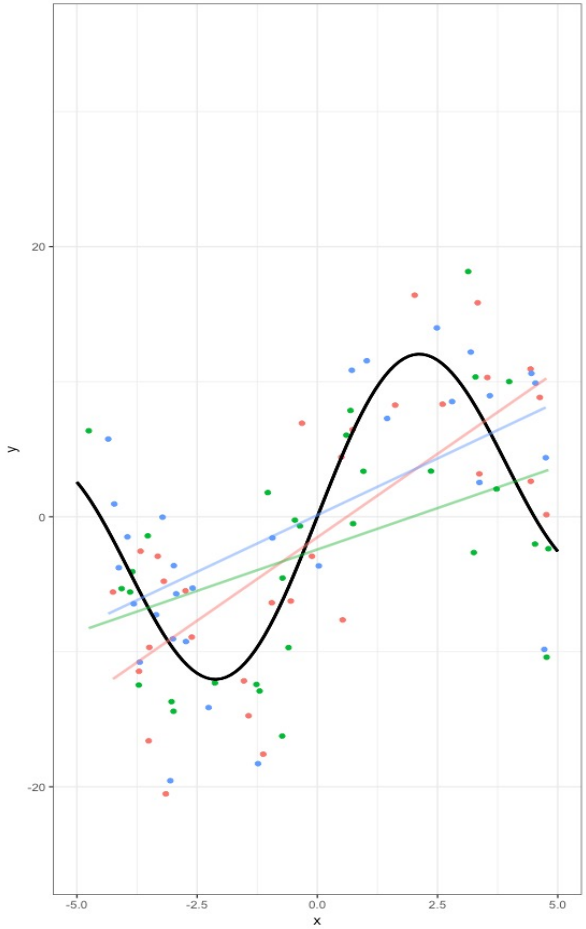
2. Evaluation Prädiktiver Modelle mit Resampling

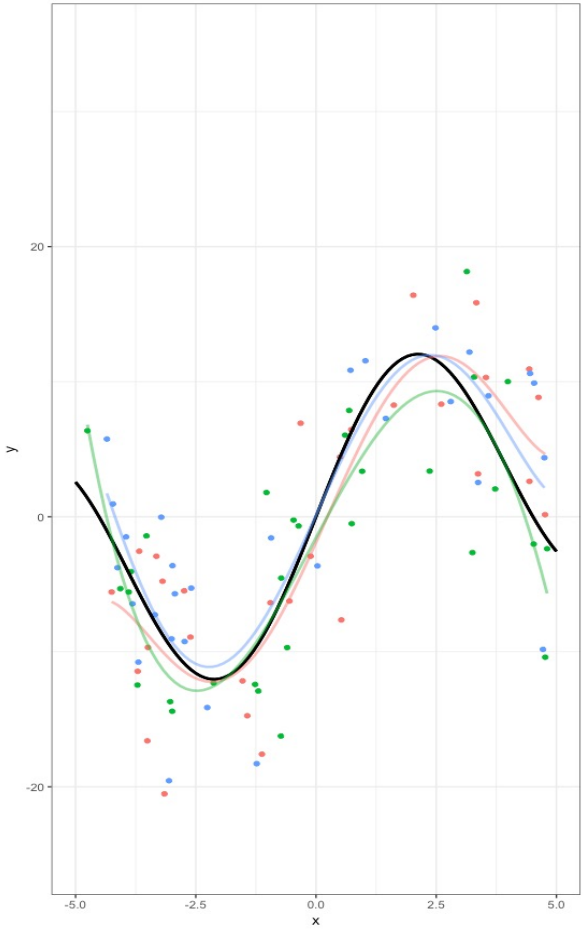
Agenda für heute

- Overfitting
- Bias-Varianz Tradeoff (Wdh.)
- Resampling
  - Grundprinzipien
  - Holdout
  - Negatives R^2
  - Kreuzvalidierung

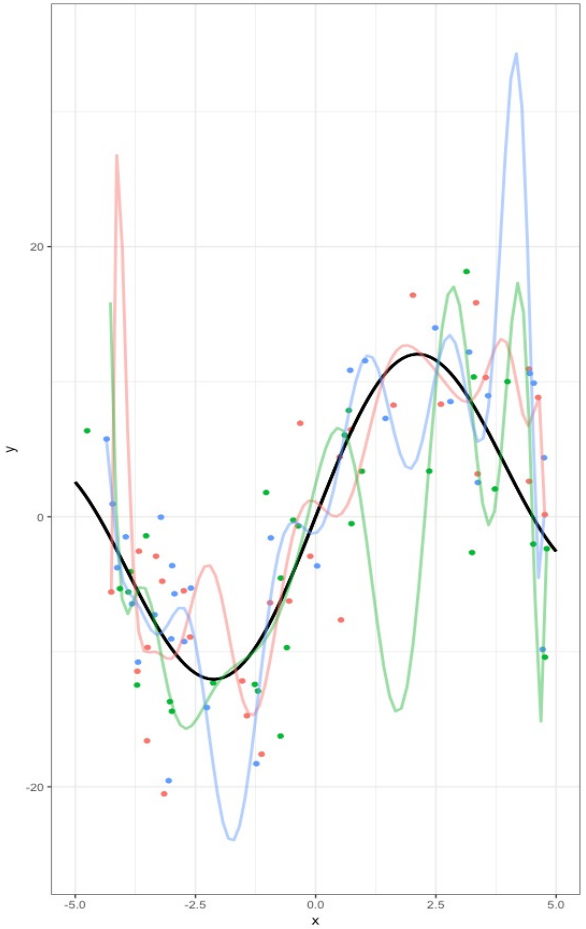
Overfitting



zu einfach



angemessene Komplexität



zu komplex

Schwarze Linie = wahrer Zusammenhang

Overfitting

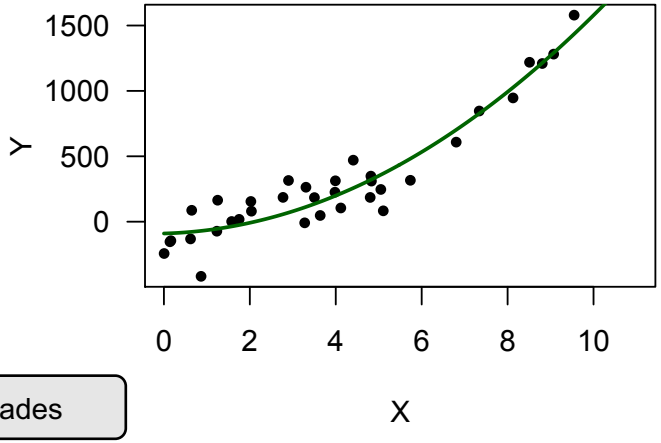
- **Beobachtung:** Manche trainierten prädiktiven Modelle beschreiben eine konkrete Stichprobe gut
  - d.h. gute Vorhersagen für diejenigen Beobachtungen, die auch zur Schätzung der Modellparameter verwendet wurden („in-sample“) …
  - … aber ohne den wahren Zusammenhang in der Population richtig abzubilden
  - → schlechte Vorhersagen für *neue* Daten („out-of-sample“).
- Ist dieses Phänomen besonders ausgeprägt, spricht man von „Overfitting“
  - Der zufällige, nicht reduzierbare Fehler (das „Rauschen“) wird fälschlicherweise mitmodelliert
  - Das Modell passt sich den Daten zu stark an, es „lernt auswendig“
  - Das prädiktive Modell „halluziniert“ Muster in den Daten, die in der Realität gar nicht existieren
- Overfitting ist schlimmer ...
  - je größer der nicht reduzierbare Fehler
  - je kleiner die Stichprobe im Training
  - je flexibler die Modellklasse

