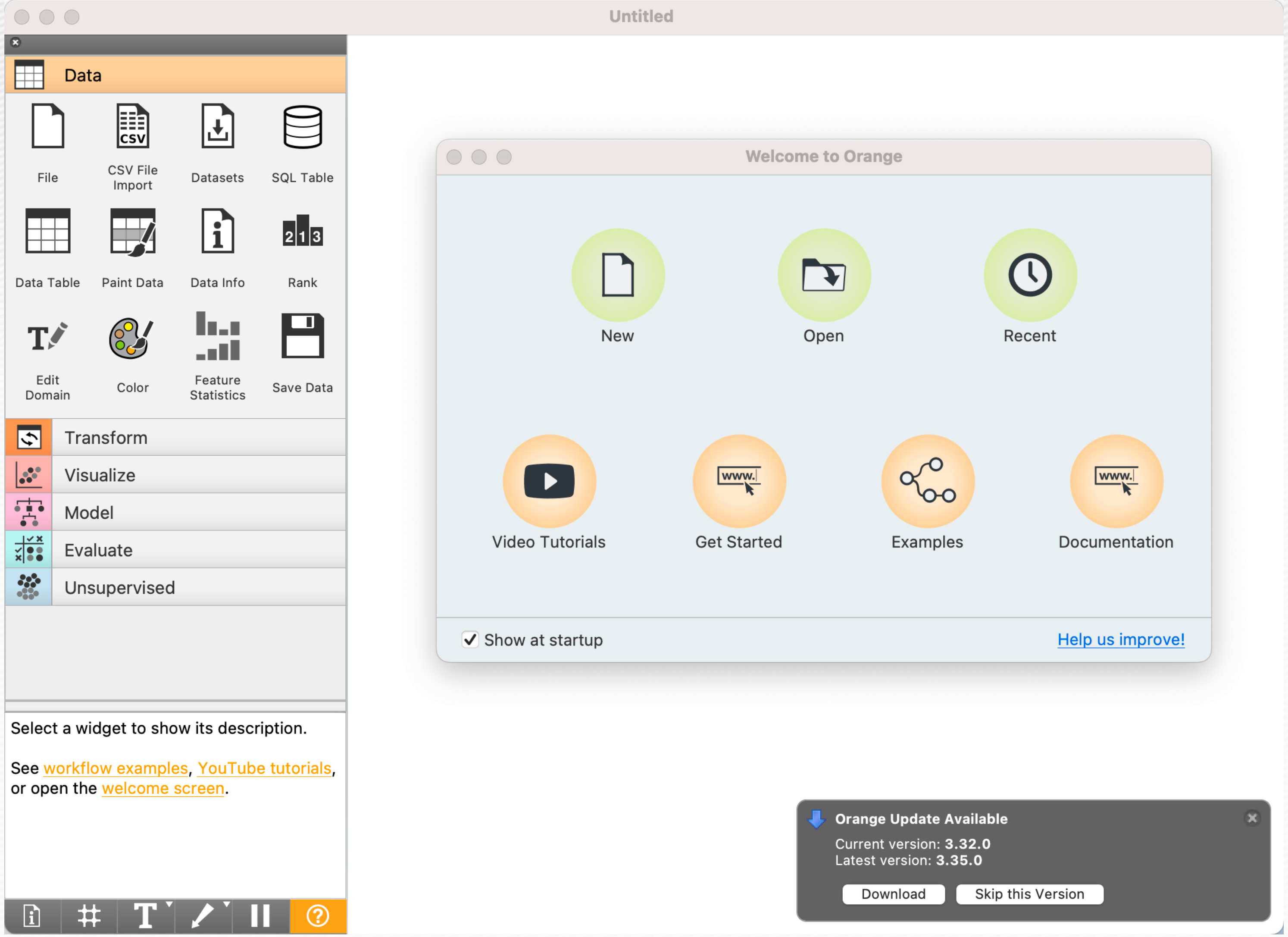
Orange 설치

or

```
conda config --add channels conda-forge
conda install pyqt
conda install orange3
```

```
pip install PyQt5 PyQtWebEngine
pip install orange3
```

Orange Visual Programming



Select a widget to show its description.

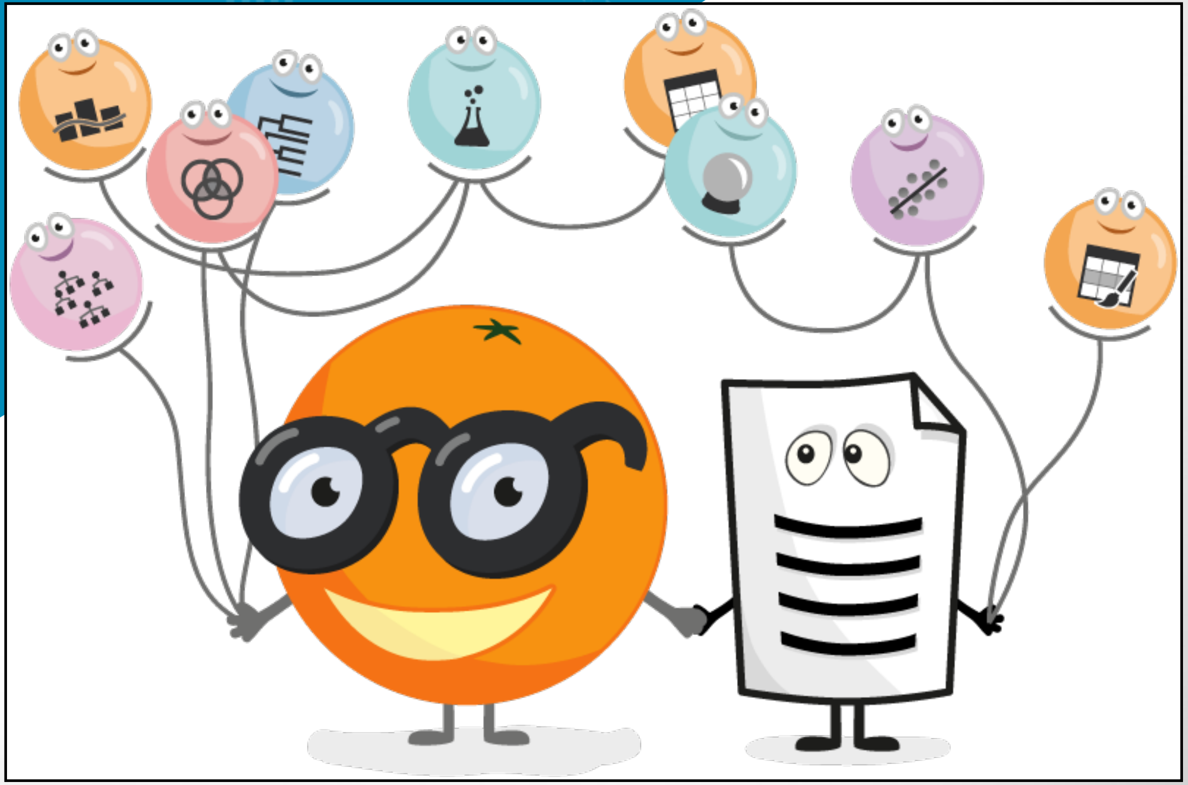
See workflow examples, YouTube tutorials, or open the welcome screen.

Orange Update Available
Current version: **3.32.0**
Latest version: **3.35.0**

https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/index.html

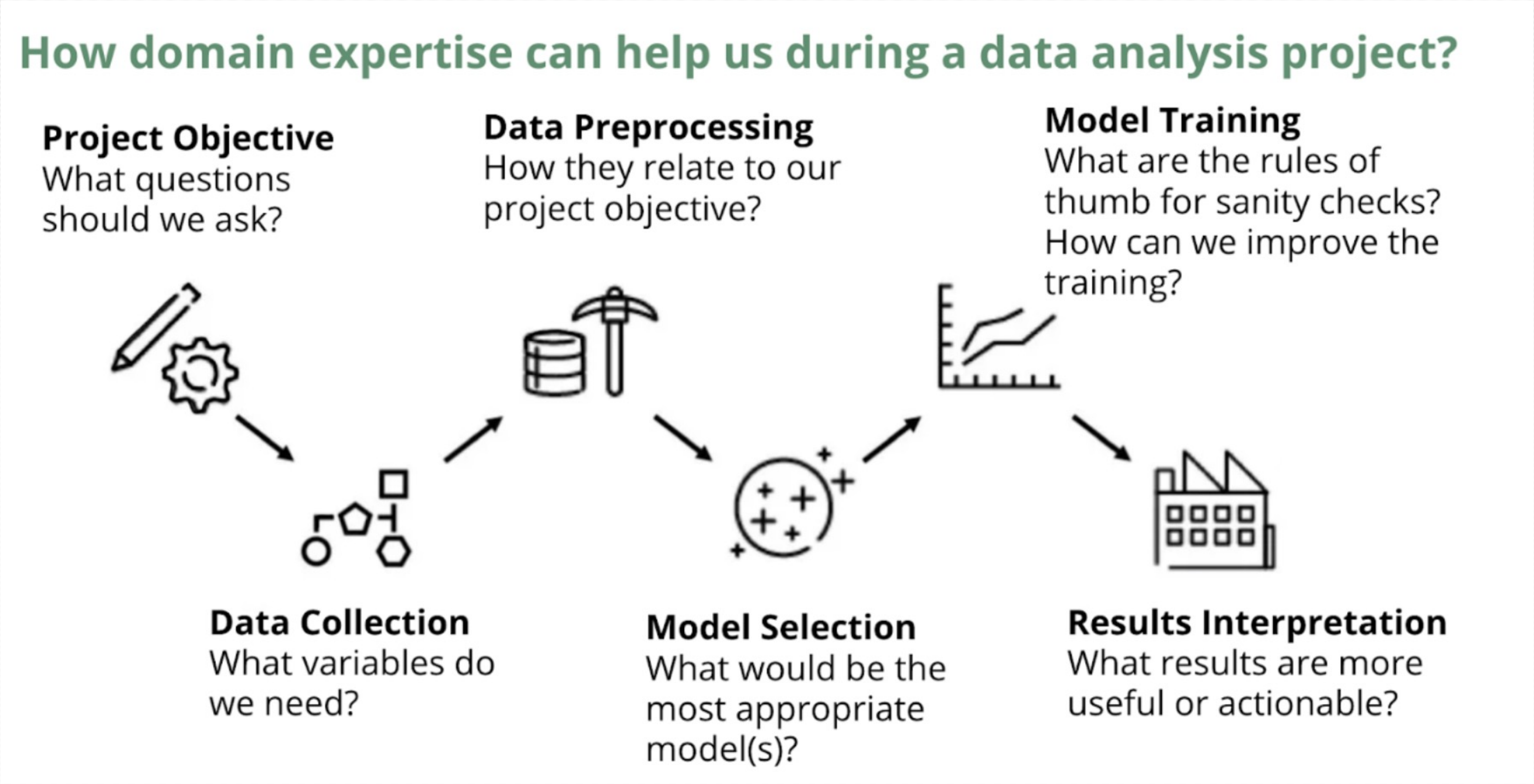
Part 3

Visual Programming with orange



11:30–12:00

Workflow for a data analysis project



How domain expertise can help us during a data analysis project?

Project Objective
What questions should we ask?

Data Collection
What variables do we need?

Data Preprocessing
How they relate to our project objective?

Model Selection
What would be the most appropriate model(s)?

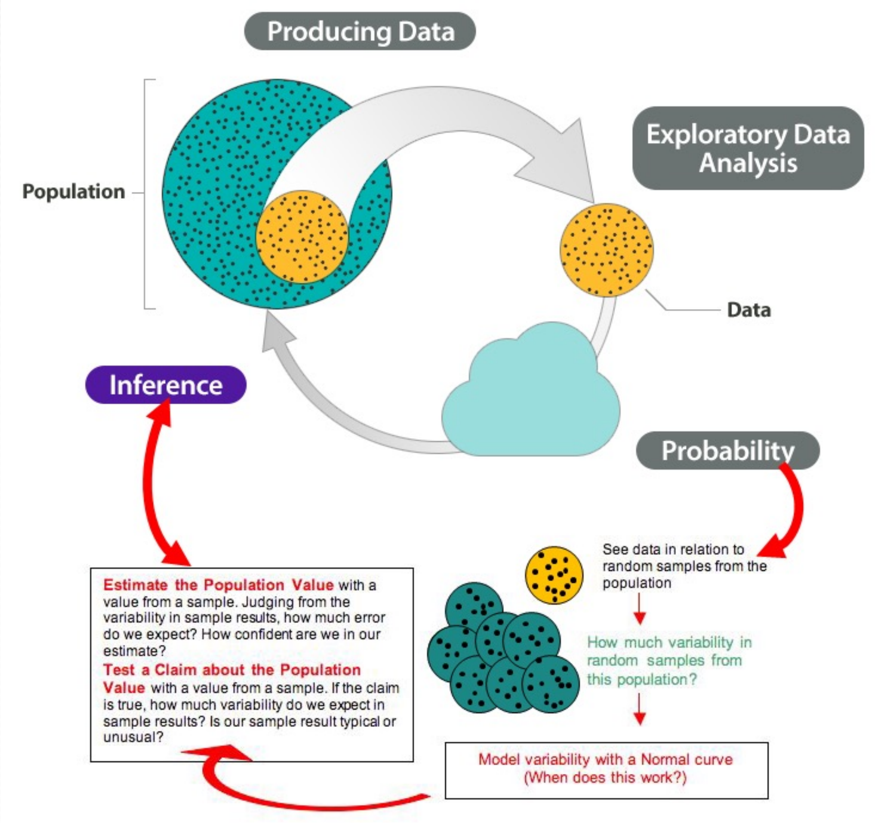
Model Training
What are the rules of thumb for sanity checks? How can we improve the training?

Results Interpretation
What results are more useful or actionable?

Source: https://blog.ml.cmu.edu/2020/08/31/1-domain-knowledge/

Statistical Inference vs. Modeling

Inference



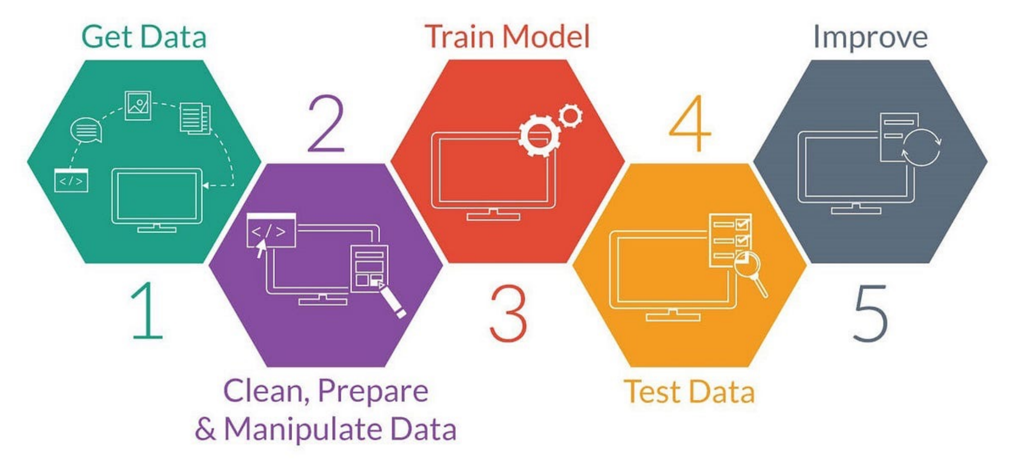
See data in relation to random samples from the population

How much variability in random samples from this population?

Model variability with a Normal curve (When does this work?)

Estimate the Population Value with a value from a sample. Judging from the variability in sample results, how much error do we expect? How confident are we in our estimate?
Test a Claim about the Population Value with a value from a sample. If the claim is true, how much variability do we expect in sample results? Is our sample result typical or unusual?

Modeling

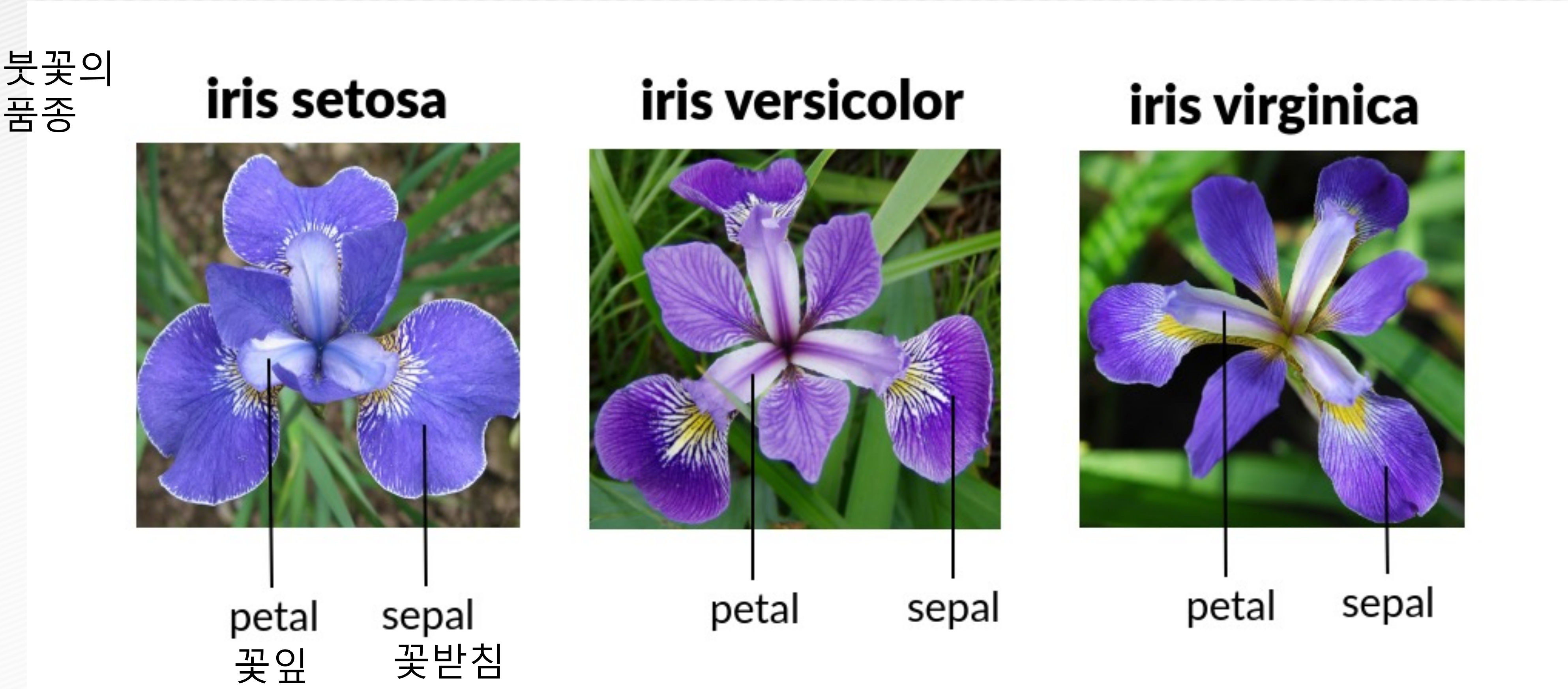


https://www.coursesidekick.com/statistics/study-guides/wmopen-concepts-statistics/wim-linking-probability-to-statistical-infe

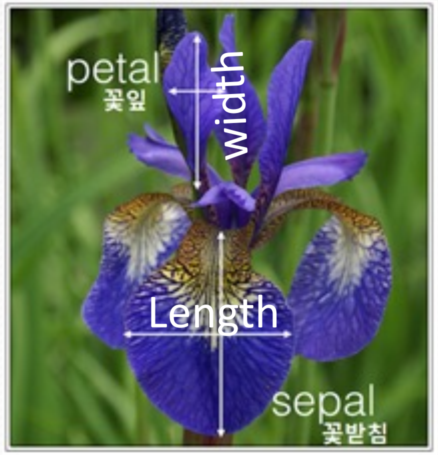
Classic Example: iris classification

https://archive.ics.uci.edu/ml/datasets/Iris

붓꽃의
품종

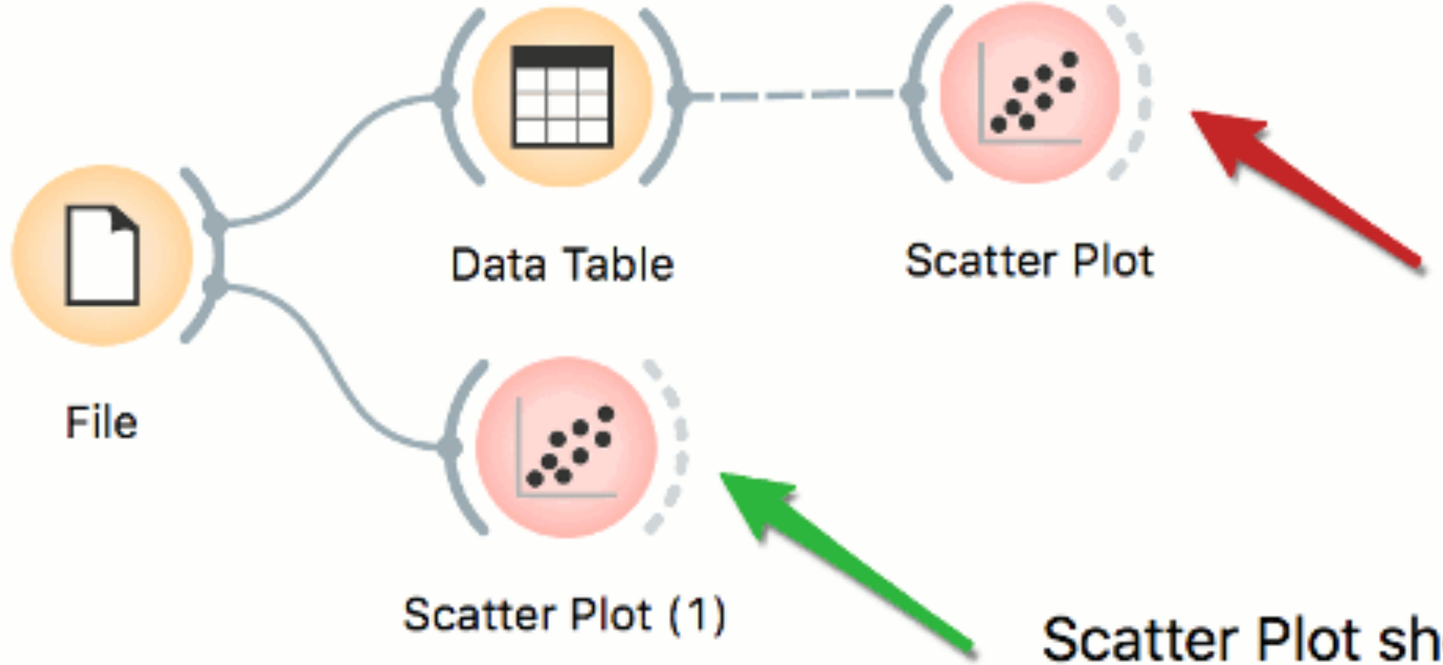


Measurement (in cm)



https://www.analyticsvidhya.com/blog/2022/06/iris-flowers-classification-using-machine-learning/

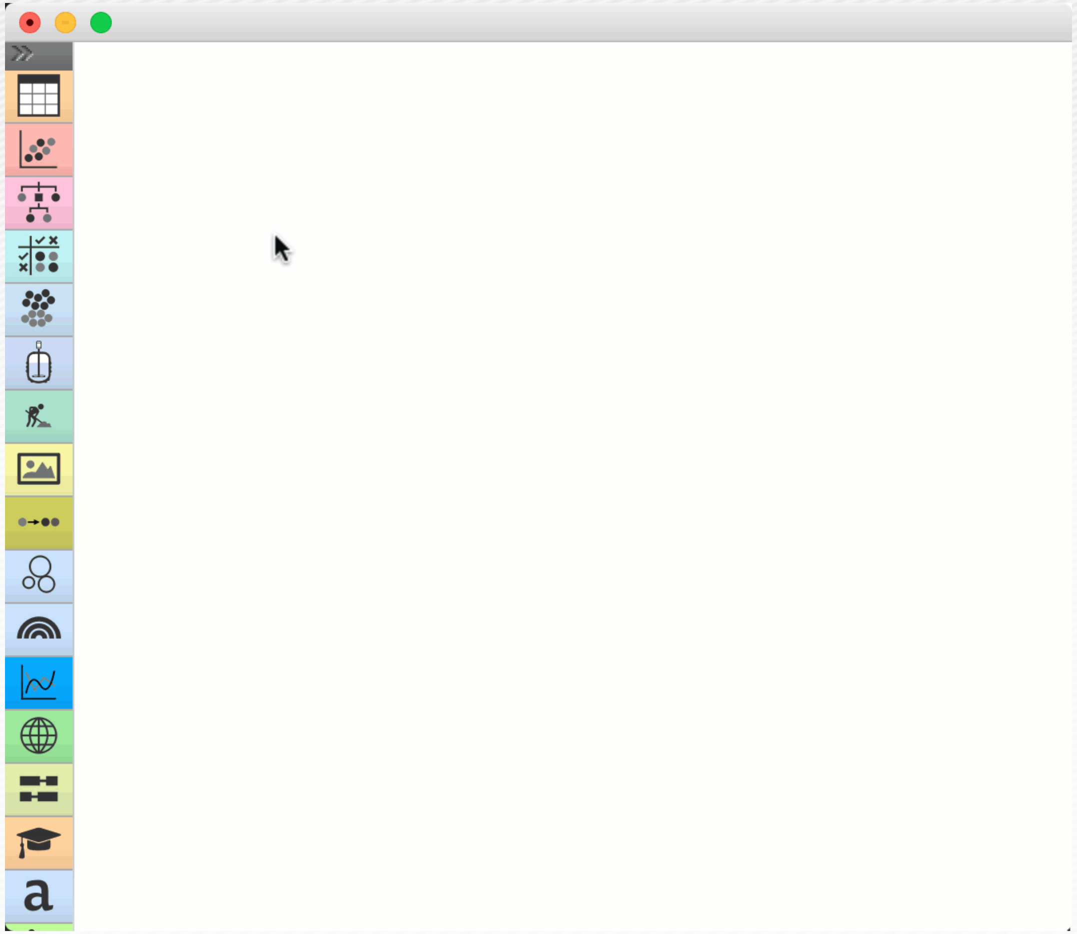
Orange: Building workflow by connecting widgets

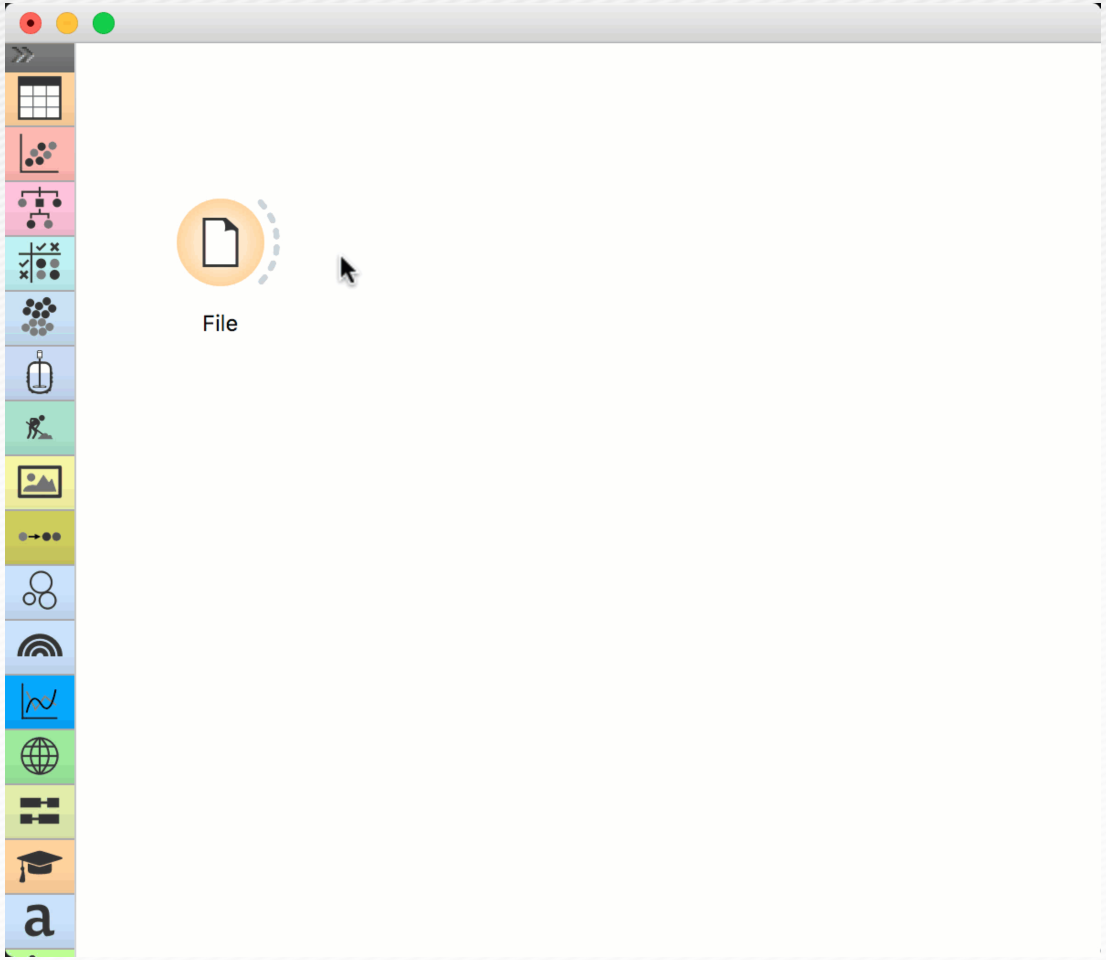


Data Table outputs selected data. Since no data is selected, the connection is empty (dashed). Scatter Plot is empty.

