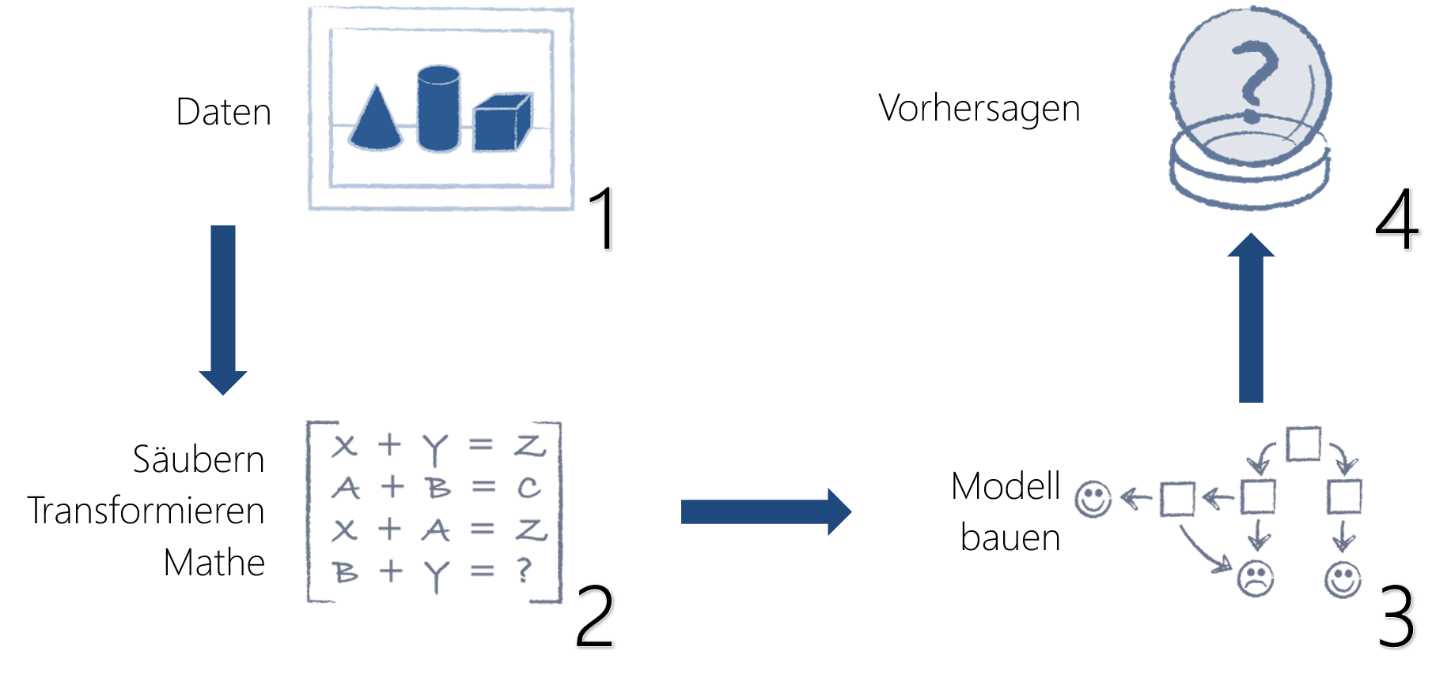
# 19. Was ist Data Mining? Was ist der Unterschied/Zusammenhang zu Machine Learning?

Unter Data Mining versteht man die systematische Anwendung von Methoden auf einen Datenbestand. Das Ziel ist die Erkennung von neuen Mustern. Beim Data Mining geht es primär um Datenbestände die so groß sind, dass sie nicht mehr von Menschenhand verarbeitet werden können.

Der Unterschied zu Machine Learning ist, dass beim Data Mining der Fokus auf dem Finden von neuen Mustern liegt. Im Gegenteil zu Machine Learning bei dem bekannte Muster in neuen Daten wiedererkannt werden sollen. Verfahren aus dem Machine Learning finden oft beim Data Mining Anwendung und umgekehrt.

# 20. Woraus besteht der Lernprozess in Machine Learning? Wie kann das Model verbessert werden?



Vor dem Modellieren sollte der Datenmüll gesäubert werden. Bei der Datenmodellierung sollten die einzelnen Spalten/Attribute/Features genauer untersucht werden, da nicht jedes davon interessant ist.

# 21. Was bedeutet Bias und was bedeutet Over-Fitting? Wie stehen die beiden Werte zueinander? Wie kann man Overfitting verhindern?

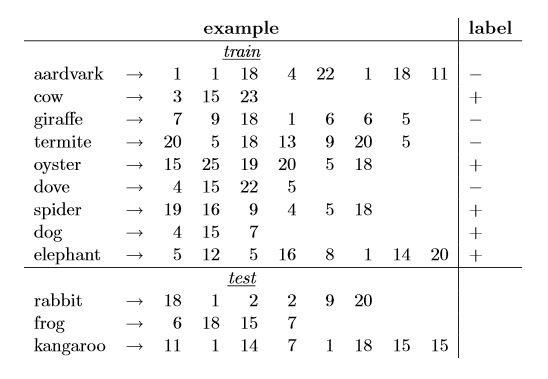
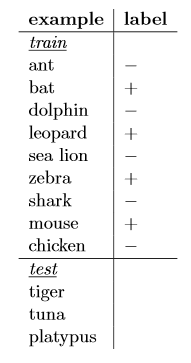
Beim Overfitting wird die Datenrepresentation zu sehr an die Trainingsdaten angepasst. Das kann entweder aufgrund von fehlerhafte Daten (Rauschen) oder zu wenig Datensätzen auftreten.

