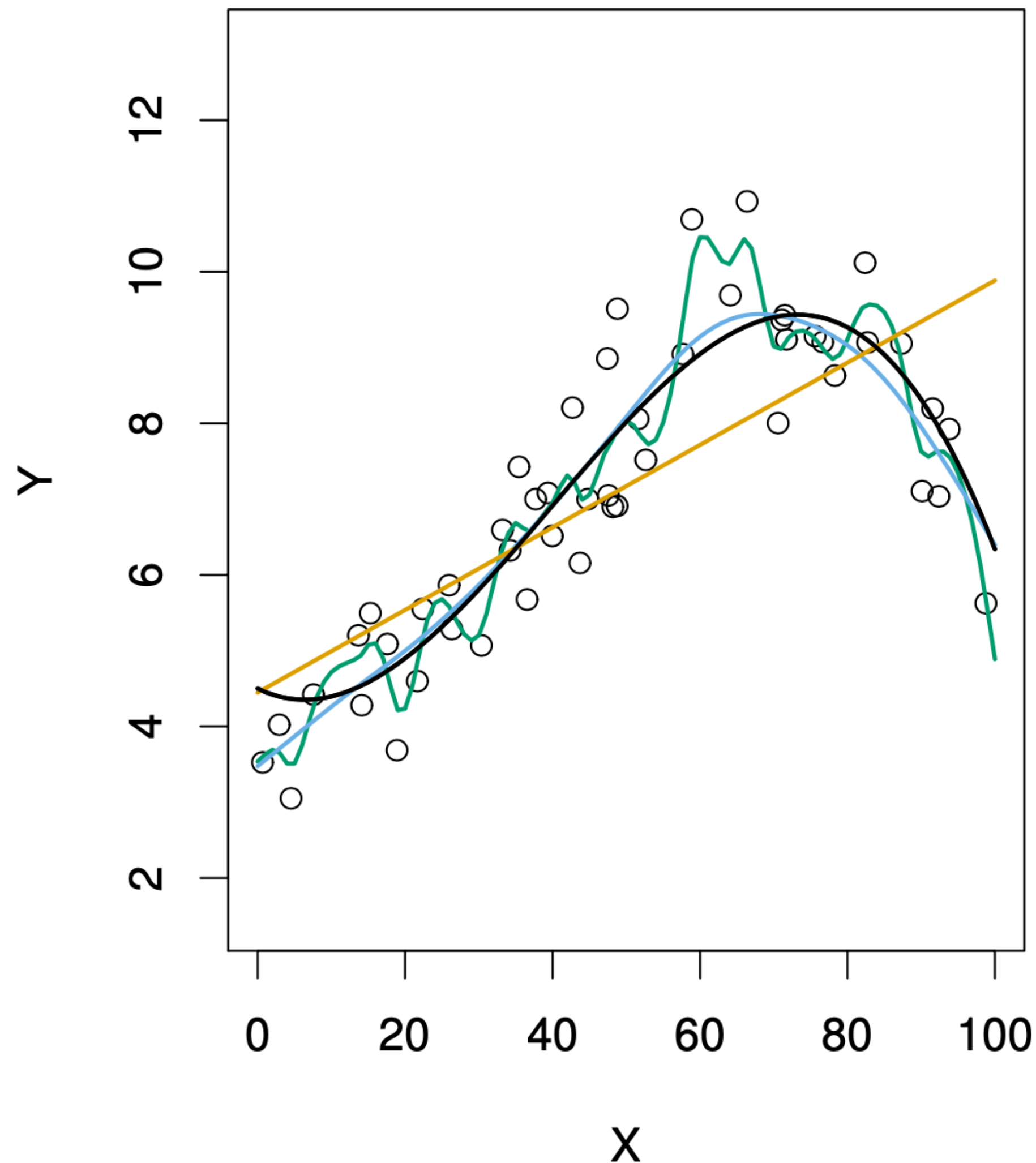


Podstawy uczenia maszynowego

Część 2

Dr Mateusz Radzimski

Bias & Variance

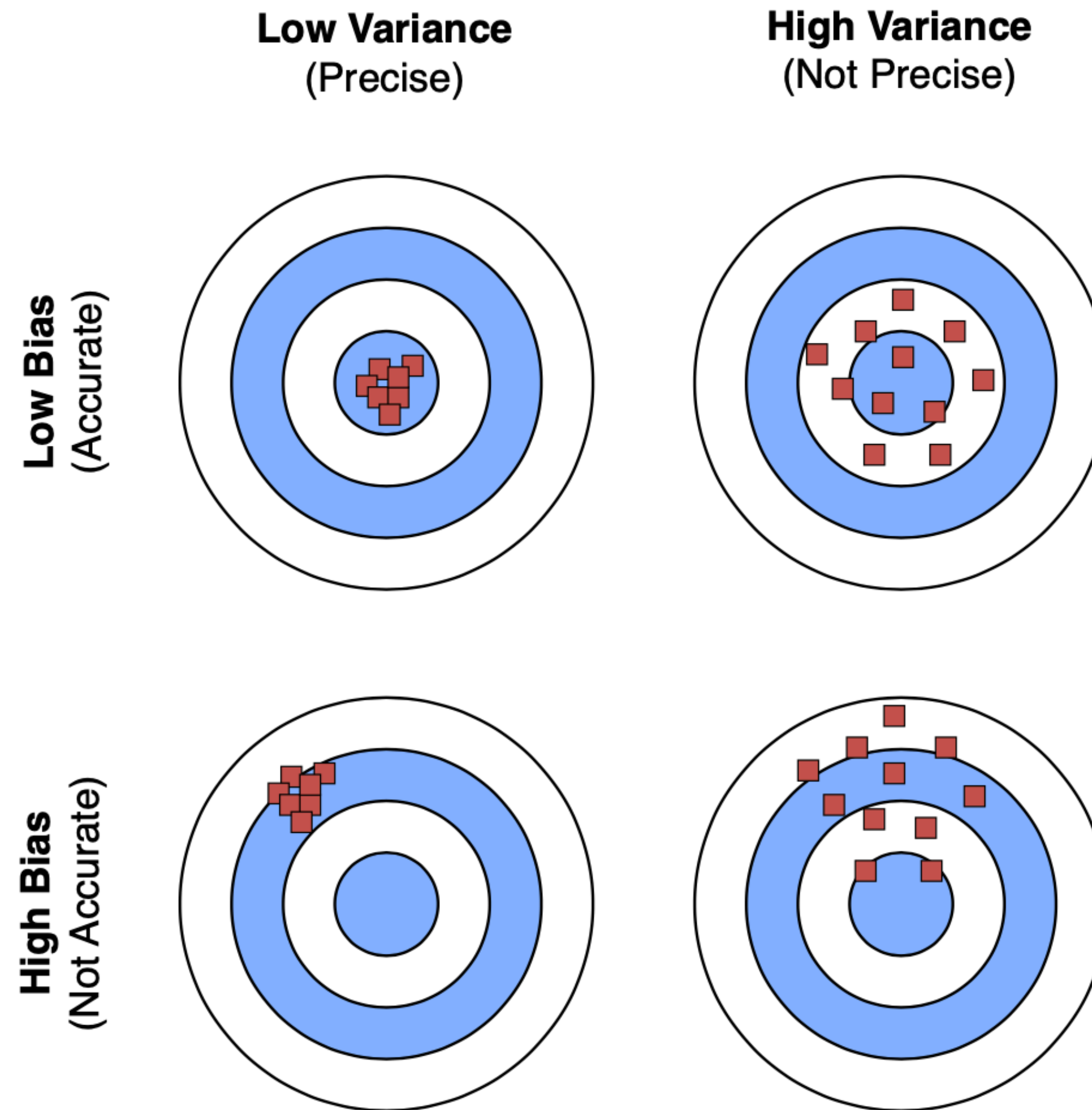


- Terminy te określają jak dobrze dopasowaliśmy model do naszego problemu
- Dla przypomnienia: w regresji liniowej używaliśmy metryki MSE i chcieliśmy aby była najmniejsza
- Można pokazać, że

$$E \left(y_0 - \hat{f}(x_0) \right)^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon).$$