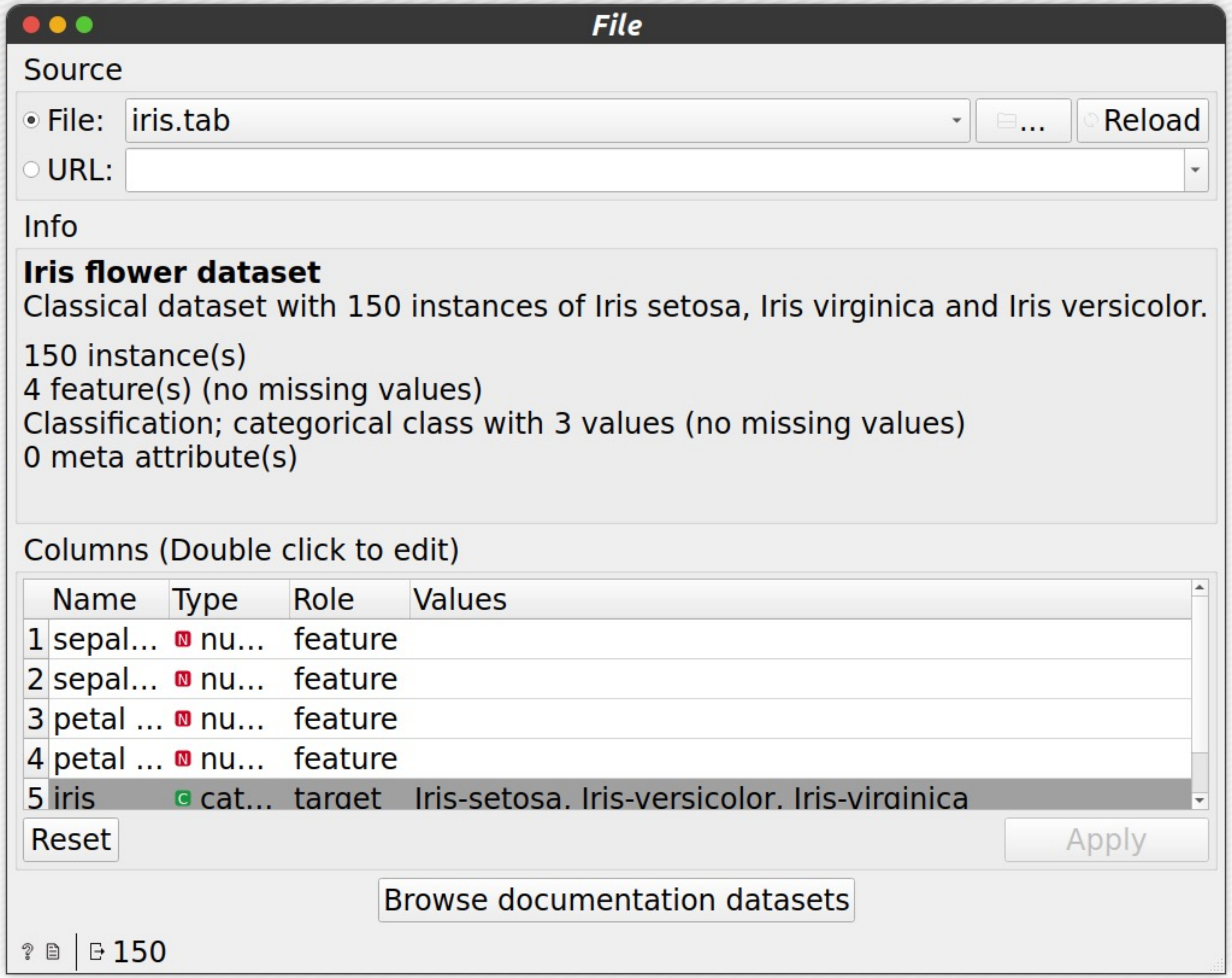
Scatter Plot show a plot. The connection is full, as the File widget is sending data directly to the widget.

How to build in Orange





Loading data using File widget



Source

● File: iris.tab ... Reload

○ URL:

Info

Iris flower dataset
Classical dataset with 150 instances of Iris setosa, Iris virginica and Iris versicolor.

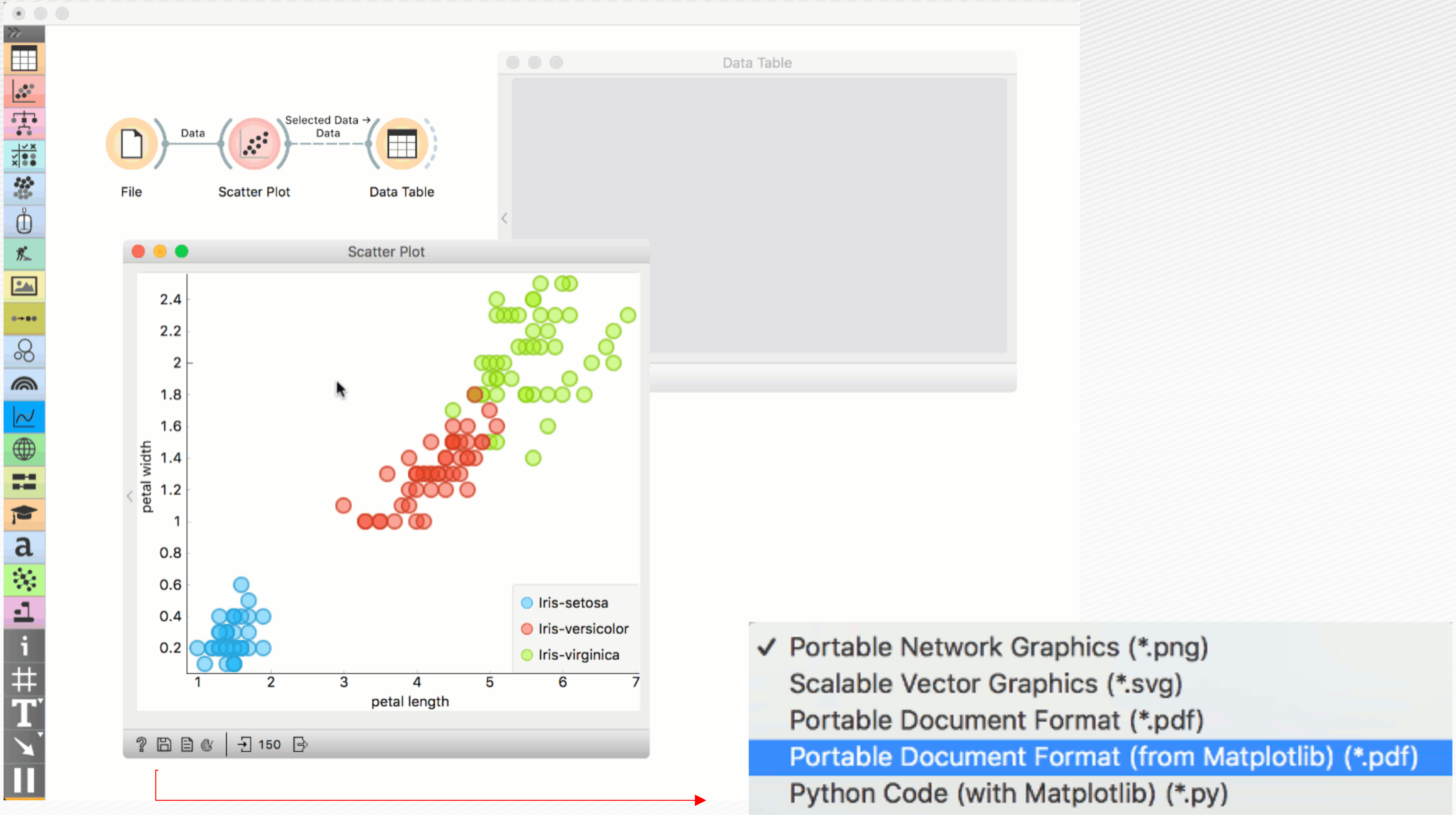
150 instance(s)
4 feature(s) (no missing values)
Classification; categorical class with 3 values (no missing values)
0 meta attribute(s)

Columns (Double click to edit)

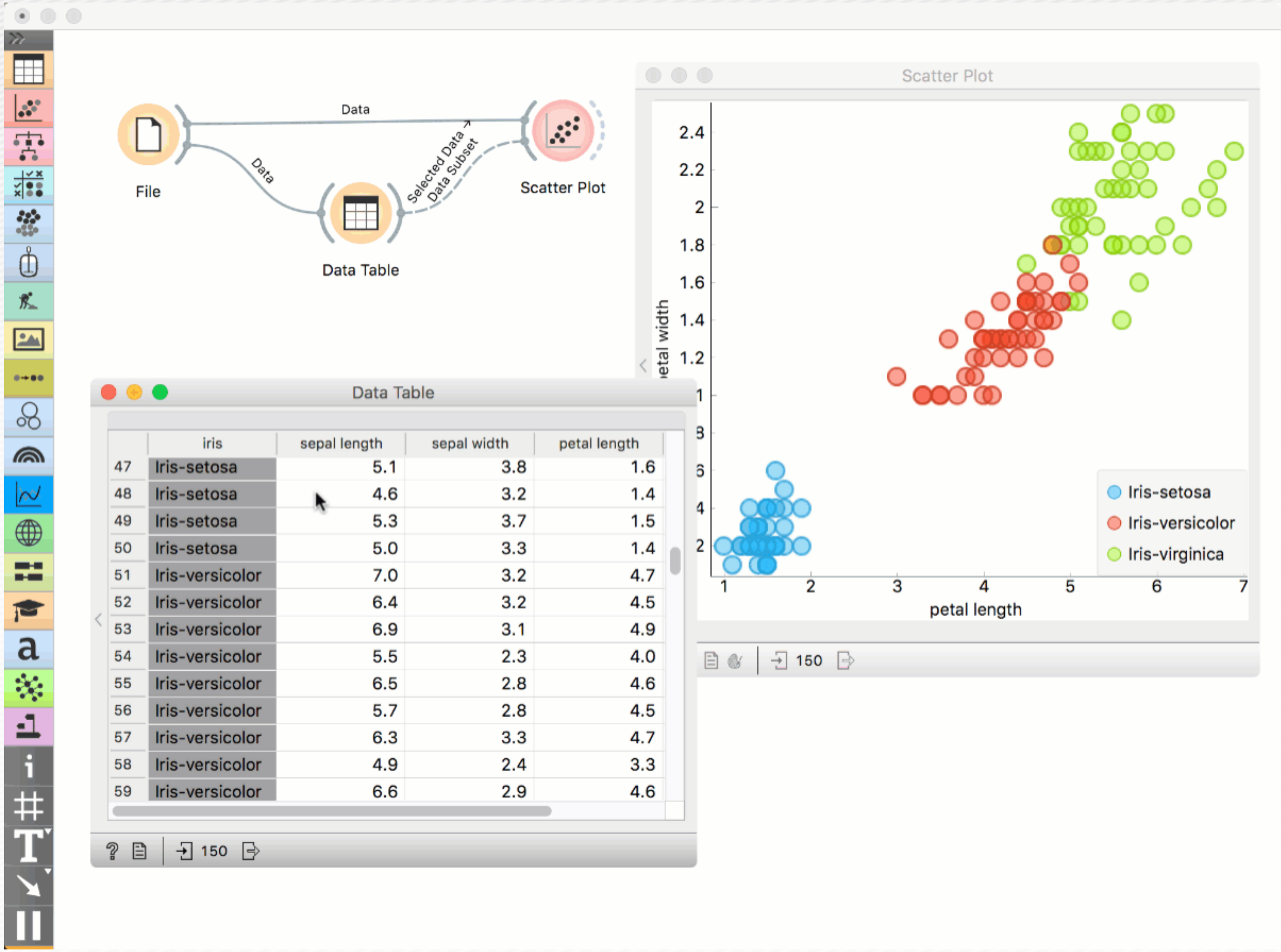
	Name	Type	Role	Values
1	sepal...	N nu...	feature	
2	sepal...	N nu...	feature	
3	petal ...	N nu...	feature	
4	petal ...	N nu...	feature	
5	iris	C cat...	target	Iris-setosa, Iris-versicolor, Iris-virginica

- Loading dataset we have to use **File Widget** which is available in Data Section.
- After clicking on File Widget, it will automatically appear on Canvas then you have to double click on that widget.

Selection subset of the data (1)

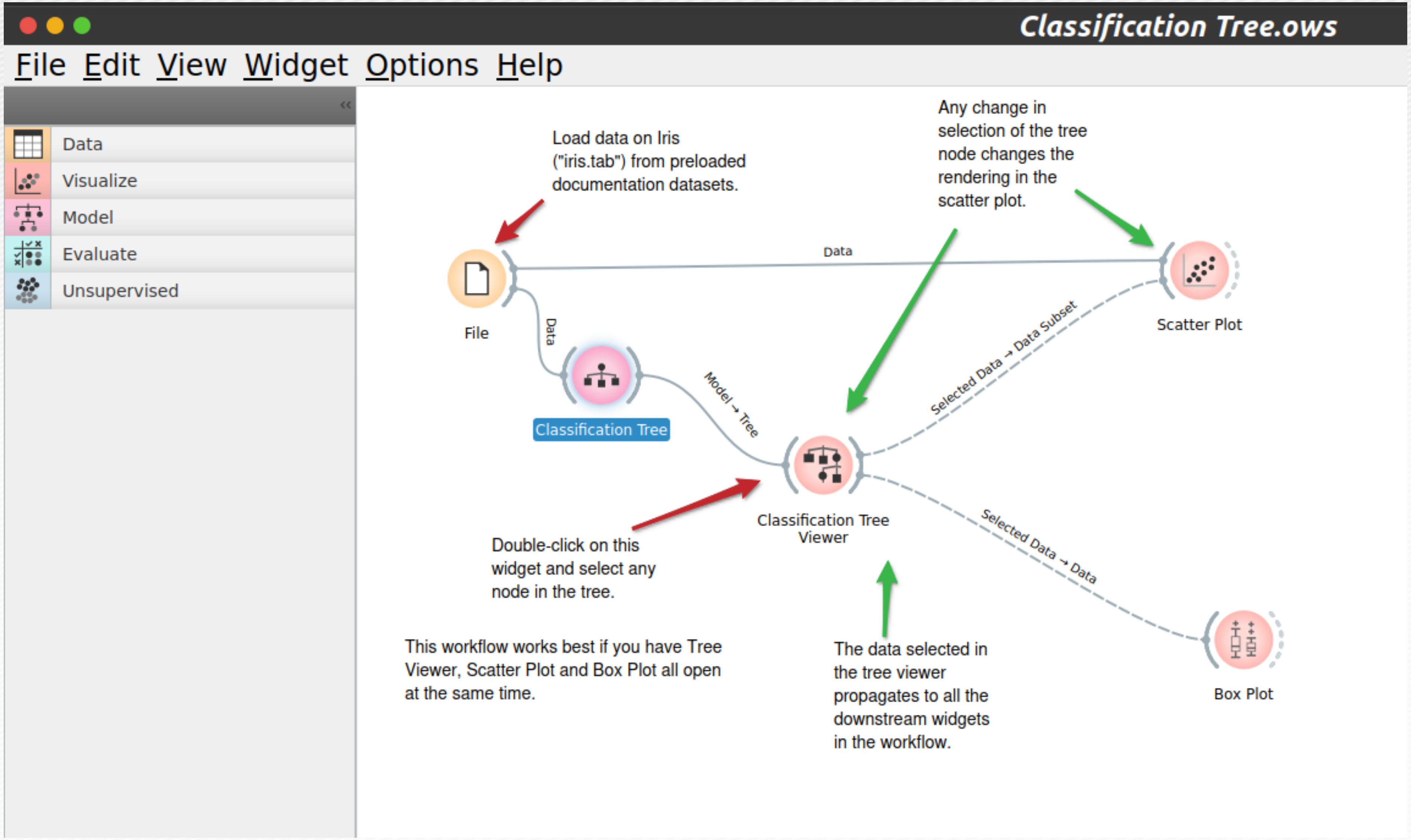

Data
Selected Data → Data
File
Scatter Plot
Data Table
Data Table
Scatter Plot
petal width
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
1
2
3
4
5
6
7
petal length
Iris-setosa
Iris-versicolor
Iris-virginica
150
✓ Portable Network Graphics (*.png)
Scalable Vector Graphics (*.svg)
Portable Document Format (*.pdf)
Portable Document Format (from Matplotlib) (*.pdf)
Python Code (with Matplotlib) (*.py)

Selection subset of the data (2)



	iris	sepal length	sepal width	petal length
47	Iris-setosa	5.1	3.8	1.6
48	Iris-setosa	4.6	3.2	1.4
49	Iris-setosa	5.3	3.7	1.5
50	Iris-setosa	5.0	3.3	1.4
51	Iris-versicolor	7.0	3.2	4.7
52	Iris-versicolor	6.4	3.2	4.5
53	Iris-versicolor	6.9	3.1	4.9
54	Iris-versicolor	5.5	2.3	4.0
55	Iris-versicolor	6.5	2.8	4.6
56	Iris-versicolor	5.7	2.8	4.5
57	Iris-versicolor	6.3	3.3	4.7
58	Iris-versicolor	4.9	2.4	3.3
59	Iris-versicolor	6.6	2.9	4.6

Can we create this workflow example?



Load data on Iris ("iris.tab") from preloaded documentation datasets.

Any change in selection of the tree node changes the rendering in the scatter plot.

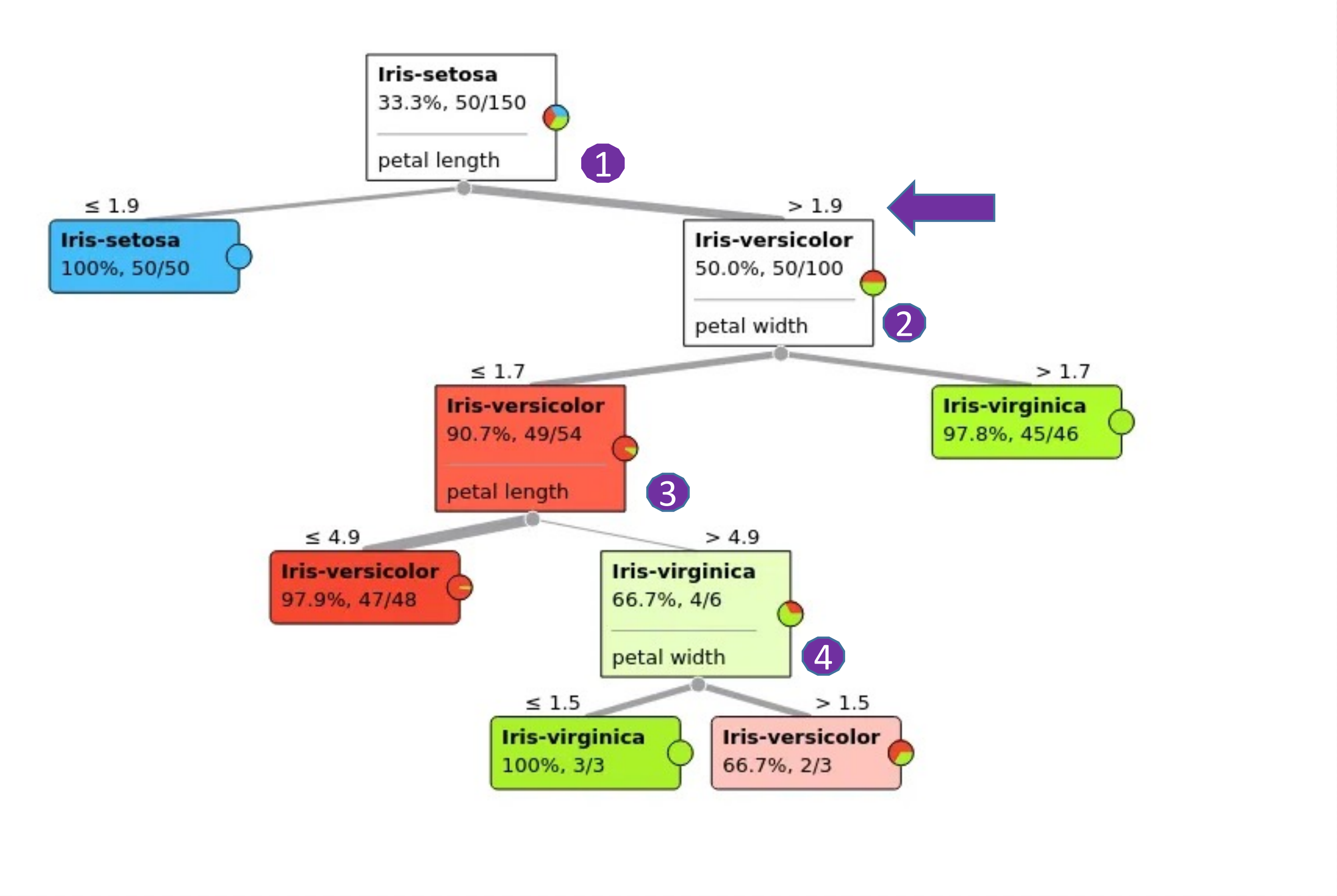
Double-click on this widget and select any node in the tree.

This workflow works best if you have Tree Viewer, Scatter Plot and Box Plot all open at the same time.

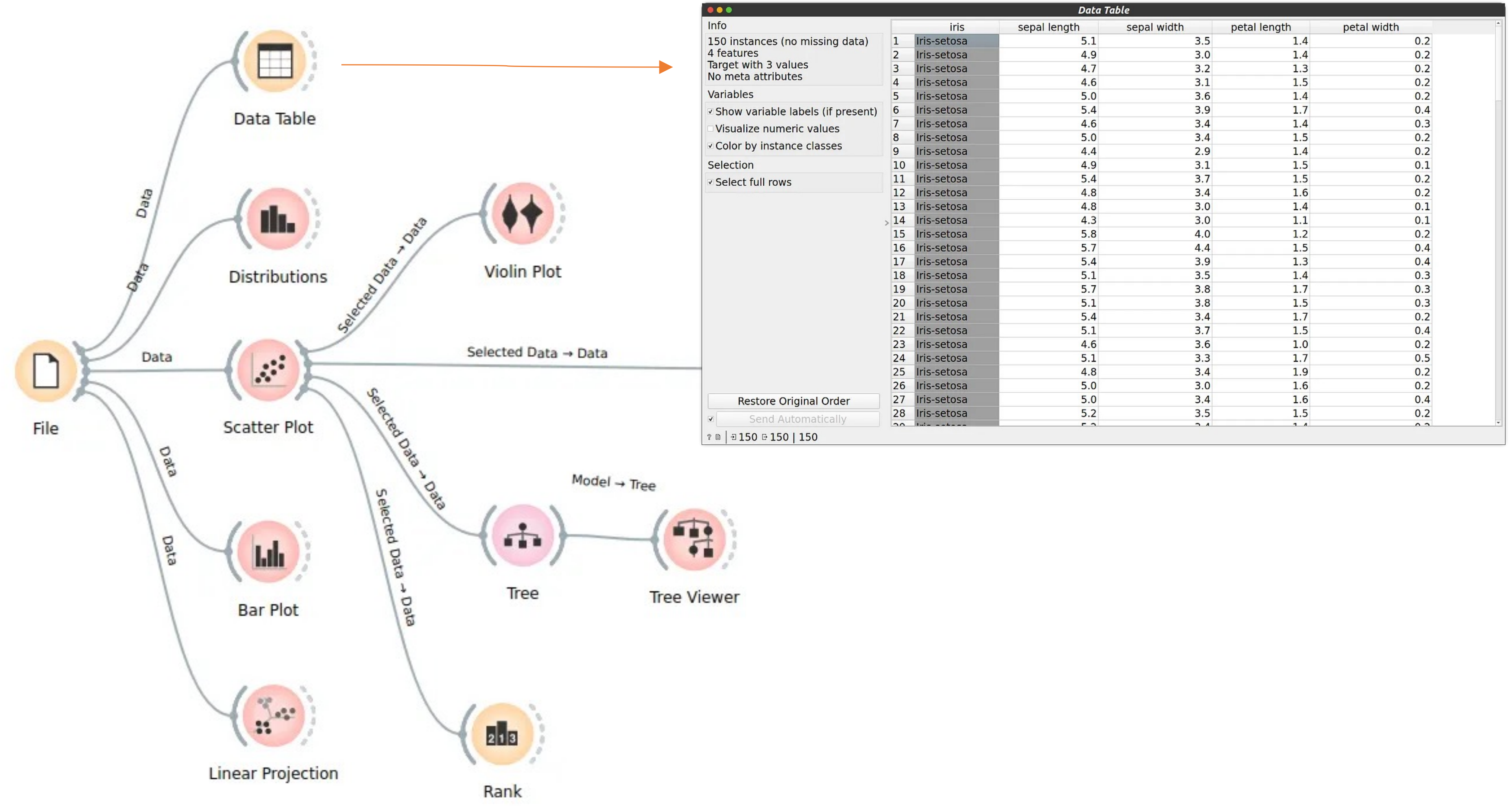
The data selected in the tree viewer propagates to all the downstream widgets in the workflow.

Source: https://medium.datadriveninvestor.com/2-data-science-data-preprocessing-using-scikit-learn-35e0cab65d13

How to read the output of classification tree



Data Visualization – Data Table



Info

150 instances (no missing data)
4 features
Target with 3 values
No meta attributes

Variables

☑ Show variable labels (if present)
☐ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

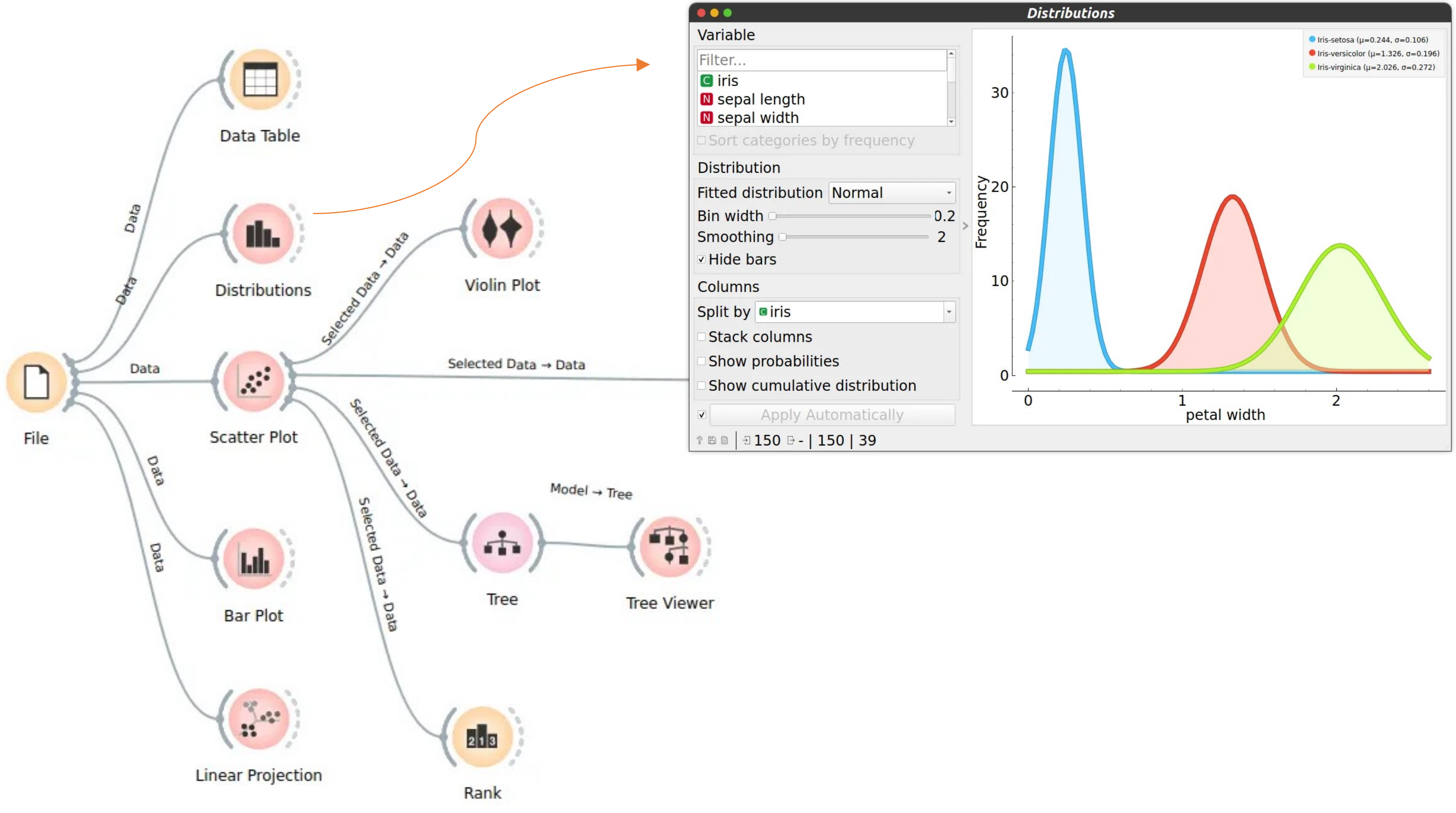
	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2
6	Iris-setosa	5.4	3.9	1.7	0.4
7	Iris-setosa	4.6	3.4	1.4	0.3
8	Iris-setosa	5.0	3.4	1.5	0.2
9	Iris-setosa	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
24	Iris-setosa	5.1	3.3	1.7	0.5
25	Iris-setosa	4.8	3.4	1.9	0.2
26	Iris-setosa	5.0	3.0	1.6	0.2
27	Iris-setosa	5.0	3.4	1.6	0.4
28	Iris-setosa	5.2	3.5	1.5	0.2

150 150 | 150

Data Visualization

Feature Enginnering

Data Visualization – Distribution



Variable

Filter...

