CRISP DM Summary:

**Was ist das Problem, das wir haben könnten?**

**Problem:   
DS/ML Projekte laufen Gefahr unwissenschaftlich zu werden und damit ist schnell unklar was genau gemacht wurde, warum, und vor allem mit welchen Ziel?   
Was wurde gemacht und warum?**

Zentraler Kern des wissenschaftlichen Arbeitsansatzes ist die Reproduzierbarkeit &   
Verifizierbarkeit.

* Was kann schief gehen?   
  <https://www.newyorker.com/news/john-cassidy/the-reinhart-and-rogoff-controversy-a-summing-up>

Das sollte recht einfach sein -> Code und Datenstatz runterladen, kompilieren und

laufen lassen.   
Warum ist es nicht so leicht??

Viel passiert im Doing! Man beginnt, testet Hypothesen, ändert variablen, ….

Bild von Bruegel

**Problem:  
Wo stehen wir, was haben wir noch vor uns, was wurde bereits gemacht? Sind wir sicher, dass wir uns in die richtige Richtung bewegen?**

CRISP DM als Dokumentationstool für den Kunden und Projektleiter:

* Es können sehr leicht, sehr präzise Updates an den/die Kunden weitergegeben werden -> Steigerung im Vertrauen, Entspannung etc.

* Falls sich das Projekt in eine falsche Richtung entwickelt, kann das grundsätzlich schnell(er) durch den Projektleiter/Kunden erkannt werden, da Handlungsprämissen im Projekt sehr transparent für den PL einsehbar sind und nach Vorlieben mit dem Kunden geteilt werden können

**Was ist die Lösung?**

* **Strukturierte, Standarisierte Prozesse**
* **Documentation & Kommunikation**

Gut, aber wer legte die Prozesse fest und warum sollten diese dann sinnvoll/hilfreich sein?

CRISP DM, das steht für   
CRoss Industry Standard Process for Data Mining

4 Abstractionslevel:

1. Phasen:  
   Bestehen aus generische Tasks
2. Generische Tasks:   
   Generisch, weil Struktur soll flexibel genug sein um alle Data Mining Situationen abdecken zu können. Dabei wurden diese Tasks nach den Prinzipen „vollständig“ und „stabil“ entworfen.   
     
    Vollständig in dem Sinne, dass es den ganzen DM Prozess abdeckt und stabil im Sinn, dass diese Tasks auch noch valide bleiben wenn neue, bisher unbekannte Entwicklungen wie z.B. neue Modellierungstechnologien auftauchen.
3. Spezialisierte Tasks

Kümmern sich um die konkrete ausführung der Generischen Tasks.  
Z.B. G: Model bauen -> S: Lineare Regression bauen:

1. Prozess Instanzen:  
   ist eine Dokumentation von Aktionen, Entscheidungen und Ergebnissen eines tatsächlichen Data-Mining-Einsatzes

Ursprünglich aus Team von:

* Daimler-Chrysler
* SPSS
* NCR

**Und wie setzt man das um?**

Mit, \*Trommelwirbelllllll\*

Laut KD Nuggets-Umfragen aus den Jahren 2007 und 2014 ist CRISP-DM die am weitesten verbreitete Methode für Data-Science- und Analyseprojekte: Über 40 % der Befragten gaben an, sie zu verwenden.

Auch laut wissenschaftlicher Literatur der Industrie Standard:   
https://www.sciencedirect.com/science/article/pii/S1877050921002416

A close-up of a document

Description automatically generated

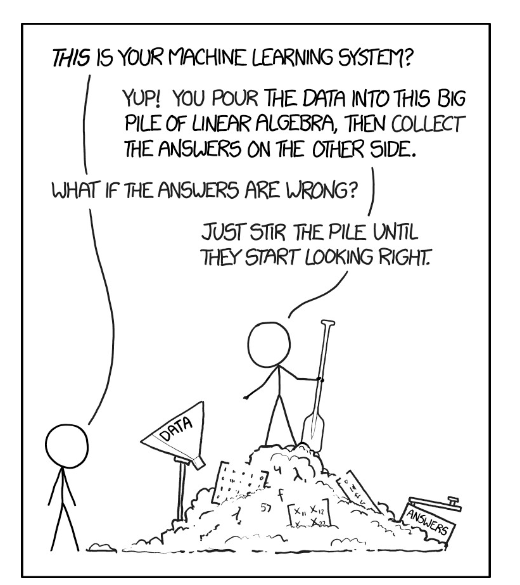
Anders als andere Prozessmodelle wie z.B.:

* Fayyad’s KDD
* SEMMA (von SAS)

Liegt der Fokus nicht nur auf die zentralen Data Science Tasks an sich, sondern beinhaltet den ganzen Projekt Prozess also auch Themen wie Deployment & Business Understanding:

A diagram of a data processing process

Description automatically generated

:

ISO 22989 als Ergänzung erwähnen!!

