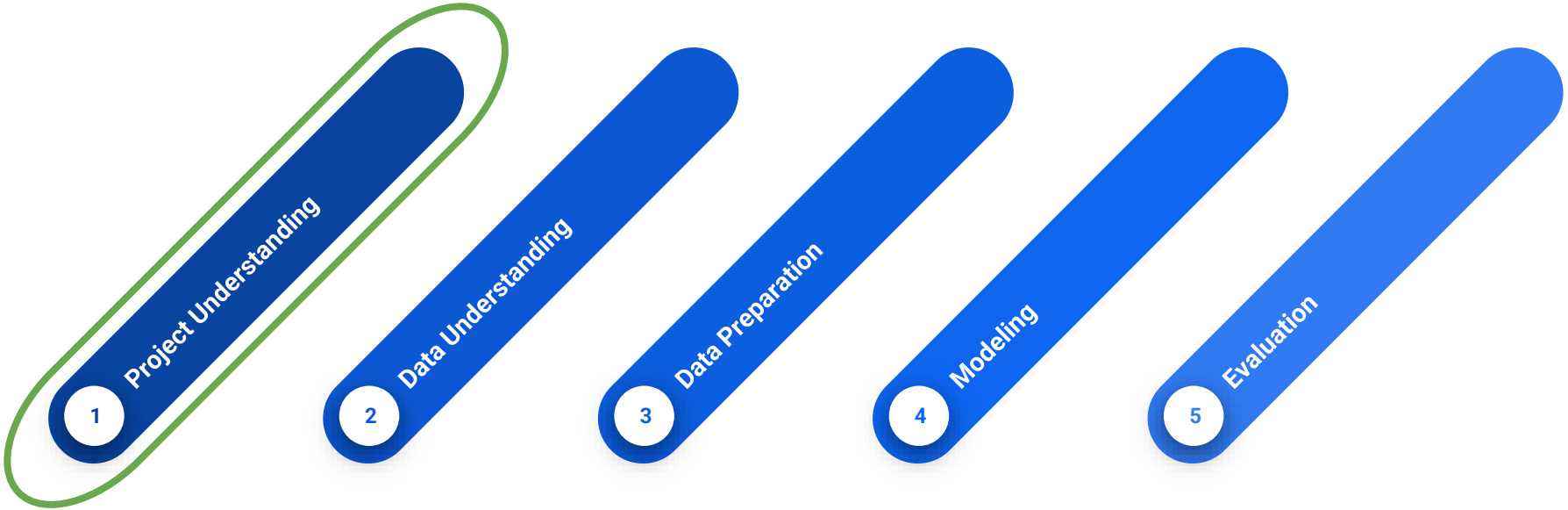
Abschlusspräsentation

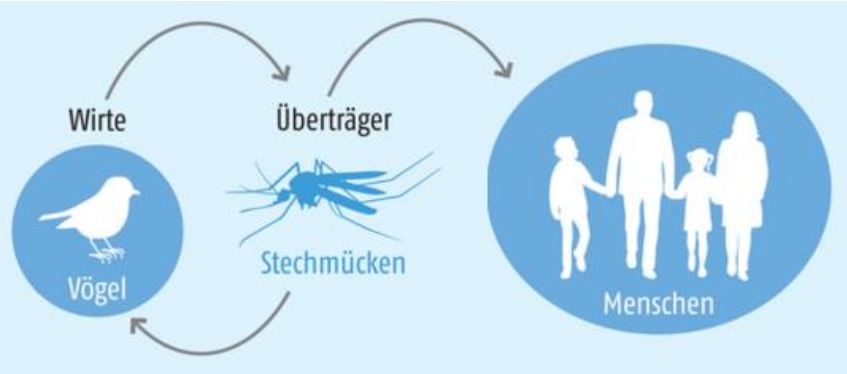
Business Intelligence SoSe ‘23

West Nil Virus - Vorhersage in Chicago

Agenda - CRISP-DM MODELL

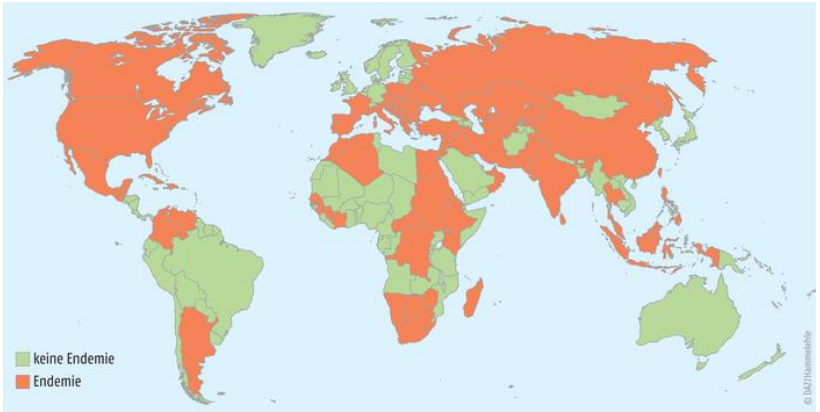


West Nil Virus



- Keine Impfung beim Menschen
- Alle Erdteile betroffen

- Inkubationszeit 3-14 Tage
- 20% grippeähnlichen Symptome

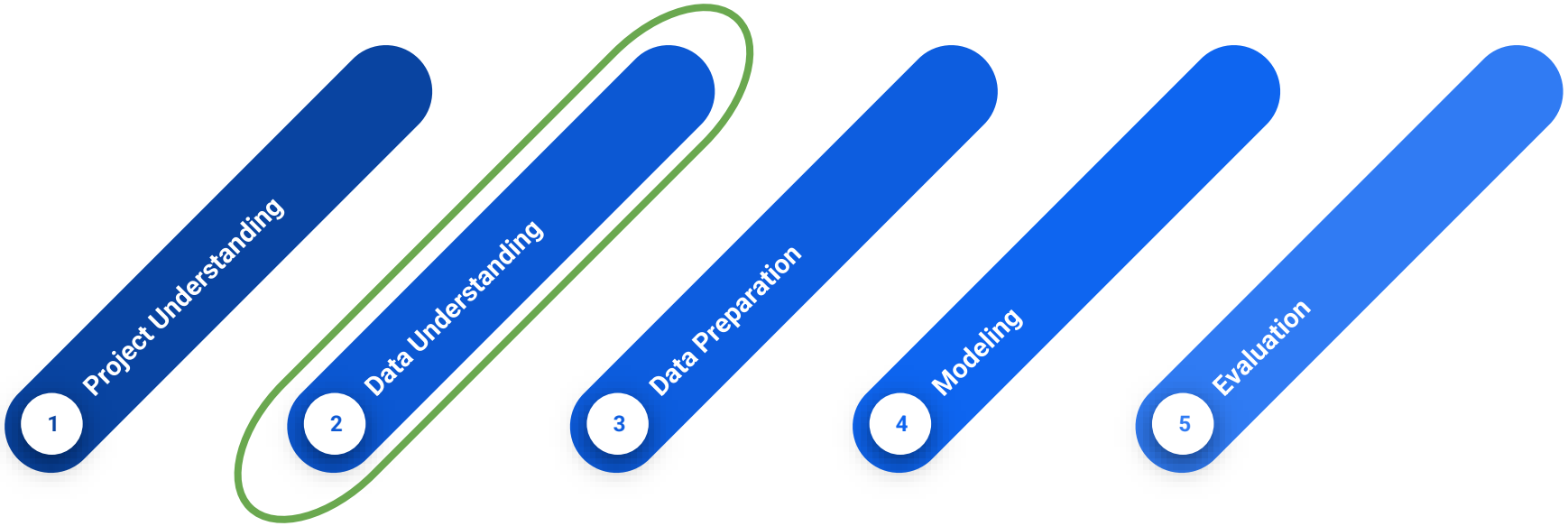


Project Understanding

- Überall in Chicago sind Mückenfallen aufgestellt
- Mücken werden auf West-Nil-Virus getestet
- Dadurch wird entschieden, wo vermehrt Pestizide eingesetzt werden müssen
- Annahme: Heißes und trockenes Wetter ist für das West-Nil-Virus günstiger als kaltes und nasses Wetter

Entwicklung eines Vorhersagemodells, ob das West-Nil-Virus für eine bestimmte Zeit und einen bestimmten Ort vorhanden ist oder nicht.

CRISP-DM MODELL