- C iris
- N sepal length
- N sepal width

☐ Sort categories by frequency

Distribution

Fitted distribution: Normal

Bin width: 0.2

Smoothing: 2

☑ Hide bars

Columns

Split by: iris

☐ Stack columns

☐ Show probabilities

☐ Show cumulative distribution

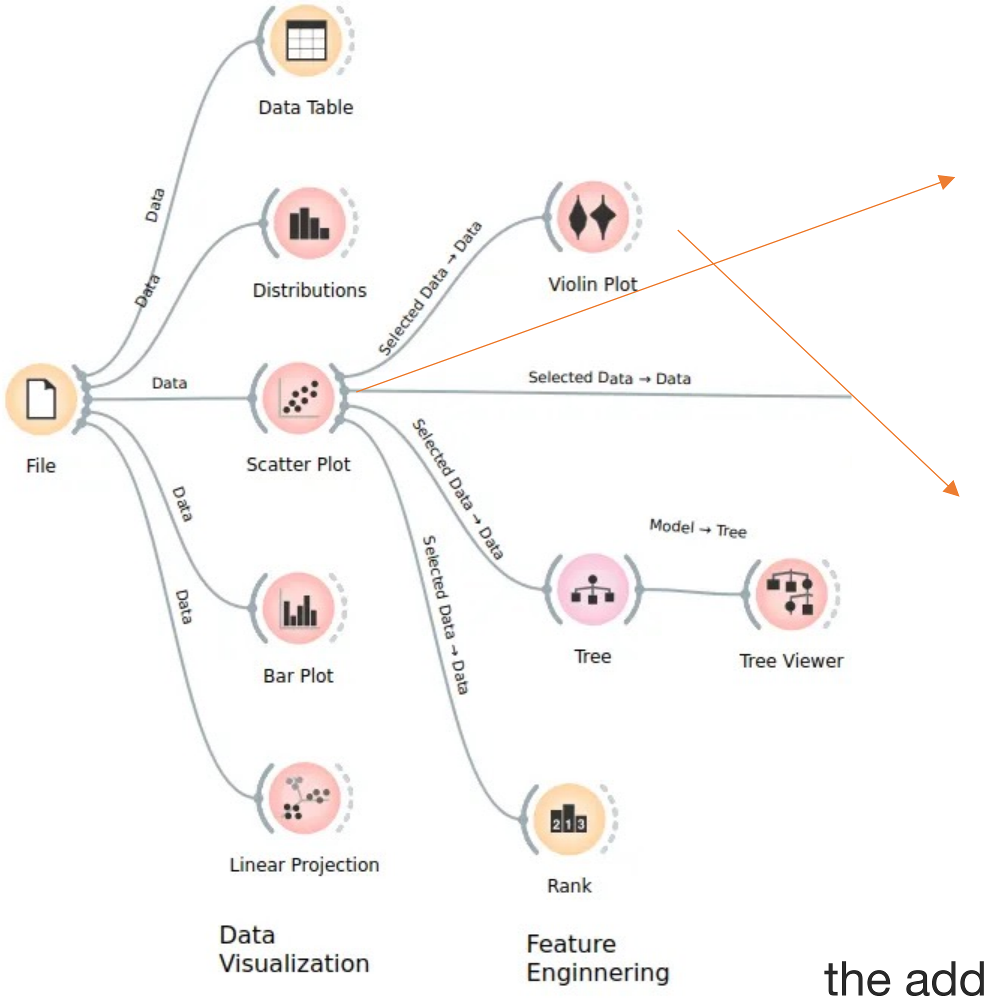
☑ Apply Automatically

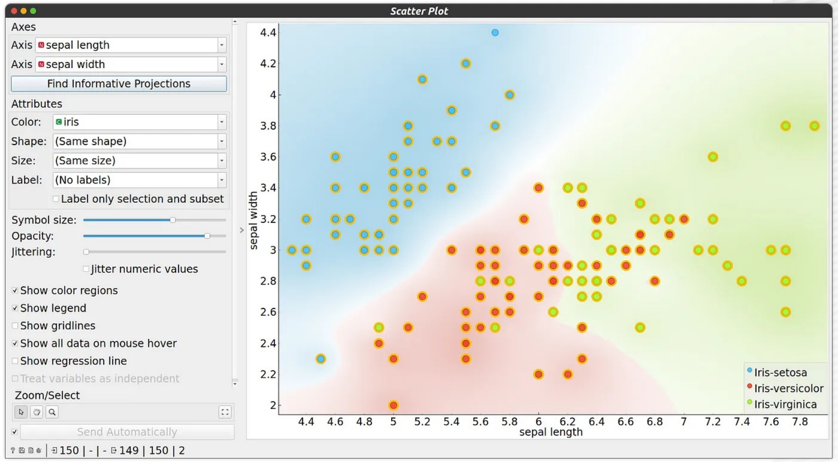
150 | - | 150 | 39

Data
Visualization

Feature
Enginnering

Data Visualization – Scatter & Violin Plots




Scatter Plot

Axes
Axis: sepal length
Axis: sepal width
Find Informative Projections

Attributes
Color: iris
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
☐ Label only selection and subset

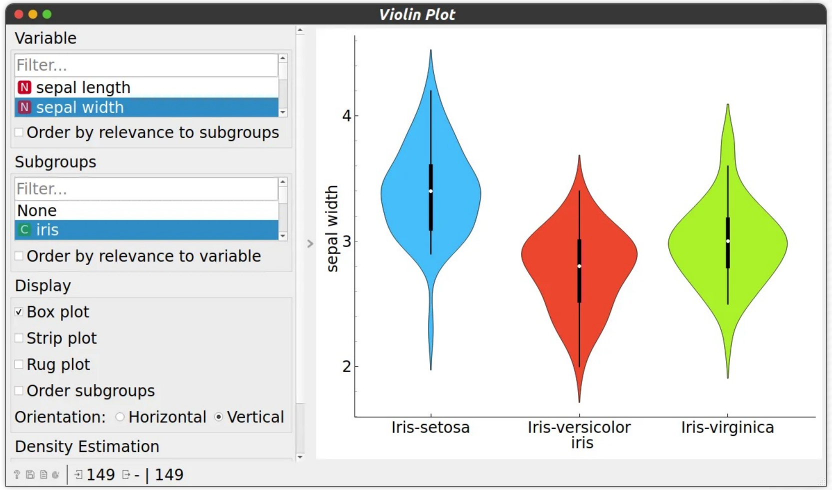
Symbol size:
Opacity:
Jittering:
☐ Jitter numeric values

☑ Show color regions
☑ Show legend
☐ Show gridlines
☑ Show all data on mouse hover
☐ Show regression line
☐ Treat variables as independent

Zoom/Select

☑ Send Automatically

150 | - | - 149 | 150 | 2


Violin Plot

Variable
Filter...
N sepal length
N sepal width
☐ Order by relevance to subgroups

Subgroups
Filter...
None
C iris
☐ Order by relevance to variable

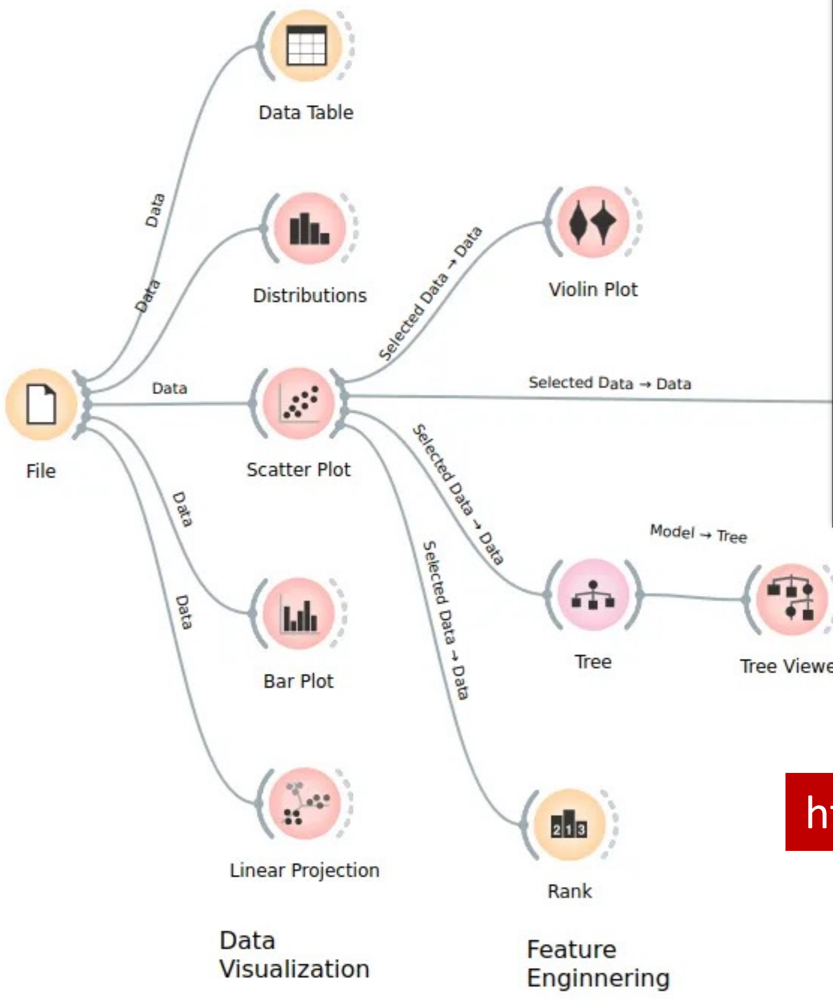
Display
☑ Box plot
☐ Strip plot
☐ Rug plot
☐ Order subgroups
Orientation: ○ Horizontal ● Vertical

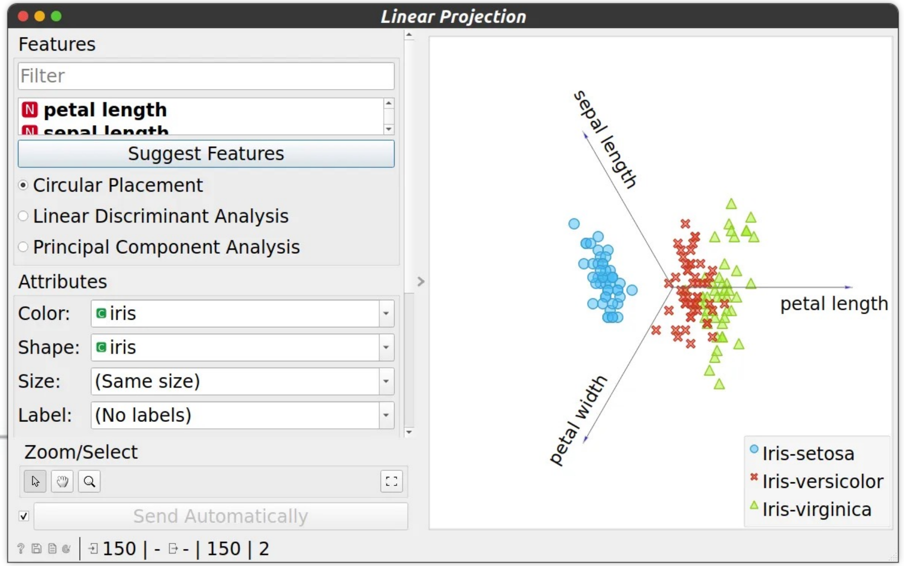
Density Estimation

149 - | 149

the addition of a rotated **kernel density plot** on each side

Data Visualization – Linear Projection





Features

Filter

N petal length

Suggest Features

● Circular Placement
○ Linear Discriminant Analysis
○ Principal Component Analysis

Attributes

Color: iris
Shape: iris
Size: (Same size)
Label: (No labels)

Zoom/Select

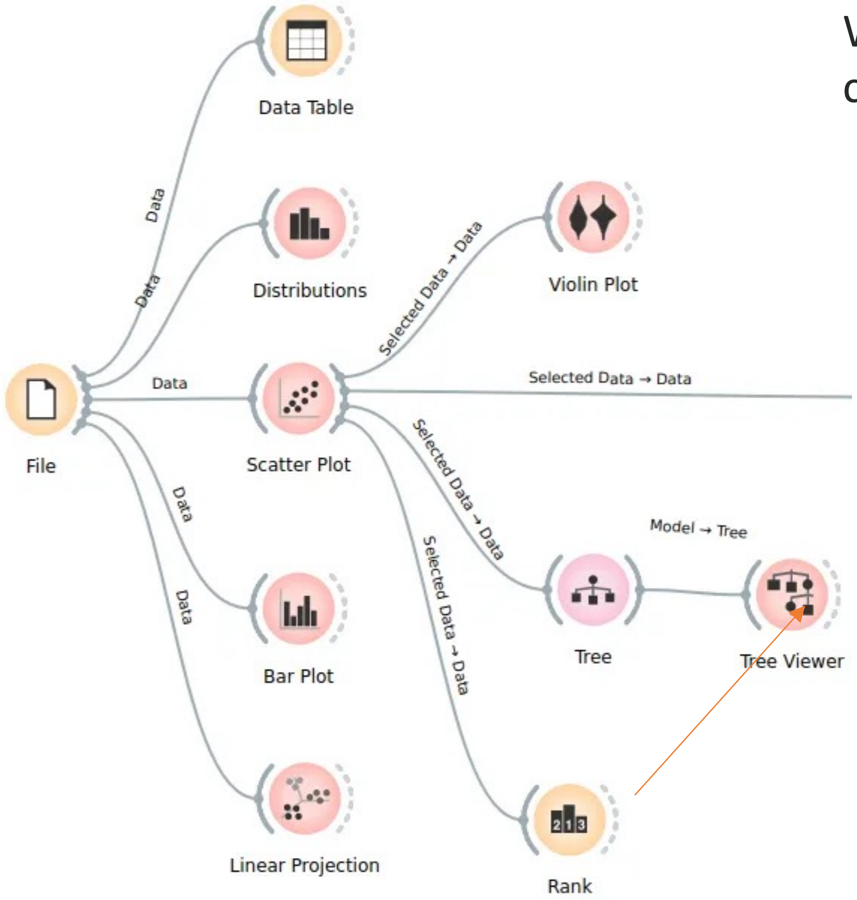
☑ Send Automatically

150 | - - | 150 | 2

visualize the data up to 3D

https://orangedatamining.com/widget-catalog/visualize/linearprojection/

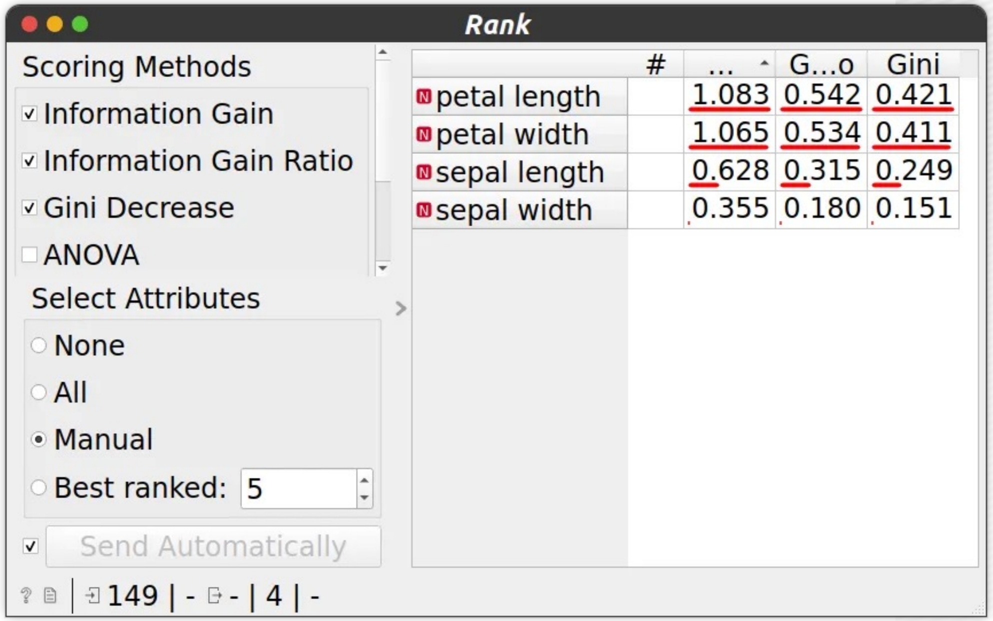
Data Visualization



Data Visualization

Feature Enginnering

Which features are most important for classification



Rank

Scoring Methods

- ☑ Information Gain
- ☑ Information Gain Ratio
- ☑ Gini Decrease
- ☐ ANOVA

Select Attributes

- ○ None
- ○ All
- ● Manual
- ○ Best ranked: 5

☑ Send Automatically

	#	...	G...o	Gini
N petal length		1.083	0.542	0.421
N petal width		1.065	0.534	0.411
N sepal length		0.628	0.315	0.249
N sepal width		0.355	0.180	0.151

149 | - | - | 4 | -

If you are interested in what widgets are available, click the blue circle above.

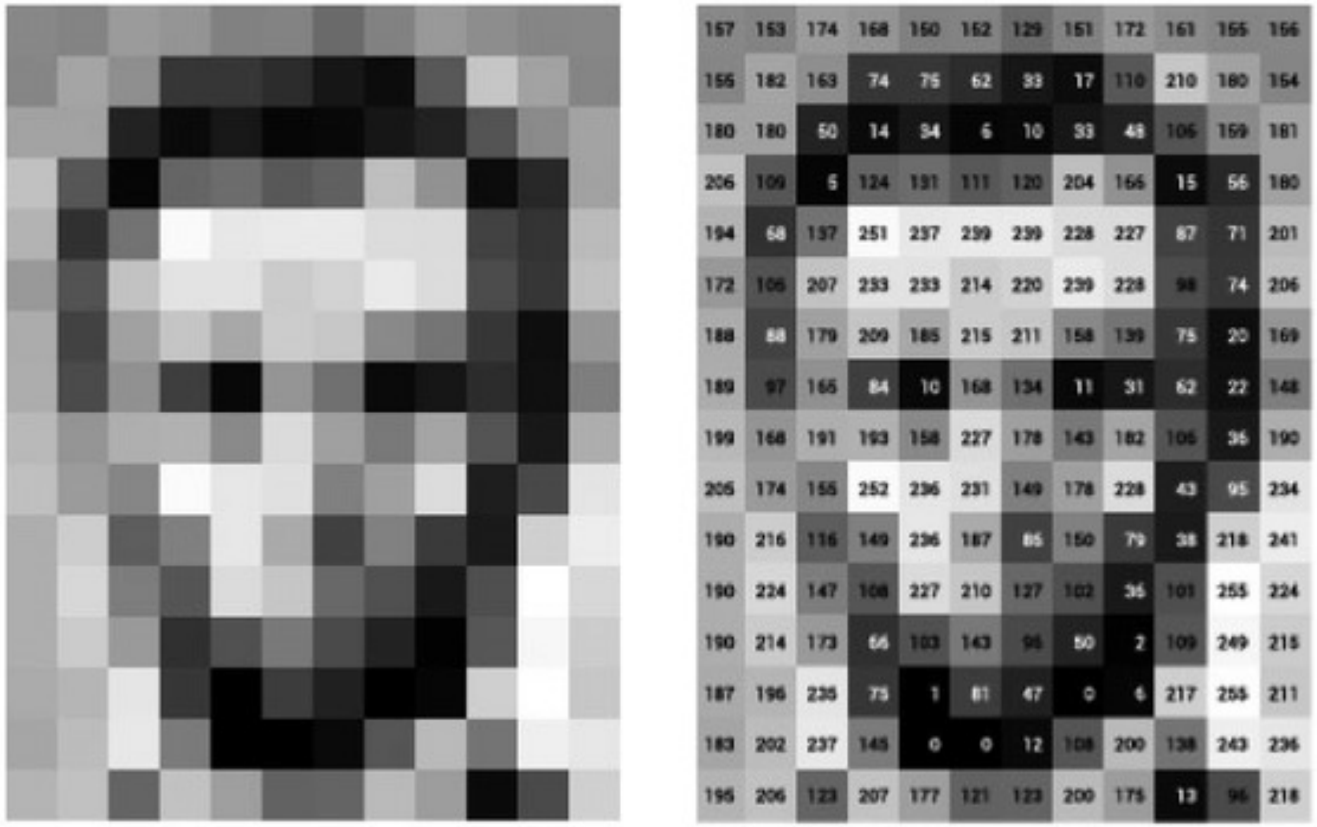
https://orangedatamining.com/widget-catalog/

Part 4
Word Cloud

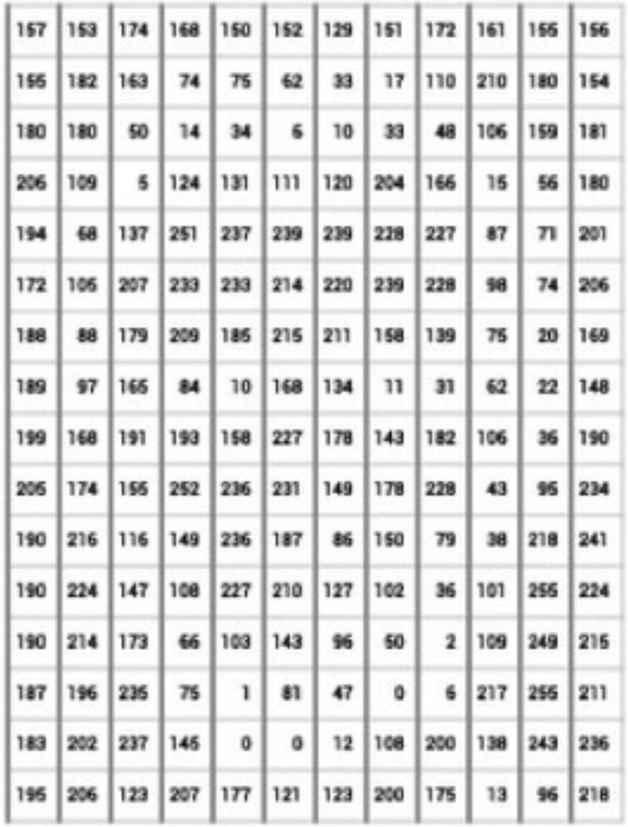
2:00–2:20

Image as numbers

- 흑백 그림은 픽셀의 밝기 값을 0-255 사이의 값으로 표현한 이미지



pixel



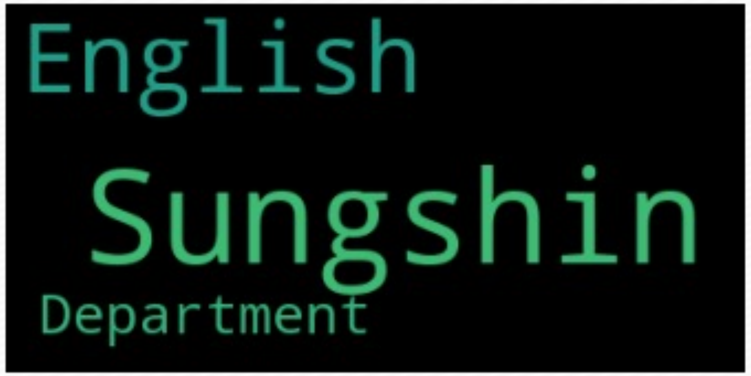
157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

https://www.researchgate.net/figure/mage-of-Abraham-Lincoln-as-a-matrix-of-pixel-values_fig1_330902210

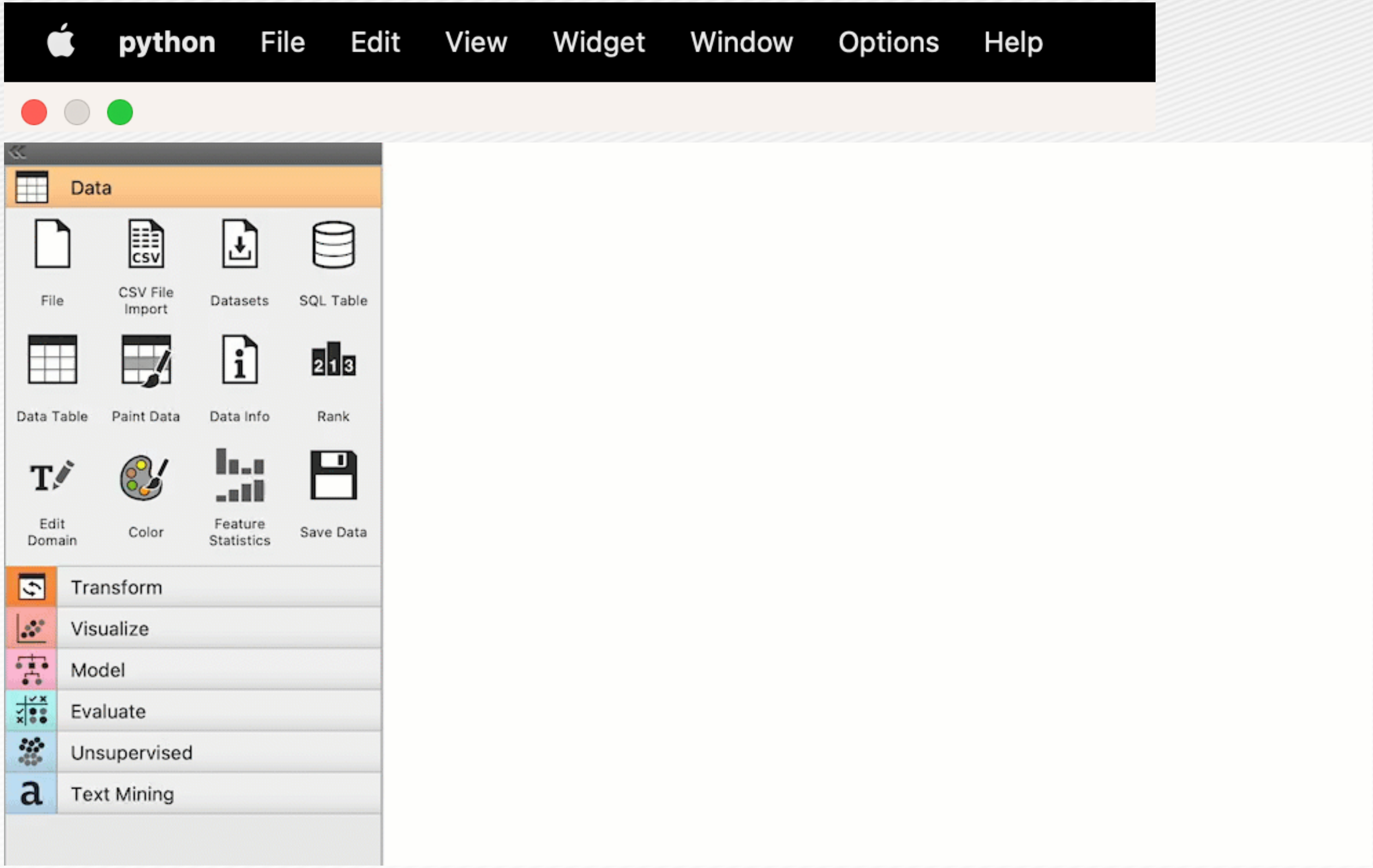
Making Word Cloud using CLI

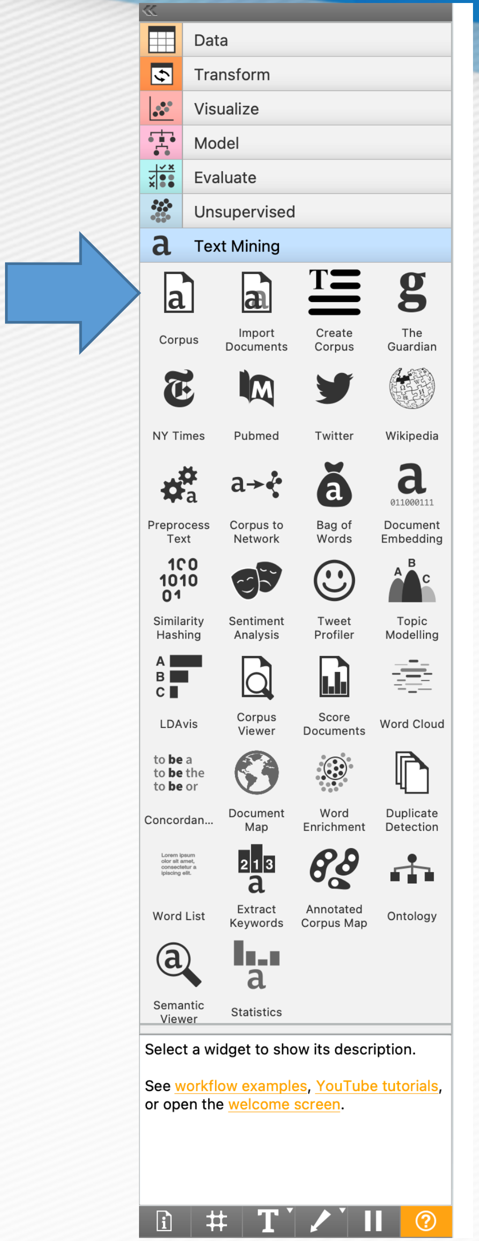
```
from wordcloud import WordCloud
text = "Sungshin Sungshin Sungshin Sungshin \
Sungshin Sungshin Sungshin Sungshin Sungshin Sungshin \
English English English English English Department"
wc = WordCloud().generate(text)
print(dir(wc))
```

```
import matplotlib.pyplot as plt
plt.figure()
plt.imshow(wc)
plt.axis(“off”)
plt.show()
#plt.savefig(“wc_sungshin.png”)
```



Options > Add-ons… > Text





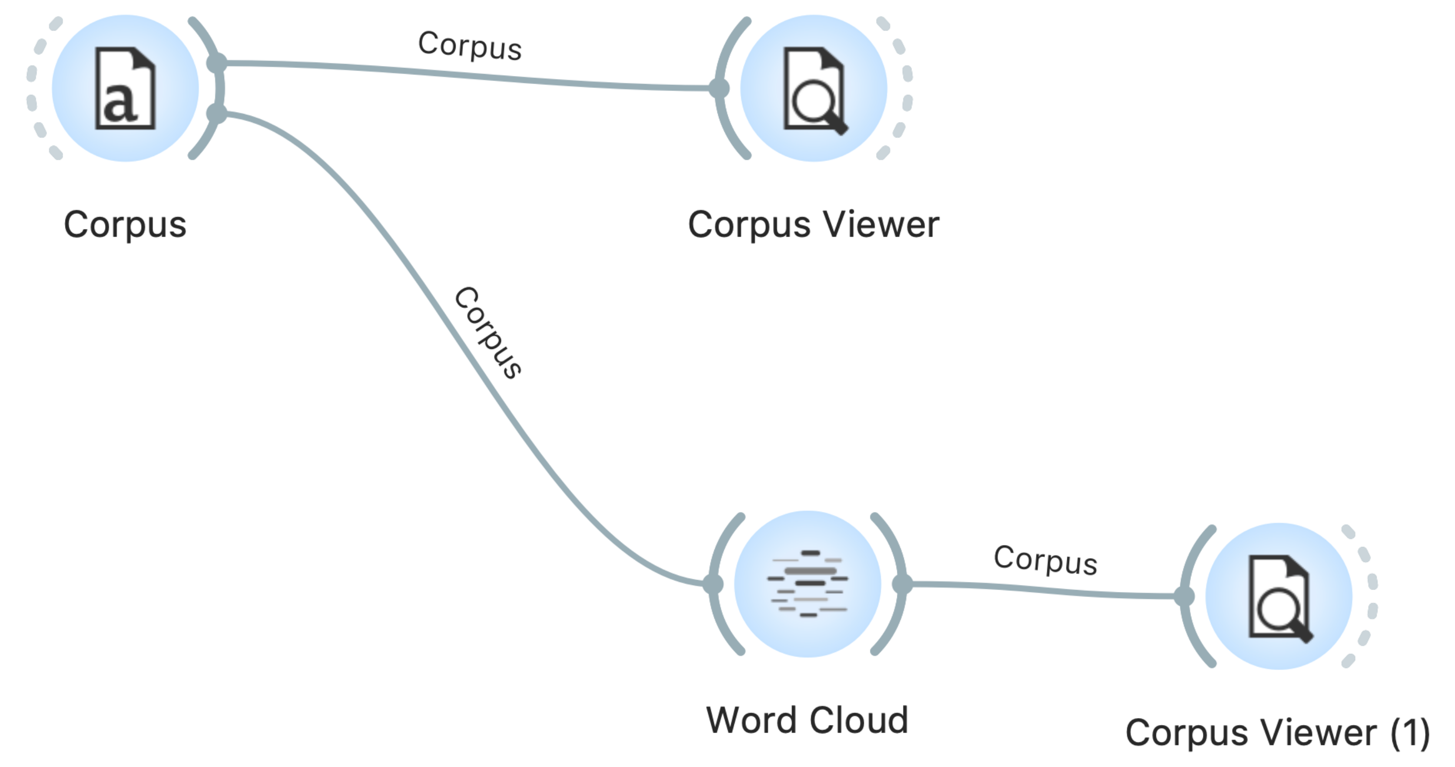
Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.

