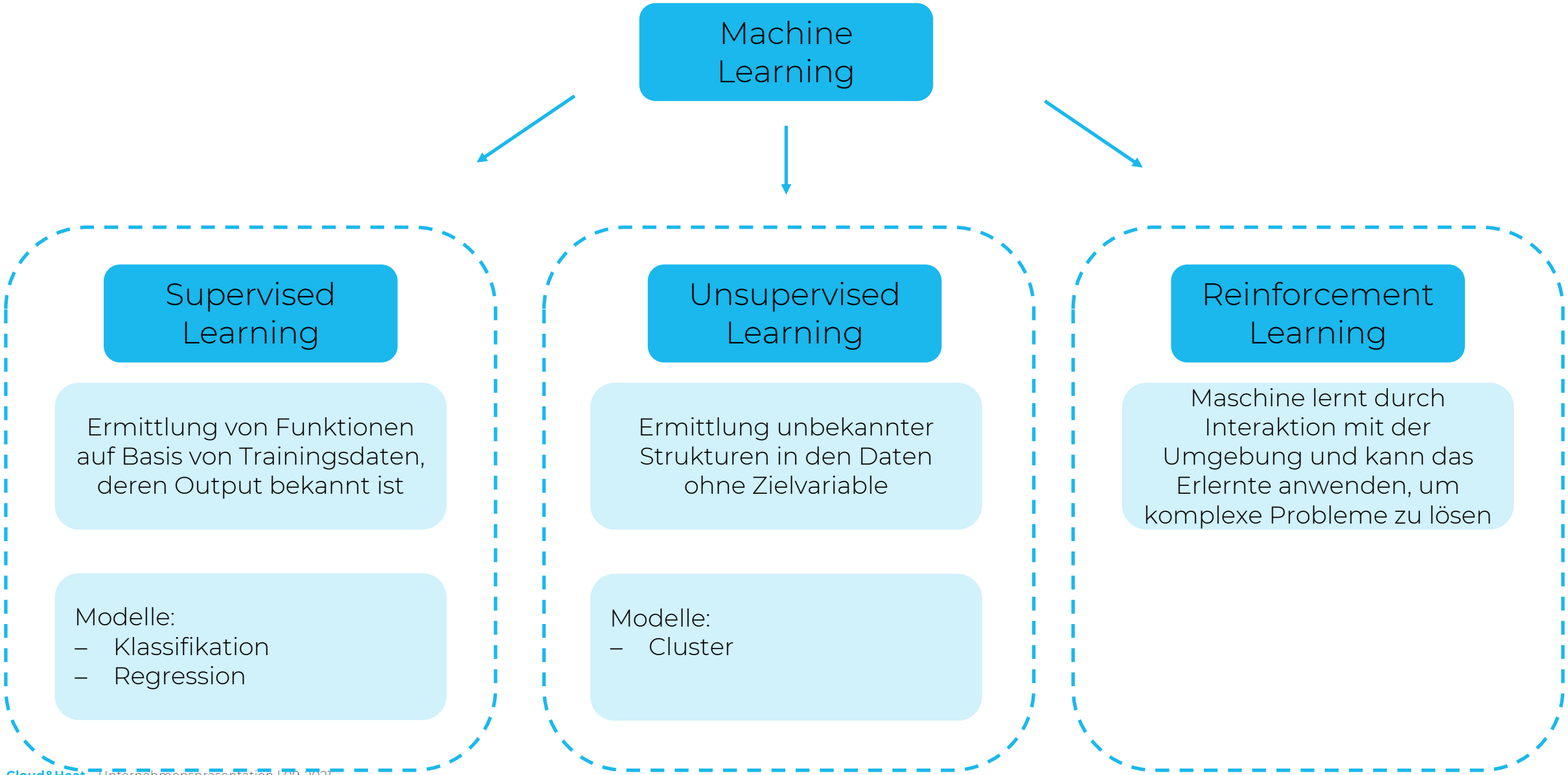
MACHINE LEARNING GRUNDLAGEN

AGENDA

- Überblick ML-Algorithmenklassen
- Beispiel
- Data Science Workflow
  - Prozess-/Datenverständnis
  - Datenvorbereitung
  - Modellierung
  - Evaluation

ÜBERSICHT


Machine Learning
Supervised Learning
Ermittlung von Funktionen auf Basis von Trainingsdaten, deren Output bekannt ist
Modelle:
– Klassifikation
– Regression
Unsupervised Learning
Ermittlung unbekannter Strukturen in den Daten ohne Zielvariable
Modelle:
– Cluster
Reinforcement Learning
Maschine lernt durch Interaktion mit der Umgebung und kann das Erlernte anwenden, um komplexe Probleme zu lösen

ANWENDUNGSBEISPIEL DATEN

Preis	Quadratmeter	Schlafzimmer	Badezimmer	Gegend
114200	188	4	2	Osten
114800	161	3	2	Osten
94700	183	3	2	Osten
119800	197	3	3	Osten
114600	165	3	2	Norden
151600	170	3	3	Westen
150700	200	4	2	Westen
119200	196	4	2	Osten
104000	160	3	3	Osten
132500	188	3	2	Osten
123000	173	2	2	Osten
102600	177	3	2	Norden
...	...	...	...	...

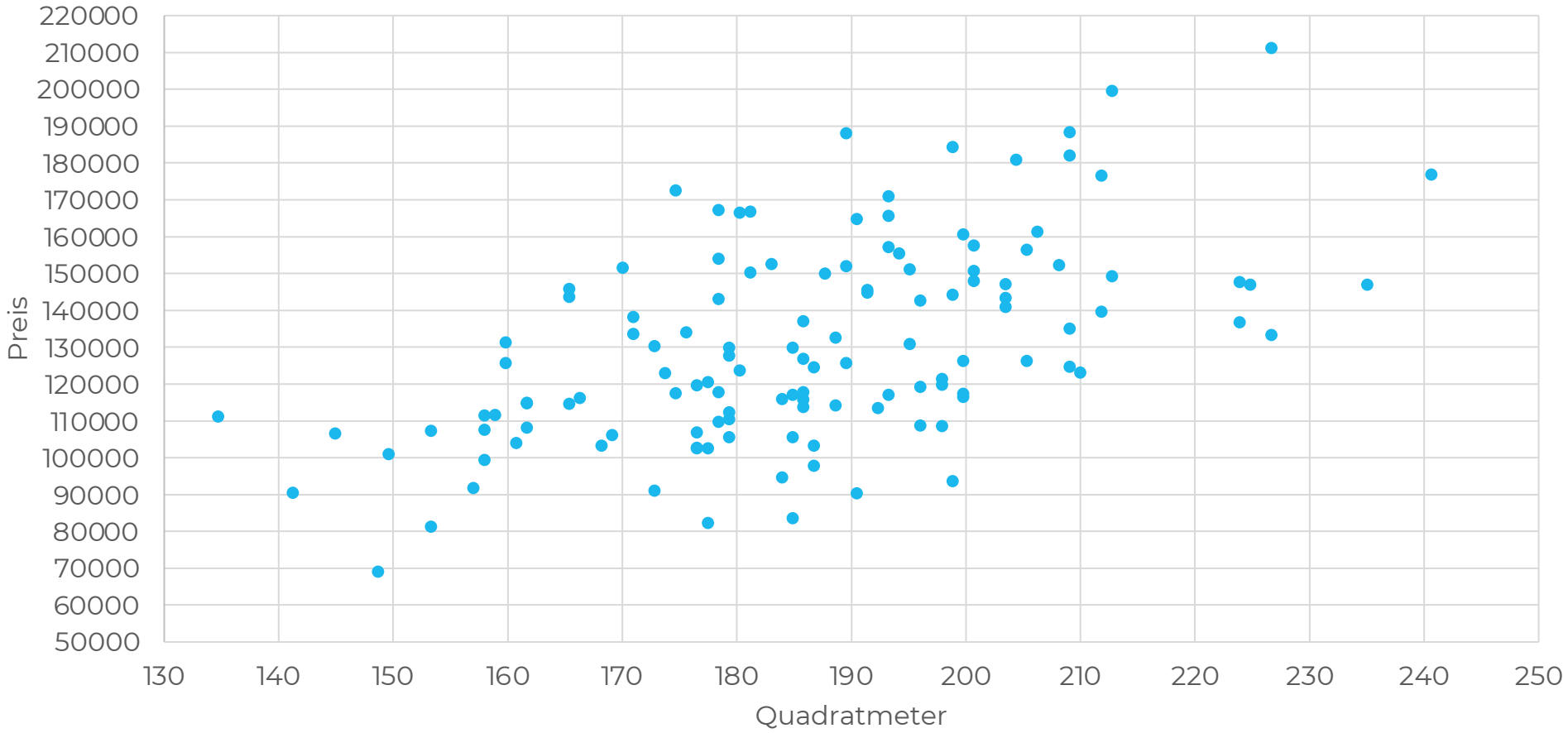
ANWENDUNGSBEISPIEL DATEN

VORHERSAGE PREIS ANH. QM

Preis	Quadratmeter
114200	188
114800	161
94700	183
119800	197
114600	165
151600	170
150700	200
119200	196
104000	160
132500	188
123000	173
102600	177
...	...

