

Künstliche Intelligenz

... kennenlernen

... ausprobieren

... selber machen

REFERENTEN:

DR.-ING. ANNE GUTSCHMIDT

M.SC. HANNES GRUNERT

01.11.2021

KURZVORSTELLUNG DER REFERENTEN

Dr.-Ing. Anne Gutschmidt



Koordinatorin
Zentrum für Künstliche
Intelligenz in MV
Tel. 0381/4987435
anne.gutschmidt@uni-rostock.de

M.Sc. Hannes Grunert



Wissenschaftlicher Mitarbeiter
Zentrum für Künstliche
Intelligenz in MV
Tel. 0381/4987436
hannes.grunert@uni-rostock.de

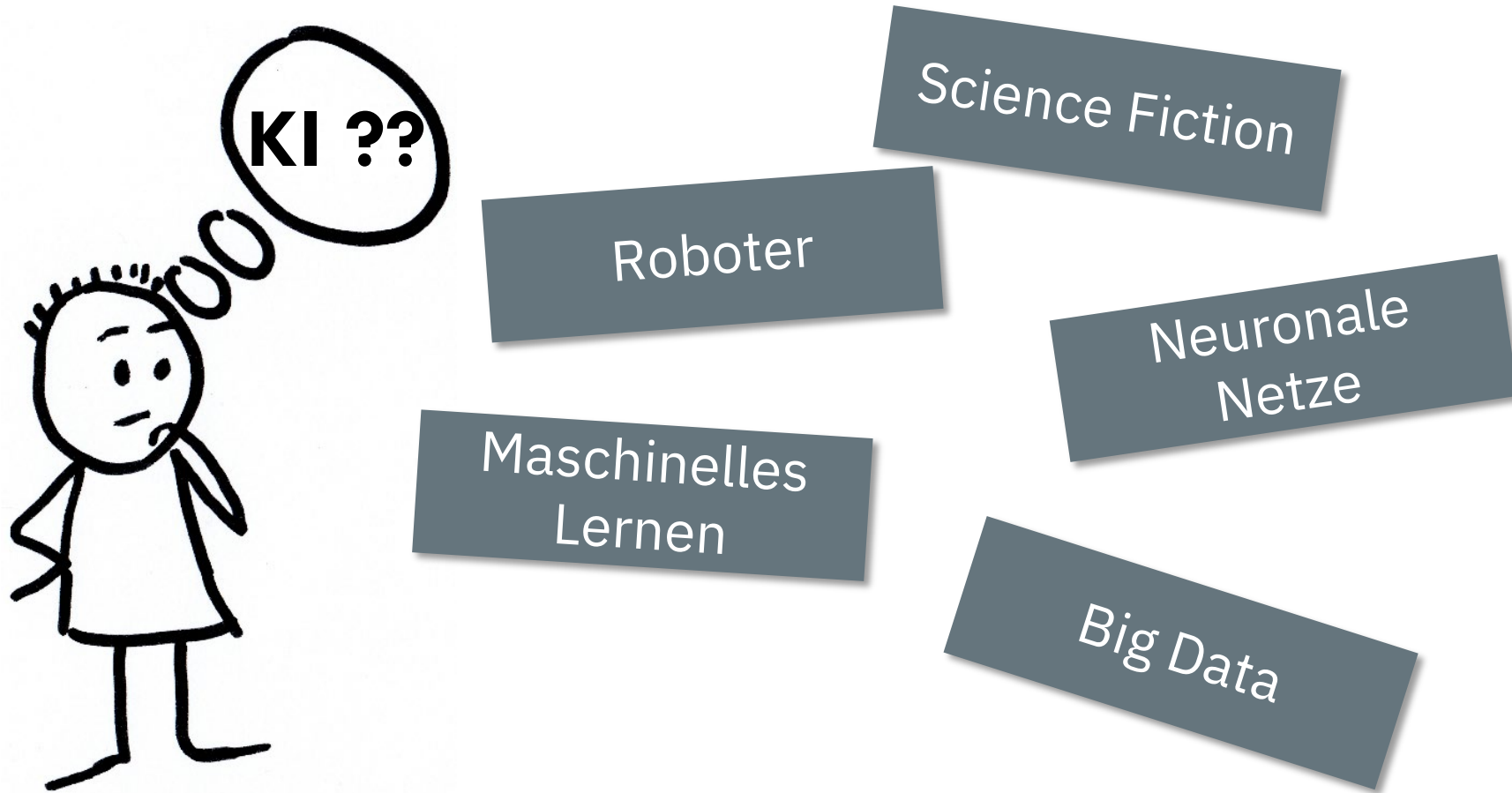


www.ki-mv.de

KÜNSTLICHE INTELLIGENZ KENNENLERNEN

1. Was ist künstliche Intelligenz?
2. Begriffe aus der KI-Welt
3. Herausforderungen beim Einsatz von KI

WAS FÄLLT IHNEN ZU KÜNSTLICHER INTELLIGENZ EIN?



DIGITALISIERUNG UND KI

Digitalisierung → Sensordaten, Textdaten, Transaktionsdaten, ...

- ⇒ Angesichts der Datenmengen bzw. der Komplexität eines Problems keine (herkömmliche) Lösung möglich
- ⇒ Einsatz künstlicher Intelligenz bei scheinbar unlösbaren oder schwer lösbaren Problemen
- ⇒ KI kann: suchen, schlussfolgern, Probleme lösen, wahrnehmen, lernen, schätzen, analytisch denken, optimieren und planen

KI als wichtiger Baustein, der die digitale Transformation vorantreibt

GEBURTSTUNDE DER „ARTIFICIAL INTELLIGENCE“

Dartmouth Summer Research Project on Artificial Intelligence 1956

Initiiert von John McCarthy:

*“Diese Untersuchung ist auf der Grundlage der Vermutung fortzuführen, dass **alle Aspekte des Lernens oder anderer Merkmale der Intelligenz im Prinzip so genau beschrieben werden können, dass sich eine Maschine bauen lässt, um sie zu simulieren.**”*

(aus dem Aufruf zum Workshop)



John McCarthy

"null0"

(https://commons.wikimedia.org/wiki/File:John_McCarthy_Stanford.jpg), „John McCarthy Stanford“,

<https://creativecommons.org/licenses/by-sa/2.0/legalcode>

WAS IST KI?

Schwache KI ...

löst Probleme eines bestimmten Anwendungsbereiches



Assistenzsysteme



Autonomes Fahren



Industrie 4.0: Qualitätsüberwachung, Wartungsassistentz etc.



Parkleitsysteme



Empfehlungssysteme

Starke KI ...

soll folgendes können:

- Logisch denken, Strategien anwenden, Rätsel lösen
- Entscheiden unter Unsicherheit
- Wissen darstellen, inkl. Allgemeinwissen
- Planen und lernen
- In natürlicher Sprache kommunizieren
- All diese Fähigkeiten zusammen nutzen, um Ziele zu erreichen
- Bewusstsein

BEGRIFFE AUS DER KI-WELT

- Machine Learning
- Neuronale Netze
- Deep Learning
- Symbolische KI

BEGRIFFE AUS DER KI-WELT

- **Machine Learning**
- Neuronale Netze
- Deep Learning
- Symbolische KI

Anhand von Trainingsbeispielen lernen Algorithmen Muster und Entscheidungsregeln, so dass sie neue Datensätze klassifizieren/einordnen können

Supervised Learning

Klassen sind in Trainingsbeispielen vorgegeben

Unsupervised Learning

Keine Klassen vorgegeben, z.B. Clustering

Reinforcement Learning

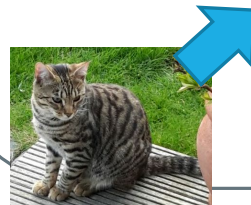
Aus Erfahrung / durch Belohnung lernen

BEGRIFFE AUS DER KI-WELT

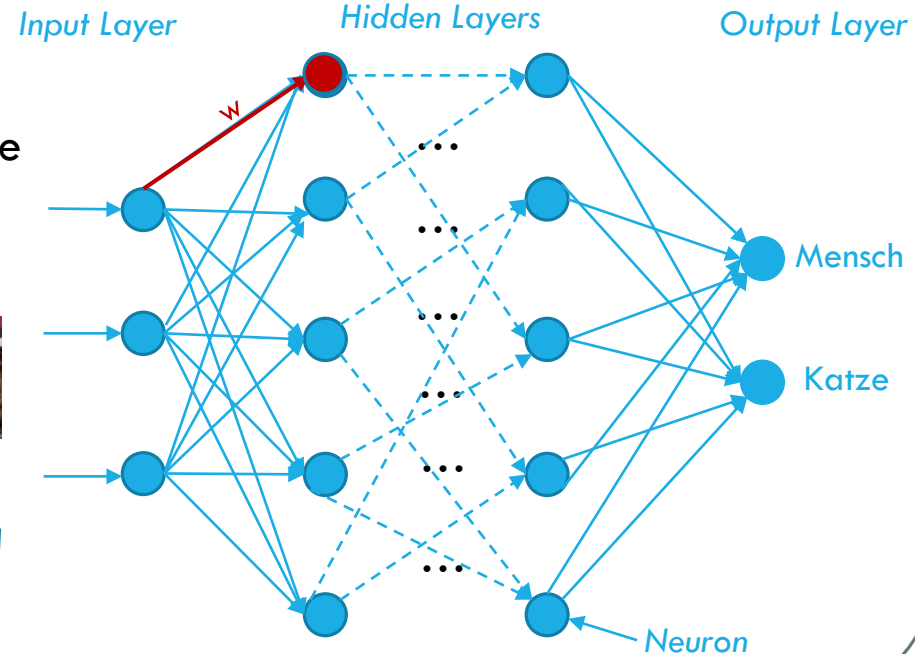
- Machine Learning
- **Neuronale Netze**
- Deep Learning
- Symbolische KI

Ein spezielles Verfahren des Maschinellen Lernens angelehnt an biologische neuronale Netze

Trainingsbeispiele



Neuer Datensatz



BEGRIFFE AUS DER KI-WELT

- Machine Learning
- Neuronale Netze
- **Deep Learning**
- Symbolische KI

Spezielle neuronale Netze, bei denen
sehr viele Hidden Layers eingesetzt
werden;
Sehr rechenintensiv!

BEGRIFFE AUS DER KI-WELT

- Machine Learning
- Neuronale Netze
- Deep Learning
- **Symbolische KI**

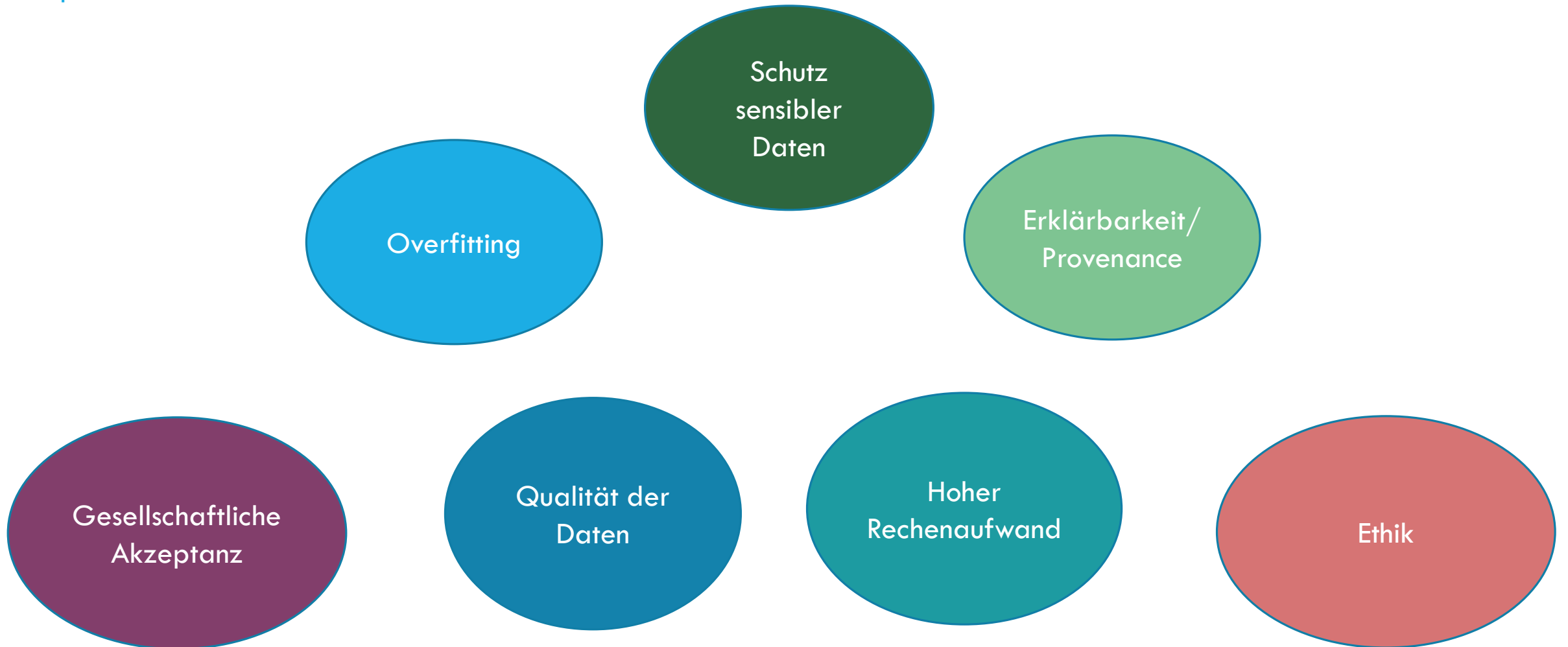
Wissensbasierte Ansätze

- Regeln und Beziehungen zwischen Konzepten
- Für Menschen nachvollziehbar

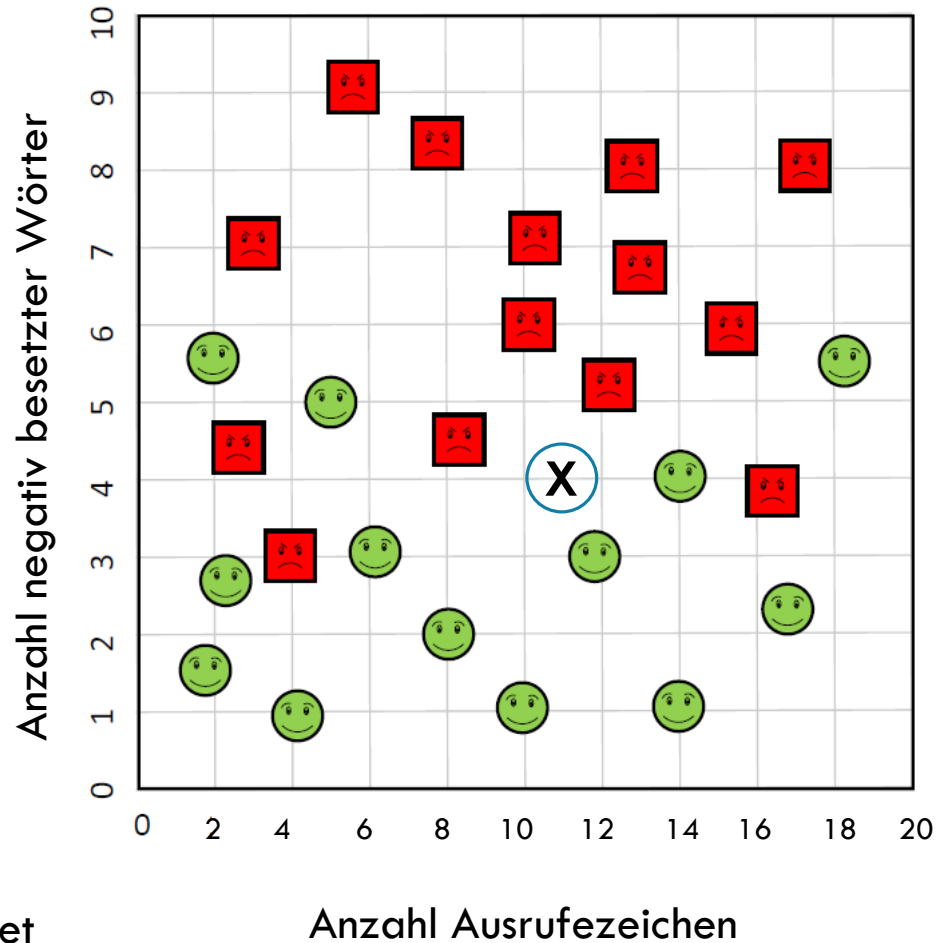
ZWISCHENDURCH KURZ ZUSAMMENGEFASST

- KI löst komplexe Probleme
- „Starke KI“ → Science Fiction
- Vielzahl an Verfahren im Bereich der KI

HERAUSFORDERUNGEN IN DER KI



HERAUSFORDERUNGEN IN DER KI – ETHIK



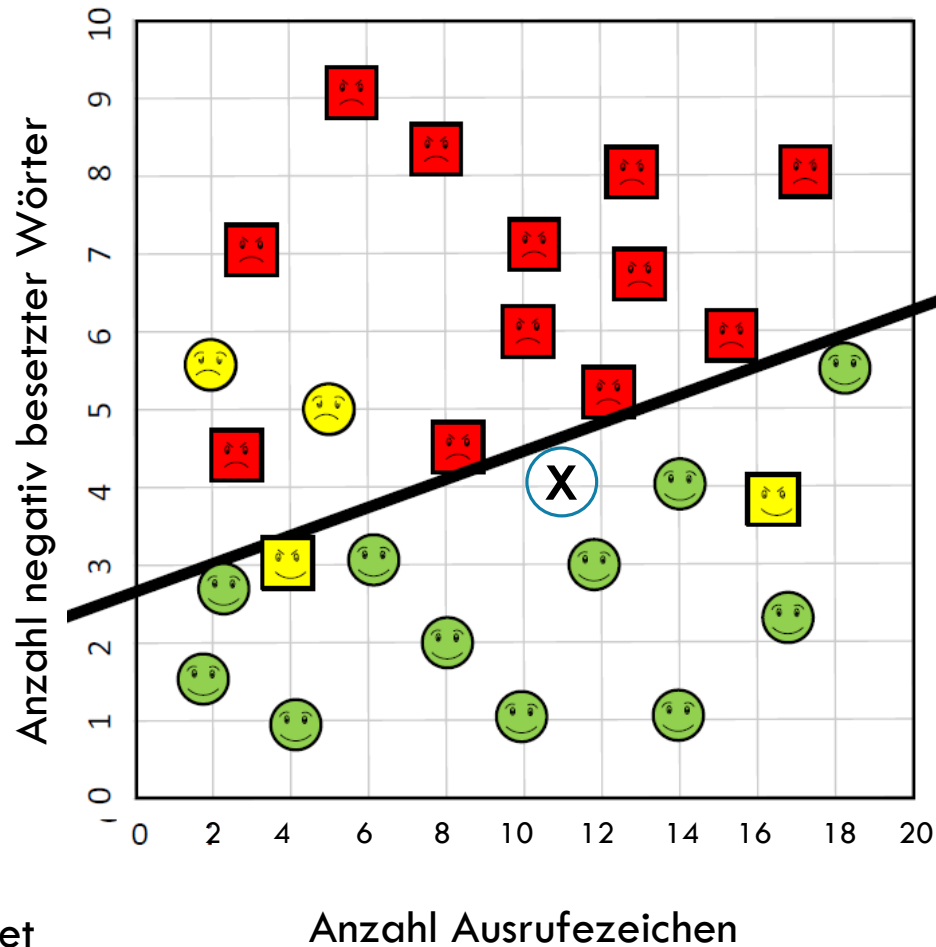
Beispiel von Prof. K.A. Zweig

Bewerten Sie den neuen Tweet von Frau Müller:
4 negativ besetzte Wörter
11 Ausrufezeichen