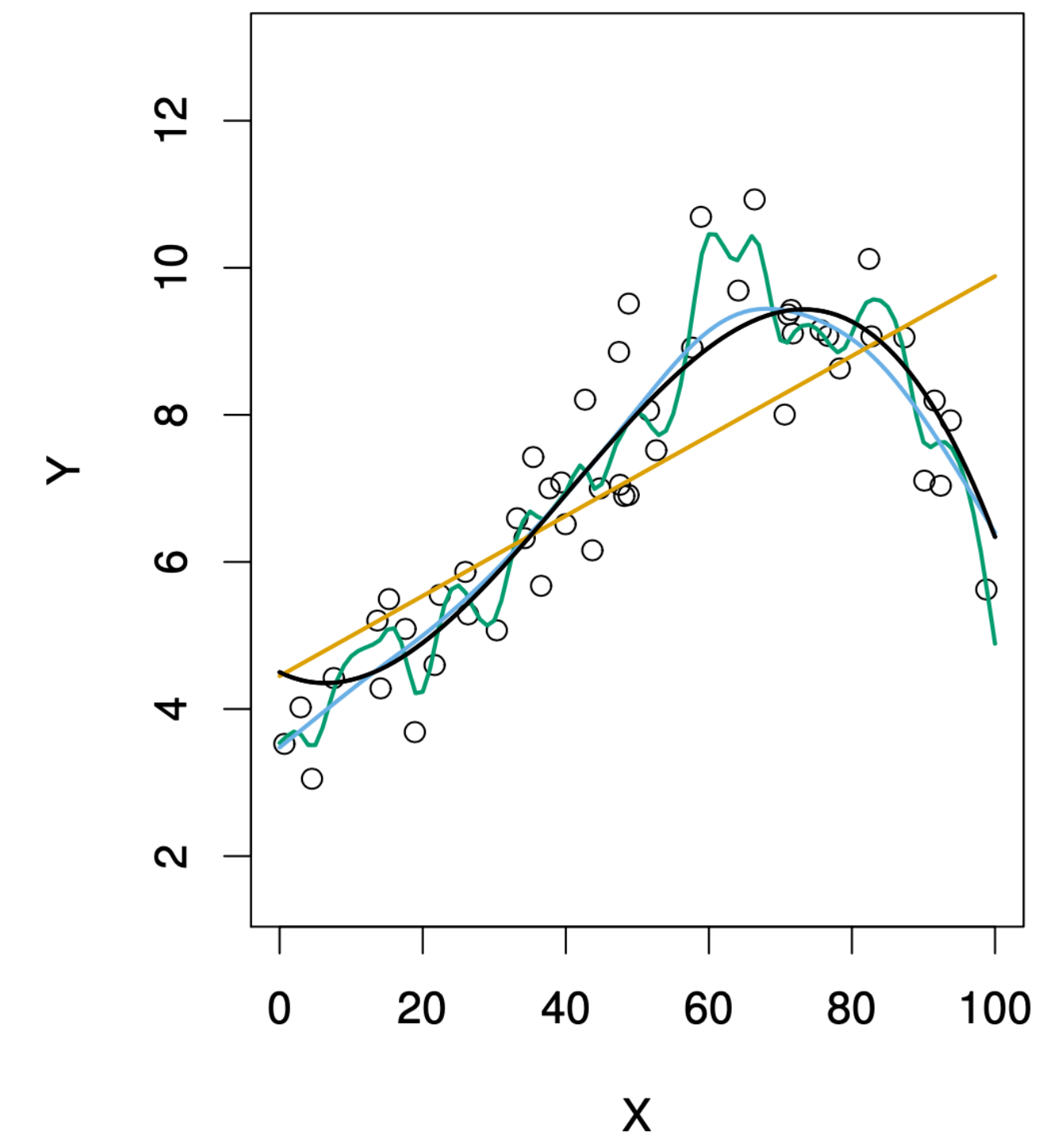
- Słownie: oczekiwane MSE zbioru testowego = Wariancja + Obciążenie + wariancja błędu nieoznaczonego (nieredukowalnego)

Obciążenie & Wariancja (Bias & Variance)