Overfitting

Trainingsset

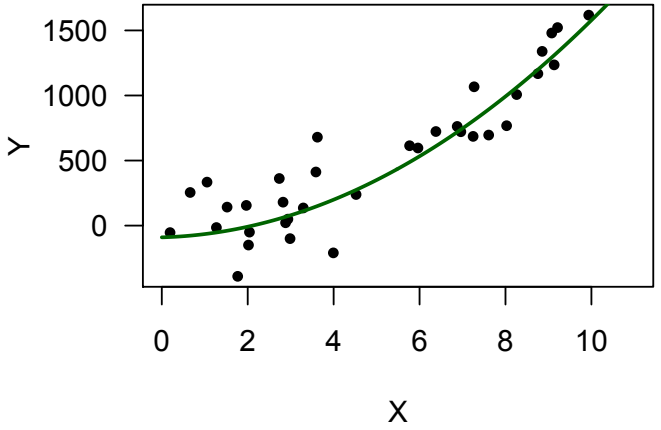
R^2 = 0.91 ; AICc = 452.63

$$y_i = b_0 + b_1 x_i + b_2 x_i^2 + \epsilon_i$$



Testset

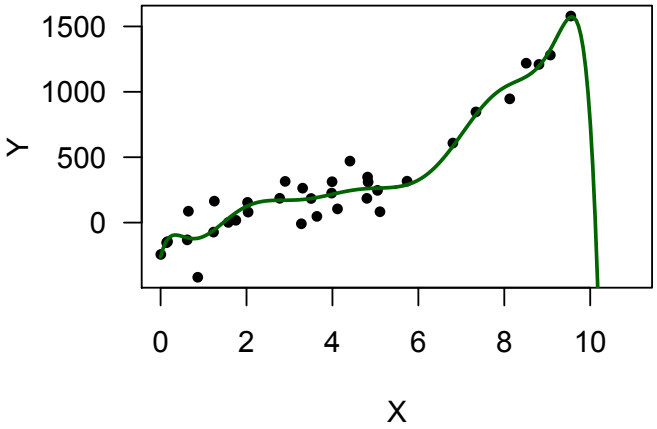
R^2 = 0.85 ; AICc = 480.87



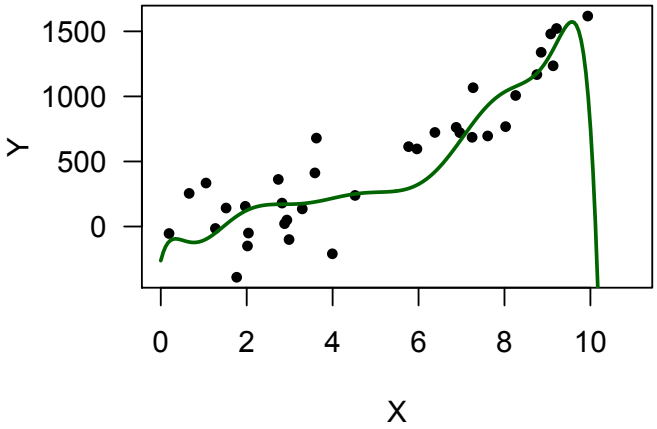
Polynom 10. Grades

R^2 = 0.94 ; AICc = 466.34

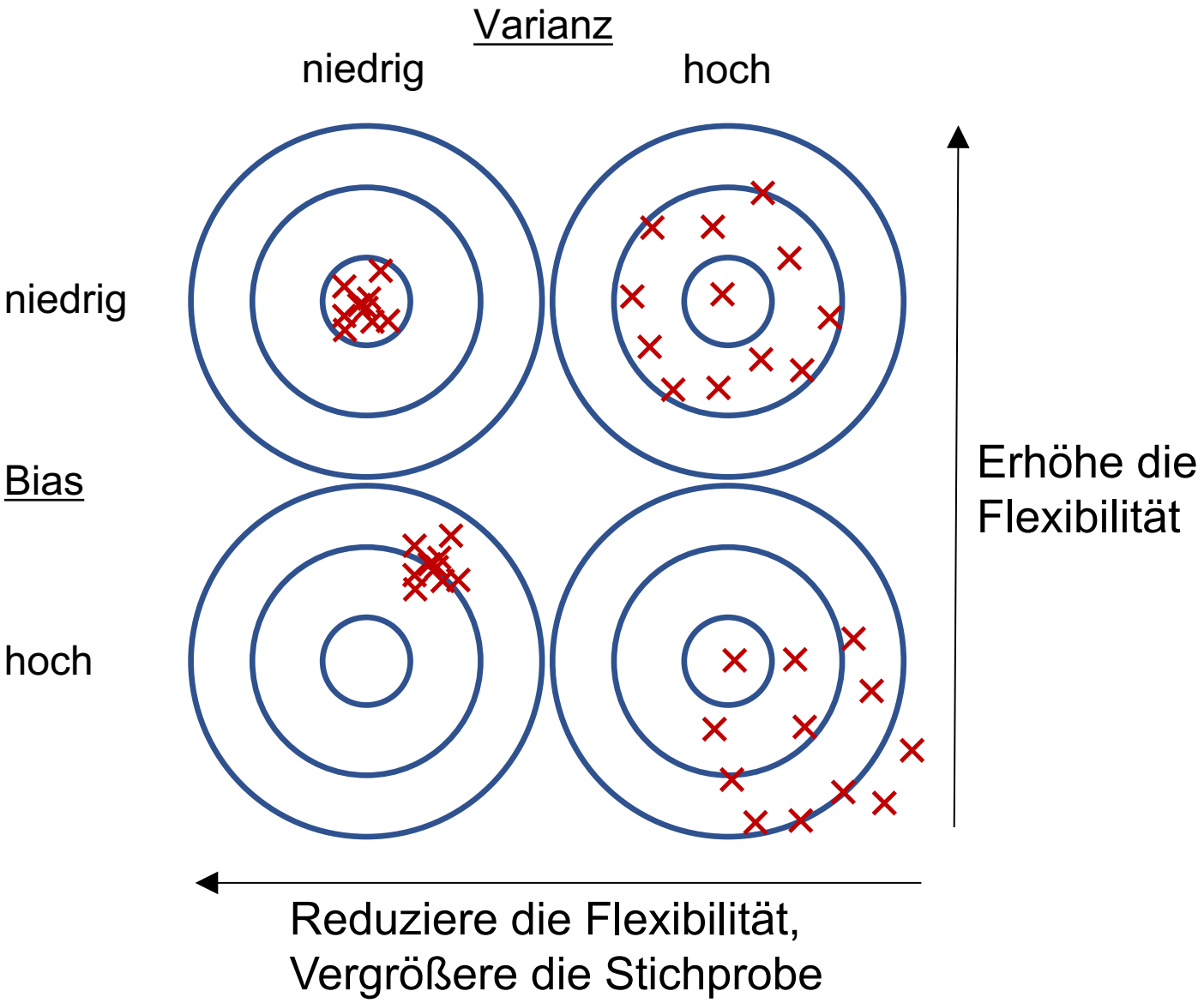
$$y_i = b_0 + b_1 x_i + b_2 x_i^2 + b_3 x_i^3 + b_4 x_i^4 + b_5 x_i^5 + ... + b_{10} x_i^{10} + \epsilon_i$$