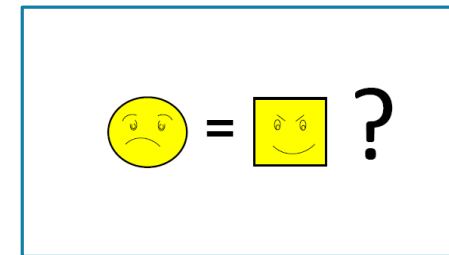
 Hass-Tweet

 Normaler Tweet

HERAUSFORDERUNGEN IN DER KI – ETHIK



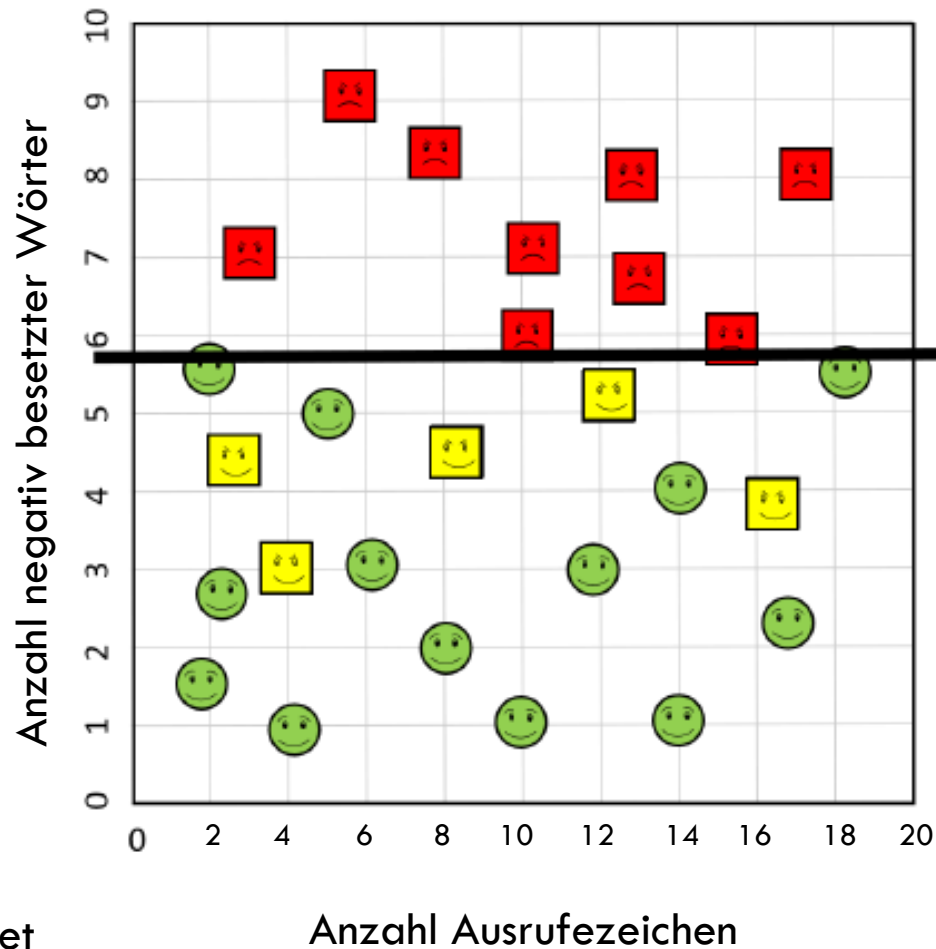
Beispiel von Prof. K.A. Zweig



 Hass-Tweet

 Normaler Tweet

HERAUSFORDERUNGEN IN DER KI – ETHIK



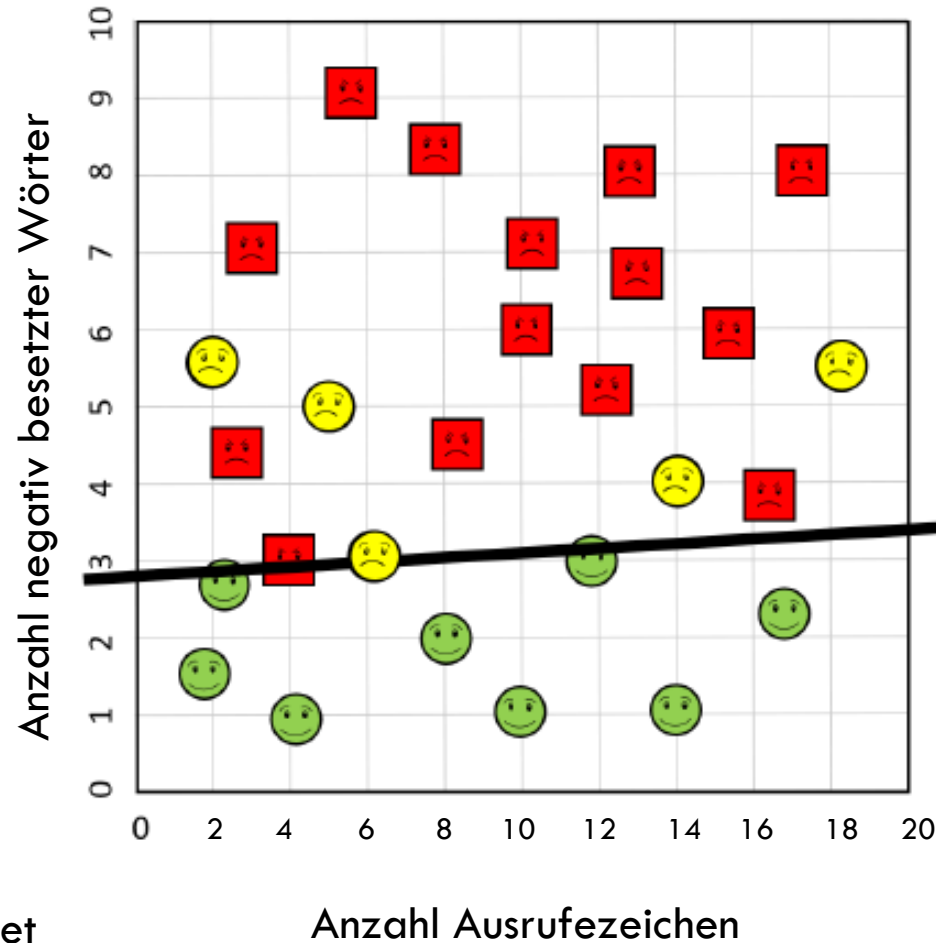
Beispiel von Prof. K.A. Zweig

Anbieter will sich nicht
Zensur vorwerfen lassen

 Hass-Tweet

 Normaler Tweet

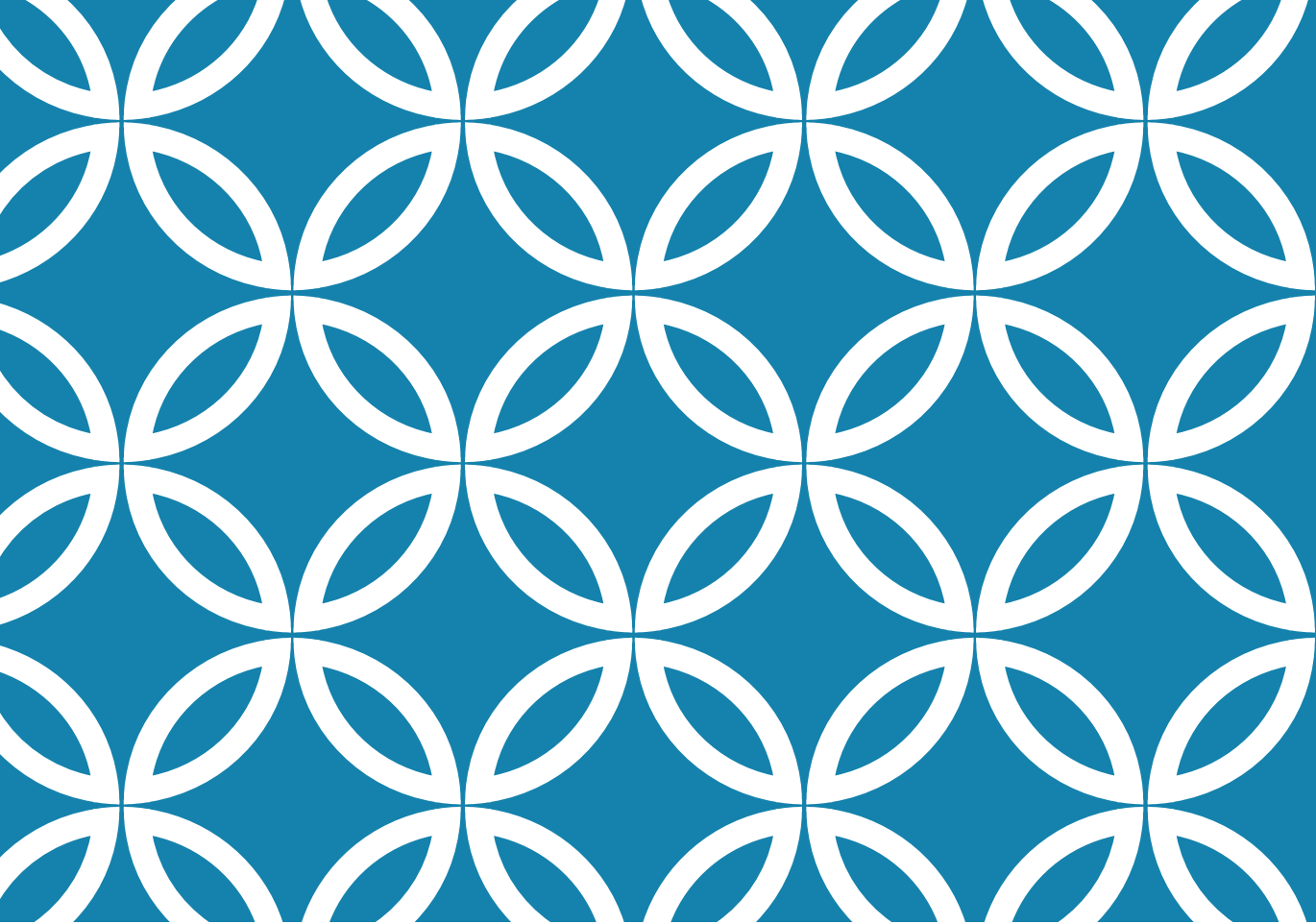
HERAUSFORDERUNGEN IN DER KI – ETHIK



Beispiel von Prof. K.A. Zweig

Auf Nummer Sicher gehen:
Hassbotschaften werden
herausgefiltert, normale Tweets
kann es fälschlicherweise auch
treffen

Was durch KI optimiert werden soll,
entscheiden wir!

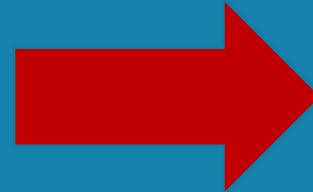


Künstliche Intelligenz

... kennenlernen

... ausprobieren

... selber machen



REFERENTEN:

DR.-ING. ANNE GUTSCHMIDT

M.SC. HANNES GRUNERT

01.11.2021

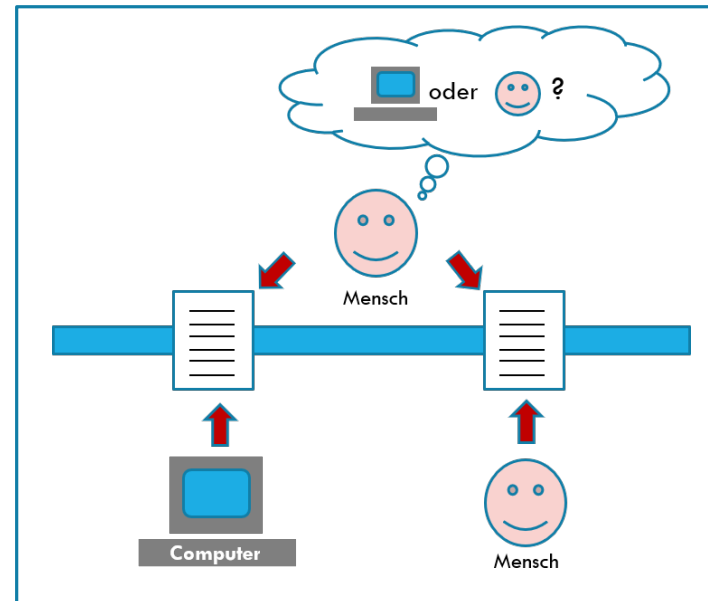
KÜNSTLICHE INTELLIGENZ AUSPROBIEREN

1. Bot or not? Ein kleiner Turing-Test
2. Text in Sprache umwandeln – Watson Text to Speech Voices
3. Sentiment-Analyse mit Orange Data Mining

BOT OR NOT – EIN KLEINER TURING-TEST

Der Turing-Test

- Mensch kommuniziert über Textnachrichten mit einem Menschen und einem Computer
- Kann der Mensch nicht entscheiden, wer „hinter der Wand“ der Mensch und wer die Maschine ist, hat die Maschine den Test bestanden



Ausprobieren:
<https://botor.no/>

PLAYING WITH DAVID 00:06

hey

hey Davis?

hi hi

are you david hasselhoff?

haha

do you know him?

what's the weather like where you are

cloudy

and your place?

Say something...

TEXT IN SPRACHE UMWANDELN – WATSON TEXT TO SPEECH VOICES

IBM Watson Text to Speech Demo

Interested in Watson Text to Speech? [Get Started on IBM Cloud](#)

Watson Text to Speech Voices

Listen to voices across languages and dialects

Language: German | Neural voice: Dieter

Use the sample text or enter your own text in German

Wir zieh'n durch die Straßen und die Clubs dieser Stadt
Das ist unsere Nacht, wie für uns beide gemacht, oho oho
Ich schließe meine Augen, lösche jedes Tabu
Küsse auf der Haut, so wie ein Liebes-Tattoo, oho, oho

Was das zwischen uns auch ist
Bilder die man nie vergisst
Und dein Blick hat mir gezeigt

Adjusted to 0.9x default speed | Adjust pitch: default | Play voice

0.2x ————— 1.7x

What is a Neural Voice?
By using Deep Neural Networks trained on human speech, Watson can produce natural-sounding and smooth voice quality.
[Learn more](#) about the science behind the service

Custom Voice Training
To distinguish your brand, work with IBM to train a voice that suits your distinct style with as little as one hour of audio.
[Learn more](#) about creating custom voices

Tune Neural Voices by Example
coming soon
Use your own voice to adjust for misplaced

FEEDBACK

Hintergrundinfos:
<https://cloud.ibm.com/docs/text-to-speech?topic=text-to-speech-science&locale=de>

Ausprobieren: <https://www.ibm.com/demos/live/tts-demo/self-service/home>

EINLEITUNG

In unserem KI-Workshop am MINT-Fachtag möchten wir Ihnen gerne die Möglichkeit geben, selbst KI-Verfahren auszuprobieren. Dazu möchten wir Sie bitten, vor unserer Veranstaltung die Software Orange Data Mining zu installieren. Auf den folgenden Seiten finden Sie eine Anleitung für die Installation.

Wir empfehlen für den Workshop die Nutzung eines zweiten Monitors. Dies ist aber keine Voraussetzung.

INSTALLATION ORANGE DATA MINING

Laden Sie sich die portable Version des Tools herunter

<https://orangedatamining.com/download>

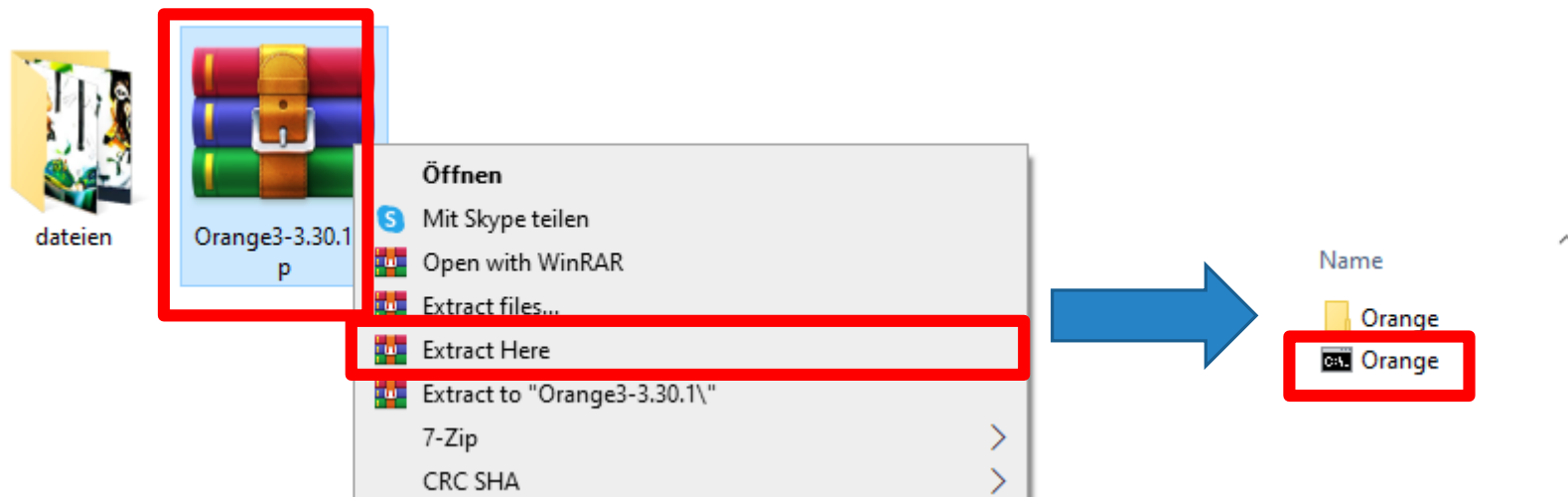
The screenshot shows the Orange Data Mining website's download page. At the top, there is a navigation bar with links for Screenshots, Workflows, Download, Blog, Docs, Workshops, and a search icon. Below the navigation bar, there are three platform icons: Windows, macOS, and Linux / Source. The Windows section is highlighted, and a blue arrow points to the 'Download Orange 3.30.1' button. Below this button, there are three download options: 'Standalone installer (default)', 'Orange3-3.30.1-Miniconda-x86_64.exe (64 bit)', and 'Portable Orange Orange3-3.30.1.zip'. The 'Portable Orange' option is highlighted with a red box and a blue arrow. Below the download options, there is a section for Anaconda with a code block containing the command: `conda config --add channels conda-forge`.