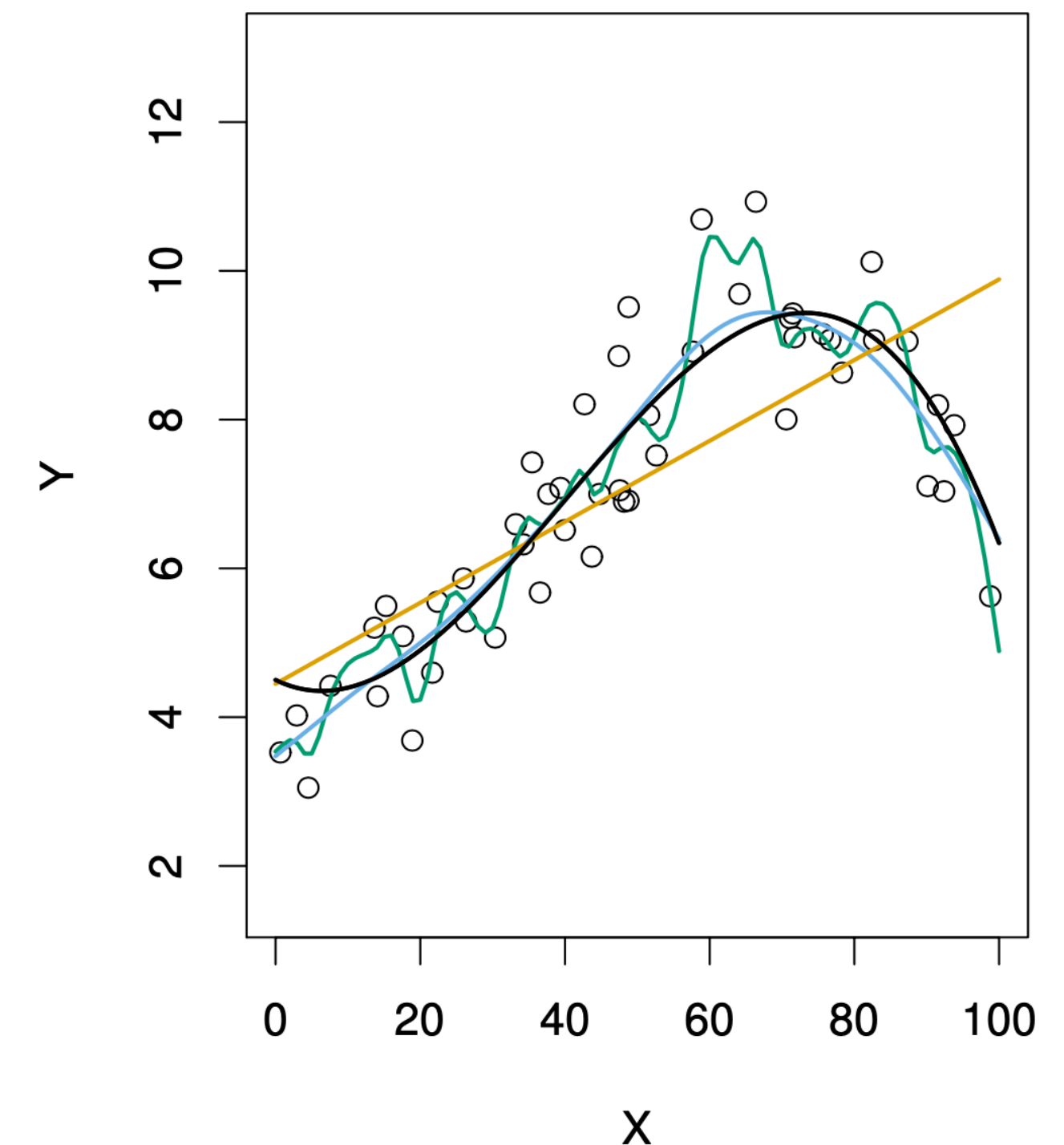
Bias & Variance - intuicja

- Wariancja — $\text{Var}(\hat{f})$ — jak zmieni się nasza funkcja \hat{f} gdy wyestymujemy ją innymi danymi testowymi
 - W idealnym przypadku inny zbiór obserwacji tego samego zjawiska nie powinien zbytnio wpłynąć na \hat{f}
 - Duża wariancja powoduje, że inny zbiór obserwacji spowoduje duże zmiany estymowanej funkcji \hat{f}
 - Przebieg zielonej krzywej obrazuje \hat{f} o dużej wariancji
 - Przebieg pomarańczowej prostej za dużo się nie zmieni gdy zmienimy kilka obserwacji - ta funkcja ma małą wariancję
- Błąd wariancji to uczenie się nieistotnych różnic, wrażliwość na lekkie odchylenia