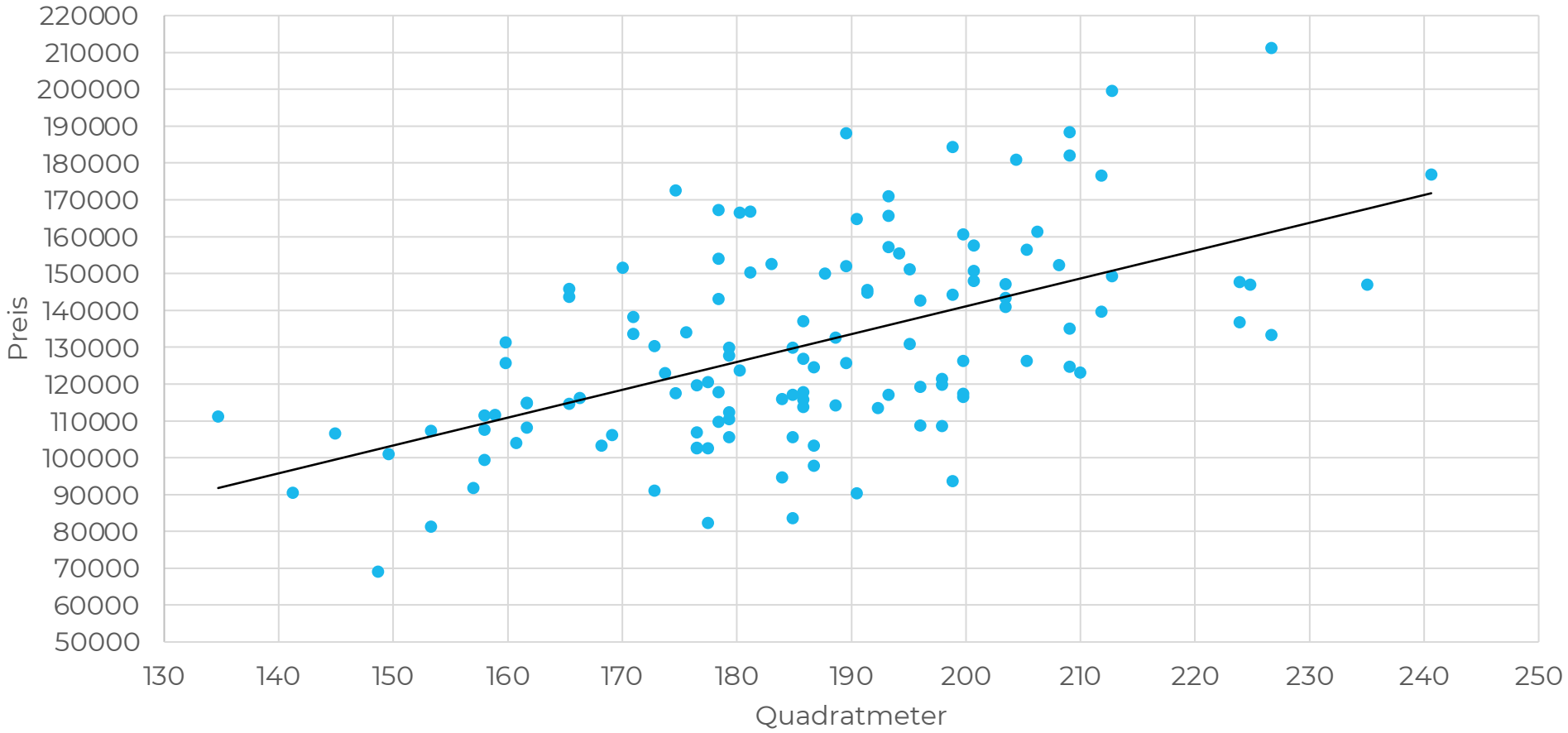
SUPERVISED LEARNING

REGRESSION



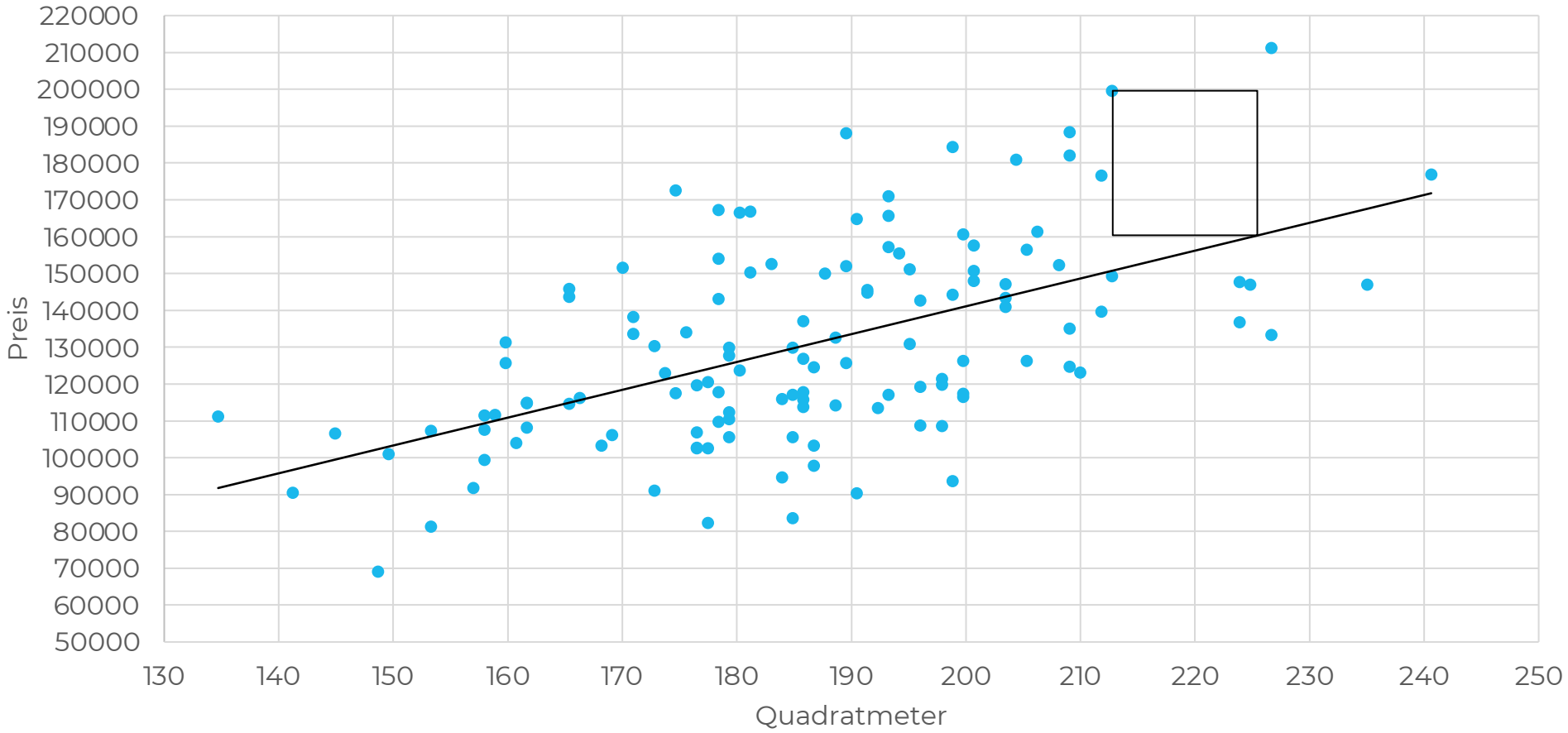
SUPERVISED LEARNING

REGRESSION



SUPERVISED LEARNING

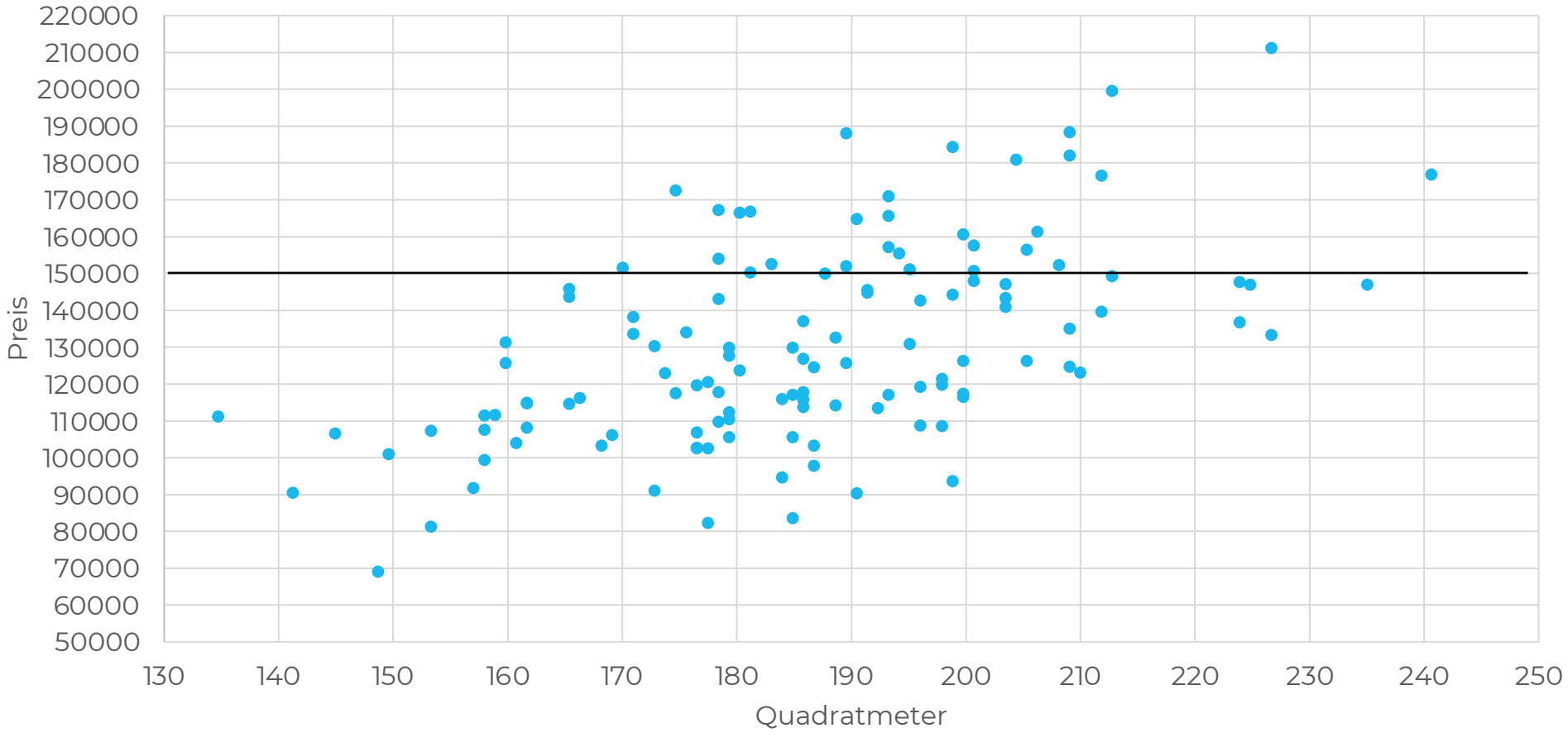
REGRESSION



DEMO

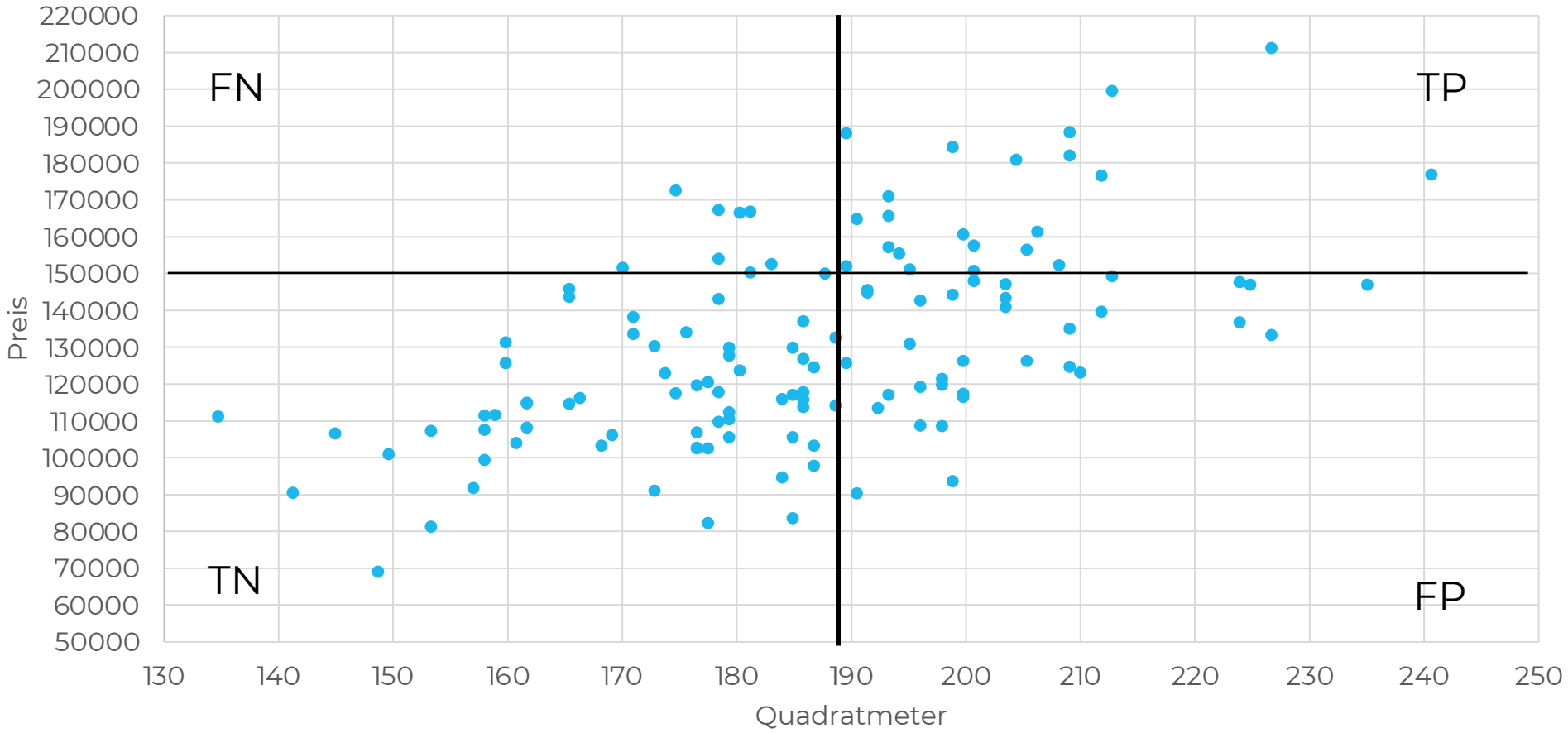
SUPERVISED LEARNING

KLASSIFIKATION



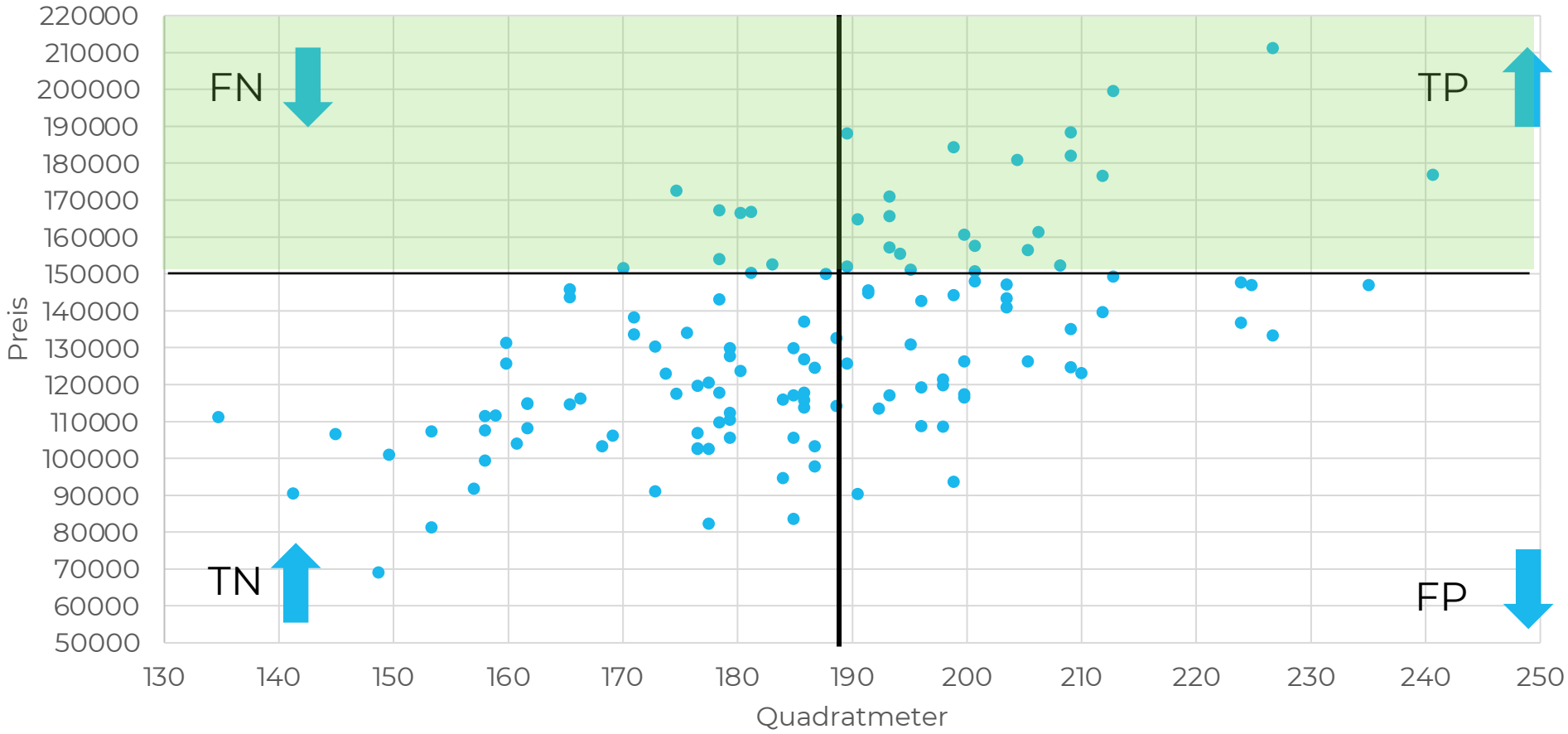
SUPERVISED LEARNING

KLASSIFIKATION



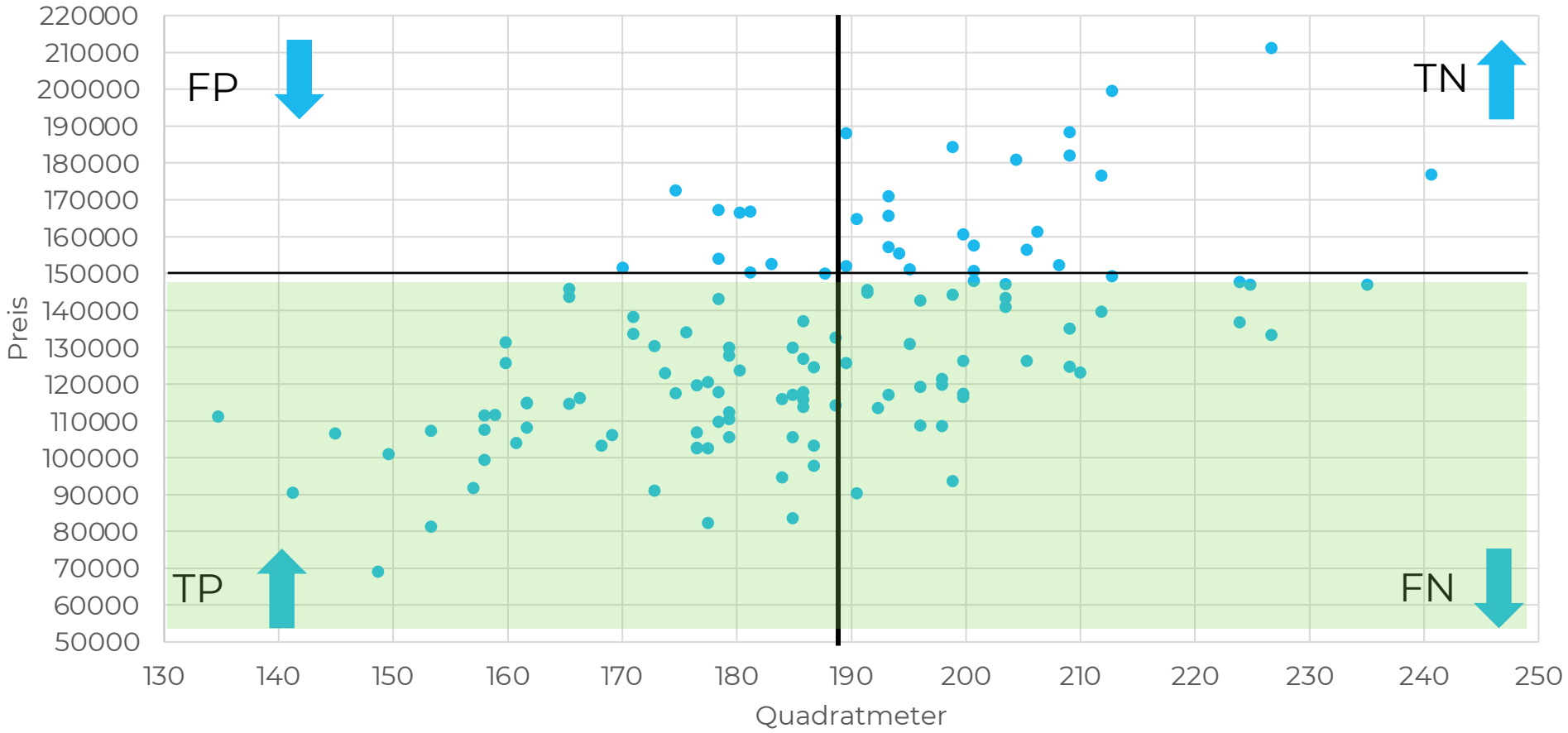
SUPERVISED LEARNING

KLASSIFIKATION – TEURE HÄUSER