The screenshot shows a Firefox file dialog box titled 'Opening Orange3-3.30.1.zip'. The dialog box displays the file name 'Orange3-3.30.1.zip' and its size '472 MB'. Below the file name, there is a dropdown menu for 'What should Firefox do with this file?'. The 'Save File' option is selected and highlighted with a red box. The 'OK' button is also highlighted with a red box. A blue arrow points from the 'OK' button in the dialog box to the download bar in the next screenshot.

The screenshot shows a browser download bar with three items: 'Orange3-3.30.1.zip' (472 MB), '20211023.pdf' (15.3 MB), and '20211025.pdf' (8.1 MB). The 'Orange3-3.30.1.zip' item is highlighted with a red box. A blue arrow points from the 'OK' button in the previous screenshot to this download bar.

INSTALLATION ORANGE DATA MINING

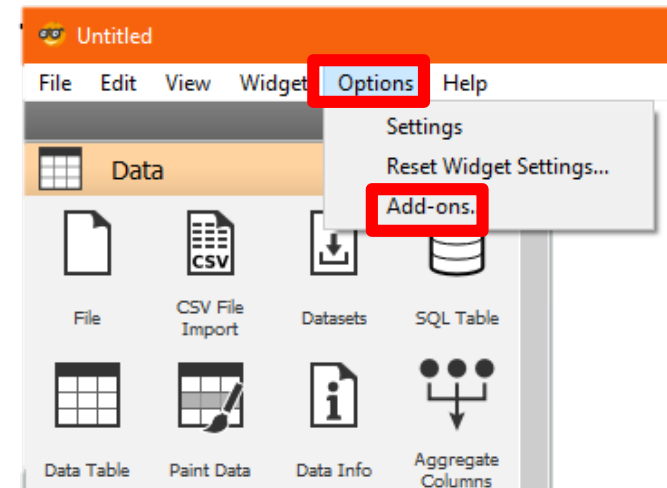
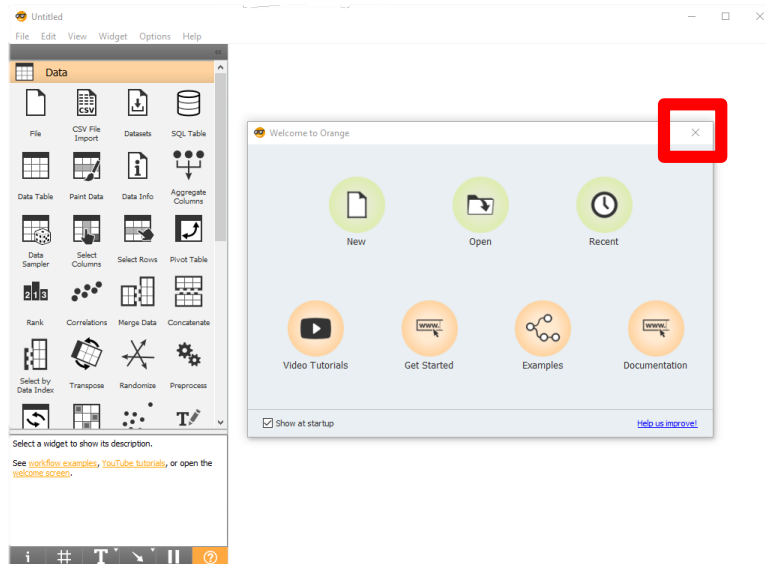
Extrahieren Sie das ZIP-Archiv und starten Sie die Anwendung



ERWEITERUNGEN HINZUFÜGEN

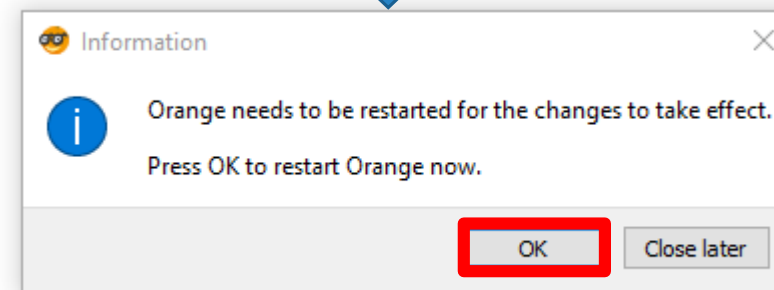
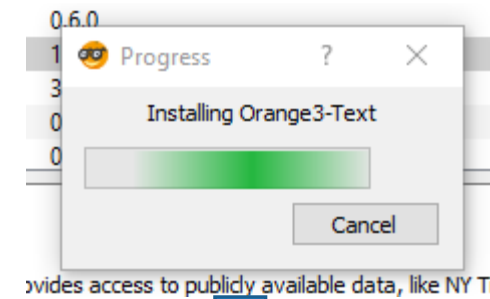
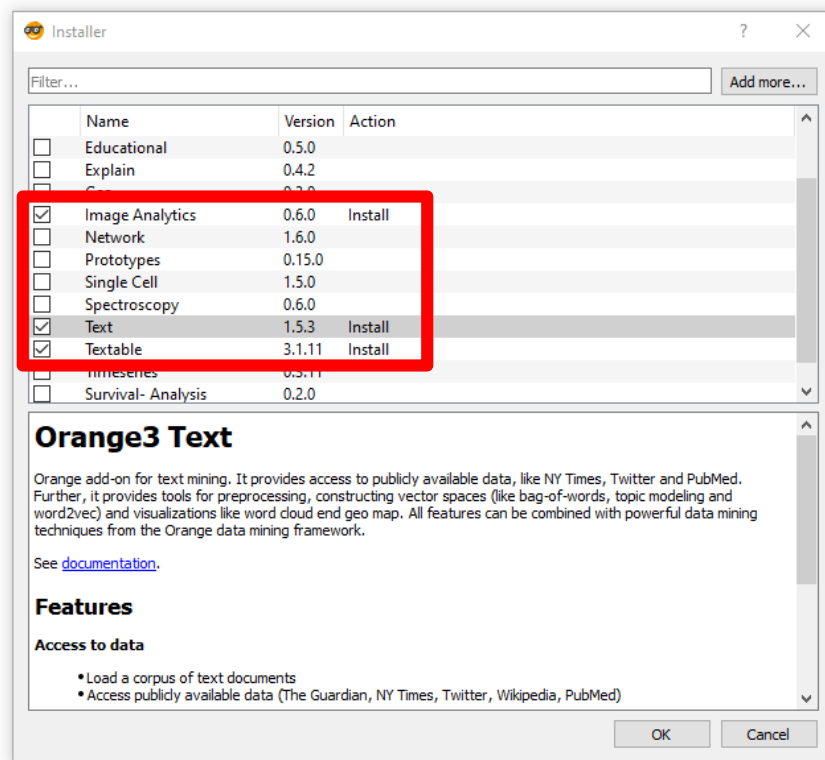
Schließen Sie die Willkommenseite und starten Sie die Anwendung

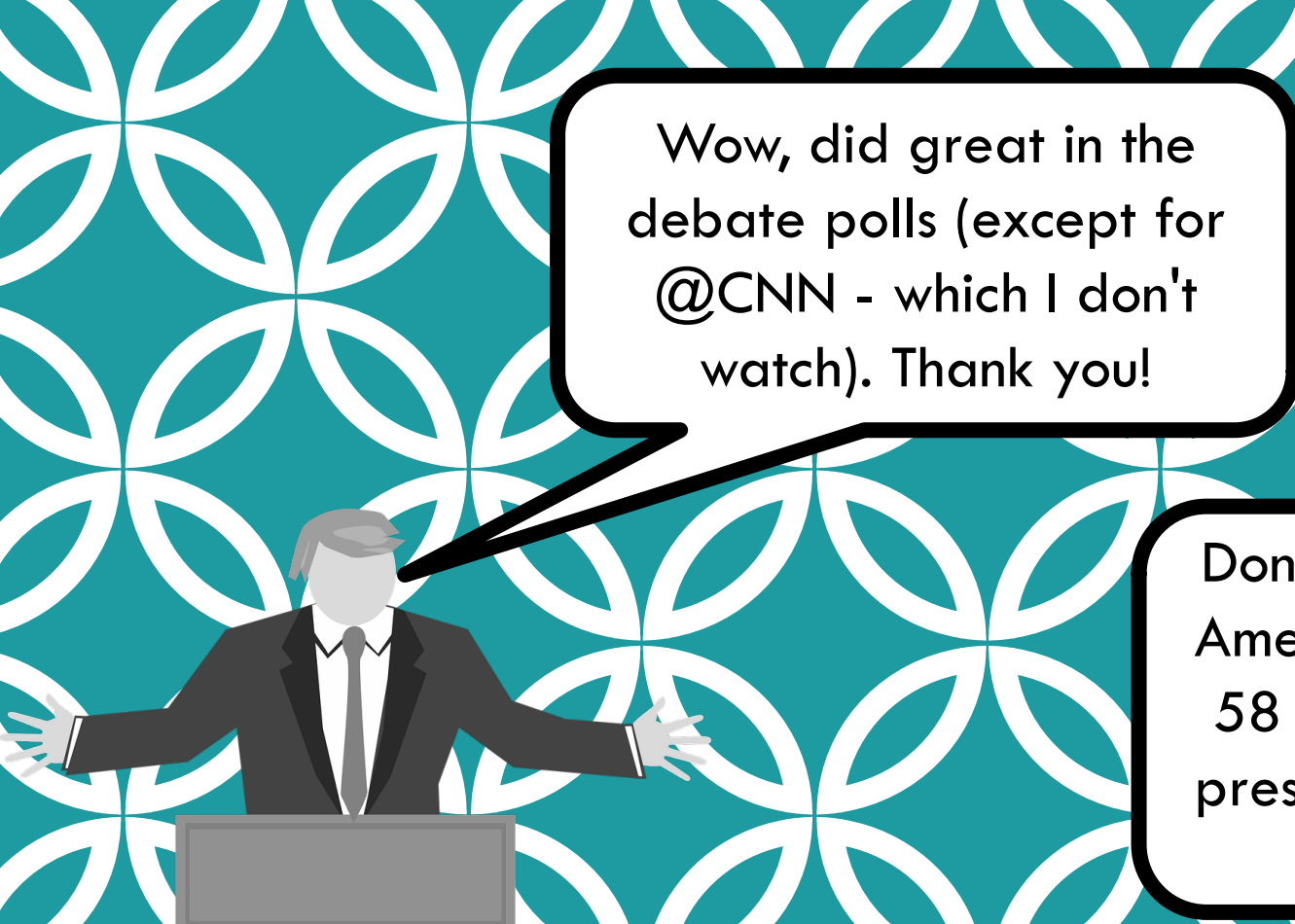
Öffnen Sie das Add-On-Menü



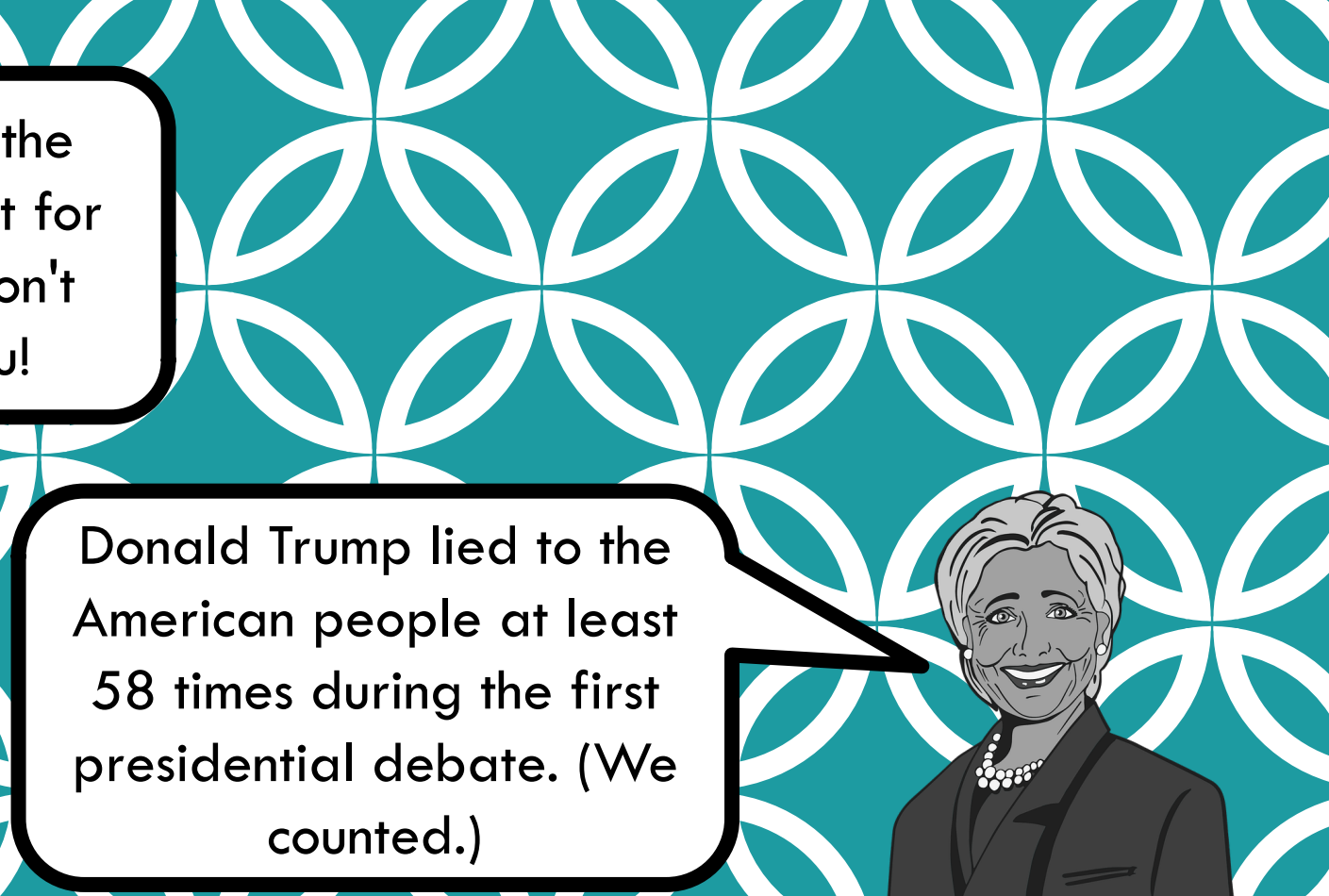
ERWEITERUNGEN HINZUFÜGEN

Image Analytics, Text und Textable auswählen





Wow, did great in the debate polls (except for @CNN - which I don't watch). Thank you!



Donald Trump lied to the American people at least 58 times during the first presidential debate. (We counted.)