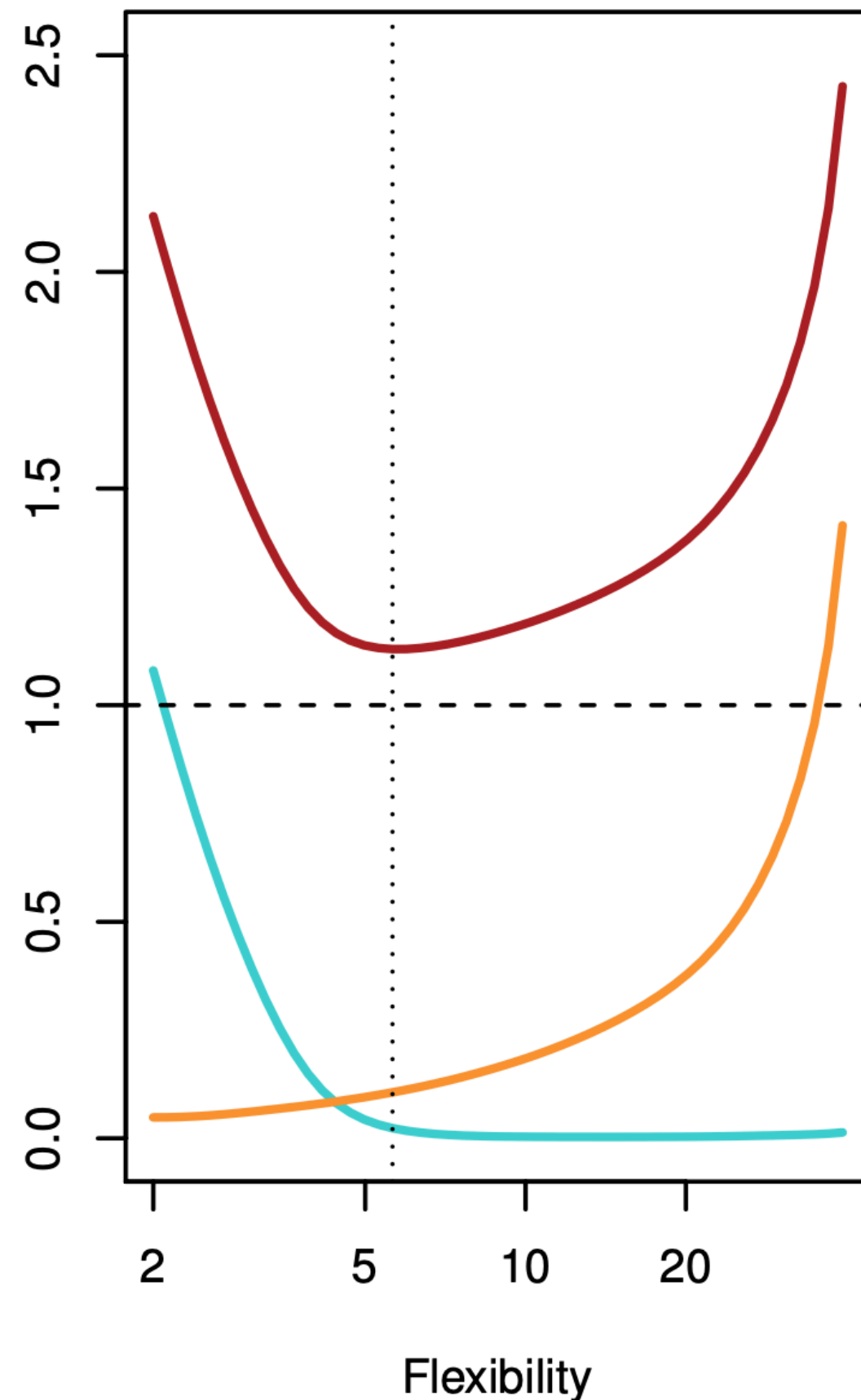


Bias & Variance - intuicja

- Obciążenie — $\text{Bias}(\hat{f})$ — błąd związany z pewnym przybliżeniem rzeczywistego zjawiska przez uproszczony model
 - Np. regresja liniowa zakłada liniową zależność między Y a $X_1, X_2, X_3 \dots, X_n$
 - W praktyce wiele zjawisk ma bardziej komplikowane zależności
 - W przykładzie na obrazku regresja liniowa nie jest dobrym dopasowaniem do zbioru uczącego
 - Funkcja \hat{f} reprezentowana przez pomarańczową prostą będzie miała wysokie obciążenie
- Błąd obciążenia to systematyczne uczenie się błędnych rzeczy, wynikające ze złych założeń/złego doboru modelu/itd.