Das schöne ist, es können auch firmenweite standards für generalisierte tasks definiert werden. Z.B Standards für die Ermittlung bzw. den Umgang mit Outliern.

Das bild erst ganz zum Schluss anzeigen.

A table of data analysis

Description automatically generated with medium confidence

Tasks sind Fett markiert, Ergebnisse sind in kursiv markiert.

1. **Business Understanding**In dieser ersten Phase geht es darum, die Projektziele und -anforderungen aus der Unternehmensperspektive zu verstehen und dieses Wissen dann in eine Data-Mining-Problemdefinition und einen vorläufigen Projektplan zur Erreichung der Ziele umzusetzen.  
   1. **Business Objectives**
      1. **Ergebnis**: Verständnis des Projektziels und der Anforderungen bzw. des Problems des Kunden aus Unternehmensperspektive das gelöst werden sollte.

 Competing objectives and constraints

 Uncover factors that can influence the final outcome

 Risk: investing a great deal of effort producing the correct answers to the wrong questions.

* + 1. **Aktivitäten**
       1. Informelles Beschreiben des zu lösenden Problems
       2. Spezifizieren des erwarteten Nutzens aus geschäftlicher Sicht
* Beispiel:   
  Kunde möchte Gründe für Kundenabwanderungen kennen um Gegenmaßnamen treffen zu können

* + - 1. Zusatz durch EU AI Act: Risikoklassifikation vornehmen  
           
         A diagram of a triangle with different colored circles and black text

         Description automatically generated
  1. **Business Success Criteria**
     1. Ergebnis:
        1. Erfolgskriterien mit folgenden Eigenschaften:
           1. Spezifisch
           2. Messbar
           3. Allgemein
           4. Subjektiv (Angabe der messenden Person)
     2. Aktivitäten:
        1. Bestimmen der Erfolgskriterien
        2. Festlegen der Person(en) die die Erfolgskriterien bewerten
        3. Achtung: Jedes Erfolgskriterium sollte sich mindestens auf eine Business Objective beziehen ☺
* Beispiel: Ziel ist es, 70% der Kunden die in den nächsten 3-6 Monaten abwandern wollen, bereits frühzeitig zu identifizieren.
  1. **Data Mining Goals  
     Ergebnisse:**Beschreibung der Ergebnisse, die das Erreichen der Unternehmensziele ermöglichen
  2. - Kann die explizite Auflistung von Nicht-Zielen beinhalten (!)  
       
     **Aktivitäten**:   
     Übersetzen Sie die Geschäftsfragen in Data-Mining-Ziele

(z.B. eine Marketingkampagne erfordert eine Segmentierung der Kunden; die Ebene/Größe der Segmente sollte spezifiziert werden).

- Spezifizieren Sie den DM-Problemtyp (z.B. Klassifizierung, Regression, Clustering).

- Hinweis: Überprüfen Sie die Übereinstimmung zwischen Geschäfts- und DM-Zielen!

- Z.B. "Verbesserung der Qualität der Produkte" ->

"bei gegebenen Prozessüberwachungsdaten die Qualität des resultierenden Produkts vorhersagen"? "Wann kann ich bei gegebenen Prozessdaten eine Abweichung vom idealen Prozess erkennen? Korrekturen empfehlen?" ...

* 1. **Data Mining Success Criterions**Ergebnisse:  
       
     - Definieren Sie Kriterien f%C3%BCr ein erfolgreiches Ergebnis in technischer hinsicht- idealerweise objektiv, quantitativ- Bei subjektiven Kriterien die Personen identifizieren, die das Urteil fällen  
       
     **Aktivitäten**

- Festlegung von Kriterien für die Modellbewertung (z. B. Modellgenauigkeit, Leistung, Robustheit, Komplexität)   
  
- Definition von Benchmarks fürr Bewertungskriterien  
Legen Sie Kriterien fest, die subjektive Bewertungskriterien berücksichtigen (z. B. die Fähigkeit, ein Modell zu erklären)- Berücksichtigung von Einsatzaspekten  
  
Hinweis: Die Erfolgskriterien für Data Mining unterscheiden sich von den geschäftlichenErfolgskriterien!