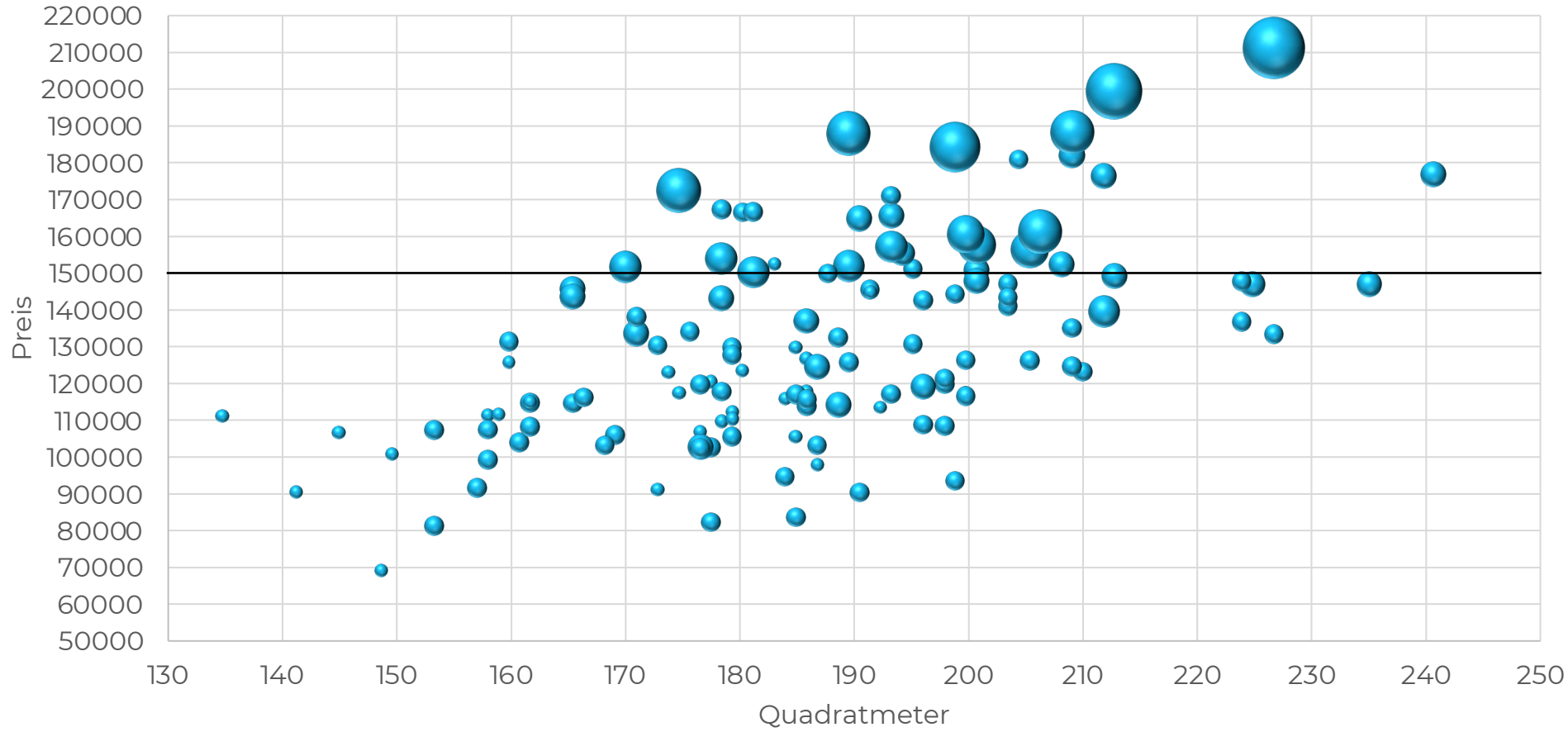
SUPERVISED LEARNING

KLASSIFIKATION – GÜNSTIGE HÄUSER



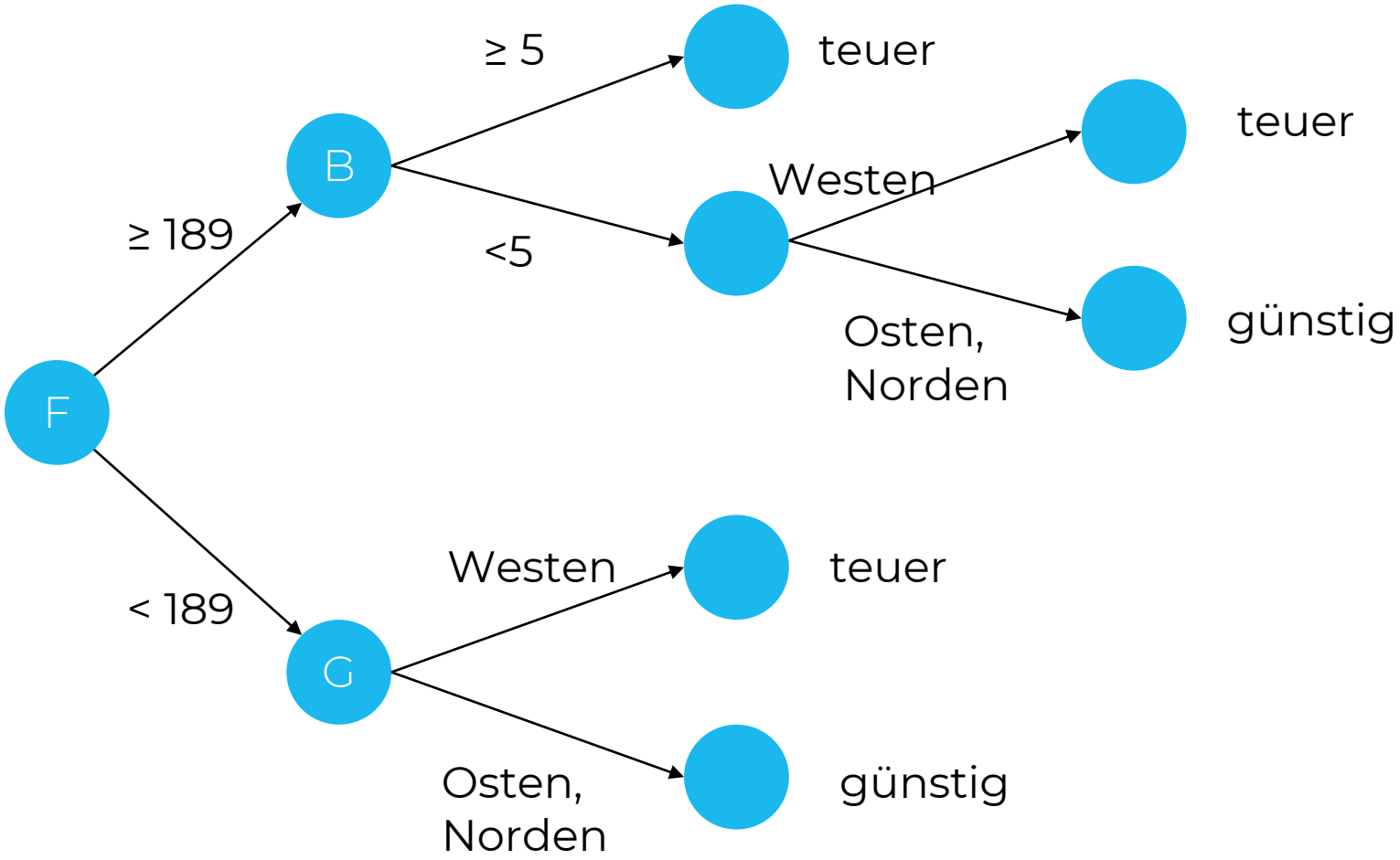
SUPERVISED LEARNING

KLASSIFIKATION

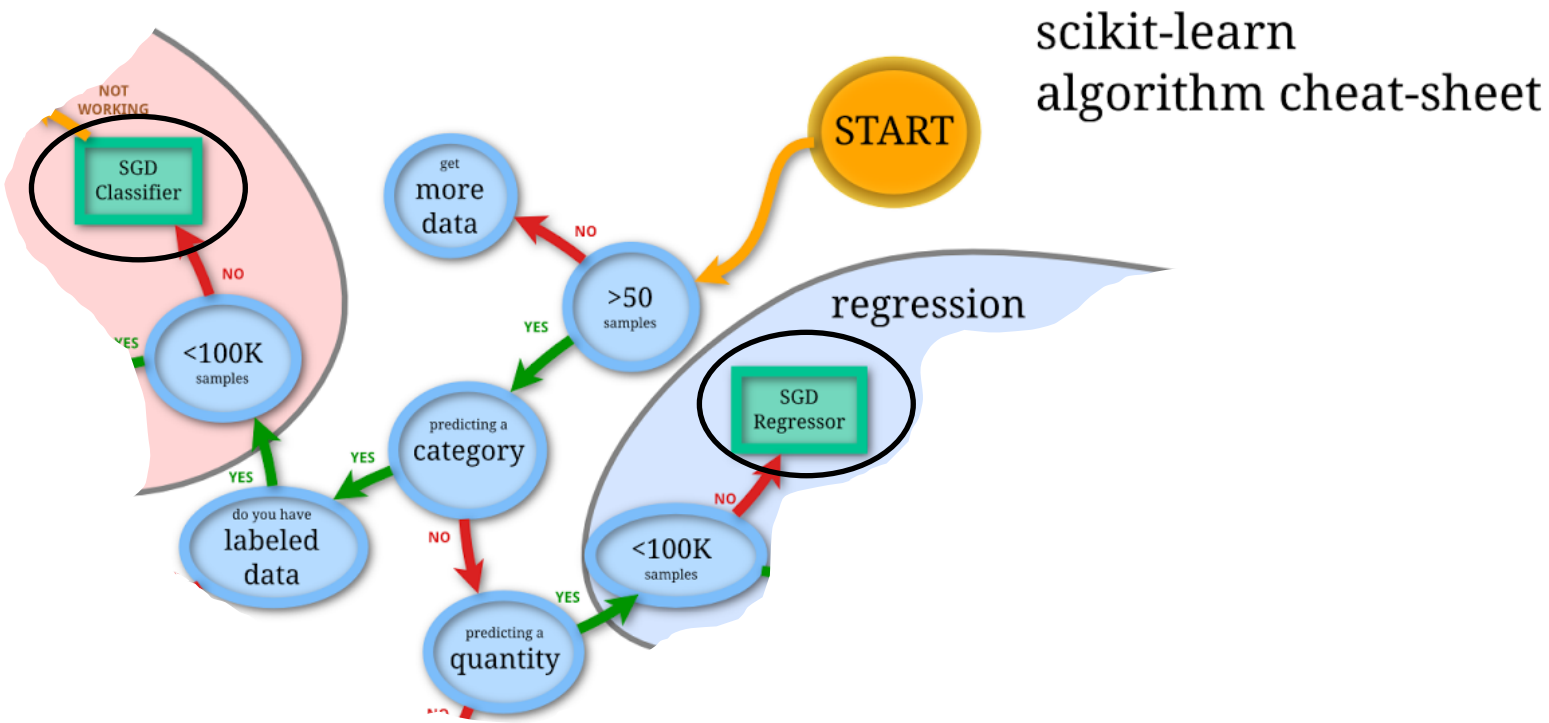


Größe der Blase: Anzahl an Schlafzimmern

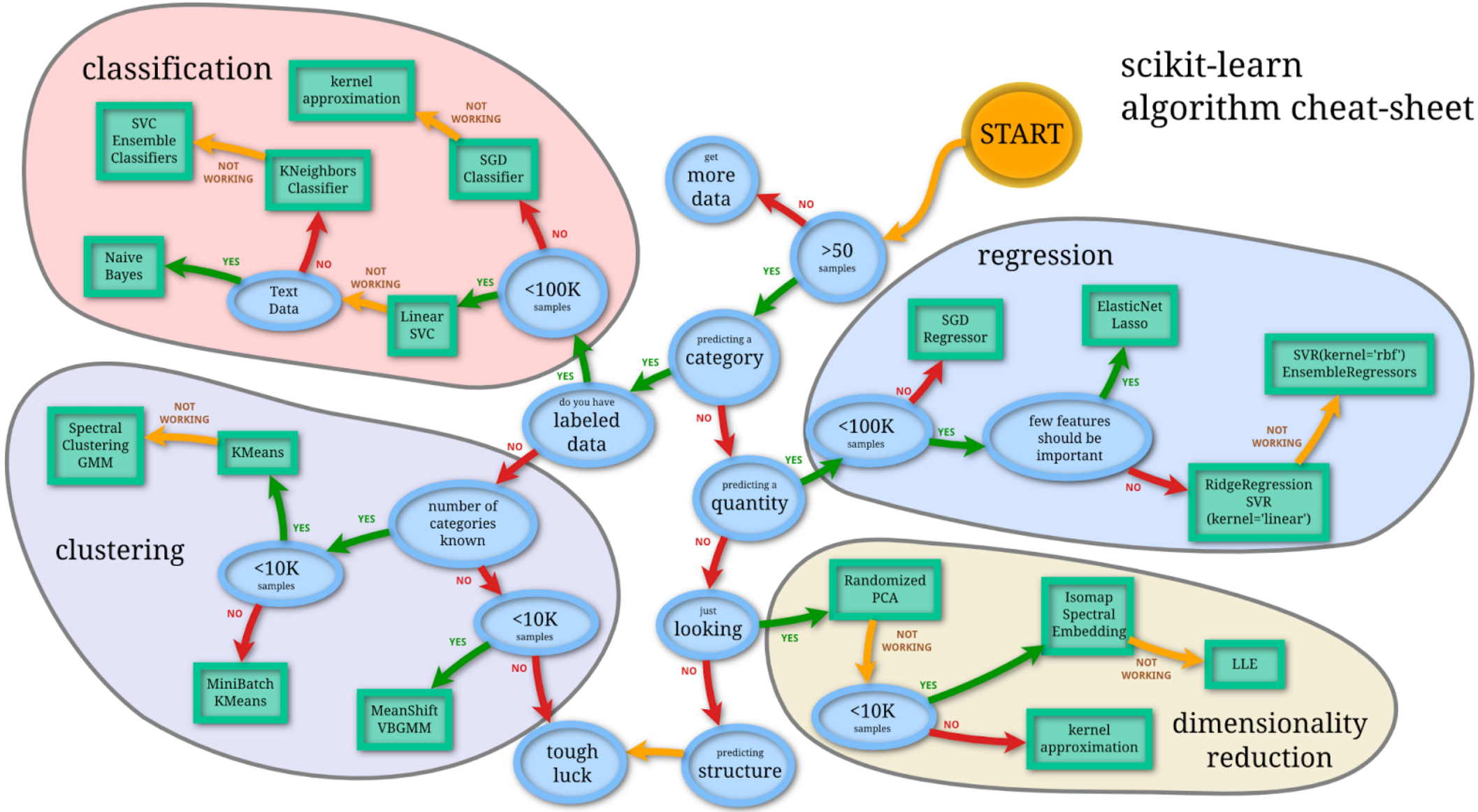
DECISION TREES



ÜBERSICHT


scikit-learn
algorithm cheat-sheet
START
get
more
data
NO
>50
samples
YES
NOT
WORKING
SGD
Classifier
NO
<100K
samples
YES
predicting a
category
YES
do you have
labeled
data
YES
NO