Bias & Variance - kompromis



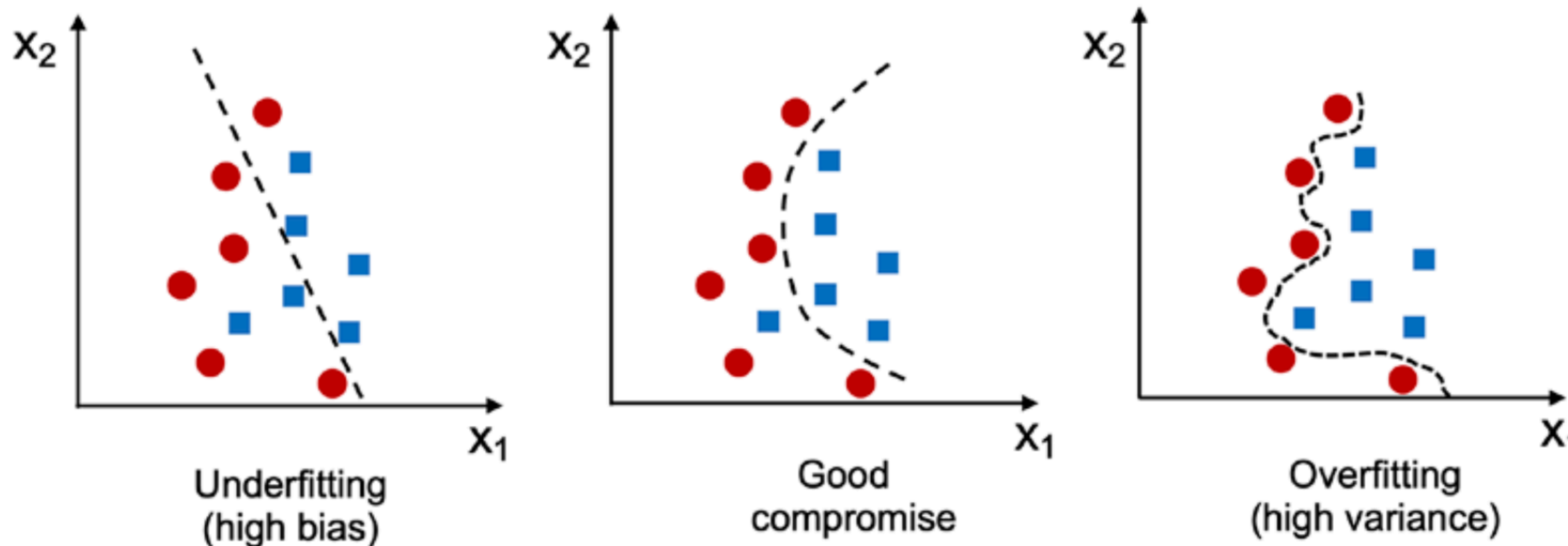
← Prostszy model
Zbyt małe dopasowanie
(Underfitting)

→ Bardziej złożony model
Nadmierne dopasowanie
(Overfitting)

- Niebieska krzywa - bias
- Pomarańczowa krzywa - wariancja
- Przerywana prosta - błąd nieoznaczony
- Czerwona prosta - oczekiwany MSE dla danych testowych
- Celem jest takie dopasowanie modelu, aby błąd był najmniejszy (minimum czerwonej prostej) - co jest kompromisem między obciążeniem a wariancją.

Underfitting / overfitting

- Kompromis między obciążeniem i wariancją wprost przekłada się na dwa istotne zjawiska:
- Problem nadmiernego dopasowania / niedotrenowania / underfitting
- Problem zbyt małego dopasowania / przetrenowania / overfitting



Sposoby na overfitting

- Regularyzacja (o tym w następnym semestrze)
- Uproszczenie modelu
- Przerwanie trenowania zanim algorytm zacznie dopasowywać model do szumu
- Cross validation - nie zapobieganie, ale pokaże że występuje

Wszystkie te metody będą ważne gdy poznamy nowe algorytmy uczenia maszynowego (szczególnie sieci neuronowe).

Zasady zaliczenia

- Zaliczeniem będzie projekt robiony w grupie 3-4 osobowej (max 5).
- Projekt polega na zastosowaniu regresji liniowej lub klasyfikacji (regresji logistycznej) na wybranym zbiorze danych
 - Zaproponuje 2 zbiory danych do wyboru: jeden na klasyfikację, drugi na regresję
 - Każdy może też przyjść ze swoim zbiorem danych (**wtedy zalecam konsultacje ze mną, żeby określić zakres**)
- Rezultatem projektu powinien być notebook (collab), który może zawierać elementy omawiane na wykładzie:
 - Wczytanie danych, może jakaś wizualizacja zbioru
 - Zastosowanie algorytmu uczenia maszynowego
 - Ewaluacja: stworzenie macierzy błędów/metryk modelu/współczynnik R^2 regresji itd.
- Zachęcam do wzorowania się na notebookach prezentowanych na wykładzie
- Projekt (notebook) najlepiej wysłać mi mejlem
- Termin do 18 czerwca