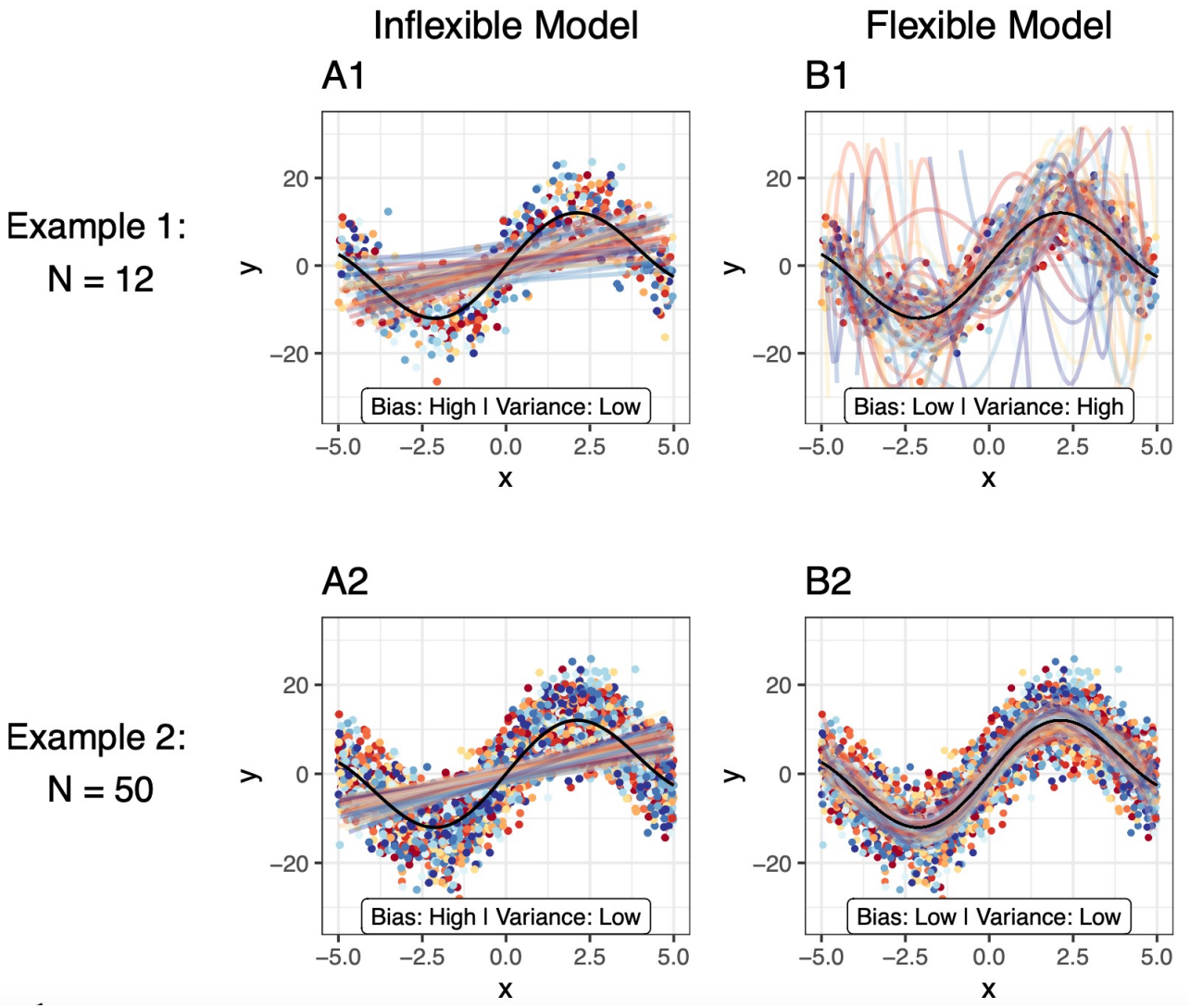
R^2 = 0.78 ; AICc = 524.43



Wiederholung: Bias & Varianz



Exkurs: Bias – Varianz Tradeoff und Overfitting (Pargent, Schödel, & Stachl, 2022)



Bei B1 sehen wir das Overfitting

Overfitting Fazit

- Insbesondere bei komplexen (=flexiblen) Modellen und/oder kleinen Stichproben fangen die Modelle an, das Rauschen in dieser konkreten Stichprobe abzubilden.
  - Diese zufälligen Schwankungen werden sich in der nächsten Stichprobe nicht mehr finden – uns interessiert nur das systematische Signal, das in allen Stichproben vorhanden ist.
- Wir können Performancemaßen nicht trauen, die auf dem Trainingsdatensatz berechnet wurden („in-sample): Sie werden, gerade bei komplexen Modellen, stark überschätzt sein.
  - Hinweis: Bei der „klassischen“ Statistik machen wir aber genau das – z.B. das R^2 für Regressionsmodell *im Trainingsdatensatz* berechnen.
- Sobald ich das einmal trainierte Prädiktionsmodell auf *neue* Daten anwende, wird die Vorhersageperformance schlechter sein.
- Wie kommt man nun zu einer Abschätzung des Vorhersagefehlers für *neue* Daten?

Evaluation Prädiktiver Modelle

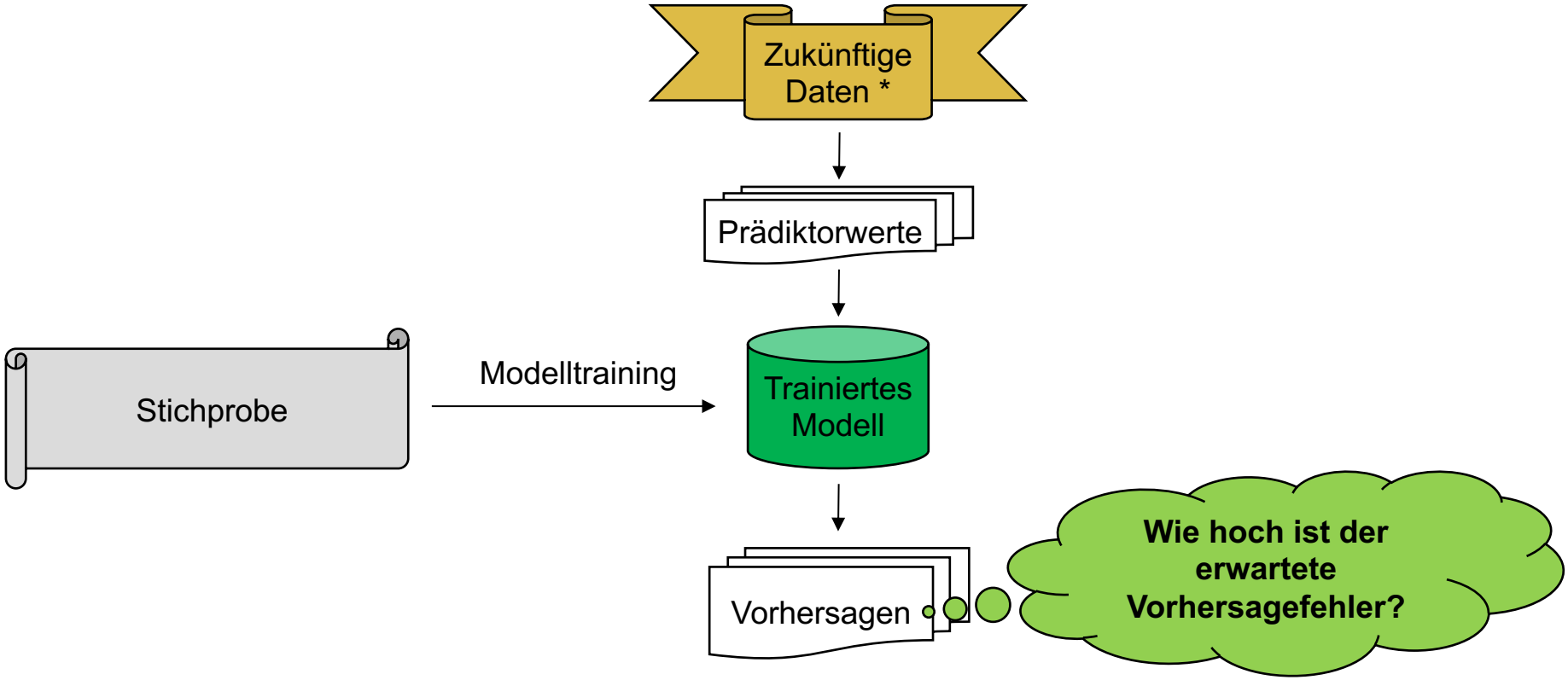
- Wie genau sind die Vorhersagen des Modells bei *neuen* Beobachtungen?
  → **Ziel:** Schätzung des erwarteten Vorhersagefehlers
  - **Utopisches Vorgehen:** Erhebe zusätzliche Beobachtungen aus der Population (Prädiktorwerte und Kriteriumswerte) und berechne ein geeignetes Performancemaß für die Vorhersagen der neuen Daten
    → In der Praxis meist nicht durchführbar
  - **Naives Vorgehen:** Schätze den erwarteten Vorhersagefehler durch Berechnung des Performancemaßes in *derselben* Stichprobe, die bereits zur Schätzung des Modells verwendet wurde
    → Unterschätzung des erwarteten Vorhersagefehlers
  - **Empfohlenes Vorgehen:** Schlaues „Recyclen“ der vorliegenden Stichprobe für eine realistische Abschätzung des Vorhersagefehlers
    → Resampling Methoden