predicting a
quantity
YES
regression
SGD
Regressor
NO
<100K
samples

ÜBERSICHT


scikit-learn
algorithm cheat-sheet
START
get more data
>50 samples
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
kernel approximation
SVC Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K samples
clustering
Spectral Clustering GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
MeanShift VBGMM
regression
SGD Regressor
ElasticNet Lasso
SVR(kernel='rbf') EnsembleRegressors
<100K samples
few features should be important
RidgeRegression SVR (kernel='linear')
dimensionality reduction
Randomized PCA
Isomap Spectral Embedding
LLE
<10K samples
kernel approximation
YES
NO
NOT WORKING

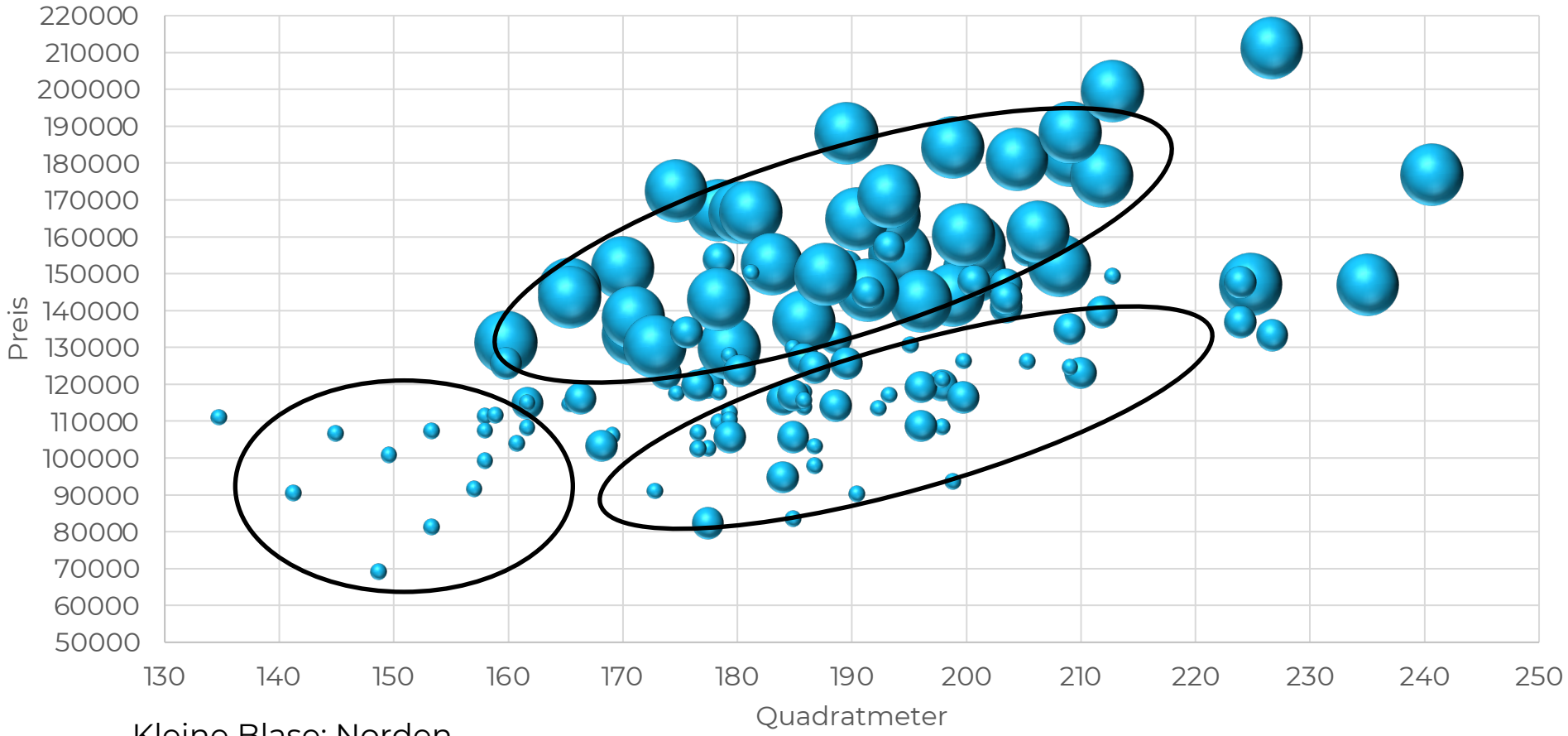
ANWENDUNGSBEISPIEL DATEN

CLUSTER

Preis	Quadratmeter	Gegend
114200	188	Osten
114800	161	Osten
94700	183	Osten
119800	197	Osten
114600	165	Norden
151600	170	Westen
150700	200	Westen
119200	196	Osten
104000	160	Osten
132500	188	Osten
123000	173	Osten
102600	177	Norden
...	...	...