1. **Data Understanding:**

**CRIPS DM als ML-Prozessframework:**

Laut einer Studie von Lee et al. (2015) erfüllen derzeit 75 bis 85 Prozent der Ergebnisse von ML-Projekten in führenden Technologieunternehmen nicht die Erwartungen der Auftraggeber.  
Auf die Frage nach der Ursache nennen Fischer et.al. (2021) dabei unter anderem die Daten- und Softwarequalität als die größten Herausforderungen im Lebenszyklus des maschinellen Lernens und als potenzielle Ursache für die häufig verfehlten Erwartungen.   
Ein weiterer Grund sei auch das Fehlen von Leitlinien in Form von Standards und Entwicklungsprozessmodellen die speziell für die gesonderten Bedürfnisse von ML-Anwendungen ausgelegt sind, denn vor allem größere Industrieunternehmen verlassen sich stark auf Standards, um eine gleichbleibende Qualität ihrer Produkte oder Dienstleistungen zu gewährleisten.

Solche Prozessmodelle, speziell für ML-Anwendungen sind aktuell aber nur bedingt verfügbar, weswegen viele Organisationen auf ML-verwandten Prozessmodellen wie beispielsweise CRISP DM (als Branchenstandard im Data Mining (Mariscal etl.al, 2010) [3]) zurückgreifen.

Wichtig ist jedoch dabei anzumerken, dass CRISP DM in seiner Urform nur bedingt für den Einsatz in Machine Learning Projekten geeignet ist.  
Die Gründe dafür werden aktuell in den folgenden vier Einschränkungen gesehen:

1. Fehlende **Qualitätssicherung**
2. Fehlende Ausrichtung auf die **Bedürfnisse/Herausforderungen des maschinellen Lernens**
3. Fehlende Monitoring- und Lösungsansätze für **Bias-Probleme**
4. Fehlende Berücksichtigung der Anforderungen des **AI-Acts.**

Laut Studer et. al (2021) [4] und Eberle (2023) [5] können dieses Einschränken jedoch beseitigt werden und die damit entstandene Weiterentwicklung von CRISP DM als neuer Industrie-Standard (unter dem Namen CRISP ML(Q)) für ML-Projekte eingesetzt werden.   
  
Der folgende Abschnitt fasst zu diesem Zweck, die aktuellen Einschränkungen von CRISP DM sowie die Weiterentwicklungs-/ Lösungsansätze der obigen Autoren zusammen:

%20x%3D%22890%22%20y%3D%22580%22%20width%3D%22542%22%20

# Fehlende Qualitätssicherung

CRISP-DM fehlt eine Anleitung zur Methodik der Qualitätssicherung (QS). Dieses Defizit zeigt sich insbesondere im Vergleich zu anderen Standards im Bereich der Informationstechnologie (wie beispielsweise dem IEEE Standard für die Entwicklung von Software-Lebenszyklus-Prozessen), aber auch bei alternativen Prozessmodellen für Data Mining (beispielsweise SEMMA von SAS).   
Denn anderes als bei Anwendungen im DM-Kontext wird bei ML-Projekten Qualität nicht nur über die Zweckdienlichkeit („fit to Purpose“) des Produkts definiert, sondern über die Qualität der Aufgabenausführungen in jeder Phase der Entwicklung einer ML-Anwendung.   
  
Dadurch wird sichergestellt, dass Fehler so früh wie möglich erkannt werden, um die Kosten in den späteren Phasen des Entwicklungsprozesses zu minimieren.

Dieses Problem kann jedoch durch die Einführung und Verankerung einer QS-Methodik im Prozessmodell behoben werden.

Dabei wird die vorgeschlagene QS-Methodik aus weit verbreiteten Standards zur Qualitätssicherung (z.B. IEEE 730-1998) übernommen, insbesondere aufbauend auf dem Prinzip des "Risk Based Thinking" (DIN EN ISO 9001). Der risikobasierte Prozess ist dabei explizit branchen- und anwendungsneutral gehalten und ist bereits frühzeitig in das ML-Projekt einbezogen, und zwar auf präventive Weise (im Gegensatz zum reaktiven Risikomanagement).

Dabei liegt der **Schwerpunkt der QS-Methodik** in erster Linie auf den technischen Aufgaben, die erforderlich sind, um den Nachweis zu erbringen, dass jeder Schritt im Entwicklungsprozess von ausreichender Qualität ist, um die Übernahme in die Geschäftsprozesse zu rechtfertigen.