Data Understanding: Trainingsdaten

df_train.head():

	Date	Address	Species	Block	Street	Trap	AddressNumberAndStreet	Latitude	Longitude	AddressAccuracy	NumMosquitos	WnvPresent
0	2007-05-29	4100 North Oak Park Avenue, Chicago, IL 60634,...	CULEX PIPIENS/RESTUANS	41	N OAK PARK AVE	T002	4100 N OAK PARK AVE, Chicago, IL	41.954690	-87.800991	9	1	0
1	2007-05-29	4100 North Oak Park Avenue, Chicago, IL 60634,...	CULEX RESTUANS	41	N OAK PARK AVE	T002	4100 N OAK PARK AVE, Chicago, IL	41.954690	-87.800991	9	1	0
2	2007-05-29	6200 North Mandell Avenue, Chicago, IL 60646, USA	CULEX RESTUANS	62	N MANDELL AVE	T007	6200 N MANDELL AVE, Chicago, IL	41.994991	-87.769279	9	1	0

WnvPresent: Ist West Nil Virus präsent?

1 = ja, 0 = nein

Data Understanding: Weitere Daten

Map-Data

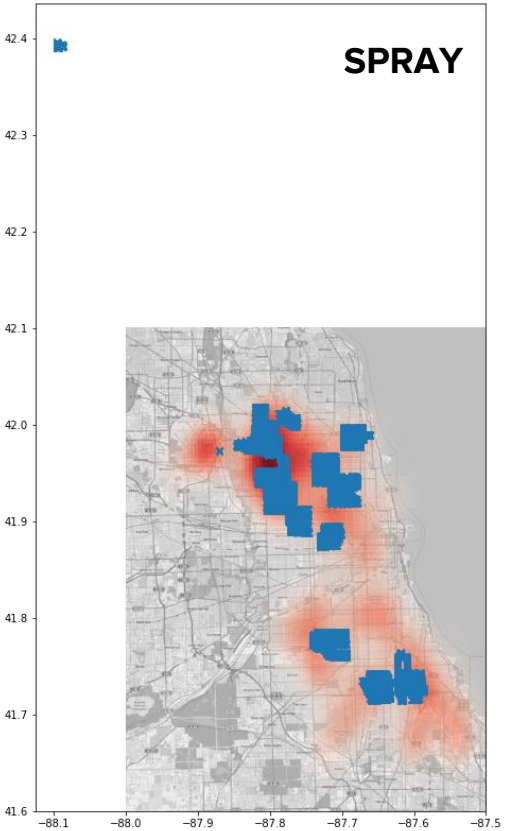
- aus Openstreet Map
  - 87.5 bis 88 Longitude
  - 41.6 bis 41.2 Latitude