Corpus is any collection of documents.

Workflow for Word Cloud

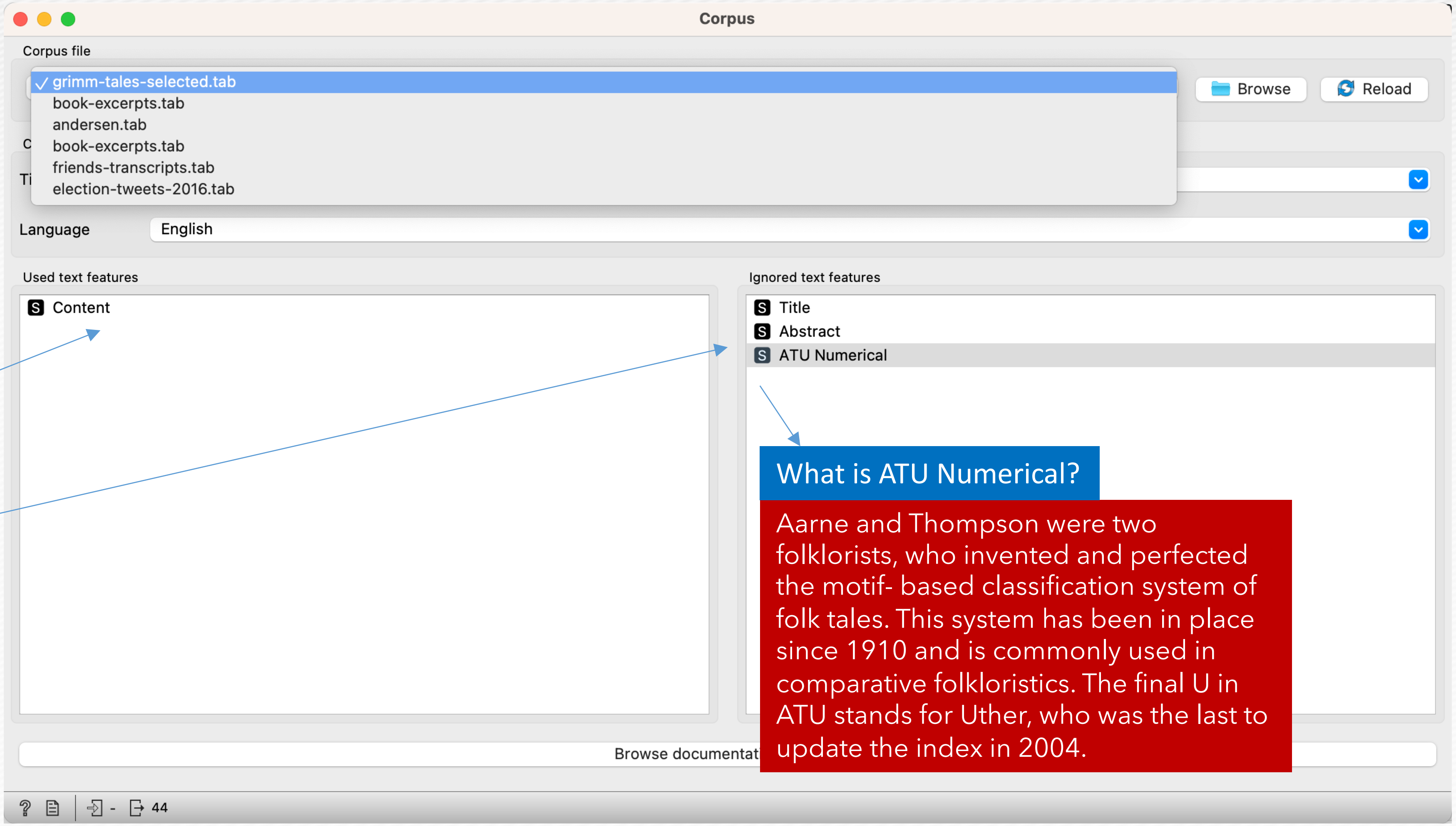
- Start by constructing a workflow that consists of a Corpus widget, a Word Cloud widget and two Corpus Viewer widgets:



Preloaded dataset

- From “Browse documentation data sets...” choose *Grimm-tales-selected.tab*, a data set containing Grimm’s selected tales.

  - The particularity of the Corpus widget is that it sets the text feature(s) to apply text mining on.
  - “Used text features” defines the content (text), while other columns contain meta attributes (title, abstract, etc.).



Corpus

Corpus file

✓ grimm-tales-selected.tab
book-excerpts.tab
andersen.tab
book-excerpts.tab
friends-transcripts.tab
election-tweets-2016.tab

Language: English

Used text features
- Content

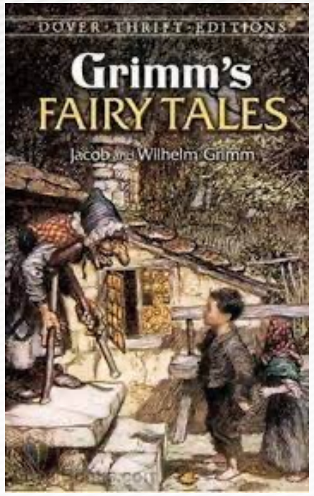
Ignored text features
- Title
- Abstract
- ATU Numerical

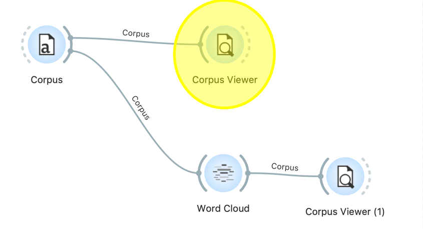
What is ATU Numerical?

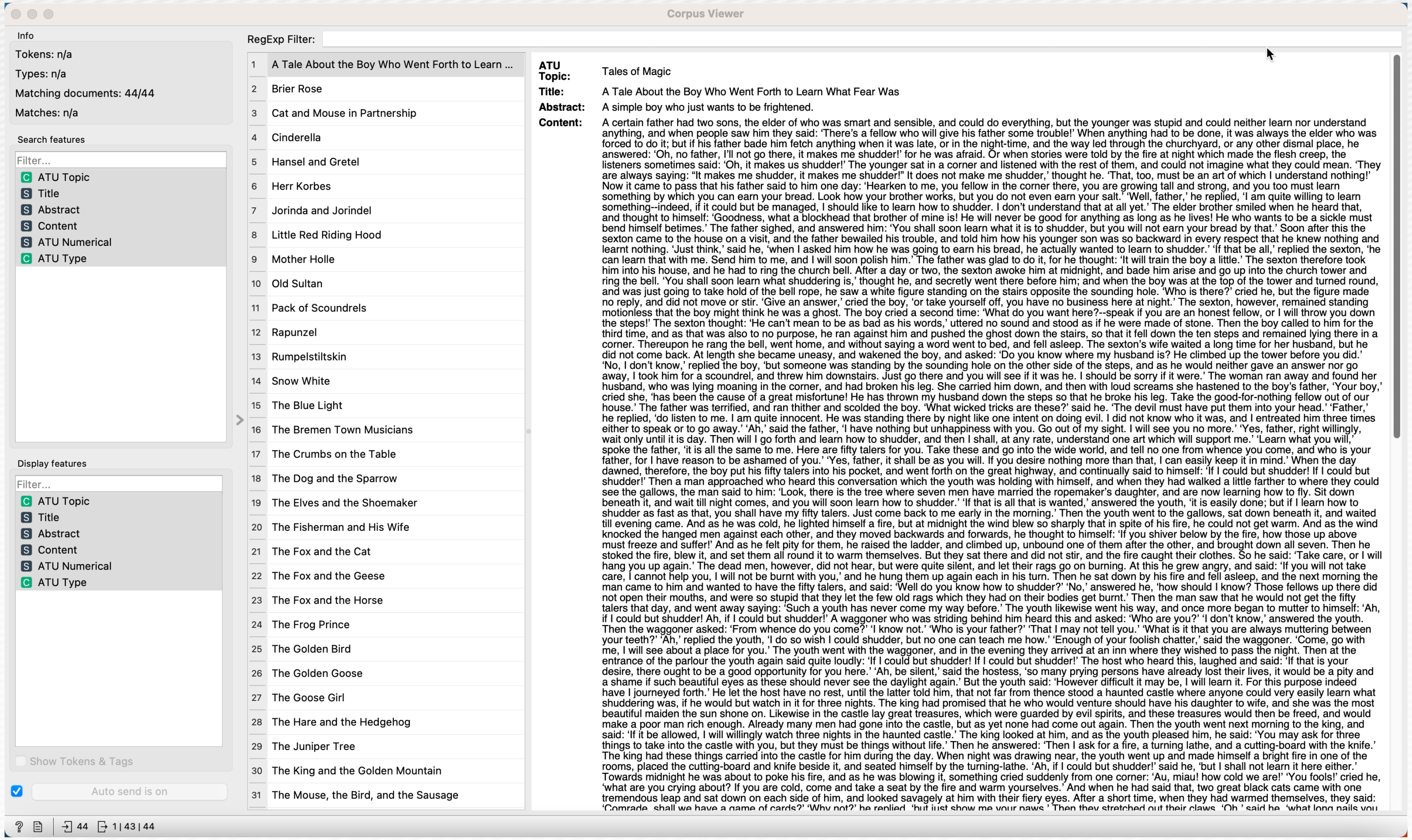
Aarne and Thompson were two folklorists, who invented and perfected the motif- based classification system of folk tales. This system has been in place since 1910 and is commonly used in comparative folkloristics. The final U in ATU stands for Uther, who was the last to update the index in 2004.

Data: Grimm-tales-selected (44 selected Grimm's tales)

- preloaded dataset







Info
Tokens: n/a
Types: n/a
Matching documents: 44/44
Matches: n/a

Search features
Filter...
- C ATU Topic
- S Title
- S Abstract
- S Content
- S ATU Numerical
- C ATU Type

Display features
Filter...
- C ATU Topic
- S Title
- S Abstract
- S Content
- S ATU Numerical
- C ATU Type

☐ Show Tokens & Tags
☑ Auto send is on

RegExp Filter:

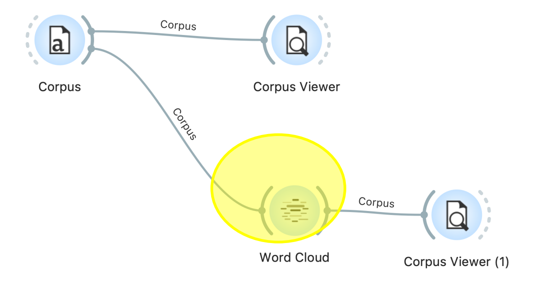
1 A Tale About the Boy Who Went Forth to Learn ...
2 Brier Rose
3 Cat and Mouse in Partnership
4 Cinderella
5 Hansel and Gretel
6 Herr Korbes
7 Jorinda and Jorindel
8 Little Red Riding Hood
9 Mother Holle
10 Old Sultan
11 Pack of Scoundrels
12 Rapunzel
13 Rumpelstiltskin
14 Snow White
15 The Blue Light
16 The Bremen Town Musicians
17 The Crumbs on the Table
18 The Dog and the Sparrow
19 The Elves and the Shoemaker
20 The Fisherman and His Wife
21 The Fox and the Cat
22 The Fox and the Geese
23 The Fox and the Horse
24 The Frog Prince
25 The Golden Bird
26 The Golden Goose
27 The Goose Girl
28 The Hare and the Hedgehog
29 The Juniper Tree
30 The King and the Golden Mountain
31 The Mouse, the Bird, and the Sausage

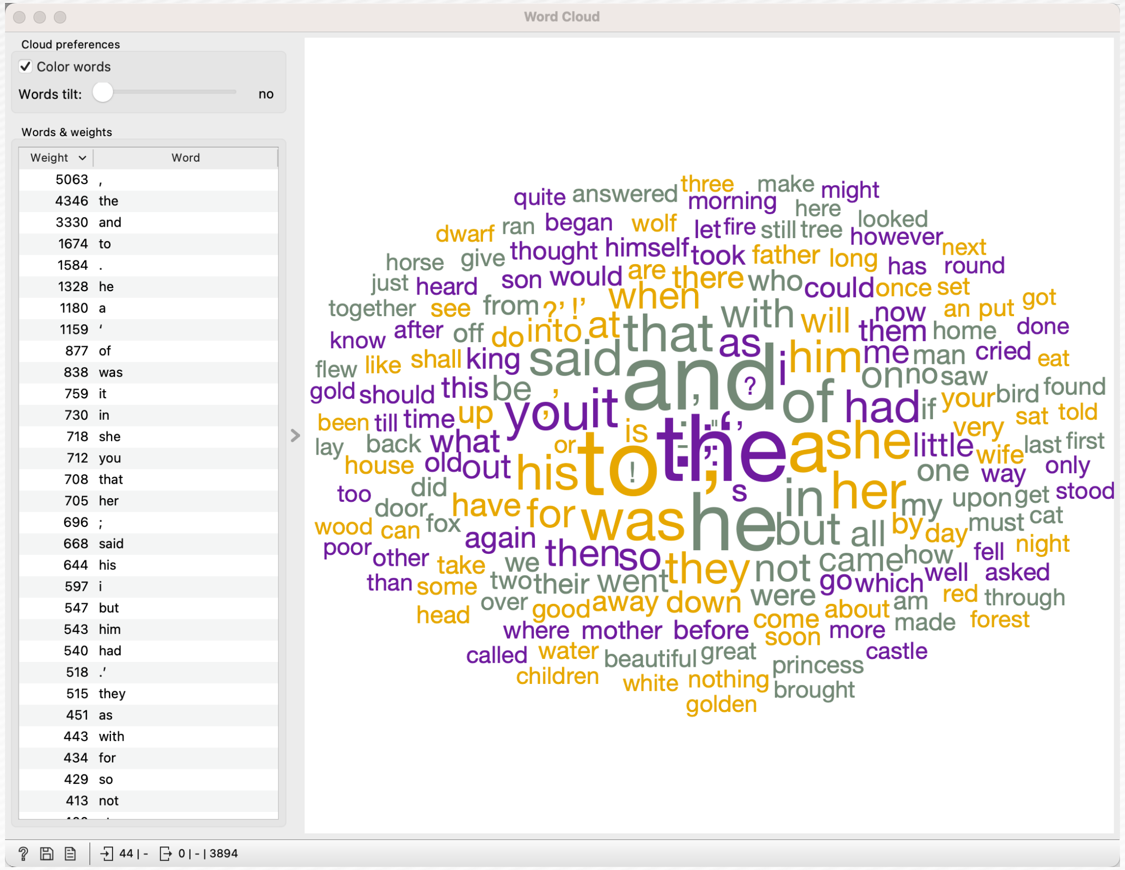
ATU Topic: Tales of Magic
Title: A Tale About the Boy Who Went Forth to Learn What Fear Was
Abstract: A simple boy who just wants to be frightened.
Content: A certain father had two sons, the elder of who was smart and sensible, and could do everything, but the younger was stupid and could neither learn nor understand anything, and when people saw him they said: 'There's a fellow who will give his father some trouble!' When anything had to be done, it was always the elder who was forced to do it; but if his father bade him fetch anything when it was late, or in the night-time, and the way led through the churchyard, or any other dismal place, he answered: 'Oh, no father, I'll not go there, it makes me shudder!' for he was afraid. Or when stories were told by the fire at night which made the flesh creep, the listeners sometimes said: 'Oh, it makes us shudder!' The younger sat in a corner and listened with the rest of them, and could not imagine what they could mean. 'They are always saying: "It makes me shudder, it makes me shudder!" It does not make me shudder,' thought he. 'That, too, must be an art of which I understand nothing!' Now it came to pass that his father said to him one day: 'Hearken to me, you fellow in the corner there, you are growing tall and strong, and you too must learn something by which you can earn your bread. Look how your brother works, but you do not even earn your salt.' 'Well, father,' he replied, 'I am quite willing to learn something--indeed, if it could but be managed, I should like to learn how to shudder. I don't understand that at all yet.' The elder brother smiled when he heard that, and thought to himself: 'Goodness, what a blockhead that brother of mine is! He will never be good for anything as long as he lives! He who wants to be a sickle must bend himself betimes.' The father sighed, and answered him: 'You shall soon learn what it is to shudder, but you will not earn your bread by that.' Soon after this the sexton came to the house on a visit, and the father bewailed his trouble, and told him how his younger son was so backward in every respect that he knew nothing and learnt nothing. 'Just think,' said he, 'when I asked him how he was going to earn his bread, he actually wanted to learn to shudder.' 'If that be all,' replied the sexton, 'he can learn that with me. Send him to me, and I will soon polish him.' The father was glad to do it, for he thought: 'It will train the boy a little.' The sexton therefore took him into his house, and he had to ring the church bell. After a day or two, the sexton awoke him at midnight, and bade him arise and go up into the church tower and ring the bell. 'You shall soon learn what shuddering is,' thought he, and secretly went there before him; and when the boy was at the top of the tower and turned round, and was just going to take hold of the bell rope, he saw a white figure standing on the stairs opposite the sounding hole. 'Who is there?' cried he, but the figure made no reply, and did not move or stir. 'Give an answer,' cried the boy, 'or take yourself off, you have no business here at night.' The sexton, however, remained standing motionless that the boy might think he was a ghost. The boy cried a second time: 'What do you want here?--speak if you are an honest fellow, or I will throw you down the steps!' The sexton thought: 'He can't mean to be as bad as his words,' uttered no sound and stood as if he were made of stone. Then the boy called to him for the third time, and as that was also to no purpose, he ran against him and pushed the ghost down the stairs, so that it fell down the ten steps and remained lying there in a corner. Thereupon he rang the bell, went home, and without saying a word went to bed, and fell asleep. The sexton's wife waited a long time for her husband, but he did not come back. At length she became uneasy, and wakened the boy, and asked: 'Do you know where my husband is? He climbed up the tower before you did.' 'No, I don't know,' replied the boy, 'but someone was standing by the sounding hole on the other side of the steps, and as he would neither gave an answer nor go away, I took him for a scoundrel, and threw him downstairs. Just go there and you will see if it was he. I should be sorry if it were.' The woman ran away and found her husband, who was lying moaning in the corner, and had broken his leg. She carried him down, and then with loud screams she hastened to the boy's father, 'Your boy,' cried she, 'has been the cause of a great misfortune! He has thrown my husband down the steps so that he broke his leg. Take the good-for-nothing fellow out of our house.' The father was terrified, and ran thither and scolded the boy. 'What wicked tricks are these?' said he. 'The devil must have put them into your head.' 'Father,' he replied, 'do listen to me. I am quite innocent. He was standing there by night like one intent on doing evil. I did not know who it was, and I entreated him three times either to speak or to go away.' 'Ah,' said the father, 'I have nothing but unhappiness with you. Go out of my sight. I will see you no more.' 'Yes, father, right willingly, wait only until it is day. Then will I go forth and learn how to shudder, and then I shall, at any rate, understand one art which will support me.' 'Learn what you will,' spoke the father, 'it is all the same to me. Here are fifty talers for you. Take these and go into the wide world, and tell no one from whence you come, and who is your father, for I have reason to be ashamed of you.' 'Yes, father, it shall be as you will. If you desire nothing more than that, I can easily keep it in mind.' When the day dawned, therefore, the boy put his fifty talers into his pocket, and went forth on the great highway, and continually said to himself: 'If I could but shudder! If I could but shudder!' Then a man approached who heard this conversation which the youth was holding with himself, and when they had walked a little farther to where they could see the gallows, the man said to him: 'Look, there is the tree where seven men have married the ropemaker's daughter, and are now learning how to fly. Sit down beneath it, and wait till night comes, and you will soon learn how to shudder.' 'If that is all that is wanted,' answered the youth, 'it is easily done; but if I learn how to shudder as fast as that, you shall have my fifty talers. Just come back to me early in the morning.' Then the youth went to the gallows, sat down beneath it, and waited till evening came. And as he was cold, he lighted himself a fire, but at midnight the wind blew so sharply that in spite of his fire, he could not get warm. And as the wind knocked the hanged men against each other, and they moved backwards and forwards, he thought to himself: 'If you shiver below by the fire, how those up above must freeze and suffer!' And as he felt pity for them, he raised the ladder, and climbed up, unbound one of them after the other, and brought down all seven. Then he stoked the fire, blew it, and set them all round it to warm themselves. But they sat there and did not stir, and the fire caught their clothes. So he said: 'Take care, or I will hang you up again.' The dead men, however, did not hear, but were quite silent, and let their rags go on burning. At this he grew angry, and said: 'If you will not take care, I cannot help you, I will not be burnt with you,' and he hung them up again each in his turn. Then he sat down by his fire and fell asleep, and the next morning the man came to him and wanted to have the fifty talers, and said: 'Well do you know how to shudder?' 'No,' answered he, 'how should I know? Those fellows up there did not open their mouths, and were so stupid that they let the few old rags which they had on their bodies get burnt.' Then the man saw that he would not get the fifty talers that day, and went away saying: 'Such a youth has never come my way before.' The youth likewise went his way, and once more began to mutter to himself: 'Ah, if I could but shudder! Ah, if I could but shudder!' A waggoner who was striding behind him heard this and asked: 'Who are you?' 'I don't know,' answered the youth. Then the waggoner asked: 'From whence do you come?' 'I know not.' 'Who is your father?' 'That I may not tell you.' 'What is it that you are always muttering between your teeth?' 'Ah,' replied the youth, 'I do so wish I could shudder, but no one can teach me how.' 'Enough of your foolish chatter,' said the waggoner. 'Come, go with me, I will see about a place for you.' The youth went with the waggoner, and in the evening they arrived at an inn where they wished to pass the night. Then at the entrance of the parlour the youth again said quite loudly: 'If I could but shudder! If I could but shudder!' The host who heard this, laughed and said: 'If that is your desire, there ought to be a good opportunity for you here.' 'Ah, be silent,' said the hostess, 'so many prying persons have already lost their lives, it would be a pity and a shame if such beautiful eyes as these should never see the daylight again.' But the youth said: 'However difficult it may be, I will learn it. For this purpose indeed have I journeyed forth.' He let the host have no rest, until the latter told him, that not far from thence stood a haunted castle where anyone could very easily learn what shuddering was, if he would but watch in it for three nights. The king had promised that he who would venture should have his daughter to wife, and she was the most beautiful maiden the sun shone on. Likewise in the castle lay great treasures, which were guarded by evil spirits, and these treasures would then be freed, and would make a poor man rich enough. Already many men had gone into the castle, but as yet none had come out again. Then the youth went next morning to the king, and said: 'If it be allowed, I will willingly watch three nights in the haunted castle.' The king looked at him, and as the youth pleased him, he said: 'You may ask for three things to take into the castle with you, but they must be things without life.' Then he answered: 'Then I ask for a fire, a turning lathe, and a cutting-board with the knife.' The king had these things carried into the castle for him during the day. When night was drawing near, the youth went up and made himself a bright fire in one of the rooms, placed the cutting-board and knife beside it, and seated himself by the turning-lathe. 'Ah, if I could but shudder!' said he, 'but I shall not learn it here either.' Towards midnight he was about to poke his fire, and as he was blowing it, something cried suddenly from one corner: 'Au, miau! how cold we are!' 'You fools!' cried he, 'what are you crying about? If you are cold, come and take a seat by the fire and warm yourselves.' And when he had said that, two great black cats came with one tremendous leap and sat down on each side of him, and looked savagely at him with their fiery eyes. After a short time, when they had warmed themselves, they said: 'Comrade, shall we have a game of cards?' 'Why not?' he replied, 'but just show me your paws.' Then they stretched out their claws. 'Oh,' said he, 'what long nails you

https://en.wikipedia.org/wiki/Grimms%27_Fairy_Tales

Word Cloud via Orange

- Open Word Cloud.
- Word Cloud displays word frequencies, where the more frequent the word, the larger the font.





Cloud preferences

☑ Color words

Words tilt: no

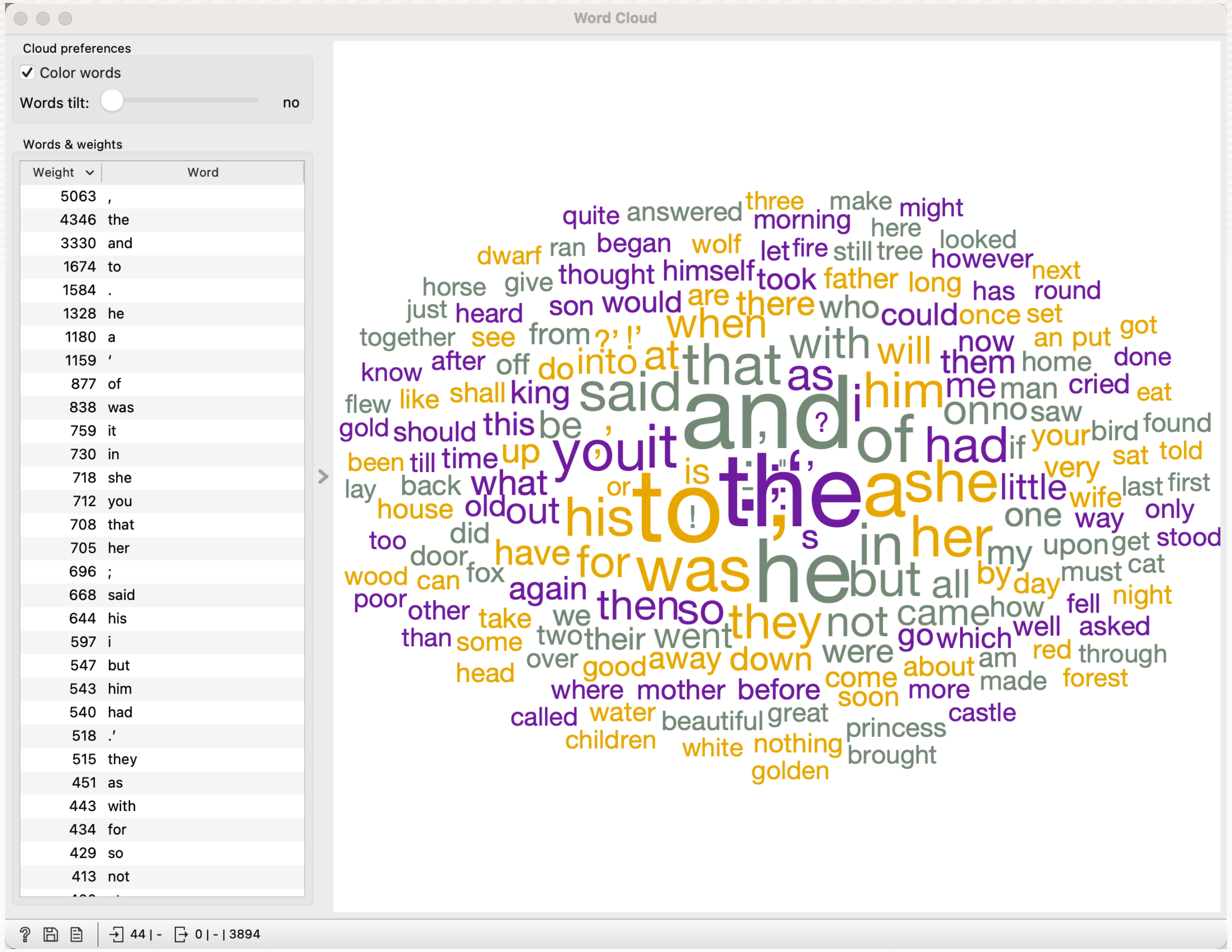
Words & weights

Weight	Word
5063	,
4346	the
3330	and
1674	to
1584	.
1328	he
1180	a
1159	‘
877	of
838	was
759	it
730	in
718	she
712	you
708	that
705	her
696	;
668	said
644	his
597	i
547	but
543	him
540	had
518	.’
515	they
451	as
443	with
434	for
429	so
413	not

quite answered three make might
morning here
dwarf ran began wolf let fire still tree looked however
horse give thought himself took father long has next round
just heard son would are there who could once set
together see from ?’ !’ when
know after off do into at that with will now them an put got home done
flew like shall king said as him me man cried eat
gold should this be and ? of had on no saw if your bird found
been till time up you it very sat told
lay back what or is the ‘ a she little wife last first
house old out his to ; he in her one way only
too did ! s my upon get stood
wood door fox have for was all by day must cat
poor can again but then so came how fell night
other take we two their went they not go which well asked
than some over good away down were about am red through
head where mother before come soon more made forest
called water beautiful great princess castle
children white nothing brought
golden
44 | -
0 | - | 3894

Text cleaning needed!

- This word cloud is a mess! We got a bunch of semantic junk in our visualization. Is there a way to clean this up?