%20x%3D%22890%22%20y%3D%22580%22%20width%3D%22542%22%20

# Fehlende Ausrichtung auf die Bedürfnisse/Herausforderungen des maschinellen Lernens

Bedingt durch seine ursprüngliche Ausrichtung konzentriert sich CRISP-DM (standardmäßig) auf Data Mining und deckt nicht das Anwendungsszenario von ML-Modellen ab, die über einen langen Zeitraum konstant qualitative Entscheidungen ableiten sollen (Abb.1).

A diagram of data mining process

Description automatically generated

Abbildung 1: Unterschied zwischen Data Mining und Machine Learning

Im Gegensatz zum DM muss sich das ML-Modell konstant an eine sich verändernde Umgebung anpassen zu können, um Peformanceabnahmen des Modells im Laufe der Zeit zu vermeiden.

Damit ist, anders als in DM Projekten nach der Einführung ein konstantes Monitoring und Wartung des ML-Modells erforderlich ist.

**Lösung**:

Dieses Problem kann durch die Entwicklung eines umfassenden Prozessmodells für praktische ML-Anwendungen gelöst werden, das alle relevanten Phasen im Lebenszyklus abdeckt. Dabei dient CRISP-DM als Grundlage, wird jedoch um zusätzliche, literaturgestützte Phasen erweitert.

Die Relevanz für ein Prozessmodell ist durch Standards im Bereich der Informationstechnologie motiviert, die sich in der Anwendung bewährt haben, aber keine ML-Spezifika abdecken (z.B. IEEE 1074-1997). Das Modell folgt den Prinzipien von CRISP-DM, indem es insbesondere branchen- und anwendungsneutral gehalten wird, wird aber an die besonderen Anforderungen von ML-Anwendungen angepasst.

# 

# eiFehlende Monitoring- und Lösungsansätze für Bias-Probleme

Bedingt durch die Ursprungsform von CRISP DM, gibt es im Standardmodell auch keine (expliziten) Prozessschritte die sich mit der Erkennung und Behandlung von Bias (Verzerrungen) auseinandersetzen. Mittlerweile, unter anderem durch die Forderungen des AI-Acts, sind ein entsprechendes Monitoring und Management-Strategien jedoch unerlässlich.   
Glücklicherweise gibt es jedoch in der aktuellen Literatur auch hier Ansätze um CRISP DM auch in diesem Aspekt um die Anforderungen der aktuellen Zeit zu erweitern.

**Bias Definition (Eberle, 2023) [5]**

- „...systematische unterschiedliche Behandlung bestimmter Objekte, Personen oder Gruppen im Vergleich zu anderen, wobei „Behandlung“ jede Art von Handlung ist, einschließlich Wahrnehmung, Beobachtung, Darstellung, Vorhersage oder Entscheidung“.

- „. . . jede Grundlage für die Wahl einer Verallgemeinerung gegenüber einer anderen, abgesehen von der absoluten Kohärenz mit den Instanzen “

- „... eine oder mehrere geschützte Fremdvariablen, die die Beziehung zwischen den Input- (unabhängigen) und den Output- (abhängigen) Variablen verzerren und somit zu falschen Schlussfolgerungen führen “

- „. . . der Unterschied in der zugrundeliegenden Verteilung des Modell-Lernergebnisses in Bezug auf eine bestimmte(n) Gruppe(n), die durch ihre Zugehörigkeit zu der spezifischen Gruppe beeinflusst wird. Bei der Gruppe kann es sich um das Geschlecht, die Ethnie, das Alter oder jedes andere geschützte Attribut handeln “

- „...Abweichung von einer Norm “

Allgemein gilt:

Ein Bias bedeutet nicht per se fehlender Fairness, jedoch kann ein unerwünschter/unbewusster Bias zu fehlender Fairness führen. Dementsprechend ist es, um möglichst fair agierende Modelle und Anwendungen zu entwickeln unerlässlich sich ihm Rahmen der Entwicklung und des Betriebs mit diesem Thema auseinander zu setzen.

Um effektive Bias-Erkennung und ggf. Vermeidungen durchführen zu können ist es an erster Stelle notwendig herauszufinden an welchen Prozess-/Betriebsschritten systematisch Bias auftreten können.   
Abbildung 2 liefert dazu einen Überblick und gibt an, in welche Interaktionen besonders häufig Quellen von Bias sind/sein können:

A diagram of a data flow

Description automatically generated

Glücklicherweise können dem folgende auch diese Interaktionen auf CRIPS DM Prozessschritte gemappt werden und so konkrete, Phasen der aktive Handlungserfordernis abgeleitet werden. Die daraus resultierenden Aktivitäten finden sich in der Arbeit von Eberle, 2023 [5].