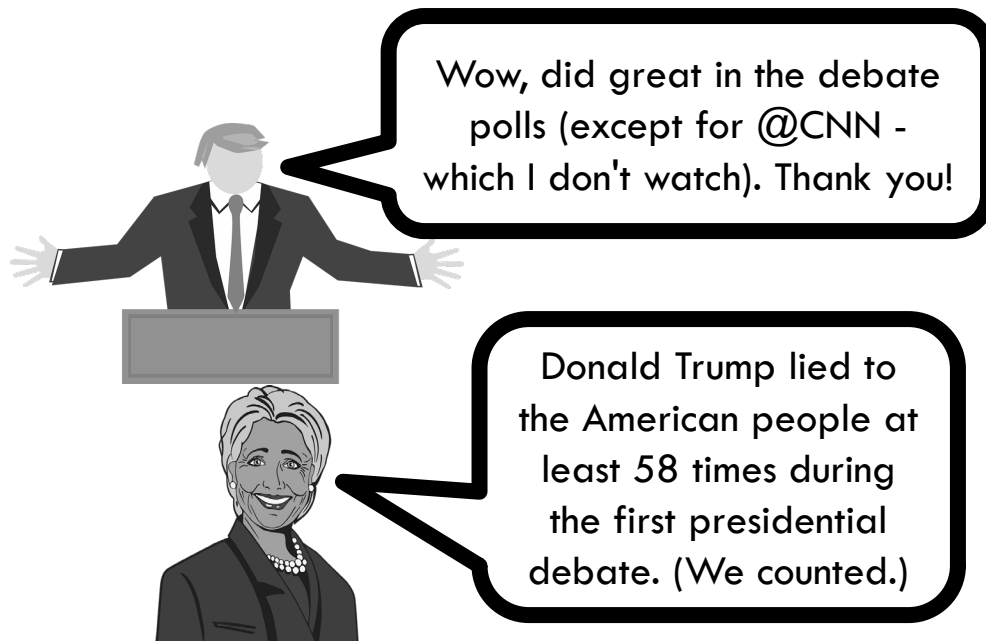
SENTIMENT-ANALYSE

Beispiel I: Tweets von Hillary Clinton und Donald Trump im Wahlkampf 2016

BEISPIEL I: SENTIMENT-ANALYSE

Standardaufgabe bei der Analyse von Texten

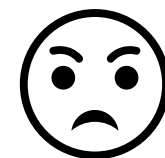
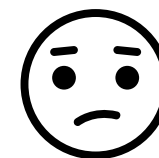
Erkennung von Stimmungen und Emotionen in Texten



Stimmung



Emotionen

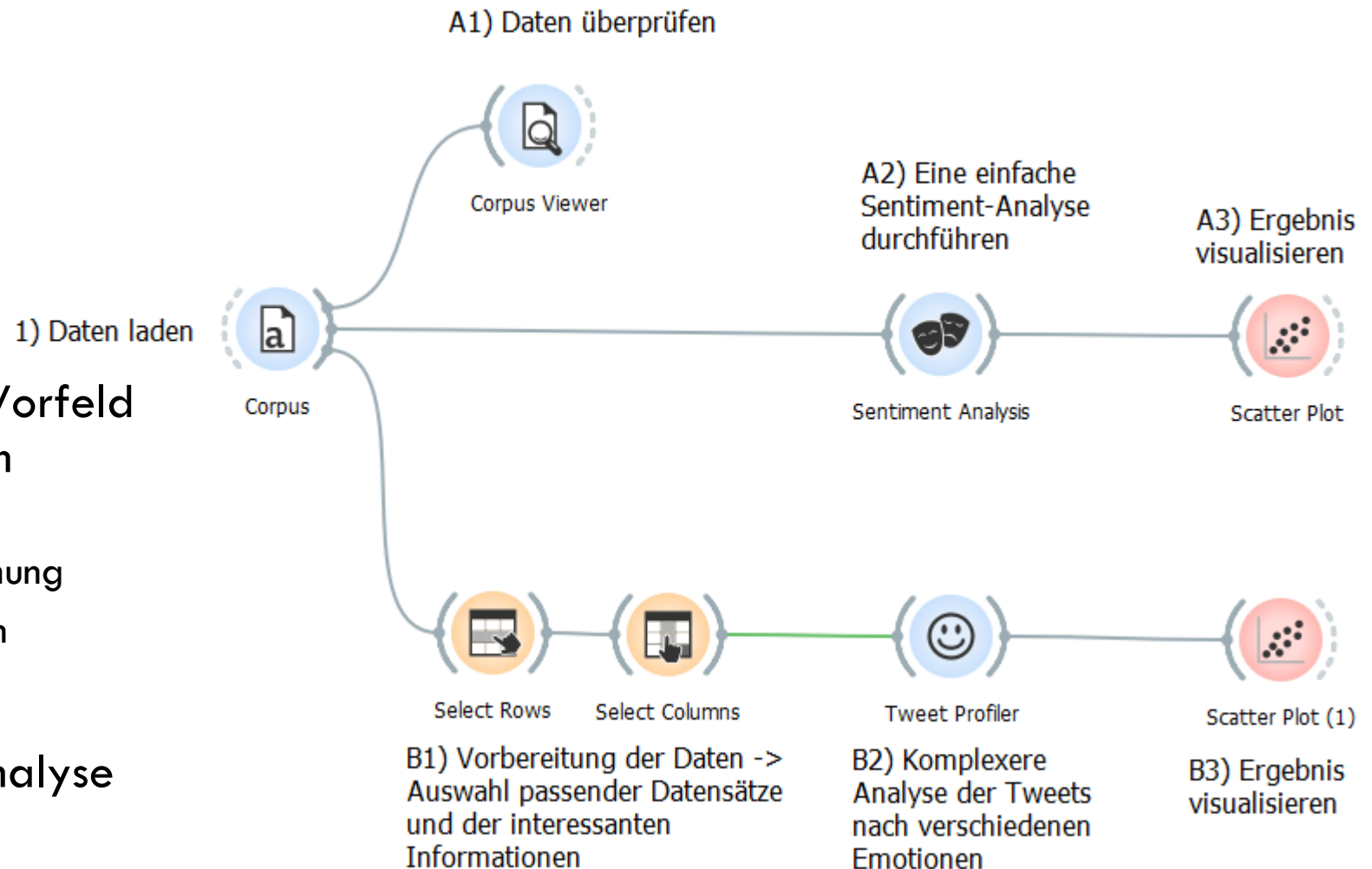


BEISPIEL I: SENTIMENT-ANALYSE

Wir nutzen zwei im Vorfeld erstellte Modelle zum Auswerten

- Positive/negative Stimmung
- Verschiedene Emotionen

Zum Abschluss:
Visualisierung der Analyse

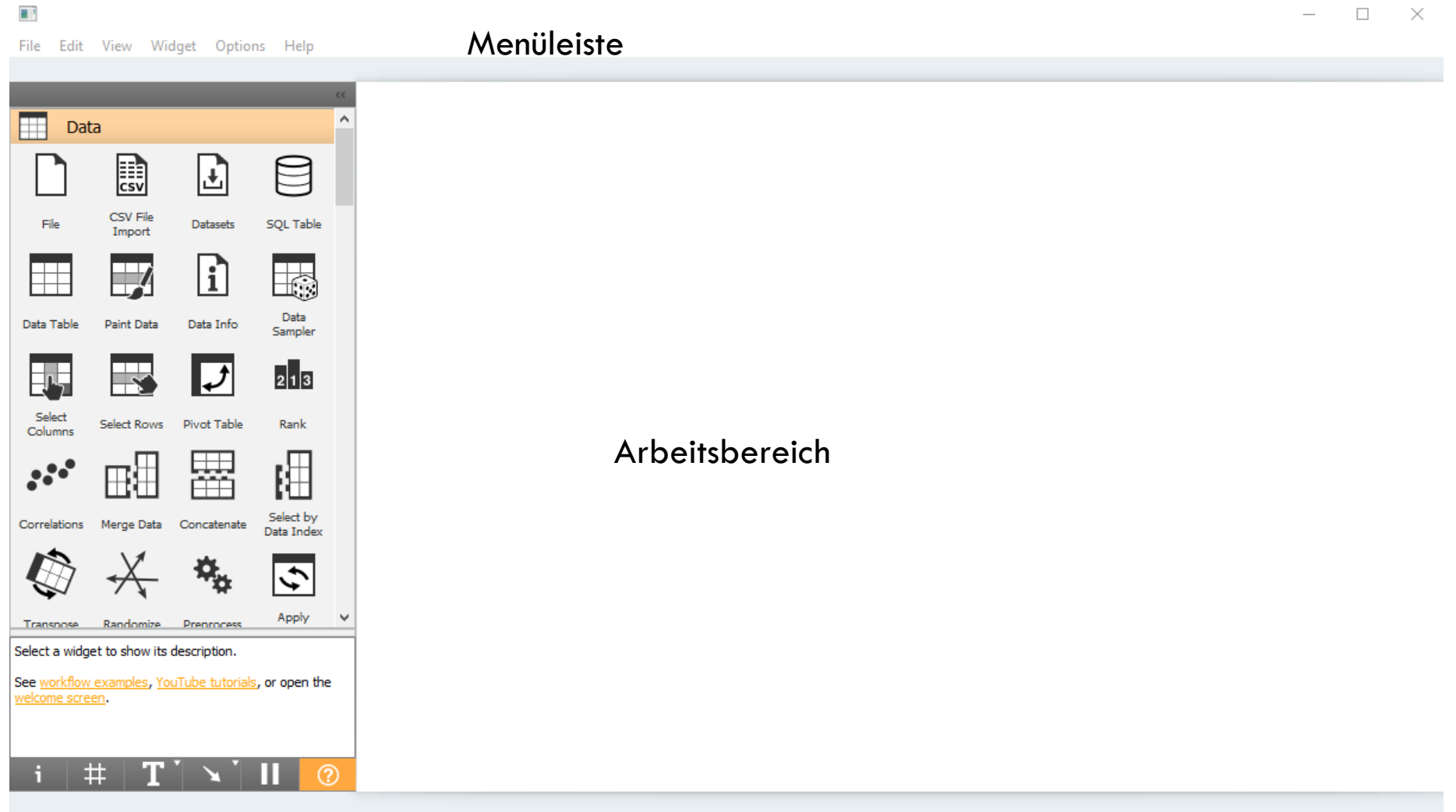


ÜBERBLICK ZU ORANGE DATA MINING

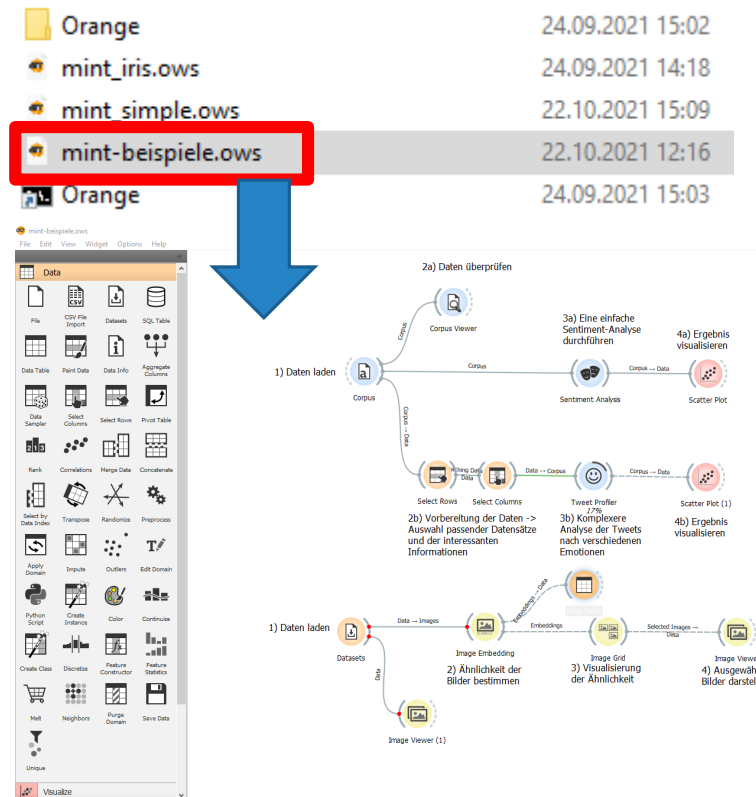
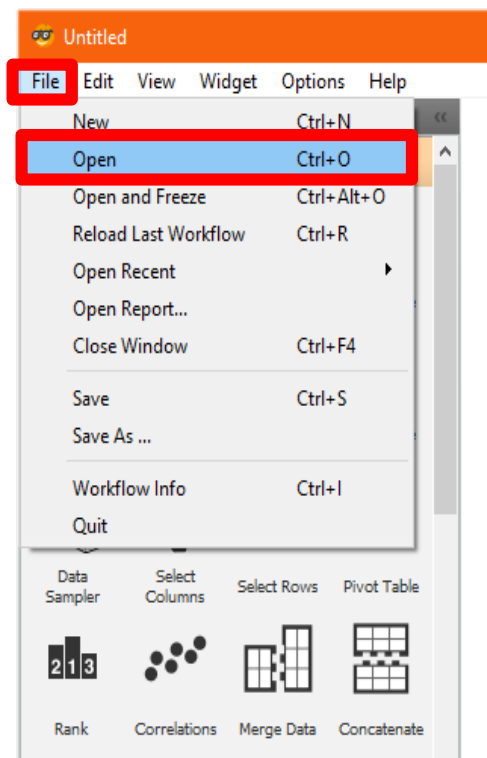
Widgets: Prozesse zum

- Datenimport
- Vorverarbeitung
- Modellieren
- Visualisieren

Informationen zum Prozess



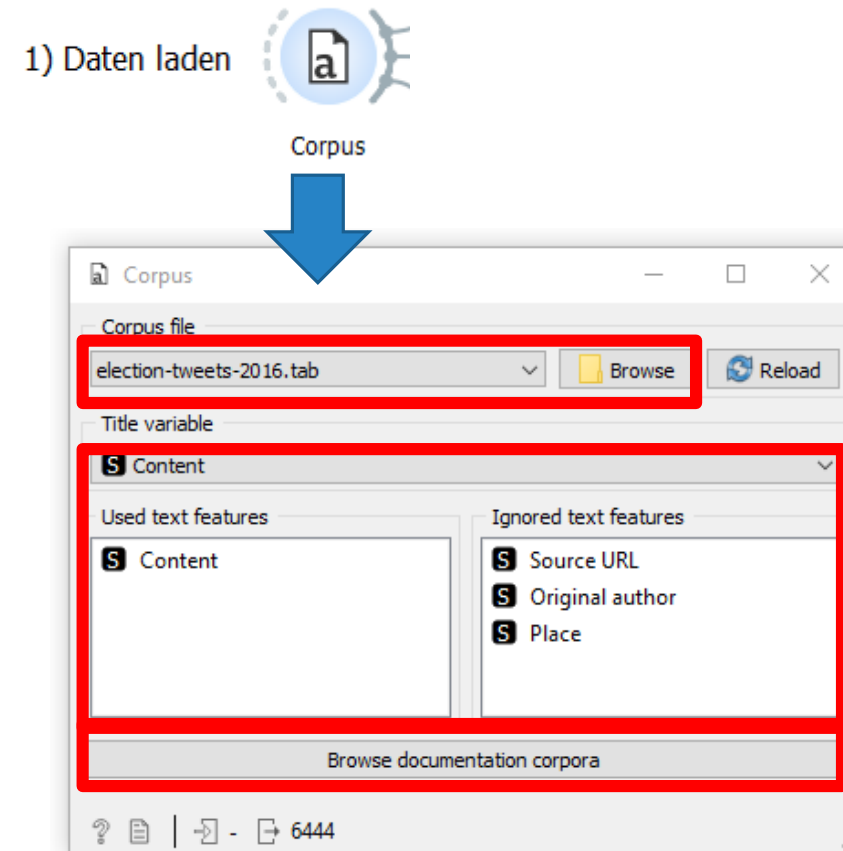
BEISPIEL LADEN



BEISPIEL I: SENTIMENT-ANALYSE

Schritt 1: Laden der Daten

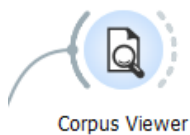
- Datei: „election-tweets-2016.tab“
 - Entweder über „Browse“ auswählen
 - Oder über „Browse documentation corpora“
- Title variable: „Content“
 - Für Anzeige im Corpus Viewer
- Used text features: „Content“
 - Für die eigentliche Analyse



BEISPIEL I: SENTIMENT-ANALYSE

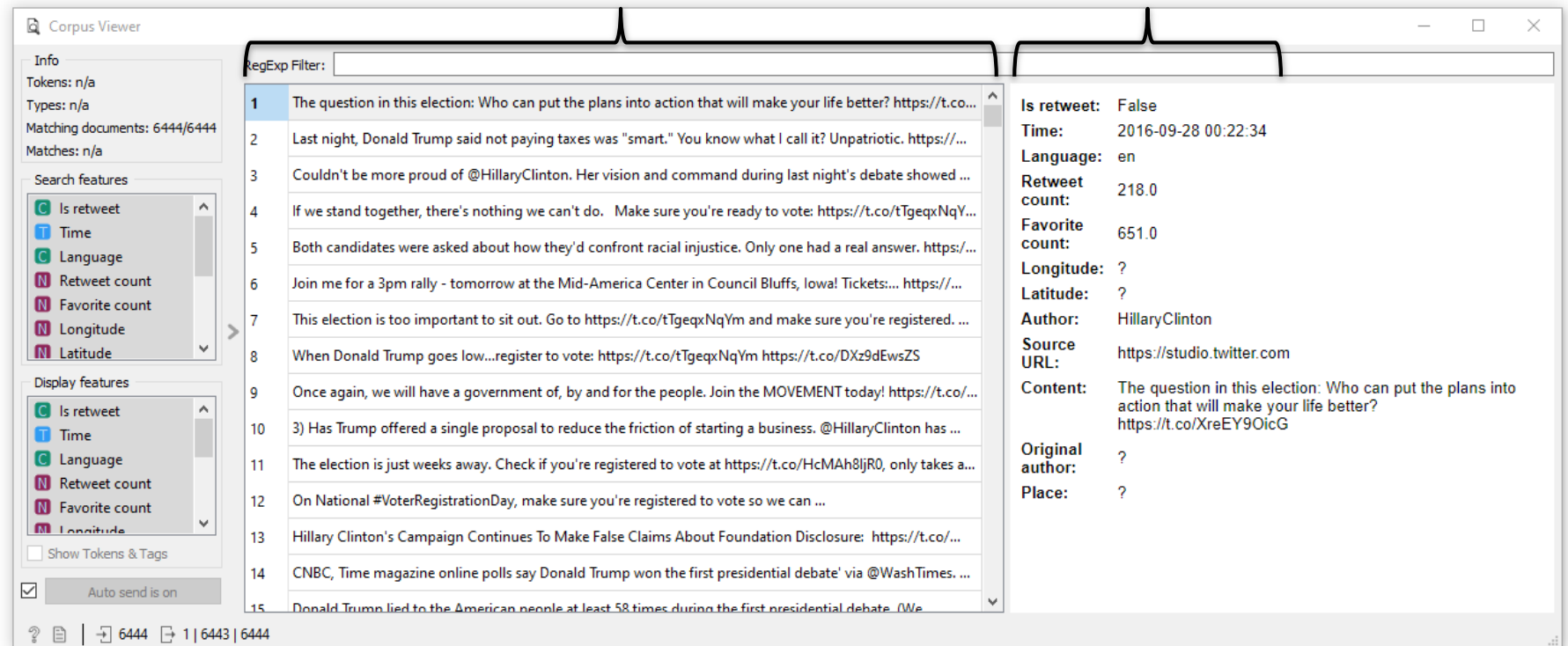
Schritt 2a: Daten überprüfen

2a) Daten überprüfen



Der eigentliche Tweet

Weitere Daten



Info

Tokens: n/a
Types: n/a
Matching documents: 6444/6444
Matches: n/a

Search features

- Is retweet
- Time
- Language
- Retweet count
- Favorite count
- Longitude
- Latitude

Display features

- Is retweet
- Time
- Language
- Retweet count
- Favorite count
- Longitude

Show Tokens & Tags

Auto send is on

RegExp Filter:

1 The question in this election: Who can put the plans into action that will make your life better? https://t.co...

2 Last night, Donald Trump said not paying taxes was "smart." You know what I call it? Unpatriotic. https://...

3 Couldn't be more proud of @HillaryClinton. Her vision and command during last night's debate showed ...

4 If we stand together, there's nothing we can't do. Make sure you're ready to vote: https://t.co/tTgeqxNqY...

5 Both candidates were asked about how they'd confront racial injustice. Only one had a real answer. https://...

6 Join me for a 3pm rally - tomorrow at the Mid-America Center in Council Bluffs, Iowa! Tickets:... https://...

7 This election is too important to sit out. Go to https://t.co/tTgeqxNqYm and make sure you're registered. ...

8 When Donald Trump goes low...register to vote: https://t.co/tTgeqxNqYm https://t.co/DXz9dEwsZS

9 Once again, we will have a government of, by and for the people. Join the MOVEMENT today! https://t.co/...

10 3) Has Trump offered a single proposal to reduce the friction of starting a business. @HillaryClinton has ...