Word Cloud
quite answered three make might
morning here
dwarf ran began wolf let fire still tree looked
however
horse give thought himself took father long has next round
just heard son would are there who could once set
together see from ?' !' when
know after off do into at that with will now an put got
them home done
flew like shall king said as him me man cried eat
gold should this be , and i ? on no saw
been till time up you it of had if your bird found
lay back what is the ; a she little very sat told
house old out his to ! ' s wife last first
too did in her one way only
door have for was he my upon get stood
wood can fox but all by day must cat
poor again then so they not came how fell night
other take we two their went go which well asked
than some over good away down were am red through
head where mother before come about more made forest
soon
called water beautiful great castle
children white nothing princess
golden brought
44 | -
0 | - | 3894

Cloud preferences

☑ Color words

Words tilt: no

Words & weights

Weight	Word
5063	,
4346	the
3330	and
1674	to
1584	.
1328	he
1180	a
1159	‘
877	of
838	was
759	it
730	in
718	she
712	you
708	that
705	her
696	;
668	said
644	his
597	i
547	but
543	him
540	had
518	.’
515	they
451	as
443	with
434	for
429	so
413	not

Preprocessing

Part 5
Text Preprocessing

2:20–2:35

Text Preprocessing

- In the Preprocess Text widget, we decided to transform all words to lowercase, treat each word as a token (and omit punctuation), and to remove the stopwords (such as “in”, “and”, and “the”).
- This preprocessing outputs the following tokens:

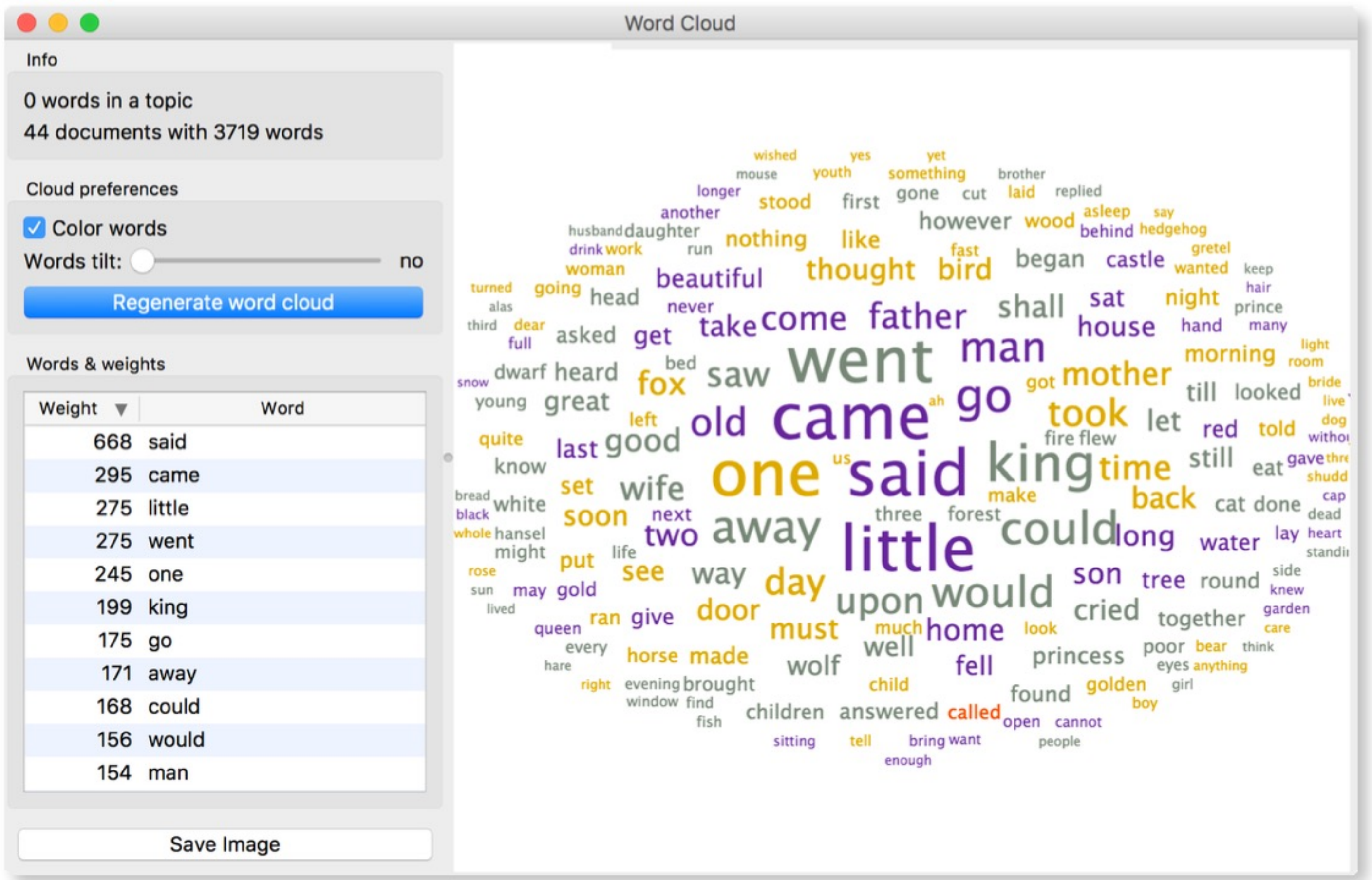
  “This is a sample sentence.” → “sample”, “sentence”

- To see the results of preprocessing, we can display the most frequent tokens in Word Cloud. Word Cloud enables us to identify redundant words and irregularities.

https://orangedatamining.com/blog/2017/06/19/text-preprocessing/

Word Cloud after preprocessing

We see the results of our preprocessing in the Word Cloud. Two of the most frequent words are "would" and "could". If we decide these two words are not important for our analysis, it would be good to omit them.



Word Cloud

Info

0 words in a topic
44 documents with 3719 words

Cloud preferences

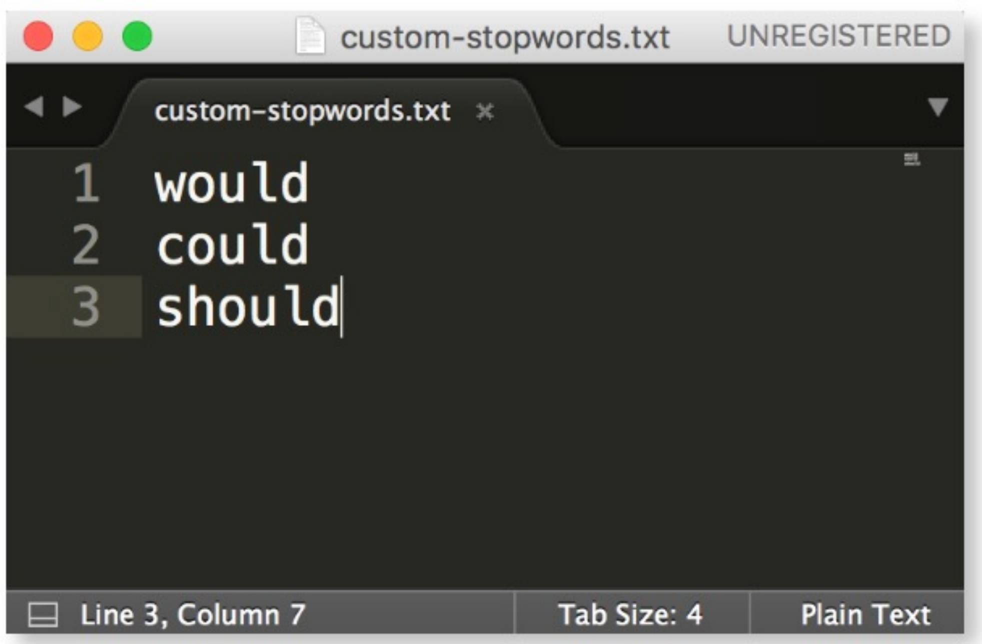
☑ Color words
Words tilt: no

Words & weights

Weight ▼	Word
668	said
295	came
275	little
275	went
245	one
199	king
175	go
171	away
168	could
156	would
154	man

Custom filtering

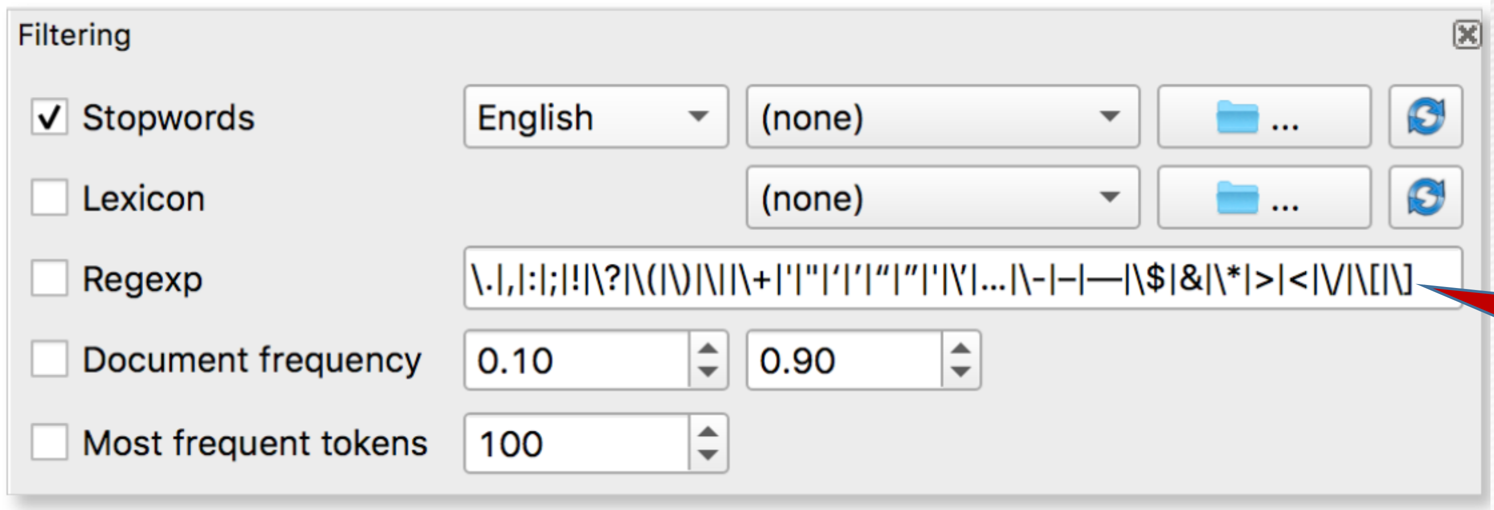
- Load the list of custom stopwords in the right-hand dropdown of the Filtering section.



```
would
could
should
```

A good plain text editor is Sublime, but you can easily work with Notepad (or 메모장).

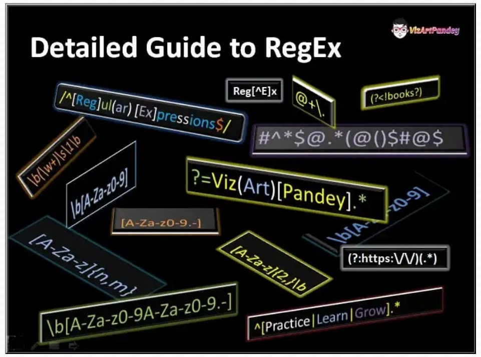
Other ways of filtering



Filtering

- ☑ Stopwords: English, (none)
- ☐ Lexicon: (none)
- ☐ Regexp: `\.|,|:|;|!|\?|\(|\)|\||\+|'|"|‘|’|“|”|'|\'|…|\-|–|—|\$|&|\*|>|<|\/|\[|\]`
- ☐ Document frequency: 0.10, 0.90
- ☐ Most frequent tokens: 100

Regular Expression



https://www.w3schools.com/python/python_regex.asp

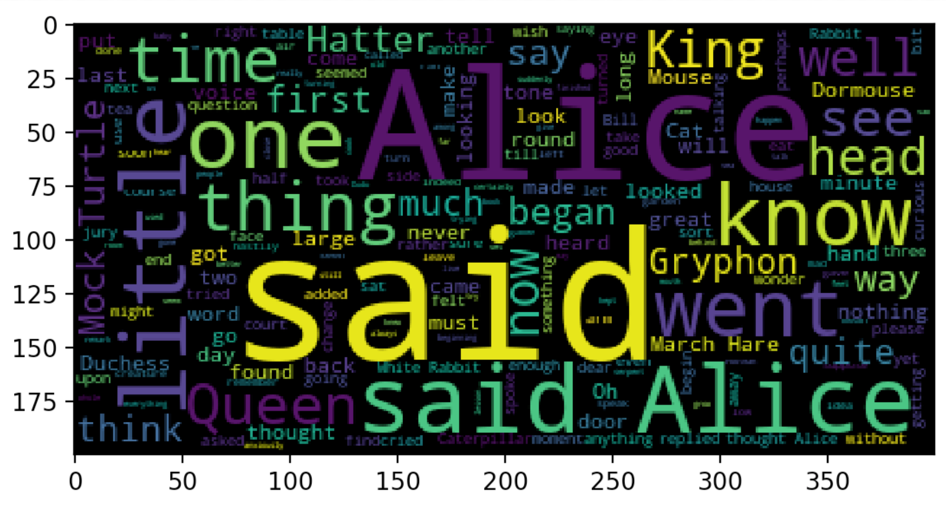
https://regexr.com

Part 6
CLI-based Word Cloud

2:00–2:20

Alice’s Adventures in Wonderland

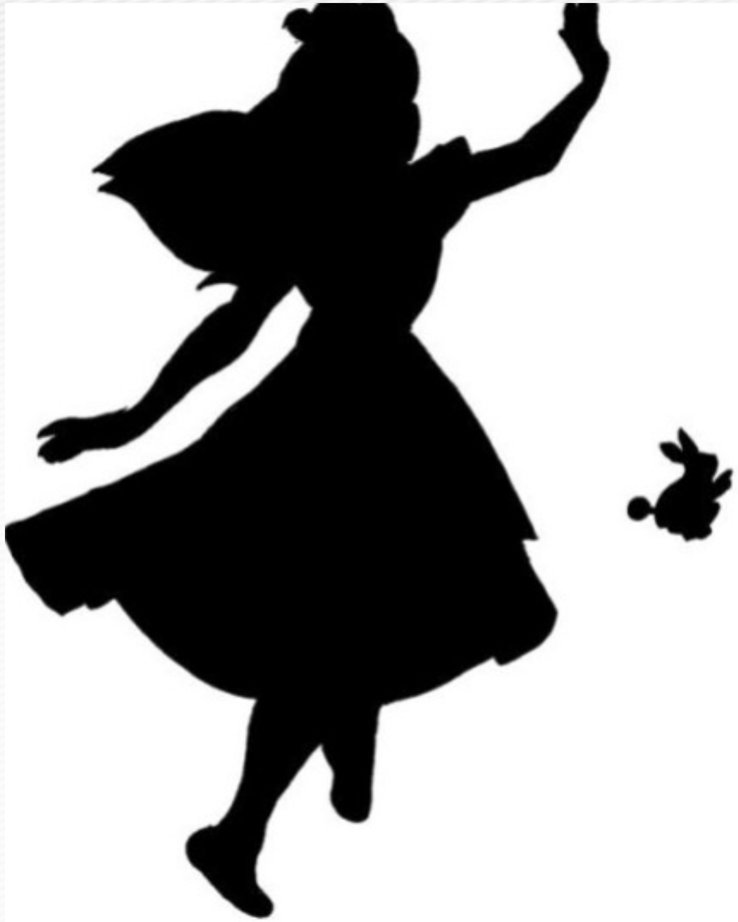
```
from wordcloud import WordCloud
text = open(‘alice.txt’).read()
wc_obj = WordCloud()
wc = wc_obj.generate(text)

import matplotlib.pyplot as plt

plt.figure()
plt.imshow(wc)
plt.axis(“off”)
plt.show()
plt.savefig(“wc_alice.png”)
```

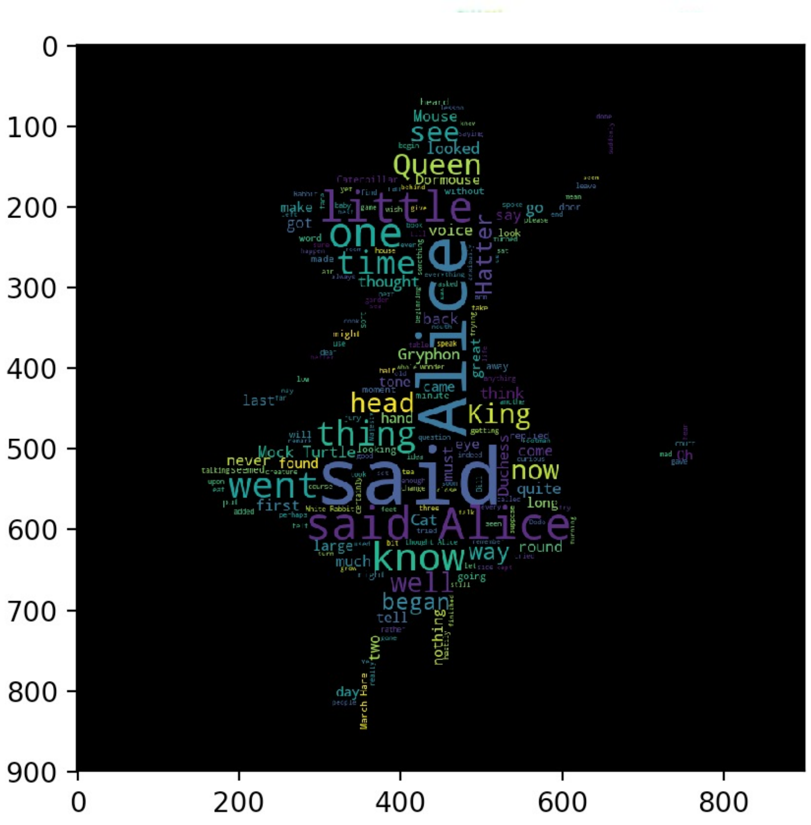


STOPWORDS in the wordcloud library

```
from wordcloud import WordCloud, STOPWORDS
stopwords = set(STOPWORDS)
stopwords.add("said")
wc = WordCloud(stopwords=stopwords)
wc = wc.generate(text)
plt.figure(figsize=(12,12))
plt.imshow(wc)
plt.axis("off")
plt.show()
```

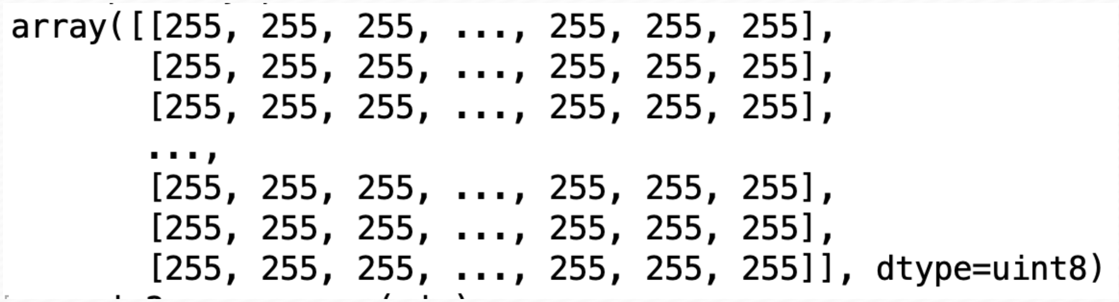
Masked Word Cloud만들기



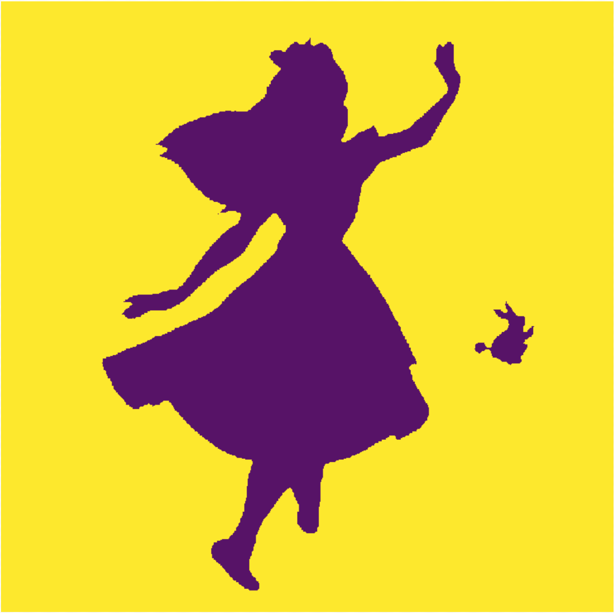


준비물: alice_mask.png

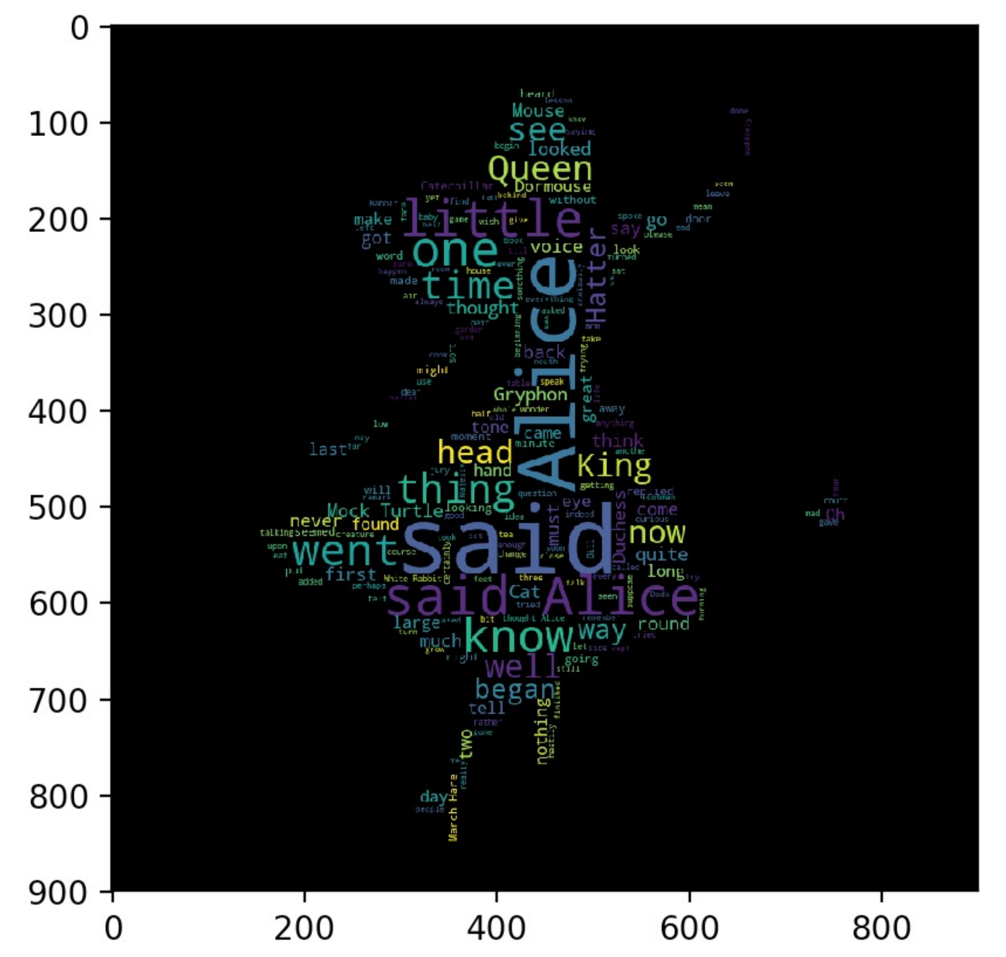
```
from PIL import Image
import numpy as np
alice_mask = np.array(Image.open(‘alice_mask.png’))
plt.figure()
plt.imshow(alice_mask)
plt.axis(“off”)
```



```
array([[255, 255, 255, ..., 255, 255, 255],
       [255, 255, 255, ..., 255, 255, 255],
       [255, 255, 255, ..., 255, 255, 255],
       ...,
       [255, 255, 255, ..., 255, 255, 255],
       [255, 255, 255, ..., 255, 255, 255],
       [255, 255, 255, ..., 255, 255, 255]], dtype=uint8)
```



```
wc_obj = WordCloud(mask=alice_mask)
wc = wc_obj.generate(text)


plt.figure()
plt.imshow(wc)
plt.imshow(alice_mask)
plt.axis(‘off’)
plt.show()
```

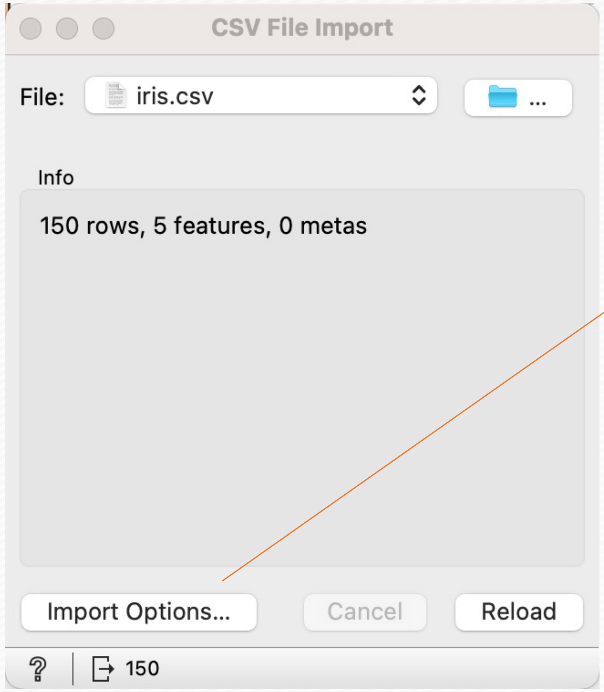


Part 7
Loading your own data

2:35–3:00

iris.csv

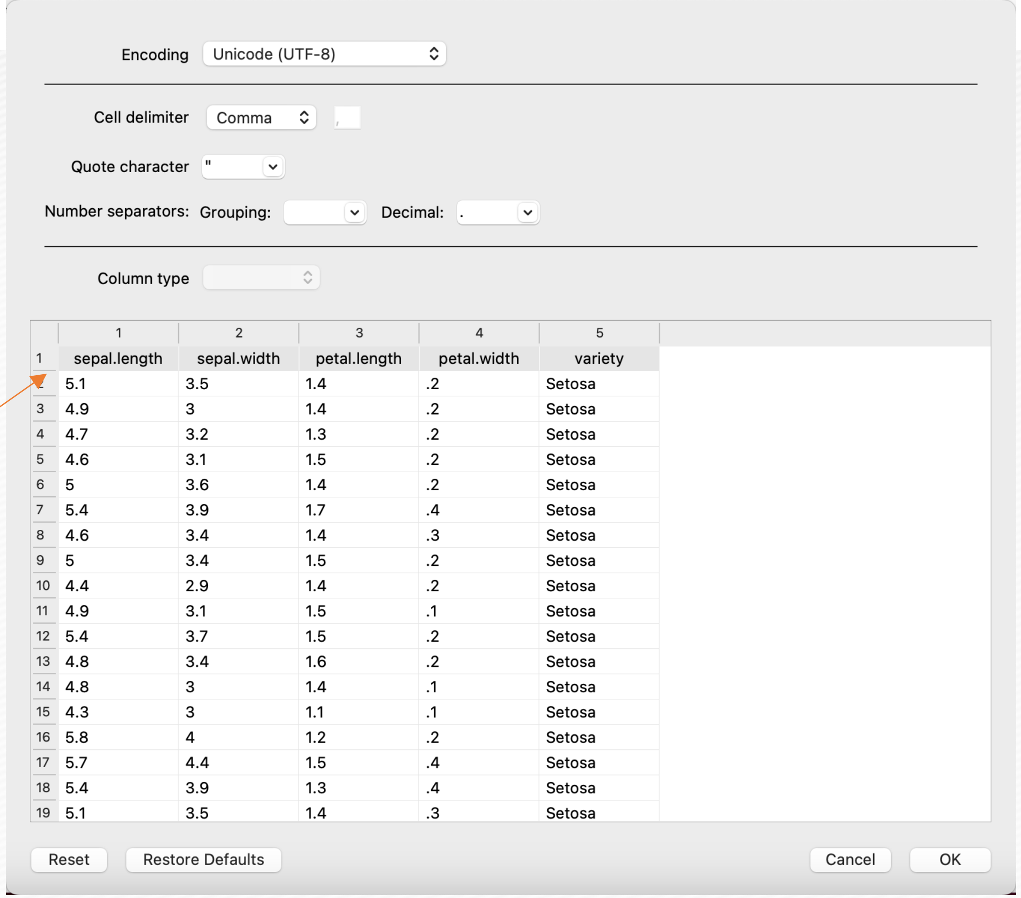
```
"sepal.length","sepal.width","petal.length","petal.width","variety"
5.1,3.5,1.4,.2,"Setosa"
4.9,3,1.4,.2,"Setosa"
4.7,3.2,1.3,.2,"Setosa"
```



File: iris.csv

Info

150 rows, 5 features, 0 metas



Encoding: Unicode (UTF-8)

Cell delimiter: Comma ,

Quote character: "

Number separators: Grouping: Decimal: .