Trainings- und Testset

- Zur Modellevaluation teilt man den Gesamtdatensatz zufällig in zwei Teile
  - Trainingsset: Wird zum Trainieren des Modells verwendet
  - Testset: Wird zur Abschätzung der Vorhersagegüte verwendet
- Realistische Abschätzung des erwarteten Vorhersagefehlers des gesamten Modells durch die Berechnung der Vorhersagegüte im Testset
  - Schätze Modellparameter anhand der Beobachtungen im Trainingsset
  - Verwende das trainierte Modell zur Berechnung von Vorhersagen für die Beobachtungen im Testset
  - Vergleiche die Vorhersagen im Testset mit den tatsächlichen Kriteriumswerten anhand eines geeigneten Performancemaßes

- **Hinweis:** Im Folgenden wird von einer Regressionsfragestellung mit dem MSE als Performancemaß ausgegangen. Klassifikation oder Verwendung anderer Performancemaße funktioniert analog

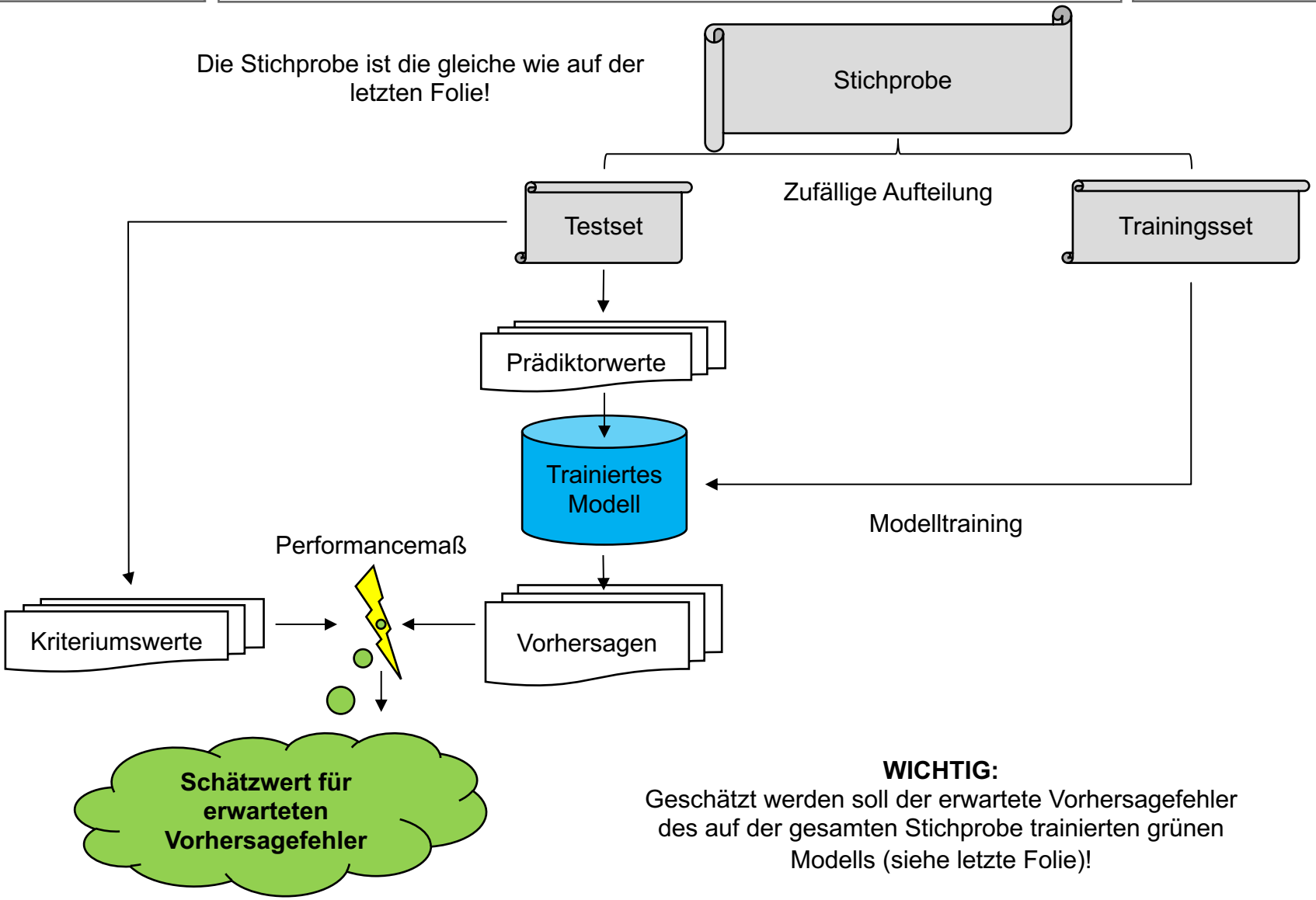
Modellevaluation



* Für zukünftige Daten aus der praktischen Anwendung liegen nur die Prädiktorwerte vor!