11 The election is just weeks away. Check if you're registered to vote at https://t.co/HcMAh8ljR0, only takes a...

12 On National #VoterRegistrationDay, make sure you're registered to vote so we can ...

13 Hillary Clinton's Campaign Continues To Make False Claims About Foundation Disclosure: https://t.co/...

14 CNBC, Time magazine online polls say Donald Trump won the first presidential debate' via @WashTimes. ...

15 Donald Trump lied to the American people at least 58 times during the first presidential debate. /We

Is retweet: False
Time: 2016-09-28 00:22:34
Language: en
Retweet count: 218.0
Favorite count: 651.0
Longitude: ?
Latitude: ?
Author: HillaryClinton
Source URL: https://studio.twitter.com
Content: The question in this election: Who can put the plans into action that will make your life better? https://t.co/XreEY9OicG
Original author: ?
Place: ?

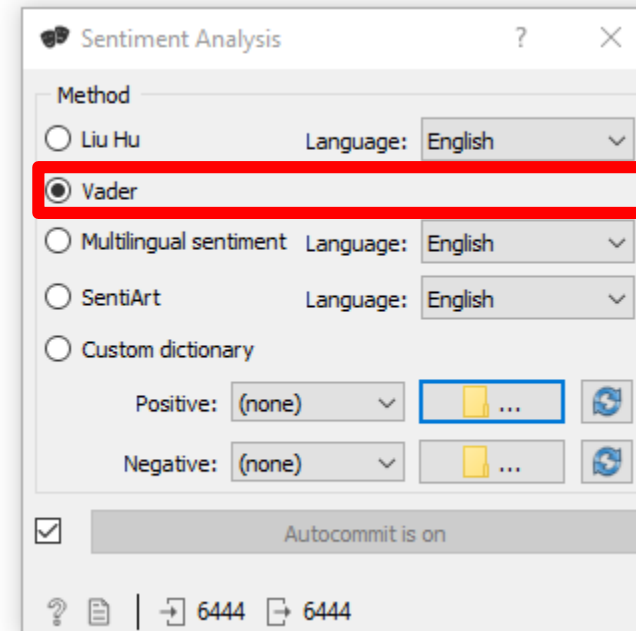
6444 | 1 | 6443 | 6444

BEISPIEL I: SENTIMENT-ANALYSE

3a) Eine einfache Sentiment-Analyse durchführen



Sentiment Analysis



Schritt 3a: Die eigentliche Sentiment-Analyse

- Es stehen mehrere einfache, vortrainierte Modelle zu Auswahl (Hier: Vader)
- Es lassen sich auch eigene Listen mit positiven und negativen Begriffen anlegen
- Ausgaben:
 - Anteil positiver, negativer und neutraler Wörter
 - Daraus abgeleitete Werte
 - Die ursprünglichen Daten

BEISPIEL I: SENTIMENT-ANALYSE

4a) Ergebnis
visualisieren



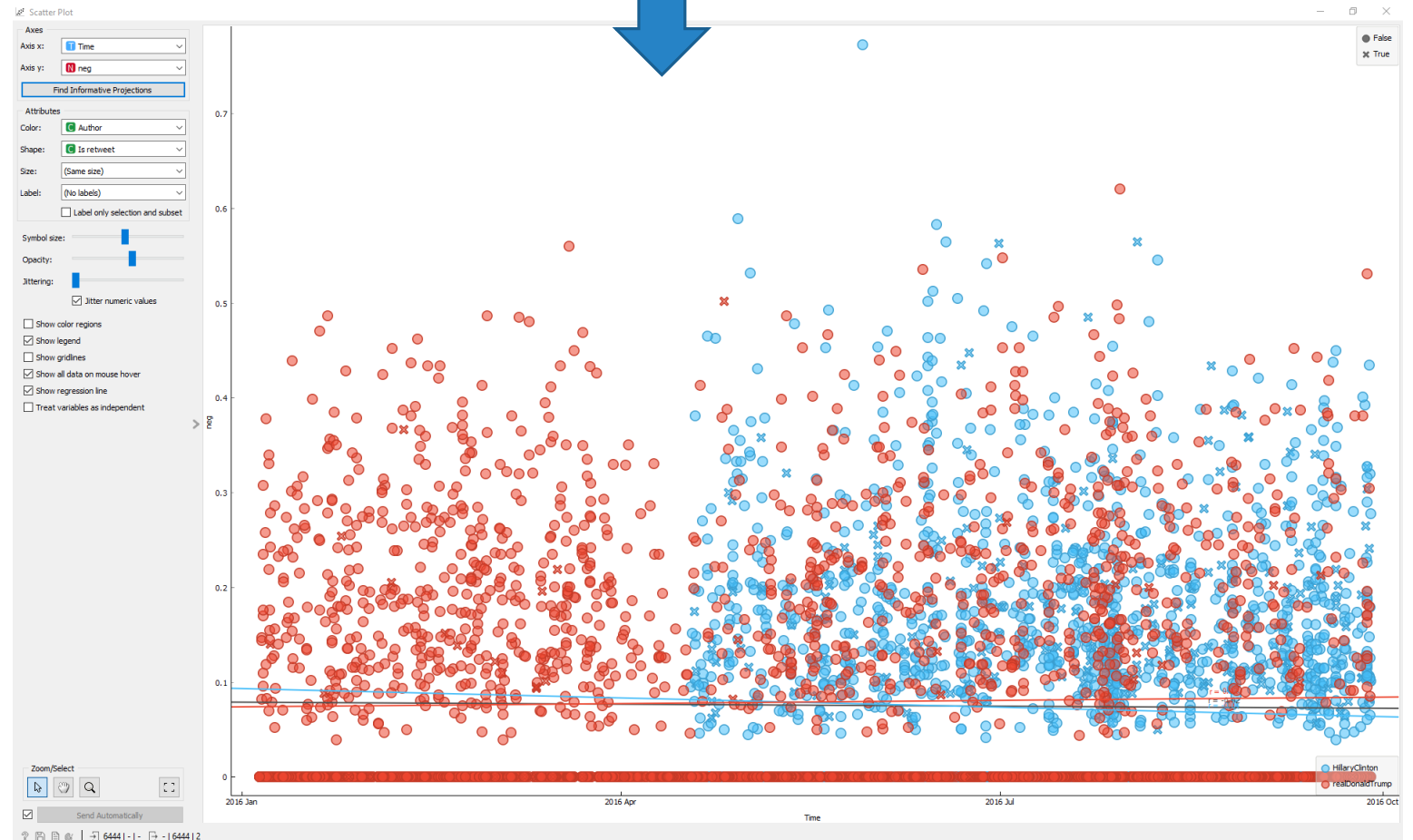
Scatter Plot



Schritt 4a:
Visualisierung

Hier: Scatter Plot

- X-Achse: Zeit
- Y-Achse: Negativer Anteil
- Farbe: Person
- Form: eigener Tweet oder weitergeleitete Nachricht?



BEISPIEL I: SENTIMENT-ANALYSE

Schritt 2b: Daten vorverarbeiten

Wichtig für Analysen

- Vorverarbeitung: 80% der Zeit
- Analyse: 20% der Zeit

Hier: Auswahl passender Datenätze

- Einfache Attribut-Wert-Vergleiche
 - Zeit, Anzahl Weiterleitungen, Autor, Text
- Auswahl interessanter Attribute
 - Zeitliche Einschränkungen
 - Sprache
 - ...

The image shows a workflow in Orange3. It starts with a 'Select Rows' widget, followed by a 'Select Columns' widget. A blue arrow points from the 'Select Columns' widget to a 'Select Columns' dialog box. This dialog box has an 'Ignored' list containing Latitude, Longitude, Favorite count, Is retweet, Language, Source URL, Original author, and Place. A red box highlights the 'Features' section of this dialog, which includes Time and Retweet count. Below the 'Select Columns' dialog is a 'Select Rows' dialog box. A blue arrow points from the 'Select Columns' dialog to the 'Select Rows' dialog. The 'Select Rows' dialog has a 'Conditions' table with three rows: 'Is retweet' is False, 'Time' is greater than '2016-06-15 03:36:53', and 'Language' is 'en'. A red box highlights this table. At the bottom of the 'Select Rows' dialog, the 'Add Condition' button is highlighted with a red box. The 'Send Automatically' checkbox is checked in both dialog boxes.

2b) Vorbereitung der Daten -> Auswahl passender Datensätze und der interessanten Informationen

Condition	Operator	Value
Is retweet	is	False
Time	is greater than	2016-06-15 03:36:53
Language	is	en

BEISPIEL I: SENTIMENT-ANALYSE

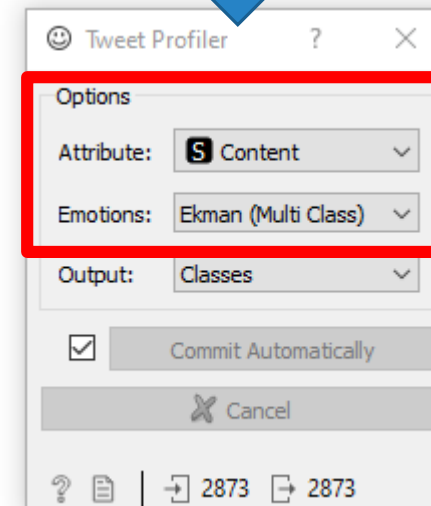
Schritt 3b: Die eigentliche Sentiment-Analyse

- Diesmal andere vortrainierte Modelle zur Unterscheidung in mehrere Emotionen wie
 - Angst
 - Freude
 - Traurigkeit
 - ...
- Analyse wieder über den „Content“



Tweet Profiler

3b) Komplexere Analyse der Tweets nach verschiedenen Emotionen



BEISPIEL I: SENTIMENT-ANALYSE

Schritt 4b)
Ergebnisse
visualisieren

- Scatter Plot
- X-Achse: Zeit
- Y-Achse: Emotion
- Farbe: Autor
- Größe: Anzahl Retweets



Scatter Plot (1)

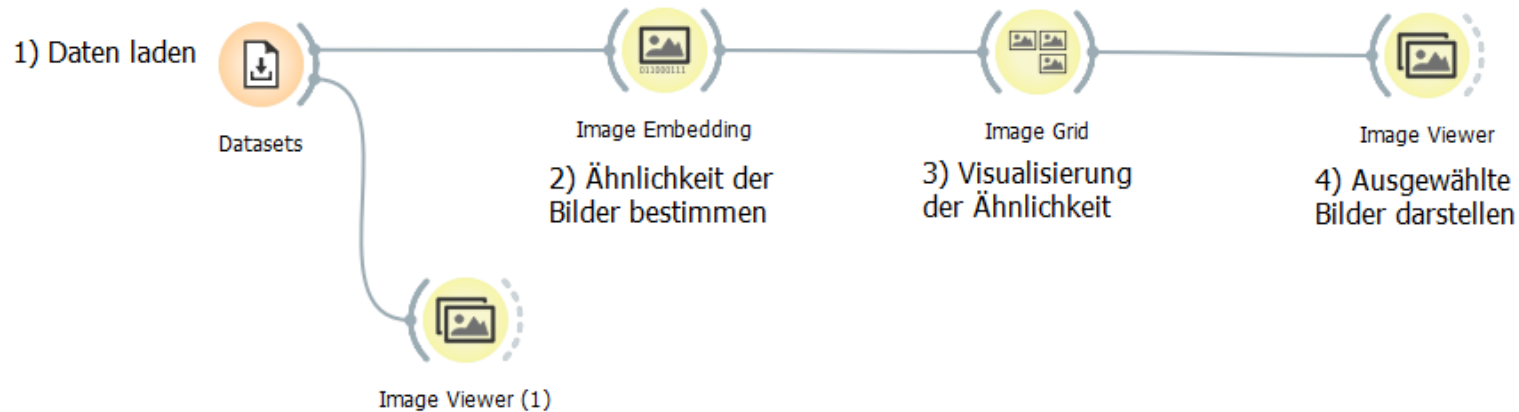
4b) Ergebnis
visualisieren



BEISPIEL 2: ÄHNLICHKEITEN VON BILDERN



BEISPIEL 2: ÄHNLICHKEITEN VON BILDERN



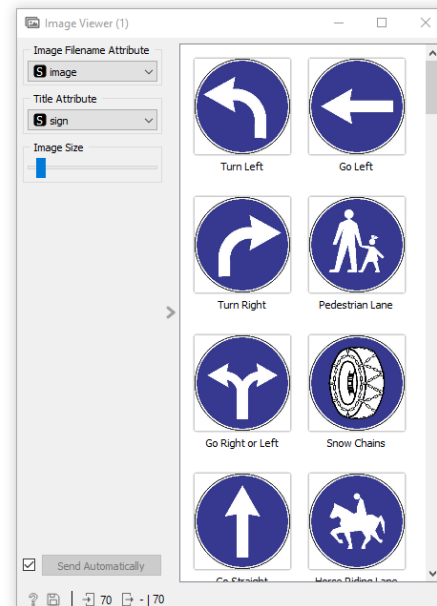
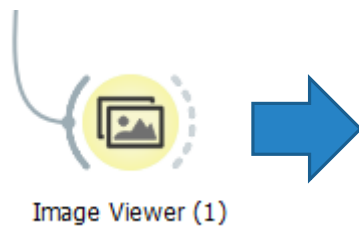
Wir

1. Laden eine Menge von Bildern,
2. Bestimmen die Ähnlichkeit zwischen den Bildern,
3. Visualisieren diese Ähnlichkeitsbeziehung und
4. Wählen eine Teilmenge der Bilder interaktiv aus

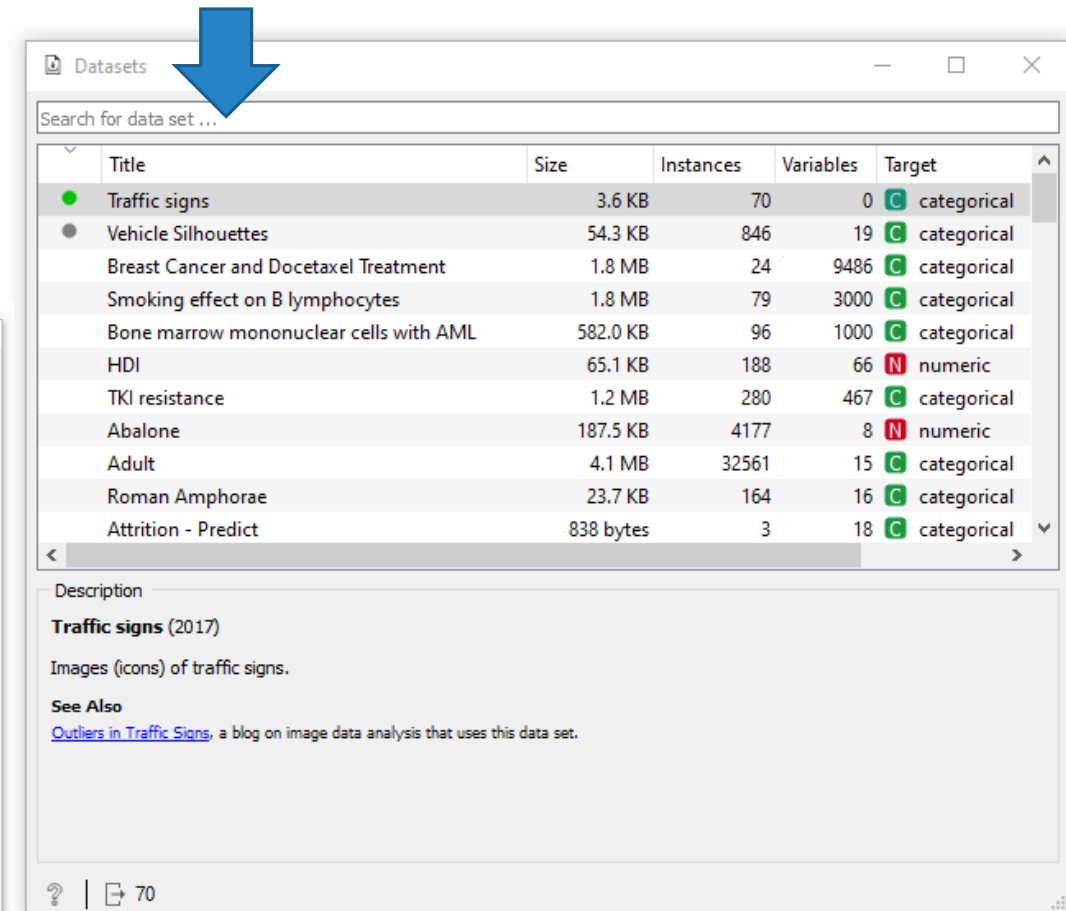
BEISPIEL 2: ÄHNLICHKEITEN VON BILDERN

Schritt 1: Laden der Daten

- „Traffic signs“: Symbole von Verkehrsschildern
- Funktioniert auch mit Fotos, allerdings dauert die Analyse dann länger



Datasets



Datasets

Search for data set ...

Title	Size	Instances	Variables	Target
Traffic signs	3.6 KB	70	0	C categorical
Vehicle Silhouettes	54.3 KB	846	19	C categorical
Breast Cancer and Docetaxel Treatment	1.8 MB	24	9486	C categorical
Smoking effect on B lymphocytes	1.8 MB	79	3000	C categorical
Bone marrow mononuclear cells with AML	582.0 KB	96	1000	C categorical
HDI	65.1 KB	188	66	N numeric
TKI resistance	1.2 MB	280	467	C categorical
Abalone	187.5 KB	4177	8	N numeric
Adult	4.1 MB	32561	15	C categorical
Roman Amphorae	23.7 KB	164	16	C categorical
Attrition - Predict	838 bytes	3	18	C categorical

Description

Traffic signs (2017)

Images (icons) of traffic signs.

