CRISP DM Summary:

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

**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

**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 PL:

* 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

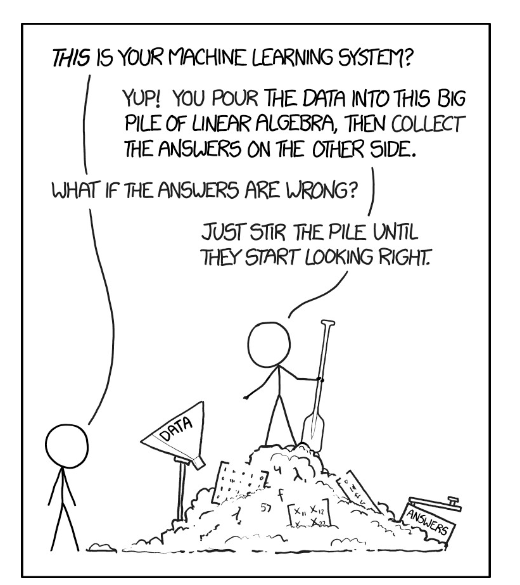
Anders als andere Prozessmodelle wie z.B.:

* 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

Phasen:

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:**

**Current Shortcomings:**

In Ursprünglicher Version gibt es nichts zu Quality Assurance

ML ist auch nicht Thema!

In der Ursprünglichen Variante gibt es keine Guidelines für die Identifikation/Vermeidung von Bias(en) und damit auch kein generisch, systematischer Zugang.

**Definition Bias:**   
Systematische Fehler im Urteilsvermögen und in der Entscheidungsfindung, die auf kognitive LImitationen, Motivationsfaktoren und/oder Anpassungen an die natürliche Umgebung zurückzuführen sein können.   
  
Beispiel: Stereotyping

**Drei Kategorien nach Mehrabi et.al (2021)**

**Siehe plot**

Für Detailierte Beschreibungen, siehe: Link zu Masterarbeit

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

But what about ML?   
  
 A diagram of data mining process

Description automatically generatedheight%3D%22400%22%20as%3D%22geometry%22%2F%3E%3C%2FmxCell%3E%3C%2Froot%3E%3C%2FmxGraphModel%3E

# CRISP ML(Q):

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.

(*checklist from Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 2003, 3, 1157–1182.)*  
  
  
1. **Do you have domain knowledge**? If yes, construct a better set of “ad hoc” features.   
  
2. **Are your features commensurate**? If no, consider normalizing them.   
  
3. **Do you suspect interdependence of features**? If yes, expand your feature set by constructing conjunctive features or products of features, as much as your computer resources allow you (see example of use in Section 4.4).  
  
 4. **Do you need to prune the input variables** (e.g. for cost, speed or data understanding reasons)? If no, construct disjunctive features or weighted sums of features (e.g. by clustering or matrix factorization, see Section 5).   
  
5. **Do you need to assess features individually** (e.g. to understand their influence on the system or because their number is so large that you need to do a first filtering)? If yes, use a variable ranking method (Section 2 and Section 7.2); else, do it anyway to get baseline results.   
  
6. **Do you need a predictor**? If no, stop.   
  
7. **Do you suspect your data is “dirty”** (has a few meaningless input patterns and/or noisy outputs or wrong class labels)? If yes, detect the outlier examples using the top ranking variables obtained in step 5 as representation; check and/or discard them.   
  
8. **Do you know what to try first?** If no, use a linear predictor.3 Use a forward selection method (Section 4.2) with the “probe” method as a stopping criterion (Section 6) or use the `0-norm embedded method (Section 4.3). For comparison, following the ranking of step 5, construct a sequence of predictors of same nature using increasing subsets of features. Can you match or improve performance with a smaller subset? If yes, try a non-linear predictor with that subset.   
  
9**. Do you have new ideas, time, computational resources, and enough examples?** If yes, compare several feature selection methods, including your new idea, correlation coefficients, backward selection and embedded methods (Section 4). Use linear and non-linear predictors. Select the best approach with model selection (Section 6).   
  
10. **Do you want a stable solution** (to improve performance and/or understanding)? If yes, subsample your data and redo your analysis for several “bootstraps” (Section 7.1)

* + 1. Data Selection
       1. Discarding features/samples should be well documented and be strictly be based on objective quality criterias
    2. Unbalanced Classes
  1. Clean Data
     1. Noise reduction
     2. Data imputation
  2. Construct Data
     1. Feature Engineering
     2. Data augmentation
  3. Standardized Data
     1. File Format
     2. Normalisation

1. **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
2. **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
3. **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
4. **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