Spray

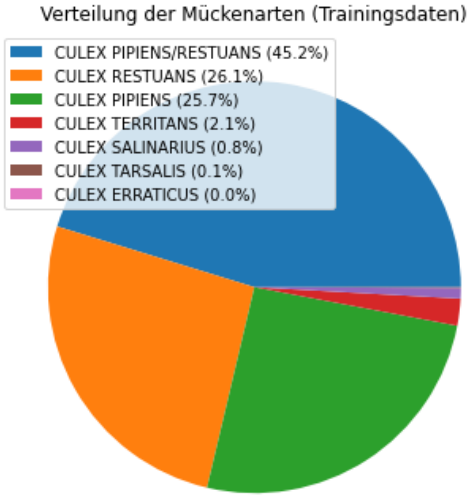
- Wo wurde gesprayed? (➜ Latitude, Longitude)
- Wann wurde gesprayed? (➜ Date, Time)

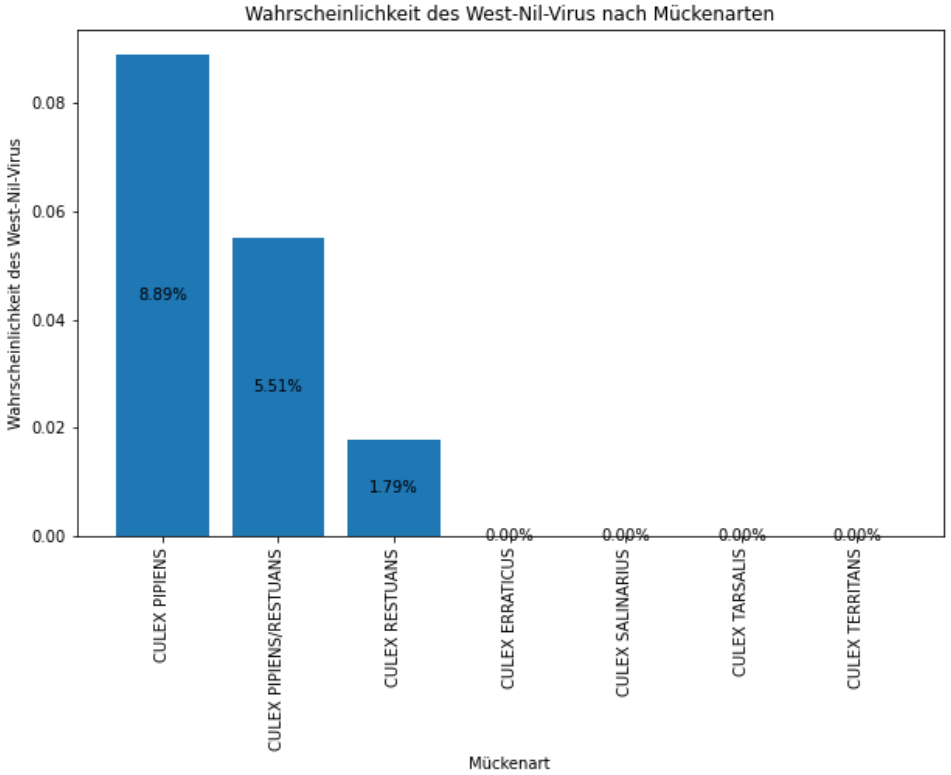
Weather

- detaillierte Wetterdaten wie:
  - Sunrise, Sunset, Temperature, SeaLevel, SnowFall