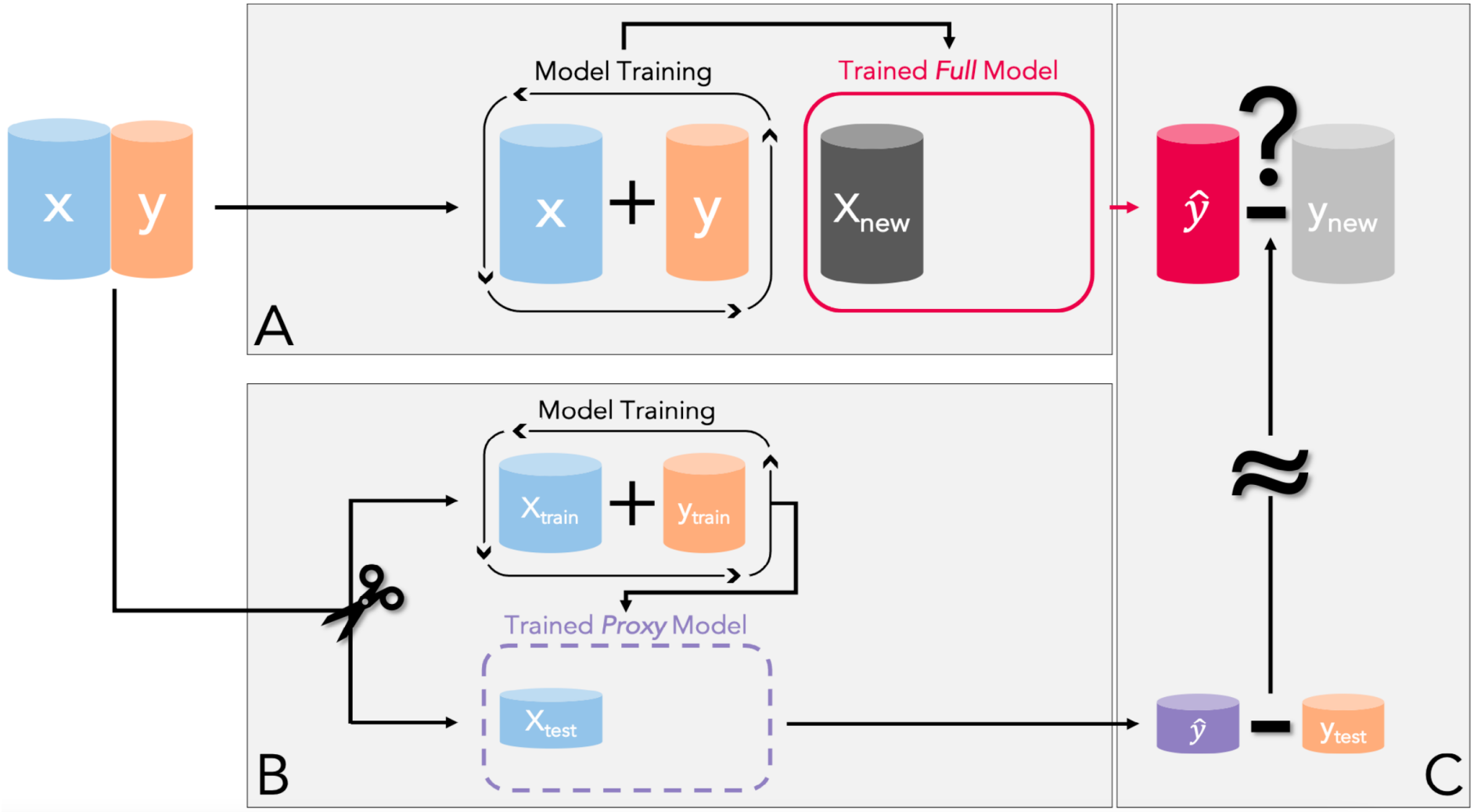
Veranschaulichung Trainings- und Testset

Die Stichprobe ist die gleiche wie auf der letzten Folie!



WICHTIG:
Geschätzt werden soll der erwartete Vorhersagefehler des auf der gesamten Stichprobe trainierten grünen Modells (siehe letzte Folie)!

Exkurs: Nochmal anders veranschaulicht (Pargent, Schödel, & Stachl, 2022)



Training und Testperformance

$$MSE_{train} = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} \left(y_{i,train} - \hat{y}_{i,train}^{(train)} \right)^2$$

$$MSE_{test} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \left(y_{i,test} - \hat{y}_{i,test}^{(train)} \right)^2$$

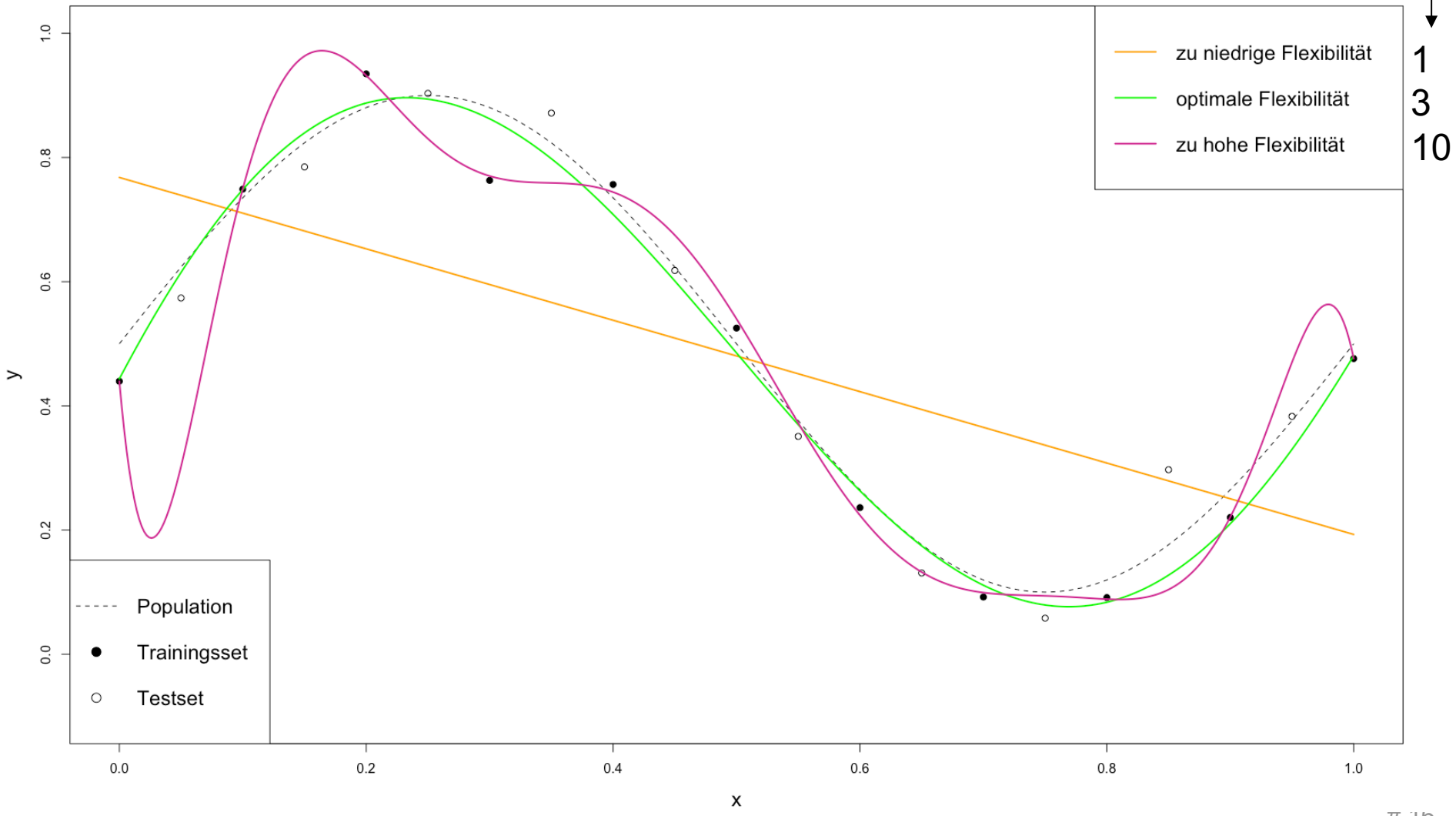
- $y_{i,train}$: Kriteriumswert von Person i aus dem Trainingsset
- $y_{i,test}$: Kriteriumswert von Person i aus dem Testset
- $\hat{y}_{i,train}^{(train)}$: Vorhergesagter Wert für Person i aus dem Trainingsset, mithilfe des prädiktiven Modells, welches am Trainingsset geschätzt wurde
- $\hat{y}_{i,test}^{(train)}$: Vorhergesagter Wert für Person i aus dem Testset, mithilfe des prädiktiven Modells, welches am Trainingsset geschätzt wurde
- $N = N_{train} + N_{test}$