Bei parametrischen Modellen sollte versucht werden, eine geringe Anzahl an Parametern anzustreben.  
Bei nichtparametrischen Modellen sollte versucht werden die Anzahl der Freiheitsgrade im Vorhinein zu beschränken. Weiters können Methoden wie die Kreuzvalidierung oder v-fache Kreuzvalidierung zur Vermeidung von Overfitting angewandt werden. Modell so klein wie nur möglich halten.

# 22. Welche Schritte im Prozess sind anzuwenden um mit machine learning Algorithmen zu arbeiten?

# 23. Was sind Features und Examples in machine learning? Beschreiben Sie diese und erläutern Sie im Anschluss anhand eines konkreten Beispiels.

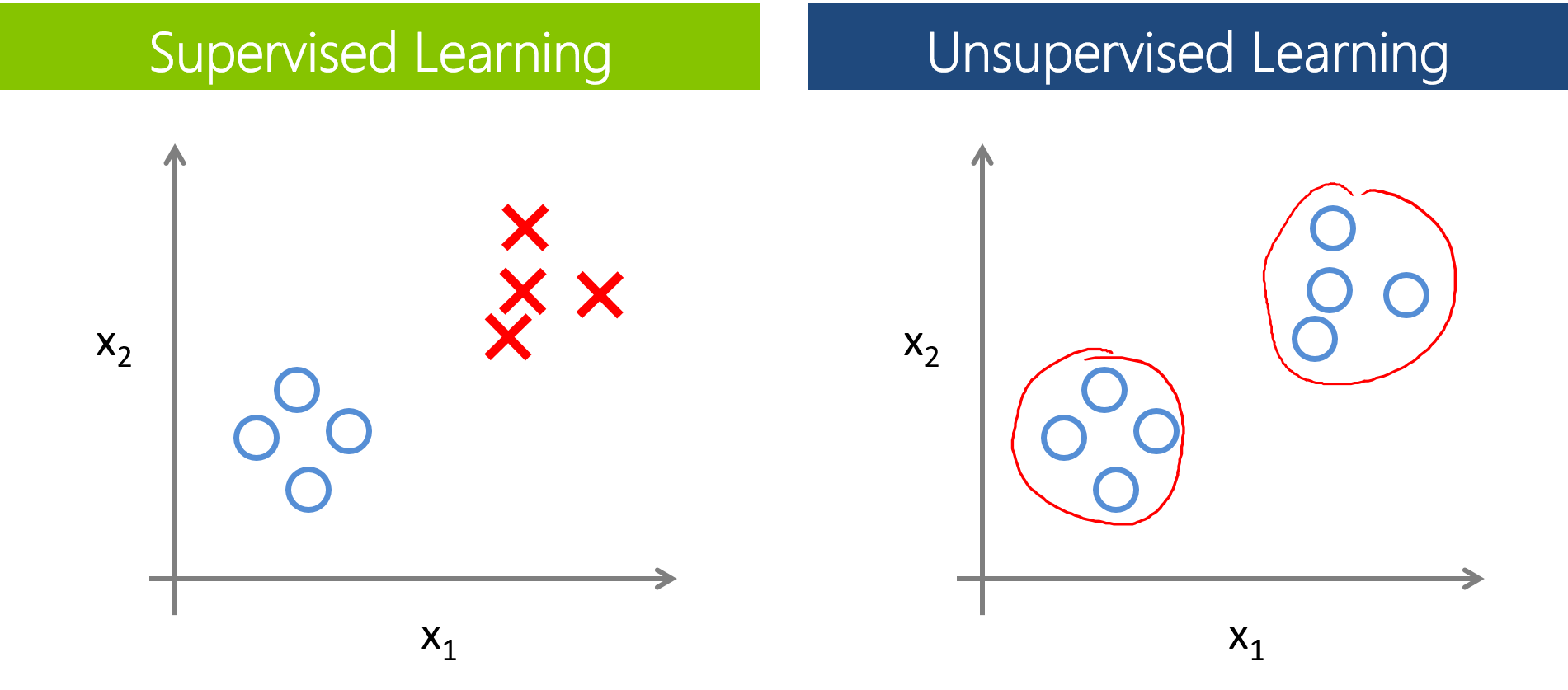
Ein Example ist ein Trainingseintrag.



Ein Feature ist eine Charakterisierung durch Attribut-Wert-Paare der Eigenschaften der Daten. Diese Features werden als Datenbasis für Algorithmen verwendet. Features werden verwendet um Daten besser zu verarbeiten.  
Beispiel >> Ein Feature beim Filtern vom Spam Mails könnte die E-Mail Struktur, grammatische Korrektheit oder die Sprache sein.

# 24. In welche 2 Kategorien können machine learning Algorithmen eingeteilt werden? Nennen Sie je ein Beispiel für eine konkrete Anwendung.

**Überwachtes Lernen:** Beim überwachten Lernen wird versucht Gesetzmäßigkeiten nachzubilden. Die Ergebnisse sind bekannt und werden verwendet um das System anzulernen. Die Ergebnisse des Lernprozesses werden mit den bekannten, richtigen Ergebnissen verglichen.  
Ziel: „Zukunft voraussagen“  
Beispiel >> Handschrifterkennung  
**Unüberwachtes Lernen:** Beim unüberwachten Lernen sind die Ergebnisse im Voraus nicht bekannt. Es wird versucht anhand der Eingaben Muster zu erkennen. Das System lernt durch die Eingabemuster und versucht diese in Kategorien zu klassifizieren (Clustern).   
Ziel: Vorhandene Daten zu vestehen  
Beispiel >> Klassifizierung von Münzen, Genomforschung

**Während wir bei Supervised Learning schon im Voraus** wissen**, dass es genau zwei Kategorien gibt (blauer Kreis oder rotes Kreuz),** ergeben **sich bei Unsupervised Learning zwei Gruppen.**