See Also

[Outliers in Traffic Signs](#), a blog on image data analysis that uses this data set.

BEISPIEL 2: ÄHNLICHKEITEN VON BILDERN

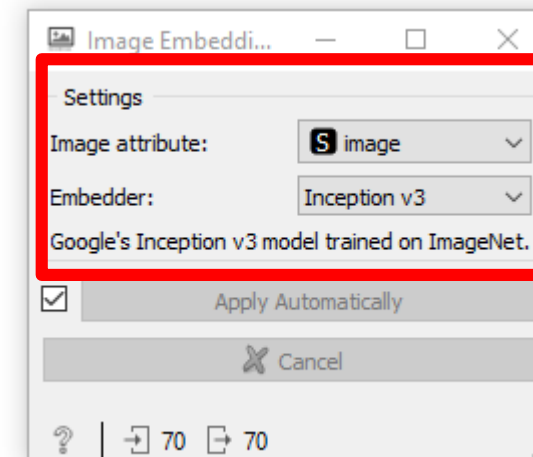
Schritt 2: Ähnlichkeit bestimmen

- Image attribute: „image“ → Das sind die Bilddaten
- Nutzen wieder vortrainiertes Modell
- Hier: Inception v3 von Google
 - Bildpunkte werden auf 2048 numerische Werte zwischen 0 und 1 abgebildet
 - Diese wurden über ein Neuronales Netz bestimmt
 - Grundlage für Bestimmung der Ähnlichkeit



Image Embedding

2) Ähnlichkeit der Bilder bestimmen



BEISPIEL 2: ÄHNLICHKEITEN VON BILDERN

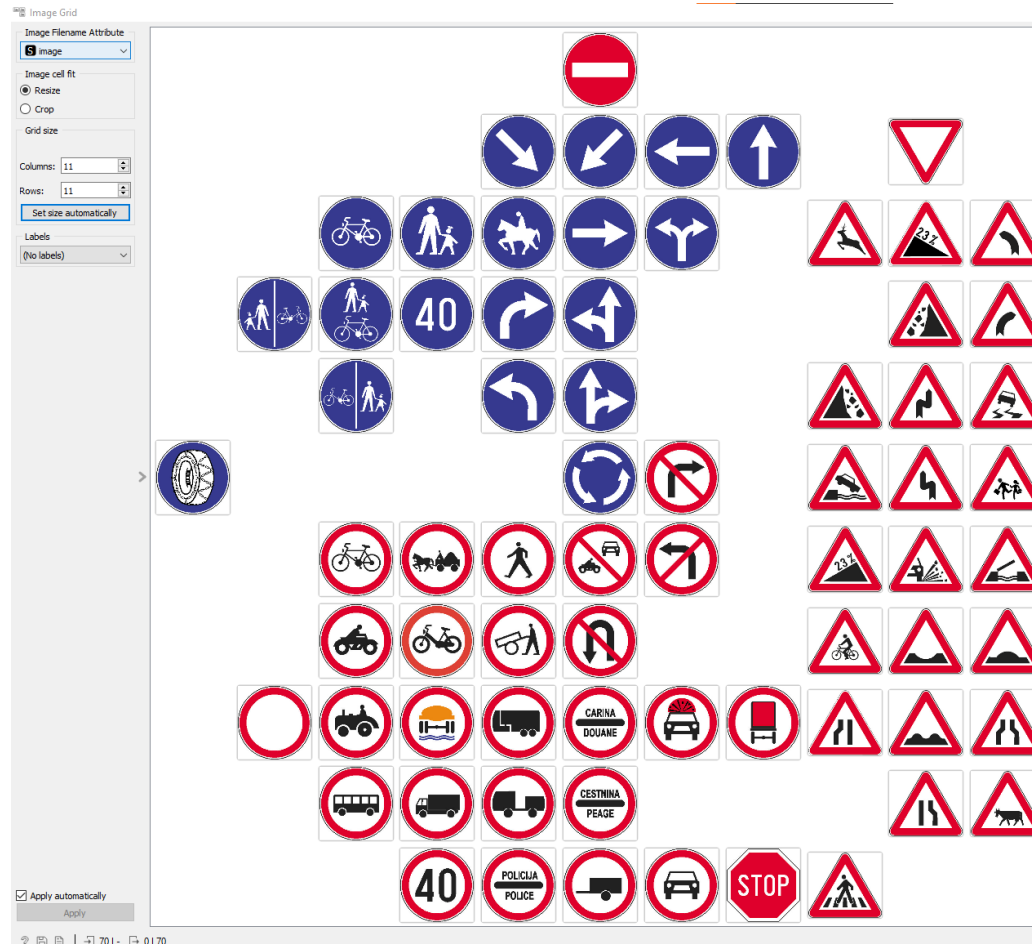
Anordnung der Bilder im Raster

- Ähnliche Bilder nahe beieinander
- Farbe
- Form



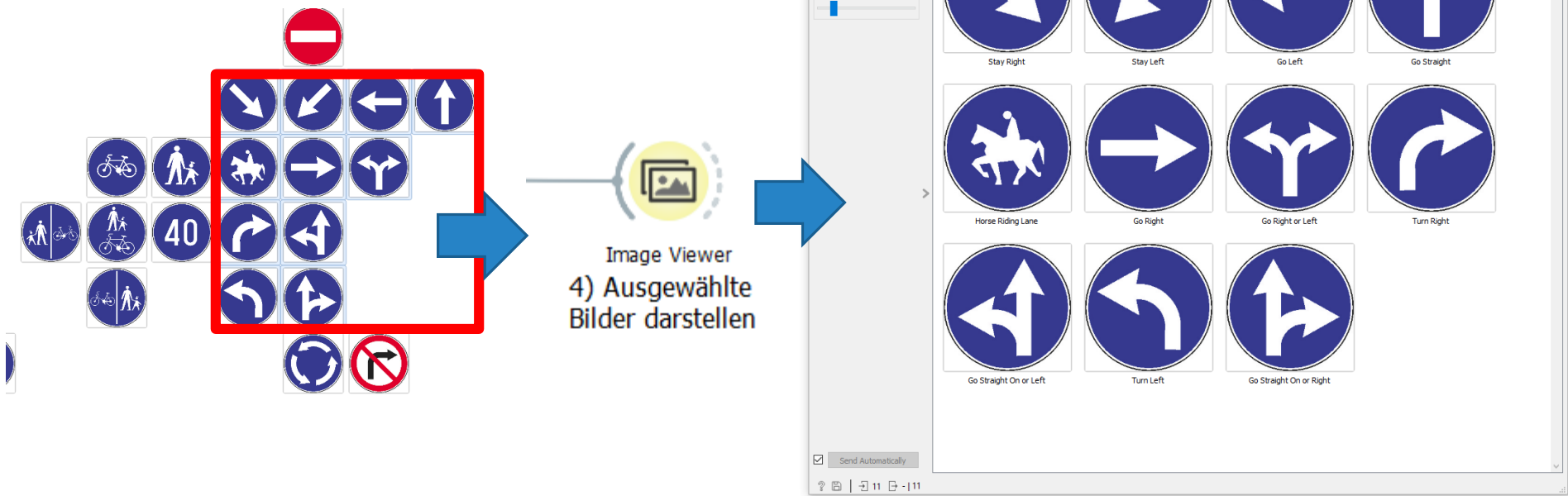
Image Grid

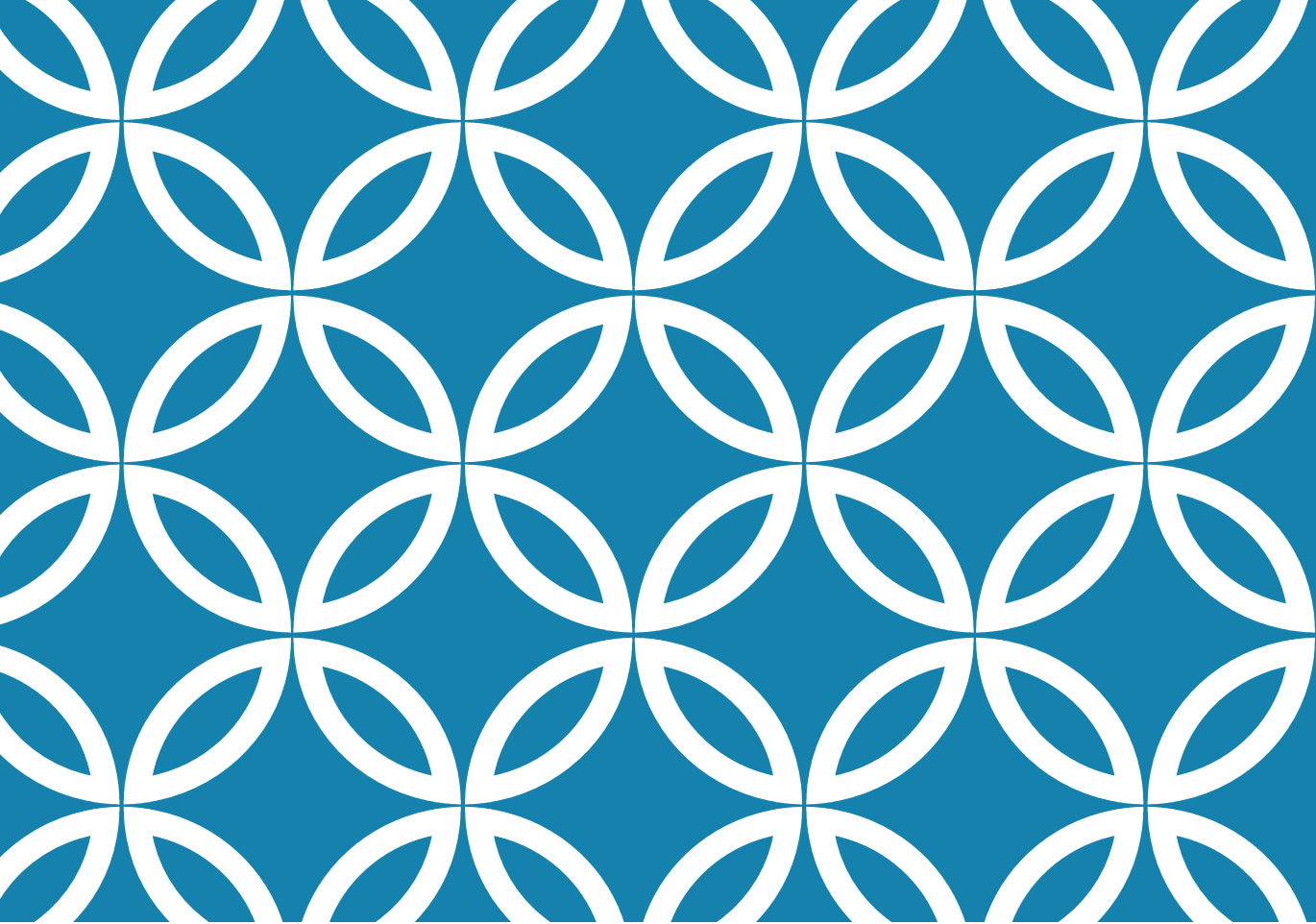
3) Visualisierung der Ähnlichkeit



BEISPIEL 2: ÄHNLICHKEITEN VON BILDERN

Schritt 4: Auswahl der Bilddaten



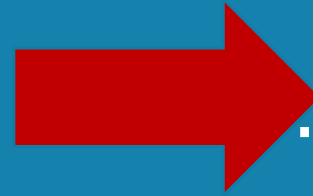


Künstliche Intelligenz

... kennenlernen

... ausprobieren

... selber machen



REFERENTEN:

DR.-ING. ANNE GUTSCHMIDT

M.SC. HANNES GRUNERT

01.11.2021

EIGENE ANALYSE: AUFGABENSTELLUNG

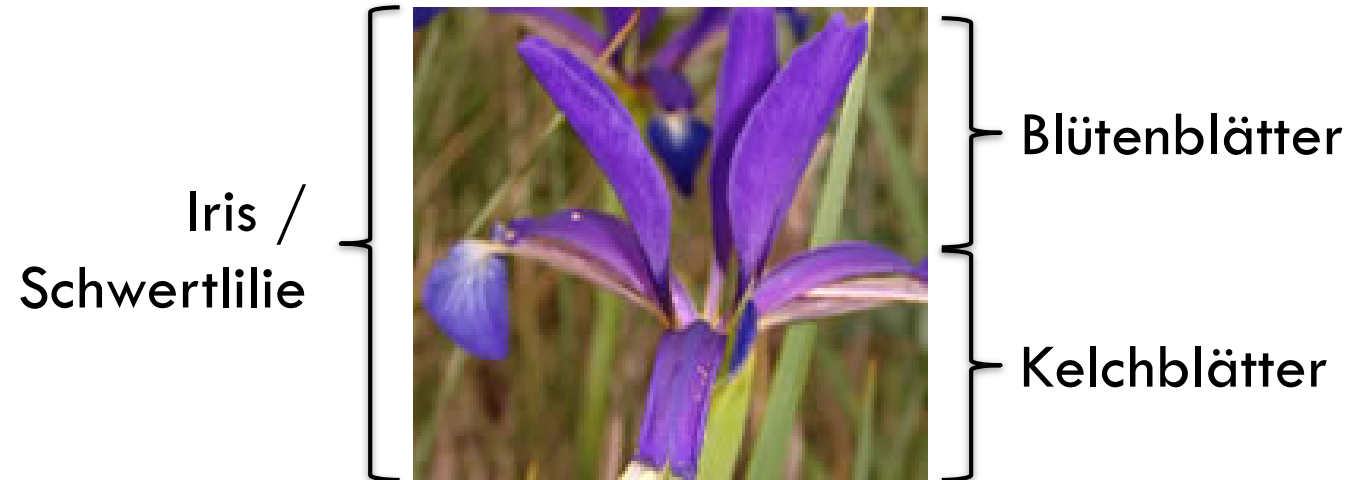
Bisher: Fertige Modelle verwendet

Jetzt: Eigenes Modell erstellen

Schritte:

1. Daten laden
2. Einen ersten Blick auf die Daten werfen
3. Einen genaueren Blick auf die Daten werfen
4. Verschiedene Modelle trainieren...
5. ... und miteinander vergleichen
6. + 7. Überprüfen, wo es noch zu Problemen kommt.

1) DER DATENSATZ



„Dies ist vielleicht die bekannteste Datenbank, die in der Literatur zur Mustererkennung zu finden ist. [...] Der Datensatz enthält 3 Klassen [...], wobei sich jede Klasse auf eine Art von Iris-Pflanze bezieht. Eine Klasse ist linear von den anderen 2 trennbar; die letzteren sind NICHT linear voneinander trennbar.“

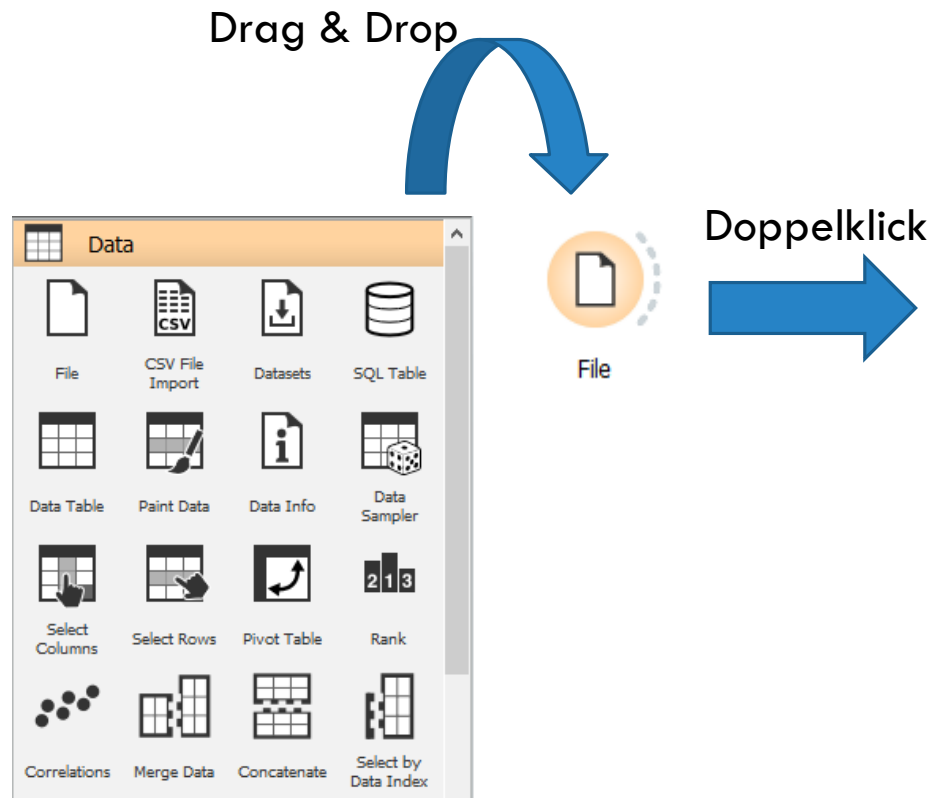
Übersetzung via DeepL; Quelle: <http://archive.ics.uci.edu/ml/datasets/Iris>

1) DER DATENSATZ

Kelchblatt		Blütenblatt		Unterart
Länge Kelchblatt	Breite Kelchblatt	Länge Blütenblatt	Breite Blütenblatt	Unterart
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica
7.1	3.0	5.9	2.1	Iris-virginica
...

150 Messwerte

1) DER DATENSATZ



File

Source

File: iris.tab 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	N numeric	feature	
2	sepal width	N numeric	feature	
3	petal length	N numeric	feature	
4	petal width	N numeric	feature	
5	iris	C categorical	target	Iris-setosa, Iris-versicolor, Iris-virginica

Reset Apply

Browse documentation datasets

? | 150

The screenshot shows the 'File' dialog box in a software application. The 'Source' section has 'File: iris.tab' selected. The 'Info' section provides details about the 'Iris flower dataset', including the number of instances (150), features (4), and classification type. The 'Columns' section displays a table with 5 columns: 'sepal length', 'sepal width', 'petal length', 'petal width', and 'iris'. The 'iris' column is highlighted as the target variable. The 'Reset' and 'Apply' buttons are visible at the bottom, along with a 'Browse documentation datasets' link and a status bar showing '150' instances.