Holdout - Schätzer

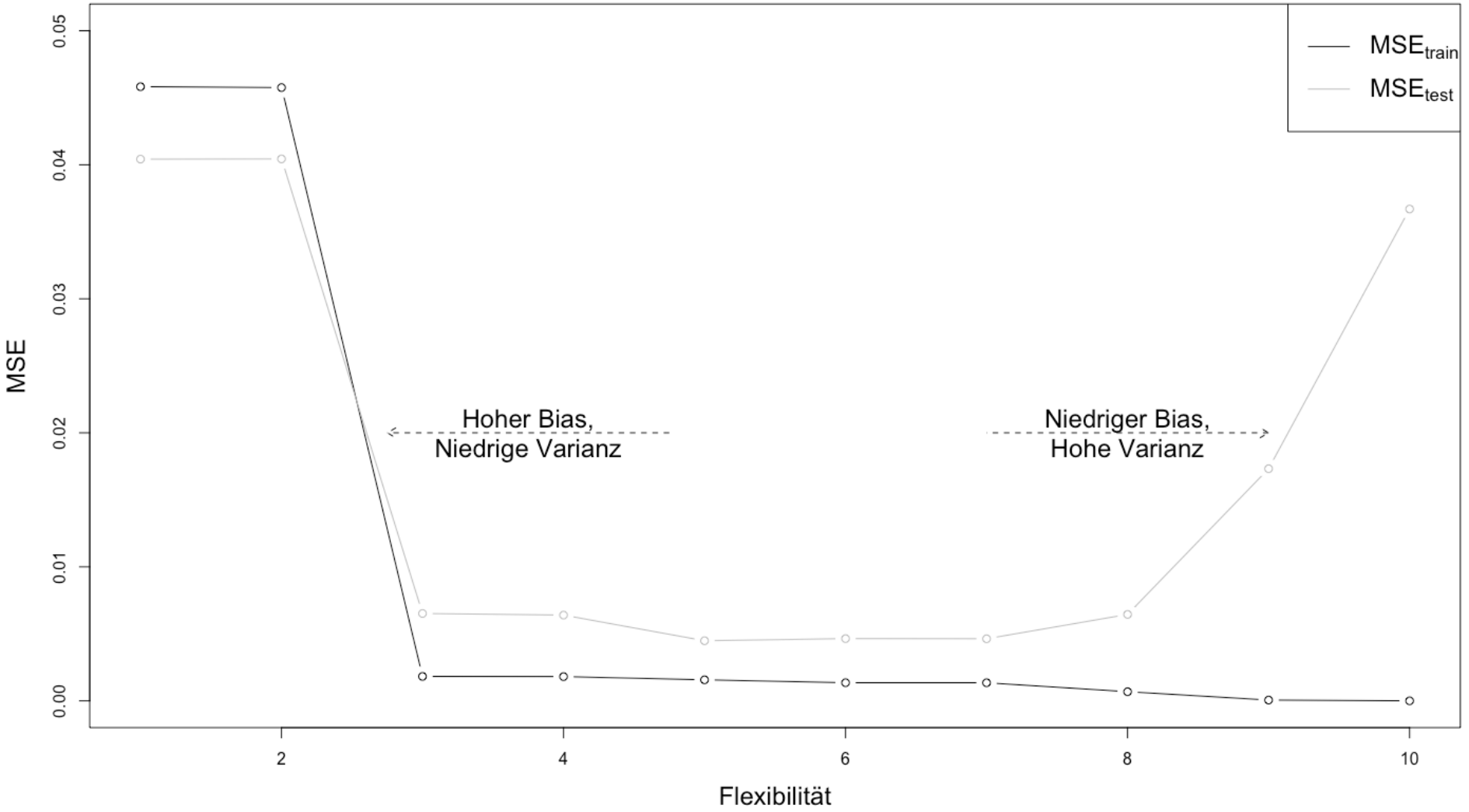
- Im oben skizzierten Vorgehen mit genau einem Trainingsset und einem Testset wird MSE_{test} auch als „Holdout-Schätzer“ bezeichnet
- Die Bezeichnung Holdout-Schätzer ist unabhängig vom verwendeten Performancemaß, gilt also auch für z.B. R^2_{test}, $MMCE_{test}$, $SPEC_{test}$, ...
- Der Holdout-Schätzer ist die einfachste Resampling Prozedur, die eine vernünftige Abschätzung des erwarteten Vorhersagefehlers erlaubt