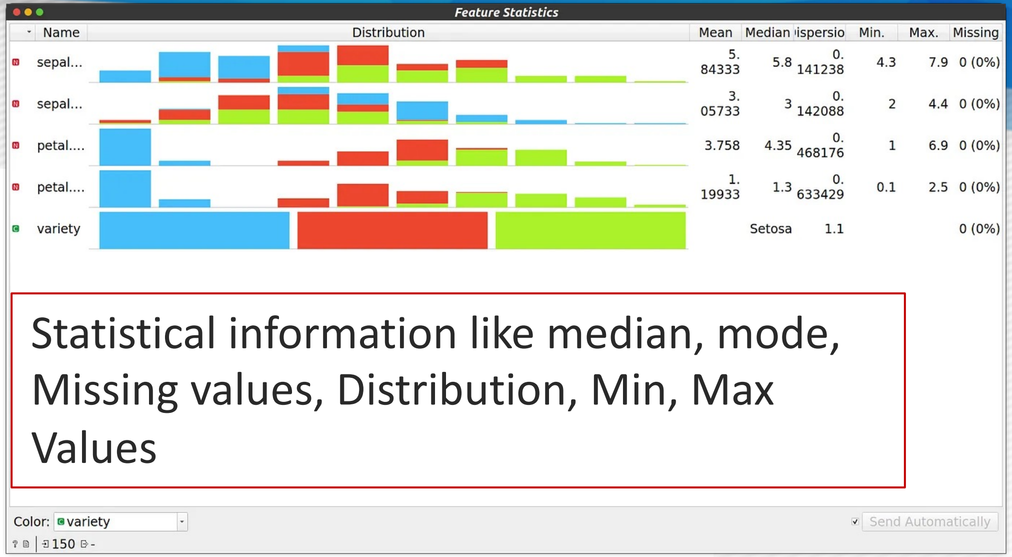
Column type

	1	2	3	4	5
1	sepal.length	sepal.width	petal.length	petal.width	variety
2	5.1	3.5	1.4	.2	Setosa
3	4.9	3	1.4	.2	Setosa
4	4.7	3.2	1.3	.2	Setosa
5	4.6	3.1	1.5	.2	Setosa
6	5	3.6	1.4	.2	Setosa
7	5.4	3.9	1.7	.4	Setosa
8	4.6	3.4	1.4	.3	Setosa
9	5	3.4	1.5	.2	Setosa
10	4.4	2.9	1.4	.2	Setosa
11	4.9	3.1	1.5	.1	Setosa
12	5.4	3.7	1.5	.2	Setosa
13	4.8	3.4	1.6	.2	Setosa
14	4.8	3	1.4	.1	Setosa
15	4.3	3	1.1	.1	Setosa
16	5.8	4	1.2	.2	Setosa
17	5.7	4.4	1.5	.4	Setosa
18	5.4	3.9	1.3	.4	Setosa
19	5.1	3.5	1.4	.3	Setosa

Attribute Type

- C: Continuous
- D: Discrete
- T: Time
- S: String

Loading a csv file



Feature Statistics

	Name	Distribution	Mean	Median	ispersio	Min.	Max.	Missing
N	sepal...		5.84333	5.8	0.141238	4.3	7.9	0 (0%)
N	sepal...		3.05733	3	0.142088	2	4.4	0 (0%)
N	petal....		3.758	4.35	0.468176	1	6.9	0 (0%)
N	petal....		1.19933	1.3	0.633429	0.1	2.5	0 (0%)
C	variety			Setosa	1.1			0 (0%)

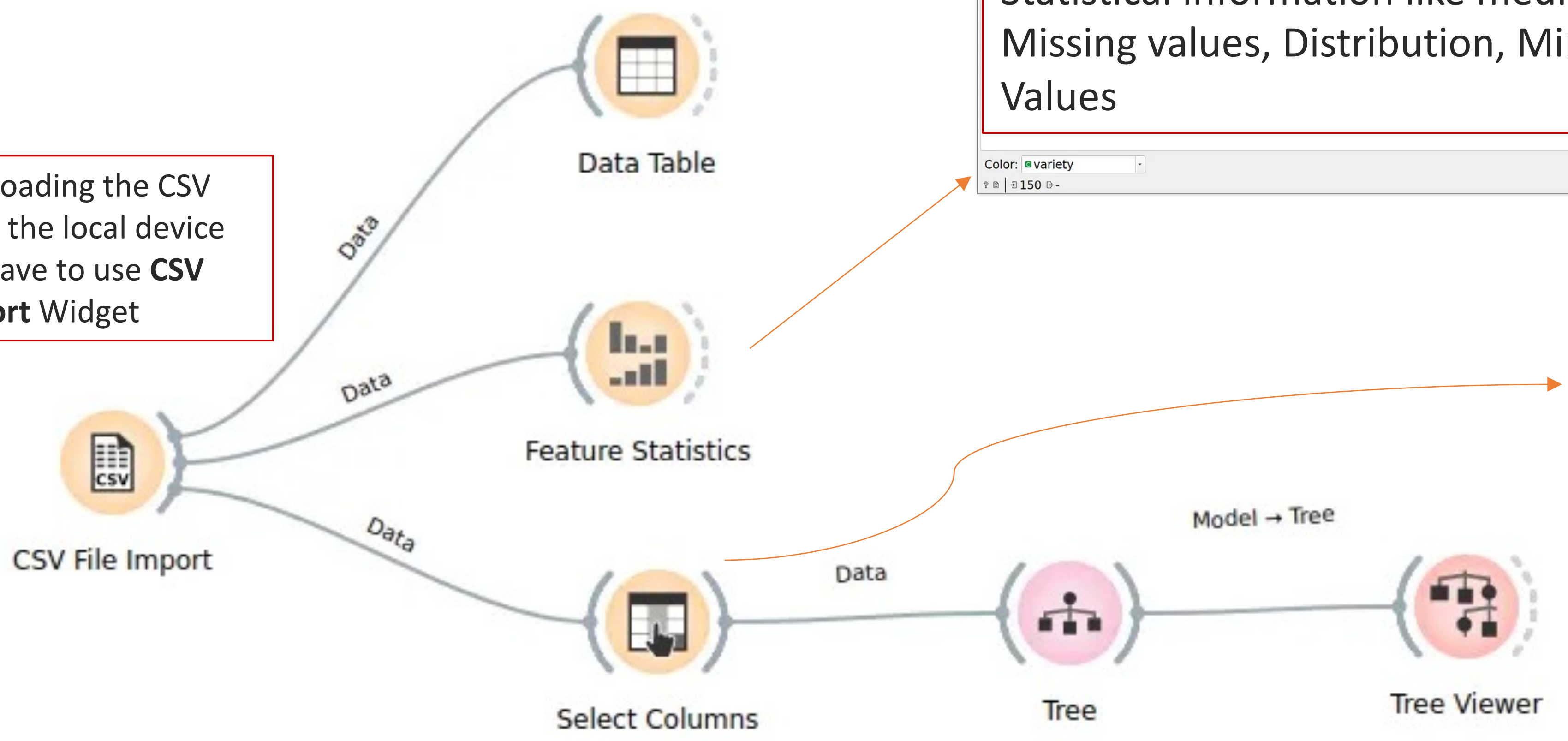
Statistical information like median, mode, Missing values, Distribution, Min, Max Values

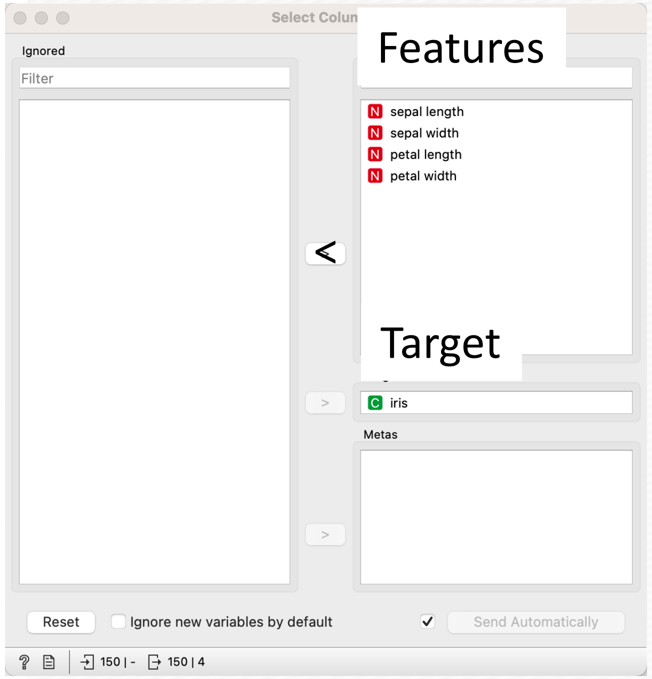
Color: variety

☑ Send Automatically

150

For Loading the CSV from the local device we have to use **CSV Import** Widget





Ignored

Features

- N sepal length
- N sepal width
- N petal length
- N petal width

Target

- C iris

Metas

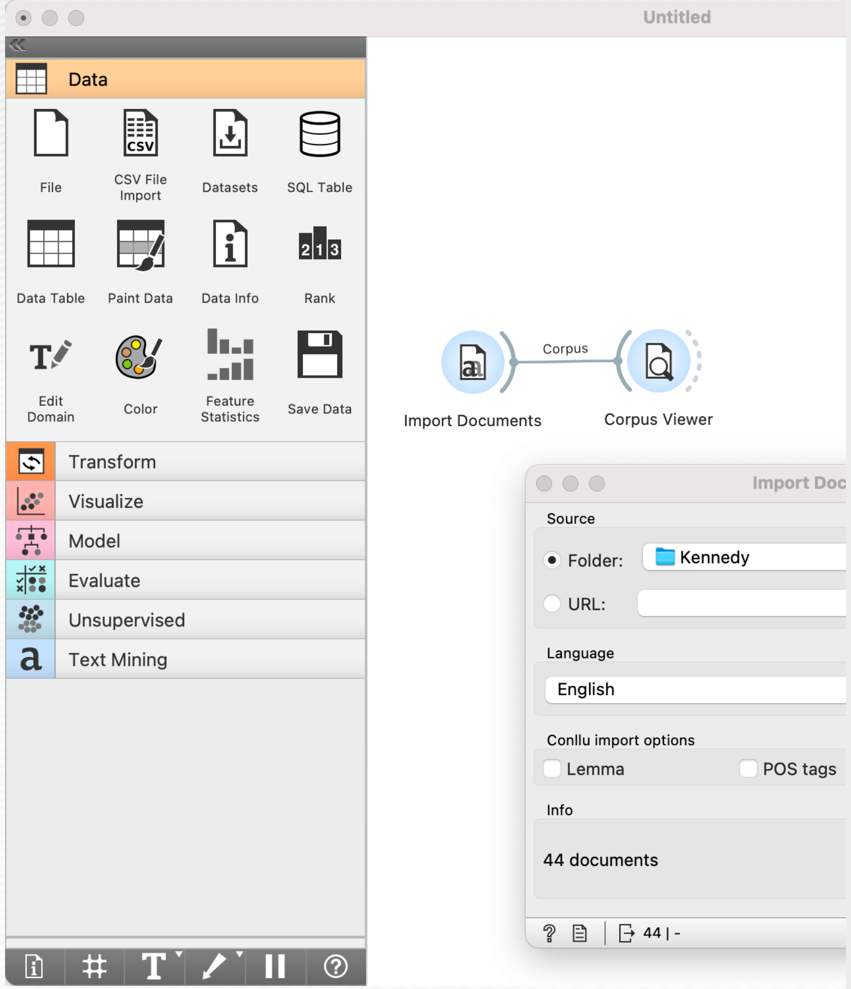
☐ Ignore new variables by default

☑ Send Automatically

150 | - 150 | 4

https://gist.githubusercontent.com/netj/8836201/raw/6f9306ad21398ea43cba4f7d537619d0e07d5ae3/iris.csv

Import your own text documents



Source

● Folder: Kennedy

○ URL:

Language

English

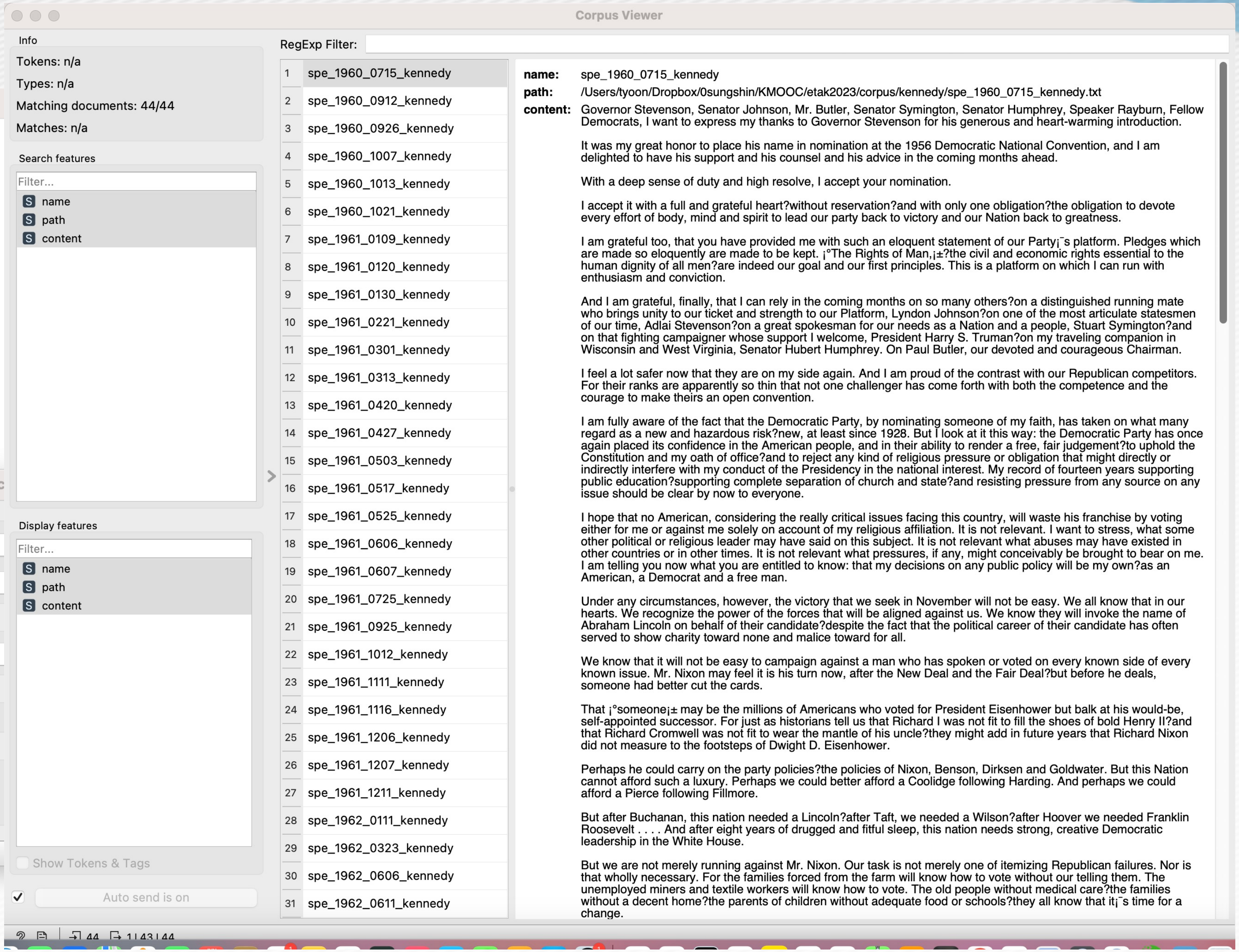
Conllu import options

☐ Lemma ☐ POS tags

Info

44 documents

44 | -



Info

Tokens: n/a

Types: n/a

Matching documents: 44/44

Matches: n/a

Search features: name, path, content

Display features: name, path, content

RegExp Filter:

1 spe_1960_0715_kennedy
2 spe_1960_0912_kennedy
3 spe_1960_0926_kennedy
4 spe_1960_1007_kennedy
5 spe_1960_1013_kennedy
6 spe_1960_1021_kennedy
7 spe_1961_0109_kennedy
8 spe_1961_0120_kennedy
9 spe_1961_0130_kennedy
10 spe_1961_0221_kennedy
11 spe_1961_0301_kennedy
12 spe_1961_0313_kennedy
13 spe_1961_0420_kennedy
14 spe_1961_0427_kennedy
15 spe_1961_0503_kennedy
16 spe_1961_0517_kennedy
17 spe_1961_0525_kennedy
18 spe_1961_0606_kennedy
19 spe_1961_0607_kennedy
20 spe_1961_0725_kennedy
21 spe_1961_0925_kennedy
22 spe_1961_1012_kennedy
23 spe_1961_1111_kennedy
24 spe_1961_1116_kennedy
25 spe_1961_1206_kennedy
26 spe_1961_1207_kennedy
27 spe_1961_1211_kennedy
28 spe_1962_0111_kennedy
29 spe_1962_0323_kennedy
30 spe_1962_0606_kennedy
31 spe_1962_0611_kennedy

name: spe_1960_0715_kennedy

path: /Users/tyoon/Dropbox/0sungshin/KMOOC/etak2023/corpus/kennedy/spe_1960_0715_kennedy.txt

content: Governor Stevenson, Senator Johnson, Mr. Butler, Senator Symington, Senator Humphrey, Speaker Rayburn, Fellow Democrats, I want to express my thanks to Governor Stevenson for his generous and heart-warming introduction.

It was my great honor to place his name in nomination at the 1956 Democratic National Convention, and I am delighted to have his support and his counsel and his advice in the coming months ahead.

With a deep sense of duty and high resolve, I accept your nomination.

I accept it with a full and grateful heart?without reservation?and with only one obligation?the obligation to devote every effort of body, mind and spirit to lead our party back to victory and our Nation back to greatness.

I am grateful too, that you have provided me with such an eloquent statement of our Party¡¯s platform. Pledges which are made so eloquently are made to be kept. ¡°The Rights of Man,¡±?the civil and economic rights essential to the human dignity of all men?are indeed our goal and our first principles. This is a platform on which I can run with enthusiasm and conviction.

And I am grateful, finally, that I can rely in the coming months on so many others?on a distinguished running mate who brings unity to our ticket and strength to our Platform, Lyndon Johnson?on one of the most articulate statesmen of our time, Adlai Stevenson?on a great spokesman for our needs as a Nation and a people, Stuart Symington?and on that fighting campaigner whose support I welcome, President Harry S. Truman?on my traveling companion in Wisconsin and West Virginia, Senator Hubert Humphrey. On Paul Butler, our devoted and courageous Chairman.

I feel a lot safer now that they are on my side again. And I am proud of the contrast with our Republican competitors. For their ranks are apparently so thin that not one challenger has come forth with both the competence and the courage to make theirs an open convention.

I am fully aware of the fact that the Democratic Party, by nominating someone of my faith, has taken on what many regard as a new and hazardous risk?new, at least since 1928. But I look at it this way: the Democratic Party has once again placed its confidence in the American people, and in their ability to render a free, fair judgement?to uphold the Constitution and my oath of office?and to reject any kind of religious pressure or obligation that might directly or indirectly interfere with my conduct of the Presidency in the national interest. My record of fourteen years supporting public education?supporting complete separation of church and state?and resisting pressure from any source on any issue should be clear by now to everyone.

I hope that no American, considering the really critical issues facing this country, will waste his franchise by voting either for me or against me solely on account of my religious affiliation. It is not relevant. I want to stress, what some other political or religious leader may have said on this subject. It is not relevant what abuses may have existed in other countries or in other times. It is not relevant what pressures, if any, might conceivably be brought to bear on me. I am telling you now what you are entitled to know: that my decisions on any public policy will be my own?as an American, a Democrat and a free man.

Under any circumstances, however, the victory that we seek in November will not be easy. We all know that in our hearts. We recognize the power of the forces that will be aligned against us. We know they will invoke the name of Abraham Lincoln on behalf of their candidate?despite the fact that the political career of their candidate has often served to show charity toward none and malice toward for all.

We know that it will not be easy to campaign against a man who has spoken or voted on every known side of every known issue. Mr. Nixon may feel it is his turn now, after the New Deal and the Fair Deal?but before he deals, someone had better cut the cards.

That ¡°someone¡± may be the millions of Americans who voted for President Eisenhower but balk at his would-be, self-appointed successor. For just as historians tell us that Richard I was not fit to fill the shoes of bold Henry II?and that Richard Cromwell was not fit to wear the mantle of his uncle?they might add in future years that Richard Nixon did not measure to the footsteps of Dwight D. Eisenhower.

Perhaps he could carry on the party policies?the policies of Nixon, Benson, Dirksen and Goldwater. But this Nation cannot afford such a luxury. Perhaps we could better afford a Coolidge following Harding. And perhaps we could afford a Pierce following Fillmore.

But after Buchanan, this nation needed a Lincoln?after Taft, we needed a Wilson?after Hoover we needed Franklin Roosevelt . . . . And after eight years of drugged and fitful sleep, this nation needs strong, creative Democratic leadership in the White House.

But we are not merely running against Mr. Nixon. Our task is not merely one of itemizing Republican failures. Nor is that wholly necessary. For the families forced from the farm will know how to vote without our telling them. The unemployed miners and textile workers will know how to vote. The old people without medical care?the families without a decent home?the parents of children without adequate food or schools?they all know that it¡¯s time for a change.

Part 8
Concordance

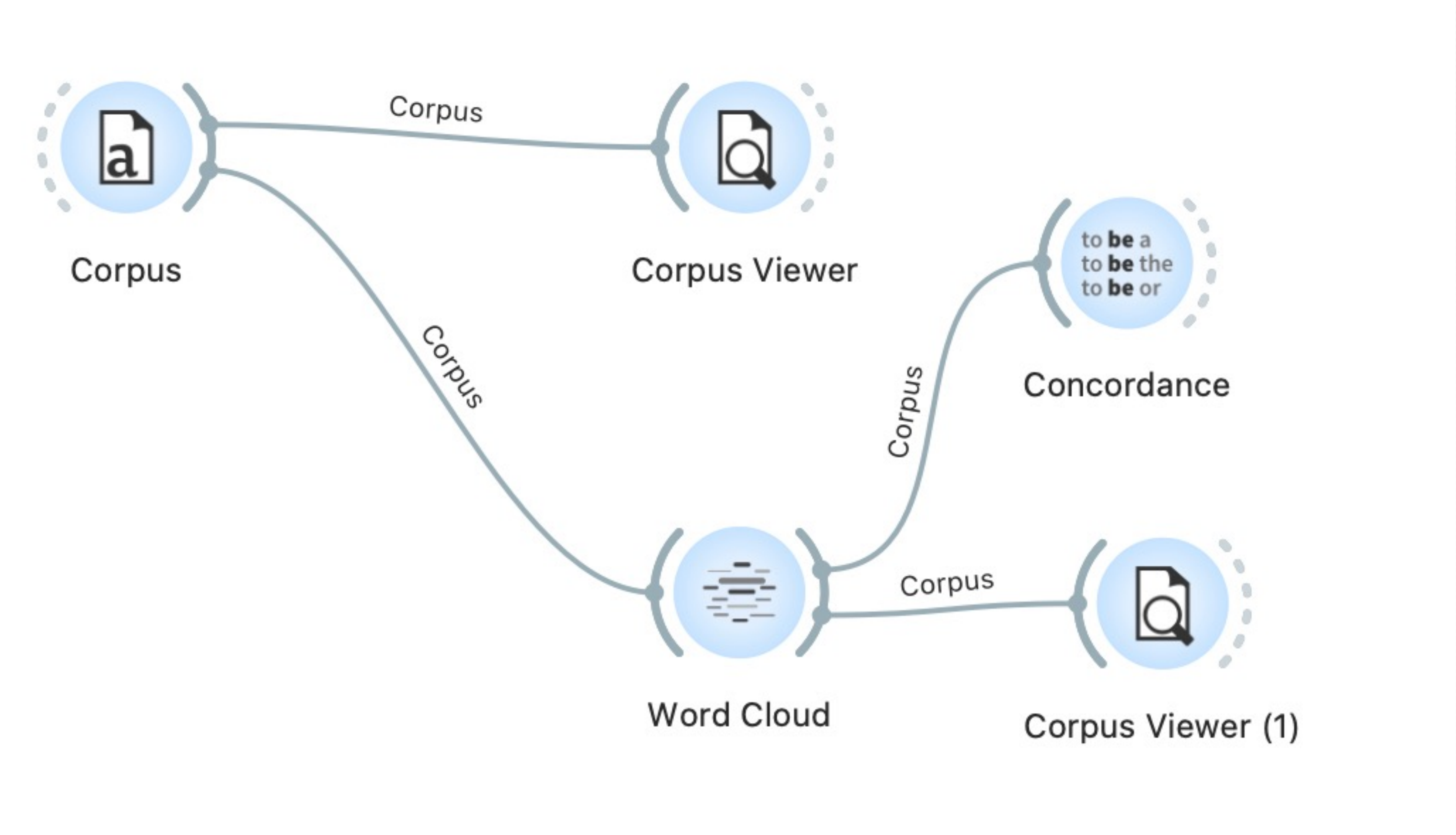
3:00–3:10

Visualizing corpus - Concordance

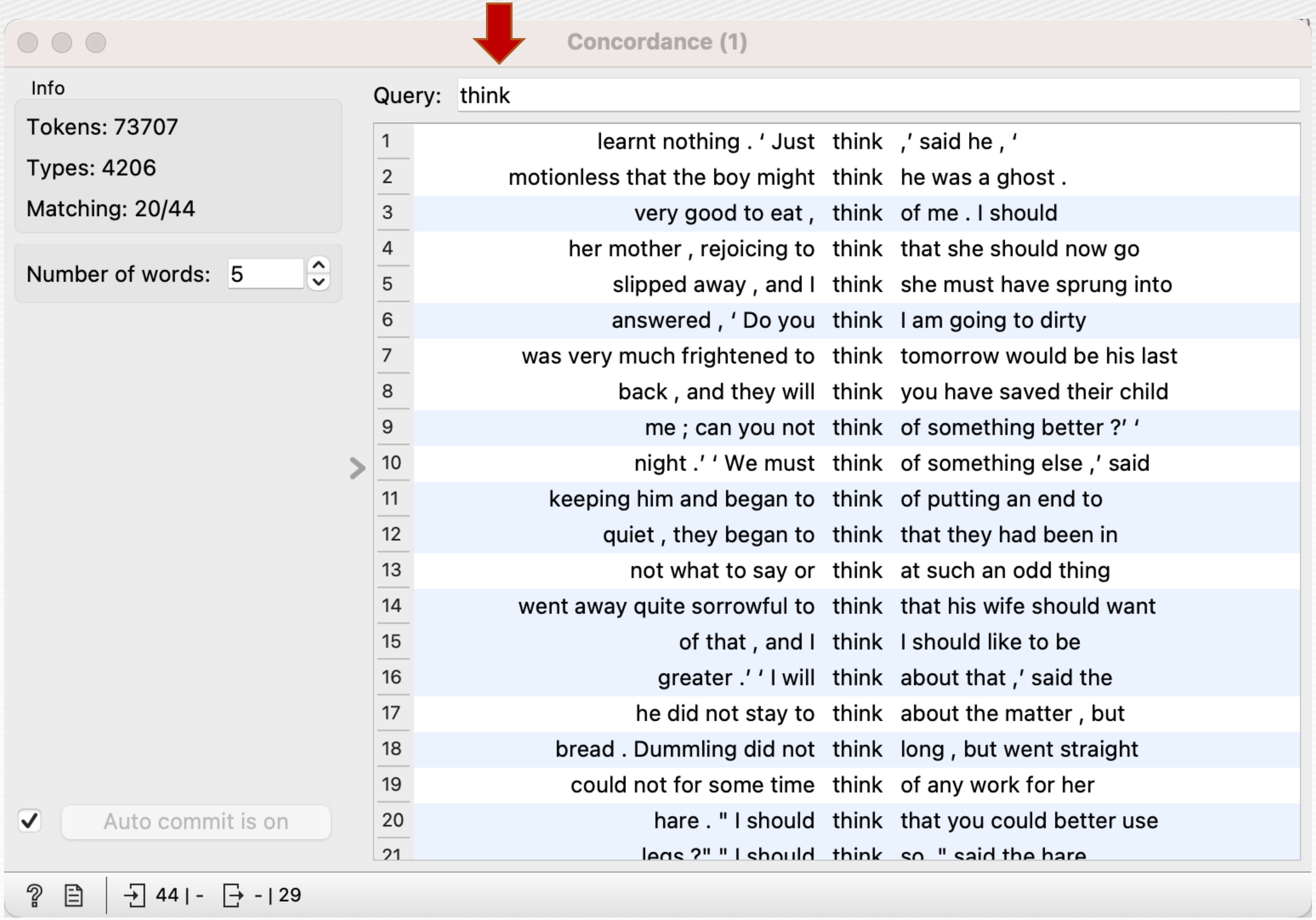
- We have already seen some of the preprocessing results in a word cloud.
  - Word Cloud shows us word frequencies.
- But we still don’t know much about the use of a specific word in a text.
  - For example, ‘oh’ could be a lowercase version of OH (the chemical compound of hydroxide), a simple exclamation ‘Oh!’ or an abbreviation for the state of Ohio.
- To check the context of a particular word we can use Concordance widget.
  - Concordance shows us the text around our word.

Concordance widget

- Connect Concordance to Corpus to pass the text to the widget.



Concordance



Concordance (1)

Info

Tokens: 73707

Types: 4206

Matching: 20/44

Number of words: 5

Query: think

1	learnt nothing . ' Just	think	,' said he , '
2	motionless that the boy might	think	he was a ghost .
3	very good to eat ,	think	of me . I should
4	her mother , rejoicing to	think	that she should now go
5	slipped away , and I	think	she must have sprung into
6	answered , ' Do you	think	I am going to dirty
7	was very much frightened to	think	tomorrow would be his last
8	back , and they will	think	you have saved their child
9	me ; can you not	think	of something better ?' '
10	night .' ' We must	think	of something else ,' said
11	keeping him and began to	think	of putting an end to
12	quiet , they began to	think	that they had been in
13	not what to say or	think	at such an odd thing
14	went away quite sorrowful to	think	that his wife should want
15	of that , and I	think	I should like to be
16	greater .' ' I will	think	about that ,' said the
17	he did not stay to	think	about the matter , but
18	bread . Dummling did not	think	long , but went straight
19	could not for some time	think	of any work for her
20	hare . " I should	think	that you could better use
21	legs ?" " I should	think	so , " said the hare

☑ Auto commit is on

44 | -   - | 29

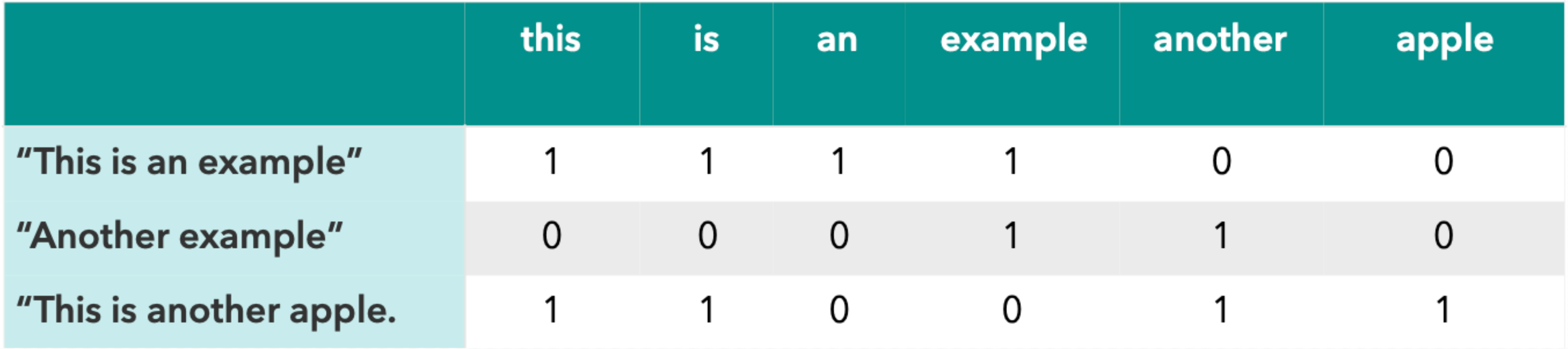
- To browse the word, type it in the query line at the top or provide it with the Word Cloud.
- Here we have selected the word 'think' and observed the context in Concordance.