UNSUPERVISED LEARNING

CLUSTER



Kleine Blase: Norden
Mittlere Blase: Osten
Große Blase: Westen

ANWENDUNGSBEISPIEL DATEN

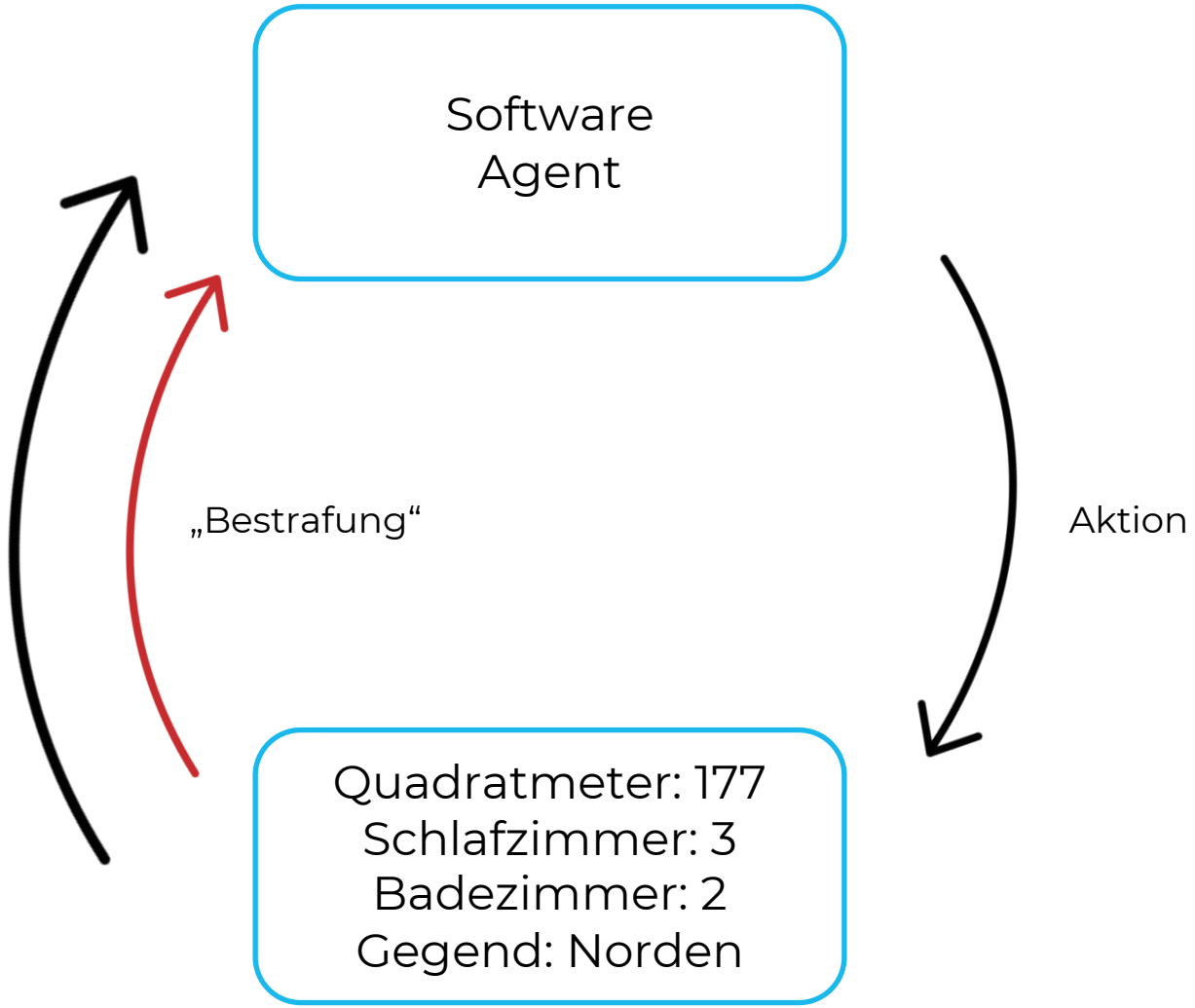
REINFORCEMENT LEARNING

Preis	Quadratmeter	Schlafzimmer	Badezimmer	Gegend
114200	188	4	2	Osten
114800	161	3	2	Osten
94700	183	3	2	Osten
119800	197	3	3	Osten
114600	165	3	2	Norden
151600	170	3	3	Westen
150700	200	4	2	Westen
119200	196	4	2	Osten
104000	160	3	3	Osten
132500	188	3	2	Osten
123000	173	2	2	Osten
102600	177	3	2	Norden

REINFORCEMENT LEARNING

MÖGLICHST HOHER PREIS FÜR EIN HAUS MIT MAX. 180 QUADRATMETERN

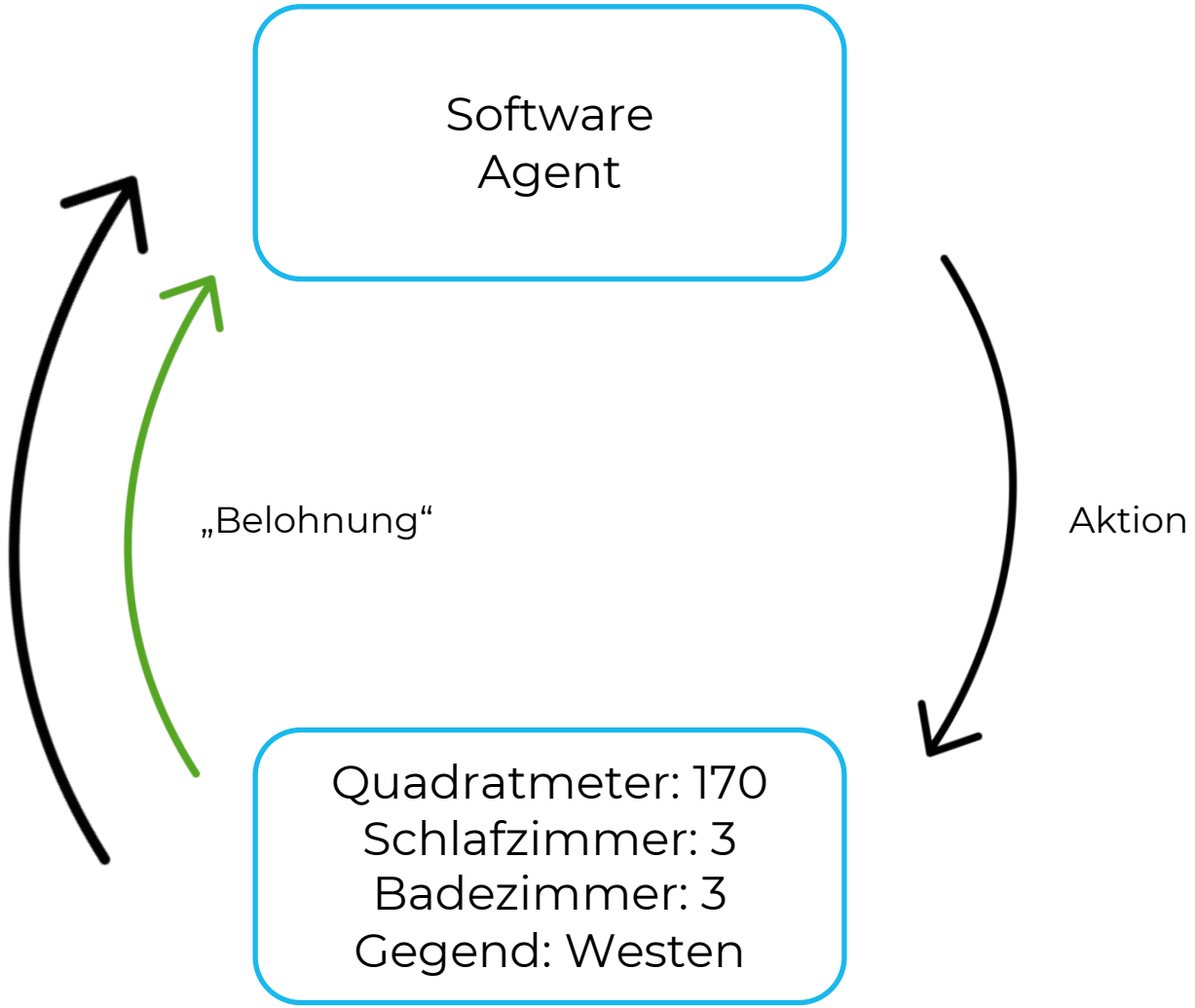
Preis	Quadratmeter	Schlafzimmer	Badezimmer	Gegend
114200	188	4	2	Osten
114800	161	3	2	Osten
94700	183	3	2	Osten
119800	197	3	3	Osten
114600	165	3	2	Norden
151600	170	3	3	Westen
150700	200	4	2	Westen
119200	196	4	2	Osten
104000	160	3	3	Osten
132500	188	3	2	Osten
123000	173	2	2	Osten
102600	177	3	2	Norden



REINFORCEMENT LEARNING

MÖGLICHST HOHER PREIS FÜR EIN HAUS MIT MAX. 180 QUADRATMETERN

Preis	Quadratmeter	Schlafzimmer	Badezimmer	Gegend
114200	188	4	2	Osten
114800	161	3	2	Osten
94700	183	3	2	Osten
119800	197	3	3	Osten
114600	165	3	2	Norden
151600	170	3	3	Westen
150700	200	4	2	Westen
119200	196	4	2	Osten
104000	160	3	3	Osten
132500	188	3	2	Osten
123000	173	2	2	Osten
102600	177	3	2	Norden