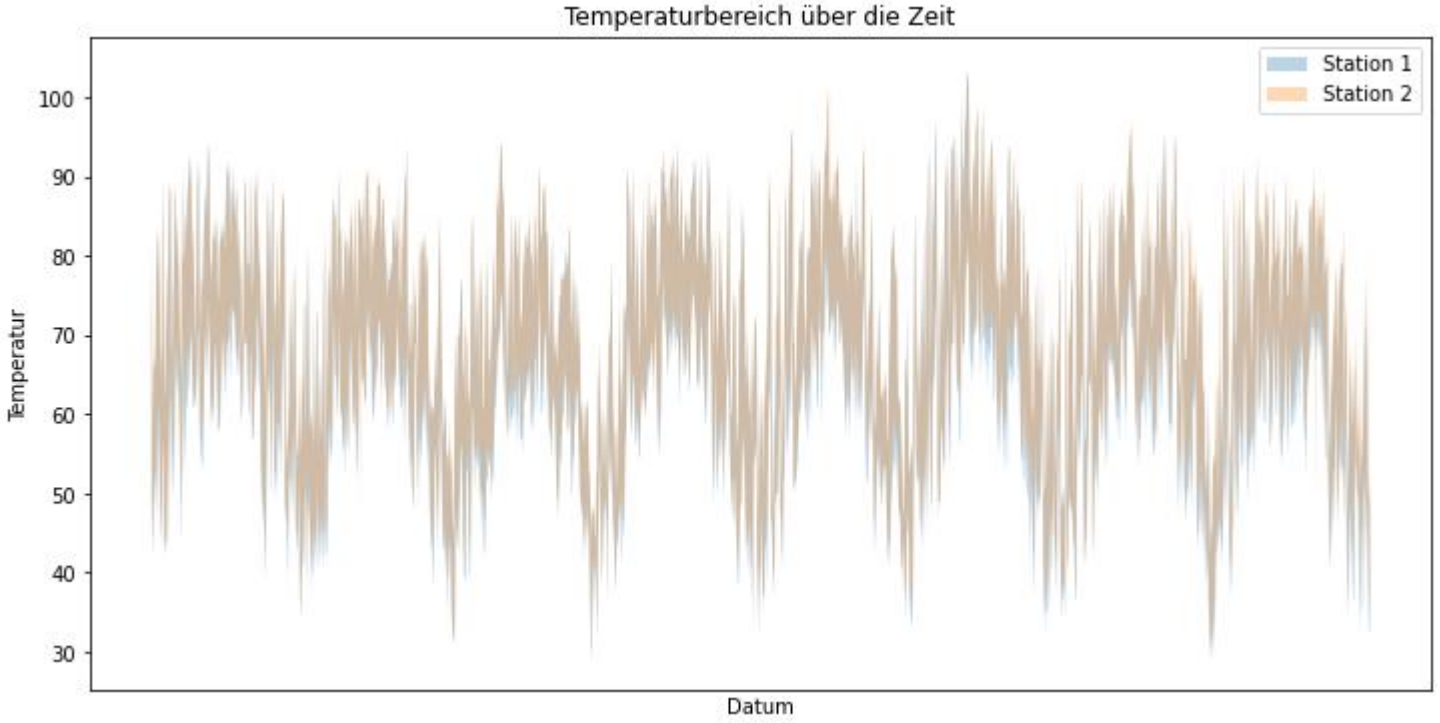
Data Understanding: Trainings- und Testdaten



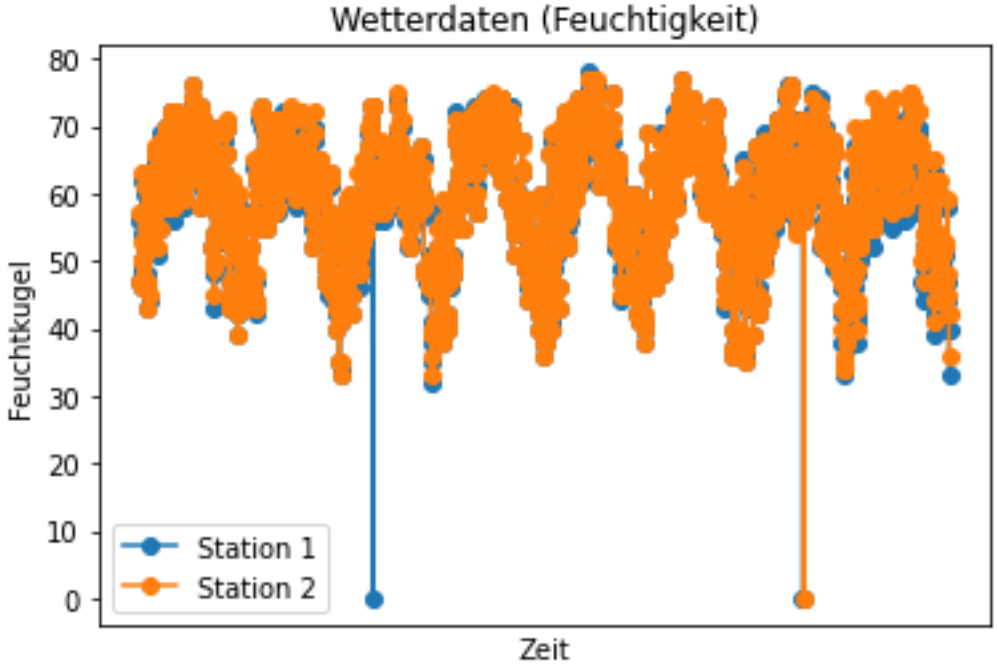


Nachweis des WestNile Virus in 5,2 % der Gesamtstichprobe (Mücken)