Part 9
Bag of Words

3:10-3:30

Finding patterns in the text

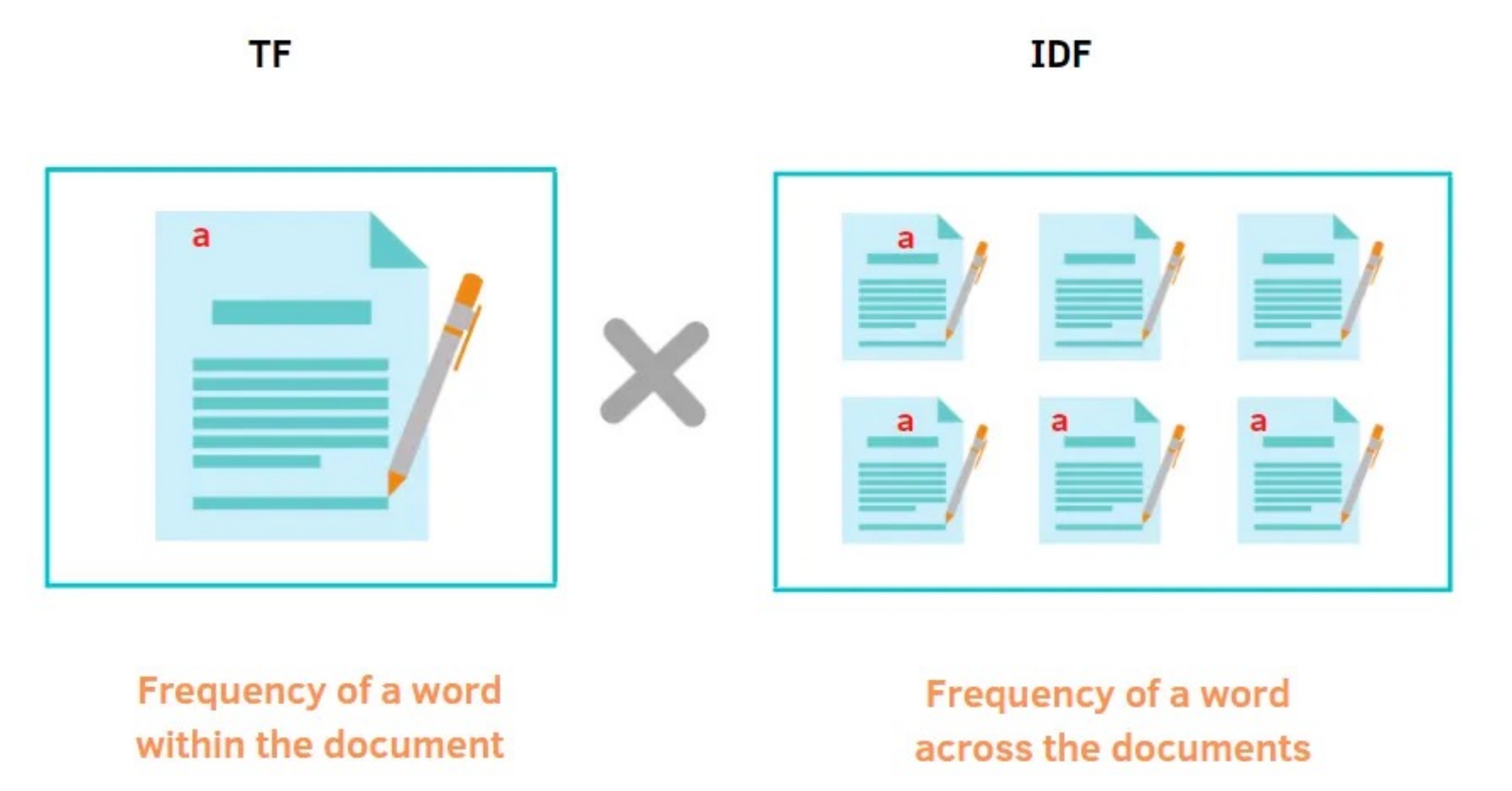
- To find any patterns in our text, we need to convert documents into numeric vectors is to count the words in each text.
- Bag of Words creates a table with words in columns and documents in rows.



	this	is	an	example	another	apple
“This is an example”	1	1	1	1	0	0
“Another example”	0	0	0	1	1	0
“This is another apple.	1	1	0	0	1	1

https://www.youtube.com/watch?v=rH_vQxQL6oM&list=PLmNPvQr9Tf-bxmjlFRg0AGjMzzd1y8_kB&index=2

TF-IDF

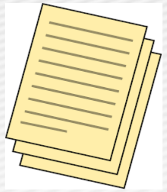

TF
IDF
a
a
a
a
a
Frequency of a word within the document
Frequency of a word across the documents

$$\mathrm{TF}(t,d) = \frac{\text{number of times t appears in d}}{\text{total number of words in d}}$$

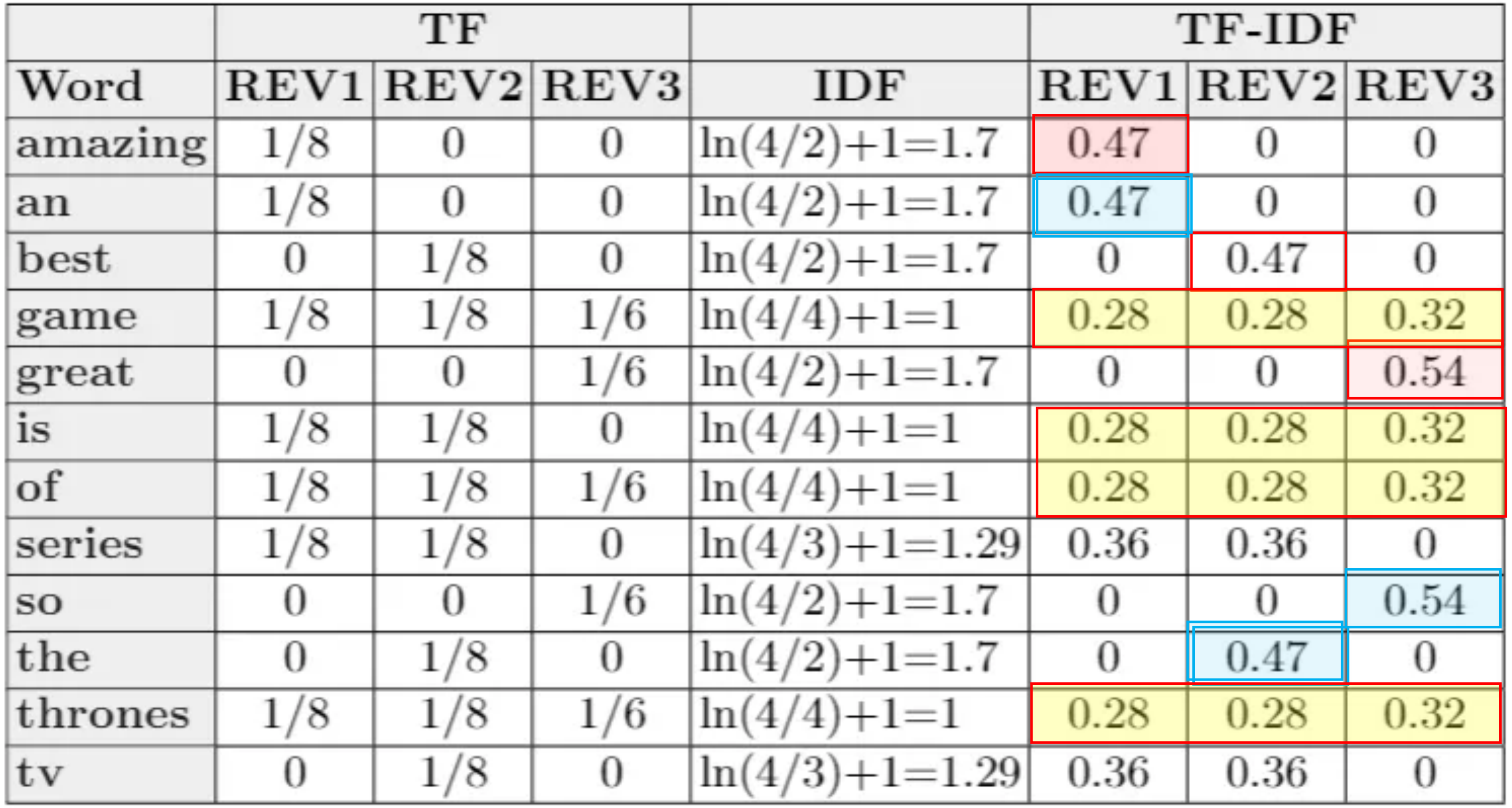
$$\mathrm{IDF}(t) = \log \frac{\text{Total number of documents}}{\text{Number of documents that contain t}}$$

https://betterprogramming.pub/a-friendly-guide-to-nlp-tf-idf-with-python-example-5fcb26286a33

TF-IDF toy example



Review 1: Game of Thrones is an amazing TV series!
Review 2: Game of Thrones is the best TV series!
Review 3: Game of Thrones is so great.

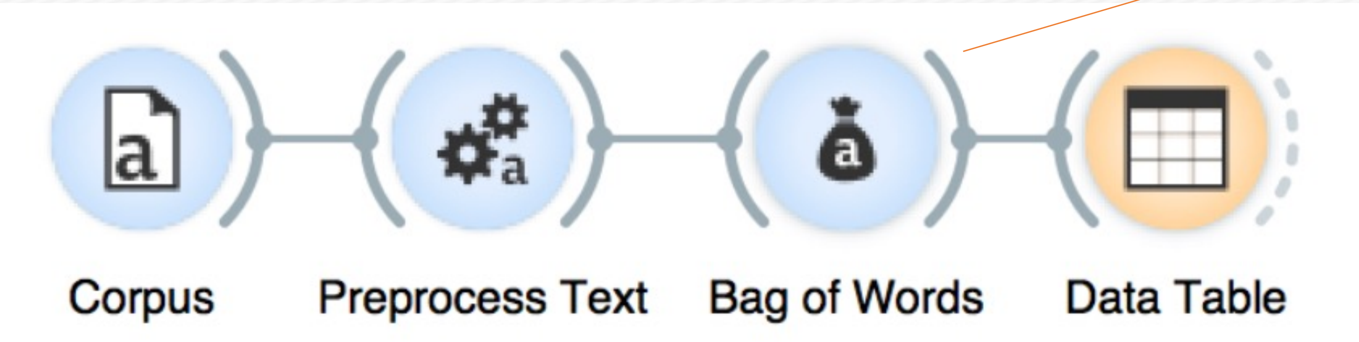


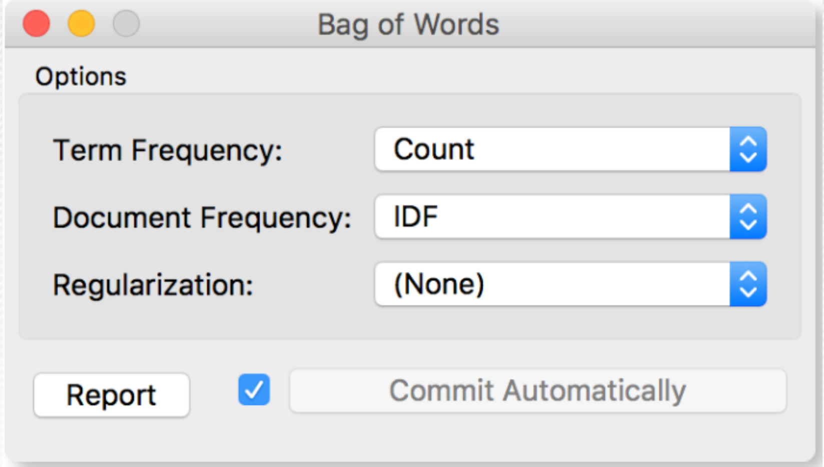
	TF				TF-IDF		
Word	REV1	REV2	REV3	IDF	REV1	REV2	REV3
amazing	1/8	0	0	$\ln(4/2)+1=1.7$	0.47	0	0
an	1/8	0	0	$\ln(4/2)+1=1.7$	0.47	0	0
best	0	1/8	0	$\ln(4/2)+1=1.7$	0	0.47	0
game	1/8	1/8	1/6	$\ln(4/4)+1=1$	0.28	0.28	0.32
great	0	0	1/6	$\ln(4/2)+1=1.7$	0	0	0.54
is	1/8	1/8	0	$\ln(4/4)+1=1$	0.28	0.28	0.32
of	1/8	1/8	1/6	$\ln(4/4)+1=1$	0.28	0.28	0.32
series	1/8	1/8	0	$\ln(4/3)+1=1.29$	0.36	0.36	0
so	0	0	1/6	$\ln(4/2)+1=1.7$	0	0	0.54
the	0	1/8	0	$\ln(4/2)+1=1.7$	0	0.47	0
thrones	1/8	1/8	1/6	$\ln(4/4)+1=1$	0.28	0.28	0.32
tv	0	1/8	0	$\ln(4/3)+1=1.29$	0.36	0.36	0

- the most present words, such as “game”, “of”, “thrones”, “is”, have the smallest IDF
- terms like “amazing” and “great” have higher TF-IDF values
- common words, like “so” and “the”, contribute more too since they aren’t present in all the sentences and we didn’t remove the stop words to keep the approach as simple as possible.

A Bag of Words Widget

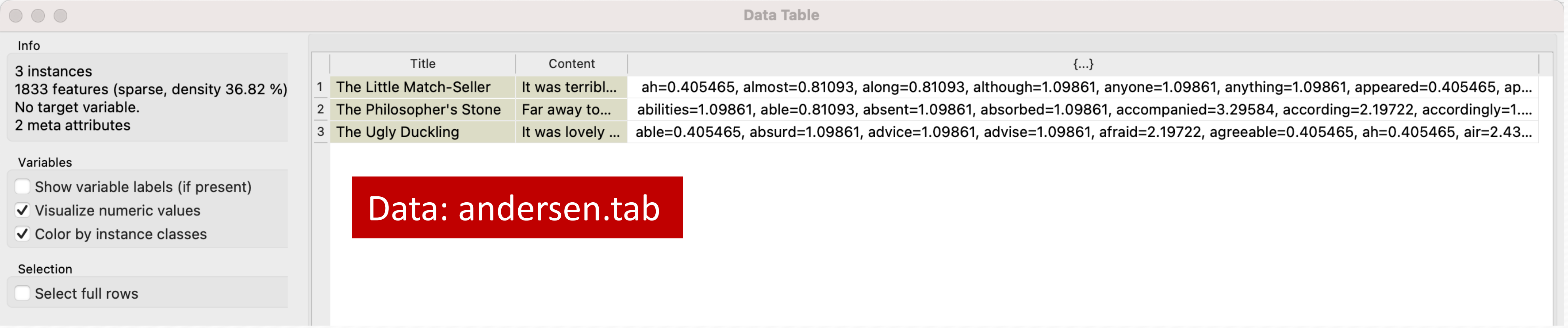
Pass the data through **a Bag of Words widget** and then again to a Data Table.
We get a new column that contains word counts for each document.





Options

Term Frequency: Count
Document Frequency: IDF
Regularization: (None)



Info

3 instances
1833 features (sparse, density 36.82 %)
No target variable.
2 meta attributes

Variables

☐ Show variable labels (if present)
☑ Visualize numeric values
☑ Color by instance classes

Selection

☐ Select full rows

	Title	Content	{...}
1	The Little Match-Seller	It was terribl...	ah=0.405465, almost=0.81093, along=0.81093, although=1.09861, anyone=1.09861, anything=1.09861, appeared=0.405465, ap...
2	The Philosopher's Stone	Far away to...	abilities=1.09861, able=0.81093, absent=1.09861, absorbed=1.09861, accompanied=3.29584, according=2.19722, accordingly=1....
3	The Ugly Duckling	It was lovely ...	able=0.405465, absurd=1.09861, advice=1.09861, advise=1.09861, afraid=2.19722, agreeable=0.405465, ah=0.405465, air=2.43...

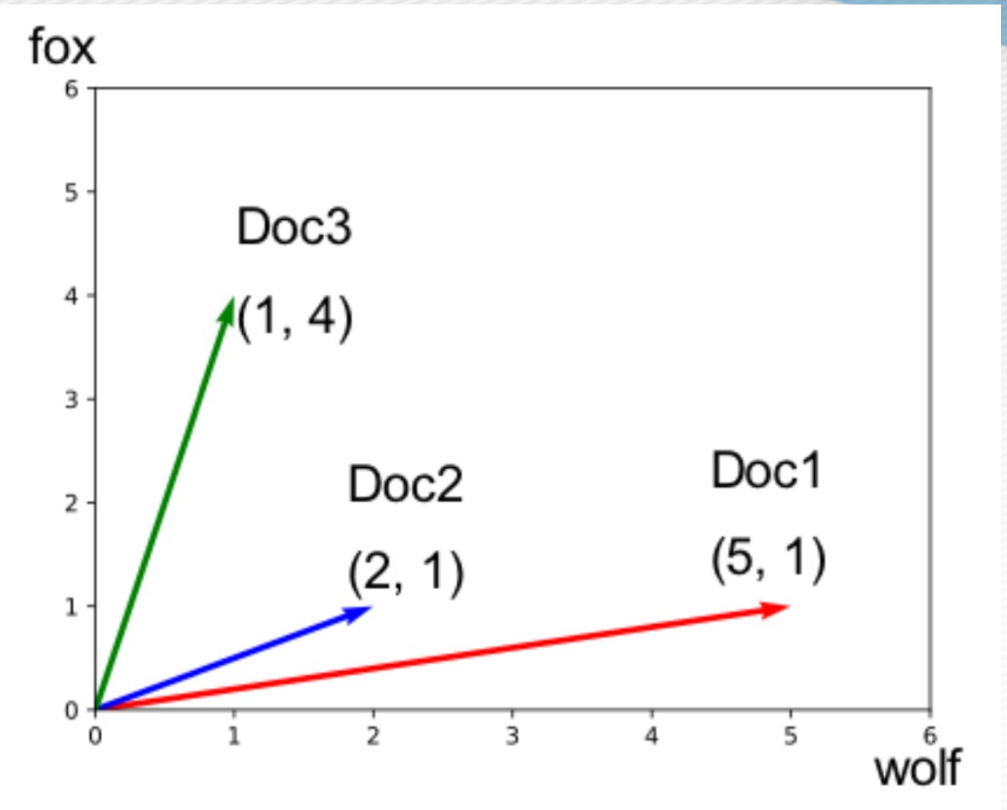
Data: andersen.tab

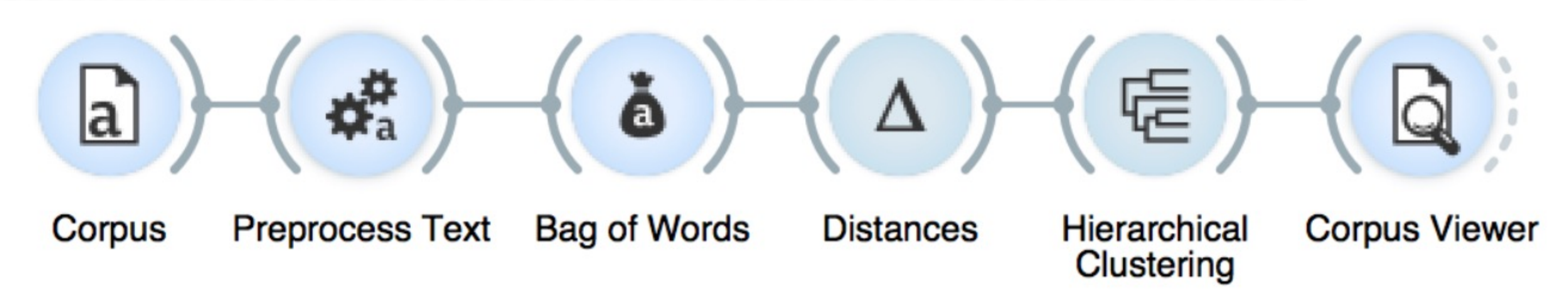
Part 10
Clustering and Distances

3:30–3:50

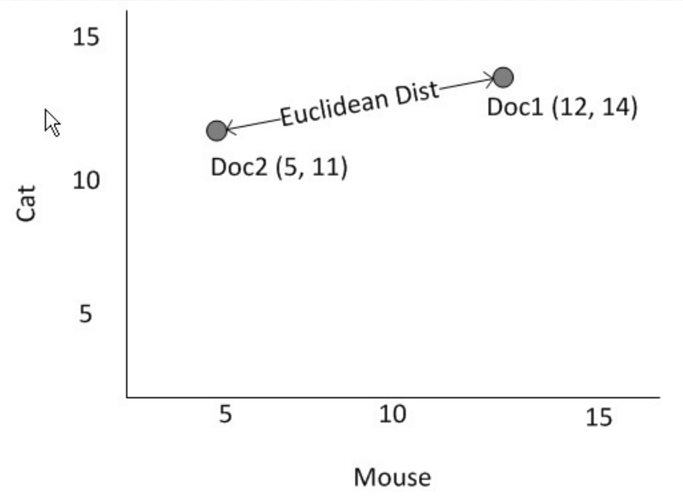
Distance and similar documents

- One common task in text mining is finding interesting groups of similar documents.
- That is, we would like to identify documents that are similar to each other. .
- We pass the data to Distances, use Euclidean distance, then to Hierarchical Clustering.

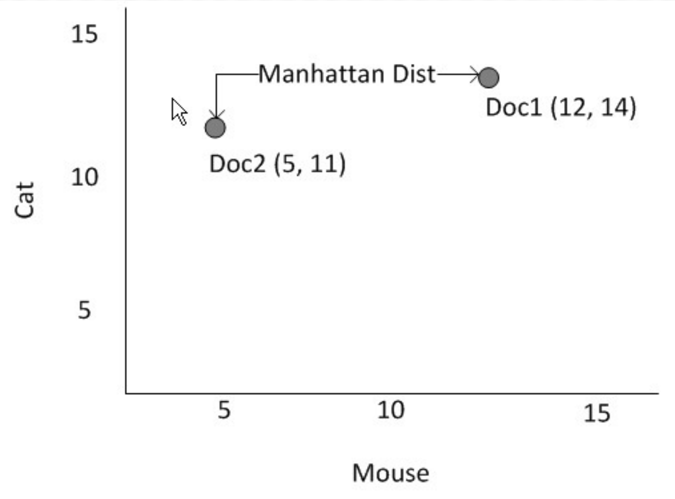




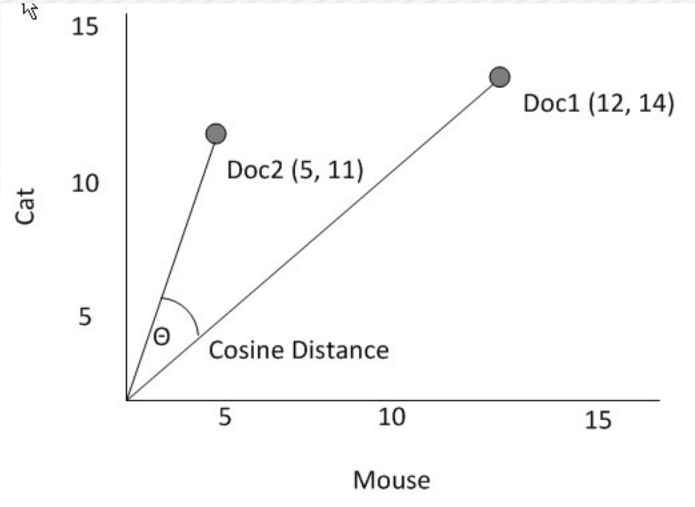
Types of Distance



The Euclidean Distance is the shortest distance between two points.



The Manhattan distance is the sum of the lengths of the rectangle formed by the two points



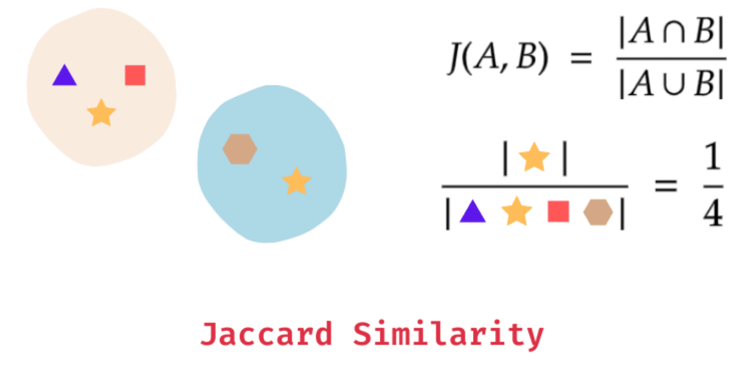
the Cosine distance is the angle subtended at the origin between the two documents.

https://nickgrattan.wordpress.com/2014/06/10/euclidean-manhattan-and-cosine-distance-measures-in-c/

Jaccard Similarity coefficient or Jaccard Index

- For text, an intuitive approach for measuring similarity would also be the number of words that two documents share.

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$$


$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$
$\frac{|★|}{|▲ ★ ■ ⬢|} = \frac{1}{4}$

Jaccard Similarity

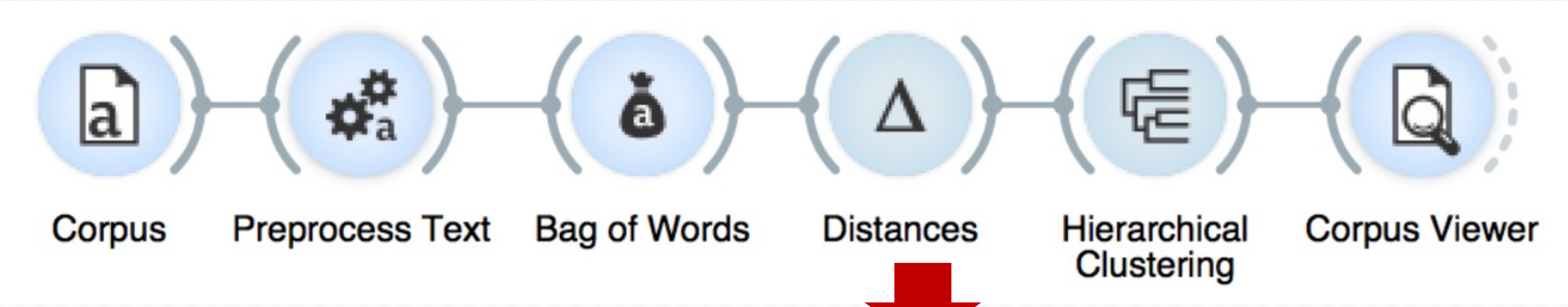
- The measure is called Jaccard similarity coefficient or Jaccard index.
- Note that in this case, we are measuring similarity, not distance.

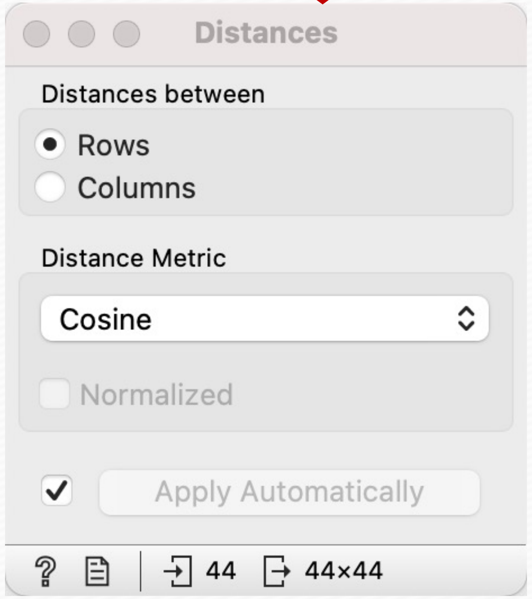
https://datascienceparichay.com/article/jaccard-similarity-python/

Workflow

- Now, let us go back to our Grimm's Tales and construct the following workflow:

You can try the same workflow on a different corpus, say ***bookexcerpt.tab***, which contains excerpts from adult and children's books.