1) DER DATENSATZ (EINSCHUB)

Download weiterer Beispiele



Datasets

- Anpassung an eigenen Lehrplan / das Unterrichtsfach

Title	Size	Instances	Variables	Target	Tags
Conferences	2.3 KB	42	5		
Cyber Security Breaches	225.0 KB	1055	10		security, time, geo
Dermatology	30.9 KB	366	35	C categorical	biology, medical
Development of Social Amoeba	15.5 KB	152	0	C categorical	image analytics, biology
Illegal waste dumpsites in Slovenia	2.8 MB	13165	25		geo, timeseries, ecology
Foodmart 2000	4.0 MB	62560	126		economy, associate, basket
Forest Fires	31.3 KB	517	12	N numeric	ecology
Glass	10.4 KB	214	10	C categorical	physics, criminology
Grades for English and Math	265 bytes	12	3		synthetic, educational
Course Grades	9.2 KB	16	7		synthetic, education
Hair section	18.2 MB	3250	831		spectral, hyperspectral
Heart Disease	23.5 KB	303	14	C categorical	biology, medicine

Description

Forest Fires (2007), from [UCI ML Repository](#)

This is a difficult regression task, where the aim is to predict the burned area of forest fires in the northeast region of Portugal. The attributes report on meteorological data (temperature, wind, rain, humidity), month and day of the status, several indices of the Forest Fire Weather Index, and spatial coordinate within the Montesinho park map. Two extra meta attributes are include in the Orange data set that encode the log of area + 1 and the binary attribute reporting if the part of the park was under fire (non-zero fire area).

References

Cortez P and Morais A (2007) [A Data Mining Approach to Predict Forest Fires using Meteorological Data](#). In Proc. of the 13th Portuguese Conference on Artificial Intelligence, Portugal, pp. 512-523.

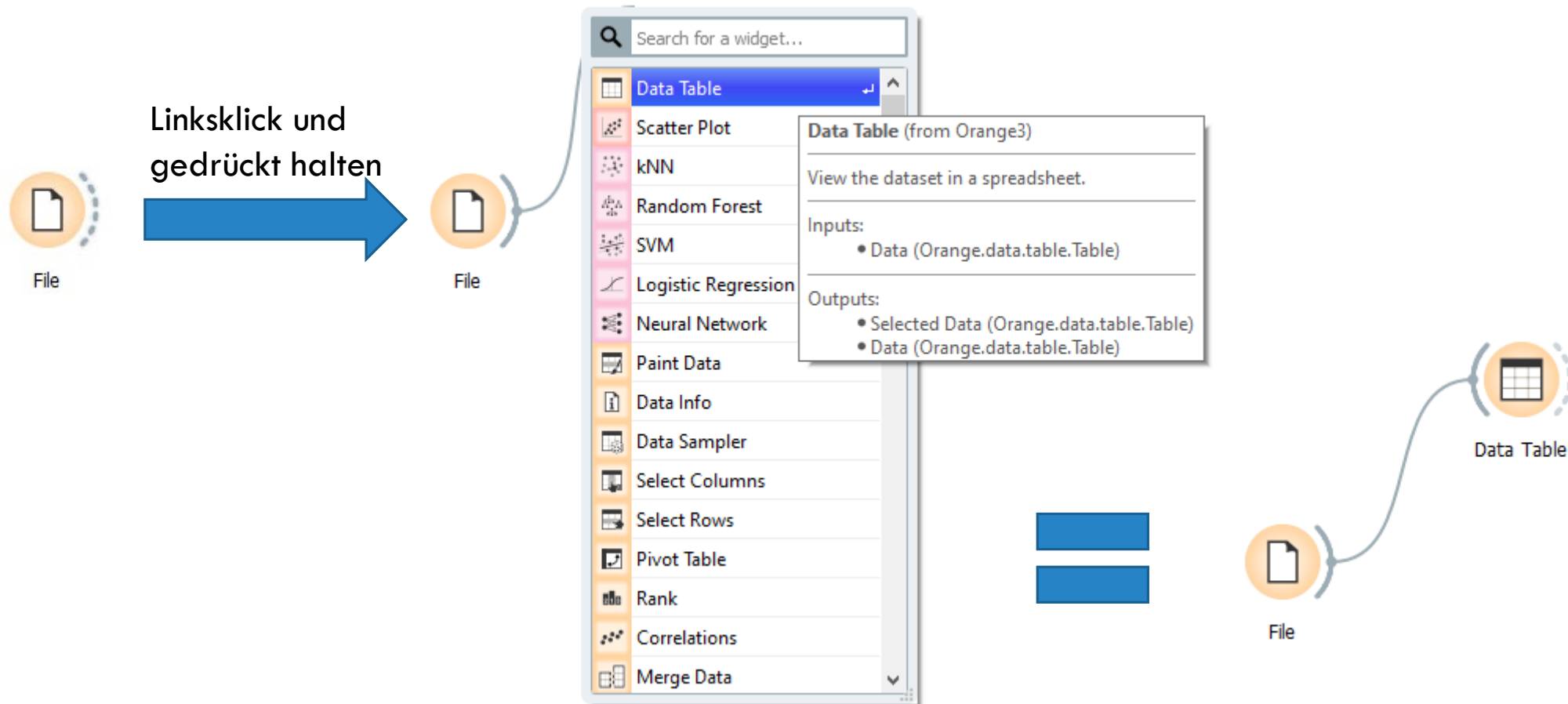


SQL Table

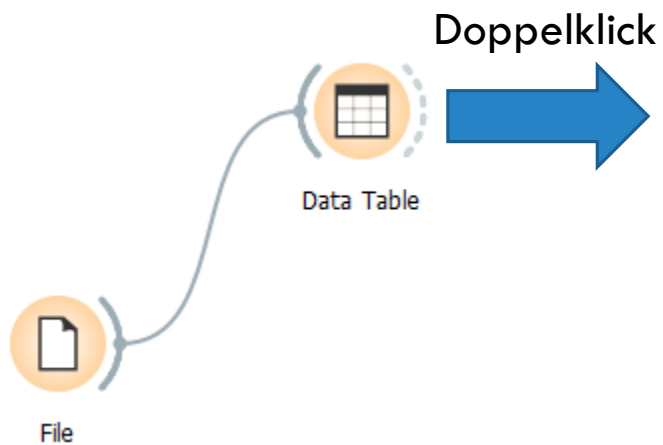
Verbindung mit eigener Datenbank

- Für Schüler mit Datenbankkenntnissen
- Erfordert vorherige Konfiguration

2) DATEN ALS TABELLE



2) DATEN ALS TABELLE



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

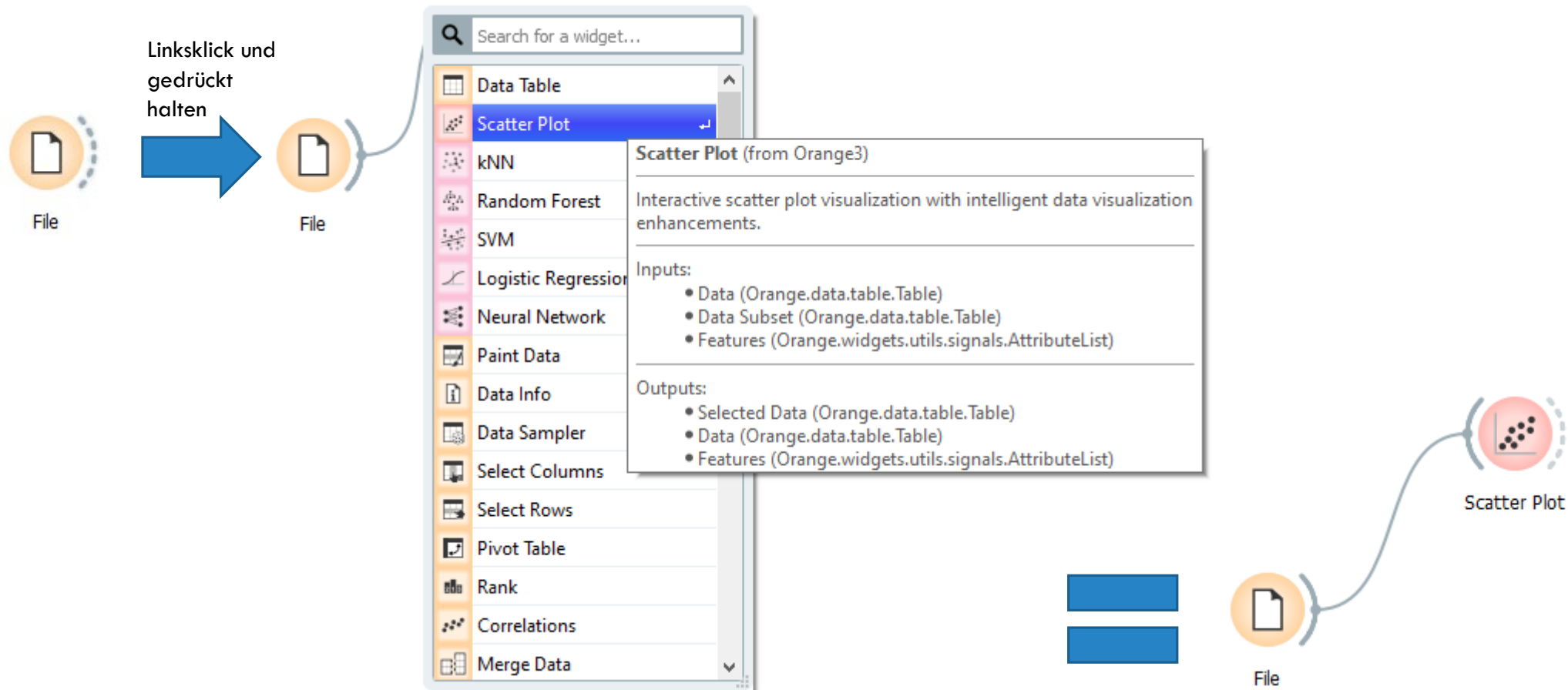
Restore Original Order

Send Automatically

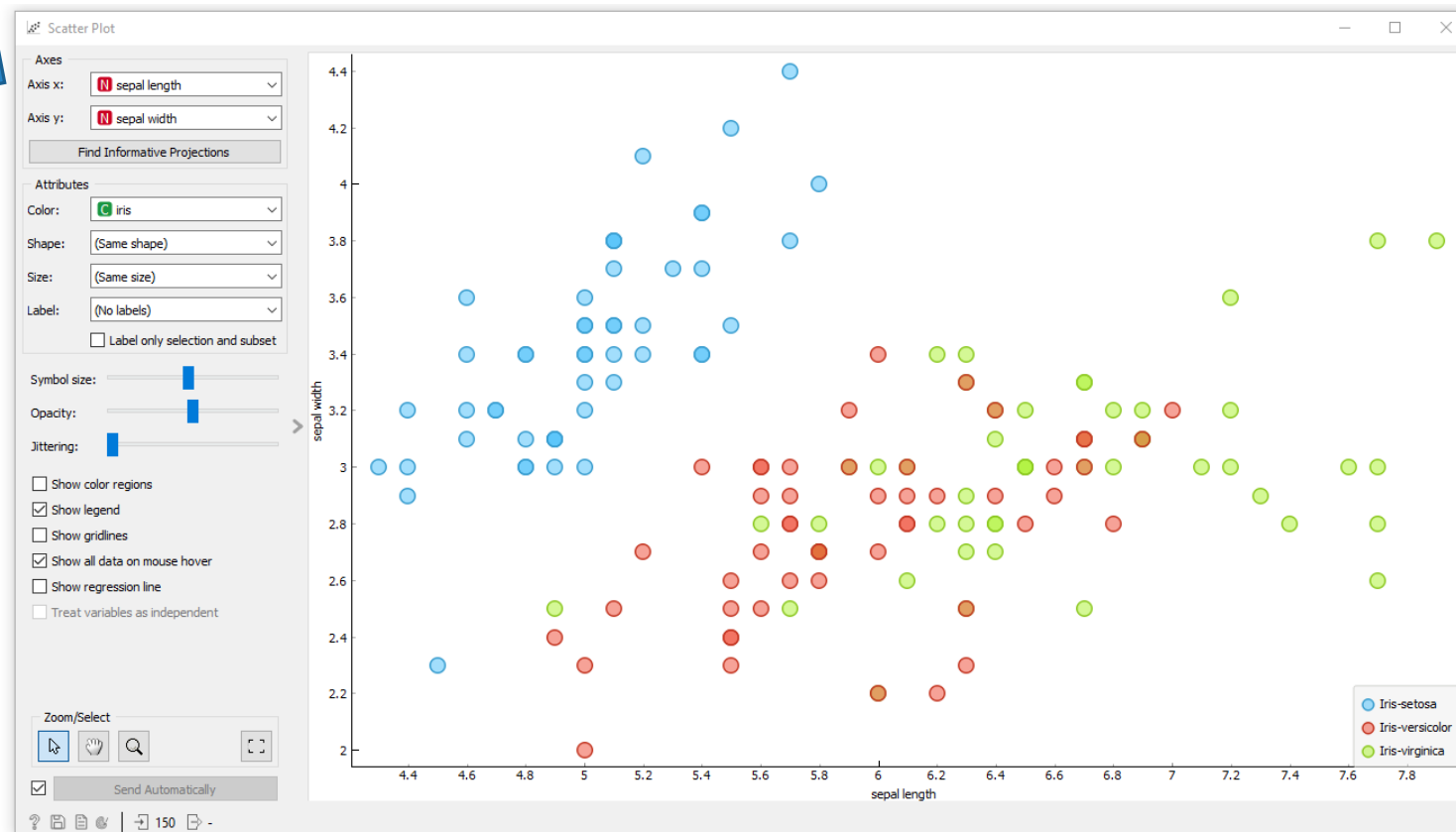
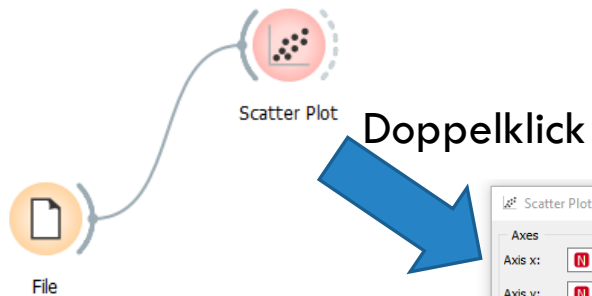
	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2
6	Iris-setosa	5.4	3.9	1.7	0.4
7	Iris-setosa	4.6	3.4	1.4	0.3
8	Iris-setosa	5.0	3.4	1.5	0.2
9	Iris-setosa	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

? | 150 | 150

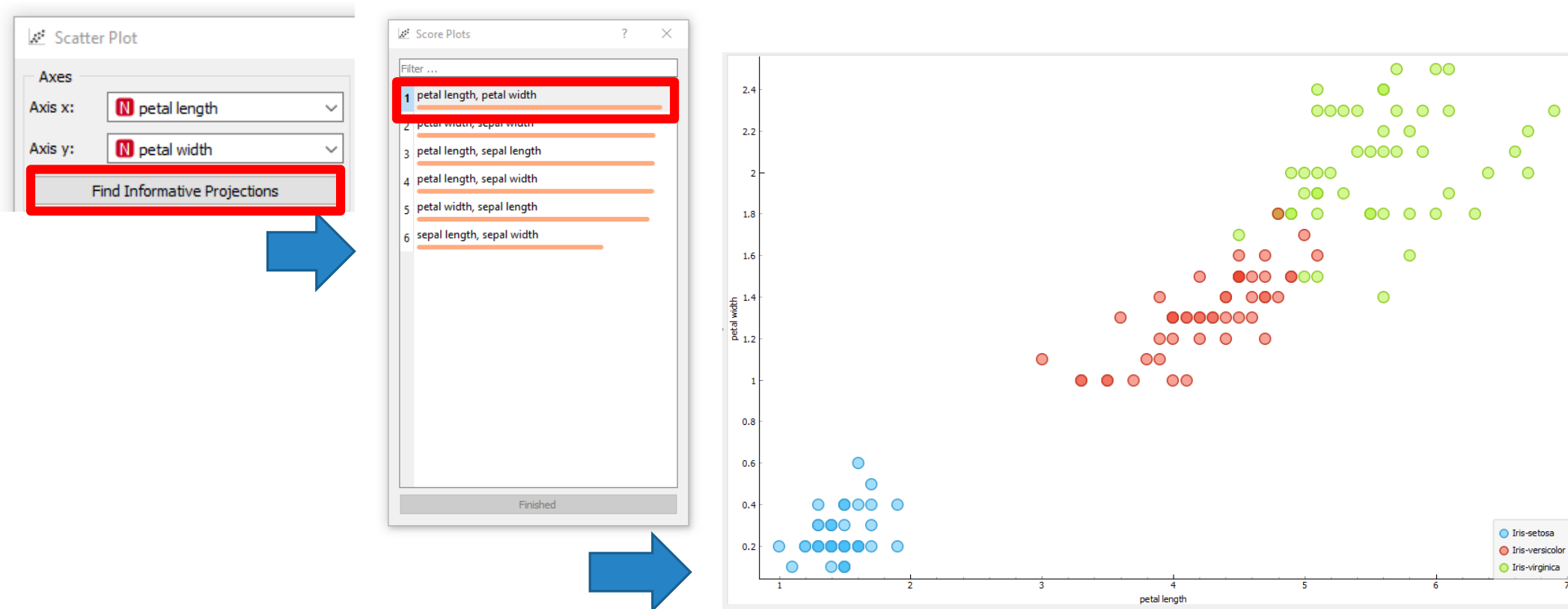
3) DATEN VISUALISIERT



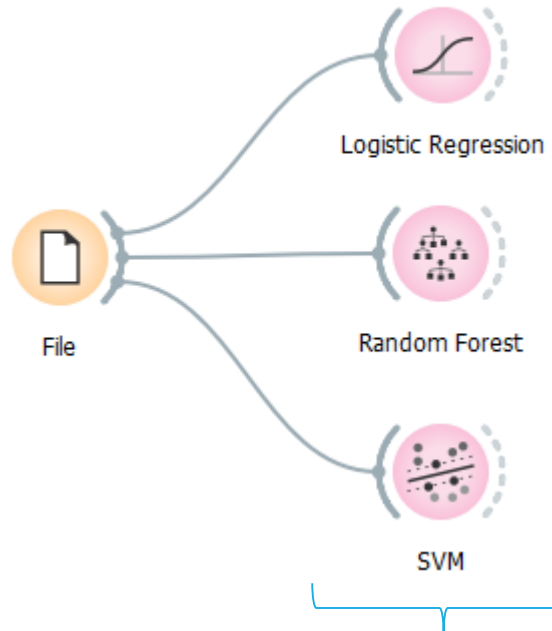
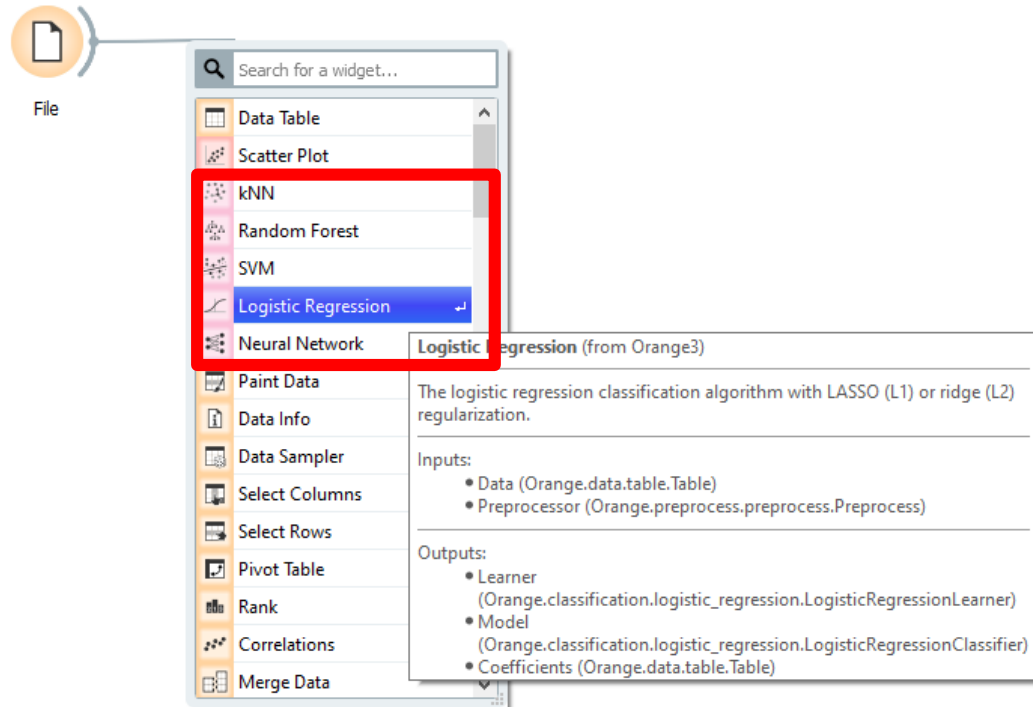
3) DATEN VISUALISIERT