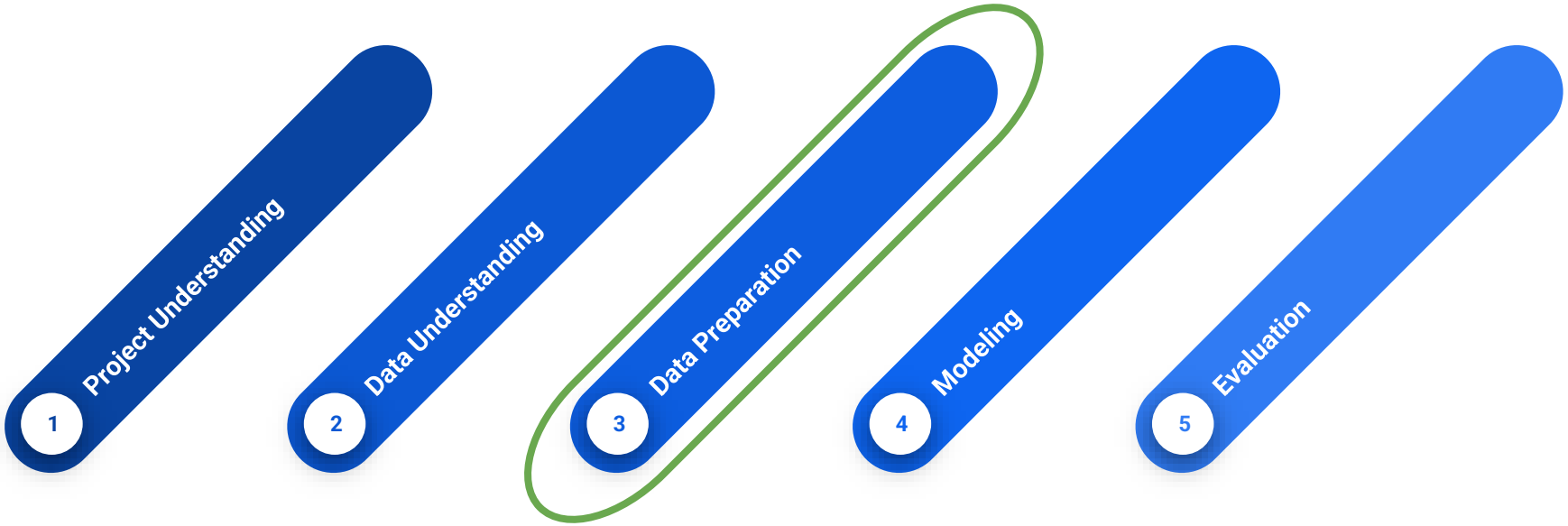
Data Understanding: Wetterdaten (Temperatur)



Data Understanding: Wetterdaten (Feuchtigkeit)



CRISP-DM MODELL



Data Preparation - Selection

RELEVANT

Trainingsdaten

- Date
- Species
- Latitude
- Longitude
- AddressAccuracy
- WnvPresent

NICHT RELEVANT

Trainingsdaten

- Address
- Block
- Street
- Trap
- AddressNumberAndStreet
- NumMosquitos

Spraydaten

Data Preparation - Selection: Wetterdaten

RELEVANT

- Station
- Date
- Tmax
- Tmin
- WetBulb
- ResultSpeed
- ResultDir

NICHT RELEVANT

- SeaLevel
- Tavg
- DewPoint
- StnPressure
- Sunset
- Sunrise
- Cool
- Heat
- Depth
- CodeSum
- Water1
- SnowFall
- PrecipTotal
- AvgSpeed
- Depart

Data Preparation: Cleaning & Transformation

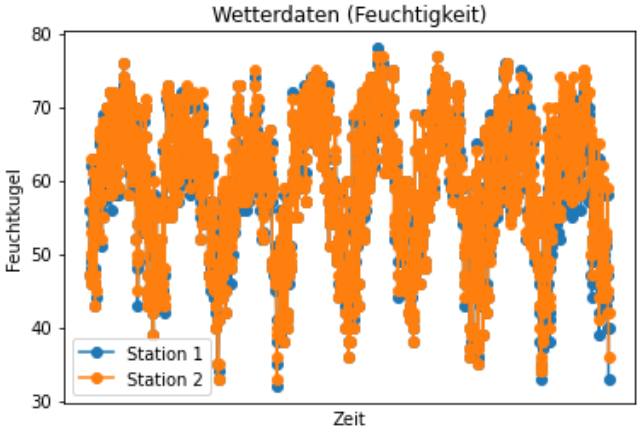
Trainingsdaten

- Mapping der Mückenarten (Integer)
- Normierung des Datums auf 01.01.1970 (Integer)

Wetterdaten

- Umwandlung “WetBulb” zu Datentyp Integer
- Ersetzung Missing Values (M) mit 0
- Ersetzung Ausreißer in der Spalte “WetBulb” durch Durchschnittswert
- Normierung des Datums auf 01.01.1970 (Integer)

```
"CULEX PIPIENS/RESTUANS": 1,
"CULEX RESTUANS": 2,
"CULEX PIPIENS": 3,
"CULEX SALINARIUS": 4,
"CULEX TERRITANS": 5,
"CULEX TARSALIS": 6,
"UNSPECIFIED CULEX": 7,
"CULEX ERRATICUS": 8
```