Bias – Varianz Tradeoff am Testset

„Flexibilität“ = Grad des Polynoms



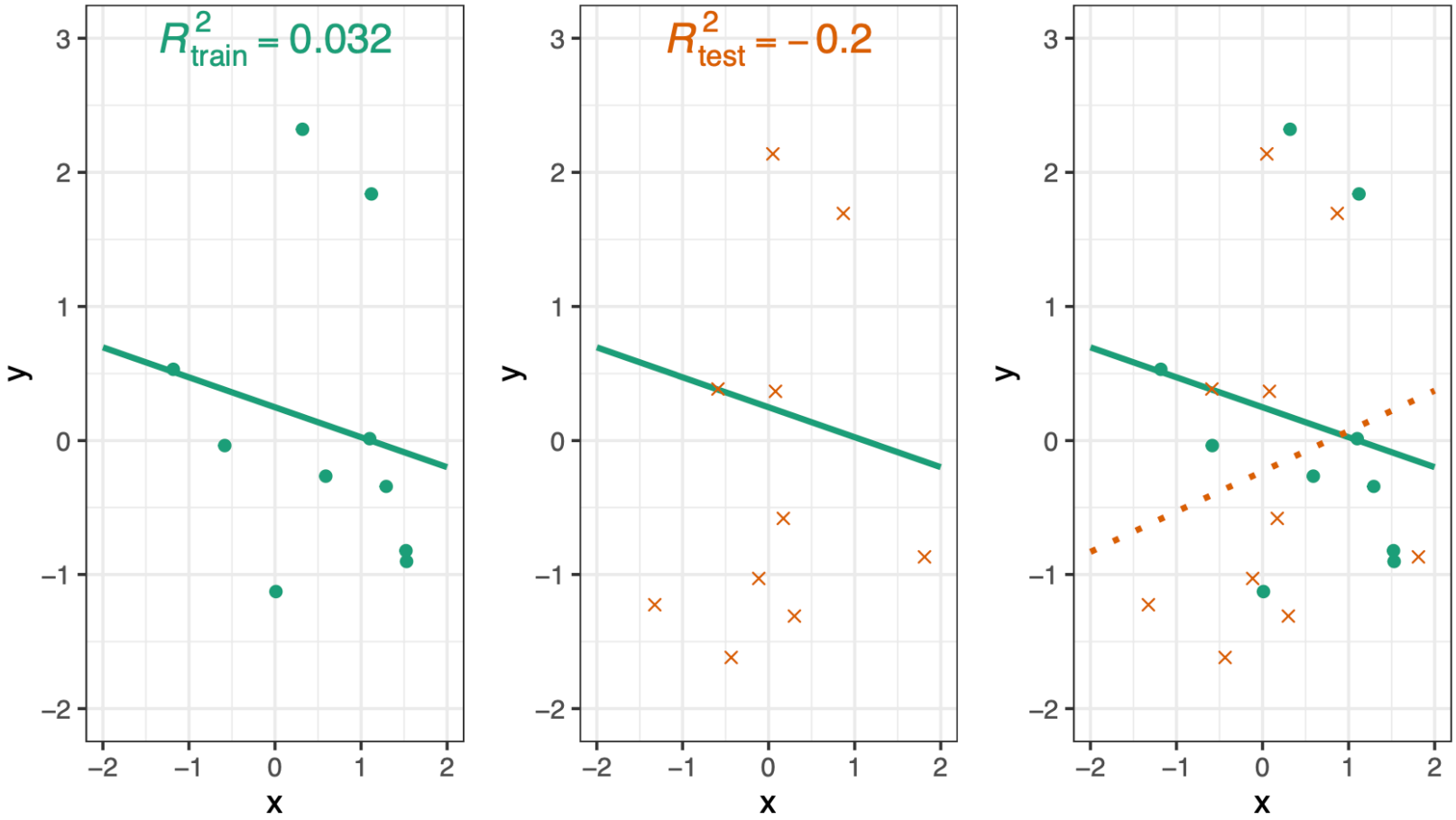
Bias – Varianz Tradeoff am Testset



Negatives R^2

- Der Wertebereich von R^2 liegt allgemein im Intervall $]-\infty, 1]$
- Nur für R^2_{train} in der multiplen linearen Regression gilt: $R^2_{train} \in [0,1]$
  - durch die Methode der kleinsten Quadrate ist die schlechtmöglichste Vorhersage der multiplen linearen Regression innerhalb des Trainingssets die konstante Vorhersage des Mittelwerts $\bar{y}$ (dann gilt: $R^2_{train} = 0$)
- Interpretation von $R^2_{test} < 0$:
  - Das prädiktive Modell trifft schlechtere Vorhersagen als ein Modell, welches für jede Beobachtung immer den Mittelwert im Kriterium $\bar{y}$ vorhersagt, d.h. die Werte auf den Prädiktoren komplett ignoriert
  - Ein nützliches Modell muss immer das triviale Modell mit konstanter Vorhersage schlagen, d.h. ein positives R^2 haben
  - Ein negatives R^2_{test} zeigt häufig Overfitting an
- **Fazit:** Bei der Anwendung prädiktiver Modelle in der Praxis ist es möglich, schlechtere Vorhersagen zu erzielen als durch „schlaues Raten“

Negatives R^2

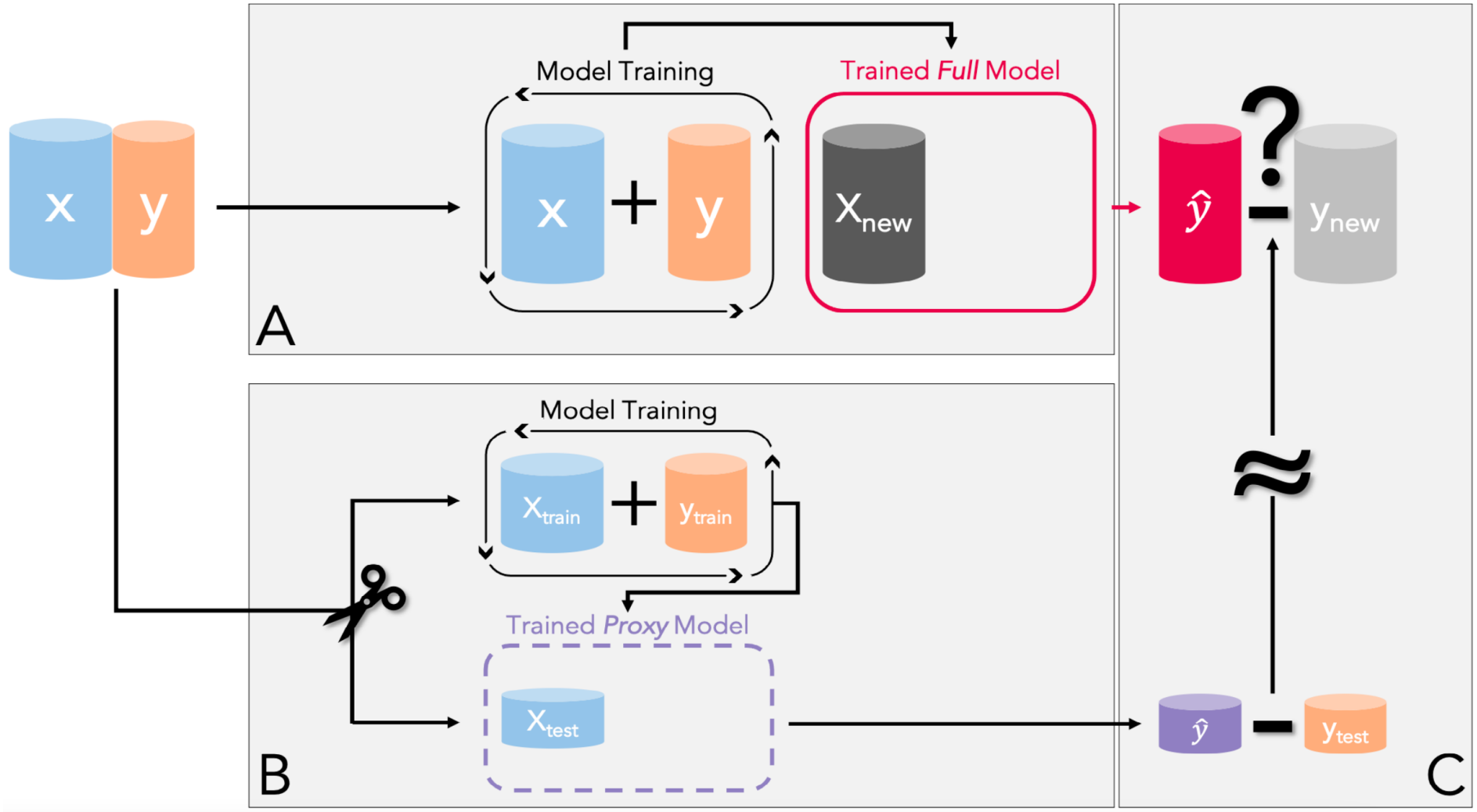


- Trainiere Modell anhand der Trainingsdaten (grün)
- Positives R^2_{train} für die Trainingsdaten

- Vorhersage der (orange) Testdaten mithilfe des trainierten Modells (grün)
- Negatives R^2_{test}

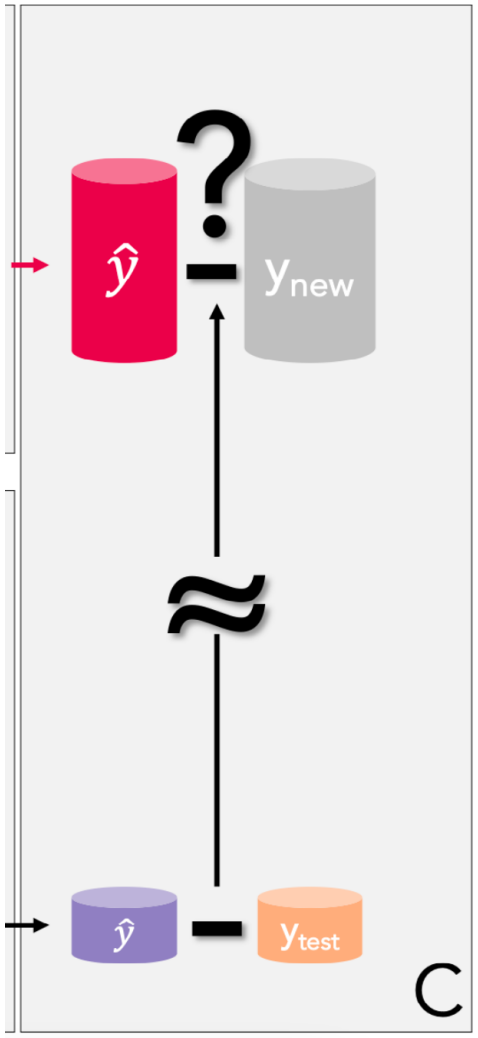
- Zusammenhang im Testset ist positiv, Zusammenhang im Trainingsset ist negativ

Wdh: Performance im Testset als Schätzer für Gesamtperformance



Bias und Varianz des Holdout-Schätzers

- Der Holdout-Schätzer stellt eine Schätzfunktion für den erwarteten Vorhersagefehler des auf der **gesamten** Stichprobe trainierten Modells dar.
- Ist das ein guter Schätzwert? Die Güte dieser Schätzfunktion hängt ab von deren ...
  - Bias: Entspricht der Mittelwert des Holdout-Schätzers bei wiederholten Stichproben dem erwarteten Vorhersagefehler?
  - Varianz: Wie stark unterscheiden sich die Schätzwerte des Holdout-Schätzers zwischen wiederholten Stichproben?
- Bias und Varianz hängen stark vom gewählten Größenverhältnis von Trainings- und Testset ab.