Distances between

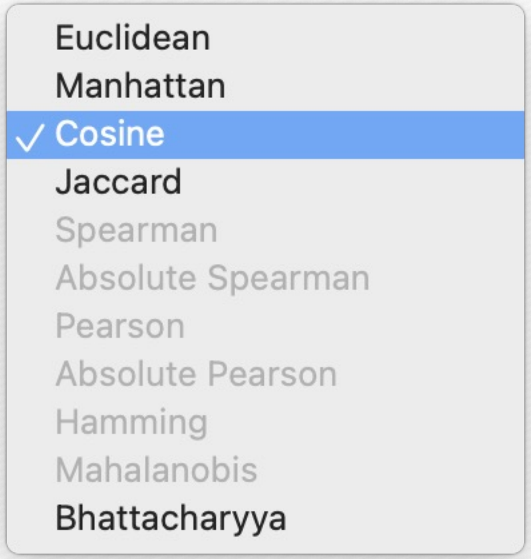
● Rows
○ Columns

Distance Metric

Cosine

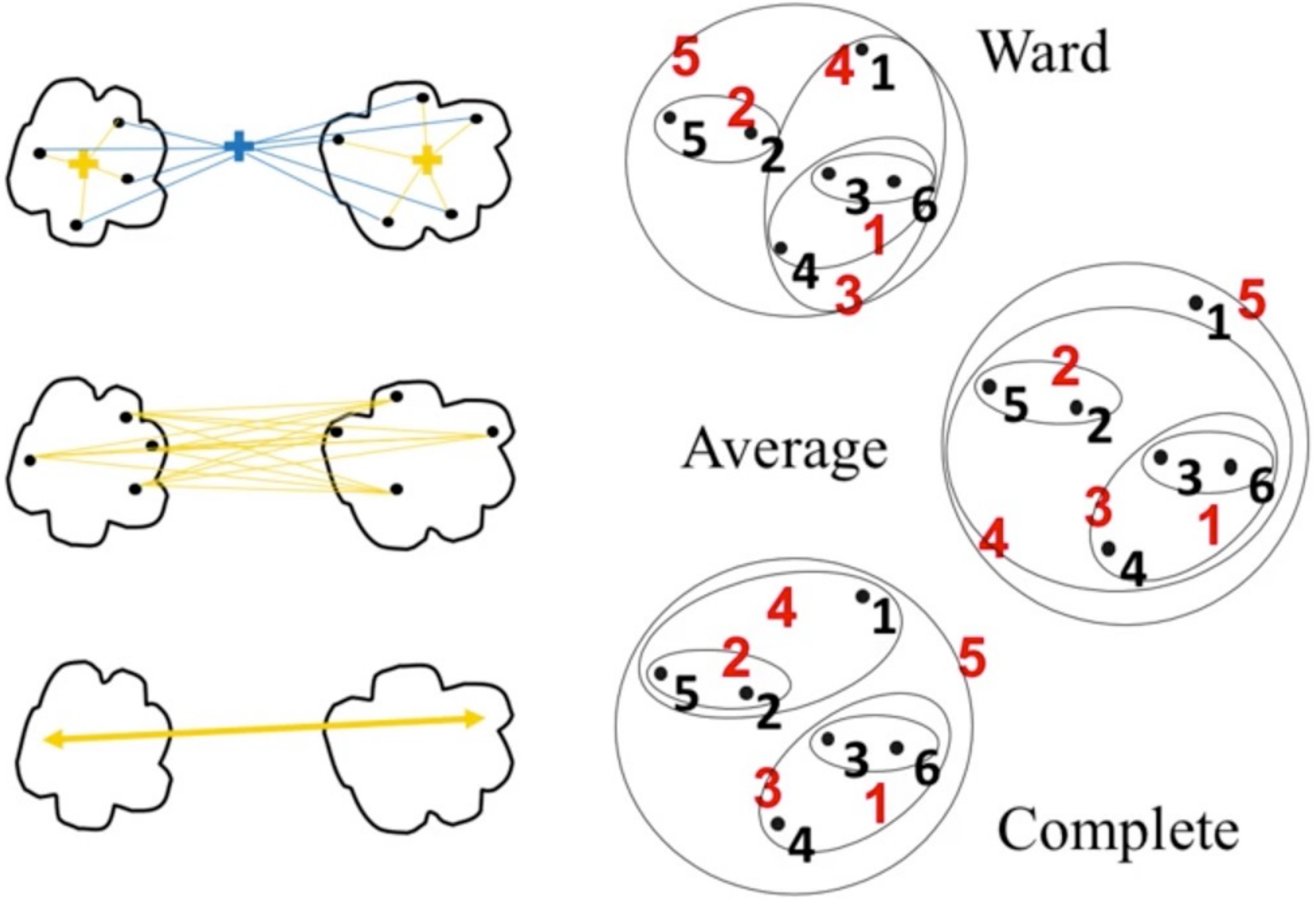
☐ Normalized

☑ Apply Automatically



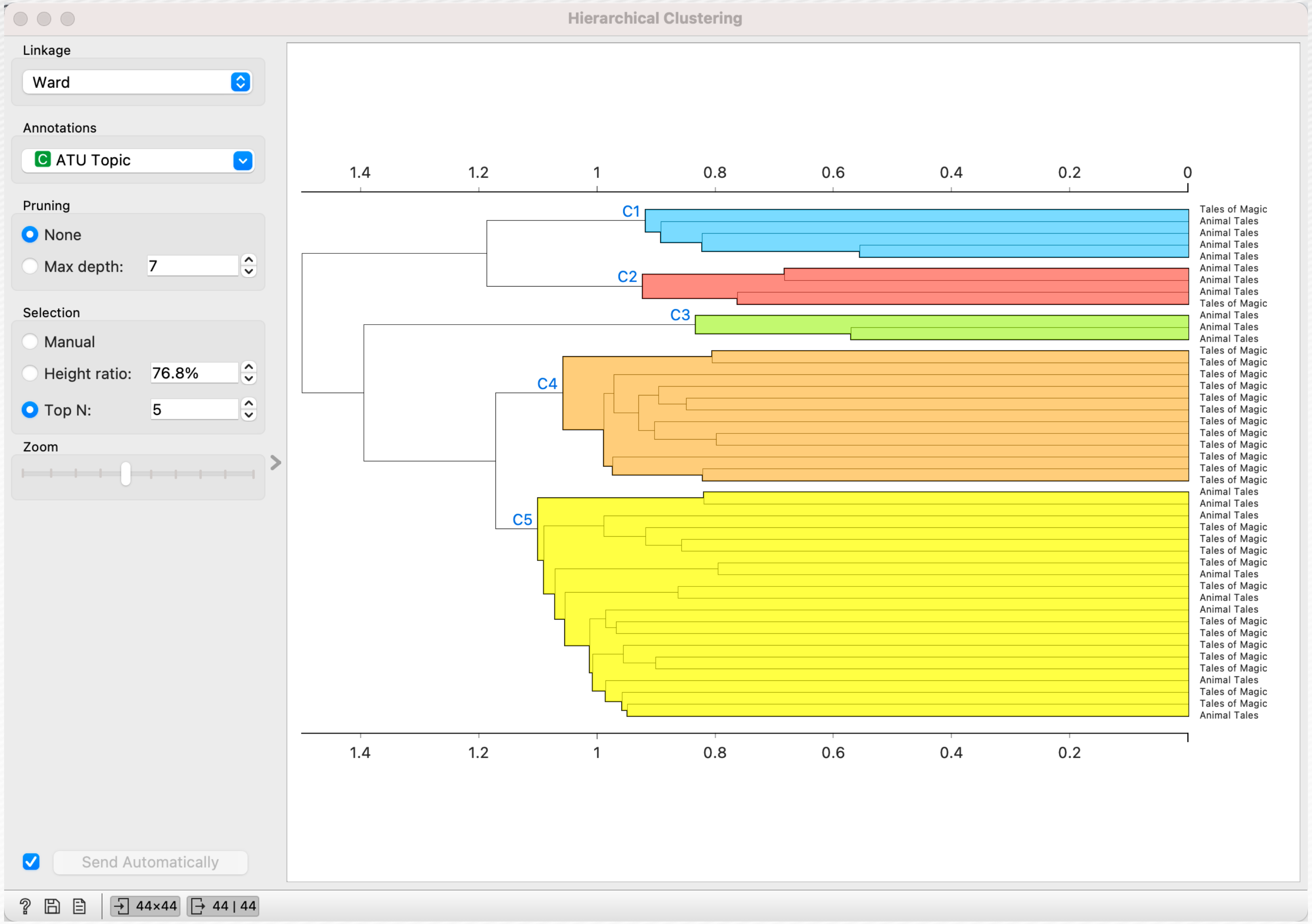
Hierarchical Clustering

- **Ward's method**
  - *Least increase in total variance (around cluster centroids)*
- **Average linkage**
  - *Average distance between clusters*
- **Complete linkage**
  - *Max distance between clusters*



https://python-data-science.readthedocs.io/en/latest/_images/aggocluster2.png

Hierarchical Clustering



Hierarchical Clustering

Linkage
Ward

Annotations
ATU Topic

Pruning
● None
○ Max depth: 7

Selection
○ Manual
○ Height ratio: 76.8%
● Top N: 5

Zoom

☑ Send Automatically

44×44 44 | 44

Why are some Tales of Magic mixed with Animal Tales? What do they have in common?

https://datascience.stackexchange.com/questions/12129/derivation-of-wards-formula-for-agglomerative-clustering

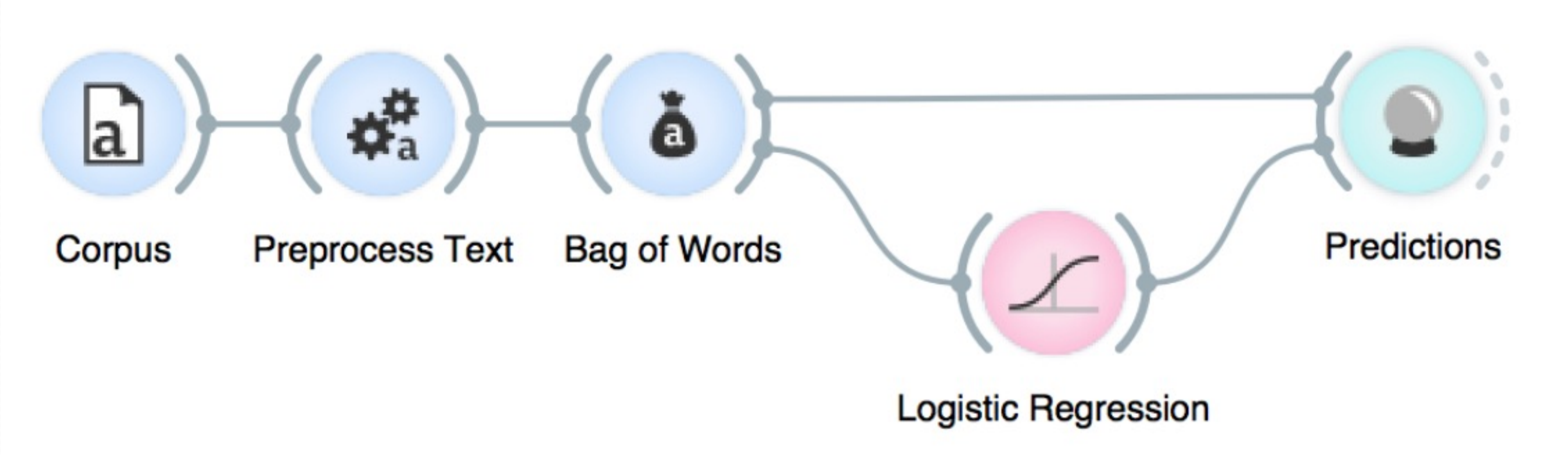
성신여자대학교
SUNGSHIN UNIVERSITY

Part 11
Classification

4:00–4:20

predict the ATU type based on the content of the tale

- The Aarne-Thompson type (ATU)
  - the index of folk-tale motifs
  - It is already marked every tale with a high-level (genre) and a mid-level ATU type (subgenre).
- Could we perhaps predict the ATU type based on the content of the tale?



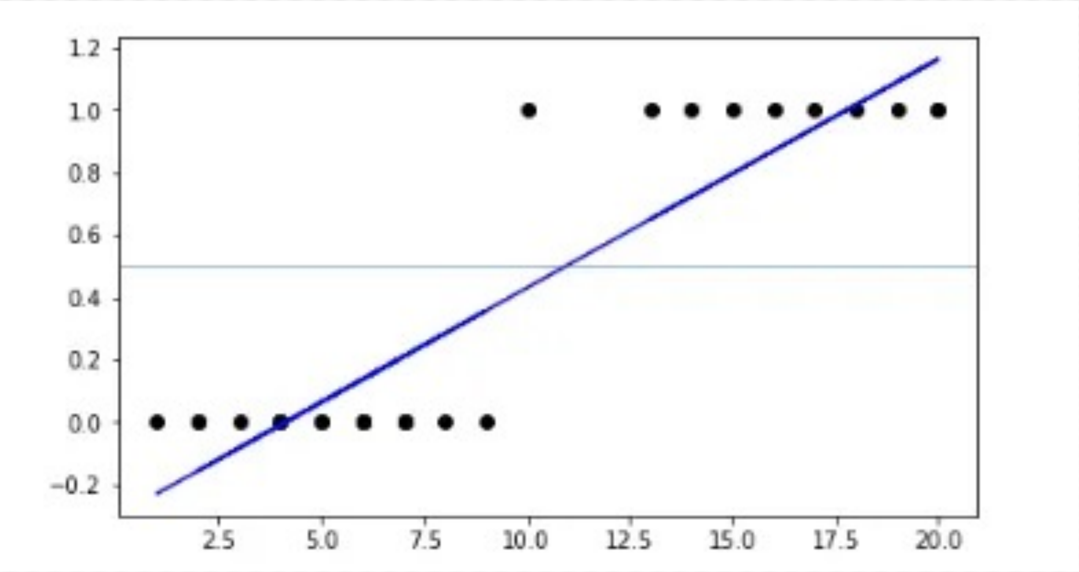
Aarne and Thompson were two folklorists, who invented and perfected the motif- based classification system of folk tales. This system has been in place since 1910 and is commonly used in comparative folkloristics. The final U in ATU stands for Uther, who was the last to update the index in 2004.

Logistic Regression

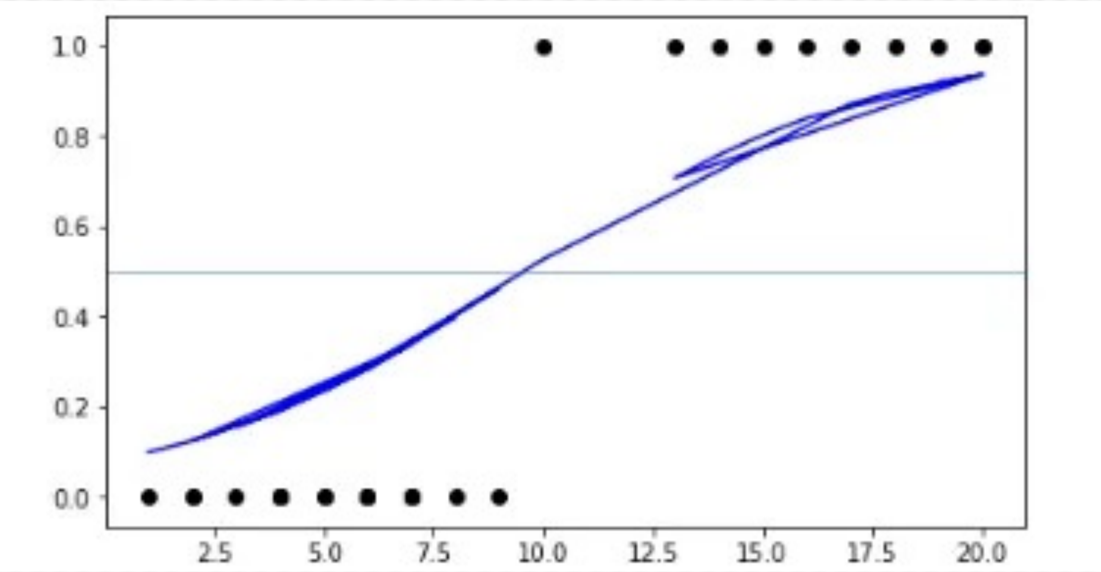
$$p(X) = \beta_0 + \beta_1 X.$$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$



Linear Regression

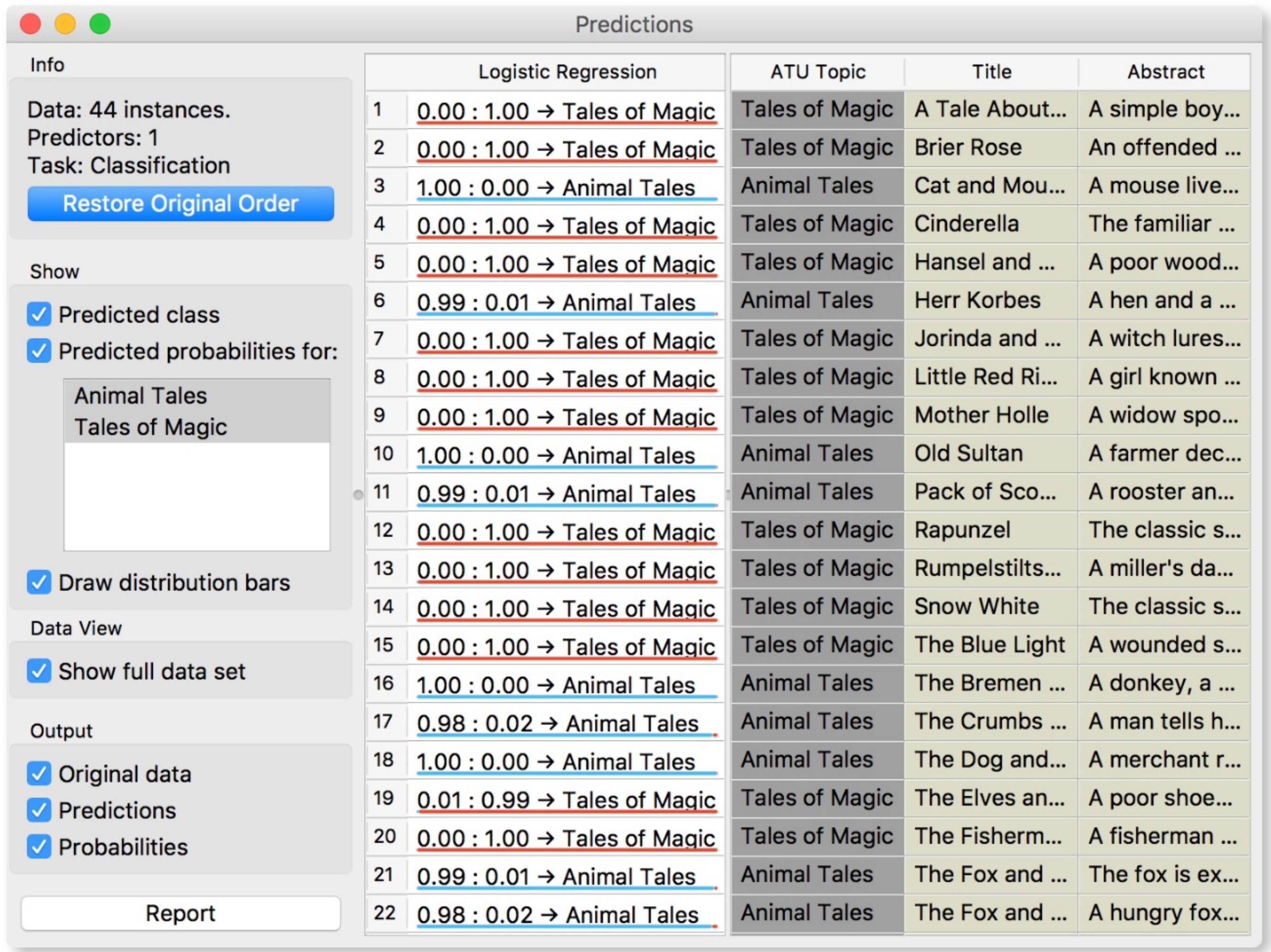


Logistic Regression

https://medium.com/data-science-group-iitr/logistic-regression-simplified-9b4efe801389

Prediction

- Target – an ATU type
- Feature – Numerical representation of each document
- Model – Logistic Regression
- Prediction - a column with predicted values from Logistic Regression



Data: 44 instances.
Predictors: 1
Task: Classification

☑ Predicted class
☑ Predicted probabilities for:

Animal Tales
Tales of Magic

☑ Draw distribution bars

☑ Show full data set

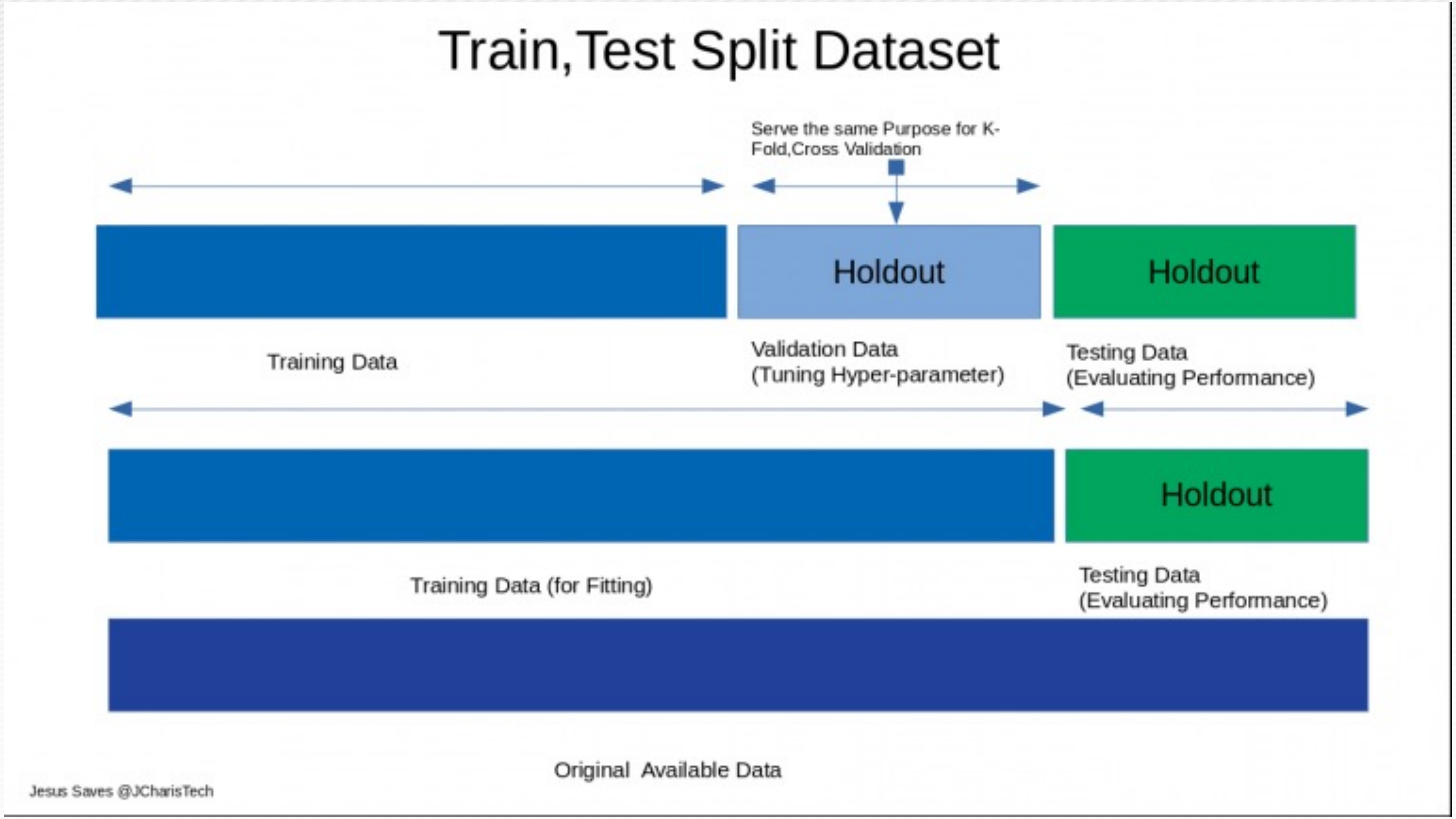
☑ Original data
☑ Predictions
☑ Probabilities

	Logistic Regression	ATU Topic	Title	Abstract
1	0.00 : 1.00 → Tales of Magic	Tales of Magic	A Tale About...	A simple boy...
2	0.00 : 1.00 → Tales of Magic	Tales of Magic	Brier Rose	An offended ...
3	1.00 : 0.00 → Animal Tales	Animal Tales	Cat and Mou...	A mouse live...
4	0.00 : 1.00 → Tales of Magic	Tales of Magic	Cinderella	The familiar ...
5	0.00 : 1.00 → Tales of Magic	Tales of Magic	Hansel and ...	A poor wood...
6	0.99 : 0.01 → Animal Tales	Animal Tales	Herr Korbes	A hen and a ...
7	0.00 : 1.00 → Tales of Magic	Tales of Magic	Jorinda and ...	A witch lures...
8	0.00 : 1.00 → Tales of Magic	Tales of Magic	Little Red Ri...	A girl known ...
9	0.00 : 1.00 → Tales of Magic	Tales of Magic	Mother Holle	A widow spo...
10	1.00 : 0.00 → Animal Tales	Animal Tales	Old Sultan	A farmer dec...
11	0.99 : 0.01 → Animal Tales	Animal Tales	Pack of Sco...	A rooster an...
12	0.00 : 1.00 → Tales of Magic	Tales of Magic	Rapunzel	The classic s...
13	0.00 : 1.00 → Tales of Magic	Tales of Magic	Rumpelstilts...	A miller's da...
14	0.00 : 1.00 → Tales of Magic	Tales of Magic	Snow White	The classic s...
15	0.00 : 1.00 → Tales of Magic	Tales of Magic	The Blue Light	A wounded s...
16	1.00 : 0.00 → Animal Tales	Animal Tales	The Bremen ...	A donkey, a ...
17	0.98 : 0.02 → Animal Tales	Animal Tales	The Crumbs ...	A man tells h...
18	1.00 : 0.00 → Animal Tales	Animal Tales	The Dog and...	A merchant r...
19	0.01 : 0.99 → Tales of Magic	Tales of Magic	The Elves an...	A poor shoe...
20	0.00 : 1.00 → Tales of Magic	Tales of Magic	The Fisherm...	A fisherman ...
21	0.99 : 0.01 → Animal Tales	Animal Tales	The Fox and ...	The fox is ex...
22	0.98 : 0.02 → Animal Tales	Animal Tales	The Fox and ...	A hungry fox...

Part 12
Predictions with test data

4:20-4:30

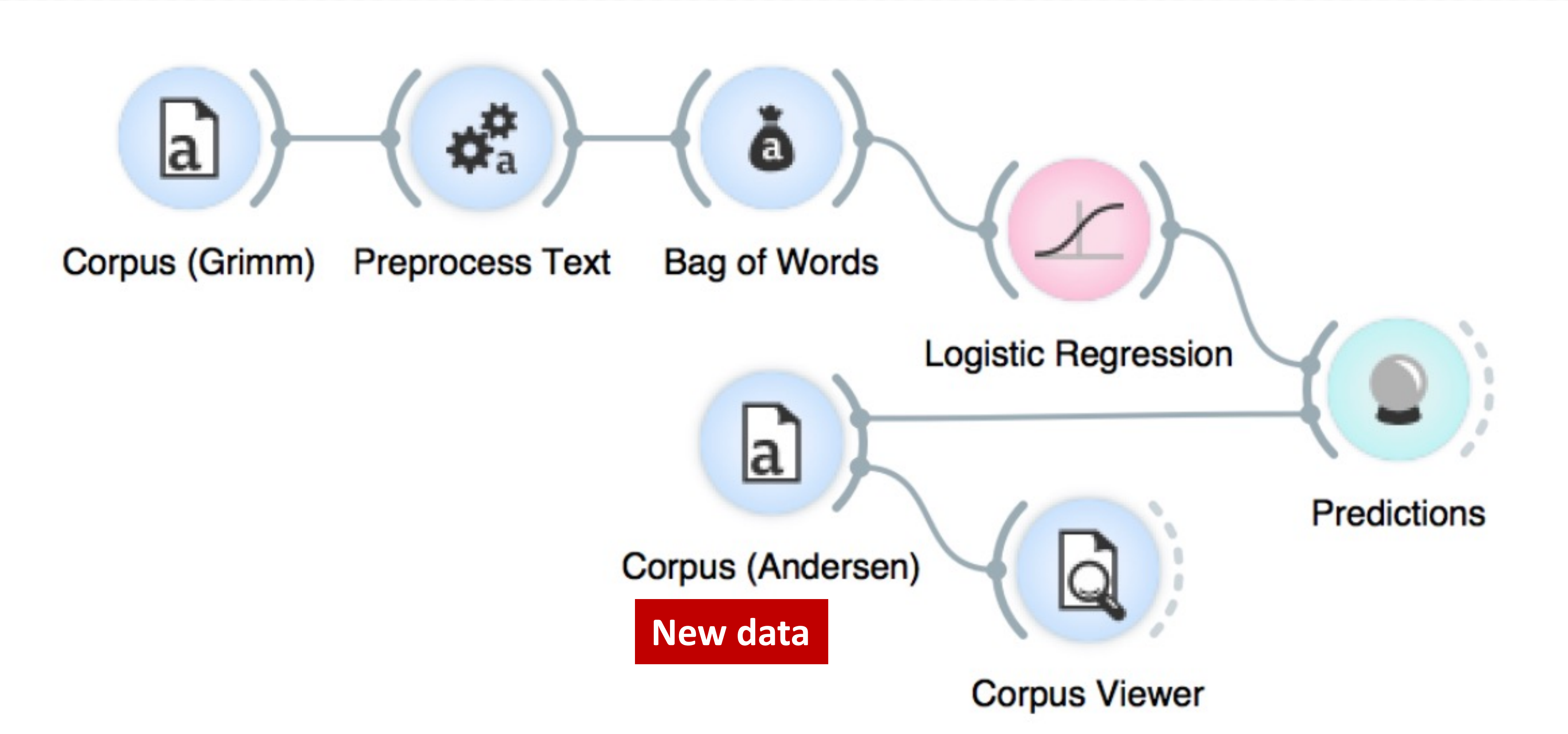
Data splitting



https://blog.jcharistech.com/2020/09/23/how-to-split-dataset-into-training-and-testing-dataset-for-machine-learning/

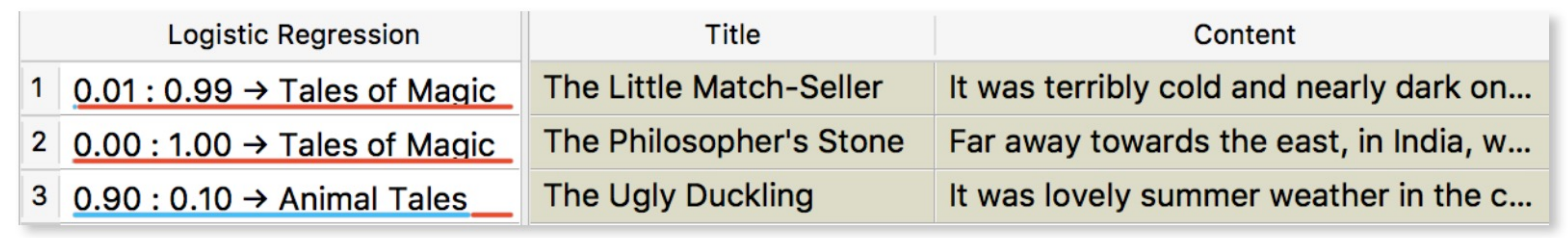
Predicting on new data

- Predicting on new data works just like for regular data.



Guess game – guess the tale type

- Open a new Corpus widget and load the ***andersen.tab*** corpus.
    - Three tales from H. C. Andersen.
- Connect them to Predictions the same way as before - with Logistic Regression passing the constructed model and the new Corpus widget passing the data for prediction.
- Logistic Regression predicted two tales to be Tales of Magic and one the Animal Tale.



	Logistic Regression	Title	Content
1	0.01 : 0.99 → Tales of Magic	The Little Match-Seller	It was terribly cold and nearly dark on...
2	0.00 : 1.00 → Tales of Magic	The Philosopher's Stone	Far away towards the east, in India, w...
3	0.90 : 0.10 → Animal Tales	The Ugly Duckling	It was lovely summer weather in the c...

Resources

4:30–5:00

References

- Orange Lecture Notes
  - https://orangedatamining.com/blog/2020/2020-02-08-lecture-notes/
- New Video Tutorials on Text Mining
  - https://orangedatamining.com/blog/2020/2020-09-28-text-tutorials/
- Observing Word Distribution
  - https://orangedatamining.com/blog/2021/2021-01-27-word-distribution/
- Machine Learning Jargon
  - https://orangedatamining.com/blog/2022/2022-02-01-machine-learning-jargon/
- Semantic Analysis of Documents
  - https://orangedatamining.com/blog/2021/2021-09-17-semantic-analysis/
- PCA vs. MDS vs. t-SNE
  - https://orangedatamining.com/blog/2021/2021-06-17-pca-mds-tsne/
- How to identify fake news with document embeddings
  - https://orangedatamining.com/blog/2020/2020-10-15-document-embedders/
- Detecting Story Arcs with Orange
  - https://orangedatamining.com/blog/2020/2020-07-27-story-arcs/