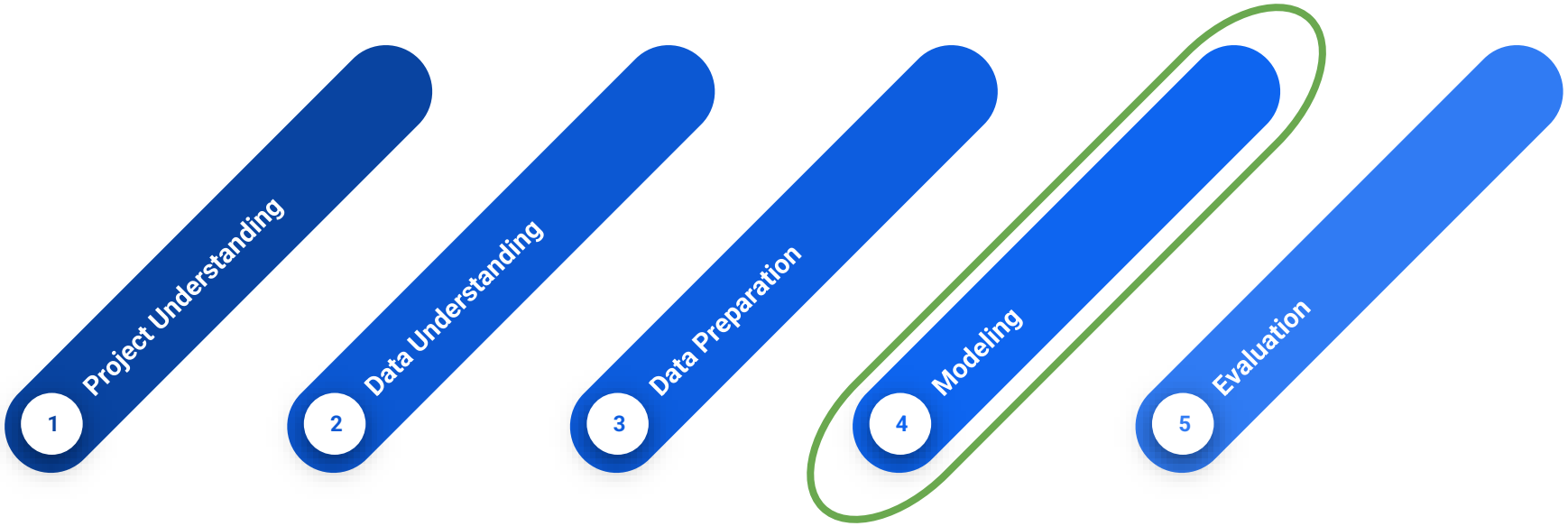


Data Preparation: Integration

- Hinzufügen der Spalte “Station” in die Trainingsdaten:
  - Latitude > 41.8905: Station 1
  - Latitude <= 41.8905: Station 2
- Merge der Wetterdaten mit den Trainingsdaten über Datum und Station

	WnvPresent	Latitude	Longitude	AddressAccuracy	Station	Tmax	Tmin	WetBulb	ResultSpeed	ResultDir	Species	Date
0	0	41.954690	-87.800991	9	1.0	88	60	65.0	5.8	18	1	1180396800
1	0	41.954690	-87.800991	9	1.0	88	60	65.0	5.8	18	2	1180396800
2	0	41.994991	-87.769279	9	1.0	88	60	65.0	5.8	18	2	1180396800
3	0	41.974089	-87.824812	8	1.0	88	60	65.0	5.8	18	1	1180396800
4	0	41.974089	-87.824812	8	1.0	88	60	65.0	5.8	18	2	1180396800

CRISP-DM MODELL



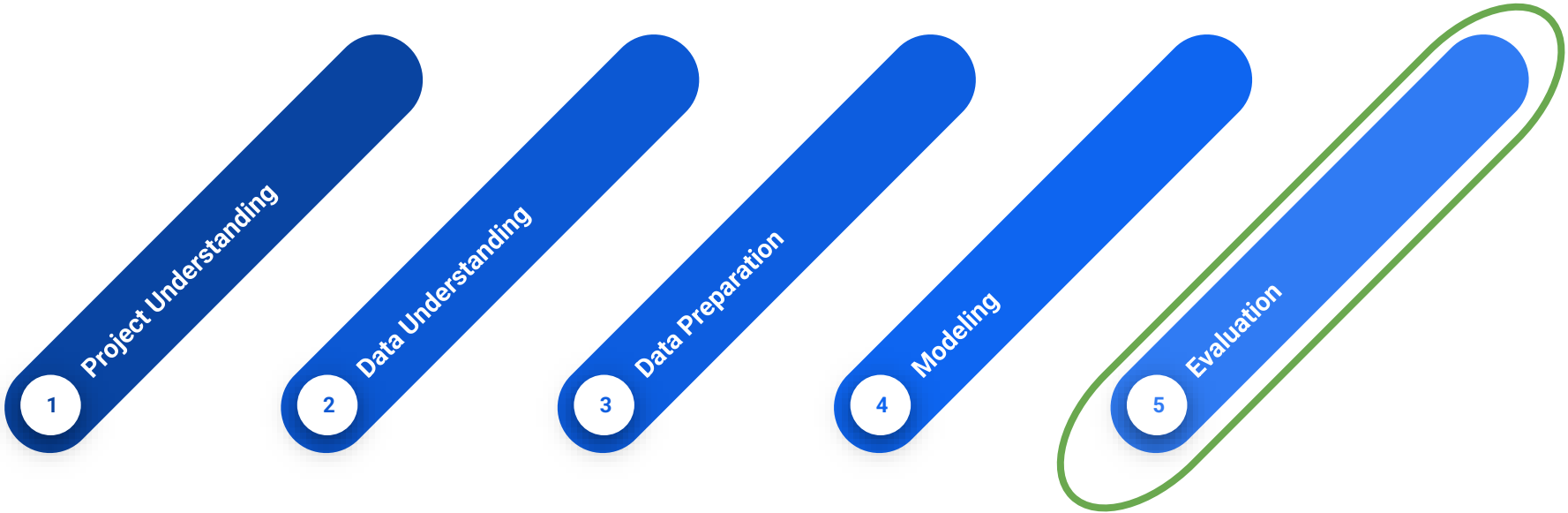
Modeling

- Aufteilung der Trainingsdaten in Trainings- und Testdaten (80/20) bei:
  - Random Forest Classifier
  - Support Vector Machine
  - Logistische Regression
  - k-NN Classifier: n_neighbors = 3
- K-fold Kreuzvalidierungsgenerator bei SVM und logistischer Regression mit k = 5