Bias und Varianz des Holdout-Schätzers

- Generell gilt: Je mehr Daten zur Schätzung eines prädiktiven Modells verwendet werden, desto präziser sind dessen Vorhersagen
- Da das Trainingsset immer kleiner ist als der Gesamtdatensatz, wird die Genauigkeit des gesamten prädiktiven Modells tendenziell unterschätzt
  - Bias ist größer, je kleiner das Trainingsset

- Generell gilt: Je mehr Daten zur Berechnung der Performance verwendet werden, desto präziser wird der erwartete Vorhersagefehler geschätzt (siehe Standardfehler des Mittelwerts bei unabhängigen Beobachtungen)
  - Varianz ist größer, je kleiner das Testset

- Es gilt immer: Je größer das Trainingsset, desto kleiner das Testset
  → Es ist nicht möglich sowohl Bias als auch Varianz zu minimieren
  → Optimaler Tradeoff von Bias und Varianz nicht eindeutig zu bestimmen

Optimale Größe von Trainings- und Testset

- Wie groß soll das Trainingsset im Verhältnis zum Testset optimalerweise gewählt werden?

- Daumenregel zur Aufteilung in Trainings- und Testset:
  - Trainingsset: $\frac{2}{3}$ des Gesamtdatensatzes
  - Testset: $\frac{1}{3}$ des Gesamtdatensatzes

Verbesserung des Holdout Schätzers

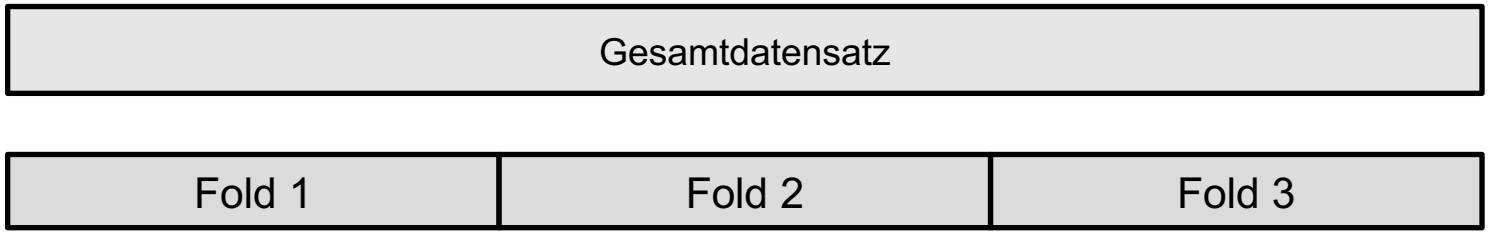
- Kann der Holdout-Schätzer durch eine schlauere Aufteilung in Trainings- und Testdaten noch verbessert werden?
- Beispiel:
  - Teile den Gesamtdatensatz auf in ein Trainings- und Testset mit gleicher Größe
  - Vergleiche den Holdout-Schätzer mit folgendem Vorgehen:
    - Berechne Holdout-Schätzer
    - Vertausche Trainings- und Testset
    - Berechne Holdout-Schätzer erneut
    - Berechne Mittelwert der beiden Holdout-Schätzer
      → 2-Fold Cross-Validation (2-Fold CV)
- Bei gleicher Größe des Trainingssets hat der 2-Fold CV Schätzer den gleichen Bias wie der Holdout-Schätzer, aber eine niedrigere Varianz
  → Grund: Mitteln über zwei Testsets der gleichen Größe wie bei Holdout

K-Fold Cross-Validation

- Zufällige Aufteilung des Datensatzes in *K* gleich große Teile („Folds“)
- Verwende jeden der *K* Teile einmal als Testset und die anderen *K*-1 Teile als das dazu gehörige Trainingsset
- Berechne MSE_{test} (oder anderes Maß) für jedes der *K* Testsets. Für die Vorhersagen verwende das Modell aus dem jeweiligen Trainingsset
- Schätze den erwarteten Vorhersagefehler durch den Mittelwert von MSE_{test} über alle *K* Testsets

- Ähnlich wie beim Holdout-Schätzer ist die optimale Anzahl an Folds nicht eindeutig zu bestimmen
  - In der Praxis werden meist 5 oder 10 Folds verwendet
  - Es gilt **nicht**: Je mehr Folds desto besser (weil bei vielen Folds ja wiederum das Testset immer kleiner wird)
- Weitere Resampling Methoden:
  Repeated Cross-Validation, Bootstrap, Subsampling

Beispiel 3-Fold Cross-Validation



Zufällige Aufteilung

$MSE_{test}^{(1)}$

$MSE_{test}^{(2)}$

$MSE_{test}^{(3)}$

$$MSE_{CV} = \frac{1}{3}\left(MSE_{test}^{(1)} + MSE_{test}^{(2)} + MSE_{test}^{(3)}\right)$$

Schätzfunktion K-Fold Cross-Validation

$$MSE_{test}^{(k)} = \frac{1}{N_k}\sum_{i=1}^{N_k}\left(y_{i,k} - \hat{y}_{i,k}^{(train_k)}\right)^2$$

MSE aus dem Fold *k*

$$MSE_{CV} = \frac{1}{K}\sum_{k=1}^{K} MSE_{test}^{(k)}$$

MSE über alle *K* folds gemittelt

- K: Anzahl der Folds
- N_k: Anzahl von Beobachtungen in Fold k
- $y_{i,k}$: Kriteriumswert der i – ten Person in Fold k
- $\hat{y}_{i,k}^{(train_k)}$: Vorhergesagter Wert für die i – te Person aus Fold k mithilfe des prädiktiven Modells, welches anhand des zu Fold k gehörenden Trainingssets (besteht aus allen Folds außer Fold k) geschätzt wurde
- $N = N_1 + \cdots + N_K$

Fazit zur Stichprobengröße

- Der beste Weg die Vorhersagegüte prädiktiver Modelle in einem konkreten Anwendungsfall zu verbessern, besteht in der Erhöhung der zur Verfügung stehenden Stichprobe.

- **Gründe:** Je größer die Stichprobe, desto...
  - geringer die Varianz der Vorhersagen
  - größer das Potential flexiblerer Modellklassen mit niedrigem Bias
  - geringer die Gefahr von Overfitting
  - mehr Prädiktoren können sinnvoll genutzt werden

- Außerdem gilt: Je größer die Stichprobe, desto genauer kann für ein trainiertes prädiktives Modell der erwartete Vorhersagefehler abgeschätzt werden (weil man dann auch das Testset entsprechend größer machen kann).

Nächste Woche

- Kennenlernen eines prädiktiven Standardmodells bei Regressions- und Klassifikationsfragestellungen, das mit sehr vielen Prädiktorvariablen umgehen kann und trotzdem noch ähnlich gut interpretierbar ist wie multiple lineare und logistische Regressionsmodelle.

→ Regularisierte lineare und logistische Regression (LASSO)