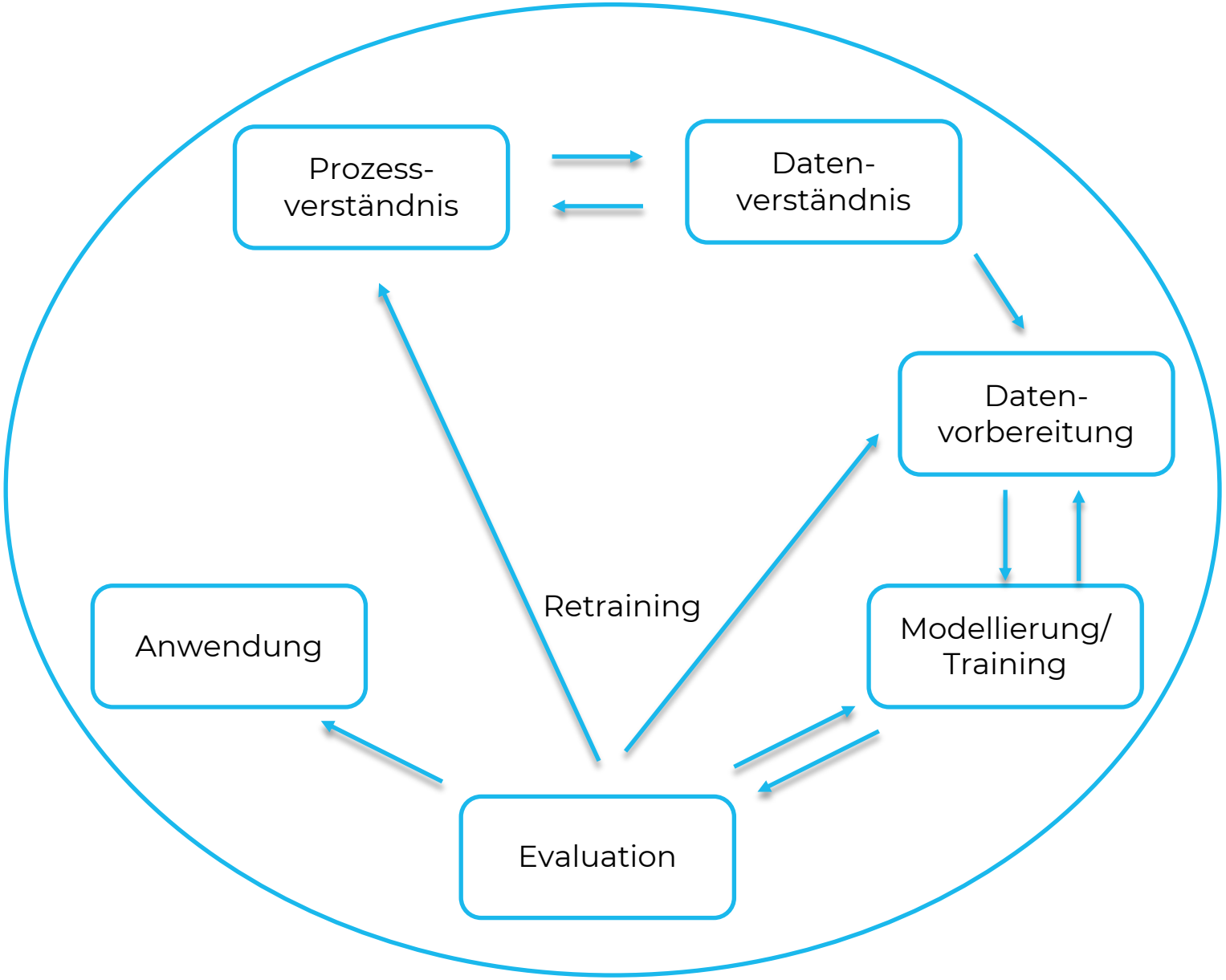


BEGRIFFSDEFINITION

<table>
<tr><th>Statistik/
Mathematik</th><th>Data Science</th><th>Künstliche Intelligenz
(letztes Webinar)</th><th>Machine Learning
(letztes Webinar)</th><th>Deep Learning
(nächstes Webinar)</th></tr>
<tr><td rowspan="3">– Sammeln, Auswerten, Interpretieren, Präsentieren und Organisieren von Daten
– Basis für alle Bereiche der KI, Neuronale Netze, Data Science</td><td rowspan="3">– Bereich an der Kreuzung zwischen Computer Science, Mathematik und Wissen von bestimmten Domänen
– Anwendung zur Gewinnung von Wissen aus Daten
– Wird bspw. genutzt um die Unternehmens-steuerung zu optimieren oder die Entscheidungs-findung zu unterstützen</td><td rowspan="3">– Überbegriff für Anwendungen, bei denen Maschinen bestimmte Entscheidungs-strukturen des Menschen nachbilden sollen
– Fähigkeit zu lernen ist die Hauptanforderung an KI-Systeme
– System muss mit Unsicherheiten und probabilistischen Informationen umgehen können</td><td colspan="2">– Teilbereich der KI</td></tr>
<tr><td>– Algorithmen, deren Leistung sich durch die Zugabe von strukturierten bekannten Zusammenhäng-en verbessert
– Modellgenauigkeit wird iterativ verbessert
– Eingriff des Menschen in die Datenanalyse und den Entscheidungspro-zess</td><td>– Algorithmen, deren Leistung sich durch die Zugabe von unstrukturierten bekannten Zusammenhängen mit Hilfe von Neuronalen Netzen verbessert
– System verknüpft das Erlernte mit neuen Daten und lernt somit ständig weiter
– Lernvorgang findet ohne Eingreifen des Menschen statt</td></tr>
<tr><td colspan="2">– Anwendung des Erlernten auf unbekannte Zusammenhänge</td></tr>
</table>

– Neuron ist eine mathematische Formel
– Input wird verarbeitet und ein Output wird generiert
– Viele künstliche Neuronen arbeiten zusammen und ergeben ein Neuronales Netz

DATA SCIENCE WORKFLOW



PROZESS-/DATENVERSTÄNDNIS

Wichtig ist das Kennenlernen und Verstehen der Daten, sowie der fachliche Bezug: **Was genau schaue ich mir überhaupt an ?**

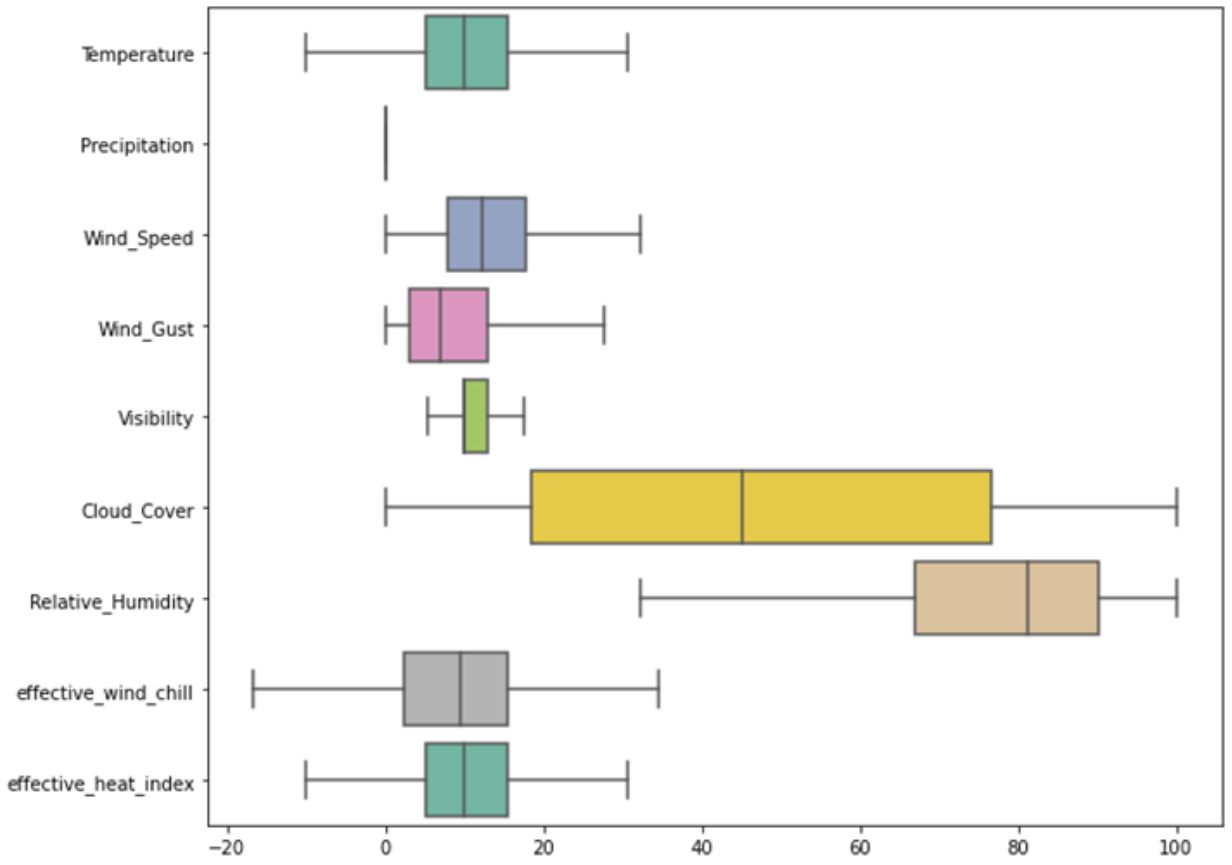
Wichtig: Die Problemstellung muss klar sein. Was soll optimiert, vorhergesagt, analysiert, oder klassifiziert werden ?

Typische Methoden:

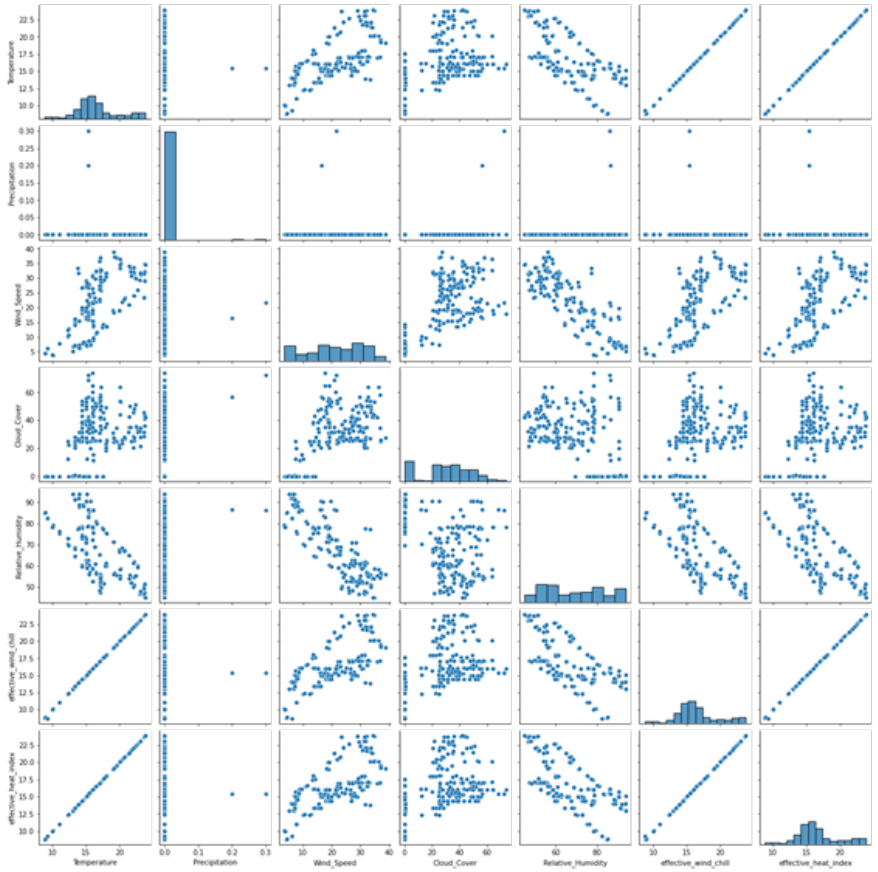
- Explorative und statistische Datenanalyse
  - **Explorativ** im Sinne von „**Erkunden der Daten**“ mit Hilfe von **Visualisierung** und Datenverläufen
  - **Statistisch** anhand von **Verteilungsfunktionen** oder anderen **Visualisierungen** und **statistischen Werkzeugen**
  - Diagramme mit gleitenden Durchschnitten, Boxplots, Pairplots, Heatmaps, Verteilungsdiagramme u.v.m.