Modeling

Modell	Genauigkeit	Präzision	Recall	F1-Score	Konfusions-matrix
k-NN Classifier	0.8919562113279391	0.7307692307692307	0.0794979079497908	0.14339622641509434	[19 7 220 1855]
Support Vector Machine	0,886244645406949	0,0	0,0	0,0	[0 0 239 1862]
Logistische Re-gression	0,886244645406949	0,0	0,0	0,0	[0 0 239 1862]
Random Forest Classifier (10 verwendete Bäume)	0.9876308277830638	0.9649122807017544	0.6111111111111112	0.7482993197278912	55 2 35 2010]
Dummy Classifier	0.4902427415516421	0.11552680221811461	0.5230125523012552	0.18925056775170326	[125 957 114 905]
Majority (Dummy) Classifier	0.886244645406949	0.0	0.0	0.0	[0 0 239 1862]

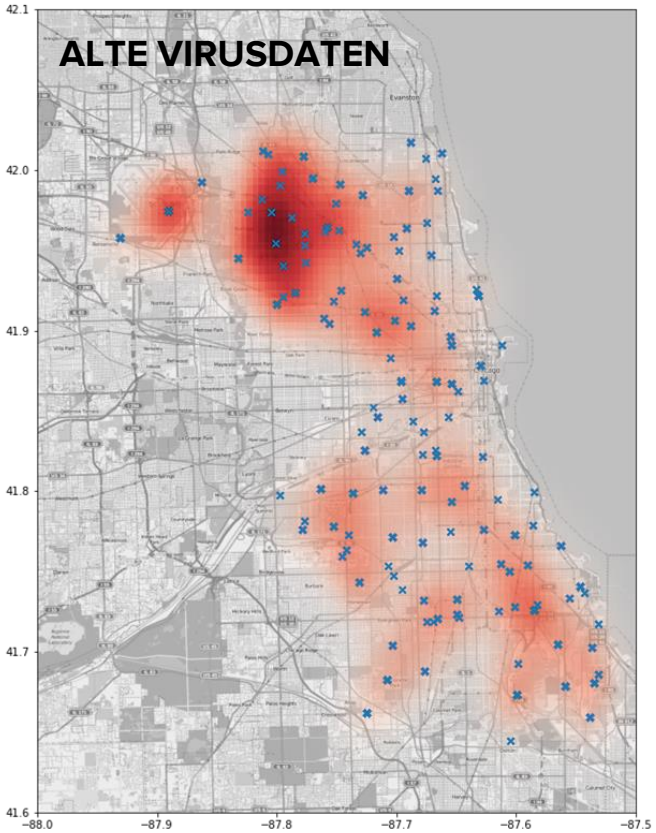
CRISP-DM MODELL

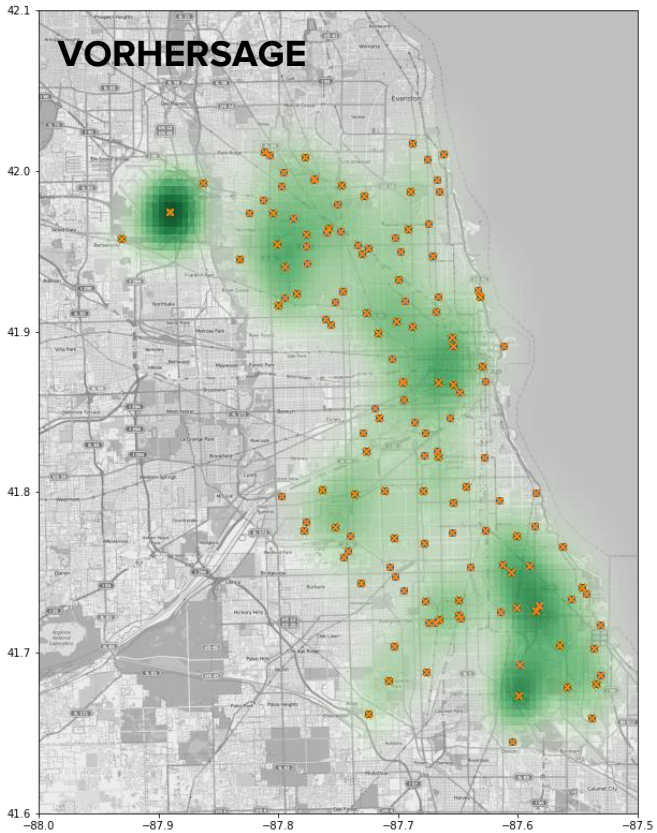


Evaluation

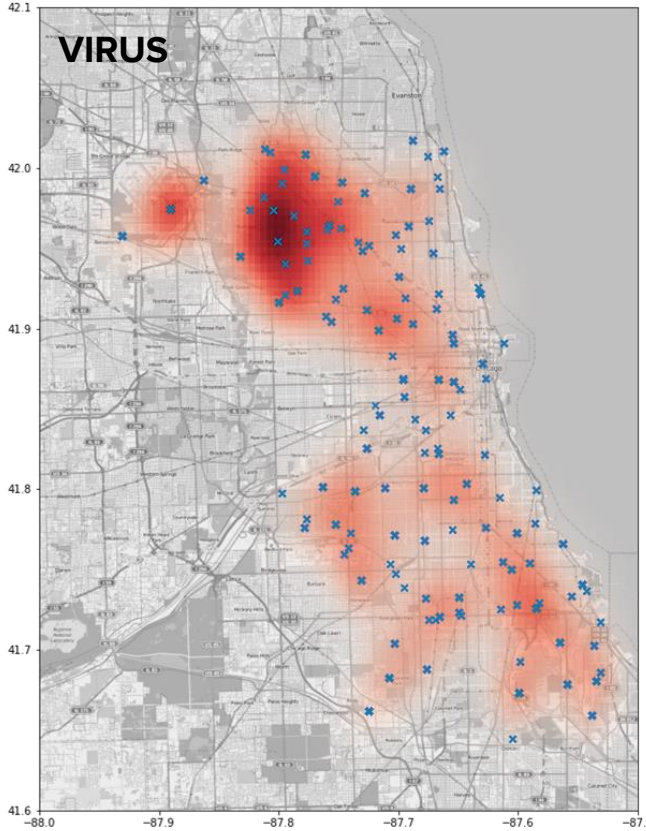
Modell	Genauigkeit	Präzision	Recall	F1-Score	Konfusions-matrix
k-NN Classifier	0.8919562113279391	0.7307692307692307	0.0794979079497908	0.14339622641509434	[19 7 220 1855]