3) DATEN VISUALISIERT



4) MODELLE AUSWÄHLEN



Mehrere Verfahren auswählen;
einfache Verfahren sind z.B. „k
Nearest Neighbours“ kNN und
Entscheidungsbäume (Tree)

4) MODELLE AUSWÄHLEN – EINSTELLUNGEN

Logistic Regre... ? X

Name
Logistic Regression

Regularization type: Ridge (L2) v

Strength:
Weak ——— Strong
C=1

Balance class distribution

Apply Automatically

? | 150

Random Forest ? X

Name
Random Forest

Basic Properties

Number of trees: 10

Number of attributes considered at each split: 5

Replicable training

Balance class distribution

Growth Control

Limit depth of individual trees: 3

Do not split subsets smaller than: 5

Apply Automatically

? | 150

SVM ? X

Name
SVM

SVM Type

SVM Cost (C): 1,00

Regression loss epsilon (ϵ): 0,10

v-SVM Regression cost (C): 1,00

Complexity bound (v): 0,50

Kernel

Linear Kernel: $\exp(-g|x-y|^2)$

Polynomial g: auto

RBF

Sigmoid

Optimization Parameters

Numerical tolerance: 0,0010

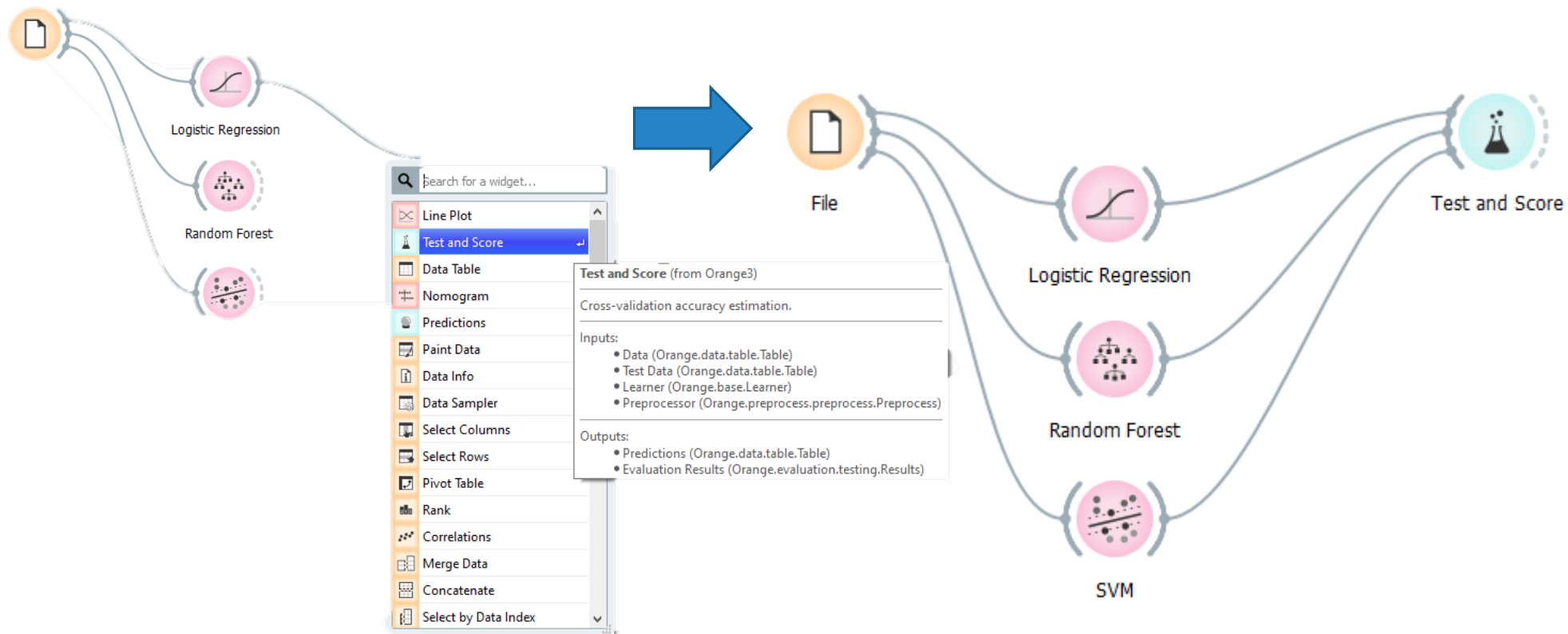
Iteration limit: 100

Apply Automatically

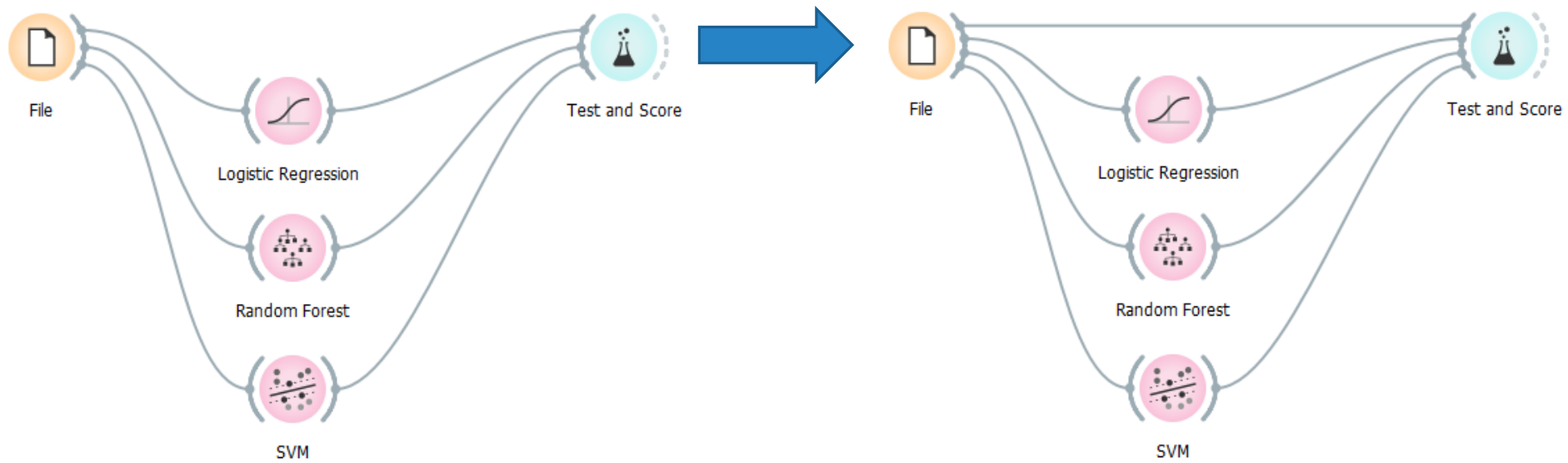
? | 150

Realität: Parameter automatisch anpassen, bis Ergebnis „gut genug“

5) MODELLE TRAINIEREN...



5) ... UND TESTEN



5) MODELLE TRAINIEREN UND TESTEN

Doppelklick

File

Logistic Regression

Random Forest

SVM

Test and Score

Test and Score

Sampling

Cross validation

Number of folds: 5

Stratified

Cross validation by feature

Random sampling

Repeat train/test: 10

Training set size: 66 %

Stratified

Leave one out

Test on train data

Test on test data

Target Class

(Average over classes)

Model Comparison

Area under ROC curve

Negligible difference: 0.1

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
SVM	0.997	0.957	0.957	0.957	0.957
Random Forest	0.991	0.957	0.957	0.957	0.957
Logistic Regression	0.997	0.963	0.963	0.963	0.963

Model Comparison by AUC

	SVM	Random Forest	Logistic Regression
SVM			
Random Forest			
Logistic Regression			

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

? | 150 | 510

6) KONFUSIONSMATRIX

The diagram illustrates a workflow starting with a 'File' node, which branches into three parallel paths for 'Logistic Regression', 'Random Forest', and 'SVM'. These paths converge into a 'Test and Score' node, which then leads to a 'Confusion Matrix' node. A blue arrow labeled 'Doppelklick' (double-click) points from the 'Confusion Matrix' node to a software window.

Confusion Matrix Window

Learnners: Logistic Regression, Random Forest, SVM

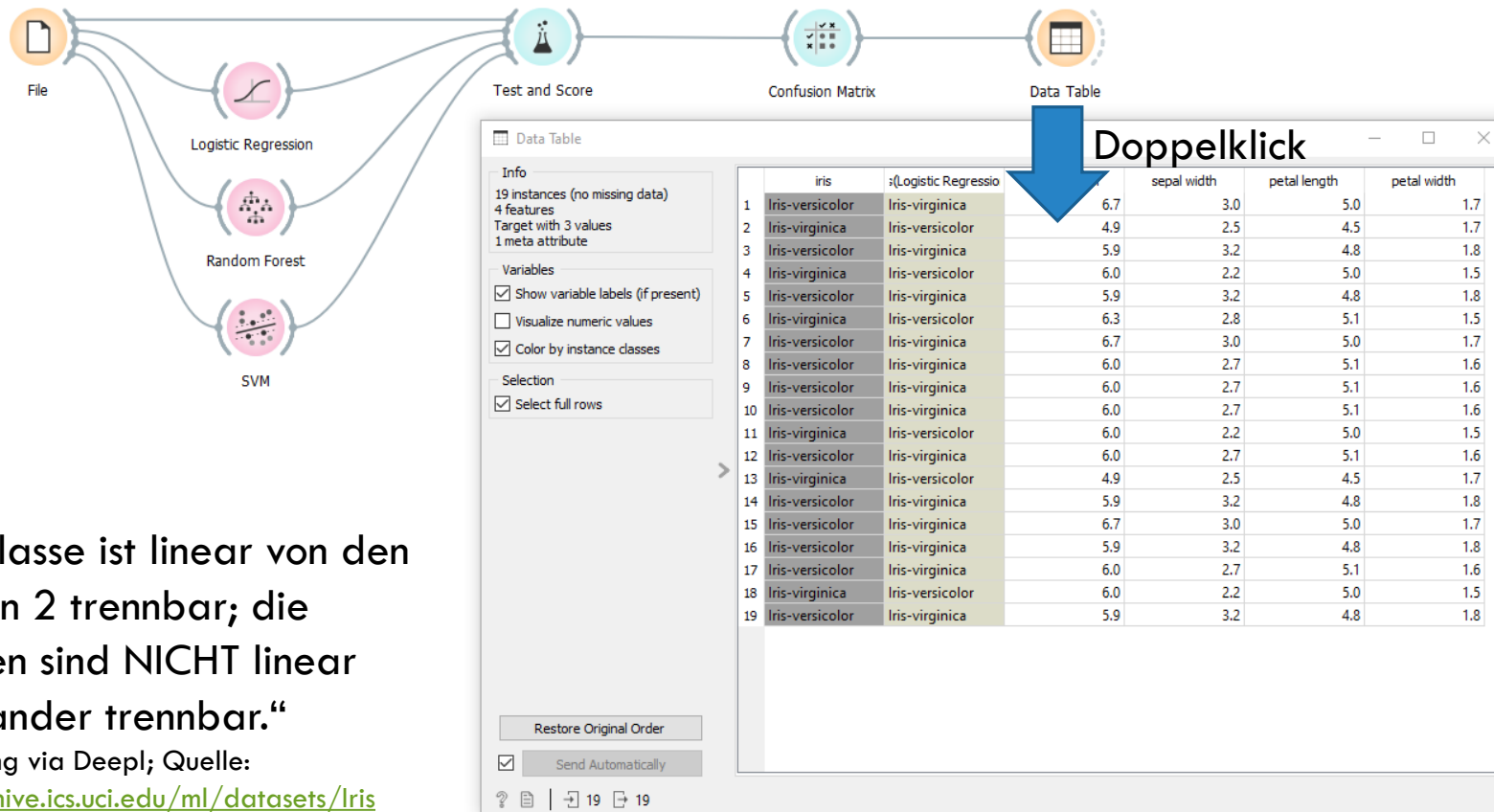
Show: Number of instances

		Predicted			Σ
		Iris-setosa	Iris-versicolor	Iris-virginica	
Actual	Iris-setosa	170	0	0	170
	Iris-versicolor	0	157	13	170
	Iris-virginica	0	6	164	170
Σ		170	163	177	510

Buttons: Select Correct, Select Misclassified, Clear Selection

Options: Predictions Probabilities, Apply Automatically

7) FALSCH KLASSIFIZIERTE DATEN



The diagram shows a workflow starting with a 'File' icon, branching into three models: 'Logistic Regression', 'Random Forest', and 'SVM'. All three models lead to a 'Test and Score' step, which then leads to a 'Confusion Matrix' and finally a 'Data Table'.

The 'Data Table' window is shown with a blue arrow pointing to the 'iris' column, labeled 'Doppelklick'. The table contains 19 rows of data with columns: iris, s(Logistic Regressio), sepal width, petal length, and petal width.

	iris	s(Logistic Regressio	sepal width	petal length	petal width	
1	Iris-versicolor	Iris-virginica	6.7	3.0	5.0	1.7
2	Iris-virginica	Iris-versicolor	4.9	2.5	4.5	1.7
3	Iris-versicolor	Iris-virginica	5.9	3.2	4.8	1.8
4	Iris-virginica	Iris-versicolor	6.0	2.2	5.0	1.5
5	Iris-versicolor	Iris-virginica	5.9	3.2	4.8	1.8
6	Iris-virginica	Iris-versicolor	6.3	2.8	5.1	1.5
7	Iris-versicolor	Iris-virginica	6.7	3.0	5.0	1.7
8	Iris-versicolor	Iris-virginica	6.0	2.7	5.1	1.6
9	Iris-versicolor	Iris-virginica	6.0	2.7	5.1	1.6
10	Iris-versicolor	Iris-virginica	6.0	2.7	5.1	1.6
11	Iris-virginica	Iris-versicolor	6.0	2.2	5.0	1.5
12	Iris-versicolor	Iris-virginica	6.0	2.7	5.1	1.6
13	Iris-virginica	Iris-versicolor	4.9	2.5	4.5	1.7
14	Iris-versicolor	Iris-virginica	5.9	3.2	4.8	1.8
15	Iris-versicolor	Iris-virginica	6.7	3.0	5.0	1.7
16	Iris-versicolor	Iris-virginica	5.9	3.2	4.8	1.8
17	Iris-versicolor	Iris-virginica	6.0	2.7	5.1	1.6
18	Iris-virginica	Iris-versicolor	6.0	2.2	5.0	1.5
19	Iris-versicolor	Iris-virginica	5.9	3.2	4.8	1.8

“Eine Klasse ist linear von den anderen 2 trennbar; die letzteren sind NICHT linear voneinander trennbar.“

Übersetzung via DeepL; Quelle:

<http://archive.ics.uci.edu/ml/datasets/Iris>

LITERATUR

- Mark Lutz: *Python – kurz & gut*, O'Reilly, 5. Auflage, 2014
- Annalyn Ng, Kenneth Soo: *Data Science – Was ist das eigentlich?*, Springer, 2018
- Jürgen Cleve, Uwe Lämmel: *Data Mining*, De Gruyter Verlag, 3. Auflage, 2020
- Uwe Lämmel, Jürgen Cleve: *Künstliche Intelligenz: Wissensverarbeitung – Neuronale Netze*, Hanser Verlag, 5. Auflage, 2020
- Stuart Russell, Peter Norvig: *Künstliche Intelligenz: Ein moderner Ansatz*, Pearson Studium, 3. Auflage, 2012
- Stuart Russell: *Human Compatible: Künstliche Intelligenz und wie der Mensch die Kontrolle über superintelligente Maschinen behält*, mitp Verlag, 1. Auflage, 2020