DATENVERSTÄNDNIS

BEISPIELE



- Boxplot



- Pairplot

DATENVORBEREITUNG

- Wodurch wird die Datenqualität beeinträchtigt?
  - **Datenlücken** (Missing Values), z.B. als **NaNs** oder im Datenkontext unrealistische Werte als Platzhalter wie -1000, 0, 1000 bezeichnet
  - **Anomalien**, z.B. **Ausreißer**, **unerklärbare Werte**
  - **Ungleiche Datenverteilung**, beispielsweise mehr Beobachtungen in der unabhängigen Variable A als in B
  - Im Datenverlauf sich plötzlich **ändernde Datenniveaus**
  - **Unterschiedliche Datentypen** je Merkmalsträger (Feature) oder sogar innerhalb eines Merkmalsträgers
  - Sehr **viele** oder sehr **wenige Features**

Kernaspekt ist das **Bereinigen** und **Vorbereiten** der Daten für die nachfolgende Phase, die Modellbildung.

Typische Vorbereitungen:

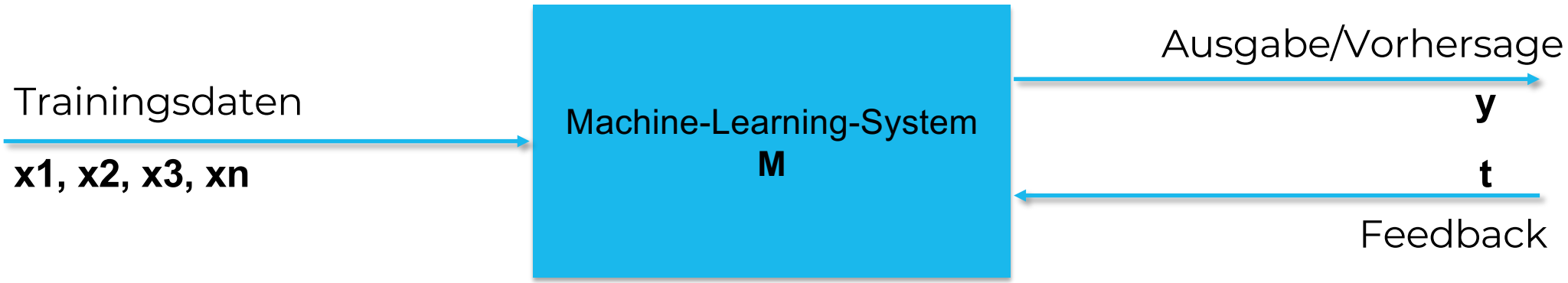
- **Normalisieren** (Anpassung der Daten, sodass alle eine gemeinsame Skala haben)
- Teilen des Datensatzes in **Trainingsdaten**, **Testdaten**

  und **Validierungsdaten**
- Datenqualität Verbessern

Ziel: Daten so vorbereiten, dass ein **Modell** möglichst **gut trainiert** werden kann.

MODELLIERUNG

PRINZIP EINES MODELLS



- Entscheidungsbaum
- Künstliches neuronales Netz
- Lineare Regression
- etc.

MODELLIERUNG

- **Kernaspekt** ist das **Entwerfen** und **Trainieren** geeigneter Algorithmen des maschinellen Lernens bzw. geeigneter künstlicher neuronaler Netze.

- **Typische Tätigkeiten**:
  - **Auswahl** eines oder mehrerer zur Problemstellung **passender Modelle** → Klassifikation/Regression
  - **Aufbau** des **Modells** mit sogenannten **Hyperparametern**
    - **Klassische Hyperparameter:**

      Iterationen/Epochen (bei KNNs), also Trainingsdurchgänge
      - **Lernrate**
      - **Stichprobe (Subsample)**
      - **Aktivierungsfunktion**

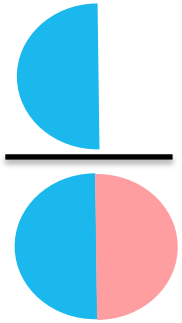
  - **Start** des **Trainingsprozesses**
    - Kann je nach Modell, Hyperparametern, Datenmenge, Featurezahl, Rechenleistung u.v.m. unterschiedlich lang dauern, von wenigen Sekunden bis zu mehreren Tagen

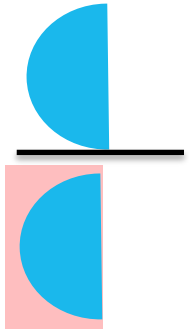
- **Ziel**: Ein **robustes**, **generalisierendes** Modell mit **guten Ergebnissen**.

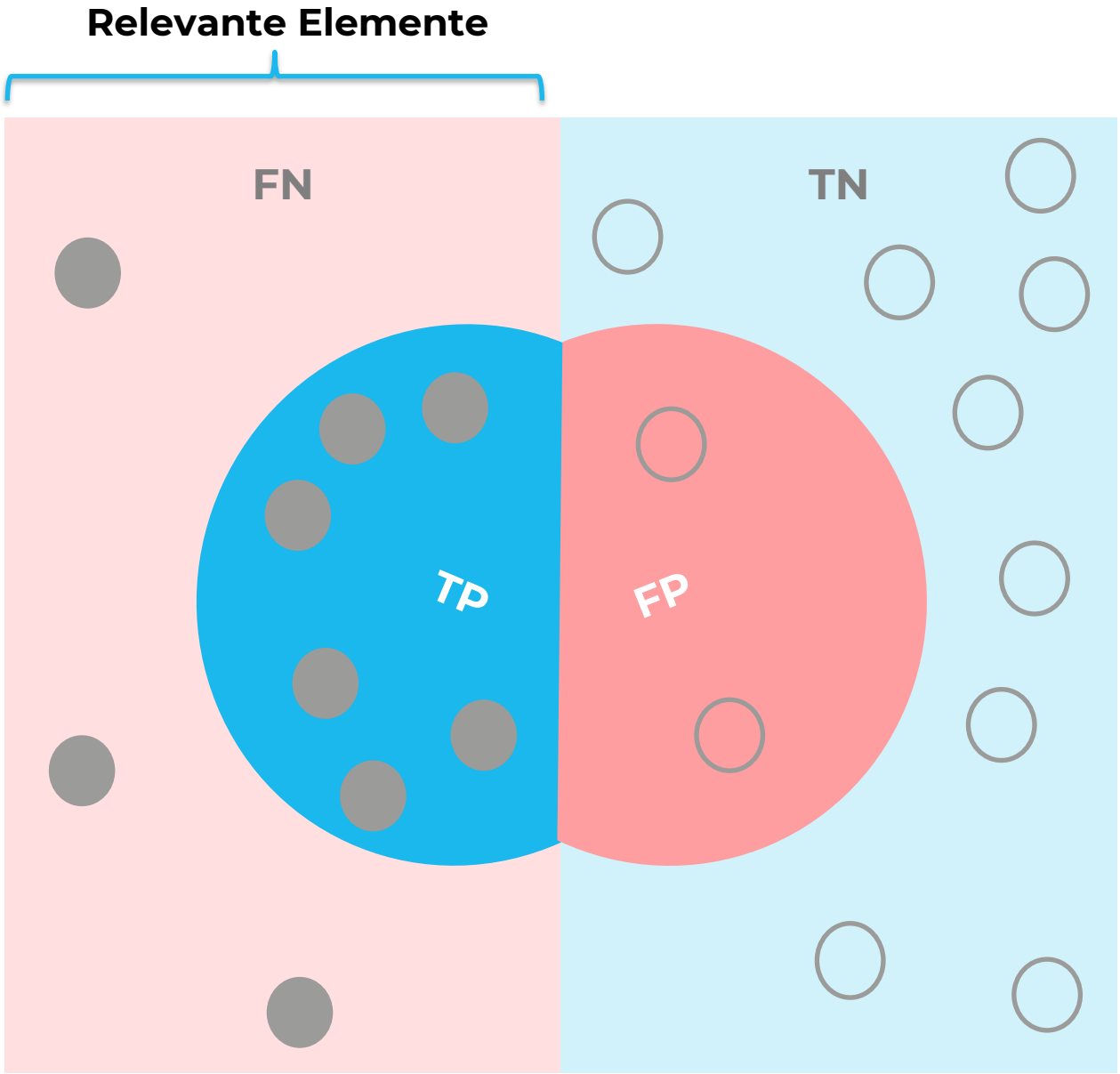
- **Außerdem**:
  - **Auswahl** und Überprüfung der Modelle **nach steigender Modellkomplexität**
    - -> Beispiel: lineare Regression ist deutlich simpler als ein komplexes Künstliches neuronales Netz

KLASSIFIKATION - EVALUATION

- Anhand eines **Testsets** kann die Qualität des Algorithmus gut eingeschätzt werden
- Verschiedene (aggregierte) Metriken helfen bei der Überprüfung und **iterativen** Verbesserung

Precision = 

Recall = 



AUSBLICK

WEBINAR-REIHE

12.01

WARUM?

Überblick, Definition,
Anwendungsfälle,
Herausforderungen

26.01.

WAS?

Grundbegriffe und
Systematisierung

Data Science Workflow

09.02.

WIE?

Neuronale Netze

Frameworks, Operations

WER WIR SIND

CLOUD&HEAT IN ZAHLEN

2011

GRÜNDUNG

in Dresden, Deutschland

97

MITARBEITER

aus 10 verschiedenen Ländern

4

STANDORTE

Dresden, Frankfurt a. Main, Madrid und Dubai