Random Forest Classifier (10 verwendete Bäume)	0.9876308277830638	0.9649122807017544	0.6111111111111112	0.7482993197278912	55 2 35 2010]
Support Vector Machine (k-Folds)	0,886244645406949	0,0	0,0	0,0	[0 0 239 1862]
Logistische Re-gression	0,886244645406949	0,0	0,0	0,0	[0 0 239 1862]
Dummy Classifier	0.4902427415516421	0.11552680221811461	0.5230125523012552	0.18925056775170326	[125 957 114 905]
Majority (Dummy) Classifier	0.886244645406949	0.0	0.0	0.0	[0 0 239 1862]

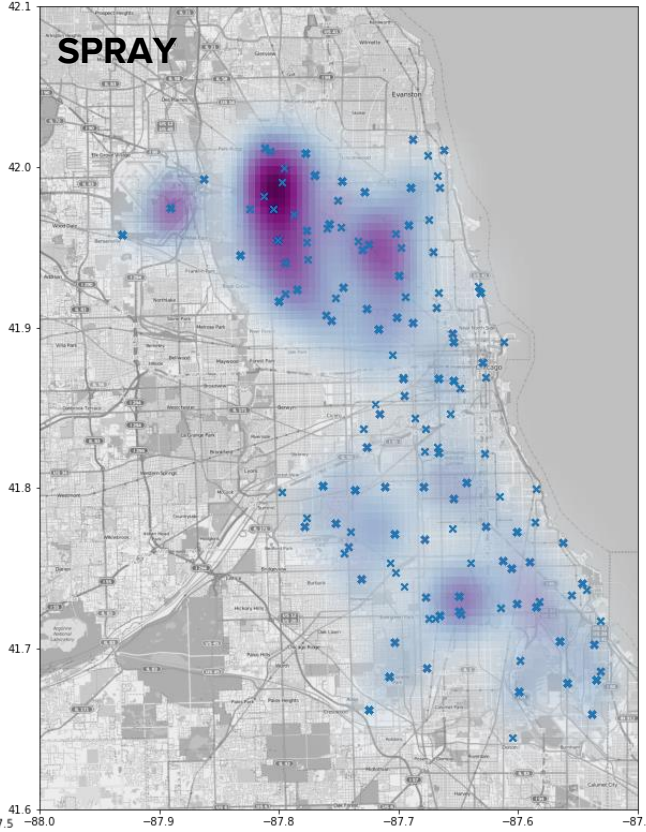
Evaluation

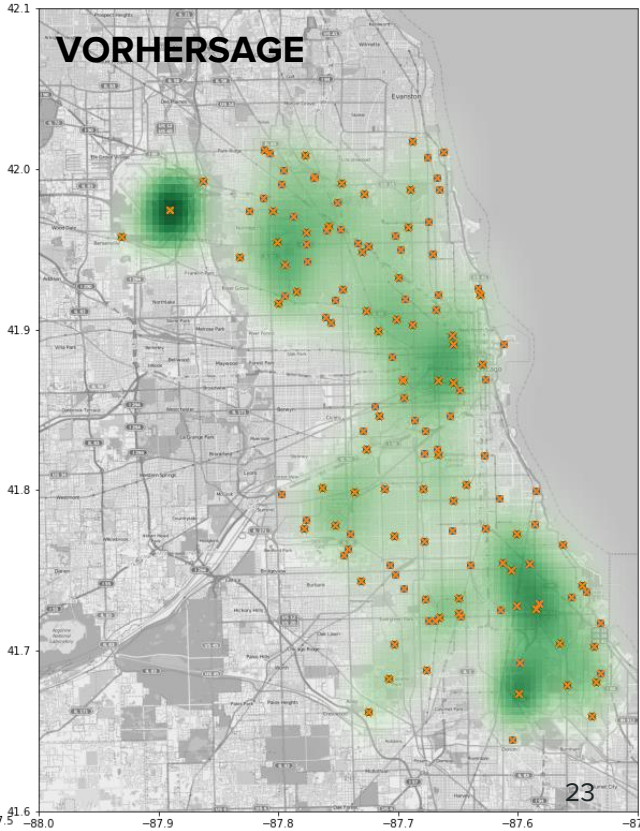

ALTE VIRUSDATEN
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5


VORHERSAGE
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5

Evaluation


VIRUS
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5


SPRAY
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5


VORHERSAGE
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6

Vielen Dank für die Aufmerksamkeit