# A yellow and white document with a checklist Description automatically generated with medium confidence

# Fehlende Berücksichtigung der Anforderungen des AI-Acts.

# Mapping des AI-Acts auf CRISP DM Phasen:

# Analog gibt es für (Teile) der AI Act Anforderungen auch Mappings auf die einzelnen CRISP DM Schritte.

A table of information

Description automatically generated

# CRISP ML(Q):

Daraus ergibt sich das folgende, gesamtheitliche Prozessmodel für ML Projekte mit

1. **Business and Data Understanding**
   1. Scope of Application:   
      Business Needs
   2. Measurable Success Criteria
      1. Business Success Criteria:
         1. Purpose and Success Criteria from a business point of view
      2. ML Success Criteria:
         1. Define minimum acceptable level of performance to meet the business goals
      3. Economic Success Criteria
         1. Objective ML Success by Introduction of KPI(s)
   3. Feasibility
      1. Applicability of ML Technology
      2. Legal Constraints
      3. Requirements on the application
         1. Robustness
         2. Scalability
         3. Explainability
         4. Resource Demand
   4. Data Collection
      1. Data Collection Plan
         1. In terms of Cost and Time needed to collect enough consistent data
      2. Data Version Control
         1. Data is collected iteratively, hence (planed) modifications of the data set should be documented
   5. Data Quality Verification
      1. Data Description
         1. Expert Knowledge regarding data sets like expected value ranges of features, maximum number of missing values. Guide to identify non-plausible data
      2. Data Requirements
      3. Data Verification
         1. Initial data, added data & production data must be checked according to the requirements
   6. Review of Output Documents
2. **Data Preperation**
   1. Select Data
   2. Select Features:
      1. Can be separated in three categories:
         1. Filter Methods
         2. Wrapper Methods
         3. Embedded Methods
      2. Would be good to have someone with expert knowledge have a look at it again.
      3. Data Selection
         1. Discarding features/samples should be well documented and strictly be based on objective quality criteria’s
      4. Unbalanced Classes
   3. Clean Data
      1. Noise reduction
      2. Data imputation
   4. Construct Data
      1. Feature Engineering
      2. Data augmentation
   5. Standardized Data
      1. File Format
      2. Normalisation
3. **Modelling**
   1. Literature Research on similar Problems
   2. Define Quality Measures of the model
      1. Robustness
      2. Explainability
      3. Scalability
      4. Resource Demand
      5. Model Complexity
   3. Model Selection
   4. Incorporate domain knowledge
   5. Model Training
   6. Model Compression
   7. Ensemble methods
   8. Result reproducibility
   9. Experimental Documentation
4. **Evaluation:** 
   1. Validate Performance
      1. Come up with a plan to validate the performance
   2. Determine robustness
   3. Increase Explainability
   4. Compare results with defined success criteria
5. **Deployment:**
   1. Define inference hardware
   2. Model evaluation under production condition
   3. Assure User Acceptance and usability
   4. Minimize risks of unforeseen errors
   5. Deployment strategy
6. **Monitoring and Maintenance:** 
   1. Identify Risks for performance degradation:
      1. Non-stationary data distributions/data drift
      2. Degradation of hardware
      3. System updates
   2. Monitor
      1. Monitor all input signals and compared to trainings data -> in this way, updates in the input data could be caught. What to do with anomalies in the input?
      2. Monitoring History of Performance
   3. Update
7. Nimdzi Insights. Artificial Intelligence: Localization Winners, Losers, Heroes, Spectators, and You. Available online: https: //www.nimdzi.com/wp-content/uploads/2019/06/Nimdzi-AI-whitepaper.pdf (accessed on 21 April 2021).
8. Fischer, L.; Ehrlinger, L.; Geist, V.; Ramler, R.; Sobiezky, F.; Zellinger, W.; Brunner, D.; Kumar, M.; Moser, B. AI System Engineering—Key Challenges and Lessons Learned. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 56–83. [CrossRef]
9. Mariscal, G.; Marbán, O.; Fernández, C. A survey of data mining and knowledge discovery process models and methodologies. *Knowl. Eng. Rev.* **2010**, *25*, 137–166. [CrossRef]
10. Studer S., Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology [[CrossRef](https://arxiv.org/abs/2003.05155)]
11. Eberle, F.; [Systematic extension of CRISP-DM by structured mapping of emerging regulatory requirements on bias in AI](https://repositum.tuwien.at/handle/20.500.12708/189389), [[CrossRef](https://repositum.tuwien.at/handle/20.500.12708/189389)]